



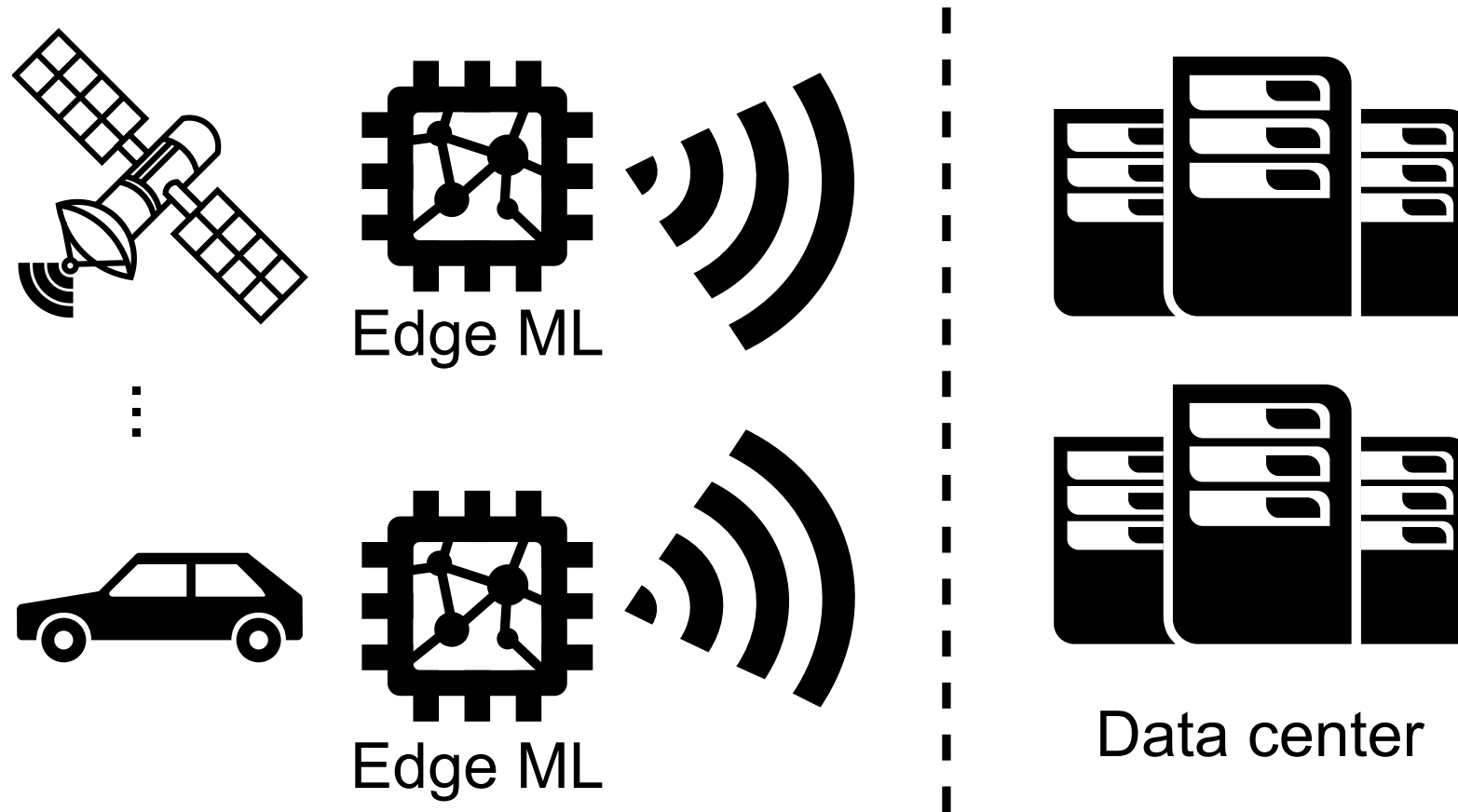
No Resource Left Behind: Machine Learning at the Edge

Robert Bayer

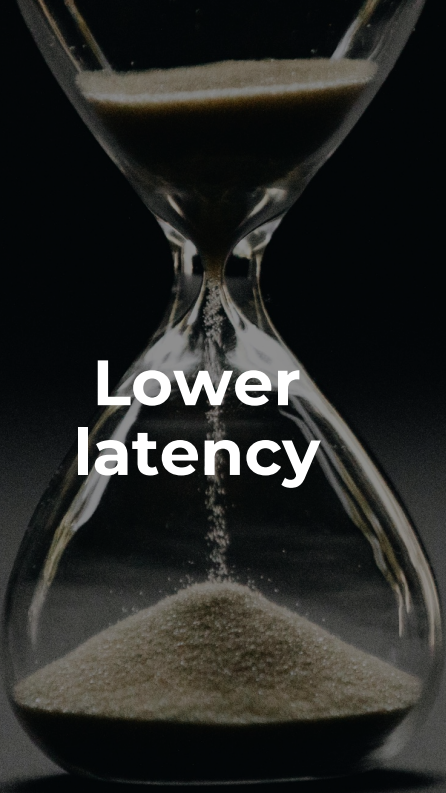
EuroSys Doctoral Workshop

22nd April 2024

Machine Learning at the Edge



Benefits

An hourglass with sand falling from the top bulb to the bottom bulb, symbolizing the passage of time and low latency.

**Lower
latency**

A dense traffic jam on a city street, with many cars and taxis stuck in a line, symbolizing network congestion.

**Reduced network
congestion**

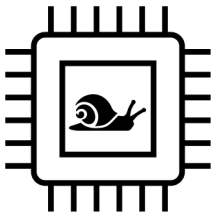
A wall covered with numerous black surveillance cameras, symbolizing increased privacy and security.

