

# THE ECONOMICS OF THE CLOUD

**Rolf Harms**

**Michael Yamartino**

*Computing is undergoing a seismic shift from client/server to the cloud, a shift similar in importance and impact to the transition from mainframe to client/server. Speculation abounds on how this new era will evolve in the coming years, and IT leaders have a critical need for a clear vision of where the industry is heading. We believe the best way to form this vision is to understand the underlying economics driving the long-term trend. In this paper, we will assess the economics of the cloud by using in-depth modeling. We then use this framework to better understand the long-term IT landscape.*

For comments or questions regarding the content of this paper, please contact  
Rolf Harms ([rolfh@microsoft.com](mailto:rolfh@microsoft.com)) or Michael Yamartino ([michael.yamartino@microsoft.com](mailto:michael.yamartino@microsoft.com))

## 1. INTRODUCTION

When cars emerged in the early 20<sup>th</sup> century, they were initially called “horseless carriages”. Understandably, people were skeptical at first, and they viewed the invention through the lens of the paradigm that had been dominant for centuries: the horse and carriage. The first cars also looked very similar to the horse and carriage (just without the horse), as engineers initially failed to understand the new possibilities of the new paradigm, such as building for higher speeds, or greater safety. Incredibly, engineers kept designing the whip holder into the early models before realizing that it wasn’t necessary anymore.

FIG. 1: HORSELESS CARRIAGE SYNDROME



Initially there was a broad failure to fully comprehend the new paradigm. Banks claimed that, “*The horse is here to stay but the automobile is only a novelty, a fad*”.

Even the early pioneers of the car didn’t fully grasp the potential impact their work could have on the world. When Daimler, arguably the inventor of the automobile, attempted to estimate the long-term auto market opportunity, he concluded there could never be more than 1 million cars, because of their high cost and the shortage of capable chauffeurs<sup>1</sup>.

By the 1920s the number of cars had already reached 8 million, and today there are over 600 million cars – proving Daimler wrong hundreds of times over. What the early pioneers failed to realize was that profound reductions in both cost and complexity of operating cars and a dramatic increase in its importance in daily life would overwhelm prior constraints and bring cars to the masses.

Today, IT is going through a similar change: the shift from client/server to the cloud. Cloud promises not just cheaper IT, but also faster, easier, more flexible, and more effective IT.

Just as in the early days of the car industry, it’s currently difficult to see where this new paradigm will take us. **The goal of this whitepaper is to help build a framework that allows IT leaders to plan for the cloud transition<sup>2</sup>.** We take a long-term view in our analysis, as this is a prerequisite when evaluating decisions and investments that could last for decades. As a result, we focus on the economics of cloud rather than on specific technologies or other driving factors like organizational change, as economics often provide a clearer understanding of transformations of this nature.

In Section 2, we outline the underlying economics of cloud, focusing on what makes it truly different from client/server. In Section 3, we will assess the implications of these economics for the future of IT. We will discuss the positive impact cloud will have but will also discuss the obstacles that still exist today. Finally, in Section 4 we will discuss what’s important to consider as IT leaders embark on the journey to the cloud.

<sup>1</sup> Source: Horseless Carriage Thinking, William Horton Consulting.

<sup>2</sup> Cloud in this context refers to cloud computing architecture, encompassing both public and private clouds.

## 2. ECONOMICS OF THE CLOUD




Economics are a powerful force in shaping industry transformations. Today's discussions on the cloud focus a great deal on technical complexities and adoption hurdles. While we acknowledge that such concerns exist and are important, historically, underlying economics have a much stronger impact on the direction and speed of disruptions, as technological challenges are resolved or overcome through the rapid innovation we've grown accustomed to (Fig. 2). During the mainframe era, client/server was initially viewed as a "toy" technology, not viable as a mainframe replacement. Yet, over time the client/server technology found its way into the enterprise (Fig. 3). Similarly, when virtualization technology was first proposed, application compatibility concerns and potential vendor lock-in were cited as barriers to adoption. Yet underlying economics of 20 to 30 percent savings<sup>3</sup> compelled CIOs to overcome these concerns, and adoption quickly accelerated.

The emergence of cloud services is again fundamentally shifting the economics of IT. Cloud technology standardizes and pools IT resources and automates many of the maintenance tasks done manually today. Cloud architectures facilitate elastic consumption, self-service, and pay-as-you-go pricing.

Cloud also allows core IT infrastructure to be brought into large data centers that take advantage of significant economies of scale in three areas:

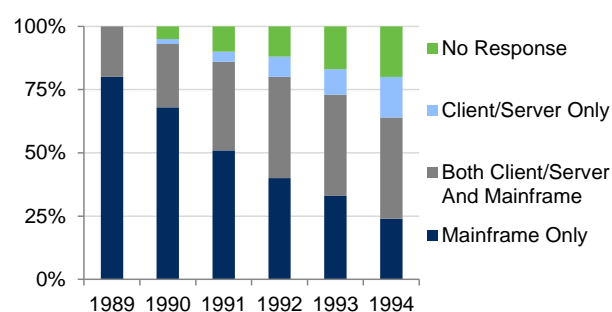
- **Supply-side savings.** Large-scale data centers (DCs) lower costs per server.
- **Demand-side aggregation.** Aggregating demand for computing smooths overall variability, allowing server utilization rates to increase.
- **Multi-tenancy efficiency.** When changing to a multitenant application model, increasing the number of tenants (i.e., customers or users) lowers the application management and server cost per tenant.

**FIG. 2: CLOUD OPPORTUNITY**

		Technology	Economic	Business Model
Mainframe		Centralized compute and storage Thin clients	Optimized for efficiency because of the high cost	High up-front costs for hardware and software
Client/Server		PCs and servers for distributed compute, storage, and so on	Optimized for agility because of the low cost	Perpetual license for OS and application software
Cloud		Large DCs, ability to scale, commodity hardware, devices	Efficiency and agility an order of magnitude better	Ability to pay as you go, and only for what you use

Source: Microsoft.

**FIG. 3: BEGINNING THE TRANSITION TO CLIENT/SERVER TECHNOLOGY**



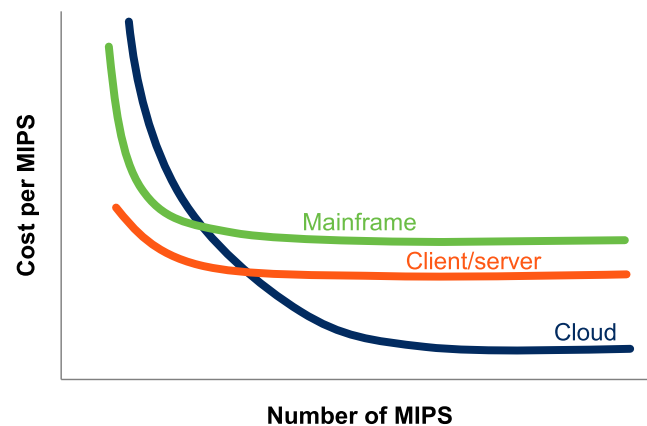
Source: "How convention shapes our market" longitudinal survey, Shana Greenstein, 1997.

<sup>3</sup> Source: "Dataquest Insight: Many Midsize Businesses Looking Toward 100% Server Virtualization". Gartner, May 8, 2009.

## 2.1 Supply-Side Economies of Scale

Cloud computing combines the best economic properties of mainframe and client/server computing. The mainframe era was characterized by significant economies of scale due to high up-front costs of mainframes and the need to hire sophisticated personnel to manage the systems. As required computing power – measured in MIPS (million instructions per second) – increased, cost declined rapidly at first (Fig. 4), but only large central IT organizations had the resources and the aggregate demand to justify the investment. Due to the high cost, resource utilization was prioritized over end-user agility. Users' requests were put in a queue and processed only when needed resources were available.

FIG. 4: ECONOMIES OF SCALE (ILLUSTRATIVE)



Source: Microsoft.

With the advent of minicomputers and later client/server technology, the minimum unit of purchase was greatly reduced, and the resources became easier to operate and maintain. This modularization significantly lowered the entry barriers to providing IT services, radically improving end-user agility. However, there was a significant utilization tradeoff, resulting in the current state of affairs: datacenters sprawling with servers purchased for whatever need existed at the time, but running at just 5%-10% utilization<sup>4</sup>.

Cloud computing is not a return to the mainframe era as is sometimes suggested, but in fact offers users economies of scale and efficiency that exceed those of a mainframe, coupled with modularity and agility beyond what client/server technology offered, thus eliminating the tradeoff.

