Internet Scale Storage

Microsoft Storage Community

James Hamilton, 2011/11/30

VP & Distinguished Engineer, Amazon Web Services

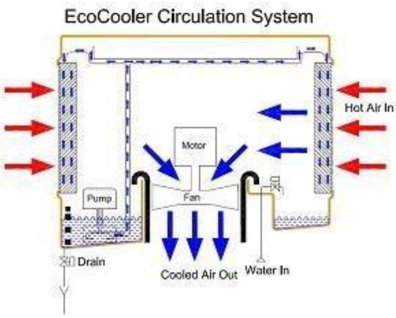
email: James@amazon.com

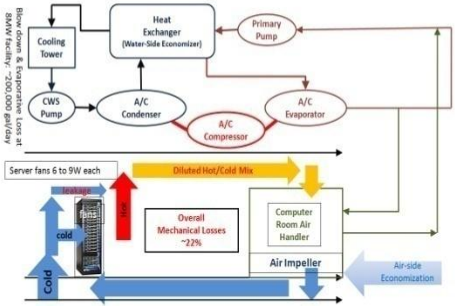
web: mvdirona.com/jrh/work

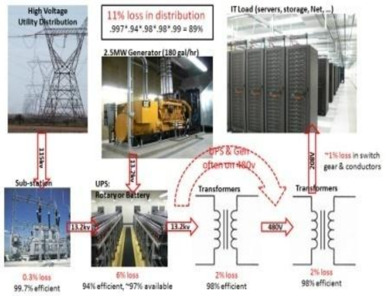
blog: perspectives.mvdirona.com

Agenda

- **Cloud & Accelerating Pace of Innovation**
- Technology Changes
  - Memory wall & Storage Chasm
  - Disk is Tape
  - Sea Change in Networking
- Data & Storage Trends
  - Map Reduce & NoSQL
  - Migration to Cloud







Talk does not necessarily represent positions of current or past employers

The DB World is on Fire Again

- One Size does not fit all
  - Stonebraker showed >3 DB companies actually possible
  - Customers willing to support multiple DBMS
- 30 year old architectural decisions no longer valid
  - Memories exploding
  - Disk IOPS density going backwards
    - 1990 Seagate ST41600: 37.5 IOPS/GB
    - 2007 Seagate ST373453 : 2.4 IOPS/GB
- Plunging cost of computing
- Cloud computing accelerates all above

Plunging Cost of Computing

- Rapidly declining cost of computing
  - Technology & cloud computing economies of scale
- Warehouse & analytical use scales inversely with cost
  - Lower costs supports more data & deeper analysis
- Traditional transactional systems scale with business
  - Purchases, ad impressions, pages served, etc.
  - Machine-to-machine transactions scale faster limited only by value of transaction & cost (e.g. computational trading)

Cloud Computing Driving Wave of Innovation & Growth

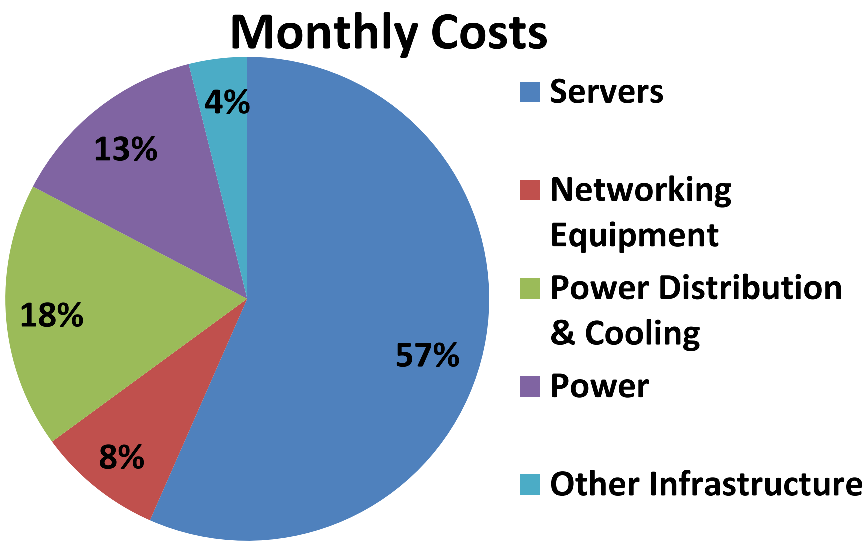
- Datacenter pace of innovation increasing
  - More innovation in last 5 years than previous 15
  - Driven by cloud service providers & very high-scale internet applications like search
- Not just a cost center
  - At scale, focus on cost
  - Mechanical, power, server, & net specialists
- Server, Storage, & infrastructure costs falling fast
- Data really is the challenge
  - Scaling stateless applications easy

