PMCA506L: Cloud Computing

Module 2 : Cloud Infrastructure

Courtesy : *Ming Lian , Dogules E Comer & Other Sources of Internet*

Elastic Computing

- Cloud computing centers on the ability of a customer to lease servers and only pay for the number of servers they need.
- A customer can choose to lease a few servers or many.
- More important, a customer can change the allocation dynamically, adding servers during peak times and decreasing the number of servers during times they are not needed.
- A customer could lease a full rack of servers, a half rack, or a quarter rack, or multiples of the sizes.

How Virtualized Servers Aid Providers ?

- From a cloud provider's point of view, the ability to virtualize servers provides the basis for elastic computing and makes cloud computing economically viable.

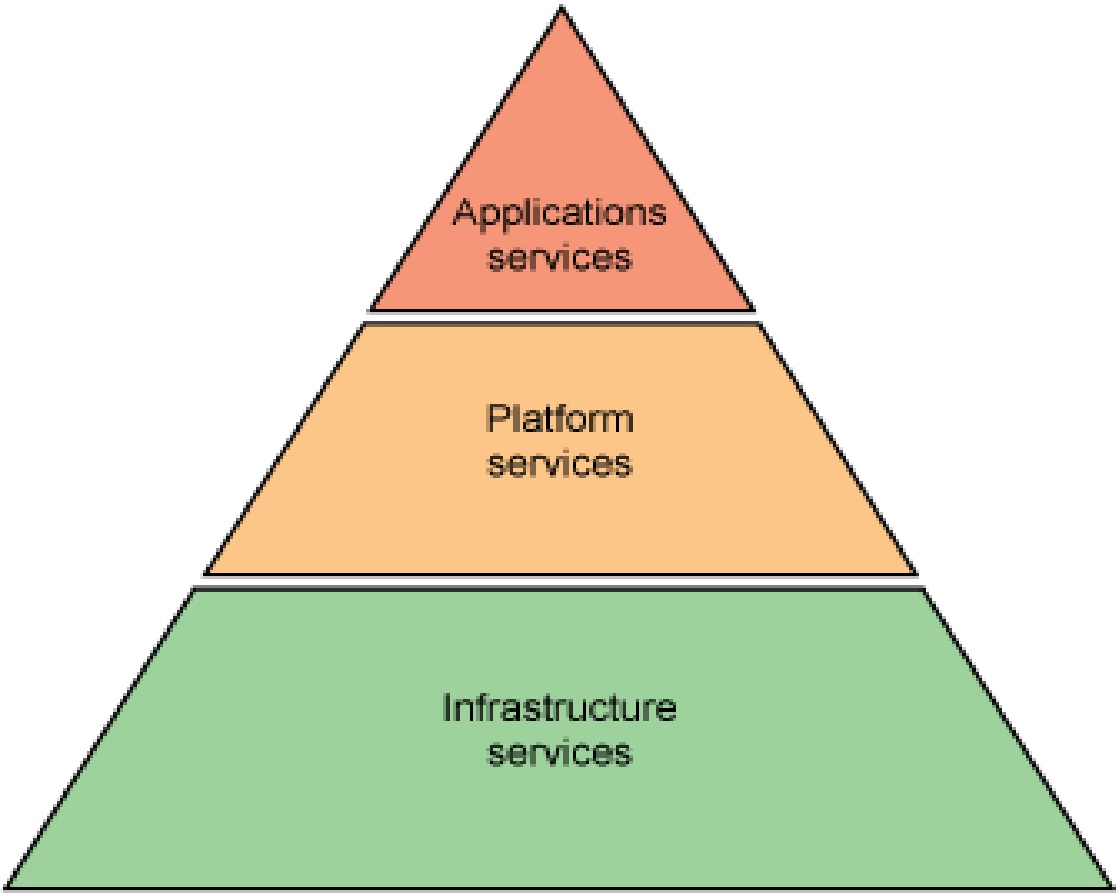
- A cloud provider only needs to use computer software to increase or decrease the number of servers a customer is leasing.

- Furthermore, the use of virtualized servers allows a provider to accommodate the changing needs of many customers.

How Virtualized Servers Help A Customer ?

- To a customer, a virtualized server appears to act like a physical server.
- A virtualized server allows apps to communicate over the Internet.
- Like a physical server, each virtualized server is assigned an Internet address.
- A virtualized server can boot a standard operating system and then allow a user to run standard apps, just as if the operating system runs on the physical server.

Business Models For Cloud Providers



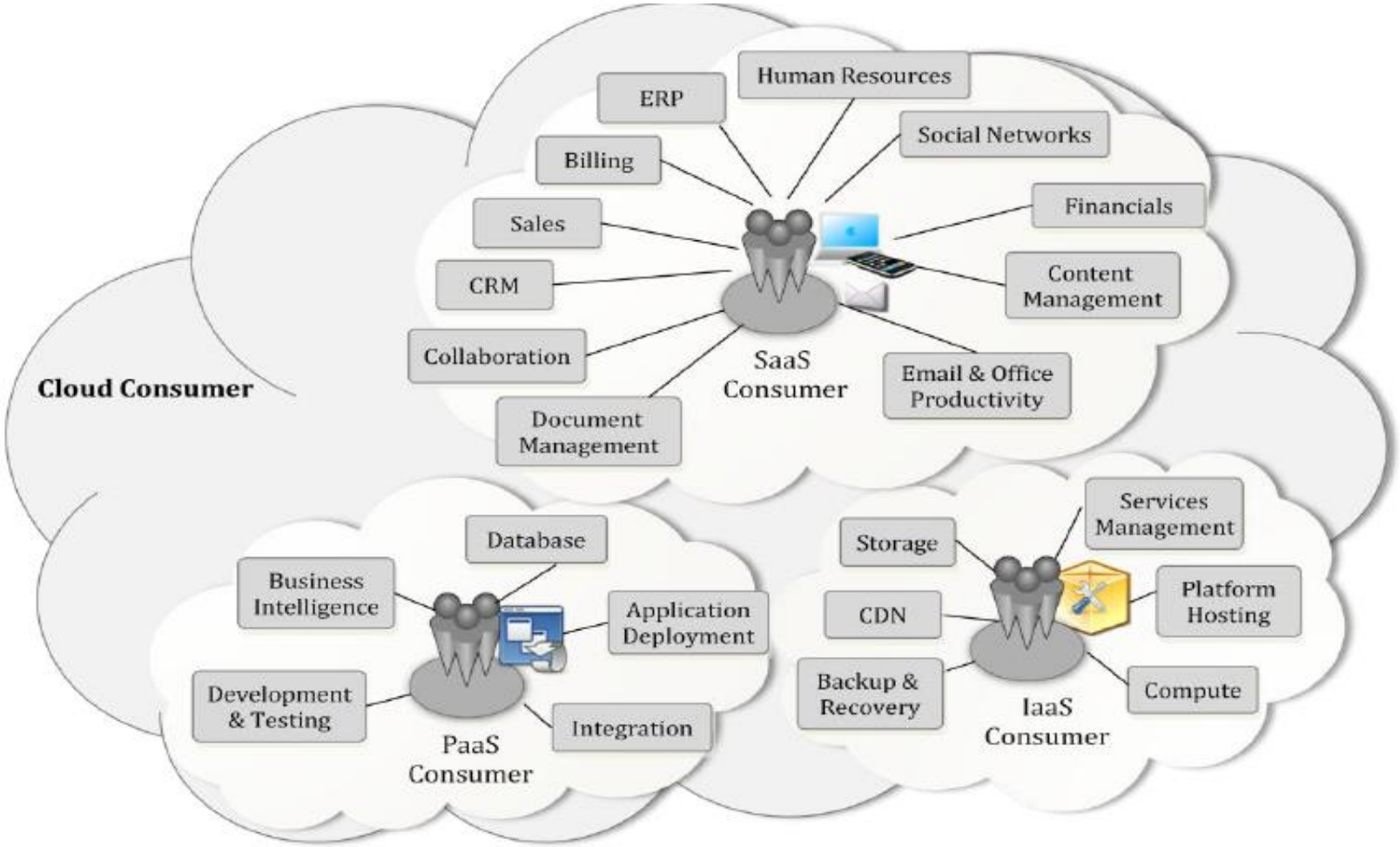
- Software as a Service (SAAS)
- Platform as a Service (PAAS)
- Infrastructure as a Service (IAAS)

Universal access

Guaranteed synchronization

High availability

Cloud Services Available



Software as a Service(SAAS)

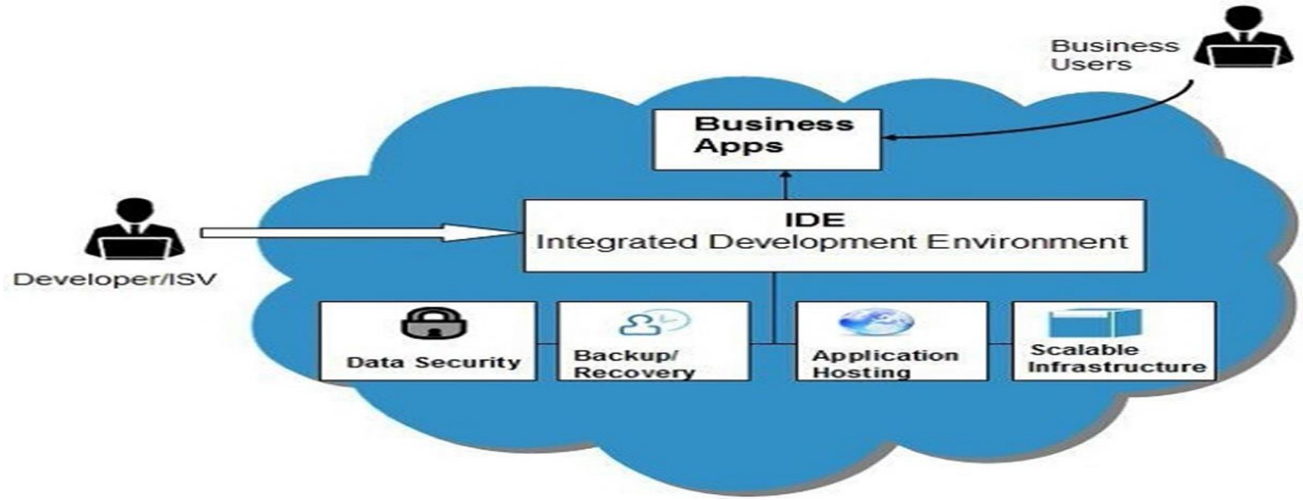
- Model-Application hosted as a service
- Customer doesn't maintain it
- Out of Customer's hand when hosting service decides to change it
- CRM ,Video Conferencing ,Accounting
- Web Analytics, IT Service Management

Benefit

- Smaller Staff
- Customization
- Security (SSL-Secured Sockets Layer)
- QOS with more bandwidth

Platform as a Service

- Application Delivery Model
- Supplies all resources to build application
- Design,Development,Testing,Deployment,Hosting
- Web service Integration, DB Integration,Security,Scalability,Storage.



Infrastructure as a Service

- Simply offers Hardware (Servers,Racks,pay for space)
- HAAS-Rent Resources
  -> Server Space
  -> Network Equipment
  -> Memory
  -> CPU Cycles
  -> Storage Space
- Billed on Utility Computing Basis
- Network(Firewalls,Routers,LB..)

Desktop as a Service(DaaS)

- System implements remote desktop access
- DaaS paints a desktop on the user's screen and allows the user to click on icons, run apps, browse files, and perform other actions exactly as if the desktop was local.
- The desktop that the user sees, the operating system that supplies the desktop, and the apps a user invokes all run on a server in the cloud instead of the user's local device.

