STORAGE DEVELOPER CONFERENCE
SDC 23
BY Developers FOR Developers

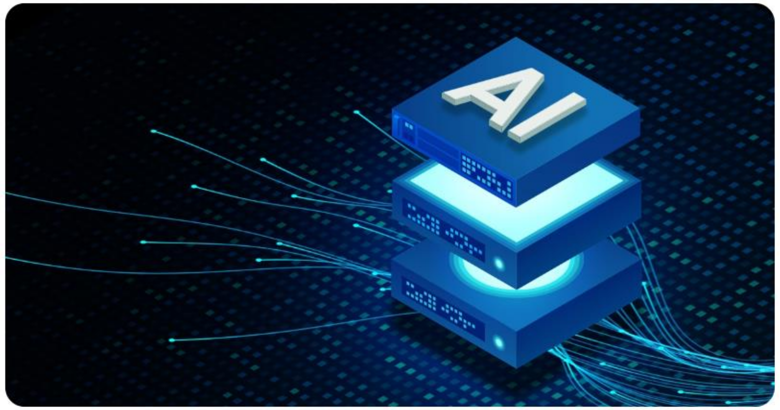
Benchmarking Storage with AI Workloads

Presented by

Devasena Inupakutika, Charles Lofton, Bridget Davis
Samsung Semiconductor Inc.

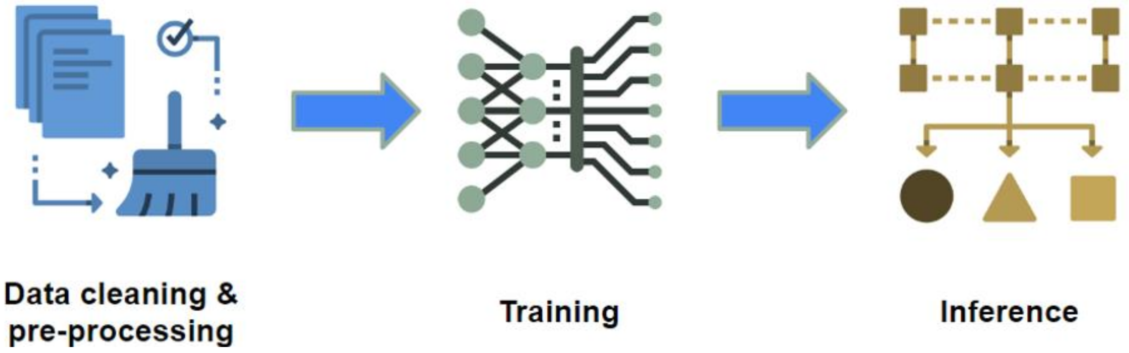
Motivation

- Growing production datasets: 10s, 100s of petabytes
- Samsung’s datacenter storage and memory products
- Research involving the impact of storage on AI/ ML pipelines is limited
- How to showcase Samsung datacenter product’s impact to real world workloads?



Introduction

- Benchmarking essential to evaluating storage systems:
  - Storage needs for large machine learning datasets are growing
- Evaluating storage for AI workloads is challenging
  - Real-world AI training requires specialized hardware
  - System resources stressed by AI application
- Do AI workloads benefit from high performance storage systems?
- Is there a realistic method to showcase high performance storage for AI workloads?
- Can the test methods be easily implemented and reproducible?



COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC23

Introduction

- Benchmark datasets are smaller whereas data is the moving force of AI algorithms
- Real-world production workloads demands huge data (both for training and generation during streaming)
- Empirical study to understand how AI workloads utilize storage devices through I/O patterns

AI Workloads I/O Characterization

- Better understanding of AI I/O profiles
- Provides insights on the design and configuration of storage systems
- Main aspects under consideration:
  - I/O Rates
  - Throughput Rates
  - Randomness
  - Locality of reference
  - I/O size distribution
  - % Reads vs Writes

COLLABORATE. INNOVATE. GROW.

Blocktrace Analysis of AI Workloads

- Gives deeper insight into I/O profile
- The block report generated by “btt” provides detail about each I/O:
  - Command (read or write), precise timestamp, starting LBA, ending LBA
  - From the above data we can derive details about:
    - Randomness: If starting address of I/O “B” equals ending address of I/O “A”, I/O is sequential
    - Read/write ratios
    - I/O size distribution: Ending LBA minus starting LBA equals block size in sectors
    - Locality of reference: Some address ranges are accessed more frequently than others

Rule of Thumb

- AI workloads are computation bound
  - Loading a 200KB image takes ~200us
  - Classify a image takes ~10ms
- Parallelize AI jobs to saturate I/O
  - Use a cluster of GPUs
  - Keep every GPU busy

I/O intensive Methodologies

Benchmarking AI workloads in a customer representative scenarios

Limiting Memory

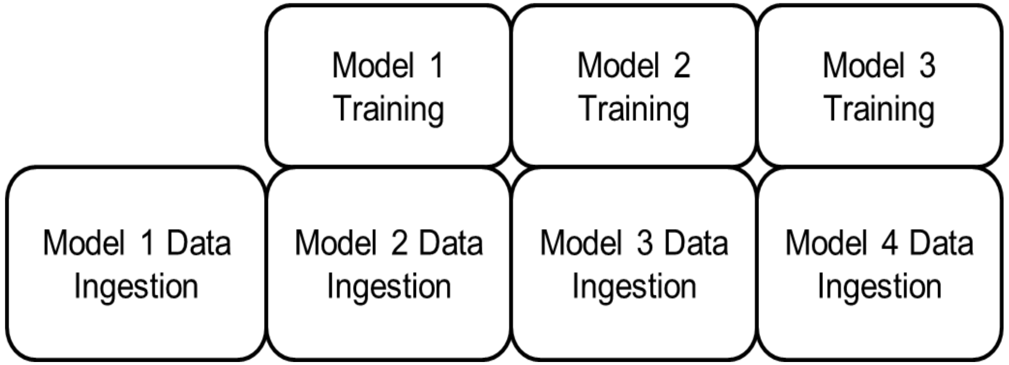
- To accurately model realistic workload with very large training dataset requirement
  - Readily available benchmark datasets are small and fit in memory
  - Goal is to stress storage in a small realistic test environment
- Control Dataset size to memory ratio
  - e.g. MLPerf ImageNet dataset (150 GB)
  - Docker memory limit options

Dataset Size (GB)	System Memory (GB)	Ratio
150	768	1:5
150	64	2.5:1

COLLABORATE. INNOVATE. GROW.

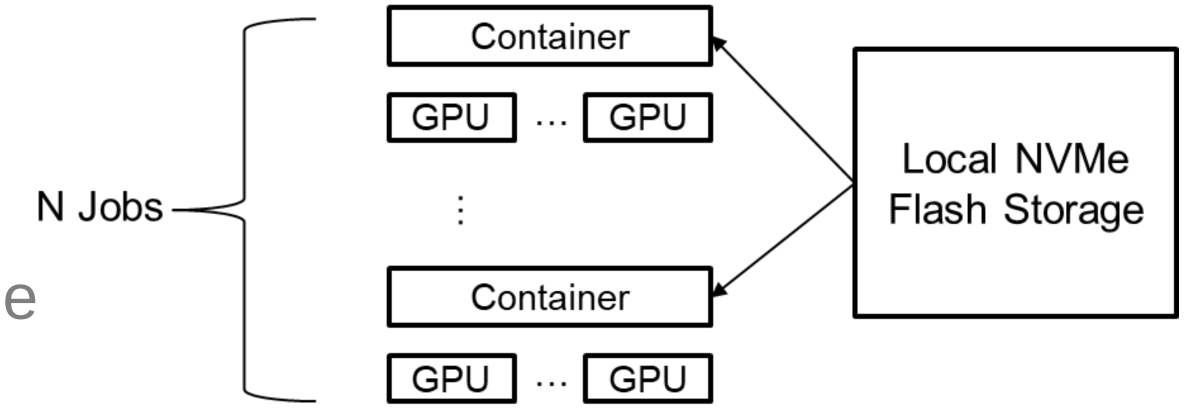
Simultaneous Data Ingestion and Training

- Normally, training is not run in isolation
- Multiple models to be trained
- Realistic scenario: data ingest and training happen together



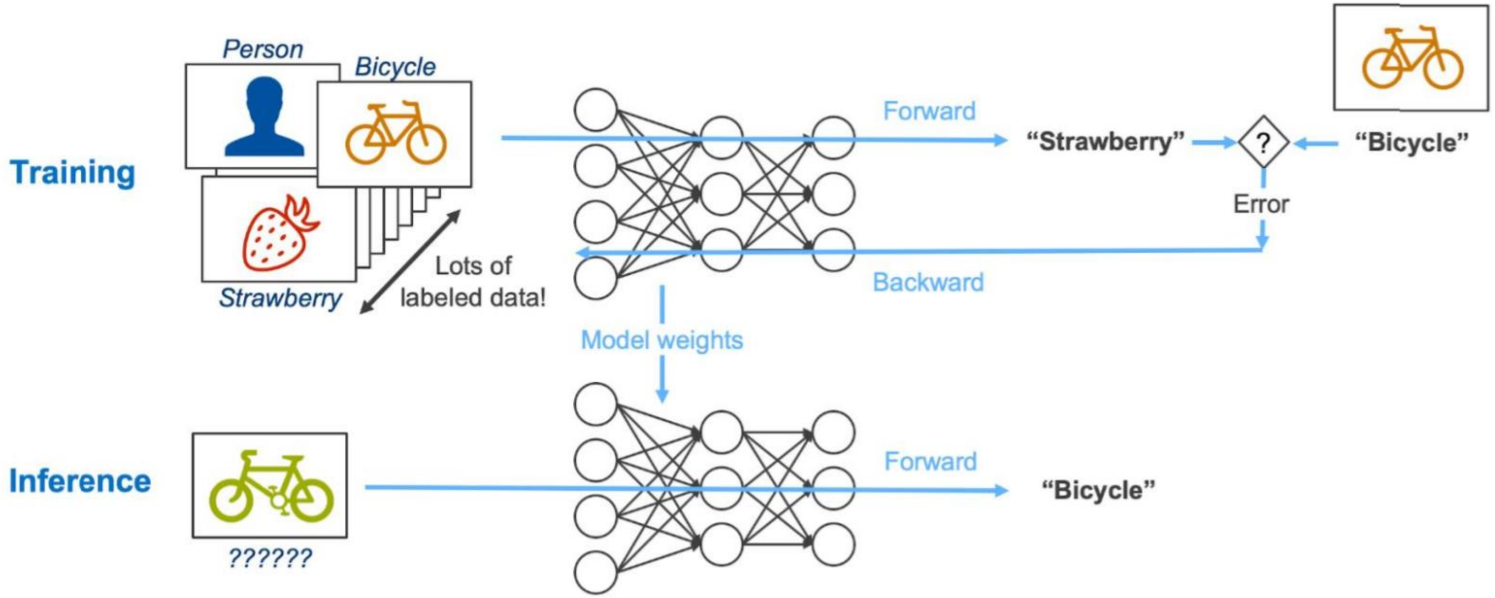
Training in parallel

- Training parallelism:
  - Storage to meet the needs of concurrent data ingest of different training jobs
- Hyper-parameter tuning:
  - Run tens of hundreds of instances of the same training job with different configuration of the model



Inference: Streaming applications

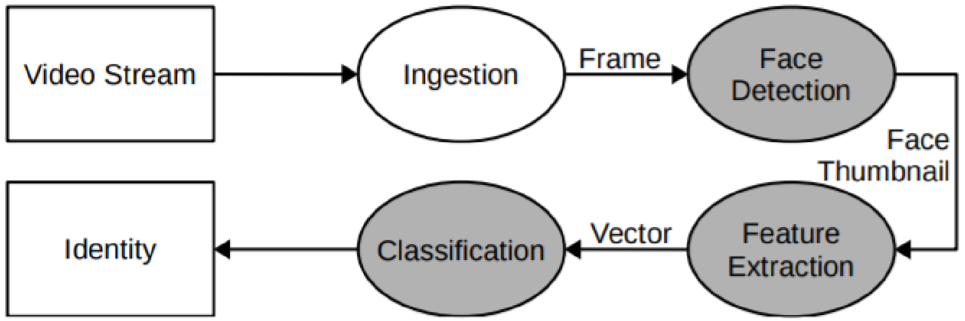
- Inference is more likely I/O bound
  - Training has 3x computations compared to Inferencing
    - Forward propagation, backward propagation, and weight updates
  - Less CPU bound implies possibility of I/O bound



 |

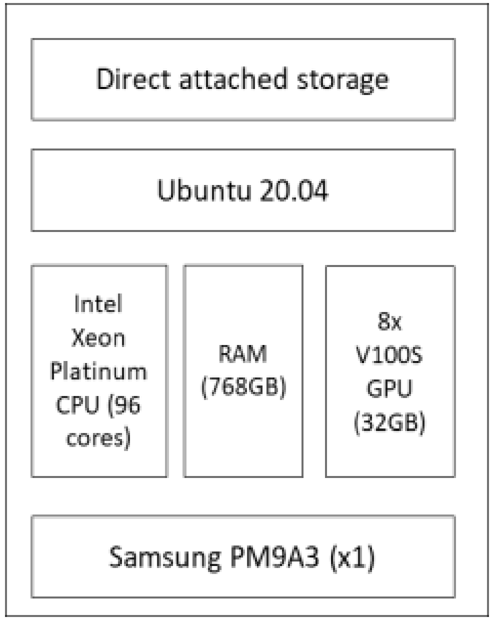
I/O Challenges for Streaming applications

