451 Research® | Advisory

CLOUD PRICING

SPIRALING CLOUD COSTS AT THE CROSSROADS OF MATURITY

A Report Based on Research Commissioned by Hewlett Packard Enterprise

Hewlett Packard Enterprise

NOVEMBER 2015

ABOUT 451 RESEARCH

451 Research is a preeminent information technology research and advisory company. With a core focus on technology innovation and market disruption, we provide essential insight for leaders of the digital economy. More than 100 analysts and consultants deliver that insight via syndicated research, advisory services and live events to over 1,000 client organizations in North America, Europe and around the world. Founded in 2000 and headquartered in New York, 451 Research is a division of The 451 Group.

New York
20 West 37th Street, 6th Floor
New York, NY 10018
Phone: 212.505.3030
Fax: 212.505.2630

London
Paxton House (5th floor), 30 Artillery Lane
London, E1 7LS, UK
Phone: +44 (0) 207 426 0219
Fax: +44 (0) 207 426 4698

San Francisco
140 Geary Street, 9th Floor
San Francisco, CA 94108
Phone: 415.989.1555
Fax: 415.989.1558

Boston
1 Liberty Square, 5th Floor
Boston, MA 02109
Phone: 617.275.8818
Fax: 617.261.0688

TABLE OF CONTENTS

INTRODUCTION

It's been almost a decade since the IT industry first saw inklings of an intense new trend in delivering and consuming IT infrastructure, and the revolution is nearly complete. As a result, expectations and habits have changed radically for almost every consumer of IT, and the techniques of IT service delivery we call 'cloud computing' – on-demand, self-service, programmatic and nearly instant access to computing, storage and networking resources – have reached a crossroads of maturity.

For IT organizations of all stripes, but especially for software-driven, high-tech service providers, a new calculus has emerged about how to consume cloud computing. It's now the default for SaaS providers, social and consumer creative media businesses, and online businesses of all kinds to begin their businesses 'in the cloud,' using one of the dozens of public cloud environments currently available.

There are sound reasons for this because cloud computing:

- Requires minimal up-front capital investment
- Allows pay-per-use resources on a consumption model
- Can utilize an opex cost model vs. traditional capex
- Can scale with the business without effort
- Is typically highly connected and highly available to end users
- Comes with access to many adjunct services required for online activity, such as Domain Name System, content delivery network, load balancing and assorted other services already integrated or nearby.

However, every IT-centric business now has a careful decision to make when that public cloud consumption starts to grow: is it time to build and operate a private cloud?

The cloud model has matured, and to a great extent, the technology for operating one's own cloud is broadly accessible – hardware, software, orchestration and interfaces are widely available now at every level from free open source, do-it-yourself (DIY) implementations to fully integrated, pre-configured racks that literally need only to be plugged in and powered on.

Public cloud providers offer agility and speed, but they charge a premium for that privilege, especially if costs are projected outward and take into account amortization. And with the ease of access for both private cloud platforms and public cloud, it can be relatively painless, if not effortless, to make the shift and radically improve the operational cost structure with private clouds. The cost to provision a virtual machine on a public cloud is pennies per hour; the cost to provision a VM privately is pennies per month. The trade-off is investment, time and business outcome.

The cost to provision a virtual machine on a public cloud is pennies per hour; the cost to provision a VM privately is pennies per month.

This report is an exploration of the potential to seize that step-change increase in efficiency for businesses that before now may have been born in the cloud. The aim: to transition to the same model of service delivery that Google, eBay, Facebook and many others do, and to stop overpaying for basic infrastructure.

PAYING FOR IT

On-demand procurement of IT resources is truly having a revolutionary effect, not just on businesses but also on individuals. The ability to purchase a potentially unlimited amount of storage, compute and other public cloud services with no capital up front has given everyone from teenagers to CEOs, startups to service providers the ability to consume powerful technology simply, cheaply and without financial risk. Ten years ago, individuals wanting to store their files in a globally accessible location would have had to use a dedicated infrastructure with associated capital and operating expenses; now, it is as simple as pressing a button on a smartphone.

Although public cloud has brought transparency and accessibility to the IT industry, private cloud is still procured in the old style of IT: platforms are typically custom designed and built based on an RFP or specification; quotes are on a negotiated, per-customer basis, and few details are made public; payment is predictable and invoiced; and capital is still typically required, or the project must be financed. What is interesting is that vendors aren't the ones making this process difficult and complex – anecdotally, we hear time and again from both service providers and enterprises that although on-demand IT procurement is attractive, predictability, variability and scalability characteristics open up new issues in terms of budgets, management and legalities.

Although on-demand IT procurement is attractive, predictability, variability and scalability characteristics open up new issues in terms of budgets, management and legalities.

