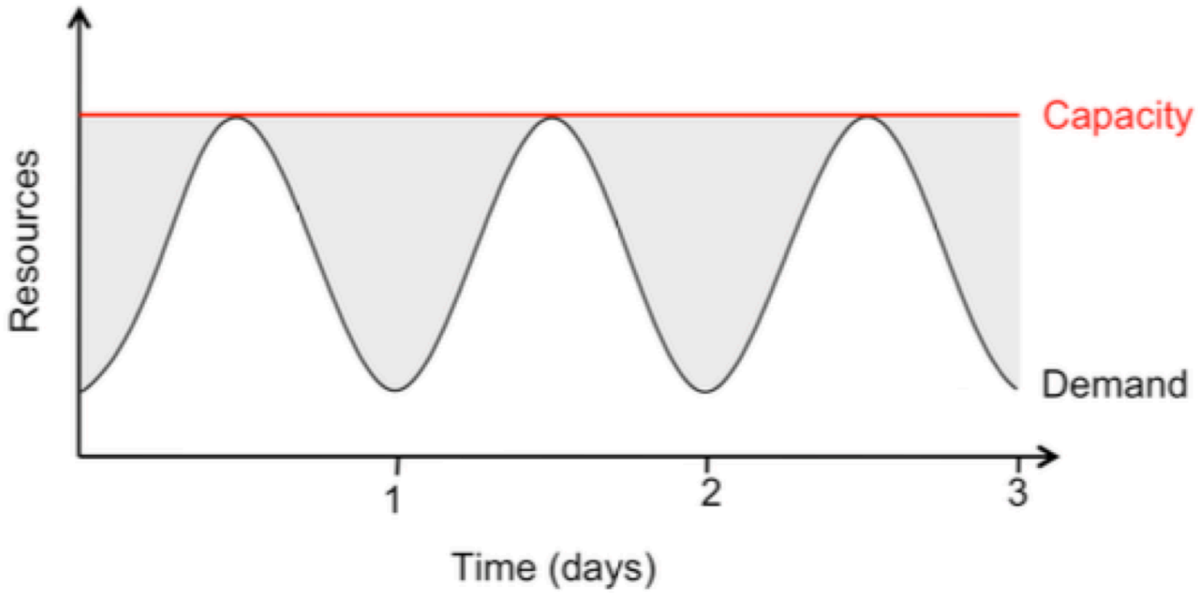
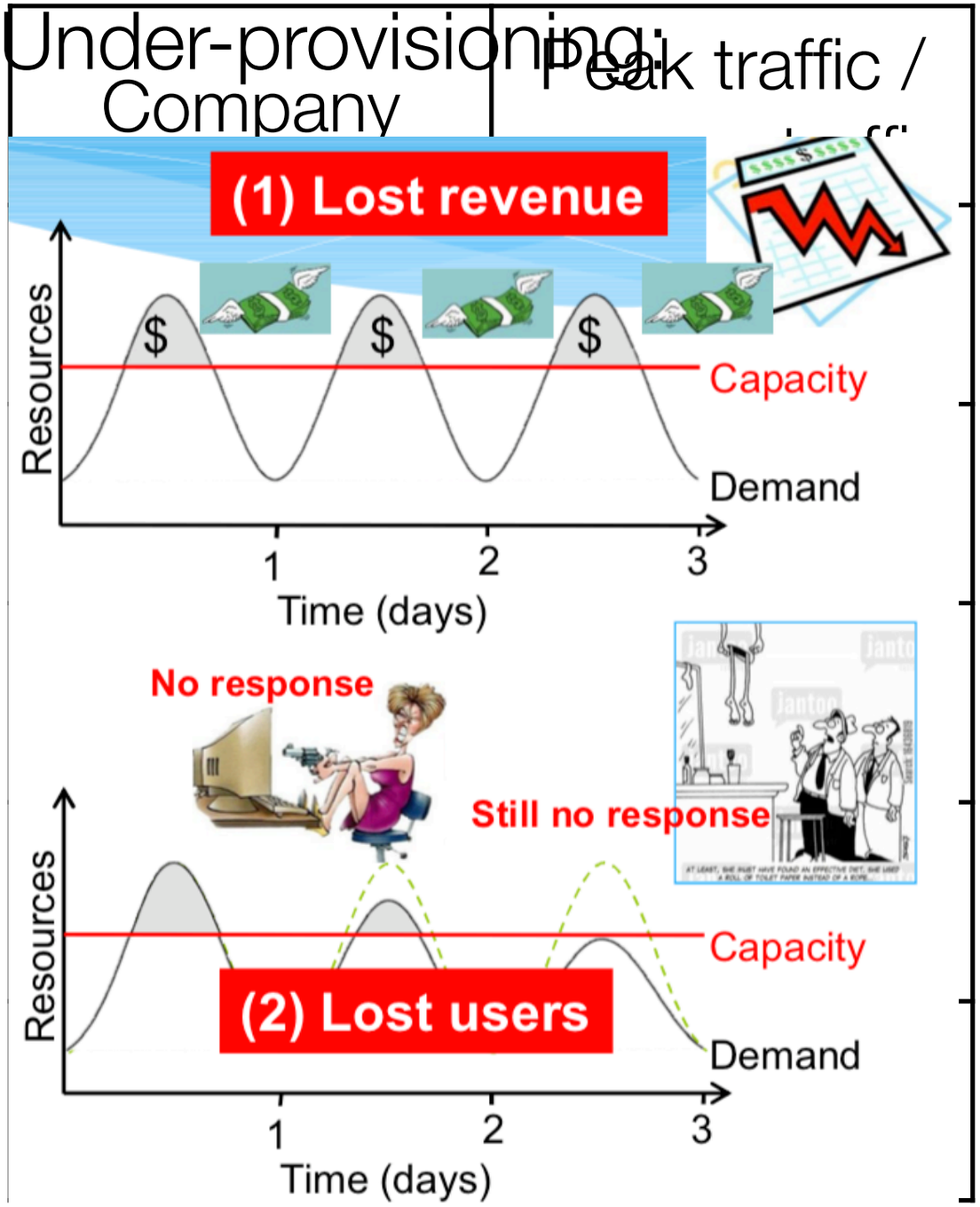
Resource planning at traditional data center

