Topic 7: The Datacenter

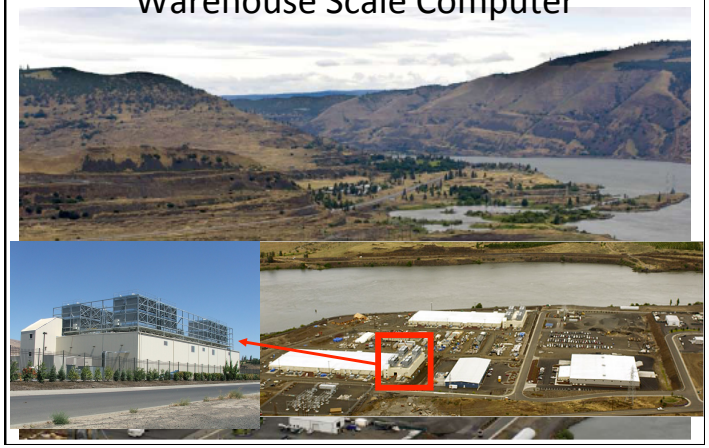
(DC/Data Center/Data Centre/….)

Our goals:

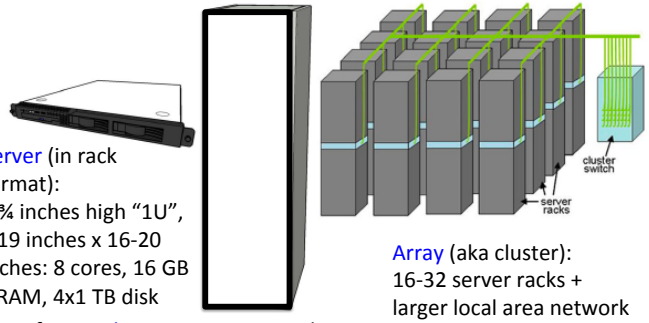
- Datacenters are the new Internet; regular Internet has become mature (ossified); datacenter along with wireless are a leading edge of new problems and new solutions
- Architectures and thoughts
  - Where do we start?
  - old ideas are new again: VL2
  - c-Through, Flyways, and all that jazz
- Transport layer obsessions:
  - TCP for the Datacenter (DCTCP)
  - recycling an idea (Multipath TCP)
  - Stragglers and Incast

2

Google Oregon Warehouse Scale Computer



Equipment Inside a WSC



Server (in rack format):
1 ¾ inches high "1U", x 19 inches x 16-20 inches: 8 cores, 16 GB DRAM, 4x1 TB disk

7 foot Rack: 40-80 servers + Ethernet local area network (1-10 Gbps) switch in middle ("rack switch")

Array (aka cluster): 16-32 server racks + larger local area network switch ("array switch") 10X faster => cost 100X: cost f(N^2)

4

Server, Rack, Array

Data warehouse?

If you have a Petabyte,
you might have a datacenter

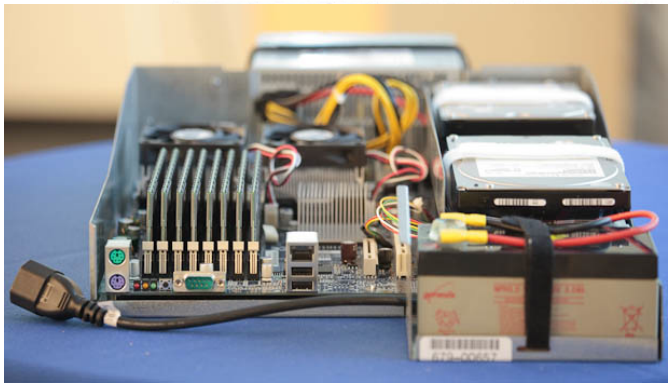
If your paged at 3am because you only have a few Petabyte left,
you might have a data warehouse

Luiz Barroso (Google), 2009

The slide most likely to get out of date…

6

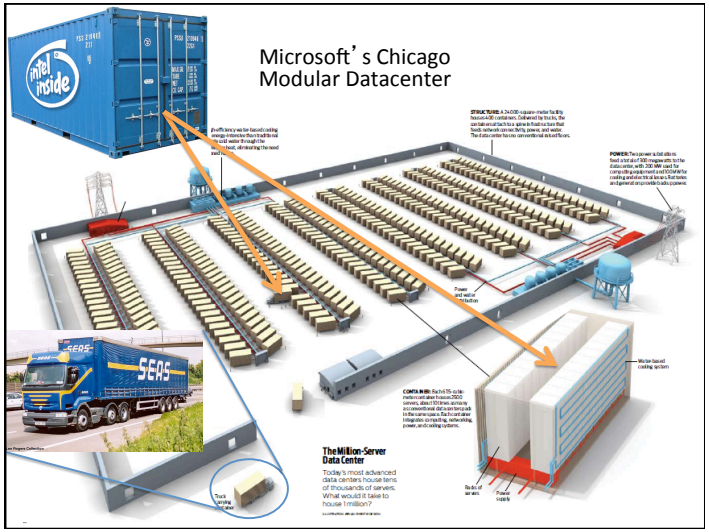
Google Server Internals



7

8

Microsoft's Chicago Modular Datacenter



Some Differences Between Commodity DC Networking and Internet/WAN

Characteristic	Internet/WAN	Commodity Datacenter
Latencies	Milliseconds to Seconds	Microseconds
Bandwidths	Kilobits to Megabits/s	Gigabits to 10's of Gbits/s
Causes of loss	Congestion, link errors, …	Congestion
Administration	Distributed	Central, single domain
Statistical Multiplexing	Significant	Minimal, 1-2 flows dominate links
Incast	Rare	Frequent, due to synchronized responses

