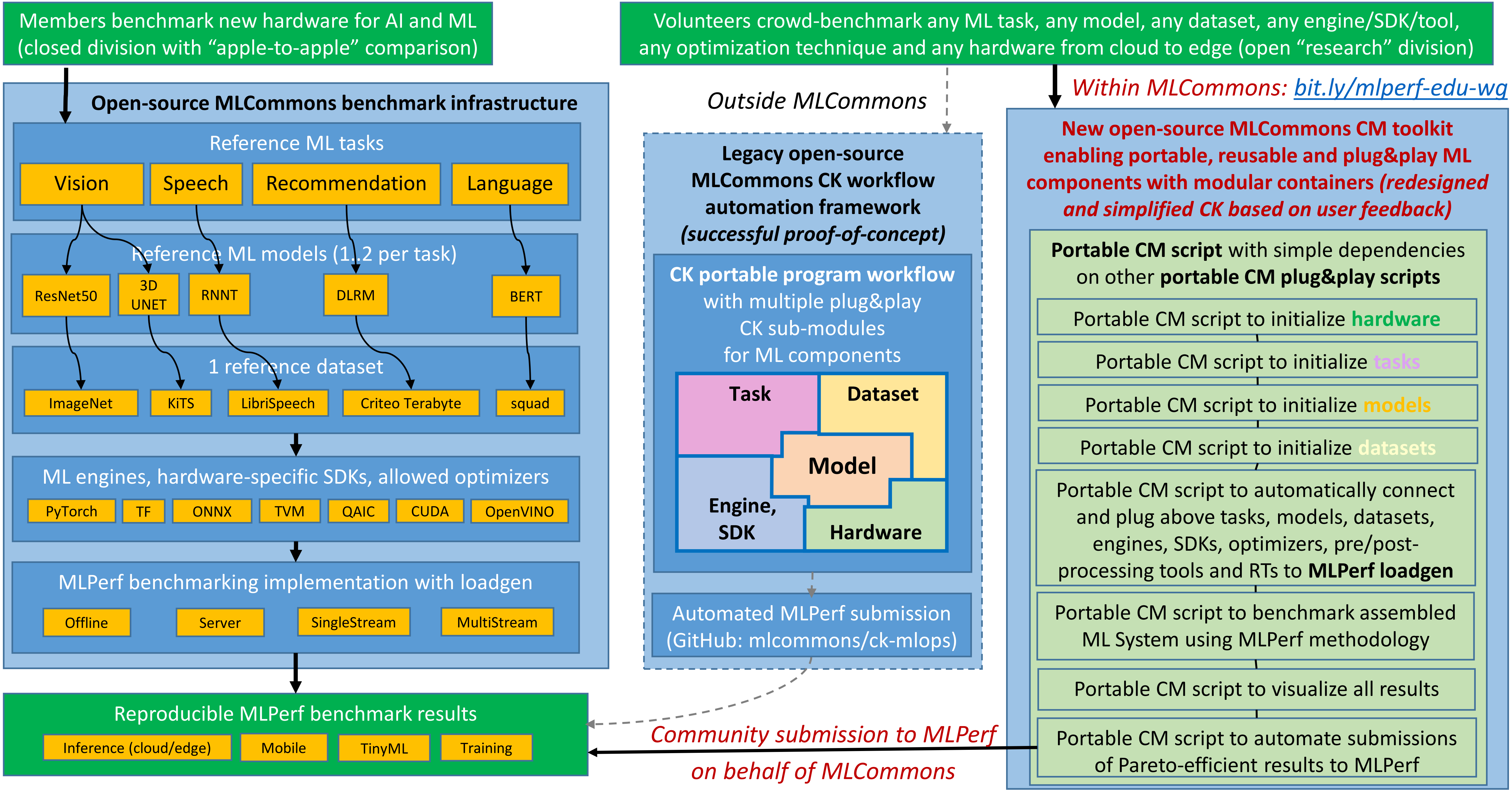
MLCommons members: mlcommons.org

MLPerf education workgroup: learn with the community how to modularize, crowd-benchmark and optimize any ML System using the MLPerf methodology



Members benchmark new hardware for AI and ML (closed division with "apple-to-apple" comparison)

Volunteers crowd-benchmark any ML task, any model, any dataset, any engine/SDK/tool, any optimization technique and any hardware from cloud to edge (open "research" division)

Open-source MLCommons benchmark infrastructure

Reference ML tasks

Vision | Speech | Recommendation | Language

Reference ML models (1..2 per task)

ResNet50 | 3D UNET | RNNT | DLRM | BERT

1 reference dataset

ImageNet | KiTS | LibriSpeech | Criteo Terabyte | squad

ML engines, hardware-specific SDKs, allowed optimizers

PyTorch | TF | ONNX | TVM | QAIC | CUDA | OpenVINO

MLPerf benchmarking implementation with loadgen

Offline | Server | SingleStream | MultiStream

Reproducible MLPerf benchmark results

Inference (cloud/edge) | Mobile | TinyML | Training

Outside MLCommons

Legacy open-source MLCommons CK workflow automation framework *(successful proof-of-concept)*

CK portable program workflow with multiple plug&play CK sub-modules for ML components

Automated MLPerf submission (GitHub: mlcommons/ck-mlops)

Within MLCommons: bit.ly/mlperf-edu-wg

New open-source MLCommons CM toolkit enabling portable, reusable and plug&play ML components with modular containers *(redesigned and simplified CK based on user feedback)*

Portable CM script with simple dependencies on other **portable CM plug&play scripts**

Portable CM script to initialize hardware

Portable CM script to initialize tasks

Portable CM script to initialize models

Portable CM script to initialize datasets

Portable CM script to automatically connect and plug above tasks, models, datasets, engines, SDKs, optimizers, pre/post-processing tools and RTs to **MLPerf loadgen**

Portable CM script to benchmark assembled ML System using MLPerf methodology

Portable CM script to visualize all results

Portable CM script to automate submissions of Pareto-efficient results to MLPerf

Community submission to MLPerf on behalf of MLCommons