Automatic Server to Circuit Mapping with The Red Pills

Jie Liu
Microsoft Research
One Microsoft Way, Redmond, WA 98052
liuj@microsoft.com

Abstract

As fine grained power monitoring becomes crucial in data centers, a challenge raises on how to correctly map server identities to power circuits. Power provisioning, power capping, and power tracking all depend on accurately accounting which server consumes power from which circuit. Manual survey is cumbersome and error prone. We describe a solution, called Red Pill, that can systematically and automatically identify the mapping. The idea is to generate a power consumption pattern (a.k.a. a signature) by controlling CPU utilization, and to reliably detect it from circuit-level power measurements. We describe our implementation of the Red Pill system and evaluate it with real data traces.

1 Introduction

A modern data center for Internet services or cloud computing can host hundreds of thousands of IT devices and consume tens of megawatts of electricity. Power represents a fundamental parameter for data center capacity, capital investment and operation cost. Due to workload fluctuation, electricity price changes, and dynamic provisioning, the power management mechanisms in data centers are becoming increasingly sophisticated [7, 2, 3, 5], involving over-provisioning, power capping, load shifting, and adaptive cooling control.

However, one question has largely been ignored by the research community. That is how to verify the power distribution network in a data center.In particular, identifying the *server-to-circuit* (S2C) mapping is hard at the data center scale. A large data center has thousands of racks. Each rack can have up to four power strips. For power load balancing and reliability reasons, these power strips are connected to different power distribution units (PDUs), which are then balanced over multiple UPS systems and sometimes different power grids.

S2C mapping is essential to data center safety and efficiency for several reasons:

- **Circuit load balancing**: Power consumptions vary significantly depending on the server types and their workloads. While the PDU output 3-phase electricity, each server is only plugged into a single phase. Balancing out power consumption across different phases improves power utilization efficiency and reduces power failure risks.
- **Over-subscription and power capping**: Servers typically never reach their nameplate power consumption. Over-subscribing power capacity [5, 6] is a way to improve data center infrastructure utilization by plugging more servers than the circuit can handle had all the servers reach their peal power consumption. However, due to asynchronous load fluctuation, the circuit is safe most times. How much to oversubscribe depends on the power profile of the servers and the applications running on them. Furthermore, when the power consumption on a circuit is close to the limit, we need to know which subset of the servers to notify so they can properly throttle down the power consumption, either by shifting workload or by slow down computation. Since throttling degrades server performance, it is desirable to only apply to the servers corresponding to the endangered circuit.
- **Fail-over analysis**: Many servers have multiple power supplies and sit on multiple circuits for reliability and redundancy reasons. If one of the circuit breakers is triggered, the entire load of those servers on that circuit will fall to the other circuits. When the power capacity margin is squeezed by over-provisioning, it is important to make sure that this kind of fail-over does not cause cascading effects and bring down more circuits.

Despite its importance, S2C mapping is usually imposed by wiring conventions and then recorded by hand.

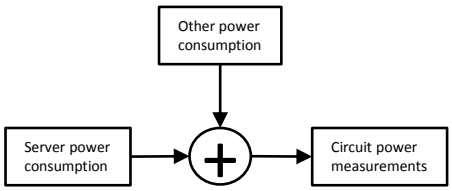

Figure 1: A channel abstraction of the power distribution network.

The sheer volume of wires and the number of servers make the manual survey process time consuming, error prone, and difficult to maintain. While errors introduced by the manual process may be tolerable in the past when data center power capacities are over provisioned and loosely managed, it cannot meet the accuracy requirement for future data centers where power redundancy and margin are tightly controlled.

Data center equipment providers, such as Raritan and APC, use Ethernet capable power strips, usually called smart PDU, to bring addressing capability to each power plug. However, it introduces high capital investment, and the confusion at server power cord level is still not solved. RFID and IP addressable power supply (inside servers) are possible solutions. But there is no successful products on market due to extreme high cost when apply at the data scale.