10

Coping with Performance in Array

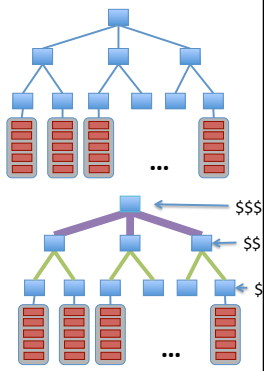
Lower latency to DRAM in another server than local disk
Higher bandwidth to local disk than to DRAM in another server

	Local	Rack	Array
Racks	--	1	30
Servers	1	80	2400
Cores (Processors)	8	640	19,200
DRAM Capacity (GB)	16	1,280	38,400
Disk Capacity (GB)	4,000	320,000	9,600,000
DRAM Latency (microseconds)	0.1	100	300
Disk Latency (microseconds)	10,000	11,000	12,000
DRAM Bandwidth (MB/sec)	20,000	100	10
Disk Bandwidth (MB/sec)	200	100	10

11

Datacenter design 101

- Naive topologies are tree-based
  same boxes, and same b/w links
  - Poor performance
  - Not fault tolerant
- An early solution; speed hierarchy (fewer expensive boxes at the top)
  - Boxes at the top run out of *capacity (bandwidth)*
  - but even the $ boxes needed
    $$$ abilities (forwarding table size)

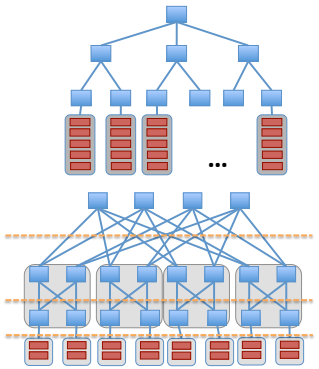


12

Data Center 102

- Tree leads to FatTree
- All bi-sections have same bandwidth

This is not the only solution…



13

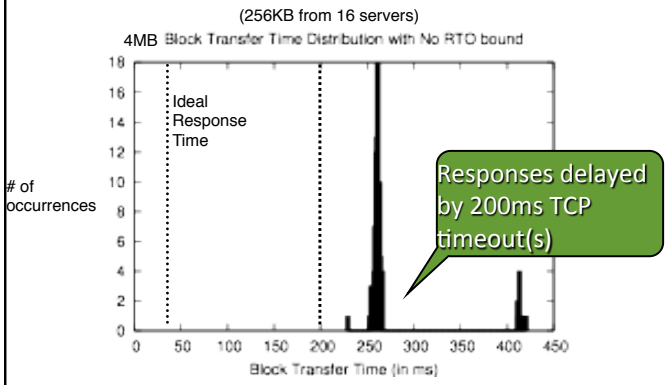
Latency-sensitive Apps

- Request for 4MB of data sharded across 16 servers (256KB each)
- How long does it take for all of the 4MB of data to return?

14

14

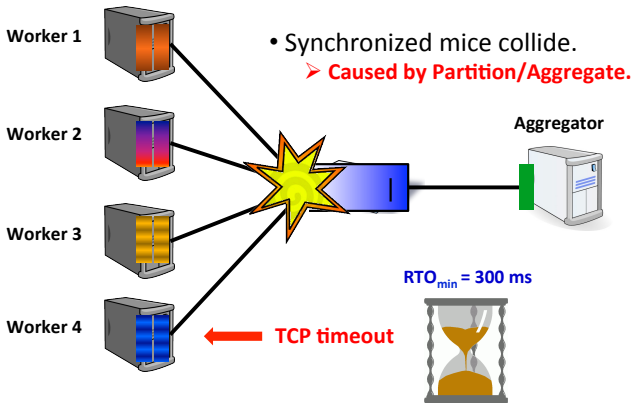
Timeouts Increase Latency



15

15

Incast



- Synchronized mice collide.
  - ➢ **Caused by Partition/Aggregate.**

16

16

Applications Sensitive to 200ms TCP Timeouts

- "Drive-bys" affecting single-flow request/ response
- Barrier-Sync workloads
  - Parallel cluster filesystems (Incast workloads)
  - Massive multi-server queries
    - Latency-sensitive, customer-facing

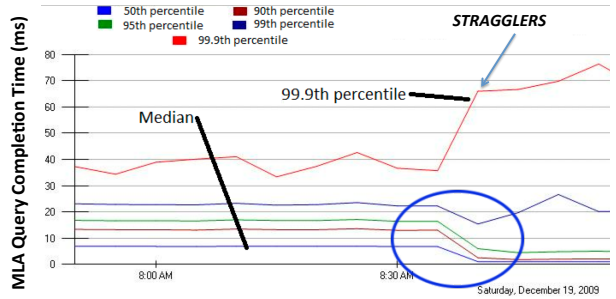
The ***last result delivered*** is referred to as a ***straggler***.

Stragglers can be caused by one off (drive-by) events but also by incast congestion which may occur for every map-reduce or database record retrieve or distributed filesystem read….

