CODY: Characterizing Power and Energy Usage with Resource Auto-Configuration in Cloud Computing

Jiacheng Jason He, Ziming Zhang, Dr. Song Fu
Dependable Computing Systems Laboratory
Department of Computer Science and Engineering, UNT

Introduction

Ever since the Internet was created in the 1980s, the world has been about sharing resources through the most efficient means. The latest innovation connecting the world is **cloud computing**. Cloud computing is essentially, for clients, the ability to access files, programs, and software from a remote location via network connection. As long as clients continue to pay their fees, they receive their rights to a portion of a storage server at a massive data center.

What does this mean for hosts of these cloud servers? **Data centers**, the hearts of large-scale, real-world cloud infrastructures, contain servers available to clients from all over the world. Within a data center there are multiple **control nodes**, **data nodes**, and **storages**. Control nodes are responsible for receiving information from client computers and directing the data nodes in following through with instructions.

Data nodes are like worker bees, constantly buzzing about, completing tasks quickly. Finally, storage databases hold all task results and relay them back to the client. The factor that determines how quickly the nodes process tasks is hardware configuration. **Configuration, though, also affects power usage**. There are many factors that can be configured, so the best options must be found.

One data center uses ten times as much power as the entire city of Denton in one year. Using power meters to find total power usage of large-scale infrastructures is an impossible demand, so finding a model to estimate usage is a necessity.

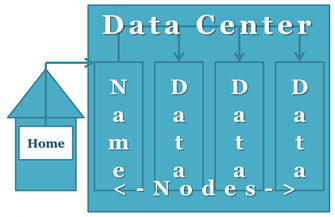
Objectives

To **estimate power consumption effectively via a power model**, two objectives had to be met.

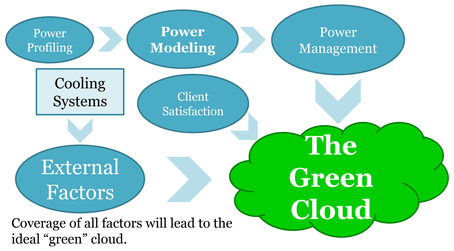
1. **Power Profiling**: Testing a training microbenchmark to gather data on CODY, the infrastructure.
    1. The first part is to compare the effect of different CPU core distributions between the two data nodes on the power usage.
    2. The second part is to compare the effect of different memory distributions between the two data nodes on the power usage.
    3. The third part is to compare the effect of different I/O distributions between the two data nodes on the power usage.
    4. The fourth part is to compare the effect of different CPU frequencies of all nodes on the power usage.
2. **Power Modeling:** Analyzing the collected data from the test bed servers to create a model that can estimate total power usage, with three of the four as variables in the model. This will remove the necessity of the power meters in further experimentation.

Abstract

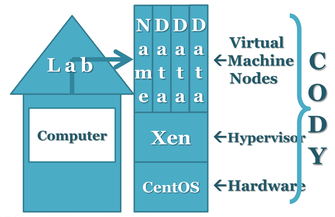
As cloud computing—a recent technology that enables worldwide communication—develops faster and faster, one remaining problem that remains unsolved is power management. One factor that plays a major role in power usage and program runtime is resource configuration. Four settings, CPU core allocation, memory allocation, I/O utilization, and CPU frequency, each affect power usage. Two steps were taken to effectively create a power model that can be applied to calculate power usage for large-scale cloud infrastructures. To simulate a large-scale environment, 16-server test bed was set up with multiple virtual machines running on each node. Different configurations for each were tested with microbenchmarks and power meters for power profiling purposes. After testing a variety of configurations multiple times, certain trends could be identified. The next step was then implemented: power modeling. A linear regression model is proved to be the best approach in estimating the total power usage with any configuration. The model conceived by this research is proved to be extremely accurate (over 90%), thus making it relevant in future discussions of cloud power saving.



A model of an industrial cloud computing data center.



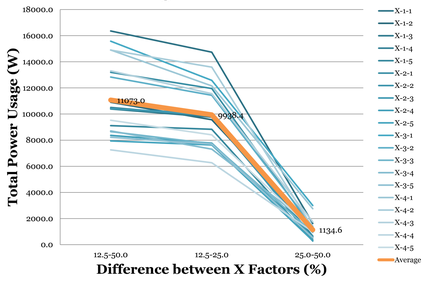
Coverage of all factors will lead to the ideal "green" cloud.



