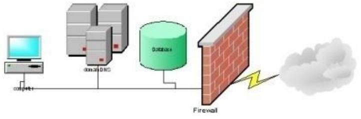
Management of Information Systems

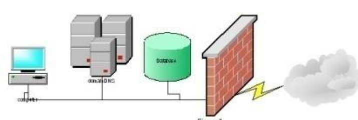
Data Centers

BME VIK TMIT

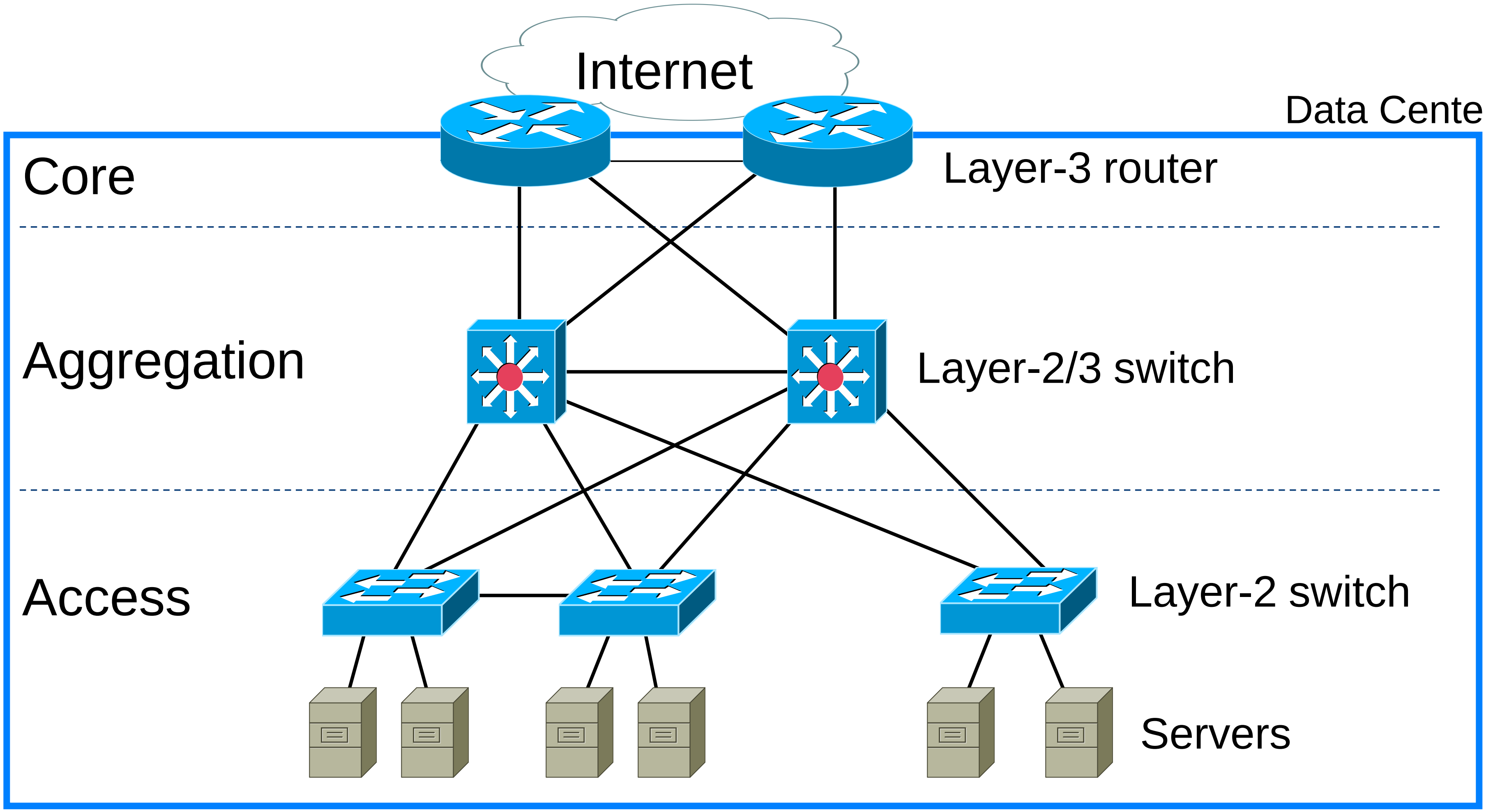


Data Centers

- A **data center** is a facility used to house computer systems and associated components
- Nowadays, data center consisting of tens of thousands of PCs are increasingly common in universities, research labs, and companies
- Data center can be used to do scientific computing, financial analysis, data analysis and warehousing, and providing large-scale network services

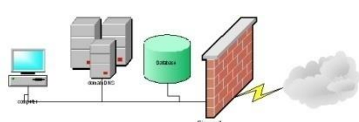


Common data center topology

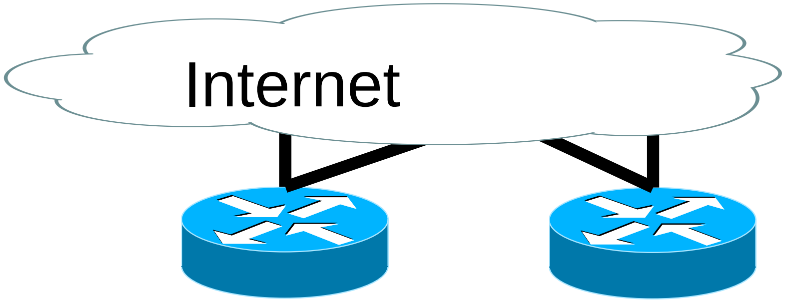


Multi-Tier Model

- Access
  - Where hosts connect to the network
- Aggregation
  - To which the access layer is redundantly connected
- Core
  - Provides routing services
  - “Connects DC to world outside”
    - to other parts of the data center
    - to services outside of the data center such as the Internet
    - geographically separated data centers
    - and other remote locations

