DNN Partitioning

APROPOS Summer School

Delft

Axel Jantsch

June 28, 2023

Institut für Computertechnik
Institute of Computer Technology

Embedded Nodes have Resource Limitations

Resource limitations

	Embedded	Data center
Computation [flop]	$30 - 1800 \cdot 10^{12}$	$86 \cdot 10^{18}$
Memory [bit]	10^{10}	10^{15}
Power [W]	5-100	$10^3 - 10^6$
Energy [Wh]	48-1000	$200 \cdot 10^6$

Computation Embedded refers to an Nvidia Jetson Nano running 1 min and 1 hour, respectively.
Computation server refers to the computation needed for the 40 day experiment with AlphaGo Zero
Energy embedded refers to a mobile phone and to a car battery, respectively.
Energy server refers to the 40 day experiment for AlphaGo Zero.

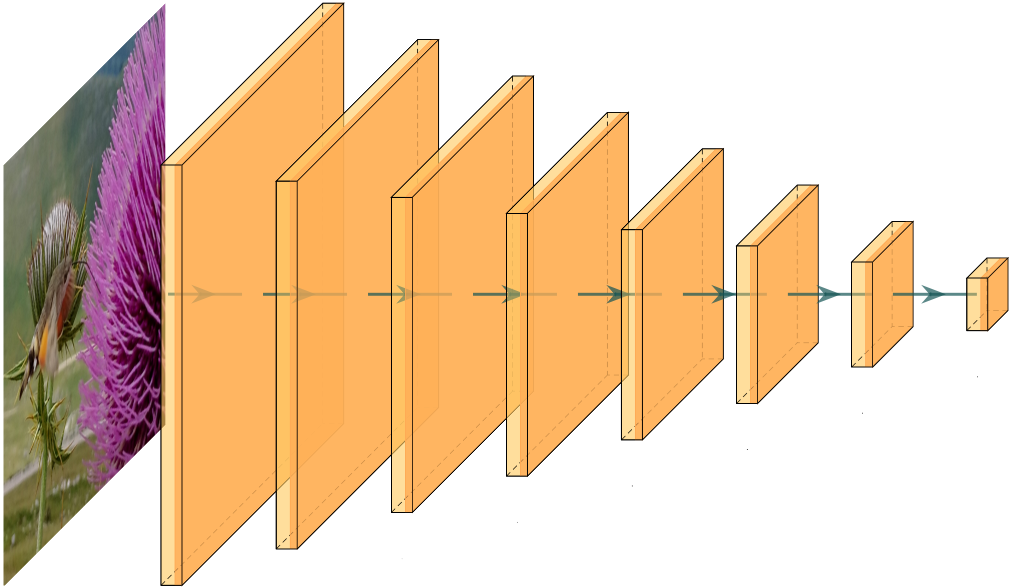
Outline

Outline

1. Partitioning for Inference
2. Impact of image size and content
3. Waist Tightening

Partitioning for Inference

Image Processing Pipelines



Partitioning of the Inference Task

- Energy depends on
  - computation platform
  - amount of computation done
  - communication protocol
  - amount of communication done
- Communication energy cost is very different for different protocols
- Communication energy (and latency tends) to dominate total energy and latency

Irida Shallari, Isaac Sánchez Leal, Silvia Krug, Axel Jantsch, and Mattias O'Nils. "Design space exploration on IoT node: Trade-offs in processing and communication". In: *IEEE Access* (2021)

Energy and Latency Model

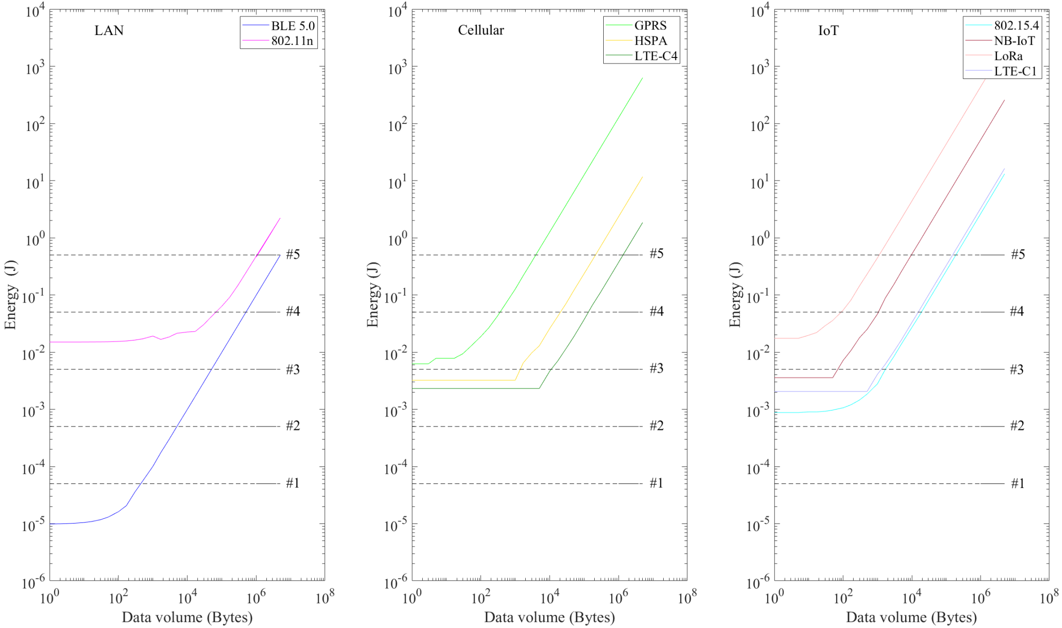
$$
\begin{aligned}
E_{\text{Node}}(x) &= E_S + E_P(T_{(0,x)}, P) + E_c(V_x, C) \\
L_{\text{Node}}(x) &= L_s + L_P(T_{(0,x)}, P) + L_c(V_x, C)
\end{aligned}
$$

L_{Node}	Node latency per sample	E_{Node}	Node energy per sample
L_S	Sensing latency	E_S	Sensing energy
L_P	Processing latency	E_P	Processing energy
L_C	Communication latency	E_C	Communication energy

x	Partitioning point in $[0, \ldots, N]$
$T_{(0,x)}$	Computation tasks of stages $0 \ldots x$
P	Hardware platform
V_x	Data volume at output of stage x
C	Communication protocol

Communication Protocols

Communication groups		
LAN	Cellular	IoT
BLE 5.0	GPRS	802.15.4 g
802.11	HSPA	NB-IoT
	LTE C. 4	LoRa
		LTE C. 1

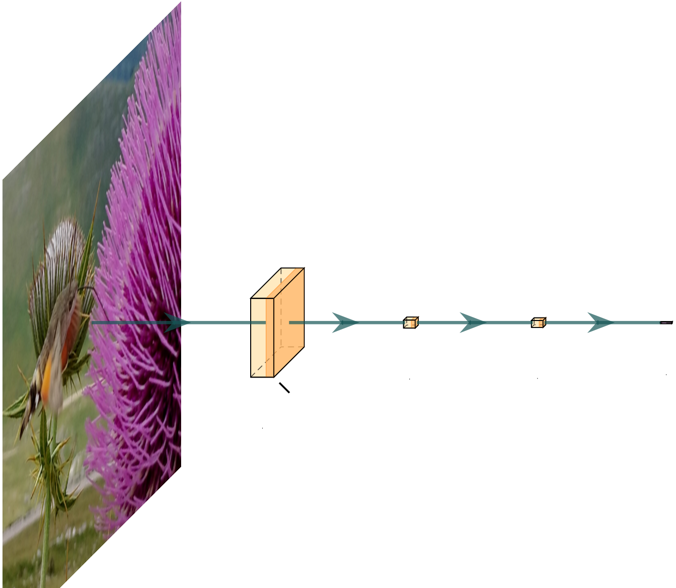


Communication energy for different protocols.