The economies of scale emanate from the following areas:

- **Cost of power.** Electricity cost is rapidly rising to become the largest element of total cost of ownership (TCO),<sup>5</sup> currently representing 15%-20%. Power Usage Effectiveness (PUE)<sup>6</sup> tends to be significantly lower in large facilities than in smaller ones. While the operators of small data centers must pay the prevailing local rate for electricity, large providers can pay less than one-fourth of the national average rate by locating its data centers in locations with inexpensive electricity supply and through bulk purchase agreements.<sup>7</sup> In addition, research has shown that operators of multiple data centers are able to take advantage of geographical variability in electricity rates, which can further reduce energy cost.

<sup>4</sup> Source: *The Economics of Virtualization: Moving Toward an Application-Based Cost Model*, IDC, November 2009.

<sup>5</sup> Not including app labor. Studies suggest that for low-efficiency datacenters, three-year spending on power and cooling, including infrastructure, already outstrips three-year server hardware spending.

<sup>6</sup> Power Utilization Effectiveness equals total power delivered into a datacenter divided by "critical power" – the power needed to actually run the servers. Thus, it measures the efficiency of the datacenter in turning electricity into computation. The best theoretical value is 1.0, with higher numbers being worse.

<sup>7</sup> Source: U.S. Energy Information Administration (July 2010) and Microsoft. While the average U.S. commercial rate is 10.15 cents per kilowatt hour, some locations offer power for as little as 2.2 cents per kilowatt hour

- **Infrastructure labor costs.** While cloud computing significantly lowers labor costs at any scale by automating many repetitive management tasks, larger facilities are able to lower them further than smaller ones. While a single system administrator can service approximately 140 servers in a traditional enterprise,<sup>8</sup> in a cloud data center the same administrator can service thousands of servers. This allows IT employees to focus on higher value-add activities like building new capabilities and working through the long queue of user requests every IT department contends with.
- **Security and reliability.** While often cited as a potential hurdle to public cloud adoption, increased need for security and reliability leads to economies of scale due to the largely fixed level of investment required to achieve operational security and reliability. Large commercial cloud providers are often better able to bring deep expertise to bear on this problem than a typical corporate IT department, thus actually making cloud systems more secure and reliable.
- **Buying power.** Operators of large data centers can get discounts on hardware purchases of up to 30 percent over smaller buyers. This is enabled by standardizing on a limited number of hardware and software architectures. Recall that for the majority of the mainframe era, more than 10 different architectures coexisted. Even client/server included nearly a dozen UNIX variants and the Windows Server OS, and x86 and a handful of RISC architectures. Large-scale buying power was difficult in this heterogeneous environment. With cloud, infrastructure homogeneity enables scale economies.

Going forward, there will likely be many additional economies of scale that we cannot yet foresee. The industry is at the early stages of building data centers at a scale we've never seen before (Fig. 5). The massive aggregate scale of these mega DCs will bring considerable and ongoing R&D to bear on running them more efficiently, and make them more efficient for their customers. Providers of large-scale DCs, for which running them is a primary business goal, are likely to benefit more from this than smaller DCs which are run inside enterprises.

**FIG. 5: RECENT LARGE DATA-CENTER PROJECTS**

Company	Location	Cost (\$ in millions)	Size (in sq. feet)
Internet Villages JUL 2009	Annandale, Scotland	1,600	3,000,000
National Security Admin. JUL 2009	Camp Williams, Utah	2,000	1,000,000
Lockerbie Data Centers DEC 2009	Lockerbie, Scotland	1,500	N/A
Microsoft SEP 2009	Chicago, Illinois	500	700,000
I/O Data Centers JUN 2009	Phoenix, Arizona	N/A	538,000
Apple MAY 2009	Maiden, North Carolina	1,000	500,000
Microsoft JUN 2010	Dublin, Ireland	500	N/A
U.S. Social Security Admin. FEB 2009	Baltimore, Maryland	400	N/A
Facebook FEB 2010	Prineville, Oregon	N/A	307,000
Next Generation Data MAR 2010	Cardiff, Wales	301	N/A

Sources: Press releases.

## 2.2 Demand-Side Economies of Scale

The overall cost of IT is determined not just by the cost of capacity, but also by the degree to which the capacity is efficiently utilized. We need to assess the impact that demand aggregation will have on costs of actually utilized resources (CPU, network, and storage).<sup>9</sup>

In the non-virtualized data center, each application/workload typically runs on its own physical server.<sup>10</sup> This means the number of servers scales linearly with the number of server workloads. In this model,

<sup>8</sup> Source: James Hamilton, Microsoft Research, 2006.

<sup>9</sup> In this paper, we talk generally about "resource" utilization. We acknowledge there are important differences among resources. For example, because storage has fewer usage spikes compared with CPU and I/O resources, the impact of some of what we discuss here will affect storage to a smaller degree.

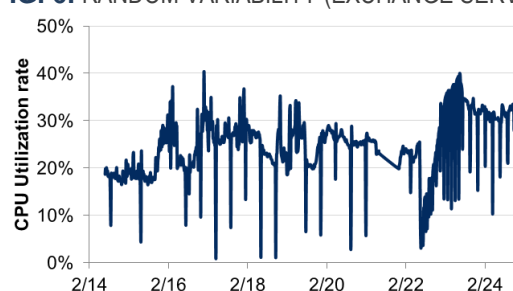
utilization of servers has traditionally been extremely low, around 5 to 10 percent.<sup>11</sup> Virtualization enables multiple applications to run on a single physical server within their optimized operating system instance, so the primary benefit of virtualization is that fewer servers are needed to carry the same number of workloads. But how will this affect economies of scale? If all workloads had constant utilization, this would entail a simple unit compression without impacting economies of scale. In reality, however, workloads are highly variable over time, often demanding large amounts of resources one minute and virtually none the next. This opens up opportunities for utilization improvement via demand-side aggregation and diversification.

We analyzed the different sources of utilization variability and then looked at the ability of the cloud to diversify it away and thus reduce costs.

We distinguish five sources of variability and assess how they might be reduced:

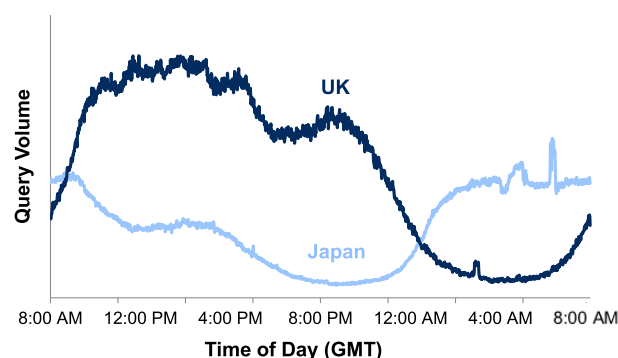
1. **Randomness.** End-user access patterns contain a certain degree of randomness. For example, people check their email at different times (Fig. 6). To meet service level agreements, capacity buffers have to be built in to account for a certain probability that many people will undertake particular tasks at the same time. If servers are pooled, this variability can be reduced.
2. **Time-of-day patterns.** There are daily recurring cycles in people's behavior: consumer services tend to peak in the evening, while workplace services tend to peak during the workday. Capacity has to be built to account for these daily peaks but will go unused during other parts of the day causing low utilization. This variability can be countered by running the same workload for multiple time zones on the same servers (Fig. 7) or by running workloads with complementary time-of-day patterns (for example, consumer services and enterprise services) on the same servers.

**FIG. 6: RANDOM VARIABILITY (EXCHANGE SERVER)**



Source: Microsoft.

**FIG. 7: TIME-OF-DAY PATTERNS FOR SEARCH**



Source: Bing Search volume over 24-hour period.

<sup>10</sup> Multiple applications can run on a single server, of course, but this is not common practice. It is very challenging to move a running application from one server to another without also moving the operating system, so running multiple applications on one operating system instance can create bottlenecks that are difficult to remedy while maintaining service, thereby limiting agility. Virtualization allows the application plus operating system to be moved at will.

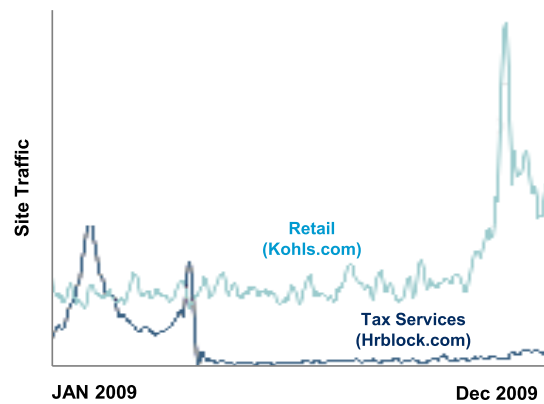
<sup>11</sup> Source: *The Economics of Virtualization: Moving Toward an Application-Based Cost Model*, IDC, November 2009.

