ETH zürich

# Finetuning and Quantization of EEG-Based Foundational Biosignal Models on ECG and PPG Data for Blood Pressure Estimation

Integrated Systems Laboratory (ETH Zürich)

**Thorir Mar Ingolfsson,** Bálint Tóth, Dominik Senti, Jeffrey Zweidler, Alexandre Elsig, Luca Benini, Yawei Li

thoriri@iis.ee.ethz.ch

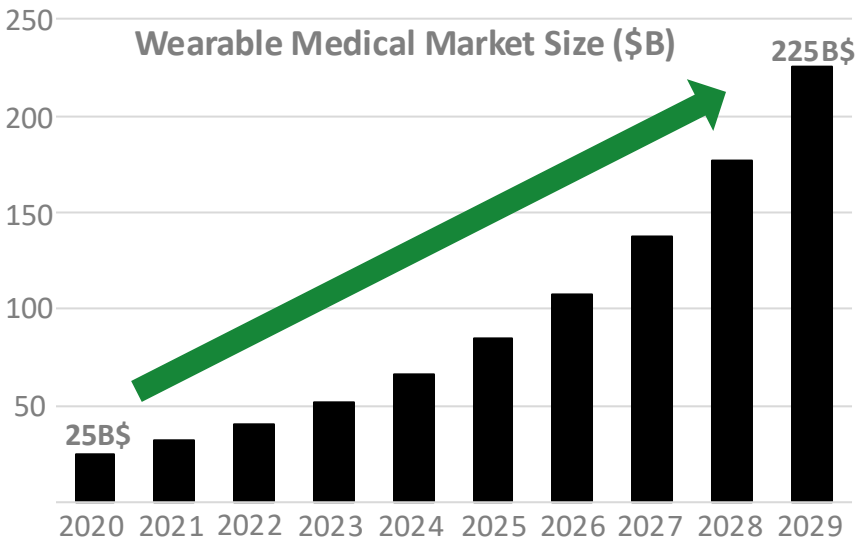

COPENHAGEN

IEEE EMBS IEEE Engineering Medicine & Biology Society

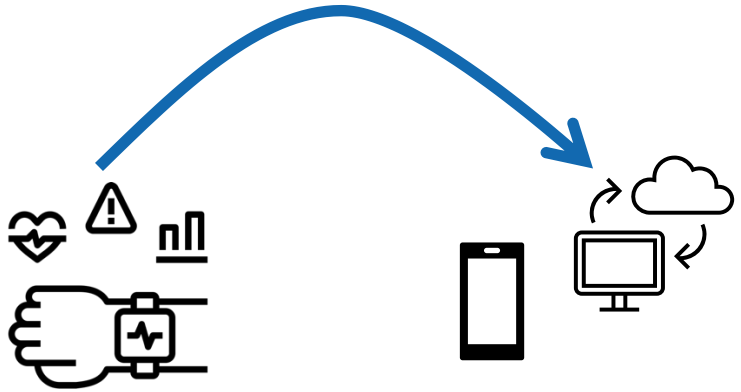
47TH ANNUAL
International Conference of the
IEEE Engineering in Medicine
and Biology Society

# The Wearable Revolution & Its Processing Gap

**Rapid Growth in Wearable Health Monitoring**

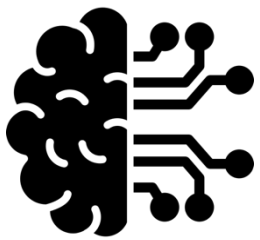



Heart Rate (**ECG,PPG**), Brain Signals (**EEG**) etc.

