Inference Accelerator Deployment at Meta

Cao Gao, Software Engineer
Presenting the work of many people across Meta

Meta AI

Agenda

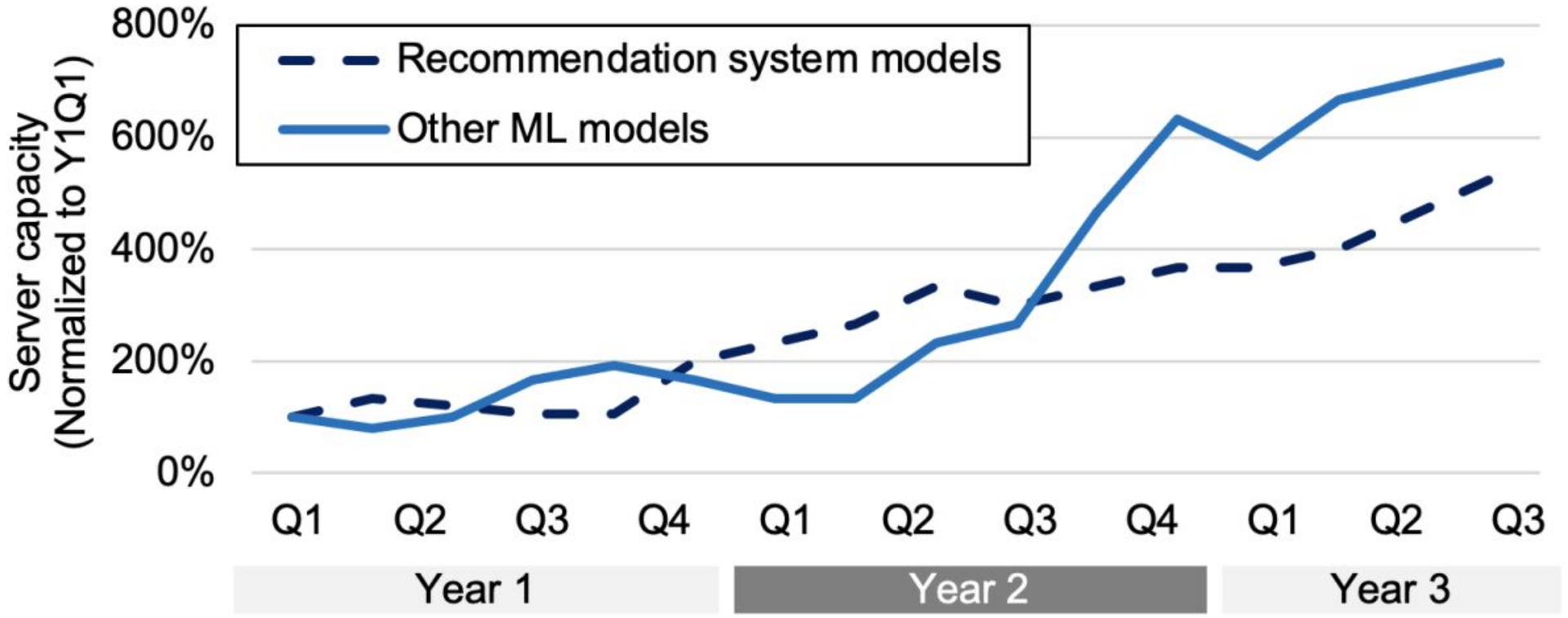
01 Introduction

02 Workloads and Requirements

03 Software / hardware co-design

04 Results and discussion

Tremendous AI growth



deploy inference accelerators to meet growing demand and model complexity

Inference workloads

TABLE I: Model Characteristics.

