Evaluating AMD Instinct™ MI300A APU: Performance Insights on LLM Training via Knowledge Distillation

Dennis Dickman (Seedbox), Philip Offenhäuser (HPE), Rishabh Saxena (HLRS), George Markomanolis (AMD), Alessandro Rigazzi (HPE), Patrick Keller (HPE), Kerem Kayabay (HLRS), Dennis Hoppe (HLRS)

CUG 2025

7 May 2025

AMD
together we advance_

Motivation

- Rise of the LLMs and their compute demands
- Is there any difference/restriction because of the new architecture?
- There are available enough material/documents for AI workflows and MI300X but not for MI300A, so this work contributes to this direction
- Key questions: Can HPC centers utilize MI300A for AI workflows? Can MI300A handle real-world LLM training/inference efficiently?
- Verifying an AI workflow works and helps scaling an AI model and uses fewer resources
- Supporting architectural decisions for the upcoming system: Herder (~2027)

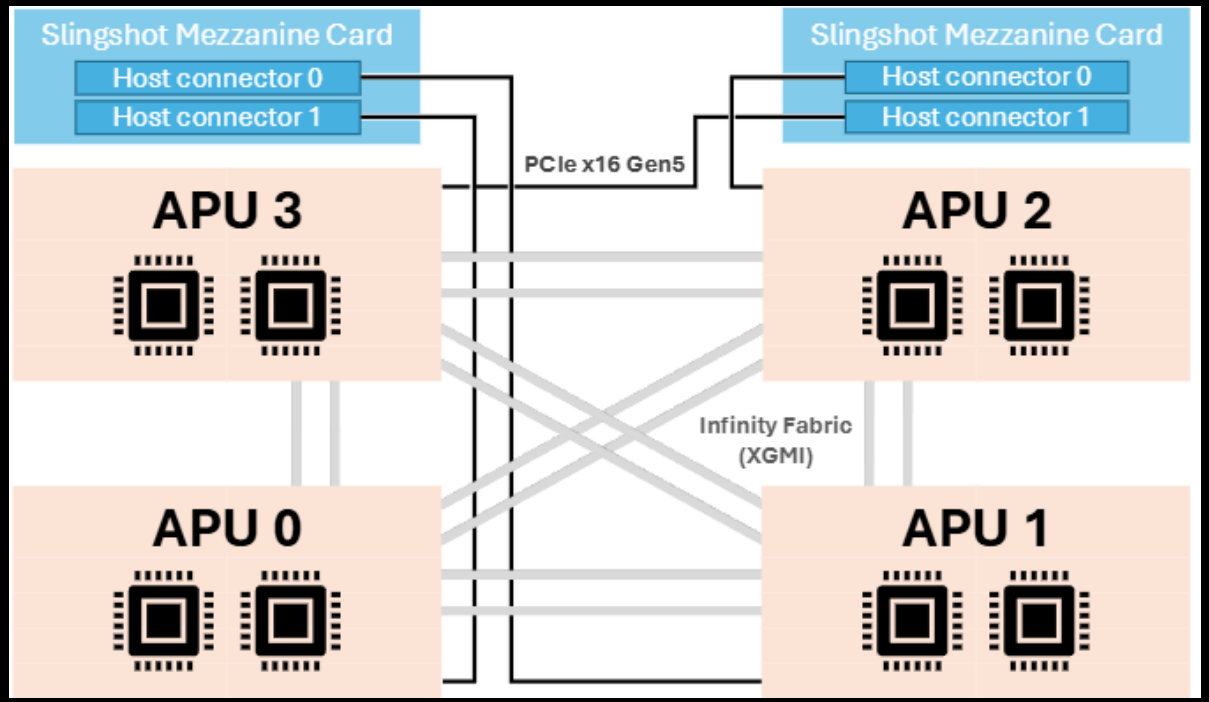
Hunter Supercomputer

- The new HLRS supercomputer was deployed and accepted end of January 2025, with a theoretical peak performance of 48.1 PFlops

- HPE/AMD system with 188 nodes of 4 x AMD Instinct™ MI300A per node

- For each Instinct™ MI300A:
  - 6 AMD CDNA 3, GPU XCD Chiplets with a total of 228 compute units
  - 3 AMD “Zen 4” chiplets with 24 CPU cores
  - 128 GB HBM3 of unified memory

- Slingshot network, 200 Gbps per APU



Software Stack

- Libraries: PyTorch (2.6.0), DeepSpeed (0.16.1), RCCL, AWS-OFI plugin (commit 17d41cb), Hugging Face Accelerate
- Optimization: ZeRO-3, NUMA-aware memory pinning
- ROCm (6.2.2) + DeepSpeed plugins

Overview of the Study

- Dataset: 20 billion token multilingual corpus spanning 24 EU languages
- Focus pre-trained models: SmolLM2-1.7B from HuggingFace, Mistral-Small-24B-Base-2501 from MistralAI

- Goal:
  - Present strong scaling results for training META AI's LLM 3.1 70B with Deepspeed
    - Learn more about the performance
    - Tune for our real-world AI workloads
  - Assess MI300A on LLM workloads using Knowledge Distillation
  - Test pipelines for compression and fine-tuning
  - Investigate the scaling of such workflows on Hunter supercomputer

- Contributions:
  - Hardware-aware model optimization
  - Implementation framework for novel architecture adoption
  - One of the first analysis of MI300A for LLM distillation
  - Decision support for novel architecture adoption
  - New model KafkaLM-15B

