ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/262405956

Variation-aware Server Placement and Task Assignment for Data Center Power Minimization

Conference Paper · July 2012

DOI: 10.1109/ISPA.2012.29

CITATIONS
9

READS
31

3 authors:

Ali Pahlevan
École Polytechnique Fédérale de Lausanne
15 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Mahmoud Momtazpour
Amirkabir University of Technology
11 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Maziar Goudarzi
Sharif University of Technology
112 PUBLICATIONS **436** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Data-Variety-Aware Big Data Processing View project

Big Data Processing for Genomics View project

All content following this page was uploaded by Maziar Goudarzi on 24 November 2016.

The user has requested enhancement of the downloaded file.

Variation-Aware Server Placement and Task Assignment for Data Center Power Minimization

Ali Pahlavan
Sharif University of Technology
Tehran, IRAN
pahlavan@ce.sharif.edu

Mahmoud Momtazpour
Sharif University of Technology
Tehran, IRAN
momtazpour@ee.sharif.edu

Maziar Goudarzi
Sharif University of Technology
Tehran, IRAN
goudarzi@sharif.edu

***Abstract*—Size and number of data centers are fast growing all over the world and their increasing total power consumption is a worldwide concern. Moreover, increase in the amount of process variation in nanometer technologies and its effect on total power consumption of servers has made it inevitable to move toward variation-aware power reduction strategies. This paper formulates a variation-aware joint server placement and task assignment method using Integer Linear Programming (ILP) to minimize total power consumption of data centers. We first determine the optimum placement of servers in the data center racks based on total power consumption of each server and the data center recirculation model obtained by Computational Fluid Dynamics (CFD) simulations. Then, we dynamically consolidate the ON servers in chassis and racks such that the use of power-greedy servers is minimized. Experimental results reveal up to 14.85% and an average of 8.92% power saving at different server utilization rates with respect to conventional methods.**

Keywords-Data center, Power reduction, Process variation, Server placement, Task assignment.

I. INTRODUCTION

Increasing power consumption cost of information and communication technology (ICT) and information technology (IT) equipments due to higher demands for data centers in IT industry has brought data center power reduction strategies into focus [1]. In recent years, there has been an increasing interest in power reduction of data centers, including both IT equipments and cooling devices. To this end, the authors in [2] have shown that energy consumption of the cooling system encompasses more than half of total energy consumption in a large scale data center by a typical breakdown of how the electrical capacity is divided among the various loads. Another research in [3] presents several thermal-aware methods for task assignment to reduce total power consumption of data centers. However, these methods are inefficient in online power management due to the computational overhead of the CFD simulations used. The work in [4], [5] are also other examples of online task assignment based on CFD simulation. There are also few examples in which a sensor-based online scheduling is used with lower computational overhead [6], [7]. Authors in [1] also developed a mathematical method to consolidate ON servers in chassis and racks.

To the best of our knowledge, this paper is the first to consider process variation in data center power reduction. As we scale to nanometer region, the amount of process variation increases which in turn results in a huge deviation in the leakage power consumption of the manufactured chips. Intel has reported that process variability can cause up to 20X variation in chip leakage power in high-end processors manufactured in 180nm technology [8]. Leakage power consumption along its variability has been considered as a major concern for the next 15 years design technology [9]. Therefore, the power consumption of identically manufactured servers in a data center cannot be considered to be identical anymore, resulting in a heterogeneous data center. Without consideration of process variation, existing struggles in data center power management are incomplete. For example, thermal-aware researches such as [3] may result in setting improper cooling temperature due to the non-uniformity of servers power consumption caused by process variation. This leads to increased hardware failure rate and response time due to overheated servers. Chassis consolidation approach introduced in [1] may also result in an inappropriate selection of ON chassis without taking process variation into account. This in turn, affects proper cooling temperature selection and the overall power consumption of data center.

In this paper, we propose a variation-aware server placement and task assignment algorithm to minimize total power consumption of data centers. The server placement method tries to find the optimum server placement in a data center based on the power characteristics of servers and the heat recirculation model of the data center. The heat recirculation in a data center can be modeled as a Cross Interference Matrix by using CFD tools as used in [6]. Then, we use variation-aware chassis consolidation technique along with task assignment process to reduce data center power consumption. This will ensure that incoming tasks are assigned to minimum number of ON chassis among those that have lower impact on total power consumption. We formulate our task assignment and chassis consolidation technique as an Integer Linear Programming method. Simulation results reveal improvement in total power consumption of data center in comparison with random server placement method. In summary, our contributions in this work are described as follows:

- To the best of our knowledge, we are the first to tackle variation effect in data center power optimization.

- We introduce a server placement method that determines the best position for each server based on its power consumption characteristics under process variation and the heat recirculation model of the data center. This ensures that power-greedy servers are placed in chassis having lower contribution to the total power consumption of the data center.
- We also consider process variation when performing chassis consolidation for task assignment. This guarantees that the use of power-greedy chassis is minimized, leading to a more efficient power reduction.

The rest of paper is organized as follows. In section 2, we explain consolidation concept and process variation conditions. Section 3 introduces server and cooling system power model to calculate total power consumption of data center via a mathematical formulation. In section 4, we propose our server placement algorithm and formulate task assignment using an Integer Linear Programming (ILP) method. Section 5 is related to simulation environment and evaluated results. Finally, we draw our conclusion and talk about future work in the last section.

