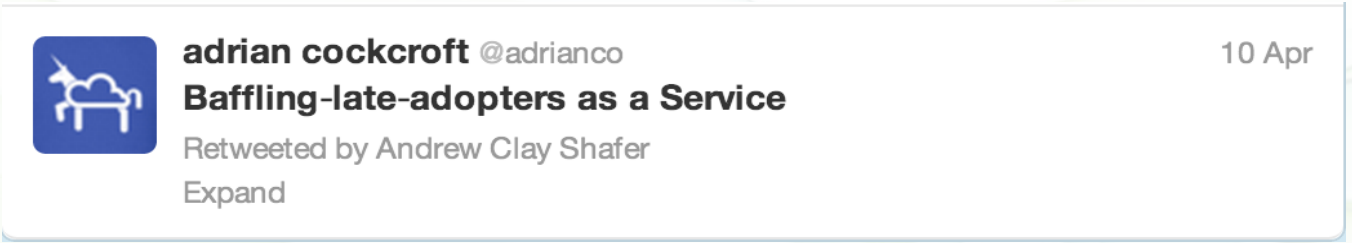
Disrupting the Storage Industry

Adrian Cockcroft @adrianco April 2014

Industry perspective
Cloud trends
Cloud native storage architecture


adrian cockcroft @adrianco
10 Apr
Baffling-late-adopters as a Service
Retweeted by Andrew Clay Shafer
Expand

Typical reactions to my Netflix talks…

What I learned from my time at Netflix

- Speed wins in the marketplace
- Remove friction from product development
- High trust, low process, no hand-offs between teams
- Freedom and responsibility culture
- Don’t do your own undifferentiated heavy lifting
- Use simple patterns automated by tooling
- Self service cloud makes impossible things instant


NEWS

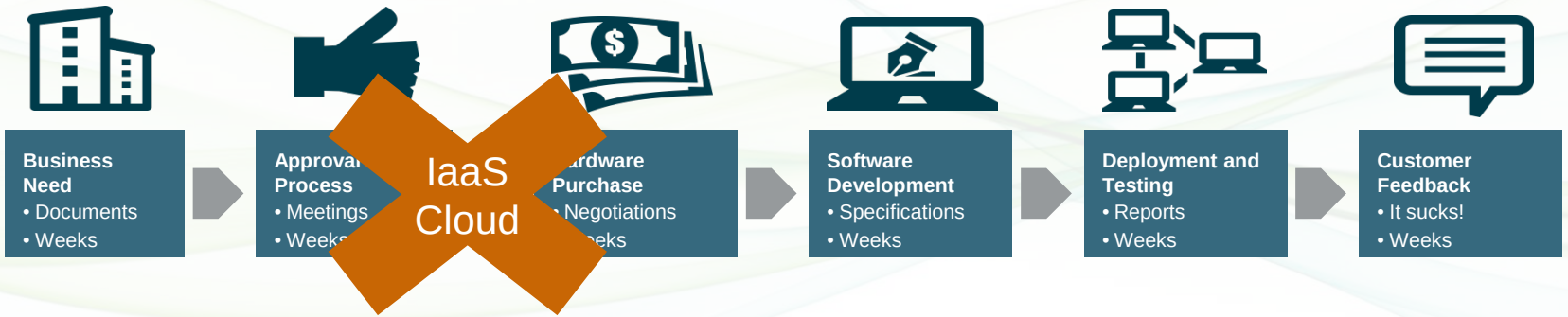
Demands on IT Increased 1000x

Compete or lose in the market!

How fast can you innovate?

