On Energy- and Cooling-Aware Data Centre Workload Management

Danuta Sorina Chisca
ALMA MATER STUDIORUM
Università di Bologna
Email: danutasorina.chisca@studio.unibo.it

Ignacio Castiñeiras Deepak Mehta Barry O'Sullivan
Insight Centre for Data Analytics
University College Cork, Ireland
{*ignacio.castineiras*|*deepak.mehta*|*barry.osullivan*}*@insight-centre.org*

***Abstract*—The power consumption of a data centre (DC) can be attributed to the power consumed for running the servers and to the computer room air conditioner (CRAC) power for cooling them. The challenge is to distribute the load among servers, controlling the number of active servers and optimally balancing IT and cooling power requirement. This goal demands integration of thermal, power and workload models that minimises a non-linear energy utilisation function. In this paper first we encode this problem using a non-linear objective function and use local search for solving it. We carry out simulation experiments using data provided by the Bluesim tool. The results encourage the effectiveness of our approach, showing that system-wide energy utilisation can be reduced using a holistic approach.**

I. Introduction

The growing demand for providing IT services has resulted in the proliferation of large-sized Data Centres (DCs) around the world, which consume enormous amounts of electricity. Therefore, the energy management has become increasingly critical for sustainable DCs not only for economic reasons but also for environmental reasons. The power consumption of a DC can be attributed to the power consumed for running the servers and to the power consumed by the computer room air conditioner (CRAC) for cooling servers. The goal is to minimise the total power consumption by distributing the load among servers, controlling the number of active servers and optimally balancing IT computing power and cooling power requirement.

Many approaches have been developed to reduce the power consumption of a DC, but a lot of work has focused either on minimising only IT power or cooling power. The former is generally achieved by consolidating workloads in as few servers as possible, and the latter is achieved by maximising the supplied temperature by balancing the workload over servers. The challenge is to resolve the trade-off between IT power and cooling power optimally.

One of the computational challenges in minimising the total power is that the energy computation function is non-linear. Consequently, some previously proposed approaches fix the temperature supplied (based on some ad-hoc reasoning) to make it linear. In [1], a Cooling-aware workload placement problem is proposed as a Mixed Integer Non-Linear Problem, but they linearised the objective function because the non-linear solvers can solve the optimality only for small instances.Their model does not consider workload constraints related to Service Level Agreements (SLAs). In [2], the power consumption is minimised by using an Integer Linear Programming problem. They fix the values of the cooling temperature supplied and focus on server consolidation. In [3], the focus is put on workload distribution over chassis instead of servers, and the heat re-circulation is minimised through a linear objective function. They perform an analysis by studying different values of temperature supplied and its impact on the concentration/dispersion of the workloads. It is important to note that, if chassis are assigned to workloads, then in some cases the resulting solution might not be feasible to implement.

In this paper we focus on the minimisation of the total power consumption of a DC, where the objective function is non-linear and the decisions are to assign workloads to servers. We present a Mixed Integer Non-Linear Programming problem that minimises the overall power consumption in the DC, and we use local search to solve the problem instances. Furthermore, we present a decomposition approach, where first we minimise the cooling power (by maximising the temperature for cooling the servers) and then we use this temperature value as a constraint to minimise the total IT power. We also present some empirical results showing the effectiveness of the proposed approach for minimising total power consumption of a DC.

II. Mixed Integer Non-Linear Model

A DC is a warehouse with several rows of racks and other system facilities, such as batteries, generators and cooling units. Typically, it embraces the hot/cold aisle model, where each row is placed between a hot and a cold aisle, as shown in Figure 1. The cold air supplied by the Computer Room Air Conditioning (CRAC) units passes through an elevated floor and picks up the heat exited by the servers. The warm exit air returns to the CRAC intakes, which are positioned at the end or on the ceiling of the hot aisles. In this section we present a non-linear mixed integer programming formulation for the assignment of workloads to servers to minimise the total power consumption of DC.

Workload Allocation Constraints. Let m be the number of servers and let n be the number of workloads. Let x_{ij} be a

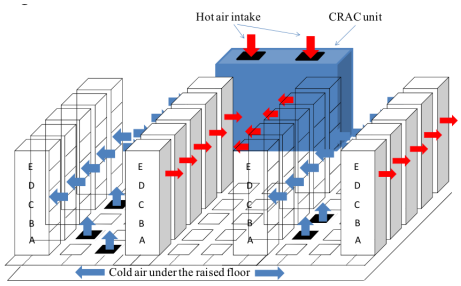

Figure 1. Hot-aisle/cold-aisle [2]

Boolean variable which is equal to 1 if the workload i is assigned to the server j. Equation (1) guarantees that each workload is assigned to a single server.