- Large amount of concurrent input data volume
  - One 4K 30 fps video stream: 45Mbps (~6MBps)
    - 1000 video streams: 45Gbps (~6GBps)
  - Massive intermediate data from different stages in a pipeline
- Video processing pipeline
  - Videos are split into frames
  - Stages are isolated into containers
  - One stage consume frames from last stage
  - Frames are passed through Apache Kafka with replicas

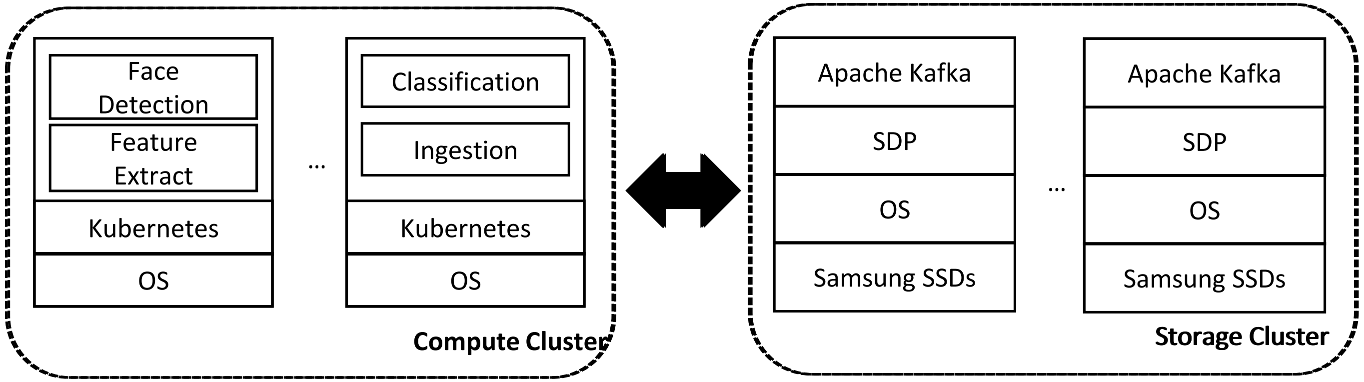


Test System

Hardware Components	Details
GPU	8x Nvidia Tesla V100S, 32 GB
CPU	Intel Xeon Platinum 8268, 2.9 GHz, 2 Sockets, 2 threads per core, 96 (24*2*2) total cores, 768 GB System Memory
Storage	Local: 1 Samsung PM9A3 (3.49 TiB) drive per host: PCI Express Gen4 x 4 interface U.2 (EXT4 file system)
Software Components	Details
Ubuntu	20.04 focal
Tensorflow (tensorflow-gpu)	MLPerf- Version: 2.4.1
Docker	Version: 20.10.12
CUDA Toolkit	Version: CUDA-11.2
FIO	Version: 3.26-59
ResNet50 v1.5 model	Distributed multi-GPU training with ImageNet ILSVRC2012 dataset
OpenMPI	Version: 3.0.0
Horovod	Version: 0.24.2



- For inference testbed:
  - Compute node cluster
    - Kubernetes
  - Storage (message broker) cluster
    - Kafka (Helm charts)



COLLABORATE. INNOVATE. GROW.

Dataset and Model details

Task	Model	Framework	Dataset details
Image classification training	ResNet50	Tensorflow-gpu: 2.4.1	ImageNet-1k
Video streaming and recognition: Inference through Image classification model	ResNet50	Tensorflow-gpu: 2.11.0	1. Videos: a. Big Buck Bunny, Frame rate: 24FPS, Resolution: 1920 x 1080, Size: 45 MB, Duration: 09:56 min b. Costa Rica, Frame rate: 60FPS, Resolution: 3840 x 2160, Size: 1.13 GB, Duration: 05:13 min 2. ImageNet-1k Validation dataset

COLLABORATE. INNOVATE. GROW.

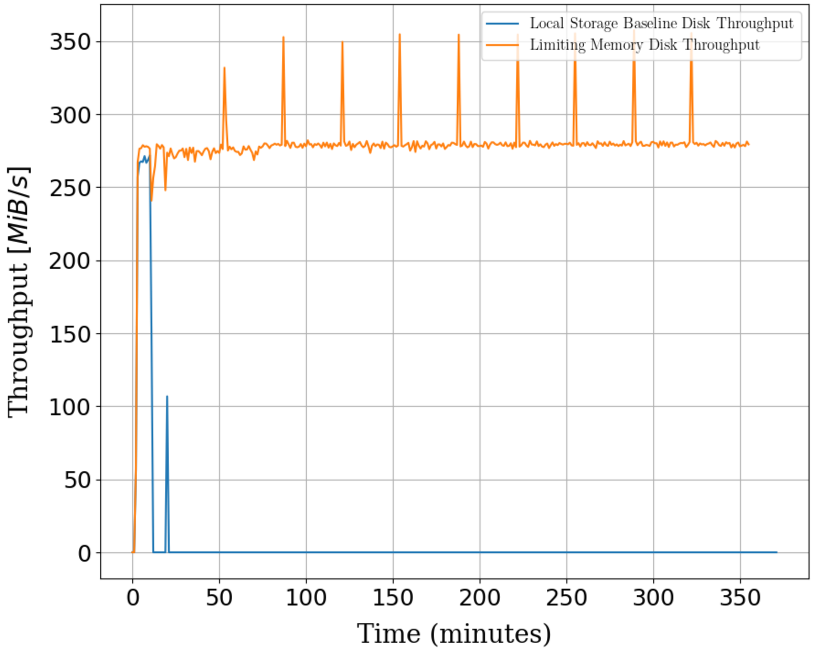
SAMSUNG SDC 23

Impact of Limiting Memory

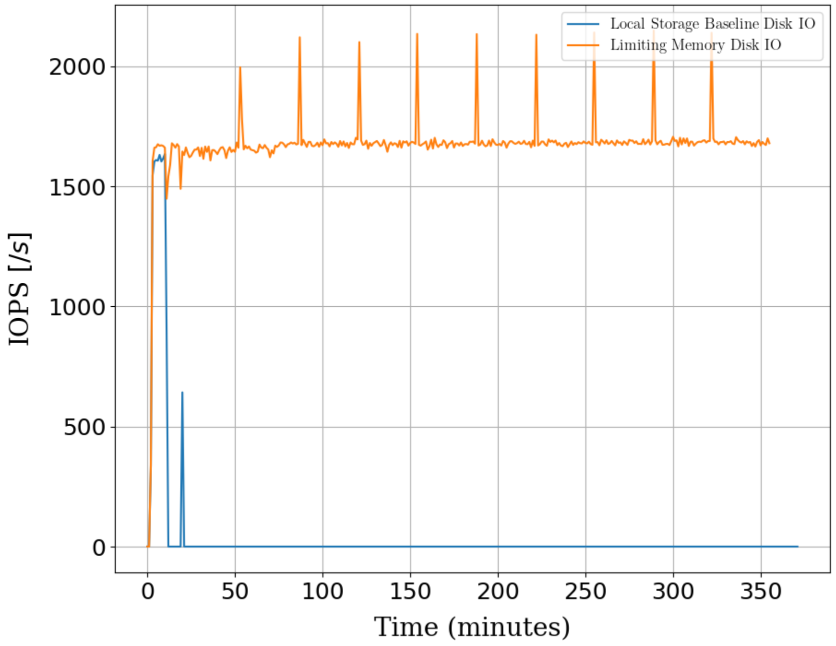
Baseline vs Limited memory: Disk profiles

Metric	Baseline	Limited Memory
Avg. IOPS	23	2,244
Avg. Throughput (MiB/s)	5.84	280.46
Avg. Block Size (KiB)	169.55	170.23
Avg. Response time (µs)	203.63	185.91
Training time (minutes)	364	357

* Zero values are discarded from disk metric statistics calculation in the tables. Disk I/O, Throughput, Block sizes, Response time, CPU and GPU utilization % are average values.



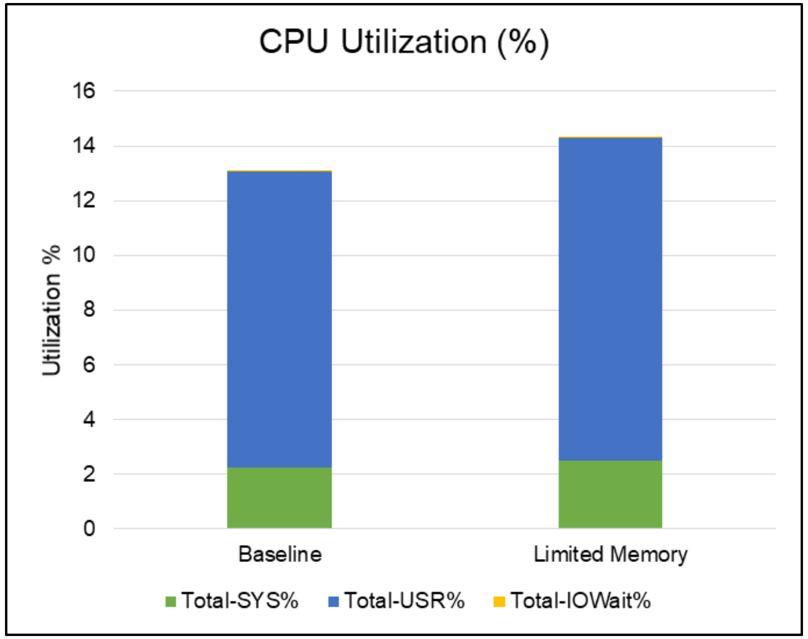
- Disk throughput is substantially increased → 48x
- Training time does not change much when limiting memory → with faster/ performant storage

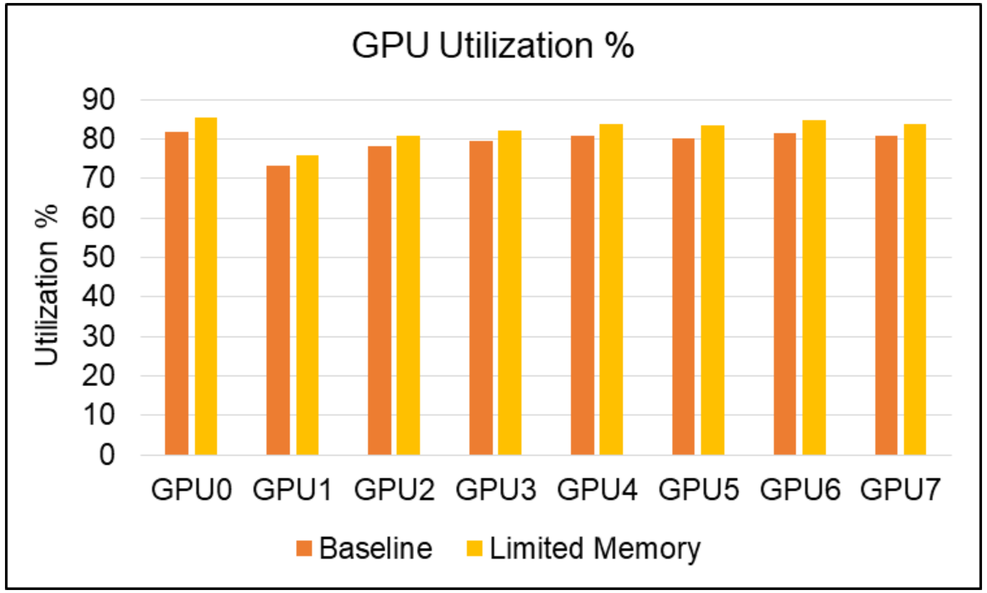


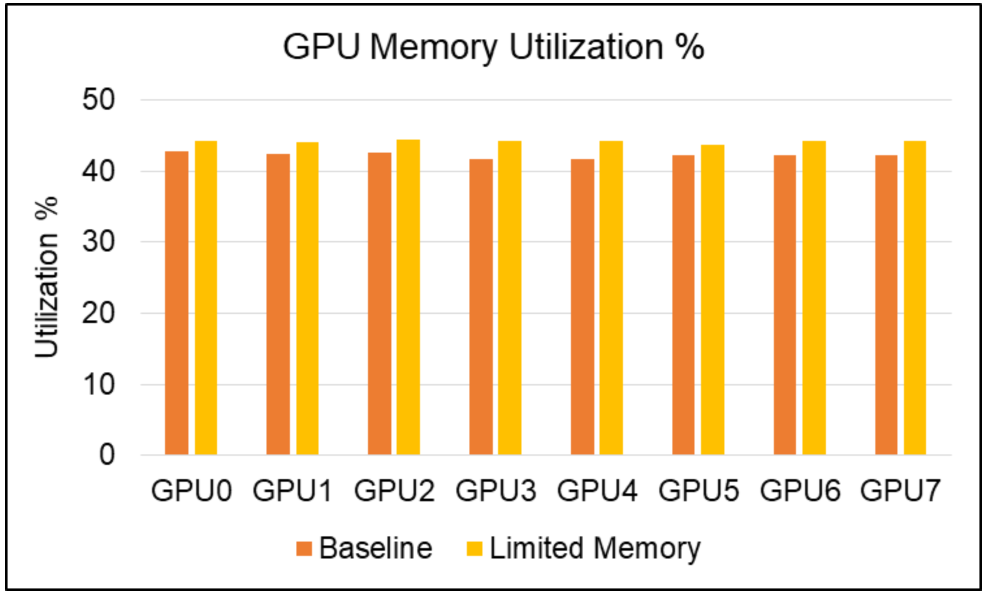
COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC23

System resources







- Baseline and Limiting memory exhibit comparable performance

COLLABORATE. INNOVATE. GROW.

