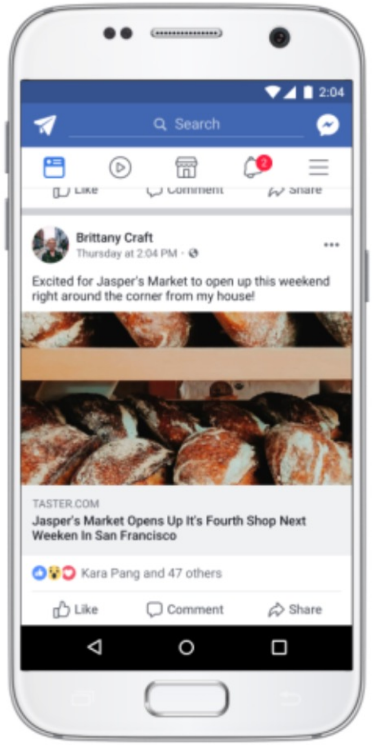
Check-n-Run: a Checkpointing System for Training Deep Learning Recommendation Models

Assaf Eisenman, Kiran Kumar Matam, Steven Ingram, Dheevatsa Mudigere,
Raghuraman Krishnamoorthi, Krishnakumar Nair, Misha Smelyanskiy, Murali Annavaram

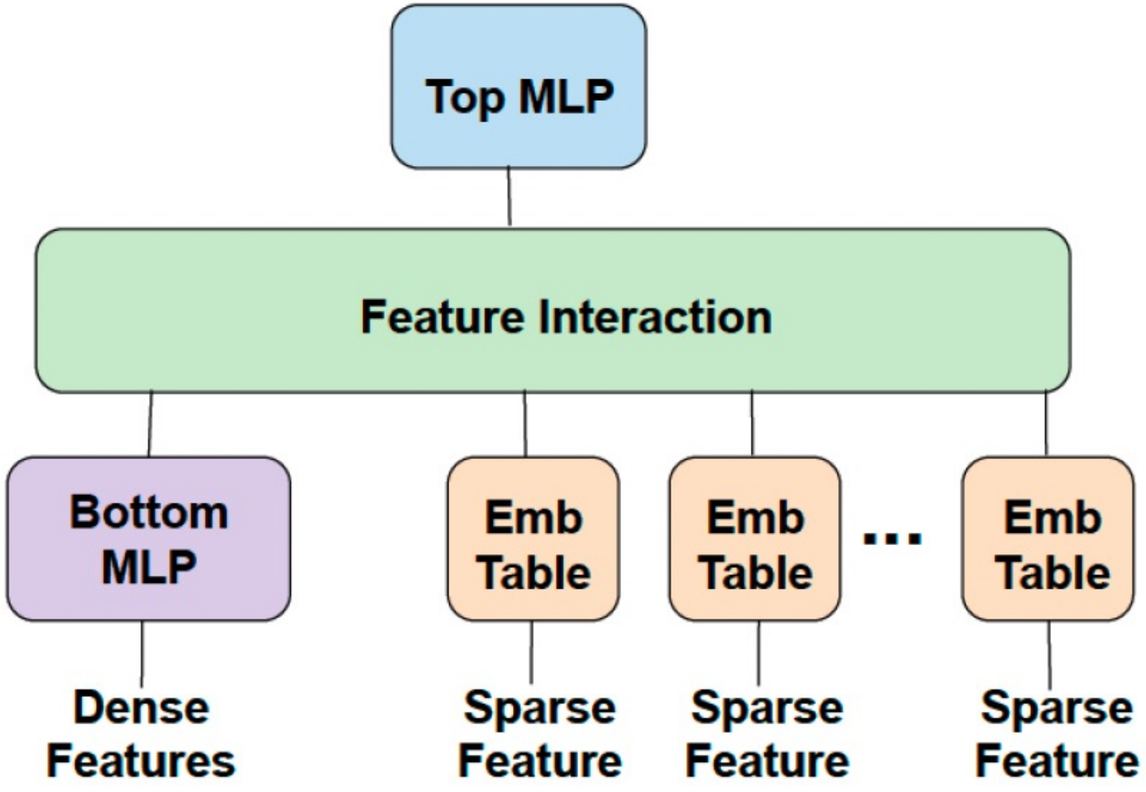
