

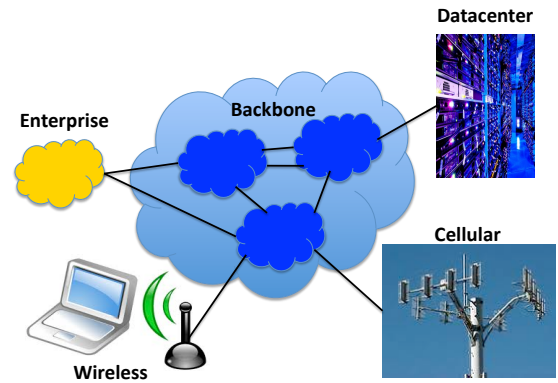


Datacenter Networks

Mike Freedman
COS 461: Computer Networks

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

Networking Case Studies



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

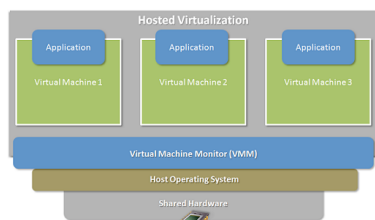
5

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

6

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**

7

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

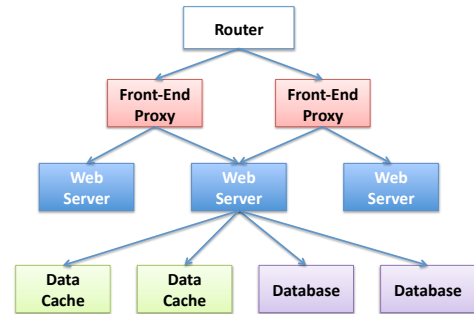
8

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)

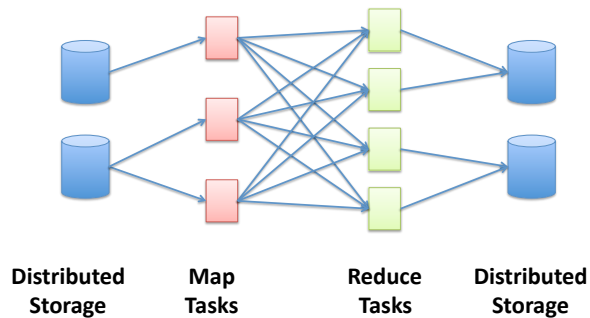
9

North-South Traffic



10

East-West Traffic

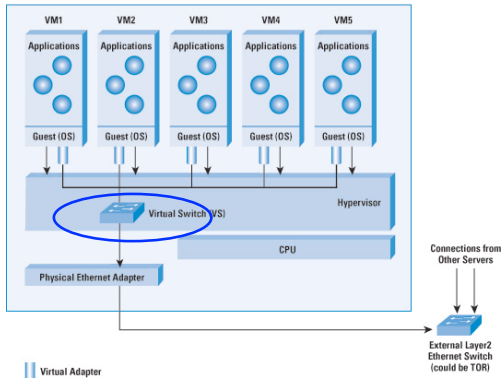


11

Datacenter Network

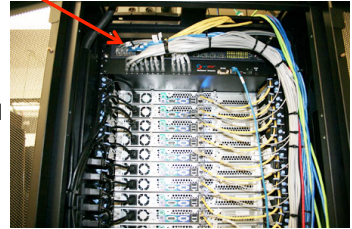
12

Virtual Switch in Server

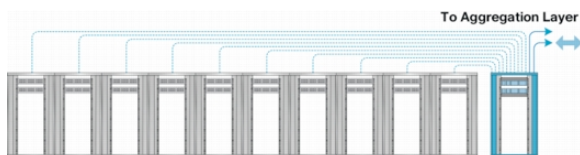


Top-of-Rack Architecture

- 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

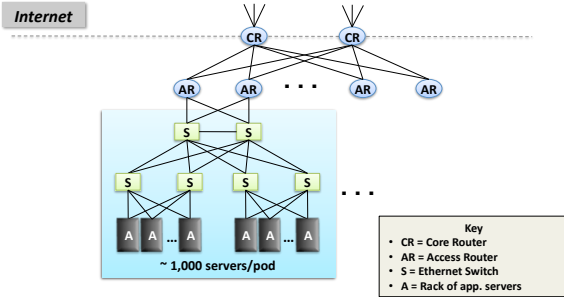
- Containers



- Many containers

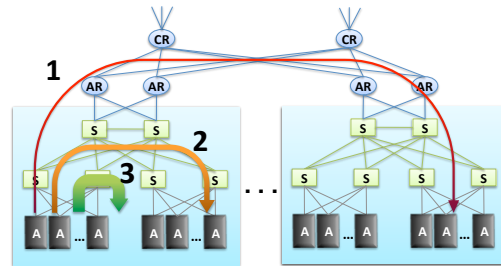


Datacenter Network Topology



17

Capacity Mismatch?