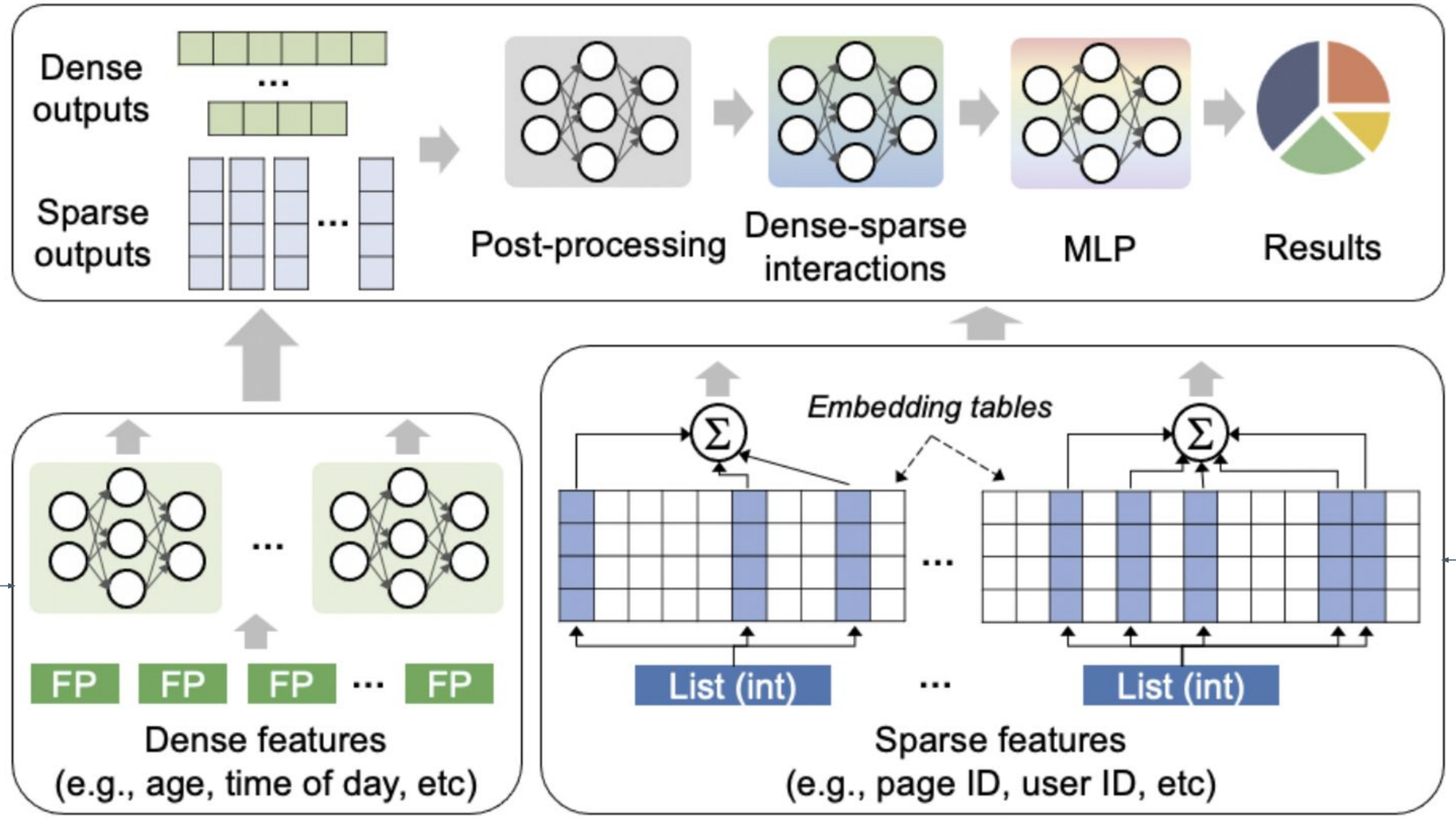
Category	Model name	Model Size (MParams)	FLOPs (GFLOPs per batch)	Typical Batch Size	Arith. Intensity (Weights+Activations)	Latency Constraints (ms)
Recommendation	Less complex model	70,000	0.02	32-64	90	100 (per 150-180 items)
	More complex model	>100,000	0.1	32-64	80	100 (per 150-180 items)
Computer Vision	ResNeXt101-32x4-48	44	15.6	1 image	355	∼1000
	RegNetY	700	256	1 image	395	∼1000
	FBNetV3 based model	28.6	72	1 image	1946	∼300
Video Understanding	ResNeXt3D based	58	3.4	4 frames	362	∼350
Natural Language Processing	XLM-R	558	20 (32 tokens)	20-70 tokens (1 sentence)	#tokens (20-70)	∼200

