Reliable and Efficient hardware for Trustworthy and Sustainable Deep Neural Networks

Alberto Bosio

alberto.bosio@ec-lyon.fr

Acknowledgments

agence nationale de la recherche

RE-TRUSING Project, Grant number: ANR-21-CE24-0015
https://inl.cnrs.fr/projects/re-trusting/

AdequetDL Project, Grant number: ANR-18-CE23-0012

HORIZON 2020

FVLLMONTI

FRANCE 2030

PEPR IA Adapting

Outline

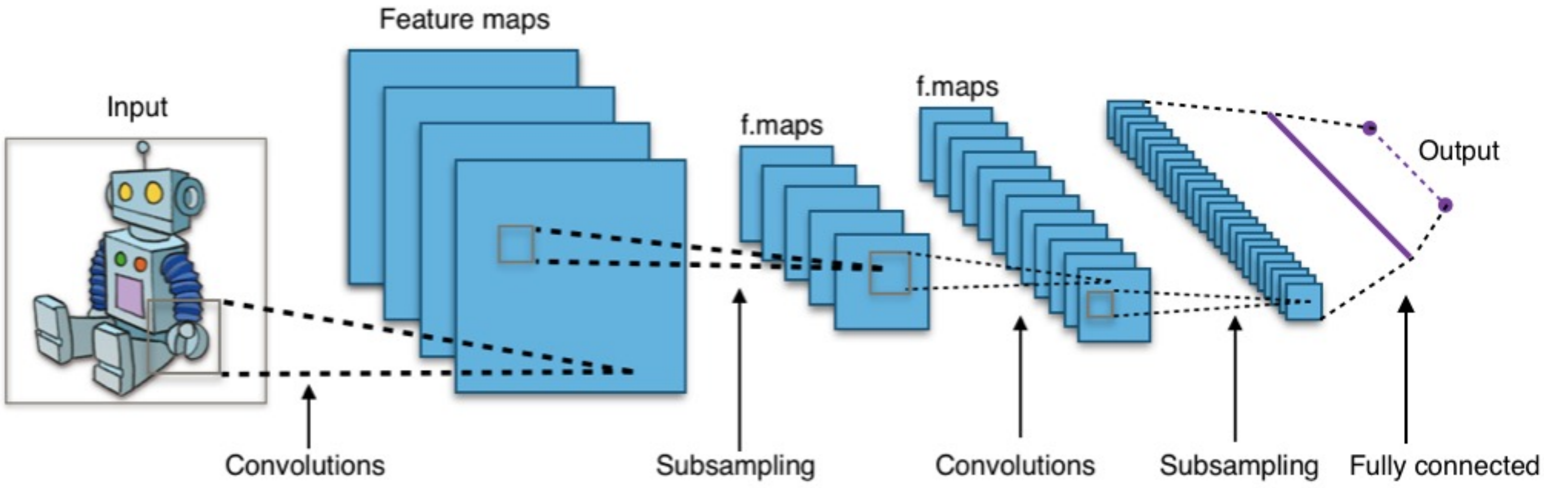
- Introduction
- Efficient HW accelerators
- Reliable HW accelerators
- Conclusions

Outline

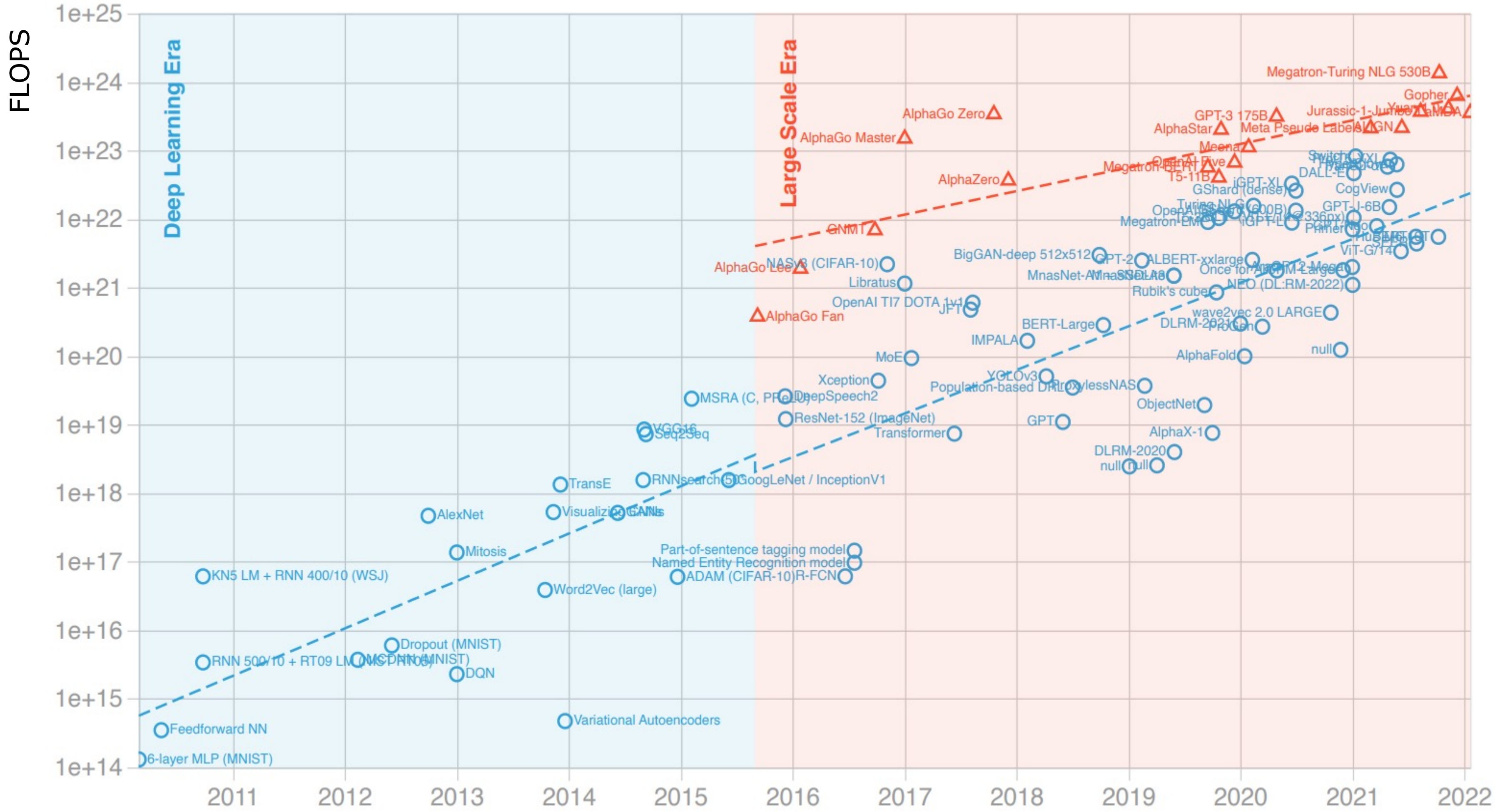
- Introduction
- Efficient HW accelerators
- Reliable HW accelerators
- Conclusions

Context

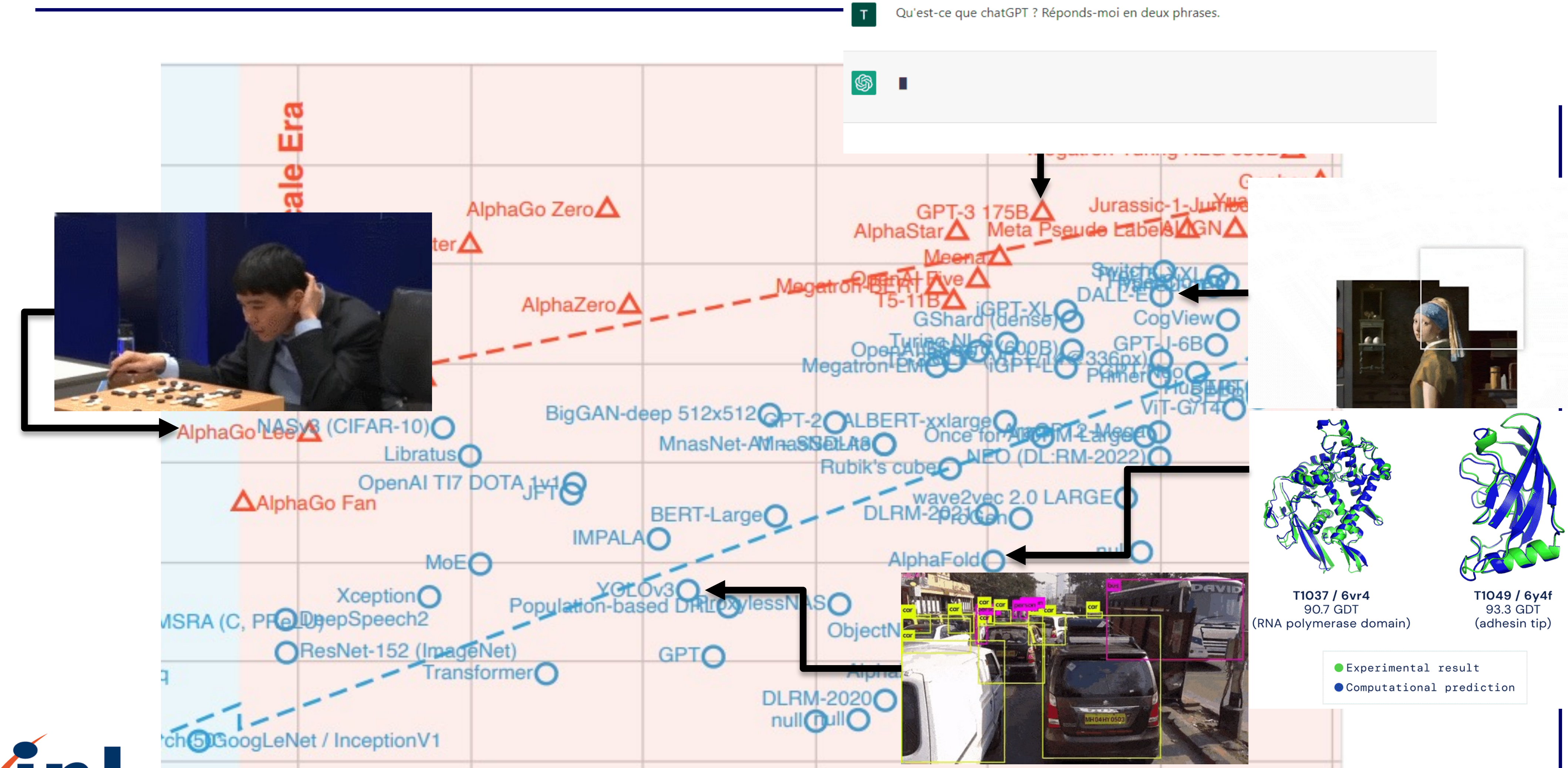
- Deep Neural Networks



Context


FLOPS
1e+25
1e+24
1e+23
1e+22
1e+21
1e+20
1e+19
1e+18
1e+17
1e+16
1e+15
1e+14
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Deep Learning Era
Large Scale Era
Megatron-Turing NLG 530B
Gopher
AlphaGo Zero
AlphaGo Master
GPT-3 175B
AlphaStar
Meta Pseudo Labels
Meena
AlphaZero
GNMT
AlphaGo Lee
AlphaGo Fan
NASv3 (CIFAR-10)
Libratus
BigGAN-deep 512x512
OpenAI TI7 DOTA 1v1
JFT
BERT-Large
IMPALA
MoE
Xception
MSRA (C, PReLU)
DeepSpeech2
ResNet-152 (ImageNet)
Transformer
GPT
ObjectNet
AlphaX-1
DLRM-2020
AlphaFold
wave2vec 2.0 LARGE
CogView
GPT-J-6B
ViT-G/14
Rubik's cube
NEO (DL:RM-2022)
VGG16
Seq2Seq
TransE
RNNsearch-50
GoogLeNet / InceptionV1
AlexNet
Mitosis
Part-of-sentence tagging model
Named Entity Recognition model
ADAM (CIFAR-10)
R-FCN
KN5 LM + RNN 400/10 (WSJ)
Word2Vec (large)
Dropout (MNIST)
RNN 500/10 + RT09 LM
DQN
Feedforward NN
Variational Autoencoders
6-layer MLP (MNIST)