3. **Industry-specific variability.** Some variability is driven by industry dynamics. Retail firms see a spike during the holiday shopping season while U.S. tax firms will see a peak before April 15 (Fig. 8). There are multiple kinds of industry variability — some recurring and predictable (such as the tax season or the Olympic Games), and others unpredictable (such as major news stories). The common result is that capacity has to be built for the expected peak (plus a margin of error). Most of this capacity will sit idle the rest of the time. Strong diversification benefits exist for industry variability.

4. **Multi-resource variability.** Compute, storage, and input/output (I/O) resources are generally bought in bundles: a server contains a certain amount of computing power (CPU), storage, and I/O (e.g., networking or disk access). Some workloads like search use a lot of CPU but relatively little storage or I/O, while other workloads like email tend to use a lot of storage but little CPU (Fig. 9). While it's possible to adjust capacity by buying servers optimized for CPU or storage, this addresses the issue only to a limited degree because it will reduce flexibility and may not be economic from a capacity perspective. This variability will lead to resources going unutilized unless workload diversification is employed by running workloads with complementary resource profiles.

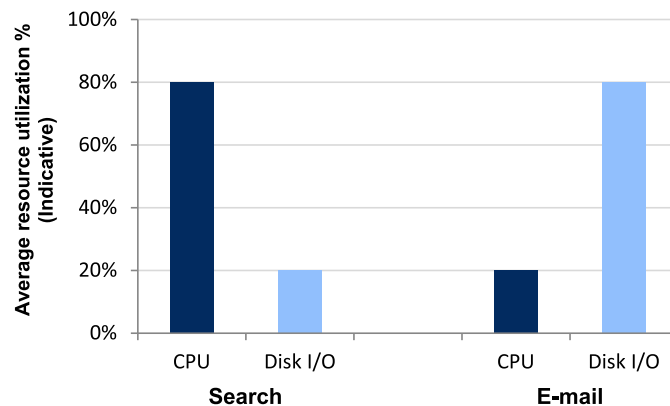
5. **Uncertain growth patterns.** The difficulty of predicting future need for computing resources and the long lead-time for bringing capacity online is another source of low utilization (Fig. 10). For startups, this is sometimes referred to as the “TechCrunch effect.” Enterprises and small businesses all need to secure

**FIG. 8: INDUSTRY-SPECIFIC VARIABILITY**



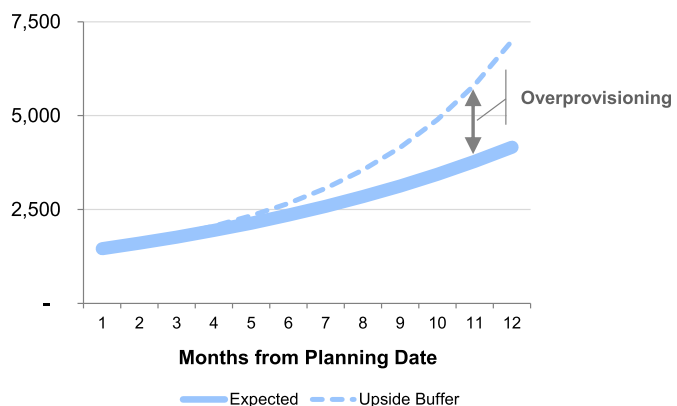
Source: Alexa Internet.

**FIG. 9: MULTIRESOURCE VARIABILITY (ILLUSTRATIVE)**



Source: Microsoft.

**FIG.10: UNCERTAIN GROWTH PATTERNS**



Source: Microsoft.

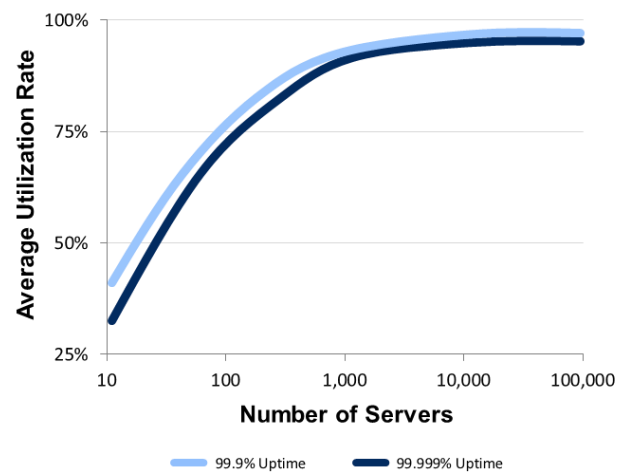


approval for IT investments well in advance of actually knowing their demand for infrastructure. Even large private companies face this challenge, with firms planning their purchases six to twelve months in advance (Fig. 10). By diversifying among workloads across multiple customers, cloud providers can reduce this variability, as higher-than-anticipated demand for some workloads is canceled out by lower-than-anticipated demand for others.

A key economic advantage of the cloud is its ability to address variability in resource utilization brought on by these factors. By pooling resources, variability is diversified away, evening out utilization patterns. The larger the pool of resources, the smoother the aggregate demand profile, the higher the overall utilization rate, and the cheaper and more efficiently the IT organization can meet its end-user demands.

We modeled the theoretical impact of **random variability** of demand on server utilization rates as we increase the number of servers.<sup>12</sup> Fig. 11 indicates that a theoretical pool of 1,000 servers could be run at approximately 90% utilization without violating its SLA. This only holds true in the hypothetical situation where random variability is the only source of variability and workloads can be migrated between physical servers instantly without interruption. Note that higher levels of uptime (as defined in a service level agreement or SLA) become much easier to deliver as scale increases.

FIG. 11: DIVERSIFYING RANDOM VARIABILITY



Source: Microsoft.

Clouds will be able to reduce **time-of-day variability** to the extent that they are diversified amongst geographies and workload types. Within an average organization, peak IT usage can be twice as high as the daily average. Even in large, multi-geography organizations, the majority of employees and users will live in similar time zones, bringing their daily cycles close to synchrony. Also, most organizations do not tend to have workload patterns that offset one another: for example, the email, network and transaction processing activity that takes place during business hours is not replaced by an equally active stream of work in the middle of the night. Pooling organizations and workloads of different types allows these peaks and troughs to be offset.

**Industry variability** results in highly correlated peaks and troughs throughout each firm (that is, most of the systems in a retail firm will be at peak capacity around the holiday season (e.g., web servers, transaction processing, payment processing, databases)).<sup>13</sup> Fig. 12 shows industry variability for a number of different industries, with peaks ranging from 1.5x to 10x average usage.

<sup>12</sup> To calculate economies of scale arising from diversifying random variability, we created a Monte Carlo model to simulate data centers of various sizes serving many random workloads. For each simulated DC, workloads (which are made to resemble hypothetical web usage patterns) were successively added until the expected availability of server resources dropped below a given uptime of 99.9 percent or 99.99 percent. The maximum number of workloads determines the maximum utilization rate at which the DC's servers can operate without compromising performance.

<sup>13</sup> Ideally, we would use the server utilization history of a large number of customers to gain more insight into such patterns. However, this data is difficult to get and often of poor quality. We therefore used web traffic as a proxy for the industry variability.



Microsoft services such as Windows Live Hotmail and Bing take advantage of **multi-resource diversification** by layering different subservices to optimize workloads with different resource profiles (such as CPU bound or storage bound). It is difficult to quantify these benefits, so we have not included multi-resource diversification in our model.

Some **uncertain growth pattern variability** can be reduced by hardware standardization and just-in-time procurement, although likely not completely. Based on our modeling, the impact of growth uncertainty for enterprises with up to 1,000 servers is 30 to 40 percent overprovisioning of servers relative to a public cloud service. For smaller companies (for example, Internet startups), the impact is far greater.

So far we have made the implicit assumption that the degree of variability will stay the same as we move to the cloud. In reality, it is likely that the variability will significantly increase, which will further increase economies of scale. There are two reasons why this may happen:

- **Higher expectation of performance.** Today, users have become accustomed to resource constraints and have learned to live with them. For example, users will schedule complex calculations to run overnight, avoid multiple model iterations, or decide to forgo time-consuming and costly supply chain optimizations. The business model of cloud allows a user to pay the same for 1 machine running for 1,000 hours as he would for 1,000 machines running for 1 hour. Today, the user would likely wait 1,000 hours or abandon the project. In the cloud, there is virtually no additional cost to choosing 1,000 machines and accelerating such processes. This will have a dramatic impact on variability. Pixar Animation Studios, for example runs its computer-animation rendering process on Windows Azure because every frame of their movies takes eight hours to render today on a single processor, meaning it would take 272 years to render an entire movie. As they said, “We are not that patient.” With Azure, they can get the job done as fast as they need. The result is huge spikes in Pixar’s usage of Azure as they render on-demand.
- **Batch processes will become real time.** Many processes — for example, accurate stock availability for online retailers — that were previously batch driven, will move to real-time. Thus, multi-stage processes that were once sequential will now occur simultaneously, such as a manufacturing firm that can tally its inventory, check its order backlog, and order new supplies at once. This will amplify utilization variability.