Perspective on Scaling

Each day Amazon Web Services adds enough new capacity to support all of Amazon.com's global infrastructure through the company's first 5 years, when it was a $2.76B annual revenue enterprise

Where Does the Money Go at Scale?

- **Assumptions:**
  - Facility: ~$88M 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

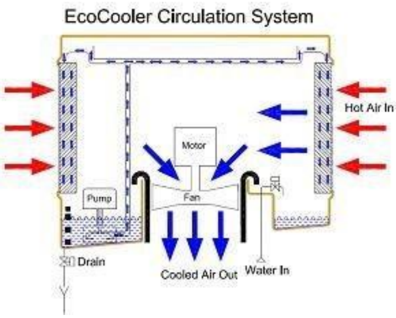
- **Observations:**
  - 31% costs functionally related to power (trending up while server costs down)
  - Networking high at 8% of overall costs & 19% of total server cost (many pay more)

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

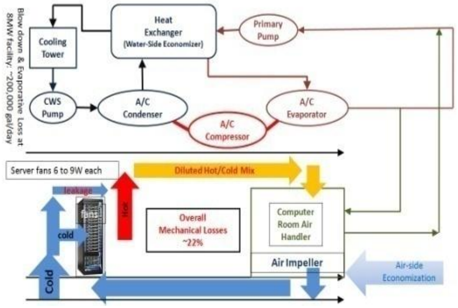
Agenda

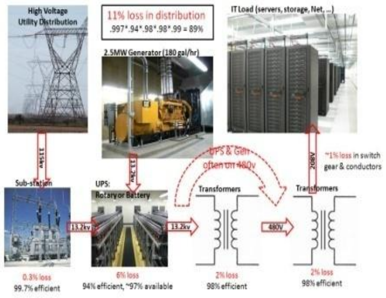
- Cloud & Accelerating Pace of Innovation

- Technology Changes
  - Memory wall & Storage Chasm
  - Disk is Tape
  - Sea Change in Networking

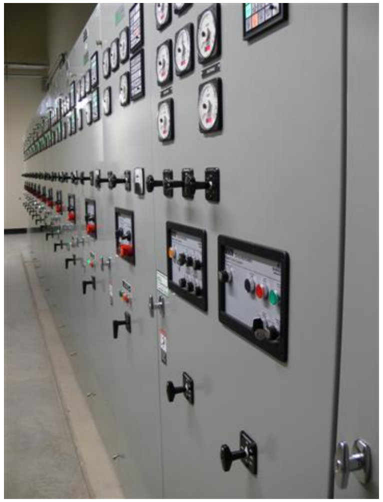


- Data & Storage Trends
  - Map Reduce & NoSQL
  - Migration to Cloud

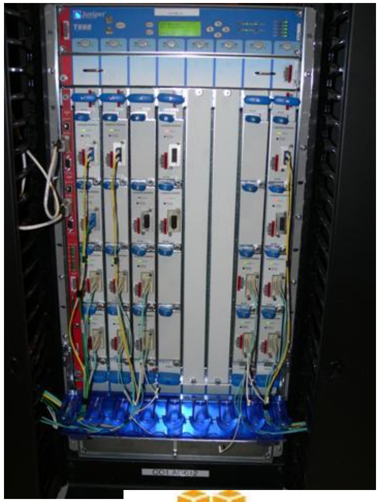




Limits to Computation



- Processor cycles are cheap and getting cheaper
- What limits application of infinite cores?
  1. **Data**: inability to get data to processor when needed
  2. **Power**: cost rising and will dominate
- Most sub-Moore attributes need most innovation
  - Infinite cores require infinite power
  - Getting data to processors in time to use next cycle:
    - Caches, multi-threading, ILP,…
      - All techniques consume power
    - More memory lanes drives b/w but more pins costs power
- Power & data movement remain key constraints