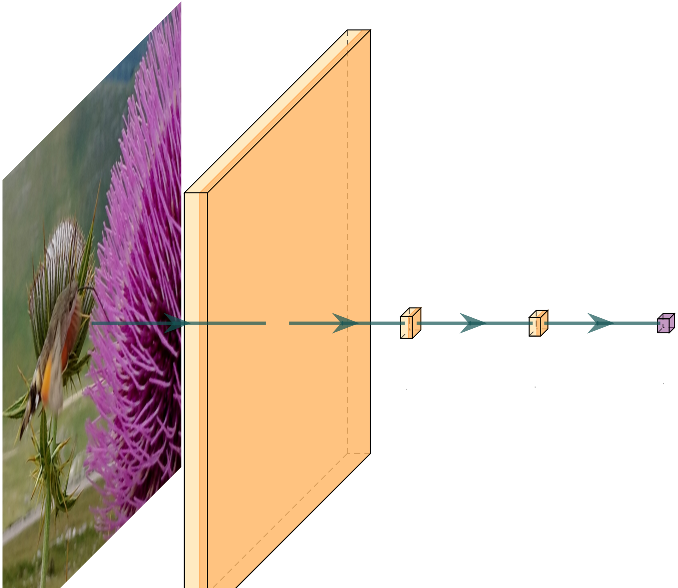
Tasks	Traditional systems		CNN systems	
	People Counting	Particle Detection	AlexNet	VGG16
0	307 200	307 200	307 200	307 200
1	8940	256 000	154 587	150 528
2	91	680	69 984	3 211 264
3	75	500	43 264	1 605 632
4	4	259	64 896	802 816
5			9216	401 408
6			4096	100 352
7			1000	25 088
8				4096
9				1000

(data volume after each processing stage in bytes)

