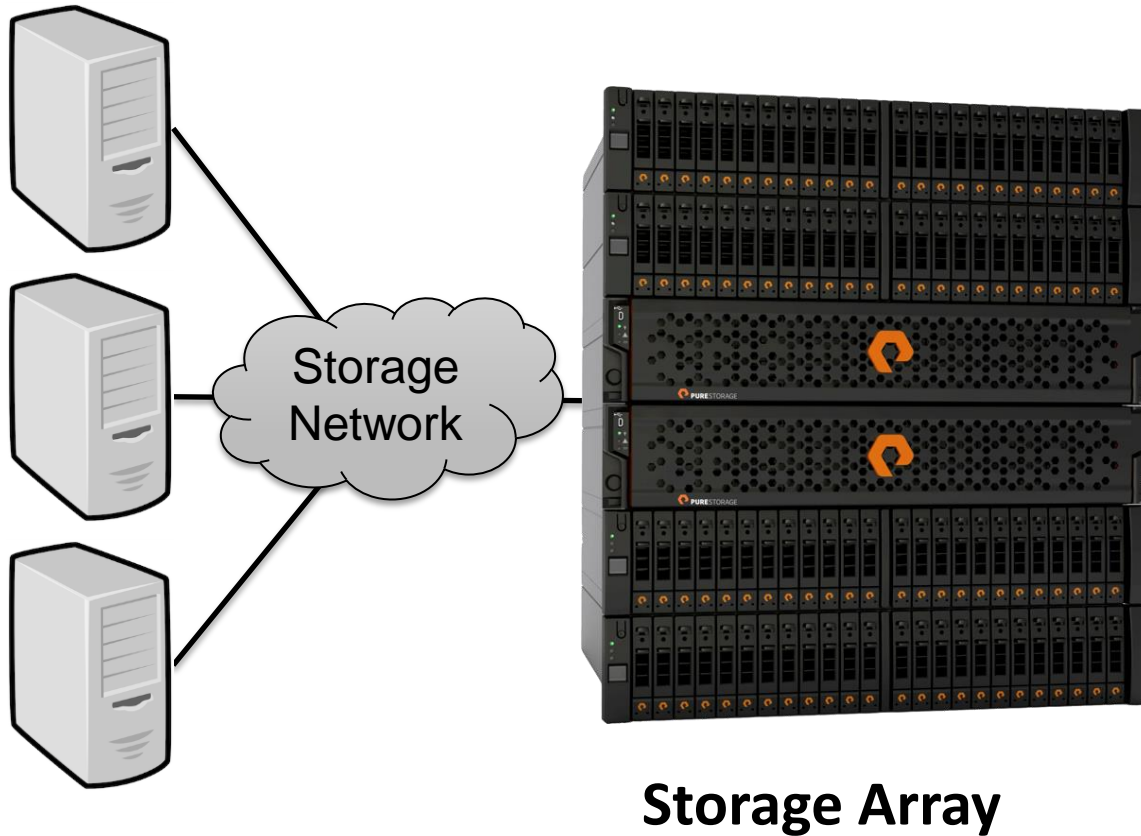




Flash in an Enterprise Storage Array - 10x Performance for Less Than Disk

Presented by Neil Vachharajani

Enterprise Storage Arrays



Consolidated, manageable, and reliable

Enterprise Storage: \$30B Market Built on Disk

- **Dominated by spinning disk**
 - Capacity is plentiful
 - Performance has stagnated
 - Random I/O workloads (virtualization) – subpar performance
- **Consumer space has transitioned to Flash**
 - Drives today's smartphones, cameras, USB drives
 - Laptops and desktops come with solid-state drives (SSD)
- **Why not just put SSDs into today's disk array?**
 - Current software systems are optimized for disk
 - Flash and disk are very different
 - Need storage arrays designed to leverage flash

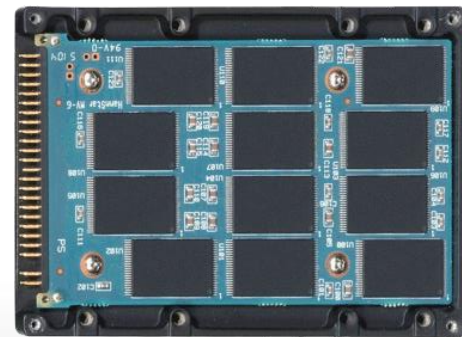
Flash: Opportunities and Challenges

- **Opportunities**

- Reads: Random access and fast
- Performance isolation
- Virtualized data layout

- **Challenges**

- Device longevity
- No in place overwrites
- Read/write performance asymmetry
- Cost \$\$

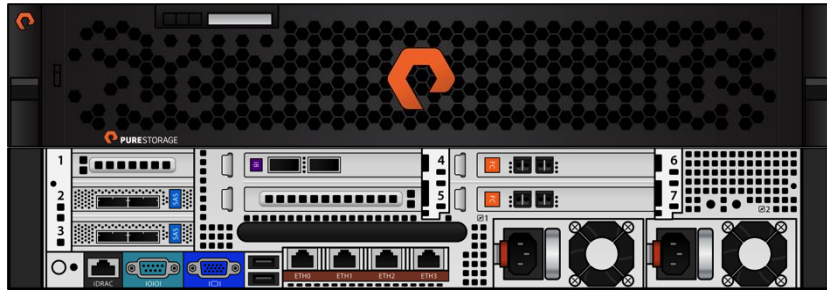


Pure Storage Architecture Principles

- **Trade raw performance for simplicity and lower cost**
 - Simplicity – prefer self-tuning system
 - Use CPU and surplus read bandwidth to reduce writes
- **Don't splurge on “enterprise” hardware**
 - Leverage cost trends in the consumer space
 - Optimize array wide, not at the individual SSD level

