Hotspot management to reduce energy consumption in containerised data centres

Mr Jose Mejia-Arredondo

Submitted in accordance with the requirements for the degree of
Master of Philosophy

The University of Leeds
Institute of Thermofluids
School of Mechanical Engineering

November 2025

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I want to thank:

My supervisors: Dr. Mark Wilson, Professor Karim Djemame, Professor Simon Rees, and Dr. Jon Summers. I could not have started or completed this work without their guidance, feedback, and support.

Professor Peter Culmer and the faculty staff (Joanne, Vicky, Michelle, Francesca, Debra, and Sanita), for their support.

Aly, Trevor, Carol, and Alder for their support and help in accessing resources, and for always being available to assist.

Dr. Alan Burns for his patience, teaching, guidance, and tutoring in CFD.

The Leeds University Union, which helped me navigate university processes and provided support in difficult moments.

CONACyT-Sener funded this work.

My family and friends have supported me emotionally and financially during this journey, offering a helping hand in times of need.

Abstract

The rapid increase in computing demand has increased the energy load of data centres, especially in containerised data centres where limited space worsens thermal problems. This thesis tests the hypothesis that the energy efficiency of containerised data centres can be improved by strategically distributing computational workloads to minimise thermal hotspots. To evaluate this theory, a computational fluid dynamics (CFD) model was developed based on the experimental containerised data centre system of Wang et al. (2017). A critical review of the literature established typical server characteristics, workload behaviours, and key energy metrics, highlighting significant gaps in validated thermal models and underscoring the need for realistic assessments of airflow, heat transfer and cooling performance in containerised data centres.

A parametric study of multiple Workload Distribution strategies was conducted to quantify their effects on hotspot formation, rack-level thermal behaviour, and cooling energy demand. Six workload configurations were analysed under steady-state conditions, complemented by a detailed examination of IT energy consumption across utilisation levels. The results demonstrate that baseline energy consumption remains substantial even at low utilisation, emphasising the importance of reducing idle capacity. Three workload strategies (cases D9, E12, and F15) successfully reduced hotspot intensity while maintaining cold-aisle temperatures within ASHRAE limits, resulting in observable reductions in total energy consumption. These findings indicate that coordinated workload allocation can enhance both thermal performance and cooling efficiency, although aggressive consolidation risks creating new hotspots that could offset potential energy savings. This research provides a validated framework for assessing the interaction between workload distribution, thermal behaviour, and energy consumption in containerised data centres. Although limited by steady-state modelling and simplified boundary conditions, the study offers a foundation for developing dynamic, environmentally aware scheduling algorithms. Future work should incorporate transient workloads, external climatic effects, variable airflow rates, and multiple cooling units to generalise the findings and support more energy-efficient containerised data centre designs.

Table of Contents

Acknowledgements ii

Abstract iii

Chapter 1 Introduction 1

1.1 Introduction 1

1.2 Aim and objectives 3

1.3 Thesis Outline 4

Chapter 2 Background and Literature Review on Thermal Management and Energy Efficiency 7

2.1 Introduction 7

2.2 Air Cooling 8

2.2.1 Cold Aisle/Hot Aisle Layout System 8

2.2.2 Cold or Hot Aisle Air Containment 12

2.2.3 Discussion 15

2.3 Liquid Cooling 20

2.4 Environmental Cooling 20

2.5 Energy consumption in data centres 25

2.5.1 Energy consumption by cooling systems in data centres 25

2.5.2 Energy efficiency metrics 31

2.5.3 Current data centres energy efficiency 33

2.6 Conclusion 41

Chapter 3 Background and Literature Review on Computational Workloads 44

3.1 Introduction 44

3.2 Data centre workload types 44

3.3 Scheduling 46

3.3.1 Basic Terminology 48

3.3.2 Classification of Scheduling Types 50

3.3.3 Local Scheduling 52

3.3.4 Geographic Scheduling 54

3.3.5 Temperature reactive 57

3.4 Conclusions from the Literature Review 58

Chapter 4 Computational Methodology 61

4.1 Introduction 61

4.2 Data centre configuration and previous work 61

4.2.1 Server-scale test problems 65

4.2.2 Boundary conditions for the server-scale model 70

4.2.3 Velocity sensitivity study 70

4.3 Numerical Methodology 73

4.3.1 Governing Equations and Turbulence Modelling 73

4.3.2 Boundary conditions 74

4.3.3 Numerical Solution Procedure 76

4.3.4 Porous medium representation of the servers 77

4.3.5 Momentum source representation 78

4.3.6 Computational cost and scaling strategy 78

4.3.7 Mesh quality...79

4.3.8 Convergence 79

4.3.9 Computational Set-up 80

4.4 Containerised data centre model 81

4.5 Model Validation 86

4.6 Summary 90

Chapter 5 Analysis of workload distribution effects on hotspots. 92

5.1 Workloads 92

5.2 Limitations and Assumptions 93

5.3 Scenario 1 94

5.4 Scenario 3 102

5.5 Analysis and Discussion of Airflow Characteristics 108

5.6 Conclusions 113

Chapter 6 Analysis of energy consumption 117

6.1 Methodology for Energy Estimation 118

6.2 Energy Consumption with IT in different states 120

6.3 Energy Consumption using this approach 122

6.4 Energy required for Air Cooling...124

6.5 Discussion 128

Chapter 7 Conclusions 130

7.1 Overall Conclusions 130

7.2 Achievement of the Research Aim and Objectives 130

7.3 Principal Contributions 133

7.4 Limitations 134

7.5 Future Work 136

7.6 Final Remarks 138

Appendix A 149

List of Tables

Table 2.1 Economic comparison for a cooling system CAC(Cold Aisle Containtment) and VED(Vertical Exhaust Duct) with a dry cooler, mechanical chiller, and cooling tower (Shrivastava et al., 2012)...................................... 14

Table 2.2 Summary of most thermal metrics and benchmark values (Ni and Bai, 2017).. 30

Table 2.3 Energy efficiency metrics and their benchmark values (Ni and Bai, 2017).. 31

Table 4.1 Independence Mesh Study detailed results. 69

Table 4.2 Results of the variation of velocity in the inlet and the effect that produces in the parts of interest. 71

Table 4.3 Numerical model settings, material properties and boundary conditions used in the CFD simulations. 76

Table 5.1 Different workload scenarios studied and their distribution in the data centre. 93

Table 5.2 Results obtained in seventeen scenarios where the inlet air flowrate for V1: 3.52 m/s, V2: 2.83 m/s, and V3: 4m/s (Detailed results can be found in Appendix A). 97

Table 6.1 Shows that the server is near the peak energy efficiency at 50% utilisation, while consuming 431 W. The idle power draw is 82.9 W (approximately 8.8% of full load power). 121

Table 6.2 presents the information corresponding to the energy consumption by the servers in each workload distribution, where the inlet air flowrate for V1: 3.52 m/s, V2: 2.83 m/s, and V3: 4m/s. 123

Table 6.3 Presents the results of numerical calculations for six different workload distributions, where cases with less total energy are highlighted for each Workload Distribution. 127

List of Figures

Figure 2.1 Schematic of De Boer et al's. (2018) benchmark case study data centre for cooling operations. Temperatures and airflows are indicated by the colour (warm – red; cool - blue) and direction of the arrows. 9

Figure 2.2 Overhead cooling container as an evaluated data centre configuration (Qouneh et al., 2011).. 10

Figure 2.3 Circular in-row cooling container as an evaluated data centre configuration (Qouneh et al., 2011).. 10

Figure 2.4 Power consumption in data centres with Cold Aisle/Hot Aisle. (Song et al., 2015) 11

Figure 2.5 This shows how the hot air is separated by adding a physical barrier (Niemann et al., 2011). 13

Figure 2.6 In-row cooling container - cold and hot aisle containment technique as an evaluated data centre configuration (Qouneh et al., 2011).......... 15

Figure 2.7 Some schemes of direct airside free cooling utilized in datacentres (Daraghmeh and Wang, 2017).......... 21

Figure 2.8 U.S. data centre electricity consumption (historical and projected scenarios), adapted from the 2024 Lawrence Berkeley National Laboratory Data Center Energy Usage Report. 25

Figure 2.9 Ranges Chart showing energy consumed by cooling systems in data centres expressed by different authors. 29

Figure 2.10 Annual average PUE of 23 data centres in Singapore (Ni and Bai, 2017)**.** 31

Figure 2.11 Annual average PUE of 44 data centres (Ni and Bai, 2017).. 32

Figure 2.12 Design of "Blackbox" by Sun Microsystems (Waldrop, 2007) 38

Figure 2.13 Schematic of fully implemented Environmentally Opportunistic Computing (Ward et al., 2012) 39

Figure 2.14 Schematic of prototype EOC node connected to a local greenhouse. The inset image shows a photo of the EOC node in-place. (Ward et al., 2012) 40

Figure 4.1 Experimental containerised data centre configuration reported by Wang et al. (2017), showing the CRAH unit, ductwork and rack arrangement used as the basis of the CFD model. 63

Figure 4.2 Computational domain used for the server-scale validation study, illustrating the porous server representation within the wind tunnel. 66
Figure 4.3 Structured computational mesh adopted for the wind-tunnel validation case. 67
Figure 4.4 Independence Mesh Study chart, the Y axis shows the average temperature in the interior of the porous block expressed in K (kelvins) and number of mesh elements in the bottom axis. 68
Figure 4.5 Plot of vectors that show the fans activated to draw air from the porous block.......... 72
Figure 4.6 Plot of vectors that show the fans and the inlet activated to draw air from the porous block 72
Figure 4.7 Contour plot that shows the temperatures reached by the porous block representing the heat generation of a server and how the heat is drawn by the fans towards the outlet 72
Figure 4.8 Illustrates the view of the mesh in the half-model. 79
Figure 4.9 Vectors showing the airflow path in the halved containerised data centre simulation 81
Figure 4.10 Design of a containerised data centre (according to Wang et al., 2017a).......... 82
Figure 4.11 Diagram that describes how both experimental and theoretical implementations intend to approximate models in real-life to achieve more closely real-world results.......... 83
Figure 4.12 Geometry of the halved containerised data centre showing the inlet and outlet's system. 84

Figure 4.13 Number distribution of the blocks representing the servers inside the containerised data centre. 84
Figure 4.14 Sideview of the halved containerised data centre with the outlet in the front side of the geometry. 85
Figure 4.15 Mirrored view of the geometry of the containerised data centre displaying an approximation of the one used by Wang et al. (2017). 86
Figure 4.16 Results by Wang et al. (2017), where they are exploring the scenario 1 pictured in Figure 4.17. 87
Figure 4.17 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre. 89
Figure 4.18 Contours of static temperature obtained from the measurement plane located in the hot aisle of the containerised data centre. 90
Figure 5.1 Representation of Workload Distribution utilising a traffic light system to depict the two different workloads, where Workload A, B and C are intended to distribute 60% corresponding to scenario 1, and D, E and F are meant to allocate 30% as per scenario 3. 94
Figure 5.2 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution A, the conditions listed as Cases 1, 2, 3 in Table 5.2 and the workload allocation as shown in Figure 5.1 A). 95
Figure 5.3 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution B, the

conditions listed as Cases 4, 5, 6 in Table 5.2 and the workload allocation as shown in Figure 5.1 B). 98

Figure 5.4 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution C, the conditions listed as Cases 7 and 8 in Table 5.2 and the workload allocation as shown in Figure 5.1 C). 101

Figure 5.5 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution D, the conditions listed as Cases 9, 10 and 11 in Figure 5.1 D). 103

Figure 5.6 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution E, the conditions listed as Cases 12, 13 and 14 in Figure 5.1 E). 105

Figure 5.7 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution F, the conditions listed as Cases 15, 16 and 17 107

Figure 5.8 Cold aisle view illustrates that airflow enters from the inlet before descending through the rack region. The path lines converge in the middle row, indicating preferential flow through regions of lower resistance. 108

Figure 5.9 The view shows the 3-D nature of the airflow. High-temperature path lines (green–yellow–red) are within and immediately downstream of the rack, illustrating the transport of heat generated by the servers. 110

Figure 5.10 Isometric view highlights the accumulation of recirculating flow near the inlet of the container. The path lines indicate that not all of the heated air exits directly through the outlet. Instead, portions of the flow remain trapped within. 111
Figure 6.1 shows a graphic with the Server Energy Consumption and the Energy Efficiency by Workload level based on the data provided by Qiu et al. (2019). 122

List of Acronyms and Abbreviations

ACM – Association for Computing Machinery

AI - Artificial Intelligence

AIRCC – Academy & Industry Research Collaboration Center

ANSYS Fluent – Computational Fluid Dynamics simulation software

ASHRAE – American Society of Heating, Refrigerating and Air-Conditioning Engineers

BHS – Best Heuristic Scheduling

CAC – Cold Aisle Containment

CCHP – Combined Cooling, Heating, and Power

CEVP – Cross Entropy-Based Virtual Machine Placement

CFD – Computational Fluid Dynamics

CNN – Convolutional Neural Network

CO2 – Carbon Dioxide

CRAC – Computer Room Air Conditioning

CRAH – Computer Room Air Handler

CUE – Carbon Usage Effectiveness

DAG – Directed Acyclic Graph

DC – Data Centre

DCiE – Data Centre Infrastructure Efficiency

DVFS – Dynamic Voltage and Frequency Scaling

EEVS – Energy Efficient Virtual Machines Scheduling

EOC – Environmentally Opportunistic Computing

ETC – Estimated Time to Compute

FCFS – First Come First Served

GPU – Graphics Process Unit

HEFT – Heterogeneous Earliest Finish Time

HP POD – Hewlett-Packard Performance Optimised Datacenter

IEEE – Institute of Electrical and Electronics Engineers

IoT – Internet of Things

IT – Information Technology

LMT – Levelized Minimum Time

MREE-PSO – Multi-resource Energy Efficiency based on Particle Swarm Optimisation

MTDP – Minimum Total Data Centre Power (MINTOTDATACENTERPOW)

POD – Performance Optimised Datacenter

PPO – Proximal Policy Optimisation

PSO – Particle Swarm Optimisation

PUE – Power Usage Effectiveness

QoS – Quality of Service

RCI – Rack Cooling Index

SDN – Software-Defined Networking

SLA – Service Level Agreement

SLR – Schedule Length Ratio

VED – Vertical Exhaust Duct

VM – Virtual Machine

Chapter 1

Introduction

1.1 Introduction

Digital services i.e., internet, telecommunications, and information systems rely on data centres and associated networks that consume large and increasing proportions of the country's power generation. There is an increasing demand in energy consumption (Shehabi, 2008; Whitehead et al., 2014a; Sharma et al., 2015a; Uddin et al., 2015; Oró et al., 2018a). The main problem with energy consumption in data centres is that much of the energy turns into heat, which is a problem since electronic components have an upper-temperature limit for safe and reliable operation; therefore, cooling is needed, heat removal is compulsory for a proper working data centre in all cases. The current requirements to reduce energy consumption and emissions are changing, driving a need for more efficient energy use and the reduction of carbon footprint generated in data centres.

Modern society makes increasing use of digital services. An example of the extreme and unexpected increase in the use of digital services is the crisis experienced due to the Coronavirus outbreak. This event made everybody remain at home. Therefore, students attended lectures from home, and non-essential workers had to work remotely, resulting in heavy bandwidth use. Another example is the popularisation of AI-driven services. AI's current use streamlines business operations, powers daily conveniences, and drives scientific breakthroughs, from writing meeting minutes, generating content for media, providing live translations, predicting traffic, suggesting travel routes, tracking road incidents, and generating optimal routes. Virtual assistance that manages daily routines, calendars, and spam filters.

This increasing business demand for computing is driving data centres to increase their hosted servers rapidly; furthermore, data centres are seeking new ways to operate the existing power facilities as close as possible to their maximum capacities, which increases overall energy consumption (Zheng et al., 2015). Consequently, developments in high-performance heterogeneous computing scheduling algorithms have been pursued to increase energy efficiency. This allows optimal use of resources with workload distribution among the different servers and processors, which reduces energy consumption. However, further research is needed to maximise the energy efficiency contribution.

Another use that motivates the research in this type of data centre is the possibility of implementing the emerging 5G cellular technology to keep the data flowing in the now-growing digital infrastructure to the current network of edge nodes. This type of data centre was designed to bring digital services closer to the end user with applications that require low latency over the internet protocol networks as well helping to reduce the ever-growing demand for high volume digital services.

This research is motivated by the necessity to reduce energy consumption within data centres, specifically containerised. Containerised data centres are subsets of data centres which interact with different components of the larger data centre network. Multiple containerised data centres can be combined to form clusters of nodes to form a farm of data centres.

Hotspots occur inside data centres, which operators must detect and monitor before they become an issue. A hotspot is an area at the inlet of the IT equipment where the temperature is greater than the 27°C recommended by the American Society of Heating Refrigerating and Air-Conditioning Engineers Inc. (ASHRAE) (2016), which is an international professional organisation dedicated to advancing heating, ventilation, air conditioning, and refrigeration (HVAC&R) systems, indoor air quality, and energy efficiency. The problem with a hotspot is that if it crosses the safety maximum allowable threshold which is 32°C, it can cause IT equipment to overheat, which puts the data centre's performance at risk during peak hours. In the worst-case scenario, the heat can damage electronic components, resulting in an interruption to the service provided by the data centre. Usually, data centres have a bigger cooling capacity than required; however, hotspots occur in those data centres because of inadequate airflow management. Hotspots inside data centres can also be predicted through computational models that combine the thermo-fluid mechanics of the system to identify the number, locations and size of hotspots in data centres. The model simulates the airflow and heat generation inside the data centre, aiding in predicting the hotspots and preparing different strategies to prevent them from becoming an issue.

1.2 Aim and objectives

The aim of this research is to test the hypothesis that the energy efficiency of containerised data centres can be improved by carefully distributing the computational workload to minimise hotspots.

The objectives for this research were as follows:

- Develop a computational fluid dynamics model of a containerised data centre, based on the experimental system of Wang et al. (2017).
- Review the relevant literature to establish typical server characteristics, workload patterns, and energy consumption metrics.
- Conduct a parametric study of different workload distributions to determine their effects on hotspot locations and intensity.
- Quantify the corresponding energy consumption of the datacentre under the different workload configurations to determine the degree to which energy can be saved through workload distribution.

1.3 Thesis Outline

The structure of the thesis to follow:

Chapter 2 outlines the three approaches to cooling systems within data centres: air cooling, liquid cooling, and environmental cooling. This is followed by explaining the process of measuring the energy consumption of the different approaches to cooling data centres and determining data centre energy efficiency.

Chapter 3 provides an in-depth overview of workload scheduling in data centres, including local and geographical scheduling allocation. This is followed by an outline of current energy efficiency algorithms used within task scheduling, highlighting research limitations.

Chapter 4 delivers the methodology applicable to this research. It explains the selected method of study, which involves computational fluid dynamics calculations, the assumptions, settings, and the validation of the CFD model against the literature.

Chapter 5 presents the computational fluid dynamics to obtain information about the localisation and magnitude of hot spots and the behaviour of thermal conditions with different workloads in two simulated scenarios.

Chapter 6 showcases the analysis of energy consumption by carefully relocating workload distribution in two simulated scenarios studied responding to localisation and magnitude of hot spots, reallocating the workload distribution and the number of servers involved while observing the thermal behaviour within the data centre.

Chapter 7: Conclusions and future work.

Chapter 2

Background and Literature Review on Thermal Management and Energy Efficiency

2.1 Introduction

This chapter provides context for thermal management and energy efficiency for different types of data centres, their layouts and thermal management. In the case of containerised data centres, most studies are performed with computational fluid dynamics (CFD) simulations, and a smaller quantity are performed in data centres experimentally. First the chapter will outline different approaches to cooling containerised data centres, which will comprise of air cooling, liquid cooling and environmental cooling. This will be followed by measuring the energy consumption of the different approaches and determining data centre energy efficiency.

Data centre cooling technologies have evolved considerably over the last decade in response to increasing rack power densities and the rapid growth of artificial intelligence (AI) workloads. While traditional air-cooling methods based on hot aisle/cold aisle arrangements remain widely deployed, modern facilities increasingly incorporate containment systems to minimise air recirculation and improve cooling efficiency. Furthermore, direct liquid cooling technologies are becoming increasingly common in hyperscale and AI-oriented data centres because conventional air cooling becomes progressively less effective at very high heat densities. Environmental or free-cooling strategies also continue to play an important role in suitable climates by reducing mechanical refrigeration requirements. Consequently, current best practice is no longer represented by a single cooling architecture but by selecting an appropriate cooling strategy according to workload density, climatic conditions and energy efficiency objectives.

Containerised data centres are equipped with cooling systems which perform thermal management differently: air cooling, liquid cooling and environmental cooling. The principal cooling architectures employed in containerised data centres are described in the following sections.

2.2 Air Cooling

Air cooling has two types of layouts, which are explored in this section: Cold aisle/hot aisle, described further in 2.2.1 and cold aisle/hot aisle air containment later on in the subsection 2.2.2.

2.2.1 Cold Aisle/Hot Aisle Layout System

The most conventional method for cooling a data centre is by drawing cooled air through servers. Typically, servers are arranged to create hot aisles and cold aisles (Song et al, 2015), as shown in Figure 2.1. There are three techniques to improve the efficiency of cooling systems. The first technique deliberately separates the cold and hot air using a plenum chamber under the floor of the room to generate cold air, which is then delivered through the perforated floor by a cooling system (CRAC).

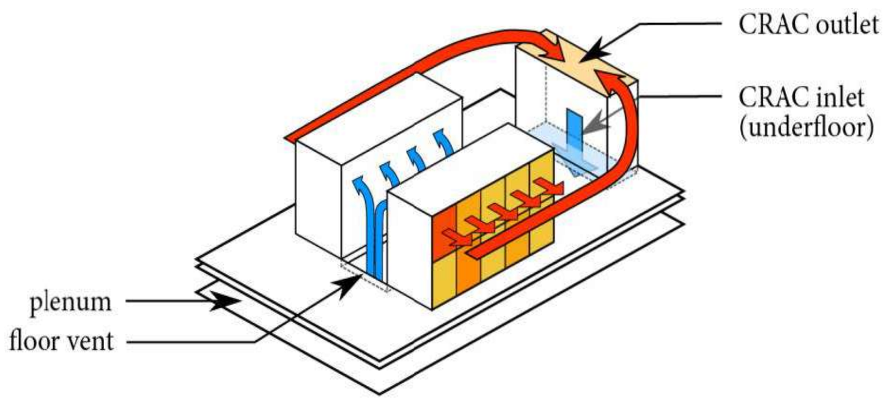

Figure 2.1 Schematic of De Boer et al's. (2018) **benchmark case study data centre for cooling operations. Temperatures and airflows are indicated by the colour (warm – red; cool - blue) and direction of the arrows.**

The second technique to improve the efficiency of cooling systems is an Overhead Cooling Container system which uses a single row of racks that divides the container into two parts; one forms a cold aisle and the other forms a hot aisle, as shown in Figure 2.2. CRAC units are installed on top of the racks to supply the cold air. The HP POD (Performance Optimised Datacenter) system uses this cooling technique (Hewlett-Packard Development Company, 2011).

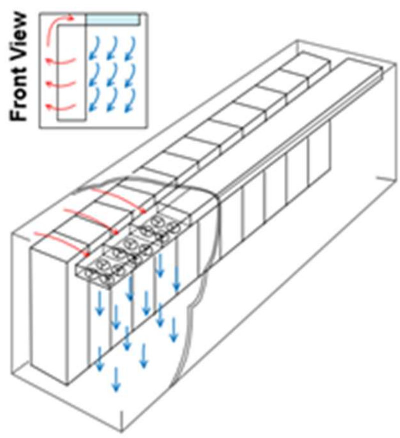

Figure 2.2 Overhead cooling container as an evaluated data centre configuration (Qouneh et al., 2011).

Finally, a third technique, is to use a Circular In-row Cooling Container, where the racks are also stacked on both sides as in the 'Overhead Cooling Container', however the racks are stacked front to back rather than side to side with a CRAC unit in-between. Air flows in a circular fashion and is the one adopted by Sun's modular data centres (Sun Microsystems inc, 2009). This configuration can be seen in Figure 2.3.

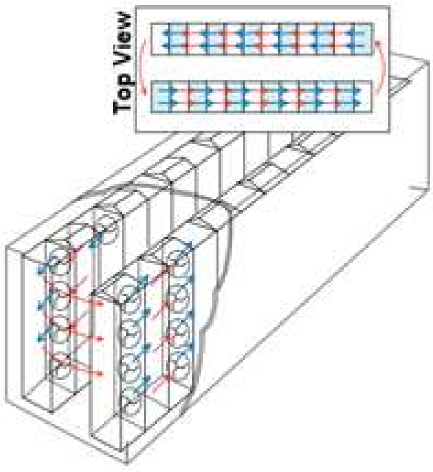

Figure 2.3 Circular in-row cooling container as an evaluated data centre configuration (Qouneh et al., 2011).

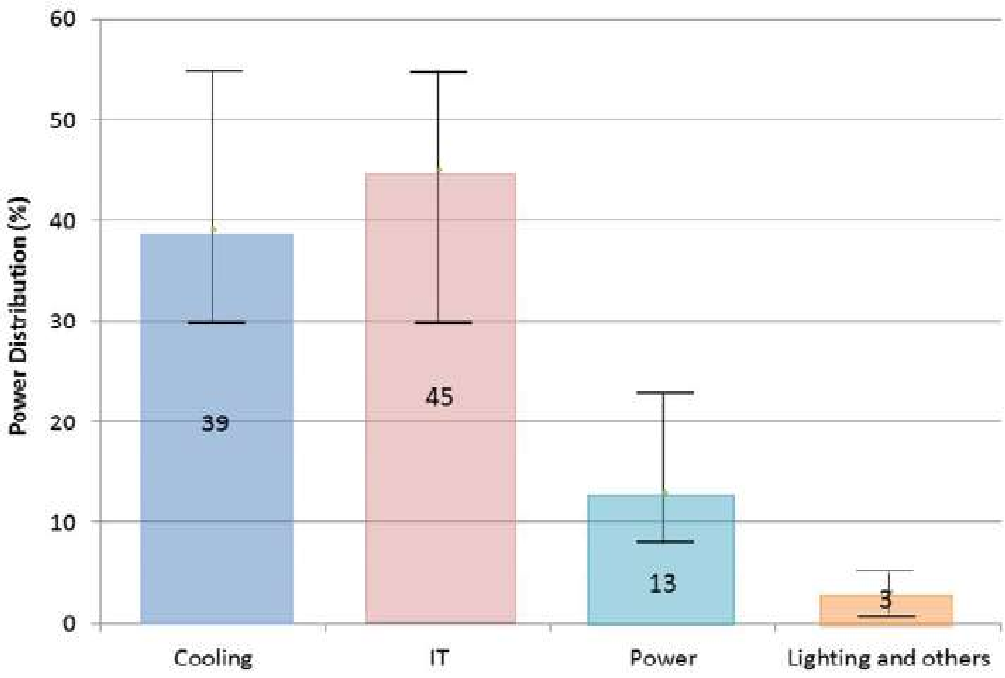

Figure 2.4 Power consumption in data centres with Cold Aisle/Hot Aisle. (Song et al., 2015)

In Figure 2.4 it can be observed that the power consumption of the cooling system can vary between 30% and 55% with an average of 39%. Whilst this is significant, the energy spent by IT also varies between 30% and 55%, but the IT average is higher than the cooling system average. This implies that energy consumption must be reduced by two factors: a) improving the cooling systems, and b) optimisation of IT resources. The fluctuations in IT resources and their energy consumption are correlated to workload, and this correlation is supported by a wide body of literature (Jia et al. 2013; Dragseth and Hanssen, 2016; Li, 2016; Yan et al., 2018).

2.2.2 Cold or Hot Aisle Air Containment

Container-based data centres rely on containment and modularity, which is achieved by barriers or racks that prevent air recirculation inside the container. Two commercially available container architectures are described here.

The main intention of Cold/Hot Aisle Air Containment is to manage airflows (Kumar and Joshi, 2012), forcing air through servers, to avoid bypassing the servers and mixing cold and hot air. Therefore, physical barriers are added to improve the airflow control as seen in Figure 2.5, which increases the cooling efficacy and energy efficiency by forcing the cooling of the servers (Arghode et al., 2013; Choo et al., 2014). While Arghode et al. (2013) and Choo et al. (2014) have modelled this system through simulations using CFD, some others like Khalili et al. (2018) have validated their results using Data Centre (DC) facilities as well as simulations.

It has also been demonstrated by Shrivastava et al. (2012) that this system can save a large amount of money. Table 2.1 shows an economic comparison between CAC(Cold Aisle Containment) and VED(Vertical Exhaust Duct) with a dry cooler, mechanical chiller, and cooling tower., made by Shrivastava et al. (2012) where it can be observed that using Cold Aisle Containment in all rows can save $195,633 per annum.

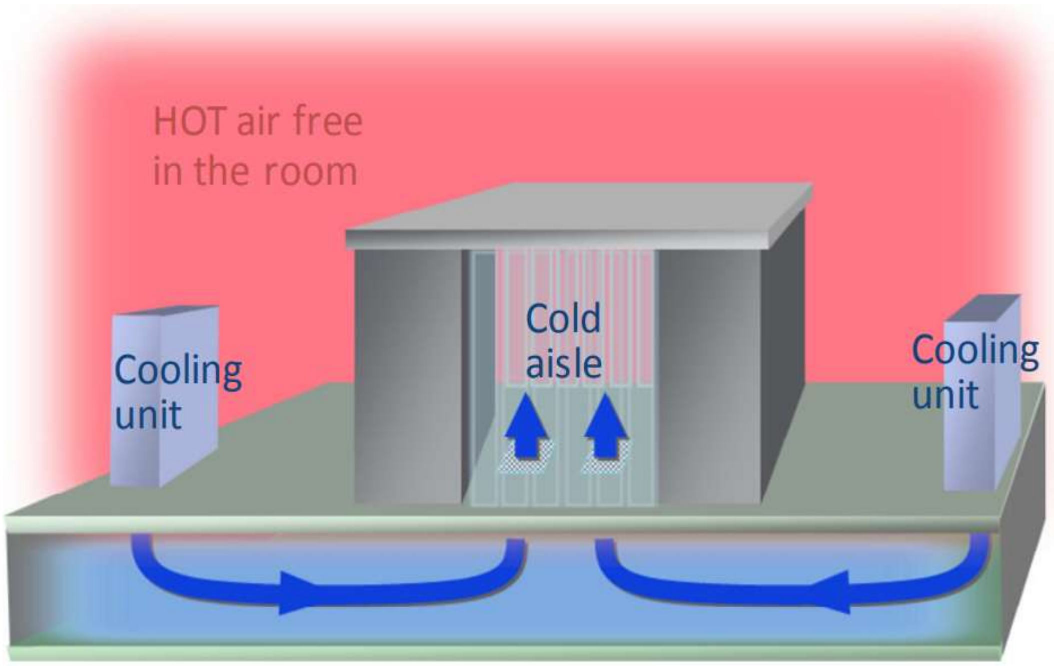

Figure 2.5 This shows how the hot air is separated by adding a physical barrier (Niemann et al., 2011).

In Figure 2.5, it is possible to distinguish how, with this technique, the cooling unit is only focused on cooling the IT, taking advantage of the physical barrier. Therefore, the energy spent on cooling is used to remove heat from the servers rather than wasted in recirculation.

	Baseline	CAC on Server Rows	VED on Server Rows	VEDs or CACs on all Rows
Cooling Airflow,cfm	288,000	235,200	235,200	207,317
CRAH Supply Air Temp. (F)	59	68	77	78
Av. CRAH Return Air Temp, (F)	77	90	99	103
Chiller Water Supply Temp. (F)	46	55	64	65
Hours of Dry Cooler	885	2,126	3,074	3,215
Hours of Chiller Operation	7,875	6,634	5,686	5,545
Total CRAH Fan Power Consumption (kWh)	1,239,365	675,046	675,046	462,301
Chiller Power Consumption (kWh)	2,157,927	1,483,971	985,731	930,278
Dry Cooler Fan Power Consumption (kWh)	66,375	159,450	230,550	241,125
Pumps Power Consumption (kWh)	334,559	314,802	299,710	297,465
Cooling Tower Fan Energy (kWh)	236,250	199,061	170,580	166,350
Annual Cooling Energy Cost ($0.101/kWh)	407,482	286,061	238,523	211,849
Annual Savings ($)	NA	121,421	168,959	195,633

Table 2.1 Economic comparison for a cooling system CAC(Cold Aisle Containtment) and VED(Vertical Exhaust Duct) with a dry cooler, mechanical chiller, and cooling tower (Shrivastava et al., 2012).

The next technique is In-Row Cooling Container which consists of stacked racks on both sides along the length of the container creating a “cold” middle aisle and two “hot” narrow aisles behind the rows, as shown in Figure 2.6. CRAC units, interspersed among the racks, supply the cold air. This configuration is the one used by SGI’s ICE Cube modular data centres (Qouneh et al., 2011).

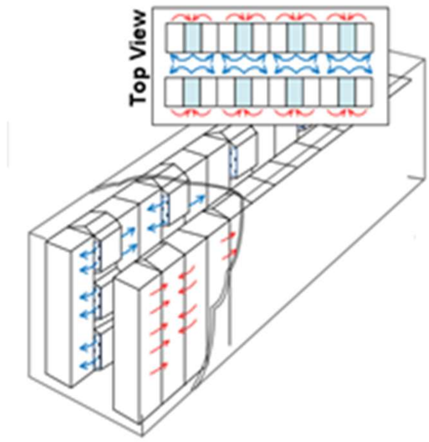

Figure 2.6 In-row cooling container - cold and hot aisle containment technique as an evaluated data centre configuration (Qouneh et al., 2011).

2.2.3 Discussion

In the study performed by Qouneh et al. (2011) there are no results regarding a comparison of the thermal behaviour of the three mentioned configurations for the cooling systems in the containerised data centre models mentioned above. However, their results were shown on energy efficiency and PUE improvements.

Therefore, the advantages and disadvantages of any of them cannot be considered to evaluate them side by side with other studies where the authors provide the snapshots of the thermal behaviour and the parameters involved in the physics to replicate said environment. On the other hand, the airflow management shown on the study presented by Qouneh et al. (2011) have been performed on containerised data centres providing a vision of what is possible to achieve if a better configuration is implemented. For example, in Figure 2.6 is shown a probed design with different configurations that allow inferring that cooling efficiency can improve according to the cooling system configuration.

The body of literature that recognises some disadvantages in the cold aisle/hot aisle technique (Schmidt et al., 2001; Kang et al., 2001; VanGilder and Schmidt, 2005; Song et al., 2015b; Arghode et al., 2017; Ling et al., 2017a; Khalili et al., 2018). The most significant disadvantage is that airflow cannot be controlled; therefore, cold, and hot air are mixed. Wang et al. (2017) states that the most popular layout for effective air-flow management in a data centre is the cold/hot aisle configuration, which, minimises mixing of hot and cold air. However, the literature partially disagrees with Wang et al. (2017). Even if this technique is applied by supplying air through underfloor raised plenums, this does not prevent the cold and hot air from mixing, since both streams are not completely isolated. i.e. the configuration that Shrivastava et al. (2012) is using for their study.

In the case study presented by Choo et al. (2014) a porous jump was used to simulate a server rack. An increase of permeability in porous medium (increased number of pores and their diameter) and reduction of resistance is studied to evaluate the effectiveness of the increasing surface to provide air into the cold aisles. However, there was a pressure loss in this scenario derived from several characteristics such as, the porosity of screen, i.e., the ratio of the open area to the overall tile section; the tile thickness t and pore diameter D, usually considered by the relative thickness t/D; the upstream section area; the number of tiles; and the number of pores. Also Choo et al. (2014) questioned the validity of the model in large spaces, given the large section that the stream passes before being discharged in the cold aisle. The implication of the air stream traveling a long distance is an increase in the temperature of the stream which could result in a waste of energy. Therefore, the present author agrees with Choo et al. (2014) that the size of the room needs to be considered in the model before validating the results, since the size of the room is important to preserve the cold stream. This position is supported by Emerson Network Power (2010) and Mastelic et al. (2014) who noted that the effectiveness of cooling is influenced by the height and width of the data centre and the way the air is supplied.

Choo et al. (2014), Mastelic et al. (2014) and Emerson Network Power (2010) during their case of studies were focused on preserving the cold stream in the room. While that Wang et al. (2017) provided their case to address the heat generation at chip level to reduce the energy consumption.