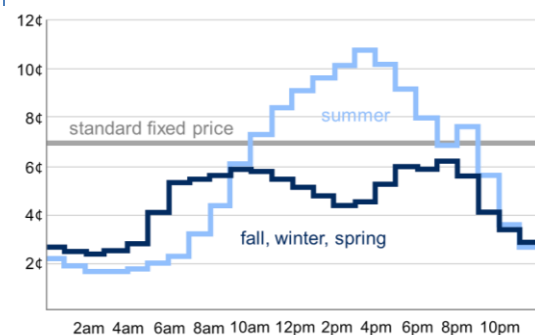
We note that even the largest public clouds will not be able to diversify away all variability; market level variability will likely remain. To further smooth demand, sophisticated pricing can be employed. For example, similar to the electricity market (Fig. 13), customers can be incented to shift their demand from high utilization periods to low utilization periods. In addition, a lower price spurs additional usage from customers due to price elasticity of demand. Demand management will further increase the economic benefits of cloud.

**FIG. 12: INDUSTRY VARIABILITY**

Company	Peak Traffic/ Average Traffic
Tax Services	10x
General Retail	4x
Sports (NFL)	2.5x
Travel (airlines, hotels)	1.5x
News	1.5x – 2.0x

Source: Microsoft, Alexa Internet, Inc.

**FIG. 13: VARIABLE PRICING IN ELECTRICITY**



Source: Ameren Illinois Utilities.

## 2.3 Multi-tenancy Economies of Scale

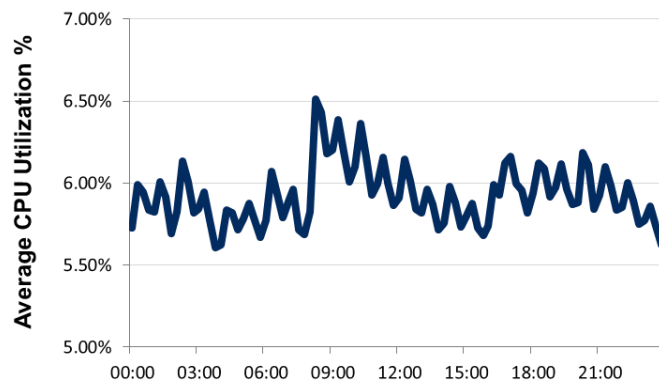
The previously described supply-side and demand-side economies of scale can be achieved independent of the application architecture, whether it be traditional scale-up or scale-out, single tenant or multitenant. There is another important source of economies of scale that can be harnessed *only* if the application is written as a multitenant application. That is, rather than running an application instance for each customer – as is done for on-premises application and most hosted applications such as dedicated instances of Microsoft Office 365 – in a multitenant application, multiple customers use a single instance of the application simultaneously, as in the case of *shared* Office 365. This has two important economic benefits:

- **Fixed application labor amortized over a large number of customers.**

In a single-tenant instance, each customer has to pay for its own application management (that is, the labor associated with update and upgrade management and incident resolution). We've examined data from customers, as well as Office 365-D and Office 365-S to assess the impact. In dedicated instances, the same activities, such as applying software patches, are performed multiple times – once for each instance. In a multi-tenant instance such as Office 365-S, that

cost is shared across a large set of customers, driving application labor costs per customer towards zero. This can result in a meaningful reduction in overall cost, especially for complex applications.

**FIG. 14: UTILIZATION OVERHEAD**



Source: Microsoft.

- **Fixed component of server utilization amortized over large number of customers.** For each application instance, there is a certain amount of server overhead. Fig. 14 shows an example from Microsoft's IT department in which intraday variability appears muted (only a 16 percent increase between peak and trough) compared to actual variability in user access. This is caused by application and runtime overhead, which is constant throughout the day. By moving to a multitenant model with a single instance, this resource overhead can be amortized across all customers. We have examined Office 365-D, Office 365-S, and Microsoft Live@edu data to estimate this overhead, but so far it has proven technically challenging to isolate this effect from other variability in the data (for example, user counts and server utilization) and architectural differences in the applications. Therefore, we currently assume no benefit from this effect in our model.

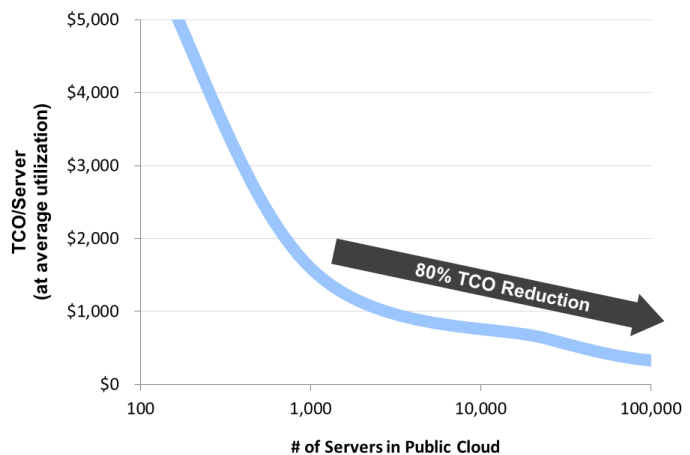
Applications can be entirely multitenant by being completely written to take advantage of these benefits, or can achieve partial multi-tenancy by leveraging shared services provided by the cloud platform. The greater the use of such shared services, the greater the application will benefit from these multi-tenancy economies of scale.

## 2.4 Overall Impact

The combination of supply-side economies of scale in server capacity (amortizing costs across more servers), demand-side aggregation of workloads (reducing variability), and the multi-tenant application model (amortizing costs across multiple customers) leads to powerful economies of scale. To estimate the magnitude, we built a cost scaling model which estimates the long term behavior of costs.

Fig. 15 shows the output for a workload that utilizes 10 percent of a traditional server. The model indicates that a 100,000-server datacenter has an 80% lower total cost of ownership (TCO) compared to a 1,000-server datacenter.

**FIG. 15: ECONOMIES OF SCALE IN THE CLOUD**

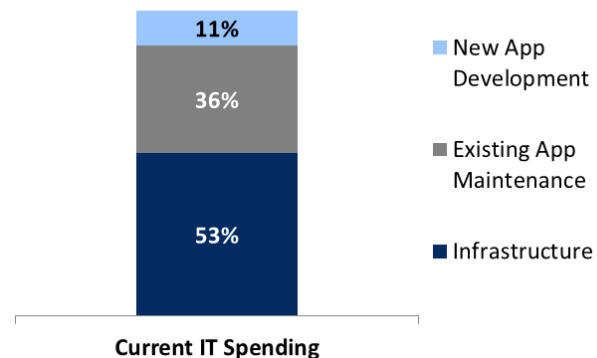


Source: Microsoft.

This raises the question: what impact will the Cloud Economics we described have on the IT budget? From customer data, we know the approximate breakdown between the infrastructure costs, costs of supporting and maintaining existing applications, and new application development costs (Fig. 16). Cloud impacts all three of these areas. The supply-side and demand-side savings impact mostly the infrastructure portion, which comprises over half of spending. Existing app maintenance costs include update and patching labor, end-user support, and license fees paid to vendors. They account for roughly a third of spending and are addressed by the multi-tenancy efficiency factor.

New application development accounts for just over a tenth of spending<sup>14</sup>, even though it is seen as the way for IT to innovate. Therefore IT leaders generally want to increase spending here. The economic benefits of cloud computing described here will enable this by freeing up room in the budget to do so. We will touch more on this aspect in the next paragraph as well as in Section 3.

**FIG. 16: IT SPENDING BREAKDOWN**



Source: Microsoft.

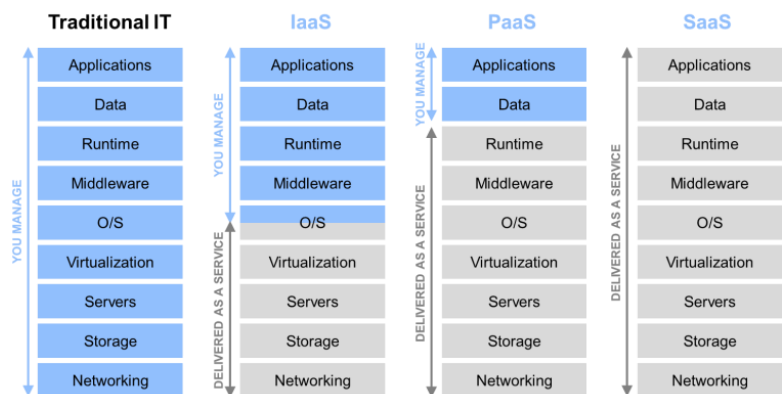
<sup>14</sup> New application development costs include only the cost of designing and writing the application and excluding the cost of hosting them on new infrastructure. Adding these costs results in the 80% / 20% split seen elsewhere.