Service Level Agreement (SLA)

- The SLA is a contract negotiated and agreed between a customer and a service provider
- Service provider is required to execute service requests from a customer within negotiated quality of service requirements for a given price
- Due to variable load, dynamically provisioning computing resources to meet an SLA and allow for an optimum resource utilization will not be an easy task

Business Model

Architectural

- Multi-tenancy
- Scalability
- Security
- Performance

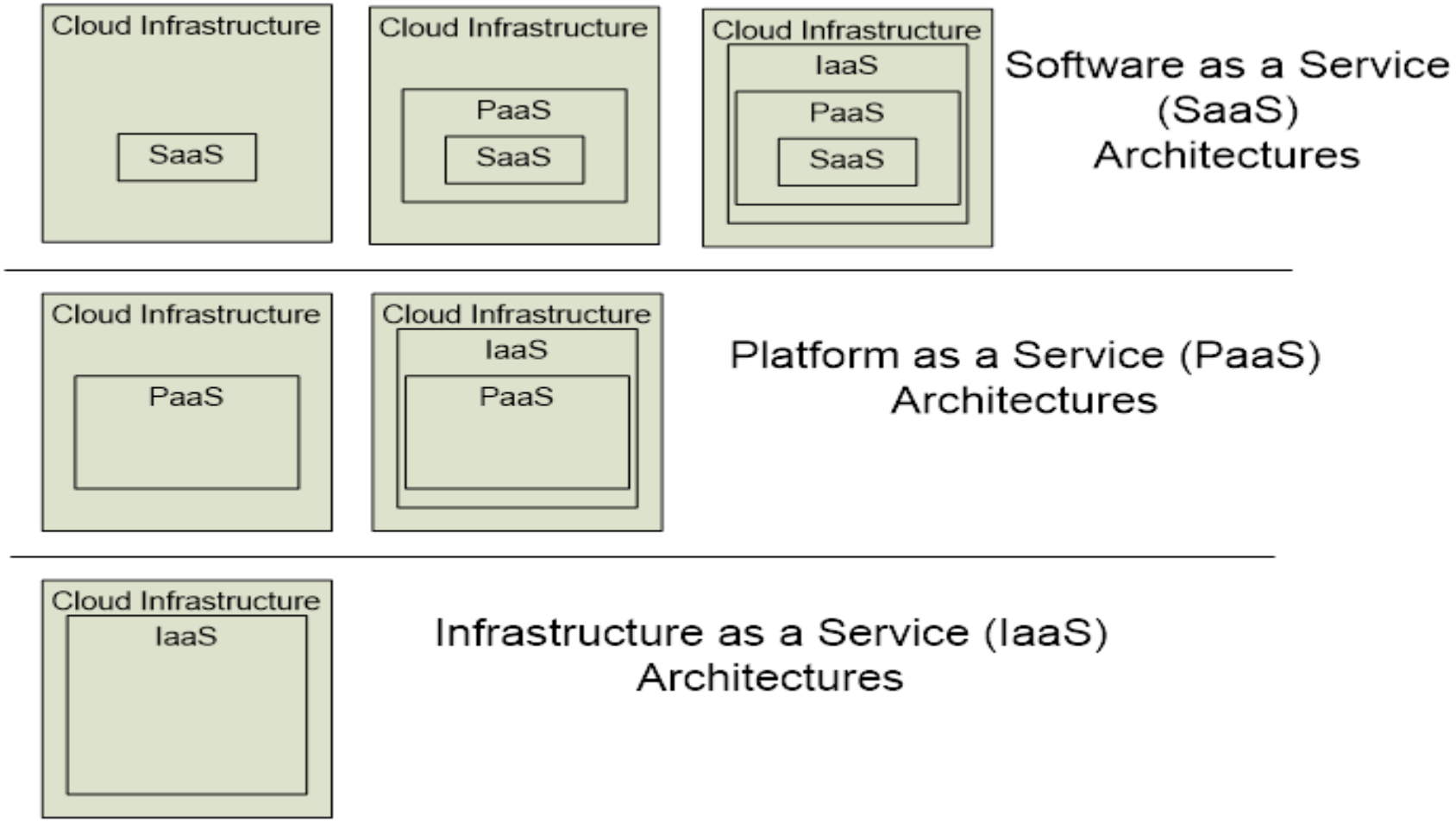
Functional

- Provisioning
- Billing
- Metering
- Monitoring

Usage-based pricing

- **Per user per month**
- **Per transaction**
- **Per GB of storage per month**

Service Model Architecture



Types of Cloud(Deployment Models)

- Public cloud
  - Sold to the public, mega-scale infrastructure
  - available to the general public

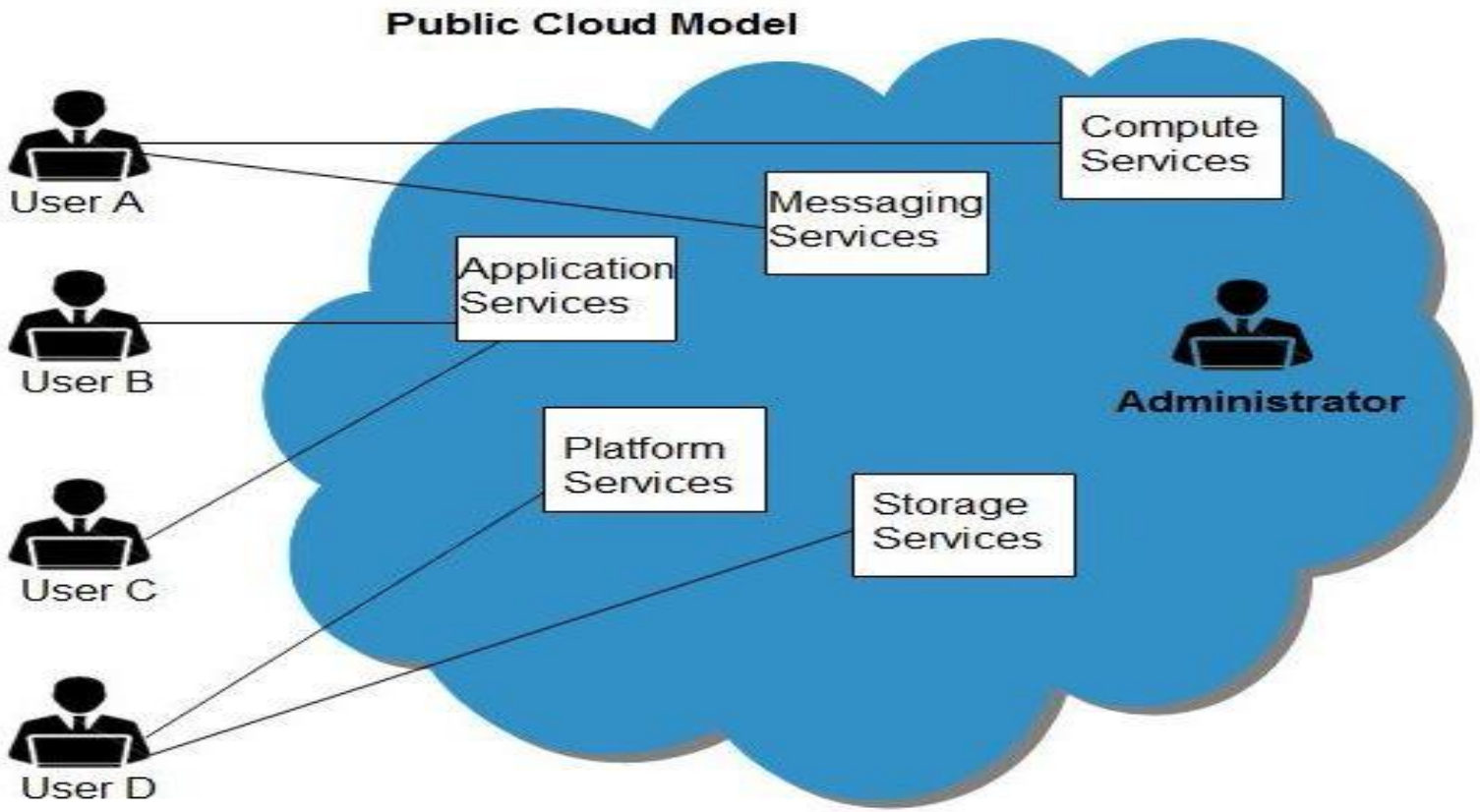
- Private cloud
  - single org only,
  - managed by the org or a 3rd party,
  - on or off premise

Types of Cloud(Deployment Models)

- Hybrid cloud
  - composition of two or more clouds
  - bound by standard or proprietary technology

- Community cloud
  - shared infrastructure for specific community
  - several orgs that have shared concerns,
  - managed by org or a 3rd party

Public Cloud

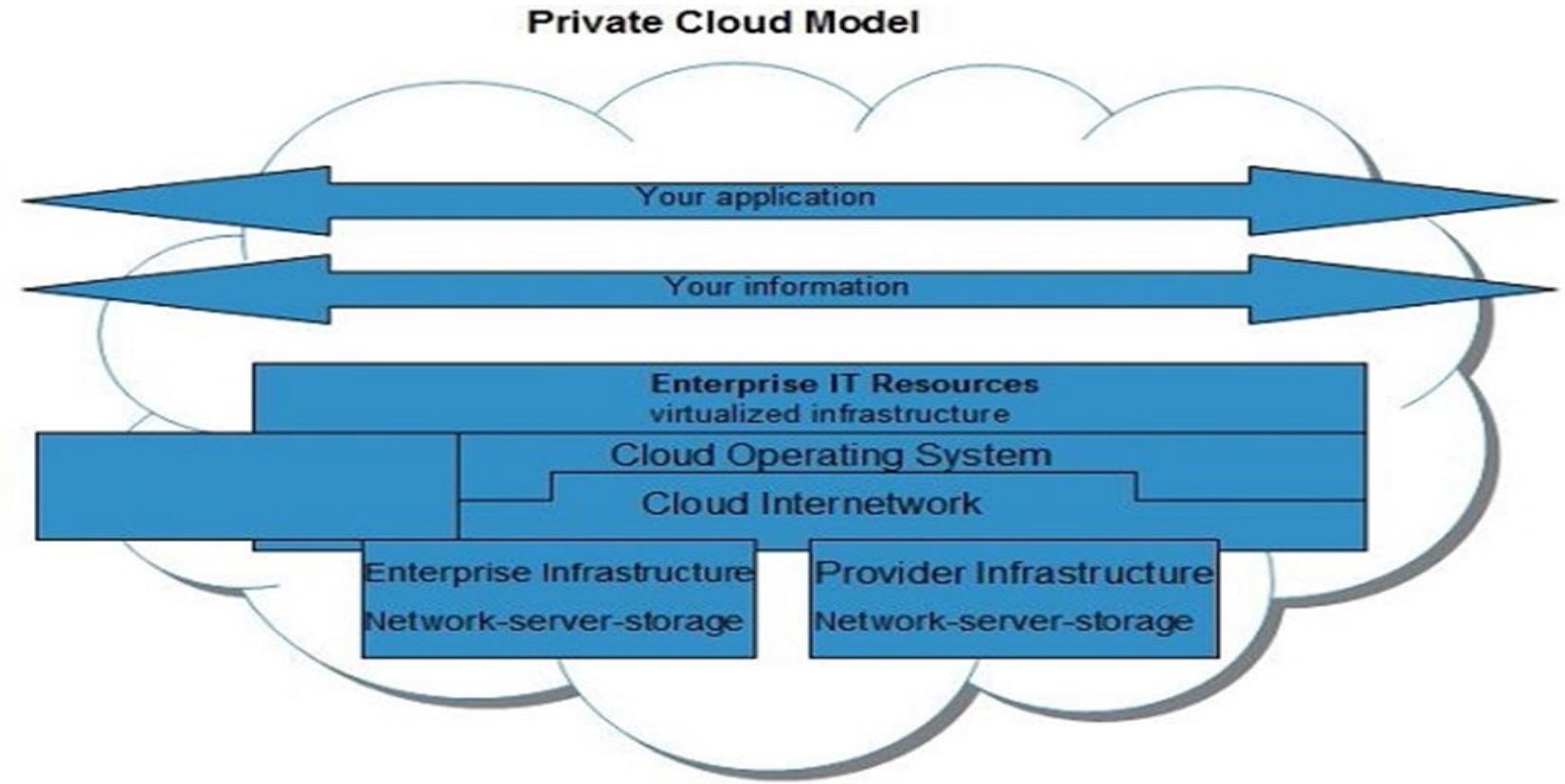


The Public Cloud Model allows systems and services to be easily accessible to general public. e.g. Google, Amazon, Microsoft offers cloud services via internet.

