Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Gonçalo Marques Pestana

Energy Efficiency in High Throughput Computing

Tools, techniques and experiments

Master's Thesis
Espoo, 1 December, 2014

DRAFT! — January 12, 2015 — DRAFT!

Supervisors: Professor Jukka K. Nurminen
Advisor: Zhonghong Ou (Post-Doc.)

Aalto University
School of Science

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author:	Gonçalo Marques Pestana		
Title: Energy Efficiency in High Throughput Computing Tools, techniques and experiments			
Date:	1 December, 2014	**Pages:**	17
Major:	Data Communication Software	**Code:**	T-110
Supervisors:	Professor Jukka K. Nurminen		
Advisor:	Zhonghong Ou (Post-Doc.)		
abstract			
Keywords:	energy efficiency, scientific computing, ARM, Intel, RAPL, tools, techiques		
Language:	English		

Acknowledgements

I wish to thank all students who use LaTeX for formatting their theses, because theses formatted with LaTeX are just so nice.

Thank you, and keep up the good work!

Espoo, 1 December, 2014

Gonçalo Marques Pestana

Abbreviations and Acronyms

2k/4k/8k mode	COFDM operation modes
3GPP	3rd Generation Partnership Project
ESP	Encapsulating Security Payload; An IPsec security protocol
FLUTE	The File Delivery over Unidirectional Transport protocol
e.g.	for example (do not list here this kind of common acronymbs or abbreviations, but only those that are essential for understanding the content of your thesis.
note	Note also, that this list is not compulsory, and should be omitted if you have only few abbreviations

Contents

Scheduling based on dynamic energy pricing

Summary

There are several studies exploring inter data center solutions to lower the electricity bill by leveraging the spacial-time dynamic of energy pricing. The emphasis is given to job scheduling across data centers that are located in different places. The main idea is to exploit the fact that energy prices are change based on location and time. The research community is mostly concerned with fairness, server availability, queue delays, bandwidth costs with job migration and quality of service.. In addition there are several research studies related with migration of cloud computing jobs. Studies that in one way on another address this perspective are [7], [1], [6], [5], [3], amongst others.

Besides inter center solutions, the research community has been addressing the power consumption of the computing nodes specifically from a data center perspective. This perspective is closer to what we are trying to achieve with our solution. For example, one work that seems closer to our solution is [9]. In this study, the authors achieve better energy performance in a dynamic pricing environment with HPC systems by judiciously scheduling parallel jobs - which have different energy profiles - depending on the energy pricing of the moment. The main difference to out solution is that the performance of the machines are not taken into consideration when scheduling the jobs, but rather the job energy profiling.

Another research study that related to our solution is [8]. They came up with an optimal algorithm and two heuristic algorithms to schedule tasks to heterogeneous processors. In addition, they also take into consideration the memory allocation in heterogeneous memory in order to minimize energy consumption while meeting the assumed deadlines. Their work, though, seems to go further than our solution since it considers heterogeneous memory allocation as well. They consider is a computing process executing several tasks in a parallel computing environment. The system consists in a variety of different computational node, each of one with a given number of processors. All computational nodes are connected by a high-speed network. Thus, all the processors can cooperate and realize complementary and parallel tasks. From the energy point of view, the processors of each computational node have an energy profile assigned and have a certain frequency, which will be taken into consideration when scheduling the task. The work dates from end of 2014, which indicates that this is a trendy and hot subject, but it seems that our approach is been used already.

A similar idea has been explored in [10]. They present only heuristic algorithms to schedule tasks on heterogeneous computing systems, based on efficiency and energy consumption. They develop heuristic algorithms due to the fact that an optimal solution for the needed scheduling is NP-complete.

Our solution takes a different perspective when compared with the inter data center solutions. Studies like [8] and [10] do not take the dynamics of electrical pricing into consideration. However, their algorithm is already quite complex and proved NP-complete, to the point they have to come up with heuristic algorithms to apply it in the real world.