$$\forall_{1\leq j\leq n}: \quad \sum_{i=1}^{m} x_{ij} = 1 \tag{1}$$

Let o be the number of chassis. Let S_j be the set of servers associated with a chassis j. Let y_i be an integer variable that denotes the chassis assigned to workload i.

$$\forall_{1\leq i\leq n}: \quad y_i = \sum_{j=1}^{o} \sum_{k\in S_j} j \times x_{ik} \tag{2}$$

A service is a set of workloads. Let S^E be the set of services such that a set of workloads associated with each service $s \in S^E$ must be assigned to same chassis.

$$\forall_{s\in S^E} \forall_{\{i,j\}\subseteq s}: \quad y_i = y_j \tag{3}$$

Let S^D be the set of services such that a set of workloads associated with each service $s \in S^D$ must be assigned to different chassis.

$$\forall_{s\in S^D} \forall_{\{i,j\}\subseteq s}: \quad y_i \neq y_j \tag{4}$$

IT Power Constraints. Let W_i be the power required for workload i. Let p_j^{idle} be the power required for keeping server j active when no workload is assigned to it. Let p_j^{max} be the maximum power consumed by server j. Let $P_j \in [0, P_j^{max}]$ be a continuous variable that denotes the power consumed by server j. The power consumed by a server is linearly proportional to its CPU utilisation [1]. Equation (5) computes the power consumed by each server.

$$\forall_{1\leq j\leq m}: \quad P_j = \sum_{i=1}^{n} x_{ij} * W_i + \lambda_j * p_j^{idle} \tag{5}$$

Let λ_i be a Boolean variable which is equal to 1 if server i has some workload assigned. Equation (6) enforces that only active servers can host workloads.

$$\forall_{1\leq j\leq m}: \quad x_{ij} \leq \lambda_j \tag{6}$$

Let Q_i be a continuous variable that denotes the power consumed by a chassis j.

$$\forall_{1\leq j\leq o}: \quad Q_j = \sum_{i\in S_j} P_i \tag{7}$$

To minimise only IT power we can minimise $\sum_{j=1}^{m} P_j$ subject to the above mentioned constraints.

Thermal Constraints. An important issue in DCs is the hot air exhausted by the servers, which recirculates to the intake of the other servers and warms up the air, thus reducing the efficiency of the cooling system. The temperature evolution in a DC can be computed using Computational Fluid Dynamic (CFD) simulators, but it is complex and time-consuming for online use. Another way is to use the abstract heat flow model of [4], [5]. Under the assumption of steady state condition, this abstract model can be reduced to a heat recirculation matrix D, where the element D_{ij} represents the temperature increment at the inlet of chassis i per unit of electrical power consumed by chassis j [3].

Let $T_j^{in} \in [11, 25]$ be a continuous variable that denotes the temperature at the inlet of the chassis j. Notice that the domain of inlet variables implicitly guarantee that the inlet temperature at chassis j must not exceed the maximum allowed inlet temperature as specified by the manufacturer. Let $t_{sup} \in [11, 25]$ be a continuous variable that denotes the temperature of the cool air supplied by CRAC. The inlet temperature at chassis i is the sum of the cool air provided by CRAC and the hot air contribution of all the chassis.

$$\forall_{1\leq j\leq o}: \quad T_j^{in} = t_{sup} + \sum_{1\leq i\leq o} D_{ij} \cdot Q_i \tag{8}$$

This equation means that applying a power distribution to the o chassis, globally cooled with air at temperature t_{sup}, their inlet temperatures will converge to T^{in} in steady state.

Total DC Power Consumption. The power consumption of a DC is the sum of the power consumed by the servers or chassis (denoted by P_{IT}) plus the power consumed by the CRAC unit (denoted by P_{CRAC}). If we assume that steady state conditions hold, then all the power drawn by servers is dissipated as heat. The amount of heat the CRAC removes from the computer room is the total IT power [2]. The efficiency of a cooling system is usually characterised by the Coefficient of Performance (COP), which can be written as the ratio between the heat removed and the cooling energy.

$$COP = \frac{\texttt{heat removed}}{\texttt{cooling energy}} = \frac{P_{IT}}{P_{CRAC}}$$