A further source of wasted energy arises from servers creating resistance to cold airflow, leading to the airflow bypassing the servers, and consequently limiting the heat transfer between cold air and servers in the Cold Aisle/Hot Aisle configuration. By not addressing the resistance of servers and allowing the mixing of hot and cold air streams, energy is wasted.

Whitehead et al. (2014) highlighted some other disadvantages, in this case related to fans; one of them is regarding energy consumption and the other is about maintenance.

Another limitation is (that most of the studies do not consider the power category) The power category expressed in the Power consumption chart specifically for cold/hot aisle in the Figure 2.4 is the powering provision system for the data centre (for example: batteries to back up the critical devices in the data centre).

The use of fans leads to a significant amount of energy consumption. Energy consumption can be affected to varying degrees by other factors. For example, the speed of the fans impacts energy consumption whereby not properly maintaining fans can lead to increased consumption. Fans and other components often fail because of air pollutants and condensation damaging hardware so air-cooled data centres use filters and other systems to reduce these failures, however, it is still important to regularly maintain fans to ensure optimal performance and avoid increased energy consumption (McFarlane et al., 2023).

Although all of the cooling layouts described in this chapter continue to be used within the data centre industry, their prevalence varies according to application. Conventional enterprise facilities typically employ contained hot aisle/cold aisle air cooling because of its relatively low cost and operational simplicity. However, hyperscale and AI-focused data centres increasingly adopt direct liquid cooling to accommodate higher rack power densities while maintaining acceptable Power Usage Effectiveness (PUE). Immersion cooling remains an emerging technology with growing adoption in specialist high-performance computing applications, but has not yet become mainstream. Consequently, aisle containment combined with efficient airflow management currently represents best practice for conventional air-cooled facilities, whereas liquid cooling is becoming the preferred solution for high-density computing environments. Although the cooling layouts described above were originally developed more than a decade ago, they remain the foundation of modern data centre thermal management. Recent studies continue to investigate methods for improving airflow distribution, reducing hot-air recirculation and enhancing cooling efficiency within these established layouts. For example, Chu et al. (2020) demonstrated that optimising airflow management within containerised data centres significantly improves Rack Cooling Index (RCI) and Supply Heat Index (SHI), while Chu et al. (2021) further showed that aisle containment and airflow resistance management can substantially reduce hot spots and improve cooling performance in containerised environments. These studies indicate that, although the underlying cooling layouts remain largely unchanged, ongoing research is focused on optimising their thermal performance rather than replacing them.

Overall, the literature indicates that hot aisle/cold aisle configurations remain the predominant air-cooling strategy for conventional data centres owing to their simplicity and effectiveness. Recent studies have focused on improving these layouts through enhanced airflow management, aisle containment and optimisation of supply airflow, particularly within containerised data centres. However, as rack power densities continue to increase, especially for AI applications, direct liquid cooling is emerging as the preferred solution for high-density deployments, whereas contained air-cooling remains current best practice for conventional facilities.

2.3 Liquid Cooling

Liquid cooling techniques implement different types of liquids to improve energy efficiency as part of their innovation. There are several different ways in which liquid can be incorporated into a cooling system. There are some thermal advantages gained through using liquid cooling, for example, higher conductivity and greater heat capacity. Such systems are also easily scalable and there is a lower risk of air pollutants; however, using cold natural resources and rejecting the heat to the environment contributes to global warming (Bian, 2020).

2.4 Environmental Cooling

The environmental impact of data centre power consumption, as well as the effect of escalating power densities on the ability to pack machines into a data centre, is discussed by Rivoire et al. (2007). Cold weather can be utilised to save energy. This type of cooling is a technique which takes into account the environmental conditions to save energy Covas et al. (2013). Therefore, if the weather outside is cool, a CRAC system may not be needed; otherwise, there will be normal energy consumption.

This method usually consists of letting in the airflow from the outside to the data centre, sometimes mixed with one of the air-cooling techniques, which are explained in Figure 2.7 Whitehead et al. (2014). Air-side free cooling involves removing the cooling load of a space using outdoor air of a lower temperature than that of the thermal environment (Lee and Chen, 2013). The significant air-side free cooling potential was achieved in Data Centres with mixed humidity climate zones reducing 2.8 - 8.5% the energy consumption (Lee and Chen, 2013). The results from the different authors show that this method is efficient in reducing energy consumption.

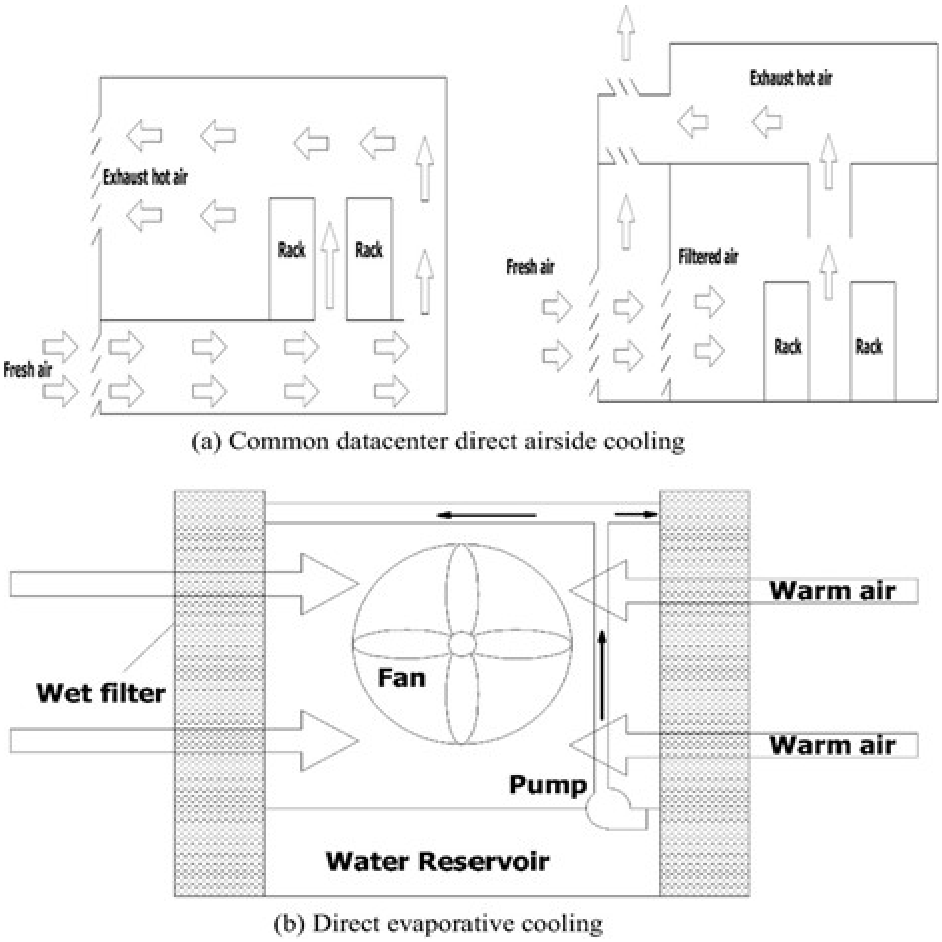

Figure 2.7 Some schemes of direct airside free cooling utilized in datacentres (Daraghmeh and Wang, 2017)

In Figure 2.7 (a) can be appreciated roughly how the air-side free cooling technique works, where the fresh air is forced inside the data centre and by forced convection the racks are cooled, and the hot air is exhausted by the top of the data centre. If this model were studied closer it's possible that mixed convection would be found.

A solar radiative heating simulation using CFD with Phoenics software package was performed by Budiyanto et al. (2017) where the impact the solar radiation has in shipping containers was studied. The highlights of their research include the average amount of energy to remove the heat load inside the container from the effect of heat penetration due to solar radiation. Budiyanto et al. (2017) reported that approximately 2.2 kWh of additional cooling energy was required over a 24-hour period to remove the heat introduced by solar radiation. Because this value represents accumulated energy over time, the appropriate unit is kilowatt-hours (kWh) rather than kilowatts (kW), which would describe instantaneous power demand. Although 2.2 kWh per day appears modest when compared with the overall daily electricity consumption of operational data centres, it represents only the additional thermal load caused by solar radiation on the container envelope. Over extended operating periods, particularly in regions with high solar irradiance, this additional cooling requirement becomes cumulative and may noticeably reduce cooling system efficiency. Consequently, solar heat gain should be incorporated into thermal models of containerised data centres located in warm climates. In consequence, this affects the cooling system efficiency, by increasing the energy consumption in environments where the solar radiation is high. The present author finds it interesting how other authors with CFD studies have neglected the thermal load added by solar radiation. This shows that the results will clearly be affected depending on the location of the container as expressed in the physics of the paper presented by Budiyanto et al. (2017).

Although environmental cooling seems a very good cooling technique for data centres, this research highlights that is unlikely to be suitable in most tropical countries such as Mexico where the weather is rather tropical than cold, and the solar radiation is high.

2.5 Energy consumption in data centres

In this section, current metrics for energy consumption in data centres are presented.

2.5.1 Energy consumption by cooling systems in data centres

Earlier projections suggested that U.S. data centre electricity consumption would reach approximately 140 billion kWh by 2020 (Chu et al., 2019). However, more recent evidence demonstrates that electricity demand has accelerated substantially beyond those earlier forecasts due largely to rapid expansion of AI workloads and GPU-accelerated computing. The latest Lawrence Berkeley National Laboratory report estimates that annual U.S. data centre electricity consumption increased from approximately 60 TWh in 2014 to 176 TWh in 2023 and could reach between 325 and 580 TWh by 2028 depending upon AI deployment and cooling technology adoption.

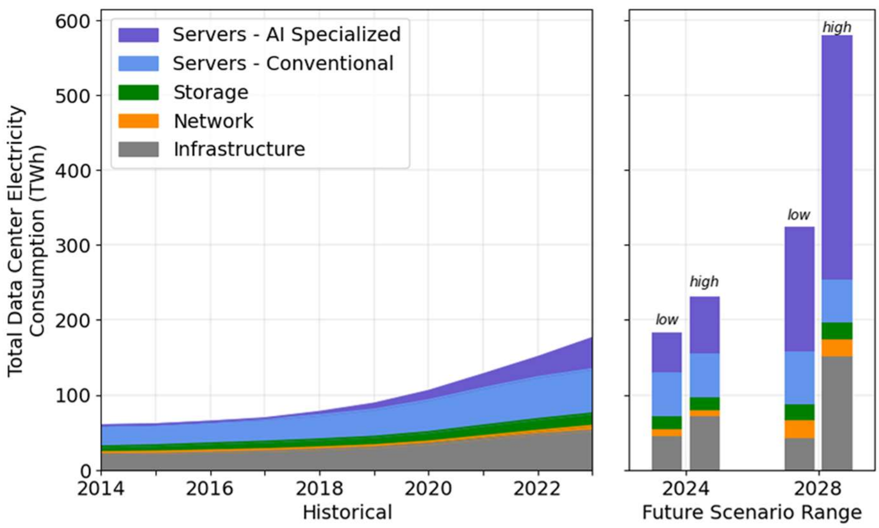

Figure 2.8 U.S. data centre electricity consumption (historical and projected scenarios), adapted from the 2024 Lawrence Berkeley National Laboratory Data Center Energy Usage Report.

Although total electricity consumption has increased significantly since 2017, this increase should not be interpreted solely as a decline in energy efficiency. Between approximately 2010 and 2017, improvements in server utilisation, virtualisation, cooling system performance and reductions in Power Usage Effectiveness (PUE) largely offset the rapid growth in digital services. More recently, however, the widespread deployment of GPU-accelerated servers for AI workloads has substantially increased rack power densities and electricity demand, resulting in a renewed upward trend in total data centre energy consumption.

Almoli et al. (2012) mentioned that roughly 40% of the total energy consumption corresponds to cooling systems requirements. This is fairly close to the average consumption in the hot aisle/cold aisle cooling technique shown in Figure 2.4 and discussed in the section 2.2.1. A major portion of the consumed energy (almost 50%) is used for cooling infrastructure to maintain the operational temperature within the allowable limits (Almoli, 2013). Although this statement is above the average, it is within the mentioned range on Figure 2.4.

One of the Energy Efficiency Trends that were mentioned by Shehabi et al. (2016) is about improving the energy efficiency in server parts: for example, microprocessor, cooling fan and power supply. Some other parts are expected to be removed: “The fan in servers using liquid cooling”. This type of technique could also be referred to as Hotspot-Targeted Cooling.

Although with the amount of information publicly provided at that time, the above-mentioned forecast was convincing, now it is public knowledge that industry secrecy affects the way that energy efficiency measures are being applied to data centre facilities. This research recognises that industry secrecy can be turned into a collaboration network to improve energy efficiency in data centres, which consequently will reduce the impact on the carbon footprint and the overall environment. The literature indicates that this collaborative data would provide insights into proposing the creation of an institute or a group that certifies the collaboration of the industry, which might work as an incentive for the industry to end the secrecy. This idea provides insights for a framework based on the gamification of collaboration, which is also a methodology used mostly in video games and social platforms to encourage users to do certain sets of activities and provide them with a badge but in this case to show their compliance with the institute.

Shehabi also expressed that not only industry secrecy and rapidly evolving technology alter forecast reports but predicted that electricity consumption would rise significantly in 2007 because electricity usage in the industry servers doubled between 2000 and 2005 (LBNL, 2007). However, in 2007 the server market was impacted by the financial crash and developments in technology and the expected increase did not happen.

Other researchers such as Chu et al. (2019) mentioned that the energy consumption of a CRAC accounts for 40 to 50% of the total energy consumption and through thermal management strategies a 10 -15 % improvement is usually likely to be achieved. This project is accounting for a wider range arguing that

30% ~ 60% of energy consumption is attributable to the cooling infrastructure according to Liu et al. (2016). Energy consumption depends on several factors: type of cooling system, setting of the cooling system, amount of IT inside the data centre and usage of the IT, amongst others. Finally, Martin et al. (2007) reported that effective strategies in the oracle data centre achieved as much as 40% power savings. All these statements are gathered in Figure 2.9 below where the fluctuation in the percentage of energy consumption in data centres according to different authors is demonstrated.

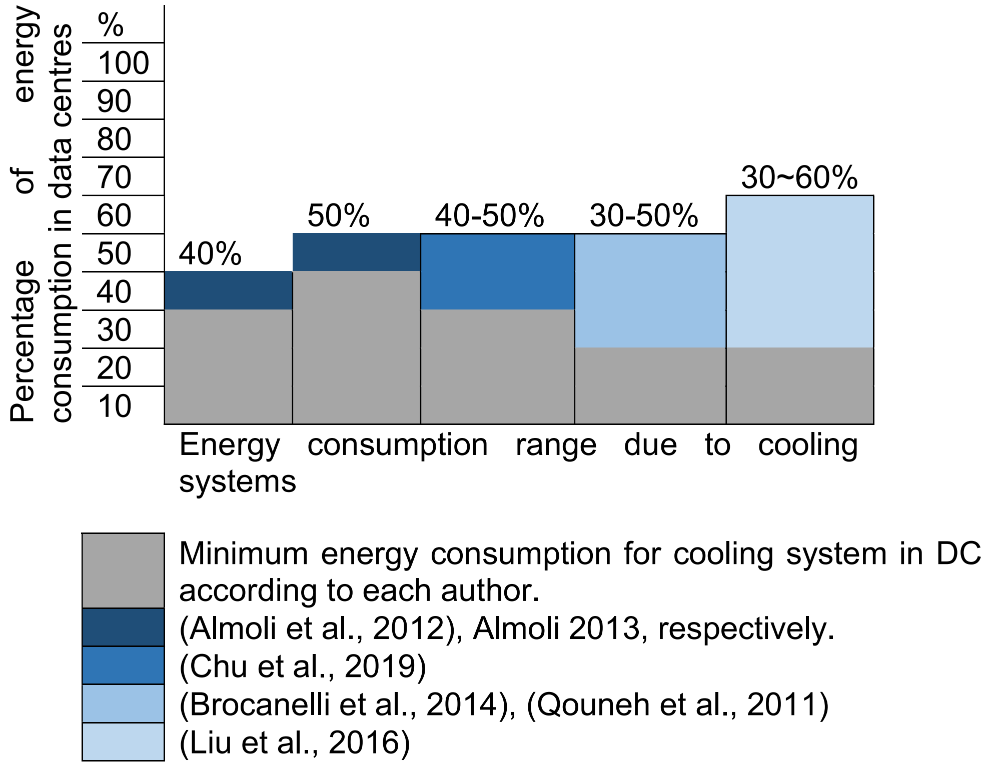

Figure 2.9 Ranges Chart showing energy consumed by cooling systems in data centres expressed by different authors.

Figure 2.9 proves that the energy spent on cooling systems can vary in a greater percentage in contrast with the variations presented before in Figure 2.4.

There are other ways to measure the cooling efficiency which at a certain point can be translated to energy efficiency, some of them were compiled by (Ni and Bai, 2017) and are summarised in Table 2.2 all of them applied to air cooling techniques.

Introduced by	Metric	Explanation	Information provided	Input parameters	Formula	Benchmark value
Sharma et al.	SHI	Supply Heat Index	Quantify the extent of mixing of return air	airflow supply temperature from CRAC unit, rack inlet and outlet airflow temperatures	$SHI = \frac{\sum_i \sum_j (T^r_{in_{i,j}} - T^C_{sup})}{\sum_i \sum_j (T^r_{out_{i,j}} - T^C_{sup})}$	ideal 0 good < 0.2
Sharma et al.	RHI	Return Heat Index	Quantify the extent of bypass of supply air	mass flow rate supply from CRAC unit, mass flow rate across equipment, airflow return temperature at CRAC, airflow supply temperature from CRAC unit, airflow temperature exhaust from rack	RHI = 1-SHI	ideal 1 good > 0.8
Schmidt et al.	β Index	Beta Index	Presence of recirculation and over heating	Local airflow inlet supply and outlet temperatures	$\beta = \frac{T^r_{in}(z) - T^C_{sup}}{T^r_{out} - T^r_{in}}$	ideal 0
Herrlin	RCI_{LO}	Rack Cooling Index Low	Quantify the rack health at the high end of the temperature range	Rack intake air temperatures distribution, maximum recommended and allowable temperature according to some guideline or standard	$RCI_{LO} = \left[1 - \frac{\sum (T_{min-rec} - T_x)_{T_x < T_{min-rec}}}{(T_{min-rec} - T_{min-all}) \times n}\right] \times 100\%$	ideal 1 good $\geq 96\%$ acceptable 91–95% poor $\leq 90\%$
Herrlin	RCI_{HI}	Rack Cooling Index High	Quantify the rack health at the low end of the temperature range	Rack intake air temperatures distribution, minimum recommended and allowable temperature according to some guideline or standard	$RCI_{HI} = \left[1 - \frac{\sum (T_x - T_{max-rec})_{T_x > T_{max-rec}}}{(T_{max-all} - T_{max-rec}) \times n}\right] \times 100\%$	ideal 1 good $\geq 96\%$ acceptable 91–95% poor $\leq 90\%$
Tozer et al.	NP	Negative Pressure ratio	Airflow infiltration into underfloor plenum	return air temperature to CRAC, discharge air temperature from CRAC, floor void temperature	$NP = \frac{T^{uf}_{sup} - T^C_{sup}}{T^C_{ret} - T^{uf}_{sup}}$	ideal 0
Tozer et al.	BP	Bypass ratio	the rate of air which leaves the floor grills and returns directly to the CRAC unit without cooling servers	return air temperature to CRAC, floor void temperature, server outlet air temperature	$BP = \frac{T^S_{out} - T^C_{ret}}{T^S_{out} - T^{uf}_{sup}}$	ideal 0 good 0–0.05 acceptable 0.05–0.2
Tozer et al.	R	Recirculation ratio	the rate of hot air which goes into the IT equipment	server inlet and outlet air temperature, floor void temperature	$P = \frac{T^S_{in} - T^{uf}_{sup}}{T^S_{out} - T^{uf}_{sup}}$	ideal 0 good < 0.2
Tozer et al.	BAL	Balance ratio	the difference between cold air produced by the cooling plant and the server request	server inlet and outlet air temperature, return air temperature to CRAC, discharge air temperature from CRAC	$BAL = \frac{T^S_{out} - T^S_{in}}{T^C_{ret} - T^C_{sup}}$	ideal 1
Herrlin	RTI	Return Temperature Index	Presence of recirculation or bypass phenomena	server inlet and outlet air temperature, return air temperature to CRAC, discharge air temperature from CRAC	$RTI = \frac{T^C_{ret} - T^C_{sup}}{T^r_{out} - T^r_{in}} \times 100\%$	ideal 100% good 95–105%

Table 2.2 Summary of most thermal metrics and benchmark values (Ni and Bai, 2017).

2.5.2 Energy efficiency metrics

The energy efficiency metrics are used to evaluate the performance and the benchmarks are made to identify the potential opportunities to reduce the use of energy in data centres, the Table 2.3 shows definitions and typical values of the energy metrics and benchmark values compiled by (Ni and Bai, 2017).

Introduced by	Metrics		Benchmark Values		
			Standard	Good	Better
The Green Grid	PUE	$PUE = \frac{Total\ Facility\ Power}{IT\ Equipment\ Power}$	2	1.4	1.1
The Green Grid	DCiE	$DCiE = \frac{IT\ Equipment\ Power}{Total\ Facility\ Power}$	0.5	0.7	0.9
NREL	HVAC Effectiveness Index	$\text{HVAC System Effectiveness} = \frac{Annual\ IT\ Equipment\ Energy}{Annual\ HVAC\ System\ Energy}$	0.7	1.4	2.5
NREL	Airflow Efficiency	$\text{Airflow Efficiency} = \frac{Total\ Fan\ Power}{Total\ Fan\ Airflow}$	1.25 W/cfm	0.75 W/cfm	0.5 W/cfm
NREL	Cooling System Efficiency	$\text{Cooling System Efficiency} = \frac{Average\ Cooling\ System\ Power}{Average\ Cooling\ Load}$	1.1 kW/ton	0.8 kW/ton	0.6 kW/ton

Table 2.3 Energy efficiency metrics and their benchmark values (Ni and Bai, 2017).

However, the most widely used and accepted efficiency metric is Power Usage Effectiveness (PUE) and the Data Centre Infrastructure Efficiency (DCiE) which is a reciprocal of PUE.

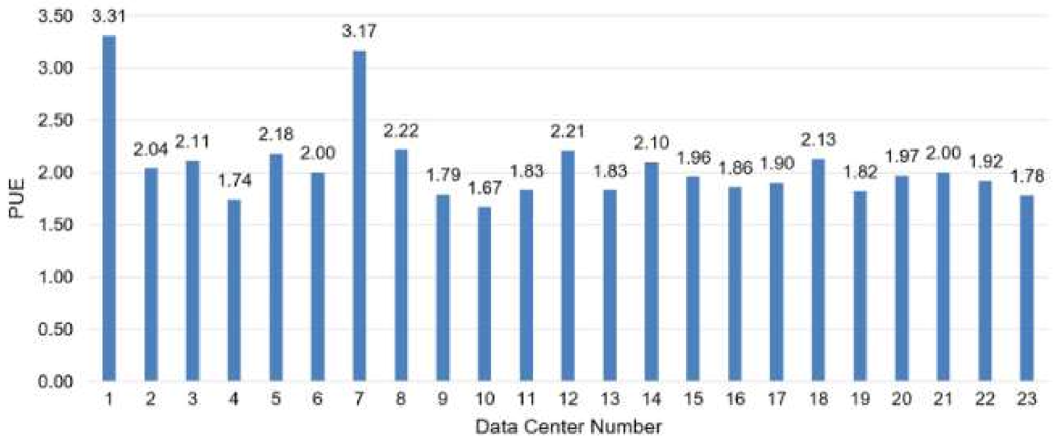

Figure 2.10 Annual average PUE of 23 data centres in Singapore (Ni and Bai, 2017)**.**

Half of the collected data from articles and reports by (Chu et al., 2019) showed that the power usage efficiency (PUE) are over 2.4 and the desirable value of PUE is 1.0, which indicates appreciable inefficiencies of the current data centres (Chu et al., 2019) As can be seen in Figure 2.10 and Figure 2.11.

(Shehabi et al., 2016) argue that in an attempt to improve the PUE of smaller data centres, the use of infrastructure and the set point temperature has been reduced, improving airflow and thermal management with cold/hot aisle containment.

However, as similar to the work done in the research mentioned above, the evidence suggests that is challenging to achieve the same results in containerised data centres due to the lack of space and less opportunities to manage the equipment and configurations on the layout to improve the airflow and thermal management.

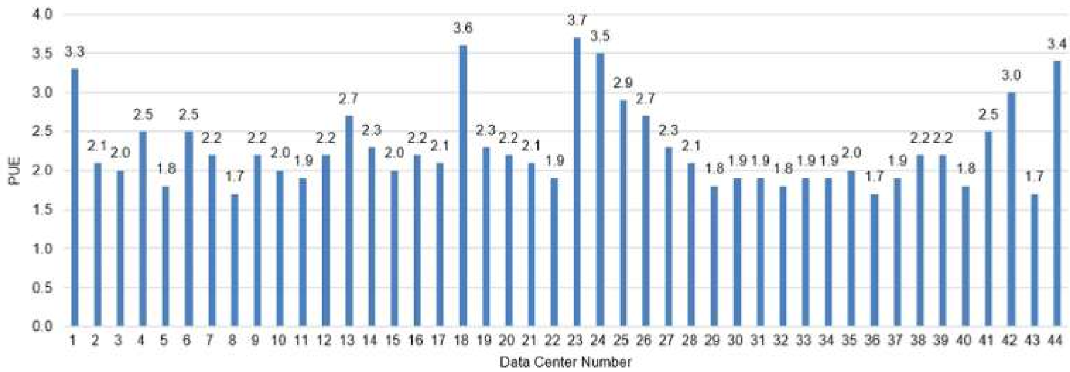

Figure 2.11 Annual average PUE of 44 data centres (Ni and Bai, 2017).

Given the annual average PUE of the data centres in Figure 2.10 and Figure 2.11 and taking into consideration the Figure 2.9 is possible to infer that energy efficiency in data centres can still be improved, which is encouraging for this project.

2.5.3 Current data centres energy efficiency

The current data centre energy efficiency is commonly measured using the Power Usage Effectiveness (PUE) metric (Brady et al., 2013; Shaikh et al., 2020). PUE is a ratio that represents the total energy consumed by a data centre divided by the energy consumed by the IT equipment alone (Brady et al., 2013). A lower PUE value indicates higher energy efficiency, as it means that a smaller proportion of energy is being used for non-IT purposes such as cooling and lighting.

The Green Grid Association has highlighted the importance of PUE and Data Centre Infrastructure Efficiency (DCiE) in understanding and improving data centre energy efficiency(Islam et al., 2019). DCiE is the reciprocal of PUE and represents the proportion of energy used by the IT equipment relative to the total energy consumed by the data centre (Islam et al., 2019). By monitoring and optimising PUE and DCiE, data centre operators can identify areas for improvement and implement energy-saving measures.

PUE primarily focuses on power efficiency and does not consider factors such as CO2 emissions and cost (Shaikh et al., 2020). However, PUE remains the most popular metric for measuring energy efficiency in data centres (Shaikh et al., 2020). Other metrics, such as Carbon Usage Effectiveness (CUE), have been proposed to address the environmental impact of data centres (Shaikh et al., 2020). CUE considers the carbon emissions associated with energy consumption and can provide a more comprehensive assessment of data centre sustainability.

Unlike the research presented by Shaikh et al. (2020) where the CUE is proposed, which found a way to address the effectiveness of the carbon usage, it is not consistent and addressing the issues in the overall picture. Therefore, the present study focuses on energy efficiency, and the use of PUE provides a better understanding of how energy efficiency plays in data centres.

The average PUE value for data centres can vary depending on design, location, and operational practices. A study analysing data from companies participating in the European Code of Conduct for Data Centre Energy Efficiency found that the average PUE in the European Union was around 1.8 (Avgerinou et al., 2017). However, this value may not be representative of all data centres globally.

In summary, the current energy efficiency of data centres is commonly measured using the PUE metric. Lower PUE values indicate higher energy efficiency, and data centre operators can use PUE and DCiE to identify areas for improvement. While PUE is the most popular metric, other factors such as

CO_2 emissions and cost should also be considered for a comprehensive assessment of data centre sustainability.

2.5.3.1 Relation between data centre energy consumption and their size

Maurice et al. (2014) highlighted that data centre energy consumption significantly contributes to the environmental impacts of ICT, including greenhouse gas emissions. There is a clear relationship between data centres' energy consumption and their size. Several studies have examined this relationship and found that larger data centres consume more energy (Shehabi et al., 2016). Dayarathna et al. (2016) found this through modelling data centre energy consumption. This is likely because larger data centres house more servers and equipment, which require more energy to operate. Hammadi and Mhamdi (2014) surveyed architectures and energy efficiency in data centre networks and found that the energy consumption of different data centre structures is influenced by their size. The study emphasized the importance of considering energy requirements in data centre architecture design.

Furthermore, Pesch et al. (2017) stated that data centres are large primary energy consumers, with energy consumed by IT equipment, server room air conditioning, and general building facilities. Shehabi et al. (2016) also highlighted the indirect energy impacts of data centres, such as the energy required to manufacture and transport equipment, which further contributes to their overall energy consumption. Space inside the data centre also accounts for the time it takes to cool or heat due to the server's usage. Thus, with regards to efficiency, the data centre's size and available space are critical part of the equation.

While larger facilities consume substantially more electricity in absolute terms, energy consumption alone does not necessarily indicate poorer efficiency. Avgerinou et al. (2017) examined data centre energy consumption trends under the European Code of Conduct for Data centre Energy Efficiency. The study found that larger data centres have higher energy consumption, requiring more cooling and infrastructure to support their operations. This confirms that larger data centres consume more electricity but are often more energy efficient.

Modern hyperscale data centres benefit from economies of scale, enabling more efficient utilisation of cooling infrastructure, electrical distribution systems, and computing resources. Centralised hyperscale facilities generally achieve lower PUE values because cooling systems can be optimised across larger installations, airflow management is more effective, server utilisation is higher and advanced cooling technologies such as direct liquid cooling can be deployed economically. Conversely, many smaller enterprise data centres operate below optimal utilisation and often rely on less efficient legacy cooling systems, leading to higher energy use per unit of computation. Therefore, although centralised facilities require considerably more electricity overall, the literature generally indicates that they are more energy efficient on a per-unit computing basis. Nevertheless, the rapid growth of AI workloads has increased power densities to such an extent that improvements in efficiency alone are no longer sufficient to offset increases in total electricity demand.

Overall, the evidence from these studies supports the claim that there is a relationship between data centres' energy consumption and their size. Larger data centres tend to consume more energy due to the increased number of servers and equipment they house and the additional cooling and infrastructure required to support their operations. A valuable addition to the literature on the relationship between the size and energy consumption of data centres has been made by the researchers above. However, it fails to address the usage effectiveness which could provide a fairer comparison in relation to energy consumption and the effectiveness of the usage of it.

2.5.3.2 Containerised Data Centres

Big companies such as Microsoft, eBay, and Google have adopted portable containerised modules to provide additional computing capacity due to low cost, modular design, relatively short time between construction and deployment, disaster recovery, sports events, or military applications (Qouneh et al, 2011). In sharp contrast to traditional (room-based) data centres, containerised data centres can be quickly built (typically within 1–6 months), shipped with trucks, and deployed with power and network in a few hours (Fried, 2009; Miller, 2012; Miller, 2013; Brocanelli et al., 2014). In 2005, Google deployed its first container-based data centre comprising 45 containers. Microsoft commissioned its own 225 container-based data centre in Chicago (Qouneh et al., 2011). In 2010, Taiwan put containerised data centres in their sight planning to apply the same principle to halve the cost of containerised data centres by creating a standard way for building data centres inside 20-foot shipping containers. However, Sun Microsystems popularised the idea in 2006 with the project Blackbox, an example of this design can be seen in Figure 2.12 (Nystedt, 2010).

As seen above some of the most famous companies in IT use this type of data centre, whose portability and capacity has been taken advantage of by various scientific communities (Qouneh et al, 2011).

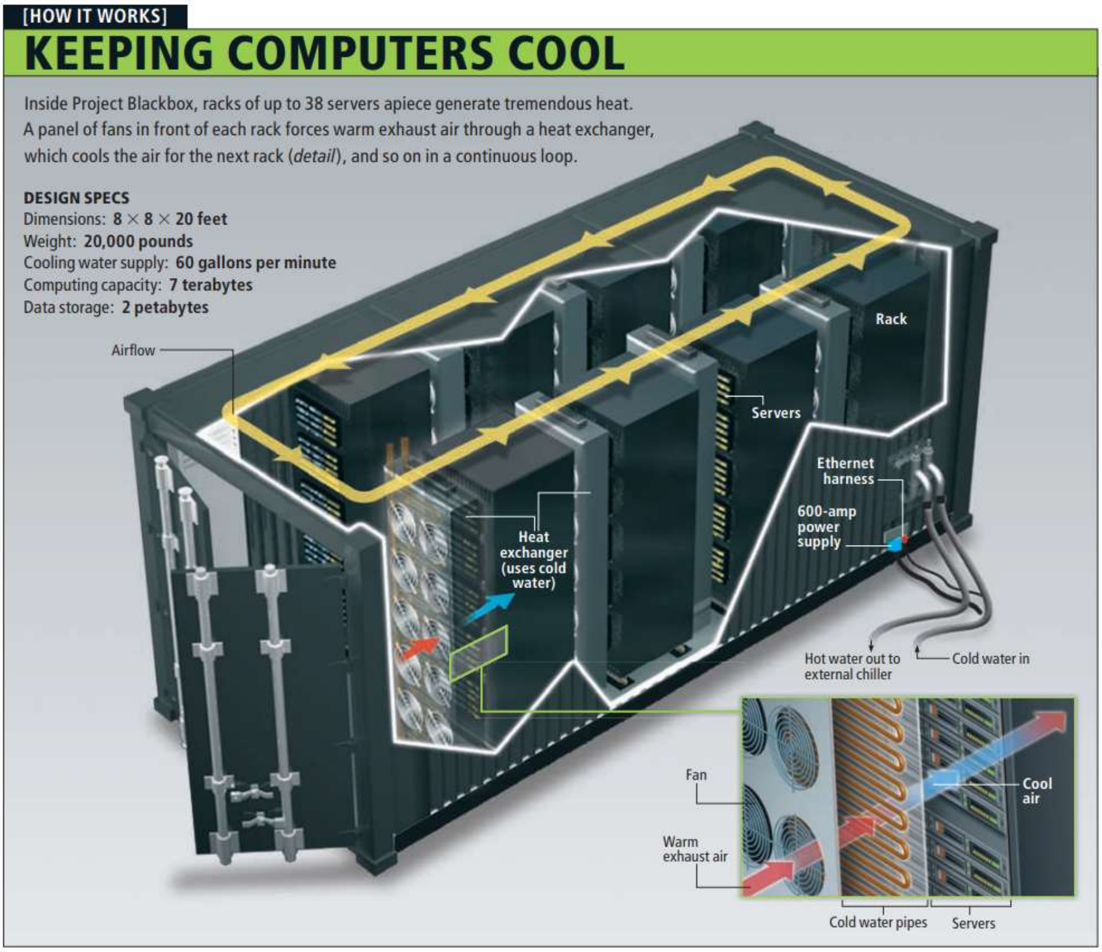

Figure 2.12 Design of "Blackbox" by Sun Microsystems (Waldrop, 2007)

The Industrial Technology Research Institute wanted to develop a standardised container data centre that costs half as much as systems used nowadays which is easy to use and saves energy (Nystedt, 2010).

Various hardware vendors have also introduced their own versions of containers like HP's Performance-Optimised Data centre (POD), Sun Microsystems' Blackbox, SGI's ICE Cube and Dell's Modular Data centre (Qouneh et al., 2011).

Another advantage of the containerised data centres is the concept mentioned by (Ward et al., 2012) and also used by Brenner et al. (2012), Brenner et al. (2009), Buccellato et al. (2011) called Environmentally Opportunistic Computing (EOC) in which instead of wasting heat, the heat is recovered to distribute it locally. The idea in this concept is to integrate the equipment with existing or new buildings where thermal waste can be used as heating as shown in Figure 2.13. Therefore, the way to reuse all the heat generated by these type of data centres brings to this project another perspective to reduce the energy consumption from a holistic approach.