Therefore, our approach may have some novelty in a really narrow and still unexplored idea: to develop a scheduling algorithm for heterogeneous HPC that takes into consideration the nodes' energy profile, the dynamic electricity price and also, eventually, the tasks' energy profiling. The algorithm would schedule the jobs in order to minimize the energy consumption and energy bill (note: energy consumption and energy bill are not the same thing), while the deadline is met.

However, there are some open points that we still have might want to consider. First, as [9] mentions, it is important to insure that the hardware existent in the data center is used at its full potential, in order to not waste the investment made when it was purchased. Our solution, though, does not insure that since the idea is to power down/idle machines that are less power efficient in high-peak times. Secondly, from a practical perspective, if we consider only the scheduling between ARM and Intel architectures, it seems not likely that the data center will haver the same software running over both architectures at the same time, give the expertise and investment needed to have the application stack running properly in both architectures (as we witness with CERN's efforts). If we decide to abstract from that point and see the machine's architectures as a black box, then that's not a problem. Thirdly, comparing with other recent research works such as [8], our algorithm model seems to be over simplifying the problem to an extent that might hinder our purposes of creating a practical and energy efficient scheduling algorithm for heterogeneous HPC under dynamic electrical pricing.

General notes

- Intra data center judicious job scheduling based on the heterogeneous architecture of the machines.

- Minimize the electricity costs in data centers by leveraging the dynamical electricity pricing models and heterogeneous computing.

- Online computation schedule the jobs. Jobs are in a queue in a serial fashion and are scheduled depending on the decision of the algorithm at a given time. On the other hand, there are the static scheduling algorithms. These algorithms know all the data they need beforehand and map the jobs to the machines taking that into consideration.

- There are several studies that aim to leverage the potential of geographical load balancing to provide significant cost savings (see [24, 28, 31, 32, 34, 39] in [2])

- Make a problem specification as in [5]

- In our solution, we could also coinsider different stages of functioning such as idling and turning of the machines, depending on the expected workload and the server configuration. The problem might be to understand if we have (or not) knwoledge of the server utilization in the future and its worklaods. Actually, this is an important factor to consider - wether we have or not idea of the future workload.

- It would be interesting to test and simulate our model and algorithm using, for example, workloads and electricity prices in a Google data center. As many studies do, show the potential of our approach by simulating scnarios based on real case data.

- As [9] briefly mentions, does Dynamic Voltage and Frequency Scaling (DVFS) have the same results than our heterogeneous approach in a homogeneous data center ? i.e. are the benefits of a more efficient processor such as ARM surpassed (or the same) as an INTEL working under a DVFS ? The principle seems the same: when energy consumption is smaller (in ARM or INTEL under low DVFS), the jobs take longer to accomplish. This said, is ARM more efficient than INTEL under low DVFS ?

- Should our work be an extension on [9] where, instead of scheduling the workload taking into consideration the job's energy profile, also consider the machine's energy performance ?

Paper's notes

In [4]:

Demand side management are programs implemented by the utility companies to control and influence the user-side behavior. For example, electrical companies often fluctuate the energy price depending on the user's demand.

There is need to encourage household owners to *shift* high demand energy consumptions outside the peak hours, in order to reduce the peak-to-average (PAR) in load demand. [- we aim towards the shifting of schedule different machines depending on the PAR]

Direct load control (DLC) gives the utility companies the possibility to remotely control the household's applications (dim or turn of lights, turn of thermal equipment, amongst others). Though, this model arises some problems related with household's privacy [- see more 'A direct load control model for virtual power plant management' - what if DLC would be implemented for servers and in a heterogeneous scheduling scenario? Would it bing any advantage or liability ?]

An alternative to DLC is smart pricing, where users are encouraged to voluntarily and individually shift their loads out of the peak-hours be increasing the energy prices when the load is big.