**Increased
privacy**

Challenges

Limited resource

Compute

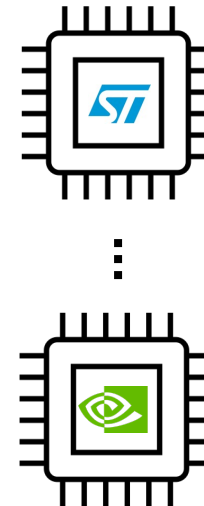


Power budget



Heterogeneity

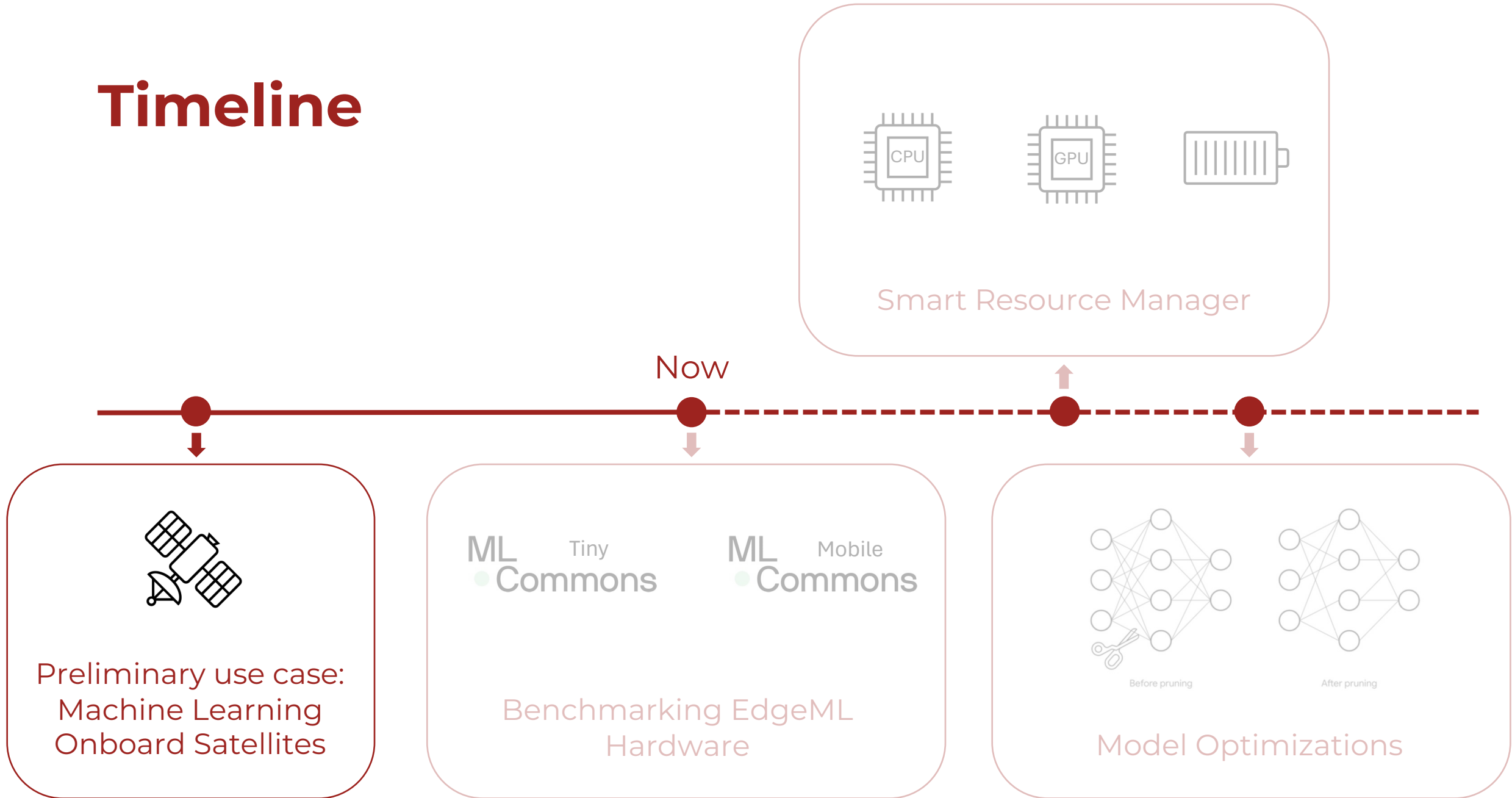
Hardware



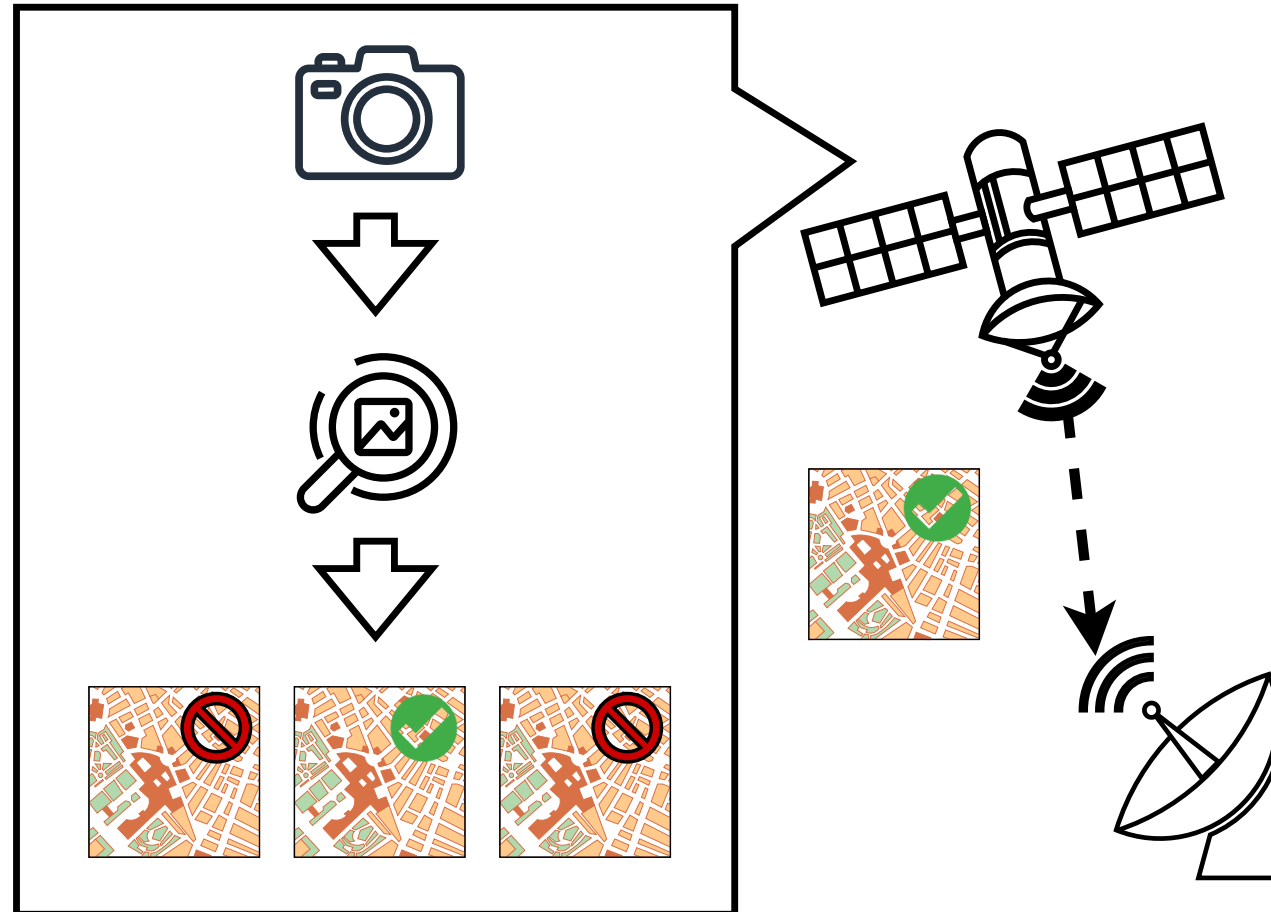
Tooling



