EDGE AI Academy

Summer School

7-8 July 2025, Pisa, Italy

Edge AI Academy Summer School

Exploring hardware-efficient architectures for reducing data transfer in DNNs

Sebastian Prebeck, Mounika Vaddeboina,
Infineon Technologies AG, Germany

7-8 July 2025 Pisa, Italy

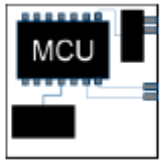
Outline

- Motivation
- Uncompressed NNs
  - Memory mapping
  - Architectural limitations
- Compressed NNs
  - Golomb-Rice(GR) algorithm
  - GR-encoding for NN compression
  - Decoder architectures(Parallel Processing)

Copyright © Infineon Technologies AG 2025. All rights reserved.

ML Platforms are Heterogeneous

Large computing continuum with possibly connectivity:



Datacenter:
Multi-Servers with
Multi-GPUs

Desktop/Workstation /Fog:
PC with GPU

Edge/Mobile:
Mobile Phone
Raspberry PI
Embedded GPU
Specialized SoCs

Extreme Edge / TinyML:
MCU
Specialized low-power SoC

Hundreds of CPUs
Hundreds of GBs of DRAM
Several GPUs with
Tens of GB of DRAM
Several TB of Storage

2-128 CPUs
Tens of GBs of DRAM
1-2 GPUs with Tens of GB of DRAM
A few TB of Storage

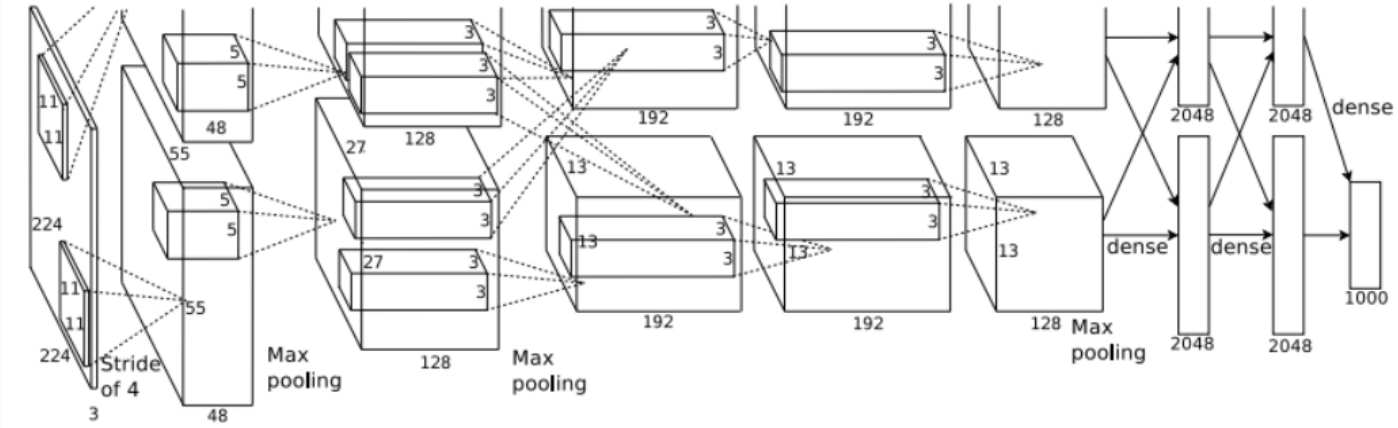
1-4 CPUs
1-4 GBs of DRAM
1 GPUs with a few GB of DRAM
Specialized Accelerators
Tens to Hundreds of GB of Storage

1 CPU
Hundreds of kB to a few MB of embedded SRAM
Low-power Acceleration / Co-proccesors
A few MB of Storage, e.g. embedded Flash

Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

AI networks may need Trillions of MACs



Source: ImageNet Classification with Deep Convolutional Neural Networks, Alex Krizhevsky et.al.

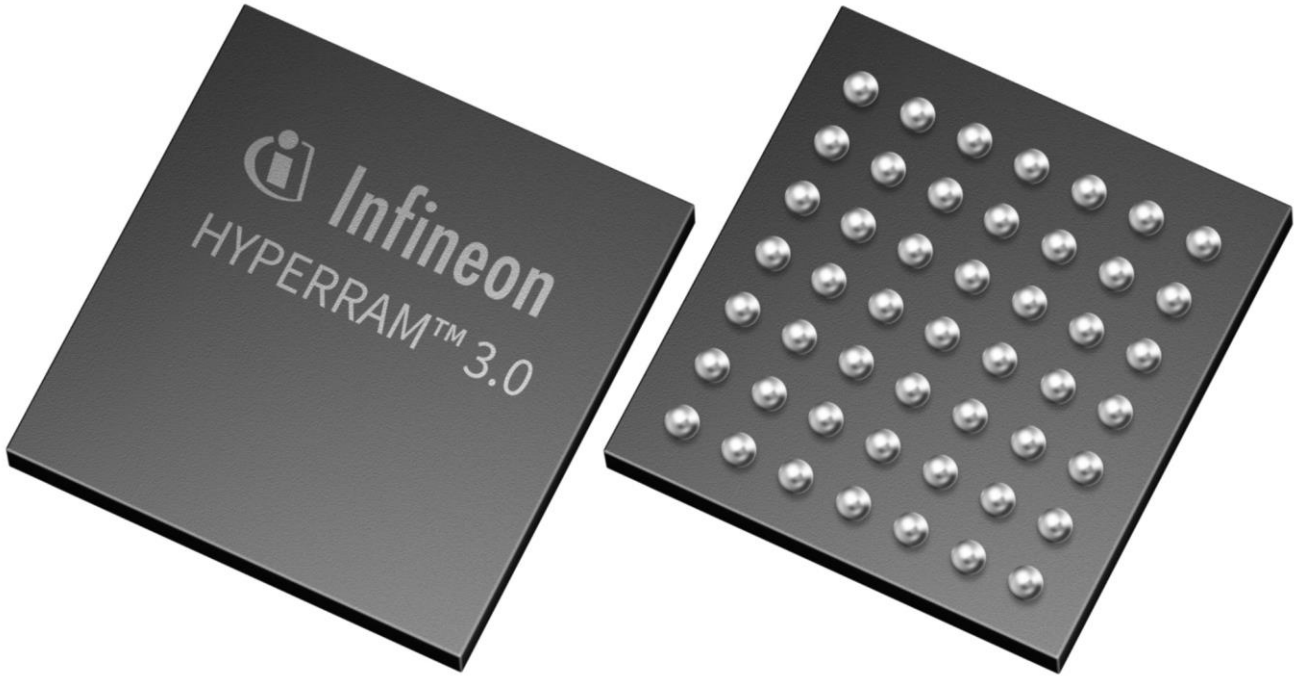
- AlexNet has **63 Million weights**
- **227 x 227 x 3 pixels** Input Image

=> **6.3 Billion MAC ops @ 100Hz**

Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

DRAM access is expensive



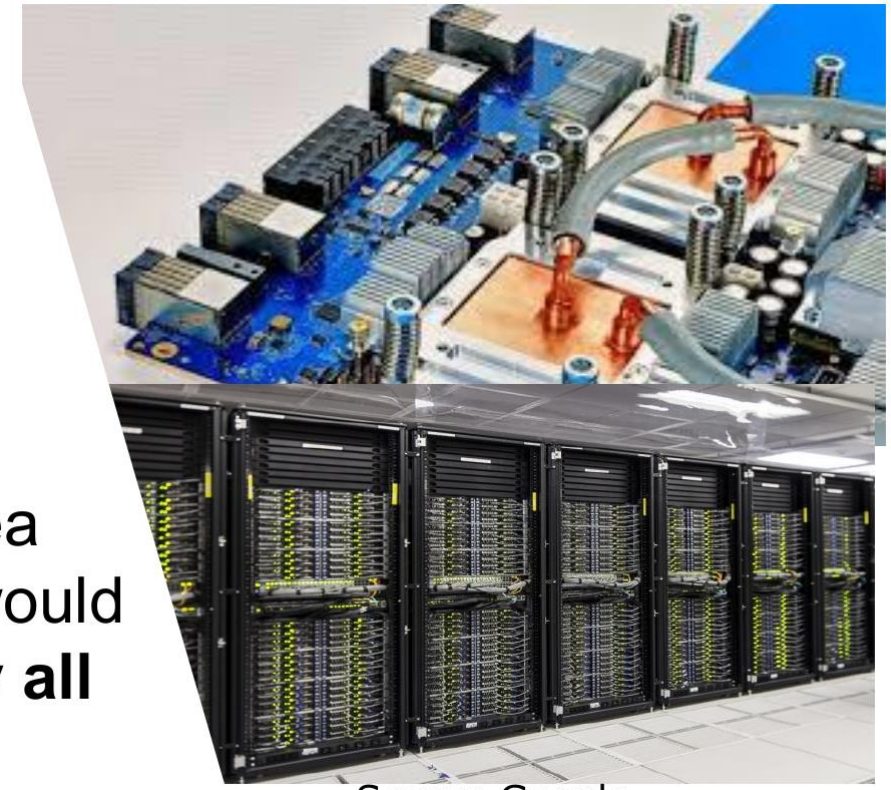
- Processing NNs requires many memory access (weights and activations)
- Internal memory access takes approx. as much Energy as a MAC operation
- External memory access takes ~100x more energy than an internal one

Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

Impact on society: Energy consumption

- Google's TPU is water cooled
- Nvidea's Titan consumes > 250 Watt, Google's Deep Mind 1 Mega-Watt
- All German households permanently running an Nvidea Titan for an intelligent home would require the energy produce by **all lignite power plants** in 2020.

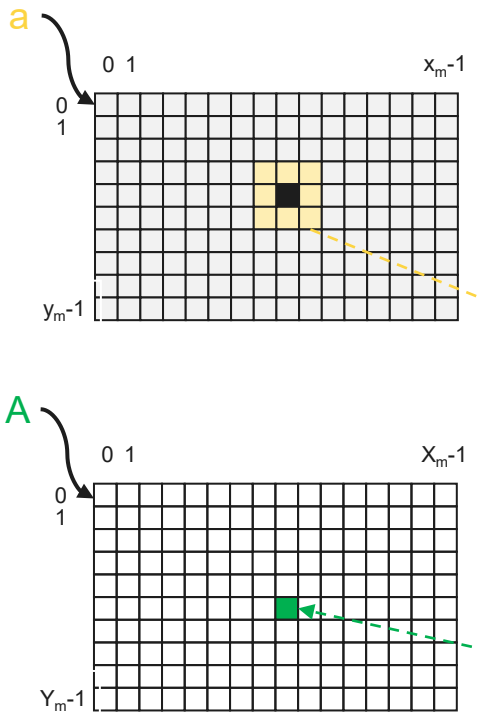


Source Google

Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

Core Belief: Memory as dominant factor (xNN Addressing)




```
// Pseudo-code for illustrating the memory access challenge
int a[x_m*y_m]; // assumption that both spaces have the same dimension
int A[X_m*Y_m];
int W[…];
int i_w=…;
for(i_x=0; i_x<x_m;i_x++) { // walk through 2D space
   for(i_y=0; i_y<y_m;i_y++) {
      s = 0;
      for(k_x=i_x-d_x; k_x<=i_x+d_x;k_x++) { // walk through kernel
        for(k_y=i_y-d_y; k_y<=i_y+d_y;k_y++) {
          if( k_x<0|| k_y<0||k_x>=x_m||k_y>=y_m){
             t=0; // simple handling of out of 2D space
          } else {
             t=a[k_x+x_m*k_y];
          }
          s+=t*W[i_w ++] // MAC and weight access
        }
      }
      s=activationFunction(s,W[i_w++]);
      A[i_x+x_m*i_y]=s;
   }
}
```

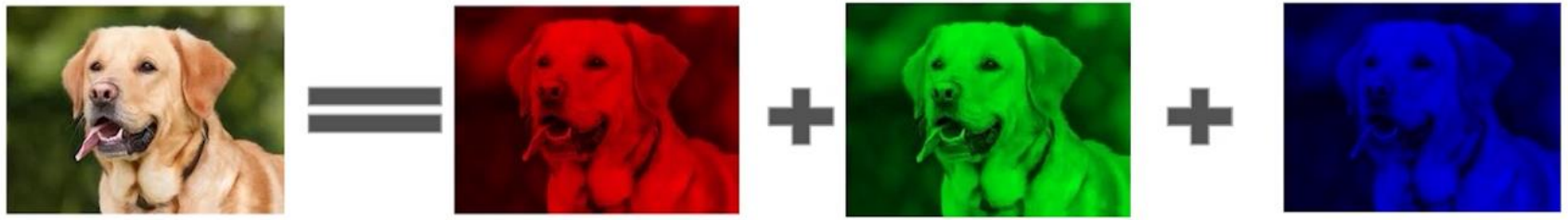
10-100 instructions for xNN addressing

More Instructions needed for
- Compression/decompression
- Up-/Downscaling
- Quantization
- Activation Function

Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

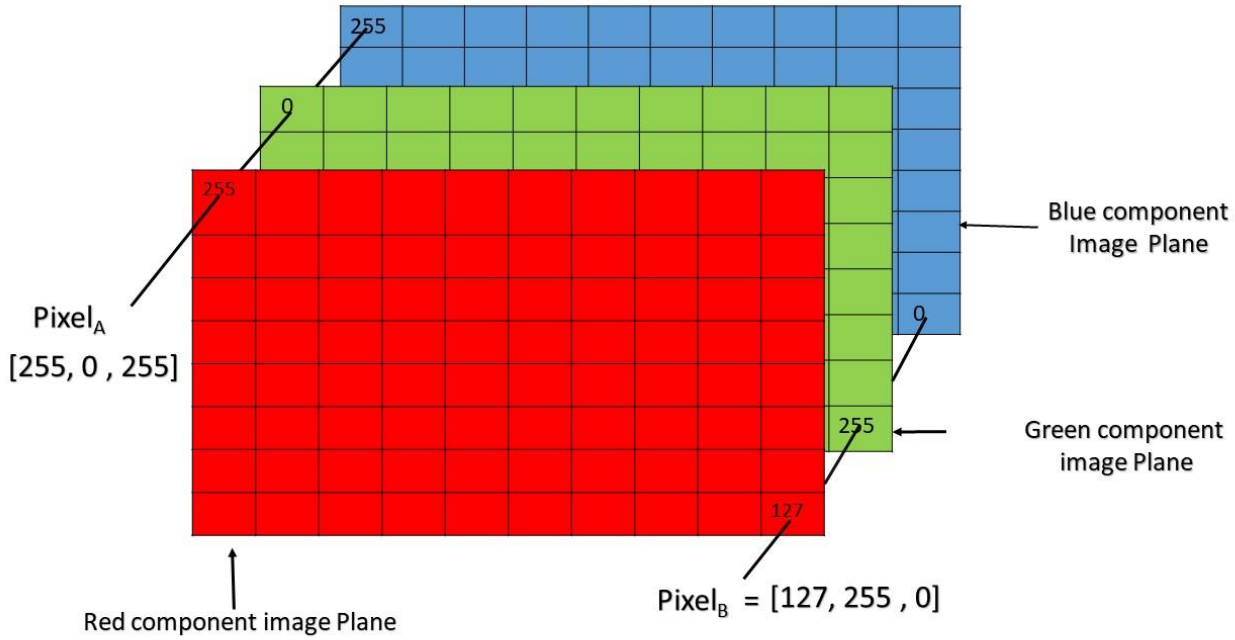
Data representation: Coloured Image



https://www.analyticsvidhya.com/blog/2021/03/grayscale-and-rgb-format-for-storing-images/

Copyright © Infineon Technologies AG 2025. All rights reserved.

Data representation: RGB Matrix



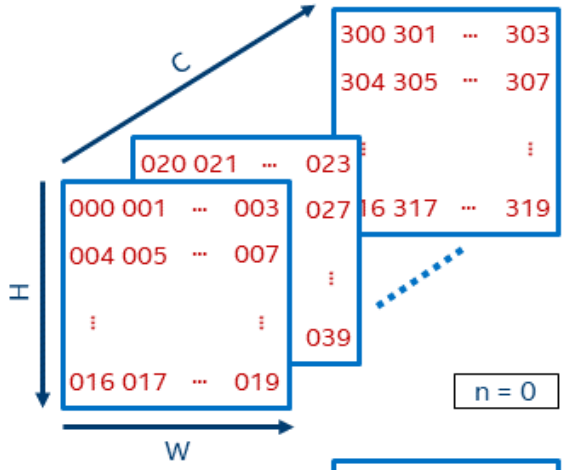
Pixel of an RGB image are formed from the corresponding pixel of the three component images

https://www.geeksforgeeks.org/matlab-rgb-image-representation/

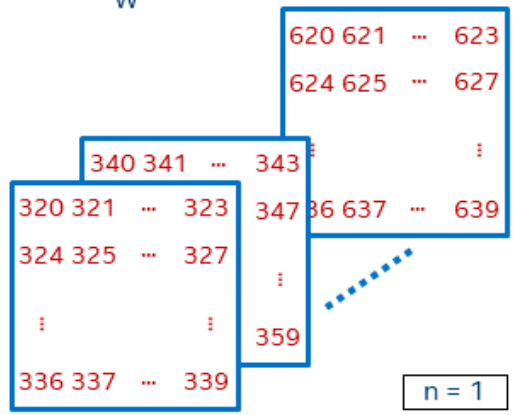
Copyright © Infineon Technologies AG 2025. All rights reserved.

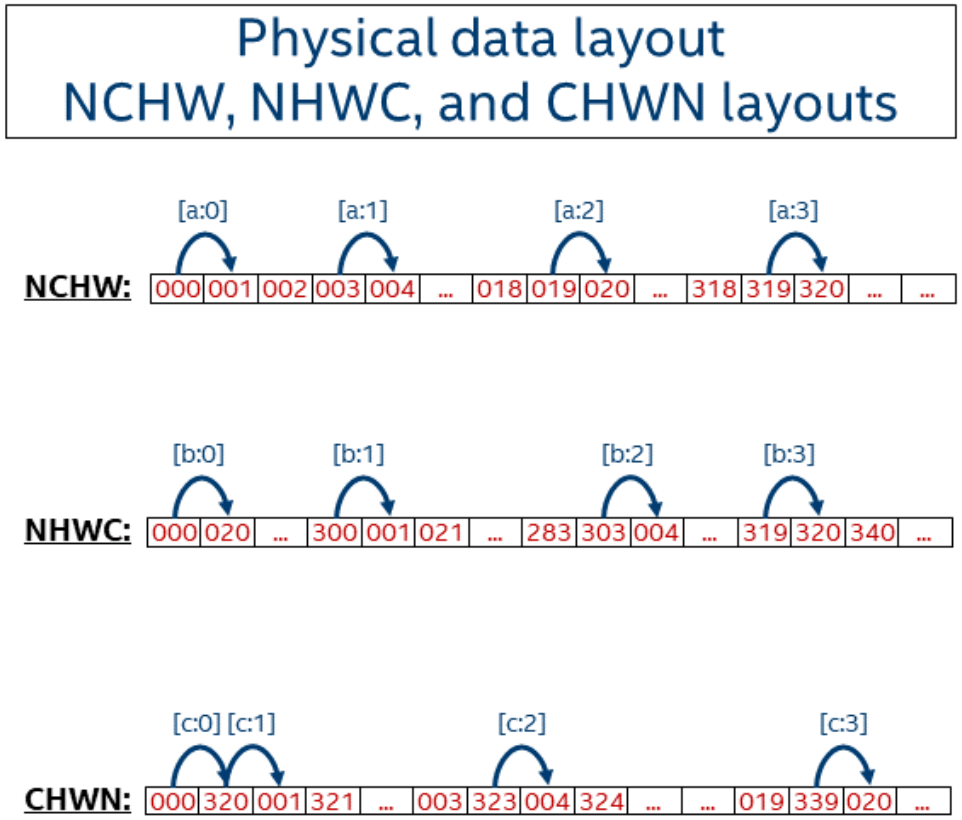
Formats

Batch 0



Batch 1



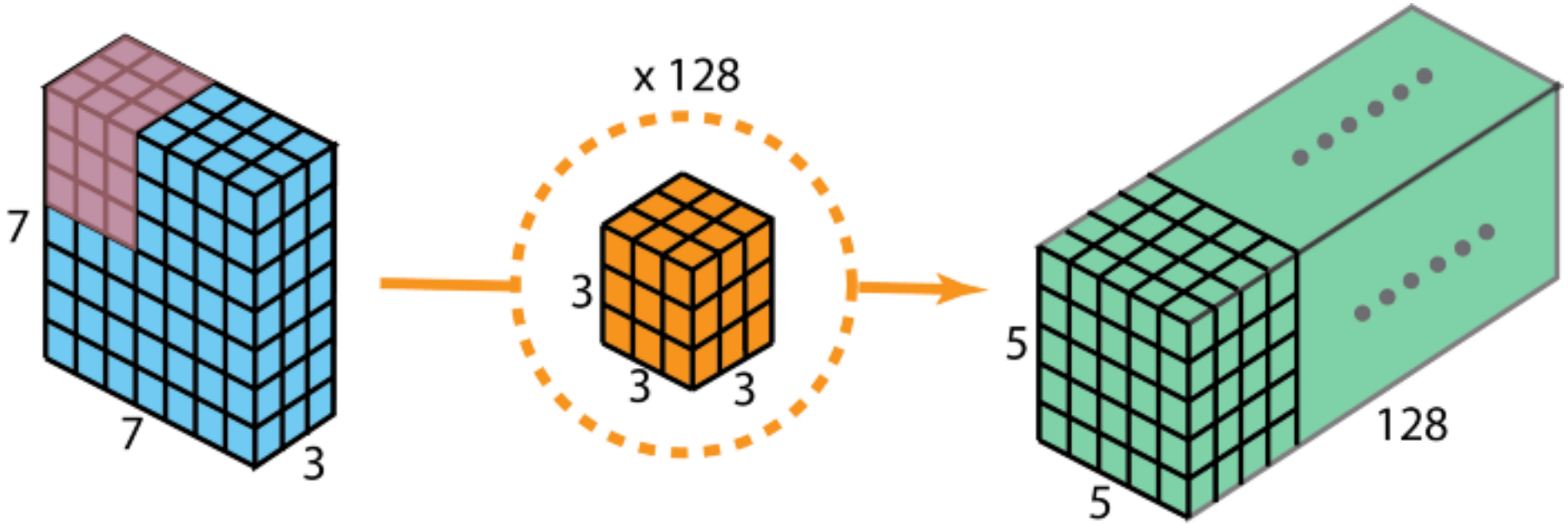



```
// channel major
for n in Batches
 for c in Channels
  for h in Heigth
   for w in Width
```

https://oneapi-src.github.io/oneDNN/dev_guide_understanding_memory_formats.html

Copyright © Infineon Technologies AG 2025. All rights reserved.

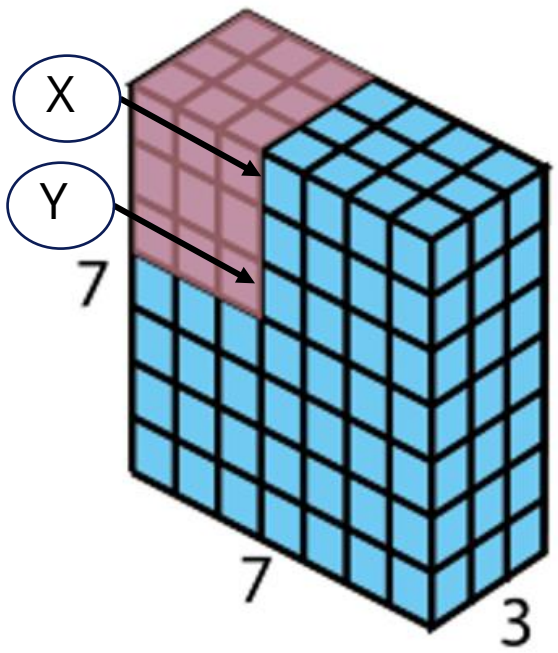
2D convolution with multiple Filters

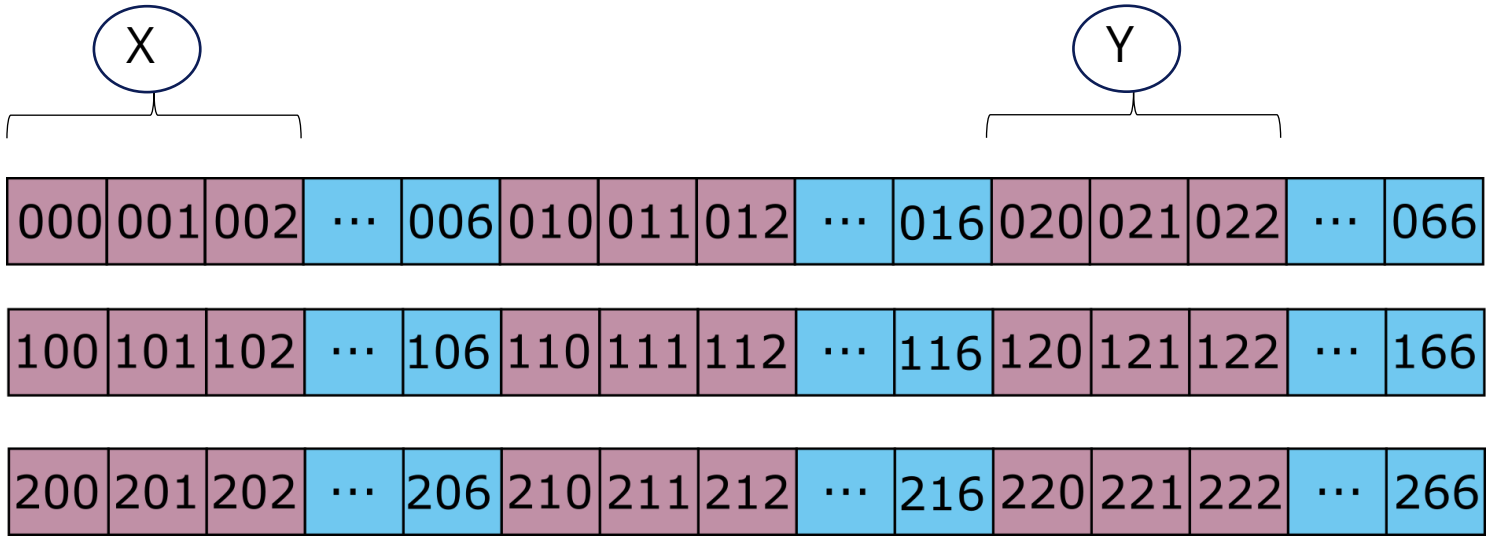


https://towardsdatascience.com/types-of-convolution-kernels-simplified-f040cb307c37

Copyright © Infineon Technologies AG 2025. All rights reserved.