One problem with this approach is synchronization: when a large number of users shift their peak at the same time for a low-peak time, the PAR may not be reduced due to the amount of users churning energy at low-peak time. [- this might happen as well with our scheduling strategy. If the amount of users running our scheduling system at the same time is the same, it does not help to reduce the PAR and the prices might get worst]

The paper suggests that households should synchronize their energy usage and schedule their energy applications not only according to the price of the energy at a given time, but also taking into consideration what others are consuming as well. Thus, by acting in synchronization, the group of users can optimize the energy the overall energy consumption and its pricing.

They propose an incentive-based energy consumption pricing model for the smart grid, where the energy source is shared by several users. The meters communicate between each other in a distributed network to find the optimal energy consumption for each user.

Based on game theory, it is shown that through an incentive-based pricing scheme, an optimal scheduling - where users consume less energy and pay less money - can be achieved.

In [7]:

Because of the magnitude of energy costs in data centers, it is important to lower the energy consumption in data centers. The servers are composed of heterogeneous machines from the performance and energy efficiency. In addition, the data centers may be disposed in different geographical locations and, thus, have different energy tariffs. The authors of [7], claim that the key idea to lower the energy bill in data centers is to have energy efficiency servers and schedule the jobs to where energy is more affordable at a given time.

In the context of servers distributed over different geographical locations, it is also important to satisfy fairness and delay constraints. This scenario is less critical when the server is not distributed, as in our case.

In [7], the authors present an online scheduler that distributes batch workloads across multiple data centers geographically distributed. The scheduler aims to minimize the energy consumption of the set of servers having into consideration fairness and delay requirements.

The scheduler is inspired on the technique developed by Lyapunov ['Resource allocation and cross-layer control in wireless networks'] that optimized time-varying systems.

The algorithm takes a queue of jobs schedule them to the different servers having in consideration the (1) server availability, (2) energy price and (3) job fairness distribution. Consequently, the algorithm is tuned to calculate the tradeoff between energy pricing, fairness and queueing delay.

- The model:

The data center model takes into consideration the possibility of the energy prices to vary over time. The state of the data center can be represented at a given time by a tuple of (i) server availability and (2) energy price.

The job model is characterized by a tuple of (1) service demand - job length - and (2) the set of data centers the job can be scheduled.

The scheduler can turn on/off a server when needed. The scheduling is done based on the server availability and job queue and thus, what matters is the energy consumed by the server when it is 'idle' or 'busy'.

The scheduler also considers he model fairness (which is not important to our study, since we focus in a non-distributed server) and queuing delay. Queuing delay defines the time a job will take to start to be processed, according to relation of the number of jobs scheduled and machine availability.

In [7], the scheduler developed takes into consideration the server availability, energy costs, fairness and queuing delay to schedule random jobs arrivals. It opportunistically schedules jobs when (and to where) energy prices are low.

Comparing to our study, though, we do not consider geographically distributed servers but rather, we have schedule the jobs based on the heterogeneous set of machines existing on the server.

In [1]:

This study aims to exploit the temporal and geographical variation of electricity prices, in the context of data centers. They study algorithms to schedule (migrate) jobs in data center based on the energy cost and availability.

When the servers are in different geographical location, costs with data migration have to be taken into consideration, namely bandwidth costs of moving the application state and data between data centers. The bandwidth costs increase proportional to the amount of data migrated between servers.

Their study focuses on inter data center optimization, rather than intra data center optimization (as our study is aiming for)

The algorithm differs from others in 3 major differences: First, they consider migration of batches of jobs. Second, the algorithm has into consideration the future influence of the job scheduling, providing robustness against any future deviations of the energy price. Finally, they also take into consideration the bandwidth costs associated with job migration across data servers.

The main point is to provide a good tradeoff between the energy pricing and the job migration, taking into consideration the bandwidth prices.

Comparing to our study, we do not approach the problem from an inter data center perspective, but rather from an intra data center, by scheduling the jobs to machines depending on their energy performance and the actual energy prices. One interesting idea from this study that can be used, is the usage of an online algorithm that takes into consideration the expected prices and also the actual prices.

In [6]: In [6], they try to systematically study the problems of how minimize the electricity cost in data centers while guaranteeing minimal quality of service. To that end, they take into consideration the local and time diversity of electricity prices.