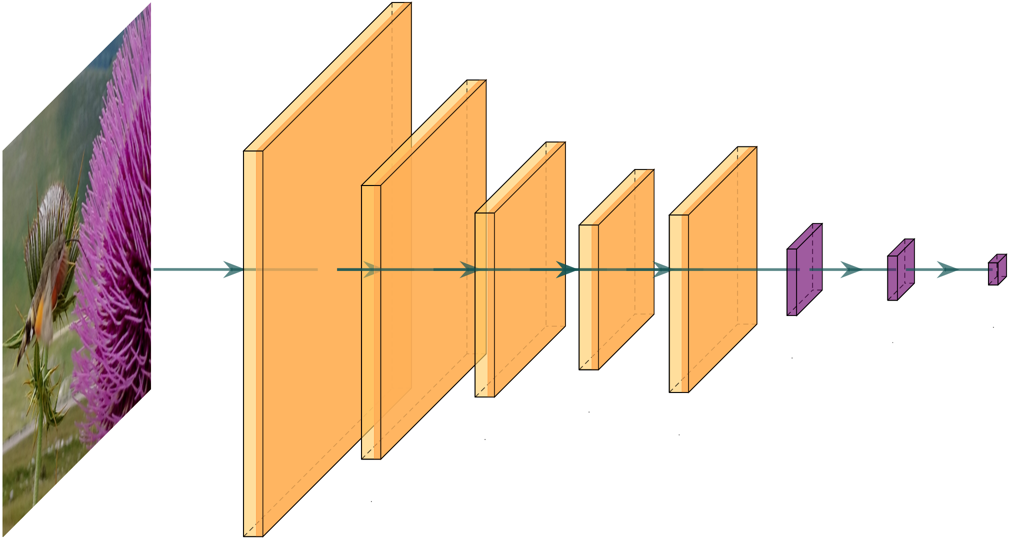
People Counting Pipeline



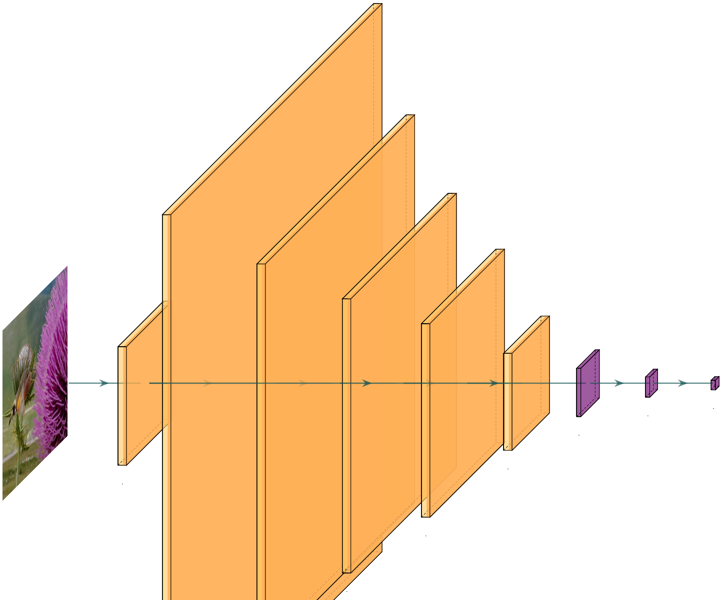
Particle Detection Pipeline

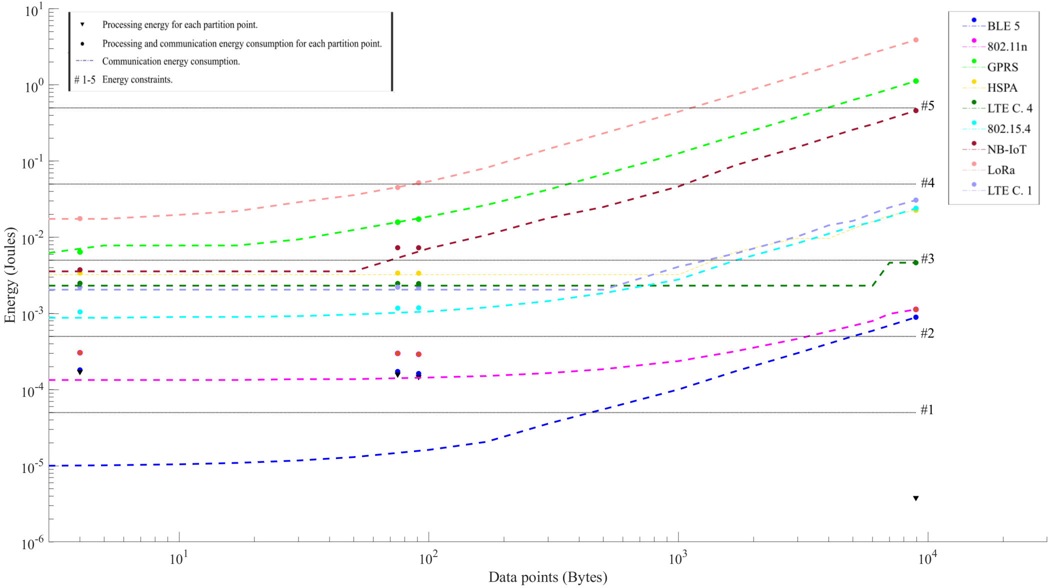


AlexNet Pipeline

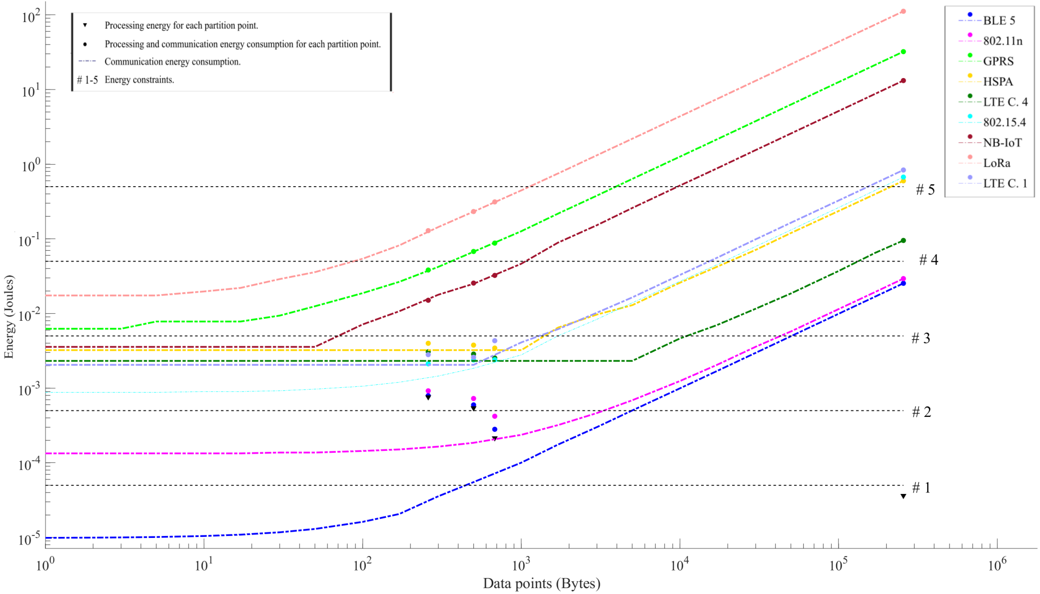


VCG16 Pipeline

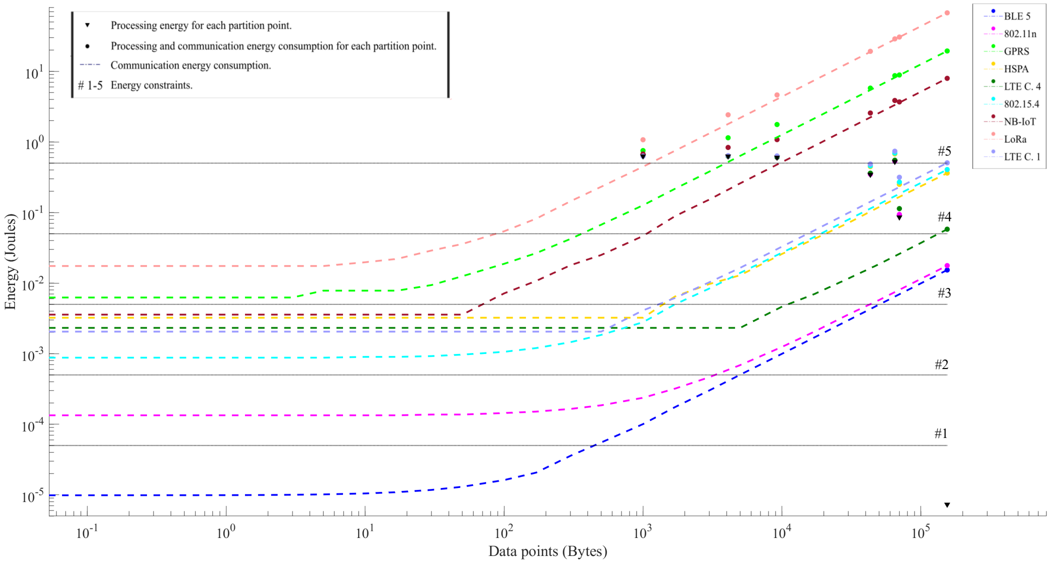




People counting application

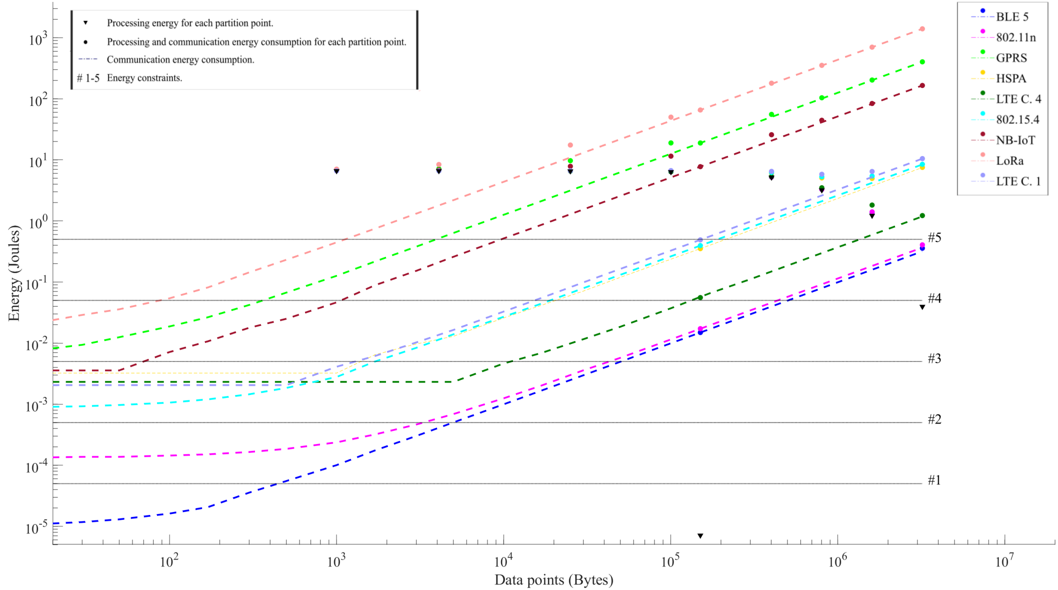


Particle detection application



Processing energy for each partition point.
Processing and communication energy consumption for each partition point.
Communication energy consumption.
1-5 Energy constraints.
BLE 5
802.11n
GPRS
HSPA
LTE C. 4
802.15.4
NB-IoT
LoRa
LTE C. 1
#5
#4
#3
#2
#1
Energy (Joules)
Data points (Bytes)

AlexNet



Processing energy for each partition point.
Processing and communication energy consumption for each partition point.
Communication energy consumption.
1-5 Energy constraints.
BLE 5
802.11n
GPRS
HSPA
LTE C. 4
802.15.4
NB-IoT
LoRa
LTE C. 1
#5
#4
#3
#2
#1
Energy (Joules)
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^2
10^3
10^4
10^5
10^6
10^7
Data points (Bytes)

VGGNet 16

Conclusions

- For high energy communication protocols the it is optimal to minimize transmitted data.
- For low energy communication protocols the sweet spot is not at the extremes.
- The optima depend on the application, the IoT platform and the communication protocol used.

Outline

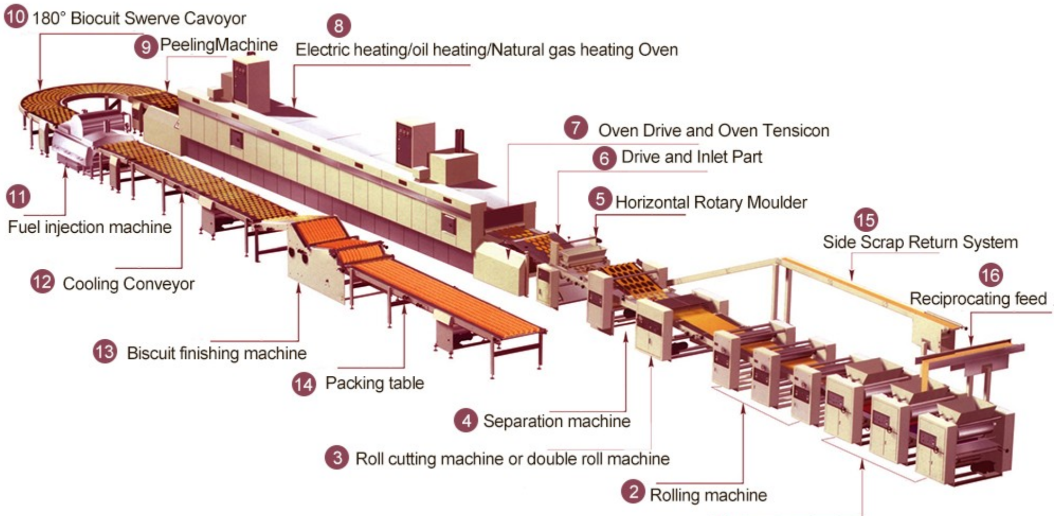
Impact of image size and content

Impact of image size and image content

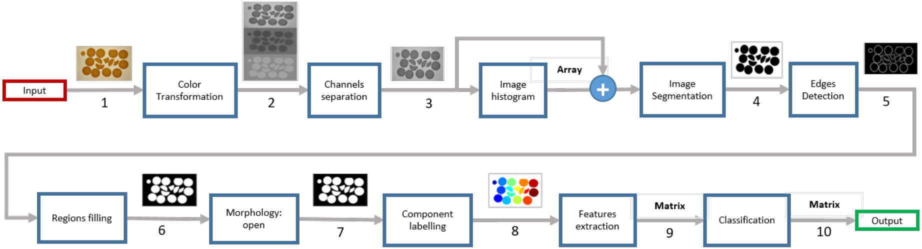
Case study: Conventional image processing pipeline - Biscuit inspection system

