CS15-319: Cloud Computing

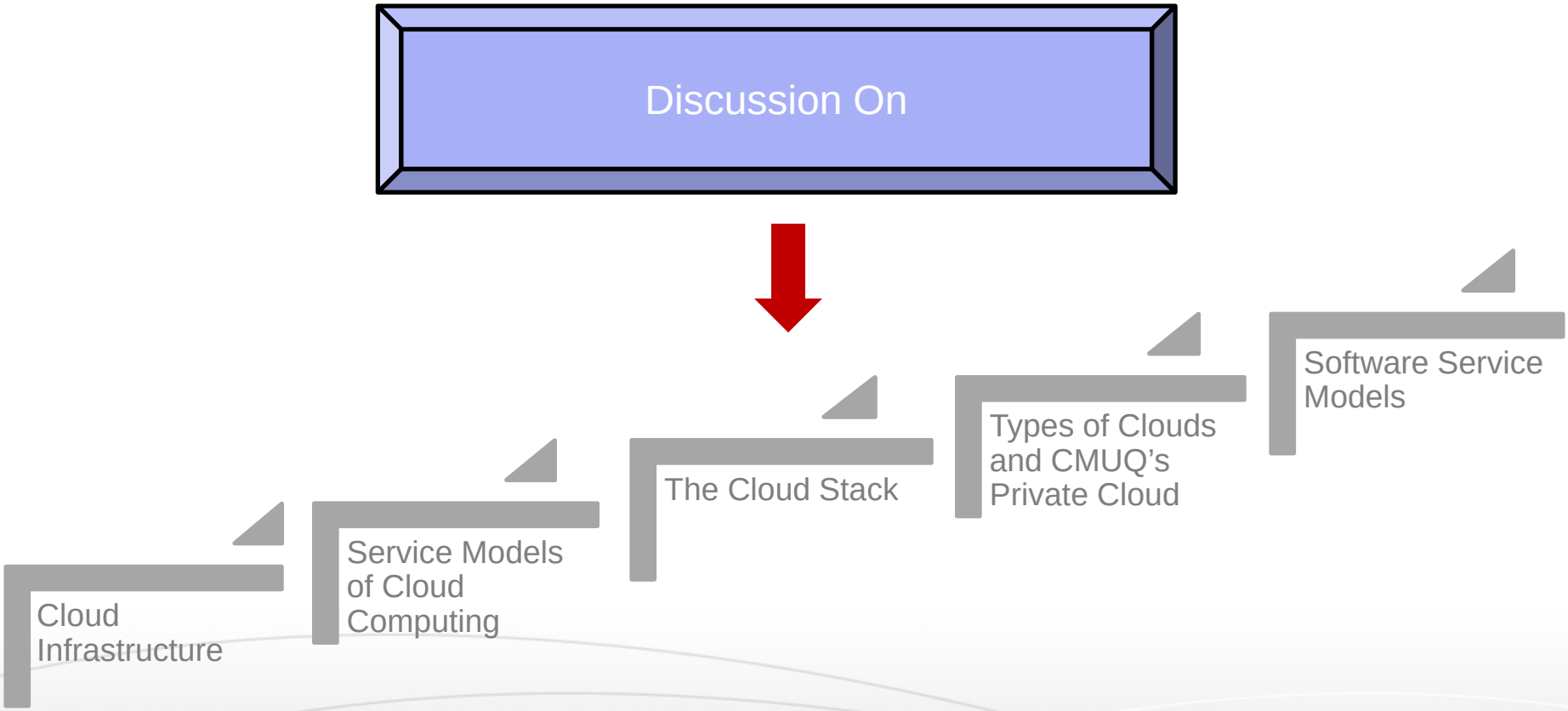
Lecture 2

Introduction to Cloud Computing

Prof. Majd F. Sakr

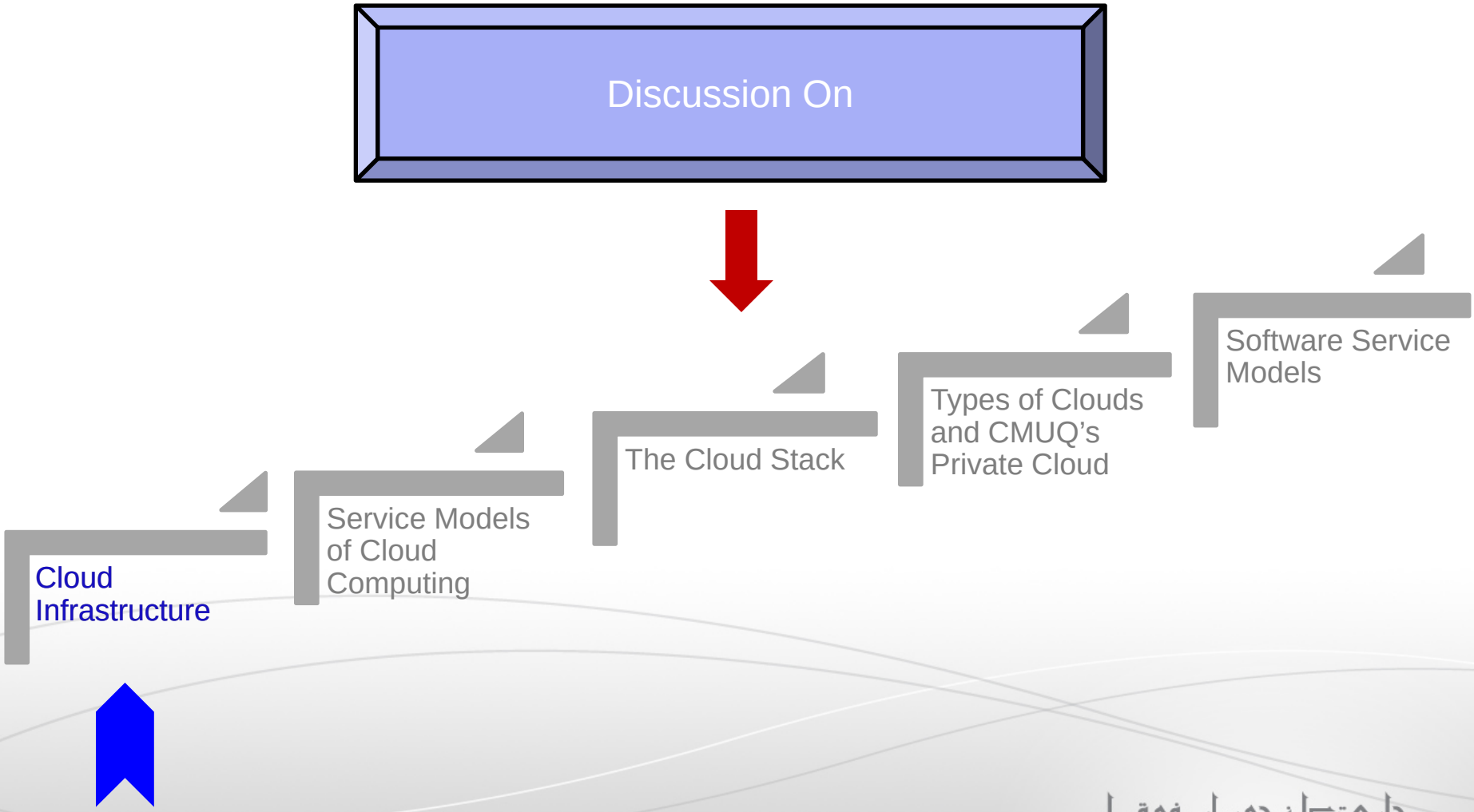
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Lecture Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Lecture Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

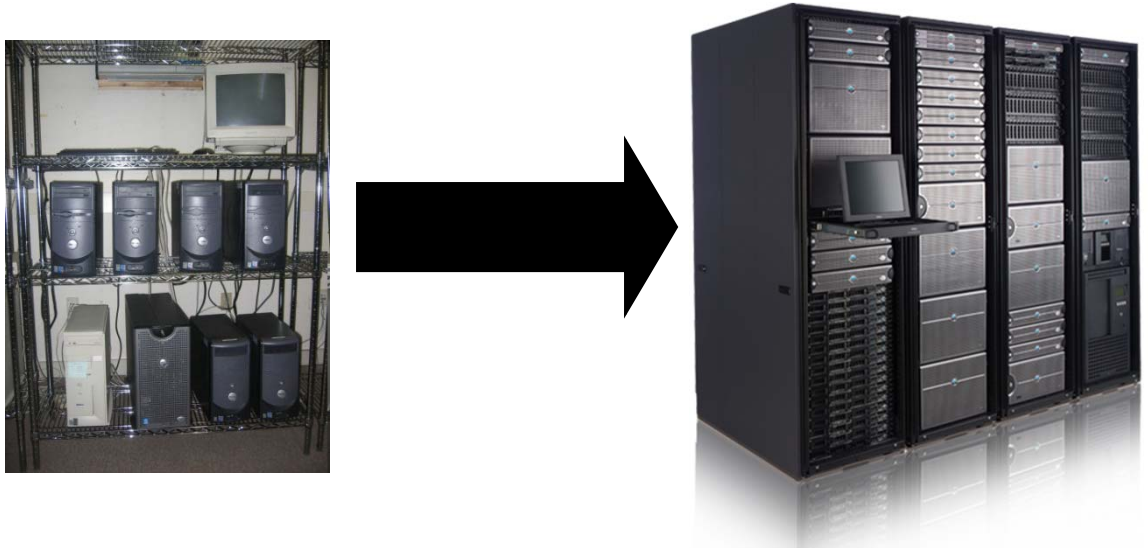
What is a Server?

- Servers are computers that provide “services” to “clients”
- They are typically designed for reliability and to service a large number of requests
- Organizations typically require many physical servers to provide various services (Web, Email, Database, etc.)
- Server hardware is becoming more powerful and compact

Compact Servers

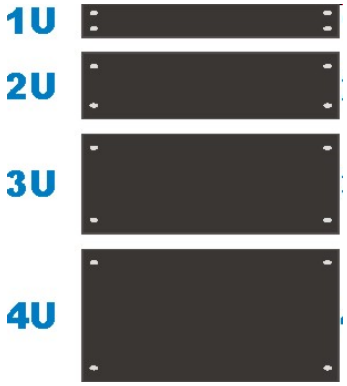
- Organizations would like to conserve the amount of floor space dedicated to their computer infrastructure

- For large-scale installations, compact servers are used. This helps with:
  - Floor Space
  - Manageability
  - Scalability
  - Power and Cooling



Racks

- Equipment (e.g., servers) are typically placed in racks
- Equipment are designed in a modular fashion to fit into rack units (1U, 2U etc.)
- A single rack can hold up to 42 1U servers



1U Server

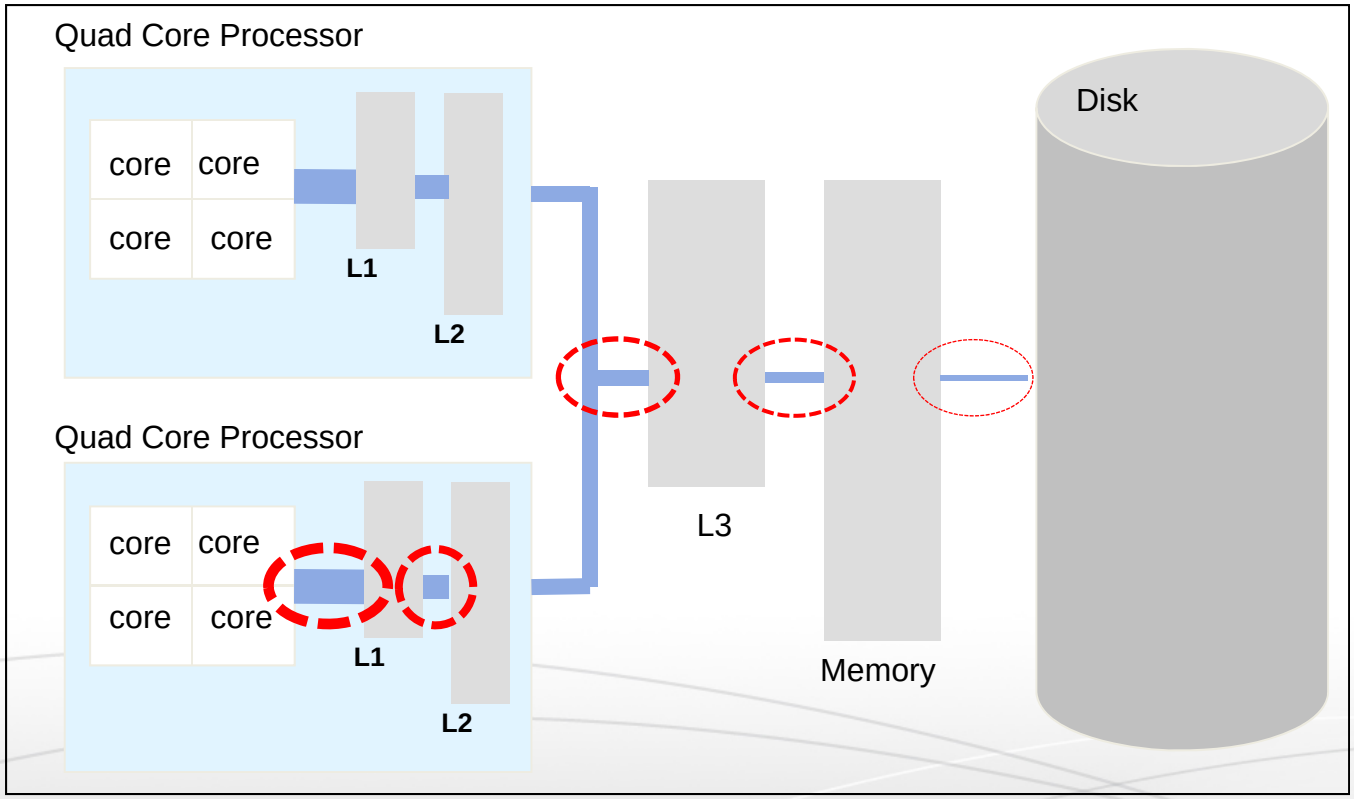
7U Blade center

Blades and Blade Enclosures

- A blade server is a stripped down computer with a modular design
- A blade enclosure holds multiple blade servers and provides power, interfaces and cooling for the individual blade servers