Non-Cloud Product Development

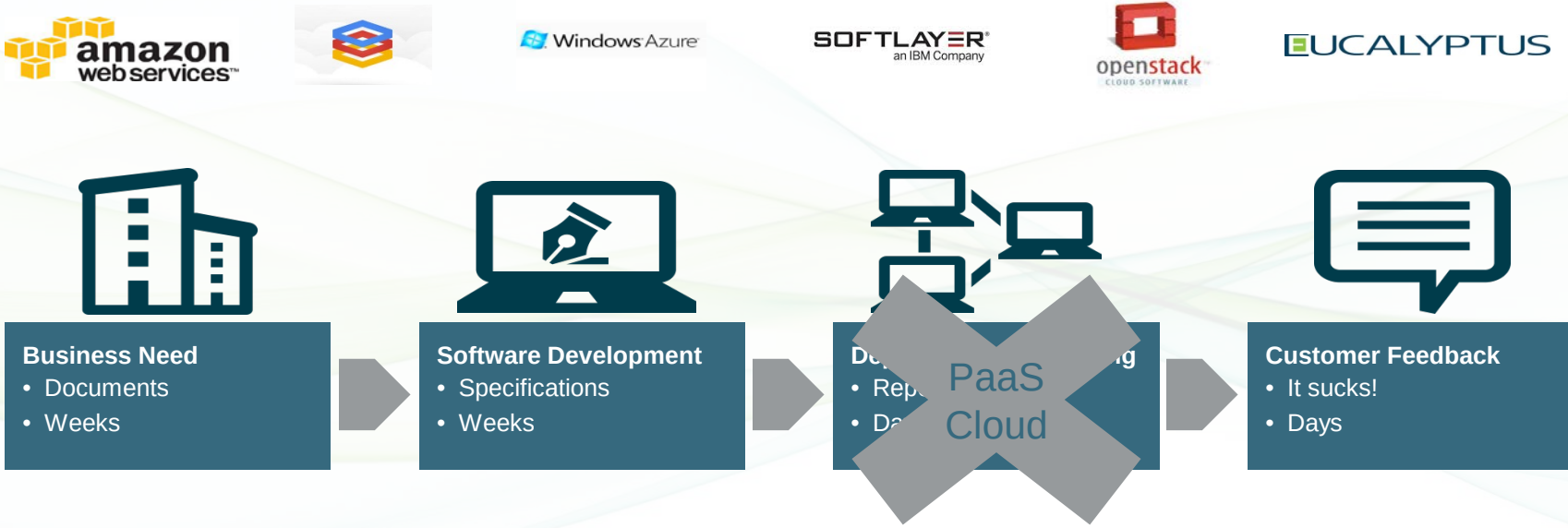
Months before you find out whether the product meets the need