I/O Profile: Resnet50 Single-Model Training

I/O	Read Pct.	Random Pct.	Average IOPS	Minimum Read Request (KiB)	Median Read Request (KiB)	Maximum Read Request (KiB)	Mean Read Request (KiB)	Standard Deviation (KiB)	Minimum Write Request (KiB)	Median Write Request (KiB)	Maximum Write Request (KiB)	Mean Write Request (KiB)	Standard Deviation (KiB)
Total	99.94%	83.88%	639	4	128	256	171	60	4	8	108	16	16
Random	99.96%	100%	536	4	128	256	177	62	4	8	108	13	13
Sequential	99.85%	0%	103	4	128	256	135	30	4	4	44	19	18

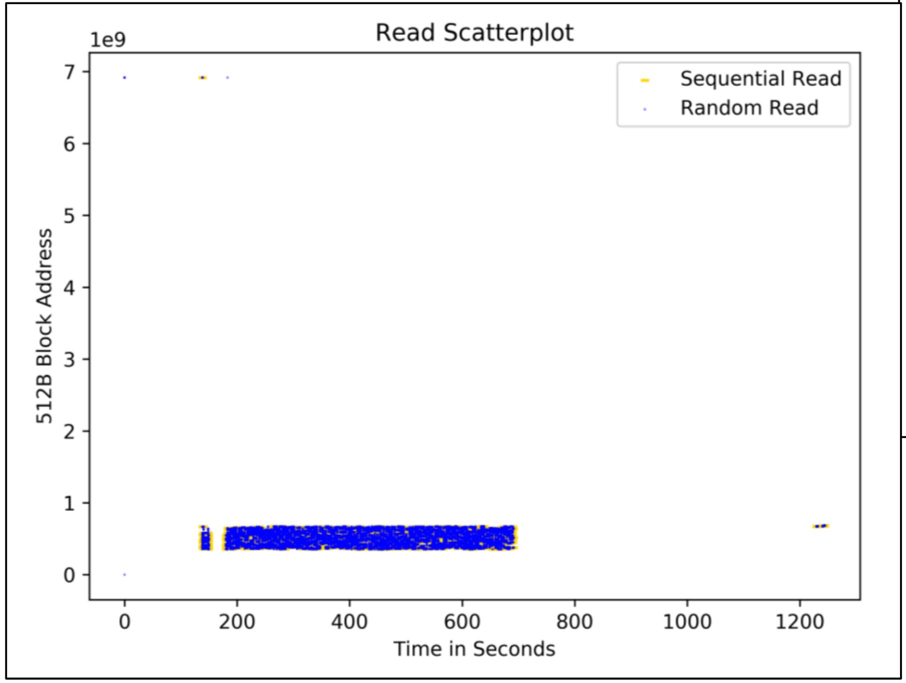
- Nearly 100% read, 84% random, with I/O sizes ranging from 4K to 256K

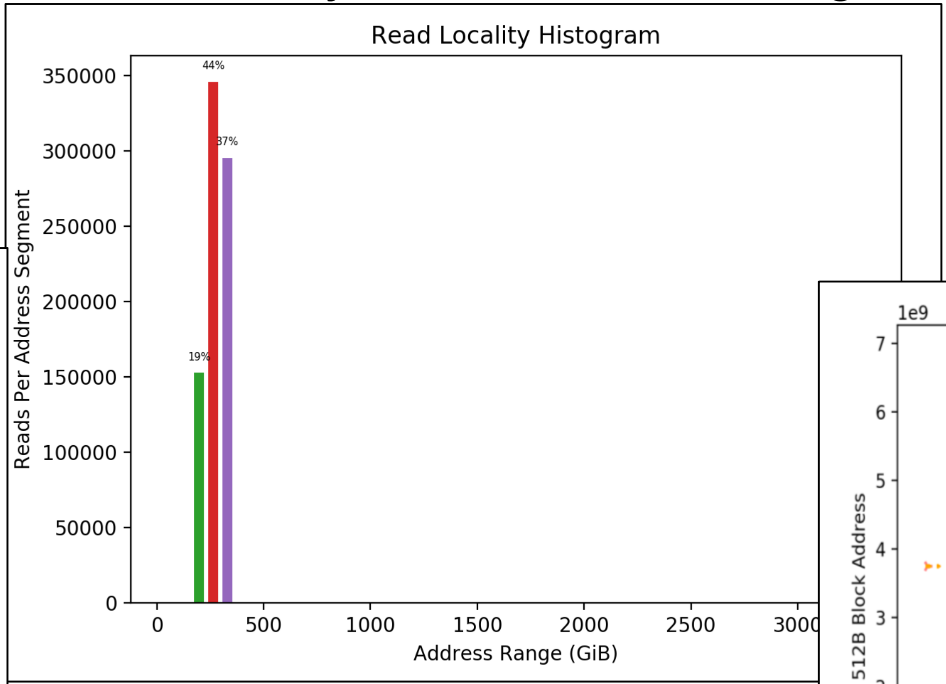
COLLABORATE. INNOVATE. GROW.

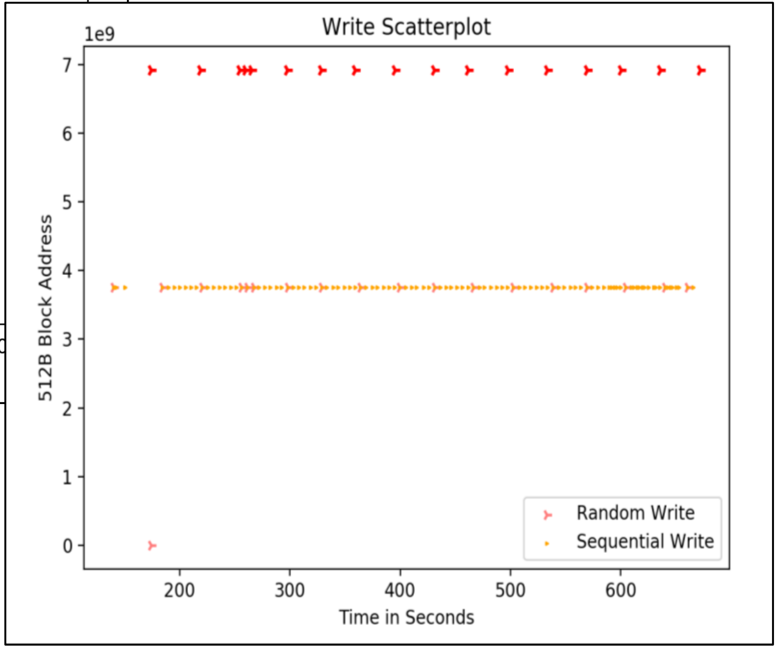
SAMSUNG SDC 23

Trace statistics: I/O plots and locality histogram

- Random and Sequential reads within a relatively narrow address range
- High locality of reference





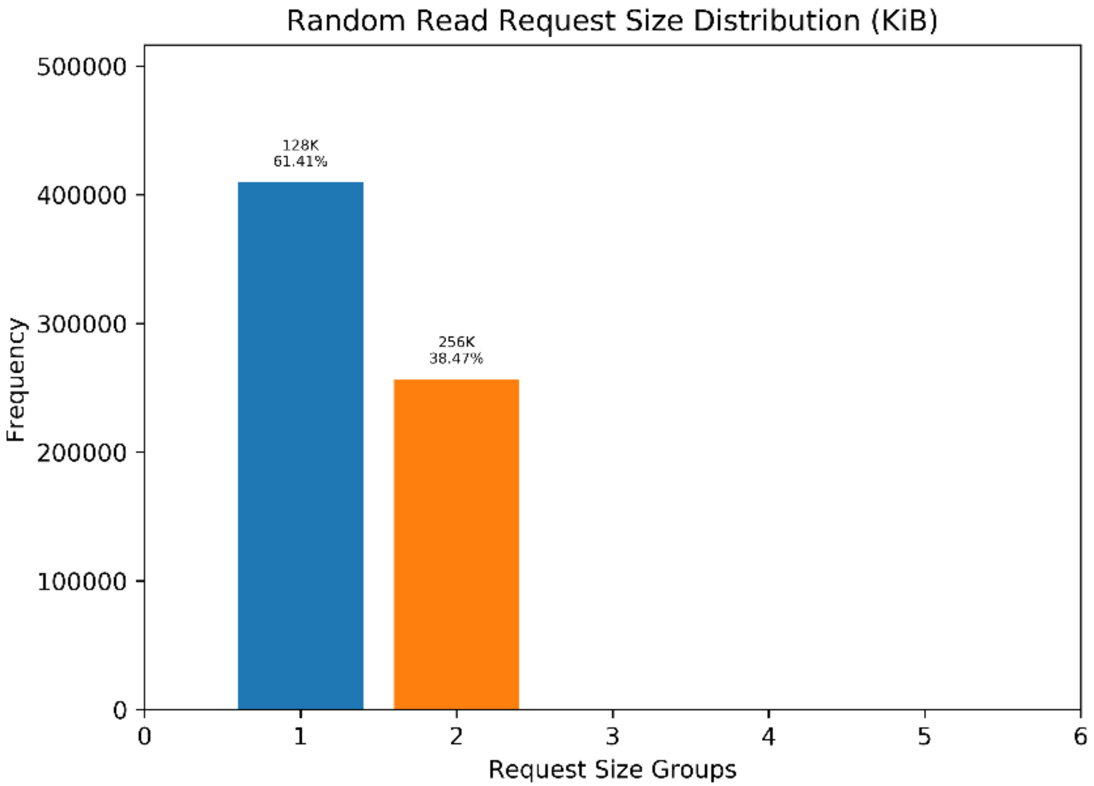


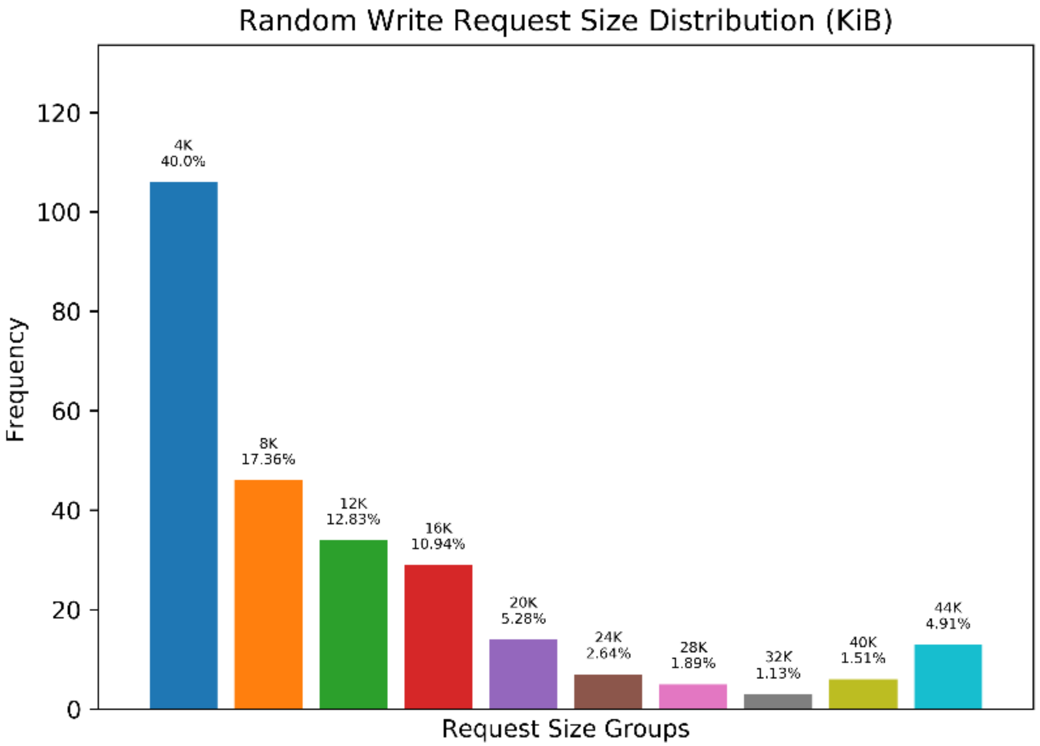
COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC23

Trace statistics: I/O Request Sizes

- Random reads ranged from 4K to 256K, but more than 99% were either 128K or 256K (left)
- Random write I/O sizes were more diverse (right). Sequential I/O size distribution was similar.