Yes, public cloud lets companies scale to meet demand, but who pays, and where does the money come from? Who controls this ability to grow and shrink, and who controls the cost? Fixed budgets, RFP processes and invoices may seem old hat, but they ensure expenditures don't get out of control and corporate risks and liabilities are managed. The desire for flexibility coupled with the need for control has triggered innovative pricing and finance models that seek to give customers the best of both words.

In 451 Research's Voice of the Enterprise Q4 2014 results, private cloud is cited as the main deployment method for every use case surveyed, including big data, web applications and enterprise applications. 451 Research's Market Monitor estimates the size of the cloud-enabling technology market (vendors that sell software to build private clouds) to be 30% greater than the size of the cloud computing market through 2018. So the message is that private cloud is desired by the market and is deemed to be superior to public cloud by many users in many scenarios, especially those with high utilization rates and those that seek full control of security and governance.

Although buyers are happy to have conversations with providers and vendors, and certainly want a 'trusted adviser' to help them maximize their use of the private cloud, they do have a challenge – when should they use private, and when should they use public? Most want the best of both worlds; they want to access the best execution venues that meet their performance, cost, contractual and other requirements. Comparing and contrasting these options is difficult, indeed, not just in terms of technical capability, but also in terms of bare-bones financials. Each option is priced, publicized and invoiced in an entirely different way. ***This problem isn't just faced by enterprises; systems integrators, distributors and service providers have to decide when should they partner with a public cloud provider, and when should they invest in their own private clouds.***

451 Research's Cloud Price Index (CPI) Private Edition is a range of benchmark indicators that show the average price of private cloud across a range of options, including OpenStack® distributions, managed services and in-house deployments. By considering the total costs of our 'basket' of hosting services, infrastructure, software and operating systems from a range of providers delivered using a range of mechanisms, we can find the average market price per virtual machine (VM) per hour of our basket. By measuring the change in the cost of our basket over time, we can evaluate how the industry is changing, and what this means for service providers and buyers. ***Crucially, it lets us analyze the financial differences between public and private cloud.***

METHODOLOGY

Private vs. Public Pricing

To derive our CPI indicators, 451 Research took on the role of an organization that wants to implement a private cloud. The organization, potentially a service provider looking to sell cloud offerings, requires its expenditure to be assessed over 36 months, and has drawn up a specification of a private cloud using two sets of requirements:

- Virtualization requirements – 500 VMs, each consisting of (roughly) an Intel E2670v2 2.5GHz core, 4GB memory, 200GB SAN storage without thin-provisioning and 0.5GB/sec peak bandwidth. Any virtualization and orchestration platform can be used, but resource contention must be as close to zero as can be achieved. As a private cloud, it must contain tools to allow self-service of resources, rescale/redistribute capacity as required, and to track/monitor consumption.
- Infrastructure requirements – 451 Research has designed a private cloud of its own using list pricing of hardware infrastructure that we believe could support the virtualization requirements. The architecture consists of 25 servers, each with 500GB local storage, 96GB RAM, 2 x Intel E2670v2 10-core CPU, resilient power supplies, a 100TB iSCSI SAN, resilient load balancer and firewall, and supporting network fabric.

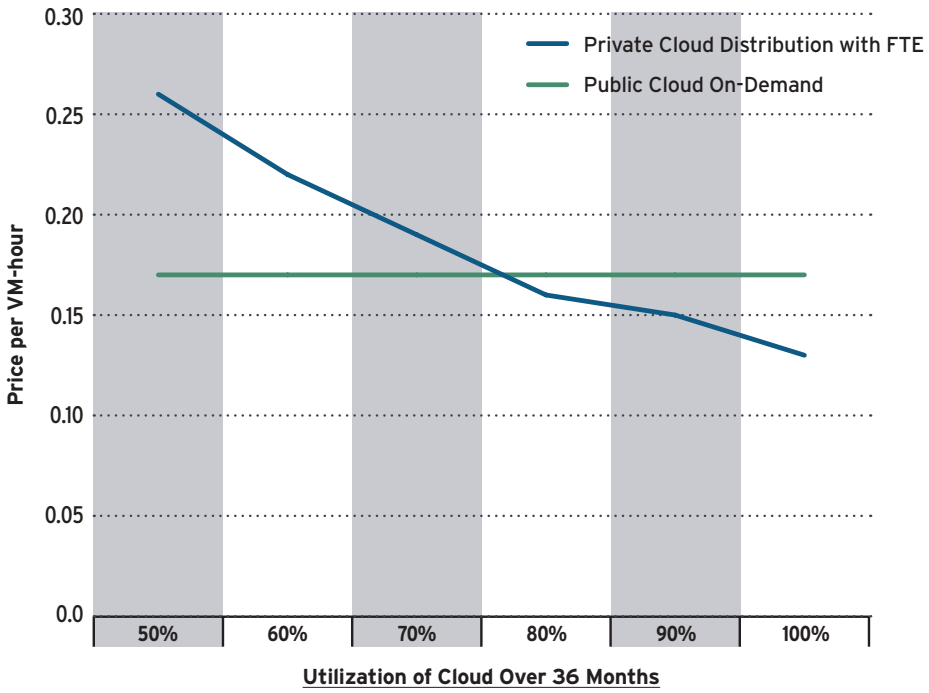
Both sets of requirements were sent to a range of cloud-orchestration vendors, which were invited to respond with a quote for virtualization and cloud-orchestration software and 24/7 support. In circumstances where vendors did not provide a response to our request, we made estimates using public pricing. Public cloud pricing was obtained through a similar process, for a similarly sized virtual machine.