Meta

Recommendation Models are Important

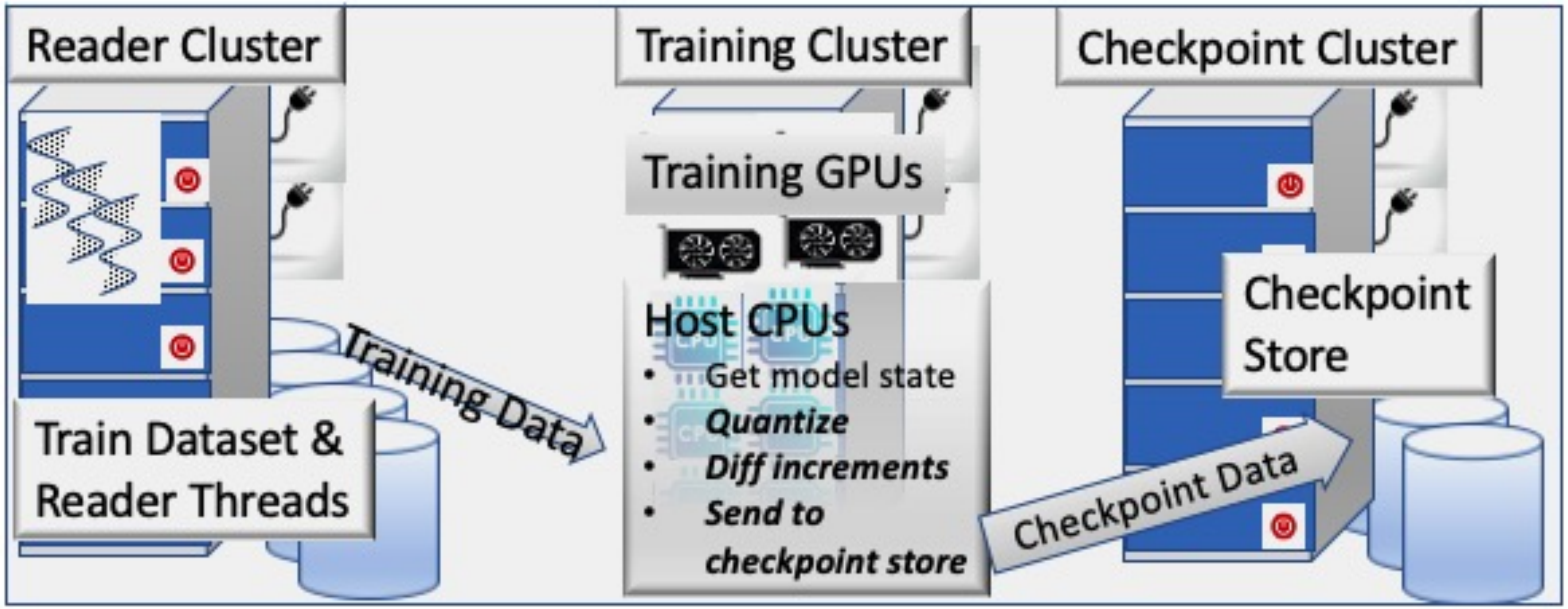
- Use cases include:
  - E-commerce marketplaces
  - Social media platforms
  - Entertainment services
- Consumes most of AI compute cycle at Meta
  - > 50% of training compute cycle
  - > 80% of inference compute cycles



