

**Modelling and Simulation for Power
Distribution Grids of 3D Tiled
Computing Arrays**

Pakon Thuphairo

PhD

University of York

Computer Science

April 2023

| /

Abstract

This thesis presents modelling and simulation developments for power distribution grids of 3D tiled computing arrays (TCAs), a novel type of paradigm for HPC systems, and tests the feasibility of such systems for HPC systems domains.

The exploration of a complex power-grid such as those found in the TCA concept requires detailed simulations of systems with hundreds and possibly thousands of modular nodes, each contributing to the collective behaviour of the system. In particular power, voltage, and current behaviours are critically important observations.

To facilitate this investigation, and test the hypothesis, which seeks to understand if scalability is feasible for such systems, a bespoke simulation platform has been developed, and (importantly) validated against hardware prototypes of small systems.

A number of systems are simulated, including systems consisting of arrays of 'balls'. Balls are collections of modular tiles that form a ball-like modular unit, and can then themselves be tiled into large scale systems. Evaluations typically involved simulation of cubic arrays of sizes ranging from 2x2x2 balls up to 10x10x10.

Larger systems require extended simulation times. Therefore models are developed to extrapolate system behaviours for higher-orders of systems and to gauge the ultimate scalability of such TCA systems. It is found that systems of 40x40x40 are quite feasible with appropriate configurations.

Data connectivity is explored to a lesser degree, but comparisons were made between TCA systems and well known comparable HPC systems, and it is concluded that TCA systems can be built with comparable data-flow and scalability, and that the electrical and engineering challenges associated with the novelty of 3D tiled systems can be met with practical solutions.

Author's declaration

I declare that this thesis is a presentation of original work and has been done in collaboration with the following:

Pakon Thuphairo: Under supervision for the design and implementation of the models and simulation framework, and prototype discussions, in this thesis.

Christopher Crispin-Bailey (Supervisor): Previous theoretical investigation of TCA concept, contributions to prototype designs and measurements, assisted with the design entry and 3D-printing of hex-tile frames.

Anthony Moulds: Senior experimental officer, assisted with the prototype PCB implementation and electrical measurements due to the requirements for a high wattage bench-testing setup.

Jim Austin: We acknowledge the previous work and patents as published by Professor James Austin (Retired).

This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Aspects of the tile modelling and a brief introduction of the simulation framework in this thesis, have been published in the following paper by this PhD candidate and co-authors:

continued...

P. Thuphairo, C. Bailey, A. Moulds and J. Austin, "Investigating Novel 3D Modular Schemes for Large Array Topologies: Power Modeling and Prototype Feasibility," 2022 25th Euromicro Conference on Digital System Design (DSD), Maspalomas, Spain, 2022, pp. 268-275, doi: 10.1109/DSD57027.2022.00044.

"In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of the University of York' s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation."

Submitted manuscripts and additional technical reports relating to the Hex-tile project are given in Appendix A.

Acknowledgements

First of all, I would like to express thanks to my supervisor, Dr Christopher Crispin-Bailey, for his continuous supervision, guidance, efforts, and support for many things from the beginning of my PhD journey with the COVID-19 pandemic starting just after my first PhD year. Without him, it would have never led to this thesis completion.

Anthony Moulds, a senior experimental officer, who joint the TCA project during the COVID-19. I would also like to thank him, and also with my supervisor, in the collaborative work on significant parts in building hardware prototypes. I really appreciate not only his collaboration in the hardware prototypes, but also the board model with the switching regulator employed in the prototypes, and his advice regarding SPICE designs and simulations.

In addition, I would also thank all the departmental members of technical staff involved with assembly of the hardware prototypes and anything in the research project. In fact, not only the departmental level, much appreciation will also go to the University of York's IT support team who constantly helped me with simulation-related resources, (IGGI machines), which more than one time, went wrong unintentionally by myself. Viking - University of York Research Computing Cluster), was also part of the simulation activities in this thesis. Unfortunately, with the limited time-frame and some complications in the simulation issues, the cluster simulation framework is still in an experimental stage. However, the experience of interacting with a high performance computing cluster is extremely valuable.

continued...

I would also like to thank the Royal Thai Government for awarding me a full PhD scholarship. Some parts of this scholarship also contribute to partial components in the hardware prototypes and also the equipment for some hardware experiments in this thesis.

I would also like to thank everyone from Rajamangala University of Technology Ratanakosin, the university I have been working for since before I started my PhD degree, who has supported me for all the preparation processes before coming to the UK for my PhD course.

Apart from the above, I wish to thank my father and mother who have always been supporting me with many things since birth. Without their kindness, it would have never been me so far. I would also like to thank everyone else both in Thailand and the UK that helped me before and throughout years of my PhD journey.

List of Contents

List of Contents

- 1 Introduction 1**
 - 1.1 Synopsis 1
 - 1.2 Motivation 2
 - 1.3 Thesis Scope 3
 - 1.4 Research Hypothesis and Objectives 4
 - 1.4.1 Definition 5
 - 1.4.2 Research Objectives 6
 - 1.5 Contributions 8
 - 1.5.1 Methodology 8
 - 1.5.2 Findings 9
 - 1.6 Thesis Organisation 10

- 2 Field Survey and Literature Review 11**
 - 2.1 Existing Concepts and Principles Relating to TCA 15
 - 2.1.1 Short Introduction to TCA 15
 - 2.1.2 TCA-concept Definition 16
 - 2.1.3 Geometrical Properties of Ball Arrays 17
 - 2.2 Mapping Between Logical and Physical Topologies 20
 - 2.3 Parallel/Distributed Computer Packaging 25
 - 2.3.1 Rack-mount Packaging 25
 - 2.3.2 Non rack-mount Packaging 27
 - 2.4 TCA Physical Scalability: Possible Constraints 36
 - 2.4.1 Power Networks in TCA Systems 36
 - 2.4.2 Physical Engineering Factors 37
 - 2.5 Related Concepts for Power Modelling 39
 - 2.5.1 Overview of Power Modelling in Parallel/Distributed Computers 39

2.5.2	Typical Inter-board Level Power Delivery in Parallel/Distributed Computers	40
2.5.3	Power Modelling at the Computing-board Level	42
2.5.4	Voltage Regulators	43
2.5.5	Switching Regulator: Average VS Complex Models	46
2.6	SPICE Simulators	48
2.7	Mapping Tasks Into a Parallel/Distributed Machine	48
2.8	Summary and Implications for Hypothesis	50
3	Hardware and Model Development	53
3.1	Relevant Research Objectives	54
3.1.1	Objective 1: Employing and Designing Models and Simulation Tools	54
3.1.2	Objective 2: Hardware Validation	55
3.2	Hardware Prototypes	57
3.3	Arbitrary Electrically-conductive Media Designs	62
3.4	Prototype Models	67
3.4.1	Tile Modelling	67
3.4.2	Connector-pin Resistance Modelling	68
3.4.3	Board Modelling	70
3.4.4	Voltage Regulator Modelling	76
3.4.5	Regulated Load Modelling	77
3.4.6	Ball-array Modelling	77
3.4.7	Power Source Modelling	79
3.5	Validation Work and Results	79
3.5.1	Study 1) Hardware Versus System Modelling	79
3.5.2	Study 2) Simplified Versus Switching Models	81
3.6	Chapter Summary	85
4	TCA Power-grid Simulation Tools	87
4.1	Relevant Research Objectives	87
4.1.1	Objective 1: Employing and Designing Models and Simulation Tools	87

4.1.2	Objective 4: Optimised Power Distribution	89
4.1.3	Objective 6: Simulation Framework Documentation	90
4.2	Overall Simulation Framework	91
4.3	Tile Naming Convention	93
4.3.1	Facet Naming Convention	94
4.3.2	Edge Numbering Convention	94
4.3.3	Further Detail	94
4.4	Connector-pin Terminal-name Replacement Convention	96
4.5	System Generator	97
4.5.1	Ball-array Generator	98
4.5.2	External Power/ground Rails Renaming	101
4.6	Manual SPICE-file Editing	102
4.7	Simulation Modules	103
4.7.1	Power-network Simulator	103
4.7.2	Power Allocation Schemes	109
4.7.3	Uniform Power Simulator	118
4.7.4	Non-uniform Power Simulator	118
4.8	SPICE Simulator	128
4.9	Chapter Summary	128
5	Scalability Evaluations	131
5.1	Relevant Research Objectives	133
5.1.1	Objective 3: Fundamental Simulation Experiments	133
5.1.2	Objective 4: Optimised Power Distribution	135
5.1.3	Objective 5: Scalability Evaluations	136
5.2	Uniform Power Allocation	137
5.3	Non-uniform Power Allocation	150
5.3.1	Brute-force Simulation for Relative-position Scheme	150
5.3.2	System-level Regulated Load-power and Connector-pin Current Optimisations	155
5.3.3	System-level Power Efficiency	163
5.4	Total System Power	169

5.5	Preliminary Topological Analyses, Simulations, and Comparisons . . .	172
5.6	Comparison of Large-scale Traditional Rack-mount Systems with TCAs	179
6	Conclusions and Future Work	183
6.1	Conclusions	183
6.1.1	Research Hypothesis and Objectives	183
6.1.2	Contributions	188
6.1.3	Tool Design Experiences	189
6.1.4	Limitations and Assumptions	191
6.2	Possible Future Work	192
6.2.1	Lower Hop Counts	193
6.2.2	Investigation of External Power Designs	193
6.2.3	Arbitrary Inter and Intra-unit Level Power Media	194
6.2.4	Node Power Model	194
6.2.5	Multiple Optimisation Algorithms for Power Allocation	194
6.2.6	Improved Visualisations	194
6.2.7	System Computing Performance Analysis	195
6.2.8	Hardware Prototype Improvements	195
6.2.9	Simulation on Computing Cluster	195
6.2.10	Cooling Systems	196
6.2.11	Other Engineering Concerns	196
6.3	Final Remarks	196
	Appendices	199
A	Published Work	201
A.1	Submitted Manuscript of DSD2022 Paper	202
A.2	Short Report for Internal Funding Award (Hex-tile Project)	211
A.3	Hex-tile PCB Layouts	217
B	Example SPICE File	219
	References	223

List of Figures

- 1.1 Mapping a couple of separate groups of tasks into different packaging systems. (b) is a representation of the two groups mapped into three horizontal computing-boards, each with 9 PEs, in a rack-mount system. (c) represents a mapping of the tasks into 27 tiles in a TCA system. . . . 2

- 2.1 Energy efficiency trends reported by [9]. Each year in the graph is a combination of biannual 10 first-rank reports in June and November. 13

- 2.2 Examples of large scale system cooling, showing (a) TPU v3 unit, (b) TPU v3 Pod. (reprinted from Figure 1(b-c) in [10]) 14

- 2.3 Submer Immersive cooling system. (reprinted from [11]) 14

- 2.4 Example of a TCA design proposed as 'Computing devices' in [12]. (a) shows tile structure, (b) shows ball structure, composed from tiles. (adapted from Figure 1 and 5 in [12], respectively) 15

- 2.5 Logical views of a 2x2x2 3D-mesh and 3D-torus topologies. (a) depicts a 3D-mesh topology (adapted from [13]), and (b) depicts a 3D-torus topology [13].) 17

- 2.6 Example implementation alternatives of the TCA concept. (a) shows 1D tile-level array of 4 modules, whilst (b) shows a group of tile-level 2D array of 2x2 modules, and a possible 3D construction from the tile level can be seen in (c) as a complete 3D array of 2x2x2 modules. Instead of tile-level composition, each module may also be alternatively implemented as a single ball-shaped object. (d) depicts a row of 1D array of 4 modules. (e) constructs a 2D-array of 4x4 modules, whilst a 3D-array of 4x4x4 modules can be constructed as shown in (f). 18

- 2.7 A logical (machine-packaging independent) computing topology of 10 tasks. 21

2.8	Black dots represent occupied nodes, and green dots are the tasks being mapped. (a) Mapping a set of tasks on a plane of 3x3 nodes, and a single node on a different board in a rack-mount system. (b) The same mapping on a TCA system with the same system size. As the last task is mapped on an adjacent node above, and with the availability of a channel in the physical third dimension, the logical computing topology can be maintained.	22
2.9	The shortest route of a data unit travelling from node 2 to node 20 between computing boards in a rack-mount system, resulting in a hop count, at least, of 6. It is assumed that all the nodes at the far-end of each board are directly connected to a routing system	23
2.10	The shortest route of a data unit travelling from node 2 to node 20 between a couple of adjacent balls in a TCA system (node 2 and node 11 reside in the same ball, as a ball consists of eight tiles), resulting in a hop count of 2. A single hop count can be achieved on 3D-torus implementation with a wrap-around channel.	24
2.11	Timeline of four recent examples of Exascale projects. The period of each project duration can be found in [20], [21], [16], [22].	26
2.12	Frontier supercomputer. (reprinted from [26])	26
2.13	A conceptual optical realization of the space-invariant five-cube network. (reprinted from Figure 7(b) in [31])	29
2.14	Sketch of the HAEC Box. (reprinted from Figure 2(a) in [33])	29
2.15	An illustration of a cluster of ball-shaped computing devices. (reprinted from [35])	31
2.16	Wire-free power transmission media. (reprinted from Figure 3.1 in [36])	32
2.17	Timeline of TCA concepts and its prior work in the series of development. All the details in the timeline can be found in [35], [36], [39], [34], [40], [41], [14], [12], and [42]. ((1)-(2) reprinted from Figure 7 in [14], and Figure 8(c) in [42], respectively)	33
2.18	A panorama of the SpiNNaker 1 million core machine. (reprinted from [43])	34
2.19	Example water cooling system for the ball-shaped computing devices discussed in Subsection 2.3.2. (reprinted from Figure 8 in [35])	38
2.20	Traditional data centre supply system. (reprinted from Figure 1 in [54])	41

2.21	Examples of power distribution units. (a) shows a power strip for AC power, and (b) shows two rails of back-plane PDUs which can be found in a rack or cabinet. ((b) is reprinted from [55])	41
2.22	Examples of switching regulator efficiencies with varying input voltages and load currents. (reprinted from 'Typical Performance Characteristics', page 4 in [65])	45
2.23	Example of grounding issue for a voltage regulator simulation model on a 2x2 conceptual computing boards in a TCA. Each of the resistors surrounding each module represents inter-node power medium resistance. (a) shows a correct modelling, whilst in (b), the ground of each regulator model is directly tied to the global ground of the SPICE simulation. (c) shows the global positive and negative (ground) rails. This example is only for illustrative purpose. In the actual TCA systems, it is considered much more complex due to the 3D meshed power network.	47
3.1	(a) shows top and bottom views of a hex-tile prototype. (b) shows a half-ball (petal) composition. (c) shows eight tile-frames composed as a ball upon a power base-plate providing power via the trapezoidal faces. (d) illustrates a ball with two tiles removed, being powered and demonstrating different power loading by LED colours. (reprinted from Figure 8 in [42])	58
3.2	Illustrations of conceptual designs and a hardware prototype. (a) illustrates a conceptual model of the tile frame. Examples of possible materials are plastic or ceramic. In (b), the tile frame is shown with an embedded PCB or Multi-Chip-Module (MCM). As shown in (c), a tile may also be covered with a partially transparent material, allowing the visibility of components inside. (d) shows a possible data I/O connectivity upon each tile edge, represented by the dashed/blue lines, whilst solid/red arrows show power and ground rails. Having composed a group of eight tiles in 3D, it can make up a truncated octahedron, a ball-like volume shown in (e). Finally, (f) illustrates a ball-frame prototype. (reprinted from Figure 1 in [42], (e) is adapted from [71])	59

3.3	Example of a double packed internal array of $2 \times 2 \times 2$ grey balls embedded in between the existing $3 \times 3 \times 3$ array. Some of the balls are removed to expose the internal balls. (reprinted from Figure 2(b) in [42])	60
3.4	Example DC conduction analysis (Electrical Potential) on a hexagonal-shaped conductive medium. (a) shows a frame view with face numbers. (b) gives an electrical potential distribution on the object. In this particular case, an electric potential of 5V is applied on a single rectangular edge, which can be obviously seen in the red area. All the other edges are applied with an electric potential of 0V.	64
3.5	Example DC conduction analysis (Current Density) on a hexagonal-shaped conductive medium. (a) shows only the y-component of the current density, and (b) combines all the x, y, and z components, of the hexagonal medium. As this medium shape is a thin 3D object, it can be roughly considered that most of the currents flowing in this object are only in x and y axes.	65
3.6	Experimental custom electrically conductive medium shape modelling. a possible design flow starts from modelling a custom shape, followed by (a), electrical potential/DC conduction analysis, and finally building an equivalent lumped-resistor network. (b) illustrates an edge of the object showing an area for current flows. (c) shows a vector-field plot of current flows in the object. This experimental design can be part in the future simulation framework.	66
3.7	(a) shows conceptualised tile model. (cropped from Figure 4 in [42]). The resistors named <code>r_p_resist</code> and <code>r_g_resist</code> represent positive and ground rails of an edge power-connection. This inter-tile electrical medium resistance model can also be found in Figure 3.9. In (b), a legend describes the rest of the simulation components.	68
3.8	Comparison of a tile prototype with the conceptualised tile model (cropped from Figure 4 in [42]). Although the voltage and current measurement modules are added for simulation purposes, these modules can also be optionally implemented in a physical tiled computing unit for power management purposes.	69

3.9 Connector-pin resistance model. The resistor symbol named 'contact' partially framed with the red dashed line does not exist in the actual model used in a SPICE simulation file, but added in this figure for illustration purposes. The contact resistance occurring between a mated pin-pair can be split into halves and equally be added to the 'internal' resistances on both sides for simulation purposes. A photograph of the red-edge connector detachable housing taken from the prototypes. (reprinted from Figure 3 in [42]))	70
3.10 Board model with the switching regulator model.	73
3.11 Overall mechanism of the circuit-based board resistance adjuster during a SPICE simulation.	74
3.12 Simulation results used for validating the simplified board model (curve-fitted model), compared against the same board model using the full manufacturer's precise LT [®] 3976 regulator SPICE representation, as shown in Figure 3.10. In this validation, a system of 3x3x3-ball was used for both simulations, as this is the smallest ball-array to contain at least an inner ball to reflect voltage drops. (a) shows the simulation result of the full prototype-board model, and (b) shows the simulation result of the simplified board model. (reprinted from Figure 5 in [42])	84
4.1 Overall simulation framework. Some details in the automated area are omitted and encapsulated for a concise view of the whole framework. . .	93
4.2 Visualisation of tile naming convention. (a) visualises the top view of a ball, whilst (b) is for the bottom view, where Facet 'E' is directly underneath Facet 'A' when a ball is viewed from above, and the same positions for the rest. The square at the centre is one of the six holes for cooling. p and n are the names for positive and negative (ground) rails on a trapezoidal facet.	95

4.3	Example of connector-pin resistance terminal name change of a ball group. (a) shows only a couple of resistor names. In (b), a newly added ball is generated with its own resistor names at the edge to be coupled. In (c), The names of the new-ball resistors replaced with the ball to which it is coupled, finally modelling a new ball coupled to an existing system.	97
4.4	Step-by-step example of balls coupled in a SPICE-file of a system generation, starting from (a) towards a complete system in (h), respectively. When a ball is added, the SPICE node names of the connector-pin resistor model at the ball-edge connected to a previously generated ball in each dimension will be replaced.	99
4.5	4x4x4-ball connection cases during a SPICE-file generation, and a legend table. Except a transparent circle illustrating the first node generated at the coordinates (0,0,0), when adding a node represented by a unique colour/symbol shown in the legend table, it will be coupled with at least one existing node, depending on which location it is added to. For instance, a red circle node is coupled to the existing nodes in X, Y, and Z dimensions.	100
4.6	Function calls of TCA system generator and external voltage-source renaming.	101
4.7	External rails renaming. In (a), this particular case, a single voltage-source of 12V is applied to all of the external pins (some external pins are invisible due to a 2D surface illustration). In real implementation, multiple pins at each edge can be used for each power/round rail to allow more current tolerance. (b) shows an incomplete code-snippet of the generated system.	102
4.8	An LTspice [®] implementation of the whole board model with the circuit-based board-resistance adjuster in this thesis. This implementation is referred to as 'adjuster' published in [42], and supports only one regulated load-resistance (power) during a simulation. Thus, a separate simulation board-model file is needed for another load power consumption. Except 'r_board_resistance', the rest of the block modules and lines of code are part of the adjuster. '+' sign at the beginning is for the continuation of the line. The line with 'b_i_diff' implements the curve-fitting equation. .	106

4.9	Illustration of the last three data-points of board resistance for calculating 'rate of change' in the variable R-step mode. (a) shows the case that the third data-point resides in the bounding up range, resulting in reverting the latest Rstep to the previous one. (b) is the acceptable range to double the Rstep value, accelerating the reduction of board-resistance to more quickly approach the board input voltage-current profile.	109
4.10	Possible hierarchical power allocation schemes in TCA.	110
4.11	Visualisations of a couple of relative-position schemed TCAs. (a) shows a 3D-view of a 4x4x4-tile, and (b) shows only a 2D-surface of a 6x6x6-tile due to the abundant quantity of relative coloured-groups of internal 3D layers. The same colours in each of the arrays in (a) and (b) represent the same allocated amount of power.	111
4.12	Example showing that a two-point distance in a TCA is not valid for checking whether nodes (tiles) are equally impacted by external fully connected power. Even though the two coordinates of (0,2,2) and (1,1,1) have the same two-point distance of 2.5981 units from the system-centre coordinates (2.5, 2.5, 2.5), the first node is in the outermost layer, whilst the second node is in the first inner one. It is noted that the coloured nodes in this example show cubic layers, whilst Figure 4.11 explains a different subset node grouping, relative-position scheme.	112
4.13	Example three cubic layers of nodes as a possible power allocation scheme for TCA systems.	113
4.14	Diagram showing relations amongst power allocation schemes for cubic-array systems.	113
4.15	TCA system illustrating voltage drops over the entire system. In this particular system, for an illustrative purpose a ball is implemented as the smallest unit. Whilst this thesis focuses on a tile as the smallest unit. . .	116
4.16	Examples of different system shapes. (a) simple 3D mesh. (b) non-complete 3D mesh with partial outer layer nodes removed. (c) pyramid shaped system (d) sphere (e) simple 3D mesh similar to (a) but using a double-packed array (pink nodes packed between yellow), and showing the flow channels for cooling highlighted in blue.	117

4.17 Function calls of TCA system uniform power simulator.	119
4.18 Function calls of TCA system non-uniform (GA) power simulator.	119
4.19 Mechanism of the objective 1 and 2 associating with each other for SPICE-simulation instances.	122
4.20 (a) TCA multiple-objective simulation result file name and extension format. (b) examples of generated result files. (c) the content format of a GA report file. Due to a long length of chromosome (the tile regulated-side load-resistance values), some of the values are omitted. The last line with a numerical value reports the worst-case maximum pin current. The line with 'areConstraintsPassed = true', can be used for post processing to check whether any constraint, e.g., voltage drop is within the acceptable range.	122
4.21 Conventional mapping of tile-by-tile power of a 2x2x2-ball to a chromosome in the genetic algorithm employed.	124
4.22 The reduction of the chromosome size compared to the mapping in Figure 4.21 when using relative-position scheme. Further more reductions can also be seen in larger sizes shown in Figure 5.11.	125
4.23 Example list of the MATLAB® gamultiobj's parameter named 'options' [85] used for the 6x6x6-ball system simulated in this thesis.	126
4.24 Workflows of uniform and non-uniform power allocation simulations.	127
5.1 Estimated best and worst-case voltage drop simulations for uniform power-allocation with 101mA assumed supply side 12V fan load.	138
5.2 Estimated maximum connector-pin currents for uniform power-allocation with 101mA assumed supply side 12V fan load.	139

5.3	Predicted estimated worst connector-pin currents for 1W regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. To estimate the feasibility of a TCA system, this connector-pin currents report can be used together with the worst-case voltage drops report as shown in Figure 5.6, which represents the difference between the supply voltage at all of the surface power connectors of the grid array (12V in this case) and the worst-case voltage drop (loss) of all of the board input voltages. A combined plot can also be seen in Figure 5.8.	144
5.4	Predicted estimated worst connector-pin currents for 5W (a) to 25w (e) regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. The sizes in each case are limited when the the worst connector-pin currents exceed the current limit of 6A.	145
5.5	Detailed processes for predicting arbitrary regulated load wattages and ball cubic-size systems. (a) shows an example of simulated-data preparation for a desired regulated load wattage of 12.5W. The sizes of 2x2x2 to 9x9x9 are omitted. (b) details the process 1 generating 10 interpolated worst connector-pin currents. (c) shows process 2, with an example of extrapolation of a couple of desired cubic sizes of 11x11x11 and 12x12x12-ball systems (shown in green) from the data generated by process 1 (shown in red).	146
5.6	Predicted worst-case voltage drops for 1W regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. To estimate the feasibility of a TCA system, this voltage drops report can be used together with the worst-case connector-pin currents report as shown in Figure 5.3, which represents the highest current experienced from all of the lumped resistors modelling single (or parallel) tile-edge power (or ground) pins. A combined plot can also be seen in Figure 5.8.	147

5.7	Predicted estimated worst voltage drops for 5W (a) to 25w (e) regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. The sizes in each case are limited when the the worst voltage drop exceeds the voltage drop limit of 6.5V (due to the external voltage supplied of 12V and the minimum input voltage of 5.5V specified in the regulator data-sheet).	148
5.8	Dual-constrained TCA scalability for cube-shaped TCA systems with 1W regulated load.	149
5.9	Example of brute-force simulation for the relation of worst-case connector-pin currents and system-level regulated load-power.	151
5.10	Visualisation of a 64-tile TCA comparing the same system-level regulated 1000W load-power but different node-level power allocations. The red dots represent highest-wattage nodes, and the lower ones are highlighted in blue. (a) shows the worst case, whilst (b) is the best one obtained shown in Figure 5.9.	153
5.11	Conventional per-tile vs Relative-position allocations.	154
5.12	(a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 1x1x1-ball system.	156
5.13	(a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 2x2x2-ball system.	157
5.14	(a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 3x3x3-ball system.	158
5.15	(a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 4x4x4-ball system.	159
5.16	(a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 5x5x5-ball system.	160
5.17	(a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 6x6x6-ball system.	161
5.18	Example of a conceptual power-managed node with a CPU and an FPGA. The other components consist of any circuitry that consume power. The power management unit communicates with all the sub-units to maintain the node-level power consumption within the upper-bound power budget.	163

5.19 1x1x1-ball system-level power efficiency on GA and uniform allocation schemes.	164
5.20 2x2x2-ball system-level power efficiency on GA and uniform allocation schemes.	165
5.21 3x3x3-ball system-level power efficiency on GA and uniform allocation schemes.	165
5.22 4x4x4-ball system-level power efficiency on GA and uniform allocation schemes.	166
5.23 5x5x5-ball system-level power efficiency on GA and uniform allocation schemes.	166
5.24 6x6x6-ball system-level power efficiency on GA and uniform allocation schemes.	167
5.25 7x7x7-ball to 10x10x10-ball system-level power efficiency on the uniform allocation scheme.	168
5.26 Estimated TCA system power of various cubic sizes based on the simplified board model, showing the additional power contributed by each increment of array size n (and thus total power for the final case), for a range of component tile power loads. Total ball power in each case will be 8 times tile load.	170
5.27 (a) a logical representation of nodes and bi-directional channels forming a 2D-mesh topology. (b) the same network with explicit routers shown. (c) example of BookSim2's anynet user-file constructing the network with uniform latency of 1 cycle. All the channel latencies in this example file are uni-directional.	173
5.28 (a) A hexagonal torus topology employed in a SpiNNaker system. (reprinted from Figure 2.7 in [5]). (b) Example of node-number convention for hop-count simulations in this thesis.	174
5.29 Hop-count distribution of a 729-node, 9x9x9, 3D mesh topology.	175
5.30 Hop-count distribution of a 729-node, 9x9x9, 3D torus topology.	176
5.31 Hop-count distribution of a 729-node, 27x27, hexagonal-torus topology.	176
5.32 Cumulative hop-counts comparison amongst 3D-mesh, 3D-torus, and hexagonal torus.	177

6.1	Examples of power connectors and enhanced designs. Left: Existing prototype connector 22mm length, 6-pin (2xVcc, 2xGND, 2xData). Middle: Larger connector (approx. 30mm length), 12 pins, example: 4xVcc, 4xGnd, 4xData, giving double the power capacity. Right: Bespoke Design © C Crispin-Bailey, University of York, Diameter 25mm, 30-way connector, exploiting connections via hexagonal H-Facets rather than trapezoidal T-Facets.	184
6.2	A conceptual design for a bypass channel between second-immediate hops. The black line shows a dedicated line medium implemented on the same or a different PCB layer of existing components.	186
A.1	Hex-Tile Board Topside, showing main component layouts.	218
A.2	Hex-Tile Board Underside, showing Power resistor PCB pads (Larger square areas), and other component layouts.	218

List of Tables

- 2.1 Single packed cubic array quantitative equations. 19
- 2.2 Double packed cubic array quantitative equations. 20
- 2.3 "Critical length of common wires at 2 GHz (without equalization)." (reproduced from Table 3.2 in [15]) 23
- 2.4 List of hardware and network configurations in the surveyed parallel/distributed computers. 35
- 2.5 "Network traffic patterns. Random traffic is described by a traffic matrix, Λ , with all entries $\lambda_{sd} = 1/N$. Permutation traffic, in which all traffic from each source is directed to one destination, can be more compactly represented by a permutation function π that maps source to destination. Bit permutations, like transpose and shuffle, are those in which each bit d_i of the b -bit destination address is a function of one bit of the source address, s_j where j is a function of i . In digit permutations, like tornado and neighbor, each (radix- k) digit of the destination address d_x is a function of a digit s_y of the source address. In the two digit permutations shown here, $x = y$. However, that is not always the case." (reproduced from Table 3.1 in [15]) 50
- 3.1 Comparison of approximate simulation times and file sizes between the full prototype-board model and the simplified board model, for the 3x3x3-ball validation case shown in Figure 3.12. 71
- 3.2 The differences between the two alternative fitting techniques for the board modelling proposed in this thesis. 75
- 3.3 Prototype/Model: Single tile, Single connector. (© 2022 IEEE, regenerated from Table II(a) in [42]) 80
- 3.4 Prototype/Model: 8-tile ball, 2 co-located power connectors. (© 2022 IEEE, regenerated from Table II(b) in [42]) 81

3.5	Prototype: grid stability (worst case voltage drop, 10W load, 12V supply). (© 2022 IEEE, regenerated from Table II(c) in [42]. Note that there are an average of one power connector for every four tiles in all three cases, to ensure uniformity and also to ensure that connector pins per connector are not overloaded).	82
3.6	Simple vs complex simulation test case. Accuracy of the hypothetical test case of 3x3x3-ball TCA system shown in Figure 3.12 for non-negligible current flows of the validation of simplified versus complex manufacturer switching models.	82
3.7	LTspice® example code and parameters. (© 2022 IEEE, regenerated from Table III in [42])	83
4.1	Example of TCA SPICE-code template generation.	100
4.2	Advantages and disadvantages of the two adjuster types categorised in this thesis.	108
4.3	Overview of the details of the two-objective GA used in this thesis. Addi- tional details specifically in one of the MATLAB® gamultiobj's parameters named 'options' can be found in Figure 4.23.	126
5.1	Two-step extrapolation processes and curve fitting formulae used in Figures 5.3 to 5.8.	141
5.2	Example of a uniform simulator report file for a size of 7x7x7-ball TCA with the board model based-on the tile prototypes in this thesis without fan for cooling. Two parallel pins are dedicated for each of power and ground rails, and the connector-pin parameter is set for 50m-Ohm mated pin-pair. The highlighted case represents the nearest equivalent power load to that of the 65kW MDGRAPE-4A system.	171
5.3	Comparison of nodes per dimension and the total number of nodes of 3D mesh, 3D torus, and hexagonal torus.	178
5.4	Comparison of the selected existing systems concerning given total system power and estimated PE-related configurations.	180

5.5 Estimation of the required number of TCA tiles with a power budget of 25W regulated-load per tile, to mimic the whole power consumed at rack level for a given system as described in Table 5.4 using Equations 5.1 to 5.3. 181

5.6 Required cube-sizes of the approximate numbers of TCA tiles in Table 5.5. 182

1.1 Synopsis

In undertaking the research work and preparation of this thesis, **Modelling and Simulation for Power Distribution Grids of 3D Tiled Computing Arrays**, a number of research questions have focused around the novel concept of multi-dimensional tiled computing arrays (TCAs), and a hypothesis with respect to their feasibility and scalability. In particular this concept envisages a modular computing device, a core, or node in other terms, which is capable of interfacing directly with other computing nodes in one dimensional (1D), two dimensional (2D) or three dimensional (3D) structures, simply by connecting nodes directly together by means of convenient connectors. The intention of such modules is to eliminate all of the supporting printed circuit board structure, and the associated limitations of traditional back-plane systems, where 3D computing topologies are obliged to map onto primarily 2D structures, racks, back-planes, etc.

From another observational perspective, the design principle of the TCA concept is *physical 3D inter-node data-centric*, allowing significant freedom concerning inter-node data I/Os for direct neighbouring connection in 3D. This contradicts traditional rack-based computing approaches, which might be described as *back-plane centric* for both data and power connectivity. Back-plane-centric models limit both inter-board power and data I/O paths routed via one or more of hierarchical back-plane levels, with inherently 2D properties. However, with 3D inter-node data-centric models, whilst data flow has the advantage of true 3D connectivity, power distribution has to conform to the same model, leading to the specific novel problems of 3D cascaded power grids, as presented in this thesis for the investigated TCA systems. In the literature review, it will be seen that some other attempts to overcome back-plane centric limitations have been explored, to gain similar benefits. It is therefore valuable to explore the TCA model thoroughly at this time.

1.2 Motivation

Performing computing tasks on any parallel or distributed computer systems require mapping the tasks with their own logical topologies into given physical machines. By doing this, the physical topology enforced by the implementation of the system dictates the number of inter-node communication hops.

Traditionally, parallel or distributed rack-mount computers lack the availability of the third physical-dimension of communication channels between processing elements (PEs), e.g., inter-chip communication typically possible only on a 2D-plane of printed circuit board. This is a potential issue which degrades the result of mapping a set of logical tasks into a physical system, meaning the communication hops may not be achieved as optimally as they are implied in the logical representation. An example case of mapping is shown in Figure 1.1.

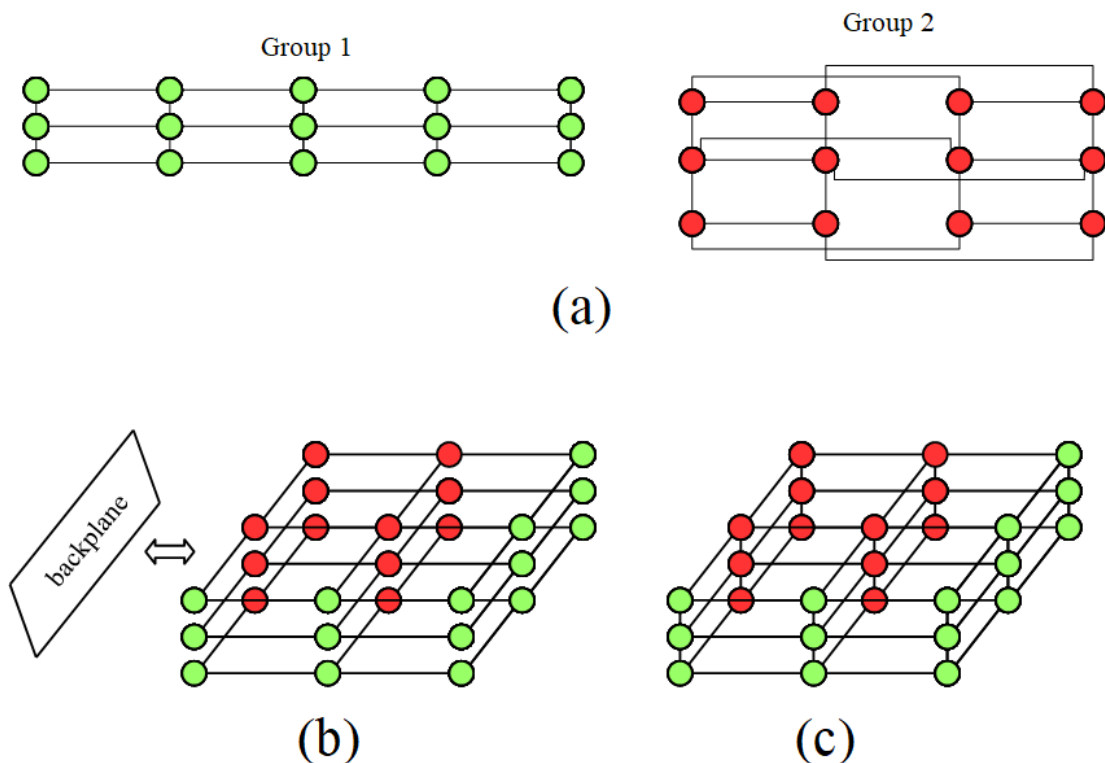


Figure 1.1: Mapping a couple of separate groups of tasks into different packaging systems. (b) is a representation of the two groups mapped into three horizontal computing-boards, each with 9 PEs, in a rack-mount system. (c) represents a mapping of the tasks into 27 tiles in a TCA system.

In Figure 1.1, two separate groups of tasks as shown in Figure 1.1(a) are mapped into a single set of 27 computing PEs. At first glance, both of Figures 1.1(b) and (c) seem to give an identical system performance. However, it can be seen in Figure 1.1(b) that every single node suffers from the lack of vertically physical data channels. This results in some *source-to-destination* pairs in a group shown in Figure 1.1(a) requiring additional physical communication-hops via other nodes and/or a communication back-plane, instead of ideally single immediate-hops. On the contrary, the two logical groups can be perfectly mapped into a TCA system as it provides the equivalent physical communication channels available for each group of the tasks.

Therefore, this advantage of the TCA packaging is the direct motivation to investigate the power-distribution grid of the tiled computing array in this thesis.

1.3 Thesis Scope

The 3D-TCA concept can be physically implemented into many practical hardware alternatives. However, in this thesis, only some aspects, abstraction levels for modelling, simulation framework, and hardware prototypes are investigated due to some research limitations such as the time-frame and financial budget for hardware building. The scope of this thesis is as follows:

► **Hardware prototypes:**

1) TCA unit-shape: A TCA unit for building a large-scale computing array is possible with many shapes, e.g., cube, triangle, hexagon. In this thesis, only a hexagonal-shaped tile is focused.

2) Power connector: Only an off-the-self connector model is employed for inter-tile power and intercommunication to demonstrate a possible practical hardware design. Whilst other practical connectors or alternative designs are also possible, however, are not in the scope for prototype implementation.

3) Voltage regulator: Whilst the voltage control at tile level is possible with the implementation of many types and numbers of voltage regulators, in this

thesis, at tile level, it consists of only a single regulator, which is a step-down switching regulator model.

► **Models:**

1) External (surface) power sources: In practice, many separate power-sources such as multiple connectors with different acceptable voltage ranges are possible to feed power into the whole system. However, for simplicity and to reduce the complexity of large-scale simulations, only a single voltage-source of 12V is assumed to be connected to all the surface input-power connections in the simulations.

2) Power-distribution grid resistance: In a power-distribution grid, resistance can technically exist in any electrically conductive media, including, the printed circuit board's trace resistance. However, only an inter-tile power medium type with a straight and uniform cross-sectional shape, e.g., circle or rectangular, along the medium, and also with the same resistance value, are assumed. Also, the heat arising from any component that may impact on resistance changes is not in the scope of this thesis.

3) Board-level power model: The power consumption at both the whole board level and also at the output-voltage of the regulator can vary over time due to various factors, for instance, the type of regulator employed and the behaviour of the onboard power loads. However, for large-scale TCA-size simulations, only constant power consumption at these two levels are assumed in this thesis for acceptable requirements such as simulation times and the machines required for simulations using a simplified board model. Also, as described earlier regarding the voltage regulator employed in the hardware prototypes, currently the board model also focuses upon the use of a single regulator and a single regulated load.

1.4 Research Hypothesis and Objectives

This thesis explores the feasibility of such systems in terms of scalability of power delivery networks, an aspect that is itself unusual as a result of the unique approaches

taken in TCA systems. An overall hypothesis for this thesis may be defined as follows:-

It is feasible to build a physical large-scale Tiled Computing Array with the power-grid constraints given, whilst still scaling up the system computing performance.

1.4.1 Definition

Regarding the research hypothesis, there are two key terms to be defined as follows:

- ▶ **large-scale:** The term *large-scale* is considered a comparative definition. In this thesis, it is defined, for a TCA system, as an estimation to achieve at least 10,000 tiles when constructed as a cube-size TCA to contain the same amount of processing-element power at rack/cabinet level of a recent traditional rack-mount system existing since a decade ago, whilst also considering the given two constraints for the power-distribution grid in this thesis. A comparison metric regarding this term will be elaborated in Chapter 5.
- ▶ **constraints:** There can be several electrical constraints when designing and implementing computing devices. However, the following two key constraints are focused in this thesis:
 - 1) **voltage drop:** In this thesis, *voltage drop* is defined as the amount of voltage reduction, dropping from the voltage of the external (surface) power sources. In many cases, this term will be frequently used when discussing the board input-voltage.
 - 2) **connector-pin current:** This term refers to an electrical current flowing through either a power (positive) or a ground (negative) pin, on a tile-edge connector.

In order to determine if this hypothesis is valid, several important research questions are investigated:

- ▶ What are the necessary design choices for constructing tileable modules?
- ▶ What are the component characteristics of the power grid in a TCA array?

- ▶ How is system computing performance influenced by the power grid design and limitations?

These questions are addressed in this thesis, and a number of methodologies are utilised in order to achieve clear conclusions. In particular, the use of complex modelling via MATLAB[®] [1] and circuit modelling/simulation tools, and the construction of validation prototypes has been employed in order to calibrate models if needed, and to permit deeper exploration of the topic. It will be demonstrated that power grids are feasible in TCA systems and that they may potentially scale to many thousands of computing nodes.

1.4.2 Research Objectives

Given the research hypothesis, in this subsection, the research objectives will be briefly given. There are two requirements prior to set out the research objectives for the modelling and simulation towards testing the research hypothesis. First, the critical components in a TCA which contribute to the limit of the power-distribution grid's scalability and computing performance need to be understood. This ensures that only relevant components will be focused upon for modelling and simulation developments.

Second, due to the existing research field of power modelling in parallel/distributed computer systems, relevant models and simulation tools should be surveyed to be employed or modified, if any exist, for the purpose of TCA scalability evaluations, thus avoiding any re-invention of existing toolsets. These two requirements will be carried out in Chapter 2. After considering the critical components, and the survey of existing relevant models and simulation tools, the research objectives carried out throughout this thesis are to be designed towards testing the research hypothesis, which are as follows:

- ▶ **Objective 1: Employing and designing models and simulation tools**

After the survey of the existing models and tools, if it is found that they are not suitable for the intended research investigation, new model designs and simulation tools need to be built for representation of the novel TCA architecture.

► **Objective 2: Hardware validation**

Whilst the models and the simulation framework are being built, both should be verified and validated. This would initially be in comparison with expected results, but simulation results should then also be validated for accuracy by comparing with real hardware prototypes as far as resources permit.

► **Objective 3: Fundamental simulation experiments**

With the simulation platform having been validated, experimental cases are to be designed to evaluate the scalability of the TCA concept in various scenarios.

► **Objective 4: Optimised power distribution**

This objective can be considered an advanced capability in simulation experiments. With heterogeneous node implementation in a parallel/distributed machine, non-uniform power allocation may be involved and this also impacts upon the behaviour of the power network in TCA systems. The optimisation performed in the non-uniform simulation adjusts each tile's regulated load-power, trying to achieve the system-level goal, i.e., the summation of all the regulated load-power from all the tiles in the system should be as high as possible, whilst all the connector-pin currents are still within the desired current limit. Non uniform power allocation might also occur where an external power connector is unplugged or has some form of failure. So this capability is useful for predicting system resilience and reliability.

► **Objective 5: Scalability evaluations**

This objective can be considered the product of the collective efforts of the preceding objectives defined, and is ultimately the purpose of the simulation platform in exploring the hypothesis. Evaluation of the data, test cases, and implications in the context of the stated hypothesis and research questions are the key aims here.

► **Objective 6: Simulation framework documentation**

Tool documentation is also an important process, not only for the system designers, but also for future tool extensions. Tool documentation can be seen as an ongoing and long-term process during the current development and in future work on the next capabilities to be added to the tool in the future. Not all of the desired detailed documentation may be possible in a limited time-frame, how-

ever, at least important features and useful guidelines should be included. Other issues such as cautions and experiences from this thesis are also additionally useful for future tool adopters and developers.

All of the research objectives will subsequently be expanded in detail in their relevant chapters, along with discussions, including success criteria.

1.5 Contributions

This thesis makes a number of novel contributions to the field of parallel/distributed computer systems, and in particular the novel area of 3D tiled modular arrays, an area that has not been well explored to date. The thesis presents a number of novel outcomes in terms of tools and methodologies that permit simulation of power grids in such systems based upon a number of hypothetical parameters, including specific characteristics of the essential components of the system. The novelty therefore lies in creating a framework and using that framework to evaluate and validate 3D cascaded power grids in novel 3D tiled modular system arrays.

Further detail on specific outcomes relating to these contributions are given below.

1.5.1 Methodology

- **Modelling:** The power-related research topics regarding traditional rack-mount systems typically focus upon various levels of component. However, it lacks the power modelling regarding the power-distribution grid for the whole parallel/-computer system employable for the unique TCA power-network focused upon in this thesis. Thus, a simplified board model is proposed to evaluate large-scale TCA power-distribution grid scalability under the assumption of constant board-level power consumption. The model simplification eliminates 1) the need for modelling a complex model at board level, and 2) the issue of the global SPICE ground-node tied to the internal grounds of a switching voltage-regulator model. Consequently, this simplified board model mitigates several large-scale

TCA simulation difficulties. Moreover, the simplified board model is validated against the hardware prototypes built purposely in this thesis. Apart from the simplified board model, the other components, which are connector pins and intra-tile power medium are also proposed and investigated in the modelling framework in this thesis.

- **Simulation framework:** Whilst a large number of existing *interconnection network performance* simulators have already been proposed, None of inter-board power-distribution grid simulation has existed specifically for the purpose of large-scale TCA simulations. In this thesis, a whole TCA power-distribution grid simulation framework is proposed, comprising of a large-set of automating simulation functions. The simulation framework is also well-structured for future extensions. In terms of board-level power allocation, two schemes are proposed, which are 1) uniform, and 2) non-uniform allocations, for optimal power allocation given the power constraints. Whilst in traditional rack-mount systems, the power consumption at computing-board level is not an obvious concern to impact on the whole power-distribution grid. However, this is not true in a TCA system, which means that board-level power consumption also impacts on the whole TCA power-distribution grid's behaviour and scalability. Additionally, parallel simulation and visualisation are also the other significant efforts made in the simulation framework.

1.5.2 Findings

The findings themselves in this thesis, as the results of the methodology proposed, are the scalability evaluations from both the uniform and non-uniform power allocation simulations. This shows that the large-scale TCA power-distribution grids based on the hardware prototypes built in this thesis, are feasible. Compared to uniform power allocation, the non-uniform power simulation framework proposed in this thesis is able to further discover additional optimal allocation cases under the given power constraints.

1.6 Thesis Organisation

This thesis will be organised as follows:-

Chapter 1 introduces the overview of this thesis, motivation, thesis scope, research hypothesis and objectives, and contributions.

Chapter 2 discusses existing concepts and principles relating to TCA, logical-to-physical mapping problems, important surveyed previous work, beginning with systems using rack-mounted traditional packaging, followed by TCA constraints, related models, and other relevant topics.

Chapter 3 details the simulation models and equivalent hardware prototypes and their comparison for simulation validation purposes.

Chapter 4 elaborates the simulation framework, ranging from the creation of SPICE-simulation entry files to the automation of scalability simulations.

Chapter 5 demonstrates and discusses key scalability simulation results, and also provides topological simulation results as preliminary work for future tool developments.

Chapter 6 concludes all the outcomes from the modelling and simulation framework, and discusses possible future work.

Field Survey and Literature Review

2

This thesis is focused towards modelling and simulation for power distribution grids of 3D tiled computing arrays. To undertake a suitable field review it is necessary to understand the basic and somewhat novel concepts of the TCA, and already established principles, alongside other surrounding relevant topics such as trends and challenges in high performance computing (HPC).

In [2], the need for unconventional HPC architectures is discussed due to the slowing down of Moore's Law, and the end of Dennard scaling. Examples of these unconventional architectures are neuromorphic computing, artificial intelligence (AI) chips, and processing in memory (PIM). Apart from these non von Neumann computing architectures, the interconnects between any levels of computing elements are also important. Co-packaged optical (CPO) is a technique to shorten the electrical paths on a PCB when using conventional pluggable optic modules, by integrating both optics and silicon onto the same package. These developing trends of both unconventional computing architectures and also I/O technologies are driving the HPC systems to new architectural designs.

Complications for high performance computing are not limited only to computing architectures and I/O improvements. In current packaging technologies employed for inter-chip or inter-board levels in computer systems, there are physical integration complexities when constructing whole parallel or distributed machines. A computing board with preset numbers of PEs, e.g., CPUs, GPUs, along with expansion slots for memory, and peripherals, are usually mounted in a rack or cabinet. Respectively, multiple boards are interconnected together via back-plane(s) and wiring systems for power and intercommunication at intra-rack or inter-rack and cabinet levels. A large-sized high performance computing board can be advantageous, supporting a large portion of workload with multi-thread and/or shared-memory model applications on a large single board. However, some types of computing workload may require less memory and also need *logical* 3D intercommunication patterns that may be more

suitably mapped on a *physical* topology and exploit the close proximity of PEs in physical 3D space. The SpiNNaker project [3], [4] is an example of machines where multiple PE-chips are located close together at intra- and inter-board levels with logical 3D connectivity (hexagonal-torus topology at system level [5]) but 2D physical implementation (inter-board level). The direct *logical-to-physical* task mapping issue is due to a large-sized board and intercommunication via a back-plane and/or via separate routing devices. However such systems limit the opportunities for direct and short chip-to-chip communication in the third physical dimension between adjacent boards.

Regarding power delivery systems, traditionally this consists of boards (e.g., rack-/blade servers, etc.) powered by AC/DC units depending on specific designs chosen. If AC-to-DC power conversion units are also part of a board, they inevitably occupy a dedicated intra-board space for the purpose. A back-plane supporting DC power rails, (e.g. at 12V) is also a possible design, distributing input voltages across stacked up boards in a rack or cabinet. However, with the idea of DC back-plane powering for multiple high-wattage boards, high electrical current distribution becomes a concern to be carefully taken care of for electrically conductive media used in the power back-plane (such as bus bars). Indeed, back-plane bus-bars can be quite substantial metal components carrying heavy currents. The voltages lost in the back-plane are also an issue to some extent.

Heat dissipation is also an important consideration. Thus, appropriate cooling systems are required. Most modern machines are equipped with air cooling systems, whilst some utilise liquid cooling, either by directed cooling or immersion. Whilst the trend of the whole-system performance per power efficiency is positively increasing as shown in Figure 2.1, it should not be underestimated for lower levels of design for heat issues due to the high-transistor count per chip and also the density of nodes per volume. With this consideration, liquid cooling techniques are a potential choice for current and future high performance machines [6], [7], [8]. Examples of some of these cases are shown in Figures 2.2 and 2.3.

Interestingly, the cooling approach taken for the Google TPU data-centre as shown in

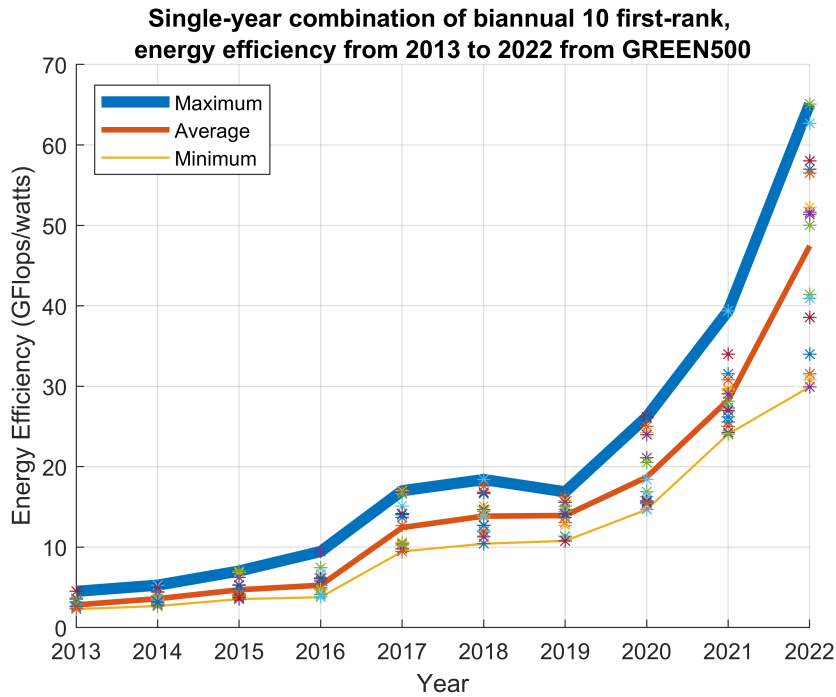


Figure 2.1: Energy efficiency trends reported by [9]. Each year in the graph is a combination of biannual 10 first-rank reports in June and November.

Figure 2.2 is a duplication of the data inter-connectivity problem encountered when a 3D logical topology is mapped onto a 2D rack and cabinet system architecture. This suggests that the TCA concept also potentially has significant advantages for this problem domain, as well as for data connectivity.

Deeply detailed investigation of cooling for TCA is beyond the scope of the thesis. However, the cooling issue is a fundamental concern for TCA node-level design and is a potentially fruitful area for future research projects.

With the introduction above, the topics relevant and essentially crucial to TCA evaluations and implementation will be discussed in the following sections.

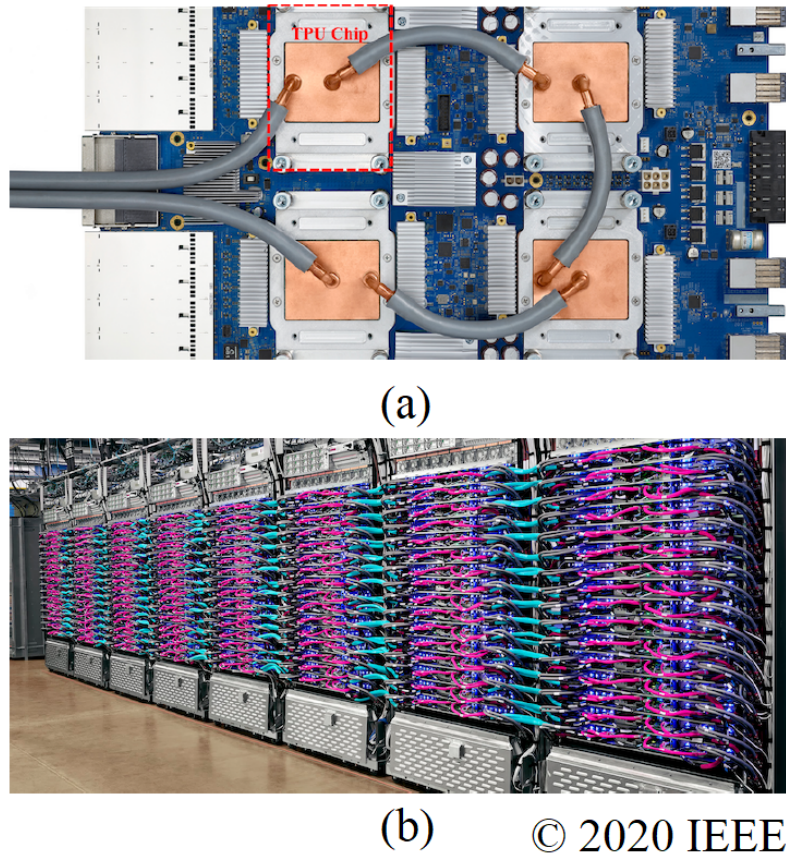


Figure 2.2: Examples of large scale system cooling, showing (a) TPU v3 unit, (b) TPU v3 Pod. (reprinted from Figure 1(b-c) in [10])

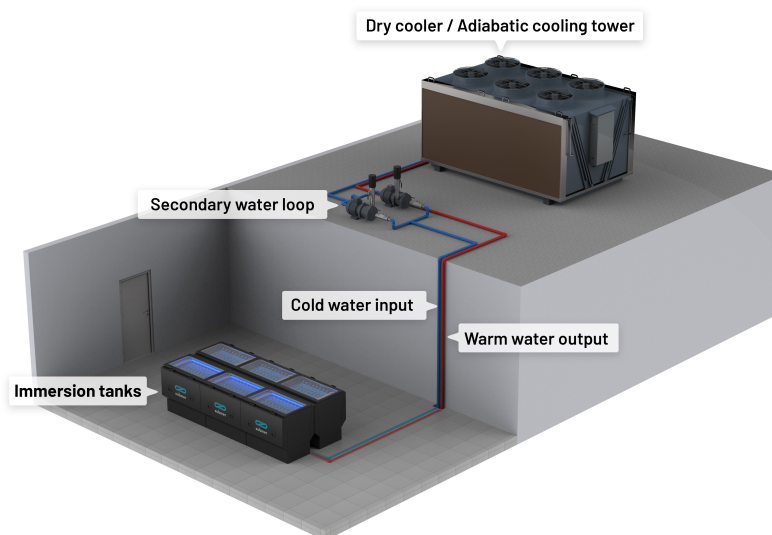


Figure 2.3: Submer Immersive cooling system. (reprinted from [11])

2.1 Existing Concepts and Principles Relating to TCA

Before implementing real systems, typically, it is worth comprehending their background ideas. This is to realise the core-value concepts of the systems to be implemented. In this thesis, the *TCA* structure can be implemented in many alternatives, ranging from intra-node components, to the whole configuration of the system.

2.1.1 Short Introduction to TCA

It can be considered that the first variant design of the TCA concepts, was officially published in a UK patent, 'Computing devices' [12]. In that design concept, the smallest unit, referred to as a *tile* or *hex-tile*, is a hexagonal shape package containing computing circuitry inside. Each unit can be coupled to another one to form a 1D, 2D, or 3D array, constructing a networked computer system. An example construction is illustrated in Figure 2.4. Constructing eight tiles into a *ball* module (actually a truncated octahedron) as shown in Figure 2.4(b), gives a uniform building block for power delivery, data communication, and pathways for cooling. With this uniformity, it eliminates traditional hierarchical infrastructures starting from this unit level towards the entire system.

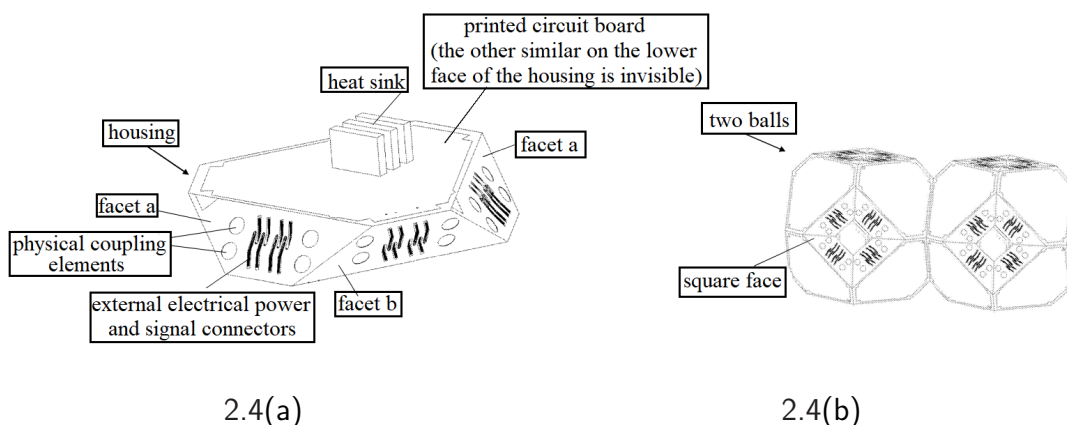


Figure 2.4: Example of a TCA design proposed as 'Computing devices' in [12]. (a) shows tile structure, (b) shows ball structure, composed from tiles. (adapted from Figure 1 and 5 in [12], respectively)

To give some advantageous examples, when comparing to conventional rack-mount packaging, it can easily be observed in a TCA that an entire machine can be rapidly composed without data cabling efforts inside the system. Except that extra wrap-around data wires are sometimes needed in some other topologies, for example, a torus. An example of a 3D-torus topology is shown in Figure 2.5. Also, uniform tileable modules remove the need for custom-system rack modules, efforts in printed circuit board (PCB) design and manufacture of 2D-board containing complicated hardware layouts, and all of the associated costs and environmental impacts of these construction overheads. Although inter-module data and power wiring is not required inside a TCA, such a machine still needs to be externally powered via connecting points on its six rectangular surfaces. With a well-designed convenient external power-connection module, e.g., plate, power connections can be rapidly completed in a short time-frame and also with low composition effort. This TCA packaging design will be thoroughly discussed along with its predecessor research and developments in this chapter. An additional advantage of a true 3D topology is that technologies such as optical fibre and I/O channels can be readily integrated into the balls with no external fibre connectivity. Ball facets simply abut and permit electrical or optical data transfer via aligned pairs of optical ports on the surface of the two facets. The principles of integrated silicon photonics and chip to chip wave-guides are well established and already in use.

2.1.2 TCA-concept Definition

TCA modules can have many forms. In this thesis, some particular design choices are assumed. Some aspects of the TCA concepts have been briefly discussed earlier. However, this subsection is dedicated to providing a concise definition of the very fundamental conceptual idea, without expanding this into any specific hardware implementation. This concept, which is referred to in this thesis as the 'Tiled Computing Array', or TCA, can be split into two parts defining the concept of construction:

- 1) **Computing Array:** A 1D, 2D, or 3D array of computing-related module. Each module may be implemented as some form of packaging, for example, the hexagonal

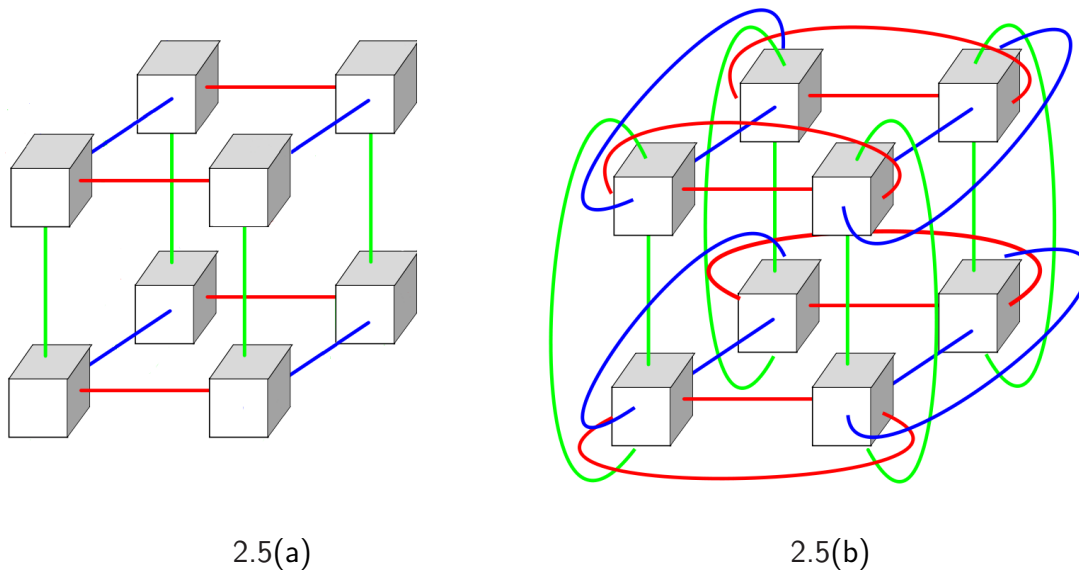


Figure 2.5: Logical views of a 2x2x2 3D-mesh and 3D-torus topologies. (a) depicts a 3D-mesh topology (adapted from [13]), and (b) depicts a 3D-torus topology [13].

tile or ball illustrated in Figure 2.6. Each of which contains single or multiple processing elements (PEs), and/or other components such as memory, router, purposely for making up the whole parallel/distributed machine.

2) **Tiled:** Physically constructing a parallel/distributed computer in 1D, 2D, or 3D space, by using the modules in 1) in *tiling-like* action, - hence, *tiled*.

With this fundamental concept to ease the composition, and eliminate inter-board wiring, racks and other hierarchies, such a system can be implemented in many physical alternatives. An example variant design of the TCA concepts can be found in [14].

2.1.3 Geometrical Properties of Ball Arrays

As mentioned earlier, hexagonal tiles may be composed into 'ball' modules by joining together eight tiles in an appropriate way. Balls are then also tileable modules in their own right.

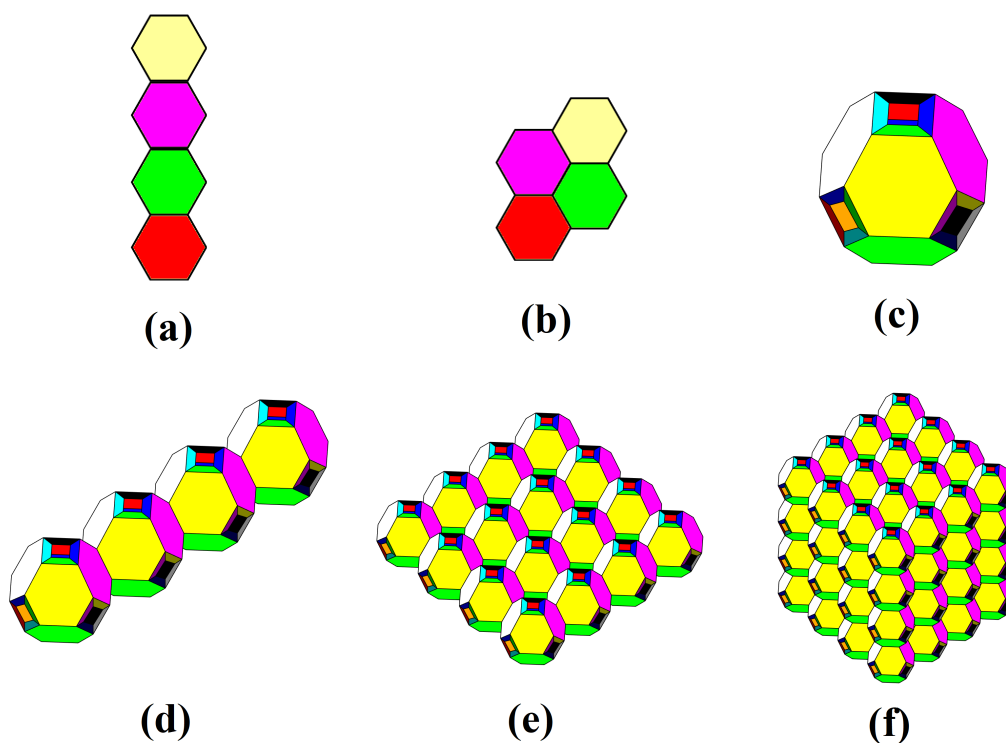


Figure 2.6: Example implementation alternatives of the TCA concept. (a) shows 1D tile-level array of 4 modules, whilst (b) shows a group of tile-level 2D array of 2x2 modules, and a possible 3D construction from the tile level can be seen in (c) as a complete 3D array of 2x2x2 modules. Instead of tile-level composition, each module may also be alternatively implemented as a single ball-shaped object. (d) depicts a row of 1D array of 4 modules. (e) constructs a 2D-array of 4x4 modules, whilst a 3D-array of 4x4x4 modules can be constructed as shown in (f).

When a number of balls are assembled into an array, the whole system acquires numerical and geometrical properties as functions of the array shape and key dimensions. A cubic array of balls is one obvious choice and as this is a focus of the thesis, several precise attributes of a ball-array system can be directly obtained by use of formulae derived from the observed properties of the cubic structure. These formulae can be employed to rapidly obtain composition-related quantities concerning array composition. A number of properties are given in Tables 2.1 and 2.2 which represent a single-packed cubic array (similar to a uniform lattice) and a double packed array (with balls packed in between balls).

The formulae define the property n as the externally visible dimension, such that a 3x3x3 array would have $n = 3$ for instance. T-facets (Trapezoidal facets) represent the 6 square faced facets present in the truncated octahedron, whilst H-facets represent

Table 2.1: Single packed cubic array quantitative equations.

Factor	Equation	Eqn. No
Total Cores (single)	$C_{TS} = n^3$	(2.1)
Total Tiles (single)	$Tl_{TS} = 8C_{TS}$	(2.2)
External Cores (single)	$C_{Es} = n^3 - (n - 2)^3$	(2.3)
Internal Cores (single)	$C_{Is} = (n - 2)^3$	(2.4)
Total T-facets (single)	$T_{TS} = 6C_{TS}$	(2.5)
Total H-facets (single)	$H_{TS} = 8C_{TS}$	(2.6)
External T facets (single)	$T_{Es} = 6n^2$	(2.7)
External H facets (single)	$H_{Es} = 24n^2 - 24n + 8$	(2.8)
Internal T facets (single)	$T_{Is} = T_{TS} - T_{Es}$	(2.9)
Internal H facets (single)	$H_{Is} = 8C_{TS} - H_{Es}$	(2.10)
T Bisection factor (single)	$B_{TS} = 0.5n^2$	(2.11)

the eight hexagonal facets of the same geometry.

The formulae allow properties such as internal, external, and total number of balls, facets, and other features to be calculated from a given value of n . Additional properties can also then be derived. For example, the bisection factor represents how many facets (of type T or H) are utilised to provide the bisection bandwidth. Simply multiplying this value by the bandwidth of a facet (e.g. 1Gbps) gives the bisection bandwidth of the whole system. Likewise, the total number of external T-facets indicates how many power connections are available, and once one knows the power capacity of a T-facet, one can calculate the entire raw power input capability of a system of given size ($n \times n \times n$) array.

However, to calculate the true behaviour of the system power distribution, internally, requires the complex modelling of connector resistance, network voltage drop, regulator behaviour and so-on, as will be explored further within later chapters of this thesis.