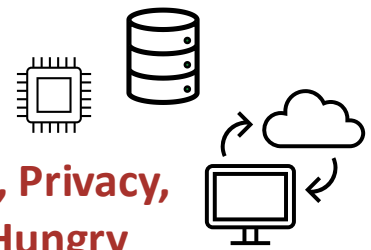


Wearables **gather data** and **send away**

**Edge devices**

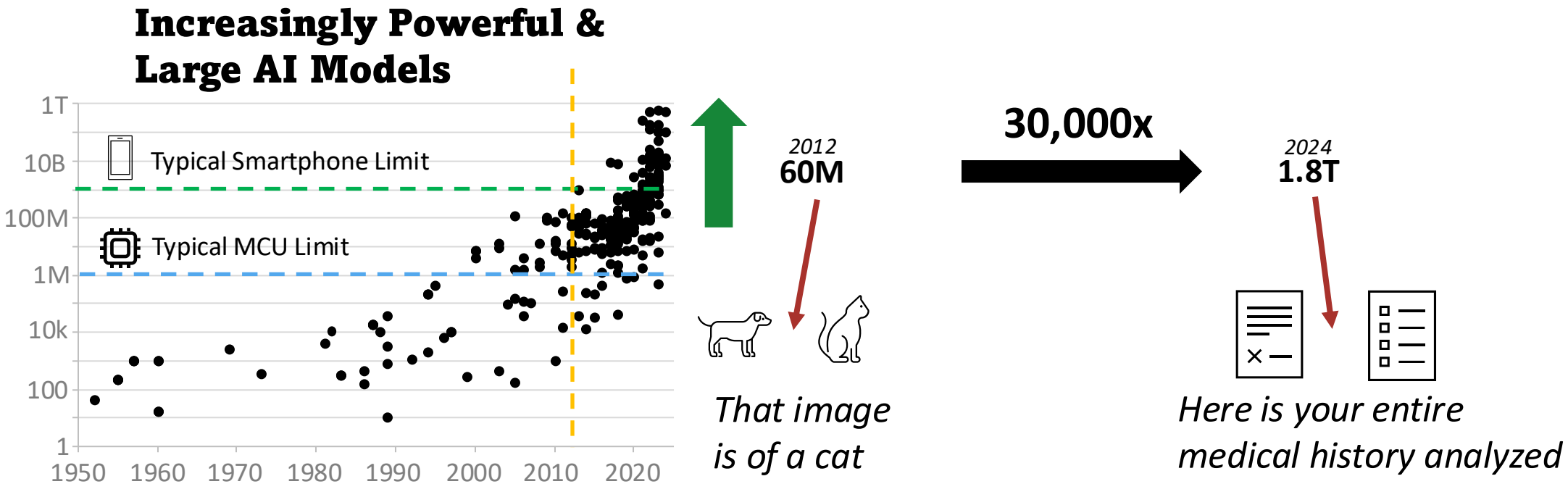


**Limited On-device processing!**



**Latency, Privacy, Power Hungry**

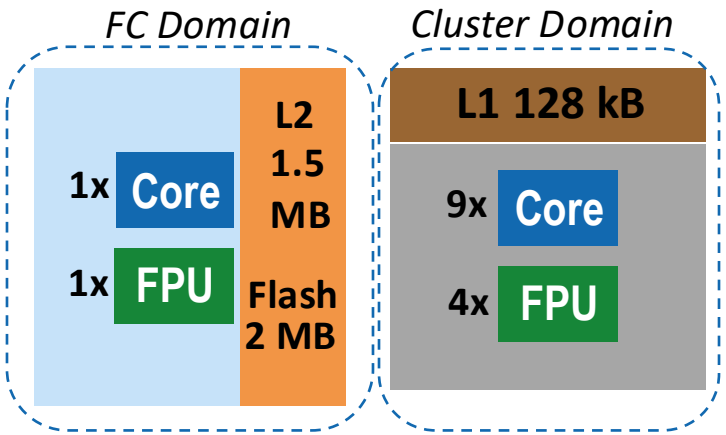
# A Widening Chasm: Powerful AI vs Tiny Devices



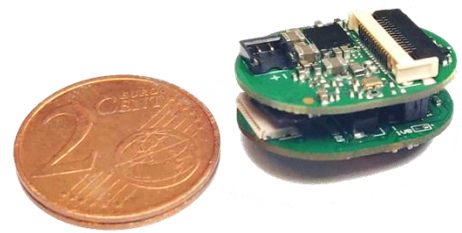

How do we push these **large brains** onto **small battery** powered **edge devices?**

