A Fast Write Buffer for All-Flash Arrays

Peter Kirkpatrick
Pure Storage

pete@purestorage.com
Motivation

- Storage performance = consistent low latency
- Value in all-flash storage: lots of processing
- So... we need a fast, non-volatile write buffer
Design Space

- Consider the whole stack…
- Performance: throughput and latency
  - Write throughput ~1GB/s .. 20GB/s
  - Latency sub-µs to ms
- Capacity
  - Few MB to many GB
- Availability, Serviceability, Scalability
  - Internal/external, stateless controllers
- Form Factor
- Protocols
  - Compatibility, performance, scaling
- Cost!
What is the Right Buffer Size?

- Top down: Cover write bursts
- Want: optimal latency to full throughput
- Client throughput * processing latency ~ capacity
  - Example: 1 GB/s * 0.1 s ~ 100 MB
  - Example: 20 GB/s * 2s ~ 40 GB

Client Writes → Buffer → Process → Persistent Store

Write Ack
Options

- Large (many GBs), slow buffer
  - Consider SSD, overprovision
- Large (many GBs), fast buffer
  - Consider PCM
- Small (MBs), fast buffer
  - Consider MRAM
- Modest (GBs), fast buffer
  - Consider DRAM+NAND+Caps
  - Optimize energy storage using NVDIMM
## Optimizing for Our Space

- Very high performance (10-100us latency, 1-10 GB/s)
- Modest capacity (several GB’s)
- External and modular, compact
- NVMe is optimal: must be dual-ported, hot-pluggable, reservations
- DRAM + NAND can trade cost (NAND) for size (caps)
  - We optimized with NVDIMM
  - hold up much lower power = lower energy storage, more compact

<table>
<thead>
<tr>
<th>External media options:</th>
<th>SSD</th>
<th>PCM</th>
<th>MRAM</th>
<th>DRAM+NAND</th>
<th>NVDIMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write Throughput</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write Latency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our Solution
NVRAM Write Throughput

Two modules

PCIe g3x8 Limit ~7 GB/s
NVRAM Write Latency (99.999%)  

Two modules

This tool has poor resolution ~50µs

True minimum latency is ~10µs
What did we Learn?

- DRAM on NVMe has excellent performance.
- External modules enhance system robustness.
- Dual port and hot plug NVMe can have high reliability.
- Optimizing for energy storage enables reasonable cost and physical size.
- Alternative media have promise, not optimal yet.