- Hardware provisioning is undifferentiated heavy lifting – replace it with IaaS

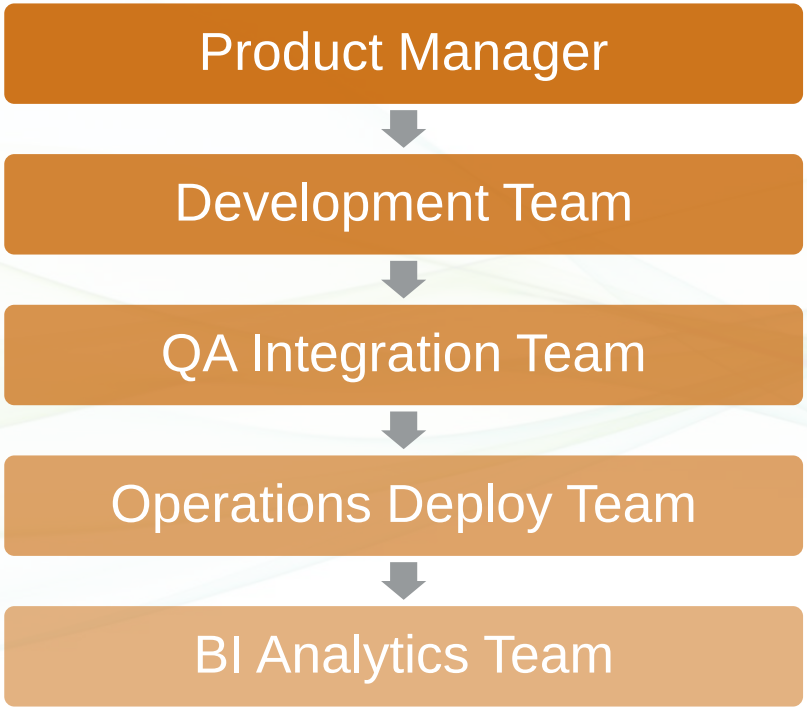
IaaS Based Product Development

Weeks before you find out whether the product meets the need

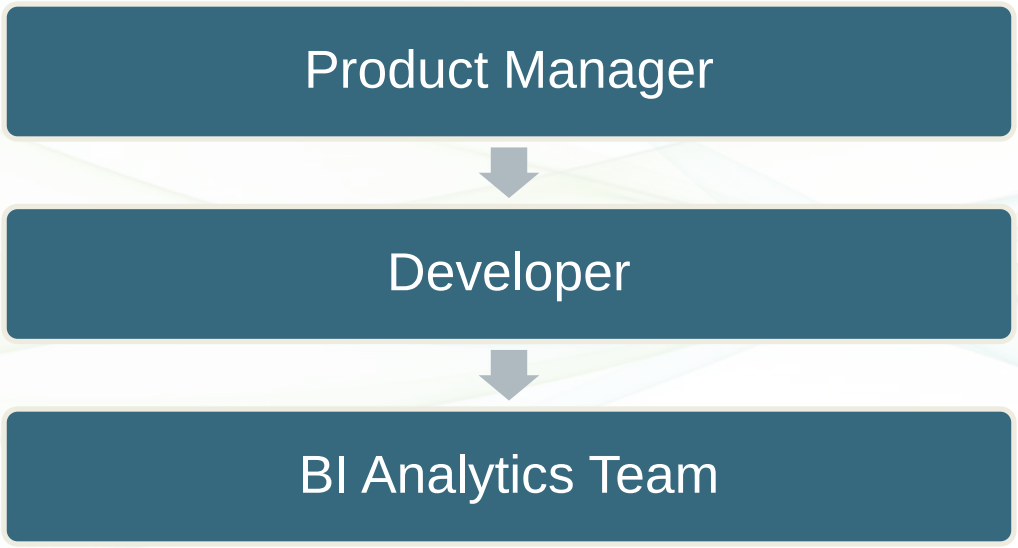


- Software provisioning is undifferentiated heavy lifting – replace it with PaaS

Process Hand-Off Steps for Product Development on IaaS



