Efficient Management of Cooling Systems in Green Datacenters

Ibrahim Safieddine, Noël de Palma
ERODS - LIG
Université Grenoble Alpes
Grenoble, France
email: {ibrahim.safieddine, noel.de-palma}@imag.fr

***Abstract*—Next-generation green datacenters were designed for optimized Power Usage Effectiveness (PUE), which is the ratio of the total power consumption of the datacenter over the computing consumption. Continuous improvement of these datacenters target the reduction of PUE in accordance with the servers' load, to approach the minimal target, i.e., PUE = 1. Datacenters must ensure in the same time a very high level of availability of resources and good management of failures, thanks to redundant equipments, while optimizing power consumption and cooling costs and reducing the environmental footprint. A datacenter consists of computing, power distribution and cooling parts. The cooling part represents the main cost that significantly increases the power bill and consequently the PUE. A datacenter can be cooled using heterogeneous cooling systems for redundancy in the case of failure. These systems have a variable consumption depending on the load and on external parameters, e.g., weather and external temperature. This paper presents an efficient cooling manager, which aims to minimize the PUE while satisfying the Service Level Agreement (SLA), by reducing the power consumption of the datacenter and opting for the most efficient cooling system according to climate conditions and by limiting temperature variations and cooling mode transitions. Our system uses an overview of all datacenter parts to provide an optimal decision that complies with regulations when using natural resources, e.g., groundwater.**

Keywords*—*Power Usage Effectiveness (PUE); cooling systems; green datacenter; autonomic computing; service level agreement (SLA)

I. INTRODUCTION

The datacenter market is growing at a rate of 15% per year globally [1]. Green computing, which aims to reduce energy consumption and the related greenhouse effect, remains a priority for datacenter managers due to the increase in the price per kWh, e.g, France's electricity bill is expected to increase by 50% in 2020 [2]. As servers become smaller without necessary consuming less energy, datacenters which have more and more servers generate more heat, and the cooling power needed will grow. The concentration of computing power per m^2 has grown very quickly over the last 10 years from 15 to 30 kW per rack, which results in significant heat dissipation and thus a higher cost for cooling.

Knowing that the electricity bill is the main operational charge in a datacenter, the new challenge of datacenters is the mastery of electrical distribution, the choice of the best cooling technologies, e.g., air-cooled or water-air, and its optimization for better performance. The intelligence in a datacenter relies on sensor networks that provide real-time measurements and robust industrial automation control, to find the best operating point.

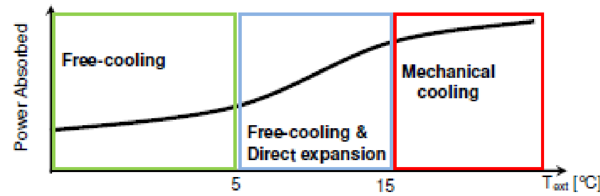

Figure 1. Hybrid cooling consumption.

The objective of datacenters is to ensure near 100% facilities availability, through redundant components which keeps the systems up even in case of the failure of an active element or during maintenance. Hence, the importance of optimizing the very complex cooling system as a whole, by analyzing a large number of parameters (ambient temperature, humidity, weather forecast, servers load, etc.) and having an overview of all datacenter equipments for a more efficient management of cooling. This will impact the overall consumption in the datacenter and the cost of operation, and thus improves the PUE and reduce costs.

There are many efficient cooling systems [3] [4] [5] [6]. Free cooling [3] is an economic method that uses low external air temperature and less power to cool water instead of mechanical refrigeration. Hybrid cooling [4] incorporates two cooling modes, free and electrical cooling, with an internal system that switches between these modes depending on the outside air temperature. Figure 1 shows the variation of the consumption of an hybrid cooling system with the external temperature [7].

Another cooling system that uses groundwater [5] and rivers' [6] water is a very economical cooling system. The system works by pumping cool water through a heat exchanger and then re-injecting the heated water back into the source, resulting in no net loss of groundwater. It uses the cold water in an open loop to cool the internal circuit water.