## 2.5 Harnessing Cloud Economics

Capturing the benefits described above is not a straightforward task with today's technology. Just as engineers had to fundamentally rethink design in the early days of the car, so too will developers have to rethink design of applications. Multi-tenancy and demand-side aggregation is often difficult for developers or even sophisticated IT departments to implement on their own. If not done correctly, it could end up either significantly raising the costs of developing applications (thus at least partially nullifying the increased budget room for new app development); or capturing only a small subset of the savings previously described. The best approach in harnessing the cloud economics is different for packaged apps vs. new/custom apps.

**Packaged applications:** While virtualizing packaged applications and moving them to cloud virtual machines (e.g., virtualized Exchange) can generate some savings, this solution is far from ideal and fails to capture the full benefits outlined in this Section. The cause is twofold. First, applications designed to be run on a single server will not easily scale up and down without significant additional programming to add load-balancing, automatic failover, redundancy, and active resource management. This limits the extent to which they are able to aggregate demand and increase server utilization. Second, traditional packaged applications are not written for multi-tenancy, and simply hosting them in the cloud does not change this. For packaged apps, the best way to harness the benefits of cloud is to use SaaS offerings like Office365, which have been architected for scale-out and multi-tenancy to capture the full benefits.

**FIG. 17: CAPTURING CLOUD BENEFITS**



Source: Microsoft.

**New/custom applications:** Infrastructure-as-a-Service (IaaS) can help capture some of the economic benefits for existing applications. Doing so is, however, a bit of a "horseless carriage" in that the underlying platform and tools were not designed specifically for the cloud. The full advantage of cloud computing can only be properly unlocked through a significant investment in intelligent resource management. The resource manager must understand both the status of the resources (networking, storage, and compute) as well as the activity of the applications being run. Therefore, when writing new apps, Platform as a Service most effectively captures the economic benefits. PaaS offers shared services, advanced management, and automation features that allow developers to focus directly on application logic rather than on engineering their application to scale.

To illustrate the impact, a startup named Animoto used Infrastructure-as-a-Service (IaaS) to enable scaling – adding over 3,500 servers to their capacity in just 3 days as they served over three-quarters of a million new users. Examining their application later, however, the Animoto team discovered that a large percentage of the resources they were paying for were often sitting idle – often over 50%, even in a supposedly elastic cloud. They re-architected their application and eventually lowered operating costs by 20%. While Animoto is a cloud success story, it was only after an investment in intelligent resource

management that they were able to harness the full benefits of cloud. PaaS would have delivered many of these benefits “out-of-the-box” without any additional tweaking required.

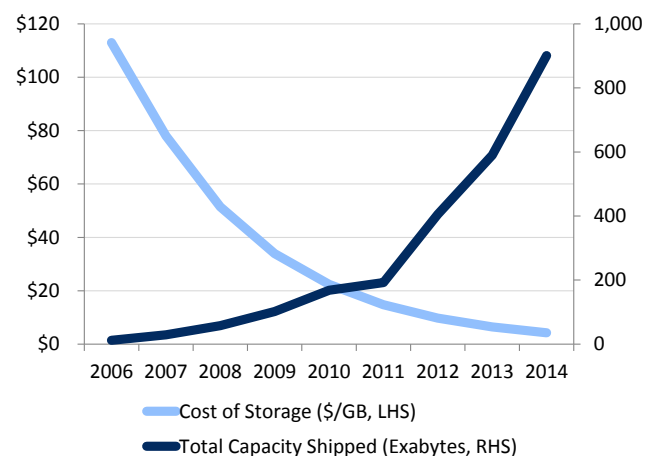
### 3. IMPLICATIONS

In this Section, we will discuss the implications of the previously described economics of cloud. We will discuss the ability of private clouds to address some of the barriers to cloud adoption and assess the cost gap between public and private clouds.

#### 3.1 Possibilities & Obstacles

The economics we described in section 2 will have a profound impact on IT. Many IT leaders today are faced with the problem that 80% of the budget is spent on “keeping the lights on,” maintaining existing services and infrastructure. This leaves few resources available for innovation or addressing the never-ending queue of new business and user requests. Cloud computing will free up significant resources that can be redirected to innovation. Demand for general purpose technologies like IT has historically proven to be very price elastic (Fig. 18). Thus, many IT projects that previously were cost prohibitive will now become viable thanks to cloud economics. However, lower TCO is only one of the key drivers that will lead to a renewed level of innovation within IT:

**FIG. 18: PRICE ELASTICITY OF STORAGE**



Source: Coughlin Associates.

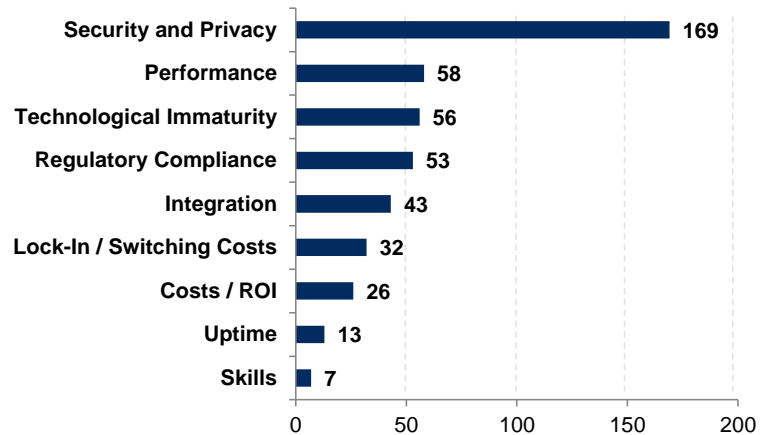
1. **Elasticity is a game-changer** because, as described before, renting 1 machine for 1,000 hours will be nearly equivalent to renting 1,000 machines for 1 hour in the cloud. This will enable users and organizations to rapidly accomplish complex tasks that were previously prohibited by cost or time constraints. Being able to both scale up and scale down resource intensity nearly instantly enables a new class of experimentation and entrepreneurship.
2. **Elimination of capital expenditure** will significantly lower the risk premium of projects, allowing for more experimentation. This both lowers the costs of starting an operation and lowers the cost of failure or exit – if an application no longer needs certain resources, they can be decommissioned with no further expense or write-off.
3. **Self-service** Provisioning servers through a simple web portal rather than through a complex IT procurement and approval chain can lower friction in the consumption model, enabling rapid provisioning and integration of new services. Such a system also allows projects to be completed in less time with less risk and lower administrative overhead than previously.
4. **Reduction of complexity.** Complexity has been a long standing inhibitor of IT innovation. From an end-user perspective SaaS is setting a new bar for user friendly software. From a developer perspective Platform as a Service (PaaS) greatly simplifies the process of writing

new applications, in the same way as cars greatly reduced the complexity of transportation by eliminating, for example, the need to care for a horse.

These factors will significantly increase the value add delivered by IT. Elasticity enables applications like yield management, complex event processing, logistics optimization, and Monte Carlo simulation, as these workloads exhibit nearly infinite demand for IT resources. The result will be massively improved experience, including scenarios like real-time business intelligence analytics and HPC for the masses.

However, many surveys show that significant concerns currently exist around cloud computing. As Figure

**FIG. 19: PUBLIC CLOUD CONCERNS**



Source: Gartner CIO survey

19 shows, security, privacy, maturity, and compliance are the top concerns. Many CIOs also worry about legacy compatibility: it is often not straightforward to move existing applications to the cloud.

- **Security and Privacy** CIOs must be able to report to their CEO and other executives how the company's data is being kept private and secure. Financially and strategically important data and processes often are protected by complex security requirements. Legacy systems have typically been highly customized to achieve these goals, and moving to a cloud architecture can be challenging. Furthermore, experience with the built-in, standardized security capabilities of cloud is still limited and many CIOs still feel more confident with legacy systems in this regard.
- **Maturity and Performance** – Cloud requires CIOs to trust others to provide reliable and highly available services. Unlike on-premises outages, cloud outages are often highly visible and may increase concerns
- **Compliance and Data Sovereignty** – Enterprises are subject to audits and oversight, both internal and external (e.g. IRS, SEC). Companies in many countries have data sovereignty requirements that severely restrict where they can host data services. CIOs ask: which clouds can comply with these systems and what needs to be done to make them compliant?

