Cloud Computing: a short introduction

3 Cloud Service Models

Cloud Software as a Service (SaaS)

Use the applications the provider supplies over a network

Cloud Platform as a Service (PaaS)

Deploy customer-created applications to a cloud

Cloud Infrastructure as a Service (IaaS)

Rent processing, storage, network capacity, and other fundamental computing resources

To be considered “cloud” they must be deployed on top of a cloud infrastructure with the key characteristics in the previous slide

SaaS – Software As A Service

- Essentially based on the concept of renting application functionality from a service provider rather than buying, installing and running software yourself.

- Offerings within this range from services such as Salesforce.com at one end, delivering the equivalent of a complete application suite, to players like MessageLabs, GoogleMail, GoogleDoc at the other, whose services are designed to complement an operational infrastructure.

Three Features of Mature SaaS Applications

- Scalable
  - Handle growing amounts of work in a graceful manner
- Multi-tenancy
  - One application instance may be serving hundreds of companies
  - Opposite of multi-instance where each customer is provisioned their own server running one instance
- Metadata driven configurability
  - Instead of customizing the application for a customer (requiring code changes), one allows the user to configure the application through metadata

Multi - Tenant
Mature SaaS Applications

- Computing resources and application code are generally shared between all the tenants on a server,
- Each tenant has its own data that remains logically isolated from those of other tenants.
- Metadata associates each database with the correct tenant,
- Database security should any tenant from accidentally or maliciously accessing other tenants' data.
- This approach tends to lead to higher costs for maintaining equipment and backing up tenant data. The number of tenants that can be housed on a server is limited by the number of databases that it can support.
- Banking or medical records management often have very strong data isolation requirements, and may not even consider an application that does not supply each tenant with its own individual database.

