Cloud Computing with FPGA-based NVMe SSDs

Bharadwaj Pudipeddi, CTO NVXL

Choice of NVMe Controllers

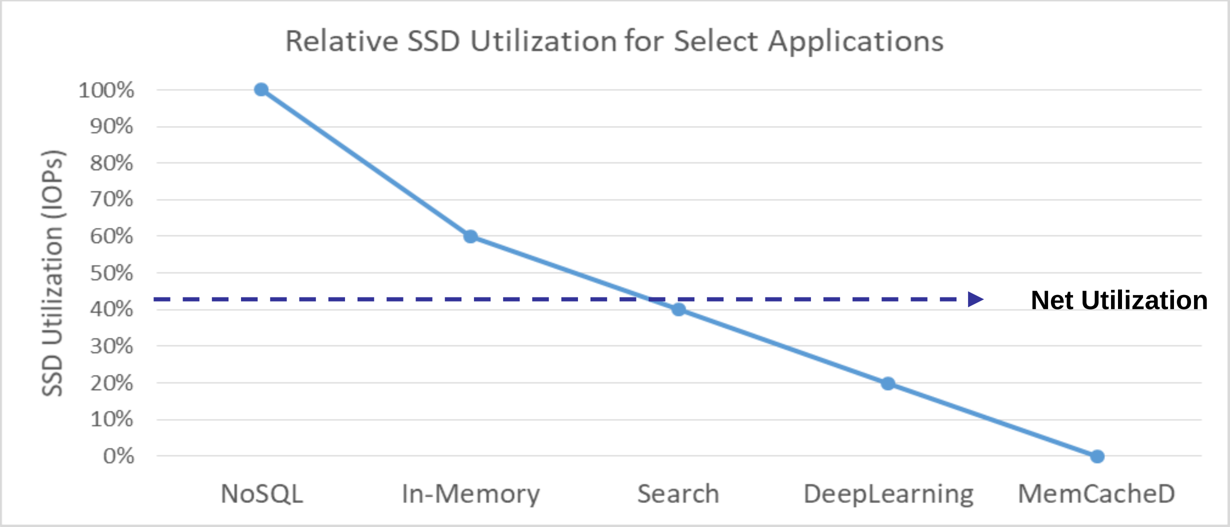
- **ASIC NVMe**: Fully off-loaded, consistent performance, M.2 or U.2 form factor
- **ASIC OpenChannel**: Host-controlled, partially off-loaded, low cost, M.2 or U.2 form factor
- **FPGA NVMe**: Fully off-loaded, consistent performance, *multi-function*, U.2 or AIC form factor

A unique feature of FPGA controllers is **multi-functionality** – the ability to reconfigure for both storage *and acceleration*.

Utilization Factor in Cloud Servers

When applications cannot use all the IOPs nor the capacity,

Is fixed function SSD the right choice in all cases?

