IT UNIVERSITY OF COPENHAGEN

From Image Analysis in Space to Complex Pipelines at the Edge

Robert Bayer

Google, Sunnyvale

19th September 2024

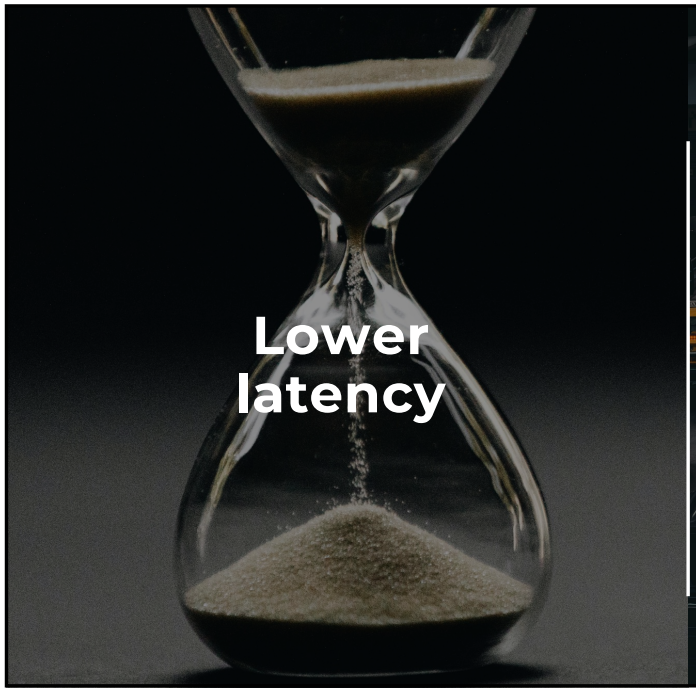
RAD
rad.itu.dk

IT UNIVERSITY OF COPENHAGEN
itu.dk

DASYA
Data-Intensive Systems and Applications
dasya.itu.dk

novo nordisk fonden

Benefits


Lower
latency

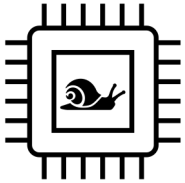

Reduced network
congestion


Increased
privacy

Challenges

Limited resource

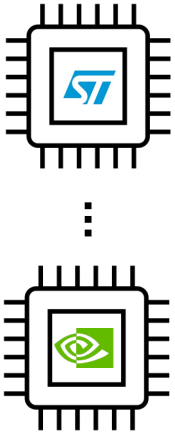
Compute



Power budget

Heterogeneity

Hardware