Timeline



Preliminary use case: Machine Learning Onboard Satellites



Requirements



Real-time imaging
4.42 s latency

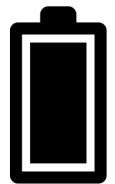


Arctic region
71.74 s latency



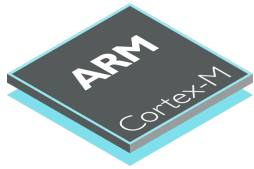
Greenland
270 s latency

<5W
MAX



<2W
AVG

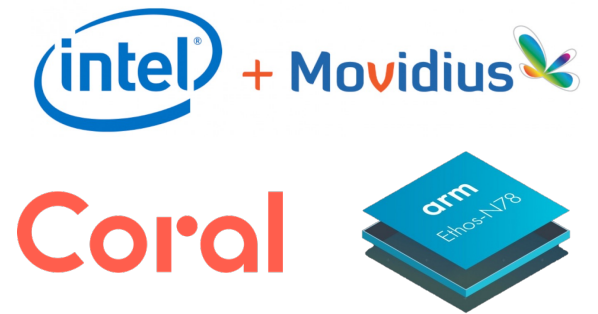
Devices under Test



Microcontroller



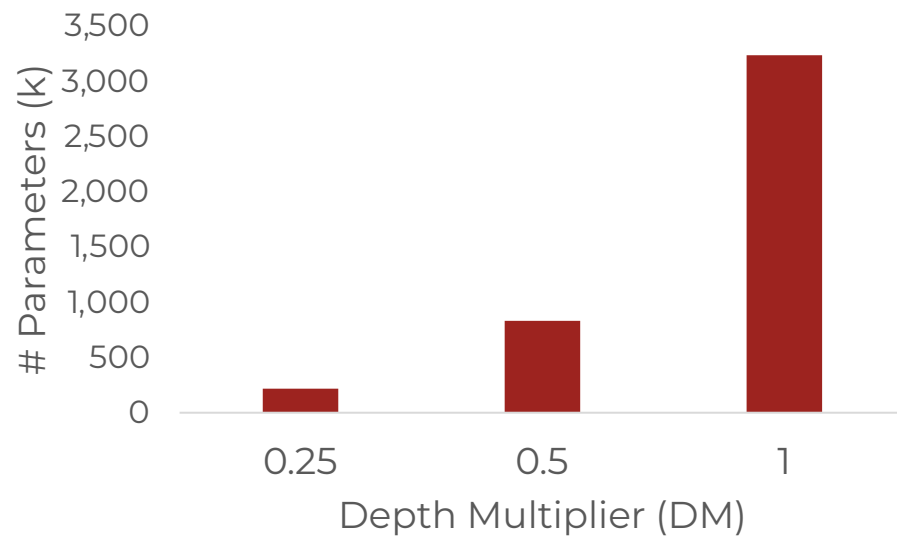
GPUs



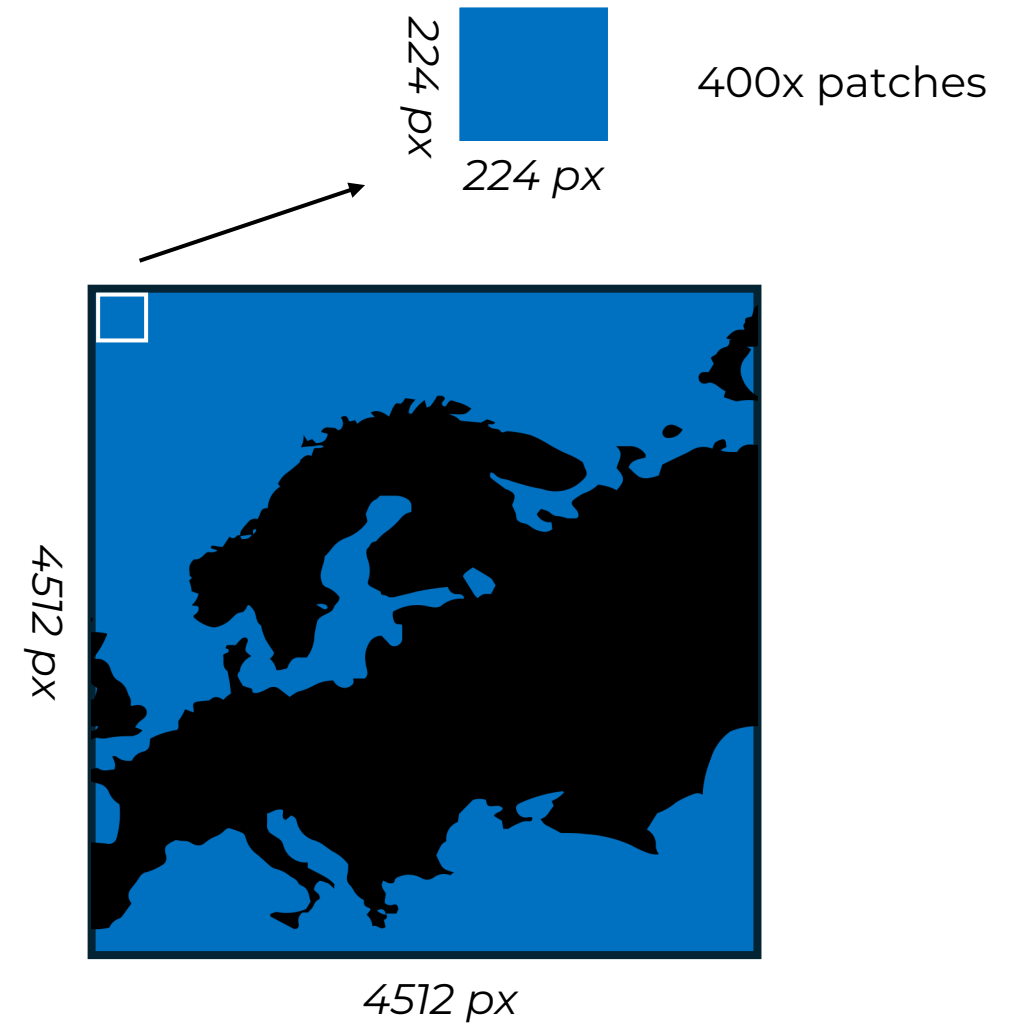
ASICs

Model