- Resource planning



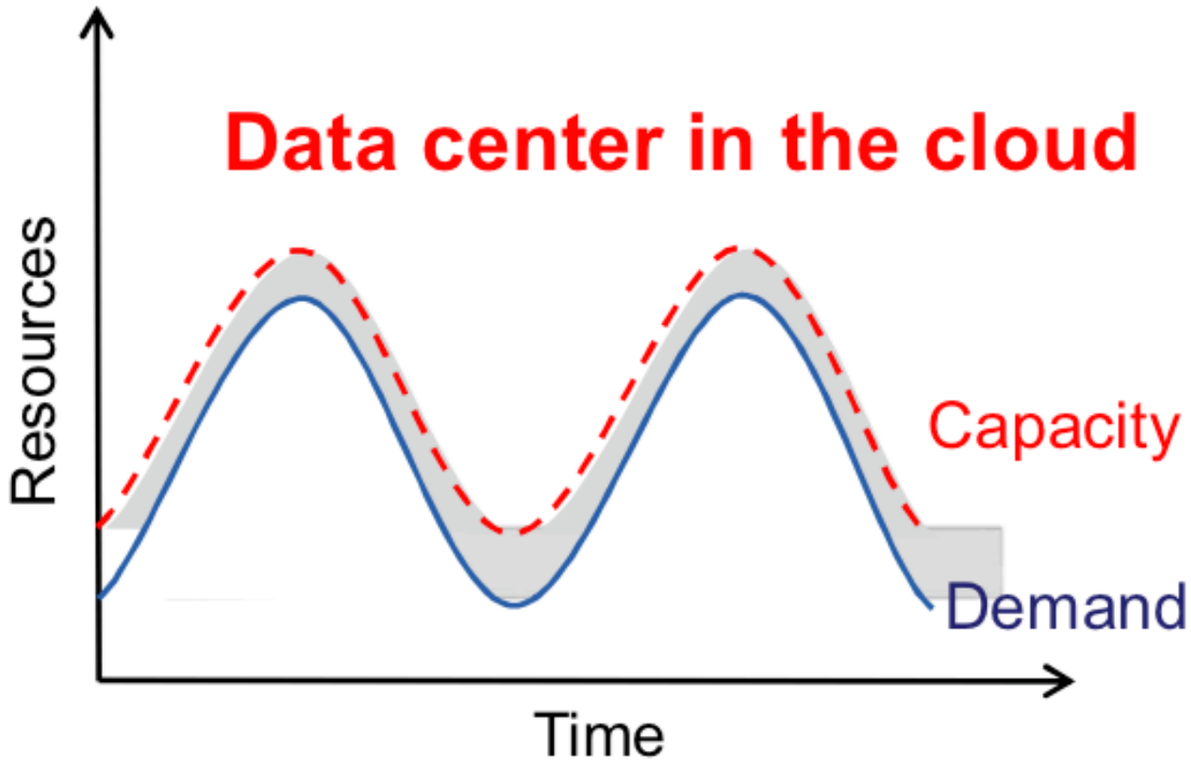
- Over-provisioning: low utilization

- Under-provisioning: Peak traffic / Company



Cloud economics

- Pay-as-you-go (usage-based) pricing:
  - Most services charge per minute, per byte, etc
  - No minimum or up-front fee
  - Helpful when apps have variable utilization



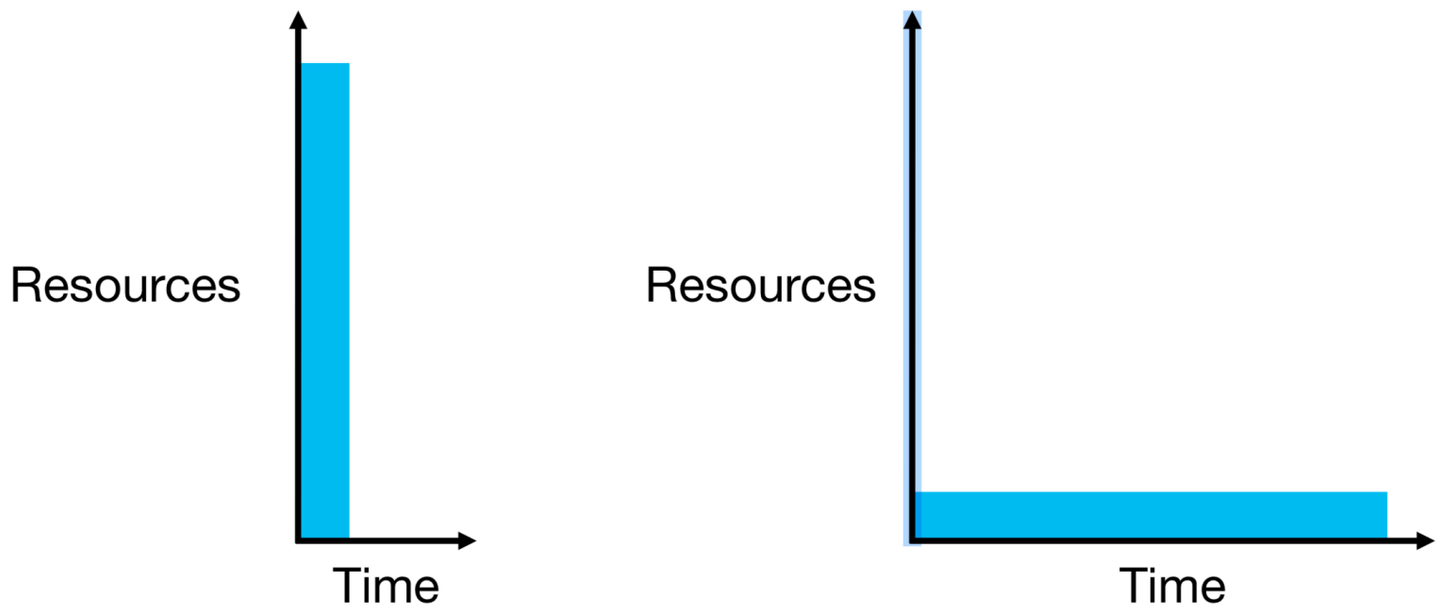
"Pay by use" (like electricity/water/gas)

Amazon EC2 On-Demand

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
a1.medium	1	N/A	2 GiB	EBS Only	$0.0255 per Hour
a1.large	2	N/A	4 GiB	EBS Only	$0.051 per Hour
a1.xlarge	4	N/A	8 GiB	EBS Only	$0.102 per Hour
a1.2xlarge	8	N/A	16 GiB	EBS Only	$0.204 per Hour
a1.4xlarge	16	N/A	32 GiB	EBS Only	$0.408 per Hour
a1.metal	16	N/A	32 GiB	EBS Only	$0.408 per Hour
t4g.nano	2	N/A	0.5 GiB	EBS Only	$0.0042 per Hour
t4g.micro	2	N/A	1 GiB	EBS Only	$0.0084 per Hour
t4g.small	2	N/A	2 GiB	EBS Only	$0.0168 per Hour
t4g.medium	2	N/A	4 GiB	EBS Only	$0.0336 per Hour
t4g.large	2	N/A	8 GiB	EBS Only	$0.0672 per Hour
t4g.xlarge	4	N/A	16 GiB	EBS Only	$0.1344 per Hour
t4g.2xlarge	8	N/A	32 GiB	EBS Only	$0.2688 per Hour
t3.nano	2	Variable	0.5 GiB	EBS Only	$0.0052 per Hour
t3.micro	2	Variable	1 GiB	EBS Only	$0.0104 per Hour

