Pay as You Go in the Cloud: One Watt at a Time

Kayo Teramoto*
Yale University
New Haven, CT, USA
kayo.teramoto@yale.edu

H. Howie Huang (Advisor)
George Washington University
District of Columbia, USA
howie@gwu.edu

***Abstract*—Advancements in virtualization have led to the construction of large data centers that host thousands of servers and to the selling of virtual machines (VMs) to consumers under a per-hour rate. This current pricing scheme employed by cloud computing providers ignores the disparities in consumer usage and in its related infrastructural costs of providing the service to different users. We thus propose a new pricing model based on the liable power consumption of the VM, which we correlate to the VM's proportion of CPU and disk I/O usage. We evaluate the fairness and practicality of our accountable power consumption model on various machines and storage types. We then demonstrate the benefits of the proposed pricing model by looking at four consumer cases. Our work is undergoing further experimentation and we hope to expand our testing using cloud services.**

***Index Terms*—cloud computing; power consumption; pricing; virtual machine;**

I. Introduction

The introduction of virtual machines (VMs) and cloud computing in the market has raised the need for a fair pricing that appropriately reflects the nature and costs of the service: providing computing resources to consumers through a data center in the form of VMs. The pricing model for household utilities, such as water and electricity, is based on the amount of the utility used; measured per gallon and per kilowatt-hour respectively. Ideally, a similar single resource based pricing model would be practiced by cloud computing services. However, the numerous types of resources involved and the varying usages of those resources by consumers complicate the distribution of costs.

Furthermore, while resource usage can be measured with a meter or system monitor on physical computers, resource usage is less transparent for virtual machines due to nature of sharing physical resources among several VMs. A perfect pricing model for cloud computing that accounts for every resource is consequently difficult to implement. Yet, a fair pricing model is necessary for survival and growth of the VM market.

In this paper, we propose a new pricing model that is more fair and more reflective of the data center's costs compared to the per-hour pricing model currently in practice. We first develop a robust VM power consumption model that can be used to infer its accountable power consumption from the measured resource utilizations, with small error. We then use the power model to propose our pricing model where, as a VM's power consumption increases, the price gradually increases. At some point, the dynamic power-based price will cross over the hour-based price, where the more highly utilized VMs will pay more than what they would currently pay, accounting for higher hardware and energy costs. Finally, we demonstrate the effectiveness of our pay-as-you-go pricing model through a comparison of consumer cases.

II. Background

Presently, cloud computer providers such as Amazon EC2 charge consumers by the full hour by rounding up the time from when the VM is launched to when it is shut down. This per-hour rate varies depending on a number of factors. In the case of Amazon EC2, the VM price depends on the allocation size of the VM (e.g., small, large, and extra large), the data center location (e.g., US east, EU, and Asia), the type of purchase (e.g. on-demand and reserved), and the software (e.g., Linux and Windows) [1]. This VM price is mainly determined by two major components: the amortized capital cost and the operating cost of the data center. To build these data centers, the cloud providers invest a significant amount of money into the facility, the hardware (e.g., servers, networking gears, and storage), power-related equipment (e.g., distribution, backup generators, and batteries), and cooling equipment. Note that different equipment will have various depreciation periods. For example, the typical period is 15 years for the facility and 3 to 5 years for the servers. These up-front investments will need to be accounted for in the per-hour VM prices. On the other hand, besides the administrative costs, the power and cooling costs constitute most of the operating expense of a data center. The power related cost is estimated to contribute to a significant percentage of the overall monthly cost, which is even more than the amortized infrastructure and server costs [2].

The varying per-hour rate considers the variety in VM specifications, but overlooks the differences in resource usage and the related infrastructural costs when running different workloads on similar VMs. Clearly, the current system gives an advantage to consumers running applications that use large amounts of resources. Providing for such a VM is more costly due to the large expenditure in power and equipment, but because the per-hour pricing is independent of resource usage (though dependent on the amount of resource allocated), less active VMs using a small percentage of hardware and

[*] Kayo Teramoto did this work as a summer 2012 undergraduate research assistant at GWU.

consuming less power must share the burden of providing for the more active VMs. This could potentially shrink the market for cloud computing as VMs become a less cost-efficient investment for consumers seeking to run less resource intensive applications.

Furthermore, the per-hour rate adds a layer of obscurity from the view of the producer. Since time does not correlate with resource usage, the expected adjustment in market price in response to change in resource price is complicated. Implementing a proportional increase in per-hour rate as a response to an increase in energy costs is unlikely to be the most rational and fair way of matching revenue with production costs. It is also well known that energy costs vary depending on several factors, which include the facility location, the type of source, and the time of day.