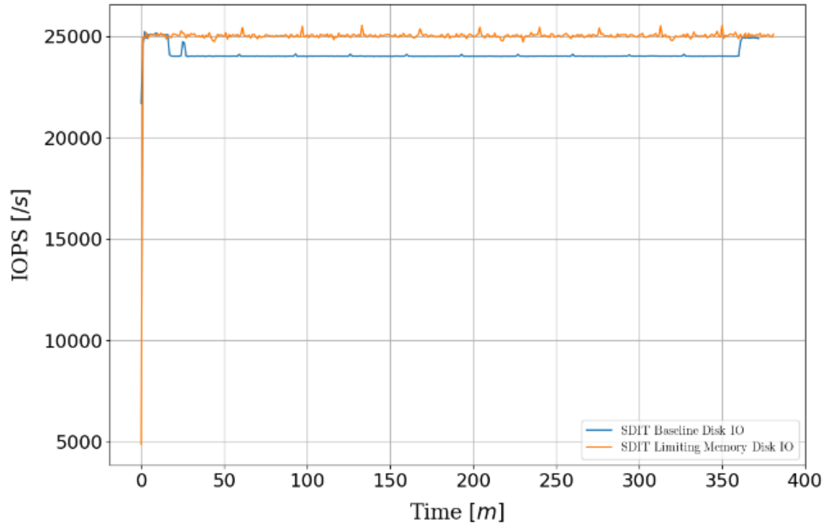


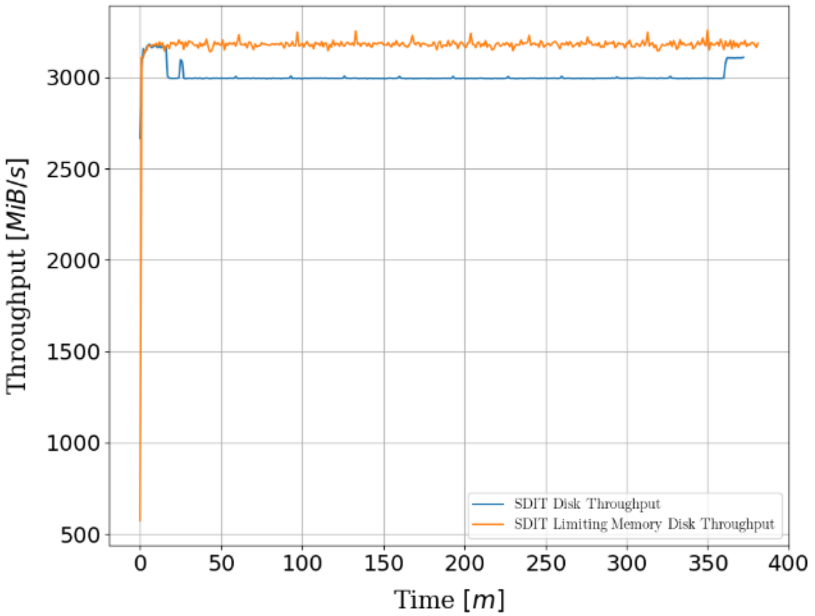
Simultaneous Data Ingestion and Training

Baseline vs Limited memory: Disk profiles

Metric	Baseline	Limited Memory
Avg. IOPS	25054	25035
Avg. Throughput (MiB/s)	3162.59	3181.91
Avg. Block Size (KiB)	Read: 169.8 Write: 128	Read: 170.4 Write: 128
Avg. Response time (ms)	79.418	75.48
Training time (minutes)	373.15	373

* Zero values are discarded from disk metric statistics calculation in the tables. Disk I/O, Throughput, Block sizes, Response time, CPU and GPU utilization % are average values.

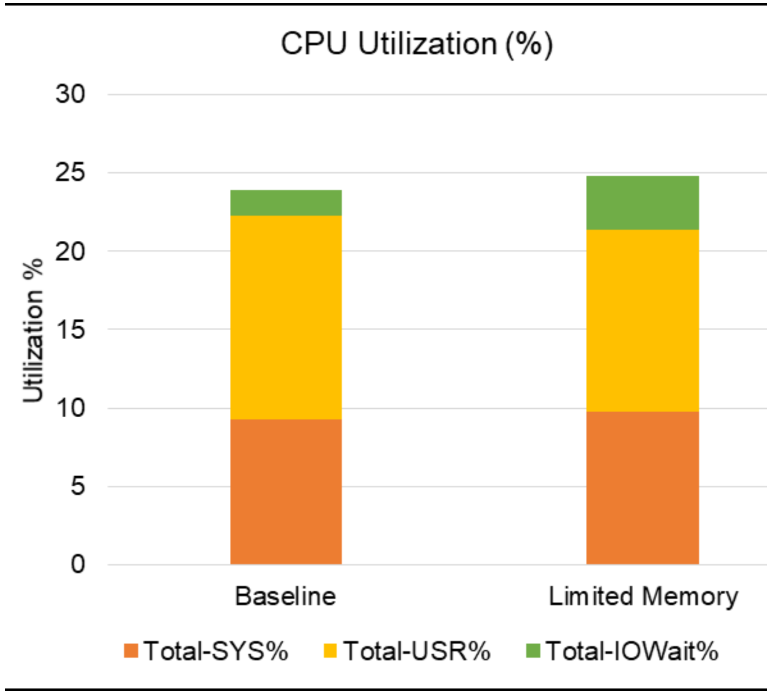


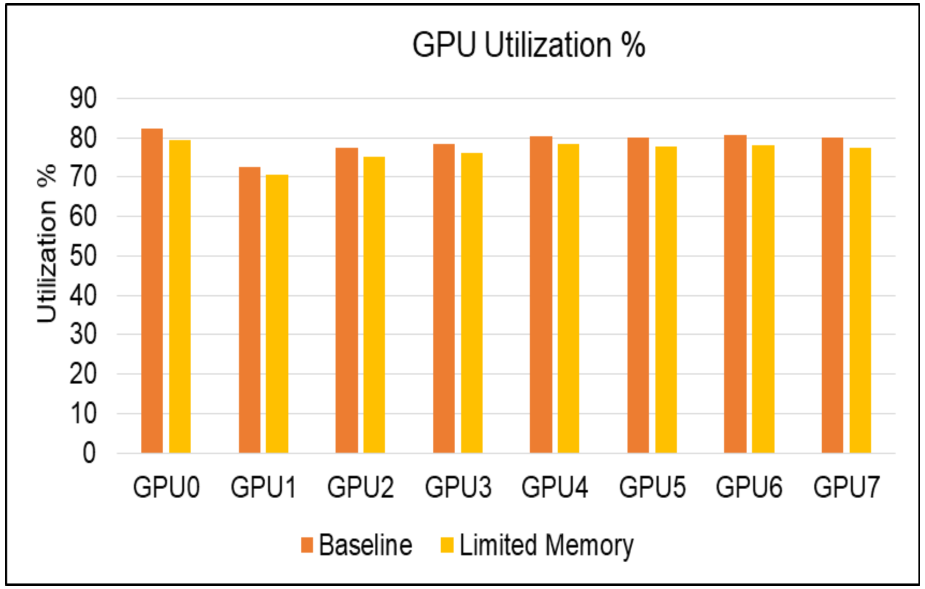


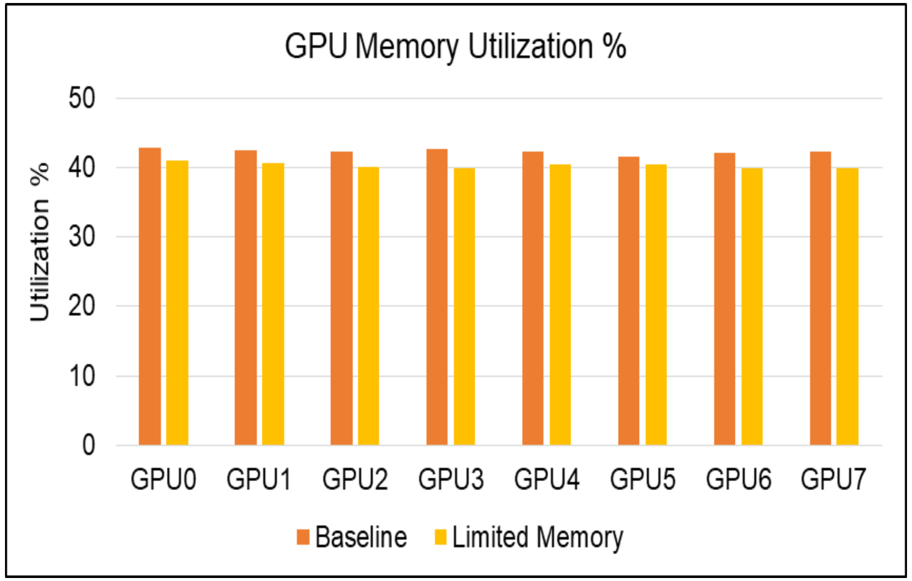
COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC23

System resources







- GPU utilization unaffected:
  - GPU not handling data ingestion operations
- CPU-IOWait increases:
  - Parallel data ingestion

COLLABORATE. INNOVATE. GROW.

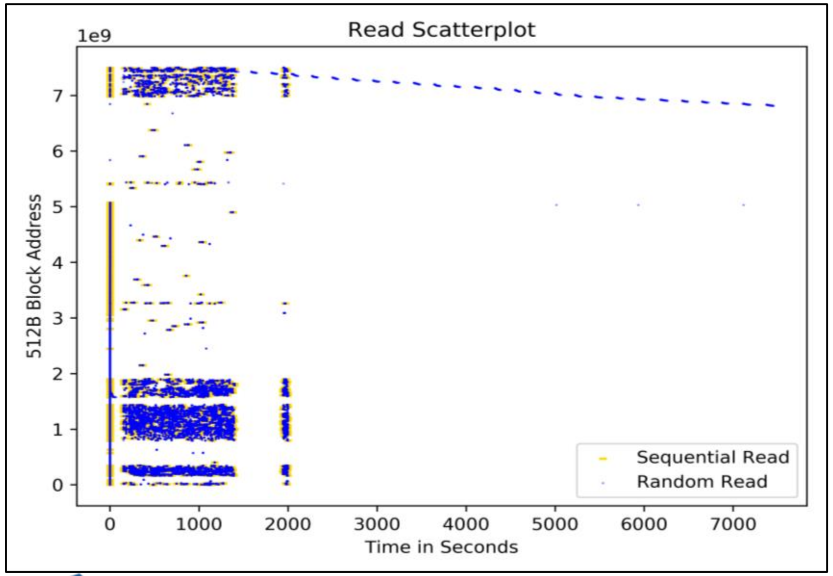
SAMSUNG SDC23

I/O Characterization

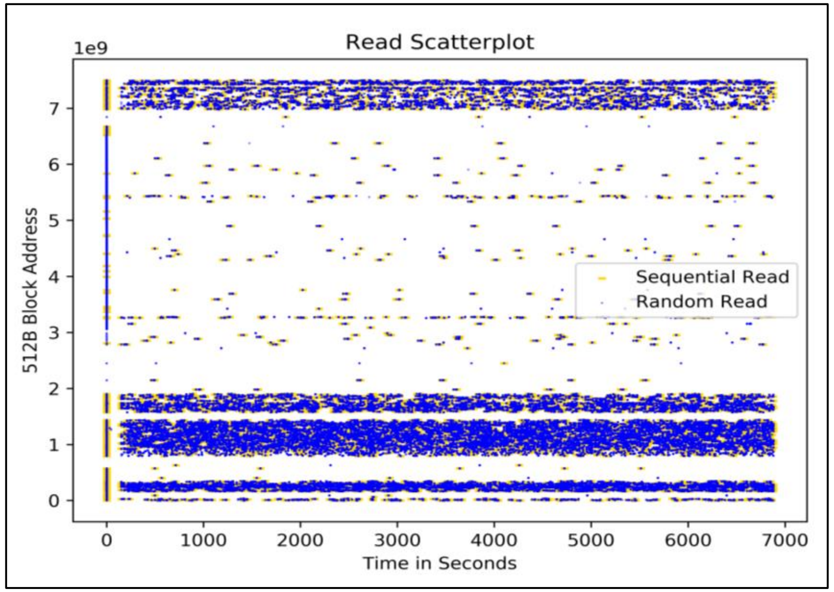
I/O	Read Percent	Random Percent	Average IOPS	Minimum Read (KiB)	Median Read (KiB)*	Mean Read (KiB)	Read Std. Dev. (KiB)	Minimum Write (KiB)	Median Write (KiB)	Maximum Write (KiB)	Mean Write (KiB)	Write Std. Dev. (KiB)
Baseline	0.33%	95.47%	24,714	4	256	247	46	4	128	508	128	6
Limited Memory	1.78%	93.86%	24,786	4	256	245	52	4	128	508	128	7