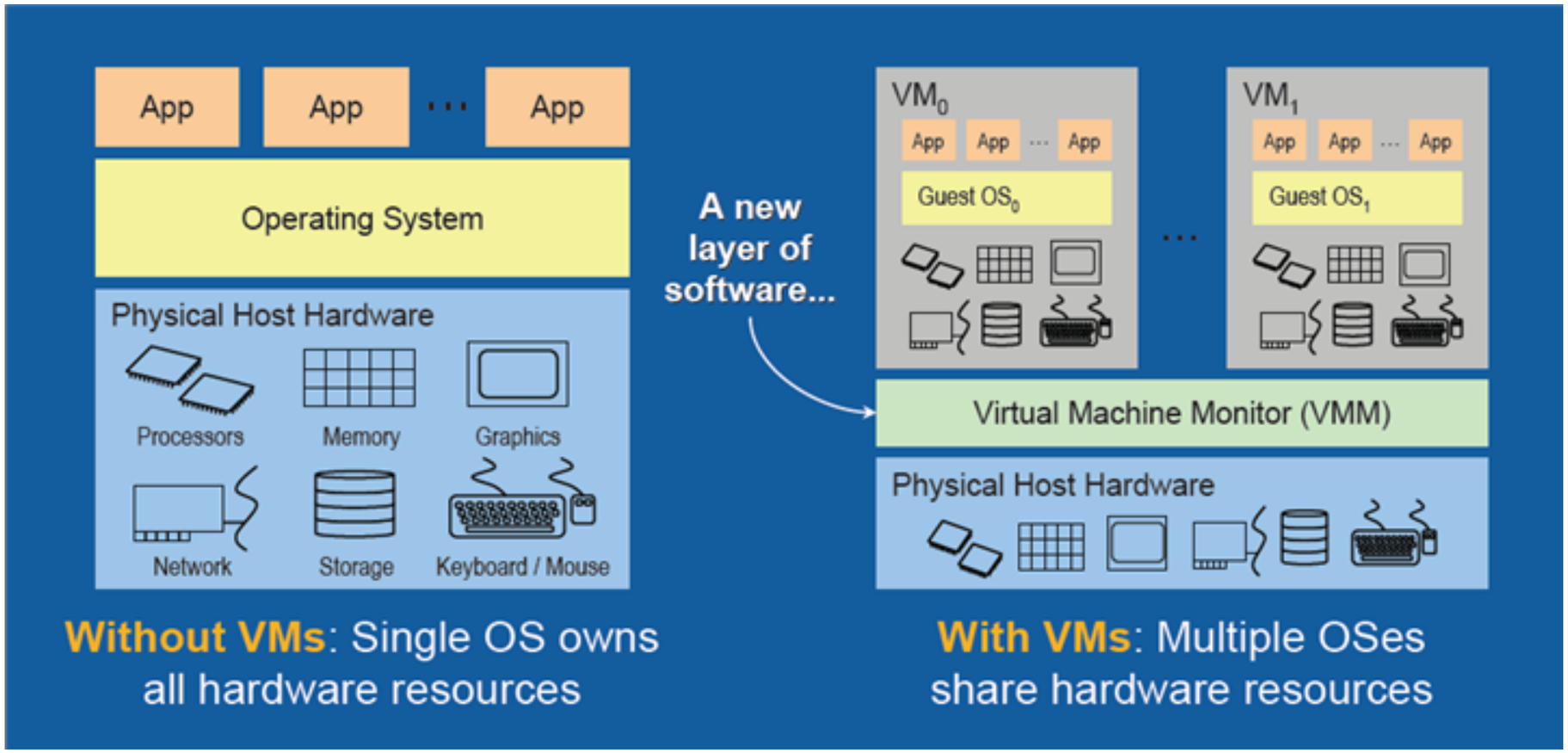
Cloud economics

- Elasticity:
  - Using 1000 servers for 1 hour costs the same as 1 server for 1000 hours
- Same price to get a result faster!



Enabling technology: virtualization

The power of virtualization



- Server consolidation: 10:1 in many cases
- Enable rapid deployment: an image can run on any hardware
- Dynamic load balancing: move hotspot to under-utilized hardware
- Disaster recovery: move affected VMs to other hardware

Major VMMs

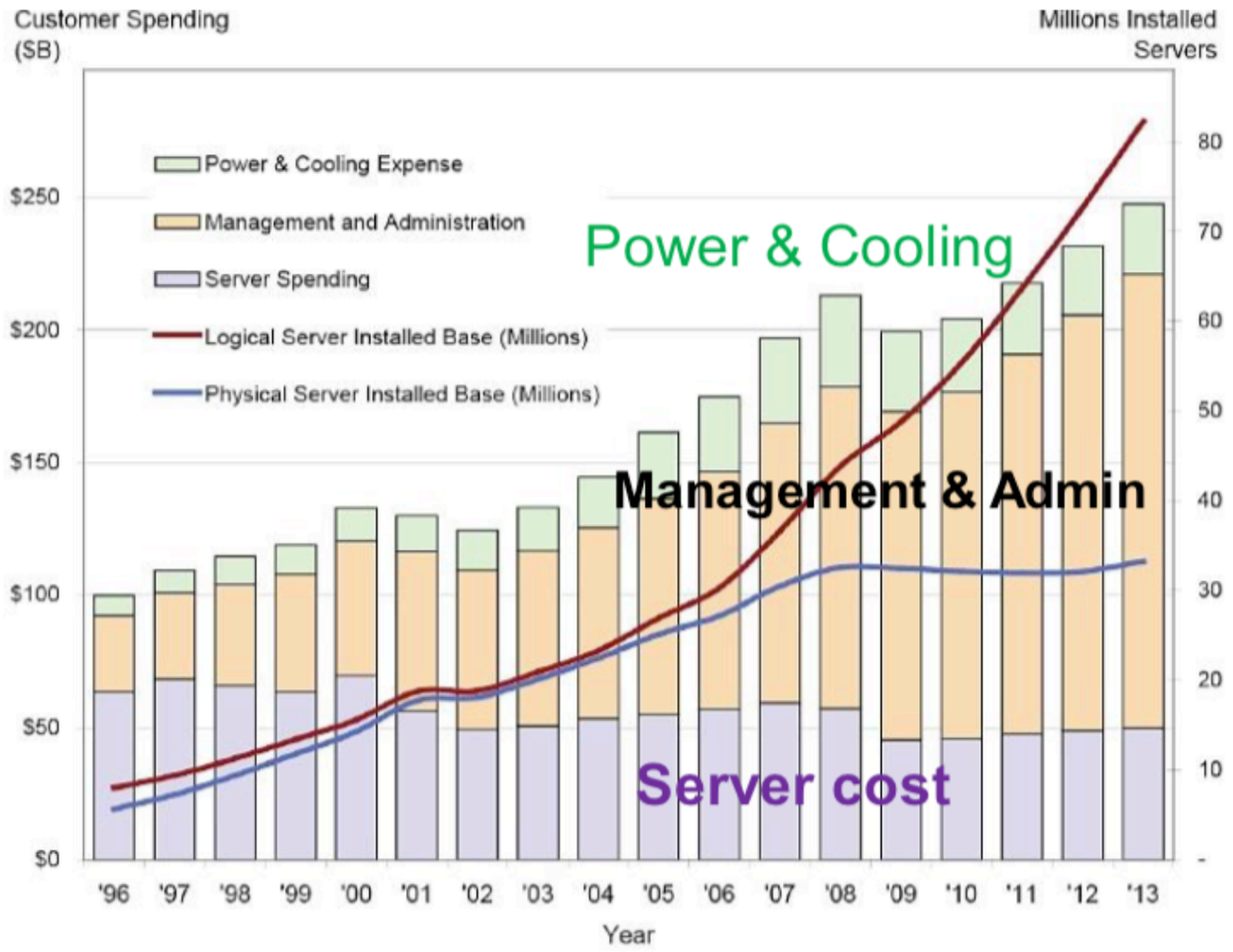
- Xen (2003):
  - University of Cambridge, Computer Laboratory, Fully open sourced (Backed by Citrix)
  - Host OS: NetBSD, Linux, Solaris
- VMware (1999) :
  - Closed source
  - Host OS: Windows, Linux, Mac OS X, no host OS (ESX Server, ESXi)
- KVM (Kernel-based Virtual Machine): Jan. 2007
  - Open source, included in Linux Kernel (>= 2.6.20)
  - HostOS: FreeBSD,Linux

Cloud economics

- Economies of scale:
  - Purchasing, powering, managing machines at scale gives lower per-unit costs than customers'

Economics of Big Scale

Traditional data centers are notoriously under-utilized, often idle **85%** of the time.



