Data Mining for Thermal Analysis of Big Dataset of HPC-Data Center

Davide De Chiara[1], Marta Chinnici[2], Ah-Lian Kor[3]

[1] ENEA-ICT Division, C.R. Portici Piazzale Enrico Fermi 1, Napoli 80055, Italy
davide.dechiara@enea.it
[2] ENEA-ICT Division, C.R Casaccia Via Anguillarese 301, ROMA 00123, Italy
marta.chinnici@enea.it
[3] Leeds Beckett University, Leeds, UK
a.kor@leedsbeckett.ac.uk

Abstract. Greening of Data Centers could be achieved through energy savings in two major areas namely: compute systems and cooling systems. A reliable cooling system is necessary to produce a persistent flow of cold air to cool the servers due to increasingly demanding computational load. Servers' dissipated heat effects a strain on the cooling systems. Consequently, it is imperative to individual servers that frequently occur in the hotspot zones. This is facilitated through the application of data mining techniques to an available big data set with thermal characteristics of HPC-ENEA-Data Center, namely Cresco 6. This work involves the implementation of an advanced algorithm on the workload management platform produces hotspots maps with the goal to reduce data centre wide thermal-gradient, and cooling effectiveness.

Keywords: Data Center, HPC, Data Mining, Big Data, Thermal, Hotspot, Cooling, Thermal management.

1 Introduction

A large volume of electricity is generated worldwide through the burning of hydrocarbons. This causes a rise in carbon emission and other Green House Gasses (GHG) in the environment, contributing to global warming. Data Center (DC) worldwide was estimated to have consumed between 203 to 271 billion kWh of electricity in the year 2010 [1]. According to [2], unless steps are taken to save energy and go-green, global DC share of carbon emission is estimated to rise from 307 million tons in 2007 to 358 million tons in 2020. Servers consume energy which is eventually converted to heat, and is proportional to the allocated computing loads. Cooling machines are deployed to maintain the computing servers at the vendor specified temperature for consistent and reliable performance. Koomey [1] showed that electricity is primarily spent on cooling and computer systems (comprising servers in chassis and racks). Hence, these two systems have been critical focal points for energy savings. Computing-load processing entails jobs and IT tasks management. DC cooling encompasses the installation of cooling systems and involvement of hot/cold aisle configurations. Incompetent thermal management is the underlying cause of inefficiency of the IT infrastructure within a data center. Microprocessors in servers are the primary electricity consumers and heat generators [4]. Generally, the vast amount of waste heat generated a high performance-oriented microprocessor architecture is beyond efficient air cooling capability. Thus, it is necessary to efficiently disperse the waste heat to avoid overheating. All this while, DCs have been spending an equal amount of electricity on cooling and computing [1]. A more effective energy savings strategy would be to reduce electricity consumption by minimising the load on the cooling system for keeping servers (in the computer system) cool. Thermal-aware scheduling is a computational workload scheduling based on waste heat. Thermal-aware schedulers adopt different thermal-aware approaches. Heat modelling provides a model that links power consumed by servers and their associated waste heat. Thermal-aware monitoring acts as a thermal-eye for the scheduling process and entails recording and evaluation of heat distribution within DCs. Thermal profiling is based on useful monitoring information on workload-related heat emission and it is used to predict the DC heat distribution. Our analysis explores the relationship between thermal-aware scheduling and computer workload scheduling. This is followed by selecting an efficient solution to evenly distribute heat within a DC to avoid hotspots and cold spots.In this work, a data mining technique is chosen for a deeper analysis of hotspots prediction and thermal-profiling to avoid them. The dataset employed for this analysis is a big dataset of ENEA-HPC-CRESCO6 nodes. The analysis encompasses the following: hotspots localisation; categorisation of users according to the job that usually submit on CRESCO6 cluster; categorisation of nodes behaviour based on internal and surrounding air temperatures with a cause-effect study on heat produced by certain workloads This analysis aims to minimise thermal gradient taking into account different level of available thermal data consumption such as nodes, CPU, IT room environment. Unsupervised learning has been employed due to the variability of thermal data and uncertainties in defining temperature thresholds for hotspots identified. In this analysis phase, the optimal workload distribution to clusters nodes is determined and available thermal characteristics (i.e. exhaust temperature, CPUs temperatures) are inputs to the clustering algorithm. Subsequently, a series of clustering results are intersected to unravel nodes (identified by IDs) that frequently fall into high-temperature areas of the cluster racks. The paper is organised as follows: Section I – Introduction; Section II – Background: Related Works; Section III – Methodology; Section IV – Analysis and Results; Section V – Conclusions and Future Works.

