

Pushing Biosignal Foundation Models to Ultra-low-power Edge Devices

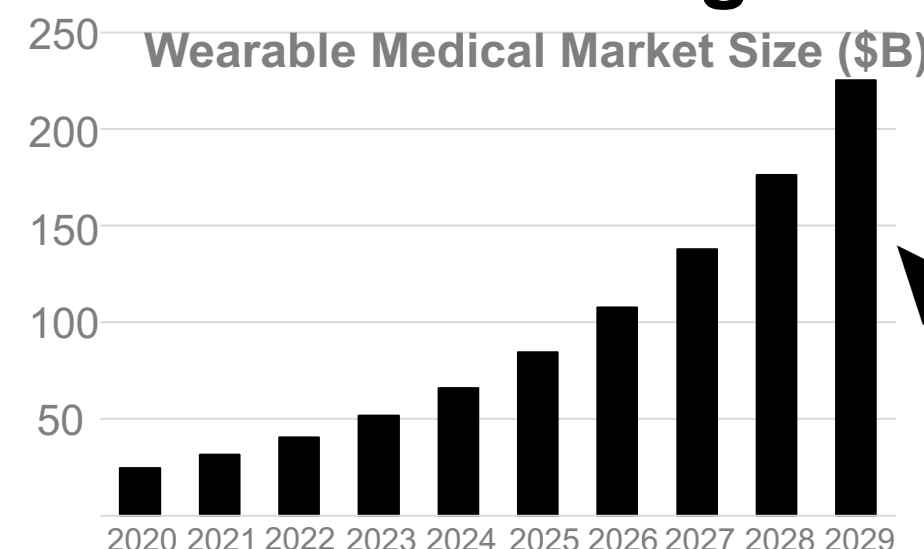
Thorir Mar Ingolfsson¹, Anna Tegon¹, Glenn Anta Bucagu¹, Luca Benini¹

¹ETH Zurich

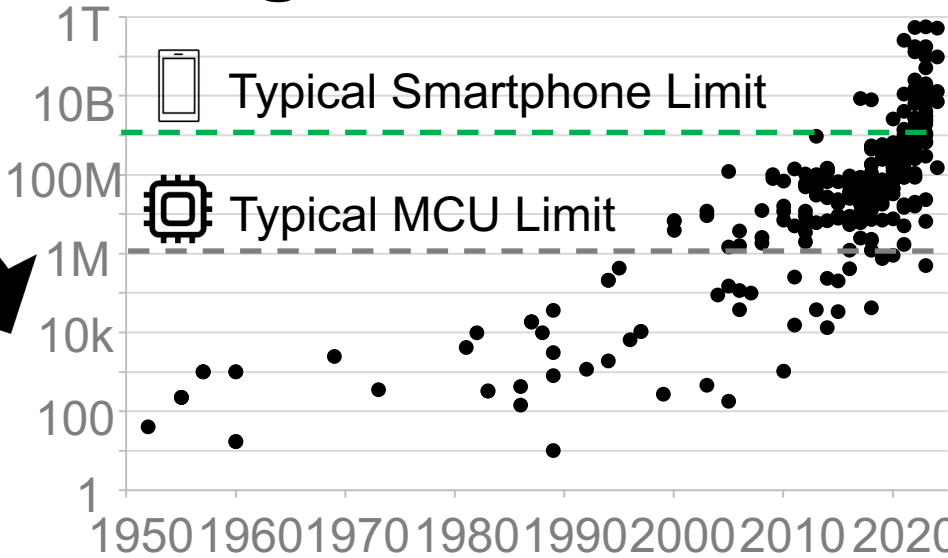
Motivation

Bridging Foundation Models and Wearable Constraints

Rapid Growth in Wearable Health Monitoring



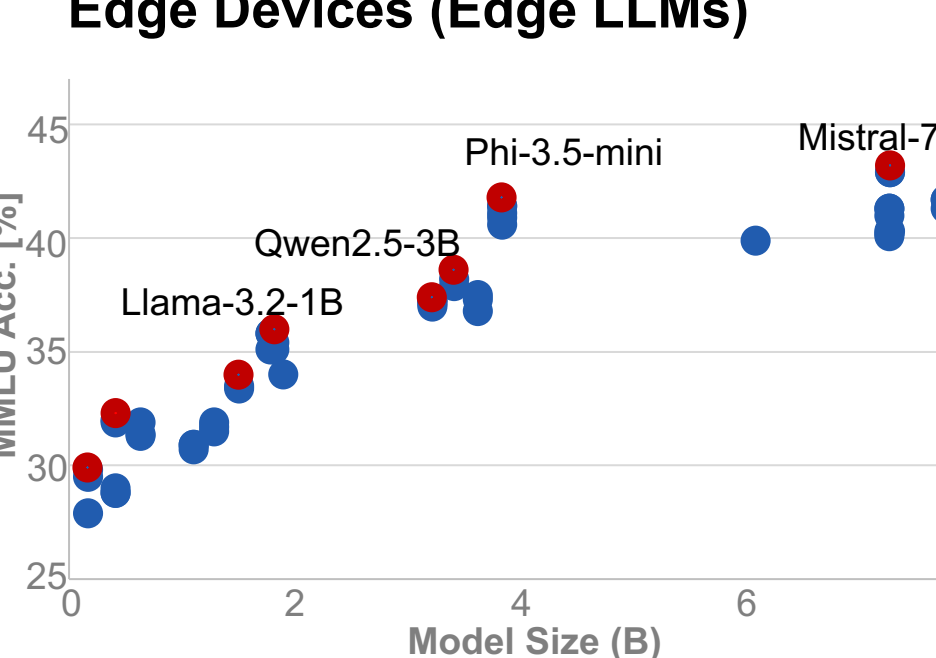
Increasingly Powerful & Large AI Models



Key Issues

Data Privacy & Security
Variable Quality of Sensor Data

Efficient Foundation Models on Edge Devices (Edge LLMs)