Recommendation Model Architecture

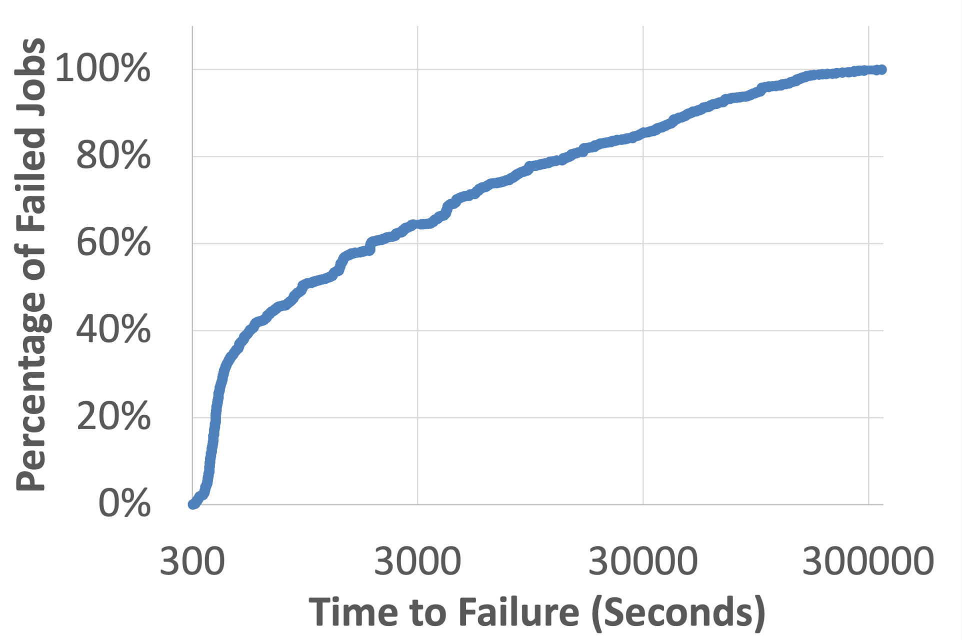


High Performance Training at Meta



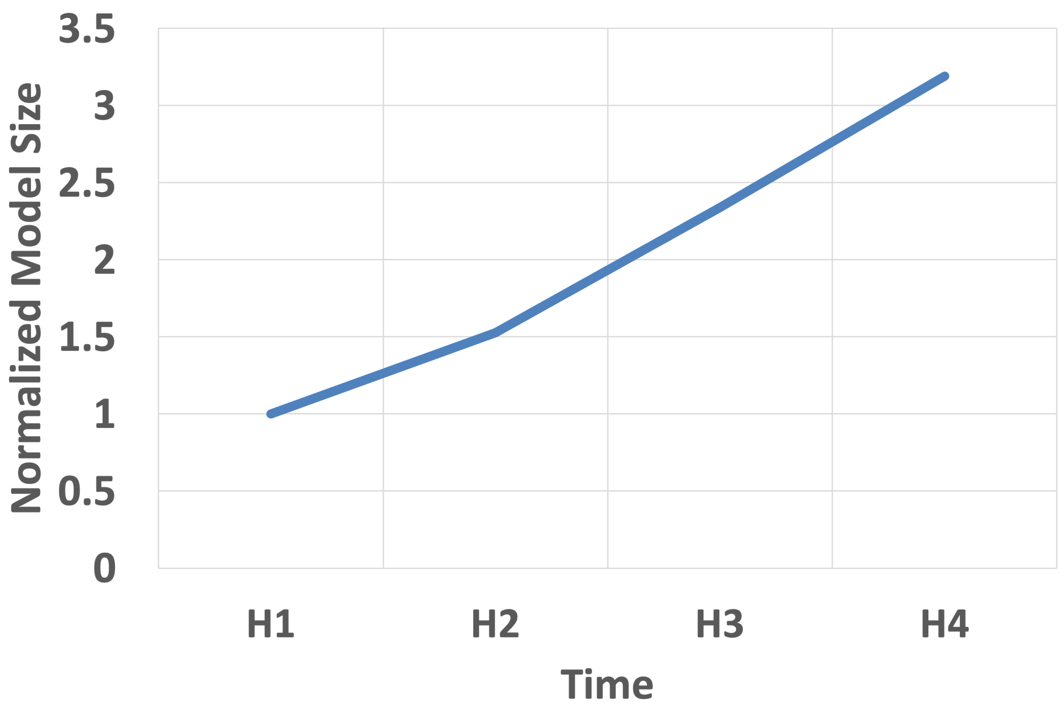
The Criticality of Checkpointing

- Failure recovery (ensure progress)
- Migrating training jobs
- Publishing snapshots
- Transfer learning