* Also Maximum Read

Baseline



Limited Memory

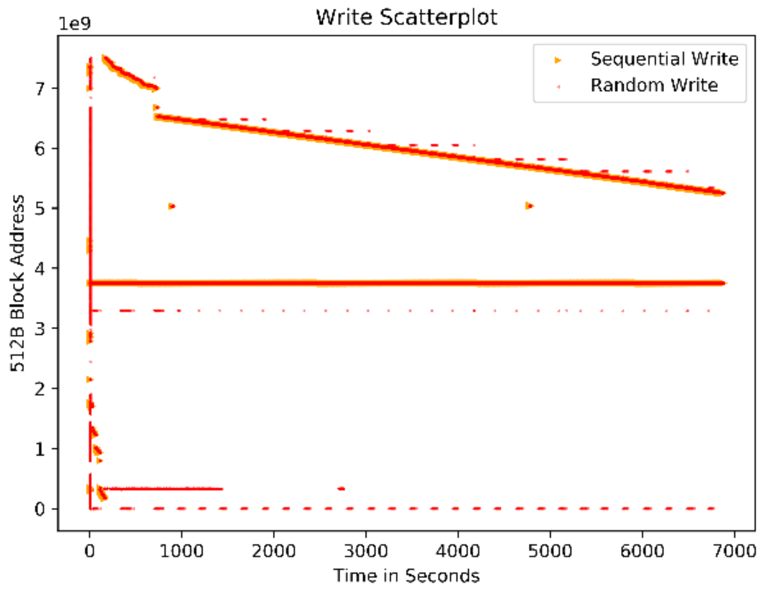


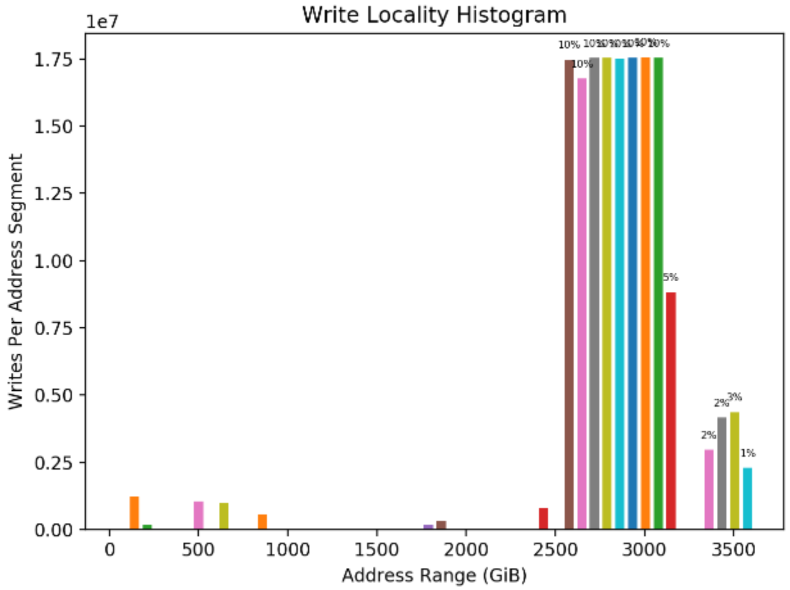
- I/O profile is mostly write and mostly random
- Primary difference between baseline and limited memory is in the read profile
- In baseline training run, disk reads occur primarily in the first epoch because the entire data set fits in memory
- In limited memory run, reads from disk occur during all training epochs

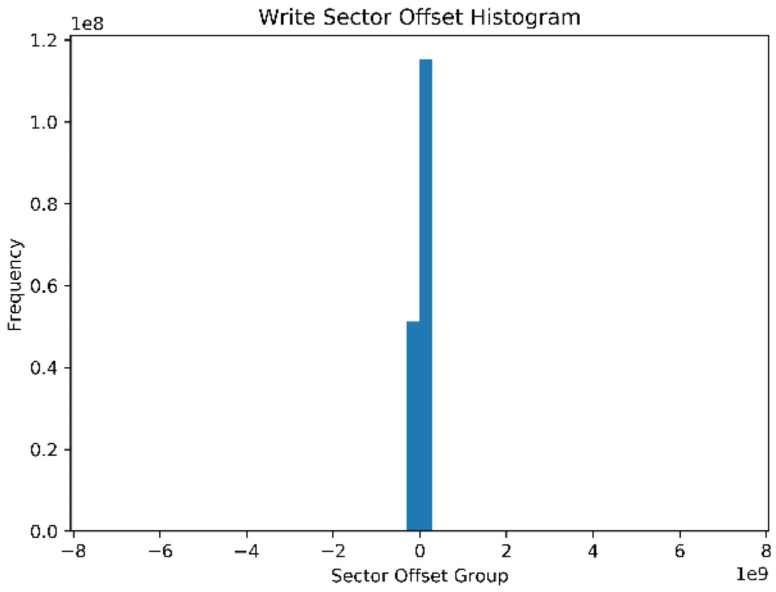
COLLABORATE. INNOVATE. GROW.

Trace statistics: Write I/O plots and locality

- Writes are ~95% random, but locality of reference is high







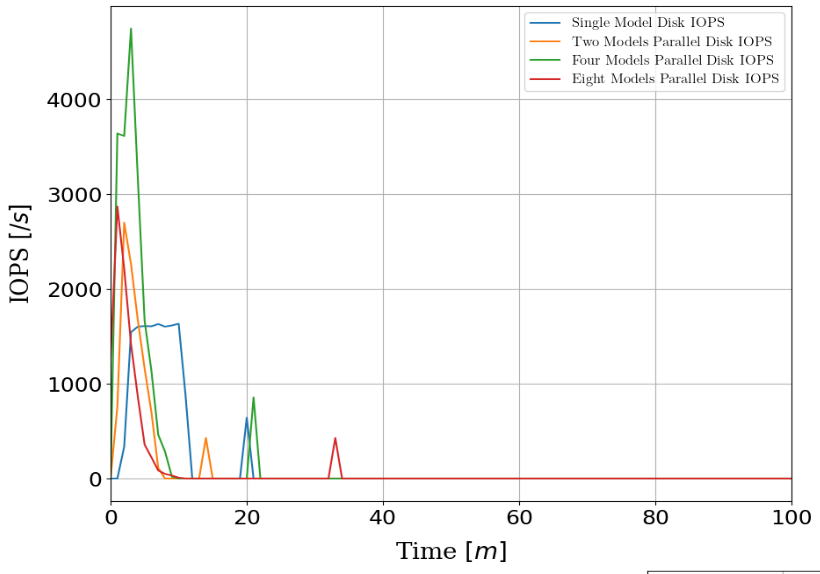
Baseline

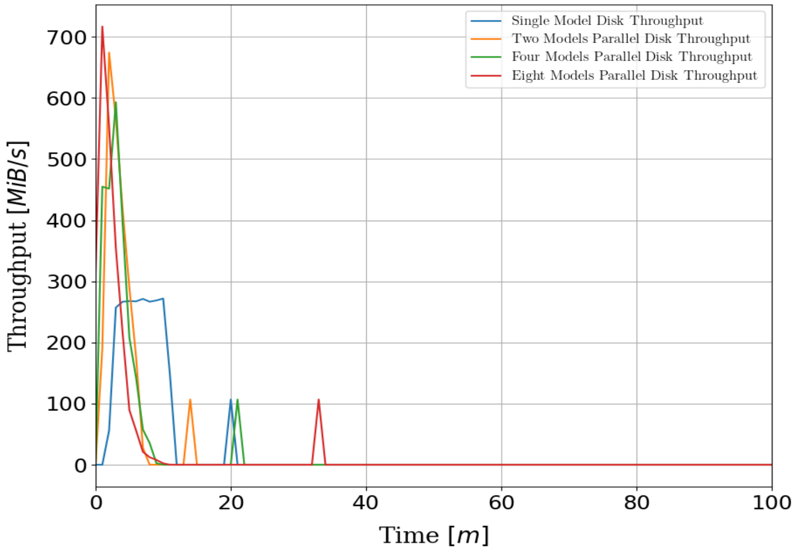
Limiting Memory

Training in Parallel

Parallel models training: Disk profiles

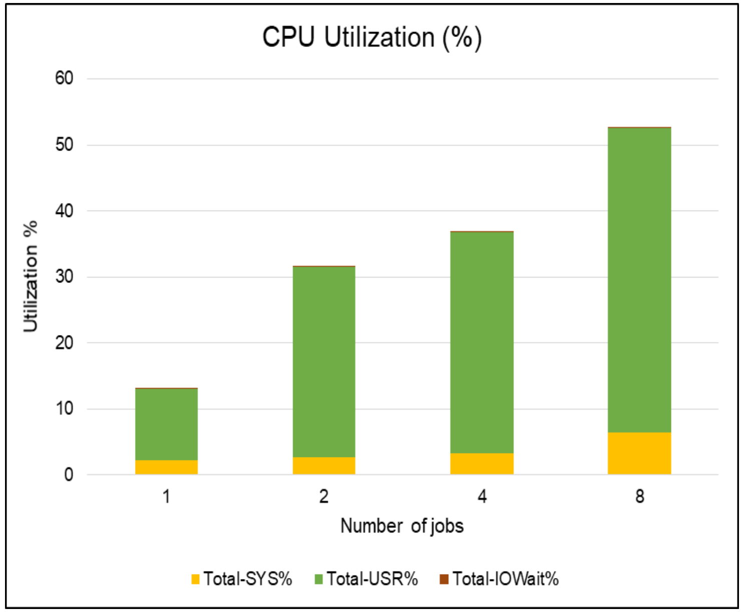
Containers/ Parallel Models	1	2	4	8
GPUs per training workload	8	4	2	1
Batch Size	1024	1024	1024	512
Disk I/O	1658.3	1679.94	2805.26	1245.34
Disk Throughput (MiB/s)	276.55	419.56	351.32	310.72
Block (KiB)	169.55	253.71	127.31	254.2
Response time (µs)	203.63	304.57	162.71	195.88
Training time (minutes)	364	258.2	441	682

