GB-RADNEXT

June 12, 2024
Rutherford Appleton Lab.

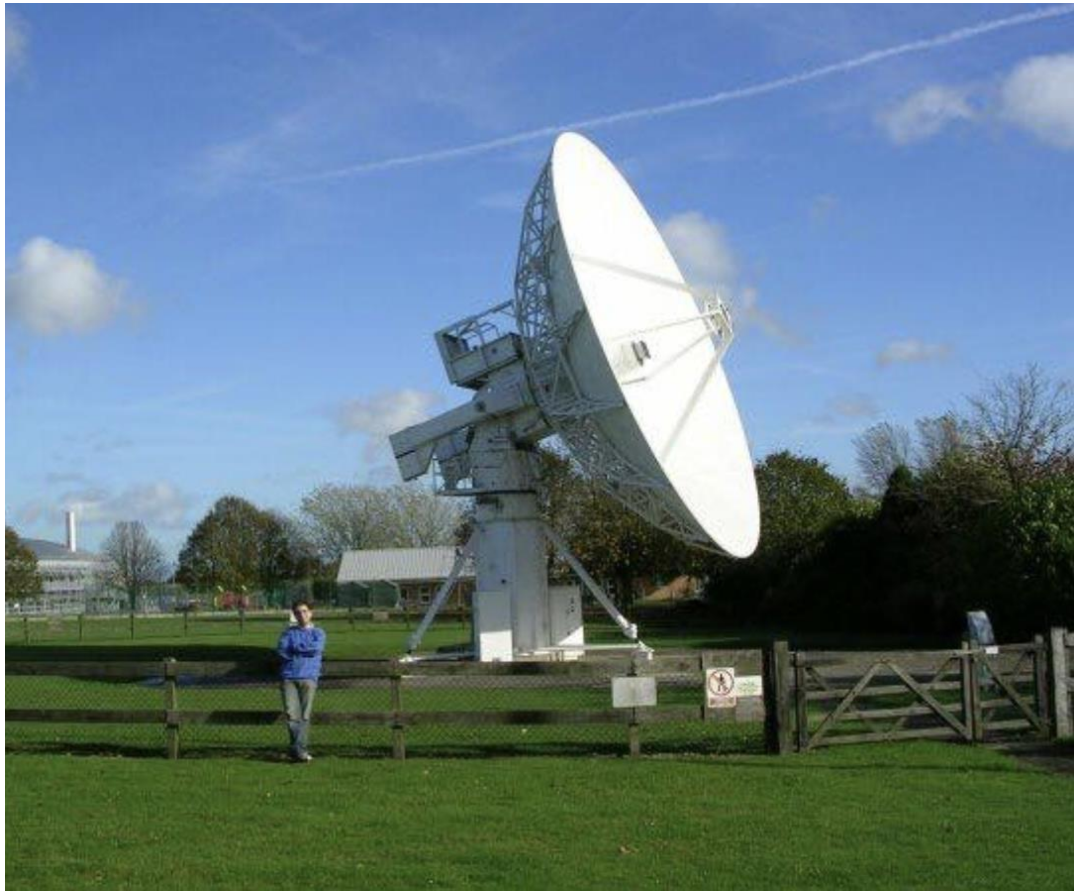
Radiation Experiments on AI accelerators: Current and Future Challenges and Opportunities

Paolo Rech
paolo.rech@unitn.it

UNIVERSITÀ DI TRENTO
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

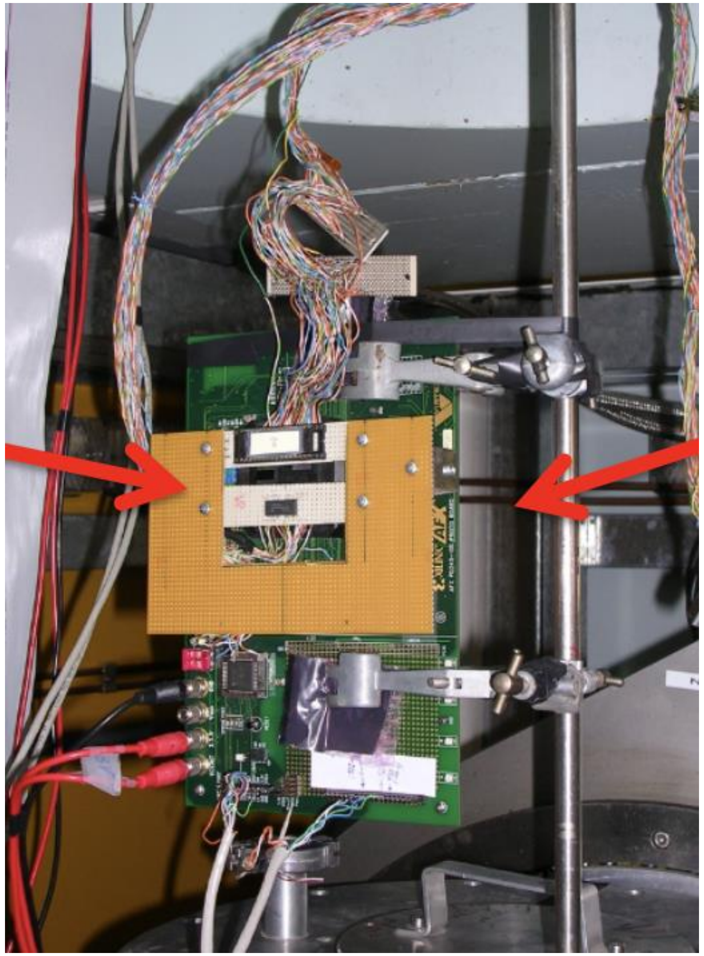
Testing at ISIS

2007



Vesuvio

SRAM

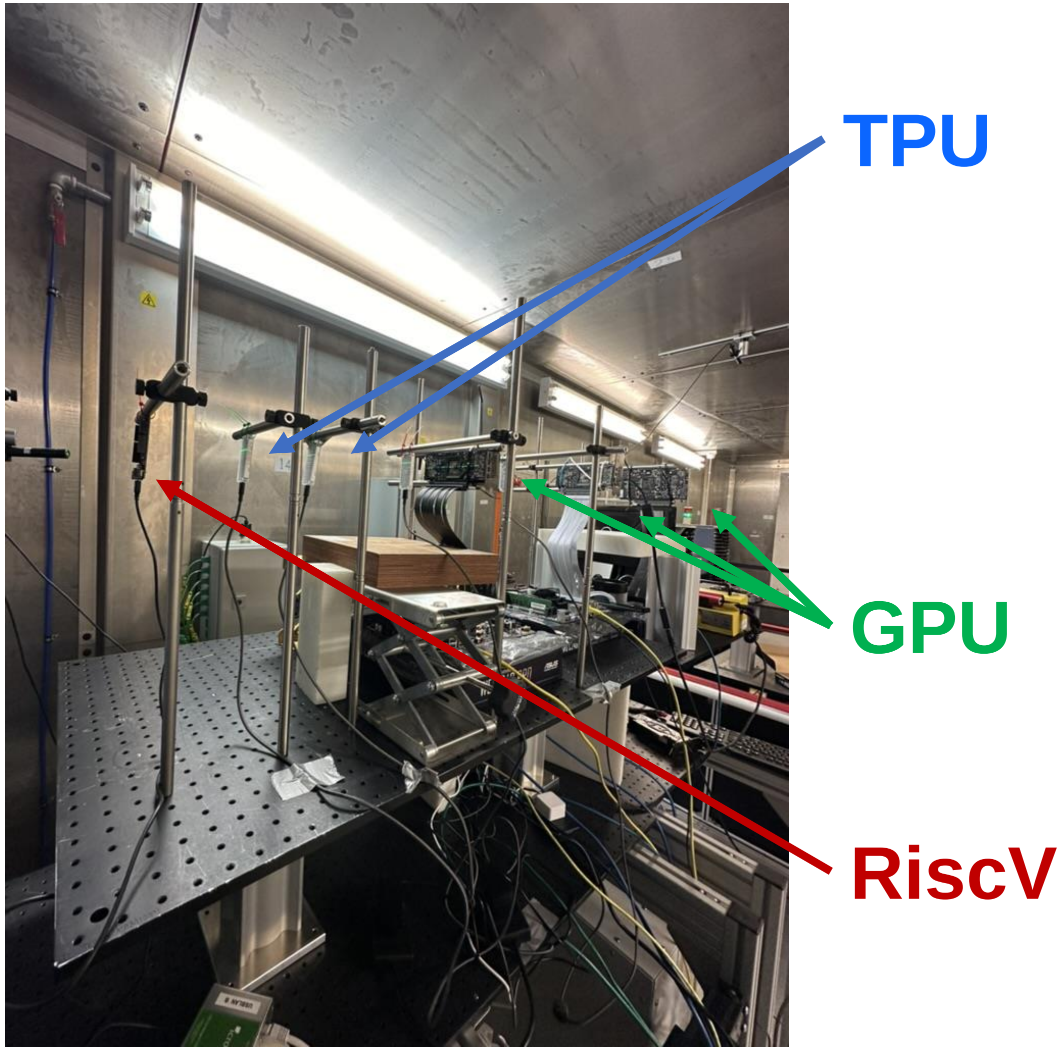


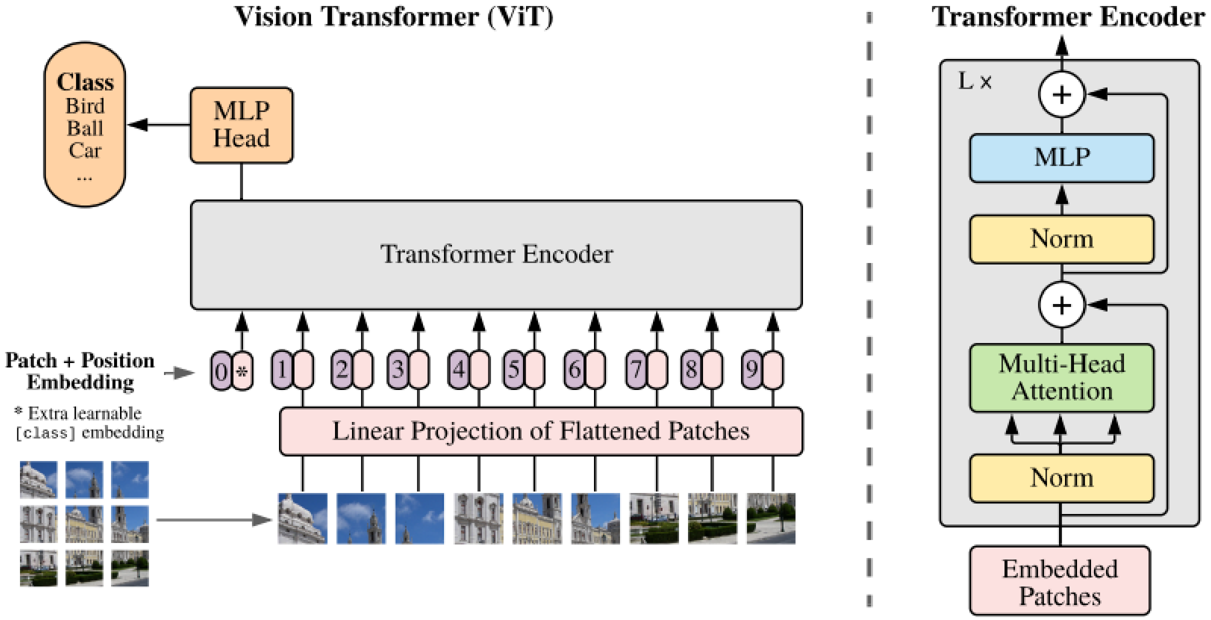
FPGA

Testing at ISIS

HiCREST UNIVERSITÀ DI TRENTO

May 2024 @ChipIR





Outline

- Introduction to AI reliability
- HW and SW for AI
- Experiments challenges
- What do we need for the (near) future?
- Conclusions and future perspective

AI Reliability

CNNs identify objects in a scene

AI Reliability

CNNs identify objects in a scene

Identification is probabilistic

AI Reliability

HiCREST UNIVERSITÀ DI TRENTO

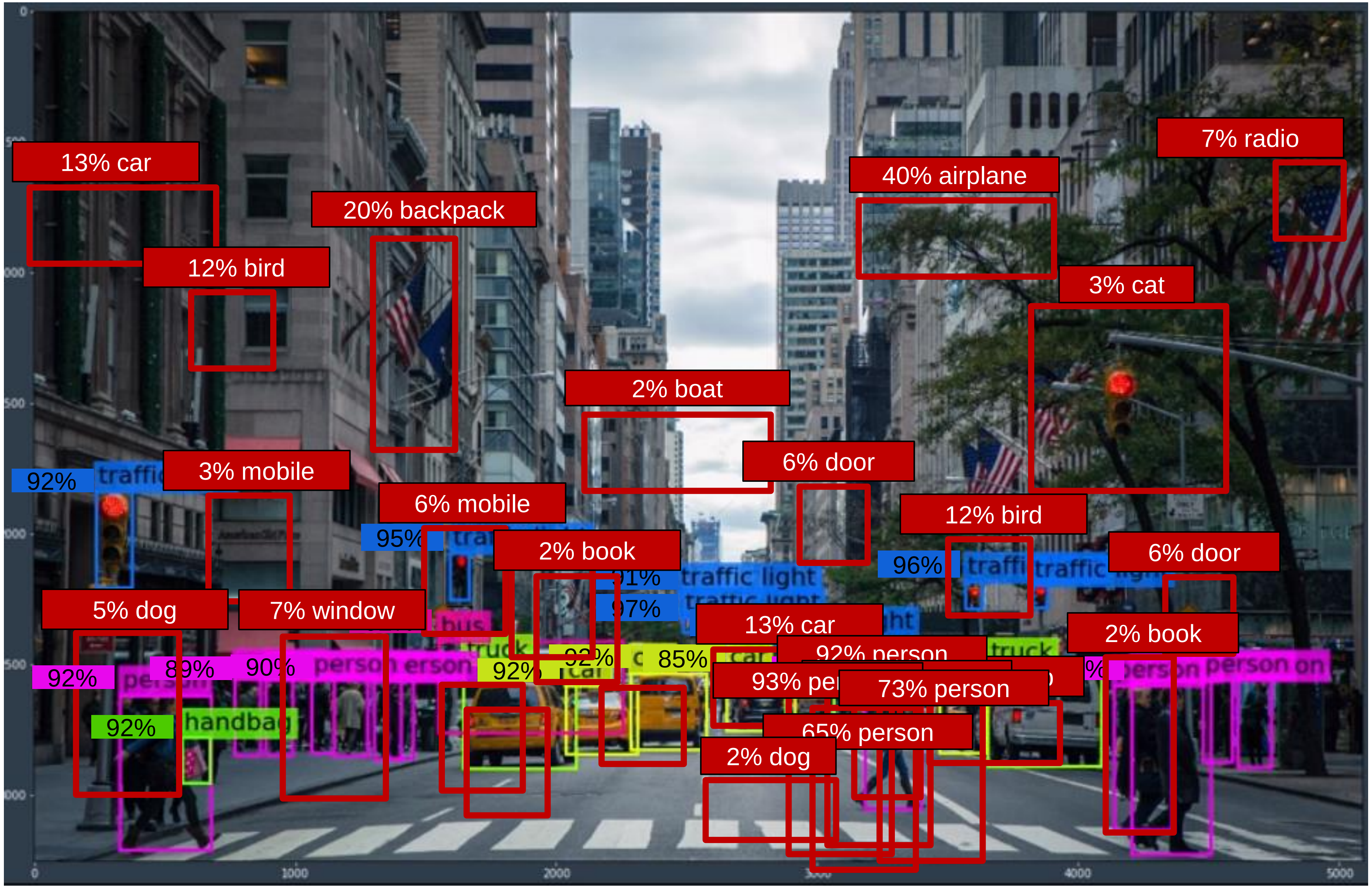
CNNs identify objects in a scene

Identification is probabilistic

Many objects with low probability are identified

AI Reliability

HiCREST UNIVERSITÀ DI TRENTO



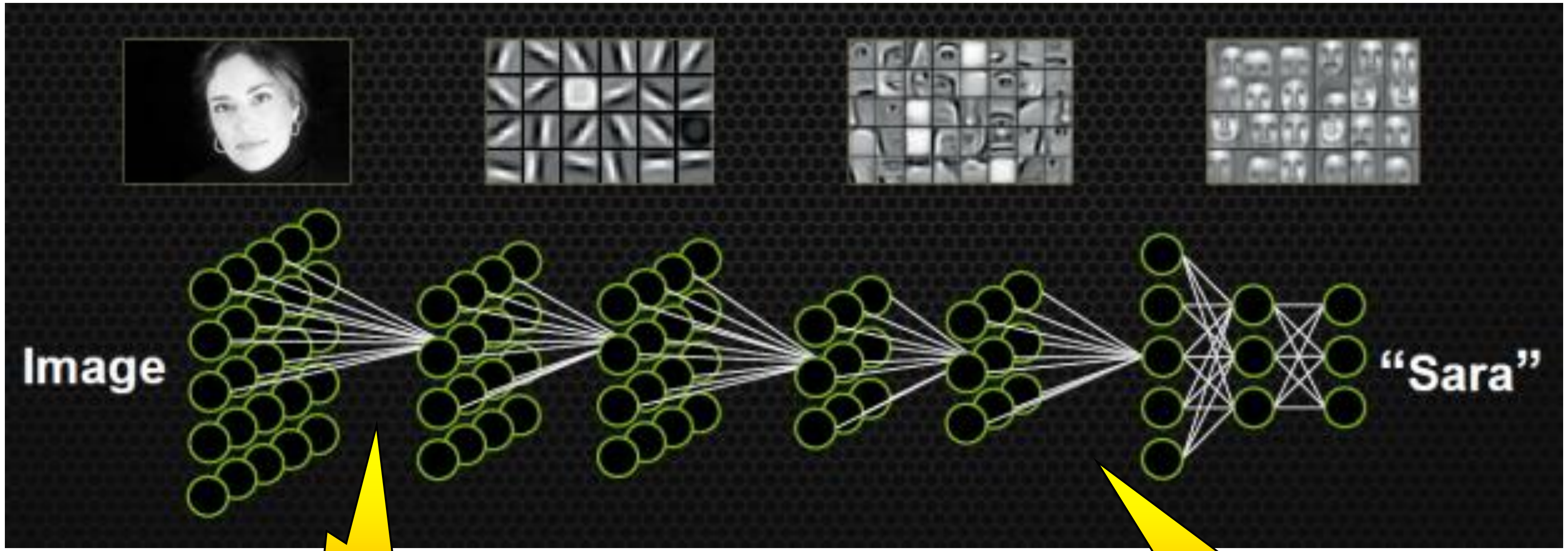
CNNs identify objects in a scene

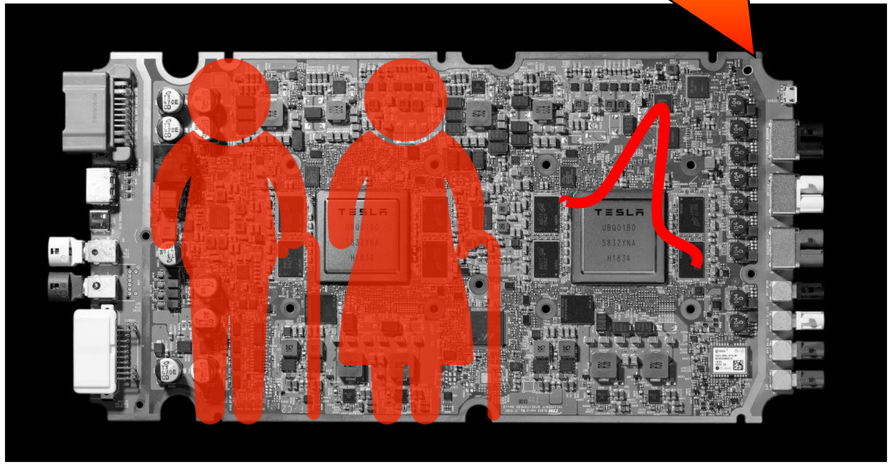
Identification is probabilistic

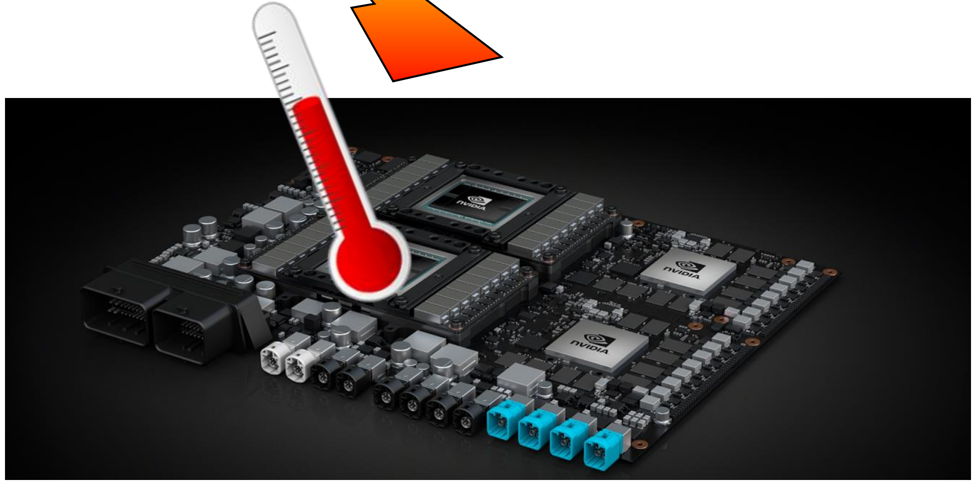
Many objects with low probability are identified

We need to select a detection threshold

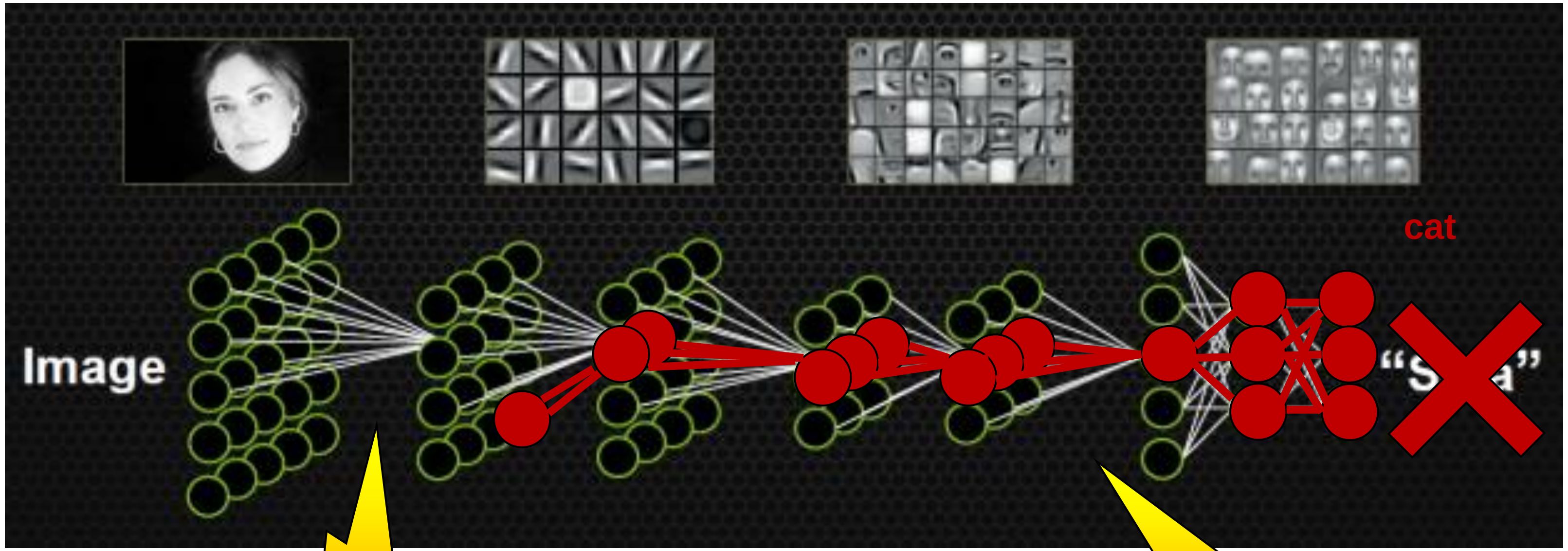
What about the Hardware?

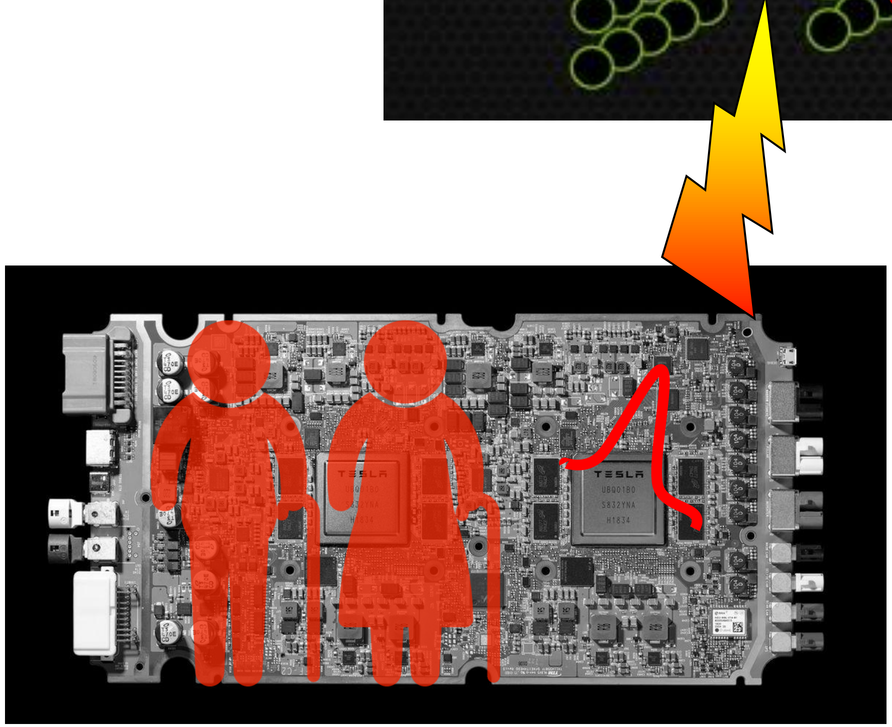


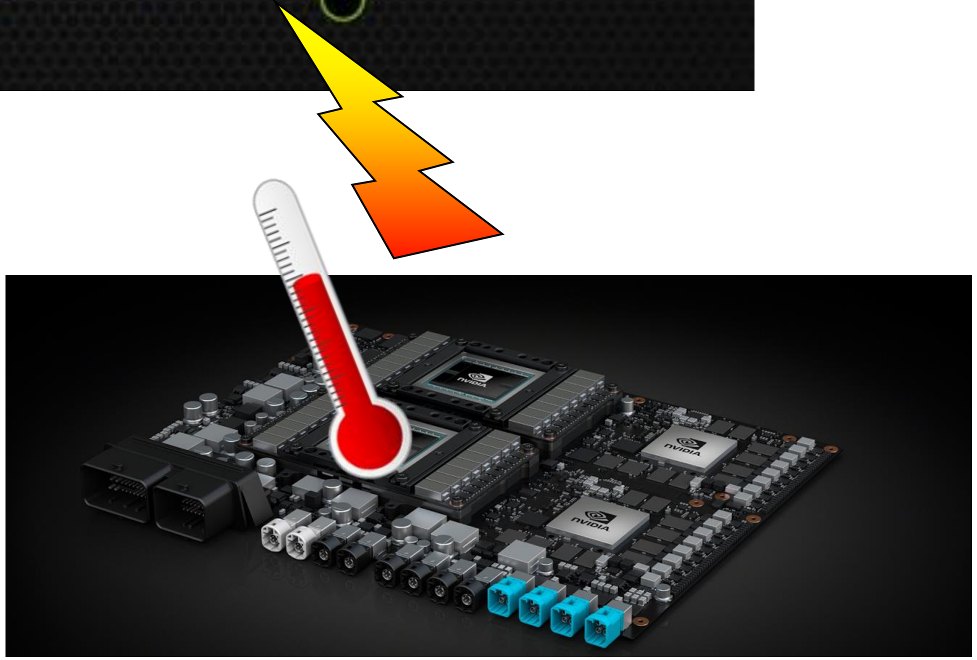




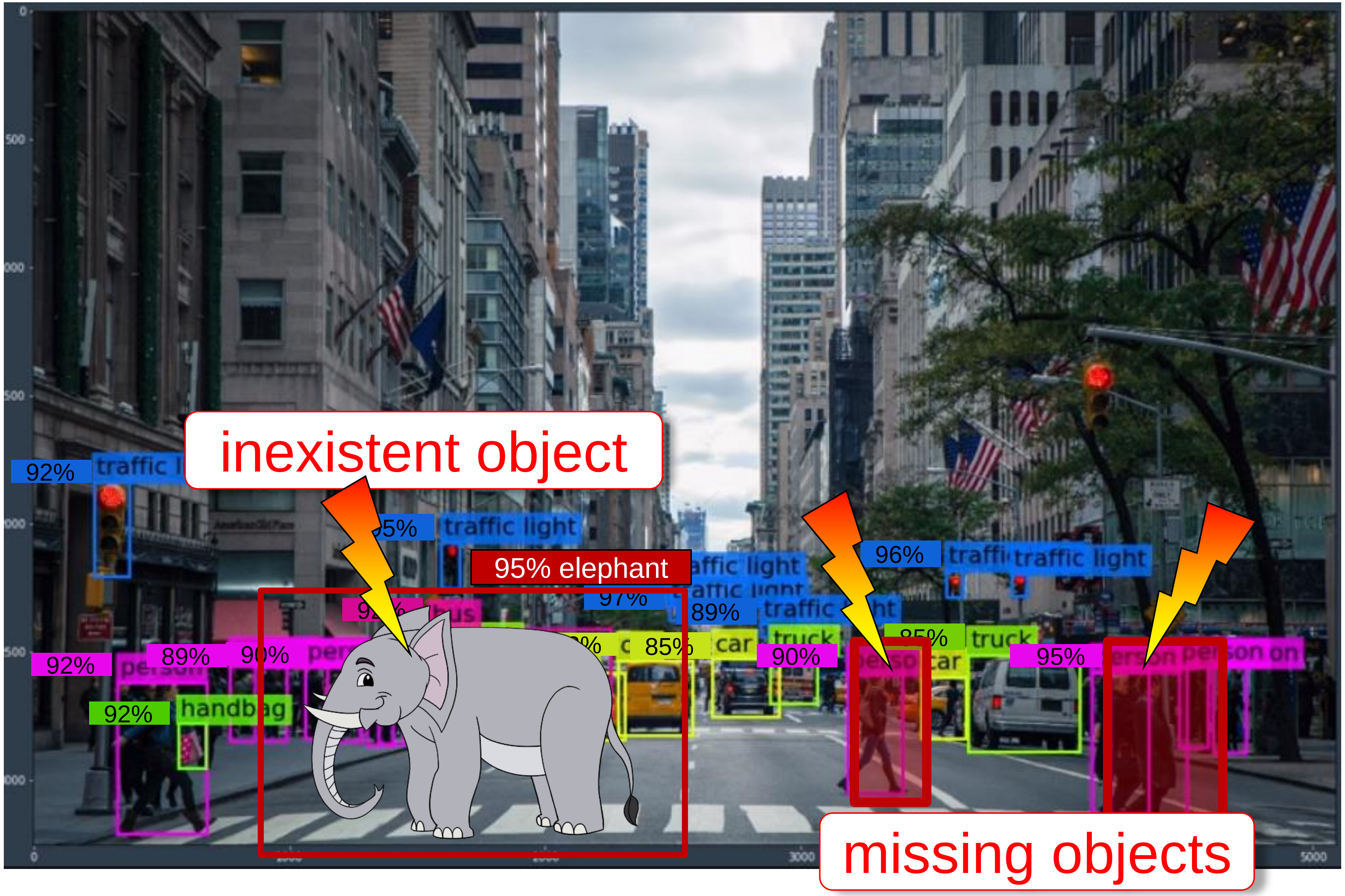
What about the Hardware?