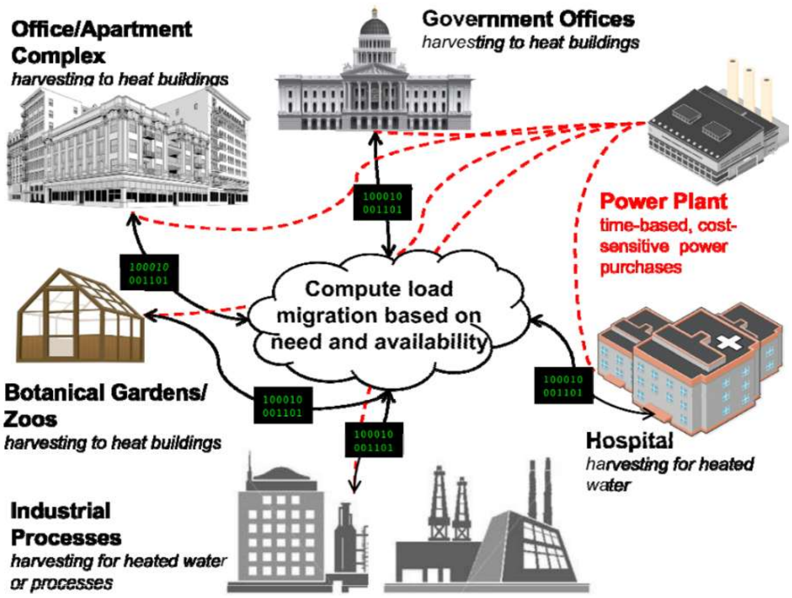

Figure 2.13 Schematic of fully implemented Environmentally Opportunistic Computing (Ward et al., 2012)

For experimental purposes, an important aspect to take into consideration is that the cooling system being used inside the containerised data centres has been established using the free cooling approach. The ambient air is pulled into the container which works as a cold aisle. The warm air at the back of the servers is used as a hot aisle in which heat is directly expelled to the ambient during summer, or it can be used to heat a specific space such as those shown in Figure 2.14 (Ward et al., 2012).

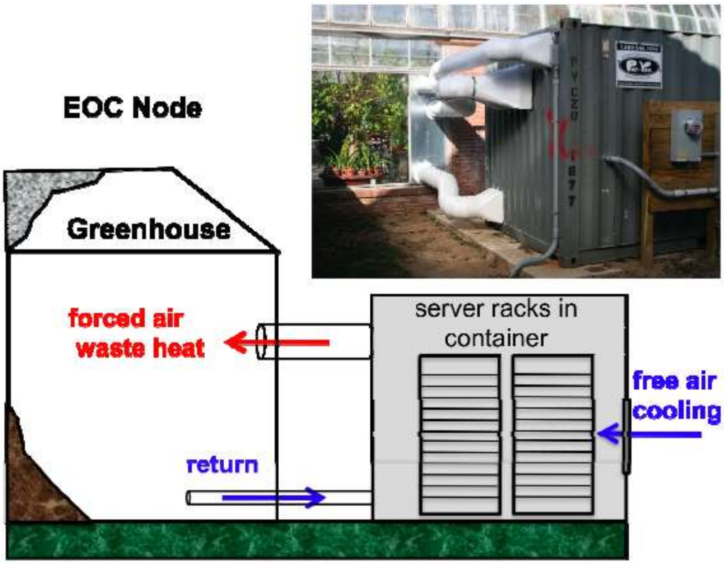

Figure 2.14 Schematic of prototype EOC node connected to a local greenhouse. The inset image shows a photo of the EOC node in-place. (Ward et al., 2012)

There are some advantages that must be considered when a containerised data centre is deployed, one of them is because of the space in the building, for example if there is no space in the building, using containerised data centres could be a great idea. The EOC concept allows the idea of the containerised data centres as distributed nodes attached to other-purpose buildings instead of big centralised data centres. Another advantage is that in the case of a disaster and a need to deploy fast thinking containerised data centres are plug and play and can be deployed quickly and at scale. However, the only way to increase their computational capacity is to add an entire other containerised data centre. The disadvantage of this modularity, is a single point of failure can eradicate a whole pod. The water and cooling systems remain distinct, so that pod cannot share the cooling capacity of neighbouring pods (Belady and Goolsbee, 2010; Kozlowicz, 2014).

2.6 Conclusion

This chapter reviews the literature on thermal management and energy efficiency in data centres, with a particular focus on containerised data centres. It highlights key theoretical perspectives, empirical findings, and methodological approaches. The review shows that, although significant progress has been made in understanding cooling behaviours, important gaps remain that justify this research. After examining air, liquid, and environmental cooling techniques, a recurring theme is the tension between theoretical advancements and real-world practicality. Most studies depend heavily on CFD simulations, often without validating thermal loads or airflow patterns in physical environments, especially in containerised data centres where space constraints and heat transfer dynamics differ considerably from traditional facilities.

The literature points to airflow mixing, resistance from servers, and limited controllability of air streams as major drawbacks of hot aisle/cold aisle configurations. Even with containment solutions, performance depends heavily on geometric and operational factors such as height, width, and layout, which underscores the need for more detailed, localised understanding of heat behaviour—not only at room level but also at chip and rack levels, where inefficiencies often originate. The difference in scale between studies focused on preserving the cold stream at room level (e.g. Choo et al., Mastelic et al.)

and those addressing heat generation at rack level (e.g. Wang et al.) illustrates the necessity of both perspectives; relying on just one is insufficient.

Energy consumption trends also reinforce the importance of this work. The wide range reported for cooling-related energy use (from around 30% to as high as 60%) indicates significant variability depending on configuration, climate, and operational practices. This suggests that improvements in thermal management could lead to notable gains in overall energy efficiency, particularly in containerised data centres where design options are constrained. PUE, as the primary metric, is another key point; while it is useful and widely adopted, it does not fully capture environmental or operational aspects. Nonetheless, PUE remains a practical and consistent metric for evaluating energy efficiency. Recent growth in AI workloads has further reinforced the importance of developing more energy-efficient thermal management strategies, particularly for containerised data centres where space and cooling capacity are inherently constrained.

The literature on containerised data centres demonstrates strong interest in modular, rapidly deployable systems, but also reveals a lack of detailed experimental work on their thermal performance. Although promising concepts—such as Environmentally Opportunistic Computing (EOC) and waste heat reuse—exist, these ideas depend on a deeper understanding of thermal behaviour than current studies typically provide. Another important factor, highlighted by Budiyanto et al., is the impact of solar radiation on containers, an often-overlooked aspect in CFD studies.

In conclusion, there is a clear need to experimentally characterise the thermal and energy performance of containerised data centres under realistic conditions, considering the unique constraints that differentiate them from traditional data centres. A more comprehensive understanding of heat behaviour can support improvements in energy efficiency. By addressing the limitations identified in existing research and prioritising real-world measurements, this work aims to contribute towards a more reliable and practical approach to thermal management in containerised data centres.

Chapter 3

Background and Literature Review on Computational Workloads

3.1 Introduction

This chapter provides an introduction to the three different types of workloads, for example, services, data analysis, and data processing, that data centres utilise and the way the scheduler should manage the number and distribution of hotspots within a containerised data centre through workload optimisation. This minimises the need for air-cooling systems, so they are activated only when overall temperatures approach levels that risk the integrity of electronic equipment.

3.2 Data centre workload types

Workload characterisation is key to efficient data centre management (Qiu et al., 2019). Data centre workloads can be classified into different types based on their characteristics and objectives, and this affects how they are managed within a data centre. One classification divides data centre workloads into services or data analysis workloads (Jia et al., 2013a). When a service workload is being processed in the data centre, as it is continuous and sustained (Han et al. 2016), which generates a lot of heat, it is concentrated in one part of the data centre, and only a smaller workload should be accepted. This also needs to be continually relocated. However, for data analysis workloads, despite the large volume of data throughput, intermittent processing and efficient data processing techniques mean less heat is generated; therefore, this does not need to be relocated that often(Zhan et al., 2012; Jia et al., 2013). Zhan et al. (2012) benchmark data analysis workloads as requests per minute and the amount of data processed per minute; whereas Han et al. (2016) focuses on benchmarking the different types of workloads in arrival patterns, using sequences of time-stamped requests or jobs to measure how long it takes the data centre to complete the request. The reason for this difference in definitions is due to the focus of the algorithms that the different groups of authors are evaluating. In relation to the current project, Han et al. (2016) view is more appropriate to what the study will be investigating, given that it involves the time that the data centre servers are active to deal with the workload and therefore it is imperative to manage the consequent heat produced inside the data centre when the servers are active or idling.

Zhan et al. (2012), Jia et al. (2013a) and Zhou et al. (2020) argue that the workload should be studied, classified and categorised before planning; however, Qiu et al. (2019) emphasise that workload characterisation can be automated post-deployment, thereby enabling scalable performance trade-offs. As an automated approach, the characterisation post-deployment improves the accuracy of workload categorisation. Ilavarasan and Thambidura (2007a) also support the post-deployment characterisation, stating that task scheduling can be performed at compile time or at run time. They both emphasise that workload characterisation is key to achieving high performance in heterogeneous computing systems, which is crucial for optimising resource allocation in heterogeneous large-scale environments. This allows a better understanding of the heat behaviour inside the data centre, and in consequence an improved hotspot management.

Qiu et al. (2019) and Ilavarasan and Thambidura (2007) conclusions that characterisation should be post-deployment is more likely suitable for data centres, given that the workload can vary often. However, this is less suitable for virtual machines and cloud computing due to greater variation in available resources. Until the data centre is fully deployed and processing, workload allocation in a cloud-based data centre cannot be efficiently managed. However, they are outside the scope of this thesis.

3.3 Scheduling

Scheduling plays a pivotal role in computing in Data Centres. At its core, scheduling involves allocating resources over time to achieve specific objectives efficiently and effectively (Liang et al., 2020). Nonetheless, the definition presented by Fan et al. (2018) includes other aspects necessary for this research, such as identifying the tasks to be completed, their attributes, and the resources required to carry them out. On the other hand, Liang et al. (2020) emphasise the use of first-come-first-served (FCFS), backfilling, and priority queues, which are basic scheduling methodologies. Although scheduling methodologies such as the ones presented by Li et al. (2020) are widely used when data centres are extremely difficult and time-consuming to configure, identifying and allocating resources over time is not enough to achieve objectives efficiently and effectively. As stated by Liang et al. (2020), when no plan is in place, these omissions severely compromise the data centre's performance, flexibility and usability. A plan that details when and how tasks will be conducted, the criteria by which to measure the quality of the task and any factors which may impact the planning of the schedule first needs to be considered (Fan et al., 2018). The Scheduler's function is assigning tasks to processors, allocating processing time for computer tasks, or orchestrating the order of task execution to optimise resource utilisation while meeting various constraints and objectives (Fan et al., 2018).

The literature widely reports that the constraints usually fall within the CPU, bandwidth, RAM, and QoS levels. Algorithms such as the one designed by Mastroianni et al. (2013a) use an approach to manage the workload evaluating the resource utilisation such as CPU and RAM. While the study presented by Mastroianni et al. (2013b) provided insights into the assignment and migration procedures for VM consolidation, it failed to address scalability with more than 100 servers, which is a disadvantage considering the rapidly increasing need to provide services in the fastly growing digital infrastructure. While this study is valuable to the literature on job scheduling, it also overlooks the energy consumption of the cooling systems in the data centres, which has been widely reported to be between 30% to 60% of the data centre's total energy consumption (Song et al. (2015a), Liu et al. (2016), Brocanelli et al. (2014), Chu et al. (2019), Almoli 2013 and Qouneh et al. (2011)). Further, these studies recognise the importance of the cooling system energy consumption when the objective is to reduce the energy consumption of data centres. Although the algorithm introduced by X. Li et al. (2018) attempts to address energy consumption by cooling systems, their focus is mostly on CPU temperature. This approach can be criticised as inefficient given that the CPUs, although contributing to the change of temperature inside the data centre, are not the only factor to be considered. Other devices, such as network racks and energy backup equipment, account for the temperature inside the data centre and the CRAC's impact on energy consumption.

3.3.1 Basic Terminology

Tasks/Jobs are the units of work that need to be scheduled. Task scheduling can have complex structures, for example those found in MapReduce require careful consideration of their interdependencies and execution order (Crăciun and Salomie, 2017; Im et al., 2018; Tim Yuan et al., 2018; Kulkarni and Li, 2018a). However, this thesis does not focus on tasks/jobs. It focuses on **Workloads**, which the tasks and jobs are part of, and their partition within the data centre.

Resources are the entities required to execute tasks. They could be processors, nodes, a different data centre, or tangible or intangible assets required for task completion. The dynamic nature of the data centre's workload means that the availability of resources can fluctuate, which demands strategies to optimise resource management (Crăciun and Salomie, 2017). The latest approaches to this issue involve resource virtualisation; this allows a more flexible allocation of resources. While the virtualisation approach seems to be an advantage, it also leaves room for improvement. For example, X. Li et al. (2018) addresses issues with virtualisation but does not contemplate the scenarios where the VMs are under-provisioned, which, depending on the demand, could cause QoS (Quality of Service) issues given the lack of resources due to their unavailability.

Schedule: A schedule represents the allocation of resources to tasks over time. It provides a detailed plan outlining when each task will be executed, by which resource, and for how long. Several factors influence the scheduling process, such as the physical distance between data centres, resources (Memory, GPU, CPU, Storage), and limitations of the network (bandwidth, latency, availability, usage, etc.), all of which play an important part in the performance of scheduling algorithms (Havet et al., 2017; F. Alves and Kabbani, 2022). While F. Alves and Kabbani's (2022) scheduling algorithm focuses more on bandwidth and latency, the algorithm developed by Havet et al. (2017) categorises the workload in containers according to the resources that the workload requires: memory-heavy or CPU-heavy. Thus, schedules are oriented to certain objectives.

Objective Function: This is the criterion used to evaluate the quality of a schedule. Typical objectives include minimising the total time to complete all tasks, maximising resource utilisation, meeting deadlines, or minimising costs. The research indicates that scheduling in data centres often involves different approaches. For example, the scheduling algorithm proposed by Mtech and Kamal (2014) focuses on balancing the overall IT energy consumption with task completion deadlines by using backfilling and lookahead techniques, which prevent the underutilisation produced by the First Come, First Served (FCFS) Scheduling technique. In this case Mtech and Kamal's (2014) focus was directed to reduce the underutilisation produced by the FCFS technique, which leaves the nodes that are available idling while the number of nodes required becomes available. On the other hand, the genetic algorithm proposed by Sui et al. (2019) load balances the workload received by forecasting the utilisation, which, in the case of cloud data centres, reduces the migration costs, network traffic and performance interference by overloaded servers. This research recognises the importance of the objectives or function of scheduling; therefore, it is note worthy that the algorithm developed in the following chapters aims to reduce energy consumption.

3.3.2 Classification of Scheduling Types

As mentioned above, different types of task scheduling occur according to a variety of features depending on the data centre capabilities, availability and usage, whether the tasks can be run discretely or whether they are dependent on each other, and their classification depends on the characteristics that suit those features. These features impact the sequence in which the tasks will be conducted.

3.3.2.1 Characteristics

Single Task Scheduling: In this scenario, tasks are processed in a single core, and the objective is typically to minimise makespan or total completion time;therefore, no other tasks are processed until the task is completed (Ilavarasan and Thambidura, 2007).

Parallel Task Scheduling: Tasks can be executed simultaneously on multiple machines. The goal is to allocate tasks to machines optimally, considering factors such as machine capabilities and task dependencies(N'Takpe et al., 2007; Mtech and Kamal, 2014; Barzegar et al., 2019).

Heterogeneous Job Scheduling involves scheduling tasks on multiple nodes with different capabilities and constraints. Each task may require a sequence of operations on different machines Different processors from different architectures and types of resources, such as GPUs, CPUs, Co-Processors, Network environments, and different types of peripheric might be required for specific tasks(Ilavarasan and Thambidura, 2007; N'Takpe et al., 2007; Szmajduch and Kolodziej, 2015; Li, 2019; Akbari et al., 2020).

Homogeneous Job Scheduling: The scheduler distributes tasks that can be run in parallel or have no dependency on one another. Tasks are scheduled to run on multiple nodes with the same capabilities and constraints.

Concurrent Job Scheduling: This type of scheduling allows a job to be split into several tasks. Some of them might depend on one another whereas others may run in parallel. The scheduler reads the properties of each job and the requirements for each task to obtain the same result regardless of the order in which the task was executed, as long as they do not depend on another task currently being processed(Kulkarni and Li, 2018b; Yin et al., 2020).

With the different characteristics presented in section 3.3.2 the set of instructions and resources available to deal with the workload can be resolved using two techniques, which are related to the type of resources available: either local or geographic. However, since this research is interested in cooling energy, these will be categorised according to their saving energy focus. Overall IT Energy Consumption or Cooling Energy Consumption.

3.3.3 Local Scheduling

First, local scheduling algorithms have a set of instructions to use the resources available in a data centre and can be defined as the way an algorithm reacts to allocate where the workload should be redirected inside a data centre. It is local because even if the data centre is comprised of a cluster of data centres, the allocation is limited to the one where the action is being triggered. Implementing local scheduling algorithms can help optimise energy usage, and the best allocation of the computational resources, which is crucial because energy efficiency in data centres is critical.

Beloglazov and Buyya (2010) research survey proposed architectural principles for the energy-efficient management of clouds, resource allocation policies, and scheduling algorithms that were limited to virtual machines (VMs). They emphasised the importance of considering the QoS expectations and power usage characteristics of devices in resource allocation decisions. However, the study primarily focused on reducing server energy use and did not explicitly consider the thermal conditions of the data centre and their associated cooling energy demand.

The multi-objective co-evolutionary scheduling algorithm proposed by Lei et al. (2016) selects computing nodes, supply voltages, and processor frequencies while considering renewable energy availability. Likewise, Iturriaga and Nesmachnow (2016) developed a scheduling approach combining

Dynamic Voltage and Frequency Scaling (DVFS) with hybrid energy sources to reduce dependence on non-renewable energy. Both approaches focus primarily on reducing the energy consumed by IT equipment rather than directly reducing cooling demand.

Other scheduling methods, such as Best Heuristic Scheduling (Peng et al., 2020), energy-saving resource allocation strategies (Li et al., 2017), and Peak Efficiency Aware Scheduling (Wong, 2016), similarly seek to improve the energy efficiency of computing resources through workload consolidation and server-level optimisation. While these methods can indirectly influence cooling requirements by lowering server power consumption, cooling energy is not explicitly considered within their scheduling objectives.

In contrast, thermal-aware scheduling approaches establish a more direct relationship between workload allocation and cooling energy demand. Varsamopoulos et al. (2009) developed thermal models to represent heat generation and recirculation within data centres and incorporated these models into scheduling decisions. Their work demonstrated that workload placement can influence thermal distribution and hotspot formation, thereby affecting the operation of cooling systems. Among the proposed models, the stepwise linear model provided the most accurate representation of heat recirculation and thermal interactions within the data centre.

The distinction between IT-energy-aware and thermal-aware scheduling is particularly important for this research. Many existing local scheduling approaches reduce the energy consumed by computing resources, which may indirectly reduce cooling requirements. However, only thermal-aware approaches explicitly consider the relationship between workload placement and cooling demand. Consequently, these studies provide a foundation for understanding how workload distribution can be used not only to improve computing efficiency but also to influence a facility's thermal behaviour and cooling requirements.

3.3.4 Geographic Scheduling

Geographic scheduling extends workload allocation beyond a single data centre by distributing workloads across multiple geographically dispersed facilities. Unlike local scheduling, geographic scheduling can exploit differences in climate conditions, electricity prices, renewable energy availability, and cooling efficiency between locations.

Grange et al. (2018) examined scheduling strategies for data centres powered by renewable energy sources, highlighting the potential Green IT Scheduling using data centres powered with renewable energy. Similarly, Liu et al. (2022) investigated scheduling policies that incorporate thermal conditions and heat recovery opportunities across different locations. Although their primary focus was not cooling energy reduction, the study demonstrated how location-specific characteristics can influence the energy performance of data centre operations. Guo et al. (2014) instead explored the use of geographical load balancing, opportunistic scheduling of delay-tolerant workloads, and thermal storage management to increase renewable energy utilisation and reduce dependence on conventional energy sources. Their findings showed that workload placement decisions can be influenced by environmental and operational conditions that vary across locations. Although the work of Grange and Liu are focused on the usage and storage of renewable energy, the geographical load balance presented by Guo shows a larger decrease in non-renewable energy usage.

Mahmood et al. (2022) presented a simulation framework for energy-efficient and secure job scheduling in geographically distributed data centres. The framework demonstrated how workload migration between sites can improve overall operational efficiency by considering differences in resource availability and energy characteristics. Similarly, Fernández-Cerero et al. (2018) examined energy policies and scheduling strategies for data centres of varying sizes and workloads, providing methods for identifying operational inefficiencies and improving energy management practices. Mahmood's and Fernandez-Cerero's models are useful to highlight operational and energy management inefficiencies, which, although can be translated into energy savings do not explicitly measure the amount of energy consumption.

Alternatively, F. Alves and Kabbani. (2022) study has investigated geo-distributed scheduling algorithms designed to optimise workload placement across multiple sites. Their proposed algorithm SWAG incorporates geographical information into scheduling decisions to improve resource utilisation and reduce energy consumption. Other approaches, including DVFS-based scheduling and energy-aware resource management frameworks (Beloglazov and Buyya, 2010; Liu et ak., 2022), similarly demonstrate the benefits of incorporating energy considerations into workload allocation decisions across distributed infrastructures.

The reviewed geographic scheduling literature highlights that workload placement decisions can be influenced by factors beyond computational efficiency, including local environmental conditions, renewable energy availability, and infrastructure characteristics. Nonetheless, most studies focus on reducing overall energy consumption or increasing renewable energy utilisation, and the models highlight inefficiencies rather than specifically targeting cooling energy demand.

This thesis builds upon these scheduling concepts by using workload relocation as a mechanism to reduce cooling energy consumption. Rather than relocating workloads solely to minimise IT energy use or maximise renewable energy utilisation, the proposed method considers the thermal performance of workload within data centres. Insights from thermal-aware scheduling demonstrate that workload placement influences heat generation and cooling requirements, while geographic scheduling studies show that workloads can be transferred between facilities with different environmental and operational conditions. By combining these perspectives, the proposed approach evaluates whether relocating workloads to data centres with more favourable cooling conditions can reduce the cooling energy required to maintain acceptable operating temperatures.

3.3.5 Temperature reactive

Energy efficiency in data centres is a critical concern due to their high energy consumption. Several algorithms have been proposed to optimise energy usage in data centres, with a focus on temperature as a reactive factor.

Yan et al. (2018) studied the energy performance of workload allocation strategies based on chip and inlet temperature awareness. Their results showed that the chip temperature-aware strategy improved energy efficiency by preventing over-cooling and optimising the cooling air supply. This approach was particularly effective in scenarios with variable-flow cooling air supply.

In terms of temperature, Arroba et al. (2018) emphasised the importance of understanding the relationship between power, consolidation, and performance for enabling energy-efficient management at data centre level. Similarly, Tang et al. (2007) developed their thermal-aware task-scheduling algorithms to minimise heat recirculation in data centres. By doing so, they found they were also minimising cooling energy costs and optimising task scheduling. Additionally, Sun et al. (2021) developed prototype energy models in support of energy-saving measures and codes for data centres. These models cover various climate zones and can be used to evaluate energy-saving measures and improve data centre energy efficiency.

The algorithms discussed above improve energy efficiency in data centres, particularly by considering temperature as a reactive factor. These algorithms include workload allocation strategies, dynamic management of computing and cooling energy and thermal-aware task scheduling. By optimising these factors, data centres can reduce energy consumption and improve overall efficiency.

3.4 Conclusions from the Literature Review

Cooling the air to the requisite temperature to lower the data centre temperature requires a lot of energy, which is becoming increasingly expensive because data centres are growing in density and not in space. It is a worldwide concern for data centre owners to find different ways to improve the cooling inside the containerised data centres and for researchers to investigate them to achieve energy efficiency through targeted air circulation cooling techniques.

Qouneh et al. (2011) provides a comparison between container-based and raised floor data centres, where different cooling configurations were tried. They put the same amount of IT equipment in both data centres; however, because the containerised data centres have less space than the raised floor data centres, this increased their density and workload, negatively affecting the cooling performance. Their conclusion was that when the containerised data centre maintains optimal IT equipment density, its workload is at mid-capacity, and a raised-floor data centre matches this density and workload, the containerised data centre achieves 80% savings in cooling power compared to a raised-floor data centre. In terms of high workload, multiple container-based data centres achieved 46% savings in cooling power compared to a raised-floor data centre. When a raised-floor data centre operates at low utilisation, it can approach the performance of a container that employs no intelligent cooling. But the paper would appear to be overambitious in its claims because the conditions are not comparable, for example: “multiple container-based data centres achieve 46% savings in cooling power compared to a raised-floor data centre when highly loaded”. The findings would have been more convincing if one container-based data centre had been compared to one raised-floor data centre. It can be perceived from this sentence that multiple container-based data centres were used together to level the same amount of IT that the raised-floor data centre was using.

The study would have been more informative if the researchers had also considered comparing the efficiency of each cooling configuration of every container-based data centre that they presented in the paper. Therefore, this provides a gap to perform some CFD experiments to study the findings for each cooling configuration, and a way to explore the behaviour of these cooling configurations using scheduling techniques. Also, it could be important to add the variable external heating radiation (sun) effect in the thermal efficiency according to the exposure time of the container to sunlight depending on its location and the season. A previous study was performed but there was only a one day simulation which could not replicate all seasonal daylight variation and how this could affect the thermal management inside the containerised data centre.

Due to the nature of a containerised data centre, it is necessary to quantify the effect of the temperature of the returned air on the data centre internal temperature, the extent of supply air bypassing the racks and its effect on the internal temperature, the rack cooling indexes to measure the health of every rack inside the data centre while testing scheduling techniques and the recirculation ratio to quantify the amount of hot air that goes into the IT equipment.

As shown in Section 3.3, researchers in the area have widely reported task scheduling as a tool to help in the distribution of workload evenly among the available processors to improve the performance and efficiency in solving tasks. This brings interesting insights for this project: for example, the hotspots inside a containerised data centre during summertime and how they change in winter; how the thermal behaviours inside the container-based data centre change; how this improves with workload scheduling; and how this affects energy consumption in the data centre.

Chapter 4

Computational Methodology

4.1 Introduction

This chapter presents the development of a computational fluid dynamics (CFD) model of a containerised data centre, based upon the experimental system described by Wang et al. (2017). The theory being explored is that managing the number and distribution of hotspots within a containerised data centre through workload optimisation can reduce energy consumption by minimising the need for air-cooling systems. These systems would only be activated when overall temperatures approach levels that risk the integrity of electronic equipment. This kind of simulation represents the conjugate heat transfer in this model. The chapter describes the experimental system and explains the associated assumptions, equations, and conditions underlying the CFD model. The CFD analysis is built up in steps of increasing geometrical complexity, leading to a model of the containerised data centre. The predictions of the model are validated against the experimental observations, and the methodology for identifying hotspots is explained.

4.2 Data centre configuration and previous work

The containerised data centre proposed by SUN is one that utilises liquid cooling (Waldrop, 2007). Given the potential hazards associated with liquid cooling in the data centre environment —specifically, the risk of leakage and related insurance complications —this configuration has been deemed unsuitable for further consideration. The type of data centre proposed by Oró et al. (2018b) depends on having a pool or a heating system to discharge the heat transfer obtained by the heat rejection. Given that this is not always possible to have or take advantage of, it was also discarded as a possibility to be studied. Alternatively, the containerised data centre proposed by Ling et al. (2017b) utilises raised tiles but these reduced the possibility of mobilising the data centre from one place to another. Since portability and ease of installation are key elements of containerised data centres, this was also discarded. Thus, the proposed model of a containerised data centre is inspired by Wang et al. (2017): a shipping container modified with an overhead air supply cooled by a chiller, which meets these requirements. Additionally, the study by Wang et al. (2017) also provides valuable experimental measurements.

The system utilised by Wang et al. (2017) is an experimental set-up based on implementing modifications to include a Computer Room Air Conditioning (CRAC) system, with an overhead duct that feeds into seven grills set on the ceiling of the container. The container's interior is fitted with heaters that approximate the heat generated by the servers inside a data centre. The layout of the racks is a Cold-Aisle, Hot-Aisle arrangement which sets the middle aisle as the cold aisle, being the front of the racks. The back of the racks represent the hot-aisle in this approximation. Thus, the present research develops a Computational Fluid Dynamics (CFD) model based on this experimental configuration to investigate whether the cooling demand of the air-conditioning system can be reduced while maintaining server inlet temperatures within acceptable operating limits.

The geometry based in the work of Wang et al. (2017), and the conditions and behaviour inside a containerised data centre are modelled in this research to find insights that might be useful to achieve energy efficiency in this kind of data centre, which comprises cold and hot aisle containment. The purpose of the next stage of these CFD simulations is to validate predictions against the literature and show that the system can be modelled; the design is illustrated in Figure 4.1.

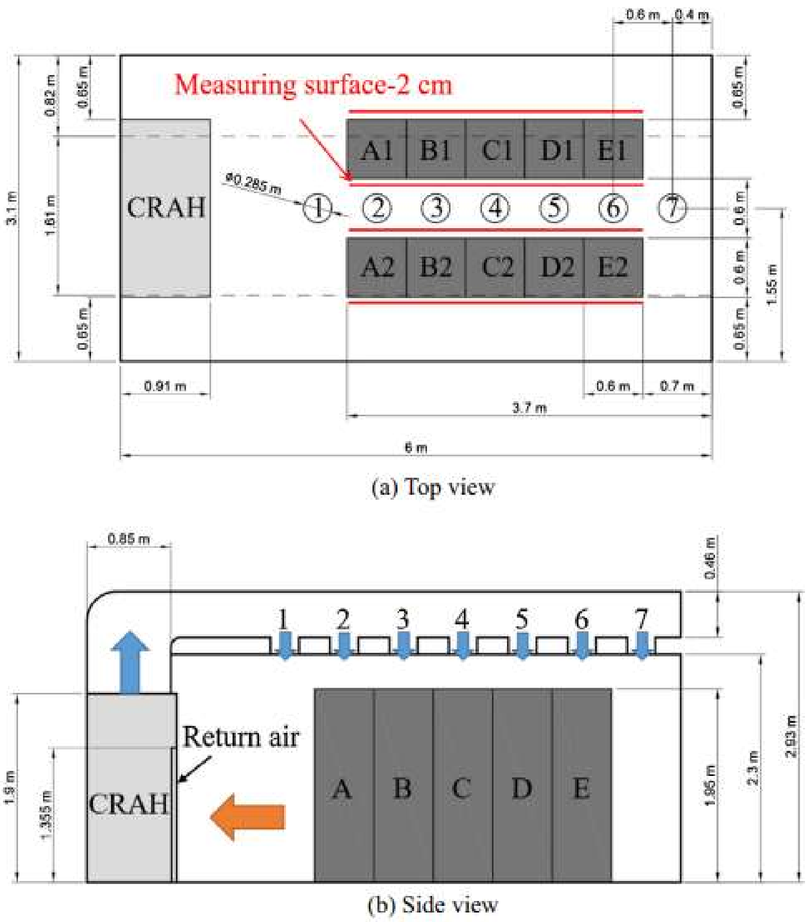

Figure 4.1 Experimental containerised data centre configuration reported by Wang et al. (2017), showing the CRAH unit, ductwork and rack arrangement used as the basis of the CFD model.

To investigate the airflow and heat transfer mechanisms within the containerised data centre, a Computational Fluid Dynamics (CFD) model was developed based on the experimental configuration proposed by Wang et al. (2017). The first stage of the modelling process involved identifying the governing physical processes, boundary conditions and modelling assumptions required to reproduce the experimental system. To reproduce the airflow and thermal behaviour of the containerised data centre, the CFD methodology was developed progressively using a hierarchy of increasingly complex computational domains. Rather than modelling the complete containerised data centre directly, the simulations began with a simplified server-scale configuration consisting of a single porous server enclosed within a wind-tunnel domain. This simplified geometry allowed the conjugate heat transfer, airflow behaviour and porous-media representation of the server to be validated while keeping the computational cost low. After the modelling approach had been verified, the geometry was progressively expanded by increasing the number of servers, first to a complete row of servers and subsequently to the half-containerised data centre. This staged approach ensured that the numerical model remained computationally efficient while preserving the dominant physical mechanisms governing airflow and heat transfer.

Once confidence had been established in the simplified numerical model, the computational domain was progressively expanded by increasing the number of servers to represent complete rack configurations before developing the half-containerised data centre model used in the subsequent investigations. This hierarchical modelling strategy ensured that each increase in model complexity was supported by a validated numerical methodology while reducing the need for repeated simulations of the complete data centre during model development.

The development of the CFD model is described in the following sections, beginning with a simplified server-scale validation case before progressing to the reduced containerised data centre used for the numerical investigations.

4.2.1 Server-scale test problems

Before modelling the complete containerised data centre, a simplified CFD model was developed to validate the physical modelling approach. Simulating the full geometry from the outset would have required a substantial computational effort while making it more difficult to identify modelling errors or inappropriate boundary conditions. Therefore, a wind-tunnel configuration containing a single porous medium representing an individual server was selected as the initial validation case. The computational geometry is illustrated in **Figure 4.2**. This configuration preserves the essential airflow and heat transfer mechanisms of a server while significantly reducing the size of the computational domain. Consequently, the influence of mesh resolution, boundary conditions and porous-media properties could be assessed independently before these modelling choices were transferred to the larger data-centre simulations.

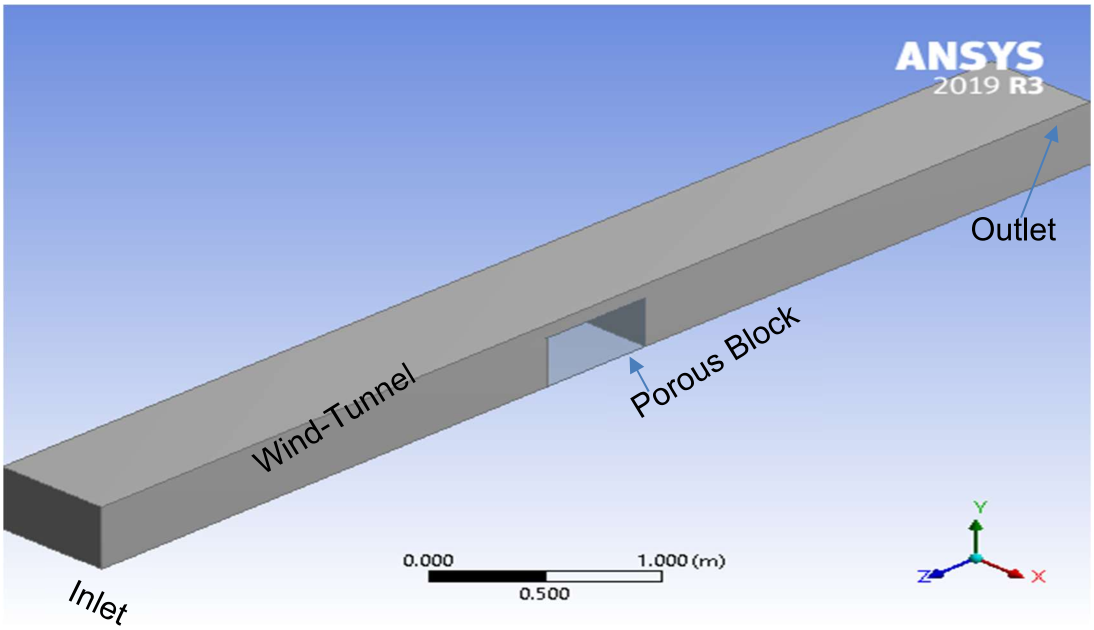

Figure 4.2 Computational domain used for the server-scale validation study, illustrating the porous server representation within the wind tunnel.

The server was represented as a porous medium block rather than explicitly modelling the detailed internal arrangement of processors, heat sinks and electronic components. This approach preserves the dominant pressure losses experienced by airflow through a server while significantly reducing the computational effort required for mesh generation and solution. Airflow resistance within the porous region was introduced through the porous medium model available in ANSYS Fluent. The porous region therefore acts as a distributed momentum sink that reproduces the pressure loss associated with airflow through the internal server components. Heat generated by the electronic equipment was represented by a uniform volumetric heat source within the porous region, allowing heat to be transferred to the airflow through conjugate heat transfer.