# Algorithms for HW and HW for Algorithms


*GAP9*




**BioGAP**



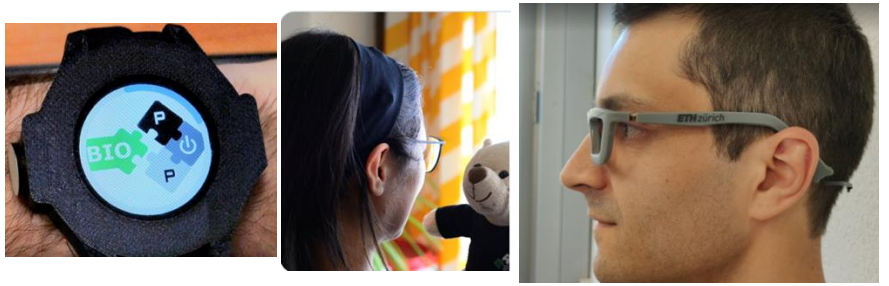
GREENWAVES TECHNOLOGIES

LUXOTTICA®

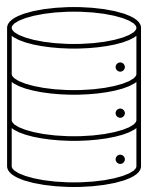
**Key limitations:**

**Memory:** **L1 128kB**
**L2 1.5MB**

**Speed:** **400MHz**
**9 Core** computation cluster



# The Next Frontier: From Specific to Universal Models



$\mathbb{X}$: Raw Data
**Cheap**

Heart Rate (ECG)
Brain Signals (EEG)
Muscle Activations (EMG)

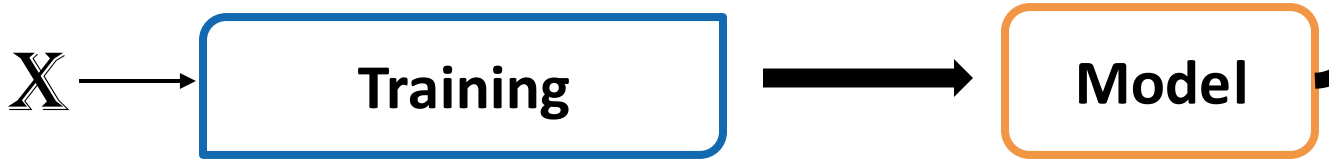
**~ 10TBs (20k+ hours).**

$\mathbb{Y}$: Labels
**Expensive**

Atrial Fibrillation, Seizure, etc.

**Supervised Learning**


**Self-Supervised Learning**

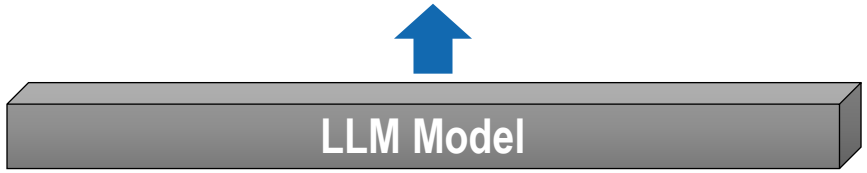



**Foundation**

# How do we train on unlabeled data?

LLMs are often pretrained in an intuitive way.

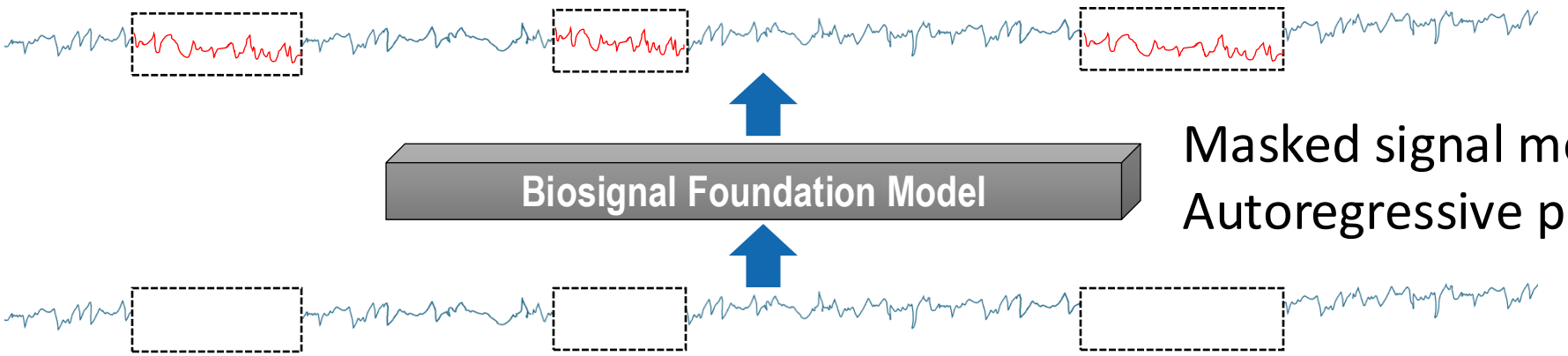
The ocean was very calm and the waves crashed into the rocks




