Usage Patterns and the Economics of the Public Cloud

Cinar Kilcioglu
Uber
San Francisco, CA
ckilcioglu16@gsb.columbia.edu

Justin M. Rao
Microsoft AI & Research
Redmond, WA
justinmrao@outlook.com

Aadharsh Kannan
Microsoft AI & Research
Redmond, WA
akannan@microsoft.com

R. Preston McAfee
Microsoft AI & Research
Redmond, WA
mcafee@microsoft.com

ABSTRACT

We examine the economics of demand and supply in cloud computing. The public cloud offers three main benefits to firms: 1) utilization can be scaled up or down easily; 2) capital expenditure (on-premises servers) can be converted to operating expenses, with the capital incurred by a specialist; 3) software can be "pay-as-you-go." These benefits increase with the firm's ability to dynamically scale resource utilization and thus point to the need for dynamic prices to shape demand to the (short-run) fixed datacenter supply. Detailed utilization analysis reveals the large swings in utilization at the hourly, daily or weekly level are very rare at the customer level and non-existent at the datacenter level. Furthermore, few customers show volatility patterns that are excessively correlated with the market. These results explain why fixed prices currently prevail despite the seeming need for time-varying dynamics. Examining the actual CPU utilization provides a lens into the future. Here utilization varies by order half the datacenter capacity, but most firms are not dynamically scaling their assigned resources at-present to take advantage of these changes. If these gains are realized, demand fluctuations would be on par with the three classic industries where dynamic pricing is important (hotels, electricity, airlines) and dynamic prices would be essential for efficiency.

1. INTRODUCTION

"Cloud computing," despite the new and fanciful name, involves two decades-old advances in computing technology. The first is virtualization, the ability to create a simulated environment that can run software just like a physical computer. Virtualization allows multiple users to share the underlying computational resource and originally allowed a single mainframe computer to be used by many co-located "terminals." The second advance was the development of network communication protocol (and the laying of the physical network) so that physically distant computers could interact with each other easily and quickly. Now the "terminal" can be anywhere on the network and the "datacenter" can be located where land and electricity are inexpensive.

Although these technologies are quite old, the "public cloud" is relatively new. Firms have historically opted to locate and operate their IT equipment in-house, known as "on-premises" deployments, and usually without a "hypervisor" to virtualize the resource. For example, an analysis group at a bank would use a combination of personal computers and co-located servers, with data stored in a dedicated database server.[1] The public cloud replaces all of these functions with shared resources, operated by a specialist firm and located to minimize costs. Instead of owning hardware, firms use "virtual machines," "instances" and "containerized compute." The computational resources utilized by each customer (or application) can be scaled up and down dynamically. Indeed cloud service providers invest in technologies such as load balancing, auto-scaling and redundancy management so firms can take advantage of cloud architecture easily.

While it is difficult to know the extent to which cloud-based IT will overtake the traditional model, most current projections have it taking the lead within ten years.[2] This disruption extends well beyond the ownership and operation of computational hardware. Software has traditionally been sold using a licensing model—software licenses provide the right to install the software on a physical computer (the price often depends on the number of cores) and use the software on this computer with zero marginal cost for a (priorly determined) specified period of time (often in perpetuity). Clearly this model is not a natural fit for the cloud because size and quantity of the computational resources can scale up and down dynamically. Accordingly, software pricing is moving to usage-based "pay-as-you-go" models. This fundamentally impacts competition—firms can try new software at low costs, they always use the latest version and are not tied to an expensive set of licenses that have already been purchased.

Adoption of the cloud thus offers three primary benefits to firms: 1) capital investments are converted into operating expenses supported by a competitive underlying market; 2) computational resources can be "elastically" scaled up or

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052707

[1]Note this not virtualization, they are simply operating the physical computer remotely.
[2]For a summary of many such projections, see a recent Forbes article on the topic http://bit.ly/18dQ1zA

down, allowing firms to only pay for what they use and expand/contract with lower adjustment costs; 3) it reduces lock-in of software, thus spurring competition to meet their needs. While the first two benefits have received greater attention in the academic literature [1, 4], all three benefits are amplified by elastic scaling of computational resources. These financial incentives would lead us to expect large swings in usage within customers under a fixed price regime. Given that customers from a specific geographic area are likely to have similar usage demand patterns (e.g. high when their employees are at work), we in turn expect there to be large swings of usage at the datacenter level.

Research in economics and operations management posits that dynamic pricing is critically important when capacity is fixed (at least in the short-run) and fixed costs represent a substantial fraction of total costs. In these markets, firms change prices so that demand is equal or close to the fixed capacity. The degree to which this can be accomplished depends on demand responsiveness and the firm's price setting ability, which jointly determine how supply goes unused. Our scientific understanding of peak-load pricing focuses on three industries: electricity, airlines and hotels [5, 6, 10, 12]. Cloud computing thus seems to offer a fourth large industry where dynamic prices are expected to be important.