In the experimental configuration of Wang et al. (2017), airflow through each server is generated by multiple axial fans located at the rear of each rack. Explicitly modelling these rotating fans would require highly refined meshes and substantially increase computational cost. Instead, the present work represents their effect using a distributed momentum source within the porous medium region. This approximation reproduces the overall pressure rise generated by the fans while avoiding the computational expense associated with modelling rotating machinery.

Once the simplified geometry had been established, a mesh independence study was carried out to ensure that the numerical solution was not influenced by the discretisation of the computational domain. A structured mesh was selected because of the regular geometry of the wind tunnel and porous block, providing good numerical accuracy while minimising computational cost. Several mesh densities were evaluated by systematically increasing the number of cells in both horizontal and vertical directions, while monitoring the predicted velocity and temperature fields.

The resulting structured computational mesh is shown in **Figure 4.3**.

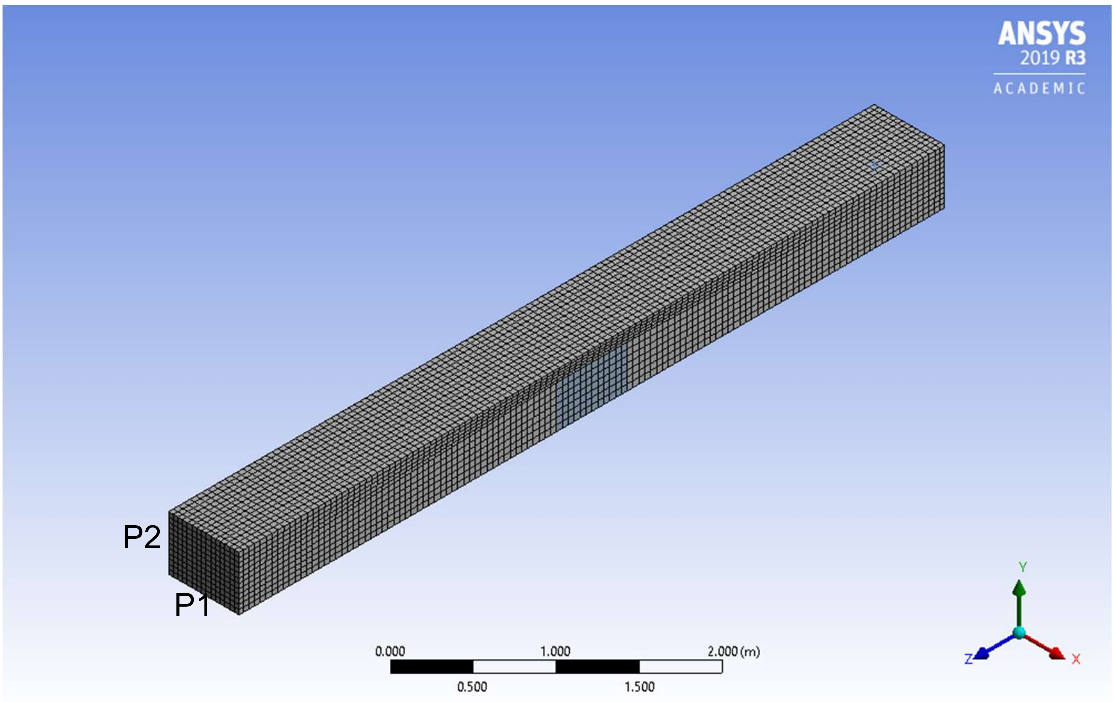

Figure 4.3 Structured computational mesh adopted for the wind-tunnel validation case.

Mesh independence was assessed by varying the number of cells along the horizontal (P1) and vertical (P2) directions, while maintaining a uniform cell size throughout the computational domain (see **Figure 4.3**). The number of cells in the remaining regions of the domain was determined automatically to preserve a consistent cell size. The resulting meshes are summarised in **Table 4.1**. The area-weighted average outlet velocity and temperature, together with the average porous-medium temperature, were monitored to assess mesh independence. Percentage errors were calculated relative to the finest mesh. This is also shown graphically in **Figure 4.4**.

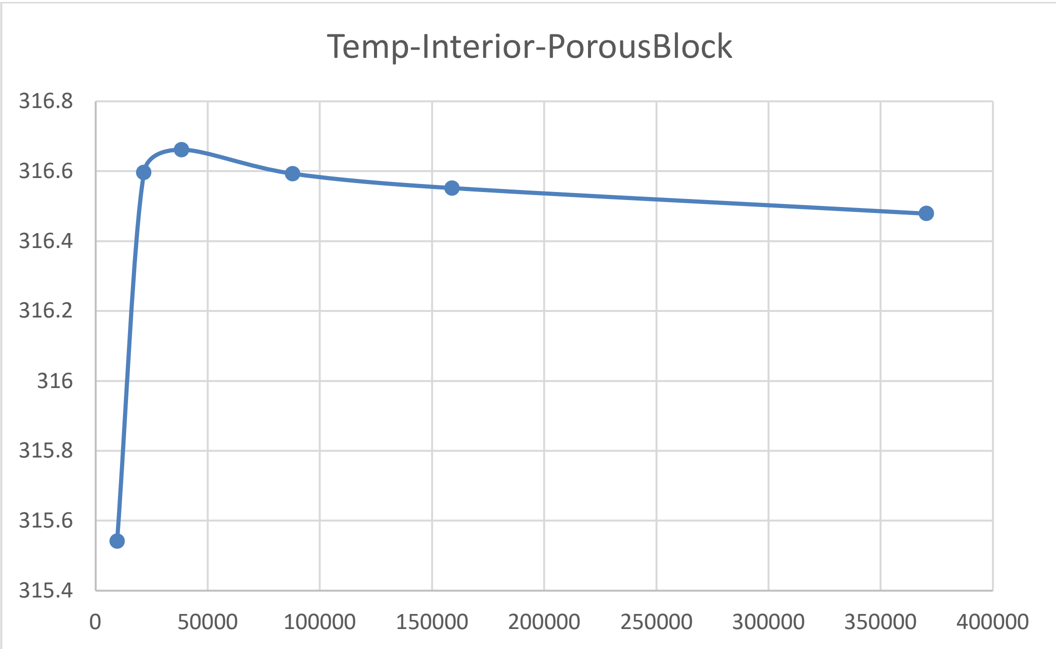

Figure 4.4 Independence Mesh Study chart, the Y axis shows the average temperature in the interior of the porous block expressed in K (kelvins) and number of mesh elements in the bottom axis.

Name	P1	P2	P3	P4	P5	P6	P7	P8	P9
DP 0	10	8	12286	9608	4.0180088	2.1321114	312.39555	315.54222	0.30
DP 1	15	12	25496	21528	4.0210183	2.0464599	312.50618	316.59652	0.04
DP 2	20	16	43646	38368	4.0197322	2.0415991	312.71345	316.66163	0.06
DP 3	30	24	95840	87768	4.0148059	2.0464336	312.58856	316.59325	0.04
DP 4	40	32	170026	158928	4.0118271	2.0490089	312.49934	316.55176	0.02
DP 5	60	48	388214	370368	4.0089071	2.0540062	312.46268	316.47903	0.00

Where:

P1 - Horizontal Sizing Number of Divisions
P2 - Vertical Sizing Number of Divisions
P3 - Mesh Nodes
P4 - Mesh Elements
P5 - awa-vel-outlet [m s^-1]
P6 - awa-vel-interiorpb [m s^-1]
P7 - awa-temp-outlet [K]
P8 - awa-temp-interiorpb [K]
P9 - Percentage error

Table 4.1 Independence Mesh Study detailed results.

The results demonstrate that further mesh refinement produced negligible changes in the predicted temperatures and velocities. Consequently, mesh DP1 was selected for the remaining simulations because it provided a satisfactory compromise between solution accuracy and computational cost.

In addition to mesh independence, the quality of the computational mesh was assessed to ensure that the discretisation was suitable for the prediction of conjugate heat transfer. The structured mesh produced predominantly orthogonal elements with smooth transitions in cell size and avoided excessive skewness throughout the computational domain. The structured mesh provided adequate resolution adjacent to the porous block surfaces to represent the thermal gradients between the porous medium and the surrounding airflow.

Following the establishment of a mesh-independent solution, a sensitivity study was conducted to investigate the influence of inlet air velocity on the airflow and thermal behaviour of the simplified server model.

4.2.2 Boundary conditions for the server-scale model

The boundary conditions for these experiments were set as it follows: The porous medium was assigned an inverse permeability of 5290000 m^2 with a 0.5 constant of porosity and an interior heat flux generation of 293W/m^3.

The inlet air temperature was fixed at 294.15 K, while the inlet velocity was varied between 1 and 4 m s^{-1} as part of the sensitivity analysis (**Table 4.2**). A pressure outlet boundary condition was applied at the downstream boundary to permit unrestricted outflow and minimise the likelihood of reverse flow. All solid walls were modelled as stationary no-slip boundaries with adiabatic thermal conditions.

The outlet was set as a pressure outlet to avoid any reversed backflow.

The walls are stationary walls with no-slip shear condition and zero heat flux set as a thermal condition.

4.2.3 Velocity sensitivity study

Following the mesh independence study, a sensitivity analysis was performed to investigate the influence of inlet air velocity on the airflow and thermal behaviour of the simplified server model. The inlet velocity was varied between 1 and 4 m s^{-1} while all remaining boundary conditions were maintained constant.

NAME	P10	P5	P6	P7	P8	P9
A)	1	1.0096017	0.24488835	300.63943	318.50483	297.84435
B)	1.5	1.5133475	0.47487086	298.53685	305.66554	296.62558
C)	2	2.0160764	0.74552697	297.5072	300.99163	296.05117
D)	2.5	2.5175166	1.0412738	296.88292	298.81221	295.66976
E)	3	3.019306	1.3600674	296.4654	297.59739	295.40452
F)	3.5	3.5209192	1.6953758	296.16598	296.84731	295.21163
G)	4	4.0198082	2.0457484	295.94382	296.34403	295.06743

Where:

P10 - velocityinlet [m s^-1]
P5 - awa-vel-outlet [m s^-1]
P6 - awa-vel-interiorpb [m s^-1]
P7 - awa-temp-outlet [K]
P8 - awa-temp-interiorpb [K]
P9 - awa-temp-interiorduct [K]

Table 4.2 Results of the variation of velocity in the inlet and the effect that produces in the parts of interest.

Increasing the inlet velocity enhanced the convective heat transfer within the porous medium, resulting in progressively lower outlet and porous-medium temperatures. The results indicate that increasing the airflow rate improves cooling performance by increasing the removal of heat from the server. These observations provided confidence that the simplified model reproduced the expected thermal response before applying the methodology to the larger containerised data-centre simulations.

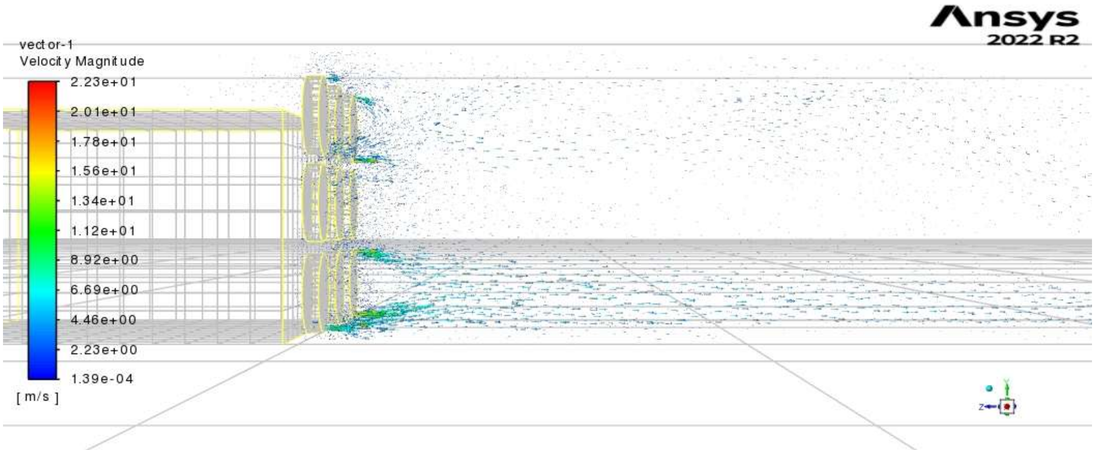

Figure 4.5 Plot of vectors that show the fans activated to draw air from the porous block

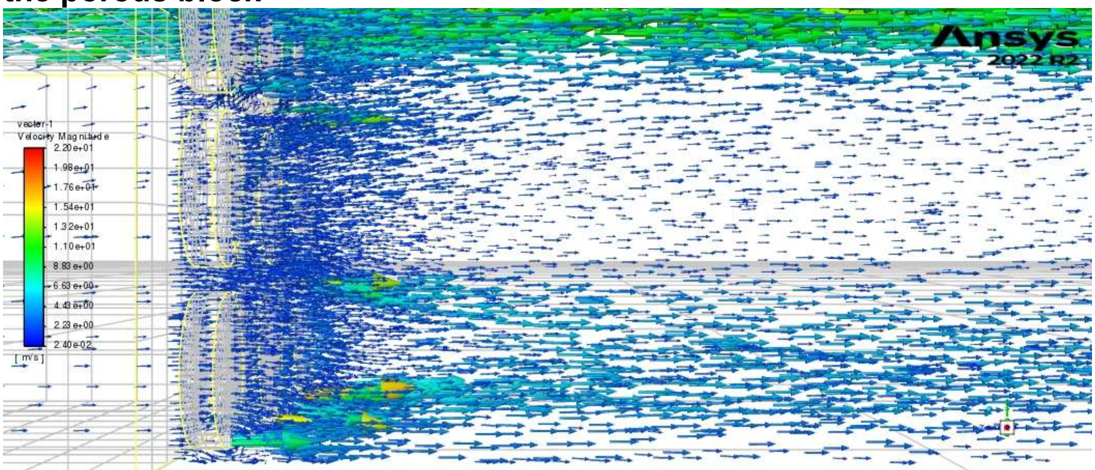

Figure 4.6 Plot of vectors that show the fans and the inlet activated to draw air from the porous block

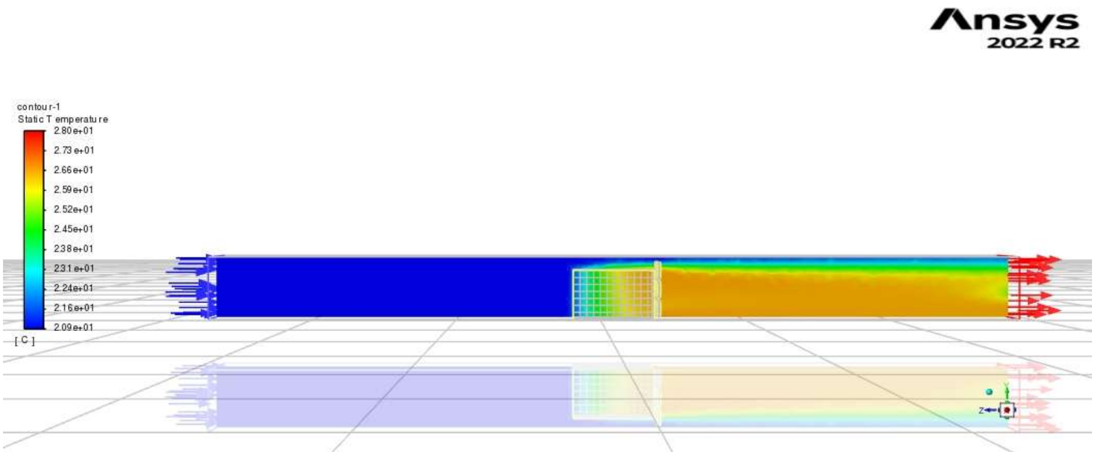

Figure 4.7 Contour plot that shows the temperatures reached by the porous block representing the heat generation of a server and how the heat is drawn by the fans towards the outlet

Figures 4.5 and 4.6 illustrate the airflow generated by the server fans and the interaction between the inlet flow and the porous medium. The velocity vectors demonstrate that air is drawn uniformly through the porous server before being discharged towards the outlet. Figure 4.7 shows the corresponding temperature distribution, confirming that heat generated within the porous medium is transported downstream by the forced airflow. The predicted flow and temperature fields are consistent with the expected behaviour of an air-cooled server.

4.3 Numerical Methodology

The CFD model was implemented in ANSYS Fluent, in which the computational mesh, governing equations, turbulence model and boundary conditions were combined to predict the airflow and heat transfer within the containerised data centre. The numerical methodology was developed to evaluate the influence of workload distribution on hotspot formation and cooling performance under steady-state operating conditions.

4.3.1 Governing Equations and Turbulence Modelling

The airflow within the containerised data centre is characterised by forced convection generated by the CRAH unit and server fans. Owing to the Reynolds numbers associated with the inlet jets and the confined airflow passages between the server racks, the flow is fully turbulent. Consequently, the governing equations are formulated using the Reynolds-Averaged Navier–Stokes (RANS) equations, in which the instantaneous velocity field is decomposed into mean and fluctuating components.

Reynolds averaging introduces additional Reynolds stress terms into the momentum equations, requiring a turbulence closure model. In this work, turbulence closure was achieved using the Renormalisation Group (RNG) $k-\varepsilon$ model. This model has previously been shown to provide good agreement for indoor airflow and data-centre cooling applications while maintaining reasonable computational cost (Almoli, 2013). Compared with the standard $k-\varepsilon$ model, the RNG formulation provides improved predictions of recirculating flows, streamline curvature and rapidly strained regions, making it well suited to the airflow patterns generated by the CRAH unit and server racks.

The governing equations therefore comprise the RANS continuity, momentum and energy equations together with the transport equations for turbulent kinetic energy (k) and turbulent dissipation rate (ε). Heat transfer was modelled by solving the energy equation simultaneously with the momentum equations, allowing conjugate heat transfer between the porous server representation and the surrounding airflow to be predicted.

4.3.2 Boundary conditions

Thermal boundary conditions were specified to reproduce the operating conditions reported by Wang et al. (2017). Heat generated by the servers was represented within the porous medium, while the inlet air temperature and velocity were prescribed according to the experimental configuration. Air humidity was neglected because its influence on sensible heat transfer is small under the operating conditions considered.

Buoyancy effects were neglected because airflow within the containerised data centre is dominated by forced convection generated by the CRAH unit and server fans. In mechanically ventilated enclosures, the relative importance of natural and forced convection is commonly assessed using dimensionless parameters such as the Reynolds and Richardson numbers. Under the operating conditions considered in the present study, the relatively high inlet air velocities indicate that inertial forces dominate buoyancy forces, such that forced convection governs the airflow distribution. Consequently, the contribution of natural convection to the overall thermal behaviour is expected to be small. Similar assumptions have been adopted in previous CFD investigations of mechanically ventilated data centres (Almoli, 2013; Wang et al., 2017) and are consistent with ASHRAE guidance for cooling aisle analyses. The Richardson number (Ri), which expresses the ratio of buoyancy to inertial forces, is expected to be significantly less than unity under the present operating conditions because of the relatively high forced-air velocities generated by the CRAH unit and server fans. This indicates that forced convection dominates over natural convection, supporting the neglect of buoyancy effects. Heat generated by the servers was represented using a uniform volumetric heat generation source applied within the porous medium. The principal numerical settings and boundary conditions adopted in the CFD model are summarised in Table 4.3. The present study employed steady-state simulations to evaluate the long-term thermal behaviour of the containerised data centre under constant operating conditions and prescribed workload distributions. Transient effects associated with short-term workload fluctuations were therefore outside the scope of the present investigation and are identified as an area for future work in Chapter 7.

Viscosity Model RNG k-epsilon (2 eqn)						
Model Constants		Materials		Boundary Conditions		
Cmu	0.0845	Fluid	Air	inlet	velocity-inlet	4 m/s
c1-epsilon	1.42	Solid	Aluminium		turbulent intensity	5%
c2-epsilon	1.68	Cell zone conditions			turbulent viscosity ratio	10%
		A-Racks	Solid	outlet	pressure-outlet	0 pa
Near wall treatment					turbulent intensity	5%
Standard wall functions		Fluid part	Air		turbulent viscosity ratio	10%

Table 4.3 Numerical model settings, material properties and boundary conditions used in the CFD simulations.

4.3.3 Numerical Solution Procedure

The numerical simulations were performed using the pressure-based steady-state solver available in ANSYS Fluent. Pressure and velocity were coupled using the Coupled algorithm to improve convergence of the strongly coupled momentum and pressure fields. Spatial gradients were computed using the Least Squares Cell-Based method, while second-order spatial discretisation was employed for pressure, momentum, turbulent kinetic energy and energy equations to minimise numerical diffusion and improve solution accuracy.

The computational domain consisted of approximately 244,344 control volumes, 1,328,668 faces, and 945,930 nodes. The simulations were solved using four parallel partitions on a workstation equipped with a 13th Generation Intel Core i7-13700 processor and 32 GB RAM operating under a 64-bit Windows environment. Parallel partitioning reduced computational time while maintaining numerical consistency throughout the solution.

Steady-state convergence was assessed using the scaled residuals for continuity, momentum, turbulent kinetic energy, turbulent dissipation rate and energy. A convergence criterion of 1×10^{-5} was specified for all solved equations. In addition to residual reduction, engineering quantities including outlet temperature and airflow patterns were monitored to ensure that the solution had reached a physically steady state.

The selected mesh size and steady-state formulation enabled all simulations to be completed within practical computational times while maintaining the accuracy required for the parametric studies presented in Chapter 6.

4.3.4 Porous medium representation of the servers

The servers were represented using porous medium blocks rather than modelling the detailed internal geometry of individual electronic components. This approach considerably reduced the computational cost while preserving the dominant pressure losses and heat transfer characteristics associated with airflow through a server chassis.

The porous medium was implemented in ANSYS Fluent by assigning an isotropic viscous resistance corresponding to an inverse permeability of $5.29 \times 10^6\ m^{-2}$ and a porosity of 0.5. These values were adopted following the methodology presented by Almoli (2013) for representing server blades in CFD simulations. Heat generated by the electronic components was introduced as a volumetric heat source within the porous region. Consequently, heat was transferred by conduction through the porous medium and subsequently removed by forced convection as air flowed through the server. This representation captures the conjugate heat transfer process while avoiding the computational expense of explicitly resolving individual heat sinks, processors, and electronic components.

4.3.5 Momentum source representation

In the experimental configuration described by Wang et al. (2017), airflow through each server is generated by several axial fans mounted at the rear of the server. Explicit modelling of rotating fan blades would substantially increase the computational cost owing to the additional geometric complexity and local mesh refinement required. Therefore, instead of modelling each fan individually, their effect was represented by a momentum source applied within the porous medium region. The momentum source reproduced the pressure rise generated by the server fans, maintaining the correct airflow direction through the server while significantly reducing the computational requirements of the simulations. This approximation preserves the dominant airflow characteristics while making the progressive scaling from a single server to a complete containerised data centre computationally feasible.

4.3.6 Computational cost and scaling strategy

As the computational domain increased from the server-scale model to the containerised data centre, the computational cost increased substantially. The hierarchical modelling strategy adopted in this work enabled mesh development, boundary condition specification and model validation to be completed using simplified geometries before progressively increasing the domain complexity. This significantly reduced the overall computational effort while maintaining confidence in the numerical methodology.

4.3.7 Mesh quality

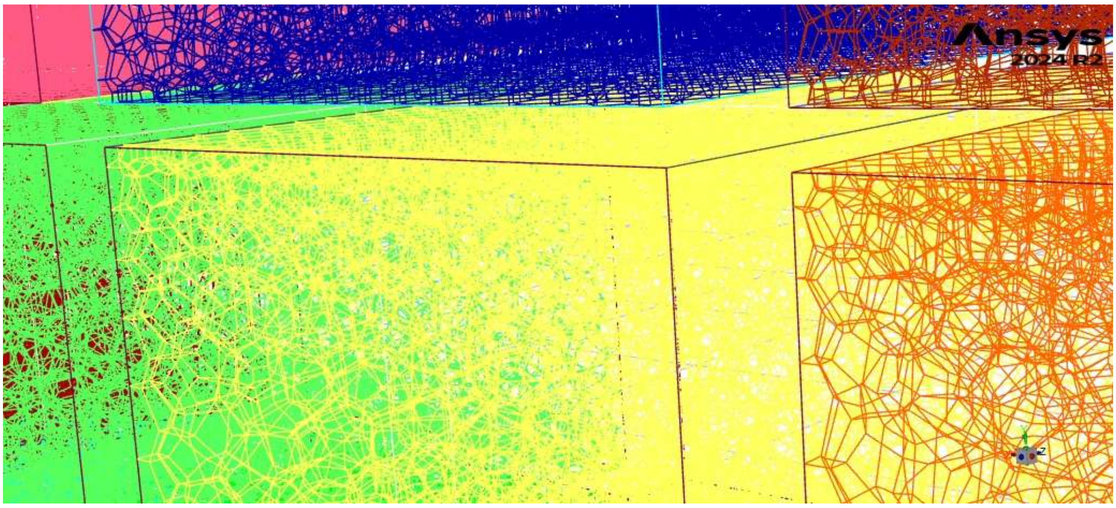

Figure 4.8 Illustrates the view of the mesh in the half-model.

The computational mesh was assessed using the built-in Fluent mesh quality diagnostics prior to solution. A structured mesh was adopted for the wind-tunnel geometry to minimise numerical diffusion and maintain good element orthogonality throughout the domain. Particular attention was given to the fluid-solid interfaces surrounding the porous medium block to ensure that the temperature gradients associated with conjugate heat transfer were adequately resolved. Mesh independence was subsequently confirmed by progressively refining the computational mesh and demonstrating negligible variation in the predicted temperatures and velocities beyond mesh DP1.

4.3.8 Convergence

Each steady-state simulation was considered converged when the scaled residuals of continuity, momentum, turbulence quantities and energy reduced below the prescribed convergence criteria and the monitored outlet temperature and velocity became independent of further iterations. In addition to residual convergence, key engineering quantities including outlet temperature, pressure drop and porous block temperature were monitored to ensure that a physically steady solution had been achieved. A convergence criterion of $\mathbf{1 \times 10^{-5}}$ was specified for all solved equations. In addition to residual reduction, engineering quantities including outlet temperature and airflow patterns were monitored to ensure that the solution had reached a physically steady state.

4.3.9 Computational Set-up

Following the successful validation of the server-scale model, the CFD methodology was extended progressively to represent increasingly realistic data-centre configurations. Additional porous servers were introduced to form complete rack rows before constructing the half-container geometry. At each stage, the boundary conditions and modelling assumptions established during the wind-tunnel validation were retained, ensuring consistency throughout the modelling process.

4.4 Containerised data centre model

Following validation of the server-scale methodology, the same modelling approach was transferred to the reduced containerised data centre.

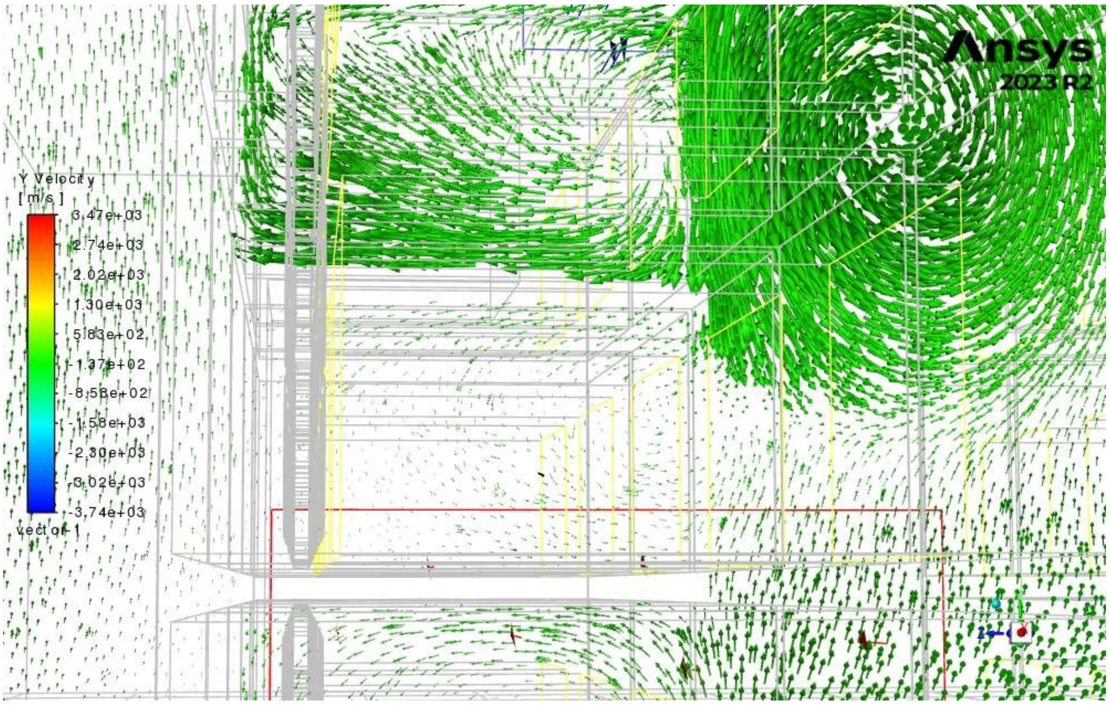

Figure 4.9 Vectors showing the airflow path in the halved containerised data centre simulation

Figure 4.9 Shows that the air now follows the behaviour achieved in the wind tunnel after extrapolating the settings used in the small simulation. The airflow now respects the boundaries the solid rack sets, which is the expected behaviour. The predicted airflow pattern is consistent with the expected flow behaviour within the containerised data centre and demonstrates that the boundary conditions established using the simplified model can be successfully transferred to the larger computational domain.

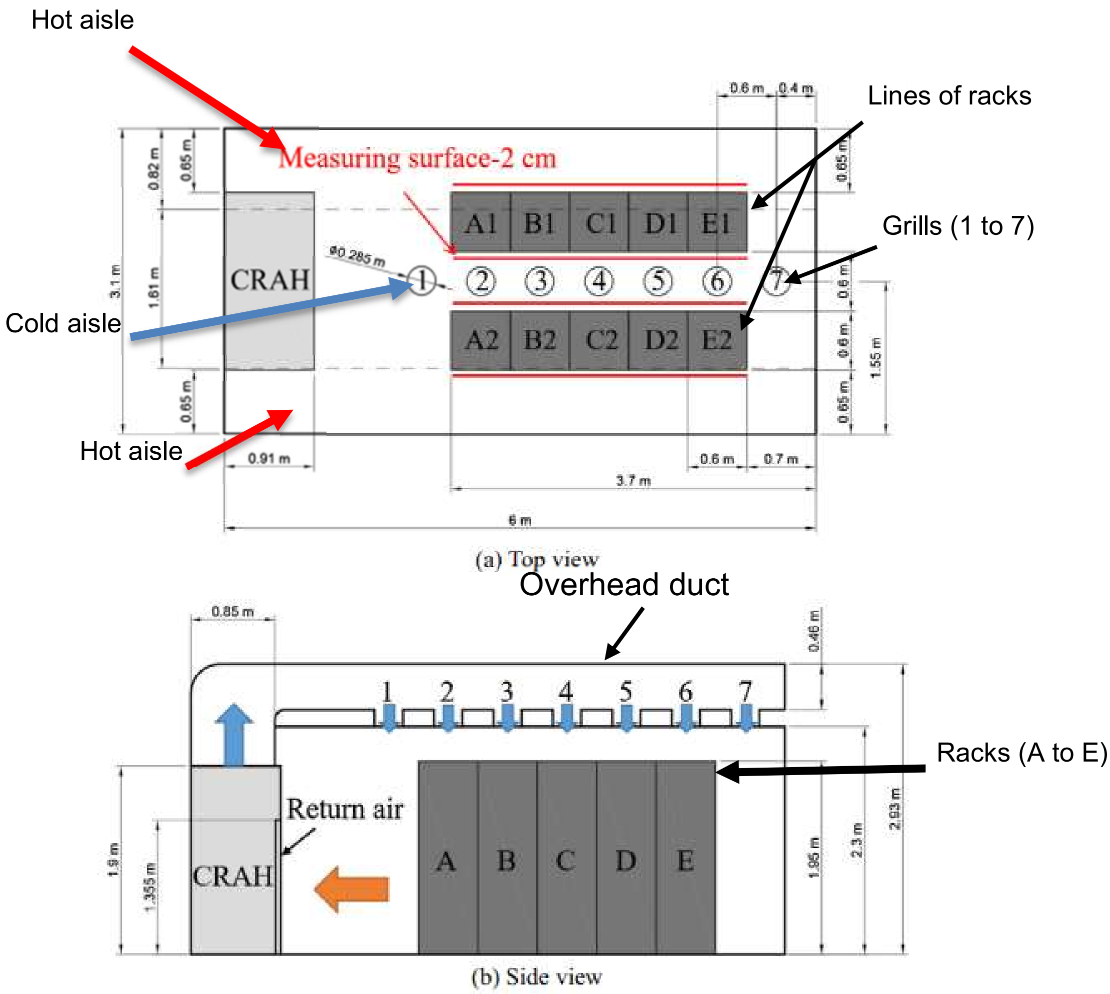

Figure 4.10 Design of a containerised data centre (according to Wang et al., 2017a).

Following validation of the server-scale model, the computational domain was progressively expanded to represent the halved containerised data centre (see Figure 4.12). This required several modelling simplifications to create an approximation of the experimental setup as described in Figure 4.11.

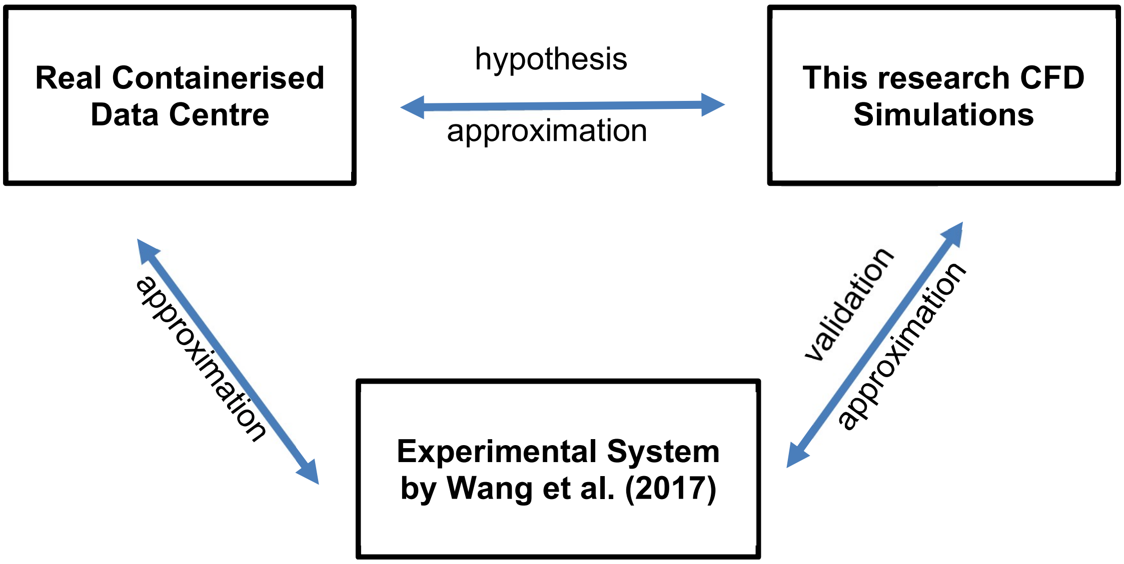

Figure 4.11 Diagram that describes how both experimental and theoretical implementations intend to approximate models in real-life to achieve more closely real-world results.

Figure 4.13 displays the number assigned to each block in the geometry to simulate servers and assign heat generation to simulate the workload in a containerised data centre. Figure 4.14 displays the geometry with the CRAH outlet side being at the front, and the sides of porous blocks representing the servers. Figure 4.15 shows the approximation to the full containerised data centre by mirroring the model of the geometry of the halved containerised data centre.