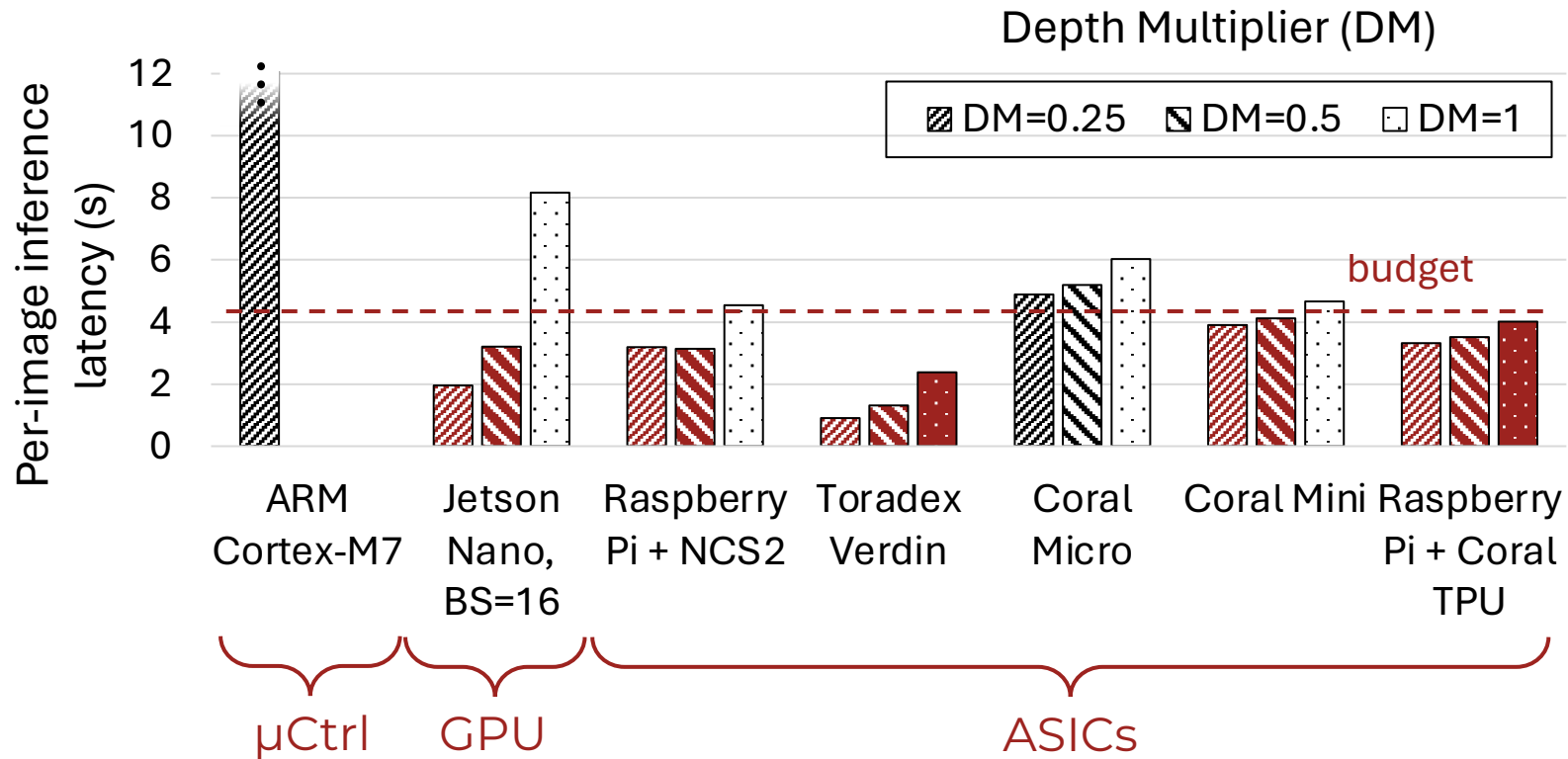
Pretrained MobileNetV1



Data

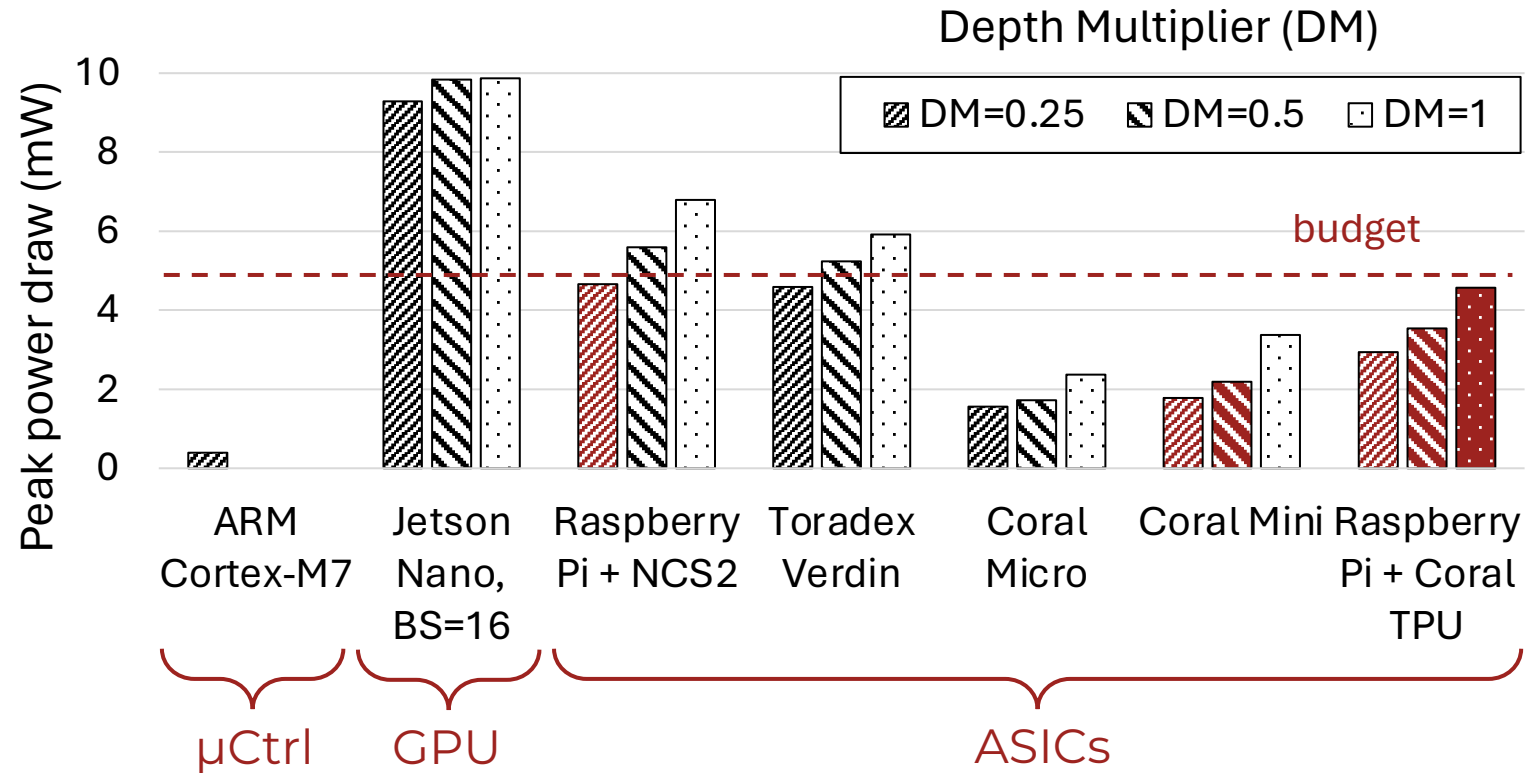


Results – Latency



High degree of parallelism necessary

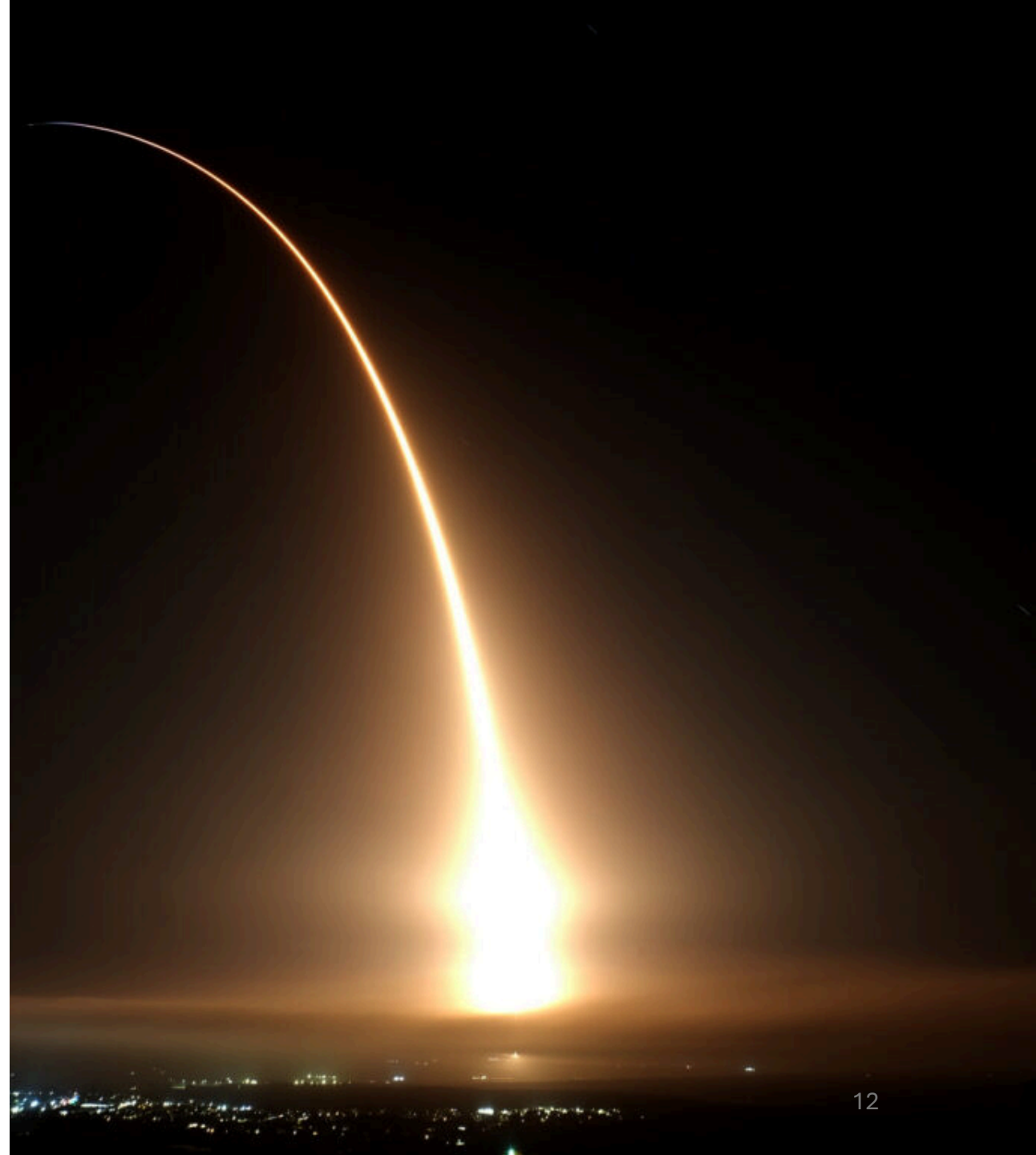
Results – Peak Power Draw