Key Issues

Integrating & Interpreting Heterogeneous Sensor Data
Resource Demands of Large AI Models (Compute, Energy, Memory)

Need for Edge AI: For Privacy & Robust On-Device Data Processing

Need for Edge AI: To Enable Efficient, Scalable Models on Device

What we need

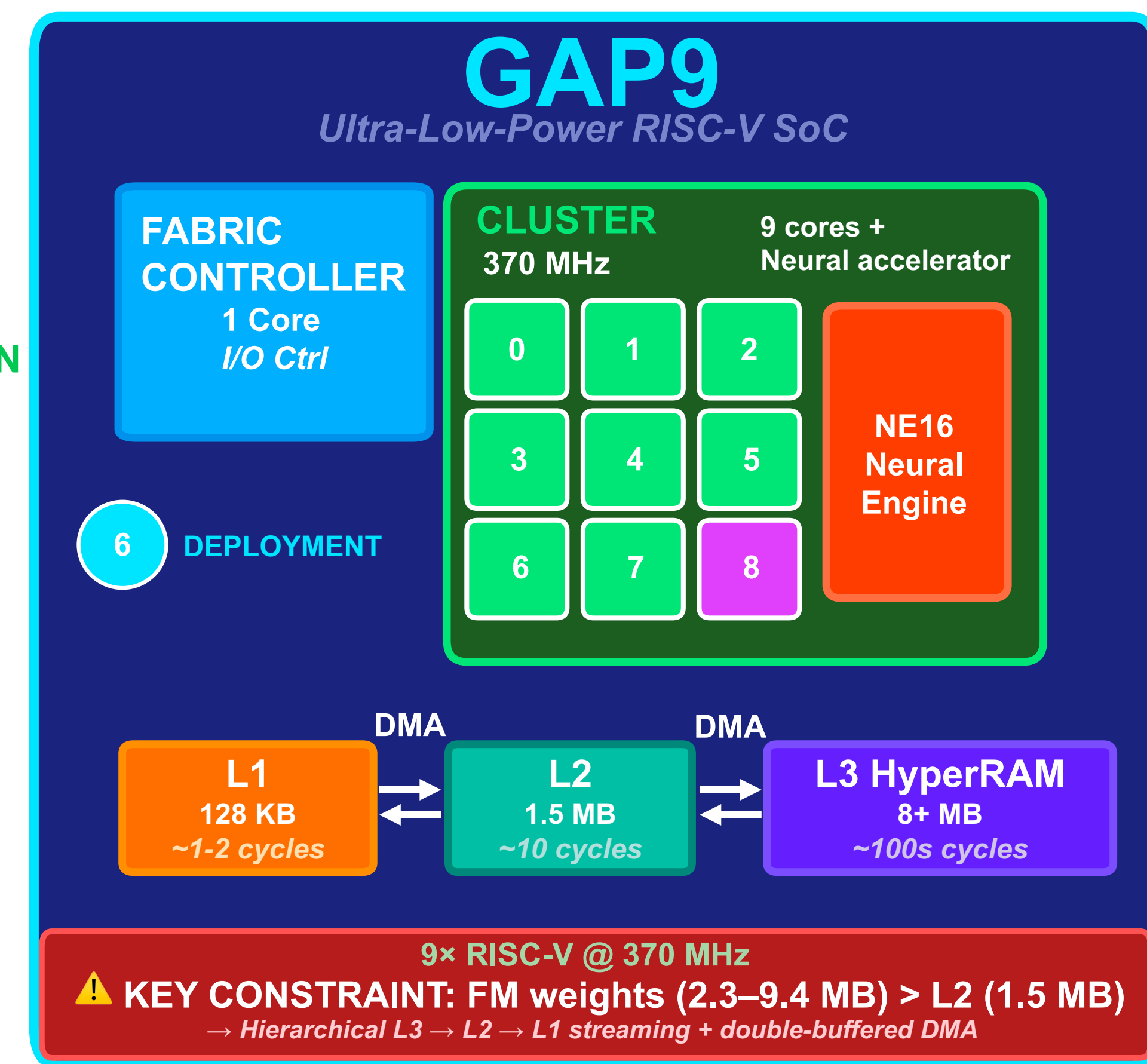
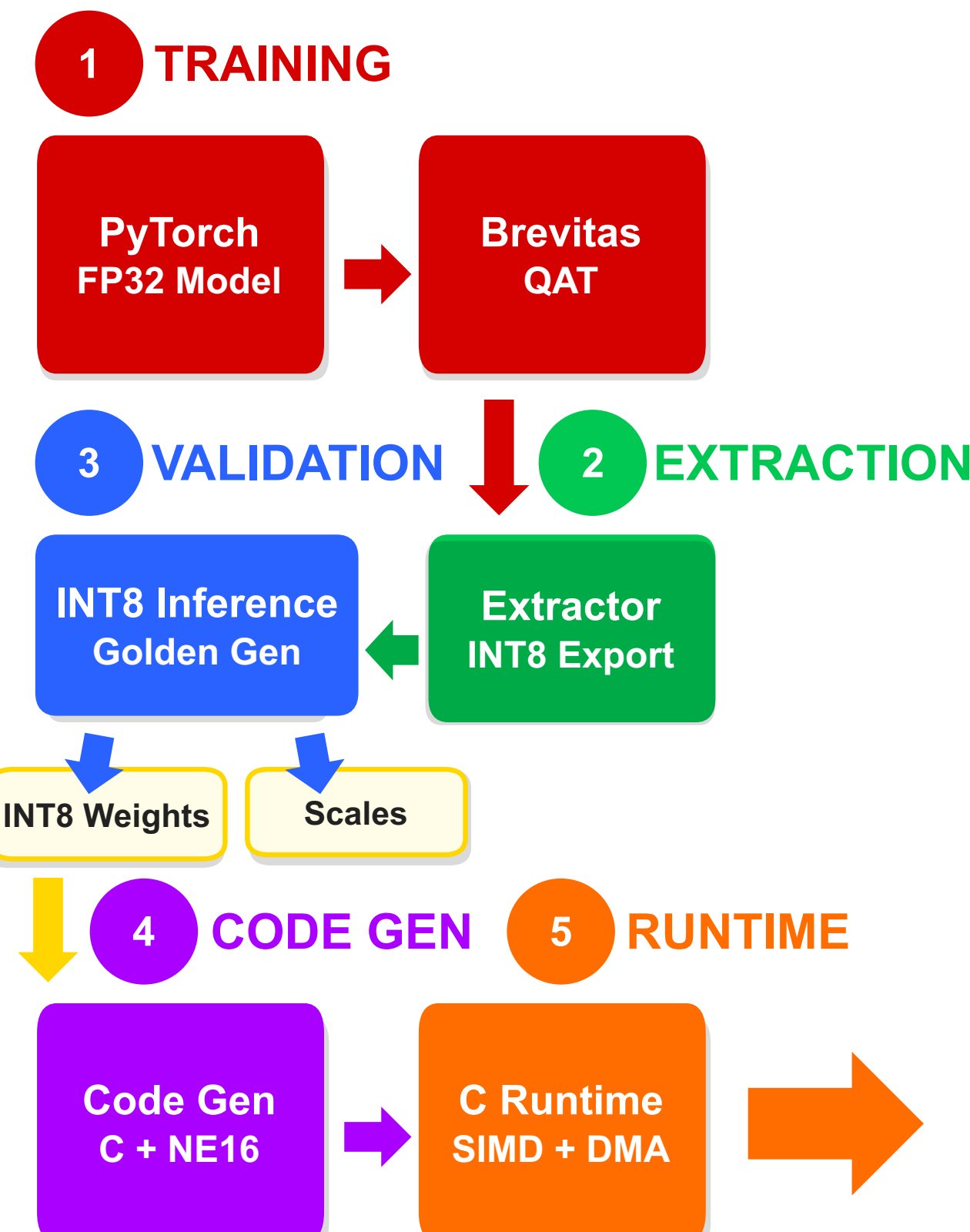
- FMs that are SoA yet compact
- End-to-End deployment pipeline
- Hardware-aware optimization