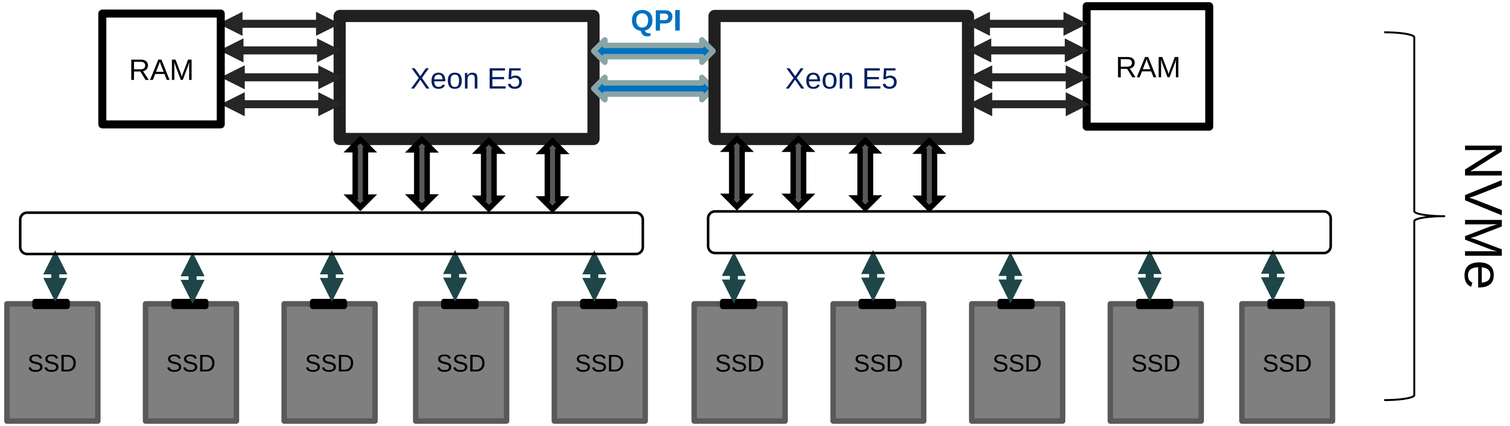


Flash Memory Summit

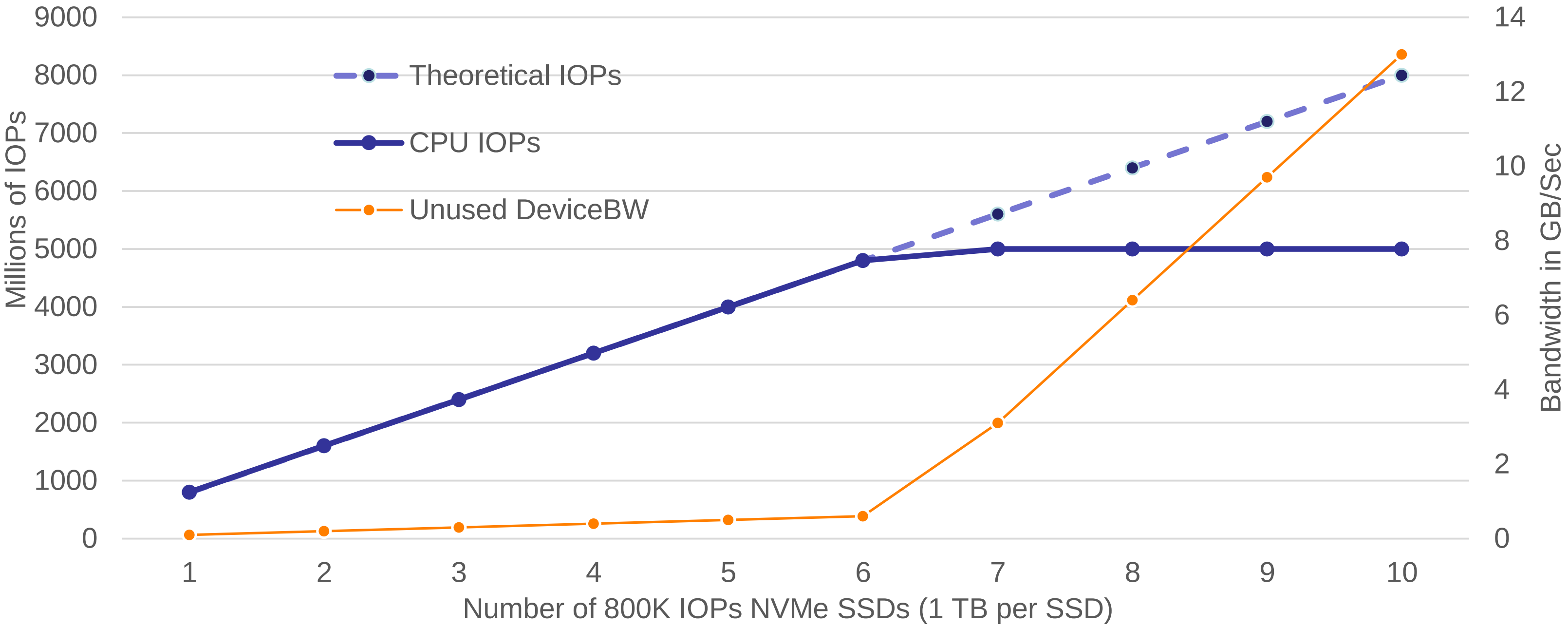
NVMe Performance Scalability

A large capacity (say 24TB) server requires many SSDs – anywhere from 3 to 24.

