An In-Depth Examination of the Workings of an Enterprise-Class SSD

David Flynn, CTO Fusion-io
Enterprise-Class SSD Design

- Basic dimensions
  - Reliability & Data Integrity
  - Capacity
  - Performance
  - Longevity

- For each discuss...
  - Metrics
  - Raw Media capabilities (today & tomorrow)
  - Integration approaches (pros & cons)

- Scalability
Reliability cannot be compromised

Other requirements vary by workload
Raw Media Reliability

- **The GOOD**
  - No moving parts
  - Post infant mortality catastrophic device failures are rare
  - Predictable wear out

- **The BAD**
  - Relatively high bit error rate, which increases with wear
  - Higher density and MLC increases bit error rate
  - Program and Read Disturbs

- **The UGLY**
  - Partial Page Programming
  - Data retention is poor at high temperature and wear
  - Infant mortality is high (large number of parts….)
Controller Reliability Management

- **In-Flight**
  - Corruption upstream disk controllers
  - Corruption in SSD controller itself
  - Flush at power loss

- **At-Rest**
  - ECC
  - Scanning & scrubbing
  - Redundancy

- **Meta-data**
  - Error correcting memory
  - Data integrity field

Poor Media + Great Controller = Great SSS Solution
Capacity Performance Relationship

Access delay in time

6 orders of magnitude

Confidential Information: Fusion-io
Performance is about ROI

Lower CapEx
- Fewer CPUs
- Less RAM
- Less Network Gear
- Fewer SW Licenses
- Less Space

Lower OpEx
- Less HW Maintenance
- Less SW Maintenance
- Greater Uptime
- Less Power/Cooling
- Fewer Diverse Skills

HIGHER Productivity
Confidential Information: Fusion-io

Traditional SSD’s

Access delay in time

5 orders of magnitude
PCle Attached SSD’s

Access delay in time

3 orders of magnitude
Performance Dimensions

- Bandwidth
- IOPS
- Latency

Niche
Access Bandwidth

- **CPU**
- **L1 Cache, L2 Cache, L3 Cache, DRAM**
- **SAN, NAS, RAIDed DAS**

Access Bandwidth

- **PB**
- **TB**
- **GB**
- **MB**
- **KB**

- **TB/S**
- **GB/s**
- **MB/s**
Traditional SSD’s are no better

Access Bandwidth
Workload Segregation

Access Bandwidth

- CPU
- L1 Cache
- L2 Cache
- L3 Cache
- DRAM
- SSD
- SAN, NAS, RAIDed DAS
- Normal access
- Niche access
A cache needs...

- Bandwidth
- Mixed reads and writes
- Writes while full (saturated)

That’s exactly what SSD’s suck at!
(well traditional ones anyway)
PCle SSD’s are more like DRAM
Raw Media Performance

- **The GOOD**
  - Performance is excellent (wrt HDDs)
  - High performance per power (IOPS/Watt)
  - Low pin count: shared command / data bus → good balance

- **The BAD**
  - Not really a random access device
    - Block oriented
    - Slow effective write (erase/transfer/program) latency
    - R/W access speed imbalance
  - Performance changes with wear

- **The UGLY**
  - Some controllers do read/erase/modify/write
  - Others use inefficient garbage collection
Controller Performance Drivers

- Interconnect
- Number of NAND Flash Chips (Die)
- Number of Buses (Real / Pipelined)
- Data Protection (internal/external RAID; DIF; ECC…)
- SLC / MLC
- Effective Block (LBA; Sector) Size
- Write Amplification
- Garbage Collection (GC) Efficiency
- Buffer Capacity & Mgmt
- Meta-data processing
800 MBytes per second peak bandwidth

Half bandwidth at 4K packet size
Scalability

- Following Slides Show
  - Scalability of \{1, 2, 4, 8\} units
  - Only 1 SATA controller is used – limiting scalability
  - Only 1 thread running

- Measurements taken at Read/Write Ratios of
  - \{100/0, 75/25, 50/50, 25/70, 0/100\}
  - RMS value is the “root mean square” of these values

- IOPS measurement taken at 512 Byte Transfers

- Bandwidth taken at 128K Byte Transfers
  - Unless shown differently
  - Linux has a 128K limit
Scalability vs RW Ratio vs Block Size

SATA-B Scalability
IOPS

SATA-B Scalability
Bandwidth (MB/s)

Block Size Increasing

Santa Clara, CA USA
August 2009

37
Thank you!