Knowledge Distillation Basics

- KD can decrease the model size, thus improve inference speed and utilize less resources
- The big model (teacher) trains the small model (student)
- White-box vs black-box KD
- We use white-box KD for this study with a dual approach that focuses on both the hidden embedding layers and output logits
  - Dual Loss: Kullback-Leibler (KL) divergence, and Mean Squared Error (MSE) objective over hidden states

Dataset

- 20B token multilingual dataset with synthetic content
  - Tailored for post-pruning knowledge distillation
  - Normal cases require 50-100B tokens
  - Our novel loss function reduces data requirements while it restores model performance after pruning
  - Because of the MI300A, and the restriction of the 512 GB memory per node, we build custom data loader that prepares small batches (pre-fetching) and feeds the APU on the fly to avoid memory overload
  - We achieve 88-90% GPU utilization

SimplePrune

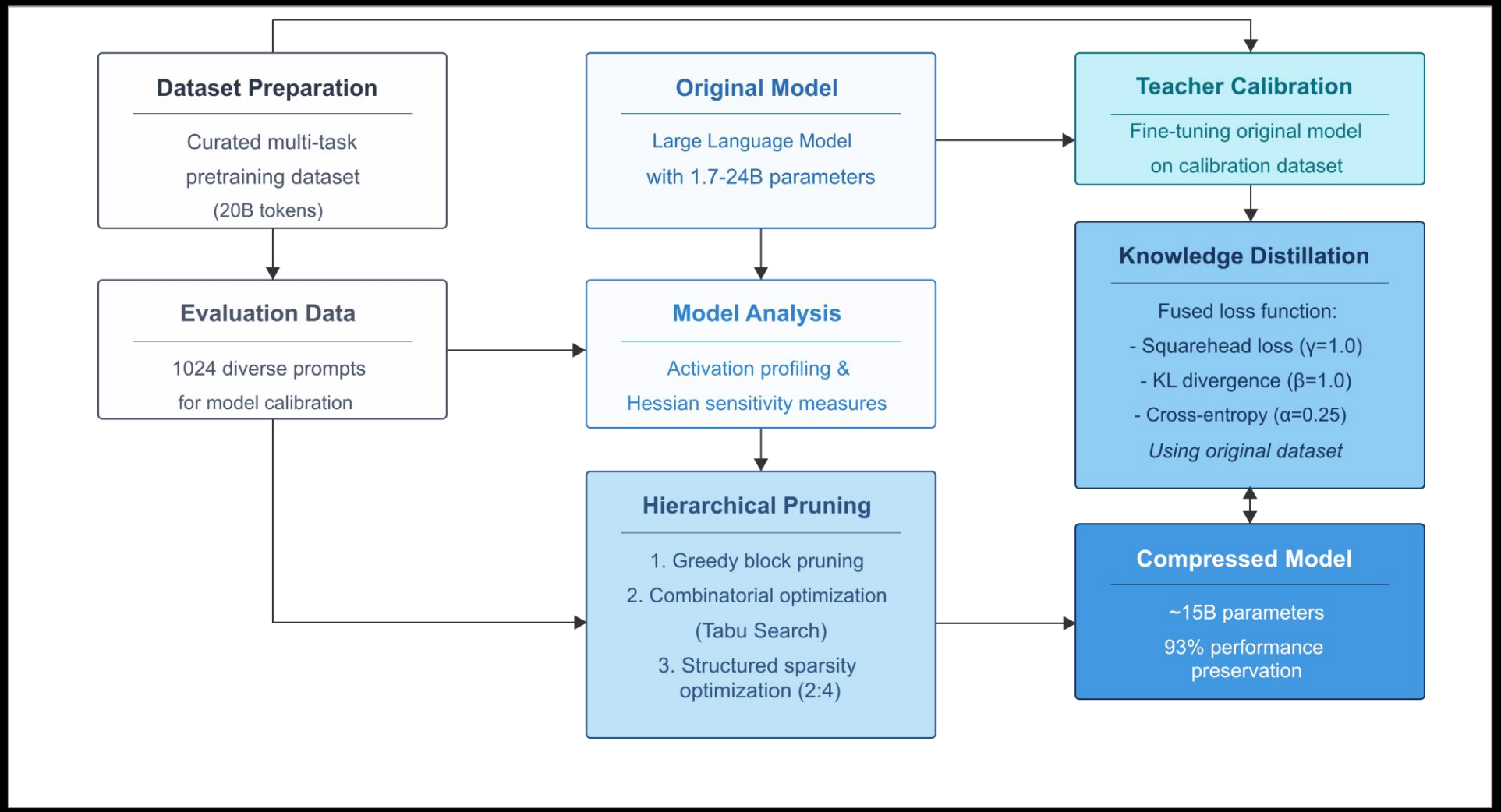
- Multi-stage hierarchical pruning pipeline
- Hardware-aware
- Hierarchical pruning framework
- Designed for efficient compression of LLMs
- The goal is to reduce model size while preserving performance, making LLM deployment feasible in memory-constrained environment
- Structured 2:4 sparsity
- Hessian-based importance, Tabu Search metaheuristics
- **Achieved up to 40% reduction with minimal loss**

