AI Scale at the Modern Era

R12631055 林東甫
R12631056 劉昕恩

What drove deep learning era?

More compute

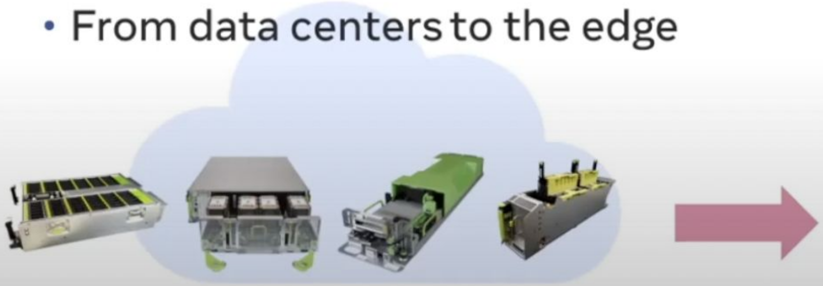
Better algorithm

Bigger and better data

Machine Learning at Facebook

- Machine learning is used extensively
  - Ranking posts
  - Content understanding
  - Object detection, segmentation, and tracking
  - Speech recognition/translation

- From data centers to the edge



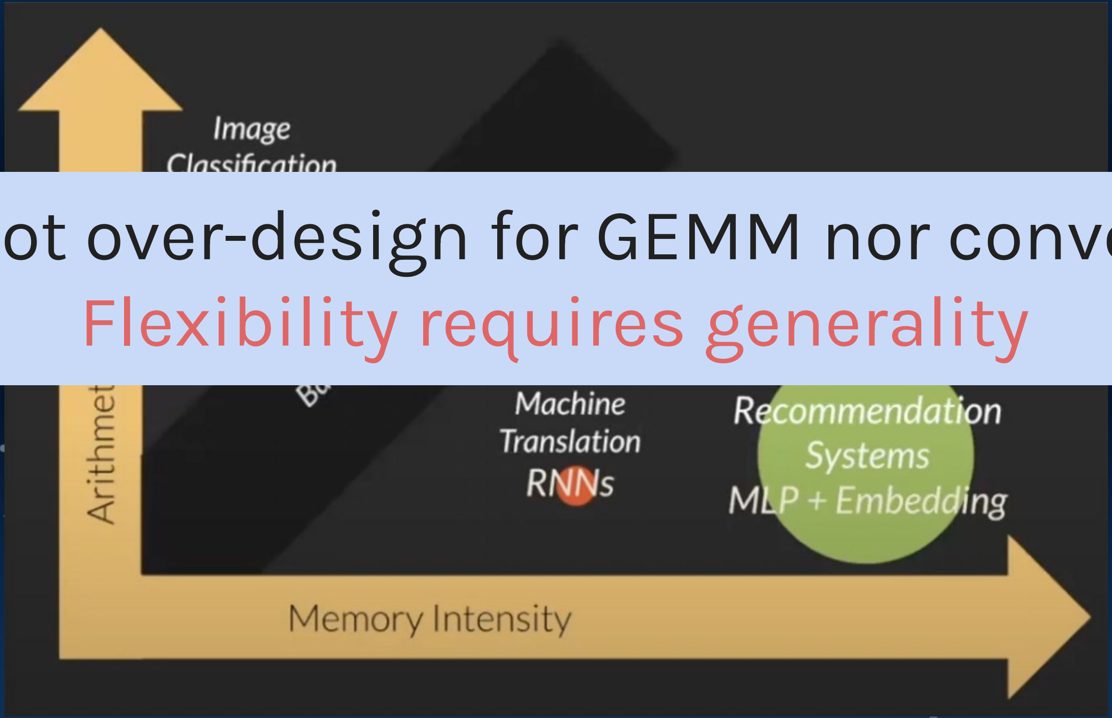
Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. Hazelwood et al. HPCA-2018.