Storage & Memory B/W lagging CPU

	CPU	DRAM	LAN	Disk
Annual bandwidth improvement (all milestones)	1.5	1.27	1.39	1.28
Annual latency Improvement (all milestones)	1.17	1.07	1.12	1.11

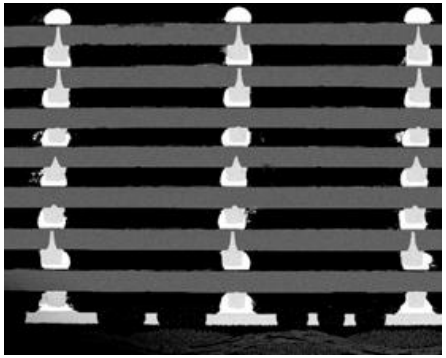
Memory wall

Storage Chasm

- CPU B/W requirements out-pacing memory and storage
- Disk & memory getting "further" away from CPU
  - Core limiting factor: power consumption & data availability
  - Powered CPU cores have no value without data
- Large sequential transfers better for both memory & disk

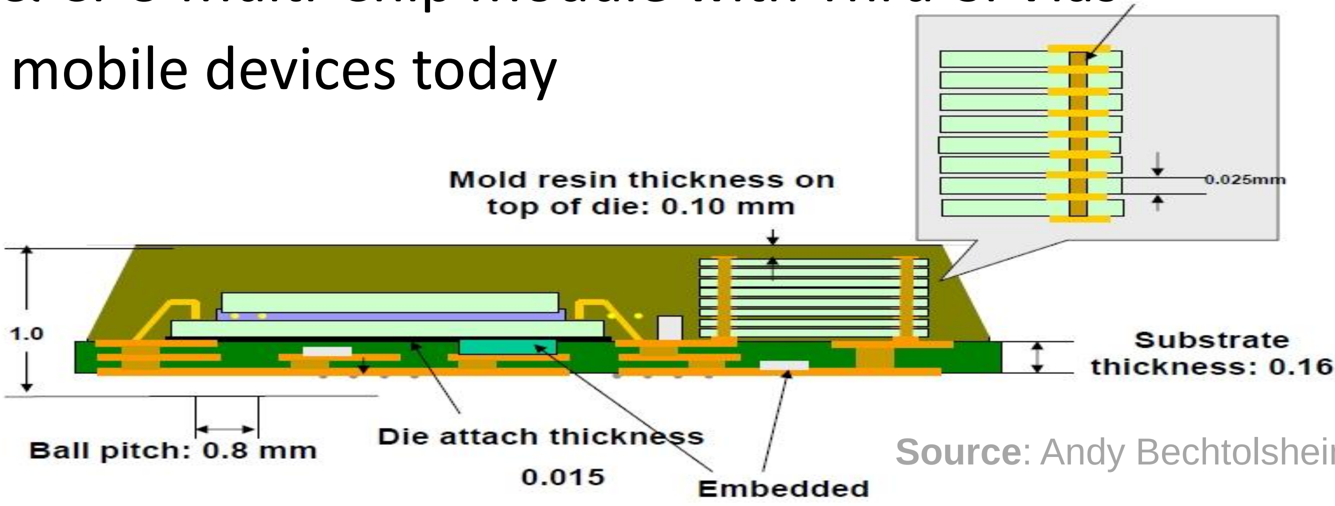
Source: Dave Patterson: Why Latency Lags Bandwidth and What It Means to Computing (25 yrs)

Memory Wall



Multi-Chip Module

- Adding processor I/O pins has a positive impact but at significant power cost
  - Positive but bounded impact
- Taming the memory wall:
  - Mem & CPU Multi-Chip Module with Thru-Si Vias
  - Lab & mobile devices today