Source: "The Cost of Retaining Aging IT Infrastructure ".

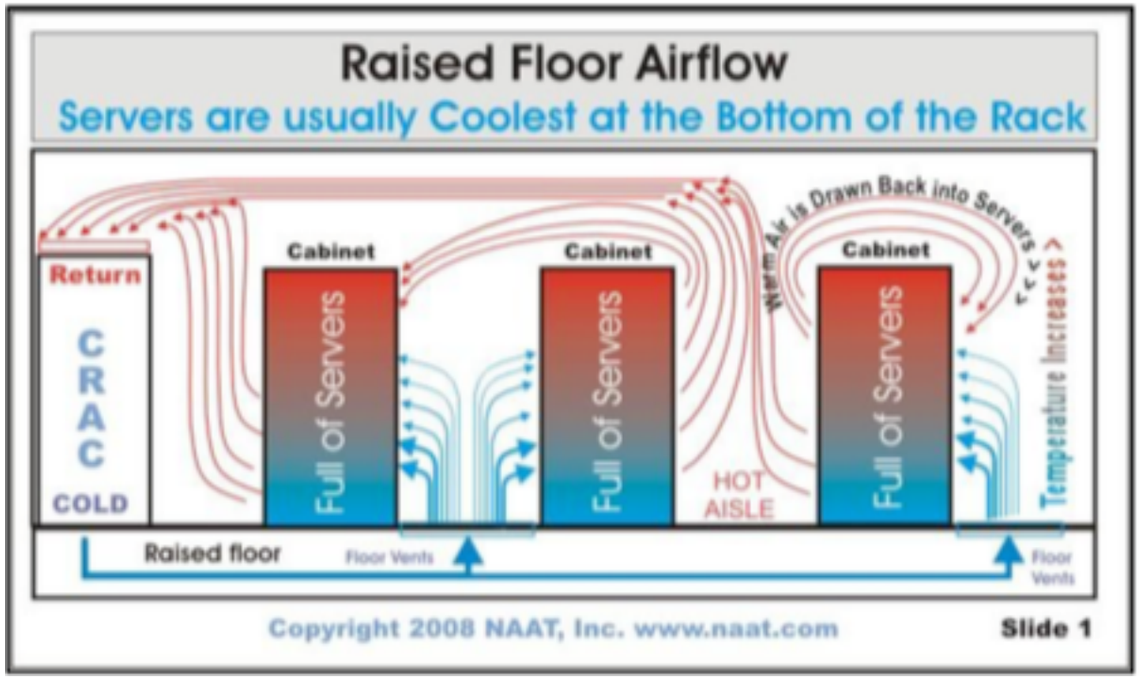
Average life of a data center: 9 years.

Data centers more than 7 years old are considered out of date as per Green Computing norms.

Data center cost break down :

Electricity: 20%
Engineering and Installation
Manpower: 18%
Power and Server Equipment: 18%
Facility Space: 15%
Service and Maintenance: 15%
HVAC Equipment: 6%
Project Management: 5%
Rack Hardware: 2%
System Monitoring: 1%

Cooling in Traditional Data Centers



Economics of “big” cloud

Resource	Cost in Medium DC	Cost in Very Large DC	Ratio
Network	$95 / Mbps / month	$13 / Mbps / month	**7.1x**
Storage	$2.20 / GB / month	$0.40 / GB / month	**5.7x**
Administration	**≈140 servers/admin**	**>1000 servers/admin**	**7.1x**

User’s Voice:

“I only care about results, not how IT capabilities are implemented”

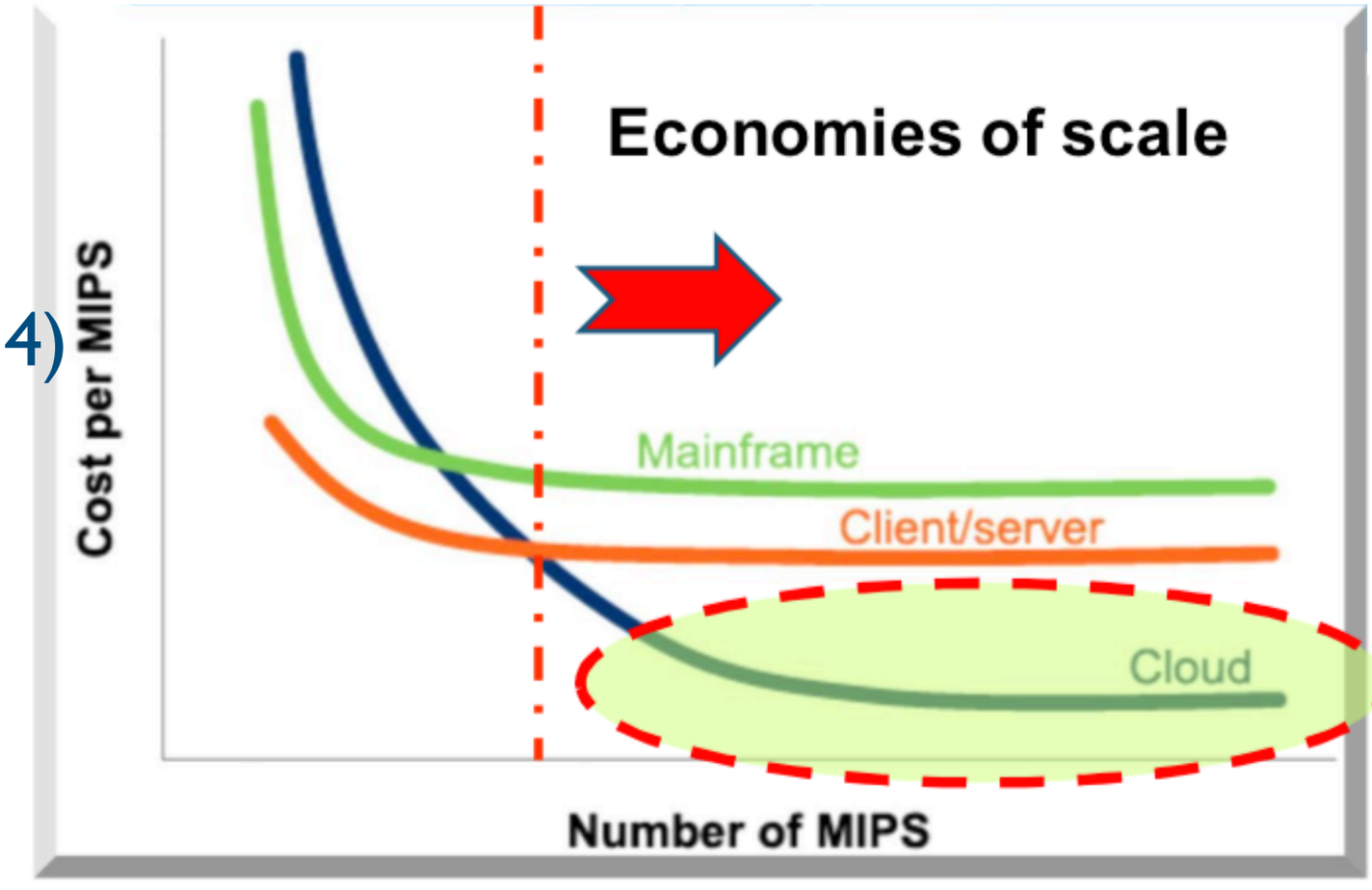
"I want to pay for what I use, like a utility“