Distillation Pipeline

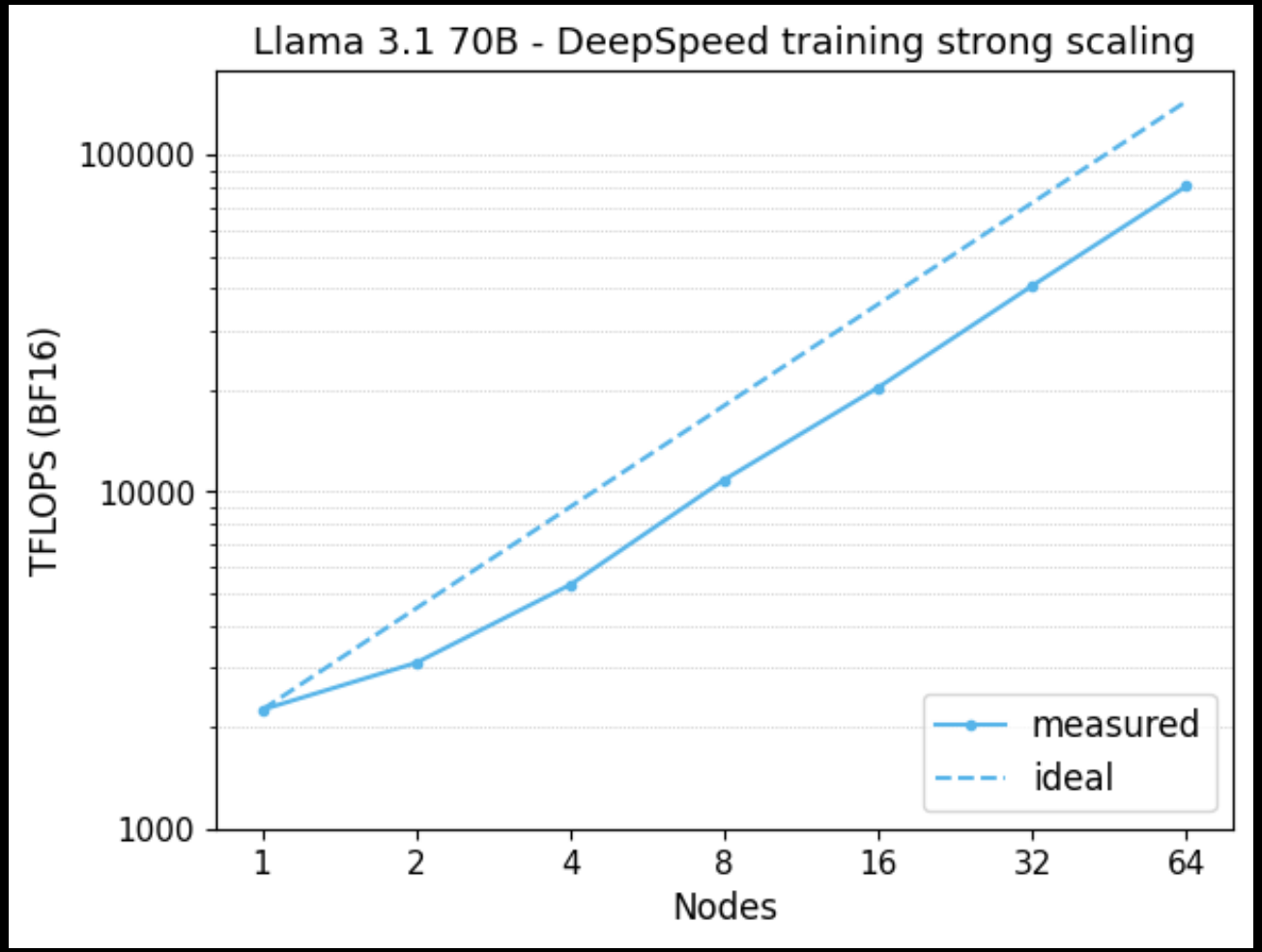
- Teacher calibration: We fine-tuned the teacher model on the distillation data first so that it gives more consistent, realistic answers. This helps the student model learn better and avoids problems caused by the teacher being too confident or confused.
- Knowledge distillation:
  - Used dual DeepSpeed plugins simultaneously orchestrated by Hugging Face’s Accelerate library
    - ZeRO-3 (Zero Redundancy Optimizer - Stage 3) for activation partitioning and optimizer state sharding
    - Disable CPU offloading considering MI300A architecture
    - NUMA-aware memory pinning and prefetching
  - Reducing memory footprint by approximately 43%
  - Achieving stable throughput exceeding 35,000 tokens/second
- Custom fused loss function
- Adaptive layer alignment

Workflow

Hierarchical Pruning and Knowledge Distillation Workflow



Llama 3.1 70B – Pre-production run



We need to study more and investigate the performance from 1 to 4 nodes.