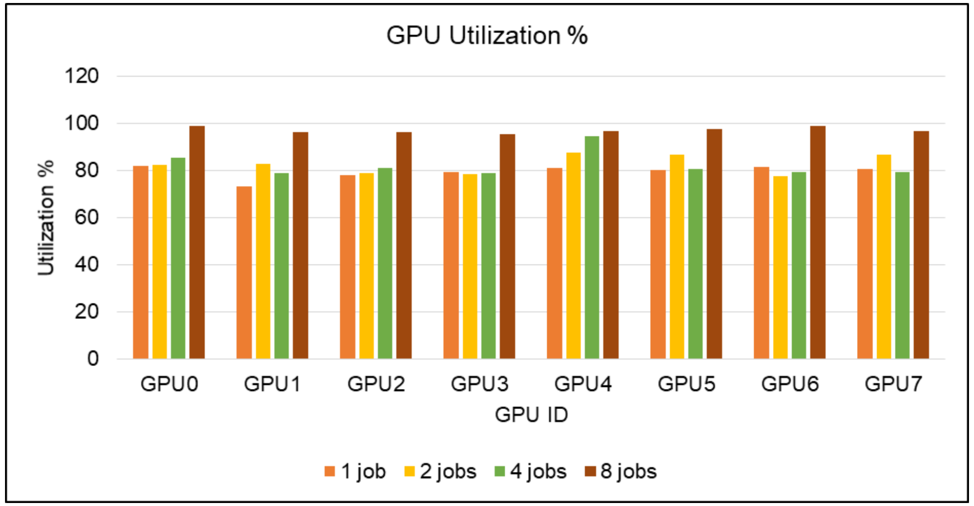


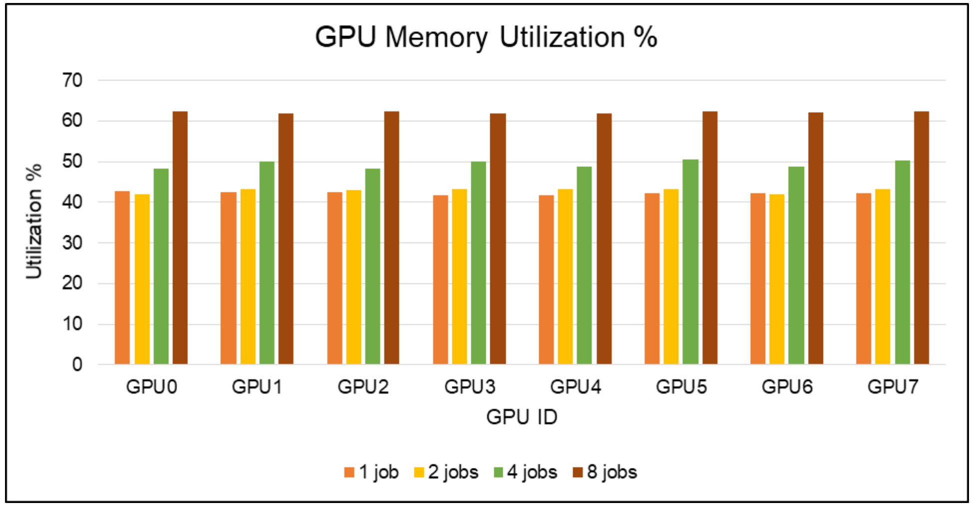


* Zero values are discarded from disk metric statistics calculation in the tables. Disk I/O, Throughput, Block sizes, Response time, CPU and GPU utilization % are average values.

System resources







- CPU and GPU utilization increases with number of read-intensive training workloads

COLLABORATE. INNOVATE. GROW.

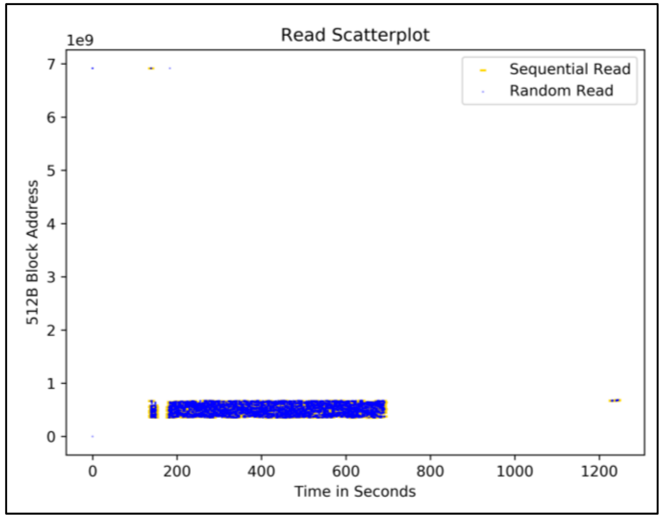
SAMSUNG SDC23

I/O Characterization

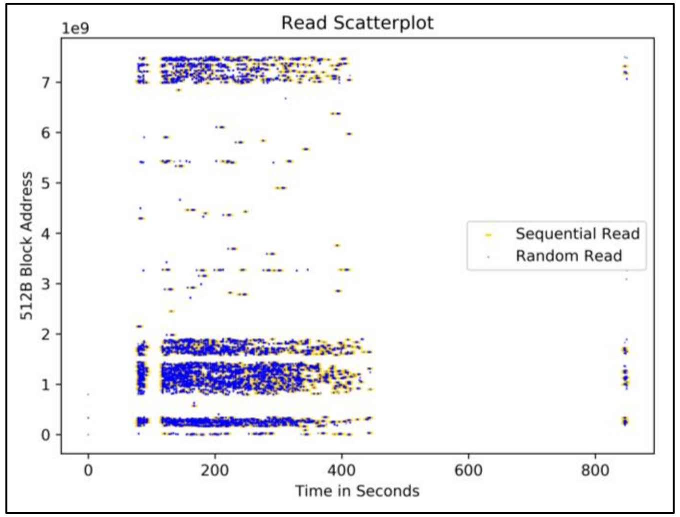
	1 Model	2 Models	4 Models	8 Models
Total Reads	794,262	509,876	1,084,946	509,674
Mean Read Request	170 KiB	256 KiB	128 KiB	256 KiB
Median Read Request	128 KiB	256 KiB	128 KiB	256 KiB
Randomness	83.9%	95.4%	74.8%	92.6%
Locality Bands	1	3	1	3
Percent of I/O received by 10% address space	99%	63%	98%	62%

- 2-models and 8-models parallel training similarities
- Average request size increased from 256 blocks to 512 blocks (256 KiB)
- 8-models training is 100% read, with randomness increasing from 75% (4-models) to 92%

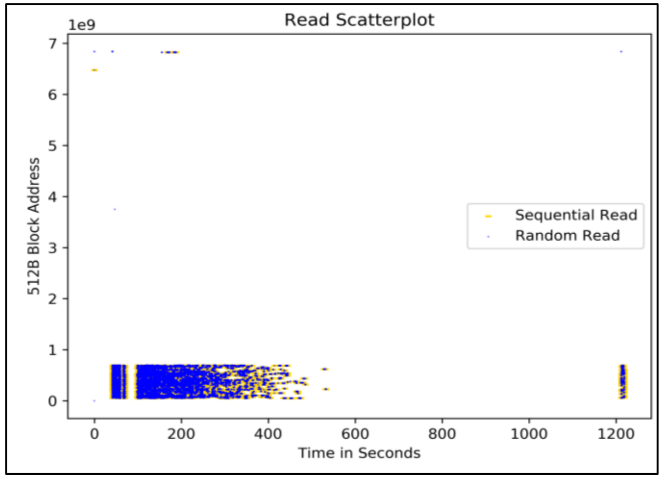
Trace statistics: I/O Plots



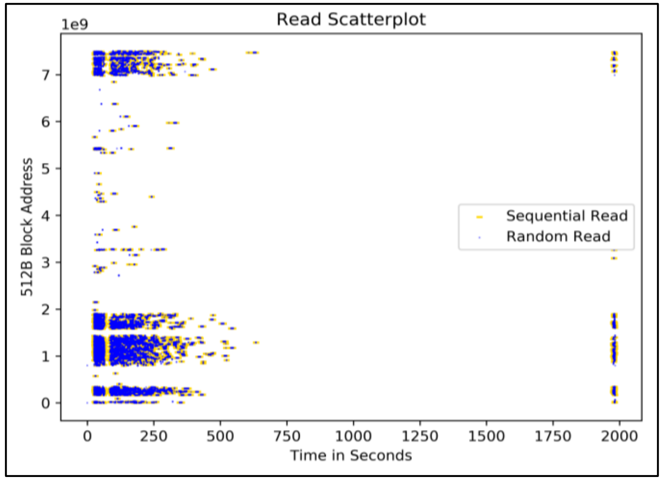
Single



Two models



Four models



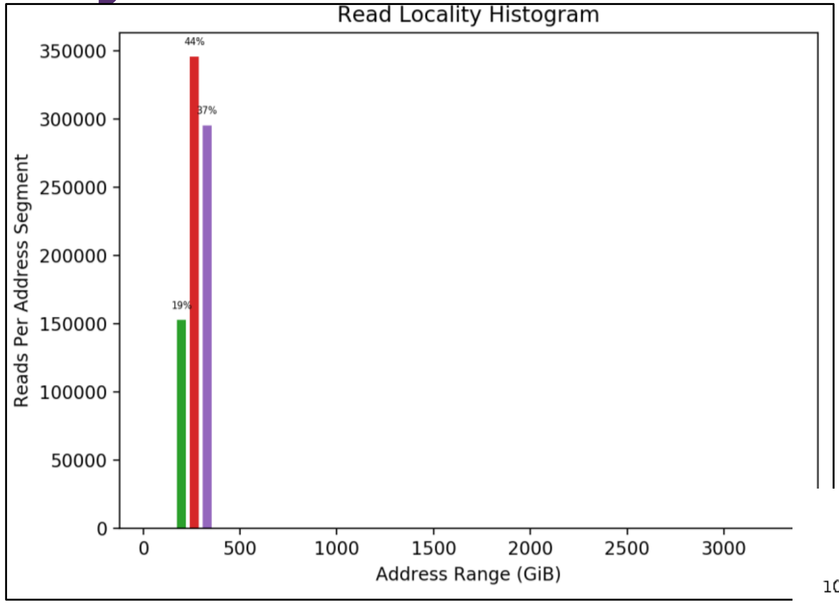
Eight models

- Two- and eight-models show several bands of activity distributed across drive's address range

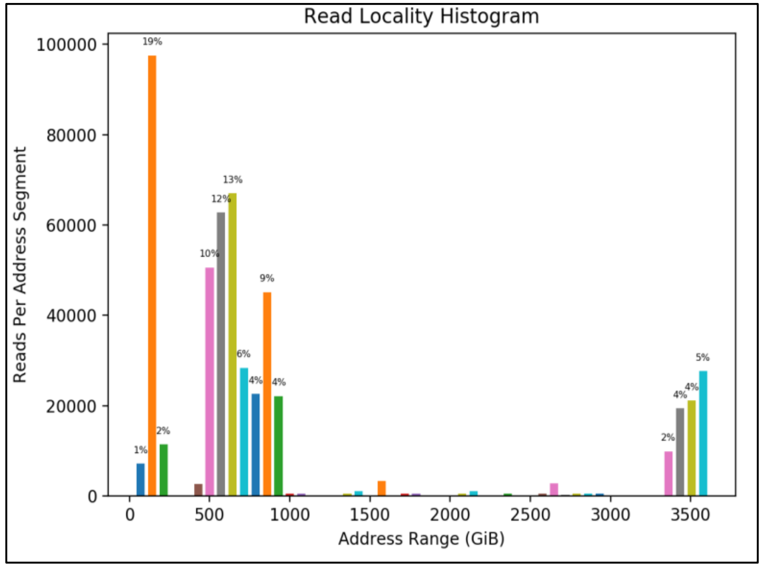
Trace statistics: Locality

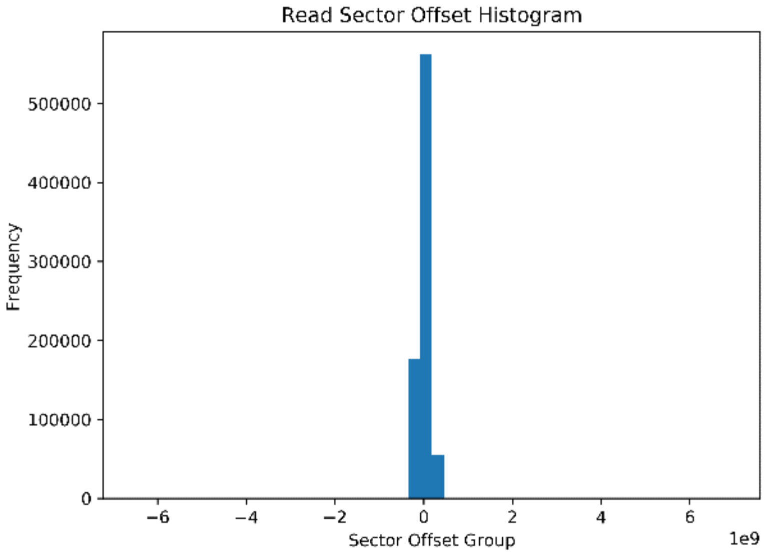
- Highest locality of reference in single model training: 6% address space receiving > 99% reads
- Two- and eight-models have reads more distributed across the drive's address range

