ML Commons

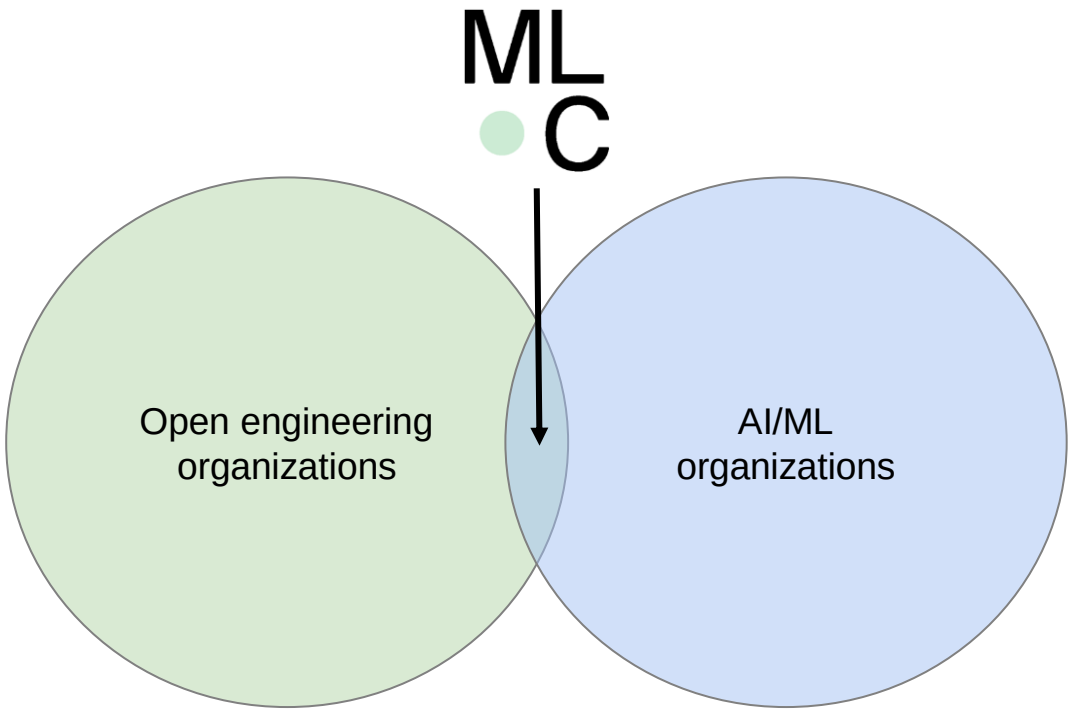
David Kanter
Executive Director
david@mlcommons.org

September 12, 2023

MLPerf: Building industry standards for LLMs

Introduction

We need an **open engineering organization** to create better ML for everyone



MLCommons is a global community

Academic Institutions

- Harvard University
- Polytechnique Montreal
- Peng Cheng Laboratory
- Stanford University
- University of California, Berkeley
- University of Toronto
- University of Tübingen
- University of Virginia
- University of York, UK
- Yonsei University
- York University, Canada

Benchmarks drive progress and transparency

“What get measured, gets improved.” — Peter Drucker

Benchmarking aligns the entire community in pursuit of the same clear objective.