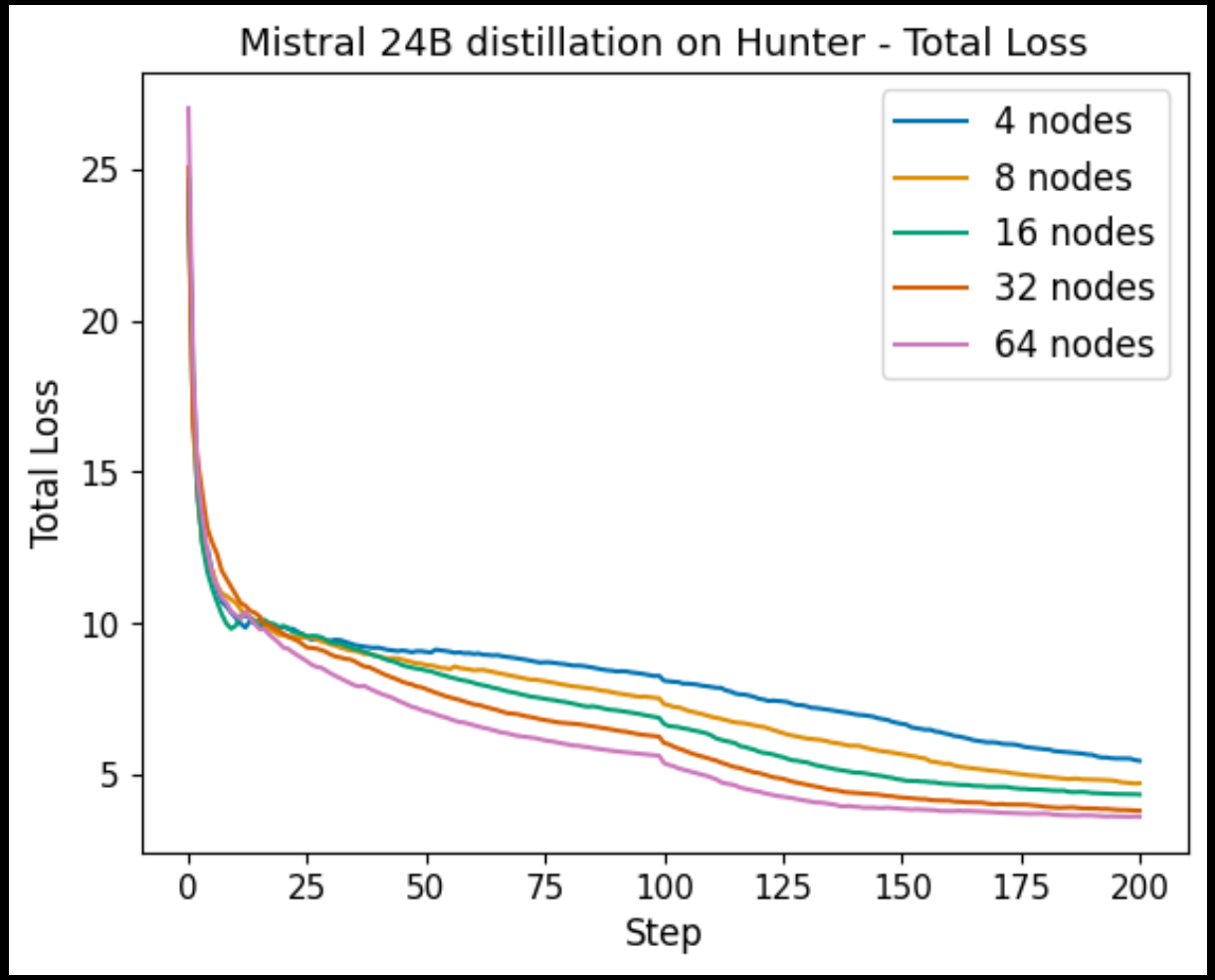
Experiments

- Fine-tuning:
  - SmolLM2-1.7B – 32 APUs
  - Mistral-Small-24B-Base-2501 – 128 APUs

- Knowledge Distillation
  - We use the calibrated teacher models to train the student
  - SmolLM2-1.7B – 32 APUs
  - Mistral-Small-24B-Base-2501 – 256 APUs