Next token prediction
Masked token prediction

The [ ] was very [ ] and the waves [ ] into the rocks

We try to play a similar game in our biosignal pretraining.




Masked signal modeling
Autoregressive prediction

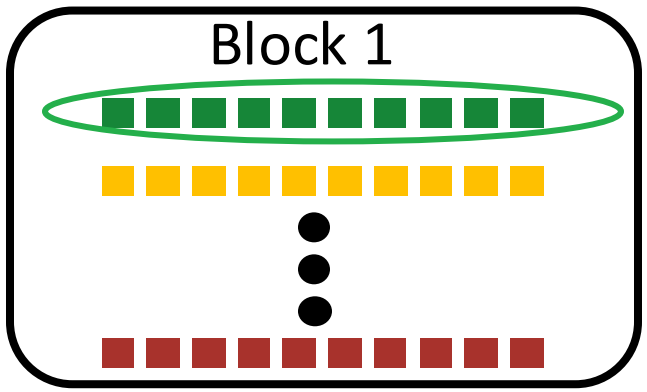
Other methods: Contrastive Learning

# CEReBrO's Alternating Attention to the Rescue


Standard Self Attention → **Expensive**

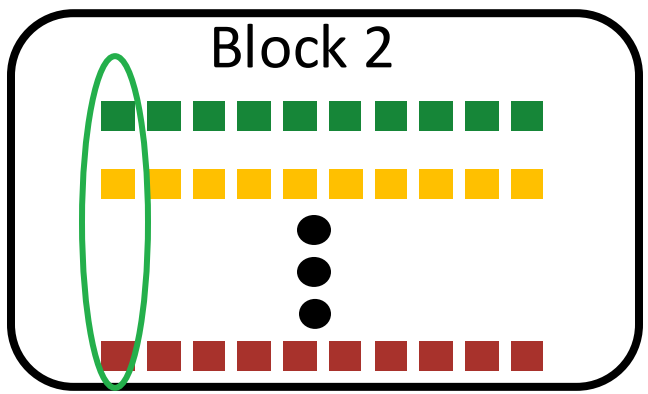
$O(N^2)$

$O(TN + CN)$



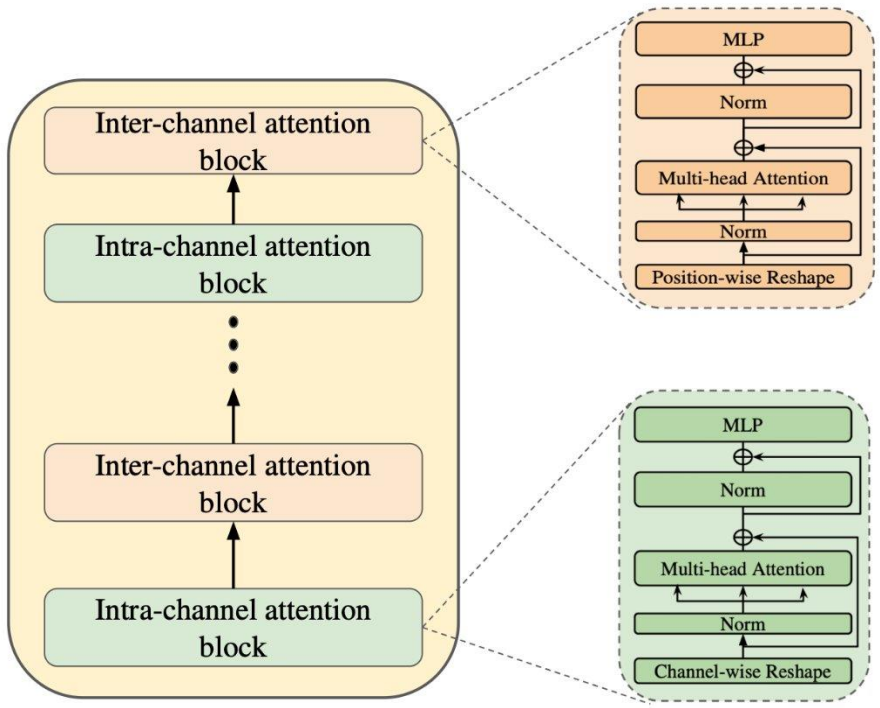