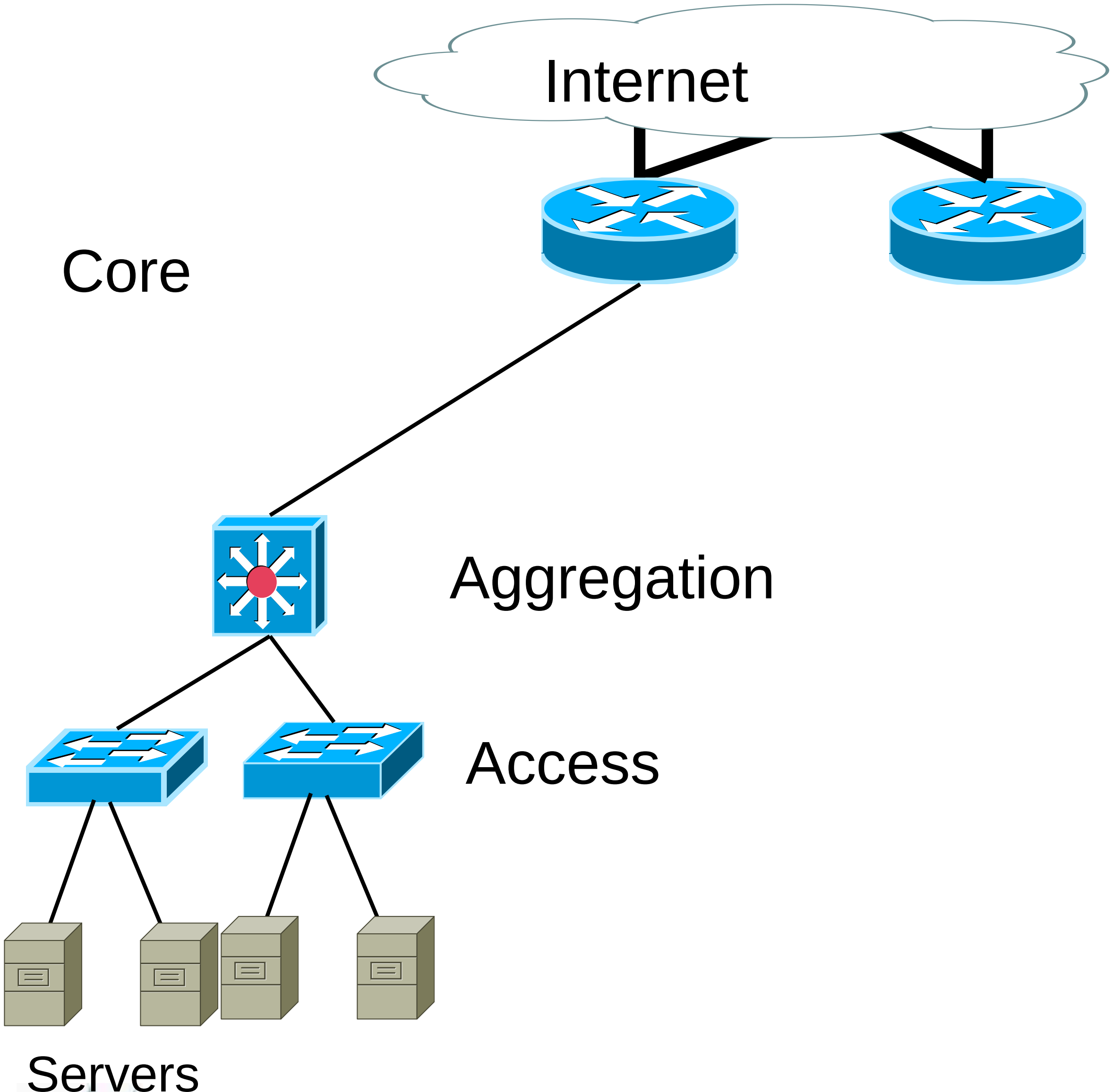


Multi-Tier Model

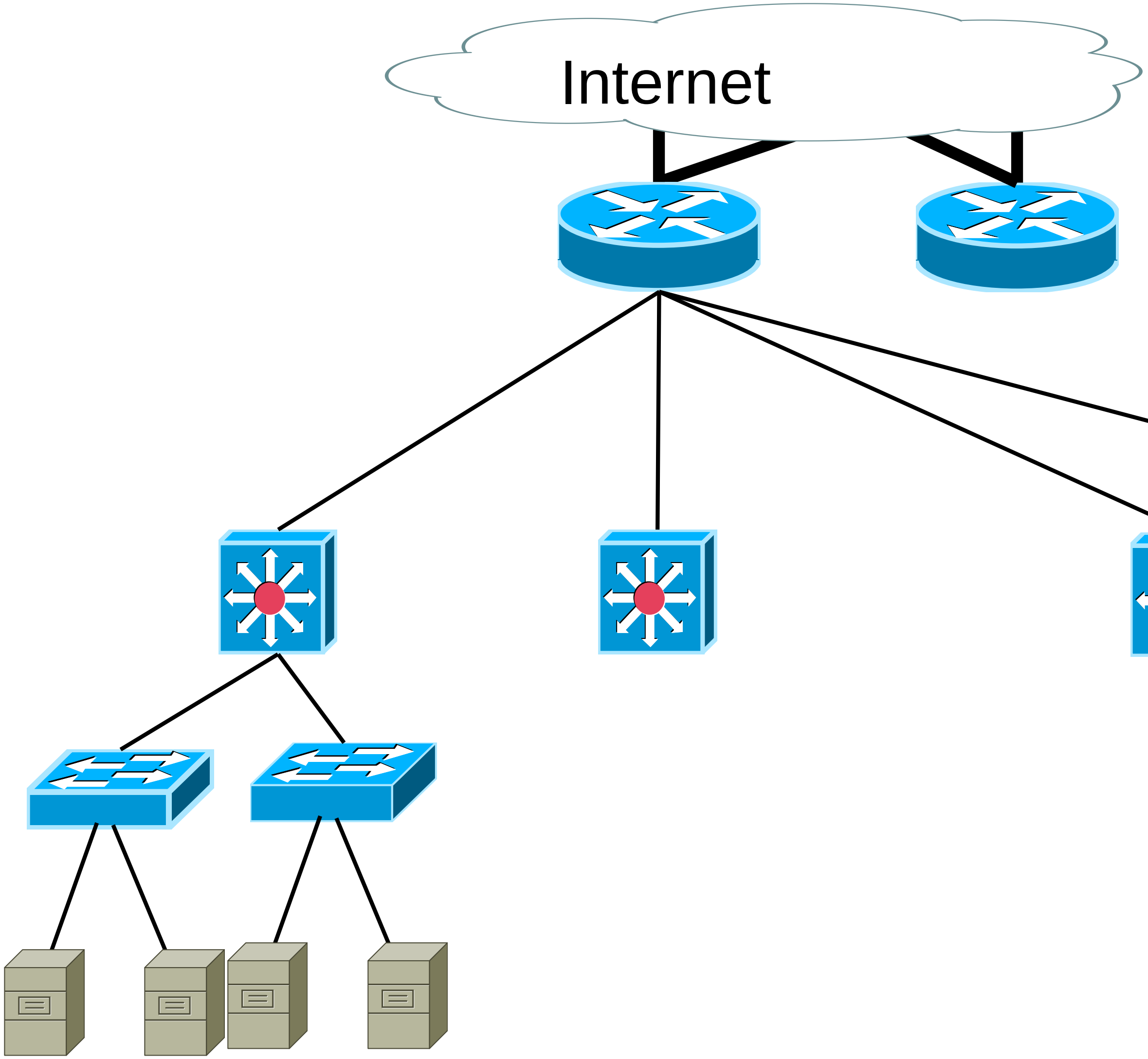


Core

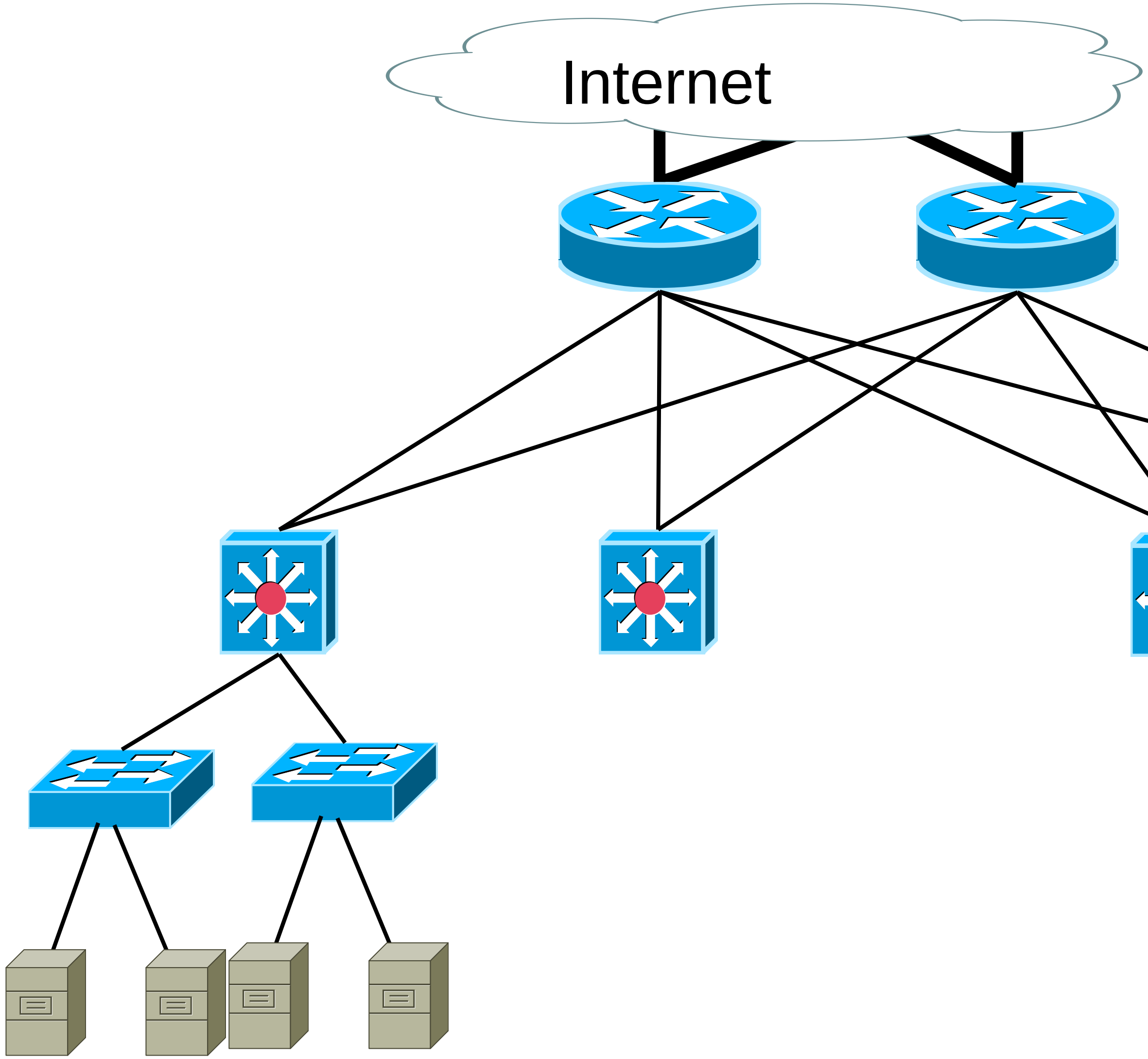
Multi-Tier Model

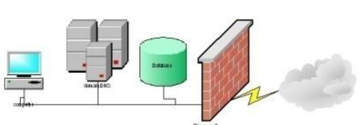


Multi-Tier Model

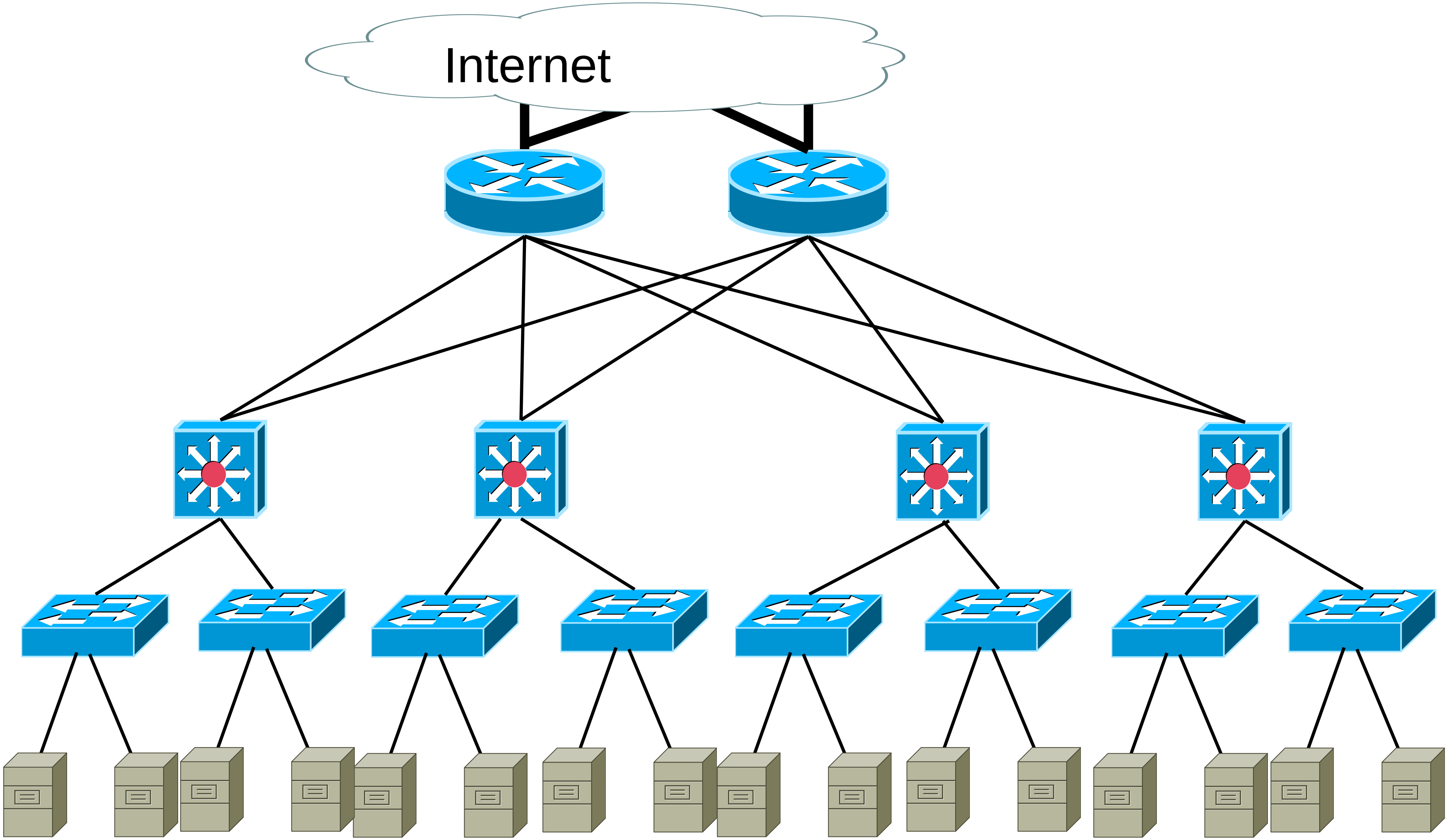


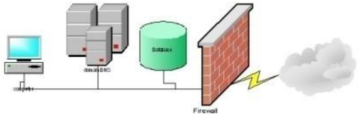
Multi-Tier Model



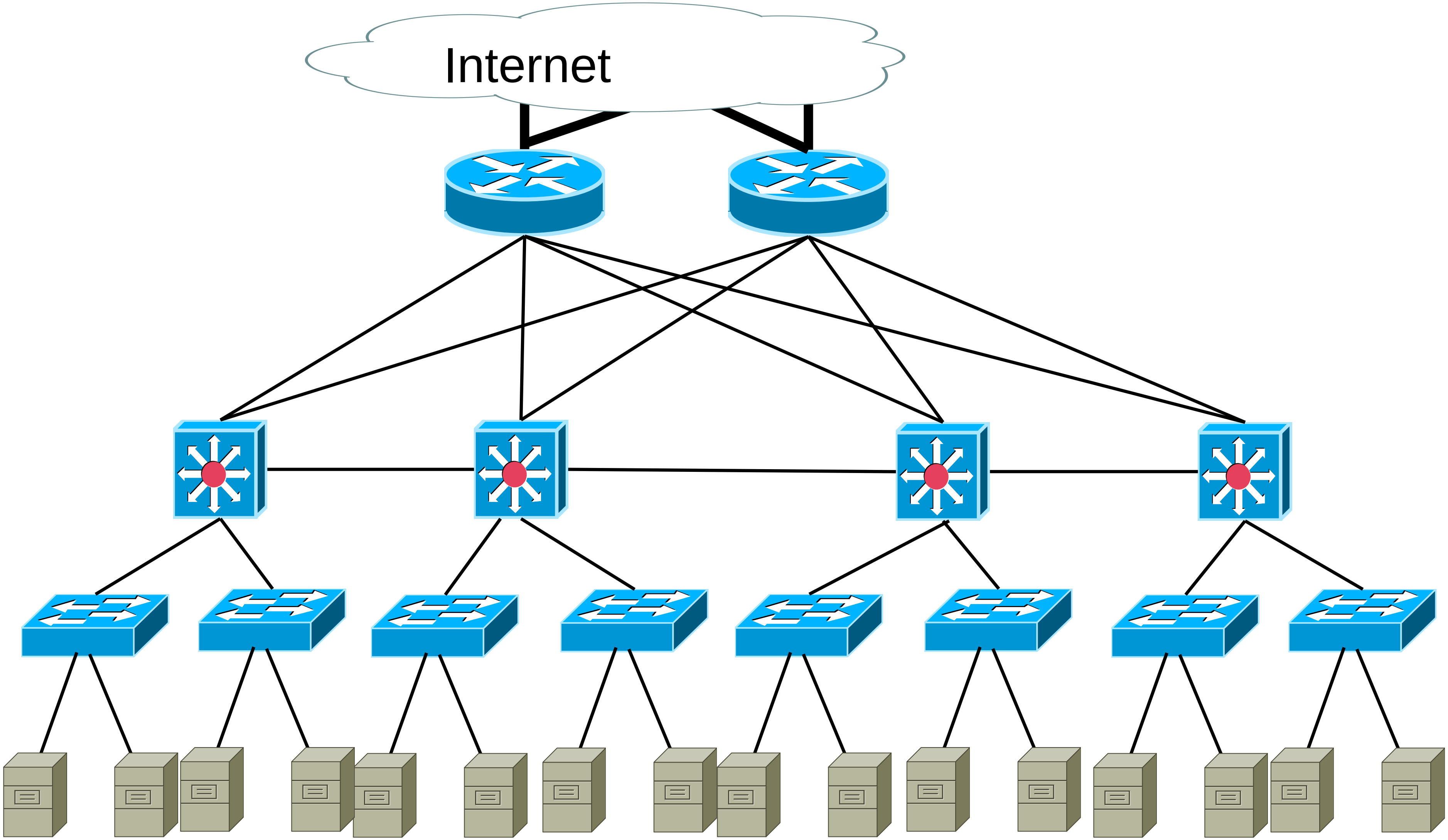


Multi-Tier Model



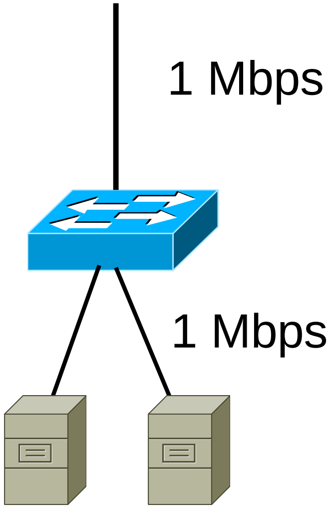


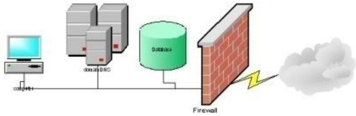
Multi-Tier Model



Oversubscription

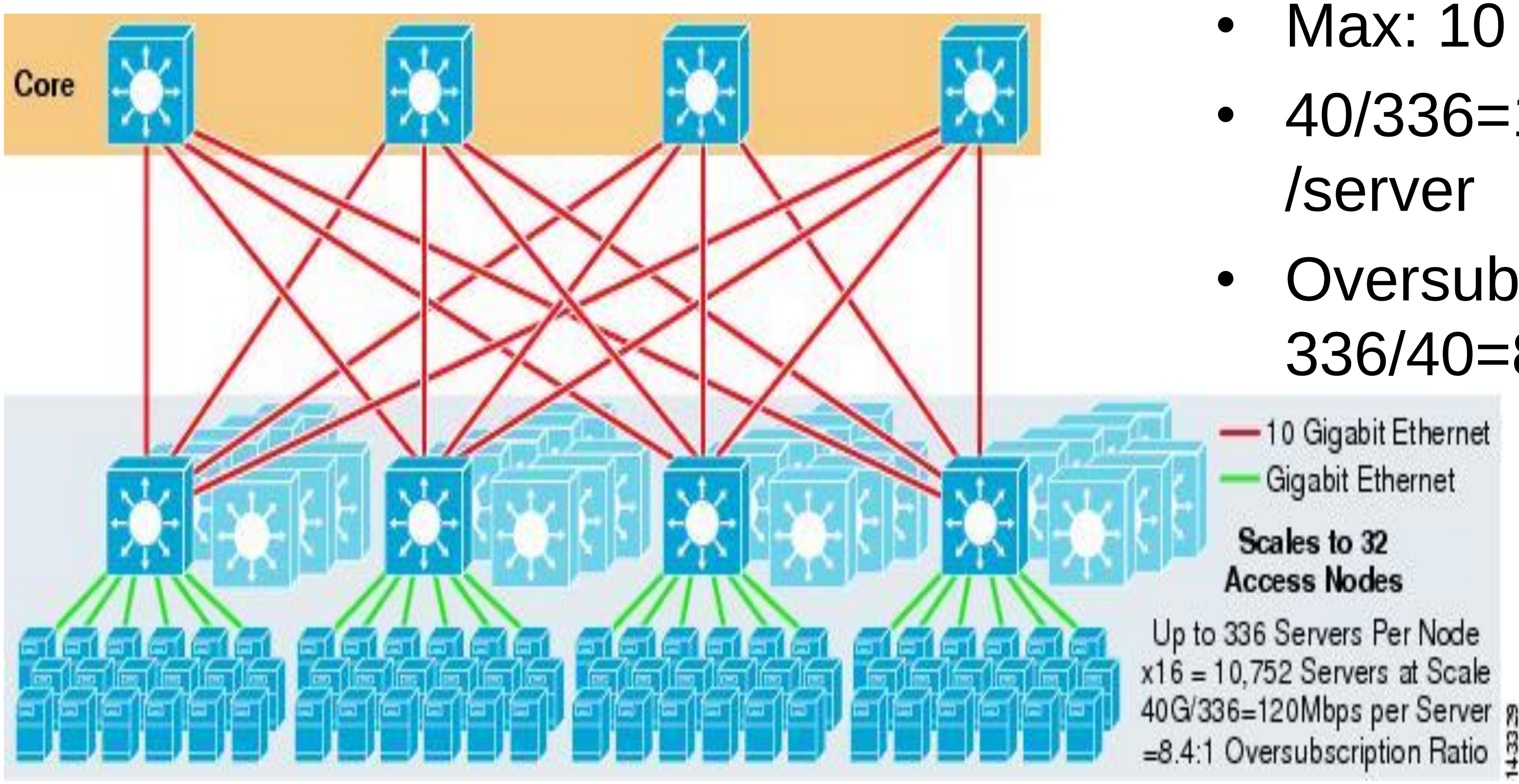
- Total output bandwidth/ input bandwidth
- 2*1 Mbps / 1 Mbps = 2
- Not likely that all bandwidth is needed at the same time





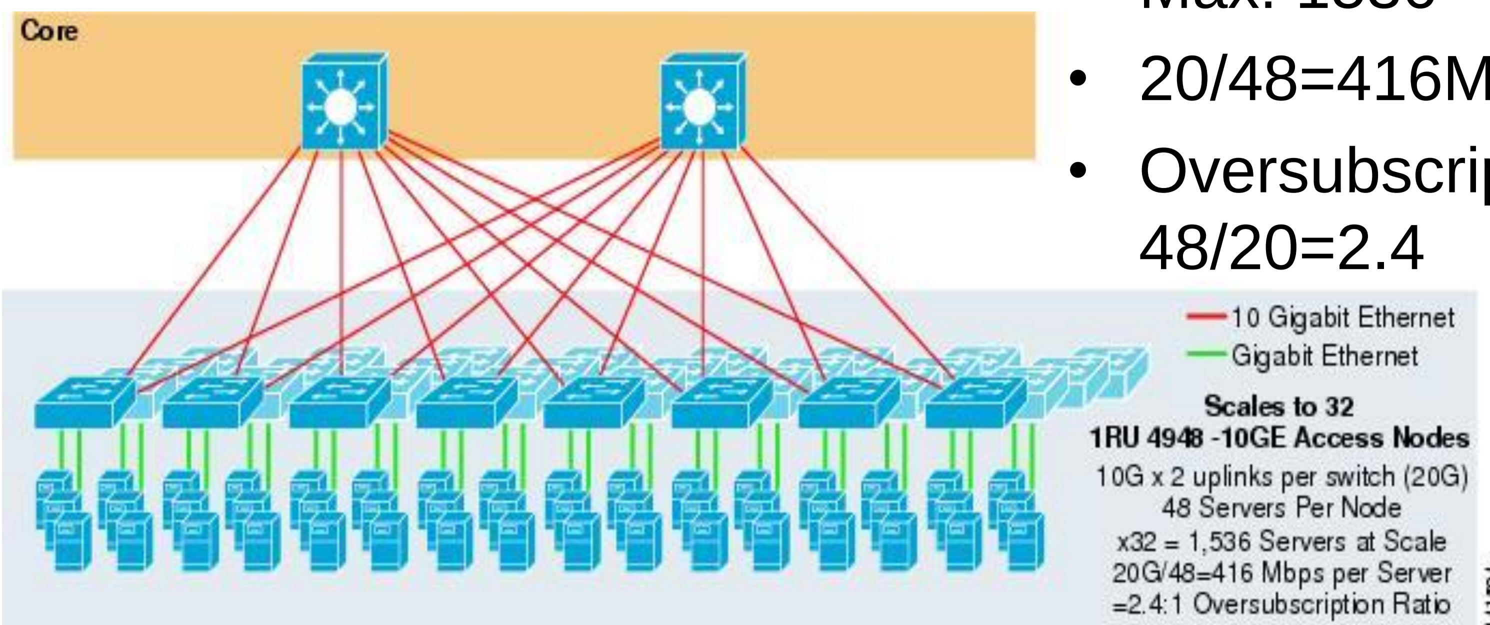
2-Tier Model

- Core: 32*10G
- Aggregation: 4*10G + 336*1G
- Max: 10 752
- 40/336=120M /server
- Oversubscription: 336/40=8.4