MLPerf Goals

Enforce performance result replicability to ensure reliable results

Use **representative workloads** reflecting production use-cases

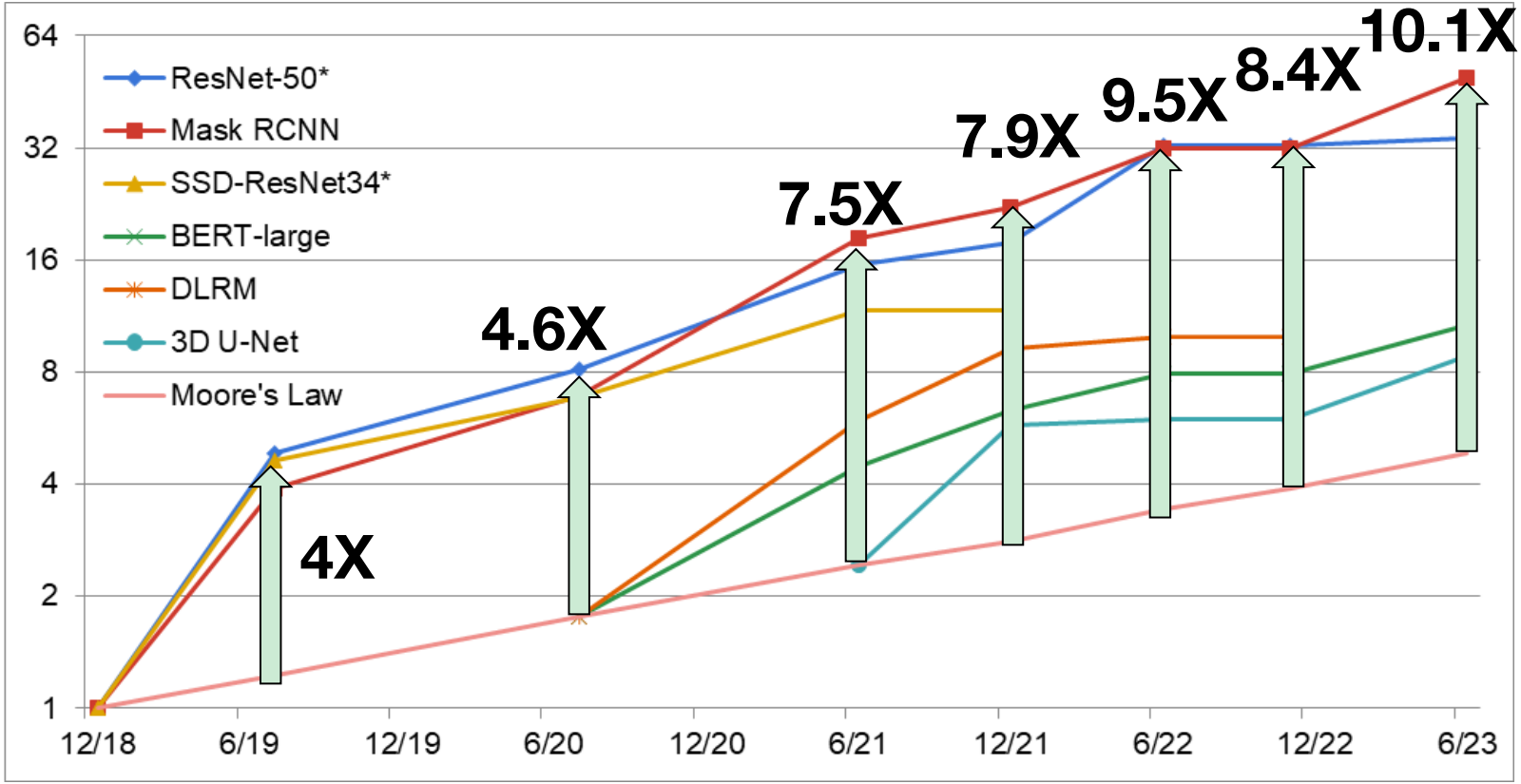
Encourage innovation to improve the state-of-the-art of ML

