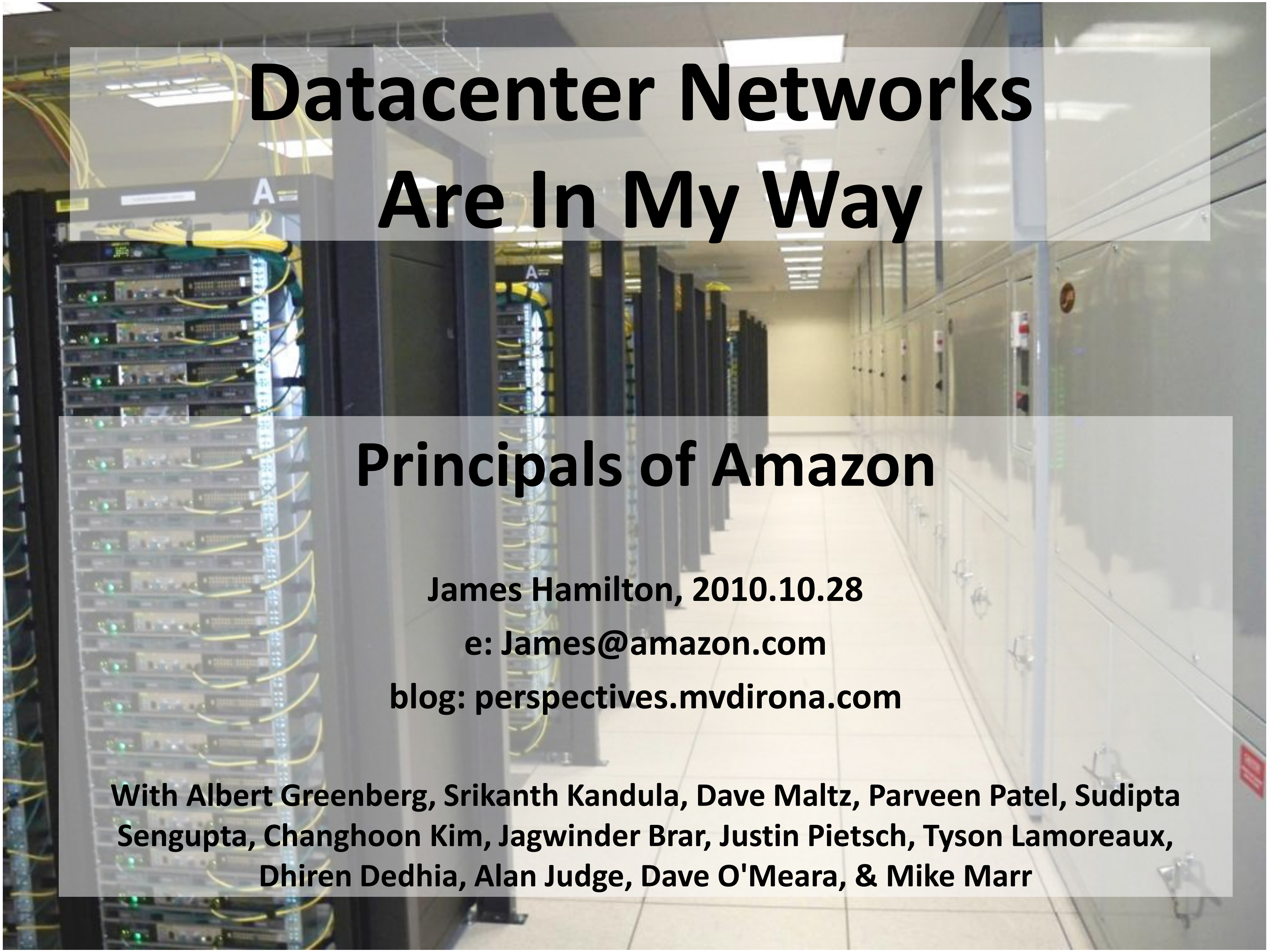


Datacenter Networks Are In My Way

Principals of Amazon

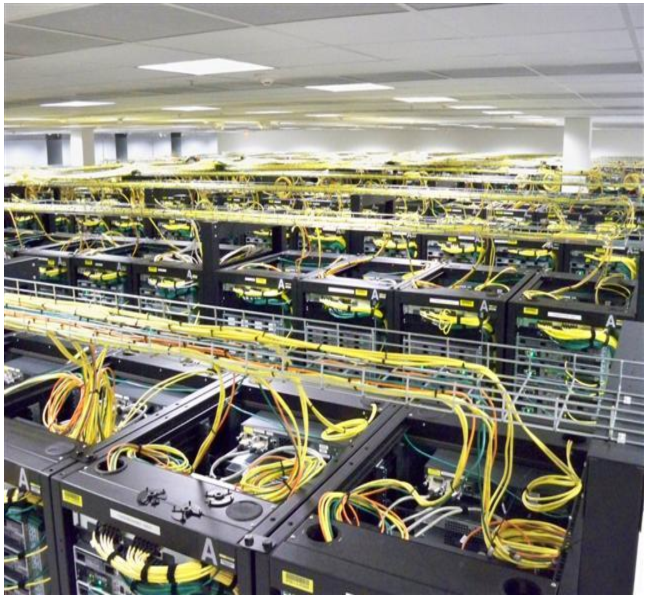
James Hamilton, 2010.10.28

e: James@amazon.com

blog: perspectives.mvdirona.com

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, & Mike Marr

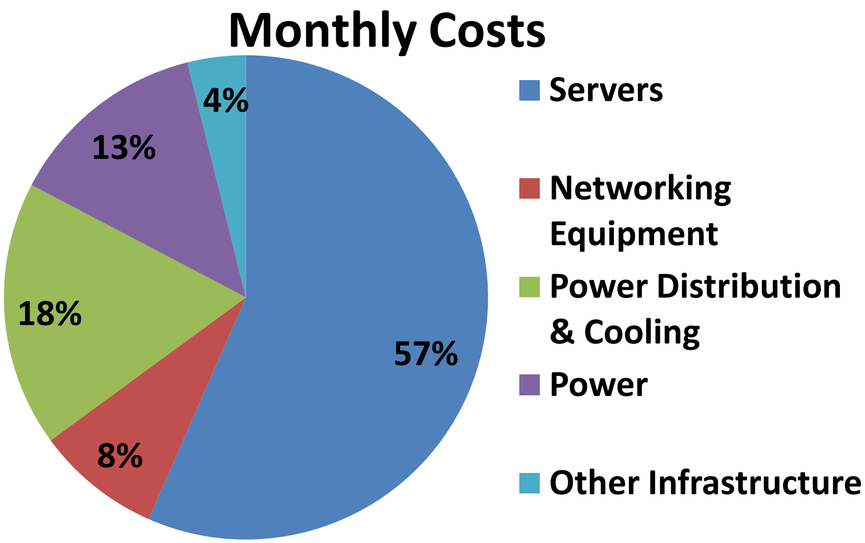
Agenda



- Datacenter Economics
- 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
- New Architecture for Networking

Datacenter Economics

- **Assumptions:**
  - Facility: ~$72M for 8MW critical power
  - Servers: 46,000 @ $1.45k each
  - Commercial Power: ~$0.07/kWhr
  - Power Usage Effectiveness: 1.45



3yr server & 10 yr infrastructure amortization

- **Observations:**
  - 31% costs functionally related to power (trending up while server costs down)
  - Networking high at 8% of costs & 19% of total server cost (much more with inter-DC)

Source: http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx

Is Net Gear Really the Problem?