We assume that the cost of capital is at a 4% annual percentage rate and that the comparison only relates to the direct cost of the cloud and does not take into account any internal costs or savings. Furthermore, the findings of this comparison do not necessarily translate to other scenarios; exact costs will vary depending on scenario, choice of cost models and manpower requirements, but this is a suitable baseline scenario. These assumptions do remove some realism from our analysis, but for the purpose of demonstrating issues, these assumptions are appropriate.

Assumptions

- The chart below shows the total cost of operating our virtualization requirements on a public cloud and a private cloud built on the CPI's average price of a pre-packaged OpenStack distribution with 24/7 support.
- All costs include infrastructure, hardware, orchestration, virtualization and a premium operating system.
- Further, we assume the private cloud is fully utilized throughout its life and can be supported by a single full-time OpenStack engineer earning a salary extracted from a third-party employment website of $126,000 per annum.

EXAMPLE PUBLIC VS. PRIVATE VIRTUAL MACHINE COST SCENARIO

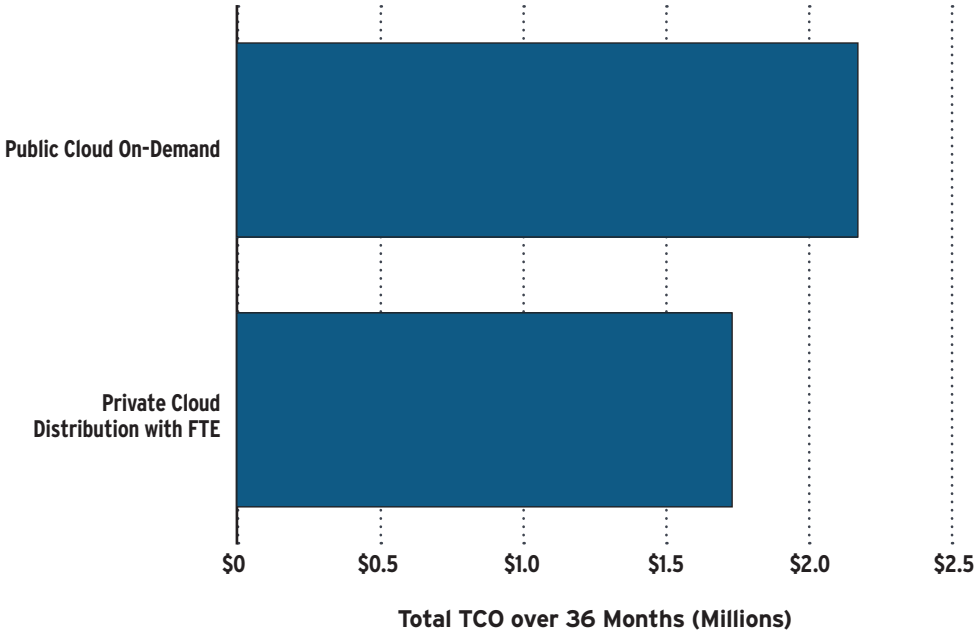


Looking at this figure, our comparison seems pretty clear-cut. ***However, we are making a huge assumption in this evaluation: we assume every VM is being used all the time.*** In some scenarios, this assumption might be reasonable – for example, in batch processing where tasks are queued, there may always be a steady stream of jobs ready to take free capacity when it is available. Applications that are not built to scale might also use a fixed capacity at all times, as would a cloud deployment that aggregates variability from many different applications to create an averaged high demand. However, most cloud applications should be able to grow and shrink as demand dictates; for these applications, it is likely that the utilization of our capacity will vary over time.

The total cost is an important consideration in building a business case for any cloud deployment, but this measure doesn't take into account how the deployment will be used. A fairer assessment of cost is how much each consumed VM costs individually. So, let's look at how the price per VM changes with utilization.

A fairer assessment of cost is how much each consumed VM costs individually. So, let's look at how the price per VM changes with utilization.

EXAMPLE PUBLIC VS. PRIVATE TCO SCENARIO



WHAT IS HAPPENING HERE?

In this scenario, the private cloud option requires capex to be paid up front and a fixed opex be paid every month. This private cloud's TCO remains the same, regardless of how many VMs are subsequently used. As the number of virtual machines, n, is consumed, the price per machine follows a 1/n relationship – the bigger the value of n, the smaller the cost per machine, and vice versa. On-demand pricing has no capex, so the cost per VM remains the same regardless of utilization.