While many of these concerns can be addressed by cloud today, concerns remain and are prompting IT leaders to explore private clouds as a way of achieving the benefits of cloud while solving these problems. Next, we will explore this in more detail and also assess the potential tradeoffs.

### 3.3 Private Clouds

Microsoft distinguishes between public and private clouds based on whether the IT resources are shared between many distinct organizations (public cloud) or dedicated to a single organization (private cloud). This taxonomy is illustrated in Fig. 20. Compared to traditional virtualized datacenters, both private and public clouds benefit from automated management (to save on repetitive labor) and homogenous



hardware (for lower cost and increased flexibility). Due to the broadly-shared nature of public clouds, a key difference between private and public clouds is the scale and scope at which they can pool demand.

**FIG. 20: COMPARING VIRTUALIZATION, PRIVATE CLOUD, AND PUBLIC CLOUD**

	Operator	Automated Management	Homogenous Hardware	New App Model
Public Cloud	Department, Central IT, Third-party Provider	✓	✓	✓
Private Cloud	Department, Central IT, Third-party Provider	✓	✓	✓
Virtual Server	Department, Central IT, Third-party Provider	✓	✗	✗
Traditional Server	Department, Central IT, Third-party Provider	✗	✗	✗

Source: Microsoft. Shaded checks indicate an optional characteristic.

specific variability, but the size of the pool and the difficulty in moving loads from one virtual machine to another (exacerbated by the lack of homogeneity in hardware configurations) limits the ability to capture the full benefits. This is one of the reasons why even virtualized data centers still suffer from low utilization. There is no app model change so the complexity of building apps is not reduced.

- **Private clouds** move beyond virtualization. Resources are now pooled across the company rather than by organizational unit,<sup>15</sup> and workloads are moved seamlessly between physical servers to ensure optimal efficiency and availability. This further reduces the impact of random, time-of-day, and workload variability. In addition, new, cloud-optimized application models (Platform as a Service such as Azure) enable more efficient app development and lower ongoing operations costs.
- **Public clouds** have all the same architectural elements as private clouds, but bring massively higher scale to bear on all sources of variability. Public clouds are also the only way to diversify away industry-specific variability, the full geographic element of time-of-day variability, and bring multi-tenancy benefits into effect.

Private clouds can address some of the previously mentioned adoption concerns. By having dedicated hardware, they are easier to bring within the corporate firewall, which may ease concerns around **security and privacy**. Bringing a private cloud on-premise can make it easier to address some of the **regulatory, compliance and sovereignty** concerns that can arise with services that cross jurisdictional boundaries. In cases where these concerns weigh heavily in an IT leader's decision, an investment in a private cloud may be the best option.

Private clouds do not really differ from public cloud regarding other concerns, such as **maturity and performance**. Public and private cloud technologies are developing in tandem and will mature together.

<sup>15</sup> Aggregation across organizational units is enabled by two key technologies: live migration, which moves virtual machines while remaining operational, thereby enabling more dynamic optimization; and self-service provisioning and billing.



A variety of performance levels will be available in both public and private form, so there is little reason to expect that one will have an advantage over another.<sup>16</sup>

While private clouds can alleviate some of the concerns, in the next paragraph we will discuss whether they will offer the same kind of savings described earlier.

### 3.4 Cost Trade-Off

While it should be clear from the prior discussion that conceptually the public cloud has the greatest ability to capture diversification benefits, we need to get a better sense of the magnitude. Fig. 21 shows that while the public cloud addresses all sources of variability the private cloud can address only a subset.

**FIG. 21: DIVERSIFICATION BENEFITS**

	Sources of Variability			
	Random	Time of day	Industry	Multiple Resource
Private Cloud	✓	✓	✗	✓
Public Cloud	✓	✓	✓	✓

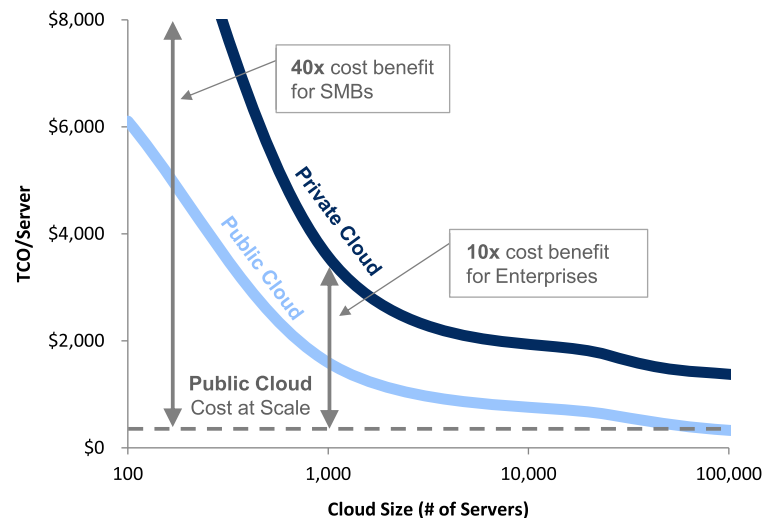
Source: Microsoft.

For example, industry variability cannot be addressed by a private cloud, while growth variability can be addressed only to a limited degree if an organization pools all its internal resources in a private cloud. We modeled all of these factors, and the output is shown in Fig. 22.

The lower curve shows the cost for a public cloud (same as the curve shown in Fig. 15). The upper curve shows the cost of a private cloud. The public cloud curve is lower at every scale due to the greater impact of demand aggregation and the multi-tenancy effect. Global scale public clouds are likely to become extremely large, at least 100,000 servers in size, or possibly much larger, whereas the size of an organization's private cloud will depend on its demand and budget for IT.

Fig. 22 also shows that for organizations with a very small installed base of servers (<100), private clouds are prohibitively expensive compared to public cloud. The only way for these small organizations or departments to share in the benefits of at-scale cloud computing is by moving to a

**FIG. 22: COST BENEFIT OF PUBLIC CLOUD**



Source: Microsoft.

<sup>16</sup> Private clouds do allow for a greater degree of customization than public clouds, which could enhance performance for a certain computational task. Customization requires R&D effort and expense, however, so it is difficult to make a direct price/performance comparison.

public cloud. **For large agencies with an installed base of approximately 1,000 servers, private clouds are feasible but come with a significant cost premium of about 10 times the cost of a public cloud for the same unit of service, due to the combined effect of scale, demand diversification and multi-tenancy.**

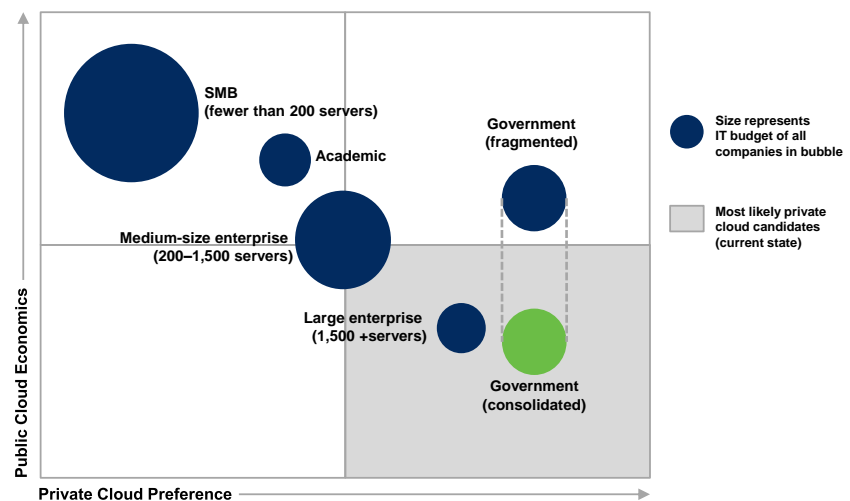
In addition to the increase in TCO, private clouds also require upfront investment to deploy – an investment that must accommodate peak demand requirements. This involves separate budgeting and commitment, increasing risk. Public clouds, on the other hand, can generally be provisioned entirely on a pay-as-you-go basis.

### 3.5 Finding Balance Today: Weighing the Benefits of Private Cloud against the Costs

We've mapped a view of how public and private clouds measure up in Figure 23. The vertical axis measures the public cloud cost advantage. From the prior analysis we know public cloud has inherent economic advantages that will partially depend on customer size, so the bubbles' vertical position is dependent on the size of the server installed base. The horizontal axis represents the organization's preference for private cloud. The size of the circles reflects the total server installed base of companies of each type. The bottom-right quadrant thus represents the most attractive areas for private clouds (relatively low cost premium, high preference).