Given the discussion thus far, it may come as a surprise that the largest cloud providers (and all providers to the best of our knowledge) overwhelmingly use static prices. Even Amazon's "spot market," which is very small relative to their standard offerings, generally trades at a fixed reserve price and prices do not vary with plausible demand patterns, indicating they are not serving the traditional aims of dynamic pricing [8]. Further "reserved instances" in which a customer pays in advance for one or three years of 24/7 operation of virtual machines (VM), have grown in popularity. Although these are early days for the public cloud, the pricing models used in practice seem to be at odds with the standard "story of the cloud" (told here and elsewhere). How can we reconcile this apparent contradiction? Are pricing models overly simplistic, leaving huge gains on the table from more sophisticated mechanisms, such as time-of-day pricing? Conversely, are demand patterns such that fixed prices are close to optimal and the main benefit is outsourcing operation of computational resources to a specialized firm, and not dynamic scaling as is commonly assumed?

We have address these questions using detailed telemetry data from a major provider's public cloud datacenters. During the period of study, prices were constant and all billing was done at the minute level (e.g. no annual reservations). We show that at the customer level, hourly and daily bill-usage volatility is small, with the majority of customers showing a min/max variation of less than 5%. This volatility tends to be smaller for large customers. We do observe a handful of outliers, both a random set of one hundred customers and the largest set of one hundred customers, that use resource much more elastically. At the datacenter level however, these outliers (firms) usage washes out—overall usage patterns are stable and predictable. Borrowing from the finance literature, we compute each tenant's "beta," the correlation of their volatility with the datacenter's demand and find that most tenants have a beta less than one, though the estimates are positive on average. In particular, we do not see large "high beta" tenants that would exacerbate demand spikes and thus be subject to intense peak load pricing.

These results on usage patterns directly inform pricing models. First, the observed patterns explain the pricing mechanisms currently used in the market, putting to rest intuition that they are too simplistic to be near-optimal. In a regime of stable usage, static prices and reserved capacity come at a low efficiency loss and have advantages of simplicity and predictability. Second, predictable usage allows providers to run datacenters at high utilization efficiency without the need to use prices to shape demand. In contrast, if firms had correlated demand spikes, for instance weekdays during business hours, similar to what is commonly observed in electricity markets [11], then the cloud provider has to "provision for the peak" just as firms must do for on-premises infrastructure. While some of this inventory could be sold off-peak at lower prices, it is nonetheless the case that the degree to which firm-level fluctuations cancel each other out directly impacts efficiency—and thus the price—in a competitive market.

We view these results as suggesting that the first generation of cloud utilization by-and-large takes the form of firms "lifting" their on-premises software stack and "placing" it in a public datacenter. If this stack was designed to run on a fixed amount of computational resources, it is explains why we see many customers having very predictable usage in the cloud. In the language of software developers, these programs are not "cloud native," in that they are not designed to dynamially provision resources to reap efficiency gains. The customers who are outliers in our data, show usage patterns that we'd expect from this type of architecture.

Since the public cloud is relatively new and evolving quickly, we do not expect the future to match the present. To provide a lens into future utilization patterns we look at the actual CPU utilization for each VM. These data give the min, max and average utilization for 5-minute intervals. Comparing the max usage in any period to 100% can be used as a proxy for optimally provisioning resources to meet the maximum computational needs within the interval (the difference between the observed max and 100% can thus be "saved"). We observe that the average max CPU utilization as compared to peak max CPU utilization ratio is less than 60% on average at datacenter level. This means that approximate potential gain from dynamic scaling is more than 40%. (Note that the customer pays based on the number of instances "up," not CPU usage directly and thus has a financial incentive to introduce such scaling.) This indicates that if we moved to a world in which resources were only turned on when needed, then we would expect far greater fluctuations in resource utilization—order half the datacenter—which is similar to demand fluctuations observed in the classic peak-load pricing industries. Relatedly, we document much more pronounced circadian cyclicality of CPU usage as compared to VM usage. Since there is a financial incentive to adopt efficient scaling, we expect dynamic prices will be necessary for efficiency at some point in the evolution of cloud computing. Taken together the results both explain current market practices and suggest that as the market matures, new economic models will have to be developed.

2. A PRIMER ON CLOUD COMPUTING

As computing and storage requirements have increased over time firms have moved IT infrastructure to dedicated facilities known as "datacenters," which serve as the backbone of the system that satisfies these requirements. A typ-

ical datacenter is designed to house thousands of "tenants" at the same time. Large IT companies such as Amazon, Apple, Facebook, Google, and Microsoft have multiple datacenters located around the world. Datacenters are organized in "clusters," each containing 10+ "racks," which contain roughly 20–80 connected servers housed in the same cabinet. Racks are connected by switches and routers. Each server is a physical computer that can accommodate multiple users via virtualization (see [3] for a detailed analysis of modern datacenters). Virtualization enables multiple users residing in the same hardware structure to have a simulated computer environment without interference from other users. Through this process, a menu of "virtual machines" is offered with differing performance levels. The hypervisor, or virtual machine monitor, is the most critical part in this technology. It can be called as the "datacenter operating system" as it equips isolation between users, supports multiple operating system, schedules and allocates resources, and provides minimal performance loss due to multi-tenancy. Moreover, it plays a critical role in managing required redundancy, autoscaling and load balancing which are the critical parts that give elasticity to the cloud (see [2] for Xen, a widely used hypervisor).