In this scenario, representing a typical small-scale use case, the private cloud is a better value at a threshold of about 76%. In other words, if 76% of the private cloud's capacity is consumed throughout its lifetime, it will have a lower TCO.

Also, in this scenario, we aren't considering alternative pricing methods where discounts are rewarded for public cloud commitments or flexible models where private cloud opex might vary with demand. This fact, together with specific requirements (including manpower), will affect the utilization break-even point, but this example is suitable for drawing general conclusions.

In practical terms, the buyer in this scenario must make a risk assessment. If the organization believes it can achieve a high level of utilization, it can pay less using private cloud. But in return for this discount, it is taking on a risk; if its demand forecast is wrong, the per-VM cost will increase. An accurate utilization forecast means the most cost-effective option can be chosen prior to implementation. On-demand pricing has no risk because no forecast is required, and the buyer can spontaneously scale up or down. In this model, the public cloud provider takes on this risk – but the buyer pays a premium for this luxury.

So the question is: how can service providers ensure they achieve such a high level of utilization that they can have a lower cost base than public cloud?

So the question is: how can service providers ensure they achieve such a high level of utilization that they can have a lower cost base than public cloud?

CREATING LEVELS OF UTILIZATION

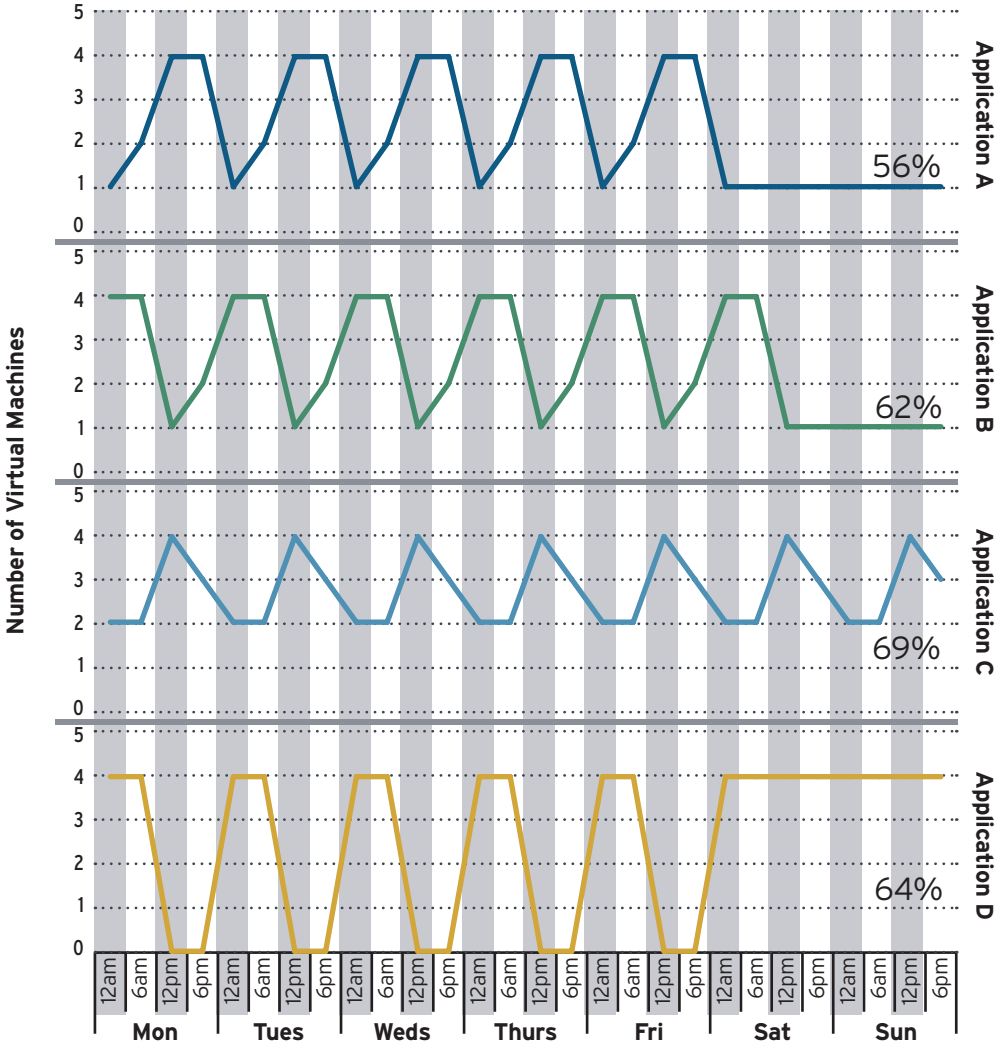
The chart below shows an example of changing VM demands for a variety of applications, with numbers intentionally chosen to be small for a simpler explanation. Over a week, these cloud-native applications consume VMs in different cycles.