To address this problem, we propose to replace the per-hour rate with a power based pay-as-you-go per-watt-hour rate. Power has been shown to dynamically change with resource usage [3] and, as mentioned previously, is a major data center cost, making it a good marker of differences among VMs.

The main advantages of an power-based rate are:

- A fairer division of cost based on VM power consumption is achieved, which is a more accurate estimate for the service consumers receive. That is, consumers will pay a price that is proportional to the cost of providing the service to them. Note that the rate will still depend on location, software, and other VM specifics since capital costs and quantity of service sought by consumers will differ.
- Cloud computing providers' profit can become more stable. With a per-hour-rate, Amazon's profit may decrease when resource usage is maximized (due to less flexibility in VM consolidation). With our model, Amazon can ensure a guaranteed profit from each user and more easily adjust for changes in resource cost. Depending on the cost of VM provision, a heavy user may be charged a premium for above-average power consumption.
- Consumers pay only for their accountable power consumption. With partial hour billing available with the power-based rate, consumers will no longer have to pay for unused minutes (e.g., Amazon EC2's full-hour billing [1]).

Previously, the lack of a means to measure VM power consumption made it impossible to implement an power-based rate. Clearly, a physical metering device cannot be connected to virtual object, in this case a VM, which would provide the most direct way of measuring power consumption. However, recent work [3] has shown that virtual machine power consumption can still be calculated at a high degree of accuracy with readily available data.

III. Power Consumption and Pricing Models

We will now discuss our proposed pricing scheme by presenting our pricing and the power consumption models. The overall scheme is presented in Figure 1.

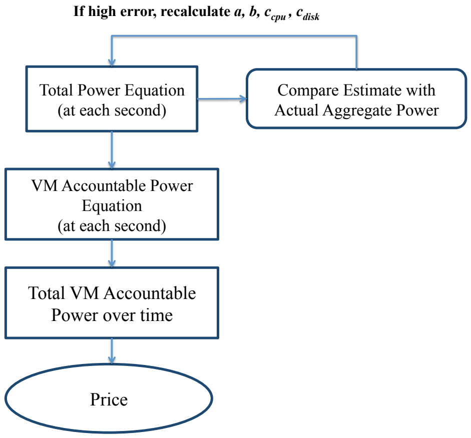
If high error, recalculate a, b, c_{cpu}, c_{disk}
Total Power Equation (at each second)
Compare Estimate with Actual Aggregate Power
VM Accountable Power Equation (at each second)
Total VM Accountable Power over time
Price

Figure 1. Calculating Total Power Consumption Per VM

A. Pricing Model

The pricing for virtual machines should be simple, yet reflect the amount of data center cost for which a VM is held accountable–a complicated and esoteric pricing model would risk losing consumers unfamiliar with the cloud computing jargon and create a entry barrier for the average consumer. Hence, the model proposed has a dynamic charge based completely on a VM's accountable power consumption, which we attain by summing the accountable power $AccPower_{VM,i}$ at each second from start time $t = 0$ to shut down $t = x$. It is unreasonable to assume that a VM's power consumption will remain constant over the entire session, hence we find the power consumed at every second rather than take samples.

$$Price_{VM,i} = r \cdot (\sum_{t=0}^{x} AccPower_{VM,i}) \tag{1}$$

i is the VM ID number, r is the dollars-per-watt-hour rate, and $AccPower_{VM,i}$ is the VM's accountable power consumption.

A baseline cost could be added to the pricing model; however, we argue that high power consuming VMs should pay a greater portion of the constant costs since the data center is providing greater services to them than to the low power consuming VMs. Furthermore, a baseline cost would disadvantage consumers who use their VMs for short time periods. We also note that many utility companies charge with only a per-resource rate (e.g., per-gallon) without the addition of a base charge and allow that dynamic rate to account for the expense of maintaining and running the facility.

B. Power Model

The practicality of using the power based pricing model depends upon the existence of an accessible and accurate per-VM power metering system. Here we propose a new model that excludes Dom0 (the host domain that manages the system) and partitions all of the power among the guest VMs in proportion to their CPU usage (CPU) and number of read and write requests ($Disk$).

$$\begin{aligned} TotalPower = a\sum_{i=1}^{n}(CPU_{VM,i}) + c_{cpu} \\ +b\sum_{i=1}^{n}(Disk_{VM,i}) + c_{disk} + Idle \end{aligned} \tag{2}$$

a and b are CPU and disk coefficients, c_{cpu} and c_{disk} are the respective constants, and $Idle$ is the idle power of the physical machine (while still logging the data).