In this paper, we propose an autonomic solution that efficiently manages and optimizes the choice of the cooling system in a high available green datacenter while satisfying the SLA. Our algorithm can manage many heterogeneous cooling systems with different cooling capacities to minimize the cooling power consumption and can use multiple cooling systems simultaneously for better efficiency. To improve our solution, we correlated the datacenter internal measurements with other indicators such as external temperatures and weather forecast.

 ISBN: 978-1-61208-405-3

We used a global vision on all datacenter layers, FaaS (Facilities as a Service)-IaaS (Infrastructure as a Service)-PaaS (Platform as a service)-SaaS (Software as a Service), to avoid conflicted decisions and process in a better and faster way. We validate our algorithm on a real architecture and real measurements from the green datacenter of Eolas[1] [8] that can host more than 13,000 servers. We then demonstrate that this solution can be scaled to include other cooling systems and can be adapted on more complicated environments. The rest of this paper is organized as follows. In Section II, we present the state of the art on cooling system optimization algorithms. In Section III, we describe our algorithm for optimizing the management of cooling systems. In Section IV, we present the architecture of a redundant green datacenter on which our work is based. Finally, in Section V, we summarize the conclusions and future perspectives of this work.

II. Related work

In order to reduce the power consumption of cooling systems, several solutions are proposed. Some solutions are based on varying water or air temperature in the servers' room to reduce the cooling power. This solution can be coupled with servers consolidation to reduce the dissipation of heat. Other global approaches aims to reduce the consumption of the cooling infrastructure by switching between free and electrical cooling according to external temperatures.

Shaoming et al. [9] discuss the impact of server consolidation and the variation of the cold aisle temperature in a servers room, on the cooling consumption of the datacenter. They focus on optimization of cooling consumption and maintenance costs. Increasing the temperature by 1°C can reduce energy cooling consumption by 2 to 5%, however high temperature reduce the reliability of electronic components, and increases the cost of hardware maintenance of CPU and Memory. In the same way, the consolidation that consists on cycles of start/stop servers, decrease the cooling costs but in the same time decreases the hard-disks lifetime.

While the previous paper focuses on increasing inlet temperature and servers consolidation in servers room, Jungsoo et al. [3] use servers consolidation and exploit time-varying servers workload and external climate conditions in order to reduce the power consumption of the entire datacenter. This system is based on the maximum usage of the free-cooling and preventing frequent cooling transition.

Ratnesh and al. [10] propose a framework for dynamic thermal management based on asymmetric workload placement that can promote uniform temperature distribution that reduces local hot spots, quickly responds to thermal emergencies, reduces energy consumption costs, reduces initial cooling system capital costs and improves equipment reliability. This framework is not related to cooling systems capacity.

[1]Eolas (www.eolas.fr) is a French company based in Grenoble, part of the international Consulting and System Integration (CSI) company Business & Decision

III. Implementation

The basic automatons implemented in most of the cooling systems are very reliable and can manage efficiently the availability of cooling in case of failure. However, they are limited to the predefined priority established by the Data Center Manager (DCM) and each automaton can manage a limited number of cooling systems and works independently from other datacenter layers. In our implementation, we used the requested cooling capacity and the external weather conditions to improve the global efficiency of the cooling system and reduce the datacenter consumption. Using weather forecast, we can limit the transition between cooling systems when the external temperature is constant or when the temperature decrease while using free-cooling. In hybrid cooling, we can predict the cooling system that will be used (free-cooling, mixed or chiller cooling) depending on the temperature forecasts before starting the system. In addition, the usage of the groundwater cooling is highly regulated: the water flow and the yearly water volume is limited by the law. Our algorithm proposes a better way to manage the groundwater cooling system based on external temperature and servers load history throughout the year. Knowing that it is possible to start several cooling system simultaneously, it is important to reduce the number of active cooling systems to reduce the consumption.

In order to prevent damage to cooling system by repetitive start/stop cycles, we defined a minimal period between two transitions, based on the systems data manual.