Isaac Sánchez Leal, Irida Shallari, Silvia Krug, Axel Jantsch, and Mattias O'Nils. "Impact of Input Data on Intelligence Partitioning Decisions for IoT Smart Camera Nodes". In: *Electronics* 10.16 (2021)

Biscuit Production



#	Processing Task	Output type	Changes in data out due to: $\Delta Img.Size$	$\Delta Img.Objects$	Processing time behavior
t1	Color transformation.	Color space YCbCr.	$\Delta Linear$	Constant	$\Delta Linear_{Size}$
t2	Channels separation.	Y channel.	$\Delta Linear$	Constant	$\Delta Linear_{Size}$
t3	Image Histogram.	Array with 256 elements.	Constant	Constant	$\Delta Linear_{Size}$
t4	Segmentation.	BW image without background.	$\Delta Linear$	Constant	$\Delta Linear_{Size}$
t5	Edges detection.	BW image with detected regions.	$\Delta Linear$	Constant	$\Delta Linear_{Size}$
t6	Regions filling.	BW image with filled regions.	$\Delta Linear$	Constant	$\Delta Linear_{Size}$
t7	Morphology: open.	BW image without particles.	$\Delta Linear$	Constant	$\Delta Linear_{Size}$
t8	Component Labelling.	Non-binary labels image.	$\Delta Linear$	Constant	$\Delta Linear_{Size}$
t9	Features extraction.	2D features matrix.	Constant	$\Delta Linear$	$\Delta Linear_{Size,Objects}$
t10	Classification.	2D coordinates matrix.	Constant	$\Delta Linear$	$\Delta Linear_{Objects}$



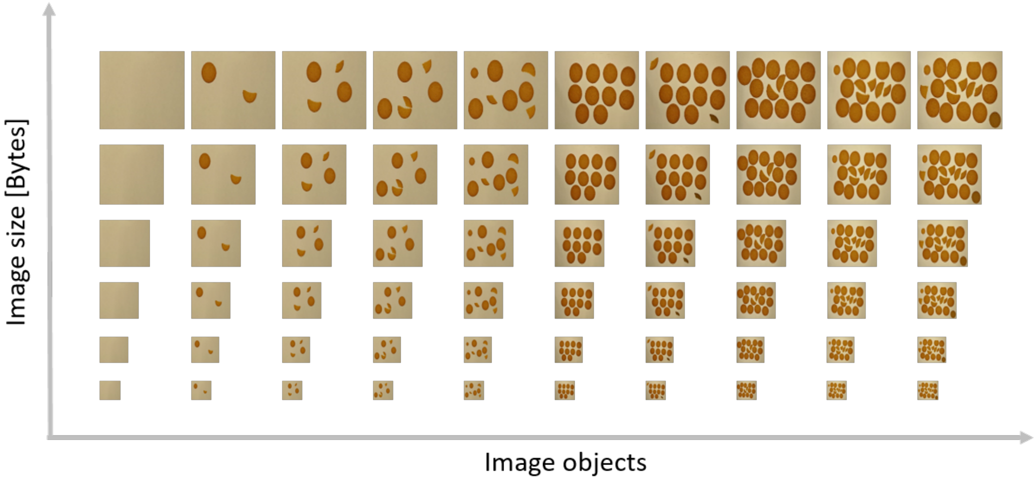
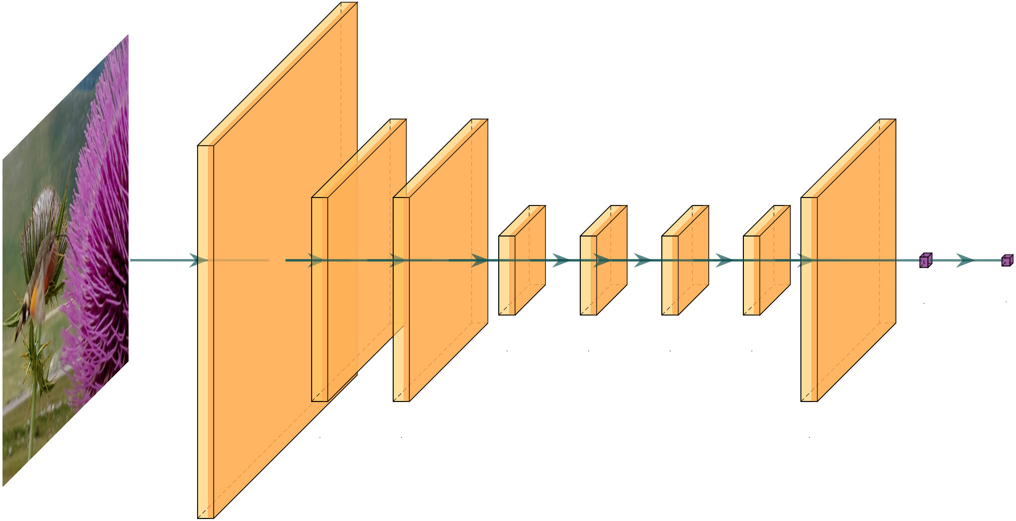


Image size [Bytes]
Image objects

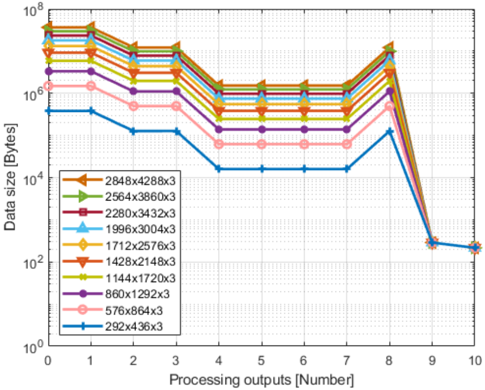
Main points

- Most tasks' processing time depends on the size of image: t1,t2,t4-t8
- Some tasks depend on number of objects: t9, t10, which are later in the pipeline
- Data Volume depends heavily on image size, only late in the pipeline (t9,t10) also on number of objects

Biscuit Image Processing Pipeline



Data Volume



18 objects

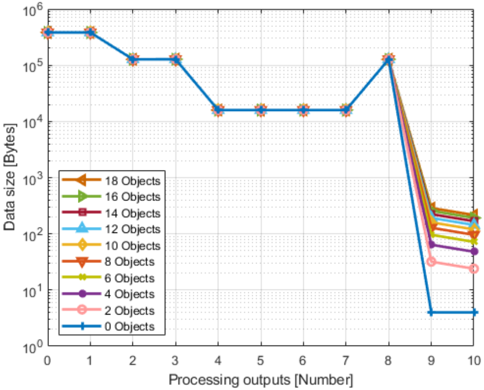
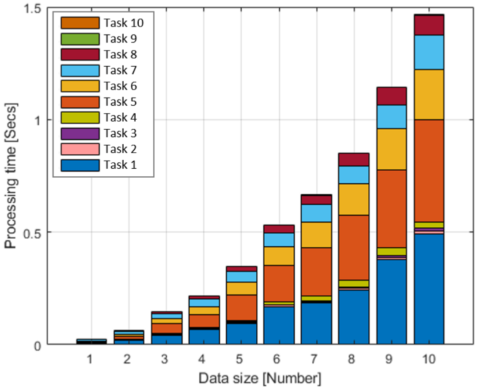
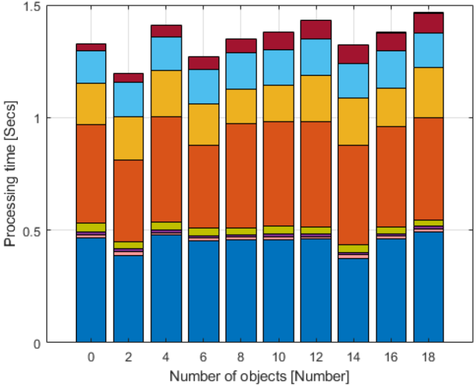


Image size: $292 \times 436 \times 3$.

