Storage in the New Age of AI/ML

Young Paik

Sr Director Product Planning

Samsung

May 21, 2019

Legal Disclaimer

This presentation is intended to provide information concerning SSD and memory industry. We do our best to make sure that information presented is accurate and fully up-to-date. However, the presentation may be subject to technical inaccuracies, information that is not up-to-date or typographical errors. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of information provided on this presentation.

The information in this presentation or accompanying oral statements may include forward-looking statements. These forward-looking statements include all matters that are not historical facts, statements regarding the Samsung Electronics' intentions, beliefs or current expectations concerning, among other things, market prospects, growth, strategies, and the industry in which Samsung operates. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements contained in this presentation or in the accompanying oral statements. In addition, even if the information contained herein or the oral statements are shown to be accurate, those developments may not be indicative developments in future periods.

Speaker Disclaimer

Sometimes accuracy is the enemy of the truth

AI/ML Workflow – So Simple

Logs

DBs

Real-time
streams

Images

Video

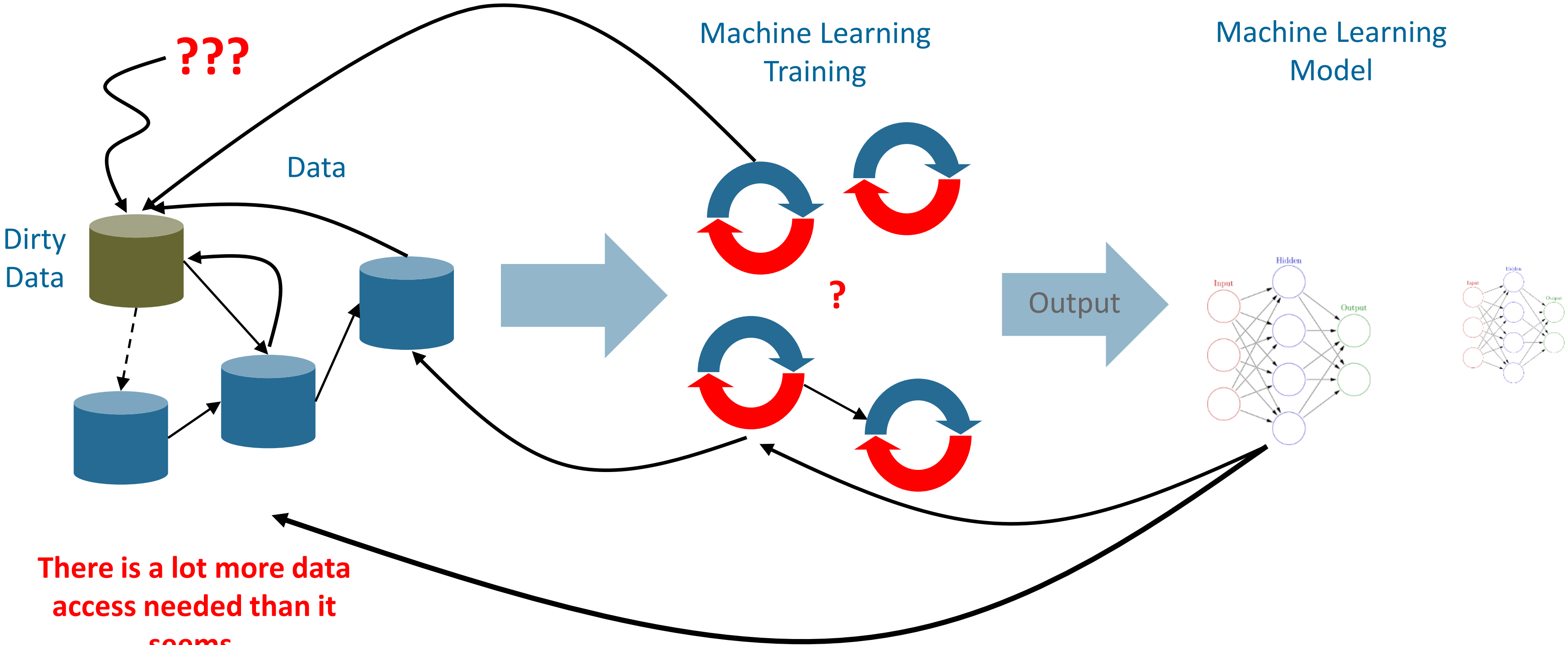
Audio

IoT

Genetic Data

But is it really this simple?