2 Background: Related Work

In the context of HPC-DC operations and management, energy efficiency comprises cooling and IT equipment utilisation optimised to maintain recommendable IT room conditions and to satisfy service level agreements with minimal energy consumption. Pursuing DC energy efficiency is a challenging task due to a large number of factors affecting DC productivity and energy efficiency. A trade-off between colder locations for the free air-cooling and sunny places for solar power plants is an issue yet to be critically analysed [8]. Another challenge concerns thermal equipment: raising the setpoint of cooling equipment or lowering the speed of CRAC (Computer Room Air Conditioning) fans to save energy used by thermal equipment may in the long-term, decrease the IT systems' reliability and thus, a balance is yet to be found [8]. Furthermore, an ongoing challenge of power overprovisioning and causing energy waste for idle servers has brought about research works on energy storage in UPS (Uninterruptible Power Supply), optimal allocation of PDUs (Power Distribution Units) with respect to servers, and multi-step algorithms for power monitoring and on-demand provisioning reviewed in [8]. Other challenges encompass workload management, network-level issues as optimal routing, VM allocation, balance between power savings and network QoS (Quality of Service) parameters as well as choice of appropriate metrics for DC evaluation. One standard metric used by a majority of industrial DCs is Power Usage Effectiveness (PUE) proposed by Green Grid Consortium [2]. It shows the ratio of total DC energy utilisation with respect to the energy consumed solely by IT equipment. A plethora of metrics currently under development evaluates thermal characteristics, a ratio of renewable energy use, energy productivity of various components and other parameters. DCs experience an urgent need for a holistic framework that would thoroughly characterise them with a fixed set of metrics and find potential pitfalls in their operation. Although such attempts have been found in existing research work, thus far, no framework has been standardised [9]–[13]. Viewing the fact that IT is the major energy consumers within a DC, its thermal characteristics should be the primary focus of an energy efficiency framework. To address this, researchers have proposed various methods for reduce the thermal production in a DC. Sungkap et al. [11] proposed an ambient temperature-aware capping to maximize power efficiency while minimising overheating. Authors analysed the composition of the energy consumption of a cloud computing DC. They put forward that the energy consumption of the DC is composed of about 45% of the computing energy consumption and 40% of the air conditioning refrigeration energy consumption. The remaining 15% of energy consumption is mainly consumed by storage and power distribution systems. This means that about half of the energy consumption of the DC is consumed by non-computing devices. In [6], Wang et al. put forward an analytical model describing DC resources with heat transfer properties and workloads with thermal features. Thermal modelling and temperature estimation from thermal sensors should consider that the increase in inlet air temperature may cause some servers hotspot states and thermal solicitation. This could be attributed to an inappropriate positioning of a rack or even inadequate room ventilation. This emerging problem is unravelled by thermal-aware location analysis. Thermal-aware server provisioning approach to minimise the total power consumption of DC calculates the value power taking into consideration of the maximum working temperature of the servers. Such calculation should also consider that the inlet temperature rise may cause the servers to reach the maximum temperature and cause thermal stress and severe damage in the long run. The thermal-aware scheduling types were identified as reactive, proactive and mixed. However, there was no mention of heat-modelling or thermal-monitoring and profiling. Kong et al. [4] discussed the concepts of thermal-aware profiling, thermal-aware monitoring and thermal-aware scheduling. Thermal-aware techniques were linked to the minimisation of heat production, heat convection to adjacent cores, task migrations, thermal-gradient across the microprocessor chip and power consumption in microprocessors. Microprocessor dynamic thermal management (DTM) techniques encompasses the following techniques: Dynamic Voltage, and Frequency Scaling (DVFS), Clock gating and task migration and Operating System (OS) based DTM and scheduling. In [5], Parolini et al. proposed a heat model and a brief overview of power and thermal efficiency that progresses from microprocessors to DCs. Due to the reasons discussed, it is essential for the energy efficiency of DCs to include the thermal awareness to provide insights into the relationship between the thermal part and the IT part in terms of workload management. In this work, the authors incorporate thermal-aware scheduling, heat modelling, thermal-aware monitoring and thermal profiling using a big thermal dataset of a HPC-Data Center. This research involves quantification, measurement, and analysis of compute nodes and refrigerating machines. The aim of the analysis is to uncover underlying causes that brings about temperatures rise that leads to the creation of thermal hotspots.