Machine Learning Execution Flow

Data Scale at Facebook(and elsewhere)

XXX PB Replicated daily	**XX PB** Ingested daily	**X TB/s** Stream processing throughput	**XXX PB** Daily shuffle
X M Machines	**X EB** Warehouse size	**XX K** Pipelines	**X K** Pipeline authors

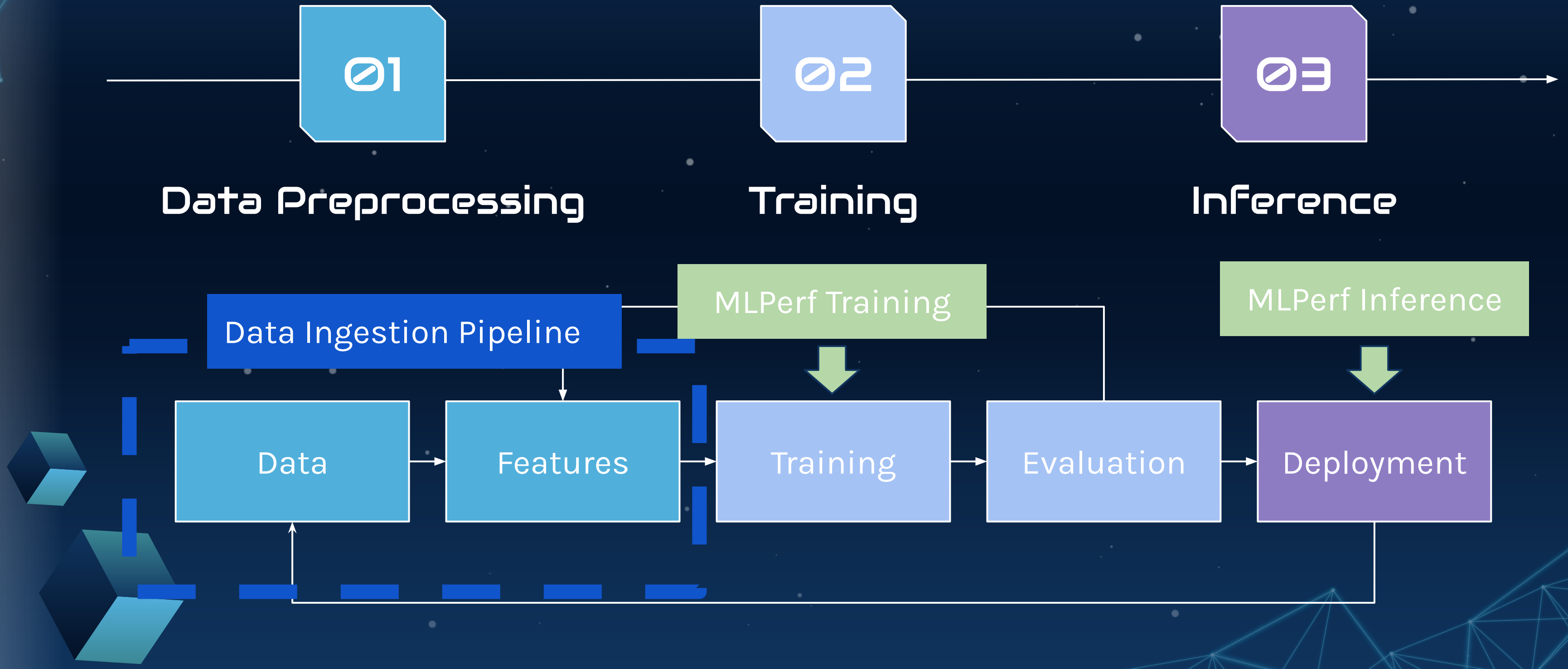
Diversity in DL Use Cases



Must not over-design for GEMM nor convolution

Flexibility requires generality

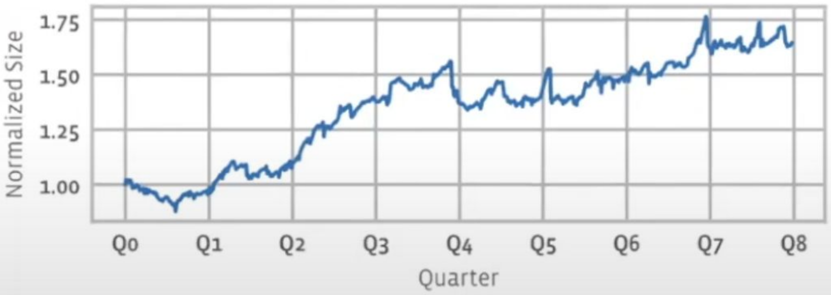
Machine Learning Execution Flow



Training Data and Feature Growth for Recommender System

Data Storage Growth

Training data for recommendation models has grown by 1.75x in 2 years



Model Memory Growth

Size of Facebook's production recommendations models has grown by an order of magnitude in 3 years[2]