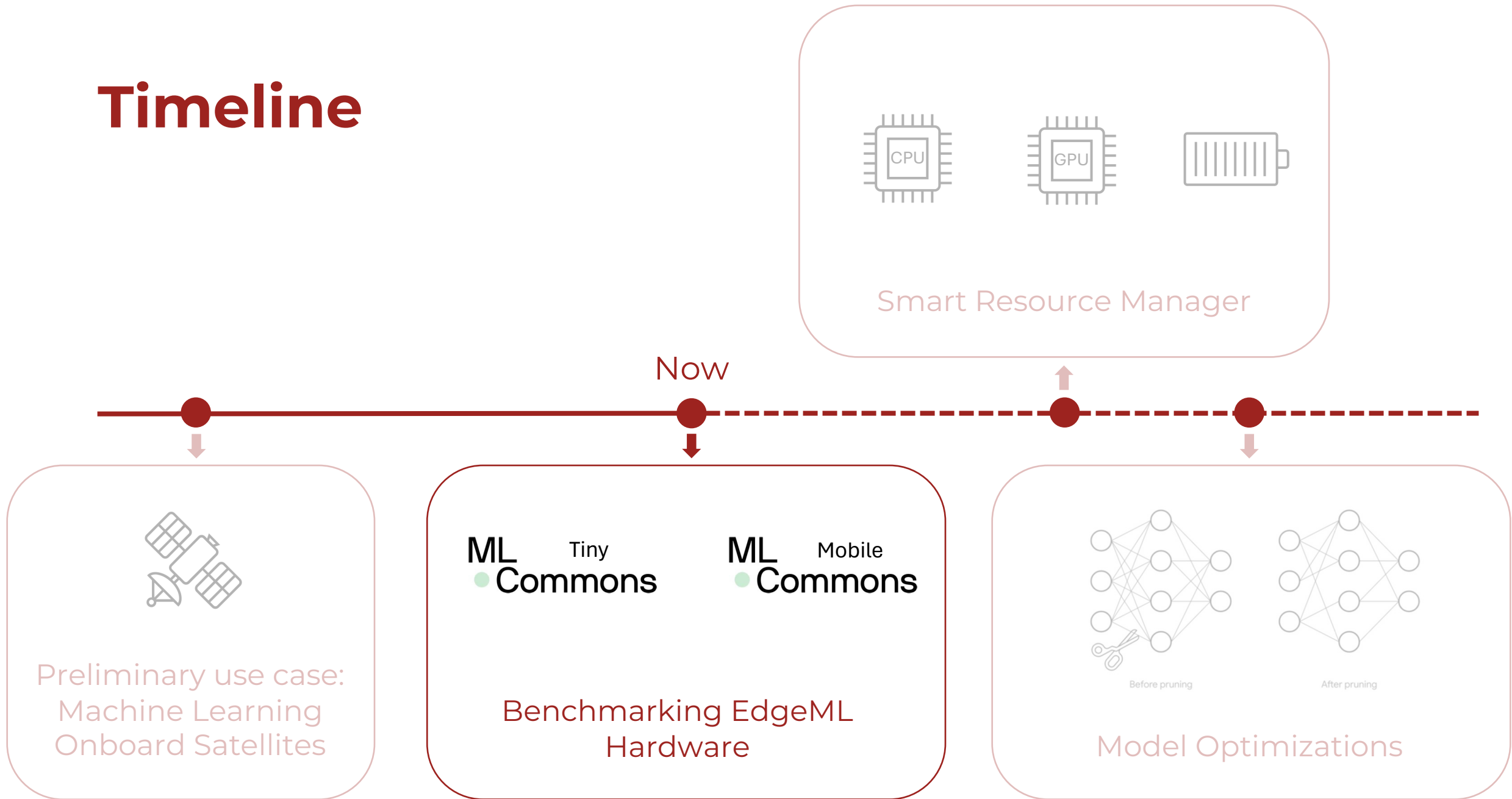
Only the highest specialization
can also fulfill power budget



***Coral Dev Board
Mini launched
onboard DISCO1 in
April 2023***

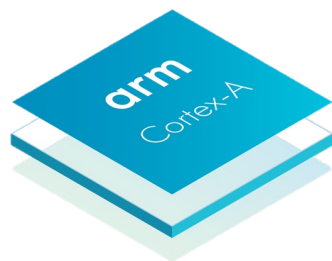
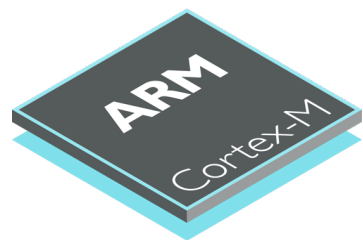