In this paper, we describes a method, called Red Pill, for automatically discovering server-to-circuit mapping without relying on any additional hardware, except for power meters at various tiers of the power distribution system, which are typical for modern data centers. Our idea is to control the servers to generate distinct power signatures that can be detected from the circuit power measurements. This mechanism requires no capital investment, can be completely automated, and only needs to run once during the servers' "burn-in" process (or any other time on demand).

In the rest of the paper, we describe the basic principle in section 2 and our design of the Red Pill system in section 3. In section 4, we evaluate our system with real data traces.

2 Basic Principle

Our basic idea of discovering S2C mapping is to view a server's power cord and the circuit it plugs into as a communication channel, as shown in Figure 1.

If we can generate a known signal (a.k.a. a *signature*) from the specified server, which is then detected by one, and only one, of the circuit power meters, then we can

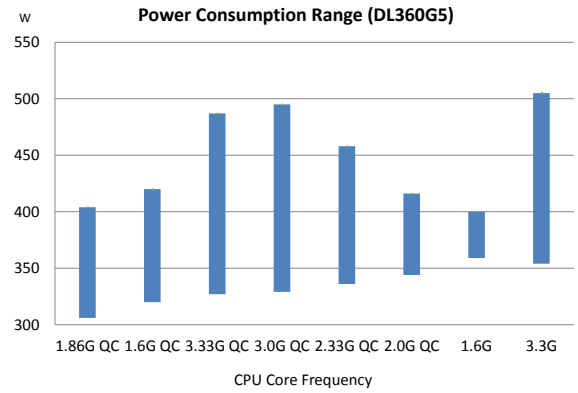

Figure 2: Power consumption ranges for some HP DL360G5 server models.

determine the mapping between the server and the circuit. Since the circuit power meters can only measure power consumption, the signature we expect to generate must be a power signature.

It is well known that a server's power consumption depends on its workload. When a server is idle, it consumes roughly 60% of its peak power [3]. On top of that, the total power consumption changes dynamically depending on its CPU power states, CPU utilization, disk IO activity, network card traffic, fan speed, etc. Figure 2 shows the power consumption range for a set of HP DL360G5 servers[4]. We see that power consumption variation can be as large as 200W on a single server.

The server inevitably shares the circuit with other equipment, and the power meter only measures the total power consumption at the circuit level. So, the signature we generate is subject to disturbances due to other activities shown on the power distribution network.

Using power line as a communication media is not new. For example Ethernet over power line and X10 are mature products and standards. However, in order not to increase hardware cost, we rely completely on off-the-shelf power meters, which are not designed for high speed communication. In fact, most meters can only sample/report data for less than 10 samples a minute. Thus, there are several challenges when realizing our basic principle.

- **Signature generation**: what kind of power signatures is desirable, i.e. easy to generate and resistance to disturbance?
- **Signal/noise ratio (SNR)**: how accurately can we detect the signature at the power meter end? That is, can the signature be significant enough to sustain the disturbance from power consumption traces

caused by other servers sharing the same circuit?

- **Detection efficiency**: How much time does it take to perform one detection experiment? Can we perform multiple detection experiments at the same time?

In the rest of this paper, we describe the Red Pill system that addresses these challenges.

3 The Red Pill System

In the Red Pill system, an agent is installed in every server, which, in response to a command, can run a stress program to alter the power consumption of the server with the goal of generating a predefined signature. For a multi-corded server, this also means that the signature will be spread into multiple circuits, which is then detected by the power meters.

3.1 Signature Selection

An ideal power signature should be easy to generate, easy to detect and short in length. In addition, if the signals are from a family of *orthogonal* signals (see, e.g. [1]), we can perform multiple experiments in parallel and reduce overall mapping time.

These requirements motivate us to look at various multiplexing techniques in communication theory, in particular frequency devision multiplexing (FDMA) and code devision multiplexing (CDMA). Although CDMA is known to use bandwidth more efficiently than FDMA, periodic signals and their frequency responses are attractive due to its simplicity, and it is easy to control the duration. Since normal server operation rarely exhibit periodic power consumption behaviors over a long period of time, if we insert an artificial periodic power consumption signal, it should be easily detectable from circuit power measurement traces through Fourier analysis.