Data Ingestion Pipeline

A Typical Data Ingestion Pipeline for MLPerf

Cloud Storage

Dataset downloaded to local storage

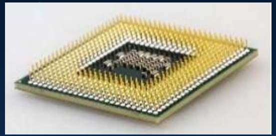
1 Local Storage

2 Host CPU

3 Training GPUs

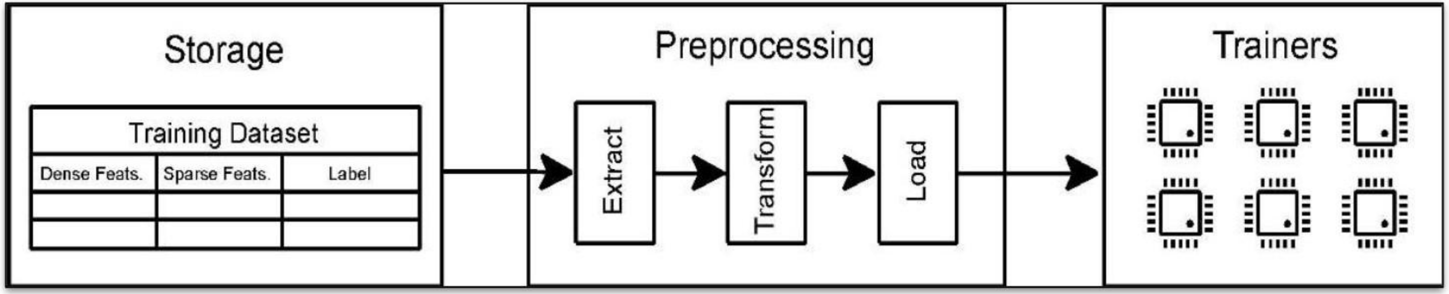
Raw batches read from local storage

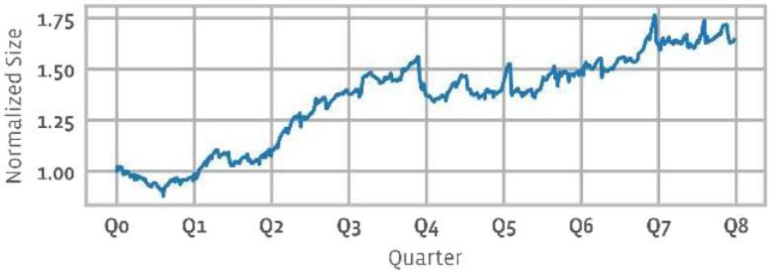
Preprocessed tensors loaded onto GPUs



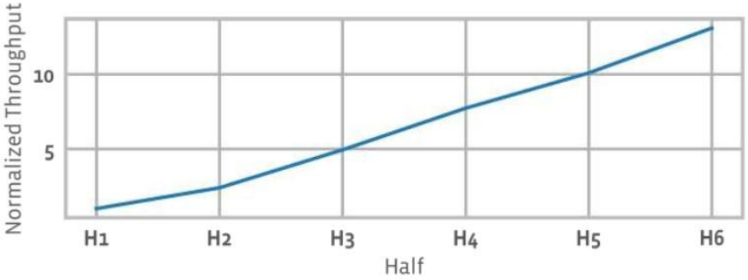
NVIDIA DGX

ML Training Storage growth @FB





~1.75x growth in training data ***storage size*** over past 2 years



~13x growth in training data ingestion ***throughput*** projected over 3 years

ML Training datasets cannot be stored locally on Trainers

Model	Table Size (PB)	Partition Size (PB)	Used Partition
RM1	13.45	0.15	11.95
RM2	29.18	0.32	25.94
RM3	2.93	0.07	1.95

ML Training Preprocessing @FB

Model	kQPS	Storage RX (GB/s)	Transform RX (GB/s)	Transform TX (GB/s)	# CPU Sockets required
RM1	11.623	0.8	1.37	0.68	24.16
RM2	7.995	1.2	0.96	0.50	9.44
RM3	36.921	0.8	1.01	0.22	55.22

ML training preprocessing compute requirements exceed trainer host capabilities

Local data storage and preprocessing doesn't work for us!