Overall, effective DC management requires energy use monitoring, particularly, energy input, IT energy consumption, metering of supply air temperature and humidity at room level, metering of air temperature as well as more granular metering at CRAC/CRAH unit level. Measurements taken should be further analysed to reveal energy use and economisation levels for the improvement of DC energy efficiency level. DC efficiency metrics will not be discussed in this paper. However, in the ensuing section will primarily mainly focus on thermal guidelines from ASHRAE [7].

3 Methodology

Our research goal is to reduce DC wide thermal-gradient, hotspots and maximise cooling effects. This would entail the identification of individual nodes of the server that frequently occur in the hotspot zones through the implementation of an advanced algorithm on workload management platform. The big dataset on thermal characteristics of ENEA Portici CRESCO6 computing cluster is employed for the analysis. It has 24 variables (or features) – Table 1 - and comprises measurements for the period from May 2018 to January 2020. Briefly, the cluster CRESCO6 is a High-Performance Computing System (HPC) consisting of 434 nodes for a total of 20832 cores. It is based on Lenovo Think System SD530 platform, an ultra-dense and economical two-socket server in a 0.5 U rack form factor inserted in a 2U four-mode enclosure. Each node is equipped with: 2 Intel Xeon Platinum 8160 CPUs, each with 24 cores with a clock frequency of 2.1 GHz; A RAM of 192 GB, corresponding to 4 GB/core; A low-latency Intel Omni-Path 100 Series Single-port PCIe 3.0 x16 HFA network interface. The nodes are interconnected by an Intel Omni-Path network with 21 Intel Edge switches 100 series of 48 ports each, bandwidth equal to 100 GB/s, latency equal to 100ns. The connections between the nodes have 2 tier 2:1 no-blocking tapered fat-tree topology. The consumption of electrical power during massive computing workloads amounts to 190 kW.

Table 1. Thermal dataset – description of features.

Node Name	server ID, integer from 1 to 434;
Timestamp	timestamp of a measurement;
System, CPU, Memory Power	one server instantaneous system, memory, CPU power use in three corresponding columns, W;
Fan 1a, Fan1b, …, Fan 5a, Fan 5b	speed of a cooling fan installed in the node, RPM;
System, CPU, Memory, I/O utilisation	ratio of component utilisation, %, missing data;
Inlet, CPU1, CPU2, Exhaust temperature	temperature at the front, inside (CPU1 and CPU2) and at the rear of every node;
SysAirFlow	speed of air traversing the node, CFM;
DC Energy	total energy that the server has used by the corresponding timestamp, kWh

3.1 Saving Energy Approach

This work incorporates thermal-aware scheduling, heat modelling and thermal-aware monitoring and thermal as well as user profiling. Thermal-aware DC scheduling could be designed based on results of data analytics conducted on real data that obtained from running cluster nodes in a real physical DC. This work is based on approximately 20 months of data collection, which include data relating to the parameters of each node, environmental parameters that measure temperatures and humidity in both the hot and cold aisles, parameters that concern cooling machines and finally, the data about individual users who use cluster nodes for their work. Data for each node relates to energy consumption of CPU, RAM, memories and internal temperatures of each node. This research focuses on the effect of dynamic workload assignment on energy consumption of both the computing and cooling systems, as well as performance. The constraint is that each arrived job must be assigned irrevocably to a server and without any information about future arriving jobs. Once the job has been assigned, no pre-emption or migration is allowed, which is typically assumed for HPC applications since they tend to incur a high cost in terms of data reallocation. In this research, we particularly look for an optimised mapping of nodes that needs to be physically and statically placed in advance to one of the available rack slots in the DC forming a matrix made of computing units with specific characteristics and certain availability of resources in a given time t. The idea is to create a list of candidate nodes to deliver “calculation performance” required by a user’s job. In choosing the candidate nodes, the job-scheduler will evaluate, among the thermally cooler nodes, which at the instant t, possess the appropriate characteristics to satisfy the calculation requested by a user. To improve and facilitate the choices made by the scheduler, it is essential to try to understand in advance, what type of work will be required from the node/s by a user. Based on this fundamental observation, starting from the study of several years' worth of data and through machine learning algorithms, we code user profiles into 4 large macro-categories:

1. CPU_intensive
2. MEMORY_intensive
3. CPU&MEMORY_intensive
4. CPU&MEMORY_not intensive

This behavioural categorisation will provide an opportunity to save energy and better allocate tasks on cluster nodes to affect even temperatures across the cluster and reduce AC costs in DCs. Though the primary aim is to reduce the overall node temperatures, it is essential to better distribute the jobs to avoid thermal hotspots, cold spots and averaging the temperature of the calculation nodes in a DC.