2D convolution memory access pattern CHW (Channel – Heigth – Width)

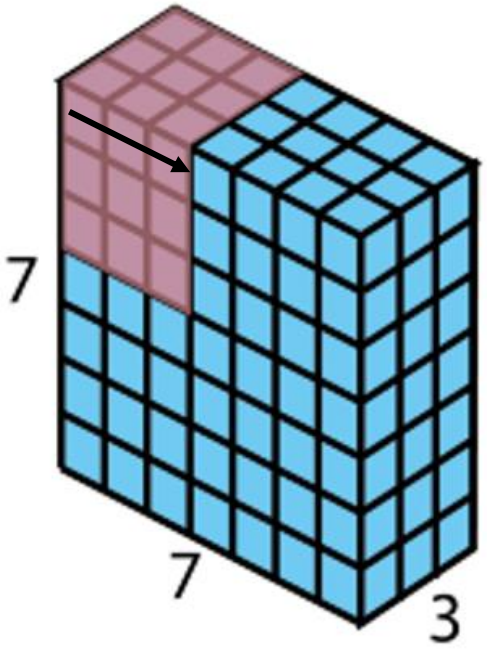




https://towardsdatascience.com/types-of-convolution-kernels-simplified-f040cb307c37

Copyright © Infineon Technologies AG 2025. All rights reserved.

2D convolution memory access pattern CHW (Channel – Heigth – Width)



1st kernel footprint:

000	001	002	⋯	006	010	011	012	⋯	016	020	021	022	⋯	066
100	101	102	⋯	106	110	111	112	⋯	116	120	121	122	⋯	166
200	201	202	⋯	206	210	211	212	⋯	216	220	221	222	⋯	266

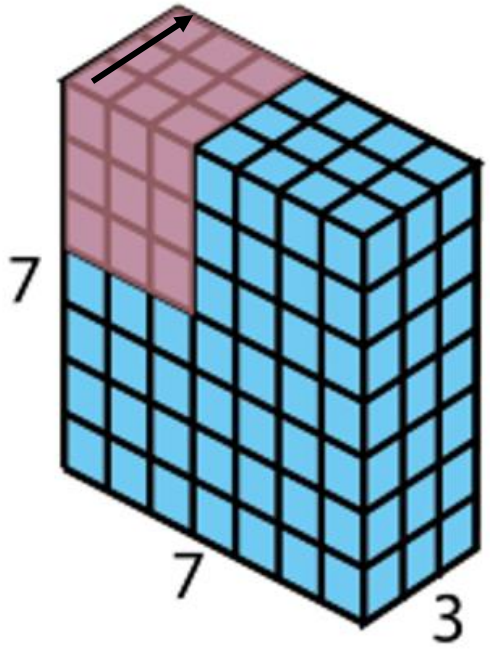
2nd kernel footprint:

000	001	002	003	⋯	006	011	012	013	⋯	016	021	022	023	⋯	066
100	101	102	103	⋯	106	111	112	113	⋯	116	121	122	123	⋯	166
200	201	202	203	⋯	206	211	212	213	⋯	216	221	222	223	⋯	266

https://towardsdatascience.com/types-of-convolution-kernels-simplified-f040cb307c37

Copyright © Infineon Technologies AG 2025. All rights reserved.

2D convolution memory access pattern HWC (Heigth – Width – Channel)



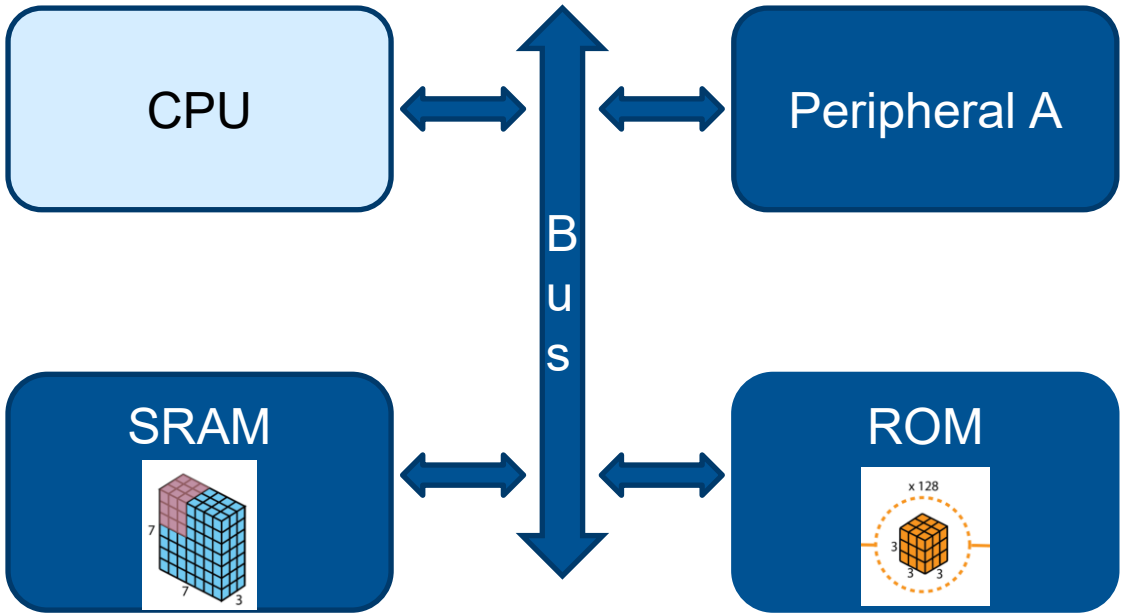
000	100	200	⋯	600	001	101	201	⋯	601	002	102	202	⋯	602
010	110	210	⋯	610	011	111	211	⋯	611	012	112	212	⋯	612
020	120	220	⋯	620	021	121	221	⋯	621	022	122	222	⋯	622
030	⋯	666												

https://towardsdatascience.com/types-of-convolution-kernels-simplified-f040cb307c37

Copyright © Infineon Technologies AG 2025. All rights reserved.

Weights and Activations Location

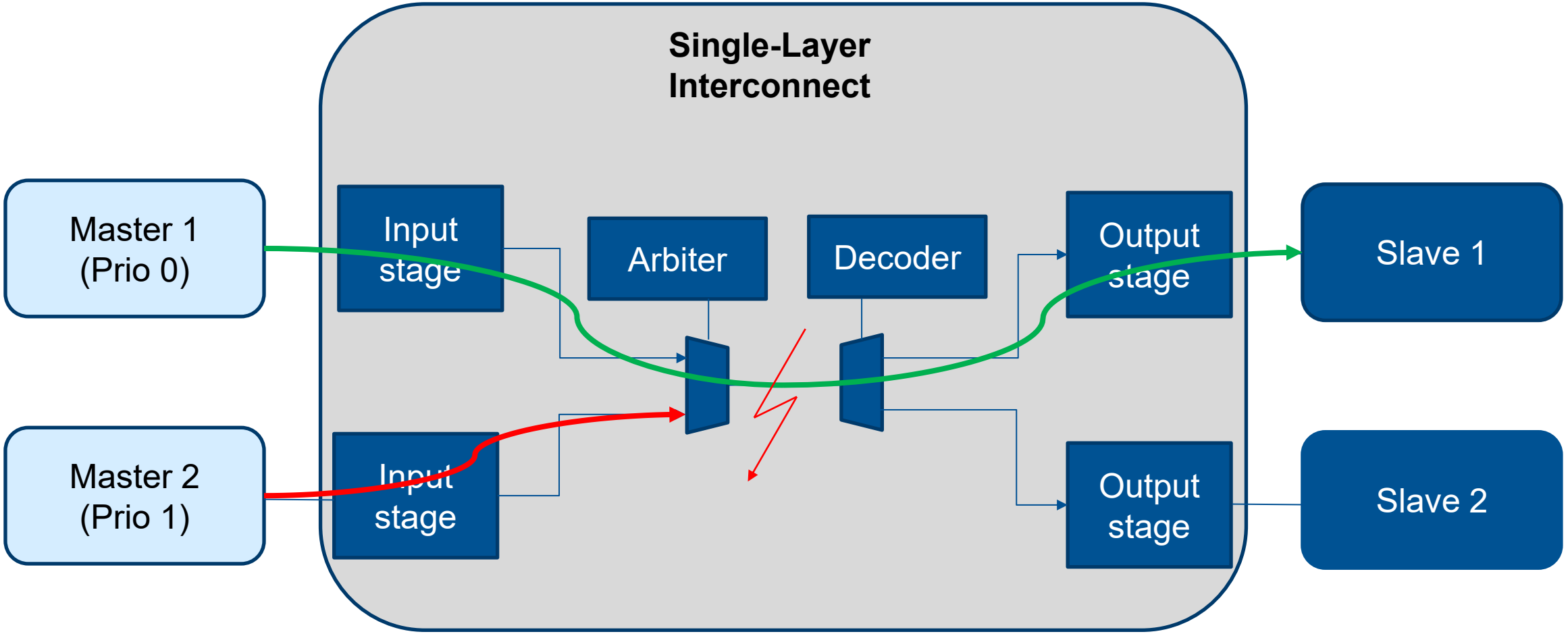
- Activations need to be in a RW Memory
- Weights are constants



Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

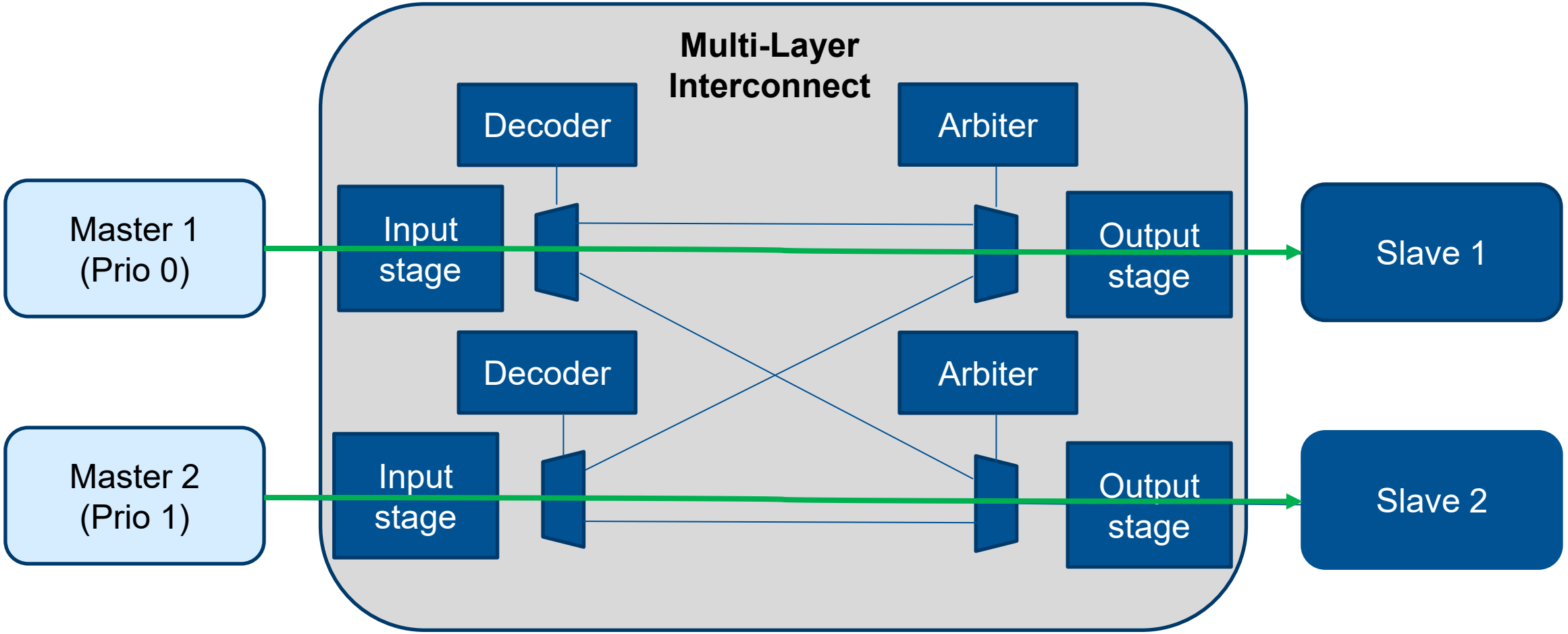
Single-layer Interconnect



Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

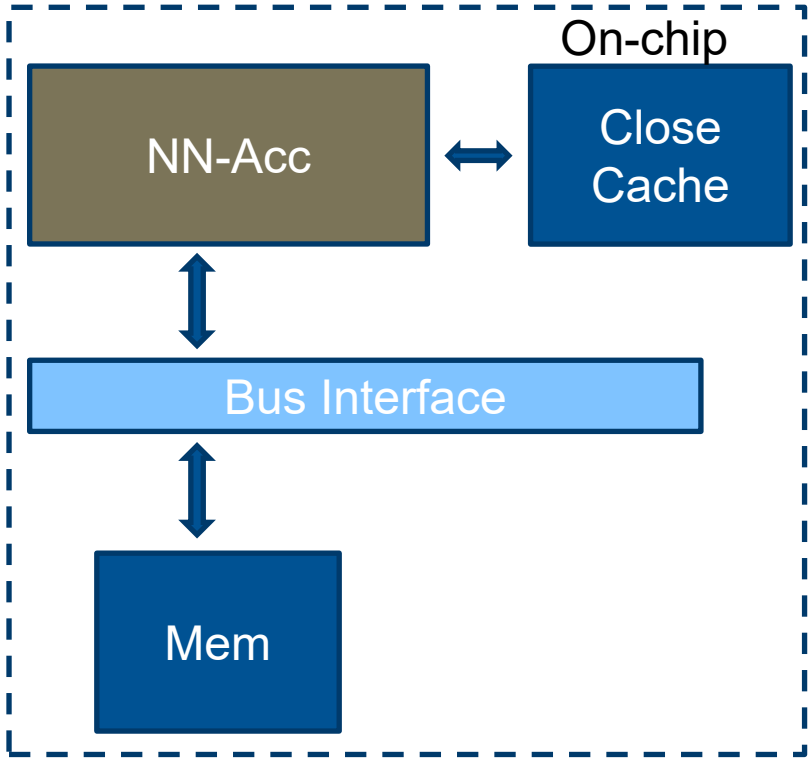
Multi-layer Interconnect



Source: https://developer.arm.com/documentation/dvi0045/latest/

Copyright © Infineon Technologies AG 2025. All rights reserved.

