Reliability Evaluation of Mixed-Precision Architectures

Fernando F. Santos, Caio Lunardi, Daniel Oliveira, Fabiano Libano, and Paolo Rech

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

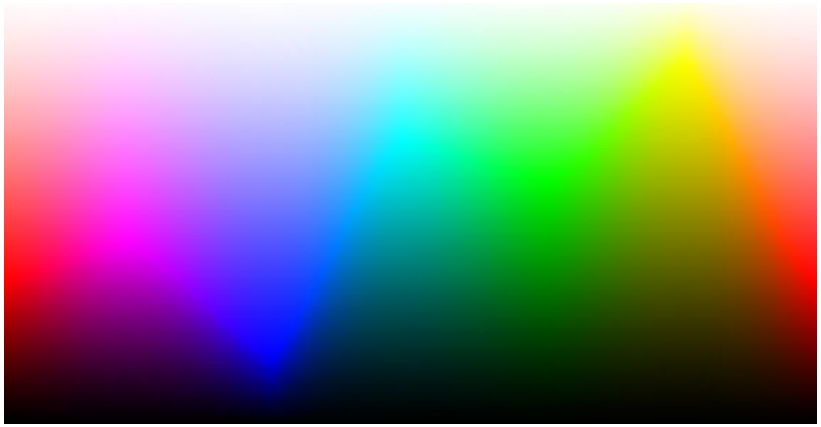
.inf
INSTITUTO DE INFORMÁTICA
UFRGS

Mixed-Precision Data and Operations

Double (64 bit)

Single (32 bit)

floating point operations increased **graphics quality** and **user experience**…

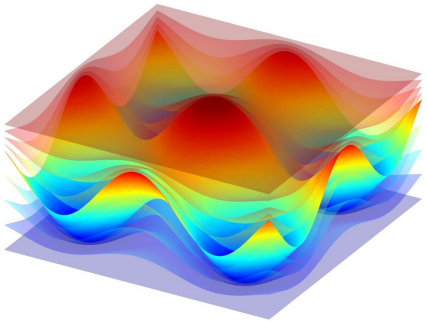


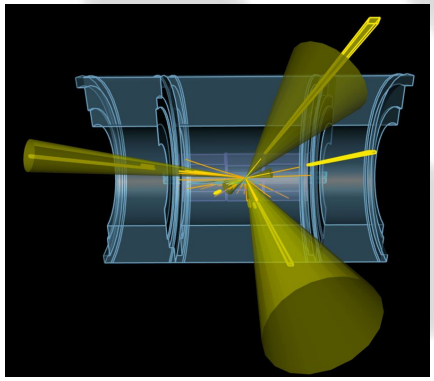
Mixed-Precision Data and Operations

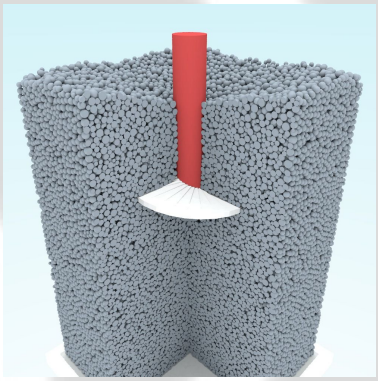
Double (64 bit)

Single (32 bit)

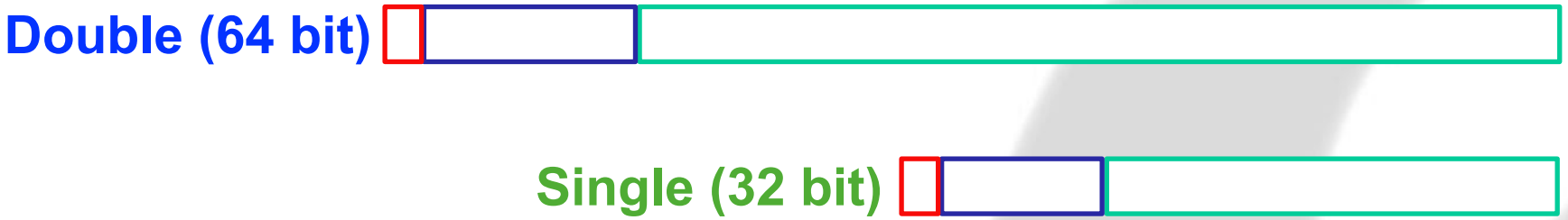
floating point operations increased **graphics quality** and **user experience**…
as well as **physical simulation precision** and **accuracy**