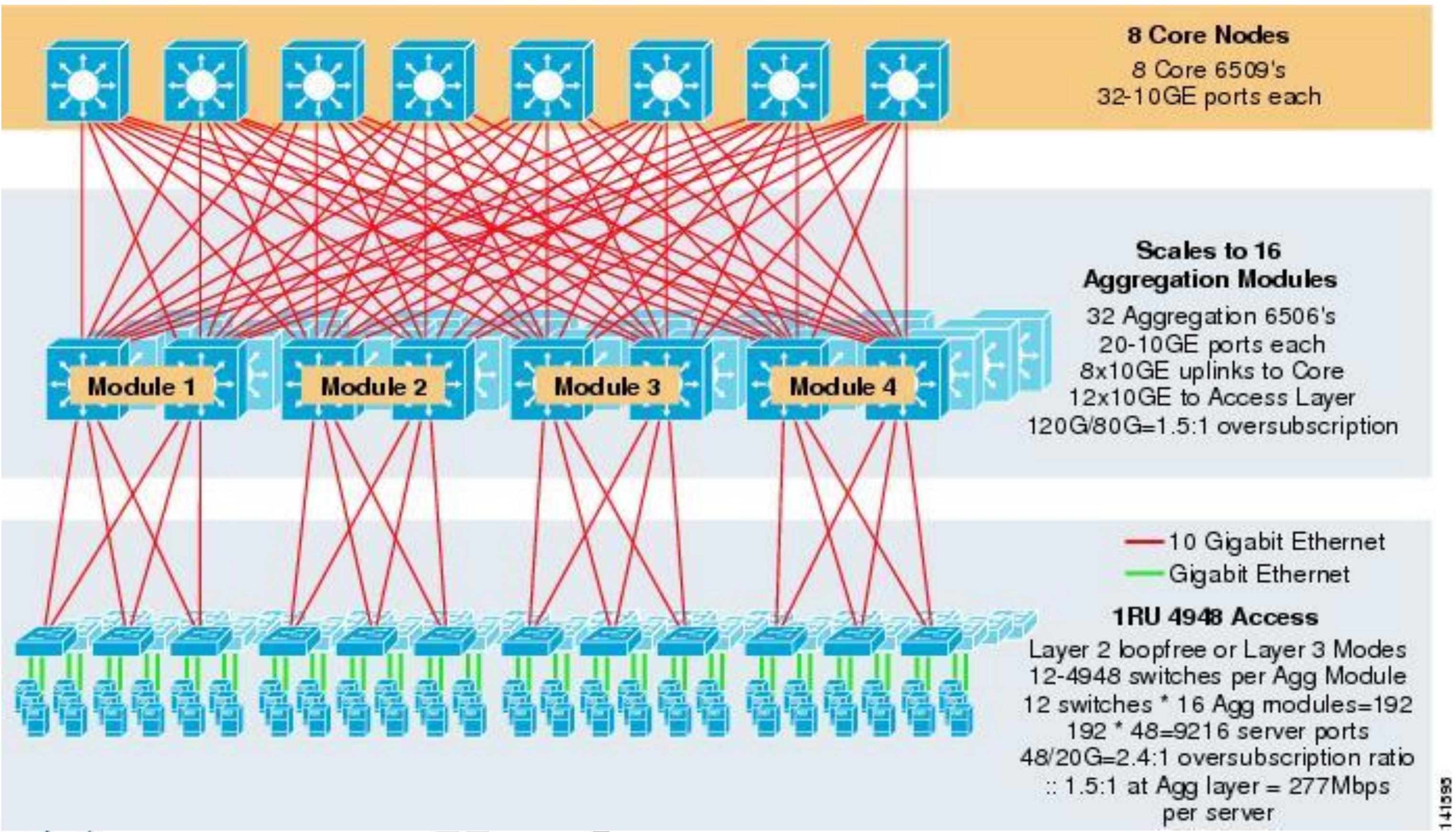
2-Tier with ToR (1RU) switches

- Core: 32*10G
- Aggregation:
  Top of Rack 2*10G
  + 48*1G
- Max: 1536
- 20/48=416M /server
- Oversubscription:
  48/20=2.4



3-Tier Model

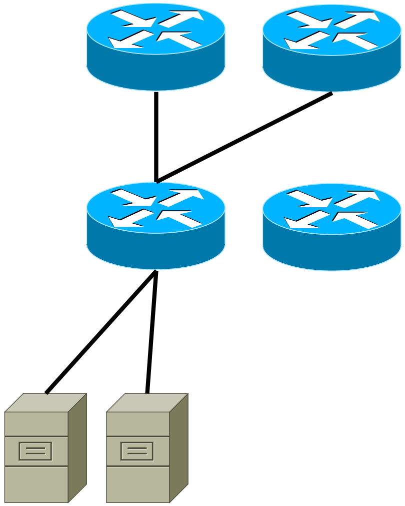
- (2*80)/(12*48) = 277M/server
- Overs.: 1.5*2.4=3.6



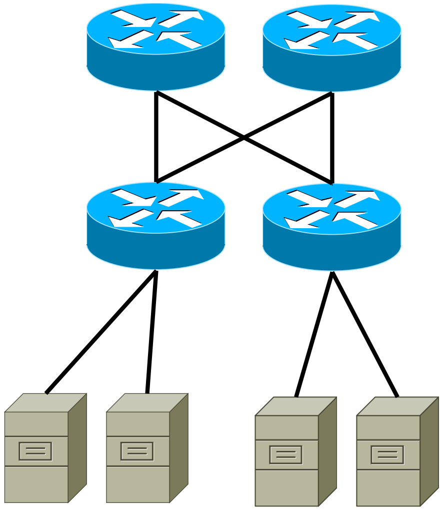
Multi-Tier Models

- Large, special routers/switches
- –> Smaller, commercial
- –> Fat Tree Topology