Table 2.2: Double packed cubic array quantitative equations.

Factor	Equation	Eqn. No
Total Cores (double)	$C_{Td} = n^3 + (n - 1)^3$	(2.12)
Total Tiles (double)	$Tl_{Td} = 8C_{Td}$	(2.13)
External Cores (double)	$C_{Ed} = n^3 - (n - 2)^3$	(2.14)
Internal Cores (double)	$C_{Id} = (n - 1)^3 + (n - 2)^3$	(2.15)
Total T-facets (double)	$T_{Td} = 6C_{Td}$	(2.16)
Total H-facets (double)	$H_{Td} = 8C_{Td}$	(2.17)
External T-facets (double)	$T_{Ed} = 6n^2 + 6(n - 1)^2$	(2.18)
External H-facets (double)	$H_{Ed} = 24n^2 - 24n + 8$	(2.19)
Internal T-facets (double)	$T_{Id} = T_{Td} - T_{Ed}$	(2.20)
Internal H-facets (double)	$H_{Id} = 8C_{Td} - H_{Ed}$	(2.21)
T Bisection factor (double)	$B_{Td} = 0.5n^2 + 0.5(n - 1)^2$	(2.22)
H Bisection factor (double)	$B_{Hd} = 2(n - 1)^2$	(2.23)

The formulae in Tables 2.1 and 2.2, have been formulated by Dr. Christopher Crispin-Bailey.

2.2 Mapping Between Logical and Physical Topologies

In this section, a brief introduction to the key problems seen in this thesis relating to traditional rack-mounted construction will be given. However, the literature of traditional rack-mount systems surveyed in this thesis will also be discussed in detail in later sections of this chapter.

1) Mapping logical 3D topologies to the physical packaging technology:

Achieving the actual logical computing tasks of a set of parallel computational requirements, (i.e. a workload) as shown in Figure 2.7, is highly dependent on the physical topology of the parallel/distributed machine, and the restrictions it has (e.g. rack mount and back-plane, 3D tiles, etc.). In the traditional rack-mount packaging approach, the non-availability of the direct data channels between nodes that are immediately located in different boards could degrade the desired logical topology as it is implemented. Figure 2.8 shows an example of mapping tasks in Figure 2.7 onto a rack-mount system and a TCA. With the restriction of only a single physical data channel between the vertical nodes at the corner in Figure 2.8(a), the desired

logical computing topology is not preserved. The single isolated node that is mapped onto a separate board above seems to be a trivial issue at first sight. However, in the situation that this particular node demands large amounts of intercommunication data, it could cause a bottle neck problem due to indirect hops between rack-mount boards. A 3D tiled system would however have no such limitation and hence is the direct motivation for TCAs to be investigated.

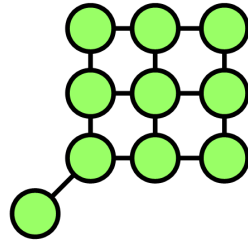


Figure 2.7: A logical (machine-packaging independent) computing topology of 10 tasks.

2) Physical communicating distance in three dimensions

The previous mapping issue focuses on the *deformed/degraded* logical computing topology and unnecessary *hop counts*, which can be considered *pure topological properties*. However, the physical distance problem in this part is from the perspective of the actual *physical travelling distances*.

This particular issue can be considered as a by-product of the previous issue mentioned due to forcing indirect hops along 2D-planes, and often necessitating dedicated routing devices or inter-rack data packetisation overheads (as observed in SpiNNaker for example [3]). For a high-performance server implemented using a large circuit board equipped with several PEs such as CPUs, GPUs, this can be seen as a single node. However, when considering the physical communication distance between a given pair of PEs that are located on different circuit boards, it is inevitable that a data unit from its 'source-node' needs to travel all the way through the intra-board circuitry implementing some form of topology, then traversing through a board-edge cable at the back-plane. Afterwards, typically the data unit is forwarded to at least a single intermediate routing device such as a separate router. Finally, the data unit is fed into the terminal board to reach its 'destination node'. This *inter-board long trip* is

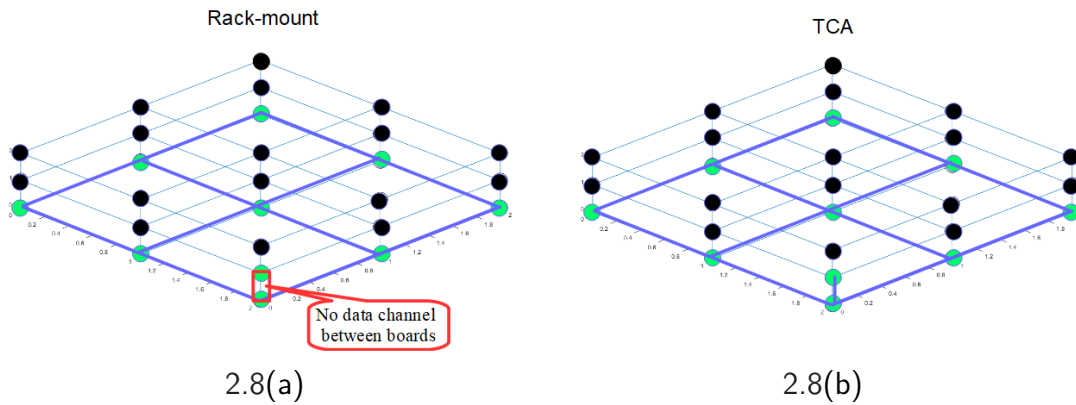


Figure 2.8: Black dots represent occupied nodes, and green dots are the tasks being mapped. (a) Mapping a set of tasks on a plane of 3x3 nodes, and a single node on a different board in a rack-mount system. (b) The same mapping on a TCA system with the same system size. As the last task is mapped on an adjacent node above, and with the availability of a channel in the physical third dimension, the logical computing topology can be maintained.

necessitated by the non-availability of a direct channel in the third dimension, resulting in a large number of hops.

This *inter-board long trip* and 3D TCA alternative can be visualised in Figure 2.9 and Figure 2.10 respectively. Therefore, a physical 3D topology permits direct connectivity whilst a back-plane system can only provide inter-board connectivity by adding a number of dedicated data channels or shared bandwidth back-plane data highways. This often has further consequences such as the need for dedicated data protocols, hardware structures and so-on to manage the data traffic.

At intra-board level, a long PCB trace can also be implemented as a single channel between the two extreme far-end nodes. However, it incurs a reduction of channel frequency. Examples of critical lengths degrading the signal quality can be found in Table 2.3. Notably, in the recent (2022) white paper "Unconventional HPC Architectures" [2], short-range high-speed localised communications links are given as one of the key requirements for future HPC systems.

3) Physical channel-lengths: Another important factor that impacts both inter-node communication latency and throughput is the length of a channel. Physically, a channel between a couple of immediate PEs can be an intra-board level channel, e.g., PCB trace, or inter-board level, e.g., cable. For intra-board level, it depends on

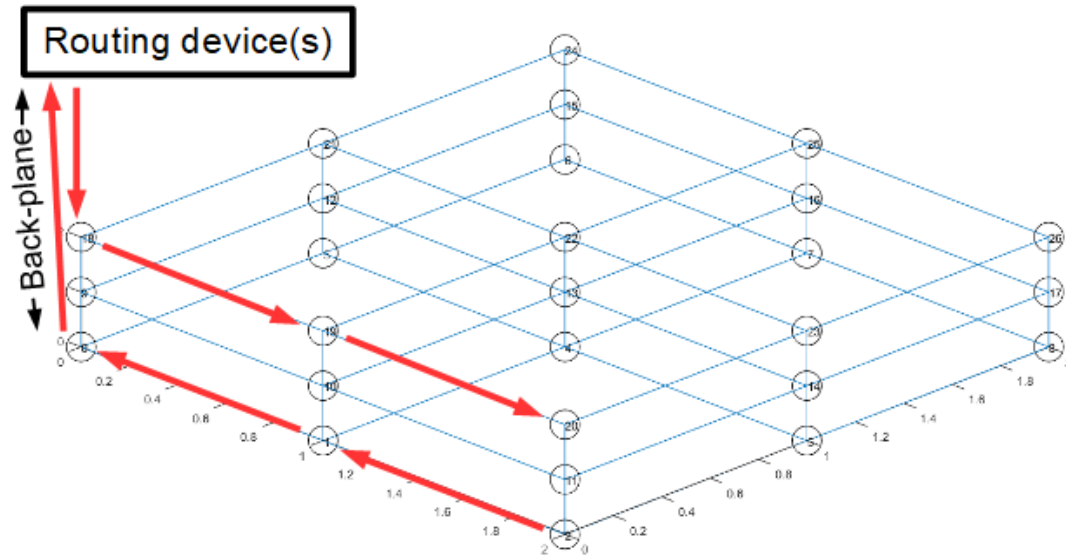


Figure 2.9: The shortest route of a data unit travelling from node 2 to node 20 between computing boards in a rack-mount system, resulting in a hop count, at least, of 6. It is assumed that all the nodes at the far-end of each board are directly connected to a routing system

Table 2.3: "Critical length of common wires at 2 GHz (without equalization)." (reproduced from Table 3.2 in [15])^a

Wire Type	l_c
5 mil strippguide	0.10 m
30 AWG pair	0.56 m
24 AWG pair	1.11 m
RG59U coax	10.00 m

^a With granted permission by Elsevier

the distance between PEs printed on a circuit board. However, for inter-board level, boards are typically interconnected via cables to convey data electronically or optically. Thus, if poor board placements are employed, or the desired logical topology is not suitable to be implemented in rack-mounted packaging, they may incur physically long channels (cables). A TCA can eliminate this inter-board level issue by coupling nodes together with convenient and direct connectors, available in both horizontal and vertical directions to form a uniform 3D-mesh topology with very short physical channels.

4) I/O-signal driver module: With cables required to construct a system, spe-

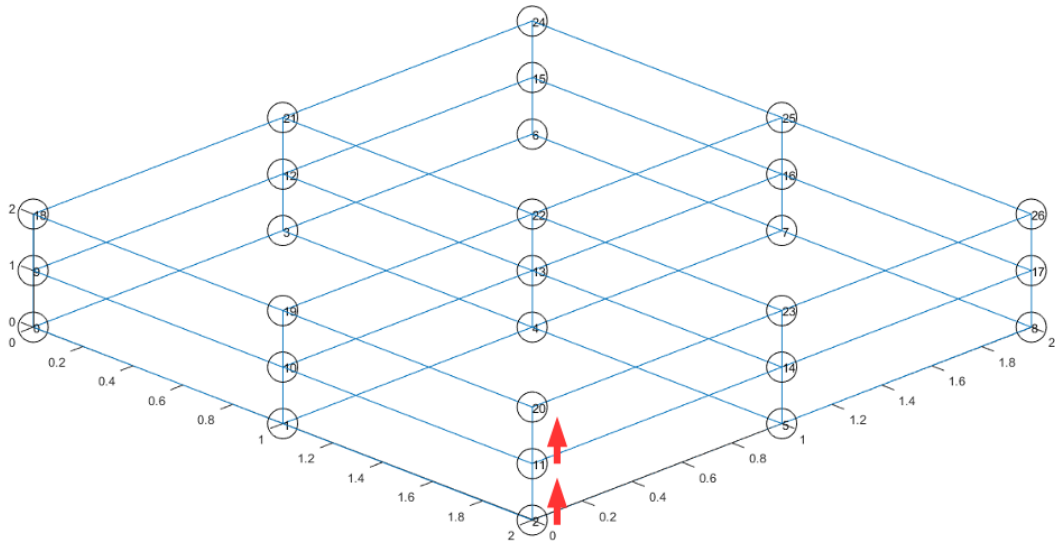


Figure 2.10: The shortest route of a data unit travelling from node 2 to node 20 between a couple of adjacent balls in a TCA system (node 2 and node 11 reside in the same ball, as a ball consists of eight tiles), resulting in a hop count of 2. A single hop count can be achieved on 3D-torus implementation with a wrap-around channel.

cialised circuitry such as an Ethernet module may be also needed to be responsible for maintaining the strength and quality of the signal travelling from a source node to its destination. On the other hand, TCA can eliminate this extra overhead (which consumes power, creates signal delays, and occupies physical circuit board space) by directly coupling units together, with a much more tightly-coupled connection.

5) Cabling effort: The cabling efforts in traditional rack-mount systems depend upon the complexity of the logical topology and physical placements of computing hardware. With hierarchical structures, a well-designed cabling plan is required. The SpiNNaker project [5] is an example that investigates the building and operating of an unconventional architecture, and provides a collection of tools for generating cabling plans. For a TCA system, the cabling inside the system is completely eliminated. However, external power cables may also still be required, if the external power sources are separated. A power plate is an optional design choice for a compact and rapid means of system power configuration.

Referring to the definition of *large-scale* in this thesis, this reference point should be well-defined and quantitative to be comparable with recent existing systems. Defining

the term large-scale can simply specify a large number of nodes in the system. However, it is not only the number of nodes, but also the power per node that should also be sensible for embedding at least a reasonable real-world PE and computing-related components per node to construct a high performance system. Thus, in order to have an agreed definition of the term 'large-scale' in this thesis, consideration is also given to the perspective of the power figures of recent existing systems. This is to prevent scaling up a TCA physical system with unrealistic node-level (tile) regulated power allocation, for example, allocating too low, e.g., the range of mW, or exceeding the maximum regulated voltage/current specified by the regulator employed in the current stage of the hardware prototypes.

2.3 Parallel/Distributed Computer Packaging

In this section, the survey of important traditional rack/cabinet systems will be discussed with other different techniques. This section will also emphasise how the TCA concept investigated in this thesis is a strong motivation to tackle the problems highlighted earlier in this chapter.

2.3.1 Rack-mount Packaging

There have been several rack-mount systems developed. However, it is considered not productive to exhaustively collect a large number of existing traditional systems as they share the same characteristic of being *rack-mount*. Instead, in this subsection, important and recent systems surveyed will be discussed. Recently, a group of projects, ExaNeSt [16], [17], [18], [19], ExaNoDe [20], ECOSCALE [21], and EuroEXA [22], developed HPC components to achieve *exascale*, 10^{18} floating-point operations per second. The timeline of the projects is shown in Figure 2.11. The cooling system in ICEOTOPE [23] is employed in the ExaNeSt project by immersing computing boards in non-conductive liquid to cool down at blade-level. Frontier [24], is another recent

HPC that achieved the first rank on High Performance LINPACK (HPL) benchmarks [25]. Cabinets in the Frontier supercomputer is shown in Figure 2.12.

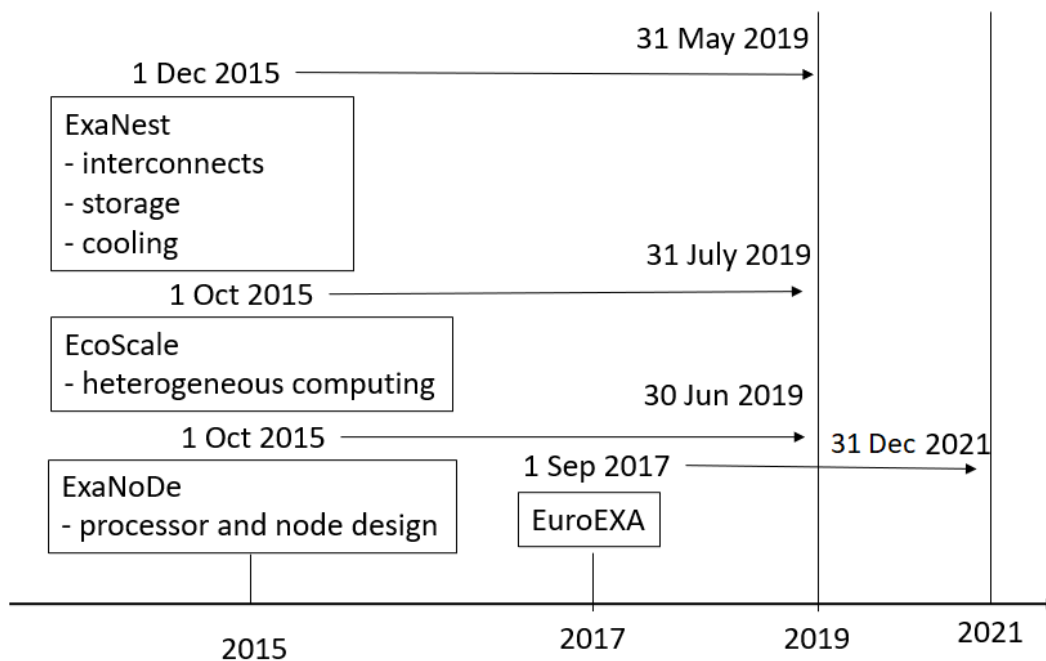


Figure 2.11: Timeline of four recent examples of Exascale projects. The period of each project duration can be found in [20], [21], [16], [22].



Figure 2.12: Frontier supercomputer. (reprinted from [26])

Another recent unconventional high-performance machine, SpiNNaker [3], is specialised for spiking neural networks. A prominent advantage regarding power consumption is that, at the chip level, it is reported that a 200MHz, 18-core chip consumes

only 1W [27]. SpiNNaker is one of the architectures surveyed in this thesis that locates multiple PE-chips close together in 3D space. However, SpiNNaker still needs substantial wiring effort to form a *3D hexagonal torus* topology. [5] investigates this wiring issue and provides a collection of tools for generating cabling plans. MDGRAPE-4A [28] is another recent high-performance computer specialised for molecular dynamics simulations. This machine is topologically relevant due to being a 3D-torus system. Even though it is designed with a similar logical topology that can also be constructed in TCA, it can be obviously seen in [29] that the same significant effort of fibre cabling is still required.

To end this subsection, it can be concluded that most of the high-performance computers at the present are based upon rack-mount, back-planed approaches, which have been considered very mature for decades. However, it can be seen that the challenging complexity of hardware structures still exist. This is due to hierarchical construction that brings about the effort of composing complete functional systems, ranging from PEs to the entire systems. From the topological point of view, rack-mount and back-plane systems provide the convenience of board removal and some flexibility for constructing variants of desired topologies by typically cabling them to routing devices. However, disadvantages are also encountered. Back-plane channel routing extends physical lengths of data channels between inter-board PE chips, which sometimes completely differ from the expected appearance of physical topology vs the logical one. Cabling does not only involve interconnection-network data channels, but also power delivery. Also, the existence of cables not only means cost, but also the actual physical length between PEs, which impacts on communication bandwidth and latency. Table 2.4 show a list of important surveyed rack-mount systems. Increasingly the environmental cost of components such as densely cabled systems and complex PCB modules are also a consideration.

2.3.2 Non rack-mount Packaging

The machines in this category are conceptually proposed or real systems. All of the systems surveyed that are not obviously and/or substantially based on the traditional

rack-mount packaging, are also grouped in this subsection. Whilst all of the systems in this subsection are considered as non rack-mount systems, TCA is separately discussed in detail to point out its unique and motivational features.

1) Relevant work in the field

In this sub-category, the TCA concept is not mainly discussed in detail, but rather in a comparison manner, to clearly distinguish the research gap left to investigate in this thesis. However, some prior related work also shares some of the goals that TCA systems envisage to tackle.

Dated back to the 90s, J-Machine [30], a fine-grain parallel computer, presents an attractive packaging architecture due to its node density and topological construction, even though, "chassis", the container used in J-Machine, has a partial appearance of rack or cabinet at a degree. However, what is more interesting is its distinctly unique method of multiple-board composition. Boards are not only stacked up as in traditional rack-mount systems, but are also vertically connected using *elastomeric connectors* [15] to construct a 3D-mesh topology. Compared with TCA, J-machine not only shares the similarity of 3D-mesh topology, but also includes the availability of direct and short physical data channels between vertically immediate 2D planes of computing nodes. However, there is also an obvious difference. A single composition unit of TCA, *tile*, occupies a lower area, whilst the J-Machine board is considerably much larger, integrating multiple chips on a single board in two dimensions.

In 1994, a 3D optical interconnection network, *Optical Multi-Mesh Hypercube* (OMMH) [31], was explored. The network topology in this system is a mixture between hypercube and mesh, aggregating the advantages of each topology, for example, small diameter, symmetry, constant node degree, and scalability. Regarding the construction, a plane, referred to as an, *Optical Interconnect Module* (OIM), is located between a couple of PE planes. A conceptual construction is shown in Figure 2.13. The OIM directs the optical beams from one plane to another. Despite a number of advantages of the optical techniques in this work, there is also a disadvantage of the immature technology for implementation at the time of this particular work [32].

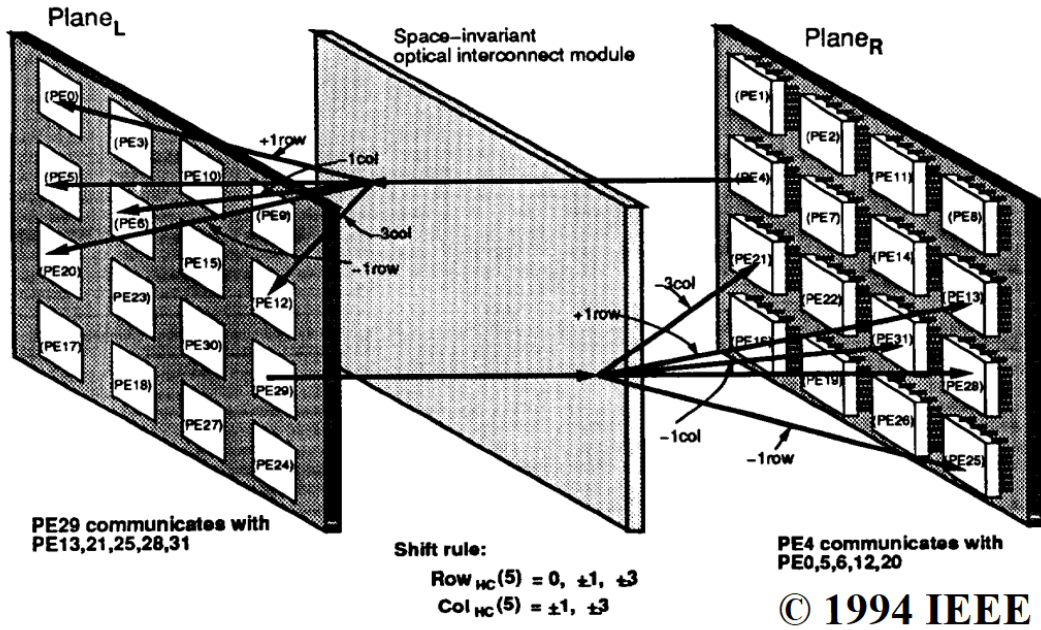


Figure 2.13: A conceptual optical realization of the space-invariant five-cube network. (reprinted from Figure 7(b) in [31])

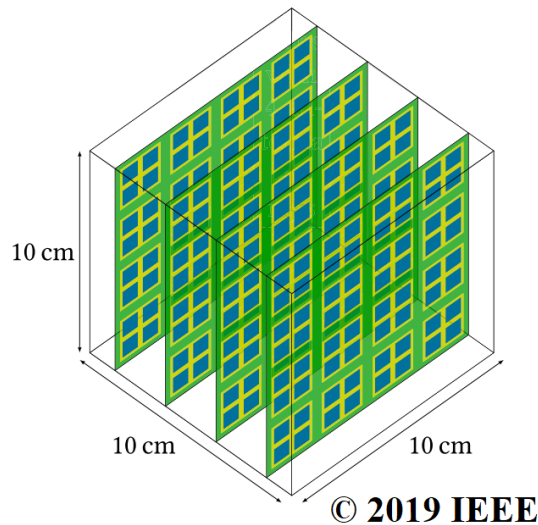


Figure 2.14: Sketch of the HAEC Box. (reprinted from Figure 2(a) in [33])

HAEC [33] is another system proposed to employ optical communication. A small conceptual implementation in this work is called HAEC Box. This system employs onboard optical waveguides for the onboard level, and wireless networking at inter-board level. Whilst the concept is reported that the boards communicate together in a rack, the sketch of the HAEC Box shown in Figure 2.14 shows a conceptual form,

which is considered as having a degree of difference from the obviously traditional form of rack-mount systems in the previous subsection. Thus, with this particular visualised small-scale implementation of the concept, it is categorised in this subsection. However, the wireless communication in this work shares the advantage of direct communication between neighbouring nodes at inter-board level to avoid network topologies to be physically implemented via a back-plane, which is similar to the core idea of the TCA concept (wired communication in the current prototypes, but also in earlier designs via wireless methods [34]).

Unlike the large number of rack-mount systems existing, whilst it is to the best effort in this thesis to survey all of the unusual inter-board packaging techniques, only a few of such systems have been found. Thus, this subsection will end with discussing a series of previous work that has paved the way and is considered highly-relevant to the TCA concept. In [35], the proposed concept is to alleviate the effort of assembling a cluster of computing devices, each named, 'ball', encapsulating computing elements wirelessly powered and intercommunicated. An illustration of the concept is shown in Figure 2.15. With non-wired power delivery and communication, these computing balls can be randomly and rapidly put into a container. The powering method envisaged in this work is not limited only to a light source, but also, theoretically, by liquid, which at the same time, serving cooling purposes. However these concepts lack the practicality of wired power grids.

Regarding the unusual wireless power and communication for clustered computer systems, [36] and [34] further investigated the conceptual idea in Figure 2.15: For the feasibility of implementation, [36] summarises wire-free power transmission into six subcategories as shown in Figure 2.16, and also provides a brief overview of two wireless technologies, which are microwave and infrared systems. Focusing on the wireless data communication, [34] investigated the viability of a wireless interconnect network for a highly parallel computer by modelling and creating simulation and visualisation tools for evaluating network performance of the same ball packaging concept. To evaluate the interconnection network performance proposed in [34], a level of abstraction of task models is proposed.

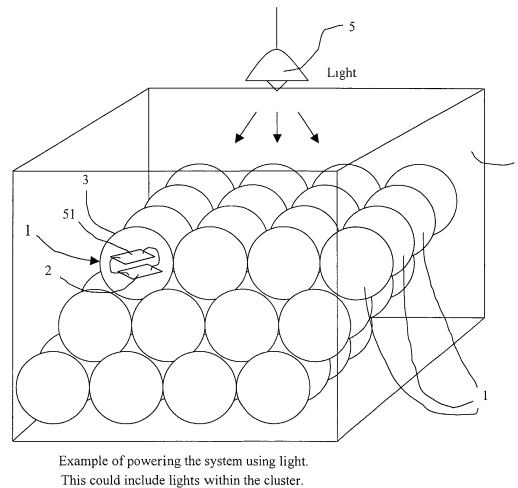


Figure 2.15: An illustration of a cluster of ball-shaped computing devices. (reprinted from [35])

With all the previous work in this series of *prior-TCA* work, it can be considered that wireless techniques for power and data communication have been attractive. However, the wireless computing idea for high-performance systems still needs several improvements such as modelling and simulation methodologies. In contrast with wireless network simulation, the modelling and simulation tools in the field of wired interconnection network have been extensively developed. A large number of these simulators, which prove to be mature, were also surveyed in an early stage of this thesis. Therefore, interfacing the *power-per-node* data to one of these existing tools could require less effort compared to systems with wireless communication. BookSim2 [37] is one of the prominent tools in the survey, which will also be discussed due to its attractiveness for interconnection network performance simulation in the future.* This leads to a motivational decision seeking simple, viable, solutions with current widely-used technologies, to employ wired communication in the hardware prototypes at the current stage of the research.

2) Tiled Computing Architectures

Having proposed the wireless ball-like shape computing devices in the previous prior-

* Although experimental modifications have already been being carried out, the author of this thesis considers that a number of good practices for tool-modification processes, for example, test cases, completeness of overall functions involved (statistical-results logging functions, etc.), should be carefully performed.

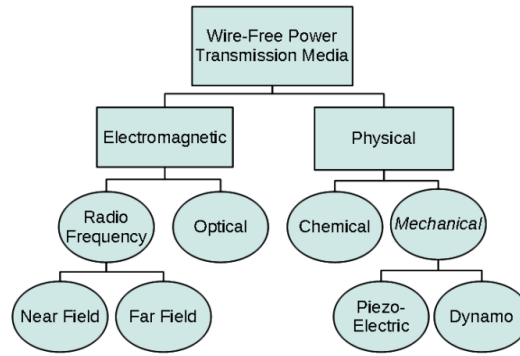


Figure 2.16: Wire-free power transmission media. (reprinted from Figure 3.1 in [36])

TCA category, [38] and [14], briefly introduced a *hexagonal* unit for constructing a 3D array of multiple nodes. Subsequently, the variant was official published as patent [12]. The hexagonal-shaped unit proposed does not focus only on dense computing nodes, but also cooling fluid concerns in order to dissipate heat. In [12], inter-module power and data connectivity is implemented at each of the module's trapezoidal edges. However, it does not detail how the power network for large-scale systems can be handled. For this hexagonal-unit variant, to mitigate the inter-node hop-count issue, [14] also discusses a possibility of wireless interconnection. However, compared to wireless networking, the traditional wired interconnect method is simpler, and cost-effective, using the mentioned trapezoidal areas to form short and direct inter-node communication, and also without any inter-node wires.

Unlike the external power rails directly fed to each board in typical rack-mount machines, the power-network in the TCA concepts tolerates faulty inter-node rails at a degree as multiple-edged power connecting points are available. A single unit, like a *hex-tile* as called in this thesis can provide not only the diversity of power rails, but also intercommunication routes by utilising the unique advantage of mesh-like topology. This unusual power-route advantage, in another way, is also a challenge for understanding of the unforeseen behaviour of the power network and electrical constraints. A comparison of different power delivery systems can be found in Table 2.4.

To emphasise the topological aspect in a parallel/distributed machine, there could be several component levels for hierarchical typologies, ranging from the entirety of a

system, down to the node level. For instance, the system level can be constructed as a 3D mesh, whilst at the node level, a sub-topology can be formed as a network of homogeneous or heterogeneous PEs such as ring, tree, or full mesh. It is also possible that in a heterogeneous system, each node may contain a different sub-topology than the others. These multiple topological levels can also be applied to TCAs. The topologies focused in this thesis are at *inter-node* level, which sometimes, is referred to as *inter-board* for clarification as the current tile prototypes are implemented using PCBs. However, for future TCA designs, a unit can possibly contain computing-related components composed in a way that may look completely different than a 2D board in the present or even a single highly integrated VLSI die. In this thesis, only the power network is intensively investigated, however, a simple wired communication method is implemented in the prototypes for simplicity in the current phase of investigation.

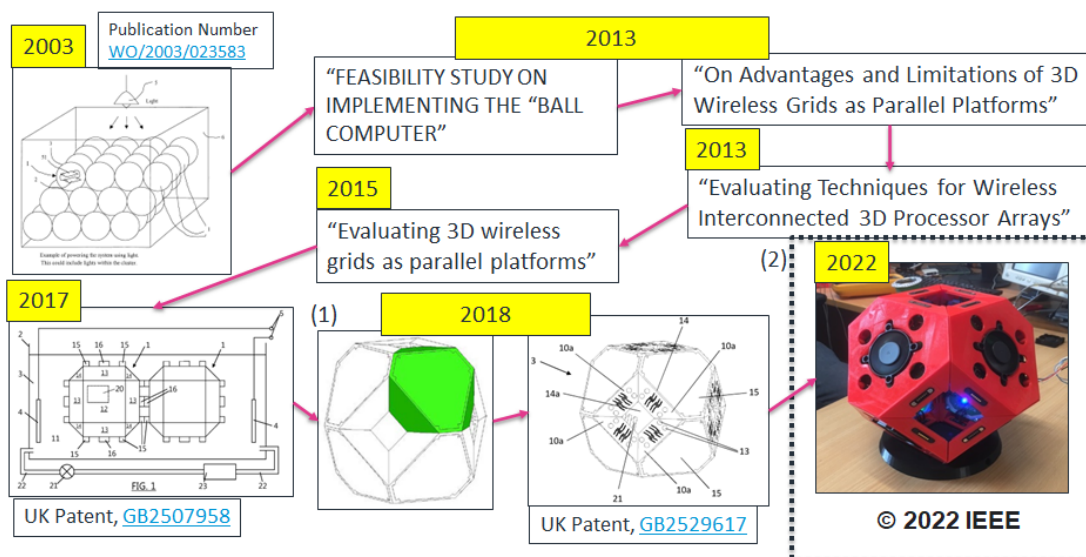


Figure 2.17: Timeline of TCA concepts and its prior work in the series of development. All the details in the timeline can be found in [35], [36], [39], [34], [40], [41], [14], [12], and [42]. ((1)-(2) reprinted from Figure 7 in [14], and Figure 8(c) in [42], respectively)

Considering a single rack as a multiple-node container, if nodes as seen in Figure 2.18 operate at low-power budgets, both traditional rack-mount and TCA machines seem to be more or less equal in terms of node density as chip-level cooling, e.g., per-chip fan is not necessary. For the traditional rack-mount systems, both the intra- and inter-rack power delivery systems occupy some physical spaces, which reduce the contiguous spaces amongst the PEs. However, for a TCA design concept, it does not

limit the contiguous size of network, as long as the power-related constraints, e.g., inter-node pin currents and node-level voltage drops are in the acceptable ranges. This means that nodes can be expanded consecutively in all the three dimensions under the required electrical, cooling, and other engineering factors. In the sense of the power-grid issue, this thesis essentially seeks to answer whether a large-scale of power-grid sizing is feasible to be implemented with current technologies. Figure 2.17 depicts the timeline of related prior work towards the TCA concepts.



Figure 2.18: A panorama of the SpiNNaker 1 million core machine. (reprinted from [43])

Table 2.4: List of hardware and network configurations in the surveyed parallel/distributed computers.

References	Topologies/ Inter-board Communication	Inter-board Packaging	Power Delivery	Implementation	Power (kW))
J-Machine ^a [30]	3D mesh/ connector	chassis	bus bars [30]	chassis	
OMMH [31]	optical multi-mesh hypercube/ wireless (optical)	not specified	not specified	conceptual	-
[32]	hypercube and mesh/ wireless (optical)	not specified	not specified	conceptual	-
MDGRAPE-4A [44]	3D torus/ cables	rack/cabinet	back-plane [29]	rack/cabinet	65 [28]
HAEC [33]	wireless configuration/ radio	HAEC Box	not specified	HAEC playground (network-protocol evaluations)	1 [33]
[35],[41], [36][34]	wireless configuration/ radio	ball-shape object	wireless	conceptual	-
ExaNest [17], [18]	hybrid [19]/ cables	rack/cabinet	back-plane	rack/cabinet	60 [17]
A variant in [14]	3D mesh, (4D hypercube at inter-PE level/ wireless (radio))	ball constructed from hexagonal module	not specified	conceptual	-
SpiNNaker [3], [4]	board level: hexagon [27] system level: hexagonal torus [5]/ cables	rack/cabinet	back-plane [45], [46]	rack/cabinet	75 [3]
SpiNNaker 2 [47] ^b		rack/cabinet		rack/cabinet	
Frontier ^c [24]	dragonfly [48], [49] cables	rack/cabinet		rack/cabinet	21,100 [50]
This thesis	3D mesh (3D torus with external data channels)/ connector	similarly to [14],[12], also investigating power-network for large-scale systems	3D power grid	hexagonal board and frame prototype	see Chapter 5

^a J-Machine is not a recent machine at the time of the research, thus, the power information is not in the interest of this thesis for comparison.

^b As Spinnaker 2 is also a recent machine at the time of the thesis, explicit information for topologies, inter-board communication, power delivery and consumption, are not found in [47]. However, [47] discusses that it is a development based-on SpiNNaker 1.

^c A single node is implemented as a blade server. It implies some form of back-plane power. However, no explicit relevant information is found in [48].

2.4 TCA Physical Scalability: Possible Constraints

Having described the research motivation in Section 1.2 and detailed comparison of various packaging systems found in the literature in Section 2.3, it elaborates how the TCA concept physically eliminates such issues arising from packaging methods employed. However, there can also be constructional constraints specifically in tiled computing structures, which will be discussed in the following subsections. Physical engineering factors are left as future work, whilst the power distribution grid is the main focus in this thesis. The investigation of this power network could not only lead to insights into how it impacts on the physical scalability, but also involves other aspects, such as overall system computing performance and comparability.

2.4.1 Power Networks in TCA Systems

Whilst traditional rack-mount systems share similar power delivery structures, e.g., back-plane, the TCA concepts introduce different possible power-network topologies relying upon the unit shape and how they are coupled to construct the whole system. The hex-tile unit as proposed in [12], is considered a variant design that constructs a 3D-mesh power-network topology. Apart from the fully-connected external power connectors at the array surfaces, internal layered-nodes are impacted by voltage drops. This is due to the fact that powering nodes in this kind of tiled structure requires inter-node electrically conductive media to be laid out in 3D space to distribute power demanded by the whole meshed computing nodes. The voltage-drop issue depends on variously collective factors, for instance, external voltage sources supplied, the design of inter and intra-node conductive media employed, the efficiency of intra-node power regulation units, and all the components consuming power. This is effectively a highly complex series-parallel resistance network. Designing a TCA system without the voltage drop issue taken into account could encounter several consequent issues for large-scale systems, e.g., more demands of higher currents and/or voltages from the external power supply units to maintain the acceptable input-voltage ranges distributed all over the computing nodes in the entire system, or brown-out problems at

the regulator output stages. Voltage drops on inter-node power media also mean the waste of power for non-computing components.

Apart from this constraint of voltage drop, another electrical constraint in TCA systems is *inter-node power rails' current limits*. With high voltage externally supplied, a TCA system may not significantly suffer from the voltage drop issue if the size of layered-array is not at a large scale. However, with high-wattage nodes in large-scale systems, they could draw large amounts of electrical currents flowing through networked power rails and finally hit the current limits of the power rails. Depending on how these power rails are implemented, the maximum currents allowed over a single rail can be a small value, e.g., under 1A, or much higher such as 10A. Amongst the other constructional concerns, these *dual constraints*, voltage drop and power-rail current limit, are non-traditional power-related factors that should be taken into account. The voltage drop constraint may only cause nodes not to operate within an acceptable range of input voltages. However, an exceeded current limit can be the undesirable root cause of many issues, ranging from degraded electrically conductive media to permanent component damage, which could consequently lead to adversely catastrophic physical failure in large-scale systems.

2.4.2 Physical Engineering Factors

The concepts of TCA try to remove non-computing components out of the internal volume of the system and, ideally, to allow the actual PEs to be located as close as possible together in a physically contiguous 3D space. However, inherent heat dissipation from nodes is another factor of concern. For this reason, the TCA concept is designed with cooling-aware structure, having free-space pathways for air or liquid to cool down each unit. Whilst a design of TCA provides empty spaces for unit-level cooling, it is considered not adequate to only pay attention at this small-scale. Finally, all the heat produced by nodes must be managed to be taken out of the system, meaning that system-level cooling is also an important issue. A notable example of liquid cooling system is [51]. This cooling solution is developed for multiple large computing boards. However, water cooling is strongly expected to be specialised

for TCA to take the cooling-aware advantage of the empty 3D spaces. The cooling containers in Figure 2.3 are examples of environments that a TCA array could be entirely submerged within. An illustration for a submerging idea from [35] is also shown in Figure 2.19.

Unlike rack-mount systems, the cooling system used in a TCA system may also indirectly affect the system power-network scalability. This is due to the altered resistances of inter-node power media, affected by the temperature in the system. Additionally, if an *in-system* cooling method itself also consumes power from the power network, e.g., node-level fan or impeller, this is also a factor impacting on the two power constraints. With appropriate designs of cooling and the media employed, for example, inter-node connector pins, the effect of heat projected to inter-node resistances would be considered negligible. This correlated cooling issue of a TCA design is not the main focus of this thesis and expected to be tackled in the future.

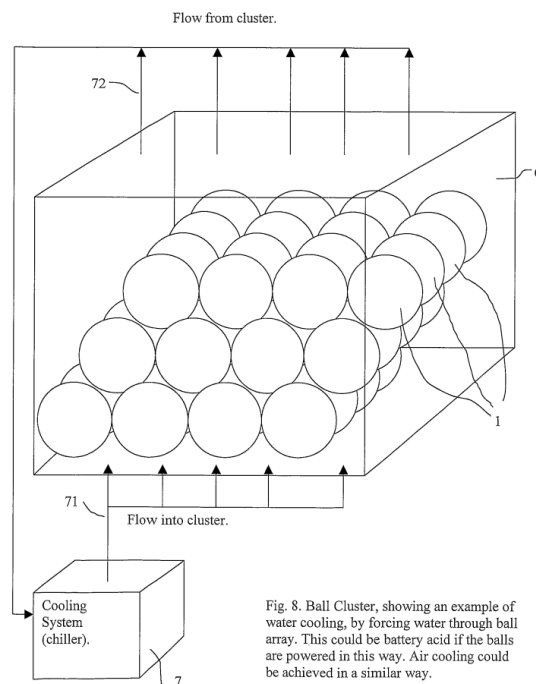


Fig. 8. Ball Cluster, showing an example of water cooling, by forcing water through ball array. This could be battery acid if the balls are powered in this way. Air cooling could be achieved in a similar way.

Figure 2.19: Example water cooling system for the ball-shaped computing devices discussed in Subsection 2.3.2. (reprinted from Figure 8 in [35])

There can be other issues related to building a physical TCA machine aside from the power network and heat dissipation. Materials and other construction-related issues

are also example factors involved. Traditionally, computing boards in rack-mounted machines may not cause weight issues as the frame of the rack itself supports all the weights of the nodes (computing boards). On the other hand, the bottom plane of a TCA array may need to carry all nodes above if there is not another weight-distribution technique involved. However, the weight issue may also be considered together with cooling. By submerging an entire TCA array in a liquid container, some form of high-density liquid may also be investigated to 1) partially support the system weight, and also 2) as its main cooling-purpose, to cool the system temperature down. Physical engineering factors are also not the main focus of this thesis. However, during the development of hardware prototypes, these factors will be empirically manifested by themselves alongside the building of prototypes.

2.5 Related Concepts for Power Modelling

This section discusses the survey of related background and power models at important levels of components in typical parallel/distributed machines. The contents in this section will not be discussed in terms of comparisons in individual subsections, but rather from the perspective of employing the relevant knowledge, methodologies, and models, required for the modelling and simulation framework proposed in this thesis and future work.

2.5.1 Overview of Power Modelling in Parallel/Distributed Computers

Several previous works in the research field of power modelling in parallel/distributed computers have paid attention to various components, ranging from the computing node itself to the power characteristics from the whole system. In [52] and [53], several power and energy models for HPC systems have been compiled, categorised by components in the systems. It can be found that the interests of the surveys focus

upon power and/or energy models for the whole power-figure of a system, interconnects, down to the node level, but not those of power delivery methods. Typically the assumption is made that traditional back-plane, e.g., bus power distribution or some standard approach will be used. However, this may not always be a realistic option, and certainly not in the case of the tiled TCA model.

2.5.2 Typical Inter-board Level Power Delivery in Parallel/Distributed Computers

In Table 2.4, some power delivery methods are summarised along with other important properties on various related works regarding packaging techniques. In this subsection, it is dedicated for a summary of power delivery in typical parallel/distributed machines. Apparently, the typical power delivery methods used at inter-board level are some forms of back-plane based power configurations. Detailed surveys of this kind of back-plane-centric power delivery are not in the scope of this thesis, as it is obviously not comparable with the mesh-like power network in the TCA concept, which laid out all over the entire a 3D computing array. However, to give an overall view of this kind of power delivery method employed in typical machines, a short discussion in this subsection is provided.

For a holistic picture of power supply system, Figure 2.20 shows the structure of traditional data centre supply system. It can be obviously seen that the system is dominated by DC power supply. Whilst all the loads are DC-powered, the original power sources can be either AC or DC. To narrow it down to power components near to the actual computing boards, various systems may employ or develop their own power supply units (PSUs), and/or power distribution units (PDUs). Generally, one of PSU's functions is converting AC to a well-regulated DC power source. However, it is also implementation specific. Some systems may only perform AC-to-AC step-down conversion. Respectively, a PDU, as its self-explanatory name, is responsible for distributing power through some sort of network. A PDU can be a simple small power strip, or some form of a back-plane bus as shown in Figure 2.21. The appearances of PDUs are not limited to these rectangular-shaped units. Some implementation

alternatives may not even be insulated. One heat-related reason is to let the conductive media cooled down by the cooling system employed in the facility.

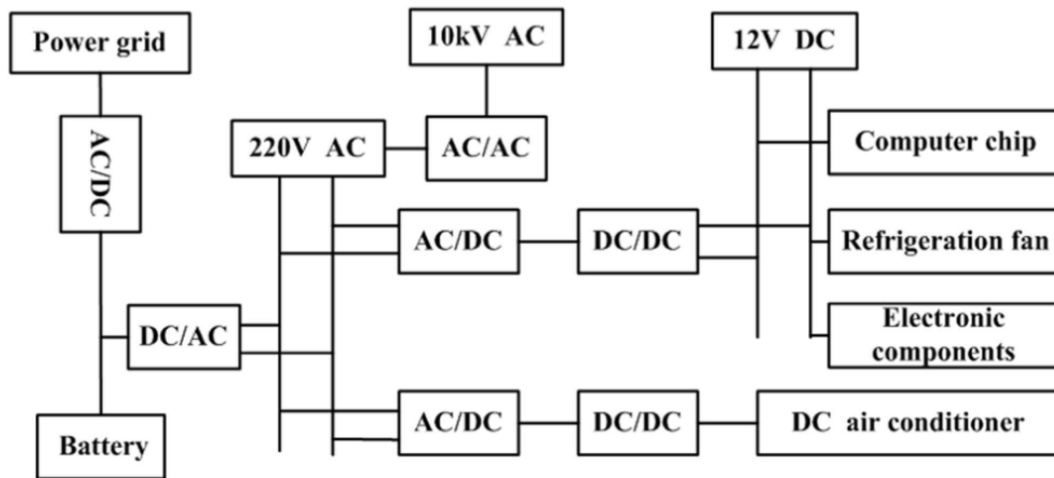


Figure 2.20: Traditional data centre supply system. (reprinted from Figure 1 in [54])

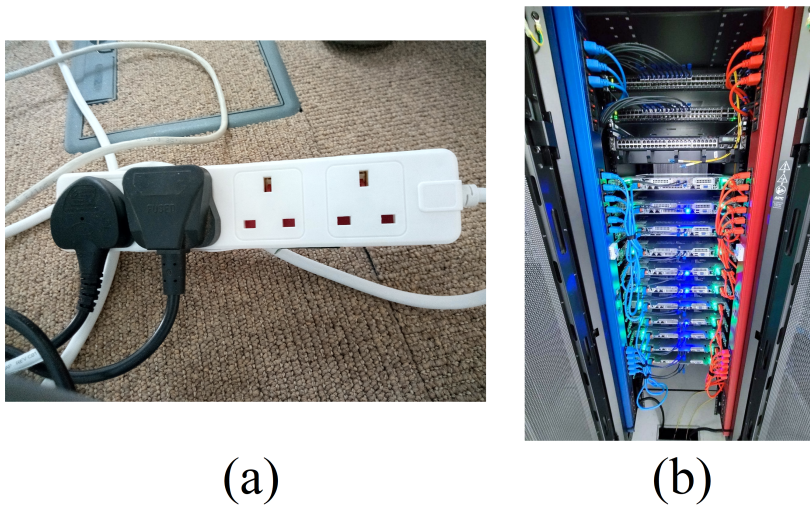


Figure 2.21: Examples of power distribution units. (a) shows a power strip for AC power, and (b) shows two rails of back-plane PDUs which can be found in a rack or cabinet. ((b) is reprinted from [55])

Relating to the inter-node power method in TCA, the meshed power network can be technically considered as a form of PDU. This is due to the fact that in traditional rack-mounted systems, the PDUs are the components located at some certain locations, but typically at the back-planes, to supply power to computing boards mounted in a rack. With the same logic, the meshed power-network in TCA distributes power to each of computing nodes all over the system, hence the name *TCA power distri-*

bution grid. Internal paths between power connector facets form an internalised PDU pathway, whereas the PDU pathways are external in traditional PCB and back-plane systems.

Given a concise picture of typical inter-board level power delivery, it is obviously considered that the methods employed in traditional rack-mounted systems are significantly different from the TCA's power delivery investigated in this thesis. In the traditional systems, a unit, typical a board, is directly supplied with the required voltage. Whereas, the input voltage for a node (tile) in TCA, can be impacted by several factors, for example, the implementation of power topology, correlated failures from power rails nearby, electrically conductive media of the power-grid network itself, plus the dynamics of the internal power loads at the node level such as PEs carrying out some computing tasks with varying workload.

2.5.3 Power Modelling at the Computing-board Level

For computing-board-level power modelling, a possible method, however, considered complicated, is to measure separate power consumed by intra-board components, such as voltage regulators (if equipped), network chips, and PEs. Using this method, the whole board-level power consumed can be broken down to individual components contributing their own portions of power. The difficulty may arise upon this method as all the components in a pre-manufactured board may have been well laid out without voltage and/or current readout points on the PCB. In another way, a simpler method is measuring the whole board power consumption during its operation. In particular, this method is preferred in this thesis as from the perspective of the power network, the whole node-level circuitry affects the input voltage and current, which is adequate for the scalability evaluations in this thesis. However, in this subsection, only some attractive PE power models are discussed as PEs are the main actual units to perform computing workloads. However, as mentioned earlier, other power and energy models can also be found in detail in [52] and [53].

There can be many scenarios of interest in terms of statistical power data, for instance, stand-by, average, or the worst case. To make the situation even more complicated, considering only the PEs, they can continuously and dynamically vary their own power loads over a fine-grained period of time domain. For CPUs, as a PE type, special processor-level registers called *performance monitoring counters (PMCs)*, can be employed to create models by relating events involving the activities of the CPU. Using this method, the model can predict power estimation at a fine-grain level over time. Examples of PMC-based models can be found in [56], [57], [58]. The advantages of the PMC technique also comes with some drawbacks. As these specialised hardware counters are required, CPUs without PMCs can obviously not take advantage of this processor-level power modelling idea. Another CPU-related power modelling technique is to use *CPU utilisation*, [59], [60]. With this method, the CPU utilisation reported by the operating system can be related to the CPU's power, without requiring PMCs.

Another widely-used type of PE discussed in this subsection is *field-programmable gate array (FPGA)*. The idea of CPU utilisation may also be feasible for relating power to computing workloads in an FPGA if the interested re-configurable area is implemented as a *soft-core* processor with an operating system running on it to report the soft-core CPU's utilisation. However, a soft-core processor is implementation specific, depending on several factors such as the micro-architecture itself and the FPGA synthesis tool employed. In another way, an FPGA vendor may provide a power estimation tool, for example, Xilinx[®] [61] Power Estimator (XPE) [62]. The tool provides a spreadsheet file helping with estimating the power consumption of an FPGA by inputting various parameters of the components inside. With this tool, a user can estimate the worst-case power by adjusting all the parameters to be both temporally and spatially active.

2.5.4 Voltage Regulators

In this subsection, voltage regulators are discussed in the context of how they are required in TCA, rather than in terms of comparison of various types of regulators.

Typically, modern PEs, for instance, CPUs and FPGAs, are designed to support low input voltages from 3.3 V or lower [63], [64]. Apart from the PEs, I/O circuitries also rely upon operating ranges of supplying voltage requirements. This results in input voltages of computing boards required to be in acceptable ranges of the voltage regulators employed. In traditional rack-mount systems, as mentioned earlier, PDUs may distribute AC or regulated-DC power directly to each node (computing board, e.g., rack/blade server). However, in TCA it incurs voltage drops in internal layered-nodes. Thus, the considerations of voltage regulator in TCA are not only concerning adequate levels of external voltages supplied at the surfaces, but also ensuring the voltage regulators employed are suitable for the voltage drop issues in the system. With this reason, *switching regulators* are preferred choices compared to *linear* ones due to their higher efficiency, lower heat dissipation. Figure 2.22 shows some examples of power efficiency curves.

Regarding the voltage drop issue in TCA, it can be seen that the system can be powered by supplying voltages at its surfaces with two different power rail models - 1) *Direct supply rail model* - the board's required input voltage supplied at system surface + without local regulators. This supplying model does not require a voltage regulator for board input voltage, but will bring about the following problems.

- ▶ **Board input-voltage stability:** This will occur due to the voltage drops caused by the *internal* and *contact* resistances of the inter-node power media such as connector pins, and also the varying power by dynamic computing workloads.
- ▶ **High currents in the power rails:** As the modern PEs in a computing board tend to be designed for low voltages but consume high wattages, directly supplying the required regulated voltage at the system surfaces could cause much higher currents all over the system. This issue will also negatively impact on the power wasted on connector pins, and also dramatically reduce their lifespans.
- ▶ **High-current capacity power supply units:** As high currents are required, it is straightforward that the power supply units employed need to support high-current capacity. This may also require large-sized external conductive media such as very-wide diameter power wires. Higher specifications of the power supply units also mean higher costs of equipment.

The other power rail model is - 2) *local regulated power model* - external higher voltage + with local node-level regulators. Using this voltage regulation model, all the problematic issues mentioned can be mitigated. Thus, local voltage regulators are employed in the current TCA design. Switching regulators, as aforementioned, have their own advantages. However, for simulation purposes they also have some drawbacks that intensively affect simulation efforts in this thesis. With this difficulty, one comes to the next subsection to discuss the suitability of modelling switching voltage regulators in this thesis.

TYPICAL PERFORMANCE CHARACTERISTICS $T_A = 25^\circ\text{C}$, unless otherwise noted.

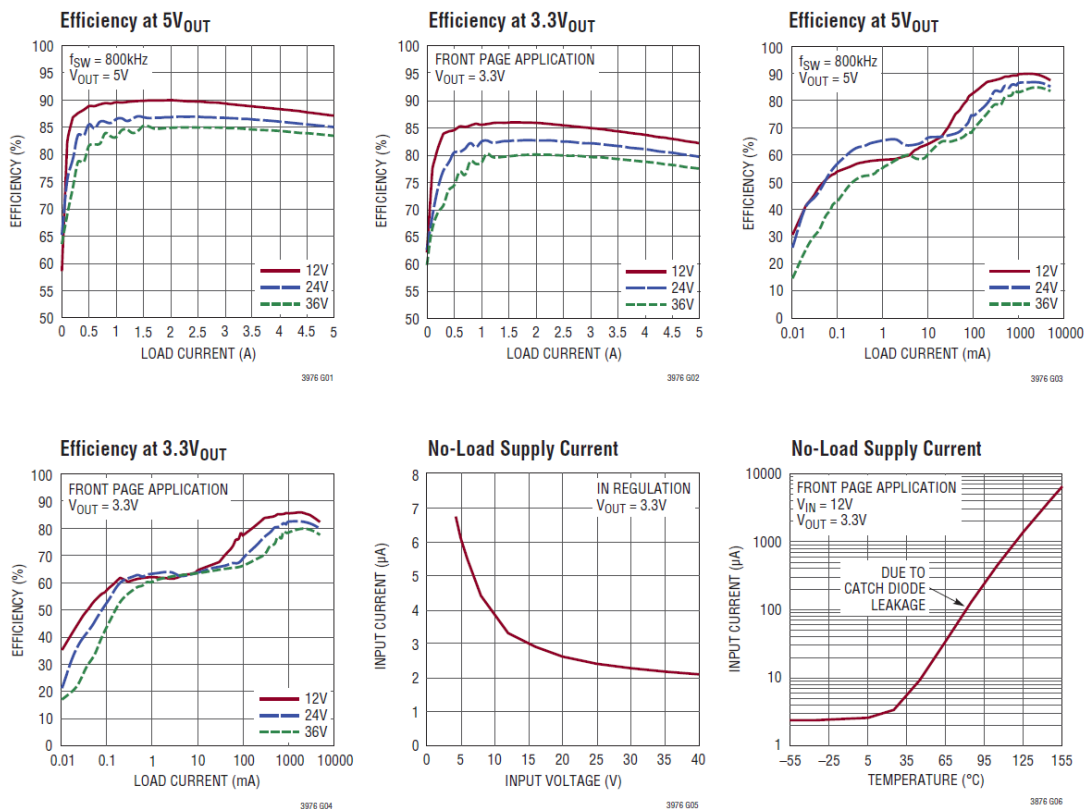


Figure 2.22: Examples of switching regulator efficiencies with varying input voltages and load currents. (reprinted from 'Typical Performance Characteristics', page 4 in [65])^a

^a With granted permission by Analog Devices, Inc.

2.5.5 Switching Regulator: Average VS Complex Models

With the complex design of switching regulators, it incurs a drawback of long simulation times. Thus, some semiconductor companies provide *average models* of their switching regulators for simulation purposes together with complex (*cycle-by-cycle*) models. One issue concerning any voltage regulator simulation models provided by any companies is that if some of the ground signals of the models are directly tied to the global ground of SPICE simulation, the simulation results will be inaccurately simulated. This issue can be resolved by spending efforts on understating the internal mechanism of the models and manually editing the simulation models.

However, for proprietary models, the circuit modeller may need to consult with the companies to provide alternative models with non-globally tied grounds. This issue can be illustrated in Figure 2.23. This grounding issue is completely eliminated by the simplified model proposed in this thesis due to the board modelling process only focusing upon the ground level of a single switching regulator simulation model itself, thus it is not affected by the globally tied grounds in the original models.

The simplified model will be discussed later in Chapter 3. Regarding the averaged models, they are able to be manually modelled by the circuit designer to simulate the average behaviours of the regulator employed in the system. [66] provides several average-model methodologies. However, the models require efforts to comprehend the internal mechanism of the regulator and the process of converting a switching model to an average one. Several works such as [67] and [68] also provide automated methodologies for modelling average models.

Concerning the objectives in this thesis, the averaged modelling of switching regulators is itself beyond the scope and not in the interest of this thesis. Rather than investigating the averaged switching models, this thesis focuses on the perspective of the inclusion of a suitable switching regulator model for power-grid simulation design. Thus, it has been briefly discussed here for the completeness as part of the surveyed related modelling methodologies.

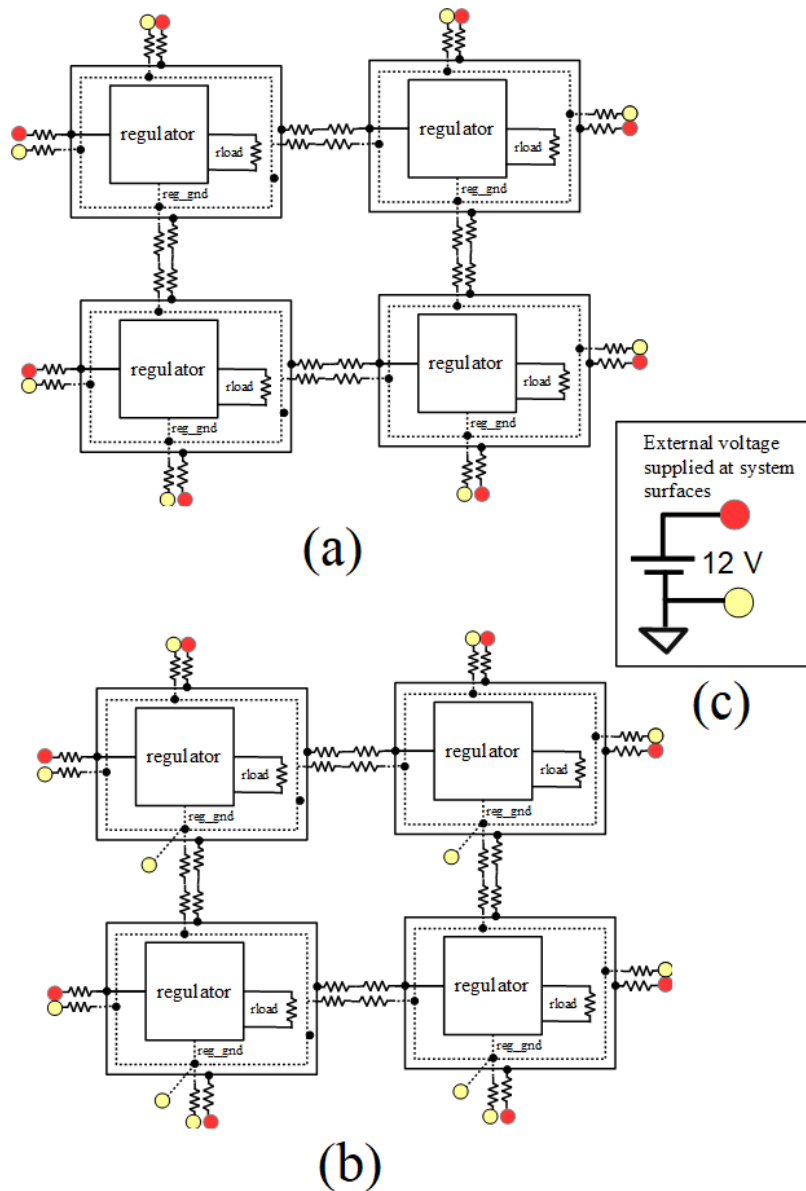


Figure 2.23: Example of grounding issue for a voltage regulator simulation model on a 2x2 conceptual computing boards in a TCA. Each of the resistors surrounding each module represents inter-node power medium resistance. (a) shows a correct modelling, whilst in (b), the ground of each regulator model is directly tied to the global ground of the SPICE simulation. (c) shows the global positive and negative (ground) rails. This example is only for illustrative purpose. In the actual TCA systems, it is considered much more complex due to the 3D meshed power network.

In this thesis, this issue of the complex model is tackled by proposing a simplified model by modelling the whole board power consumption as a *single lumped-resistor**. This technique totally eliminates vendor-specifics, for both simulation models and the

* As DC power can be formulated as $P = I^2R$, thus, the current following through a specific device is also involved. This will be demonstrated later to show how the models and simulation framework proposed in this thesis achieve this lumped-resistor value in Chapters 3 and 4.

simulator itself, the effort of average modelling, and mitigates long-simulation time issues. It will be demonstrated later that *curve fitting* was adequately effective under the assumption of constant regulated load power [42]. The proposed simplified model will be thoroughly discussed in Chapter 3.

2.6 SPICE Simulators

Given that the surveyed previous work on tiled and meshed power-network models shows that there have not been any such suitable models developed that are applicable, a new simulation tool-set was required, which would in this case relied upon an existing SPICE simulator as part of the custom power-network simulation in this thesis. This is one of the important contributions of this thesis, which will be demonstrated later. Concerning circuit simulators to be employed in this thesis, several SPICE simulators are available. LTspice[®] [69] is an example of free proprietary simulators. Whilst some are available as open-source simulators, e.g., ngspice [70]. In this thesis, LTspice[®] is required due to the switching regulator's simulation model employed. ngspice is also another SPICE simulator employed to demonstrate the SPICE-simulator portability of the simplified models proposed in this thesis.

2.7 Mapping Tasks Into a Parallel/Distributed Machine

Task mapping in this section will be discussed from the point of view of system packaging, which impacts on how computing tasks will be actually mapped in physical locations, rather than exhaustively discuss mapping algorithms as part of literature. In parallel/distributed computers, different application domains can have many forms of data communication patterns as shown in Table 2.5. A single computing job submitted to be executed on a parallel/distributed computer can be broken down into *tasks* communicating together to complete a desired set of computations. Each task in a

job is mapped into a portion of computing elements, e.g., CPU core. This means that an inefficiently-designed mapping algorithm is one of the factors that can adversely impact on the physical distances and hop counts amongst those tasks allocated. Apart from the algorithms performing task-PE mapping, in the hardware-design perspective, the computing-node packaging systems at any level such as the PCB designs in computing boards, how they are actually located, and hardware channels implemented in physical 3D space, can also be factors that impact on physical distances. These are the issues that the TCA concepts essentially aim to tackle.

In Table 2.5, It can be seen that not only *node*, *channel*, and *topology*, as implemented in an interconnection network, will impact on a machine's computing performance, but also *traffic patterns* of the applications running on the network. A traffic pattern represents how each node demands to send data units to a destination node. To perform these traffic patterns effectively from the hardware point of view, the real hardware design structure should be part of considerations for optimising hop counts and channel latency to be as low as possible. On the contrary, channel bandwidth should be maximised. Apart from the task-mapping issue, *wire lengths*, are also another factor, which have already been mentioned in Table 2.3, providing examples of wires' critical lengths that impact on channel bandwidth. Short links and the high availability of low hop-count data pathways are therefore two very valuable attributes for HPC systems.

Table 2.5: "Network traffic patterns. Random traffic is described by a traffic matrix, Λ , with all entries $\lambda_{sd} = 1/N$. Permutation traffic, in which all traffic from each source is directed to one destination, can be more compactly represented by a permutation function π that maps source to destination. Bit permutations, like transpose and shuffle, are those in which each bit d_i of the b -bit destination address is a function of one bit of the source address, s_j where j is a function of i . In digit permutations, like tornado and neighbor, each (radix- k) digit of the destination address d_x is a function of a digit s_y of the source address. In the two digit permutations shown here, $x = y$. However, that is not always the case." (reproduced from Table 3.1 in [15])^a

Name	Pattern
Random	$\lambda_{sd} = 1/N$
Permutation	$d = \pi(s)$
Bit permutation	$d_i = s_{f(i)} \oplus g(i)$
Bit complement	$d_i = \neg s_i$
Bit reverse	$d_i = s_{b-i-1}$
Bit rotation	$d_i = s_{i+1} \pmod b$
Shuffle	$d_i = s_{i-1} \pmod b$
Transpose	$d_i = s_{i+b/2} \pmod b$
Digit permutations	$d_x = f(s_{g(x)})$
Tornado	$d_x = s_x + (\lceil k/2 \rceil - 1) \pmod k$
Neighbor	$d_x = s_x + 1 \pmod k$

^aWith granted permission by Elsevier

2.8 Summary and Implications for Hypothesis

This chapter discussed previous work related to traditional rack-mount systems, the relatively new and novel TCA concepts and highly relevant work, and also other related topics required for the modelling and simulation framework for evaluating the feasibility of TCA power distribution grids. Even though a TCA hardware-variant has been proposed in [12], it has not yet been implemented as a practical large-scale system by any researcher. Thus, several aspects, for instance, engineering factors for practical building, external power supply systems, and also in particular the power-distribution grid, are left as research gaps in non-conventional packaging techniques for parallel/distributed computers. Apart from the physical construction, interconnection network performance is also a subsequent issue that is impacted by different methods of system packaging.

Having evaluated the literature in the field and the techniques and concepts that relate to the original research hypothesis, it will now be clear that there are some key

areas of work that need to be addressed in order to test the research aims set out in Chapter 1, and the question:

Is it feasible to build a large-scale power-grid network of Tiled Computing Array, whilst still scaling up the system computing performance?

And now clearly, the underlying questions include the following sub-questions:

- ▶ What are the necessary design choices for constructing tile-able modules?
- ▶ What are the component characteristics of the power grid in a TCA array?
- ▶ How is system computing performance influenced by the power grid design and limitations?

In order to address these questions, a combination of theory, practical hardware construction and evaluation, and system simulation and modelling will be necessary. The remainder of this thesis explores these issues and then makes suitable conclusions.

The abstract concept of TCA can be applied to construct many variant hardware designs and implementation alternatives. In this thesis, hardware prototypes are built for simulation purposes, for instance, to validate models for accuracy, and also for gaining insights into unforeseen practical issues and suitability in various aspects.

In particular this chapter will attempt to address one of the key research questions and two of the the three related sub-questions as stated below:

- ▶ *Is it feasible to build a large-scale power-grid network of Tiled Computing Array, whilst still scaling up the system computing performance?*
 - What are the necessary design choices for constructing tile-able modules?
 - What are the component characteristics of the power grid in a TCA array?

Answering these questions, in practice, requires the development of suitable simulation and modelling frameworks, and their validation against real prototypes (in order that the simulation framework can be considered accurate and capable of making projections of system behaviour at scale). An important part of that overall aim is the modelling of viable component systems such as the board level models of hexagonal tiles.

From the simulation point of view, an inter-node power-rail medium can be modelled as a *lumped resistor* if considered only the pure resistive property. However, hardware implementation can involve a range of choices, e.g., an off-the-shelf connector with some form of materials used to facilitate power rails or grids. For example, connector pins, a custom design of hexagonal copper-plate used for the same reason (as discussed later in Section 3.3), both capable of supporting large amounts of electrical current, but also incurring manufacturing difficulties.

With the aim of being able to build real prototypes during the PhD research project, various factors were considered such as the research objectives, construction difficulties, expertise collaboration, funding, and research time-frame, some off-the-shelf hardware components are selected for simplicity at this stage of the research, whilst some other parts, for instance, hexagonal tile-frames and PCBs, are built as custom components. Although there can be many possible implementation choices for a single concept, the methodologies for modelling and creating the simulation framework are intentionally as simple but well-structured as possible to support the simulation demands for the hardware prototypes built in this thesis, and also for the adaptation of future variant designs.

In this chapter, the hardware prototypes and each of the models proposed in this thesis will be discussed in detail, whilst the whole discussion of the simulation framework will be separately organised in the next chapter.

3.1 Relevant Research Objectives

Both the research objectives 1 and 2 are relevant to this chapter. Each of the objectives will be elaborated and the sections in which they are achieved are also given as follows:

3.1.1 Objective 1: Employing and Designing Models and Simulation Tools

Methodologies/Activities:

- ▶ **Electrical circuit design:** As the scalability evaluations in this thesis focus upon electrical quantities in the power-distribution grid model, the circuit models involve both of building the full prototype-board model and the simplified board model for fast simulations. These can be found in Sections 3.2 to 3.4.

- ▶ **SPICE simulation:** Understanding SPICE coding for circuit elements and sub-circuit creation, are essential for good practice of simulation tools with modular design. Although the final product of prototype-building is hardware, the prior processes also involve the original voltage-regulator model SPICE-simulation. Thus, this item can be found in Sections 3.2, 3.4, and 3.5.
- ▶ **Model simplification:** As the main purpose of the thesis is to evaluate the scalability of the power-grid model, the drawbacks of the inclusion of switching model of the regulator adversely affect simulation activities in terms of simulation times and machine's performance requirements. A model simplification technique, e.g., curve fitting, is required for large-scale evaluations. A resistor network as a result of the experimental non-conventional conductive shape is also considered a model simplification. Thus, the main relevant parts are Sections 3.3 and 3.4.

Expectations and outcomes:

- ▶ **Parameterised SPICE-model files for simulation:** This is the key outcome, and can be found in Section 3.4.

Success Criteria:

The models need to be designed at abstraction levels adequate for reasonable simulation times and precision. Thus, the models are to be compared against real hardware for validation of accuracy.

