SSD Design Considerations for Ultra-Low Latency

Bernie Rub
This presentation will touch on a few design considerations related to making SSDs faster and more responsive:

- How do latency and IOPS specs relate to actual usage?
- In the evolution towards lower latency, what’s after PCIe?
- Is it better to have one high capacity device or many small ones?
- Why write intensive workloads benefit most from reduced latency
- How increasing IOPS impacts endurance
• Some of the materials for this presentation have been provided by Diablo Technologies
Increasing Demand for Fast & Responsive Storage

- **Financial Services**
  - Low, deterministic latency transactions
  - Fast Interactive Data Analysis

- **Database/Cloud**
  - Increase Transactions per Second
  - Memcached consolidation

- **Virtualization**
  - Enable increased VMs per Node
  - Reduce capex and opex
  - Fast response times per VM

- **Blade Server**
  - Enable high density storage blades
  - Utilize empty DIMM slots

- **Big Data Analytics**
  - Increase transactions per second
  - Reduce response times for analytics queries

**Higher IOPS & Reduced, Deterministic Response Time**
Example: SPC Benchmark 1 Executive Summary
TMS RamSan-630

Typical Performance Curve

Opposite Trends for IOPS and Response Time

IOPS Specs are for very large queue depths. Results in unacceptably high response time.

Latency Specs are for Q=1
Apply only to unrealistically light workloads

Typical Usage
The Path to Ultra Low Latency

- Lower Latency
- Increased IOPS
- Deterministic Response Time
The Path to Ultra Low Latency

- Lower Latency
- Increased IOPS
- Deterministic Response Time

**Timeline**

- 100’s usec
- 10’s usec
- 1’s usec

**Flash DIMM**
Memory Controller Path provides > 10X lower latency and 2X higher bandwidth
Reducing Latency and Improving Predictability
Reducing Latency and Improving Predictability

Connecting Flash to the Memory Bus eliminates arbitration and data contention on the I/O hub
• **Mixed Workload: 70% Read / 30% Writes**
• **Response time with very light loads** (latency) is dominated by flash read time
• **Huge increase in response time** with heavy loads
For mixed workloads, latency is dominated by flash read time.

In this example, FlashDIMM with MLC has higher latency than PCIe SSD with SLC.

* ULLtraDIMM is a Flash DIMM product jointly developed by Smart Storage Systems and Diablo Technologies.
**PCIe SSD vs. Flash DIMM**
*(70/30 Workload – Expanded View)*

For read or mixed workloads, latency is dominated by flash read time.

In this example, FlashDIMM with MLC has higher latency than PCIe SSD with SLC.

*ULLtraDIMM is a Flash DIMM product jointly developed by Smart Storage Systems and Diablo Technologies*
PCle SSD vs. Flash DIMM

Linear scaling of IOPS with constant Response Time
Scaling with 8 Flash DIMMs

Linear scaling of IOPS with constant Response Time
• Many applications are sensitive to write latency
  – Processes that wait for writes to complete, e.g.
    ✓ Transactional Databases logging
    ✓ Virtual Desktop check pointing

• Write latencies in storage devices
  – Most storage devices incur significant latency performing the physical write to the media
  – High performance SSDs significantly reduce this latency
    ✓ Write data is immediately saved in power safe storage within controller
    ✓ Return of status is not gated by writing to flash
    ✓ Response time is dominated by IO Path

• By drastically reducing the IO path delays, Flash DIMM can achieve write latency << 10 µs!
Impact of Increasing Access Density

Access density quantifies the IO intensity of a workload.

Access Density = \( \frac{\text{IOPS}}{\text{Capacity}} \)

- **HOT**
- **COLD**
Increasing access density
## Endurance vs. Access Density

### Drive Writes per Day for Sustained 70/30 Read/Write Workload

\[
DWPD = \frac{\text{Write IOPS} \times \text{IOSize} \times 86,400}{\text{Capacity (B)}}
\]

- **Write IOPS** = **Read IOPS** \(\times\) \(\rho\) \(\times\) \%Write

  - Relative IOPS for 70/30 workload: 30%

\[
DWPD = \frac{\text{Read IOPS} \times 5.3\times10^{-2}}{\text{Capacity (GB)}}
\]

- **Access Density**
Endurance vs. Access Density

Increasing DWPD
Summary

- FlashDIMM reduces the IO overhead by connecting directly to the memory controller and bypassing the IO hub
- Modular storage elements scale IOPS without increasing latency
- Multiple FlashDIMMs outperform a single larger capacity PCIe SSD
- Required endurance is proportional to the access density
- Higher endurance is required to service “hotter” workloads