3.2 Users and workload understanding: profiled log

Based on thermal data, it is necessary to better understand in depth what users do and how they manage to solicit the calculation nodes with their jobs.The three main objectives of understanding users' behaviour are as follows: Identify parameters based on the diversity of submitted tasks to describe user behaviour; Analyse the predictability of various resources (CPU, Memory) and identify the presence of time patterns in the usage of various resource types; Implement models for prediction of future CPU and memory usage based on historical data carry out Load Sharing Facility (LSF) platform which provide accounting. Identification of behavioural patterns in the task submission and resource consumption to is necessary to predict future resource requirements. This is exceptionally vital for dynamic resource provisioning in a DC. Gathered profiling information can be utilised to categorise each job task based on the previous 4 macro categories: 1) CPU-intensive, 2) disk-intensive, 3) both CPU and MEMORY- intensive, or 4) neither CPU- nor MEMORY -intensive. We set a utilisation threshold, beyond or below which an application can be marked as one of the job types. For instance, if the CPU load is high (e.g., larger than 90%) during almost 60% of the job running time for an application, then the job can be labelled as a CPU-intensive one. The idea of job-scheduler optimises task scheduling when a job with the same AppID or same username is submitted to a cluster again. In case of a match with the previous AppID or username, utilisation stats from the profiled log are retrieved. Based on the utilisation patterns, this user/application is categorised as one of these: CPU-intensive, memory-intensive, neither CPU or memory intensive, and both CPU- and memory-intensive ones. Once the tasks of a job are categorised, search for a node where CPU nor disk load is sufficiently light to handle the next task. A task with high CPU and memory requirement will not be immediately processed until the node temperatures are well under a safe temperature threshold. Node temperature refers to the difference between the temperature of the outlet air and the temperature entering the node (it generally corresponds to the air temperature in the aisles cooled by the air conditioners).

3.3 Real-time workload management based on Thermal awareness: Cluster evaluation

It is necessary to have a snapshot - with all thermal parameters such as temperatures of each component inside the calculation nodes - the cluster to allow the job-scheduler to allocate jobs in efficiency manner. Generally, a snapshot is obtained by directly interrogating the nodes and all the sensors installed near the DC or inside the calculation nodes. For each individual node, the temperatures of the CPUs, memories, instantaneous consumption and the speed of the cooling fans are evaluated. Obviously from these data, the amount of effort that the calculation node is making can be obtained on the fly. Undoubtedly, the parameter with more weight is the difference between the temperature of the air entering the node and that which exists. A very marked difference provides evidence that the node is very busy (jobs that require a lot of CPU or memory). Therefore, for each calculation node, every useful data is detected on the fly, and the data is virtually stored in a matrix that represents the entire cluster. In each cell of the matrix, there is the node corresponding to that position. The general idea is to include the parameters of the nodes among the selection criteria in the scheduling algorithm, in order to allow the allocation of new jobs in the most recent and distant nodes. In this way, the heat is distributed evenly over the entire "matrix" of calculation nodes and the creation of hotspots is significantly reduced. Obviously, user profiling is equally important since user profiles provide insights into user consumption patterns and also the type of job that will be run and other type of associated parameters. For example, if we know that a user will perform CPU-intensive jobs for 24 hours, we will allocate the job in a "cell" (calculation node) or a group of cells (when the number of resources requires many calculation nodes) far away among them, with an antipodal choice. The main goal is to share the cooling load by spreading high-density nodes around. This will help to minimise DC hot spots and ensure efficient cooling without additional costs.

4 Results and Discussion

As previously discussed, we have extracted information on the users' behaviour. Undeniably, classification based on the listed 4 categories improves every time a user proposes jobs to the cluster. This implies that the outcome of a user's classification is not permanent, and if for example, a user has been classified as "CPU intensive" and for a certain period, the user's work is no longer CPU intensive, the user will certainly be placed in another category. With our scheduler in place, we aim to reduce the overall CPU/memory temperatures in general, and outlet temperatures of cluster nodes in particular. The following design principles enable us to design and implement our schedulers: 1)Job categories. Classify tasks into 1 of these 4 categories: CPU-intensive, memory-intensive, neither CPU nor memory-intensive, and both CPU- and memory-intensive tasks; 2) monitoring utilisation. Monitoring CPU and memory utilisation while making scheduling decisions;3)Controlling redline temperature. Operating CPUs and memory under threshold temperatures always updated and improved over time; 4) Maintaining average temperatures. Keeping track of average CPU and memory temperatures in a node; managing an average outlet temperature across a cluster. We consider maintaining a log profile of both CPU and memory usage for every job that has been processed on the cluster, thereby, making it possible to categorise the users into CPU-intensive, memory-intensive, either CPU or memory intensive, and both CPU- and memory-intensive ones. Every user recorded in a log file could have the following attributes: (1) user ID; (2) Application identification; (3) the number of tasks submitted; (4) CPU utilization; (5) memory utilisation.