In contrast to elastic scaling in the cloud, the performance of traditional, on-premise computers is bounded by the physical hardware (software must respect resource constraints). If the hardware is not used at capacity, the unused part is wasted. "Cloud native" applications provide flexible computing power that can be adjusted based on computing needs by allocating resources dynamically. [1, 13] provide a more technical treatment of techniques to adaptively scale when computing requirements fluctuate over time. Many popular methods utilize auto-scaling and load balancing functionality provided by major cloud providers, but other methods exist. For example, a firm could run scripts to identify idle VMs and kill the deployment programmatically via command line tools. If data is always saved to a network disk (persistent storage), then this simple policy would reduce compute utilization without loss of data.

While the cloud offers these capabilities, it can also function just like traditional equipment. If a firm simply "lifted" their software applications and "placed" them in the cloud, then little dynamic scaling would occur. This highlights the difference between traditional software, where computational resources are *constraints* and marginal costs are essentially zero, and cloud-native software, where there are no hard constraints, but marginal costs are positive, since the user is paying by the minute per virtual core. Not only are the design incentives fundamentally different, but the capital expense is converted to an operating expense that is outsourced to a specialist firm.

The degree to which firms architect their software to dynamically scale directly impacts demand volatility at the datacenter, which is, in turn, a key determinant of the optimal pricing schedule. If demand is stable over time, there is minimal efficiency loss with fixed prices. If demand fluctuates widely, dynamic pricing models perform better and the extent of performance improvement is depends on the magnitude and predictability of the variation. Economic models of "peak load pricing" predict the welfare loss proportional to the square of gap between peak usage to average usage—this gaps captures lost efficiency when one has "provision for the peak." If the variation is predictable, a menu of a few fixed prices, for instance peak and off-peak price, can achieve nearly the efficiency of fully dynamic prices. Peak-load pricing is discussed extensively in the context of electricity markets. Closest to our case, [7] discusses that if the firms anticipate Cournot competition, they invest in capacity more compared to fully competitive case, which results more ample capacity during off-peak periods.

For this study, our data consist of datacenters logs from a major cloud provider. These detailed logs give information on the computational resources customers are utilizing. In general, we will use the term "utilizing" or "demand" to referred to the deployed, and thus paid for, resources, not the actual utilization of compute cycles. The billing log data is recorded per-minute at the VM level and aggregated to the hour for our purposes. We do have data to examine how intensely the deployed resources our used. These data are record summary statistics (e.g. max, min, mean) at 5-minute intervals. All records are fully anonymized before any analysis is conducted. Our study time period is four months unless otherwise noted.

3. REGION LEVEL ANALYSIS

Table 1 shows the demand volatility in region (e.g. "US East") level.[3] We normalize usage to be relative to the maximally observed value at the region (e.g. 0.04 indicates 4% of the maximally observed value for that region during the time period).[4] We treat supply as fixed for the purposes of our study. In practice this assumption fails because new datacenters can be built and existing ones can be expanded. However, this adjustment takes longer than 4 months, the period of our study, and so is not directly relevant. Further, it is often not possible for a provider to expand supply in a given region, due to various constraints such as land availability, water allowances and local regulations. Finally we note that rack failures sometimes do occur, leading to small fluctuations in available supply, but the size of these failures relative to the whole is negligible.

Currently there are not any major providers that have dynamically adjusting prices for standard workloads. While prices do change over time, it is in the form of a major announcement by the provider and those prices prevail for many months. We study a period which did not involve any price changes by the three largest American cloud providers.[5] The fact that prices are constant, supply is fixed and global capacity did not run out during our analysis period greatly simplifies our analysis because it means usage behavior is a reliable measure of demand.

We measure demand volatility at daily, weekly and monthly period lengths at the regional level. These time periods have natural ties to variation in human activity and have capture

[3]In cloud computing customers typically choose a "region," for example "US West," to deploy their workloads. A region typically contains a few co-located datacenters (physical structures) which are connected to each other at very low latency via high throughput fiberoptics. The provider can thus easily load balance between these physical units and the entire complex is commonly referred to as a "datacenter."

[4]This was a requirement for publication as it preserves confidential information about datacenter efficiency.

[5]Amazon Web Services has a "spot market" for "evictable workloads" (VMs can be shutdown without notice). Past work has shown that this market can best be understood as a secondary market where a relatively small amount of an inferior product is sold at discounted prices [8].

predictable variation in other markets, such as electricity. We use two measures to capture volatility. First is the average max-min range, which captures average peak-to-trough variation within the unit of time (e.g., daily range gives the average difference between the min and max value observed in a day). Second is the standard deviation, which captures how different a randomly selected time-interval is expected to differ from the time-interval average.

