Origin of Million IOPs throughput and microsecond latencies in NVMe Enterprise SSDs

Andrei Konan
Enterprise SSD Firmware Engineer
Topics to cover

- Introduction
- SSD Generic Architecture
  - NAND Flash basics
  - Over-provisioning
  - Read and Write Operation Flow
- Workloads definition
- Bottlenecks
- Affecting Enterprise features
- Performance calculator
Abstract

Creation of new SSD product is driven by two main aspects: cost reduction and performance increase. Presentation will cover second major aspect of new SSD development – Performance and Quality-of-Service (QoS) improvements. Having SSD cost at the same level as competitors leaves only single opportunity to win the business – better throughput, consistency and latency. Nowadays Enterprise SSDs have broken Million IOPs throughput level and advertising microsecond latencies for some workloads.

Full text paper can be found at the following link
Introduction
Introduction

- Two main aspects of new product development
  - Cost reduction
  - Performance improvement
- PCIe 4.0 speed is 8GB/s (2 million IOPs of 4KB size)
- Majority of SSDs barely meet even PCIe 3.0 speeds
SSD Generic Architecture
NAND Flash basics – operation

- Page based Program, Block based Erase

**TLC Physical Block example**
NAND Flash basics – geometry

- Multiple working in parallel units (Planes) in one chip

Can Read and Program in Parallel with some limitations
NAND Flash basics – metrics

- Page Read (tR) = 50us
- One-Shot TLC Page Program (tPROG) = 2000us
- Block Erase (tBERS) = 6000us

- QoS (Latency) related features:
  - Program / Erase Suspend for Read
  - Plane Interleaving for Read
Over-provisioning – definition

A. Flash chip size is fixed by NAND Vendor

B. Drive external capacity is fixed by IDEMA standard

C. (A) – (B) is SSD over-provisioning (OP)
   - Read Intensive Drives OP is ~10%
   - Mixed Workload Drives OP is ~30%
Over-provisioning – impact

- Bigger OP = Less garbage collection (GC) in background
- Less GC = More time to Handle Host Write Traffic
- GC vs Host Write Traffic ratio is named Write Amplification (WA)
  - Read Intensive Drive (OP~10%): WA is ~ 4:1
  - Mixed Workload Drive (OP~30%): WA is ~ 2:1
Write and Read Operation
Write Operation Flow

1. Write
2. Status
3. Flush
4. GC

Controller

Front End

Flash Translation Layer

Back End

SSD

Cache

Garbage Collector

ECC

NAND

NAND

NAND

SRAM

DRAM

Power Loss Protection POSCAPs

Host

PCIe
Read Operation Flow

Host

SSD

Front End

(1) Host Read

Flash Translation Layer

(2) NAND Read

(2') Retry

Controller

Back End

NAND

NAND

NAND

SRAM

DRAM

PCIe

Power Loss Protection POSCAPs
Workloads and Bottlenecks
Workloads definition – parameters

- Queue Depth – maximal amount of SSD outstanding commands
- Block Size – 128KB for Sequential access, 4KB for Random
- Access Type – Read, Write, Trim, Mixed
- Drive State – Fresh (Formatted), Steady (Sustained)
- Examples
  - QD256 4K RND WR
  - QD1 128K SEQ RD
Workloads definition – metrics

- Bandwidth / Throughput
  - Sequential - GB/s, Random - KIOPs/MIOPs

- Consistency
  - Enterprise requires over 95%

- QoS (Latency)
  - Average
  - Low nines: 99%, 99.9%, 99.99%
  - High nines: 99.999%, 99.9999%, 99.99999%
Bottlenecks

- Host interface (PCIe, SATA, SAS)
- ECC encoder / decoder bandwidth
- DRAM interface speed
- NAND operation (tPROG, tR, tBERS)
- NAND Endurance (Retry Count)
- NAND interface speed
- NAND Channel and Die count
- Internal CPU instruction bandwidth
- Enterprise Feature
Affecting Enterprise Features

- Power Loss Protection
- Power Consumption
- Consistent Performance is more preferred than Peak one
- User Data Reliability and Security is of top priority
- Attention up to 10-nines QoS requirements
- Extra high densities (up to 64TB)
- Overheating
Performance calculator – Sequential

**Sequential Write**

- **Host**
  - PCIe Gen4
  - 6.5 GB/s

- **Controller**
  - Front End
  - Back End
  - ECC
  - 500MHz, 8b
  - 8 GB/s

- **NAND**
  - tPROG : 2ms
  - 98KB(*) / tPROG
  - 128 NAND Dies
  - 6.2 GB/s

  \[ X \text{ GB/s} \]

**Sequential Read**

- **Host**
  - PCIe Gen4
  - 6.5 GB/s

- **Controller**
  - Front End
  - Back End
  - ECC
  - 500MHz, 8b
  - 8 GB/s

- **NAND**
  - tR : 50us
  - 32KB(**) / tR
  - 128 NAND Dies
  - >10 GB/s

  \[ X\pm50\% \text{ GB/s} \]

(*) – 8KB physical page * 4 planes * 3 pages per one-shot, (**) – 8KB physical page * 4 planes
Performance calculator – Random

**4KB Random Write**

- **Host** to **Controller**:
  - PCIe Gen4: 1.7 MIOPs
  - Front End: Cache
  - Back End: GC
  - ECC
  - 500MHz, 8b
  - 2 MIOPs
  - 400 KIOPs

- **NAND**:
  - tPROG: 2ms
  - 98KB(*) / tPROG
  - 128 NAND Dies
  - 1.6 MIOPs
  - 320 KIOPs

- **Y MIOPs**

**4KB Random Read**

- **Host** to **Controller**:
  - PCIe Gen4: 1.7 MIOPs
  - Front End
  - Back End
  - ECC
  - 500MHz, 8b
  - 2 MIOPs
  - 4KB(/**) / tR

- **NAND**:
  - tR: 50us
  - 4KB(**) / tR
  - 128 NAND Dies
  - 2.5 MIOPs

- **< 1.5 MIOPs**

- **Y±50% MIOPs**

(*) – 8KB physical page * 4 planes * 3 pages per one-shot, (***) – single 4KB Host Read will occupy entire Die

Only 20% throughput is for Host Write, 80% for GC

Write Amplification is 1:4
Thanks a lot!

Q & A