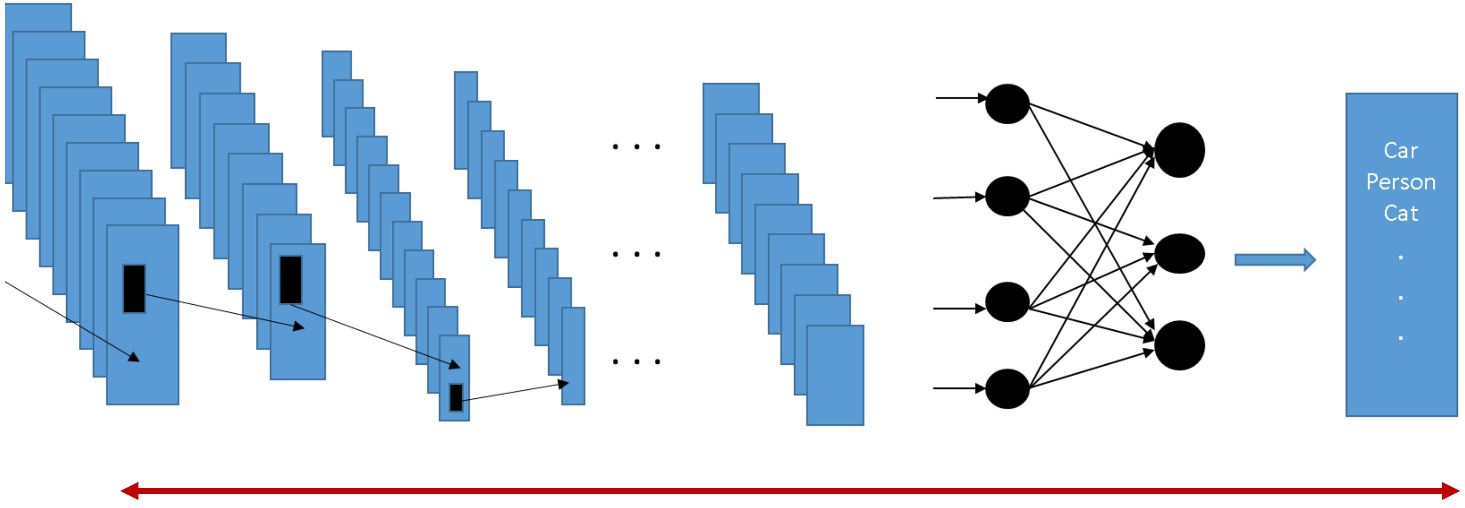




What about the Hardware?

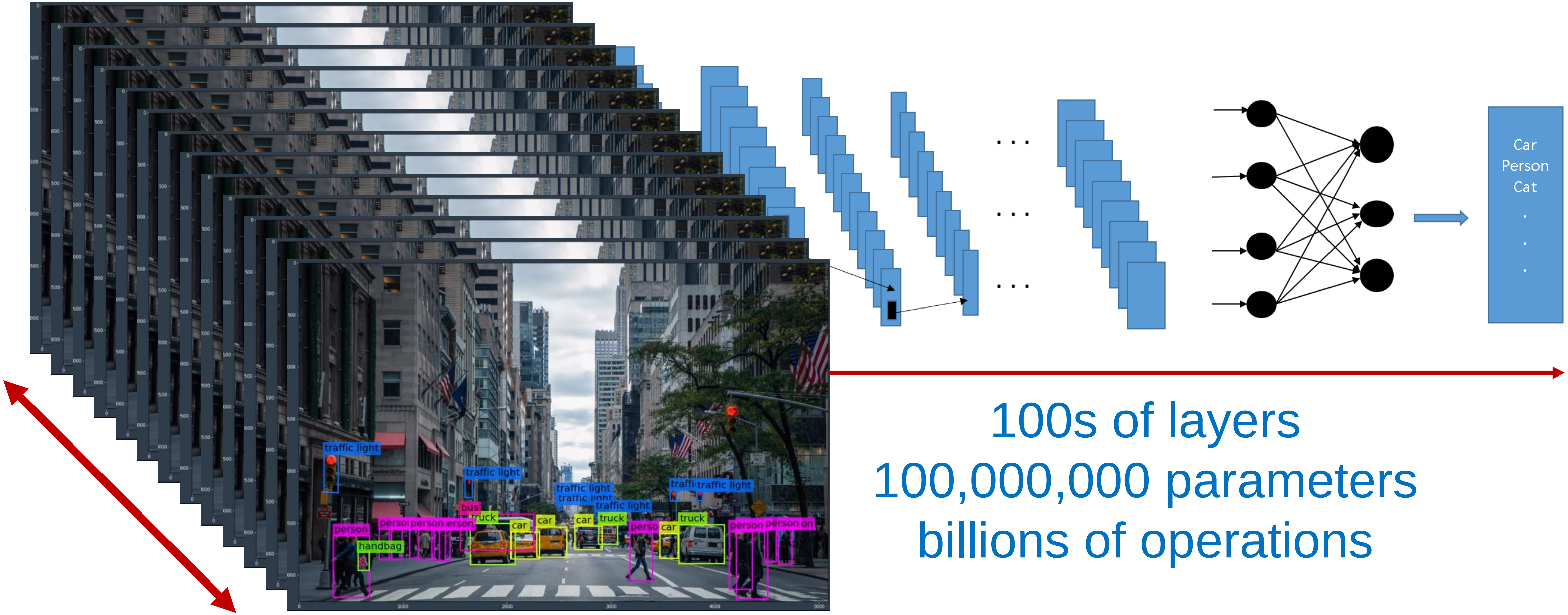


HW complexity



100s of layers
100,000,000 parameters
billions of operations

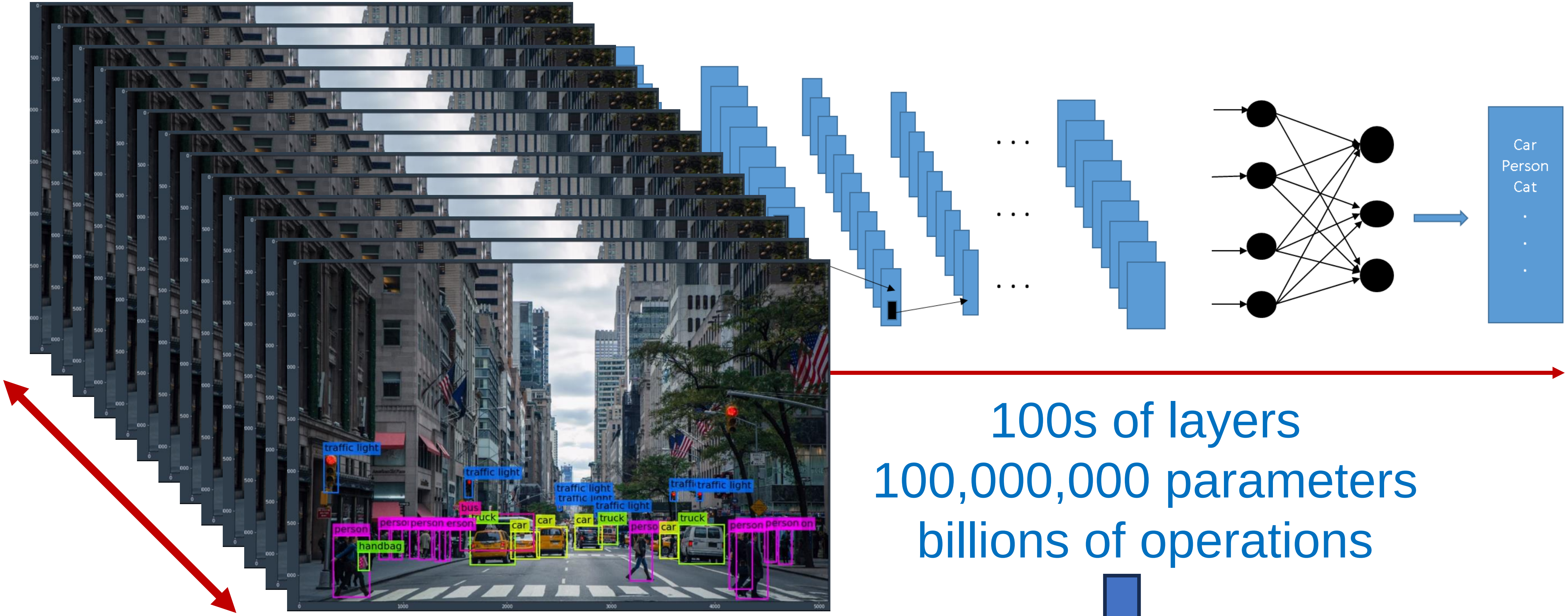
HW complexity



100s of layers
100,000,000 parameters
billions of operations

detection in real-time:
at least 40 frames per second

HW complexity

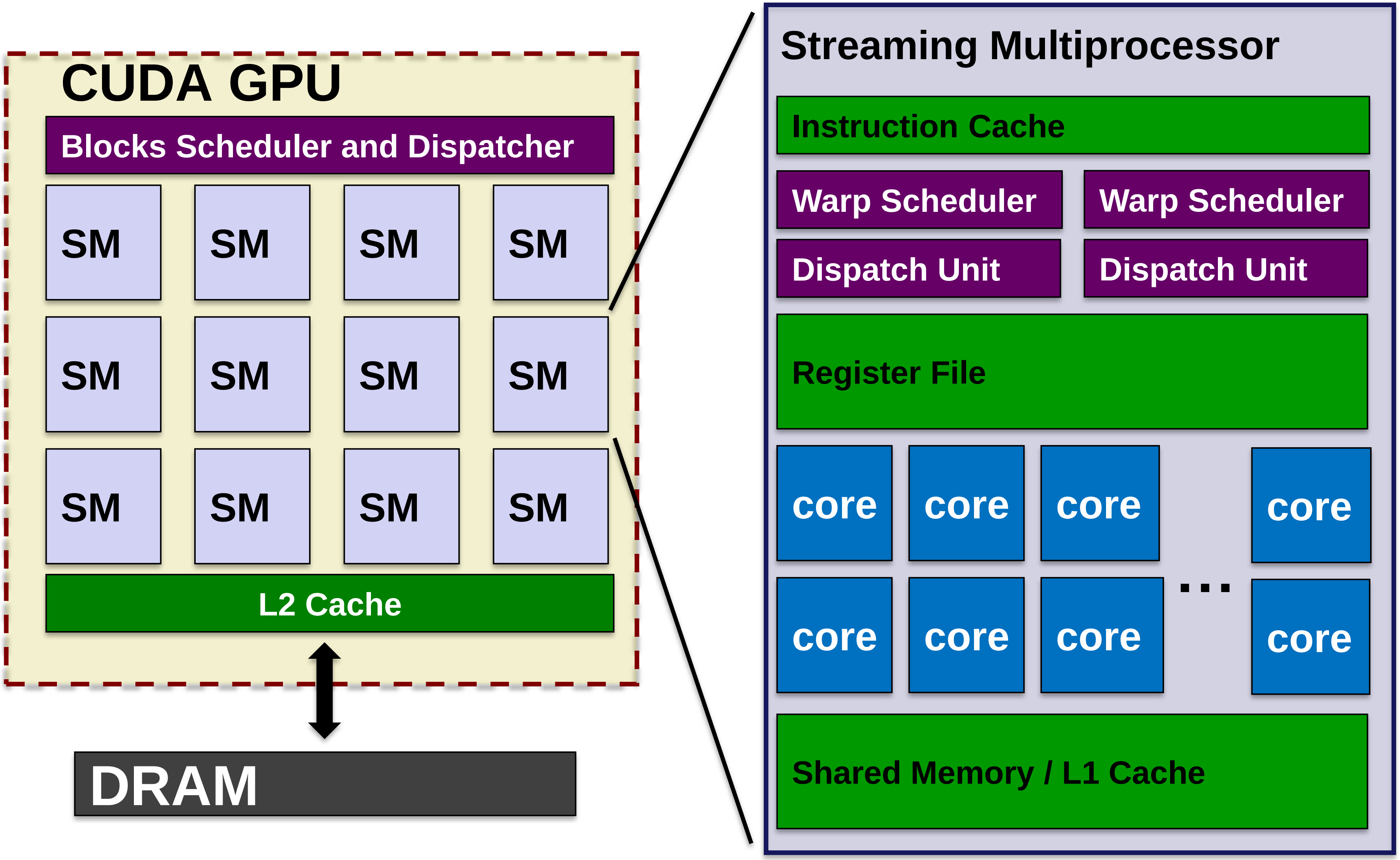


100s of layers
100,000,000 parameters
billions of operations

detection in real-time:
at least 40 frames per second

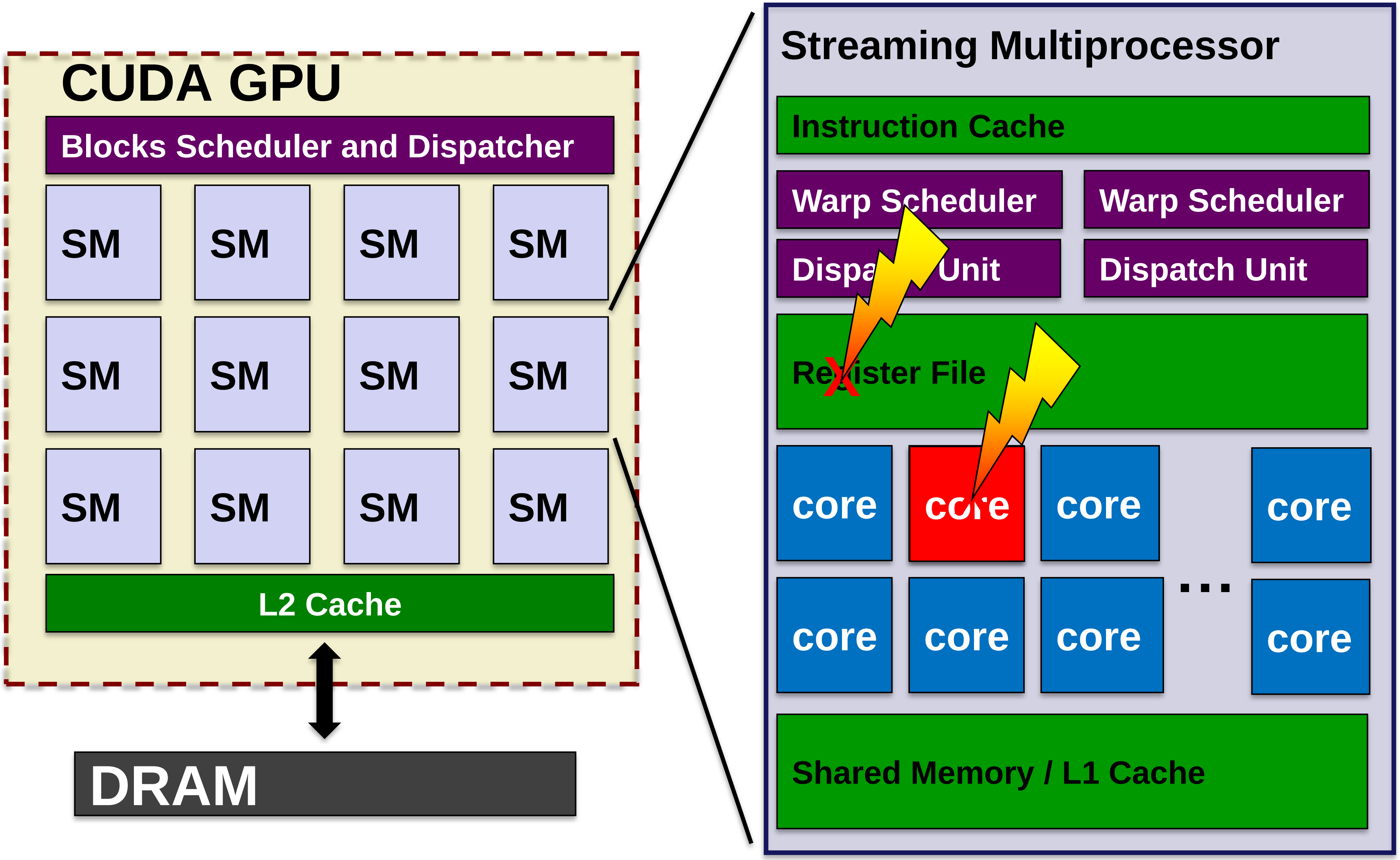
we need highly performant HW: GFLOPS!

Parallel Accelerators



High Performance HW accelerators are required to execute CNN in real-time.

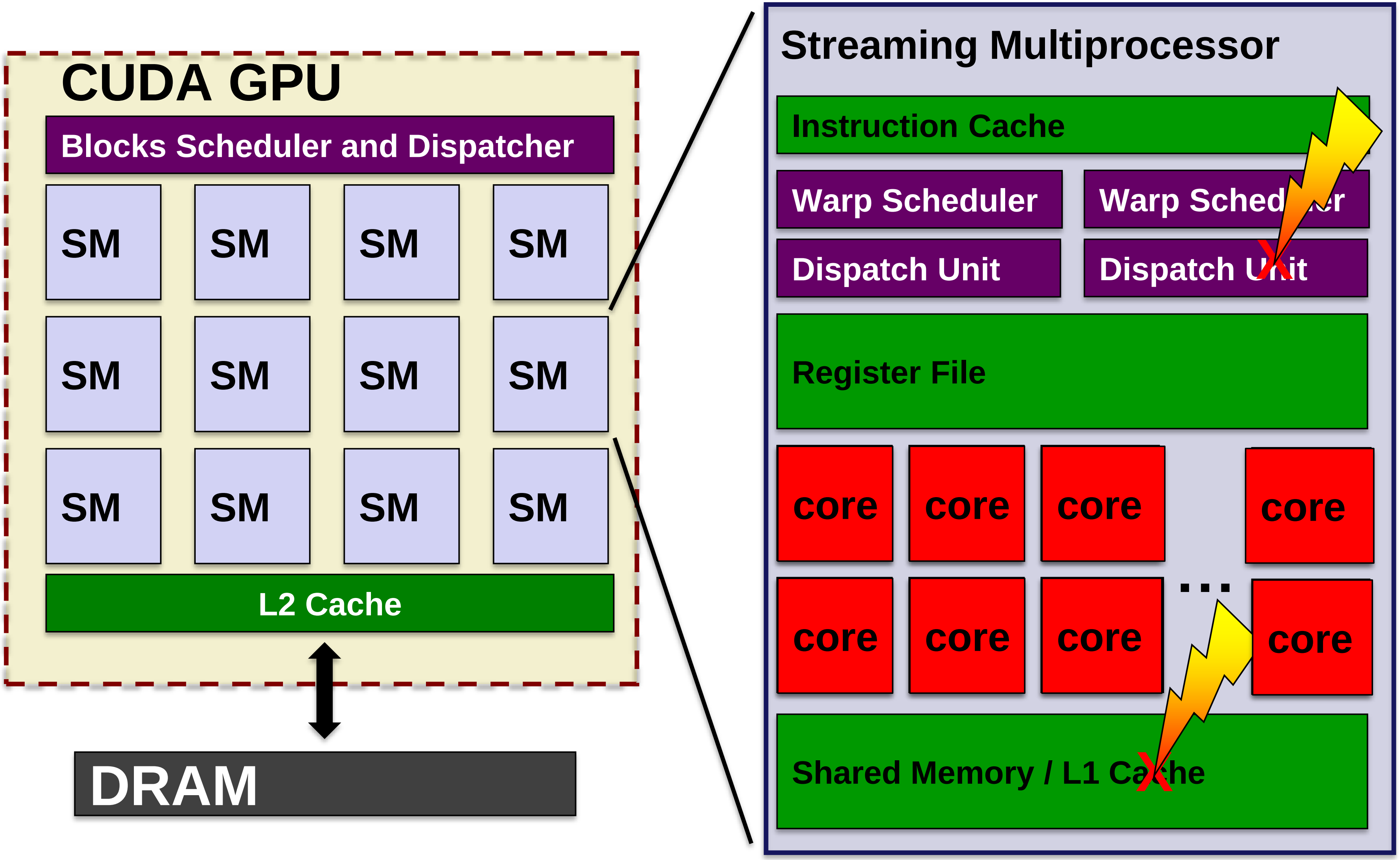
Parallel Accelerators



High Performance HW accelerators are required to execute CNN in real-time.

Large area = high error rate.

Parallel Accelerators



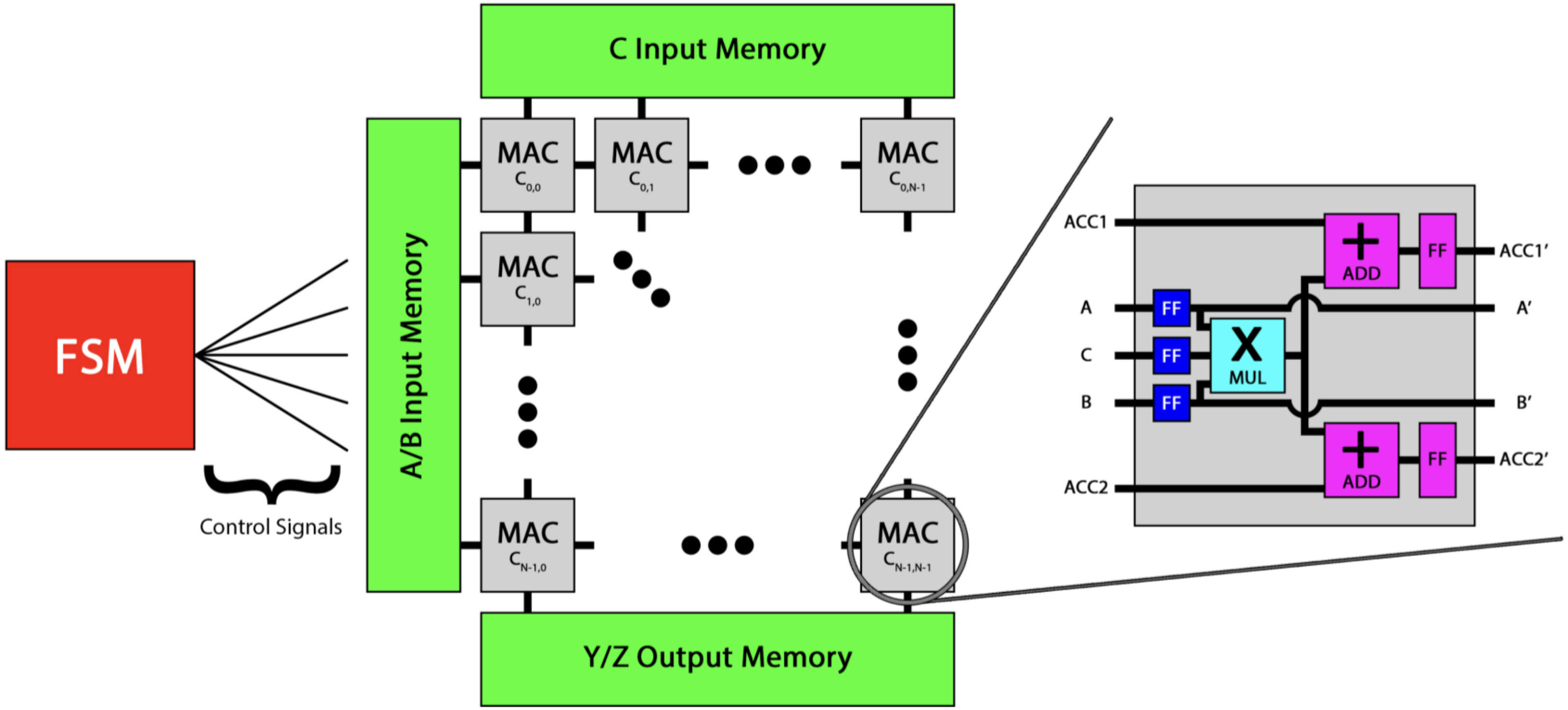
High Performance HW accelerators are required to execute CNN in real-time.

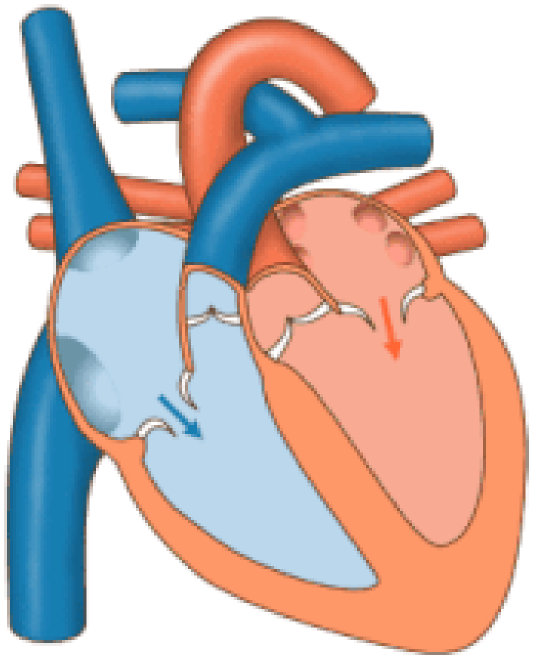
Large area = high error rate.

Shared resources corruption can lead to multiple output elements corruption.

Systolic Arrays

Systolic arrays are big functional units to compute matrices **MAC**. They are composed of an array of connected **mul and add units**.
Data is pumped in while previous data is being processed.