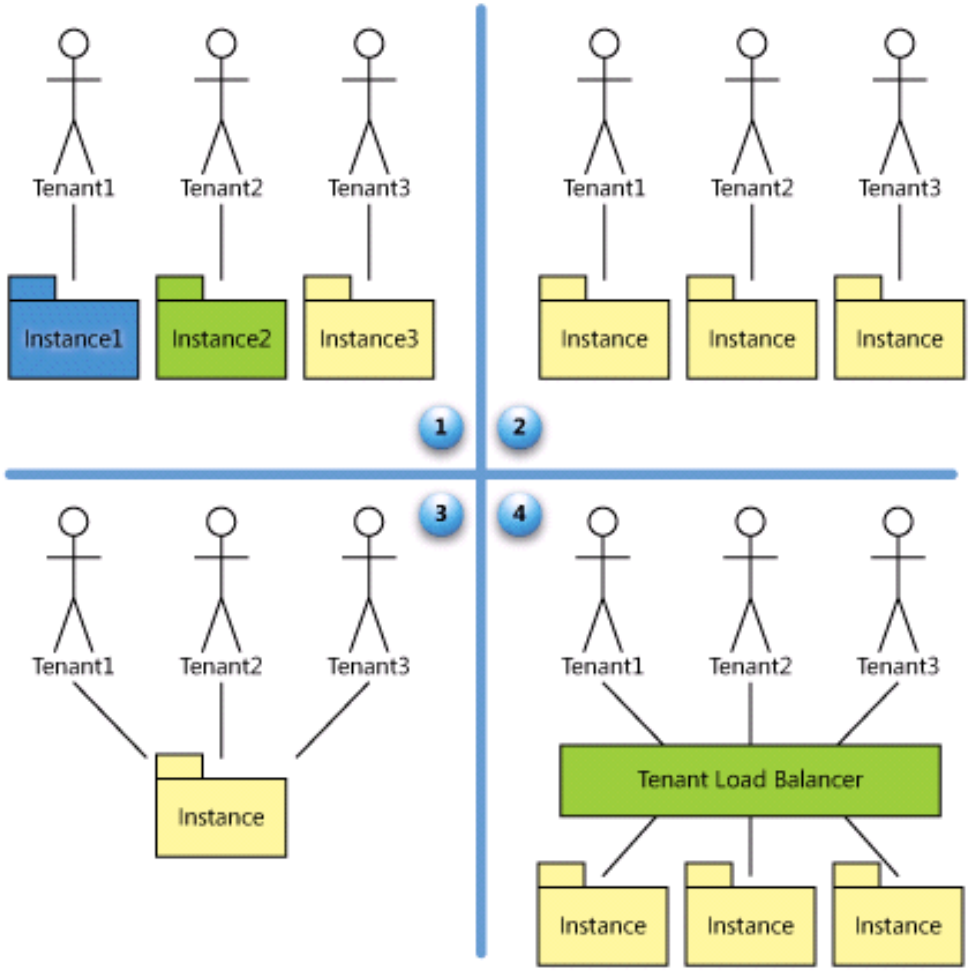
SaaS Maturity Levels

Level 1: Ad-Hoc/Custom

Level 2: Configurable

Level 3: Configurable, Multi-Tenant-Efficient

Level 4: Scalable, Configurable, Multi-Tenant-Efficient



Source: Microsoft MSDN Architecture Center

PaaS – Platform As A Service

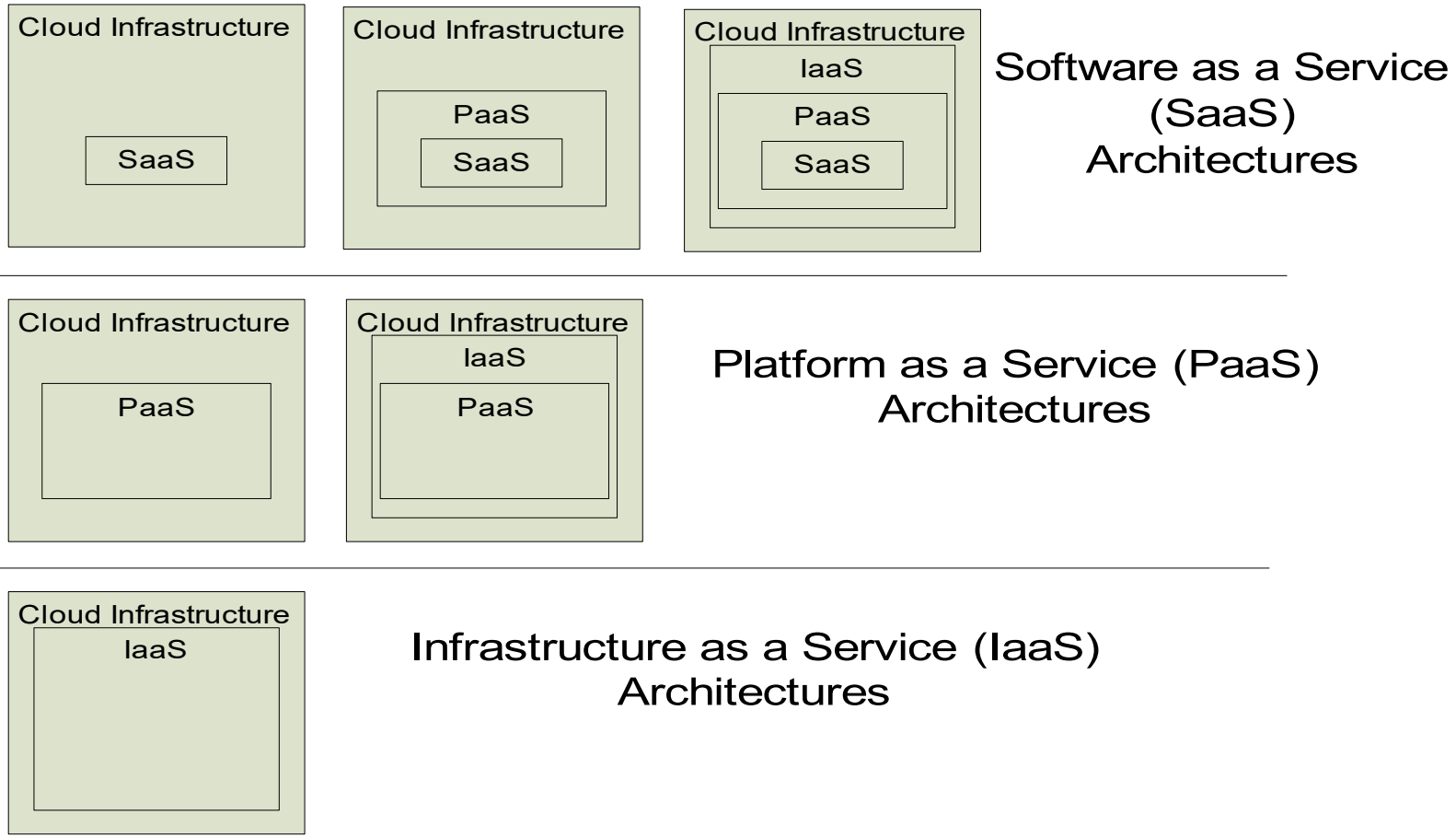
- Platform as a service (PaaS), which is all about providing, a platform in the cloud, upon which applications can be developed and executed.
- Players like Google, again Salesforce.com (this time with Force.com), and Microsoft (with Azure) exist in this space.
- Facilities provided include things like database management, security, workflow management, application serving, and so on.
- PaaS is a set of services that helps the development and the testing of apps without worrying about the underlying infrastructure. Developers don't need to worry about provisioning the servers, storage and backup associated with developing and launching an app. They write code, test the app, launch the app, and make changes to fix bugs. All the back-end stuff about setting up servers occurs automatically and transparently in the background

IaaS – Infrastructure As A Service

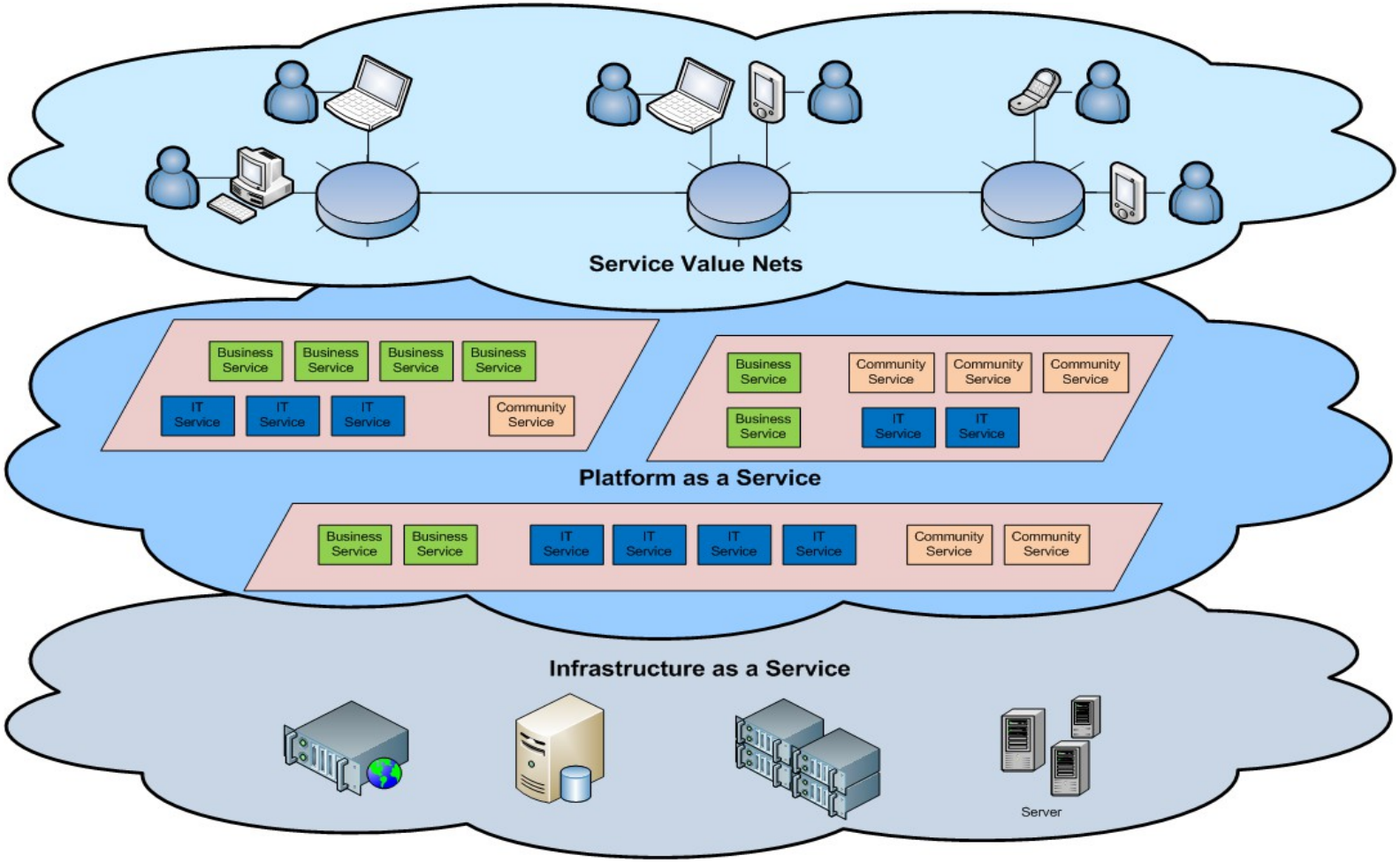
The proposition here is the offering of compute power and storage space on demand.