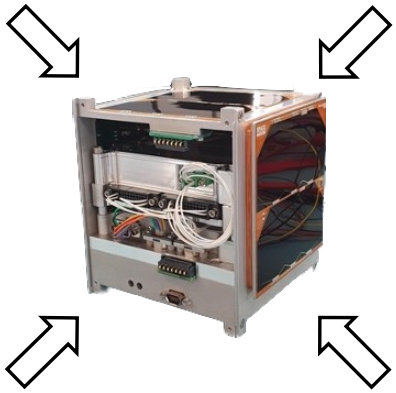
Tooling

Small Satellites

Benefits

$$
⇩
$

Reduced cost



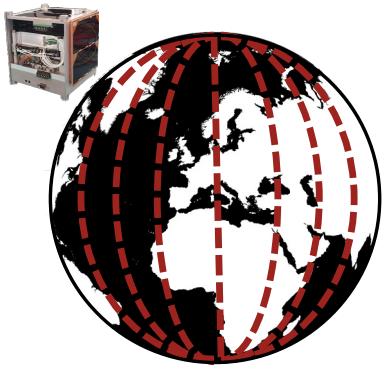
Shrinking
+
Standardization

Compromises

Reduced
Power
Generation

Problem

Can transfer MAX
49.1 images/day



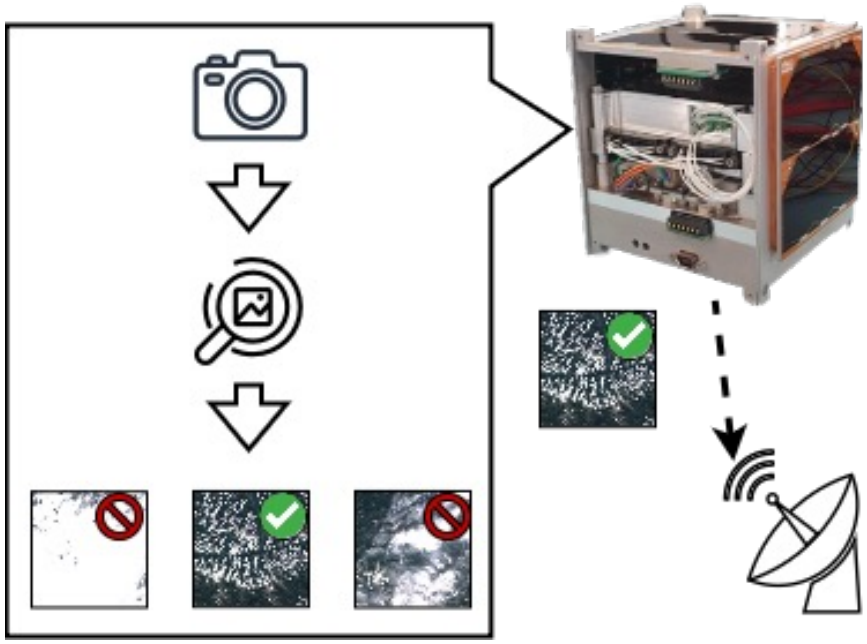
Real-time imaging
4.42 s latency

19,547 images
captured / day

<5W
MAX

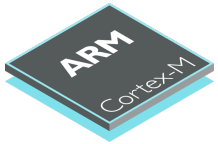
<2W
AVG

Solution - Machine Learning



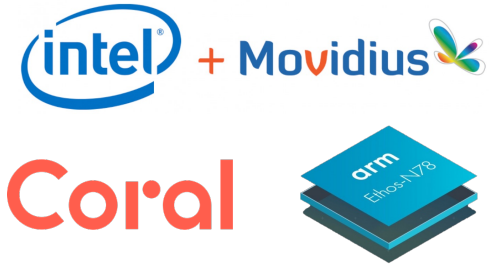
Our goal: Determine the right edge device to deploy on the satellite for this task.