Tensor Processing Unit

Google's Tensor Processing Unit (TPU) is an accelerator able to execute elementary machine learning operations (convolution, pooling, ReLU)

A host (and a good SW framework) is required to execute the neural network!

Reduced power consumption

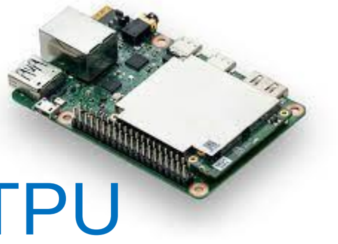
Elementary ML operations in low data precision

Used at scale → **Important to investigate their reliability**

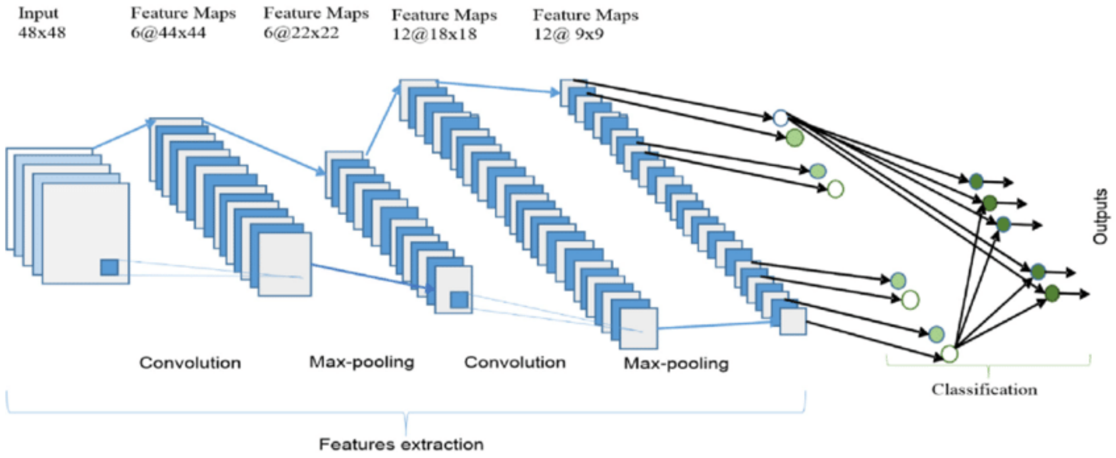
Fault propagation

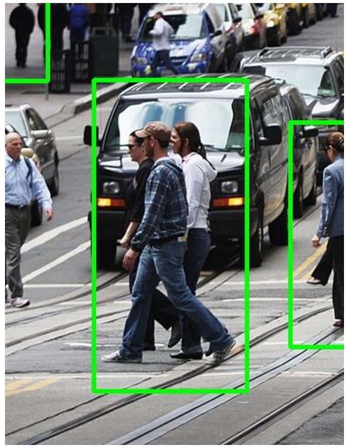
GPU

FPGA

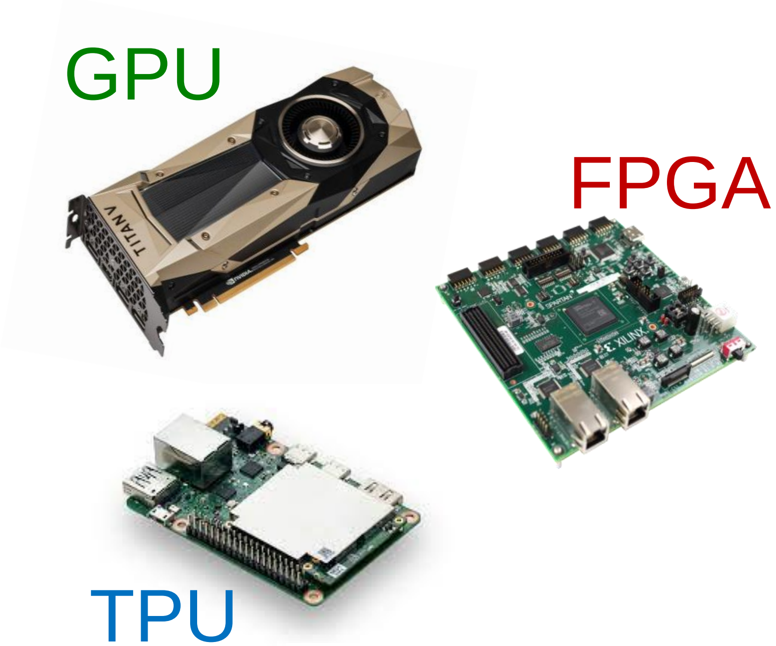


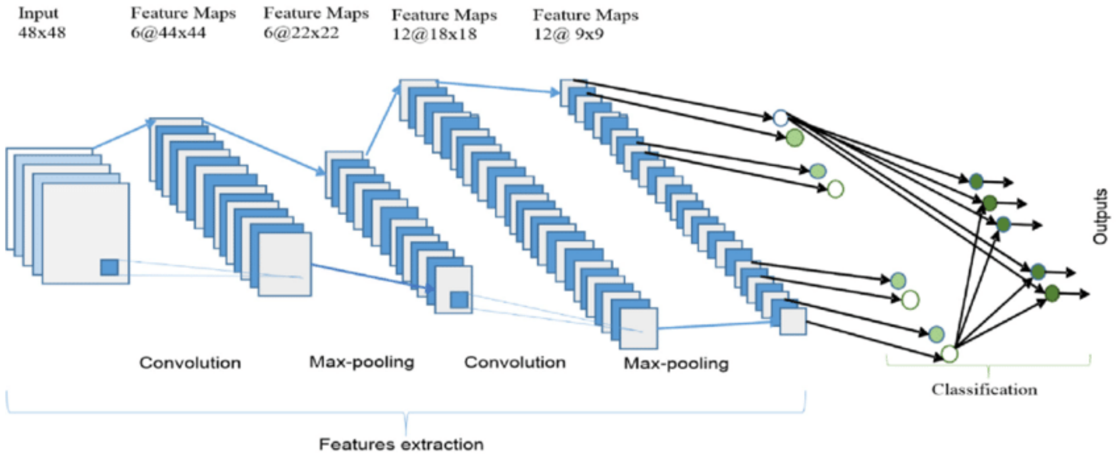
TPU

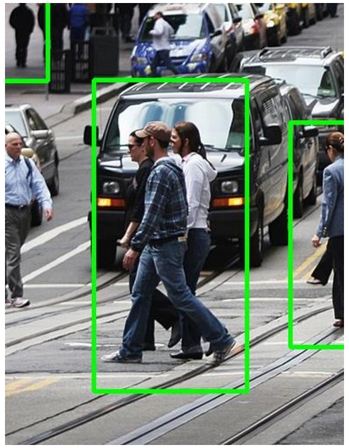


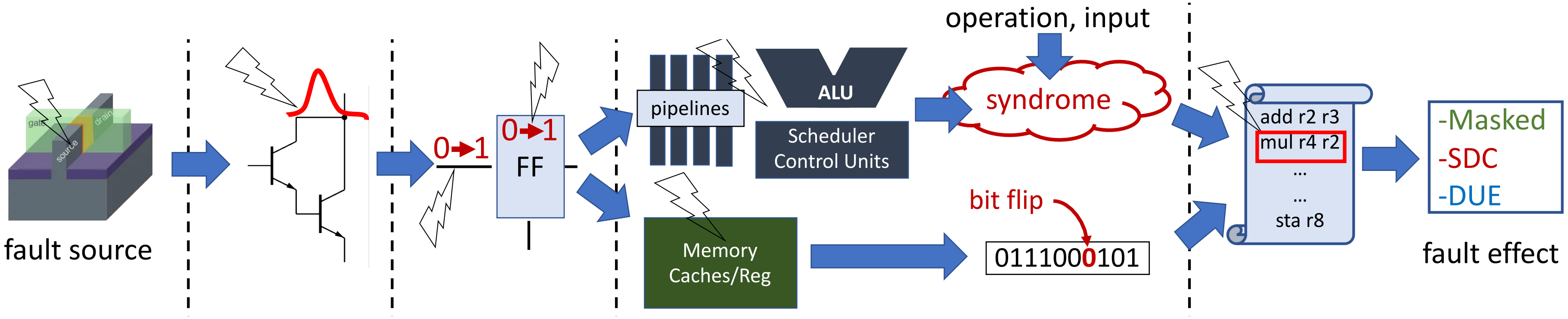


Fault propagation

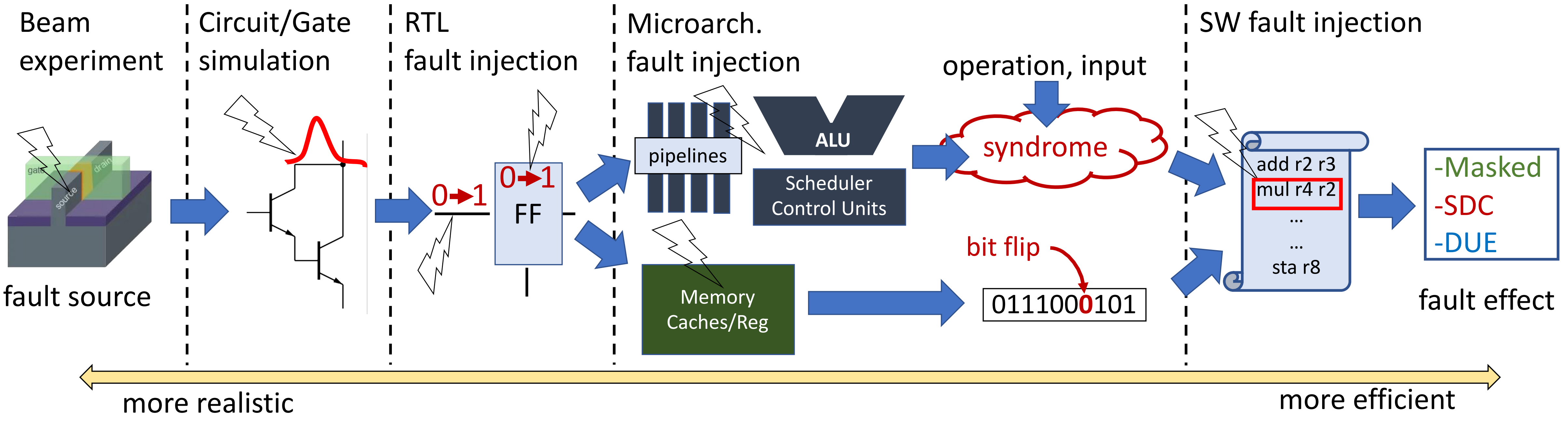




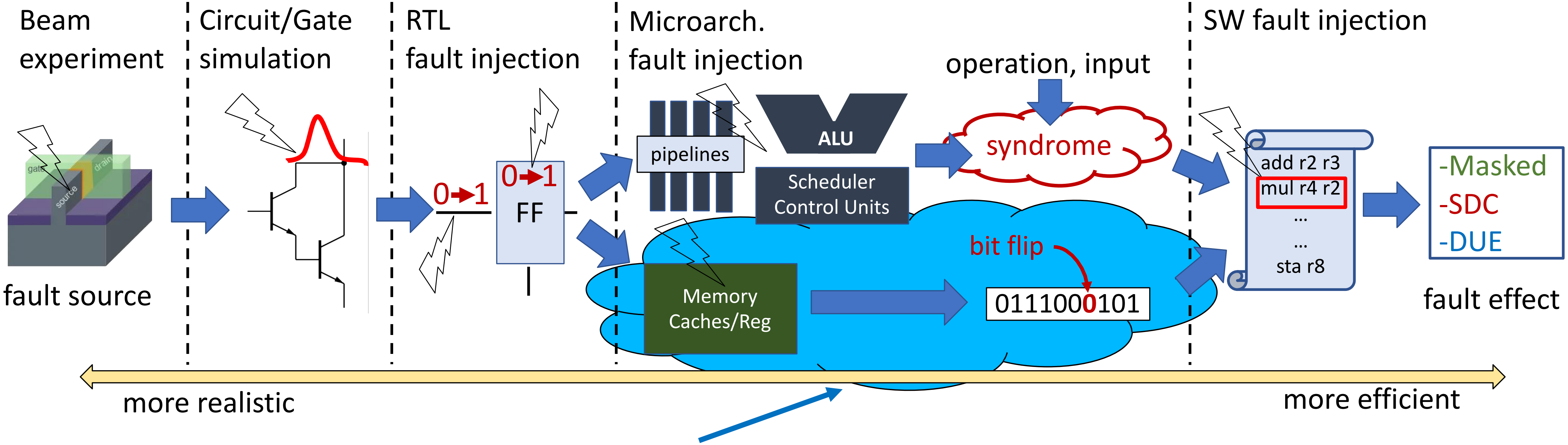




Evaluation methodologies



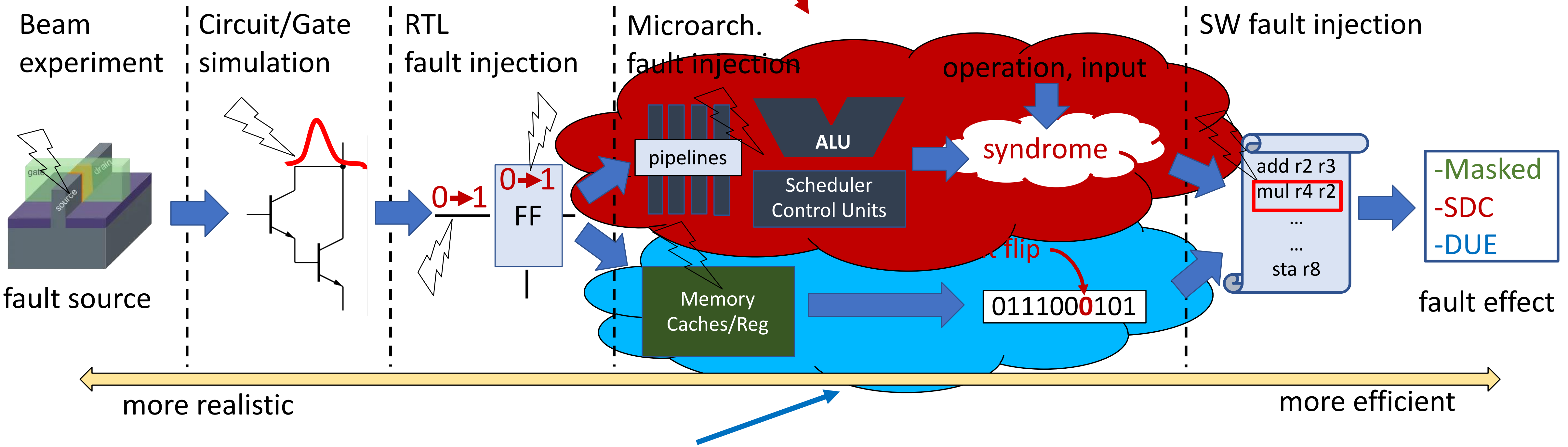
Evaluation methodologies



Memory has a naïve fault model: single bit flips
Well studied for SRAM and DDR (since the 80s)
Memory is easily protectable (ECC)

Evaluation methodologies

Faults in logic have not-trivial syndrome on the output
Largely unknown for complex devices
No efficient protection available

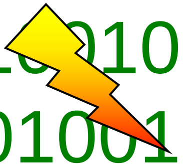


Memory has a naïve fault model: single bit flips
Well studied for SRAM and DDR (since the 80s)
Memory is easily protectable (ECC)