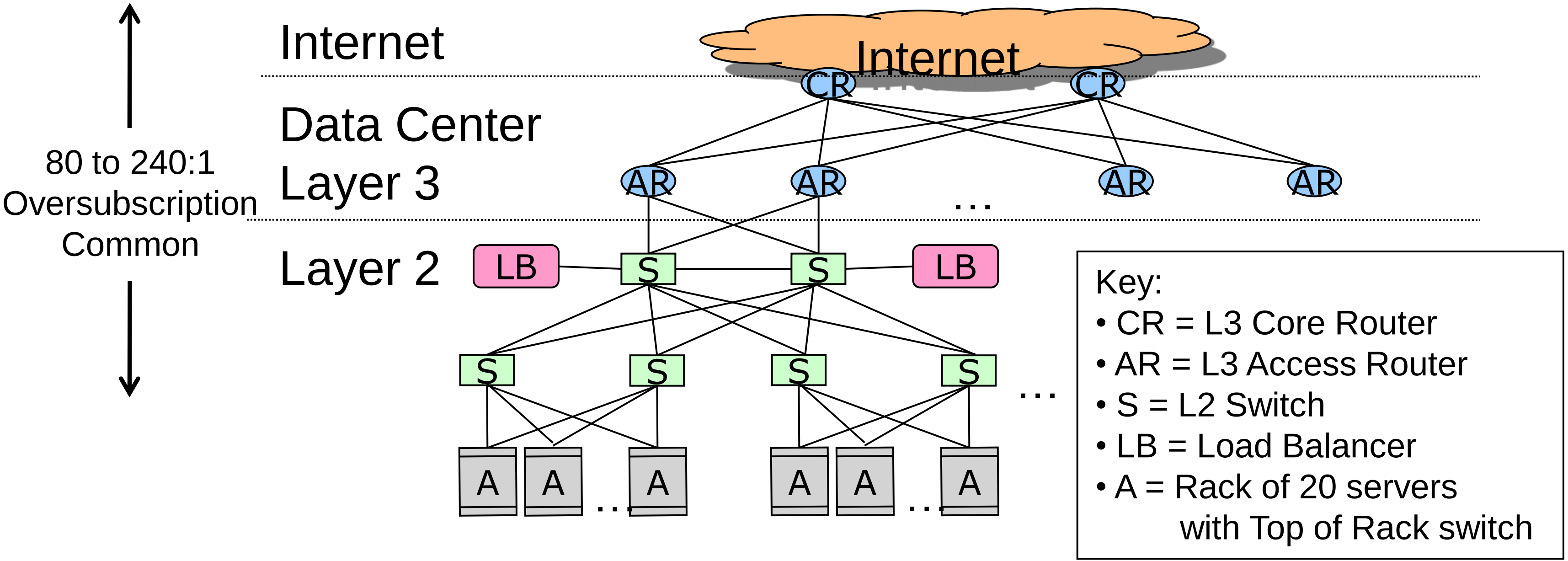
- Inside the DC net gear represents only:
  - 8% of the monthly cost
  - 5.2% of the power
- Improvement needed but not dominant
- Servers: 64% Power & 57% monthly cost
  - Low server utilization Low: 30% good, 10% common
- Networking in way of the most vital optimizations
  - Improving server utilization
  - Supporting data intensive analytic workloads

Source: http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx

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
  - Optimize for server utilization rather than locality
- We are allowing the network to constrain optimization of the most valuable assets