- The difference between this and the other two categories of cloud is that the software that executes is essentially yours.
- In practical terms, the model is based on the same principles of virtualisation in the context of server partitioning or flexible storage. Rather than running a virtual image on a partition on a physical server in your data centre, you spin it up on a virtual machine that you have created in the cloud. Virtual disks can be created in a similar manner, to deal with the storage side of things.
- You can also run/map your software onto a network of virtual machines

Service Model Architectures



Cloud Architecture



Cloud Provider

- ***Cloud Provider***: Person, organization or higher-level system responsible for making a service available to *Cloud Consumers*
- The providers perform different tasks for different service types.

Provider Type	Major Activities
SaaS	Installs, manages, maintains and supports the software
PaaS	Manages cloud infrastructure and other middleware the for the platform
IaaS	Maintains the storage, networking and the hosting environment for virtual machines

- The operations of service providers are discussed in further details from the following perspectives: *Service Deployment*, Service Orchestration, *Business Support* and *Operational Support*.

Service Level Agreements (SLAs)

- Contract between customers and service providers of the level of service to be provided
- Contains performance metrics (e.g., uptime, throughput, response time)
- Problem management details
- Documented security capabilities
- Contains penalties for non-performance

Consumer and Provider Roles

The differences in scope and control between the cloud consumers and cloud providers, for each of the service models:

- SaaS: The cloud consumer does not manage or control the underlying cloud infrastructure or individual applications, except for preference selections and limited administrative application settings. Security provisions are carried out mainly by the cloud provider.
- PaaS: The cloud consumer has control over applications and application environment settings of the platform. Security provisions are split between the cloud provider and the cloud consumer.
- IaaS: The cloud consumer generally has broad freedom to choose the operating system and development environment to be hosted. Security provisions beyond the basic infrastructure are carried out mainly by the cloud consumer.

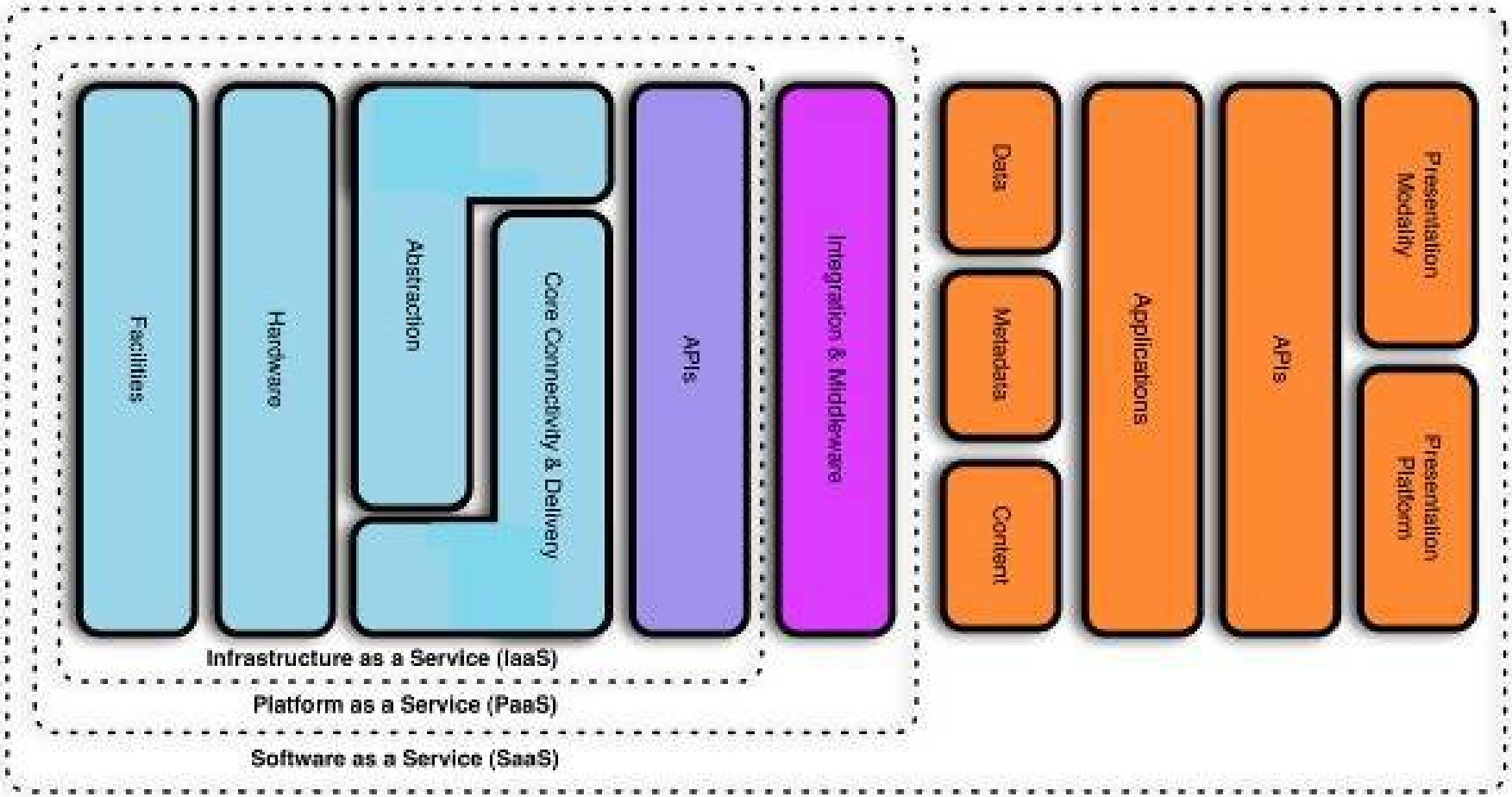
The resulting levels of a cloud system

The five conceptual layers of a generalized cloud environment:

- ***Facility Layer:*** Heating, ventilation, air conditioning (HVAC), power, communications, and other aspects of the physical plant comprise the lowest layer, the facility layer.
- ***Hardware Layer:*** Includes computers (CPU, memory), network (router, firewall, switch, network link and interface) and storage components (hard disk), and other physical computing infrastructure elements.
- ***Virtualized Infrastructure Layer:*** Entails software elements, such as hypervisors, virtual machines, virtual data storage, and supporting middleware components used to realize the infrastructure upon which a computing platform can be established. While virtual machine technology is commonly used at this layer, other means of providing the necessary software abstractions are not precluded.
- ***Platform Architecture Layer:*** Entails compilers, libraries, utilities, and other software tools and development environments needed to implement applications.
- ***Application Layer***: Represents deployed software applications targeted towards end-user software clients or other programs, and made available via the cloud.