Processing Time



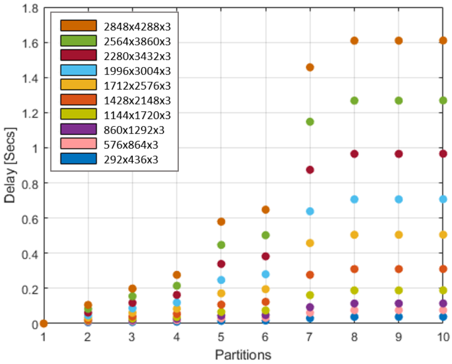


18 objects

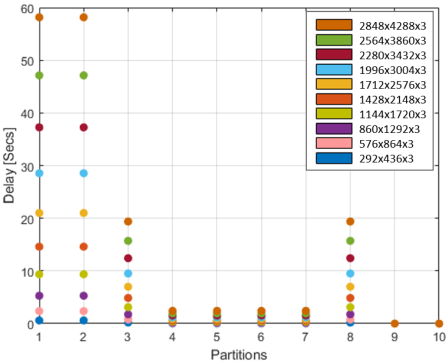
Image size: $2848 \times 4288 \times 3$

Latency

Proccesing latency with RasberryPi.



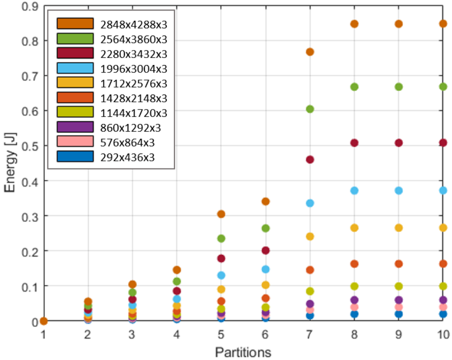
Communication latency with LTE Cat.1.



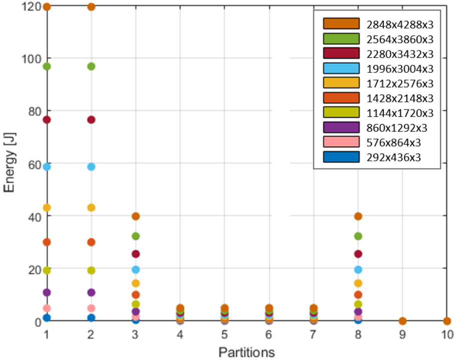
18 objects

Energy

Proccesing energy on RasberryPi.



Communication energy with LTE Cat.1.



18 objects

Optimization Problems

1. Miminizing latency:

$$L_{\text{Node}} \to \mathit{Min}$$

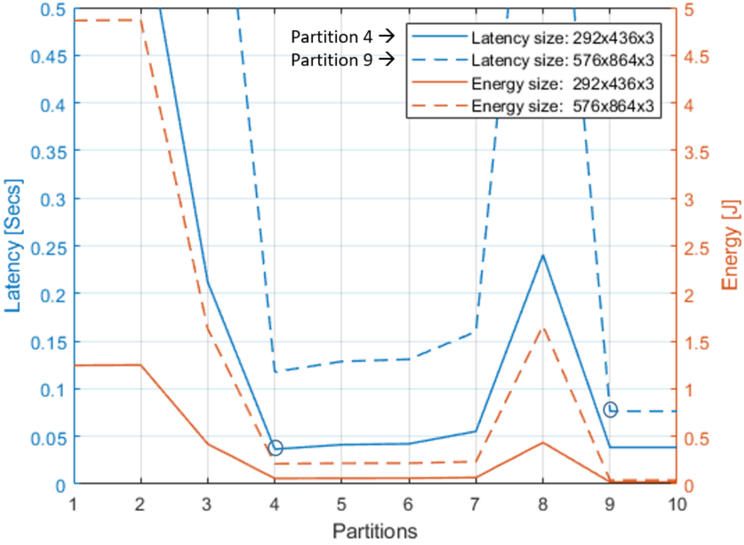
2. Minimizing energy:

$$E_{\text{Node}} \to \mathit{Min}$$

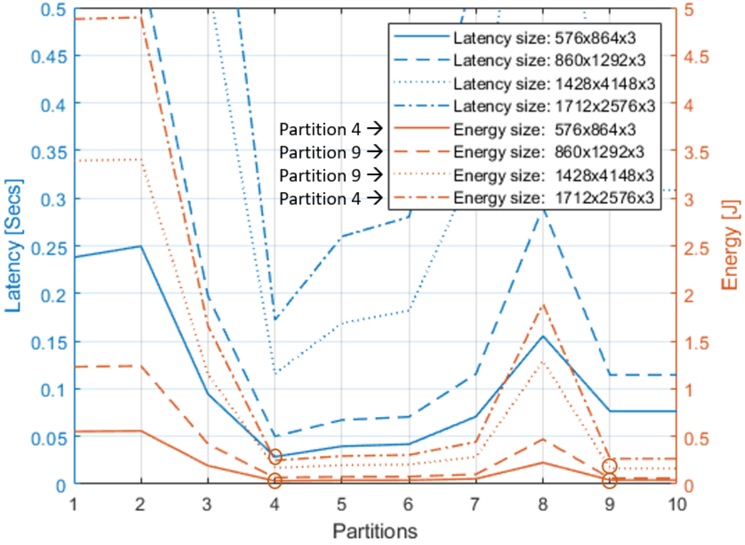
3. Minimizing energy under a latency constraint:

$$L_{\text{Node}} \leq \mathit{MaxDelayConstraint}$$
$$E_{\text{Node}} \to \mathit{Min}$$

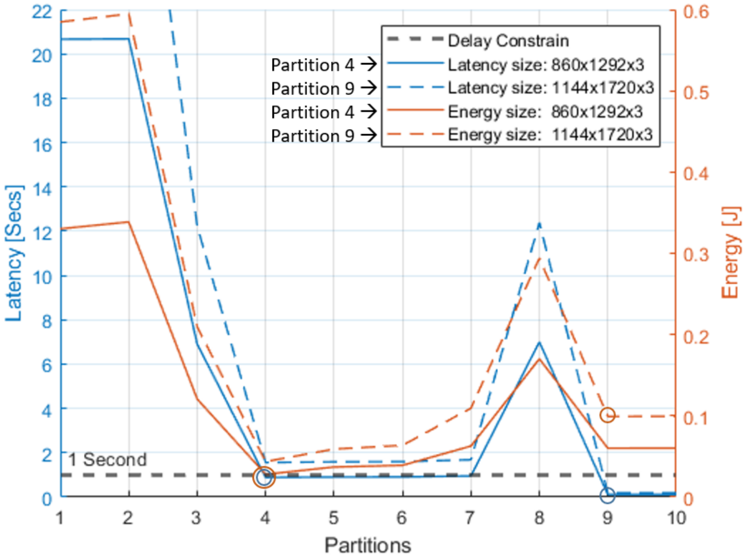
Latency minimization for LTE Cat.1



Energy minimization for LTE Cat.4



Energy minimization under a delay constraint for BLE5



Outline

Waist Tightening

Methodology

Assumptions

- Energy and delay model
- Optimization goal
- Constraints on energy, latency and accuracy

Selecting the partitioning point

- Candidate points have low data volume
- For each candidate point
  - Quantization,Pruning and compression
  - Retraining
- Best point according to constraints and optimization criteria is selected.