**Intra-Channel Attention**

Self-attention between tokens of the same channel




**Inter-Channel Attention**

Self-attention between tokens from the same timestep



# Biosignal Cross-Modality Knowledge Transfer

CEReBrO model trained **on EEG** – Can we utilize it for more tasks?

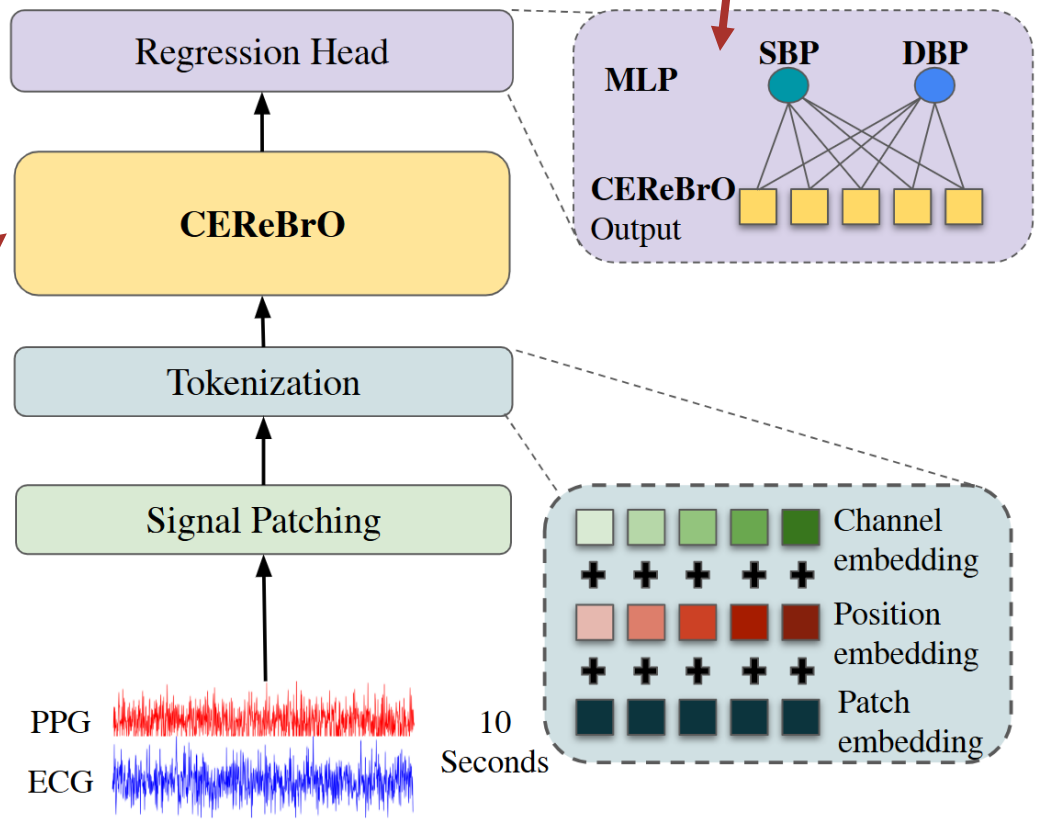
Can knowledge learned from **EEG waveforms** transfer effectively to other biosignals, such as **ECG** and **PPG**, for tasks such as BP estimation?