Hierarchical & Over-Subscribed



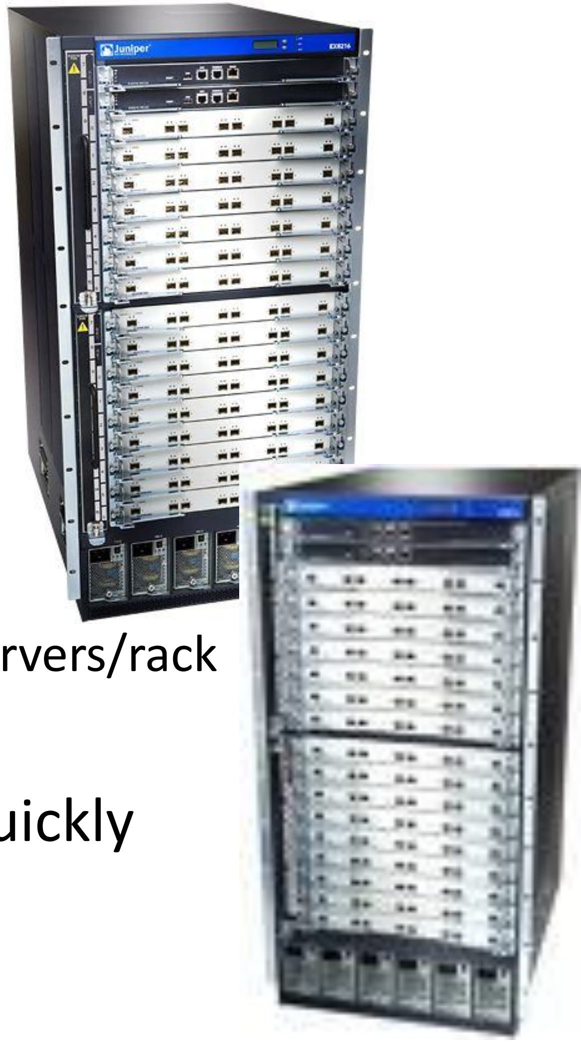
80 to 240:1
Oversubscription
Common

Key:
- CR = L3 Core Router
- AR = L3 Access Router
- S = L2 Switch
- LB = Load Balancer
- A = Rack of 20 servers with Top of Rack switch

Albert Greenberg Diagram

- Poor net gear price/performance forces oversubscription
- Oversubscription:
  - Constrains workload placement
  - Support for data intensive workloads poorly
    - MapReduce often moves entire multi-PB dataset during single job
    - MapReduce, HPC, Analysis, MPP database,…
- **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
- Often used as aggregation switch
  - Assume pair, each with 110 ports “down” & 40 servers/rack
  - Only: 4.4W/server port in pair configuration
- Network power consumption is increasing quickly
- Far from dominant data center issue but still conspicuous consumption

Mainframe Business Model

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

Net Equipment

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

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

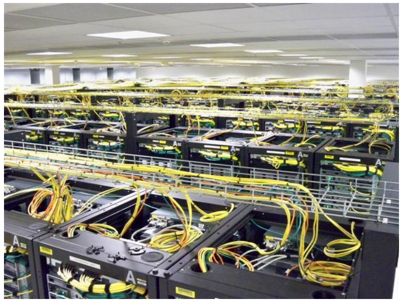
System Software Supply
- Linux (many distros/support)
- Windows & other proprietary offerings

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

Commodity Server

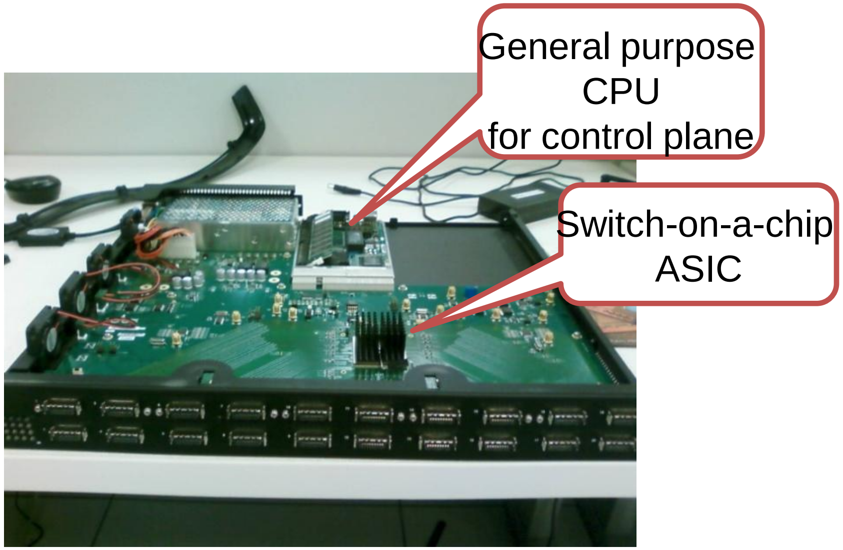
- **Example**:
  - Juniper EX 8216 (used in core or aggregation layers)
  - Fully configured list: $716k w/o optics and $908k with optics
- **Solution**: Merchant silicon, H/W independence, open source protocol/mgmt stack

