ebay Storage, From Good to Great

Farid Yavari

Sr. Storage Architect - Global Platform & Infrastructure

September 11,2014

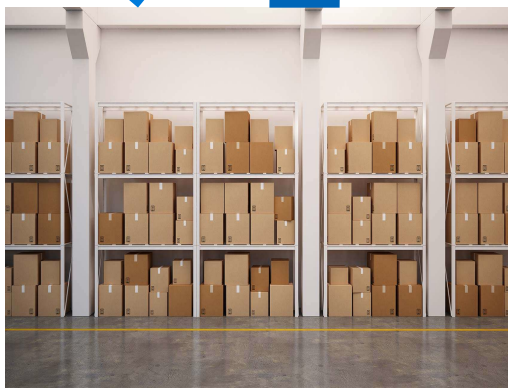
ebay Journey from Good to Great

2009 to 2011
TURNAROUND

2011 to 2013
POSITIONING FOR THE FUTURE

2013 to 2015
CAPITALIZE ON THE OPPORTUNITY

From Storage Perspective



20% Y/Y OLTP Growth

300% Y/Y Analytics Growth

Size of the Managed Infrastructure

- SAN ~ FC OLTP environment 80PB
- NAS ~ 6PB
- Object Store ~1EB by 2015
- Analytics environment ~270PB

- ~130 enterprise SAN/NAS/FLASH storage arrays in 3 major DCs
- ~1.4mil peak hour IOPS in SAN environment
- Thousands of servers with external storage

Storage Ecosystems

OLTP	Cloud (C3)	Analytics / Big Data	NoSQL
• FC/NFS → ISCSI, Object Store • Centralized Storage • SQL transactional • > 16TB/U → 150TB/RU • ~80PB	• ISCSI, Block + Object Store • Local attached → external Storage • KVM Openstack • Cinder (5PB) • Swift • >20TB/U → 120TB/U	• HDFS → Object Store • Local Attached → Disaggregate compute (Map Reducers) • 128TB/U → 256TB/U • Uneven tiering on Flash • Swift object store • ~270PB → ~500PB	• ISCI Block • MongoDB • Local Attached → external storage • Cinder (10PB) • Cassandra • Couch Base • Supports mobile apps

2017 Vision

- Flash Everywhere
- Memory Class Storage emerging
- Storage Density Scales to 1Pb/RU
- Storage networking moves to Ethernet
- Software Defined Everything
- Hyperscale OpenStack
- Exabytes of Analytics
- 2 Flash Tiers: Performance Flash and Big Data Flash

Foundation of an Elastic Infrastructure

Definition: An infrastructure that can spawn, destroy, grow, shrink and move processes dynamically and efficiently within and across data centers.

- Automated Control Plane
- Resource Pool
  - Compute
  - Memory
  - Storage
  - Low latency, High bandwidth interconnect
- Traffic Management
  - PCI Compliance
  - Security
  - QoS

Key Technologies to Enable an Elastic Infrastructure

- Control Plane
  - Virtualization / Containers/ Hardware
  - Orchestration of infrastructure resources
  - Normalization of resources
- Resource Pool
  - High Speed Networking (10Gbe, 40Gbe, 100Gbe, beyond)
    - RDMA enabled (routable layer3)
    - Lossless flow control
  - Memory Class Storage
  - New Media beyond Virtical Nand
- Traffic
  - Virtual Lan
  - Access Control