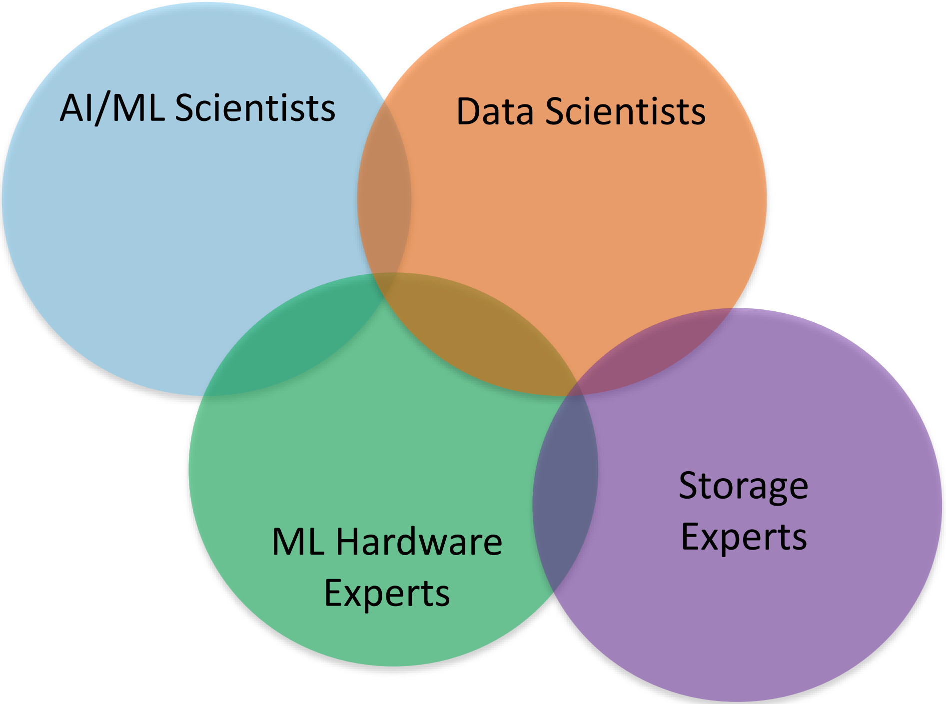
AI/ML Workflow – It's Never Easy



There is a lot more data access needed than it seems.

Disparate Groups of Experts

Skill sets are highly specialized, often without overlapping skill sets



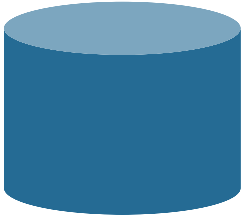
Artificial Intelligence Workflow – Major Challenges

1) Hard to parallelize ML
< ~150 compute nodes per model

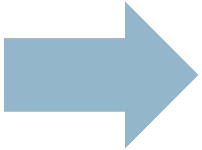
2) Servers with multiple GPUs have PCIe limitations
Often very expensive

3) Models are growing quickly
Up to 2 TB
Can be shrunk but initial training can be big

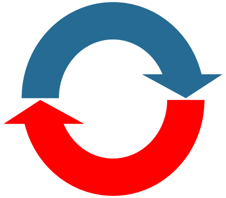
Data

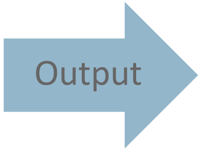


4) High network bandwidth
~ 1 GB/GPU (up to 16 GB/host)
Storage B/W much higher

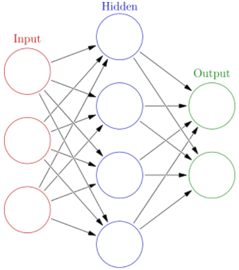


Machine Learning Training





Machine Learning Model



5) Data must be pre-processed
May require accelerators (GPU/FPGA/ASICs)

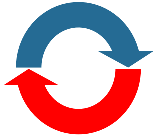
What does preprocessing look like?

Artificial Intelligence Workflow – Facial Recognition

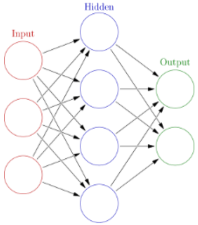
Images

Facial Recognition Training

Facial Recognition Model



Output



- Photo by rawpixel.com from Pexels

Deep Learning models need the same facial form

AI/ML Training servers may cost up to $400K

Facial Recognition Example of Preprocessing

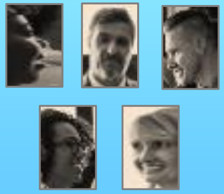
1) Find faces

P
by rawpixel.com from P
exels

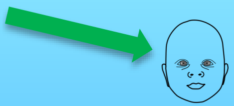
Photo by rawpixel.com from Pexels

2) Extract faces

3) Resize image and color

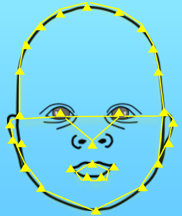


4) Rotate face



(My sincere apologies to the model for this rendering)

5) Extract features



To recognize the identity of a face, you must first isolate every face.