Context



Qu'est-ce que chatGPT ? Réponds-moi en deux phrases.
AlphaGo Zero
AlphaZero
AlphaGo Lee
AlphaGo Fan
AlphaStar
GPT-3 175B
Meena
T5-11B
Libratus
OpenAI TI7 DOTA 1v1
BigGAN-deep 512x512
BERT-Large
IMPALA
MoE
Xception
DeepSpeech2
ResNet-152 (ImageNet)
Transformer
GPT
YOLOv3
AlphaFold
DLRM-2020
GoogLeNet / InceptionV1
CogView
GPT-J-6B
DALL-E
ViT-G/14
wave2vec 2.0 LARGE
Rubik's cube
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)
T1049 / 6y4f
93.3 GDT
(adhesin tip)
Experimental result
Computational prediction

Why we talk about Trustworthy and Sustainable AI?

- AI ethics: "the study of ethical and societal issues facing developers, producers, consumers, citizens, policy makers, and civil society organizations."

https://link.springer.com/article/10.1007/s43681-021-00043-6

Why we talk about Trustworthy and Sustainable AI?

- Waves:
  1. Fanciful scenarios of robot uprisings

  2. The problem of **explainability**
     - The lack of equal representation in training data and the resulting biases in AI models (https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people)
     - Hardware malfunctions:
       - Intentional: Adv Attacks
       - Un-intentional: Hardware Faults
  3. The **sustainable** development
     1. AlphaGo Zero generated 96 tonnes of CO_2 over 40 days of research training which amounts to 1000 h of air travel or a carbon footprint of 23 American homes
     2. Energy usage during ChatGPT's training has been estimated to be equivalent to that of an American household for over 700 years

Why we talk about Trustworthy and Sustainable AI?

- Waves:
  1. Fanciful scenarios of robot uprisings

  2. The problem of **explainability**
     - The lack of equal representation in training data g biases in AI models (https://www.theguardian.com/tech /12/google-racism-ban-gorilla-black-people)
     - Hardware malfunctions:
       - Intentional: Adv Attacks
       - Un-intentional: Hardware Faults

  3. The **sustainable** development
     1. AlphaGo Zero generated 96 tonnes of CO2 over h training which amounts to 1000 h of air travel or a carbo merican homes
     2. Energy usage during ChatGPT's training has bee equivalent to that of an American household for over 700 y

Outline

- Introduction
- Efficient HW accelerators
- Reliable HW accelerators
- Conclusions

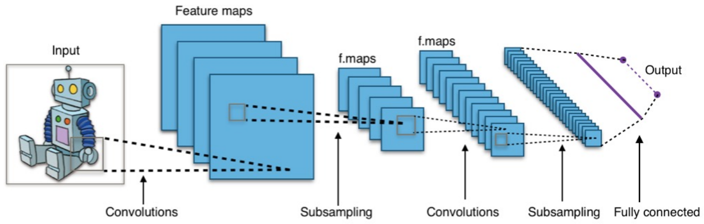
State of the Art

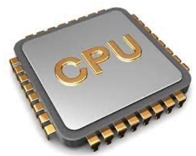
Software

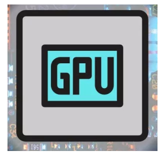
- Application software
- Virtual machines
- Operating system/Hypervisor
- Firmware

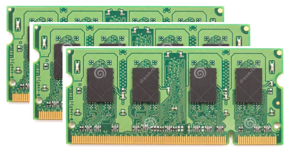
Hardware

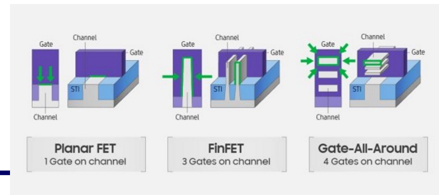
- Architecture
- Technology



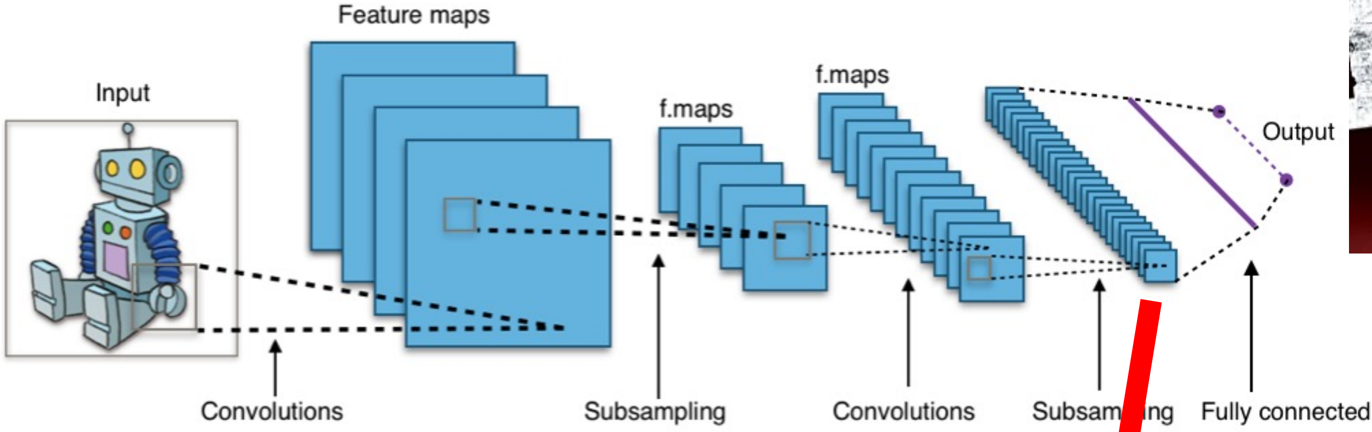




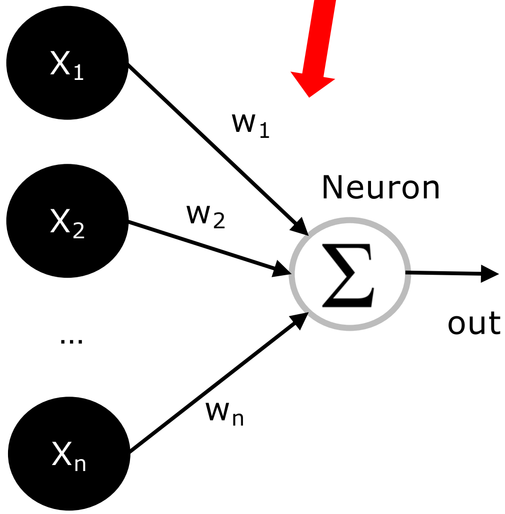




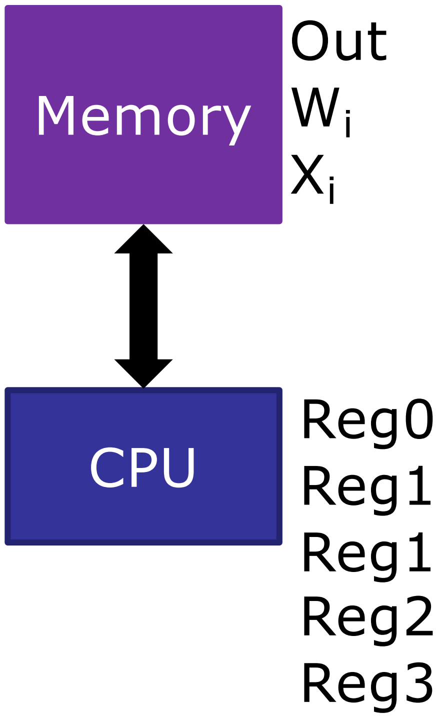
Why DNN are so complex?




```
for (I = 1 to N)
    out += Wi * Xi
```



Computer Architecture




```
for (I = 1 to N)
  out += Wi * Xi
```

assembly

```
Load Wi, reg0
Load Xi,  reg1
Load out, reg2
Mul reg3, reg1, reg0
Add reg2, reg3, reg2
Store reg2, out
```

6 memory accesses for instructions
4 memory accesses for data
2 operations

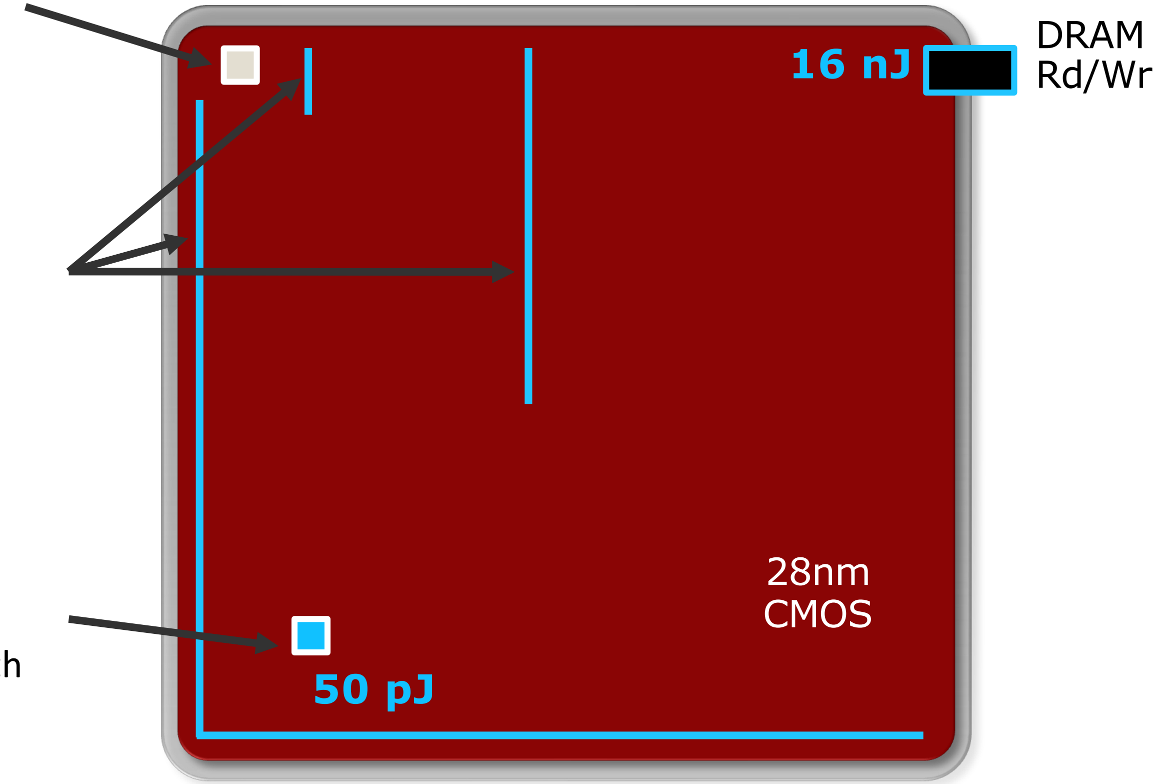
Energy Cost in a Processor

[Adapted from Dally, IPDPS'11]