The Pure Storage Flash Array

FA-400 Controller



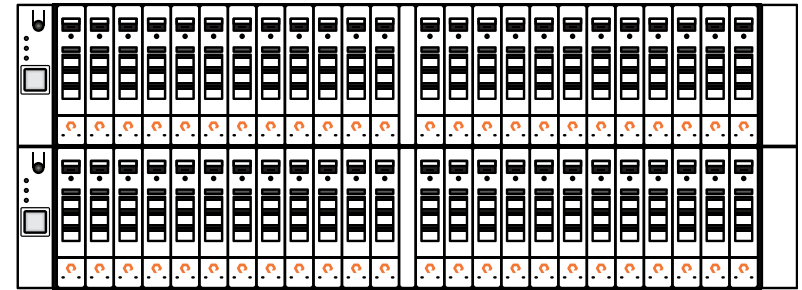
Performance*

- 400,000 8K IOPS
- 5 GB/s bandwidth
- <1ms average latency

Specifications

- 2x Intel “Sandy Bridge” 8-core CPUs
- 256GB DRAM
- 8Gb/s FC or 10Gb/s Ethernet
- 56Gb/s InfiniBand & 6Gb/s SAS
- 2U, 420W

Storage Shelves



SSDs

- 256 GB or 512 GB SSDs
- 100% MLC Flash

NVRAM

- Up to 2 NVRAM devices per shelf

Scale

- 20 – 100 TBs usable**
- 11 – 23 TBs raw flash (and growing!)

How Purity Works

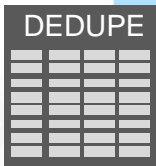


Protect

- Checksum
- Copy to NV-RAM

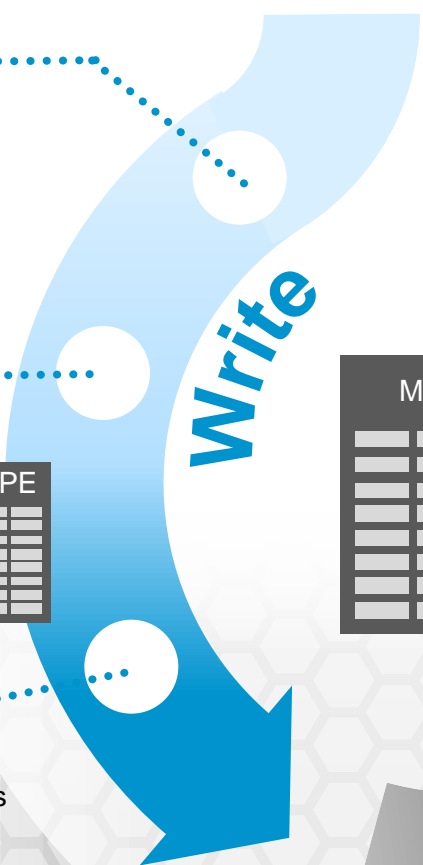
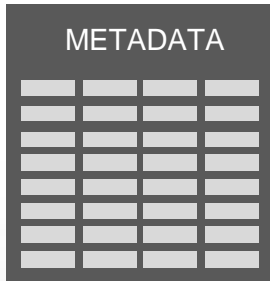
Reduce

- Pattern removal
- Deduplication
- Compression



Store

- Create RAID-3D segments
- Flush writes via flash scheduler



Validate & Serve

- Validate checksums



Multi-Path Read

- Read from fastest path via scheduler
- Decompress



Flash Management

- Global wear leveling & refresh
- Global deletion management
- Integrity checking



Flash Memory

Caring For Your Flash: Writing in Log Structure

SSD 1	SSD 2	SSD 3	SSD 4	SSD 5
D	D	D	P	Q
Spare Block	Spare Block	Spare Block	Spare Block	Spare Block
D	P	Q	D	D
P	Q	D	D	D

- **Aligning to SSD Geometry**

- Spare blocks encompass one or more SSD erase blocks
- Writes encompass one or more SSD pages

- **Contiguous sectors not contiguous on flash**

- Flash has great random read performance

FlashCare™: Optimizing Flash Globally

100% Virtualized Wide-Dispersed Data Layout

- No performance hot spots
- Evenly wears flash
- Bonus: no hot spares!

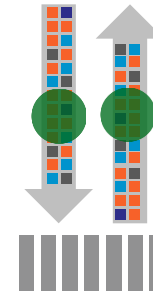


Flash Geometry-Aligned Writes



- Aligns with erase block boundaries
- Minimizes data movement “work” by SSD controller

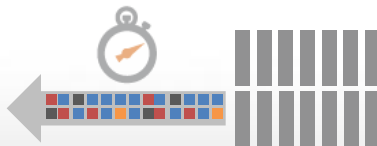
Non-Blocking Reads & Writes



- Micro-schedules each SSD
- Isolates reads and writes to a SSD
- Re-issues IO to alternate location if SLA exceeded

Deep Write Pipelining

- Manage volatile SSD caches
- SW tolerates flush latency
- Optimized to leverage SSD bandwidth



Continuous Background Optimization

- Handles garbage collection and wear management globally
- Periodically refreshes flash cells for longer retention
- Verifies data integrity



Flash Personality Layer

- Understands ideal IO fingerprint of each SSD
- Allows for mixing multiple generations of flash in one system



Conclusions

- **Data reduction in the field**



- Makes flash affordable for all
- **Enterprise system from commodity components**
 - Non-disruptive everything
 - From software updates to hardware upgrades
 - With the Purity OS, components have proven extremely reliable
 - High performance - < 1 ms latency typical