Saving on Cooling: The Thermal Scheduling Problem

Koyel Mukherjee
University of Maryland
College Park, USA
koyelm@cs.umd.edu

Samir Khuller
University of Maryland
College Park, USA
samir@cs.umd.edu

Amol Deshpande
University of Maryland
College Park, USA
amol@cs.umd.edu

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: modeling techniques; F.2.2 [**Nonnumerical Algorithms and Problems**]: sequencing & scheduling

General Terms

Algorithms, Design, Performance

Keywords

Scheduling, Parallel, Cooling energy, Data centers, Multi-core

1. INTRODUCTION

In this abstract we define some very basic scheduling problems motivated by increasing power density and consequent cooling considerations in data centers and multi-core chips. Modern data centers consist of thousands of computers closely packed in a dense space, typically arranged as hundreds of *racks* of processors. The energy costs of a data center have been compared to that of a small town, with a significant portion contributed by the cost incurred in cooling the machines [1]. The energy cost of cooling is directly driven by the *supply temperature* (denoted T_{sup}) of the cold air being blown in to cool the data center – the incoming air is often kept at a lower than necessary temperature to prevent hotspots from forming since those can damage the hardware. For instance, it has been observed that servers near the top of a rack often run hotter and are subject to higher failure rates [3]. Thermal balancing through judicious task scheduling can lead to fewer hotspots and thus lower overall cooling costs and lower failure rates. Similarly in multi-core chip architectures, the increasing density of cores and a movement toward 3D architectures [7] has made *dynamic thermal management* a key challenge. Increasing temperatures affect circuit reliability and longevity over the long term, and result in increased power consumption (because of increased leakage power) and overall high cooling costs.

In this framework we study a basic scheduling problem, called the *thermal scheduling* problem. We would like to minimize the maximum temperature of the machines in a data center while executing a set of jobs, or maximize assigned jobs while keeping the maximum temperature below a pre-specified *red-line* temperature (T_{red}). The key differentiating factor here from much of the prior work in scheduling is the notion of *spatial cross-interference*: the heat generated by jobs running on a machine raises its own temperature as well as the temperatures of nearby machines due to recirculation effects. Such effects are well-documented both for data centers [6, 5] and for multi-core chips [7]. In addition, the *geometry* of the data center plays a significant role in determining the cross-effect parameters, which are often asymmetric. Although several recent works have recognized the increasing importance of such cross-effects, that prior work (a survey can be found in [4]) has presented either reactive policies that take corrective action, or heuristics without any approximation guarantees. In this work, we initiate a formal study of the problem of thermal scheduling in presence of spatial cross-interference and present analytical results for a 1-dimensional asymmetric model where the machines are arranged in a linear array, with the cold air blowing in from one end.

2. PROBLEM DEFINITION AND MODEL

We base our model on the abstract heat circulation model suggested by Mukherjee et al. [5]. As with much of the prior work on thermal scheduling, we assume that the system is in steady state; i.e., we assume the jobs are long-lived, and analyse the system state when all the jobs have arrived and the temperatures have stabilized. According to the model, the temperature of a machine i is given by: $T_i = T_{sup} + D_i\mathbf{L}$, where D_i is the i^{th} row of the heat distribution matrix D, and $\mathbf{L}$ is the load vector. The vector $\mathbf{L} = \{L_1, \cdots, L_m\}$, where m is the number of machines, denotes the loads on the machines in terms of the *power* consumed by the jobs assigned to the machine. The matrix D represents how the heat or load of any machine j affects machine i (called *cross-interference*). Since the temperature of no machine should exceed T_{red}, we have that: $T_{sup} + \max_{i \in [1,\ldots,m]} D_i\mathbf{L} \leq T_{red}$. Thus, given a set of jobs, our goal is to schedule them so as to either: (a) *maximize* T_{sup} (or equivalently, minimize $\max_{i \in [1,\ldots,m]} D_i\mathbf{L}$), or (b) given a constraint on $\max_{i \in [1,\ldots,m]} D_i\mathbf{L}$, maximize the number of assigned jobs.

