Article development led by **acmqueue**
queue.acm.org

The competition among cloud providers may drive prices downward, but at what cost?

BY DAVE DURKEE

Why Cloud Computing Will Never Be Free

THE LAST TIME the IT industry delivered outsourced shared-resource computing to the enterprise was with timesharing in the 1980s when it evolved to a high art, delivering the reliability, performance, and service the enterprise demanded. Today, cloud computing is poised to address the needs of the same market, based on a revolution of new technologies, significant unused computing capacity in corporate data centers, and the development of a highly capable Internet data communications infrastructure. The economies of scale of delivering computing from a centralized, shared infrastructure have set the expectation among customers that cloud computing costs will be significantly lower than those incurred from providing their own computing. Together with the reduced deployment costs of open source software and the

perfect competition characteristics of remote computing, these expectations set the stage for fierce pressure on cloud providers to continuously lower prices.

This pricing pressure results in a commoditization of cloud services that deemphasizes enterprise requirements such as guaranteed levels of performance, uptime, and vendor responsiveness, much as has been the case with the Web hosting industry. Notwithstanding, it is the expectation of enterprise management that operating expenses be reduced through the use of cloud computing to replace new and existing IT infrastructure. This difference between expectation and what the industry can deliver at today's near-zero price points represents a challenge, both technical and organizational, which will have to be overcome to ensure large-scale adoption of cloud computing by the enterprise.

The Essential Characteristics of Cloud Computing

This is where we come full circle and timesharing is reborn. The same forces are at work that made timesharing a viable option 30 years ago: the high cost of computing (far exceeding the cost of the physical systems), and the highly specialized labor needed to keep it running well. The essential characteristics of cloud computing that address these needs are:⁴

- *On-demand access.* Rapid fulfillment of demand for computing and continuing ability to fulfill that demand as required.

- *Elasticity.* Computing is provided in the amount required and disposed of when no longer needed.

- *Pay-per-use.* Much like a utility, cloud resource charges are based on the quantity used.

- *Connectivity.* All of the servers are connected to a high-speed network that allows data to flow to the Internet as well as between computing and storage elements.

- *Resource pooling.* The cloud provider's infrastructure is shared across

some number of end customers, providing economies of scale at the computing and services layers

- *Abstracted infrastructure.* The cloud end customer does not know the exact location or the type of computer(s) their applications are running on. Instead, the cloud provider provides performance metrics to guarantee a minimum performance level.

- *Little or no commitment.* This is an important aspect of today's cloud computing offerings, but as we will see here, interferes with delivery of the services the enterprise demands.

Service Models

Cloud is divided into three basic service models. Each model addresses a specific business need.

- *Infrastructure as a Service (IaaS).* This is the most basic of the cloud

service models. The end customer is purchasing raw compute, storage, and network transfer. Offerings of this type are delivered as an operating system on a server with some amount of storage and network transfer. These offerings can be delivered as a single server or as part of a collection of servers integrated into a virtual private data center (VPDC).

- *Platform as a Service (PaaS).* This is the next layer up where the end customer is purchasing an application environment on top of the bare bones infrastructure. Examples of this would be application stacks: Ruby on Rails, Java, or LAMP. The advantage of PaaS is that the developer can buy a fully functional development and/or production environment.

- *Software as a Service (SaaS).* This currently is the highest layer in the cloud

stack. The end customer is purchasing the use of a working application. Some examples of this are NetSuite and Salesforce.com. (This service is not the focus of this article.)

Perfect Competition Determines Cloud Pricing Strategies

In our experience providing cloud services, many of the current cloud end customers use price as their primary decision criteria. As a result, service providers' offerings tend toward a least common denominator, determined by the realities of providing cloud service at the lowest possible price. At the same time, the cloud computing market is becoming more crowded with large providers entering the playing field, each one of which trying to differentiate itself from the already established players. The result of many pro-

viders competing to deliver very similar product in a highly price-competitive environment is termed *perfect competition* by economists. Perfectly competitive markets, such as those for milk, gasoline, airline seats, and cellphone service, are characterized by a number of supplier behaviors aimed at avoiding the downsides of perfect competition, including:

- ▶ Artificially differentiating the product through advertising rather than unique product characteristics
- ▶ Obscuring pricing through the use of additional or hidden fees and complex pricing methodologies
- ▶ Controlling information about the product through obfuscation of its specifications
- ▶ Compromising product quality in an effort to increase profits by cutting corners in the value delivery system
- ▶ Locking customers into long-term commitments, without delivering obvious benefits.