The Advantages Of Public Cloud

- **Economic** – much lower cost than a private cloud
- **Expertise** –access to a staff with expertise on many topics
- **Advanced services**– offerings not available elsewhere

Private Cloud

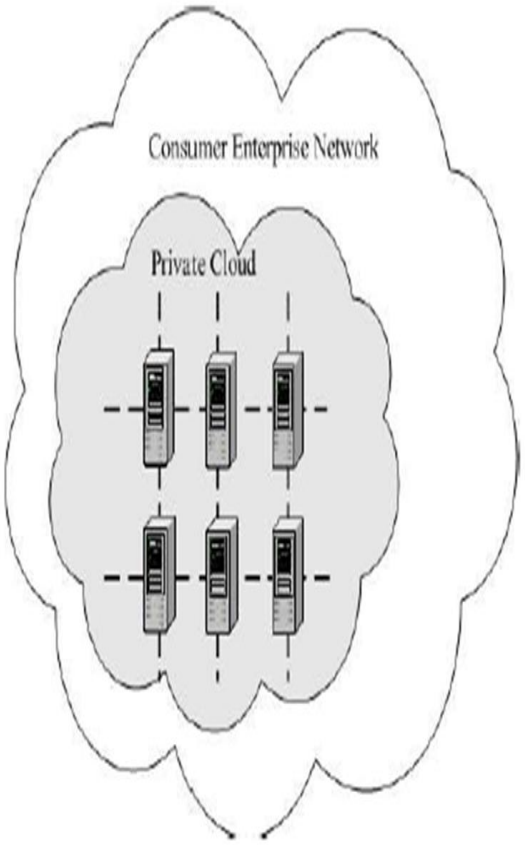


A private cloud is a cloud computing environment dedicated to a single organization.

A private cloud can be hosted either at an organization's own data center, at a third party colocation facility, or via a private cloud provider who offers private cloud hosting services

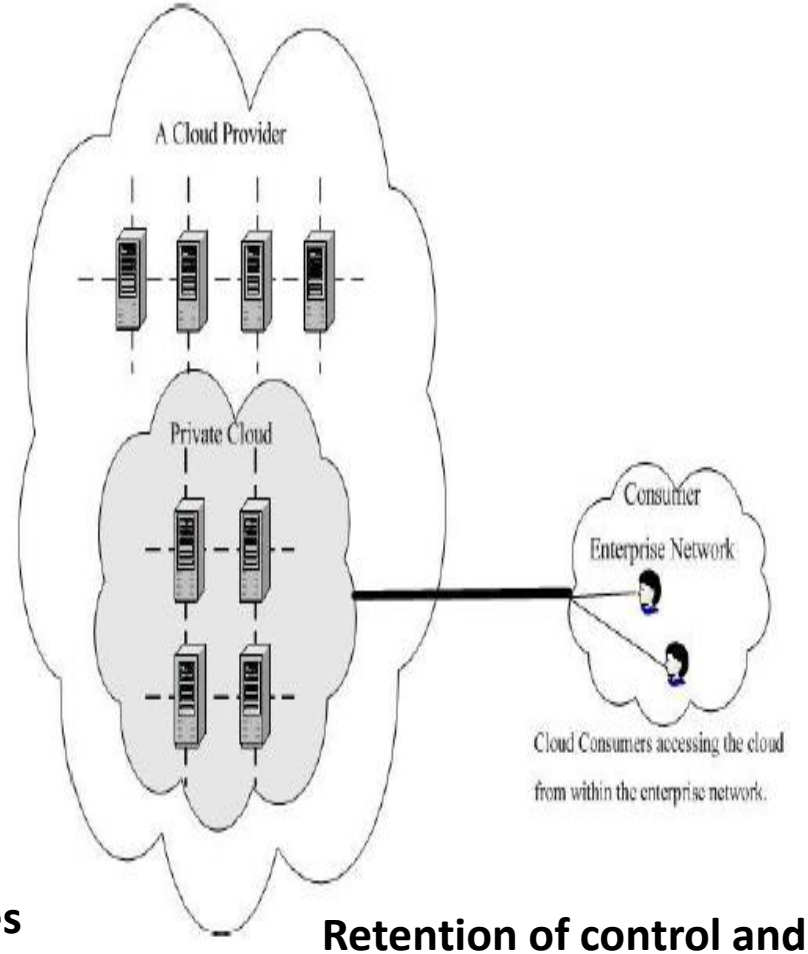
Private Cloud

ONSITE



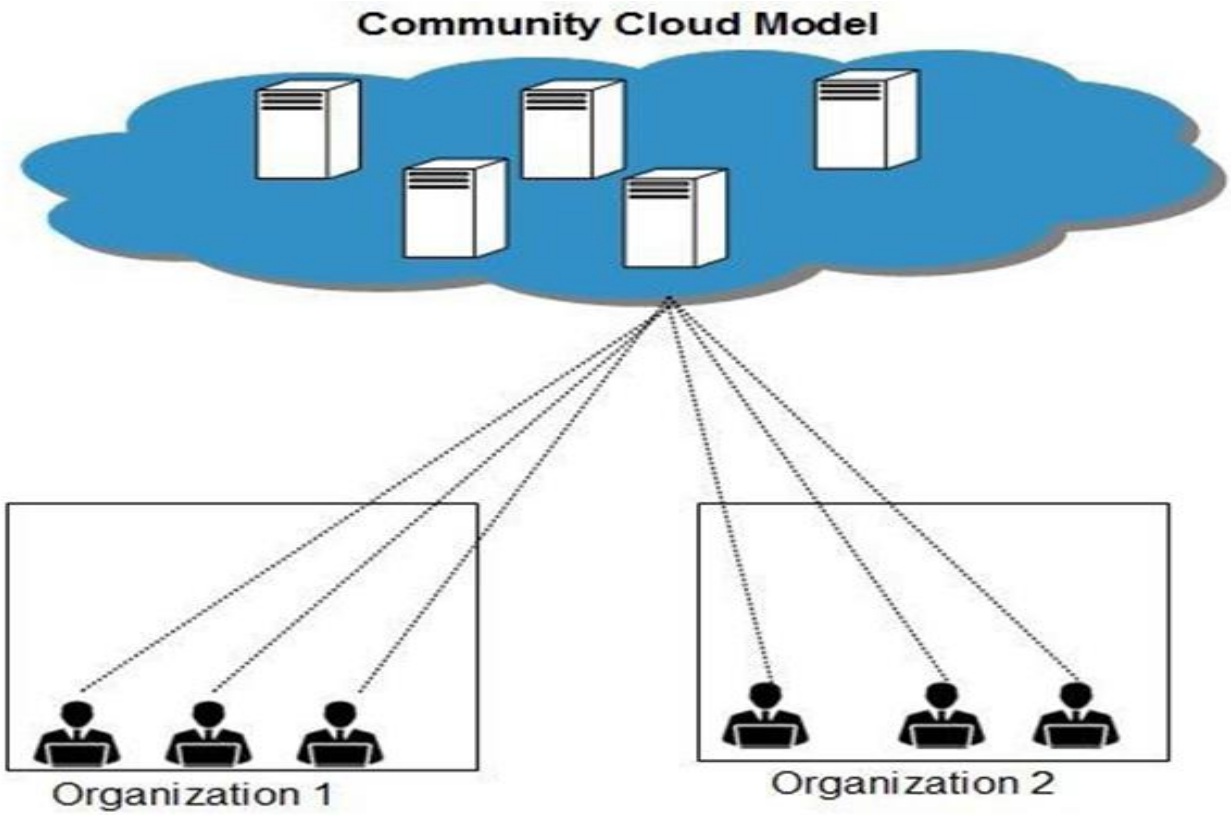
Reduced latency with on-premises facilities

OUT-SOURCED



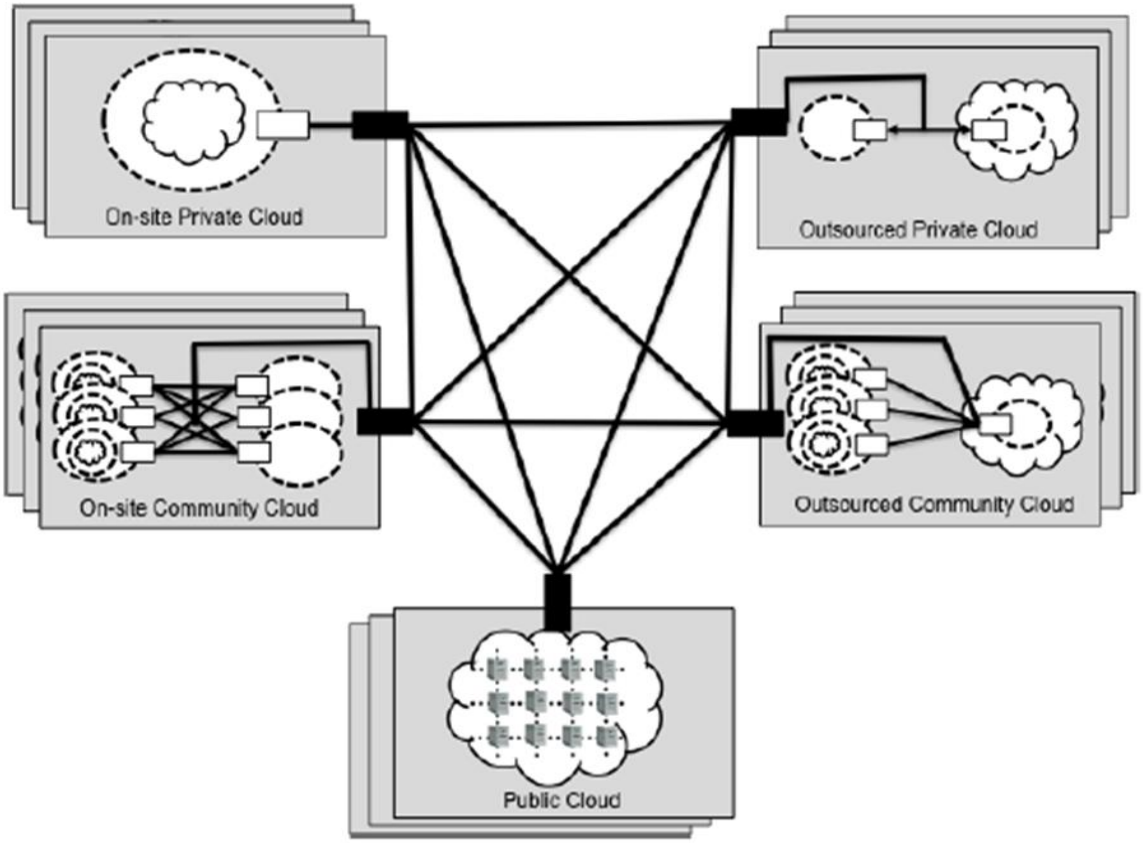
Retention of control and visibility

Community Cloud



The **Community Cloud** allows system and services to be accessible by group of organizations. It shares the infrastructure between several organizations from a specific community. It may be managed internally or by the third-party.