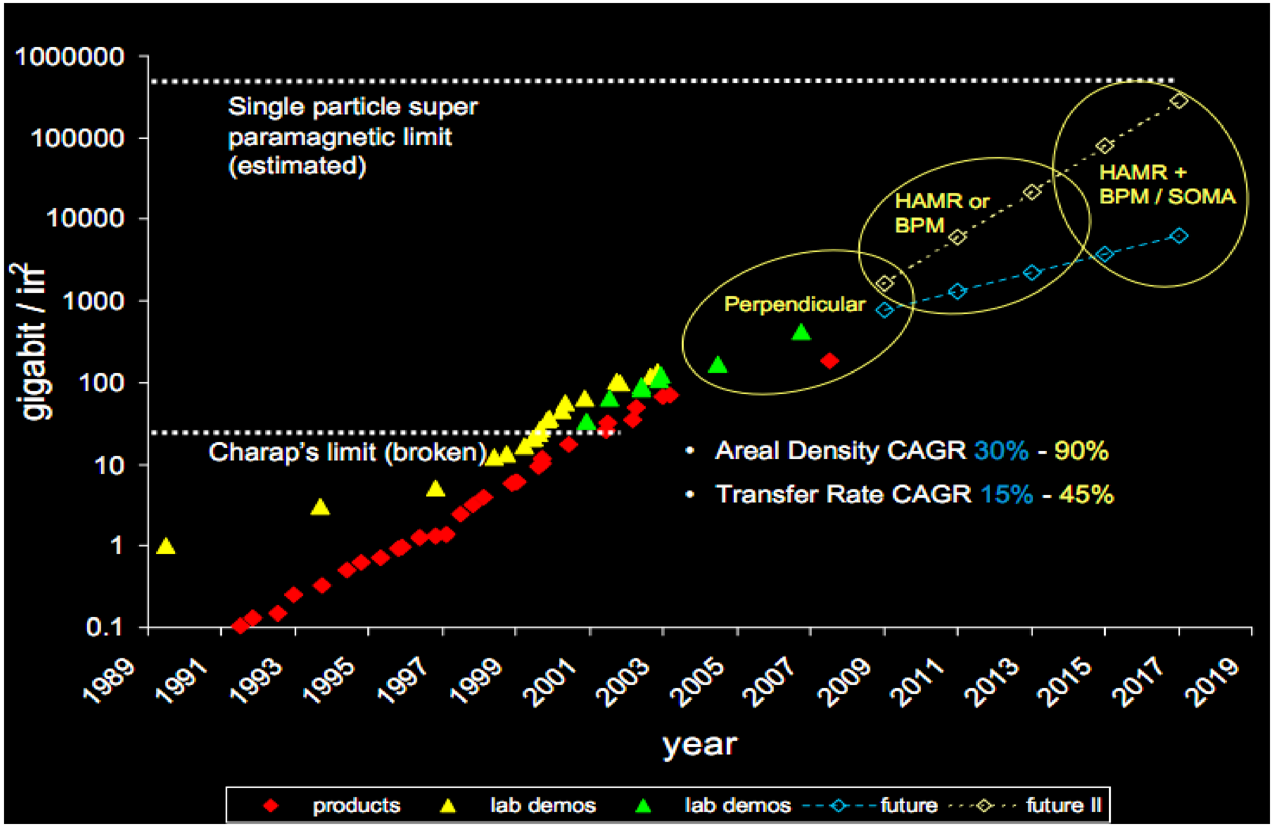


Source: Andy Bechtolsheim

- But what about HDD & storage chasm?

HDD: Capacity

- Capacity growth continues unabated

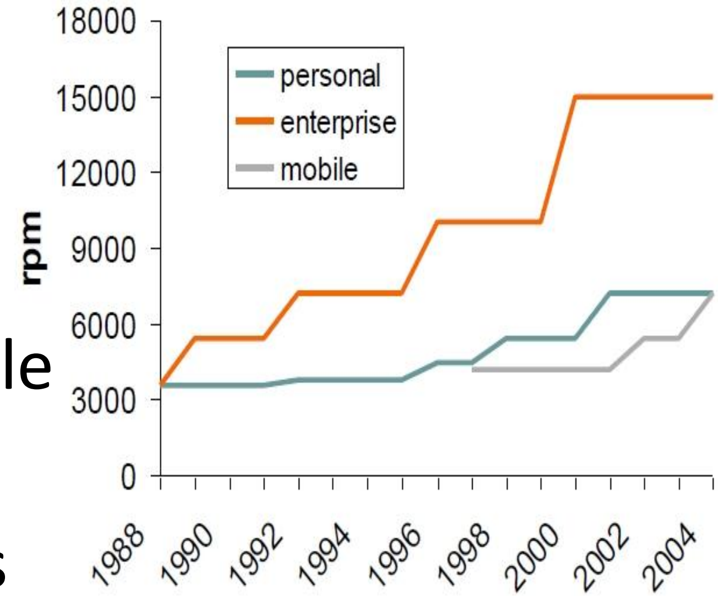


Source: Dave Anderson

- Capacity isn't the problem
  - What about bandwidth and IOPS?