Memory vs Logic

radiation corrupts some bits

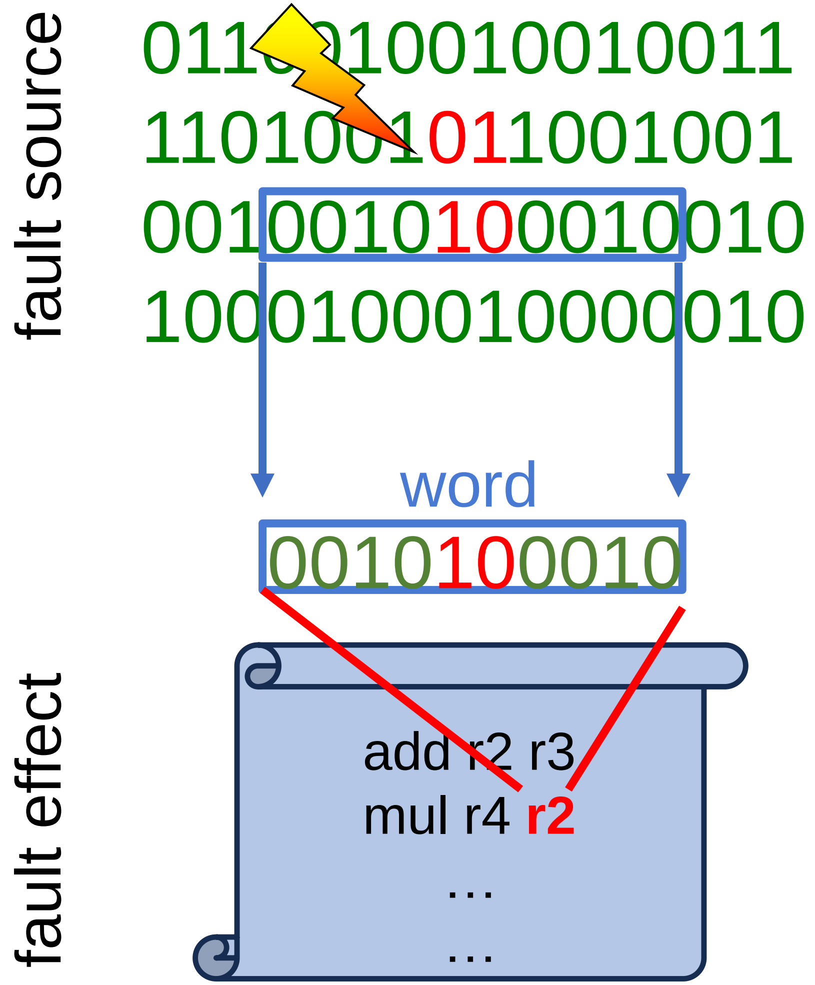
fault source



0110010010010011
1101001011001001
0010010100010010
1000100010000010

Memory vs Logic

radiation corrupts some bits

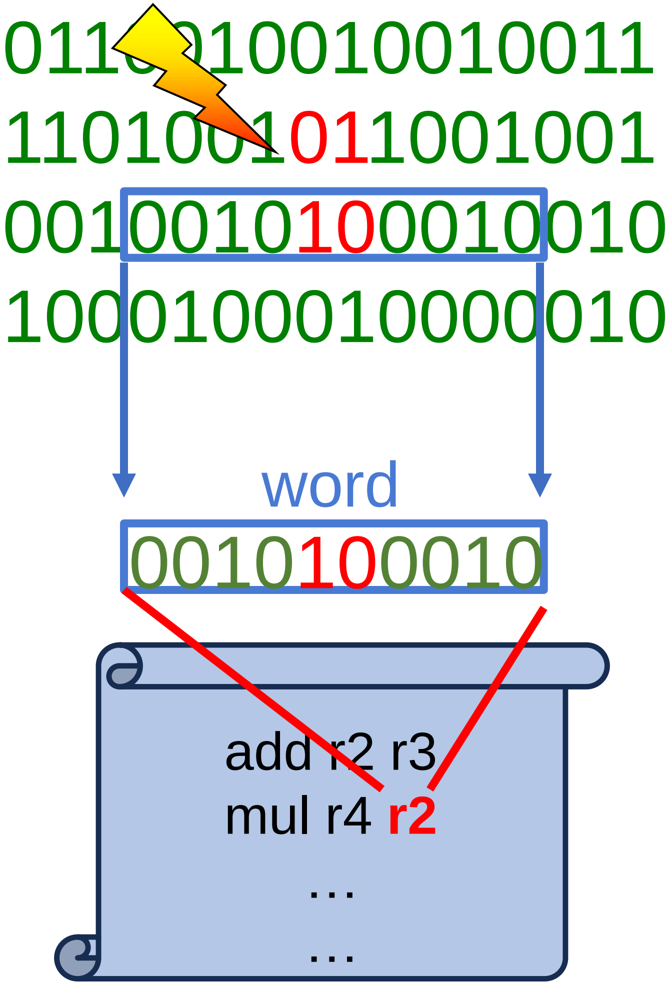


Memory vs Logic

radiation corrupts some bits

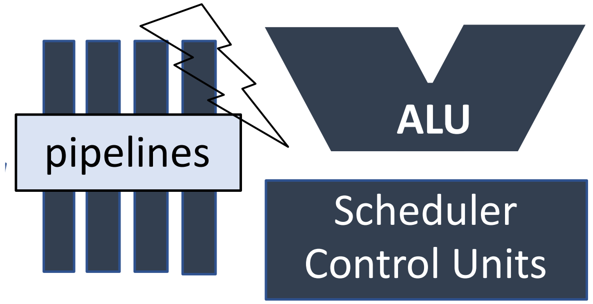
radiation corrupts logic

fault source



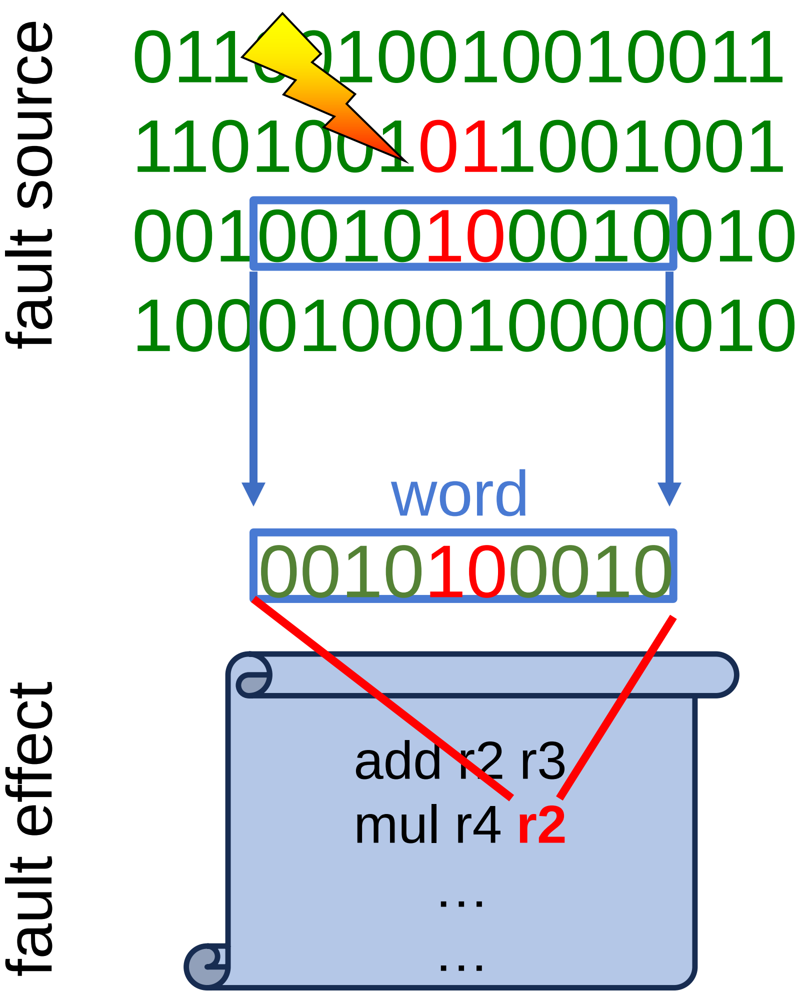
fault effect

fault source

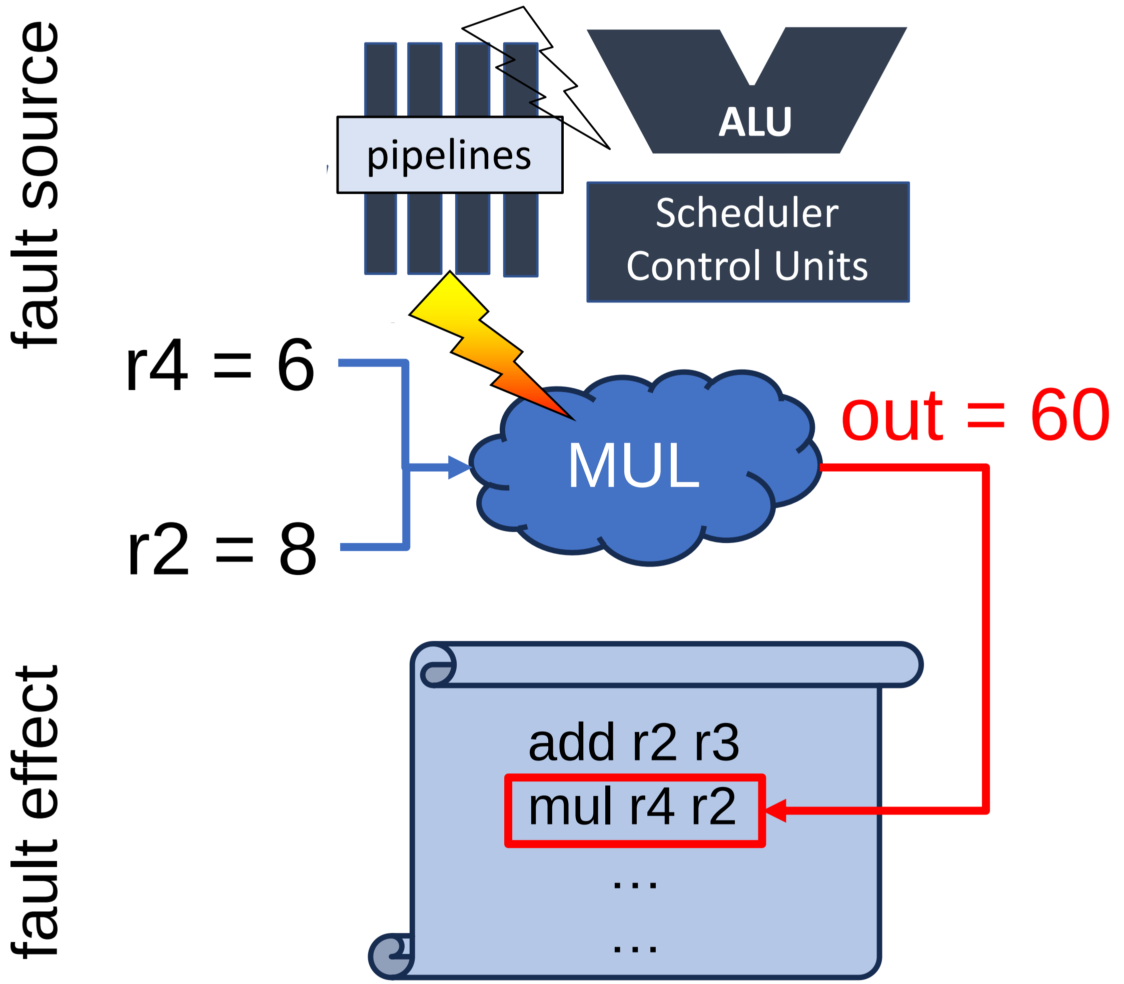


Memory vs Logic

radiation corrupts some bits

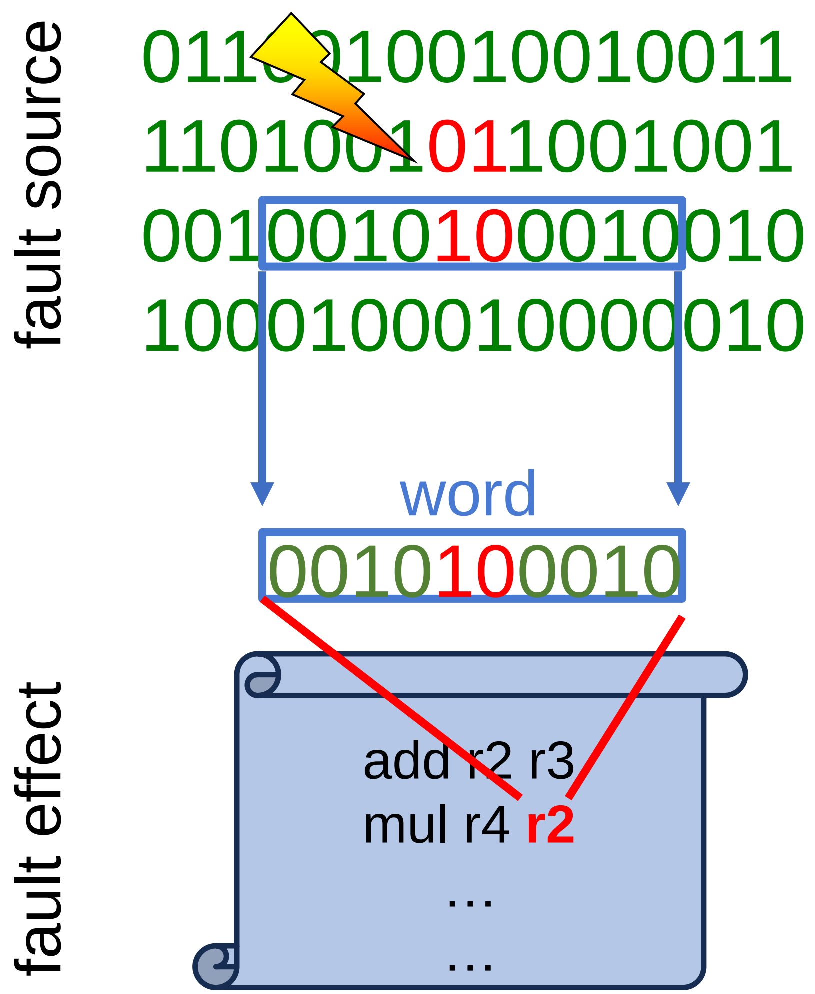


radiation corrupts logic

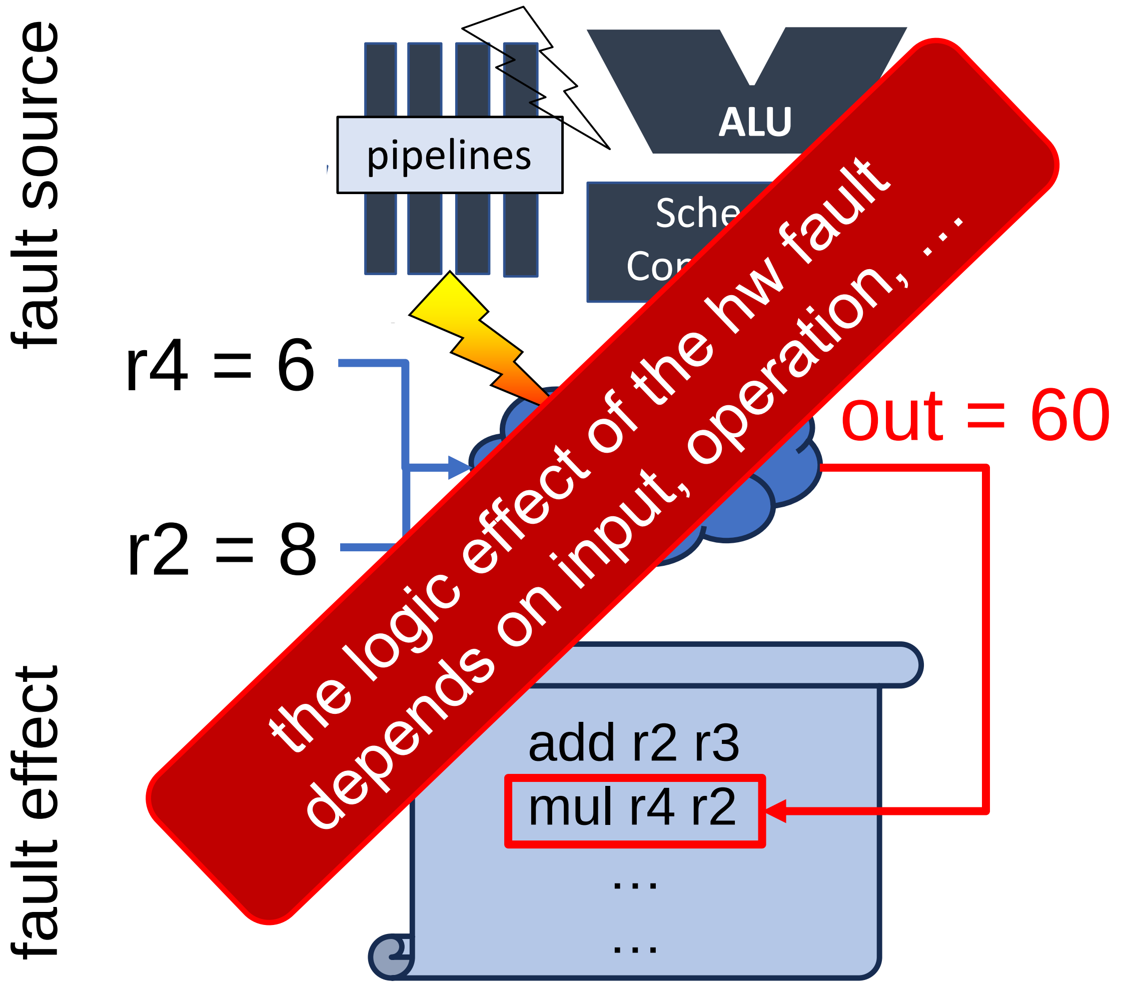


Memory vs Logic

radiation corrupts some bits

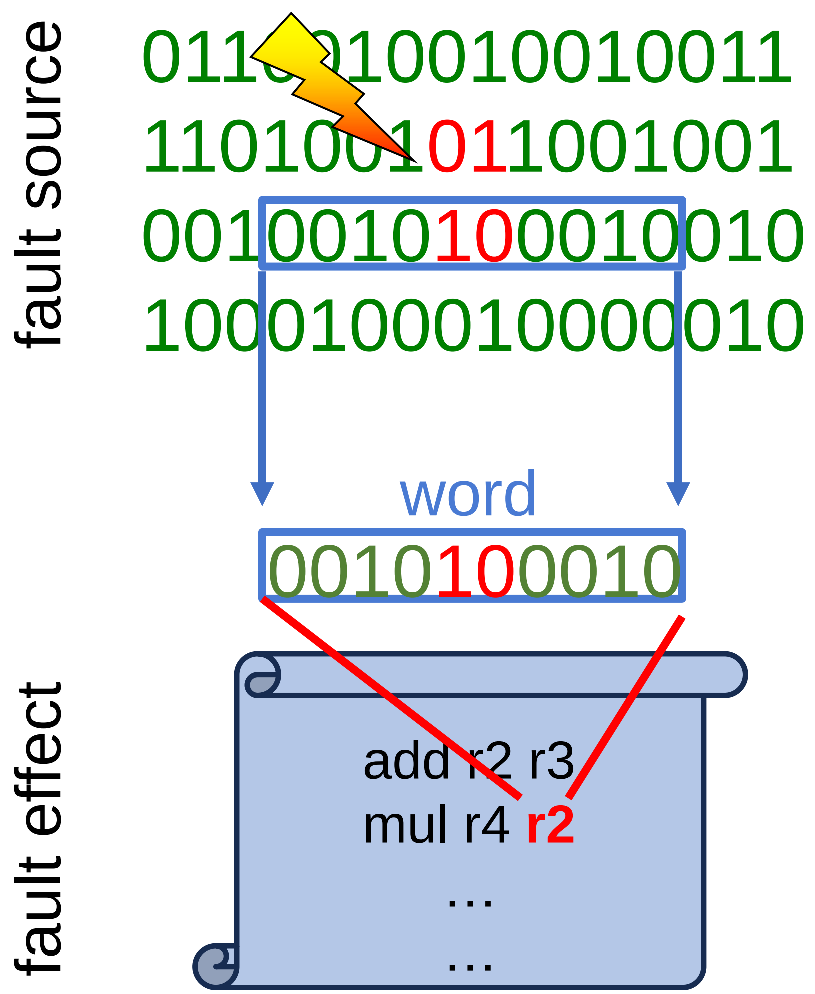


radiation corrupts logic

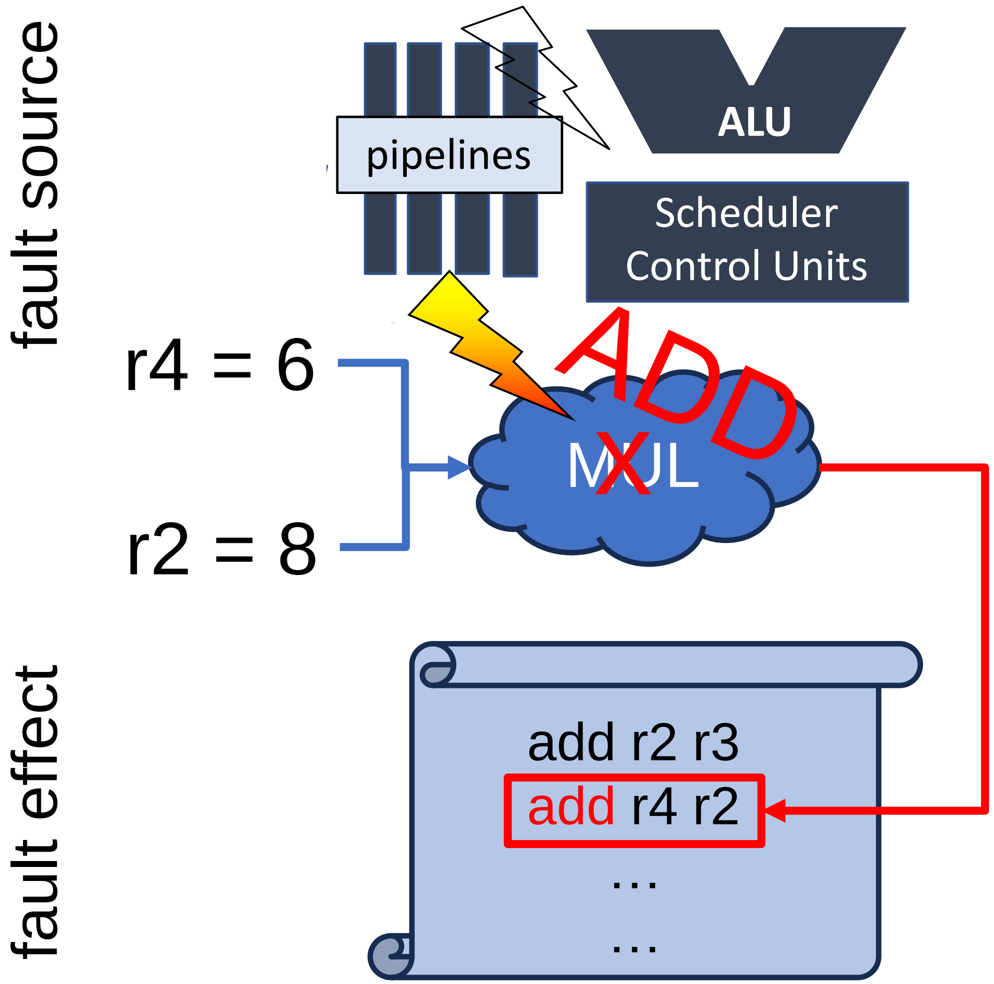


Memory vs Logic

radiation corrupts some bits

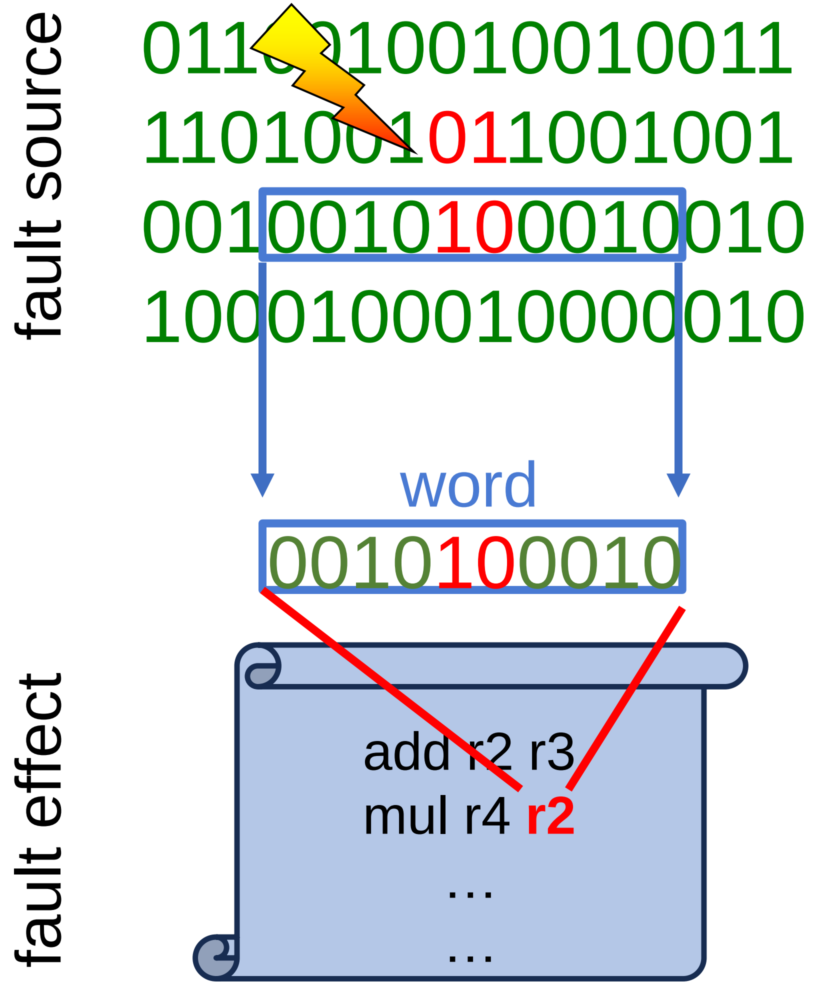


radiation corrupts logic

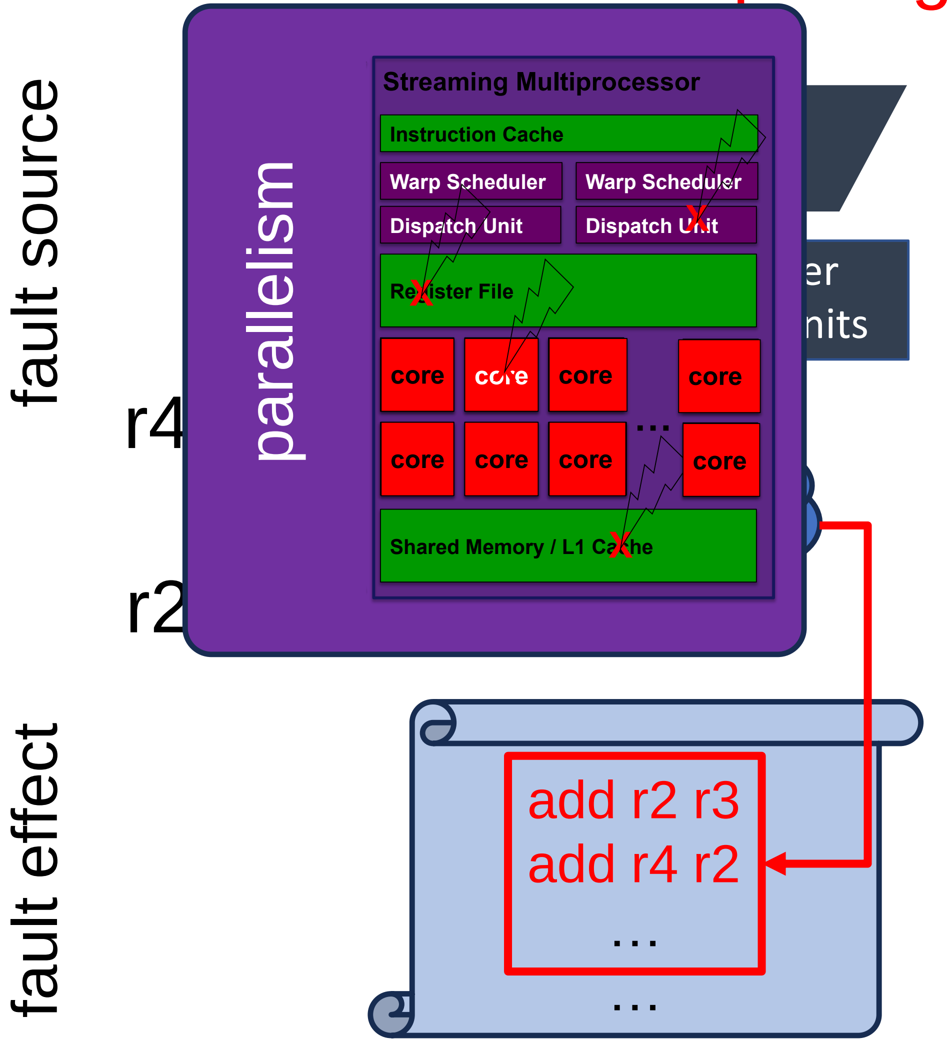


Memory vs Logic

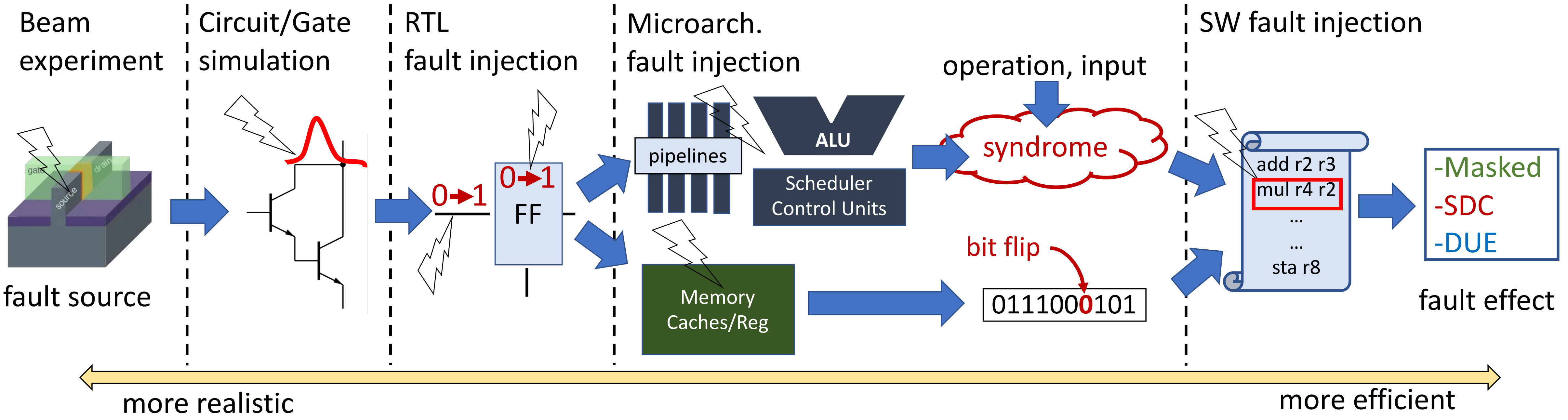
radiation corrupts some bits



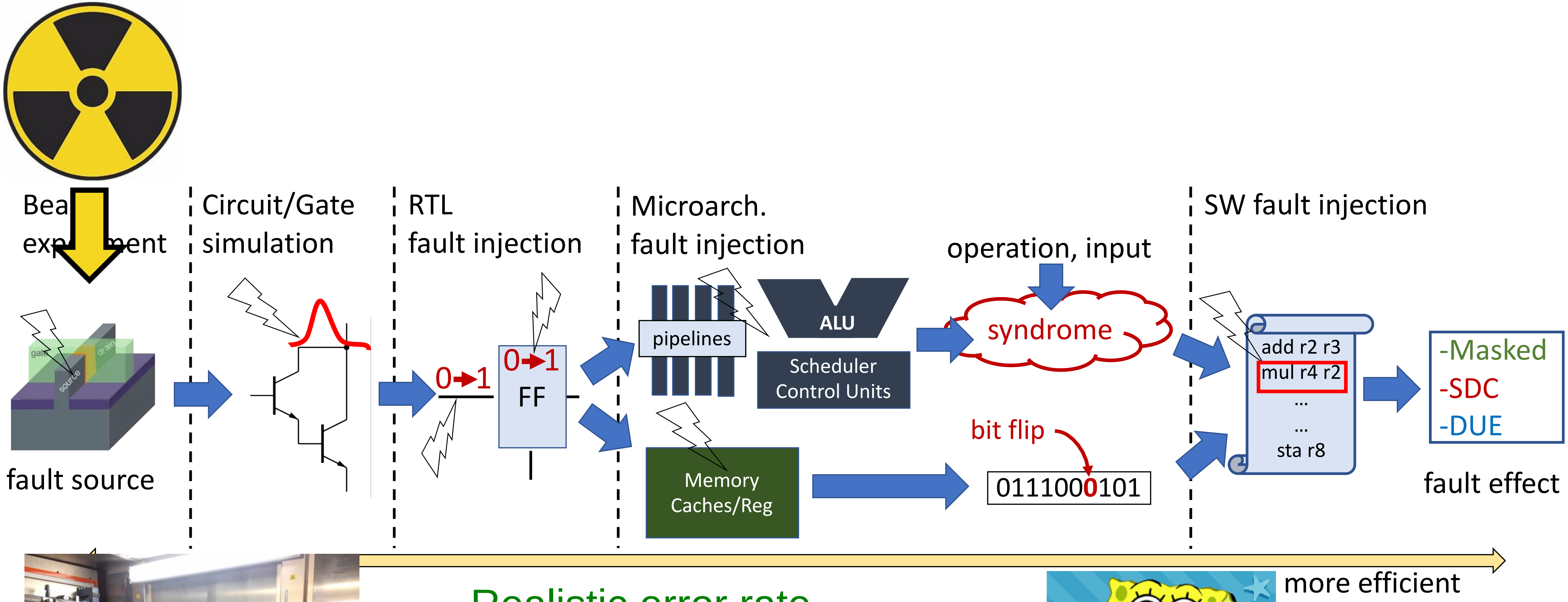
radiation corrupts logic

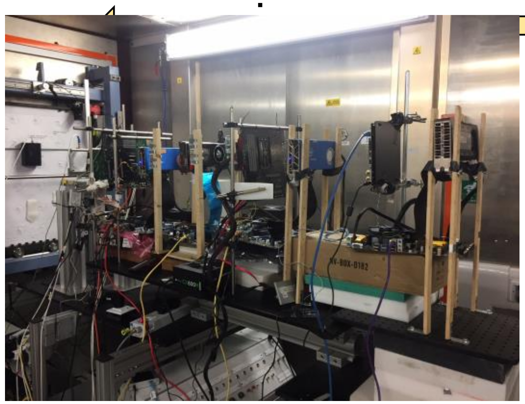


Evaluation methodologies

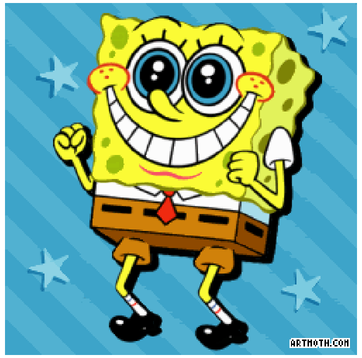


Beam experiments



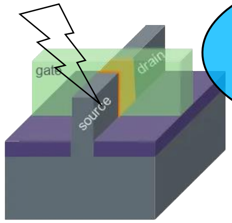


- Realistic error rate
- Realistic fault model
- All HW is exposed



Beam experiments

Beam experiment



fault source

Circuit/Gate simulation

RTL fault injection

Micro fault

input

syndrome

pipelines

Scheduler Control Units

Memory Caches/Reg

bit flip

0111000101

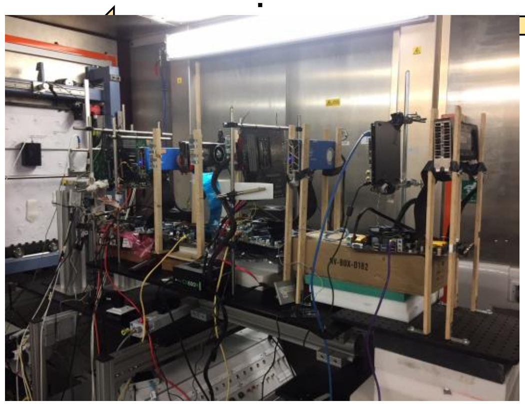
we do not know what happens here

SW fault injection

add r2 r3
mul r4 r2
...
...
sta r8

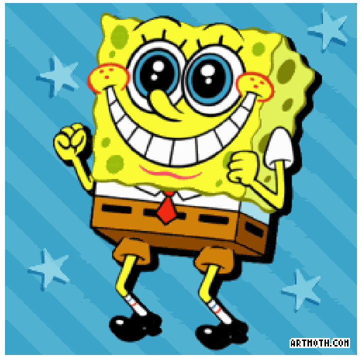
- -Masked
- -SDC
- -DUE

fault effect



more efficient

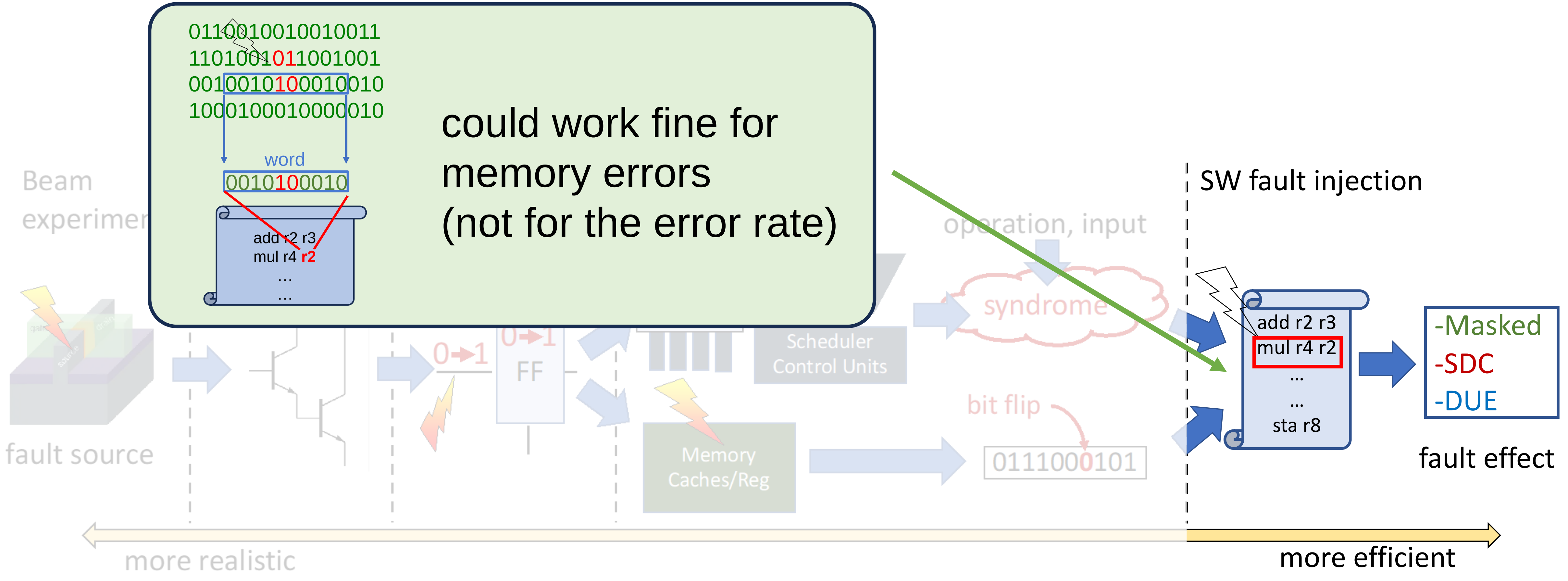
- -Realistic error rate
- -Realistic fault model
- -All HW is exposed



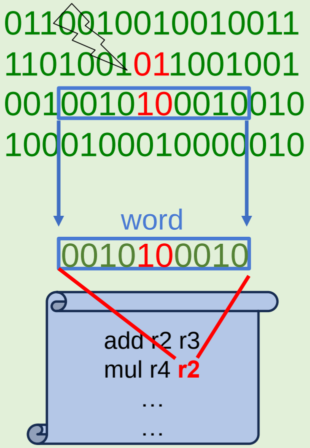
Paolo Rech

SW fault injection

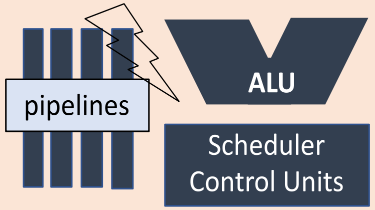
could work fine for memory errors (not for the error rate)



SW fault injection

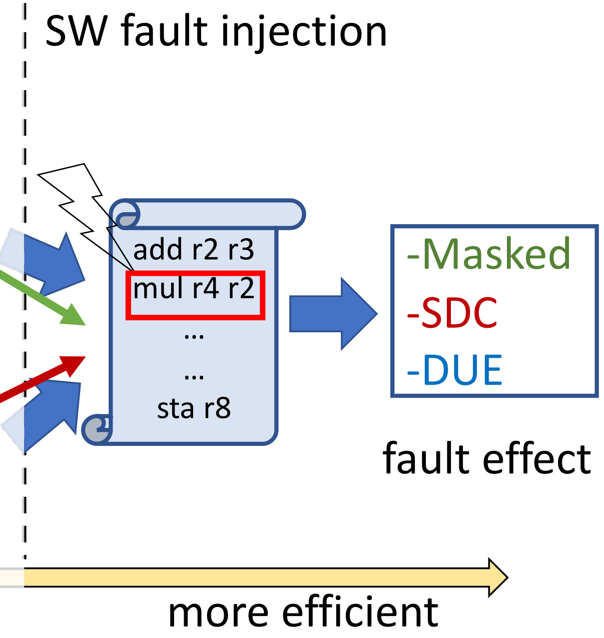


could work fine for memory errors (not for the error rate)



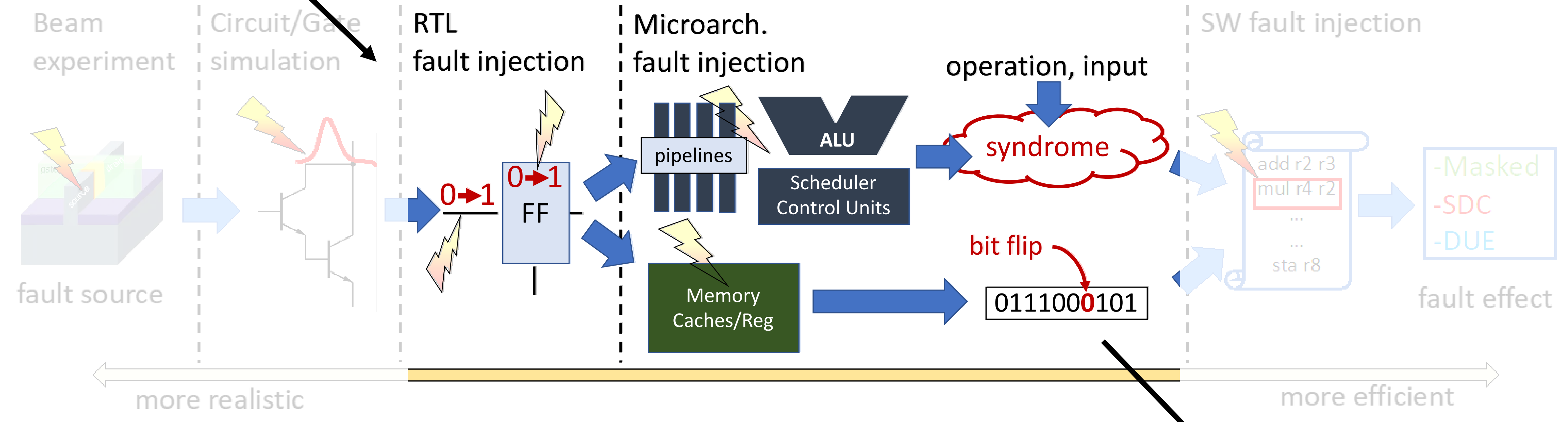
less accurate for logic or data-path faults.

