Data Center Networks

Xia Zhou
May 10, 2018

Network/Cluster Computing Has Been Around for a While

- Grid computing, cluster computing (Beowulf cluster)


1961, Information Processing Center at the National Bank of Arizona

- All of a sudden, data centers are extremely hot
  - **Why?**


The Scale!

- Everyone wants to operate at Internet scale
  - Millions of users
  - Zetabytes of data to analyze
    - Web server logs, Ads reviews/clicks, social networks, blogs, Twitter, video…
- But not everyone has the expertise to build a cluster
  - **Let someone else do it for you!**

Cloud Computing

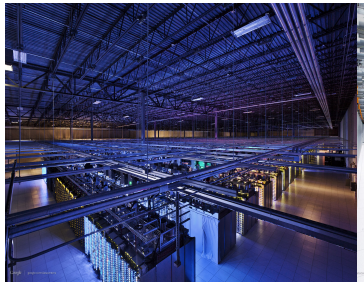
- Everything is a service
  - Infrastructure
  - Platform
  - Software
  - Storage

- What actually powers the cloud?

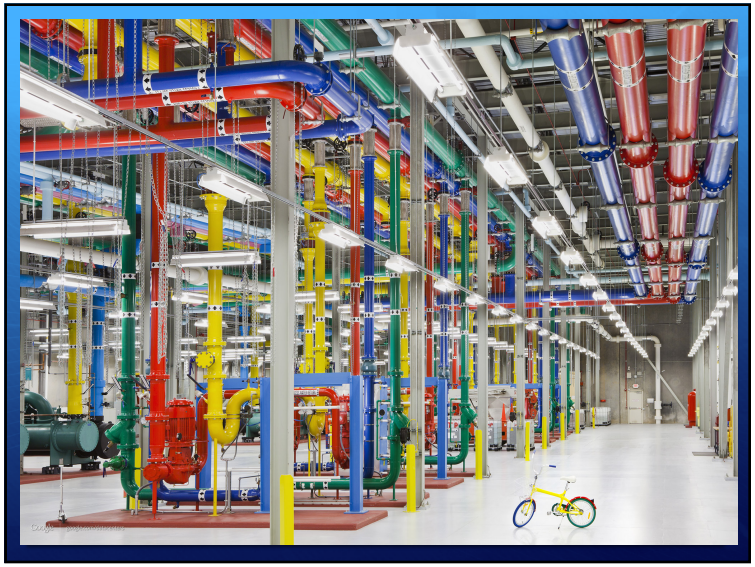
5

Today's Data Centers

- Warehouse of servers



6

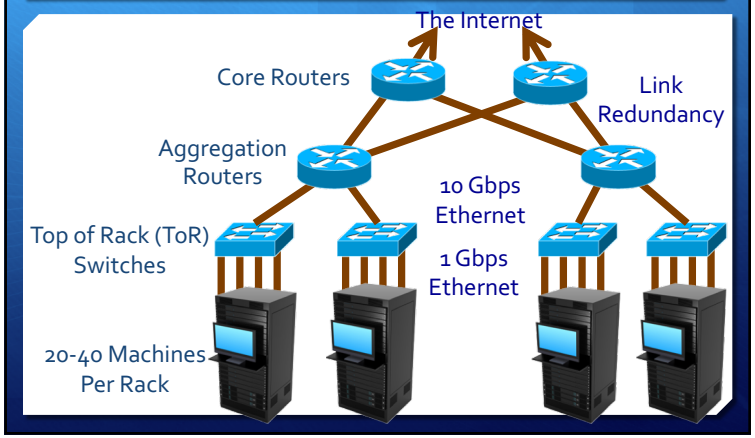


What Goes into a Data Center?

- Servers organized in racks
- Each rack has a "Top of Rack" switch
- An aggregation fabric interconnects ToR switches

8

Typical Data Center Topology



What's Different about Data Center Networks?