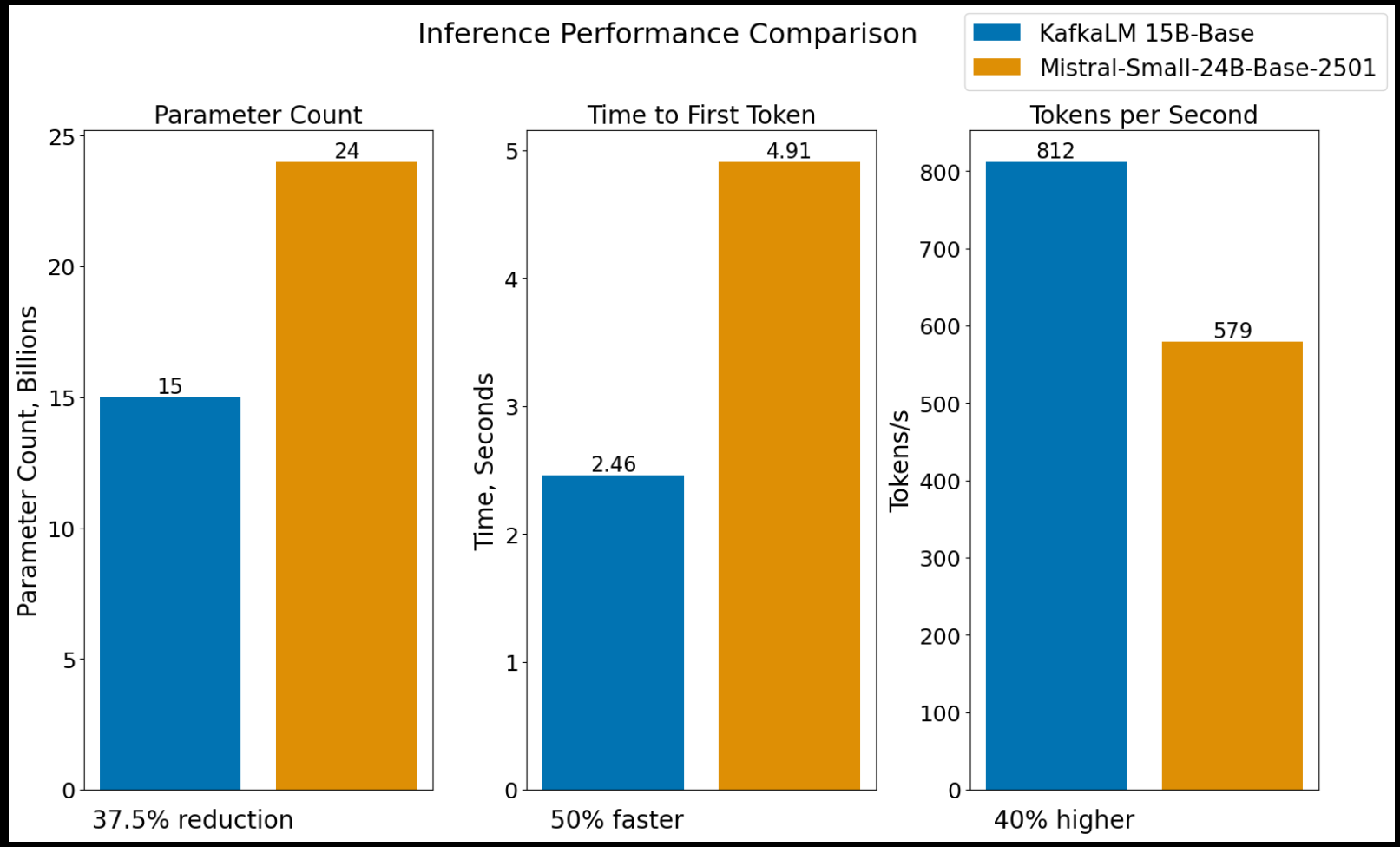
Results – Training Performance

- Convergence during KD training of the 15B pruned model across different setups
- The best convergence is achieved with a global batch size of 512 (64 nodes)
- The distilled model achieved 96% accuracy recovery across various performance metrics
- We plan for further performance benchmarks in the post-training pipeline such as supervised fine-tuning, reinforcement learning and other



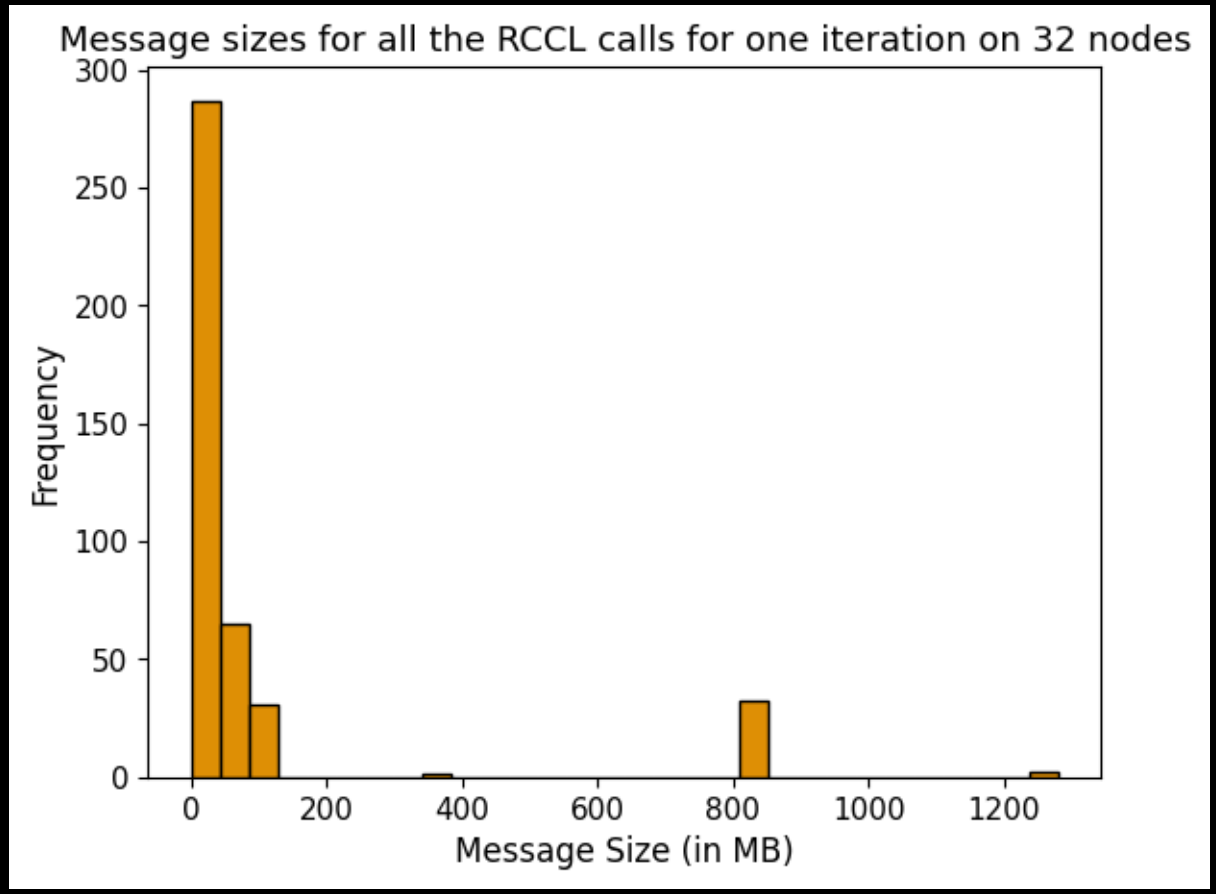
Results – Inference Performance

- KafkaLM vs Mistral: Faster inference despite 37.5% fewer parameters
- 50% lower time to first token, 40% higher token throughput