- 64-bit FPU: 20pJ/op
- 32-bit addition: 0.05pJ
- 16-bit multiply: 0.25pJ

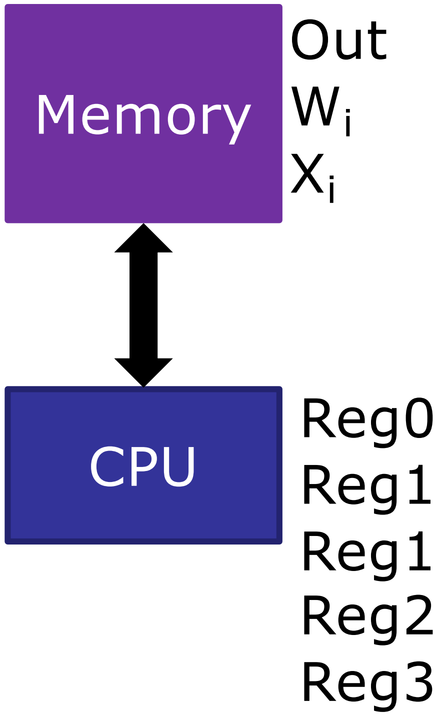
- Wire energy
  - 32 bits: 40pJ/word/mm
  - 8 bits: 10pJ/word/mm

- Register-File
  - Depends on word-length



Courtesy of O. Sentieys

Computer Architecture




```
for (I = 1 to N)
  out += Wi * Xi
```

assembly

nJ	`Load Wi, reg0`
nJ	`Load Xi,  reg1`
nJ	`Load out, reg2`
pJ	`Mul reg3, reg1, reg0`
pJ	`Add reg2, reg3, reg2`
nJ	`Store reg2, out`

6 memory accesses for instructions
4 memory accesses for data
2 operations

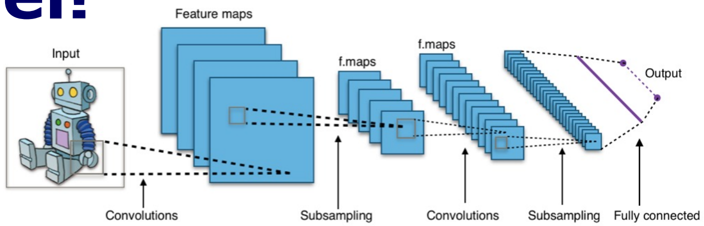
We can target any abstraction level!

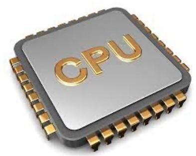
Software

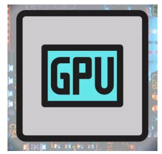
- Application software
- Virtual machines
- Operating system/Hypervisor
- Firmware

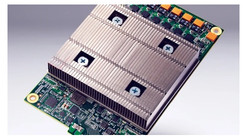
Hardware

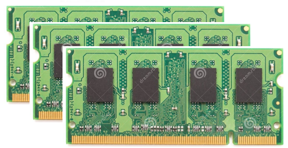
- Architecture
- Technology

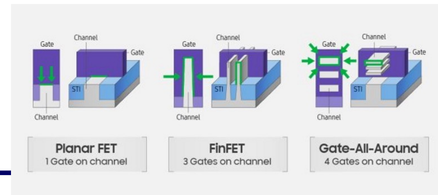




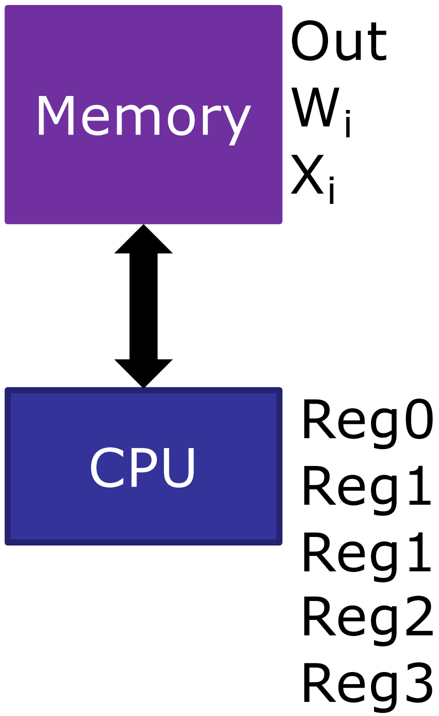








Software Level




```
for (I = 1 to N)
```

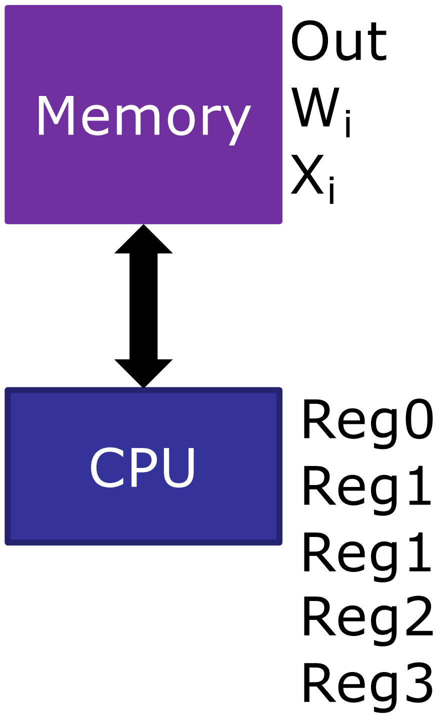
```
out += Wi * Xi
```

assembly

```
Load Wi, reg0
Load Xi,  reg1
Load out, reg2
Mul reg3, reg1, reg0
Add reg2, reg3, reg2
Store reg2, out
```

6 memory accesses for instructions
4 memory accesses for data
2 operations

Software Level



`for (I = 1 to N)`

`out += W_i * X_i`

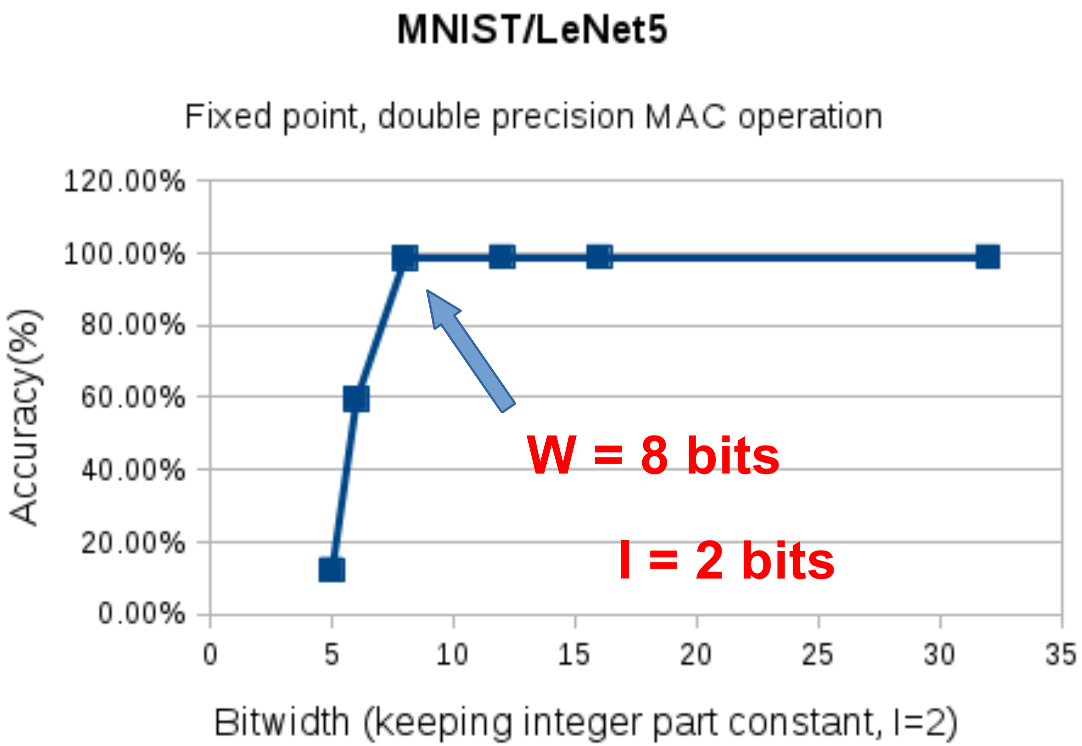
assembly

Load Wi, reg0
Load Xi, reg1
~~Load out, reg2~~
Mul reg3, reg1, reg0
Add reg2, reg3, reg2
~~Store reg2, out~~

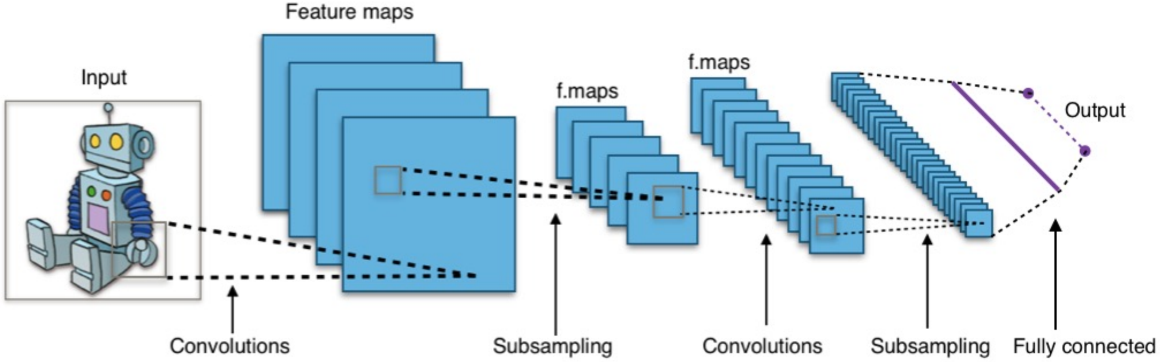
~~6~~ 4memory accesses for instructions
~~4~~ 2 memory accesses for data
2 operations

