Networking Case Studies

Enterprise → Backbone → Datacenter

Wireless
Cloud Computing
Cloud Computing

• **Demand-elastic resources**
  - Expand & contract resources as demand dictates
    • Pay-per-use; Infrastructure on demand

• **Multi-tenancy**
  - Multiple independent users
  - Security and resource isolation
  - Amortize the (shared) infrastructure cost
  - 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 & estimating resource needs
Enabling Technology: Virtualization

- Multiple virtual machines on one physical machine
- Applications run unmodified as on real machine
- Recently: Lighter-weight virtualization through “containers”
- Can migrate from one machine to another
- Autoscale by spinning up/down VMs & containers
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, written back to storage
  – Traffic may shift on small timescales (e.g., minutes)
North-South Traffic

Router

Front-End Proxy
- Web Server
  - Data Cache
- Web Server
  - Data Cache
- Web Server
  - Database

Front-End Proxy
- Web Server
  - Database
East-West Traffic

Distributed Storage

Map Tasks

Reduce Tasks

Distributed Storage
Datacenter Network
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
Datacenter Network Topology

Internet

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

~ 1,000 servers/pod

Key

- CR = Core Router
- AR = Access Router
- S = Ethernet Switch
- A = Rack of app. servers
Capacity Mismatch?

“Oversubscription”: Much more demand vs. supply for higher links
Capacity Mismatch!

Particularly bad for east-west traffic
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

Key
• CR = Core Router (L3)
• AR = Access Router (L3)
• S = Ethernet Switch (L2)
• A = Rack of app. servers

-~ 1,000 servers/pod == IP subnet
New datacenter networking problems have emerged...
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 potentially exceed switch's buffer, causing drops
Network Incast

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
Network Bandwidth Measurements

- Bisection bandwidth: Split nodes into two halves such that bandwidth between the halves is minimal, that is the bisection b/w

- Full bisection bandwidth: $\frac{1}{2}$ of the nodes can communicate simultaneously with the other $\frac{1}{2}$
Full Bisection Bandwidth

• Eliminate oversubscription?
  – Enter FatTrees
  – Provide static capacity
  – Heterogeneous Links
    • 1-10 GB in racks
    • 40-100GB to core
Full Bisection Bandwidth

- But “scale up” link capacity has limits
- New scale out architectures
  - 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