



Why Scale Matters and how the Cloud Really is Different

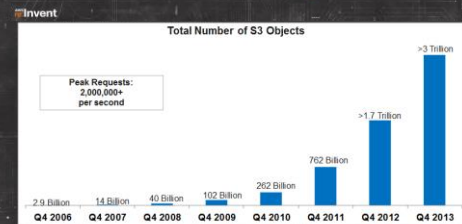
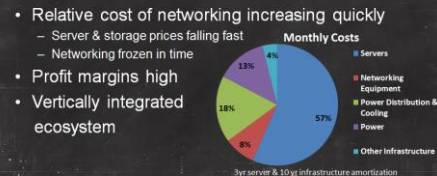
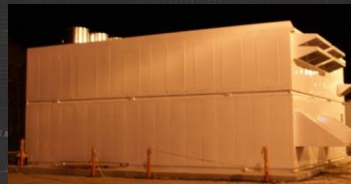
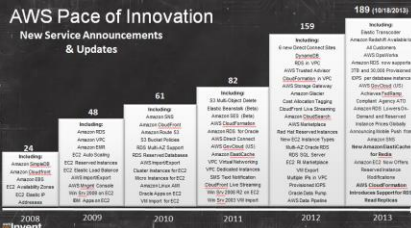
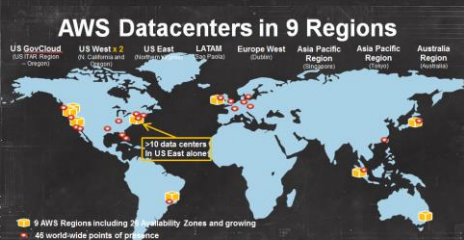
James Hamilton, AWS VP & Distinguished Engineer

SPOT205: November 23, 2013



Agenda

- Redefining Scale at AWS
- AWS Designed Hardware & Infrastructure
 - AWS Specific Power Distribution and Optimizations
 - Custom Network Hardware & Protocol Stack
 - Workload targeted Server Designs
 - AWS Custom NIC & other H/W Components



Perspective on Scaling



Every day, AWS adds enough new server capacity to support all of Amazon's global infrastructure when it was a \$7B annual revenue enterprise

AWS Datacenters in 9 Regions

US GovCloud
(US ITAR Region
-- Oregon)

US West x 2
(N. California and
Oregon)

US East
(Northern Virginia)

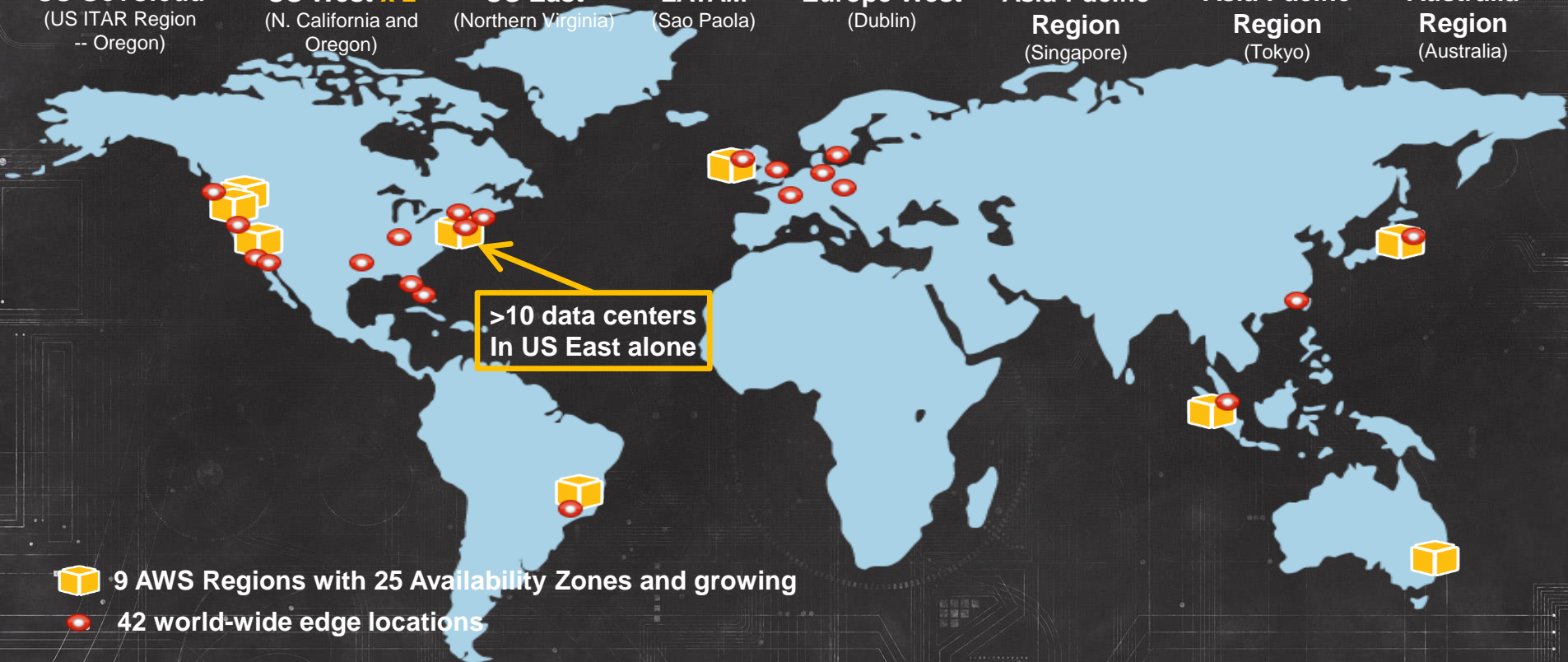
LATAM
(Sao Paulo)