Table 1 summarizes these measures—we report the min, max, median, and mean across regions. Across all time-units, mean/median standard deviation is roughly constant 0.05–0.06. This indicates, for instance, that a randomly selected day or week tends to be about 5% different than the datacenter's average day or week. The ranges tell a similar story. Daily and weekly ranges are 2–3% on average. Since the weekly range must be greater than the daily range (if the largest differences occur within a day, they also occur within a week), the fact that weekly range exceeds daily range by roughly 30% reveals that the majority of the variation occurs with a period length of a day. Monthly range is greater, coming in at 7.2%, with a max of 11% at the datacenter level. In the supplementary material, we show that these figures are driven by growth via new customers, not a predictable monthly fluctuation, which makes sense given that past work would not lead us to expect seasonality at the monthly level. Taken together these results indicate while there is variation over time, it is small relative to the size of the datacenter.

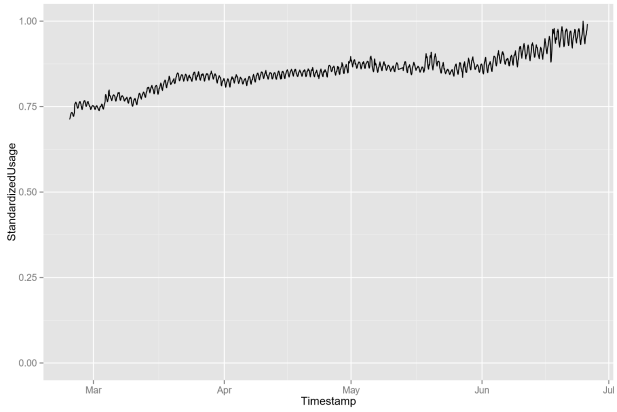

Figure 1: Standardized hourly usage in a typical region

Demand variation can be understood visually by examining Figure 1, which shows usage over time for typical region. A rather predictable time trend in the region is evident, along with small daily fluctuations of around 2%. Weekly fluctuations are not easy to make out. Indeed utilization not only has low variance, but the variance we do observe appears to be predictable. To verify this formally we split regional usage into train and test sets (i.e., test sets are not used to train the model). For each region, we first fit a predictor on daily usage with day-of-week dummies and a time trend. Then, we fit a regression line on hourly usage with hour-of-day dummies and a time trend and evaluate model predictions on the test set. The results are given in Table 2. The mean absolute percentage error has a mean/median of about 2% in both models. When compared to the variation in Table 1, we can see that the model explains roughly 70% of the observed variation. The fact that even this simple model can produce good results confirms our initial observations on predictability.

4. CUSTOMER LEVEL ANALYSIS

We now examine demand at the customer level. We will conduct our analysis on two samples of firms: 1) the top 100 customers, which captures behavior of large enterprises; 2) a random set of 100 customers,[6] which consists mostly of small businesses. We start by examining the relationship between each customer's demand and the entire demand for the region they are deployed in. To do so we run a simple linear regression, $IndivUsage_t = \alpha + \beta \times RegionUsage_t$, where the coefficient β captures the linear relationship of a customer's utilization at time t with the overall region. β has a natural analog in the Capital Asset Pricing Model, a widely used model of a security's risk and returns, in that it captures the degree and magnitude to which a customer's demand spikes tend to co-occur with market spikes. All data is de-trended and normalized so that $\beta = 1$ signifies that when the datacenter demand increases by 1% the customer's demand tends to increase by the same percentage amount.

Estimates for each individual customer are given in Figure 2. The histogram shows that the customer level usage tends to be postively related to market demand (most values are positive). Values above 1 are rare, indicates that there are very few customers who exacerbate regional level fluctuations. Further, many customers are close to zero or negative, indicating they are either negatively correlated or uncorrelated with market-level demand shocks. These findings helps explain relatively smooth utilization at the datacenter level.

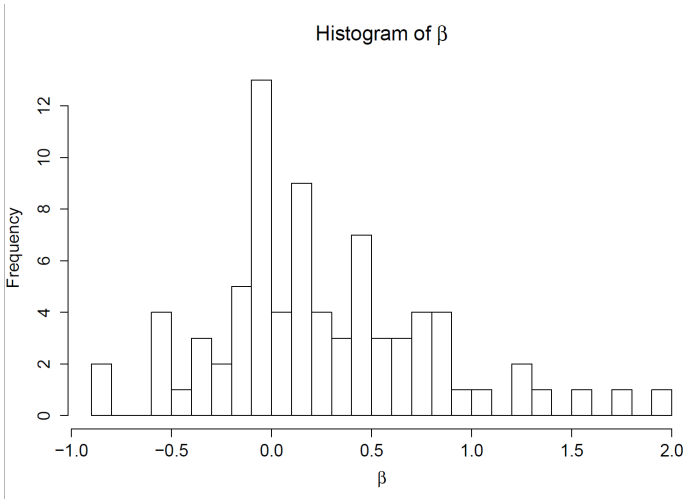

Figure 2: Histogram of β

[6]Customers are required to be active for at least 15 days during our sample period.