3.1.2 Objective 2: Hardware Validation

Methodologies/Activities:

- ▶ **Electrical circuit design:** Whilst the circuit design in the objective 1 can be carried out solely on a SPICE simulator, the circuit design in this objective needs some other careful consideration regarding the hardware prototypes. For example, the limitations of circuit design for hardware validation are primarily driven by the availability of funds, practicality of construction techniques,

and timescales, available in a university research department. Nonetheless prototypes needed to meet certain purposes, particularly allowing power grid behaviour to be physically tested and then providing suitable data for simulator validations. The relevant parts for hardware-prototype circuit design can be found in both Sections 3.2 and 3.5.

- ▶ **3D printing:** Regarding the non-conventional packaging form (e.g. hexagonal frame) existing in the TCA design in this thesis, 3D printing is employed for customising the hardware prototype mounting frames according to the novel tiled geometries. This task can be found in Section 3.2.
- ▶ **Geometry:** The geometric considerations in this objective require the precision of the 3D-printed frame. Imprecise geometrically designed 3D-hexagonal tiles could lead to sub-standard physical mating, resulting in grid-coupling issues. Also, this task can be found in Section 3.2.
- ▶ **Hardware prototype construction:** Section 3.2 is dedicated for this item of task. After all of the designs of circuits and packaging frames are considered, the final process is to assemble the hardware prototypes. For example, this step involves PCB manufacturing, soldering, connector attachment, packaging frame refinements due to 3D-printing errors, etc. This work involved considerable interaction with the Department of Computer Science technical team and their support in construction procedures.

Expectations and outcomes:

- ▶ **A set of working hardware prototypes that can be measured:** The prototypes successfully built can be found in Section 3.2.
- ▶ **Validation simulations:** This set of simulations is to validate the accuracy of important parameters, starting from a single tile, to multiple combinations, and then to complete ball modules. The simulation results are given in Section 3.5.

Success Criteria:

The accuracy of the models running on the simulation framework should be within acceptable error percentages to gain better understanding of real hardware and the variability of actual components in practice (for example, connector resistance, actual

versus data-sheet and so-on). Importantly, a degree of error can be tolerated, but being able to quantify the error range allows the simulation models to be understood to offer realistic projections of performance within the same error range.

Given the relevant research objectives, the rest of this chapter will discuss all the details relevant, then closing the chapter with discussing the success criteria.

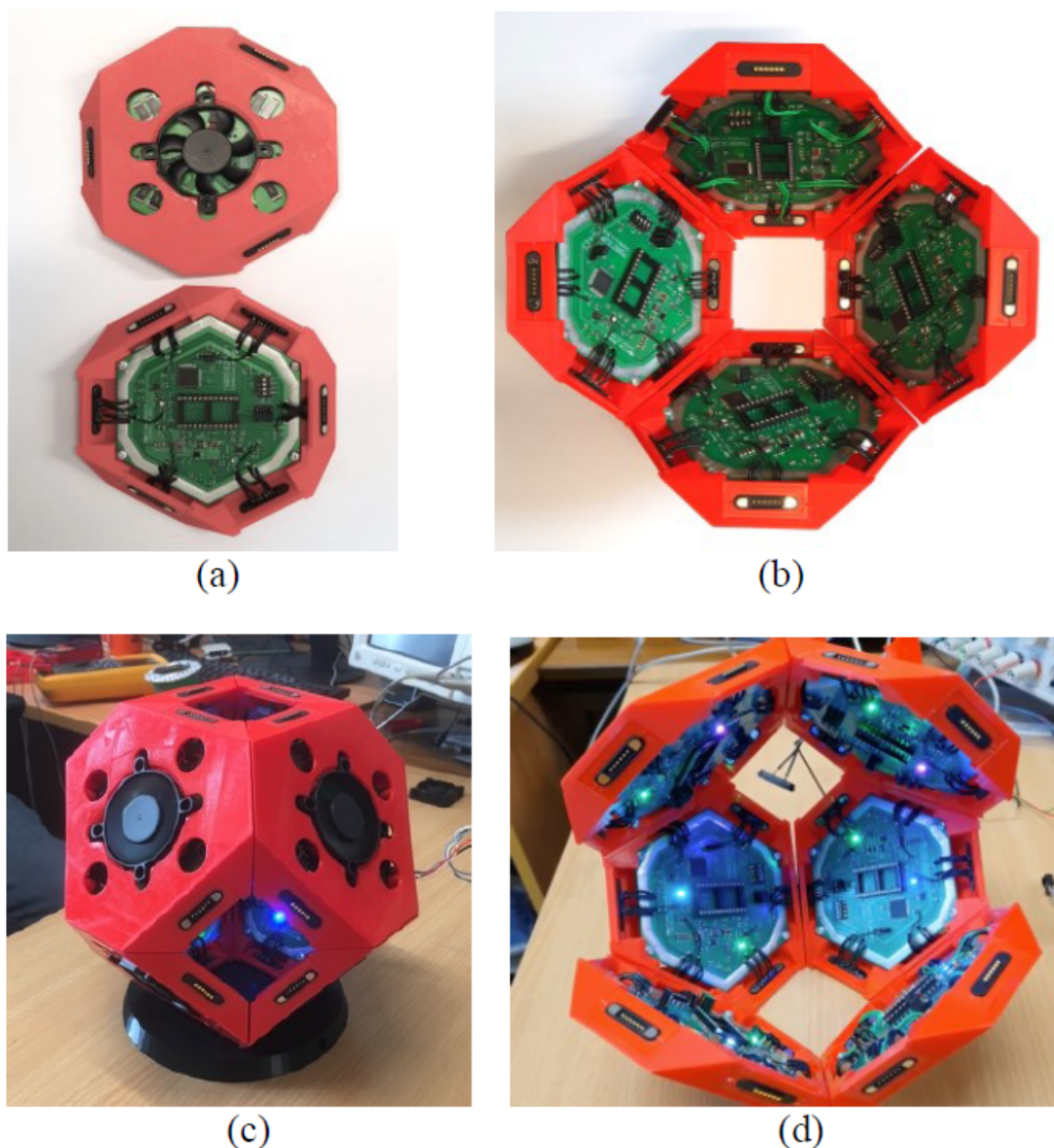
3.2 Hardware Prototypes

The hardware prototypes in this thesis* are considered based-on variant designs in terms of the unit shape and edge power-connection in [38], [14], and [12]. Although the hardware prototype designs in this thesis share some similarities with the aforementioned works, there are some distinctive differences that can be seen in Figure 3.1 such as tile-fan for cooling, a variant of off-the-shelf magnetic connector with pins for data I/O and power. The embedded hexagonal-board inside is also built purposely to support model validations in this thesis, but also provides some flexibility for assembling additional components via an IC socket. The conceptual design of this variant can also be seen in Figure 3.2.

The smallest building block of the prototypes in this thesis, referred to as *hex-tile*, has alternate angles at tile-edges as depicted in Figure 3.2(a)-(c). Processing elements (PEs), e.g., CPUs, and also memory, routing units, and power-conversion circuitry such as voltage regulators, can be embedded inside as illustrated in Figure 3.2(b).

All six bevelled tile-edges can be used for power/ground rails and communication-channel ports, conceptualised as shown in Figure 3.2(d). Each of the communication I/O rails is typically routed from a dedicated pin of the embedded computing chip to another one in an immediate neighbouring tile, however, a shared data channel like a bus is also a possible design to broadcast a physical signal to multiple destination nodes.

* All of the collaborative efforts on the hardware prototypes are given in Author's declaration.



© 2022 IEEE

Figure 3.1: (a) shows top and bottom views of a hex-tile prototype. (b) shows a half-ball (petal) composition. (c) shows eight tile-frames composed as a ball upon a power base-plate providing power via the trapezoidal faces. (d) illustrates a ball with two tiles removed, being powered and demonstrating different power loading by LED colours. (reprinted from Figure 8 in [42])^a

^aIn the photographs, the tiles were powered by wires as the power base-plate were still in an early stage of design.

The outer hexagonal metal grey ring which can be observed in Figure 3.1(a) is a shared power (or ground) rail integrated with the PCB layout, and appears on both sides of the PCB (for positive and ground rails). This tile-level metal ring can be connected to its immediate neighbours via all the six edge-connector pins, forming a unique complex

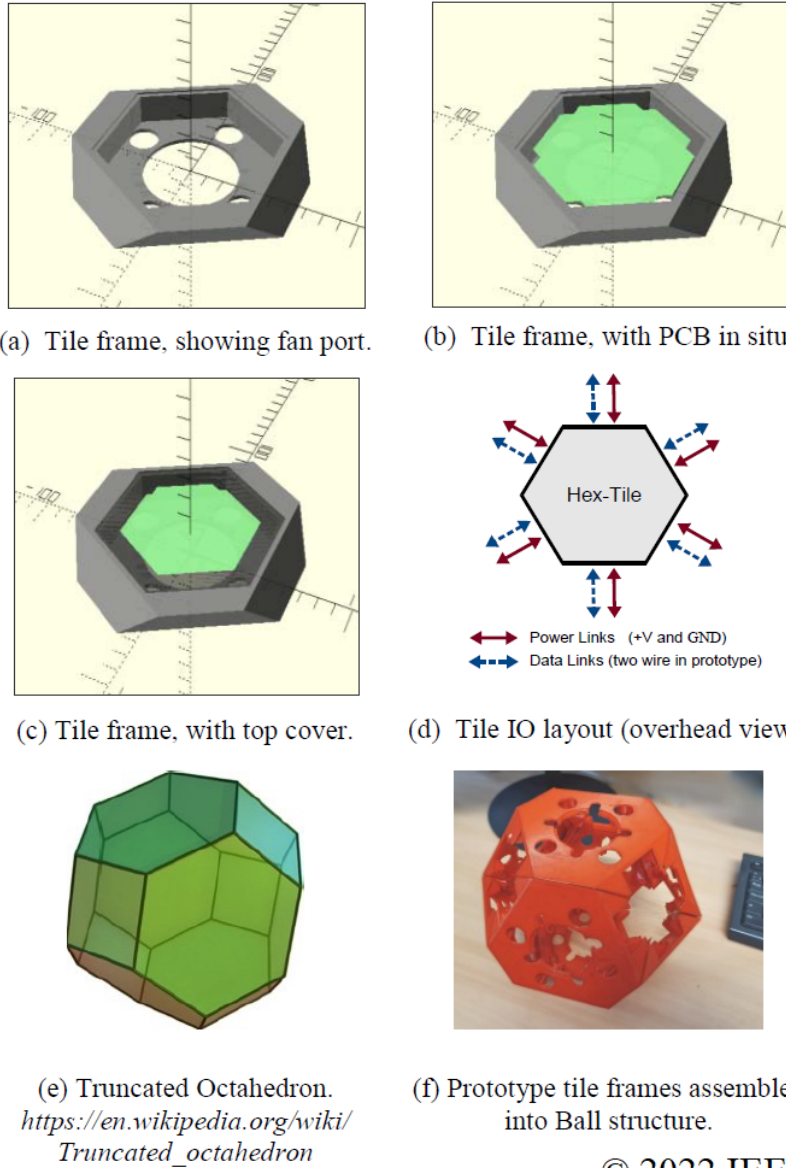
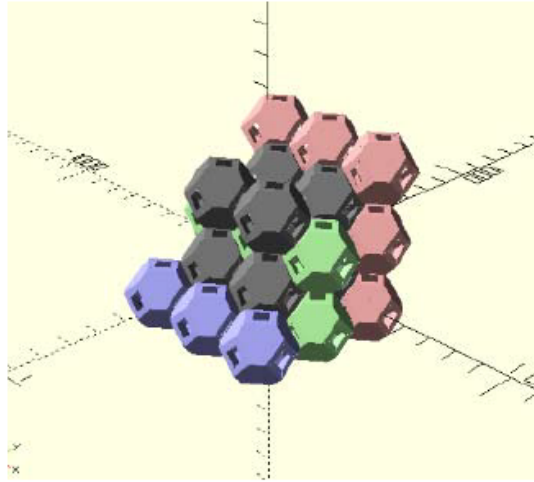


Figure 3.2: Illustrations of conceptual designs and a hardware prototype. (a) illustrates a conceptual model of the tile frame. Examples of possible materials are plastic or ceramic. In (b), the tile frame is shown with an embedded PCB or Multi-Chip-Module (MCM). As shown in (c), a tile may also be covered with a partially transparent material, allowing the visibility of components inside. (d) shows a possible data I/O connectivity upon each tile edge, represented by the dashed/blue lines, whilst solid/red arrows show power and ground rails. Having composed a group of eight tiles in 3D, it can make up a truncated octahedron, a ball-like volume shown in (e). Finally, (f) illustrates a ball-frame prototype. (reprinted from Figure 1 in [42]^a, (e) is adapted from [71])

^a (a)-(c), (f), are generated by Christopher Crispin-Bailey.

power-network of a TCA design. Regarding the alternate angled edge-connectors, a group of tiles can be formed as a 1D, 2D, or 3D-array topology of eight-tiled ball-like unit. A theoretically truncated octahedron shape, which is equivalent to the ball



© 2022 IEEE

Figure 3.3: Example of a double packed internal array of $2 \times 2 \times 2$ grey balls embedded in between the existing $3 \times 3 \times 3$ array. Some of the balls are removed to expose the internal balls. (reprinted from Figure 2(b) in [42])^a

^a Generated by Christopher Crispin-Bailey, using OpenSCAD [72] modelling tool-set

structure can be found in [73] and [74], and is well known to be a permutahedron with high three-dimensional packing density.

The current design of the hardware prototypes allows balls to be coupled together via the trapezoidal edges. However, in future variants, instead of a tile, a smallest unit of a truncated-octahedron shape is also possible. With this shape, both the square and hexagonal surfaces can be options for different power, data communication, and cooling purposes. To increase the node density in this ball-array system, a second group of identical balls can be embedded using the space between balls in the existing array. An example of 'doubled-array' in a single TCA system is shown in Figure 3.3. The total number of balls of a single cubic-array system with n balls per dimension can be formulated as n^3 , whilst $n^3 + (n - 1)^3$ is for a doubled-array system, increasing the density almost twice for large systems.

To supply power to a TCA ball-array, it can be conveniently facilitated at the outer surfaces of the array, where those external balls present trapezoidal connection points (T-facets). Full connection of all of these external T-facets maximises the power capacity to the system and also guarantees the best current distribution in the entire

system. However, given the power-network structure and the power demands inside the system, an exhaustive connection scheme may be unnecessarily overpopulated, suggesting that some other connection patterns with lower connection counts may allow adequate power, whilst still within the constraints of voltage drop and connector-pin current. In theory there will likely be an ideal pattern of external connections to meet any scenario.

With this unique TCA power-network structure, and unlike traditional rack-based system power back-planes, inter-tile voltages and currents behave differently, even when those tiles are consuming the same amount of power. This attribute is due to inter-node electrical resistances from the implemented conductive materials such as connector pins, node-level power consumption itself, and the cascaded networked power structure. It is not only the current-flow pattern in the system that is unusual compared to rack-mount systems as a concern, but also the connector-pin voltage/current specifications themselves at each intermediate point within the grid that must be complied with. This power-related modelling challenge is a key motivation in this thesis, aiming to investigate the models, simulation framework, and hardware prototypes for the scalability of the TCA power network.

Given the detailed concerns regarding the power network, it leads one to define important key constraints for viable TCA systems to operate within those power attributes. The constraints defined should ensure that every tile must be operating without electrical requirements and violations. In this thesis, the electrical constraints focused upon in the power network are defined as follows:

- ▶ 1) The board-level regulated output voltage is within the specified levels, for instance, in many typical computing boards, at 5V.
- ▶ 2) The input voltage of the board-level power-conversion unit, e.g., voltage regulator, is in the specified operating ranges. For example, 7V to 24V. This ensures that the power-conversion unit correctly operates to regulate the desired output voltage in 1).
- ▶ 3) Every current flowing through connector pins does not exceed the current limit specified by the pin manufacturer, or some chosen lower limit.

Heterogeneity is also a factor. In a real TCA system, each tile-able unit may contain different components from the others, e.g., CPUs, memory, pure networking, re-configurable devices such as FPGAs, power-storage, shared memory banks, SSDs, GPUs, etc. Therefore, each of the constraints 1) and 2) above may not be only a single uniform value for all nodes.

Consider for instance, tile *A* may require 5V output-load, whilst tile *B* is regulated at 3.3V with a complete different voltage regulator profile and supplier. Whilst the tool-sets developed are able to accept any variations in constraints, in this thesis it is assumed, unless stated otherwise, that all tiles have the same voltage regulator. Likewise, every regulated load-resistance in the entire system may be different, but each is also assumed constant over a given time domain in a single simulation.

To close this introductory section for the hardware prototypes in this thesis, the prototypes are currently designed to employ wired communication channels. However, it is not limited to a regular 3D-mesh topology, as variant designs can be possible via both the square and hexagonal faces of tile and/or ball compositions. An inherent drawback of a 3D-mesh topology is the maximum hop-count for large mesh sizes. A possible mitigation is to additionally add *wrap-around* channels to construct a 3D-torus topology instead. More advance techniques such as wireless communication is also a potential technique to mitigate wired hops. [31], [32], [34], [14], [33] also investigated optical or wireless (radio) communication techniques. Optical data links are also highly promising, with silicon photonics and bonding of wave-guides to silicon die, providing a road-map for highly integrated I/O in future systems.

3.3 Arbitrary Electrically-conductive Media Designs

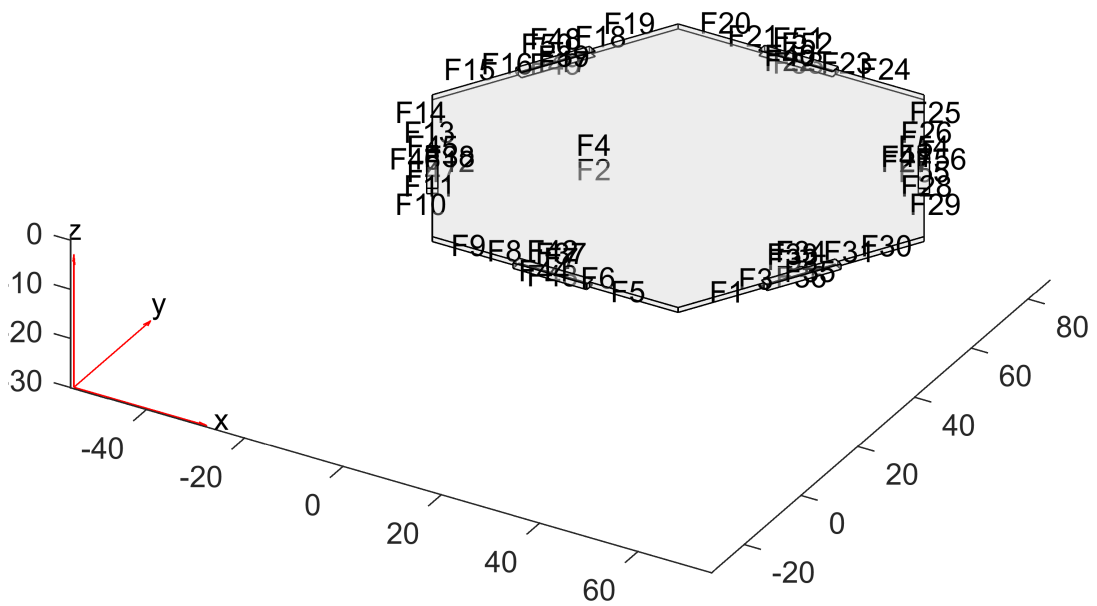
All the main evaluations in this thesis are based on the models built upon an off-the-shelf connectors employed in the prototypes. However, in future work at intra-module and inter-module (e.g., tile) developments, the models are not limited to conventional medium shapes such as rounded/rectangular pins or internal PCB power rings or planes. In this section, a design flow for simulating arbitrary conductive shapes is

briefly discussed to provide an overall view for building custom shapes to delivery power to TCA nodes. This particular piece of development would also further open more investigations into both power capacity and heat issues, as the conductive media may also be employed for cooling.

Apart from the power-delivery network, a specific type of the electrically conductive medium employed in a computing unit may not only serve power delivery purposes, but can also help with heat dissipation, for instance, by direct contact with the IC packages to draw heat being dissipated. Thus, designing a power medium can also simultaneously involve cooling purposes. At an early stage of this research, some methodologies and software packages/platforms for this design choice were also investigated. The investigation utilised existing specialised tools, rather than deeply investigating into the properties of materials. As the current stage of the TCA prototypes focuses on DC current, therefore, DC conduction analysis will be discussed for the possibility of custom-shaped conductive media.

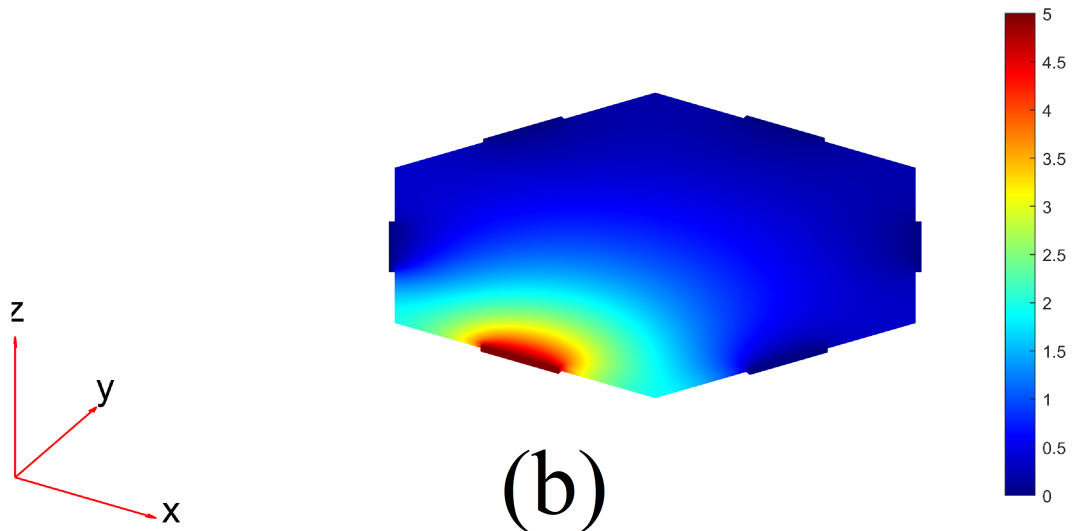
MATLAB[®] DC conduction solution [75] and QuickField[™] DC conduction analysis [76] are the two main tools surveyed in this thesis as part of the exploration of methodologies and relevant software. Using these design environments, a custom-shaped electrically conductive medium can be initially modelled as a 3D model using a separate free and open-source 3D computer-aided design (CAD) modeller such as FreeCAD [77]. Examples of DC conduction analysis on MATLAB[®] are shown in Figures 3.4 and 3.5. This work may be of particular interest for future researchers wishing to explore ways to increase the internal current capacity of the tiles or balls as part of a TCA cascaded power grid. Options include a simple full-area PCB layer (power-plane) dedicated to each of the power rails (positive or ground) to increase current capacity and lower cascaded resistances, or metal components of thicker metal form that could act as both a heat-sink and a power plane at the same time.

Having modelled a custom-shaped medium and taken a DC conduction analysis, the obtained electrical-current data can then be used for building a simplified equivalent circuit of resistor-meshed network. This circuit can be easily added into the hierarchy



(a)

Electric Potential



(b)

Figure 3.4: Example DC conduction analysis (Electrical Potential) on a hexagonal-shaped conductive medium. (a) shows a frame view with face numbers. (b) gives an electrical potential distribution on the object. In this particular case, an electric potential of 5V is applied on a single rectangular edge, which can be obviously seen in the red area. All the other edges are applied with an electric potential of 0V.

of SPICE sub-circuits for voltage/current analyses in a TCA system. A possible design flow can be illustrated in Figure 3.6.

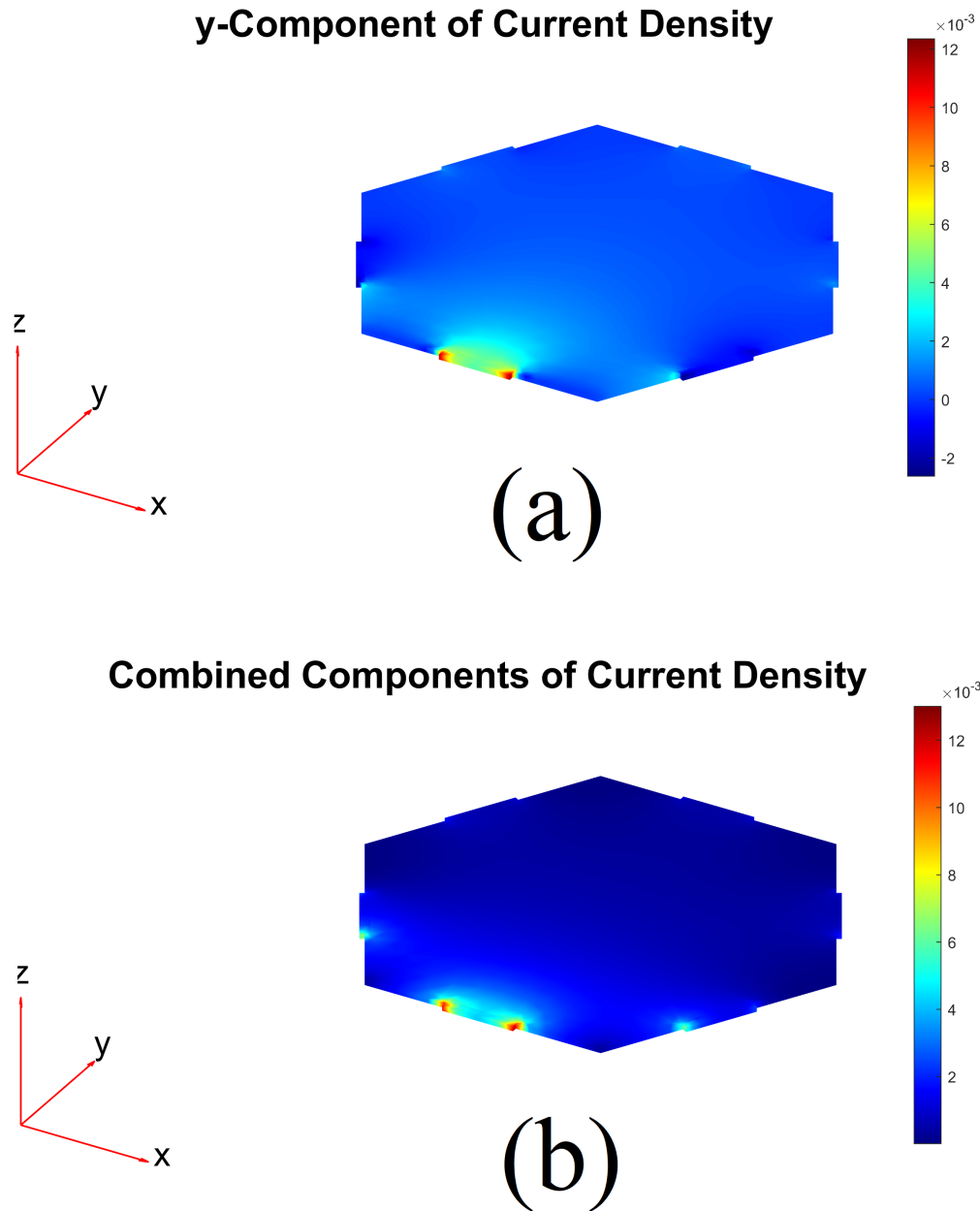


Figure 3.5: Example DC conduction analysis (Current Density) on a hexagonal-shaped conductive medium. (a) shows only the y-component of the current density, and (b) combines all the x, y, and z components, of the hexagonal medium. As this medium shape is a thin 3D object, it can be roughly considered that most of the currents flowing in this object are only in x and y axes.

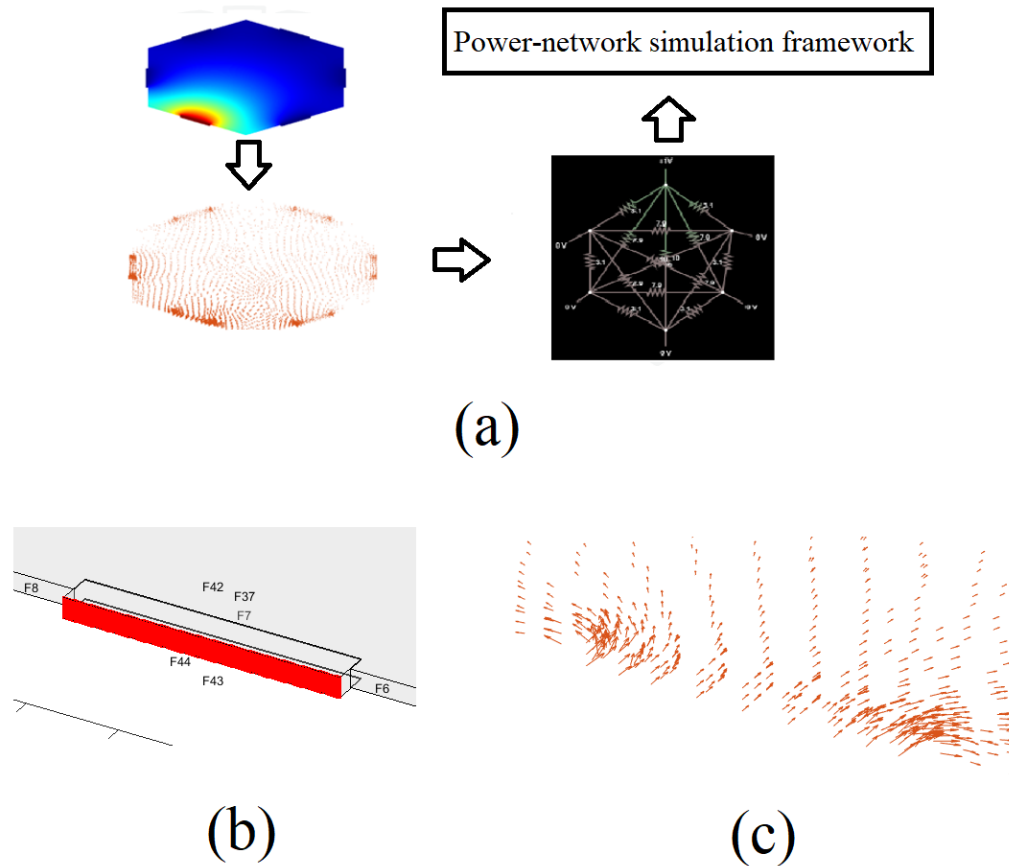


Figure 3.6: Experimental custom electrically conductive medium shape modelling. a possible design flow starts from modelling a custom shape, followed by (a), electrical potential/DC conduction analysis, and finally building an equivalent lumped-resistor network. (b) illustrates an edge of the object showing an area for current flows. (c) shows a vector-field plot of current flows in the object. This experimental design can be part in the future simulation framework.

3.4 Prototype Models

Traditionally, the power-delivery topologies found in rack-mount systems utilise a busbar system [78] or some form of power network via the back-plane, distributed to each computing board in a rack/cabinet. On the other hand, the power network in TCA is rather a grid-like topology. Concerning power-grid models, the appearance of the RLC network model in [79] is a good example of electrical network that has some similarity to the power network in a TCA system. However, the model was not intentionally investigated at the inter-board level focused in this thesis.

At a higher level, [53] surveyed several energy and power models in HPC systems, categorised by the components of systems. It is found that the models are for either PEs, interconnects, or at the system level, rather than the model of the power delivery itself. Due to no identifiable existing power-network modelling and simulation tools being found which fit the expectations of the unconventional TCA power grid, custom circuit models and simulation tools were required to be built for the constraint evaluations. These models are also validated against the hardware prototypes. Each of the models will be discussed in the following subsections.

3.4.1 Tile Modelling

The top-level module of the tile model consists of two sub-modules, 1) inter-tile power medium resistance-model, which represents connector-pin resistance in the present prototypes, and 2) the board model, representing the board power consumption of processing element(s), and all the other components regulated by the voltage regulator, including the power loss in the regulator itself. The separation of the sub-modules allows for modular flexibility. For instance, changing the characteristics of inter-tile coupling resistances does not affect the board model itself. A conceptualised model of the tile simulation model is shown in Figures 3.7 and 3.8. The conceptualised tile model represents the 'circuit-based board-resistance adjuster' described in Subsection 4.7.1, and the software implementation, 'post-processing based', is also discussed.

In this thesis, the current/voltage measurement modules, and the equivalent board-resistance adjuster are proposed with two different alternatives. These elements will also be discussed in Subsection 4.7.1. For simulation purposes, they can be considered auxiliary simulation components during board-resistance adjusting simulations. However, in a real hardware tile, the concept of these elements can be applied to build a power monitoring unit to control the power consumption of the regulated load.

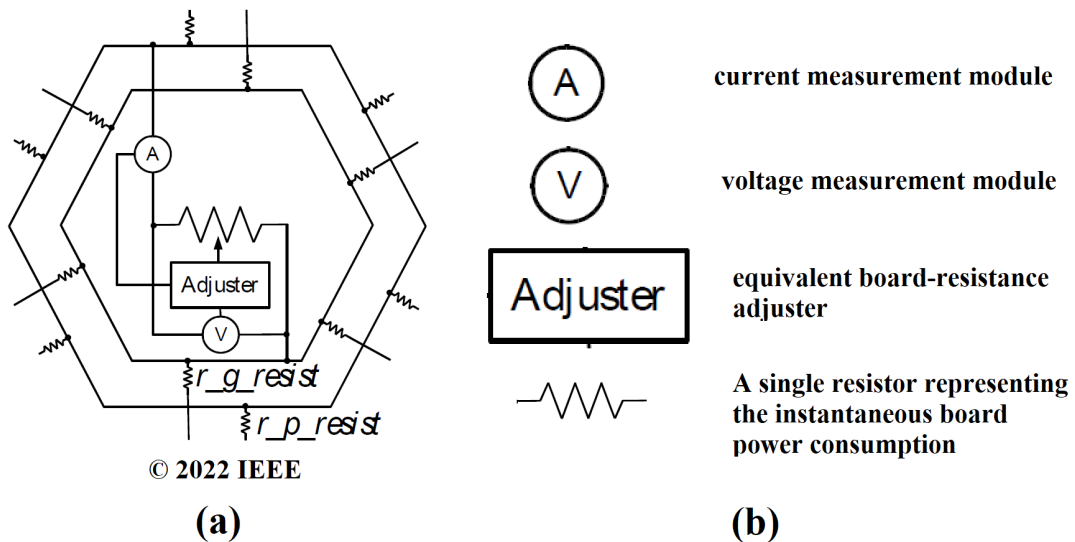


Figure 3.7: (a) shows conceptualised tile model. (cropped from Figure 4 in [42]). The resistors named r_p_resist and r_g_resist represent positive and ground rails of an edge power-connection. This inter-tile electrical medium resistance model can also be found in Figure 3.9. In (b), a legend describes the rest of the simulation components.

3.4.2 Connector-pin Resistance Modelling

Modelling the connector-pin resistances is one of the most important modelling-related tasks required in this thesis. The tiny resistance value of a fraction of 1Ω seems to be trivial at first glance. However, the total cascaded resistances of connector pins all over the power network should not be underestimated for large-scale systems. In Figure 3.9, the total resistance per mated pin-pair is a combination of 1) the internal resistance of a single pin, plus 2) the contact resistance phenomenon occurring in between.

Aside from the internal resistance of a pin, the additional surface contact-resistance

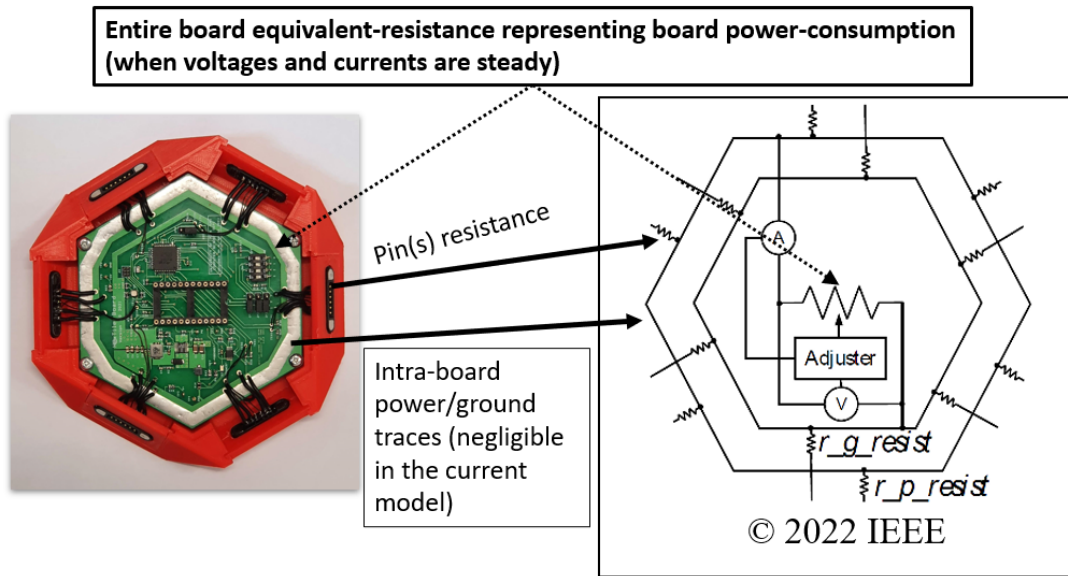


Figure 3.8: Comparison of a tile prototype with the conceptualised tile model (cropped from Figure 4 in [42]). Although the voltage and current measurement modules are added for simulation purposes, these modules can also be optionally implemented in a physical tiled computing unit for power management purposes.

does not only inherently affect the electrical quantities, voltages and currents, in the power network. It can also cause problematic system-level issues if not well-assembled to be appropriately mated. This contact resistance can be obtained from 1) the supplier's datasheet, or 2) manual measurements with adequate high-precision equipment. To design a model representing these two types of resistance, a single lumped resistor, named r_{p_resist} , models either, a single tile-edge power (positive) pin, or a collective parallel pins on the same connector. For the ground pins, they are modelled in the same way with the resistor named r_{g_resist} .

Additional parallel pins do not only have a beneficial effect on a higher current capacity, but also the reduction of resistance due to resistances in parallel. With the simulation framework created in this thesis, simulations can estimate whether a single or multiple parallel pin arrangement is required for all the currents flowing through each connector. These currents flowing through connectors will thus be simulated for connector-pin constraint evaluations. In this thesis, to simplify the model for large-scale simulations, all of connector resistances are assumed to be uniform and constant. However, an automated SPICE file can be later manually edited for resistance varia-

tions, for instance, the effects of pin failure situations or variations in manufacturing. Given a connector-pin arrangement at the tile edges, the next subsection will elaborate how the power consumption of the hexagonal board is modelled.

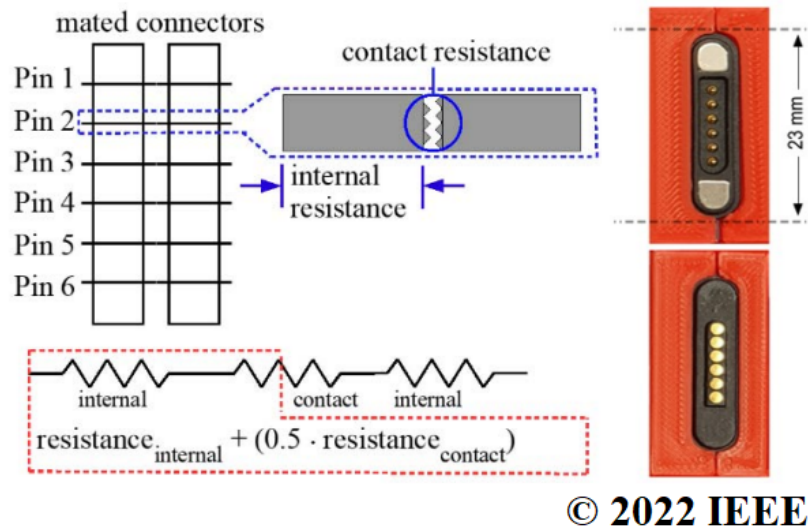


Figure 3.9: Connector-pin resistance model. The resistor symbol named 'contact' partially framed with the red dashed line does not exist in the actual model used in a SPICE simulation file, but added in this figure for illustration purposes. The contact resistance occurring between a mated pin-pair can be split into halves and equally be added to the 'internal' resistances on both sides for simulation purposes. A photograph of the red-edge connector detachable housing taken from the prototypes. (reprinted from Figure 3 in [42])

3.4.3 Board Modelling

At an early stage of the research, some *linear* voltage regulator models were investigated as options for converting the external supply voltage down to a desired on-board voltage level for powering tiles. However, due to the drawback of power inefficiency, this choice not only wastes a large amount of energy, especially when a TCA is supplied with a much higher external voltage, but also could cause problematic heat issues. With careful consideration, therefore, it is decided that *switching* voltage regulator models are a preferred option. In practice, a switching regulator obviously offers the advantage of high efficiency, however, from the simulation perspective, it raises some difficulties from its circuit complexity. One of the prominent drawbacks is tremendously long simulation times. Moreover, with the large amount of TCA tiles

Table 3.1: Comparison of approximate simulation times and file sizes between the full prototype-board model and the simplified board model, for the 3x3x3-ball validation case shown in Figure 3.12.

Model	Simulation time	Simulation file size
Full prototype	21.68 hours	38.56 gigabytes
Simplified	72.23 seconds	74.45 megabytes

for simulation, it could cause a huge amount of memory required and also SPICE convergence issues.

To mitigate the simulation complexity issue of switching regulators, some existing techniques, for example, [66], provide a number of average-behaviour models, whilst [67] and [68] investigated the automation of the modelling processes of switching regulators. After thoroughly investigating these existing techniques, large portions of detailed components in the models and processes were found to be unnecessary for the purpose of the scalability evaluations in this thesis.

As a result of that analysis, a higher-level simplified model, is proposed as an alternative option for large-scale TCA simulations. The advantages of the proposed simplified model are as follows:

Simplification:

The technique employed to simplify the board model was *curve fitting*, which was found to be adequate to evaluate the relationship of board input voltage and current. This is due to that the focus is how the whole board-level power consumption impacts upon the power-distribution grid. The curve-fitted model also significantly reduces simulation times and simulation result file sizes. A comparison of a test case is shown in Table 3.1. The voltage regulator $LT^{\circledR}3976^*$ [65] is employed in the tile prototypes. This switching regulator, along with all the components in the tile, are converted into a simplified board model to evaluate the whole TCA system when all voltages and currents become steady under the assumption of constant regulated loads.

To elaborate how curve fitting is applied to derive a simplified model, in Figure 3.7, the modelled resistor at the centre pointed by the arrow represents the whole board

* $LT^{\circledR}3976$ is a power-management product of Analog devices, Inc.

resistance, indicating an instantaneous power consumption of the board. The adjuster, as a *virtual element** for simulation purposes, periodically samples the input voltage, vin_s , and current, I_{board_s} . The adjuster alters the board-resistance value if a sampled board input-current is found to be *out of range* with the given expected input current, I_{board_e} , for the input voltage being supplied, as formulated in Equation 3.1. This equation is automatically generated by MATLAB® Curve Fitter tool [80] by analysing given multiple pairs of steady-state board input-voltage and average input-current values extracted from a simulation of the complex board model shown in Figure 3.10 given a regulated power-load. I_{diff_thres} , *input-current difference threshold*, is the parameter specifying the acceptable difference between the curve-fitted profile and the sampled input current during a simulation.

$$I_{board_e} = p_1 vin_s^3 + p_2 vin_s^2 + p_3 vin_s + p_4 \quad (3.1)$$

where:

vin_s = Sampled board input-voltage during a simulation

$p_{1..4}$ = Coefficients of the curve-fitting equation

I_{board_e} = Expected input current (equation auto-generated by MATLAB®)

Now consider I_{diff} , as in Equation 3.2, which representing the difference between the sampled and the expected input-currents based upon some changing adjuster value. Once I_{diff} converges into the interval of $(-I_{diff_thres}, +I_{diff_thres})$, the board resistance is maintained and then not further altered by the adjuster. There are also two additional parameters. tr_init initialises the initial resistance for a period of time. Afterwards, $Rstep$, controls the distance of board-resistance alteration in each clock cycle. A flowchart describing the overall mechanism of this implementation of the adjuster is shown in Figure 3.11.

* As the adjuster is a virtual element, technically, it is rather considered part of the simulation framework. The detailed circuit diagram of the board model implemented with 'circuit-based adjuster' can be found in Figure 4.8. However, it is also discussed in this chapter to provide the overview of how board resistance is adjusted.

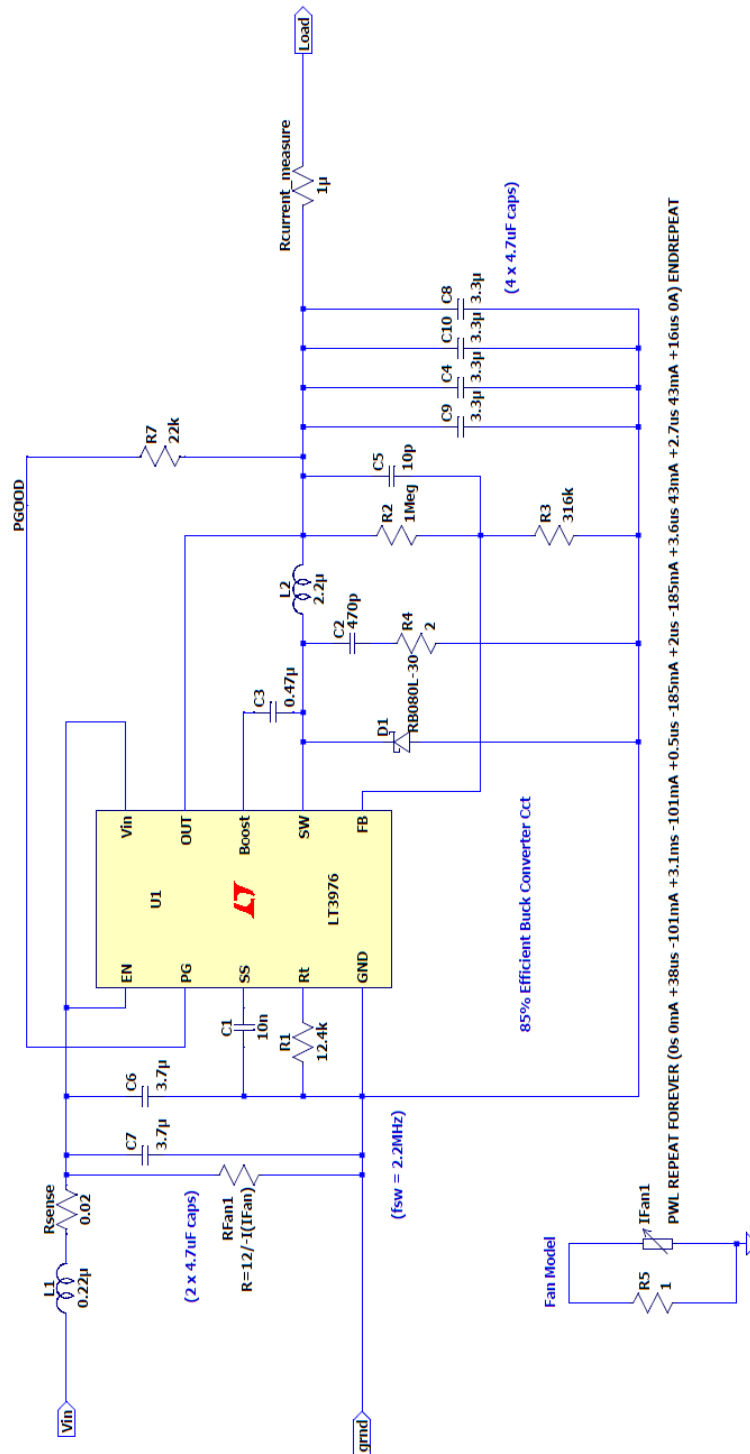


Figure 3.10: Board model with the switching regulator model.^a

^a The original version of the schematic designed by Anthony Moulds contains some additional devices, e.g., MOSFETs, to vary the regulated load in the prototypes. In this schematic, the load-controlling circuit area has been removed, and the load itself is moved to a higher-level sub-circuit for modular design.

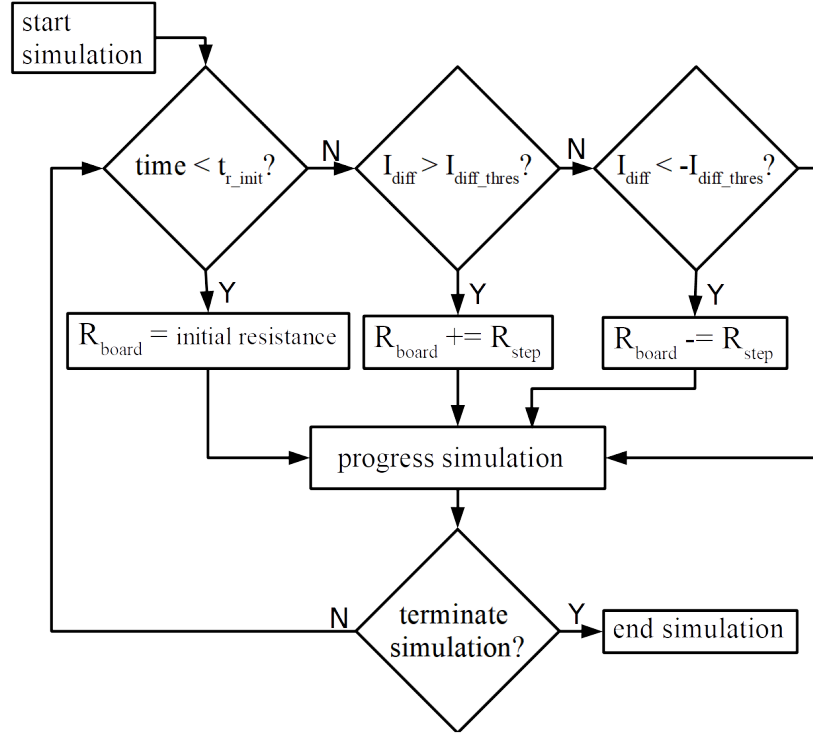


Figure 3.11: Overall mechanism of the circuit-based board resistance adjuster during a SPICE simulation.

$$I_{diff} = I_{board_s} - I_{board_e} \quad (3.2)$$

where:

I_{board_s} = Sampled board input-current during a simulation

I_{diff} = Difference between I_{board_s} and I_{board_e}

After all of the board resistance values are in a steady state, this indicates that all the board-model instances in a TCA are mimicking the averaged board input voltages and currents of the switching-model simulations when all the switching activities are steady. Having settled into this *black box* input voltage and current profile, the simulation can be terminated, and board input voltages and currents, and connector-pin currents, can be read out for constraint evaluations.

Regarding the Equation 3.1 aforementioned, the equation is an early form of curve fitting used in this thesis. Afterwards, an improved fitting technique is also proposed

Table 3.2: The differences between the two alternative fitting techniques for the board modelling proposed in this thesis.

Curve Fitting technique	Input	Output	Rload-value support
Equation 3.1 as published in [42]	Board input-voltage	Board input-current	Discrete: A fitting equation for an Rload-value existing in a board-model SPICE file.
Equation 3.3	Rload	Board input-current	Continuous: Multiple fitting-equations in a single board-model SPICE file. Each equation is for a single board input-voltage.

in this thesis and later used to run *post-processing based* power-distribution grid simulations on ngspice as an alternative choice for open-source SPICE simulator. This improved curve-fitting technique for board modelling is used to produce scalability simulation results in Chapter 5. The equation form of the improved curve-fitting technique can be seen in Equation 3.3.

$$I_{board_e} = p_1 Rload^{p_2} + p_3 \quad (3.3)$$

The differences between the early and the improved fitting techniques are described in Table 3.2. It can be seen that the input variable of the improved curve-fitting equation only takes *5V-regulated load resistance* (Rload) as an input. This is due to that each of the multiple fitting-equations existing in a single board-model SPICE file is generated for a single board input-voltage. In practice, the board input-voltage can also be an arbitrary value during simulations and in hardware operation, thus an interpolation process is required for a board input-voltage that resides in between a couple of consecutive-stepped fitting equations.

Accuracy:

The simplified approach was found to still maintain high accuracy after validating a ball-array simulation with the simplified model in comparison to the original more complex switching simulation model of the regulator, as can be seen in Figure 3.10. The model validation data can be found later in Figure 3.12, as discussed in Section 3.5. Regarding the I_{diff_thres} , *input-current difference threshold* mentioned earlier,

this parameter also consequently results in the accuracy of the simulation results, ranging in the interval of $(-I_{diff_thres}, +I_{diff_thres})$. As published in [42], this parameterisable I_{diff_thres} is set to 0.01A during the adjusting process, and regarding the curve-fitting model in Equation 3.1, accuracy is determined by selecting the optimal number of polynomials used to fit the curve. Third-degree polynomial fitting has been found to be adequate for modelling the board input voltage and current profiles. The maximum values of the curve-fitting error thresholds of the sum of squares due to error (SSE) and root mean squared error (RMSE) [81] for both the curve fitting techniques were found to be approximately at 0.01.

Applicability:

As mentioned earlier, the modelling technique proposed focuses upon how the whole board-level power consumption impacts on the power-distribution grid under the assumption of constant regulated loads. Thus, the underlying implementation of intra-board level components can be encapsulated as a *black-box*, resulting in mitigating the difficulties of unnecessary intra-board complex modelling tasks. In practice, the power consumption at this board level can vary due to dynamic computational loads. However, the *worst-case* power consumption is one of the applicable power scenarios, as it also guarantees the upper-bound of the whole board-level power consumption.

3.4.4 Voltage Regulator Modelling

As the voltage regulator is already an embedded part of the board model, it is unnecessary to separately model it. However, in future work if multiple voltage regulators or power circuits reside in a single board for some advantageous reasons such as multiple voltage requirements, then each of the regulators may also be simplified as a high-level abstract model for voltage and current characteristics.

3.4.5 Regulated Load Modelling

The regulated load is also another element in the board model, and therefore not needing to be treated as a separate modular model. For the completeness of component description, it will be briefly discussed in this subsection. In both circuit and post-processing based versions of the adjuster, the regulated load, R_{load} , can be seen as a modelled single lumped-resistor. However, instead of being a SPICE-simulation device of a resistor, the R_{load} value is used as part of performing the curve-fitting profiling. In the same way as voltage regulator modelling, if future board-model implementations wish to contain multiple loads, e.g., different PE types, each of the loads may also be separately modelled. It is worth emphasising again that a key concern of SPICE simulation is that the greater number of SPICE elements, the higher circuit-complexity for simulation and the more limited the opportunity to explore large scale systems simulations (due to infeasible compute time and resource requirements). This is also an important reason why the board model is designed to contain only single lumped-resistor, *board resistance*, for large-scale simulations.

3.4.6 Ball-array Modelling

In the previous subsections, all of the models were discussed in a *top-down* approach. This is due to the fact that most of the modelling tasks are actually at the tile level. On the other hand, as a *bottom-up* approach, the *TCA ball* could be the starting point as a possible unit for constructing a ball array, therefore modelling at the ball level is also of interest. Most of the difficulties of modelling noted in this thesis are from the tile level down to the regulated load, however, forming a TCA ball-array also involves some considerations.

Modelling of an entire TCA system is analogous to a real hardware construction. Eight tiles are coupled to be a ball, then balls are simply connected to construct a TCA ball-array. From the power-network model's point of view, at this point, a complex network of series-parallel resistors is constructed. Afterwards, external positive and negative terminals of power sources are connected, the entire resistor network is ready for

simulation. It is also possible that a TCA system is not fully filled with consecutive balls in each of the three dimensions. For instance, balls with the coordinates of odd numbers might be deliberately absent in a topology in order to allow more free volume for cooling, or to create specialised non-cubic topologies.

For a partially-filled array, an example is constructing a conventional 3D-array volume with some balls absent from selected coordinates, perhaps for enhanced cooling, or to facilitate internal bypass cabling to build arbitrary topologies to suit some specific purposes. However, non-regular topologies need custom routing algorithms, and if not well-designed, can lead to *deadlock* in packet switching. Data communication and routing algorithms are not the main focus in thesis. However, some preliminary topological investigation will be provided in Chapter 5.

As mentioned previously, the topology focused in the current stage of the research presented is also a simple 3D-mesh, aka, a TCA system of a fully filled 3D cubic ball array. As the automated process of system construction is also a separate code function, future researchers or developers only need some modification effort at the level of inter-ball coupling generation in order to generate any desired topology and then run simulations.

3.4.7 Power Source Modelling

Theoretically, power supplies connected to a TCA system can be a single voltage source or multiple ones. For a small system size, a single power supply unit could be sufficient, whilst larger systems may need multiple power supply units to distribute currents to the system at various external connections. This thesis focuses on the constraints of voltage drops and connector-pin currents. Thus, for simplicity and relaxing the complexity of the simulation models, the external power source is modelled as a single voltage source. In future work, power-sourcing topics are also an important area. Multiple power sources, for instance, paralleled switching regulators are one of the most promising choices.

3.5 Validation Work and Results

There are two main validation studies presented in this thesis:

- 1) Validating the full simulation model with the *switching* model of the voltage regulator *LT[®]3976* as shown in Figure 3.10, against real prototypes
- 2) Validating a *simplified* simulation model as shown in the internal hexagonal area in Figure 3.7(a), against the full simulation model

Validation study 1) allows the full system modelling capability to be set against the real hardware prototypes so that true accuracy of results can be measured, and also the correctness of the model confirmed, whilst study 2) allows the modelling techniques for simplified (faster) simulation to be demonstrated to be within acceptable margins of error.

3.5.1 Study 1) Hardware Versus System Modelling

In Tables 3.3 and 3.4, two switching simulation cases of a single-tile and an eight-tile ball are performed and validated against the hardware prototypes. In the tables, the

typical errors were found to be in the approximate range of 1-2%, except the first cases of no load, which produce a degree of higher errors up to approximately 13.3%. This is due to the lower range of less than 100mA, which could be caused by several sensitive factors, e.g., the precision of the laboratory measuring equipment, or the model itself.

Voltage effects on the 12V power source in the measurements are shown in Table 3.5. The measurement results are as expected: both 2D and 3D configurations provide a better voltage stability compared to a 1D arrangement. As the increased parallelism of the power-network can also reduce the overall power grid resistance, this also provides a better current distribution and reduces the voltage drop experienced between points in the power network. This is important since the way voltage-drop scales as a function of the size of the grid will be a key factor in the scalability of such a system. More sophisticated investigations for these two effects are expected to be carried out in the future.

Table 3.3: Prototype/Model: Single tile, Single connector. (© 2022 IEEE, regenerated from Table II(a) in [42])

	Min (base) ~ 0W	Low +2.5W	Med +5.0W	High +10.0W	Max +17.5W
$I_P \pm 5\text{mA}^1$	60 mA	310 mA	540 mA	1000 mA	1760 mA
I_M^2	62.29 mA	310.82 mA	539.93 mA	1012.81 mA	1753.82 mA
Error ³ (ave) (min, max)	4.5% 13.3%, -4.2%	0.3% 1.9%, -1.3%	0.0% 0.9%, -0.9%	1.3% 1.8%, 0.8%	-0.4% -0.1%, -0.6%

¹ Each value in this row is a measured value, I_P , meaning an input current flowing into the tile being measured in the experiment. These electrical currents were measured using a device with a display with three digits after the decimal point, resulting in a rounded number with an error of $\pm 5\text{mA}$.

² I_M is a tile input-current extracted from a simulation result with the complex board model.

³ Equation 3.4 is also used to calculate *min* and *max* of the values in this row, based upon the known range of input currents observed under test (with the $\pm 5\text{mA}$ range). As the models were developed to predict the scalability of the prototypes, the *Experimental* (observed) value in this particular case is I_M , whilst the *Theoretical* (expected) values are split into two values of $I_P - 5\text{mA}$ (*min*) and $I_P + 5\text{mA}$ (*max*), respectively. *ave* is simply an average of both the *min max*. For example, at $60\text{mA} \pm 5\text{mA}$, the measured 62.29mA compares to the input range of 55mA to 65mA , giving the $+13.3\%$ and -4.2% errors respectively, this averages to 4.5% nominal error.

3.5.2 Study 2) Simplified Versus Switching Models

For the simplified-switching validation, the average percent-error from the simulation results in Figure 3.12 of a $3 \times 3 \times 3$ ball-array were found to be less than 1%. In this particular validation, load resistances in tiles were set to 1Ω , mimicking approximately 25W per tile regulated at 5V. External power connections are fully-connected with a 12V voltage source.

In Figure 3.12, the initial external voltage starts at 0V, then increased up to 12V. This is to help the SPICE simulator with achieving a DC operating point more easily. Inner tile-level input voltages are impacted by the connector-pin resistances in the power network, thus, receiving voltages under 12V. At this small array-size, voltage drops are not obviously visible as the number of cubic layers is small. However, in large-scale systems such as a $10 \times 10 \times 10$ ball-array, the meshed network of pin resistances can cause dramatic voltage drop issues. With the regulator input-voltage under the minimum specified, the load voltage regulation can be unstable. At the end period of simulation, both the switching and simplified models have converged into their steady states.

All of the board input voltages and currents, and connector-pin currents, in the simplified-model simulation are expected to be equal to the averages of those in the switching one. As can be observed in Table 3.6, the differences between the two models at steady state are very small. The accuracy of these voltages and currents depends on both a) the quality of the curve fitting method employed, and b) the simulation

Table 3.4: Prototype/Model: 8-tile ball, 2 co-located power connectors. (© 2022 IEEE, regenerated from Table II(b) in [42])

	Min (base) ~ 0W	Low +2.5W	Med +5.0W	High +10.0W	Max +17.5W
$I_P \pm 5\text{mA}^1$	530 mA	2550 mA	4370 mA	8070 mA	14010 mA
I_M^2	501.67 mA	2493.57 mA	4328.48 mA	8121.95 mA	14079.9 mA
Error ³ (ave)	-5.3%	-2.2%	-0.9%	0.6%	0.5%
(min, max)	-4.4%, -6.2%	-2.0%, -2.4%	-0.8%, -1.1%	0.7%, 0.6%	0.5%, 0.5%

^{1,2,3} See the corresponding notes in Table 3.3

Table 3.5: Prototype: grid stability (worst case voltage drop, 10W load, 12V supply). (© 2022 IEEE, regenerated from Table II(c) in [42]. Note that there are an average of one power connector for every four tiles in all three cases, to ensure uniformity and also to ensure that connector pins per connector are not overloaded).

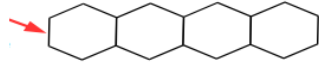
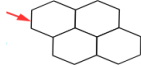
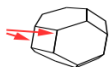
Tiling	Configuration	Prototype
	1D tiling: 4 tiles, 1 connector	1.25%, 150mV
	2D tiling: 4 tiles, 1 connector	0.33%, 40mV
	3D tiling: 8 tiles, 2 connectors	0.17%, 20mV

Table 3.6: Simple vs complex simulation test case. Accuracy of the hypothetical test case of 3x3x3-ball TCA system shown in Figure 3.12 for non-negligible current flows of the validation of simplified versus complex manufacturer switching models.

Quantities	Percent error: simple vs complex	
	minimum	maximum
Board input voltages	0.0007%	0.0071%
Board input currents	-0.5989%	-0.1992%
Connector-pin currents	-0.6408%	0.3154%

parameter I_{diff_thres} . Table 3.6 shows the accuracy of the test case shown in Figure 3.12. The accuracy values in the table are reported as percent errors calculated using Equation 3.4 [82], where, in this particular case, *Experimental* is a simulation result with the simplified board model, whilst *Theoretical* is that of the full switching board model. All the three types of quantities in the table extracted from the simulations to calculate the percent errors are positive values, and both the positive and negative percent errors are reported to show the directions of distance from the *Theoretical* values, thus the absolute operation ($| |$) is absent in the equation. Example lines of SPICE code and parameter values as a part of the circuit-based implementation of the adjuster are shown in Table 3.7.

$$\% Error = \frac{Experimental - Theoretical}{Theoretical} \times 100 \quad (3.4)$$

Table 3.7: LTspice[®] example code and parameters. (© 2022 IEEE, regenerated from Table III in [42])

<p>Example parameter values: Initial resistance period: 6 Ohms, held for 21 us, then 0.005 Ohm steps</p>
<p>Example LTspice[®] code with the above parameter values</p> <pre> b_i_board i_board v = i(r_board_resistance) b_i_diff i_diff 0 v = v (i_board_s) - ((-0.006025)*(v(vin_s)**3) + + 0.2087*(v(vin_s)**2) - 2.623*v(vin_s) + 14.39) b_r_board r_board 0 v = if(time<21us, 6 ,if(v(i_diff) > 0.01, v(r_board_s) + 0.005, if(v(i_diff) <-0.01, v(r_board_s)-0.005, v(r_board_s)))) </pre>

NOTE: 21 us and 0.005 Ohm, are the hard-coded values for tr_{init} , and $Rstep$, respectively.

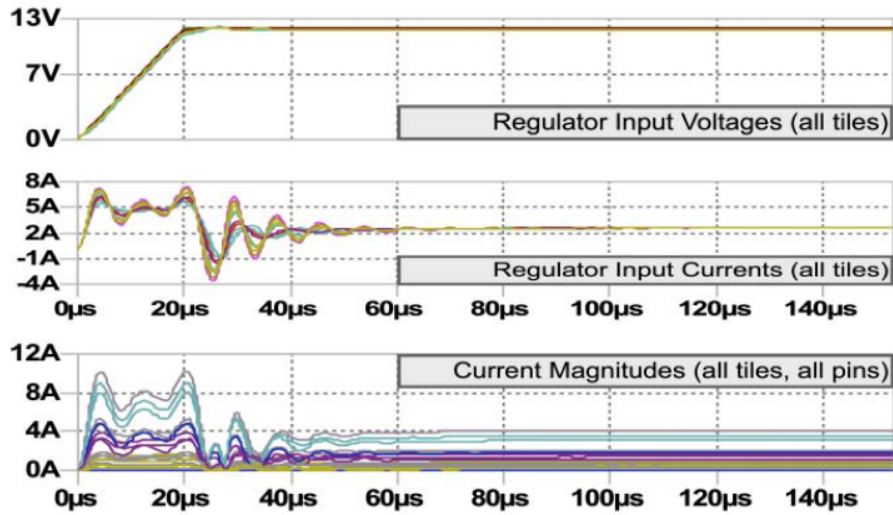
Having discussed the hardware prototypes, models, and validations, the sub-research questions addressed in Chapter 1 are restated and discussed with the relevant sections as follows:

► *What are the necessary design choices for constructing tile-able modules?*

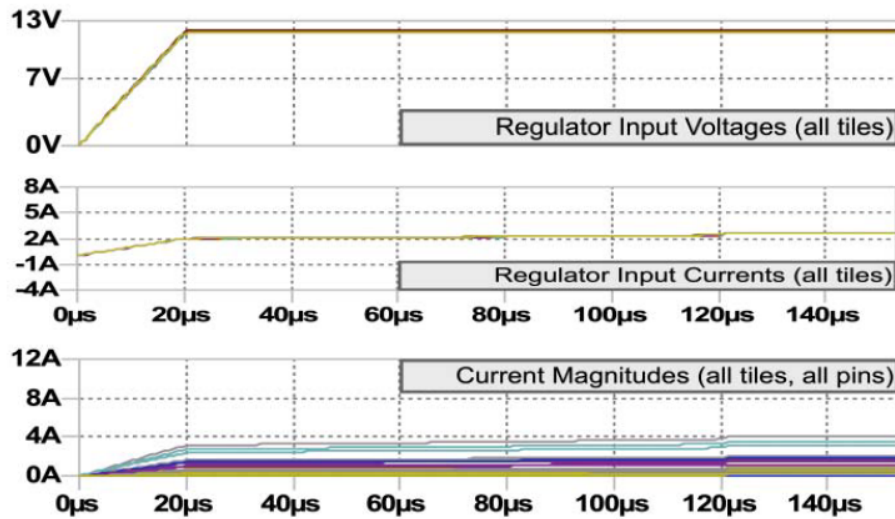
Regarding the intra-tile level, it can be seen that both the prototypes built in this thesis with off-the-shelf connectors in Section 3.2, and arbitrary electrically-conductive media in Section 3.3, are discussed. In Subsection 3.4.2, it is also a consideration specifically for how many connector pins are required for tolerating the currents flowing through themselves given a desired TCA size. Moreover, at the intra-board level implementation, the choices of linear and switching regulators are also discussed in Subsection 3.4.3. The regulators employed can dictate several factors, for instance, input/output voltage ranges, maximum output current and power.

► *What are the component characteristics of the power grid in a TCA array?*

For this question, the most relevant parts are the electrical conductive media employed, how they are structured, and also how the voltages and currents of the power grid itself are impacted by the tile-able modules. First, it can be seen in Figures 3.4 and 3.5 that designing a custom conductive shape has the complex characteristics of both the electrical potential and current distribution to be considered. Figure 3.1(a) also shows an intra-board hexagonal conductive shape, allowing high currents flowing between tiles as part of the whole power-grid network. The



(a) Simulation based upon LT3976 regulator model



(b) Simulation using simplified (faster) model

© 2022 IEEE

Figure 3.12: Simulation results used for validating the simplified board model (curve-fitted model), compared against the same board model using the full manufacturer's precise LT[®]3976 regulator SPICE representation, as shown in Figure 3.10. In this validation, a system of 3x3x3-ball was used for both simulations, as this is the smallest ball-array to contain at least an inner ball to reflect voltage drops. (a) shows the simulation result of the full prototype-board model, and (b) shows the simulation result of the simplified board model. (reprinted from Figure 5 in [42])

prototype measurements in Table 3.5 also show how different composition configurations affect the characteristics of voltage drops. Additionally, Figure 3.12 depicts the possible fluctuations of electrical currents when switching regulators are employed in the system.

3.6 Chapter Summary

This chapter reflects upon the hardware prototypes and the models of relevant components for TCA power-distribution grid scalability evaluations. The hardware prototypes have been successfully built at the scale of a single ball. This is due to the funding limitation. However, one of the essential purposes of modelling and the simulation framework is to predict the scalability of a TCA system of interest based upon a small scale, prior to building a physical large system.