These factors, when applied to the cloud computing market, result in a product that does not meet the enterprise requirements for deterministic behavior and predictable pricing. The resulting price war potentially threatens the long-term viability of the cloud vendors. Let's take a closer look at how perfect competition affects the cloud computing market.

Variable performance. We frequently see advertisements for cloud computing breaking through the previous price floor for a virtual server instance. It makes one wonder how cloud providers can do this and stay in business. The answer is that they over commit their computing resources and cut corners on infrastructure. The result is variable and unpredictable performance of the virtual infrastructure.⁵

Many cloud providers are vague on the specifics of the underlying hardware and software stack they use to deliver a virtual server to the end customer, which allows for overcommitment. Techniques for overcommitting hardware include (but are not limited to):

- ▶ Specify memory allocation and leave CPU allocation unspecified, allowing total hardware memory to dictate the number of customers the hardware can support;
- ▶ Quote shared resource maximums instead of private allocations;

The cloud computing market is becoming more crowded with large providers entering the playing field, each one of which trying to differentiate itself from the already established players.

▶ Offer a range of performance for a particular instance, such as a range of GHz; and

▶ Overallocate resources on a physical server, or “thin provisioning.” Commercial virtualization management software such as VMWare or Virtuozzo offer the ability to overallocate resources on the underlying hardware, resulting in reduced performance during peak loads.

Like overcommitment, limiting access to infrastructure resources or choosing lower-priced, lower-performance (and potentially older) infrastructure is used by vendors to make providing cloud computing at rock-bottom prices viable. We entered the cloud provider business after discovering we could not guarantee enterprise-grade performance to our customers by reselling other vendors' cloud services due to their corner-cutting. Here is a list of some of the strategies the author has seen over the years:

▶ *Traffic shaping.* A new client asked us to move their existing VPDC to our infrastructure. Shortly after initiating the transfer, the data rate dropped from approximately 10Mbit/sec to 1Mbit, where it remained for the duration of the transfer. This behavior speaks pretty strongly of traffic shaping. Because the client based their downtime window for the data center move on the assumption that the connecting network was gigabit Ethernet, they missed their estimate by over a week.

▶ *Using older gigabit or fast Ethernet networking.* An ISP that was selling an Amazon-like cloud computing product connected their servers to the Internet using fast Ethernet switches. If a customer wanted faster connectivity, there was an up charge per port.

▶ *Recycling failed disk drives.* A client leased several servers in a private cloud from a very large ISP. He had previously experienced several drive failures with this ISP, so he decided to check up on the hardware by running `smartctl` to assess the health of the drives. What he found was shocking to him: the ‘new’ servers he had just received had disk drives in them that were over three years old! When he challenged the ISP, he was told their policy was to replace a drive only when it fails.

▶ *Deploying older CPU technology.* We were asked to manage a client's ap-

plications that were hosted at another ISP. As part of our initial assessment of the client's environment, we discovered that he had received Athlon desktop processors in the servers that he was paying top dollar for.

This difference between advertised and provided value is possible because cloud computing delivers abstracted hardware that relieves the client of the responsibility for managing the hardware, offering an opportunity for situations such as those listed here to occur. As our experience in the marketplace shows, the customer base is inexperienced with purchasing this commodity and overwhelmed with the complexity of selecting and determining the cost of the service, as well being hamstrung by the lack of accurate benchmarking and reporting tools. Customer emphasis on pricing levels over results drives selection of poorly performing cloud products. However, the enterprise will not be satisfied with this state of affairs.

Extra charges. For example, ingress and egress bandwidth is often charged separately and using different rates, overages on included baseline storage, or bandwidth quantities are charged at much higher prices than the advertised base rates, charges are applied to the number of IOPS used on the storage system, and charges are levied on HTTP get/put/post/list operations, to name but a few. These additional charges cannot be predicted by the end user when evaluating the service, and are another way the cloud providers are able to make the necessary money to keep their businesses growing because the prices they are charging for compute aren't able to support the costs of providing the service. The price of the raw compute has become a loss-leader.