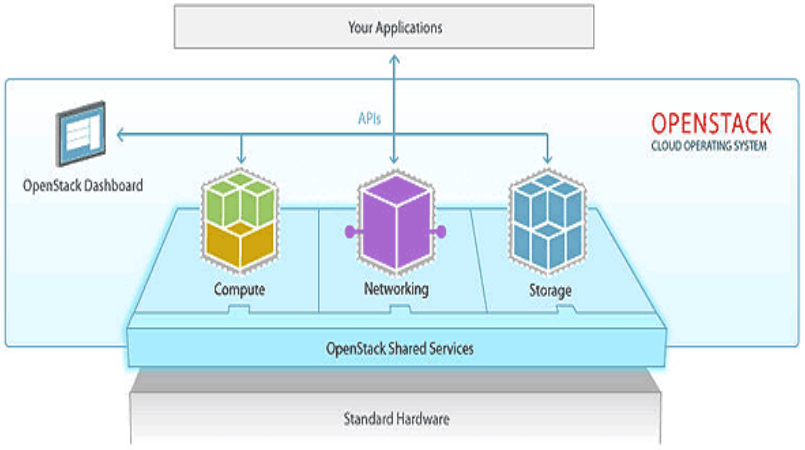


Image Credit: Open Stack

Key Initiatives to Enable an Elastic Infrastructure

- Separation of Storage and Compute
  - Hadoop use case
- Software defined storage, software defined network
- Cloud, SLA, OLA based services
  - Standardization
  - Automation
  - Show/Chargeback
  - Self Service

Shifting Paradigm of Storage

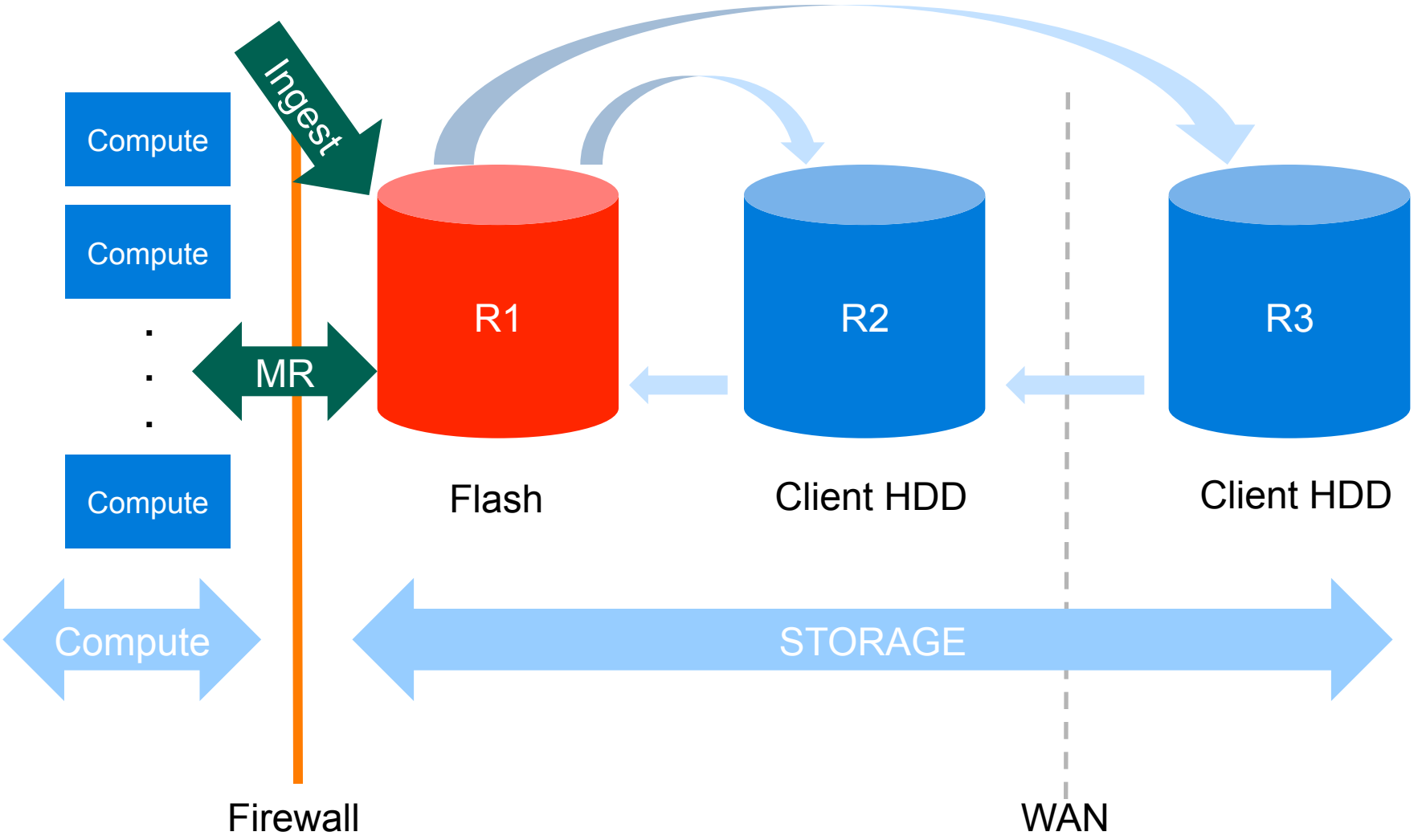
Tech	Bus	BW	Latency	Power	2014	2 yr	3 yr	4 yr
HDD (LFF)	SAS/ SATA	600MB/s 1.2GB/s	3 - 12ms	6-15 W	5-6TB	7-8TB	10TB	?
SSD (SFF)	SAS/ SATA	600MB/s 1.2GB/s	0.3 - 0.8ms	2-12 W	4-8TB	8-16TB	24TB	36-48TB
Flash	PCIE	2GB/s 3GB/s	2µs - 150µs	25 W	1-16TB	16TB+		
Storage Class Memory	PCIE	2+GB/s	1µs - 40µs 100's ns			8-24TB	24TB+	

