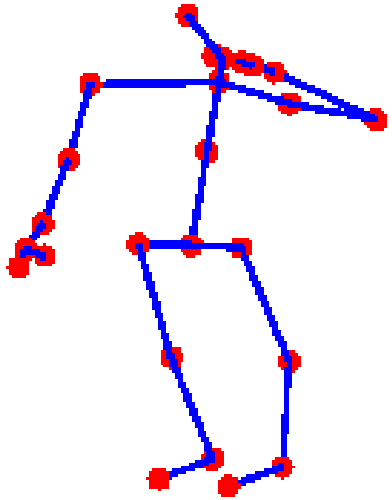
GENN: ENABLE FLEXIBLE AND EFFICIENT AI FOR RESOURCE-CONSTRAINED PLATFORMS

Yan Zhu, Kaija Mikes, Karthik Ganesan, Natalie Enright Jerger

UNIVERSITY OF TORONTO

Why AI on Resource-Constrained Platform



Posture Detection

Health Monitoring

Signal Classification

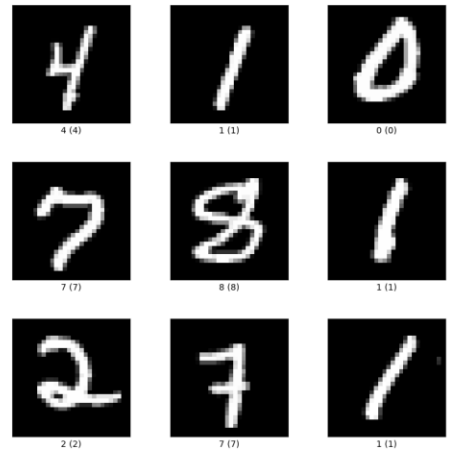
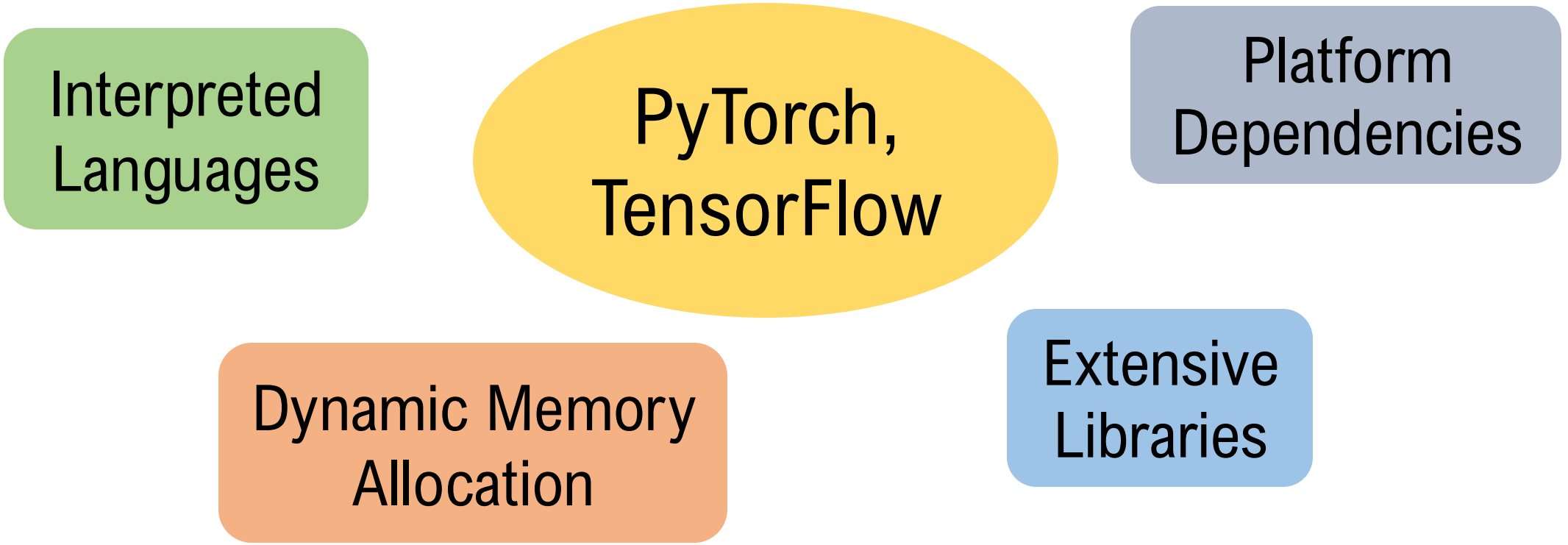


Image Classification

Motivation I: Limitations of Dominant Deep Learning Frameworks for Low-Performance Devices