NN-Accelerator close Cache



Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

Accelerator-close Cache (CHW)

In Memory:

000	001	002	⋯	006	010	011	012	⋯	016	020	021	022	⋯	066
100	101	102	⋯	106	110	111	112	⋯	116	120	121	122	⋯	166
200	201	202	⋯	206	210	211	212	⋯	216	220	221	222	⋯	266

In Cache:

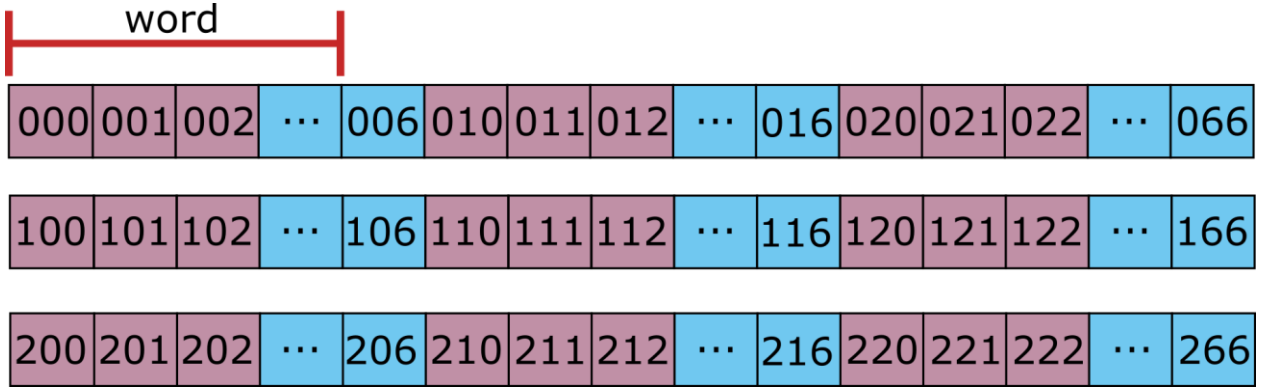
001	002	011	012	021	022
101	102	111	112	121	122
201	202	211	212	221	222

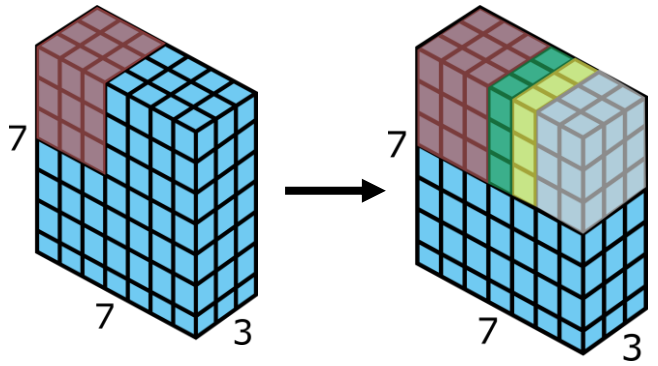
Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

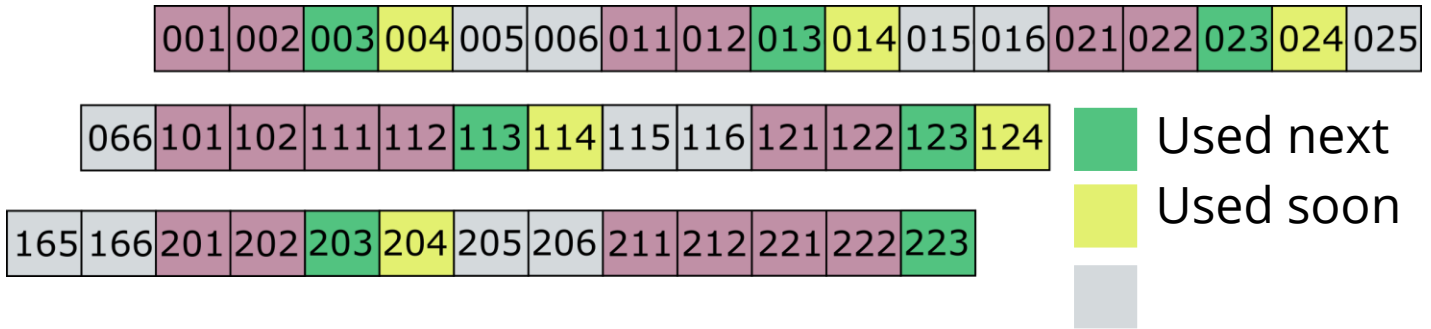
Preload close Cache (CHW) (without additional Memory accesses)

In Memory:





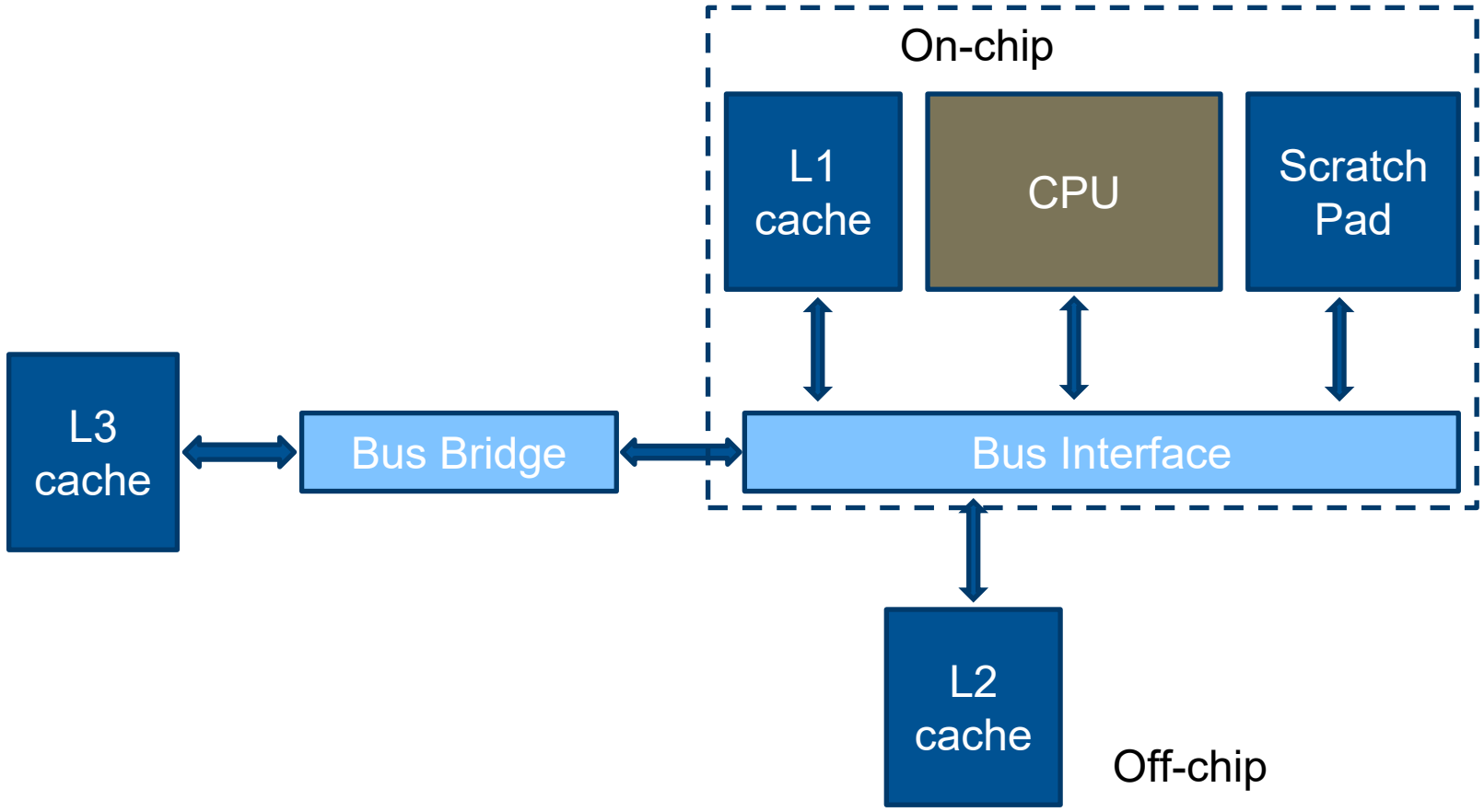
In Cache:



Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

Scratch Pad Topology



Source: Embedded Systems for Machine learning lecture @ TUM

Copyright © Infineon Technologies AG 2025. All rights reserved.

Energy saving by using a Scratch Pad

Cache per access (2 kbytes)	4.57 nJ
Scratch pad per access (2 kbytes)	1.53 nJ
Main memory read access, 2 bytes	24.00 nJ
Main memory read access, 4 bytes	49.30 nJ
Main memory write access, 4 bytes	41.10 nJ

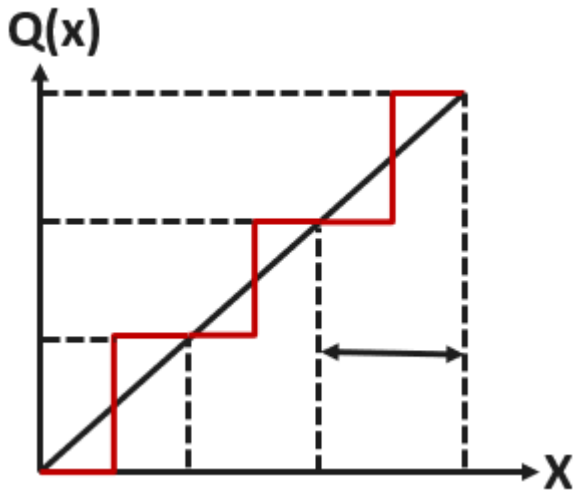
approx. 3x

Source: Banakar, Rajeshwari, et al. "Scratchpad memory: A design alternative for cache on-chip memory in embedded systems." *Proceedings of the Tenth International Symposium on Hardware/Software Codesign. CODES 2002 (IEEE Cat. No. 02TH8627)*. IEEE, 2002.

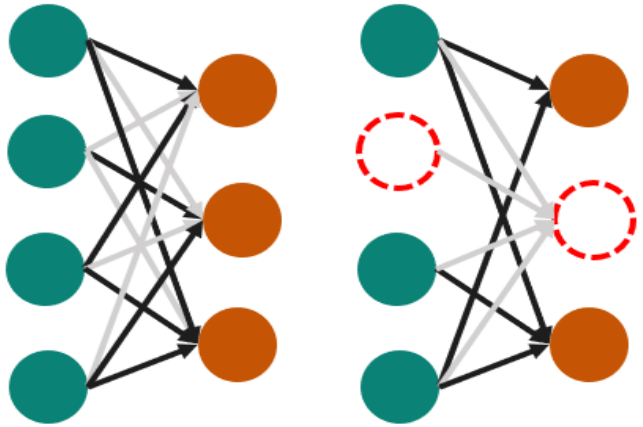
Copyright © Infineon Technologies AG 2025. All rights reserved.