Devices under Test



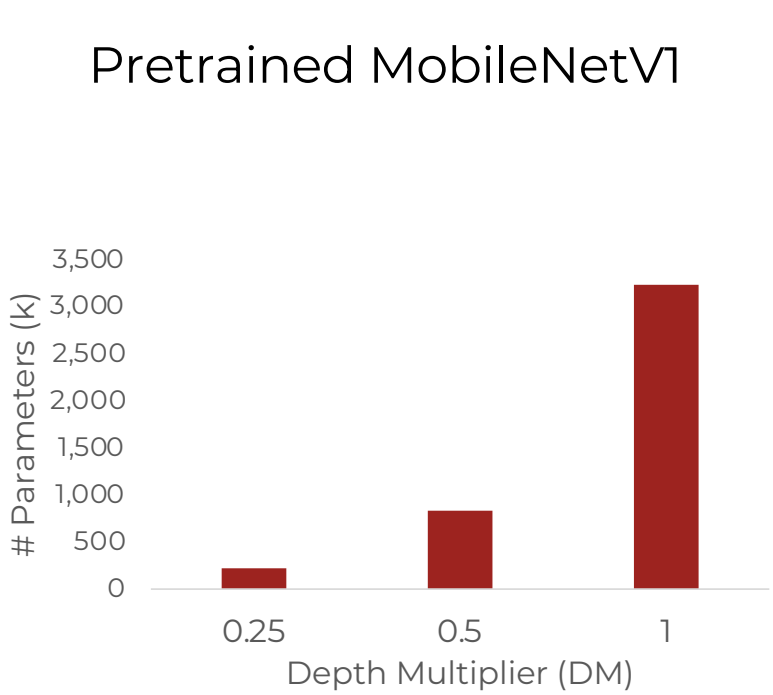
Microcontroller

GPUs

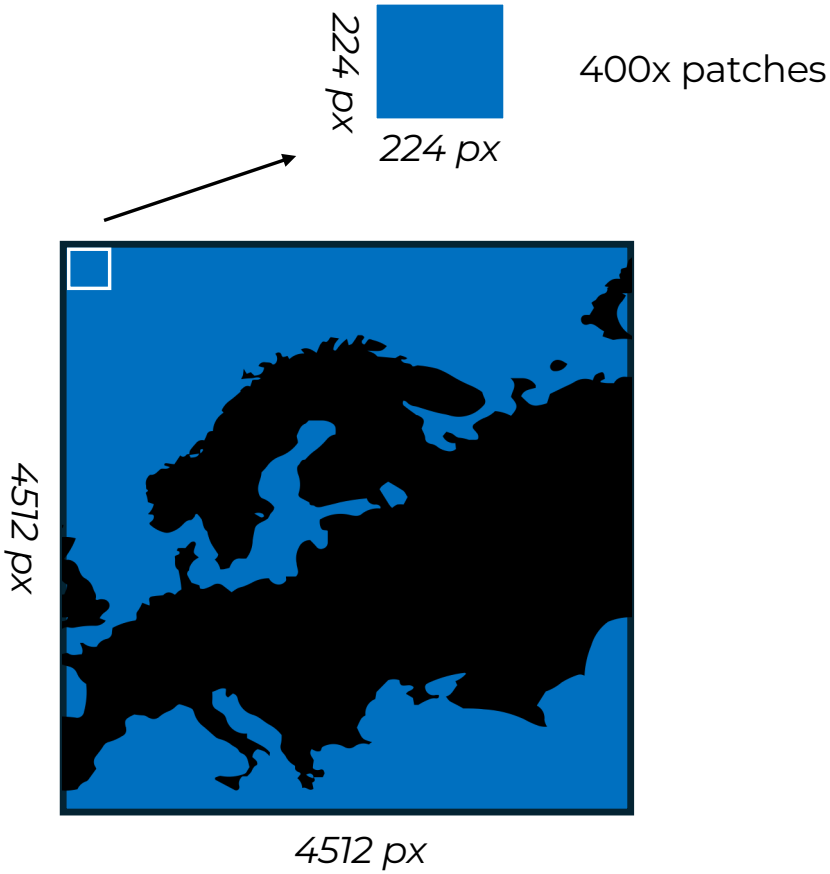


ASICs

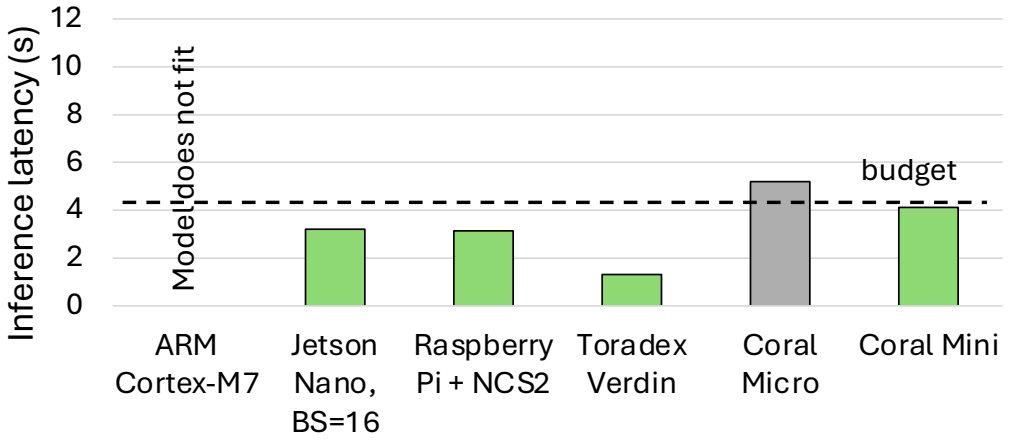
Model

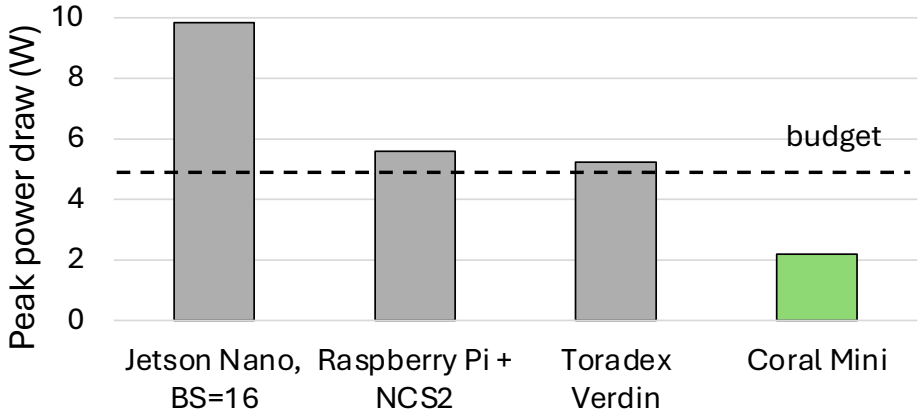

Pretrained MobileNetV1
Parameters (k)
3,500
3,000
2,500
2,000
1,500
1,000
500
0
0.25
0.5
1
Depth Multiplier (DM)

Data

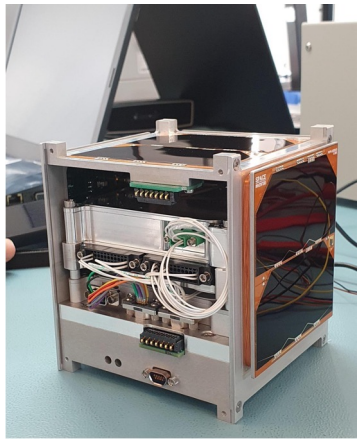

224 px
224 px
400x patches
4512 px
4512 px

Results





Coral Mini best candidate!