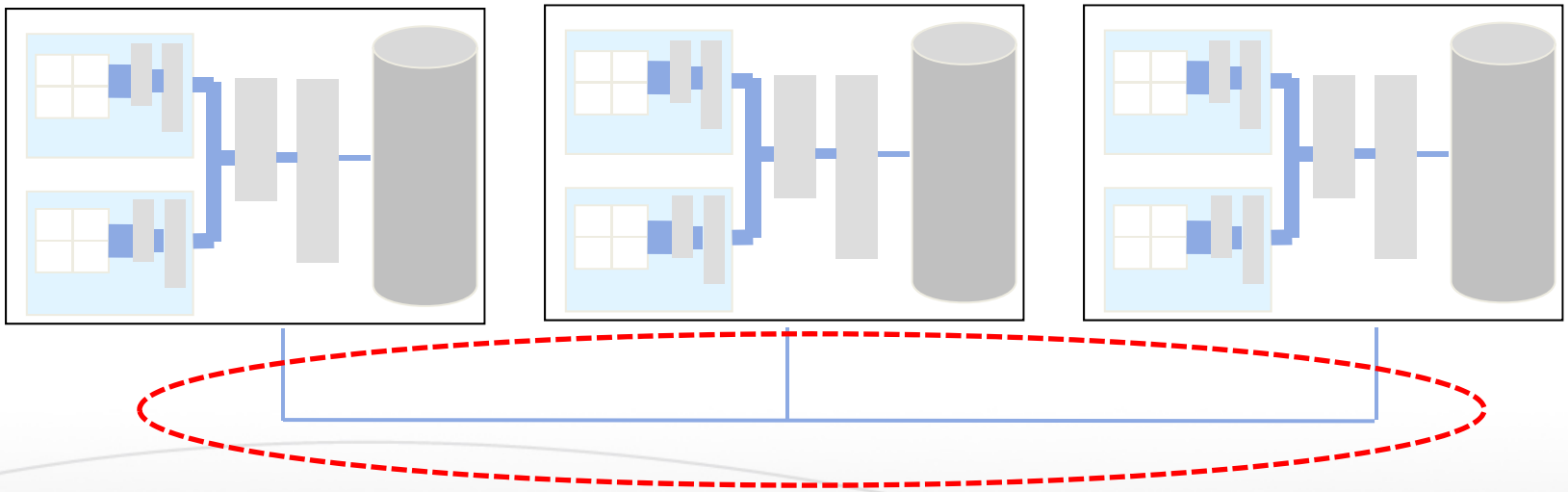
Blade Performance

- Consider bandwidth and latency between these layers



…Performance across blades

- Consider bandwidth and latency across blades



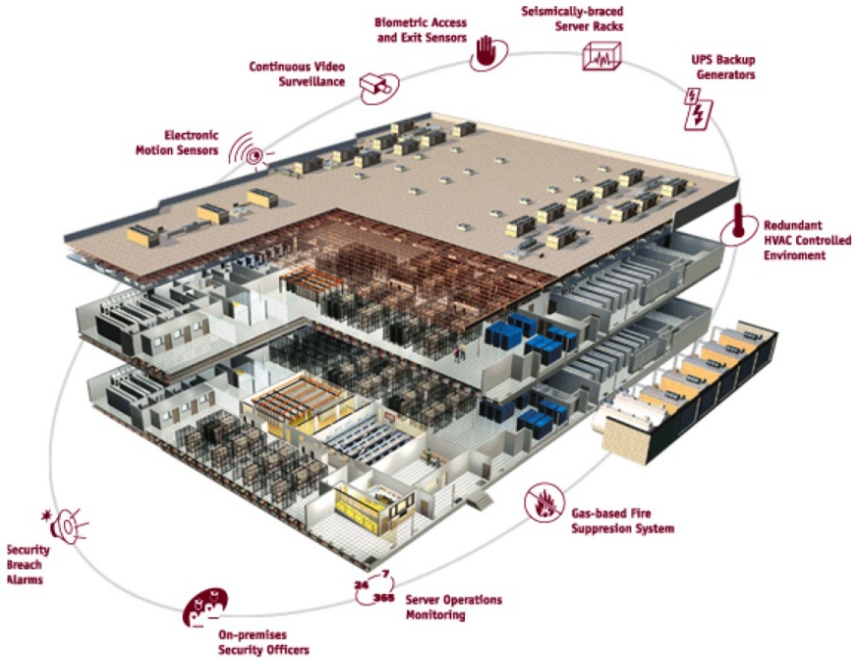
Network is usually the bottleneck

What is a Data Center?

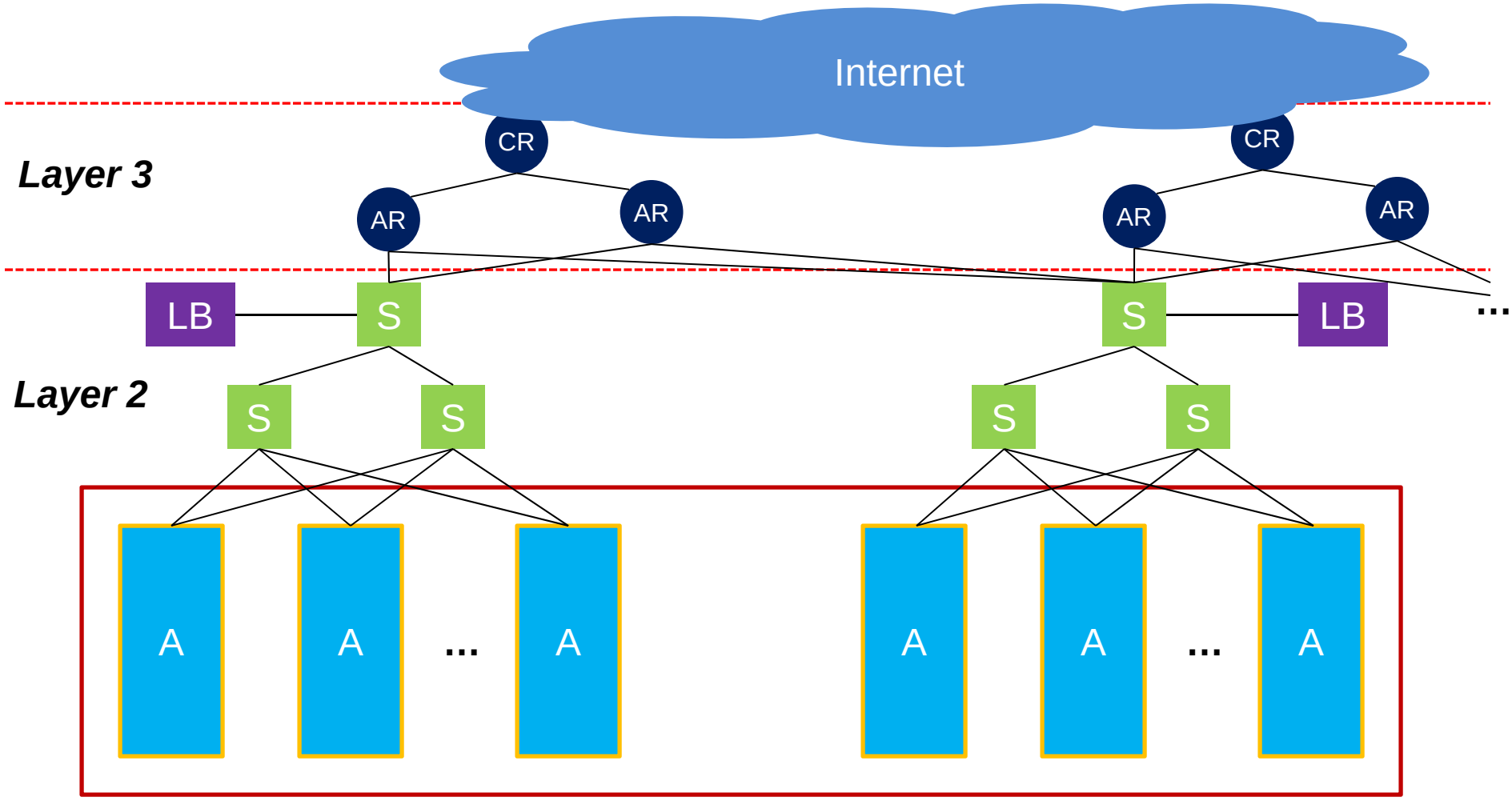
- A data center is a facility used to house computer systems and associated components, such as networking and storage systems, cooling, uninterruptable power supply, air filters…
- A data center typically houses a large number of heterogeneous networked computer systems
- A data center can occupy one room of a building, one or more floors, or an entire building

Data Center Components

- Air conditioning
  - Keep all components in the manufacturer's recommended temperature range
- Redundant Power
  - UPS/Generators
  - Multiple power feeds
- Fire protection
- Physical security
  - CCTV/Access Control
- Monitoring Systems
- Connectivity
  - Multiple ISPs/Leased Lines



The Network of a Modern Data Center



- ***CR*** = L3 Core Router, ***AR*** = L3 Access Router, ***S*** = L2 Switch, ***LB*** = Load Balancer, ***A*** = Rack of 20 servers (Cisco with ~ 4,000 servers)

Communication In Data Centers

- Communication in data centers are most often based on networks running the IP protocol suite
- Data centers contain a set of routers and switches that transport traffic between the servers and to the outside world
- Traffic in today's data centers:
  - 80% of the packets stay inside the data center
  - Trend is towards even more internal communication
- Typically, data centers run two kinds of applications:
  - Outward facing (serving web pages to users)
  - Internal computation (data mining and index computations–think of MapReduce and HPC)

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Communication Latency

- Propagation delay in the data center is essentially 0
  - Light goes a foot in a nanosecond
- End to end latency comes from
  - Switching latency
    - 10G to 10G:~ 2.5 usec (store&fwd); 2 usec (cut-thru)
  - Queuing latency
    - Depends on size of queues and network load
- Typical times across a quiet data center: 10-20usec

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Elasticity and Performance

- Bare data centers make it hard for applications to grow/shrink
  - VLANs can be used to isolate applications from each other
    - IP addresses are topologically determined by Access Routers
    - Reconfiguration of IPs and VLAN trunks is painful, error-prone, slow, and often manual
    - In addition, no performance isolation is provided:
      - VLANs typically provide reachability isolation only
      - One service sending/receiving too much traffic hurts all services sharing its subtree

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Power in Data Centers

- Pretty good data centers have efficiency of 1.7
  - 0.7 Watts lost for each 1W delivered to the servers
- How can we reduce power costs?
  - Create servers that use less power?
    - Conventional server uses 200 to 500W
    - Reductions have ripple effects across entire data center
    - Mostly a problem for scientists to tackle!!
  - Eliminate power redundancy?
    - Allow entire data centers to fail
  - Reduce power usage of network gear?
    - Total power consumed by switches amortizes to 10-20W per server

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Utilization In Data Centers

- Utilization of 10% to 30% is considered “good” in data centers

- Causes:
  - Uneven application fit:
    - Each server has CPU, memory, and disk: most applications exhaust one resource, stranding the others
  - Long provisioning timescales
  - Uncertainty in demand:
    - Demand for a new service can spike quickly
  - Risk management:
    - Not having spare servers to meet application demands leads to failure

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

What About?