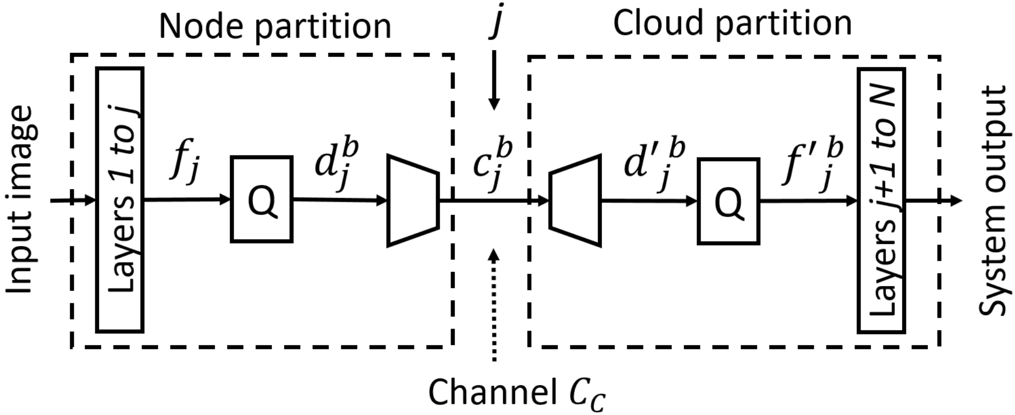
Energy and Latency Model

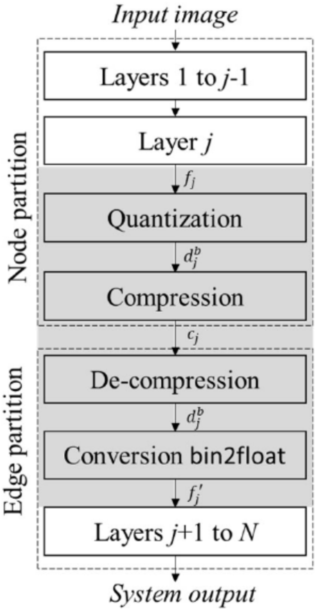
$$
\begin{aligned}
E_{\text{Node}}(x) &= E_S + E_P(T_{(0,x)}, P) + E_c(V_x, C) \\
L_{\text{Node}}(x) &= L_s + L_P(T_{(0,x)}, P) + L_c(V_x, C)
\end{aligned}
$$

L_{Node}	Node latency per sample	E_{Node}	Node energy per sample
L_S	Sensing latency	E_S	Sensing energy
L_P	Processing latency	E_P	Processing energy
L_C	Communication latency	E_C	Communication energy

x	Partitioning point in $[0, \ldots, N]$
$T_{(0,x)}$	Computation tasks of stages $0 \ldots x$
P	Hardware platform
V_x	Data volume at output of stage x
C	Communication protocol

Partitioned DNNs





DNN Partitioning:

$$M_p = M_{1..j} \cup I \cup M_{j+1..N}$$

Optimization problem

Minimizing energy under accuracy constraint:

$$\begin{aligned} E_{\text{Node}} &\rightarrow Min \\ mAP(M_p) &\geq mAP(M) - L_{\text{th}}. \end{aligned}$$

M	Original DNN
M_p	DNN partitioned at point j
I	Interface
mAP	mean average precision
L_{th}	Threshold for acceptable loss of precision

Eiraj Saqib, Isaac Sánchez Leal, Irida Shallari, Axel Jantsch, Silvia Krug, and Mattias O'Nils. "Optimizing the IoT Performance: A Case Study on Pruning a Distributed CNN". In: *Proceedings of the IEEE Sensors Applications Symposium (SAS)*. 2023

Quantization

Two quantization methods:

$$
\begin{aligned}
Q_1: \quad d_j^b &= \left\lfloor \frac{f_j - S}{M - S} \times 2^{b-1} \right\rfloor, && \text{with } b \in [1-8]. \\
f_j' &= d_j^b \times \left(\frac{M-S}{2^{b-1}}\right) + S, && \text{where } b \in [1-8].
\end{aligned}
$$

$$
\begin{aligned}
Q_2: \quad d_j^b &= \lfloor (f_j - \mu) \,/\, M \rfloor && \text{with } b = 1. \\
f_j' &= \left(d_j^b \times M\right) + \mu && \text{with } b = 1.
\end{aligned}
$$

f_j	Floating point number in feature map at output of stage j
d_j^b	Quantized number to b bit
f_j'	De-quantized value
M	Maximum value of the dynamic range
m	Minimum value of the dynamic range
S	$(M - m)/2$
μ	Mean value

Pruning

Two step pruning:

1. Pruning of filters in Layer j, subject to

$$mAP(M'_p) \geq mAP(M) - L_{\text{th}}$$

2. Pruning of layers $i = 1, \ldots, j-1$, subject to

$$mAP(M''_p) \geq mAP(M) - L_{\text{th}}$$

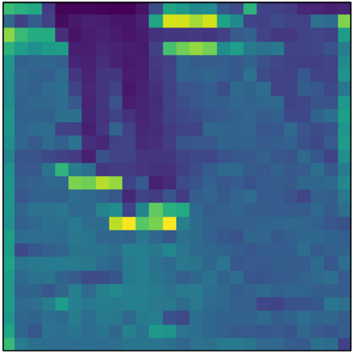
M' DNN with layer j pruned
M'' DNN with layers $i = 1, \ldots, j$ pruned

Compression

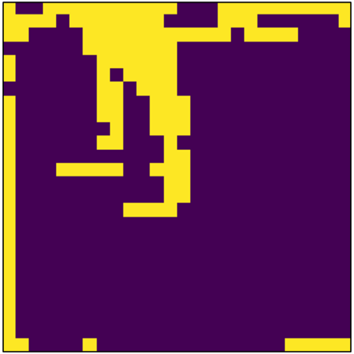
Compression to minimize the data transmitted over the channel
We use the zip and JPEG compression algorithms.

Wheel Chair Steering Case Study

Input image



32-bits

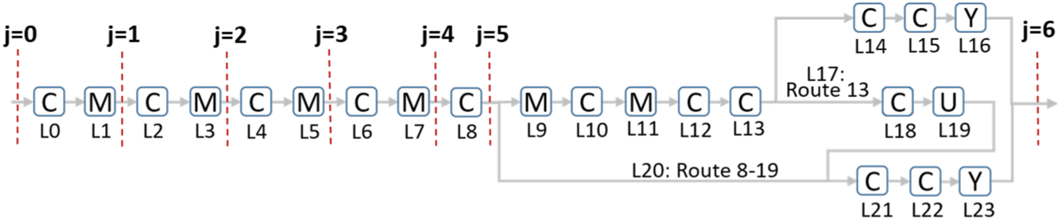


1-bit

Cristian Vilar Giménez, Silvia Krug, Faisal Z. Qureshi, and Mattias O'Nils. "Evaluation of 2D-/3D-Feet-Detection Methods for Semi-Autonomous Powered Wheelchair Navigation". In: *Journal of Imaging* 7.12 (2021)