Process Hand-Off Steps for Feature Development on PaaS

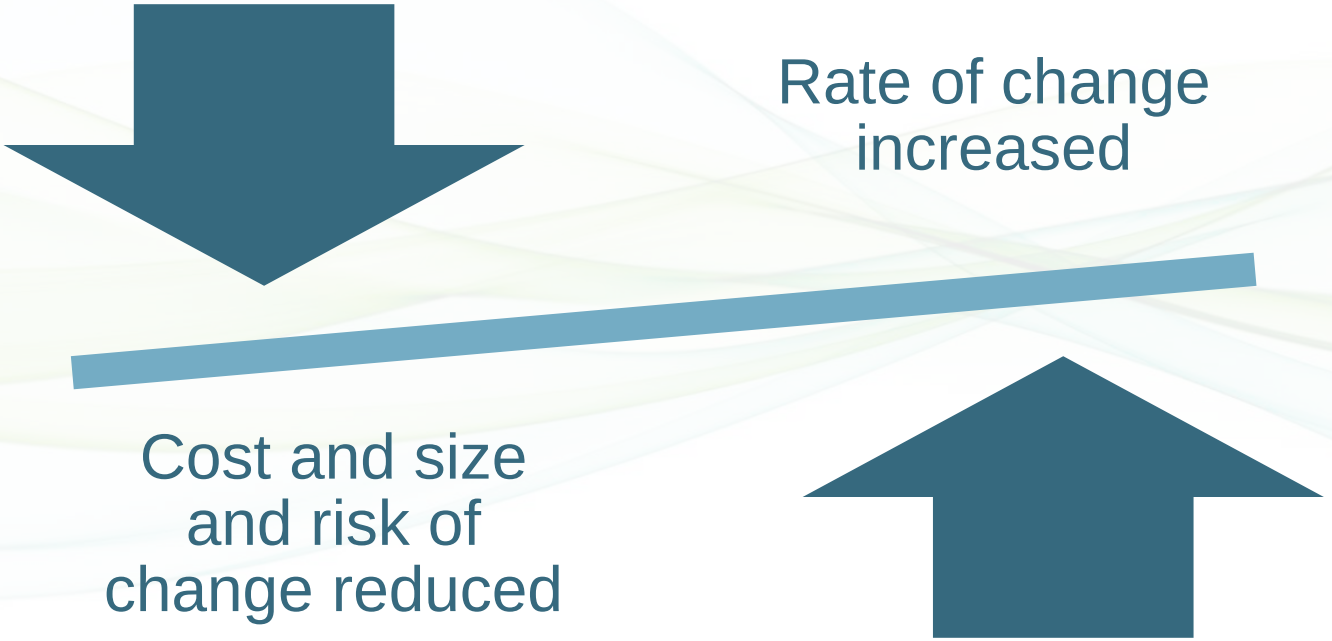


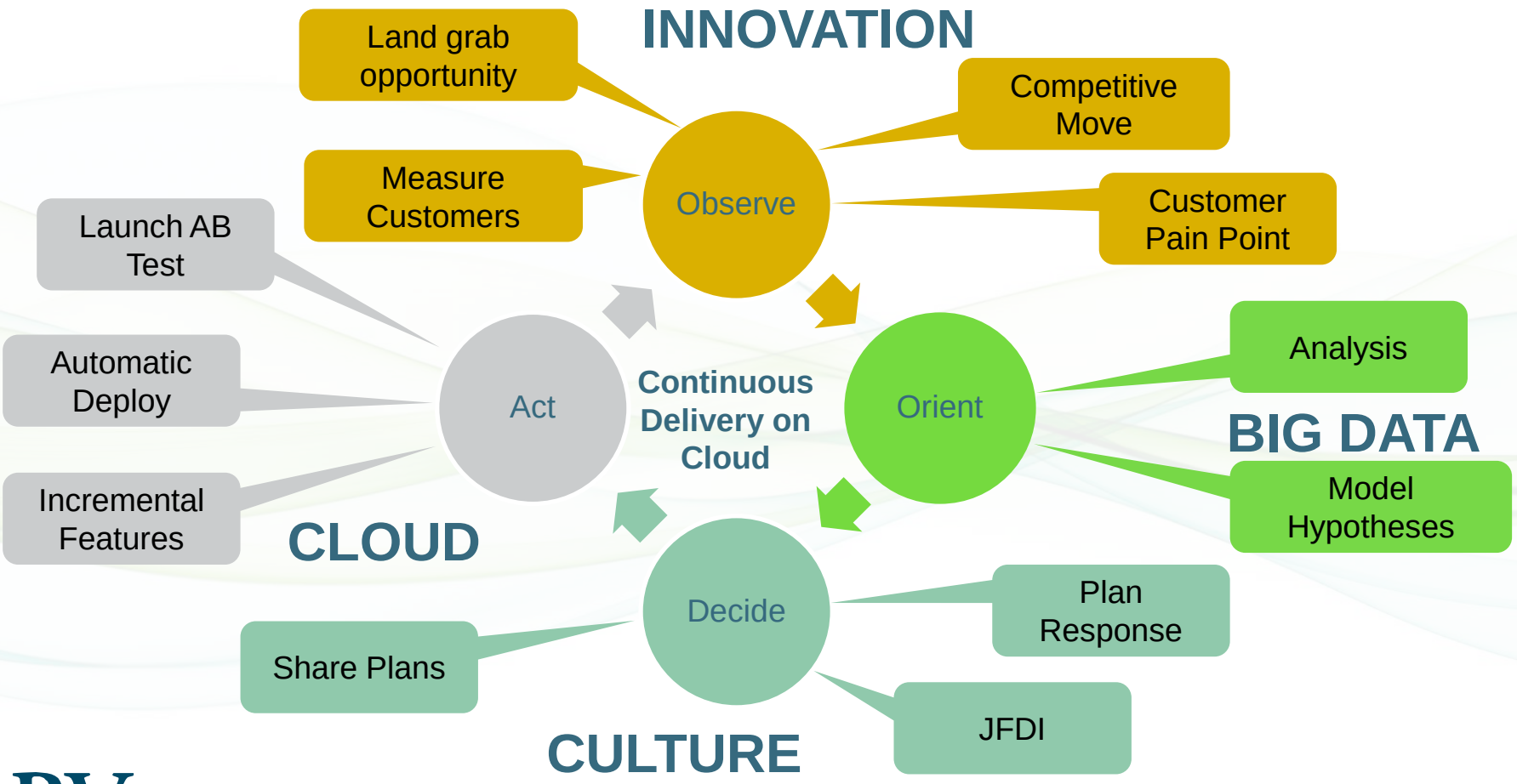
PaaS Based Product Feature Development

Days before you find out whether the feature meets the need

- Building your own business apps is undifferentiated heavy lifting – use SaaS

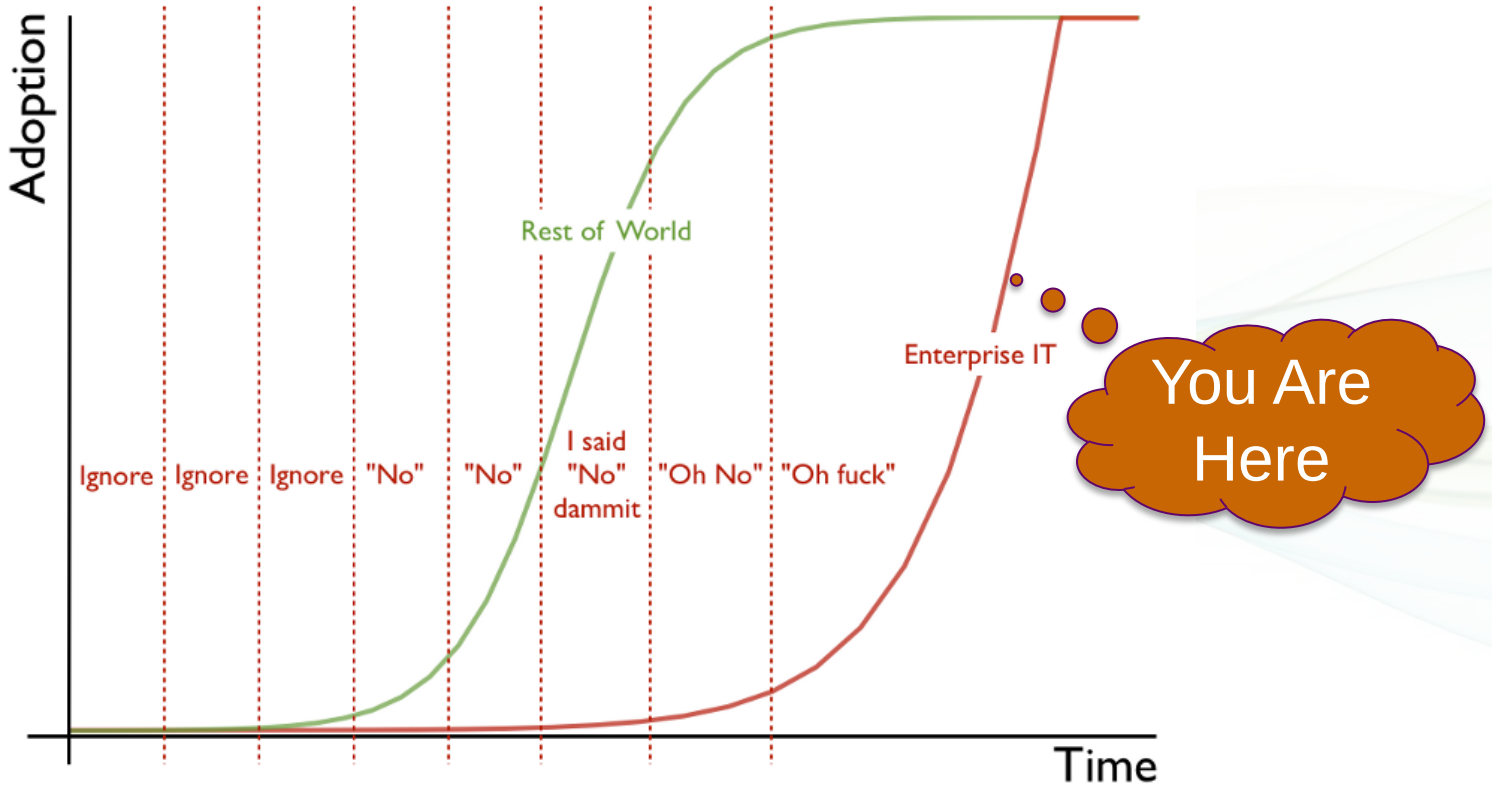
What Happened?