This coefficient depends directly on the temperature of the cold air supplied by the CRAC, denoted by t_{sup}. So the power consumed by the CRAC is P_{IT}/COP, and it varies quadratically with t_{sup}, and described in [3]

$$COP = 0.0068 * t_{sup}^2 + 0.0008 * t_{sup} + 0.458 \tag{9}$$

In order to minimise the total energy consumption, we minimise the following objective function:

$$\left(1 + \frac{1}{COP}\right) \sum_{j=1}^{o} Q_j \tag{10}$$

III. Emprical Results

A. Experimental Setup

We implemented the non-linear model for minimising the total power consumption of the DC using LocalSolver. To perform the experiments we used the data provided by Bluesim tool [6]. We considered a DC with 200 servers, where each server belongs to the IBM series 350M2 model with idle power of 100Watts and maximum power of 300Watts. For the heat recirculation, we used the D matrix provided by the BlueSim tool [6]. The workloads were generated randomly: We generated 10 instances for each value of $n \in \{250, 300, 350\}$, where the power footprint was randomly selected between 1 and 200W. We further enforced that the aggregate power footprint requirement is approximately a certain percentage utilisation of the DC computing capacity. More specifically, we used 3 values of DC utilisation percentage: 50%, 60% and 70%. In total we carried out experiments on 90 instances. The simulations were carried out on a computer with an 2.4 CPU GHz, i5 Cores with 8 GB of physical RAM. For all the experiments the timeout was set to 300 seconds.

B. Thermal-aware vs Non-thermal-aware

In this section we first present some results to show the drawbacks of solving the workload allocation problem in a non-thermal-aware setting, and then present results for total power consumption when both IT power and cooling power are considered along with workload allocation, affinity and anti-affinity constraints. To this end, two solving approaches are compared: A first one, denoted by MIT (of minimising IT power), which is only concerned on minimising the IT power of running the servers. We use this approach as a baseline. The second approach, denoted by MTP (of minimising total power), deals with the non-linear problem, by considering both the power for running and cooling the servers.

Although the first approach is simpler, the lack of control at the temperature needed to run the servers can lead to an overall higher cost. More specifically, MIT basically tries to allocate the workloads in the fewer servers as possible, given that all of them have the same characteristics. However, this can lead to higher inlet temperatuqres in some of the chassis of the DC. Given that the DC contains a single CRAC unit, this lead to an overall lower temperature to be supplied, and thus to a higher cooling cost. Maintaining higher values of supplied temperature in a non-thermal aware setting can be risky and cause system-wide failures. Figure 2 plots the results of total power consumption of DC when only IT power is minimised. It uses different values of DC utilisation and different temperatures. It can be seen that, if DC is utilised 50%, 60% and 70%, then its maximum supplied temperature can be at most 19.5, 18.3 and 17 degree Celsius, resp.

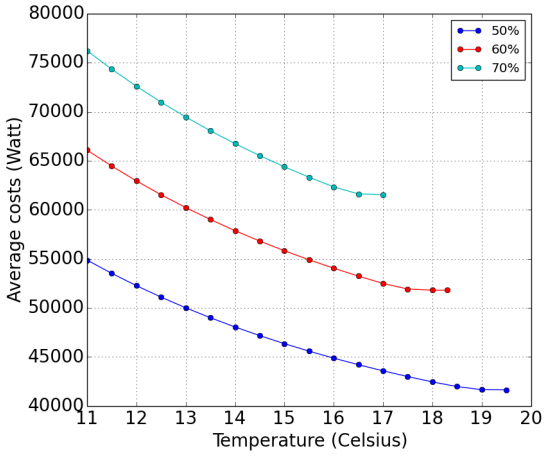

Figure 2. Total Power Consumption of DC in a non-thermal aware setting for different percentage values of DC utilisation and different temperatures.

C. Total Power Consumption

MTP considers both the IT and cooling constraints. Although it deals with a harder model, it can resolve the tradeoff between the two costs and thus lead to overall energy savings. For the fairness of the results, if a DC performing workload allocation uses MIT to solve a benchmark of $M = \{M_1, M_2, \ldots, M_k\}$ instances, then one can not consider k maximum temperatures allowed, but a single one $k' = \min(M_1, M_2, \ldots, M_k)$, avoiding violating the inlet temperature of any chassis irrespectively of the concrete instance M_i being run. We use this temperature k' to compute the results of MIT and compare them with the MTP approach.

The results for MIT and MTP of Table I show that MTP requires less power consumption than MIT. Notice that the results measurements reflect the power (Watts), but not the energy (Watts/second). This lead even small power differences between MIT and MTP solutions to high energy savings (e.g., consider a static consolidation during several hours).

We further implemented a decomposition approach for minimising the total power consumption. It first maximises the value of supplied temperature, which would effectively minimise the cooling power requirement. Then, it minimises the total IT power by fixing the supplied temperature to the value previously obtained. This decomposition approach, which has the advantage of being linear, is denoted by `MCP`. For the fairness of the results, the timeout was split (150 seconds for each step).

