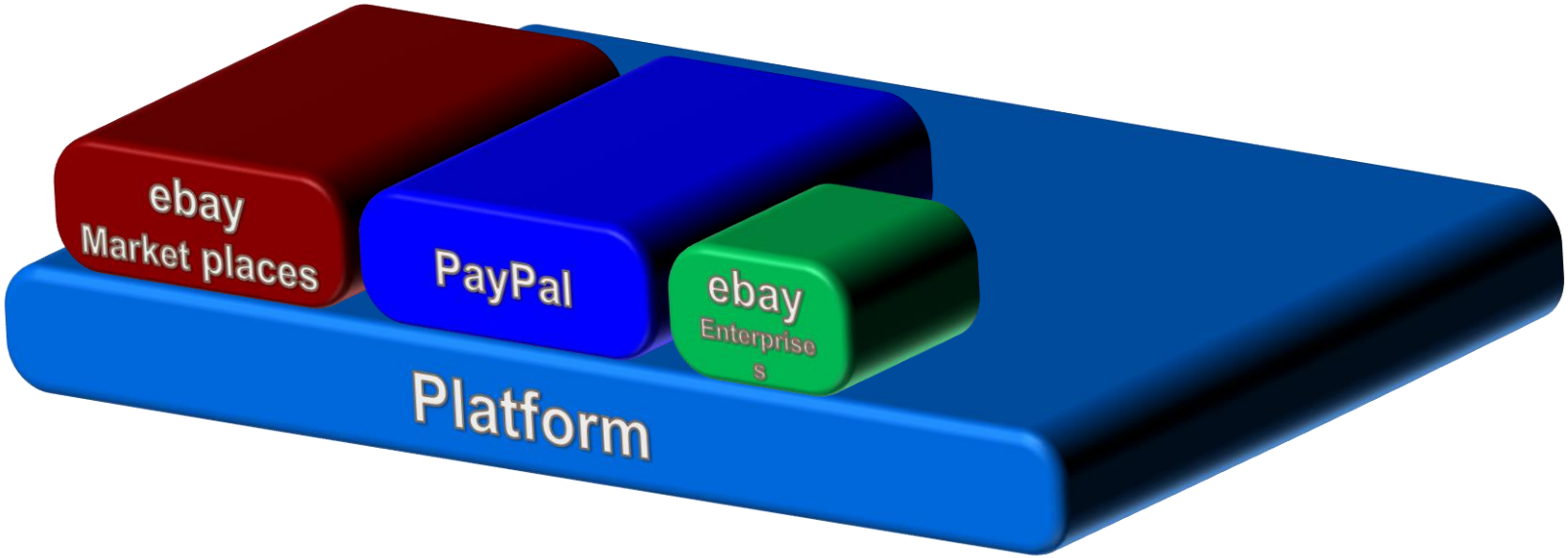
Scalable Digital Data Center Design

Russ McElroy

June 4th 2014

The eBay Inc. Portfolio



2017 Vision

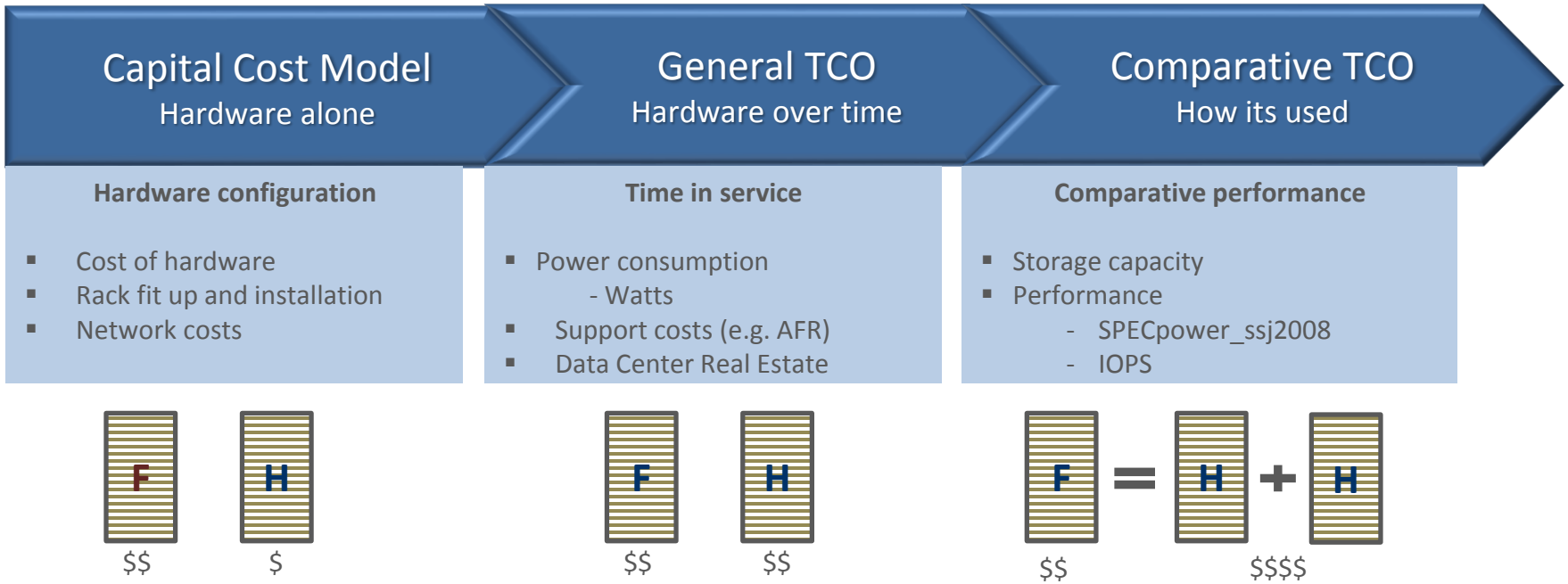
- Flash Everywhere
- Memory Class Storage emerging
- Storage Density Scales to 1Pb/RU