Accelerate progress in ML via **fair and useful measurement**

Serve both the **commercial and research communities**

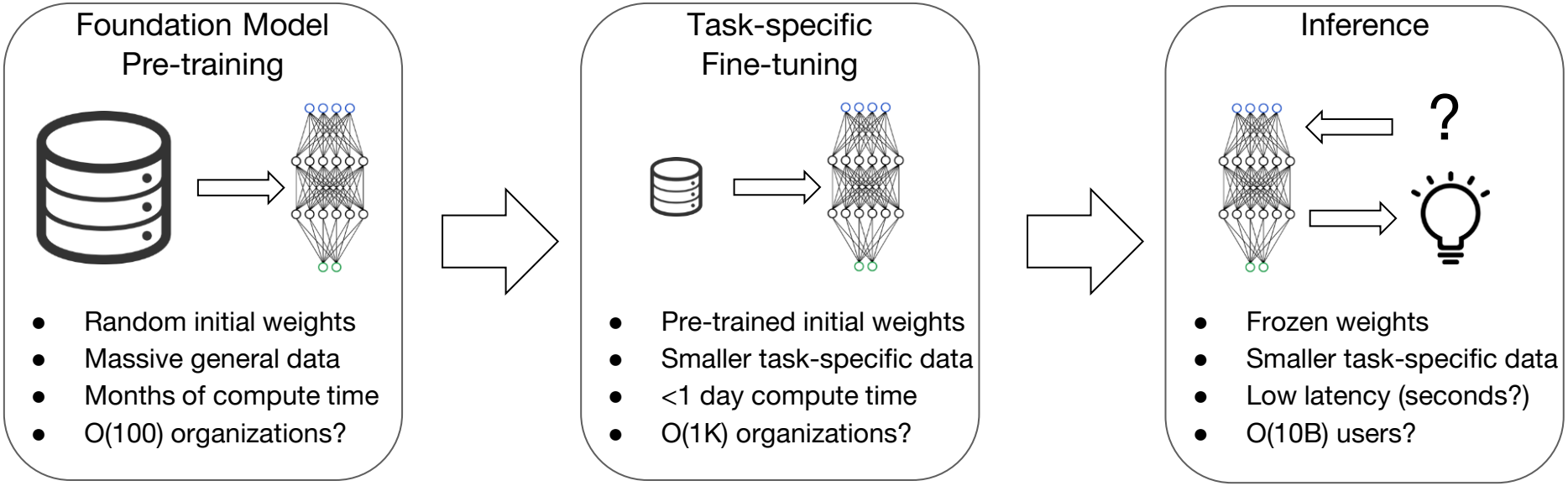
Keep **benchmarking affordable** so that all can participate

MLPerf Training - Ahead of Moore's Law



*indicates benchmark increased accuracy target in MLPerf Training v0.6

LLM Lifecycle and MLPerf Benchmarks



- MLPerf Training v3.0 includes a benchmark for Foundation Model pre-training (GPT3 175B)
- Future MLPerf Training round will include a benchmark for fine-tuning an LLM
- MLPerf Inference v3.1 includes a benchmark for LLM inference (GPT-J 6B)

LLM Training Benchmark Details

C4 Dataset

- c4/en/3.0.1 - Colossal, Cleaned, Common Crawl corpus dataset hosted on HuggingFace
  - 305GB and ~174B tokens

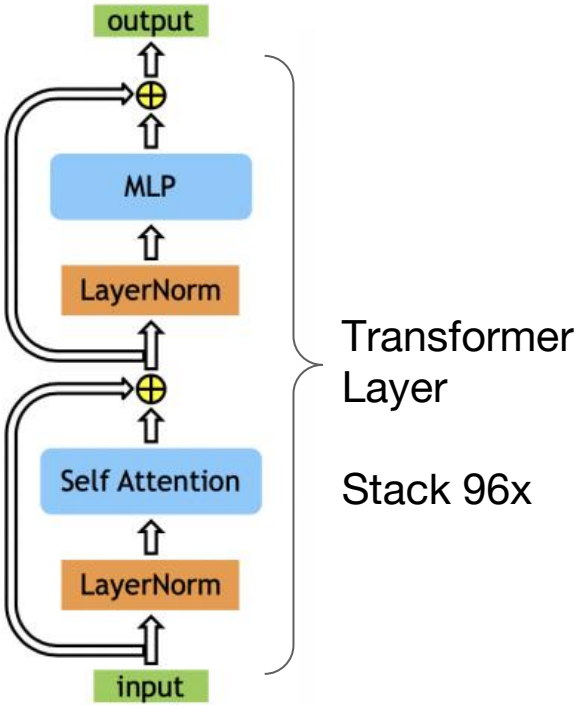
- Benchmark trains on a portion of the training set
  - Start from an initial checkpoint that is trained on 12.5B tokens
  - Benchmark measures training on ~1.3B tokens
- Model Accuracy evaluation happens on ~1/20th (11.5M tokens) of the validation dataset