- The first application might be an enterprise application that needs capacity during the day.
- The second application might be exactly the same app, but consumed in a different time zone.
- The third app might handle backup or archiving.
- The fourth app might process data overnight.

If each of these applications were to have its own private cloud that is architected to support its typical peak capacity, the utilization shown in the blue boxes would be reflected. Capacity planning such that performance can be maintained is important for end users' experience as applications scale, but it must also be kept as small as possible so that capacity isn't wasted – unused resources are sunk costs.

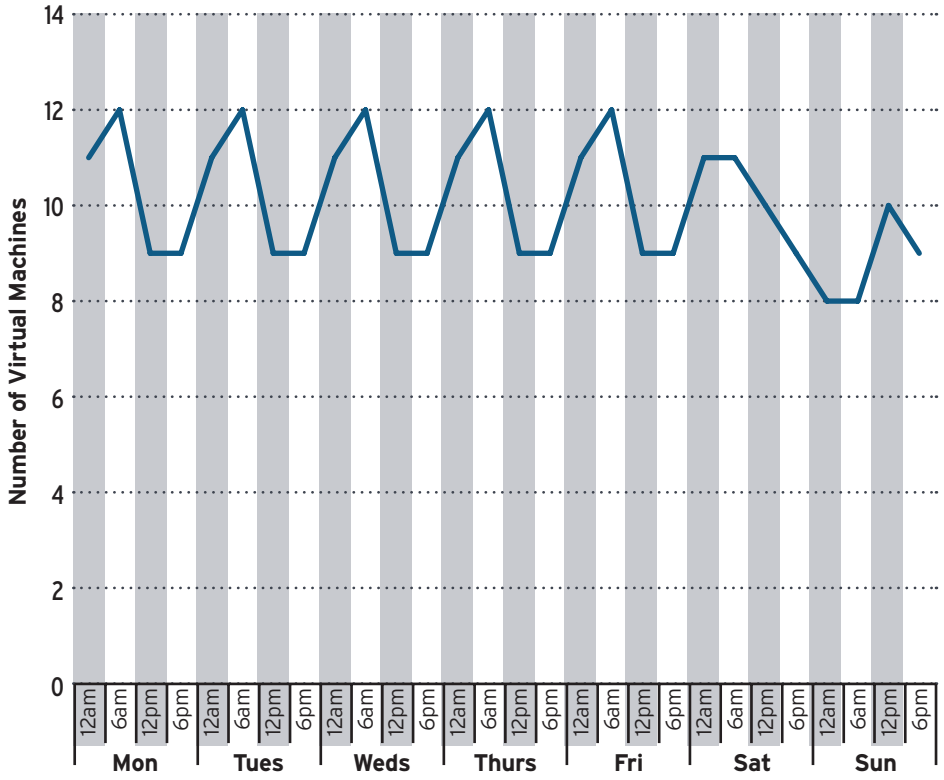
In our pricing scenario, none of these applications would have a TCO advantage by using private cloud over public cloud – from a purely financial point of view, it seems it would be better to deploy these applications on a public cloud.

EXAMPLE APPLICATION DEMAND PROFILES WITH UTILIZATION %



However, one private cloud, potentially hosted and managed by a service provider, could support all of these applications, so why not operate them as a single infrastructure? How does this affect utilization, and what does our total VM demand look like now?

EXAMPLE AGGREGATED DEMAND PROFILE



We haven't cut any corners. All of our applications are receiving exactly the same virtual machines they required previously, each containing exactly the same resources. The difference is that by aggregating workloads, we've increased our utilization. ***If pricing were similar to our CPI Private Cloud specification, we'd have crossed the TCO threshold, and private cloud would now be cheaper than public cloud. This workload superimposition means that our private cloud requires less capacity than the sum of its parts.***

