I/O Acceleration from the Bottom Up

How will new SSD technologies shape future data serving infrastructures?

Sangyeun Cho

Memory Business
Samsung Electronics Co.
A little bit of history
Areal Density Trend, NAND Flash Media

CAGR (over past 20 years) 83%

Year

MB/mm²

1000

100

10

1


Tb devices have arrived!

Efforts in QLC

(ISSCC Press Kit, 2020)
## Today’s NAND Flash Memory (in Production)

<table>
<thead>
<tr>
<th></th>
<th>TLC</th>
<th>QLC</th>
<th>SLC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Die Capacity</strong></td>
<td>512Gb</td>
<td>1Tb</td>
<td>64Gb</td>
</tr>
<tr>
<td><strong>Areal Density</strong></td>
<td>5Gbit/mm²</td>
<td>7.53Gbit/mm²</td>
<td>-</td>
</tr>
<tr>
<td><strong>Page Read Latency</strong></td>
<td>45μs</td>
<td>110μs</td>
<td>3μs</td>
</tr>
<tr>
<td><strong>Program Throughput</strong></td>
<td>82MB/s</td>
<td>18MB/s</td>
<td>160MB/s</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>ISSCC 2019</td>
<td>ISSCC 2020</td>
<td>ISSCC 2018</td>
</tr>
</tbody>
</table>
Demise of Performance Hard Drives

• In 2016~2017, Samsung introduced industry’s 1st enterprise SSDs built with 3D VNAND TLC
  – Status quo was to use planar SLC or eMLC

• Compelling MB/s, IOPS/$, IOPS/GB, and AFR advantages
• A 2.5” SSD offered capacity points from 0.5~16TB

<table>
<thead>
<tr>
<th></th>
<th>Performance HDD</th>
<th>SSD (PM1633a, 2017)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interface</strong></td>
<td>Dual-port SAS (6G~12G)</td>
<td>Dual-port SAS (12G)</td>
</tr>
<tr>
<td><strong>Density</strong></td>
<td>250~600GB</td>
<td>0.5~16TB</td>
</tr>
<tr>
<td><strong>Sequential</strong></td>
<td>&lt;400MB/s</td>
<td>1,200MB/s (Read); 900MB/s (Write)</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IOPS</strong></td>
<td>&lt;1K</td>
<td>200K (Read); 31K (Write)</td>
</tr>
</tbody>
</table>
Achieving High Density

- When mass-produced in 2017, 16TB PM1633a was the world’s highest capacity drive (yes, including HDDs)
- A novel “scale-out” architecture
  - Main controller + many sub-controllers
  - Industry’s 1st use of LPDDR4 DRAM in enterprise storage
Short-circuiting data to compute
Moving Data vs. Moving Compute

Traditional data processing

- Data
- Storage Server
- Host interface / Network
- Processing Server
- Client

Near-data processing

- Data
- Storage Server
- Host interface / Network
- Processing Server
- Client

Issues file (block) requests to storage servers

Database Server

Flash Storage Servers

Fast NVMe SSDs

Alternatively, storage servers may perform database scanning ("smart scan")

AWS AQUA Architecture

EXADATA

Compute nodes

High speed networking

Sub query

Filtered & aggregated results

Parallel execution

AQUA layer

AWS-designed analytics processor

SSD

scale-out architecture

Amazon S3

Durable storage
Pushing Compute to the Far End

Host Interface Controller

CPUs

On-Chip SRAM

Flash Memory Channel #0

Flash Memory Channel # (n-1)

NAND Flash Array

CPU(s)

DRAM Controller

DRAM

Flash Memory Controller

ECC

Flash Memory Controller

ECC
Pushing Compute to the Far End

[Cho, Park, Oh, Kim, Yi, and Ganger. “Active disk meets flash: a case for intelligent SSDs.” ICS 2013]
Data Processing Throughput

ISSD-XL: intelligent SSD with an accelerator (stream processor) per flash memory channel
ISSD-800: intelligent SSD with an embedded processor per flash memory channel running @800MHz
ISSD-400: intelligent SSD with an embedded processor per flash memory channel running @400MHz
Host-*: host server processing with I/O bandwidth of *

[Cho, Park, Oh, Kim, Yi, and Ganger. “Active disk meets flash: a case for intelligent SSDs.” ICS 2013]
Throughput Efficiency

Solid lines capture "iso-performance" points with intelligent SSD processing (# channels) vs. host CPUs (# cores)
Dotted lines capture "iso-performance" points with intelligent SSD processing + acceleration vs. host CPUs

[Cho, Park, Oh, Kim, Yi, and Ganger. “Active disk meets flash: a case for intelligent SSDs.” ICS 2013]
Energy Efficiency

[Cho, Park, Oh, Kim, Yi, and Ganger. “Active disk meets flash: a case for intelligent SSDs.” ICS 2013]
Near Data Processing with Biscuit

- An intelligent SSD for In-Storage Compute (ISC)
- Strong emphasis on programmability
  - User-friendly C++11 based programming model
  - Dynamic loading of user binary onto SSD
  - Seamless support for hardware acceleration

Biscuit Programming Model

- Biscuit follows a **data-flow model**
  - Data movement through ISC tasks determines their order of execution
  - On receiving all required inputs, an ISC task produces output and passes it to the next ISC tasks in the dataflow path

- A Biscuit program is composed of **ISC tasks and a host-side program**
  - An ISC task is a unit of work that runs on an ISC-enabled SSD
  - Both run concurrently in the SSD and the host, respectively

![Diagram of Biscuit Programming Model]
Due to the interface speed limit (PCIe x4 in this case), SSD’s internal bandwidth is ~30% higher.

