Workload–Based Power Usage Modeling

Christopher Hundt Ariel Kleiner

Introduction

As highlighted by growing interest in both academia and industry, the electrical power usage stemming from computing hardware has become a great concern. This issue of power usage is particularly acute for heavy users of large-scale computational resources such as data centers. Unnecessarily high or ill-distributed power usage can lead to increased costs, heat generation, and environmental damage. One can imagine workload and power usage modeling and prediction being used to better allocate servers and manage their use within a data center or even across multiple data centers. To successfully achieve this goal, it will be necessary to have an understanding of the power consumption behavior of individual servers and other computing hardware. Thus, we adopt this perspective and focus on the power usage characteristics of a single server.

Prior Work:

- Power consumption as a function of system utilization metrics (e.g., [1, 2])
- Contributions of individual hardware components to power consumption (e.g., [3])

Our idea: Model power consumption as a **direct function of the workload** to which the server is subjected, rather than adopting a lower-level view.

Results

The Model

Input: Average rates (in requests per second) of requests of nine different types

Output: Predicted average power usage, computed as an **affine function** of the rates

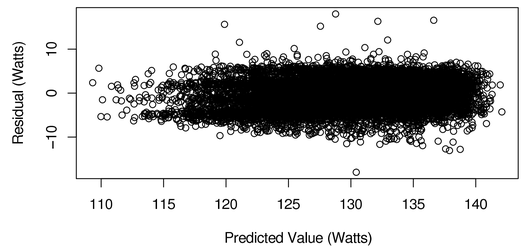
Training data: Power usage data from 30,000 synthetically generated inputs

Test data: Synthetic and real-world

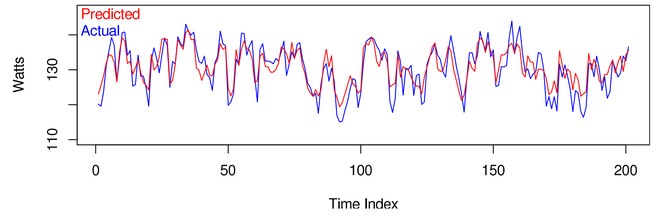
Learned Coefficients

Intercept	108.95374
0-10k	0.03017
10k-25k	0.04166
25k-200k	0.12315
0-10k (PHP)	0.07381
10k-25k (PHP)	0.08848
0-10k + small array sort (PHP)	0.07303
10k-25k + small array sort (PHP)	0.08632
0-10k + large array sort (PHP)	0.07476
10k-25k + large array sort (PHP)	0.08186

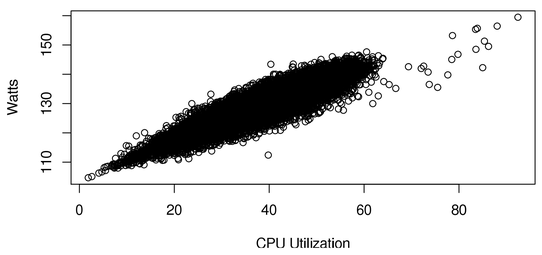
Model Residuals on Synthetic Test Data



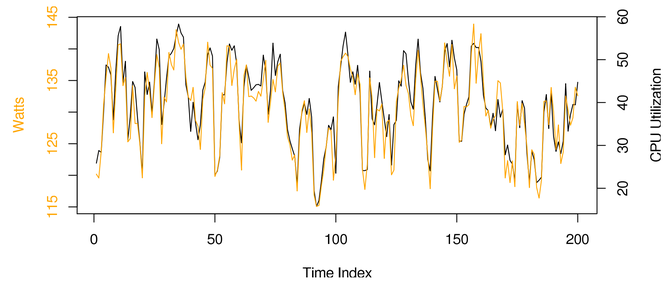
Prediction Accuracy Over Time (Synthetic Test Data)



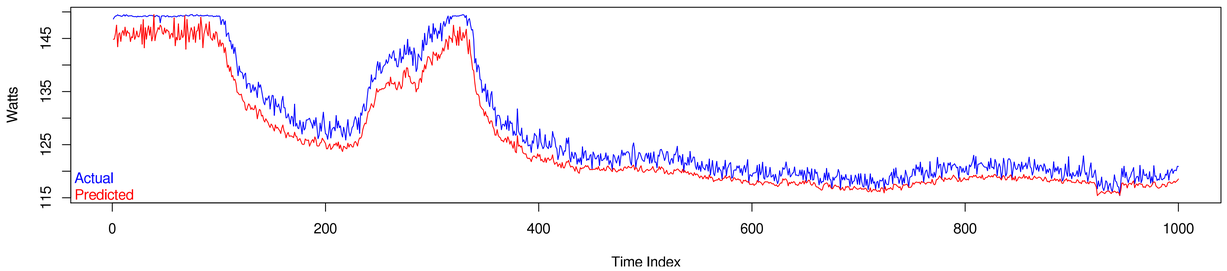
Synthetic Test Data: CPU Utilization vs. Power