Long-term commitments. Commitment hasn't been a prominent feature of cloud customer-vendor relationships so far, even to the point that pundits will tell you that "no commitment" is an essential part of the definition of cloud computing. However, the economics of providing cloud computing at low margins is changing the landscape. For example, Amazon AWS introduced reserved instances that require a one- or three-year long commitment.

There are other industries that offer their services with a nearly identical

delivery model, most obviously cellular telephone providers and to some extent electrical utilities. However, for some reason, cloud computing is not delivered with the same pricing models as those developed over the last hundred years to deliver electricity. These providers all use long-term commitments to ensure their economic viability by matching their pricing to customer resource usage that determines their costs. Long-term commitments—in other words, contracts—allow for time-of-use pricing and quantity discounts. We feel these characteristics will become ubiquitous features of cloud computing in the near future. For cloud computing delivered as SaaS, long-term commitments are already prevalent.

Navigating Today's Perfectly Competitive Cloud Computing Market

Today's price-focused cloud computing market, which is moving rapidly toward perfect competition, presents

challenges to the end customer in purchasing services that will meet their needs. This first-generation cloud offering, essentially Cloud 1.0, requires the end customer to understand the trade-offs that the service provider has made in order to offer computing to them at such a low price.

Service-Level Agreements. Cloud computing service providers typically define an SLA as some guarantee of how much of the time the server, platform, or application will be available. In the cloud market space, meaningful SLAs are few and far between, and even when a vendor does have one, most of the time it is toothless. For example, a well-known cloud provider guarantees an availability level of 99.999% uptime, or five minutes a year, with a 10% discount on their charges for any month in which it is not achieved. However, since their infrastructure is not designed to reach five-nines of uptime, they are effectively offering a 10% discount on their services in exchange for the benefit of claiming that level of reli-

ability. If a customer really needs five-nines of uptime, a 10% discount is not going to even come close to the cost of lost revenue, breach of end-user service levels, or loss of market share due to credibility issues.

Another trick the service providers play on their customers is to compute the SLA on an annualized basis. This means that customers are only eligible for a service credit after one year has passed. Clearly the end user should pay close attention to the details of the SLA being provided and weigh that against what business impact it will have if the service provider misses the committed SLA. From what we have seen in the last four years of providing IaaS and PaaS, most customers do not have a strong understanding of how much downtime their businesses can tolerate or what the costs are for such downtime. This creates a carnival atmosphere in the cloud community where ever higher SLAs are offered at lower prices without the due diligence needed to achieve them...another race to the bottom.

Taking advantage of the low prices of Cloud 1.0 requires an honest assessment by the end customer of the level of reliability they actually need.

Performance is almost never discussed. One of the hazards of shared infrastructure is that one customer's usage patterns may affect other customers' performance. While this interference between customers can be engineered out of the system, addressing this problem is an expense that vendors must balance against the selling price. As a result, repeatable benchmarks of cloud performance are few and far between because they are not easily achieved, and Cloud 1.0 infrastructure is rarely capable of performance levels that the enterprise is accustomed to.

While it makes intuitive sense to quiz the cloud provider on the design of their infrastructure, the universe of possibilities for constraining performance to achieve a \$0.03/hour instance price defies easy analysis, even for the hardware-savvy consumer. At

best, it makes sense to ask about performance SLAs, though at this time we have not seen any in the industry. In most cases, the only way to determine if the service meets a specific application need is to deploy and run it in production, which is prohibitively expensive for most organizations.

In my experience, most customers use CPU-hour pricing as their primary driver during the decision-making process. While the resulting performance is adequate for many applications, we have also seen many enterprise-grade applications that failed to operate acceptably on Cloud 1.0.

Service and support. One of the great attractions of cloud computing is that it democratizes access to production computing by making it available to a much larger segment of the business community. In addition, the elimination of the responsibility for physical hardware removes the need for data center administration staff. As a result, there is an ever-increasing number of people responsible for production computing who do not have system administration backgrounds, which creates demand for comprehensive cloud vendor support offerings. Round-the-clock live support staff costs a great deal and commodity cloud pricing models cannot support that cost. Many commodity cloud offerings have only email or Web-based support, or only support the usage of their service, rather than the end-customer's needs.