Table 1: Demand volatility at the region level

Region	Hourly s.d.	Average daily range	Daily s.d.	Average weekly range	Weekly s.d.	Average monthly range	Monthly s.d.
Min	0.031	0.014	0.031	0.021	0.032	0.041	0.038
Max	0.156	0.056	0.159	0.059	0.169	0.110	0.202
Mean	0.063	0.026	0.064	0.032	0.067	0.072	0.079
Median	0.052	0.023	0.053	0.031	0.054	0.065	0.061

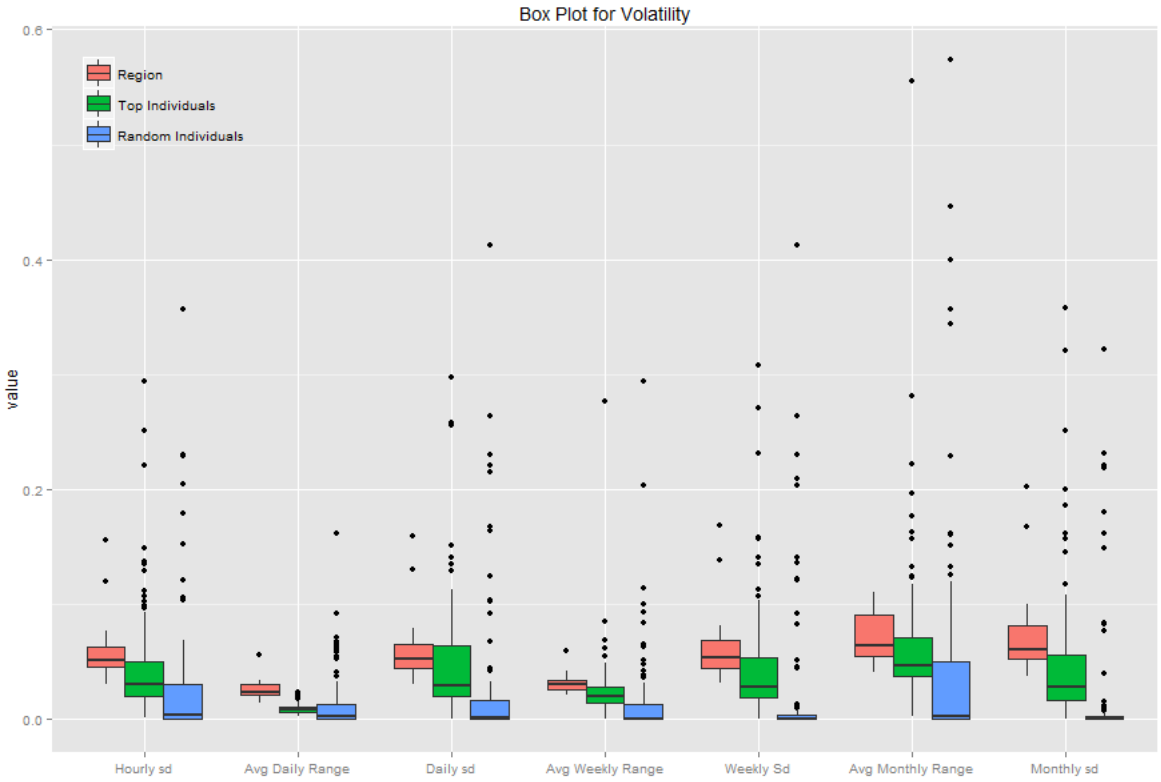

Figure 3: Summary of volatility

Table 2: Mean absolute percentage error on the test set at the region level

Region	Daily MAPE	Hourly MAPE
Min	0.007	0.010
Max	0.045	0.042
Mean	0.022	0.023
Median	0.020	0.022

Figure 3 gives customer level volatility metrics, where all measures are normalized by the maximum observed utilization at the customer level to make them comparable to the overall datacenter (which are normalized with the datacenter max). The green (middle) bars give the top 100 customers and the blue (right) bars give a randomly selected 100. The red (left) bars give the region for comparison purposes. It is immediately obvious that the typical customer level shows similar demand fluctuations to the region as a whole. Interestingly, there are a relatively high number of outlier customers in both the random 100 and top 100 that exhibit relatively large daily swings in usage. Tables 3 and 4 provide a bit more detail in confirming these conclusions. These customers are likely using dynamic scaling techniques to minimize costs and thus appear as outliers relative to the typical customer. However, since these customers are not the norm and, as we previously showed, the bursts are not strongly correlated with broader market demand, we do not observe fluctuations of this magnitude at the regional level. This means at present, datacenters can be operated with fixed prices at high capacity utilization, this efficiency is largely passed through to customers via lower prices in a competitive market.