Cloud Storage

Dataset downloaded to local storage

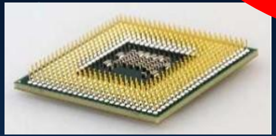
1

Local Storage

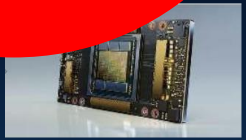
Raw batches read from local storage

2

Hos

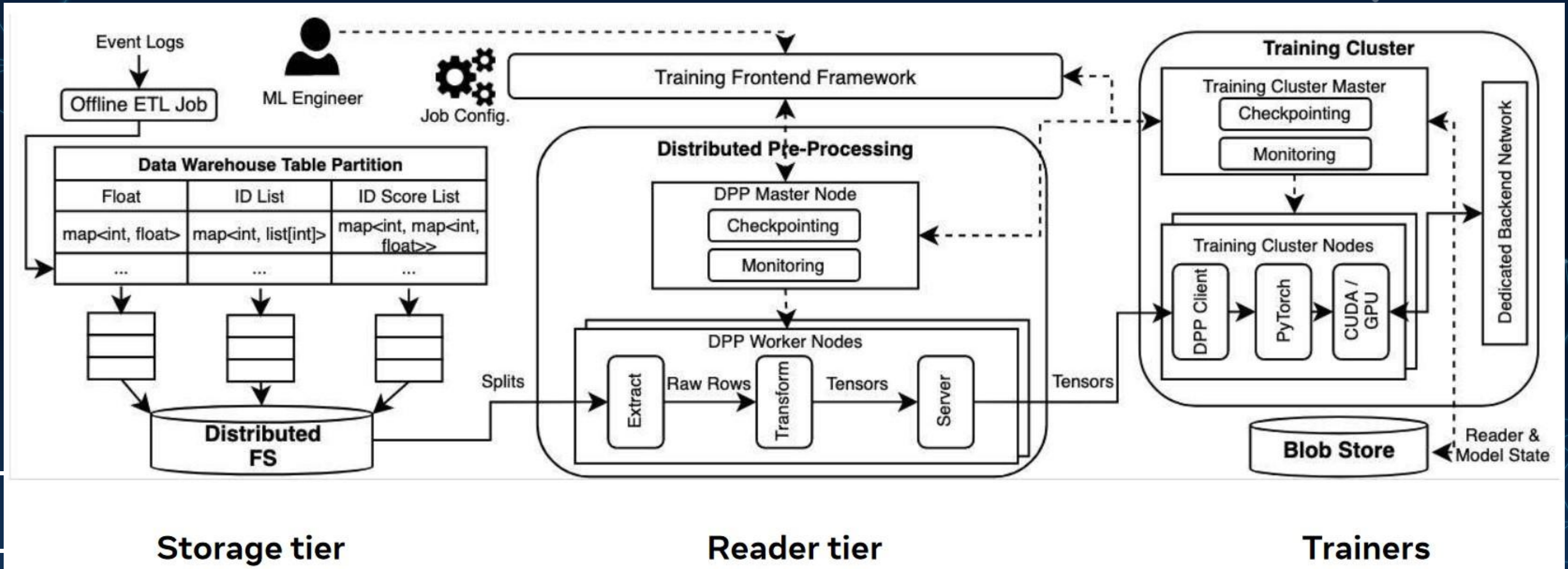


proces ed