The coefficients and constants are found using the $TotalPower$ equation. From here, the accountable power of VM $i = 1...n$ at a given second is calculated in proportion to the other VMs running on the machine:

$$\begin{aligned} AccPower_{VM,i} = a \cdot CPU_{VM,i} + b \cdot Disk_{VM,i} + \frac{Idle}{T} \\ + \left(\frac{CPU_{VM,i}}{\sum_{i=1}^{n}(CPU_{VM,i})}\right)c_{cpu} \\ + \left(\frac{Disk_{VM,i}}{\sum_{i=1}^{n} Disk_{VM,i}}\right)c_{disk} \end{aligned} \tag{3}$$

T is the number of VMs of a specific instance/size that can be run on one physical machine.

The number of VMs running on a machine can vary: a VM running an application that utilizes 10% of the CPU could share a machine with 9 other low resource-utilizing VMs or with 1 resource-intensive VM depending on scheduling and the workloads of the other VMs. Dividing the idle power equally among all VMs sharing a machine would thus be unfair since the consumer has no say in whether his VM is running on a machine with many other VMs or a few other VMs; the VM is always placed with the goal of achieving maximum scheduling. We hence divide the idle power by the total number of similar VMs that could share a machine for fairness, and T is dependent on the type of VM instance and size (e.g., standard small, high-memory extra large, and high-CPU medium) that a consumer chooses to purchase.

Since many of the unobserved resource states are correlated to CPU and disk I/O usage, the model recalculates the coefficients and constants when errors in estimated total power increase (generally from workload changes) and cross a predetermined threshold.

IV. Evaluation

In evaluating our proposed pricing model, we first assessed the accuracy of our power model and then compared our pricing model with the current model.

A. Experimental Setup

Hardware:

We tested the power model on three configurations: a Dell machine with 2.93 GHz Intel Core2 Duo E7500 processor, 4GB RAM, and a 3.5-inch Samsung HDD, the same Dell machine with a 2.5-inch Octane SSD, and a low-power ZOTAC machine with 1.6GHz Dual Core ATOM N330 processor, 2GB RAM, and a 2.5-inch Samsung HDD. We connected the power cable to a Watts up? Pro ES power meter to measure the actual total physical power consumption.

Software:

On each of the Samsung and Octane drives, we installed the 12.04 Ubuntu desktop and Xen 4.1.0 hypervisor. Guest VMs were created using the installed virt-manager application. Per-VM CPU and disk I/O data were collected at every second using a perl script.

Benchmarks:

To simulate varied workloads, we installed SysBench and ran the CPU, memory, and file I/O benchmarks on each of the VMs.

B. Accountable Power Model

We looked at both the aggregate power and individual VM power to validate our power model.

Aggregate Power:

To calculate the resource specific coefficients (a and b) and constants (c_{cpu} and c_{disk}), we first simultaneously ran a varied workload of Sysbench benchmarks on each of the VMs and collected the CPU, disk I/O, and total physical machine power data at one second intervals. The CPU and memory benchmark led to high CPU usage while file I/O led to high disk I/O and moderate CPU usage. Both CPU and disk I/O seemingly affect the power in the same way: the power measured when only CPU is very high is within 1-2 watts of the power measured when disk I/O is high and CPU usage is moderate.

After collecting the data, we excluded times when one or more of the data values were not captured. The first and last 3 seconds of every benchmark run were also excluded due to the sudden power jumps and falls at the beginning and ends of benchmark runs that sometimes lacked a corresponding jump or fall in CPU or disk I/O (which could be attributed to a data capture delay of a few milliseconds between the power meter and perl script).

Linear regression using MATLAB yielded the coefficients. We compared the estimated aggregate power consumption with the actual total power measured by the Watts Up? meter. We verified the specific coefficients and constants in the equation by re-running the same workload on the VMs and then running different workloads. When the estimated aggregate power was compared with the measurements from the power meter, all three machines yielded average errors of less than 1.5% (see Table I).

TABLE I
Aggregate Power Error

	Avg Error	Std Dev
2.5” Samsung	0.0087	0.0073
3.5” Samsung	0.013	0.013
2.5” Octane	0.0055	0.0076

These low average errors were achieved without adjusting the aggregate power equation when the benchmarks changed or when the error increased. We can assume that with a self-updating equation, the error will be even lower.