Manually Configured & Fragile at Scale



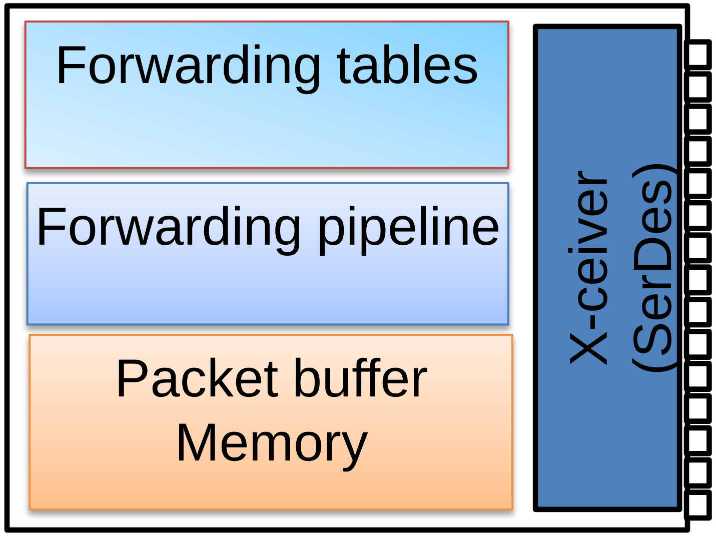
- Unaffordable, scale-up model yields 2-way redundancy
  - ROC needs more than 2-way
- Brownout & partial failure common
  - Unhealthy equipment continues to operate & drop packets
- Complex protocol stacks, proprietary extensions, and proprietary mgmt
  - Norm is error-prone manual configuration
- Networking distributed management model
  - Complex & slow to converge
  - Central, net & app aware mgmt is practical even in large DCs (50k+ servers)
  - Need application input (QOS, priorities, requirements, ….)
- Scale-up reliability gets expensive faster than reliable
  - Asymptotically approaches "unaffordable" but never gets to "good enough"
  - ROC management techniques work best with more than 2-way redundancy

What Enables New Solution Now?