**MIMIC-III**
30,000 ICU Subjects

**VitalDB**
6153 Subjects

**PulseDB**
4941 + 3458 Records
125Hz ECG & PPG (10s Windows)

**Load**



# Cross-Modality knowledge Transfer brings SoA Results

## VitalDB

| Method | Diastolic Blood Pressure MAE | Systolic Blood Pressure MAE |
|---|---|---|
| ResUNet+Attention | 2.95 | 4.64 |
| CiGNN | 2.79 | 4.15 |
| Fusion+Attention | 3.76 | 5.32 |
| **Our CEReBrO** | **1.92** | **3.14** |

## MIMIC-III

| Method | Diastolic Blood Pressure MAE | Systolic Blood Pressure MAE |
|---|---|---|
| ResUNet+Attention | **1.13** | 4.55 |
| ResNet+LSTM | 2.23 | 4.12 |
| Conv-LSTM | 3.29 | 4.30 |
| **Our CEReBrO** | 1.57 | **2.72** |

**Lower → better**

**SoA results** → Cross Modality Knowledge Transfer **works!**

Best CEReBrO models quantized down to around **3.8 MB** → **Fits on edge devices**

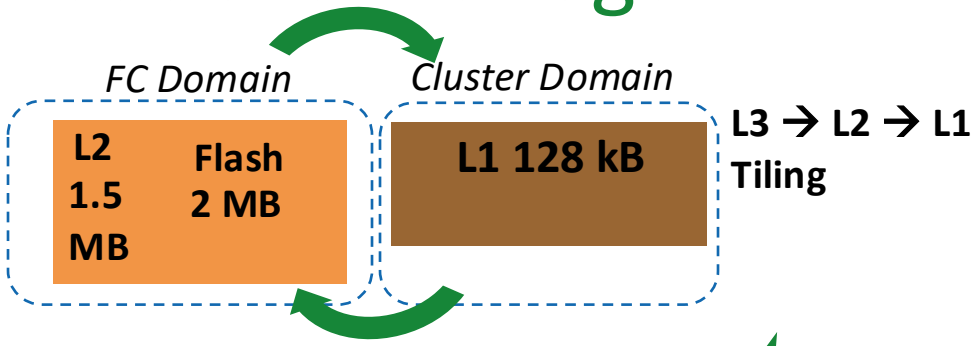
# Bringing Foundation Models to the Edge