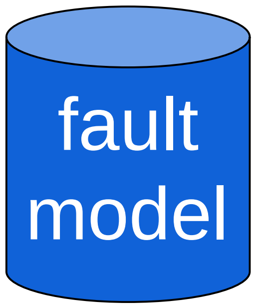
We do not know the high-level effect!



Cross-layer evaluation

FlexGrip+ GPU model (@PoliTo)
GeFIN ARM model (@UAthens)



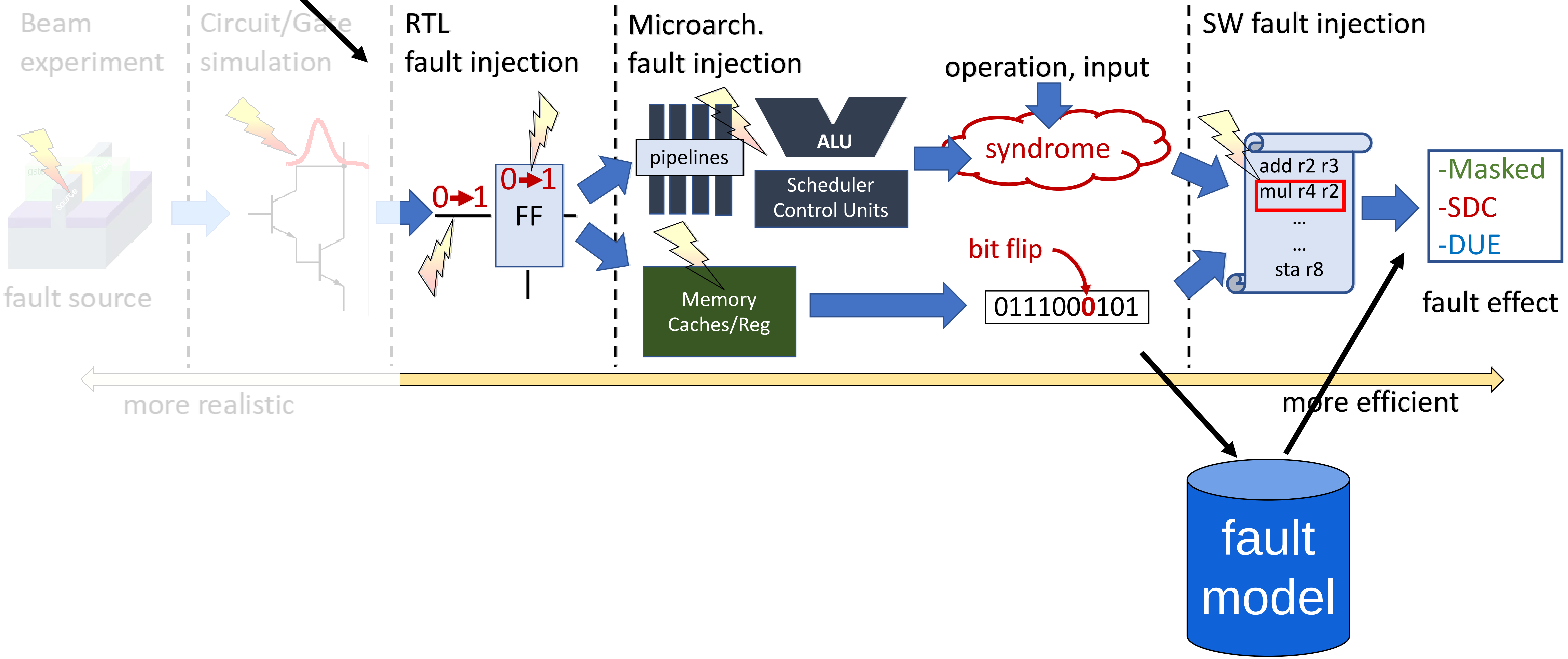


Characterization of the effects on micro-instructions

Paolo Rech

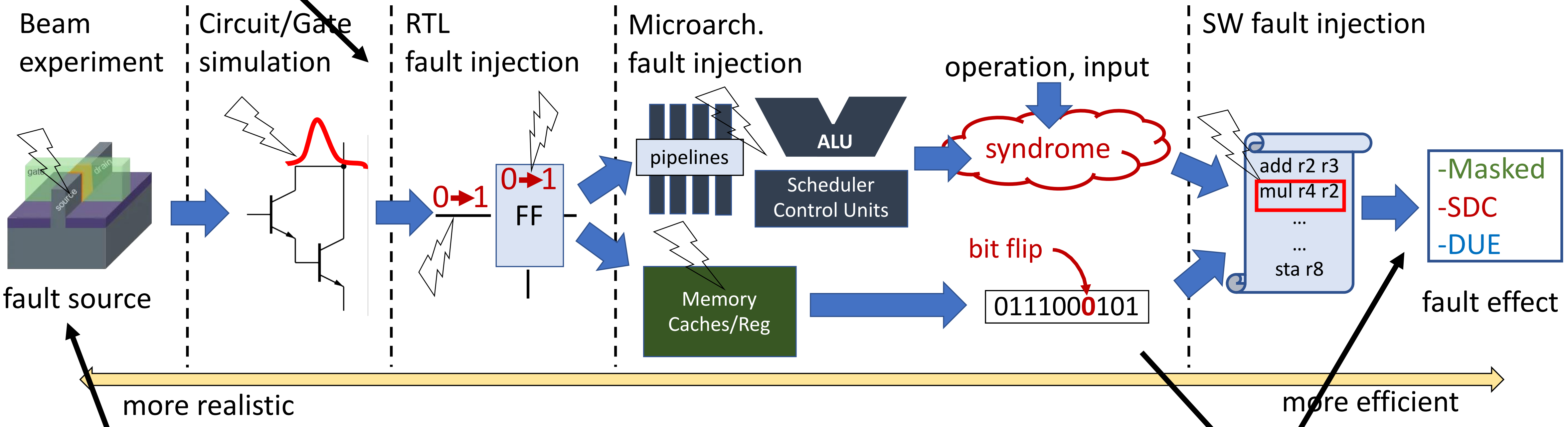
Cross-layer evaluation

FlexGrip+ GPU model (@PoliTo)
GeFIN ARM model (@UAthens)

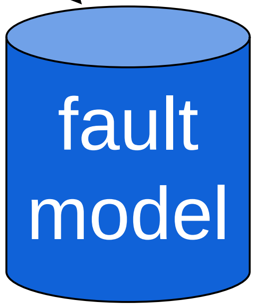


Cross-layer evaluation

FlexGrip+ GPU model (@PoliTo)
GeFIN ARM model (@UAthens)

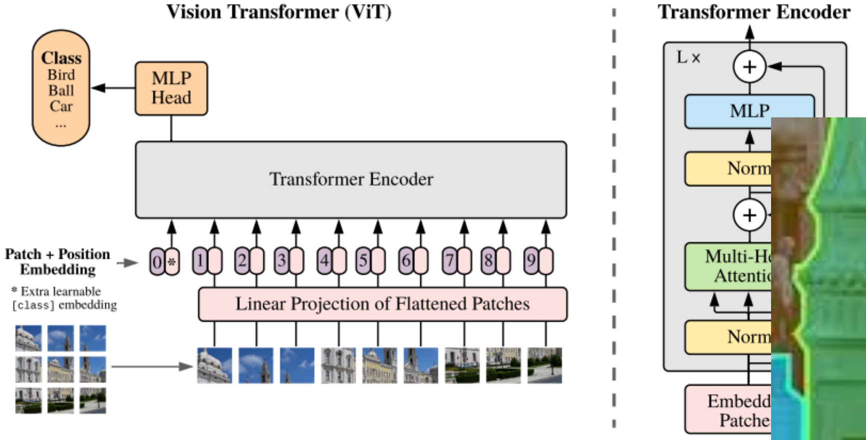


Beam experiments
on micro-instructions to measure
the hardware cross section

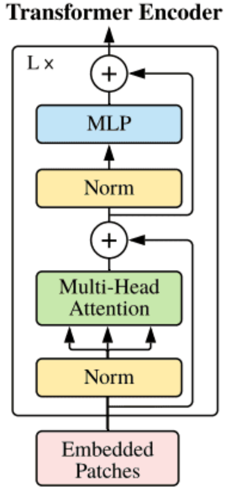


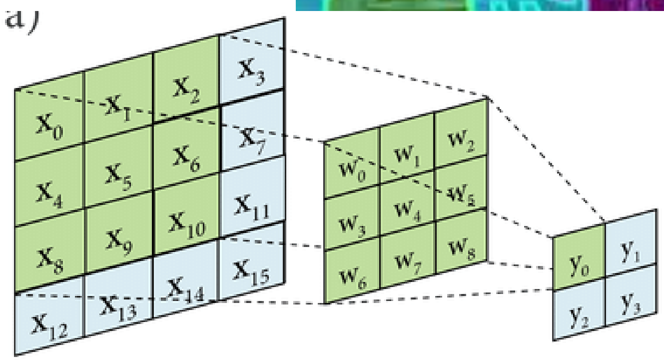
What do we need to test?

Complex models



(Micro)-benchmarks





Atomic operations

A+B

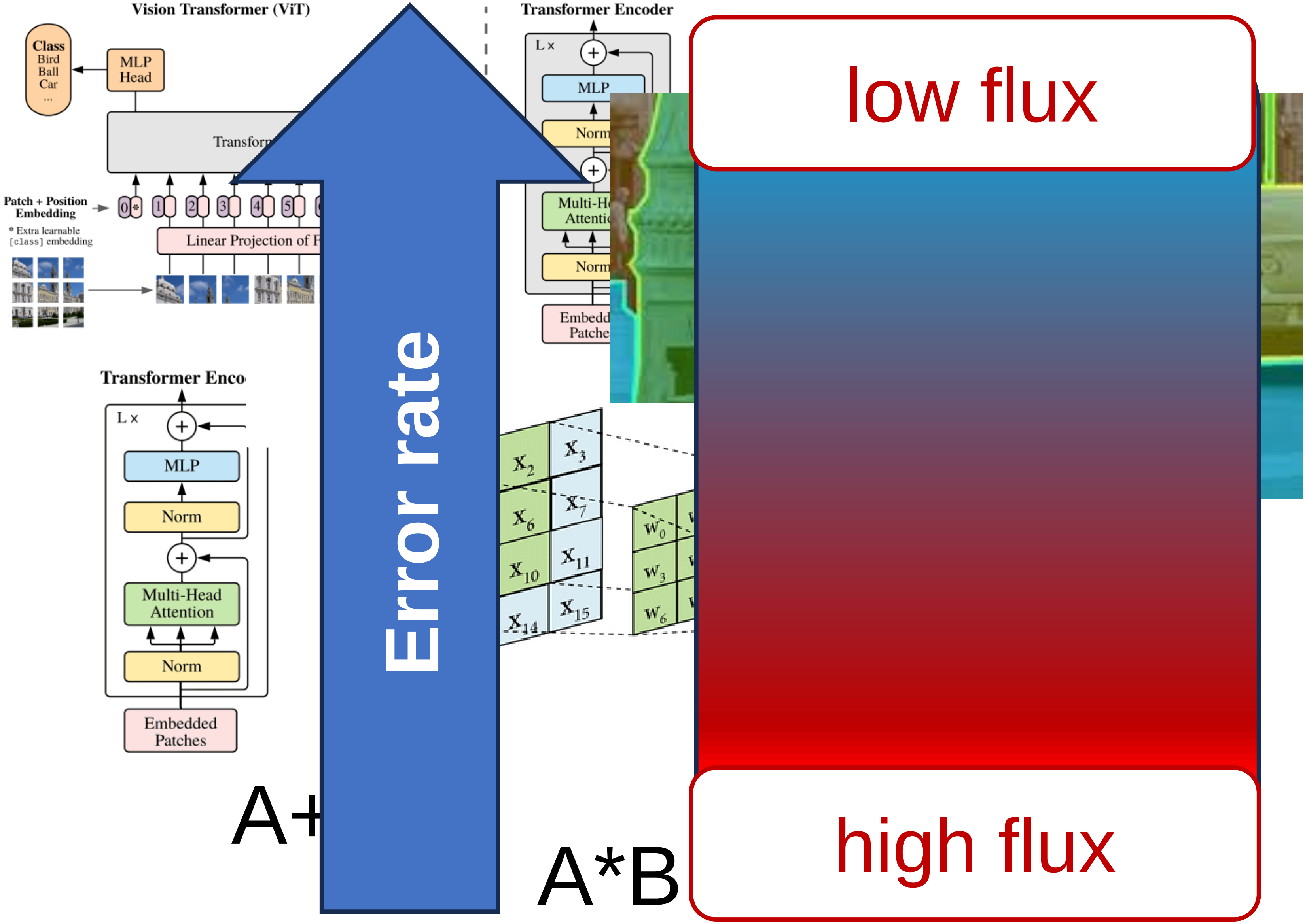
A*B

What do we need to test?

Complex models

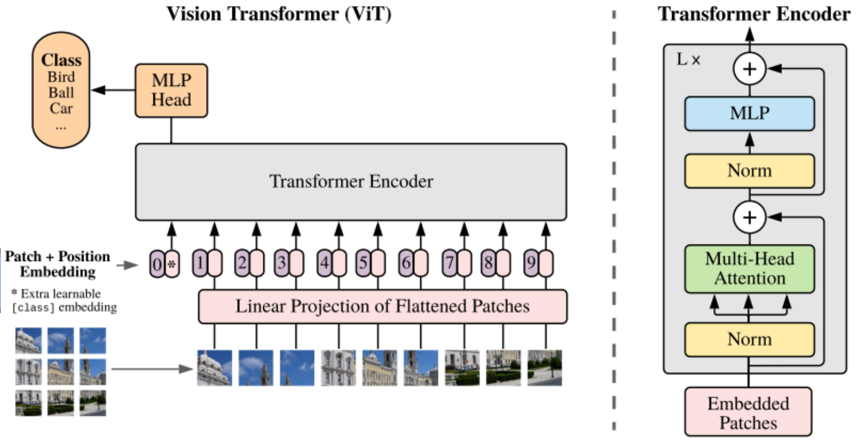
(Micro)-benchmarks

Atomic operations



Setup challenges

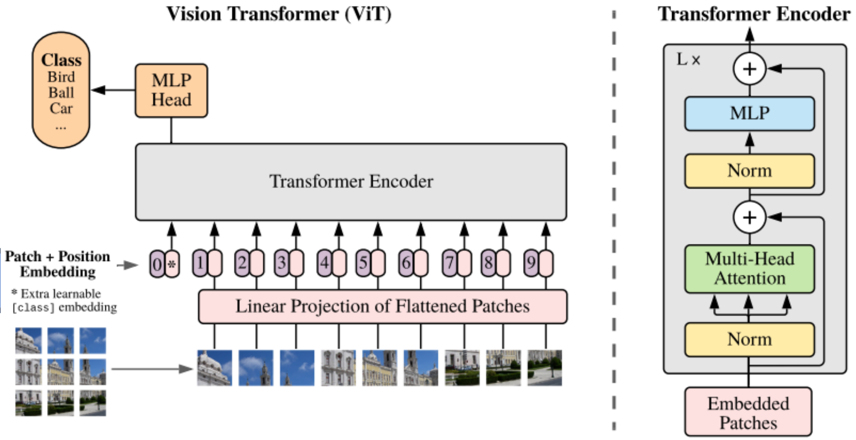
Complex models



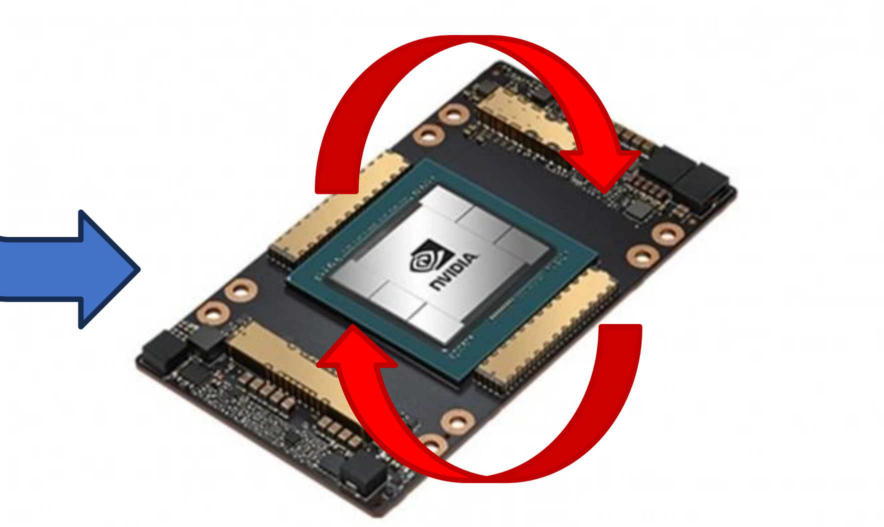
load the model
(1 Trillion params)

Setup challenges

Complex models



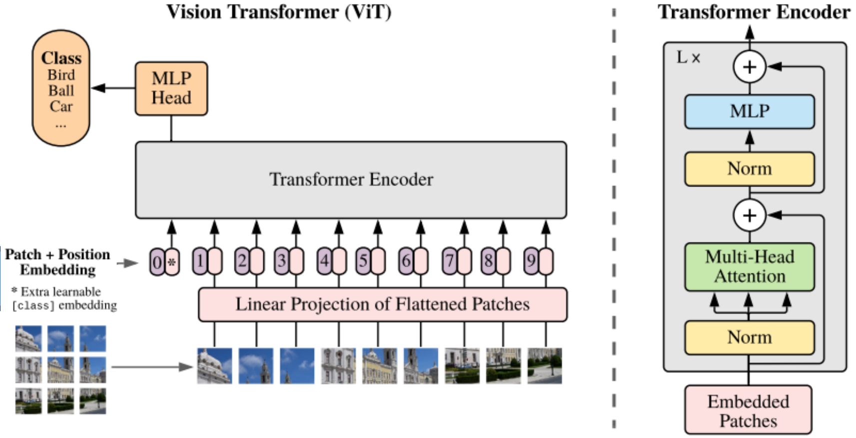
load the model
(1 Trillion params)



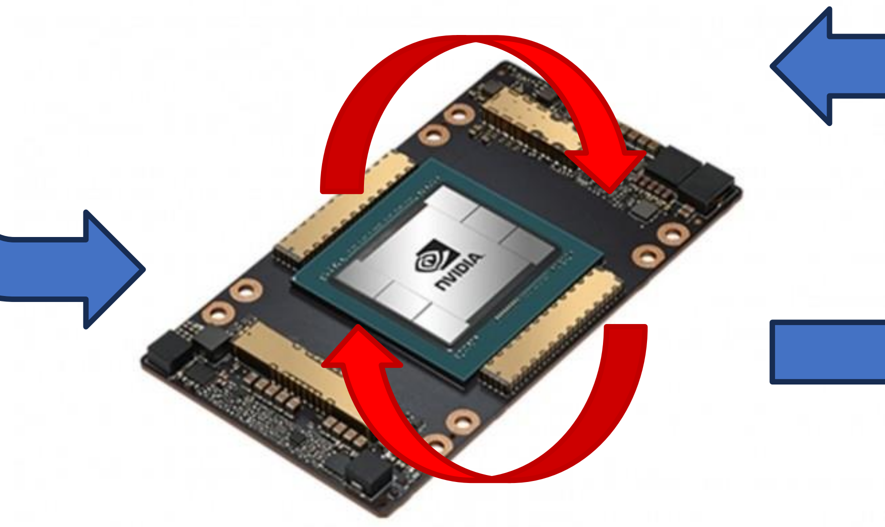
execute the model (~seconds)

Setup challenges

Complex models



load the model
(1 Trillion params)

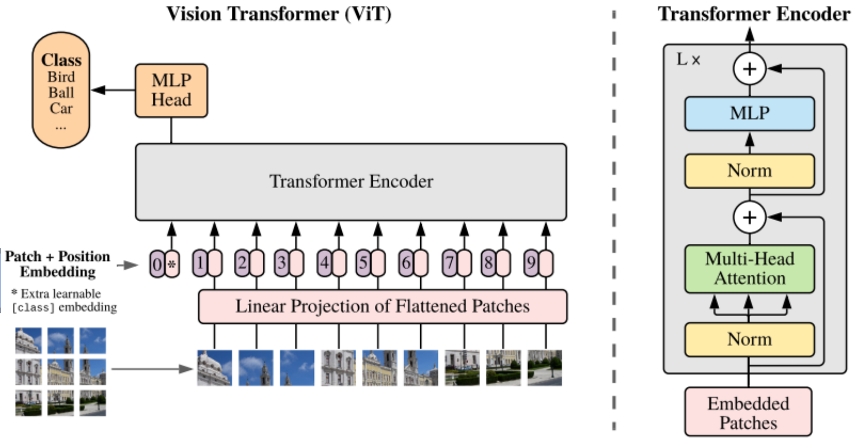


check for errors
(comparing detection)

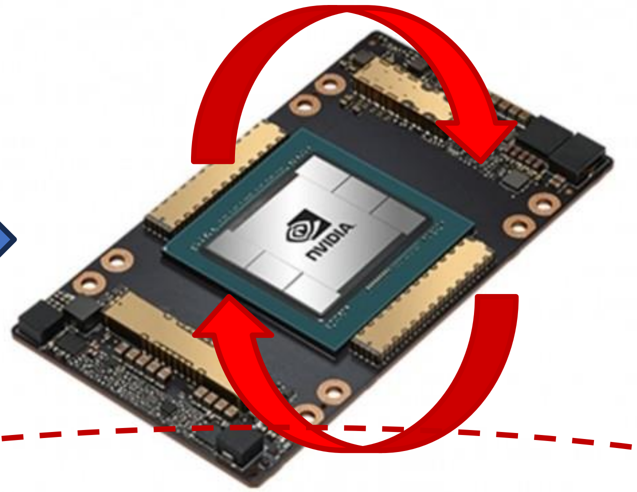
execute the model (~seconds)

Setup challenges

Complex models



load the model
(1 Trillion params)



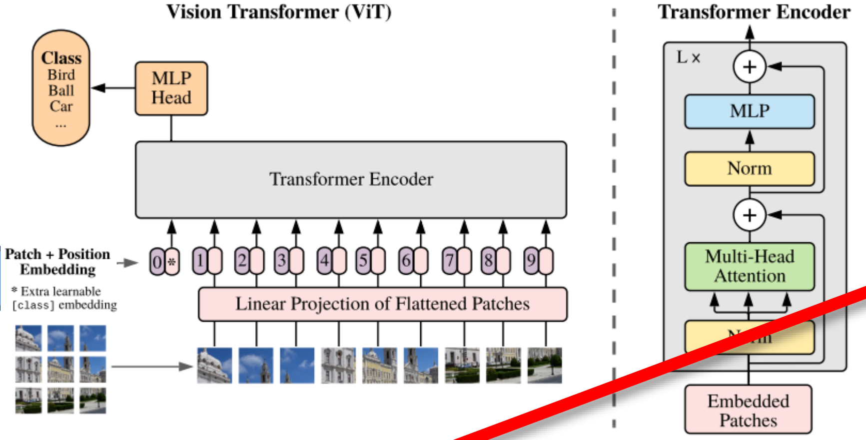
check for errors
(comparing detection)

execute the model (~seconds)

neutrons
are useful

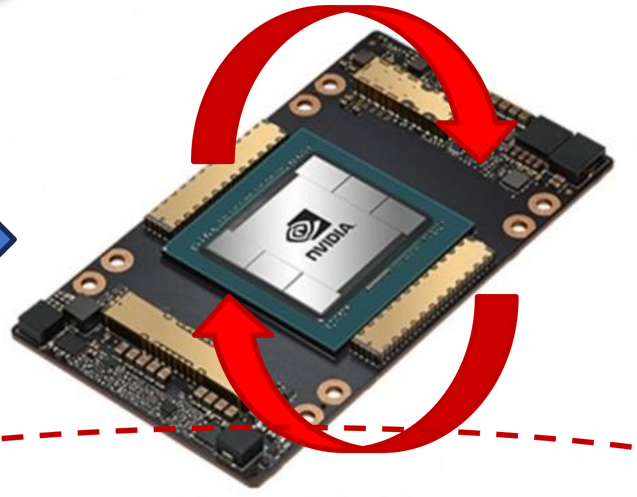
Setup challenges

Complex models



neutrons are wasted

load the model (1 Trillion params)

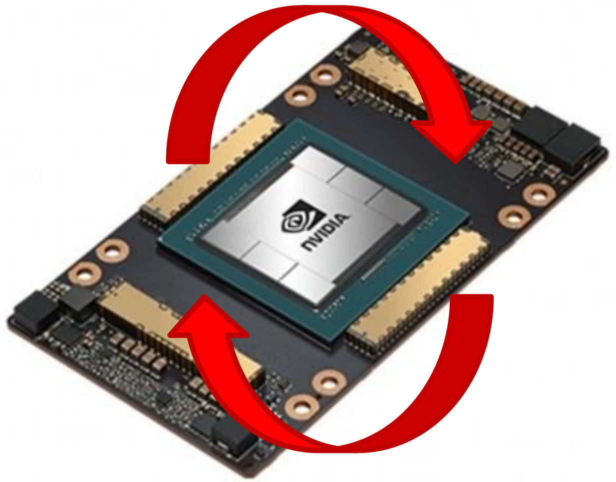


check for errors (comparing detection)

execute the model (~seconds)

neutrons are useful

Setup challenges



check for errors

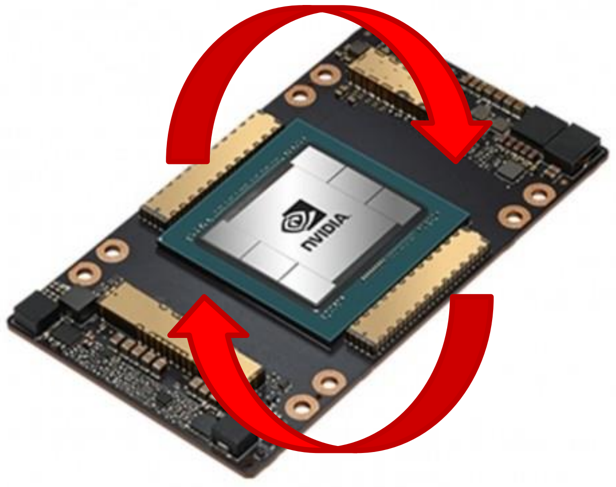
A+B = C ?