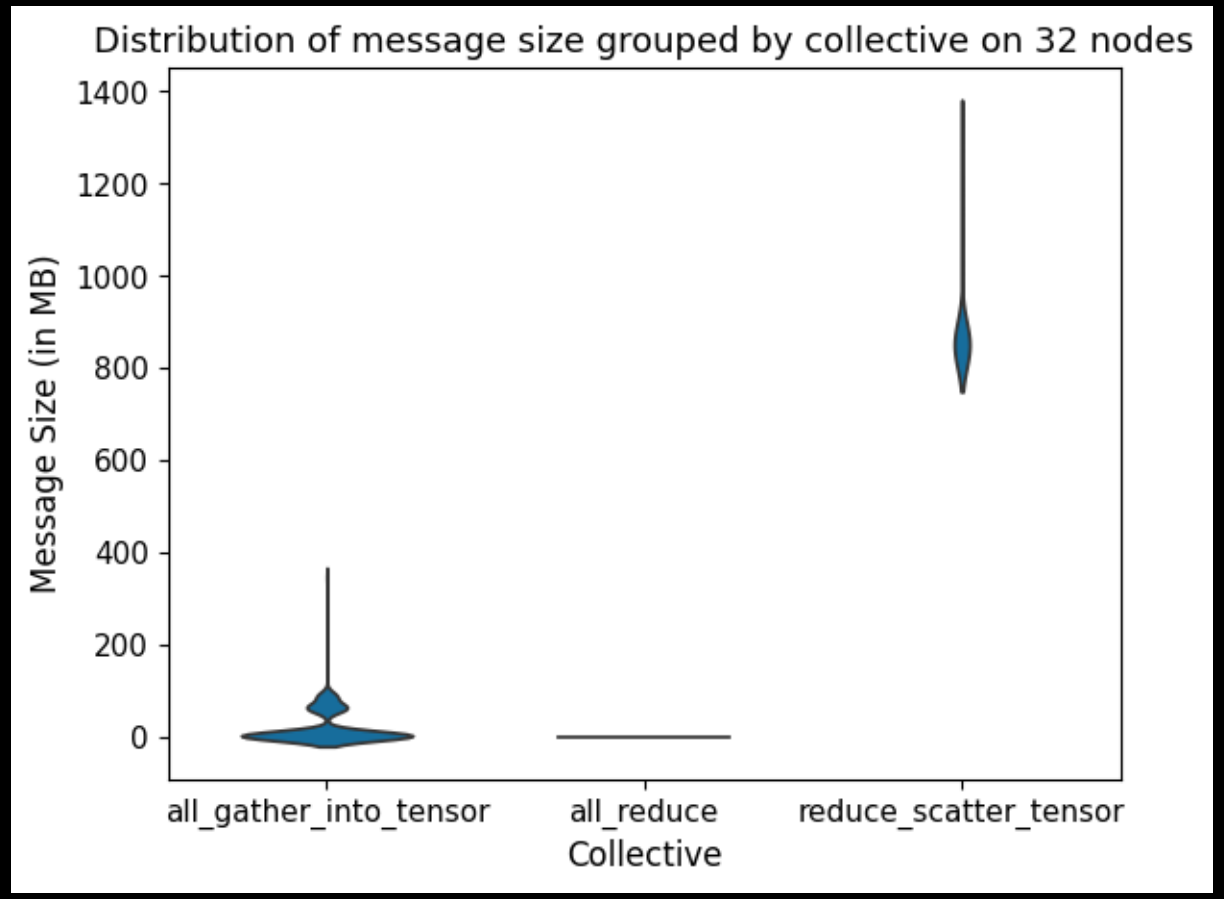


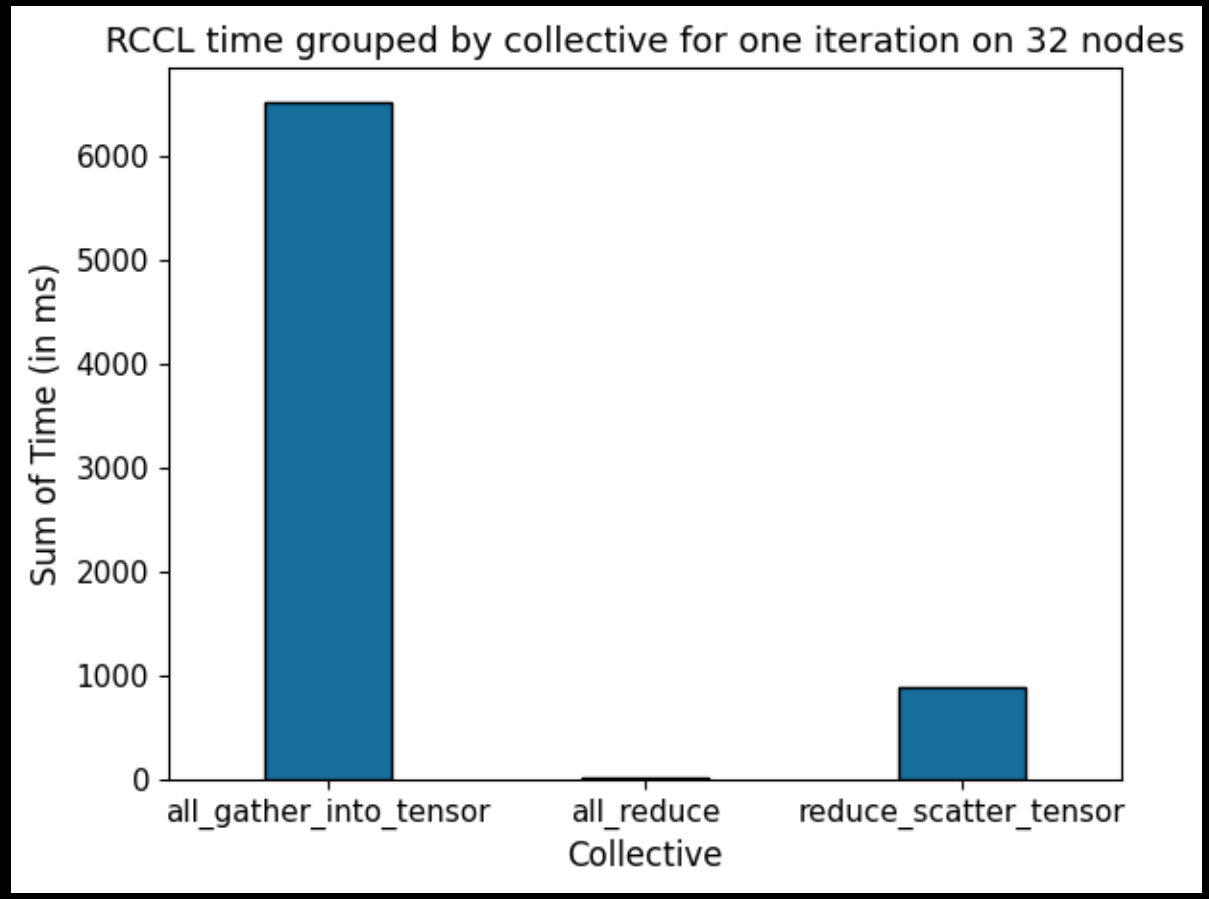
Scaling (I)