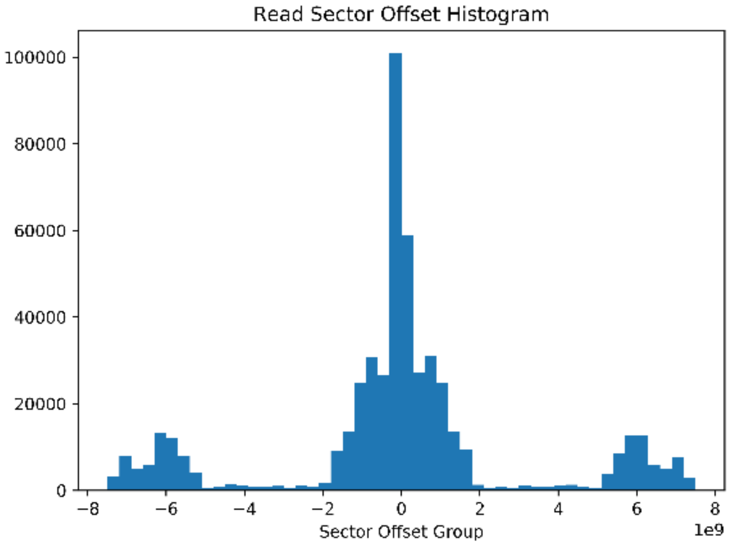
1 & 4 models



2 & 8 models



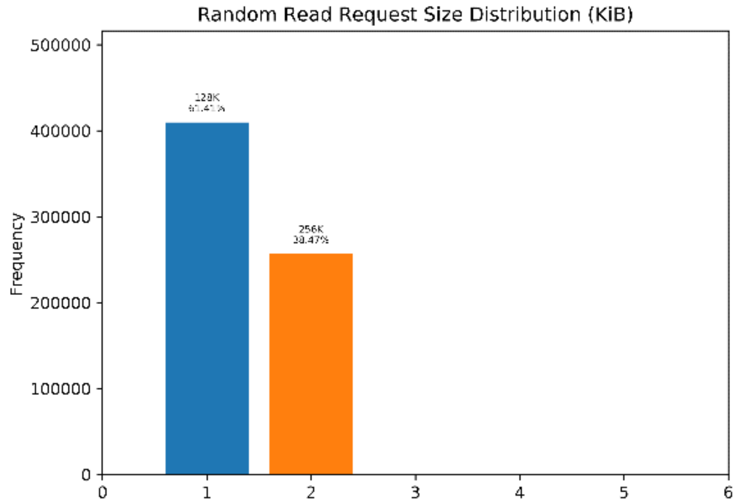


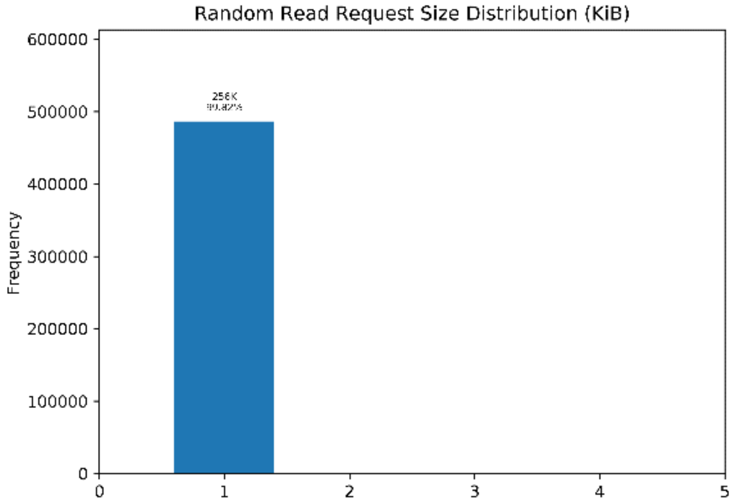


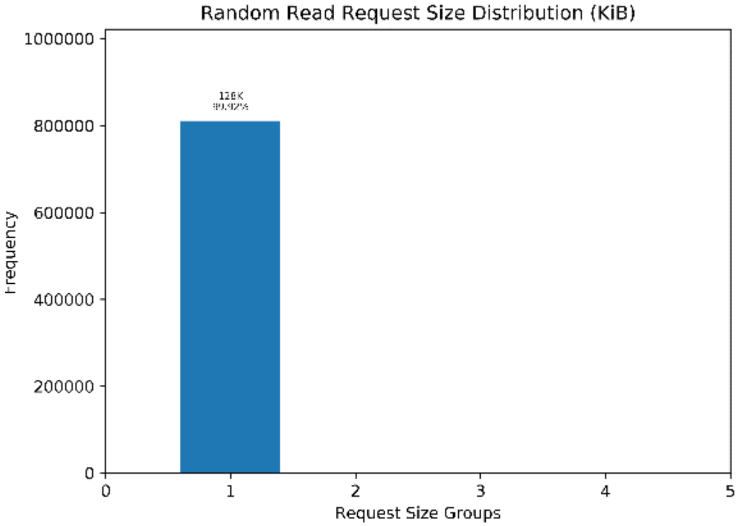
COLLABORATE. INNOVATE. GROW.

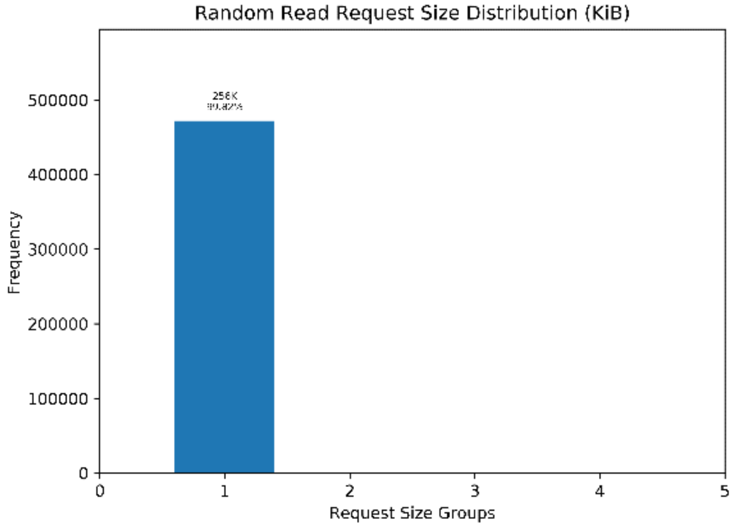
SAMSUNG SDC 23

Trace statistics: I/O Request Sizes









Four models

Eight models

- Single model: Random read request sizes ranged from 4KiB to 256KiB
  - Mainly either 4KiB or 256KiB
- Four models: Most reads are 128 KiB

COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC23

Inference: Streaming workload

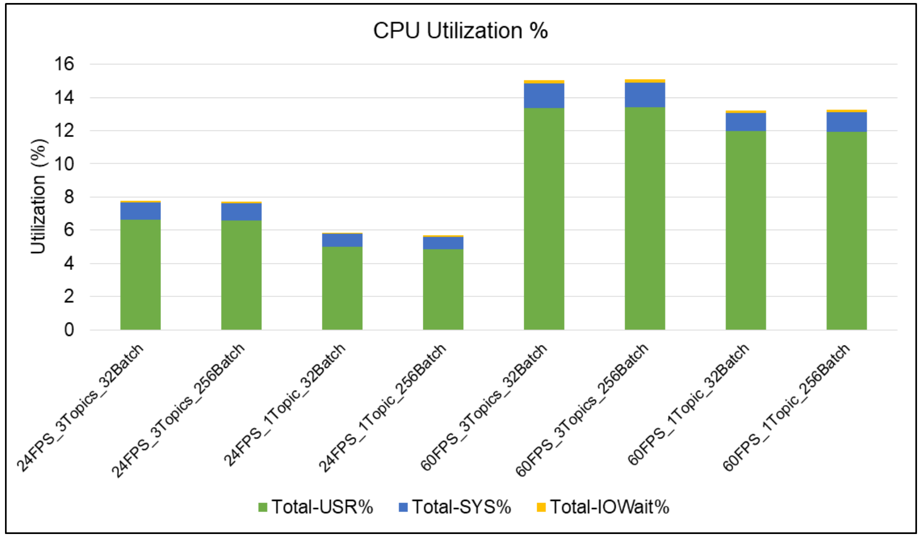
Data Ingestion Disk Metrics

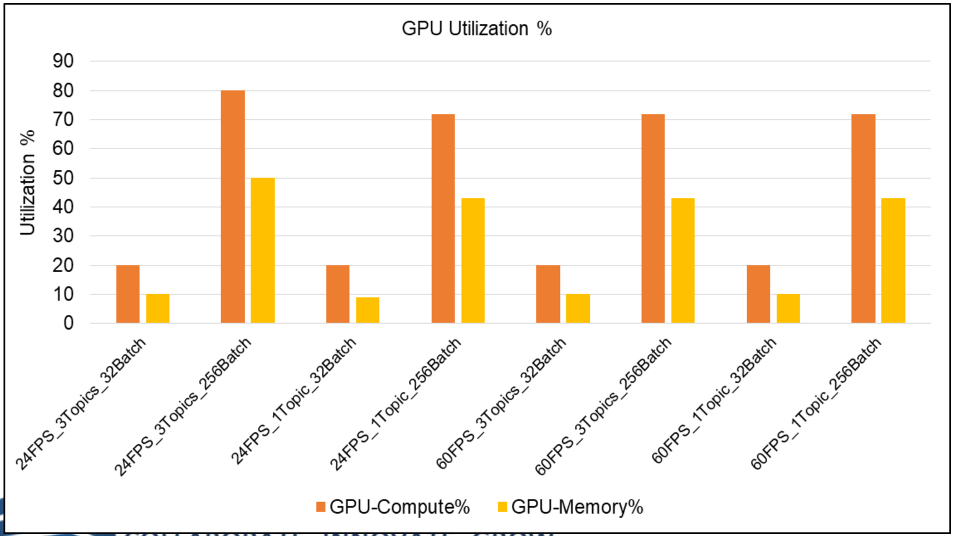
Metric/ Concurrent Streams	300, 24 FPS Videos, 3 RF (6 partitions) - 1 topics	300, 24 FPS Videos, 3 RF (6 partitions) - 3 topics	300, 60 FPS Videos, 3 RF (6 partitions) - 1 topic	300, 60 FPS Videos, 3 RF (6 partitions) - 3 topics
Avg. IOPS	4471.79	7327.74	27637.63	13234
Avg. Throughput (MiB/s)	46.77	152.69	407.75	306.63
Avg. Block Size (KiB)	Read: 110.87 Write: 11.69	Read: 44 Write: 18	Read: 157.7 Write: 13.2	Read:125 Write: 21.18
Avg. Response time (µs)	838.37	1489.38	975.29	1223.09

- Frame extraction from 300 concurrent streams and publish to topic: ~27K IOPS
- Disk I/O and Throughput increase with great parallelism

COLLABORATE. INNOVATE. GROW.

System Resources





- CPU overhead increased with increasing partitions from 3 to 6 but remained constant with further increase to 12 partitions.
- Videos with higher frame rate (FPS) and resolution showed relatively higher CPU utilization.

COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC 23

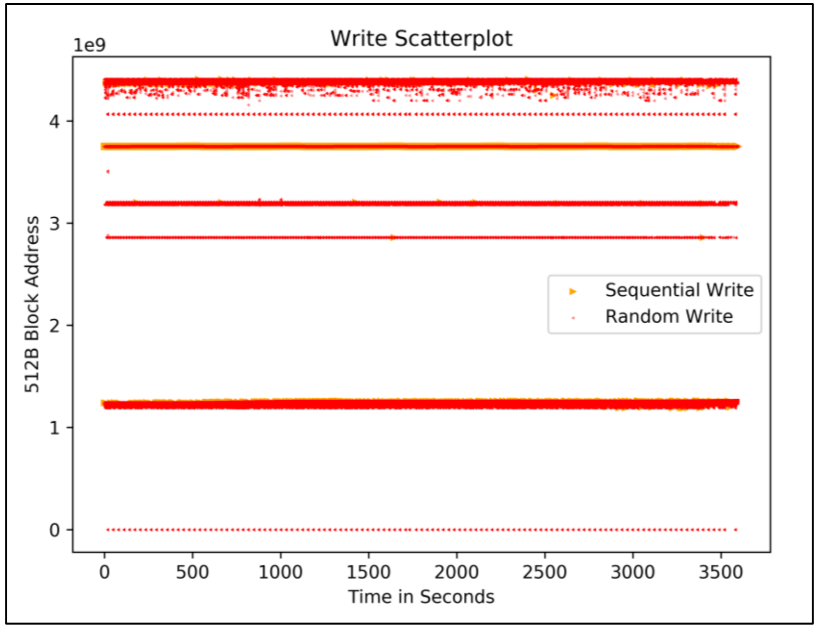
Data Ingestion I/O Characterization

I/O	Read Percent	Random Percent	Average IOPS	Minimum Write (KiB)	Median Write (KiB)	Maximum Write (KiB)	Mean Write (KiB)	Std. Dev. (KiB)
30 Streams	0.08%	71.43%	281	4	4	764	32	96
100 Streams	0.54%	69.92%	422	4	8	764	64	140