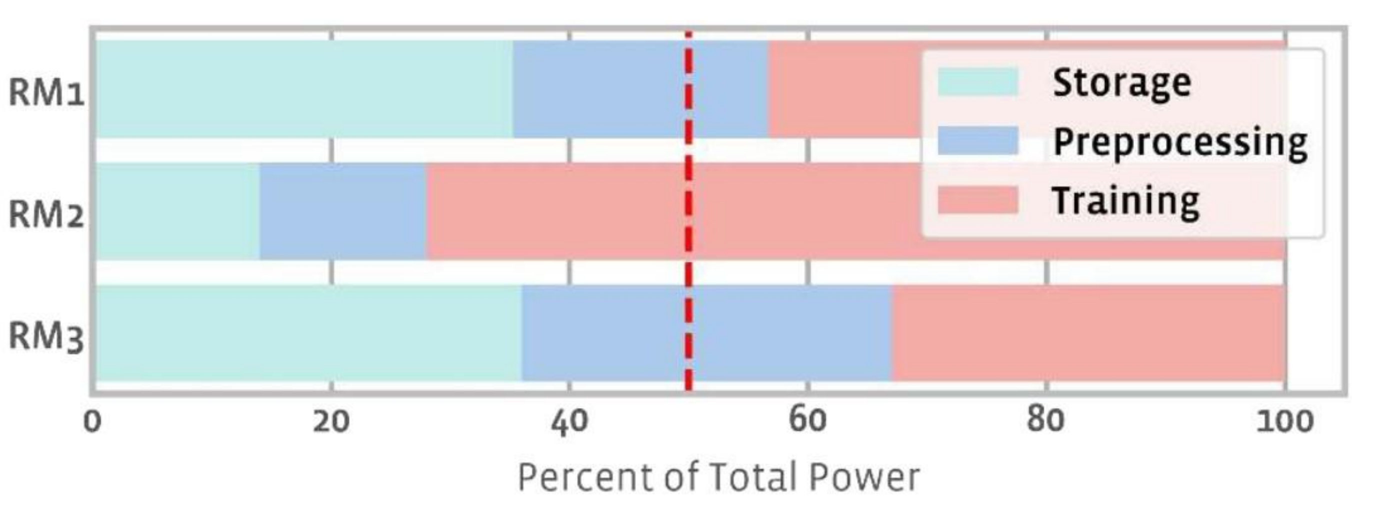


NVIDIA DGX

Disaggregated Training Data Ingestion @FB



Disaggregation is not enough: Training Data Ingestion Challenges



Data ingestion (Storage + Preprocessing) represents a significant, and growing, component of training capacity.