- Maximize useful work per dollar spent – 59% of dollars are spent on servers with very low utilization (10%)
- Turn the servers into a single large resource pool and let services "breathe" : dynamically expand and contract their footprint as needed
- Two main requirements:
  - Means for rapidly and dynamically satisfying application fluctuating resource needs

Enabled by Virtualization

Enabled by Programming Models and Distributed File Systems

  - Means for servers to quickly and reliably access shared and persistent data
    - Data too large to copy during provisioning process

A Cloud is …

- A data center hardware and software that the vendors use to offer the computing resources and services

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Cloud Computing

"Cloud Computing is the transformation of IT from a product to a service"

Innovation → Product → Service

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Cloud Computing

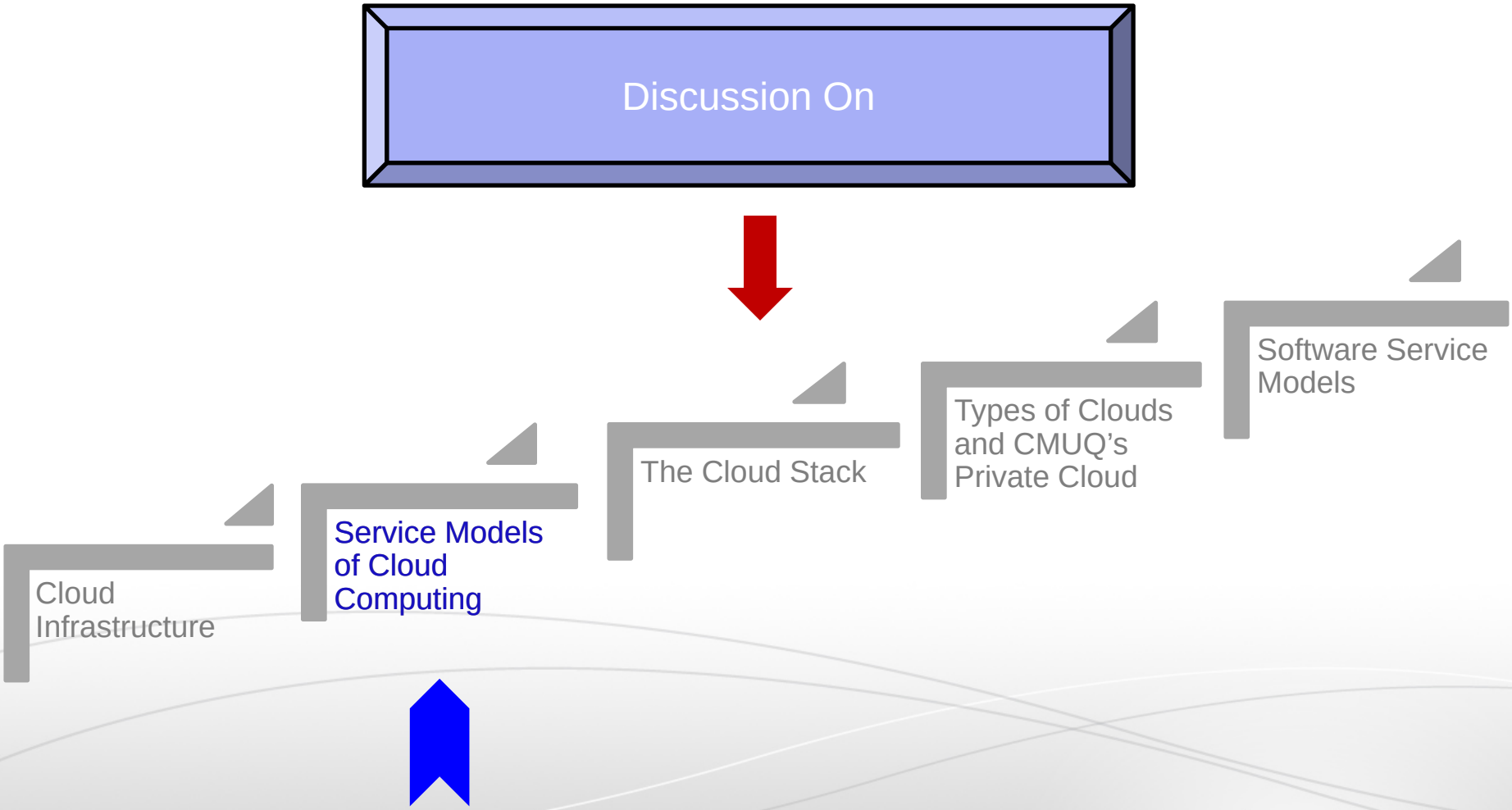
Cloud Computing is the delivery of computing as a **service** rather than a **product**,

whereby **shared resources, software**, and **information** are provided to computers and other devices,

as a **metered service** over a **network.**

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Lecture Outline



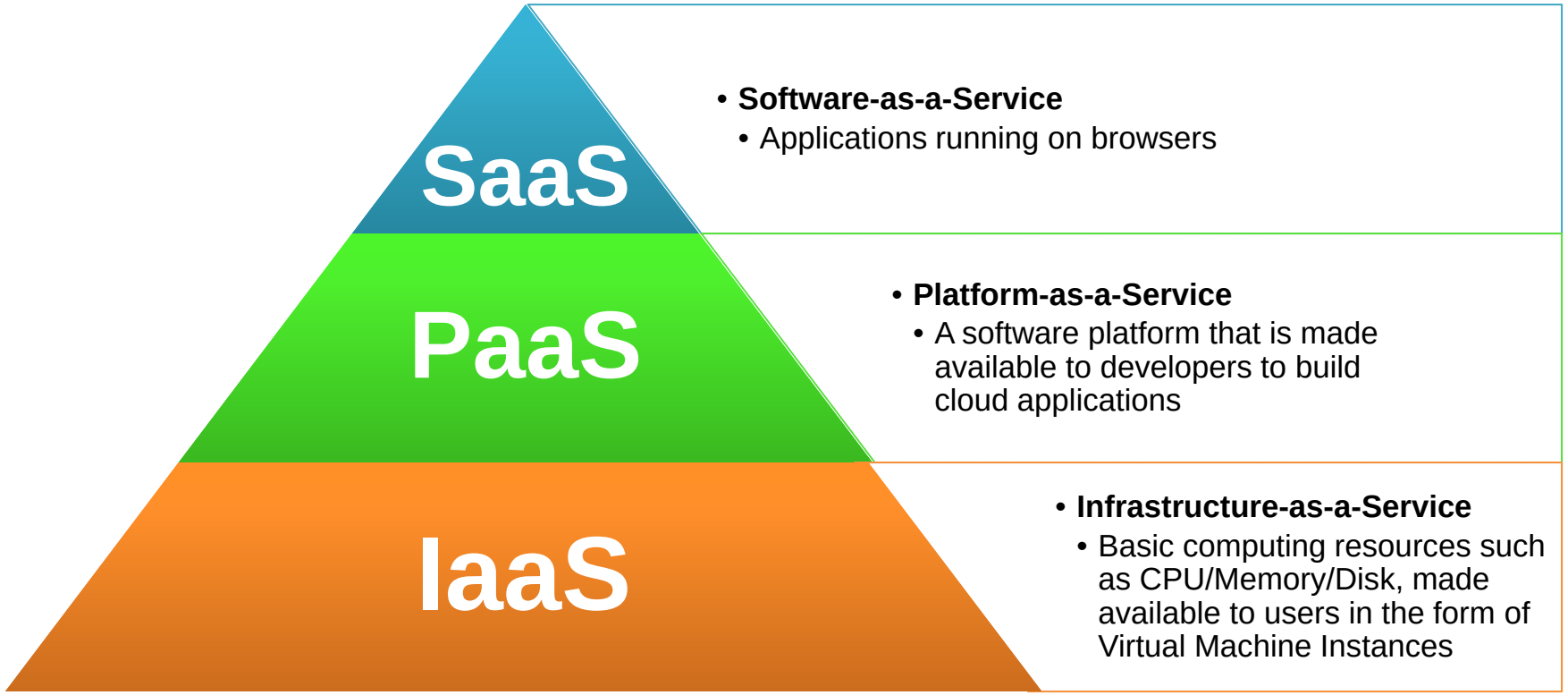
IT as a Service

- How do you offer IT as a service?
- Different users have different needs
- Consider the needs of:
  - Average End User
  - Mobile Application Developer
  - Enterprise System Architect

Let us look at some of the typical service models

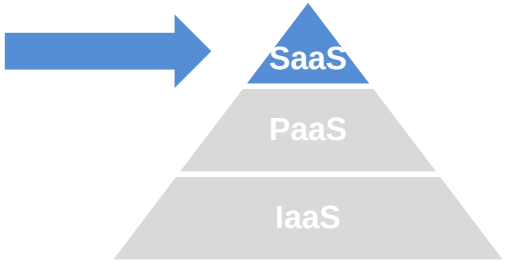
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

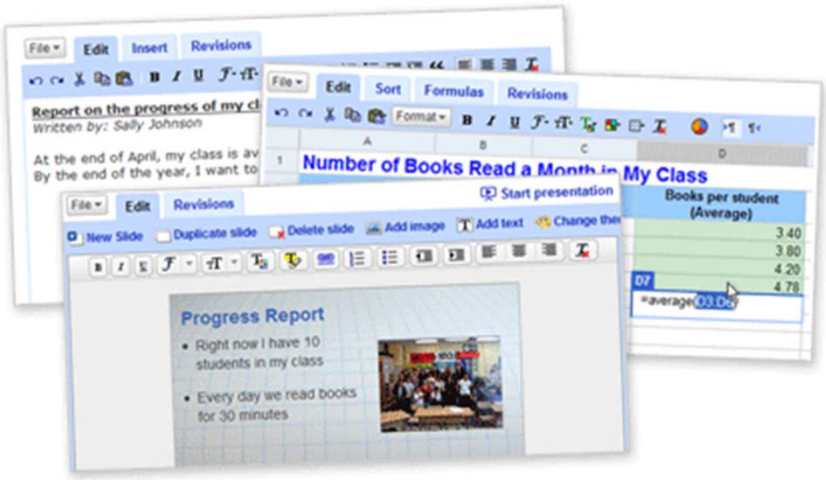
Cloud Service Models



SaaS

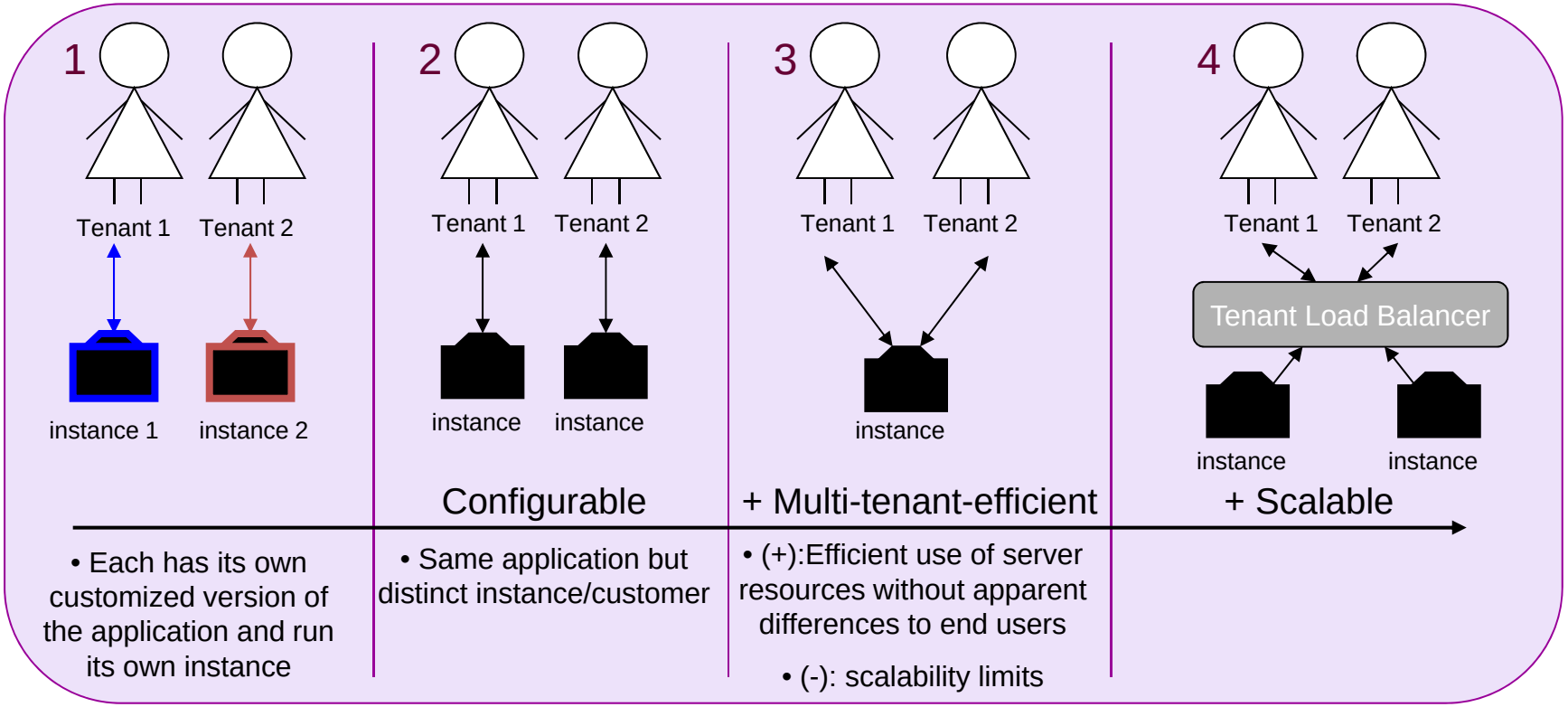
- You are most familiar with this!
- Software is delivered as a service over the Internet, eliminating the need to install and run the application on the customer's own computer
- This simplifies maintenance and support
- Examples: Gmail, YouTube, and Google Docs, among others





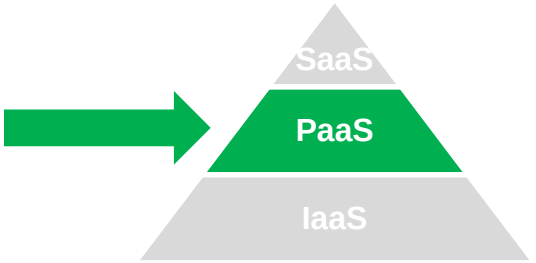
SaaS Maturity Levels

- **Distinguishing attributes:** configurability, multi-tenant efficiency, scalability

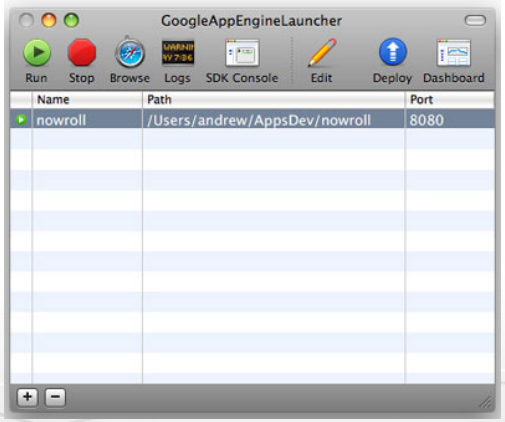


PaaS

- The Cloud provider exposes a set of tools (a platform) which allows users to create SaaS applications