Quantization

- 10k images, MNIST/LeNet-5
- Fixed-Point Arithmetic



Weight Sharing



- 8 bits x W
  - 200 bits

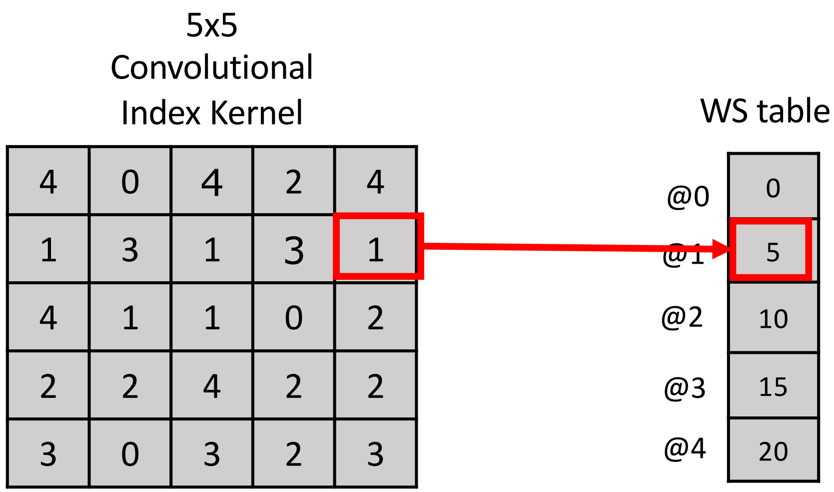
5x5 Convolutional Kernel

17	3	18	11	19
5	14	9	13	7
19	5	7	0	12
10	12	20	8	10
14	1	16	10	14

K-means Clustering

17	3	18	11	19
5	14	9	13	7
19	5	7	0	12
10	12	20	8	10
14	1	16	10	14

Weight Sharing



- 3 bits x W
  - 75 + 40 = 115 bits (instead of 200)
- ~42% bits reduction

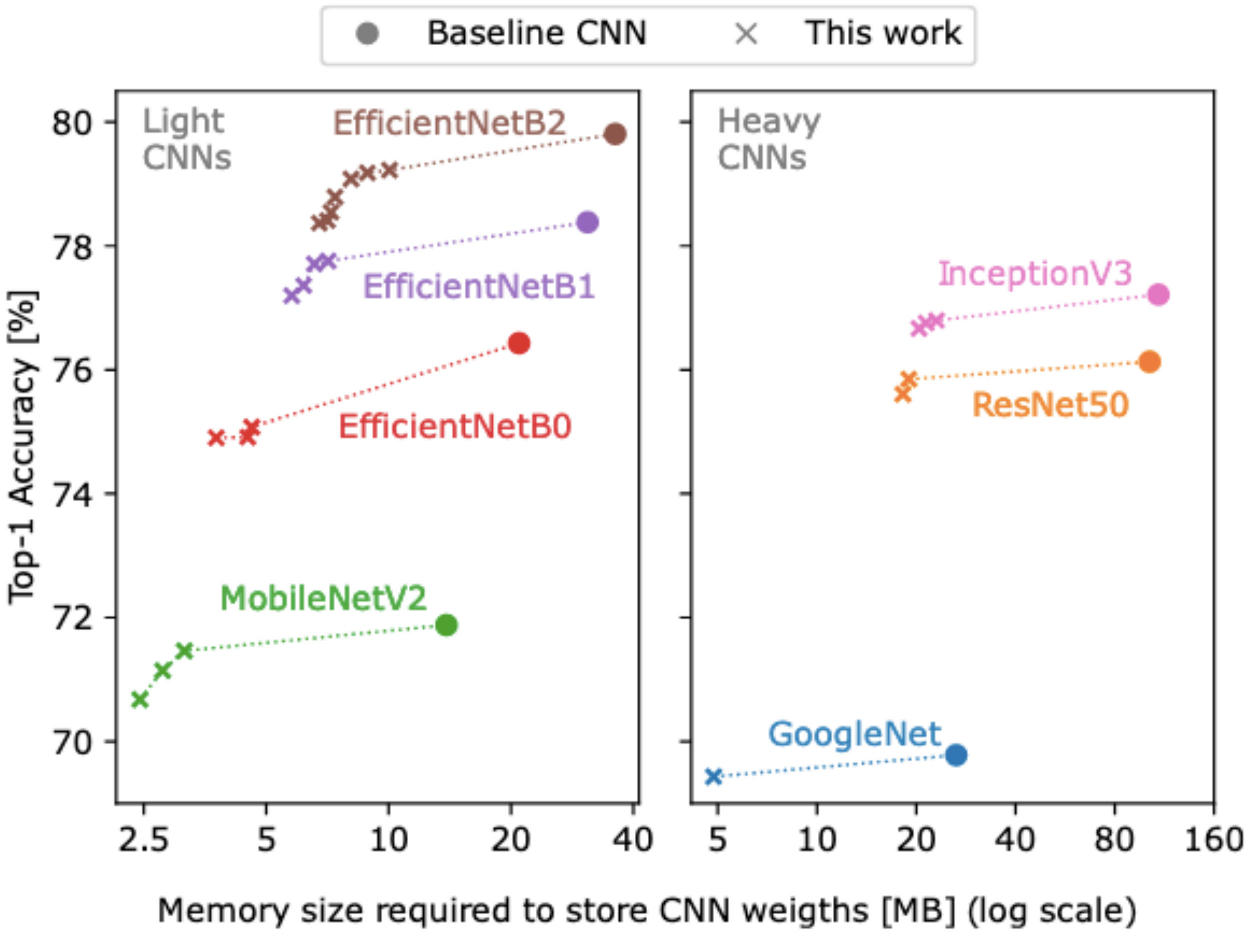
10.23919/DATE48585.2020.9116350

Results

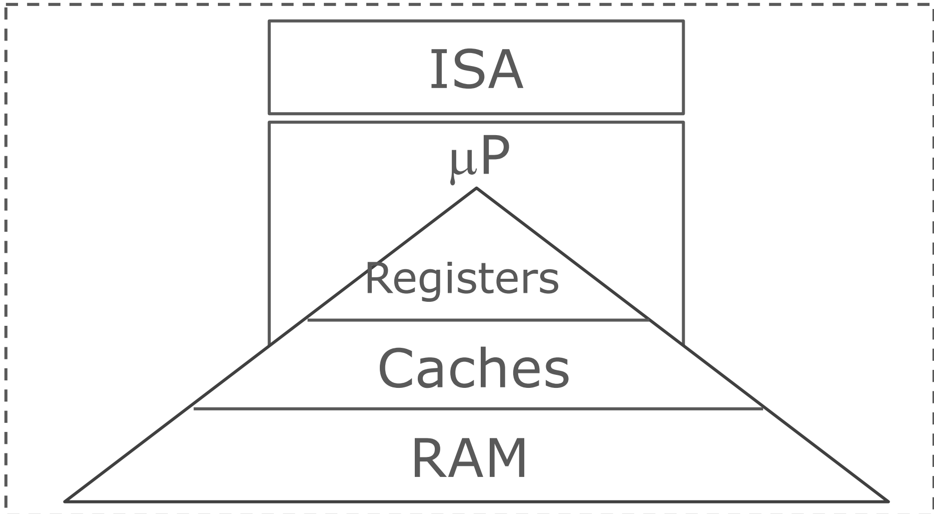
10.1109/ASP-DAC52403.2022.9712487

anr agence nationale de la recherche

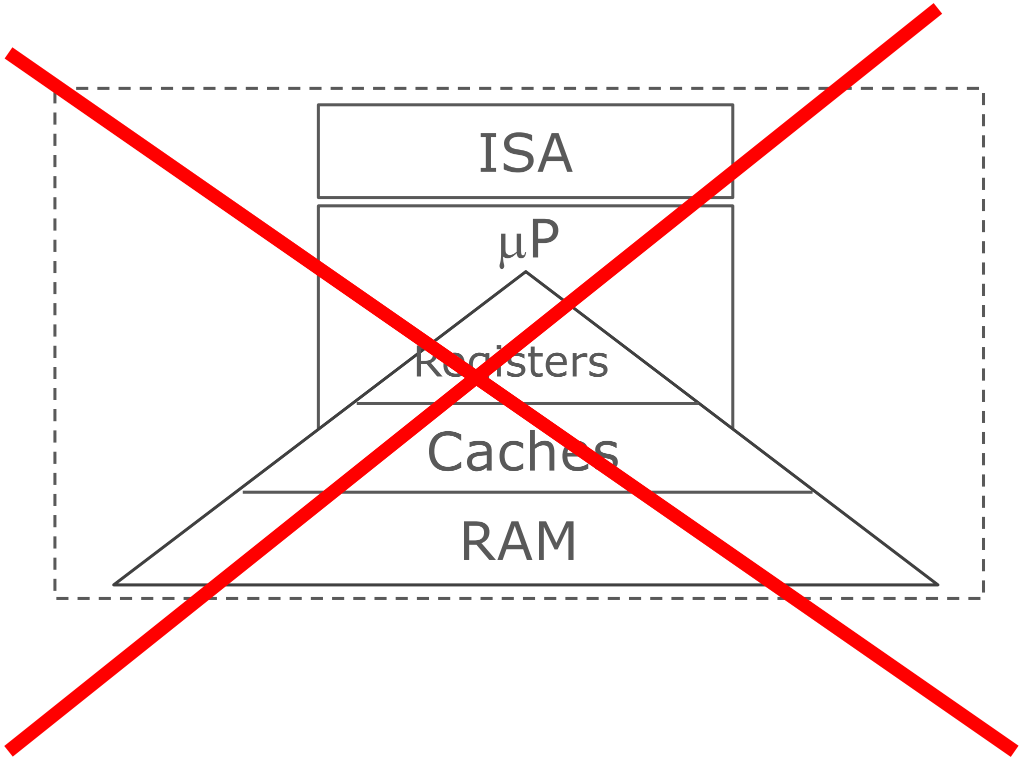
AdequetDL