- Huge scale
  - 1M servers (Microsoft), > $1B to build one site (Facebook)
- Speed, speed, speed…
  - High bandwidth 10/40/100Gbps
  - RTT: **10s of microseconds**
- Single administrative domain
  - Can invent your own design, control the traffic placement
- Regular/planned topologies

10

Lots of Open Problems

- Diverse applications
  - Heterogeneous, unpredictable traffic patterns
  - Competition over resources
  - Isolation
  - Reliability issues
  - Privacy
- Management, diagnosis and debugging at scale
- Heat and power

11

Today's Topic: Network Problems

- Data centers are data-intensive
- Hardware can handle it
  - CPUs scale with Moore's Law, RAM is almost as fast as CPU, RAID and SSDs are pretty fast
- Current network cannot handle it
  - Slow, not keeping pace over time
  - Wiring is a nightmare
  - Expensive
  - Hard to manage
  - Non-optimal protocols

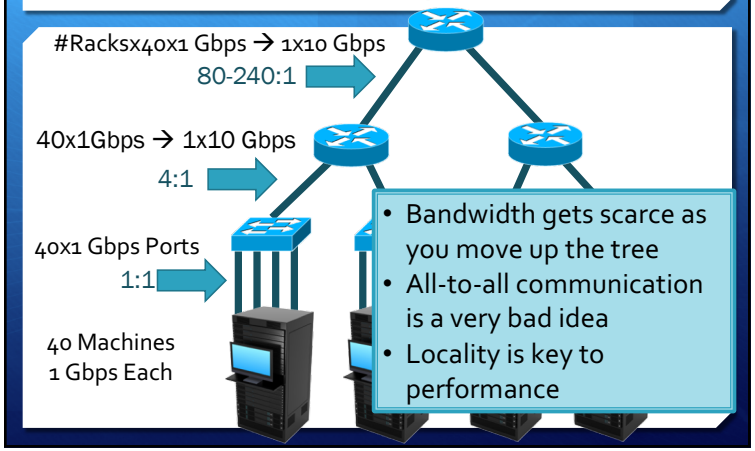
12

Outline

- Introduction
- Network topology
  - Fat tree
  - Wireless in data centers
  - Optical in data centers
- Transport protocols

13

Problem: Oversubscription



Oversubscription can be Harmful

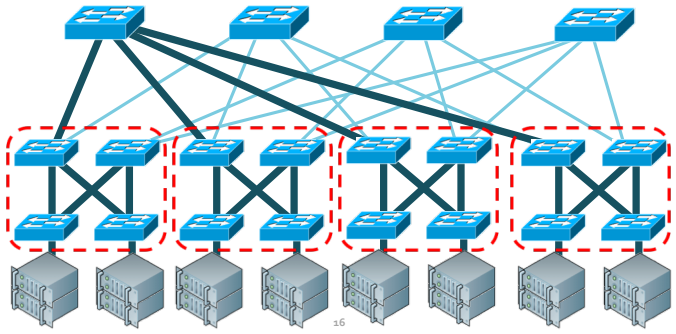
- Ruin your network
  - Limits application scalability
- Problem is about to get worse
  - 10 GigE servers are more affordable
  - 128 port 10 GigE routers are not
- An issue of the core routers
- Get rid of the core routers by using cheap switches?
  - Maintain 1:1 subscription ratio

15

Fat-tree

To build a K-ary fat tree
- K-port switches
- K pods, each with K switches
- $K^3/4$ hosts
- $(K/2)^2$ core switches

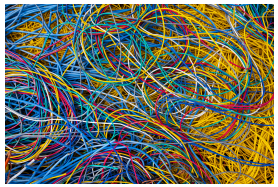
In this example K=4
- 4-port switches
- 4 pods, each with 4 switches
- $K^3/4 = 16$ hosts
- $(K/2)^2 = 4$ core switches



16

Fat-tree

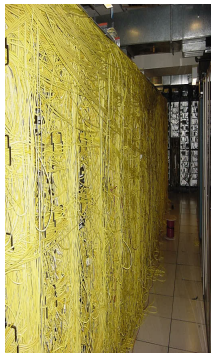
- The good
  - Full bisection bandwidth
  - Low-cost, commodity hardware
  - Redundancy for failover
- The bad
  - Custom routing (NetFPGA)
  - Wiring is a nightmare
    # of wires: → $3K^3/4$
    - 48 port switches = 82944



17

Limitations of Wired Interconnects

- Wiring is complex and costly
  - Planning, deploying, testing 10K+ fibers
  - Takes several weeks or even months
- Difficult to change wiring
  - High labor cost
  - Significant interruptions to operations
- Overprovisioning is difficult
  - Traffic demands unpredictable
  - Limited by hardware costs



18

Dealing with Traffic Hotspots

- Measurements show sporadic congestion losses caused by traffic hotspots
  - Traffic hotspots are unpredictable, can appear anywhere
  - Can double failure rate for some jobs

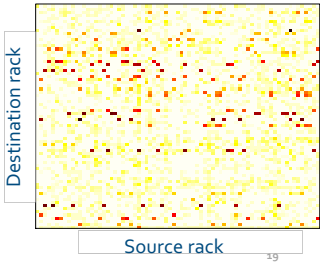


Figure source: Halperin, D., et al. Augmenting data center networks with multi-gigabit wireless links. In *Proc. of SIGCOMM* (2011)