Here is a server with 10 to 12 NVMe drives:



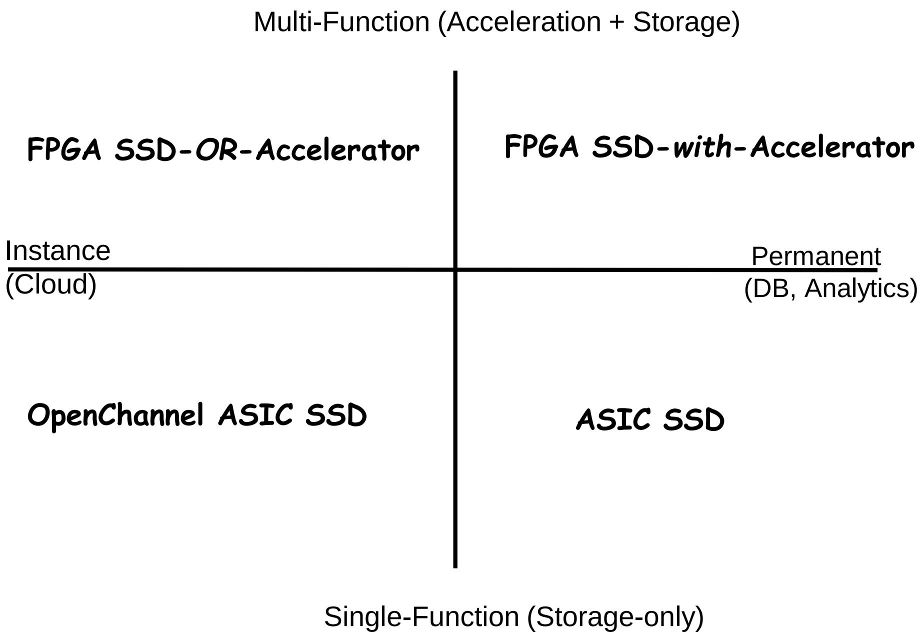
Problem: System-level IOPs constrained by CPU



NVMe Application Spectrum

In the simplified 2x2, ASICs fit all quadrants for *continuous, dedicated* usage.

FPGAs fit variable workloads in cloud and data analytics where acceleration and storage are often ***both*** in need.



Multi-Function (Acceleration + Storage)

FPGA SSD-OR-Accelerator

FPGA SSD-with-Accelerator

Instance (Cloud)

Permanent (DB, Analytics)

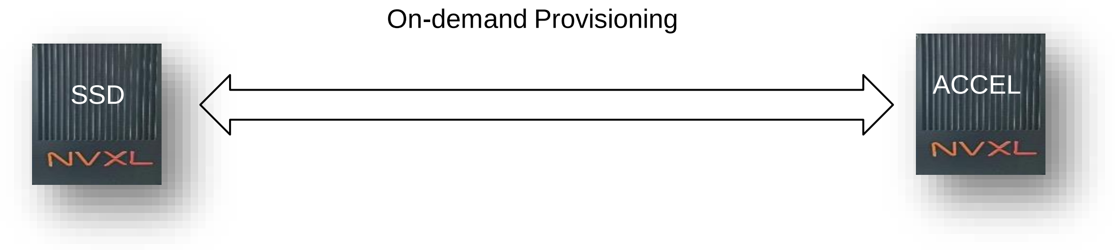
OpenChannel ASIC SSD

ASIC SSD

Single-Function (Storage-only)

FPGA-based Controller

- Reconfigurable multi-function device (U.2 form factor)
- RTL-optimized acceleration and performance
- SCM-class latencies (sub-6 uSec) by using high bandwidth memory designed for accelerator (SuperRAM)
- Compute acceleration ranging from 700 GFLOPs to 8 TFLOPs
- On-demand provisioning by Software Abstraction Layer (NAL)