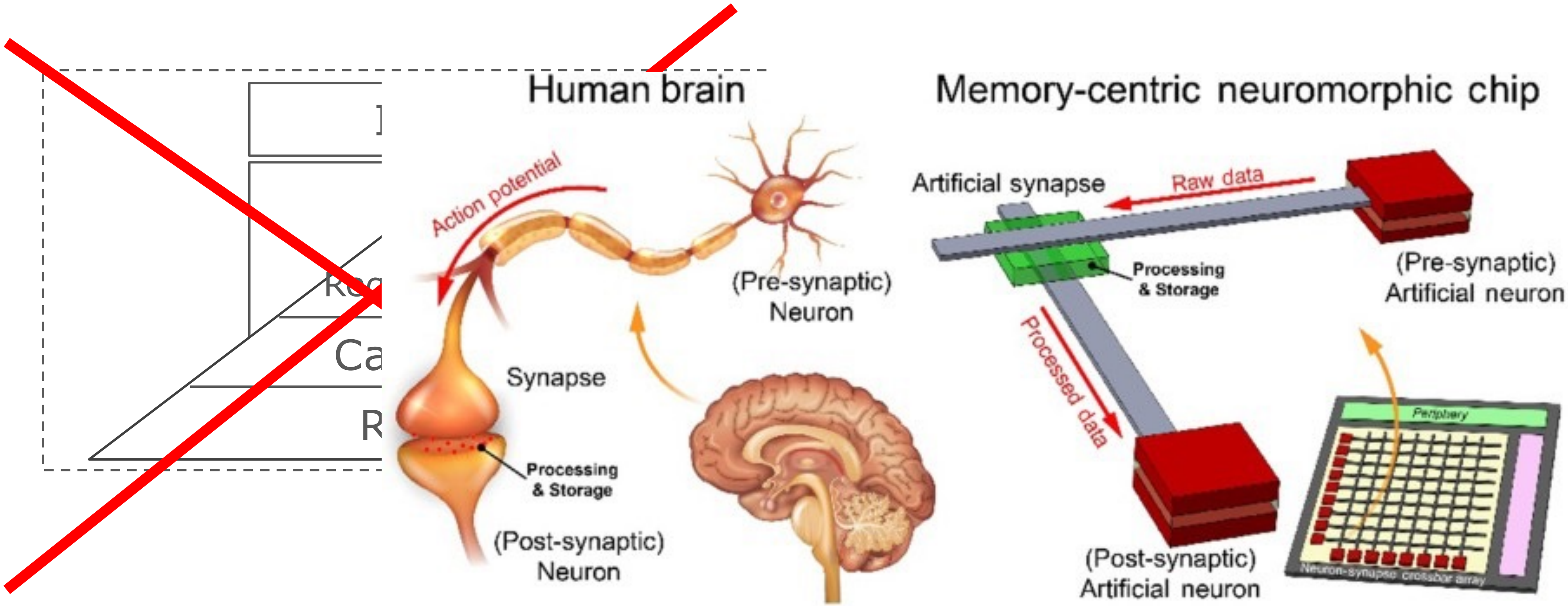
Hardware Level



Hardware Level

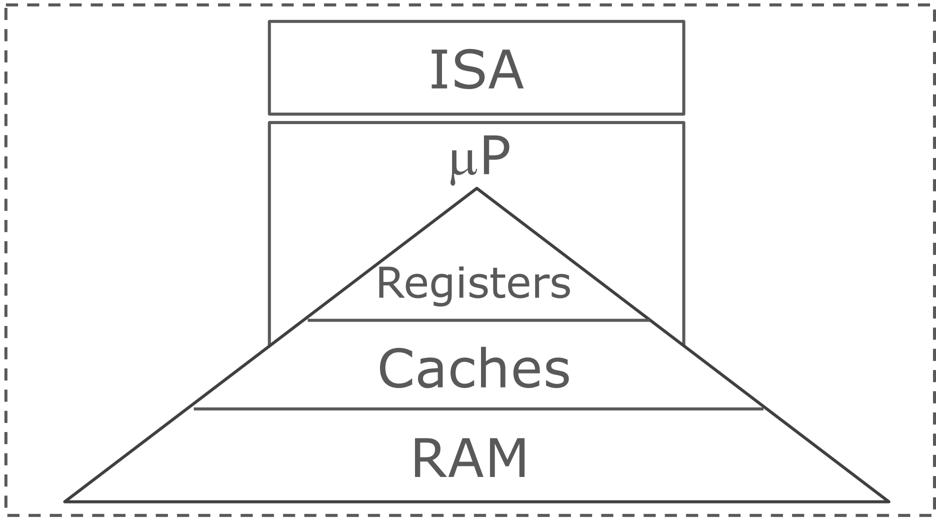


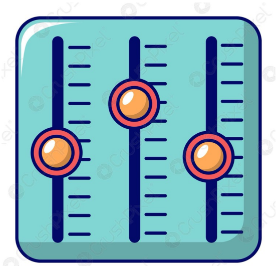
Hardware Level



https://link.springer.com/article/10.1007/s12274-021-3452-6

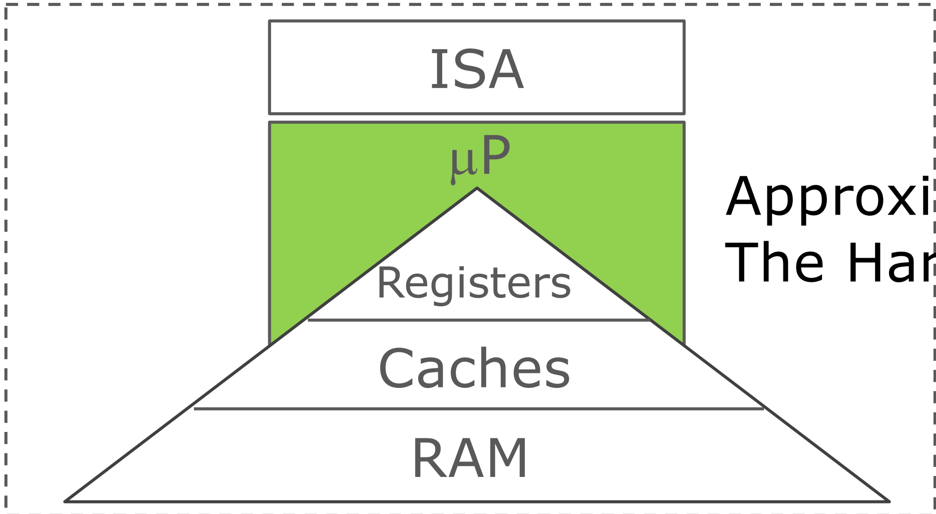
Hardware Level



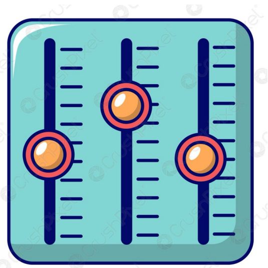


Tune Vdd

Hardware Level

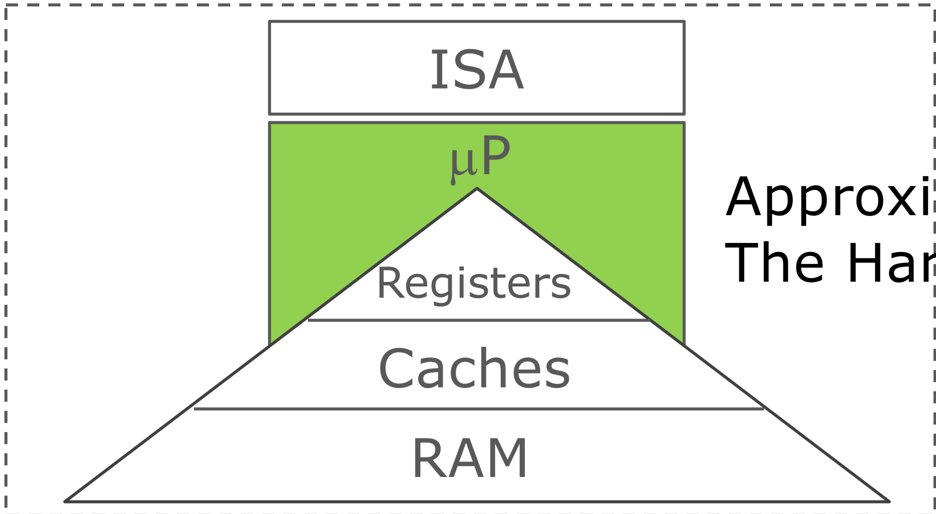


Approximate
The Hardware

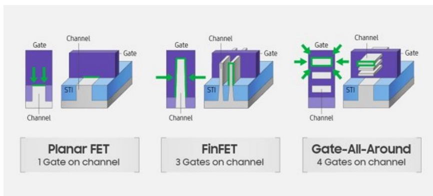


Tune Vdd

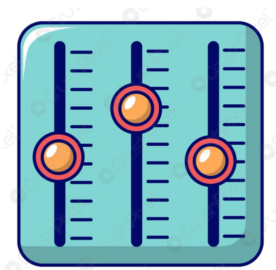
Hardware Level



Approximate
The Hardware



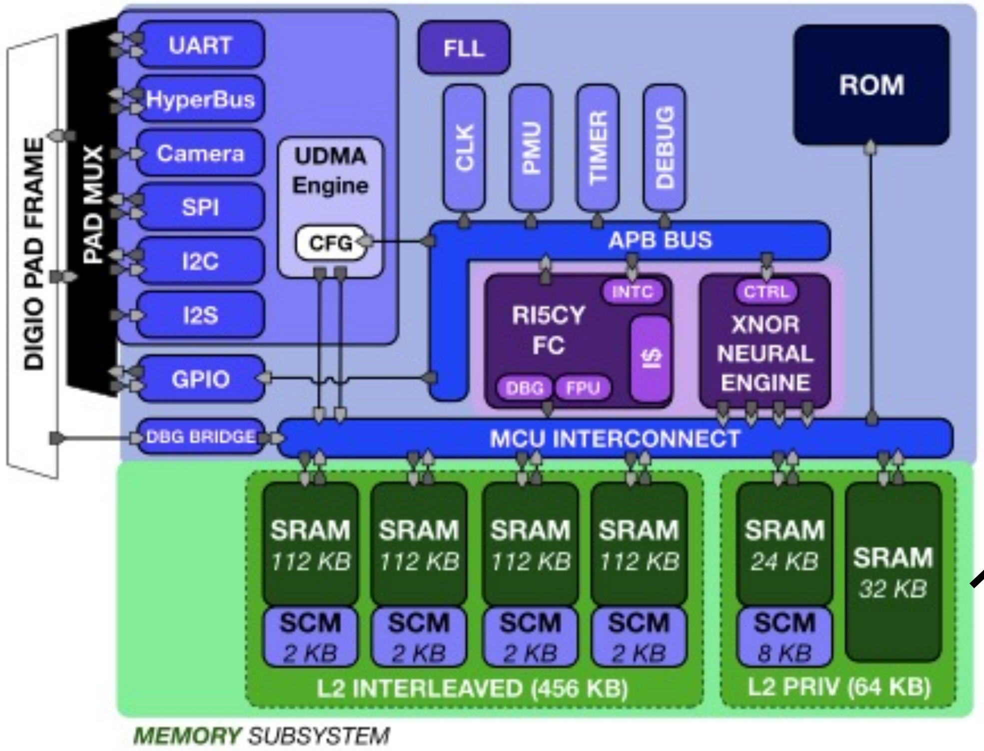
Improve the technology



Tune Vdd

Some examples: Over-Scaling