The contributions are twofold: In one hand, they show that local and time dependent electricity pricing can be leveraged to minimize total energy price of clusters of data centers. On the other hand, they present a mixed-integer optimization formula with linear programming formulation to show that the energy pricing of clustered data centers can be improved under such conditions.

To model the total of electricity costs, they assume that all the servers have a similar power profile - which means that all the servers, disregarding their locations, have the same workload. They calculate the power consumed by the server by multiplying the total of servers at a certain region by the total of workload they have.

Again, the time constraints and delays considered in a inter data center study does not need to be considered in out work.

They obtain the most efficient solution, they approximate an optimization problem through a linear programming formulation and then, convert the linear programming formulation to a minimum cost flow problem.

This work dates from 2010 and don't take into consideration the bandwidth costs of migrating the batches between data centers. Even though that is not an issue in our study, this is taken into consideration in other works such as [1]. Again, it is part of the set of studies on inter datacenter and electrical costs optimizations that location and time based pricing allows.

In [5]: The authors of [5] show that existing systems may be able to save millions of dollars by judiciously schedule workload to servers taking into consideration the temporal and geographical variation of energy prices. The results are based in historical data collected on Akamai's CDN.

In [9]: In [9], the authors leverage the fact that parallel jobs have distinct energy profiles. Taking it into consideration, they study the impact of scheduling jobs according to the energy prices at a given moment and the job's energy profiles. So, the study aims to reduce the electricity bill by scheduling and dispatching jobs according to their energy profile. Their solution has a negligible impact on the system's utilization and scheduling fairness.

Their basic idea is to schedule jobs with low energy profile during on-peak electricity time and, on the other hand, schedule jobs with high energy profile during the off-peak electricity time. In addition, the scheduling is done in such a way that it is guaranteed that there is no degradation of the overall system performance.

The authors take an intra data center approach, since it considers a solution that can be put into practice at a data center level.

The authors claim that "A key challenge in HPC scheduling is that system utilization should not be impacted. HPC systems require a tremendous capital investment, hence taking full advantage of this expensive resources is of great importance to HPC centers.". This may make impractical and wreck our solution, because of the inevitability of turning off (or idle) great amounts of computing resources. Although, internet data centers (cloud data centers) may be a good match to our solution: usually there are much less resources being used at a given time than in HTC computing [need confirmation, partially mentioned in this article].

The scheduling algorithm used places jobs in a time-window. The jobs are chosen to run based on job fairness, job energy profile and energy prices at a given time. A greedy algorithm and 0-1 Knapsack based policy are used to minimize the electrical costs.

Their results show that gains in the order of 23% can be obtained without impact on the overall system.

According to the survey carried by [9], the dynamic energy pricing has been implemented in the biggest markets in Europe, North America, Oceania and China, while Japan was at the time starting to test it on its major cities.

They develop two power aware job policies: 1) greedy approach, where jobs are allocated based on their energy profiles and 2) 0-1 Knapsack based policy, where both job profile and system utilization are taking into consideration.

In [3]:

The authors of [3] present a novel task scheduling algorithm for HPC systems which considers two main points: reducing the energy consuption of the overall system and minimize the schedule length. An HP system is defined by the authors as set of distributed computing machines with differrent configurations connected through a high speed link to compute paralell applications.

They assume a that all the information needed to schedule the taksk is known beforehand. The scheduling algorithm assigns then the jobs to the different machines. Thus, the scheduling algorithm is said to be static, in opposition to, for example, the online algorithms.

One of the particularities of the algorithm is to reduce the impact of duplication-based algorithms. The duplication-based algorithms schedule jobs across machines redudantly, in order to maximize performance by eliminating intercommunication between tasks. However, from the energy consumption point of view, it is not th ideal situation since more than one processor are performing the same job.

Once again, this research work aims at improve the energy efficiency of HPC systems at a distributed level and do not focus, as our approach, on inter data center solutions.