19

Dealing with Traffic Hotspots

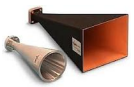
- Measurements show sporadic congestion losses caused by traffic hotspots
  - Traffic hotspots are unpredictable, can appear anywhere
  - Can double failure rate for some jobs
- Hard to add bandwidth using wires
  - Do not know where and when to add capacity
  - Rewiring is complex, high labor cost, interrupts current operation

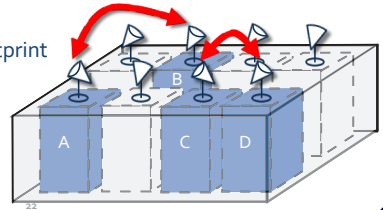
Need alternative solutions!

20

Augmenting via Wireless Links

- Key benefit: on-demand links
  - Create links on-the-fly at congestion hotspots
  - Adapt to traffic dynamics
- New wireless technology: 60 GHz beamforming
  - Multi-Gbps data rate
  - Small interference footprint

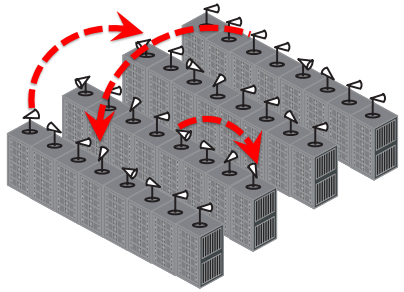




22

Flexible Wireless Links in Data Centers

- Connect any rack pair wirelessly to address dynamic traffic hotspots

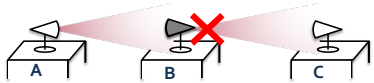


Hard to do so using 60GHz transmissions

23

Key Challenges

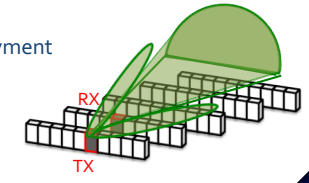
- **Link blockage**: small obstacles block the link



→ Must use multi-hop forwarding

- **Radio interference**: beam interferes with racks in its direction
  - Exacerbated by dense rack deployment
  - Signal leakage makes it worse

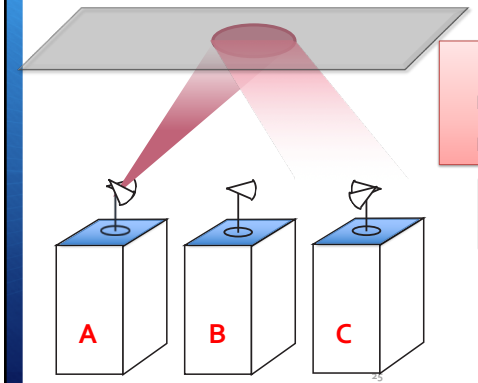
→ Links interfere with each other



24

Wireless 3D Beamforming

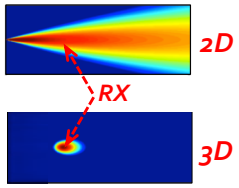
Connect racks by reflecting signals off the ceiling!



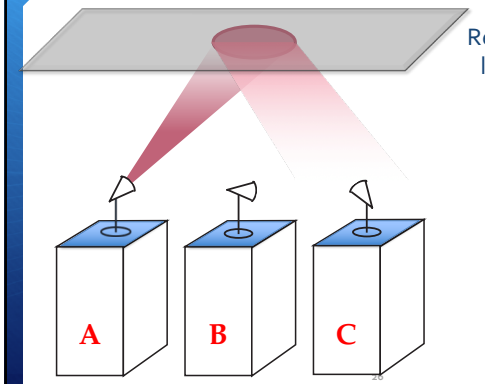
Key Benefits

No more link blockage

Much smaller interference



Easy Setup



Reuse existing hardware, low maintenance cost!