Inference workloads: DLRM

TABLE I: Model Characteristics.

Category	Model name	Model Size (MParams)	FLOPs (GFLOPs per batch)	Typical Batch Size	Arith. Intensity (Weights+Activations)	Latency Constraints (ms)
Recommendation	Less complex model	70,000	0.02	32-64	90	100 (per 150-180 items)
	More complex model	>100,000	0.1	32-64	80	100 (per 150-180 items)
Computer Vision	ResNeXt101-32x4-48	44	15.6	1 image	355	~1000
	RegNetY	700	256	1 image	395	~1000
	FBNetV3 based model	28.6	72	1 image	1946	~300
Video Understanding	ResNeXt3D based	58	3.4	4 frames	362	~350
Natural Language Processing	XLM-R	558	20 (32 tokens)	20-70 tokens (1 sentence)	#tokens (20-70)	~200

Inference workloads: DLRM



Arithmetic intensity is low

Benefit from on-chip memory

1000s of embedding tables, varying size (MB ~ TB), require fast look-up

Mem capacity and BW intensive

Input features can be fairly large

I/O intensive

Inference workloads: Content Understanding

TABLE I: Model Characteristics.

Category	Model name	Model Size (MParams)	FLOPs (GFLOPs per batch)	Typical Batch Size	Arith. Intensity (Weights+Activations)	Latency Constraints (ms)
Recommendation	Less complex model	70,000	0.02	32-64	90	100 (per 150-180 items)
	More complex model	>100,000	0.1	32-64	80	100 (per 150-180 items)
Computer Vision	ResNeXt101-32x4-48	44	15.6	1 image	355	~1000
	RegNetY	700	256	1 image	395	~1000
	FBNetV3 based model	28.6	72	1 image	1946	~300
Video Understanding	ResNeXt3D based	58	3.4	4 frames	362	~350
Natural Language Processing	XLM-R	558	20 (32 tokens)	20-70 tokens (1 sentence)	#tokens (20-70)	~200