17

17

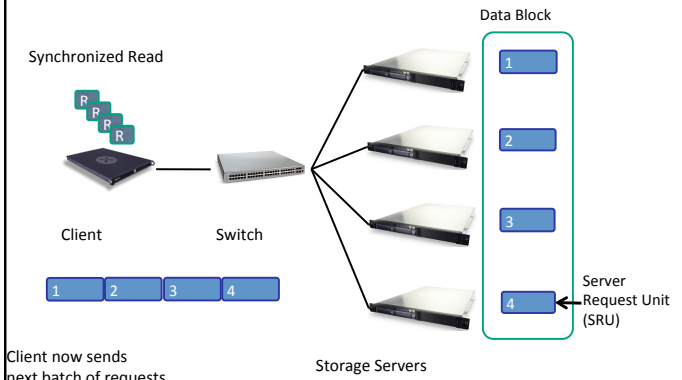
Incast Really Happens



Jittering 99.9th percentile is being tracked. ntiles.

18

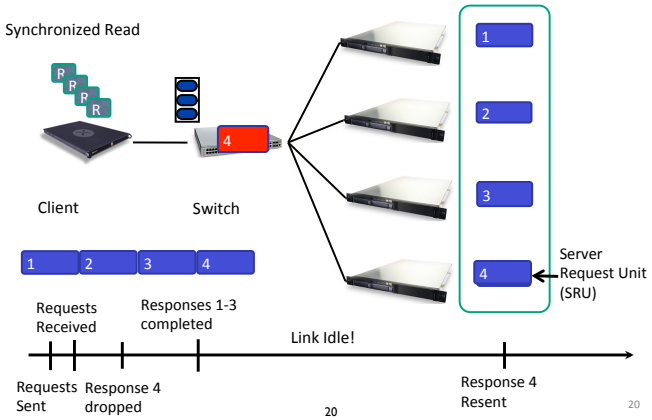
The Incast Workload



19

19

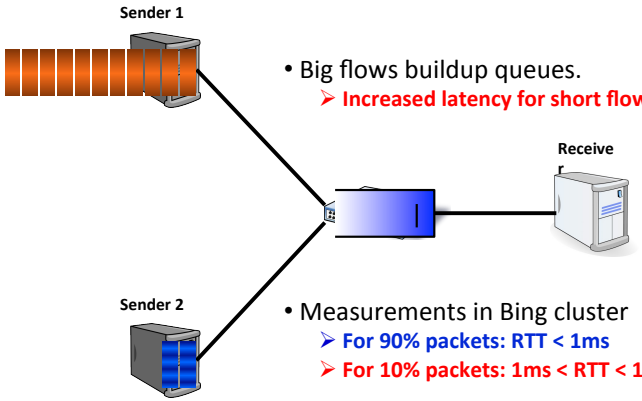
Incast Workload Overfills Buffers



20

20

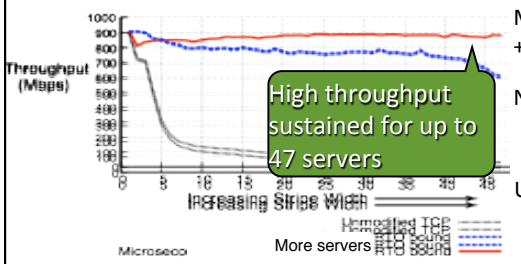
Queue Buildup



- Big flows buildup queues.
  - Increased latency for short flows.
- Measurements in Bing cluster
  - For 90% packets: RTT < 1ms
  - For 10% packets: 1ms < RTT < 15ms

21

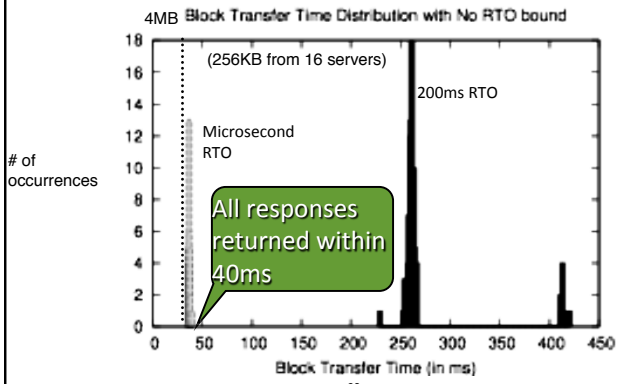
Microsecond timeouts are necessary



22

22

Improvement to Latency



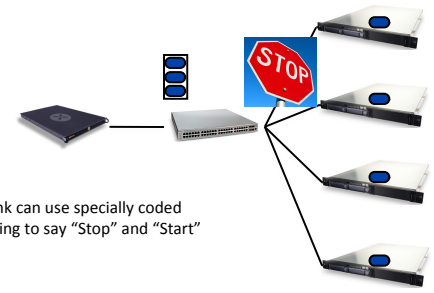
23

23

Link-Layer Flow Control

Common between switches but this is flow-control to the end host too…

- Another idea to reduce incast is to employ Link-Layer Flow Control…..

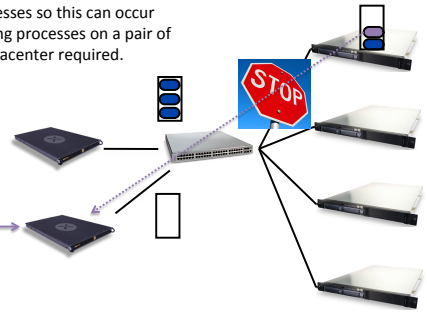


Recall: the Data-Link can use specially coded symbols in the coding to say "Stop" and "Start"

24

Link Layer Flow Control – The Dark side Head of Line Blocking….