Note:

- Benchmark is a portion of LLM pre-training (0.4% of full GPT3)
- Need to keep run time reasonable to allow wider participation

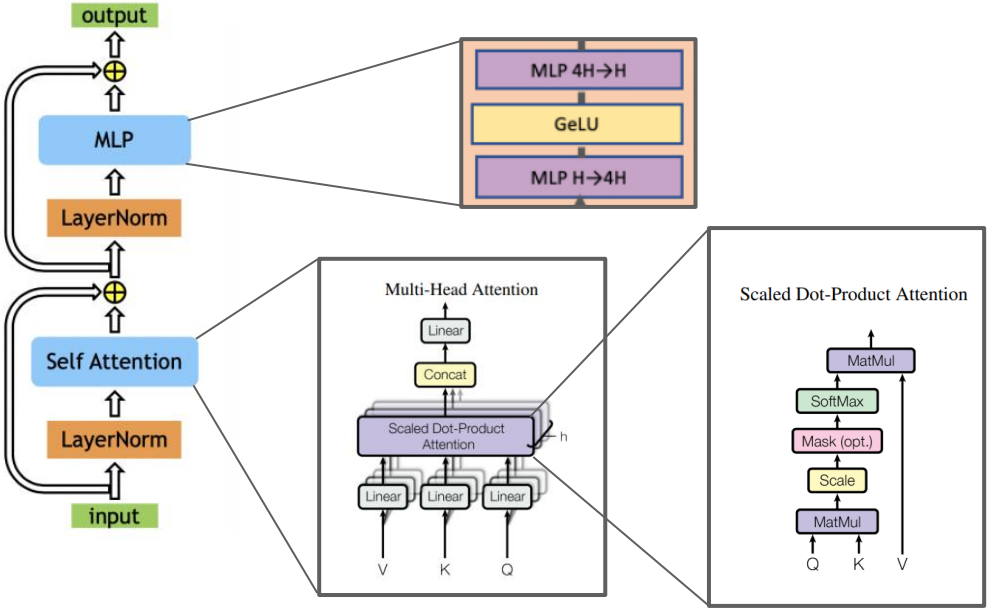
GPT-3 175B Reference Model

- 96 Transformer Layers (decoder only)
- SentencePiece with BPE Tokenizer (not in diagram)
- Sequence Length of 2048 Tokens
  - Input tokens used to make a prediction
- Adam Optimizer: Adaptive learning rate optimized for large problems
- Starts from initial checkpoint and reaches target accuracy (2.69 log perplexity) by training over ~1.3B tokens



https://arxiv.org/abs/2005.14165

Transformer Layer



- 96 stacked Transformer Layers:
  - Multi-Head Self Attention (memory intense)
  - MLP (compute intense)

- GPT-3 has 175B parameters (~350GB in BF16)
- Plus more to save model optimizer states, activations

- Must split across processors for training (e.g., reference requires at least 64 accelerators)

Benchmark construction challenges

- Realistic runtime
  - Select checkpoints and convergence region within full run
  - Select realistic target accuracy and eval frequency

- Reproducibility and fairness
  - Multi-framework convergence analysis and debugging
  - Identify hyperparameters that are stable across different numerics
  - Distributing dataset and checkpoint

- Scale
  - Minimum system size 64 accelerators
  - Many experiments required O(100) runs
  - Compute time and availability (>600K accelerator hours)

Submission challenges

- ~100X compute (~1 month) vs. prior benchmarks
  - New debug, optimization techniques needed

- Systems from 256 to 3,584 accelerators
  - Interconnect plays a significant role at this scale
- Partitioning is necessary to achieve performance
  - Data, Tensor and Pipeline parallelism
  - Sophisticated approaches necessary for good utilization

