Data Center Networks Are in My Way

Stanford Clean Slate CTO Summit

James Hamilton, 2009.10.23

VP & Distinguished Engineer, Amazon Web Services

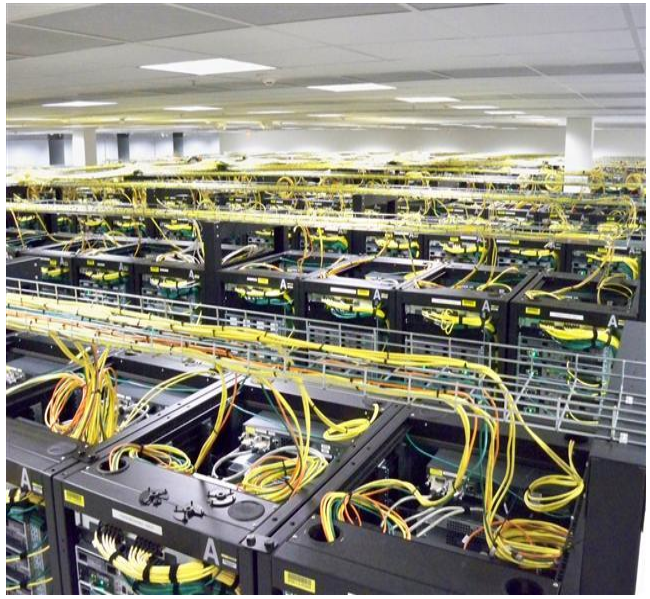
e: James@amazon.com

web: mvdirona.com/jrh/work

blog: perspectives.mvdirona.com

work with Albert Greenberg, Srikanth Kandula, Dave Maltz, Parveen Patel, Sudipta Sengupta, Changhoon Kim, Jagwinder Brar, Justin Pietsch, Tyson Lamoreaux, Dhiren Dedhia, Alan Judge, & Dave O'Meara

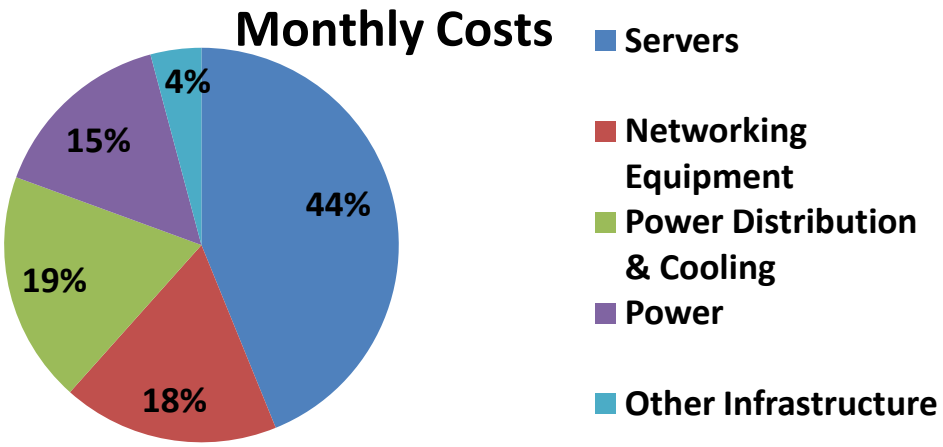
Agenda



- Where Does the Money Go?
  - Is net gear really the problem?
- Workload Placement Restrictions
- Hierarchical & Over-Subscribed
- Net Gear: SUV of the Data Center
- Mainframe Business Model
- Manually Configured & Fragile at Scale
- Problems on the Border
- Summary

Where Does the Money Go?

- **Assumptions:**
  - Facility: ~$200M for 15MW facility, 82% is power dist & mechanical (15-year amort.)
  - Servers: ~$2k/each, roughly 50,000 (3-year amort.)
  - Average server power draw at 30% utilization: 80%
  - Server to Networking equipment ratio: 2.5:1 ("Cost of a Cloud" data)
  - Commercial Power: ~$0.07/kWhr



3yr server & 15 yr infrastructure amortization

- **Observations:**
  - 62% per month in IT gear of which 44% in servers & storage
  - Networking 18% of overall monthly infrastructure spend

Details at: http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx
& http://perspectives.mvdirona.com/2009/03/07/CostOfACloudResearchProblemsInDataCenterNetworks.aspx

Where Does the Power Go?