Atomic operations

A+B

A*B

Setup challenges



check for errors

A+B = C ?

The complexity of the operation is similar to the complexity of error detection.

risk: waste 50% of neutrons

Atomic operations

A+B

A*B

Setup challenges

The complexity of the

he

Golden rule:

In at least 90% time we must do something useful for the experiment (something that gives us data)

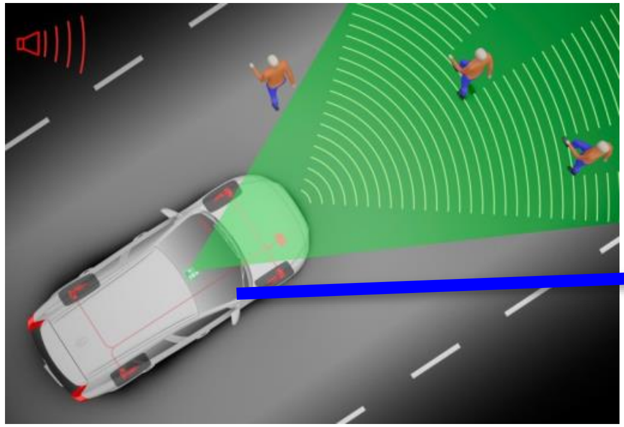
s

Atomic operations

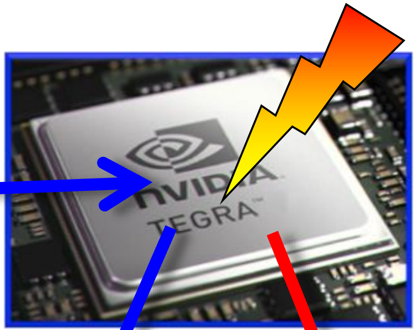
A+B

A*B

Setup challenges

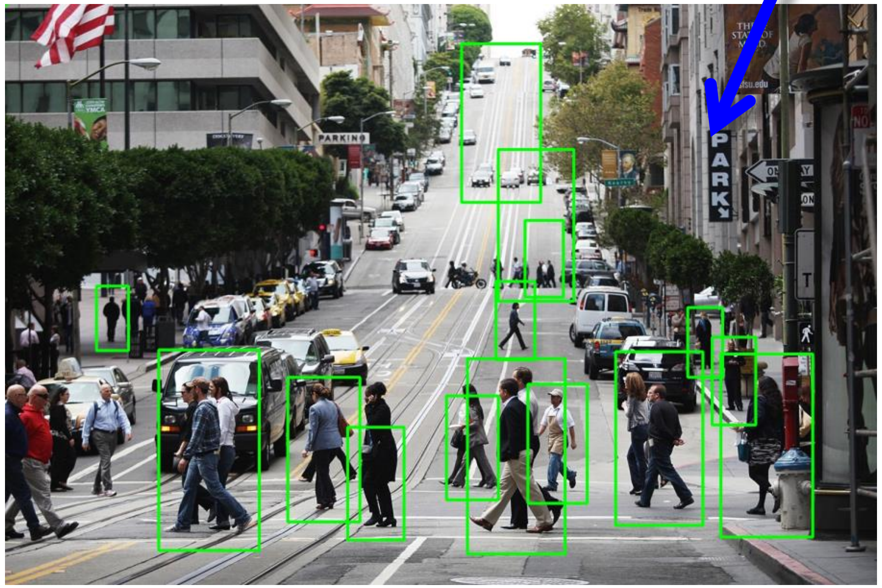


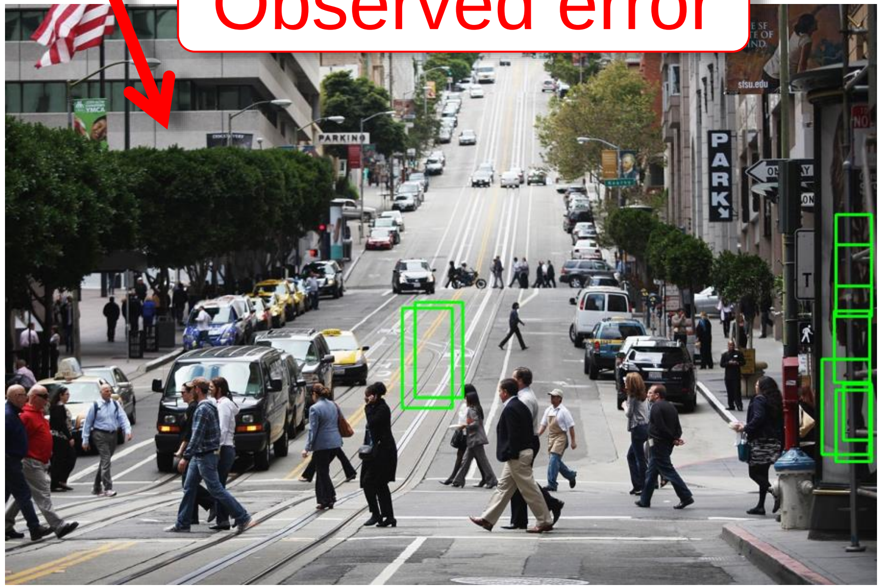
Objects Detection System:



ML accelerator

Observed error



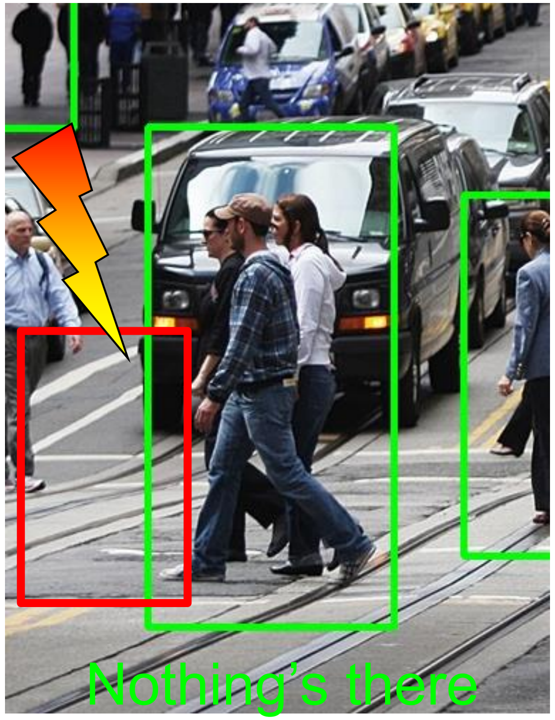


Setup challenges

Expected

Tolerable

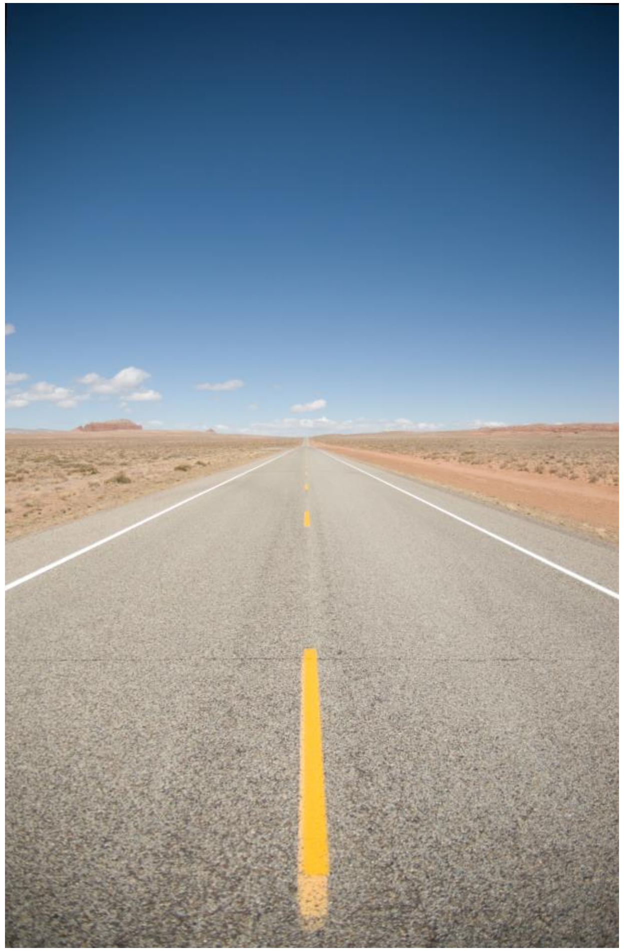
Slight modification of detection



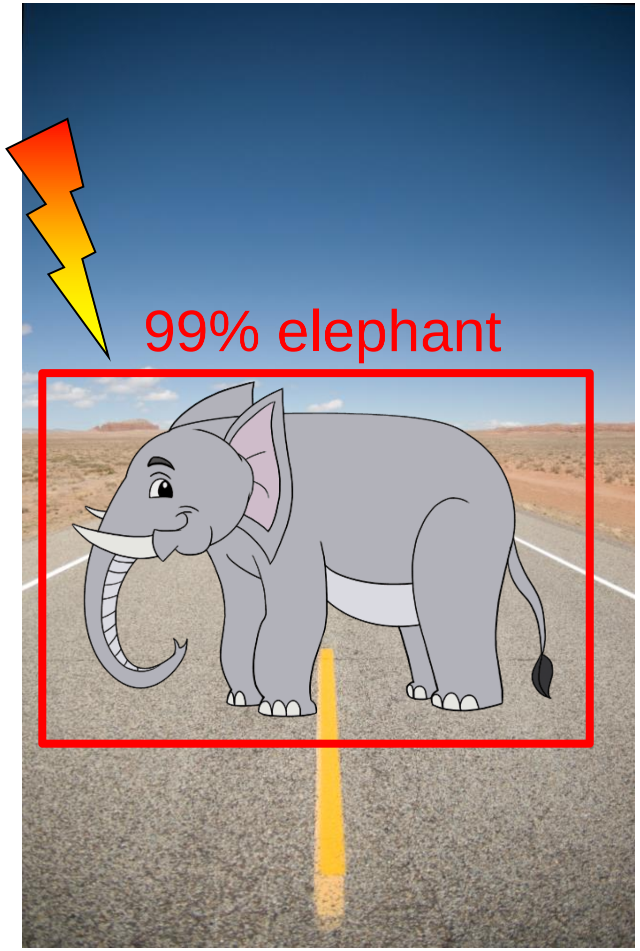
Critical

Missing an object

Examples of errors



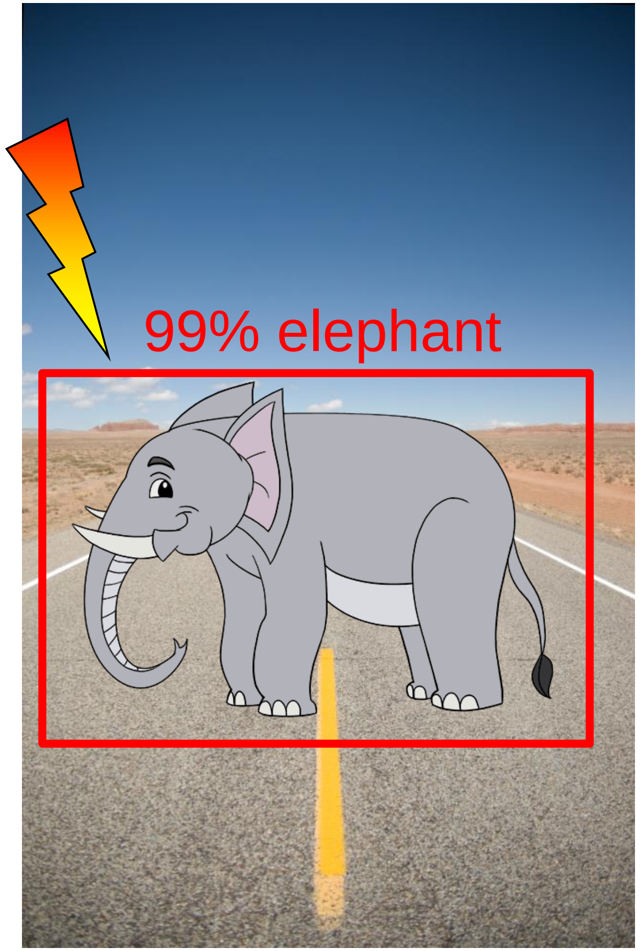
Examples of errors



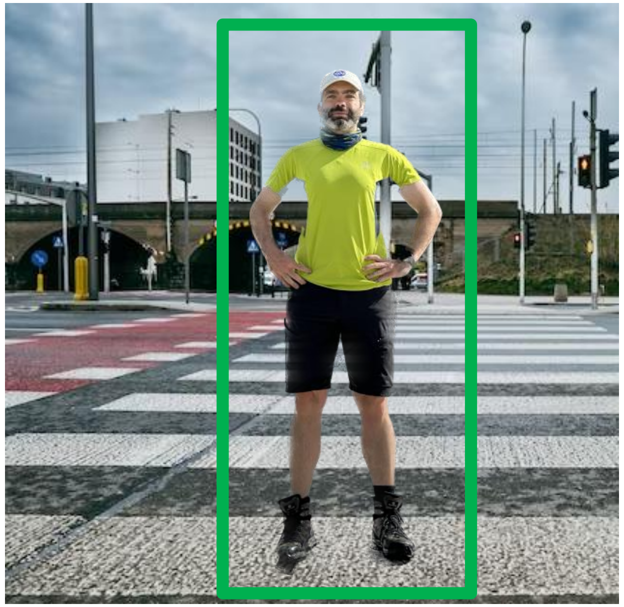
False positive

Unnecessary stops

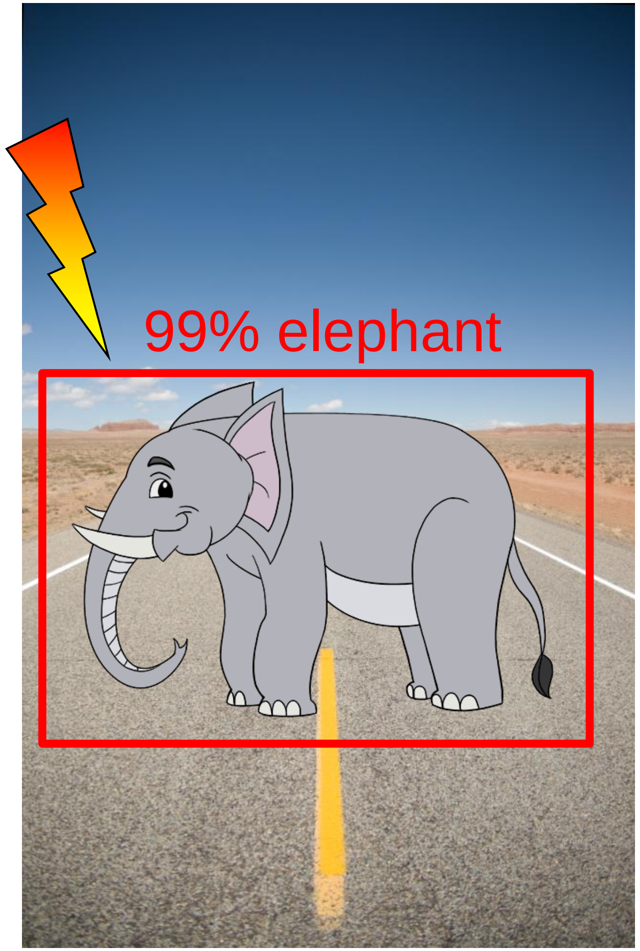
Examples of errors



False positive
Unnecessary stops

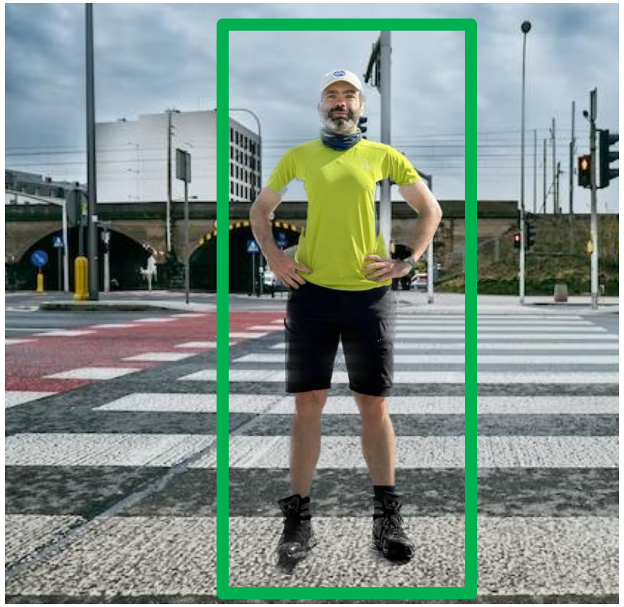


Examples of errors



False positive

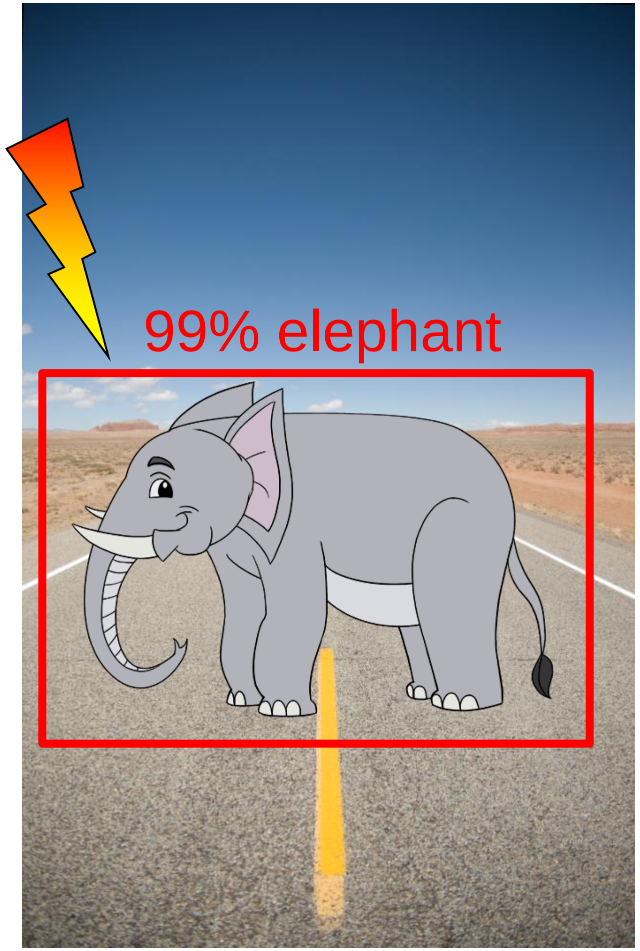
Unnecessary stops



Classification Error

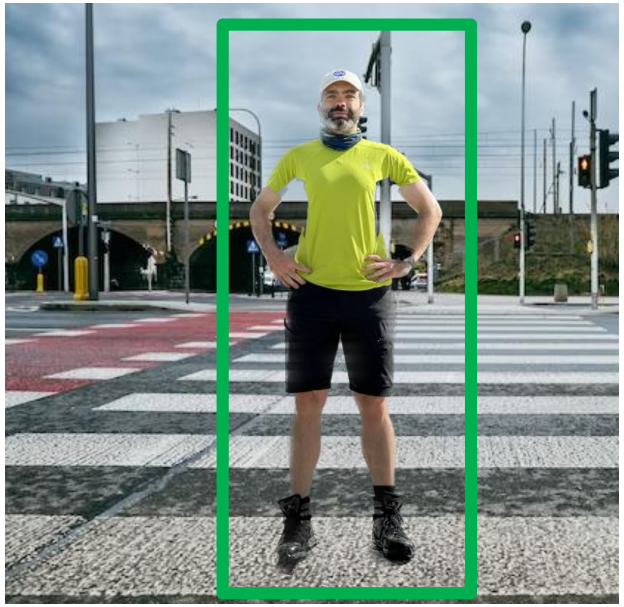
wrong object detected

Examples of errors



False positive

Unnecessary stops

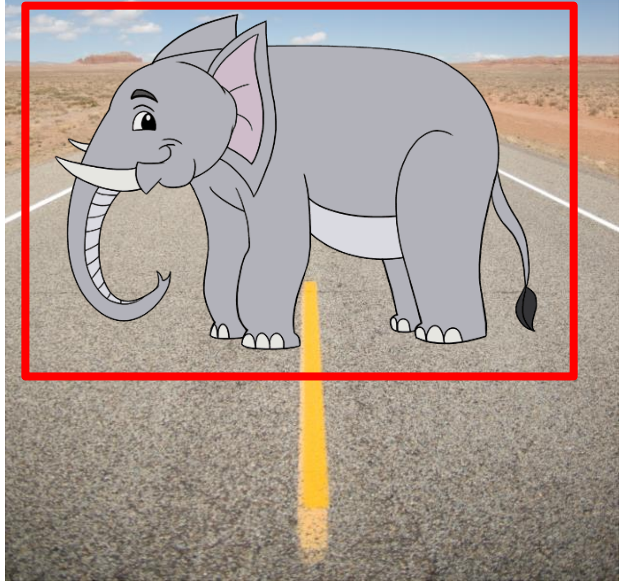


Classification Error

wrong object detected

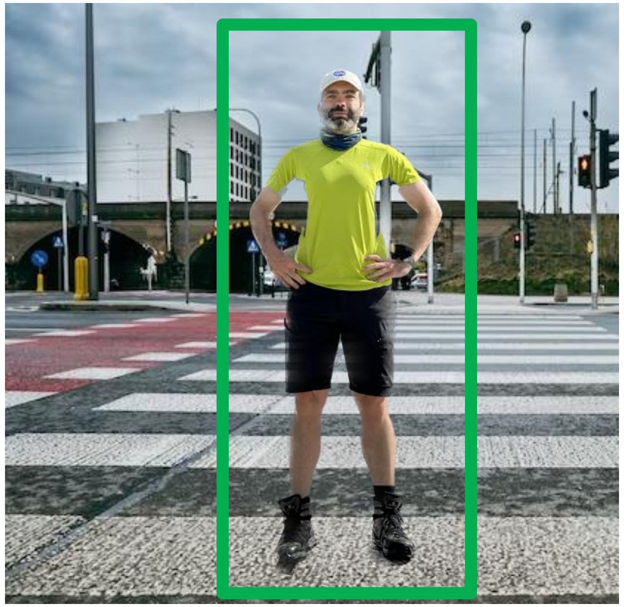
Examples of errors

99% elephant



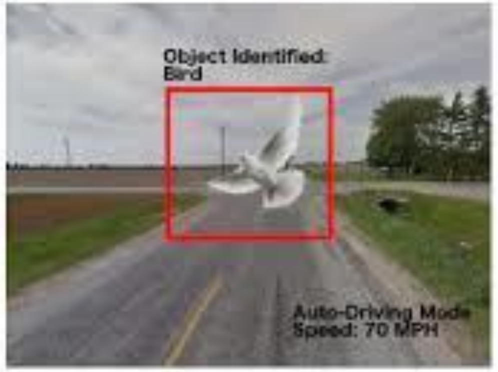
False positive

Unnecessary stops



Classification Error

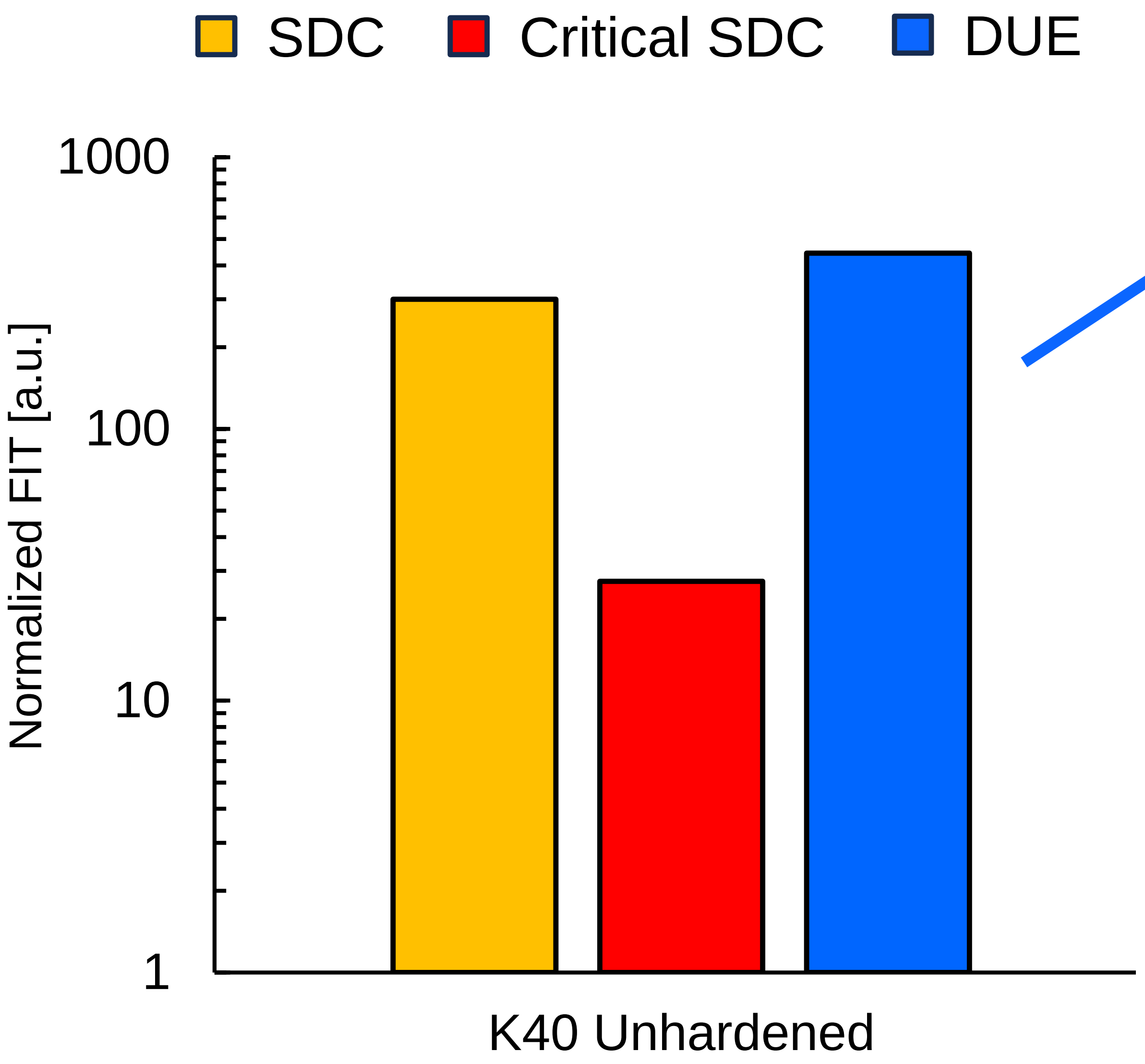
wrong object detected



*SC17 paper by BCU

Results – FIT rates

YOLO
@ChipIR



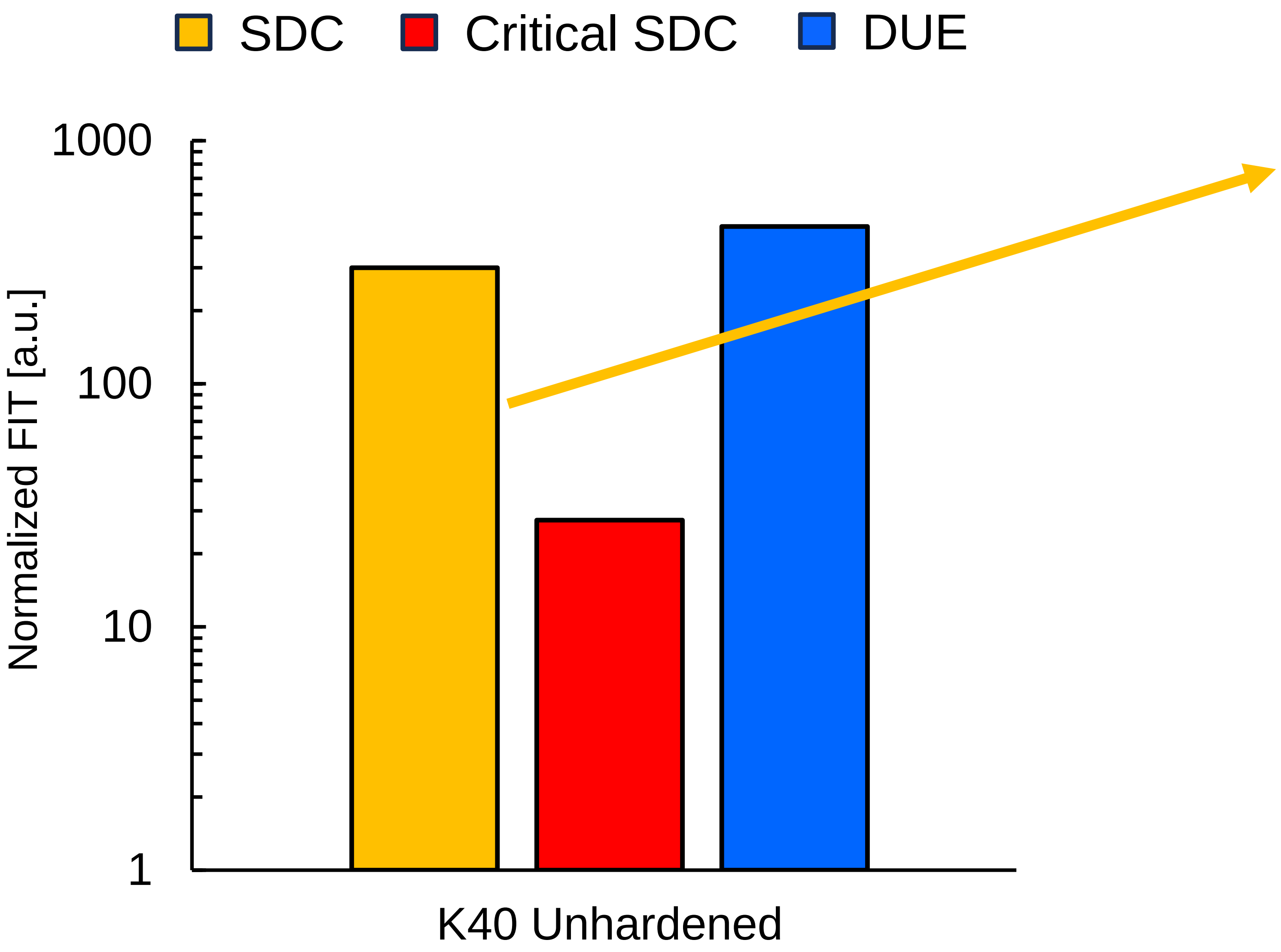
DUEs are the most common failures.