II. PRELIMINARIES

A. Related Work

The authors in [3] proposed a method to minimize the total energy consumption of data center while considering thermal management for their reliable function. They have used CFD-based simulations in order to evaluate their task scheduling methods such as Minimal Computing Energy (MCE) which is the best one among others. The aforementioned algorithm tries to minimize the total number of running servers and turn off all other idle ones. The algorithm first assigns the tasks to the nodes having the lowest inlet temperature. In another CFD-based work [5] an analysis of a data center with temperature variation has been presented. The authors have obtained static provisioning for an arbitrary distribution of cooling resources that will lead to a reference state. They try to minimize the inlet temperature by dividing the workload on other available systems.

CFD-based techniques are often time consuming and complex. This issue will not lead to tackling online scheduling problem efficiently [10]. Therefore, fast thermal evaluation models are developed such as [6], [7]. In [6], the goal is to reduce the peak inlet temperature in order to obtain the lowest power consumption of cooling system using heat recirculation model. As a result, 20% to 30% cooling power saving will occur in different data center utilization rates. In [7] fast prediction of temperature distribution is done using distributed sensors to reduce energy consumption in high performance data centers considering recirculation properties. So, this method is suitable for real time and online management. The authors in [10] demonstrated a thermal aware resource management method considering heat transfer properties and workloads having thermal features. Their scheduling algorithm will result in reduced power consumption without performance degradation.

Reference [1] presented the chassis consolidation technique as a mathematical optimization problem and a heuristic algorithm. Optimization problem has been solved via ILP. Their experiments show that they gain 13% power saving for different utilization rates in comparison with the technique lacking consolidation.

None of the above works considers process variation effects on total power consumption of high performance servers. As will be explained below, variability effects in nanometer-scale technologies has caused dramatic variations in leakage and total power consumption of processor cores especially in high performance processors [8]. Such variation effects are visible in today processors and are expected to further rise with technology scaling when further approaching atomic scales. Process variation effects are already studied in high-performance multiprocessor and embedded systems [11], [12], [13], but to the best of our knowledge such effects have not been previously considered at data center scales.

B. Consolidation Effect

Server consolidation is the process of assigning incoming tasks to minimum number of servers and shutting down the idle servers to save power [14]. In the new data center generation, blade servers are placed in chassis with shared power and cooling units [3]. Due to high power consumption of these chassis, the concept of chassis consolidation has been introduced [1]. In this technique, we try to assign tasks to minimum number of servers such that the number of ON chassis is also minimized. In a homogenous data center with identically manufactured chassis and servers, minimizing the number of ON chassis and servers will directly result in total data center power minimization. However, as the semiconductor technology scales into ever deeper regimes, the increasing variation in the power consumption of servers may cause the data center to be more heterogeneous. Therefore, the ideal non-variation-aware chassis consolidation strategy may not be effective anymore since the power consumption of servers is not identical due to process variation effect. In this paper, we introduce an ILP-based variation-aware chassis consolidation method to find the optimal consolidation plan for the data center in the presence of process variation.

C. Process Variation

Sources of variations and uncertainties are referred to three categories: Process, Environmental and Design variations. Process variation is defined as the inability to precisely control the transistor or interconnection parameters such as channel width (W_{eff}), channel length (L_{eff}) and threshold voltage (V_{th}) or width and spacing of interconnects. Environmental variation includes temperature, supply voltage and interconnect factors. Design variation is related to timing analysis, transistor models and circuit simulations [15]. Process Variation is divided into die-to-die and intra-die or within-die forms. Within-die variation refers to variation in transistor parameters across identically designed neighboring transistors in a single chip while die-to-die variation refers to deviation in process parameters across different identically designed chips in a wafer. Intel has reported that process variability can cause up to 30% deviation in frequency and up to 20X variation in chip leakage power in high-end processors

manufactured in 180nm technology [8]. So, the power consumption of identically designed servers in a data center cannot be identical anymore. For example, a Blade Server with 50W total power consumption (considering that leakage power can contribute to almost 40% of total power consumption [16]) can have up to 400W leakage power at worst case. The process variation can also be categorized into systematic and random effects. Systematic effects include lithographic lens aberration, and random ones comprise doping density fluctuation as a sample [17]. In summary, process variation is categorized into systematic die-to-die variation, random die-to-die variation, systematic within-die variation and random within-die variation.

D. Data Center Configuration

A data center comprises a hierarchical structure typically from top to down: Rows, Racks, Chassis and Servers. Fig. 1 depicts a data center configuration with four rows of racks forming hot and cold aisles. Each rack contains several chassis. Each chassis consists of several single or multi-core servers along with a shared power unit and a cooling fan. The data center also has a Computer Room Air Conditioning (CRAC) unit that circulates hot and cold airs in order to keep the data center temperature below a certain limit. The cold air current is dispensed through the leaky tiles located on the elevated floor, and then it is sucked by the chassis fans from the cold aisles. The hot air exits the other side of chassis toward hot aisles and then leaves the data center's room by the available intakes of air conditioning unit on the ceiling located above the hot aisles. This data center configuration is similar to one used in [1].