Reflector

Absorber

26

3D Beamforming Summary

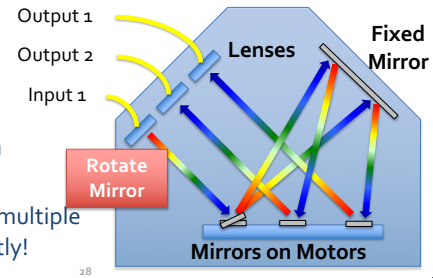
- New wireless primitive added into data center toolbox
  - Overcomes key challenges of using wireless in data centers
  - Opens up more opportunities of using wireless in data centers
- Practical issues
  - Traffic scheduling
  - Reflector placement, impact on data center cooling

http://www.cs.dartmouth.edu/~xia/papers/sigcomm12_beam3d.pdf

27

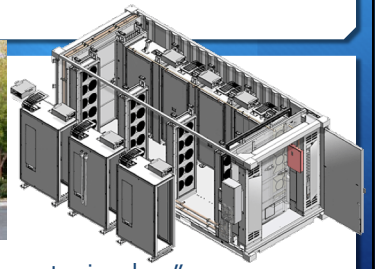
Augmenting via Optical Links

- Optical circuit switch
  - Uses mirrors to bounce light from port to port
  - No decoding
  - Mechanically adjust mirrors in milliseconds
- Wavelength division multiplexing
  - Single port carries multiple streams concurrently!



28

Example: Connecting Modular Datacenters



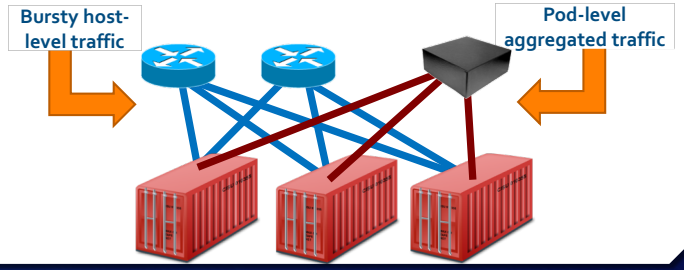
- Shipping container: "datacenter in a box"
  - 250-1K servers per container
- How do you connect the containers?
  - Copper cable constraint: physical distance matters (10GigE -> 10 meters)

29

Helios: Datacenters at Light Speed

Packet switch network
- Electrical packet switches
- Connect all containers

Optical circuit network
- Optical circuit switches
- Direct container-to-container links on demand



Outline

- Introduction
- Network topology
- Transport protocols
  - Google and Facebook
  - DCTCP
  - D3

31

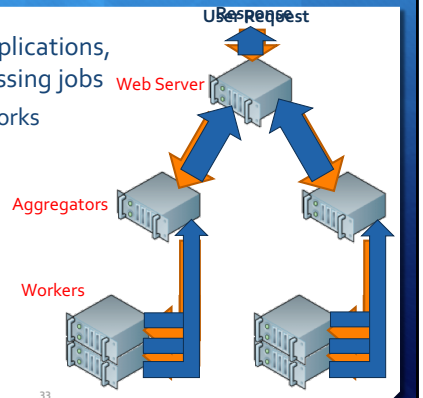
Transport in Data Center Networks

- TCP optimized for wide area networks
  - Slow-start, AIMD convergence
  - Zero knowledge congestion control
    - Self-induces congestion, loss equals congestion
- But, data center is not the Internet
  - Latency is tiny (250µs in absence of queuing)

32

Partition/Aggregate Pattern

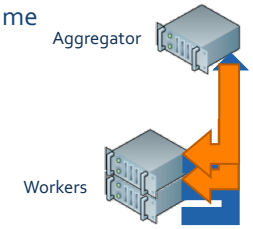
- Common for web applications, and even data processing jobs
  - Search, social networks
  - Dryad, MapReduce
- Responses under deadline
  - 230~300ms



33

Problem: Incast

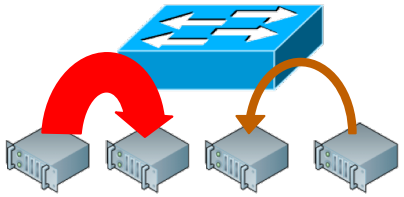
- Aggregator sends out queries to a rack of workers
  - 1 aggregator, 39 workers
- Each query takes the same time to complete
- All workers answer at the same time
  - 39 flows → 1 port
  - Limited switch memory
  - Limited buffer at aggregator
- Result: packet losses ☹



34

Problem: Buffer Pressure

- In theory, each port on a switch should have its own buffer
- Cheap switches share buffer memory across ports
  - Fat flows can congest the thin flow!



35

Problem: Queue Buildup

- Long TCP flows congest the network
  - Ramp up, past slow start
  - Don't stop until they induce queuing + loss
  - Oscillate around max utilization
- Short flows can't compete
  - Never get out of slow start
  - Deadline sensitive!