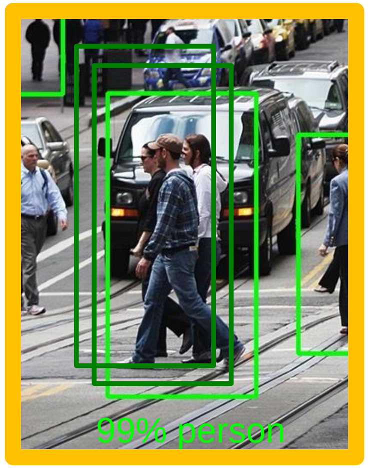
A setup with a very quick DUE detection and recovery is fundamental.

Booting up a GPU can take minutes… a minute lost is a lot of neutrons lost.

Results – FIT rates

YOLO
@ChipIR

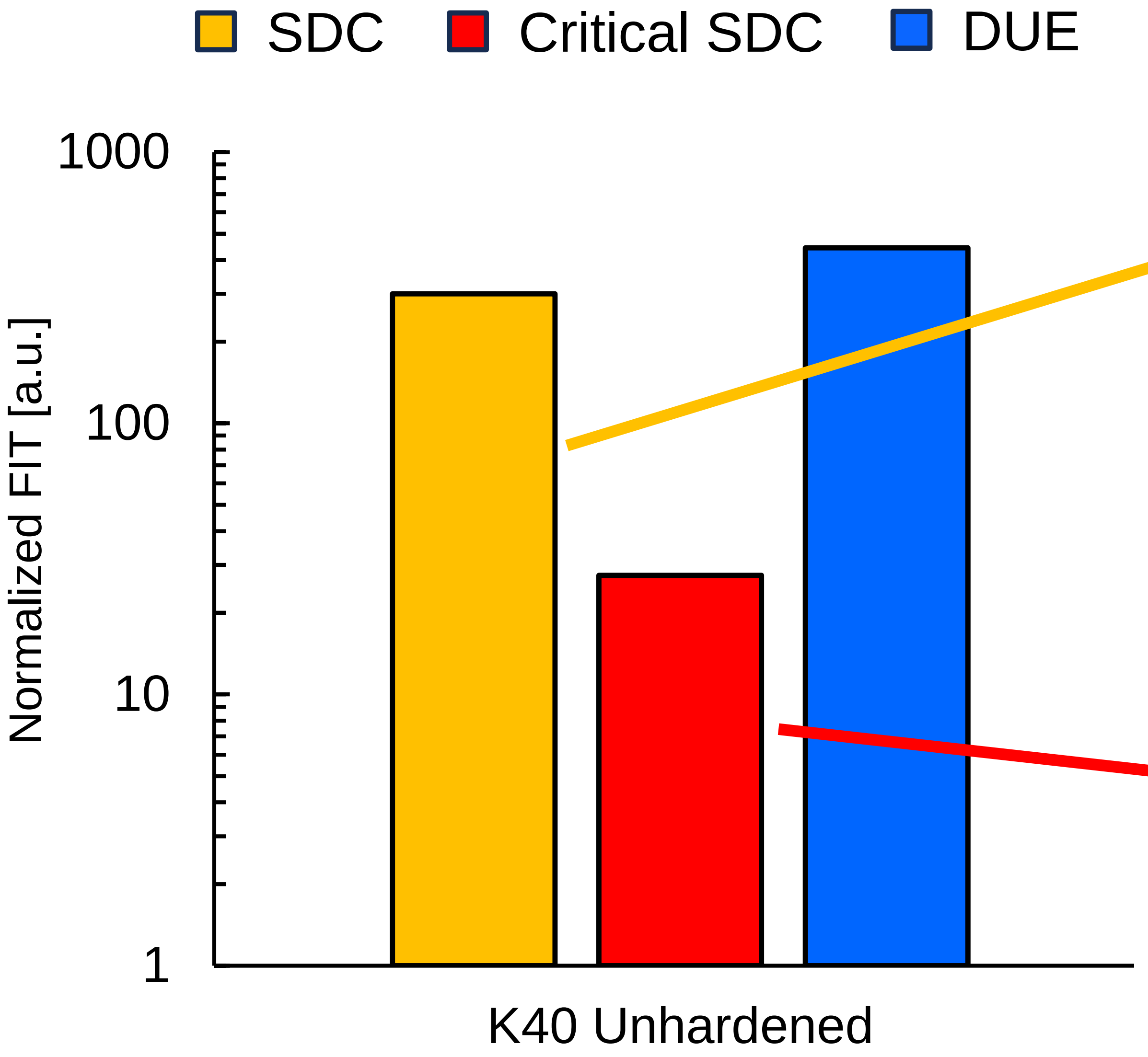


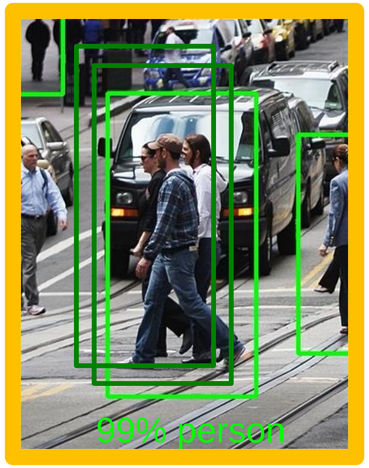


Tolerable

Results – FIT rates

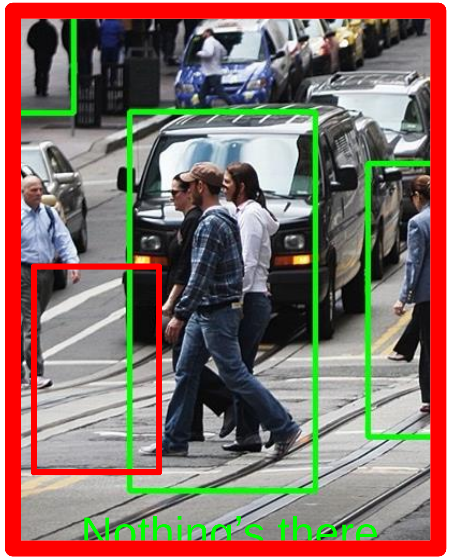
YOLO
@ChipIR

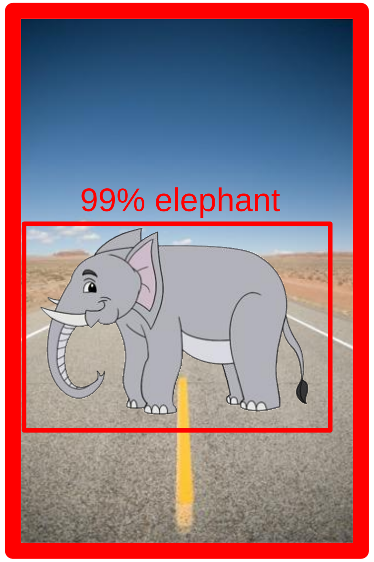




Tolerable

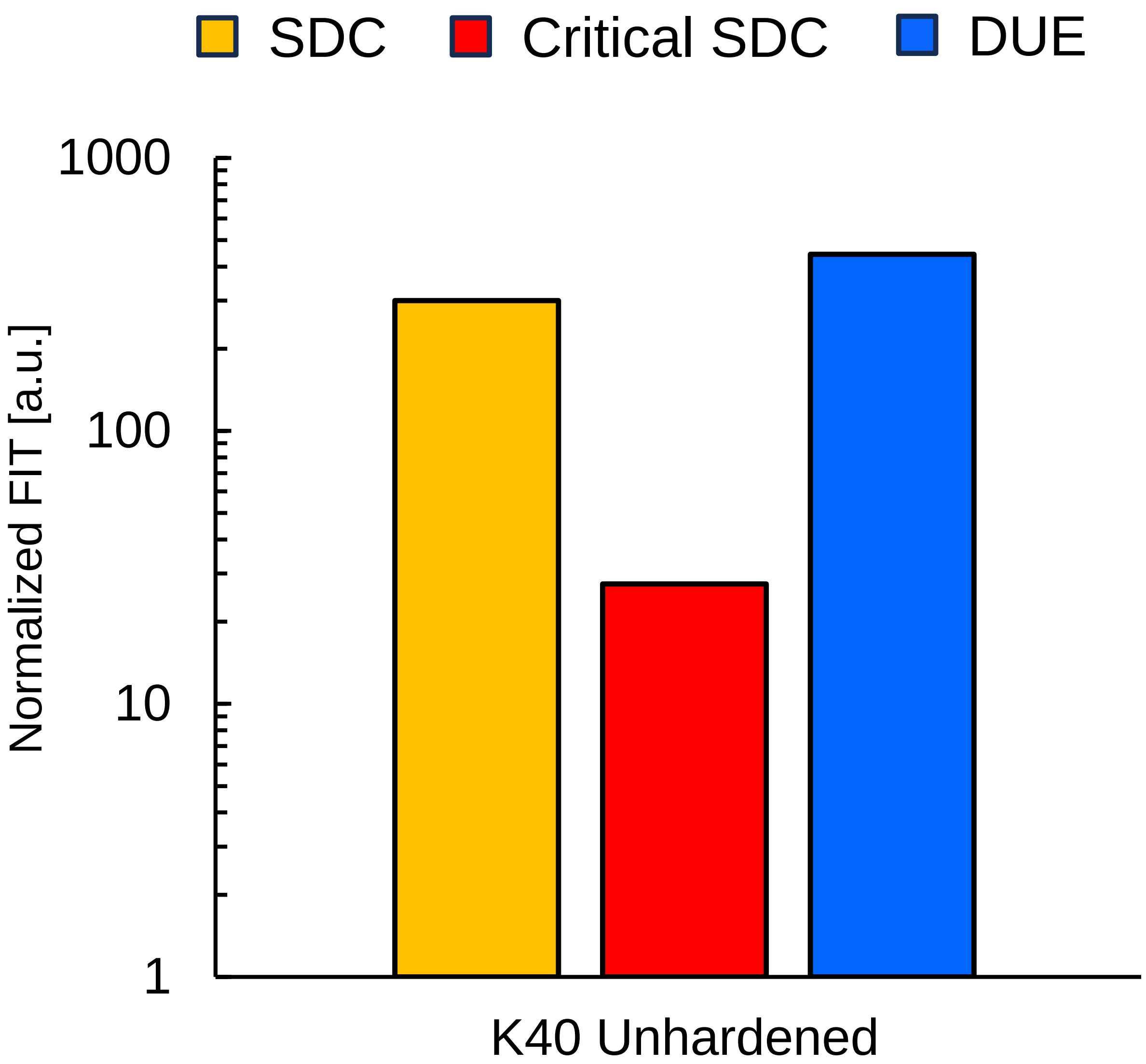
Critical





Results – FIT rates

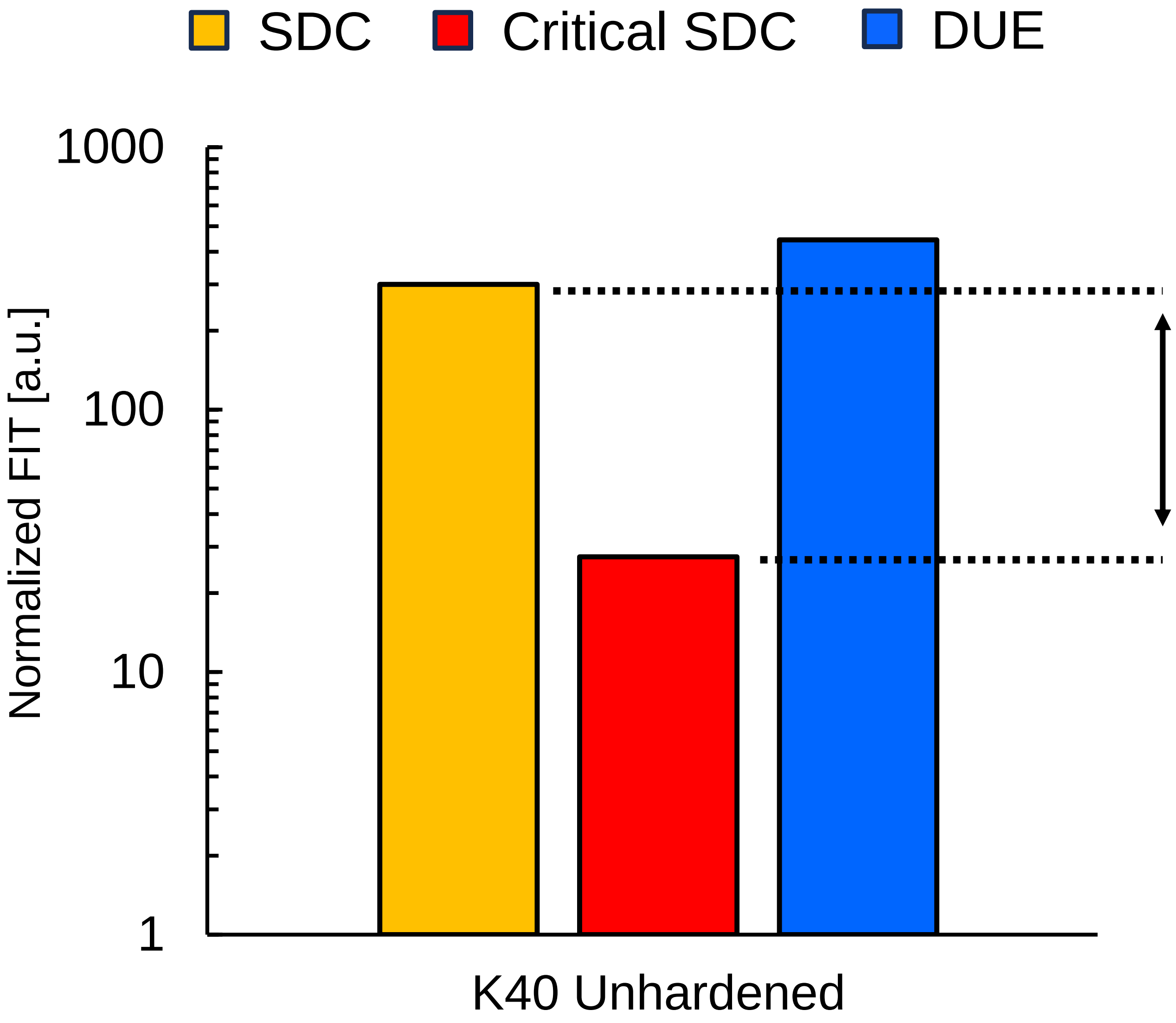
YOLO
@ChipIR



Not all SDCs affect detection/classification!

Results – FIT rates

YOLO
@ChipIR

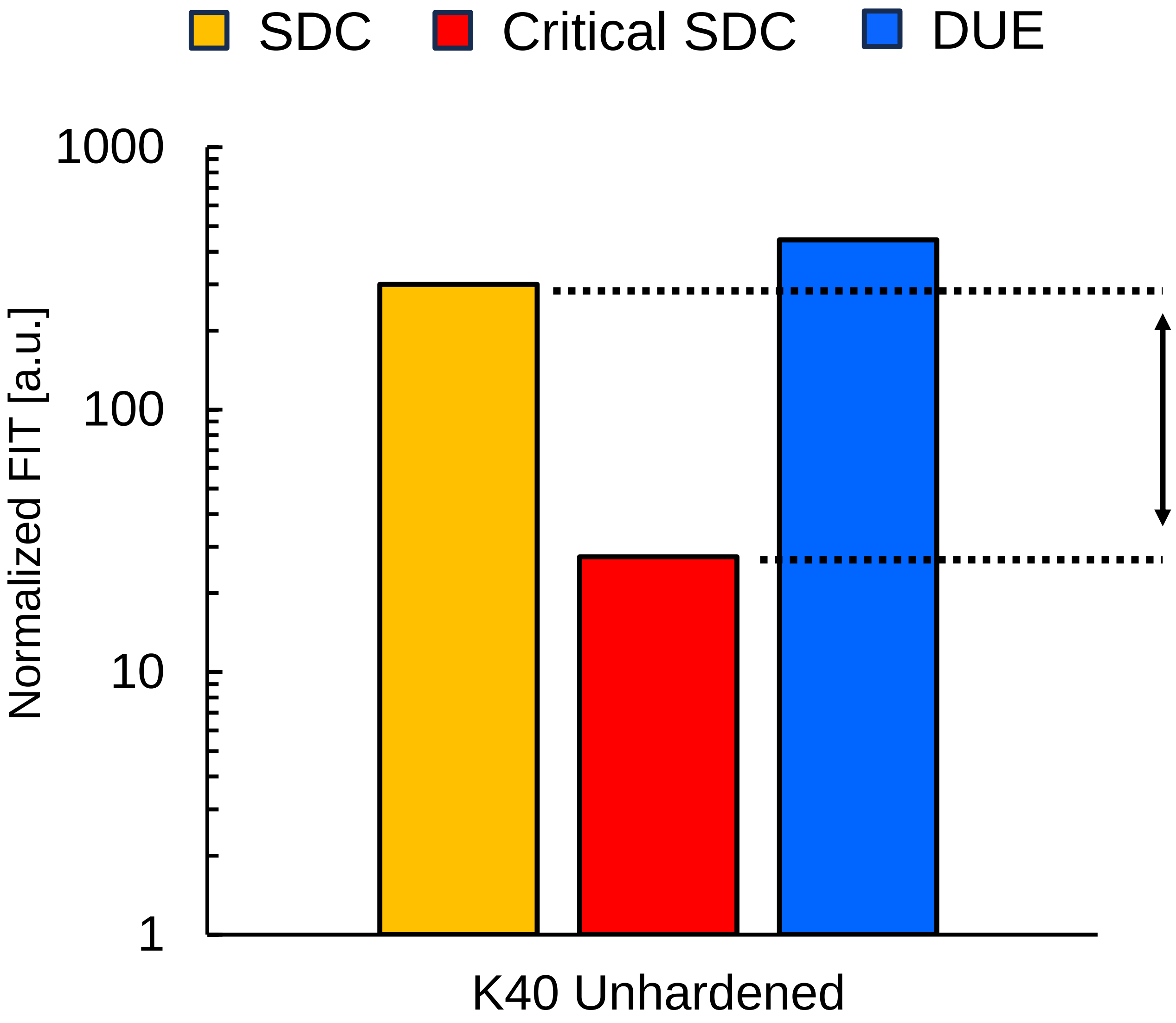


Not all SDCs affect detection/classification!

There is ~1 order of magnitude of difference between **SDCs** and **Critical SDCs**.

Results – FIT rates

YOLO
@ChipIR



Not all SDCs affect detection/classification!

There is ~1 order of magnitude of difference between SDCs and Critical SDCs.

The **fluence** to get a sufficient amount of SDCs can be quite high

Neutrons experiment challenges

TPU

GPU

HOST device (in beam room)

Neutrons experiment challenges

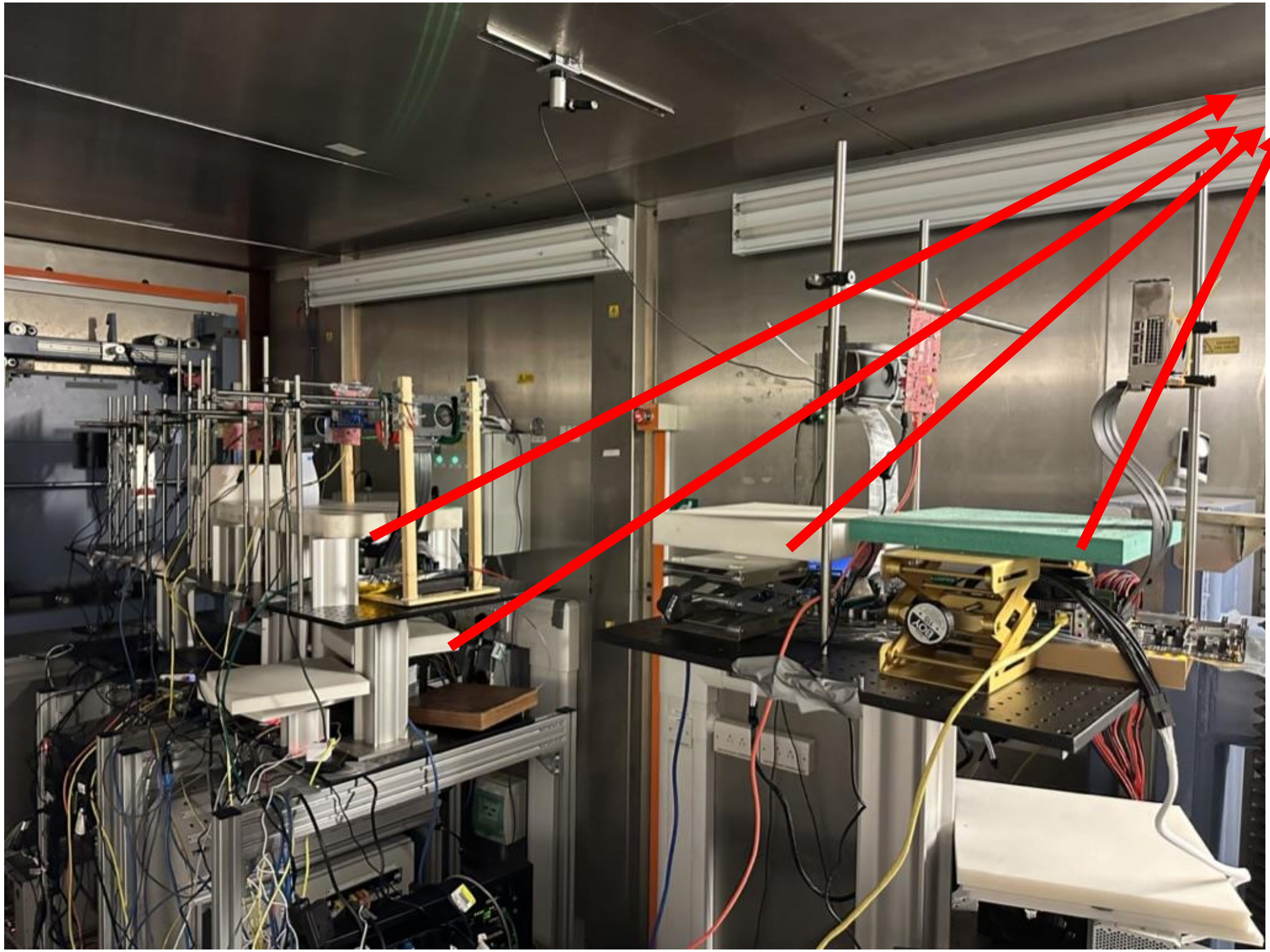
TPU

GPU

The host device must be protected.
Avoid scattering neutrons or thermal neutrons to corrupt the host

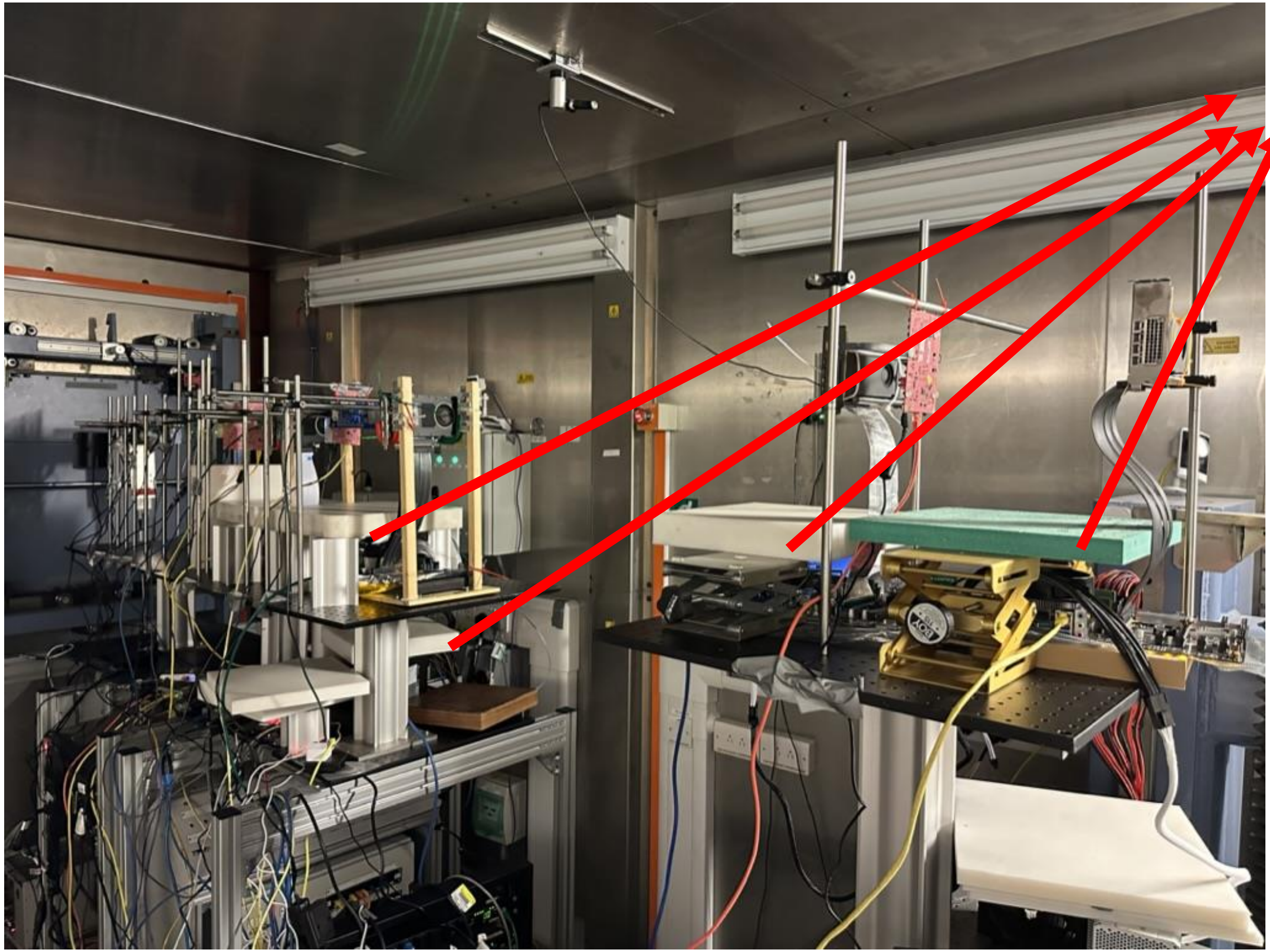
HOST device (in beam room)