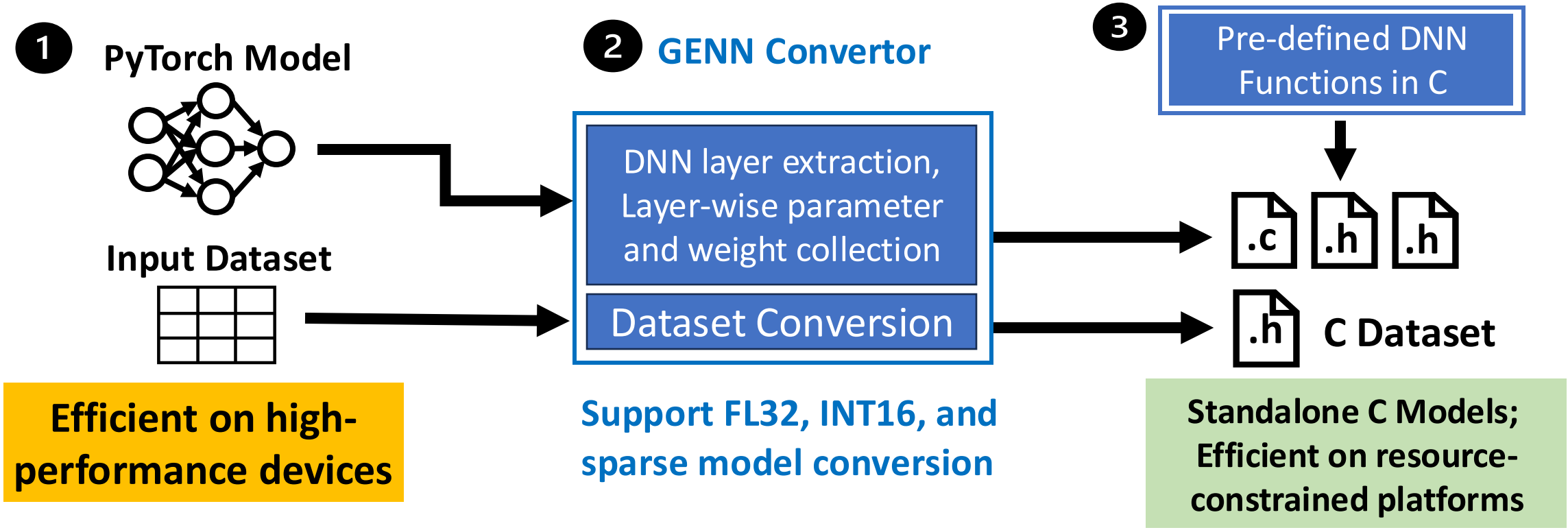
Motivation II: Drawbacks of Prior Works

	GENN (Our Work)	CMSIS-NN	uTensor	TensorFlow Lite Micro
Platform Requirement	None	ARM Cortex-M Processor	Mbed OS	Dynamic Memory Allocation
Auto Generation	✓	✗	✓	✓
DL Framework	PyTorch	None	TensorFlow	TensorFlow
Usability	Easy	Very Hard	Medium	Hard
Sparsity	✓	✗	✗	✗

Comparison of GENN and prior works.

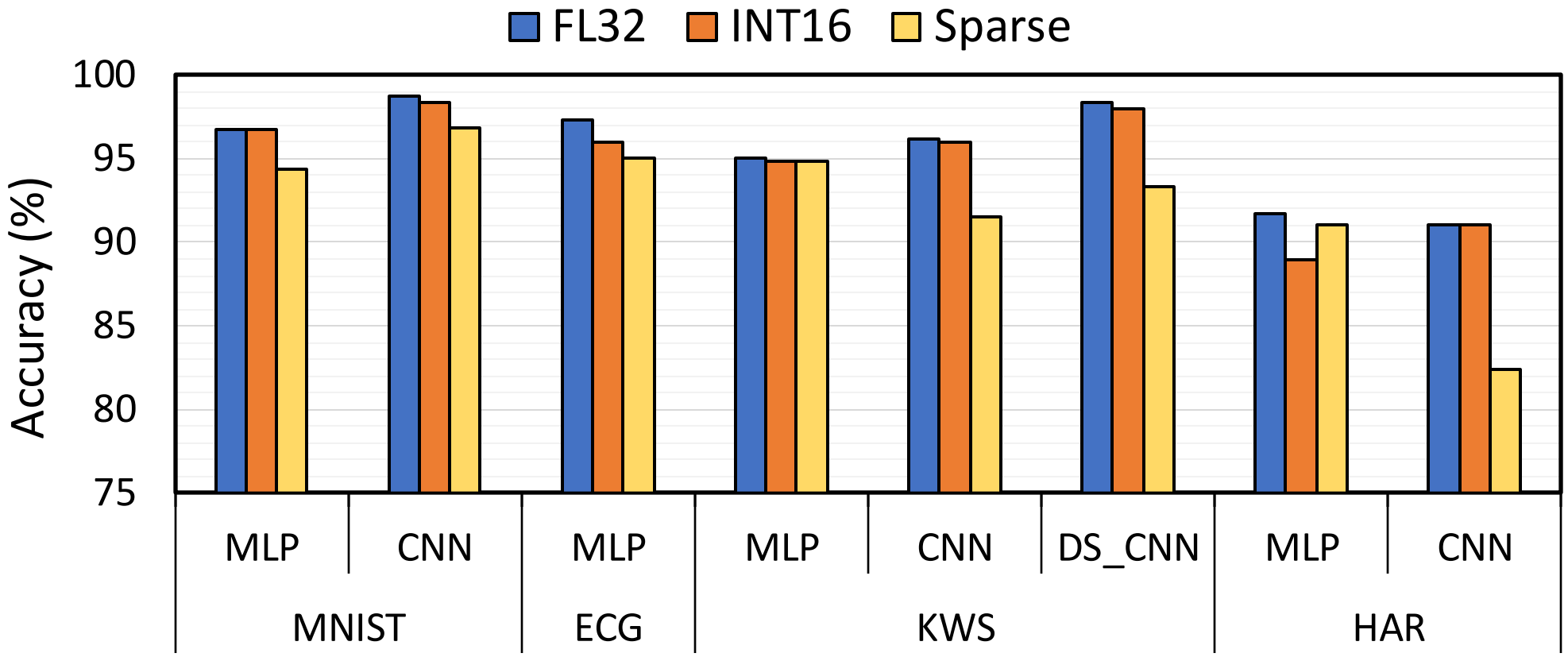
Research Goal: Develop a more portable and flexible PyTorch conversion tool that can better serve AI research on low-performance platforms.