INNOVATION
Land grab opportunity
Competitive Move
Measure Customers
Observe
Customer Pain Point
Launch AB Test
Automatic Deploy
Act
Continuous Delivery on Cloud
Orient
Analysis
BIG DATA
Model Hypotheses
Incremental Features
CLOUD
Decide
Plan Response
Share Plans
JFDI
CULTURE

Cloud Trends

Cloud Enterprise IT Adoption



By Simon Wardley http://enterpriseitadoption.com/

What happened in the “price war” announcements?

Google Cloud

- Demo: Live instance migration
- Sustained Usage Pricing
  - Discount for over 25%/month
- New Google DNS Service
- New MS Windows Support
- Price cuts
  - Storage 2.6 cents/GB/month
  - Storage access 1 cent/10k ops
  - 32% reduction in instance cost
  - Cost for n1-standard == m3

Amazon Web Services

- Demo: Workspaces Cloud VDI
- New High Memory Instances
  - r3 cheaper & bigger than m2
- Updated Storage Instances
  - i2 cheaper & bigger than hi1
- Price cuts
  - S3 2.75-3.0 cents/GB/month
  - S3 access 0.4 cent/10k ops
  - m3 cheaper&faster than m1
  - c3 cheaper&faster than c1

What Was Missing Last Week

Google Cloud

- No big enterprise customers
- No reservation options
- Need more regions and zones
- Need lower inter-zone latency
- No SSD options

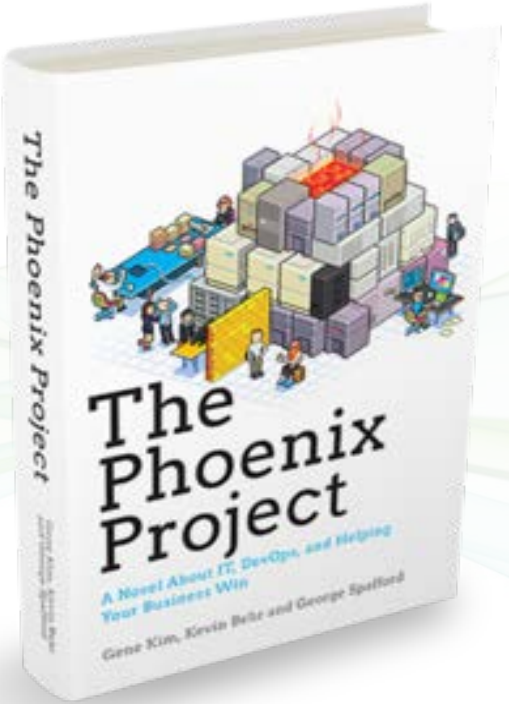
Amazon Web Services

- No per minute billing
- Need simpler discount options
- Need more regions and zones
- No integrated PaaS strategy
- No instance migration
- Need update for m1

Too many architectural differences make using both interchangeably tricky

How does IT get there?