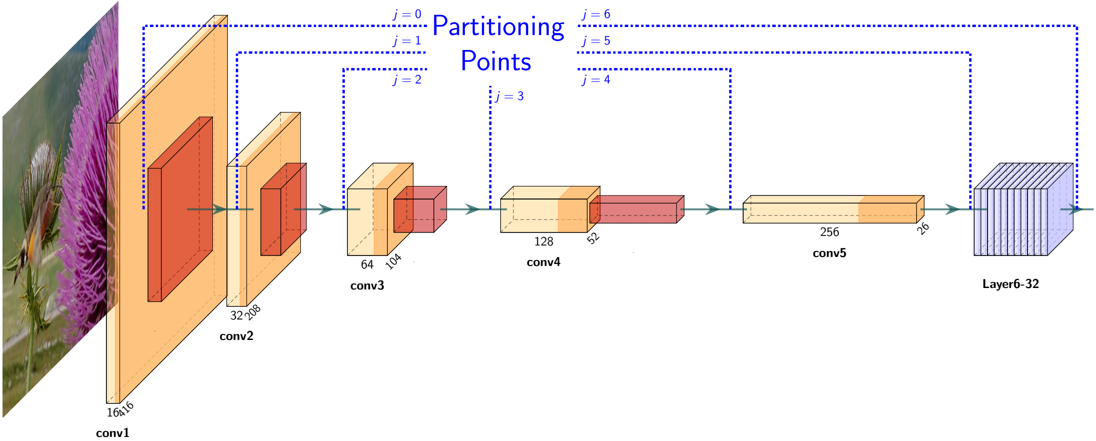
Isaac Sanchez Leal, Eiraj Saqib, Irida Shallari, Axel Jantsch, Silvia Krug, and Mattias O'Nils. "Waist Tightening of CNNs: A Case study on Tiny YOLOv3 for Distributed IoT Implementations". In: *Proceedings of the Real-time And intelliGent Edge computing workshop (RAGE)*. San Antonio, Texas, May 2023

TinyYolo V3

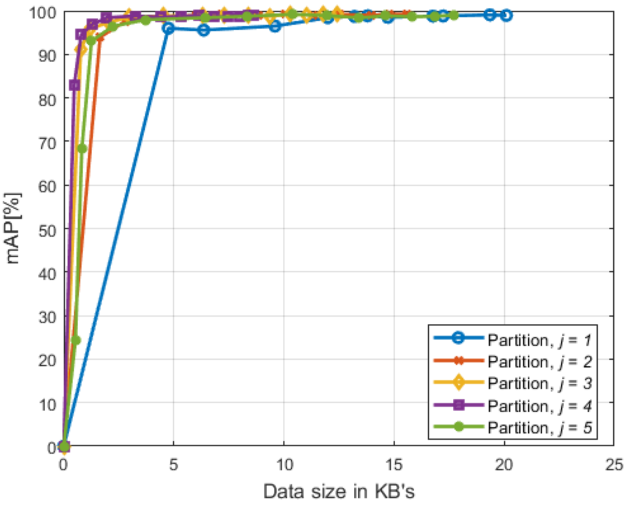


Li: Layer "i"	C: Convolutional	U: Up-sampling
j: Partition	M: Maxpool	Y: YOLO region

TinyYolo V3

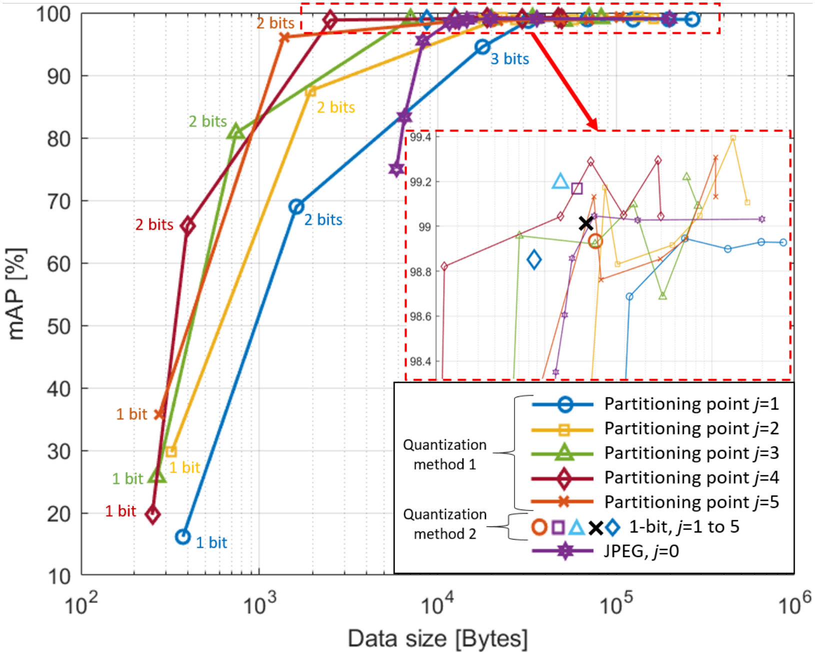

$j = 0$
$j = 1$
$j = 2$
Partitioning Points
$j = 3$
$j = 6$
$j = 5$
$j = 4$
16
416
conv1
32
208
conv2
64
104
conv3
128
52
conv4
256
26
conv5
Layer6-32