Checkpoint Challenges

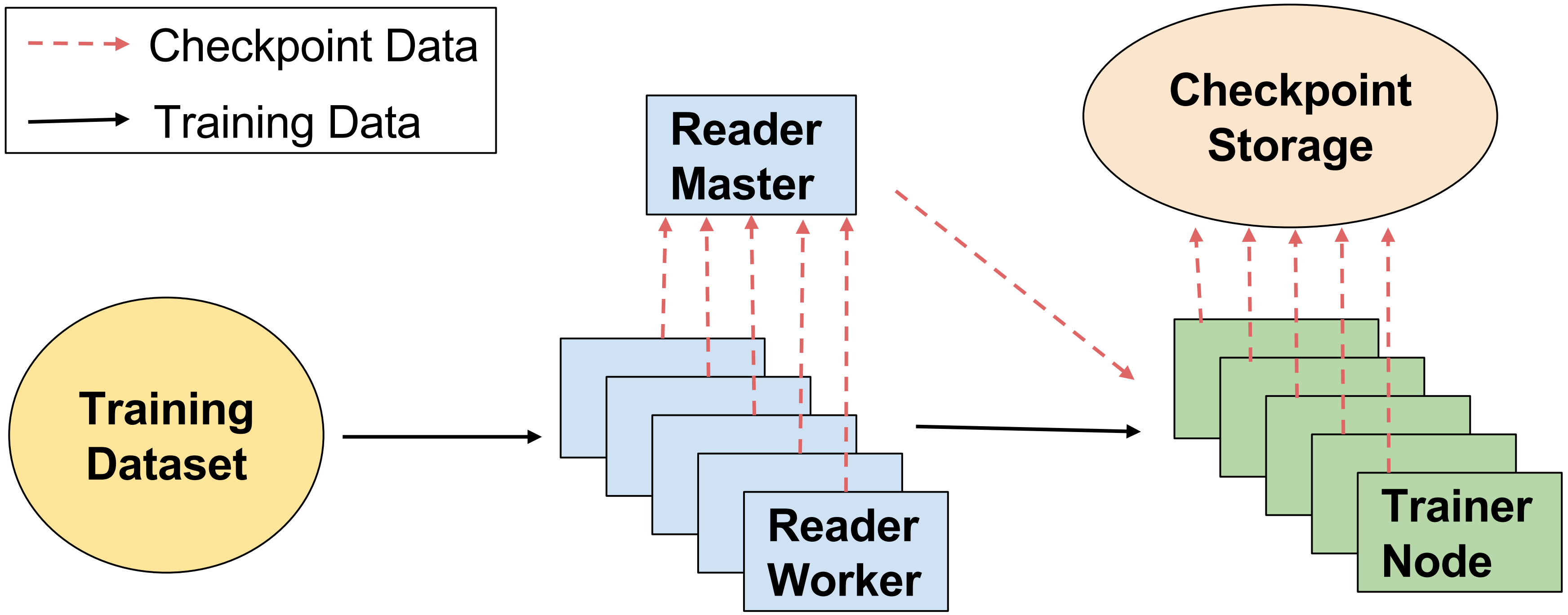
- Accuracy
- Frequency
- Write bandwidth
- Storage capacity



Check-n-Run

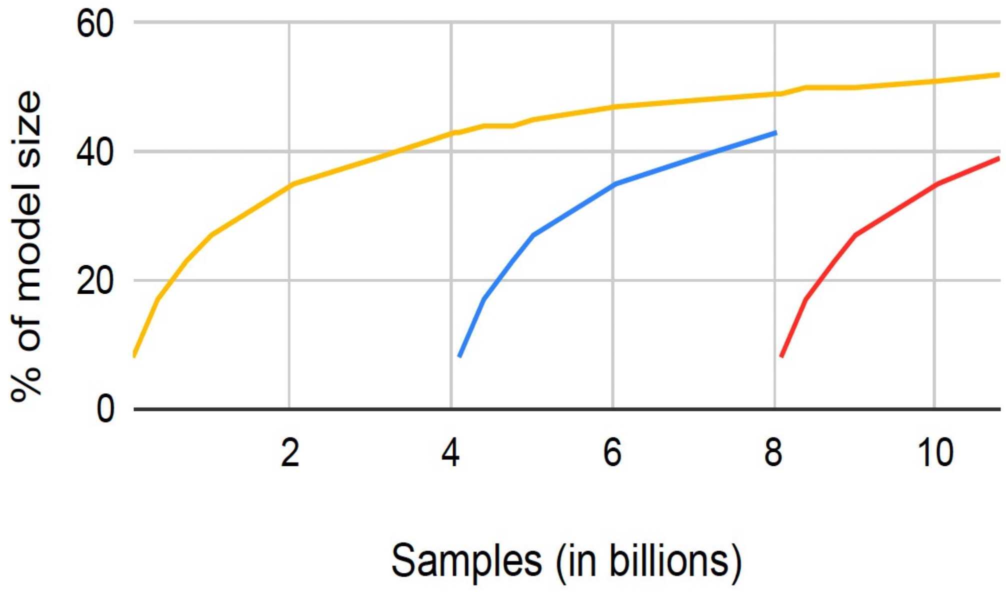
- Goal: a checkpointing system that significantly reduces the required write-bandwidth and storage capacity, without degrading accuracy
- What to Checkpoint?
- Decoupled Checkpointing
- Reducing write-bandwidth (WB) and storage capacity