“Oversubscription”: Demand/Supply

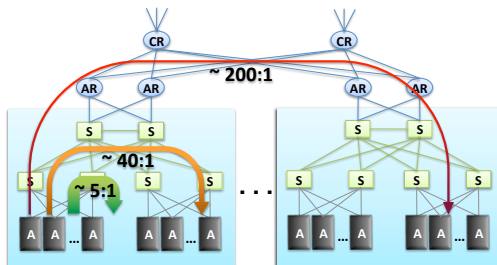
A. $1 > 2 > 3$

B. $1 < 2 < 3$

C. $1 = 2 = 3$

18

Capacity Mismatch!



Particularly bad for east-west traffic



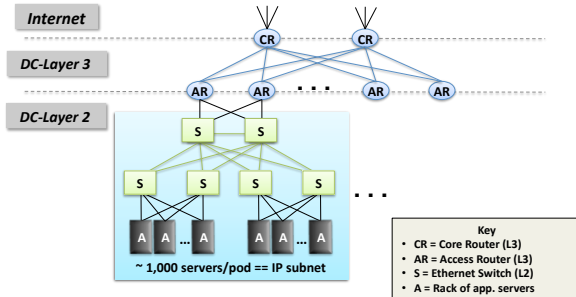
19

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

20

Datcenter Routing

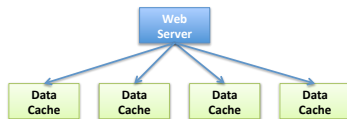


21

New datacenter networking problems have emerged...

22

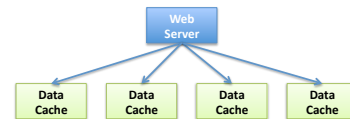
Network Incast



- 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

23

Network Incast

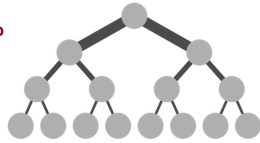


- Solutions mitigating network incast
 - Reduce TCP’s min RTO (often use 200ms >> DC RTT)
 - Increase buffer size
 - Add small randomized delay at node before reply
 - Use ECN with instantaneous queue size
 - All of above

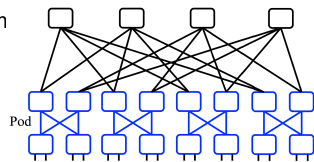
24

Full Bisection Bandwidth

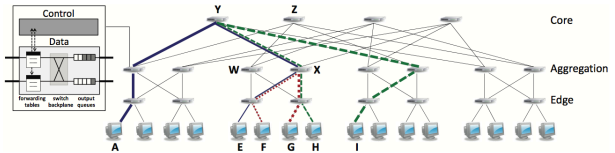
- Eliminate oversubscription?
 - Enter FatTrees
 - Provide static capacity



- 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

26

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

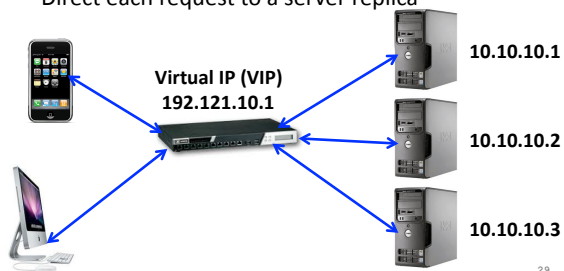
27

Load Balancing

28

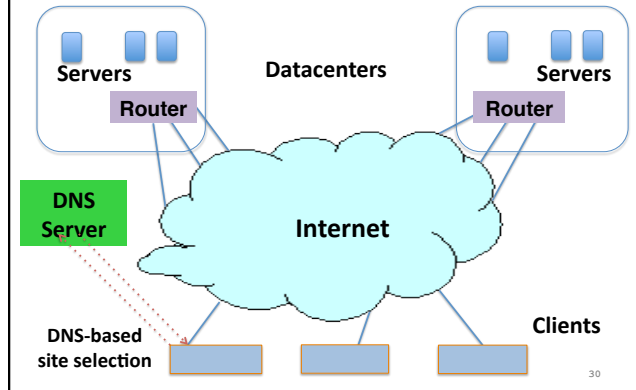
Load Balancers

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



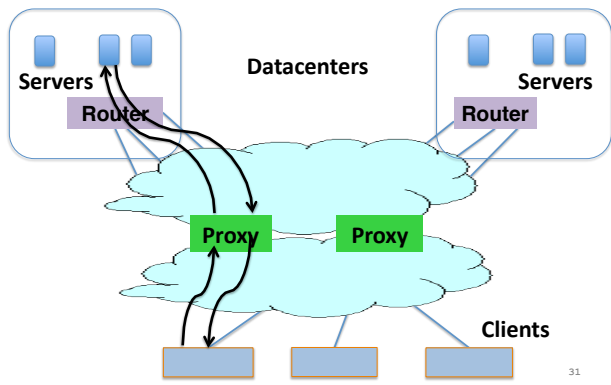
29

Wide-Area Network



30

Wide-Area Network: Ingress Proxies



31