Training must work on individual faces

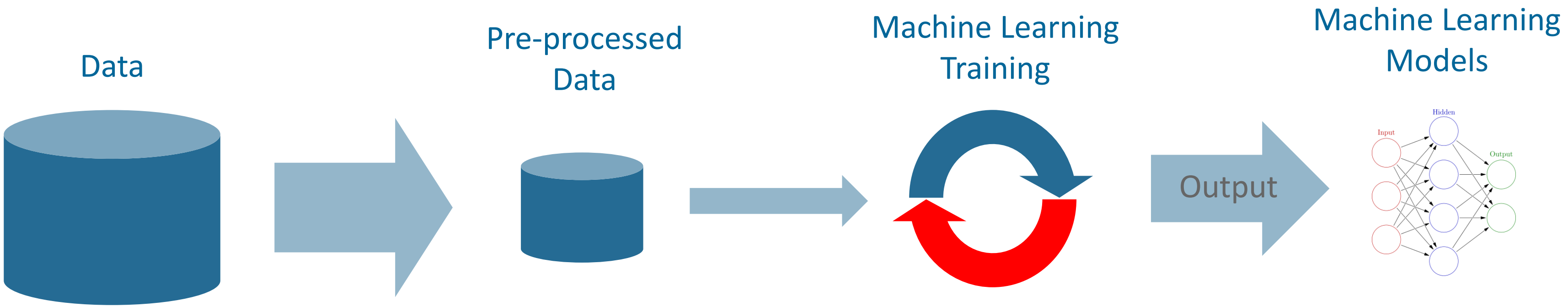
Images must conform to the same pixel and color resolution

Face must be front (there are algorithms that do this)

You can now extract the facial features and begin the training

All of this is parallelizable and does not need to be done on the training server