Architectura levels

CSA Cloud Reference Model



4 Cloud Deployment Models

Private cloud

enterprise owned or leased

Community cloud

shared infrastructure for specific community

Public cloud

Sold to the public, mega-scale infrastructure

Hybrid cloud

composition of two or more clouds

Private Cloud

- Intentionally limits access to its resources to service consumers that belong to the same organization that owns the cloud
- to maintain a consistent level of control over security, privacy, and governancethe infrastructure is managed and operated for one organization only
- also know as internal cloud or on-premise cloud

Essential characteristics of a private cloud typically include:

- heterogeneous infrastructure
- customized and tailored policies
- dedicated resources
- in-house infrastructure (capital expenditure cost model)
- end-to-end control

Pubblic Cloud

Also known as external cloud or multi-tenant cloud

It generally provides an IT infrastructure in a third-party physical data center that can be utilized to deliver services without having to be concerned with the underlying technical complexities.

Essential characteristics of a public cloud typically include:

- homogeneous infrastructure
- common policies
- shared resources and multi-tenant
- leased or rented infrastructure; operational expenditure cost model
- economies of scale
- can host individual services or collections of services, allow for the deployment of service compositions and even entire service inventories.

Community Cloud and others

This deployment model typically refers to special-purpose cloud computing environments shared and managed by a number of related organizations participating in a common domain or vertical market.

Other Deployment Models (Variations of the models)

Hybrid cloud = a model merging private and public cloud environments

Dedicated cloud (also known as the hosted cloud or virtual private cloud) = cloud computing environments hosted and managed off-premise or in public cloud environments, but where dedicated resources are provisioned solely for an organization's private use.

The Intercloud (Cloud of Clouds)

- The intercloud is not as much a deployment model as it is a concept based on the aggregation of deployed clouds. Just like the Internet, which is a network of networks; intercloud refers to an inter-connected global cloud of clouds.
- As the World Wide Web, intercloud represents a massive collection of services that organizations can explore and consume.
- From a service consumer's perspective, an intercloud is an on-demand SOA environment where useful services managed by other organizations can become a part of the aggregate portfolio of services of those same organizations.

Public Statistics on Cloud Economics

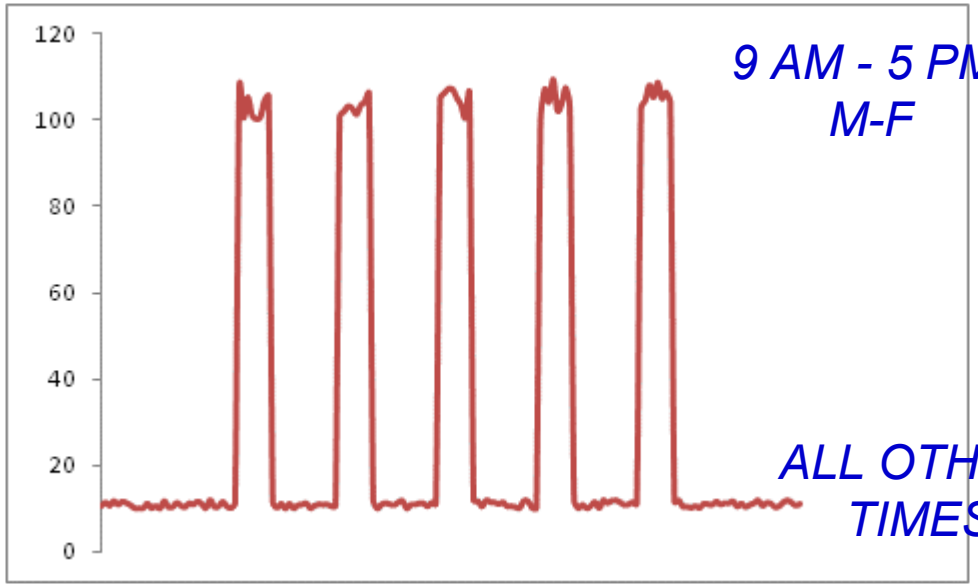
Cost of Traditional Data Centers

- 11.8 million servers in data centers
- Servers are used at only 15% of their capacity
- 800 billion dollars spent yearly on purchasing and maintaining enterprise software
- 80% of enterprise software expenditure is on installation and maintenance of software
- Data centers typically consume up to 100 times more per square foot than a typical office building
- Average power consumption per server quadrupled from 2001 to 2006.
- Number of servers doubled from 2001 to 2006

Suppose you are Forbes.com

- You offer on-line real time stock market data
- Why pay for capacity weekends, overnight?

Rate of Server Accesses



Forbes' Solution

Host the web site in Amazon's EC2 *Elastic Compute Cloud*

Provision new servers every day, and deprovision them every night

Pay just $0.10* per server per hour

* more for higher capacity servers

Let Amazon worry about the hardware!