Checkpointing Workflow



Reducing WB with Differential Checkpointing

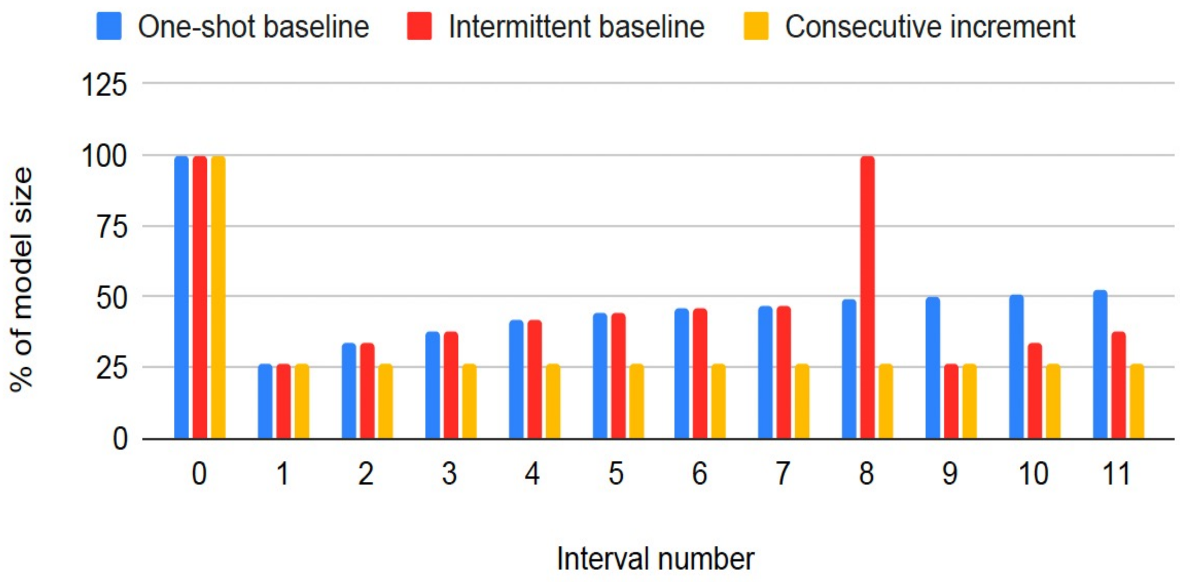
- Motivation: model accesses are sparse



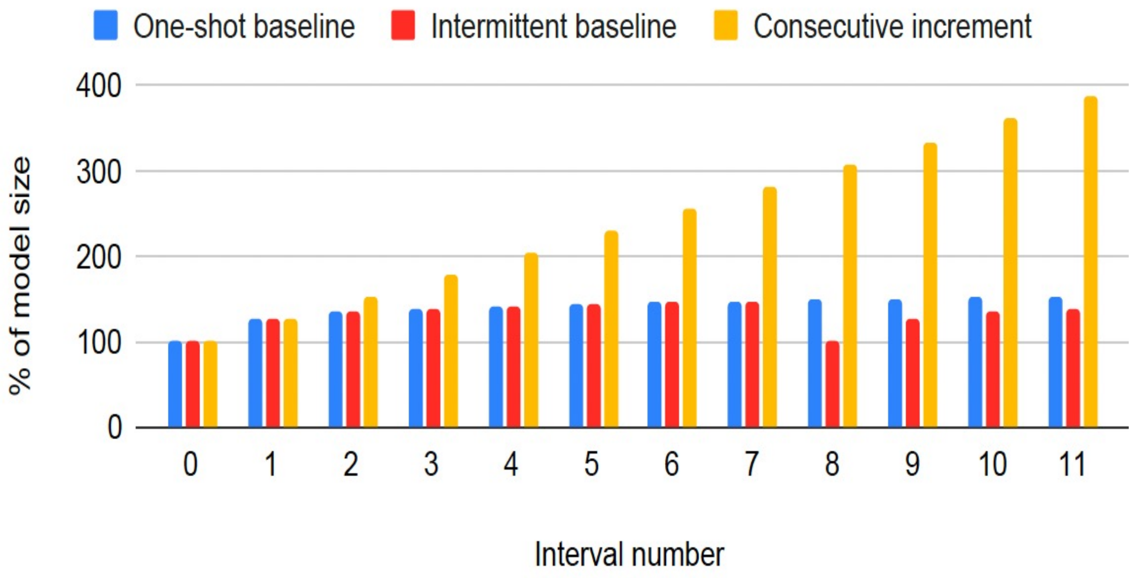
Approaches for Differential Checkpointing

- One-Shot Differential Checkpoint
- Consecutive Incremental Checkpoint
- Intermittent Differential Checkpoint