Information based on Currently available products on the market and industry roadmap information

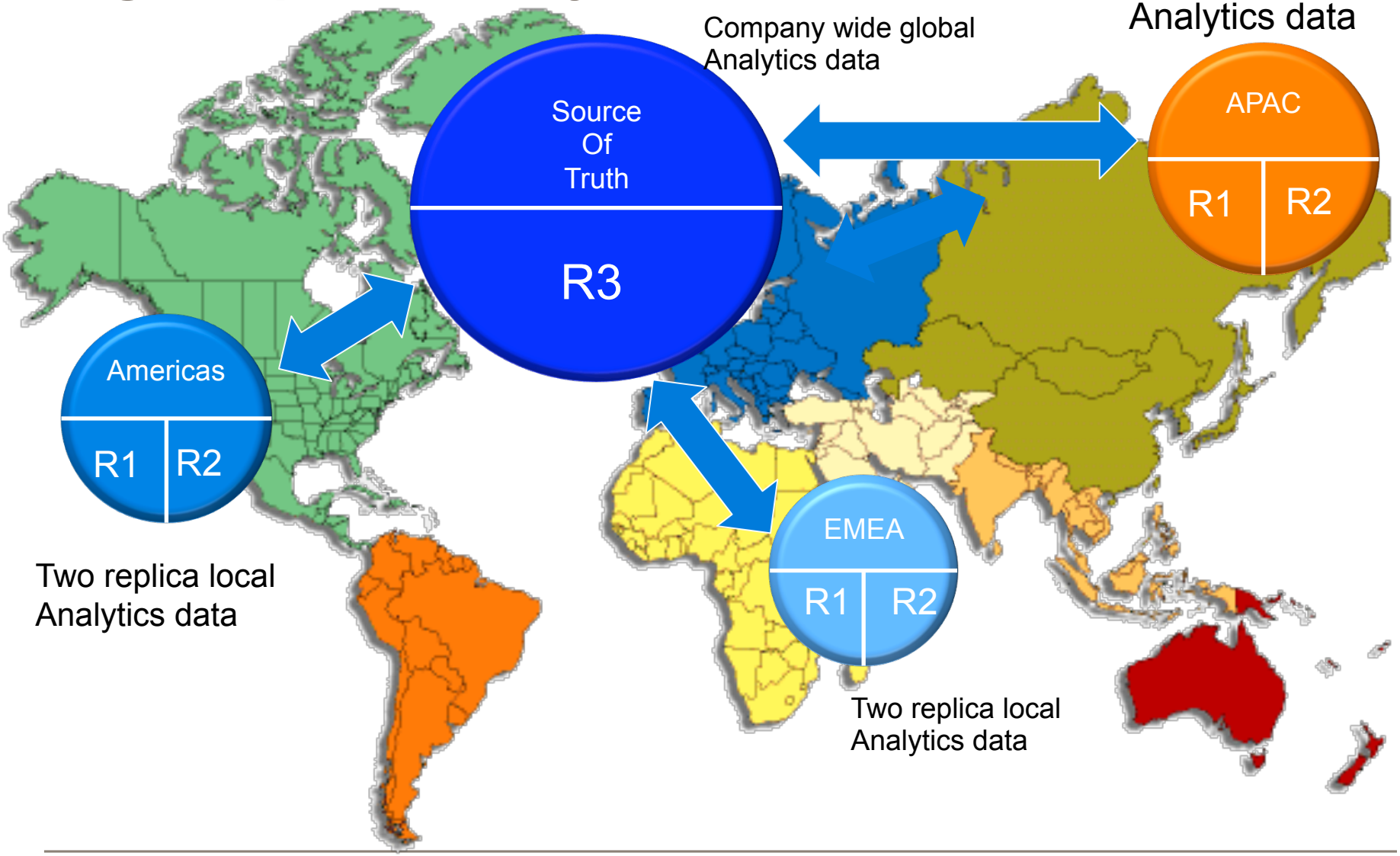
Big Data Flash

- Design Criteria
  - Merely beat disk in IOPs per TB
  - Heavy read workload, write seldom
  - $.30/Gb Target
  - 30 day power off retention and < 200ms power on response
  - 6Gb/s throughput
  - 8PB in a rack
- Hadoop Requirements
  - Tiered replicas
  - Abstract storage: Data Nodes separated from Map Reducers
  - Line Rate Networking

Hadoop Disaggregated Storage Model



Region Specific Analytics Model



Storage Class Memory

ReRAM/Memrister
- High Storage Density
- Low Power
- Memory Class latency (~100ns)
- Higher Endurance
- Cost < Flash
- Standard CMOS
- Long Scalability Roadmap
- Excellent Retention

Phase Change Memory (PCM)
- Moderate Density
- High Power
- Higher write latency
- Limited Endurance
- Cost < DRAM
- Non-CMOS
- Unknown scalability
- Write Disturbance

Spintronics (STT-RAM)
- Low Density
- High Power
- Lower Latency
- Higher Endurance
- Cost > DRAM
- Non-CMOS
- Limited scalability
- Good Retention

The Challenges

1. Storage growth in areas not traditionally designed for solid state such as Big Data will cause scaling challenges. How do we realize the benefits of flash while controlling the costs.
2. Storage Systems not understanding Flash
3. Software not understanding Flash
4. Capex Gap between technologies can make financial acceptance difficult.
5. Aggregated resources cause greater utilization imbalance
6. When will Flash need to pass the baton to Storage Class Memory.

The Opportunities

1. Flash designed with Big Data in mind will enable this technology to address the scaling and cost issues.
2. Scalable Software Defined Storage multi-rack ecosystems are well suited to data placement on flash vs. RAID. Tiering and metadata enhancements will introduce flash in more areas.
3. Gradually, applications, kernels and filesystems are beginning to treat flash as flash.
4. The Capex Gap can be met with Opex and comparative advantages. Quantify it! Additionally, Capex gap will close.
5. Networking advances will enable low cost storage disaggregation. (RDMA/IP, 40Gbps+, PFC)
6. Work is being done on optimizing for non-volatile memory but this will be an evolution. Doing this will allow for low cost memory scaling.

Q&A

ebay inc™