- The SaaS application runs on the provider's infrastructure

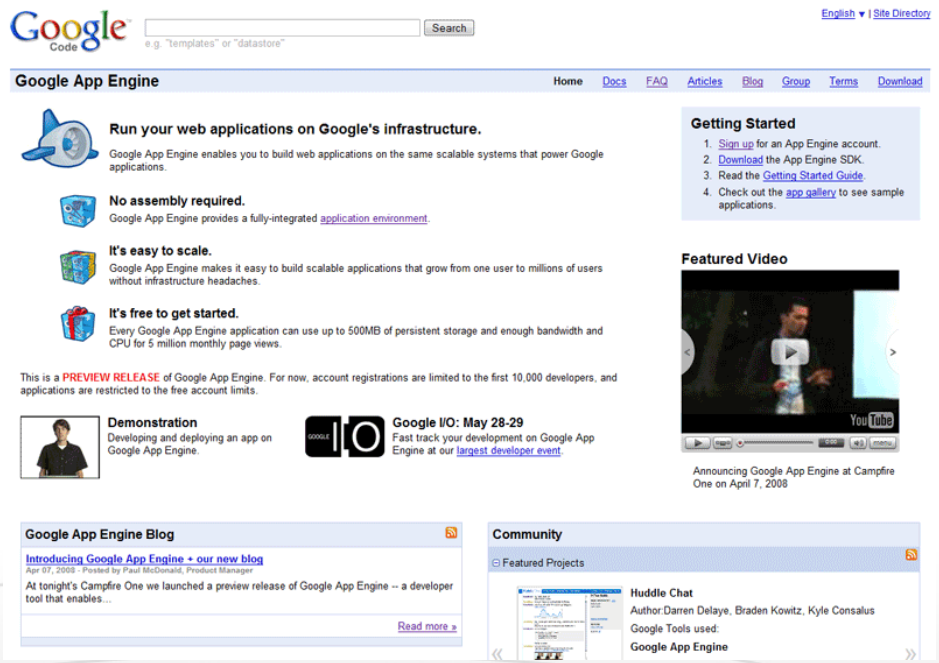


- The cloud provider manages the underlying hardware and requirements

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

PaaS Example I

- Google App Engine

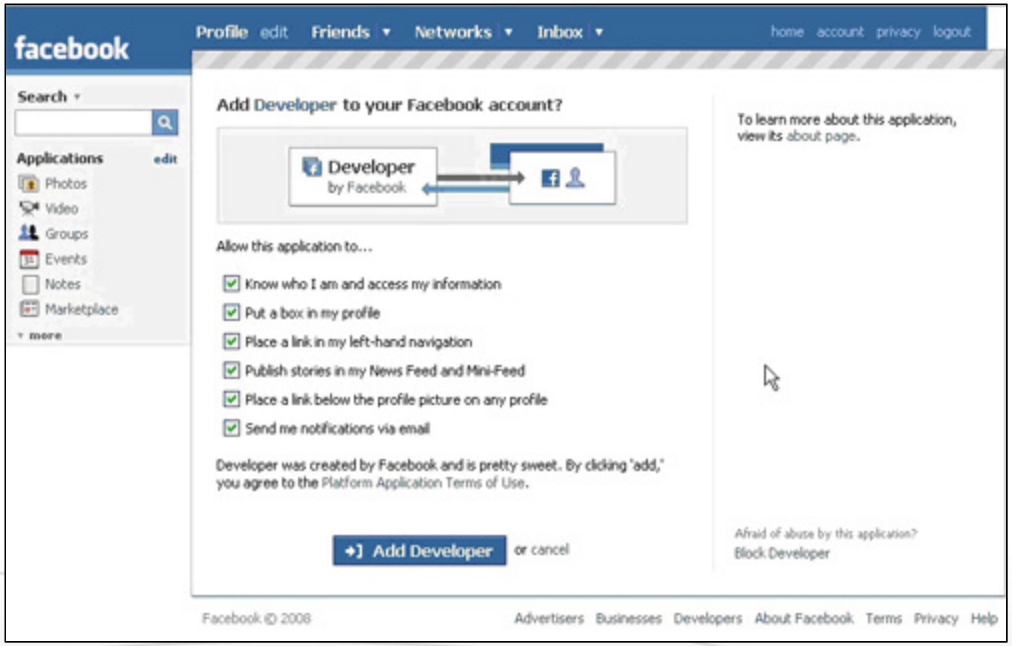


Build web applications on Google's Infrastructure

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

PaaS Example II

- The Facebook Developer Platform



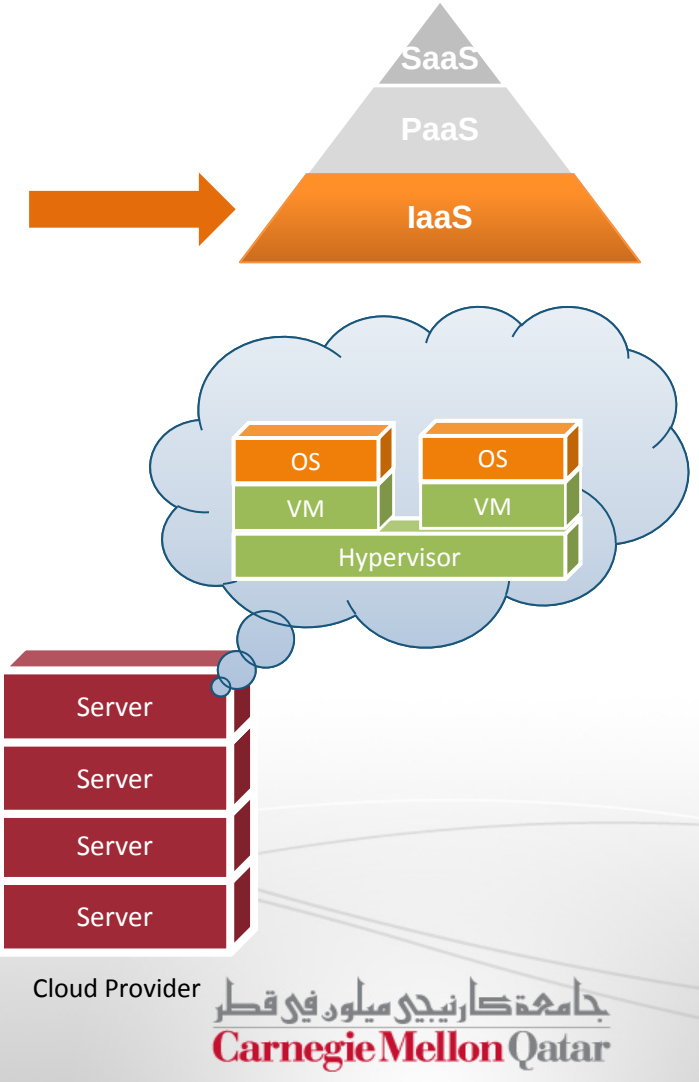
Set of APIs that allow you to create Facebook Applications

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

IaaS (1/3)

- The cloud provider leases to users Virtual Machine Instances (i.e., computer infrastructure) using the ***virtualization*** technology

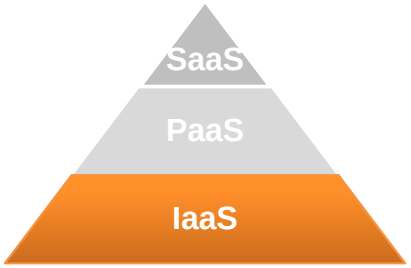
- The user has access to a standard Operating System environment and can install and configure all the layers above it

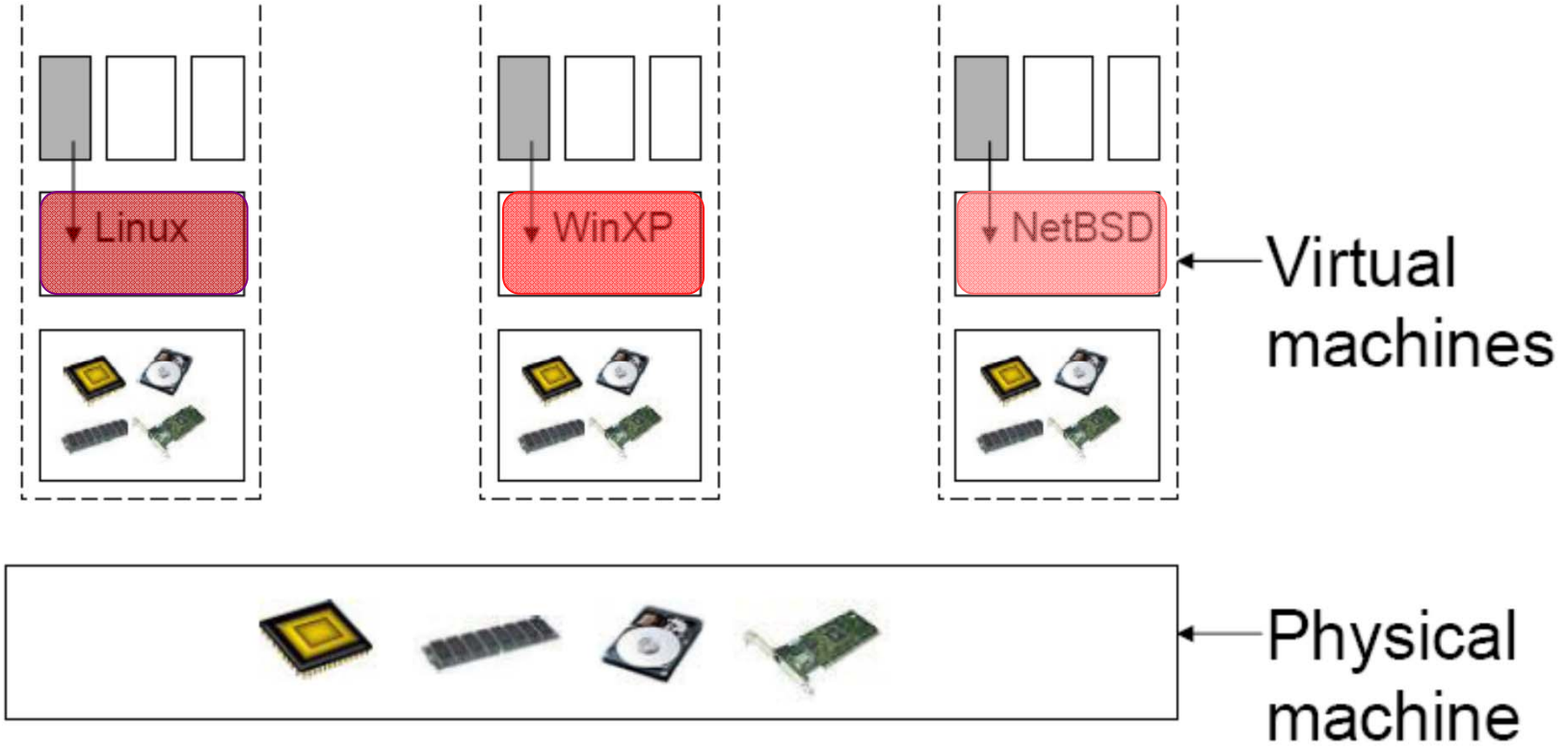


جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

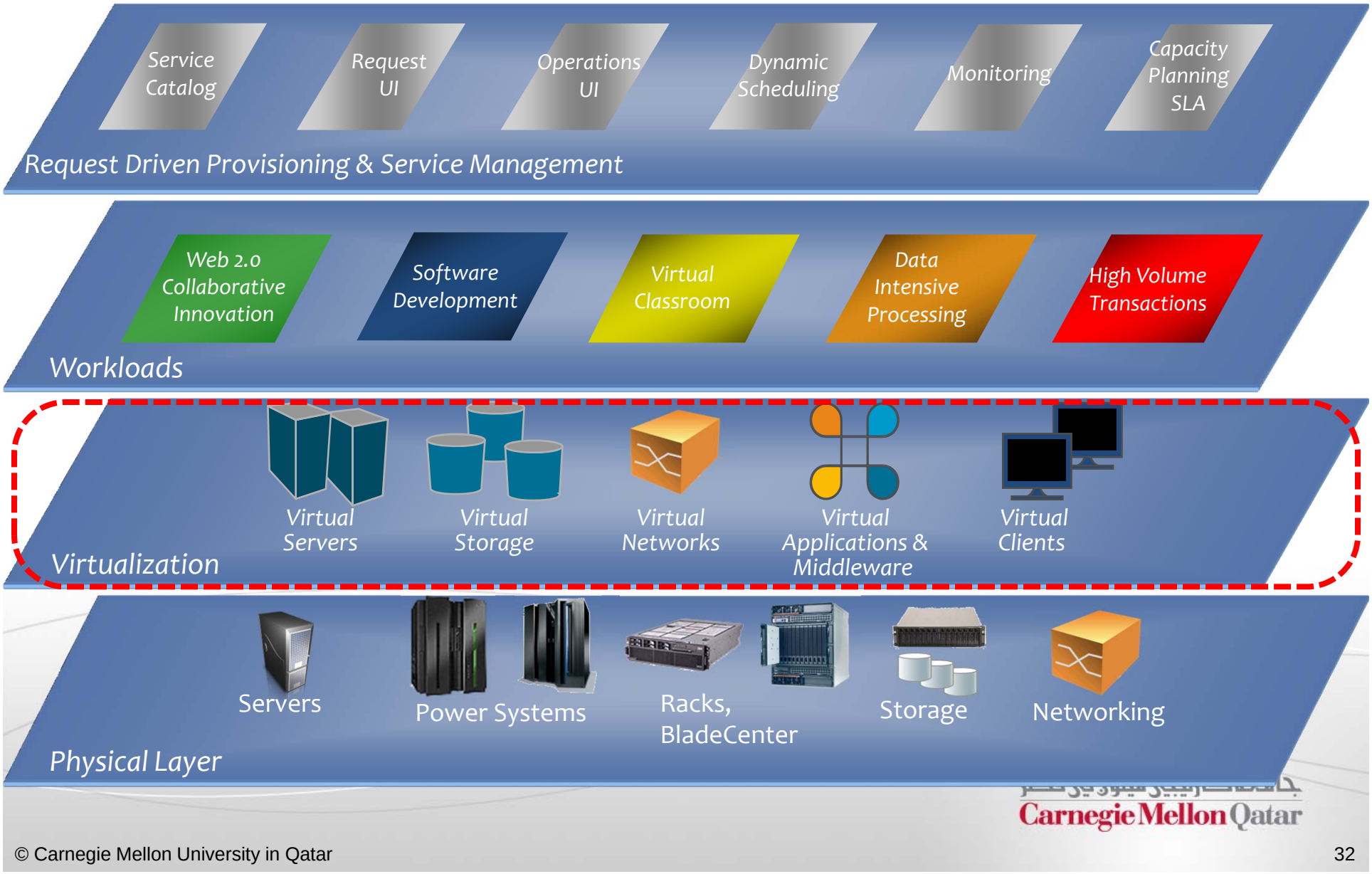
IaaS (2/3)