Europe West
(Dublin)

**Asia Pacific
Region**
(Singapore)

**Asia Pacific
Region**
(Tokyo)

**Australia
Region**
(Australia)

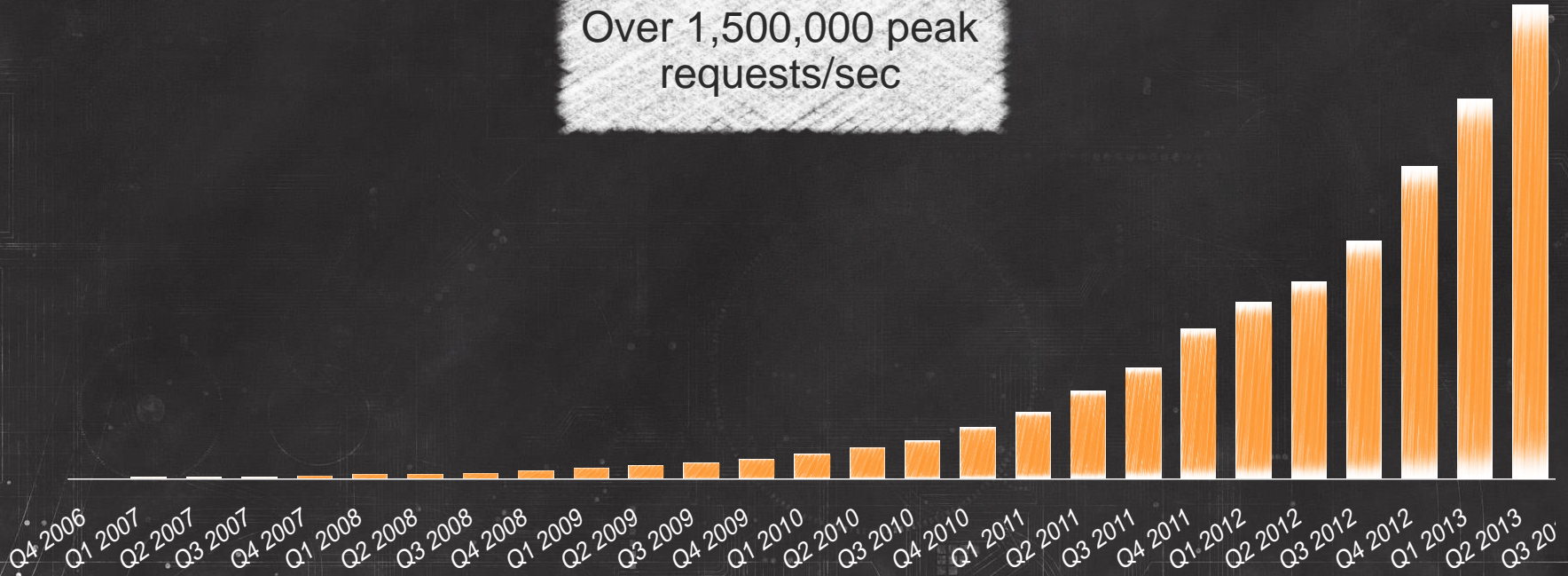


 9 AWS Regions with 25 Availability Zones and growing