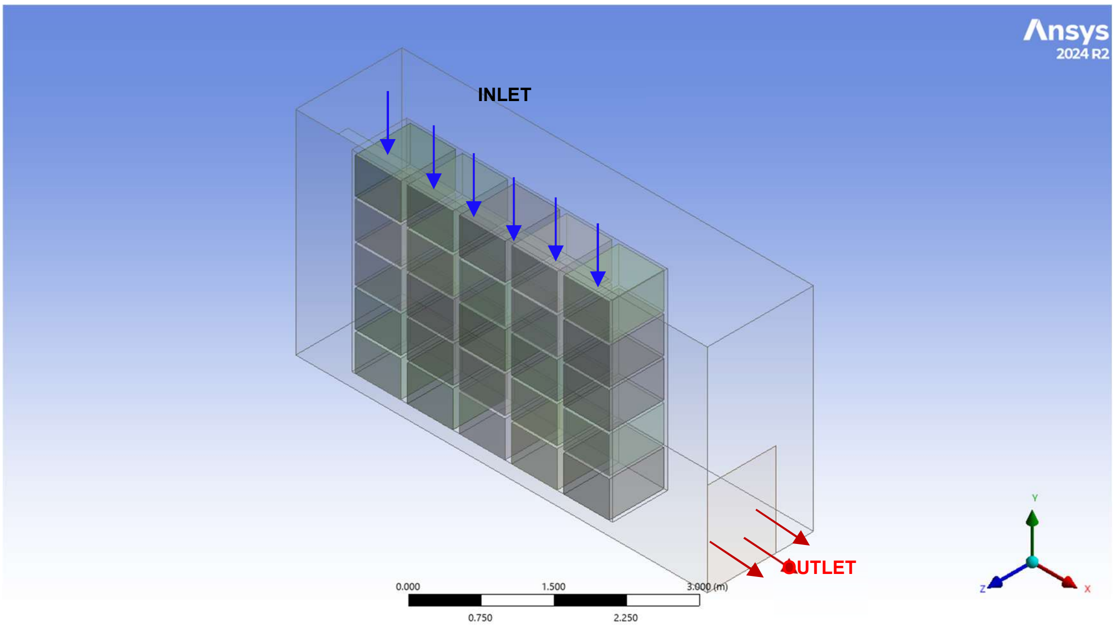

Figure 4.12 Geometry of the halved containerised data centre showing the inlet and outlet's system.

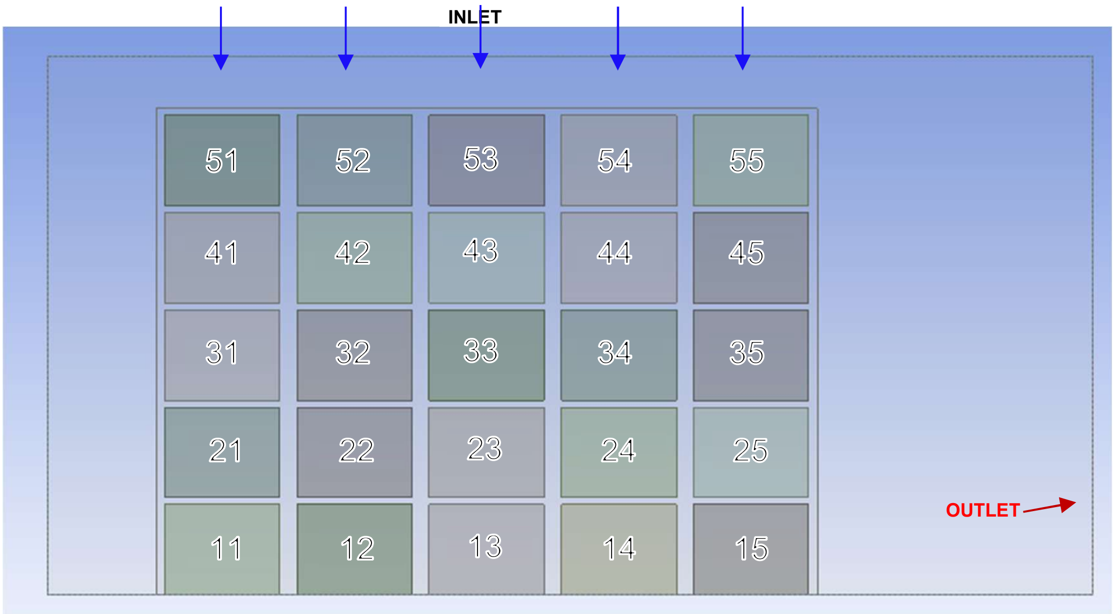

Figure 4.13 Number distribution of the blocks representing the servers inside the containerised data centre.

The boundary conditions for this problem are described in Table 4.3.

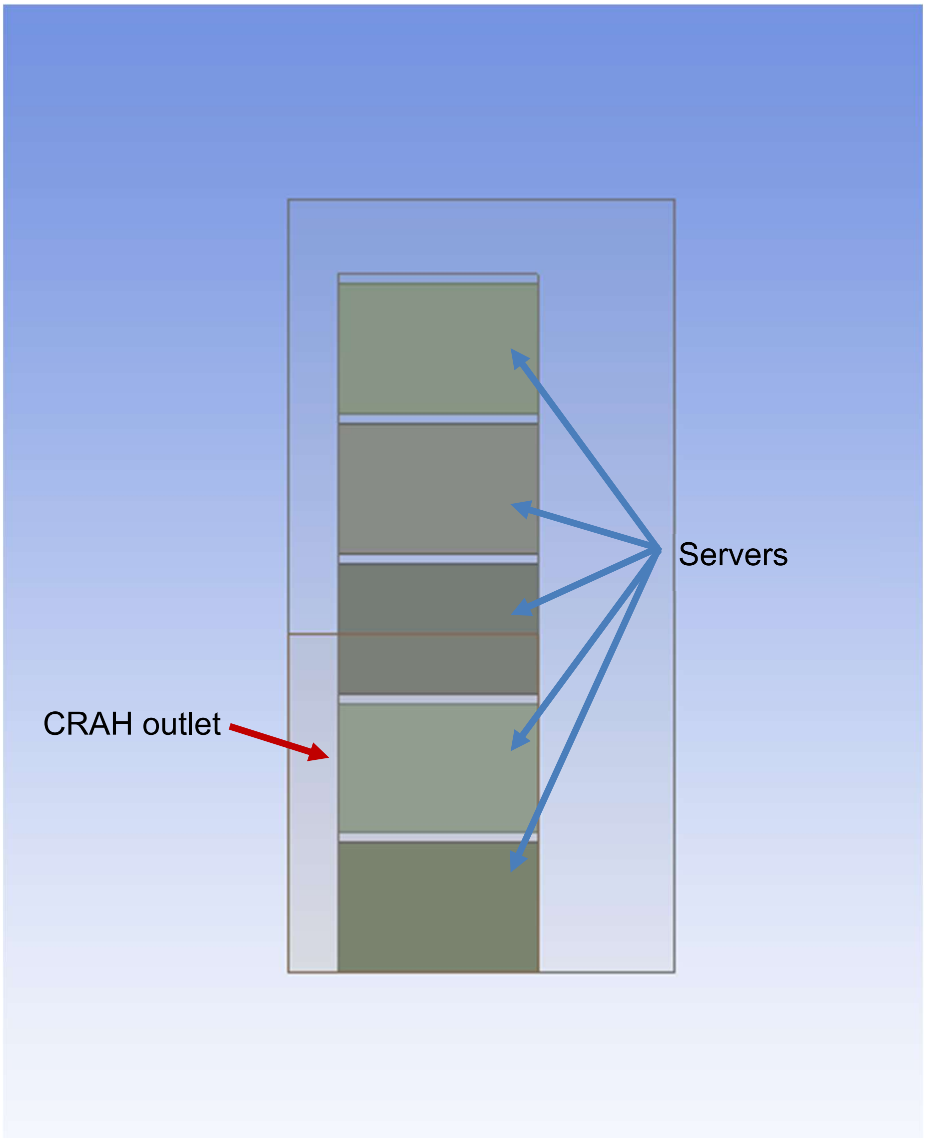

Figure 4.14 Sideview of the halved containerised data centre with the outlet in the front side of the geometry.

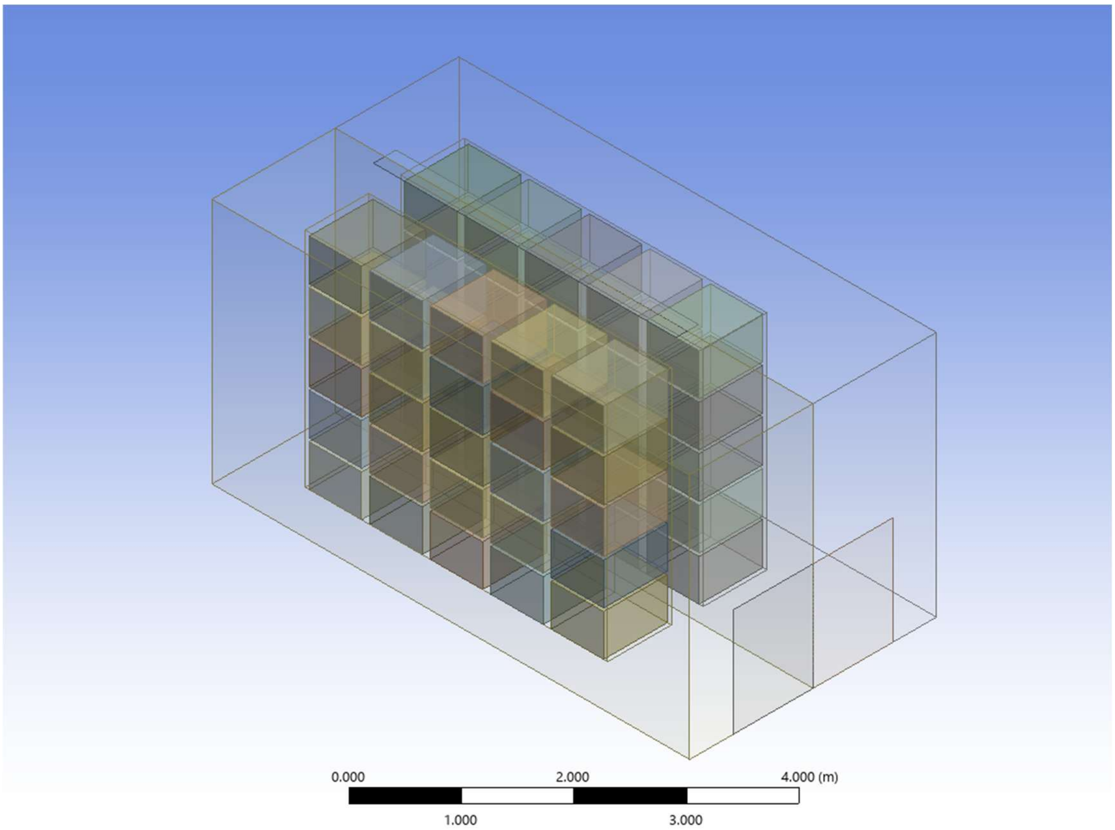

Figure 4.15 Mirrored view of the geometry of the containerised data centre displaying an approximation of the one used by Wang et al. (2017).

The computational domain comprised five complete racks, representing approximately half of the experimental containerised data centre. This reduction was introduced solely to limit the computational requirements while preserving the dominant airflow paths, thermal interactions and hotspot formation within the central region of the data centre.

4.5 Model Validation

The CFD model was validated by comparing the predicted airflow and temperature distributions with the experimental measurements reported by Wang et al. (2017).

Although only half of the containerised data centre was modelled, this was not implemented using a symmetry boundary condition. Instead, the computational domain represents a physically reduced model comprising five complete racks. The illustration of the full containerised data centre (Figure 4.15) is included solely to demonstrate the relationship between the computational model and the experimental arrangement reported by Wang et al. (2017), rather than to indicate the application of geometric symmetry within the CFD model.

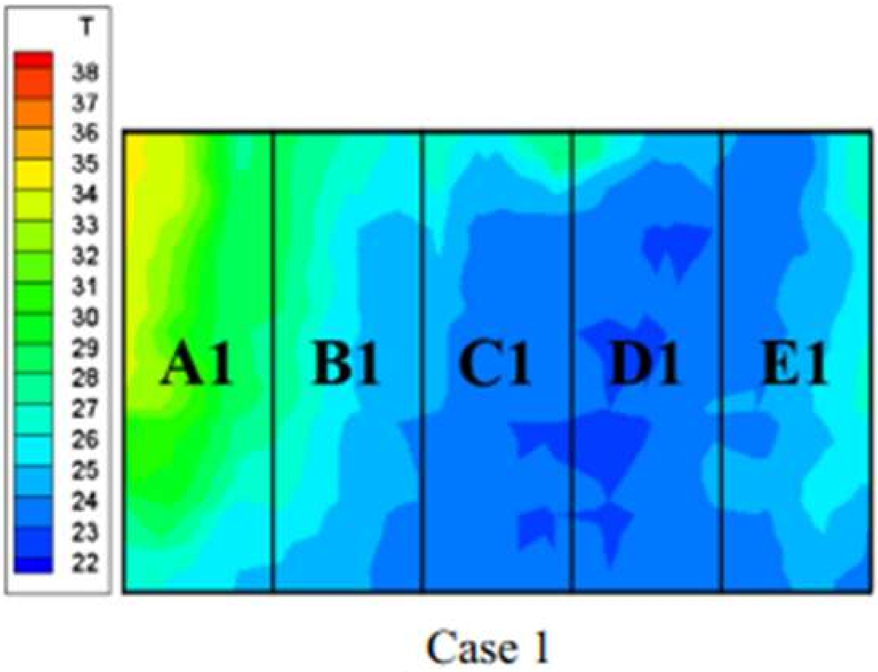

Figure 4.16 Results by Wang et al. (2017), where they are exploring the scenario 1 pictured in Figure 4.17.

Figure 4.16 shows the results of the published experiment that was conducted with different configurations in a test container data centre where hot air recirculation prevails mostly at the entrance of the cold aisle. The authors of the papers (Wang et al., 2017; Wang et al., 2017a) tried different layouts (configurations) as well as enclosing the end of the cold aisle. The latter performed the worst in cooling the data centre. Further full containment of the cold aisle was also tried with high jet air flowrate from the grilles which showed even worse performance. The best scenario was when the grills outside the containment area comprising the cold aisle were blocked, and the flow rate was reduced from 3.52 to 2.83 m^3 s^{-1}.

The model presented by Wang et al. (2017) has a duct on top. Considering the heat radiation received by the sun, one limitation of the experimental configuration is having a duct on top, which is not ideal and could affect some of the findings discussed above; this would be a worthy consideration for future research: trying other configurations for the duct or the way that the air flows inside the container. Using the configurations described above to preserve the fidelity of the simulated environment led to complications, thereby requiring extensive computational resources. These were not available to simulate the setting, resulting in computational constraints.

Figure 4.17 indicates that the numerical simulations, considering all the restrictions described by Wang et al. (2017), have an agreement within about ±4 °C in the maximum temperature reached under the same conditions outlined in the study of Wang et al. (2017), compared to the results shown in Figure 4.16. This agreement is significant because their experiments were conducted in real-world facilities, whereas the methodology used in this research is a numerical representation. For example, in Figure 4.17, according to Wang et al. (2017), Rack A1 is positioned towards the outlet. Additionally, the observed difference in hot spot generation follows a different pattern, considering that in Figure 5.2, there are nine fans drawing air towards the hot aisle. However, to reduce the computational resources needed to simulate the environment, given the computational limitations described in section 4.5, a simplification was implemented by removing the nine fans at the back of each radiator. These were replaced with a momentum generated inside the porous block, which consequently directed hot air in all directions of the porous block. Hence, the variation in heat patterns and the maximum temperature reached in the same measurement area.

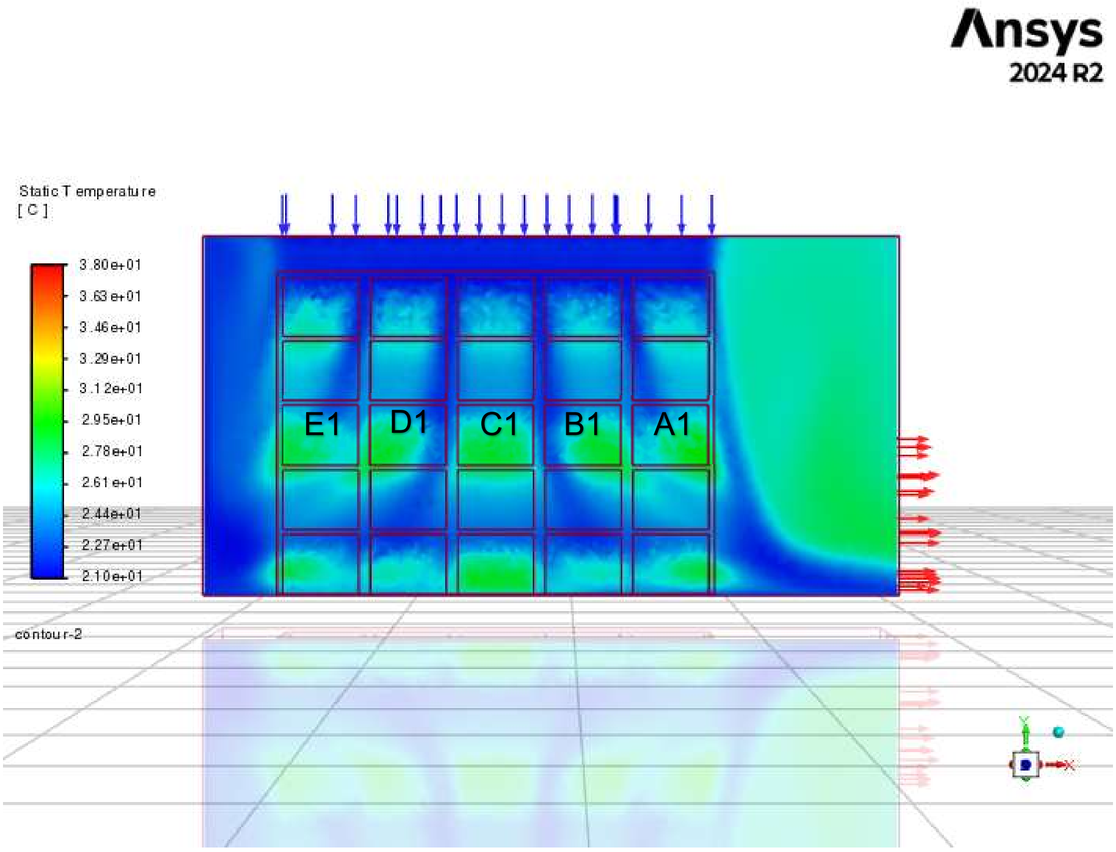

Figure 4.17 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre.

Figure 4.18 shows the hot aisle, which is the back part of Figure 4.17. This shows that most of the hotspots are being identified at the bottom line of the racks. This behaviour is due to the inlet cold air Jetstream driving the heat towards the bottom to be naturally driven towards the outlet.

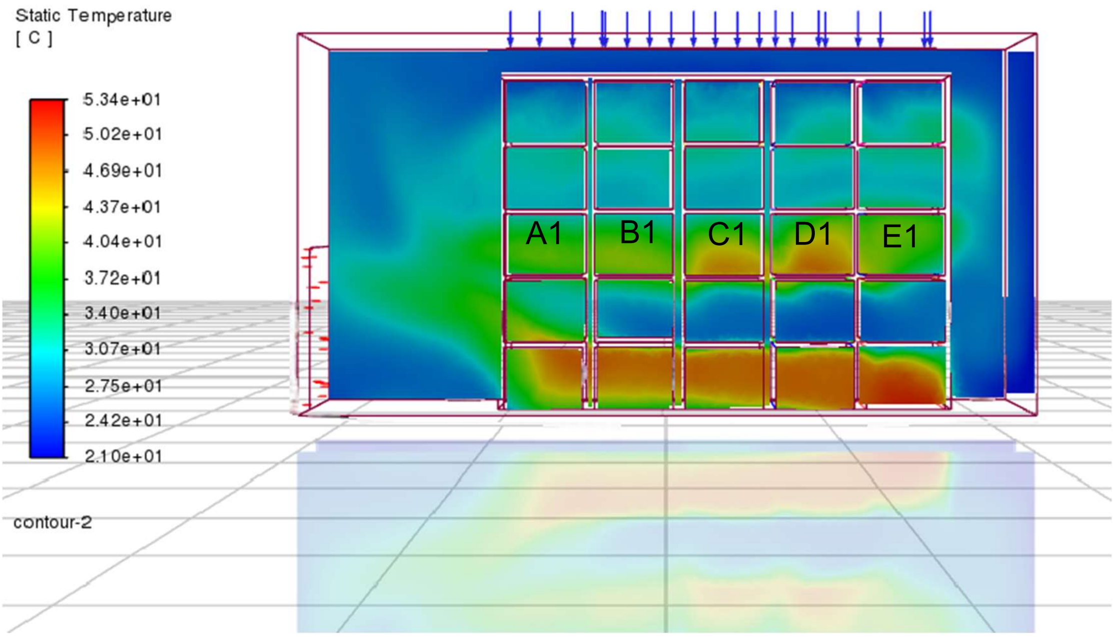

Figure 4.18 Contours of static temperature obtained from the measurement plane located in the hot aisle of the containerised data centre.

4.6 Summary

This chapter has presented the complete computational methodology adopted for the numerical investigation of the containerised data centre. A hierarchical modelling strategy was developed, beginning with a simplified server-scale validation case before progressively extending the computational domain to the containerised data centre. The numerical methodology, turbulence modelling, boundary conditions and solution procedures were described, and the CFD predictions were validated against the experimental measurements of Wang et al. (2017). Although several modelling simplifications were introduced to reduce computational cost, the numerical model reproduced the principal airflow and thermal characteristics of the experimental system with acceptable agreement, providing confidence for the parametric investigations presented in Chapter 6.

Chapter 5 Analysis of workload distribution effects on hotspots.

5.1 Workloads

The chapter describes the simulation of a number of data centre scenarios in relation to different possible workloads and their distributions within data centres. The workloads in this thesis are common patterns within data centres explored by various authors (e.g. Yan et al. (2018), Liu et al. (2012), Li et al. (2019)). The chapter will provide the rationale for the five chosen workload scenarios used to run the data centre simulations, along with the server temperature distributions achieved by different strategies under constant-flow cooling air supply:

- Scenario 1 provides a case study where the workload is 60% of the total capacity of the data centre, where the load is evenly distributed between four rows, five racks and 20 different servers.
- Scenario 2 outlines a case study of 50% of the total capacity of the data centre distributed between the top and bottom rows of five racks among 25 different servers.
- Scenario 3 portrays a case study where the workload is only at 30% of the total capacity of the data centre and is strategically and evenly distributed in five racks among 11 different servers.
- Scenario 4 describes a case study where the workload is at 70% of the total capacity and is distributed between five racks and 25 different servers.
- Scenario 5 shows a case study where the workload is at 90% of the total capacity in the data centre and is distributed between five racks and 25 different servers.

Scenario	Workload	Distribution
1	60%	evenly between four rows, five racks and 20 different servers.
2	50%	distributed between the top and bottom rows of five racks among 25 different servers.
3	30%	strategically and evenly distributed in five racks among 11 different servers.
4	70%	distributed between five racks and 25 different servers.
5	90%	distributed between five racks and 25 different servers.

Table 5.1 Different workload scenarios studied and their distribution in the data centre.

5.2 Limitations and Assumptions

There were computational limitations regarding the resources needed to represent the different scenarios, so modifications had to be made. First, instead of the full containerised data centre, only half of it was modelled. Instead of including nine fans at the back of each server, only the momentum estimated to be generated by these was used. However, instead of rejecting the heat towards the hot aisle, it was directed towards all sides of the porous medium. Also, instead of leaving space between the rack and the porous medium in this research, the size of the porous medium was increased to fit exactly within the space to simulate a more accurate setting, as is the case in racks and data centres. Scenarios 2, 4 and 5 are included to define the complete experimental framework but were excluded from detailed CFD analysis because of computational resource limitations in this work; to explore all the other scenarios, more computational resources and a full containerised data centre model should be considered.

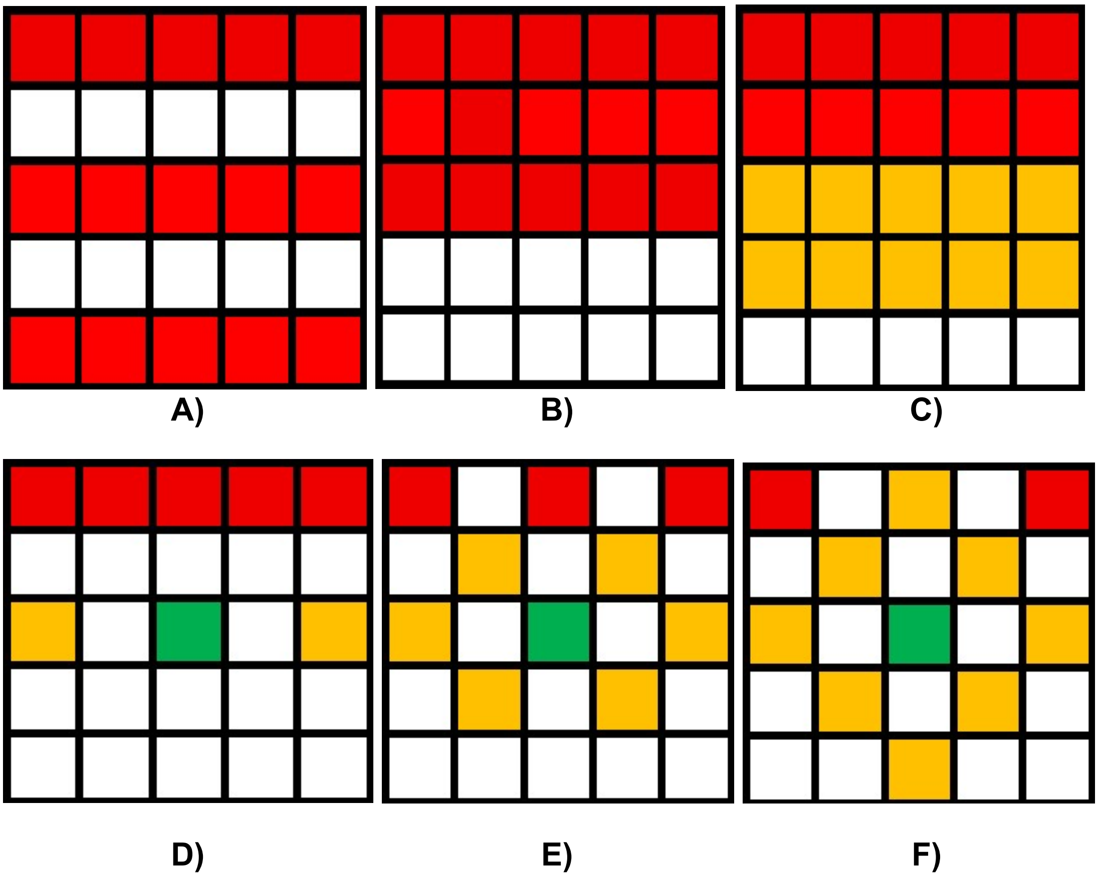

Figure 5.1 Representation of Workload Distribution utilising a traffic light system to depict the two different workloads, where Workload A, B and C are intended to distribute 60% corresponding to scenario 1, and D, E and F are meant to allocate 30% as per scenario 3.

Figure 5.1 is a grid representing a rack for scenario 1 described above. The different colours indicate the workload each server handles in each scenario. A traffic light system is used to depict these workloads: red for full usage 70-100%, yellow for 26-69% usage, and green for 1-25%. Servers not coloured are turned off in the respective scenario. Each letter represents a different workload distribution.

5.3 Scenario 1

The purpose of this series of simulations is to reduce the hot spots in areas in front of the rack, to avoid unnecessary triggering of the air conditioner, and in consequence reduce the energy consumption of the cooling system. It is expected that the temperature should be kept at 27 ºC as recommended by the ASHRAE, therefore, all the spots in front of the rack (cold aisle) that exceed this temperature are considered to possibly trigger the air conditioner. However, the ASHRAE maximum allowable temperature is 32 ºC.

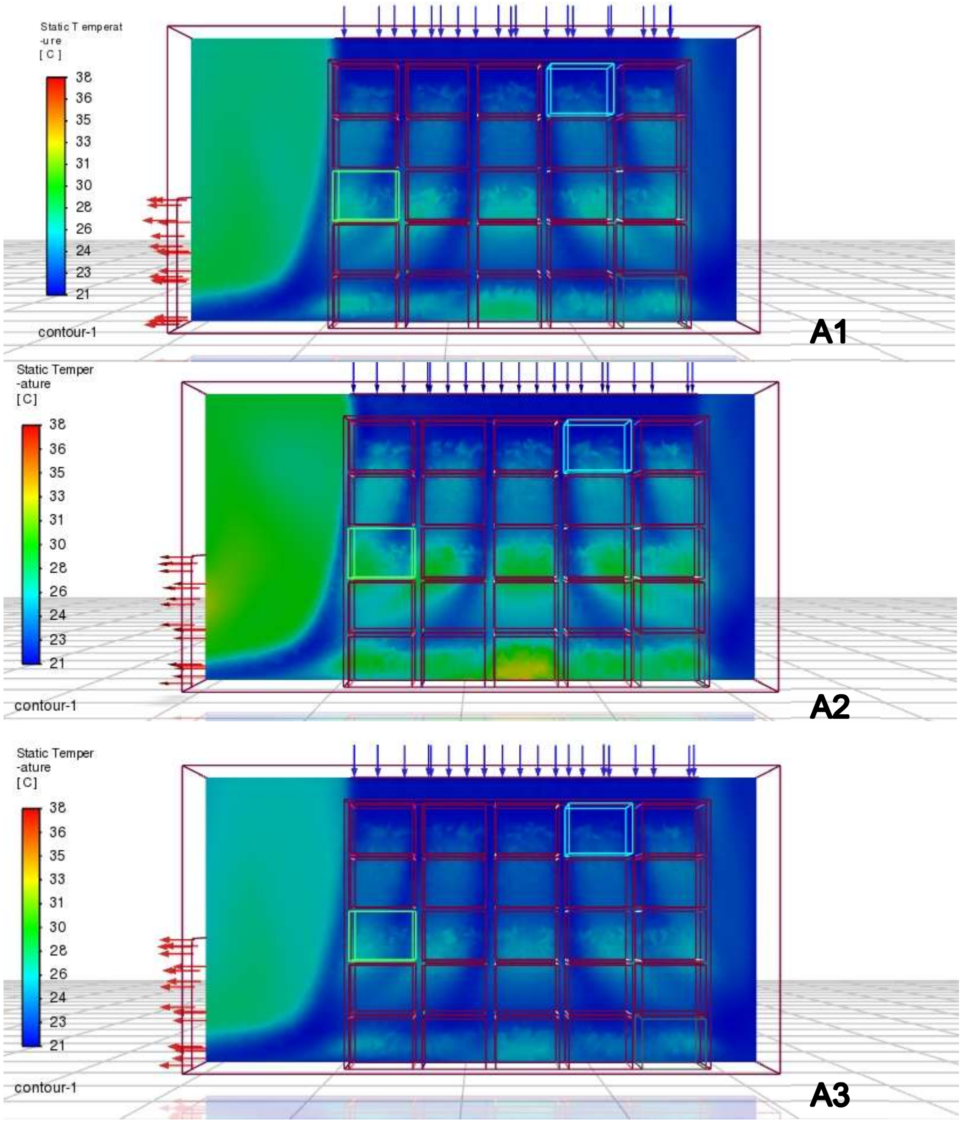

Figure 5.2 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution A, the conditions listed as Cases 1, 2, 3 in Table 5.2 and the workload allocation as shown in Figure 5.1 A).

Figure 5.2 is a static temperature contour plotted for three different cases following Workload Distribution A of scenario 1 (A in Figure 5.1). The simulation results are described in Table 5.2 where the measurement plane set in the cold aisle can reach a maximum of 35 ºC in the server area and a maximum of 38 ºC in the same plane towards the outlet area. It was hypothesised that the temperature would be equilibrated by distributing the workload in three rows (top, middle and bottom) of the containerised data centre; however, this hypothesis was rejected because the hotspot with the maximum temperature was situated mostly in the bottom row. Additionally, the maximum temperature achieved within the servers rose to 65 ºC in the hot aisle suggesting that some of the chips may have gone above this temperature, risking the integrity of the electronic circuits and triggering an increased usage of the CRAH.

	Workload Distribution	Case	Inlet Air Flowrate	Inlet Temp in °C	Avg. Temp in Outlet in °C	Max Temp in Cold Aisle in °C	Max Temp in Hot Aisle in °C	Max Temp Inside Blocks in °C
Scenario 1	A	1	V1	21	27.47	30.59	53.83	51.97
		2	V2	21	29.20	34.77	64.82	62.92
		3	V3	21	26.66	28.92	48.42	48.73
	B	4	V1	21	27.20	34.30	58.00	57.41
		5	V2	21	29.45	39.93	65.60	66.35
		6	V3	21	26.53	32.12	52.66	53.84
	C	7	V1	21	27.21	33.69	52.14	56.04
		8	V3	21	26.52	32.05	49.93	52.22
Scenario 3	D	9	V1	21	23.33	27.59	41.72	46.18
		10	V2	21	23.66	28.51	44.15	48.44
		11	V3	21	24.52	30.36	49.06	54.47
	E	12	V2	21	23.32	27.34	39.00	45.53
		13	V1	21	23.68	28.30	43.86	48.14
		14	V3	21	24.49	30.11	48.81	52.87
	F	15	V2	21	23.34	27.34	38.39	45.52
		16	V1	21	23.71	28.29	39.90	47.85
		17	V3	21	24.47	30.10	43.06	52.30

Table 5.2 Results obtained in seventeen scenarios where the inlet air flowrate for V1: 3.52 m/s, V2: 2.83 m/s, and V3: 4m/s (Detailed results can be found in Appendix A).

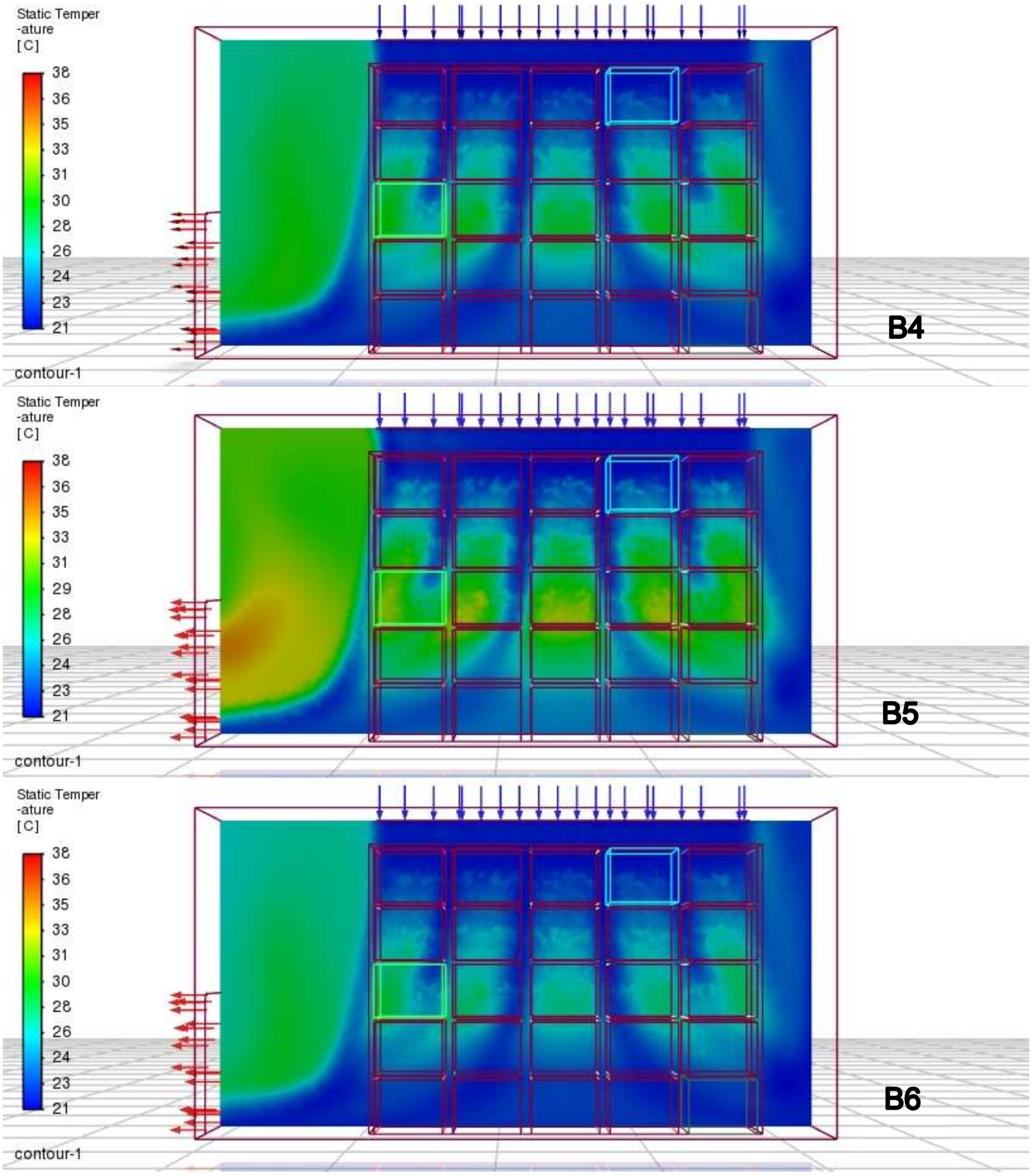

Figure 5.3 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution B, the conditions listed as Cases 4, 5, 6 in Table 5.2 and the workload allocation as shown in Figure 5.1 B).