The figure also reveals that regions exhibit higher average volatility than is observed at the customer level. While this seems a bit curious at first, it is easily explained by the fact that regions have an additional source of variation in the form of new customer acquisition. When new customers join, even if their usage is perfectly stable, this contributes

Table 3: Demand volatility at the tenant level (top 100 tenants)

Tenant Rank	Hourly s.d.	Average daily range	Daily s.d.	Average weekly range	Weekly s.d.	Average monthly range	Monthly s.d.
1	0.002	0.003	0.001	0.001	0.000	0.003	0.000
2	0.002	0.003	0.001	0.001	0.001	0.004	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	0.107	0.006	0.107	0.015	0.107	0.055	0.107
Min	0.002	0.003	0.001	0.001	0.000	0.003	0.000
Max	0.294	0.024	0.298	0.277	0.308	0.555	0.358
Mean	0.048	0.009	0.049	0.025	0.048	0.066	0.051
Median	0.031	0.009	0.030	0.020	0.028	0.047	0.028

Table 4: Demand volatility at the tenant level (randomly selected 100 tenants)

Tenant #	Hourly s.d.	Average daily range	Daily s.d.	Average weekly range	Weekly s.d.	Average monthly range	Monthly s.d.
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.005	0.004	0.001	0.001	0.001	0.005	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	0.001	0.001	0.000	0.000	0.000	0.001	0.000
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Max	0.357	0.162	0.412	0.294	0.412	0.574	0.321
Mean	0.029	0.015	0.027	0.017	0.023	0.049	0.020
Median	0.004	0.003	0.001	0.001	0.000	0.003	0.000

positively to our measures of regional volatility, but not our measures of customer volatility. We could have de-trended regional demand, but this would have made a source of variation that providers must deal with, namely an uncertain growth rate due to new customers.

To better understand customer behavior, we give three archetypes in Figure 4. A low volatility customer typically sits at a preferred level with minimal deviations. As discussed, this likely corresponds to a customer that is using software that is used treating comptuational resources as a constraint. In this case, the size of deployment in the cloud or number of servers purchased in a traditional set-up is set to ensure a desired level of performance. A moderate volatility customer floats around a relatively small range, indicating some dynamic scaling, whereas a high volatility customer shows much larger jumps utilization. While these archetypes by no means capture all usage patterns, they provide a good mental model of customer types.

5. CPU UTILIZATION

The preceding analysis focused on deployed resources, not actual utilization of the resources that are deployed. A VM instance can be "up" and not fully utilized. Cloud technologies such as auto-scaling, load-balancing and containers (an environment that allows code execution without full operating system functionality) help draw billed utilization and actual resource utilization closer together, but these technologies are not yet fully adopted, as they often require custom development solutions when moving to the cloud. Since deployments are billed by the minute, there is a financial incentive to re-architect software to make use of these benefits and thus we expect them to more widely adopted as time goes. Actual CPU utilization offers a lens into a potential future where adoption is widespread. We could imagine, for instance, a world in which these technologies allowed customers to only be billed for computational cycles (an analogy would be to pay only for what your laptop "does," not keeping the OS "ready for use").

In this section we use detailed telemetry data on CPU usage. These data give the min, max and average utilization for every 5 minutes at the VM level. One likely familiar example of this type of data data is the "system performance" feedback interface on a personal computer. We summarize these data at the region level for each 5 minute interval by taking the observed max for each VM, adding these up and comparing this to the maximum total computational usage. The max is the most attractive measure because of the resources necessary to satisfy demand within this narrow time interval. We however note that scaling could occur at time frequencies below 5 minutes, meaning our estimates may understate the impact of dynamic scaling technologies.

Figure 5 shows CPU utilization for a representative week at a typical datacenter.[7] Both day-of-week (the first two days are the weekend) and time-of-day effects are now clearly evident and far more pronounced that those observed in Figure 1. CPU utilization is steady and lower during the weekend than the weekdays. Moreover, midday has higher CPU utilization for all weekdays. The overall peak-to-trough variation is 30%, with a value closer to 20% if we ignore a few large spikes (we notably did not observe significant spikes for billed usage). Variation in max CPU utilization measure at 5-minute intervals is 10x higher than the variation in VM usage. If CPU usage is indeed lens into future behavior, then we should expect datacenter utilization to transition from the current regime of low volatility, to much more meaningful swings in demand. This would in-turn raise the

[7]Note that the data is in datacenter level, not in region level. The regional results would be similar.