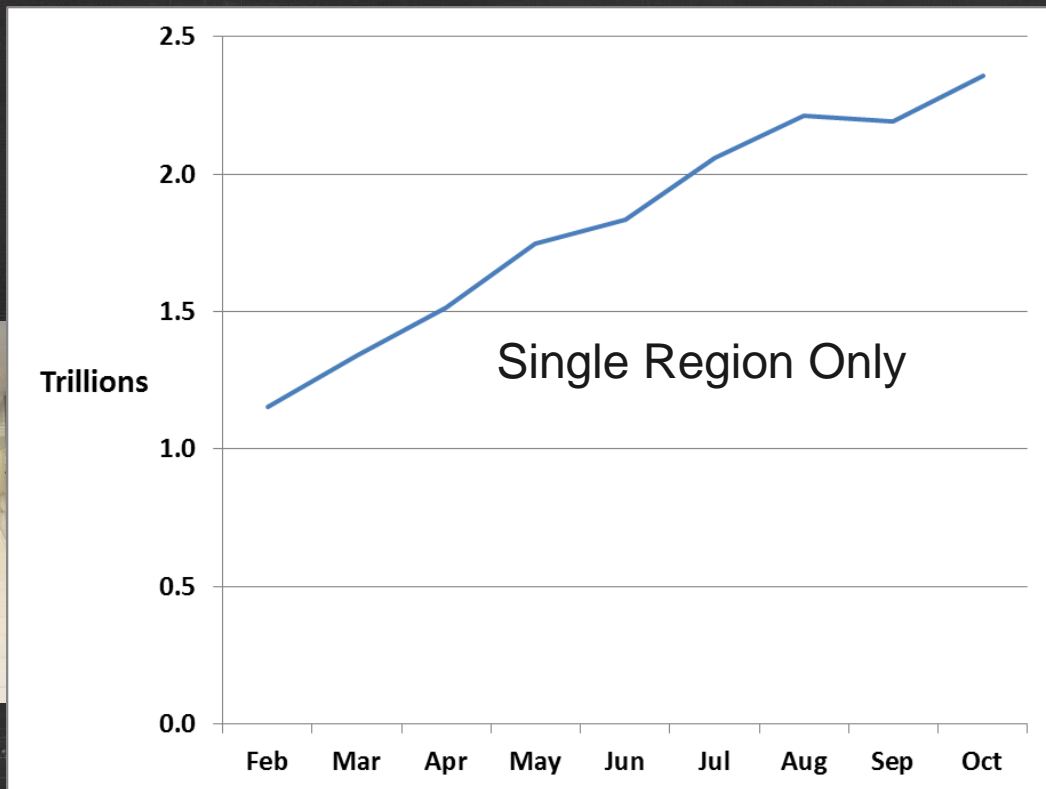
 42 world-wide edge locations

Amazon S3: Trillions of Total Objects

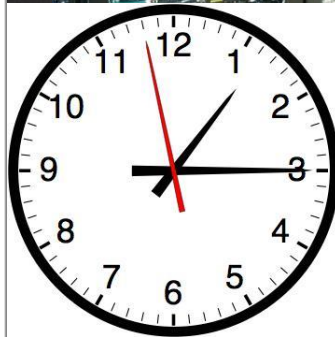
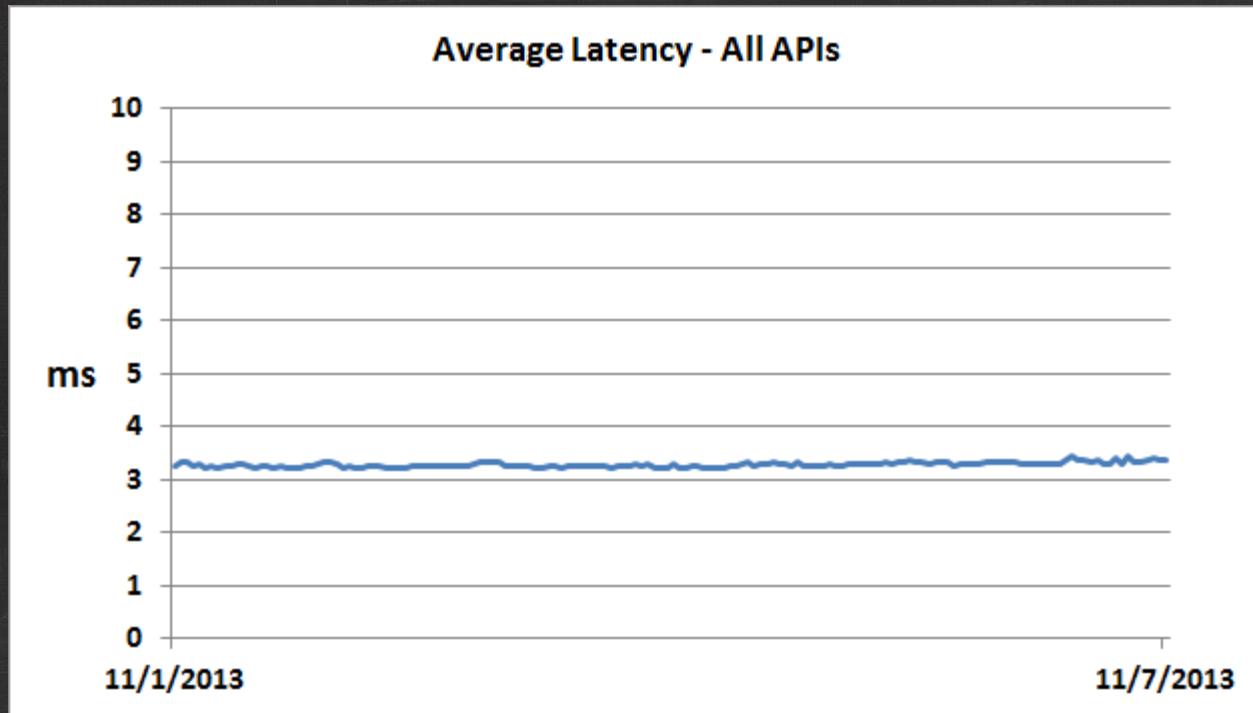
Over 1,500,000 peak
requests/sec