When you can't reach your server just before that important demo for the new client, what do you do? Because of the mismatch between the support levels needed by cloud customers and those delivered by Cloud 1.0 vendors, we have seen many customers who replaced internal IT with cloud, firing their system administrators, only to hire cloud administrators shortly thereafter. Commercial enterprises running production applications need the rapid response of phone support delivered under guaranteed SLAs.

Before making the jump to Cloud 1.0, it is appropriate to consider the costs involved in supporting its deployment in your business.

The Advent of Cloud 2.0: The Value-Based Cloud

The current myopic focus on price has

created a cloud computing product that has left a lot on the table for the customer seeking enterprise-grade results. While many business problems can be adequately addressed by Cloud 1.0, there are a large number of business applications running in purpose-built data centers today for which a price-focused infrastructure and delivery model will not suffice. For that reason, we see the necessity for a new cloud service offering focused on providing value to the SME and large enterprise markets. This second-generation value-based cloud is focused on delivering a high performance, highly available, and secure computing infrastructure for business-critical production applications, much like the mission of today's corporate IT departments.

This new model will be specifically designed to meet or exceed enterprise expectations, based on the knowledge that the true cost to the enterprise is not measured by the cost per CPU cycle alone. The reasons most often given by industry surveys of CIOs for holding back on adopting the current public cloud offerings are that they do not address complex production application requirements such as compliance, regulatory, and/or compatibility issues. To address these issues, the value-based cloud will be focused on providing solutions rather than just compute cycles.

Cloud 2.0 will not offer CPU at \$0.04/hour. Mission-critical enterprise applications carry with them a high cost of downtime.² Indeed, many SaaS vendors offer expensive guarantees to their customers for downtime. As a result, enterprises typically require four-nines (52 minutes unavailable a year) or more of uptime. Highly available computing is expensive, and historically, each additional nine of availability doubles the cost to deliver that service. This is because infrastructure built to provide five-nines of availability has no single points of failure *and* is always deployed in more than one physical location. Current cloud deployment technologies use n+1 redundancy to improve on these economies up to the three-nines mark, but they still rule past this point. Because the cost of reliability goes up geometrically as the 100% mark is neared, many consider five-nines and above to be nearly unachievable (and unaffordable), only

The current myopic focus on price has created a cloud computing product that has left a lot on the table for the customer seeking enterprise-grade results.

deserving of the most mission-critical applications. In addition, there are significant infrastructure challenges to meet the performance requirements of the enterprise, which significantly raise resource prices.

Technology challenges faced by Cloud 2.0 providers. The number-one problem that Cloud 2.0 providers face is supplying their enterprise customers with storage that can match the performance and reliability they are accustomed to from their purpose-built data centers at a price point that is significantly lower. When traditional storage technologies are used in a cloud infrastructure, they fail to deliver adequate performance because the workload is considerably less predictable than what they were designed for. In particular, the randomness of disk accesses as well as the working set size are both proportional to the number of different applications that the storage system is serving at once. Traditionally SANs have solved the problem of disk read caching by using RAM caches. However, in a cloud application, the designed maximum RAM cache sizes are completely inadequate to meet the requirement of caching the total working sets of all customer applications. This problem is compounded on the write side, where the caches have traditionally been battery-backed RAM, which is causing storage vendors to move to SSD technology to support cloud applications.

Once the storage caching problem has been solved, the next issue is getting cloud applications' large volumes of data out of the SAN into the server. Legacy interconnect, such as fiber-channel with which most SANs are currently shipped, cannot meet the needs of data-hungry Cloud 2.0 infrastructures. Both Ethernet and Infiniband offer improved performance, with currently shipping Infiniband technology holding the title of fastest available interconnect. Storage vendors who eschew Infiniband are relegating their products to second-tier status in the Cloud 2.0 world. Additionally, fast interconnect is a virtual requirement between servers, since enterprise applications are typically deployed as virtual networks of collaborating instances that cannot be guaranteed to be on the same physical servers.¹

With an increasing number of clouds being deployed in private data centers or small-to-medium MSPs, the approach to build a cloud used by Amazon, in which hardware and software were all developed in-house, is no longer practical. Instead, clouds are being built out of commercial technology stacks with the aim of enabling the cloud vendor to go to market rapidly while providing high-quality service. However, finding component technologies that are cost competitive while offering reliability, 24x7 support, adequate quality (especially in software), and easy integration is extremely difficult, given that most legacy technologies were not built or priced for cloud deployment. As a result, we expect some spectacular Cloud 2.0 technology failures, as was the case with Cloud 1.0. Another issue with this approach is that the technology stack must provide native reliability in a cloud configuration that actually provides the reliability advertised by the cloud vendor.