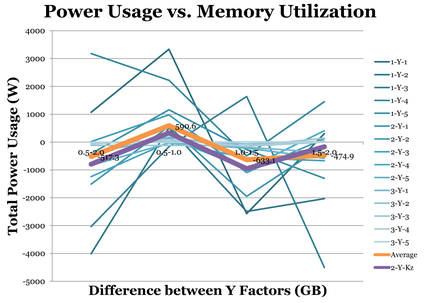
CODY, the lab's model of a large-scale cloud infrastructure.

Objective 1: Power Profiling

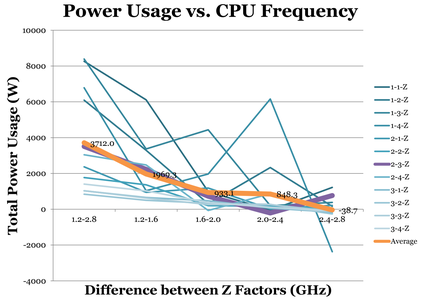
Power Usage vs. CPU Utilization



Power Usage vs. Memory Utilization



Power Usage vs. CPU Frequency



Power Usage vs. CPU Allocation
Min memory utilization, min CPU frequency

CPU Allocation (%)	Time Elapsed (s)	Total Power Usage (W)
12.5	421	21029.3
25.0	121	6297.7
50.0	87	4674.4

Power Usage vs. CPU Allocation
Max memory utilization, max CPU frequency

CPU Allocation (%)	Time Elapsed (s)	Total Power Usage (W)
12.5	205	13175.5
25.0	67	4765.6
50.0	51	3652.6

Power Usage vs. Memory Allocation
Min CPU utilization, min CPU frequency

Memory Allocation (GB)	Time Elapsed (s)	Total Power Usage (W)
0.5	421	21029.3
1.0	357	17694.1
1.5	407	20265.2
2.0	400	19960.9

Power Usage vs. Memory Allocation
Max memory utilization, max CPU frequency

Memory Allocation (GB)	Time Elapsed (s)	Total Power Usage (W)
0.5	50	3645.5
1.0	51	3659.8
1.5	51	3652.2
2.0	51	3652.6

Power Usage vs. CPU Frequency
Min CPU utilization, min memory utilization

CPU Frequency (GHz)	Time Elapsed (s)	Total Power Usage (W)
1.2	421	21029.3
1.6	282	14916.5
2.0	246	13926.9
2.4	228	13983.9
2.8	192	12761.2

Power Usage vs. CPU Frequency
Max CPU utilization, max memory utilization

CPU Frequency (GHz)	Time Elapsed (s)	Total Power Usage (W)
1.2	96	5058.6
1.6	71	4033.6
2.0	60	3658.9
2.4	54	3548.7
2.8	51	3652.6

Objective 2: Power Modeling

The power profiling portion of the project is crucial in creating a training set of data to improve the power model.

$$Pw(k) = w_{CPU} \cdot U_{CPU}(k) + w_{MEM} \cdot U_{MEM}(k) + w_{IO} \cdot U_{IO}(k) + \varepsilon$$

After the initial power model is created, it is tested on more advanced benchmarks. The training benchmarks are purposely designed to train each part of the model: the CPU utilization's effects, the memory utilization's effects, the I/O utilization's effects, and the CPU frequency's effects. The testing benchmarks, though, are random, complex benchmarks to evaluate the overall effectiveness in a realistic environment.

The model is proved to be very effective, as seen in the figure above. At points of inaccuracy, the model's estimation is only a small degree higher, which is better than underestimating the power.

Cpu Freq	w_{CPU}	w_{MEM}	w_{IO}	ε
2.8GHz	55.30	18.35	8.34	55.60
2.4GHz	37.47	16.84	7.55	50.97
2.0GHz	22.78	12.45	6.62	50.71
1.6GHz	18.56	10.50	3.12	47.32
1.2GHz	12.15	6.25	1.45	46.99

Conclusions

Total power usage of a cloud computing infrastructure can effectively be modeled using linear regression. Any large-scale infrastructure can apply a similar approach and technique to estimate and control power consumption. Furthermore, any similar infrastructure can implement reinforcement-learning techniques to adapt to optimize power usage.

References

Zhang, Z., Qiang, G., He, J., Fu, S. "Adaptive Power and Performance Management with Resource Auto-Configuration in Cloud Computing Systems ", accepted by *Journal of Communications*, pp.1-12, 2013.

Zhang, Z., Fu, S. "macropower: A Coarse-Grain Power Profiling Framework for Energy-Efficient Cloud Computing", in *Proceedings of the 30th IEEE International Performance Computing and Communications Conference (IPCCC)*, 2011.