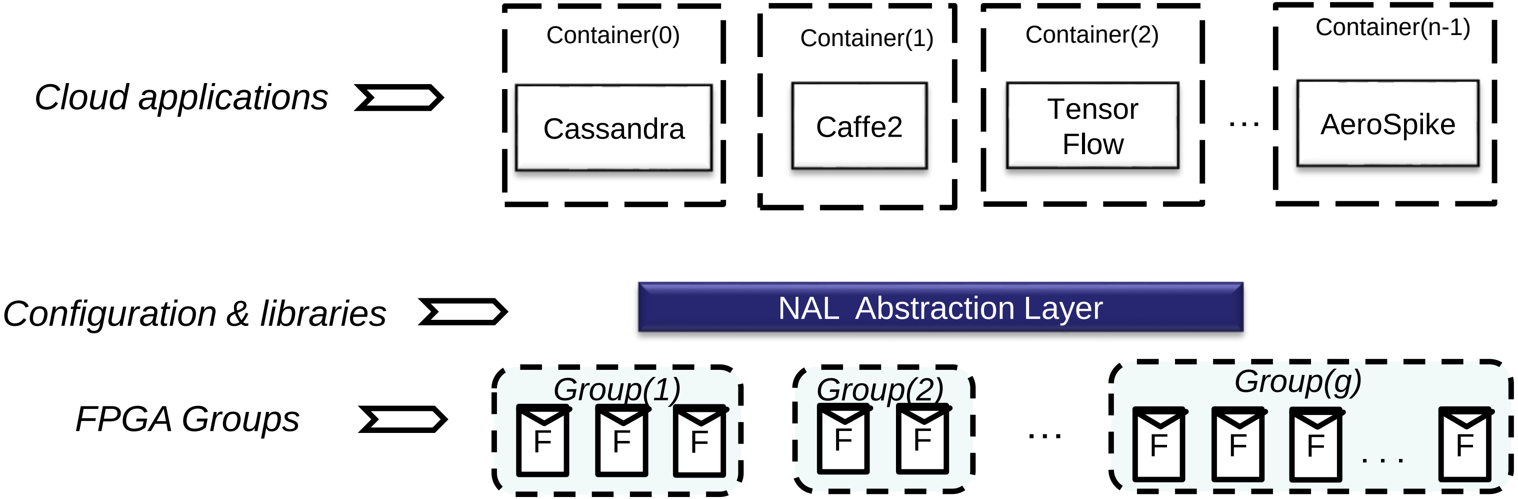


On-demand Provisioning

SSD

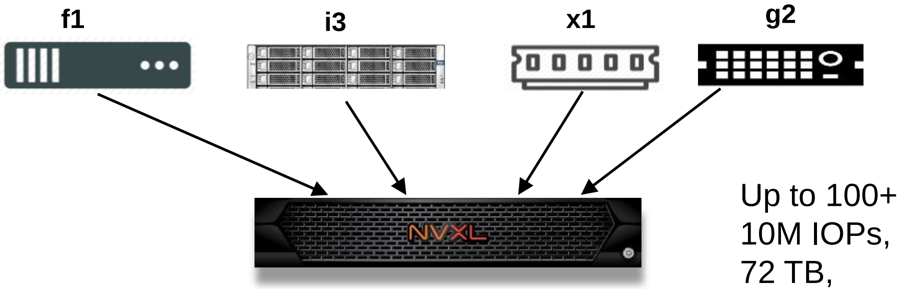
ACCEL

Scaling in the box: Grouping sets of FPGAs for Tasks



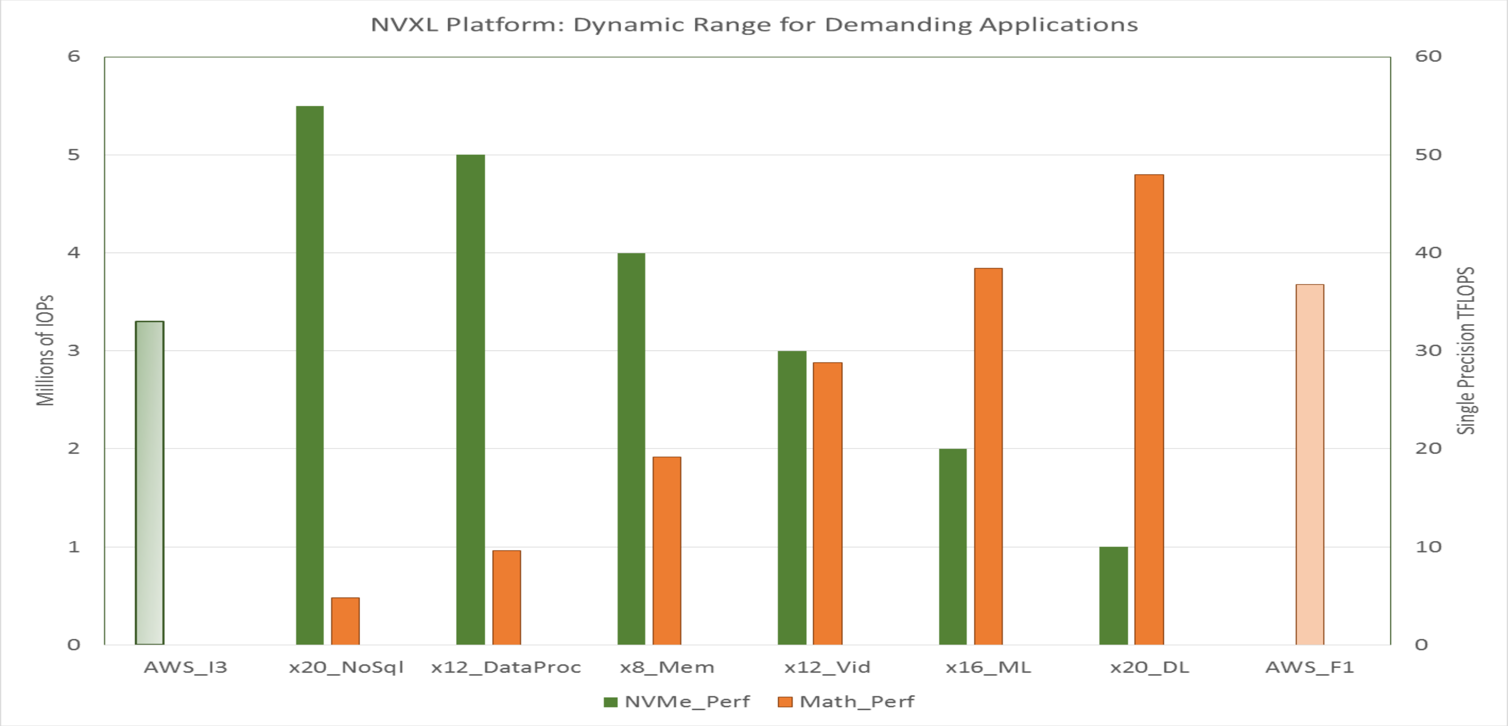
Cloud SKU Compaction from AWS

- Four popular AWS performance SKUs: f1 (fpga), i3 (NVMe), x1 (in-memory), g2 (GPU)
- With NVXL provisioning, all SKUs can be consolidated into 1 SKU
- Better utilization, easier management



Up to 100+ TFLOPs,
10M IOPs,
72 TB,
SCM-class 1TB of NVMe memory
(1W1Q < 4 uSec latency)

Different modes against AWS i3 and f1 SKUs



