Cloud Computing Driving Infrastructure Innovation

Intel DCSG Distinguished Speaker Series

James Hamilton, 2012/9/25
VP & Distinguished Engineer, Amazon Web Services
email: James@amazon.com
web: mvdirona.com/jrh/work
blog: perspectives.mvdirona.com
Agenda

- Cloud Computing Scaling & Costs
- Cloud Computing Economics
- Infrastructure Innovation
  - Power Distribution
  - Mechanical Systems
  - Data Center Building Design
  - Networking
  - Storage
- Cloud Computing Drives H/W Feature Use
Pace of Innovation

• Datacenter pace of innovation increasing
  – More innovation in last 5 years than previous 15
  – Driven by cloud service providers and very high-scale internet applications like search
  – Cost of infrastructure dominates service cost
  – Not just a cost center

• High focus on infrastructure innovation
  – Driving down cost
  – Increasing aggregate reliability
  – Reducing resource consumption footprint
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.
The Cloud Scales: Amazon S3 Growth

Peak Requests: 500,000+ per second

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Number of S3 Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q4 2006</td>
<td>2.9 Billion</td>
</tr>
<tr>
<td>Q4 2007</td>
<td>14 Billion</td>
</tr>
<tr>
<td>Q4 2008</td>
<td>40 Billion</td>
</tr>
<tr>
<td>Q4 2009</td>
<td>102 Billion</td>
</tr>
<tr>
<td>Q4 2010</td>
<td>262 Billion</td>
</tr>
<tr>
<td>Q4 2011</td>
<td>762 Billion</td>
</tr>
<tr>
<td>Q4 2012</td>
<td>&gt;1 Trillion</td>
</tr>
</tbody>
</table>

http://perspectives.mvdirona.com
AWS Datacenters in 8 Regions

- **US GovCloud** (US ITAR Region -- Oregon)
- **US West x 2** (N. California and Oregon)
- **US East** (Northern Virginia)
- **Europe West** (Dublin)
- **Asia Pacific Region** (Singapore)
- **Asia Pacific Region** (Tokyo)

- **LATAM** (Sao Paola)

>10 datacenters In US East alone

- **8 AWS Regions and growing**
- **21 AWS Edge Locations for CloudFront (CDN) & Route 53 (DNS)**
Where Does the Money Go?

**Assumptions:**
- Facility: ~$88M for 8MW critical power
- Servers: 46,000 @ $1.45k each
- Commercial Power: ~$0.07/kWhr
- Power Usage Effectiveness: 1.45

**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)

Agenda

• Cloud Computing Scaling & Costs
• Cloud Computing Economics
• Infrastructure Innovation
  – Power Distribution
  – Mechanical Systems
  – Data Center Building Design
  – Networking
  – Storage
• Cloud Computing Drives H/W Feature Use
Utilization & Economics

• Server utilization problem
  – 30% utilization VERY good & 10% to 20% common
    • Expensive & not good for environment
  – Solution: pool number of heterogeneous services
    • Non-correlated peaks & law of large numbers
• Pay as you go & pay as you grow model
  – Don’t block the business
  – Don’t over buy
  – Transfers capital expense to variable expense
  – Apply capital for business investments rather than infrastructure
• Charge back models drive good application owner behavior
  – Cost encourages prioritization of work by application developers
  – High scale needed to make a market for low priority work
Data Center Efficiency

• Datacenter design efficiency
  – Average datacenter efficiency low with PUE over 2.0 (Source: EPA)
    • Many with PUE over 3.0
  – High-scale cloud services in 1.2 to 1.5 range
  – Lowers computing cost & better for environment

• Multiple datacenters
  – At scale multiple datacenters can be used
    • Close to customer
    • Cross datacenter data redundancy
    • Address international markets efficiently

• Avoid upfront datacenter cost with years to fully utilize
  – Scale supports pervasive automation investment
Scale Effects

• Custom service-optimized hardware
  – ODM sourced

• Purchasing power at volume

• Supply chain optimization
  – Shorter supply chain drives higher server utilization
    • Predicting next week easier than 4 to 6 months out
  – Less over buy & less capacity risk

• Networking transit costs strongly rewards volume

• Cloud services unblocks new business & growth
  – Remove dependence on precise capacity plan
Amazon Cycle of Innovation

• 15+ years of operational excellence
  – Managing secure, highly available, multi-datacenter infrastructure

• Experienced at low margin cycle of innovation:
  – Innovate
  – Listen to customers
  – Drive down costs & improve processes
  – Pass on value to customers

• 19 AWS price reductions so far
  – Expected to continue
Agenda

- Cloud Computing Scaling & Costs
- Cloud Computing Economics
- Infrastructure Innovation
  - Power Distribution
  - Mechanical Systems
  - Data Center Building Design
  - Networking
  - Storage
- Cloud Computing Drives H/W Feature Use
Power Distribution

High Voltage Utility Distribution

Sub-station

115kV

13.2kv

115kV

0.3% loss
99.7% efficient

Generators

13.2kv

UPS:

Rotary or Battery

13.2kv

6% loss
94% efficient, ~97% available

Transformers

13.2kv

2% loss
98% efficient

IT Load (servers, storage, Net, ...)