In [8]:

In [8], the authors address the problem of an energy aware scheduling for heteregenous data allocation and task scheduling. The problem consists in finding the best taks scheduling in a heterogeneous system that meet the deadlines while minimizing the energy consumption.

The processors and memories come in different flavors nowadays in HPC systems, making complex the task of efficiently scheadule processor power and memory space in an energy efficient way. The problem of finding an optimal processor and data scheduling becomes critical when trying to minimize energy consumption and meet imposed deadlines.

As the study shows, there are several research efforts tackling the task scheduling problems on heterogenous computing and, most notably for our research, [**?**].

They present an optimal algorithm and two heuristical algorithms to solve the HDATS problem, since the optimal algorithm takes too long to solve problems until 100 nodes. The optimal solution has two phases: First is uses the DFG_Assign_CP algorithm to better map each task to node. Secondly, it choses the data assignment to whose total energy consumed is reduced and the deadlines met.
They consider:

- Heteregenous processors
- Heteregenous memories
- Precedence constrained inputs

- Input/output of each task
- Processor execution times
- Data access times
- Time constraints
- and Energy consumption

When solving the data allocation and task scheduling problem, which is an approach much more solid and complete than ours.

In [10]

The authors of [10] claim that, unfortunatelly, there are not many studies of processor scheduling algorithms that take into consideration both time and energy. In this study, they explore heuristical scheduling algorithms focused on high performance computing and green computing. They work on heuristical algorithms and not in the optimal algorithm, because the optimal algorithm is proven to be NP-complete.

Bibliography

[1] Buchbinder, N., Jain, N., and Menache, I. Online job-migration for reducing the electricity bill in the cloud. In *NETWORKING 2011*. Springer Berlin Heidelberg, 2011, pp. 172–185.

[2] Liu, Z., Lin, M., Wierman, A., Low, S. H., and Andrew, L. L. Greening geographical load balancing. In *Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems* (New York, NY, USA, 2011), SIGMETRICS '11, ACM, pp. 233–244.

[3] Mei, J., and Li, K. Energy-aware scheduling algorithm with duplication on heterogeneous computing systems. In *Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing* (Washington, DC, USA, 2012), GRID '12, IEEE Computer Society, pp. 122–129.

[4] Mohsenian-Rad, A.-H., Wong, V., Jatskevich, J., Schober, R., and Leon-Garcia, A. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *Smart Grid, IEEE Transactions on 1*, 3 (Dec 2010), 320–331.

[5] Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J., and Maggs, B. Cutting the electric bill for internet-scale systems. *ACM SIGCOMM Computer Communication Review 39*, 4 (2009), 123–134.

[6] Rao, L., Liu, X., Xie, L., and Liu, W. Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment. In *INFOCOM, 2010 Proceedings IEEE* (2010), IEEE, pp. 1–9.

[7] Ren, S., He, Y., and Xu, F. Provably-efficient job scheduling for energy and fairness in geographically distributed data centers. In *Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on* (2012), IEEE, pp. 22–31.

[8] Wang, Y., Li, K., Chen, H., He, L., and Li, K. Energy-aware data allocation and task scheduling on heterogeneous multiprocessor systems with time constraints. *Emerging Topics in Computing, IEEE Transactions on 2*, 2 (June 2014), 134–148.

[9] Yang, X., Zhou, Z., Wallace, S., Lan, Z., Tang, W., Coghlan, S., and Papka, M. Integrating dynamic pricing of electricity into energy aware scheduling for hpc systems. In *High Performance Computing, Networking, Storage and Analysis (SC), 2013 International Conference for* (Nov 2013), pp. 1–11.

[10] Zeng, G., Yu, L., and Ding, C. An executing method for time and energy optimization in heterogeneous computing. In *Green Computing and Communications (GreenCom), 2011 IEEE/ACM International Conference on* (Aug 2011), pp. 69–74.

Appendix A

First appendix

This is the first appendix. You could put some test images or verbose data in an appendix, if there is too much data to fit in the actual text nicely.