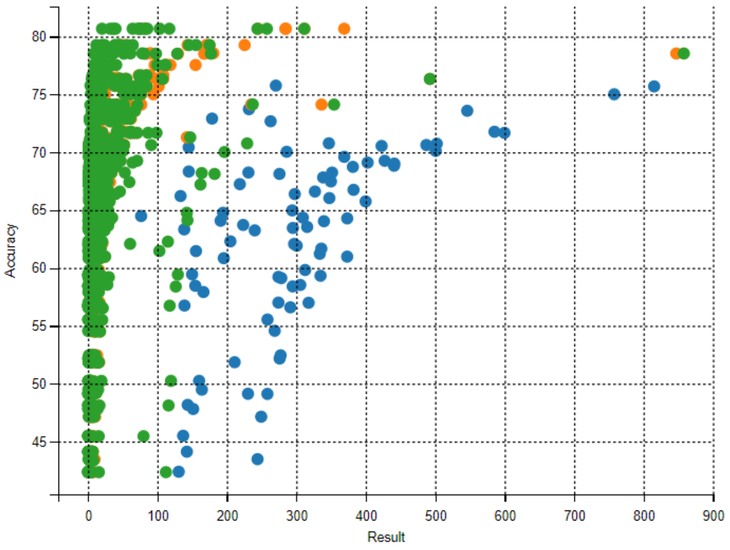
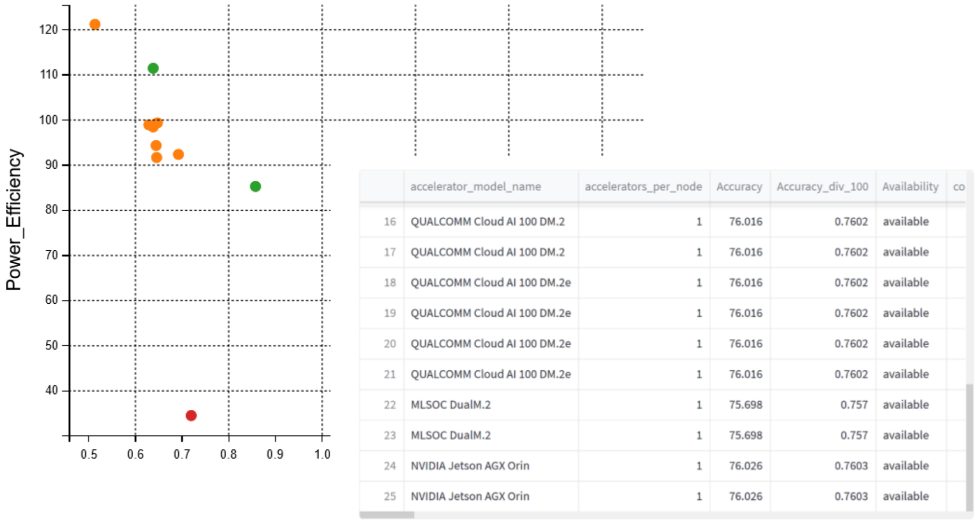
Automation and Reproducibility Task Force
Collective Knowledge Playground

access.cKnowledge.org

A free, open-source, technology-agnostic and on-prem automation platform for collaborative and reproducible MLPerf inference benchmarking, optimization and comparison across any software, hardware, models and data sets from any vendor: https://github.com/mlcommons/ck/tree/master/platform





	accelerator_model_name	accelerators_per_node	Accuracy	Accuracy_div_100	Availability	co
16	QUALCOMM Cloud AI 100 DM.2	1	76.016	0.7602	available	
17	QUALCOMM Cloud AI 100 DM.2	1	76.016	0.7602	available	
18	QUALCOMM Cloud AI 100 DM.2e	1	76.016	0.7602	available	
19	QUALCOMM Cloud AI 100 DM.2e	1	76.016	0.7602	available	
20	QUALCOMM Cloud AI 100 DM.2e	1	76.016	0.7602	available	
21	QUALCOMM Cloud AI 100 DM.2e	1	76.016	0.7602	available	
22	MLSOC DualM.2	1	75.698	0.757	available	
23	MLSOC DualM.2	1	75.698	0.757	available	
24	NVIDIA Jetson AGX Orin	1	76.026	0.7603	available	
25	NVIDIA Jetson AGX Orin	1	76.026	0.7603	available	

Simple GUI to analyze, compare and reproduce MLPerf v3.0, 2.1 and 2.0 results with any derived metric such as Performance/Watt or Performance/$: https://github.com/mlcommons/cm_inference_results

Our 1st MLPerf inf v3.0 community submission

We thank **Neural Magic (Michael Goin)**, **Pablo Gonzalez Mesa**, students (**Himanshu Dutta, Aditya Kumar Shaw, Sachin Mudaliyar, Thomas Zhu**) and other great contributors to help us validate the MLCommons CK technology (including CM aka CK2 - the new version of our portable workflow framework) to unify, automate and reproduce MLPerf inference submissions:

- 80% of all results and 98% of power results
- Diverse CPUs, GPUs and DSPs with PyTorch, ONNX, QAIC, TF/TFLite, TVM and TensorRT
- Hardware from Nvidia (including 4090 workstation and Jetson AGX Orin edge device), Qualcomm, AMD, Intel and Apple
- Deep Sparse optimization from Neural Magic and models from the Hugging Face Zoo
- Cloud submissions on AWS and GCP
- 1st end-to-end student submissions including on Apple Metal

cKnowledge.org/mlperf-inf-v3.0-forbes

cKnowledge.org/mlperf-inf-v3.0-report

Next: join the 1st public optimization tournament for MLPerf inference v3.1!

cKnowledge.org/challenges

Contact Grigori and Arjun (automation and reproducibility task force co-chairs) and/or join our Discord server to learn about how to participate in the upcoming 1st reproducible optimization tournament for MLPerf inference v3.1 and suggest your own challenges: discord.gg/JjWNWXKxwT

We will continue working with all MLCommons members and researchers to adapt MLCommons CK/CM to their needs, reduce their benchmarking and optimization costs, and improve MLPerf/MLCommons value:

- Integrate their software and inference engines into portable CK-MLPerf workflows
- Improve CK platform to automate their MLPerf experiments and optimization
- Automatically generate containers for MLPerf benchmarks with CK/CM workflows and unified CLI

Based on your feedback, we plan to enhance the CK playground to generate Pareto-efficient end-to-end AI and ML-based applications using MLPerf results, CK technology and modular CK/CM containers - prototype is available and will be integrated with the CK playground by Q3 2023!