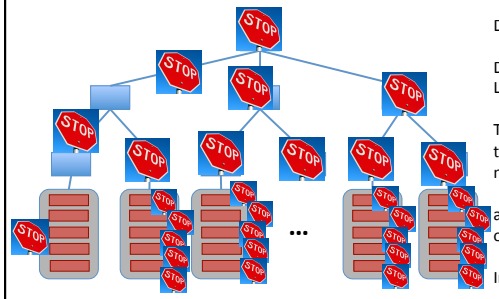
Such HOL blocking does not even differentiate processes so this can occur between competing processes on a pair of machines – no datacenter required.



Waiting for no good reason….

25

Link Layer Flow Control
But its worse that you imagine….



Double down on trouble….

Did I mention this is Link-Layer!

That means no (IP) control traffic, no routing messages….

a whole system waiting for one machine

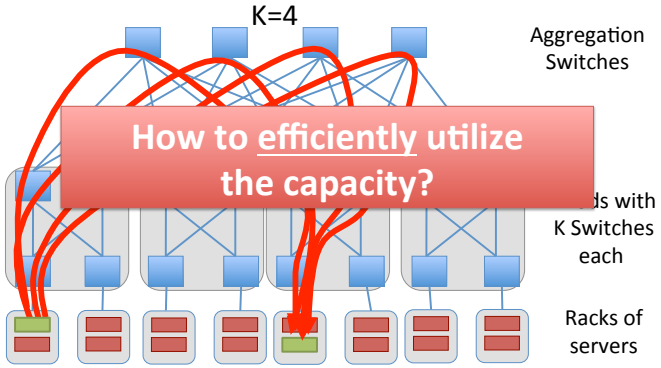
Incast is very unpleasant.

Reducing the impact of HOL in Link Layer Flow Control can be done through priority queues and *overtaking*….

26

Fat Tree Topology

(Fares et al., 2008; Clos, 1953)

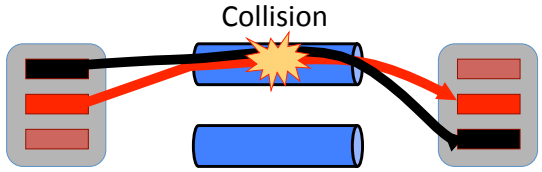


27

State of the Art

(as discussed in Hedera)

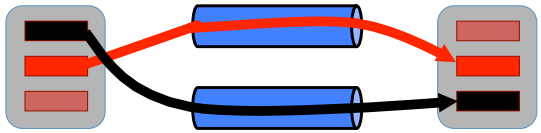
Statically stripe flows across available paths using ECMP



ECMP: Equal Cost Multi-Path Routing is common in Data Centers but network ninjas may be required to configure it correctly…

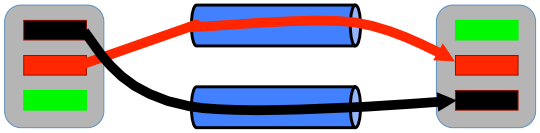
28

How about mapping each flow to a different path?



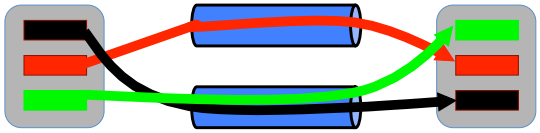
29

How about mapping each flow to a different path?



30

How about mapping each flow to a different path?

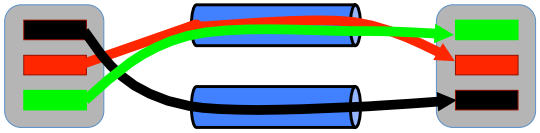


Not fair

31

How about mapping each flow to a different path?

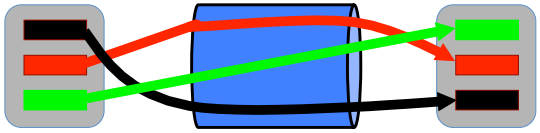
Not fair



Mapping each flow to a path is the wrong approach!

32

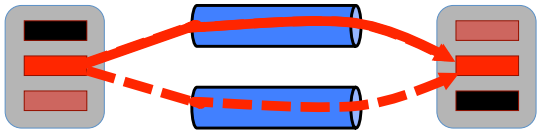
Instead, pool capacity from links



33

Multipath TCP Primer
(IETF MPTCP WG)

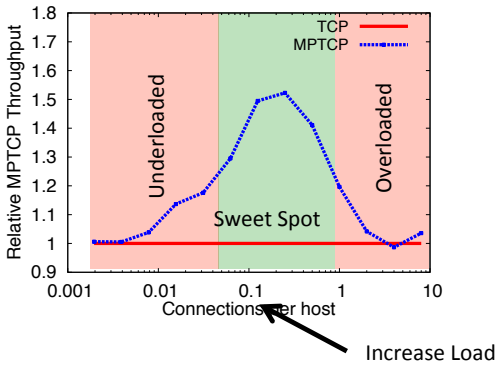
- A drop in replacement for TCP
- Spreads application data over multiple sub flows



- For each ACK on sub-flow r, increase the window w_r by $\min(\alpha/w_{total}, 1/w_r)$
- For each loss on sub-flow r, decrease the window w_r by $w_r/2$