- The virtualization technology is a major enabler of IaaS





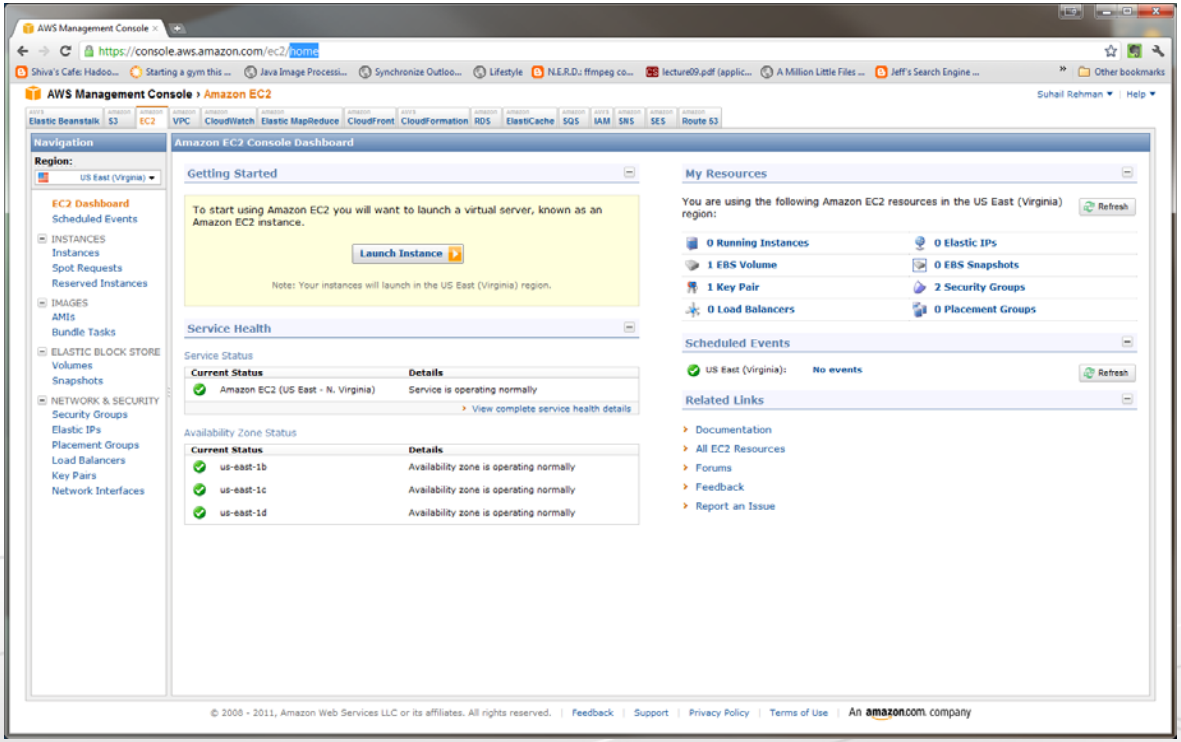
IaaS (3/3)



Carnegie Mellon Qatar

IaaS Example

- Amazon Web Service Elastic Compute Cloud (EC2)



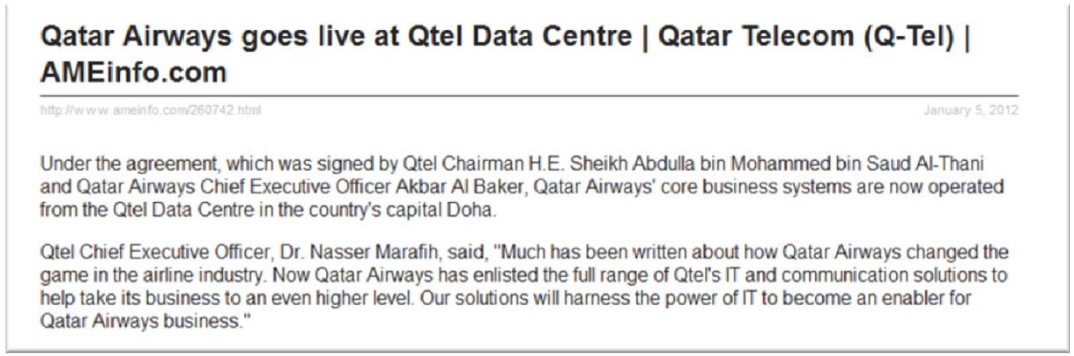
Other Service Models

- Hardware-as-a-Service
- Communication-as-a-Service
- XaaS
  - “X” as a Service

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Datacenter-as-a-Service

- Increasing Number of Servers
- Manpower, Electricity, Cooling, Security?
  - Management Nightmare
- Why not give it to someone else?



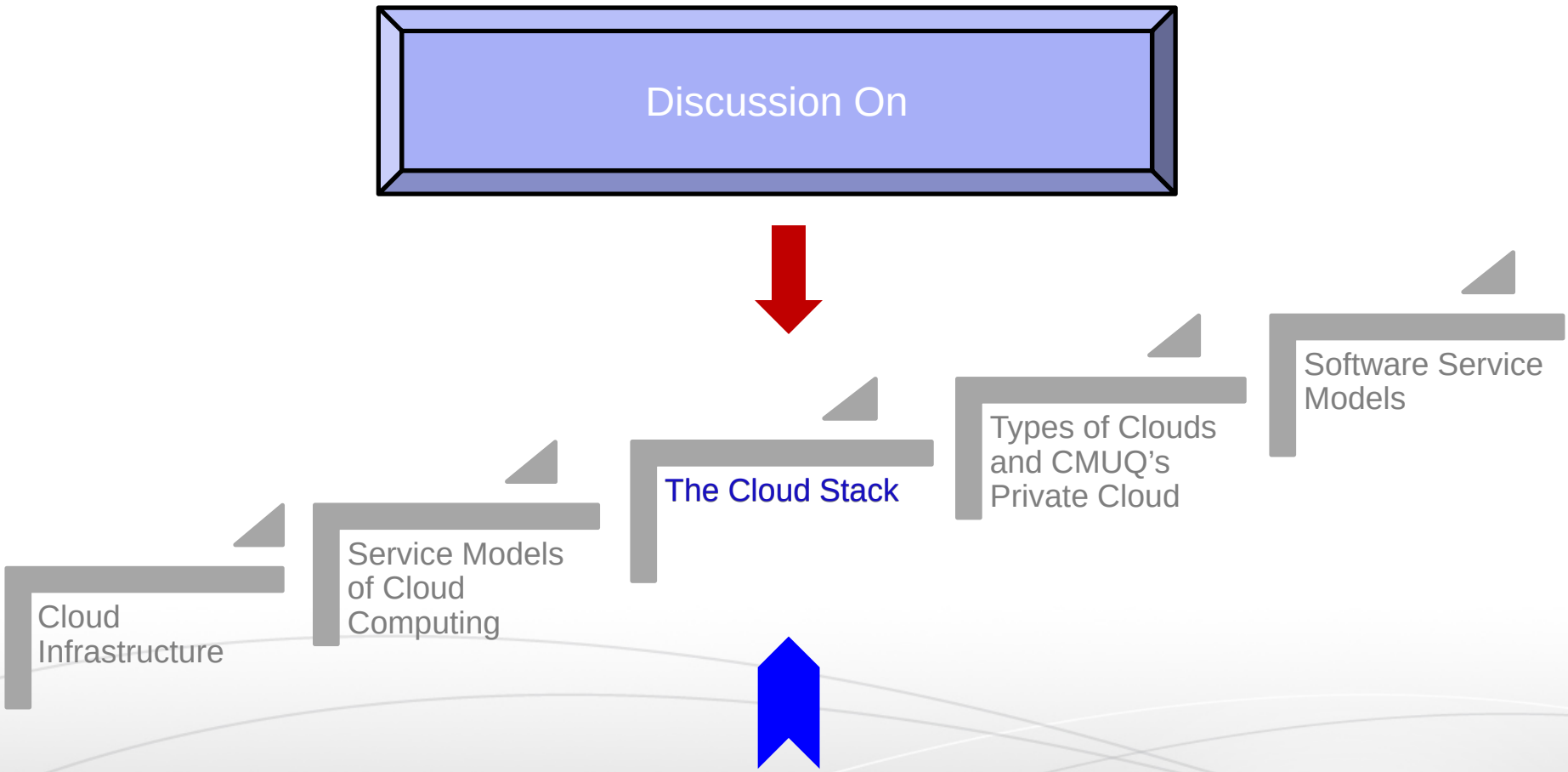
Qatar Airways goes live at Qtel Data Centre | Qatar Telecom (Q-Tel) | AMEinfo.com

http://www.ameinfo.com/260742.html January 5, 2012

Under the agreement, which was signed by Qtel Chairman H.E. Sheikh Abdulla bin Mohammed bin Saud Al-Thani and Qatar Airways Chief Executive Officer Akbar Al Baker, Qatar Airways' core business systems are now operated from the Qtel Data Centre in the country's capital Doha.

Qtel Chief Executive Officer, Dr. Nasser Marafih, said, "Much has been written about how Qatar Airways changed the game in the airline industry. Now Qatar Airways has enlisted the full range of Qtel's IT and communication solutions to help take its business to an even higher level. Our solutions will harness the power of IT to become an enabler for Qatar Airways business."

Lecture Outline

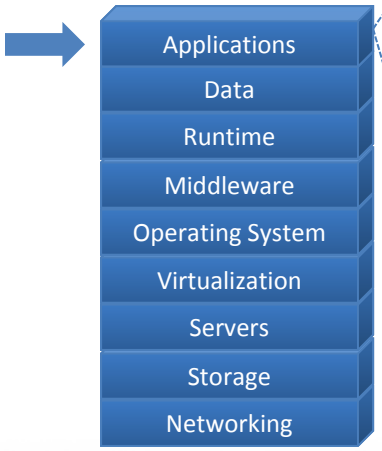


جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

The Cloud Stack

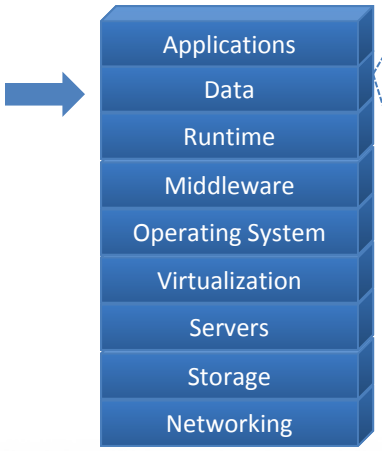
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Applications



- Cloud applications can range from Web applications to scientific computational jobs

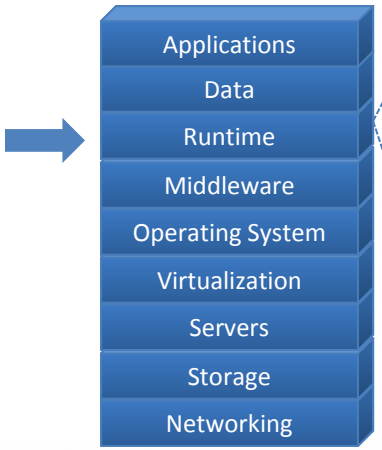
Data



- Data Management
- New generation cloud-specific databases and management systems
- E.g., Hbase, Cassandra, Hive, Pig etc.

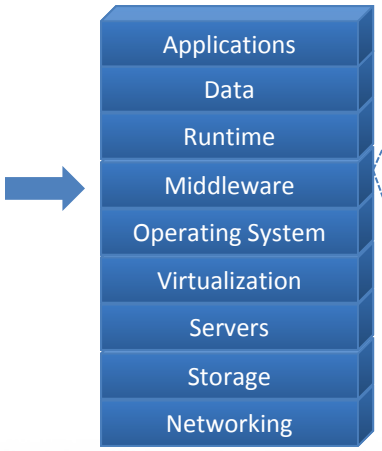
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Runtime Environment



- Runtime platforms to support cloud programming models
- E.g., MPI, MapReduce, Pregel etc.