End-to-end Co-design for Data Ingestion Efficiency

Co-design

- **Feature Flattening** — Push data filtering to storage nodes
- **In-Memory Flatmap** — Optimized data formats
- **Merged Reads** — Increased disk throughput
- **Feature Reordering** — Mitigate unnecessary reads

Regular Map Reads

Hive Table

Row idx	Features (map<str: int>)
1	A: 1, B: 1, C: 3, D: 1, E: 3, F: 3
2	A: 2, B: 1, C: 2, D: 1, E: 2, F: 6

A: 1, B: 1, C: 3, D: 1, E: 3, F: 3

A: 2, B: 1, C: 2, D: 1, E: 2, F: 6

Read Features (A, D)

Entire rows are read

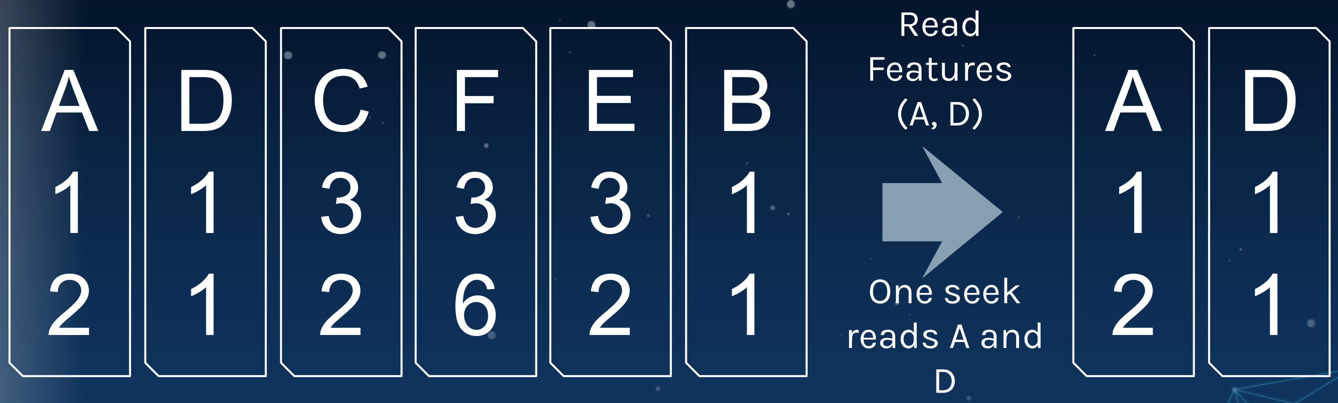
A: 1, B: 1, C: 3, D: 1, E: 3, F: 3

A: 1, B: 1, C: 3, D: 1, E: 3, F: 3

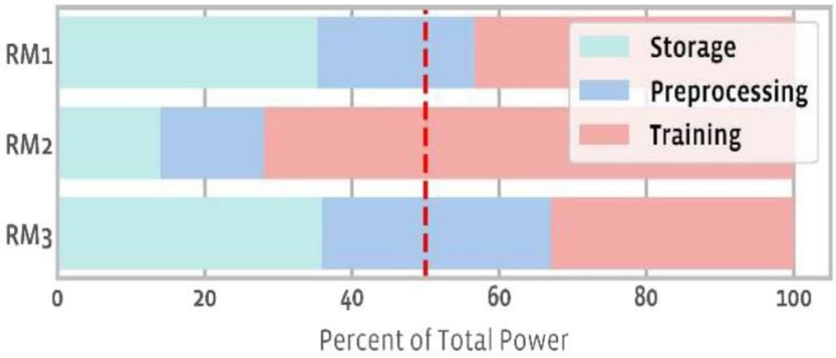
Feature Flattening + Merged Reads + Feature Reordering

Hive Table

Row idx	Features (map<str: int>)
1	A: 1, B: 1, C: 3, D: 1, E: 3, F: 3
2	A: 2, B: 1, C: 2, D: 1, E: 2, F: 6