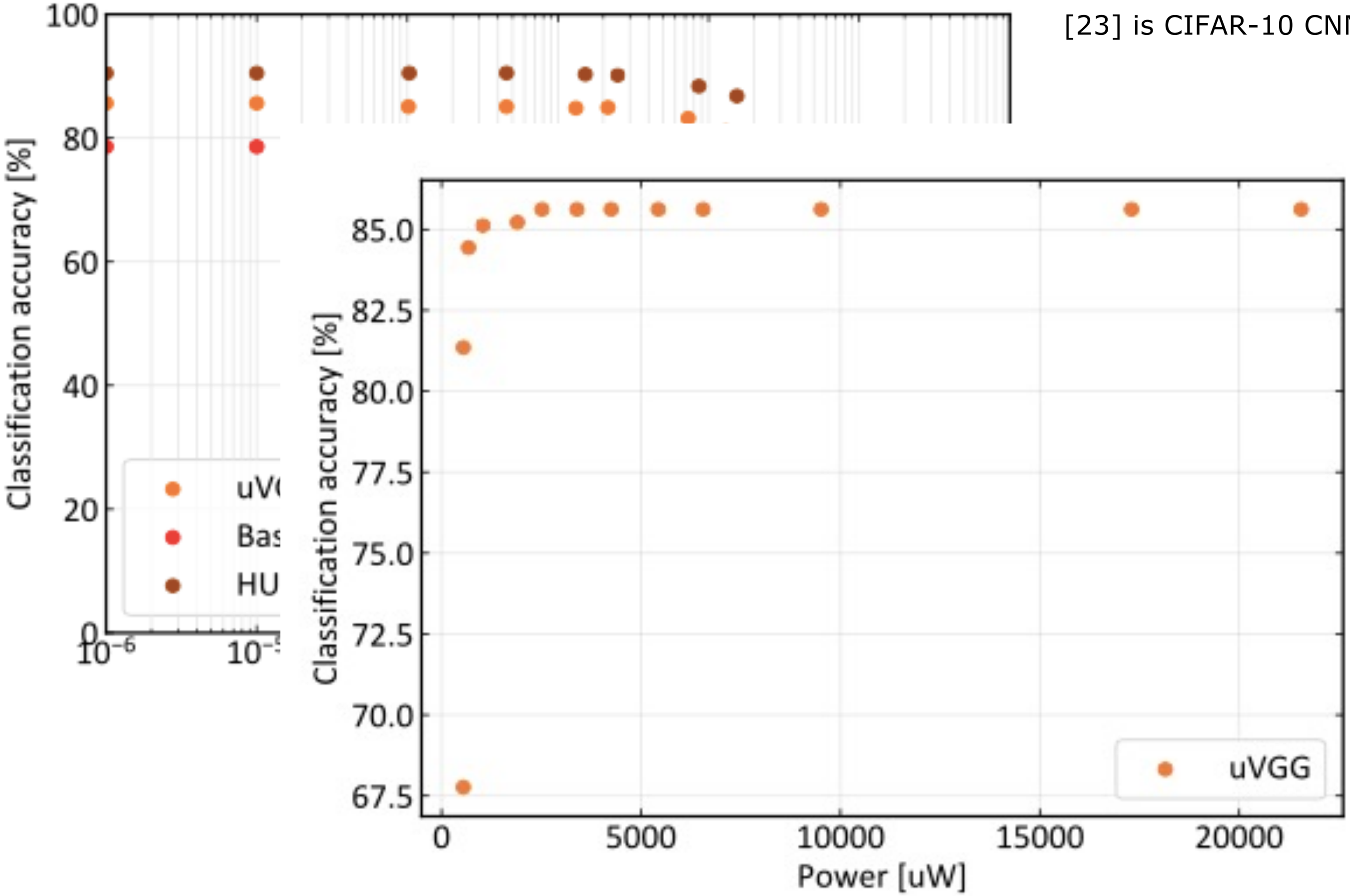
- Quentin SoC (based on PULPissimo system)



Over-Scaling:
reduce Vdd of SRAM

10.1109/TCSI.2020.3012576

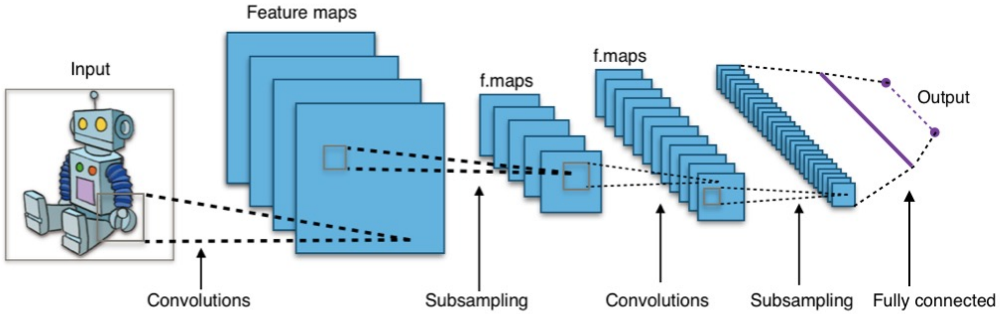
Results



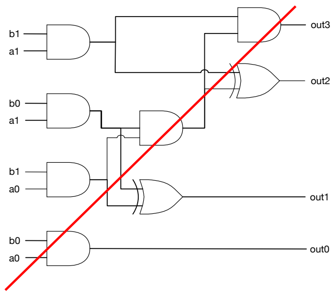
[23] is CIFAR-10 CNN

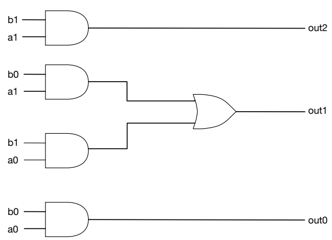
10.1109/TCSI.2020.3012576

Functional Approximation



Multiplier	ER	Accuracy (%)	
		MNIST	SVHN
Exact	0	97.69	86.93
mul8-350	99.0	97.70	**87.00**
mul8-439	97.8	97.71	86.96
mul8-120	98.5	97.70	**87.00**
mul8-183	97.2	97.70	86.98
mul8-134	93.9	**97.72**	86.95



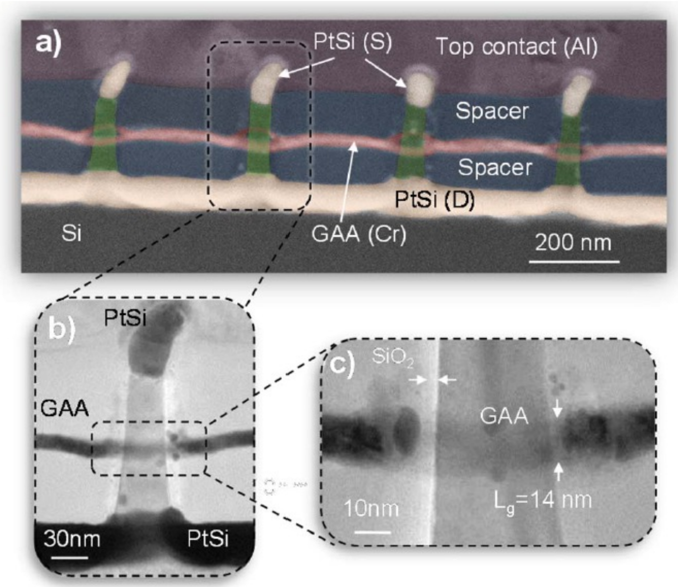


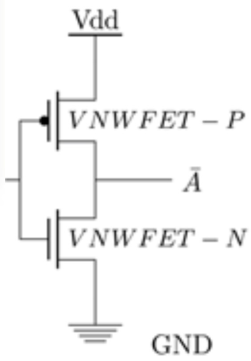
- Up to 71.45% more energy-efficient
- Up to 61.55% smaller