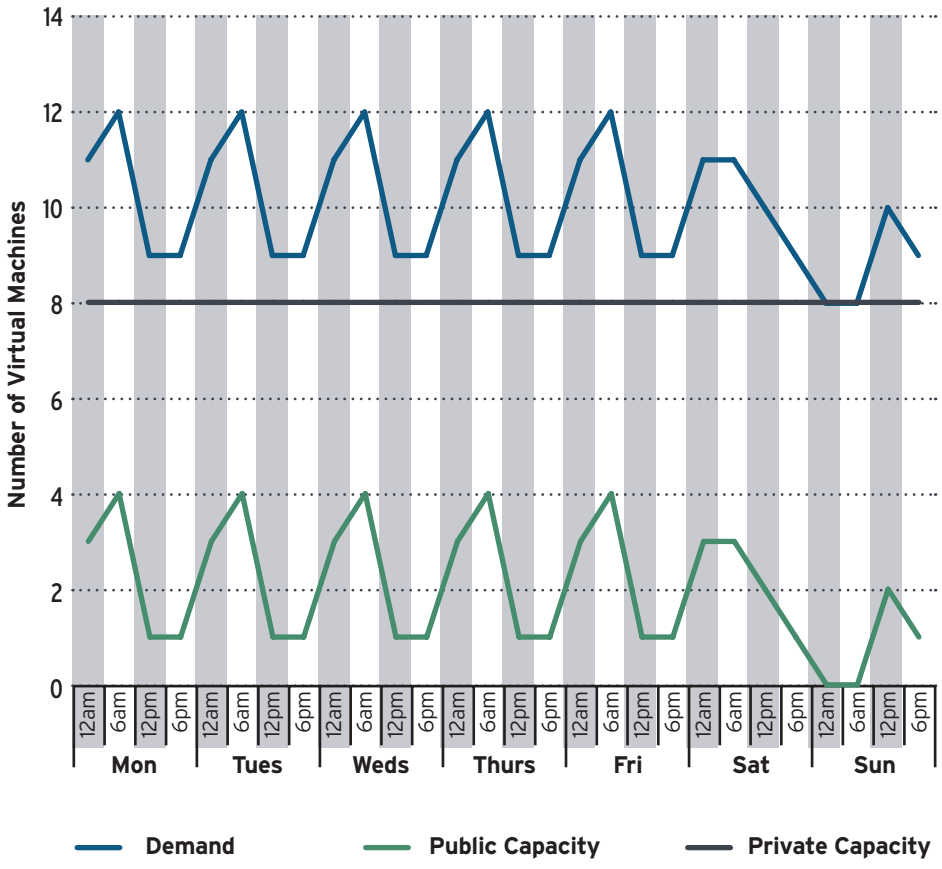
In fact, the more randomly impacted, independent applications situated together on a private cloud, the closer the private cloud will be utilized to a fixed capacity. And by having a fixed capacity, we can control our maximum utilization so costs don't spiral out of control. We can also dictate contention and over-provisioning – anecdotally, we hear of public cloud resources not being turned down when they are serving no useful purpose, essentially generating a cost implication without generating value. In this private cloud model, capacity that isn't being utilized can be automatically turned over to other applications.

HYBRID CLOUD

Is there a way we can increase our utilization even further? Yes, in a hybrid cloud model, we can use a private cloud for delivering a stable and constant capacity, and a public cloud for bursting above our maximum capacity. In the chart below, the private cloud is providing a minimum capacity of eight VMs, which are 100% utilized. When additional capacity is needed, the applications utilize VMs from a public cloud. With this hybrid approach, maximum efficiency is achieved: there is no sunk cost, no performance or resource constraints, and the cheapest resources are used when required.

In a hybrid cloud model, we can use a private cloud for delivering a stable and constant capacity, and a public cloud for bursting above our maximum capacity.

EXAMPLE AGGREGATED DEMAND PROFILE



OTHER APPLICATIONS

Workloads that aren't time-sensitive and that can be scheduled ad hoc are likely to help a private cloud reach higher utilization thresholds. In this manner, the private cloud can constantly be fully utilized – any spare capacity can be reallocated to applications already executing to speed up the result, or can be reallocated to a new application. This approach is ideal for big data; critical time-to-results processing; batch processing; economic, financial and scientific analytics; and even overnight archiving and indexing.

Service providers running private clouds could potentially offer spare capacity for a discounted rate, the proviso being that the VMs can be terminated spontaneously to make way for customers paying full price with real-time requirements. This means the infrastructure can be highly utilized, but customers that spontaneously need more capacity don't have to queue.