General purpose CPU for control plane

Switch-on-a-chip ASIC



Forwarding tables

Forwarding pipeline

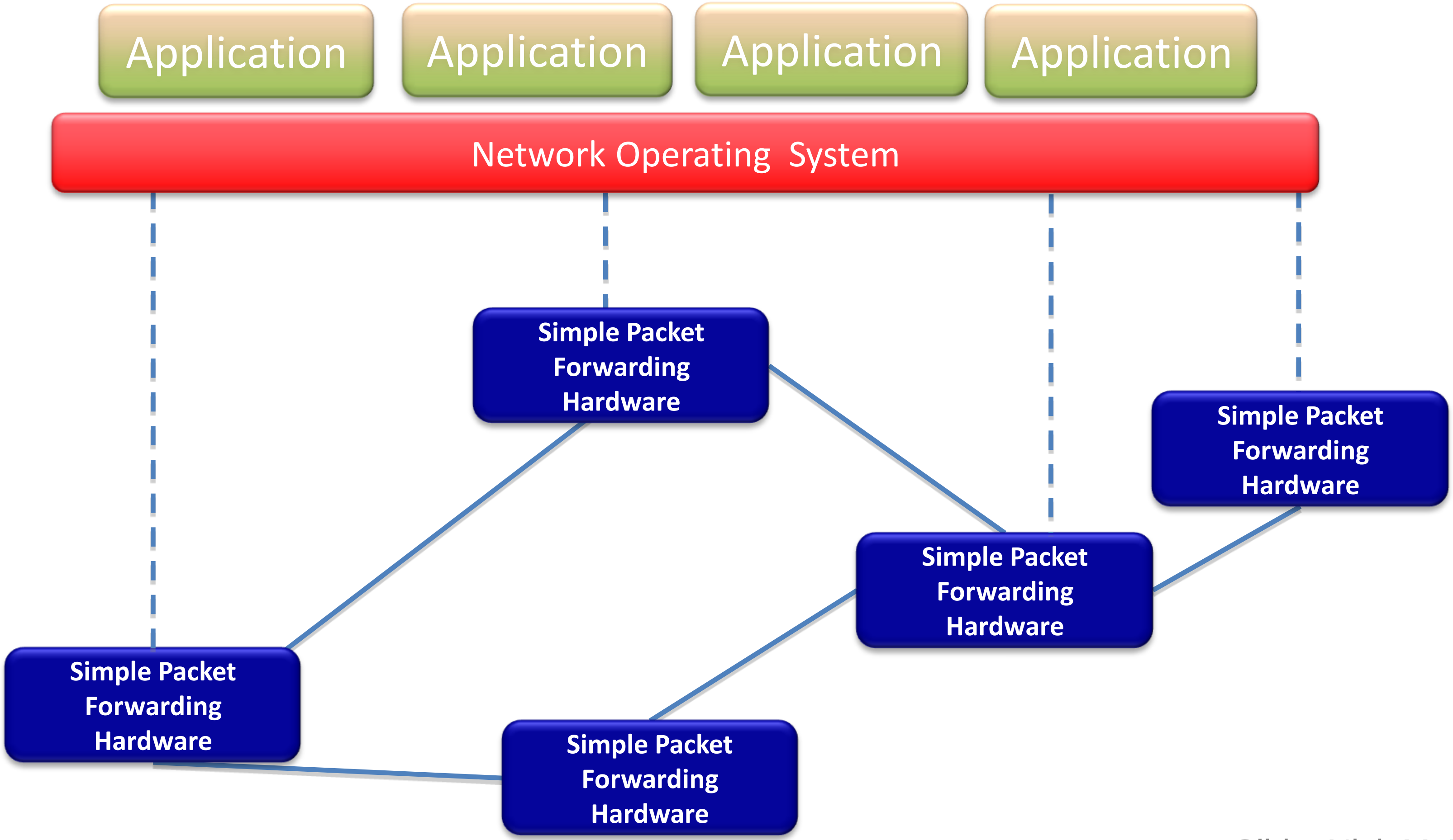
Packet buffer Memory

X-ceiver (SerDes)

- Last year:
  - 24 port 1G, 4 10G 16K IPv4 fwd entries, 2 MB buff
- This year:
  - 48 port 10G, 16K fwd entries, 4 MB buff
- Next year:
  - 64 to 96 port 10G
- ASIC cost roughly constant while port count scales every 18 to 24 months