Hybrid Cloud



Cloud Definition Framework

Deployment Models

Service Models

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)

Essential Characteristics

- On Demand Self-Service
- Broad Network Access
- Rapid Elasticity
- Resource Pooling
- Measured Service

Common Characteristics

- Massive Scale
- Resilient Computing
- Homogeneity
- Geographic Distribution
- Virtualization
- Service Orientation
- Low Cost Software
- Advanced Security

Essential Cloud Characteristics

- On-demand self-service
  - Get computing capabilities as needed automatically
- Broad network access
  - Services available over the net using desktop, laptop, PDA, mobile phone

Essential Cloud Characteristics (Cont.)

- Resource pooling
  - Location independence
  - Provider resources pooled to server multiple clients
- Rapid elasticity
  - Ability to quickly scale in/out service
- Measured service
  - control, optimize services based on metering

Cloud computing - Characteristics

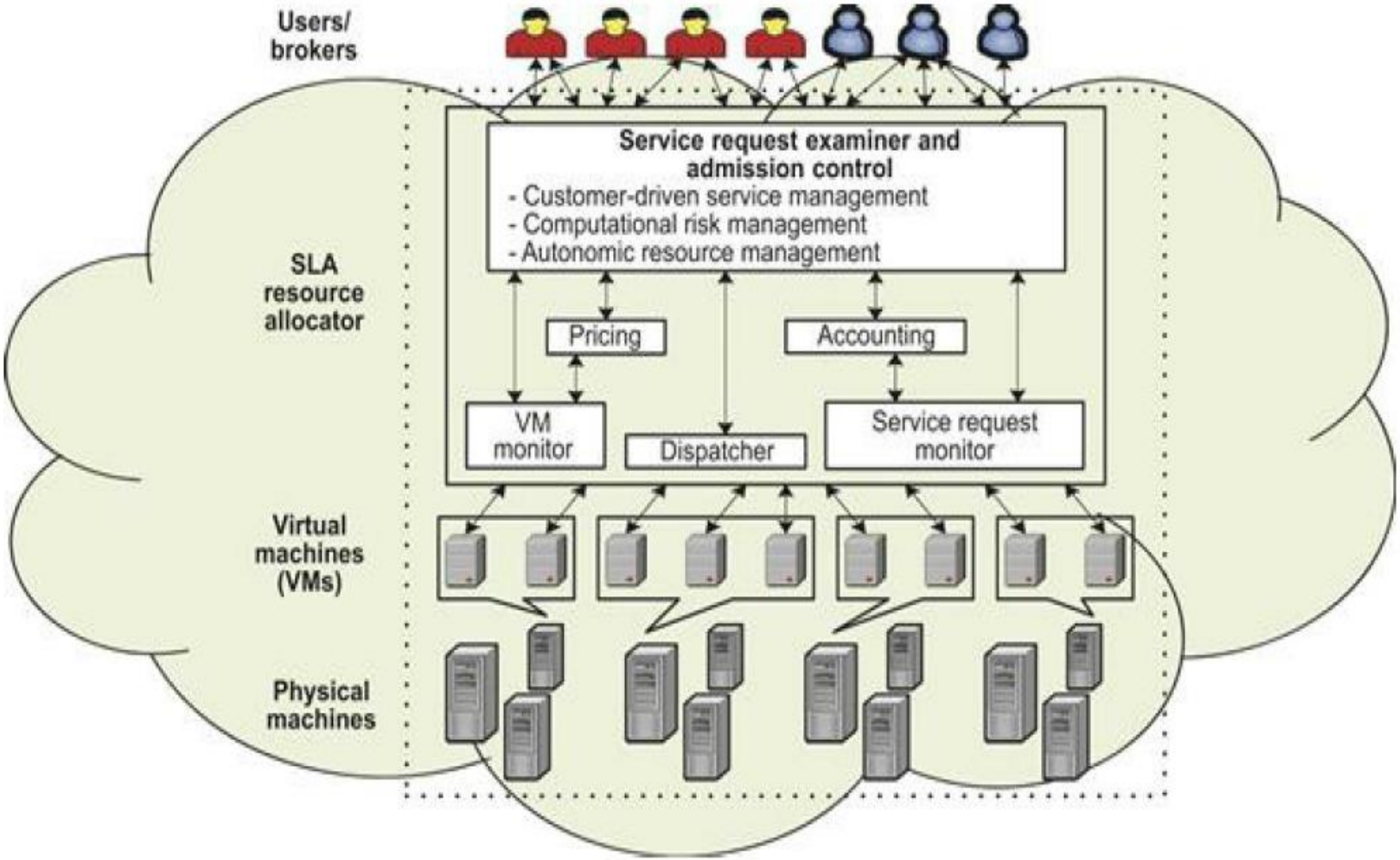
- Agility – On demand computing infrastructure
  - Linearly scalable – challenge
- Reliability and fault tolerance
  - Self healing – Hot backups, etc
  - SLA driven – Policies on how quickly requests are processed
- Multi-tenancy – Several customers share infrastructure, without compromising privacy and security of each of the customer's data
- Service-oriented – compose applications out of loosely coupled services. One service failure will not disrupt other services. Expose these services as API's

Cloud computing - Characteristics

- Virtualized – decoupled from underlying hardware. Multiple applications can run in one computer

- Data, Data, Data
  - Distributing, partitioning, security, and synchronization

Market Oriented Cloud



Conventional Computing
vs.
Cloud Computing

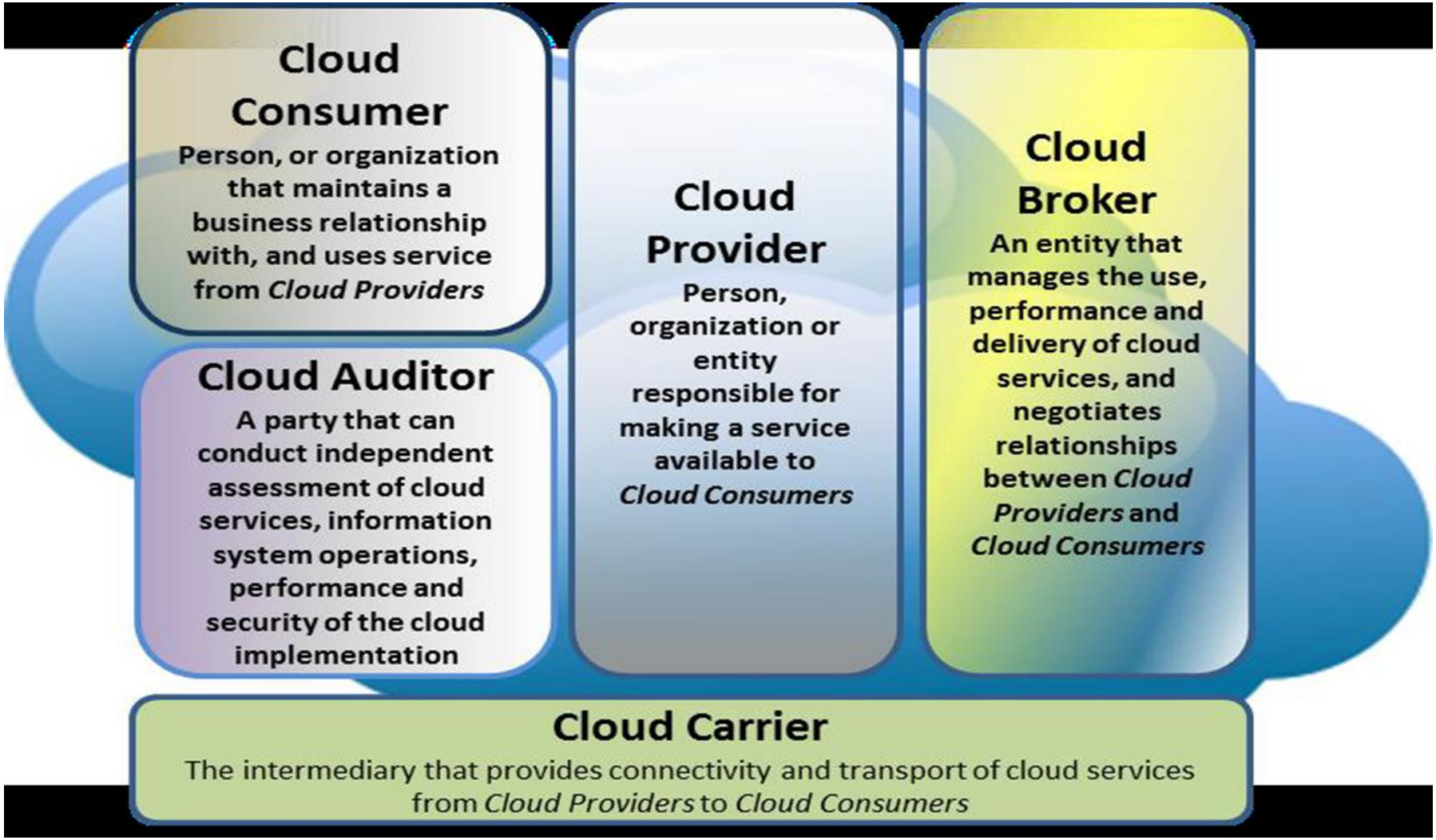
- Conventional

- Manually Provisioned
- Dedicated Hardware
- Fixed Capacity
- Pay for Capacity
- Capital & Operational Expenses

Cloud

- Self-provisioned
- Shared Hardware
- Elastic Capacity
- Pay for Use
- Operational Expenses

NIST- Cloud Computing Standards
(National Institute of Standards and Technology)



Provider Lock-In

- Cloud providers usually offer a *migration service* that makes it easy for a corporate customer to move their computing into the provider's public cloud.

- *cloud-to-cloud migration services*.

- Industry uses the term lock-in to refer to the practice of using enticements and obstacles that make it in convenient ore expensive for customers to move to another cloud provider.

Multi-Cloud

- An organization becomes a customer of more than one public cloud provider.
- The division of computation among providers depends on the structure of the organization and its IT needs.
- An organization can also adopt a multi-cloud approach in which the organization uses multiple public cloud providers.

Hyperscalers

- Companies that own and operate the largest data centers are known as hyperscalers.

- Among public hyperscale cloud providers, Amazon's AWS, Microsoft's Azure Cloud, and Google's GCP have attracted enterprise customers.

Data Center Infrastructure And Equipment

Racks, Aisles, And Pods

- Physically, racks holding equipment are placed side by side in rows, leaving *aisles* between them.
- Data center is built by replicating a basic set of equipment known as a *pod.*
- A pod or a cluster is simply a set of computers linked by high-speed networks into a single unit
- A point of delivery, or PoD, is "**a module of network, compute, storage, and application components that work together to deliver networking services**".

Pod Size

- An early design created pods with over 200 racks per pod.
- The industry has moved to smaller sizes, where a pod with 48 racks is considered “large,” and an average-size pod contains 12 to 16 racks.