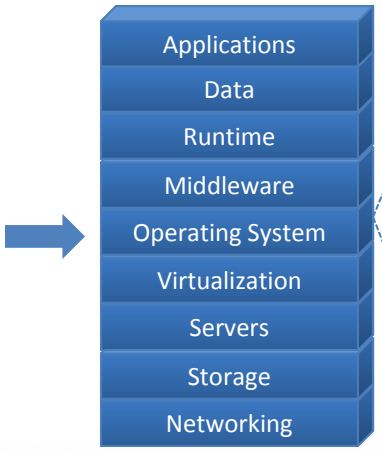
Middleware for Clouds



- Management platforms that enable:
  - Resource Management
  - Monitoring
  - Provisioning
  - Identity Management and Security

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

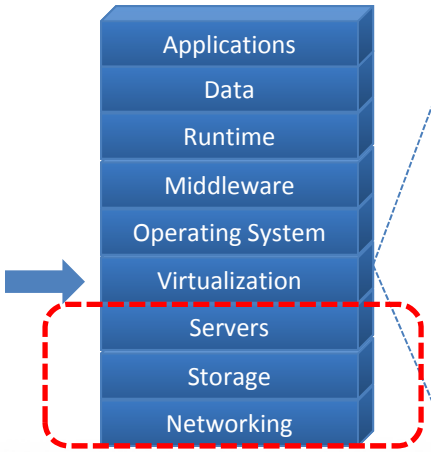
Operating Systems



- Standard Operating Systems used in Personal Computing
- Packaged with libraries and software for quick deployment and provisioning
- E.g., Amazon Machine Images (AMI) contain OS as well as required software packages as a “snapshot” for instant deployment

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

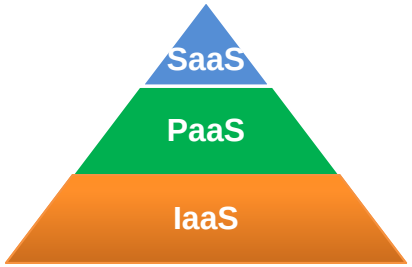
Virtualization

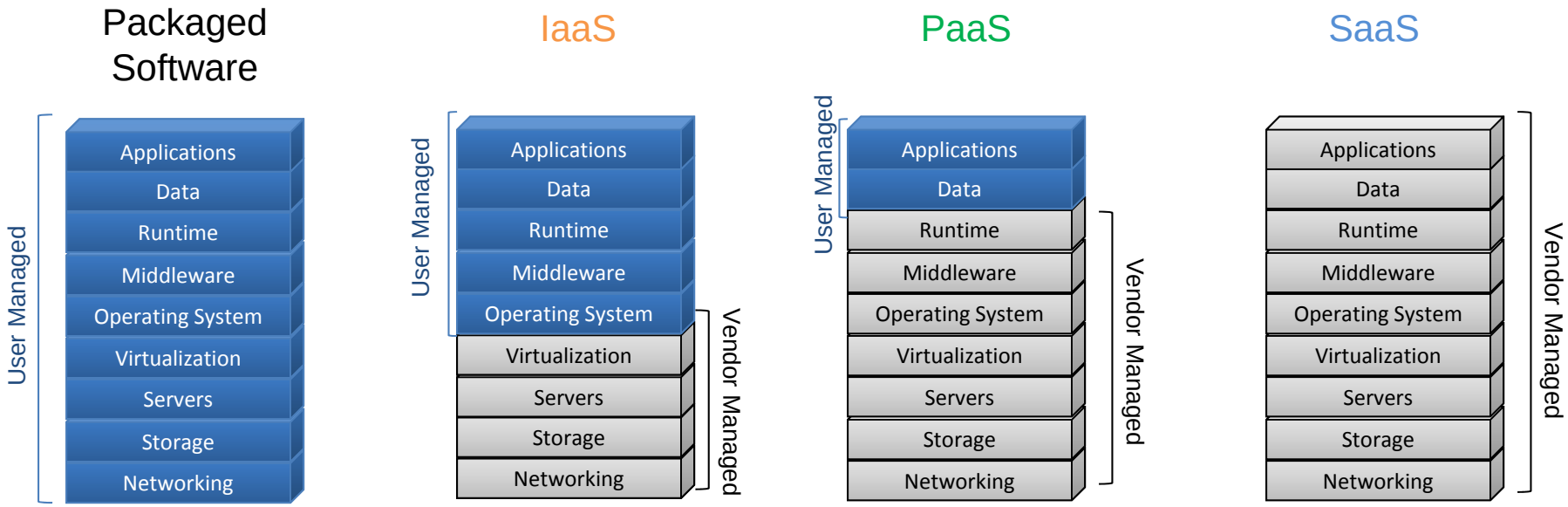


- Key Component
- Resource Virtualization
- Amazon EC2 is based on the Xen virtualization platform

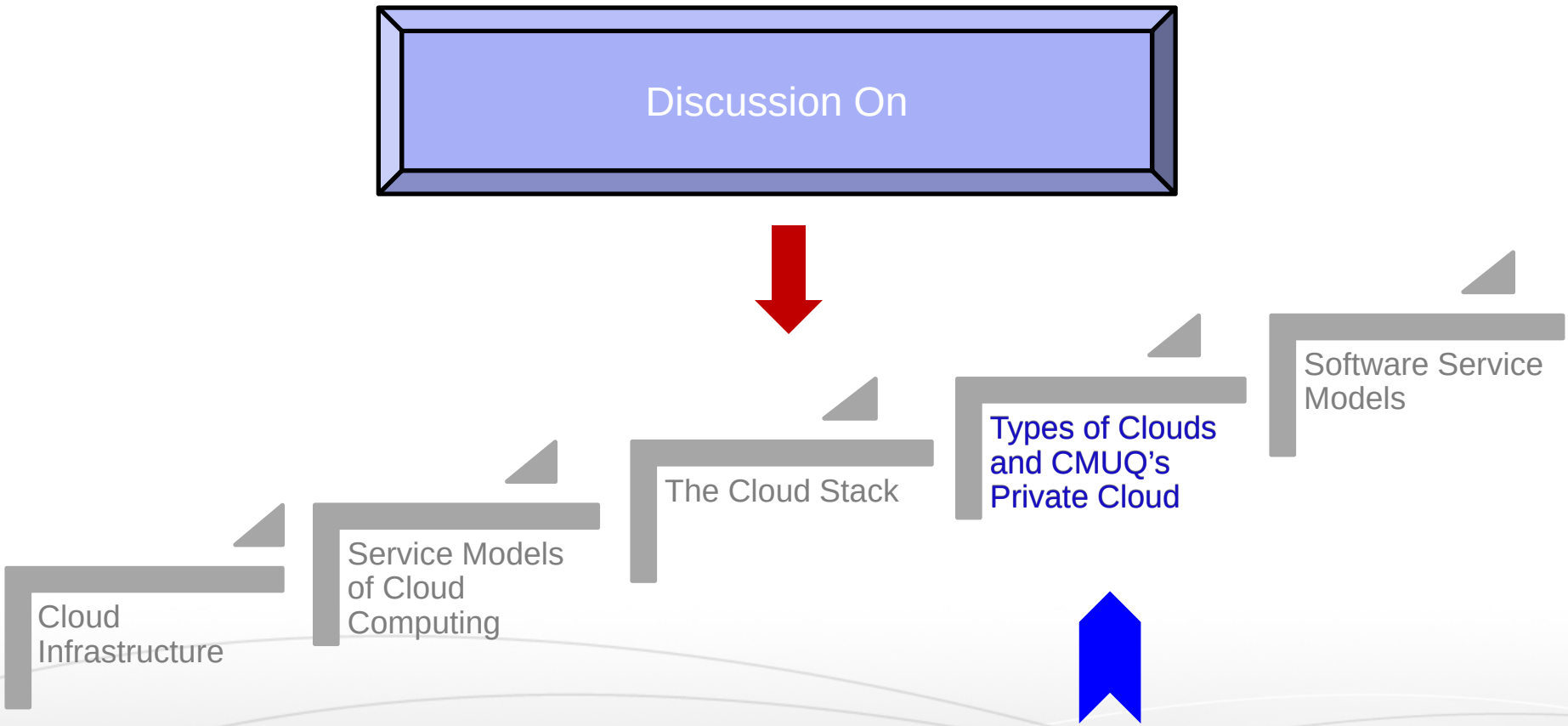
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Cloud Service Layers in the Service Levels



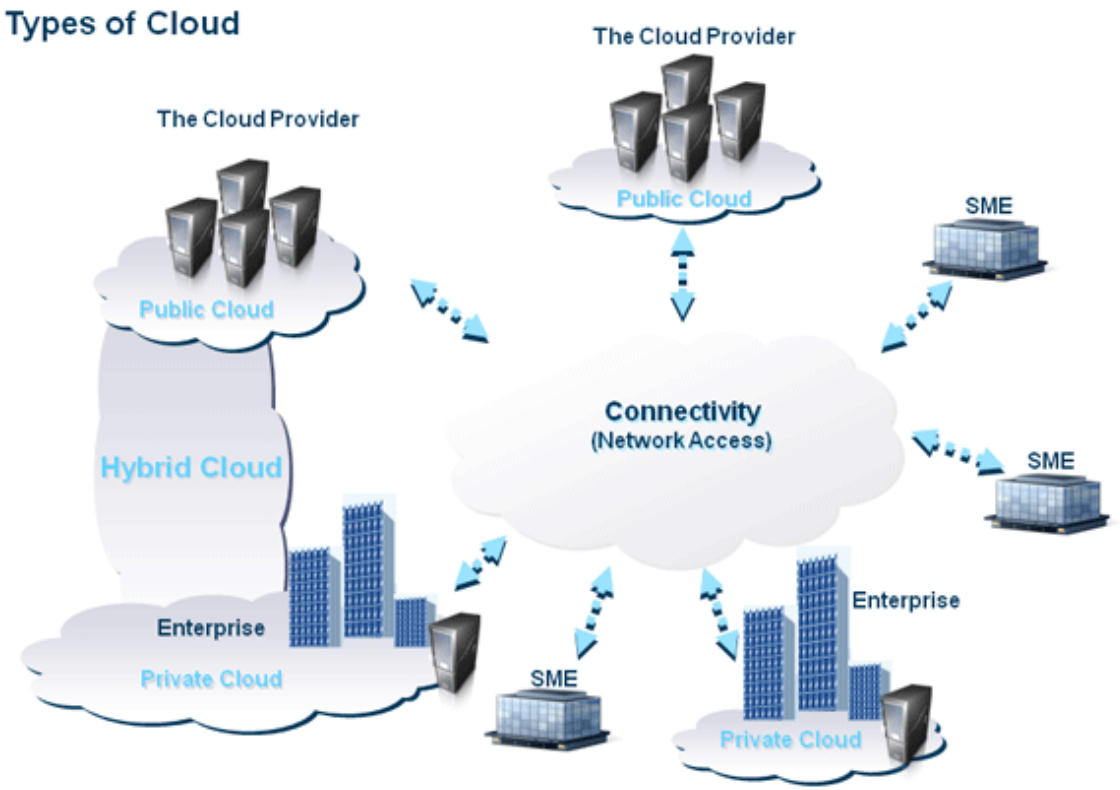


Lecture Outline



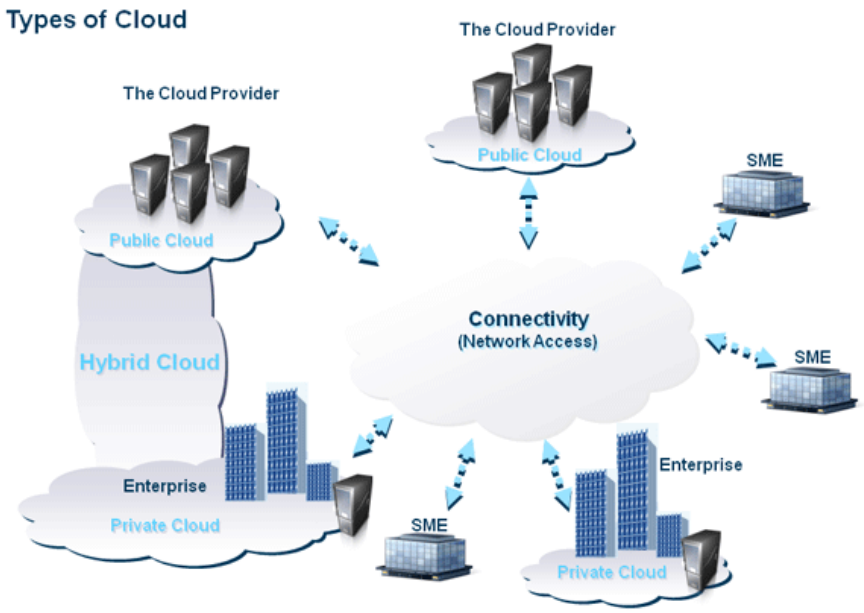
Types of Clouds (1/4)

- Public
- Private
- Hybrid



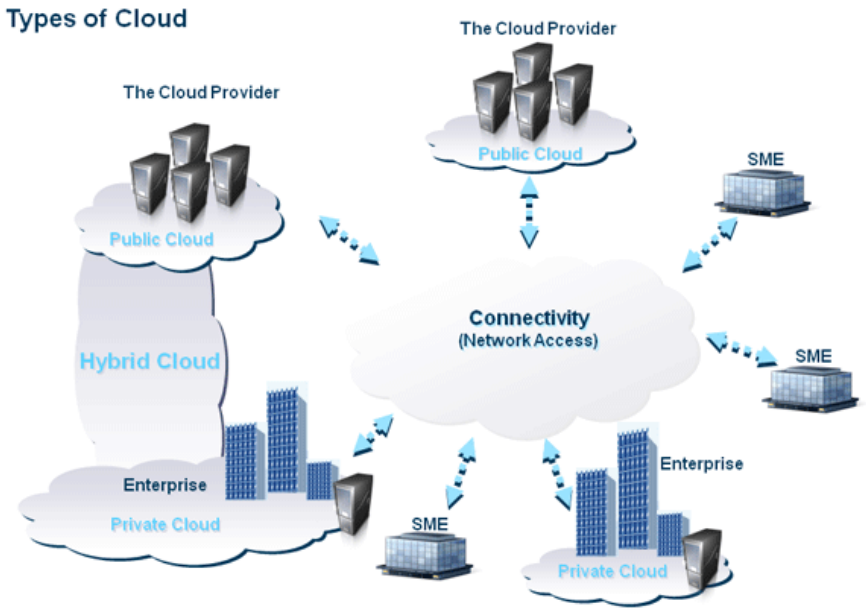
Types of Clouds (2/4)