"I can access services from anywhere, from any device"

“I can scale up or down capacity, as needed"

Economics of “big” cloud

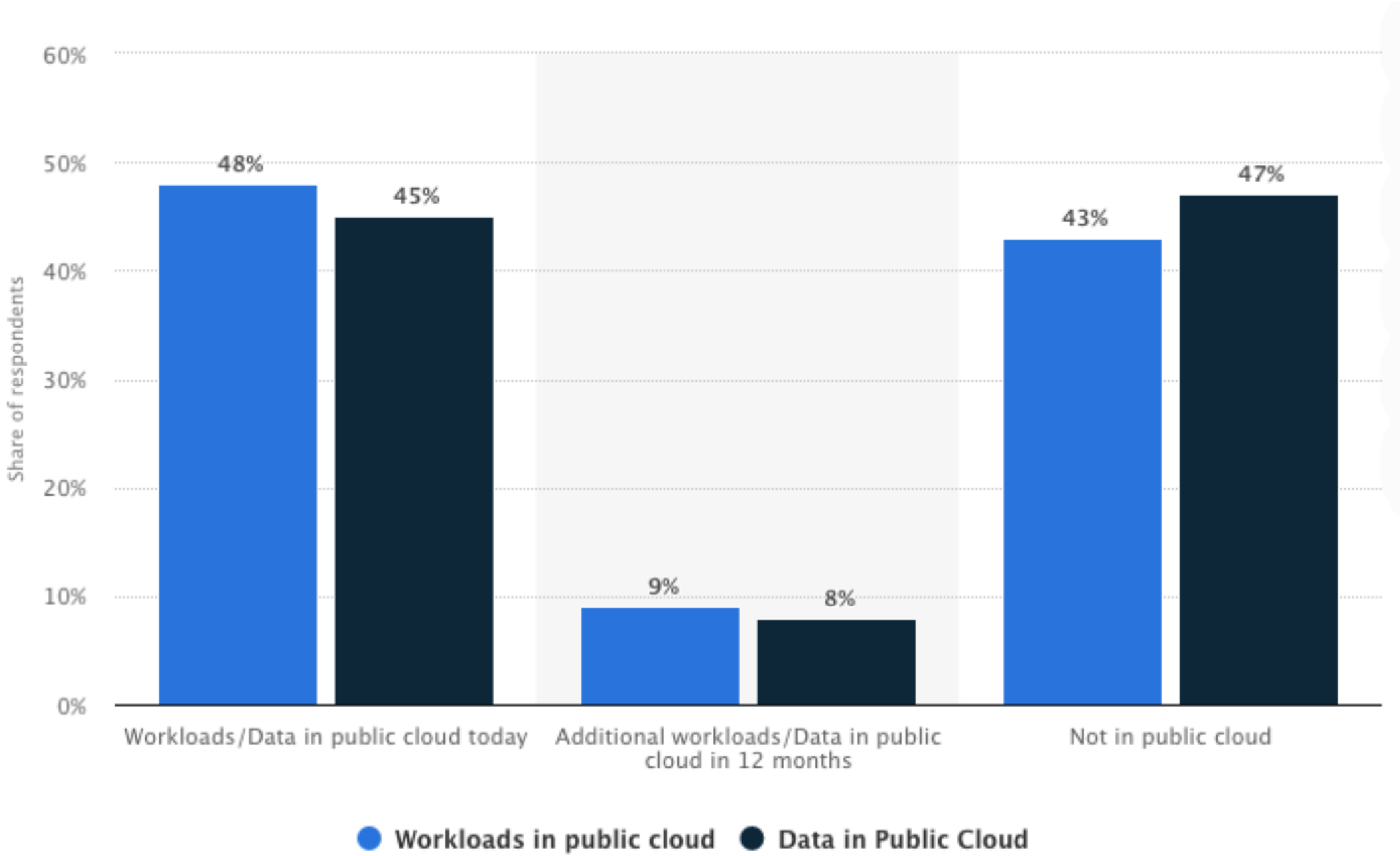
* **Google: > 1 million.**
  * 900,000 in 2010
* **Microsoft: > 1 million (2013)**
* **Amazon : ~~158,000 ? (Netcraft)~~**
  * ~~4,600 (2009)~~ 1.4 million (2014)
* **Facebook: 180,000**
* **Intel: 100,000**
* **Yahoo: 50,000 (2010)**
* **Ebay: 50,000**
* **Rackspace: 50,000**
* **Time Warner: 25,000**
* **AT&T: 20,000**



MIPS (million instructions per second)

- **83% Of Enterprise Workloads Will Be In The Cloud By 2020; 41% of enterprise workloads will be run on public cloud!**

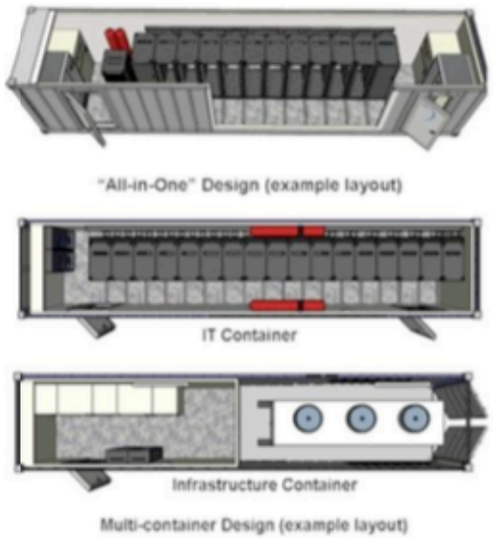
Worldwide enterprise workload/data in public cloud 2020 (Source: statista)

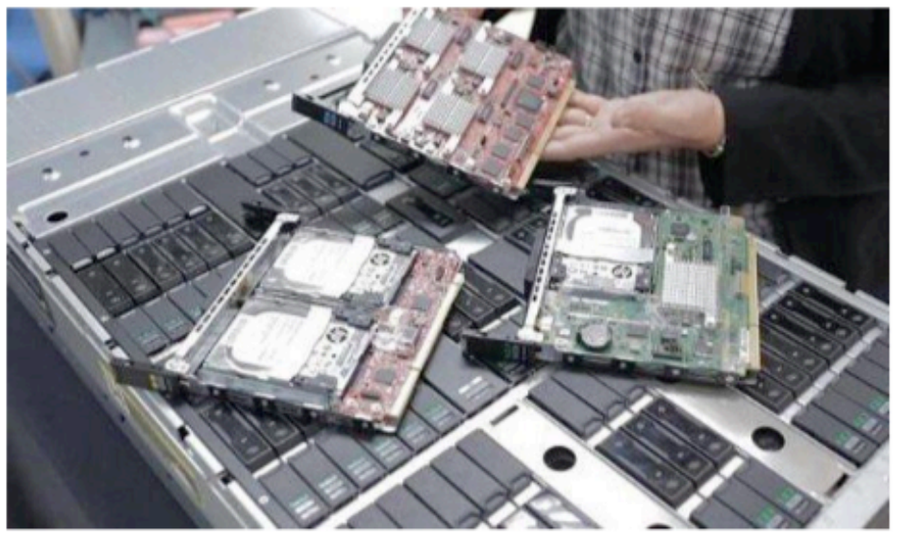


Containerized data centers

According to IBM, it can cost **30 per cent** less to design and build a **containerized data centre** than a traditional one with air-conditioning and possibly raised floors or dropped ceilings.