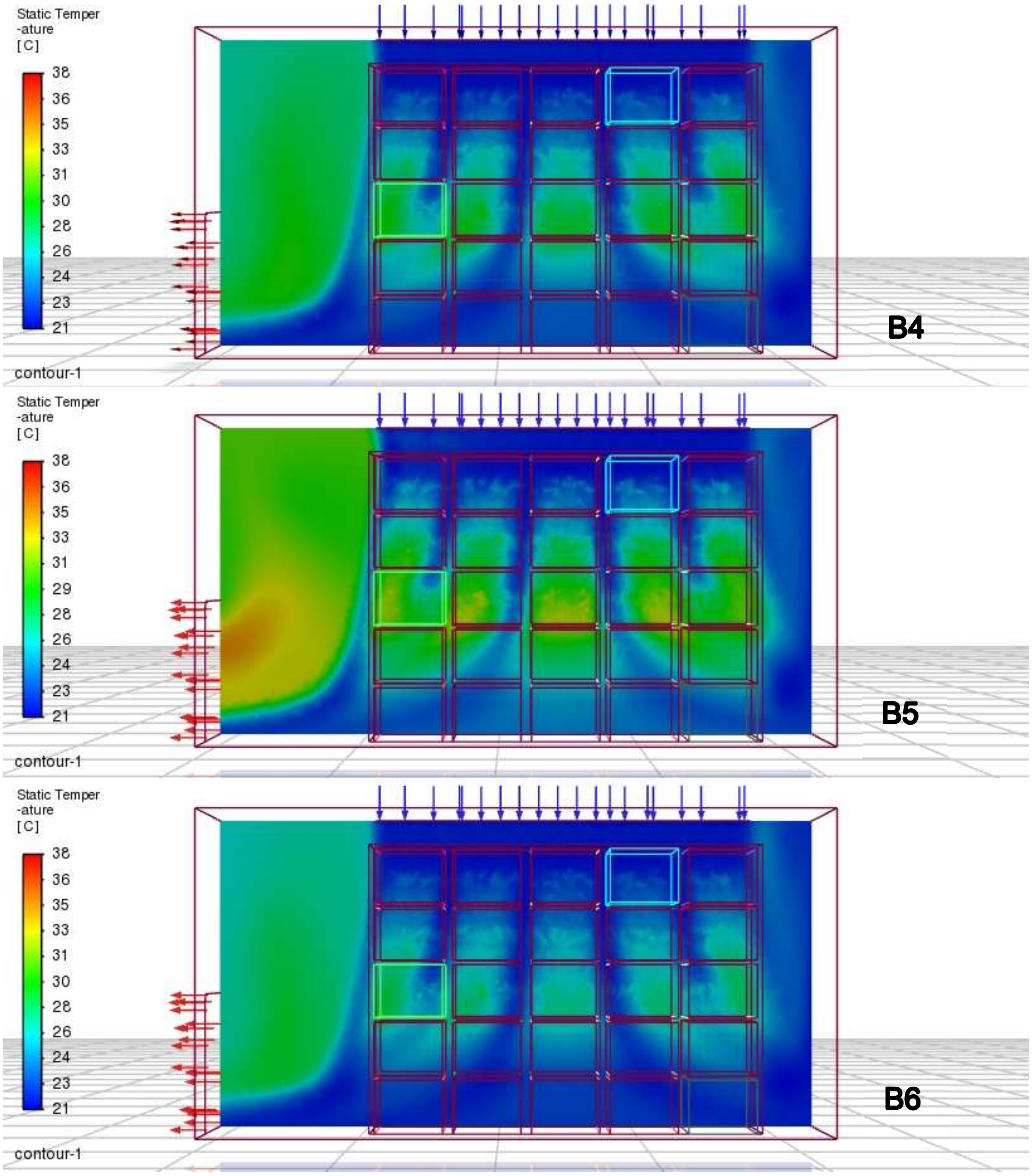

Figure 5.3 is a static temperature contour plotted with the simulations results described in cases B4, B5 and B6 in Table 5.2 for scenario B in Figure 5.1, where the measurement plane in the cold aisle can reach a maximum of 40 ºC in the server area and 38 ºC in the same plane towards the outlet area. It was hypothesised that by moving the workload into the three top rows of the containerised data centre, the temperature would reduce, given that the A Workload Distribution was generating a big hot spot at the bottom; however, this hypothesis is rejected, because it is not better than the original distribution

observed in the cases A1, A2 and A3. The maximum temperature achieved within the servers rose to 66 ºC in the hot aisle area. This temperature suggests that some of the chips may have gone above the threshold because the air in the hot aisle has been mixed with bypassed cold air from the servers, and it is expected that the air temperature in the hot aisle is cooler than some of the chips on the inside of the server, in consequence risking the integrity of the electronic circuits and triggering an increased usage of the CRAH.

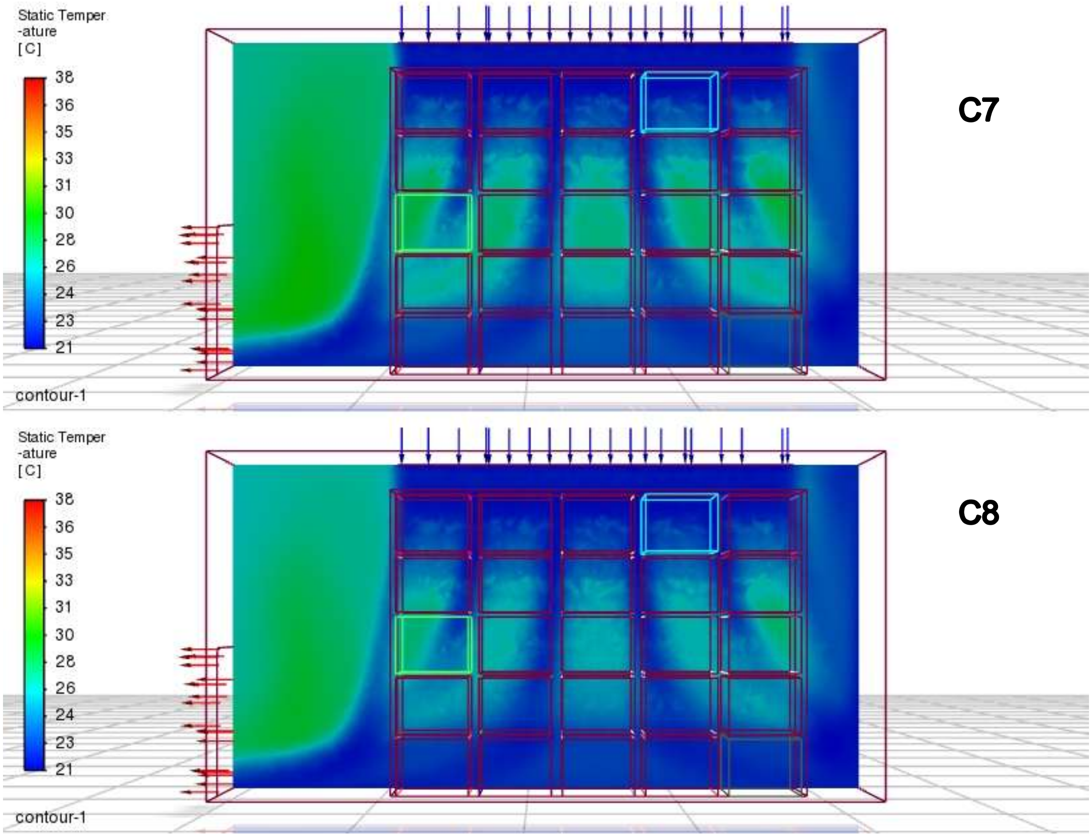

Figure 5.4 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution C, the conditions listed as Cases 7 and 8 in Table 5.2 and the workload allocation as shown in Figure 5.1 C).

Figure 5.4 is a static temperature contour plotted with the simulation results described in Table 5.2 for C7 and C8 cases, where the measurement plane in the cold aisle can reach a maximum of 34 ºC in the server area and 31 ºC in the same plane towards the outlet area. It was hypothesised that the temperature would reduce by distributing the workload from the middle row of the containerised data centre into two penultimate rows (rows 3 and 4) given that the B Workload Distribution was generating a big hot spot in the middle row of the containerised data centre; however, this hypothesis of dividing the workload from row 3 between rows 3 and 4 the cold area on the top area of the data centre is rejected, because it is not better than the original distribution observed in the cases A1, A2 and A3. The maximum temperature achieved within the servers rose to 52 ºC in the hot aisle area; therefore, this Workload Distribution also suggested that some of the chips may have gone above this temperature risking the integrity of the electronic circuits and triggering an increased usage of the CRAH.

5.4 Scenario 3

In the following subsection, three workload distributions are tested to reduce hot spots at the front of the rack, aiming to prevent unnecessary triggering of the air conditioning system and consequently decrease energy consumption. It is recommended that the temperature should be maintained at 27 °C as advised by ASHRAE; therefore, all spots in front of the rack (cold aisle) exceeding this temperature are considered potential triggers for the air conditioning. However, the maximum allowable temperature according to ASHRAE can be up to 32 °C. For this scenario, it is hypothesised that the maximum temperature in the cold aisle should be kept at 27 °C.

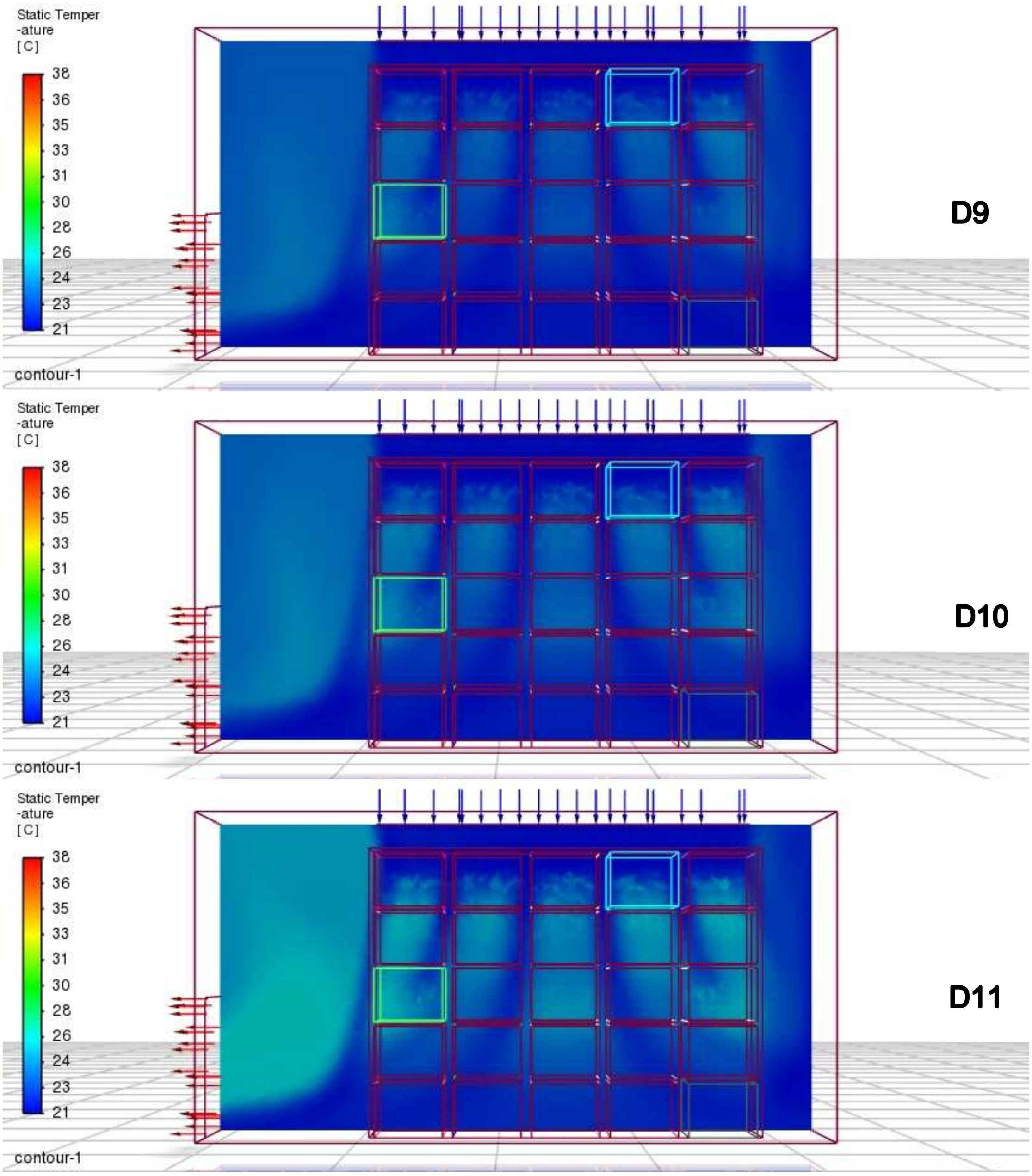

Figure 5.5 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution D, the conditions listed as Cases 9, 10 and 11 in Figure 5.1 D).

Figure 5.5 is a static temperature contour plotted using the simulation results described in Table 5.2 for the D9, D10, and D11 cases. The measurement plane in the cold aisle can reach a maximum of 30 ºC in the server area and 32 ºC in the same plane towards the outlet area. It was hypothesised that the temperature would be reduced by distributing the full workload servers to the top row of the containerised data centre. The workload from the middle row was divided between blocks 31, 33, and 35, given that the C Workload Distribution was generating hot spots above the maximum allowable temperature of the containerised data centre. This hypothesis of dividing the workload using Workload Distribution D proves to be better than A, B and C Workload Distributions because it maintains the cold aisle server area within the maximum allowable temperature; however, the maximum temperature in the cold aisle area of the data centre is still above the maximum recommended; therefore, this hypothesis is rejected. The maximum temperature achieved within the servers rose to 49 ºC in the hot aisle area. Thus, this Workload Distribution also suggested that some of the chips may have gone above this temperature, risking the integrity of the electronic circuits and triggering an increased usage of the CRAH. Because the hypothesis was rejected, a further set of simulation workloads from blocks 52 and 54 was removed and redistributed among blocks 42, 44, 22, and 24. A graphic representation of this Workload Distribution is shown in Figure 5.1.

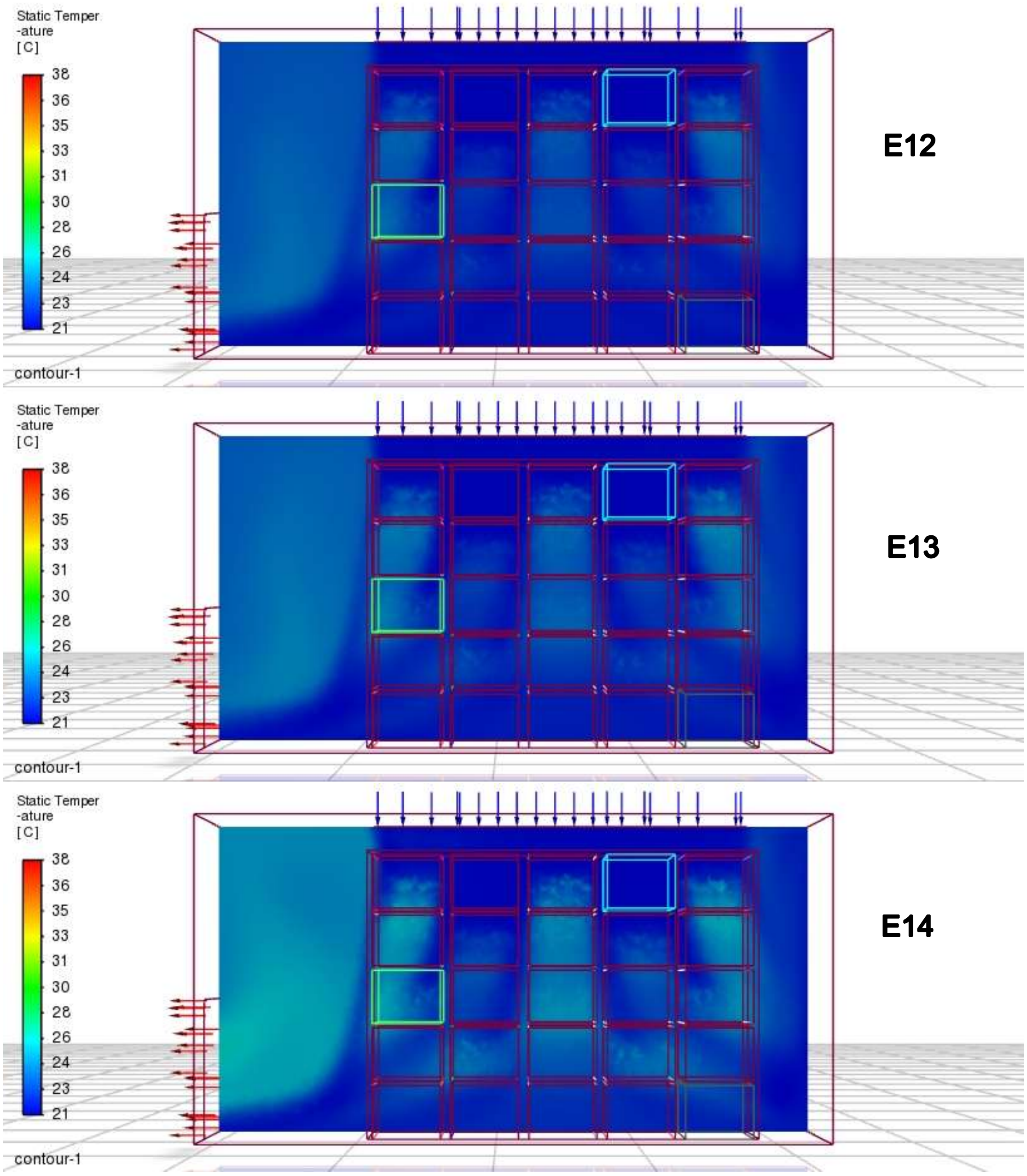

Figure 5.6 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution E, the conditions listed as Cases 12, 13 and 14 in Figure 5.1 E).

Figure 5.6 is a static temperature contour plotted using the simulation results described in Table 5.2 for the E12, E13 and E14 cases. The measurement plane in the cold aisle can reach a maximum of 28 ºC in the server area and 30 ºC in the same plane towards the outlet area. It was hypothesised that the temperature would be reduced by distributing the full workload servers (52, 54) from the top row of the containerised data centre. The workload from the top is divided between four servers (42, 44, 22, and 24). The middle row was divided between blocks 31, 33, and 35, given that the D Workload Distribution was generating hot spots above the maximum recommended temperature of the containerised data centre. This hypothesis of dividing the workload using Workload Distribution E proves to be better than A, B, C and D Workload Distributions because it maintains the cold aisle server area within the maximum allowable temperature; however, the maximum temperature in the cold aisle area of the data centre is still above that recommended; therefore, this hypothesis is rejected. The maximum temperature achieved within the servers rose to 48 ºC in the hot aisle area. Thus, this Workload Distribution also suggested that some of the chips may have gone above this temperature, risking the integrity of the electronic circuits and triggering an increased usage of the CRAH. Because the hypothesis was rejected, a further set of simulation workload from block 53 was divided, and half of the workload was relocated to block 13. A graphic representation of this Workload Distribution is shown in Figure 5.1.

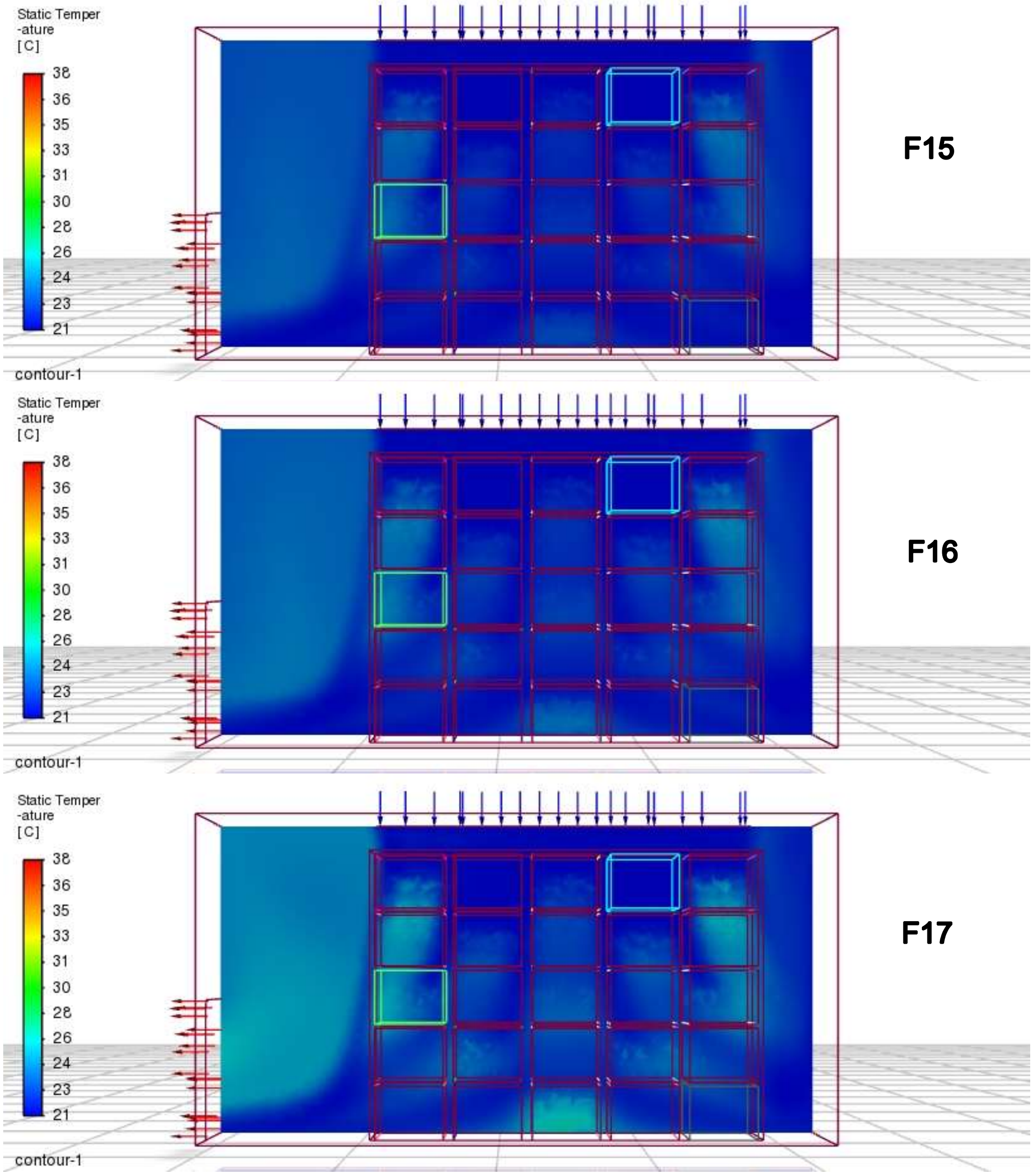

Figure 5.7 Contours of static temperature obtained from the measurement plane located in the cold aisle of the containerised data centre, the results plotted in this figure were obtained utilising Workload Distribution F, the conditions listed as Cases 15, 16 and 17.

Figure 5.7 shows a static temperature contour based on the simulation results described in Table 5.2 for the cases F15, F16, and F17. The measurement plane in the cold aisle can reach a maximum of 30 ºC. It was hypothesised that the temperature could be reduced by distributing workload servers according to the F Workload Distribution, as in case F15, within the maximum recommended temperature of 27 ºC for the containerised data centre. This hypothesis dividing the workload using Workload Distribution F proves to be better than Distributions A, B, C, D, and E because it keeps the cold aisle server area within the recommended temperature limit; therefore, it is accepted. The maximum temperature within the servers increased to 38 ºC in the hot aisle area. Thus, this Workload Distribution also suggests that the chips remain within the recommended temperature range, resulting in a more efficient use of the CRAH than in the other workload distributions.

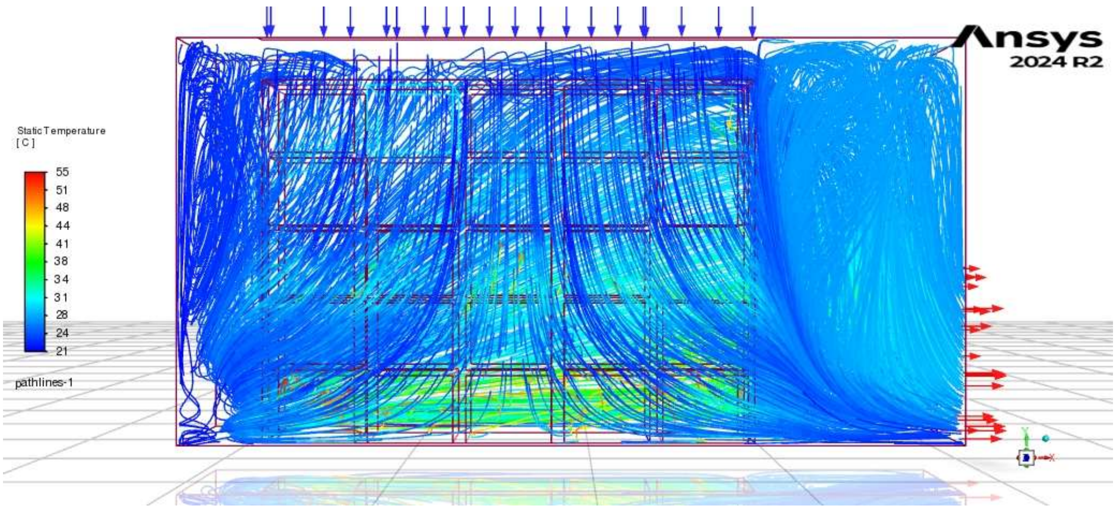

Figure 5.8 Cold aisle view illustrates that airflow enters from the inlet before descending through the rack region. The path lines converge in the middle row, indicating preferential flow through regions of lower resistance.

5.5 Analysis and Discussion of Airflow Characteristics

The airflow inside the data centre was analysed. In addition to the thermal performance, the airflow behaviour inside the containerised data centre provides important insight into the mechanisms governing heat removal and cooling efficiency. Figure 5.8 illustrates the airflow path lines coloured according to the static temperature (blue: coldest and red: hottest). The path lines reveal the interaction between the inlet, the server racks, and the containerised data centre walls, highlighting regions of efficient cooling as well as areas susceptible to flow recirculation and thermal accumulation. The airflow is characterised by a predominantly forced-convection regime in which air enters the containerised data centre through the inlet and is redirected around the server racks before exiting through the outlet. Rather than travelling uniformly across the container, the incoming flow undergoes significant deflection as it encounters the rack, resulting in complex three-dimensional circulation patterns. In Figure 5.9 the path lines indicate that the racks act as major flow obstructions, producing acceleration through the rack rows while simultaneously generating wake regions downstream. These wakes are associated with reduced local velocities and increased flow mixing, which influence the spatial distribution of temperature throughout the containerised data centre. One of the most prominent features visible in the path lines is the formation of large recirculation zones adjacent to the side walls and near the corners of the container. The recirculating flow is characterised by closed-loop path lines that remain within the enclosure for relatively long residence times before reaching the outlet. Such behaviour indicates that portions of the cooling air are repeatedly circulated rather than contributing effectively to heat removal. The recirculation is particularly evident in three areas: near the downstream bottom side of the cold aisle, close to the side walls, and underneath the fourth row of the rack. These regions generally correspond to relatively low-momentum flow where inertial forces are insufficient to overcome adverse pressure gradients created by the obstacles. The existence of these circulation zones reduces cooling effectiveness because warmer air remains trapped within the data centre instead of being transported directly towards the CRAH.

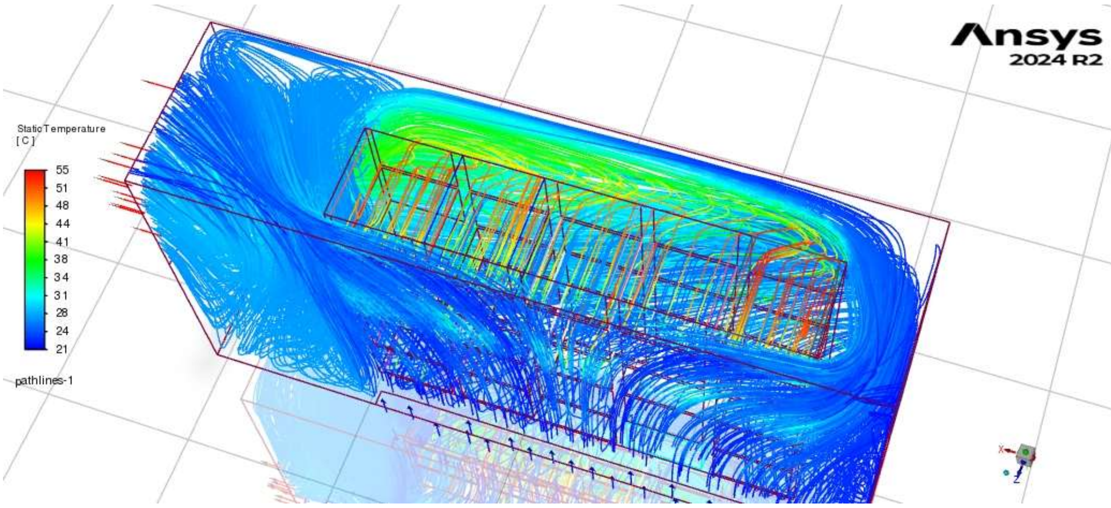

Figure 5.9 The view shows the 3-D nature of the airflow. High-temperature path lines (green–yellow–red) are within and immediately downstream of the rack, illustrating the transport of heat generated by the servers.

The airflow distribution through the rack appears non-uniform. The central area of the rack receives comparatively stronger airflow, as indicated by the denser and more continuous path lines passing through the middle of the server. Conversely, peripheral regions experience weaker flow penetration. This non-uniformity suggests that some server locations receive greater cooling capacity than others.

Areas supplied with higher local velocity benefit from enhanced convective heat transfer, whereas regions experiencing lower airflow are more susceptible to elevated operating temperatures. The temperature distribution reveals several localised hotspots, particularly around the upper portions of the rack and downstream of the heat-generating components. These hotspots are associated with two principal mechanisms:

Insufficient Air Velocity: Where airflow velocity decreases due to flow separation or recirculation, the local convective heat transfer coefficient is reduced. Consequently, heat generated by the servers is removed less efficiently, allowing local temperatures to increase.

Hot Air Recirculation: Some of the hot exhaust air appears to be entrapped back towards the rack rather than being immediately removed through the outlet. This recirculation mixes hot air with the incoming cooling air, increasing the inlet temperature experienced by downstream equipment. This is particularly evident from the elevated-temperature path lines (green to yellow) that remain within the enclosure before exiting.

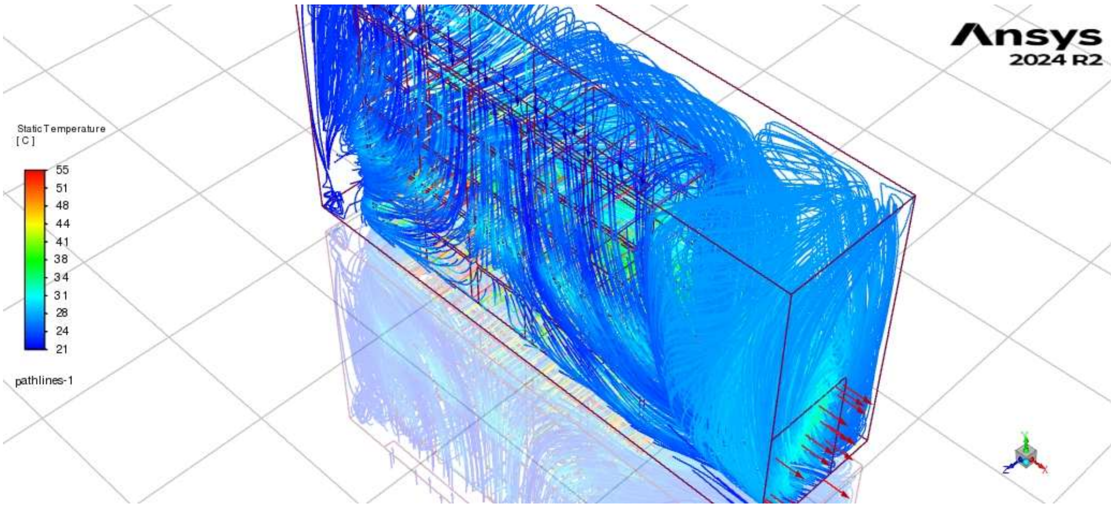

Figure 5.10 Isometric view highlights the accumulation of recirculating flow near the inlet of the container. The path lines indicate that not all of the heated air exits directly through the outlet. Instead, portions of the flow remain trapped within.

The path lines in Figure 5.10 also demonstrate substantial mixing between the cold air and hot air. From a thermal management perspective, moderate mixing can improve temperature uniformity within the data centre. However, excessive mixing reduces the temperature difference between the cooling air and the server; therefore, decreasing the driving force for convective heat transfer.

Consequently, although increased turbulence improves local heat transfer coefficients, excessive recirculation may reduce this benefit by raising the air temperature. The airflow shows greater momentum through the rack row compared with low-velocity conditions. Increased inlet velocity enhances penetration of the cooling air into the rack, reducing stagnant regions and improving heat extraction. However, increasing velocity also strengthens the large circulation cells formed after the airflow exits the rack. These vortical structures are particularly noticeable near the downstream wall, where the high-momentum supply air on the enclosure before turning towards the outlet. This suggests that there exists an optimal operating velocity that balances improved convective cooling with minimised recirculation losses.

The thermal field observed within the data centre is closely coupled to the underlying airflow structure. Regions exhibiting high airflow velocities correspond to lower temperatures owing to enhanced convective heat transfer, whereas areas characterised by recirculation and reduced momentum coincide with localised thermal accumulation. The results therefore demonstrate that cooling effectiveness depends not only on the supplied airflow rate but also on its spatial distribution and the degree of flow uniformity through the rack. Finally, from an energy perspective, increasing the supply air velocity generally enhances convective heat removal by increasing the local heat transfer coefficient. However, the airflow visualisations indicate that higher velocities may also intensify recirculation and mixing within the containerised data centre. Consequently, further increases in fan speed may yield diminishing thermal benefits while incurring a disproportionate increase in fan power consumption.

This highlights the importance of identifying an optimal operating velocity that achieves effective cooling with minimal energy expenditure.

5.6 Conclusions

This chapter evaluated the influence of workload distribution on the thermal performance of a containerised data centre using Computational Fluid Dynamics (CFD) simulations under a constant-flow cooling strategy. Five workload scenarios were initially defined in Table 5.1 to represent different overall utilisation levels and workload placement strategies commonly reported in the literature. These scenarios were designed to reflect representative data centre operating conditions as discussed in Chapter 3, where workload characterisation and thermal-aware scheduling were identified as important mechanisms for reducing cooling energy demand while maintaining acceptable operating temperatures.