Slide: Albert Greenberg

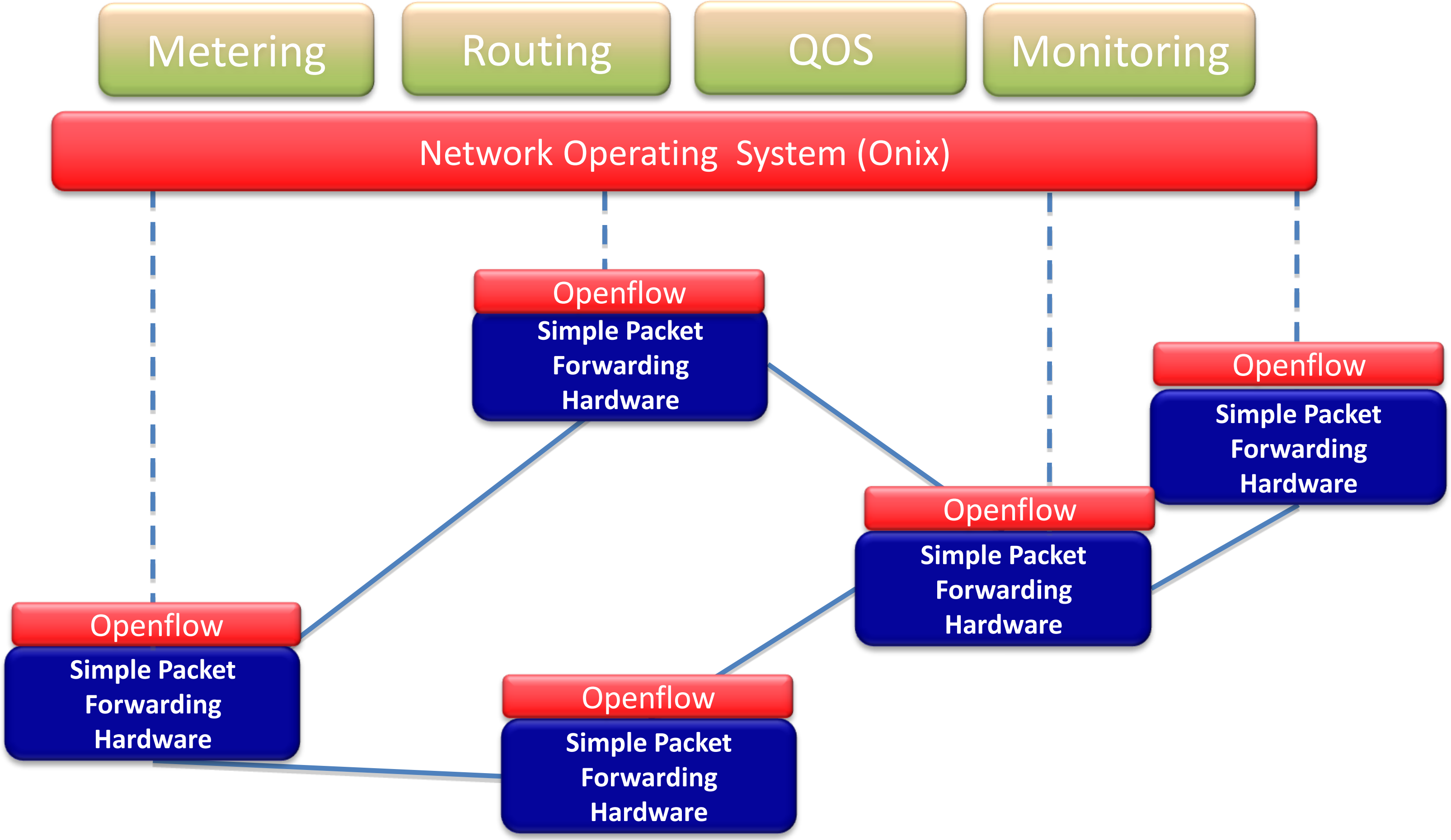
Software-Defined Networks



Application
Application
Application
Application
Network Operating System
Simple Packet Forwarding Hardware
Simple Packet Forwarding Hardware
Simple Packet Forwarding Hardware
Simple Packet Forwarding Hardware
Simple Packet Forwarding Hardware

Slide: Nick McKeown

OpenFlow/Onix



Metering

Routing

QOS

Monitoring

Network Operating System (Onix)

Openflow

Simple Packet Forwarding Hardware

Openflow

Simple Packet Forwarding Hardware

Openflow

Simple Packet Forwarding Hardware

Openflow

Simple Packet Forwarding Hardware

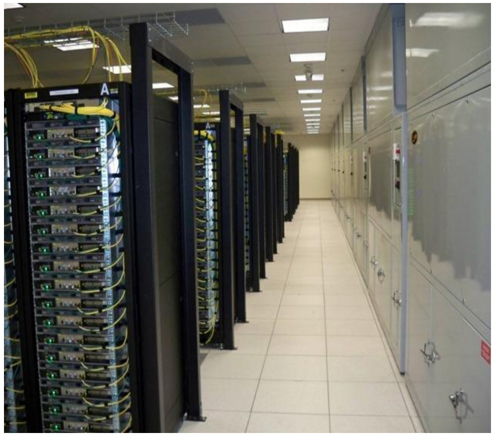
Openflow

Simple Packet Forwarding Hardware

Summary

- Again we learn:
  - scale-up doesn’t work
  - single-source, vertically integrated is bad idea
- Ingredients for solution near:
  - Merchant silicon broadly available
  - Distributed systems techniques
  - Standardized H/W platform layer (OpenFlow)
- Need an open source protocol & mgmt stack

More Information



- **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
- **Blog & Email:**
  - http://perspectives.mvdirona.com & James@amazon.com