Write Bandwidth:



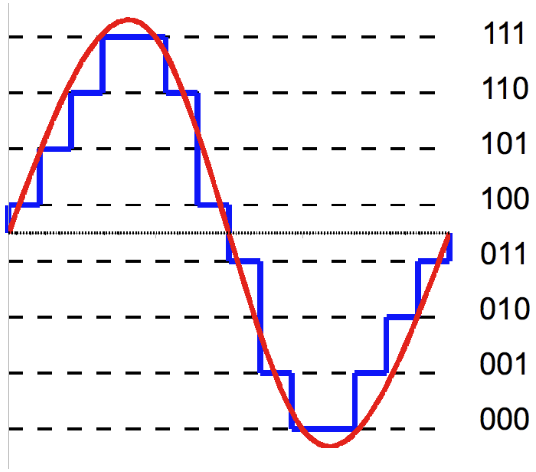
Storage:



Checkpoint Quantization

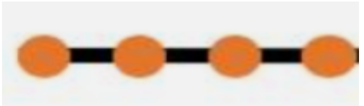
- Compress checkpoint without degrading training accuracy

- Approaches:



…	..	…	…	…
0.21	-0.31	0.03	0.01	-0.05
…	..	…	…	…
…	..	…	…	…

Uniform:



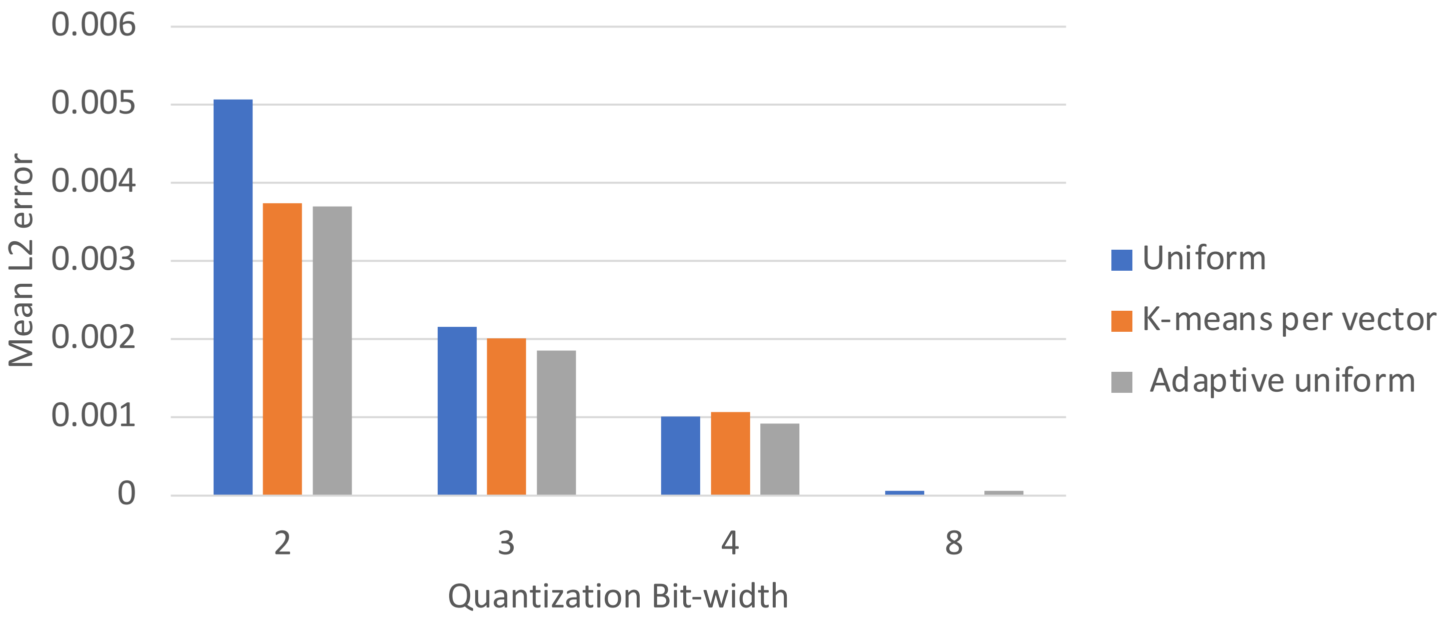
(-0.31) () (-0.05, 0.03, 0.01) (0.21)