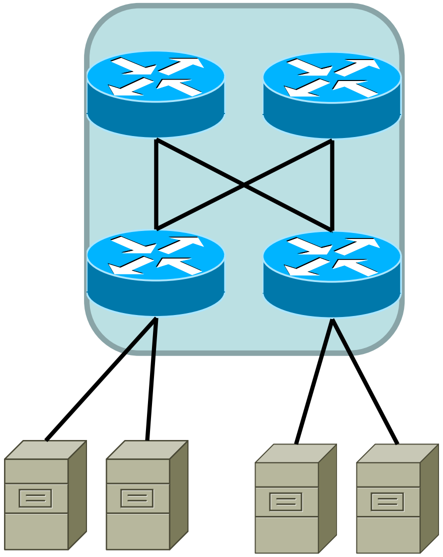
Fat Tree Topology



Fat Tree Topology

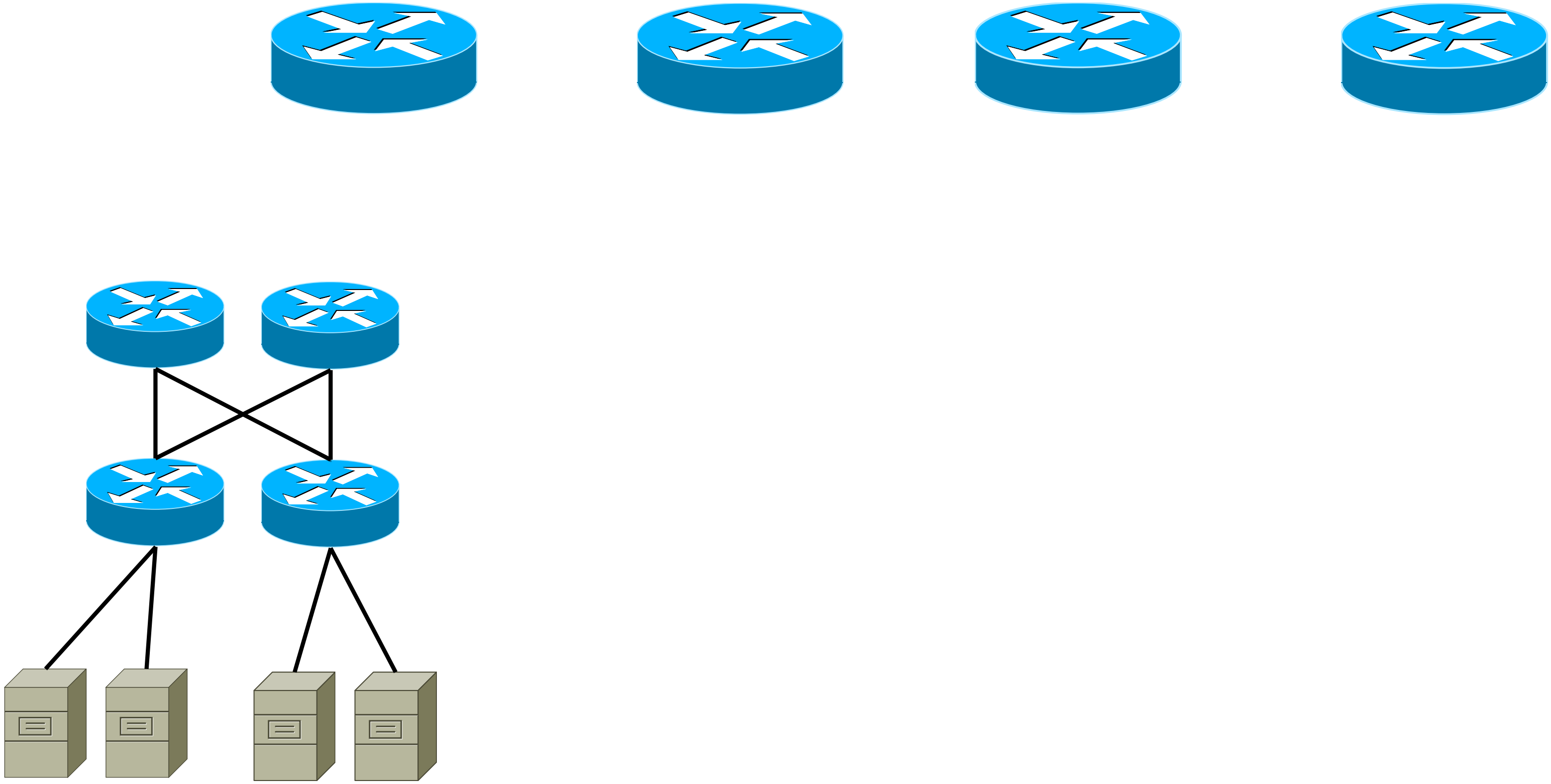


Fat Tree Topology

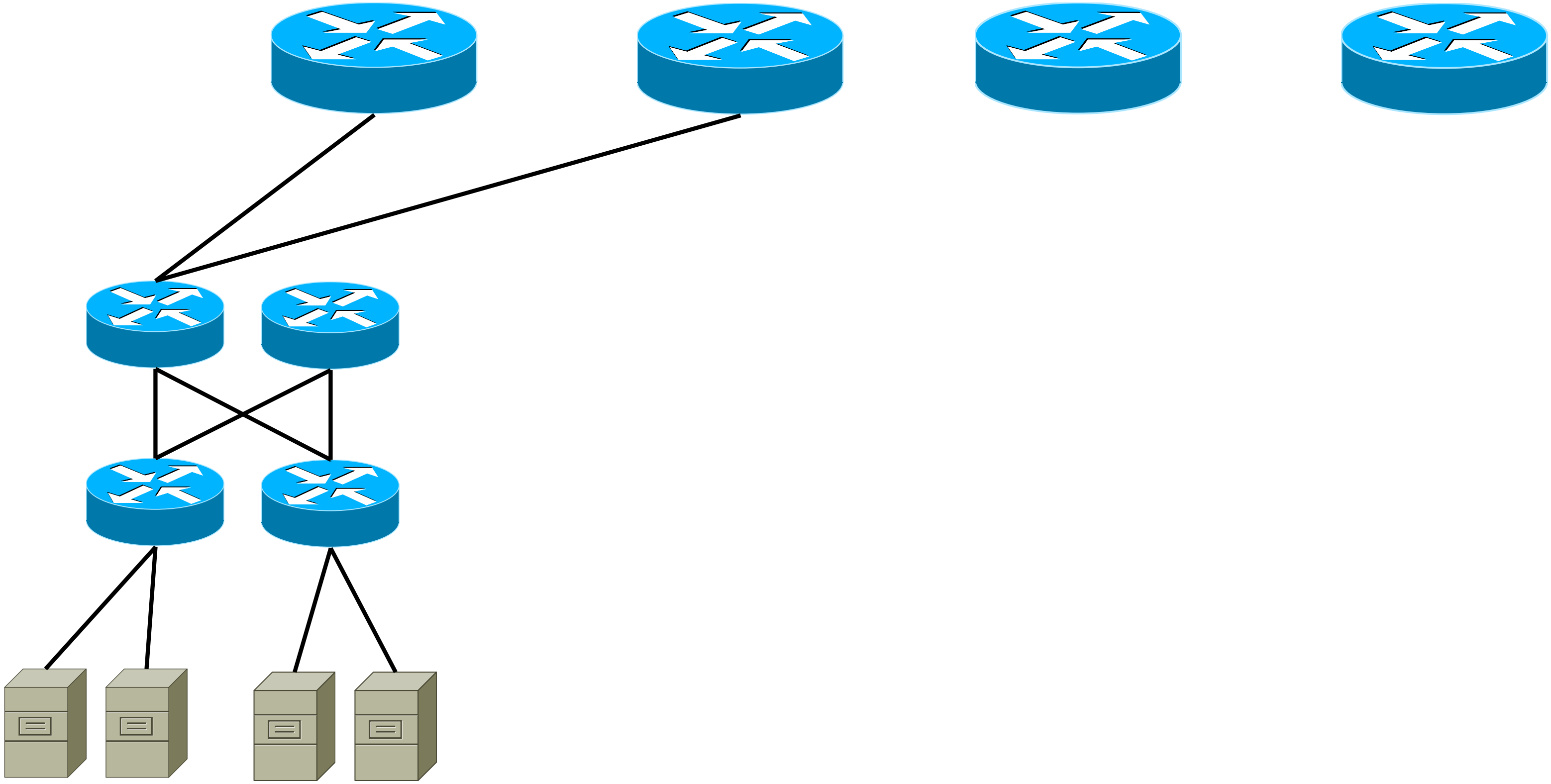


Pod

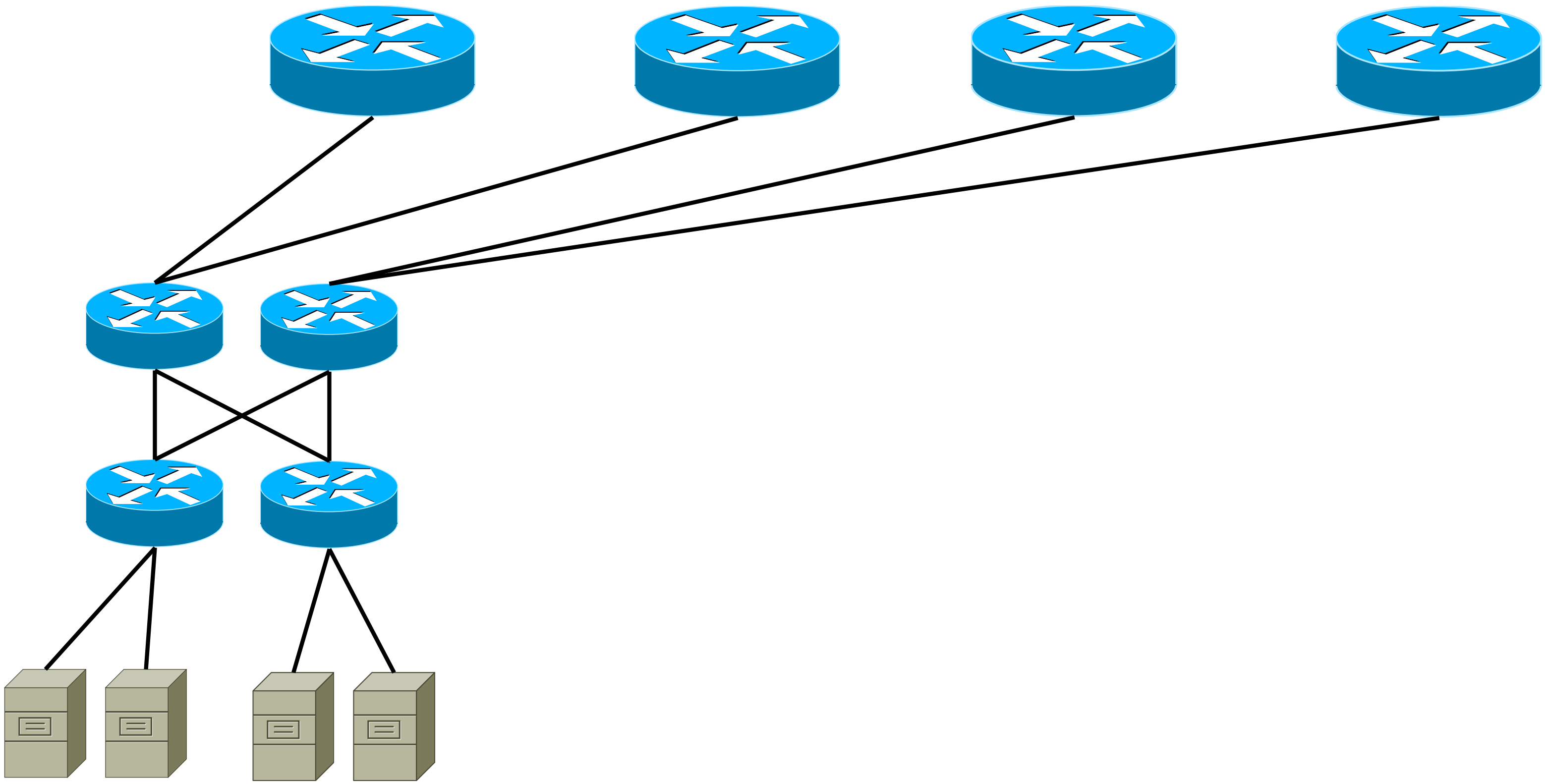
Fat Tree Topology



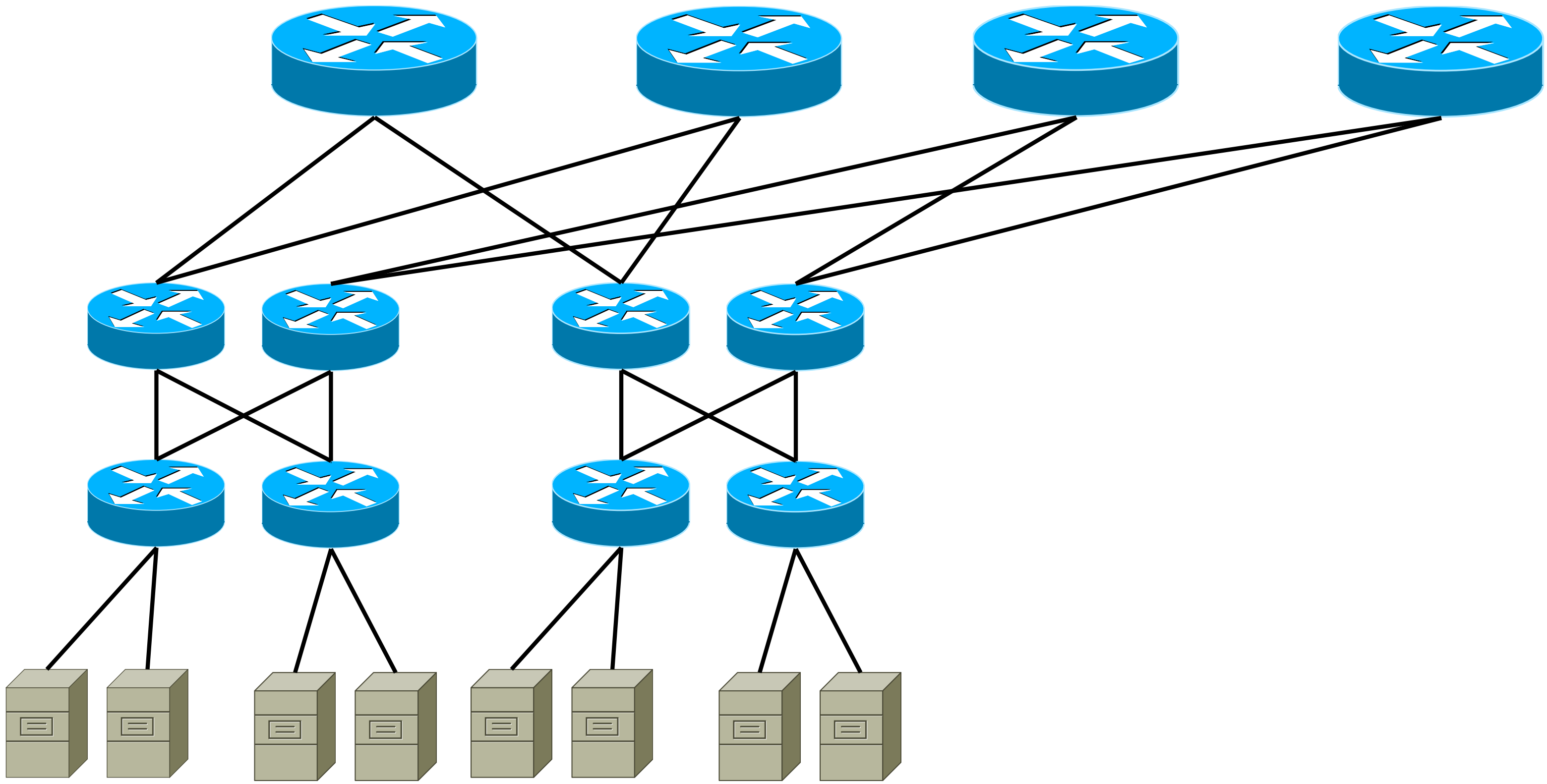
Fat Tree Topology



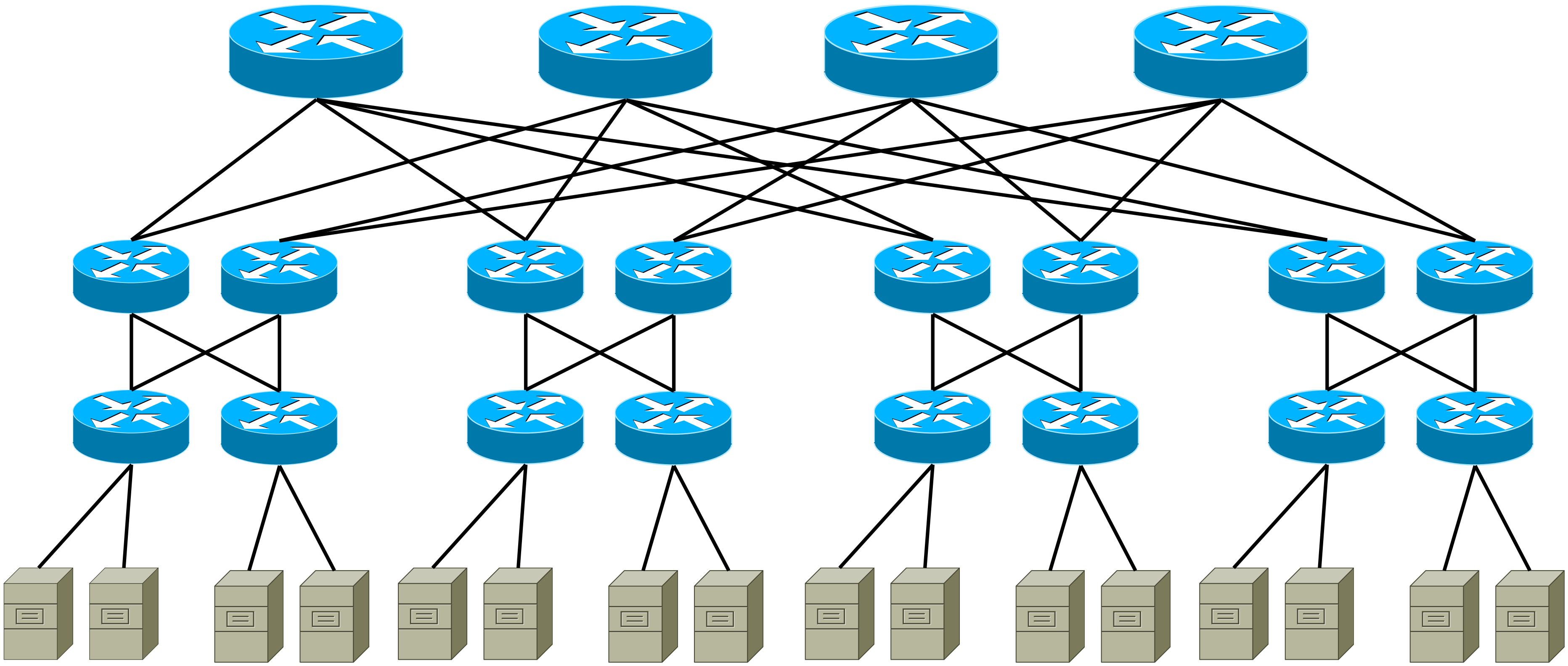
Fat Tree Topology



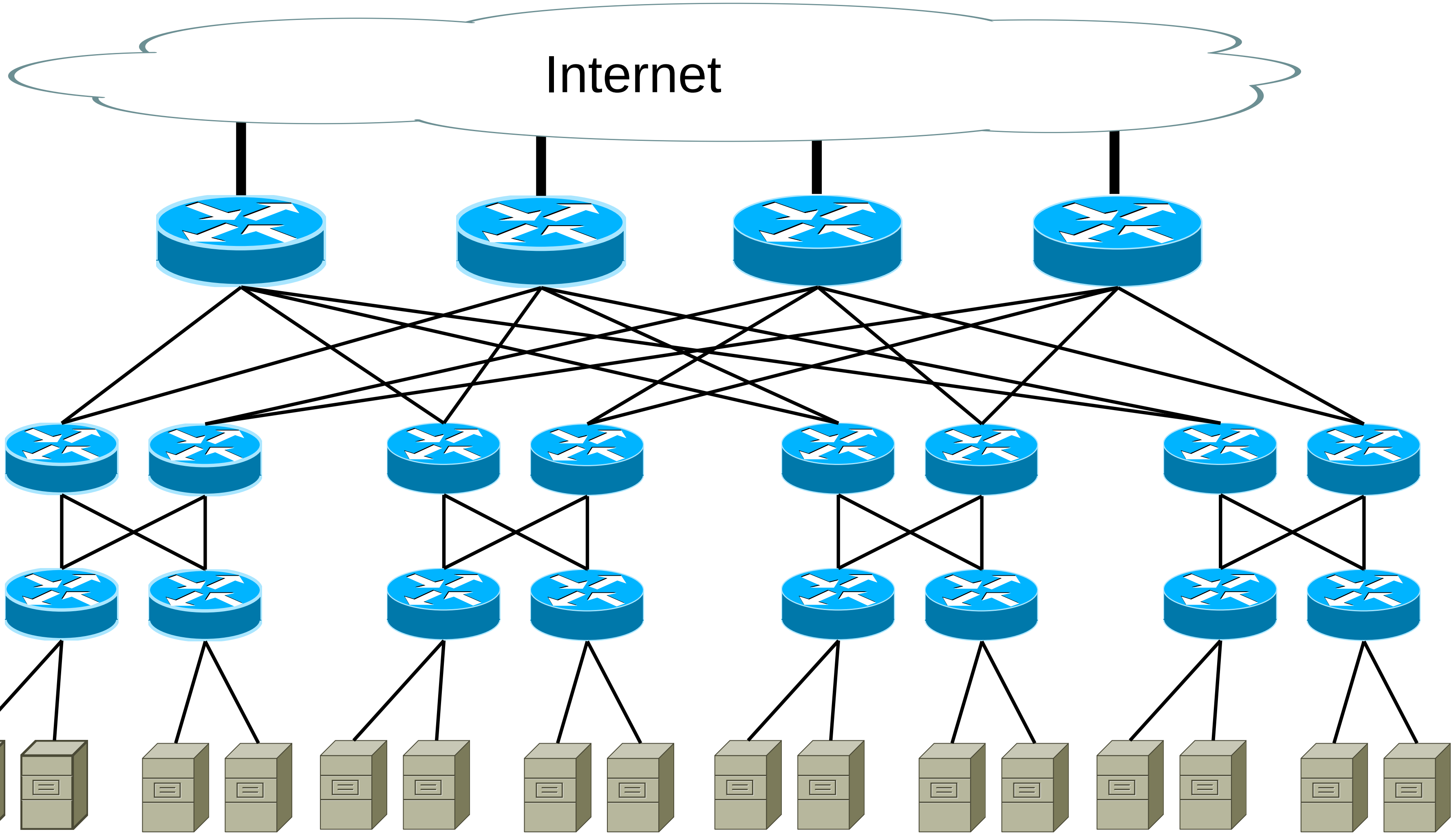
Fat Tree Topology



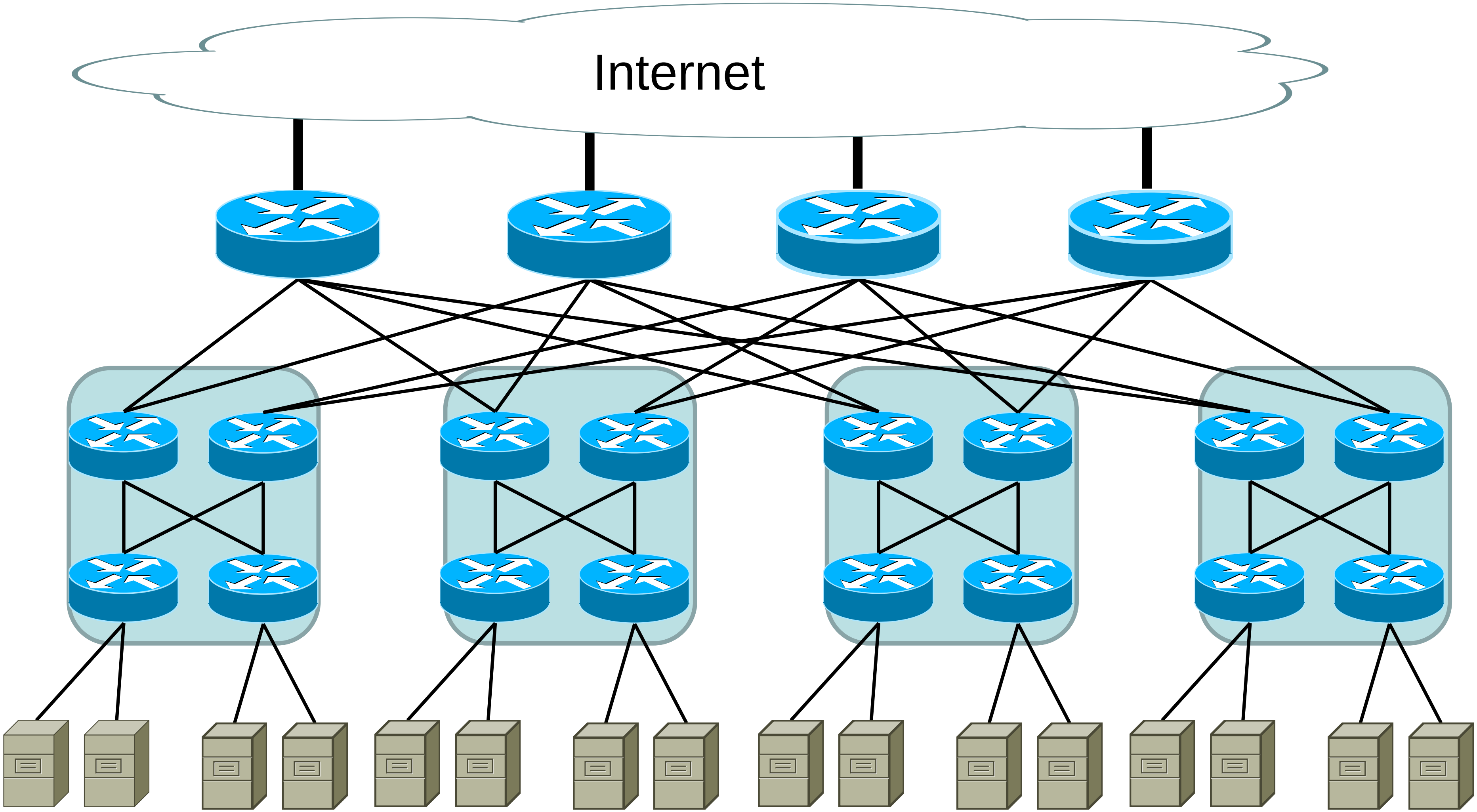
Fat Tree Topology



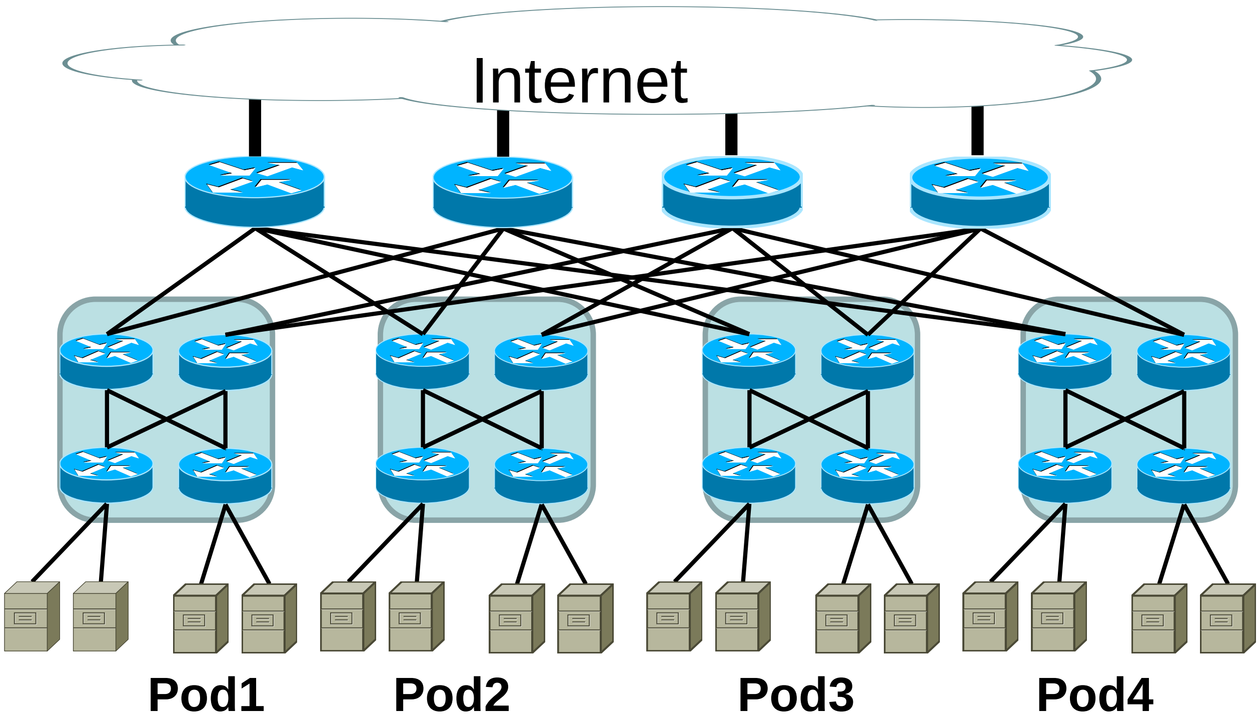
Fat Tree Topology



Fat Tree Topology



Fat Tree Topology

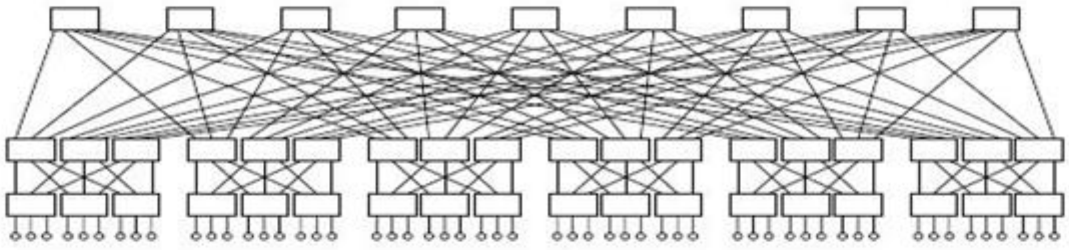


- 1 switch: k ports
- k pods
- 1 pod contains:
  - k/2 aggregate switches
  - k/2 access switches
  - (k/2)*(k/2) = $(k/2)^2$ servers
- (k/2)*(k/2) = $(k/2)^2$ core switches

Fat Tree Topology

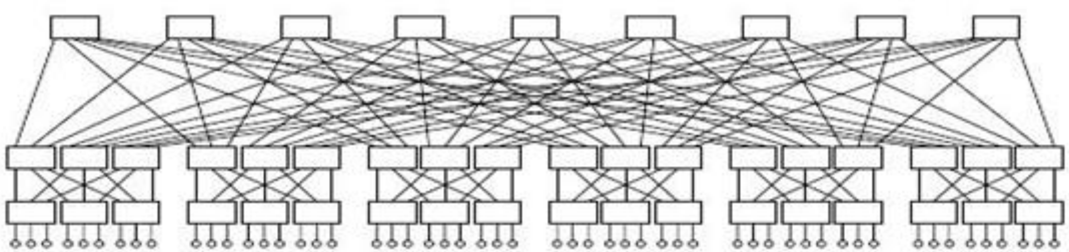
k-ary fat tree:

- three-layer topology (access, aggregation and core)
- each switch has k ports (-> k pods)
  - half of them up, other half down
- each access switch connects to k/2 servers & k/2 aggregation switches
- each aggregation switch connects to k/2 access & k/2 core switches
- $(k/2)^2$ core switches: each connects to k pods
- each pod consists of $(k/2)^2$ servers & 2 layers of k/2 k-port switches
- altogether $k * (k/2)^2 = k^3/4$ servers



6-ary Fat Tree Topology

- each switch has 6 ports -> 6 pods
- each access switch connects to 6/2 = 3 servers & 6/2= 3 aggregation switches
- each aggregation switch connects to 6/2 = 3 access & 6/2 = 3 core switches
- $(6/2)^2$ = 9 core switches: each connects to 6 pods
- each pod consists of $(6/2)^2$ = 9 servers & 2 layers of 6/2 = 3 6-port switches
- altogether: 6 * 2 * (6/2) = 36 access+aggregation switches
- altogether 6 * $(6/2)^2$ = $6^3/4$ = 54 servers