We bridge this gap with 4 SoA Biosignal Foundation Models



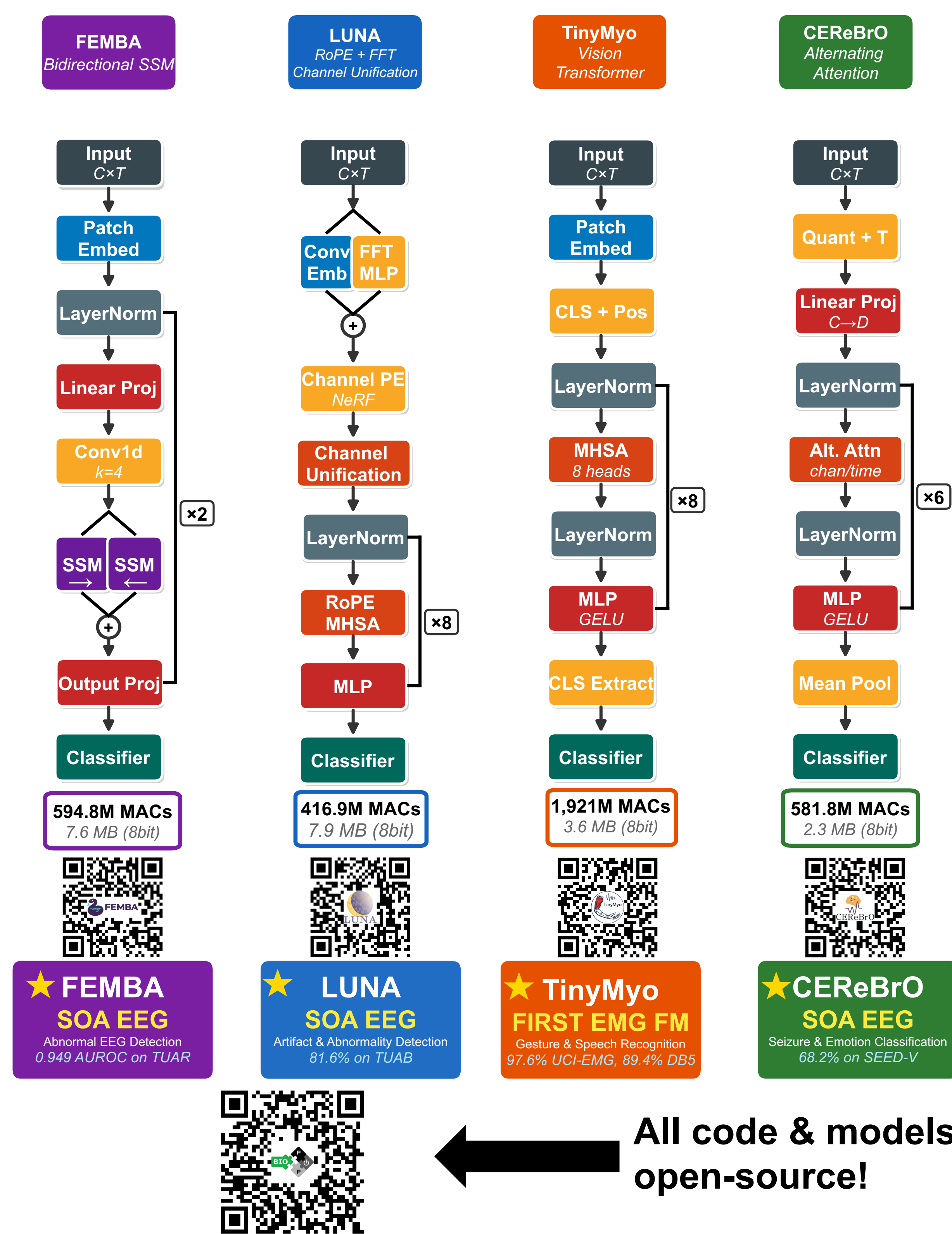
End-to-End Foundation Model Deployment Pipeline