- Numerical stability - FP8 is on the frontier of feasibility today
- General software stack maturity and robustness, e.g., flash attention

Congratulations to: Intel-Habana Labs, NVIDIA, and NVIDIA+Coreweave

LLM Inference Benchmark Details

LLM Inference Task: Text summarization

- Text Summarization of a news article
- Model output is evaluated using the ROUGE metrics
- Tokenizer isn't counted in timing

Model Input – "`<instruction/prompt>:<prefix>`"

```
Summarize the following news article:(CNN)Following last year's successful U.K.
tour, Prince and 3rdEyeGirl are bringing the Hit & Run Tour to the U.S. for the
first time. The first -- and so far only -- scheduled show will take place in
Louisville, Kentucky, the hometown of 3rdEyeGirl drummer Hannah Welton. Slated
for March 14, tickets will go on sale Monday, March 9 at 10 a.m. local time.
Prince crowns dual rock charts. A venue has yet to be announced. When the Hit &
Run worked its way through the U.K. in 2014, concert venues were revealed via
Twitter prior to each show. Portions of the ticket sales will be donated to
various Louisville charities. See the original story at Billboard.com. ©2015
Billboard. All Rights Reserved.
```

CNN-DailyMail Fine-tuning Dataset

- cnn_dailymail/en/3.0.0 - dataset hosted on HuggingFace
  - 287k training articles, 13.4k validation articles and 11.5k testing articles
  - Each article is ~781 tokens, capped at 2K tokens
  - Each summary is ~56 tokens (~3.75 sentences), capped at 128 tokens

- Total fine-tuning dataset is 224M tokens
  - 1785X smaller than pre-training data
  - Did not use entire fine-tuning dataset to reach sufficient accuracy

- Benchmark uses the validation dataset to run the summarization task
- Model Accuracy evaluation also uses the validation dataset

GPT-J 6B Reference Model

- GPT-J is a 6B parameter OSS autoregressive language model (~22.55GB in FP32)
- Hugging Face checkpoint for the Pre-trained Language Model (PLM)
  - Trained on 400B tokens from the PILE dataset.
  - Fine-tuned until rouge1=42.9865, rouge2=20.1235, rougeL=29.9881

- GPT-2/3 tokenizer with 50257 different tokens
  - Tokenizations excluded from performance measurement
- 28 transformer layers with dimension 4K, feedforward dimension of 16K
- Sequence Length of 2048 tokens
  - Input tokens used to make a prediction
- Similar to training model, but subtle architectural differences
  - Rotary embedding (RoPE)
  - Parallel attention and feedforward layer for decreased communication

Inference Details

- Quality target: 99.9% or 99% of the original FP32 reference
  - rouge1=42.9865, rouge2=20.1235, rougeL=29.9881
  - Generated text length must be >90% of reference (4,0186,878)

- GPT-J in Datacenter and Edge (Single stream, server, offline scenarios)
  - Server latency constraint is 99% of queries below 20 seconds

- Generations Parameters:
  - num_beams: 4
  - min_new_tokens: 30
  - max_new_tokens: 128
  - early_stopping: True

Benchmark construction challenges

- Which LLM?
  - Size: **6B**, 40-70B, 175B, 1T
    - Capture broader user-base, e.g., enterprises
  - Task: chatbot vs. **summarization** vs. code-gen vs. agents
  - Availability of a (high quality) pre-trained checkpoint, dataset