Computer vision models: heavy in compute

Video understanding: 3D convolutions and bandwidth bound operations

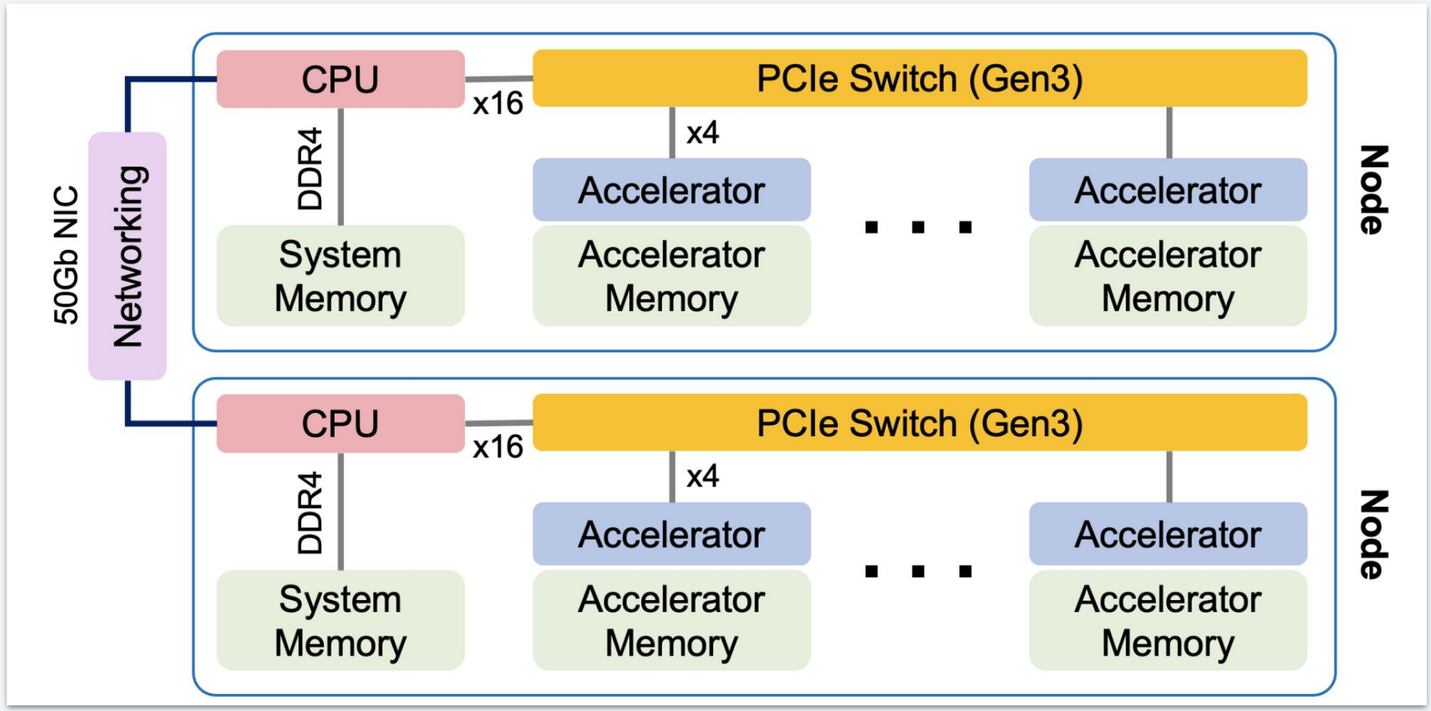
Natural Language Processing: dominated by matrix-multiple compute; however, arithmetic intensity is not high

Design requirements

- High perf/watt ratio
- Support large models (e.g. state-of-the-art DLRM)
- Adapt to fast-evolving ML framework landscape (e.g. PyTorch)