However, a pure (single frequency) signal requires the server to generate a power trace that resembles a sine wave. This is hard to achieve since there are other activities on the system, e.g. background OS tasks and thread scheduling that are not under users' control. Alternatively, a square wave is both easy to generate, and exhibit a strong peak in the frequency domain. To illustrate this, Figure 3 compares the DFT spectrum of a sine wave and a square wave with the same amplitude, frequency, and length.

3.2 Signature Detection

To detect the signature, we perform hypothesis testing on the energy exhibits at the specific frequency. Let k be the Fourier index corresponding to the frequency of

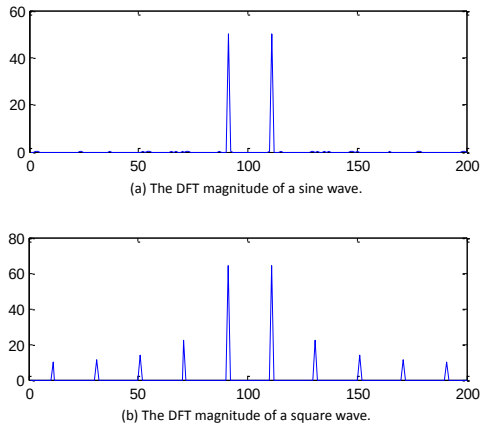

Figure 3: The frequency analysis of a sine wave and a square wave with same frequency.

the square wave $x = \{x_0, x_1, ...x_{N-1}\}$ generated by the Red Pill agent, and $T = N * s$ be the duration of the signature, where s is the sampling rate of the circuit power meter and N is the number of samples the meter collected during time T.

Then, by Discrete Fourier Transform (DFT),

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi kn}{N} i} \quad (1)$$

Clearly, the longer the signal N, the stronger is $|X_k|$. In fact, the energy of the signature $|X_k|^2$ is proportional to $|x|^2 N$, where $|x|$ is the amplitude of the square wave, i.e. the maximum dynamic power consumption we can generate.

Let Y_k by the random variable for the k-th Fourier coefficient of the circuit-level power measurement. Then for each circuit i we can formulate hypothesis testing with:

- H_0: Server j is not plugged into circuit i, and
- H_1: Server j is plugged into circuit i.

Our goal is to minimize Type I error, which implies that a server is wrongly assigned to a circuit. (Type II error implies we fail to detect the signature from any circuit and have to repeat the test). In other words, given threshold α, we want to find the threshold on X_k such that $Prob(Y_k > X_k) \leq \alpha$. For example, if Y_k is a Gaussian random variable $N(\mu, \sigma)$, then $\alpha = 0.01$ implies $X_k > 3\sigma$, the commonly known 3-sigma rule. In reality, when server operations are independent, the Gaussian assumption is reasonable during a short duration.

3.3 Implementation

Figure 4 shows an overview of the Red Pill system implementation. An RPAgent is installed in every server, which listens at a TPC port. The agent, upon receiving a command for generating a square wave with period p and duration D, stresses CPU of all cores accordingly. The system is controlled by the RPManager, which can send commands to each RPAgent under its control and collect power meter measurements from each circuit.