Timeline



Benchmarking EdgeML Hardware

CPUs



Accelerators



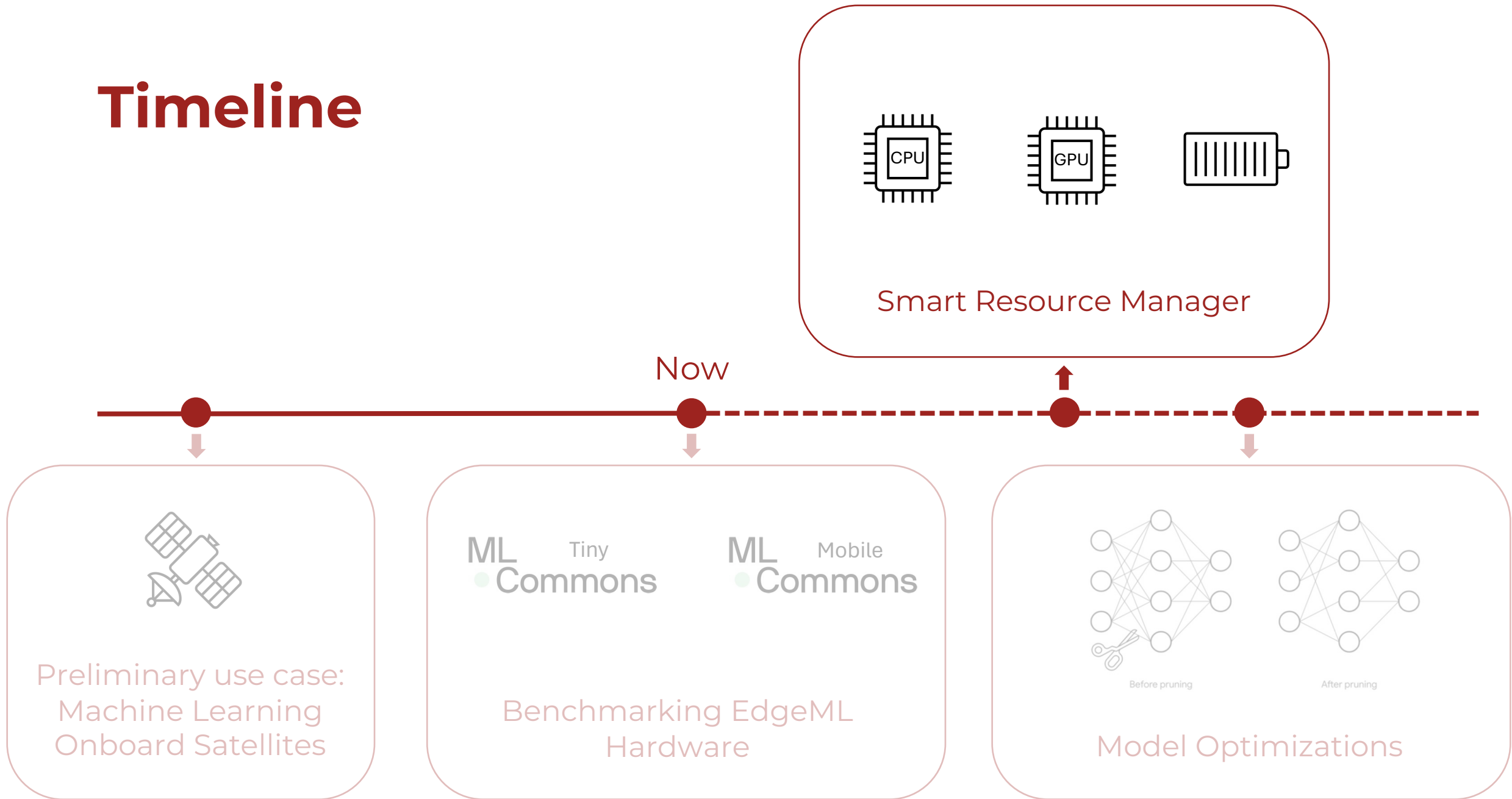
Coral



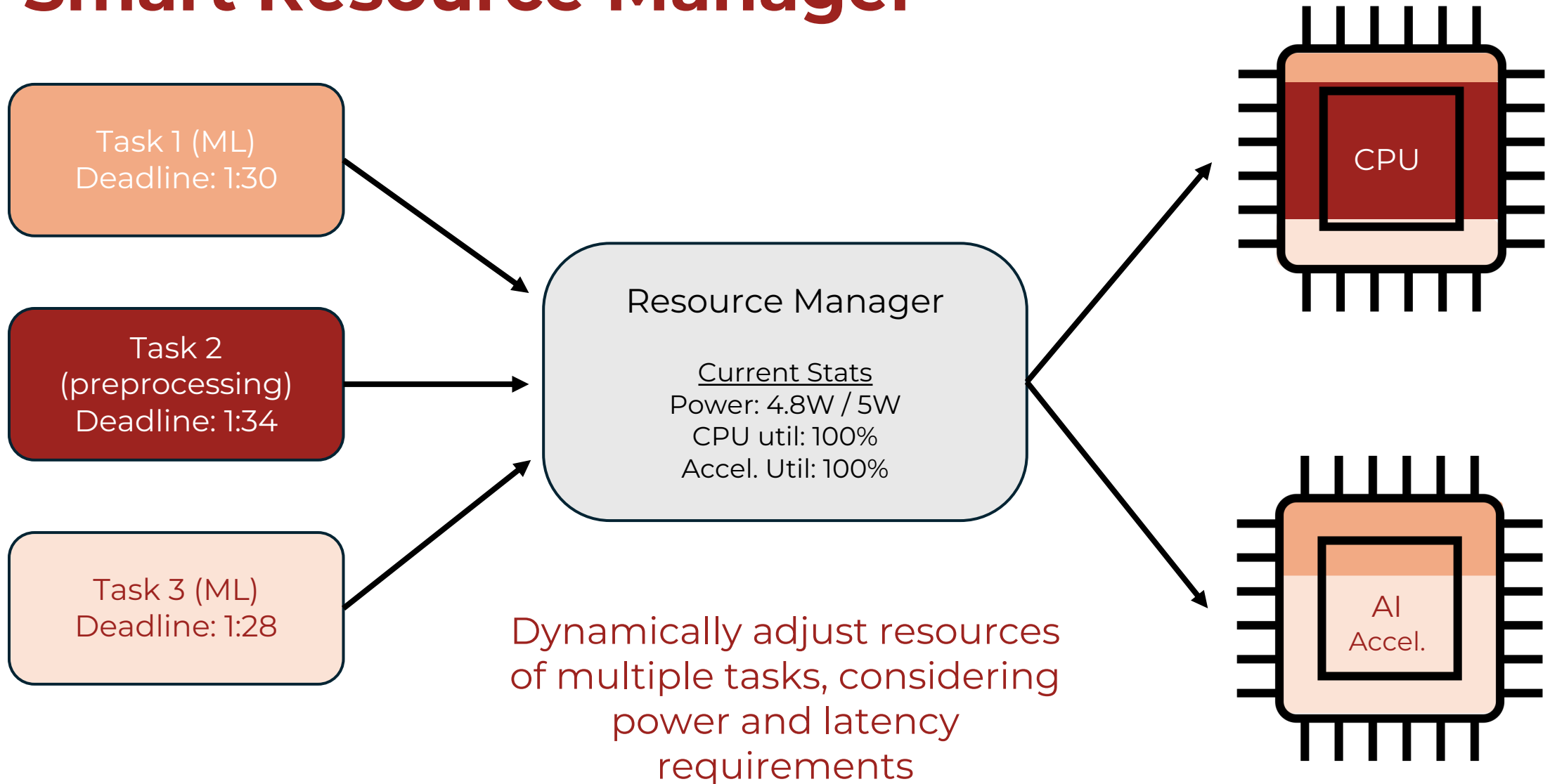
ML
● Tiny
Commons

ML
● Mobile
Commons

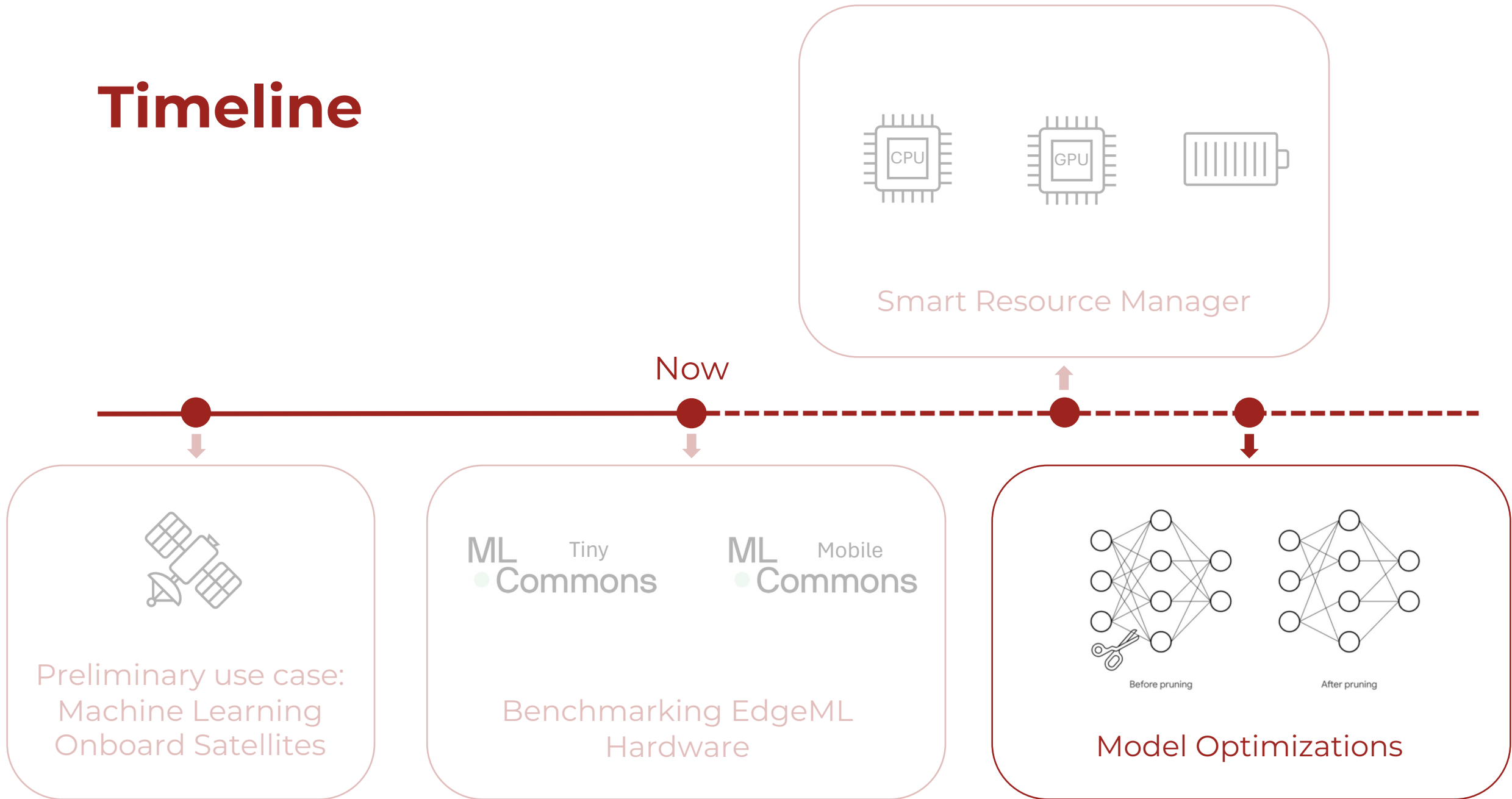
Timeline



Smart Resource Manager



Timeline

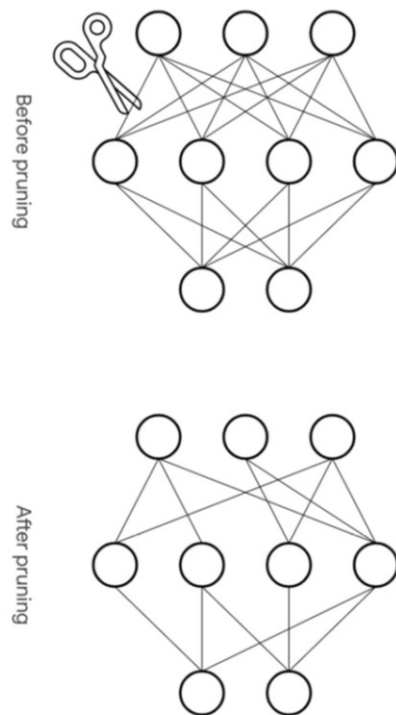


Model Optimizations

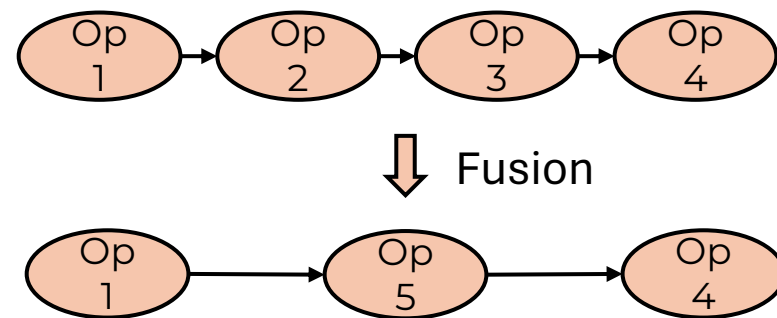
Quantization

$FP32 \rightarrow INT8$

Pruning



Operator fusion and other graph optimizations



Summary

Machine Learning Onboard Satellites

Benefit of reducing network congestion

Benchmarking EdgeML Hardware

Understanding trade-offs is expensive and difficult

Smart Resource Manager

Concurrent tasks, deadlines, power budgets

Model Optimizations

Make most out of devices

Thank you!



Robert Bayer
roba@itu.dk

EuroSys Doctoral
Workshop

22nd April 2024