III. Data Center Power Model

A. Power Model of Blade Servers

In this work, we model the total power consumption of a data center with two components: IT equipment's power consumption and cooling device's power consumption. The power consumption of IT equipment is comprised of server power consumption and chassis power consumption. The chassis power accounts for its cooling fan and AC to DC power convertor. The server power comprises two parts: *core* power consumption and *uncore* power consumption. The uncore power consumption of a server is the power consumed by its memory controller, I/O drivers and any other components except the processing cores [1]. The power consumption of the processing cores is also referred as core power. The difference between power consumption values for

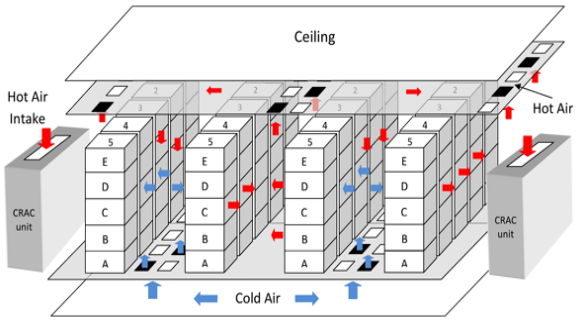

Figure 1. Data center configuration.

different utilization rates is small compared to the total power consumption of the chassis, and hence we simply assume that the active power consumption of a server is largely independent of its utilization level as also assumed in similar researches [6]. Also the tasks on the servers can be run in different voltage frequency scaling level (L_{vf}). The voltage frequency (V-F) level of each server is adjusted based on the type of running tasks in the process of assigning the tasks and is fixed until the end of the execution interval of that task. Therefore, the power consumption of blade servers is calculated as follow:

$$P_i = P_{iBase} + P_{iUncore} + P_{iCore}. \tag{1}$$

Each term in (1) is specified as:

$$\begin{cases} P_{iBase} = \delta_i \\ P_{iUncore} = a_i s_i \\ P_{iCore} = \sum_{j=1}^{N_s} \sum_{k=1}^{L_{vf}} b_{i,j,k} r_{i,j,k} \end{cases} \tag{2}$$

where P_{iBase} denotes ith chassis power consumption excluding the power consumption of its inner servers. $P_{iUncore}$ presents uncore power consumption of the ith chassis which equals uncore power consumption of all ON servers within it. So, a_i and s_i denote the uncore power consumption of a server and the number of ON servers in the ith chassis respectively. In (2), P_{iCore}, $b_{i,j,k}$ denotes power consumption of jth active core in an ith chassis which is set to kth voltage frequency (V-F) level. This scaling is due to process variation that causes different power consumption of servers in any V-F scaling. N_s is the number of servers in a chassis. We define binary variable r such that, $r_{i,j,k}$=1 if jth server in ith chassis runs at kth V-F level and $r_{i,j,k}$=0 for other V-F levels (i.e., for any i and j). The number of ON servers doesn't exceed the number of available servers in a chassis (N_s) automatically with considering the dimension and type (binary) of variable r. Finally, the power consumption of blade servers in N_c chassis is calculated as a vector as follow:

$$P = \delta + a \odot s + C, \quad C = \sum_{j=1}^{N_s} \sum_{k=1}^{L_{vf}} b \odot r. \tag{3}$$

Equation (3) is adopted from [1] which P, δ and C denote $[P_1,P_2,\ldots,P_{Nc}]^T$, $[\delta_1,\delta_2,\ldots,\delta_{Nc}]^T$ and $[C_1,C_2,\ldots,C_{Nc}]^T$ vectors respectively. Operator $\odot$ presents element-by-element matrix and array product. Therefore, $a \odot s$ and $b \odot r$ denote $[a_i s_i]_{N\times 1}$ and $[b_{i,j,k} r_{i,j,k}]_{Nc\times Ns\times Lvf}$ respectively. Array $b \odot r$ contains core power consumption of all chassis; and C is computed as the summation of power of ON servers in all chassis. Finally, the total power consumption of IT equipments (P_s) in a data center is calculated as:

$$P_s = \sum_{i=1}^{N_c} P_i. \tag{4}$$

where P_i represents power consumption of ith chassis.

B. Power Consumption of Cooling Unit

CRAC performance as a cold air supplier of data center room depends on several factors such as outgoing air speed and main material used in construction of CRAC [1].

Energy consumption of CRAC is determined according to Coefficient of Performance (COP) criterion. COP shows the efficiency of CRAC and is defined with the ratio of removed amount of heat (Q_s) by the cooling system to total energy consumption of CRAC (E_{CRAC}) for cooling air process [18]. So, COP is specified as:

$$COP = Q_s / E_{CRAC}. \quad (5)$$

In our work, we use COP model of HP Utility Data Center's CRAC. The COP varies with supplied cold temperature of CRAC (T_s) [18]. The COP model is defined as:

$$COP = 0.0068T_s^2 + 0.0008T_s + 0.458. \quad (6)$$

So the cooling power consumption may be specified as follows [18]:

$$P_{CRAC} = P_s / COP. \quad (7)$$

where P_s is IT power consumption and COP denotes coefficient of performance of CRAC.

C. Total Power Cost

Total power of data center contains IT power consumption and CRAC power that are accounted in part A and B respectively. IT power consumption contains core and uncore power consumption of servers. Therefore, data center power cost is the sum of all chassis and CRAC power consumption without considering power losses due to electrical power conversion and distribution network comprising UPS, AC-DC and DC-DC converters and also the switch gear and conductors [1]. Total power cost (P_{DCTC}) is the summation of (4) and (7) and can be written as:

$$P_{DCTC} = (1 + \frac{1}{COP})\sum_{i=1}^{N_c} P_i. \quad (8)$$

Finally, by replacing (1) in (7), we obtain:

$$P_{DCTC} = (1 + \frac{1}{COP})(\sum_{i=1}^{N_c} \delta_i + \sum_{i=1}^{N_c} a_i s_i + \sum_{i=1}^{N_c}\sum_{j=1}^{N_s}\sum_{k=1}^{L_{sf}} b_{i,j,k} r_{i,j,k}). \quad (9)$$

IV. Proposed Server Placement and Task Assignment Method

In this section, we demonstrate our proposed server placement algorithm based on heat recirculation effects. Then, we present a variation-aware chassis consolidation and task assignment method based on the obtained server placement strategy.