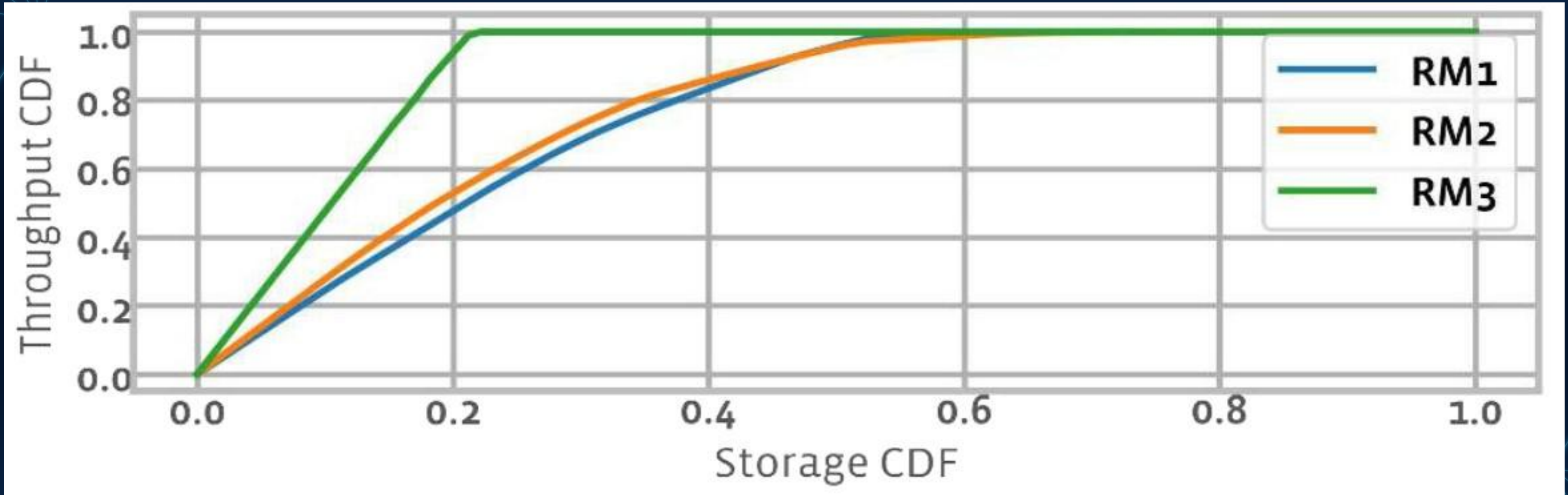


Training Data Efficiency Impact through co-design



2X power and cost savings for Data Ingestion

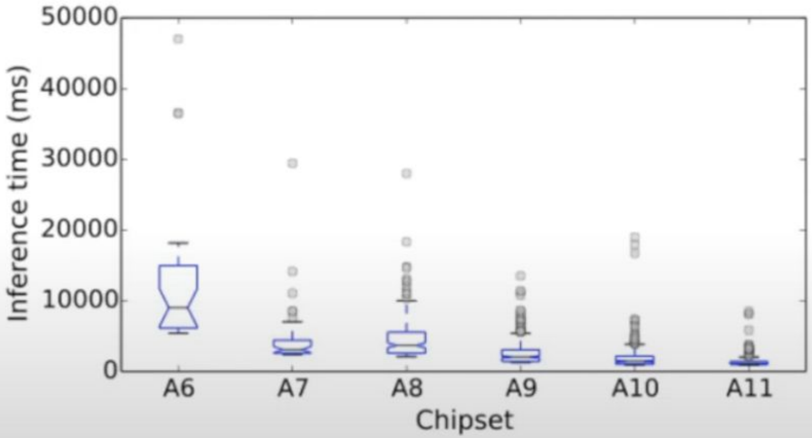
Future Opportunities: Training Data Reuse and Flash Caching



- A subset of bytes (20-40%) contribute to most of Storage IO
- Opportunity for Flash to absorb the IO more efficiently

High System Diversity for ML at the Edge

The diversity of mobile hardware and software is not found in the controlled datacenter environment.



Conclusion

Ever-Increasing AI Growth

Diverse ML System Requirement

Compute, Memory, Networking