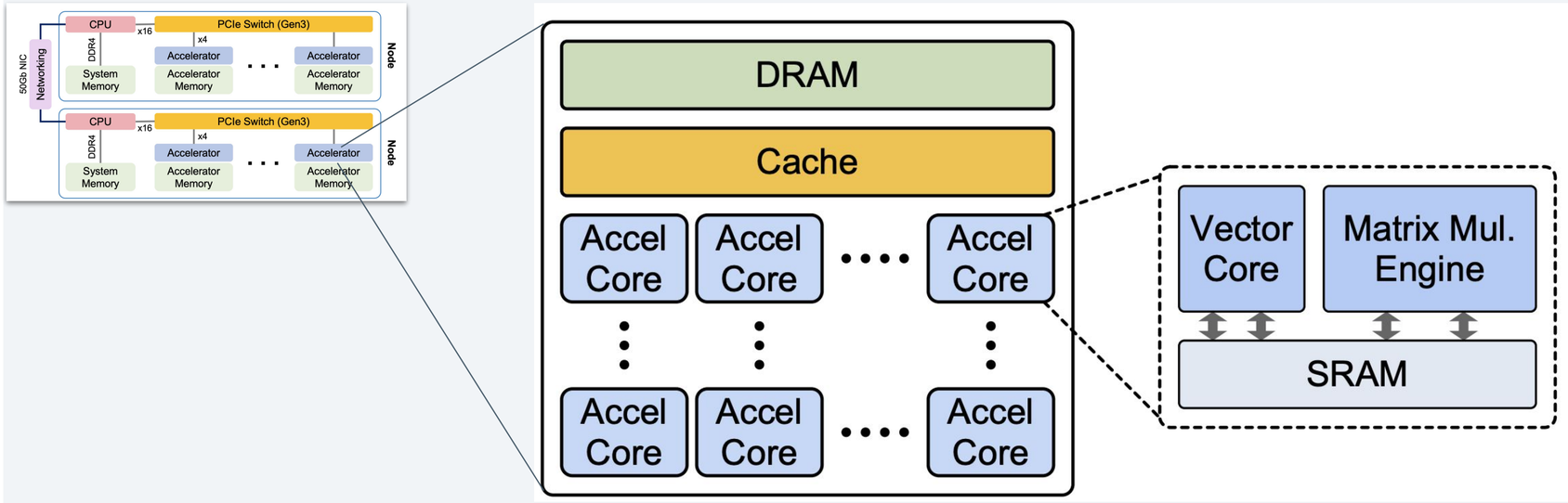
System design

- Designed for high perf/W
- Provide enough memory bandwidth for large models with a multi-card setup
- Peer to peer communication through PCIe



Accelerator card

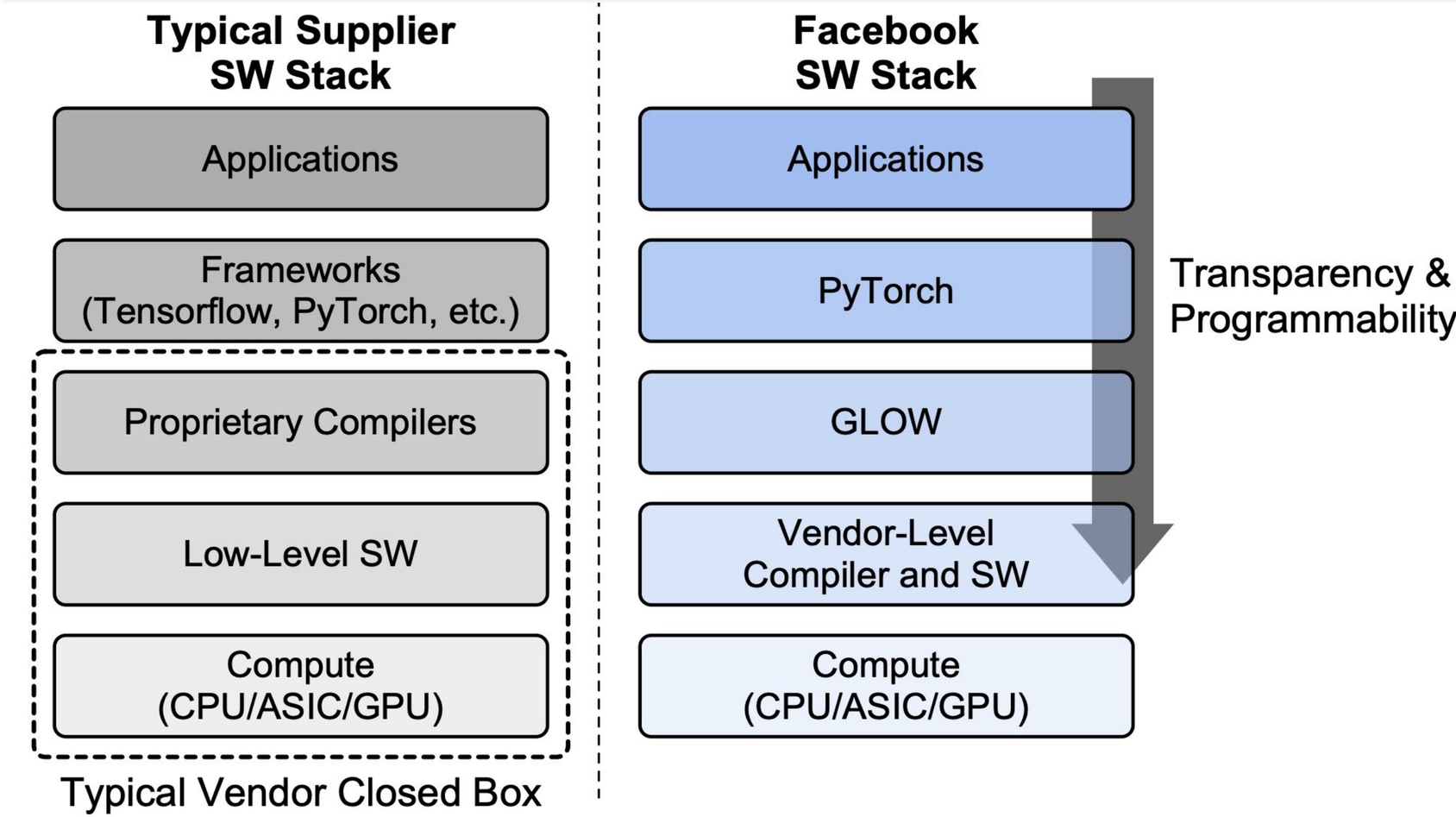
- Dedicated core for matrix multiplies and convolutions
- On-chip storage for fast memory access