- Public (external) cloud
  - Open market for on demand computing and IT resources
  - Concerns: Limited SLA, reliability, availability, security, trust and confidence
  - Examples: IBM, Google, Amazon, …



Types of Clouds (3/4)

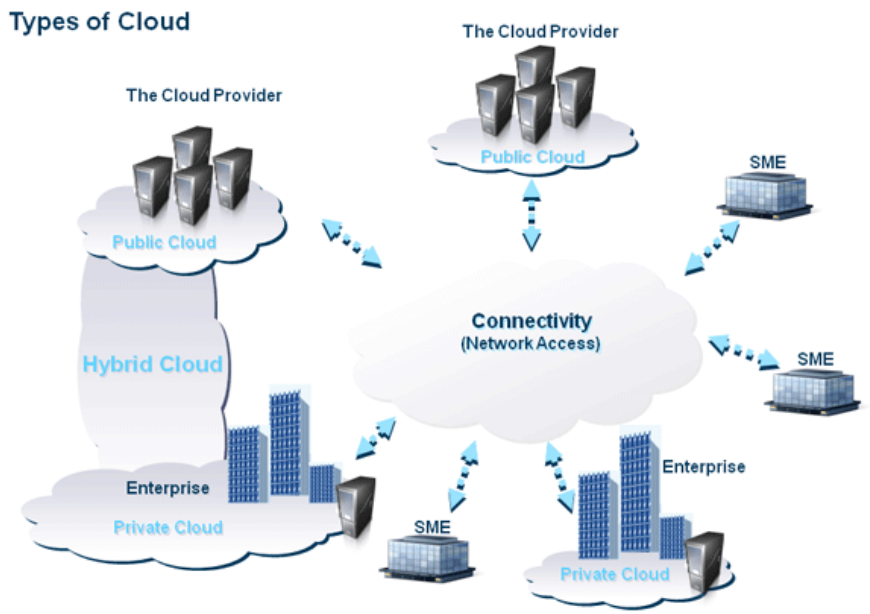
- Private (Internal) cloud
  - For enterprises/corporations with large scale IT



Types of Clouds (4/4)

- Hybrid cloud
  - Extend the private cloud(s) by connecting it to other external cloud vendors to make use of their available cloud services

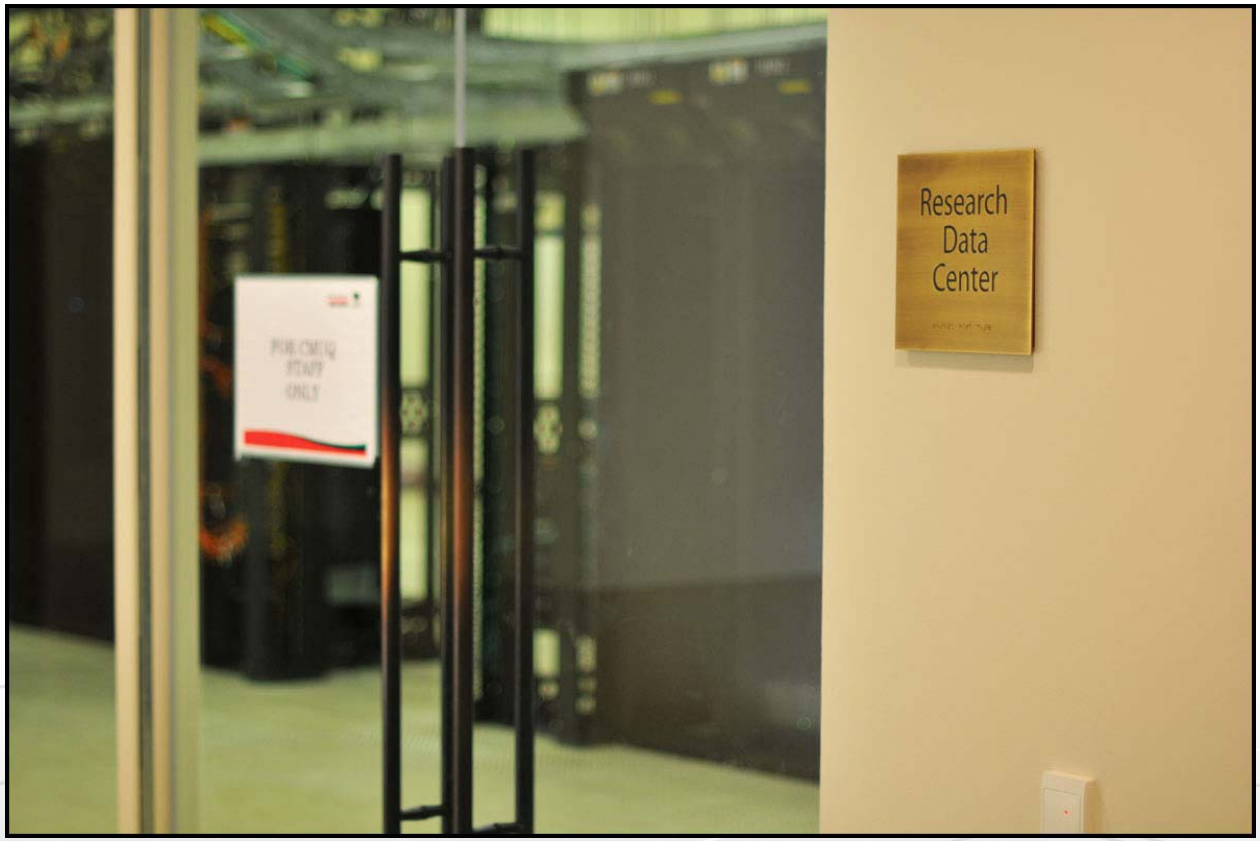
- Cloud Burst
  - Use the local cloud, and when you need more resources, burst into the public cloud



The Qloud:
CMUQ's Private Cloud

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

CMU-Q's Research Data Center



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

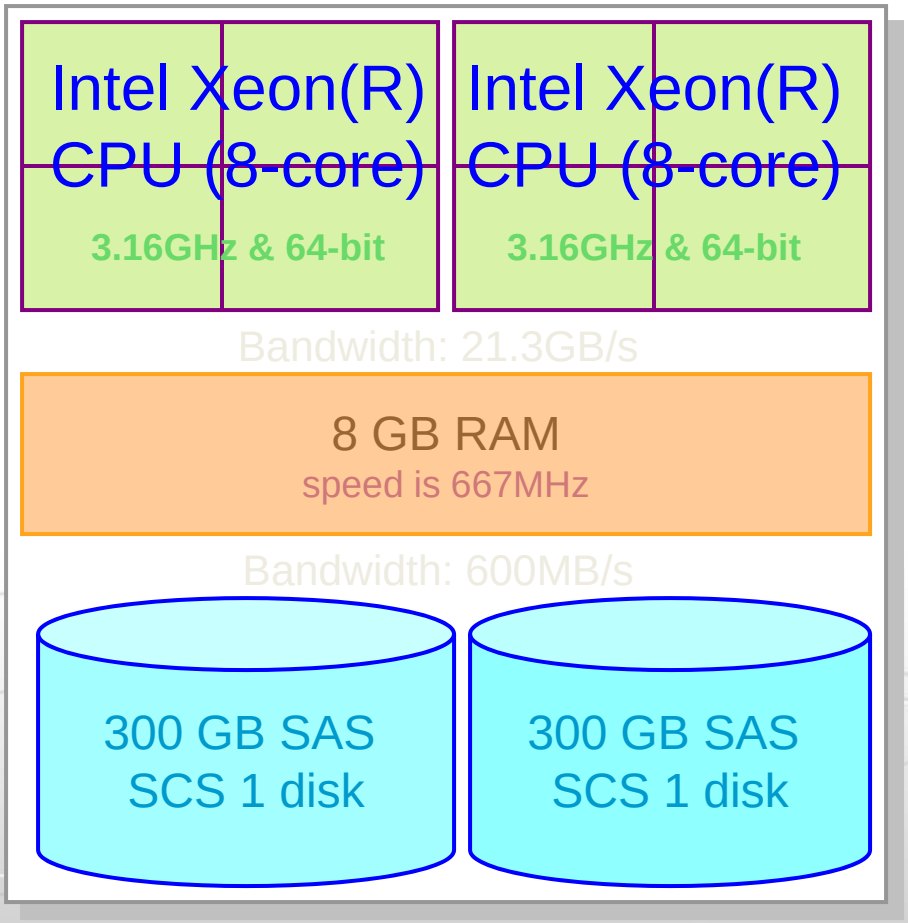
One of Qloud Hardware (1/3)

- IBM Bladecenter H
  - Advanced Management Module
  - Two Nortel Gigabit Switches
  - 112 cores
  - 7TB of storage
  - 14 Blades

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

One of Qloud Hardware (2/3)

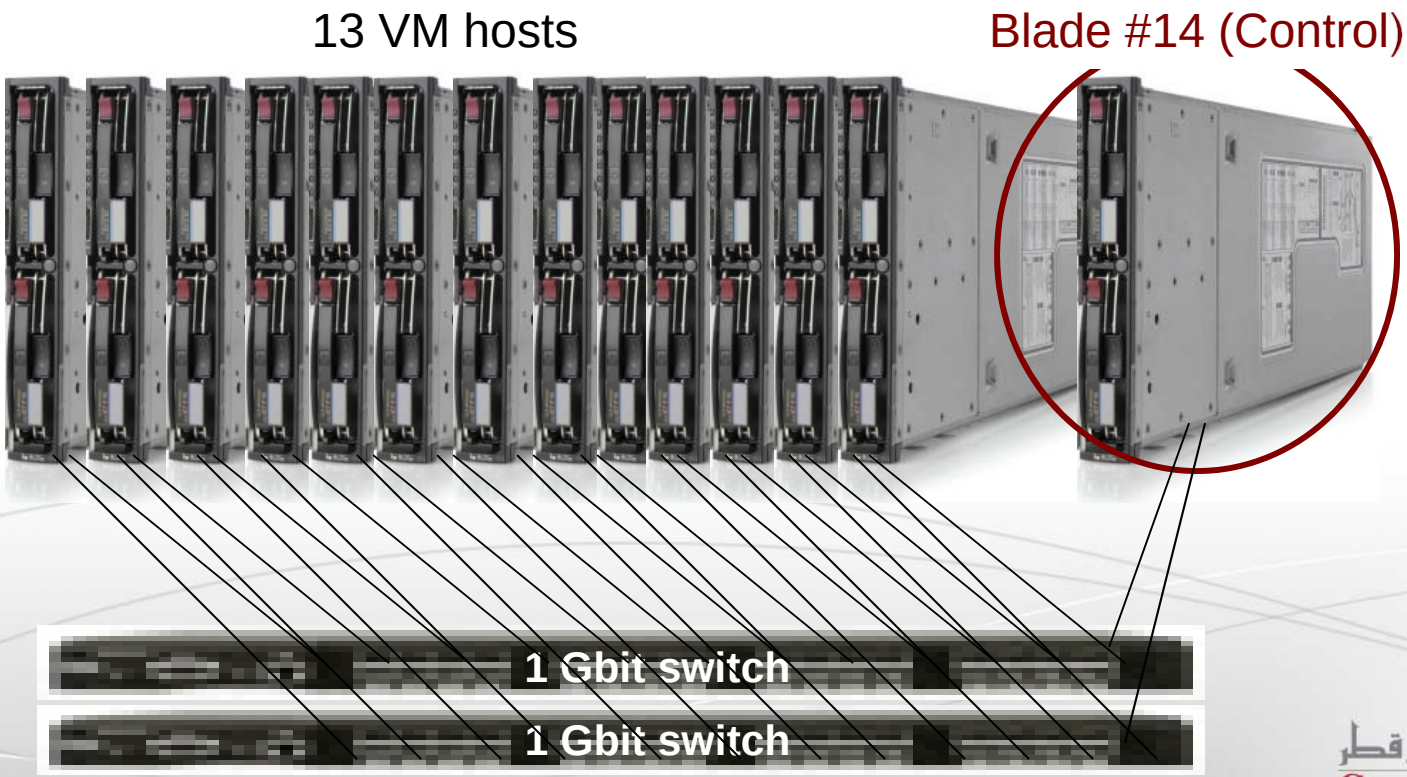
- Each Blade



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

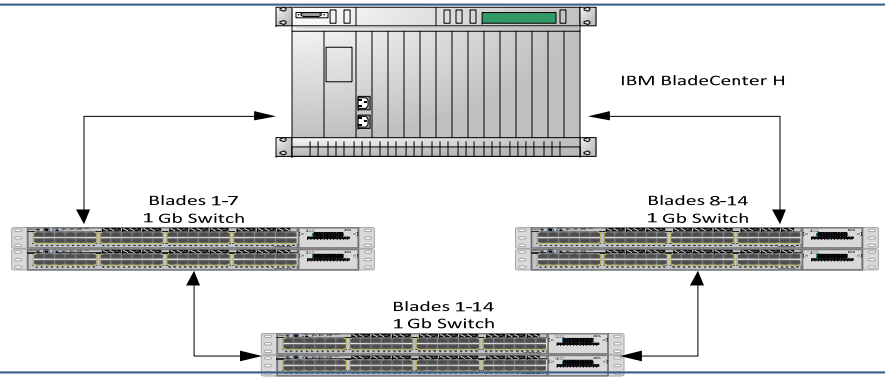
One of Qloud Hardware (3/3)