HDD: Rotational Speed

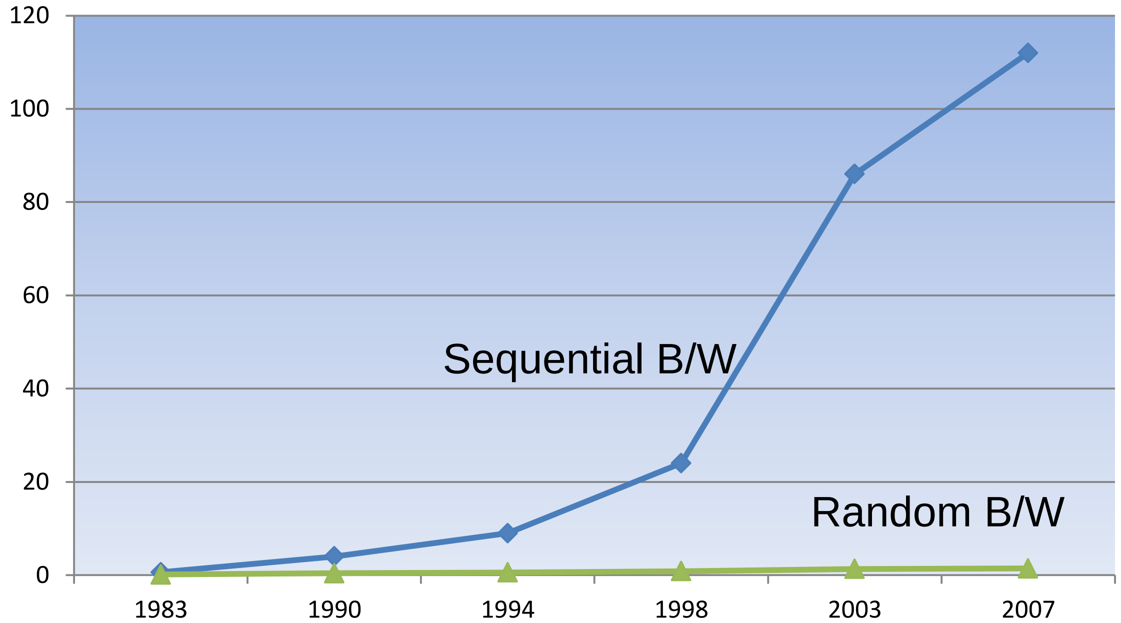
- RPM contributes negatively to:
  - rotational vibration
  - Non-Repeating Run Out (NRRO)
  - Power cubically related to RPM
- >15k RPM not economically viable
  - no improvement in sight
- RPM not improving & seek times only improving very slowly
- IOPS improvements looking forward remain slow



product information for Seagate and Control Data disc drives since 1988, mobile includes Toshiba drives since 1997