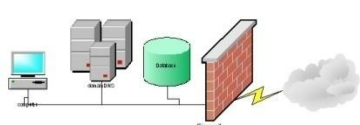


64-ary Fat Tree Topology

- Number of servers?
- $k * (k/2)^2 = k^3/4$ servers $= 64 * 32^2 = 65536$ servers

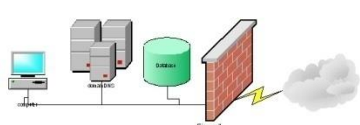
Evaluation of Fat Tree Topology

- Bandwidth
  - Fat tree has identical bandwidth at any bisections
  - Each layer has the same aggregated bandwidth
    - Oversubscription = 1
- Can be built using cheap devices with uniform capacity
  - Each port supports same speed as end host
  - All devices can transmit at line speed if packets are distributed uniform along available paths
- Great scalability: k-port switch supports $k^3/4$ servers
- Smaller power consumption
  - Lower heat / air condition



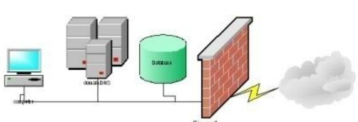
Problems of the Traditional Architectures

- Change in the traffic pattern
  - Traditionally: north-south
  - Now: east-west as well
    - Why?
    - Virtualisation / Cloud
      - Any (virtual) server can be placed at any physical
      - Reallocation of workload
      - Communication is required between them: SAME speed between ANY cluster
      - Higher reliability

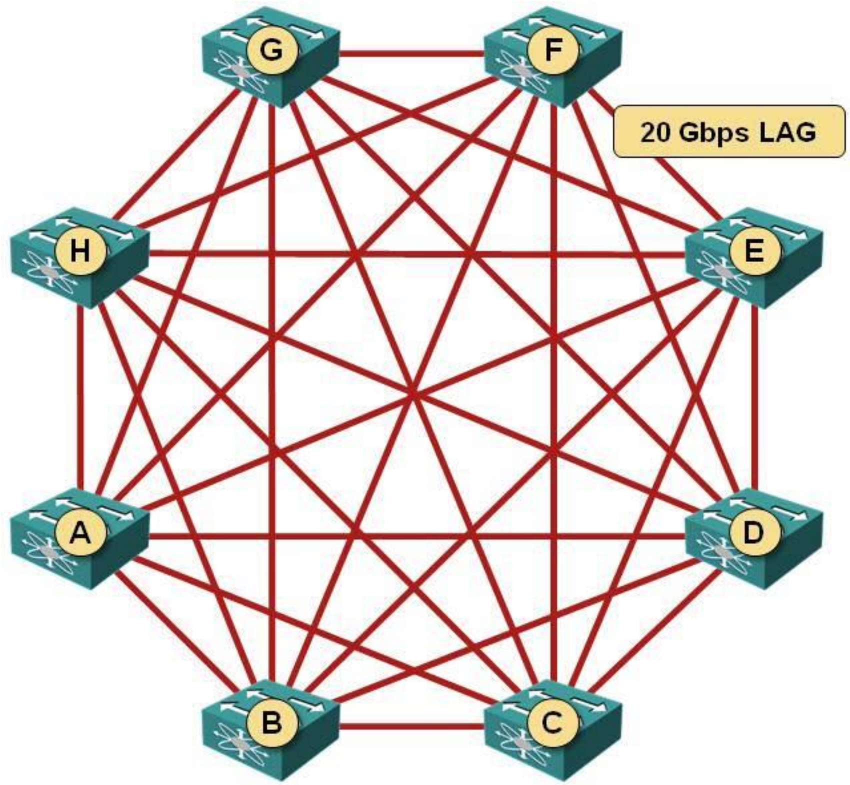


Full Mesh

- Typical data center switches:
  - 48 x 10GE ports and
  - 4 x 40GE ports that can be used as 16 x 10GE ports
  - Altogether 64*10GE ports
- Use:
  - 48 ports for servers
  - 16 (14) ports for intra fabric connections
- Oversubscription =
  - 3:1



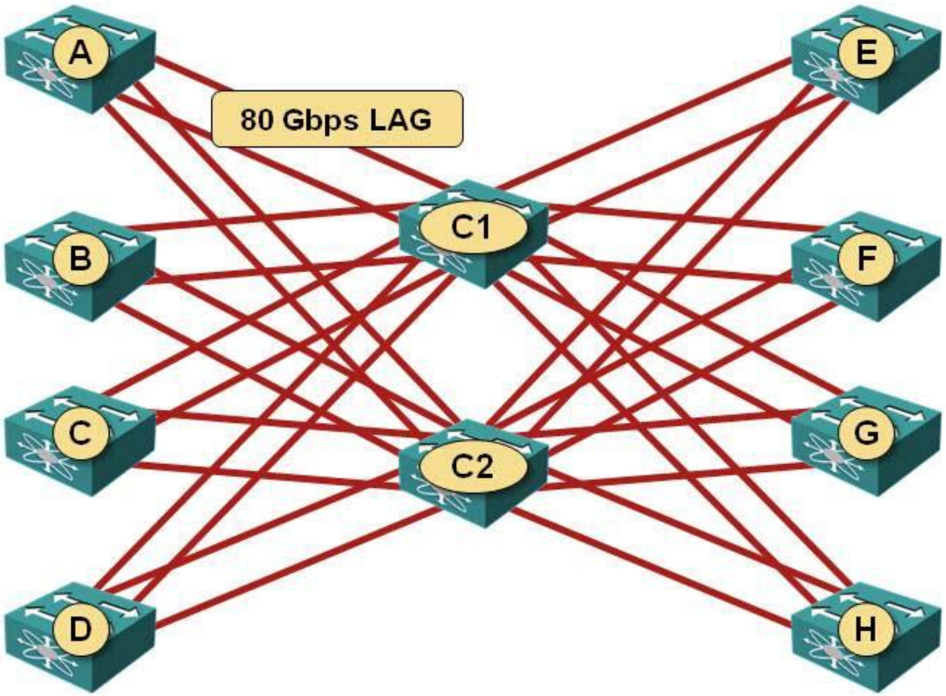
Full Mesh



- Though we have 140 Gbps uplink capacity, we can achieve only 20 Gbps between any two nodes
- No alternative routes
  - Error prone
- Lot of (n*(n-1)/2) links

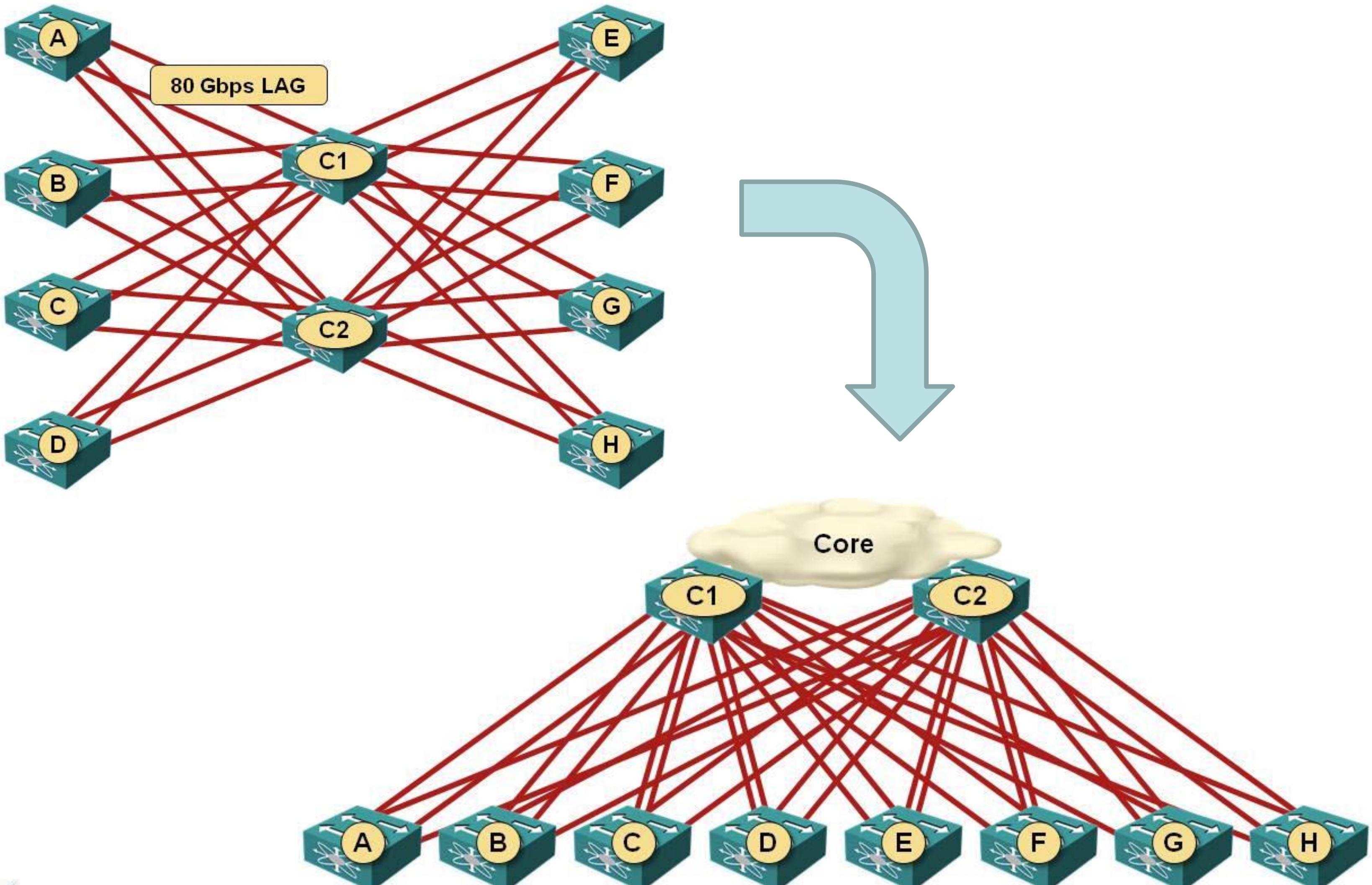
- LAG: Link aggregation

More Efficient Architecture



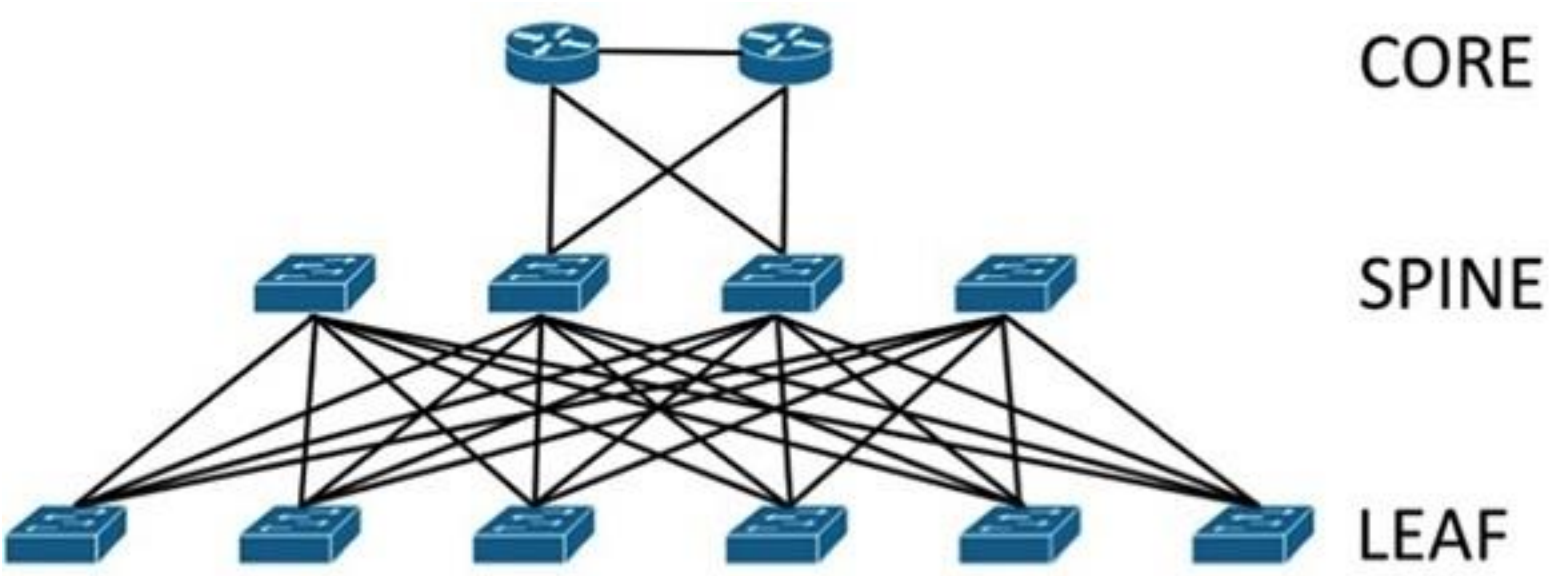
- 160 Gbps between any two nodes
- Redundant
- BUT:
  - more switches
  - slower