Regarding the models, whilst an experimental hexagonal conductive-shape is also investigated for future work, the models based upon the hardware prototypes have been proposed and thoroughly discussed. The main research question and relevant success criteria will be restated as follows:

The main research question:

Is it feasible to build a large-scale power-grid network of Tiled Computing Array, whilst still scaling up the system computing performance?

Referring to the two sub-research questions discussed in the previous section, the main research question is now considered partially fulfilled at this point due to the models validated against the hardware prototypes ready to be part of scalability evaluations in Chapter 5. However, several of the other relevant research objectives have now been fully achieved as follows:

Objective 1 - Success Criteria:

The models need to be designed at abstraction levels adequate for reasonable simulation times and precision. Thus, the models are to be compared against real hardware for validation of accuracy.

This criteria set is found to be met with the following evidence:

- Models are proposed in well-structured and hierarchical levels as discussed in Section 3.4.

- ▶ The simulation times and file sizes of a 3x3x3-ball model validation have been reported in Table 3.1.
- ▶ The models have been developed for future convenience parameterisation as described in Section 3.4, which also means that the parameter values can be set for validating against the hardware prototypes built in this thesis.

Objective 2 - Success Criteria:

The accuracy of the models running on the simulation framework should be within acceptable error percentages to gain better understanding of real hardware and the variability of actual components in practice (for example, connector resistance, actual versus data-sheet and so-on). Importantly, a degree of error can be tolerated, but being able to quantify the error range allows the simulation models to be understood to offer realistic projections of performance within the same error range.

This criteria set is also found to be met with the following evidence:

- ▶ The models have been validated against real hardware prototypes, and the accuracy results have also been reported in Section 3.5.

This chapter is particularly related to the research objectives 1 and 4. The simulation models used here have previously been discussed in Chapter 3. Moving forward, this chapter focuses on the simulation framework, explaining the reasoning for some of design choices in the framework, difficulties, and various simulation-related issues. The simulation topics in this chapter mainly involve the power-network of TCA, whilst the modifications for preliminary interconnection network performance evaluation will be separately discussed in Chapter 5. The simulation framework will be thoroughly detailed in terms of high-level conceptions; however, some correspondingly important pseudo/source-code examples will also be given where beneficial for detailed understanding of the implementation. For the complete provision of reproducibility purposes, all of the important source code files are expected to be published in an online repository.

4.1 Relevant Research Objectives

Both of the research objectives 1 and 4 are relevant to this chapter. The research objective 6 as the purpose of tool documentation, however, is also an additional relevant objective. This objective is also included in this chapter for completeness as part of a good practice of long-term tool development. The objectives along with the sections in which they are achieved are given as follows:

4.1.1 Objective 1: Employing and Designing Models and Simulation Tools

Methodologies/Activities:

- ▶ **High-level programming:** As the large sets of evaluation in this thesis require automated processes for simulation, not only the designs of circuit models are required. MATLAB[®] is selected for creating simulation tools as it provides several useful data-manipulation functions, and also scientific and engineering toolboxes that can be employed in future developments. Several generators and other simulation tools can be found in Sections 4.5 and 4.7.
- ▶ **Geometry:** Trigonometry, as a subset of geometry will be used for calculating the lengths and angles in a hexagonal tile. This is to correctly construct 3D-model shapes, for example, trapezoids and hexagons appearing in tiles, balls, or the whole system. This will be used in visualisation capabilities as part of simulation tools. Every visualisation in this chapter that has hexagonal or trapezoidal shapes as part of generation requires the geometrical calculations described.
- ▶ **Visualisation:** Visualisation can be considered for 1) understanding the appearance when constructing tiles, as the smallest module, and also balls as combined modules for constructing a TCA system, and 2) verifying the correctness of the automation tools built in this thesis. Only preliminary visualisation capabilities such as voltage drops or pin currents at tiles are experimentally designed in this thesis. A large portion of the figures in this chapter are produced by the visualisation capabilities developed in this thesis. A prominent and meaningful visualisation can be seen in Figure 4.15.
- ▶ **SPICE simulation:** A SPICE simulator is required for evaluating the scalability and optimisations in this thesis. Two SPICE simulators are employed in the simulation framework, which can be seen in Figure 4.1, and will be discussed in Section 4.8.

Expectations and outcomes:

- ▶ **Automating scripts:** In addition to Sections 4.5 and 4.7, examples of several functions can be found in Figures 4.17 and 4.18.
- ▶ **Visualisation capabilities for tool verification and meaningful resulting representations:** Examples of 2D and 3D visualisation examples can be seen in

Figures 4.2, 4.4, and especially 4.15, which obviously helps verify the expected voltage-drop trend of a simulation result.

Success Criteria: The simulation framework, may include components of some existing tools, and also custom tools built as part of the PhD research presented. They should also correctly automate internal processes without significant user effort in terms of parameterising, running, and modification for future work. An open-source SPICE simulator is expected to be used in this thesis, as it can provide some flexibility in terms of tool modification for specific purposes. However, power regulator simulation models may be required to be simulated only in some specific simulator configurations. Therefore, both open-source and proprietary SPICE simulators will be together investigated to determine what is needed. Another important point is that it is a tremendously labour-intensive task to manually write large input files for simulations. Thus, automating scripts are to be built. This is considered one of the essential core parts in the entire simulation work-flow. Ideally, scripts should be also split into modular-design parts to support maintainability and future re-use.

4.1.2 Objective 4: Optimised Power Distribution

Methodologies/Activities:

- ▶ **SPICE simulation:** SPICE-simulation activities are also involved for optimised-power results. Several SPICE-related processes can be seen in Figure 4.18.
- ▶ **Parallel simulation:** The parallel simulation in the simulation framework shown in Figure 4.19 helps reduce the completion time of a simulation set.
- ▶ **Visualisation:** As nodes (tiles) are not allocated the same amount of power in the non-uniform power allocation, Figure 4.15 is an example clarifying how nodes are allocated under a proposed power scheme (*relative position*).
- ▶ **Genetic Algorithm (GA) problem formulations:** The power-allocation problem mapping can be shown in Figures 4.21 and 4.22.

- ▶ **Non-uniform power allocation:** The relative-position scheme discussed in Subsection 4.7.2.2, is proposed as a possible non-uniform power allocation scheme.

Expectations and outcomes:

- ▶ **An optimisation framework for non-uniform power allocation:** The GA-optimised simulation will be discussed in Subsection 4.7.4.2. Figure 4.24 also shows the non-uniform simulation workflow.

Success Criteria: The optimisation framework should demonstrate how uniform and non-uniform power allocation schemes can differently impact on the TCA scalability.

4.1.3 Objective 6: Simulation Framework Documentation

This additional objective is given in this subsection without linking to any other sections, as the process of documentation is considered structural and time-consuming for long-term simulation-framework development. During the time of the simulation framework development, the on-going processes are given as follows:

Methodologies/Activities:

- ▶ **Structurally documenting the simulation tools developed in this thesis:**

Expectations and outcomes:

- ▶ **Function descriptions (visualisations may be required for some complex processes):**
- ▶ **Example test cases with results:**
- ▶ **Usage warnings and cautions:**
- ▶ **Tool limitations:**
- ▶ **Guidelines for capability extensions and integration with other tools:**

Success Criteria: The documentation should be well structured and understandable with enough information to establish the basics of the tool-set.

Given the relevant research objectives, the rest of this chapter will discuss all the details relevant.

4.2 Overall Simulation Framework

The entire *simulation framework* in this thesis is comprised of several hierarchical parts as shown in Figure 4.1. Some of these parts are of modular design and also are capable of running as stand-alone functions for specific purposes by a user, who may want to focus on particular investigations. Although the whole simulation framework is expected to be fully automated, a few sub-processes still need to be manually performed by the user. Figure 4.1 shows the overall view of the power-distribution grid simulation framework, with the hierarchical boxes summarised as follows:

- ▶ **Ball:** In this box, the tool user needs to manually write several SPICE *sub-circuits*. From a bottom-up approach, it starts with the *board model*, which has already been discussed in Chapter 3. At the same level, there is another component model, *connector sub-circuit*, which represents the resistance values of both the power and ground pins of a connector. These two models are shown in a sub-box, which represents the internal sub-circuits of a *tile model*. At the tile-model level, there are six connector-sub-circuit instances, due to the existence of the six connectors in a hexagonal tile. Finally, a single-ball model is composed of eight instances of the tile model. All the sub-circuits described can take some parameters such as regulated load-resistance and connector-pin resistance. At the current stage of the research, ball is the smallest level allowed for automated generation of a complete TCA cubic-system. Given a ball-model SPICE file, it will be later used in semi-automated power-distribution grid simulations.
- ▶ **System generator:** This box can be split into two internal parts, First, *Ball-array generator* is responsible for generating the whole cubic-array. *single-ball*

template is a single SPICE line of code, containing a ball-instance name, its predefined sub-circuit ports, the name of the ball-model sub-circuit which has already discussed in the previous box, and also some textual placeholders for later manual modifications by the user. Each template-line will then be gradually added into SPICE lines of code during multiple-ball generation, together with renaming some inter-ball signals necessary for connecting them together. Second, the step of *External power/ground rails renaming* takes the whole lines of code generated for the whole ball-array to replace only the external (surface) power connections, i.e., power and ground signal names plane-by-plane to be effectively connected to an external voltage source. Finally, the output of this sub-process still contains textual placeholders, which will be systematically replaced with user-defined sub-circuit ports, name, and parameters. This task can be easily done by the tool-user using a text editor.

- ▶ **Manual SPICE-file editing:** In this box, it is the last step prior to perform power-distribution grid simulations. A tool user can flexibility change the textual placeholders as a whole if all the ball instances are from the same model. Moreover, if some specific balls are specialised, those ball-model names, and also other ball-specific parameters, can later be separately altered. The output of this sub-process is a ready SPICE file to be simulated.
- ▶ **Simulation modules:** This box combines several functions implemented to perform both uniform and non-uniform power-distribution grid simulations as parallel executions of a SPICE simulator. The visualisation capabilities are also included in this box.
- ▶ **SPICE Simulator:** A SPICE simulator is executed via an operating-system command. Thus, the TCA SPICE-simulation file is taken as an input of the SPICE simulator chosen. During power-grid simulations, after a single SPICE simulation is complete, the simulation results will be read by either uniform or non-uniform simulation functions for subsequent processes.

Having introduced several parts in the entire simulation framework, the following sections will elaborate important underlying details for the power-distribution grid simulation.

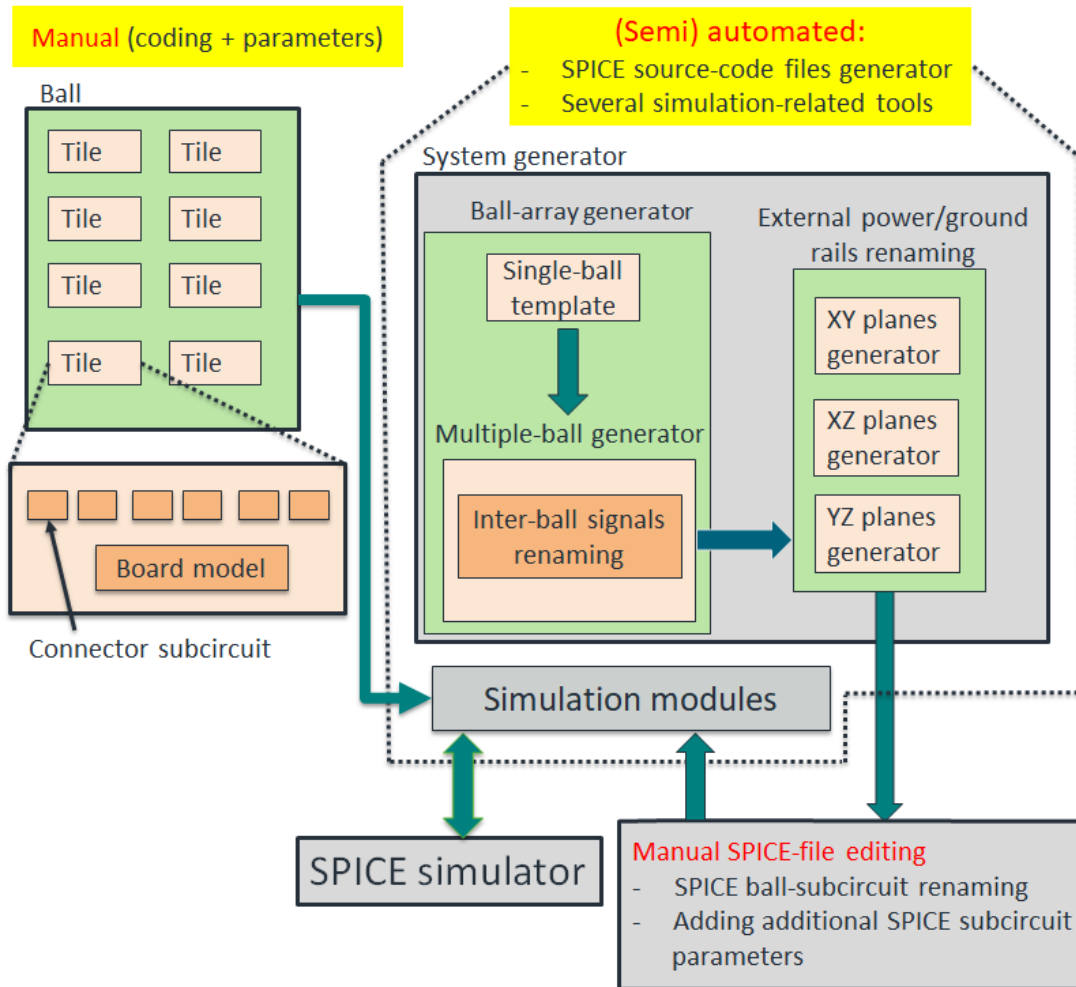


Figure 4.1: Overall simulation framework. Some details in the automated area are omitted and encapsulated for a concise view of the whole framework.

4.3 Tile Naming Convention

Prior to generating SPICE simulation files, in this section, the naming convention of tiles is discussed. This tile naming is one of the very important fundamental concepts for constructing the entire TCA power-network for simulations. It consists of a facet labelling convention, and an edge numbering convention. These conventions are related to the *Ball* box in Figure 4.1.

4.3.1 Facet Naming Convention

As shown in Figure 4.2, naming all of the eight tiles in a ball starts from the four upper tiles of a ball as shown in Figure 4.2(a), with H-Facets facing diagonally outward, as labelled $\{A,B,C,D\}$. The same process is performed for the four lower tiles of a ball as shown in Figure 4.2(b), with H-facets labelled $\{E,F,G,H\}$ completing all of the tile names in a single ball. Each of these eight A-to-H labels, will be used to name their corresponding tile-instances, which can be found in the *Ball* box in Figure 4.1.

4.3.2 Edge Numbering Convention

With a side view of a single ball, for a tile, the edge-numbers begin at the top-edge with the edge 0, moving on clockwise all of the way around the six tile edges, then stopping at the edge 5. As shown in Figure 4.2(a) with a top view, the numbering convention starts at the inside edge, whilst in Figure 4.2(b) as a bottom view, it starts at the outside edge. Regarding the pin-resistance model, p and n , each represents the outside terminal(s) of a single or a combination of parallel pins for positive (power) or negative (ground) rails on each trapezoidal facet. These outside terminals will then be connected to those of another ball. For the inside terminals, they are connected to the board model. The outside and inside terminals described are the terminals of `r_p_resist`, or `r_g_resist` connector-pin resistance models as shown in Figure 3.8. The edge numbering convention in this subsection will be internally used in each of the eight tile-instances in the *Ball* box in Figure 4.1. Moreover, the edge numbers are also used at the ball level, at which the eight tiles are instantiated to form a complete ball.

4.3.3 Further Detail

With facet labels and edge numbering, it is possible to reference any connection point on a tile, and thus in a composition of tiles, including edges that connect to each other.

As described in the previous subsection, eight tiles are connected together to form a single ball. In a SPICE file of the ball level, the user needs to substitute the signal names of the pins' outside-terminals mated to their neighbouring tiles with some suitable names to effectively make them connected together. Fortunately, this step at the ball level has already been completed and provided as a ball-template file in this thesis. For the next upper hierarchical-level, balls connected into arrays, this step is automated by the ball-array generator. For ball-array generation, only p-n signal names upon all of the 24 trapezoidal facets around the six holes, will be concatenated with edge numbers, tile names, and ball coordinates. These concatenated names are then visible at inter-ball level. A detailed example of signal renaming is given in the next section, Section 4.4, with examples of this scenario: see Figure 4.3 and Table 4.1 for instance.

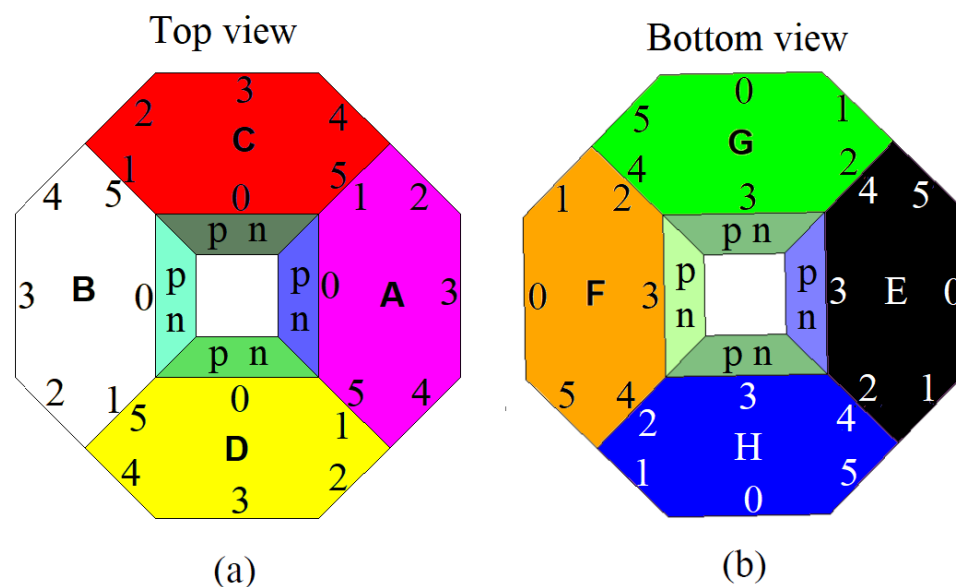


Figure 4.2: Visualisation of tile naming convention. (a) visualises the top view of a ball, whilst (b) is for the bottom view, where Facet 'E' is directly underneath Facet 'A' when a ball is viewed from above, and the same positions for the rest. The square at the centre is one of the six holes for cooling. p and n are the names for positive and negative (ground) rails on a trapezoidal facet.

4.4 Connector-pin Terminal-name Replacement Convention

Another important process during the generation of a whole system model is the replacement of SPICE node names representing the inter-ball power-rail terminals. This replacement convention will be performed in the *inter-ball signals renaming* box as shown in Figure 4.1. This process permits a new ball to be added into the system in terms of SPICE elements. When a ball is about to be added to an incomplete array, there are four alternatives for effectively connecting together two resistors modelling a mated connector-pin pair:

- ▶ 1) Combine these two resistor names together,
- ▶ 2) Replace those of the new ball with the existing ones,
- ▶ 3) Vice versa, replace the existing resistor names with those of the new ball,
- ▶ 4) Rename these mated terminals with a completely new name.

In this thesis, choice 2), replacing those of the new ball with the existing ones, is utilised. The rationale for this decision is that choice 1) makes the SPICE node name longer. It is however good practice to make node names as concise as possible. This is primarily for readability but also has some small simulation performance and memory requirements benefits, for example if a third-party SPICE simulator employed is sensitive to parsing long node names and memory management this would improve code portability.

Of the other options, choice 3) and choice 2) appear to be very similar. However, maintaining the previously generated names will, later, notify the user of the sequence of balls added into the system. Choice 4) is another alternative, but will make the names arising from this inter-ball connection inconsistent with other names all over the system. This *new-to-old name* (choice 2) replacement convention can be visually explained in Figure 4.3.

Having discussed all of the conventions and the difficulties involved, it is obviously seen that the manual preparation of multiple instances of the ball subcircuit as shown

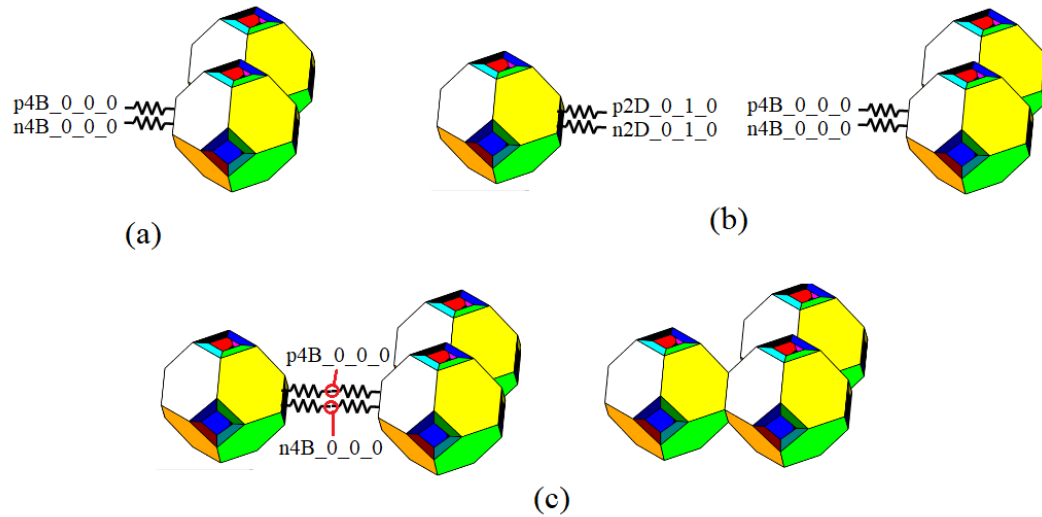


Figure 4.3: Example of connector-pin resistance terminal name change of a ball group. (a) shows only a couple of resistor names. In (b), a newly added ball is generated with its own resistor names at the edge to be coupled. In (c), The names of the new-ball resistors replaced with the ball to which it is coupled, finally modelling a new ball coupled to an existing system.

in the *ball* box in Figure 4.1 for the entire TCA system power-network model files is likely to be unfeasible for large-scale simulations. Moreover, at the penultimate step prior to invoking a SPICE simulation, an additional process for an external voltage supplied to the system is also required. Therefore, to make these tasks manageable, a system generator was developed, which will be discussed in the following section.

4.5 System Generator

The system generator combines the two sub-processes of *Ball-array generator* and *External power/ground rails renaming process* to be referred to as *System Generator* shown in Figure 4.1. The important parts of these two sub-processes will be discussed. All of the models, henceforth, are discussed here from the perspective of how they are manipulated to be ready for simulations, rather than detailing the mechanism of the models themselves, which have already been discussed in the previous chapter.

Although this aspect of the framework still requires the user to perform some processes manually, this is mainly by calling separate functions with input parameters,

rather than manually coding the SPICE files themselves. These two sub-processes are comprised of a number of hierarchical sub-functions generating several partial results of SPICE-code modification. Each of the two boxes as shown in Figure 4.1 will be described as follows.

4.5.1 Ball-array Generator

After the conventions of tile naming and pin terminal-name replacements have been agreed, and a single-ball SPICE file has been manually coded, the next step is to generate the whole system by connecting balls together one-by-one. Figure 4.4 shows a system gradually generated by first adding balls in the X direction, then Y, and then Z. This action is performed by the *Multiple-ball generator* process as shown in Figure 4.1. This generator gradually adds SPICE lines of code, which are instances of *Single-ball template*.

With this filling convention, all of the balls in the bottom plane of a 3D array are completed first, then successive layers grow upward towards the top plane, making up the entire system. In this 3-axis structural mapping, there is another positional referencing convention for ball-array generation. The directions of ball connection is not aligned with the actual X, Y, and Z axes for visualisation. This is due to the simplicity of creating the first ball as shown in Figure 4.4(a) with tile A and B as shown in Figure 4.2, to be parallel with the X axis.

As balls are added to an array during the automated array generation, there can be cases where a new ball is connected to more than one existing ball. These multiple neighbouring-coupling cases are shown in Figure 4.5. Only the first ball at the coordinates (0,0,0) is not connected to any existing ball as it is the first element generated. Whilst the rest of the ball connections will depend on where they are added in an array structure. For example, the balls generated where their Y and Z coordinates are zero, are connected to the previous ones in the X dimension only. The number of connections per ball for all of the other cases is also shown in the tabular legend in

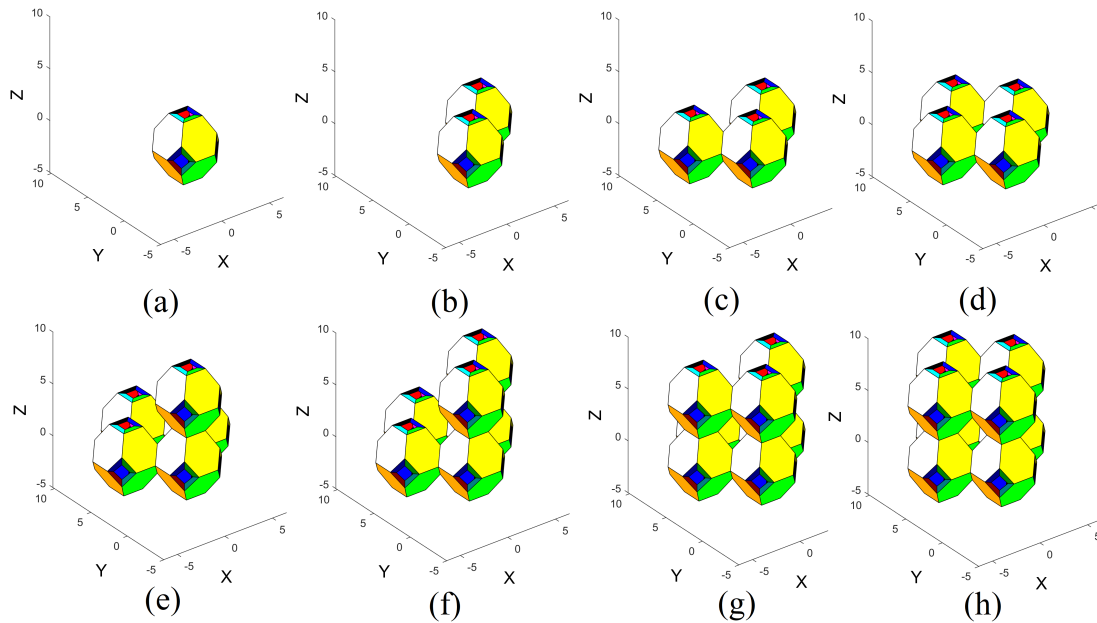


Figure 4.4: Step-by-step example of balls coupled in a SPICE-file of a system generation, starting from (a) towards a complete system in (h), respectively. When a ball is added, the SPICE node names of the connector-pin resistor model at the ball-edge connected to a previously generated ball in each dimension will be replaced.

Figure 4.5. To give a further explanation, Figures 4.3 and 4.4(c) illustrate an example case of the *dimensional-connection Y*, as represented by a black-rhombus shape shown in the tabular legend. This means that only the trapezoidal facet of the newly added ball in the Y dimension will be connected to an existing ball. The understanding of ball-connection cases simplifies the implementation of *Inter-ball signals renaming* sub-process as shown in Figure 4.1, which is the connector-pin resistance terminal replacement process for all of the six trapezoidal facets as shown in Figure 4.3.

As described earlier, a TCA system configured as a ball array is auto-generated by the framework. A single ball and a TCA-system of 2x2x1 balls, are given as examples of the resulting SPICE-code generation in Table 4.1. In the actual SPICE files, each of a single-ball instance contains only one line of code. The section of 'THE_REST_OF_YOUR_BALL_PORTS', 'YOUR_BALL_SUBCIRCUIT_NAME', and 'YOUR_BALL_PARAMETERS', gives the flexibility to the user to manually edit this part for variants of ball design in the future, without the need for modifying the ball-array automating functions shown in Figure 4.6.

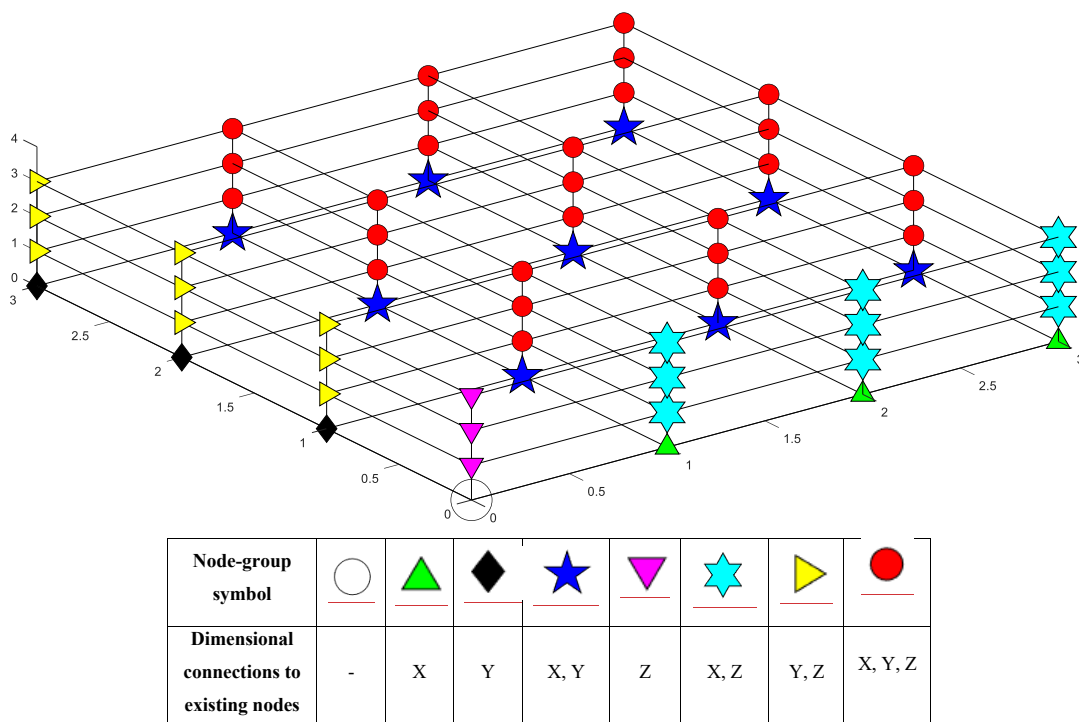


Figure 4.5: 4x4x4-ball connection cases during a SPICE-file generation, and a legend table. Except a transparent circle illustrating the first node generated at the coordinates (0,0,0), when adding a node represented by a unique colour/symbol shown in the legend table, it will be coupled with at least one existing node, depending on which location it is added to. For instance, a red circle node is coupled to the existing nodes in X, Y, and Z dimensions.

In the case that all of the balls in the system are of the same model, these three placeholders are a very trivial-effort task to modify by using a simple text editor or indeed via simple additional tools that can be written by future users in a command line environment for example.

Table 4.1: Example of TCA SPICE-code template generation.

Function	Example code (Some portions are omitted or split into new lines)
gen_ball	X_0_0_0 p0A_0_0_0 ... THE_REST_OF_YOUR_BALL_PORTS YOUR_BALL_SUBCIRCUIT_NAME YOUR_BALL_PARAMETERS
gen_system	X_0_0_0 p0A_0_0_0 ... THE_REST_OF_YOUR_BALL_PORTS YOUR_BALL_SUBCIRCUIT_NAME YOUR_BALL_PARAMETERS X_1_0_0 p0A_1_0_0 ... THE_REST_OF_YOUR_BALL_PORTS YOUR_BALL_SUBCIRCUIT_NAME YOUR_BALL_PARAMETERS X_0_1_0 p0A_0_1_0 ... THE_REST_OF_YOUR_BALL_PORTS YOUR_BALL_SUBCIRCUIT_NAME YOUR_BALL_PARAMETERS X_1_1_0 p0A_1_1_0 ... THE_REST_OF_YOUR_BALL_PORTS YOUR_BALL_SUBCIRCUIT_NAME YOUR_BALL_PARAMETERS

4.5.2 External Power/ground Rails Renaming

Following the ball-array generation process, the next process is the external power/ground rails renaming. This is the penultimate step prior to running a power-network simulation. There can be many methods to connect power sources at the surfaces of a TCA. However, the *Fully-connected* configuration is the main focus in this thesis. An example of fully-connected power configuration is illustrated in Figure 4.7(a). In Figure 4.1, there are three internal XY, XZ, and YZ plane generators inside the *External power/ground rails renaming* box. These three sub-processes are responsible for generating external power/ground SPICE node-names on all the six cubic surfaces. All the generated node names will be renamed for connecting with the external voltage source as shown in Figure 4.7(a).

As seen in Figure 4.7(b), some of the p and n names are replaced with $Vsrc$ and 0 , respectively. An example of a fully auto-generated file can be found in Appendix B*. Alternative connection configurations, which do not utilise full surface connections, are also possible. For instance, a scheme might have connections just at the eight corners, or the centre of each surface, or some combination. Although the investigation

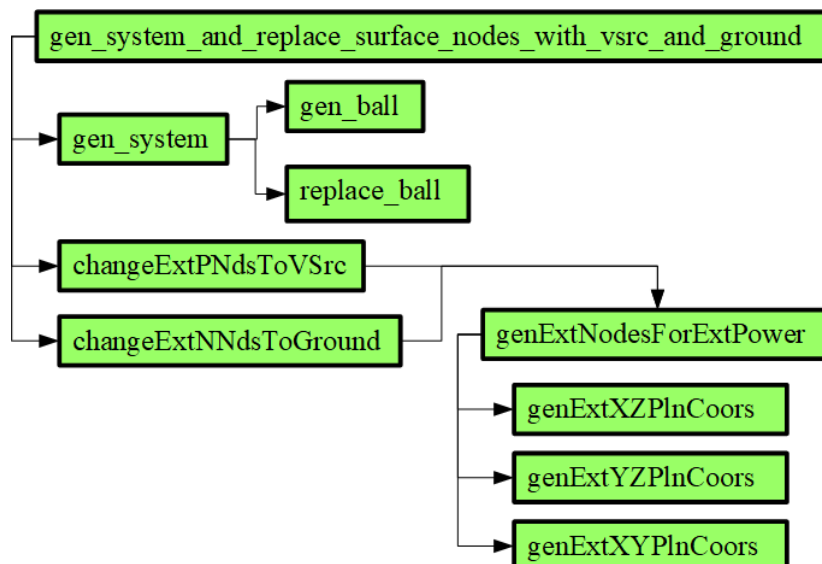


Figure 4.6: Function calls of TCA system generator and external voltage-source renaming.

* Most of the other file types generated by the system are not easily understandable to the reader and therefore are not included in the appendices.

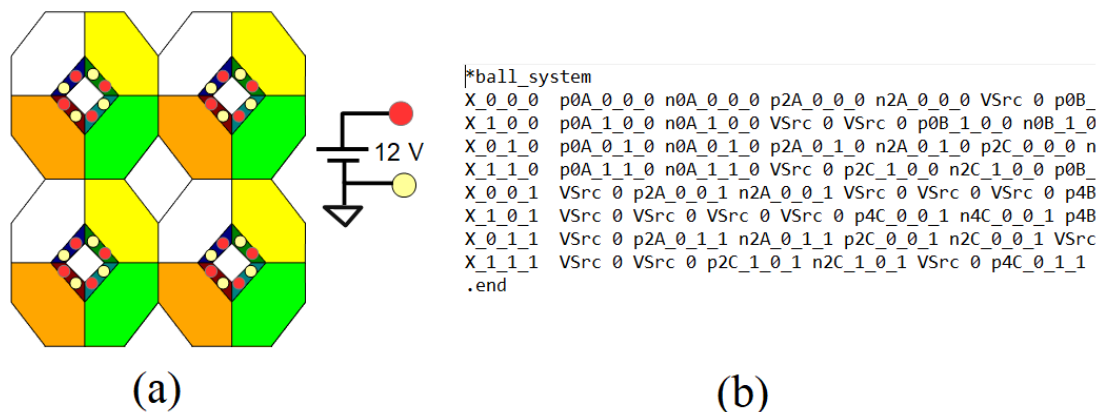


Figure 4.7: External rails renaming. In (a), this particular case, a single voltage-source of 12V is applied to all of the external pins (some external pins are invisible due to a 2D surface illustration). In real implementation, multiple pins at each edge can be used for each power/round rail to allow more current tolerance. (b) shows an incomplete code-snippet of the generated system.

of these partial connections are left for future modification of the framework for advance power connecting-point simulations, an advantage of the simulation tool-set developed is that this only requires modifications of a few functions in Figure 4.6.

4.6 Manual SPICE-file Editing

As shown in *Manual SPICE-file editing* box in Figure 4.1, this is the final stage performed in order to output a single complete SPICE-file, before running an actual SPICE simulation. It requires the user to manually replace the three placeholders for the ball sub-circuit ports, sub-circuit names, and all of the parameters, with their own custom cases, or else use default cases as provided in this thesis.

Theoretically, each ball instance in an array is not required to contain the same model. For example, in a heterogeneous system, balls, or even tile-level units, could conceivably utilise a variety of sub-systems, such as CPUs, FPGAs, DSPs, neural accelerator chips, memory banks, SSDs, etc. Fortunately, with the unique $\{X,Y,Z\}$ coordinates assigned to each ball instance, a *wrapper generator* like the external power/ground renaming function, can be additionally built to use these coordinates to perform a systematic renaming process to assign different ball sub-circuit-related

code sections for each case, and thus compose any possible heterogeneous array simulation desired.

With the completion of a TCA SPICE-file, it is now possible to perform a power-distribution grid simulation by either the uniform or non-uniform power allocation simulators, which will be detailed in the following section.

4.7 Simulation Modules

The *simulation modules* box, as shown in Figure 4.1, encompasses all of the other simulation tools. *Power-network simulator* is the core simulator, which is called for execution by both the uniform and non-uniform power allocation simulators. Visualisation tools to produce all the visualisations in this thesis are also categorised in this box. The most important parts of the simulation modules will be detailed in this section.

4.7.1 Power-network Simulator

The power-network represents a number of different attributes of a TCA system. At the stage of the research, only simple lumped-resistors and a single voltage-source models exist in this simplified power-network model. However, at the board-level model, a transient model of a switching regulator can also be embedded instead of the simplified board-model proposed in this thesis. With the transient model, the spikes of voltage and current can be observed for any particular interests, e.g., how they might impact on the lifespans of important components such as connector pins. This investigation obviously involves testing behaviours against specifications of components if provided, or real measurements on a hardware test bench. The detailed discussions on the simplified model itself can be revisited in Chapter 3, thus this subsection rather focuses on how the power-network simulator performs in detail.

An interesting example area for future research would be to examine the dynamic behaviour of a power network to understand more about the potential for spikes and transients and how resilient the regulators are in augmenting these effects. Alongside this is the possibility for some nodes in an array to operate as power-banks, acquiring charge in low power demand periods and providing localised additional power at other times.

Before deeply discussing the power-network simulator, it is worth summarising that a resistor in the network can present some or all of the following items:

- ▶ equivalent board-resistance when the board input voltage and current are steady. This resistor will be used to calculate the board-level steady power-consumption,
- ▶ the modelled resistance of a single or parallel pins used for power or ground rails on a tile edge,
- ▶ intra-board PCB-trace resistance. This model is not currently focused on in this thesis, but can also be modelled as discussed in Chapter 3.

With these pure-resistor elements for SPICE simulation, it does not only allow for faster simulation times for the purpose of scalability investigations, but also provides portability when a tool developer wants to migrate the models to be simulated on another SPICE simulator for some reason, e.g., parallel-simulation capability.

The *power network*, sometimes referred to as the *resistor network* in terms of SPICE simulation representations, is the basis of power network simulations reported in this thesis. The power-network simulator is a subset of the simulation framework and can be considered the *core engine*. This simulator can be run directly by the user, and can also be utilised to perform both uniform and non-uniform power simulations, which will be discussed in Subsections 4.7.3 and 4.7.4.

Apart from modelling the resistor-network itself, as described in Chapter 3, this resistor-network simulator also involves trade-offs in terms of implementation difficulties and the final achievable simulation accuracy. The *adjuster* module is responsible for adjusting a board resistance to conform to a board input voltage and current relation, which is profiled as a curve-fitting equation.

In prototype hardware instances, the adjuster allows the hex-tile PCB to emulate a range of power loads either statically or dynamically in bench-tests. However, in this chapter, the adjuster is discussed from the simulation point of view, as it can be considered a *virtual element* adjusting the simulated board resistance up or down until it becomes steady within the profile-threshold parameter.

In this thesis, two alternatives of its implementation were explored, 1) a circuit-based representation of the adjuster, and 2) A post-processing based approach. Each has advantages.

4.7.1.1 Circuit-based Board-resistance Adjuster

A circuit-based board-resistance adjuster can be implemented in a SPICE simulation file using analogue and/or digital devices. Different SPICE simulators may not provide these devices for simulation purposes with the same level of circuit abstraction, for example, behavioural or logic levels. Some SPICE simulators may also provide proprietary *black-box* modules, which are convenient for the tool designers. However, they may not always be portable when they need to migrate to another SPICE simulator that does not provide an equivalent module.

LTspice[®] is an example SPICE simulator that provides a useful *sample-and-hold* module, which is employed in the circuit-based adjuster in this thesis. It may also be possible to design an equivalent module with an amount of effort by using custom methods in each SPICE simulator, such as the *XSPICE framework* [83] provided in ngspice. The detail of building such a custom module is not in the scope of this thesis. Figure 4.8 shows the detailed implementation of the circuit-based adjuster proposed in this thesis.

As shown in Figure 4.8, there are three sub modules, which are *behavioural Sample and Hold function blocks* available as standard in LTspice[®]. These blocks are activated by a clock signal to periodically sample board input-voltage (v_{in}), board input-current (i_{board}), and board resistance (r_{board}). In a ball array, each of the board models contains all of these blocks and code elements.

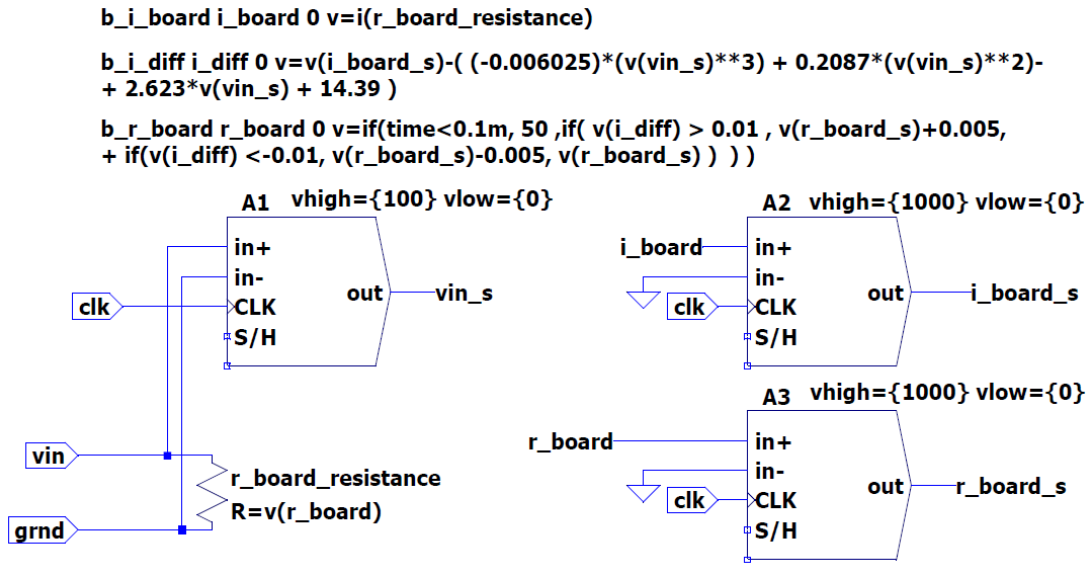


Figure 4.8: An LTspice[®] implementation of the whole board model with the circuit-based board-resistance adjuster in this thesis. This implementation is referred to as 'adjuster' published in [42], and supports only one regulated load-resistance (power) during a simulation. Thus, a separate simulation board-model file is needed for another load power consumption. Except 'r_board_resistance', the rest of the block modules and lines of code are part of the adjuster. '+' sign at the beginning is for the continuation of the line. The line with 'b_i_diff' implements the curve-fitting equation.

During a SPICE simulation, a free-running clock is fed to all of these board-model instances, allowing each to adjust its own board resistance to converge to a solution. Over a period during a time-domain simulation, if all of the board resistances in the entire system are found to conform with the *input-current difference threshold* (hard-coded as 0.01 in Figure 4.8), it indicates that the entire system is with that constraint, and the model is ready to be read out that solution and all of the measured voltages and currents for scalability analysis.

4.7.1.2 Post-processing Based Board-resistance Adjuster

An alternative implementation for the adjuster is to use a non-SPICE simulator, and instead employ a separate tool to read out a SPICE simulation file, and process the simulation data. In this thesis, MATLAB[®] is also employed for implementing this implementation of the adjuster.

With the board-resistance adjusting offloaded onto an external software tool, the

cycles of the clock signal are re-implemented as *iterations of single SPICE simulations*. At first glance, this technique seems to be awkward, however, the offloaded sequential circuit elements (sample and holds) are removed from the simulation model. Thus, this dramatically simplifies the contents in the final SPICE devices of the whole TCA power-network model, and was found to be preferred at the stage of the research due to the focus of simplifying the complexity of the simulation circuits to reduce simulation times. The implementation as software-based function call hierarchies are described in detail later, see for example Figures 4.17, and 4.18, in Subsections 4.7.3 and 4.7.4, respectively.

Having discussed the alternatives, it cannot be concluded that there is only a single absolute-design choice, depending on several factors such as background knowledge of design, the SPICE simulator and external tools employed, and tool maintainability. Some important advantages and disadvantages can be summarised in Table 4.2.

Regarding the post-processing based board-resistance adjuster, there are two adjusting modes developed in this thesis, which are as follows:

- **Static R-step mode:** Static R-step mode is a simple mechanism that the adjuster uses with a constant parameter, R_{step} , given by the user to adjust up or down all of the board resistances in the entire system. With the simplicity of this mode, the value must be low enough allowing each of the board resistance to reach the threshold interval of $(-I_{diff_thres}, +I_{diff_thres})$, which is the error distances from a point in the curve-fitting profile. The lower the I_{diff_thres} is, the better accuracy the voltages and currents in the simulation are. However, with a low value of I_{diff_thres} , it also indirectly requires a lower value of R_{step} . Otherwise, with an unsuitably high value, the adjusting process may be led into *oscillation state*, not converging into the $(-I_{diff_thres}, +I_{diff_thres})$ interval. One drawback of this mode is that if the initial board-resistance is wrongly *guessed* far from the value of steadiness, and if the R_{step} is set very relatively low, it incurs a large number of iterations of SPICE simulations. Thus, an alternative improved mode is also proposed, *Variable R-step mode*.

Table 4.2: Advantages and disadvantages of the two adjuster types categorised in this thesis.

Implementation	Advantages	Disadvantages
Circuit-based adjuster	<ul style="list-style-type: none"> ▶ More self-contained, an external tool is not required during board-resistance adjustment, reducing the processes of signal readouts to external post-processing tools ▶ Portable between SPICE simulators with no or small effort of modifications, if implemented with basic or equivalent SPICE devices ▶ Better reflecting on potential future real hardware capabilities such as additional circuitries for intratile level damage prevention like current/voltage threshold detection. 	<ul style="list-style-type: none"> ▶ Increase the complexity of the SPICE simulation file. ▶ Manually checking whether the entire system has converged into the regulator profile by visually monitoring a plot of a TCA system being simulated in time domain. ▶ Poor designs may lead to undesirable simulation circumstances such as convergence issues, memory-hungry simulation instances. ▶ difficulties for cross-compatibility if proprietary modules are used.
Post-processing based adjuster	<ul style="list-style-type: none"> ▶ Just 'run and wait-for-notification' until the external tool completes adjusting board resistances in iterations of TCA SPICE-simulations, each solving a DC-solution. This is also useful to notify an optimisation framework to process a completed simulation instance. ▶ Simplify the circuit complexity, offloading the profile checking tasks to the external tool. ▶ More availability of several useful tools by the external tools, e.g., data manipulations, etc. ▶ Additionally to parallel simulation (multi-core, etc.) feature that may be feasible in some SPICE simulators, reading out simulation results into an external tool also further expose data for the tool developer to employ parallel data processing available by the external tool. 	<ul style="list-style-type: none"> ▶ Each SPICE simulation instance solves a DC solution. If a visualisation of board resistances progressing towards convergence is required, an external tool needs to read each SPICE instance and plot signals in the same way like the circuit-based method does. ▶ May also incur additional tool costs, if not provided by the institution/organisation. ▶ May be error-prone for signal values if not carefully coded and signal-name rechecked after readouts.

- ▶ **Variable R-step mode:** is an advance mechanism which allows the initial *Rstep* parameter value to dynamically change during consecutive SPICE simulations for

adjusting board resistances. After a simulation, all of the I_{diff} values from the entire system are read out and used to obtain the maximum. This mode continuously performs *three data-point slope* as shown in Figure 4.9, determining which direction, up or down, and how far, $Rstep$ should take. If the maximum of I_{diff} in the current simulation is lower than the previous one, the adjuster increases the $Rstep$ value, accelerating all of the board resistances to approach the final stage of adjusting. Any initial board-resistance value is also mitigated, as this mode doubles $Rstep$ if the latest maximum of I_{diff} keeps going down.

In this mode, at the near-end stage of adjusting, an oscillation state may also occur, however, with the adaptability of this mode, finally $Rstep$ will be set to a low value that is adequate for reaching the $(-I_{diff_thres}, +I_{diff_thres})$ interval.

Having described both manual and automated parts for SPICE-file generation, and the power-network simulator, the next subsection will discuss the power allocation schemes investigated in this thesis.

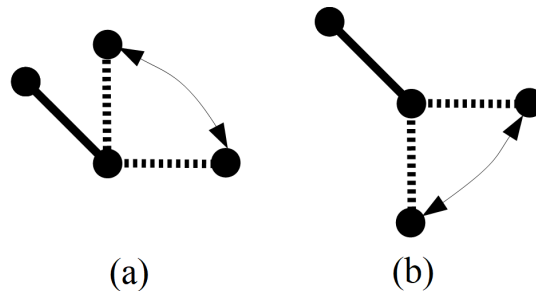


Figure 4.9: Illustration of the last three data-points of board resistance for calculating 'rate of change' in the variable R-step mode. (a) shows the case that the third data-point resides in the bounding up range, resulting in reverting the latest Rstep to the previous one. (b) is the acceptable range to double the Rstep value, accelerating the reduction of board-resistance to more quickly approach the board input voltage-current profile.

4.7.2 Power Allocation Schemes

Considering power allocation as being allocated at the level of the smallest packaged unit in TCA, *tile*, there can be many possible schemes applicable for many requirements and limitations. This involves both the design-time and run-time aspects of a

TCA system. The tree diagram in Figure 4.10 shows possible hierarchical categories of the power allocation schemes. In this thesis, uniform and a specific non-uniform type of allocation (*relative-position*) are proposed, and will be discussed in the following subsections.

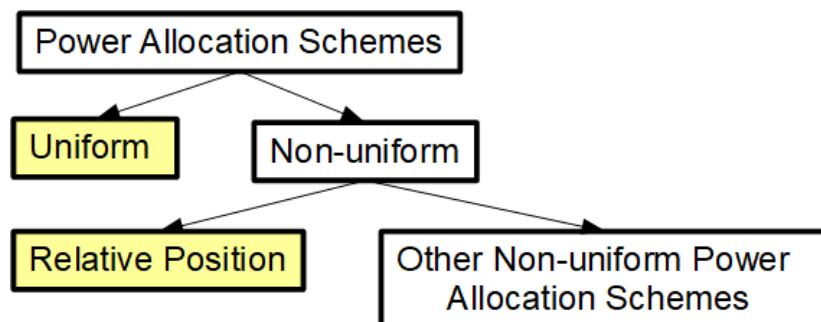


Figure 4.10: Possible hierarchical power allocation schemes in TCA.

4.7.2.1 Uniform Power Allocation Scheme

The *Uniform* allocation scheme can be considered the simplest method for allocating power in a TCA system. It consists of allocating all of the units with the same amount of power demand. This allocation scheme may be enforced due to a pre-designation of known-in-advance power-consumption upper bound per unit at design-time, or other requirements. For instance, uniformly rearranging power-limit per tile at run-time.

In terms of PE types existing in the system, this scheme is straightforward for *homogeneous* units, e.g., tiles are built to contain the same types of CPUs, FPGAs, etc. However, the uniform scheme can also be employed for *heterogeneous* systems. For example, a tile may utilise up to 100% of its allocated power solely by a single CPU at peak throughput, whilst another tile might comprise of an embedded lower-powered CPU coupled with a GPU for example, and thus that tile has combined power also within the total allocated amount of power.

4.7.2.2 Relative-position Allocation Scheme

Whilst the uniform scheme is straightforward, another possible scheme, *non-uniform*, which varies the allocated amount of power per tile, has some advantages over the uniform scheme.

In this thesis, *relative position* is proposed as a subset-scheme to intentionally reduce the number of simulation cases when a TCA is powered with all of the external power connectors. With the fully-connected power on a symmetric TCA dimension, e.g., cubic array shape, it therefore also implies that an optimal power allocation pattern across the system would also follow symmetric properties. This allocation scheme is proposed for simulation purposes, however, it can also be practically employed for allocating power in a real hardware system (for example where the computational load at each node can be predicted according to the workload being operated). Examples of nodes with a relative-position scheme can be illustrated in Figure 4.11. In Chapter 5, the reduction of simulation cases will also be discussed in Subsection 5.3.1 when performing brute-force simulations. A comparison of tile-by-tile non-uniform and relative-position schemes can also be found in Figure 5.11.

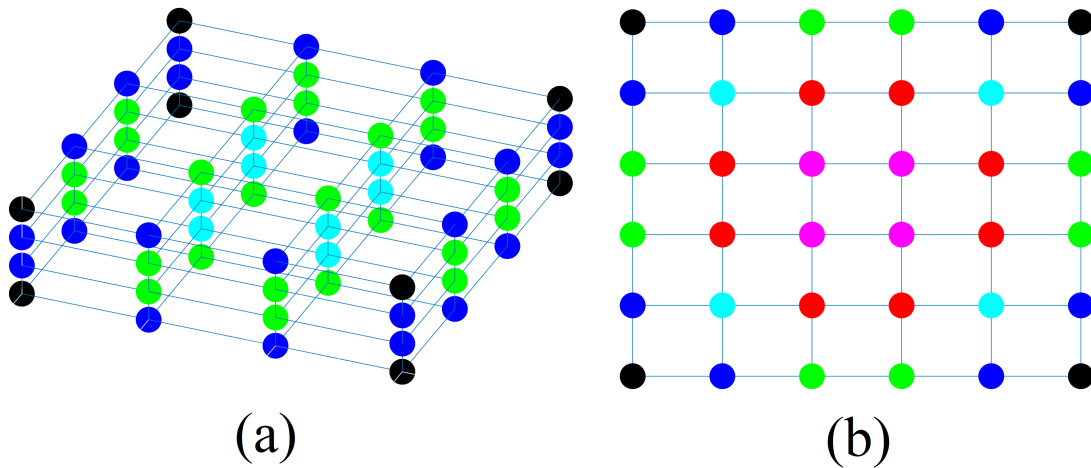


Figure 4.11: Visualisations of a couple of relative-position schemed TCAs. (a) shows a 3D-view of a 4x4x4-tile, and (b) shows only a 2D-surface of a 6x6x6-tile due to the abundant quantity of relative coloured-groups of internal 3D layers. The same colours in each of the arrays in (a) and (b) represent the same allocated amount of power.

As will be explained, a concept referred to as *two-point distance* is useful to consider at this point. At first glance, it may be believed that Equation 4.1, the standard

two-point distance equation (derived from the standard Pythagorean theorem), may be employed to determine a node distance from the system centre-point.

$$distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (4.1)$$

However, by the visual proof in Figure 4.12, it can be observed that this is not the case, and in-fact nodes in the same relative-positioned group cannot be simply calculated via the equation for the distance from the system centre to the desired coordinates. This is due to the fact that in a cube-shaped TCA, there are some nodes (tiles) with the same two-point distance but they are not in the same cubic layer as shown in Figure 4.13.

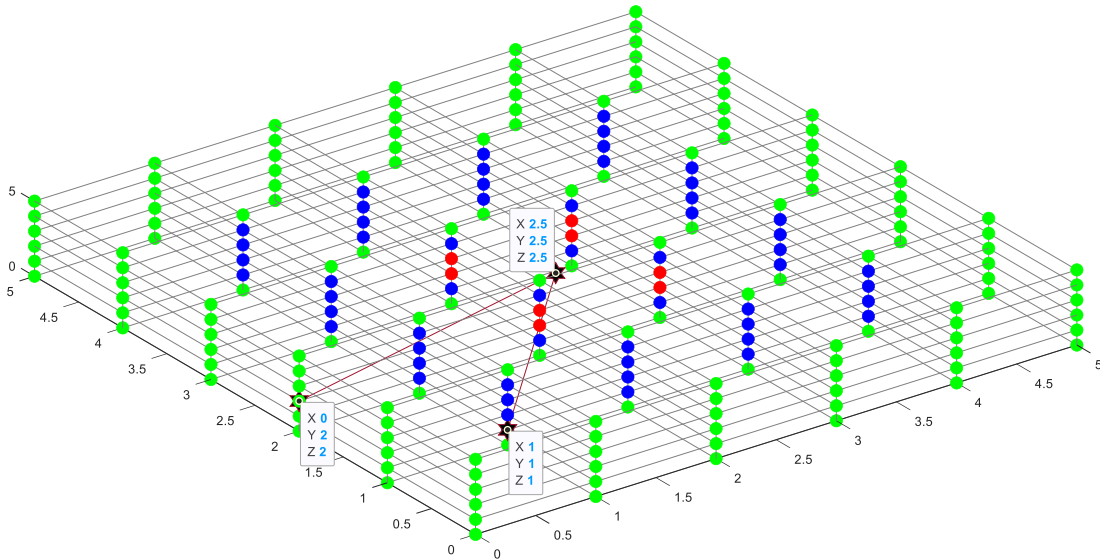


Figure 4.12: Example showing that a two-point distance in a TCA is not valid for checking whether nodes (tiles) are equally impacted by external fully connected power. Even though the two coordinates of (0,2,2) and (1,1,1) have the same two-point distance of 2.5981 units from the system-centre coordinates (2.5, 2.5, 2.5), the first node is in the outermost layer, whilst the second node is in the first inner one. It is noted that the coloured nodes in this example show cubic layers, whilst Figure 4.11 explains a different subset node grouping, relative-position scheme.

It may be further complicated if (as might be the case) in a cubic layer, not all the nodes are equally impacted by voltage drops. As a result, the relative-position scheme is proposed in this thesis. This can be seen in Figure 4.11. To clarify the inclusion of the relative-positioning amongst the power allocation schemes discussed in this thesis, a diagram showing allocation schemes is given in Figure 4.14. The relative-position

scheme refines the cubic allocation approach to achieve a better power allocation overall. Thus, if it is desired to allocate power to nodes in terms of their layers within the array, an automating module is created purposely for this power allocation scheme.

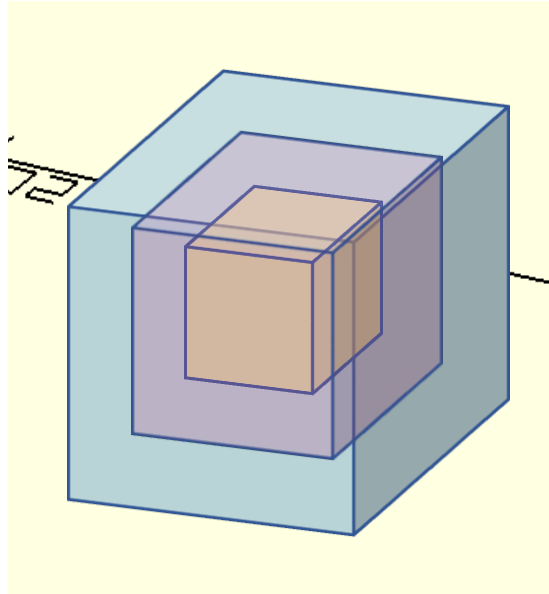


Figure 4.13: Example three cubic layers of nodes as a possible power allocation scheme for TCA systems.

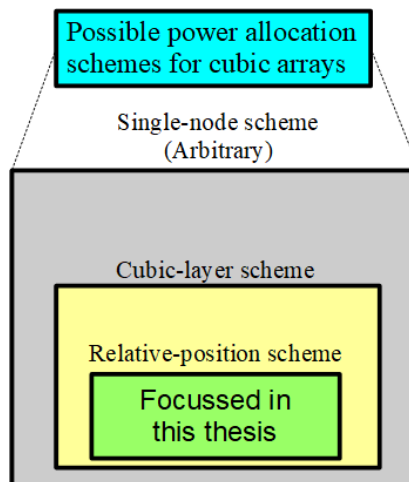


Figure 4.14: Diagram showing relations amongst power allocation schemes for cubic-array systems.

Algorithm 1 implements the generation of the total number of node groups, and a data structure containing pairs of tile-coordinates and the group-ID to which they belong. To give an example of node groups generation, generating node groups for a 4x4x4 tiles in Figure 4.11(a) starts off with the surface layer, grouping the four corner-tiles highlighted in black, followed by the dark-blue, and the green ones. At this stage, all the groups in the layer have been completed. For the next inner layer, which is also the deepest one in this particular case, the four tiles are given the same group-ID highlighted in light-blue as the last step. In Figure 4.11(b), it shows only one of the six surfaces. Following the same process per layer, the order of grouping the tiles in the outermost layer is the tiles highlighted in black, dark-blue, green, light-blue, red, and pink, respectively.

To further explain, In Figure 4.12, two straight-line measurements are given to visually prove that these two nodes at the coordinates of (0,2,2) and (1,1,1) have the same distance of 2.5981 units measured from the system centre at the coordinates (2.5, 2.5, 2.5), but they are in different layers. The node at (0,2,2) is directly supplied by the external power and no voltage drop exists, whilst the node at (1,1,1) resides in the first outside-in internal layer, experiencing a voltage drop. This can be obviously seen in a custom-designed visualisation in Figure 4.15.

It is shown that having discussed only the cubic TCA, and only with a power connection model in which all external power connectors are fully connected to power supplies, there are several factors for both simulation and practicality issues. Thus, it is unsurprising that arbitrarily constructed systems would further introduce non-uniform and even more unpredictable nodal voltage drops and current magnitudes and directions all over the system. For example, a mesh of heterogeneous tiles with different characteristics, or a non-cubic structure. Examples of arbitrarily-shaped systems are shown in Figure 4.16.

The concept of TCA does not limit the flexibility of inter-node level construction. However, it should be borne in mind that non-uniform power supply models and arbitrarily-shaped systems are more complicated in terms of simulation-case reduction, and also possibly increasing the complexity of SPICE simulation issues. This underlines

Algorithm 1: Generate relative-position node groups.

Data: Number of tiles S in each dimension of a cube-shaped system

Result: Number of groups N generated in the relative-position scheme,
A key-value data structure $M(x, y, z)$, where a key (x, y, z) is the coordinates of a tile to find its associated group number

```

1  $x \leftarrow 0$ ;
2  $y \leftarrow 0$ ;
3  $z \leftarrow 0$ ;
4  $N \leftarrow 0$ ;
5  $minCoorForLayer \leftarrow 0$ ;
6  $maxCoorForLayer \leftarrow S - 1$ ;
7 while  $minCoorForLayer < maxCoorForLayer$  do
8    $groupNumsInPlane \leftarrow NULL$ ;
9    $minCoorForPlane \leftarrow minCoorForLayer$ ;
10   $maxCoorForPlane \leftarrow maxCoorForLayer$ ;
11  while  $minCoorForPlane < maxCoorForPlane$  do
12     $groupNumsForFrame \leftarrow NULL$ ;
13     $maxGroupIdxForFrameOfPlane \leftarrow$ 
14       $((maxCoorForPlane - minCoorForPlane + 1)/2) - 1$ ;
15    for  $groupIdx \leftarrow 0$  to  $maxGroupIdxForFrameOfPlane$  do
16       $groupNumsForFrame(groupIdx + 1) \leftarrow groupCnt$ ;
17       $groupCnt \leftarrow groupCnt + 1$ ;
18    Assign group numbers to nodes in X axis;
19    if  $minCoorForPlane \neq (S/2) - 1$  then
20      Assign group numbers to nodes in Z axis;
21       $minCoorForPlane \leftarrow minCoorForPlane + 1$ ;
22       $maxCoorForPlane \leftarrow maxCoorForPlane - 1$ ;
23
24    // Plane XZ_YMin has already been generated by the above section.
25
26    Serialise group numbers in a plane for generating the remaining planes;
27
28    Generate plane XZ_YMax;
29    Generate plane YZ_XMin;
30    Generate plane YZ_XMax;
31    Generate plane XY_ZMin;
32    Generate plane XY_ZMax;
33
34     $y \leftarrow y + 1$ ;
35
36     $minCoorForLayer \leftarrow minCoorForLayer + 1$ ;
37     $maxCoorForLayer \leftarrow maxCoorForLayer - 1$ ;

```

Maximum voltage-drop over a unit is 0.3426592993119062 volt(s)

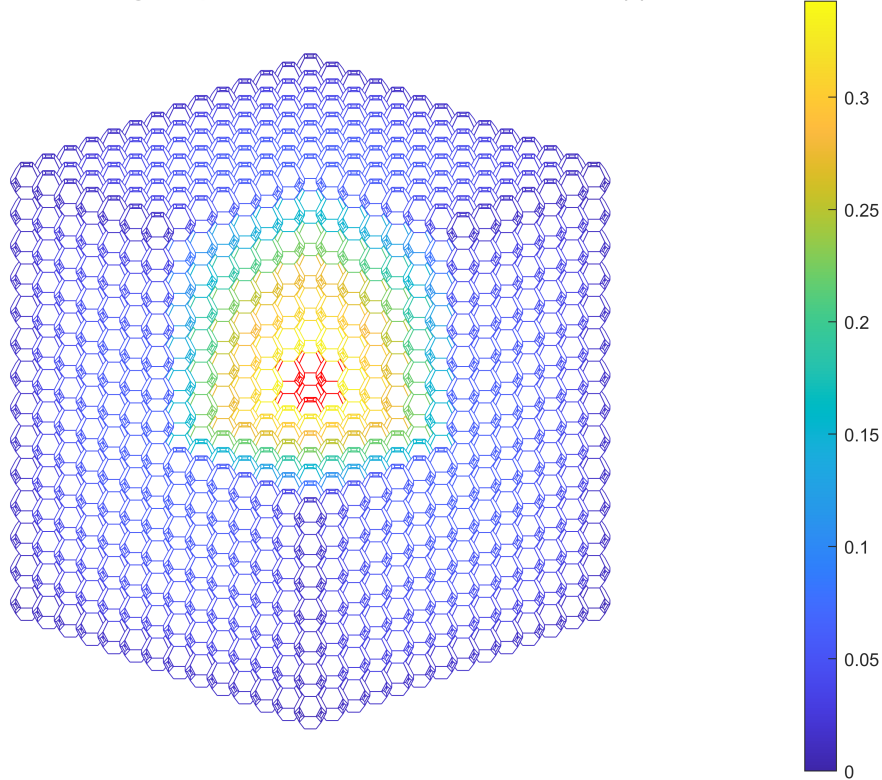


Figure 4.15: TCA system illustrating voltage drops over the entire system. In this particular system, for an illustrative purpose a ball is implemented as the smallest unit. Whilst this thesis focuses on a tile as the smallest unit.^a

^a The actual voltage drops at the innermost balls are slightly different from their immediate outer layers coloured in yellow. However, they are highlighted in red to explicitly show the direction of the voltage-drop trend.

the importance of having a simulation tool, therefore, to assist in such power-network evaluations when needed.

To summarise this subsection, as seen in Figure 4.10, only uniform and relative-position power allocation schemes are focused upon in this thesis and the point of view taken in the hierarchical schemes is pure *power-oriented*, not taking any *inter-connection network traffic patterns* as shown in Table 2.5 into consideration.

Without communication patterns involved, an optimal power allocation result may not also mean an optimal task mapping for a particular workload partially or entirely submitted into the system. Nevertheless, in future work, more sophisticated schemes, for instance, *submitted-job based* allocation can also be taken into account as a co-decision scheme along with these two, or other power-oriented schemes.