A. Problem Statement

In Fig. 2 we see the design flow for our proposed server placement and task assignment algorithm. We consider the data center specification as an input for our server placement block in the first step. Our server placement algorithm solves the ILP problem in order to select 10% of servers and specifies the optimum placement in such a way that the data center power consumption is minimized in each step. This process continues until 100% of servers are selected and placed. Ultimately, we solve ILP to find optimum task scheduling and consolidation plan for a given workload and placement.

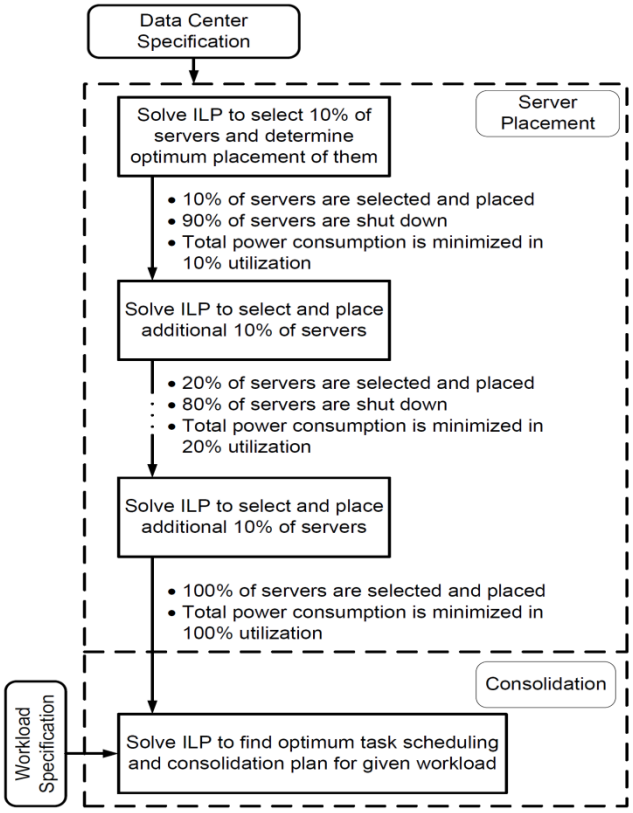

Figure 2. Server placement and task assignment design flow.

B. Server Placement Method

In this section, we formulate our server placement method by using ILP. Since total power consumption of a data center depends on server utilization, finding an optimal server placement for a fixed server utilization rate may not be optimum for other values of server utilization at run-time. Therefore, the objective of the ILP problem is considered to be the average power consumption of data center over different server utilization values in the interval 10%-100%. In order to solve the IPL problem, we use a partitioning technique.

To this end, we first considered a specified percent of data centers utilization (u_i) and obtained the best selection and placement of servers leading to minimized total power consumption of data center including both IT equipment and cooling system power consumption. For example, in a 10% utilized data center, the ILP selects 10% of servers and finds an optimum placement for them. In the next step, considering that 10% of servers are optimally placed, the ILP tries to find the optimum placement for an additional 10% of servers for a 20% utilized data center. This procedure continues until 100 percent utilization (u_n) is achieved and all of the servers are placed in chassis.

The optimal server placement problem for serving different data center utilization rates is as follows:

OBJECTIVE FUNCTION:

$$MIN\left\{P^i_{DCSP} = (1 + \frac{1}{COP})\sum_{j=1}^{N_c}(X^i_j \delta_j + a_j e^i_j + \sum_{k=1}^{S_n} Place^j_{j,k} Pv_k)\right\}. \quad (10)$$

$$for\ i = u_1\ to\ u_n$$

CONSTRAINTS: (11)

$$
\begin{cases}
1.\ Pc_j^i = X_j^i \delta_j + a_j e_j^i + \sum_{k=1}^{S_n} Place_{j,k}^i Pv_k & i = u_1, u_2, ..., u_n\ ,\ j = 1, 2, ..., N_c \\
2.\ T_s^i + D.Pc_j^i \le T_c & \\
3.\ \sum_{j=1}^{N_c} \sum_{k=1}^{S_n} Place_{j,k}^i \le N_s & i = u_1, u_2, ..., u_n \\
4.\ \sum_{k=1}^{S_n} \sum_{j=1}^{N_c} Place_{j,k}^i \le 1 & i = u_1, u_2, ..., u_n \\
5.\ e_j^i = \sum_{k=1}^{S_n} Place_{j,k}^i & i = u_1, u_2, ..., u_n,\ j = 1, 2, ..., N_c \\
6.\ \sum_{j=1}^{N_c} e_j^i = M_i & i = u_1, u_2, ..., u_n \\
7.\ Place^{i+1} - Place^i \ge 0 & i = u_1, u_2, ..., u_{n-1} \\
8.\ X_j^i \le e_j^i \le X_j^i N_s & i = u_1, u_2, ..., u_n\ ,\ j = 1, 2, ..., N_c \\
9.\ Place_{j,k} \in \{0, 1\} & j = 1, 2, ..., N_c\ ,\ k = 1, 2, ..., S_n \\
10.\ X_j^i \in \{0, 1\} & i = u_1, u_2, ..., u_n\ ,\ j = 1, 2, ..., N_c
\end{cases}
$$

where P_{DCSP}^i and Pc_j^i respectively denote total and j[th] chassis power consumption in utilization u_i. Vector Pv shows the server power consumption under process variation effect. Binary variable X_j^i is defined to indicate whether the j[th] chassis is on ($X_j = 1$) or off ($X_j = 0$). Also, binary variable $Place^i$ indicates which servers are placed in different chassis under a specific utilization u_i. On the other hand, e_j is used to specify the number of placed servers in j[th] chassis which does not exceed the server capacity. Parameters S_n and M_i respectively determine total number of servers and number of needed running servers in utilization u_i.