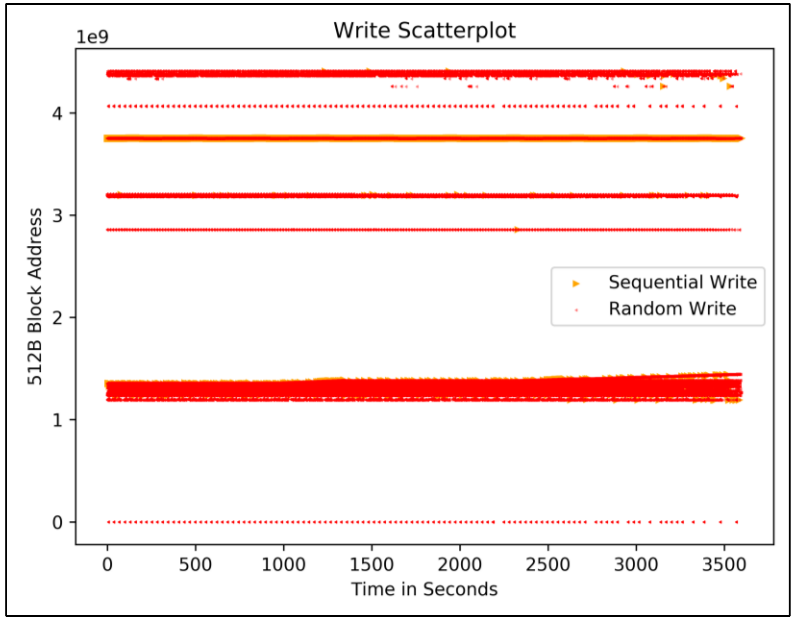
Standard deviation suggests high diversity of write sizes

- Writes more widely distributed across SSD's address range with increased streams

- Nearly 100% write, ~70% random



30 streams

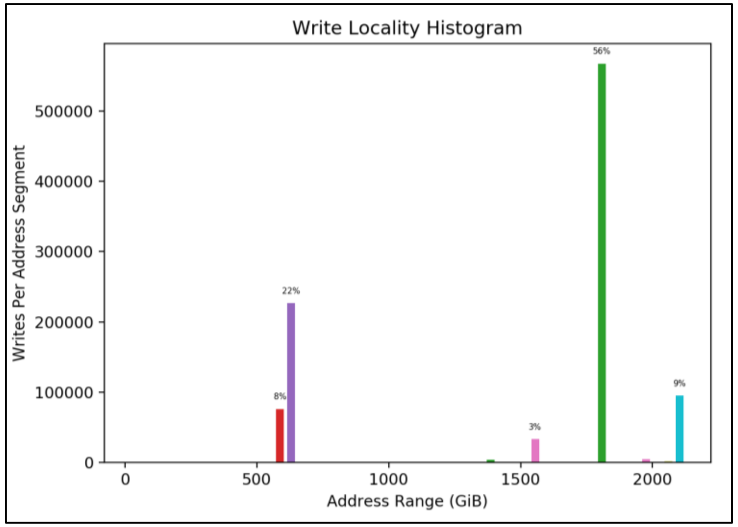


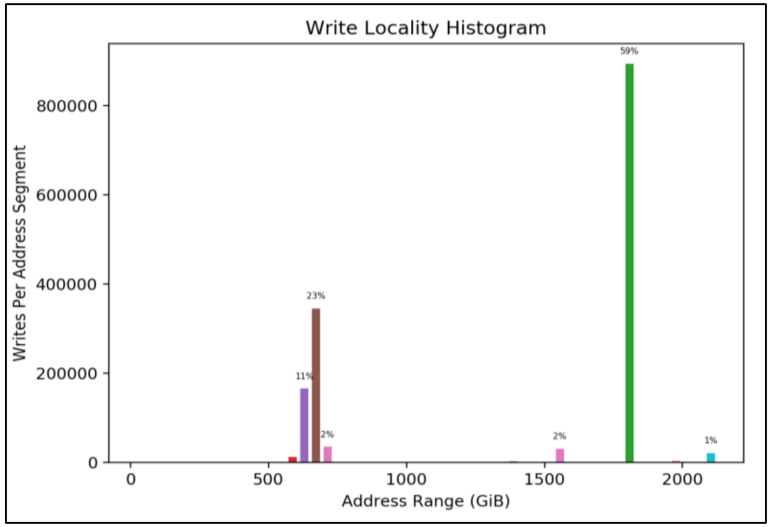
100 streams

COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC 23

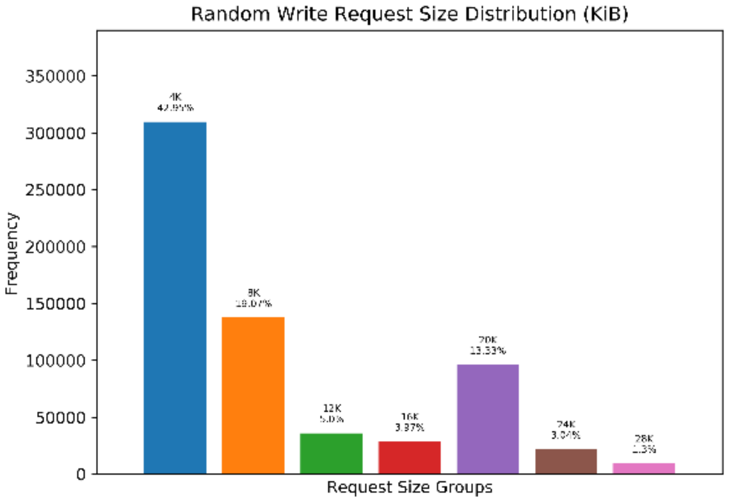
Trace statistics: Locality of reference and I/O sizes distribution



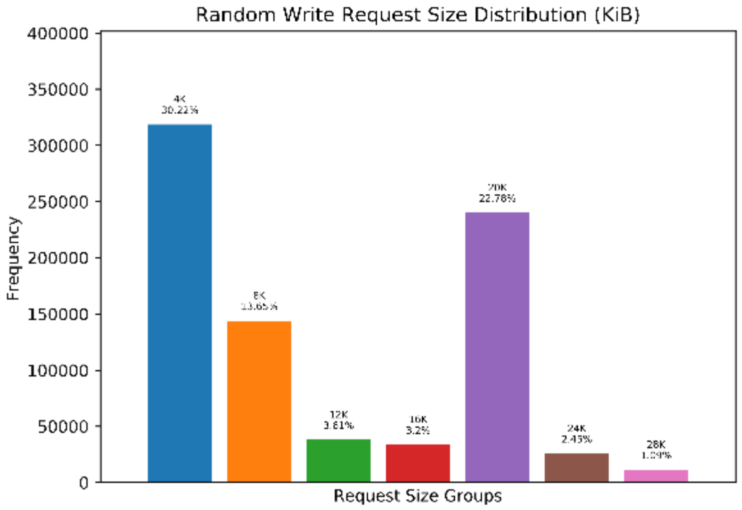


- Write locality high both for 30 and 100 streams with 6% address space receiving 87% and 93% writes respectively.

- Random write request size distribution was quite varied
- 70% of random writes were 28K or less, but the remaining 30% ranged up to 764K



30 streams



100 streams

COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC 23

System Implications and Discussion

- The majority of the workloads studied were primarily random, with relatively high locality of reference
  - Suitable for testing optimizations such as read caching and write coalesce
- Some workloads (e.g. inference streaming) exhibited a very diverse write I/O size distribution
  - Useful “real-world” benchmarking tool for challenging high performance storage systems

COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC 23

Conclusion

- Simultaneous data ingestion and training, and inference were particularly effective benchmarks
  - These approaches present challenging, “real-world” workloads to storage
- Our testing indicates that high-performance storage allows I/O-intensive and computationally-intensive portions of the AI pipeline to run in parallel with minimal impact on training and inference times.

Thank You!

Backup Slides

 |

Summary statistics

Workload Description	Read Percentage	Random Percentage	Average IOPS	Minimum Read Request (KiB)	Median Read Request (KiB)	Maximum Read Request (KiB)	Mean Read Request (KiB)	Standard Deviation (KiB)	Minimum Write Request (KiB)	Median Write Request (KiB)	Maximum Write Request (KiB)	Mean Write Request (KiB)	Standard Deviation (KiB)	Random Read Operations	Random Write Operations	Sequential Read Operations	Sequential Write Operations	Trace Length Seconds
Resnet50 Training Single Model	99.94%	83.88%	639	4	128	256	171	60	4	8	108	16	16	666,340	265	127,922	194	1,244
Resnet50 Training Two Models	100.00%	95.43%	600	4	256	256	256	6	4	4	8	2	2	486,584	2	23,292	2	850
Resnet50 Training Two Models LM	100.00%	96.20%	2,308	4	256	256	172	113	4	4	136	6	6	46,231,316	1,312	1,824,854	744	20,823
Resnet50 Training Four Models	99.95%	74.79%	890	4	128	128	128	2	4	4	128	11	20	811,309	471	273,637	52	1,220
Resnet50 Training Eight Models	100.00%	92.59%	257	4	256	256	256	7	0	0	0	0	0	471,924	0	37,746	0	1,983
Inference Baseline, Video Streaming, Ingestion Phase (30 Streams, 3 Partitions)	0.08%	71.43%	281	4	128	128	102	50	4	4	764	32	96	773	720,927	40	288,605	3,599
Inference Baseline, Video Streaming, Ingestion Phase (100 Streams, 3 Partitions)	0.54%	69.92%	422	4	128	128	118	32	4	8	764	64	140	8,016	1,054,351	260	456,703	3,599
Simultaneous Data Ingestion and Training (5 Epochs)	0.33%	95.47%	24,714	4	256	256	247	46	4	128	508	128	6	574,458	175,355,092	33,960	8,305,481	7,456
Simultaneous Data Ingestion and Training (5 Epochs Limited Memory)	1.78%	93.86%	24,786	4	256	256	245	52	4	128	508	128	7	2,879,201	157,200,319	154,185	10,321,862	6,881
Training with Checkpointing Every 100 Steps	93.27%	92.61%	165	4	256	256	255	14	4	16	1,280	431	567	507,355	12,527	16,214	25,255	3,408
Training with Checkpointing Every 1252 Steps (Default Interval)	99.68%	96.78%	151	4	256	256	256	7	4	16	1,280	134	362	501,256	297	15,348	1,351	3,438
BERT 2000-Step Default Checkpoint Interval PM983	0.22%	4.38%	26	4	128	128	126	15	4	128	128	128	5	69	2,740	74	61,185	2,511
BERT 2000-Step Default Checkpoint Interval PM9A3	0.11%	60.38%	43	4	128	256	168	66	4	8	1,280	36	176	215	164,878	92	108,218	6,395
BERT 2000-Step Default Checkpoint Interval PM9A3 + Preconditioning + New FS	0.23%	0.49%	2	4	128	256	129	89	4	1,280	1,280	1,127	326	9	16	3	5,113	2,163
BERT 2000-Step Default Checkpoint Interval PM9A3 + New FS + Pytorch Framework	0.00%	3.47%	181	0	0	0	0	0	4	508	1,280	579	443	0	7,382	0	205,078	1,176
BERT 2000-Step Limited Memory Default Checkpoint Interval PM983	0.27%	3.63%	26	4	128	128	126	5	4	128	128	128	5	107	2,149	60	59,818	2,380
BERT 2000-Step Limited Memory Default Checkpoint Interval PM9A3	0.12%	58.17%	45	4	128	256	169	63	4	8	1,280	36	174	219	158,072	106	113,707	6,110
BERT 2000-Step With 250-Step Checkpoint Interval PM983	0.10%	3.70%	106	4	128	128	123	25	4	128	128	128	5	133	9,655	119	254,328	2,504
BERT 2000-Step With 250-Step Checkpoint Interval PM9A3	0.08%	57.94%	131	4	128	256	172	64	4	8	1,280	89	285	196	202,814	99	147,279	2,680
BERT 2000-Step With Simultaneous Data Ingestion PM983	0.05%	97.63%	4,470	4	4	128	7	20	4	128	128	127	8	17,135	33,601,880	1,471	814,030	7,704
BERT 2000-Step With Simultaneous Data Ingestion PM9A3	0.04%	99.32%	24,311	4	4	256	10	31	4	128	1,280	127	12	6,949	62,821,436	16,860	411,639	2,602

COLLABORATE. INNOVATE. GROW.

SAMSUNG SDC23

Please take a moment to rate this session.

Your feedback is important to us.

COLLABORATE. INNOVATE. GROW.