Microsoft's Chicago Data Center (2,000 servers per container)





45 X-Gene cartridges in one 4.3U chassis. Each has 64GB DDR3 DRAM and 480GB flash and Ubuntu server 14.04 LTS

Cloud Deployment Models

Cloud deployment mode

- **Public Cloud**
  Cloud infrastructure is available to the general public, owned by organization selling cloud services (Google, Amazon)

- **Private Cloud**
  Cloud infrastructure for single organization only, may be managed by the org or a 3rd party (Virtual Private Cloud).

- **Community Cloud**
  Cloud infrastructure shared by several organizations that have shared concerns, managed by org or 3rd party (e.g., Amazon)

- **Hybrid Cloud**
  Combo of >=2 clouds bound by standard or proprietary technology

Public cloud

- Large scale infrastructure available on a rental basis
  - Customers access resources remotely via the Internet under some Service Level Agreements (SLAs)
  - E.g., Amazon, Google, Microsoft, Rackspace, Apple's iCloud, HP, Oracle, IBM, VMware
    - usually has a global network of data centres
- Fully customer self-service
  - Requests are accepted and resources granted via web services
- Accountability is e-commerce based
  - Web-based transaction
  - Utility style costing
  - "Pay-as-you-go" customer service, refunds, etc.

Private clouds

- A cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party and hosted internally or externally.
- Key techniques:
  - Virtualization techniques (VMWare, Xen, KVM),
  - Virtual private network (VPN) (for corporations across many industries).
- More expensive, but more secure (behind a firewall) when compared to public clouds.
  - "still have to buy, build, and manage them”
  - May not free you from the responsibility for procuring (hardware + software upgrade $$) and maintenance (in-house expertise $$$)

Private clouds

- Examples: The New York Times' "TimesMachine": 15 million articles were put 'into the cloud' on servers owned by Amazon.

  https://open.nytimes.com/the-new-york-times-archives-amazon-web-services-timesmachine-e00518ebf6da

- Open-source Software for building private and public clouds:
  - OpenStack: https://www.openstack.org/
  - Apache CloudStack: http://cloudstack.apache.org/

Community clouds

- A community cloud is a multi-tenant infrastructure that is shared among several organizations from a specific community with common concerns, such as
  - High Security: Access to the cloud is granted only after a trusted validation of identity (required by regulating bodies).
  - High Availability: Resources are 99.999% available (or better), e.g., banking
  - High Performance: optimized for high transaction rates and extremely low-latency, e.g., high-frequency trading (HFT).

Community clouds: examples

- GovernmentClouds:
  - Government organizations may share computing infrastructure on the cloud to manage data related to citizens.
- E.g., Amazon AWS GovCloud, IBM SmartCloud for Government (SCG).

IBM Cloud for Government and IBM SmartCloud for Government meet FedRAMP security requirements.

Government | **Industry**

The Department of Veterans Affairs issued a FISMA High Authority to Operate (ATO) for AWS GovCloud (US), using the regions to store and protect patient data critical to America's veterans. Learn more »

Cloud.gov, built by GSA's 18F, helps other government agencies build, buy, and share technology products, while minimizing the FedRAMP compliance work they need to execute themselves. Learn more »

The Department of Justice leverages AWS GovCloud (US) for mission-critical workloads, DevTest, and delivery of advanced capabilities. Learn more »

Defense Digital Services manages the U.S. Air Force's Next Generation GPS Operational Control system of satellites, securely running 200+ dedicated hosts and 1,000 individual virtual machines in AWS GovCloud (US). Learn more and read about their DoD IL5 workload in AWS GovCloud (US).

The U.S. Department of Treasury delivers mission assurance while enabling digital transformation in AWS GovCloud (US). Learn more »

The U.S. Department of Homeland Security's HSIN information sharing platform is a FedRAMP High system, securely enabled in AWS GovCloud (US). Learn more »

The State of Kansas and AWS Partner, PayIt, enhanced the citizen experience with government services by deploying an online and mobile license renewal app in less than 60 days. Learn more »

NASA Jet Propulsion Laboratory innovates while improving governance, security, and compliance in AWS GovCloud (US). Learn more »

Government | **Industry**

Lockheed Martin lowers capex and addresses its ITAR requirements by moving its SAP HANA ERP suite of applications to the AWS GovCloud (US) Regions. Learn more »

FIGmd operates one of the largest clinical-data registries in the U.S. See how they achieve HIPAA, HITECH, ACA, and FedRAMP compliance and reduce risks associated with the transmission of sensitive data. Learn more »

Raytheon deploys test environments in 15 minutes instead of four months, in AWS GovCloud (US). Learn more »

Motorola Solutions improves public safety and addresses CJIS requirements with their platform, which detects missing persons with Amazon Rekognition in AWS GovCloud (US). Learn more »

Government Cloud Plus runs on AWS GovCloud (US) and meets FedRAMP's High Baseline, enabling enhanced security and compliance controls that allow customers to use Salesforce for the most sensitive, unclassified data. Learn more »

After acquiring a series of independent power producers (IPPs) and their IT assets, Talen Energy migrated to AWS GovCloud (US) to meet its NRC, FERC, and 10 CFR 810 requirements. Learn more »

The Cobham Advanced Electronic Solutions InfoSec team replaced its no-cloud policy with a cloud-first policy, improving agility, innovation, and security in the process. Learn more »

GDIT leverages its hands-on experience with AWS GovCloud (US) to better serve government customers with strict regulatory compliance needs, including FedRAMP High requirements. Learn more »

Community clouds: examples

- HPC Clouds**: Cloud for High-Performance Computing
  - A different set of requirements: (1) Close to the “metal”, (2) User-space communication (bypass OS), (3) high-speed interconnect (e.g., InfiniBand)
    - Clients share a common set of "Big Data" - ranging in size from Terabytes to Petabytes. Need a very fast I/O subsystem (SSD-based storage)

- ** Moving HPC to the Cloud: http://www.admin-magazine.com/HPC/Articles/Moving-HPC-to-the-Cloud