- Profiling RCCL communication through DeepSpeed
- The data are the same for each iteration
- We plot the data for a single iteration



Scaling (II)

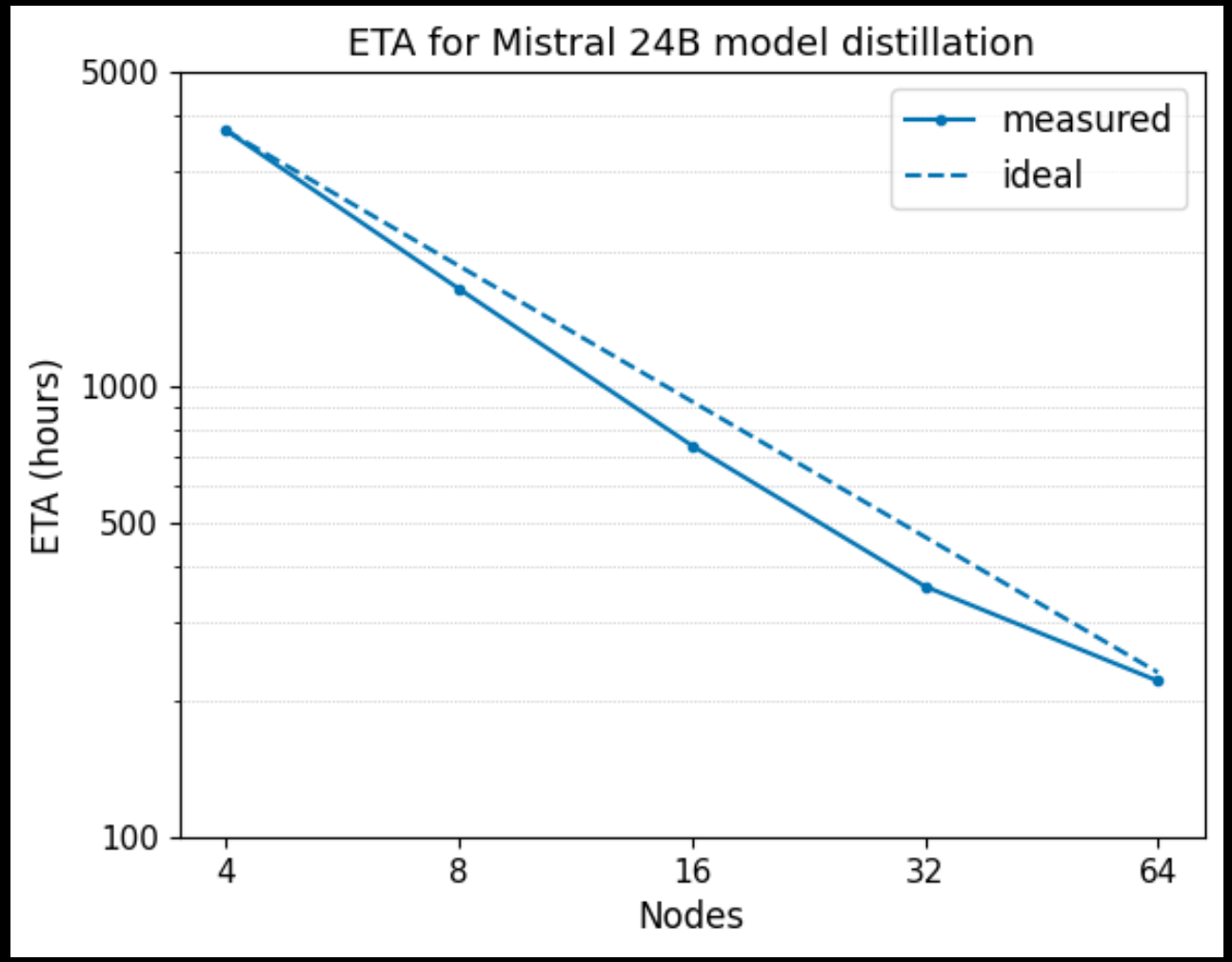




NCCL_MIN_NCHANNELS=2
NCCL_MAX_NCHANNELS=16

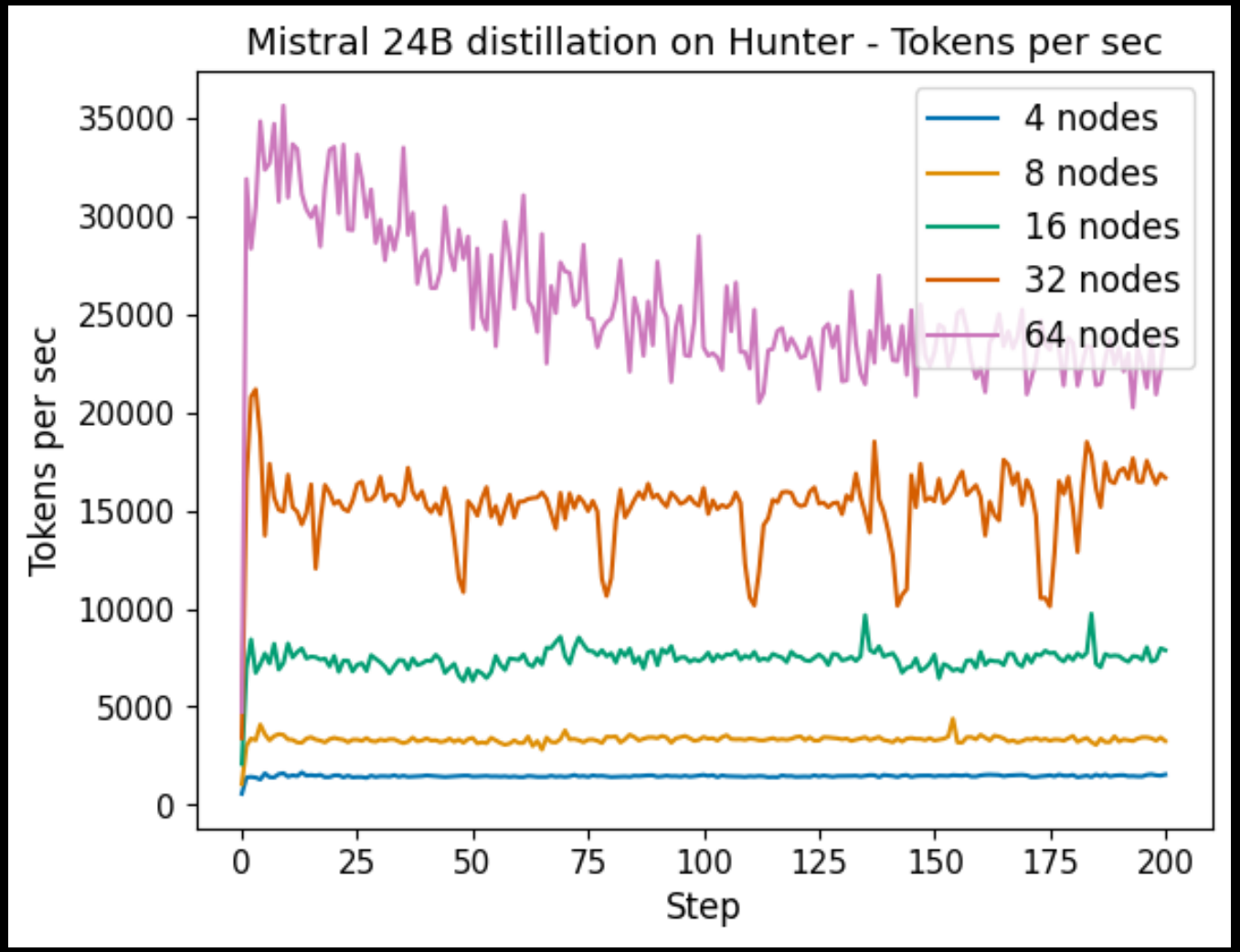
Scaling (III)

- ETA scaling from 4 to 64 nodes of Mistral-Small-24B-Base-2501
- Performance increased almost linearly due to the time reduction for distillation



Scaling (IV)

- Consistent performance improvement with increased node count
- The 64 nodes run is not so stable but the system is fresh after acceptance, and there could be some not healthy nodes
- The 64 nodes case needs further investigation



Discussion & Challenges

- Hunter did not allow containers, all the tools were built on virtual environment
- RCCL variables for tuning
- The nodes on Hunter do not have host memory, so we had to think about the workflow and the memory requirements
- The workflow we use can be guide for similar work with MI300A devices
- We presented inference results from this state-of-art knowledge distillation pipeline and demonstrated that models with reduced parameters can remain effective
- We introduced SimplePrune, a multi-stage pipeline designed for the MI300A architecture
- The KafkaLM-15B model proves that with this workflow can help in deploying AI models at scale

- Repos:
  - https://github.com/Seedbox-Ventures/kafkalm
  - https://huggingface.co/seedboxai/KafkaLM-15B-Base

Future Work

- Optimize this workflow to avoid significant inefficiencies in the time to converge and the optimal model configuration
- Explain further the SimplePrune framework
- We could also do further hardware-aware optimizations, such as for FLOPs, memory bandwidth, etc.
- As we see more and more such architecture on Supercomputers it is important to analyze and benchmark AI/ML workloads including hybrid simulation-AI pipelines
- Extend to Reinforcement Learning from Human Feedback (RLHF), inference optimization, further tuning

Questions?

DISCLAIMERS AND ATTRIBUTIONS

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

THIS INFORMATION IS PROVIDED 'AS IS." AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

AMD, the AMD Arrow logo, Radeon™, Instinct™, EPYC, Infinity Fabric, ROCm™, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.


AMD