Artificial Intelligence Workflow – Add Preprocessing



More Complicated Issues

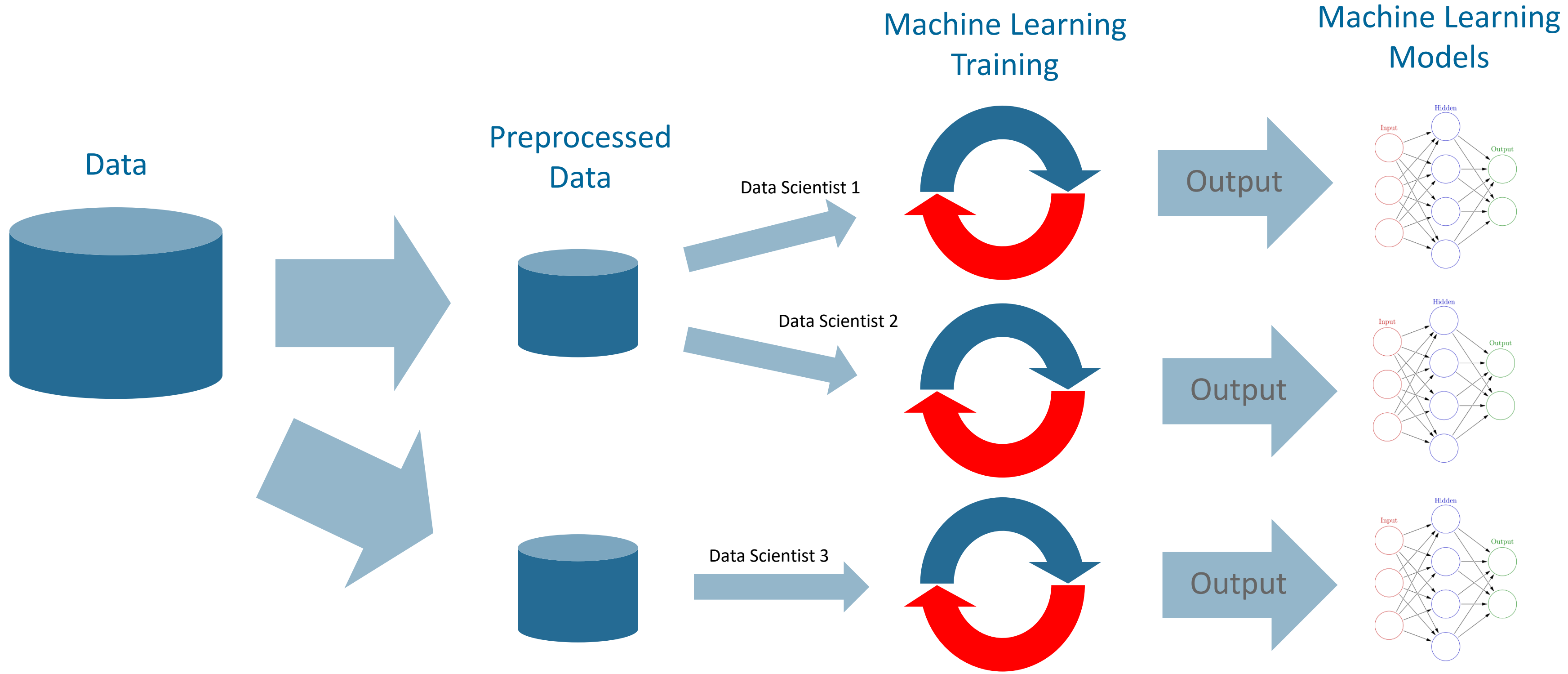
Multiple AI Scientists

Improved data processing

Dealing with long training times

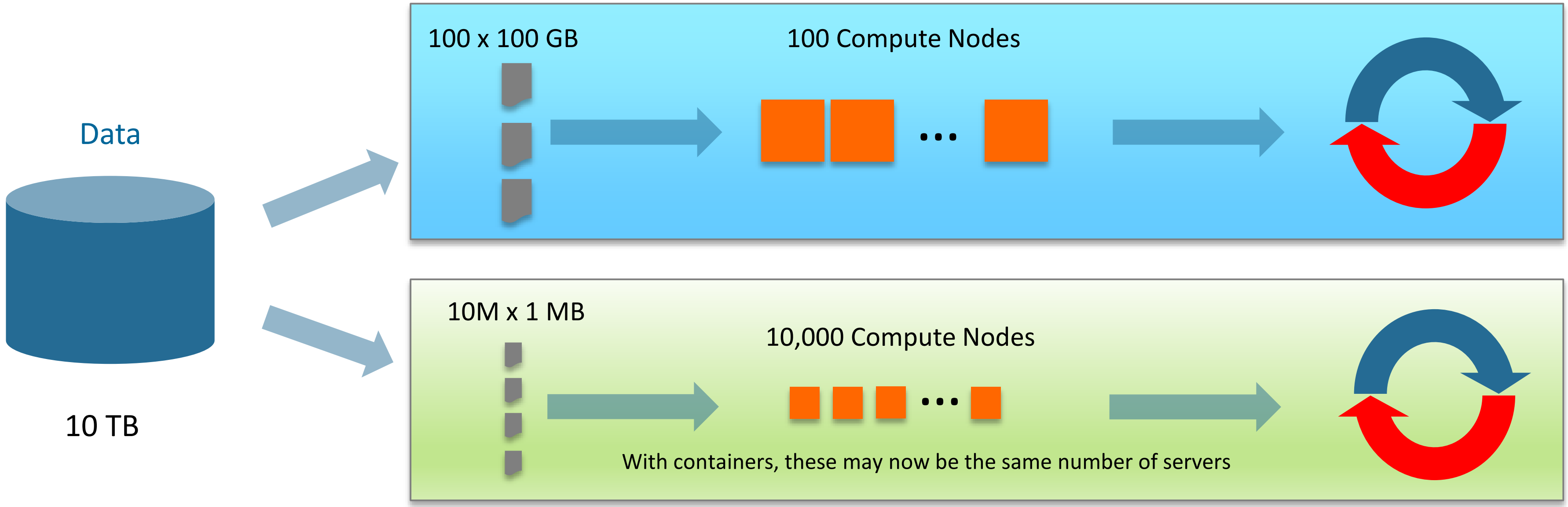
Multiple Data Scientists

Data scientist 1 and 2 want the same features, but different models
Data scientist 3 is trying a new experiment and must start from raw data