Industry Hacks

Google

- Use heavy compression to maximize data efficiency

facebook

- Custom engineer to use UDP
- Connection pooling: share buffer pool per thread

37

Dirty Slate Approach: DCTCP*

- Goals
  - Alter TCP to achieve low latency
  - Work with shallow buffered switches
  - Do not modify apps, switches, or routers
- Idea
  - Scale window in proportion to congestion
  - Use existing ECN functionality
  - Turn single-bit congestion info to multi-bit

*Data Center TCP (DCTCP). SIGCOMM'10.
http://www.sigcomm.org/sites/default/files/ccr/papers/2010/October/1851275-1851192.pdf

38

ECN and ECN++

- Original ECN
  - Switches mark EC bit of packet if there is congestion
  - Receivers echo the EC bit in ACK
  - EC in ACK stays set until sender clears with CWR
- Problem: feedback is binary
- DCTCP:
  - Receiver echoes the actual EC bits
  - Sender estimate congestion ($0 \le \alpha \le 1$) each RTT based on the fraction of marked packets
  - cwnd = cwnd * $(1 - \alpha/2)$

39

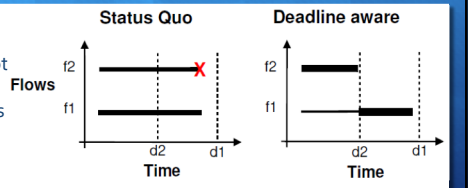
Shortcomings of DCTCP

- No scheduling, cannot solve incast
- Queries may still miss deadlines
  - Flows do not help if they miss the deadline
- Network throughput is not the right metric
  - Application goodput is
- TCP/DCTCP is oblivious to deadline

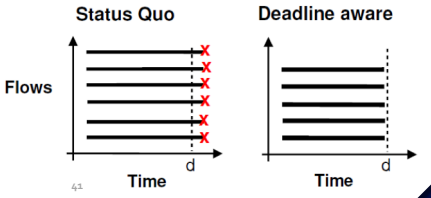
40

The Need of Deadline-Awareness

- Case #1
  - Fair share causes bot to fail
  - Unfair share enables both to succeed



- Case #2
  - Allowing all makes all fail
  - Quenching one leads to better goodput



41

Clean Slate Approach: D3*

- **Key insight**: ask for the bandwidth required to meet the deadline
  - Hosts use flow size and deadline to request bandwidth
  - Routers measure utilization and make soft-reservations
- Challenges ahead
  - Not for incremental deployment, may not play nice with TCP
  - Complexity in the switch
  - Application level changes

*Better Never than Late: Meeting Deadlines in Datacenter Networks. SIGCOMM'11. *http://conferences.sigcomm.org/sigcomm/2011/papers/sigcomm/p50.pdf*

42

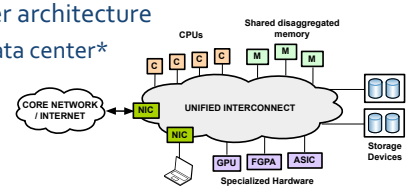
Conclusion

- Data center is a super hot topic right now
  - Topology, routing
  - Network stack
  - Heat, power
  - Management
  - Applications: Hadoop, Dynamo, Cassandra, NoSQL
- Tough for academic research
  - No real data centers to test around

43

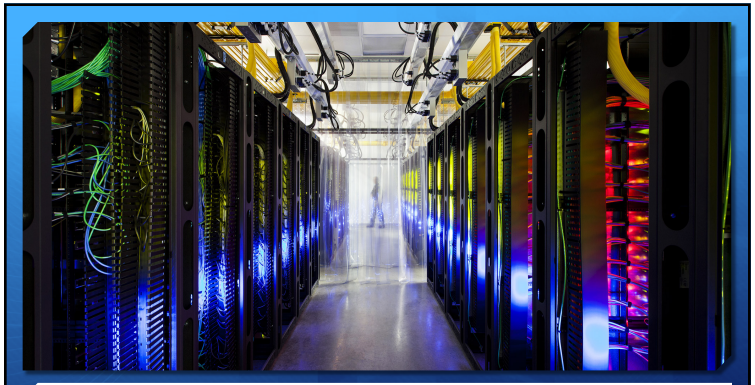
Open Problems

- Data center measurement data
  - Real traces are really hard to get
  - Hard to quantify research results
- Rethink data center architecture
  - Disaggregated data center*



*Network Support for Resource Disaggregation in Next-Generation Datacenters. HotNets'13. *http://conferences.sigcomm.org/hotnets/2013/papers/hotnets-final40.pdf*

44



We are done!

Good luck to the last assignment!

45