For each job, we assume that we can estimate the power that will be consumed to execute it on a machine; this can be computed using the estimated resources required to execute the job, its time duration, and standard system power modeling techniques [2].

The total energy consumption is obtained by adding the energy for processing and the energy for cooling, and can be modeled as: $E = L(1 + \frac{1}{CoP(T_{sup})})$, where $L = \Sigma L_i$ is the total load on all the machines, and CoP (*coefficient of performance*) is a super linear function of the supply temperature. We formulate the problem of thermal scheduling in terms of minimizing what we call the "effective load" on a machine. Effective load on a machine is a linear combination of the load of the machine itself and the load of other machines. Specifically, given that the load of machine i is L_i, and the effect of machine j's load on machine i is captured through the **cross-interference coefficient** D_{ij}, the effective load EL_i is computed as follows: $EL_i = \sum_j D_{ij}L_j$ where $0 \leq D_{ij} \leq 1$ and $D_{ii} = 1$. Our optimization problem of minimizing the maximum

SIGMETRICS'12, June 11–15, 2012, London, England, UK.
ACM 978-1-4503-1097-0/12/06.

temperature can now be seen as minimizing maximum effective load instead, an easier quantity to reason about.

In this abstract, we consider a model of machines in a linear array with the cold air blowing from one end (capturing either a rack in a data center, or a stack in 3D multi-core chips with a heat sink at one end). The i^{th} machine is affected only by the heat recirculated from the machines located below it, closer to the source of the cold air. We number machines from bottom to top, in increasing order from the cold air source. Machine i is only affected by machines $j \leq i$. We assume the heat falls off in an exponential manner. Specifically, the heat felt by a machine i due to machine j is a fraction $\frac{1}{K^d}$ of the load of j, where $d = |i - j|$ is the distance between i and j and K is a constant > 1. More formally, $D_{ij} = \frac{1}{K^{|i-j|}}$. For technical reasons we assume that $K \geq 2$. The *effective load* of the i^{th} machine, where $1 \leq i \leq m$, is then given as $EL_i = \sum_{j=1}^{i} \frac{L_j}{K^{i-j}}$.

3. RESULTS

It is NP-hard to find the optimal *integral* schedule that minimizes the effective load for a given set of jobs or that maximizes the number of jobs assigned without exceeding a certain effective load. We therefore relax the problem to the case when jobs are splittable between machines (i.e., fractional assignments are permitted), find an optimal solution, and then use this solution to devise approximations for the dual problem of maximizing the number of jobs, integrally assigned, given a hard constraint on the effective load. For the 1-D model with exponential heat fall-off, one can show that: $EL_i = L_i + \frac{EL_{i-1}}{K}$. Given this, we can show that:

LEMMA 1. *An optimal strategy for minimizing the maximum effective load for fractional assignments, with total load L, would result in uniform effective load of* $EL_i = EL = \frac{L}{m - \frac{m-1}{K}} \ \forall i$.

The optimal strategy to minimize the effective load for a given load, would be to place a higher load on machine 1 and lower on the rest. Specifically, $L_1 = \frac{L}{m - \frac{m-1}{K}}$ and $L_i = \frac{L}{m - \frac{m-1}{K}} \left(1 - \frac{1}{K}\right)$ for $i > 1$. A non-thermal strategy on the other hand would split load uniformly in minimizing the loads of machines, and this would result in highest effective load on the last machine on the rack from the bottom.

The following theorem states the thermal savings possible.

THEOREM 1. *The reduction in effective load between a thermally aware scheduler and a naive load balanced strategy is* $\geq \frac{L}{m(K-1)+1}(1 - \frac{1}{K})^{m-1} > 0$.

Figure 1 shows the percentage savings in effective load for different values of K and m.