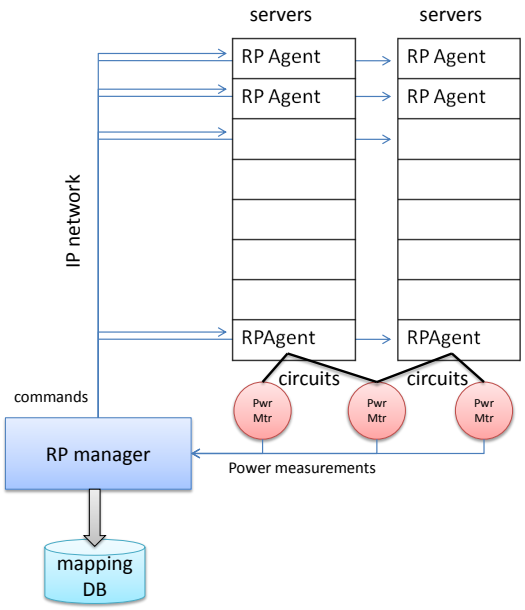

Figure 4: The implementation of the Red Pill System

To improve detection reliability, the RPManager performs Fourier analysis to collect statistics of the coefficients at various frequencies. When the RPManager needs to test a server, it picks a frequency that is not significant in the measurements. After that, it collects the power measurements from all circuit meters and starts the detection process by applying FFT (or match filters) to each signal of length D and perform statistical hypothesis testing. RPManager looks at the magnitude of the controlled frequency, and compares it to the standard deviation of the rest of the frequency components, after removing the DC component. If the magnitude of the signature frequency is three times the standard deviation, then we call it significant, and claim a detection. When multiple circuits exhibits high value at the corresponding coefficient, the RPManager can pick the highest one, or to improve reliability, the RPManager may perform multiple independent tests and merge the results by a voting mechanism. The recorded is stored in a database.

4 Evaluation

We deployed the Red Pill system in an experimental data center, which has a break circuit monitoring system at rack level granularity. During normal operation, the circuit monitoring system record about 4KW of power consumption sampled every 15 seconds. A typical trace is shown in Figure 5.

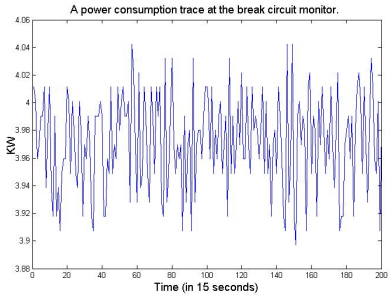

Figure 5: A power consumption trace at a circuit breaker.

We first evaluate how long the square wave signature should be for reliable detections. Figure 6 shows the detection success rate as a function of the signature lengths and the amplitudes of the signature. Every point is the success rate averaged from 33 experiments. In these experiments, we control the amplitude of the signature by exciting the CPUs: on one machine, we excited one core and two cores to get 25W and 50W amplitude respectively. On another one core machine, we obtained 40W amplitude. The period of the wave form are 2 minutes (or 8 samples), i.e. one minute of idle and one minute of full utilization.

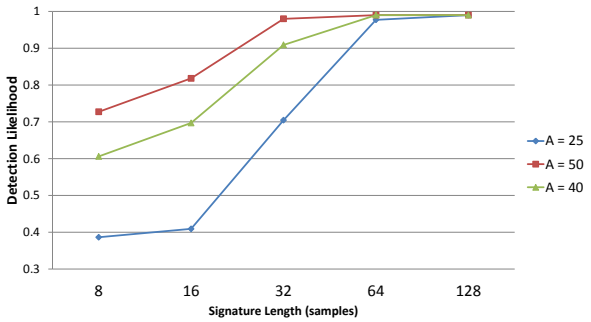

Figure 6: The detection successful rate as a function of signature length.

It is clear that longer signals and higher amplitude gives higher detection rates. However, longer signals mean longer time for the mapping process. In practice, one may want to start with a shorter sequence, e.g. 32 samples for a machine that can generate 40W power am-

plitude, and increase the signature length only if the hypothesis testing does not yield a confident result.

It is interesting to point out that we also experienced abnormal behaviors when a large number of servers, in this case almost half of the rack, were rebooted due to software upgrade. This causes a huge downward spike in the power trace, as shown in Figure 7 (a). Such spikes have a wide spectrum in the frequency domain 7 (b). Clearly, we cannot detect the signature.