- Storage networking moves to Ethernet
- Software Defined Everything
- Hyperscale OpenStack
- Exabytes of Analytics
- 2 Flash Tiers: Performance Flash and Big Data Flash

Data Center Efficiency (dse.ebay.com)

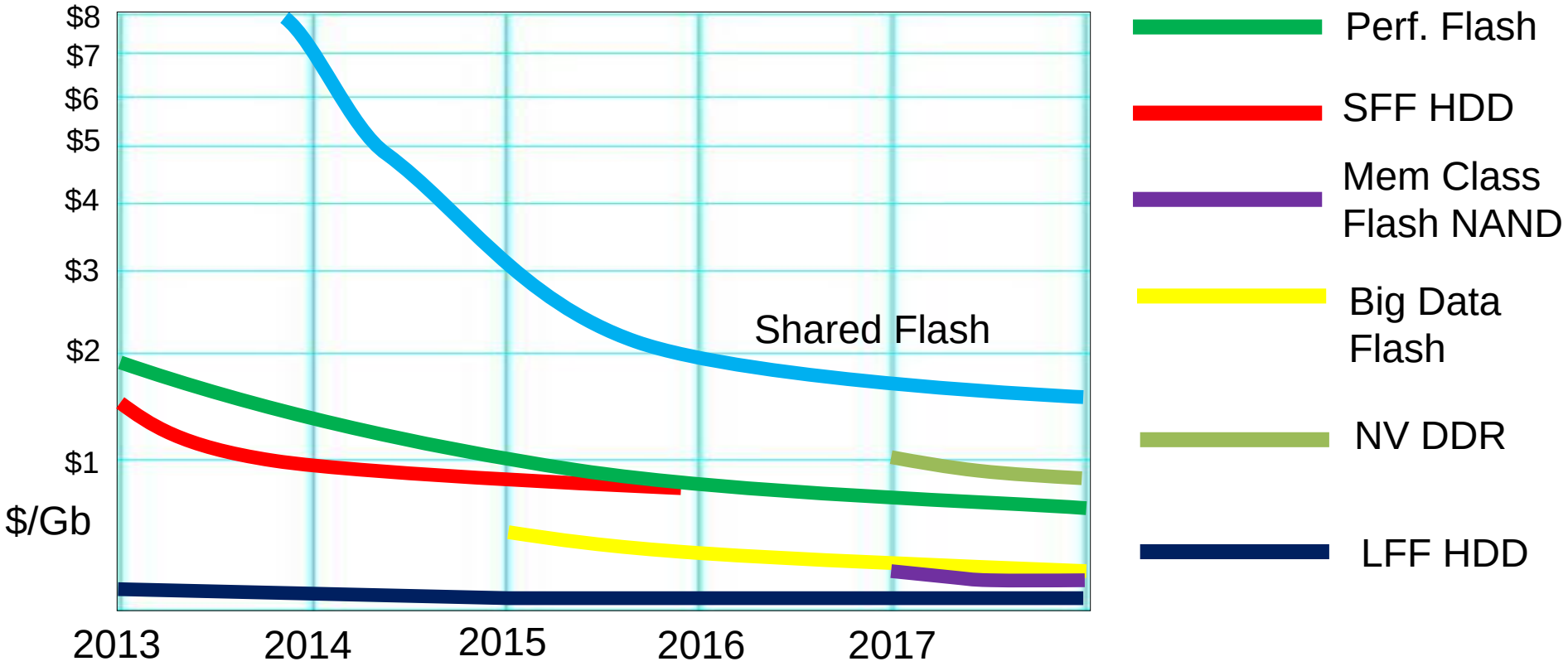
"You can't manage what you can't measure" Peter Drucker