Why transparency is important. Transparency is one of the first steps to developing trust in a relationship. As we discussed earlier, the price-focused cloud has obscured the details of its operation behind its pricing model. With Cloud 2.0, this cannot be the case. The end customer must have a quantitative model of the cloud's behavior. The cloud provider must provide details, under NDA if necessary, of the inner workings of their cloud architecture as part of developing a closer relationship with the customer. Insight into the cloud provider's roadmap and objectives also brings the customer into the process of evolving the cloud infrastructure of the provider. Transparency allows the customer to gain a level of trust as to the expected performance of the infrastructure and the vendor. Taking this step may also be necessary for the vendor to meet enterprise compliance, and/or regulatory requirements.

This transparency can only be achieved if the billing models for Cloud 2.0 clearly communicate the value (and hence avoided costs) of using the service. To achieve such clarity, the cloud vendor has to be able to measure the true cost of computing operations that the customer executes and bill for them. Yet today's hardware, as well as management, monitoring, and billing

Transparency allows the customer to gain a level of trust as to the expected performance of the infrastructure and the vendor. Taking this step may also be necessary for the vendor to meet enterprise compliance, and/or regulatory requirements.

software are not designed to provide this information. For example, billing for IOPs in a multitenant environment is a very deep technological problem, impacting not only the design of the cloud service, but the technologies it rests on such as operating systems, device drivers, and network infrastructure. Another example is computing and minimizing the costs of fragmentation of computing resources across one or more clusters of compute nodes while taking into consideration the time dependence of individual customers' loads and resource requirements.

The role of services. When cloud infrastructure reduces the barriers to deployment, what still stands in the way? That would be services, such as ongoing administration, incident response, SLA assurance, software updates, security hardening, and performance tuning. Since 80% of downtime is caused by factors other than hardware,³ services are essential to reliable production computing. Traditionally these services have been delivered by the enterprise's IT department, and simply replacing their servers with remote servers in the cloud doesn't solve the services problem. Because services delivered with cloud computing will necessarily be outsourced, they must be delivered within the context of a long-term commitment that allows the vendor to become familiar with the customer's needs, which will retire today's Cloud 1.0 customer expectation of little or no commitment. At the same time, the move toward long-term commitments will drive vendors to focus on customer satisfaction rather than the more prevalent churn visible in perfectly competitive markets.

Service-level management. Service-level agreements are the name of the game in Cloud 2.0. Enterprise customers typically have obligations to provide services to their customers within a contracted SLA. The service delivery infrastructure's SLA must meet or exceed the service levels that the enterprise has committed to provide. All aspects of the service delivery infrastructure (compute fabric, storage fabric, and network fabric) should be monitored by a monitoring system. In addition, all of the customer's cloud instances should be monitored as well. VMs must be monitored at the

system level as well as the application level. The monitoring system's rich data collection mechanisms are then fed as inputs to the service providers' processes so that they can manage service-level compliance. A rich reporting capability to define and present the SLA compliance data is essential for enterprise customers. Typically, SLAs comprise of some number of service-level objectives (SLOs). These SLOs are then rolled up to compute the overall SLA. It pays to remember the overall SLA depends on the entire value delivery system, from the vendor's hardware and software to the SLOs for the vendor's support and operations services offerings. To provide real value to the enterprise customer, the cloud provider must negotiate with the customer to deliver their services at the appropriate level of abstraction to meet the customer's needs, and then manage those services to an overall application SLA.

The role of automation. In order to obtain high quality and minimize costs the value-based cloud must rely on a high degree of automation. During the early days of SaaS clouds, when the author was building the NetSuite data center, they had over 750 physical servers that were divided into three major functions: Web delivery, business logic, and database. Machine-image templates were used to create each of the servers in each tier. However, as time went on, the systems would diverge from the template image because of ad hoc updates and fixes. Then, during a deployment window, updates would be applied to the production site, often causing it to break, which resulted in a violation of end-customer SLAs. As a consequence, extensive effort was applied to finding random causes for the failed updates. The root cause was that the QA tests were run on servers that were exact copies of the templates; however, some of the production systems were unique, which caused faults during the deployment window. These types of issues can break even the tightest of deployment processes. The moral of the story is to never log in to the boxes. This can only be accomplished by automating all routine system administration activities.