Community clouds: examples

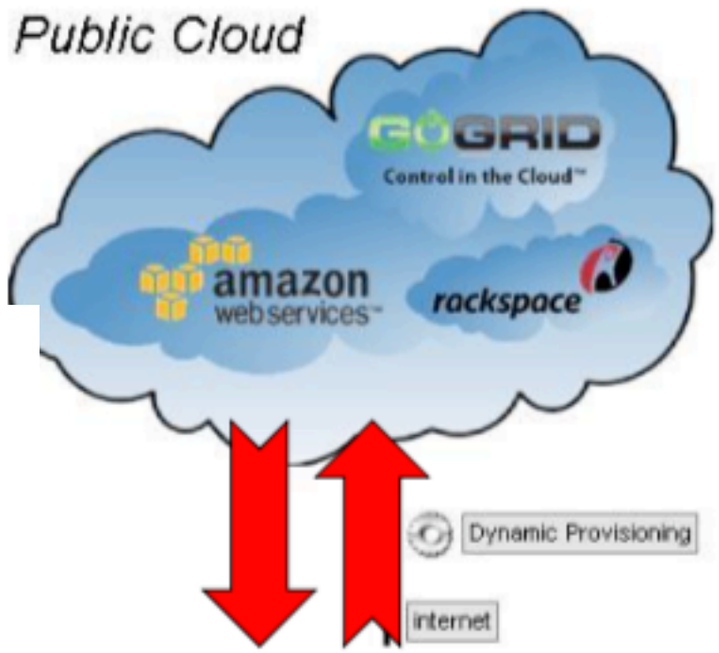
- Financial Services Clouds:
  - Require microseconds or at least milliseconds of response time and latency measurements.

- NYSE Technologies: Financial Services Community Cloud (electronic trading,
  market data analysis, etc.)
  (Read https://www.businesswire.com/news/home/20110601006045/en/NYSE-Technologies-Introduces-the-World%E2%80%99s-First-Capital-Markets-Community-Platform)

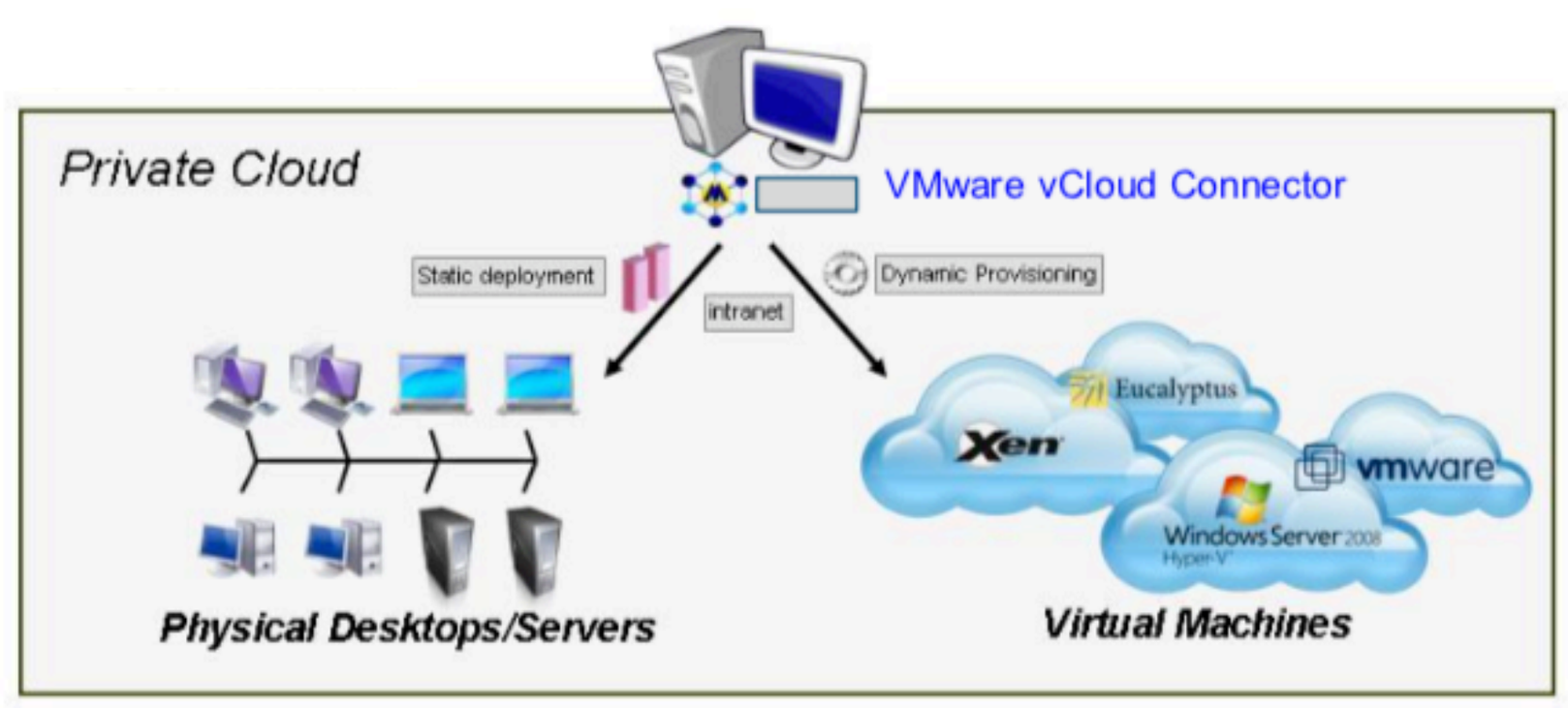
Hybrid clouds

- An integrated cloud service **utilizing both private and public clouds** to perform distinct functions within the same organization
  - e.g., non-sensitive operations on public cloud; and sensitive operations handled in-house (private cloud).
  - bound together by standardized or proprietary technology that enables data and application portability.
- Take advantage of the **scalability** and **cost-effectiveness** of the public cloud, while keeping sensitive data in the **secure** environment of a private cloud without exposing to the public cloud.

Hybrid Cloud:
seamlessly combine **Private** and **Public** Cloud resources, and deploy applications in these heterogeneous and hybrid environments.



Automatically grow workloads into public cloud resources for a period of time, and then decommission them once the heavy loads subside