Our goal (10) is to minimize data center total power consumption in each step by determining T_s. As formulated in constraints, several conditions should be taken into consideration. Power consumption is obtained using the variable X in such a way that for each chassis, it determines whether the chassis is on or off. Condition on ($X = 1$) occurs when at least one server is placed in the specified chassis. The second constraint talks about the fact that inlet temperature should not exceed critical value. The third one emphasizes that the number of placed servers should not violate the chassis capacity. Constraint (4) limits each server to be placed solely in one chassis. Constraint (5) is used to determine the number of servers placed in the j[th] chassis and Constraint (6) is used to determine chassis capacity limit. Constraint (7) is used in partitioning technique and guarantees that the optimal solution in utilization u_i is transferred to the next step for placing the rest of servers in utilization u_{i+1}. Constraint (8) states that if a chassis is turned off, no servers can be placed into it consequently.

Equation (10) is a non-linear objective function due to the existence of coefficient *COP* which is a function of T_s. Hence, in order to solve such a problem, we call the ILP solver for fixed values of T_s in the range of 0 to 25 with an accuracy of 0.1. By comparing feasible solutions, the best placement having minimum power consumption is selected as the final solution.

C. Heat Recirculation Model

In this section, first we investigate the effect of heat recirculation on the power consumption of chassis. Then, we introduce a thermal model of the data center based on the heat recirculation which is used in the proposed task assignment method in the next section.

Power consumption is equivalent to the amount of heat that is transferred per unit time. This fact is proved according to the law of energy conservation. Therefore, the amount of heat in a unit of time is carried by an air flow is calculated as [19]:

$$Q = \rho f c_p T. \tag{12}$$

where Q is defined as the heat rate in units of Watt (W), ρ is defined as air density in units of Kg/m^3, f is air flow rate in term of m^3/s, c_p denotes the specific heat of air in units of J Kg^{-1} K^{-1} and T denotes air temperature in terms of Kelvin (K).

Any chassis consumes energy according to law of energy conservation based on the difference of outgoing temperature of hot air (T_{out}) and cold air temperature entering the chassis (T_{in}). So the power consumption of the i[th] chassis (P_i) is defined as follow:

$$P_i = \Delta Q \Rightarrow P_i = \rho f_i c_p (T_{out}^i - T_{in}^i). \tag{13}$$

In order to obtain chassis power distribution, stable condition of chassis temperature and cold air temperature of CRAC is needed for reaching steady state of data center due to heat recirculation among the chassis. So, similar to [1], we assume that the tasks are running for a long period of time on the data center resulting in a stationary temperature profile. This is usually the case for using High Performance Computing (HPC) task such as HSPICE simulation that take several hours or days for running because the time granularity will change due to different incoming workload.

Heat recirculation can be described as the amount of outlet heat rate of a chassis that affects the inlet heat rate of another chassis. The authors in [20] show that the heat recirculation can be defined as a Cross Interference Matrix. Cross interference matrix is denoted as $A_{Nc \times Nc}$ in chassis granularity in which A_{ij} is the coefficient of thermal effect of i[th] chassis on the j[th] chassis. Therefore, the inlet heat rate of i[th] chassis (Q^i_{in}) is calculated as [7]:

$$Q_{in}^i = P_i + \sum_{j=1}^{N_c} A_{ji} Q_{out}^j + Q_s. \tag{14}$$

where Q^i_{in}, Q^i_{out}, P_i and A_{ij} are known and Q_s denotes the cold air rate of CRAC. We can transform heat rate to the temperature using thermodynamic constants in a vector form because, the inlet temperature of any chassis should be less than a specific value (T_c) to avoid overheating and eventually failing servers [10]. So, the relation between temperature and power consumption can be defined similar to [7] as:

$$T_{in} = T_s + D.P, \quad D = [(K - A^T K)^{-1} - K^{-1}]. \tag{15}$$

where T_{in} denotes inlet temperature vector of N_c chassis as $[T_{in}^1, T_{in}^2, \ldots, T_{in}^{N_c}]^T$, P denotes power consumption vector of N_c chassis as $[P_1, P_2, \ldots, P_{Nc}]^T$, and T_s is the column vector with same values in all entries. Matrix K is the $N_c \times N_c$ diagonal one ($K = \text{diag}(K_1, K_2, \ldots, K_{Nc})$) in which $K_i = \rho f_i c_p$. Substituting P with (3) into (15), we have:

$$T_{in} = T_s + D(\delta + a \odot s + C). \tag{16}$$

where T_{in}, should not exceed T_c for all chassis. Our goal is to minimize total power consumption of data center (9) while determining optimum T_s to reduce both CRAC and whole data center power consumption. In this paper, typical T_c is 25°C [7].