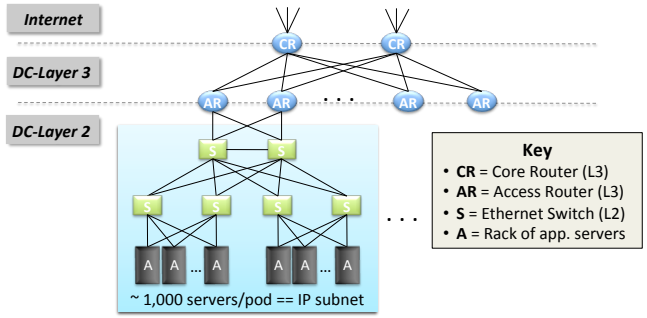
34

Performance improvements depend on traffic matrix



35

DC: lets add some of redundancy
The 3-layer Data Center

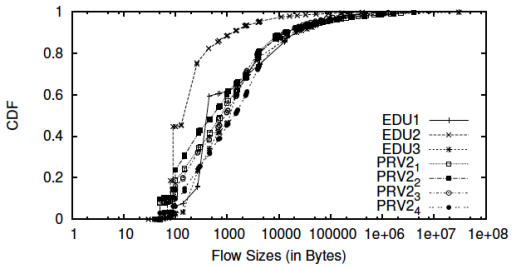


Reference – "Data Center: Load balancing Data Center Services", Cisco 2004

36

Understanding Datacenter Traffic

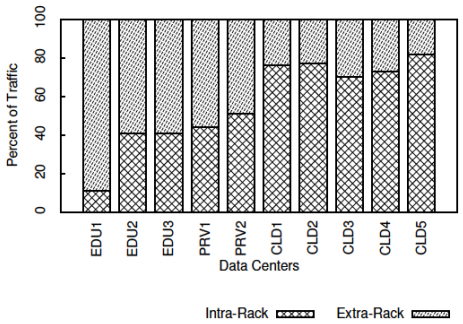
"Most flows in the data centers are **small in size** (<10KB)...". In other words, elephants are a very small fraction.



37

Understanding Datacenter Traffic

Majority of the traffic in cloud datacenters stay within the rack.



38

Today, Computation Constrained by Network*

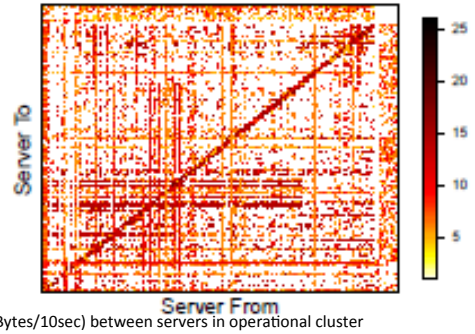


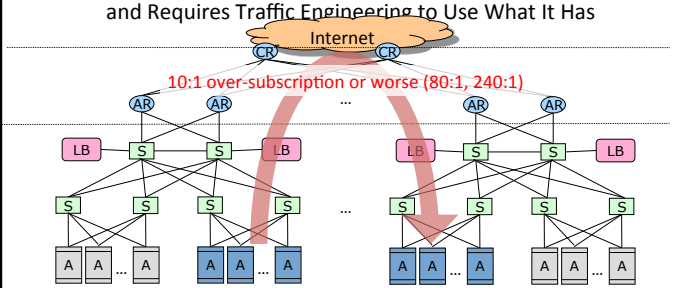
Figure: ln(Bytes/10sec) between servers in operational cluster

- Great efforts required to place communicating servers under the same ToR ➔ Most traffic lies on the diagonal (w/o log scale all you see is the diagonal)
- Stripes show there is need for inter-ToR communication

*Kandula, Sengupta, Greenberg,Patel

39

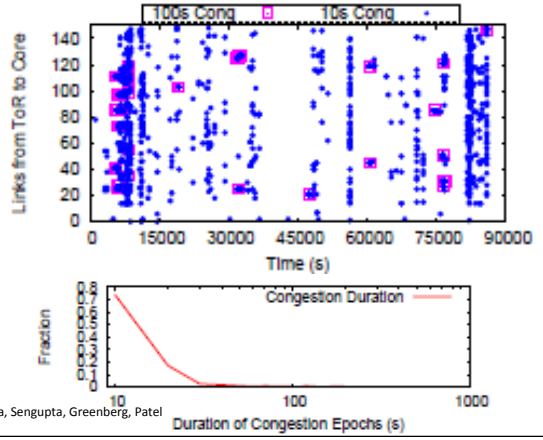
Network has Limited Server-to-Server Capacity, and Requires Traffic Engineering to Use What It Has



- Data centers run two kinds of applications:
  - Outward facing (serving web pages to users)
  - Internal computation (computing search index – think HPC)

40

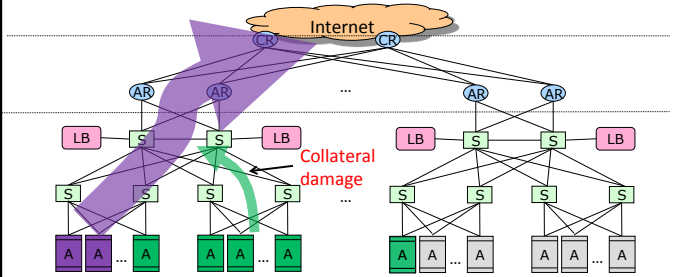
Congestion: Hits Hard When it Hits*



*Kandula, Sengupta, Greenberg, Patel

41

No Performance Isolation

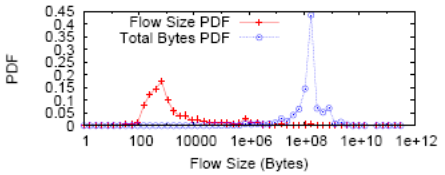


- VLANs typically provide reachability isolation only
- One service sending/receiving too much traffic hurts all services sharing its subtree