Application Benchmarks: Aerospike

- Aerospike is an **In-Memory** Key-Value Database *designed* for DRAM and **Flash**
- Hybrid architecture keeps index in DRAM and records in SSD via Write Buffer
- Write buffer is flushed when full for large updates to SSD for even wear

AeroSpike is ideally suited for acceleration with NVXL platform – by using module SuperRAM (each with 24GB of memory) for Write Buffers and Record caching

Certain Aerospike features can also be accelerated by FPGAs.

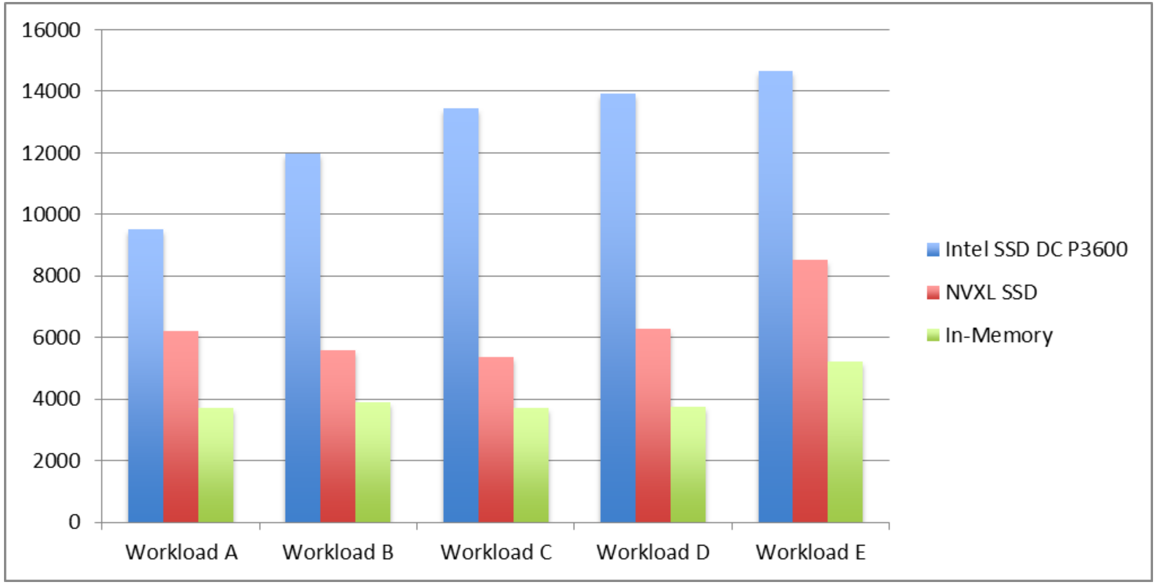
Examples: Group Bys, and Joins.