- **Assuming a conventional data center with PUE ~1.7**
  - Each watt to server loses ~0.7W to power distribution losses & cooling
  - IT load (servers): 1/1.7=> 59%
  - Networking Equipment => 3.4% (part of 59% above)
- **Power losses are easier to track than cooling:**
  - Power transmission & switching losses: 8%
  - Cooling losses remainder:100-(59+8) => 33%
- **Observations:**
  - Server efficiency & utilization improvements highly leveraged
  - Cooling costs unreasonably high
  - Networking power small at <4%

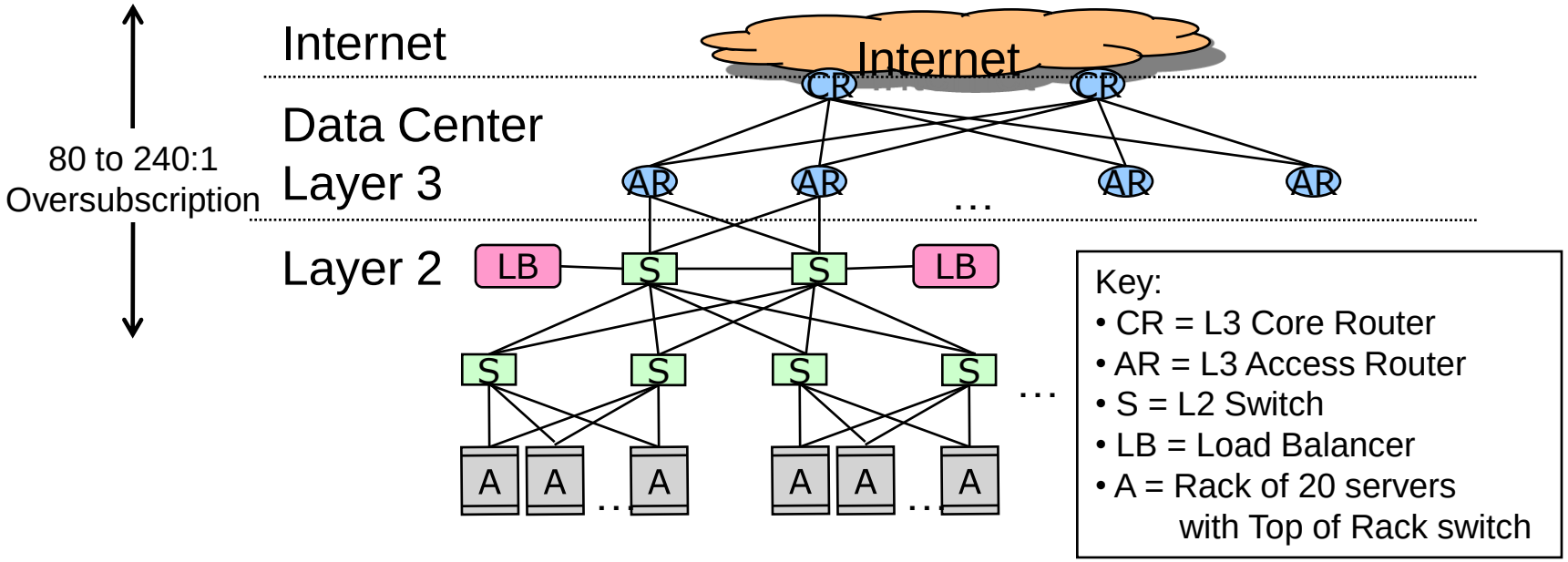
Is Net Gear Really the Problem?

- Networking represents only:
  - 18% of the monthly cost
  - 3.4% of the power
- Much improvement room but not dominant
  - Do we care?
- Servers: 55% Power & 44% monthly cost
  - Server utilization: 30% is good & 10% common
- Networking in way of the most vital optimizations
  - Improving server utilization
  - Supporting data intensive analytic workloads

Workload placement restrictions

- Workload placement over-constrained problem
  - Near storage, near app tiers, distant from redundant instances, near customer, same subnet (LB & VM Migration restrictions), …
- Goal: all data center locations equidistant
  - High bandwidth between servers anywhere in DC
  - Any workload any place
  - Need to exploit non-correlated growth/shrinkage in workload through dynamic over-provisioning
    - Resource consumption shaping
  - Optimize for server utilization rather than locality
- We are allowing the network to constrain optimization of the most valuable assets

Hierarchical & over-subscribed



- Poor net gear price/performance forces 80 to 240:1 oversubscription
- Constraints W/L placement and poor support for data intensive W/L
  - MapReduce, Data Warehousing, HPC, Analysis, ..
- MapReduce often moves entire multi-PB dataset during single job
- MapReduce code often not executing on node where data resides
- **Conclusion**: Need cheap, non-oversubscribed 10Gbps

Net gear: SUV of the data center

- Net gear incredibly power inefficient
- Continuing with Juniper EX8216 example:
  - Power consumption: 19.2kW/pair
  - Entire server racks commonly 8kW to 10kW
- But at 128 ports per switch pair, 150W/port
- Typically used as aggregation switch
  - Assume pair, each with 110 ports “down” & 40 servers/rack
  - Only: 4.4W/server port in pair configuration
- Far from dominant data center issue but still conspicuous consumption

Mainframe Business Model

Net Equipment

Central Logic Manufacture
- Proprietary & closely guarded
- Single source

Finished Hardware Supply
- Proprietary & closely guarded
- Single source

System Software Supply
- Proprietary & closely guarded
- Single source

Application Stack
- Not supported
- No programming tools
- No 3rd party ecosystem