Existing solutions for Weight Compression

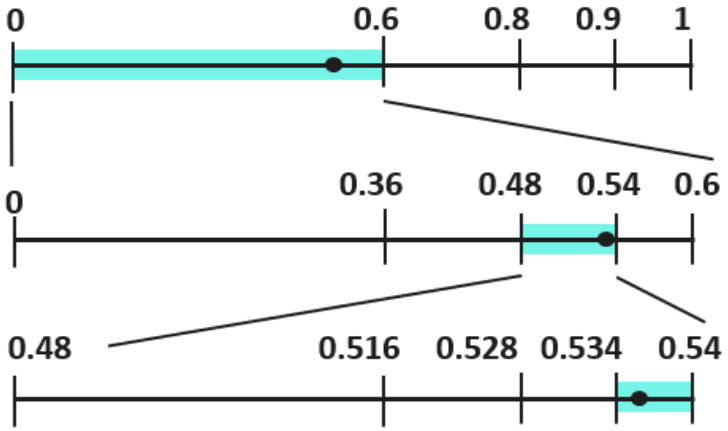
Quantization



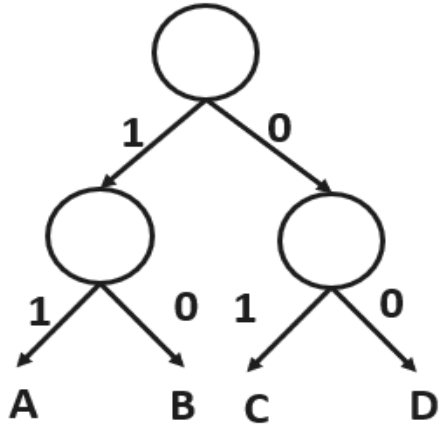
Weight and Network Pruning



Arithmetic coding

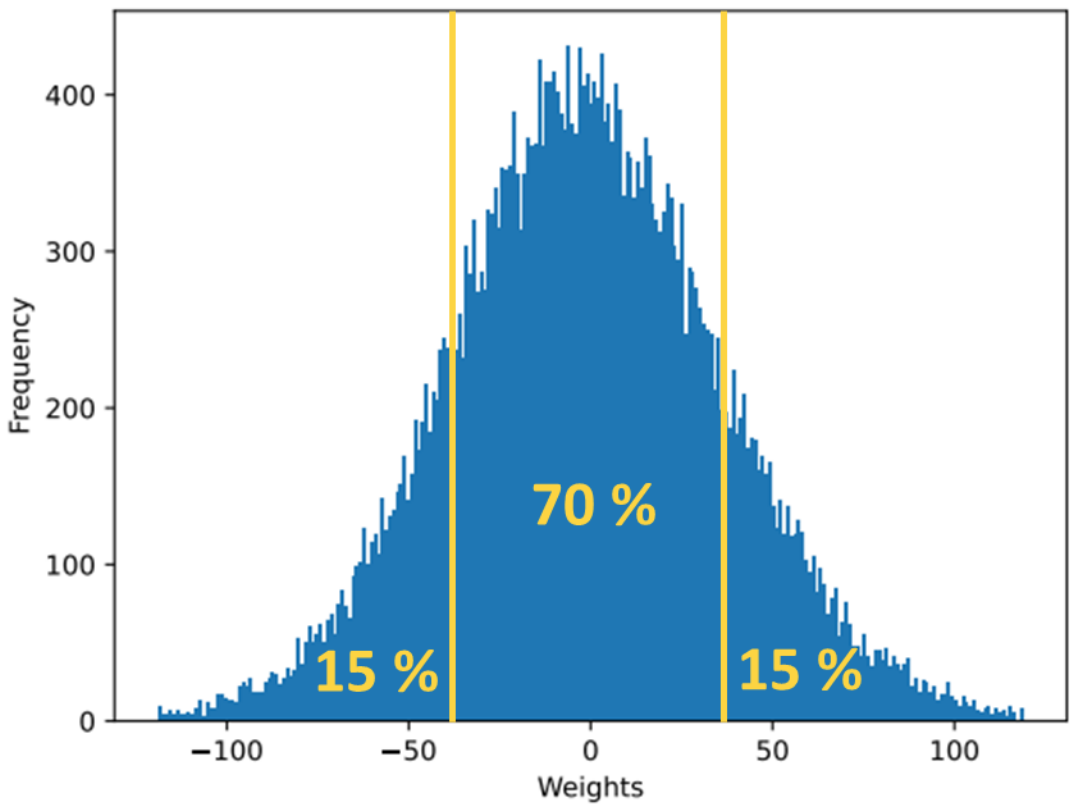


Huffman coding



Copyright © Infineon Technologies AG 2025. All rights reserved.

Weight distribution of a typical Convolution layer



Copyright © Infineon Technologies AG 2025. All rights reserved.

Golomb-Rice(GR) algorithm

- In GR coding, the GR_parameter is restricted to the powers of 2
- If N is the number to be encoded, then calculate
  - Quotient (Q) = N/ GR_parameter
  - Remainder (R) = N modulo GR_parameter
- The generated code: <Q(Unary coding)><R(Binary coding)>
  - Unary coding:
    - Q length of 1s or 0s
    - With termination of 0 or 1
  - Binary coding:
    - b = $\log_2(GR_parameter)$
    - Write R in binary representation using b bits

Example:

N = 21, GR_parameter = 8

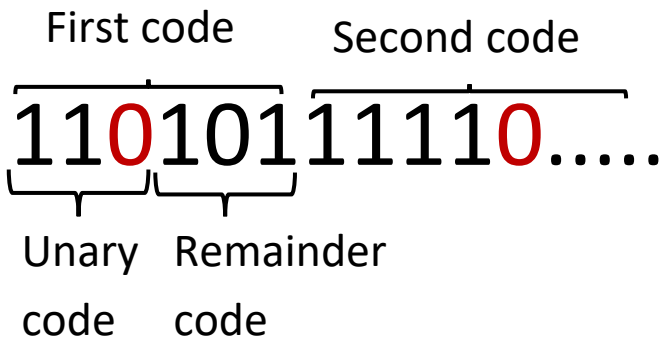
Q = 2

R = 5

Unary code = 110

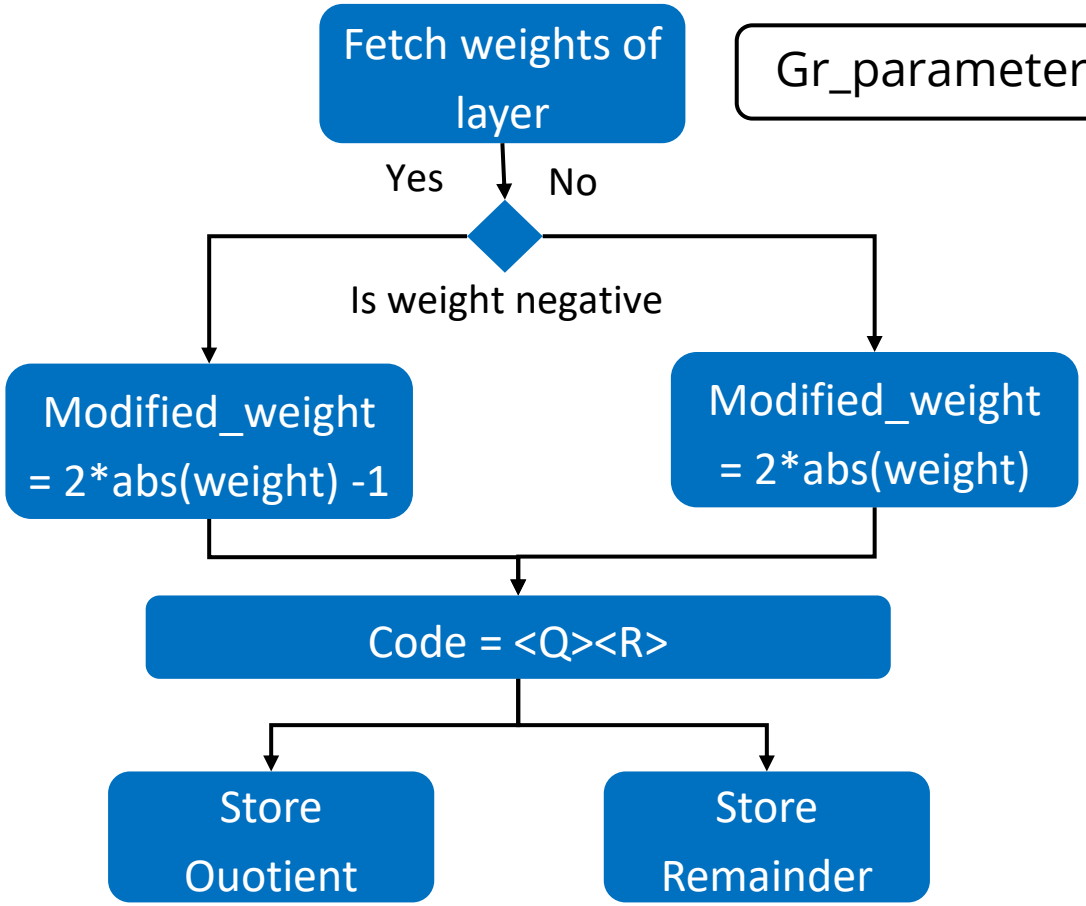
b = 3

Remainder code = 101



Copyright © Infineon Technologies AG 2025. All rights reserved.

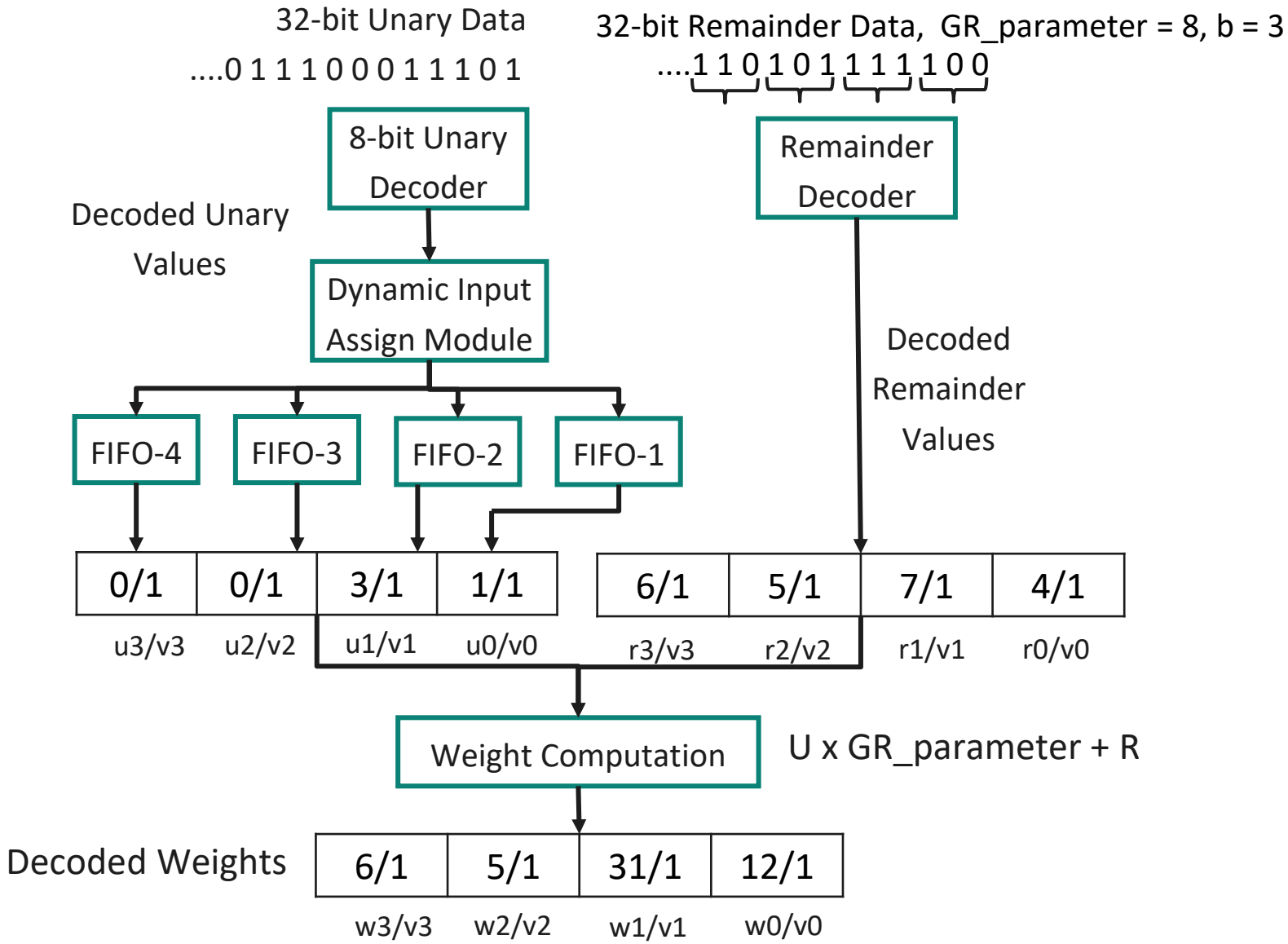
GR-encoding for non-pruned NN compression



Gr_parameter = pow(2,k), where k = 2,3,4,5,6

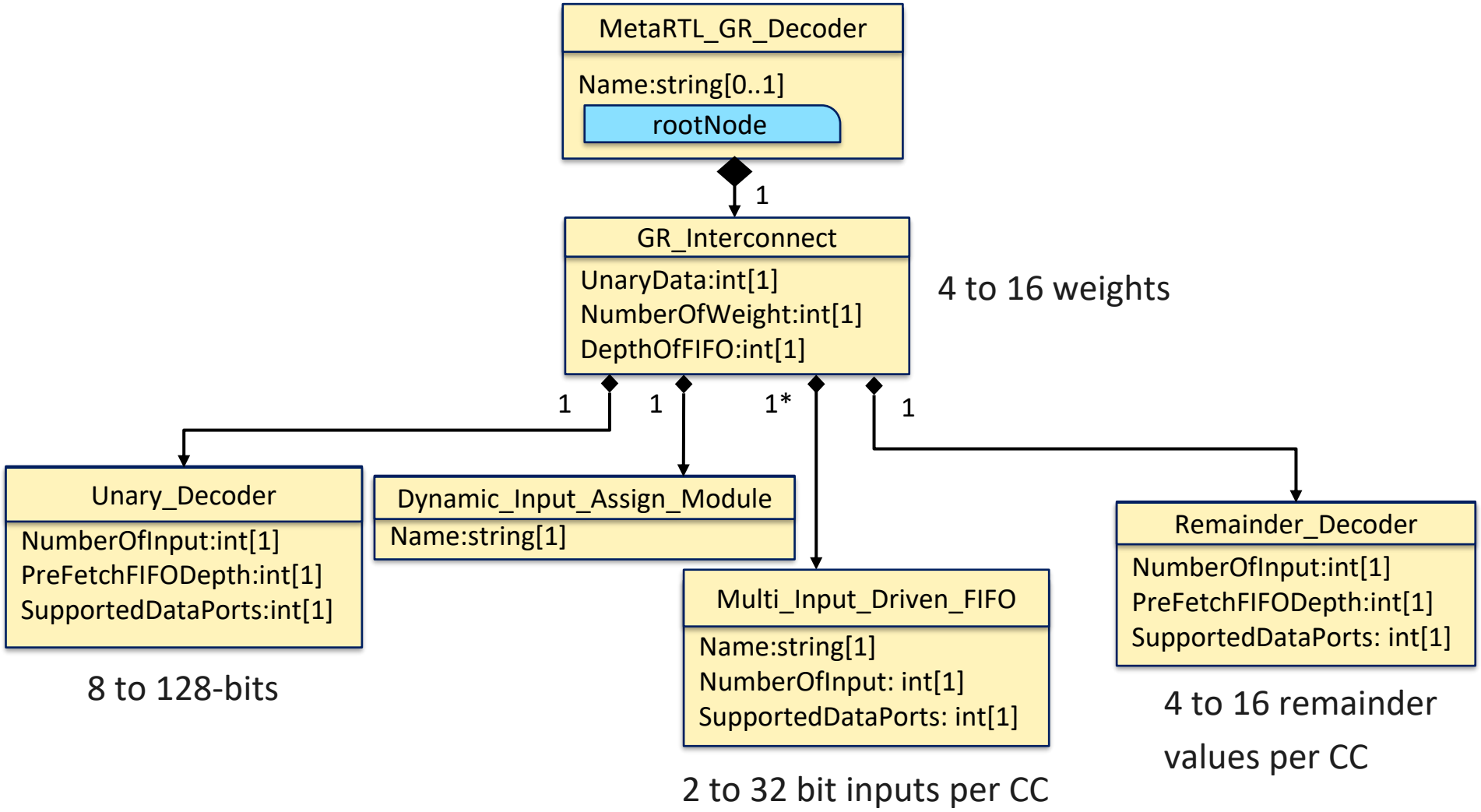
Copyright © Infineon Technologies AG 2025. All rights reserved.