✓ Incremental growth

✓ Manageability

✓ Power and cooling

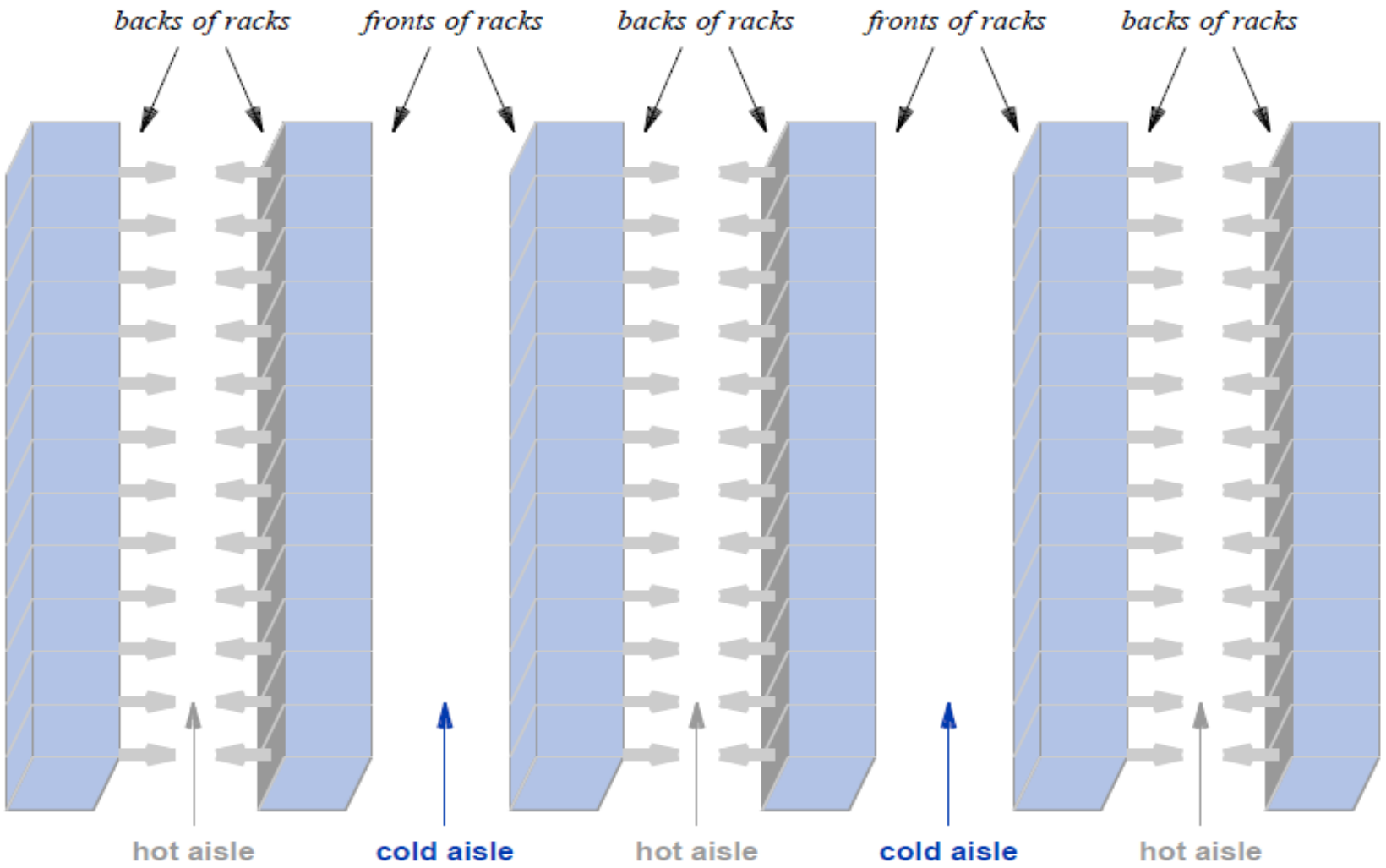
Design – Heat Reducing

- Raised floor pathways and air cooling
- Thermal containment and hot/cold aisles
- Exhaust ducts(chimneys)
- Lights-out datacenters

Thermal Containment And Hot/Cold Aisles

- Hot air leaves each piece of equipment, venting into the data center.
- Overall, air flow in the data center must be designed carefully to move hot air away from the racks.
- Ensuring that it cannot be accidentally drawn back into another piece of electronic equipment.

Thermal Containment And Hot/Cold Aisles



Exhaust Ducts (Chimneys)

- Despite fans in the ceiling that draw hot air upward, the temperature near racks with high power density can be higher than other areas of a datacenter.
- Designers refer to such areas as *hot spots*.
- For areas that generate inordinate heat, a vertical duct with a fan can be placed over the area with a fan to move hot air upward.

Lights-Out Data Centers

- An operational paradigm has been invented that helps reduce heat in a data center: minimize any extraneous use of electricity.
- *lights-out datacenter*, the scheme means that entire parts of the data center operate in the dark.
- To minimize the time lights must be on, servers, network switches, and storage equipment are accessed and managed over a network.
- The availability of reliable, automated failure recovery and maintenance systems has further enabled the lights-out approach.
- Automated systems are used for routing, monitoring tasks, and handling fast cut-over during failures.

Lights out Approach

- In addition to reducing energy costs, The lights-out approach has three advantages.
-
- Using automated systems to monitor a data center offers owners cost savings by reducing the staff size; automation is less likely than human operators to misconfigure equipment; and restricting personnel in the data center reduces the threat of malicious attacks.

A Possible Future Of Liquid Cooling

- When it changes from air cooling to liquid cooling, a data center must install hydraulic equipment to circulate cold liquid refrigerant to the racks and return heated refrigerant to the cooling unit.

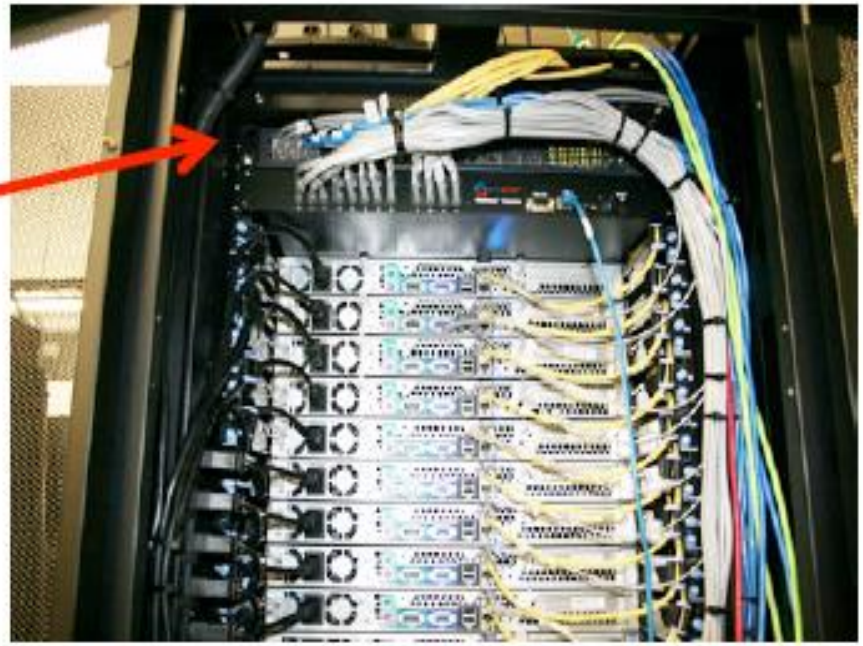
- In addition, all servers and network equipment must be replaced with units that have hydraulic fittings to accommodate liquid cooling.

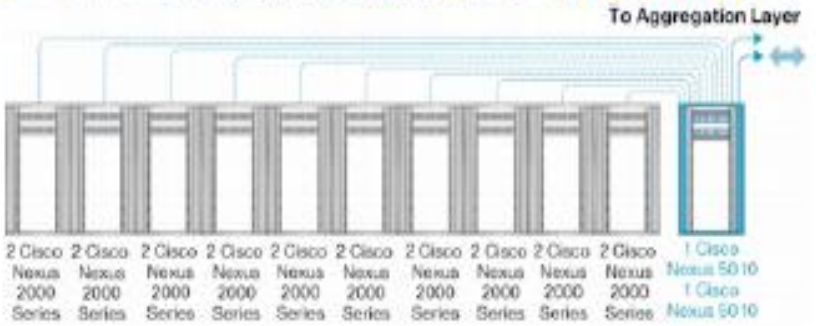
Network Equipment And Multi-Port Server Interfaces

- Servers (Physical machines)
- Storage
- Network devices (switch, router, cables)
- Topology
- Routing / switching equipment
- Protocols

Top-Of-Rack Architecture

- Rack of servers
  - Commodity servers
  - And top-of-rack switch
- Modular design
  - Preconfigured racks
  - Power, network, and storage cabling
- Aggregate to the next level

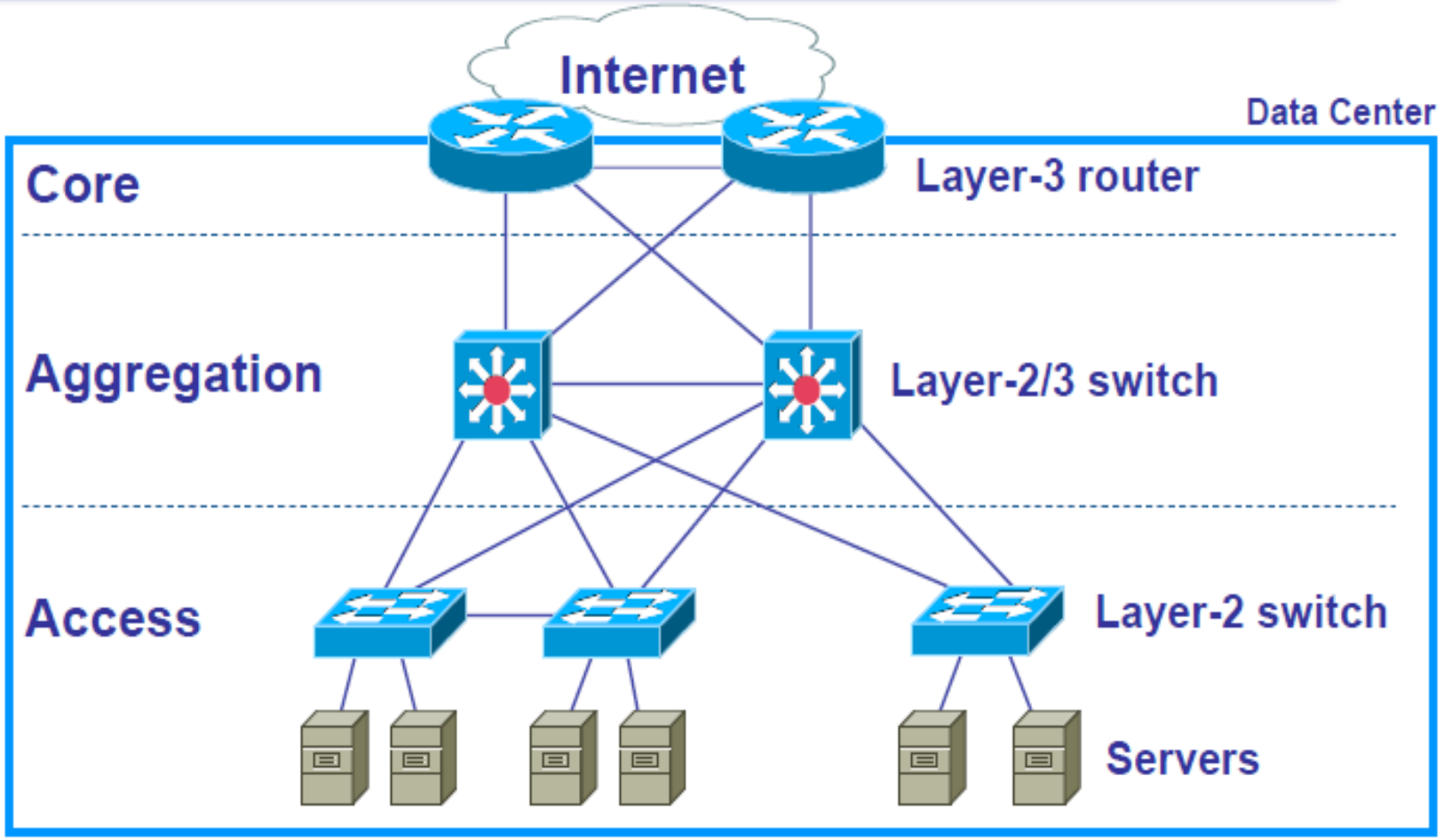




Rapid Data Transfer

- The connections between the ToR switch and each server must operate at high speed (Gbps)
- Data centers - 10 Gbps and 40 Gbps
- To further increase the rate at which data can be sent, each server can use a ***multi-port network interface card*** (*multi-port NIC*).
- Each of the ports connects to the ToR switch, and each operates independently and in parallel.
- A multi-port NIC works well with a multi-core server because it allows the server to send and receive more data.