Software Design

Transparency & programmability through Glow

- Glow runtime and compiler
- Provides APIs to expose vendor-specific logics



Openness

- Hardware specifications available through Open Compute Project (OCP)
- Open-source frameworks (PyTorch, Caffe2) and AI accelerator tooling (Glow, onnx)
- Open-source benchmarks (DLRM, PyText, Classy Vision, etc)

Numerics

- Leverage low-precision numerics to achieve high compute throughput, within strict accuracy constraints

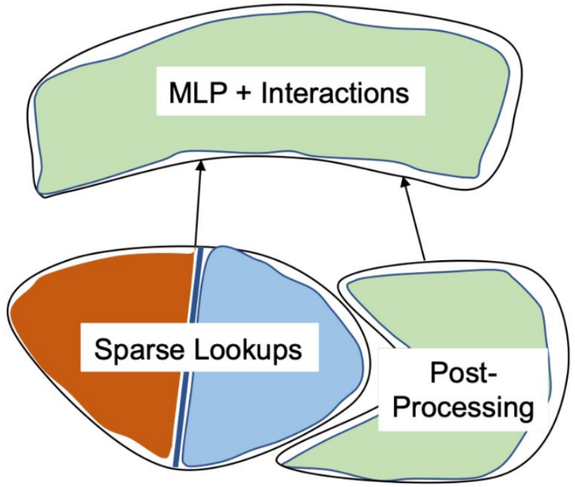
- Quantization
  - Recommendation mode: per-layer quantization for both FC and EmbeddingBag
  - CV and NLP: int8 channel-wise quantization

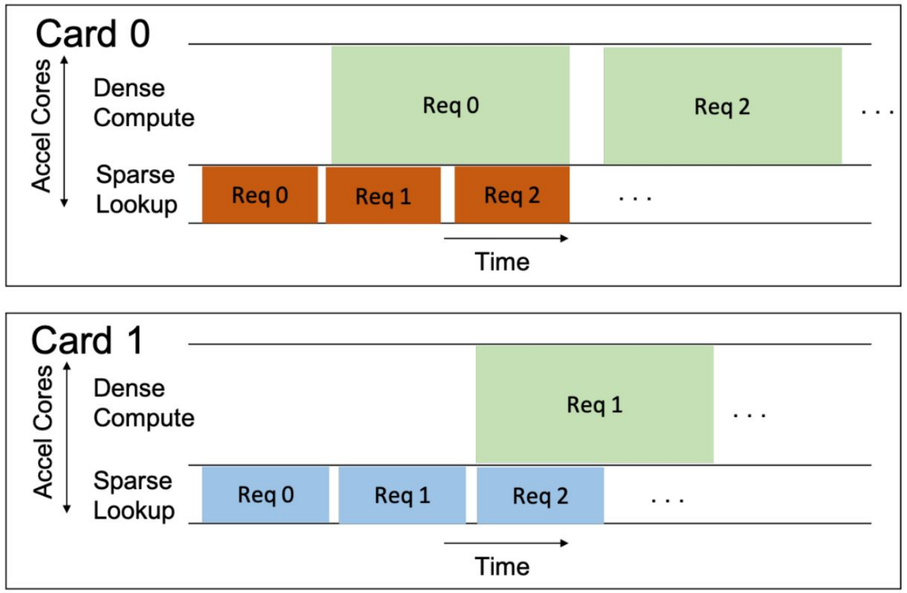
- Numerical validation
  - CPU reference kernel
  - Continuous accuracy monitoring