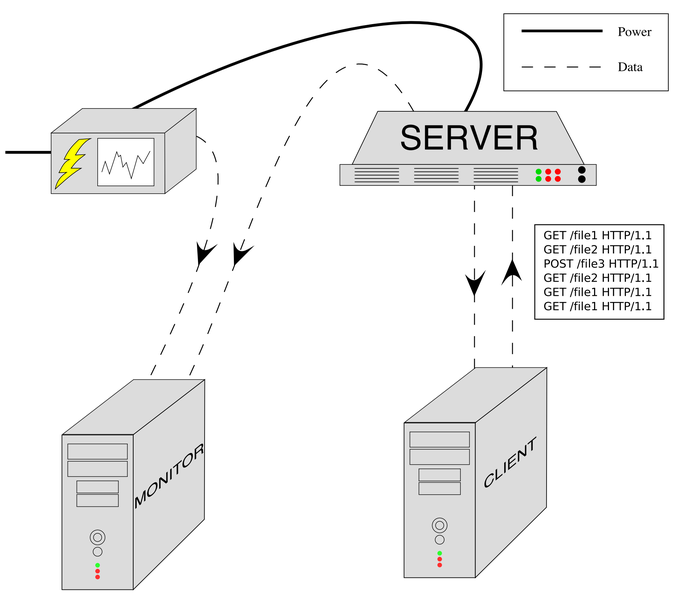
CPU Util. vs. Power Over Time (Synthetic Test Data)



Prediction Accuracy Over Time (World Cup Data, 10x Time Quantum)



Experimental Setup



Server Hardware

- Processor: 2 × 1 GHz Pentium III
- Memory: 1.5 GB ECC PC133
- Hard Drive: 2 × 36 GB IBM UltraStar 36LZX, RAID 1
- Network: Intel PRO/1000F Server Adapter

Request Types

- Size of file transferred: small, medium, large
- Computation performed: none, some, a lot
- Combinations of the above

Workload Generator (Synthetic)

- Input: Rate of each type of request
- Effect: Requests sent to server at desired rates

Real-world Data

- Trace from the 1998 World Cup Website
- All GET requests

Discussion

The above results indicate that a simple linear model can capture the dependence of server power usage on workload, which we define in terms of client request rates. This observation is somewhat surprising because it is not a priori clear that one should expect different workloads to compose additively in their power consumption. It follows that, beyond powering down as many machines as possible, little improvement in power consumption can be gained through further manipulation of the allocation of requests among machines. Furthermore, the ability to predict power usage from workload specifications rather than from system metrics is valuable in that the former are the true determining quantities of the system behavior. Our model coupled with a workload prediction module would allow prediction of expected power consumption.

Sources of Error

- Finite power meter accuracy
- Low measurement frequency (sampling error)
- Time synchronization between monitor and power meter

Future Work

- Refine analysis of World Cup data
- Use more sensitive power measurement equipment
- Further investigate workloads near saturation point
- More complex workload and request types (e.g., requests requiring database queries)
- Experiment on different server hardware
- Active and online learning

Conclusions

- A simple linear model effectively predicts server power usage under synthetically generated workloads.
- This model performs promisingly when applied to power consumption data from a real-world workload.
- Further work will be required to refine our model and investigate various extensions of it.

Acknowledgements

We thank Jon Kuroda and Mike Howard of the RAD Lab support team as well as Peter Bodik for their invaluable help in obtaining and managing the hardware resources used for this work. Furthermore, we are grateful to Michael Armbrust for his sharing of facilities for simulating workloads from the World Cup 1998 web site. Finally, Armando Fox and Michael Jordan provided valuable help in obtaining resources and evaluating ideas.

References

[1] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan. Full-system Power Analysis and Modeling for Server Environments. In *Workshop on Modeling, Benchmarking, and Simulation (MoBS)*, June 2006.

[2] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan. Zesti: Full-System Power Modeling and Estimation. `http://www.hpl.hp.com/personal/Partha_Ranganathan/papers/2006/SUBMIT_2006_hpca_zesti.pdf`. Under review, June 2006.

[3] V. Pandey, W. Jiang, Y. Zhou, and R. Bianchini. DMA-Aware Memory Energy Conservation. In *Proceedings of the 12th International Symposium on High-Performance Computer Architecture (HPCA 12)*, Feb. 2006.