PyTorch → Quantization → Code Generation → GAP9 Execution



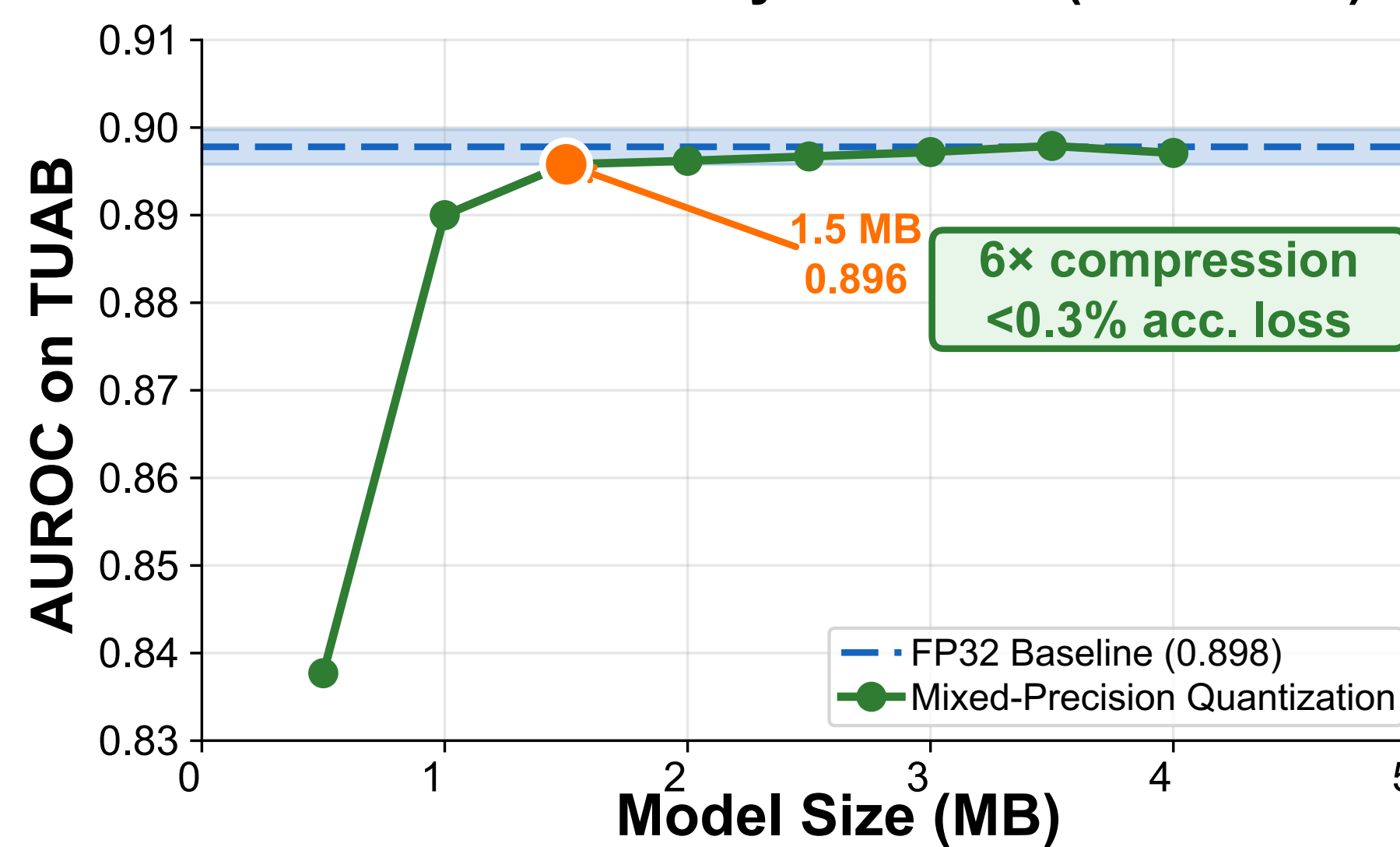
Model Architectures

FEMBA (EEG) | LUNA (EEG) | TinyMyo (EMG) | CEReBrO (EEG/ECG/PPG)