Per-VM Accountable Power:

We specifically looked at the Octane SSD on the Dell machine in evaluating the per-VM Accountable Power model because of it's relatively low idle power in comparison to its maximum power (≈44.7W and $<$ 64W, whereas for the 2.5-inch Samsung on the ZOTAC board: ≈27.8W and ≈31W when running 4 VMs). A proportionately low maximum power makes it difficult to assess the accuracy of the per-VM accountable power since CPU and disk I/O usage has less measurable impact on the power. Furthermore, the ZOTAC board lacks the capabilities of hosting many VMs without causing a decline in performance due to its low memory and few computing units.

Unlike our aggregate power model, there lacks a ground truth to our per-VM accountable power model, in part because the power consumed by Dom0 is partitioned among the guest VMs in our model. Even if a 100% accurate VM power meter existed, our model does not look at a VM's actual power consumption, but gauges it's accountable power consumption. However, we reason that our model is fair because we calculate a VM's accountable power consumption relative to the other VMs sharing the physical resources. By keeping the coefficients a and b and the constants c_{cpu} and c_{disk} the same for all of the VMs sharing the resources at any given time, a VM's resource usage and power can be directly compared to those of the other VMs. 40% CPU usage by one VM consumes the same amount of power as another VM consuming 40% CPU on the same machine at the same time. We thus believe our partitioning of power among the VMs is fair.

To further validate the aggregate power consumption model as well as the per-VM accountable power model, we ran a benchmark on one VM while allowing the other three VMs to idle (shutting down the three VMs would cause the idle power of the physical machine to decrease, so we left them idling). Since there is only one VM actively running and consuming resources, it is fair to make that VM liable for the power difference between the aggregate power and the machine's idle power. We set $TotalPower - Idle$ as the accurate VM accountable power in this case and graphed it alongside the estimated accountable VM power (see Figure 2). We recalculated the coefficients with every benchmark, and the estimated power values were fairly accurate since the estimated values landed on or near the actual power values. In Table II, we show the average difference between the actual and estimated VM accountable power in watts, the average error, and the standard deviation collected for each benchmark run on VM_1 on the Octance SSD while the other VMs idled.

All 4 benchmarks yielded an average difference of less than 1W. While the fileio run benchmark had a 20% average error, this can be attributed to the short length of the benchmark (only 10 seconds) which didn't allow the power to settle, the small amount power consumed (one tenth difference between small numbers yield high errors), and a small time offset between the power meter and the CPU and disk I/O data logger. Nonetheless, the average difference was under 1W and the actual and estimated sum of power consumed differ only by 3.5W (48.3W and 44.7W respectively).

TABLE II
BENCHMARK TESTS

Benchmark	Avg Diff (W)	Avg Error	Std Dev
CPU run	0.068	0.0066	0.0035
Memory run	0.083	0.0075	0.053
Fileio prepare	0.11	0.0083	0.0073
Fileio run	0.81	0.20	0.18

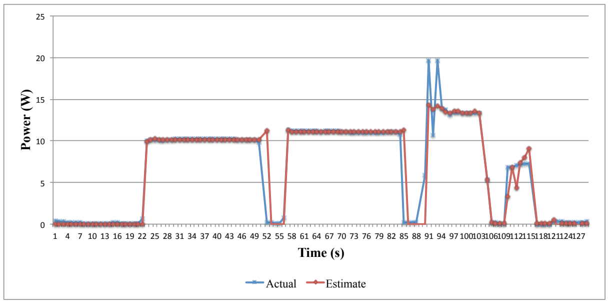

Figure 2. VM Power Consumption: Actual vs. Estimate

Further improvements in VM power modeling can improve our model. However, we conclude that our accountable power consumption model is accurate to a high degree, fair, and ultimately usable for our pricing models.

C. *Comparison of Pricing Models*

To evaluate the impact of the new proposed pricing model, we compare four consumers running identical VMs in the same data center for 200 hours with usages specified in Table III.

TABLE III
CONSUMER USAGES

Consumer	CPU %	IO Req. No.(%)
User1	5	417 (5)
User2	30	2505 (30)
User3	55	4594 (55)
User4	80	6682 (80)

As shown in Fig. 3, with the per-hour model, each of the users would pay the same amount at the end of the month ($16 with the rate of $0.080/hr for a standard on-demand small instance). In contrast, using the SSD models, User4 is consuming 37.3% more power than User1 and shall consequently be charged 37.3% more ($4.87 difference when using a $r = \$0.00142/W \cdot hr$ which was calculated on the assumption that providers expect a $64 total revenue from 4 VMs running an average of 50% CPU and 50% disk I/O).