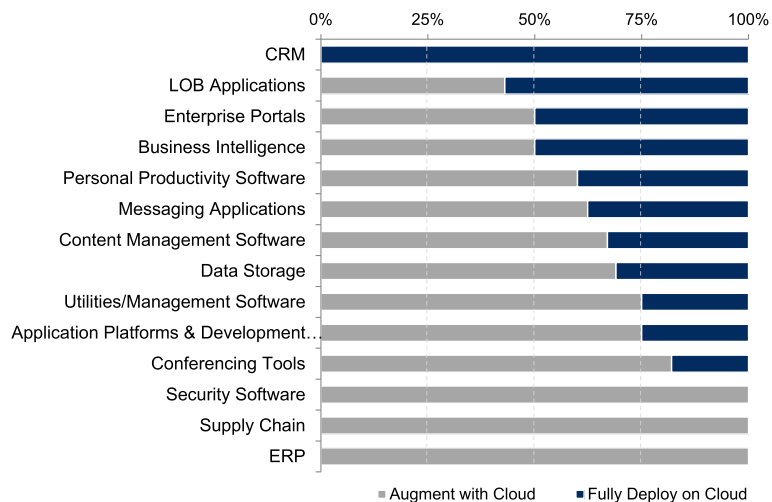
We acknowledge that Figure 23 provides a simplified view. IT is not monolithic within any of these industry segments. Each organization's IT operation is segmented into workload types, such as email or ERP. Each of these has a different level of sensitivity and scale, and CIO surveys reveal that preference for public

FIGURE 23: COST AND BENEFITS OF PRIVATE CLOUDS



Source: Microsoft

FIGURE 24: CLOUD-READY WORKLOADS (2010)



Source: Microsoft survey question "In the next 12-24 months, please indicate if a cloud offering would augment on-premise offering or completely replace it"

cloud solutions currently varies greatly across workloads (Figure 24).

An additional factor is that many app portfolios have been developed over the past 15-30 years and are tightly woven together. This particularly holds true for ERP and related custom applications at larger companies who have more sizable application portfolios. Apps that are more 'isolated' such as CRM, collaboration, or new custom apps may be more easily deployed in the cloud. Some of those apps may need to be integrated back to current on-premises apps.

Before we draw final conclusions, we need to make sure we avoid the “horseless carriage syndrome” and consider the likely shift along the two axes (economics and private preference).

### 3.6 The Long View: Cloud Transition Over Time

As we pointed out in the introduction of this paper, it is dangerous to make decisions during the early stages of a disruption without a clear vision of the end state. IT leaders need to design their architecture with a long term vision in mind. We therefore need to consider how the long term forces will impact the position of the bubbles on Fig. 23.

We expect two important shifts to take place. First, **the economic benefit of public cloud will grow over time**. As more and more work is done on public clouds, the economies of scale we described in Section 2 will kick

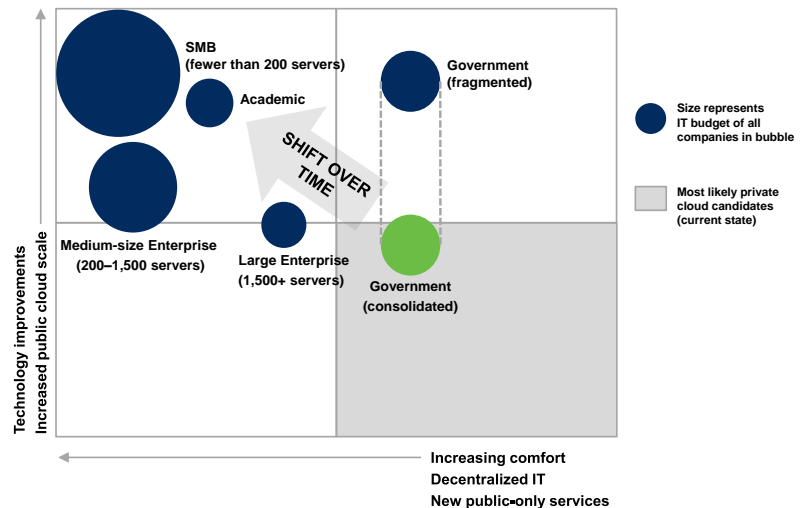
in, and the cost premium on private clouds will increase over time. Customers will increasingly be able to tap into the supply-side, demand-side and multi-tenancy savings as discussed previously. As shown in Fig. 27 this leads to an upward shift along the vertical axis.

At the same time, **some of the barriers to cloud adoption will begin to fall**. Many technology case studies show that, over time, concerns over issues like compatibility, security, reliability, and privacy will be addressed. This will likely also happen for the cloud, which would represent a shift to the left on Fig. 27. Below we will explore some of the factors that cause this latter shift.

#### Cloud security will evolve

Public clouds are in a relatively early stage of development, so naturally critical areas like reliability and security will continue to improve. Data already suggests that public cloud email is more reliable than most on-premises implementations. In PaaS, the automatic patching and updating of cloud systems greatly improves the security of all data and applications, as the majority of exploited vulnerabilities take advantage of systems that are out-of-date. Many security experts argue there are no fundamental

**FIG. 27: EXPECTED PREFERENCE SHIFT FOR PUBLIC AND PRIVATE CLOUD**



Source: Microsoft.

reasons why public clouds would be less secure; in fact, they are likely to become more secure than on premises due to the intense scrutiny providers must place on security and the deep level of expertise they are developing.

### ***Clouds will become more compliant***

Compliance requirements can come from within an organizational, industry, or government (e.g., European Data Protection Directive) and may currently be challenging to achieve with cloud without a robust development platform designed for enterprise needs. As cloud technologies improve, and as compliance requirements adapt to accommodate cloud architectures, cloud will continue to become more compliant, and therefore feasible for more organizations and workloads. This was the case, for example, with, e-signatures, which were not accepted for many contracts and documents in the early days of the Internet. As authentication and encryption technology improved and as compliance requirements changed, e-signatures became more acceptable. Today, most contracts (including those for opening bank accounts and taking out loans) can be signed with an e-signature.

The large group of customers who are rapidly increasing reliance on public clouds—small and medium businesses (SMBs) and consumers of Software as a Service (SaaS)—will be a formidable force of change in this area. This growing constituency will continue to ask governments to accommodate the shift to cloud by modernizing legislation. This regulatory evolution will make public cloud a more viable alternative for large enterprises and thus move segments along the horizontal axis toward public cloud preference.

### ***Decentralized IT (also known as ‘rogue IT’) will continue to lead the charge***

Many prior technology transitions were led not by CIOs but by departments, business decision makers, developers, and end users – often in spite of the objections of CIOs. For example, both PCs and servers were initially adopted by end users and departments before they were officially embraced by corporate IT policies. More recently, we saw this with the adoption of mobile phones, where consumer adoption is driving IT to support these devices. We’re seeing a similar pattern in the cloud: developers and departments have started using cloud services, often without the knowledge of the IT group (hence the name “rogue clouds”). Many business users will not wait for their IT group to provide them with a private cloud; for these users, productivity and convenience often trump policy.

It is not just impatience that drives “rogue clouds”; ever-increasing budgetary constraints can lead users and even departments to adopt cheaper public cloud solutions that would not be affordable from traditional channels. For example, when Derek Gottfrid wanted to process all 4TB of the *New York Times* archives and host them online, he went to the cloud without the knowledge of the Times’ IT department.<sup>17</sup> Similarly, the unprecedented pricing transparency that the public cloud offers will put further pressure from the CEO and CFO on CIOs to move to the public cloud.

CIOs should acknowledge that these behaviors are commonplace early in a disruption and either rapidly develop and implement a private cloud with the same capabilities or adopt policies which incorporate some of this behavior, where appropriate, in IT standards.

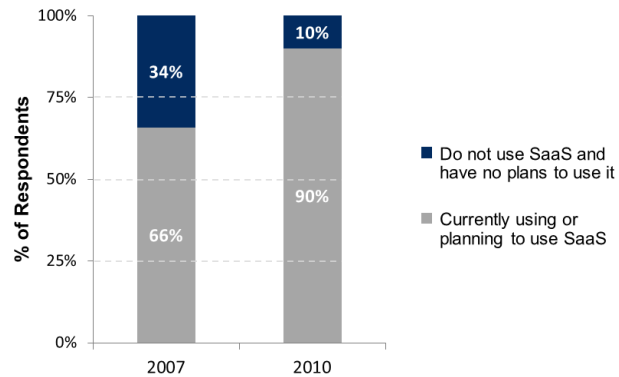
<sup>17</sup> <http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/>

### Perceptions are rapidly changing

Strength in SaaS adoption in large enterprises serves as proof of changing perceptions (Fig. 28) and indicates that even large, demanding enterprises are moving to the left on the horizontal axis (i.e., reduced private preference). Just a few years ago, very few large companies were willing to shift their email, with all the confidential data that it contains, to a cloud model. Yet this is exactly what is happening today.