Why cloud computing? Savings

Set up cost

	On-premises				Cloud Based			
	Unit	Quantity	per Unit	Total	Unit	Quantity	per Unit	Total
CAPEX (initial)				**$4,128,133**				**$66,665**
Server	Nos	100	$5,000	$500,000	Nos	0	$0	$0
Networking Equipments	Nos	50	$1,000	$50,000	Nos	0	$0	$0
Storage	TB	50	$3,500	$175,000	TB	0	$0	$0
Storage (Backup)	TB	350	$1,000	$350,000	TB	0	$0	$0
Software (OS + IIS)	Nos	100	$3,999	$399,900	Nos	0	$0	$0
Software (DB)	Nos	100	$24,999	$2,499,900	Nos	0	$0	$0
Software (AV+ Mgmt)	Nos	100	$200	$20,000	Nos	0	$0	$0
Labor	$/resource/pm	10	$13,333	$133,333	$/resource/pm	5	$13,333	$66,665
Real Estate	$/sft	0	$1,000	$0	$/sft	0	$0	$0
OPEX (annual)				**$1,985,717**				**$1,079,963**
Computing Power		0	$0	$0	$/hr	3504000	$0.125	$438,000
Storage	TB	0	$3,500	$0	$/pGB	1533000	$0.15	$229,950
Bandwidth	$/pa	3	$30,000	$90,000	$/pGB	600000	$0.22	$132,000
Staff Salary	resources/year	8	$160,000	$1,280,000	resources/year	3	$160,000	$480,000
Infrastructure Maintenance	% of total cost	15	$7,250	$108,750	% of total cost	0	$0	$0
Software Maintenance	% of total cost	33	$3,999	$131,967	% of total cost	0	$0	
Electricity	$/pa/dc	1	$120,000	$120,000	$/pm	0	$0	$0
Rent for Real Estate	$/sft/pm	1000	$135	$135,000	$/sft/pm	0	$0	$0
Other Maintenance	$/year	3	$40,000	$120,000	$/year	0	$0	$0
Pay-per-Use Savings (-)		0	$0	$0	%	25	$8,000	$199,988
Total				**6,113,850**				**1,079,963**
Savings				**$5,033,888**				

Annual cost

A simpler analysis

TABLE 1-3. Comparing the cost of different IT infrastructures

	Internal IT	Managed services	The cloud
Capital investment	$40,000	$0	$0
Setup costs	$10,000	$5,000	$1,000
Monthly service fees	$0	$4,000	$2,400
Monthly staff costs	$3,200	$0	$1,000
Net cost over three years	$149,000	$129,000	$106,000

A simpler analysis: The 10 Laws of Cloudonomics

Public utility cloud services differ from traditional data center and private enterprise clouds for 3 reasons

a) they provide true on-demand services, by multiplexing demand from enterprises into a common pool of dynamically allocated resources

b) large cloud providers operate at a scale much greater than even the largest private enterprises

c) enterprise data centers are naturally driven to reduce cost via consolidation and concentration, clouds — whether content, application or infrastructure — benefit from dispersion.

Rule 1

Utility services cost less even though they cost more.

An on-demand service provider typically charges a utility premium
= a higher cost per unit time for a resource than if it were owned, financed or leased.

However, although utilities cost more *when they are used*, they cost nothing *when they are not used*.

Consequently, customers save money by replacing fixed infrastructure with clouds when workloads are spiky, specifically when the peak-to-average ratio is greater than the utility premium.

Rule 2

On-demand trumps forecasting

The ability to rapidly provision capacity means that any unexpected demand can be serviced, and the revenue associated with it captured. The ability to rapidly de-provision capacity means that companies do not need to pay good money for non-productive assets.

Forecasting is often wrong, especially for black swans *, so the ability to react instantaneously means higher revenues, and lower costs.

*The black swan: The impact of the highly improbable, N.N. Taleb and D Sornette, Dragon-Kings, Black Swans and the Prediction of Crises, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1470006

Rule 3

The peak of the sum is never greater than the sum of the peaks.

Enterprises deploy capacity to handle their peak demands – a tax firm worries about April 15th, a retailer about Black Friday, an online sports broadcaster about Super Sunday.

The total capacity of a cloud depends upon the *peak of the sum*

↔

The reallocation of resources across many enterprises with distinct peak periods implies that a cloud needs to deploy less capacity.

This also holds for a large private cloud, e.g. a governmental cloud

Rule 4

Aggregate demand is smoother than individual.

Aggregating demand from multiple customers tends to smooth out variation. Specifically, the "coefficient of variation" (the ratio of the standard deviation to the mean) of a sum of random variables is always less than or equal to that of any of the individual variables.

coefficient of variation= statistical measure of the dispersion of data points in a data series around the mean

Therefore, clouds get higher utilization, enabling better economics

Rule 5

Average unit costs are reduced by distributing fixed costs over more units of output.

While large enterprises benefit from economies of scale, larger cloud service providers can benefit from even greater economies of scale, such as volume purchasing, network bandwidth, operations, administration and maintenance tooling.

Rule 6

Superiority in numbers is the most important factor in the result of a combat (Clausewitz).

The classic military strategist Carl von Clausewitz argued that, above all, numerical superiority was key to winning battles. In the cloud theater, battles are waged between botnets and DDoS defenses. A botnet of 100,000 servers, each with a megabit per second of uplink bandwidth, can launch 100 gigabits per second of attack bandwidth. An enterprise IT shop would be overwhelmed by such an attack, whereas a large cloud service provider that is also an integrated network service provider has the scale to repel it.

Rule 7

Space-time is a continuum (Einstein/Minkowski)

A real-time enterprise derives competitive advantage from responding to changing business conditions and opportunities faster than the competition. Decision-making depends on computing, e.g., business intelligence, risk analysis, portfolio optimization and so forth. Assuming that the compute job is elastic such computing tasks can often trade off space and time, for example a batch job may run on one server for a thousand hours, or a thousand servers for one hour.

Since an ideal cloud provides unbounded on-demand scalability, for the same cost, a business can accelerate its decision-making.

Rule 8

Dispersion is the inverse square of latency = *Worst-Case and Expected Latency* is inversely proportional to the square root of the number of nodes (data centers of the provider) in an area.

Reduced latency - the request-response delay - is increasingly essential to delivering a range of services, among them online gaming, remote virtualized desktops, and interactive collaboration. However, to cut latency in half requires not twice as many nodes, but four times. For example, growing from one service node to dozens can cut global latency (e.g., New York to Hong Kong) from 150 milliseconds to below 20. However, shaving the next 15 milliseconds requires a thousand more nodes. There is thus a natural sweet spot for dispersion aimed at latency reduction, that of a few dozen nodes — more than an enterprise would want to deploy, especially because of lower utilization = a provider can deploy a larger number of data centers that a single corporation

In short ..

The Law of Cloud Response Time: The response time for an interactive transaction served by a distributed, elastic cloud is

$$T = F + \frac{N}{\sqrt{n}} + \frac{P}{p}$$

Where

- F = fixed cost to set up a transaction
- n = number of locations that can be used
- N = worst case latency
- P = computation overhead
- p = actual degree of parallelism (p≤n)

$$n = \sqrt[3]{\left(\frac{QN}{2P}\right)}^{2}$$

Optimal number of locations with Q/n processor at each location

In short ..

$$T = F + \frac{N}{\sqrt{n}} + \frac{P}{p}$$

Simply put, the total response time T for an interactive networked application in which a client application requests and receives an interactive response from a cloud application over a network is a function of three components. F is a fixed interval that can't be accelerated, N is the worst case round-trip latency for an environment with a single node, n is the number of (evenly-distributed) processing nodes, P is the time for the parallelizable portion of the application to run on one processor, and p is the number of processors.