Coral launched
onboard DISCO1 in
April 2023

DISCO 2

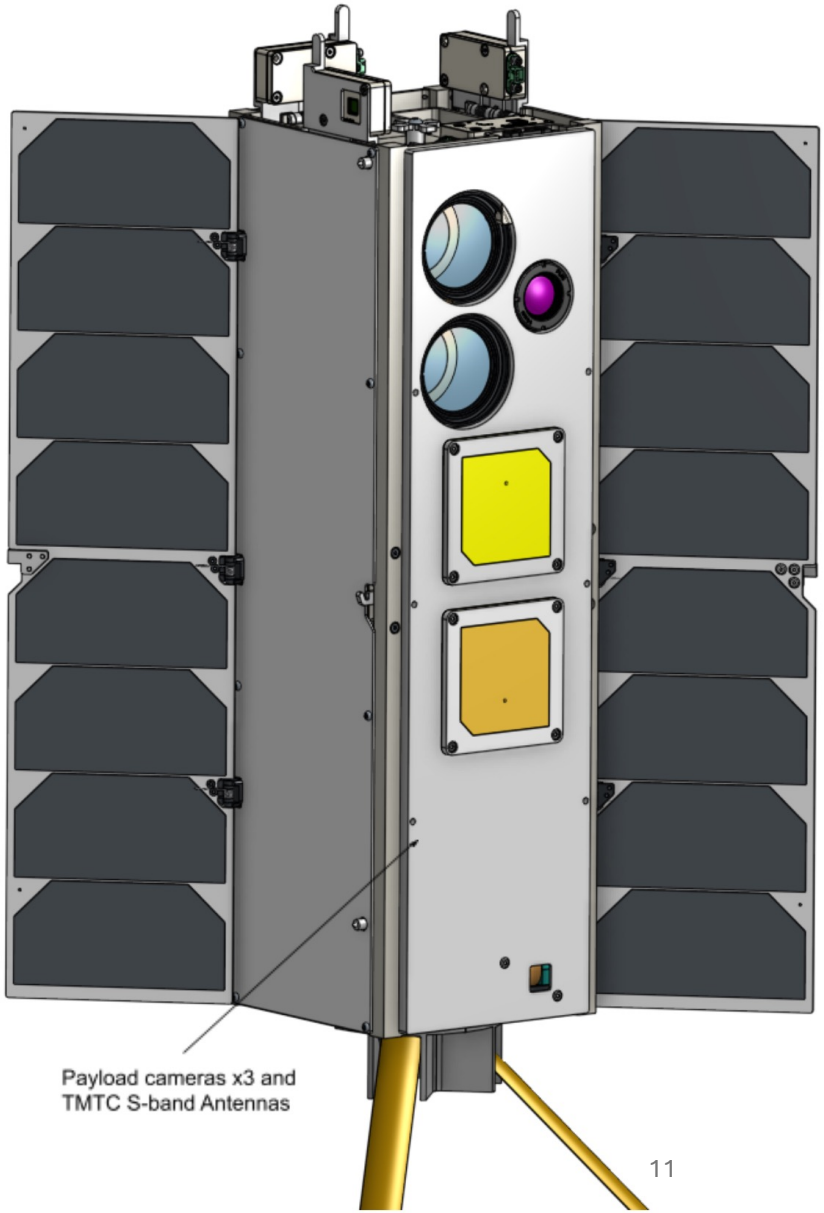
Successor of DISCO 1 with 3 cameras

MLOps become problem on constrained network

Saving bandwidth by discarding data

Mission objectives and models change frequently

Smaller models often required



Robustness

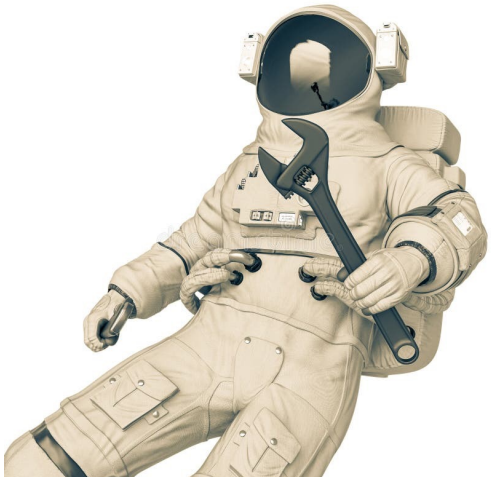
Astronaut technicians are expensive

Particles flying around space with energy equivalent to baseball travelling 100 mph

Bitflips

Corrupted file systems

Damaged subsystems



Age of AI computers / phones

What happens when we move ML to the Edge?

Latency and power - the two most important metrics

Devices will be expected to run more tasks concurrently

- How do we analyze performance?
- How do we assign resources to the tasks and prioritize them?
- How do we collocate them?

Analyzing performance

What benchmarks do we have at disposal?