Benchmark: YCSB (Yahoo Cloud Server Benchmarks)

Against: Full In-Memory, Intel SSD, and NVXL SuperRAM.

YCSB Runtime (Lower is better)

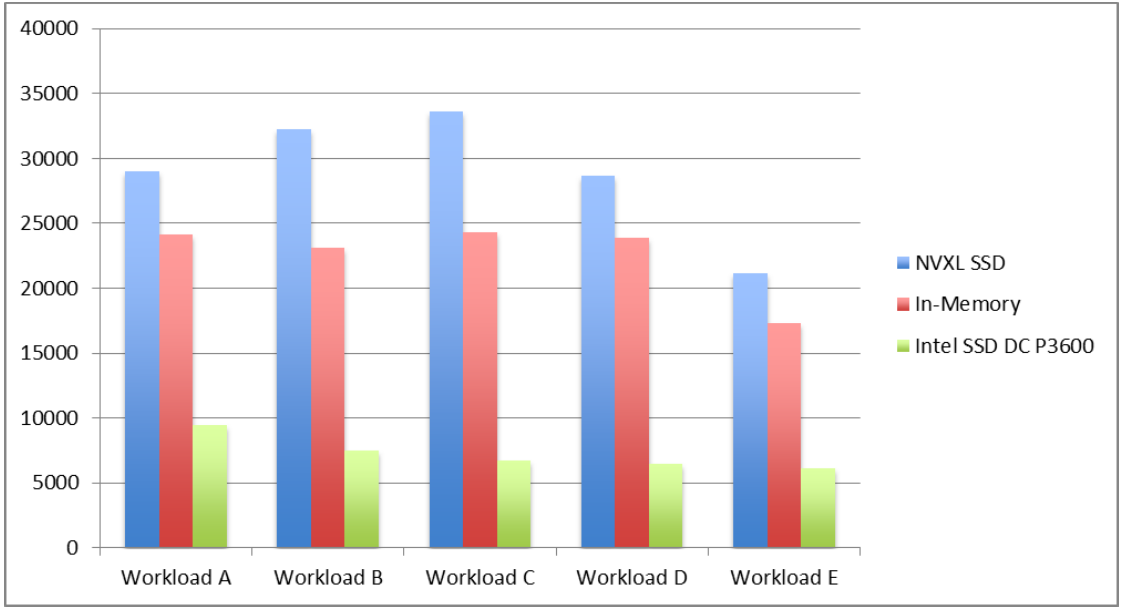
Workload	Read/Write Mix
A	50/50
B	95/5
C	100/0
D	Read Latest
E	RMW



NVXL “SSD” cuts latency by half over pure SSD due to a giant “smart” cache (“borrowed” from accelerator mode)

YCSB Throughput (Higher is better)

