POLITECNICO DI MILANO

Computing Infrastructures

Networking

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures
- Addressing and routing in DCN

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures
- Addressing and routing in DCN

Stages of enterprise infrastructures

1 - Monolithic app

- Minimal network demands
- Proprietary protocols

Stages of enterprise infrastructures
2 - Client server

- High network demands inside the enterprise
- Applications walled within the enterprise
- TCP/IP + Proprietary protocols

Stages of enterprise infrastructures
3 - Web applications

- Ubiquitous TCP/IP
- Access from anywhere
- Servers are broken into multiple units

Stages of enterprise infrastructures
4 - Microservices

- Infrastructure moved to cloud providers
- Servers broken into microservices
- Increase of server-to-server traffic

Data center infrastructure

Intra-Datacenter networking
(aka DataCenter [interconnection] Network - DCN)

The need for effective networking in WSCs

- The performance of servers increases over time, the demand for inter-server bandwidth naturally increases as well!!!
  -

The need for effective networking in WSCs

- The performance of servers increases over time, the demand for inter-server bandwidth naturally increases as well!!!

  We can double the aggregate compute capacity or the aggregate storage simply by doubling the number of compute or storage elements

The need for effective networking in WSCs

- Networking has no straightforward horizontal scaling solution.
- Doubling leaf bandwidth is easy:
  - with twice as many servers, we'll have twice as many network ports and thus twice as much bandwidth.

x 1

x 2

- But if we assume that every server needs to talk to every other server, we need to deal with bisection bandwidth

Bisection bandwidth

- The bandwidth across the narrowest line that equally divides the cluster into two parts
- Characterizes network capacity since randomly communicating processors must send data across the "middle" of the network

If we assume that every server needs to talk to every other server, we need to double not just leaf bandwidth but *bisection bandwidth*

The design of the data center network

- Design principles
  - very scalable in order to support a very large number of servers
  - minimum cost in terms of basic building blocks (e.g., switches)
  - modular to reuse simple basic modules
  - reliable and resilient
  - may exploit novel/proprietary technologies and protocols not compatible with legacy Internet

Classes of DCN

DCNs can be classified into three main categories:

- **Switch-centric** architectures
  - Uses switches to perform packet forwarding
- **Server-centric** architecture
  - Uses servers with multiple Network Interface Cards (NICs) to act as switches in addition to performing other computational functions
- **Hybrid** architectures
  - Combine switches and servers for packet forwarding

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures
- Addressing and routing in DCN

Server-centric architectures

East-West traffic

- East-West traffic
  - storage replication (few flows, many data)
    - in Hadhoop distributed filesystem, at least 3 copies of the same data, usually two in the same rack and one in another rack
  - VM migration
  - Network Function Virtualization (NFV)
    - data is processed through a sequence of VMs (e.g., firewall, web server, parental control, accounting server)
- East-West traffic usually larger than North-South traffic
  - Some citations:
    - A 1 byte transaction in North-South traffic generates on average a 100 bytes transaction in East-West traffic
    - According to Cisco's Global Cloud Index (http://blogs.cisco.com/security/trends-in-data-center-security-part-1-traffic-trends, May 2014):
      - In a data center: East-West traffic (76%), North-South traffic (17%), inter-data center traffic (7%).
      - In campus networks: North-South traffic (>90%).

Three-Tier (or layer) "Classical" Network

- Three layer architecture configures the network in three different layers:

Three-Tier (or layer) "Classical" Network

- A simple DCN topology
- Servers are connected to the DCN through access switches.
- Each access-level switch is connected to at least two aggregation-level switches.
- Aggregation-level switches are connected to core-level switches (gateways).

Data-center network architectures

- Three layer architecture reflects the topology of the data center:

Server packing in a rack

- Standard 19 inch rack 42 EIA Units (pizza box)
  - 40 server blades
    - possible single /26 subnet
  - 1 ToR (Top of Rack) switch
- without oversubscription: NB = nb
  - example
    - 40 ports @ 1 Gbit/s to the servers
    - 4 ports @ 10 Gbit/s to the other switches
- with oversubscription: NB < nb
  - example with oversubscription 1:4
    - 40 ports @ 1 Gbit/s to the servers
    - 1 ports @ 10 Gbit/s to the other switches