The savings in maximum effective load translate into savings in energy of the cooling system. Let the maximum effective load without thermal aware scheduling be EL_{old} and the one with thermal savings be EL_{new}. Let us assume a simple function for CoP: $CoP(T) = T^{1+\delta}$, for some $\delta > 0$. The following theorem states the amount of energy savings possible.

THEOREM 2. *Let the energy consumed for cooling with a naive strategy of splitting the load uniformly without any thermal awareness be* E_{old}*, and the energy consumed by our thermally aware strategy be* E_{new}*. Let the difference be* $\Delta E = E_{old} - E_{new}$*. The fraction of energy that can be saved in cooling purposes is* $\frac{\Delta E}{E_{old}} > \frac{\Delta EL}{T_{red} - EL_{new}}$*, where,* ΔEL *is the savings in effective load,* EL_{new} *is the maximum thermal aware effective load, and* T_{red} *is the red-line temperature.*

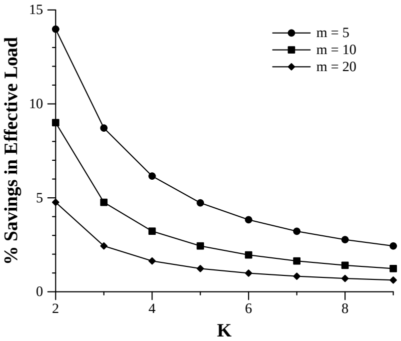

Figure 1: Percentage savings in maximum effective load for varying K and m

Now we look at the dual problem of maximizing the total assigned load, given a hard thermal or effective load constraint, which we assume is identical for all machines. Further, we assume there are no assignment restrictions. In the fractional case, by intelligently scheduling more load on the first machine than the rest, we can schedule more load overall than a non-thermally aware policy for the same effective load constraint. Let this constraint be called effective load capacity c.

LEMMA 2. *Without loss of generality, an optimal policy would assign load c to machine* 1 *and* $c\left(1 - \frac{1}{K}\right)$ *to all machines* $i > 1$ *till it exhausts either the load L or the available machines.*

THEOREM 3. *The fraction of extra load we can assign by a thermal aware strategy is* $\frac{1 - \frac{1}{K^m}}{m(K-1)} - \frac{1}{K^m}$.

The above problem becomes hard when jobs need to be assigned integrally to machines. Further, deriving the structure of the optimum fractional solution becomes much harder when we consider a 2-dimensional model with both vertical and lateral heat coupling. We show the structure of the optimum fractional solution for three different 2-dimensional heat flow models, provide analogous bounds for thermal savings, and also provide approximation algorithms for the integral assignment problem for all three heat flow models [4].

Acknowledgments: This work has been partially supported by NSF Grants CCF-0728839 and CCF-0937865.

4. REFERENCES

[1] C. Belady. In the data center, power and cooling costs more than the IT equipment it supports. *Electronics Cooling*, 2007.

[2] D. Economou, S. Rivoire, and C. Kozyrakis. Full-system power analysis and modeling for server environments. In *Workshop on Modeling Benchmarking and Simulation*, 2006.

[3] M.K. Herrlin. Gravity-assisted air mixing in data centers and how it affects the rack cooling effectiveness. *ITHERM*, 2006.

[4] K. Mukherjee, S. Khuller and A. Deshpande. Approximation Algorithms for the Thermal Scheduling Problem. URL: `http://www.cs.umd.edu/~samir/grant/thermal.pdf`, 2012

[5] T. Mukherjee, A. Banerjee, G. Varsamopoulos, S. Gupta, and S. Rungta. Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. *Computer Networks*, 53(17), 2009.

[6] R. Schmidt, E. Cruz. Raised floor computer data center: effect on rack inlet temperatures of chilled air exiting both the hot and cold aisles. In *The Eighth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, 2002.

[7] C. Zhu, Z. Gu, L. Shang, R.P. Dick, R. Joseph. Three-Dimensional Chip-Multiprocessor Run-Time Thermal Management. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2008.