SFD8 - Day 3 - Intel

@intelitcenter

#intelstorage

Optimizing Storage Solutions Using the Intel® Intelligent Storage Acceleration Library - https://software.intel.com/en-us/articles/optimizing-storage-solutions-using-the-intel-intelligent-storage-acceleration-library

*Dave Cohen, Senior Principal Engineer - Cloudification of Storage - @davecohn01

What's your workload today?
Enterprise SPs
SDI
HPC, Big Data

- new protocols optimised for data tiering
- data resident computing

"Cloudification" - the continuation of the value creation created by the internet

Evolving storage architecture landscape
Traditional
- over-provisioned, redundant networks
- replication internal to storage appliance
- no SPoF
Cloud
- flat networks, resiliency in Clos Fabric
- rack is failure domain, data centre is appliance
- direct attached storage (DAS)

these differ in how they deal with availability

[photo of Dave talking about tiering]

Hot tier
- authoritative source of data
- driven by app bandwidth
Warm tier
- fast refill for hot tier
- Read BW
- Write BW
- higher node count based on erasure code
Cold Tier
- Low $/GB
- high node count based on erasure code
- nothing ever gets deleted

- storage tiering enables the use of higher performance, more expensive storage media

the next wave of innovations will focus on addressing latency

so what does the 5th cloud wave mean?

“Breaking Through: A New class of non-volatile memory”
NVM Express and 3D XPoint Technology

3D XPoint (crosspoint)

More characteristics of memory than it does storage

Word (cache line)
Crosspoint structure - selectors allow dense packing and individual access to bits

Large memory capacity
crosspoint and scalable - memory layers can be stacked in a 3d manner

NVM breakthrough material advances - compatible switch and memory cell materials

immediately available
high performance - cell and array architecture that can switch states 1000x faster than NAND

1000x performance and 1000x endurance of NAND, 10x denser than RAM

NVMe? - the standard interface for PCIe SSDs
Architected for SSDs - eliminates HDD legacy that SATA and SAS are based upon
Works out of the box - drivers for all major OSes
efficient and scalable - streamlined protocol and command set, fewer clock cycles per IO
optimised for 3D XPoint - new storage stack with low latency to take full advantage

NVMe provides > 10x the bandwidth of SATA today
Scaling will continue with PCIe Generation 4

SSD NAND technology offers ~100X reduction in latency versus HDD

*Brian Hausauer - Low Latency networking for storage

network latency requirements driven by storage based on NV memory
network bandwidth requirements driven by
- cloud storage architecture
- storage based on NV memory
Intel response to requirements

the network latency goal fro NVMe with 3D XPoint is easier to achieve with network acceleration like RDMA
persistent memory DIMM network latency goal with as-is block or file network

protocols. new protocol needs:
- reduced number of network messages per IO and max single RTT
- no per-IO CPU interaction on the target

network bandwidth driven by cloud storage architecture
how many NVMe devices are required to saturate various network link speeds using NVMe over Fabrics protocol at the storage server home node?
- each NVMe drive is capable of 2GB/s sustained writes
- traffic pattern is 100% write, 3x replication, 4KB block IO
[table - photo]

in the cloud storage architecture, a modest number of NVMe devices can
- generate massive network bandwidth
- drive packet rates high enough to saturate network acceleration / RDMA solutions very attractive

Intel Response to requirements [refer to photo for notes]

(Bonus Slide) - [Intel_RDMA.png]
What is RDMA? A networking performance optimisation
- enables servers to communicate across a network using high-performance, low-latency, zero-copy DMA semantics
- reduces host CPU utilisation, network-related host memory bandwidth, and network latency compared to traditional networking stacks (sockets with TCP/IP)
- RDMA NIC resources (send and receive queues, doorbells, etc) are mapped directly into User or Kernel application address space, enabling OS bypass

*Roger Jeppsen - Software Architect - Storage Features in Intel Architectures Orchestrated DC

Processor Enhancements

[photo of Roger presenting]

*Nate Marushak - Enabling the storage transformation Intel ISA-L & SPDK

Address the bottleneck pendulum
- 25/50/100GbE
- Intel 3D XPoint
- RDMA
Enable the storage transformation

Storage Performance development Kit

Built on Data Plane Development Kit (DPDK)
- sw infrastructure to accelerate the packet IO to Intel CPU
Userspace Network Services (UNS)
- TCP/IP stack implemeted as polling, lock-light library, bypassing kernel bottlenecks, and enabling accessibility
Userspace NVMe, Intel Xeon / Intel Atom Processors DMA and Linux AIO drivers

- optimises back end driver performance and prevents kernel bottlenecks from forming at the back end of the IO chain
Reference Software and Example Application
- customer-relevant example app leveraging ISA-L is included, support provided on best-effort basis

SPDK
What is Provided?
- builds upon optimised DPDK technology
- Optimized UNS TCP/IP technology
- optimised storage target SW stack
- optimised persistent media SW stack
- supports linux OS

How it helps?
- avoids legacy SW bottlenecks
- removes overhead due to interrupt processing (use polling)
- removes overhead due to kernel transitions
- removes overhead due to locking
- enables greater system level performance
- enables lower system level latency

“Other names and brands may be claimed as the property of others”

Intel Intelligent Storage Acceleration Library
Algorithmic Library to address key storage market segment needs
- optimised libraries for Xeon, Atom architectures
- enhances performance for data integrity, security / encryption, data protection, deduplication and compression
- has available C language demo functions to increase library comprehension
- tested on Linux, FreeBSD, MacOS and Windows Server OS

ISA-L Functions
Performance Optimising
- Data Protection - XOR (r5), P+Q (r6), Reed-solomon Erasure Code)
- Data Integrity - CRC-T10, CRC-IEEE (802.3), CRC32-iSCSI
- Cryptographic Hashing - Multi-buffer: SHA-1, SHA-256, SHA-512, MD5
- Compression “Deflate” - IGZIP: Fast Compression
- Encryption

Takeaways - [refer to photo for notes]

*Mike Reed - Modernising the DC through SDS

Traditional Storage Challenges
Stroage Silos
- Apps (A, B, C, etc) mapped to specific appliance
- storage optimised to a specific workload

Challenges

- Limited / no sharing, interoperability
- different mgmt interfaces
- different support contracts
- limited scalability (oversubscription, forklift upgrades)

SDS [refer photo for notes]

Control / Mgmt (“SDS Controller”) is key

SDS Addresses End User Pain Points [refer photo for notes]

Coppered [photo for notes] (Copperhead)
An open source SDS controller that discovers, pools, and automates the mgmt of a heterogeneous storage ecosystem

Cinder vs CoprHD Comparison [refer to photo for table]

“Cinder is broadly adopted with a mature community; CoprHD is ~2 yrs ahead for enterprise readiness”

Enabling requirements
- automated control and mgmt
- ecosystem support

[Lunch and Q&A]