4.1 Data Analysis: Results & Discussion

A list of important terms used in the thermal management formulation of a DC and throughout these manuscripts is as follows: 1)CPU-intensive. Applications that spend a vast majority of time doing computations; 2)Memory-intensive. A significant portion of these applications is spent in processing RAM and disk operations;3)Max (redline) temperature. The maximum allowed operational temperature specified by a device manufacturer or a system administrator. 4)Inlet temperature. The temperature of the air flowing into the data node (temperature of the air sucked in from the front of the node). 5)Outlet temperature. The temperature of the air coming out from a node (the temperature of the air extracted from the back of the node). By applying these evaluation criteria, we have built an automated procedure that provides insight into the 4 user associated categories (based on present and historical data. Obviously, the algorithm will always make a comparison between a job just submitted by the user and the time series (if any) of the same user. If the application launched or the type of job remains the same, then the user will be grouped into one of the 4 categories. During each job execution, the procedure records the temperature variations of the CPUs and memories at pre-established time intervals. Finally, it continuously improves in the appreciation of a user, particularly, the length of time a user keeps the job on average. In doing so, we get a distinct picture of each user, with a reasonably reliable estimate of the type of work that the calculation node will perform and for how long it will do it. This information will be beneficial for the job scheduler which will be able to better place a job in the ideal array of calculation nodes of the cluster. A preliminary study was conducted on the functioning of the clusters by carrying out a series of experiments. For months, we have observed the various temperature and power consumption responses of the nodes that were subjected to workloads (figure1,2,3).

Figure 1. The representative shape of Power profile' portion on average for all available nodes consumption dataset for a subset of 200days.

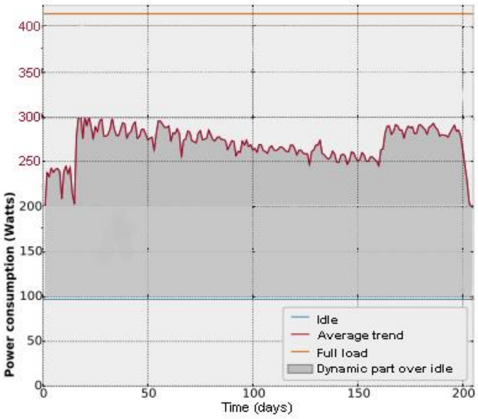


Figure 2. Temperature profiles (subset of 1 month) on average for all available nodes. Nodes are sorted in the order of exhaust temperature increase.

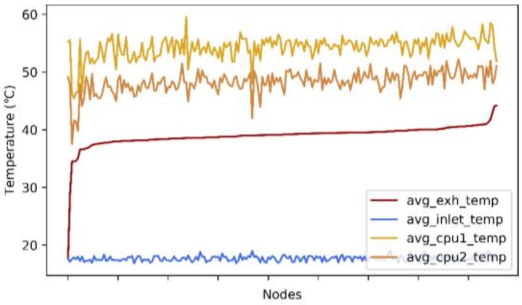


Figure 3. Energy partioning parts on average for all nodes of cluster CRESCO6.

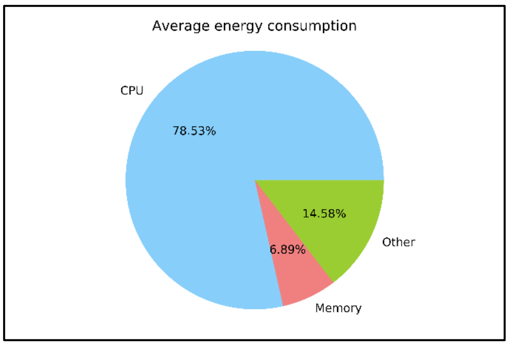
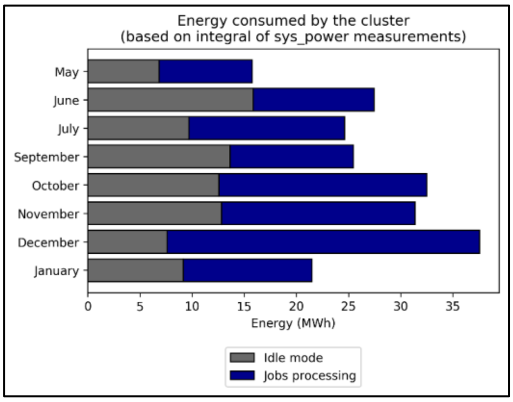
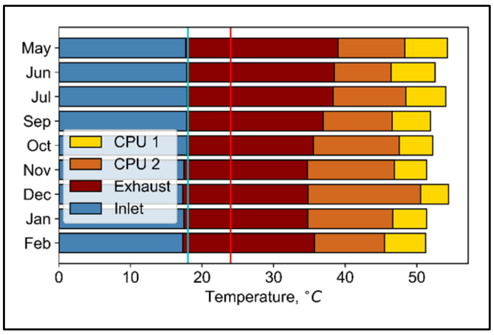