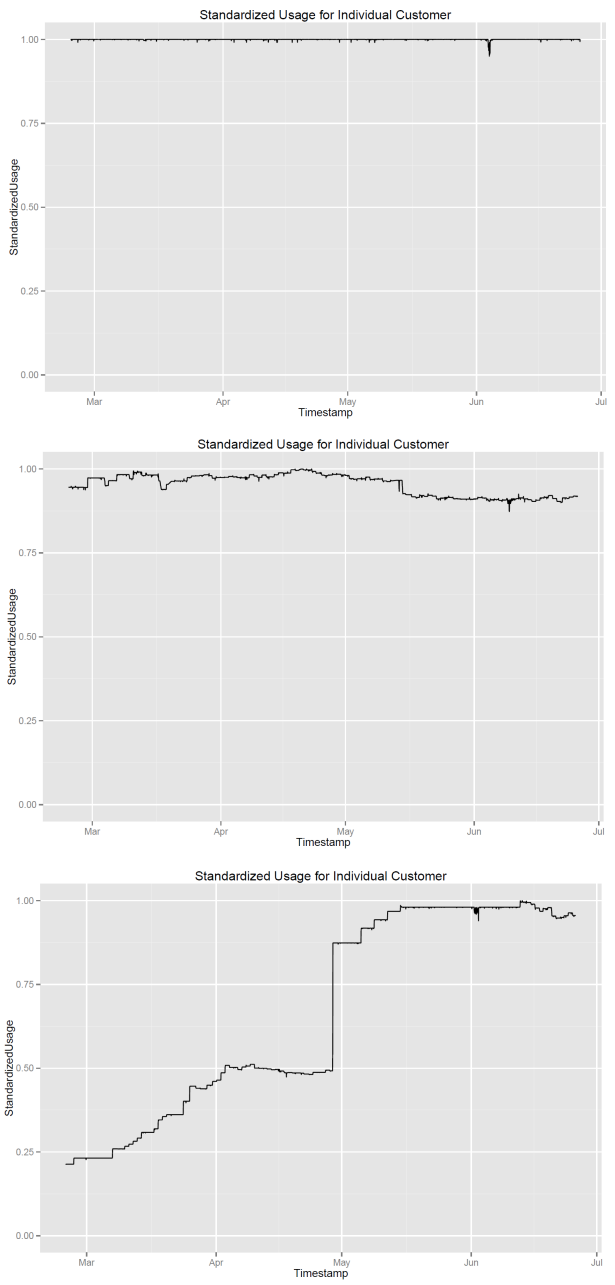

Figure 4: Standardized hourly usage of the non-volatile, median and volatile tenant

efficiency gains from using dynamic prices. Given that static prices currently prevail, it would require a reorganization of the marketplace. However, we note that while these data provide clues about the future, they are by no means dispositive.

6. DISCUSSION AND CONCLUSION

Our scientific understanding of peak-load pricing focuses on three industries: electricity, airlines and hotels [5, 6, 10, 12]. In all three of these industries, there are substantial demand fluctuations and high capacity costs, so that increasing utilization via pricing is economically important. Cloud computing is the fourth large industry that is an example of fixed capacity (in the short-run) with fixed costs representing a substantial fraction of total costs. Electricity has the property that failure to balance supply and demand either blows out transformers (over-supply) or appliances (under-supply), so that pricing cannot be used effectively as a balancing mechanism without additional controls. Moreover, peak electricity demand is generally determined by air conditioning, which is weather-dependent and therefore can be accurately forecast days in advance, which allows for a planned supply response, such as turning on peak-load generators. Airlines and hotels also have industry-specific features that limit the ability to generalize findings. Specifically, the desire of the majority of leisure travelers to book well in advance and lower price sensitivity of business travelers are salient features of the pricing problem. Especially for airlines, the discreteness of the problem is first order, with overbooking and buying back seats critical to efficiency [10]. In summary, these industries all possess quirky, industry-specific features that directly affect the utility of peak-load pricing.

Cloud computing offers great potential as a fourth major empirical example of peak-load pricing. The integer constraint is irrelevant (VMs can be threaded) and a failure to balance supply and demand is not disastrous, as with electricity. Thus, empirical work on demand for cloud computing offers potential insight into the operation of peak-load pricing. An early look at demand fluctuations [9] found peak to trough variation, for the internal demand of a single customer, on the order of 300%, suggesting using peak-load pricing would be economically critical. This variation was primarily driven by web traffic and reflects the fact that the majority of people sleep at roughly the same time in a given geographic area. Given this type of variation in demand and the standard "story of the cloud," the fact that the three largest providers don't use time of day or peak load pricing would appear to be a puzzle.

Our empirical findings offer a resolution to this puzzle and also indicate that circumstances may well change moving forward. We find that there is little variation on datacenter demand when looking at billed usage. Average variation is around 2% and the largest variation found is still under 6%. Because the efficiency loss is proportional to the square of the variation, even 6% produces a tiny efficiency loss. Moreover, there is a meaningful cost to offering a complex pricing structure—people may choose the wrong items to purchase, or shy away from purchasing due to the complexity, especially for a relatively new product—so simplicity in pricing seems entirely justified. That is, we find that, while theoretically optimal, use of an auction system or time of day pricing is unnecessary at the present time. Moreover, what variation exists is mostly (70%) predictable based on time alone; rather than using an auction, a simple, predictable time of day pricing mechanism is adequate to obtain most of the (already negligible) efficiency gains created by peak load pricing. In particular, just having two prices, peak and off-peak, with off-peak set in advance (like cellular telephone plans with their "free nights and weekends" option) would obtain the majority of the efficiency gains made available from an auction system. This is important because auctions have a substantial downside. A buyer of computing resources in an auction faces a "sudden death" loss of computing from losing the auction, and has to either write code