Draw it in a modified way…



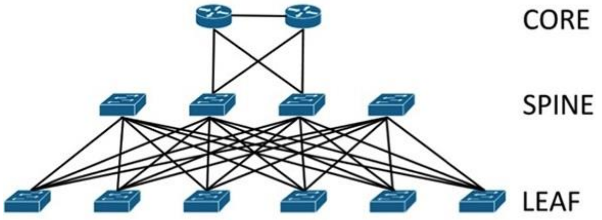
New trend

- Clos Network / Spine and Leaf architecture

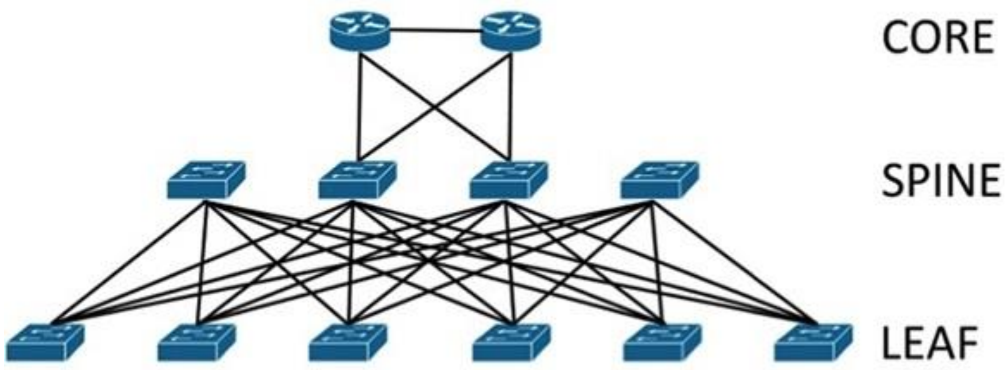


Spine and Leaf architecture

- Leaf ~ access
- Spine ~ aggregation
- But: each leaf is connected to each spine
  - – Ethernet fabric
  - – Maximum distance: 3 hops (4 links)
    - High performance clusters (HPC)



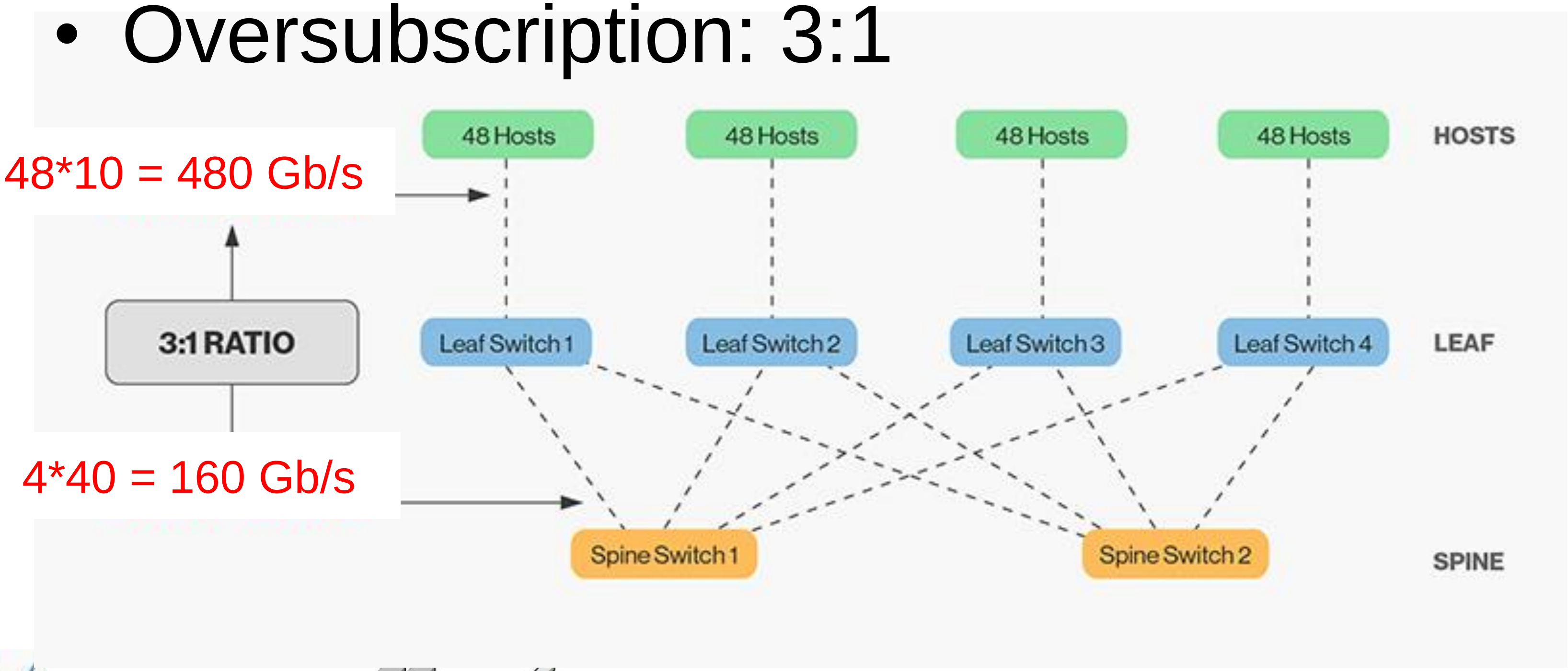
Spine and Leaf rules



- Total # of interconnections = #leaf * #spine
- #spine = #port_of_a_leaf
- #port_of_a_spine = #leaf

Typical architecture

- Leaf:
  - To Servers: 48 * 10 GbE ports
  - To Spine: 4 * 40 GbE ports
- Oversubscription: 3:1

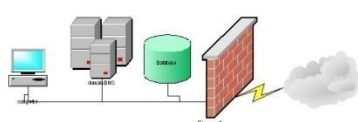


Load Balancing, Parallel Paths

- Traditional routing algorithm – Spanning Tree Protocol (STP)
  - To avoid loops finds the best path to every node
  - What is the main disadvantage?
    - Only ONE path between any two nodes
- Equal-Cost Multipath (ECMP) protocol
  - Finds more paths with the same length (cost)
    - In practice typically 8 or 16
    - Load distribution
    - Higher bandwidth utilization
    - Packet reordering

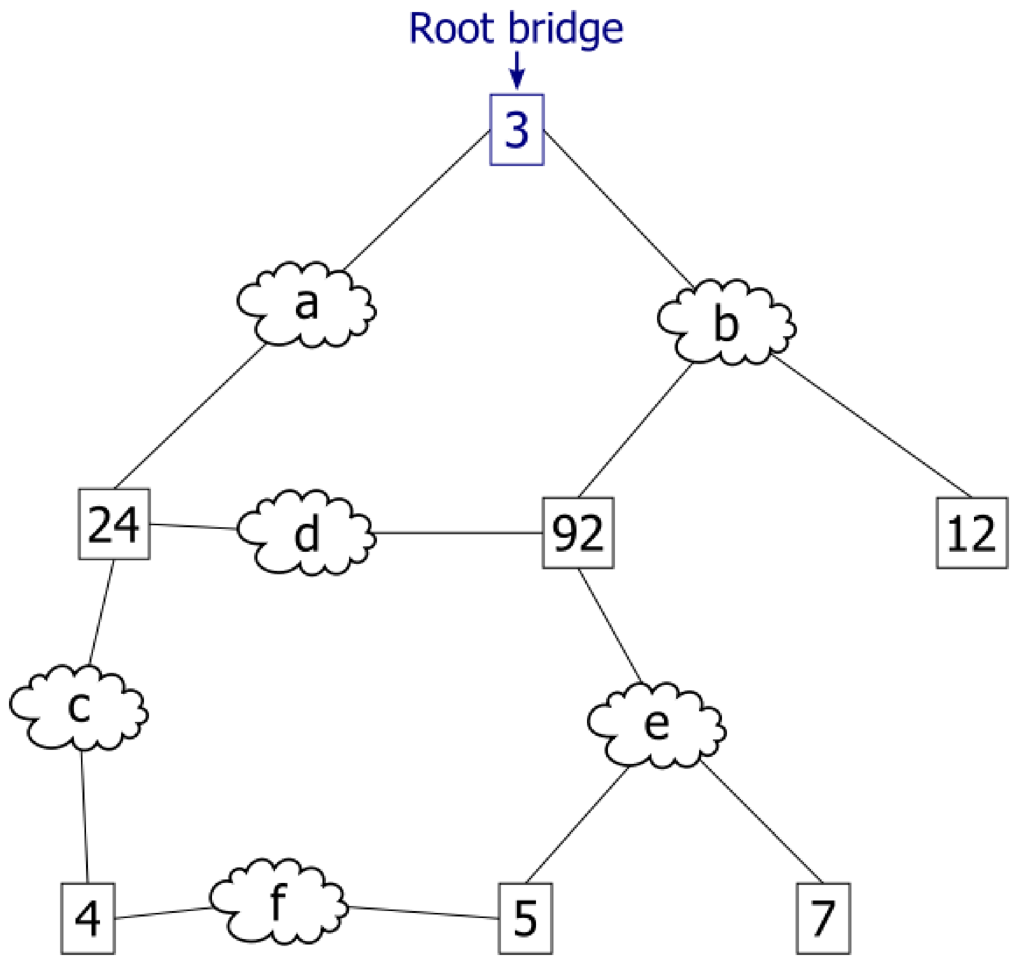
Spanning Tree Protocol (STP)

- Layer 2 protocol
  - runs on bridges and switches.
  - IEEE 802.1D (original)
  - IEEE 802.1Q-2014 (incl. different new versions)
- Builds a logical loop-free topology for Ethernet networks



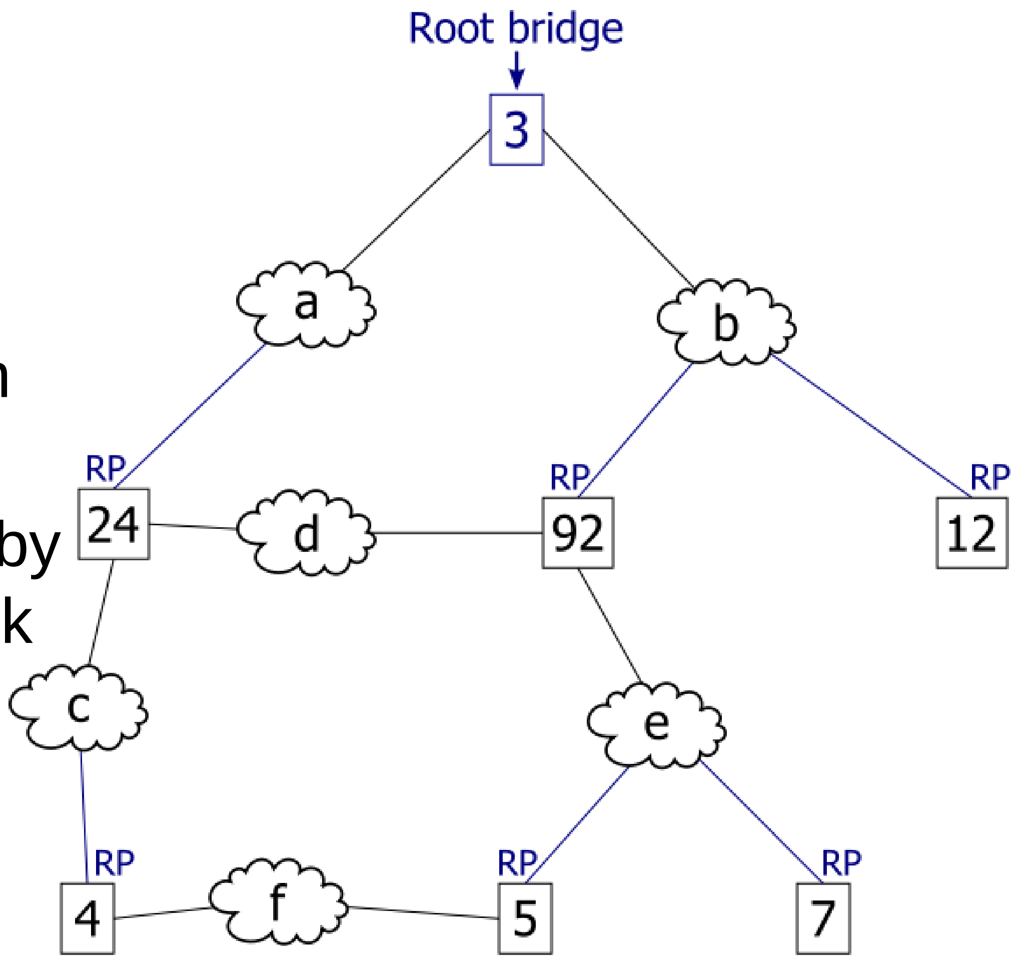
Spanning Tree Protocol (STP)

- Select a root bridge/switch
  - With smallest ID
  - ID: priority (default 32768) + MAC address
  - Good if central, no end station connected (backbone switch)



Spanning Tree Protocol (STP)

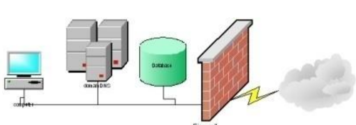
- Least cost path from each bridge
  - Bridge protocol data units (BPDUs) transmitted by root with cost=0
  - Others increment cost by the cost of incoming link (all the same on the figure)
  - Find the smallest
  - Root Port (RP)
  - When tie: from bridge with smaller ID



Default costs

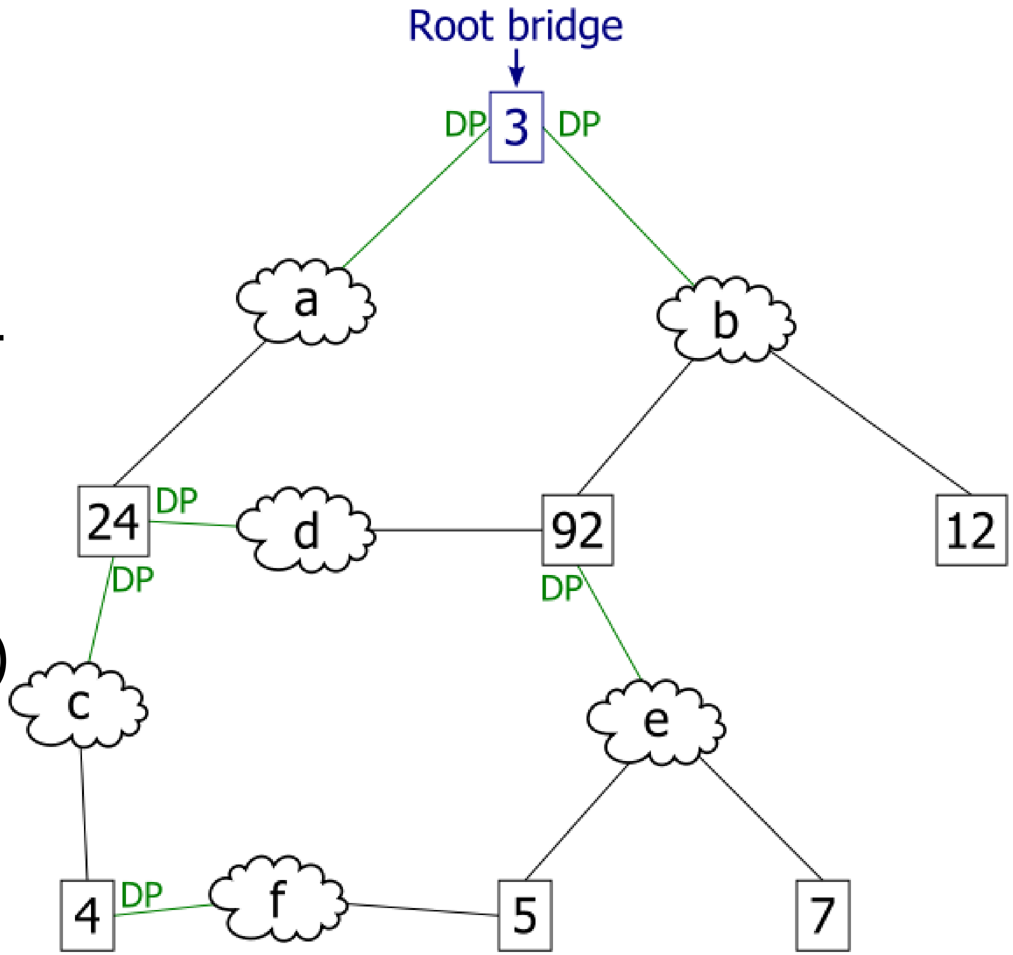
Data rate	STP cost (802.1D-1998)	RSTP cost (802.1W-2004, default value)
4 Mbit/s	250	5,000,000
10 Mbit/s	100	2,000,000
16 Mbit/s	62	1,250,000
100 Mbit/s	19	200,000
1 Gbit/s	4	20,000
2 Gbit/s	3	10,000
10 Gbit/s	2	2,000