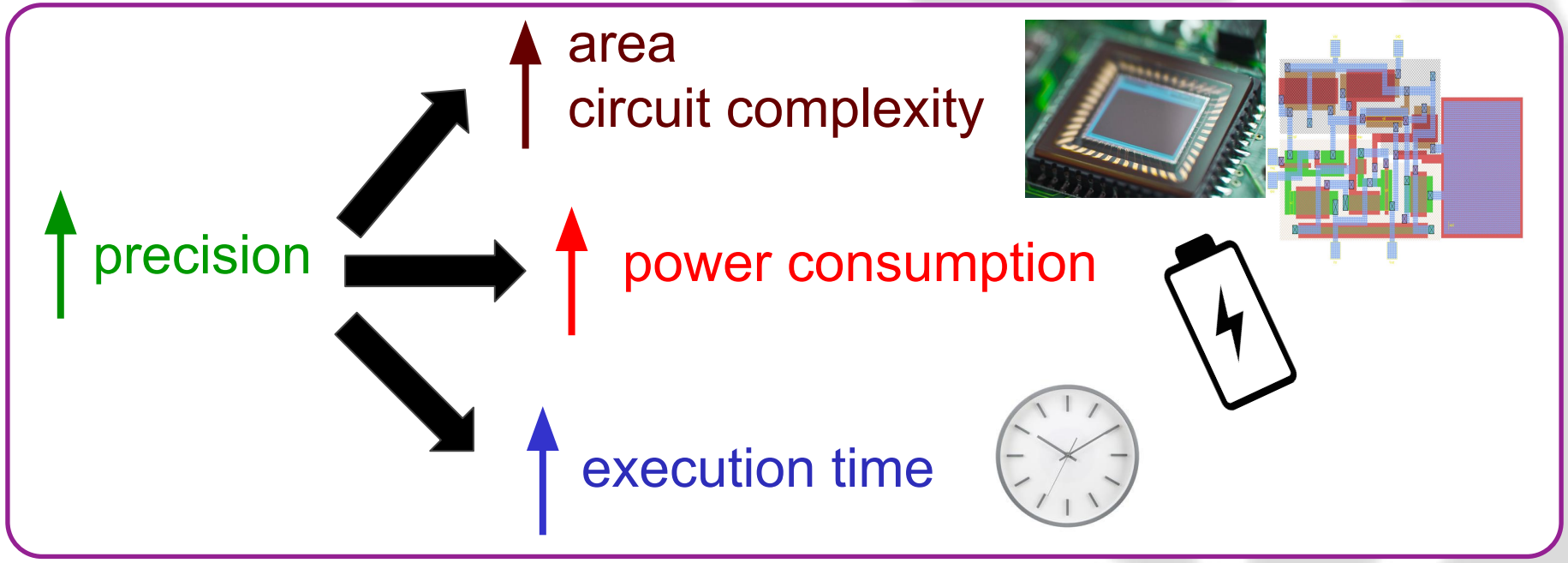




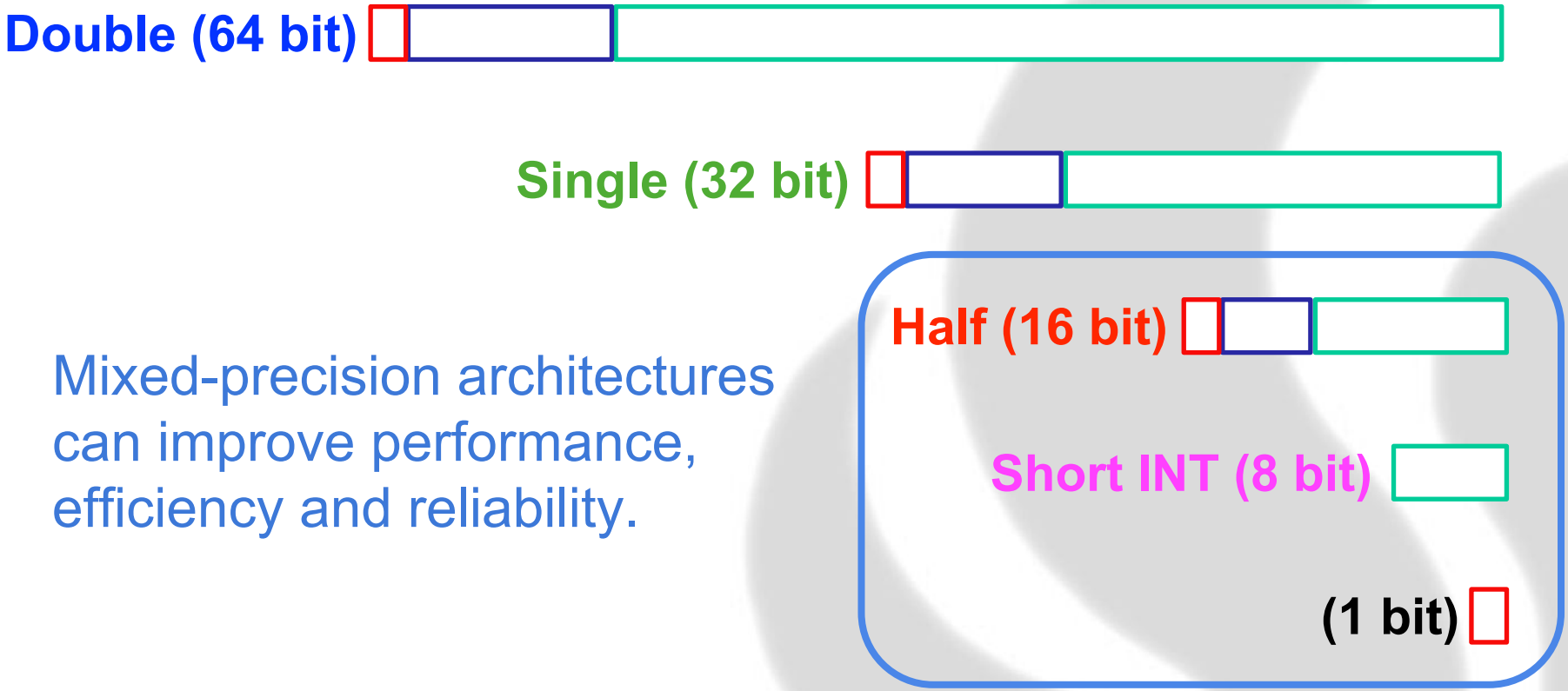


Mixed-Precision Data and Operations





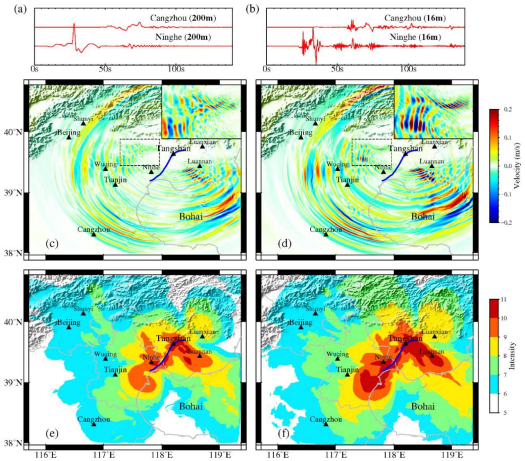
Mixed-Precision Data and Operations



Mixed-precision architectures can improve performance, efficiency and reliability.