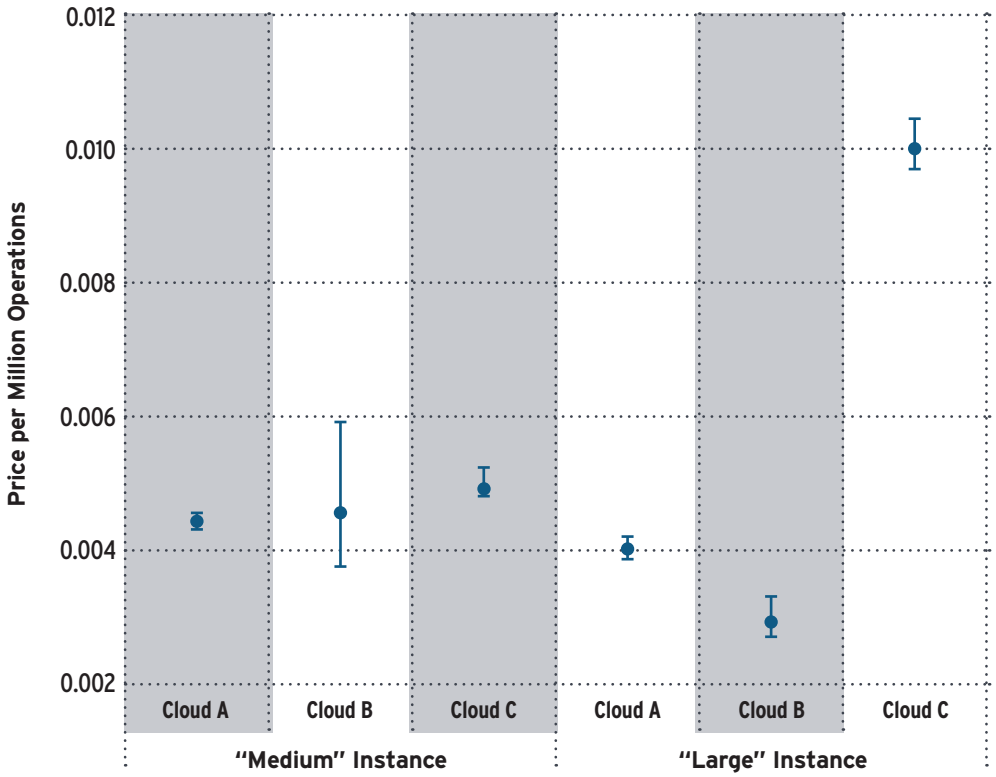
Not all applications need to change their scalability every few hours or to cycle every few days. Applications that have a fixed number of users or accounts, each requiring a fixed amount of resources, will have a fairly constant utilization. A company's workforce, for example, will usually stay constant over shorter time frames, or at least vary within a certain range.

Enterprise applications used extensively by all members of the workforce (or even teams) will typically have large utilizations, subject to good capacity planning. Of course, there will also be people entering and leaving a business, but huge spikes are unlikely to occur overnight. If there is a need to add a significant number of employees because of a large-scale recruitment or merger, the private cloud can scale by adding a bulk-capacity increment. Because there would typically be some notice that such a recruitment drive would take place, the capacity increase can be planned for. A private cloud partner that can rapidly provide additional hardware in response to additional demand can help ensure high utilization rates while maintaining levels of performance.

PERFORMANCE

One of the challenges of architecting any scalable application is consistent performance – how can the application grow and shrink while end users' experience remains the same? Much of this challenge lies in the architecture of the application to decouple components, but the infrastructure also has a role to play. The figure below shows anonymous price-per-million Cassandra transactions obtained on similar-sized instances on Azure, Google and Amazon Web Services (AWS) using performance benchmarking tool Perfkit and CPI data. The error bars show maximum and minimum price/performance from the five tests operating on each virtual machine under assessment.

CLOUD PRICE-PERFORMANCE ASSESSMENT OF CASSANDRA



PREDICTABILITY

The key takeaway from this diagram is inconsistency. Price/performance not only varies among cloud providers, but also sometimes even within the same provider depending on the time of day. Architecting an application to provide a consistent performance when the performance of individual nodes can vary is a challenge. In a private cloud, this challenge still exists but is far more controllable because the private cloud owner has access to the hypervisor and orchestration layer, and thus can control the allocation of resources. Private clouds by their nature are different in how they address end customers.

Public clouds provide generic infrastructure and services, which result in higher performance deviation. Private clouds are more nimble in that respect and can provide resources specific to the application to improve efficiency and reduce performance deviation. Therefore, service providers that deliver good value and consistent private clouds have a differentiator in the market.

THE BEST OF BOTH WORLDS

So, is private or public cloud cheaper?

The simple answer is that it depends on how the platform will be used and how it will be purchased, which is why private and public cloud users vary in their opinions on the cost-effectiveness of public and private clouds. In broad financial terms, on-demand public cloud is best for very spiky and unpredictable applications – applications can scale on-demand with no sunk cost brought about by unused capacity or degraded performance caused by a maximum capacity.

The crucial disclaimer here is that this scalability must be related to a material benefit: an application whose costs spiral is simply creating an expense if these costs aren't related to a revenue gain, for example. Private cloud and committed models are best for predictable *aggregated* capacity requirements. Aggregated is the key term here.

An application may be spiky or unpredictable, but if these spikes are offset by other applications operating on the same private cloud such that the combined capacity requirements are stable or more linear, then private cloud typically has the cost advantage. This is especially true for service providers that, by their very nature, aggregate applications. Batch processing, big data or low-/no-scalability applications are likely to exhibit predictable demand.

By combining the burst capability of public cloud with the guaranteed capacity of private cloud, end users get the best of both worlds. But it's not just in financial terms that public and private vary: control, security, compliance and integration are just some factors that will impact a buyer's decision.

Public and private clouds are both good options, depending on use case: buyers need to weigh the pros and cons of each. We hope our guidance helps buyers with that process.