1 Gigabit/second/bandwidth   20 Terabit/second/bandwidth



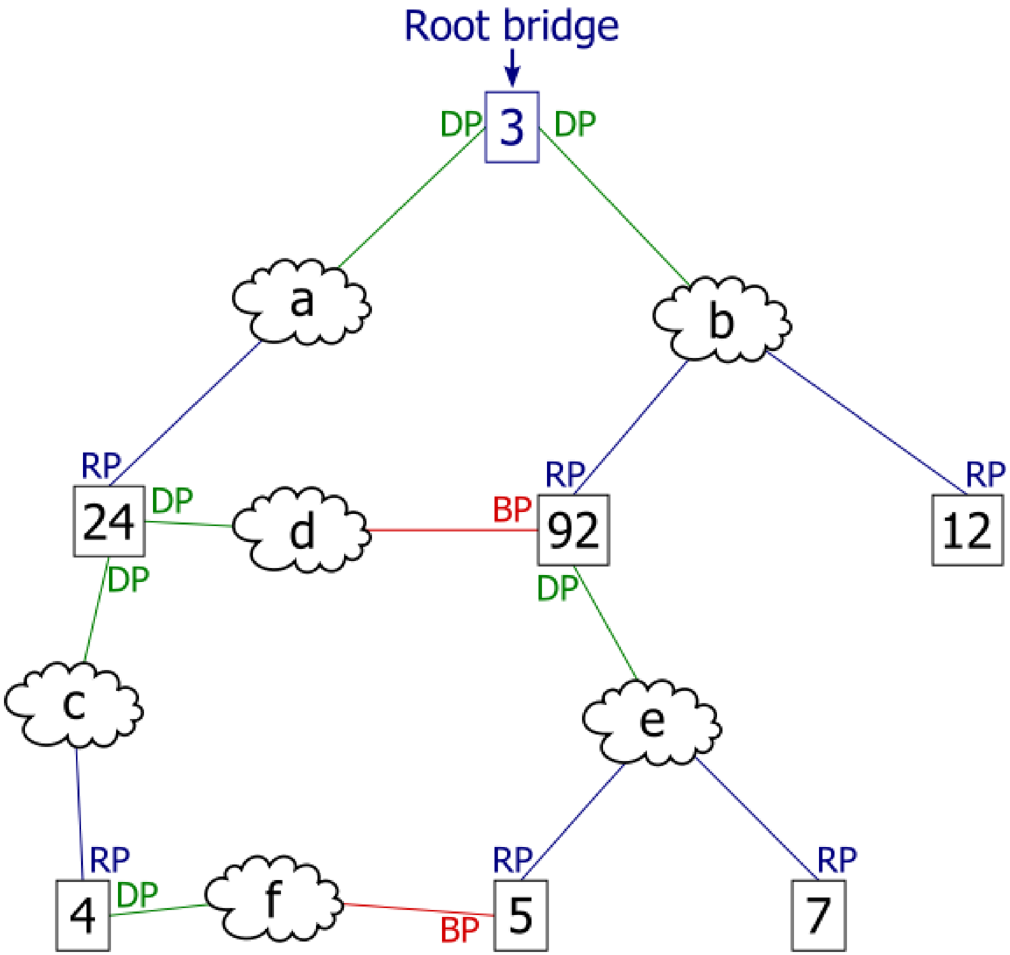
Spanning Tree Protocol (STP)

- Least cost path from each network segment
  - Bridge with the least-cost path from the network segment to the root
  - Designated *port* (DP) for the segment
  - When tie: from bridge with smaller ID (see d)



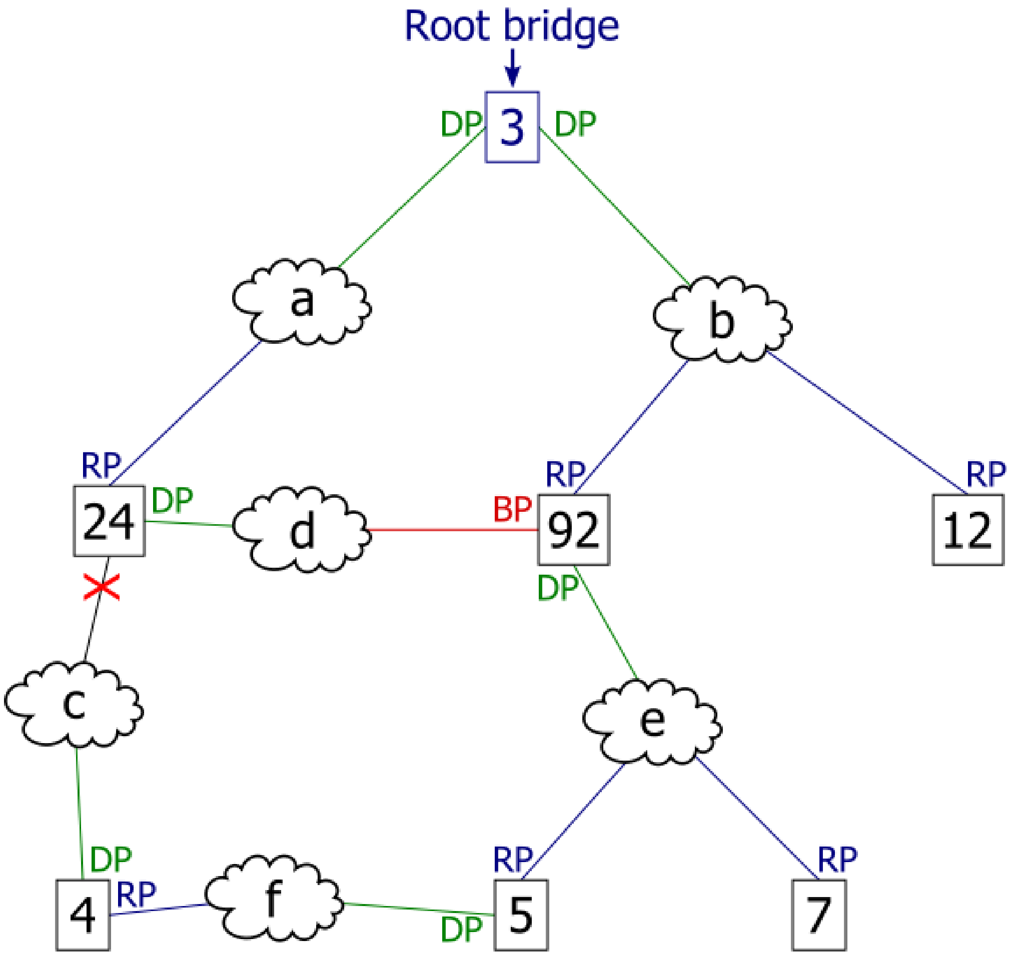
Spanning Tree Protocol (STP)

- The DPs will be set to forwarding mode
- The rest: blocked (BP)



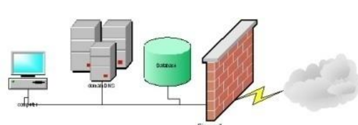
Spanning Tree Protocol (STP)

- At link failure:
  - reconfiguration

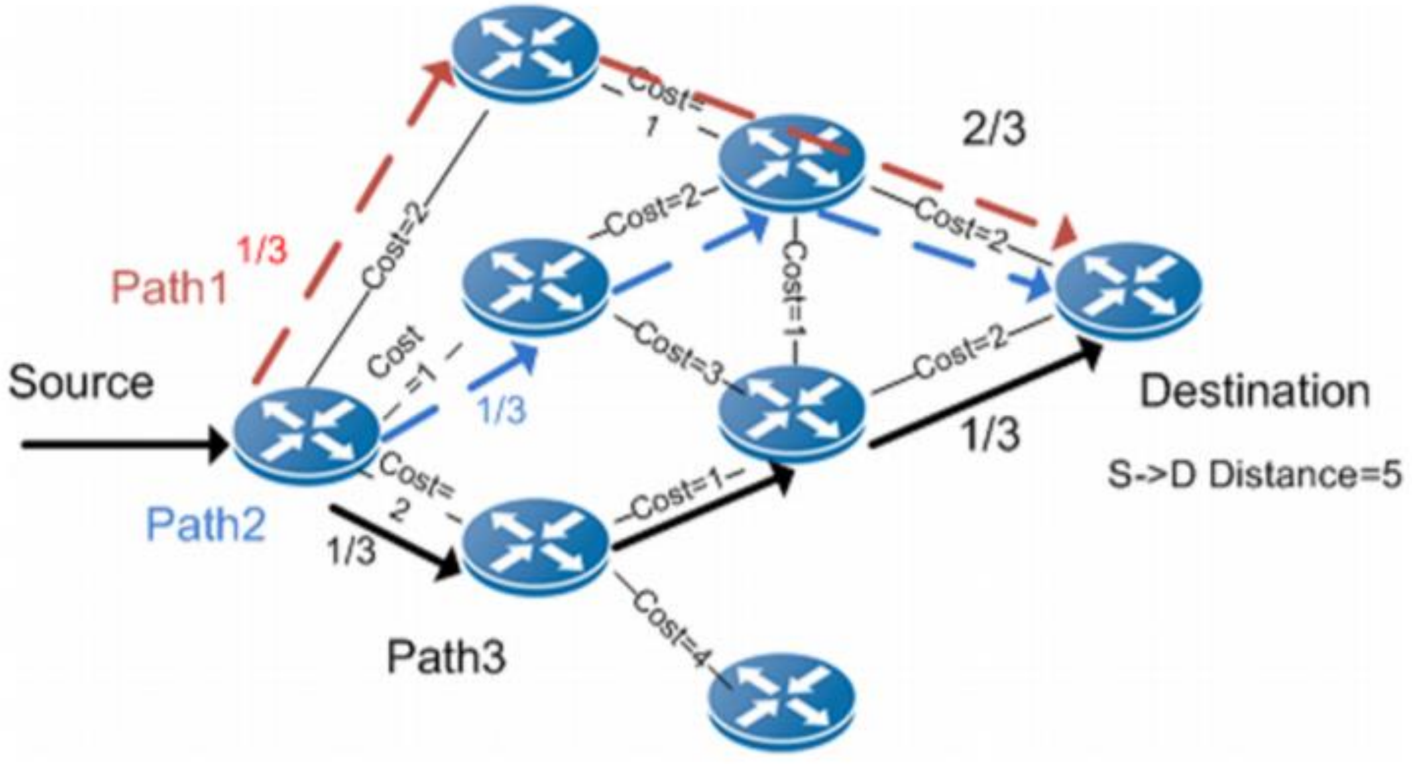


Equal-Cost Multi-Path (ECMP)

- For Layer2 and Layer3
- If n path exist between two nodes
  - Modulo n hash algorithm used to select
  - ~ remainder at division
  - Different hash algorithms
  - From header
    - Source/destination MAC/port
    - Source/destination IP+port and Protocol number
      - To direct same TCP stream onto the same link



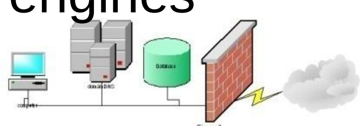
ECMP



Source
Destination
S->D Distance=5
Path1
1/3
Path2
1/3
Path3
1/3
2/3
1/3
Cost=2
Cost=1
Cost=2
Cost=2
Cost=2
Cost=1
Cost=3
Cost=2
Cost=2
Cost=1
Cost=1
Cost=4

High Performance Clusters (HPC)

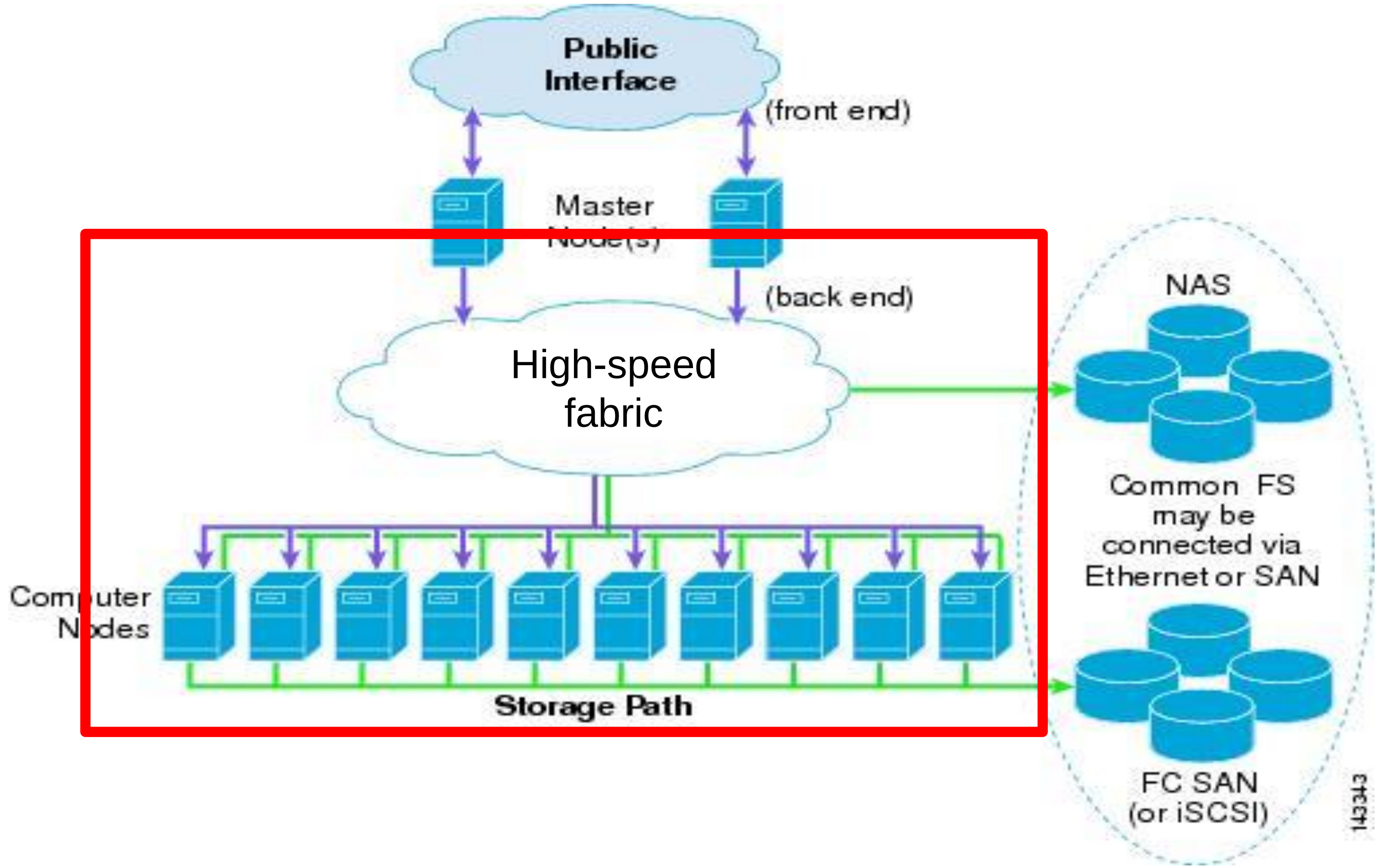
- Clusters have common goal
  - Combining multiple CPUs
  - To appear as a unified, high performance system
    - High Availability
    - Load Balancing
    - Increased Computing Power
- E.g.:
  - Meteorology (weather simulation)
  - Military research
  - Financial trending analysis
  - Film animation
  - Manufacturing (automotive/aircraft design, aerodynamic simulation)
  - Search engines