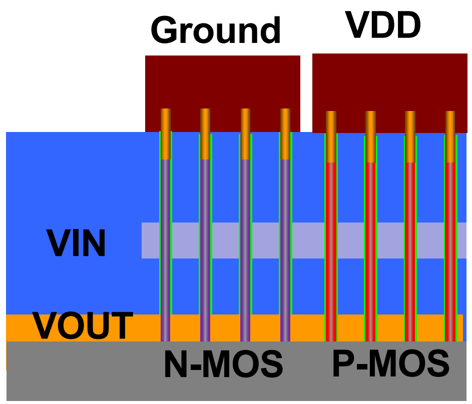
10.1109/TVLSI.2019.2940943

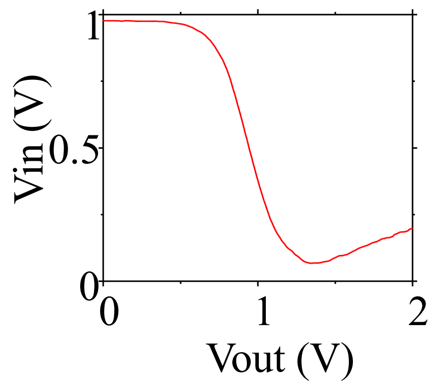
Technology

First hardware demo of 14 nm VGAA inverter



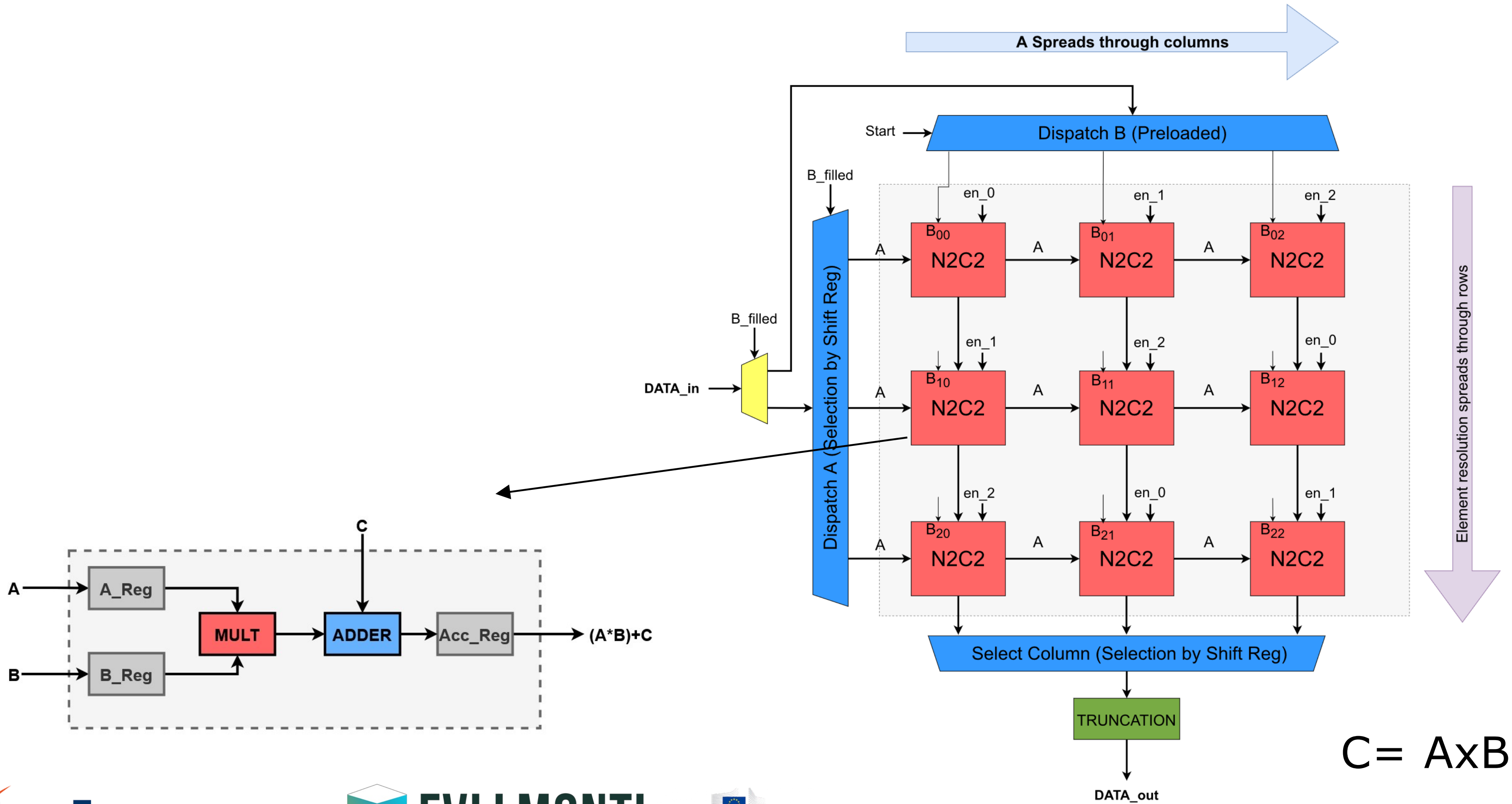






Y. Guerfi et al., Nanoscale research letters 11, 210 (2016)
G. Larrieu et al., Solid-State Electronics 130, 9 (2017)

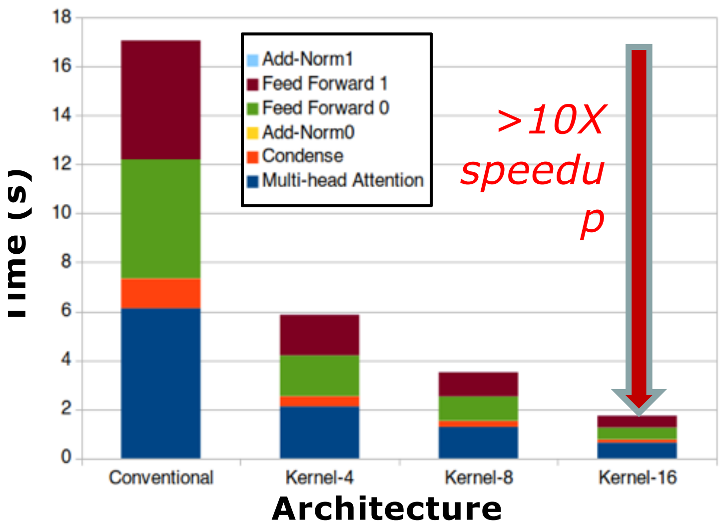
Systolic Array*

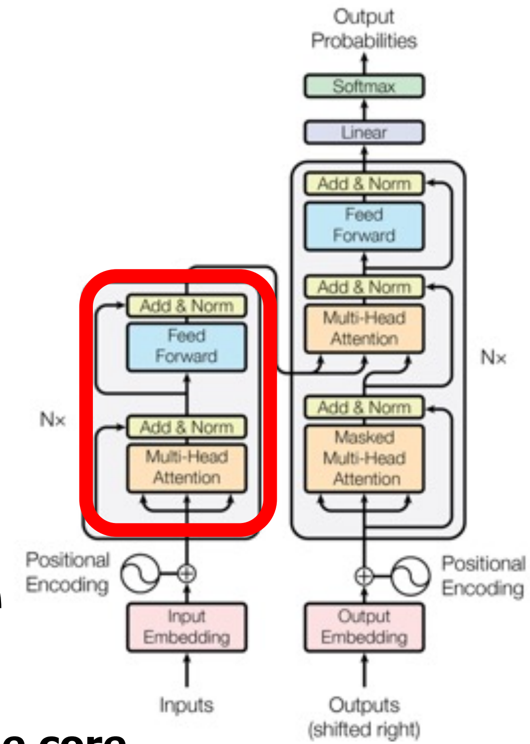


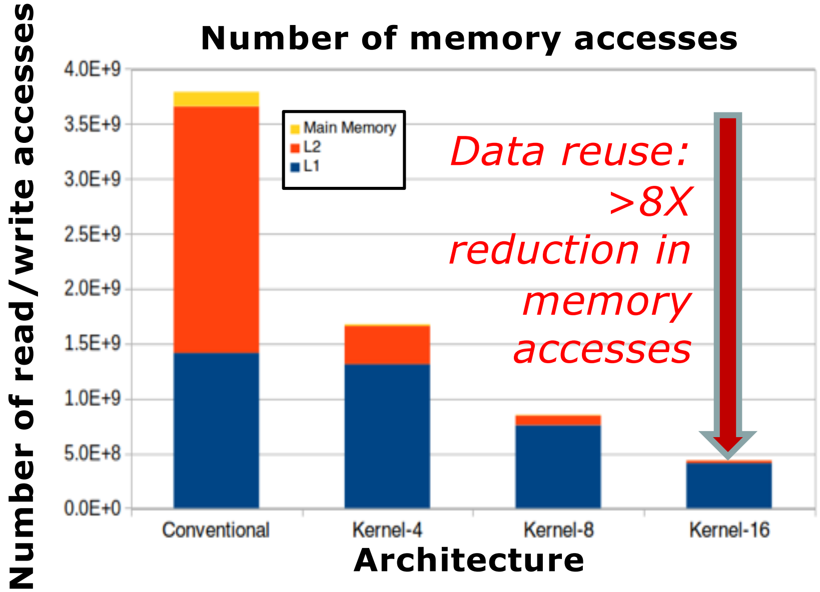
*in-house SA

Performances

Inference time of a transformer block





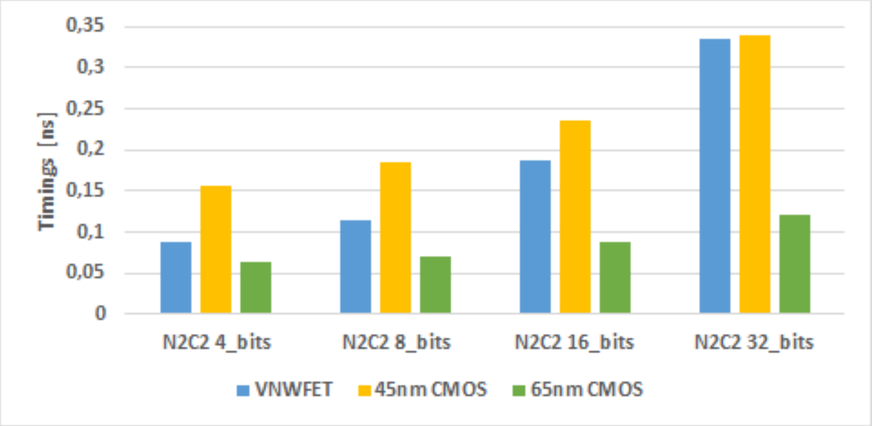


Kernel-N = NxN N^2C^2 accelerator

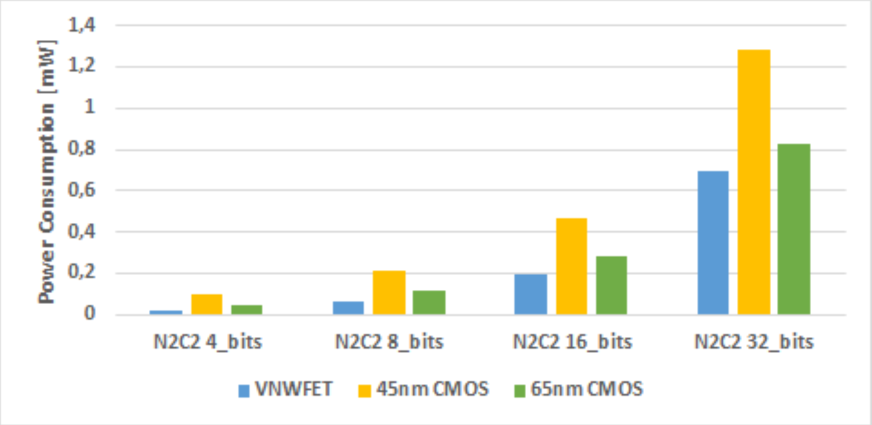
Conventional =
Non-accelerated ARM ISA - single core
Clock frequency: 1GHz - Memory type: DDR4_2400_4x16
Cache L1 size: 32KB - Cache L2 size: 1MB

Tech library and comparison

Library	INV	NAND	NOR	XOR
VNWFET	4	3	3	3
45nm CMOS	4	1	1	1
65nm CMOS	20	15	15	10



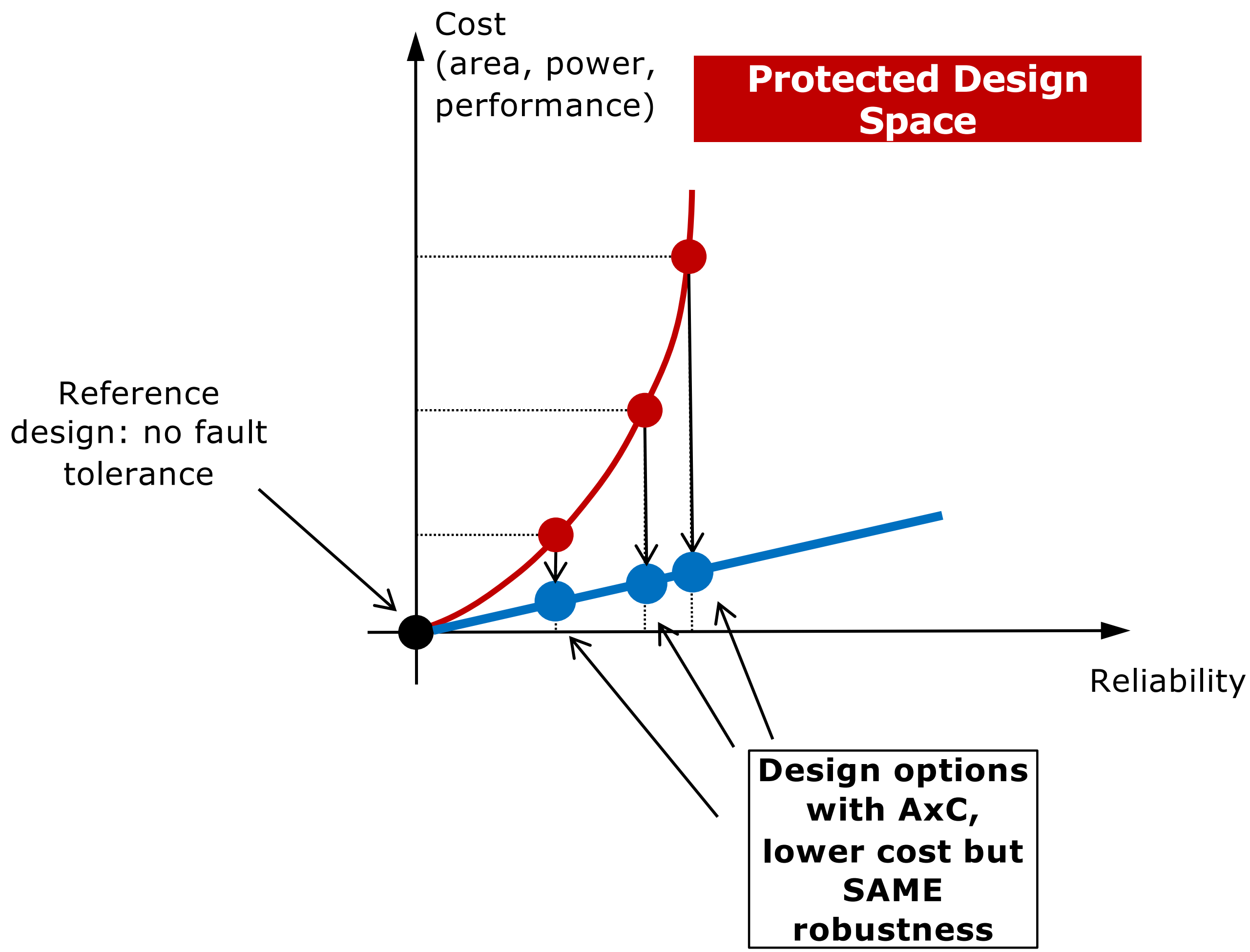
Circuit	NVWFET	45nm CMOS	65nm CMOS
4-bits PE	230	530	221
8-bits PE	737	1674	705
16-bits PE	2553	5697	2321
32-bits PE	9395	20608	8492