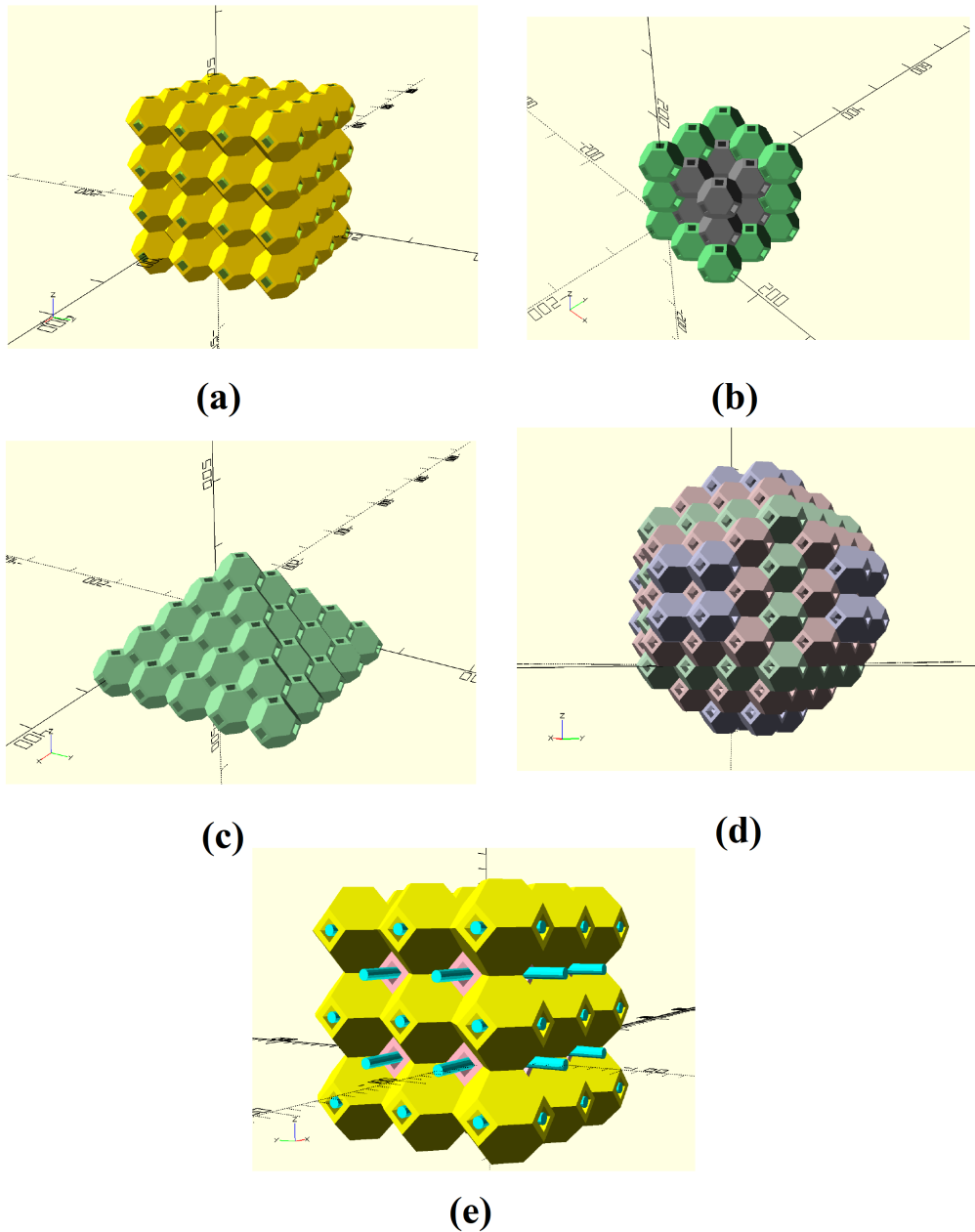


Figure 4.16: ^a Examples of different system shapes. (a) simple 3D mesh. (b) non-complete 3D mesh with partial outer layer nodes removed. (c) pyramid shaped system (d) sphere (e) simple 3D mesh similar to (a) but using a double-packed array (pink nodes packed between yellow), and showing the flow channels for cooling highlighted in blue.

^a Generated by Christopher Crispin-Bailey with OpenSCAD [72] toolset

In the following subsections, two higher-level simulation modules that invoke power-network simulation, with uniform and non-uniform power simulators, will be discussed.

4.7.3 Uniform Power Simulator

The uniform power simulator is one of the two main top-level modules in the whole simulation framework, invoking the power-network simulator and various other modules. This simulator automates a number of sub-processes, ranging from simulation cases, SPICE-file preparation, and finally issuing a set of parallel SPICE simulation instances.

In the hierarchical function calls shown in Figure 4.17, it can be seen that *uniform_power_param_sim*, the actual function of this simulator, calls *Power-network simulator*. This simulator, implemented as a software module, can be instantiated as *parallel processes*. Each of the instances then issues a background SPICE simulation running in parallel with the others. With this parallel simulation capability, multiple simulation instances can be completed in a relatively faster timescale compared to successive serial simulation runs.

4.7.4 Non-uniform Power Simulator

The *Non-uniform power simulator* is another top-level simulator. This simulator is designed to optimise power allocation in a TCA system incorporating the *relative-position* scheme. An elitist genetic algorithm (GA), which is a variant of NSGA-II [84] provided in a MATLAB[®] toolbox [85], is employed for the multiple-objective optimisation of system-level regulated load-power and the worst-case connector-pin current. As seen in Figure 4.18, *ga_sim_gamultiobj_ngspice_assigned_board_resist* is the actual main function implementing this simulator. This specific function name is due to *ngspice*, an open-source SPICE simulator, and is employed for the post-processing based board-resistance adjuster at this stage of the research. *gen_relative_position_node_groups* is an auxiliary function, calculating the number of node groups when using the relative-position scheme for cubic-array systems.

In the GA, a chromosome represents a number of possible attributes that can be varied independently to create an overall outcome. The number of node groups will



Figure 4.17: Function calls of TCA system uniform power simulator.

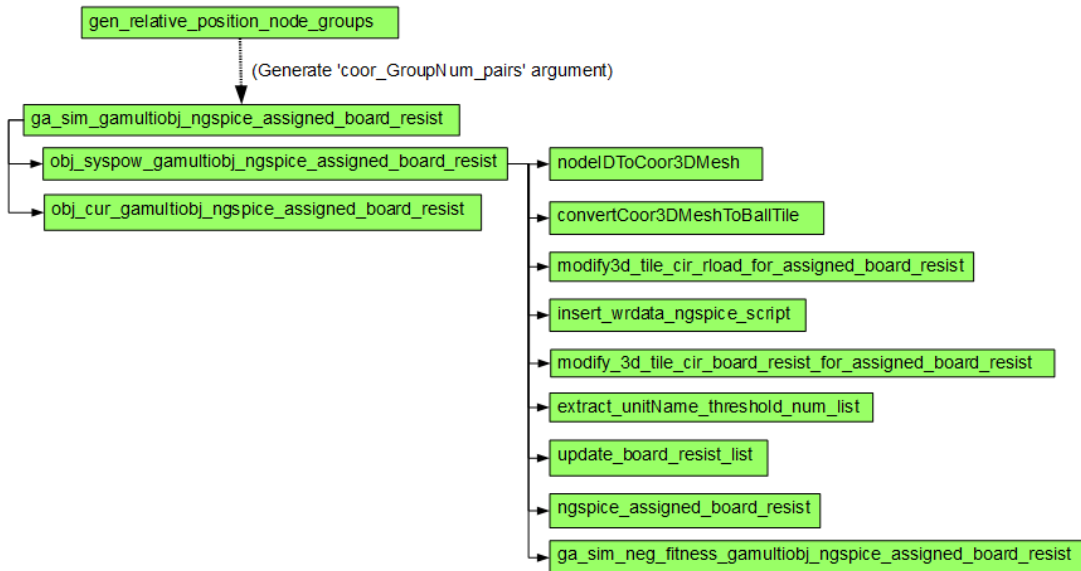


Figure 4.18: Function calls of TCA system non-uniform (GA) power simulator.

dictate the length of the chromosome in the GA employed.

The GA simulation framework is investigated here with both single and multi-objective goals. These cases are described in the next subsections.

4.7.4.1 Single-objective GA-optimised Simulator

The single-objective GA-optimisation will be briefly discussed in this subsection as it was developed in an early stage of the research for an initial framework for a non-uniform power simulator. A single-objective optimisation tool may be adequate for approaching only a single goal. For example, a TCA may be desired to allocate a minimum of total system-level load-power of 1000W, whilst allocating power all over the system to optimise (lower) connector-pin current as much as possible.

Conversely, a user may start off with the requirement of connector-pin current limit of 3A and might desire to maximise the system to gain the total regulated load-power as high as possible. On this single-objective manner, when the constraints of interest as aforementioned are to be varied to observe a range of potential solutions, multiple simulation instances are required. This thesis focuses on the two main important factors, *system-level regulated load-power* and *connector-pin current*. For *voltage-drop*, it is also another important constraint, however, it can be separately mitigated by increasing the external voltage level. In fact, higher external voltages do not only help with voltage drop, but also reduce the worst-case connector-pin current itself. More than that, even though voltage drop is not included as part of the objectives, the simulation results of *power and current* mentioned above can also be read out to recheck, as a post processing, whether it violates the voltage-drop constraint. With all of the factors considered, therefore, a multiple-objective GA framework is proposed and will be discussed in the following subsection.

4.7.4.2 Multiple-objective GA-optimised Simulator

As mentioned earlier, a given system design goal may not be only to comply with single constraints such as connector-pin current, or voltage drop, in isolation. Desired design objectives may seek to satisfy multiple quantitative-goals in a solution. In this thesis, a two-objective simulator is proposed to demonstrate this kind of optimisation.

Consider system-level regulated-power and the worst-case connector-pin current. Those two quantities can be extracted from a single TCA SPICE-simulation result file. Thus, there is no need to repeat the same simulation twice for each of the two objectives. In consequence, only a mechanism to detect the completion of a single simulation instance is required. The multiple-objective simulator proposed in this thesis operates as follows:

- ▶ Evaluation begins by calling the first objective function to initiate a TCA SPICE-simulation, and then waits until it is completed, reads out the simulation results, and calculates system-level regulated load-power.
- ▶ Meanwhile the second objective case waits until the first-objective function has produced a separate worst-case connector-pin current report file before issuing its results.

An example is illustrated in Figure 4.19: The first-objective function issues SPICE simulations. As the second-objective function needs to know when each of the simulations is completed, there must be a method to check whether a given simulation result is ready. For this purpose, MD5 [86] is employed to generate a unique file name. Some meta-data such as, date-time, and a random number are also appended to ensure that the file name generated is not repeated, due to the fact that in different GA iterations some chromosomes may be encountered multiple times.

The file name and extension format and examples of the report files can be seen in Figure 4.20. In the present version of the GA simulation framework, it is possible that the objective 1 can produce the same individuals that have already been evaluated in previous populations. This can be further improved in future work in several ways: 1) use or modify the GA that does not produce redundant chromosomes in the same

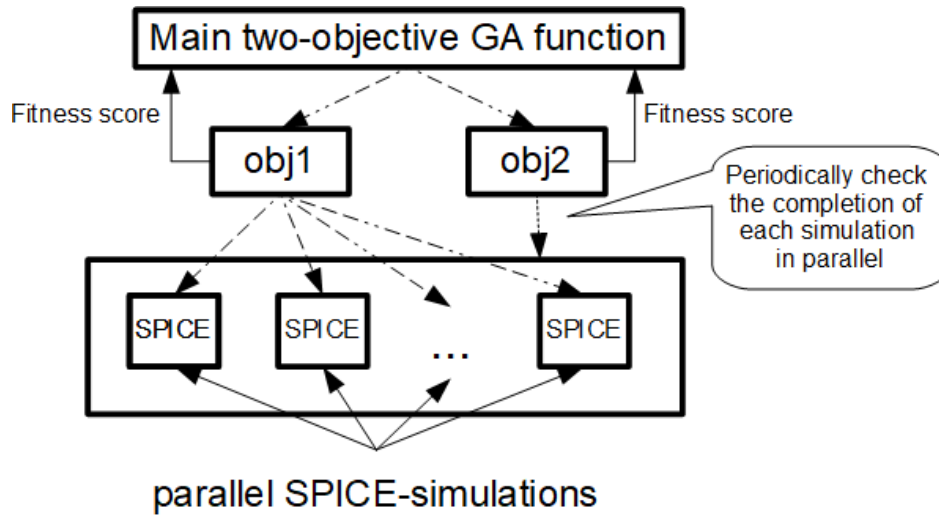


Figure 4.19: Mechanism of the objective 1 and 2 associating with each other for SPICE-simulation instances.

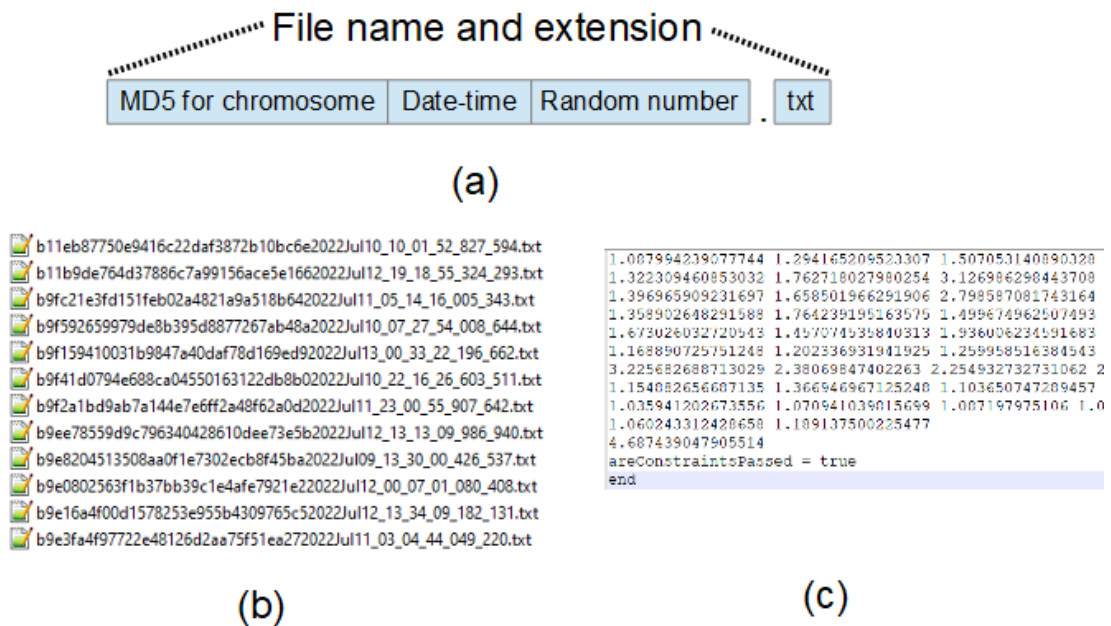


Figure 4.20: (a) TCA multiple-objective simulation result file name and extension format. (b) examples of generated result files. (c) the content format of a GA report file. Due to a long length of chromosome (the tile regulated-side load-resistance values), some of the values are omitted. The last line with a numerical value reports the worst-case maximum pin current. The line with 'areConstraintsPassed = true', can be used for post processing to check whether any constraint, e.g., voltage drop is within the acceptable range.

or different populations, or 2) modify the current TCA simulation framework itself to check whether the current population produces any redundant individuals before invoking any parallel SPICE simulations.

These approaches would therefore eliminate unnecessary computing time and resources, speeding up convergence, particularly for large array cases.

The relative-position power allocation scheme is not only beneficial for reducing cases to observe for brute-force simulations, but also for the GA simulation framework. Incorporating the relative-position scheme, it can reduce the size of each chromosome in a population, which also means decreasing the memory requirements of the machine performing the simulations. The comparison between the conventional tile-by-tile mapping and the relative-position alternative can be seen in Figures 4.21 and 4.22, showing the advantage of the shorter chromosome-length in the latter allocation-scheme. In this particular case of a 2x2x2-ball system, the chromosome size is dramatically reduced from 64 to only 4 genes per chromosome. The process of genotype-phenotype mapping is internally implemented in the first objective function, as it is responsible for starting a TCA-SPICE simulation. Whilst the conventional tile-by-tile mapping is also implemented in the objective function, the non-uniform simulation results reported in this thesis are based on the relative-position mapping. Table 4.3 and Figure 4.23 detail the objectives, constraints, and parameters in the two-objective GA used in this thesis.

Given the details of both the uniform and non-uniform power allocation simulators, a concise illustration of the overall workflows can be seen in Figure 4.24. Each workflow summarises the internal mechanism of the main function of the simulator implemented as a MATLAB[®] function. In Figure 4.24, whilst a series of steps in the yellow frame for the uniform simulation is implemented in a MATLAB[®] *parfor* loop [87], the two-objective GA-based power-distribution grid simulations are run with the support of computing fitness functions in parallel provided by the MATLAB[®] *gamultiobj* [85]. For the uniform simulation, the final result is a single textual report-file, containing all the three quantities of the maximum connector-pin current, minimum board input-voltage, and the electrical current fed into the whole system by the external voltage-source. The system-level power consumption is also included in the report file. However, extracting the system-level power consumption and connector-pin current in the non-uniform works in a different way, which can be extracted from a complete pareto-plot after the simulation is complete. The processes described above are for the tool

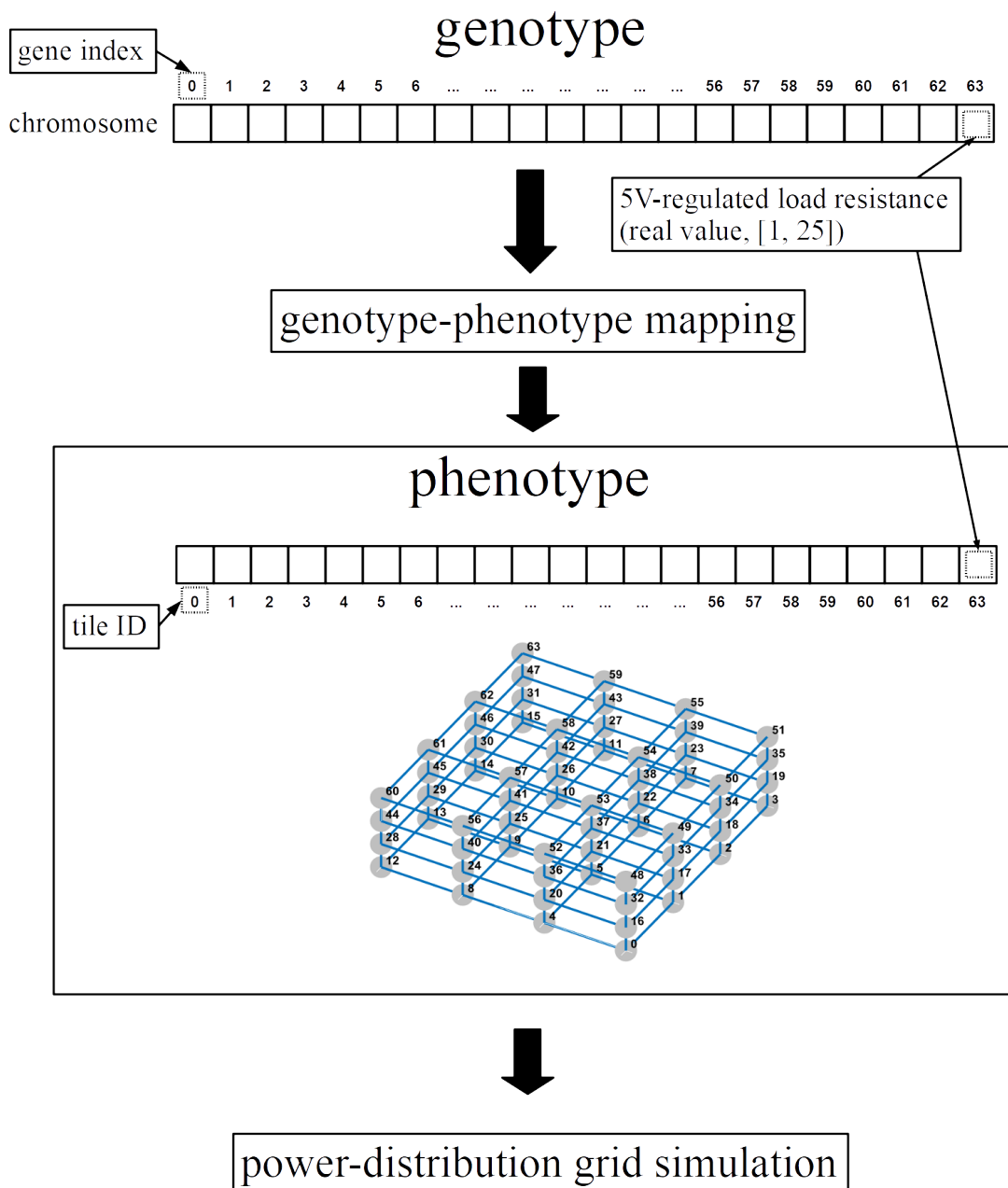


Figure 4.21: Conventional mapping of tile-by-tile power of a 2x2x2-ball to a chromosome in the genetic algorithm employed.

developer's perspective. However, from a tool user's point of view, both the uniform and non-uniform simulations can be performed by only running the main functions of the two types of simulator.

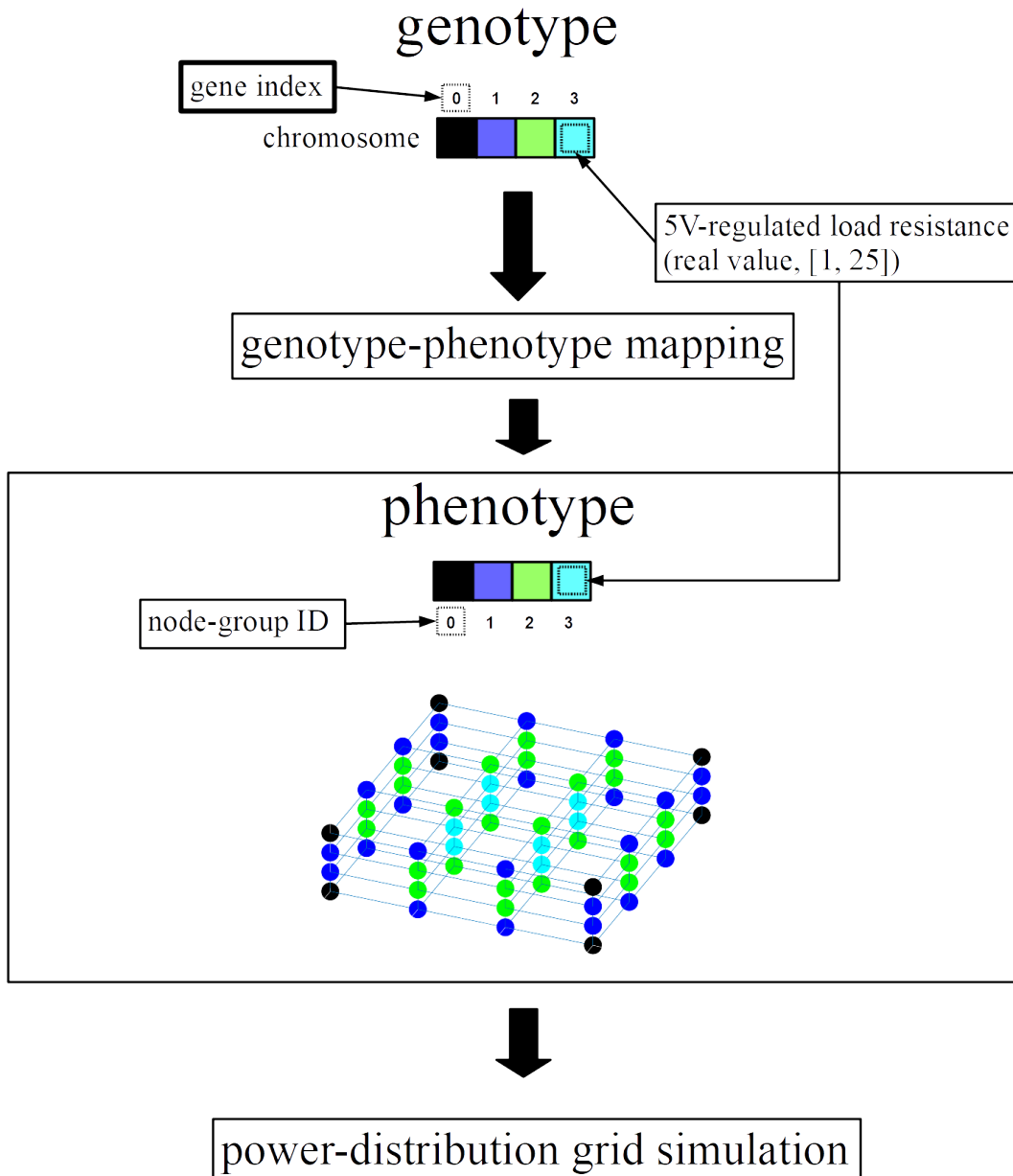


Figure 4.22: The reduction of the chromosome size compared to the mapping in Figure 4.21 when using relative-position scheme. Further more reductions can also be seen in larger sizes shown in Figure 5.11.

Table 4.3: Overview of the details of the two-objective GA used in this thesis. Additional details specifically in one of the MATLAB® gamultiobj's parameters named 'options' can be found in Figure 4.23.

List	Related Input/Output Values	MATLAB® implementation (gamultiobj parameters [85])
Objective 1: Maximising the system-level regulated-power	Type: Output minimum: Total number of tiles × 1W maximum: Total number of tiles × 25W	fun: Fitness functions to optimize
Objective 2: Minimising the worst-case connector-pin current	Type: Output Unknown minimum or maximum, as the quantities are part of the simulation results.	
The two-objective constraints: Only bound constraints	Type: Input 5V-Regulated load-resistance: - Lower bound: 1Ω - Upper bound: 25Ω	lb: Lower bounds ub: Upper bounds
Objective-function parameter: Acceptable worst-case board input-voltage	Type: Input - 6V Note: Fitness scores for violation (under 6V detected): Objective 1: 0 Objective 2: MATLAB®'s realmax	Not used: (Internally implemented in the two objective functions)

```

Set properties:
  InitialPopulationMatrix: [2×56 double]
      PlotFcn: {@gaplotpareto}
      UseParallel: 1
      UseVectorized: 0

Default properties:
  ConstraintTolerance: 1.0000000000000000e-03
  CreationFcn: []
  CrossoverFcn: []
  CrossoverFraction: 0.8000000000000000
  Display: 'final'
  DistanceMeasureFcn: {@distancecrowding 'phenotype'}
  FunctionTolerance: 1.0000000000000000e-04
  HybridFcn: []
  InitialPopulationRange: [2×1 double]
  InitialScoresMatrix: []
  MaxGenerations: '200*numberOfVariables'
  MaxStallGenerations: 100
  MaxTime: Inf
  MutationFcn: []
  OutputFcn: []
  ParetoFraction: 0.3500000000000000
  PopulationSize: '50 when numberOfVariables <= 5, else 200'
  PopulationType: 'doubleVector'
  SelectionFcn: {@selectiontournament [2]}

```

Figure 4.23: Example list of the MATLAB® gamultiobj's parameter named 'options' [85] used for the 6x6x6-ball system simulated in this thesis.

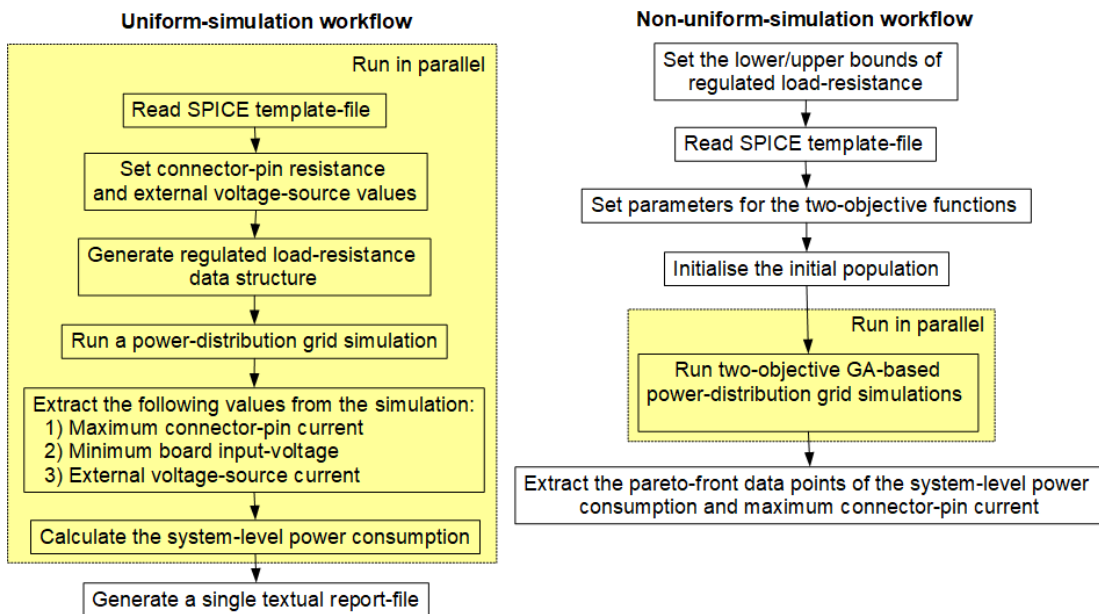


Figure 4.24: Workflows of uniform and non-uniform power allocation simulations.

4.8 SPICE Simulator

Referring to Figure 4.1, the SPICE simulator is called via an operating-system command. With this method, it is considered being loosely attached into the framework, which can be flexibly replaced with another SPICE simulators being used. ngspice [70] is chosen in this thesis, as an example of open-source SPICE simulator. However, a non open-source freeware SPICE simulator such as LTspice[®], is a powerful SPICE simulator providing a number of useful simulation blocks. For instance, the *sample and hold* is employed in the version of circuit-based board-resistance adjuster proposed in this thesis. Using this alternative of the adjuster, a single time-domain power-network simulation can be run purely on a SPICE simulator towards the completion of simulation at the equivalent steady-state of a voltage regulator board input voltage-and-current profile. This contrasts with the post-processing based adjuster being part of an external function, which is more flexible but requires an external tool to issue it for running.

4.9 Chapter Summary

In this chapter, the TCA power-grid simulation framework has been discussed in detail. Including the simplified model which has already been discussed in Chapter 3, it can be concluded that the modelling and simulation framework tackle the problematic simulation issues with the following contributions:-

- **Simulation difficulties:** The simplified models tackle long simulation times, large simulation result files, and model portability. Whilst the simulation framework includes a coordinate-aware power allocation scheme, relative-position, to reduce simulation efforts for non-uniform power allocation optimisation. Two choices of board resistance adjusters are also proposed for flexible implementations both in the circuit-design style, or a software-equivalent alternative. Visualisation tools are also provided both for meaningful representations and tool verification purposes.

- ▶ **Labour-intensive simulation efforts:** Manually creating sequences of simulation processes for large-scale simulations does not only take tremendously efforts, but also incur error-prone tasks. The simulation framework tackles these issues by providing a large number of flexible automating simulation tools to ease several processes of fundamental and optimisation simulations.
- ▶ **Simulation-framework extension and portability:** A good simulation tool may not be considered only for its purpose at present, but also involves its easiness for extension in the future.

This chapter fulfils the research objectives 3, 4, and 5. Following the model validation, and detailed simulation framework, as already discussed in Chapters 3 and 4, this chapter is dedicated to evaluating the scalability of example systems and also to certain topological aspects. The results in this chapter are an essential contribution to answer whether the main stated research hypothesis is true or not.

For interconnection network performance, *latency* and *throughput*, are the two key metrics [15], [37]. Whilst in this thesis, for TCA power network scalability the key metrics are *voltage drop* and *connector-pin current*.

It is important to emphasise that in real implementations, there are likely to be broader sets of electrical concerns, for example: voltage and current spikes during system start-up and real-time varying power demands due to computational load changes, among others. However, as described in Chapter 3, for the large-scale simulation perspective, it is assumed that the board-resistance model mimics a node's constant power consumption at a steady state. This model can be employed for evaluating a system with nodes consuming constant power in the situations of interest, for example, worst-case, or limited power-level consumption. Therefore, the simulation results in this thesis predict system behaviour when all the nodes are steadily operational.

A future goal may well be to extend the existing implementation to a more advanced approach in which dynamic changes in tile or ball power loads are emulated in order to assess aspects of dynamic power network performance. However this is outside the scope of the work presented here.

In this chapter, power-network simulation, as the main focal aspect of this thesis, will be thoroughly reported. Some initial topological analyses and simulations are briefly discussed with perspectives for further detailed investigations in future work. The power-network evaluation results in this chapter are mainly based on Analog Devices LT[®]3976 voltage regulator [65], with the simulation parameters of

$I_{diff_thres} = 0.01A$, 50 m Ω mated pin-pair (of inter-tile connection), and two parallel-pins for each of the power or ground rails on the same connector, to demonstrate how TCAs are evaluated for power-network scalability. However, in practice a TCA system can also be implemented using any other switching regulators of choice, depending on different power and other electrical characteristics required, or indeed with other types of on-board power management. As mentioned earlier in Section 4.5, the simulator platform's modularity allows this to be easily changed.

Optimisation strategies for TCAs are not only limited to achieving the possible maximum system size with the constraints of connector-pin resistance and required voltages supplied across tiles. Sometimes, these two scalability-affecting factors are more than adequate for an expected scale of system required, and then other desirable constraints or goals might be considered. Examples for other evaluation scenarios that may arise from practical constructions are as follows:

1. Where the available external power unit may only be able to provide a limited amount of power, or a desired system-level regulated load-power target/limit is given.

This objective can be applied to certain practical cases, for example, the power budget allocated for a high-performance facility in which a TCA is employed, or when a researcher/developer needs to investigate how an amount of power given to a TCA impacts on overall computing performance of a certain workload.

2. Where a set of non-uniform load-wattages is required to be efficiently allocated to the entire system or sub-regions in a TCA, with a goal of achieving as low as possible connector-pin currents and/or voltage drops.

With this objective, a TCA system composed of *pre-designated* power profiles for different computing-related unit types, e.g., CPUs, FPGAs, storage, can be effectively added to their suitable system coordinates. However, an optimal power solution may not always be an optimal computing-performance solution. For example, a group of TCA balls may need to contain a single CPU surrounded by FPGA nodes to support some specialised workload with maximum computational efficiency. A pure power-aware optimal solution may allocate

node types in such a way that results in optimal power goals, but poor computing performance affected by multiple communication hops.

3. Both the system-level regulated load-power and the connector-pin current constraints are considered. This is an example of a two-objective optimisation problem. This chapter also provides simulation results for this specific goal.

Apart from the three example cases above, there can of course be many more evaluation requirements arising in the construction phase of a system. In this chapter, some GA-based simulation results are given as examples for tackling these kinds of multiple-objective requirements.

5.1 Relevant Research Objectives

The research objectives 3, 4, and 5 are relevant to this chapter. The sections in which they are achieved are given as follows:

5.1.1 Objective 3: Fundamental Simulation Experiments

Methodologies/Activities:

- ▶ **Parallel simulation:** As large simulation sizes incur long simulation times, parallel simulations on single or multiple machines reduce the total times used for varied-parameter cases. For example, for each system size in the range of 1x1x1-ball to 10x10x10-ball used to generate Figures 5.3 to 5.8, multiple SPICE simulations (each with different regulated-load wattage) can be simulated in parallel.
- ▶ **A list of key experimental themes and objectives:** An initial small-set of simulations in the range of 1x1x1-ball to 5x5x5-ball showing the first size violating the maximum connector-pin current allowed in the data-sheet, can be found in Section 5.2. A larger set of ten TCA cube sizes used for scalability predictions

can also be found in the same section. Moreover, Section 5.4 provides a whole view of total system power for the uniform power-allocation scheme.

- ▶ **Evaluation of power connection schemes:** (i) fully-connected external power connection and (ii) uniform power allocation test cases. These two schemes are used in Section 5.2.

Expectations and outcomes:

- ▶ **Parameterizable experimental cases:** As can be seen in Section 5.2, the values of pin resistance, regulated-load wattage, and external voltage supplied, are parameterised in the simulation framework.
- ▶ **A fundamental set of simulations, useful as baselines for simulation limitations such as the estimated simulation time for given system sizes:** As will be mentioned in Section 5.2, only the range of 1x1x1-ball to 10x10x10-ball systems are simulated due to long simulation times for larger sizes.

Success Criteria:

The cases of experiments should discover meaningful simulation results such as trends, upper/lower bounds, or other factors that affect the limit of scalability. All the experiments should be carried out in reasonable periods of simulation time.

Some fundamental power schemes may be used to evaluate a system when node power budgets are equally distributed. Some practical designs such as homogeneous node implementation are suitable with this power scheme. During investigations of models and simulation tools, the planned configurations of simulation cases may change over time. However, the initial plan was as briefly summarised below:

- ▶ **Regulated load power:** Coverage of the range from 0W node loading up to the maximum power capacity. If possible, this should be expected for simulations of the switching regulator model employed.

This is important for scalability evaluations, as a non-uniformly powered TCA system may gain some beneficial power-related results compared to a system

where all nodes have uniform power demand. Also, as mentioned earlier, the regulated load power should also be designed as a parameter within the simulation.

- ▶ **Inter-node electrical resistance:** There can be various methods of implementing inter-node power delivery, ranging from simple connectors embedding metal pins, to an in-house complex plate design for high-current capacity. In this thesis, the model is expected to be simple, but nonetheless the simulator can easily be reconfigured to represent the resistance behaviours of a number of inter-node power connection media.
- ▶ **External power-supply:** The voltage level of the external power source should reflect practical power supply units available for powering the system. This will ensure that the simulation model can be validated against hardware prototypes. By default a 12 volt DC supply rail is assumed, equivalent to a standard PSU module.
- ▶ **System sizes:** A range of system sizes should be chosen to be large enough to produce a meaningful quantity of simulation cases to observe the trends of scalability. This is also helpful to verify whether the tools are working correctly. Very large system sizes may also be possible for simulation-file creations, but can incur undesirably long simulation times, depending on the machine used for simulation. Therefore there are practical limits to system sizes simulated for the purposes of this research project.

5.1.2 Objective 4: Optimised Power Distribution

Methodologies/Activities:

- ▶ **SPICE simulation:** The GA-based simulations can be found in Subsections 5.3.2 and 5.3.3.
- ▶ **Parallel simulation:** Also, the individuals in a population of the simulations in Subsections 5.3.2 and 5.3.3 can be performed in parallel.
- ▶ **Visualisation:** Whilst the visualisations in Chapter 4 are for the simulation-framework development purposes, the visualisations in this chapter are for the

data extracted from GA-based simulation results, which can be found as 2D representations in Section 5.3.

- ▶ **Non-uniform power allocation:** Section 5.3 is dedicated for the proposed *relative position* scheme, which are used to perform the GA-based simulations.

Expectations and outcomes:

- ▶ **An optimisation framework for non-uniform node allocation:** All the simulations in Section 5.3 are performed using the GA-based optimisation framework developed.

Success Criteria:

The optimisation framework should demonstrate how uniform and non-uniform power allocation schemes can differently impact on the TCA scalability.

5.1.3 Objective 5: Scalability Evaluations

Methodologies/Activities:

- ▶ **Data analysis (interpretation/extrapolation, etc.):** For instance, extrapolations can be found in Section 5.2 for scalability, and interpolated GA simulation results are in Section 5.3. The interpretation of TCA-scalability comparing with traditional rack-mount systems can also be found in Section 5.6.
- ▶ **Tabulated, 2D, and 3D data visualised data:** Apart from various 2D visualisations and a 3D visualisation depicting a specific case in Figure 5.10, the conventional tabulated data such as Table 5.2 is also meaningful to detail simulation results given a series of varied input parameters.

Expectations and outcomes:

- ▶ **Both hardware measurement and simulation results in tabulated data, 2D and/or 3D visualisation:** The hardware-measurement results shown in Tables 3.3 to 3.5 have been previously discussed in Section 3.5, but are also

referred to in this chapter due to the measurements for validation are prior tasks as part of modelling towards the scalability evaluations in this chapter.

- ▶ **The discussions, interpretations, conclusions of simulation results:** These can be found throughout Sections 5.2 to 5.4.
- ▶ **All the simulations should be reproducible:** Referring back to the files generated in Figure 4.20(b), These are practical examples of simulation results which have been archived, along with the MATLAB[®] functions to generate the test cases to produce all of the evaluations in Sections 5.2 and 5.3 for future extensions and comparison.

Success Criteria:

From this thesis's perspective, the results in this objective should be able to test whether the research hypothesis is true or not. Also for future use, the results should also be meaningful and self-explanatory information for the decision making in the design-phase of practical systems, and help to establish future guidelines for next-stage research for variants of the TCA concepts and packaging designs.

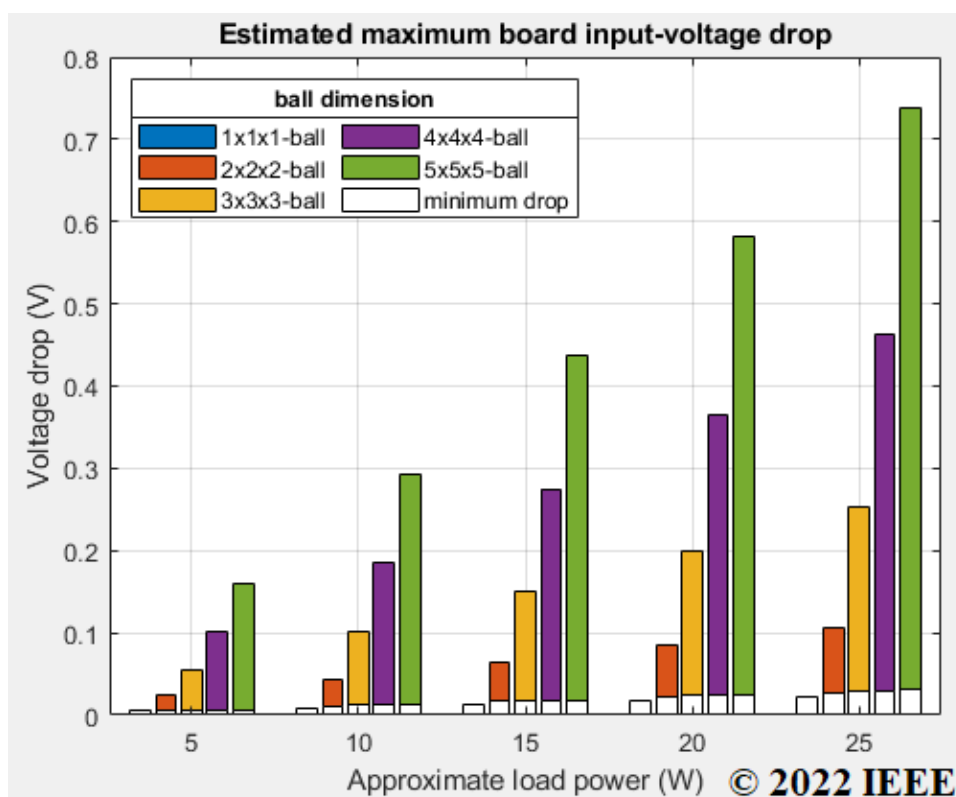
Given the relevant research objectives, the rest of this chapter will discuss all the details relevant.

5.2 Uniform Power Allocation

Uniform power-allocated TCAs can be considered as the simplest approach both in terms of practicality and simulation activities in this thesis. Prior to performing scalability evaluations, uniform simulations allow one to picture some basic behaviours of the two main constraints, voltage drop, and connector-pin current. This simulation configuration does not only investigate the constraints themselves, but also permits initial checks as to whether the models and simulation framework correctly perform for simple simulation cases.

To evaluate the scalability of cube-shaped TCA systems under the uniform power allocation using SPICE-modelled power-grid with the simplified board model, multiple

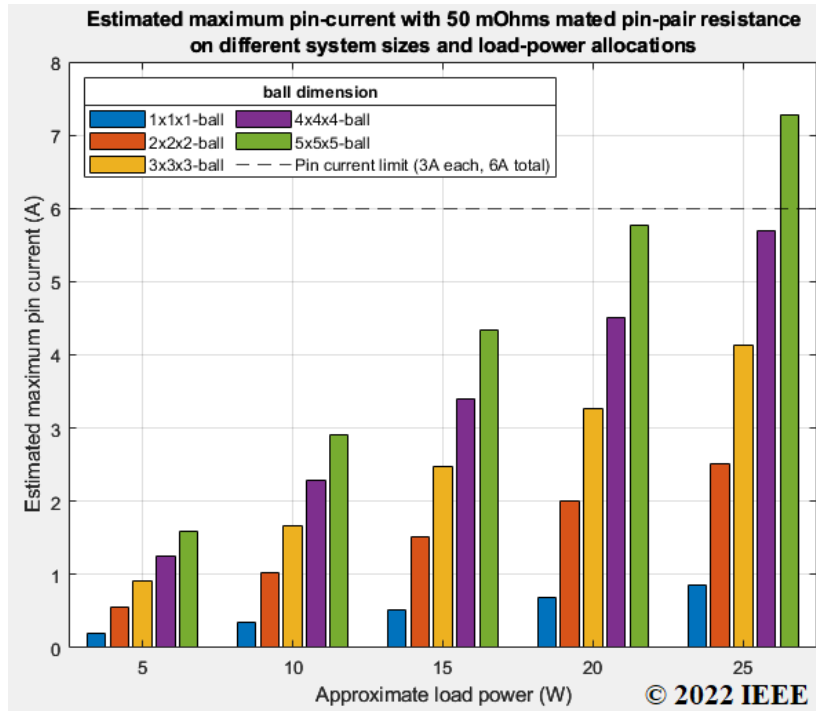
SPICE simulations are performed. Each of the simulations is assigned with a desired cubic size, e.g., 5x5x5-ball, and a 5V-regulated load wattage, for instance, 25W. After a SPICE simulation is complete, All of the board input-voltages and connector-pin currents from the simulation-result file can be extracted using the post-processing performed in MATLAB[®]. Afterwards, it searches for the worst-case (highest) voltage drop (difference from the system-surface voltage of 12V) occurring at all of the board input-voltages, and the worst-case (highest) connector-pin current all over the power-grid network. To give an example, each of these two worst-case values can be shown as a single bar in Figures 5.1 and 5.2, respectively.



(Regenerated from Figure 6(a) in [42])

Figure 5.1: Estimated best and worst-case voltage drop simulations for uniform power-allocation with 101mA assumed supply side 12V fan load.

Referring to Figure 5.1, it can be observed that this shows the trends of voltage drops for cubic TCAs as a function of cubic dimensions for $n = 1$ to 5, and arrays of $(n \times n \times n)$ ball configuration. Each power connector has a 50-mOhm mated pin-pair resistance as per maximum data-sheet specifications, and a 12V power-source is supplied to every system-surface connector. Two tile-edge pins are used in parallel for each of the power



(Regenerated from Figure 6(b) in [42])

Figure 5.2: Estimated maximum connector-pin currents for uniform power-allocation with 101mA assumed supply side 12V fan load.

and ground rails. The maximum size is limited at 5x5x5-ball due to this representing the maximum system size possible with specification of connector-pin current limit of 3A per pin in [88]. This can also be seen in Figure 5.2. The 5-step power values per node are also adequate to illustrate both the approximately-linear trends of voltage drops and pin-currents within this system-size range.

The approximate load power values in Figures 5.1 and 5.2 refer to the regulated power load for each tile. As a single ball is comprised of eight tiles, therefore, the total ball power values in these cases are, 40W, 80W, 120W, 160W, and 200W, respectively. Concerning these power budgets per ball volume, a cooling system is required for heat dissipation. The investigation of potential cooling systems is left as future work at this stage of the research. Regarding this initial set of simulations, it can be observed that with both the fully-connected and much-higher regulated-voltage externally supplied at 12V, the voltage drops occurring in this system-size range are not significant, ranging within under 1V. Whilst the worst-case connector-pin current of the 5x5x5-ball TCA with 25W-load regulated at 5V shown in Figure 5.2 begins to exceed the

standard prototype connector-pin current limit of 6A.

Having introduced two simple power-allocated simulation cases, with the symmetry of the cube shapes and systematically increasing the system sizes, the quantitative trends obtained from the simulations can be further extrapolated to predict the non-simulated estimated voltage drops and worst-case connector-pin currents of different load-wattages and system sizes. This approach appears to be less reliable at low wattages. Therefore to ensure that predictions are not unrealistic when extrapolating from a small set of smaller array sizes*, curve fitting is used with pessimistic and optimistic cases, to give a range rather than a definitive prediction.

Figures 5.3 and 5.4 show a set of worst connector-pin current predictions based on the hardware prototypes without the power-consumption profile of the cooling fan in Figure 3.1. The extrapolated cases of the sizes of 11x11x11-ball and larger ones are from another set of ten simulated cases of 1x1x1 to 10x10x10-ball systems and also with a finer watt-step of 1W performed using the simplified board model with the post-processing based adjuster. These predictions reflect the maximum cubic sizes allowed within the connector-pin current constraint under a desired specific regulated load power per tile. The predictions are based upon curve-fitting to trends established in the ten SPICE-model simulated cases, in order to estimate behaviour for higher order array sizes, with pessimistic and optimistic extrapolations representing lower and upper bounds respectively. These extrapolations are necessary since higher order array sizes require unreasonably long simulation times.

The five curve-fitting equations are given in Table 5.1. The curve-fitting process, relying upon process 1, as shown in Figure 5.5 extracts the global scaling trend from the voltage drop or pin current cases respectively. From this a local data behaviour for any given size ' n ' can be interpolated for power loads up to 25W, or extrapolated for theoretical power loads above 25W. The data can then be subject to a further curve fitting (process 2) to allow local scaling at size n to be represented for any power load. The two described processes will be used to generate all the predictions in Figures 5.3, 5.4, 5.6, 5.7, and 5.8.

* i.e. due to simulation times, only arrays of 1x1x1 to 10x10x10 were simulated.

Table 5.1: Two-step extrapolation processes and curve fitting formulae used in Figures 5.3 to 5.8.

Plot Legend (model)	Formula for Process 1 (global trend)	Formula for Process 2 (local trend)
poly2 , poly1	$p_1x^2 + p_2x + p_3$ [89]	$p_1x + p_2$ [89]
poly2 , poly2		$p_1x^2 + p_2x + p_3$ [89]
poly2 , poly3		$p_1x^3 + p_2x^2 + p_3x + p_4$ [89]
poly2 , power2		$ax^b + c$ [90]
poly2 , exp2		$ae^{bx} + ce^{dx}$ [91]

In the legends of these scalability prediction graphs, they show two fitting-equation models used to predict each sub-result. For example, *poly2 , power2* means that polynomial of degree 2 is used for the first process of approximate *regulated load-power per-tile vs voltage drop (or pin current)* fittings, whilst the power model with 2 numbers of terms, fits the second process for the relation of *cubic size vs voltage drop (or pin current)*.

In Figure 5.3, it shows the prediction where each tile is allocated 1W regulated load. Interestingly, with the different fitting models and their degrees of polynomial, power, and exponential, the most unrealistic model seems to be the exponential one, approaching the connector-pin current limit of 6A at the size of 27x27x27 balls. The other three models, hit this current limit at approximate sizes of 60x60x60, 80x80x80, and 90x90x90 balls. Thus, these four models predict the scalability (from highly pessimistic to highly optimistic) of the total nodes ranging from 157,464 to 5,832,000 tiles. The best case in this set of simulation results is tremendously large and is far beyond the expected large-scale of the degree of thousands or tens of thousands of nodes in this thesis. Indeed, even the pessimistic projection is an order of ten or twenty larger than that baseline.

The next prediction cases in Figure 5.4 repeat the same extrapolation from data with tile power values ranging from 5W to 25W regulated power per tile. It is noticeable that the divergence of the curve fitting models is much closer together, meaning that the scalability can be predicted with a much smaller degree of uncertainty.

The 5W allocation case allows a TCA system based on the hardware prototypes can scale up approximately to the sizes of 19x19x19 (54,872 tiles) to 23x23x23 balls

(97,336 tiles), according to the lower and upper-bounds of the fitting models.

Considering the cases from 10W towards 25W, all the fitting models closely agree together on predicted results as the maximum sizes shown in the figures are close to the the data set obtained from SPICE simulations. To suggest some applicability, the case of 15W can be assumed that this moderate wattage budget can allocate a small computing board containing a Raspberry Pi [92] chip, considering the maximum power specification of an official power supply [93]. Finally, the maximum regulated load power per tile allowed by the specific of the regulator is shown in Figure 5.4(e).

Interestingly, observing this approximately linear trend and with a given current limit, the maximum size allowed can be rapidly estimated. These can be obviously seen by the decreasing trend of the maximum sizes shown from Figure 5.4 (a)-(e).

The extrapolated scales aforementioned only focus on the constraint of connector-pin current. However, the other important constraint, *voltage drop*, should not also be underestimated for large-scale systems. Whilst the current constraint holds approximate linear trends in most of the fitting models when scaling in lower ranges observed, this is not the case for the voltage drop constraint as shown in Figures 5.6 and 5.7 which explicitly show non-linear trends. In a real system, a regulated load can consume any arbitrary wattage up to the maximum power consumption specified in the data-sheet. However, only six quantities of regulated power are selected to demonstrate their connector-pin current limits and voltage-drop trends. Each simulation set is given with different curve fitting models.

Given both the predictions of voltage and current constraints, Figure 5.8 shows an example of combining the two constraints for considering the mutual impacts on the TCA scalability for 1W regulated load. In this particular scaling consideration, it can be obviously seen that the voltage-drop issue is the primarily limiting factor, forcing the scalability approximately at 40 balls per dimension. Next, including the current constraint, it predicts that the maximum size is approximate at 60 balls per dimension. It can be concluded that, even though the ten simulated simulations (1x1x1 to 10x10x10 cubic sizes) are the data set used to generate the predictions in Figures 5.3, 5.4, 5.6, 5.7, and 5.8, the non-linearity existing makes the fitting models not agree

to each other when the system sizes are growing to larger scales. These variations of scaling confidence implies that 1) the simulated data set may not be adequate for low wattage node cases, or 2) additional further parameter analyses for narrowing down the most suitable fitting models are required.

Whilst all of the predicted results seem to satisfy the scaling expectations anticipated, the awareness of fitting-model selections is left for deeper investigations in future work. The next section will discuss another power scheme, non-uniform allocation, which is also feasible in practical systems and possesses some attractive characteristics.

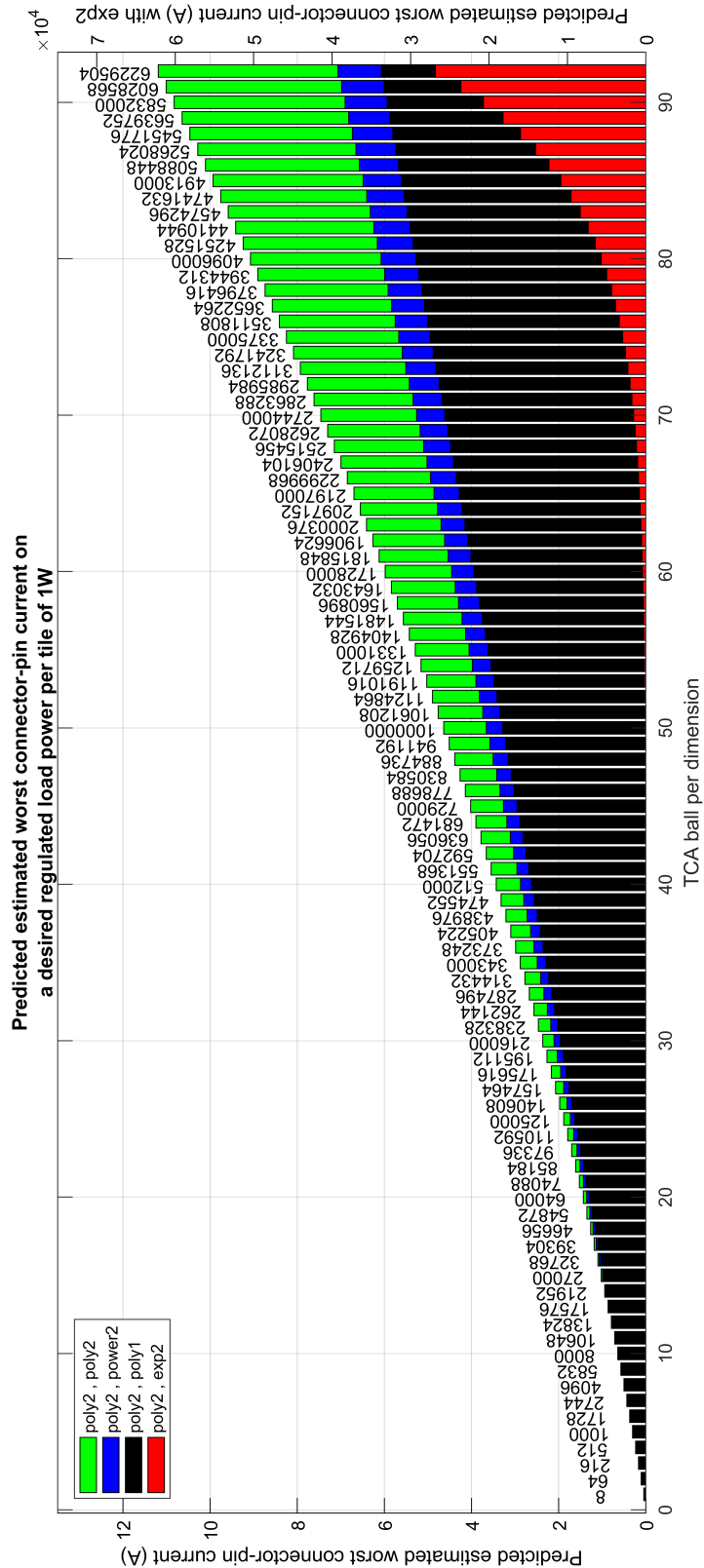


Figure 5.3: Predicted estimated worst connector-pin currents for 1W regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. To estimate the feasibility of a TCA system, this connector-pin currents report can be used together with the worst-case voltage drops report as shown in Figure 5.6, which represents the difference between the supply voltage at all of the surface power connectors of the grid array (12V in this case) and the worst-case voltage drop (loss) of all of the board input voltages. A combined plot can also be seen in Figure 5.8.

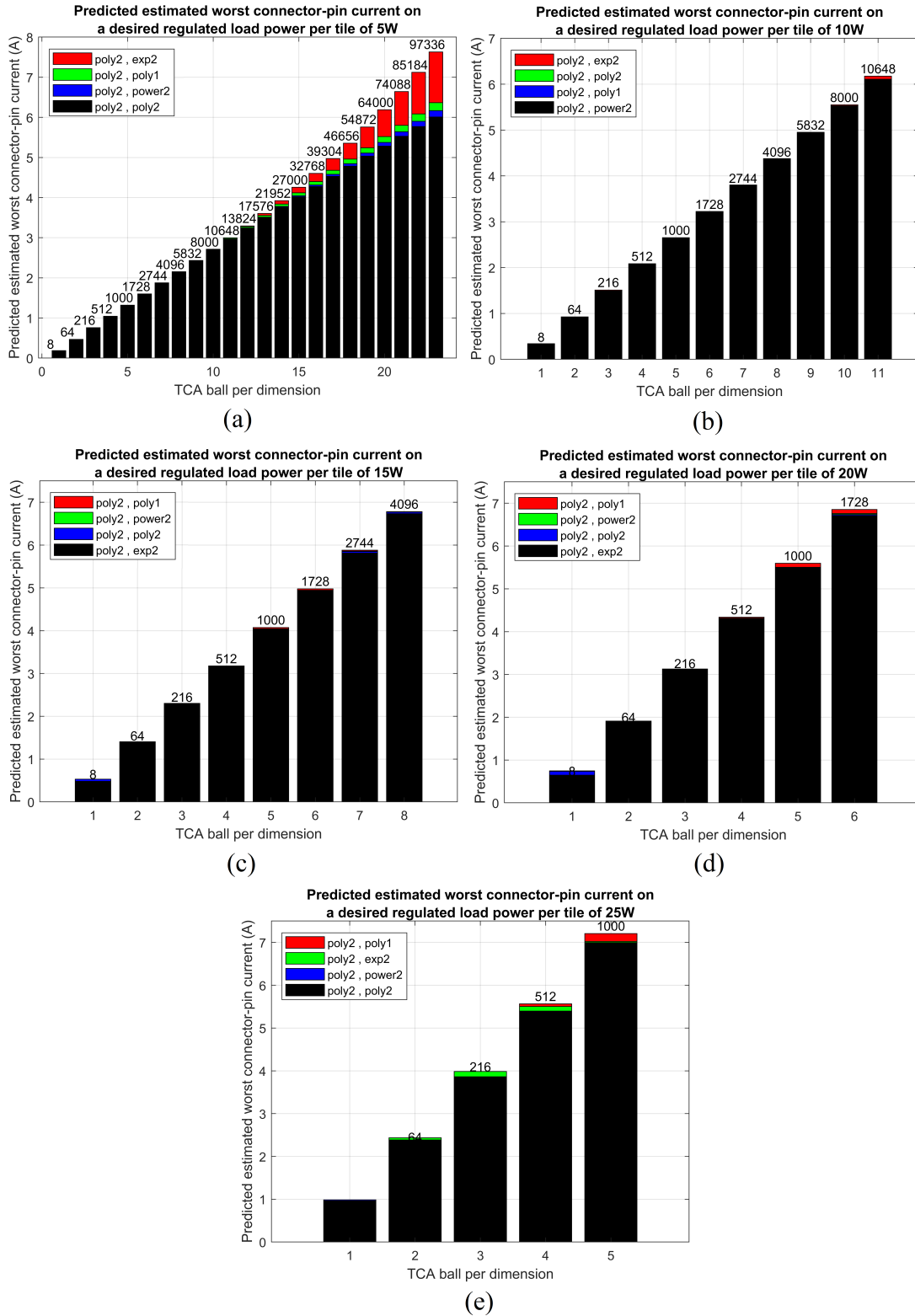


Figure 5.4: Predicted estimated worst connector-pin currents for 5W (a) to 25w (e) regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. The sizes in each case are limited when the the worst connector-pin currents exceed the current limit of 6A.

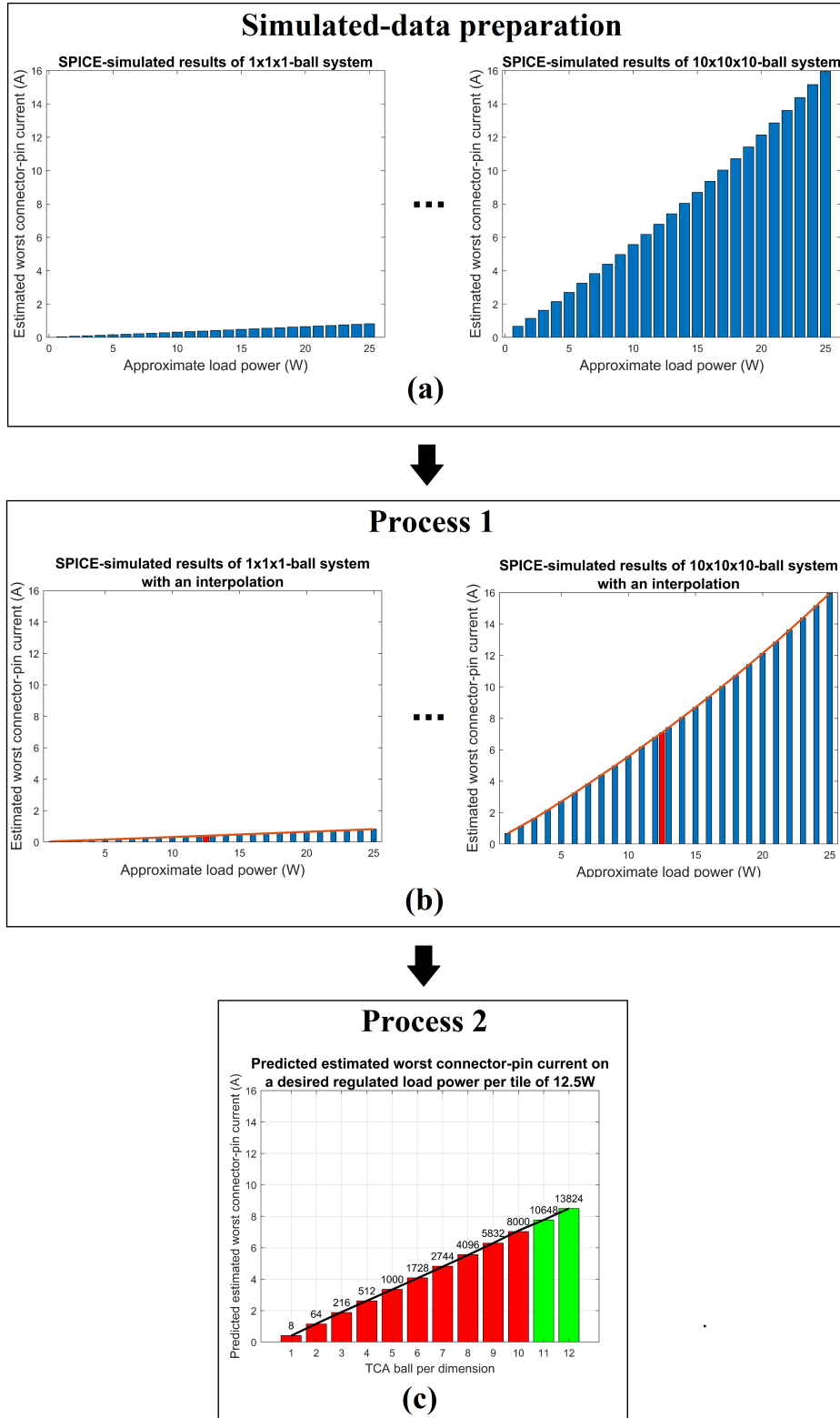


Figure 5.5: Detailed processes for predicting arbitrary regulated load wattages and ball cubic-size systems. (a) shows an example of simulated-data preparation for a desired regulated load wattage of 12.5W. The sizes of 2x2x2 to 9x9x9 are omitted. (b) details the process 1 generating 10 interpolated worst connector-pin currents. (c) shows process 2, with an example of extrapolation of a couple of desired cubic sizes of 11x11x11 and 12x12x12-ball systems (shown in green) from the data generated by process 1 (shown in red).

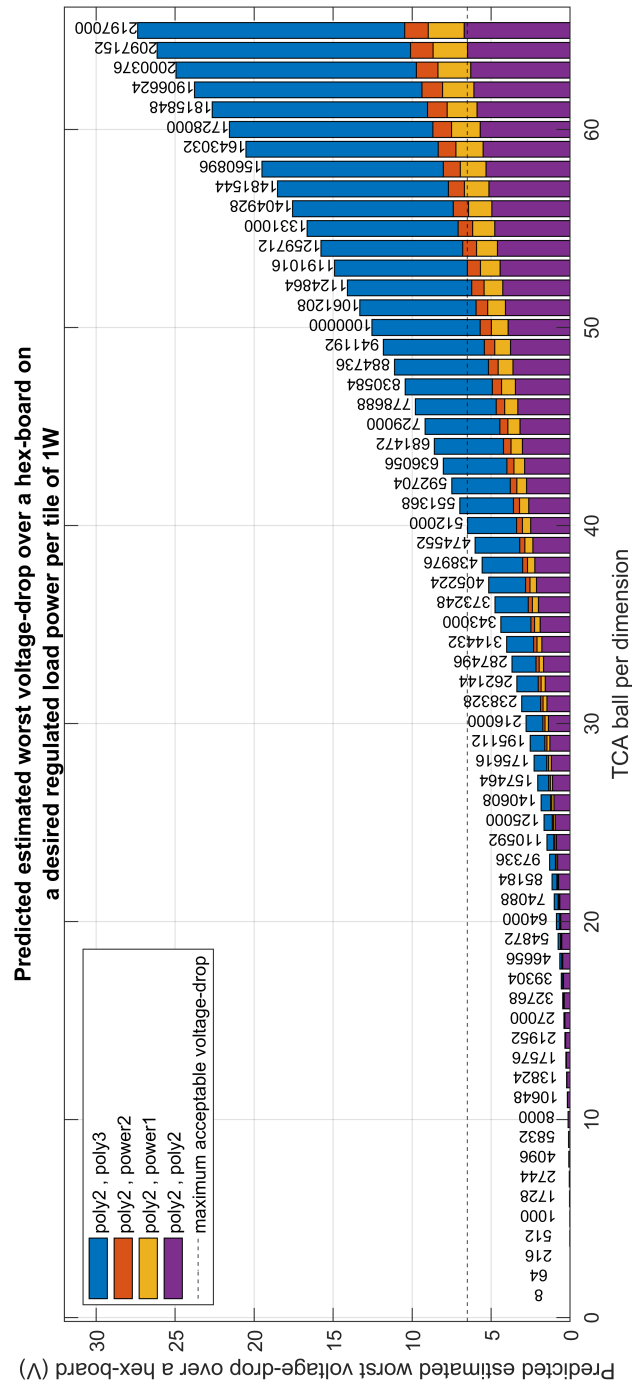


Figure 5.6: Predicted worst-case voltage drops for 1W regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. To estimate the feasibility of a TCA system, this voltage drops report can be used together with the worst-case connector-pin currents report as shown in Figure 5.3, which represents the highest current experienced from all of the lumped resistors modelling single (or parallel) tile-edge power (or ground) pins. A combined plot can also be seen in Figure 5.8.

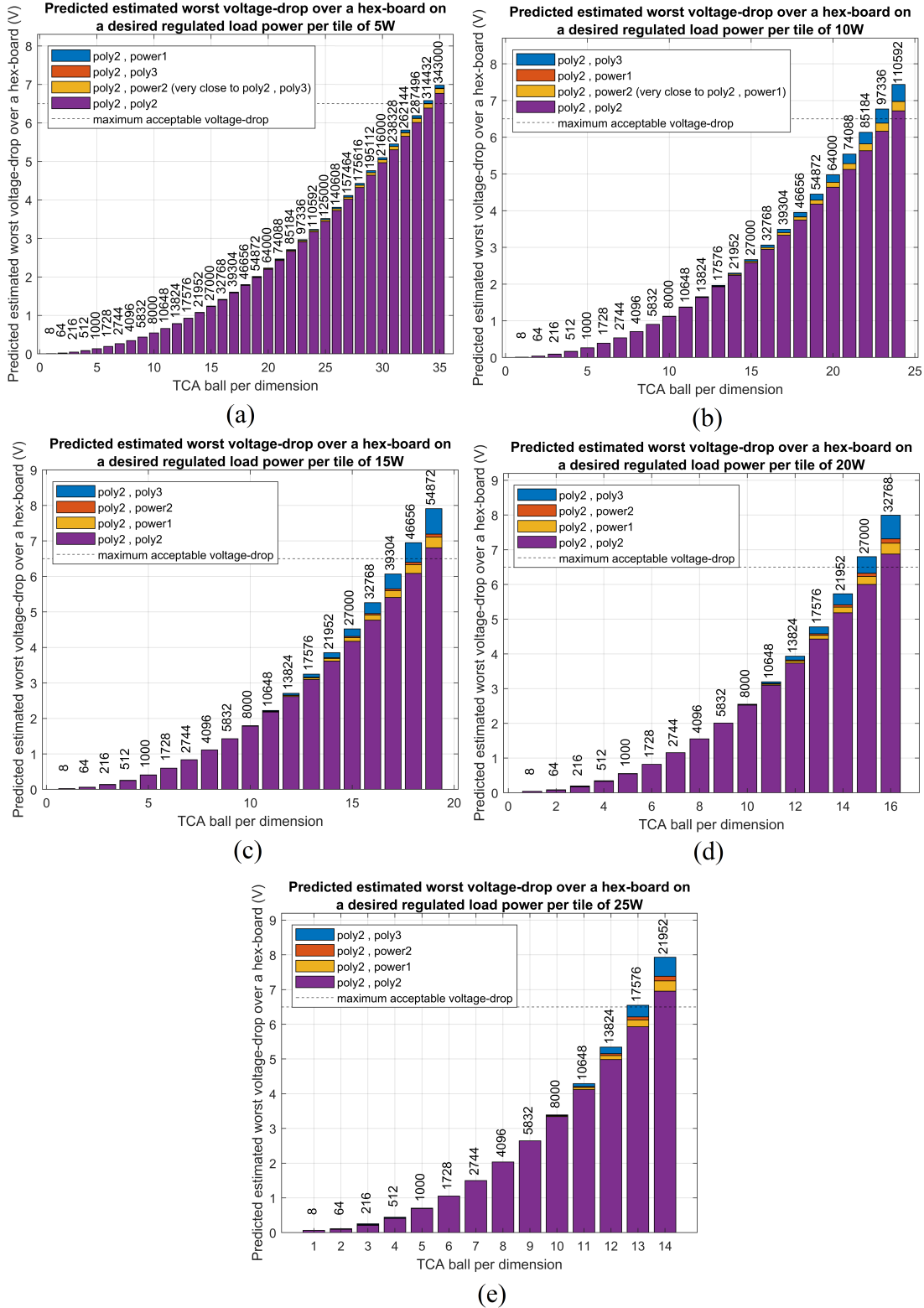


Figure 5.7: Predicted estimated worst voltage drops for 5W (a) to 25w (e) regulated load per tile on various cubic sizes. The number above each bar represents the total number of tiles in each system. The sizes in each case are limited when the the worst voltage drop exceeds the voltage drop limit of 6.5V (due to the external voltage supplied of 12V and the minimum input voltage of 5.5V specified in the regulator data-sheet).