42

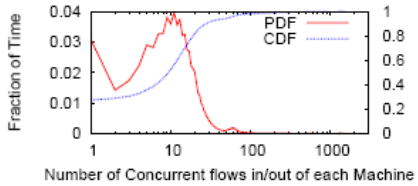
Flow Characteristics

DC traffic != Internet traffic



Most of the flows: various mice

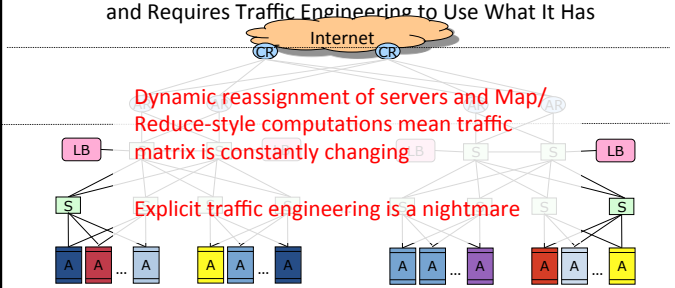
Most of the bytes: within 100MB flows



Median of 10 concurrent flows per server

43

Network Needs Greater Bisection BW, and Requires Traffic Engineering to Use What It Has

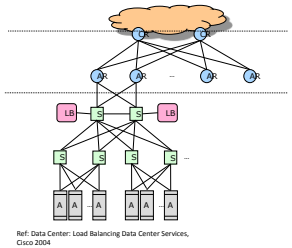


- Data centers run two kinds of applications:
  - Outward facing (serving web pages to users)
  - Internal computation (computing search index – think HPC)

44

What Do Data Center Faults Look Like?

- Need very high reliability near top of the tree
  - Very hard to achieve
    - Example: failure of a temporarily unpaired core switch affected ten million users for four hours
  - 0.3% of failure events knocked out all members of a network redundancy group



Ref: Data Center: Load Balancing Data Center Services, Cisco 2004

45

Data Center: Challenges

- From a large cluster used for data mining and identified distinctive traffic patterns
- Traffic patterns are **highly volatile**
  - A large number of distinctive patterns even in a day
- Traffic patterns are **unpredictable**
  - Correlation between patterns very weak

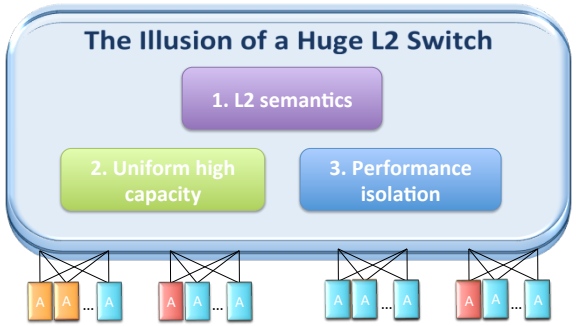
46

Data Center: Opportunities

- DC controller knows **everything** about **hosts**
- Host OS's are easily **customizable**
- **Probabilistic** flow distribution would work well enough, because …
  - Flows are numerous and not huge – no elephants!
  - Commodity switch-to-switch links are substantially thicker (~ 10x) than the maximum thickness of a flow

47

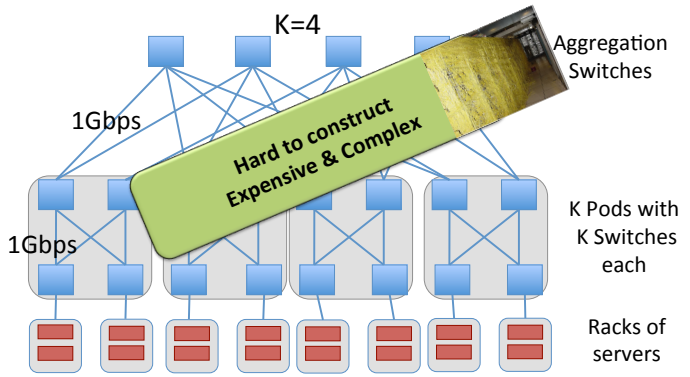
Sometimes we just wish we had a huge L2 switch (L2: data-link)



48

Fat Tree Topology

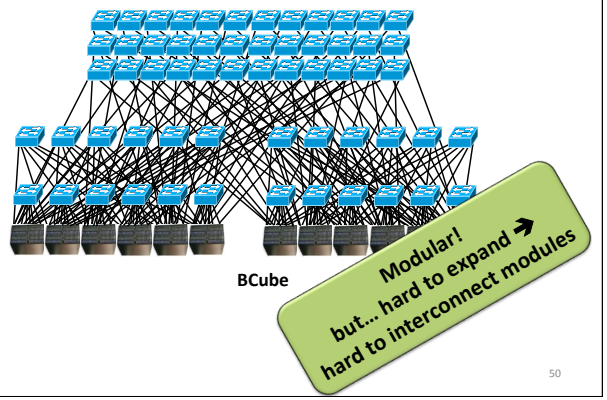
(Fares et al., 2008; Clos, 1953)



49

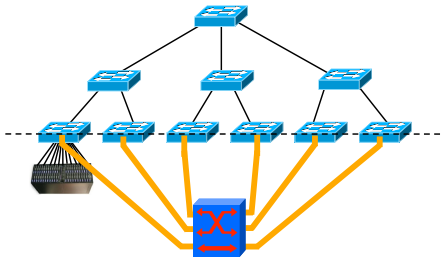
B-Tree Topology

(Guo et al., 2009)