Source: Dave Anderson

Disk Becomes Tape



Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

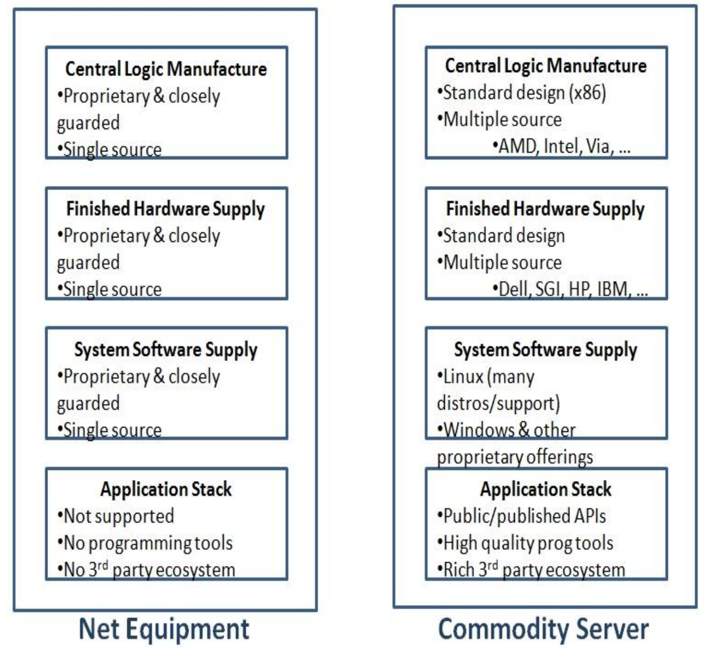
Jim Gray
Microsoft
December 2006

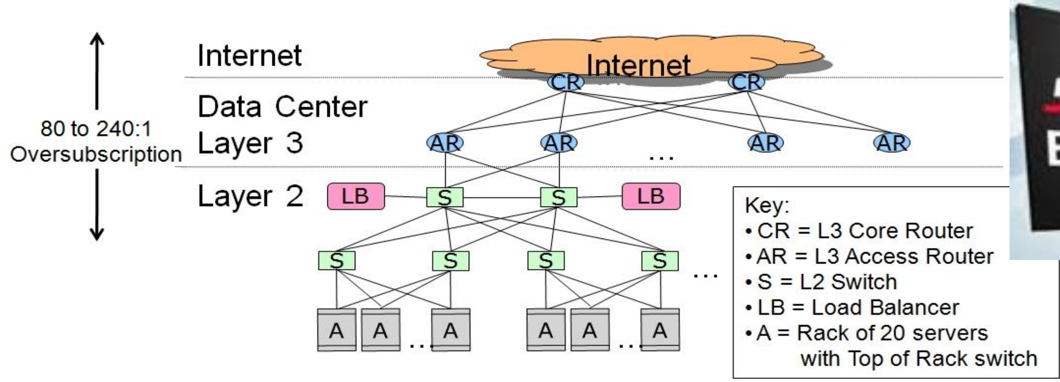
- Disk random access B/W growth ~10% of sequential B/W
- Random read 3TB disk: 31 days @ 140 IOPS (8kb)
  - 8.3 hours sequentially
- Disk only practical for sequential workloads
- Random workloads 100% to NAND flash & follow-on

Source: Dave Patterson with James Hamilton updates

Sea Change in Networking

- Current networks over-subscribed
  - Forces workload placement restrictions
  - Goal: all points in datacenter equidistant
- Mainframe model goes commodity
  - Competition at each layer over vertical integ.
- Get onto networking on Moores Law path
  - ASIC port count growth at near constant cost
  - Competition: Broadcom, Marvell, Fulcrum,…

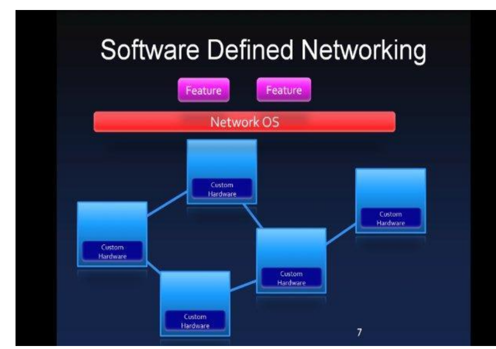




amazon webservices™

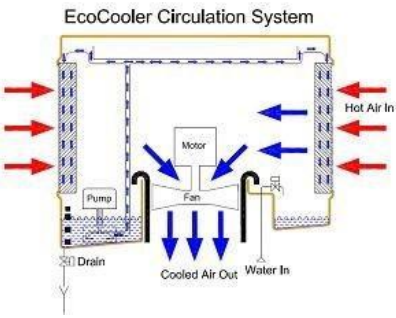
Networking Looking Forward

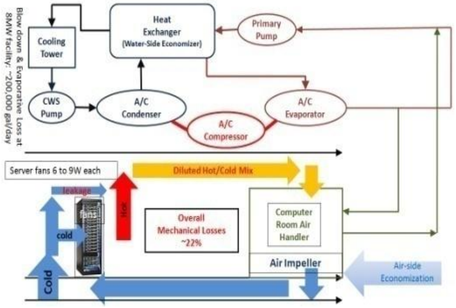
- Move to commodity routing:
  - Much less expensive & lower power
  - More redundancy & bandwidth
  - Get on Moore's law Path (ASIC port count growth)
- Centralized control plane
  - OpenFlow/Software Defined Networking
- Client side:
  - Virtualized NIC: Avoid hypervisor tax
  - ROCEE & iWarp: Avoid O/S transition
  - Cut-through routing: Avoid store and forward delay
  - B/W increases continue: 10GigE commodity

