Datacenter Networks

Mike Freedman
COS 461: Computer Networks

http://www.cs.princeton.edu/courses/archive/spr14/cos461/

Cloud Computing

- **Elastic resources**
  - Expand and contract resources
  - Pay-per-use
  - Infrastructure on demand

- **Multi-tenancy**
  - Multiple independent users
  - Security and resource isolation
  - Amortize the cost of the (shared) infrastructure

- **Flexible service management**
Cloud Service Models

• Software as a Service
  — Provider licenses applications to users as a service
  — E.g., customer relationship management, e-mail, ...
  — Avoid costs of installation, maintenance, patches, ...

• Platform as a Service
  — Provider offers platform for building applications
  — E.g., Google’s App-Engine, Amazon S3 storage
  — Avoid worrying about scalability of platform

Cloud Service Models

• Infrastructure as a Service
  — Provider offers raw computing, storage, and network
  — E.g., Amazon’s Elastic Computing Cloud (EC2)
  — Avoid buying servers and estimating resource needs

Enabling Technology: Virtualization

• Multiple virtual machines on one physical machine
• Applications run unmodified as on real machine
• VM can migrate from one computer to another

Multi-Tier Applications

• Applications consist of tasks
  — Many separate components
  — Running on different machines

• Commodity computers
  — Many general-purpose computers
  — Not one big mainframe
  — Easier scaling
Componentization leads to different types of network traffic

- **“North-South traffic”**
  - Traffic to/from external clients (outside of datacenter)
  - Handled by front-end (web) servers, mid-tier application servers, and back-end databases
  - Traffic patterns fairly stable, though diurnal variations

- **“East-West traffic”**
  - Traffic within data-parallel computations within datacenter (e.g. “Partition/Aggregate” programs like Map Reduce)
  - Data in distributed storage, partitions transferred to compute nodes, results joined at aggregation points, stored back into FS
  - Traffic may shift on small timescales (e.g., minutes)
Virtual Switch in Server

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

Modularity, Modularity, Modularity
- Containers
- Many containers
Datacenter Network Topology

Key:
- CR = Core Router
- AR = Access Router
- S = Ethernet Switch
- A = Rack of app. servers

~ 1,000 servers/pod

Capacity Mismatch?

“Oversubscription”: Demand/Supply
A. \(1 > 2 > 3\)
B. \(1 < 2 < 3\)
C. \(1 = 2 = 3\)

Capacity Mismatch!

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
Datacenter Routing

New datacenter networking problems have emerged...

Network Incast

- Incast arises from synchronized parallel requests
  - Web server sends out parallel request ("which friends of Johnny are online?")
  - Nodes reply at same time, cause traffic burst
  - Replies potential exceed switch's buffer, causing drops

Solutions mitigating network incast
A. Reduce TCP's min RTO (often use 200ms >> DC RTT)
B. Increase buffer size
C. Add small randomized delay at node before reply
D. Use ECN with instantaneous queue size
E. All of above
Full Bisection Bandwidth

- Eliminate oversubscription?
  - Enter FatTrees
  - Provide static capacity

- But link capacity doesn’t “scale-up”. Scale out?
  - Build multi-stage FatTree out of k-port switches
  - k/2 ports up, k/2 down
  - Supports k^3/4 hosts:
    - 48 ports, 27,648 hosts

Full Bisection Bandwidth Not Sufficient

- Must choose good paths for full bisectional throughput
- Load-agnostic routing
  - Use ECMP across multiple potential paths
  - Can collide, but ephemeral? Not if long-lived, large elephants
- Load-aware routing
  - Centralized flow scheduling, end-host congestion feedback, switch local algorithms

Conclusion

- Cloud computing
  - Major trend in IT industry
  - Today’s equivalent of factories

- Datacenter networking
  - Regular topologies interconnecting VMs
  - Mix of Ethernet and IP networking

- Modular, multi-tier applications
  - New ways of building applications
  - New performance challenges

Load Balancing
Load Balancers

- Spread load over server replicas
  - Present a single public address (VIP) for a service
  - Direct each request to a server replica

Wide-Area Network

- Ingress Proxies

Virtual IP (VIP)
192.121.10.1
10.10.10.1
10.10.10.2
10.10.10.3