Due to the computational requirements associated with modelling the full set of scenarios, only Scenario 1 (60% utilisation) and Scenario 3 (30% utilisation) were investigated in detail. This decision was based on the available computational resources and the complexity of the CFD model, which represented half of the containerised data centre using several simplifying assumptions. The remaining scenarios (Scenarios 2, 4 and 5) were defined to provide a broader framework for future investigation and to demonstrate how the proposed methodology can be extended to alternative workload intensities and distributions. Consequently, no conclusions are drawn regarding their thermal performance.

Within the two investigated scenarios, six workload distributions were examined. Workload Distributions A, B and C correspond to Scenario 1 (60% workload), while Workload Distributions D, E and F correspond to Scenario 3 (30% workload), as defined in Figure 5.1 and Table 5.2. Figure 5.2 and the subsequent figures should therefore be interpreted as different workload allocation strategies within each scenario rather than as separate scenarios. The successive workload distributions represent an iterative redistribution of computational load, where the results obtained from one distribution informed the development of the next in an effort to minimise hotspot formation and improve thermal uniformity.

The simulation results demonstrate that workload placement has a significant influence on the spatial distribution of temperature inside the containerised data centre. The initial workload distributions (A, B and C) produced localised hotspots in the cold aisle that exceeded the ASHRAE recommended temperature, indicating that simply balancing computational load across servers does not necessarily produce an optimal thermal distribution. Progressive redistribution of the workload towards thermally favourable regions (Distributions D, E and ultimately F) reduced hotspot intensity, improved airflow utilisation and maintained cold-aisle temperatures within the recommended operating range. Among the investigated cases, Workload Distribution F provided the most favourable thermal performance, producing the lowest hotspot temperatures and reducing the likelihood of unnecessary CRAH operation.

The airflow analysis further demonstrated that thermal performance depends not only on the cooling airflow rate but also on the interaction between airflow distribution, heat generation and workload placement. Regions exhibiting recirculation and reduced airflow momentum consistently corresponded to areas of thermal accumulation, whereas improved workload placement reduced these effects and enhanced convective heat removal. These findings support the hypothesis that intelligent workload allocation can complement conventional cooling strategies by reducing thermal hotspots without increasing cooling airflow.

The findings presented in this chapter are consistent with the literature reviewed in Chapter 3 supporting the principles of thermal-aware scheduling, which suggest that workload allocation should consider the thermal state of the data centre rather than relying solely on computational resource utilisation. Whereas much of the existing literature focuses on reducing IT energy consumption through resource consolidation or processor utilisation, the present work demonstrates that workload distribution can also be employed to improve cooling effectiveness by influencing the spatial temperature distribution within the facility. The results therefore provide additional evidence that workload scheduling can be integrated with thermal management strategies to reduce cooling energy requirements while maintaining acceptable operating conditions.

The conclusions of this chapter should be interpreted within the assumptions and limitations of the numerical model. To reduce computational cost, the simulations represented only half of the containerised data centre, server fans were modelled using equivalent momentum sources rather than explicit fan geometries, and the porous medium representation was simplified to reproduce the resistance of the servers. Furthermore, the simulations were performed under constant inlet airflow conditions and did not consider transient workloads, dynamic scheduling, seasonal environmental effects or variable cooling system operation. These assumptions necessarily limit the direct generalisation of the quantitative results; however, they do not diminish the value of the qualitative trends observed regarding the relationship between workload distribution, airflow behaviour and hotspot formation. This chapter demonstrates that careful redistribution of computational workloads can substantially improve the thermal behaviour of a containerised data centre. The results establish a clear relationship between workload placement, airflow characteristics and cooling effectiveness, providing a foundation for the development of thermal-aware scheduling strategies capable of reducing cooling energy consumption. These findings form the basis for the subsequent chapter, where the implications of the proposed methodology are discussed in the broader context of energy efficient data centre operation.

Chapter 6 Analysis of energy consumption

Chapter 5 demonstrated how different workload distributions influence airflow behaviour, hotspot formation and temperature distribution inside the containerised data centre through CFD simulations. The simulations provided detailed thermal results for each workload distribution, including outlet air temperature, maximum cold-aisle and hot-aisle temperatures, and server temperatures.

The present chapter builds directly upon those CFD results by translating the predicted thermal behaviour into electrical energy requirements. Two complementary aspects of energy consumption are considered. First, the electrical power consumed by the IT equipment is estimated from the workload allocated to each server using published server power characteristics. Secondly, the cooling energy required to remove the heat generated by the servers is estimated using the airflow rates and temperatures obtained from the CFD simulations in Chapter 5.

Consequently, the thermal outputs obtained in Chapter 5 provide the input parameters for the energy analysis presented in this chapter. The workload distributions that produced lower hotspot temperatures are expected to require less cooling energy while maintaining acceptable operating temperatures, allowing the overall energy implications of the proposed workload allocation strategy to be evaluated.

6.1 Methodology for Energy Estimation

The CFD simulations presented in Chapter 5 predict the thermal behaviour of the containerised data centre for each workload distribution under different inlet airflow conditions. The simulations provide temperatures, airflow velocities and heat transfer characteristics but do not directly calculate electrical energy consumption. Therefore, an energy model was developed to convert the CFD results into estimates of IT power consumption and cooling energy demand. The methodology consists of two independent calculations:

The first stage of the methodology estimates the electrical power consumed by the servers. Each server in the CFD model is assigned one of the five operating states defined in Table 6.1 (Off, Idle, Quarter, Medium and Full Workload) according to the workload allocated during each simulation. The electrical power associated with each operating state is taken from the SPECpower_ssj2008 benchmark and the measurements reported by Qiu et al. (2019). The total IT electrical power is then obtained by summing the power consumption of all servers operating within a particular workload distribution.

Because almost all the electrical energy consumed by modern servers is eventually converted into heat through processors, memory, storage devices and power supply losses, the IT electrical power is assumed to be equal to the heat generation used in the CFD simulations. This assumption is commonly adopted in thermal analyses of data centres.

The second stage estimates the amount of cooling required to remove the heat generated by the servers. The CFD simulations provide inlet air temperature, outlet air temperature, airflow rate and temperature rise across the rack. These outputs are then used to calculate the rate of heat removed by the cooling air using:

$$\dot{Q} = \dot{m} c_p \Delta T$$

Where:

$\dot{Q}$ = heat removal rate

$\dot{m}$ = mass air flow rate (kg s^{-1})

c_p = specific heat capacity of air (1005 $J\ kg^{-1} K^{-1}$)

ΔT = temperature increase of the cooling air (K or °C)

$$\Delta T = T_{OUT} - T_{IN}$$

Where:

T_{OUT} = average outlet air temperature obtained from the CFD simulations (°C)

T_{IN} = inlet air temperature supplied by the CRAH (21°C)

$$\dot{m} = \rho \dot{V}$$

Where:

ρ = air density (kg m^{-3})

$\dot{V}$ = volumetric airflow supply ($m^3 s^{-1}$)

The mass flow rate represents the amount of air entering the data centre every second. Since heat is transported by the moving air, the mass flow rate determines the cooling capacity available. A higher airflow rate increases the amount of heat that can be removed from the servers. For the present study, the air density was assumed constant at:

$$\rho = 1.2\ kg/m^3$$

which is representative of air at approximately room temperature.

The resulting heat removal rate represents the thermal load that must be removed by the CRAH system.

The analysis assumes:

- constant inlet air temperature of 21°C
- steady-state operation
- constant air properties
- no heat losses to the surroundings
- all server electrical power is ultimately converted into heat.

6.2 Energy Consumption with IT in different states

This section presents the power consumption model used to estimate the IT energy demand associated with the workload distributions analysed in Chapter 5. Unlike the five workload scenarios introduced in Chapter 5, which describe different overall workload allocations across the data centre (30%, 50%, 60%, 70%, and 90% utilisation with different spatial distributions), Table 6.1 defines five server operating states used for the energy calculations based on the actual workload as specified by SPECpower_ssj2008 and Qiu et al. (2019) and forms the basis for estimating energy use in each state, as in Table 6.1

Each server in the CFD simulations is assigned to one of five operating states according to the workload allocated to that server:

- **Off:** server switched off (0 W)
- **Idle:** powered on but processing no workload
- **Quarter workload:** approximately 25% server utilisation
- **Medium workload:** approximately 50% server utilisation
- **Full workload:** 100% server utilisation

The corresponding heat generation and electrical power consumption for each operating state is taken from the SPECpower_ssj2008 benchmark and Qiu et al. (2019) and are summarised in Table 6.1. These operating states provide the basis for estimating the total IT energy consumption reported later in Tables 6.2 and 6.3.

The five operating states of the servers should therefore not be confused with the five workload scenarios presented in Chapter 5. The Chapter 5 scenarios define the overall workload level and its spatial distribution across the data centre, whereas Table 6.1 defines the power model applied to individual servers when calculating the energy consumption for those scenarios.

Table 6.1 defines the server power model used throughout this chapter. Each server in the CFD model is assigned to one of five operating states according to its allocated workload. The corresponding electrical power and heat generation are obtained from published measurements reported by SPECpower_ssj2008 and Qiu et al. (2019). These values are subsequently used to estimate the total IT electrical power for each workload distribution analysed in Chapter 5.

Workload	Target Workload	Heat Generation	Power Consumption (W)
Full	100%	1000	944
Medium	50%	500	431
Quarter	25%	250	300
Idle	0%	83	82.9
Off	0%	0	0

Table 6.1 Shows that the server is near the peak energy efficiency at 50% utilisation, while consuming 431 W. The idle power draw is 82.9 W (approximately 8.8% of full load power).

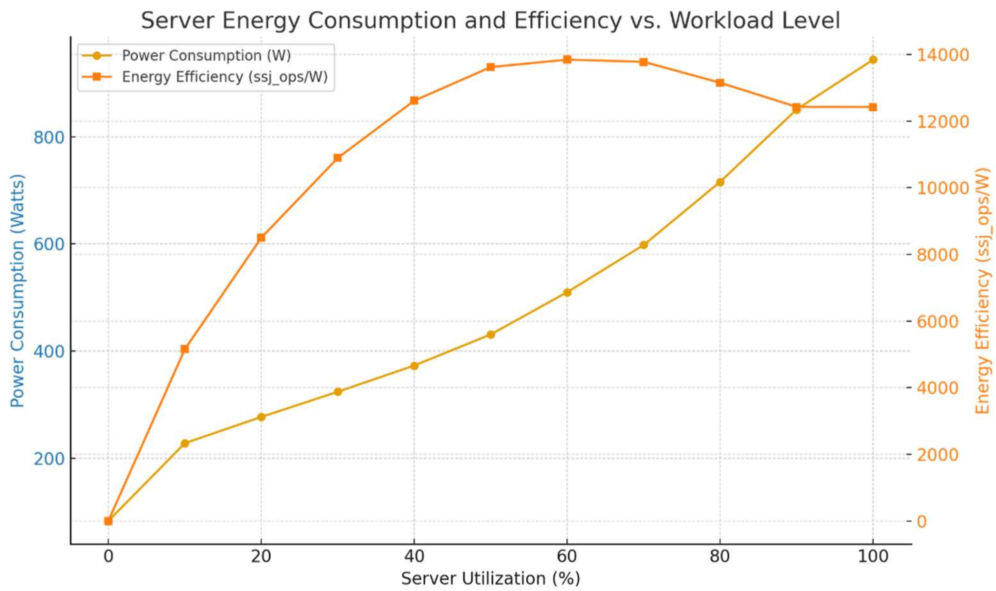

Figure 6.1 shows a graphic with the Server Energy Consumption and the Energy Efficiency by Workload level based on the data provided by Qiu et al. (2019)**.**

6.3 Energy Consumption using this approach

According to the workload distribution described in Figure 5.1 A) and B) and considering the information on power consumption in Table 6.1, the amount of energy consumption by the IT Equipment (Servers) can be inferred and is estimated at 14160 Watts. This inference is made by cross referencing all the cases shown in A) Workload Distribution, along with the level of workload the servers are expected to be processing, and information from Table 6.1. Therefore, a compilation of the different workload distributions and their IT Energy Consumption shown in Watts can be seen in Table 6.2 which presents that Workload Distribution C uses 410 Watts less and has the same Workload usage than Workload Distribution A and B. Additionally, Workload Distribution F should consume 82 Watts less than Workload Distribution E and Workload Distribution E should consume 164 Watts less than Workload Distribution D.

	Workload Distribution	Case	Inlet Air Flowrate	Inlet Temp in °C	Avg. Temp in Outlet in °C	Max Temp in Cold Aisle in °C	Max Temp in Hot Aisle in °C	Max Temp Inside Blocks in °C	IT Energy Consumption in W
Scenario 1	A	1	V1	21	27.47	30.59	53.83	51.97	14160
		2	V2	21	29.20	34.77	64.82	62.92	14160
		3	V3	21	26.66	28.92	48.42	48.73	14160
	B	4	V1	21	27.20	34.30	58.00	57.41	14160
		5	V2	21	29.45	39.93	65.60	66.35	14160
		6	V3	21	26.53	32.12	52.66	53.84	14160
	C	7	V1	21	27.21	33.69	52.14	56.04	13750
		8	V3	21	26.52	32.05	49.93	52.22	13750
Scenario 3	D	9	V3	21	23.33	27.59	41.72	46.18	5882
		10	V1	21	23.66	28.51	44.15	48.44	5882
		11	V2	21	24.52	30.36	49.06	54.47	5882
	E	12	V3	21	23.32	27.34	39.00	45.53	5718
		13	V1	21	23.68	28.30	43.86	48.14	5718
		14	V2	21	24.49	30.11	48.81	52.87	5718
	F	15	V3	21	23.34	27.34	38.39	45.52	5636
		16	V1	21	23.71	28.29	39.90	47.85	5636
		17	V2	21	24.47	30.10	43.06	52.30	5636

Table 6.2 presents the information corresponding to the energy consumption by the servers in each workload distribution, where the inlet air flowrate for V1: 3.52 m/s, V2: 2.83 m/s, and V3: 4m/s.

6.4 Energy required for Air Cooling

Calculating the energy consumption for a containerised data centre equipped with a CRAH involves more than just the duration of the CRAH's operation; it also depends on the cooling air temperature. Therefore, the maximum temperature in the Cold Aisle should be maintained at no more than 27 °C. Another important factor is the flow rate, which also influences the amount of energy consumed to achieve these results.

This research investigates a halved containerised data centre with 5 racks, where each one of them was delivering 3 kW of power for a total IT load of 15 kW in scenario 1. Cooling is provided by a CRAH supplied by a chilled water system. The air handler's flow rate is adjustable, and the numerical simulations were conducted at different airflow rates (2.83, 3.52 and 4 m^3/s). The inlet air temperature to the racks is kept constant at 21°C during all tests.

The cooling system energy consumption according to Wang et al. (2017) is almost 50% of the total energy consumption in data centres. Assuming a total of 30 kW IT load, the cooling system would typically consume up to 15 kW (assuming a 50% share), but the actual value depends on the efficiency of the cooling design.

The performance metrics to quantify the cooling effectiveness is the Rack Cooling Index (RCI) and Supplied Heat Index (SHI). Higher RCI and lower SHI values indicate better cooling efficiency and less energy wasted on recirculation or bypass.

$$RCI_{LO} = \left[1 - \frac{\sum (T_{min-rec} - T_x)_{T_x < T_{min-rec}}}{(T_{min-rec} - T_{min-all}) \times n}\right] \times 100\%$$

$$\mathrm{SHI} = \frac{\sum_i \sum_j (T^T_{in_{i,j}} - T^C_{sup})}{\sum_i \sum_j (T^T_{out_{i,j}} - T^C_{sup})}$$

$$RCI_{HI} = \left[1 - \frac{\sum (T_x - T_{max-rec})_{T_x > T_{max-rec}}}{(T_{max-all} - T_{max-rec}) \times n}\right] \times 100\%$$

Where:
T_x: Intake temperature for each rack
n: Number of racks
$T_{\text{max-rec}}$: Maximum recommended temperature (ASHRAE: 27 °C)
$T_{\text{max-all}}$: Maximum allowable temperature (ASHRAE: 32 °C)
$T_{\text{min-rec}}$: Minimum recommended temperature (ASHRAE: 18 °C)
$T_{\text{min-all}}$: Minimum allowable temperature (ASHRAE: 15 °C)
However, this research focus on quantifying the amount of energy required to cool down the air rejected by the servers.

- Airflow rate: Variable: (V1: 3.52 m^3/s, V2: 4 m^3/s and V3: 2.83 m^3/s)
- Air density (approx.):

$$\rho = 1.2 \text{ kg/m}^3$$

- Specific heat of air:

$$c_p = 1005 \text{ kJ/(kg·K)}$$

- Temperature drop:

$$\Delta T = Air\ temperature - Target\ air\ temperature$$

- Mass Flow Rate:

$$\dot{m} = \rho \times Volumetric\ Flow\ Rate$$

- Heat transfer rate (Power):

$$\dot{Q} = \dot{m} \times c_p \times \Delta T$$

The heat removal rate obtained from the heat transfer rate is expressed in Watts (W), which is equivalent to Joules per second (J s^{-1}). Dividing by 1000 converts this value to kilowatts (kW). Assuming steady-state operation for one hour, the cooling energy is:

$$Ecool = \dot{Q} \times 1\ h$$

Example Calculation (Case A1)

The procedure used to estimate the cooling energy is illustrated using **Case A1**, where the inlet airflow velocity is **V1 (3.52 m^3 s^{-1})**.

The CFD simulation predicted an average outlet air temperature of **27.47 °C**, while the inlet temperature was maintained at **21 °C**. Therefore, the temperature rise of the cooling air is:

$$\Delta T = T_{OUT} - T_{IN}$$

$$\Delta T = 27.47℃ - 21℃$$

$$\Delta T = 6.47℃$$

Using an air density of **1.2 kg m⁻³**, the mass flow rate is:

$$\dot{m} = \rho \dot{V}$$

$$\dot{m} = (1.2)(3.52)$$

$$\dot{m} = 4.22\, kg/s$$

The rate of heat removed by the cooling air is then calculated as:

$$\dot{Q} = \dot{m} \times c_p \times \Delta T$$

where the specific heat capacity of air is:

$$c_p = 1005 kJ/(kg{\cdot}K)$$

$$\dot{Q} = (4.22 kg/s) \times (1005 kJ/(kg{\cdot}K)) \times (6.47℃)$$

$$\dot{Q} = 27.48\, kW$$

Assuming steady-state operation over one hour, the cooling energy required is:

$$E_{COOL} = 27.48\, kWh$$

The IT equipment operating under Workload Distribution A consumes **14.16 kW**, corresponding to **14.16 kWh** over one hour.

Finally, the total energy consumption for Case A1 is:

$$E_{TOTAL} = E_{IT} + E_{COOL}$$
$$E_{TOTAL} = 14.16 kWh + 27.48 kWh$$
$$E_{TOTAL} = 41.64 kWh$$

This value corresponds to the total energy consumption reported for Case A1 in Table 6.3.

	Workload Distribution	Case	Inlet Air Flowrate	Inlet Temp in ºC	Avg. Temp in Outlet in ºC	Max Temp in Cold Aisle in °C	Max Temp in Hot Aisle in °C	Max Temp Inside Blocks in °C	IT Energy Consumption (kWh)	Temperature difference (K)	Mass Flowrate (kg/s)	Heat transfer rate (W)	Cooling Energy (kWh)	Total Energy Consumption per case (kWh)
Scenario 1	A	1	V1	21	27.47	30.59	53.83	51.97	14.16	6.47	4.2	27482.91	27.48	41.64
		2	V2	21	29.20	34.77	64.82	62.92	14.16	8.20	3.4	27999.61	28.00	42.16
		3	V3	21	26.66	28.92	48.42	48.73	14.16	5.66	4.8	27310.11	27.31	41.47
	B	4	V1	21	27.20	34.30	58.00	57.41	14.16	6.20	4.2	26324.84	26.32	40.48
		5	V2	21	29.45	39.93	65.60	66.35	14.16	8.45	3.4	28827.67	28.83	42.99
		6	V3	21	26.53	32.12	52.66	53.84	14.16	5.53	4.8	26669.68	26.67	40.83
	C	7	V1	21	27.21	33.69	52.14	56.04	13.75	6.21	4.2	26346.83	26.35	40.10
		8	V3	21	26.52	32.05	49.93	52.22	13.75	5.52	4.8	26622.74	26.62	40.37
Scenario 3	D	9	V3	21	23.33	27.59	41.72	46.18	5.88	2.33	4.8	11254.49	11.25	17.14
		10	V1	21	23.66	28.51	44.15	48.44	5.88	2.66	4.2	11308.70	11.31	17.19
		11	V2	21	24.52	30.36	49.06	54.47	5.88	3.52	3.4	12006.42	12.01	17.89
	E	12	V3	21	23.32	27.34	39.00	45.53	5.72	2.32	4.8	11192.55	11.19	16.91
		13	V1	21	23.68	28.30	43.86	48.14	5.72	2.68	4.2	11360.71	11.36	17.08
		14	V2	21	24.49	30.11	48.81	52.87	5.72	3.49	3.4	11917.03	11.92	17.64
	F	15	V3	21	23.34	27.34	38.39	45.52	5.64	2.34	4.8	11289.51	11.29	16.93
		16	V1	21	23.71	28.29	39.90	47.85	5.64	2.71	4.2	11487.46	11.49	17.12
		17	V2	21	24.47	30.10	43.06	52.30	5.64	3.47	3.4	11858.06	11.86	17.49

Table 6.3 Presents the results of numerical calculations for six different workload distributions, where cases with less total energy are highlighted for each Workload Distribution.

6.5 Discussion

The findings from this analysis highlight a complex and interdependent nature of energy consumption within a containerised data centre environment. First, examining IT equipment across different operational states – off, active 25%, 50% and 100% utilisation – reveals that a substantial proportion of total energy demand is driven by baseline consumption. Even when servers are underutilised, they continue to draw significant power; for example, if the server was idling, the energy consumption of the server would be equivalent to that of a server at 8.8% utilisation, which corroborates widely reported inefficiencies in traditional data centre operations. This underscores the importance of strategies that minimise idle capacity and ensure that deployed hardware is used as efficiently as possible.

The evaluation of the six different workload distribution strategies explored in this chapter demonstrates the potential for energy savings when computational resources are carefully allocated. The results suggest that relatively simple scheduling or consolidation mechanisms can show measurable reductions in energy consumption without compromising service quality. However, the benefits are not uniform; the magnitude of savings depends heavily on workload characteristics, such as temporal variability, peak to average ratios, and the degree to which tasks can be migrated or deferred. Furthermore, aggressive consolidation must be balanced against the risk of creating thermal hotspots or overloading specific systems, for example Workload Distribution A2 and B5, which offset energy gains with higher cooling requirements or reduced hardware longevity due to the higher temperature in the hotspots generated by the aggressive consolidation in these case studies.

Cooling system energy consumption emerges as a critical component of total energy consumption in data centres. The analysis in this chapter shows that, within a containerised architecture, the confined physical space can intensify thermal loads generating hotspots and making efficient cooling essential to maintaining operational reliability. The energy required to cool the internal environment is strongly influenced by several factors including external climatic conditions, rack density, airflow design and the heat rejected by the servers themselves. Although in this set of simulations the external climatic conditions were neglected, the results show that reductions in IT load through workload distribution and consolidation also translate to reduced cooling demand, which, illustrates the cascading impact of the IT optimisations on facility-level energy use. Moreover, the effectiveness of cooling systems is highly sensitive to design choices: inadequate airflow management or inefficient heat exchange can significantly reduce the gains achieved through workload optimisation.

Overall, the relation between IT energy consumption, workload distribution strategies, and cooling requirements, highlights the need for a holistic approach to energy management in containerised data centres. Optimising any single component in isolation proves to be useful as shown in Table 6.2, but this research highlights that meaningful improvements arise from coordinated strategies that align IT utilisation, thermal management and facility design as observed in case E12 in Table 6.3.

Chapter 7 Conclusions

7.1 Overall Conclusions

This thesis investigated the hypothesis that the energy efficiency of a containerised data centre can be improved through the careful distribution of computational workloads to minimise hotspot formation. Computational Fluid Dynamics (CFD) was employed to investigate the interaction between workload allocation, airflow behaviour and cooling performance, while an energy analysis quantified the impact of different workload distributions on overall energy consumption.

The results presented throughout this thesis demonstrate that the research aim has been achieved. The study showed that thermally aware workload allocation significantly influences airflow distribution, hotspot formation and cooling energy demand. By strategically redistributing workloads away from thermally unfavourable regions, hotspot intensity was reduced, airflow uniformity improved and overall energy consumption decreased without requiring additional cooling capacity. These findings demonstrate that workload scheduling should be considered not only as an IT resource management strategy but also as an effective thermal management approach for improving the energy efficiency of containerised data centres.

7.2 Achievement of the Research Aim and Objectives

The research presented in this thesis successfully addressed each of the objectives established in Chapter 1.

Objective 1: Develop a Computational Fluid Dynamics model of a containerised data centre based on the experimental system of Wang et al. (2017). This objective was successfully achieved through the development and validation of a three-dimensional CFD model capable of predicting airflow distribution, temperature fields and hotspot formation within a containerised data centre. The model incorporated realistic representations of server racks, cooling airflow and heat generation, while maintaining computational efficiency through the use of porous medium and momentum source formulations. Validation against published experimental measurements demonstrated that the model provided sufficient accuracy for investigating alternative workload allocation strategies.

Objective 2: Review the relevant literature to establish typical server characteristics, workload patterns and energy consumption metrics.

The literature review identified the principal cooling technologies used within modern data centres, together with existing approaches for workload scheduling and energy-efficient operation. The review highlighted that most scheduling strategies focus primarily on computational performance or power consumption, with relatively limited consideration of thermal behaviour inside containerised data centres. This knowledge gap provided the motivation for developing a thermal-aware workload allocation methodology capable of integrating airflow behaviour into workload scheduling decisions.

Objective 3: Conduct a parametric study of different workload distributions to determine their effects on hotspot locations and intensity.

This objective was achieved through the CFD investigations presented in Chapter 5. Multiple workload allocation strategies were examined under representative operating conditions, demonstrating that workload placement has a significant influence on airflow recirculation, cold-aisle temperatures and hotspot formation.

The initial workload distributions generated severe thermal non-uniformity, producing maximum cold-aisle temperatures approaching **39.93°C**, substantially exceeding both the ASHRAE recommended temperature of **27°C** and the maximum allowable limit of **32°C**. Maximum internal server temperatures reached **66.35°C**, indicating significant hotspot formation. Through successive redistribution of computational workloads, Workload Distribution F reduced the maximum cold-aisle temperature to **27.34°C** while reducing maximum server temperatures to **45.52°C**, demonstrating that intelligent workload placement can substantially improve thermal performance without increasing cooling airflow.

Objective 4: Quantify the corresponding energy consumption under different workload configurations to determine the potential energy savings achievable through workload distribution. The energy analysis presented in Chapter 6 demonstrated that improved thermal performance translated directly into reduced energy consumption. Under the optimised workload configurations, IT energy consumption decreased from **14.16 kWh** to **13.75 kWh** in Scenario 1, while total facility energy consumption decreased from **17.89 kWh** to **16.91 kWh**, representing approximately a **5.5% reduction**. These findings demonstrate that carefully distributing computational workloads can simultaneously improve thermal performance and reduce overall energy demand. The results therefore support the research hypothesis that thermal-aware workload allocation represents an effective strategy for improving the energy efficiency of containerised data centres.

Collectively, the results presented in Chapters 5 and 6 demonstrate that workload distribution and airflow management should be considered together when developing energy-efficient cooling strategies. Rather than increasing cooling capacity, redistributing computational workloads away from thermally sensitive regions provides a more sustainable approach to reducing hotspot formation while maintaining acceptable operating temperatures.

7.3 Principal Contributions

The principal contributions of this research are summarised as follows.

• Development of a validated CFD methodology capable of predicting airflow behaviour, temperature distributions and hotspot formation within a containerised data centre.

• Demonstration that computational workload allocation significantly influences thermal behaviour and cooling performance.

• Quantification of the relationship between workload redistribution and energy consumption, showing that thermally aware workload allocation can reduce total energy consumption by approximately 5.5%.

• Establishment of thermal-aware workload scheduling as a practical cooling optimisation strategy that complements conventional airflow management techniques.

These contributions provide a framework for integrating thermal analysis into future workload scheduling algorithms and intelligent cooling control systems.

7.4 Limitations

The findings presented in this thesis should be interpreted within the context of several modelling assumptions and computational limitations. Although the Computational Fluid Dynamics (CFD) simulations provide valuable insight into the relationship between workload allocation, airflow behaviour and thermal performance, a number of simplifications were necessary to ensure that the simulations remained computationally feasible.

First, only half of the containerised data centre was represented within the numerical model. This simplification reduced the computational cost while preserving the principal airflow and heat transfer mechanisms within the server racks. Nevertheless, modelling only half of the facility may not fully capture large-scale airflow interactions, pressure distributions and thermal recirculation patterns that could occur within a complete containerised data centre. Consequently, while the qualitative trends observed are expected to remain valid, the absolute temperature distributions and cooling requirements may differ slightly from those obtained in a full-scale model.

Secondly, several geometric simplifications were introduced in the server representation. Rather than modelling each individual server fan, equivalent momentum sources were applied within the porous medium representation of the servers. Similarly, the porous medium were used to reproduce the overall flow resistance of the racks rather than explicitly modelling the internal server components. These assumptions significantly reduced computational requirements but inevitably simplified the complex airflow structures generated inside real servers.

A further limitation is that all simulations were conducted under steady-state conditions using a constant inlet air temperature of 21 °C and fixed airflow rates of 2.83, 3.52 and 4.00 $m^3 s^{-1}$. In practical data centre environments, workloads fluctuate continuously, server utilisation changes over time and cooling systems dynamically adjust their operating conditions in response to thermal demand. Consequently, the present study does not capture transient thermal behaviour, workload migration or the dynamic response of the Computer Room Air Handler (CRAH) system.

The study also investigated only two of the five proposed workload scenarios in detail. Although five representative utilisation scenarios (30%, 50%, 60%, 70% and 90%) were defined to establish a comprehensive experimental framework, computational resource limitations restricted the detailed CFD investigation to the 30% and 60% utilisation cases. While these scenarios demonstrate the effectiveness of thermal-aware workload allocation, further investigation of the remaining utilisation levels would provide a more comprehensive understanding of workload scheduling across a wider operating range.

Furthermore, the energy analysis presented in Chapter 6 was derived from the thermal outputs predicted by the CFD simulations rather than direct experimental measurements. The IT power consumption was estimated using published SPECpower benchmark data and the assumption that virtually all electrical energy consumed by the servers is converted into heat. Although this assumption is widely accepted in data centre thermal analysis, experimental validation would improve confidence in the predicted cooling energy requirements.

Despite these limitations, the numerical model successfully demonstrated consistent relationships between workload distribution, airflow behaviour, hotspot formation and cooling energy demand. The principal contribution of this work, therefore, lies in identifying the mechanisms through which thermal-aware workload allocation can improve cooling effectiveness, providing a sound basis for future experimental and numerical investigations.

7.5 Future Work

The most important extension of this research is the development of transient CFD simulations capable of representing realistic operating conditions over extended periods. Rather than investigating isolated steady-state operating points, future work should simulate every hour of daily operation over an entire year (8,760 hourly simulations) using meteorological data representative of different geographical locations. This would enable seasonal and diurnal variations in environmental conditions to be incorporated into the thermal analysis and would provide a more representative assessment of cooling performance under real operating conditions.

To support these simulations, the CFD model should incorporate dynamic boundary conditions that vary continuously according to monitored environmental variables. Hourly weather datasets should provide ambient air temperature, solar radiation, wind speed and relative humidity for each geographical location. The ambient temperature would be used to determine the overhead cooling-air inlet temperature supplied to the containerised data centre, while solar radiation would be applied as transient heat fluxes on the external walls and roof receiving direct sunlight. The external air surrounding the container would therefore influence both the conductive heat transfer through the enclosure and the cooling requirements of the internal environment throughout the year.

Future work should also implement feedback-controlled boundary conditions, allowing the inlet cooling conditions to respond automatically to monitored operating variables. For example, the overhead supply-air temperature and airflow rate could be adjusted according to measured outdoor ambient temperature, server inlet temperatures or hotspot development within the cold aisle. Such a control strategy would more accurately represent the operation of modern CRAH systems, where cooling conditions continuously adapt to changing environmental and thermal loads.

The remaining utilisation scenarios of 50%, 70% and 90% should also be investigated to determine whether the thermal and energy benefits observed in this research remain consistent across the full operational range of the data centre.

Future studies should further integrate the CFD model with intelligent workload scheduling algorithms, including optimisation techniques, machine learning methods or digital twin frameworks capable of predicting hotspot formation and dynamically redistributing computational workloads before thermal limits are exceeded.

Experimental validation should also be undertaken using either a physical containerised data centre or operational measurements obtained from industrial facilities. Continuous measurements of temperatures, airflow rates and cooling energy consumption would enable the transient CFD model to be calibrated and validated under realistic operating conditions.

Finally, the methodology should be extended to larger facilities containing multiple rows of racks, different rack power densities and alternative cooling architectures. Such investigations would enable the proposed thermal-aware workload allocation methodology to be evaluated under increasingly realistic data centre configurations and support its implementation within next-generation energy-efficient data centres.

7.6 Final Remarks

This research has demonstrated that thermal-aware workload allocation, combined with appropriate airflow management, provides an effective means of improving the thermal and energy performance of containerised data centres. Through the integration of Computational Fluid Dynamics and energy analysis, the study showed that intelligent workload placement reduces hotspot formation, improves temperature uniformity and decreases cooling energy demand while maintaining safe operating conditions. Although the work has been limited to steady-state simulations of a single containerised configuration, it establishes a validated methodology that can be extended to transient operation, intelligent control and climate-specific analyses. Consequently, this thesis provides a strong foundation for the development of future thermal-aware workload scheduling strategies capable of improving the efficiency, reliability and sustainability of modern containerised data centres.

Bibliography

Akbari, A., Khonsari, A. and Ghoreyshi, S.M. 2020. Thermal-Aware Virtual Machine Allocation for Heterogeneous Cloud Data Centers. *Energies*.

Al-Fatlawi, A.A. and Barazanchi, I. Al 2023. A Novel Approach for New Architecture for Green Data Centre. *Bulletin of Electrical Engineering and Informatics*.

Almoli, A., Thompson, A., Kapur, N., Summers, J., Thompson, H. and Hannah, G. 2012. Computational fluid dynamic investigation of liquid rack cooling in data centres. *Applied Energy*.

Almoli, A.M. 2013. *Air flow management inside data centres*. University of Leeds.

Alqudah, M., Ahmed, I., Ahmad, F., Naseem, S. and Nisar, K.S. 2021. Energy Reduction Through Memory Aware Real-Time Scheduling on Virtual Machine in Multi-Cores Server. *Ieee Access*.

American Society of Heating Refrigerating and Air-Conditioning Engineers Inc. (ASHRAE) 2016. ASHRAE TC9.9 Data Center Power Equipment Thermal Guidelines and Best Practices. *Mission Critical Facilities, Data Centers, Technology Spaces, and Electronic Equipment*. (9.9).

Arghode, V.K., Sundaralingam, V., Joshi, Y. and Phelps, W. 2013. Thermal Characteristics of Open and Contained Data Center Cold Aisle. *Journal of Heat Transfer*. **135**(6).

Arroba, P., Risco-Martín, J.L., Moya, J.M. and Ayala, J.L. 2018. Heuristics and metaheuristics for dynamic management of computing and cooling energy in cloud data centers. *Software: Practice and Experience*. **48**(10), pp.1775–1804.

Avgerinou, M.D., Paolo, B. and Luca, C. 2017. Trends in Data Centre Energy Consumption Under the European Code of Conduct for Data Centre Energy Efficiency. *Energies*.

Azimzadeh, A. and Tabrizi, N. 2017. A Dynamic Power Management Schema for Multi-Tier Data Centers.

Barroso, L.A. and Hölzle, U. 2007. The Case for Energy-Proportional Computing. *Computer*.

Barzegar, B., Motameni, H. and Movaghar, A. 2019. EATSDCD: A green energy-aware scheduling algorithm for parallel task-based application using clustering, duplication and DVFS technique in cloud datacenters. *Journal of Intelligent & Fuzzy Systems*. **36**(6), pp.5135–5152.

Belady, C. and Goolsbee, C. 2010. Containerized data centers: Is a box a good fit? *TechTarget*. [Online]. [Accessed 11 December 2018]. Available from: https://searchdatacenter.techtarget.com/feature/Containerized-data-centers-Is-a-box-a-good-fit.