Mixed-Precision Data and Operations

Approximate computing has shown that some operations in some algorithms can be approximated without affecting significantly the final result



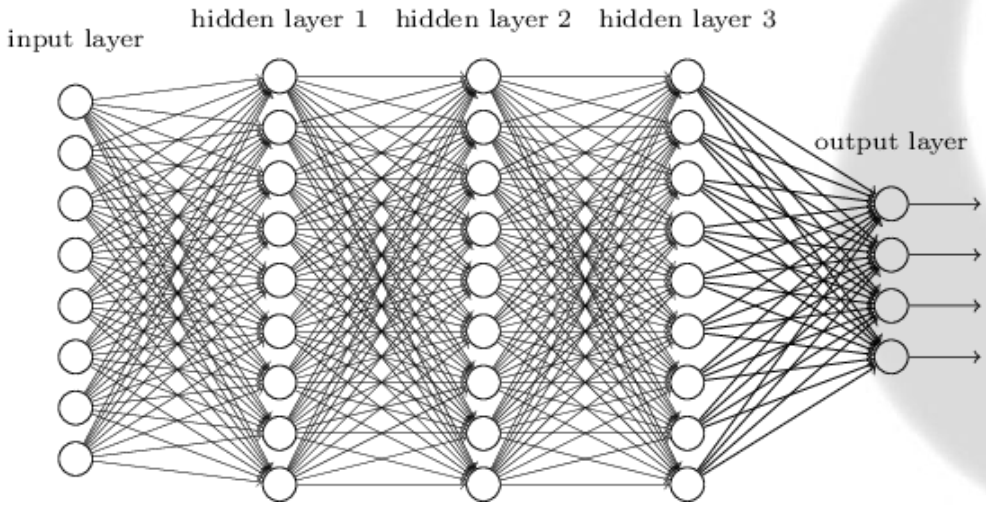
Half (16 bit)