For example, consider a search query requested via an end-user client. A time F is needed for client processing and other serial tasks; sharding and parallelization can reduce processing time via the P/p component; and replicating this service in n multiple physical locations can lower the network latency, reducing the $N/\sqrt{n}$ component. In fact, if Q processors are available for deployment, the optimum latency is reached when the number of nodes is $n = \sqrt[3]{\left(\frac{QN}{2P}\right)^2}$.

Rule 9

Never put all your eggs in one basket.

The reliability of a system with n redundant components, each with reliability r, is $1-(1-r)^n$. If the reliability of a single data center is 99 %, two centers provide four nines (99.99 %) and three centers provide six nines (99.9999 percent).

No finite number of data centers can reach 100 % reliability, but a few data centers result in an extremely high reliability architecture

To provide high availability services globally for latency-sensitive applications, there must be a few data centers in each region.

Rule 10

An object at rest tends to stay at rest (Newton).

A data center is a very, very large object. While theoretically, any company can site data centers in globally optimal locations that are located on a core network backbone with cheap access to power, cooling and acreage, few do. Instead, they remain in locations for reasons such as where the company or an acquired unit was founded, or where they got a good deal on distressed but conditioned space. A cloud service provider can locate greenfield sites optimally.

Energy Conservation and Data Centers

Standard 9000 square foot costs $21.3 million to build with $1 million in electricity costs/year

Data centers consume 1.5% of our Nation's electricity (EPA)

.6% worldwide in 2000 and 1% in 2005

Green technologies can reduce energy costs by 50%

IT produces 2% of global carbon dioxide emissions

Cloud computing = outsourcing to the next step

- **You don't have to own the hardware**
- You "rent" it as needed from a cloud
- There are public cloud e.g. Amazon EC2, and now many others (Microsoft, IBM, Sun, and others ...)
- A company can create a private one with more control over security, etc.

Goal 1 – Cost Control

- Cost
  - Many systems have variable demands
    - Batch processing (e.g. New York Times)
    - Web sites with peaks (e.g. Forbes)
    - Startups with unknown demand (e.g. the *Cash for Clunkers* program)
  - Reduce risk in acquisition
    - Don't need to buy hardware until you need it

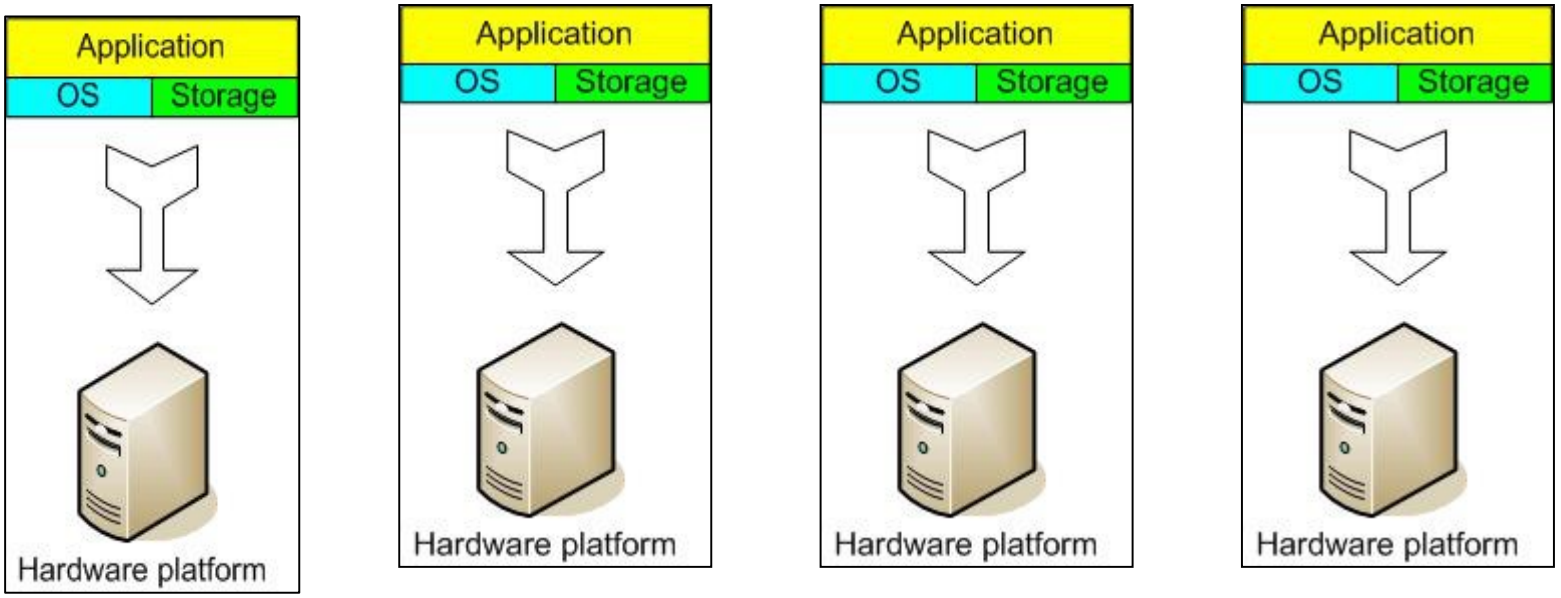
Goal 2 - Business Agility

- More than scalability - ***elasticity***!
  - Ely Lilly in rapidly changing health care business
    - Used to take 3 - 4 months to give a department a server cluster, then they would hoard it!
  - Using EC2, about 5 minutes!
    - And they give it back when they are done!
- Scaling back is as important as scaling up