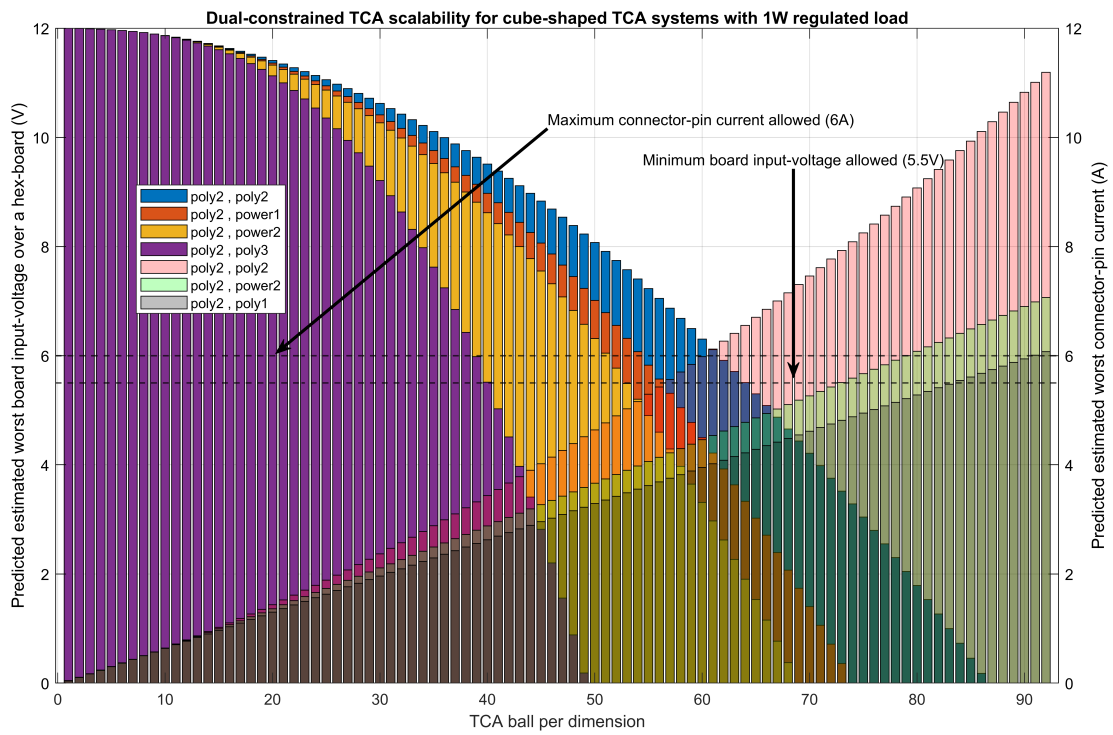


Figure 5.8: Dual-constrained TCA scalability for cube-shaped TCA systems with 1W regulated load.

5.3 Non-uniform Power Allocation

There can be many possible non-uniform allocation schemes as discussed in Subsection 4.7.2. *Relative position*, a scheme of coordinate-aware allocation proposed in this thesis, is mainly focused in the non-uniform allocation category for system evaluation, as it reduces simulation cases and is also a possible practical non-uniform type of scheme implementable in real hardware. TCA developers may need to employ this scheme when constructing physical heterogeneous systems, even including homogeneous ones that allow multiple continuous or discrete amounts of power per node.

Concerning the relative-position allocation scheme: even though the purpose is to reduce simulation cases when seeking optimal solutions where the external power sources are systematically supplied (for instance, full connection at all surfaces), the increasing of system size for simulations still adversely impacts on simulation times, making brute-force approaches impractical. Thus, a *Genetic Algorithm* (GA) is employed in this thesis to mitigate this issue.

In the following subsections, an example of *Brute-force* simulation is discussed, followed by equivalent GA simulations, which it will be noted, will significantly speed up the optimisation in the simulation framework.

5.3.1 Brute-force Simulation for Relative-position Scheme

To demonstrate the characteristic of power loading and worst-case connector-pin currents in a TCA system, in Figure 5.9, a brute-force simulation set of 64 tiles arranged into a 2x2x2 cubic ball array TCA shows the relation of system-level regulated load-power allocations and the estimated worst-case connector-pin currents. In this particular case, the relative-position scheme and 5-step regulated power allocation per tile is applied, with power in the range {5w,10w,15w,20w,25w}, starting at 5W an stepping upward finally to 25W.

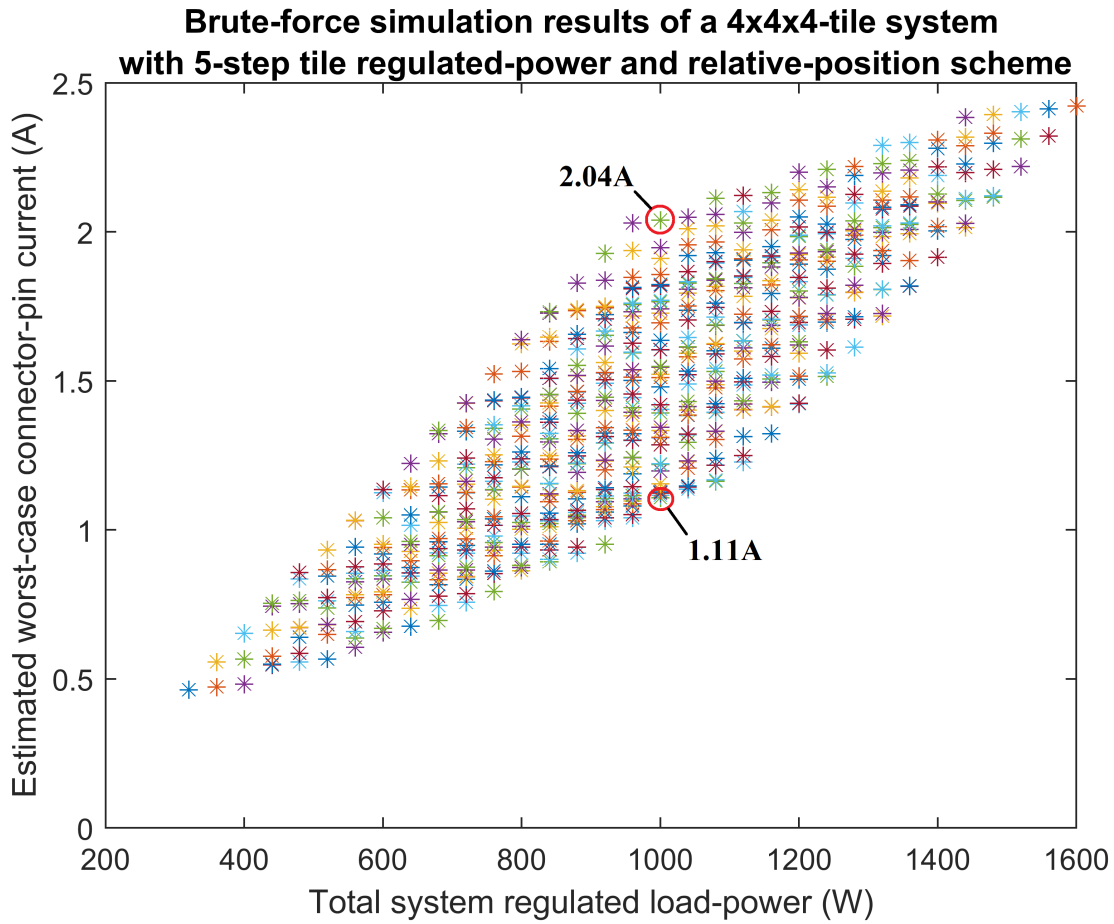


Figure 5.9: Example of brute-force simulation for the relation of worst-case connector-pin currents and system-level regulated load-power.

The leftmost mark in Figure 5.9 represents the allocation where each node-level regulated load-power operates at 5W, making up 320W for all the 64 tiles. Whilst the rightmost one is 25W, totally consuming 1600W for the whole-system regulated power.

It is observable in Figure 5.9 that for each of the two-extreme power cases of 320W and 1600W, there can only be a single possibility of allocation. The lowest case is 5W per node, and 25W for the highest extreme. In fact, these two cases including some other allocation instances in the figure are uniformly allocated due to being part of all possible cases.

However, the other cases, being non-uniformly allocated, can produce the same system-level regulated power as shown in the vertical marks for a single desired power case in the figure. With the whole picture of this characteristic pattern within stepped-

power allocations, given a desired system-level regulated power, the best allocations for the lowest worst-case connector-pin current can always be achieved, with enough computational effort.

Even though Figure 5.9 provides some insightful relational pattern, this specific simulation set is only performed with 5-step power. With this non-continuous-value power allocation, the actual exhaustive distribution may not have the same appearance. The shape of the pattern can be improved by adjusting the stepped power to a finer value of step, but with the penalty of the increasing of simulation cases, which consequently incurs longer simulation times.

Regarding this issue of stepped power granularity, including the difficulties encountered by large-number simulation instances from exhaustively searching in the brute-force method, an alternative has been investigated. A set of GA simulation modules supporting continuous power-values are therefore considered next, as described in Chapter 4.

The objective of the GA sub-framework is not to produce all the exhaustive cases of allocation-patterns, but to assist the simulation framework to find optimal solutions for large-scale TCA-system evaluations with much lower effort*

The following GA simulation results in this chapter have been performed using MATLAB® *gamultiobj* function [85] to find the pareto front of the two-objective fitness functions in each of the simulation cases.

The initial population in each case also includes an individual of the minimum power allocation for each tile of 1W, and another one with the maximum power of 25W. All the GA parameters are set default values given by the MATLAB® function at this stage of the research as the main focus is to heavily build the internal mechanism of the simulation framework instead of dedicating the whole effort for adjusting the GA parameters themselves. However, in future work, the investigation of the GA parameters, or some other optimisation algorithms such as *particle swarm optimisation* [94] can also be attractive areas to improve this specific part of optimisation framework.

* Though it should be noted that an optimal solution can take indeterminate time, the advantage of GA is that a near-optimal solution will generally be found in a reasonable timescale.

To illustrate an alternative meaningful representation of the power and constraints above, in Figure 5.10, two different power allocation results are illustrated for the same total power of 1000W regulated load-power. The visualisation gives more meaningfully geometrical results, showing that there can be several allocation cases (also shown as vertical asterisks in Figure 5.9) that provide the same power required. Other factors may then allow the most appropriate choice to be made.

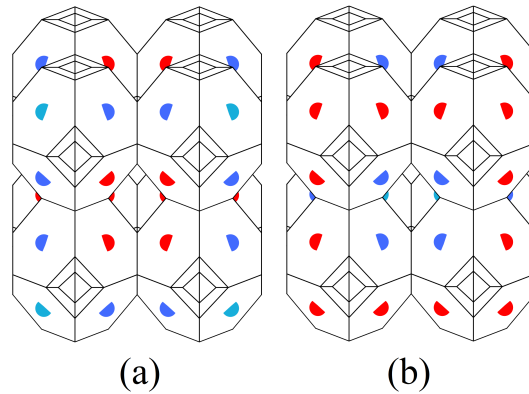


Figure 5.10: Visualisation of a 64-tile TCA comparing the same system-level regulated 1000W load-power but different node-level power allocations. The red dots represent highest-wattage nodes, and the lower ones are highlighted in blue. (a) shows the worst case, whilst (b) is the best one obtained shown in Figure 5.9.

For example: even though both the cases comply with the connector-pin current constraint, maximum voltage drop allowance, and the required total regulated load-power, the best and the worst allocations of the 1000W cases shown in Figure 5.9 can be visualised in Figure 5.10, where (a) the worst case requires a maximum of 2.04A at the most heavily loaded pin in the grid, whilst (b) requires only a maximum connector-pin current of 1.11A, an improvement of over 46%.

Therefore the best case not only gives a lower maximum connector-pin current, but also allocates higher wattage nodes at the surfaces of the system, allowing better cooling for high heat-dissipating nodes. This consideration may seem to be trivial for small system sizes but might be significantly beneficial in terms of power and heat management in large-scale systems, in which multiple layers of nodes exist.

For a large system size, a brute-force simulation will require excessive simulation time. Theoretically, if some small consecutive ball system-sizes, e.g., 2x2x2, 3x3x3, 4x4x4,

can be exhaustively simulated, the characteristics shown in Figure 5.9 for larger sizes, e.g., 5x5x5, can be estimated using some mathematical methods such as extrapolation from smaller-size results (e.g., 1x1x1 to 4x4x4).

However, even for a system with a small ball-number per dimension such as 4x4x4-ball (64 balls, 512 tile-nodes), with five-step load power, it takes 5^{512} cases, which is not feasible for a low-end simulation facility for this large set of simulations. Moreover, even though the relative-position scheme dramatically reduces the options for allocating power for a system size of 4x4x4-ball to 20 combinations, as shown in Figure 5.11, it still results in a tremendously large number of 5^{20} cases. A solution for this issue is to use an optimisation technique such as genetic algorithm.

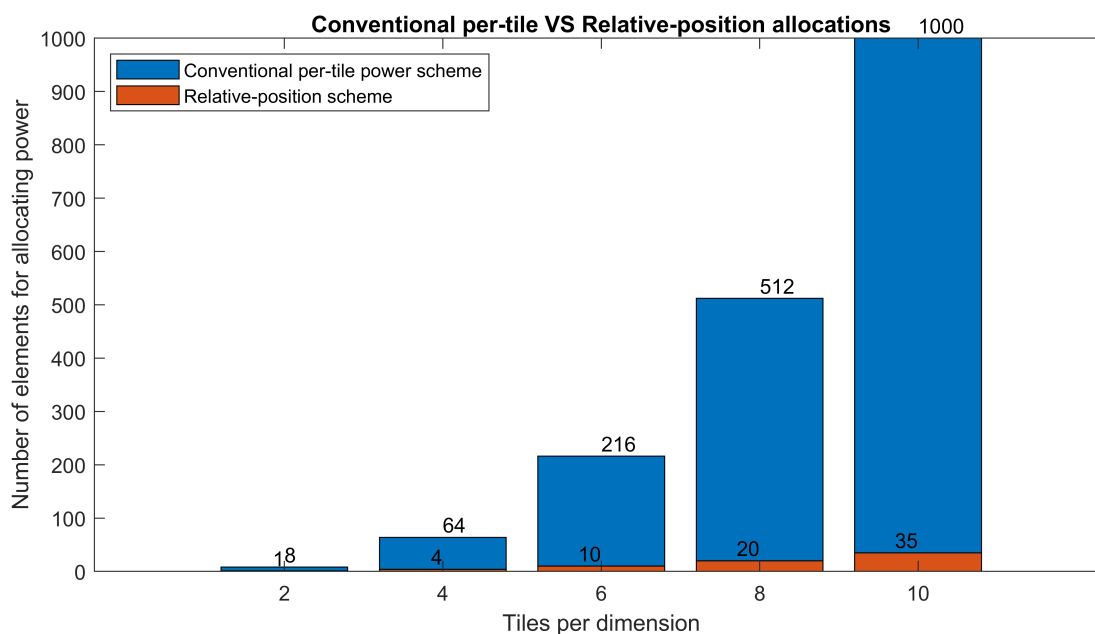


Figure 5.11: Conventional per-tile vs Relative-position allocations.

The GA simulation framework has been discussed earlier in Subsection 4.7.4. The following subsection will discuss the simulation results for scalability evaluations, providing meaningful results for constructing TCA systems, both power-vs-constraint and power efficiency simulation results.

5.3.2 System-level Regulated Load-power and Connector-pin Current Optimisations

The system sizes for evaluations should reflect the feasibility of large-scale TCA systems. In this thesis, the sizes investigated are from 1x1x1 to 6x6x6 ball-array systems. It is limited as this largest size due to suffering from a very long simulating time, taking approximately 13 days for the GA to stop given the criteria set. However, a 6x6x6-ball system is equal to 216 balls (1,728 tiles), which is considered adequate for a large-scale investigation.

All of the GA simulation results in this section are compared with their uniform-allocation counterparts. It can be seen that the 1x1x1-ball in Figure 5.12 case has no advantage from the GA optimisation as only a single ball resides in the system. Thus, each of the eight tiles are equally adjusted using the relative-position scheme. On the other hand, cube-shaped sizes starting from 2x2x2-ball, as in Figures 5.13 to 5.17, possess ball layers. In these layered systems, there are multiple nodes differently affected by node-level voltage drops, directions and quantities of connector-pin currents all over the system. The size of 2x2x2-ball in Figure 5.13 shows an exceptional case with some negative improvements. This suggests that the system size may not be large enough for the GA to discover the expected trend of the optimised results. This outlier case is expected to be investigated further in future work.

Starting from the size of 3x3x3-ball, all of the evaluated outcomes manifest similar resulting trends, gaining positive improvements in all of the range of system-level regulated load-power values. Therefore one can conclude that, at the design time, non-uniform allocation plays an important role when low connector-pin current constraints are given to achieve a desired system-level regulated power.

On the other hand, for the system's run-time, the non-uniform allocation can be also employed to maintain the design-time, or a possible new desired regulated power within the connector-pin current constraints by *re-optimising* when some of the pins encounter failure situations.

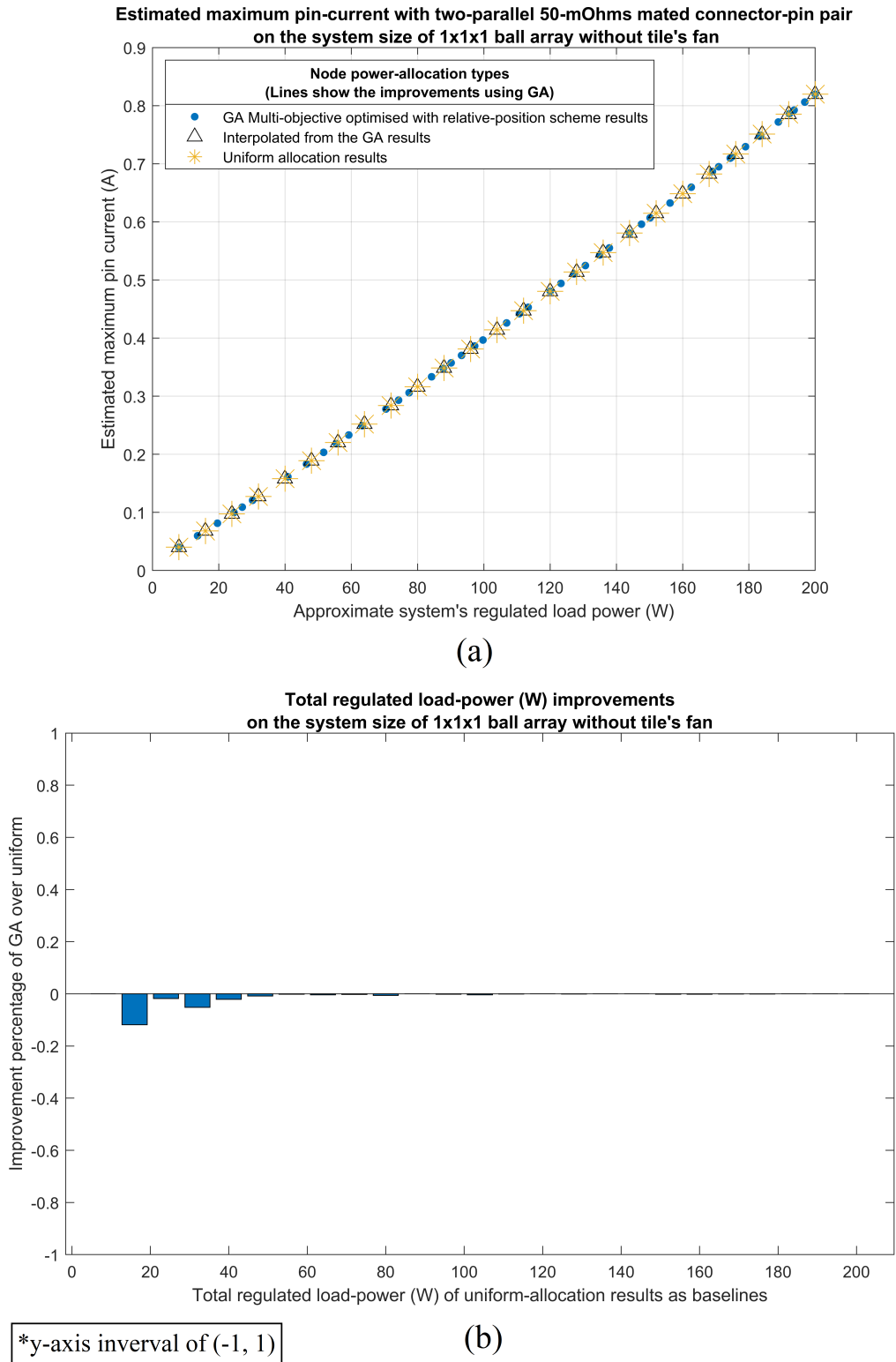
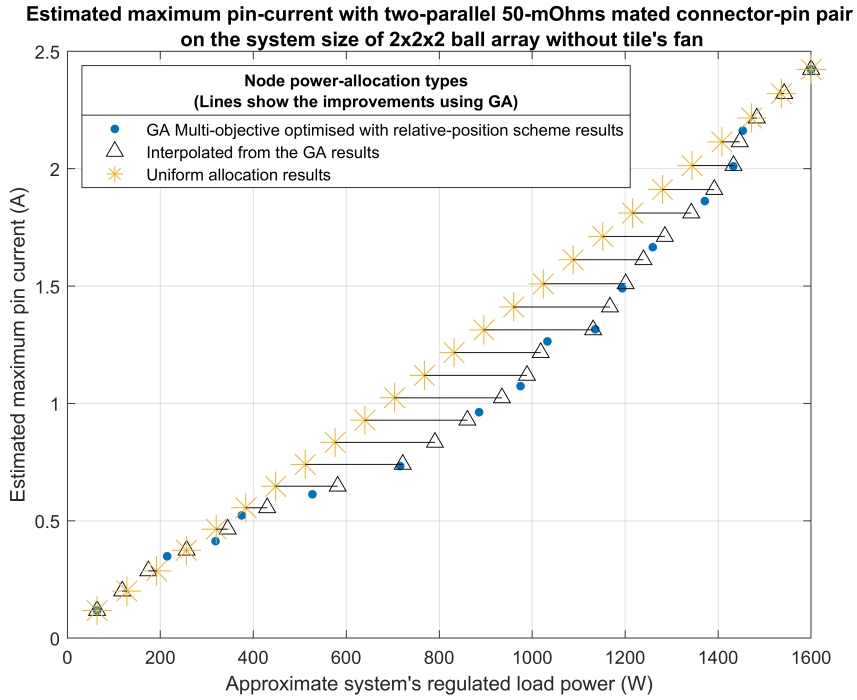
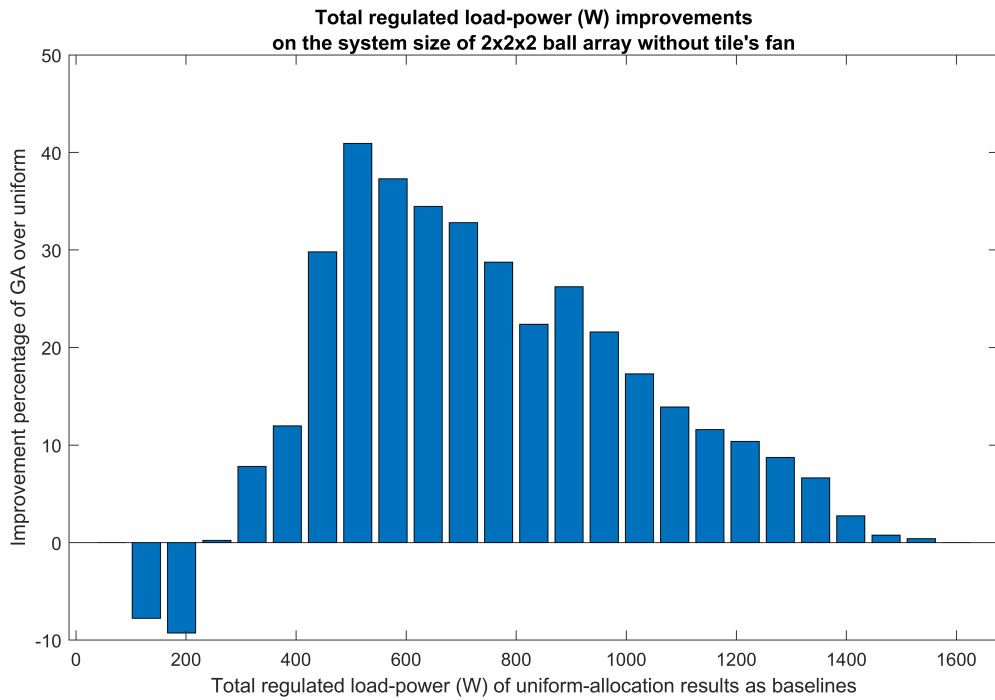


Figure 5.12: (a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 1x1x1-ball system.



(a)



(b)

Figure 5.13: (a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 2x2x2-ball system.

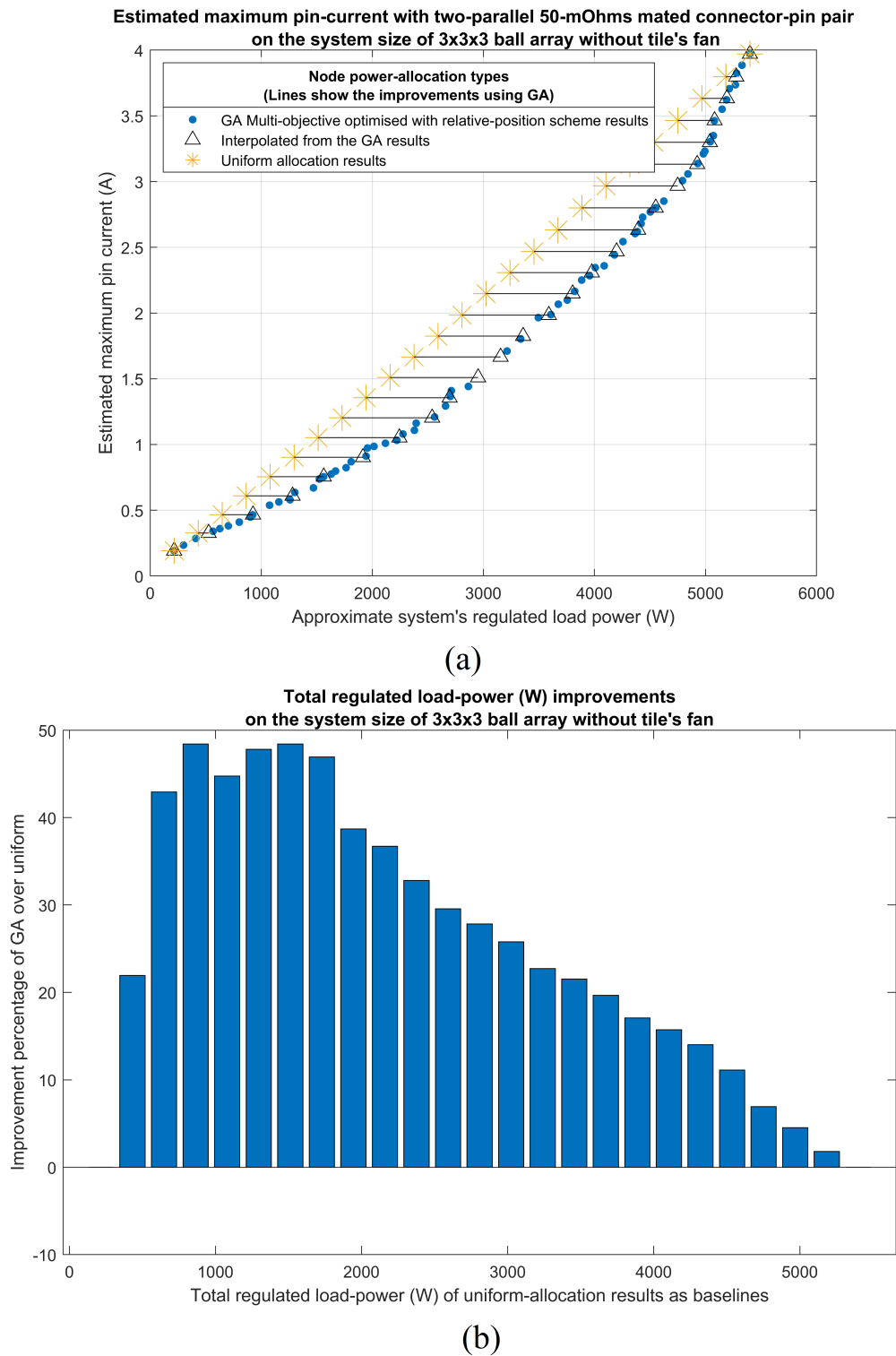


Figure 5.14: (a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 3x3x3-ball system.

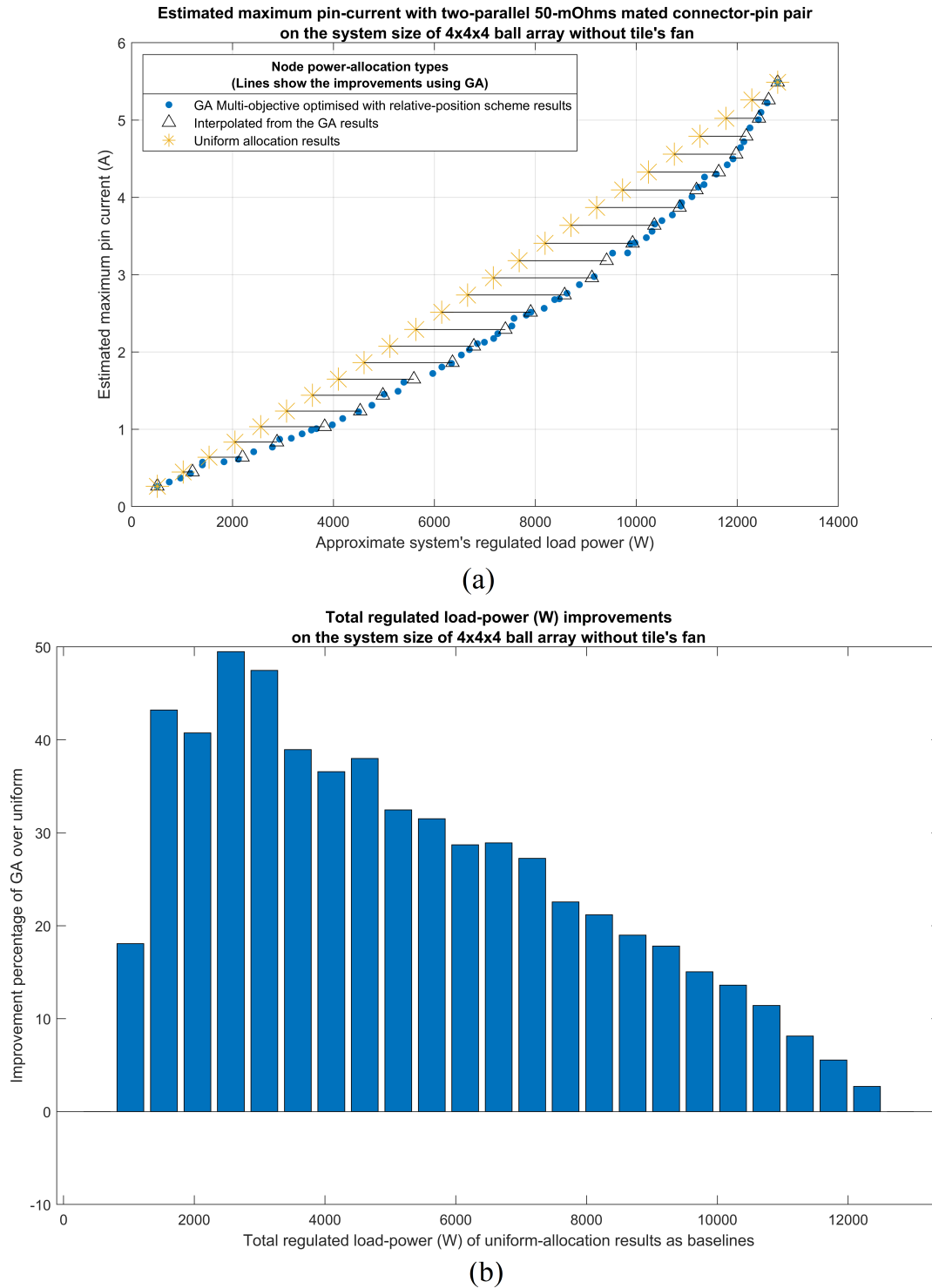


Figure 5.15: (a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 4x4x4-ball system.

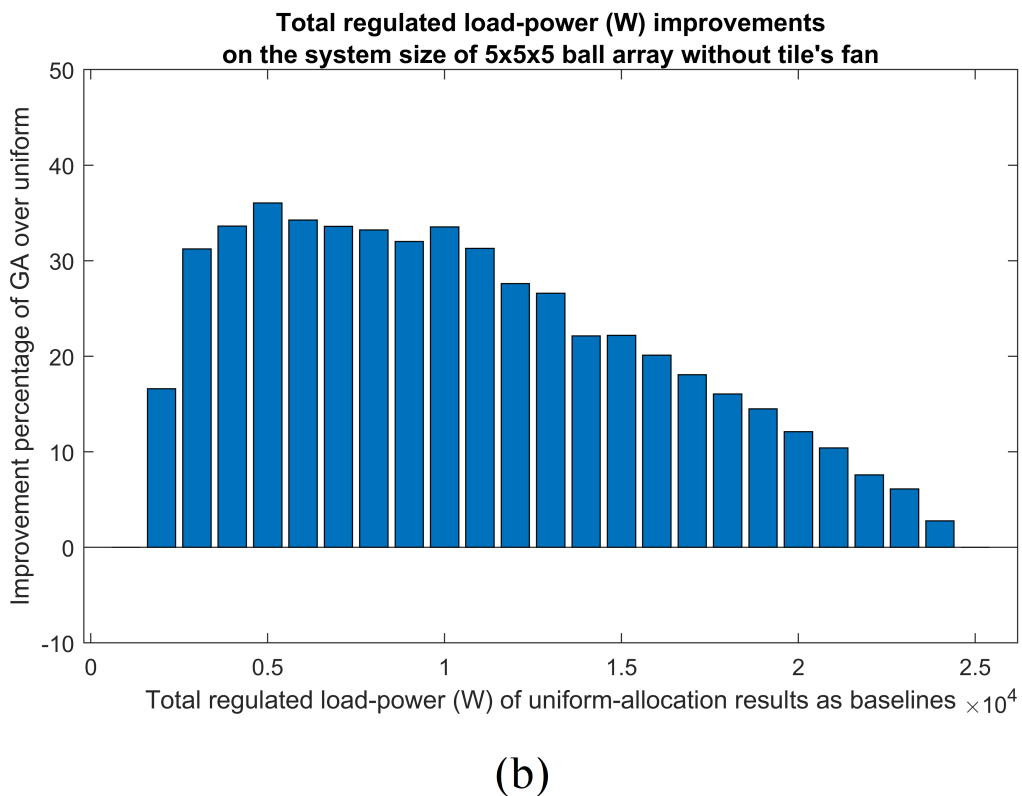
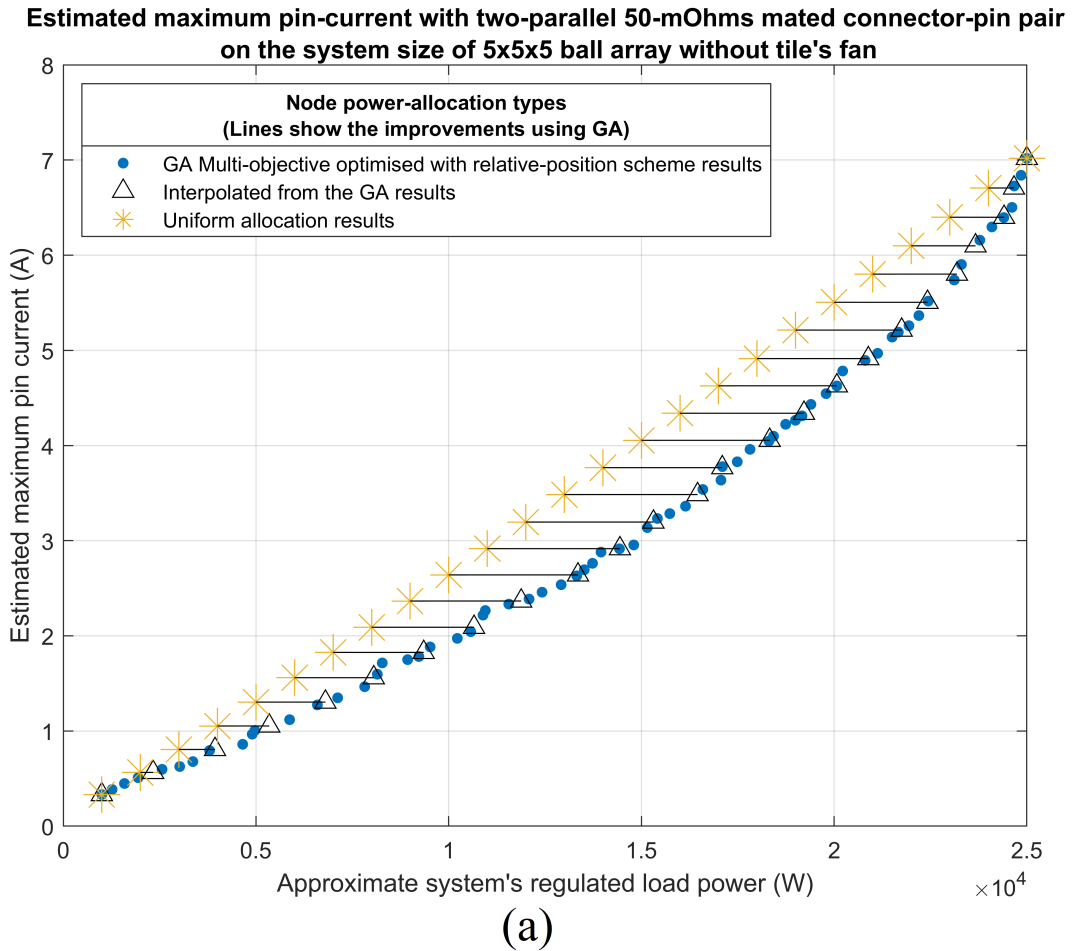


Figure 5.16: (a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 5x5x5-ball system.

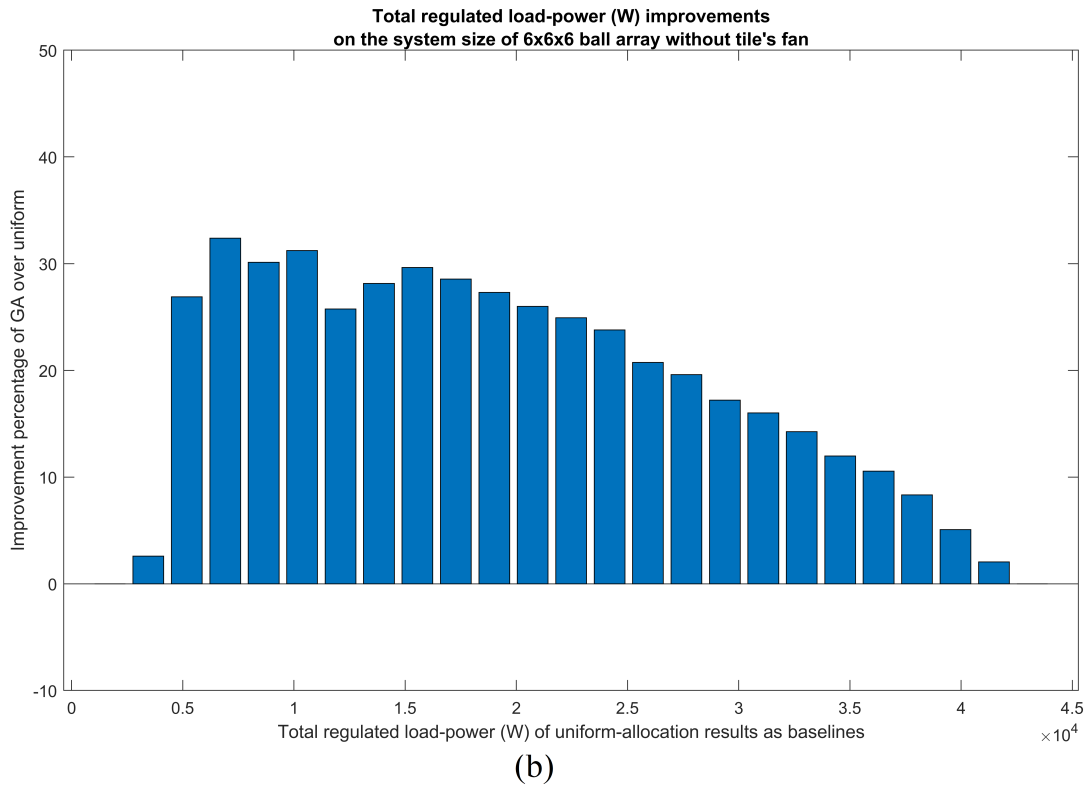
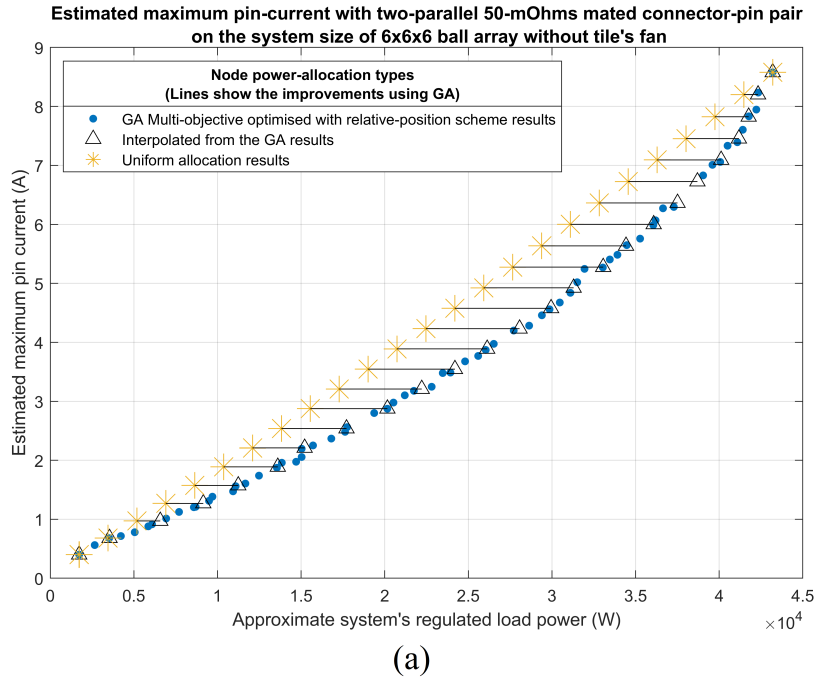


Figure 5.17: (a) uniform and GA-optimised power allocation, (b) improvement percentages, of a 6x6x6-ball system.

From the point of view of computing performance, achieving more system power budget can imply more computing power. In compute-bound workloads, this means that nodes can operate faster at higher clock-frequencies or the number of computing elements such as cores can be more active. This improved power budget may also improve I/O-bound workloads if the power consumption of non-static networking speed is quite sensitive to node-level power budget. However, relating power and computing performance is considered complicated. A single node may not only contain general-purpose PEs such as CPUs, and there are static and dynamic elements within a typical computational component power utilisation.

In addition, given a single PE-type node, modelling the relation of a submitted computing task into a node and its power allocated may also be even more complicated as the computing load may vary its instantaneous power consumption. In any case, the assumption of constant node-power allocation can be applicable for adjusting all the node-level computing units to maintain the total node power being consumed. For instance, FPGA units may be designed to be capable of dynamically adjustable clock speed, whilst the CPUs employed require some form of CPU-frequency programmability. A conceptual illustration for dynamic power adjustment is shown in Figure 5.18.

Concerning the allocation improvements found by GA, these may not be always the best solutions for some node-level power requirements. For instance, practical implementation may consist of a minimum requirement of 10W in each single node. In this particular case, one of the discovered solutions by GA can be sought to obtain the best possible case to satisfy this constraint. Alternatively, the simulation framework can be further modified to limit this lower-bound of node-level power required.

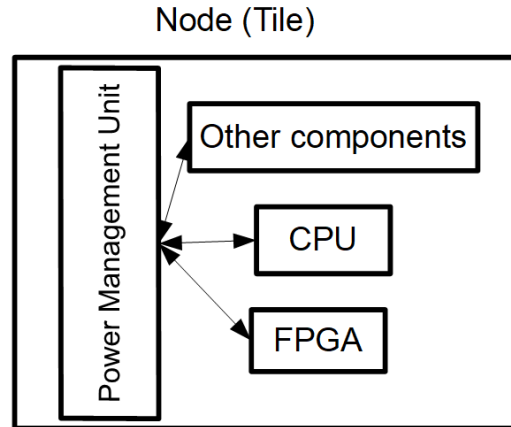


Figure 5.18: Example of a conceptual power-managed node with a CPU and an FPGA. The other components consist of any circuitry that consume power. The power management unit communicates with all the sub-units to maintain the node-level power consumption within the upper-bound power budget.

5.3.3 System-level Power Efficiency

Power efficiency is not only important at the node level, but should also be considered at the entire system scale. This is due to the fact that not only nodes themselves that consume power, but also all of the resistive loads, including the large number of connector pins existing in a large-scale system. To analyse the holistic-picture of this system-level power efficiency, the same simulation sets discussed in Subsection 5.3.2. are able to be processed and manipulated to illustrate these power efficiency questions.

There is a similarity found in Figures 5.12 and 5.19. With a single ball allocated using the relative-position scheme, no multiple combinations of power-per-tile exist. Therefore, both the uniform and GA results follow the same thin-line pattern as shown in Figure 5.19. On the other hand, all of the larger-dimensioned systems explored by GA simulation, from Figure 5.20 to Figure 5.24, produce regions of multiple allocation-cases for a desired regulated power (shown in red on the graphs).

Surprisingly, in all of the simulation cases, the uniform scheme seems to provide the best efficiencies. The inferior efficiencies in red regions produced by the GA results are due to the mixing of low and high wattage nodes in a single system. This can

partly be explained by the nature of the power regulator, and its inherent efficiency with respect to output load and input voltage.

In Figure 2.22, the regulator-efficiency graph shows that when the load current is approaching the maximum-current specification of 5A, the efficiency slightly decreases. This is the reason for some of the GA results having lower efficiencies caused by high-wattage nodes existing in those regions. Also, as the system-level efficiency results follow the regulator-efficiency curves, this can also be considered as a validation of the models and simulation framework. The efficiency on the system sizes of 7x7x7-ball to 10x10x10-ball for the uniform scheme are also provided in Figure 5.25(a)-(d). Unsurprisingly, for 25W regulated load, the size of 10x10x10-ball gives the lowest power efficiency due to the voltage losses on a large number of connector pins, resulting in the approximate ratio of 0.76 shown in Figure 5.25(d).

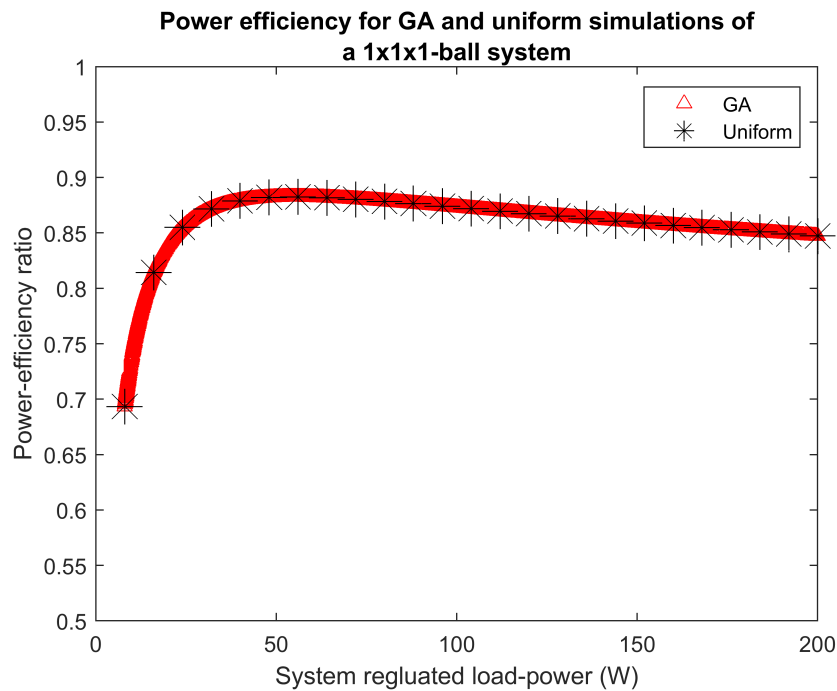


Figure 5.19: 1x1x1-ball system-level power efficiency on GA and uniform allocation schemes.

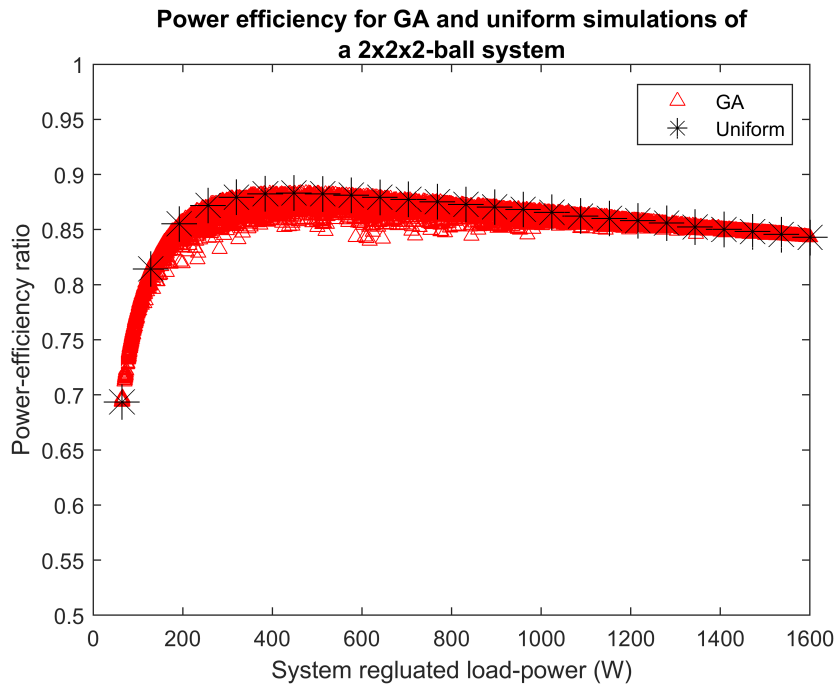


Figure 5.20: 2x2x2-ball system-level power efficiency on GA and uniform allocation schemes.

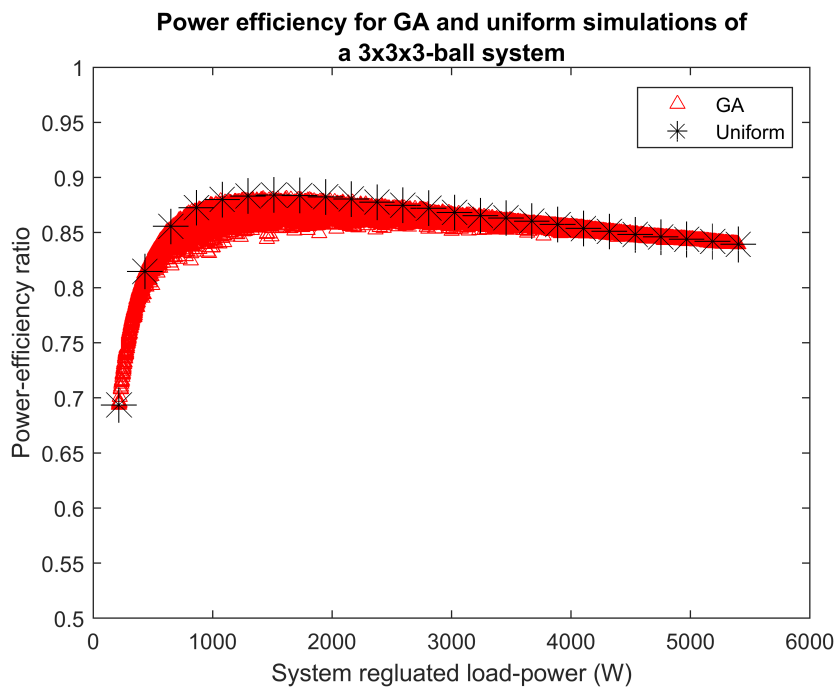


Figure 5.21: 3x3x3-ball system-level power efficiency on GA and uniform allocation schemes.

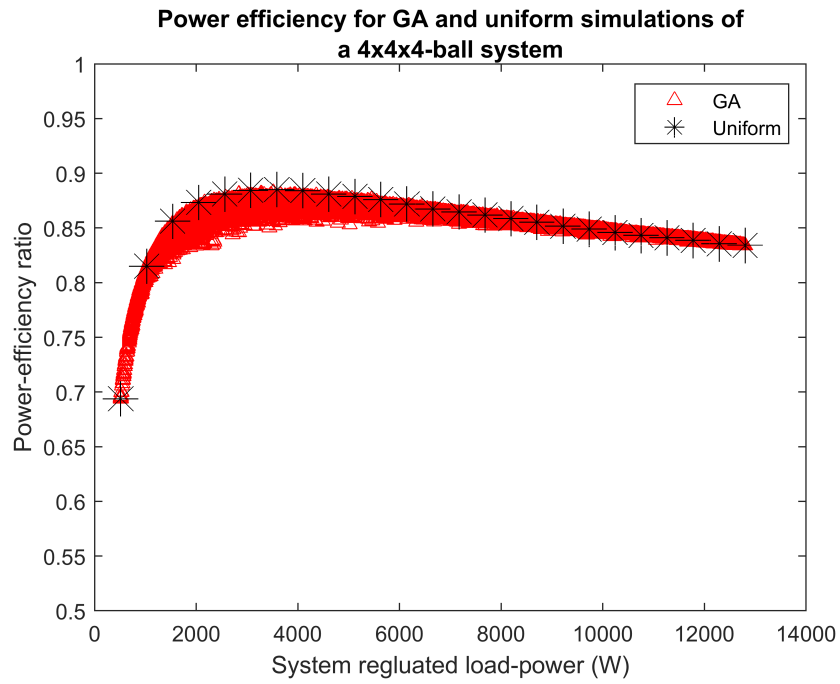


Figure 5.22: 4x4x4-ball system-level power efficiency on GA and uniform allocation schemes.

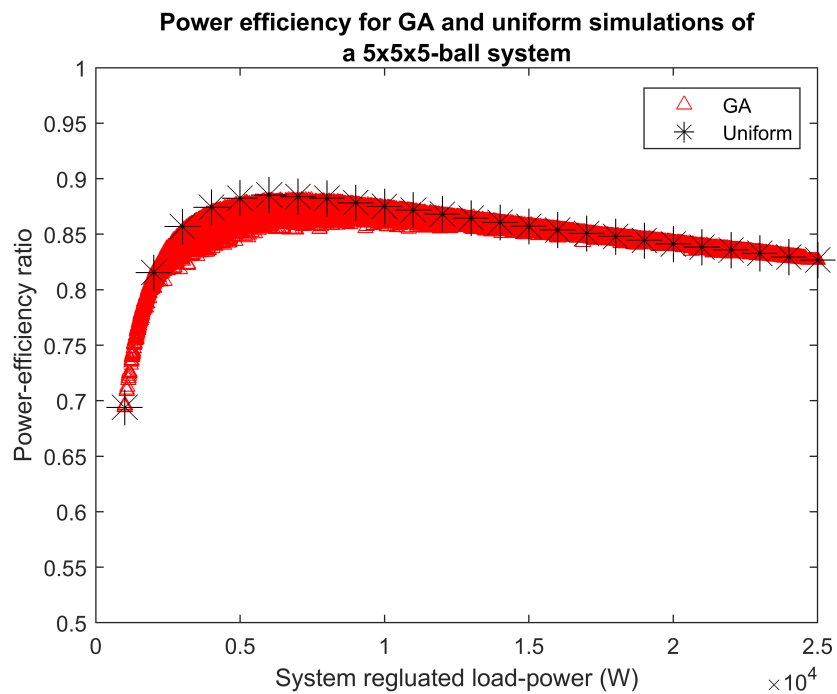


Figure 5.23: 5x5x5-ball system-level power efficiency on GA and uniform allocation schemes.

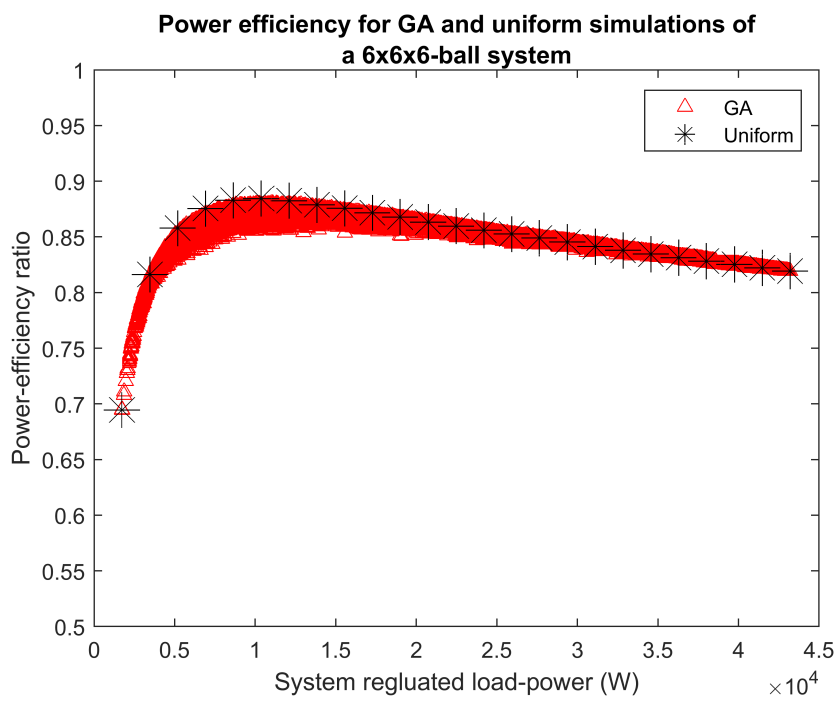


Figure 5.24: 6x6x6-ball system-level power efficiency on GA and uniform allocation schemes.

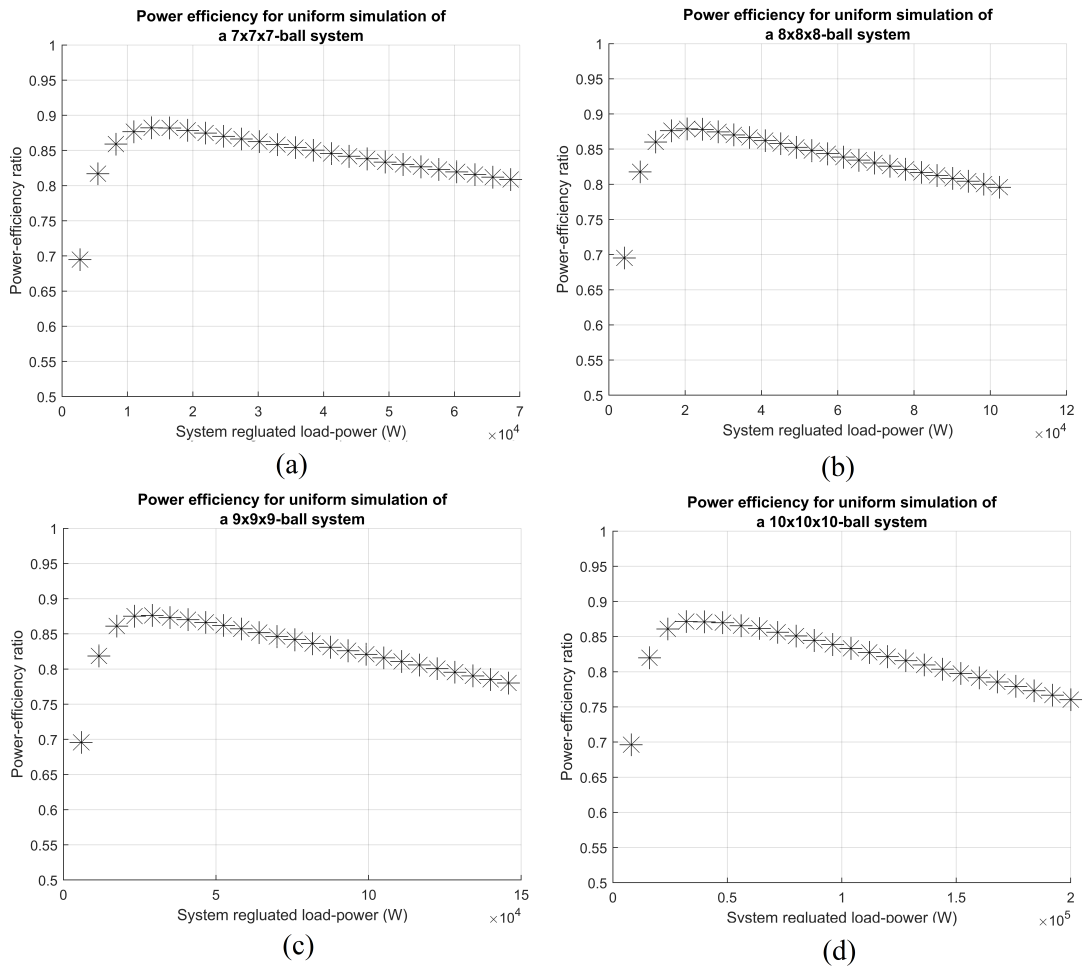


Figure 5.25: 7x7x7-ball to 10x10x10-ball system-level power efficiency on the uniform allocation scheme.

5.4 Total System Power

Having considered constraint and power efficiency simulations, another holistically meaningful data are the total system power. These power figures include all the power consumed by tile-level components, and also the power wasted on connector-pin resistances and the simulated power grid.

Figure 5.26 shows a range of cubic TCA systems. This figure is generated using the same uniform-simulation set of the TCA system-sizes ranging from 1x1x1-ball to 10x10x10-ball with the simplified board model earlier mentioned in Section 5.2. For any single cube-size TCA, 25 different load-wattage simulations are performed. The total system power of a cube-size TCA is equal to the single external voltage-source of 12V multiplied by the total input-current fed to the whole system. All the 25 total system power values are extracted into a single report file. With ten cube-size simulations, it results in ten report files generated. To obtain the power trends in Figure 5.26, each report file is read for plotting the total system power values as shown with the same coloured bars. The highest total system-power is approximately 263kW for the TCA cube-size of 10x10x10-ball, with each tile-level regulated load consuming an approximate power of 25W. With these estimated values, a TCA can be physically compared in terms of power consumption to some existing parallel/distributed computer systems. A 7x7x7-ball (343 balls, 2,744 tiles) TCA with 20W-per-tile (160W per ball) regulated power, consumes approximately the same amount of power to a 65Kw 64-board MDGRAPE-4A system reported in [28].

A report file for the 7x7x7 TCA cubic system can be found in Table 5.2. The report file shows that, with a total power-load nearest to the 65kW MDGRAPE-4A case above (at 66.4kW), the worst-case connector-pin current is approximately 8A, whilst the lowest board input-voltage is approximately 10.86V, which is still much higher than the lowest acceptable input voltage of approximately 5.5V reported in [65]. In the tile prototypes, two parallel-pins are for power or ground rails, thus doubling the maximum current allowed from 3A specification reported in [88] to 6A.

Obviously, the worst-case pin current of the 7x7x7-ball case is beyond this limit,

however, can be easily mitigated by occupying only an additional pin or pins for each of power/ground rails, for example by using the larger connector such as that illustrated in Figure 6.1 in Chapter 6. This allows us to conclude that it is possible to scale systems to large sizes, provided that care is taken to design the power interconnects between tiles or balls accordingly.

At the current stage of research and also with the different inter- and intra-node implementation choices, it is not straightforward to compare the related interconnection network performance as various factors, PE-types, routing algorithms and buffering, etc., are involved. However, comparability between TCA and equivalent logical topologies is possible. One similar system configuration is MDGRAPE-4A, a 3D torus, which can be implemented with TCA tiles or balls by adding additional wrap-around channels on the cubic faces at the system surfaces of a cubic TCA array.

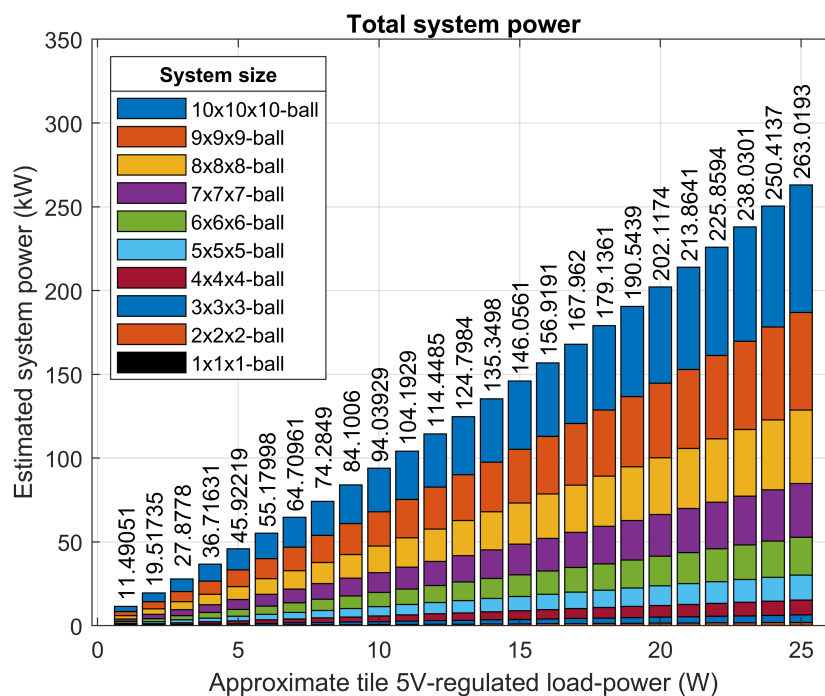


Figure 5.26: Estimated TCA system power of various cubic sizes based on the simplified board model, showing the additional power contributed by each increment of array size n (and thus total power for the final case), for a range of component tile power loads. Total ball power in each case will be 8 times tile load.

Table 5.2: Example of a uniform simulator report file for a size of 7x7x7-ball TCA with the board model based-on the tile prototypes in this thesis without fan for cooling. Two parallel pins are dedicated for each of power and ground rails, and the connector-pin parameter is set for 50m-Ohm mated pin-pair. The highlighted case represents the nearest equivalent power load to that of the 65kW MDGRAPE-4A system.

ExtVolt Supply	Connector PinModel Resist ¹	RLoad	ballPer Dim	minV Across PCB ²	max PinCur ³	iDiffThrs	Total System Power
12	0.0125	1.0000	7	10.5332	10.2146	0.0100	84821.1665
12	0.0125	1.0417	7	10.6006	9.7558	0.0100	81078.4954
12	0.0125	1.0870	7	10.6676	9.3023	0.0100	77355.4770
12	0.0125	1.1364	7	10.7334	8.8534	0.0100	73662.6173
12	0.0125	1.1905	7	10.7985	8.4082	0.0100	70018.8395
12	0.0125	1.2500	7	10.8627	7.9677	0.0100	66378.4142
12	0.0125	1.3158	7	10.9259	7.5331	0.0100	62795.3438
12	0.0125	1.3889	7	10.9891	7.1014	0.0100	59236.1195
12	0.0125	1.4706	7	11.0515	6.6765	0.0100	55666.2281
12	0.0125	1.5625	7	11.1125	6.2557	0.0100	52163.5242
12	0.0125	1.6667	7	11.1731	5.8342	0.0100	48661.8274
12	0.0125	1.7857	7	11.2333	5.4120	0.0100	45174.0826
12	0.0125	1.9231	7	11.2926	4.9955	0.0100	41761.8222
12	0.0125	2.0833	7	11.3509	4.5855	0.0100	38367.2571
12	0.0125	2.2727	7	11.4090	4.1816	0.0100	34991.5153
12	0.0125	2.5000	7	11.4660	3.7918	0.0100	31674.4193
12	0.0125	2.7778	7	11.5217	3.3931	0.0100	28379.3997
12	0.0125	3.1250	7	11.5770	2.9959	0.0100	25094.9176
12	0.0125	3.5714	7	11.6321	2.6039	0.0100	21859.4959
12	0.0125	4.1667	7	11.6868	2.2207	0.0100	18668.0568
12	0.0125	5.0000	7	11.7406	1.8456	0.0100	15550.5035
12	0.0125	6.2500	7	11.7927	1.4834	0.0100	12520.7969
12	0.0125	8.3333	7	11.8416	1.1357	0.0100	9584.7211
12	0.0125	12.5000	7	11.8888	0.7964	0.0100	6717.7470
12	0.0125	25.0000	7	11.9346	0.4684	0.0100	3949.3261

¹ The parameter 'Connector PinModel Resist' is set 0.0125Ω due to a resistance of 0.050Ω mated pin-pair is effectively reduced to 0.025Ω by two parallel pins used. Then, in the simulation model, each tile-edge has a single lumped-resistor. Thus, this 0.025Ω is respectively split into 0.0125Ω to each side of the inter-tile connection.