**CEReBrO**

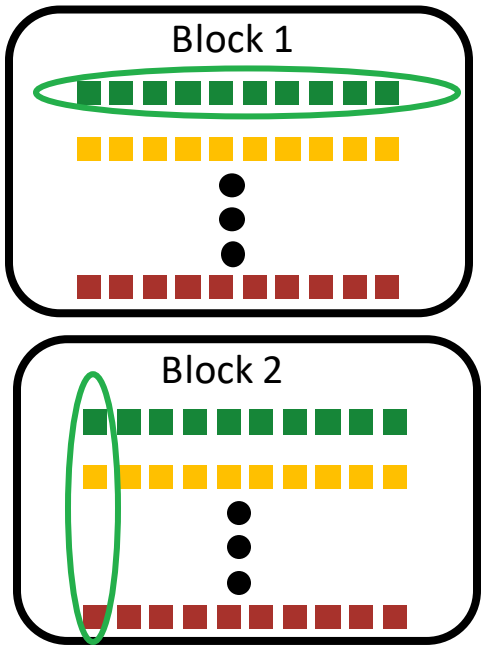
Smallest model:
3.5M parameters

15.2 MB in FP32 → **3.8 MB in INT8**


*GAP9*



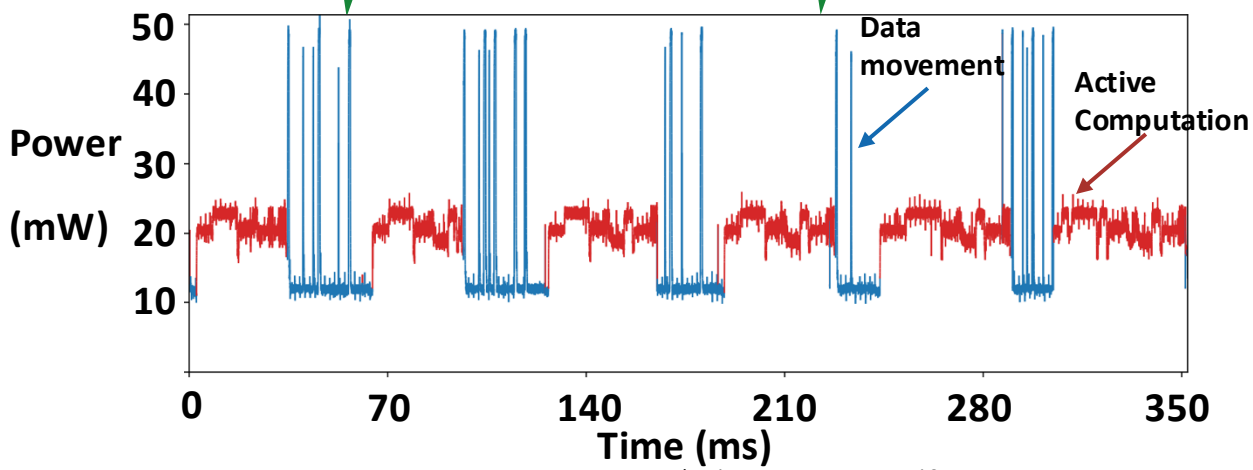

**L3 → L2 → L1**
**Tiling**




**Specific kernels written to handle alternating attention**

**10 mJ per inference**

**0.35s Runtime!**



# Cross-Modality and State-of-the-art Results

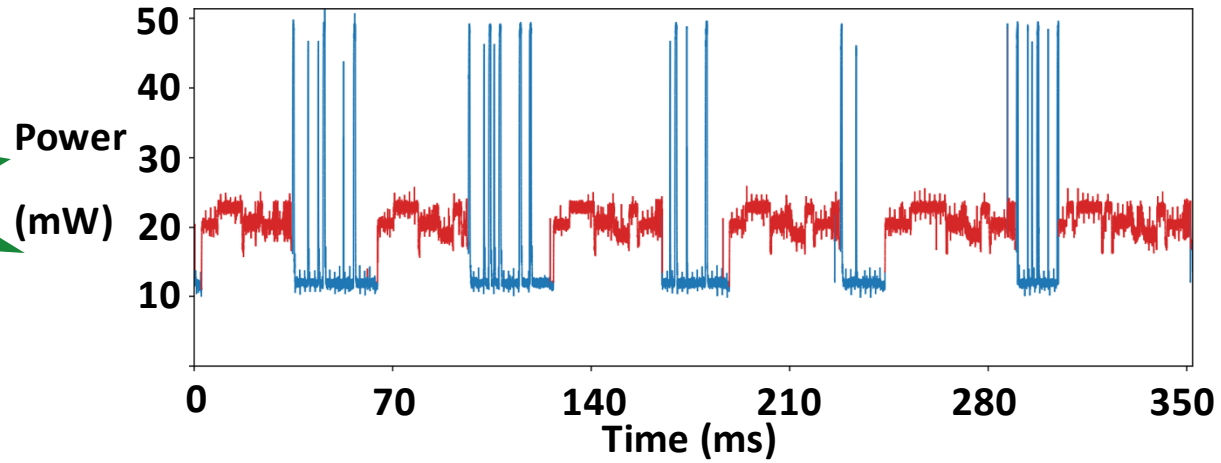
**First EEG-based Foundation Model** used for ECG/PPG usage
→ **Cross-Modality Knowledge Transfer**

**State-of-the-art accuracy** diastolic BP (**MAE 1.57 mmHg**) and surpasses **by 1.5×** the accuracy of prior works for systolic BP (**MAE 2.72 mmHg**).

**Contact info**

**10 mJ per inference**

**0.35s Runtime!**



Thank you!