"This is the IT swamp draining manual for anyone who is neck deep in alligators." **Adrian Cockcroft, Cloud Architect at Netflix**



New conference led by Gene Kim: Enterprise Devops – SF Oct 2014

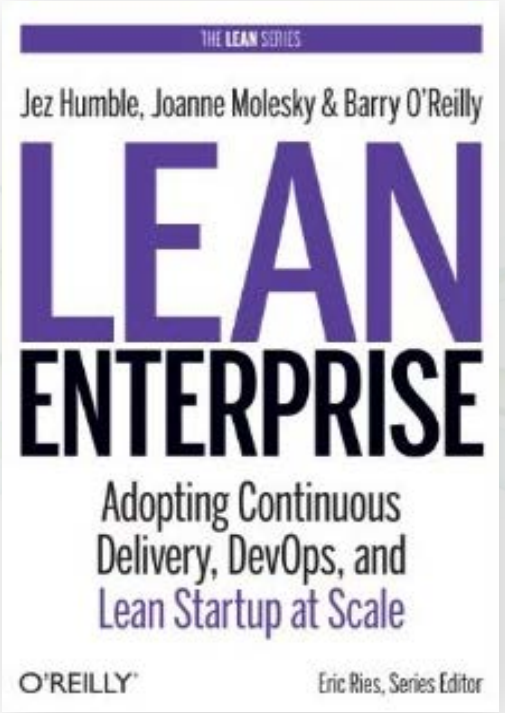
Continuous Deployment for Speed

New book 2014

Flow Conference:

Flowcon 2013 – See videos

Flowcon 2014 – September…



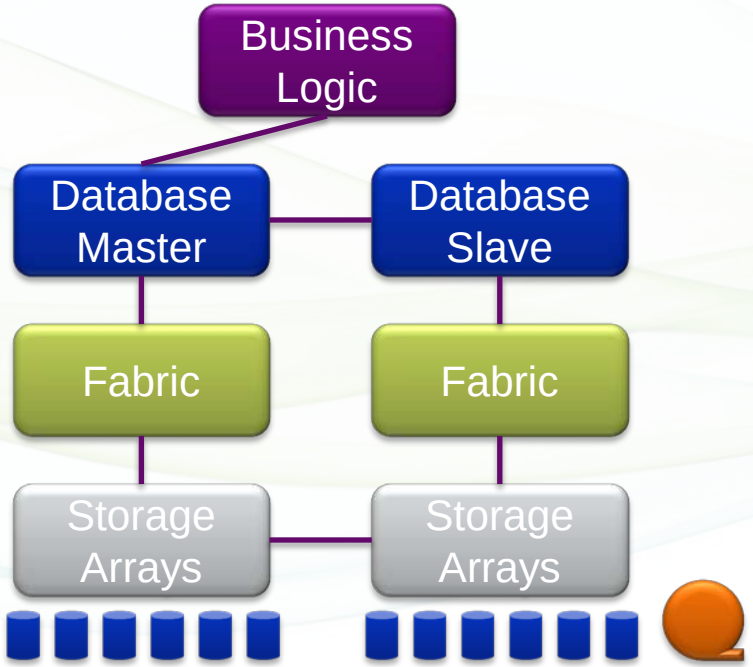
Speed wins in the marketplace

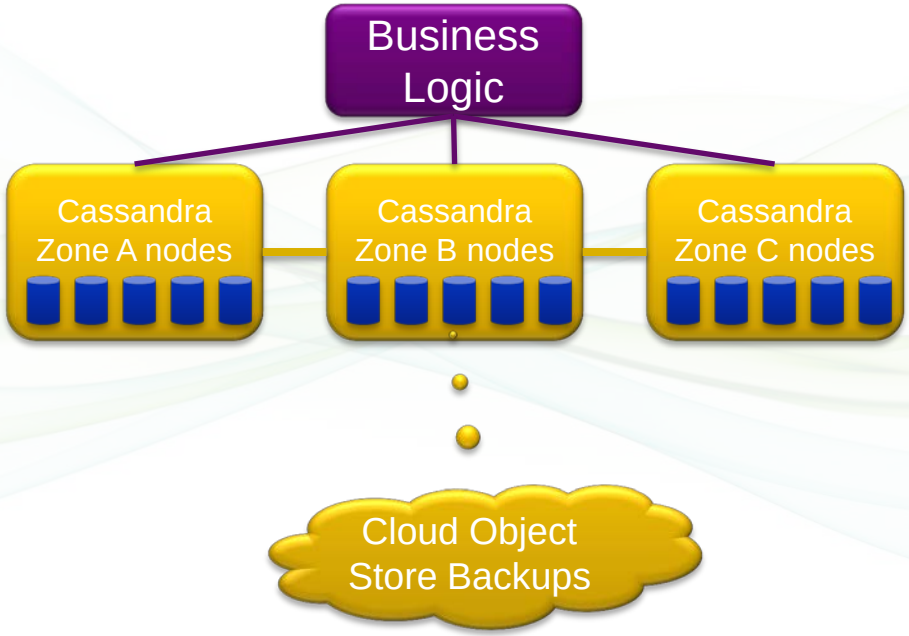
Cloud Native Storage Architecture

Cloud Native Storage Requirements

- Write data quickly
- Don’t lose data
- Read back quickly
- Read back what I wrote…
- Distribute data globally
- With low cost
- And not run out of space.
- ~~Make big profits for existing vendors~~