Dealing With Long Training Times

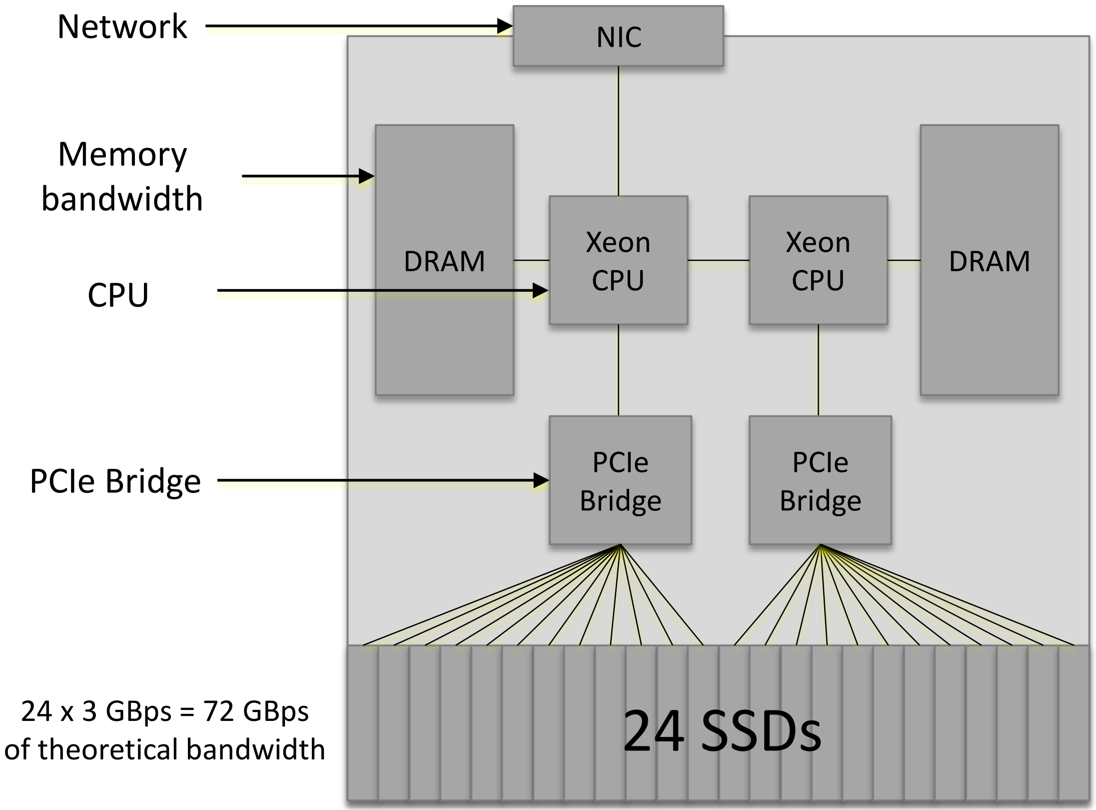
Training times may take weeks.
How can we deal with changes in workload dictated by changing priority?



Challenges:

- Minimum size for jobs (not all jobs can be shrunk)
- Scheduling is huge → Kubernetes
- Jobs are not always parallelizable (database joins)

Data Flow Limits of Modern Storage



Modern SSDs are limited by server architecture

Samsung has looked into 2 different technologies:

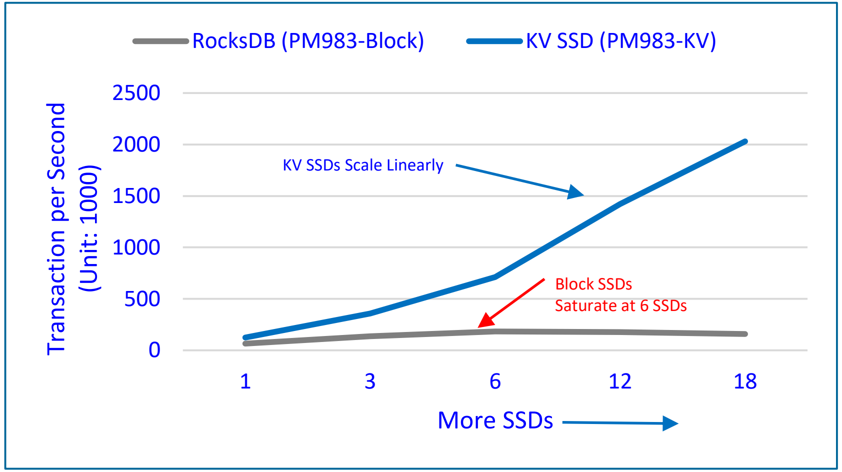
- KV SSD
- SmartSSD

KV SSD - Motivation

KV API now a SNIA Specification

https://www.snia.org/tech_activities/standards/curr_standards/kvsapi

	Block SSD	KV SSD
CPU	Overloaded with block and compaction	Freed for other tasks
Scalability	Limited to 4-6 SSDs/host	Linear performance with 18+ SSDs/host
Disk utilization	Must leave room for compaction	GC managed internally
SSD Lifetime	High WAF	Low WAF leads to greatly improved SSD lifetime