50

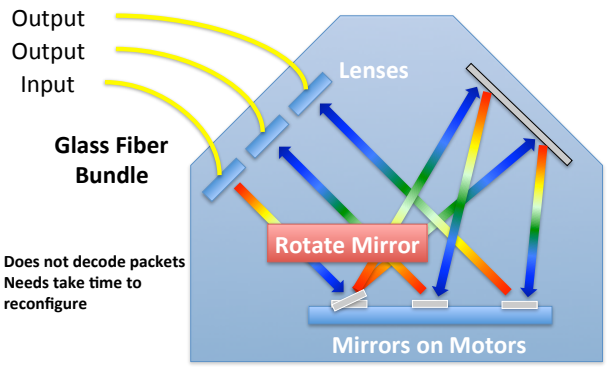
An alternative: hybrid packet/circuit switched data center network



- **Goal of this work:**
  - Feasibility: software design that enables efficient use of optical circuit
  - Applicability: application performance over a hybrid network

51

Optical Circuit Switch



- Does not decode packets
- Needs take time to reconfigure

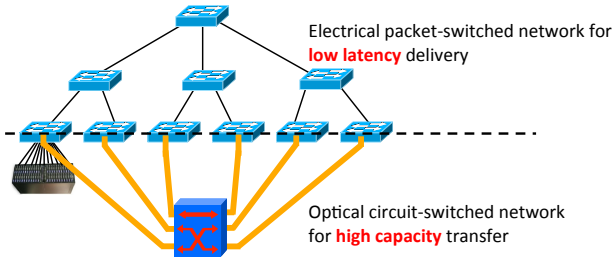
52

Optical circuit switching v.s. Electrical packet switching

	Electrical packet switching	Optical circuit switching
Switching technology	Store and forward	Circuit switching
Switching capacity	16x40Gbps at high end e.g. Cisco CRS-1	320x100Gbps on market, e.g. Calient FiberConnect
Switching time	Packet granularity	Less than 10ms
Switching traffic	For bursty, uniform traffic	For stable, pair-wise traffic

53

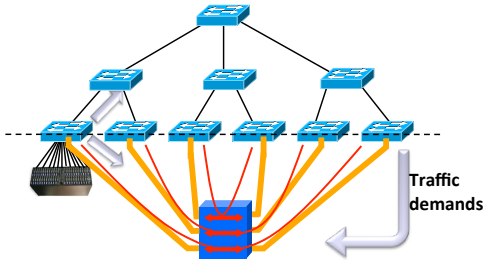
Hybrid packet/circuit switched network architecture



- Optical paths are provisioned rack-to-rack
  - A simple and cost-effective choice
  - Aggregate traffic on per-rack basis to better utilize optical circuits

54

Design requirements

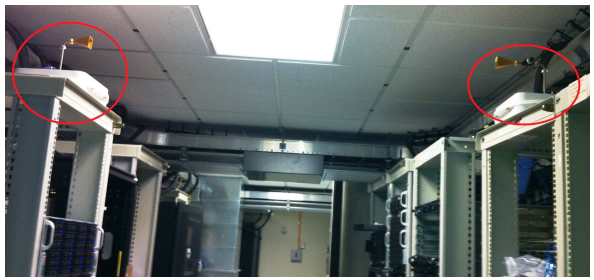


- Control plane:
  - Traffic demand estimation
  - Optical circuit configuration
- Data plane:
  - Dynamic traffic de-multiplexing
  - Optimizing circuit utilization (optional)

55

An alternative to Optical links: Link Racks by radio

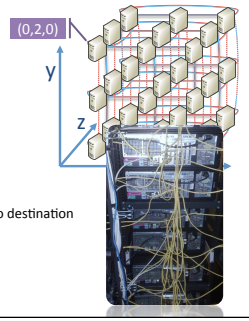
(60GHz gives about 2Gbps at 5m)



56

CamCube

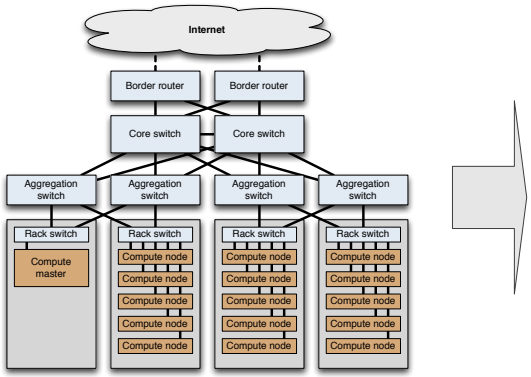
- Nodes are commodity x86 servers with local storage
  - Container-based model 1,500-2,500 servers
- Direct-connect 3D torus topology
  - Six Ethernet ports / server
  - Servers have (x,y,z) coordinates
    - Defines coordinate space
  - Simple 1-hop API
    - Send/receive packets to/from 1-hop neighbours
    - Not using TCP/IP
- Everything is a service
  - Run on all servers
- Multi-hop routing is a service
  - Simple link state protocol
  - Route packets along shortest paths from source to destination