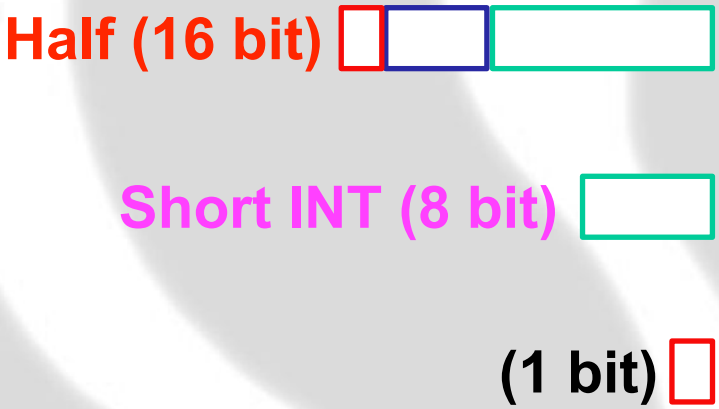
Short INT (8 bit)

(1 bit)

Mixed-Precision Data and Operations

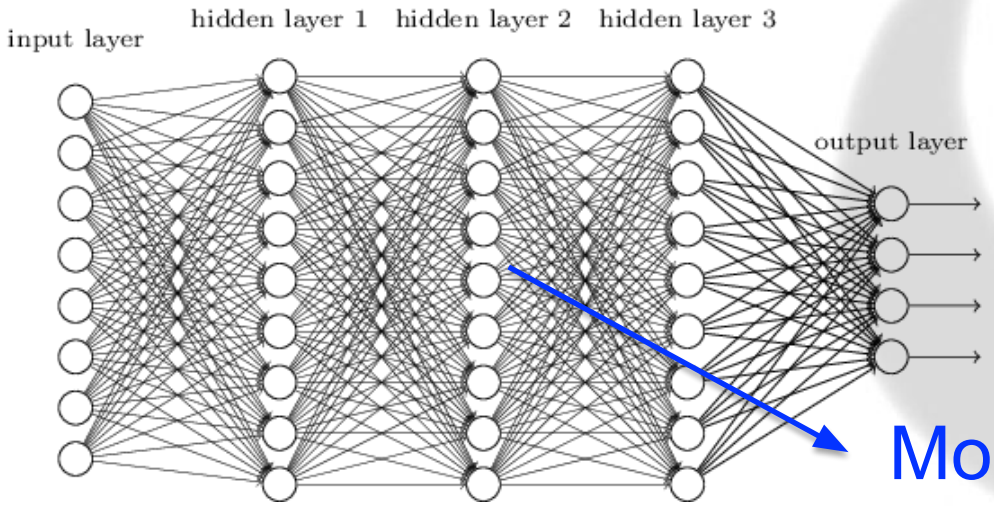
Reduced precision operations are particularly interesting for **Neural Networks training and execution**

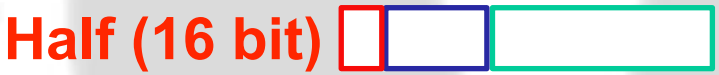




Mixed-Precision Data and Operations

Reduced precision operations are particularly interesting for **Neural Networks training and execution**





Most of CNNs weights can be represented in half

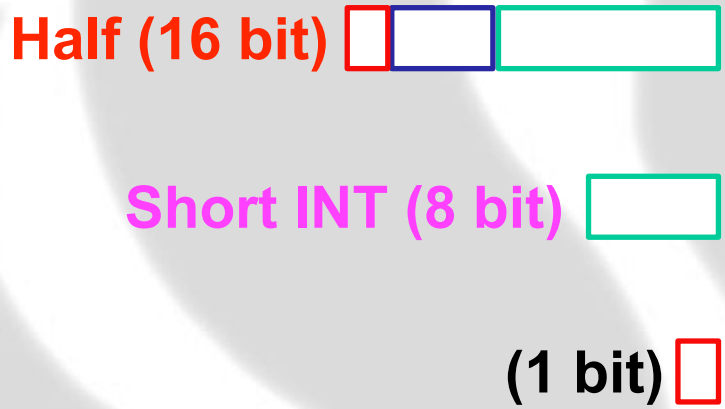
Mixed-Precision Data and Operations

Reduced precision operations are particularly interesting for **Neural Networks training and execution**