Commodity Server

Central Logic Manufacture
- Standard design (x86)
- Multiple source
  - AMD, Intel, Via, …

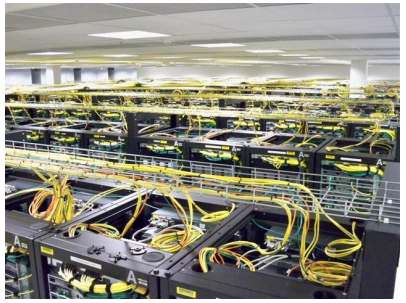
Finished Hardware Supply
- Standard design
- Multiple source
  - Dell, SGI, HP, IBM, …

System Software Supply
- Linux (many distros/support)
- Windows & other proprietary offerings

Application Stack
- Public/published APIs
- High quality prog tools
- Rich 3rd party ecosystem

- **Example**:
  - Juniper EX 8216 (used in core or aggregation layers)
  - Fully configured list: $716k w/o optics and $908k with optics
- **Solution**: Merchant silicon, H/W independence, open source protocol/mgmt stack

Manually Configured & Fragile at Scale



- Unaffordable, scale-up model leads to 2-way redundancy
  - Recovery oriented computing (ROC) better beyond 2-way
- Brownout & partial failure common
  - Neither false positives nor negatives acceptable & perfect is really hard
  - Unhealthy equipment continues to operate & drop packets
- Complex protocol stacks, proprietary extensions, and proprietary mgmt
  - Norm is error-prone manual configuration
- Networking uses a distributed management model
  - Complex & slow to converge
  - Central, net & app aware mgmt is practical even in large DCs (50k+ servers)
  - Want application input (priorities, requirements, ….)
- Scale-up reliability gets expensive faster than reliable
  - Asymptotically approaches "unaffordable" but never "good enough"
  - ROC management techniques work best with more than 2-way redundancy

Problems on the Border

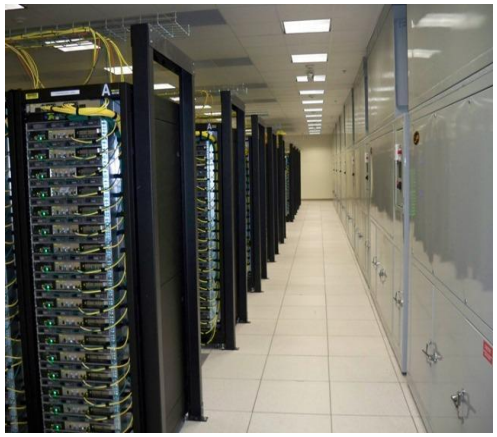
- All the problems of internal network but more:
  - Need large routing tables (FIBS in 512k to 1M range)
  - "Need" large packet buffers (power & cost)
  - Mainframe Router price point
    - Example: Cisco 7609
    - Fairly inexpensive border router
    - List price ~$350k for 32 ports or $11k/port
  - Mainframe DWDM optical price point
    - Example: Cisco 15454
    - List ~$489k for 8 ports or $61k/lambda (10Gbps)
    - Better at higher lambda counts but usually not needed
- High cost of WAN bandwidth serious industry issue
- DNS & Routing fragility (attacks & errors common)

Summary

- We are learning (again) scale-up doesn't work
  - Costly
  - Insufficiently robust
- We are learning (again) that a single-source, vertically integrated supply chain is a bad idea

- The ingredients for solution near:
  - Merchant silicon broadly available
  - Distributed systems techniques
    - Central control not particularly hard even at 10^5 servers
  - Standardized H/W platform layer (OpenFlow)
- Need an open source protocol & mgmt stack

More Information



- **This Slide Deck:**
  - I will post these slides to http://mvdirona.com/jrh/work later this week
- **VL2: A Scalable and Flexible Data Center Network**
  - http://research.microsoft.com/pubs/80693/vl2-sigcomm09-final.pdf
- **Cost of a Cloud: Research Problems in Data Center Networks**
  - http://ccr.sigcomm.org/online/files/p68-v39n1o-greenberg.pdf
- **PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric**
  - http://cseweb.ucsd.edu/~vahdat/papers/portland-sigcomm09.pdf
- **OpenFlow Switch Consortium**
  - http://www.openflowswitch.org/
- **Next Generation Data Center Architecture: Scalability & Commoditization**
  - http://research.microsoft.com/en-us/um/people/dmaltz/papers/monsoon-presto08.pdf
- **A Scalable, Commodity Data Center Network**
  - http://cseweb.ucsd.edu/~vahdat/papers/sigcomm08.pdf
- **Data Center Switch Architecture in the Age of Merchant Silicone**
  - http://www.nathanfarrington.com/pdf/merchant_silicon-hoti09.pdf
- **Berkeley Above the Clouds**
  - http://perspectives.mvdirona.com/2009/02/13/BerkeleyAboveTheClouds.aspx
- **James' Blog:**
  - http://perspectives.mvdirona.com
- **James' Email:**
  - James@amazon.com