D. Chassis Consolidation and Task Assignment Method

We consider the steady state of data center to solve our task assignment method. We assume that the characteristics of HPC tasks are specified as prior knowledge for setting V-F levels of servers. Desired V-F level for each server is determined from domain name servers (DNS) of the demands for instance [1]. Assume that N_l specifies the number of servers put in each V-F level as a problem input. Therefore, the total number of tasks (N_t) is calculated as $\sum_{l=1}^{L_{vf}} N_l = N_t$. We modeled chassis consolidation technique as an ILP formulation to minimize power consumption of data center. Similar to [1], we defined a binary variable Y to determine ON/OFF chassis. If $Y_i=1$, the ith chassis is ON, else ($Y_i = 0$) is OFF. So, the power consumption of chassis is modified as follow:

$$P_i = Y_i(\delta_i + a_i s_i + \sum_{j=1}^{N_s}\sum_{k=1}^{L_{vf}} b_{i,j,k} r_{i,j,k}). \quad (17)$$

For ILP method, we correct the nonlinear relationship in (17) due to production of variables by adding other constraint $Y_i \le s_i \le Y_i N_c$ similar to [1]. This constraint states that if $Y_i=0$, the number of ON servers in ith chassis (s_i) will be zero, else $1 \le s_i \le N_s$. Therefore, $Y_i r_{i,j,k} = r_{i,j,k}$ as well as $Y_i s_i = s_i$ because, $s_i = \sum_{j=1}^{N_s}\sum_{k=1}^{L_{vf}} r_{i,j,k}$. In summary, we formulate chassis consolidation technique for power optimization as:

OBJECTIVE FUNCTION:

$$MIN\{\sum_{i=1}^{N_c}(Y_i\delta_i + a_i s_i + \sum_{j=1}^{N_s}\sum_{k=1}^{L_{vf}} b_{i,j,k} r_{i,j,k})\}. \quad (18)$$

CONSTRAINTS:

$$\begin{cases} 1.\ T_s + D(Y_i\delta_i + a_i s_i + C_i) \le T_c \\ 2.\ Y_i \le s_i \le Y_i N_c & i=1,2,...,N_c \\ 3.\ s_i = \sum_{j=1}^{N_s}\sum_{k=1}^{L_{vf}} r_{i,j,k} & i=1,2,...,N_c \\ 4.\ \sum_{k=1}^{L_{vf}} r_{i,j,k} \le 1 & i=1,2,...,N_c,\ j=1,2,...,N_s \\ 5.\ \sum_{i=1}^{N_c}\sum_{j=1}^{N_s} r_{i,j,l} = N_l & l=1,2,...,L_{vf} \\ 6.\ \sum_{l=1}^{L_{vf}} N_l = N_t \\ 7.\ Y_i \in \{0,1\} & i=1,2,...,N_c \\ 8.\ r_{i,j,k} \in \{0,1\} & i=1,2,...,N_c,\ j=1,2,...,N_s,\ k=1,2,...,L_{vf} \end{cases}$$

where Y denotes ON/OFF chassis as a vector $[Y_1, Y_2, \ldots, Y_{Nc}]^T$.

We do not consider cooling power consumption in the objective function due to the nonlinear relationship in total power consumption (9). To solve this problem, we call ILP solver for each T_s in the range 0 to 25 with an accuracy of 0.1. Then with the comparison of feasible solutions for each T_s, the optimum total power consumption is returned as the solution similar [1].

V. Simulation and Analysis

In this section, we present the simulation environment and experiments. We also demonstrate the simulation results and compare them via specific tables.

A. Simulation Environment

We use GAMS [21] to solve the ILP problems in this paper. The specification and parameters for data centers are as follows:

The physical dimensions of the data center in the experiments are 9.6m × 8.4m × 3.6m which is considered a small scale one. There are two rows in the structure. Each row consists of five 42U racks. Each rack has five chassis with ten server slots. The type for each Blade Server is 7U Dell PowerEdge 1855 dual processor. Hence, the data center has a total number of 1000 single core servers of the same type [6]. There are two V-F levels ($L_{VF} = 2$) except the zero V-F one which among $L_{VF}1$ and $L_{VF}2$, the first one has higher V-F level. Values for power of each chassis and server are illustrated in Table I [1].

The CRAC unit is responsible for providing cold air in the room with the rate of f = 8m^3/s. The flow rate of ith chassis, air density and specific heat of air are considered as f_i = 0.2454 m^3/s, ρ = 1.19 Kg/m^3 and c_p = 1005 J Kg^{-1} K^{-1} respectively [6]. In our process variation model, 40 percent of total core power is related to leakage power [16]. We use systematic die-to-die variation and a Lognormal distribution for the corresponding model [17], [22]. Lognormal distribution parameters are μ = 3.024 and σ = 0.85 which are the mean and standard deviation respectively. These parameters are selected in such a way that the nominal value is the 40% of total power consumption and the leakage power continues up to 20 times the nominal value.