Data Volume

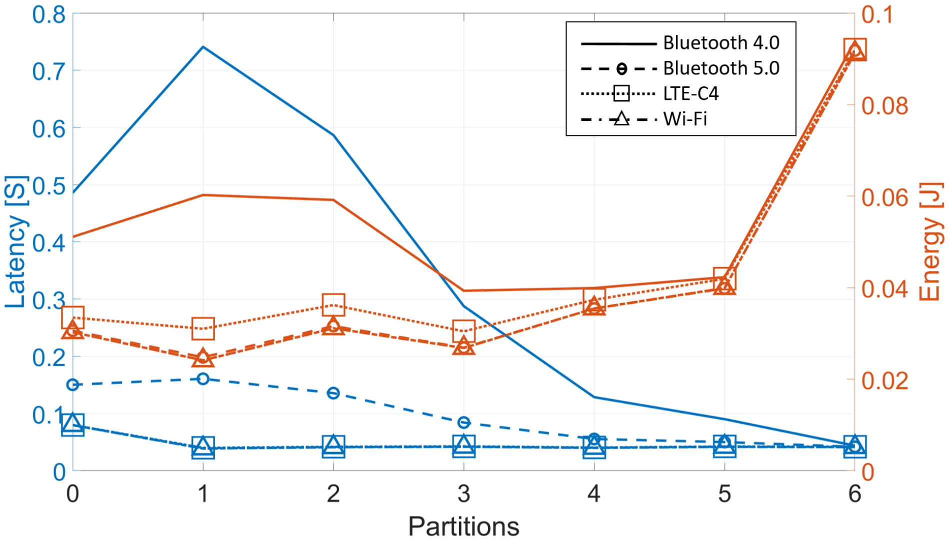


Data volume at partition point after quantization, pruning and compression

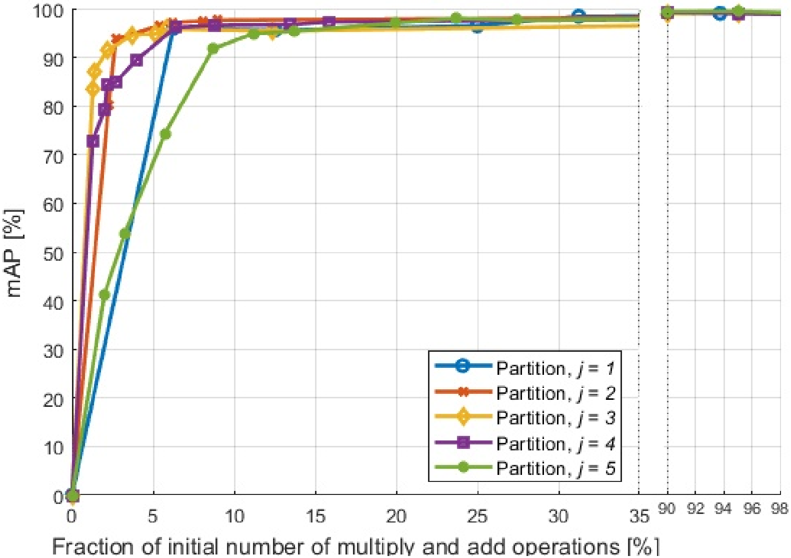
Quantization and Compression



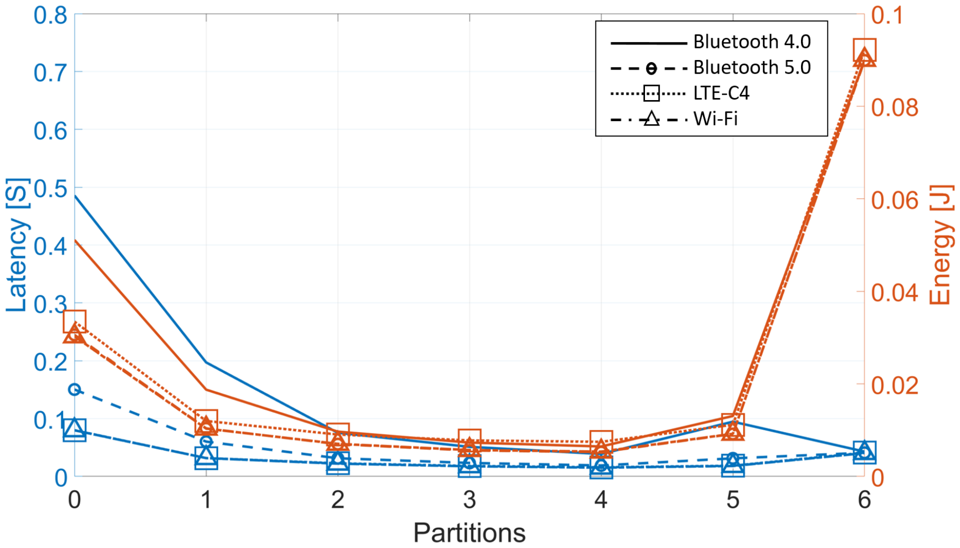
Quantization Q_2 and Compression



Quantization Q_2, Pruning and Compression



Quantization Q_2, Pruning and Compression



Partitioning and Pruning Benefits

	Quantization and Compression		Quantization, Compression and Pruning	
	All In-Edge	**All In-Node**	**All In-Edge**	**All In-Node**
Energy saving	x1.26	x3.8	x5.74	x17
System speed-up	x2.05	x1.05	x5.24	x2.65

Optimal partitioning versus all-in-edge and all-in-node reference solutions.

DNN Partitioning - Summary

Summary

- Effective DNN partitioning is feasible

DNN Partitioning - Summary

Summary

- Effective DNN partitioning is feasible
- DNN partitioning opens a considerable design space for DNN based IoT applications

DNN Partitioning - Summary

Summary

- Effective DNN partitioning is feasible
- DNN partitioning opens a considerable design space for DNN based IoT applications
- Next steps is to explore more aggressive DNN adaptions for partitioning

References I

[1] Irida Shallari, Isaac Sánchez Leal, Silvia Krug, Axel Jantsch, and Mattias O'Nils. "Design space exploration on IoT node: Trade-offs in processing and communication". In: *IEEE Access* (2021).

[2] Isaac Sánchez Leal, Irida Shallari, Silvia Krug, Axel Jantsch, and Mattias O'Nils. "Impact of Input Data on Intelligence Partitioning Decisions for IoT Smart Camera Nodes". In: *Electronics* 10.16 (2021).

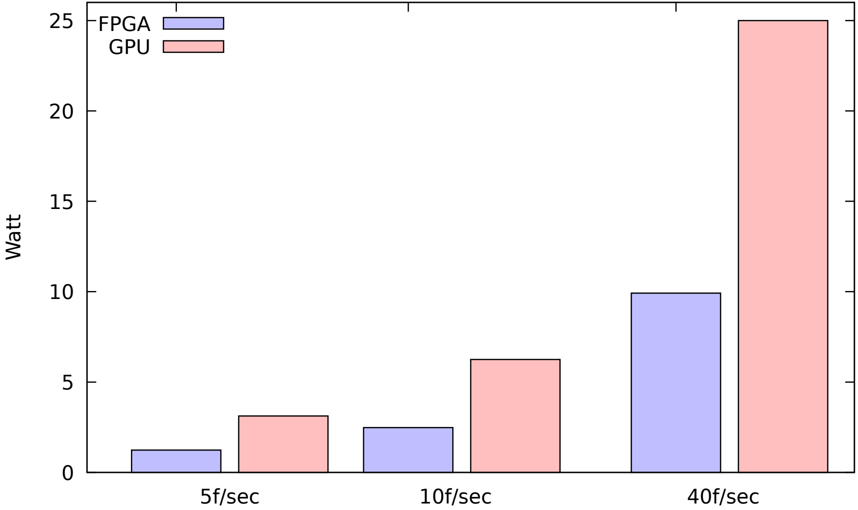
[3] Isaac Sanchez Leal, Eiraj Saqib, Irida Shallari, Axel Jantsch, Silvia Krug, and Mattias O'Nils. "Waist Tightening of CNNs: A Case study on Tiny YOLOv3 for Distributed IoT Implementations". In: *Proceedings of the Real-time And intelliGent Edge computing workshop (RAGE)*. San Antonio, Texas, May 2023.

[4] Cristian Vilar Giménez, Silvia Krug, Faisal Z. Qureshi, and Mattias O'Nils. "Evaluation of 2D-/3D-Feet-Detection Methods for Semi-Autonomous Powered Wheelchair Navigation". In: *Journal of Imaging* 7.12 (2021).

[5] Eiraj Saqib, Isaac Sánchez Leal, Irida Shallari, Axel Jantsch, Silvia Krug, and Mattias O'Nils. "Optimizing the IoT Performance: A Case Study on Pruning a Distributed CNN". In: *Proceedings of the IEEE Sensors Applications Symposium (SAS)*. 2023.

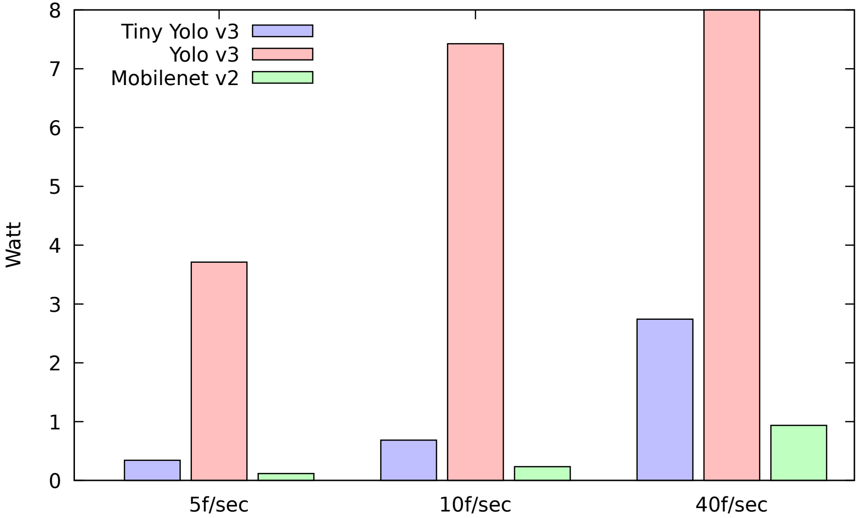

TECHNISCHE HOCHSCHULE

Power Consumption in Inference



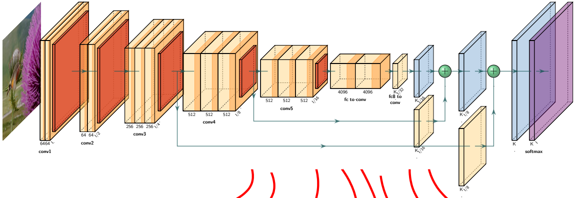
VGG16 applied to the ImageNet data set based on published papers.

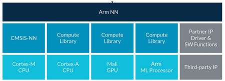
Power Consumption in Inference



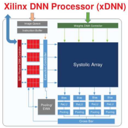
Object detection on the NCS2 platform; own measurements.

Design Space





ARM NN



Nvidia Turing

DNN Choices

Convolutional layers
Filter kernels
Number of filters
Pooling layers
Filter shape
Stride
Fully connected layer
Number of layers
Regularization
etc.

Mapping Choices

Neuron pruning
Data type selection
Approximation
Retraining
Connection pruning
Weight sparsifying
Regularization
etc.

Platform Choices

Platform Selection
Reconfiguration
Batch processing
Deep pipelining
Resource reuse
Hierarchical control
Processing unit selection
Memory allocation
Memory reuse
etc.


TECHNISCHE HOCHSCHULE