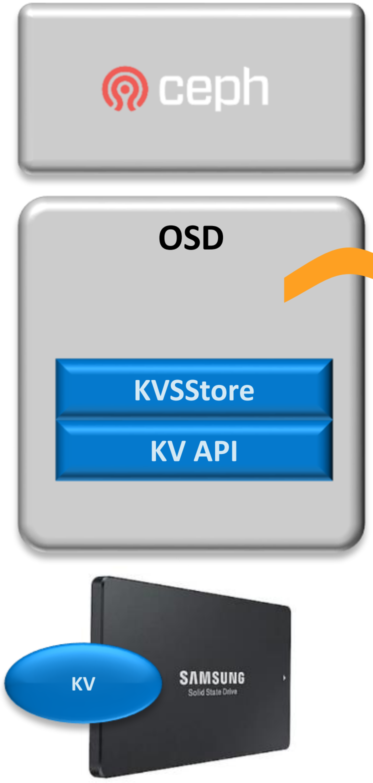


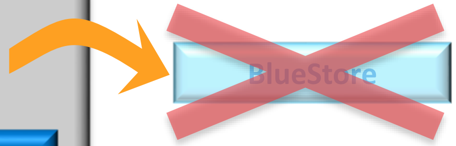
* Testing was done on a server with 2 x Intel Xeon E5-2600 v5 servers with 384 GB of DRAM, and 18 PM983 (in block or KV mode) SSDs
** Workload: 4KB uniform random writes

Main Use Cases:

- Object storage
- NoSQL databases

KV SSD – Direct Use on Ceph

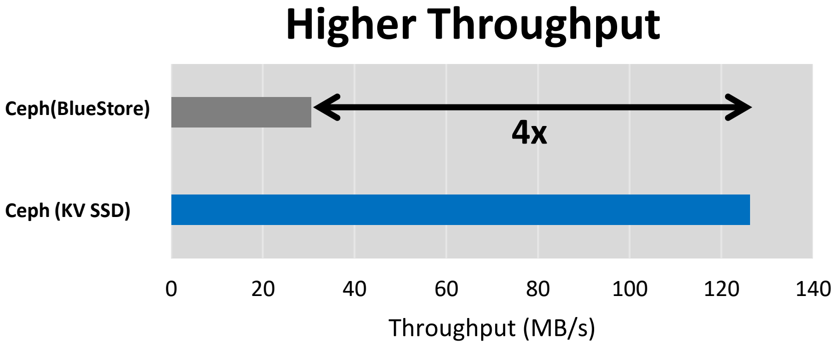




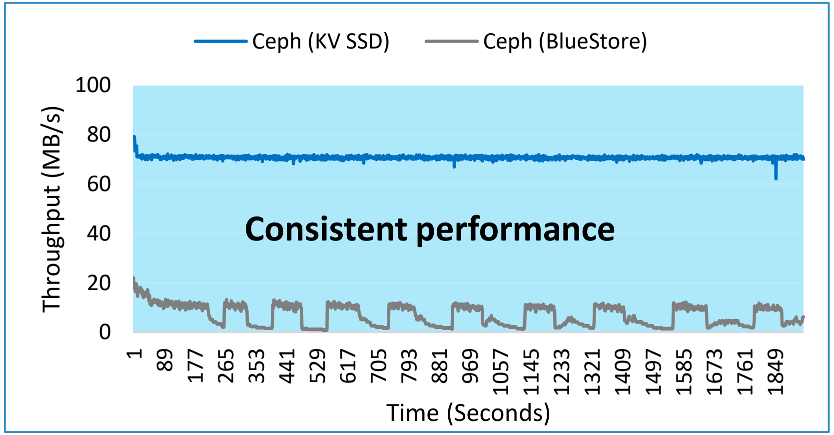
KVSStore uses the newly Open Sourced KV API to access the KV SSD

*https://github.com/OpenMPDK/KVSSD

Biggest challenge is that this requires a change in software.