- Legal risks: Dataset, distributing a frozen model, future legislation

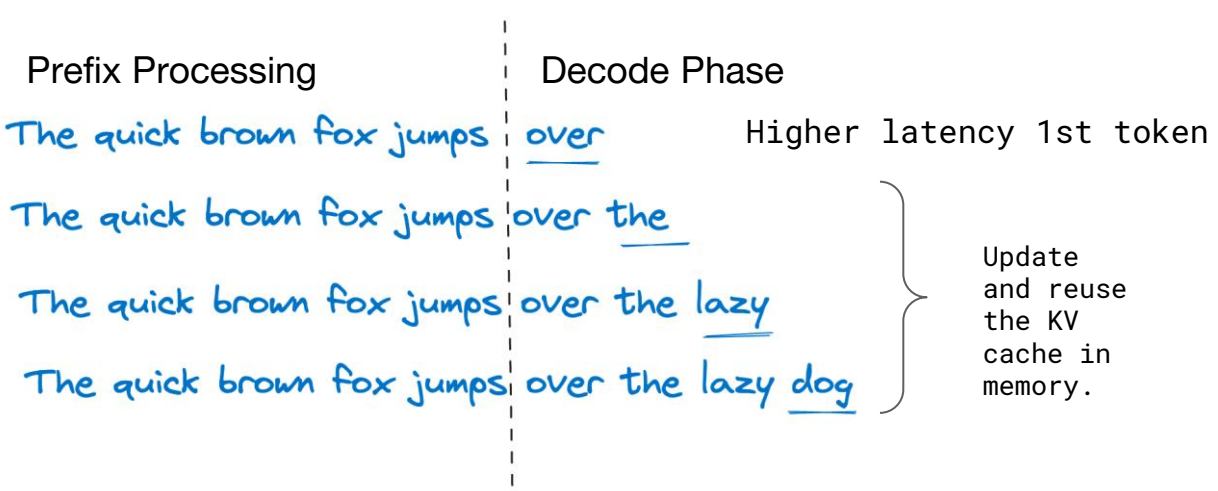
- Fine-tuning accuracy: Our model is currently #6 in world-wide leaderboard

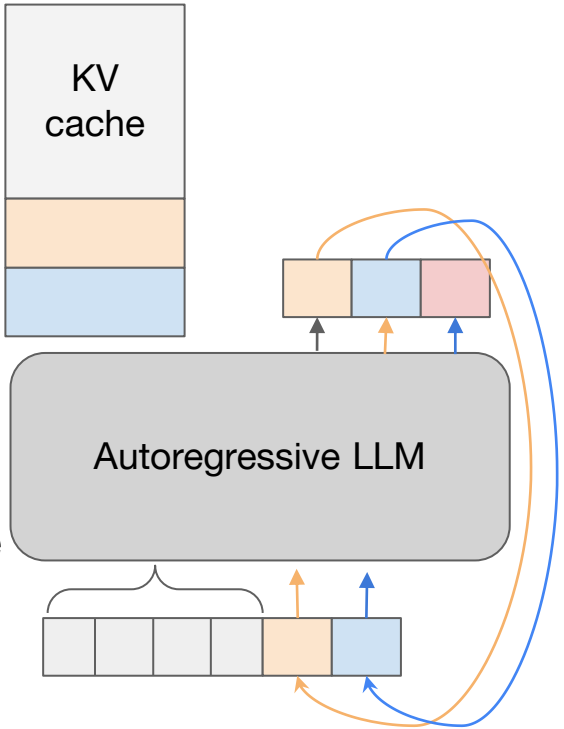
- Metrics: Token throughput? 1st token latency, **end-to-end latency**?