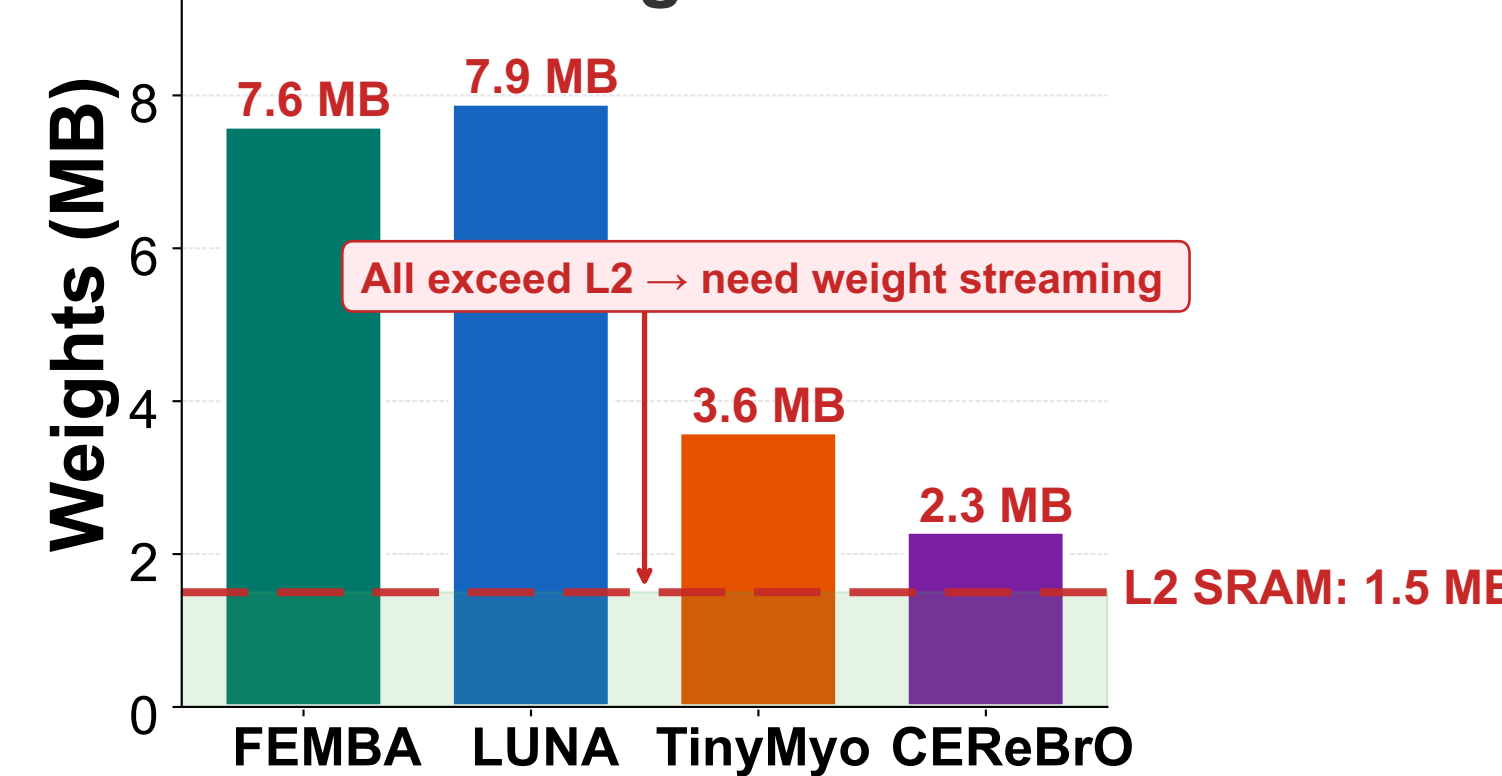


Performance Analysis

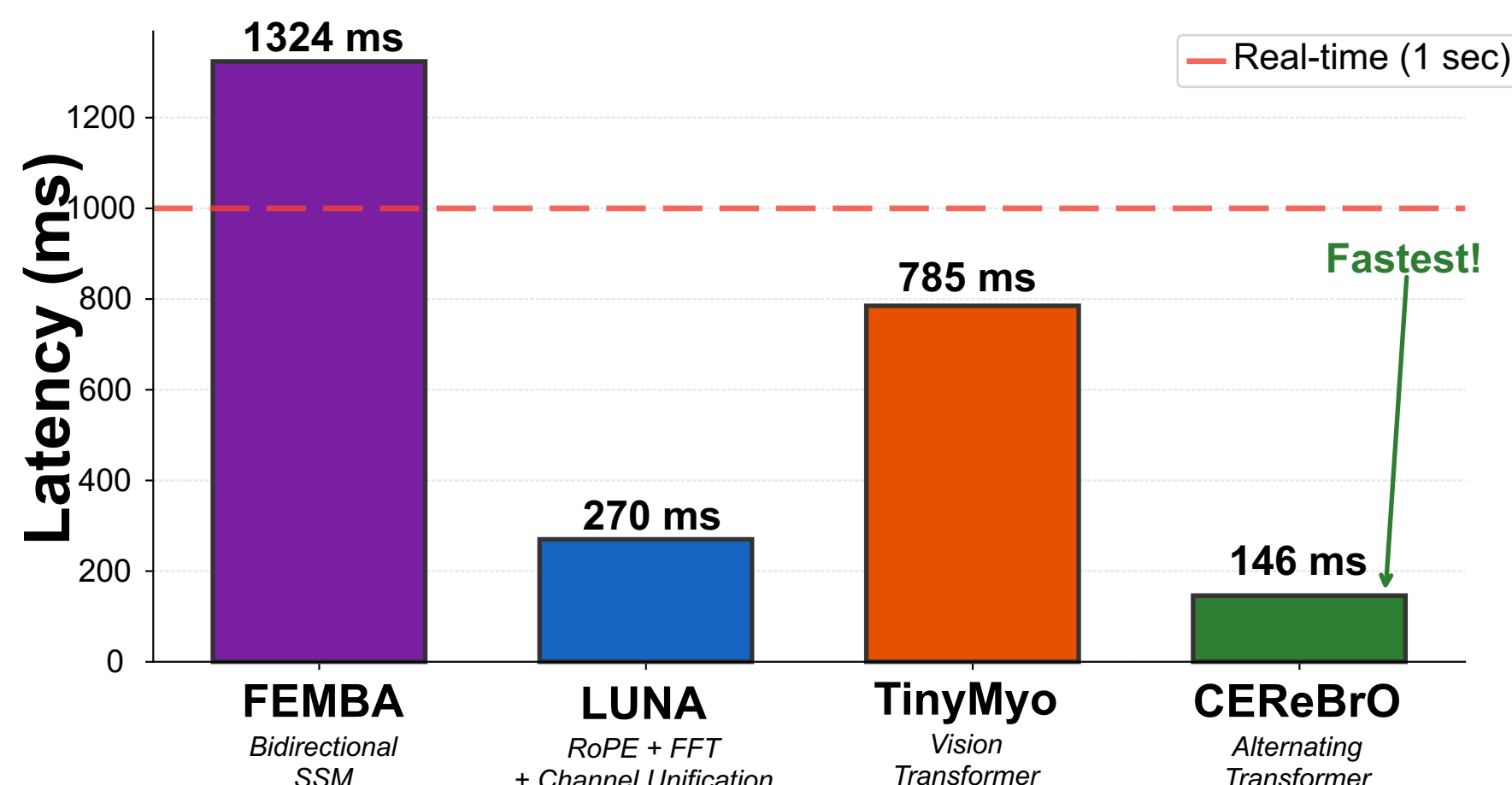
Quantization-Accuracy Trade-off (CEReBrO)