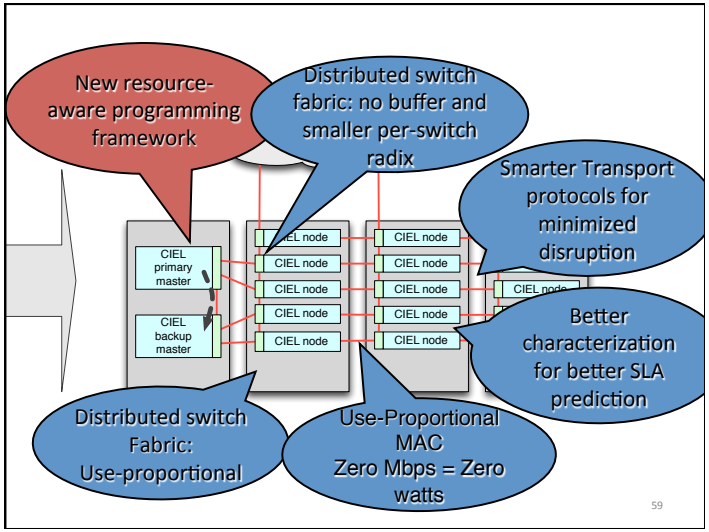
Two Downsides (there are others):
- complex to use (programming models…)
- messy to wire…

57

MRC^2: A ***new*** data center approach



58



59

Other problems

Special class problems/Solutions?

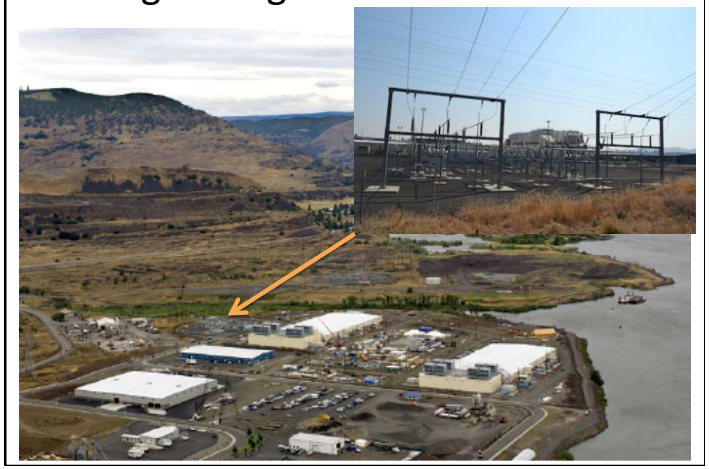
Datacenters are computers too…

What do datacenters to anyway?

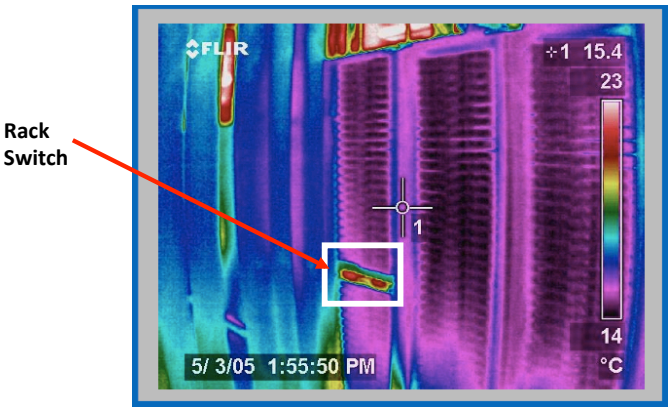
- Special class problems
- Special class data-structures
- Special class languages
- Special class hardware
- Special class operating systems
- Special class networks ✔
- Special class day-to-day operations

60

Google Oregon Warehouse Scale



Thermal Image of Typical Data Centre Rack



M. K. Patterson, A. Pratt, P. Kumar,
"From UPS to Silicon: an end-to-end evaluation of datacenter efficiency", Intel Corporation

62

DC futures

Warehouse-Scale Computing: Entering the Teenage Decade

Luiz Barroso (Google)

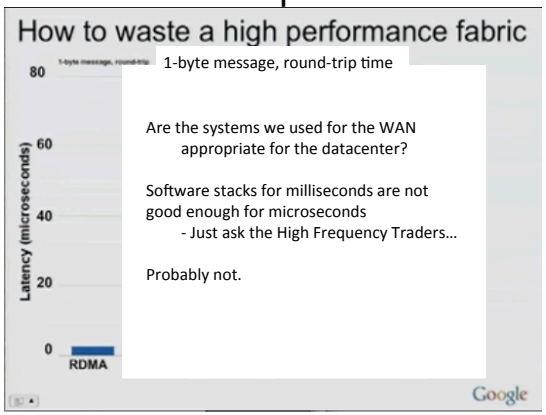
ISCA Keynote 2011

http://dl.acm.org/citation.cfm?id=2019527

It's a video. Watch it.

63

Sample



Are the systems we used for the WAN appropriate for the datacenter?

Software stacks for milliseconds are not good enough for microseconds
- Just ask the High Frequency Traders…

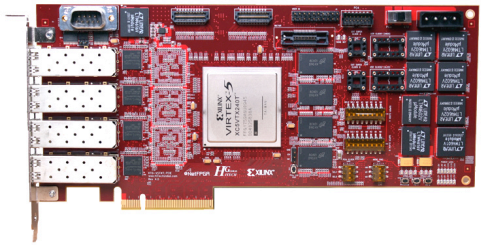
Probably not.

L. Barroso's excellent ISCA2011 keynote – go watch it! http://dl.acm.org/citation.cfm?id=2019527

64

Fancy doing somet exciting computer science? at 10Gbps? 40Gbps? 240Gbps?
Talk to me.

NetFPGA



65