Traditional vs. Cloud Native Storage Architectures





Storage Node Options

- AWS – Create clusters of 100s of Cassandra nodes automatically
  - From nothing exists to writing data 10min after launching (see live demo)
- Older Node Specifications
  - AWS m2.4xlarge – 68GB RAM, 1Gbit net, 2x840GB disks, 500 iops
  - AWS h1.4xlarge – 60GB RAM, 10Gbit net, 2TB SSD, 100000 iops
  - AWS hs1.8xlarge – 117GB RAM, 10Gbit net, 24x2TB disk, 2.6Gbyte/s
- Current Node Specifications – Intel Ivybridge v2
  - AWS i2.xlarge – 30GB RAM, 10Gbit net, 800GB SSD, 45000 iops
  - AWS i2.8xlarge – 244GB RAM, 10Gbit net, 8x800GB SSD, 365000 iops

BV

Write Data Quickly

How it works

- Apache Cassandra
- Write to RAM on local node
- Duplicate to remote node RAM
- Flush to local disk every 10 sec
- Huge sequential writes
- Immutable pre-sorted files
- Infrequent compaction merge

Speed and Scale

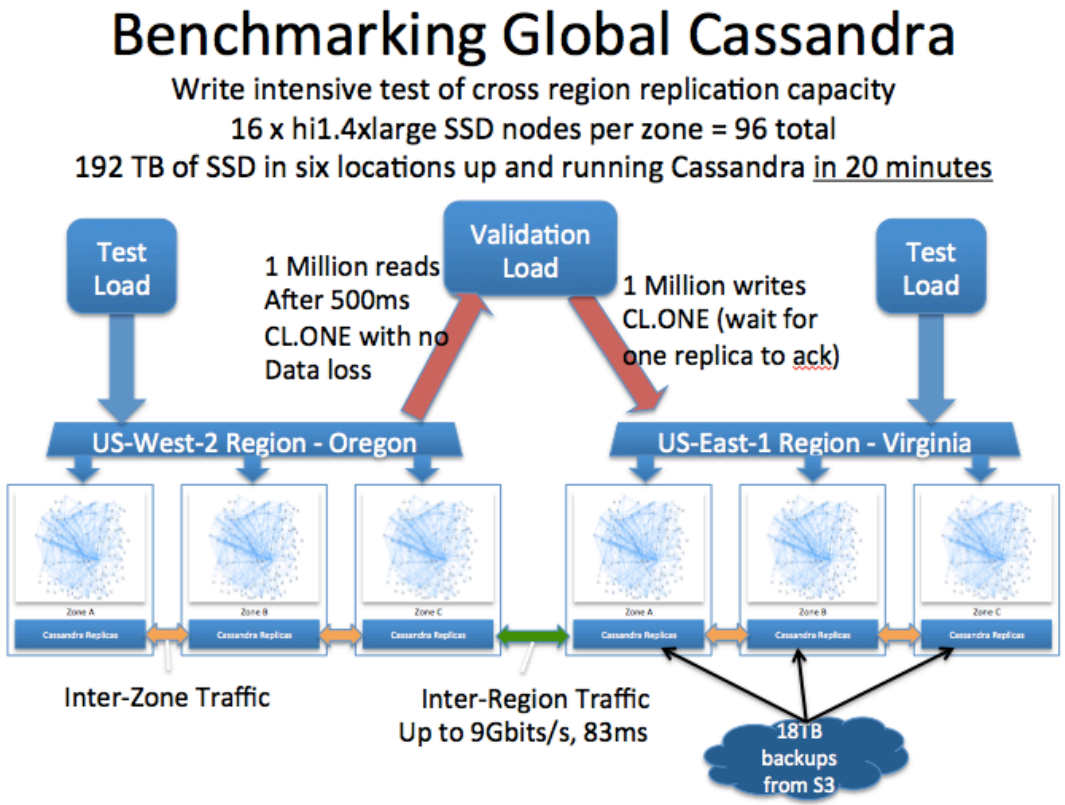
- Ack local copy, microseconds
- Ack remote copy, net latency
- Quorum write 2 out of 3 option
- Over 1M writes/s in production
- Scales linearly with nodes
- Netflix runs to 288 nodes/cluster
- Others over 1000 nodes

BV

Don’t Lose Data - Durability

- Triple replication of data - one replica per building (Availability Zone)
- Hinted handoff for down nodes
- Copy immutable files to S3 for backup, remote regions for archive
- Replica checksum compares
- Efficient anti-entropy repair
- Support option from Datastax

Distribute Data Globally – Netflix Benchmark Test



Do the Math on Costs

- Traditional Architecture Costs

  Nodes + Database licenses + SAN fabric switches + Storage arrays + Management tools + Replication tools + Backup tools + Tape backup drives + Off-site tape storage + Network switches + Power/cooling etc.