8-bit Tree-Unary based Parallel GR Decoder



Copyright © Infineon Technologies AG 2025. All rights reserved.

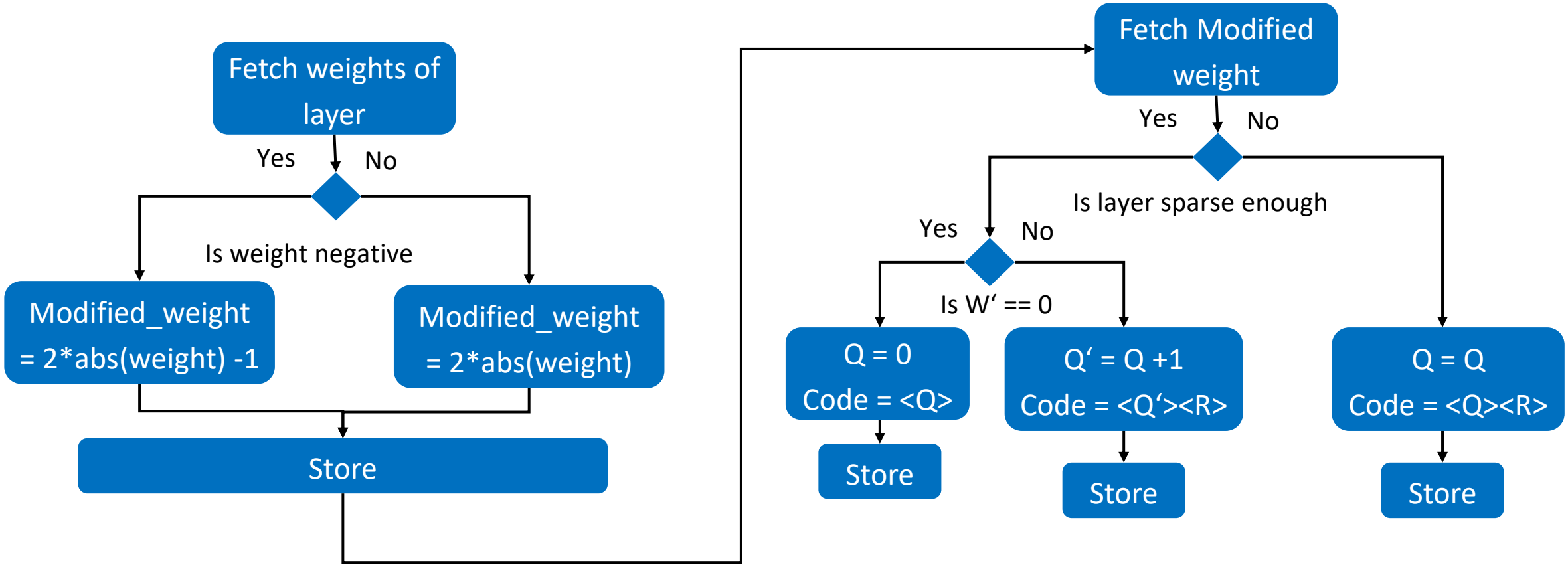
Metamodel for Scalable Parallel GR Decoder



Copyright © Infineon Technologies AG 2025. All rights reserved.

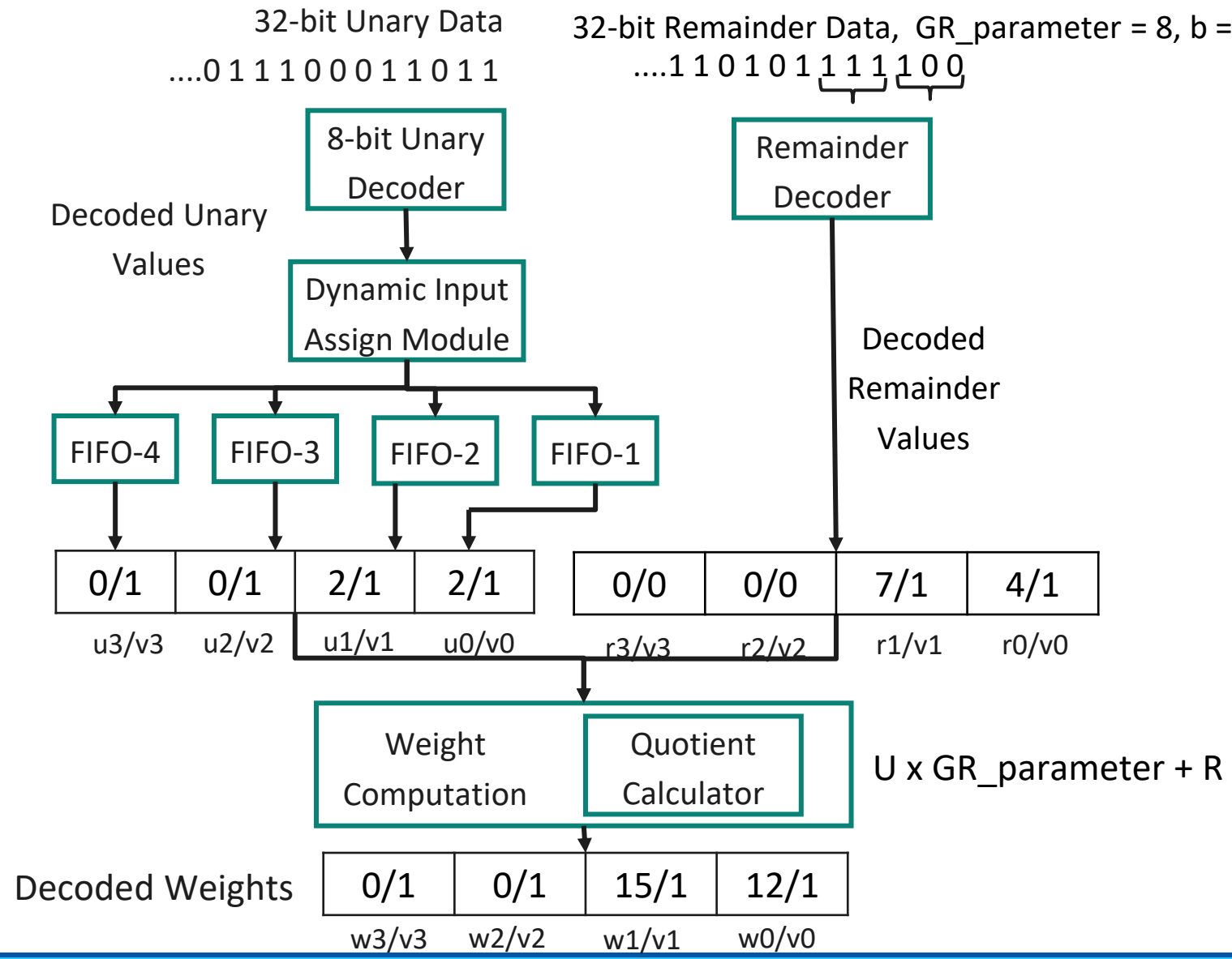
GR-encoding for non-pruned and pruned NN compression

Gr_parameter = pow(2,k), where k = 2,3,4,5,6



Copyright © Infineon Technologies AG 2025. All rights reserved.

Optimized Golomb-Rice Parallel Decoder



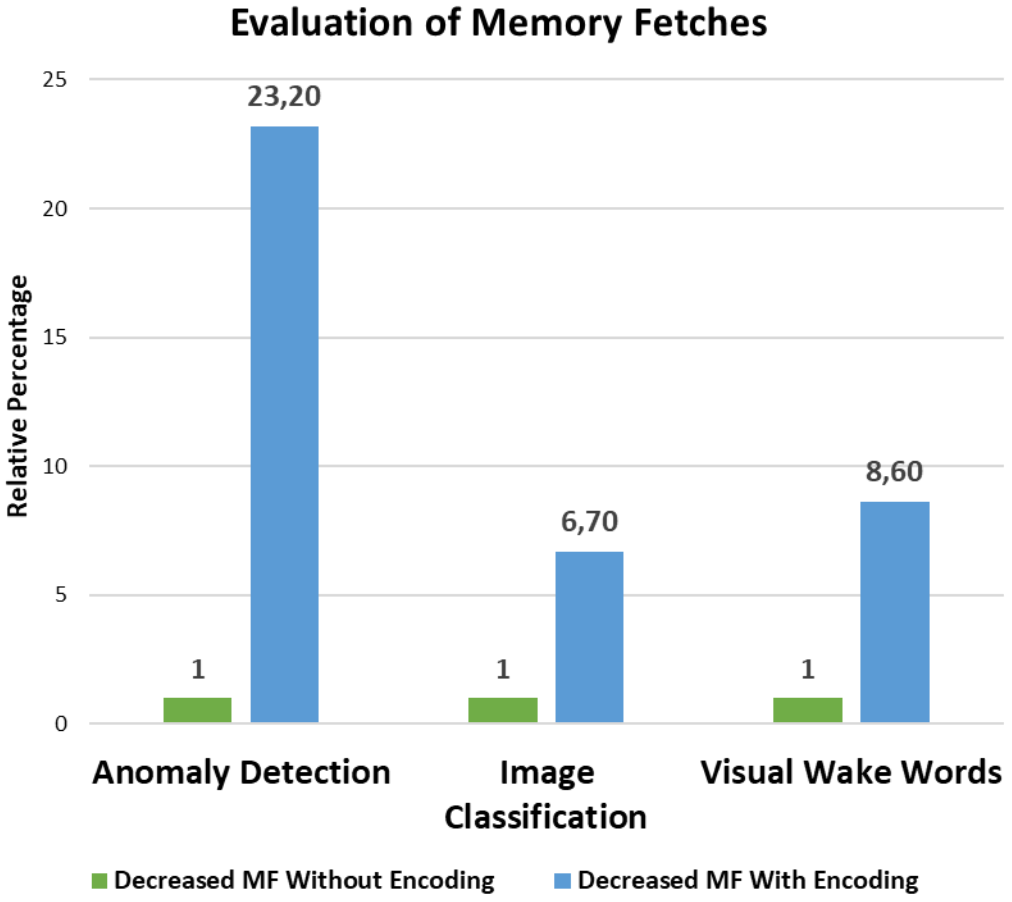
Copyright © Infineon Technologies AG 2025. All rights reserved.