The results in Table I show that MCP can outperform both MIT and MTP. It is seen that, despite increasing the complexity of the model for MTP, LocalSolver manages to produce good results within a short span of 300 seconds.

Table I
TOTAL POWER CONSUMPTION FOR THE DIFFERENT APPROACHES. THE POWER IS IN TERMS OF WATTS AND THE TEMPERATURE IS IN TERMS OF DEGREE CELSIUS.

DC %Utilisation	n	MIT Power.	MIT Temp.	MTP Power.	MTP Temp.	MCP Power.	MCP Temp.	Lower Bounds
50%	250	44723	16	39703	21.28	39703	21.25	36621
	300	44707	16	39546	21.28	39666	21.28	36644
	350	44806	16	39490	21.28	39527	21.29	36678
60%	250	53834	16	48917	20.33	48927	20.34	43872
	300	53630	16	48544	20.41	48809	20.42	43919
	350	53677	16	48202	20.46	48126	20.48	43946
70%	250	61675	16	57228	19.26	57317	19.28	50916
	300	62787	16	58871	18.92	58834	18.90	51182
	350	62621	16	58246	19.07	58570	19.05	51209

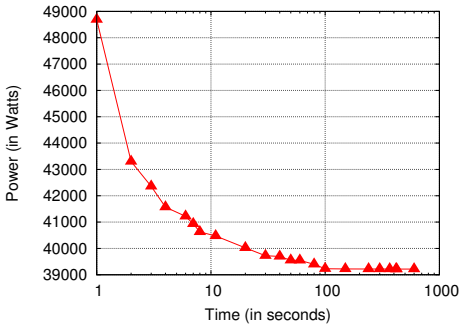

Figure 3. LocalSolver: Solution cost vs Time

The results show that, as the utilisation of DC increases, the power requirement increases for all the approaches. Nevertheless, `MTP` and `MCP` approaches are consistently better than the `MIT` approach. Table I also shows that the required cooling temperature computed using `MTP` and `MCP` approaches is consistently higher than the 15.8 degree Celcisus used by MIT approach. Thus, the power requirements of the latter is more than that of the former. Figure 3 shows the improvement in cost over time using local search. We can see that local search converges quickly and there is hardly any improvement after 100 seconds. To verify the efficiency of LocalSolver we computed very simple lower bounds. On them, the COP value was computed using the maximum temperature of 25 degree Celsius, which was multiplied by the lower bound of the IT power computed using the following equation:

$$\left(\left\lceil \frac{m * (maxp - minp)}{\sum_{1 \le i \le n} W_i} \right\rceil \cdot minp + \sum_{1 \le i \le n} W_i\right)$$

Here, `minp` denotes the minimum power requirement of a server when no workload is assigned to it, whereas `maxp` denotes the maximum power requirement when the machine is running at its full capacity. Despite being a simple and probably weak lower bound, the results suggest that LocalSolver can compute good quality solutions in a very short span of time.

IV. CONCLUSION

We have presented our ongoing work on minimising the total power consumption of a DC. As a novelty w.r.t. the related literature, we have considered the non-linear objective function underlying this total power consumption subject to workload, power and thermal constraints. We have formulated and solved the problem using LocalSolver. The results have revealed that it is a very effective and scalable approach, which is able to find quality solutions for all the problem instances within very limited time.

Acknowledgement. This work was supported by SFI-PI Grant 10/IN.1/I302, and FP7 Grant 608826 (GENiC - Globally Optimised Energy Efficient Data Centres). The Insight Centre for Data Analytics is supported by SFI Grant SFI/12/RC/2289.

REFERENCES

[1] P. Cremonesi, A. Sansottera, and S. Gualandi, "On the cooling-aware workload placement problem," in *AI for Data Center Management and Cloud Computing, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 7, 2011*.

[2] E. Pakbaznia and M. Pedram, "Minimizing data center cooling and server power costs," in *Proceedings of the 2009 ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '09. New York, NY, USA: ACM, 2009, pp. 145–150. [Online]. Available: http://doi.acm.org/10.1145/1594233.1594268

[3] H. Shamalizadeh, L. Almeida, S. Wan, P. Amaral, S. Fu, and S. Prabh, "Optimized thermal-aware workload distribution considering allocation constraints in data centers," in *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, Aug 2013, pp. 208–214.

[4] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, "Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 19, no. 11, pp. 1458–1472, Nov 2008.

[5] Q. Tang, T. Mukherjee, S. K. S. Gupta, and P. Cayton, "Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters," in *Intelligent Sensing and Information Processing, 2006. ICISIP 2006. Fourth International Conference on*, Oct 2006, pp. 203–208.

[6] "Bluetool," https://impact.asu.edu/BlueTool/wiki/index.php/BlueSim.