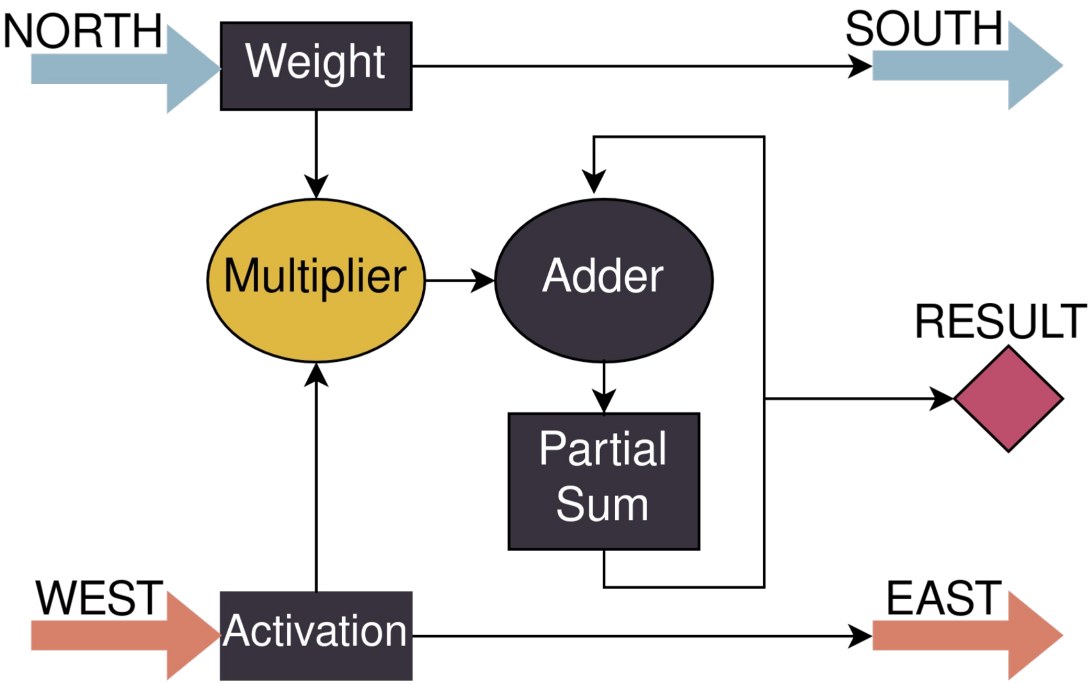
Outline

- Introduction
- Efficient HW accelerators
- **Reliable HW accelerators**
- Conclusions

Approximation is the key

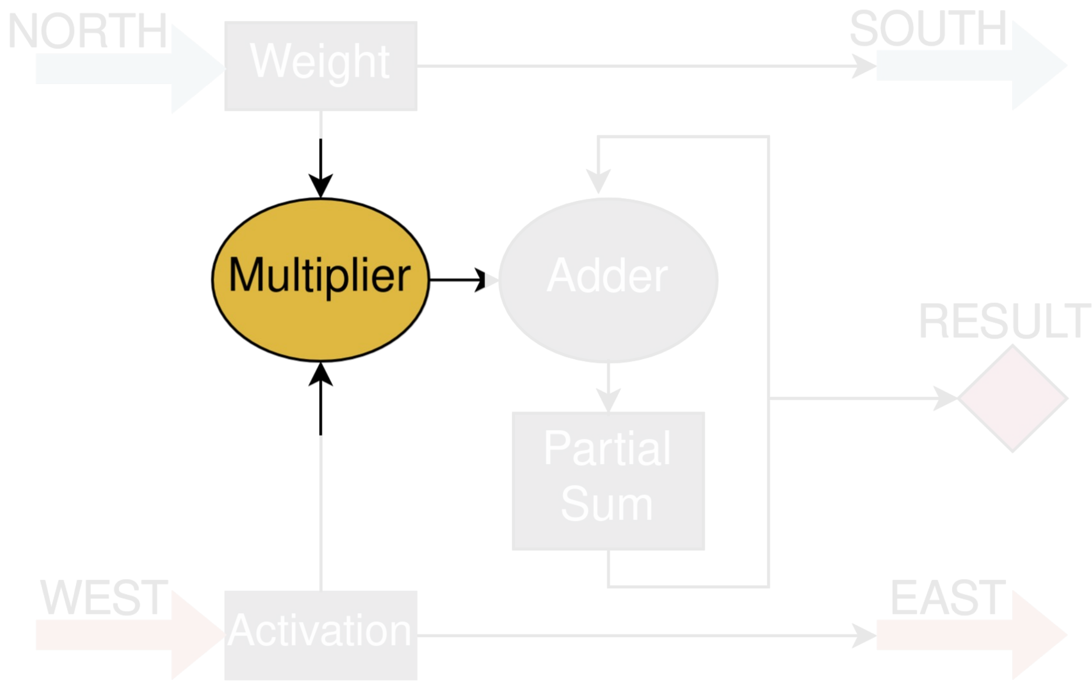


Approximate Systolic Array



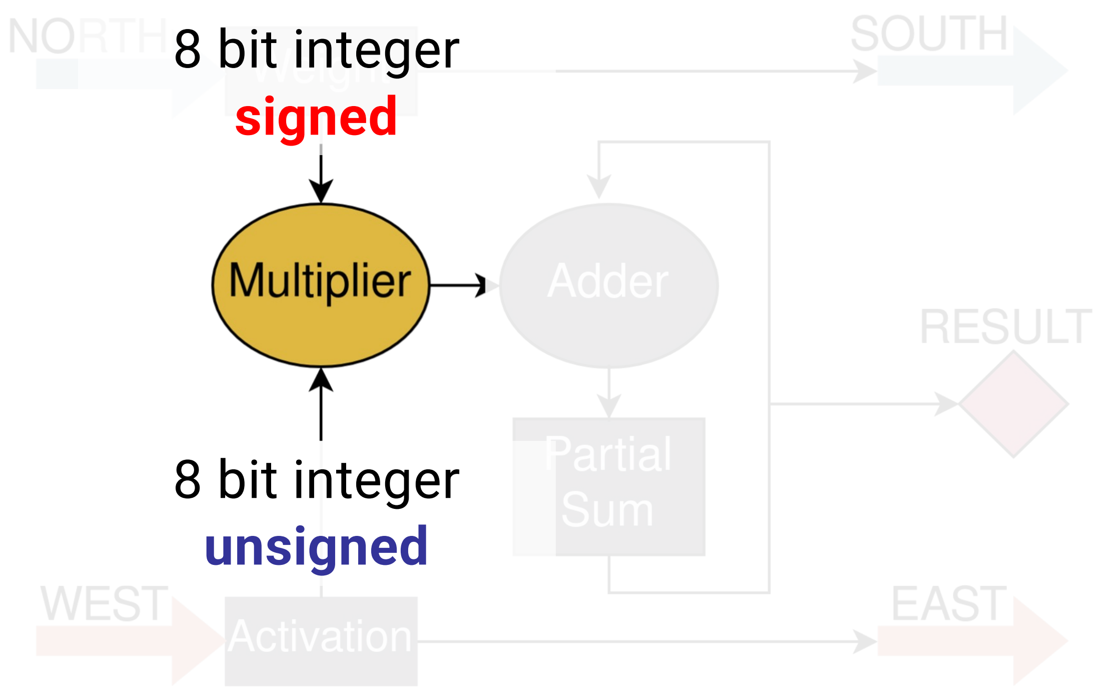
Processing Element Functional View

Approximate Systolic Array



Processing Element Functional View

Approximate Systolic Array



Processing Element Functional View

Used AxC Multipliers

Name	MAE	MAE-8
mul12s_2PT	0.000073	0.019
mul12s_2QH	0.0031	0.134
mul12s_2R5	0.0092	0.315
mul12s_34P	0.032	0.785
mul12s_2TE	0.19	6.080

MAE: Mean Absolute Error

https://ehw.fit.vutbr.cz/evoapproxlib/?folder=multiplers/12x12_signed

Results

#	Bit-width	Multiplier	Energy Reduction [%]	Accuracy [%]	Injected Faults [%]	Masked [%]	Tolerable [%]	Critical [%]
Baseline	16	Precise	–	99.07%	10%	47.58%	29,18%	23.24%
1	8	Precise	50%	99.05%	19%	64.65%	23.01%	12.34%
2	8	mul12s_2PT	50.3%	99.08%	19%	63.9%	23.79%	12.3%
3	8	mul12s_2QH	51.21%	99.1%	19%	38.92%	44.85%	16.23%
4	8	mul12s_2R5	52%	99.06%	19%	26.96%	55.69%	17.34%
5	8	mul12s_34P	55%	**98.24%**	19%	**74.16%**	**23.01%**	**2.83%**
6	8	mul12s_2TE	55.6%	9.8%	19%	3.94%	27.5%	68.76%

anr agence nationale de la recherche

RE-TRUSTING

Conclusions & Future works

- We need a holistic approach to achieve a HW-SW co-design methology to design **sober** and **reliable** AI applications
- How to reach this goal?

Still a lot of work to reach a sober system

We have to build a novel flow