Performance Optimization

Model level optimizations

- Quantization
- Net-split between host and device
- Variable-size padding





Card-level optimizations

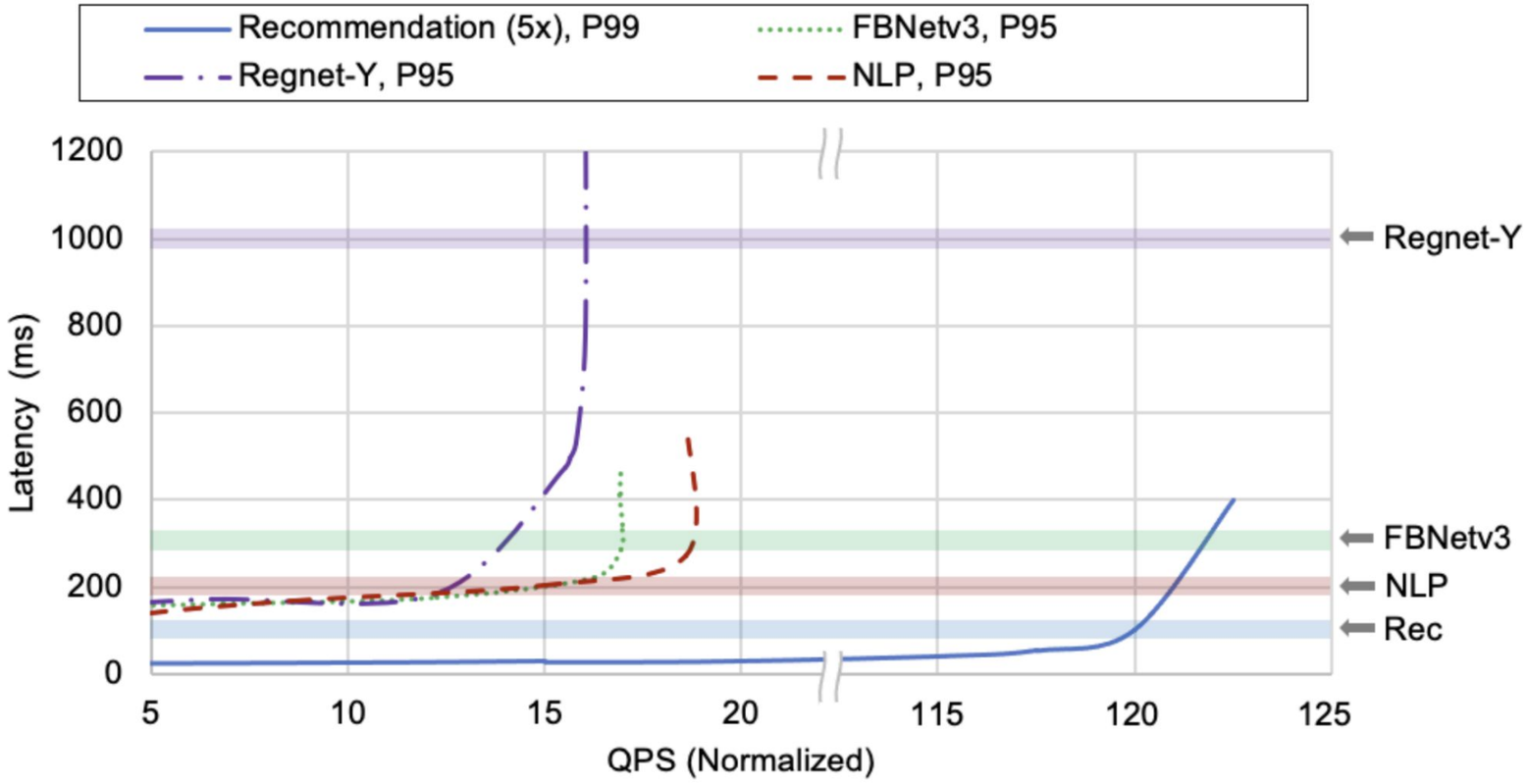
- Partitioning the model across accelerators
- Resource allocation