~11% lost in distribution

.997*.94*.98*.98*.99 = 89%

Transformers

480V

2% loss
98% efficient

~1% loss in switch gear & conductors

UP & Gen often on 480V

~11% lost in distribution

.997*.94*.98*.98*.99 = 89%

Generators

2% loss
98% efficient

~1% loss in switch gear & conductors

Note: Two more levels of power conversion at server

http://perspectives.mvdirona.com
Mechanical Systems

- **Cooling Tower**
- **CWS Pump**
- **A/C Condenser**
- **Heat Exchanger (Water-Side Economizer)**
- **Primary Pump**
- **A/C Evaporator**
- **A/C Compressor**
- Server fans 6 to 9W each
- **Diluted Hot/Cold Mix**
- **Overall Mechanical Losses ~22%**
- **Computer Room Air Handler**
- **Air Impeller**

Blow-down & Evaporative Loss at 8MW facility: ~200,000 gal/day
Innovative Building Designs

• Evaporative cooling only
  – Right: High pressure misting
  – Below: Wet media cooler

• Ductless full building cooling

Facebook Prineville above & below

EcoCooling
Modular and Pre-fab DC Designs

- Fast & economic deployments
- Sub-1.15 PUE designs
- Air-side economized
  - No mechanical cooling
- ISO standard shipping containers offered by Dell, HP, SGI, IBM, ...

Microsoft ITPAC

Amazon Perdix

http://perspectives.mvdirona.com
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,...

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
HDD: Capacity

• Capacity growth continues unabated

• Capacity isn’t the problem
  – What about bandwidth and IOPS?

Source: Dave Anderson/Seagate
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
• Even sequential BW growth insufficient

Source: Dave Anderson
• Disk sequential BW growth slow
• Disk random access BW growth roughly 10% of sequential
• Storage Chasm widening
  – BW a long term problem &IOPS growth very slow

Source: Dave Patterson with James Hamilton updates
2012/9/25
http://perspectives.mvdirona.com
Disk Becomes Tape

• Random access disk latency increasingly impractical
• Random read 4TB disk:
  – 41.3 days @ 140 IOPS with 8kb page
  – Disk increasingly impractical for random workloads
• Sequential read is over 11 hours
• Trending below tape price point
  – Tape only cost effective at very high scale
  – Disk wins at top and scales down better

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

Jim Gray
Microsoft
December 2006
Flash Becomes Disk

- All random workloads to Flash
- Flash 4 to 6x more expensive capacity
- Log structured block store
  - Compress
  - De-dupe
  - Sparse provision
- Approaches HDD capacity price point

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

Jim Gray
Microsoft
December 2006
Client Storage Migration

• Client device disk replaced by semiconductor caches
  – Much higher performance, Lower power dissipation, smaller form factor, greater shock resistance, scale down below HDD cost floor, greater humidity range, wider temp range, lower service costs, ...
  – Flash is primary client storage media

• Clients storage drives cloud storage
  – Value added services, many data copies, shared access, indexed, classified, analyzed, monetized, reported, ...
  – Overall client storage continuing to expand rapidly but primarily off device in the cloud
Enterprise to Cloud

• Cloud computing 5x to 10x improved price point
  – Low margin, high volume business
  – Yet still profitable, sustainable, & supporting re-investment
  – Incompatible with on-premise enterprise S/W & H/W profit margins
• Expect many cloud winners rather than single provider
• Direct component supplier relationship with major operators rather than via distribution channel
• Custom server & networking equipment
Agenda

• Cloud Computing Scaling & Costs
• Cloud Computing Economics
• Infrastructure Innovation
  – Power Distribution
  – Mechanical Systems
  – Data Center Building Design
  – Networking
  – Storage
• Cloud Computing Drives H/W Feature Use
Accelerating Compute & Storage Growth

• Rapidly declining cost of computing
  – Driven by technology improvements & cloud computing economies of scale

• Traditional transactional systems scale with business
  – Purchases, ad impressions, pages served, etc.
  – Computational trading & machine-to-machine transactions limited only by value of transaction & cost of infrastructure

• Warehousing & analytical systems scale inversely with cost
  – Cheaper storage allows more data to be analyzed
  – Lower compute costs allows deeper analysis
Questions?

• Perspectives Blog:
  – http://perspectives.mvdirona.com/

• Email:
  – James@amazon.com