Set up a server in the cloud

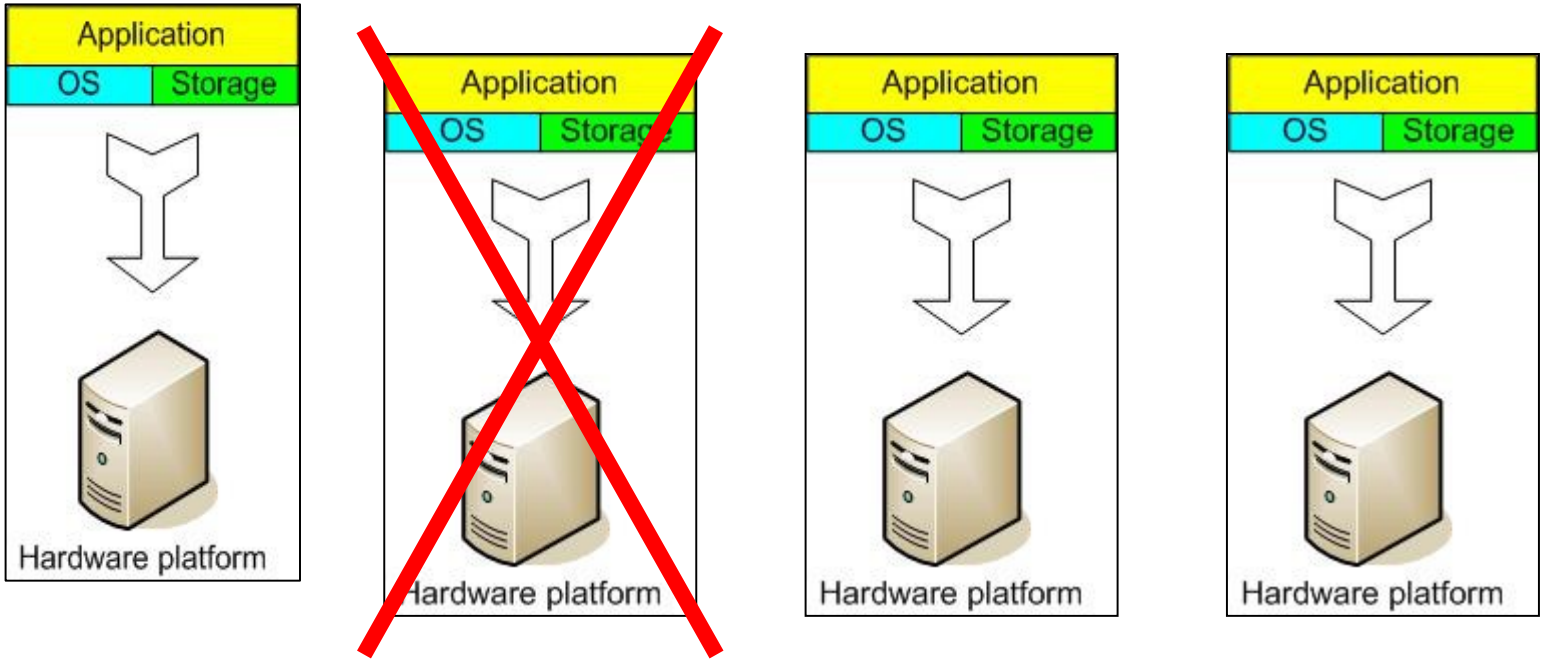
- Various providers let you create virtual servers
  - Set up an account, perhaps just with a credit card
- You create a network of virtual servers ("virtualization")
  - Choose OS software of each server
  - It will run on a large server farm located somewhere
  - You can instantiate more on a few minutes' notice
  - You can shut down instances in a minute or so
- They send you a bill for what you use
- The differences are strongly related to the savings

The Traditional Server Concept



Web Server	App Server	DB Server	EMail
Windows	Linux	Linux	Windows
IIS	Glassfish	MySQL	Exchange

And if something goes wrong ...



Web Server	App Server	DB Server	EMail
Windows	DOWN!	Linux	Windows
IIS		MySQL	Exchange

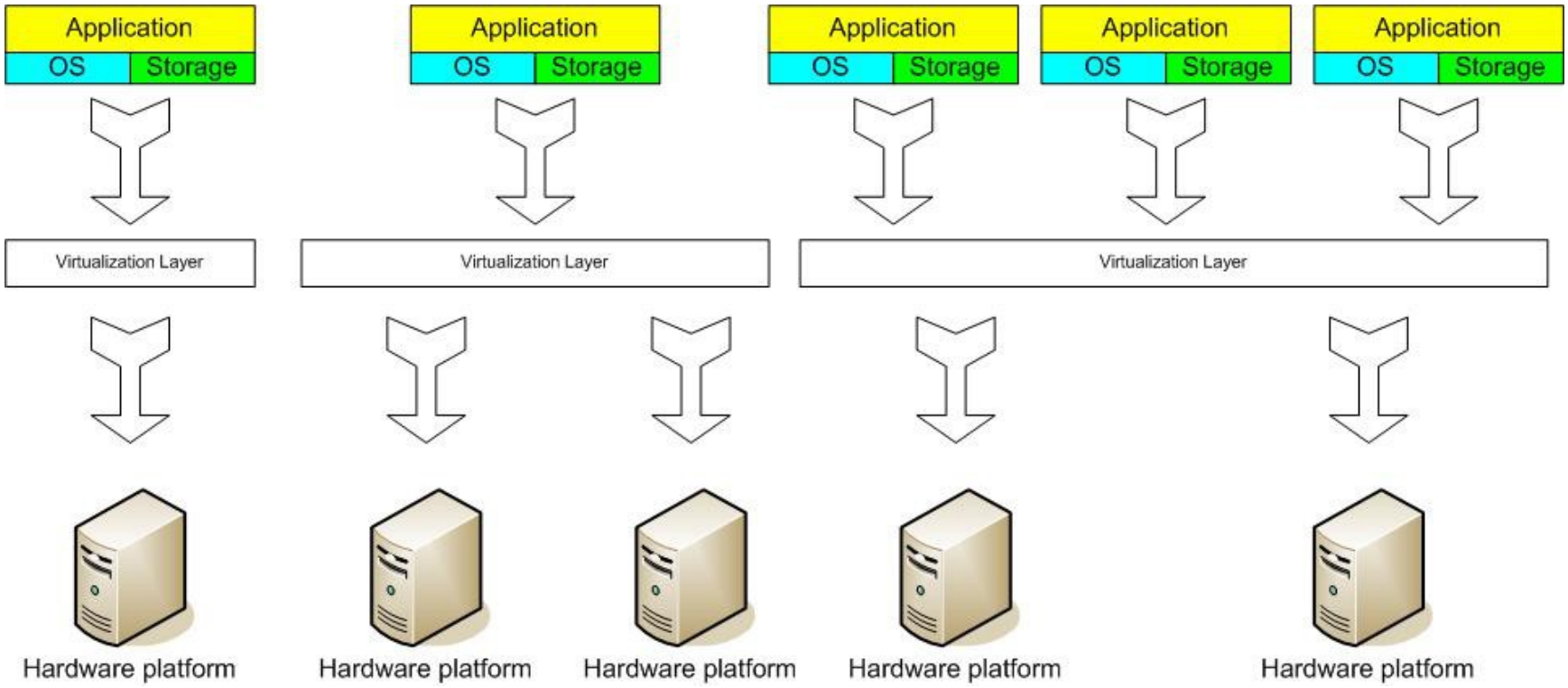
The Traditional Server Concept

- Traditional server = a whole unit that includes the hardware, the OS, the storage, and the applications.
- Servers are often referred to by their function i.e. the Exchange server, the SQL server, the File server, etc.
- If the File server fills up, or the Exchange server becomes overloaded, the System Administrators adds in a new server.
- Unless there are multiple servers, if a service experiences a hardware failure, then the service is down.
- System Admins can implement clusters of servers to make them more fault tolerant. However, even clusters have limits on their scalability, and not all applications work in a clustered environment.