Non-Uniform:

(-0.31) (-0.05) (0.03, 0.01) (0.21)

Comparing Quantization Strategies

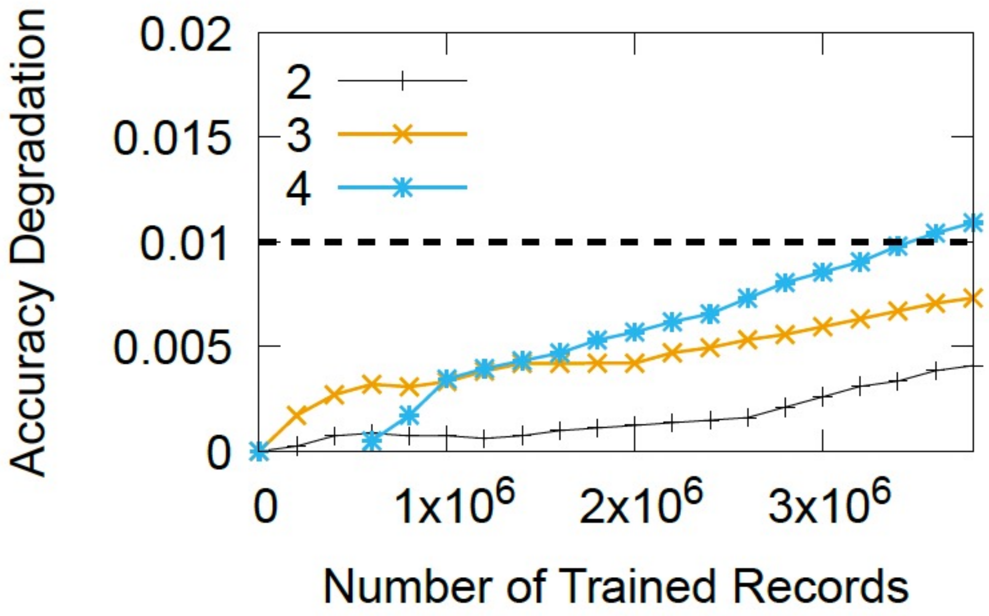
- Uniform quantization
- Non-uniform quantization using k-means
- Adaptive uniform quantization