² 'minV Across PCB' is the parameter to store the lowest board input-voltage found in the entire power-distribution grid simulated.

³ The parameter 'max PinCur' is the parameter to store the worst-case (maximum) connector-pin current found in the entire power-distribution grid simulated. As can be seen in Figure 3.7(a), only a single lumped-resistor exists at a tile edge to model a single or parallel hardware pins. Thus, if parallel hardware pins are used for a tile edge's power (or ground) rail, the actual estimated electrical-current per hardware pin equals to this parameter's value divided by the number of the pins.

5.5 Preliminary Topological Analyses, Simulations, and Comparisons

Considering physical scalability can also imply the scalability of the overall computing performance. However, in reality, measuring a parallel/distributed machine in terms of computing performance is not straightforward and encompasses a large research field as various types of workloads may possess different interconnection network traffic patterns as shown in Table 2.5. In this thesis, evaluating system workload computing performance in deep detail is not the main focus of the hypothesis. However, preliminary investigations were conducted and are discussed to highlight possible potential for future work.

The approach taken was to consider how the developed power simulation tools might work alongside some existing computational performance toolsets.

As a preliminary feasibility study, Booksim2 [37], a widely employed interconnection network tool, has been adopted, and modified for the purpose of topological analyses. This also demonstrates the opportunity of how an existing interconnection network tool can be interfaced with the simulation framework proposed in this thesis in the future.

In this feasibility study, the TCA will be compared to SpiNNaker [5], a well known large scale parallel machine. SpiNNaker is constructed using rack/cabinet based packaging, and therefore it illustrates the 'traditional' rack and back-plane approach to large scale processing arrays outlined in the initial chapters of the thesis, where the drawbacks of such systems were highlighted.

SpiNNaker was also chosen as an existing system for comparison as it shares some of similarities with TCA - having multiple PE chips located close together in 3D space, albeit at rack/cabinet-level. Secondly, it also employs a torus-like topology, and attempts to support a logical 3D interconnection topology between nodes.

Another existing machine of interest is the MDGRAPE-4A [28] system. Although the number of board-level PE chips in MDGRAPE-4A is not as dense as SpiNNaker, as the

topology is obviously similar to a TCA implemented as a 3D torus, it draws interest to also include this system in terms of topology being equivalent.

Detailed simulations for interconnection network performance are out of the thesis' s scope as it requires portions of tool modification and relevant processes, for instance, relating node power allocation to injection rates, and tool verification. Additionally, It can be said that if data I/O of TCA is implemented using trapezoidal facet (T-facet), the performance characteristics will be of the conventional 3D mesh/torus topology. However, in future work, the hexagonal facet (H-facet) is another sub-surface area that can be further investigated for an alternative of inter-unit data I/O. This can be part of decision-making prior to implementing an equivalent logical topology in an interconnection network tool.

BookSim2 provides an advantageous topological construction feature, called *anynet*. This topological construction capability allows users to create a custom topology by writing a textual file containing the connectivity of nodes and routers. An example of a simple 2D-mesh topology with a user-file is shown in Figure 5.27.

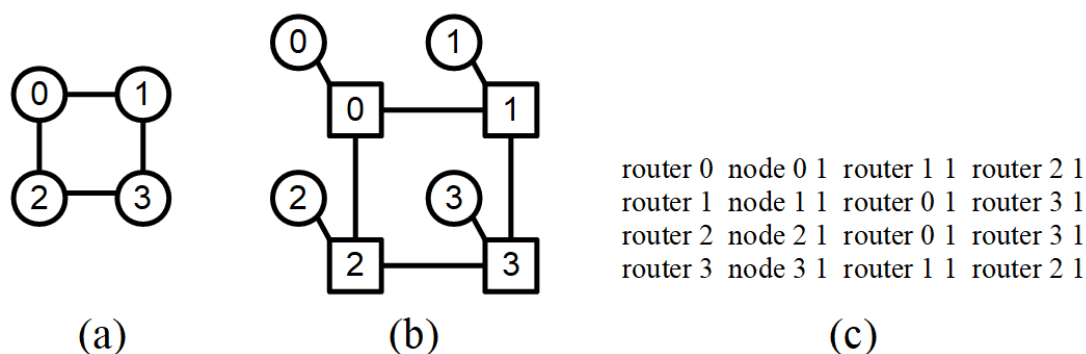


Figure 5.27: (a) a logical representation of nodes and bi-directional channels forming a 2D-mesh topology. (b) the same network with explicit routers shown. (c) example of BookSim2's anynet user-file constructing the network with uniform latency of 1 cycle. All the channel latencies in this example file are uni-directional.

An example of a user file is shown in Figure 5.27(c), each line starting with a router and its number, followed by its neighbouring devices, which can be either nodes or routers. The second numerical value immediately after, a node or router number, is the channel latency between a couple of router-node or router-router connections.

BookSim2 already provides built-in mesh and torus topologies for simulation purposes, whilst *hexagonal torus* employed in SpiNNaker does not exist.

Therefore, to assist with creating test cases, an automating tool to generate an anynet topology-file for this specific topology was also created in this thesis.

As an initial topological investigation, hop counts amongst the three systems are considered. To simulate the hop counts in a SpiNNaker system implemented with a hexagonal torus topology shown in Figure 5.28(a), a node-number convention is required for constructing a system using the anynet user-file entry in BookSim2. As a convention used in this thesis, an example of nodes sequentially numbered is shown in Figure 5.28(b).

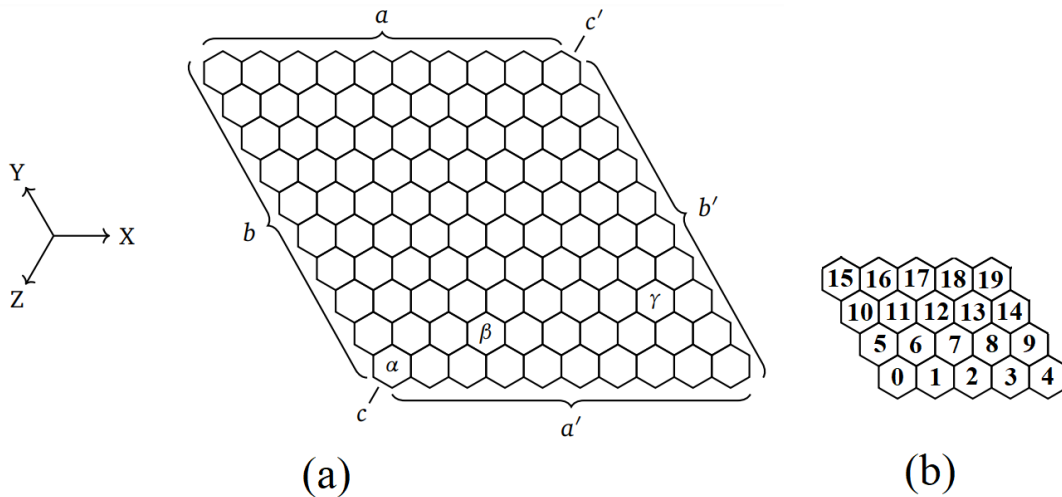


Figure 5.28: (a) A hexagonal torus topology employed in a SpiNNaker system. (reprinted from Figure 2.7 in [5])^a. (b) Example of node-number convention for hop-count simulations in this thesis.

^aWith granted permission by Jonathan Heathcote, the author of [5]

Hop counts of a systematic topology may be manually analysed by formulating the topological properties into mathematical equations, however, it is not only the nodes and channel connectivity in a topology contributing to hop counting, but also routing algorithms. In BookSim2, Dijkstra's algorithm is employed to search for one of the shortest paths for each pair of source-destination nodes.

In the current stage of research, the number of shortest paths possible for each pair is not evaluated. However, in future work multiple shortest paths that exist for each pair can be beneficial for distributing data units (packets/flits) sent from a source to destination node where the shortest possible hops are required. This extendable routing capability is beneficial for arbitrary topologies and routing algorithms, though this obviously requires an effort for tool modification.

Figures 5.29, 5.30, and 5.31, show hop counts for the three systems, TCA, MDGRAPE-4A and SpiNNaker. For hop-count simulations, all the systems are constructed with 729 nodes as this size is smallest in the range of hundreds and also give exactly the same number of nodes for each system. Table 5.3 shows a list of nodes per dimension and the total number nodes of 3D mesh/torus and hexagonal torus.

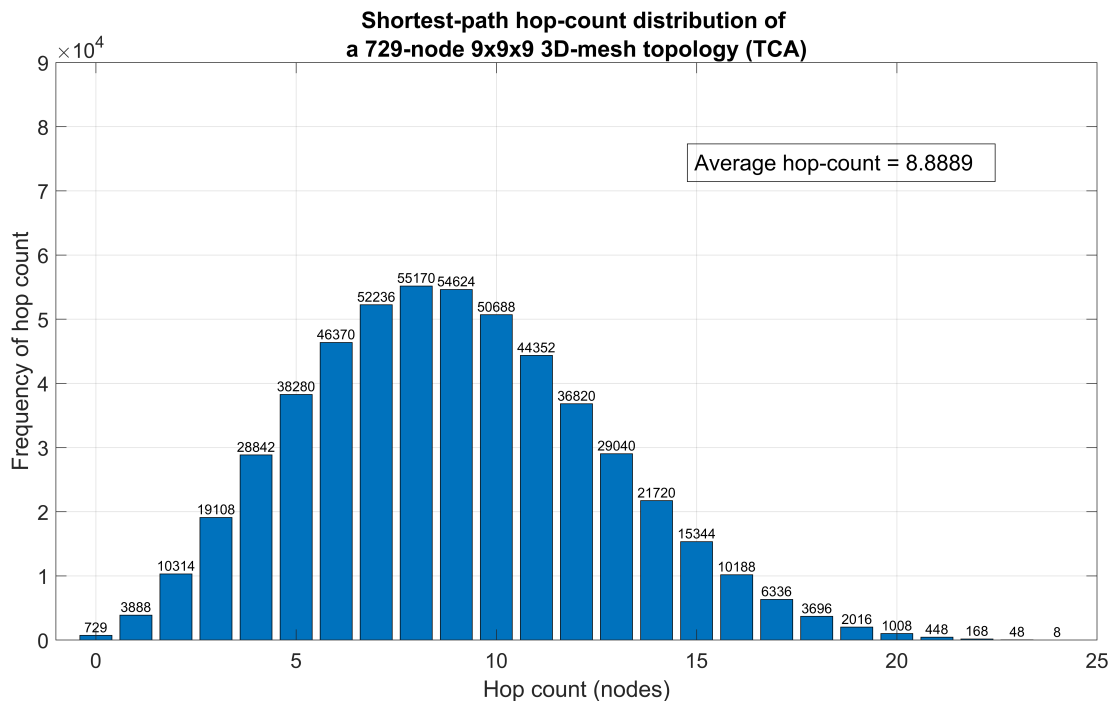


Figure 5.29: Hop-count distribution of a 729-node, 9x9x9, 3D mesh topology.

The conventional topology of TCA in the current stage of research is 3D mesh, therefore, its hop-count distribution can be seen in Figure 5.29. For MDGRAPE-4A, as it also shares the same 3D-torus topology with wrap-around channelled version of TCA, thus the hop-count distribution of those two cases is equivalent and it reduces the maximum number of hops, which is shown in Figure 5.30. Whilst TCA and

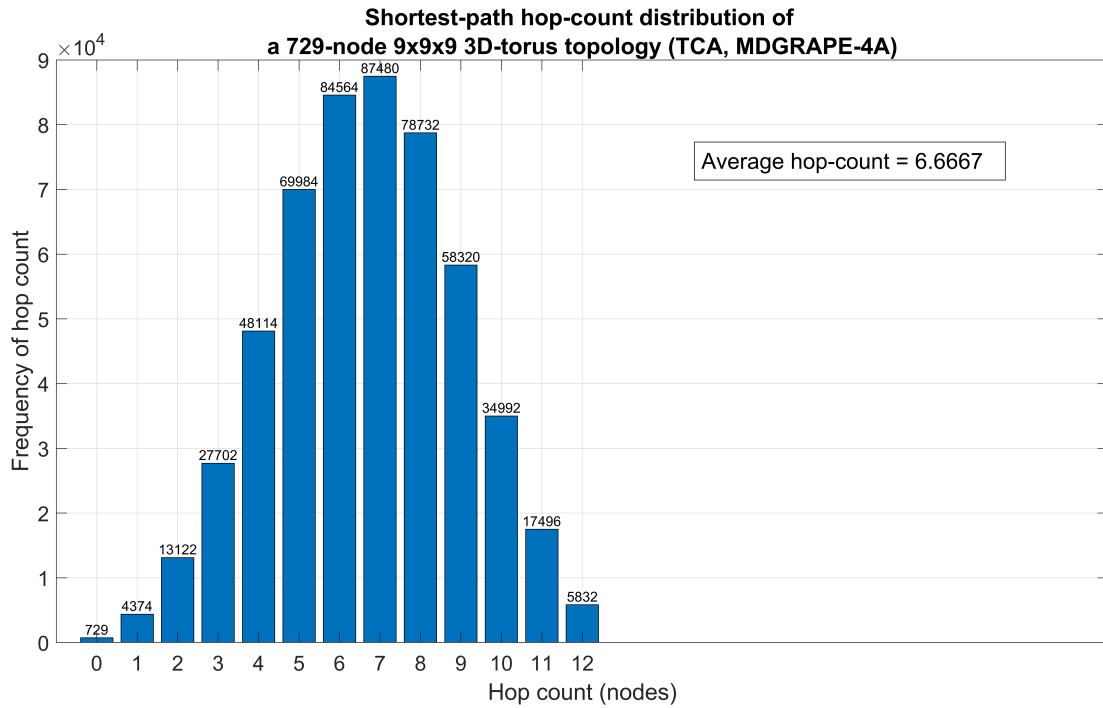


Figure 5.30: Hop-count distribution of a 729-node, 9x9x9, 3D torus topology.

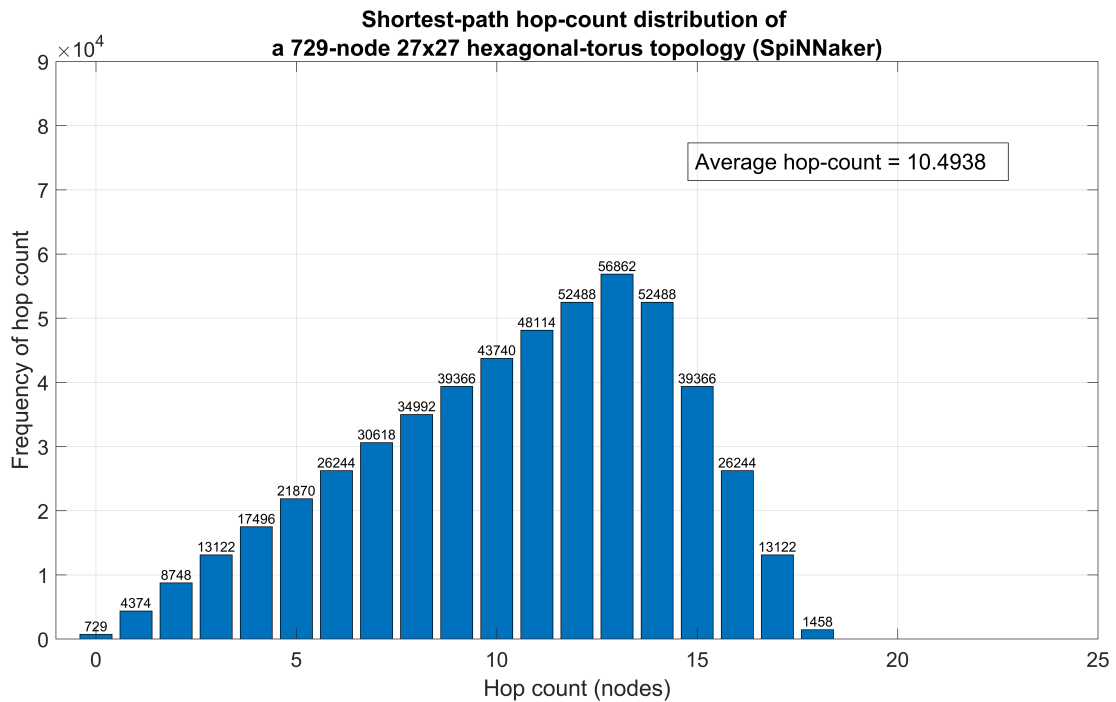


Figure 5.31: Hop-count distribution of a 729-node, 27x27, hexagonal-torus topology.

MDGRAPE-4A share some similarity in the distribution shape, SpiNNaker provides a distinct hop-count characteristic. SpiNNaker is better in terms of maximum hop

count than a 3D-mesh TCA but much less favourable than a torus TCA.

Having separately shown each hop-distribution, Figure 5.32 depicts all the hop-count data in another intuitive comparison. It can be obviously seen that a 3D-torus of either TCA or MDGRAPE-4A case outperforms the other two topologies considering both the maximum number of hops and the occupation range of lower hops (i.e. availability/multiplicity of short hops (i.e. availability/multiplicity of short hops)). Interestingly, 3D mesh/torus dominate larger portions of lower hops. However, hexagonal torus provides a lower maximum hop-count.

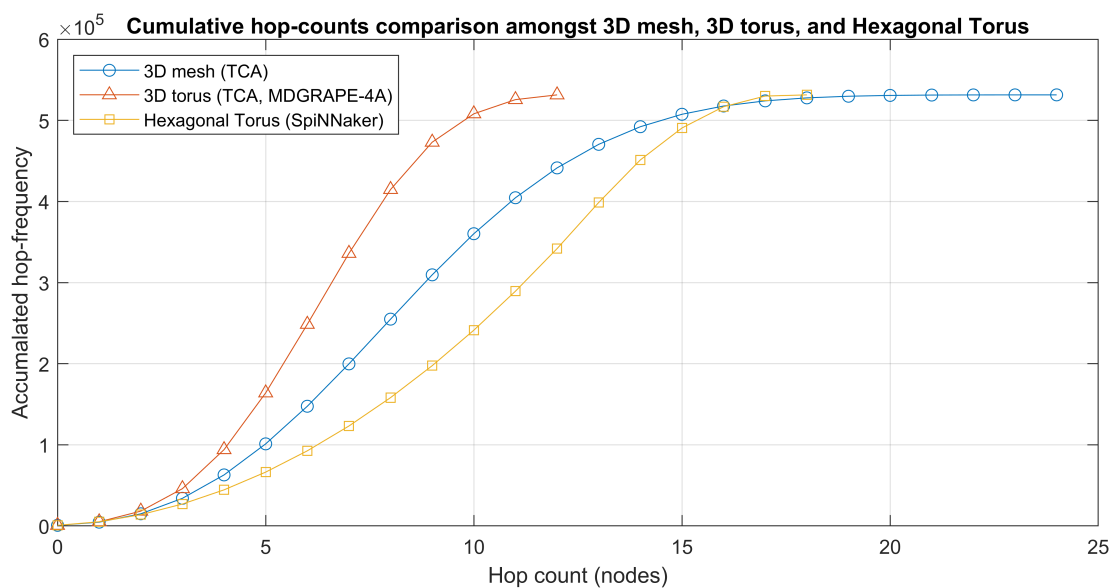


Figure 5.32: Cumulative hop-counts comparison amongst 3D-mesh, 3D-torus, and hexagonal torus.

Every topology has its own advantages and disadvantages. However, regardless of a topology having some attractive purely logical properties, it ultimately depends upon constraints of physical packaging technologies employed in their assembly. In the previous preliminary analyses, only a small set of example machines are selected as an initial consideration for tool integration. However, there are many more topologies having been investigated in the field, for instance, ring, star, tree, hypercube, dragonfly [95], etc. These topological variants for intercommunication are not in the scope of this stage of the research as only the investigation of power network is focused upon. However, more complex topologies can be applied to TCA such as the *double packing* described in Chapter 2.

Table 5.3: Comparison of nodes per dimension and the total number of nodes of 3D mesh, 3D torus, and hexagonal torus.

Nodes per dimension	3D mesh/torus	Hexagonal torus
1	1	1
2	8	4
3	27	9
4	64	16
5	125	25
6	216	36
7	343	49
8	512	64
9	729	81
10	1000	100
11	1331	121
12	1728	144
13	2197	169
14	2744	196
15	3375	225
16	4096	256
17	4913	289
18	5832	324
19	6859	361
20	8000	400
21	9261	441
22	10648	484
23	12167	529
24	13824	576
25	15625	625
26	17576	676
27	19683	729

5.6 Comparison of Large-scale Traditional Rack-mount Systems with TCAs

Traditional rack-mount systems can be implemented at large scales using multiple *discrete* racks/cabinets and potentially occupy the entire room for the HPC, with a complex mix of short and long range connections. Thus, it is not straightforwardly comparable to the unique power network in a *contiguous array* of TCA investigated in this thesis. Thus, focusing upon a sub-level of contiguous node-composition is considered a more sensible solution. Therefore, the rack/cabinet level will be compared with a cube-shaped TCA array. Taking into account the considerations above, a *comparison metric* is provided in this thesis to consider whether a TCA is considered to be 'large-scale' as achieved in recent traditional rack-mount systems, and can be formulated as per Equations 5.1, 5.2, and 5.3.

$$P_{e_PE} = \frac{P_d}{N_{PE_d}} \quad (5.1)$$

$$Mfactor_{(a,b)} = \frac{P_{e_PE_a}}{P_{e_PE_b}} \quad (5.2)$$

$$N_{PE_{(a)}} = Mfactor_{(a,b)} \times N_{PE_{(b)}} \quad (5.3)$$

where:

- P_d = Total power in the domain of interest
- N_{PE_d} = Number of PEs in the domain, meaning packaged chips in this comparison
- P_{e_PE} = Estimated average amount of power per the number of PEs in the domain
- $Mfactor_{(a,b)}$ = Multiplying factor, the ratio of P_{e_PE} in domain a over b
- $N_{PE_{(a)}}$ = Number of PEs in domain a equating with the power of PEs in domain b

In practice, the actual amounts of power per PE can vary by the PE types (CPU, GPU, etc.) employed in the system. Also, a single PE can vary its power consumption over time due to the dynamic computing workloads. However, the comparison model

Table 5.4: Comparison of the selected existing systems concerning given total system power and estimated PE-related configurations.

System	Power (kW)	Number of racks	boards/rack	PEs/board	PEs	PEs/rack approx.	Power/rack (kW)	Power/number of PEs* (W)
SpiNNaker	100	10	120	48	57,600	5,760	10	1.74
HAEC**	1	1	4	16	64	64	1	15.63
ExaNest	60	1	72	16	1,152	1,152	60	52.08
MDGRAPE-4A	65	4	16	8	512	128	16.25	126.95
Frontier	21,100	74	64	10	47,360	640	285.14	445.52 ^a

* Power/number of PEs represents an estimated power budget required per the number of PEs for the total system power reported. Thus, this does not reflect the actual power per PEs that may vary at run-time, or by the possible maximum consumption.

** The HAEC box is considered a small scale, thus, not selected for comparison at this stage.

^a Interestingly, with the TDP and TBP reported in [96], and [97], the average power per the combination of 1 CPU, and 4 GPUs, are very similar to this value.

mentioned focuses upon the estimated average amount of power per the number of PEs in Equation 5.1, by using the reported power for each machine. This metric is considered adequate for the purpose of TCA 'physical scalability' to compare with existing systems implemented with the different node power allocations.

To determine whether a TCA based on the hardware prototypes is able to be scalable to a 'large-scale' system, some important and recent traditional systems are selected as shown in Table 5.4 as use-case scenarios. This table shows a quantitative comparison concerning given total system power and estimated PE-related configurations. In this comparison, the parameter named 'Power/number of PEs' is used to estimate the ratio of the whole power consumption and the number of PEs existing at rack level of a given system of interest. This level of construction is preferred due to the fairness of comparison with TCA regarding the 'contiguity' of nodes in the system. Without this agreement, a virtually infinite large system can be constructed by loosely discrete multiple racks/cabinets, or separate groups of TCA arrays.

The 'Power/number of PEs' values will then be used to further calculate the equivalent number of TCA tiles required to contain all the PEs at the rack level of a desired traditional system in Table 5.5. The resulting parameter in Table 5.5 is the approximate equivalent TCA tiles, indicating how many tiles are required to mimic the same amount of power at the rack level of a given system. In this estimation,

Table 5.5: Estimation of the required number of TCA tiles with a power budget of 25W regulated-load per tile, to mimic the whole power consumed at rack level for a given system as described in Table 5.4 using Equations 5.1 to 5.3.

System	PEs/ rack approx.	Power/ rack (kW)	Power/ number of PEs (W)	Approx. equivalent TCA tiles
SpiNNaker	5,760	10	1.74	400 (411) ^a
ExaNest	1,152	60	52.08	2,400
MDGRAPE-4A	128	16.25	126.95	650
Frontier	640	285.14	445.52	11,406

^a Based upon a budget of 25W TCA regulated-load per tile, and a power/number of PEs of 1.74W, the actual number of PEs/tile is 14.37. Thus, this is required to be floored to 14 for an integer number of PEs located in a tile, which equals to 24.36W regulated-load per tile. Effectively, under a contiguous array of 10kW, this results in an approximate number of 411 tiles.

SpiNNaker is a special case, having 'Power/number of PEs' at rack level is under the 25W regulated-load, thus multiple chips could be located in a single tile. This obviously causes a sub-topology, for instance, partial/full mesh, ring, to be implemented for tile-level PE-to-PE interconnection network. Whilst it is feasible to migrate multiple SpiNNaker's chips to a single TCA tile, the other systems' PEs require power beyond the 25W budget, therefore a slightly different estimation is required. With this tile-level regulated-load power limit, it is assumed that a desired PE from a traditional system can operate upon a power-management technique such as variable clock frequencies to sustain the 25W power-budget.

Finally, as only cube-sizes of the TCA system are focused upon in this thesis, Table 5.6 shows the minimum cube-sizes for the approximate TCA tiles required from the previous calculation. It can be seen that in the case of MDGRAPE-4A, the connector-pin current limit of 6A is violated by an approximate maximum connector-pin current of 7.0171A. However, this can be mitigated by using some simple techniques, for instance, adding more edge-pins, or increasing the external voltage supplied, if complying with the regulator input-voltage requirements.

Whilst the connector-pin constraint is violated based upon the hardware prototypes, both the simulated and extrapolated results show that the minimum board input-voltages are still above the minimum input-voltage requirements (approximate 5.5V)

Table 5.6: Required cube-sizes of the approximate numbers of TCA tiles in Table 5.5.

System	Approx. equivalent TCA tiles	Required TCA cube-size (balls per dimension)	Maximum pin current (A)	Minimum board input-voltage (V)
SpiNNaker	400 (411)*	4x4x4 (512 tiles)	5.4886	11.5556
MDGRAPE-4A	650	5x5x5 (1,000 tiles)	7.0171 ^a	11.2887
ExaNest	2,400	7x7x7 (2,744 tiles)	10.2146	10.5332
Frontier	11,406	12x12x12 ^b (13,824 tiles)	18.7227, 20.6744	6.6602, 7.0172

* See the table-note of Table 5.5.

^a The sizes of 5x5x5-ball, and larger configurations, violate the expected maximum connector-pin current limit of 6A in the hardware prototypes. However, this can be mitigated by 1) adding more power/ground pins, or extending their the cross-sectional areas on the areas available in each trapezoidal facet, or 2) increasing the external voltage supplied at the cubic surfaces. For instance, some examples of increased pin arrangements are shown in Figure 6.1 in Chapter 6.

^b As the sizes of 11x11x11 balls and larger ones are not simulated, both the maximum connector-pin currents and lowest board input voltages are given with the lower- and upper-bounds by the extrapolations using the fitting models.

as specified in the regulator data-sheet [65]. The extrapolated minimum board input-voltages can be calculated using the voltage drops reported in Figure 5.7(e).

At this point it is worth emphasising that, apart from the models and simulation framework themselves, the hardware prototypes built in this thesis also play an important role for validating those models and therefore allowing predictions for scaling up to large-scale systems. Concerning the hardware prototypes, the best effort made possible was only at the scale of a single ball. This is due to the funding requirements for building larger scales of prototypes. However, with the validation work carried out, and in particular focusing on the role of connector resistance in the power grid, it shows that high accuracy can be achieved after validating the models against the affordable small-scale of the hardware prototypes. Moreover, the simulation results have also been considered as a range of predicted scales, rather than relying upon a single optimistic solution.

To conclude this section, an important contribution to the thesis is made here, validating the idea that TCA systems are electrically competent in meeting large scale system expectations and also in ways that are comparable to other more conventional systems in terms of computational topologies.

Conclusions and Future Work

In this chapter, the research hypothesis and objectives, contributions, useful experience of the author regarding creating novel simulation platforms will be detailed. Additionally, several possible opportunities for future work are also given.

6.1 Conclusions

A number of important aspects have been recognised and investigated in this thesis. The following subsections outline the key outcomes and conclusions in more detail.

6.1.1 Research Hypothesis and Objectives

Having proposed the models, simulation framework, hardware prototypes, and TCA scalability evaluations, the research hypothesis will be restated as follows:

In the first part:

It is feasible to build a physical large-scale Tiled Computing Array within the power-grid constraints given

Given the quantitative comparison in Section 5.6, it can be concluded that a 'large-scale' TCA power distribution grid is feasible to be constructed with the unconventional power network investigated in this thesis. This is due to the fact that the recent systems selected for comparison can be equivalently achieved to the scale of 10,000 tiles. Whilst the current constraint is not within the specification, this is due to the initial prototypes built in the current stage of the research. Also, alternative higher-wattage switching regulators are also available. Moreover, the external voltage supplied to be investigated in this thesis is 12V, which is the voltage rated in the connector data-sheet. Deep investigations for more sophisticated inter-node power



Figure 6.1: Examples of power connectors and enhanced designs. Left: Existing prototype connector 22mm length, 6-pin (2xVcc, 2xGND, 2xData). Middle: Larger connector (approx. 30mm length), 12 pins, example: 4xVcc, 4xGnd, 4xData, giving double the power capacity. Right: Bespoke Design © C Crispin-Bailey, University of York, Diameter 25mm, 30-way connector, exploiting connections via hexagonal H-Facets rather than trapezoidal T-Facets.

media in the future can lead to improved alternative designs to support higher voltages and currents. Allowing higher external (cubic-surface) voltages supplies not only decreases the worst-case inter-node current, but also consequently eases the design of inter-node power media in terms of the current limit.

As shown in Figure 6.1, there are many possibilities to enhance inter-tile connectors in this respect, and a further advantage of more pins in parallel is that their overall resistance is reduced. In other words, existing limits of scalability can be overcome by improving the design of tiles and connectors without a step-change in technology.

For the second part of the hypothesis, the further criterion was added such that (in bold):

*It is feasible to build a physical large-scale Tiled Computing Array within the power-grid constraints given, **whilst still scaling up the system computing performance.***

Quantitatively, this additional expectation extends upon the first part of the research hypothesis in terms of the overall computing performance measures of the system. This is evident in the sense that if a large number of nodes is achievable, then the requirements are potentially met with which to obtain a high performance computer system at present. To perform quantitative investigations for system computing performance requires sophisticated models to relate power allocation and detailed in-

terconnection network performance together with workload analysis, and this is not within the scope of this thesis.

However, to make some degree of evaluation on this point, in a qualitative way, the TCA concept also provides advantageous constructional properties discussed in the earlier chapters do support this hypothesis and can be summarised as follows:

- ▶ **Data channels:** TCA provides non-cabling and direct coupled nodes. This implies low latency and high bandwidth achievable in physical implementation. The convenience of facet-to-facet I/O connections also permit parallel data channels to be implemented very easily compared to rack and back-plane solutions.
- ▶ **Hop counts:** Although TCA or any other systems that are implemented with 3D mesh/torus topologies suffer from the high maximum hop-counts when the system size is growing, this is nonetheless comparable to many existing HPC systems, such as MDGRAPE-4A.
- ▶ **Future capability:** A unique opportunity within the 3D geometry of TCA systems is that hop-counts can be further reduced in future work by exploiting the idea of direct node-coupling to nearby non-neighbour nodes using *bypass channels* [98] to the second immediate nodes. An example of conceptual design is depicted in Figure 6.2. These additional channels not only reduce hop counts, but also provide a higher degree of tolerance for the channel-failure situations, and increased internal data bandwidth. The 3D structure of TCA potentially allows this to be achieved with a much higher degree of efficiency than any rack-mount system topology.

Given what has been discussed and the above points, and in particular the fact that TCA can achieve identical logical network connectivities to well known HPC systems such as MDGRAPE-4A, it can be considered that the second criterion can be met.

With the research hypothesis having been tested, all of the research objectives will also be restated with relevant success criteria to discuss how pieces of the evidence in previous chapters allow to answer them as follows:

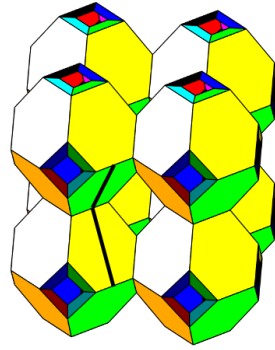


Figure 6.2: A conceptual design for a bypass channel between second-immediate hops. The black line shows a dedicated line medium implemented on the same or a different PCB layer of existing components.

► **Objective 1: Employing and designing models and simulation tools**

The hierarchical and parameterised models for validation have been designed and discussed in Section 3.4. An example of the simulation times and files sizes are reported in Table 3.1. Regarding the existing simulation tools, as shown in Figure 4.1 and Section 4.8, both LTspice[®] and ngspice are examples of employing existing tools in the simulation framework. The SPICE simulations are semi-automated with various parameterised models and modular generators for mitigating labour-intensive tasks.

► **Objective 2: Hardware validation**

The accuracy results have been reported in Section 3.5. For the validation of simplified vs complex models, the overall percent errors were found to be less than 1%, whilst the case of the complex models validated against the hardware prototypes can achieve the overall averaged errors within an approximate range of 1-2%. This excludes the no-load case, which has the overall average value of approximate 5% for error. This idle power-load is considered not a significant concern in the scenarios of high-percentage PE utilisation. However, this issue is one of the aspects to be further investigated in the future. These reports have also been published in [42].

► **Objective 3: Fundamental simulation experiments**

Prediction trends with lower and upper bounds with different fitting equations are provided in Section 5.2. It is unfeasible to simulate TCA cases with very large sizes. Thus, to mitigate this issue, a range of ten TCA sizes are simulated

and used as base data to perform extrapolations for desired larger scales. The initial plan for regulated-load power allocation should cover a sensible range of power starting from 0W. Unfortunately, it was found in the experiments that the data points in the case of 0W regulated-load caused some difficulties in the fitting processes. However, A zero-wattage power budget is not typically practical in real usages as nodes without regulated power consumption would waste the area occupation of the system. This issue is considered trivial as the simulated cases start from 1W, which is very close to the idle 0W earlier planned. The inter-node electrical resistance is based upon a simple model, created as a single lumped-resistor. The external power-supply is set at 12V as a widely typical use of power supply voltage. These two items have been carried out as planned.

► **Objective 4: Optimised power distribution**

Considering the GA-based simulation results, which can be found in Subsection 5.3.2, it clearly demonstrates that the optimisation framework is able to discover multiple power-allocation cases that are superior to the conventional uniform ones. With these simulation results, it suggests that heterogeneously power-allocated nodes are an attractive choice when concerning system-level power under a constraint of connector-pin current limit.

► **Objective 5: Scalability evaluations**

Based upon the scalability comparison provided in Section 5.6, and the research hypothesis tested true, it gives high confidence to move forward to continue with various aspects towards larger scales of prototype building, design variants, and also interconnection-network performance aspects.

► **Objective 6: Simulation framework documentation**

This objective has been included for the completeness as part of long-term tool design-practice, and is still in an ongoing stage. However, during the time of simulation-framework development, examples of the documentation processes currently achieved are as follows:

- In-line comments for important lines of code
- Examples of step-by-step running of functions, provided in corresponding

source files

- Separate test-case wrapper calling functions, if required.
- Some separate files containing visualisations explaining some complex internal processes

6.1.2 Contributions

The contributions previously highlighted in Chapter 1 are discussed with a review perspective as follows:

- ▶ **Modelling Framework:** Referring back to Section 3.3 and Subsection 3.4.2, both theoretical, and practical models based on the hardware prototypes, for inter-node power media have been developed. In the theoretical conductive-medium, a *hexagonal* plate was only initially investigated for the possibility of sophisticated conductive designs. This hexagonal plate does not actually only act as inter-node, but also, intra-node medium at the same time. For the practical conductive model in this thesis, parameters are based on an off-the-shelf connector pin selected for building the hardware prototypes but being commonly used in a wide range of similar connectors. With this typical pin-shape, a derived inter-node resistance can be easily formed compared to arbitrary ones that may require a complex current conduction simulation tool.

In a complete top-down abstraction of intra-tile level modelling, various hierarchical models ranging from, tile, the pin-resistance earlier mentioned, board, voltage regulator, and regulated load, are discussed from Subsection 3.4.1 to 3.4.5. Ball-array and power source models can also be found in Subsections 3.4.6 and 3.4.7.

- ▶ **Simulation Framework:** In Chapter 4, a large number of simulation tools have been carefully developed as a modular design. This also facilitates future extensions without tremendously restructuring the whole framework. As can be seen in Figure 4.1, without the automation of this simulation framework, it is virtually impractical to evaluate the scalability manually by tremendously labour-intensive SPICE-code generations and data analyses.

To accelerate simulations, as shown in Figure 4.19, automation scripts were implemented for a single-machine to run simulation instances in parallel over multiple processing cores. For optimisation of node power allocation, in Sub-section 4.7.4, a GA-based non-uniform simulation framework has also been proposed, implemented and demonstrated.

All of the scripted visualisations throughout this thesis are one of the important features. Without this capability, some meaningful representations cannot be easily understood. Visualising simulation results can also help verifying simulation tools, in addition to the levels of 'unit' and 'integration' testing.

- ▶ **Hardware Prototypes:** In Section 3.2, a variant of the theoretical hexagonal tiles, as real fabricated hardware prototypes, have been built specifically to support the validation of the scalability evaluations in this thesis. Whilst primarily for validation purposes, various useful practical issues have been manifested during their design, and various stages of decision making. This is a highly valuable case-study, and a lesson learnt for future developments.
- ▶ **TCA power-distribution grids scalability and optimisation evaluations:** In Chapter 5, this can be seen as the outcomes of the three items above, which is the essential purpose of the thesis. For scalability analysis, as can be seen in Section 5.2, without these results, it would not be predictable as to whether scalability could be achieved for large scales and high performance computing. Understanding the large-scale power network has not only benefited scalability validation, but also for recognising aspects such as fault tolerance, resilience, and optimality of system configuration and design. Moreover, further GA-based optimisation results are also provided in Section 5.3. These several simulation results also play an important role when, in practice, non-uniform power allocation is expected.

6.1.3 Tool Design Experiences

Unlike some other research work that may be conducted based on well-established simulation tools, this thesis has undoubtedly given considerable attention to building

both the models and a simulation framework to carry out simulations required for their utilisation. Any tool developers may have different developing styles, but the following recommendations from the author's experiences during this thesis should be used as general guidelines.

- ▶ Well structure the simulation modules, starting using the top-down view. Having broken down all hierarchies of the modules, each of them can be separately focused on for completion on its own. This is not only to clarify what a specific module should perform, but also for modularity and portability in the future.
- ▶ When facing problematic simulation runs, start with simple checks, for instance, path/function names, the correct number of variables, the existence of required simulation files, rather than deeply digging into source code in the first place.
- ▶ Log useful statistical information such as simulation times, processing utilisation, and memory and storage requirements, during simulations. These pieces of information, sometimes, may not be able to be recorded in a fine precision such as CPU utilisation since it may vary during a simulation.

Moreover, the statistical logging modules themselves, sometimes, can also interfere in the statistical data if not well considered such as too frequently logging CPU times. In this thesis, only the simulation results are focused on. However, the simulation framework can be improved to add these additional simulation statistical logging modules in future work. Dedicated simulation resources provided by the organisation/institution are also highly recommended, which benefit both simulation-related demands and accurate logging data.

- ▶ Avoid module-name confusion in the future. During a specific period of time, the tool developer may obviously comprehend all the modules being focused upon. However, leaving the development for a period of time, it could lead to unnecessary confusion to oneself. Good naming conventions and documentation can alleviate this issue.

6.1.4 Limitations and Assumptions

Even though several efforts made to the modelling and simulation for power distribution grids of 3D Tiled Computing Arrays, there are some limitations and assumptions worth mentioning as follows:

Limitations:

- ▶ As described in Subsection 2.1.2, there can be many forms of module and system compositions. Only a design of hexagonal tile is chosen to be investigated in this thesis. However this design, forming a truncated octahedron, is considered one of the most optimal choices for modular packing and therefore a natural first choice.
- ▶ In practice, the fluctuations of voltage and current all over the system can also be of interest. This thesis focuses upon these two types of quantity only when the system is in a steady state.
- ▶ Further simulated TCA sizes could lead to more obvious predicted trends. However, due to the research time-frame and simulation facilities available, only the sizes of 1x1x1 to 10x10x10 balls are simulated for predictions in Chapter 5.
- ▶ The parallel-simulation capabilities in this thesis help with reducing total time required for all of simulation instances given. However, it currently supports only the distribution of simulation instances in a single machine.
- ▶ Ideally, hardware prototypes for validations would be multiple-ball configurations. In this thesis, to the best effort made possible, only a single ball can be achieved. New work planned and being undertaken at the University of York includes a new 'K1' prototype ball array and the work undertaken will be valuable in supporting that system validation.

Assumptions:

- ▶ In practice, whilst practical power consumption at the voltage regulator's output can vary due to dynamic computational loads, it is assumed to be constant in this thesis. A more thorough dynamic analysis would be worthwhile in the future,

although the power connector pin specifications do also permit transient loads far in excess of the steady state current ratings.

- ▶ The voltage regulator in a particular tile can differ from the others employed in the system. Only a single voltage regulator model is employed in this thesis.
- ▶ At the intra-tile level, it is assumed that the conductive-medium resistances (for example internal PCB tracks or buses) are negligible.
- ▶ The values of connector-pin resistance may differ within specified tolerances amongst all the pins used in the whole system. For simplicity, a uniform value is assumed.
- ▶ Partial external power connections are possible, whilst it is assumed in this thesis that all of the power connectors are fully connected for simulations. However, a benefit of the simulation platform is that such scenarios can easily be evaluated in the future.
- ▶ In this thesis the TCA is assumed to be externally supplied with a 12v supply (and same for the hardware prototypes). However there are other voltage rail choices. Again, the simulation tools allow this to be explored if desired.

With the limitations and assumptions above, it can be concluded that there are still many interesting challenges and opportunities to be tackled in the research field for future work, some of which have become apparent as a result of the work presented in this thesis.

6.2 Possible Future Work

In this section of possible future work, both the simulation framework and hardware prototypes will be discussed to foresee how they can be improved for future investigations.

6.2.1 Lower Hop Counts

Inherently, large-scale 3D mesh topologies encounter the issue of high hop counts. A 3D torus can also partially mitigate this issue. However, the wrap-around channels are physically long for large systems and could lead to non-uniform regions of lower bandwidth and high latency. Travelling through communication hops does not only face the physical distance issues, but also the traffic congestion in some cases, if routing devices cannot effectively manage the traffic. In most systems, a hop involves some form of packet transfer protocol and potentially a routing decision point per hop, both of which imply significant bandwidth overheads.

For TCA systems, as suggested earlier, *bypass channels*, can be an attractive idea to mitigate this hop issue. Systematic bypassing channels, or arbitrary additions are both possible alternatives. A well-known shared data path, bus, is also another possible physical solution. The key point here is that connectivity to neighbour nodes and secondary neighbours is far easiest to achieve in 3D than in 2D rack and backplane systems. With a one-level bypass between a node and its secondary neighbours (using a hardwired path through the nearest neighbour intermediate node) worst case hop counts could be approximately halved in large systems.

6.2.2 Investigation of External Power Designs

This thesis focuses on the behaviour of power grid itself, not the supply units employed outside a TCA system. The total power consumed by a large-scale TCA or any traditional system draws a large amount of total electrical current and hence system input power. However, a technique such as parallelisation of power supply units can be investigated in future work for distributing currents to various points of power connections.

In addition, the capability to simulate any external power connection pattern and evaluate the efficiency and optimality of such cases is available within the existing tool-set has been developed.

6.2.3 Arbitrary Inter and Intra-unit Level Power Media

The theoretically hexagonal plate has been initially discussed in this thesis. However, it is not limited to this shape, any conductive medium designs that suit a variant of TCA-unit design in terms of electrical properties can also be possible. This information could be obtained from the PCB design tools used to develop a hexagonal tile main-board component and incorporated into the simulation parameters in future experiments.

6.2.4 Node Power Model

The board model in this thesis focuses on constant power model. This is due to the need to simplify simulation efforts for large-scale simulations. In future work, a more sophisticated model to observe the voltage/current spikes caused by dynamic workloads would also be beneficial to investigate the impacts on the conductive media in some circumstances.

6.2.5 Multiple Optimisation Algorithms for Power Allocation

Only a GA-based simulation framework has been proposed in this thesis. However, other optimisation algorithms can be employed for further investigations, e.g., particle swarm [94].

6.2.6 Improved Visualisations

Even though some visualisation capabilities have already been developed. Some specific visualisation modules are still considered experimental. Thus, further improvements can be carried out. Visualising power allocation is for physical scalability, whilst animating intercommunication traffic, which may be related to power information, is also another attractive and meaningful tool to deeply understand how power allocation affects the traffic characteristics.

6.2.7 System Computing Performance Analysis

This thesis confirms the feasibility of TCA power-grid and its scalability. However, relating power to interconnection network performance is also beneficial to understand how it impacts on overall computing performance. BookSim2, a widely-employed interconnection network tool has also been briefly discussed in Chapter 5. Further work on this topic would be valuable.

6.2.8 Hardware Prototype Improvements

The current hardware prototypes are based on, hexagonal tiles. In future work, several variant designs could be possible.

Reductions in size will ultimately lead to balls being the fundamental components in systems, and meanwhile, the power network and data intercommunication connectivities are also not limited to 3D mesh topology through Trapezoidal facets. As briefly illustrated in Figure 6.1, it is possible to utilise hexagonal facets as interfacing areas for different inter-node power and communication solutions. An updated simulation tool could simulate these cases just as easily as the T-facet connectivity model used in this thesis.

6.2.9 Simulation on Computing Cluster

The simulation framework proposed is currently performed on a single machine supporting multi-core processors. However, in future work, a framework to run large simulation instances on a computing cluster truly managed by a job scheduler, e.g., SLURM (Simple Linux Utility for Resource Management) [99] can also be expected.

6.2.10 Cooling Systems

Cooling techniques are also important in the implementation of high performance computers. The TCA concept considers this issue in the design principle itself by incorporating cooling channels into the ball structures formed by tiles. However, deeper understanding cooling behaviours in a tiled computing array could lead to more viable and effective cooling solutions for large-scale systems. There is clear opportunity to couple power per node data to a cooling (air/fluid) dynamics modelling toolset if one is available.

6.2.11 Other Engineering Concerns

Finally, not only node, intercommunication, power-distribution grid, and cooling designs, other engineering concerns regarding physical constructions should also be taken into account in the future. The materials used for physical construction, and the design decisions made for issues such as cooling (air, liquid, etc.) are essential future challenges to be addressed.

6.3 Final Remarks

Given all the findings, the author of this thesis strongly believes that the modelling and simulation for power distribution grids of 3D tiled computing arrays proposed, shows they are highly feasible, and not only beneficial for the results and outcomes carried out, but will also be useful in the future for further investigations and improvements. There is good reason to believe such systems could be a viable concept for future unconventional HPC systems.

The continued publication of research work in this field, including the paper produced as part of this PhD, and the many previous PhD projects at York, show that this field is open and capable of much further interesting research in future years. It is hoped

that the tool-sets developed here and the results examined should form a valuable part of that work.

Appendices

A

Published Work

- ▶ A.1 Conference Paper (Submitted Manuscript)
- ▶ A.2 Short Report on Hex-tile Project
- ▶ A.3 Hex-tile PCB Layouts

A.1 Submitted Manuscript of DSD2022 Paper

Tiled Computing Array: A 3D Modular Scheme for the Interconnection of Large Array Topologies: System Modeling and Prototype Feasibility.

Pakon Thuphairo
Computer Science
University of York
York, United Kingdom
pt795@york.ac.uk

Christopher Bailey
Computer Science
University of York
York, United Kingdom
chrisher.crispin-bailey@york.ac.uk

Anthony Moulds
Computer Science
University of York
York, United Kingdom
anthony.moulds@york.ac.uk

Jim Austin
Computer Science
University of York (retired)
York, United Kingdom
jim.austin@york.ac.uk

Abstract— This paper presents the Tiled Computing Array (TCA), a simple, uniform, 3D-mesh packaging at inter-board level, for massively parallel computers. TCA eliminates the need for hierarchical rackmount-structures and introduces short and immediate data channels in multiple physical orientations, allowing a more direct physical mapping of 3D computational topology to real hardware. A dedicated simulation platform has been developed, and an engineered prototype demonstrator has been built. This paper explores the feasibility of the TCA concept for current hardware technologies and systems, evaluates power modeling and validation, and highlights some of the novel design challenges associated with such a system. Evaluations of physical scalability toward large-scale systems are reported, showing that TCA is a promising approach for large-scale processor arrays.

Keywords— *computing array, interconnection network, massively parallel computers, scalability, simulation*

I. INTRODUCTION

The complexity of hardware structures in the building of parallel computers is not insignificant in terms of the effort of composing complete functional systems, starting with the processor chip as a fundamental building block, alongside memory devices, SSD and communication ICs, and power conditioning components. There are inherent complications in addressing this via board-level design, rackmount, and modular system hierarchies, and these physical demands create topological compromises between the logical processing structure and the physical equivalent. These are evident in terms of wiring constraints, power delivery and heat dissipation, and in terms of computational density of such systems.

In this paper, we aim to investigate a completely different approach, aiming to address such difficulties with a completely different structural paradigm, based upon fundamental building blocks, referred to as tiles, or ‘hex-tiles’. Tiles are therefore modules containing one or more chips, perhaps ultimately embodied as an adaptation of existing well established IC packaging technology encapsulating with a single SoC die or a perhaps a multi-chip module (MCM). Initial prototypes are necessarily less sophisticated and rely upon PCB level IC integration to create tile modules, an order of magnitude larger in scale, but capable of demonstrating concepts and principles.

Tiles as fundamental building blocks are capable of being tessellated in multiple ways. Due to a novel angled edge-interface arrangement, a group of eight tiles may be composed into a 3D structure which we equate to a ‘ball’. Balls may then be coupled to each other to build larger systems, also extending directly in three dimensions as uniform arrays.

Hex-tiles directly connect power and IO to one another, completing the power and data grids without circuit boards, racks or other physical needs. This of course results in power delivery challenges and requirements for the consideration of data distribution, connectivity, and latency in this new and different model. In order to extend the knowledge of such systems and assess their viability, we present a conceptual model, a prototype, and a simulation tool which is used to investigate how these electrical constraints impact upon the scalability and feasibility of the system.

II. MOTIVATION

Current state of the art massively parallel computing systems relies heavily upon the well-established technologies of back-plane, rackmount, and server cabinet infrastructures, along with the associated power bus architectures and interconnection strategies. Obviously, most of the systems are comprised of the supporting infrastructure, and relatively small parts of the system are the actual CPU, memory bank, SSD or other resources. In effect, the desired high-density collection of processing elements is forced to map onto a variety of physical inter-board level construction constraints, many if not all of which then impact upon other critical factors such as interconnection length, cooling strategies, granularity of local versus inter-module communications, and so-on.

The motivation for the tiled computing array (TCA) stems from this observation, and the question ‘how can we interface maximum processing elements with minimal infrastructure and constraints’. The TCA concept eliminates the need for rackmount architectures and permits a more direct physical mapping of 3D computational topology to real hardware. Eliminating rackmount infrastructure also means potentially much higher processing density. Interconnections are not constrained by granularities relating to cards, racks, cabinets, and so-on.

TABLE I. SUMMARY OF IMPORTANT CHARACTERISTICS OF THE SURVEYED PACKAGING SYSTEMS

Packaging	Topologies	Inter-board Packaging	Power delivery	Inter-board Communication	Hardware Implementation
[14]	optical multi-mesh hypercube	not specified	not specified	wireless (optical)	conceptual
[15]	hypercube and mesh	not specified	not specified	wireless (optical)	conceptual
HAEC [9]	wireless configuration	HAEC Box	not specified	wireless	HAEC playground (network-protocol evaluations)
[10,11,12]	wireless configuration	ball-shape object	wireless	wireless	conceptual
ExaNest [1,2]	hybrid [16]	rack/cabinet	backplane	wired	rack/cabinet
a variant in [13]	3D mesh, (4D hypercube at inter-processor level)	hexagonal-shape module, composed to a ball.	not specified	wireless	conceptual
this paper	3D mesh (3D torus with external data channels)	- as [13], investigating module's coupling and large-scale composition	3D power grid	direct via mated connectors	hexagonal board and frame prototype

III. DESIGN

Naturally, the tiled system has its own constraints, and its own unique properties. One of the most important is the notion of a decentralized power grid property, rather than parallelized backplane power bus, but there are others. Therefore, investigating the feasibility of such systems and understanding those properties and constraints is the key concern of this research. The goal is to determine if such systems are physically feasible when extended to large scale systems. Questions we particularly wish to answer in this research challenge include:

- Can a collective power grid sustain systems of large scale?
- Are we able to manage and predict power behaviors?
- Can such a system feasibly be physically constructed?
- Can workloads be varied node by node to optimize power distribution and computational throughput across a TCA?

Our work in some of these areas, as reported here, are a progression toward answering these questions individually and collectively.

IV. RELATED WORK

In this section, we provide some of relevant previous work surveyed concerning inter-board level packaging technologies.

A. Rack-mount Packaging

We briefly mention some parallel computers built with rack-based packaging in this category as it is considered a traditional inter-board level method. Recently, a number of projects have targeted large computing system challenges to achieve the next step of computing power at a minimum of billion-billion floating-point operations per second, i.e., exascale. ExaNeSt [1,2], ExaNoDe [3], ECOSCALE [4], and EuroEXA [5], are four example projects closely collaborating for the purpose. ExaNeSt focused on developing interconnection networks, storage, and cooling. The project employed the cooling system of ICEOTOPE [6]. The electronic circuit boards were submerged in warm non-conductive (dielectric) liquid flowing into and out of each of the blades contained in a rack. Another recent parallel computer was Supercomputer Fugaku [7]. The machine achieved the first rank in High Performance LINPACK (HPL) benchmark on TOP500 project [8], which was also built on rack-based packaging.

B. Non rack-mount Packaging

The packaging techniques in this subsection are more directly relevant to our work as they share some common configuration with our design. Thus, our work is considered a subset of this category. HAEC [9], was a project proposing a holistic energy-efficient computing system with both optical and wireless communication. In the project, a group of boards was named as HAEC Box. Another design of this category was conceptualized with a wireless computing system [10], to mitigate the complexity of data communication wiring, heat dissipation, power lines, and system composition effort. Afterwards, [11,12] further investigated the techniques. In [12], a level of abstraction of wireless interconnection network was designed for the concept. Dedicated simulation and visualization tools were also built to evaluate the performance of the wireless system behavior, it was concluded that at the time of the research, technologies of radio devices still consumed a large amount of energy, leaving a challenge space for lower energy improvements. For performance analysis of [12], it was reported that a reasonable performance can be achieved on particular tasks executed on certain networks.

Subsequently, [13], proposed a variant of the concept. The packaging technique allowed cooling fluid to pass through a level of composition in order to dissipate heat from each unit. For the packaging in [13], we envisaged the feasibility of two alternative designs of both wired and wireless communication are possible to implement. For wireless communication, transceivers can be embedded in the smallest unit. On the other hand, in a wired design each edge of the unit can be used as an interfacing area for both data communication and power lines routed into the internal components.

The power-route network enables a node to tolerate some faulty power-route situations. With a single unit added to the system, it provides the diversity of both powering and data communication networks. With such a method of powering nodes in the system, a challenge regarding electrical constraints emerges, which does not exist in traditional rack-based systems. A survey and comparison of related technologies is given in Table I.

To investigate how practical the TCA is, in terms of physical scalability prior to a concrete implementation phase of a large system, work reported in this paper focuses upon wired communication for simplicity in our first investigation. Nevertheless, an ultimate choice of mixing both the methods is potential to aggregate their strengths to obtain an optimal design in future research.

The TCA concept relies upon a fundamental building block – the ‘hex-tile’, and abstract views of which are shown in Fig. 1. Each tile is a hexagonal planar structure, with edges having alternate angles, as illustrated in Fig. 1a. The space inside a tile may contain power-conversion units, computing, and communication elements such as CPU, memory unit, and a router, as illustrated by Fig. 1b. Power and ground lines and physical data channels can be routed via each of the six edges, creating the IO connectivity showing in Fig. 1d. IO lines typically act as independent point-to-point channels, while all tile power inputs are shared via common rails within each tile.

Meanwhile, each tile is capable of joining to other tiles via the angled edge connectors, permitting a number of tiling schemes, including 2D planar tiling, and 3D topologies, including a ball-like structure comprising 8 tiles. Fig. 1e shows the shape when tiles are formed into a ball (a truncated octahedron, also known as a tetrakaidecahedron, or Kelvin Bubble [24] and Fig. 1f shows an actual equivalent prototype tile structure (more discussion of which later). Balls may then also form tileable structures, tiling in 3 dimensions via the trapezoidal faces of the structure, as shown in Fig. 2d. A system of 3x3x3-ball size is illustrated in Fig. 2a.

Interestingly, provided that the hexagonal tile edges are ‘equilateral and equiangular’, the space between tiled balls may be filled with a second 3D grid of balls, similarly interconnected, as illustrated in Fig. 2b. This agrees with the principle of packed truncated octahedrons [24]. While single packed arrays have up to (n^3) nodes in cubic space, a doubly-packed array can approach almost twice that of a singular array for large dimensions of n , with up to $(n^3) + (n - 1)^3$ nodes.

The outer balls of the array present trapezoidal connection points to be used in the most convenient power delivery arrangement. The most aggressive approach is to connect power and ground lines to all connectors available at the outer perimeter of the system. This allows the best-possible electrical current delivery and distribution throughout the system, but with a considerable degree of redundancy in power connections. The total number of connections could however be reduced significantly while maintaining a viable power grid.

Obviously, with the unique power network topology of a ball-grid, the voltage, current, and power delivery available to tiles in the different locations within a structure will vary to a degree, affected by connector-pin resistances, power capacity, and the overall collective power grid pathways. Moreover, connector-pin currents are also a special concern as the current-flow network are not obvious compared to those in rack-mount systems, and pin power/current carrying capacities have upper limits that must be respected. These concerns introduce the unique challenge of the electrical constraints, and thus ultimately a need to predict such behaviors within a dynamically work-loaded system.

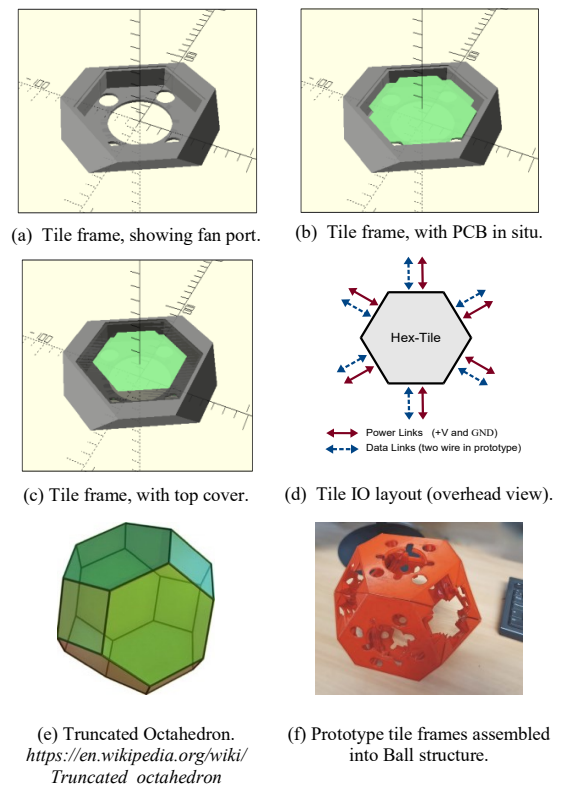


Fig. 1. Illustrations of the hex-tile. (a) shows a 3D model of the tile module frame, plastic or ceramic package material, (b) shows the tile frame with PCB or MCM in situ, (c) shows the prototype module top plate, (d) shows the IO connectivity of each tile edge, where solid/red arrows representing power and ground lines, and the dashed/blue lines are data channels, (e) shows the ball arrangement when 8 tiles are combined (forming a truncated octahedron with hexagonal and trapezoidal faces), and (f) shows an actual prototype tile-frame combination of 8 tiles into a ball (shows unpopulated tile frames).

C. Electrical Constraints

It is essential to ensure that all of the tiles in the system can operate without violating any electrical constraints, as defined within the specifications of their connectors and components. In this paper, we define three key constraints as follows.

- The regulator output-load voltages are regulated at the specified levels.
- The board input-voltages are in the operating ranges specified by the power-conversion units.
- The connector-pin currents do not exceed the levels specified by the limits of the connectors.

Practically, a system can be heterogeneously designed, composed of different tile types (SSD, Memory, CPU, FPGA, DSP, TPU to name a few). Thus, each tile may contain different load requirements, power-conversion units, and the limits of connector currents. However, in this first analysis report, we assume all the tiles are of uniform type, and the load resistances are steady, with constant power load.

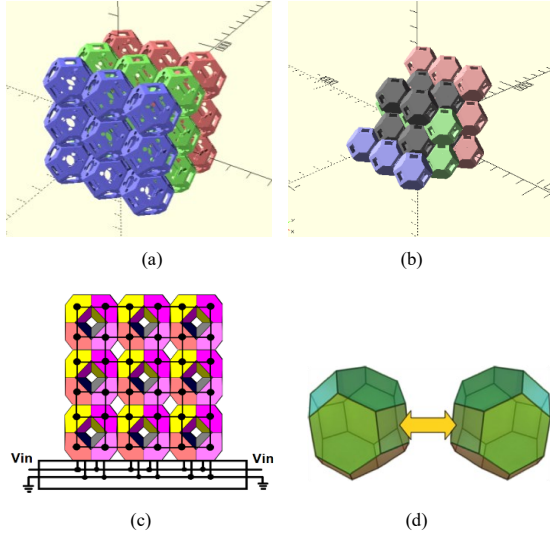


Fig. 2. (a) Visualized TCA system of 27 balls with the dimension of (3,3,3) in isometric view. (b) Example of ‘intra-grid packing’, $2 \times 2 \times 2$ gray balls can pack in between the existing $3 \times 3 \times 3$ grid (cutaway view). (c) Top view of array shown in (a) illustrating power distribution. (d) Trapezoid edges as inter-ball connection points. In theory, all external tile edges forming trapezoid ball faces can be connected to power sources, evenly distributing the power inputs.

D. Models

Most power delivery models assume a bus and tree-like power distribution network, unlike the scheme employed in TCA. The nearest model in [17] is an example in our survey that holds a similar idea in terms of circuit components we expect for simulation. However, that power model is applied in the large-scale integration (VLSI) design level, where power grids are common. To give more details concerning our survey, in [18], a large amount of power and energy models related to HPC systems have been surveyed and classified in terms of system components. In their survey, we found that researchers paid more interest to the power modeling of either nodes, interconnects, or the whole system, rather than how power-delivery mediums are modeled. For this reason, we decided to design our own circuit model and simulation tool for our constraint evaluations, and ultimately to validate this against a real physical prototype. This is described as follows:

1) *Pin-resistance model*: Due to the cascading effect of connectors in the envisaged power grid, it is important to evaluate how the bulk conductor and contact resistances of connector pins impact on the scalability of the TCA system, thus a suitable model is required. In Fig. 3, the connectors, and their respective resistance models are depicted. Apart from the fairly constant bulk resistance of a pin, the contact resistance is also an important factor of the stability of the system, and can vary under several conditions. These quantities can be observed by measurement or obtained from the connector datasheet, if provided. In our model, we use a single lumped resistor, named r_{p_resist} to model either a single tile-edge power pin, or collectively model multiple power-pins, if used in the same connector (power pins can optionally be doubled up in parallel to give higher current capacity).

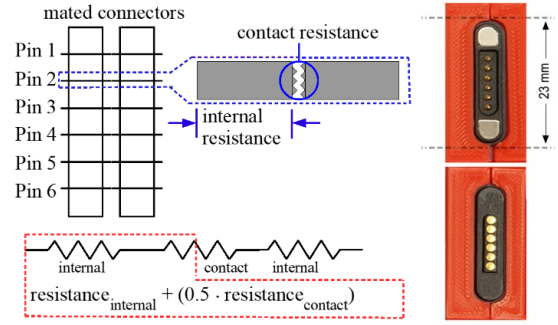


Fig. 3. Magnetic connector pair, and individual pin resistance modeling detail. Resistance r_{p_resist} (in the red, dashed frame) comprises the bulk internal pin resistance and a 50% share of pin-mating contact resistance as defined earlier.

Thus, a parallel-resistor calculation can be simply applied to assign a single resistance value to this r_{p_resist} . For a ground pin, an equivalent single resistor is named r_{g_resist} . All of the currents passing through these resistors will be collated for connector-constraint evaluations.

2) *Board model*: The inclusion of a switching regulator circuit model in our tile prototype results in excessively long simulation times for a large system. Thus, we sought a simplified model to evaluate the entire system in a steady state with constant regulator load(s). It was noted that [19] provides several average-model methodologies, and [20,21] also automate the modeling processes of an average model for switching regulators. It was determined that the curve fitting method was adequately effective for a simplified model to evaluate the system in its steady state while dramatically reducing simulation times (by a factor of hundreds), and result file size, without significant loss of accuracy (typically less than 1% for tested cases). The simulator tools can select and use either approach according to accuracy and time constraints.

In Fig. 4, the board model, as a subcircuit of the tile, can be depicted in the inner hexagon. the resistor at the center, $board_resistance$, represents the varying instantaneous equivalent-resistance of the entire board. The adjuster unit imitates the operation of a switching regulator, periodically samples both the input voltage, vin_s , and current, I_{board_s} , of the board. The adjuster adapts the value of $board_resistance$ when a sampled board input-current is not ‘close enough’ to the expected instantaneous input-current, I_{board_es} , as shown in (1). The parameter I_{diff_thres} (input-current difference threshold) controls this alignment, resulting in the accuracy of the simulation results. When the difference between the sampled and the expected input-currents, I_{diff} , shown in (2), is within the interval of $(-I_{diff_thres}, I_{diff_thres})$, the adjuster maintains $board_resistance$ value. Once every $board_resistance$ in the system is stable, the entire system reaches the steady state. At this point of simulation, all the connector-pin currents, board input voltages, and currents can be collected for constraint

evaluations. The parameter tr_init sets the period of the initial resistance before the step resistance, R_{step} , takes the role of gradually altering board_resistance. The actual power and ground materials in the board may differ from this abstract simulation model. In our simulation model, we consider the trace resistances in these power and ground lines are negligible. The complexity of the curve fitting method used to profile the steady state of the board also affects both the simulation times and the accuracy of simulation results. In this paper, we consider the polynomial fitting of degree three is adequate for our evaluations. The equations regarding the board model are given in (1) and (2). The equations of the board model can be implemented in an LTspice [22] simulation file using built-in symbolic sample-and-hold function blocks. In the simplified model, the parameter initial resistance may impact both the time required for the LTspice simulator to achieve the DC operating point and the simulation time before every tile reaches the steady state. This can be seen at the simulation time of approximately 120 μ s in Fig. 5.

$$I_{board_e} = p_1 v_{in_s}^3 + p_2 v_{in_s}^2 + p_3 v_{in_s} + p_4 \quad (1)$$

$$I_{diff} = I_{board_s} - I_{board_e} \quad (2)$$

where,

v_{in_s}	Sampled instantaneous board input-voltage.
$P_{1..4}$	Coefficients of curve-fitting equation for a constant regulator-load power.
I_{board_e}	Expected instantaneous input-current at steady state.
I_{board_s}	Sampled instantaneous board input-current.
I_{diff}	Difference between I_{board_s} and I_{board_e}

E. Model Validations

The system with the simplified board model was validated against the 'complex' LT3976 spice model, with a $3 \times 3 \times 3$ array, and found to be averaging less than 1% margin of error for examined cases under load conditions. To simplify the validation model and shorten the simulation time, the soft-start mode of the regulator was disabled, and the load resistance was set to 1 Ohm, representing approximately 25 W being regulated at 5V (25W being the maximum permitted for this particular regulator). A single 12 V external voltage source is supplied to all the system-surface power and ground pin models. The initial value of the external (surface) was set at 0 V, then ramped up to 12 V. This permitted the LTspice simulator to more quickly achieve a DC operating point, reducing simulation times. Example parameter values and the LTspice code, as a part of adjuster unit are presented in Table III.

As noted in Table II (a) and (b), the complex model is compared to real prototypes for single tile and 8-tile ball. Simulator and hardware prototype results were found to closely agree in these, with typical agreement within the region of 1-2% for all simulated power-load cases (i.e., excluding no-load). Voltage stability across sample tile networks, as given in Table II (c), was excellent, and well below 0.5% where tiles are composed as a 2D or 3D tiled cases tested. As expected, group-tiled arrangements are more stable due to the parallelism and sharing of current paths across the power grid.