The Traditional Server Concept

- Pros
  - Easy to conceptualize
  - Fairly easy to deploy
  - Easy to backup
  - Virtually any application/service can be run from this type of setup
- Cons
  - Expensive to acquire and maintain hardware
  - Not very scalable
  - Difficult to replicate
  - Redundancy is difficult to implement
  - Vulnerable to hardware outages
  - In many cases, processor is under-utilized

The Virtual Server Concept

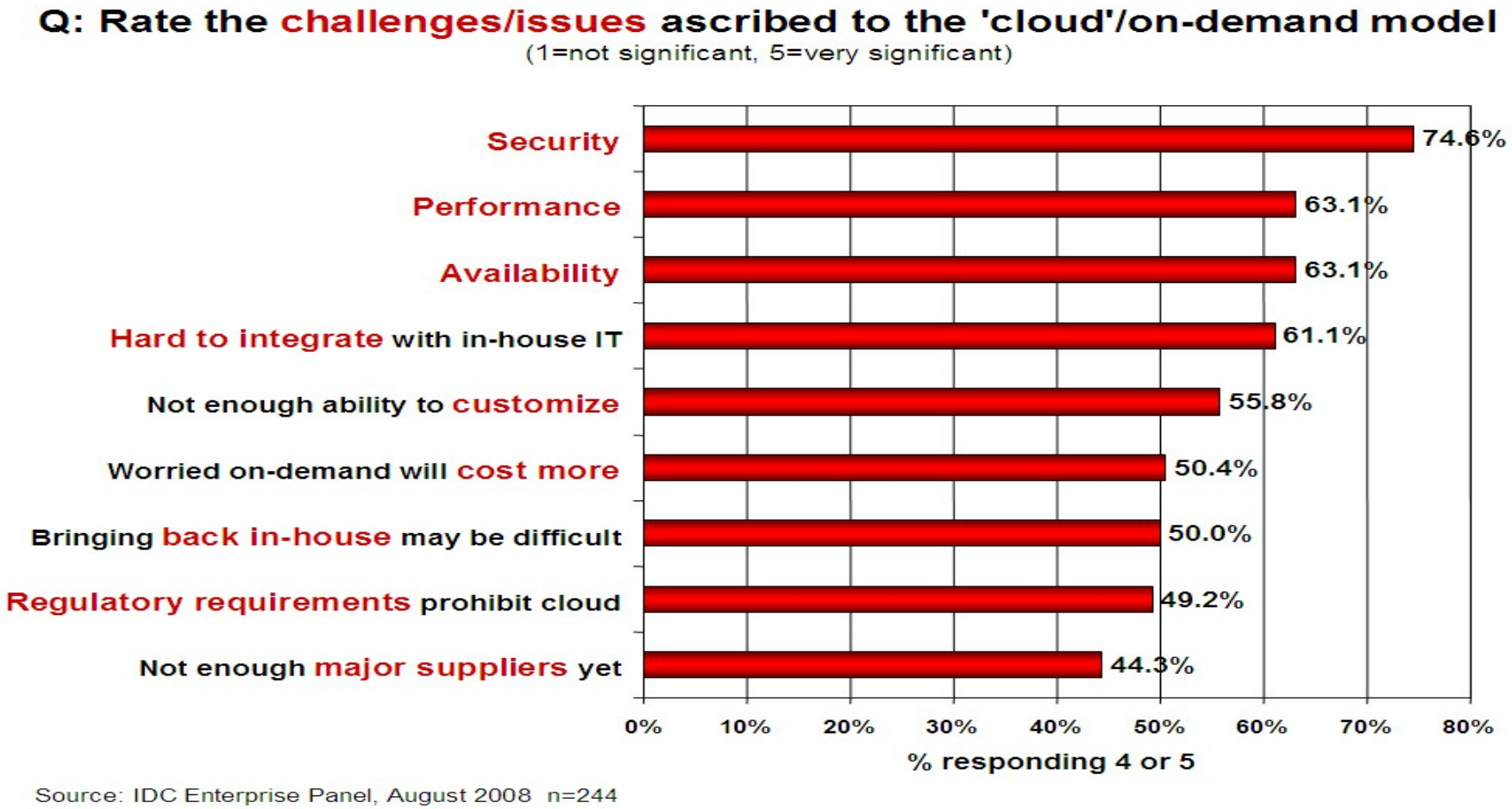


Virtual Machine Monitor layer between *Guest OS* and hardware

The Virtual Server Concept

- Virtual server
  = encapsulates the server software away from the hardware
  = OS, the applications, and the storage for that server.
- A virtual server can be mapped onto one or more hosts, and one host may house more than one virtual server.
- Virtual servers can still be referred to by their function i.e. email server, database server, etc.
- If the environment is built correctly, virtual servers will not be affected by the loss of a host (server migration).
- Hosts may be removed and introduced almost at will to accommodate maintenance
- Virtual servers can be scaled out easily.
  = If the resources supporting a virtual server are too expensive, the amount of resources allocated to the virtual server may be optimized

Security is the Major Issue



Cloud Security

Advantages

- Shifting public data to a external cloud reduces the exposure of the internal sensitive data
- Cloud homogeneity makes security auditing/testing simpler
- Clouds enable automated security management
- Redundancy / Disaster Recovery
- Dedicated Security Team
- Greater Investment in Security Infrastructure
- Hypervisor Protection Against Network Attacks

Challenges

- Trusting provider security model
- Customer inability to respond to audit findings
    - Obtaining support for investigations
- Indirect administrator accountability
- Proprietary implementations cannot be examined
- Loss of physical control
- Legislation