DynamoDB Requests Served/Month



DynamoDB: Consistent Performance at Scale



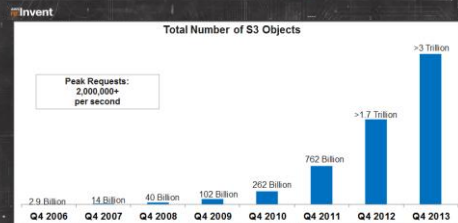
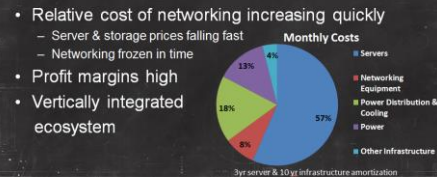
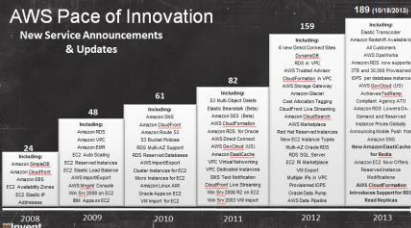
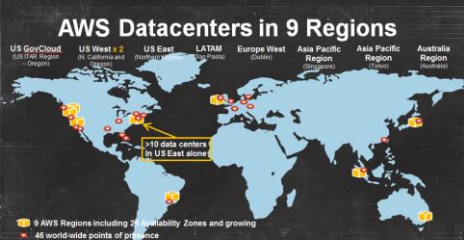
“AWS is the overwhelming market share leader, with **more than five times the compute capacity** in use than the aggregate total of the other fourteen providers.”

Gartner



Agenda

- Redefining Scale at AWS
- AWS Designed Hardware & Infrastructure
 - AWS Specific Power Distribution and Optimizations
 - Custom Network Hardware & Protocol Stack
 - Workload targeted Server Designs
 - AWS Custom NIC & other H/W Components



Pace of Innovation

- Infrastructure pace of innovation increasing
 - Driven by cloud service providers and high-scale internet applications such as search
 - Cost of datacenter & H/W infrastructure dominates
 - Infrastructure not just a cost center
- High focus on infrastructure innovation
 - Driving down cost
 - Increasing aggregate reliability
 - Reducing resource consumption footprint



facebook



Microsoft®



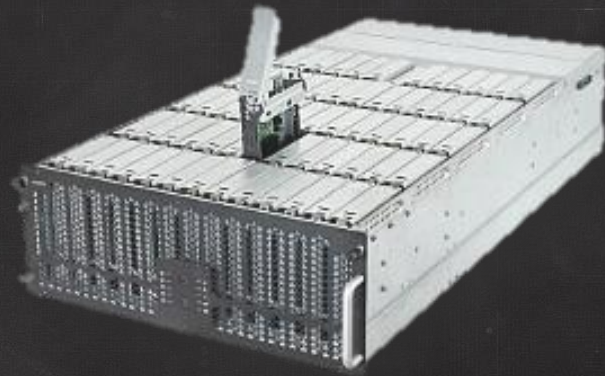
AWS Custom Server Designs

- OEM server ecosystem:
 - Optimized for 10s to 100s of thousands of customers
 - Vast, expensive, world-wide distribution network
 - Rapidly aging inventory throughout network
 - Broadly applicable servers able to run a wide variety of workloads
 - e.g. 64 PCI lanes & 24 memory slots
 - Power supply & voltage regulators tuned to worst case usage
- Cloud server ecosystem:
 - Optimized only for AWS deployment
 - Highly specialized servers optimized for a specific workload
 - Large scale deployments allow hardware specialization
 - Move hot s/w kernels to hardware implementations
 - Datacenters, servers, networking, storage to designed to integrated specification