- Cloud Native Architecture Costs

  Nodes (includes storage) + Cassandra support + S3 backup

- Cloud Native Benchmarking

  Pay only for the hours that the cluster is running

- Cloud Native ~~Capacity Planning~~

  Start small, grow the cluster only when it's filling up, no downtime

Cloud Native Storage Requirements Met

- Write data quickly – millions of low latency ops/s
- Don’t lose data – triple AZ replication, backups to S3
- Read back quickly – millions of low latency ops/s
- Read back what I wrote… - checksums and automatic repairs
- Distribute globally – Cassandra cross-regional replication
- With low cost – using lots of cheap internal SSD
- And not run out of space – scales up to “plaid” (beyond ludicrous)
- ~~Make big profits for existing vendors~~ – sorry…

Cloud Native for High Availability at Scale

NetflixOSS at netflix.github.com and techblog.netflix.com

Over 40 projects, PaaS, NoSQL, Big Data, etc.

Priam – Cassandra co-process

- Runs alongside Cassandra on each instance
- Fully distributed, no central master coordination
- S3 Based backup and recovery automation
- Bootstrapping and automated token assignment.
- Centralized configuration management
- RESTful monitoring and metrics
- Automated online re-size to double node count

Cassandra Astyanax Java Client Recipes

- Distributed row lock (without needing zookeeper)
- Multi-region row lock
- Uniqueness constraint
- Multi-row uniqueness constraint
- Chunked and multi-threaded large file storage
- Reverse index search
- All rows query
- Durable message queue
- Contributed: High cardinality reverse index

Staash - Generic Data Access Layer Microservice

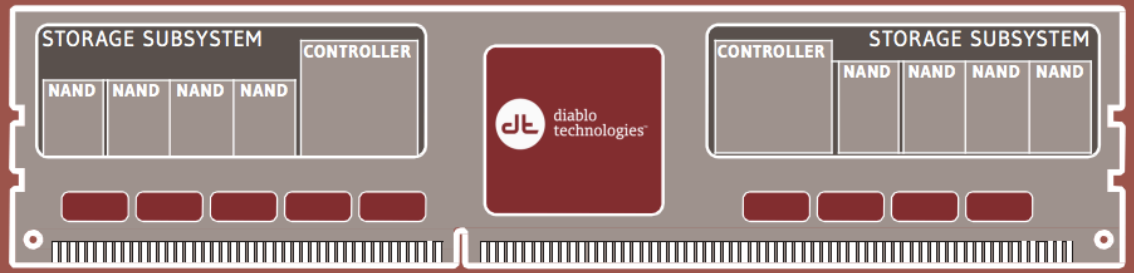
- Storage Tier As A Service over Http “STaaSH”
- Polyglot persistence via unified REST API
- Cassandra/Astyanax Recipe Implementations
- Cassandra and MySQL supported now
- More datastores under development…
- Will allow polyglot “join” across datastores

What's Next?

- Cassandra with local disk replaces Oracle/MySQL, SAN, Array etc.
- AWS S3 and Google Data Store replace tape for backup/archive
- Cloud prices halve every two years
- Epic Google and AWS price war (spoiler: everyone else dies…)
- How and when will Google compete with AWS SSD options?
- SSD moves to the memory channel, for even lower latency

SO, WHAT IS **MEMORY CHANNEL STORAGE**?

- An Architecture (not a single product)
  - Enables Flash Storage to Directly Interface on the Memory Channel
- Presents as a Block I/O Device
  - Can be Managed just like Existing Storage Devices
- DDR3 Interface, Standard RDIMM Physical Form Factor
  - Plugs into Standard DIMM Slots
  - Self-contained, No External Connections Required



diablo technologies

Disclosure: Diablo Technologies is a Battery Ventures Portfolio Company
See www.battery.com for a list of portfolio investments

Any Questions? Presentations by @adrianco

- Battery Ventures http://www.battery.com
- Adrian's Blog http://perfcap.blogspot.com
- Netflix Tech Blog http://techblog.netflix.com
- Netflix Slideshare http://slideshare.com/netflix

- Monitorama Opening Keynote Portland OR - May 7th, 2014
- GOTO Chicago Opening Keynote May 20th, 2014
- DevOps Summit at Cloud Expo New York – June 10th, 2014
- Qcon New York – June 11th, 2014
- GOTO Copenhagen/Aarhus – Denmark – Oct 25th, 2014

BV