Content Furnished by Hewlett Packard Enterprise

THE HEWLETT PACKARD ENTERPRISE VIEW

EVALUATING SERVICE PROVIDERS' IT NEEDS

With demonstrated inconsistency in performance and predictability, determining the most cost-effective approach to a service provider's IT needs is not a simple task. To obtain an accurate assessment, you must consider a number of factors, including the following:

- Performance and availability requirements
- Application and system dependencies
- Financial requirements and preferences
- Business growth and workload forecasts
- Compliance and security requirements
- Other anticipated changes

Hewlett Packard Enterprise takes a consultative approach to this challenge, working closely with service providers to understand how they are using cloud services before providing a detailed evaluation and recommendations. One goal of this process is to determine at what point a public cloud offering ceases to be cost-effective, and how much can be saved by moving workloads running on a public cloud to a different platform, including private cloud, colocation and hybrid options.

MAINTAINING A FLEXIBLE APPROACH TO UTILIZATION

Through recent innovations in financing and consumption models, HPE can help service providers implement flexible capacity, which enables control of on-site assets, data and management at a variable cost – where providers can adjust capacity based on their needs. This approach helps service providers control costs, improve agility and minimize risks. Flexible capacity strikes a balance between private and public – the control, security and performance of private cloud with the benefits of the elastic, opex model of public cloud.

IT INVESTMENT REVIEW AND PLANNING

As service providers review their infrastructure needs from a price/performance perspective, it makes sense to take a fresh look at systems through an IT investment filter. HPE Financial Services helps service providers evaluate issues such as variable-consumption models, traditional capital purchasing vs. aligning costs with revenue, pay-as-you-go/utilization considerations, opex vs. capex, and hardware/software suitability, regardless of the manufacturer. HPE can help craft an investment solution customized to each service provider's specific needs.

COST EFFICIENCY THROUGH CLOUD-BURSTING

In many cases, service providers have determined that the most cost-effective option for their data-processing needs is a hybrid approach, often with the capability of bursting into public cloud when needed. One advantage of hybrid cloud deployment is that an organization pays for extra computing resources only when they are needed. This is an important feature for service providers that experience utilization peaks around end-of-month processing requirements, holiday season business increases, or other seasonality-based utilization volumes.

A PUBLIC CLOUD PLATFORM ALTERNATIVE

One of the key HPE tools that service providers can use is *HPE Helion Eucalyptus*. Eucalyptus is one of the only open source cloud platforms designed to match the AWS API, enabling service providers to easily move public cloud workloads behind the firewall. This results in greater control over costs, data governance, application performance and workload mobility.

Service providers can continue to develop cloud applications using AWS design patterns and safely operate those applications within the firewall to protect customer data and comply with local regulations. Eucalyptus gives IT the ability to tune cloud infrastructure to match the needs of latency-sensitive applications and those requiring scale-up capabilities – without requiring changes to existing applications.

Eucalyptus also allows service providers to maintain their toolsets and performance levels while becoming more cost-effective. In fact, some Eucalyptus users have cut public cloud costs by as much as 50%, saving hundreds of thousands of dollars a year[1]. This offering allows service providers to burst into a public cloud when their utilization exceeds a predetermined level while users maintain the same application program interface and ecosystem for seamless scalability.

MASSIVE SCALABILITY AT A MOMENT'S NOTICE

As noted in a recent HPE case study, one service provider startup tried three major public cloud providers – all within a year – and then worked with HPE to implement a hybrid solution that grows, evolves and moves geographically as the service provider's business needs change. According to the service provider's CEO, "We needed to find a platform for our stack that would enable us to scale massively at a moment's notice. [A hybrid approach] gives me the ability to scale instantaneously – wherever our geographic and business needs might be." In this case, HPE was able to facilitate a workload-optimized solution that achieved twice the performance at a lower cost.

1. *See https://www.eucalyptus.com/customers/case-studies.*

FLEXIBLE MIGRATION ASSISTANCE

When system changes are required, HPE provides non-disruptive migration assistance, including redeployment behind the service provider's firewall and a cutover test before going live. HPE Workload Migration Service includes consultative services and a cloud migration SaaS platform, which automates the workload-migration process. Migration can be to the service provider's in-house resources or to a remote location. HPE can also help service providers understand and plan for storage, moving and transit costs.

Migration assistance can come in the form of accelerated migration services. By leveraging these capabilities, service providers can establish additional flexibility with existing infrastructure, speed transformation efforts and introduce technology more quickly into existing infrastructures.

THE BIG PICTURE

With its breadth of resources and extensive experience in providing solutions and consulting expertise, HPE is uniquely positioned to help service providers determine the best deployment model for a given service, with the ability to evaluate, migrate and federate workloads to meet cost, performance and security/compliance requirements.

NEXT STEPS

HPE offers a variety of resources for service providers, including:

- Case studies of service providers that have successfully migrated
- Workload-migration assessment service
- Solutions designed specifically for service providers

Access this content and get additional information at: *hpe.com/info/sp.*