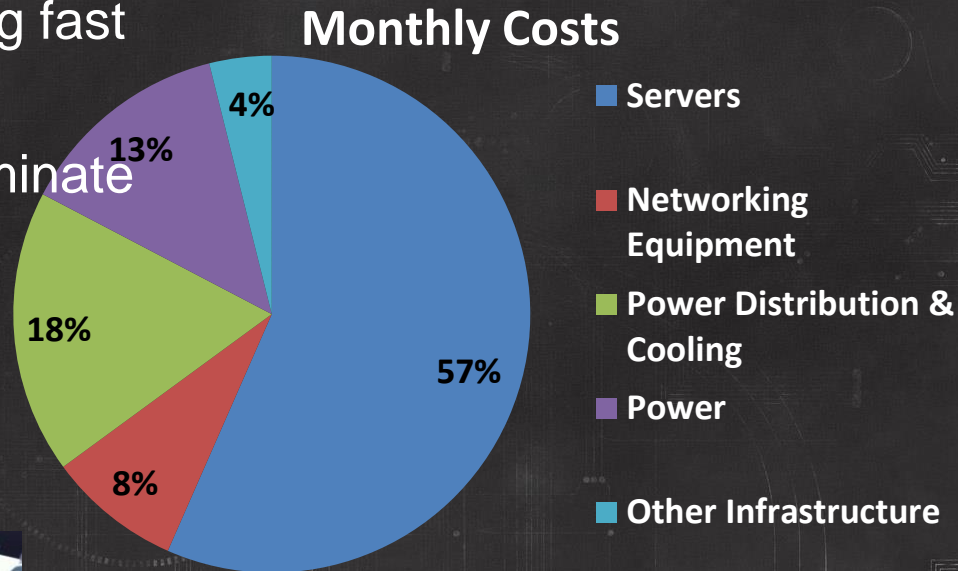
AWS Custom Storage Designs

- Commercial high-density storage:
 - Quanta M4600H 4U Disk Enclosure
 - Impressive best in class general purpose design
 - We use custom design with still higher density
- OEM storage & servers must target vast workload diversity
- High scale supports AWS-specific optimizations
 - Less power
 - Less space
 - More efficient
 - Less expensive



Networking Equipment

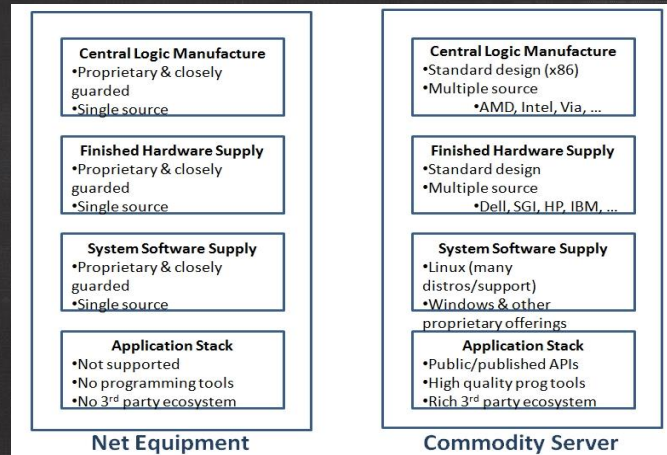
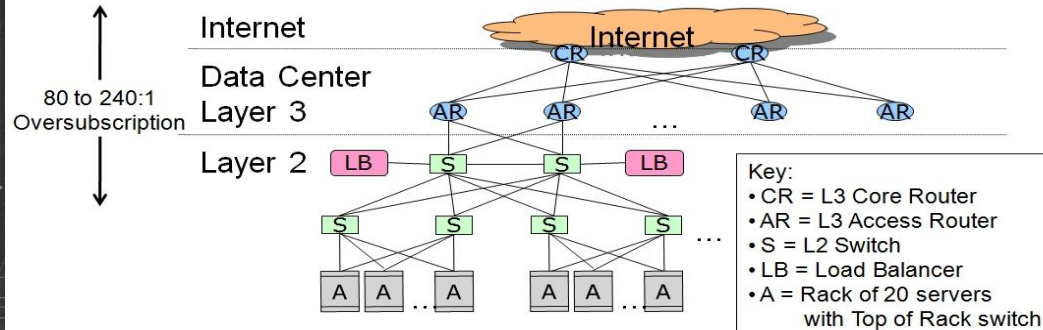
- Relative cost of networking increasing quickly
 - Server & storage prices falling fast
 - Networking frozen in time
 - Network costs on path to dominate
- Profit margins high
- Vertically integrated ecosystem