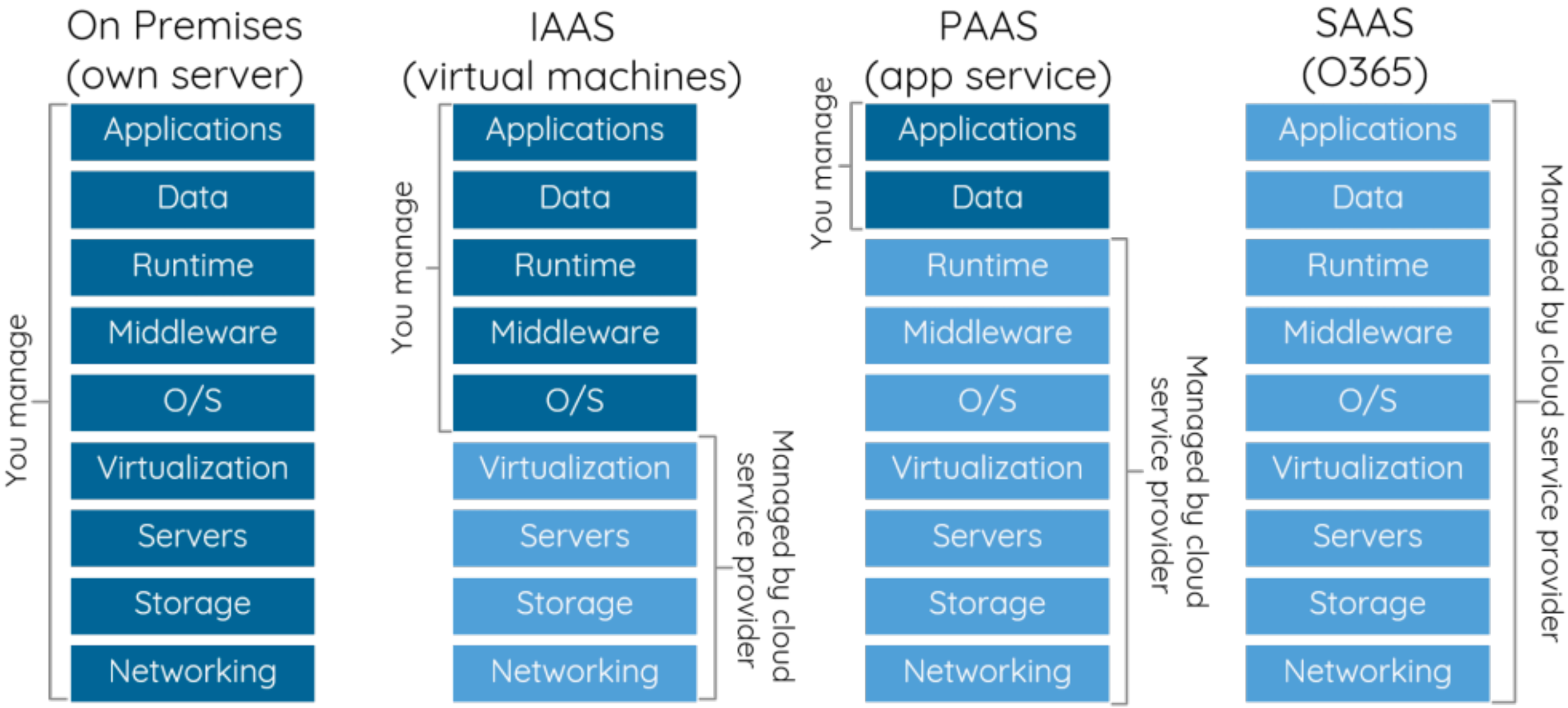
Cloud service models

- Software as a Service (SaaS): A provider licenses an **application** to customers for use as a service on demand. (Gmail/Hotmail, Web hosting, etc.)

- Platform as a Service (PaaS): Provide the **software platform** where systems run on. The sizing of the hardware resources demanded by the execution of the services is made in a transparent manner.

- Infrastructure as a Service (IaaS): Through **virtualization**, split, assign and dynamically resize the resources to build ad-hoc systems as demanded by customers, or the service providers (SPs).

Separation of responsibility

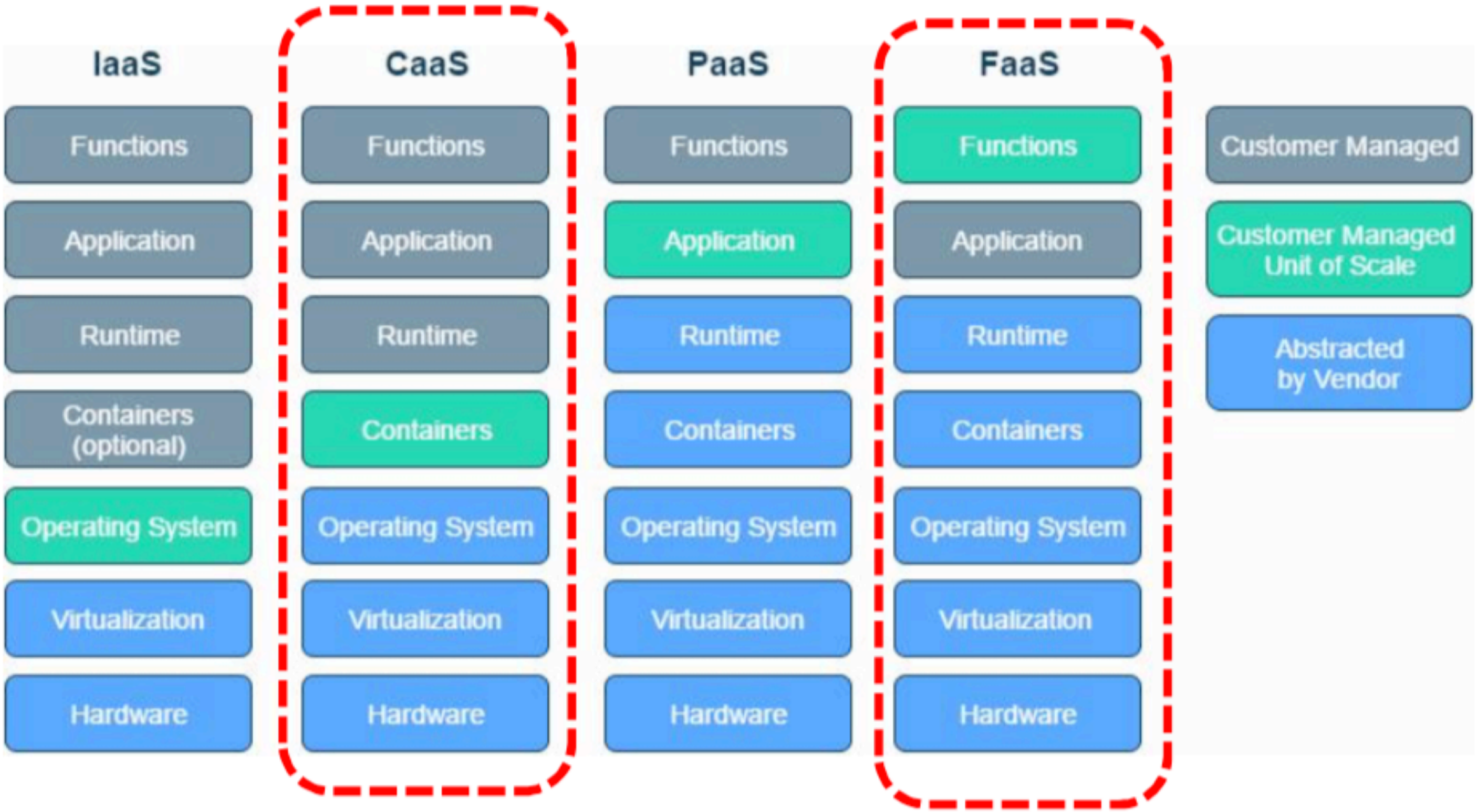


Could you name some example?

Newer models

- Containers as a Service (Caas) : Docker Cloud, Amazon Amazon Elastic Container Service (ECS).

- Functions as a Service (FaaS): AWS Lambda, Google Cloud Functions, Azure Functions, etc. Also known as serverless


IaaS
CaaS
PaaS
FaaS
Functions
Application
Runtime
Containers (optional)
Operating System
Virtualization
Hardware
Functions
Application
Runtime
Containers
Operating System
Virtualization
Hardware
Functions
Application
Runtime
Containers
Operating System
Virtualization
Hardware
Functions
Application
Runtime
Containers
Operating System
Virtualization
Hardware
Customer Managed
Customer Managed Unit of Scale
Abstracted by Vendor

Next: Data Center