Traditional DC Topology



Layer 2 vs. Layer 3?

Ethernet switching (layer 2)

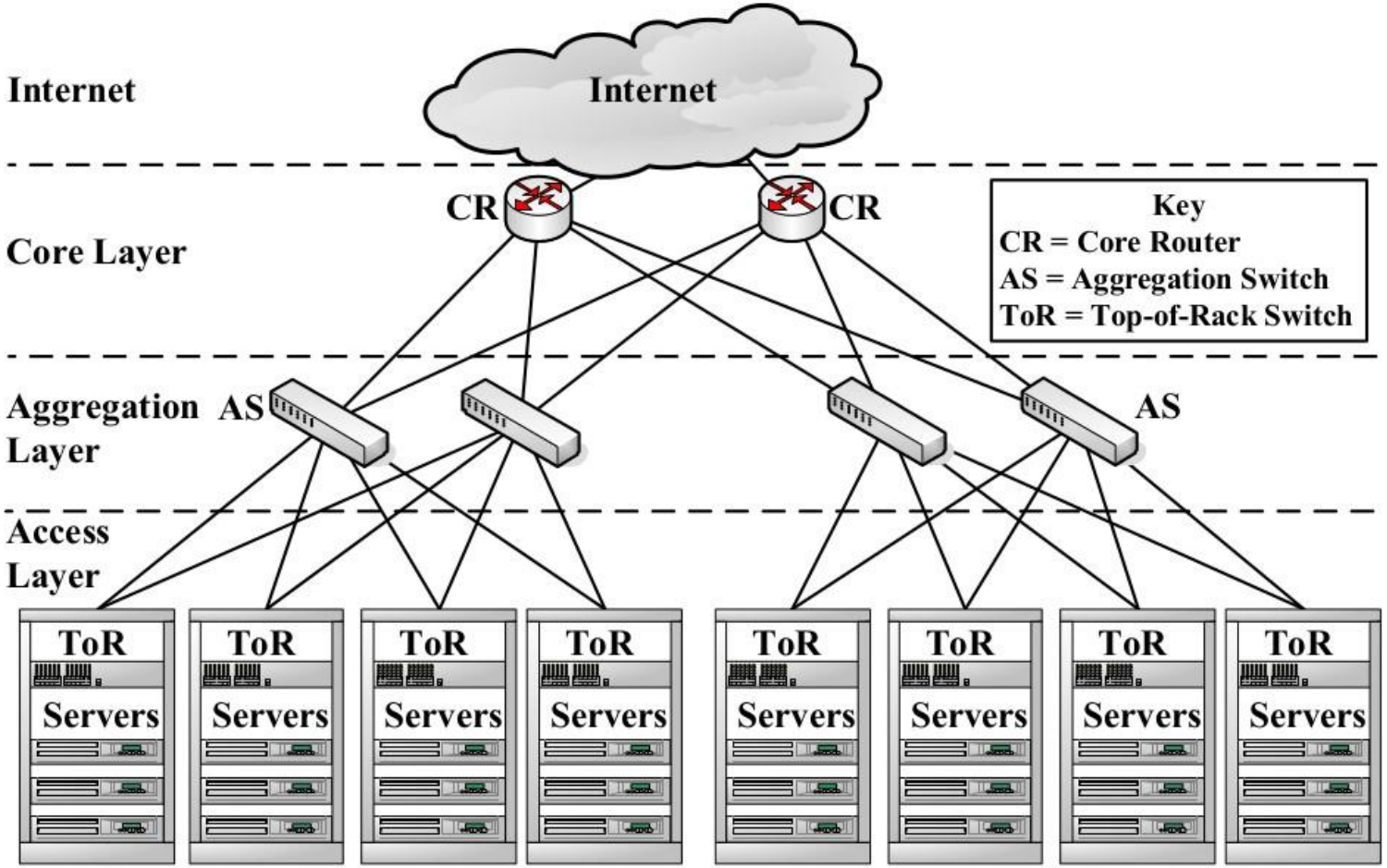
- Cheaper switch equipment
- Fixed addresses and auto-configuration
- Seamless mobility, migration, and failover

IP routing (layer 3)

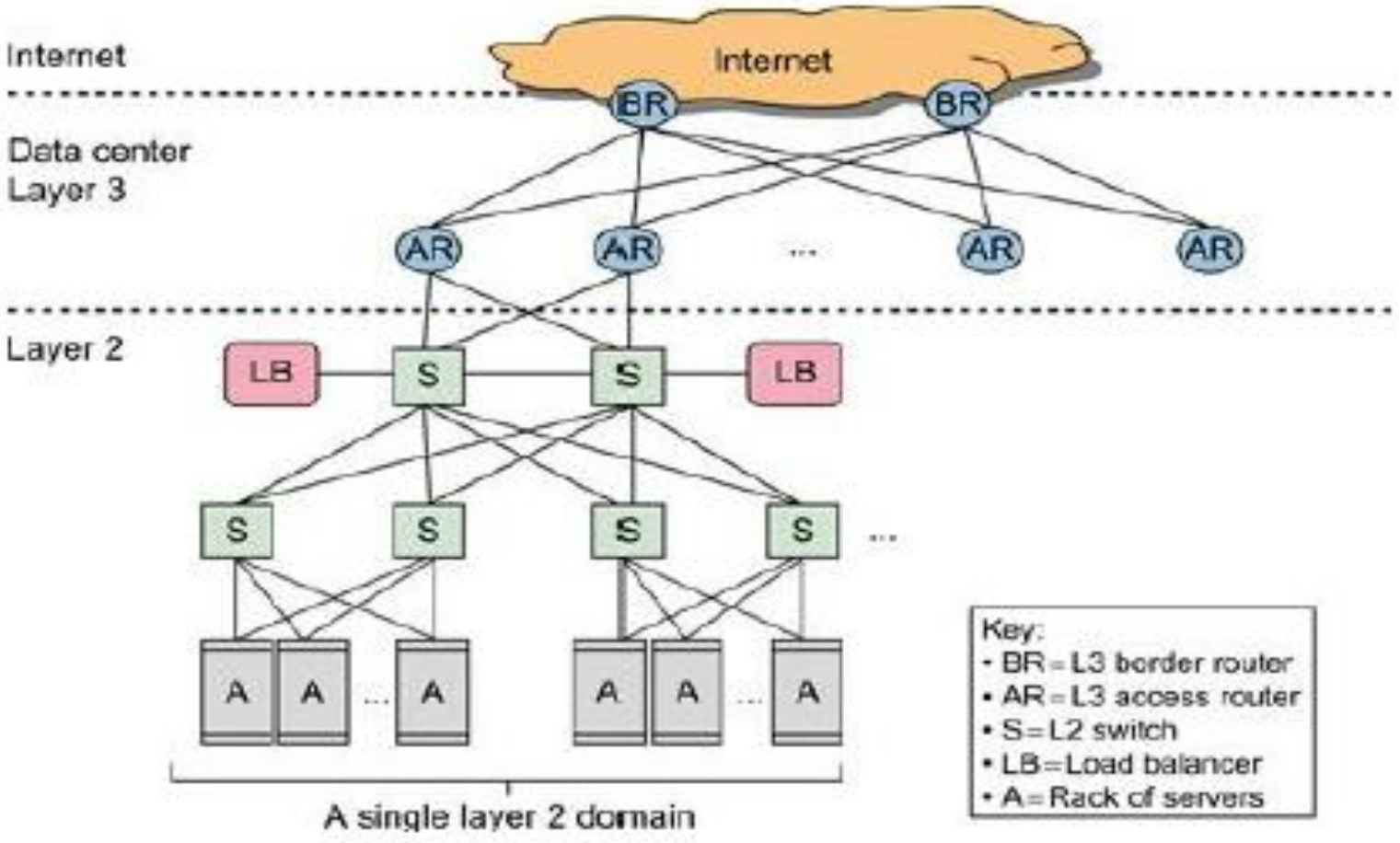
- Scalability through hierarchical addressing
- Efficiency through shortest-path routing
- Multipath routing through equal-cost multipath

- **Data centers often connect layer-2 islands by IP routers**

Conventional Topology



Standard Data Center Routing



North-South And East-West Network Traffic

- How should the ToR switches in all the racks be interconnected to form a network in the data center?
- How should the data center network connect to the Internet?
- A variety of network architectures have been used in data centers

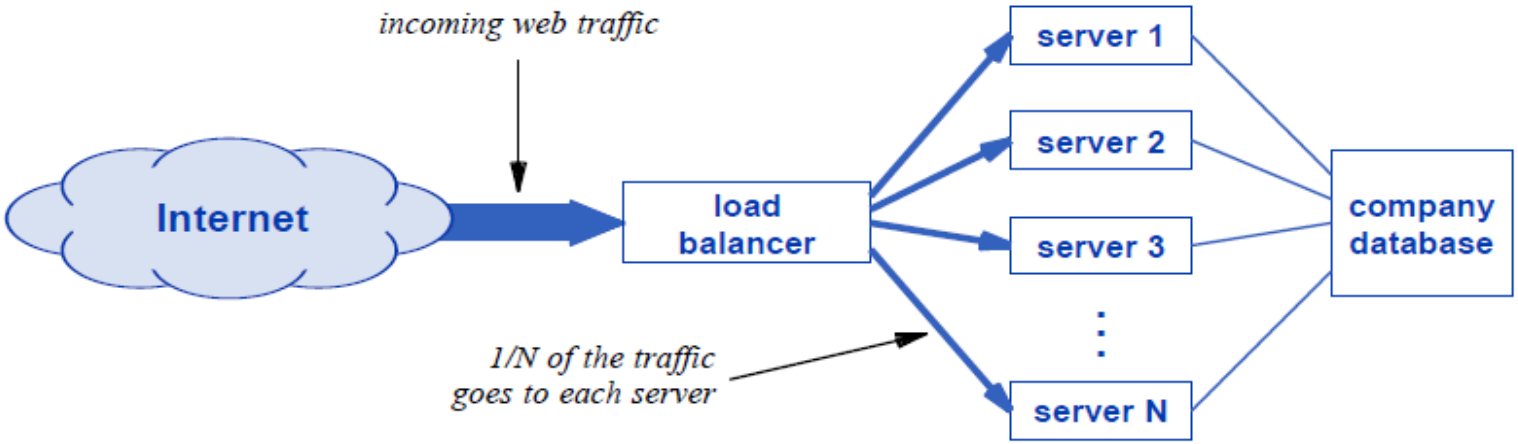
  North-south traffic

  East-west traffic

North-South traffic

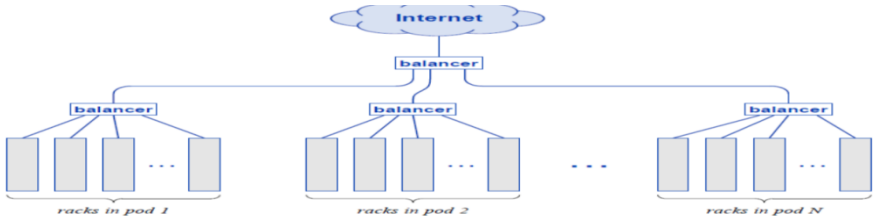
- Industry uses the term north-south traffic to describe traffic sent between arbitrary computers on the Internet and servers in a datacenter.