Proposed Solution: GENN



- *No external library dependencies, OS constraints, or dynamic memory allocation requirements.*
- *Quantization operations require zero floating-point operations.*

GENN Benchmark

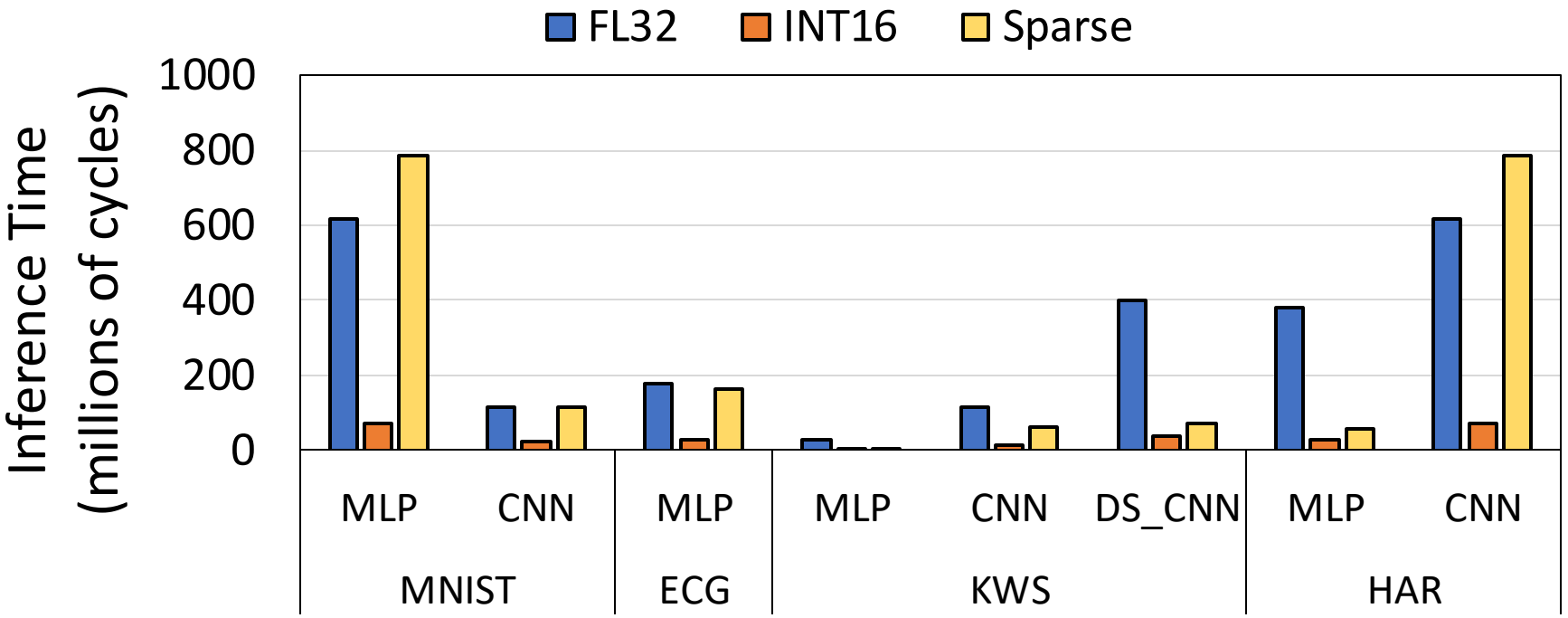


- *Eight deep learning models that are commonly used for IoT devices.*
- *Generated by GENN converter and can be used out of the box.*

Evaluation I: Simulation Results

Simulation on the Thumbulator -- a cycle-accurate simulator for the ARM Cortex-M0+ CPU, running at 24 MHz.