ToR vs EoR architectures

- ToR (Top-of-Rack) architecture
  - All servers in a rack are connected to a ToR access switch within the same rack
  - Aggregation switches are in dedicated racks or in shared racks with other ToR switches and servers
  - The number of cables is limited → simpler cabling. The number of ports per switch is also limited (lower costs)
  - Higher complexity for switch management

logical topology

physical topology

ToR vs EoR architectures

- EoR (End-of-Row) architecture
  - Aggregation Switches are positioned one per corridor, at the end of a line of rack.
  - servers in a racks are connected directly to the aggregation switch in another rack
  - Aggregation switches must have a larger number of ports
  - more complex cabling, longer cables are required (higher costs)
  - Patch panel to connect the servers to the aggregation switch
  - Simpler switch management

logical topology

physical topology

Three-Tier (or layer) "Classical" Network

- Bandwidth can be increased by increasing the switches at the core and aggregation layers, and by using routing protocols such as Equal Cost Multiple Path (ECMP) that equally shares the traffic among different routes

Three-Tier (or layer) "Classical" Network

- This solution is very simple, but can be very expensive in large data-centers since:
  - Upper layers require faster network equipments.
  - For example:
    - 1 GB Ethernet at the access layer
    - 10 GB Ethernet at the aggregation layer
    - 25 GB Optical connections at the core layer
  - The cost in term of acquisition and energy consumption can be very high

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures
- Addressing and routing in DCN

DCN Clos topology (a.k.a. Leaf-Spine)

- Two stage interconnections
  - Leaf: ToR switch
  - Spine: dedicated switches (aggregation switches)
- In practice: servers have two interfaces connected to two ToR switches to provide fault-tolerance

Clos networks

- Spine-leaf topologies are borrowed from the telephone world
- Non-folded Clos structure

Clos networks in the telephone world

- Let m be the number of middle stage switches
- Let n be the number of input and outputs
  - If m ≥ n there is always a way to rearrange communications to free a path between any pair of idle input/output
  - If m ≥ 2n - 1 there is always a free path between any pair of idle input/output

- But a DCN is a PACKET-SWITCHED network!!

From Clos to "Leaf and Spine" topology

- Clos topology (n = m = k)
  - each switching module is unidirectional
    - k input + k output ports per module
  - each path traverses 3 modules
- Leaf and spine topology
  - each switching module is bidirectional
    - Leaf: 2k bidirectional ports per module
    - Spine: k bidirectional ports per module
  - each path traverses either 1 or 3 modules

Examples of DCN design

- 3072 servers
  - 3072 ports at 10 Gbit/s ⇒ 30.72 Tbit/s
    alternative designs
    - 96 switches with 64 ports and 32 switches with 96 ports
    - 96 switches with 64 ports and 8 switches with 384 ports

Example taken from "Cisco's Massively Scalable Data Center", 2009

Examples of DCN design

- 6144 servers
  - 6144 ports at 10 Gbit/s ⇒ 61.44 Tbit/s
    alternative designs
    - 192 switches with 64 ports and 32 switches with 192 ports
    - 192 switches with 64 ports and 16 switches with 384 ports

Example taken from “Cisco’s Massively Scalable Data Center”, 2009

Examples of DCN design

- Interesting case: all switches have an equal number of ports p = 2k
  - Leaves: k x k switches (2k ports), k servers per TOR
  - Spine: 2k-port switches; # spine switches = k
  - → # leaves = 2k → $2k^2$ servers

- k = 2 → 8 servers
  - 8 ports at 10 Gbit/s ⇒ 80 Gbit/s
  - 4 switches with 4 ports and 2 switches with 4 ports

(a) Two-tier Clos using four-port switches

- k = 32 → 2048 servers
  - 2048 ports at 10 Gbit/s ⇒ 20.48 Tbit/s
  - 64 switches with 64 ports and 32 switches with 64 ports

# of Clos design in DNC

- Advantages
  - Use of homogeneous equipment
  - Simple Routing
  - The number of hops is the same for any pair of nodes
  - Small blast radius

- What About Scaling further?
- Can we scale to multi-tier designs?

(a) Two-tier Clos using four-port switches

Start with a two-tier network and add an additional row of switches

Scaling to a three-tier network

Pod-based model, aka the Fat Tree

- An option: transform each spine-leaf group into a «pod» and add a super-spine tier
  - POD: Point Of Delivery
- A highly scalable and cost-efficient DCN architecture that aims to maximize bisection bandwidth.
- It can be built using commodity Gigabit Ethernet switches with the same number of ports.
- Used by Microsoft, Amazon

(a) Two-tier Clos using four-port switches

(c) Pod-based three-tier Clos using four-port switches

Point Of Delivery (POD)

- A module or group of network, compute, storage, and application components that work together to deliver a network service
- The PoD is a repeatable pattern, and its components increase the modularity, scalability, and manageability of data
  - (taken from Wikipedia)