- For example, that early data centers focused on large-scale web sites. Web traffic falls into the category of north-south traffic.



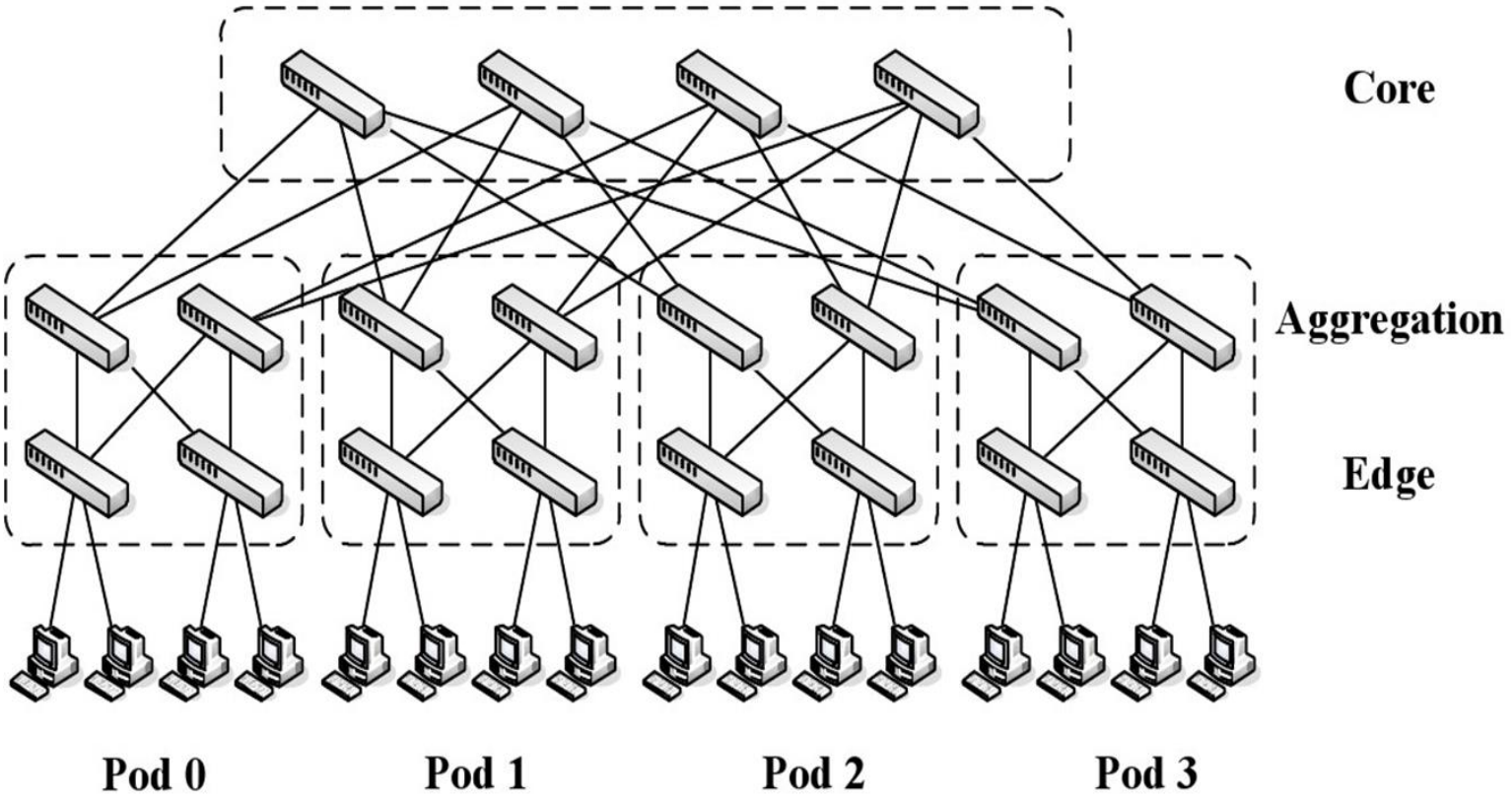
East-west traffic

- Consider a company using cloud computing. When the company fills an order, software may need to access both a catalog of products as well as a customer database.
- Similarly, when a manager approves time off, software may need to access an employee's record, payroll data, and the company's accounting system.
- Communication within the company means network traffic will travel among the servers the company has leased.
- Communication proceeds left and right, which leads to the name east-west traffic.