Analyzing performance

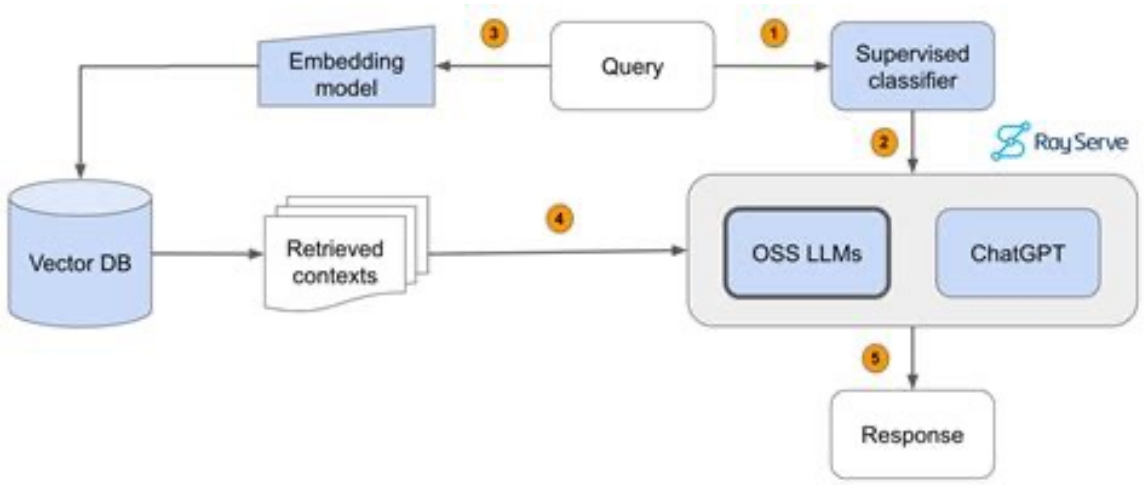
What we would like to see

Benchmark that covers different scales

Measurement of E2E performance

Possibility for collocation

Modularity, flexibility, reuse



https://github.com/ray-project/llm-applications

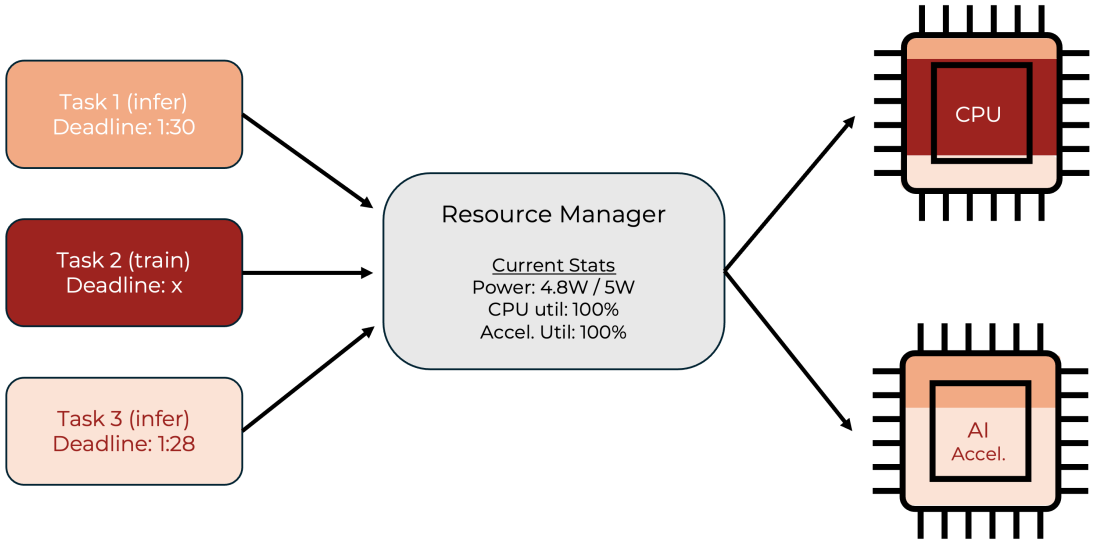
Resource Management

Especially on battery-powered devices

Dynamically assign, adjust and reallocate resources

Inference is latency-critical

Training is resource-hungry



Conclusion

- *EdgeML promises a lot, but challenging in practice*
- *ASICs at the edge necessary*
- *ML is moving to end-user devices FAST*
- *Current benchmarks do not cover complex pipelines, collocation or mixed-workloads*
- *Strict latency and power requirements will require careful resource management*

Thank you!

Do you have interesting use cases for our benchmark?

Get in touch!
roba@itu.dk

robertbayer.github.io
rad.itu.dk