Figure 4. Energy consumption in idle and active nodes (subset of 8 months).



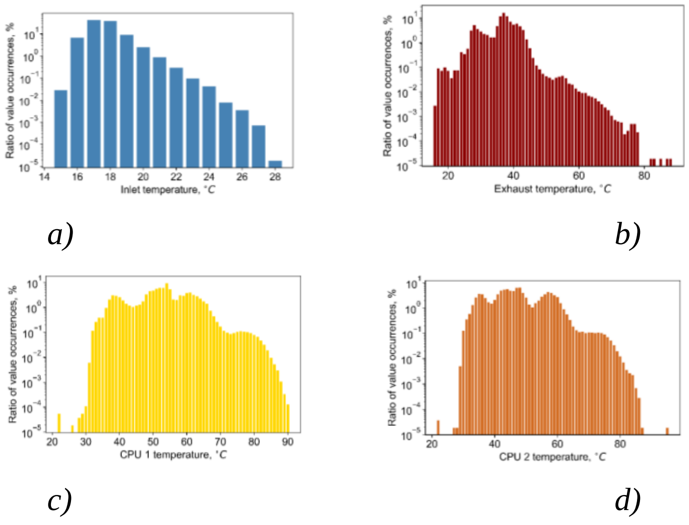
For each load of the node, the temperature increases between incoming air and outgoing air from the calculation nodes were considered. In Fig. 5 average temperature observed at the inlet of the nodes – blue segment- (in the cold aisle), and exhaust temperature at their rear side - amaranth segment - (in the hot aisle). The temperature measurements were also taken next to two CPUs of every node. The setpoints of the cooling system were fixed approximately 18°C at the output and 24°C at the input of the cooling system, which is represented in Fig. 5 as blue and red vertical lines respectively. It is subsequently unravelled that the lower setpoint is variable and provides supply air at 15-18°C as well as high setpoint varies between 24-26°C.

Figure 5. Temperature observed on average in all nodes during 9 months with vertical lines corresponding to cold and hot aisle setpoints.



As observed from the graph, cold aisle preserves the setpoint temperature at the inlet of the node, which affirms the efficient design of the cold aisle (i.e. supported by existing plastic panels isolating cold aisle from other spaces in the IT room of the DC). However, the exhaust temperature is registered on average at 10°C higher level than the hot aisle setpoint. Notably, exhaust temperature sensors are directly located at the rear of the node (i.e. in the hottest parts of the hot aisle). Therefore, the air in the hot aisle is distributed in such a way that the hotspots are immediately located at the back of server racks, and the hot aisle air is cooled down to the 24-26°C input level of the cooling system at the CRAC intake due to air circulation and mix in the hot aisle. Meanwhile, the previously mentioned difference of 10°C between the hotspots and the ambient temperature unravels the cooling system weak points, since it does not account for hotspots directional cooling. In the long term, the constant presence of the hot spots might affect the servers' performance which should be carefully addressed by the DC operator. Remarkably, although the hotspots are present at the rear of the nodes, the cooling system does influence temperature around the nodes. Cold air flows through the node and is measured at the inlet, then at CPU 2 and CPU 1 locations and finally at the exhaust point of the server. The differences between observed temperature ranges in these locations are averaged for all the nodes. Following the airflow, inlet air temperature is heated by 30℃ inside the servers until it reaches CPU sensors. It continues to increase by 4-6℃ while moving from CPU 2 to CPU 1 sensors and, due to internal server fans, drops by 15-20℃ by the moment it reaches the rear of the nodes. The study of observed temperature values distribution contributes to the overall understanding of the thermal characteristics, as it gives a more detailed overview of the prevailing temperature shown in Fig. 5 and Fig. 6. For every type of thermal sensors, the temperature values are recorded as an integer number, so the percentage of occurrences of each value is calculated. The inlet temperature is registered around 18°C in the majority of cases and has risen up to 28°C in around 0.0001% of cases. The inference is that cold aisle temperature stays around the 15-18°C setpoint within most of the monitored period. Ranges of the exhaust temperature and those of CPUs 1 and 2 are in the range 20-60°C with most frequently monitored values in the intervals of 18-50°C, 35-75°C and 30-70°C respectively. Although these observations might contain measurement errors, they reveal the possibility of servers' risks as they are frequently found to be overheated.