Neutrons experiment challenges



Boron to (try to) protect
-motherboard
-SSD
-Raspberry (for TPU)
-Power supply

Neutrons experiment challenges



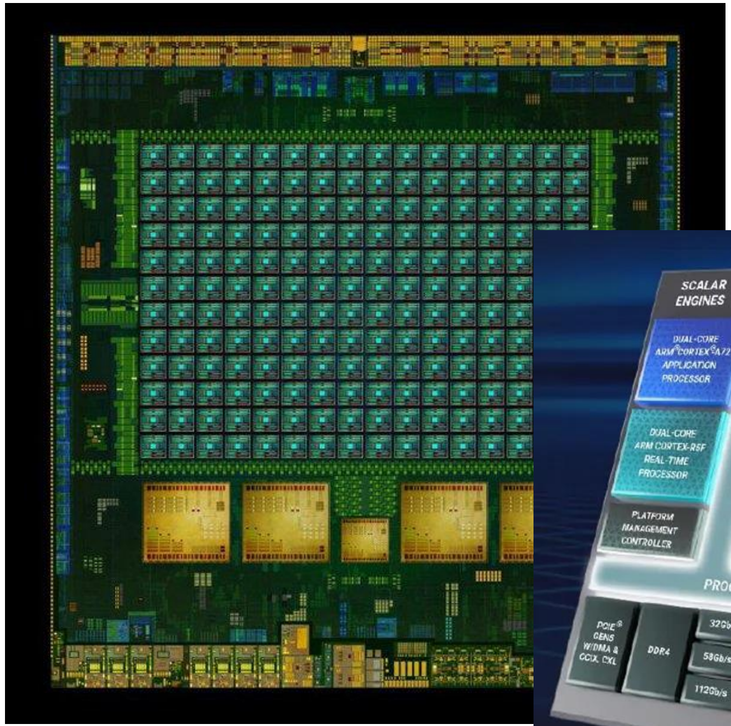
Boron to (try to) protect
-motherboard
-SSD
-Raspberry (for TPU)
-Power supply

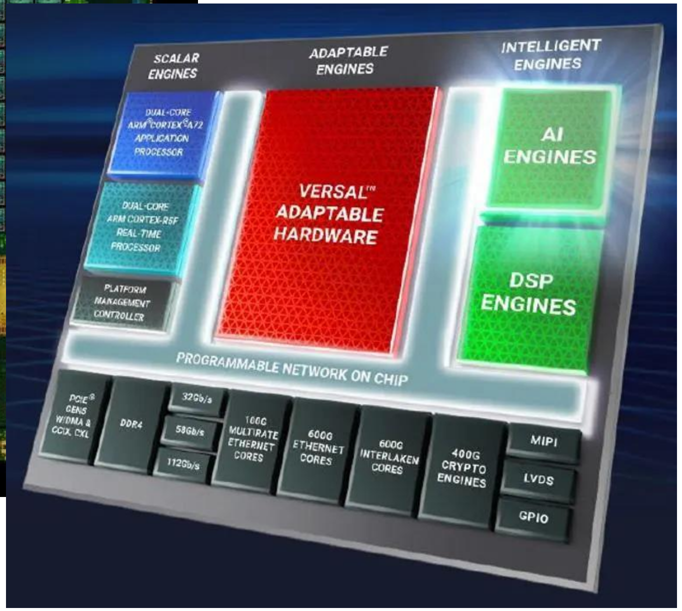
Still, we experience 2-3 failures per run…

…usually at 3am

Neutrons experiment challenges

SoC



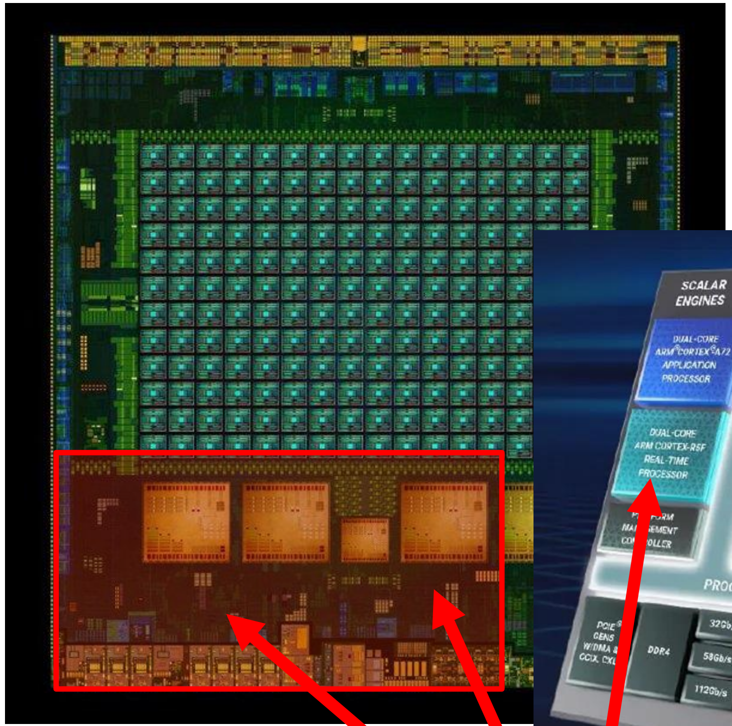


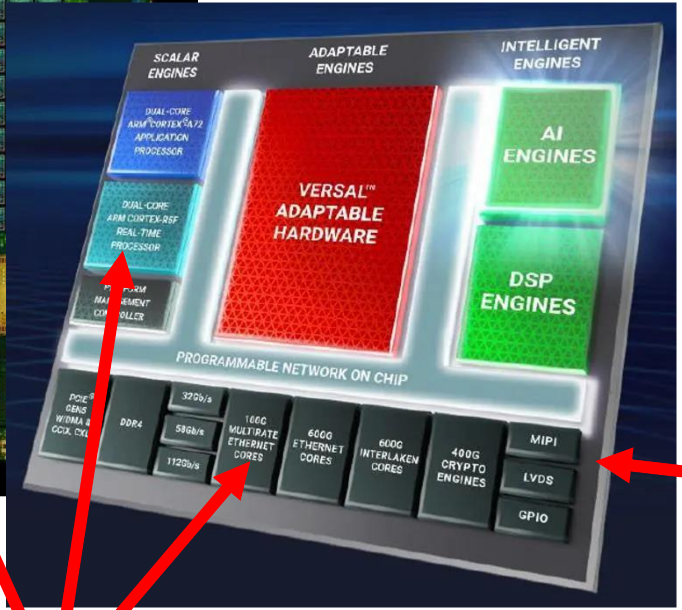
GPU

Neutrons experiment challenges

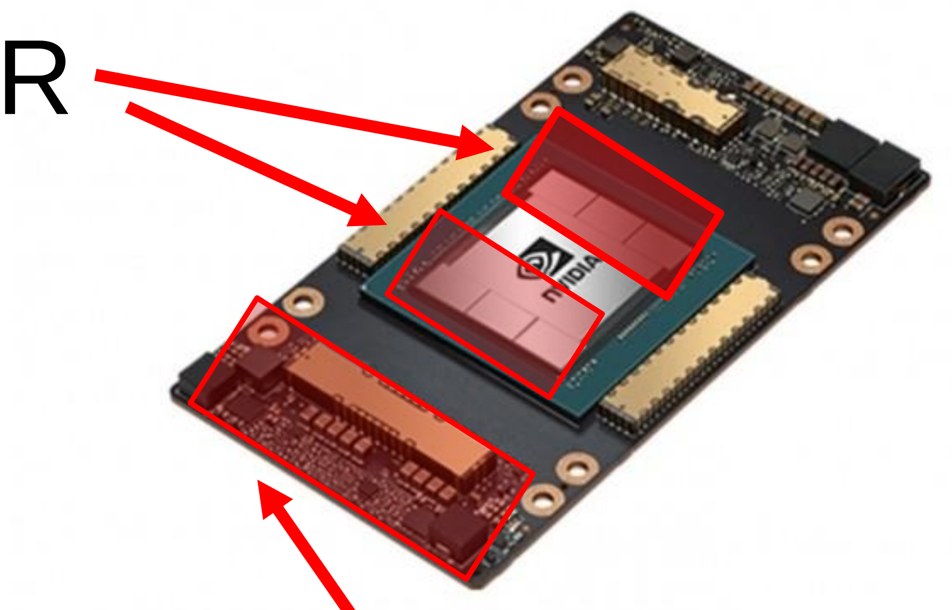
SoC

GPU





DDR

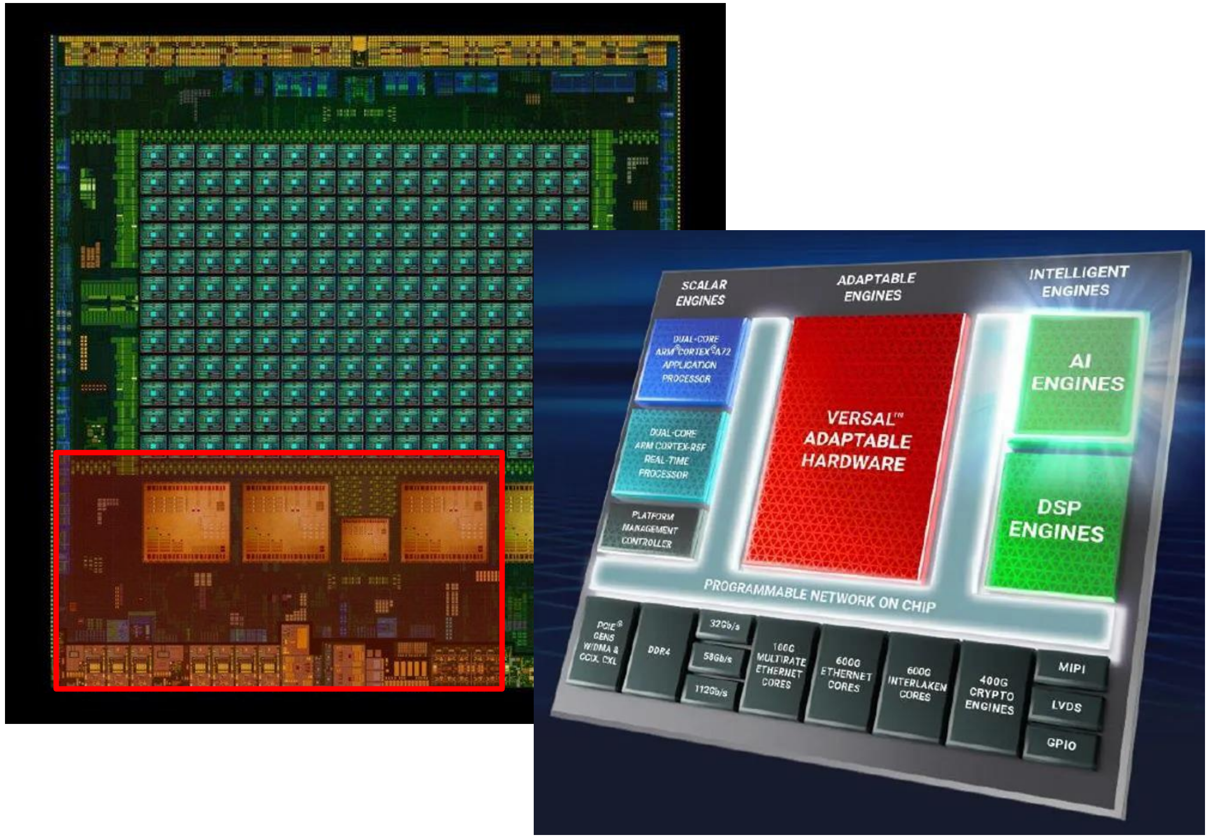


control-power circuitry

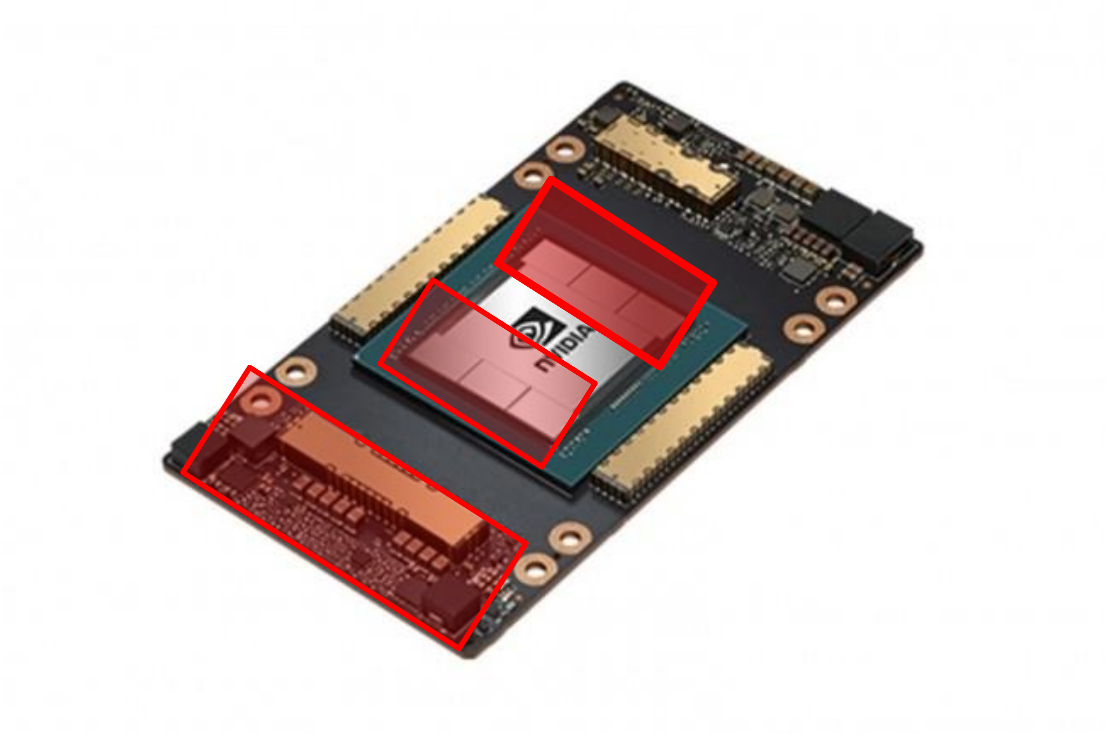
CPU(s)

Neutrons experiment challenges

SoC



GPU

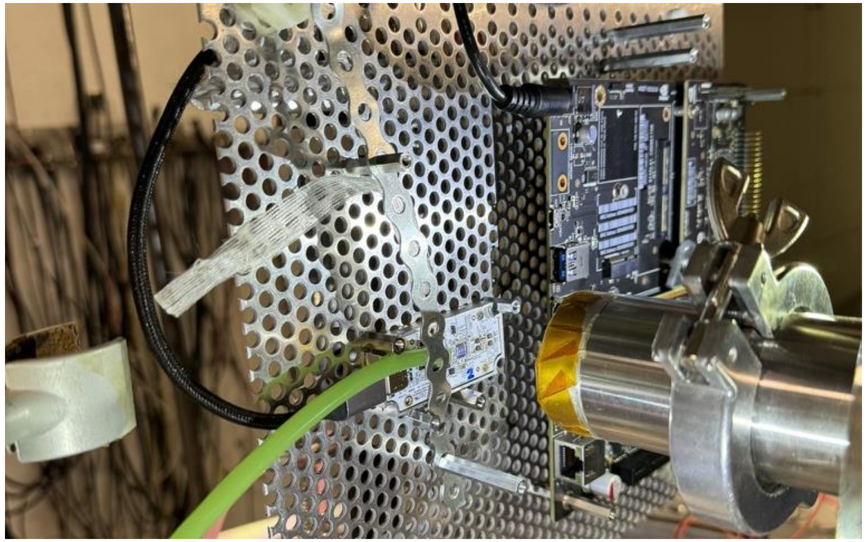


We need an extremely focused (neutron) beam

Neutrons vs Heavy Ions Experiment

ChipIR

RADEF



2x SSD/SD failures
2x Rasp4 failure
3x host DDR permanent failures
4x GPUs ECC/DDR problems

Multiple boards/tests
Relaxed environment

1x TPU stuck (solved by reboot)
1 board (thinned…)

Error rate tuneable
Quick beam shutdown

Take-away messages

-Reliability is a serious issue for safety-critical applications such as autonomous vehicles

-The SW is complex (and probabilistic), the HW is complex (and parallel)

-We need flexible beam and focused beam

-We need to focus on critical errors, critical variables, critical resources to have efficient hardening

-Novel technologies (HW, SW) are continuously being developed, we need to test them

What’s next?

– Generalization.

We test one model, trained in one way, with one data set, with few frames…
What if something changes? We need to do it all over again.

What’s next?

– Generalization.

We test one model, trained in one way, with one data set, with few frames… What if something changes? We need to do it all over again.

– Be at speed with novelty.

ML community is faster than us. What is novel to us, it’s obsolete to them. Segmentation, Transformer… trillion parameters neural network… QNN

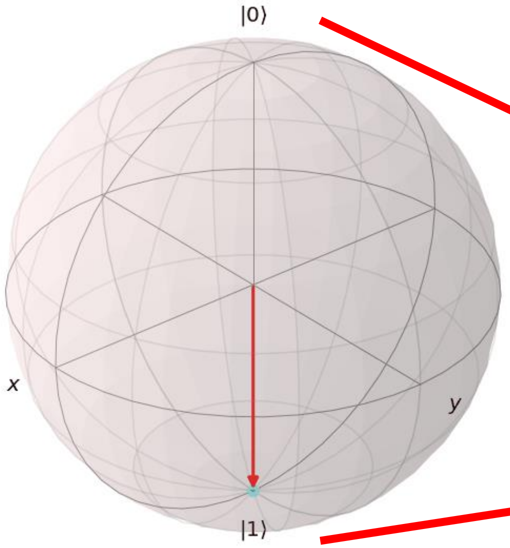
What’s next?

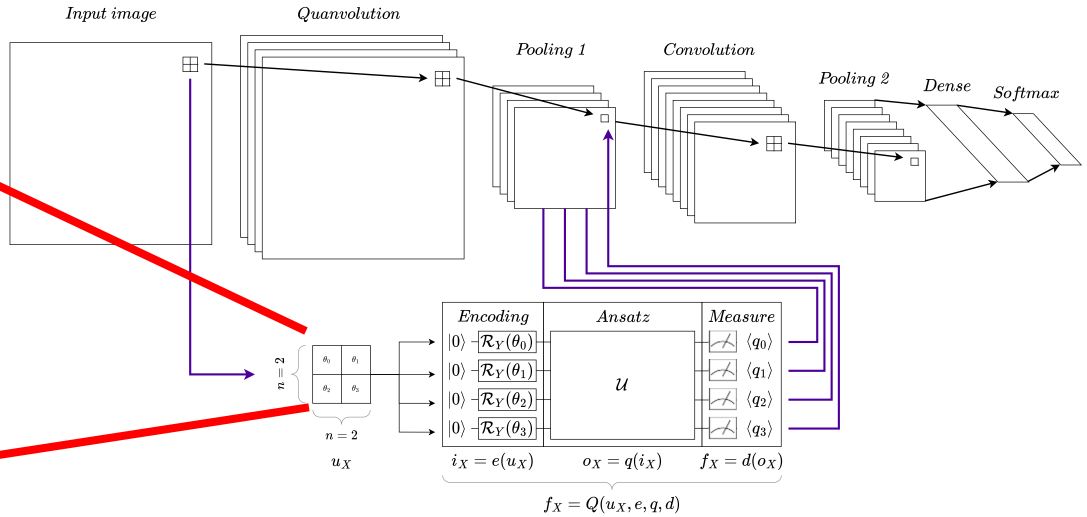
– Generalization.

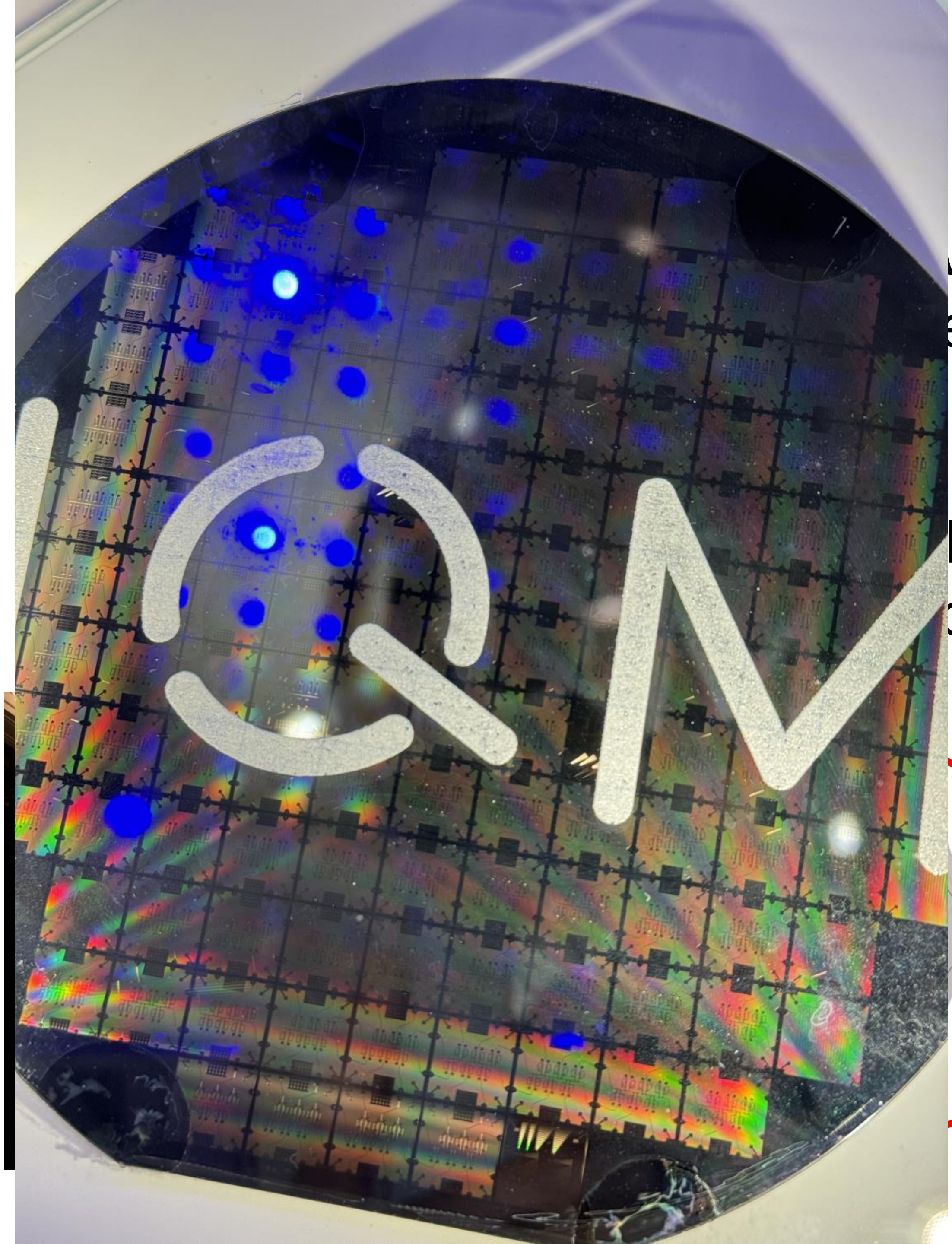
We test one model, trained in one way, with one data set, with few frames… What if something changes? We need to do it all over again.

– Be at speed with novelty.

ML community is faster than us. What is novel to us, it’s obsolete to them. Segmentation, Transformer… trillion parameters neural network… QNN

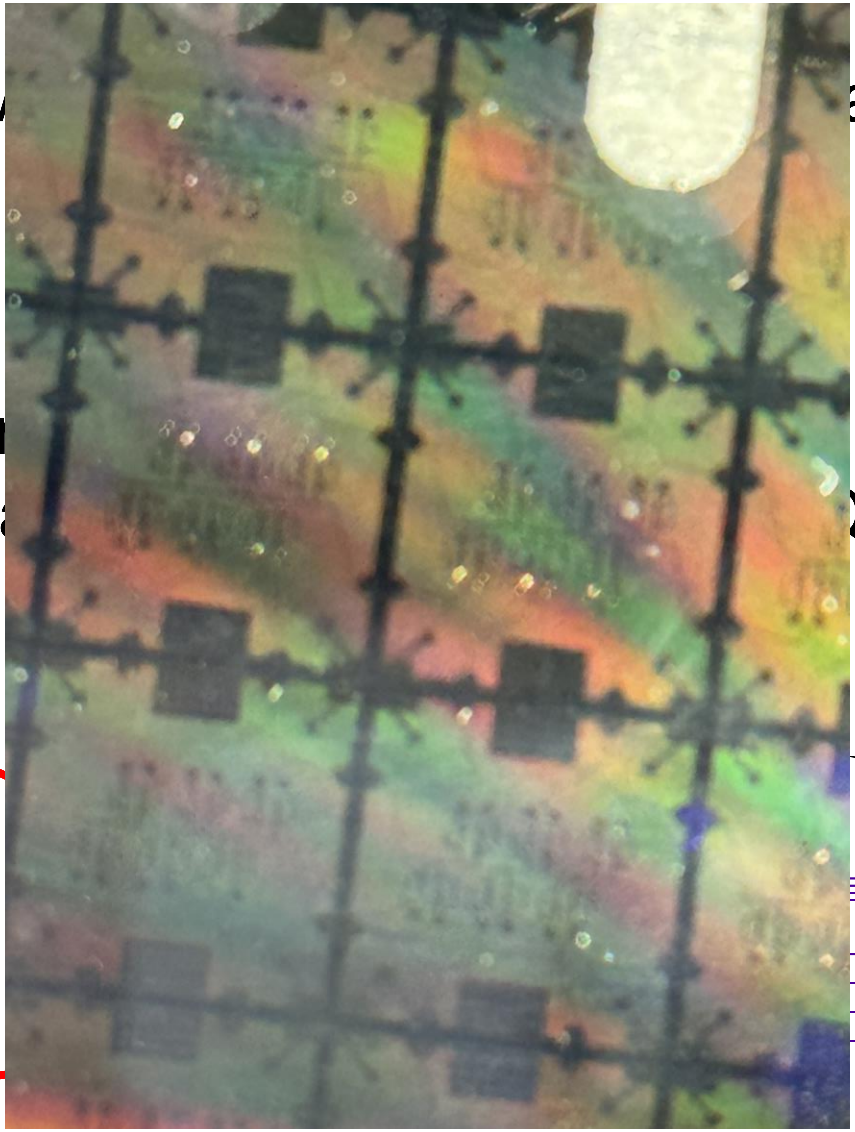


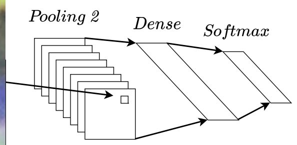




way, w ames…
eed to

hat is r hem.
on para QNN





What’s next?

– Generalization.

We test one model, trained in one way, with one data set, with few frames… What if something changes? We need to do it all over again.

– Be at speed with novelty.

ML community is faster than us. What is novel to us, it’s obsolete to them. Segmentation, Transformer… trillion parameters neural network… QNN

– HW to fit the model.

When the model changes, the HW needs to be improved to fit the model, but with a big delay. How can we predict the effect of the HW on the software.

What's next?

– Generalization.

We test one model, trained in one way, with one data set, with few frames… What if something changes? We need to do it all over again.

– Be at speed with novelty.

ML community is faster than us. What is novel to us, it's obsolete to them. Segmentation, Transformer… trillion parameters neural network… QNN

– HW to fit the model.

When the model changes, the HW needs to be improved to fit the model, but with a big delay. How can we predict the effect of the HW on the software.

– Avoid to be a Mickey Mouse.

We need to have control over what is really executed in HW.

Acknowledgements

Dario Petri
Flavio Vella
Matteo Saveriano
Marzio Vallero
Bruno Coelho
Gioele Casagranda

Heather Quinn
Elizabeth Auden
Thomas Fairbanks
Nathan DeBardeleben
Sean Blanchard
Steve Wender
Gus Sinnis

Chris Frost
Carlo Cazzaniga
Maria Kastriotou
ISIS User Office

Timothy Tsai
Siva Hari
Michael Sullivan
Steve Keckler

Matteo Sonza Reorda
Luca Sterpone
Edoardo Giusto
Emanuel Dri

Caio Lunardi
Daniel Oliveira
Pablo Bodmann
Philippe Navaux
Luigi Carro

Pete Harrold

Fernando F. Santos
Marcello Traiola
Angeliki Kritikakou

Acknowledgements

Dario Petri
Flavio Vella
Matteo Saveriano
Marzio Vallero
Bruno Coelho

Heather Quinn
Elizabeth Auden
Thomas Fairbanks
Nathan DeBardeleben
Sean Blanchard

ISIS User Office

Steve Keckler

Emanuel Dri

Caio Lunardi
Daniel Oliveira
Pablo Bodmann
Philippe Navaux
Luigi Carro

Pete Harrold

Fernando F. Santos
Marcello Traiola
Angeliki Kritikakou