Recursive Leaf and Spine

- Leaf with $2k^2$ bidirectional ports
  - k^2 ports to the servers and k^2 ports to the data center network
    - note that this cannot be used to interconnect directly $2k^2$ servers since the network would be blocking
  - built with 2k switches EACH with 2k ports

Intra-pod and inter-pod communications

- k^2P servers
  - 2kP switches with 2k ports
  - k^2 switches with P ports
- Fat-tree: choose P = 2k
  - $2k^3$ servers
  - $(k^2 + 2k \cdot 2k) = 5k^2$ switches with 2k ports

The Fat-Tree Network

- At the edge layer, there are 2k pods (groups of servers), each with k^2 servers.
- Each edge switch is directly connected to k servers in a pod and k aggregation switches.
- A fat-tree network with 2k-port commodity switches can accommodate 2k^3 servers in total
- k^2 core switches with 2k-port each, each one connected to 2k pods
- Each aggregation switch is connected to k core switches and to k edge switches.
  - Note the partial connectivity at switch level

k = 2
4 pods
16 servers
20 switches

Fat tree example

- Data center with 65 536 servers in 64 pods
  - 65 536 ports at 10 Gbit/s ⇒ 655 Tbit/s
  - P = 64 pods, k = 32
  - in total 5 120 switches with 64 ports

Virtual chassis model (e.g. Facebook)

(a) Two-tier Clos using four-port switches

(b) Virtual chassis-based three-tier Clos using four-port switches

Google scenario → Goolge Datacenter

- Worldwide coverage with tens of sites
- Data center traffic bandwidth demand doubles every 12-15 months (faster than Internet)
  - larger datasets (photo/video, logs, Internet-connected sensors, etc.)
  - web services
  - internal applications (index generation, web search, serving ads, etc.)
- **Google Design approach**
  - multistage Clos topologies on commodity switch silicon
  - centralized control
    - one configuration pushed to all the switches
    - SDN approach
  - modular hardware design with simple, robust software
  - general-purpose, off-the-shelf switch components

A. Singh, et al., "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", ACM SIGCOMM Computer Communication Reviews, Oct. 2015

Jupiter topology

Building blocks used in the Jupiter topology.

Jupiter Middle blocks housed in racks.

Options to connect to the external network layer.

Figures reproduced from [*Google*].

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- <u>Server-centric and hybrid architectures</u>
- Addressing and routing in DCN

Server-centric architectures: CamCube

- A server-centric architecture proposed for building container-sized data centers.
- It may reduce implementation and maintenance costs by using only servers to build the DCN.
- It uses a 3D-Torus topology to interconnect the servers directly.
- As a torus-based architecture, it exploits network locality to increase communication efficiency.
- **Drawbacks**: CamCube requires servers with multiple NICs to assemble a 3D Tours network, long paths, and high routing complexity

3D Torus with 64 servers

Hybrid architectures

DCell

BCube

MDCube

Hybrid architectures: DCell

- A scalable and cost-efficient hybrid architecture that uses switches and servers for packet forwarding.
- **Recursive Architecture.** It uses a basic building block called $DCell_0$ to construct larger DCells.
- $DCell_k$ (k>0) denotes a level-k DCell constructed by combining n+1 servers in $DCell_0$.
- A $DCell_0$ has n (n<8) servers directly connected by a commodity switch
- Drawbacks: Long communication paths, many required NICs, and increased cabling costs

Hybrid architectures: BCube

- A hybrid and cost-effective architecture that can scale up through recursion.
- It provides high bisection bandwidth and graceful degradation of throughput.
- It uses BCube as a building block, which consists of *n* servers connected to an *n*-port switch.
- A $BCube_k$ (*k*>0) is constructed with *n* $BCube_{(k-1)}$ *s* and n^k *n*-port switches.
- In a $BCube_k$ there are $n^{(k+1)}$ *k*+1-port servers and *k*+1 layers of switches.
- Drawbacks: limited scalability and high cabling costs, many required NICs.

(a) Request phase.

(b) Grant phase.