Figure 6. Distribution of monitored temperature values taken for all nodes and months.



a) *b)*

c) *d)*

Further, the study focuses on variation between subsequent thermal measurements with the aim to explore stability of the temperature around the nodes. All temperature types have distinct peaks of zero variation which decreases symmetrically and assumes a Gaussian distribution. It could be concluded that temperature tends to be stable in the majority of monitored cases. However, the graphs for exhaust and CPUs 1 and 2 temperature variation (Fig. 6 reveal that less than 0.001% of the recorded measurements show an amplitude of air temperature changes of 20°C or more occurring at corresponding locations. Sudden infrequent temperature fluctuations are less dangerous than long periods of constant high temperature stabilization. Nevertheless, further investigation is needed to identify causes of abrupt temperature changes so that measures could be undertaken by DC operator to maintain longer periods of constant favourable conditions. We propose a scheduler upgrade which aims to optimise CPU, memories, and outlet temperatures without relying on the profile information.

Table 2. Schema with prefixed target for improved scheduler.

	Proposed scheduler
Strategy	Schedules task based on utilisation and temperature information gathered at run-time
Task Assignment	Assigns tasks to the coolest node in a cluster at any point
Task Scheduling	Schedules tasks on the coolest node in a cluster
Temperature control	Maintains uniform temperate across a cluster
Node Activity	At least 50% are active nodes at any given time in a cluster
Pros	Works better with a large cluster
Cons	Overhead of communication of temperature and utilisation information

During the design of the scheduler, we address the following four issues. 1)Differentiate between CPU-intensive tasks and memory-intensive tasks; 2) Consider CPU and memory utilisation during the scheduling process; 3)Maintain CPU and memory temperatures under the threshold redline temperatures; 4) Minimise the average outlet temperature nodes to reduce cooling cost. The scheduler receives feedback of node status through queried Confluent (monitoring software installed on each node) value. The scheduler (when all the nodes are busy) will manage the temperatures, embarks on a load balancing procedure by keeping track of the coolest nodes in the cluster. In doing so, the scheduler continues job executions even in hot conditions. The scheduler maintains the average cluster CPU and memory utilisation represented by U- CPU- avg and U-MEM-avg and CPU, memory temperatures represented by T-CPU-avg, T-MEM-avg, respectively. The goal of our “upgraded” scheduler is to maximise the COP (coefficient of performance). Below are the 7 constraints that are subjected to our upgraded scheduler:

1. check constraint $T^{i}_{CPU} < T_{CPUAvg}$
2. otherwise, check constraint $T^{i}_{Mem} < T_{Memavg}$

3. $T_{Memavg} < T_{MemMax}$ & $T_{CPUavg} < T_{CPUMax}$
4. $T^{i}_{out} \leqslant (\sum_{i=1}^{N} Tout)$ / N
5. Each NodeManager is assigned only one task at a time
6. Each task is assigned to utmost one node
7. Minimise response time of job

With the first and second constraints in place, it is ensured that memory and CPU temperatures should always stay below the threshold temperatures. If a cluster's nodes exceed the redline threshold, then optimise the temperature by assigning tasks to the coolest node in the cluster. The third constraint specifies that if the average memory or CPU temperature rises above the maximum temperature, then the scheduler should stop scheduling tasks as it might encounter hardware failures. The fourth constraint signifies that the outlet temperature of a node should be the same as the average outlet temperature of the cluster. The fifth and sixth constraints ensure that a node gets utmost one task and a task is running on utmost one node at a time. The last point aims at reducing the completion time of a job to achieve optimal performance. The algorithm feeds into the node matrix considering the physical arrangement of every single node inside the racks. Obtain the profile of the user who requested for resources by retrieving the user's profile from a list of stored profiles.
The algorithm passes through all the nodes to understand the level of use and the respective temperatures for each node. If the profile does not exist, then when a user executes a job for the first time, the algorithm calculates a profile on the fly. All the indicated threshold values are operating values calculated for each cluster configuration and are periodically recalculated and revised according to the use of the cluster nodes. Subsequently, some temperature calculations are made from the current state of the cluster (through a snapshot of thermal status). Finally, the last step is to assign the job to the node based on the expected type of job expected. In this way, the algorithm avoids hotspot and coldspots situations by distributing the jobs uniformly in the cluster, so that the temperature remains more or less constant for all the nodes.