- Decoding algorithm and parameters? Greedy vs. **Beam search** vs. TopK

Submission Challenges

Generative Inference is Decoder-centric





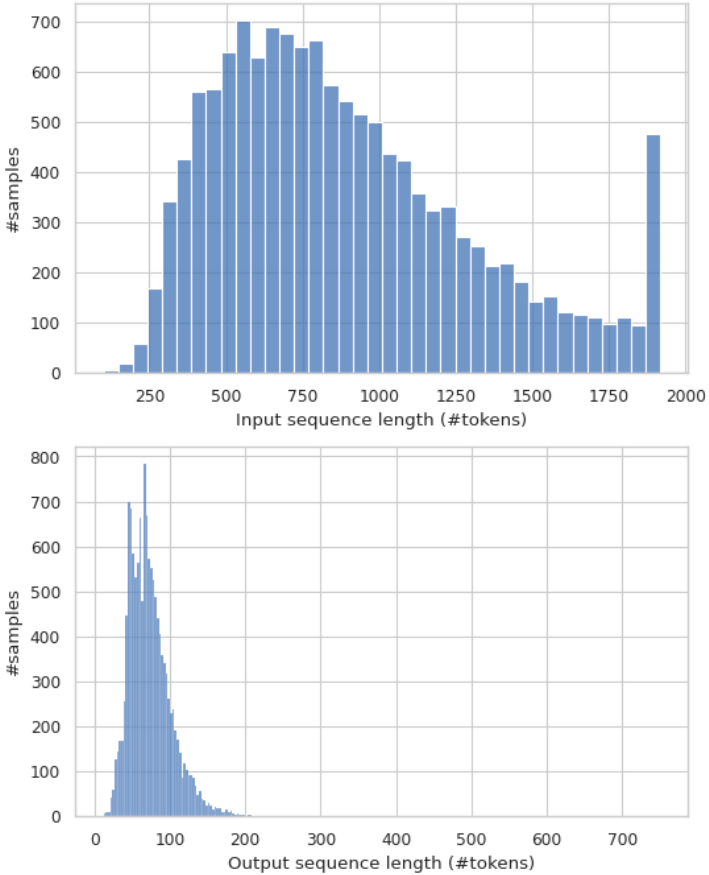
- Prefix processing is parallelizable, decode is auto-regressive
- KV cache integral optimization for transformer block
  - Size: 2 x precision x n_{layers} x d_{model} x seqlen x batch

Variable Cost Inference

- Variable length input → variable inference cost (unlike e.g. vision)

- GPT-J Benchmark
  - Input sequence < 2K tokens
  - Output sequence < 128 tokens

- Ideally bucket similar length queries
- Latency constraints limit bucketing
- Multiple processors can ***super-linearly*** improve performance
  - For some input distributions, YMMV



More Challenges

- Serving infrastructure and optimizations
- FlashAttention implementation
- Multi-processor partitioning strategy
  - Transparent or explicit?
- Numerical stability: FP8, INT4
- Reimplementation to different framework
- General software stack maturity and robustness

Congratulations to: Azure, Dell, CTuning, Giga Computing, Google, HPE, Intel, Intel-Habana Labs, Moffett, NVIDIA, Oracle, Quanta Cloud Technology, Supermicro, TTA, xFusion

Thank you

Amazing MLCommons Teamwork

Training: Yuechao Pan (Google), Shriya Palsamudram, Anmol Gupta, Ritika Borkar (NVIDIA), Itay Hubara (Intel-Habana Labs), Eric Han (Meta)

Submitters: Intel-Habana Labs, NVIDIA, NVIDIA+Coreweave

Inference: Itay Hubara (Intel-Habana Labs), Ramesh Chukka, Thomas Atta-fosu, Badhri Suresh Narayanan (Intel), Ashwin Nanjappa, Zhihan Jiang (NVIDIA), Akhil Arunkumar (d-Matrix), Yavuz Yetim, Mitchelle Rasquinha (Google), Miro Hodak (AMD)

Submitters: Azure, Dell, CTuning, Giga Computing, Google, HPE, Intel, Intel-Habana Labs, Moffett, NVIDIA, Oracle, Quanta Cloud Technology, Supermicro, TTA, xFusion

Questions?

Executive Director: David Kanter

David Kanter is a Founder, Board Member, and the Executive Director of MLCommons where he helps lead the MLPerf benchmarks and other initiatives.

He previously led the MLPerf Inference, Mobile, and Power working groups. He has 16+ years of experience in semiconductors, computing, and machine learning. He founded a microprocessor and compiler startup, was an early employee at Aster Data Systems, and has consulted for industry leaders such as Intel, Nvidia, KLA, Applied Materials, Qualcomm, and Microsoft.

David holds a Bachelor of Science degree with honors in Mathematics with a specialization in Computer Science, and a Bachelor of Arts with honors in Economics from the University of Chicago. He has a handful of patents on imaging systems and applications and has consulted on >10 patent litigation matters.