B. Simulation Results

We suppose that each task is assigned only to a certain server. We define different data center utilization rates based on varying number of assigned tasks. For instance, having a total number of 1000 servers, a 40% utilized data center means that 400 servers are executing the same number of tasks [1]. In this paper, we compare total power consumption and cooling cost in the case of Random Server Placement (RSP; several random placements were tested and the average was computed) and Optimal Server Placement (OSP) both having variation-aware chassis consolidation. We consider four different workloads which are determined based on the ratio N_l to N_t. This ratio demonstrates what percentage of tasks works with $L_{VF}1$ and what portion is applied for $L_{VF}2$. Therefore, workload 1 denotes that all tasks (servers) work with $L_{VF}1$. Similarly, workload 2 refers to the conditions in which 50% of tasks work with $L_{VF}1$ and the rest with $L_{VF}2$. The third workload is defined as 10% of tasks work with $L_{VF}1$ and the rest with $L_{VF}2$ and the last one (workload 4) is expressed in such a way that all tasks work with $L_{VF}2$.

TABLE I. VOLTAGE-FREQUENCY PARAMETERS [1]

Voltage-Frequency Level	Chassis Overhead (δ) (W)	Uncore Power of Server (a) (W)	Core Power Consumption (b) (W)
$L_{VF}1$	820	60	25
$L_{VF}2$	820	60	12.5

In the following tables (Table II-V), total power consumption and T_s with power improvement percentage in different data center utilization rates are exhibited for the mentioned workloads. Table II shows more than 13% power saving for less than 50% utilized data center. We obtained up to 14.85% and 8.92% power improvement on average for workload 1. According to table III, under conditions which 50% of tasks work with $L_{VF}1$ and the rest with $L_{VF}2$, results reveal 6.64% power saving on average and up to 9.57% improvement. Table IV demonstrates up to 8.6% power saving and 5.97% improvement on average for workload 3. Finally, Table V specifies 6.24% power enhancement on average and up to 10.13% for workload 4. In general, we observe a descending order of power saving percentage in different data center utilization rates. On the other hand, results show that by using the proposed server placement method, we obtain more power saving in lower utilization rates. This is mainly due to the fact that our server placement method is incremental. This means that we try to place low power servers in the best location available in 10% utilization and continue to place other servers until we reach 100% utilization. Therefore, in higher utilization rates, the remaining servers are all high power servers which are going to be placed in worst locations available. As shown in the results, data center power consumption is even worsened in 100% utilization with respect to conventional random server placement. However, this is acceptable since in typical data centers, the average utilization is generally 20-30% [23].

TABLE II. TOTAL POWER CONSUMPTION AND T_S WITH POWER SAVING PERCENTAGE FOR WORKLOAD 1

Data Center Utilization (%)	RSP T_s (°C)	OSP T_s (°C)	RSP Power Consumption (KW)	OSP Power Consumption (KW)	Power Saving (%)
10	24.1	24.8	17.26	14.69	**14.85**
20	23.82	24.5	34.54	29.95	**13.3**
30	23.39	24.1	52.69	45.79	**13.11**
40	22.73	23.4	71.37	62.49	**12.43**
50	21.68	22.2	91.03	80.71	**11.34**
60	20.35	21	112.04	100.47	**10.33**
70	18.97	19.2	135.19	123.79	**8.44**
80	17.35	17.6	160.97	150.1	**6.75**
90	15.73	15.5	190.69	184.77	**3.11**
100	12.78	11.8	240.97	251.83	**-4.51**
Power Improvement Average (%)	**8.92**				

TABLE III. TOTAL POWER CONSUMPTION AND T_S WITH POWER SAVING PERCENTAGE FOR WORKLOAD 2

Data Center Utilization (%)	RSP T_s (°C)	OSP T_s (°C)	RSP Power Consumption (KW)	OSP Power Consumption (KW)	Power Saving (%)
10	24.07	24.8	15.54	14.05	**9.57**
20	23.62	24.5	31.51	28.51	**9.52**
30	23.4	24.2	47.92	43.36	**9.5**
40	22.98	23.6	64.68	58.84	**9.02**
50	22.11	22.5	82.13	75.47	**8.1**
60	20.94	21.4	100.61	93.12	**7.45**
70	19.6	19.9	120.58	113.09	**6.21**
80	18.26	18.5	142.36	135.06	**5.13**
90	16.76	16.9	166.85	161.24	**3.36**
100	14.55	14.1	200.35	203.33	**-1.49**
Power Improvement Average (%)	**6.64**				

TABLE IV. TOTAL POWER CONSUMPTION AND T_S WITH POWER SAVING PERCENTAGE FOR WORKLOAD 3

Data Center Utilization (%)	RSP T_s (°C)	OSP T_s (°C)	RSP Power Consumption (KW)	OSP Power Consumption (KW)	Power Saving (%)
10	23.95	24.8	14.84	13.58	**8.46**
20	23.79	24.5	30.09	27.5	**8.59**
30	23.51	24.2	45.7	41.78	**8.58**
40	22.98	23.6	61.81	56.61	**8.41**
50	22.18	22.5	78.41	72.44	**7.61**
60	21.17	21.4	95.91	89.35	**6.84**
70	19.89	20	114.69	108.16	**5.69**
80	18.55	18.6	135.22	128.98	**4.61**
90	17.13	17	157.93	153.68	**2.69**
100	14.7	14.2	189.91	193.32	**-1.79**
Power Saving Average (%)	**5.97**				