- Qloud Network
  - 2 1-Gbit switches
  - Each blade is connected to each of the switches



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Qloud Stack

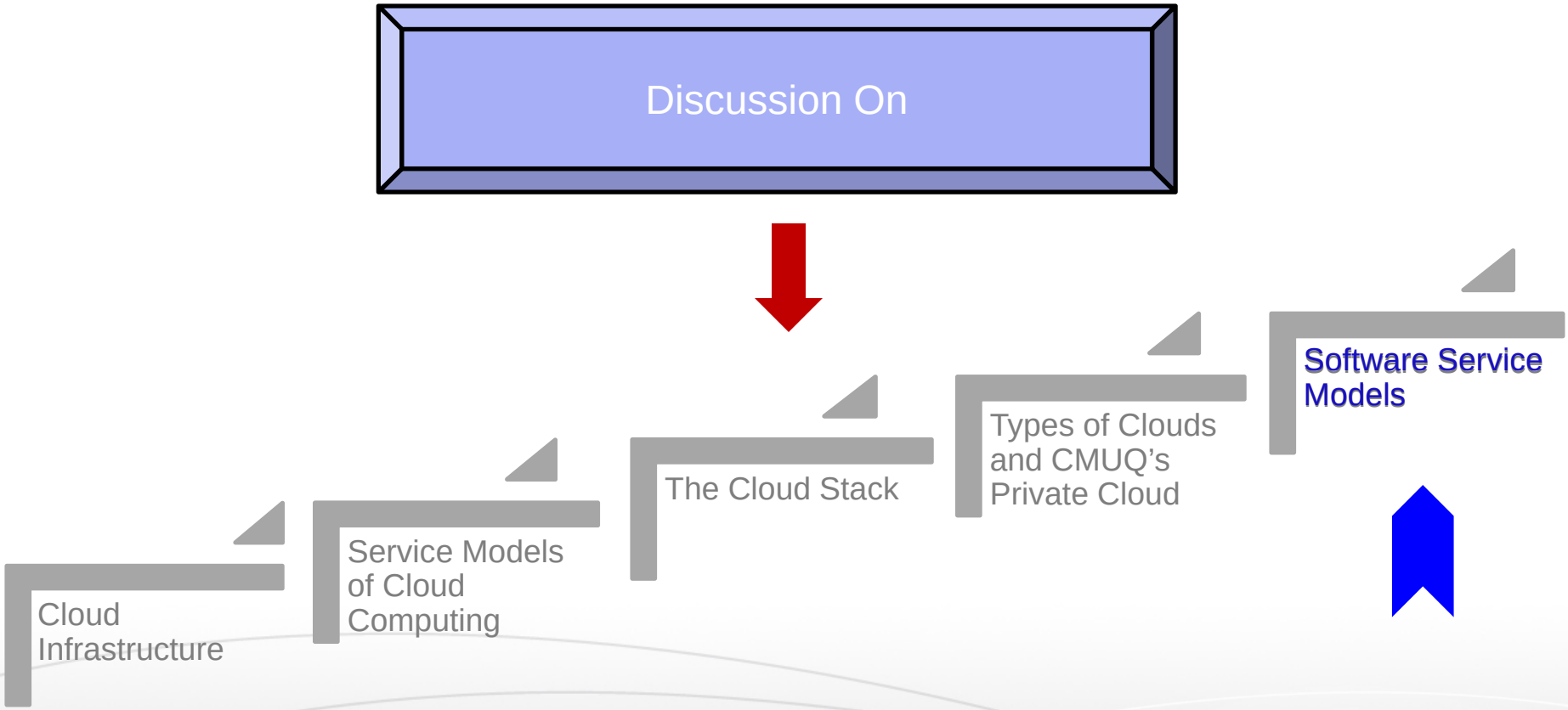
Applications	Bayesian Classification, K-Means, etc.
Data	HDFS
Runtime	Apache Hadoop 0.20.1
Middleware	Ganglia cluster monitoring system and the VMware vSphere Client
Operating System	64-Bit Fedora 13
Virtualization	Vmware vSphere 4.1/ESXi 4.1
Servers	14 IBM Quad Core (E5420) Blades
Storage	Storage/Blade = 2 x 300 GB SAS & RAM/Blade = 8 GB RAM
Networking	

A New Cloud at CMU-Q

Total Installed Capacity:
20 Servers
240 Cores, 960 GB Memory,
18 TB local storage, 20 TB
SAN Storage
VMWare vSphere 4.x
Virtualization Environment

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Lecture Outline



Economics of Cloud Computing

- Evolution of Software Service Models
- What is the Value Proposition for Cloud Computing?
- How did Cloud Computing emerge from business / industry rather than from Academia?

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Cost of Information Technology

- When you are using IT there are three primary costs associated with it:
  - Software Cost (Media + License cost/user)
  - Support Cost (Vendor Support, Updates and Patches etc.)
  - Management Cost (IT Infrastructure costs, Manpower, etc.)

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Traditional Model

- Classical Model
- Software provider develops software and charges a license fee per user for the client
- The provider may charge a support fee /user
- The management of the software is the clients responsibility
  - Up to 4x the cost of the actual software per year!
  - Infrastructure, Manpower, software maintenance
- Traditional Software – Oracle etc.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Software Service Models

	Traditional
Software Cost	$4000 /user (one-time)
Support Cost	$800 /user /year
Management Cost	Up to 4x the cost of Software!
Deployment Location	Client Side

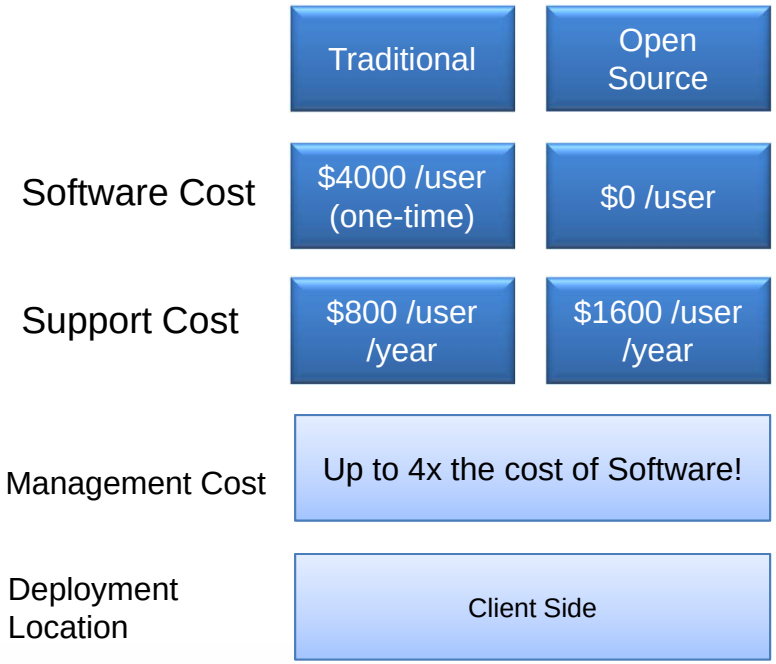
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Open Source Model

- “Free” Model
- Software provider packages Open Source Software and provides it at little or no cost to the client
- The provider makes money on support – charges a higher fee than traditional model
- The cost of Managing the software remains the same as Traditional Model
  - Up to 4x the cost of the actual software per year!
  - Infrastructure, Manpower, software maintenance

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Software Service Models

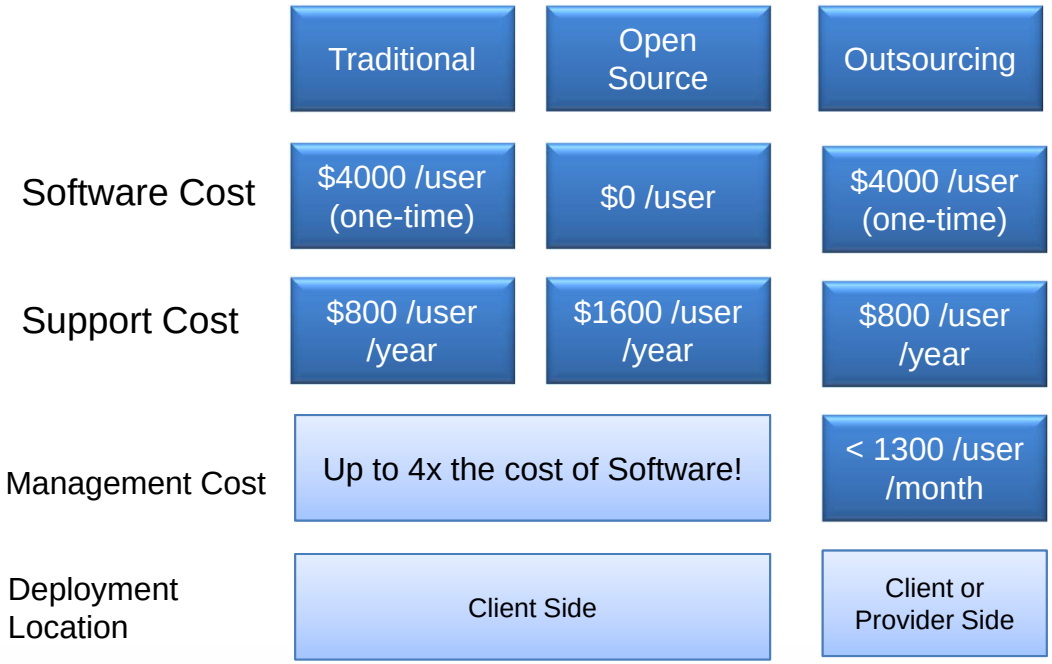


Outsourcing Model

- Primary cost of Software Management is in Manpower
- Why not delegate the management of software to a country with cheaper labor costs
  - India, China etc.
- Outsource the management of software for a flat fee – keep IT management costs under control

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Software Service Models

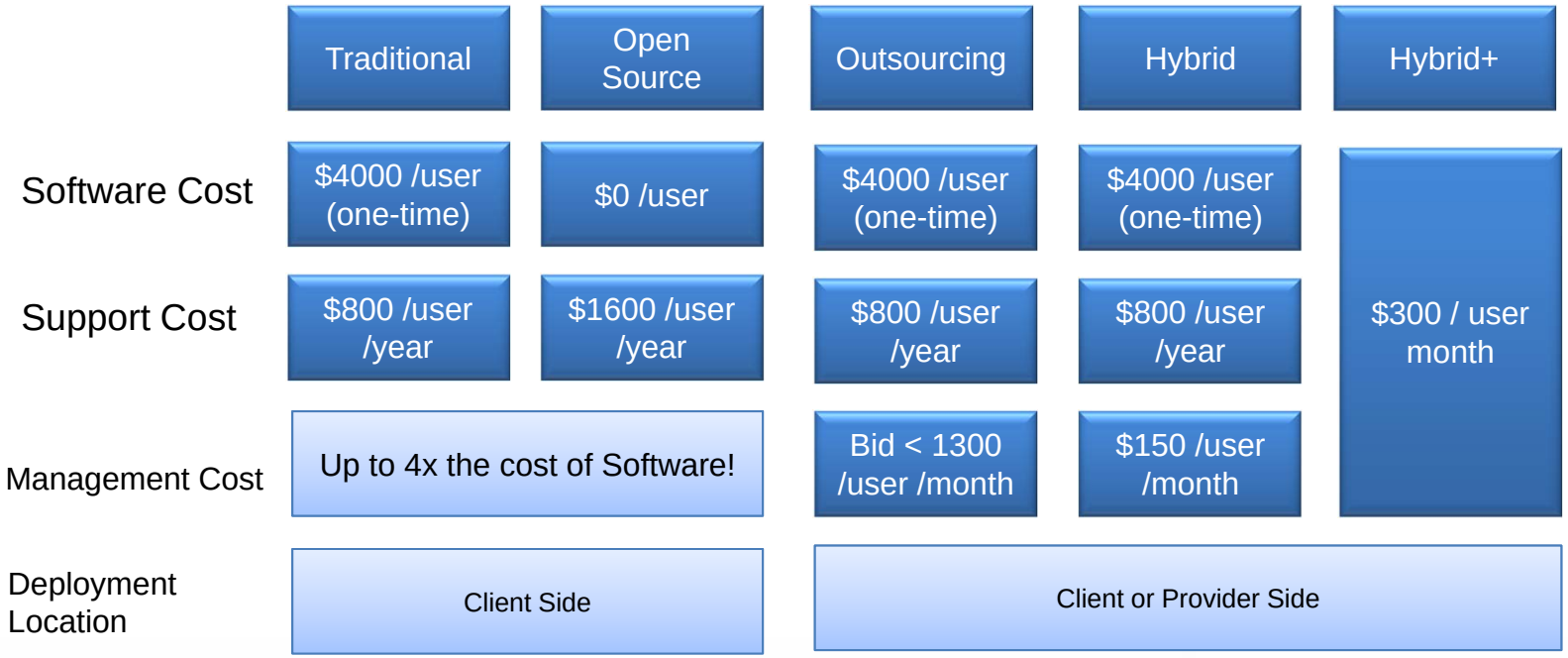


Hybrid and Hybrid+ Model

- Business Software Requirements do not change often.
  - ERP/Financials/CRM etc.
- Why reinvent the wheel?
- Standardize, Specialize and Repeat
  - Create a flexible version of the Software that can be quickly configured and deployed.
  - Automate support through remote access.
- Sell easy to deploy software to many clients.
  - Decrease the Margin
  - Increase the Customers
- Hybrid+ is more advanced – charge a flat monthly fee for the software, support and management

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Software Service Models



Software as a Service Cloud Computing

- Develop Web Application
- Offer to customers over Internet
- No deployment costs
- Amortize Management and Support costs over many clients

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Software Service Models