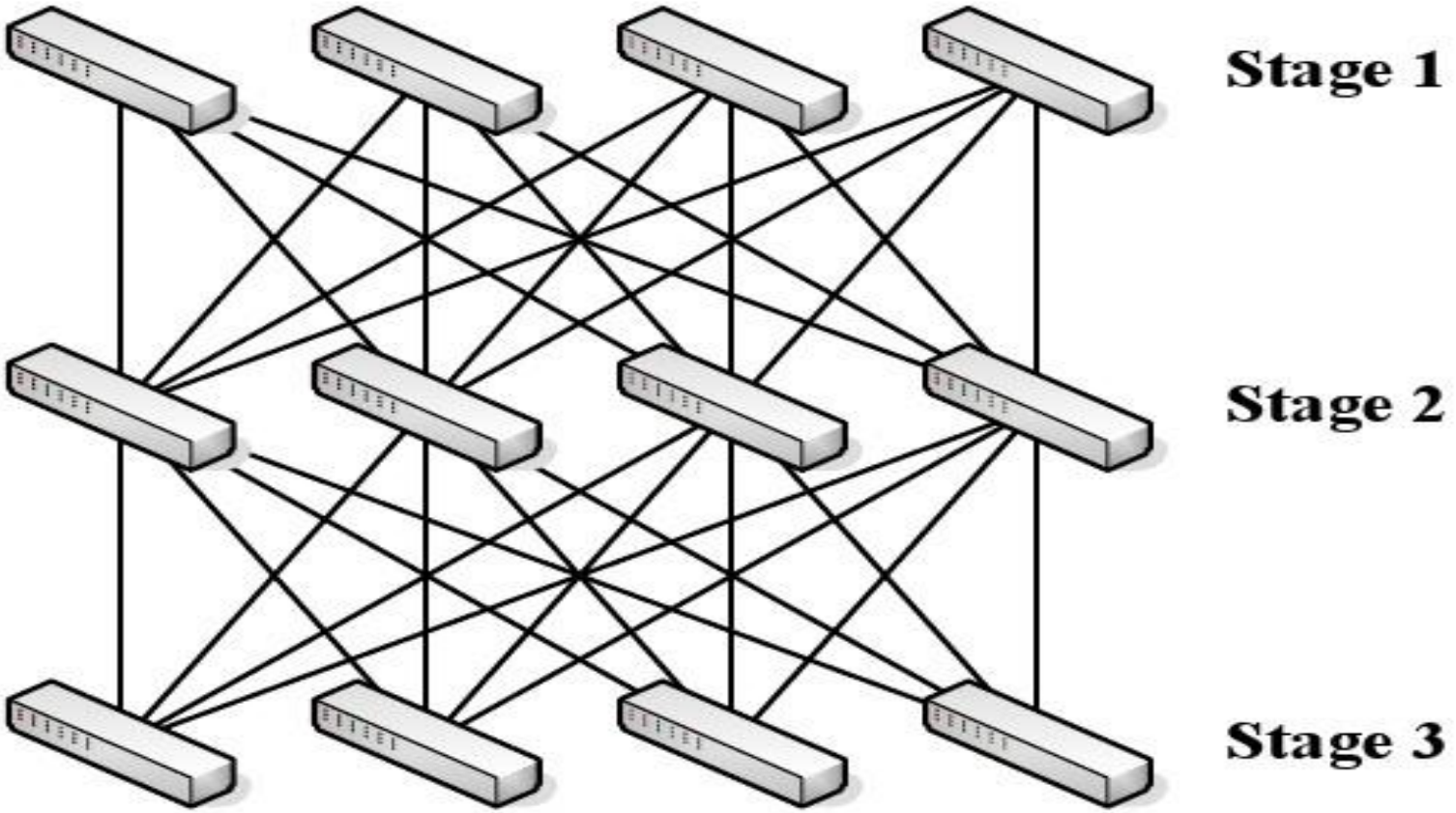


Network Hierarchies, Capacity, And Fat Tree Designs

- Hierarchical / staged/layered/fat tree



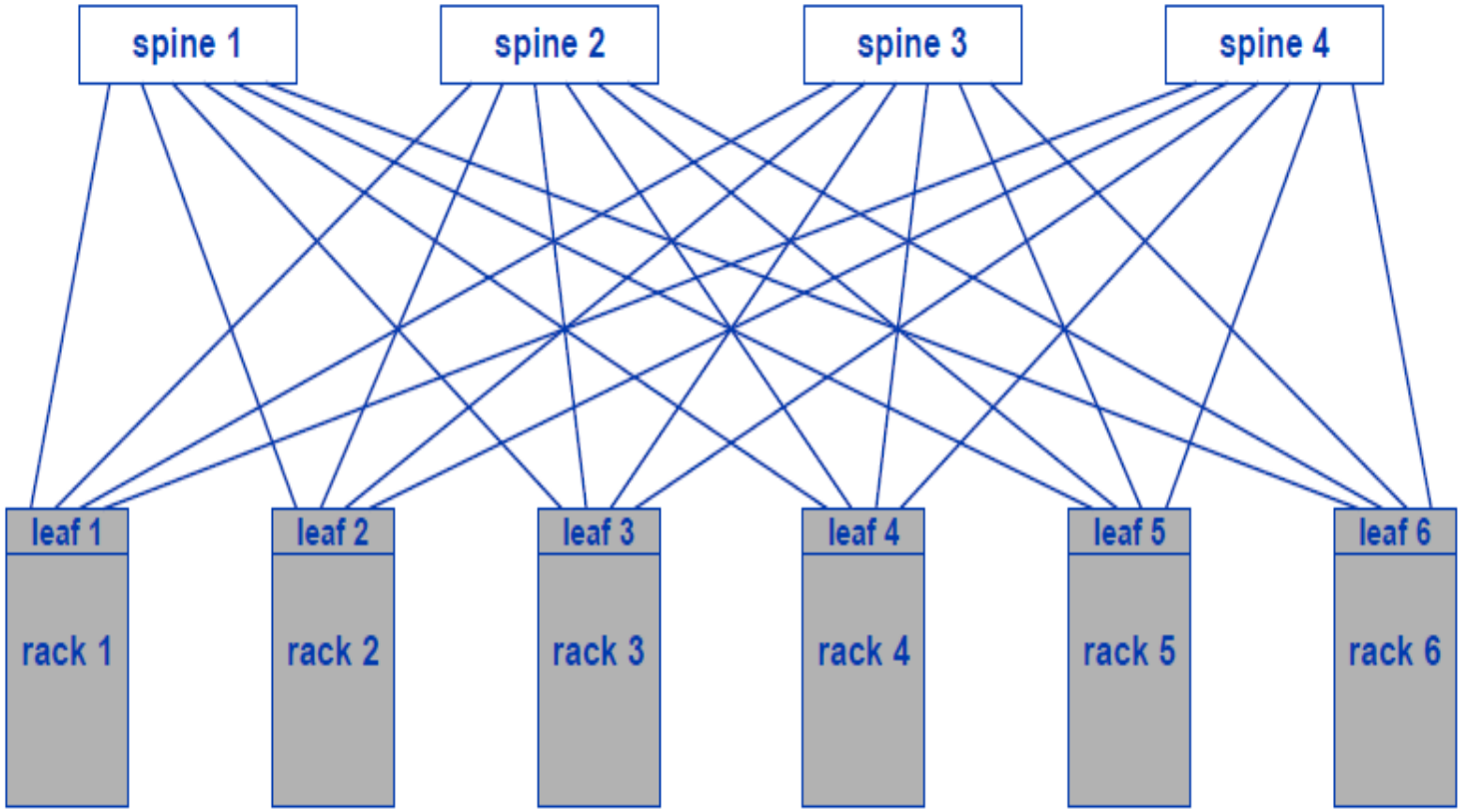
Clos Topology



Leaf-Spine Architecture

- ***How can a data center network be designed that handles large volumes of east-west traffic without using a hierarchical design?***
- The answer lies in parallelism and a form of load balancing.
- The specific approach used in data center is known as a *leaf-spine network architecture*
- In leaf-spine terminology, each Top-of-Rack switch is called a *leaf*.
- An additional set of *spine* switches and connects each leaf switch to each spine switch.

Leaf-Spine Architecture



Leaf-Spine Architecture

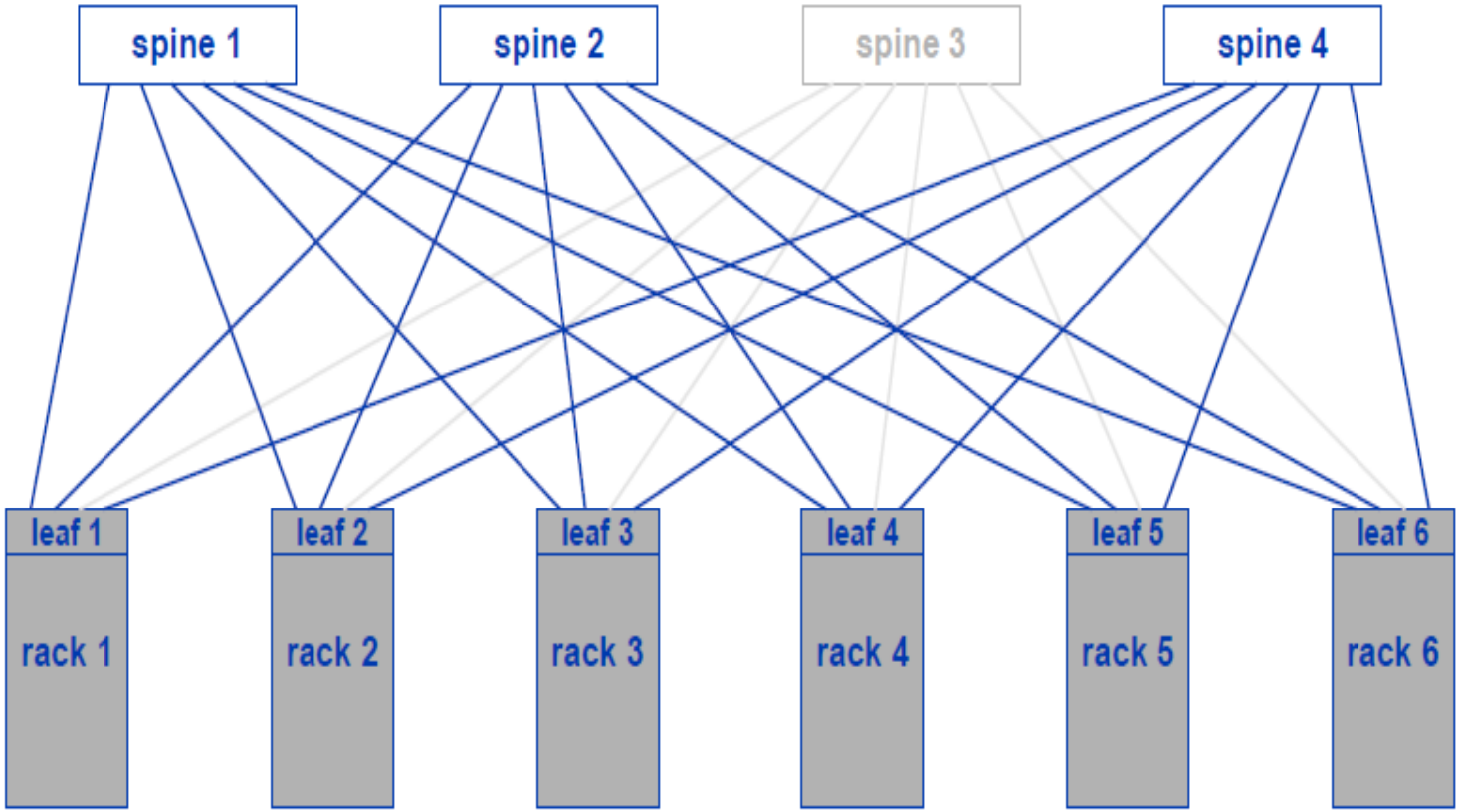
- **Higher capacity for east-west traffic**
- **Redundant paths to handle failures**

Because both the source and destination racks connect to all four spine switches, four independent paths exist between each pair of racks, one path through each spine switch

A leaf switch equipped with *Equal Cost Multipath Routing* (*ECMP*) technology can be configured to divide traffic equally among the paths.

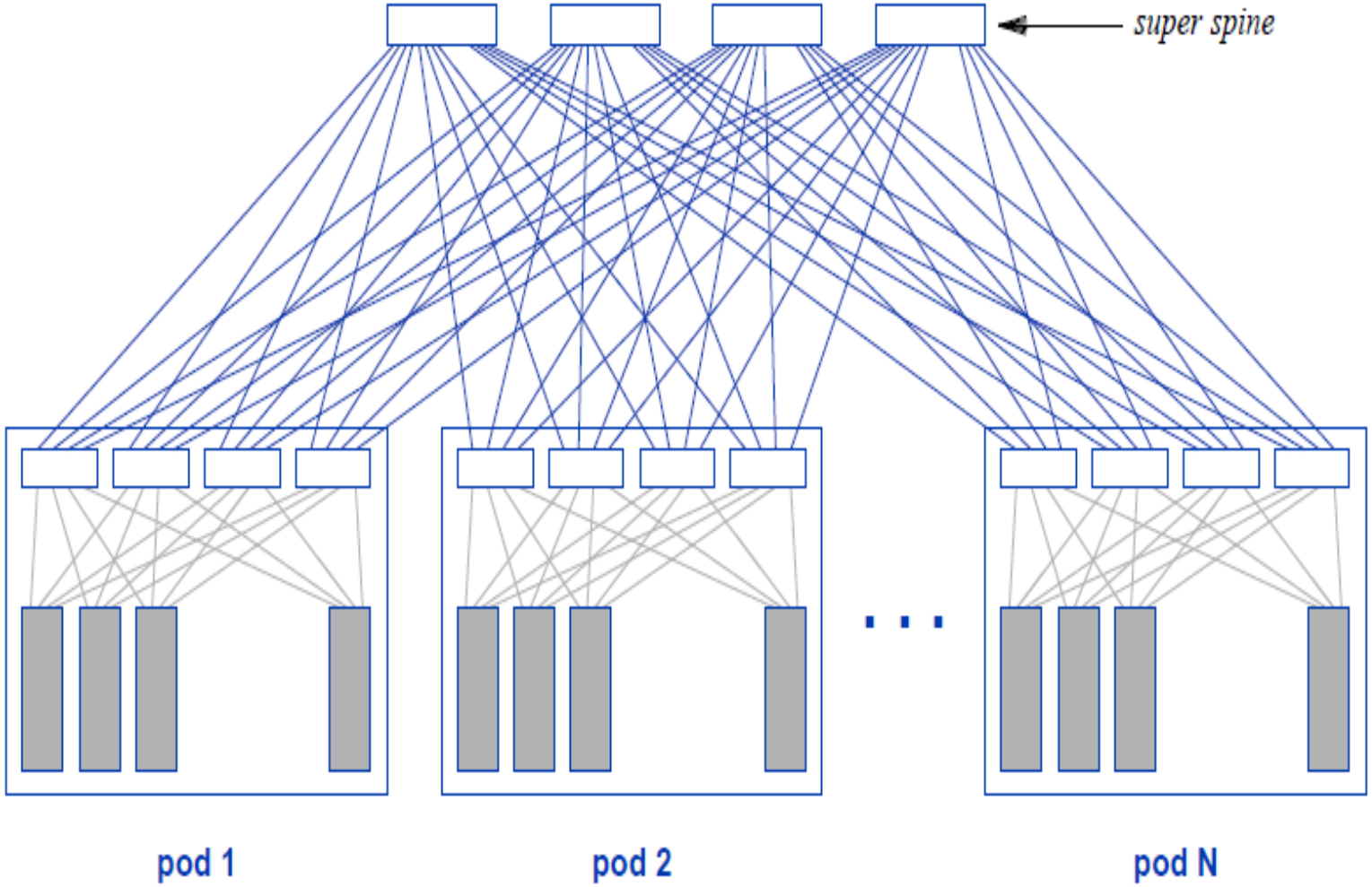
ECMP means one- fourth of the data will travel through spine 1, another fourth of the data will travel through spine 2, and so on

Redundant paths to handle failures



Network Management Software

Scaling A Leaf-Spine Architecture With A Super Spine



Storage In A Data Center

- Data center providers follow the same basic approach for storage facilities as they do for computational facilities: parallelism.
- Modern data centers use *Solid State Disks* (*SSDs*).
- Virtualized disk.
- Industry uses the term *block storage* to refer to virtualized disks.
- As software on a virtualized server accesses or stores data on its disk, requests travel across the data center network to the storage facility, and replies travel back over the network.
- The higher reliability of solid state disks has lowered failure rates, making replacement much less frequent.

Unified Data Center Networks

- Early data center storage facilities used specialized network hardware that was designed to optimize remote storage access. The specialized hardware was expensive and required running extra cables.

- Data centers are now using a single network for storage access as well as other communication.

- *The availability of low-cost Ethernet hardware and a leaf-spine network architecture has allowed data centers to eliminate special- purpose storage networks and move to a single, unified network that carries storage access traffic as well as other traffic.*

Summary

- Elastic Computing
- Business Model
- Types of Cloud
- Data Centers
- Leaf-Spine Architecture
- Storage
- Unified Datacenter Networks