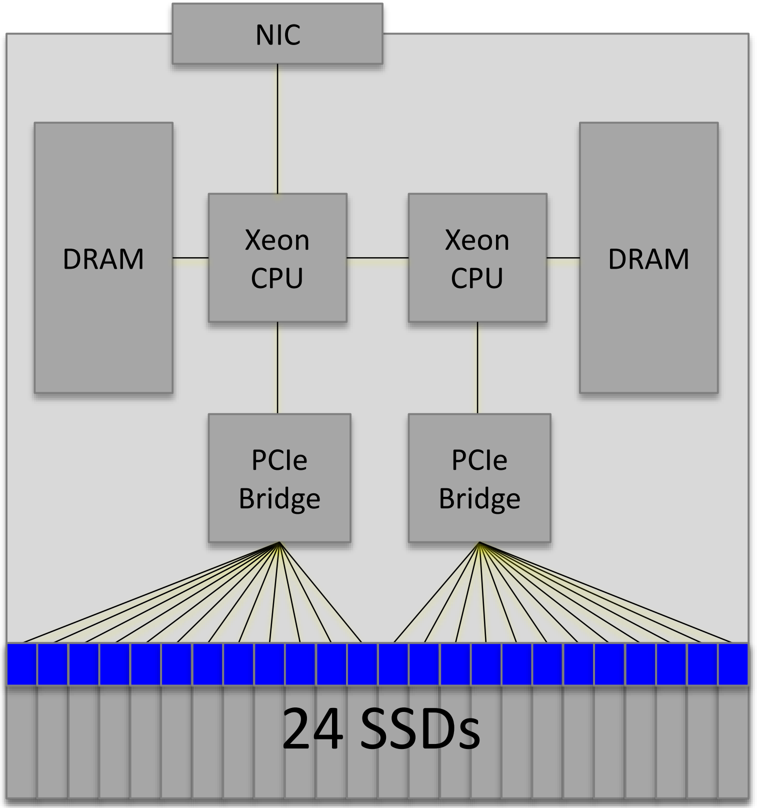


* 4096 block write Default (Sharded), 8 clients 2 OSDs- queue depth 128
* Testing was done on two servers with 2 x Intel Xeon E5-2695 v4 CPUs with 128 GB of DRAM, and a PM983 (in block or KV mode) SSD with 40 GbE



* 4096 block write Default (Sharded), 1 client 1 OSD - queue depth 128
* Testing was done on a server with 2 x Intel Xeon E5-2695 v4 CPUs with 128 GB of DRAM, and a PM983 (in block or KV mode) SSD with 40 GbE

SmartSSD-based Server Architecture



SmartSSDs process data in-storage

Allows:
- Pre-filtering
- On-disk transcoding
- Compression
- …

Challenges:
- Encryption
- RAID/Erasure Coding
- New programming model

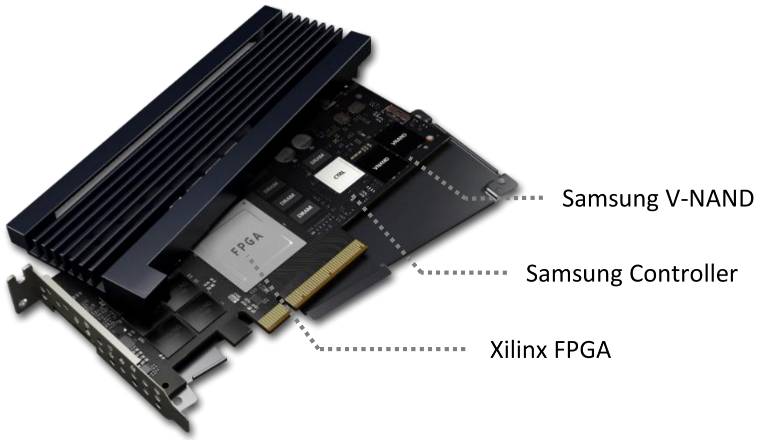
Compute occurs on storage
Parallel scans at full speed of SSDs
CPUs freed for additional work

SmartSSD

SmartSSD PM983F announced at Samsung Tech Day 2018

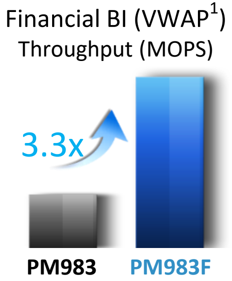
PM983F AIC

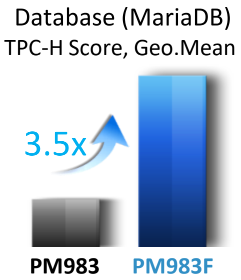
- SmartSSD PCIe add-in card
- Shown successfully integrated with Bigstream
- Several data-intensive workloads easily ported



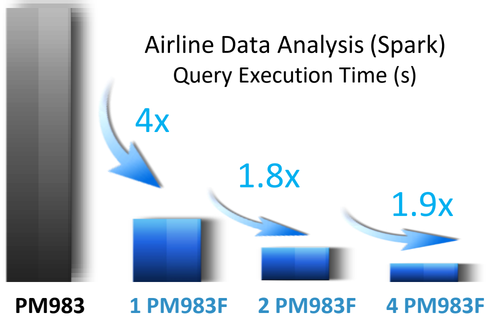
PoC Results

- For I/O-bound workloads, SmartSSD showed 3x to 4x better performance with scalability





* VWAP: Volume Weighted Average Price



New Technologies Not Covered

Technology	Description	Pros	Cons
Nvidia GPUDirect	GPUs can directly access another PCIe device	Bypasses CPU and system memory	Some people use system memory as a cache
NVMe over Fabric	Allows for very low latency to network-attached storage with RDMA latencies	Gives performance similar to direct-attach	Requires very solid network coordination
SmartNICs	These NICs have CPU offload facilities. Many have the ability to handle Reed-Solomon.	Low latency at a much lower price point.	Still very new


Thank You
Young.Paik@Samsung.com
COLLABORATE. INNOVATE. GROW.
SAMSUNG

How Important Is It To Fix Dirty Data?