There are several data-center-run book-automation tools on the market

today for use in corporate data centers. These tools allow for the complete automation of every aspect of the server life cycle from creation of a virtual infrastructure through scaling, service-level management, and disposal of the systems when the customer has finished with them. While automation has made significant progress in the corporate data center, it is only in its infancy in the cloud. Yet, to replace the corporate data center, Cloud 2.0 must include automation. This capability allows both the cloud provider and the customer to obtain some unprecedented benefits:

- ▶ *Very high service levels.* The system is managing itself, and humans get involved only as needed, both at the service provider and customer-level processes.

- ▶ *Problems and solutions become methodological rather than random.* This allows you to fix all instances of a problem with a code change.

- ▶ *Automatically scalable infrastructure.* Allows customers to pay for only what they need when they need it without additional system administration effort to maintain service levels.

- ▶ *Automatic disaster recovery.* Automation handles the manual tasks of failover to a backup data center as well as failing back to the primary data center.

- ▶ *Minimize staffing.* The automation framework uses feedback from the monitoring system to automatically address common solutions to common problems, as well as automatically execute repetitive processes. Escalation to staff occurs only when the automation framework can't address a fault.

- ▶ *Power savings.* The automation framework concentrates the workloads onto the minimum number of servers necessary to maintain service levels, and turns off the rest.

Betting your Business on Cloud 2.0

By offering value beyond simply providing CPU cycles, the cloud provider is becoming a part of the end customers' business. This requires a level of trust that is commensurate with hiring an employee or outsourcing your operations. Do you know who you are hiring? This vendor-partner must understand what the enterprise holds important, and must be able to oper-

ate in a way that will support the cloud end customer's business. By taking on the role of operations services provider to the enterprise, the vendor enables the end customer to gain all of the benefits of cloud computing without the specialized skills needed to run a production data center. However, it is unrealistic to expect outsourced IT that eliminates the need for in-house staffing to be delivered at today's cloud computing prices.

For the Cloud 2.0 revolution to take hold, two transformations must occur, which we are already seeing in our sales and marketing activities: cloud vendors must prepare themselves to provide value to the enterprise that entices them out of their purpose-built data centers and proprietary IT departments; and customers must perceive and demand from cloud vendors the combination of fast and reliable cloud computing with operations services that their end users require.

Related articles on queue.acm.org

CTO Roundtable: Cloud Computing

Mache Creeger

<http://queue.acm.org/detail.cfm?id=1536633>

Building Scalable Web Services

Tom Killalea

<http://queue.acm.org/detail.cfm?id=1466447>

Describing the Elephant: The Different Faces of IT as Service

Ian Foster and Steven Tuecke

<http://queue.acm.org/detail.cfm?id=1080874>

References

1. *EE Times Asia*. Sun grooms Infiniband for Ethernet face-off; http://www.eetasia.com/ART_8800504679_590626_NT_e979f375.HTM
2. Hiles, A. Five nines: chasing the dream?; <http://www.continuitycentral.com/feature0267.htm>
3. Jayaswal, K. *Administering Data Centers: Servers, Storage, and Voice Over IP*. John Wiley & Sons, Chicago, IL, 2005; http://searchdatamanagement.techtarget.com/generic/0,295582,sid91_gcl1150917,00.html
4. National Institute of Science and Technology. NIST Definition of Cloud Computing; <http://csrc.nist.gov/groups/SNS/cloud-computing/>
5. Winterford, B. Stress tests rain on Amazon's cloud. *IT News*; <http://www.itnews.com.au/News/153451,stress-tests-rain-on-amazons-cloud.aspx>

Dave Durkee (dave@enikiconsulting.net) is founder and technical director of ENKI, a managed cloud computing services provider in Mountain View, CA. Durkee has over 25 years of experience in IT infrastructure, networking, business applications development, and executive corporate management. He has held several senior management IT positions, including CIO of a successful hosted ERP application service provider, Netsuite.com

© 2010 ACM 0001-0782/10/0500 \$10.00