3yr server & 10 yr infrastructure amortization

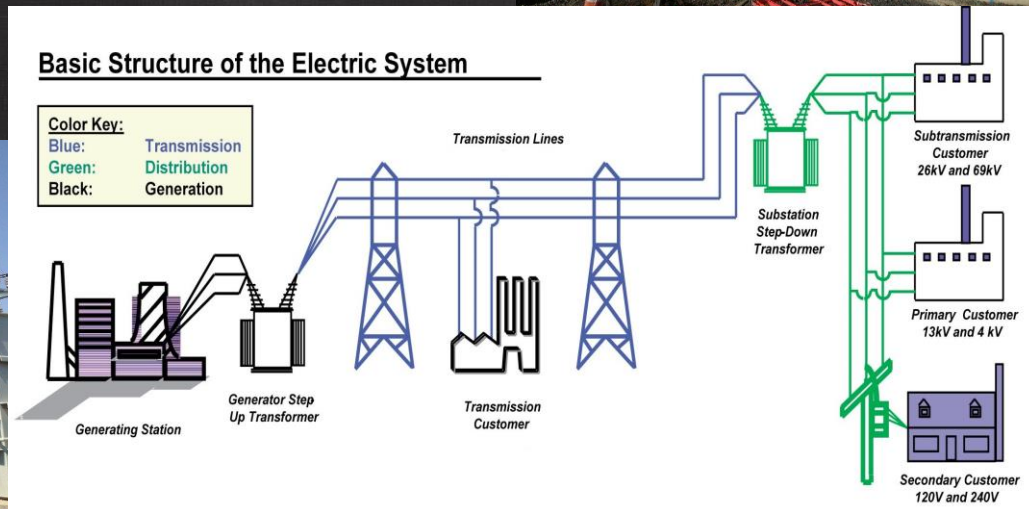
Get the Network Out of the Way

- Mainframe cost model
 - Amazon custom routers & protocol stacks
- Networks typically over-subscribed
 - Forces workload placement restrictions
 - Goal: all points in datacenter equidistant



Power Infrastructure

- Negotiated power purchasing agreements
- AWS custom high-voltage sub-stations in some regions
 - lower power cost
 - build more quickly



Super Bowl Power Outage

34 minute outage that very nearly changed the 2013 game

*“A piece of equipment that was designed to monitor electrical load **sensed an abnormality** in the system. The equipment **operated as designed** and opened a breaker that partially cut power to the Superdome in order **to isolate the issue**. Backup generators kicked in immediately as designed.”*

Lights without immediate backup power

- Restarting gas discharge lights takes 15+ min

Highly likely backup power wouldn't have helped

- Switchgear lockout

- We design & deploy custom switch firmware



Carbon Neutral Power Choice

Most companies rarely build new datacenters so there are few new power procurement options

The entire multi-datacenter US-WEST (Oregon) is 100% carbon neutral

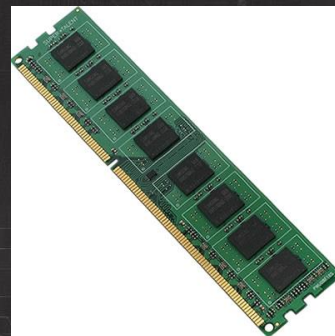
One of the largest AWS regions world-wide

- Fastest growing



Supply Chain & Procurement Optimization

- Purchasing power at volume utilizing our global demand
- Direct component purchasing
 - Disk drive, processors, memory, NICs, ...
 - Precise inventory control, better pricing, & optimized designs
- Supply chain optimization
 - Demand signals immediately feeding component buying, server & network builds, real estate purchasing, datacenter builds,...
 - Shorter cycle time drives much higher utilization
 - Predicting next week easier than 4 to 6 months out
 - Avoid overbuy & capacity risk