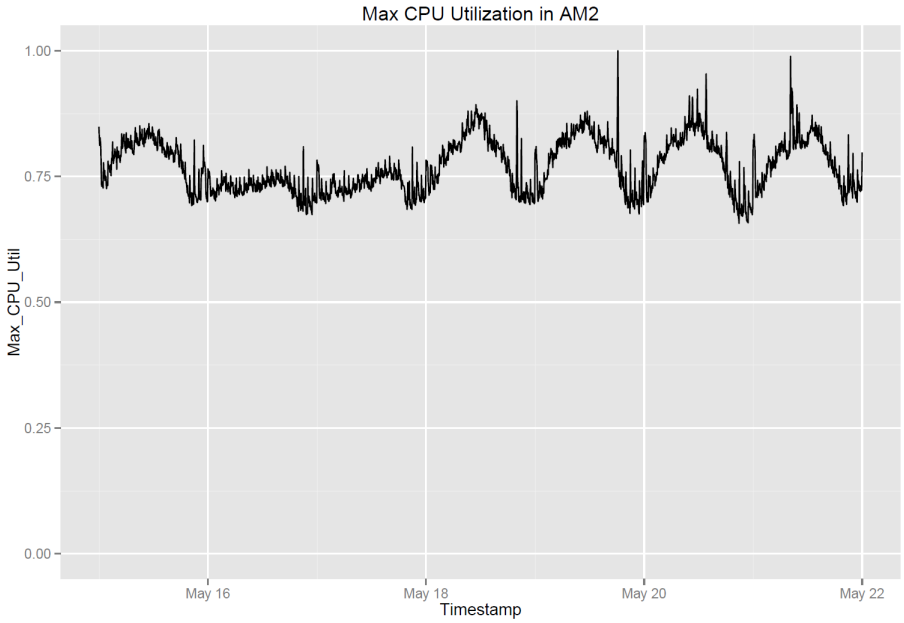

Figure 5: Max CPU utilization in a datacenter for a period of one week

to not lose the ongoing state of computation or suffer the loss of partial computation; either is costly and represents a real efficiency loss. Thus, the ability to use predictable time of day pricing in preference to an auction is an important advantage, provided the auction is not needed due to demand fluctuations.

Little variation at the datacenter level does not imply that customers have steady demand. Indeed, a fair summary is that there are three kinds of customers: steady, mild variation and large variation. Most customers are of the first two categories, but there are notable outliers in the third category that appear to be making effective use of adaptive scaling techniques. However, we showed that the correlation of these spikes with the broader market demand tends to be weak, which helps the fluctuations somewhat average out in a large sample. Indeed, a significant source of datacenter fluctuation, not present in the individual customer variation, is the entry of new customers. The steadiness of the customer demand probably has to do with low prices —a customer can save a few dollars by turning off their use, but then have to turn it back on when they need it; the convenience of "always on" overwhelms the dollar saving.

Finally, many customers may be bringing programs to the cloud that were written for an internal data center and may not take advantage of the elastic nature of cloud computing. The best indication of whether future demand will show greater fluctuation than present demand is whether usage, rather than purchasing, fluctuates more. In the same way that CPU-usage for personal computers usage fluctuates dramatically, idling at night, reveals the ability to scale back purchase, perhaps customers are just leaving their VMs on, but not actually using them. Indeed, we find that there is much greater fluctuation in usage than in purchase, 15-20% versus 2-6% for for purchases. However, this fluctuation is quite predictable, which means time of day pricing works well, even for the future. Thus, the conclusion that auctions, with their undesirable unpredictability, are unnecessary persists even with usage as a proxy for demand.

Acknowledgments

Any views and opinions expressed herein are solely those of the authors and do not reflect those of the Microsoft Corporation or Uber.

7. REFERENCES

[1] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, et al. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

[2] Paul Barham, Boris Dragovic, Keir Fraser, Steven Hand, Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, and Andrew Warfield. Xen and the art of virtualization. *ACM SIGOPS Operating Systems Review*, 37(5):164–177, 2003.

[3] Luiz André Barroso, Jimmy Clidaras, and Urs Hölzle. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture*, 8(3):1–154, 2013.

[4] Ergin Bayrak, John P Conley, and Simon Wilkie. The economics of cloud computing. *The Korean Economic Review*, 27(2):203–230, 2011.

[5] Wedad Elmaghraby and Pınar Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science*, 49(10):1287–1309, 2003.

[6] Guillermo Gallego and Garrett Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

[7] Veronika Grimm and Gregor Zoettl. Investment incentives and electricity spot market competition. *Journal of Economics & Management Strategy*, 22(4):832–851, 2013.

[8] Cinar Kilcioglu and Costis Maglaras. Revenue maximization for cloud computing services. *Mimeo*, 2015.

[9] John Langford, Lihong Li, R Preston McAfee, and Kishore Papineni. Cloud control: voluntary admission control for intranet traffic management. *Information Systems and e-Business Management*, 10(3):295–308, 2012.

[10] R Preston McAfee and Vera Velde. Dynamic pricing with constant demand elasticity. *Production and Operations Management*, 17(4):432–438, 2008.

[11] Kathleen Spees and Lester B Lave. Demand response and electricity market efficiency. *The Electricity Journal*, 20(3):69–85, 2007.

[12] Lawrence R Weatherford and Samuel E Bodily. A taxonomy and research overview of perishable-asset revenue management: yield management, overbooking, and pricing. *Operations Research*, 40(5):831–844, 1992.

[13] Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: state-of-the-art and research challenges. *Journal of internet services and applications*, 1(1):7–18, 2010.