To minimize the electricity cost and then reduce the PUE, we setup the cooling power requirement of the datacenter model which estimate the impact of servers consumption on the cooling power needed, as presented in (1):

$$\sum C_P(kW) = 0.9 * \sum S_C(kW) \tag{1}$$

where C_P is the cooling power needed in kW and S_C the total electrical consumption of servers in kW. Almost, all the power consumed by the server is transformed to heat.
For the optimization problem, we used a linear optimization program, where the goal is to minimize the cooling electrical power consumption linear function while respecting operation constraints. Linear programming is a technique for the optimization of a linear objective function, subject to linear equality and linear inequality constraints.

Equation (2) represents the cost in € to produce one kW of cooling. P_{elec} and $k€$ respectively denote the power consumed by a cooling system to produce 1kW of cooling and kWh Billing rate which comes to 6 cents/kWh in France.

$$C = P_{elec} * k€ \tag{2}$$

The global optimization program can be presented as follows:

$$Min \sum_{i=0}^{n} a_i.C_i \tag{3}$$

$$\text{Subject to}: \begin{cases} \sum_{i=0}^{n} a_i >= C_P & i \in 0, n \\ \sum Cm_i \in \{\text{possible cooling modes}\} \end{cases} \quad (4)$$

Equation (3) represents the objective function of the linear optimization program where a_i represent the amount of cooling produced by the cooling system i in kW and C_i the cost in € of 1kW of cooling using the cooling system i. n represents the maximum number of cooling systems that can be started simultaneously. We look to minimize the total consumption of started cooling systems and limit the number of started systems. While minimizing the cooling systems consumption, we minimize the datacenter consumption and then reduce the PUE. Equation (4) represents the constraints of the optimization problem. We need to satisfy the SLA by delivering as match cooling power as needed, so the total cooling power generated must be ideally equal to the needed cooling power. Cm_i is the cooling mode of the cooling system i. The possible cooling modes list is limited by many parameters as the external temperature or a high temperature of the groundwater.

Weather forecasts are used to estimate at each future period the best cooling system to be used and then limit cooling mode transitions with temperature variations.

IV. EVALUATION

In this section, we present the Eolas datacenter cooling architecture used to evaluate our algorithm. The green datacenter of Eolas is Tier 4, and therefore, designed to host mission critical computer systems, with fully redundant subsystems. Figure 2 shows that 3 principal cooling systems are used for maximum redundancy. All cooling equipment is independently dual-powered, including chillers, ventilation and air-conditioning systems. Those systems are heterogeneous: groundwater cooling (with two independent pumps), hybrid cooling and chiller cooling. Eolas uses an ultimate cooling source: city water. When all cooling systems fall down or in case of power failure, the city water (having a temperature of 12 - 14 °C) can be used to cool up to 3000 servers. This ultimate cooling source is very expensive and increase the WUE (Water Usage Effectiveness) of the datacenter.

Stopping the cooling system can be dramatic for this datacenter, i.e., every minute, the room temperature increases by 1°C, knowing that the hot aisle is fixed at 35°C, thus the room temperature may reach 60°C in just 25 minutes.

Actually, the transition between cooling modes is done manually in this datacenter. Using our algorithm, the transition is full automatic and based on several external sensors for more efficiency.

Figure 3 shows the automatic cooling mode transition between groundwater and Free cooling systems, using our optimisation algorithm, without weather forecasts in a day of May. When the temperature is too low, the Free cooling system is very efficient, i.e., his Energy Efficiency Rating (EER) is two high. When the temperature increases, the EER start decreasing and the groundwater system became more efficient. In this example, the system switches to groundwater cooling mode when the external temperature reaches 5°C. This transition is useless since the duration before returning to the Free cooling mode is too short. Using weather forecasts, the datacenter will be cooled using free cooling all the day, with no cooling systems transition.