System level optimizations

- Partial tensor transfers
- Command batching

Results

The accelerator is able to serve all of these complex models (5X than previous models) within the latency budget.



Going forward....

- Much larger / complex models
- Numerics support and numerics/performance co-design -- reduced precision, sparsity, embedding table compression
- Accelerator programmability

Michael Anderson, Benny Chen, Stephen Chen, Summer Deng, Jordan Fix, Michael Gschwind, Aravind Kalaiah, Changkyu Kim, Jaewon Lee, Jason Liang, Haixin Liu, Yinghai Lu, Jack Montgomery, Arun Moorthy, Satish Nadathur, Sam Naghshineh, Avinash Nayak, Jongsoo Park, Chris Petersen, Martin Schatz, Narayanan Sundaram, Bangsheng Tang, Peter Tang, Amy Yang, Jiecao Yu, Hector Yuen, Ying Zhang, Aravind Anbudurai, Vandana Balan, Harsha Bojja, Joe Boyd, Matthew Breitbach, Claudio Caldato, Anna Calvo, Garret Catron, Sneh Chandwani, Panos Christeas, Brad Cottel, Brian Coutinho, Arun Dalli, Abhishek Dhanotia, Oniel Duncan, Roman Dzhabarov, Simon Elmir, Chunli Fu, Wenyin Fu, Michael Fulthorp, Adi Gangidi, Nick Gibson, Sean Gordon, Beatriz Padilla Hernandez, Daniel Ho, Yu-Cheng Huang, Olof Johansson, Shishir Juluri, Shobhit Kanaujia, Manali Kesarkar, Jonathan Killinger, Ben Kim, Rohan Kulkarni, Meghan Lele, Huayu Li, Huamin Li, Yueming Li, Cynthia Liu, Jerry Liu, Bert Maher, Chandra Mallipedi, Seema Mangla, Kiran Kumar Matam, Jubin Mehta, Shobhit Mehta, Christopher Mitchell, Bharath Muthiah, Nitin Nagarkatte, Ashwin Narasimha, Bernard Nguyen, Thiara Ortiz, Soumya Padmanabha, Deng Pan, Ashwin Poojary, Ye (Charlotte) Qi, Olivier Raginel, Dwarak Rajagopal, Tristan Rice, Craig Ross, Nadav Rotem, Scott Russ, Kushal Shah, Baohua Shan, Hao Shen, Pavan Shetty, Krish Skandakumaran, Kutta Srinivasan, Roshan Sumbaly, Michael Tauberg, Mor Tzur, Sidharth Verma, Hao Wang, Man Wang, Ben Wei, Alex Xia, Chenyu Xu, Martin Yang, Kai Zhang, Ruoxi Zhang, Ming Zhao, Whitney Zhao, Rui Zhu, Ajit Mathews, Lin Qiao, Misha Smelyanskiy, Bill Jia, Vijay Rao