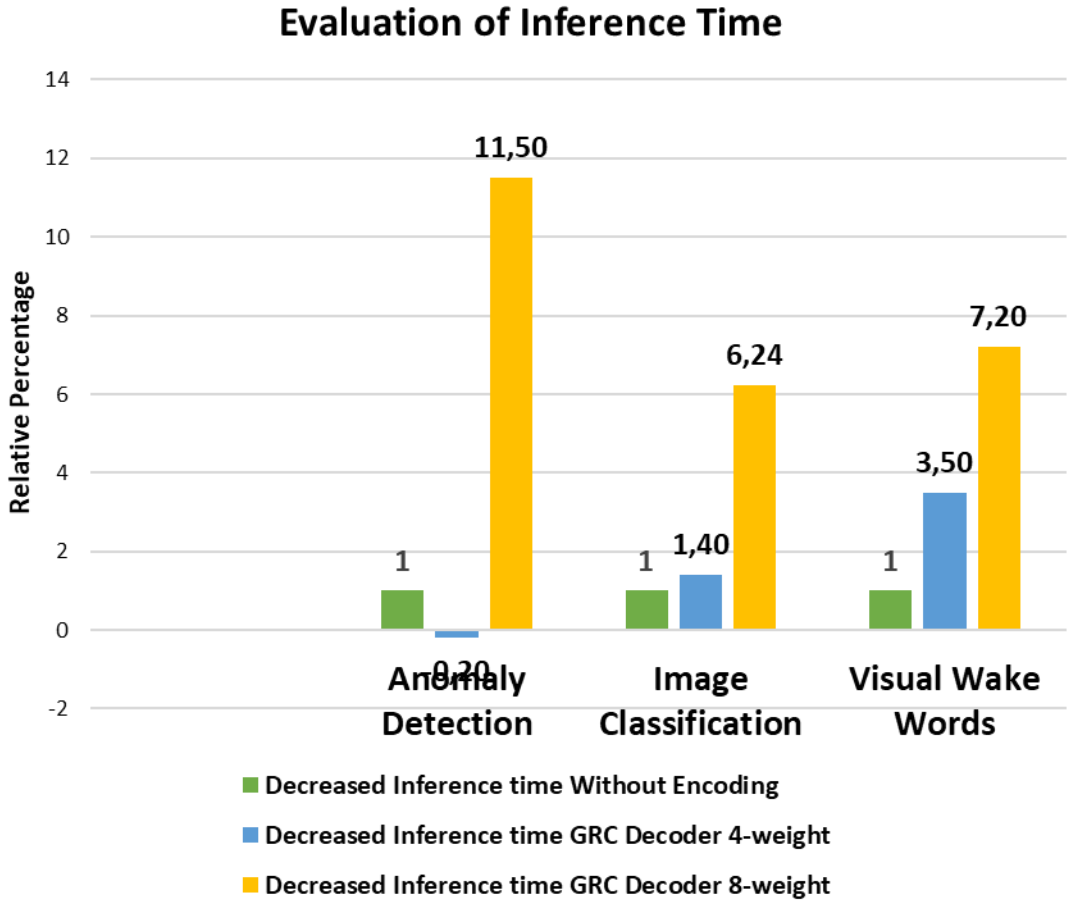
Test cases for performance evaluation

Tiny ML Benchmark Model [1]	Model	Dataset
Anomaly Detection	AutoEncoder [2]	TOYADMOS [2] [3] [4]
Image Classification	ResNet [7]	CIFAR10 [8]
Visual Wake Words	MobileNetV1 [9]	Visual Wake Word Dataset [10]

Copyright © Infineon Technologies AG 2025. All rights reserved.

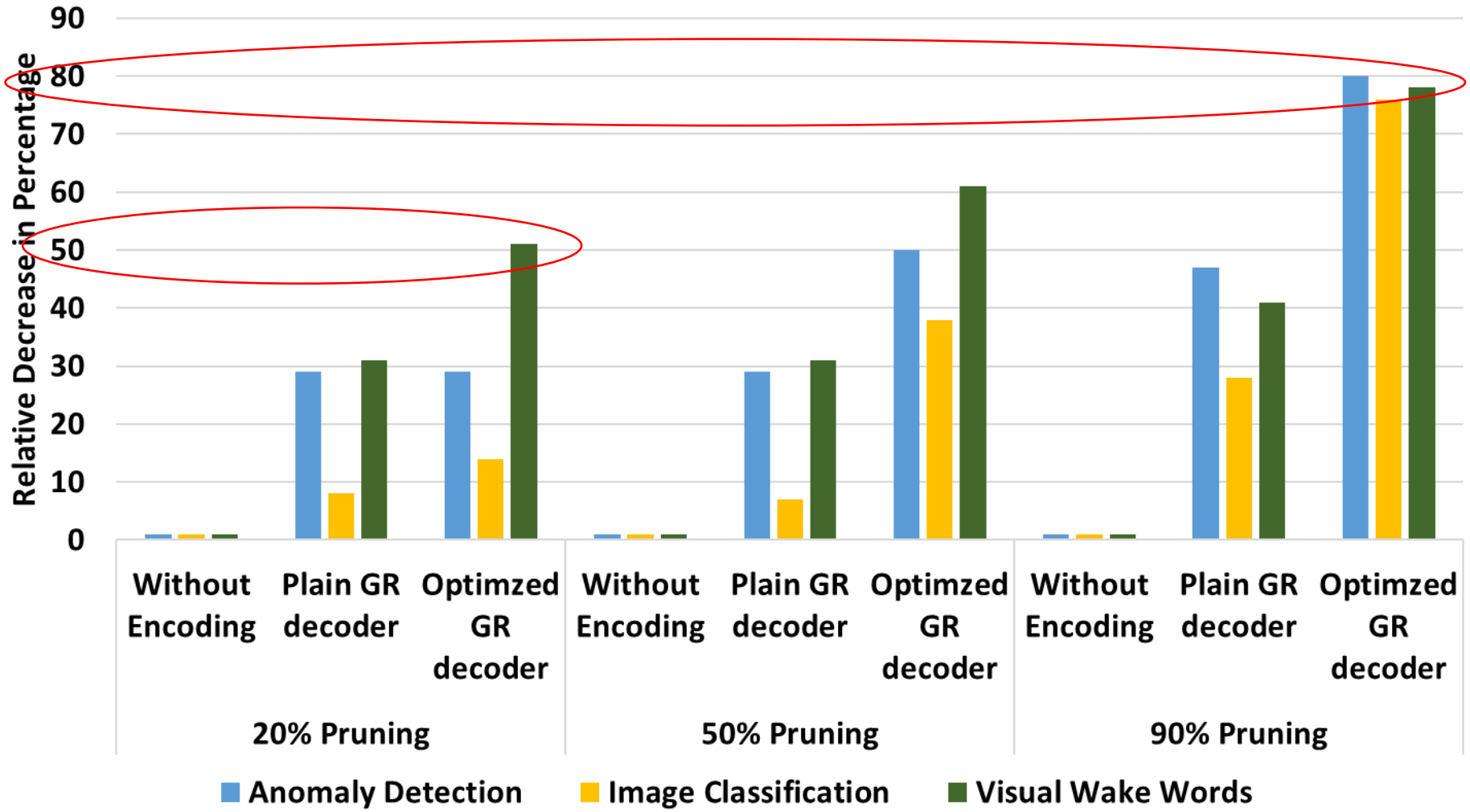
Compression and performance evaluation for non-pruned models





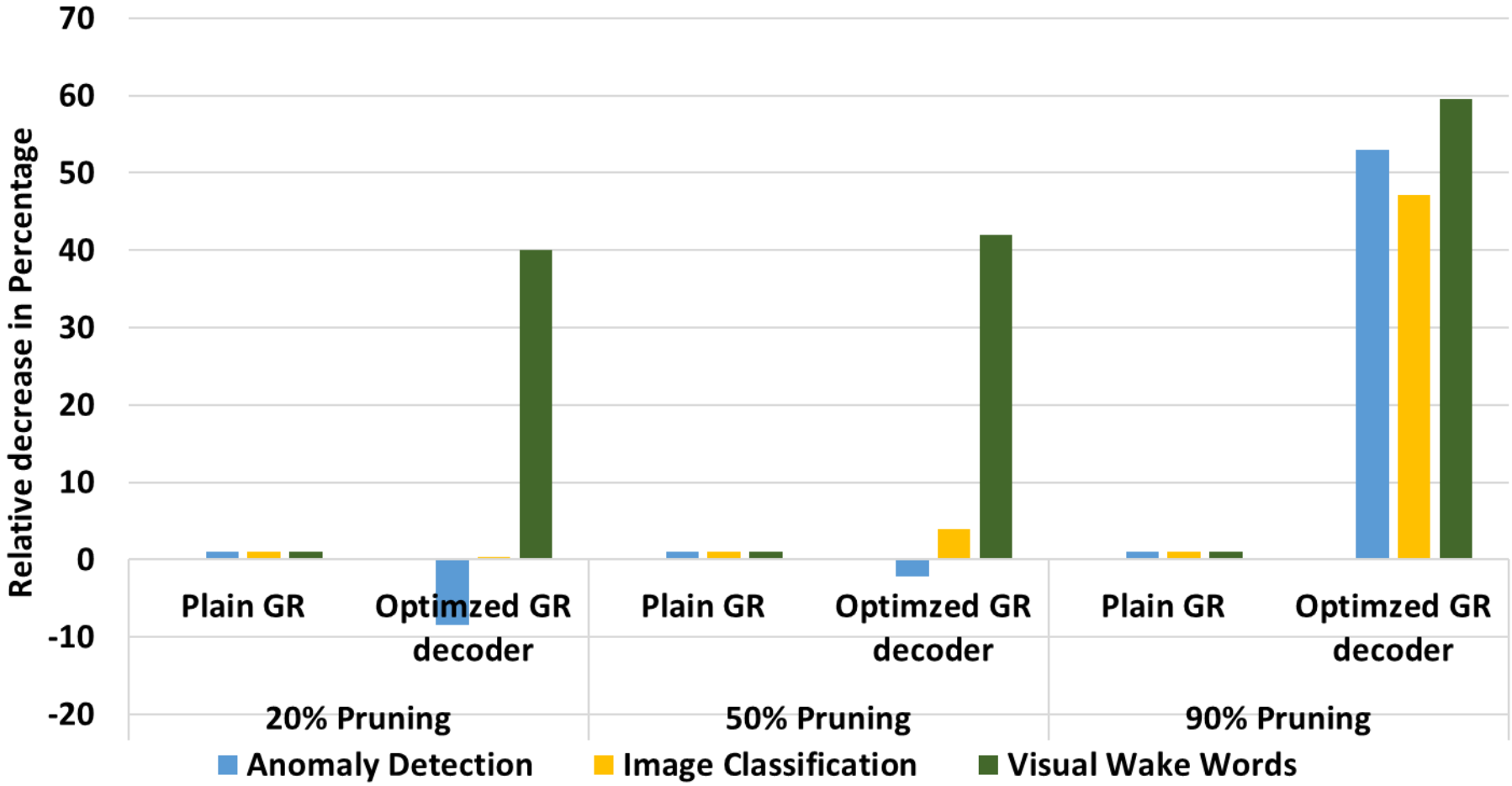
Copyright © Infineon Technologies AG 2025. All rights reserved.

Evaluation of Memory Fetches for pruned models



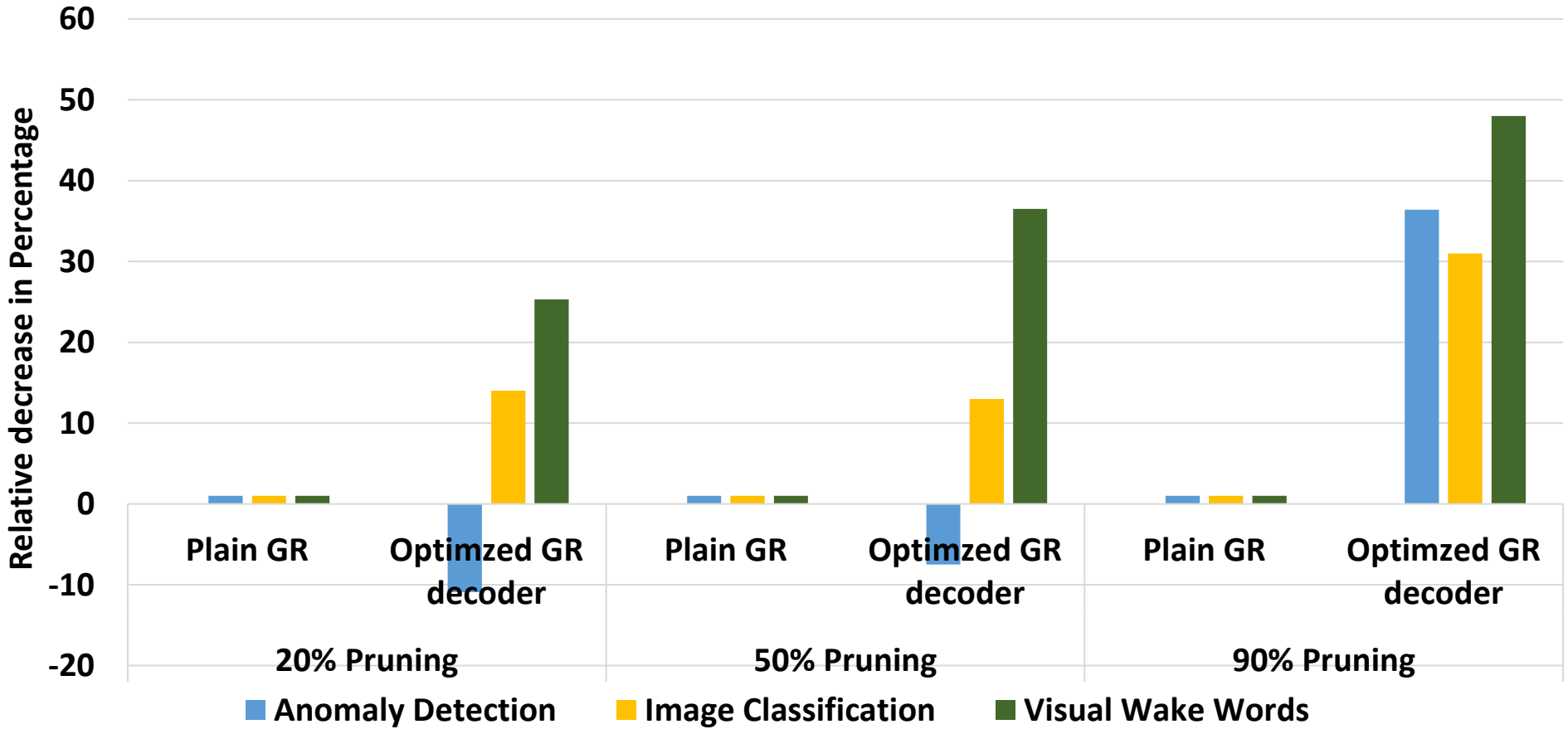
Copyright © Infineon Technologies AG 2025. All rights reserved.