Data inspection and I/O inside the SSD results in 10~20% reduction in latency + resilience against host CPU loads.
YourSQL on Biscuit

Key design considerations

- Partitioning of host/ISC tasks
- Defining interfaces between the host and ISC tasks
- Optimized query planner for ISC
- Reorganized datapath for ISC

[Jo et al. “YourSQL: A High-Performance Database System Leveraging In-Storage Computing.” VLDB 2016]
Evaluation Results

Up to 167\times gain!

15\times gain for top 5 queries

24\times lower energy

On dual Intel Xeon E5-2640 w/ 64 GB DRAM
Samsung PM1725 1TB SSD
MariaDB 5.5.42 w/ modifications using Biscuit
Workload is TPC-H w/ a scale factor of 100

[Jo et al. “YourSQL: A High-Performance Database System Leveraging In-Storage Computing.” VLDB 2016]
Samsung SmartSSD™

Performance scales as we add SSDs
Moving Forward

- “Computational Storage” is being standardized at SNIA/NVMe
  - What target applications?
  - What programming models?
  - How to coordinate and maximize the use of all platform resources?
  - Which data access mode?
  - Which interconnect technologies?
  - How to best utilize many computational storage devices?
Getting the most from the media
Logical View of Physical Media

• The LBA interface (introduced circa 1986) has helped straightforward switching to SSD
Logical View of Physical Media

**SSD**
- Read Page()
- Write Page()
- Erase Block()
- Copy Back Page()

**HDD**
- Move Arm
- Read Sector
- Write Sector

**Generic Block Layer**

**Host**
- Application
- OS
- File System

**Disk**

**NAND Flash Memory**

**OS**

**Logical Block Address**

**Sector base**
Fresh vs. Sustained Performance

Throughput

SSD-internal Operations

Multi-Streamed SSD

- Published at HotStorage 2014
- Standardized in 2017 (SAS/NVMe)
  - Linux support since 2017
- Product debut in 2016~2017

Simple, intuitive, additive model; Model concrete enough to predict effects

Model is still abstract; host can’t control data placement on specific physical units

Open-Channel SSD

- Philosophy-wise, OC-SSD aims to expose the media to the host software for direct management
  - Eliminate (parts of) FTL and give full control of data placement and access schedule to media units

Host has complete control over data placement on NAND flash media (no LBA);
Opportunities exposed for “cross-layer” optimizations between applications, file system, and FTL

Media idiosyncrasies underestimated;
Would you go back to CHS addressing from LBA?

[“Open-Channel Solid State Drives NVMe Specification.”
Revision 1.2, April 2016]
Zoned Namespace (ZNS) SSD

- SSD capacity is split into “zones” that are sequentially written
  - An SSD zone is analogous to that of shingled magnetic recording HDDs

Host has control over data placement on NAND flash media;
Complicated media management resides within the SSD

Host software must be aware of zones (SMR support is leveraged);
Design trade-offs still being explored

(zonedstorage.io)
Samsung Introduces Its First ZNS SSD With Maximized User Capacity and Enhanced Lifespan

Maximum available storage capacity and 3-4x longer lifespan enable server systems to run big data and AI applications more reliably and efficiently.

- ZNS SSDs are available and are poised to offer strong use cases for large storage systems
  - Very concrete interface
  - Good fit for many-bit cell technologies

- Software availability and readiness remains a challenge for users

- More end-to-end software building and design trade-off studies are needed
Physical Isolation of Storage Resources

VM1  VM2  VM3  ...

SSD Controller

Capacity partitioning, data placement

Hardware acceleration of I/O and key functions

Per partition (VM) bandwidth allocation, per partition performance QoS/schedule guarantee, per partition house keeping, ...

TEE* I/O Capable Host

Firmware Attestation

Expected value  =  Measured value

Hypervisor  ↪  Platform RoT

Main Memory

VM #1  VM #2

Data-in-use

Firmware Attestation

Hardware Isolation

Fine-grained Encryption

Hardware acceleration of I/O and key functions

Per partition (VM) bandwidth allocation, per partition performance QoS/schedule guarantee, per partition house keeping, ...
Outro

• SSDs offer the density and performance required by modern workloads and infrastructures
  – In turn, SSD idiosyncrasies affect how systems are designed

• System changes are expected to realize ideas around SSDs
  – Short-circuiting of data and compute
  – NAND flash media aware storing/retrieving of data
  – Hardware-level isolation support for multi-tenancy

• Future SSDs offer system level optimization opportunities
  – Further end-to-end software building efforts are needed
  – Novel data-compute mapping/cooordination ideas are wanted
I/O Acceleration from the Bottom Up

How will new SSD technologies shape future data serving infrastructures?

Sangyeun Cho
Memory Business
Samsung Electronics Co.