Model Weights vs GAP9 L2



Inference Latency @ 370 MHz

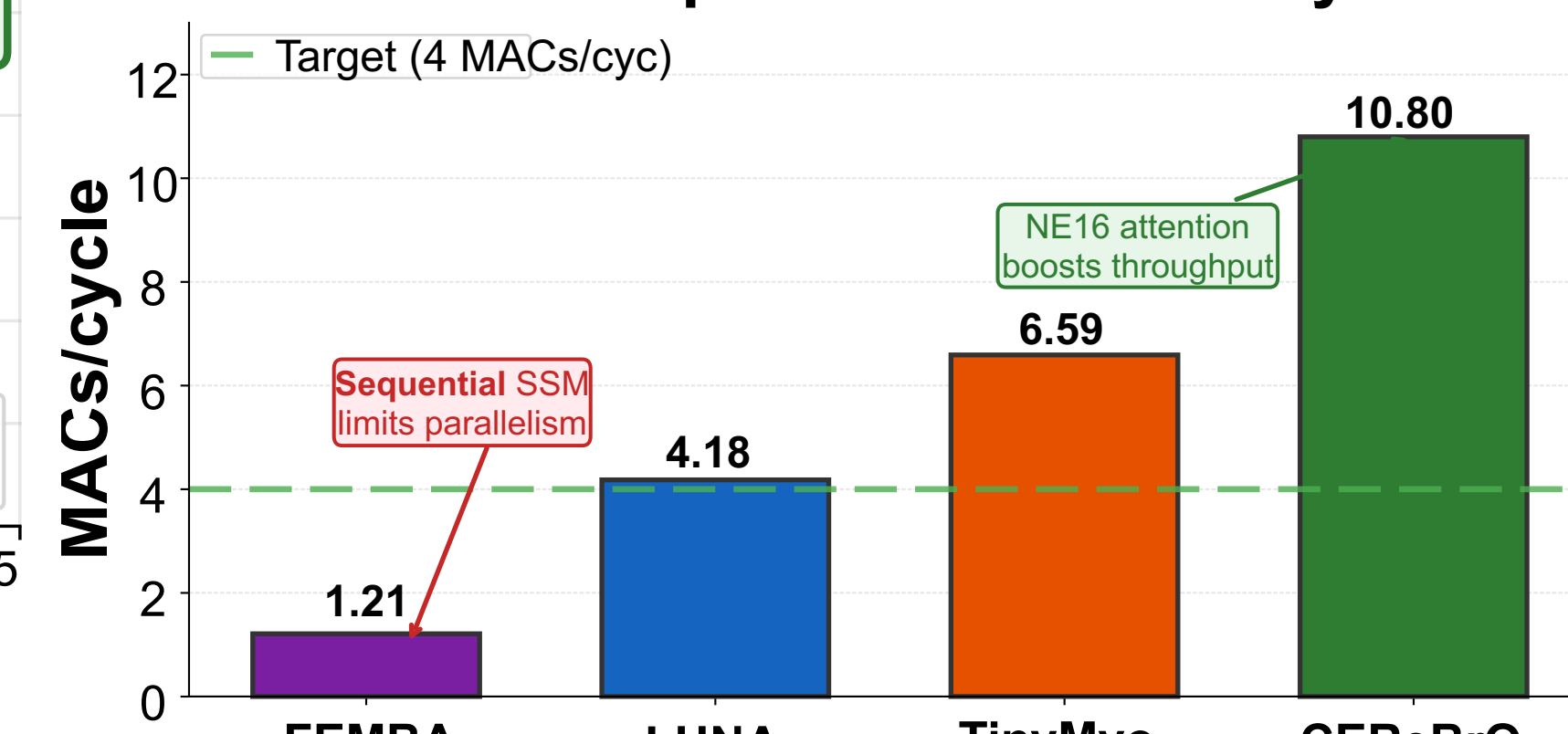


Quantization: FP32 → INT8

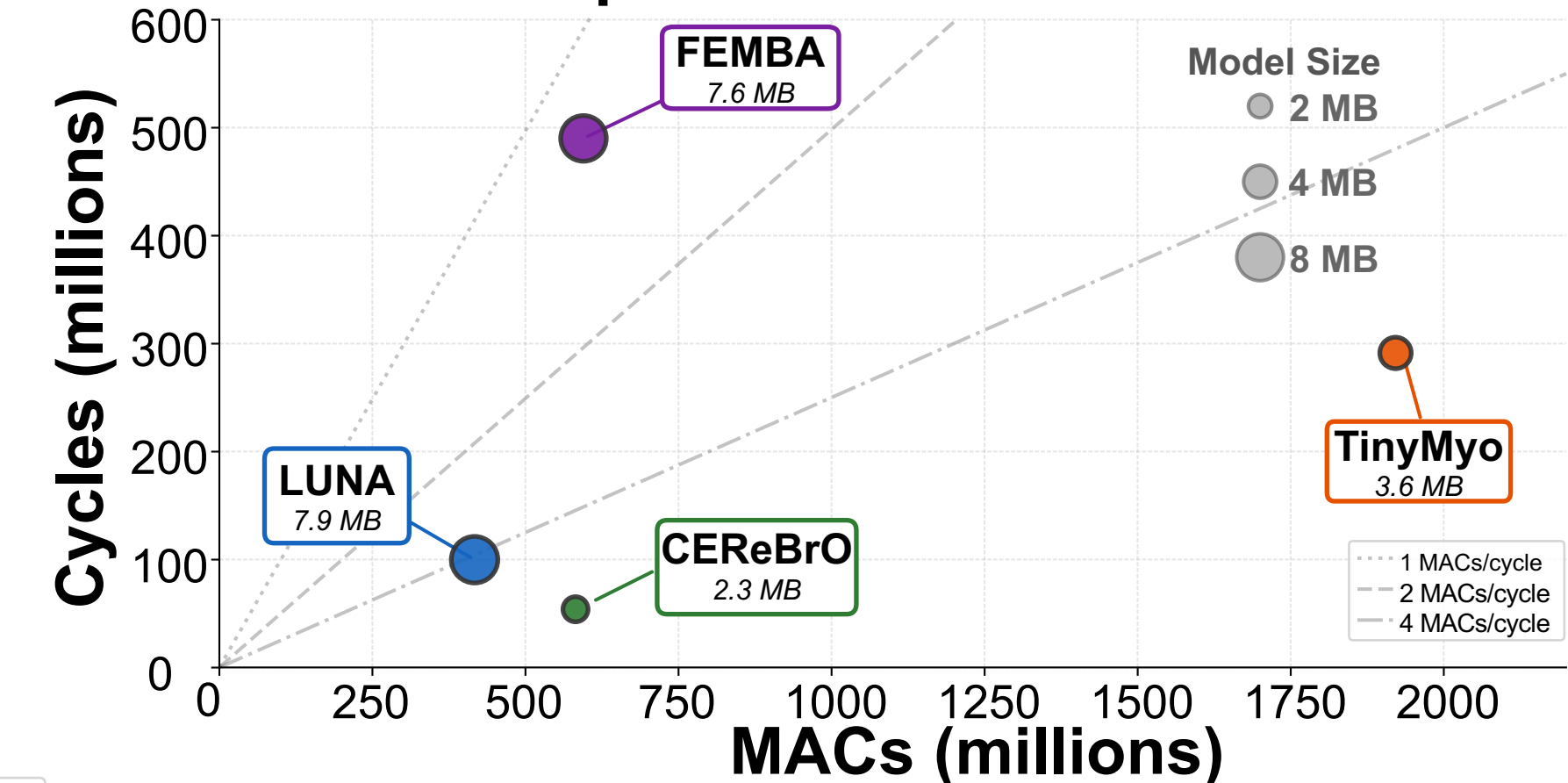
9x smaller model, minimal accuracy loss

Dataset	FP32 (9MB)	INT8 (1MB)	Δ
TUAB	0.898	0.890	-0.9%
CHB-MIT	0.877	0.863	-1.6%
SEED-VIG	0.933	0.930	-0.3%

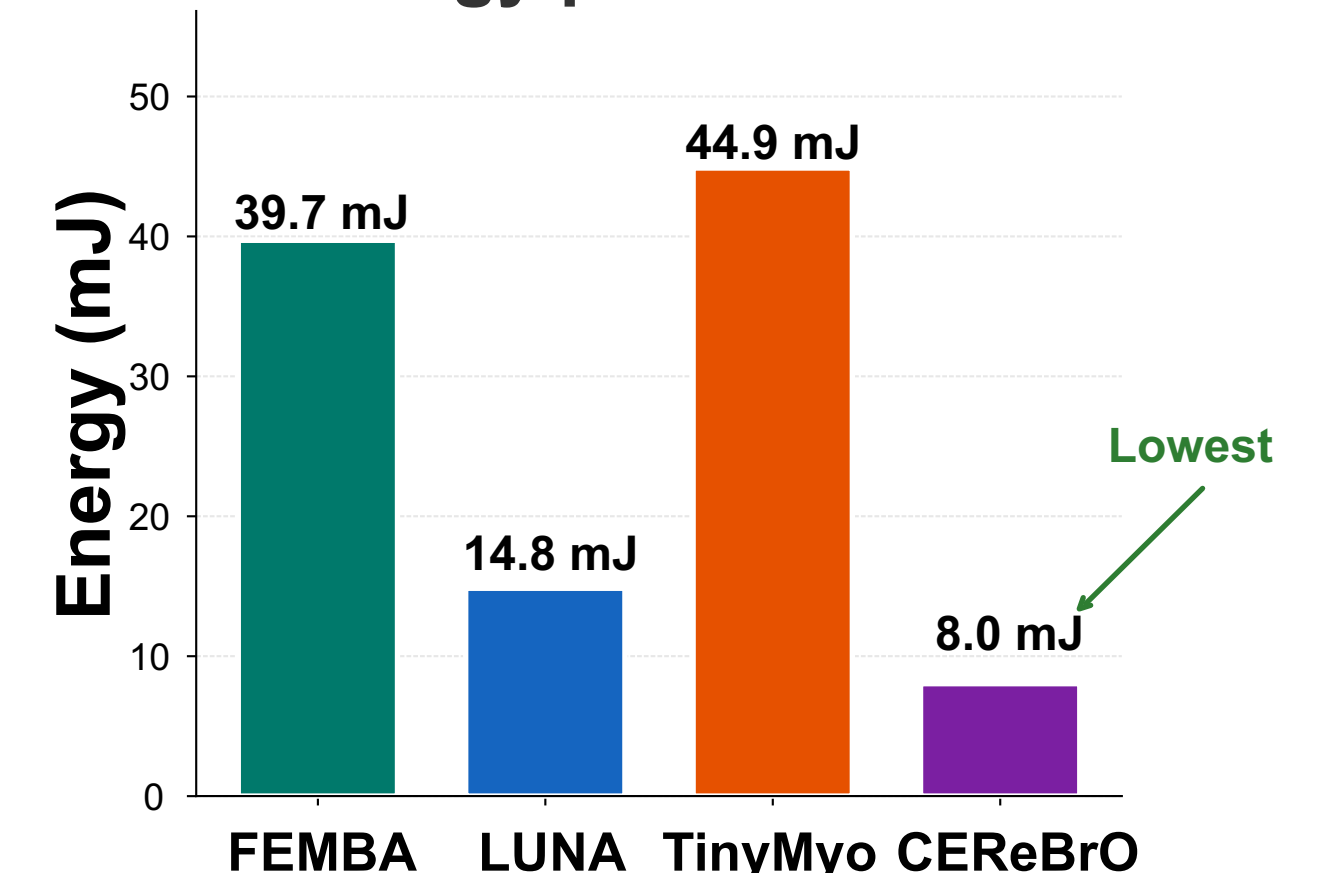
Computational Efficiency



Compute vs Execution Cost



Energy per Inference



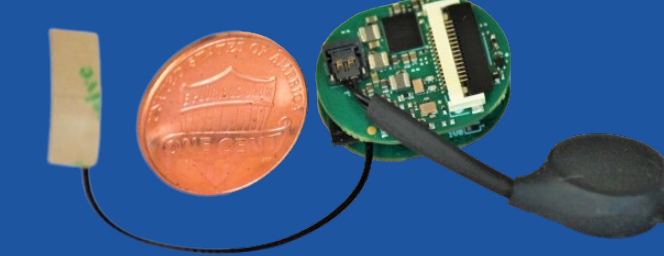
First Real-Time Deployment of State-of-the-Art Biosignal Foundation Models on Ultra-low-power Edge Device

CEReBrO: 146ms, 8mJ • FEMBA: up to 2bit quantized • TinyMyo: ~36mW always-on • LUNA Channel unification

Quantization Results
9x smaller models
<2% accuracy drop on TUAB/CHB-MIT

Efficient computation
>10 MACs/cycle
Boosted with Neural Accelerator

Energy efficient
~8mJ per inference
BioGap integration would allow for 2 days of Battery-Life



Stay in touch!