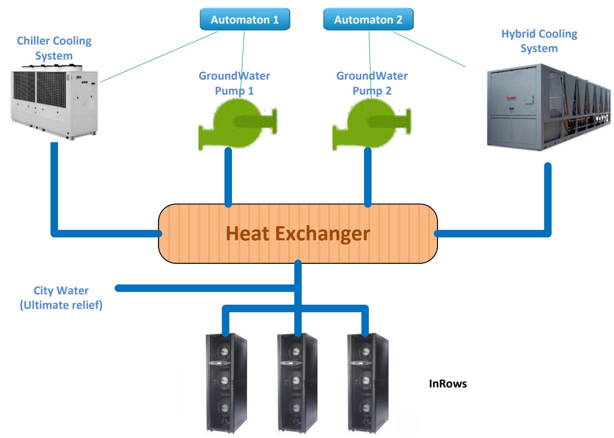

Figure 2. The global cooling architecture.

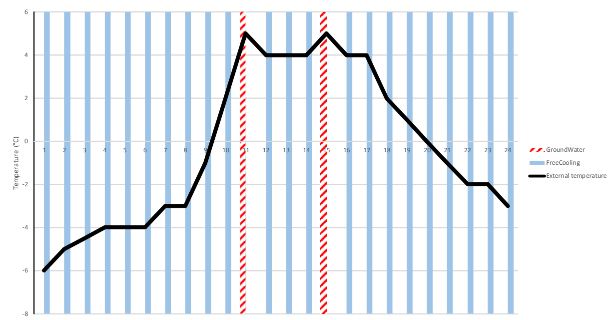

Figure 3. Cooling modes transitions in May in normal mode.

First results for a small example illustrate the potential for a coordinated control strategy to achieve better energy management than traditional industrial automatons that control the cooling systems separately. We can save up to 38% of cooling power using our algorithm.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an autonomic optimization system for heterogeneous cooling systems in a Tier 4 green datacenter. Our algorithm, connects to existing automatons and all datacenter sensors and uses external conditions and weather forecast to choose the best cooling systems combination to reduce the overall power consumption in the datacenter and limiting cooling mode transitions. We experimented this work using real data, collected from the Eolas green datacenter at Grenoble, France. As future enhancements of our solution, we intend to integrate a predictive model of the datacenter activity to predict the future cooling power needs and minimize the transition between cooling systems accordingly, in order to reduce systems failure.

Acknowledgment

This research is partially supported by ANR INFRA under a grant for the project Ctrl-Green (ANR-11-INFR 012-11) [11] and by the Datalyse project [12].

References

[1] J. G. Koomey, Growth in data center electricity use 2005 to 2010, Analytics Press, october, 2011.

[2] La commission d'enquête sur le coût réel de l'eĺectricité, SENAT - France, July 2012.

[3] J. Kim, M. Ruggiero, and D. Atienza, "Free cooling-aware dynamic power management for green datacenters," High Performance Computing and Simulation (HPCS), 2012 International Conference on, 2012, pp. 140-146.

[4] M. K. Patterson, D. Atwood, and J. G. Miner, "Evaluation of air-side economizer use in a compute-intensive data center," ASME 2009 InterPACK Conference, Vol. 2, pp. 1009-1014, July 2009.

[5] M. P. 1, "Sustainable groundwater-source cooling systems for buildings," Proceedings of the ICE - Engineering Sustainability, Vol. 161, Iss. 2, June 2008 , pp 123 133.

[6] P. Dalin, "Free cooling/natural cooling is crucial for a successful district cooling development," Teaming Up for renewable heating and cooling, April 2012.

[7] Engineering Data Manual - Aquaflair, Uniflair, October 2011.

[8] (2015, March) Eolas. [Online]. Available: http://www.businessdecision-eolas.com/544-green-it.htm

[9] S. Chen, Y. Hu, and L. Peng, "Optimization of electricity and server maintenance costs in hybrid cooling data centers," 2013 IEEE Sixth International Conference on Cloud Computing, 2008, pp. 526-533.

[10] R. Sharma, C. Bash, C. Patel, and R. Friedrich, "Balance of power: Dynamic thermal management for internet data centers," Internet Computing, IEEE (Vol.9 , Iss. 1). Feb 2005, pp. 42-49.

[11] (2015, March) Ctrlgreen. [Online]. Available: http://www.ctrlgreen.org/

[12] (2015, March) Datalyse. [Online]. Available: http://www.datalyse.fr/

 ISBN: 978-1-61208-405-3