As positive use cases continue to spur more interest in cloud technology, this virtuous cycle will accelerate, driving greater interest in and acceptance of cloud.

**FIG. 28: INCREASING ADOPTION OF SOFTWARE AS A SERVICE (SAAS)**



Source: Gartner.

In summary, while there are real hurdles to cloud adoption today, these will likely diminish over time. While new, unforeseen hurdles to public cloud adoption may appear, the public cloud economic advantage will grow stronger with time as cloud providers unlock the benefits of economics we discussed in Section 2. While the desire for a private cloud is mostly driven by security and compliance concerns around *existing* workloads, the cost effectiveness and agility of the public cloud will enable *new* workloads.

Revisiting our “horseless carriage” analogy, we see that cars became a huge success not simply because they were faster and better (and eventually more affordable) than horse-drawn carriages. The entire transportation ecosystem had to change. Highway systems, driver training programs, accurate maps and signage, targeted safety regulation, and a worldwide network of fueling infrastructure all had to be developed to enable this transition. Each successive development improved the value proposition of the car. In the end, even people’s living habits changed around the automobile, resulting in the explosion of the suburbs in the middle part of the 20<sup>th</sup> century. This created “net new” demand for cars by giving rise to the commuting professional class. This behavioral change represented a massive positive feedback loop that inexorably made the automobile an essential, irreplaceable component of modern life.

Similarly, we believe cloud will be enabled and driven not just by economics and qualitative developments in technology and perception, but by a series of shifts from IT professionals, regulators, telecom operators, ISVs, systems integrators, and cloud platform providers. As cloud is embraced more thoroughly, its value will increase.

## 4. THE JOURNEY TO THE CLOUD

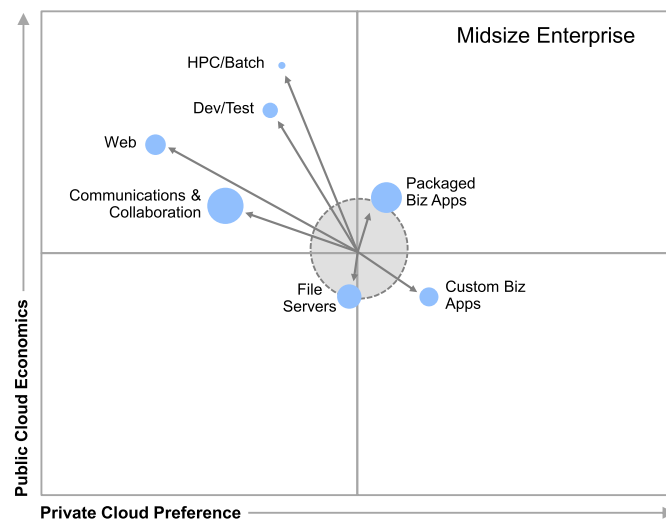
Because we are in the early days of the cloud paradigm shift, there is much confusion about the direction of this ongoing transformation. In this paper, we looked beyond the current technology and focused on the underlying economics of cloud to define the destination – where all of this disruption and innovation is leading our industry. **Based on our analysis, we see a long-term shift to cloud driven by three important economies of scale:** (1) larger datacenters can deploy computational resources at

significantly lower cost than smaller ones; (2) demand pooling improves the utilization of these resources, especially in public clouds; and (3) multi-tenancy lowers application maintenance labor costs for large public clouds. Finally, the cloud offers unparalleled levels of elasticity and agility that will enable exciting new solutions and applications.

For businesses of all sizes, the cloud represents tremendous opportunity. It represents an opportunity to break out of the longstanding tradition of IT professionals spending 80 percent of their time and budget “keeping the lights on,” with few resources left to focus on innovation. **Cloud services will enable IT groups to focus more on innovation while leaving non-differentiating activities to reliable and cost-effective providers.** Cloud services will enable IT leaders to offer new solutions that were previously seen as either cost prohibitive or too difficult to implement. This is especially true of cloud platforms (Platform as a Service), which significantly reduce the time and complexity of building new apps that take advantage of all the benefits of the cloud.

This future won’t materialize overnight. IT leaders need to develop a new 5- to 10-year vision of the future, recognizing that they and their organizations will play a fundamentally new role in their company. They need to plot a path that connects where they are today to that future. An important first step in this is to segment their portfolio of existing applications (Fig. 29). For some apps the economic and agility benefits may be very strong so they should be migrated quickly. However, barriers do exist today, and while we outlined in section 3 that many of them will be overcome over time, cloud may not be ready for some apps today. For tightly integrated apps with fairly stable usage patterns, it may not make sense to move them at all, similar to how some mainframe apps were never migrated to client/server. While new custom apps don’t have the legacy problem, designing them in a scalable, robust fashion is not always an easy task. Cloud optimized platforms (Platform as a Service) can dramatically simplify this task.

**FIG. 29: SEGMENTING IT PORTFOLIO**



Source: Microsoft.

This transition is a delicate balancing act. If the IT organization moves too quickly in areas where the cloud is not ready, it can compromise business continuity, security, and compliance. If it moves too slowly, it can put the company at a significant competitive disadvantage versus competitors who do take full advantage of cloud capabilities, giving up a cost, agility, or value advantage. Moving too slowly also increases the risk that different groups or individuals within the company will each adopt their own cloud solution in a fragmented and uncontrolled fashion (“rogue IT”), wresting control over IT from the CIO. IT leaders who stay ahead of the cloud trend will be able to control and shape this transition; those who lag behind will increasingly lose control.

To lead the transition, IT leaders need to think about the long term architecture of their IT. Some see a new role emerging, that of a Cloud Services Architect, who determines which applications and services move to the cloud and exactly when such a move takes place based on a business case and a detailed understanding of the cloud capabilities available. This should start by taking inventory of the



organization's resources and policies. This includes an application and data classification exercise to determine which policy or performance requirements (such as confidential or top secret data retention requirements) apply to which applications and data. Based on this, IT leaders can determine what parts of their IT operation are suitable for public cloud and what might justify an investment in private cloud. Beginning in this manner takes advantage of the opportunity of cloud while striking balance between economics and security, performance, and risk.

To accomplish this, IT leaders need a partner who is firmly committed to the long-term vision of the cloud and its opportunities, one who is not hanging on to legacy IT architectures. At the same time, this partner needs to be firmly rooted in the realities of today's IT so it understands current challenges and how to best navigate the journey to the cloud. IT leaders need a partner who is neither incentivized to push for change faster than is responsible nor to keep IT the same. Customers need a partner who has done the hard work of figuring out how best to marry legacy IT with the cloud, rather than placing that burden on the customer by ignoring the complexities of this transformation.

At Microsoft, we are "all in" on the cloud. We provide both commercial SaaS (Office 365) and a cloud computing platform (Windows Azure Platform). Office 365 features the applications customers are familiar with like Exchange email and SharePoint collaboration, delivered through Microsoft's cloud. Windows Azure is our cloud computing platform, which enables customers to build their own applications and IT operations in a secure, scalable way in the cloud. Writing scalable and robust cloud applications is no easy feat, so we built Windows Azure to harness Microsoft's expertise in building our cloud-optimized applications like Office 365, Bing, and Windows Live Hotmail. Rather than just moving virtual machines to the cloud, we build a Platform as a Service that reduces complexity for developers and IT administrators.

Microsoft also brings to the cloud the richest partner community in the world. We have over 600,000 partners in more than 200 countries servicing millions of businesses. We are already collaborating with thousands of our partners on the cloud transition. Together we are building the most secure, reliable, scalable, available, cloud in the world.

Over the last three decades, Microsoft has developed strong relationships with IT organizations, their partners, and their advisors. This offers us an unparalleled understanding of the challenges faced by today's IT organizations. Microsoft is both committed to the cloud vision *and* has the experience to help IT leaders on the journey.

Microsoft has a long history of bringing to life powerful visions of the future. Bill Gates founded Microsoft on the vision of putting a PC in every home and on every desktop in an era when only the largest corporations could afford computers. In the journey that followed, Microsoft and our partners helped bring PCs to over one billion homes and desktops. Millions of developers and businesses make their living on PCs and we are fortunate to play a role in that.

Now, we have a vision of bringing the power of cloud computing to every home, every office, and every mobile device. The powerful economics of cloud drive all of us towards this vision. Join Microsoft and our partners on the journey to bring this vision to life.