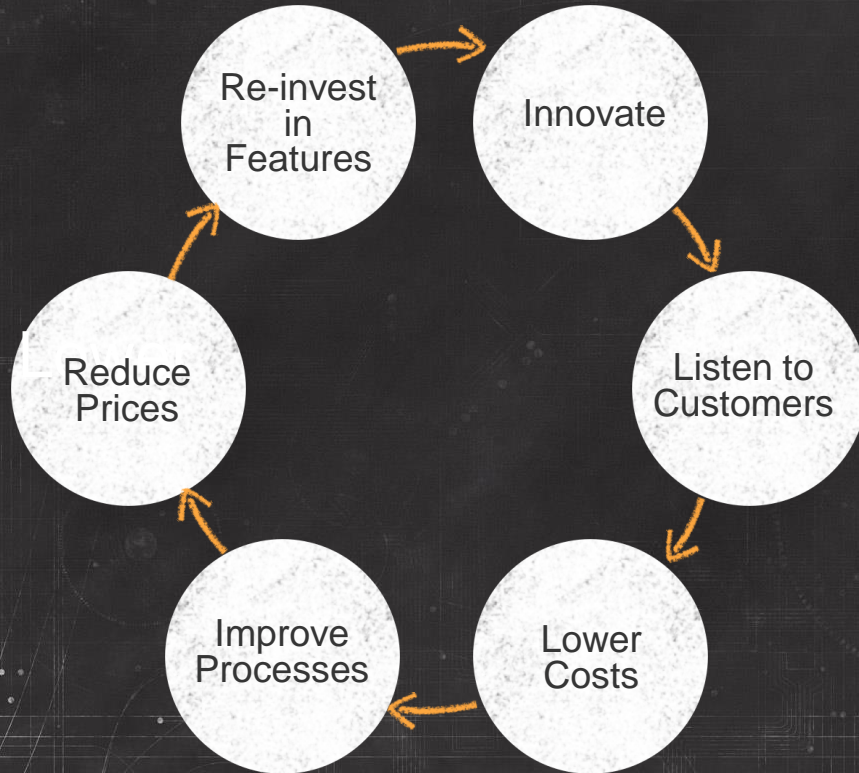


Utilization & Economics

- **Server utilization problem**
 - On premise 30% utilization VERY good & 10% to 20% more common
 - Expensive & not good for environment
 - Solution: pool number of heterogeneous services
 - Single reserve capacity pool far more efficient
 - 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 spot market for low priority work