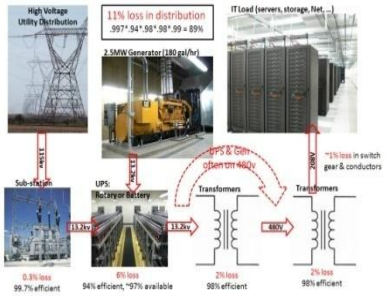


Agenda

- Cloud & Accelerating Pace of Innovation
- Technology Changes
  - Memory wall & Storage Chasm
  - Disk is Tape
  - Sea Change in Networking
- Data & Storage Trends
  - Map Reduce & NoSQL
  - Migration to Cloud

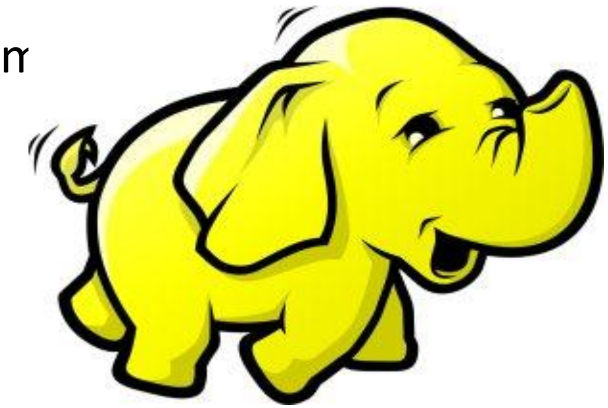






MapReduce

- Reaction to “RDBMs don’t scale” & admin costs
  - System community solution to big data problem
- MapReduce success fueled by:
  - Exploding data sizes
  - Scales (4,000 node single cluster at Yahoo)
  - Declining cost of computing
  - Sequential access pattern coupled with brute force
- MapReduce great for:
  - Extract, Transform and Load
  - Dirty data, weak schema, & access patterns not well suited to indexes
  - Executing arbitrary or complex functions over all data
- MR re-implementing indexes, materialized views, hash join, pipelined operators, …



NoSQL Movement

Everybody knows that relational databases don't scale because they use joins and write to disk…

- Another reaction "RDBMS don't scale" & admin complexity
- Unpredictable RDBMS response times dangerous at scale
- Relax a subset of ACID to achieve scale:
  - Eventually consistent
  - Non-durable on commit
  - Don't fully isolate conflicting txns
  - Don't support multi-item atomic update
  - Light to no schema enforcement
  - No complex query, no joins, no aggregates, no RI, no…
- Simple programming model and administration
  - Eventual consistency often not "really" understood
  - App code required for complex queries
- Good for some workloads at scale:
  - Cassandra, MongoDB, CouchDB, SimpleDB, …

Client Storage Migration to Cloud

- Client disk rapidly replaced by local semiconductor caches
  - Flash becoming primary client storage media
  - Higher performance, Lower power, smaller form factor, greater shock resistance, scale below HDD cost floor, greater humidity range, wider temp range, lower service costs, …
- Same trend in embedded devices
  - Well connected with cloud-hosted storage
- Clients storage drives cloud storage
  - Value added services, many data copies, shared access, indexed, classified, analyzed, monetized, reported, …

Open Source & Cloud Influence

- Open Source DBs inexpensive
  - Encourages sharding rather than scale-up
- Cloud removes DB admin cost
  - Further fueling increased used of sharding
- DBs Ideal workload for the cloud:
  - DB admin is hard but at scale it can be automated
  - Admin scales up well & down poorly
- Massive amount of data in cloud
  - Bring the query to data rather than data to query

Summary

- Cloud scale driving quickening pace of innovation
- Plunging costs driving bigger data sets and more complex analysis
  - Data moving up memory hierarchy
  - Data moving up the storage hierarchy
- Networking costs & capabilities changing fundamentally
- Most difficult scaling problems always data related
- Exciting time to be in the storage world

Questions?

- **Slides will be posted to:**
  - http://mvdirona.com/jrh/work
- **Perspectives Blog:**
  - http://perspectives.mvdirona.com/
- **Email:**
  - James@amazon.com