Beloglazov, A., Abawajy, J. and Buyya, R. 2012. Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. *Future Generation Computer Systems*. **28**(5), pp.755–768.

Beloglazov, A. and Buyya, R. 2010. Energy Efficient Resource Management in Virtualized Cloud Data Centers *In*: *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing* [Online]. IEEE, pp.826–831. Available from: http://ieeexplore.ieee.org/document/5493373/.

Berral, J.Ll., Goiri, Í., Nou, R., Julia, F., Guitart, J., Gavaldà, R. and Torres, J. 2010. Towards Energy-Aware Scheduling in Data Centers Using Machine Learning.

Bian, Q. 2020. Waste heat: the dominating root cause of current global warming. *Environmental Systems Research*. **9**(1).
Bianchini, R., Khan, S.U. and Mastroianni, C. 2017. Guest Editors' Introduction: Special Issue on Green and Energy-Efficient Cloud Computing Part II. *Ieee Transactions on Cloud Computing*.
De Boer, G.N., Johns, A., Delbosc, N., Burdett, D., Tatchell-Evans, M., Summers, J. and Baudot, R. 2018. Three computational methods for analysing thermal airflow distributions in the cooling of data centres *In*: *International Journal of Numerical Methods for Heat and Fluid Flow*.
Bouaouda, A., Afdel, K. and Abounacer, R. 2023. Meta-Heuristic and Heuristic Algorithms for Forecasting Workload Placement and Energy Consumption in Cloud Data Centers. *Advances in Science Technology and Engineering Systems Journal*.
Brady, G.A., Kapur, N., Summers, J.L. and Thompson, H.M. 2013. A case study and critical assessment in calculating power usage effectiveness for a data centre. *Energy Conversion and Management*. **76**, pp.155–161.
Brenner, P., Thain, D., Buccellato, A. and Go, D.B. 2012. Environmentally Opportunistic Computing *In*: I. Ahmad and S. Ranka, eds. *Handbook of Energy-Aware and Green Computing - Two Volume Set.* [Online]. Chapman and Hall/CRC, pp.949–964. Available from: http://www.crcnetbase.com/doi/abs/10.1201/b16631-50.
Brenner, P., Thain, D. and Latimer, D. 2009. Grid Heating Clusters: Transforming Cooling Constraints Into Thermal Benefits *In*: *The Uptime Institute Green Enterprise IT Award Paper*.
Brocanelli, M., Zheng, W. and Wang, X. 2014. Reducing the expenses of geo-distributed data centers with portable containerized modules. *Performance Evaluation*.
Buccellato, A.P.C., Brenner, P., Go, D.B., Jansen, R. and Ward, E.M. 2011. Environmentally opportunistic computing: Computation as catalyst for sustainable design *In*: *ASHRAE Transactions*.
Budiyanto, M.A., Shinoda, T. and Nasruddin 2017. Study on the CFD simulation of refrigerated container. *IOP Conference Series: Materials Science and Engineering*. **257**, p.012042.
Chen, Z., Hu, J., Min, G., Zomaya, A.Y. and El-Ghazawi, T. 2020. Towards Accurate Prediction for High-Dimensional and Highly-Variable Cloud Workloads with Deep Learning. *IEEE Transactions on Parallel and Distributed Systems*. **31**(4), pp.923–934.
Choo, K., Galante, R.M. and Ohadi, M.M. 2014. Energy consumption analysis of a medium-size primary data center in an academic campus. *Energy and Buildings*. **76**, pp.414–421.
Chu, W.-X., Hsu, C.-S., Tsui, Y.-Y. and Wang, C.-C. 2019. Experimental investigation on thermal management for small container data center. *Journal of Building Engineering*. **21**, pp.317–327.
Chu, W.X. and Wang, C.C. 2019. A review on airflow management in data centers. *Applied Energy*. **240**, pp.84–119.
Covas, M.T., Silva, C.A. and Dias, L.C. 2013. Multicriteria decision analysis for sustainable data centers location. *International Transactions in Operational Research*.

Crăciun, C. and Salomie, I. 2017. A Filter-Based Dynamic Resource Management Framework for Virtualized Data Centers. *Studia Universitatis Babeș-Bolyai Informatica*. **62**(1), pp.32–48.
Cupertino, L.F., Costa, G. Da, Oleksiak, A., Piątek, W., Pierson, J., Salom, J., Siso, L., Stolf, P., Sun, H. and Zilio, T. 2015. Energy-Efficient, Thermal-Aware Modeling and Simulation of Data Centers: The CoolEmAll Approach and Evaluation Results. *Ad Hoc Networks*.
Damme, T. Van, Persis, C. De and Tesi, P. 2019. Optimized Thermal-Aware Job Scheduling and Control of Data Centers. *Ieee Transactions on Control Systems Technology*.
Daraghmeh, H.M. and Wang, C.-C. 2017. A review of current status of free cooling in datacenters. *Applied Thermal Engineering*. **114**, pp.1224–1239.
Dayarathna, M., Wen, Y. and Fan, R. 2016. Data Center Energy Consumption Modeling: A Survey. *Ieee Communications Surveys & Tutorials*.
Dragseth, R.E. and Hanssen, S.A. 2016. *Direct-to-chip liquid cooling for reducing power consumption in a subarctic supercomputer centre* [Online]. [Accessed 14 September 2018]. Available from: http://weather.cs.uit.no/.
Emerson Network Power 2010. *Data Center Users Group Special Report: Growing Concerns over Data Center Infrastructure Monitoring and Management*. Ohio.
Escher, W., Brunschwiler, T., Shalkevich, N., Shalkevich, A., Burgi, T., Michel, B. and Poulikakos, D. 2011. On the Cooling of Electronics With Nanofluids. *Journal of Heat Transfer*.
F. Alves, D.S. and Kabbani, A. 2022. An Open-Source Simulation Platform for Benchmarking Geo-Distributed Data Center Schedulers.
Fan, G., Chen, L., Yu, H. and Liu, D. 2018. Formally modeling and analyzing cost-aware job scheduling for cloud data center. *Software: Practice and Experience*. **48**(9), pp.1536–1559.
Fatima, A., Javaid, N., Sultana, T., Hussain, W., Bilal, M., Shabbir, S., Asim, Y., Akbar, M. and Ilahi, M. 2018. Virtual Machine Placement via Bin Packing in Cloud Data Centers. *Electronics*.
Fernández-Cerero, D., Fernández-Montes, A. and Velasco, F. 2018. Productive Efficiency of Energy-Aware Data Centers. *Energies*.
Fried, I. 2009. Inside one of the world's largest data centers. *CNet*. [Online]. [Accessed 11 December 2018]. Available from: https://www.cnet.com/news/inside-one-of-the-worlds-largest-data-centers/.
Goel, B., McKee, S.A. and Själander, M. 2012. Techniques to Measure, Model, and Manage Power. *Advances in Computers*. **87**, pp.7–54.
Grange, L., Da Costa, G. and Stolf, P. 2018. Green IT scheduling for data center powered with renewable energy. *Future Generation Computer Systems*. **86**, pp.99–120.
Guo, Y., Gong, Y., Fang, Y., Khargonekar, P.P. and Geng, X. 2014. Energy and Network Aware Workload Management for Sustainable Data Centers with Thermal Storage. *IEEE Transactions on Parallel and Distributed Systems*. **25**(8), pp.2030–2042.
Hammadi, A. and Mhamdi, L. 2014. A Survey on Architectures and Energy Efficiency in Data Center Networks. *Computer Communications*.
Han, R., Zhan, S., Shao, C., Wang, J., John, L.K., Xu, J., Lu, G. and Wang, L. 2016. BigDataBench-MT: A Benchmark Tool for Generating Realistic Mixed Data Center Workloads *In*:, pp.10–21. Available from: http://link.springer.com/10.1007/978-3-319-29006-5_2.

Hankendi, C., Reda, S. and Coskun, A.K. 2013. vCap: Adaptive Power Capping for Virtualized Servers.
Havet, A., Schiavoni, V., Felber, P., Colmant, M., Rouvoy, R. and Fetzer, C. 2017. GENPACK: A generational scheduler for cloud data centers *In*: *Proceedings - 2017 IEEE International Conference on Cloud Engineering, IC2E 2017*. Institute of Electrical and Electronics Engineers Inc., pp.95–104.
Hewlett-Packard Development Company 2011. HP Performance Optimized Datacenter (POD) 240a. , pp.1–10.
Huang, S., Yan, D. and Chen, Y. 2023. An Online Algorithm for Combined Computing Workload and Energy Coordination of Data Centers.
Ilavarasan, E. and Thambidura, P. 2007a. Low Complexity Performance Effective Task Scheduling Algorithm for Heterogeneous Computing Environments. *Journal of Computer Science*. **3**(2), pp.94–103.
Ilavarasan, E. and Thambidura, P. 2007b. Low Complexity Performance Effective Task Scheduling Algorithm for Heterogeneous Computing Environments. *Journal of Computer Science*. **3**(2), pp.94–103.
Im, S., Panigrahi, D., Kell, N. and Shadloo, M. 2018. Online load balancing on related machines *In*: *Proceedings of the Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, pp.978–989.
Islam, R.U., Ruci, X., Hossain, M.S., Andersson, K. and Kor, A.-L. 2019. Capacity Management of Hyperscale Data Centers Using Predictive Modelling. *Energies*.
Iturriaga, S. and Nesmachnow, S. 2016. Scheduling Energy Efficient Data Centers Using Renewable Energy. *Electronics*. **5**(4), p.71.
Jia, Z., Wang, L., Zhan, J., Zhang, L. and Luo, C. 2013a. Characterizing data analysis workloads in data centers *In*: *2013 IEEE International Symposium on Workload Characterization (IISWC)* [Online]. IEEE, pp.66–76. Available from: http://ieeexplore.ieee.org/document/6704671/.
Jia, Z., Wang, L., Zhan, J., Zhang, L. and Luo, C. 2013b. Characterizing data analysis workloads in data centers *In*: *2013 IEEE International Symposium on Workload Characterization (IISWC)* [Online]. IEEE, pp.66–76. Available from: http://ieeexplore.ieee.org/document/6704671/.
Kelefouras, V. and Djemame, K. 2018. Workflow Simulation Aware and Multi-threading Effective Task Scheduling for Heterogeneous Computing *In*: *2018 IEEE 25th International Conference on High Performance Computing (HiPC)* [Online]. IEEE, pp.215–224. [Accessed 30 May 2020]. Available from: https://ieeexplore.ieee.org/document/8638155/.
Khalili, S., Tradat, M.I., Nemati, K., Seymour, M. and Sammakia, B. 2018. Impact of Tile Design on the Thermal Performance of Open and Enclosed Aisles. *Journal of Electronic Packaging*. **140**(1).
Kliazovich, D., Bouvry, P., Granelli, F. and da Fonseca, N.L.S. 2015. Energy Consumption Optimization in Cloud Data Centers *In*: *Cloud Services, Networking, and Management* [Online]. Wiley, pp.191–215. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9781119042655.ch8.
Kozlowicz, J. 2014. Difference Between Modular and Containerized Data Centers. *Green House Data*. [Online]. [Accessed 11 December 2018]. Available from: https://www.greenhousedata.com/blog/whats-the-difference-between-modular-and-containerized-data-centers.
Kulkarni, J. and Li, S. 2018a. Flow-time Optimization For Concurrent Open-Shop and Precedence Constrained Scheduling Models.

Kulkarni, J. and Li, S. 2018b. Flow-time Optimization For Concurrent Open-Shop and Precedence Constrained Scheduling Models.
Kumar, P. and Joshi, Y. 2012. Fundamentals of data center airflow management *In*: *Energy Efficient Thermal Management of Data Centers*.
Lei, H., Wang, R., Zhang, T., Liu, Y. and Zha, Y. 2016. A multi-objective co-evolutionary algorithm for energy-efficient scheduling on a green data center. *Computers & Operations Research*. **75**, pp.103–117.
Li, D., Wang, W., Li, Q. and Cheng, J. 2017. A Comprehensive Evaluation of Scheduling Methods of Virtual Machine Migration for Energy Conservation. *Ieee Systems Journal*.
Li, D., Yu, Y., He, W., Zheng, K. and He, B. 2015. Willow: Saving Data Center Network Energy for Network-Limited Flows. *Ieee Transactions on Parallel and Distributed Systems*.
Li, H., Ding, S., Zhang, P. and Lai, J. 2018. Energy-Efficient Resource Allocation Strategy Based on Task Classification in Data Center.
Li, K. 2016. Improving Multicore Server Performance and Reducing Energy Consumption by Workload Dependent Dynamic Power Management. *Ieee Transactions on Cloud Computing*.
Li, K. 2019. Optimal Power and Performance Management for Heterogeneous and Arbitrary Cloud Servers. *Ieee Access*.
Li, X., Garraghan, P., Jiang, X., Wu, Z. and Xu, J. 2018. Holistic Virtual Machine Scheduling in Cloud Datacenters towards Minimizing Total Energy. *IEEE Transactions on Parallel and Distributed Systems*. **29**(6), pp.1317–1331.
Li, X., Wang, L., Abawajy, J., Qin, X., Pau, G. and You, I. 2020. Data-Intensive Task Scheduling for Heterogeneous Big Data Analytics in IoT System. *Energies*.
Li, Y., Wang, X., Luo, P. and Pan, Q. 2019. Thermal-Aware Hybrid Workload Management in a Green Datacenter towards Renewable Energy Utilization. *Energies*. **12**(8), p.1494.
Liang, S., Yang, Z., Jin, F. and Chen, Y. 2020. Data Centers Job Scheduling with Deep Reinforcement Learning *In*:, pp.906–917.
Ling, Y.Z., Zhang, X.S., Zhang, K. and Jin, X. 2017. On the characteristics of airflow through the perforated tiles for raised-floor data centers. *Journal of Building Engineering*. **10**, pp.60–68.
Liu, H., Xie, T., Ran, J. and Gao, S. 2016. A modified approach for thermal distribution monitoring of the container data center by thermal image registration *In*: *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* [Online]. IEEE, pp.259–263. Available from: http://ieeexplore.ieee.org/document/7852718/.
Liu, X., Hou, G. and Yang, L. 2022. Dynamic Combined Optimal Scheduling of Electric Energy and Natural Gas Energy Consumption in Data Center W. Zhang, ed. *Discrete Dynamics in Nature and Society*. **2022**, pp.1–8.
Liu, Z., Chen, Y., Bash, C., Wierman, A., Gmach, D., Wang, Z., Marwah, M. and Hyser, C. 2012. Renewable and cooling aware workload management for sustainable data centers. *ACM SIGMETRICS Performance Evaluation Review*. **40**(1), pp.175–186.
Lu, Z., Lei, J., He, Y., Li, Z., Deng, S. and Gao, X. 2019. Energy Optimization for Software-Defined Data Center Networks Based on Flow Allocation Strategies. *Electronics 2019, Vol. 8, Page 1014*. **8**(9), p.1014.

Ma, J. and Ma, X. 2023. Distributed Control of HVAC-BESS Under Solar Power Forecasts in Microgrid System. *Ieee Transactions on Industrial Informatics*.
Mahmood, F., Zeeshan Khan, F., Ahmed, M., Ahmad, I. and Gupta, B.B. 2022. GreenCloudNet++: Simulation framework for energy efficient and secure, green job scheduling in geographically distributed data centers. *Transactions on Emerging Telecommunications Technologies*. **33**(4).
Martin, M., Mukesh Khattar PhD, P.E. and Germagian, M. 2007. High-density heat containment. *ASHRAE Journal*. **49**(12), p.38.
Mastelic, T., Oleksiak, A., Claussen, H., Brandic, I., Pierson, J.M. and Vasilakos, A. V. 2014. Cloud Computing. *ACM Computing Surveys (CSUR)*. **47**(2).
Mastroianni, C., Meo, M. and Papuzzo, G. 2013a. Probabilistic Consolidation of Virtual Machines in Self-Organizing Cloud Data Centers. *IEEE Transactions on Cloud Computing*. **1**(2), pp.215–228.
Mastroianni, C., Meo, M. and Papuzzo, G. 2013b. Probabilistic Consolidation of Virtual Machines in Self-Organizing Cloud Data Centers. *IEEE Transactions on Cloud Computing*. **1**(2), pp.215–228.
Maurice, E., Dandres, T., Samson, R., Moghaddam, R.F., Nguyen, K.K., Cheriet, M. and Lemieux, Y. 2014. Modelling of Electricity Mix in Temporal Differentiated Life-Cycle-Assessment to Minimize Carbon Footprint of a Cloud Computing Service.
Mazidi, A., Damghanijazi, E. and Tofighy, S. 2017. An Energy-Efficient Virtual Machine Placement Algorithm Based Service Level Agreement in Cloud Computing Environments. *Circulation in Computer Science*.
McFarlane, R., Milsom, S. and LLC, W. 2023. Data center air quality: The air servers breathe | TechTarget. [Accessed 17 November 2024]. Available from: https://www.techtarget.com/searchdatacenter/tip/Data-center-air-quality-The-air-servers-breathe.
Meisner, D., Sadler, C., Barroso, L.A., Weber, W.-D. and Wenisch, T.F. 2011. Power Management of Online Data-Intensive Services.
Miller, R. 2012. eBay Containers Push Frontiers of Efficiency. *DataCenter Knowledge*. [Online]. [Accessed 11 December 2018]. Available from: https://www.datacenterknowledge.com/archives/2012/02/27/ebay-containers-push-frontiers-of-efficiency.
Miller, R. 2013. Google's Infrastructure Boom Continues: Expansion Ahead in Oregon. *DataCenter Knowledge*. [Online]. [Accessed 11 December 2018]. Available from: https://www.datacenterknowledge.com/archives/2013/04/24/googles-infrastructure-boom-continues-expansion-ahead-in-oregon.
Mondal, J., Kv, P. and Vijayakumar 2017. A Review on Energy Efficient Job Scheduling Algorithms in Green Cloud Computing. *Asian Journal of Pharmaceutical and Clinical Research*.
Moss, S. 2019. Data center and cloud companies need to be more transparent about energy usage - DCD. *Data Centre Dynamics Ltd.* [Online]. [Accessed 9 May 2020]. Available from: https://www.datacenterdynamics.com/en/analysis/data-center-and-cloud-companies-need-be-more-transparent-about-energy-usage/.
Mtech, S.S. and Kamal, S. 2014. *Parallel Job Scheduling in Cloud with Lookahead and Workload Consolidation*.

Ngo, H.Q., Tran, L.-N., Duong, T.Q., Matthaiou, M. and Larsson, E.G. 2018. On the Total Energy Efficiency of Cell-Free Massive MIMO. *Ieee Transactions on Green Communications and Networking*.

Ni, J. and Bai, X. 2017. A review of air conditioning energy performance in data centers. *Renewable and Sustainable Energy Reviews*.

Nonde, L., H. El-Gorashi, T.E. and H. Elmirghani, J.M. 2015. Energy Efficient Virtual Network Embedding for Cloud Networks. *Journal of Lightwave Technology*.

N'Takpe, T., Suter, F. and Casanova, H. 2007. A Comparison of Scheduling Approaches for Mixed-Parallel Applications on Heterogeneous Platforms *In*: *Sixth International Symposium on Parallel and Distributed Computing (ISPDC'07)* [Online]. IEEE, pp.35–35. [Accessed 30 May 2020]. Available from: http://ieeexplore.ieee.org/document/4271925/.

Nystedt, D. 2010. Taiwan Eyes Containerized Data Centers. *Computerworld*., p.6.

Oró, E., Allepuz, R., Martorell, I. and Salom, J. 2018. Design and economic analysis of liquid cooled data centres for waste heat recovery: A case study for an indoor swimming pool. *Sustainable Cities and Society*. **36**, pp.185–203.

Peng, Z., Barzegar, B., Yarahmadi, M., Motameni, H. and Pirouzmand, P. 2020. Energy-Aware Scheduling of Workflow Using a Heuristic Method on Green Cloud. *Scientific Programming*.

Pesch, D., Rea, S., Torrens, J.I., Zavrel, V.V., Hensen, J.J., Grimes, D., O'Sullivan, B., Scherer, T., Birke, R., Chen, L.Y., Engbersen, T., Lopez, L., Pages, E., Mehta, D., Townley, J. and Tsachouridis, V.A. 2017. Globally Optimised Energy-Efficient Data Centres.

Qiu, Y., Jiang, C., Wang, Y., Ou, D., Li, Y. and Wan, J. 2019. Energy Aware Virtual Machine Scheduling in Data Centers. *Energies*. **12**(4), p.646.

Qouneh, A., Li, C. and Li, T. 2011. A quantitative analysis of cooling power in container-based data centers *In*: *2011 IEEE International Symposium on Workload Characterization (IISWC)* [Online]. IEEE, pp.61–71. Available from: http://ieeexplore.ieee.org/document/6114197/.

Rivoire, S., Shah, M.A., Ranganathan, P., Kozyrakis, C. and Meza, J. 2007. Models and metrics to enable energy-efficiency optimizations. *Computer*.

Shaikh, A., Uddin, M., Elmagzoub, M.A. and Alghamdi, A. 2020. PEMC: Power Efficiency Measurement Calculator to Compute Power Efficiency and CO_2 Emissions in Cloud Data Centers. *Ieee Access*.

Sharma, C.S., Tiwari, M.K., Zimmermann, S., Brunschwiler, T., Schlottig, G., Michel, B. and Poulikakos, D. 2015. Energy efficient hotspot-targeted embedded liquid cooling of electronics. *Applied Energy*.

Shehabi, A., Smith, S., Sartor, D., Brown, R., Herrlin, M., Koomey, J., Masanet, E., Horner, N., Azevedo, I. and Lintner, W. 2016. *United States Data Center Energy Usage Report* [Online]. Berkeley, CA (United States). [Accessed 16 May 2020]. Available from: http://www.osti.gov/servlets/purl/1372902/.

Shehabi, A., Smith, S.J., Masanet, E. and Koomey, J. 2018. Data center growth in the United States: decoupling the demand for services from electricity use. *Environmental Research Letters*. **13**(12), p.124030.

Shrivastava, S.K., Calder, A.R. and Ibrahim, M. 2012. Quantitative comparison of air containment systems *In*: *InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, ITHERM*.

Song, Z., Zhang, X. and Eriksson, C. 2015. Data Center Energy and Cost Saving Evaluation *In*: *Energy Procedia*.

Sui, X., Liu, D., Li, L., Wang, H. and Yang, H. 2019. Virtual machine scheduling strategy based on machine learning algorithms for load balancing. *Eurasip Journal on Wireless Communications and Networking*. **2019**(1).

Sun, C. 2018. Eco-Friendly Power Cost Minimization for Geo-Distributed Data Centers Considering Workload Scheduling.

Sun, K., Luo, N., Luo, X. and Hong, T. 2021. Prototype Energy Models for Data Centers. *Energy and Buildings*.

Sun Microsystems inc 2009. Sun Modular Datacenter - Technical Specifications. *http://sun.com/products/sunmd/s20/specifications.jsp*. [Online]. [Accessed 13 December 2024]. Available from: https://web.archive.org/web/20090226142103/http://www.sun.com/products/sunmd/s20/specifications.jsp.

Suresh, T. and Murugan, A. 2018. Cold LOGIK and RDHX Solution for Data Center Energy Optimization. *International Journal of Engineering & Technology*.

Szmajduch, M. and Kolodziej, J. 2015. Data-aware Scheduling in Massive Heterogeneous Systems *In*: *Proceedings, 29th European Conference on Modelling and Simulation, ECMS 2015 : May 26th - May 29th, 2015, Albena (Varna), Bulgaria*. European Council for Modelling and Simulation, p.700.

Tang, Q.-Z., Gupta, S. and Varsamopoulos, G. 2007. Thermal-Aware Task Scheduling for Data Centers Through Minimizing Heat Recirculation.

Tang, S., Huang, Q., Li, X. and Wu, D. 2013. Smoothing the Energy Consumption: Peak Demand Reduction in Smart Grid.

Tim Yuan, T., Xu, C., Huang, T. and Li, J. 2018. HOLMES : Holistic Mice-Elephants Stochastic Scheduling in Data Center Networks *In*: *Computer Science & Information Technology* [Online]. Academy & Industry Research Collaboration Center (AIRCC), pp.01–21. Available from: https://airccj.org/CSCP/vol8/csit88901.pdf.

Varsamopoulos, G., Banerjee, A. and Gupta, S. 2009. Energy Efficiency of Thermal-Aware Job Scheduling Algorithms Under Various Cooling Models.

Waldrop, M.M. 2007. Data Center in a Box. *Scientific American*. **297**(2), pp.90–93.

Wang, C.-H., Tsui, Y.-Y. and Wang, C.-C. 2017. Airflow Management on the Efficiency Index of a Container Data Center Having Overhead Air Supply. *Journal of Electronic Packaging*. **139**(4).

Wang, C.-H., Tsui, Y.-Y. and Wang, C.-C. 2017a. On cold-aisle containment of a container datacenter. *Applied Thermal Engineering*. **112**, pp.133–142.

Wang, D., Xie, C., Wu, R., Lai, C.S., Li, X., Zhao, Z., Wu, X., Xu, Y., Lai, L.L. and Wei, J. 2021. Optimal Energy Scheduling for Data Center With Energy Nets Including CCHP and Demand Response. *IEEE Access*. **9**, pp.6137–6151.

Wang, H., Huang, J., Lin, X. and Mohsenian-Rad, H. 2016. Proactive Demand Response for Data Centers: A Win-Win Solution. *Ieee Transactions on Smart Grid*.

Wang, L., Zhang, F., Vasilakos, A. V, Hou, C. and Li, Y. 2014. Joint Virtual Machine Assignment and Traffic Engineering for Green Data Center Networks. *Acm Sigmetrics Performance Evaluation Review*.

Ward, E.M., Goedke, M.L., Brenner, P. and Go, D.B. 2012. A simplified thermodynamic model for waste heat utilization from a containerized data center experimental platform *In*: *InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, ITHERM*.

Wong, D. 2016. Peak Efficiency Aware Scheduling for Highly Energy Proportional Servers. *Acm Sigarch Computer Architecture News*.
Wood, L. 2019. Data Center Market in US - Industry Outlook and Forecast 2018-2019 & 2024: Market is Expected to Reach Revenues of Over $69 Billion. *PR Newswire*. [Online]. [Accessed 13 May 2020]. Available from: https://www.prnewswire.com/news-releases/data-center-market-in-us---industry-outlook-and-forecast-2018-2019--2024-market-is-expected-to-reach-revenues-of-over-69-billion-300926800.html.
Xiong, A. and Xu, C. 2014. Energy Efficient Multiresource Allocation of Virtual Machine Based on PSO in Cloud Data Center. *Mathematical Problems in Engineering*.
Yan, B., Gu, L. and Qi, X. 2018. Comparative Study of Energy Performance Between Chip and Inlet Temperature-Aware Workload Allocation in Air-Cooled Data Center. *Energies*.
Yang, T.J., Chen, Y.H. and Sze, V. 2017. Designing energy-efficient convolutional neural networks using energy-aware pruning. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. **2017-January**, pp.6071–6079.
Yang, Y., He, C., Yin, B., Wei, Z., and, B.H.-K.T. on I. and 2022, undefined n.d. Cloud Task Scheduling Based on Proximal Policy Optimization Algorithm for Lowering Energy Consumption of Data Center. *itiis.orgY Yang, C He, B Yin, Z Wei, B HongKSII Transactions on Internet and Information Systems (TIIS), 2022•itiis.org*.
Yao, L. and Huang, J. 2019. Multi-Objective Optimization of Energy Saving Control for Air Conditioning System in Data Center. *Energies*.
Yin, C., Liu, J. and Jin, S. 2020. A Virtualized Data Center Energy-saving Mechanism Based on Switching Operating Mode of Physical Servers and Reserving Virtual Machines. *Concurrency and Computation Practice and Experience*.
Yuan, C. and Sun, X. 2019. Server Consolidation Based on Culture Multiple-Ant-Colony Algorithm in Cloud Computing. *Sensors 2019, Vol. 19, Page 2724*. **19**(12), p.2724.
Zhan, J., Zhang, L., Sun, N., Wang, L., Jia, Z. and Luo, C. 2012. High Volume Throughput Computing: Identifying and Characterizing Throughput Oriented Workloads in Data Centers *In*: *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum* [Online]. IEEE, pp.1712–1721. [Accessed 30 October 2024]. Available from: http://ieeexplore.ieee.org/document/6270846/.
ZHANG, M.-H., REN, H.-L. and XIA, C.-H. 2017. Research of Dynamic Scheduling of Resources Based on Virtual Machine Migration in Cloud Environment. *DEStech Transactions on Engineering and Technology Research*. (iceeac).
Zheng, W., Ma, K. and Wang, X. 2015. TE-Shave: Reducing Data Center Capital and Operating Expenses with Thermal Energy Storage. *IEEE Transactions on Computers*.
Zhou, Q., Xu, M., Gill, S.S., Gao, C., Tian, W., Xu, C.-Z. and Buyya, R. 2020. Energy Efficient Algorithms Based on VM Consolidation for Cloud Computing: Comparisons and Evaluations.
Zhou, R., Shi, Y. and Zhu, C. 2013. AxPUE: Application Level Metrics for Power Usage Effectiveness in Data Centers.

Zhou, Z., Hu, Z. and Li, K. 2016. Virtual Machine Placement Algorithm for Both Energy-Awareness and SLA Violation Reduction in Cloud Data Centers. *Scientific Programming*.
Zong, Y., Ou, Y., Hammad, A., Kondepu, K., Nejabati, R., Simeonidou, D., Liu, Y. and Guo, L. 2018. Location-Aware Energy Efficient Virtual Network Embedding in Software-Defined Optical Data Center Networks. *Journal of Optical Communications and Networking*.

Appendix A

Contains a table divided in several sections with the raw data obtained from ANSYS Fluent. This information helps the reader to reproduce the results for each scenario explored in this research.

Name	flowrate	inlet_temp	block_11	block_12	block_13	block_14	block_15
DP 0	3.52	294.15	1000	1000	1000	1000	1000
DP 2	2.83	294.15	1000	1000	1000	1000	1000
DP 4	4	294.15	1000	1000	1000	1000	1000
DP 1	3.52	294.15	0	0	0	0	0
DP 3	2.83	294.15	0	0	0	0	0
DP 5	4	294.15	0	0	0	0	0
DP 6	3.52	294.15	0	0	0	0	0
DP 7	4	294.15	0	0	0	0	0
DP 9	4	294.15	0	0	0	0	0
DP 8	3.52	294.15	0	0	0	0	0
DP 10	2.83	294.15	0	0	0	0	0
DP 12	4	294.15	0	0	0	0	0
DP 11	3.52	294.15	0	0	0	0	0
DP 13	2.83	294.15	0	0	0	0	0
DP 15	4	294.15	0	0	500	0	0
DP 14	3.52	294.15	0	0	500	0	0
DP 16	2.83	294.15	0	0	500	0	0
Name	flowrate	inlet_temp	block_21	block_22	block_23	block_24	block_25
DP 0	3.52	294.15	0	0	0	0	0
DP 2	2.83	294.15	0	0	0	0	0
DP 4	4	294.15	0	0	0	0	0
DP 1	3.52	294.15	0	0	0	0	0
DP 3	2.83	294.15	0	0	0	0	0
DP 5	4	294.15	0	0	0	0	0
DP 6	3.52	294.15	500	500	500	500	500
DP 7	4	294.15	500	500	500	500	500
DP 9	4	294.15	0	0	0	0	0
DP 8	3.52	294.15	0	0	0	0	0
DP 10	2.83	294.15	0	0	0	0	0
DP 12	4	294.15	0	500	0	500	0
DP 11	3.52	294.15	0	500	0	500	0
DP 13	2.83	294.15	0	500	0	500	0
DP 15	4	294.15	0	500	0	500	0
DP 14	3.52	294.15	0	500	0	500	0
DP 16	2.83	294.15	0	500	0	500	0

Name	flowrate	inlet_temp	block_31	block_32	block_33	block_34	block_35
DP 0	3.52	294.15	1000	1000	1000	1000	1000
DP 2	2.83	294.15	1000	1000	1000	1000	1000
DP 4	4	294.15	1000	1000	1000	1000	1000
DP 1	3.52	294.15	1000	1000	1000	1000	1000
DP 3	2.83	294.15	1000	1000	1000	1000	1000
DP 5	4	294.15	1000	1000	1000	1000	1000
DP 6	3.52	294.15	500	500	500	500	500
DP 7	4	294.15	500	500	500	500	500
DP 9	4	294.15	500	0	250	0	500
DP 8	3.52	294.15	500	0	250	0	500
DP 10	2.83	294.15	500	0	250	0	500
DP 12	4	294.15	500	0	250	0	500
DP 11	3.52	294.15	500	0	250	0	500
DP 13	2.83	294.15	500	0	250	0	500
DP 15	4	294.15	500	0	250	0	500
DP 14	3.52	294.15	500	0	250	0	500
DP 16	2.83	294.15	500	0	250	0	500
Name	flowrate	inlet_temp	block_41	block_42	block_43	block_44	block_45
DP 0	3.52	294.15	0	0	0	0	0
DP 2	2.83	294.15	0	0	0	0	0
DP 4	4	294.15	0	0	0	0	0
DP 1	3.52	294.15	1000	1000	1000	1000	1000
DP 3	2.83	294.15	1000	1000	1000	1000	1000
DP 5	4	294.15	1000	1000	1000	1000	1000
DP 6	3.52	294.15	1000	1000	1000	1000	1000
DP 7	4	294.15	1000	1000	1000	1000	1000
DP 9	4	294.15	0	0	0	0	0
DP 8	3.52	294.15	0	0	0	0	0
DP 10	2.83	294.15	0	0	0	0	0
DP 12	4	294.15	0	500	0	500	0
DP 11	3.52	294.15	0	500	0	500	0
DP 13	2.83	294.15	0	500	0	500	0
DP 15	4	294.15	0	500	0	500	0
DP 14	3.52	294.15	0	500	0	500	0
DP 16	2.83	294.15	0	500	0	500	0

Name	flowrate	inlet_temp	block_51	block_52	block_53	block_54	block_55
DP 0	3.52	294.15	1000	1000	1000	1000	1000
DP 2	2.83	294.15	1000	1000	1000	1000	1000
DP 4	4	294.15	1000	1000	1000	1000	1000
DP 1	3.52	294.15	1000	1000	1000	1000	1000
DP 3	2.83	294.15	1000	1000	1000	1000	1000
DP 5	4	294.15	1000	1000	1000	1000	1000
DP 6	3.52	294.15	1000	1000	1000	1000	1000
DP 7	4	294.15	1000	1000	1000	1000	1000
DP 9	4	294.15	1000	1000	1000	1000	1000
DP 8	3.52	294.15	1000	1000	1000	1000	1000
DP 10	2.83	294.15	1000	1000	1000	1000	1000
DP 12	4	294.15	1000	0	1000	0	1000
DP 11	3.52	294.15	1000	0	1000	0	1000
DP 13	2.83	294.15	1000	0	1000	0	1000
DP 15	4	294.15	1000	0	500	0	1000
DP 14	3.52	294.15	1000	0	500	0	1000
DP 16	2.83	294.15	1000	0	500	0	1000

Name	flowrate	inlet_temp	avg_outlet_temp	max_temp_cold_aisle	max_temp_hot_aisle	max_temp_mid_blocks
DP 0	3.52	294.15	300.62	303.74	326.98	325.12
DP 2	2.83	294.15	302.35	307.92	337.97	336.07
DP 4	4	294.15	299.81	302.07	321.57	321.88
DP 1	3.52	294.15	300.35	307.45	331.15	330.56
DP 3	2.83	294.15	302.60	313.08	338.75	339.50
DP 5	4	294.15	299.68	305.27	325.81	326.99
DP 6	3.52	294.15	300.36	306.84	325.29	329.19
DP 7	4	294.15	299.67	305.20	323.08	325.37
DP 9	4	294.15	296.48	300.74	314.87	319.33
DP 8	3.52	294.15	296.81	301.66	317.30	321.59
DP 10	2.83	294.15	297.67	303.51	322.21	327.62
DP 12	4	294.15	296.47	300.49	312.15	318.68
DP 11	3.52	294.15	296.83	301.45	317.01	321.29
DP 13	2.83	294.15	297.64	303.26	321.96	326.02
DP 15	4	294.15	296.49	300.49	311.54	318.67
DP 14	3.52	294.15	296.86	301.44	313.05	321.00
DP 16	2.83	294.15	297.62	303.25	316.21	325.45