Jonathan Stern
The future of storage - Storage Performance Development Kit (SPDK)

Hyper-scale cloud
Hyper-convergence
Non-volatile Memory

“catalysing” the shift that’s happening here

The rise of the big clouds (the super 7 - baud, alibaba, google, Facebook, microsoft, maybe Apple?)

[Intel_Traditional_Storage_rev_Predictions.png]

50% already leveraging SSDs

Tape isn’t dead, nor is disk, but there are physics limitations

Architecture optimisation: divergent metrics
$/GB - storage efficiency
$/IOPS - data path simplicity

“Adding power makes you faster on the straights. Subtracting weight makes you faster everywhere. Simplify, then add lightness” - Colin Chapman, founder of Lotus F1

CPU cycles per IO

NVM Express Driver throughput scalability
- Systems with multiple NVMe SSDs capable of millions of I/O per second
- results in many cores of software overhead with kernel-based interrupt-driven driver model
- SPDK enables: more CPU cycles for storage services, lower I/O latency

Almost no premium now on capacity to do IOPS with a system

What matters besides IOPS?

What does the future hold?
NAND comes with challenges

“In the immortal words of Olivia Newton-John - Let’s get physical!”
Intel SSD DC P3700 Series:
450K Random Read 4K IOPS
256 4K IOPS per 1MB/s
128 MB/s per Gbps
450K IOPS = around 1750MB/s = around 15Gbps
*24 drives
= approx. 42000MB/s = 360Gbps
Intel SSD DC P3700 series:
175K random write 4K IOPS
175K Random write 4K IOPS
175K IOPS = around 685MB/s = around 5Gbps
*24 drives
= approx. 16440MB/s = approx. 120Gbps


What are the forces that shape how people use this plethora of IOPS?
People find ways of consuming resources you provide to them

pmem - persistent bugs too

hyper-scale cloud
disaggregation - NVMe over Fabrics
You can now scale these quantities independently
cost or “power-optimise” high dollar enclosures

hyper-convergence - the “old cloud model”
commodity hardware
VM I/O path - SPDK virtualisation

SPDK VM I/O Efficiency
Serve SPDK storage to local VMs
- NVMe ephemeral storage
- SPDK-based 3rd party storage services
Leverage existing infrastructure for
- QEMU vhost-scsi
- QEMU/DPDK vhost-net user

Features and benefits
- high performance storage virtualisation
- reduced VM exit
- lower latency
- increased VM density
- reduced tail latencies
- higher throughput

Ali Cloud sees ~300% improvement in IOPS and latency using SPDK

VM Ephemeral Storage
- Improves Storage virtualisation
- works with KVM/QEMU
- 6x efficiency vs kernel host
- 10x efficiency vs QEMU virtuo
- increased VM density

SPDK and NVMe over Fabrics
VM Remote Storage
- enable disaggregation and migration of VMs using remote storage
- improves storage virtualisation and flexibility
- works with KVM/QEMU

NVMe over Fabrics

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilises NVM Express (NVMe) Polled Mode Driver</td>
<td>Reduced overhead per NVMe I/O</td>
</tr>
<tr>
<td>RDMA Queue Pair Polling</td>
<td>No interrupt overhead</td>
</tr>
<tr>
<td>Connections pinned to CPU cores</td>
<td>No synchronisation overhead</td>
</tr>
</tbody>
</table>

NVMe-oF Key Takeaways
- preserves the latency-optimised NVMe protocol through network hops
- potentially radically efficient, depending on implementation
- actually fabric agnostic: InfinBand, RDMA, TCP/IP, FC … all ok!
- underlying protocol for existing and emerging technologies
- using SPDK, can integrate NVMe and NVMe-oF directly into applications

VM I/O Efficiency Key Takeaways
- huge improvement in latency for VM workloads
- application-level sees 3-4X performance gains
- application unmodified: it’s all under the covers
- virtuous cycle with VM density
- fully compatible with NVMe-oF!

*Tony Luck - Intel Resource Director Technology (RDT) for Storage
“Data is the fuel of the new tech revolution”
keep operating expenses low - improve workload utilisation
Enterprise – deliver business value as fast as possible with lowest total cost of ownership
Communication Infrastructure – transition to software-focused architectures
Communications Service Providers – reduce network cost of operation and enable business innovation to drive incremental revenue
Cloud Service Providers – scale and maintain SLAs in face of shifting customer demands while optimizing cost of service delivery

Software Defined Infrastructure
- SDI Enables consolidation of a variety of workloads with different characteristics
- Latency, Throughput, Compute
- Supporting SDI requires a hardware architecture that supports Quality of Service
- SDS – Emerging requirements

Shared resource contention likely to occur when executing a variety of workloads in parallel
Emerging Challenge - Shared Resource Contention
- last level cache is shared to make best use of the resources in the platform
- however certain types of applications can cause noise and slow down others
- applications streaming in nature can cause excessive LLC evictions

Solution: RDT

RDT Attributes
Flexible
- interfaces are based on Model-specific Registers (MSRs)
- can be used by Operating systems, hypervisors, or privileged advanced software
- example usages span the cloud DC (prioritising VMs), communications, etc
Architectural
- MSR interfaces standardised
- consistent support across generations - once written, software can be reused
all parameters which can change are enumerated via CPUID
Dynamic
- settings can be updated at any time
- enables advanced usages such as closed-loop controllers

RDT Potential Storage Usages
Bandwidth
- prioritise important storage apps
- isolate them from compute-hungry apps
- optimise at a global cross-node level and enable SDS
Latencies
- isolate resources across usages / apps to control tail latencies

Where to learn about it?
Intel.com - RDT page
software.intel.com - RDT blogs
Intel software developer’s manual - volume 3B, part 2

Links?
Intel RDT landing page
Intel SDM
NPG Product literature

Summary
- consolidation of a variety of workloads can cause shared resource causing potentially impacting performance
- Intel RDT introduces hardware hooks to monitor & allocate resources on a per thread/core/app basis
- policy-driven smarter workload placement is now possible
- results from variety of use cases prove Intel RDT effectiveness across VMs, containers, comms apps and other usages
- the full ecosystem is being enabled to take full advantage of Intel RDT