Development Techniques and Tips for Maximizing NVMe Performance

Tom Friend
Toshiba Memory Corporation
You need the right attitude

- It’s the little things that count
- Be prepared to spend time on arcane details
- Take lots of notes, investigate all your questions
You need the right tools

- Begin with a PCIe analyzer
  - The only way to see what’s going on

- Why?
  - For the same reason(s) you would use a SAS or SATA analyzer
More useful tools

- Kernel level debugger
  - Allows you to track resource issues
- Test system with SATA or SAS drives
  - Stable filesystem can be useful
- Paper notebook
  - Always remembers, easily accessible.
Software Tools

“lspci”- lists all PCIe devices and details

00:1c.2 PCI bridge: Intel Corporation Wildcat Point-1F PCI Express Root Port #3 (rev e3) (prog-if 00 [Normal decode])
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAAbort- <TAAbort- <MAbort- >SERR- <PERR- INTx+
Latency: 0, Cache Line Size: 64 bytes
Interrupt: pin C routed to IRQ 0
Bus: primary=00, secondary=02, subordinate=02, sec-latency=0
Memory behind bridge: f1000000-f10ffffff
Secondary status: 66MHz- FastB2B- ParErr- DEVSEL=fast >TAAbort- <TAAbort- <MAbort- <SERR- <PERR-
BridgeCtl: Parity- SERR- NoISA- VGA- MAbort- >Reset- FastB2B-+
PriDiscTmr- SecDiscTmr- DiscTmrStat- DiscTmrSERREn-
Capabilities: [40] Express (v2) Root Port (Slot+), MSI 00
DevCap: MaxPayload 128 bytes, PhantFunc 0
ExtTag- RBE+
DevCtl: Report errors: Correctable- Non-Fatal- Fatal- Unsupported- RlxBfrd- ExtTag- PhantFunc- AuxPwr- NoSnoop-
MaxPayload 128 bytes, MaxReadReq 128 bytes
DevSta: CorrErr- UncorrErr- FatalErr- UnsuppReq- AuxPwr- TransPend-
LnkCap: Port #3, Speed 5GT/s, Width x1, ASPM L0s L1, Exit Latency L0s <512ns, L1 <16us
ClockPM- Surprise- LLActRep+ BWNot+ ASPMOptComp-
ASPM L1 Enabled RMB 64 bytes Disabled- CommClk+
ExtSynch- ClockPM- AutWidDis- BWInt- AutoBWInt-
LnkSta: Speed 2.5GT/s, Width x1, TrErr- Train- SlotClk+ DLActive+ BMgmt+ ABMgmt-
SltCap: AttnBtm- PwrCtl+ MRL- AttnInd- PwrInd- HotPlug- Surprise-
Slot #0, PowerLimit 10.000W; Interlock- NoCompl-
SltCtl: Enable: AttnBtm- PwrFlt- MRL- PresDet- CmdCplt- HPReq- LinkChg- Control: AttnInd Unknown, PwrInd Unknown, Power- Interlock-
SltSta: Status: AttnBtm- PowerFlt- MRL- CmdCplt- PresDet+ Interlock-
Gather the command line tools

- “smart-ctl” – Full device query for NVMe

smartctl 6.5 2016-04-27 r4312 [x86_64-w64-mingw32-win10] (daily-20160427)
Copyright (C) 2002-16, Bruce Allen, Christian Franke, www.smartmontools.org

=== START OF INFORMATION SECTION ===
Model Number: Samsung SSD 950 PRO 256GB
Serial Number: ...
Firmware Version: 1B0QBXX7
PCI Vendor/Subsystem ID: 0x144d
IEEE OUI Identifier: 0x002538
Controller ID: 1
Number of Namespaces: 1
Namespace 1 Size/Capacity: 256,060,514,304 [256 GB]
Namespace 1 Utilization: 117,410,267,136 [117 GB]
Namespace 1 Formatted LBA Size: 512 Local Time is: Thu Apr 28 19:32:07 2016 CEST
Firmware Updates (0x06): 3 Slots
Optional Admin Commands (0x0007): Security Format Frmw_DL
Optional NVM Commands (0x001f): Comp Wr_Unc DS_Mngmt Wr_Zero Sav/Sel_Feat
Maximum Data Transfer Size: 32 Pages
Gather the command line tools

- "nvme-cli" – send commands to an NVMe device
  - Test individual commands
  - Print out response(s)
  - Gather log entries
Time- that most precious commodity

- Optimize your access time but...

- Storage devices need time for media maintenance

- Schedule time gaps for this to occur
Priority queues for extreme conditions

- Needs more input from the system world
- Current scheme (Priority Round Robin) has its limits in extreme high IO conditions
- Per command priority?
Firmware

- Panic handler - handling crashes
- Telemetrics - reporting those crashes