Hybrid architectures
MDCube

- It is designed to reduce the number of cables in the interconnection of containers
- Each container has an ID mapped to a multidimensional tuple.
- Each container is connected to a neighbor container with a different tuple in one dimension.
- There are two types of links: Intra-container and high-speed inter-container links

Outline

- Fundamental concepts
- Switch-centric architectures
  - Classical 3-tier architecture
  - Leaf-Spine architectures
- Server-centric and hybrid architectures
- <u>Addressing and routing in DCN</u>

Addressing and routing in a data center

- Scenario
  - 100,000 servers,
  - 32 VM each
  - => $3 \cdot 10^6$ MAC and IPs (#servers * #VMs_per_Server)
- Addressing and routing schemes challenging
  - standard schemes do not provide efficient solutions
  - e.g. consider two possible alternative options:
    - Layer2 – Single LAN limitations
      - Routing as a tree topology
      - knowledge of the network discovered with broadcast messages
    - Layer3 – Many VLAN and one subnet per VLAN
      - Large number of DHCP servers, switches and routers (including configurations)
  - Ad hoc solutions are needed
    - E.g. VXLAN - Virtual Extensible Lan
      - Multiple Layer 2 network into a single one
        - Extends the VLAN addresses

- ***A dedicated course is needed for understanding such solutions***

The interplay of storage and networking technology

Balanced design

- The success of WSC distributed storage systems can be partially attributed to the evolution of data center networking fabrics:
  - **disk locality is no longer relevant in intra-data center computations.**
- This observation enables dramatic simplifications in the design of
  - ✓ distributed disk-based storage systems
  - ✓ utilization improvements,

  since any disk byte in a WSC facility can, in principle, be utilized by any task regardless of their relative locality

Computer architects are trained to solve the problem of finding the right combination of performance and capacity from the various building blocks that make up a WSC

An example: programmer's view of storage hierarchy of a hypothetical WSC _ (Part1)

- We assume a system with **5,000 servers, each with 256 GB of DRAM, one 4 TB SSD, and eight 10 TB disk drives.**
- Each group of **40 servers** is connected through a **40-Gbps link to a rack-level switch (TOR)**
- Each rack has **an additional 10-Gbps uplink** bandwidth per machine for connecting the rack to the cluster-level switch (**AGGREGATION**)
- Network latency numbers assume a TCP/IP transport, and networking bandwidth values assume that each server is using its fair share of the available cluster-level bandwidth.
- For disks, typical commodity disk drive (SATA) latencies and transfer rates are considered.

Storage hierarchy of a WSC _ (Part2)

Quantifying Latency, Bandwidth and Capacity _ (Part3)

The bandwidth from local SSDs is 3 GB/s, whereas the one from off-rack SSDs is 1.25 GB/s

Total disk storage in the cluster is more than one million times larger than local DRAM.

- **A large application** that requires servers in many racks to operate **must deal effectively with large discrepancies in latency, bandwidth, and capacity**.

- These discrepancies are much larger than those seen on a single machine, making it more difficult to program a WSC:
  - **A key challenge for architects of WSCs** is to smooth out these discrepancies in a cost-efficient manner.
  - **A key challenge for software architects** is to build SW infrastructure and services that hide this complexity

Backup slides

- Additional slides with hints on Routing from PROF. MAIER

Layer-2 design: One single-LAN scheme

Drawbacks

- very large forwarding tables in L2 switches
- shared broadcast domain and lots of broadcast traffic (e.g., ARP) routing loops
  - (unfortunately) no TTL in Ethernet packets
  - standard Spanning Tree Protocol (STP) not suitable
    - routing across multiple paths cannot be exploited
    - slow to converge
- large failures
  - due to possible improper cabling, misconfiguration, or flawed software on a single device
  - affect the entire spanning-tree domain

Layer-2 enhanced design

- Rapid Spanning Tree Protocols (RSTP)
  - to converge faster than STP
- Multiple Spanning Tree Protocol (MST)
  - to exploit multiple paths
- Multi-Chassis Link-Aggregation (MC-LAG)
  - to virtualize multiple links and switches into a single link and switch lack of ability to scale to > 2 links or switches
  - lack of standard implementations
  - failure risk of syncing the states between the switches

Layer-3 design: One subnet per VLAN

Drawbacks

- many DHCP servers and VLANs
- very large number of switches and routers (around 10,000)
- Interior Gateway Protocol IGP (e.g., OSPF) in each router
  - manual administrator configuration and oversight
  - flood prefix advertisements, controlled only by a basic partition into areas
- routing loops
  - (fortunately) TTL in IP packet
- VM migration
  - when changing LAN, a new IP address is required and existing TCP connections break

Addressing/Routing DC schemes

- VXLAN
  - scalable tunneling scheme similar to VLAN
- LISP
  - provides IP address mobility across layer-3 subnets
- IPv6
- BGP
- but many other solutions have been proposed:
  - FabricPath, TRILL, NVGRE, OTV, Shortest Path Bridging (SPB), Segment Routing (SRv6), etc.