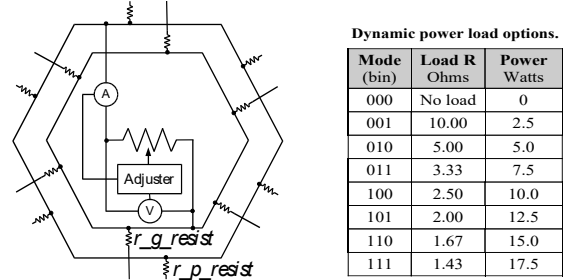


Fig. 4. Conceptual representation of tile model. As per the real tile, 'power consumption' above base load can be dynamically adjusted via a CPU-selectable load resistance. Additional CPU load (regulator 5V output ~ 25mW). Tile cooling fan (~60mA, 12V rail, ~700mW) is separately modeled.

TABLE II. ACCURACY OF THE 'COMPLEX' LT3976 MODEL SIMULATION, VERSUS ACTUAL PROTOTYPE AND SELECTED SYSTEM CHARACTERISTICS.

(a) Prototype/Model: Single tile, Single connector:

	Min (base) ~ 0W	Low +2.5W	Med +5.0W	High +10.0W	Max +17.5W
$I_p \pm 5mA$	60 mA	310 mA	540 mA	1000 mA	1760 mA
I_M	62.29 mA	310.82 mA	539.93 mA	1012.81 mA	1753.82 mA
Error (ave)	-4.5%	-0.3%	0.0%	-1.3%	0.4%
(min,max)	-13.3%, 4.2%	-1.9%, 1.3%	-0.9%, 0.9%	-1.8%, 0.8%	0.1%, 0.6%

(b) Prototype/Model: 8-tile ball, 2 co-located power connectors

	Min (base) ~ 0W	Low +2.5W	Med +5.0W	High +10.0W	Max +17.5W
$I_p \pm 5mA$	530 mA	2550 mA	4370 mA	8070 mA	14010 mA
I_M	501.67 mA	2493.57 mA	4328.48 mA	8121.95 mA	14079.9 mA
Error (ave)	5.3%	2.2%	0.9%	-0.6%	-0.5%
(min,max)	4.4%, 6.2%	2.0%, 2.4%	0.8%, 1.1%	-0.7%, -0.6%	-0.5%, -0.5%

(c) Prototype: grid stability (worst case voltage drop, 10W load, 12V supply)

Tiling	Configuration	Prototype
	1D tiling: 4 tiles, 1 connector	1.25%, 150mV
	2D tiling: 4 tiles, 1 connector	0.33%, 40mV
	3D tiling: 8 tiles, 2 connectors	0.17%, 20mV

TABLE III. LTSPICE EXAMPLE PARAMETERS

Example parameter values:
Initial resistance period: 6 Ohms, held for 21 us, then 0.005 Ohm steps
Example LTspice code with the above parameter values
<code>b_i_board i_board v = i(r_board_resistance)</code>
<code>b_i_diff i_diff 0 v = v(i_board_s) - ((-0.006025)*(v(in_s)**3) + 0.2087*(v(in_s)**2) - 2.623*v(in_s) + 14.39)</code>
<code>b_r_board r_board 0 v = if(time<21us, 6, if(v(i_diff) > 0.01, v(r_board_s) + 0.005, if(v(i_diff) < -0.01, v(r_board_s) - 0.005, v(r_board_s))))</code>

As shown in Fig. 5a, after the external voltage source reaches 12 V, the board input-voltages are at certain voltages. All the voltages are below 12 V, affected by the resistances of connectors located in different layers of the system. At this point, both the detailed and simplified models, Fig. 5a and 5b

respectively, continue to converge into the steady state, with both models very similar at 120-140 μ s.

F. Simulation Framework

In this paper, we focus on the feasibility of TCA, however, we also briefly describe the simulation framework to demonstrate how instances of the tile model can be composed into a complete system. To evaluate a large system means that a hierarchically complex resistor-network model needs to be generated and manually creating an LTspice simulation file is a tremendously labor-intensive task. Thus, we automate this process by building our own source-code file generators, which can generate a complete simulation model for any set of ball dimension parameters. The automation of LTspice code generation starts at the inter-ball level of composition. Firstly, inter-ball power and ground lines are locally named. The sequence of adding balls starts from the coordinates of (0,0,0), then follows the adding-rule of “X first, then Y, and Z”, fulfilling rows, planes and until the whole system topology is entirely generated (see Fig. 2a).

V. SIMULATION RESULTS

The number of balls in each dimension is parameterizable in our simulation framework, thus, arbitrary ball sizes of system can be generated. However, in this paper, only cube-shaped systems with 50 mOhms mated pin-pair resistance, with a single 12 V power source common to every surface connector, is evaluated and reported. Given that the individual pin current-limit is 3 Amps, power and ground pins are configured as doubled-up pairs, to permit up to 6 Amps. Fig. 6a and 6b show simulation results for multiple ball-array configurations, ranging from a single ball up to an $(n \times n \times n)$ array size of $n = 5$, with 125 balls and 1000 tiles. Power loadings per tile are varied from 5W to 25W (on the regulator output side).

It can be observed in Fig. 6a that, as expected, voltage drops across the power grid of each array will scale up as the cascaded effects of tile-to-tile pin connection resistances accumulate. In Fig. 6b, the maximum observed pin-currents across the grid are presented for the same range of ball-array configurations. Here it is observed that pin currents remain within the specifications of $I \leq 6$ Amps, until the cubic ball dimension reaches $n = 5$, at which point the pin current is exceeded in one or more pins across the array (for all tiles at full power load). However, this may well be happening in only a limited number of pins, and by moderating the power consumption on a tile-by-tile basis, for instance where some tiles operate at perhaps 20W rather than 25W, it should be possible to return pin currents to within specified limits.

An important advantage, therefore, of the availability of a modeling and simulation framework, is that it permits more advanced power management strategies to be explored. For example, a predictive power optimization model based upon a genetic-algorithm (GA) has already been implemented as part of this research work. The use of a genetic algorithm is more efficient than a brute-force approach for large array sizes. Employing more sophisticated power distribution strategies, coupled with a 3D visualization capability, which is also being developed, system behaviors and optimization strategies can be explored more deeply.

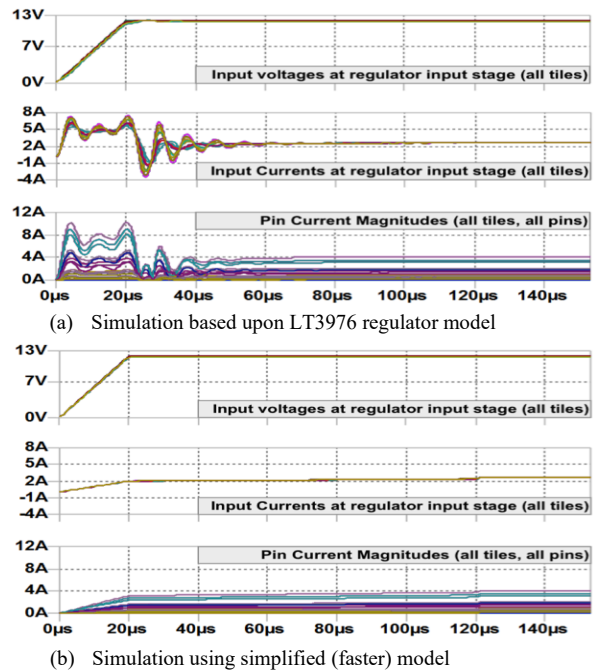


Fig. 5. Validation of the simplified board steady-load model with 3x3x3-ball system (27 balls, 216 tiles), showing (a) the LT3976 model, and (b) with the simplified (much faster) simulation model.

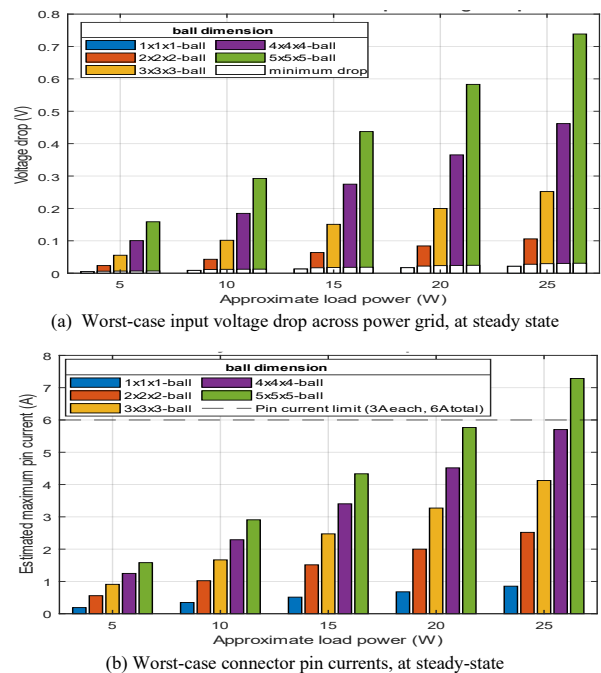


Fig. 6. Constraints simulation results (with 101mA assumed supply side 12V fan load in this case). (a) Estimated maximum board input-voltage drop and (b) Estimated maximum pin-currents for different load-power allocations and system sizes.

For example, Fig. 7 shows the visualization of the result of the experimental GA power optimization. It is observed that initial power loading distribution, as shown in Fig. 7a has been significantly improved in GA-enhanced loading of Fig. 7b, after GA algorithm converged to within a set margin of optimality.

VI. PROTOTYPE SYSTEM

To explore and validate feasibility, and simulator accuracy, a hardware prototype has been developed, examples of which are showing in Fig. 8 in various levels of assembly and operation. Each prototype tile utilizes an LT3976 power regulator, onboard ATMEGA324PB microcontroller, acting mainly as a ‘house-keeping’ control node, data IO intermediary, and also able to dynamically control a dummy power load, emulating heavier power usage at the tile level. Magnetically coupled 6-pin power/IO connectors (as shown in earlier Fig. 3) permit tile-to-tile connection, with two IO lines, two positive supply pins and two ground rails (to achieve 6A current capacity). Current prototypes include a complete 8-tile ball, a base mounting platform, and relocatable surface power leads.

The system has been tested with dynamic power ranging across tiles up to maximum system power loading. Fig. 8d shows a snapshot (from video) of a ball (two tiles removed to permit interior view) under test conditions with power loading dynamically stressed across the grid. Onboard cooling for these prototypes is achieved via an air-flow fan (visible in Fig. 8a). At Maximum power load (17.5W for prototype power configuration options), with 14 Amps supplied to the ball, an interior air-space temperature of around 15c above ambient (~21c) was observed. This power loading could also be achieved by hosting a suitable CPU in the extension socket, with similar results.

VII. FURTHER HARDWARE DEVELOPMENT DIRECTIONS

The prototype is necessarily over-sized given the construction methods available. However, the ultimate goal would be to reduce tiles to something of the order of 50-60 mm major planar dimension, and to utilize ceramic chip packaging technology to encapsulate single SOC or MCM modules representing processors, SSD, memory, or power reservoirs. A custom connector redesign should be able to accommodate this form factor with similar, and potentially improved power capabilities and significantly better IO options.

There are also possibilities to manufacture the balls as complete components and use these as the fundamental building blocks, with the same principles applying at a coarser granularity. Combination with liquid cooling systems would then be envisaged, as investigated in previous related work [10]. At this level of physical size, individual tile cooling would be dropped, and air/fluid flow-assistance via inter-ball modules located and interspersed at the trapezoid connectors would permit controllable dynamic (fluid or air) flow control across any array topology constructed. This concept is illustrated in Fig. 9. Notably, even in the case of the double-packed array of Fig. 2b, the cooling model still supports appropriate capacity to remove heat, since the cooling network is duplicated as two independent flow networks in the two interleaved arrays, with proportionate increase in cooling.

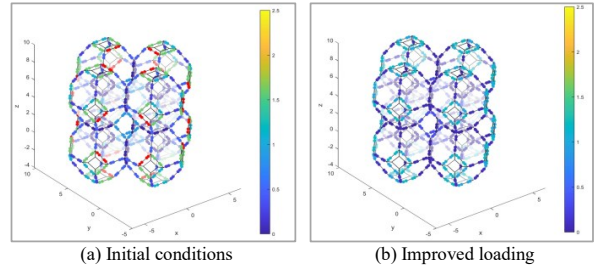


Fig. 7. 3D edge-connector current visualization for a 2x2x2 ball array (64 tiles). The colored dots represent edge-connector pin currents (colored blue through to yellow for normal loadings). Red dots highlight exceeded pin current locations. The genetic algorithm achieves better power distribution within the grid by changing the power utilization on each tile while maintaining the overall target power consumption (and thus computational capacity).

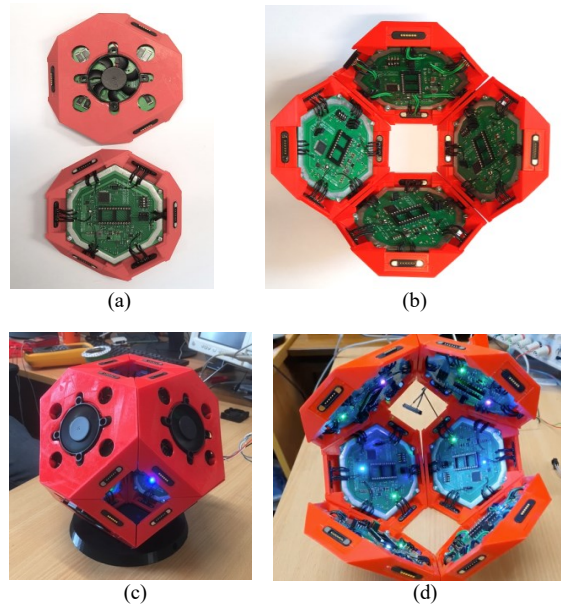


Fig. 8. (a) Hex-tile prototype (top and reverse views), (b) Four hex-tiles linked into a half-ball (petal) formation, (c) Eight tile-frames comprising a ball with a base-plate, with trapezoidal connection faces visible, (d) A ball, powered-up with shared power distribution between tiles (top two tiles removed for interior view, LED colors relate to power loading).

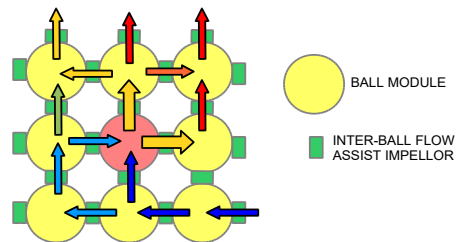


Fig. 9. Inter-ball fan/pump/impeller and air/liquid flow control principle. This shows a 2D view, but in practice would operate in 3 dimensions to modulate thermal flow dynamically under system monitoring and control. Cool air/fluid flows into the grid (blue) accumulates heat transfer in a directed fashion according to localized need, and exits system (Red). Existing prototypes can already operate in a similar mode, though in a less optimal fashion.

VIII. CONCLUSION AND FUTURE OPPORTUNITIES

A concept for extensible 3D processor array topologies has been presented, comprising of a novel hexagonal tile design, permitting the assembly into modular truncated octahedron ‘ball’ modules, and which may be combined into larger scale arrays in a variety of topologies. It has been demonstrated that power delivery within the unusual structure, and particularly the distributed power distribution, is a workable model and may be effectively predicted and managed.

A physical prototype has been briefly described and demonstrates that the system concept is practically realizable. The behavior of the prototype hardware was found to be typically within a few percent of simulation predictions, suggesting that the simulation model is representative of similar systems at larger scales, and that engineering constraints involving power and current densities may be identified and managed appropriately.

This work, alongside others [10,11,12,13] takes an important step toward the realization of large-scale systems based upon tiled modules without host-system circuit board and rack-mount architecture overheads and constraints. To progress further there are several avenues for this concept to be pursued. The use of optimal workload balancing across a topological array, in order to manage optimal power distribution versus workload throughput, dynamic power and thermal management strategies, and the exploration of thermal management technologies including airflow and fluid systems. Communications channels are currently physical point to point. However, work has already been done in the field in relation to short near-field communications at high data rates using localized wireless data links, with point-to-point, multicast and broadcast potentials.

Meanwhile, the level of integration and physical size of the hex-tile requires a further step-change. Ultimately, the basic building block may be a smaller tile, or a complete ball on smaller scales. Such modules would likely utilize relatively well-established manufacturing technologies: Ceramic chip packaging materials and custom chip-carrier designs, employing single-chip systems with complete processors, memory, storage, routing

As these areas are advanced incrementally, the authors expect to see feasibility of large-scale tiled arrays becoming greatly improved, ultimately moving toward realizable commercial systems.

ACKNOWLEDGEMENTS

1. This work was supported by Royal Thai Government NSTDA PhD scholarship (OEA ID No. ST_G5599).
2. The authors also wish to thank the University of York, Department of Computer Science, technical team for their assistance with 3D printing and board assembly work.

REFERENCES

- [1] M. Katevenis et al., "The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems," 2016 Euromicro Conference on Digital System Design (DSD), 2016, pp. 60-67, doi: 10.1109/DSD.2016.106.
- [2] R. Ammendola et al., "The Next Generation of Exascale-Class Systems: The ExaNeSt Project," 2017 Euromicro Conference on Digital System Design (DSD), 2017, pp. 510-515, doi: 10.1109/DSD.2017.20.
- [3] A. Rigo et al., "Paving the Way Towards a Highly Energy-Efficient and Highly Integrated Compute Node for the Exascale Revolution: The ExaNoDe Approach," 2017 Euromicro Conference on Digital System Design (DSD), 2017, pp. 486-493, doi: 10.1109/DSD.2017.37.
- [4] I. Mavroidis et al., "ECOSCALE: Reconfigurable computing and runtime system for future exascale systems," 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016, pp. 696-701.
- [5] "EUROEXA." euroexa.eu. <https://euroexa.eu> (accessed Apr. 26, 2022).
- [6] "Precision immersion cooling from the Cloud to the Edge | Iceotope." iceotope.com. <https://www.iceotope.com> (accessed Apr. 26, 2022).
- [7] "Fugaku | RIKEN Center for Computational Science RIKEN Website." <https://www.r-ccs.riken.jp/en/fugaku> (accessed Apr. 26, 2022).
- [8] "November 2021 | TOP500." top500.org. <https://www.top500.org/lists/top500/2021/11> (accessed Mar. 29, 2022).
- [9] G. P. Fettweis et al., "Architecture and Advanced Electronics Pathways Toward Highly Adaptive Energy-Efficient Computing," in Proc. of the IEEE, vol. 107, no. 1, pp. 204-231, Jan. 2019, doi: 10.1109/JPROC.2018.2874895.
- [10] J. Austin (Cybula Ltd), "Computing Devices" GB Patent No. GB02/04104, September 10, 2002.
- [11] R. Hind, "Feasibility Study on implementing the "Ball Computer"," M.S. thesis, Dept. Comput. Sci., Univ. of York, York, 2013.
- [12] A. M. Kamali Sarvestani, "Evaluating Techniques for Wireless Interconnected 3D Processor Arrays," Ph.D. thesis, Dept. Comput. Sci., Univ. of York, York, 2013.
- [13] A. M. Kamali Sarvestani, C. Crispin-Bailey, and J. Austin, "Performance Analysis of a 3D Wireless Massively Parallel Computer," J. Sensor and Actuator Networks, vol. 7, no. 2, pp. 1-20, Apr. 2018, doi:10.3390/jsan7020018.
- [14] A. Louri and H. Sung, "An optical multi-mesh hypercube: a scalable optical interconnection network for massively parallel computing," in Journal of Lightwave Technology, vol. 12, no. 4, pp. 704-716, April 1994, doi: 10.1109/50.285368.
- [15] A. Louri and Hongki Sung, "3D optical interconnects for high-speed interchip and interboard communications," in Computer, vol. 27, no. 10, pp. 27-37, Oct. 1994, doi: 10.1109/2.318581.
- [16] J. Navaridas, J. Lant, J. A. Pascual, M. Luján, and J. Goodacre, "Design Exploration of Multi-Tier Interconnection Networks for Exascale Systems," 2019. doi: 10.1145/3337821.3337903.
- [17] Z. Zhang, X. Hu, C. Cheng and N. Wong, "A block-diagonal structured model reduction scheme for power grid networks," 2011 Design, Automation & Test in Europe, 2011, pp. 1-6, doi: 10.1109/DATE.2011.5763014.
- [18] K. O'brien, et al., "A Survey of Power and Energy Predictive Models in HPC Systems and Applications," ACM Comput. Surv., vol. 50, no. 3, Jun. 2017, doi: 10.1145/3078811.
- [19] C. P. Basso, Switch-Mode Power Supplies, 2nd ed. New York, NY, USA: McGraw Hill Education, 2014.
- [20] M. H. Leonard, "Automated Behavioral Modeling of Switching Voltage Regulators," B.S. Thesis, Dept. Elect. Eng., Univ. Arkansas, USA, 2013.
- [21] M. H. Leonard, "Semi-Automated Switching Regulator Modeling Method and Tool," M.S. Thesis, Dept. Elect. Eng., Univ. Arkansas, Fayetteville, AR, USA, 2015.
- [22] "LTspice Simulator | Analog Devices." analog.com. <https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html> (accessed Apr. 26, 2022).
- [23] Linear Technol. Corporation, Milpitas, CA, USA. *LT3976 - 40V, 5A, 2MHz Step-Down Switching Regulator with 3.3µA Quiescent Current*, (2013). Accessed: Apr. 26, 2022. [Online]. Available: <https://www.analog.com/media/en/technical-documentation/data-sheets/3976f.pdf>
- [24] William Thomson. "On the division of space with minimum partitional area." Acta Math. 11 121 - 134, 1887-1888. DOI: 10.1007/BF02612322.

A.2 Short Report for Internal Funding Award (Hex-tile Project)

Chris Crispin-Bailey, Pakon Thuphairo, Anthony Moulds, Jim Austin.

University of York, Department of Computer Science

Version 1.00 18-08-2022

1. Executive Summary:

This document is an initial short summary of the work undertaken by the authors in completion of the hex-tile prototype demonstrator project. It is therefore concise in detail. A more detailed report and/or research publications will arise in time to supplement this report.

A novel hardware design concept for scalable computing/processor arrays has been translated into a real-world physical prototype, which has been constructed and tested successfully.

The work undertaken demonstrates the feasibility of such systems (known as tiled computing or ball computers in related research). Specific contributions include the following:-

1. *The realisation of a functioning system prototype.*
2. *Support to validation of PhD work on simulators for the same platform concepts*
3. *Contributions to peer reviewed research output(s)*
4. *Increased credibility for research funding proposals to achieve follow-on goals.*

This work was undertaken on a limited budget (approximately £500 departmental funds, plus a similar amount of miscellaneous spending from PHD and Personal research budgets). The low cost compared to the outcomes demonstrates the high value of selectively priming funds, and the support provided by Mr Moulds, as a dedicated research technical officer is an essential factor in the success of this project.

2. Dissemination to date:

Currently this work has been disseminated in the following outputs:

- a. This mini-report
- b. Research conference paper and proceedings DSD2022¹.

¹ **Investigating Novel 3D Modular Schemes for Large Array Topologies: Power Modeling and Prototype Feasibility**, Conference proceedings of the 2022 DSD conference, Euromicro DSD 2022 Conference proceedings, maspalomas, Gran Canaria, DOI and page numbers to be confirmed (8 pages).

3. Technical Summary

The general concept of hexagonal computing tiles, and their composition into tileable 'balls' (actually truncated octahedrons) is well documented in prior publications involving the authors [1,2,3,4].

The basis for the idea is that a planar hexagonal tile, with edges bevelled to specific angles, and alternating positive or negative angle, can form tileable structures, capable of one, two and three dimensional tiling properties. Specific three-dimensional (3D) topologies include the creation of a closed surface, equivalent to a ball, and more accurately a truncated octahedron. This structure is also recognised as a kelvin bubble as per Lord Kelvin's 1888 publication on packing of three dimensional volumes [6].

The intention of using these shapes as computing elements is that they may be interconnected without further circuit-board and hierarchical structures. Traditional large scale structures rely upon rack-mount cabinets, in which a number of circuit boards (blades) are slotted in a common back-bone structure for power and data connectivity. The authors observed that this means that a significant majority proportion of the volume utilised by these systems is this supporting structure, and a very small proportion of the volume is actual processing elements such as CPU chips. Indeed cpu chips themselves are of the order of cubic-centimetre magnitudes of volume and size, yet processor arrays often require tens of cubic centimetres of volume per CPU, a significant detriment to theoretical performance density of systems and also a cost factor.

Of course, traditional processor array structures have a number of practical feasibilities which make them the standard solution. Alternatives are not well understood, and hence the purpose of the Hex-Tile project is to demonstrate new and deeper understanding, and feasibility of such alternative systems.

Figures 3a-d show various examples of the hardware constructed, including the major components (Figs 3a), the assembled tile (Fig 3b), and two 3D structures including a petal (Fig 3C) and a ball (Fig 3d).

The hardware was designed collectively by Moulds, Thuphairo, and Crispin-Bailey, and Anthony Moulds undertook the PCB design and contract fabrication work to realise the PCB and its assembly. Dr Crispin-Bailey designed and implemented the plastic tile-frame structure using 3D rapid prototyping resources available within The Department of Computer Science.

Assistance from technical support staff, particularly Pete Cooper and John Mowbray is acknowledged and appreciated.

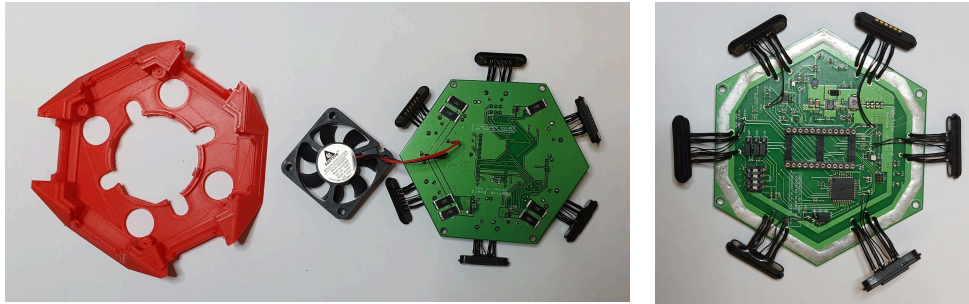


Fig 3.a. Examples of major hex-tile components (Tile frame, cooling fan, and main PCB assembly, with edge connectors assembled).

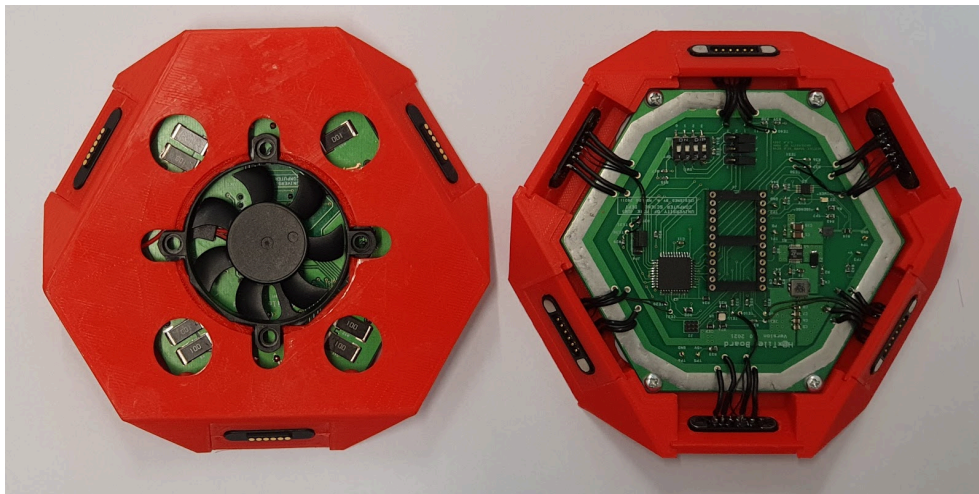


Fig 3b. Assembled Tile (left: underside, Right: Upper side). Note the bevelled edge orientations.

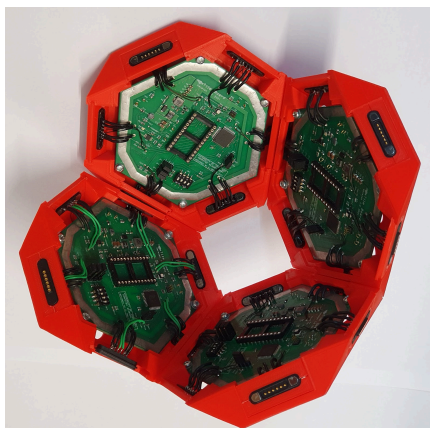


Fig 3c. 3D petal Structure.



Fig 3d. Ball structure on base plate

4. PRELIMINARY TESTING

Examples of test data are given in Fig 3.e below, including comparison to simulated system behaviour. System behaviour with 8 tiles powered up in a ball configuration have been gathered and performance is within feasible ranges. This demonstrates that a real system prototype can be constructed.

TABLE II. ACCURACY OF THE ‘COMPLEX’ LT3976 MODEL SIMULATION, VERSUS ACTUAL PROTOTYPE AND SELECTED SYSTEM CHARACTERISTICS.

(a) Prototype/Model: Single tile, Single connector

	Min (base) ~ 0W	Low +2.5W	Med +5.0W	High +10.0W	Max +17.5W
$I_P \pm 5mA$	60 mA	310 mA	540 mA	1000 mA	1760 mA
I_M	62.29 mA	310.82 mA	539.93 mA	1012.81 mA	1753.82 mA
Error (ave)	4.5%	0.3%	0.0%	1.3%	-0.4%
(min, max)	13.3%, -4.2%	1.9%, -1.3%	0.9%, -0.9%	1.8%, 0.8%	-0.1%, -0.6%

(b) Prototype/Model: 8-tile ball, 2 co-located power connectors

	Min (base) ~ 0W	Low +2.5W	Med +5.0W	High +10.0W	Max +17.5W
$I_P \pm 5mA$	530 mA	2550 mA	4370 mA	8070 mA	14010 mA
I_M	501.67 mA	2493.57 mA	4328.48 mA	8121.95 mA	14079.9 mA
Error (ave)	-5.3%	-2.2%	-0.9%	0.6%	0.5%
(min, max)	-4.4%, -6.2%	-2.0%, -2.4%	-0.8%, -1.1%	0.7%, 0.6%	0.5%, 0.5%

(c) Prototype: grid stability (worst case voltage drop, 10W load, 12V supply)

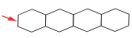
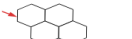
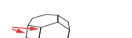
Tiling	Configuration	Prototype
	1D tiling: 4 tiles, 1 connector	1.25%, 150mV
	2D tiling: 4 tiles, 1 connector	0.33%, 40mV
	3D tiling: 8 tiles, 2 connectors	0.17%, 20mV

Fig 3e (data table extract from [5]), showing test data for 1D, 2D and 3D tiling scenarios.

The Prototype ball was able to operate at power levels up to 180W whilst maintaining a viable interconnection network within connector specifications and thermal limits. Indeed cooling in the prototype was highly efficient.

5. REFERENCES

- [1] J. Austin (Cybula Ltd), "Computing Devices" GB Patent No. GB02/04104, September 10, 2002.
- [2] R.Hind, "Feasibility Study on implementing the "Ball Computer", "M.S. thesis, Dept. Comput. Sci., Univ. of York, York, 2013.
- [3] A. M. Kamali Sarvestani, "Evaluating Techniques for Wireless Interconnected 3D Processor Arrays," Ph.D. thesis, Dept. Comput. Sci., Univ. of York, York, 2013.
- [4] A. M. Kamali Sarvestani, C. Crispin-Bailey, and J. Austin, "Performance Analysis of a 3D Wireless Massively Parallel Computer," J. Sensor and Actuator Networks, vol. 7, no. 2, pp. 1-20, Apr. 2018, doi: 10.3390/jsan7020018.
- [5] Investigating Novel 3D Modular Schemes for Large Array Topologies: Power Modeling and Prototype Feasibility, Conference proceedings of the 2022 DSD conference, Euromicro DSD 2022 Conference proceedings, maspalomas, Gran Canaria, DOI and page numbers to be confirmed (8 pages).
- [6] W. Thomson. (Lord Kelvin). "On the division of space with minimum partitional area," Acta Mathematica, vol. 11, pp. 121–134, Mar. 1887.

A.3 Hex-tile PCB Layouts

(Produced by Anthony Moulds, Senior
Research Officer, University of York)

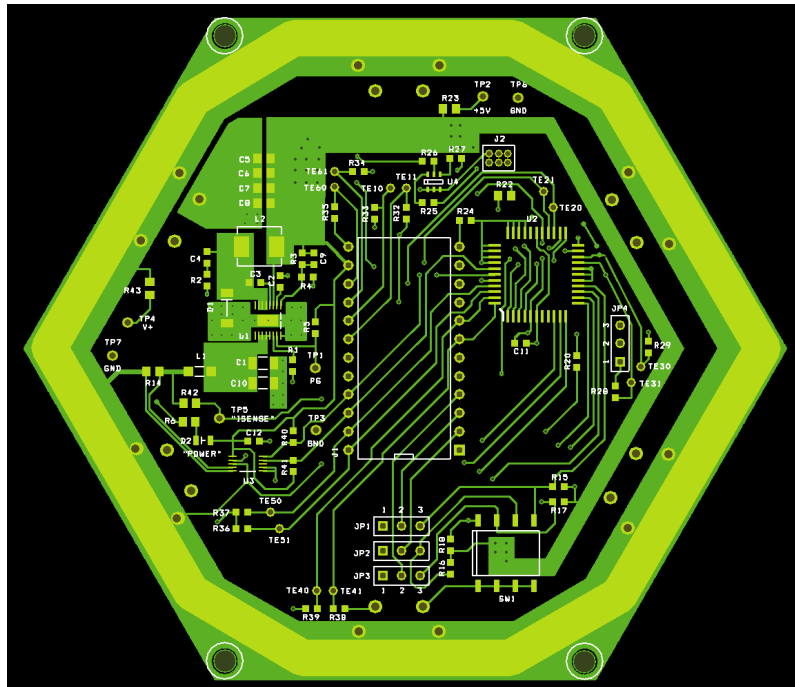


Figure A.1: Hex-Tile Board Topside, showing main component layouts.

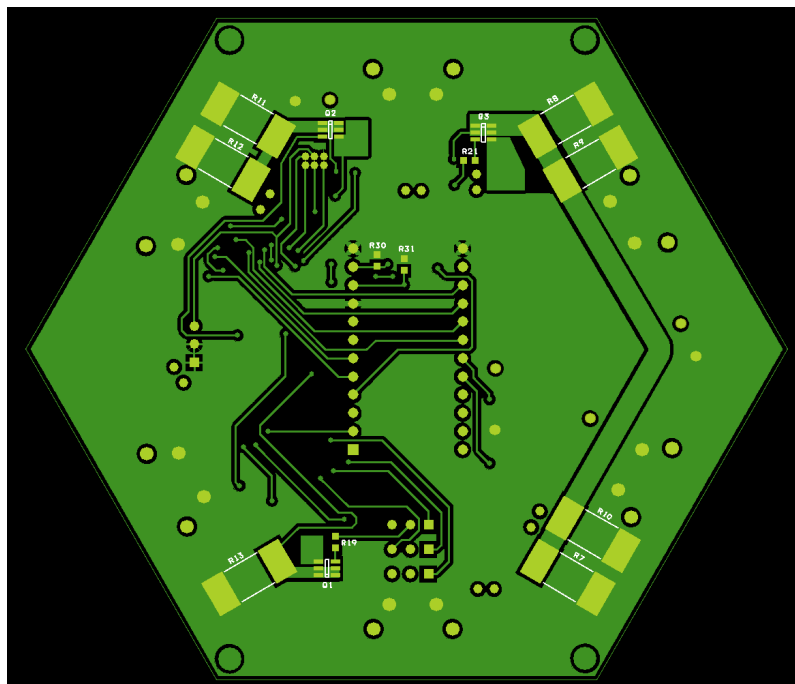


Figure A.2: Hex-Tile Board Underside, showing Power resistor PCB pads (Larger square areas), and other component layouts.

B

Example SPICE File

Example template SPICE-file auto-generated by the toolset, with small post-generation amendments by user.

Details:

- ▶ Uniform simulation framework.
- ▶ Size of 2x2x2 balls. (64 tiles)
- ▶ After changing the ball subcircuit names and their parameters by user.
- ▶ The placeholders 'x.y' can be assigned by the uniform simulation framework for parallel-simulation instances.

*A uniform sim for a 2x2x2-ball system.

```
.include  
\3D_TileSimulator\src\ngspice_src\ball\ng_ball_assign_board_resist_c  
tno_powerNOFAN_v2.cir
```

```
v_src Vsrc 0 x.y
```

```
*ball_system
```

```
X_0_0_0 p0A_0_0_0 n0A_0_0_0 p2A_0_0_0 n2A_0_0_0 VSrc 0 p0B_0_0_0  
n0B_0_0_0 VSrc 0 p4B_0_0_0 n4B_0_0_0 p0C_0_0_0 n0C_0_0_0 p2C_0_0_0  
n2C_0_0_0 p4C_0_0_0 n4C_0_0_0 p0D_0_0_0 n0D_0_0_0 VSrc 0 VSrc 0  
p1E_0_0_0 n1E_0_0_0 VSrc 0 VSrc 0 VSrc 0 VSrc 0 p5F_0_0_0 n5F_0_0_0  
p1G_0_0_0 n1G_0_0_0 VSrc 0 p5G_0_0_0 n5G_0_0_0 VSrc 0 VSrc 0 VSrc 0  
ng_ball_assign_board_resist_ctno_powerNOFAN_v2 con_resist={0.0125}  
tileA_board_r_val={x.y} tileB_board_r_val={x.y}  
tileC_board_r_val={x.y} tileD_board_r_val={x.y}  
tileE_board_r_val={x.y} tileF_board_r_val={x.y}  
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}  
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}  
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}  
tileH_rload={x.y} ball_i_diff_thrs={x.y}  
X_1_0_0 p0A_1_0_0 n0A_1_0_0 VSrc 0 VSrc 0 p0B_1_0_0 n0B_1_0_0  
p4C_0_0_0 n4C_0_0_0 p4B_1_0_0 n4B_1_0_0 p0C_1_0_0 n0C_1_0_0  
p2C_1_0_0 n2C_1_0_0 VSrc 0 p0D_1_0_0 n0D_1_0_0 VSrc 0 p2A_0_0_0  
n2A_0_0_0 VSrc 0 VSrc 0 VSrc 0 p5G_0_0_0 n5G_0_0_0 VSrc 0 p5F_1_0_0  
n5F_1_0_0 p1G_1_0_0 n1G_1_0_0 VSrc 0 VSrc 0 VSrc 0 VSrc 0 p1E_0_0_0  
n1E_0_0_0 ng_ball_assign_board_resist_ctno_powerNOFAN_v2  
con_resist={0.0125} tileA_board_r_val={x.y} tileB_board_r_val={x.y}  
tileC_board_r_val={x.y} tileD_board_r_val={x.y}  
tileE_board_r_val={x.y} tileF_board_r_val={x.y}  
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}  
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}  
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}  
tileH_rload={x.y} ball_i_diff_thrs={x.y}  
X_0_1_0 p0A_0_1_0 n0A_0_1_0 p2A_0_1_0 n2A_0_1_0 p2C_0_0_0 n2C_0_0_0  
p0B_0_1_0 n0B_0_1_0 VSrc 0 VSrc 0 p0C_0_1_0 n0C_0_1_0 VSrc 0  
p4C_0_1_0 n4C_0_1_0 p0D_0_1_0 n0D_0_1_0 p4B_0_0_0 n4B_0_0_0 VSrc 0  
p1E_0_1_0 n1E_0_1_0 VSrc 0 p1G_0_0_0 n1G_0_0_0 VSrc 0 VSrc 0 VSrc 0  
VSrc 0 VSrc 0 p5G_0_1_0 n5G_0_1_0 p5F_0_0_0 n5F_0_0_0 VSrc 0 VSrc 0  
ng_ball_assign_board_resist_ctno_powerNOFAN_v2 con_resist={0.0125}  
tileA_board_r_val={x.y} tileB_board_r_val={x.y}  
tileC_board_r_val={x.y} tileD_board_r_val={x.y}  
tileE_board_r_val={x.y} tileF_board_r_val={x.y}  
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}  
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}  
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}  
tileH_rload={x.y} ball_i_diff_thrs={x.y}  
X_1_1_0 p0A_1_1_0 n0A_1_1_0 VSrc 0 p2C_1_0_0 n2C_1_0_0 p0B_1_1_0  
n0B_1_1_0 p4C_0_1_0 n4C_0_1_0 VSrc 0 p0C_1_1_0 n0C_1_1_0 VSrc 0 VSrc  
0 p0D_1_1_0 n0D_1_1_0 p4B_1_0_0 n4B_1_0_0 p2A_0_1_0 n2A_0_1_0 VSrc 0  
VSrc 0 p1G_1_0_0 n1G_1_0_0 p5G_0_1_0 n5G_0_1_0 VSrc 0 VSrc 0 VSrc 0  
VSrc 0 VSrc 0 p5F_1_0_0 n5F_1_0_0 VSrc 0 p1E_0_1_0 n1E_0_1_0  
ng_ball_assign_board_resist_ctno_powerNOFAN_v2 con_resist={0.0125}
```

tileA_board_r_val={x.y} tileB_board_r_val={x.y}
tileC_board_r_val={x.y} tileD_board_r_val={x.y}
tileE_board_r_val={x.y} tileF_board_r_val={x.y}
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}
tileH_rload={x.y} ball_i_diff_thrs={x.y}
X_0_0_1 VSrc 0 p2A_0_0_1 n2A_0_0_1 VSrc 0 VSrc 0 VSrc 0 p4B_0_0_1
n4B_0_0_1 VSrc 0 p2C_0_0_1 n2C_0_0_1 p4C_0_0_1 n4C_0_0_1 VSrc 0 VSrc
0 VSrc 0 p1E_0_0_1 n1E_0_0_1 p0A_0_0_0 n0A_0_0_0 VSrc 0 VSrc 0
p0B_0_0_0 n0B_0_0_0 p5F_0_0_1 n5F_0_0_1 p1G_0_0_1 n1G_0_0_1
p0C_0_0_0 n0C_0_0_0 p5G_0_0_1 n5G_0_0_1 VSrc 0 p0D_0_0_0 n0D_0_0_0
VSrc 0 ng_ball_assign_board_resist_ctno_powerNOFAN_v2
con_resist={0.0125} tileA_board_r_val={x.y} tileB_board_r_val={x.y}
tileC_board_r_val={x.y} tileD_board_r_val={x.y}
tileE_board_r_val={x.y} tileF_board_r_val={x.y}
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}
tileH_rload={x.y} ball_i_diff_thrs={x.y}
X_1_0_1 VSrc 0 VSrc 0 VSrc 0 VSrc 0 p4C_0_0_1 n4C_0_0_1 p4B_1_0_1
n4B_1_0_1 VSrc 0 p2C_1_0_1 n2C_1_0_1 VSrc 0 VSrc 0 VSrc 0 p2A_0_0_1
n2A_0_0_1 VSrc 0 p0A_1_0_0 n0A_1_0_0 VSrc 0 p5G_0_0_1 n5G_0_0_1
p0B_1_0_0 n0B_1_0_0 p5F_1_0_1 n5F_1_0_1 p1G_1_0_1 n1G_1_0_1
p0C_1_0_0 n0C_1_0_0 VSrc 0 VSrc 0 p0D_1_0_0 n0D_1_0_0 p1E_0_0_1
n1E_0_0_1 ng_ball_assign_board_resist_ctno_powerNOFAN_v2
con_resist={0.0125} tileA_board_r_val={x.y} tileB_board_r_val={x.y}
tileC_board_r_val={x.y} tileD_board_r_val={x.y}
tileE_board_r_val={x.y} tileF_board_r_val={x.y}
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}
tileH_rload={x.y} ball_i_diff_thrs={x.y}
X_0_1_1 VSrc 0 p2A_0_1_1 n2A_0_1_1 p2C_0_0_1 n2C_0_0_1 VSrc 0 VSrc
0 VSrc 0 VSrc 0 VSrc 0 p4C_0_1_1 n4C_0_1_1 VSrc 0 p4B_0_0_1
n4B_0_0_1 VSrc 0 p1E_0_1_1 n1E_0_1_1 p0A_0_1_0 n0A_0_1_0 p1G_0_0_1
n1G_0_0_1 VSrc 0 p0B_0_1_0 n0B_0_1_0 VSrc 0 VSrc 0 p0C_0_1_0
n0C_0_1_0 p5G_0_1_1 n5G_0_1_1 p5F_0_0_1 n5F_0_0_1 p0D_0_1_0
n0D_0_1_0 VSrc 0 ng_ball_assign_board_resist_ctno_powerNOFAN_v2
con_resist={0.0125} tileA_board_r_val={x.y} tileB_board_r_val={x.y}
tileC_board_r_val={x.y} tileD_board_r_val={x.y}
tileE_board_r_val={x.y} tileF_board_r_val={x.y}
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}
tileH_rload={x.y} ball_i_diff_thrs={x.y}
X_1_1_1 VSrc 0 VSrc 0 p2C_1_0_1 n2C_1_0_1 VSrc 0 p4C_0_1_1
n4C_0_1_1 VSrc 0 VSrc 0 VSrc 0 VSrc 0 VSrc 0 p4B_1_0_1 n4B_1_0_1
p2A_0_1_1 n2A_0_1_1 VSrc 0 p0A_1_1_0 n0A_1_1_0 p1G_1_0_1 n1G_1_0_1
p5G_0_1_1 n5G_0_1_1 p0B_1_1_0 n0B_1_1_0 VSrc 0 VSrc 0 p0C_1_1_0
n0C_1_1_0 VSrc 0 p5F_1_0_1 n5F_1_0_1 p0D_1_1_0 n0D_1_1_0 p1E_0_1_1
n1E_0_1_1 ng_ball_assign_board_resist_ctno_powerNOFAN_v2
con_resist={0.0125} tileA_board_r_val={x.y} tileB_board_r_val={x.y}
tileC_board_r_val={x.y} tileD_board_r_val={x.y}


```
tileE_board_r_val={x.y} tileF_board_r_val={x.y}
tileG_board_r_val={x.y} tileH_board_r_val={x.y} tileA_rload={x.y}
tileB_rload={x.y} tileC_rload={x.y} tileD_rload={x.y}
tileE_rload={x.y} tileF_rload={x.y} tileG_rload={x.y}
tileH_rload={x.y} ball_i_diff_thrs={x.y}
.end
```

References

Here are the references in citation order.

- [1] *MATLAB - MathWorks*. <https://www.mathworks.com/products/matlab.html>. (Accessed Apr. 25, 2023) (cited on page 6).
- [2] Tobias Becker et al. *Unconventional HPC Architectures*. Apr. 2022. DOI: [10.5281/zenodo.6470840](https://doi.org/10.5281/zenodo.6470840) (cited on pages 11, 22).
- [3] Steve B. Furber et al. 'The SpiNNaker Project'. In: *Proceedings of the IEEE* 102.5 (2014), pp. 652–665. DOI: [10.1109/JPROC.2014.2304638](https://doi.org/10.1109/JPROC.2014.2304638) (cited on pages 12, 21, 26, 35).
- [4] Steve Furber (ed.) and Petrut Bogdan (ed.) *SpiNNaker: A Spiking Neural Network Architecture*. Boston - Delft: now publishers, 2020 (cited on pages 12, 35).
- [5] Jonathan Heathcote. 'Building and Operating Large-Scale SpiNNaker Machines'. PhD thesis. Manchester, UK: Department of Computer Science, 2016 (cited on pages 12, 24, 27, 35, 172, 174).
- [6] Doug Black. *Liquid Cooling Trends in HPC*. <https://insidehpc.com/2020/01/liquid-cooling-trends-in-hpc>. (Accessed Apr. 25, 2023). 2020 (cited on page 12).
- [7] *Data center power and cooling strategies for increasing rack power density*. <https://www.deltapowersolutions.com/en/mcis/technical-article-data-center-power-and-cooling-strategies-for-increasing-rack-power-density.php>. (Accessed Apr. 25, 2023) (cited on page 12).
- [8] Andy Lawrence. *Rack Density is Rising*. <https://journal.uptimeinstitute.com/rack-density-is-rising>. (Accessed Apr. 25, 2023). 2020 (cited on page 12).
- [9] *Green500 | TOP500*. <https://www.top500.org/lists/green500>. (Accessed Apr. 25, 2023) (cited on page 13).

- [10] Tianjian Lu et al. 'Accelerating MRI Reconstruction on TPUs'. In: *2020 IEEE High Performance Extreme Computing Conference (HPEC)*. 2020, pp. 1–9. DOI: [10.1109/HPEC43674.2020.9286192](https://doi.org/10.1109/HPEC43674.2020.9286192) (cited on page 14).
- [11] Submer Immersion Cooling. *Typical immersion cooling installation and deployment example with dry cooler or adiabatic cooling towers*. <https://upload.wikimedia.org/wikipedia/commons/thumb/4/41/Immersion-cooling-installation.jpg/2560px-Immersion-cooling-installation.jpg>. is licensed under CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0/deed.en> (Accessed Apr. 25, 2023). 2019 (cited on page 14).
- [12] J. Austin (Cybula Ltd). 'Computing devices'. UK Patent No. GB 2529617. Nov. 2018 (cited on pages 15, 32, 33, 35, 36, 50, 57).
- [13] おむこさん志望. *2x2x2 Three-Dimensional Torus Network*. *This topology is often used by High Performance Computing System, e.g. Blue Gene/L, Cray XT3*. <https://upload.wikimedia.org/wikipedia/commons/thumb/3/3f/2x2x2torus.svg/800px-2x2x2torus.svg.png>. is licensed under CC BY 2.5 <https://creativecommons.org/licenses/by/2.5/deed.en> (Accessed Apr. 24, 2023). 2007 (cited on page 17).
- [14] Amir Mansoor Kamali Sarvestani, Christopher Bailey, and Jim Austin. 'Performance Analysis of a 3D Wireless Massively Parallel Computer'. In: *Journal of Sensor and Actuator Networks* 7.2 (2018). DOI: [10.3390/jsan7020018](https://doi.org/10.3390/jsan7020018) (cited on pages 17, 32, 33, 35, 57, 62).
- [15] William James Dally and Brian Patrick Towles. *Principles and Practices of Interconnection Networks*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2004 (cited on pages 23, 28, 50, 131).
- [16] *European Exascale System Interconnect and Storage*. <https://cordis.europa.eu/project/id/671553>. (Accessed Apr. 25, 2023) (cited on pages 25, 26).
- [17] M. Katevenis et al. 'The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems'. In: *2016 Euromicro Conference on Digital System Design (DSD)*. 2016, pp. 60–67. DOI: [10.1109/DSD.2016.106](https://doi.org/10.1109/DSD.2016.106) (cited on pages 25, 35).
- [18] Roberto Ammendola et al. 'The Next Generation of Exascale-Class Systems: The ExaNeSt Project'. In: *2017 Euromicro Conference on Digital System Design (DSD)*. 2017, pp. 510–515. DOI: [10.1109/DSD.2017.20](https://doi.org/10.1109/DSD.2017.20) (cited on pages 25, 35).

- [19] Javier Navaridas et al. 'Design Exploration of Multi-Tier Interconnection Networks for Exascale Systems'. In: *Proceedings of the 48th International Conference on Parallel Processing*. ICPP '19. Kyoto, Japan: Association for Computing Machinery, 2019. DOI: [10.1145/3337821.3337903](https://doi.org/10.1145/3337821.3337903) (cited on pages 25, 35).
- [20] *European Exascale Processor Memory Node Design*. <https://cordis.europa.eu/project/id/671578>. (Accessed Apr. 25, 2023) (cited on pages 25, 26).
- [21] *Energy-efficient Heterogeneous COmputing at exaSCALE*. <https://cordis.europa.eu/project/id/671632>. (Accessed Apr. 25, 2023) (cited on pages 25, 26).
- [22] *Co-designed Innovation and System for Resilient Exascale Computing in Europe: From Applications to Silicon*. <https://cordis.europa.eu/project/id/754337>. (Accessed Apr. 25, 2023) (cited on pages 25, 26).
- [23] *Precision immersion cooling from the Cloud to the Edge | Iceotope*. <https://www.iceotope.com>. (Accessed Feb. 27, 2023) (cited on page 25).
- [24] *Frontier*. <https://www.olcf.ornl.gov/frontier>. (Accessed Apr. 25, 2023) (cited on pages 25, 35).
- [25] *November 2022 | TOP500*. <https://www.top500.org/lists/top500/2022/11>. (Accessed Apr. 11, 2023) (cited on page 26).
- [26] OLCF at ORNL. *The Exascale-class HPE Cray EX Supercomputer at Oak Ridge National Laboratory*. https://upload.wikimedia.org/wikipedia/commons/thumb/e/e0/Frontier_Supercomputer_%282%29.jpg/878px-Frontier_Supercomputer_%282%29.jpg. is licensed under CC BY 2.0 <https://creativecommons.org/licenses/by/2.0/deed.en> (Accessed Apr. 25, 2023). 2022 (cited on page 26).
- [27] Luis A. Plana et al. 'spiNNlink: FPGA-Based Interconnect for the Million-Core SpiNNaker System'. In: *IEEE Access* 8 (2020), pp. 84918–84928. DOI: [10.1109/ACCESS.2020.2991038](https://doi.org/10.1109/ACCESS.2020.2991038) (cited on pages 27, 35).
- [28] *MDGRAPE-4A | Laboratory for Computational Molecular Design | RIKEN BDR*. <https://www.bdr.riken.jp/en/research/labs/taiji-m/mdgrape4.html>. (Accessed Apr. 25, 2023) (cited on pages 27, 35, 169, 172).

- [29] *MDGRAPE—the special-purpose computer for molecular dynamics simulations (RIKEN BDR)*. https://youtu.be/0_kA3V22yT8. Accessed: Apr. 12, 2023 (cited on pages 27, 35).
- [30] W.J. Dally et al. 'The J-Machine: A fine-grain parallel computer'. In: *Computing Systems in Engineering* 3.1 (1992). High-Performance Computing for Flight Vehicles, pp. 7–15. DOI: [https://doi.org/10.1016/0956-0521\(92\)90089-2](https://doi.org/10.1016/0956-0521(92)90089-2) (cited on pages 28, 35).
- [31] A. Louri and H. Sung. 'An optical multi-mesh hypercube: a scalable optical interconnection network for massively parallel computing'. In: *Journal of Lightwave Technology* 12.4 (1994), pp. 704–716. DOI: [10.1109/50.285368](https://doi.org/10.1109/50.285368) (cited on pages 28, 29, 35, 62).
- [32] A. Louri and Hongki Sung. '3D optical interconnects for high-speed interchip and inter-board communications'. In: *Computer* 27.10 (1994), pp. 27–37. DOI: [10.1109/2.318581](https://doi.org/10.1109/2.318581) (cited on pages 28, 35, 62).
- [33] Gerhard P. Fettweis et al. 'Architecture and Advanced Electronics Pathways Toward Highly Adaptive Energy-Efficient Computing'. In: *Proceedings of the IEEE* 107.1 (2019), pp. 204–231. DOI: [10.1109/JPROC.2018.2874895](https://doi.org/10.1109/JPROC.2018.2874895) (cited on pages 29, 35, 62).
- [34] Amir Mansoor Kamali Sarvestani. 'Evaluating Techniques for Wireless Interconnected 3D Processor Arrays'. PhD thesis. York, UK: Department of Computer Science, Sept. 2013 (cited on pages 30, 33, 35, 62).
- [35] J. Austin (Cybula Ltd). 'Computing devices'. Patent publication No. WO/2003/023583. Mar. 2003 (cited on pages 30, 31, 33, 35, 38).
- [36] Richard Hind. 'FEASIBILITY STUDY ON IMPLEMENTING THE "BALL COMPUTER"'. MA thesis. York, UK: Department of Computer Science, Apr. 2013 (cited on pages 30, 32, 33, 35).
- [37] Nan Jiang et al. 'A detailed and flexible cycle-accurate Network-on-Chip simulator'. In: *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 2013, pp. 86–96. DOI: [10.1109/ISPASS.2013.6557149](https://doi.org/10.1109/ISPASS.2013.6557149) (cited on pages 31, 131, 172).
- [38] *Collectaholics, Series 2, Episode 6, BBC 2* (cited on pages 32, 57).

- [39] Amir Mansoor Kamali, Christopher Crispin-Bailey, and Jim Austin. 'On advantages and limitations of 3D wireless grids as parallel platforms'. In: *2013 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*. 2013, pp. 48–55. DOI: [10.1109/MoWNeT.2013.6613796](https://doi.org/10.1109/MoWNeT.2013.6613796) (cited on page 33).
- [40] Amir Mansoor Kamali, Christopher Bailey, and Jim Austin. 'Evaluating 3D wireless grids as parallel platforms'. In: *International Journal of Ad Hoc and Ubiquitous Computing* 19.3-4 (2015), pp. 279–289. DOI: [10.1504/IJAHUC.2015.070593](https://doi.org/10.1504/IJAHUC.2015.070593) (cited on page 33).
- [41] J. Austin (Cybula Ltd). 'Computing devices that both intercommunicate and receive power by wireless methods'. UK Patent No. GB 2507958. Apr. 2017 (cited on pages 33, 35).
- [42] Pakon Thuphairo et al. 'Investigating Novel 3D Modular Schemes for Large Array Topologies: Power Modeling and Prototype Feasibility'. In: *2022 25th Euromicro Conference on Digital System Design (DSD)*. 2022, pp. 268–275. DOI: [10.1109/DSD57027.2022.00044](https://doi.org/10.1109/DSD57027.2022.00044) (cited on pages 33, 48, 58–60, 68–70, 75, 76, 80–84, 106, 138, 139, 186).
- [43] Pabogdan. *A panorama of the SpiNNaker 1 million core machine*. https://upload.wikimedia.org/wikipedia/commons/thumb/9/97/Spinn_1m_pano.jpg/2560px-Spinn_1m_pano.jpg. is licensed under CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0/deed.en> (Accessed Apr. 25, 2023). 2018 (cited on page 34).
- [44] Gentaro Morimoto et al. 'Hardware Acceleration of Tensor-Structured Multilevel Ewald Summation Method on MDGRAPE-4A, a Special-Purpose Computer System for Molecular Dynamics Simulations'. In: *SC21: International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021, pp. 1–15. DOI: [10.1145/3458817.3476190](https://doi.org/10.1145/3458817.3476190) (cited on page 35).
- [45] Steve Temple. *AppNote 9 - SpiNN-5 Quick Start Guide*. 1.02. SpiNNaker Group, School of Computer Science, University of Manchester. Manchester, United Kingdom, Feb. 2015 (cited on page 35).
- [46] Indar Sugiarto et al. 'High performance computing on SpiNNaker neuromorphic platform: A case study for energy efficient image processing'. In: *2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC)*. 2016, pp. 1–8. DOI: [10.1109/PCCC.2016.7820645](https://doi.org/10.1109/PCCC.2016.7820645) (cited on page 35).

- [47] Christian Mayr, Sebastian Hoepfner, and Steve Furber. *SpiNNaker 2: A 10 Million Core Processor System for Brain Simulation and Machine Learning*. 2019 (cited on page 35).
- [48] *FRONTIER Spec Sheet*. https://www.olcf.ornl.gov/wp-content/uploads/2019/05/frontier_specsheet.pdf. (Accessed Apr. 25, 2023) (cited on page 35).
- [49] Scott Atchley. *Frontier's Architecture*. <https://olcf.ornl.gov/wp-content/uploads/Frontiers-Architecture-Frontier-Training-Series-final.pdf>. (Accessed Apr. 25, 2023) (cited on page 35).
- [50] *Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11 | TOP500*. <https://www.top500.org/system/180047>. (Accessed Apr. 25, 2023) (cited on page 35).
- [51] *Submer | Smart Solutions for Next Generation Datacenters*. <https://submer.com>. (Accessed Feb. 27, 2023) (cited on page 37).
- [52] Miyuru Dayarathna, Yonggang Wen, and Rui Fan. 'Data Center Energy Consumption Modeling: A Survey'. In: *IEEE Communications Surveys Tutorials* 18.1 (2016), pp. 732–794. DOI: [10.1109/COMST.2015.2481183](https://doi.org/10.1109/COMST.2015.2481183) (cited on pages 39, 42).
- [53] Kenneth O'brien et al. 'A Survey of Power and Energy Predictive Models in HPC Systems and Applications'. In: *ACM Comput. Surv.* 50.3 (June 2017). DOI: [10.1145/3078811](https://doi.org/10.1145/3078811) (cited on pages 39, 42, 67).
- [54] Wenzhi He et al. 'Research on AC & DC hybrid power supply system with high-proportion renewable energy of data centre'. In: *The Journal of Engineering* 2019.16 (2019). Is licensed under CC BY 3.0 <https://creativecommons.org/licenses/by/3.0>, pp. 3230–3233. DOI: <https://doi.org/10.1049/joe.2018.8925> (cited on page 41).
- [55] Schleifenbauer. *Active Power Distribution Units A and B feed in datacenter server rack*. https://upload.wikimedia.org/wikipedia/commons/f/fa/Schleifenbauer_PDU_Power_Distribution_Unit_in_operation.jpg. is licensed under CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0/deed.en> (Accessed Apr. 25, 2023). 2022 (cited on page 41).

- [56] Matthew J. Walker et al. 'Accurate and Stable Run-Time Power Modeling for Mobile and Embedded CPUs'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 36.1 (2017), pp. 106–119. DOI: [10.1109/TCAD.2016.2562920](https://doi.org/10.1109/TCAD.2016.2562920) (cited on page 43).
- [57] Matthew J. Walker et al. 'Run-time power estimation for mobile and embedded asymmetric multi-core CPUs'. In: *HIPEAC Workshop on Energy Efficiency with Heterogenous Computing*. 2015 (cited on page 43).
- [58] Mihai Pricopi et al. 'Power-performance modeling on asymmetric multi-cores'. In: *2013 International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES)*. 2013, pp. 1–10. DOI: [10.1109/CASES.2013.6662519](https://doi.org/10.1109/CASES.2013.6662519) (cited on page 43).
- [59] Christoph Möbius, Walteneagus Dargie, and Alexander Schill. 'Power Consumption Estimation Models for Processors, Virtual Machines, and Servers'. In: *IEEE Transactions on Parallel and Distributed Systems* 25.6 (2014), pp. 1600–1614. DOI: [10.1109/TPDS.2013.183](https://doi.org/10.1109/TPDS.2013.183) (cited on page 43).
- [60] Walteneagus Dargie. 'A Stochastic Model for Estimating the Power Consumption of a Processor'. In: *IEEE Transactions on Computers* 64.5 (2015), pp. 1311–1322. DOI: [10.1109/TC.2014.2315629](https://doi.org/10.1109/TC.2014.2315629) (cited on page 43).
- [61] *Xilinx - Adaptable. Intelligent | together we advance*. <https://www.xilinx.com>. (Accessed Apr. 25, 2023) (cited on page 43).
- [62] *Xilinx Power Estimator (XPE)*. <https://www.xilinx.com/products/technology/power/xpe.html>. (Accessed Apr. 25, 2023) (cited on page 43).
- [63] *VCCCORE DC Specifications - 004 - ID:743844 | 13th Generation Intel174; Core™ Processors*. <https://edc.intel.com/content/www/us/en/design/products/platforms/details/raptor-lake-s/13th-generation-core-processors-datasheet-volume-1-of-2/004/vcccore-dc-specifications>. (Accessed Apr. 25, 2023) (cited on page 44).
- [64] *Supervisory Devices Complementary Parts Guide for Xilinx FPGAs*. https://www.analog.com/media/en/product-associations/complementary-part-guides/supervisory_xilinx_82359_200902.pdf. (Accessed Apr. 25, 2023) (cited on page 44).
- [65] *LT3976 - 40V, 5A, 2MHz Step-Down Switching Regulator with 3.3 A Quiescent Current*. LT3976. Analog Devices, Inc. 2013 (cited on pages 45, 71, 131, 169, 182).

- [66] C.P. Basso. *Switch-Mode Power Supplies, Second Edition: SPICE Simulations and Practical Designs*. McGraw Hill LLC, 2014 (cited on pages 46, 71).
- [67] Michael Leonard. 'Automated Behavioral Modeling of Switching Voltage Regulators'. Bachelor's Thesis. Fayetteville, USA: College of Engineering, University of Arkansas, 2013 (cited on pages 46, 71).
- [68] Michael Leonard. 'Semi-Automated Switching Regulator Modeling Method and Tool'. MA thesis. Fayetteville, USA: College of Engineering, University of Arkansas, 2015 (cited on pages 46, 71).
- [69] *LTspice Information Center | Analog Devices*. <https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html>. (Accessed Apr. 25, 2023) (cited on page 48).
- [70] *Ngspice, the open source Spice circuit simulator - Intro*. <https://ngspice.sourceforge.io>. (Accessed Apr. 25, 2023) (cited on pages 48, 128).
- [71] *Truncatedoctahedron*. <https://upload.wikimedia.org/wikipedia/commons/thumb/2/20/Truncatedoctahedron.jpg/261px-Truncatedoctahedron.jpg>. is licensed under CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0/deed.en> (Accessed Apr. 28, 2023). 2023 (cited on page 59).
- [72] *OpenSCAD*. <https://openscad.org>. (Accessed Apr. 25, 2023) (cited on pages 60, 117).
- [73] William Thomson. 'On the division of space with minimum partitional area'. In: *Acta Mathematica* 11.none (1900), pp. 121–134. DOI: [10.1007/BF02612322](https://doi.org/10.1007/BF02612322) (cited on page 60).
- [74] *Truncated octahedron*. https://www.wikipedia.org/wiki/Truncated_octahedron. (Accessed Apr. 25, 2023) (cited on page 60).
- [75] *DC conduction solution*. <https://www.mathworks.com/help/pde/ug/pde.conductionresults.html>. (Accessed Apr. 25, 2023) (cited on page 63).
- [76] *DC conduction analysis –QuickField FEA Software*. <https://quickfield.com/cflow.htm>. (Accessed Apr. 25, 2023) (cited on page 63).
- [77] *FreeCAD: Your own 3D parametric modeler*. <https://www.freecad.org>. (Accessed Apr. 25, 2023) (cited on page 63).

- [78] Richard Yin. *Power Tips 109: Five major trends in power supply design for servers*. <https://www.edn.com/five-major-trends-in-power-supply-design-for-servers>. (Accessed Apr. 25, 2023). 2022 (cited on page 67).
- [79] Zheng Zhang et al. 'A block-diagonal structured model reduction scheme for power grid networks'. In: *2011 Design, Automation Test in Europe*. 2011, pp. 1–6. DOI: [10.1109/DATE.2011.5763014](https://doi.org/10.1109/DATE.2011.5763014) (cited on page 67).
- [80] *Fit curves and surfaces to data - MATLAB*. <https://www.mathworks.com/help/curvefit/curvefitter-app.html>. (Accessed Sep. 24, 2023) (cited on page 72).
- [81] *Evaluating Goodness of Fit - MATLAB Simulink*. <https://www.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html>. (Accessed Sep. 24, 2023) (cited on page 76).
- [82] *Relative change and difference*. https://www.wikipedia.org/wiki/Relative_change_and_difference. (Accessed Sep. 5, 2023) (cited on page 82).
- [83] *Ngspice User's Manual*. <https://ngspice.sourceforge.io/docs/ngspice-manual.pdf>. (Accessed Apr. 25, 2023) (cited on page 105).
- [84] Kalyanmoy Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. Chichester, England: John Wiley Sons, Ltd, 2001 (cited on page 118).
- [85] *gamultiobj Algorithm - MATLAB Simulink*. <https://www.mathworks.com/help/gads/gamultiobj-algorithm.html>. (Accessed Apr. 25, 2023) (cited on pages 118, 123, 126, 152).
- [86] *MD5*. <https://www.wikipedia.org/wiki/MD5>. (Accessed Apr. 25, 2023) (cited on page 121).
- [87] *Execute for-loop iterations in parallel on workers - MATLAB parfor*. <https://www.mathworks.com/help/parallel-computing/parfor.html>. (Accessed Oct. 5, 2023) (cited on page 123).
- [88] *1 Pair 3a Magnetic Pogo Pin Connector 6 Positions Pitch 2.2 Mm Spring Loaded Header Contact Strip Power Charge Data Transfer - Connectors - AliExpress*. <https://www.aliexpress.com/item/1005002137597704.html>. (Accessed Apr. 25, 2023) (cited on pages 139, 169).

- [89] *Polynomial Models - MATLAB Simulink*. <https://www.mathworks.com/help/curvefit/polynomial.html>. (Accessed Aug. 7, 2023) (cited on page 141).
- [90] *Power Series - MATLAB Simulink*. <https://www.mathworks.com/help/curvefit/power.html>. (Accessed Aug. 7, 2023) (cited on page 141).
- [91] *Exponential Models - MATLAB Simulink*. <https://www.mathworks.com/help/curvefit/exponential.html>. (Accessed Aug. 7, 2023) (cited on page 141).
- [92] *Teach, learn, and make with the Raspberry Pi Foundation*. <https://www.raspberrypi.org/>. (Accessed Apr. 25, 2023) (cited on page 142).
- [93] *Raspberry Pi 15W USB-C Power Supply*. <https://datasheets.raspberrypi.com/power-supply/usb-c-power-supply-product-brief.pdf>. (Accessed Apr. 25, 2023) (cited on page 142).
- [94] *Particle swarm optimization*. https://www.wikipedia.org/wiki/Particle_swarm_optimization. (Accessed Sep. 5, 2023) (cited on pages 152, 194).
- [95] John Kim et al. 'Technology-Driven, Highly-Scalable Dragonfly Topology'. In: *2008 International Symposium on Computer Architecture*. 2008, pp. 77–88. DOI: [10.1109/ISCA.2008.19](https://doi.org/10.1109/ISCA.2008.19) (cited on page 177).
- [96] *AMD EPYC™ 7453 | AMD*. <https://www.amd.com/en/products/cpu/amd-epyc-7453>. (Accessed Apr. 25, 2023) (cited on page 180).
- [97] *AMD Instinct™ MI250X Accelerator | AMD*. <https://www.amd.com/en/products/server-accelerators/instinct-mi250x>. (Accessed Apr. 25, 2023) (cited on page 180).
- [98] John Kim, James Balfour, and William Dally. 'Flattened Butterfly Topology for On-Chip Networks'. In: *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*. 2007, pp. 172–182. DOI: [10.1109/MICRO.2007.29](https://doi.org/10.1109/MICRO.2007.29) (cited on page 185).
- [99] *Slurm Workload Manager - Documentation*. <https://slurm.schedmd.com>. (Accessed Apr. 25, 2023) (cited on page 195).