Scenario: One healthcare insurance company was looking at data on charges for treatments.

We tested by looking at diseases by code and tried to guess what the disease was.

Disease Code 1

Average age:	63
Gender	
Male	33%
Female	66%
Unspecified	1%

Diagnosis: Osteoperosis

Disease Code 2

Average age:	47
Gender	
Male	98%
Female	0.5%
Unspecified	1.5%

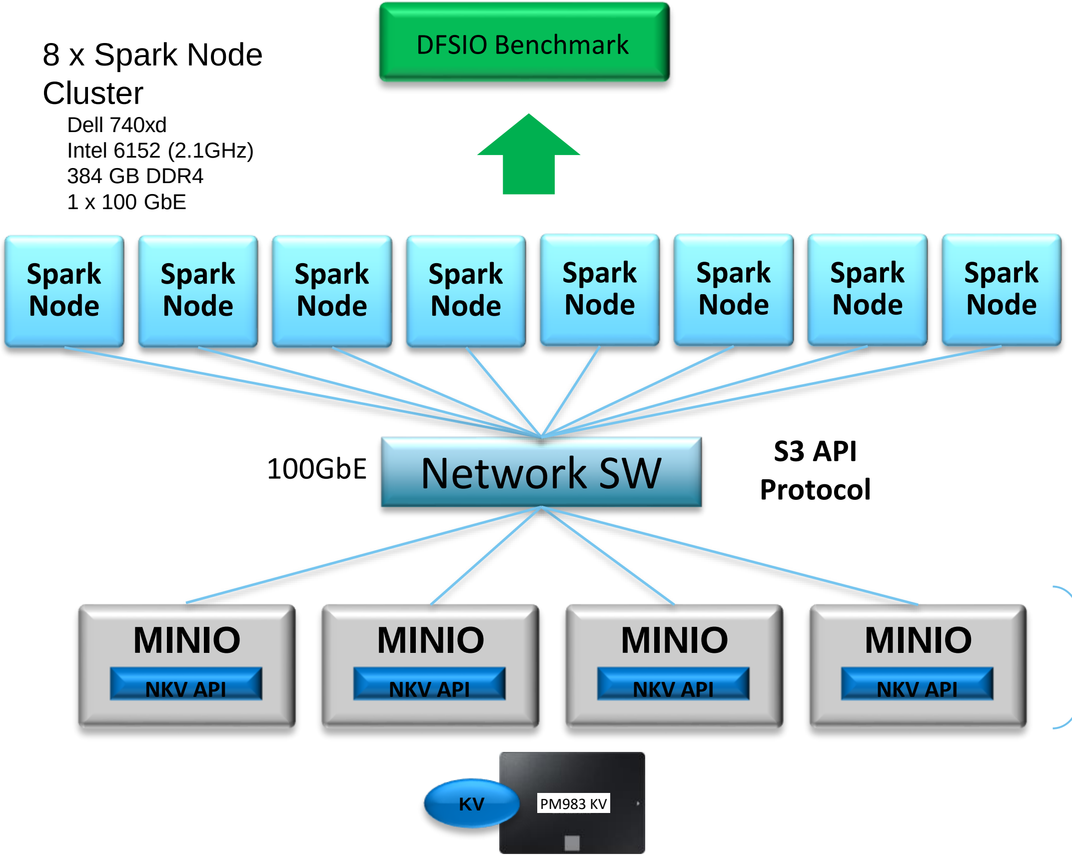
Diagnosis: ???

Moral to the story: **It is important to thoroughly process data.**

This requires much more storage I/O than people think.

MINIO + KV SSD Object Storage Performance

8 x Spark Node Cluster
- Dell 740xd
- Intel 6152 (2.1GHz)
- 384 GB DDR4
- 1 x 100 GbE



4 x MINIO + KV SSD Cluster
- 2 x Intel 6152 (2.1 GHz)
- 12 x 4 TB KV SSD
- 1 x 100 GbE
- 384 GB DDR4 (2400 MHz)
- 12 + 4 Erasure Code

Minio Bandwidth on DFSIO on Spark with 4 Nodes

* Performance tests were run with cache enabled for directory listing