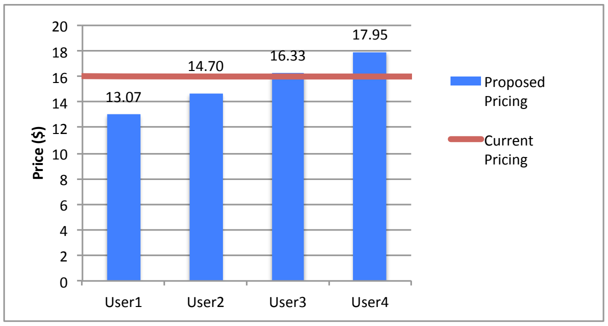

Figure 3. Current Pricing vs. Proposed Pricing

From the 4 case comparison, we confirm the impact a per-watt-hour rate pricing can have on consumer billings. Consumers with the same VM instance will be charged different amounts and this differentiation will hold a positive correlation with the quantity of power they consumed. Overall, the power-based pricing will lead to fairness and greater transparency in billings.

V. Contribution

The proposed per-watt-hour pricing model will lead to greater fairness in the market for cloud computing and VM services. For consumers, their bill will be more reflective of the costs of their received service. On the other side of the market, the producers–both resource providers and cloud computing providers–can also benefit from the energy-based pricing model.

The use of a per-watt-hour rate allows price differentiation to occur. In the context of cloud computing services, identical VMs of same instance size, location, and other fixed factors, would be offered on the market at different prices that depend on individual consumer usages. For electrical utilities, the fixed costs are greater than the variable costs. Hence, for many electrical companies, it costs a negligible amount per kilowatt-hour delivered. This means that it is beneficial for electric companies to pursue a strategy of maximizing the usage of their fixed capital, that is, to recruit as many customers as possible or to increase the energy consumption of their customers.

Data centers are one of the many consumers of the electrical utilities market. Increase in electricity demands by data centers would benefit the electricity companies following the argument presented above. Likewise, data centers would profit from following a similar price discrimination strategy. Among potential users of VMs are the price-sensitive consumers who would like to receive the service, but may not be able to afford or are skeptical of the amount of computing work they can receive from the per-hour rate. There are also consumers who have the capacity to pay more but are not paying as much as they should be. Under a per-watt-hour rate, there may be some consumers that pay less than the amount they would pay under a per-hour rate, but the data center would overall still profit from the maximization of their capital that would result from the increase in consumers. In economic terms, this is a decrease in dead weight loss.

We note that Amazon does attempt to maximize capital usage by offering spot instances, which allow consumers to bid for spare instances [1]. However, this practice can be made more profitable with an energy based rate. Since power consumption can be measured at a finer rate than per-hour, Amazon could take in revenue even when a customer's instance is prematurely interrupted. Its current practice is to not charge consumers for the interrupted hour, but with a power-based rate, they would not be constrained by the large time granularity of the hour. Furthermore, consumers would not necessarily be penalized by this practice since they already make savings by bidding for prices.

In summary, we proposed a per-watt-hour pricing model based off of our per-VM power consumption model. We argued that the division of resource costs would be fairer and that producers could profit from a rate more reflective of resource cost. This research demonstrates the application of technological and scientific advances to world markets in such a way as to benefit consumers, producers, and the economy as a whole. Research in computing can raise the level of fairness practiced by market participants that would not be possible otherwise. The exchange of goods and services in the industrial world is reliant on prices, and the state of the economy is crucial to the stability and survival of a state. Hence, the development and study of fairer pricing models is an important field of research.

Future work will further evaluate our accountable power consumption model and experiment using cloud services. Additional variables, such as network usage, may also be considered in future VM power consumption models to improve accuracy. Finally, various pricing models can be explored, such as a second degree price discrimination model that vary the rate by quality of resource consumed. This could serve as an incentive for consumers to integrate virtual machines into their work, and various other pricing schemes should be studied.

VI. Acknowledgment

This work is in part supported by the National Science Foundation grant OCI-0937875.

References

[1] Amazon EC2, http://aws.amazon.com/ec2/.

[2] C. Belady, "In the data center, power and cooling costs more than the it equipment it supports," *Electronics cooling*, vol. 13, no. 1, p. 24, 2007.

[3] A. Kansal, F. Zhao, J. Liu, N. Kothari, and A. A. Bhattacharya, "Virtual machine power metering and provisioning," in *Proceedings of the 1st ACM symposium on Cloud computing*, ser. SoCC '10. New York, NY, USA: ACM, 2010, pp. 39–50. [Online]. Available: http://doi.acm.org/10.1145/1807128.1807136