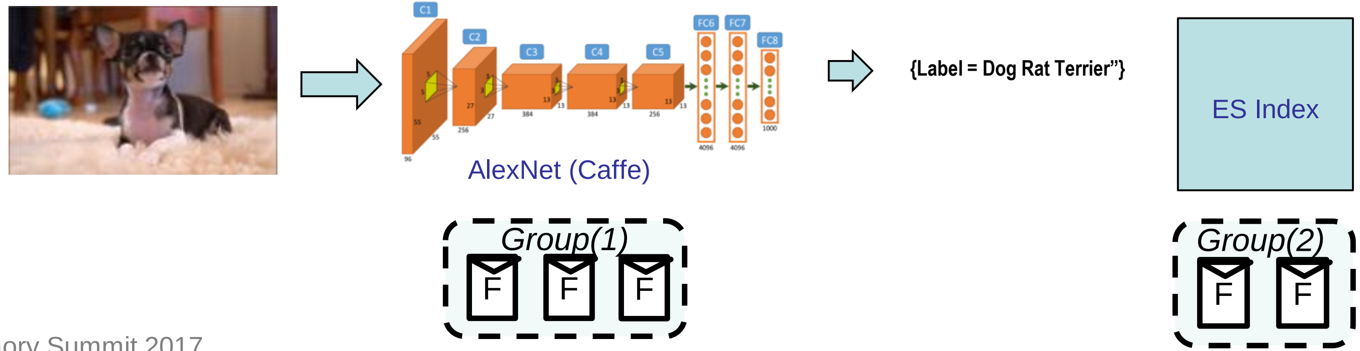
Workload	Read/Write Mix
A	50/50
B	95/5
C	100/0
D	Read Latest
E	RMW



With just two modules, NVXL SSD has even better throughput than in-memory and more predictable performance.

Searching Images with Elastic Search

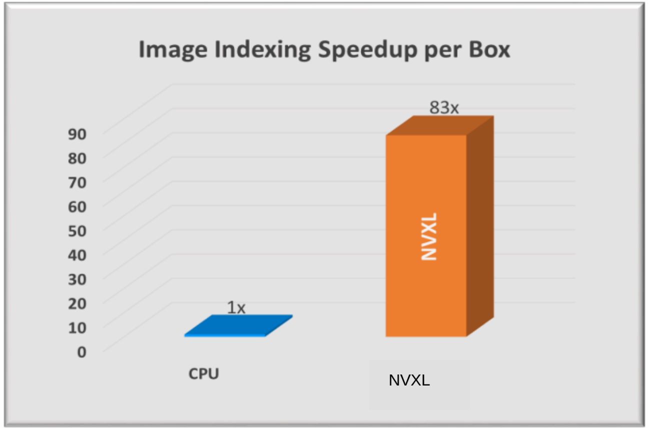
- ElasticSearch (ES) is an Apache Lucene-based distributed string search engine using schema-free JSON documents (extremely popular in retail and services)
- Application: Connect a Deep Learning inference model such as Caffe/AlexNet to ES for ***automatic tagging*** of anonymous images
- Note: indexing is normally infrequent and bursty
- During indexing, AlexNet is run on provisioned DL FPGA group (1) while ES uses Group(2)
- After indexing, Group(1) is dismantled and recirculated for other applications

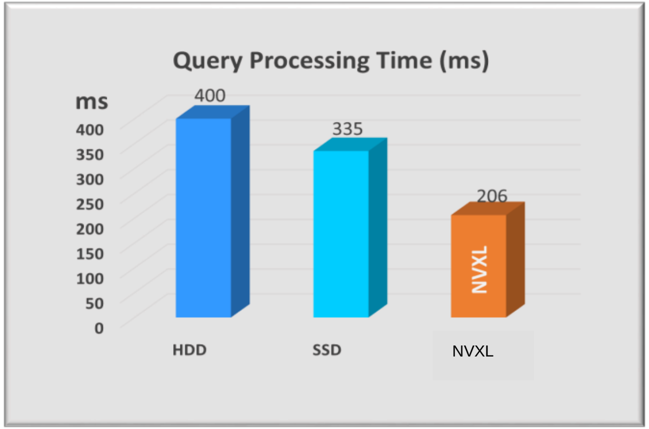


ElasticSearch Results

- Indexing speedup of 83x
- Fully optimized: 1000x speedup!
- Query processing: 40% faster latency
- TCO benefits: Low power, high utilization

Similar approach can also be used for search-by-image and ES or in-memory DB for latency about sub-5 mSec (5x faster than CPU).





Conclusion

- We present a new class of cloud server using multi-function devices in NVMe U.2 form factor
- The device is designed for storage and acceleration
- NVMe Storage benefits from SCM-class latencies due to richer memory resources
- Software layer provides grouping and reconfigurability features enabling a wide range of cloud applications from NoSQL to Deep Learning
- TCO benefits from this server cut data center costs for performance instances by 50% through increased utilization and SKU consolidation