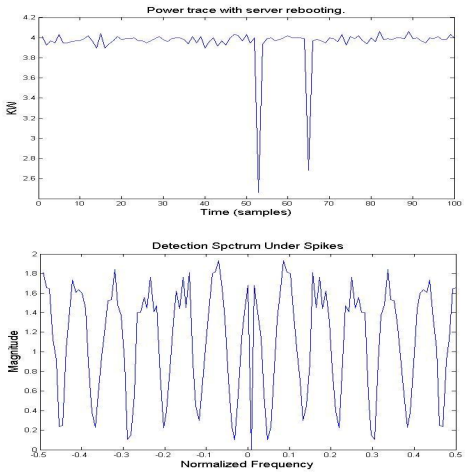

Figure 7: A rack-level power measurement trace when a large number of servers had a reboot, and its FFT magnitudes.

This example shows the importance of tolerating abnormal behaviors, e.g. by abandoning the process if we observe abnormal spikes during the detection phase. It also shows that Type II errors can happen in the detection process and experiments must be repeated.

Next, we evaluate the scalability of the approach by simulating the detection process using data traces from Windows Live Messenger. We use 32-sample signatures with amplitude 40W, and try to detect the server among a number of 20 to 100 servers. We randomly select the targeted number of the servers among the Messenger cluster and repeat the experiments.

Figure 8 shows the results averaged over 40 experiments. The more servers on the circuit, the more noise they add to the collective measurements, and the likelihood of successful detection decreases. These situations call for increasing the signature length when the total number of servers on the circuit is high.

5 Conclusion

Facing the challenge of accurately mapping servers in side a data center to their corresponding circuits, we design the Red Pill signature generation and detection system. We borrow ideas from communication theory and view the power consumption relationship between individual servers and the circuit power meters as communication channels. By exciting power consuming components on the specified server, we generate a signature that can be detected among power measurements. We proposed a simple square wave signature and studied its detection reliability using real traces.

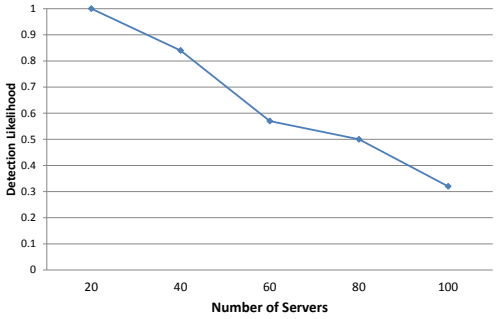

Figure 8: The likelihood of detecting a single server (40W) among a number of servers.

S2C mapping is a building block to perform power failure analysis and fine-grained power control. As part of the future work, we seek ways to deploy the tool at a data center scale and study data center power failure characteristics.

References

[1] BARRY, J. R., MESSERSCHMITT, D. G., AND LEE, E. A. *Digital Communication: Third Edition*. Springer, 2003.

[2] CHASE, J. S., ANDERSON, D. C., THAKAR, P. N., VAHDAT, A. M., AND DOYLE, R. P. Managing energy and server resources in hosting centers. In *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles* (New York, NY, USA, 2001), ACM, pp. 103–116.

[3] CHEN, G., HE, W., LIU, J., NATH, S., RIGAS, L., XIAO, L., AND ZHAO, F. Energy-aware server provisioning and load dispatching for connection-intensive internet services. In *NSDI'08: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation* (Berkeley, CA, USA, 2008), USENIX Association, pp. 337–350.

[4] COMPANY, H.-P. D. Hp power calculator utility: a tool for estimating power requirements for hp proliant rack-mounted systems, 2007.

[5] FAN, X., WEBER, W.-D., AND BARROSO, L. A. Power provisioning for a warehouse-sized computer. In *ISCA '07: Proceedings of the 34th annual international symposium on Computer architecture* (New York, NY, USA, 2007), ACM, pp. 13–23.

[6] LEFURGY, C., WANG, X., AND WARE, M. Power capping: a prelude to power shifting. *Cluster Computing 11*, 2 (2008), 183–195.

[7] RAGHAVENDRA, R., RANGANATHAN, P., TALWAR, V., WANG, Z., AND ZHU, X. No power struggles: Coordinated multi-level power management for the data center. In *ASPLOS'08* (New York, NY, USA, 2008), ACM.