Data Driven Decisions

TCO: Total Cost of Ownership
Example: Compare TCO of hardware F and hardware H



Capital Cost Estimates



Shifting Paradigm of Storage

Tech	Bus	BW	Latency	Power	1 yr	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+	

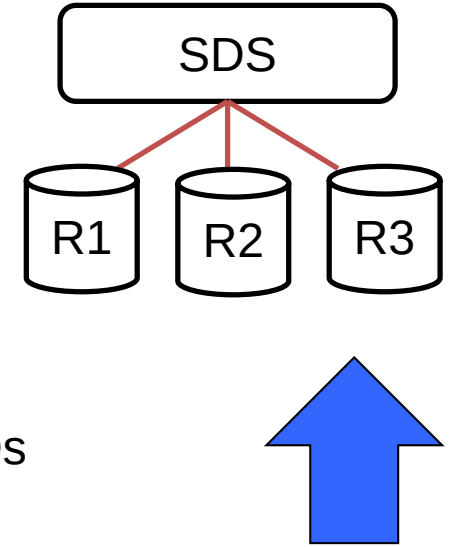
Commoditization of Flash Arrays

Software Defined Storage Ecosystem

ODM Dense Storage Enclosures

Rapidly increasing density in performance SSDs

RDMA/IP 40Gbps Storage Networking



= 100-200Tb / RU All Flash Performance Array Right Now

Flash is Flash (not disk)

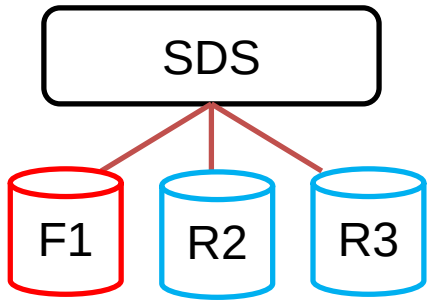
The Software Question: How do you use flash?

Optimizations are needed in:

- Kernels: Paging assumptions and NV MCS
- Applications: Legacy HDD Optimizations are everywhere
- Controllers: Fail flash as flash
- File systems: Re-represent flash
- Software Defined Storage: Design with flash in mind

Big Data Flash

- Design Criteria
  - Merely beat disk in IOPs per Tb
  - Heavy read workload, write seldom
  - 30c/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



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 Opportunity

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.


ebay inc™