5. Conclusions and Future Works

In DCs scenario, energy efficiency represents the essential goal for a sizeable high-performance computing facility to operate within society. In terms of DC operations, energy efficiency could be interpreted in a number of ways. This work primarily focuses on two of major aspects: IT equipment energy productivity (workloads) and thermal characteristics of an IT room and equipment. The findings of this work are based on analysis of available monitoring thermal data characterising the ENEA-HPC DC, CRESCO6. This analysis has unravelled possible improvements for thermal design and load management. In this work, using the big dataset for CRECO6 IT room temperature measurements, sequential clustering has been performed to group nodes based on thermal ranges in which they reside most frequently during the period of observations. Moreover, a data mining algorithm has been employed to locate the hotspots. Several measures to mitigate risks associated with the issue of hotspots have been recommended: directional cooling, load management, and continuous monitoring of the IT room thermal conditions. This research brings about two positive effects in terms of DC energy efficiency. Firstly, being a thermal design pitfall, hotspots pose as a risk of local overheating and deterioration of servers exposed to high temperature for prolonged periods of time. In this regard, localisation of hotspots is crucial for better overview and control of the IT room temperature distribution. It provides a direction of future thermal management improvement that would mitigate the specified risk. Secondly, with less computational power (and thus energy consumption) analysis techniques has brought about sufficient information to incentivise improvement of thermal conditions. Finally, the results imply that the majority of the servers operated in the medium and hot temperature ranges. Given that 8% of all cluster servers have been most frequently labelled as hot range nodes, a list of recommendations is suggested below to address the issue of hotspots.

References

1. J. Koomey: Growth in Data Center Electricity use 2005 to 2010. Analytics Press., 1–24. https://doi.org/10.1088/1748-9326/3/3/034008, 2011.
2. Greenpeace: How Dirty Is Your Data? A Look at the Energy Choices That Power Cloud Computing, 2011.
3. V. D. Reddy, et al: Metrics for Sustainable Data Centers, IEEE Trans. Sustain. Comput., vol. 2, no. 3, pp. 290–303, Jul. 2017.
4. J Kong, S. W. Chung, k. Skadron: Recent thermal management techniques for microprocessors. ACM Computing Surveys, vol. 44, no 3. https://doi.org/10.1145/2187671.2187675, 2012.
5. L. Parolini, B. Sinopoli, B. H. Krogh, Z. Wang: A Cyber-Physical Systems Approach to Data Center Modeling and Control for Energy Efficiency. Proceedings of the IEEE. 100. 254-268. 10.1109/JPROC.2011.2161244, 2012.
6. L. Wang, et al: Thermal aware workload placement with task-temperature profiles in a datacenter. J. Supercomput. 2012, 61, 780–803, doi:10.1007/s11227-011-0635-z, 2012.
7. ASHRAE Technical Committee 9.9, “Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance,” 2011.
8. X. Jin, F. Zhang, A. V. Vasilakos, and Z. Liu: Green Data Centers: A Survey, Perspectives, and Future Directions,” arXiv, vol. 1608, no. 00687, 2016.
9. M. Chinnici, A. Capozzoli, and G. Serale: Measuring energy efficiency in data centers. In Pervasive Computing Next Generation Platforms for Intelligent Data Collection, Chapter 10, pp. 299-351, 2016.
10. M. Chinnici, et al: Data Center, a Cyber-Physical System: Improving Energy Efficiency Through the Power Management,” in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf (DASC/PiCom/DataCom/CyberSciTech)*, 2017, pp. 269–272.
11. S. Yeo, et al.: ATAC: Ambient Temperature-Aware Capping for Power Efficient Datacenters, in Proceedings of the 5th ACM Symposium on Cloud Computing, Seattle, WA, USA, doi:10.1145/2670979.2670966, 2014.
12. A. Capozzoli, et al: Thermal Metrics for Data Centers: A Critical Review,” *Energy Procedia*, vol. 62, pp. 391–400, Jan. 2014.
13. A. Capozzoli, et al: Review on performance metrics for energy efficiency in data center: The role of thermal management,” *LNC,* vol. 8945, pp. 135–151, 2015.