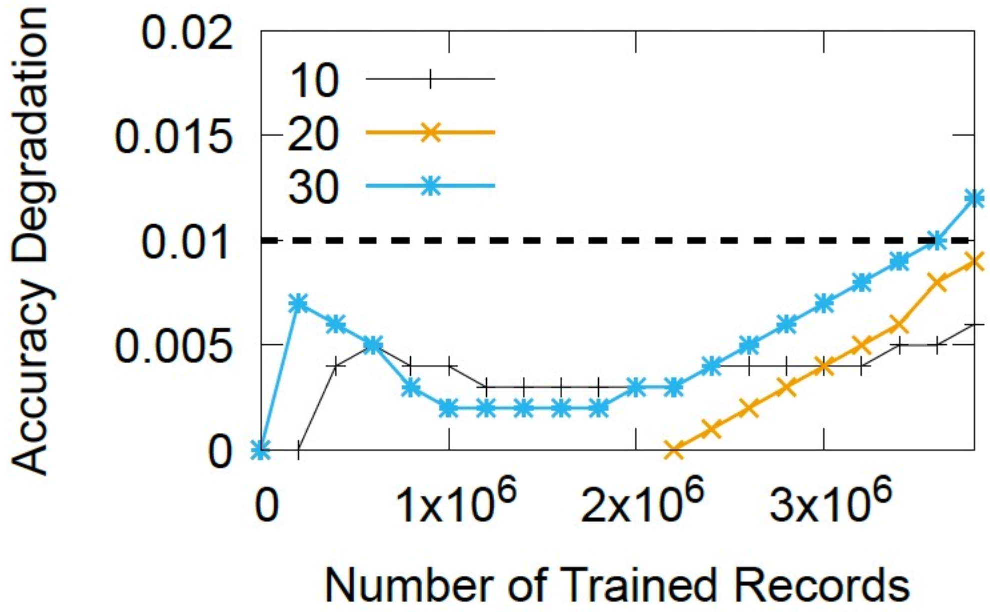
Quantization Bit-width Selection

- Quantization error may accumulate
- Select bit-width based on the probability of a failure

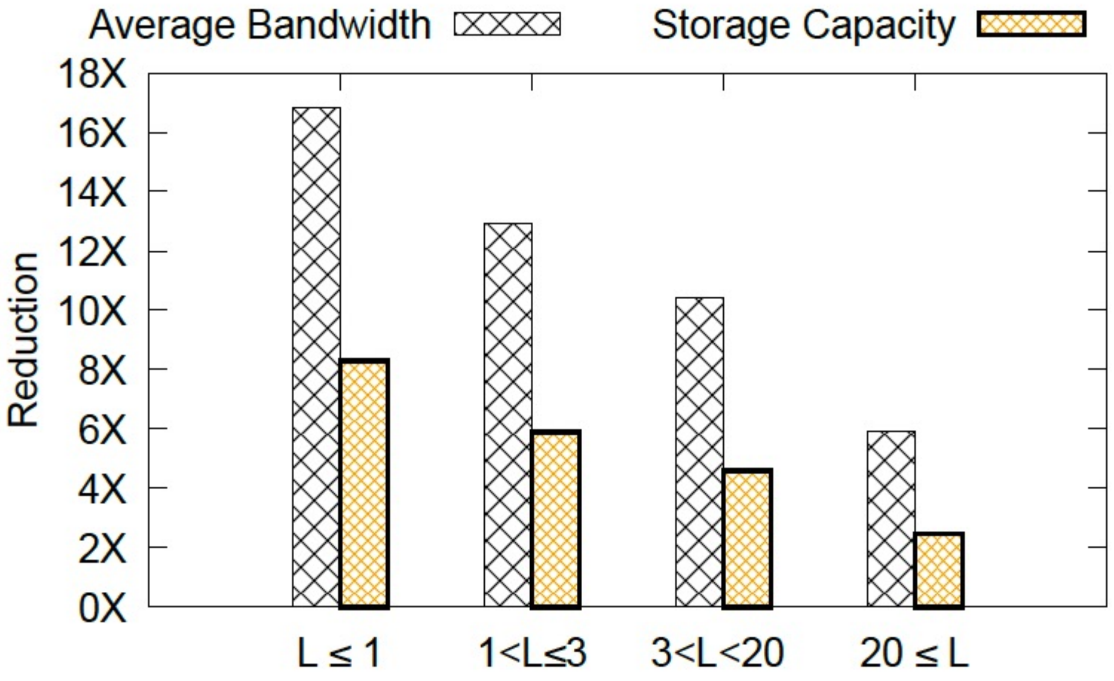
3-Bit Checkpoint Quantization:



4-Bit Checkpoint Quantization:



Overall Reduction



Summary

- The checkpointing of large recommendation systems at scale is challenging
- Check-n-run:
  - High performance checkpointing
  - Significantly reduces the required write-bandwidth and storage capacity
- Questions? aeisenman@fb.com