Evaluation of Inference time for 4-weight dec. (pruned)



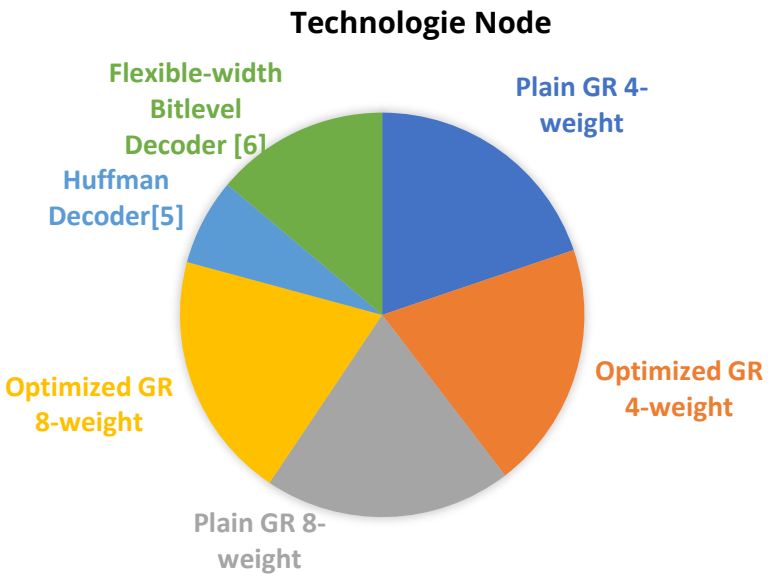
Copyright © Infineon Technologies AG 2025. All rights reserved.

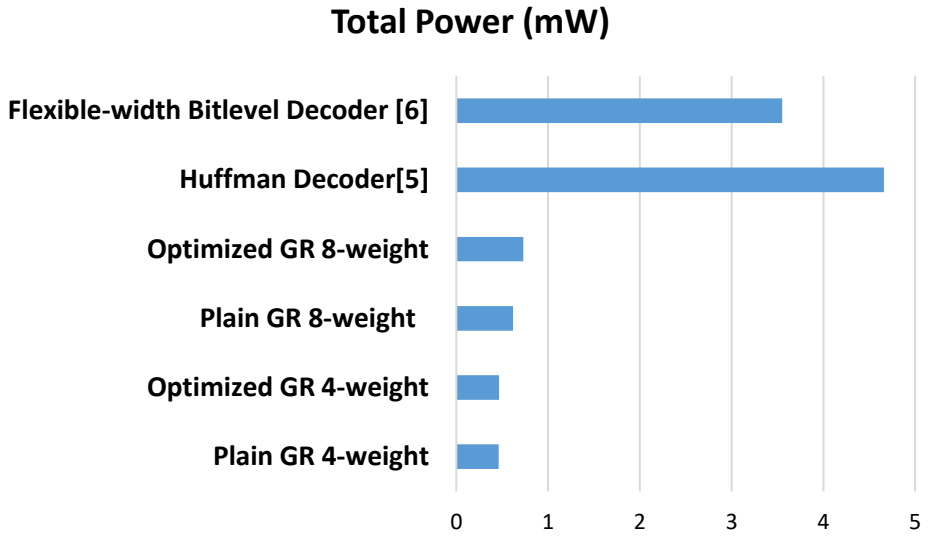
Evaluation of Inference time for 8-weight decoding

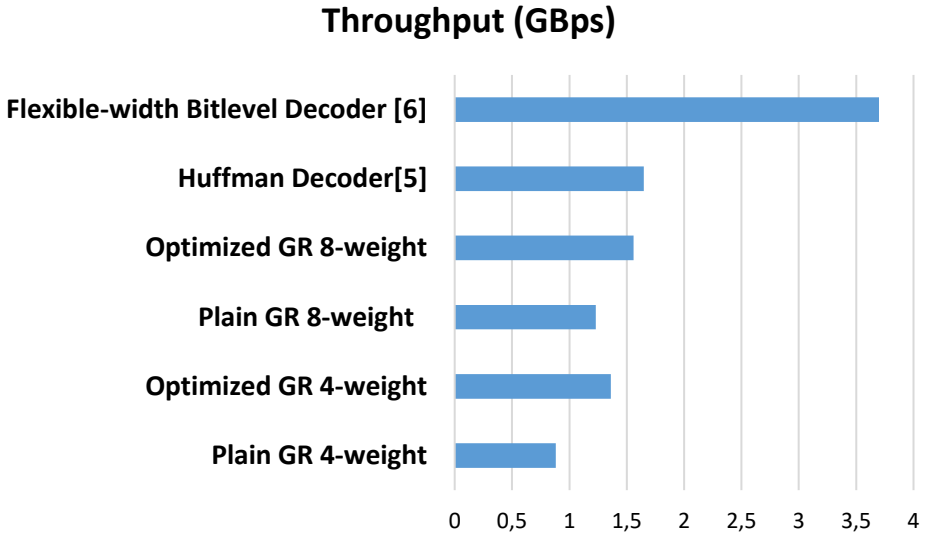


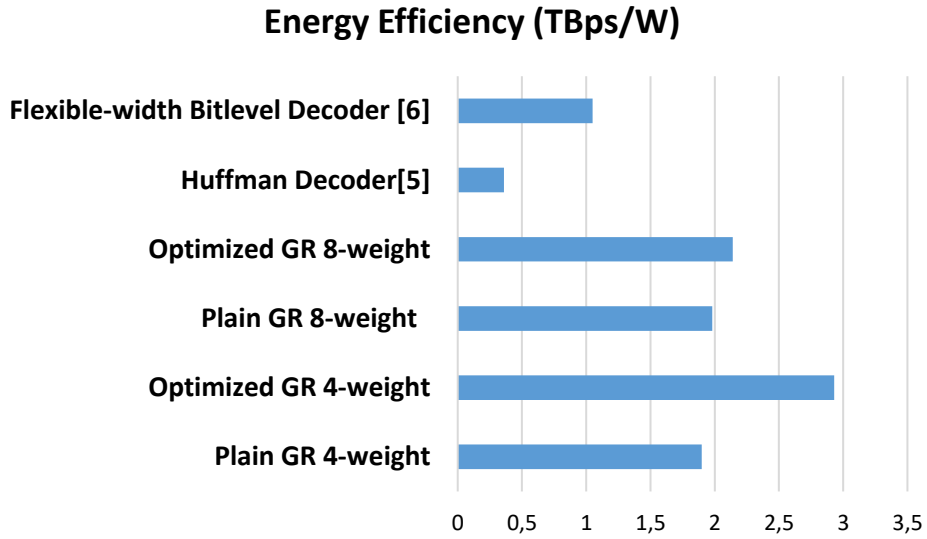
Copyright © Infineon Technologies AG 2025. All rights reserved.

Analysis of Synthesis Results



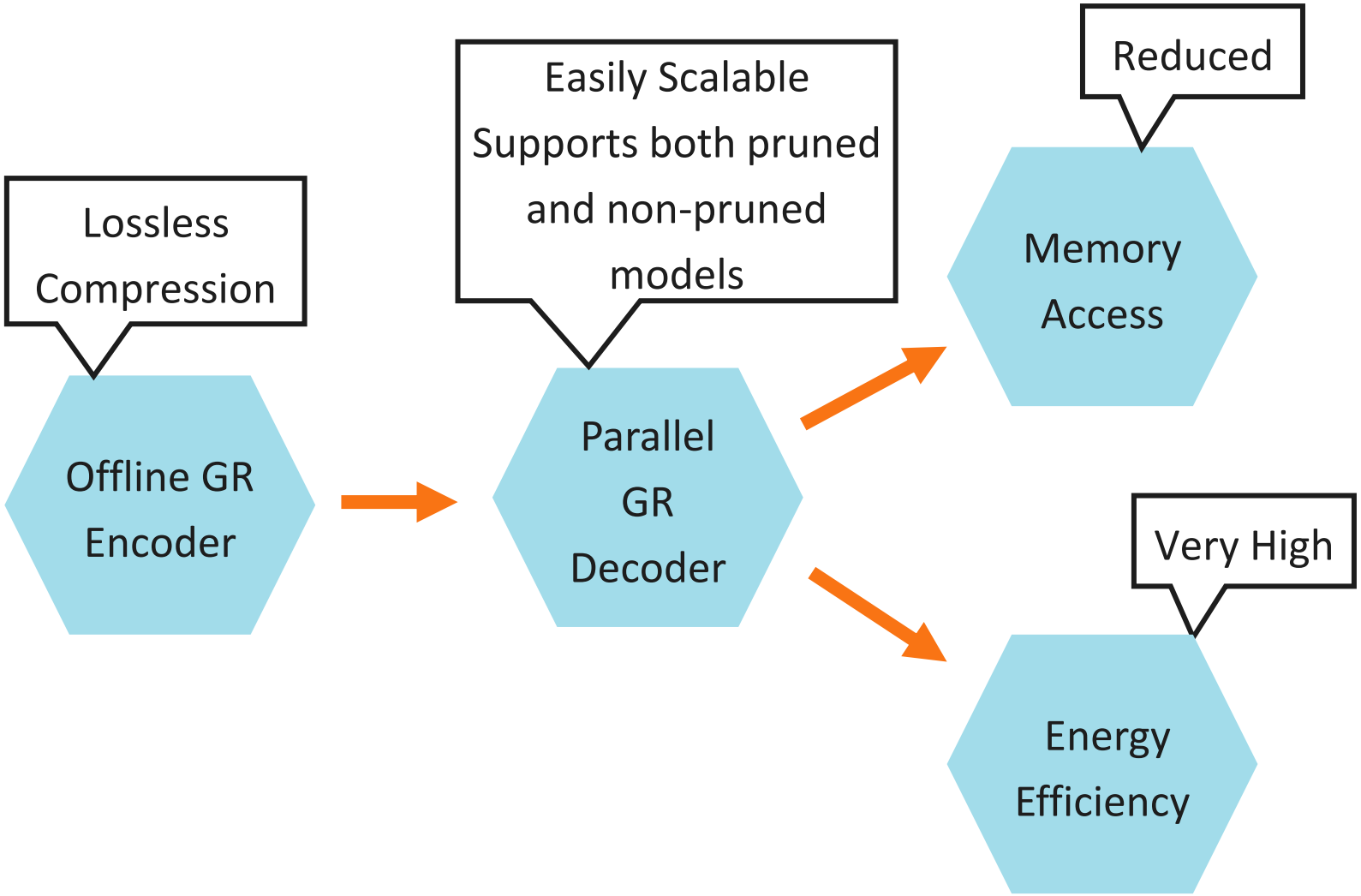






Copyright © Infineon Technologies AG 2025. All rights reserved.

Weight Compression through GR Encoder and Decoder



Copyright © Infineon Technologies AG 2025. All rights reserved.

Thank You

Sebastian.Prebeck@infineon.com

EAI

https://edge-ai-tech.eu/

The two-day innovative educational initiative is designed for students, young researchers, and professionals eager to expand their understanding and skills in these cutting-edge technologies.

Participants will engage in interactive discussions, hands-on sessions, and lectures covering foundational concepts and innovative applications of edge AI across various domains with leading experts, gain insights into real-world applications, and work on team-based projects to foster collaboration and innovation.

Copyright © Infineon Technologies AG 2025. All rights reserved.

EDGE AI Academy 2025

Summer School Objectives

- The EDGE AI Academy Summer School aims to empower participants with advanced knowledge, hands-on experience and an in-depth understanding of edge AI, which combines Artificial Intelligence (AI) with edge computing and the Internet of Things (IoT) to process data closer to its source.
- Edge AI technology is gaining prominence due to its ability to enhance real-time decision-making, reduce latency, improve data privacy, and diminish reliance on continuous internet connectivity.
- The topics will extend to RISC-V architectures as ultra-low-power high-performance RISC-V-based processors and ultra-low-power vector-based accelerators are advancing smart IoT and edge AI computing capabilities.

Copyright © Infineon Technologies AG 2025. All rights reserved.

References

[1] Colby R. Banbury, Vijay Janapa Reddi, Max Lam, William Fu, Amin Fazel, Jeremy Holleman, Xinyuan Huang, Robert Hurtado, David Kanter, Anton Lokhmotov, David A. Patterson, Danilo Pau, Jae-sun Seo, Jeff Sieracki, Urmish Thakker, Marian Verhelst, and Poonam Yadav. Benchmarking tinyml systems: Challenges and direction. CoRR, abs/2003.04821, 2020.

[2] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pages 313–317, 2019.

[3] Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. CoRR, abs/1909.09347, 2019.

[4] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, and Noboru Harada. Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring, 2020.

[5] Sudhir Satpathy, Sanu Mathew, Vikram Suresh, Vinodh Gopal, James Guilford, Mark Anders, Himanshu Kaul, Amit Agarwal, Steven Hsu, and Kam Krisnnamurthy. 34.4mbps 1.56tbps/w deflate decompression accelerator featuring block-adaptive huffman decoder in 14nm tri-gate cmos for iot platforms. In ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC), pages 90–93, 2018.

[6] Junhan Zhu, Xiaoliang Chen, Li Du, Haoran Geng, Yichuan Bai, Yuandong Li, Yuan Du, and Zhongfeng Wang. Flexible-width bitlevel compressor for convolutional neural network. In 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), pages 1–4, 2021.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[8] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[9] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.

[10] Aakanksha Chowdhery, Pete Warden, Jonathon Shlens, Andrew Howard, and Rocky Rhodes. Visual wake words dataset. CoRR, abs/1906.05721, 2019.

Copyright © Infineon Technologies AG 2025. All rights reserved.