HPC types

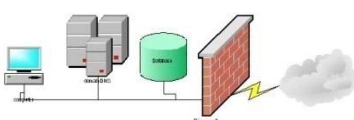
- Type 1 (tightly coupled)
  - Parallel message passing
    - Applications runs on each node in parallel
    - A Master Node determines the input for each node
    - Communication between nodes
- Type 2
  - Distributed I/O processing
    - E.g.: search engines
    - Master Node sprayes the request
- Type 3 (loosely coupled)
  - Parallel file processing
    - Source file divided up and distributed for manipulation in parallel

Logical view of a Server Cluster



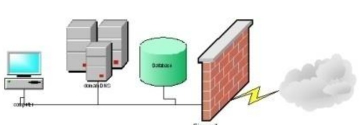
Evaluation of Spine and Leaf

- Advantages:
  - Supports east-west traffic
    - Supports virtualization (cloud)
  - High and uniform east-west (inter-cluster) speed
  - Uses all interconnection links
    - Spanning Tree Protocol (STP) -> Equal-Cost Multipath Protocol (ECMP)



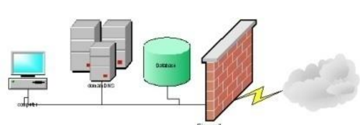
Evaluation of Spine and Leaf

- Advantages:
  - Uniform switches
    - Fix configuration switches
    - Smaller power consumption
    - Inexpensive devices -> high performance, reliable construct
      - If a switch/link fails, only slightly degrades the performance



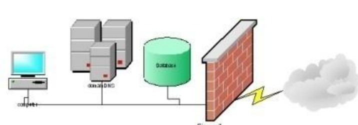
Evaluation of Spine and Leaf

- Disadvantages:
  - More switches
    - Number of ports of a switch limits the scale
  - More cables
    - When adding a new spine new cables from all leafs
    - Long cables
      - Fibre optic, optical modules (expensive)
      - Instead of coaxial cables (only for short distance, but cheap)



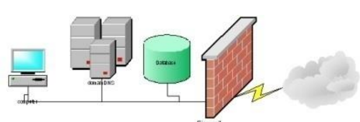
Demilitarized Zones

- Hosts most vulnerable to attack: provide services to users outside LAN
  - E-mail, web, DNS, FTP, VoIP, …
  - Place them in a separate subnetwork - Demilitarized zone (DMZ) or Perimeter Network
    - More secure than Internet, less secure than Intranet
    - Hosts in DMZ have limited connectivity to (some) internal servers and reduced/controlled communication to Internet
    - Firewall(s)
    - Protects against external attacks, but does not deal with internal attacks



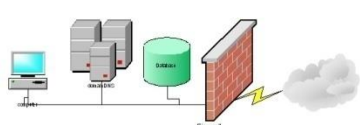
Proxy Servers

- In DMZ
  - Obliges (internal) users to use the proxy to Internet
    - Security
    - Monitoring
    - Centralized web content filtering
    - Reduces Internet traffic - caching



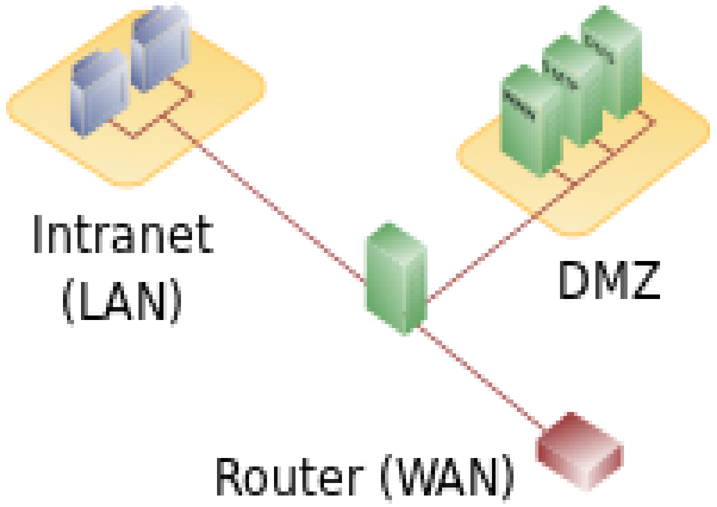
Reverse Proxy Servers

- Provides indirect access for an external network to internal resources
  - To read e-mails from outside the company
- Only the reverse proxy server can have an access to mail server
- Extra layer of security
- Typically by an application layer firewall



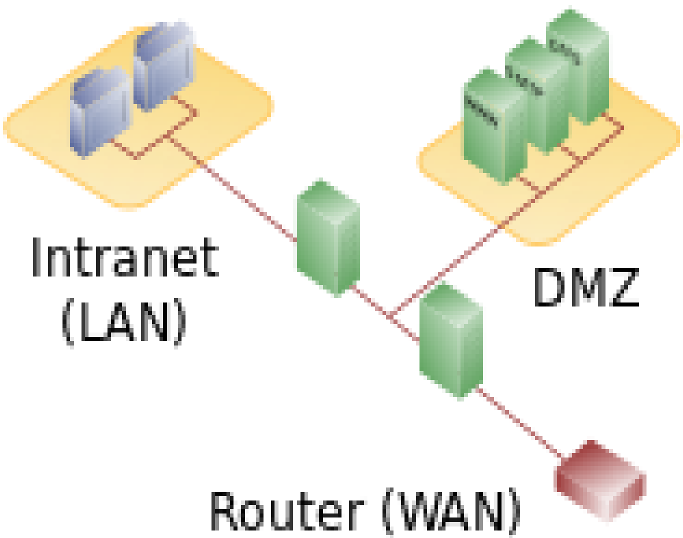
DMZ architecture I.

- Single firewall – three legged model
  - Handles traffic between DMZ-Intranet and DMZ-Internet
  - Single point of failure



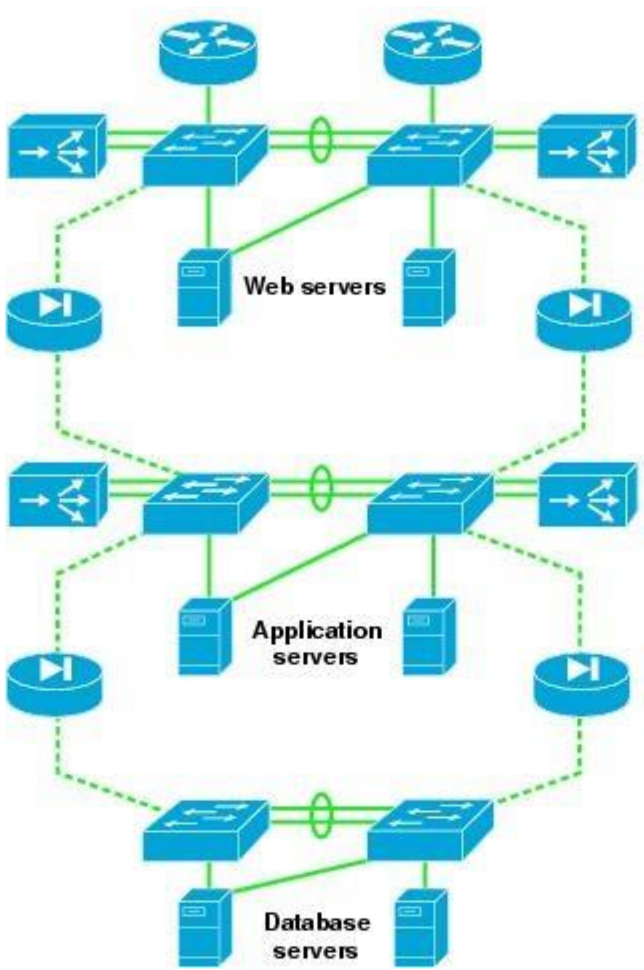
DMZ architecture II.

- Dual firewall
  – Front-end/Perimeter
  – Back-end/Internal
- Typically from different vendors
  – Not the same security holes
  – Not the same configuration methods
  – More expensive



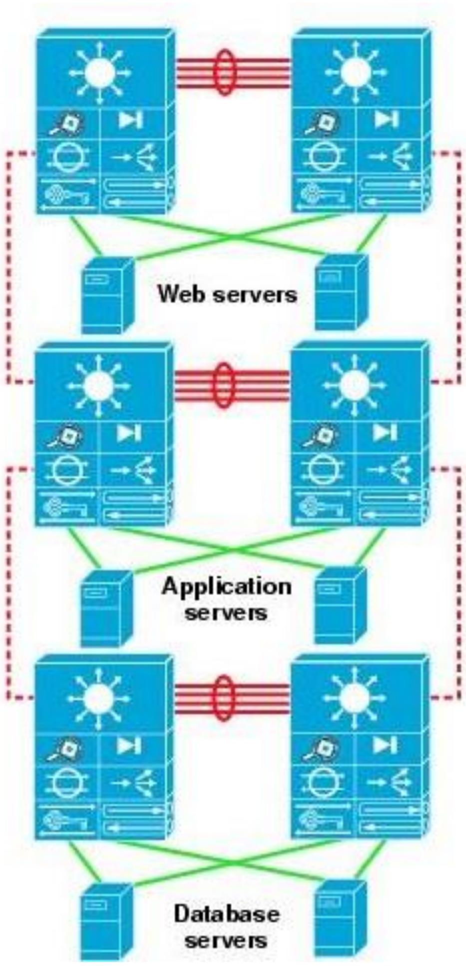
HTTP-based Applications I.

- 3 tiers
  - Web servers
  - Application servers
  - Database servers
- Separated by firewalls



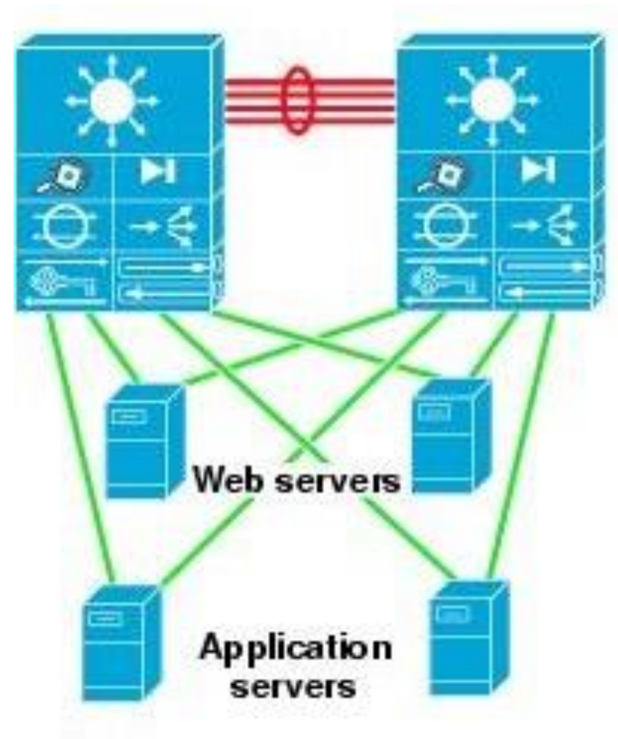
HTTP-based Applications II.

- Integrated Service Modules
  - Router/switch
  - Firewall
  - Load balancing
  - Security
  - IDS (Intrusion Detection System)

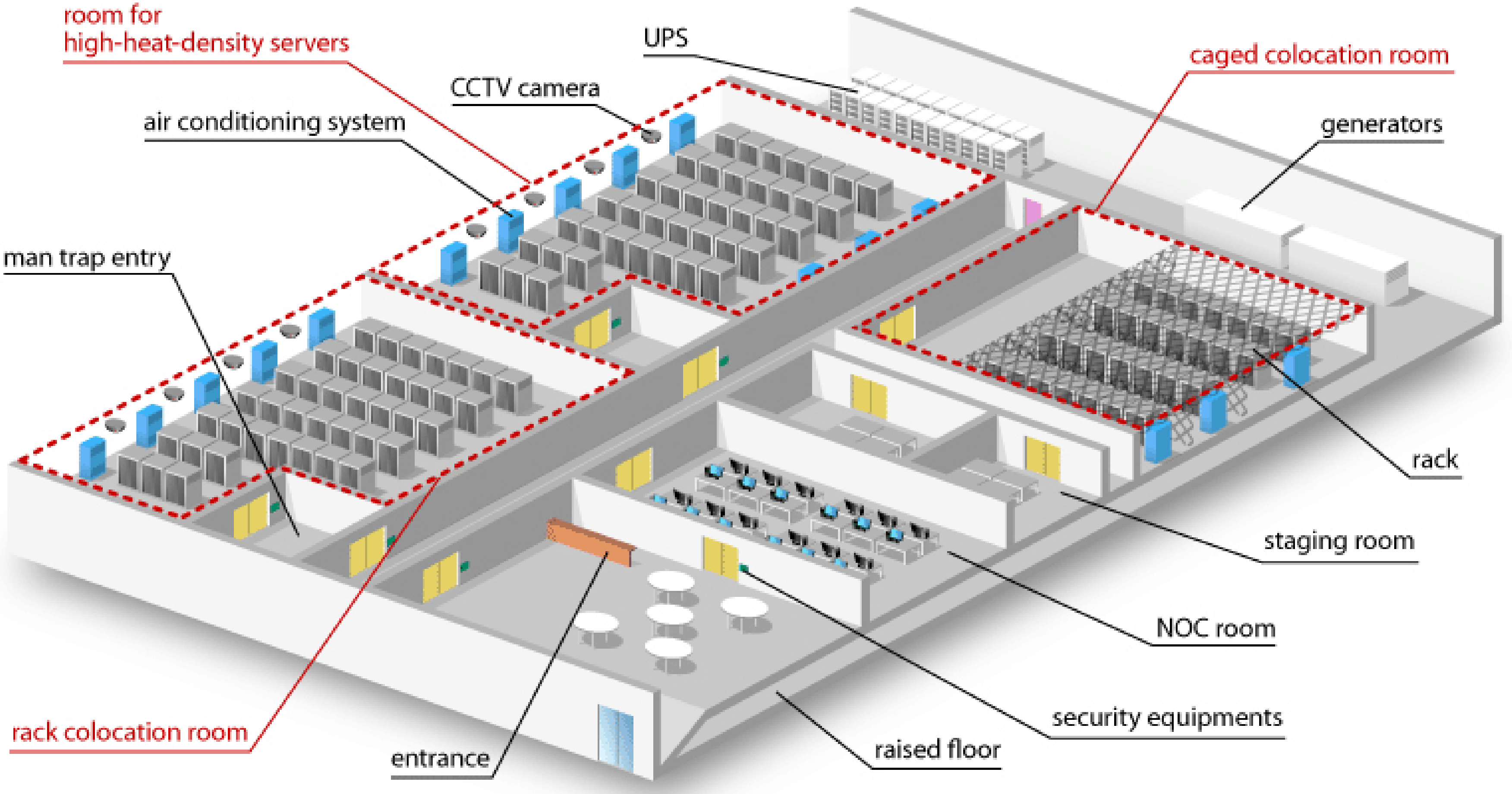


HTTP-based Applications III.

- Logical segregation
  - With VLANs
  - Reduced performance
  - But fewer devices
- + Database servers – typically physically separated

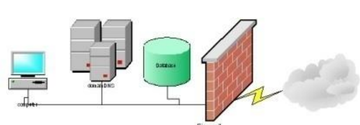


Data Center Overlook



Colocation Room (Colo)

- Place where the customers can place their equipment – “carrier hotel”
  - Closed rooms
    - Cages
    - Racks (lockable)
  - Device owned by the customer
  - Data Center provides the operation / maintenance
    - Space, power, cooling, physical security
    - Connections to telecommunications and network service providers



Caged Colocation Room

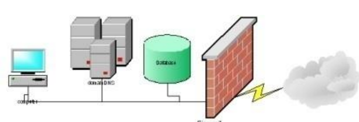
Rack Colocation Room

Staging Room

- Equipment
  - Packing / unpacking
  - Verification tests
  - Configuration
- Outside the DC

Data Center Services

- Colocation / Server rental
- 24/7 supervision
- High reliable environment
  - georedundancy
- Operation
- Design and implementation of movement
- Data back-up and restoration
- Achivation



Data Center Services

- Redundant power supply from two different systems
- Redundant transformers
- Redundant diesel generator
  – Automatic switching
- DC, n+1 redundant
- HVAC, n+1 redundant
- Raised floor, cable tables
- Redundant fibre optic ring