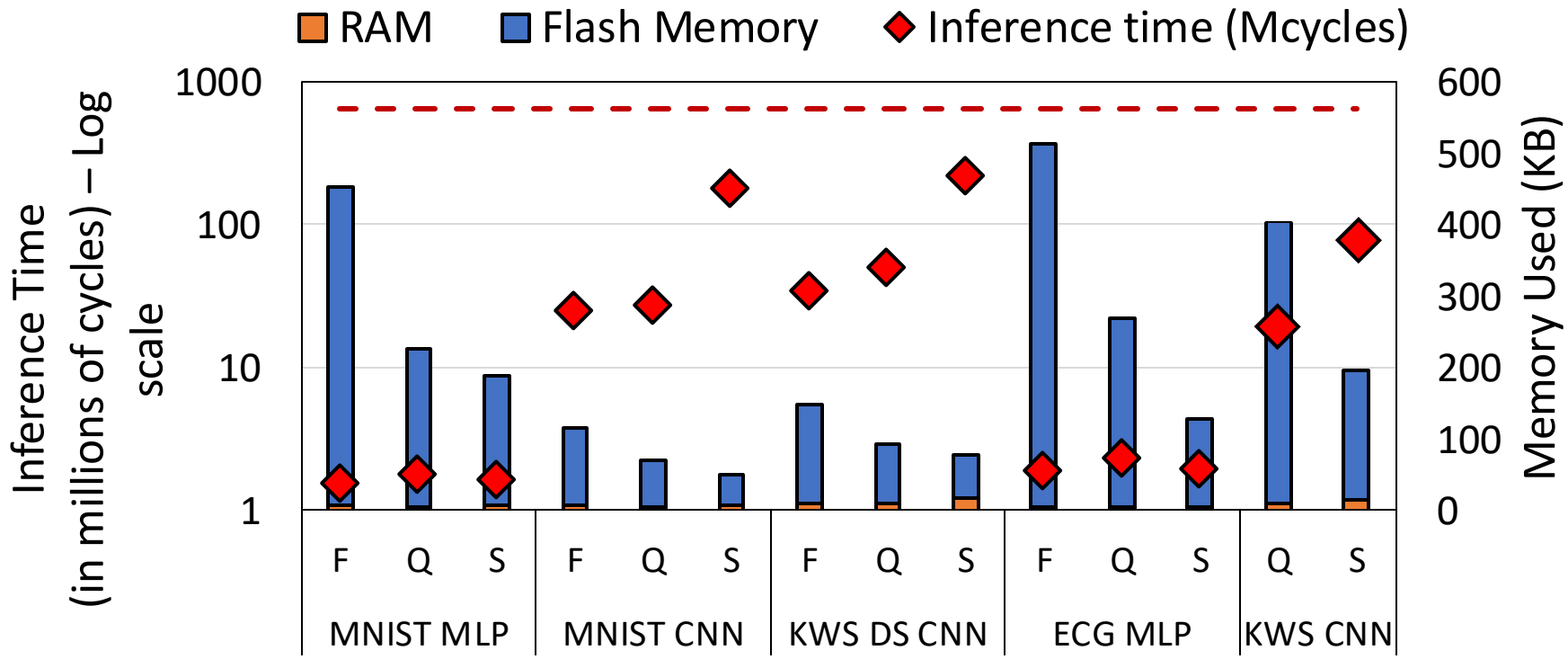
Significant speedup of quantized models due to the lack of HW support for floating point operations on the Cortex-M0+ CPU.



Evaluation II: Real Device Results

Evaluation on the STM32-NUCLEO-F411RE board with a 32-bit ARM Cortex-M4 CPU, 128 KB of RAM and 512 KB of flash memory, 100 MHz.

Quantization and sparsity reduce the memory by 44% and 60%.



Next Steps & Potential Values

Compare the inference time and memory usage of GENN benchmark to the experiment results of the prior works.

Support the quantized sparse model conversion.

Enrich GENN converter’s features to further improve its usability.

GENN: Enable Flexible and Efficient AI For Resource-Constrained Platforms

THANK YOU!

This research was undertaken, in part,

thanks to funding from the Canada Research Chairs Program.