TABLE V. TOTAL POWER CONSUMPTION AND T_S WITH POWER SAVING PERCENTAGE FOR WORKLOAD 4

Data Center Utilization (%)	RSP T_s (°C)	OSP T_s (°C)	RSP Power Consumption (KW)	OSP Power Consumption (KW)	Power Saving (%)
10	24.22	24.8	14.99	13.48	**10.13**
20	24.12	24.5	30.15	27.29	**9.48**
30	23.75	24.2	45.54	41.44	**8.99**
40	23.24	23.6	61.34	56.15	**8.47**
50	22.29	22.5	77.71	71.85	**7.53**
60	21.16	21.4	95.08	88.61	**6.8**
70	19.91	20	113.63	107.27	**5.6**
80	18.56	18.6	134.03	127.92	**4.56**
90	17.14	17	156.49	152.42	**2.59**
100	14.7	14.2	188.34	191.78	**-1.82**
Power Saving Average (%)	**6.24**				

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a variation-aware server placement and task assignment method for data center power reduction. We introduced an ILP-based server placement method to find the best location of each server in the data center. We also used a variation-aware chassis consolidation technique in the task assignment process to effectively shut down idle chassis in different utilization rates in order to save power. Experimental results showed that by utilizing our proposed server placement method along with variation-aware chassis consolidation technique, up to 14.85% power reduction can be obtained in comparison with conventional random server placement.

As technology scales into deep submicron regimes, the effect of process variation on performance and power consumption of processors increases rapidly. Analyzing the effectiveness of our approach in higher leakage variations is our future work. We also plan to utilize a more precise and structured process variation model based on real measurements.

ACKNOWLEDGMENT

This research is partially supported by grant number 17179/500/T from Research Institute for ICT of Iran. We are grateful for their support.

REFERENCES

[1] E. Pakbaznia and M. Pedram, "Minimizing data center cooling and server power costs," ISLPED09 Proceeding of the 14th ACM/IEEE International Symposium on Low Power Electronics and Design, pp. 145-150, 2009.

[2] R. Sawyer, "Calculating total power requirements for data centers," White Paper, 2004.

[3] Q. Tang, K. S. Gupta, D. Stanzione, and P. Cayton, "Thermal-aware task scheduling to minimize energy usage of blade server based datacenters," 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing, pp. 195-202, 2006.

[4] J. Choi, Y. Kim, A. Sivaubramaniam, J. Srebric, Q. Wang, and J. Lce, "A CFD-based tool for studying temperdture in rack-mounted server," IEEE Trans. Comput, vol. 57, pp. 1129-1142, 2008.

[5] A. H. Beitelmal and C. D. Patel, "Thermo-Fluids provisioning of a high performance high density data center," Distributed and Parallel Databases, vol. 21, no. 2-3, pp. 227-238, 2007.

[6] Q. Tang, K. Kumar, S. Gupta, and V. Georgios, "Energy-Efficient Thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach," IEEE Transaction on Parallel and Distributed Systems, vol. 19, no. 11, pp. 1458-1472, 2008.

[7] Q. Tang, T. Mukherjee, K. S. Sandeep, K. S. Gupta, and P. Cayton, "Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters," Proc. Int'l Conf. on Intelligent Sensing and Info. Process, pp. 203-208, 2006.

[8] S. Borkar, et al., "Parameter variations and impact on circuits and microarchitectures," Design Automation Conference, pp. 338-342, 2003.

[9] "International technology roadmap for semiconductors overview," ITRS, 2008 Update, url: http://www.itrs.net/reports.html

[10] L. Wang, G. Laszewski, J. Dayal, and T. R. Furlani, "Thermal aware workload scheduling with backfilling for green data centers," IEEE 28th Performance Computing and Communications Conference (IPCCC), pp. 289-296, 2009.

[11] F. Wang, C. Nicopoulos, X. Wu, Y. Xie, and N. Vijaykrishnan, "Variation-aware task allocation and scheduling for MPSoC," IEEE/ACM International Conference on Computer-Aided Design (ICCAD'07), Nov. 2007.

[12] M. Momtazpour, M. Goudarzi, E. Sanaei, "Variation-Aware Task and Communication Scheduling in MPSoCs for Power-Yield Maximization," IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences (Special Section on "VLSI Design and CAD Algorithms"), E93-A(12), pp. 2542-2550, Dec. 2010.

[13] M. Momtazpour, M. Goudarzi, E. Sanaei, "Static Statistical MPSoC Power Optimization by Variation-Aware Task and Communication Scheduling," Elsevier Journal of Microprocessors and Microsystems, in press, Feb. 2012.

[14] http://www.sun.com/x64/intel/consolidateusingquadcore.pdf

[15] H. Chang, "Circuit timing leakage power analysis under process variations," A Phd Dissertation of University of Minnesota, Feb. 2006.

[16] G. B. Sirsi, "Leakage power optimization flow," *International Cadence Usergroup Conference*, 2004.

[17] S. Ghosh and K. Roy, "Parameter variation tolerance and error resilency: new design paradigm for the nanoscale era," Proceedings of the IEEE, vol. 98, no. 10, pp. 1718-1751, 2010.

[18] J. Moore, J. Chase, P. Ranganathan, and R. Sharma, "Making scheduling 'Cool': Temperature-aware resource assignment in data centers," Usenix Annual Technical Conference, 2005.

[19] Y. A. Cengel, "Heat transfer: a practical approach," 2nd edition, McGraw-Hill, 2003.

[20] http://www.newservers.com/1855-specs.pdf

[21] R. E. Rosenthal, "GAMS-a user guide," GAMS Development Corporation, 2010.

[22] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical estimation of leakage current considering inter- and intra-die process variation," Low Power Electronics and Design, ISLPED, pp. 84- 89, 2003.

[23] L. Barroso and U. Holzle, "The case for energy-proportional computing," IEEE Computer, Jan 2007.

View publication stats