Amazon Cycle of Innovation



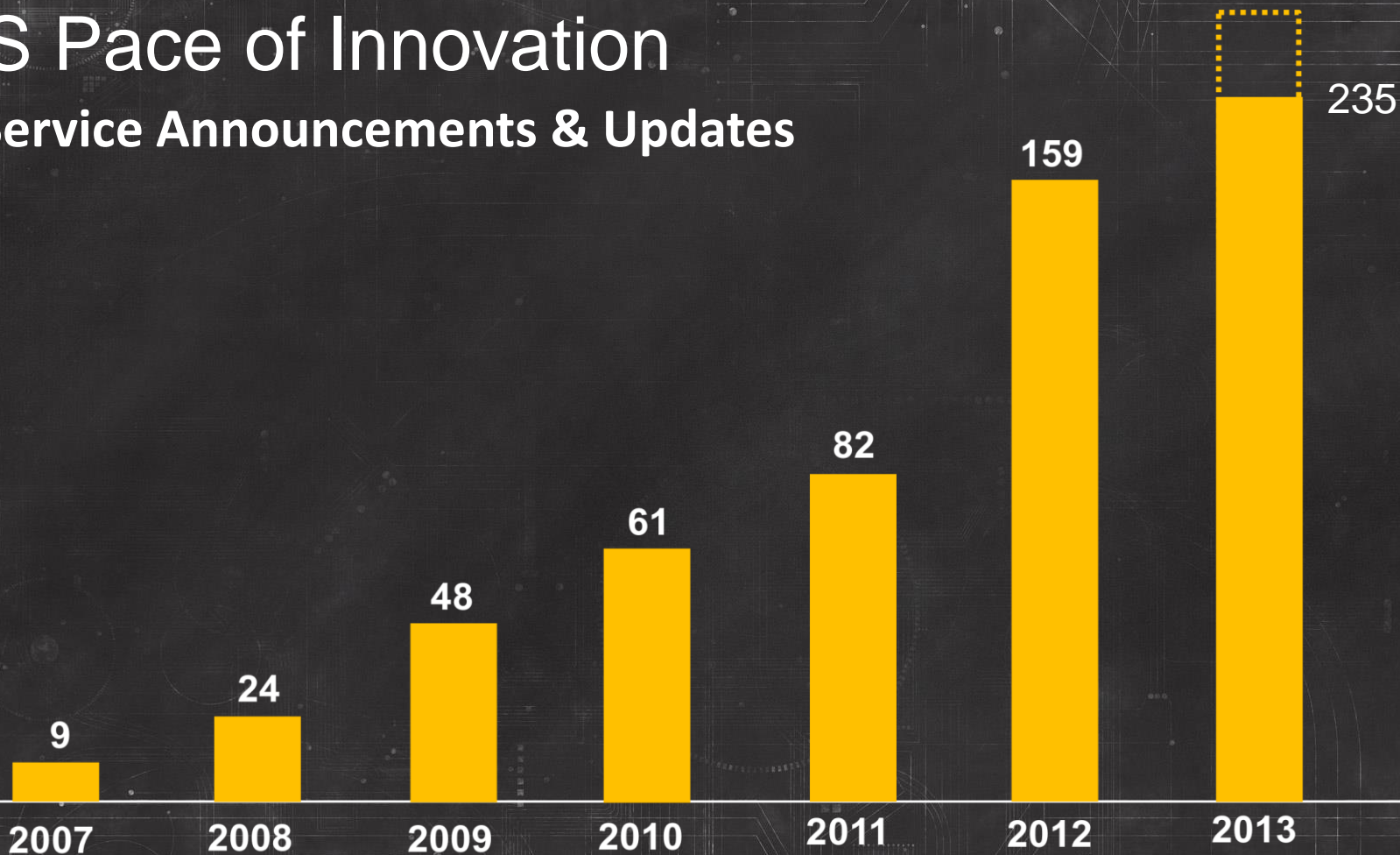
18+ years of operational excellence

38 AWS price reductions since 2006

700k trusted advisor recommendations saving customers \$140m

AWS Pace of Innovation

New Service Announcements & Updates



New Research: Customers Improve Availability by Migrating Apps to AWS

32% reduction in total application downtime

2013 AWS Customer Survey



NUCLEUS
RESEARCH

November 2013

Document N168

RESEARCH NOTE BENCHMARKING AVAILABILITY AND RELIABILITY IN THE CLOUD: AMAZON WEB SERVICES

THE BOTTOM LINE

Companies are increasingly moving applications to the public cloud to take advantage of increased agility, lower initial and ongoing cost, and reduce capital expenditures. However, high-profile efforts to spread fear, uncertainty, and doubt about the reliability and availability of the cloud — particularly Amazon Web Services (AWS) — have given some decision makers pause. To better understand the real-world reliability of cloud and on-premise reliability and availability, Nucleus surveyed Amazon Web Services customers that had moved existing on-premise applications to the Amazon cloud and found that many companies were able to completely eliminate both planned and unplanned downtime; in fact, the number of customers indicating no planned downtime increased by 53 percent with AWS. Analysts found customers were able to reduce unplanned downtime days by 32 percent and reduced planned downtime by 29 percent.

In the past decade, many organizations have moved all or part of their application footprint to the cloud, to take advantage of the inherent economies of scale, lower CAPEX requirements, increased agility, and other ROI advantages (Nucleus Research *m108 – Cloud delivers 1.7 times more ROI*, September 2012). During this time, Amazon Web Services (AWS) has also grown its public cloud business since its launch in 2006, to serve more than 100,000 customers including large enterprise clients such as Pfizer, General Electric, Merck, and Unilever.

However, Nucleus has also seen resistance to the public cloud based on purported perceptions of low availability and reliability compared to traditional on-premise computing. Naturally, those supporting the on-premise argument have been quick to broadcast reports of highly visible service disruptions of leading cloud vendors such as Google, Salesforce.com, and AWS. However, while data on service disruptions for cloud providers is readily available, comparable hard data is rarely, if ever, available to support anti-cloud proponents' claims that on-premise infrastructures deliver more reliability and availability than public clouds.

This raises the natural question: How does the availability and reliability of infrastructure and applications deployed in public clouds compare to those deployed on premises? To

Nucleus Research
Inc. 100 State Street
Boston, MA 02109

NucleusResearch.com
Phone: +1 617.725.2000

Research Note: Benchmarking availability and reliability in the cloud: Amazon Web Services Nucleus Research, November 2013, Document N168

Hosting on-premise less expensive?


- Utilization fundamentally higher in cloud
 - Aggregating non-correlated workloads, scale, spot market,...
- Amazon specific H/W designs
 - ODM acquisition of custom servers & net gear
 - Direct purchasing of disk, memory, & CPU
 - AWS controlled hypervisor & net protocol layers
- Deep R&D: Many new data centers built each year
- Immense scale
 - Volume purchasing, highly automated, specialists in all areas
- Amazon margins tiny compared with enterprise margins



- AWS Economics driven by scale & singular focus

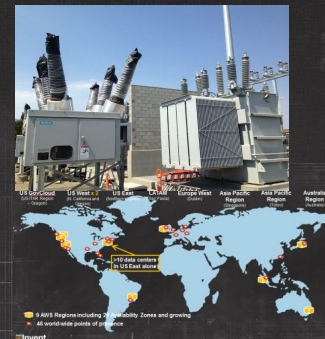
- Economies of scale
- Increased availability through multiple-datacenter deployment
- Steadily declining price

- Mega-scale advantages available to all customers regardless of size

- Datacenter presence near all customers world-wide
 - Multiple datacenters in each region for high availability
 - Deeper R&D investment & operational focus in datacenter, server, storage, & networking than any IT organization in the world
 - Buying power that rivals the biggest in the world
- 

- Cloud Model Fundamentally different from the last 30 years

- Even if rebranded as “cloud enabled”, “private cloud”, “cloud-like”, ...



AWS re:Invent

Please give us your feedback on this presentation

SPOT205

As a thank you, we will select prize winners daily for completed surveys!