Neural Nets accuracy*

Network	FP32 Baseline	Mixed precision
AlexNet	56.8%	56.9%
VGG-D	65.4%	65.4%
GoogLeNet	68.3%	68.4%
Inception v2	70.0%	70.0%
Inception v3	73.9%	74.1%
Resnet 50	75.9%	76.0%
ResNeXt 50	77.3%	77.5%

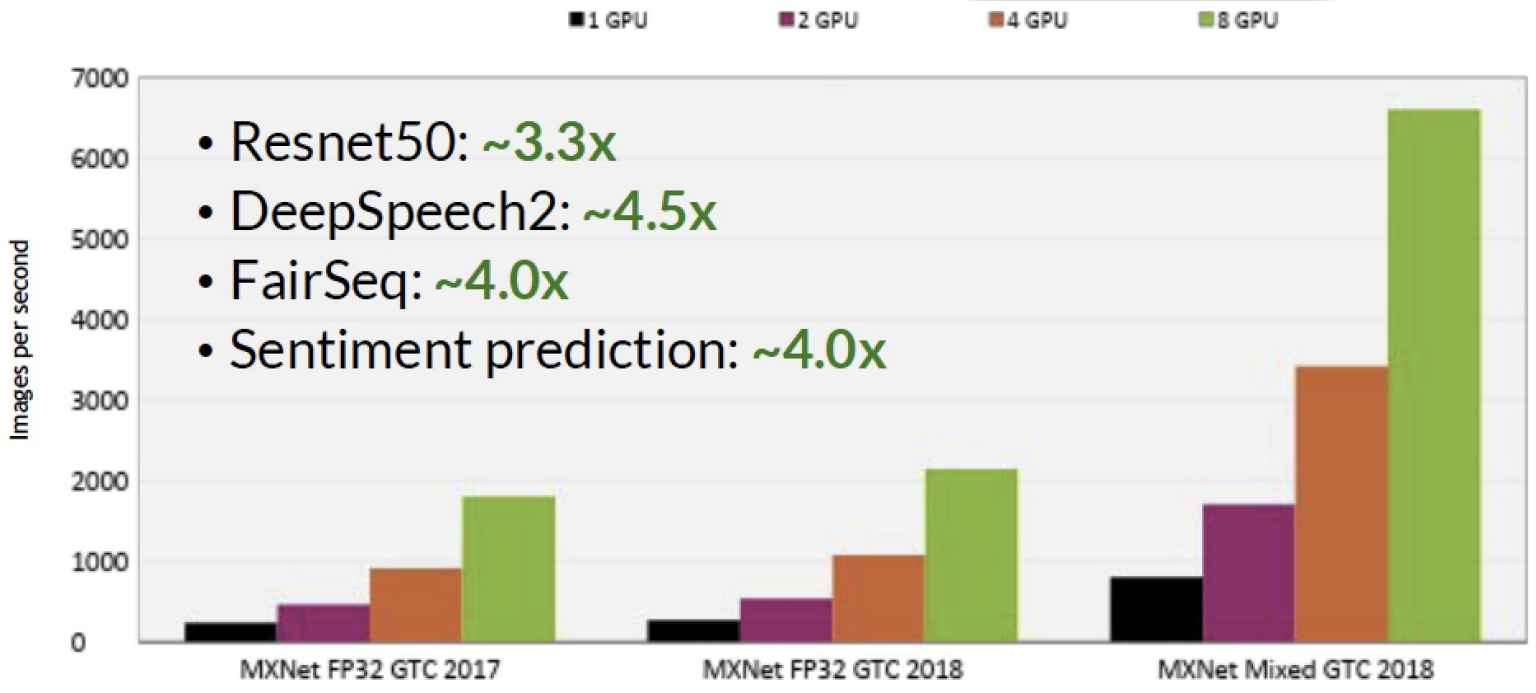
*data from nvidia



NVIDIA

Mixed-Precision Data and Operations

Reduced precision operations are particularly interesting for **Neural Networks training and execution**

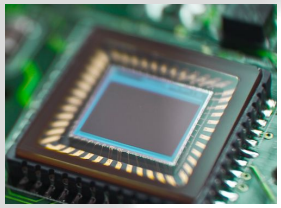


*data from nvidia

Reliability of Mixed-Precision

Mixed-Precision delivers:

- -Higher **Performance**
- -Smaller **Area**
- -Lower **Power Consumption**

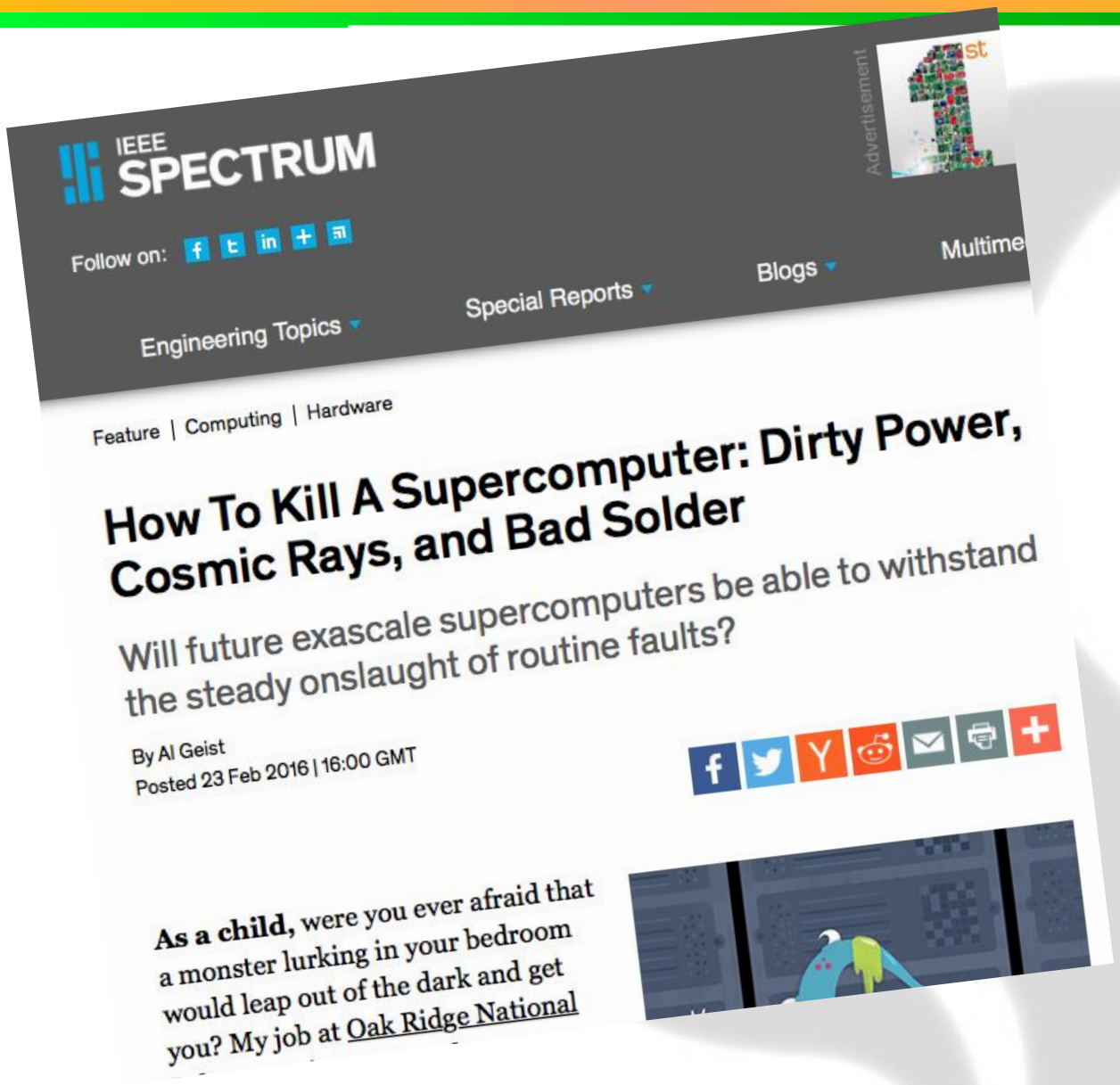


But how does it affect overall system Reliability?

Outline

- Radiation Effects Introduction
- Experimental Methodology
- Reliability of Mixed-Precision Architectures
  - FPGAs
  - x86
  - GPUs
- Conclusions

The Importance of Reliability



IEEE SPECTRUM

Follow on:

Engineering Topics | Special Reports | Blogs | Multime

Advertisement

Feature | Computing | Hardware

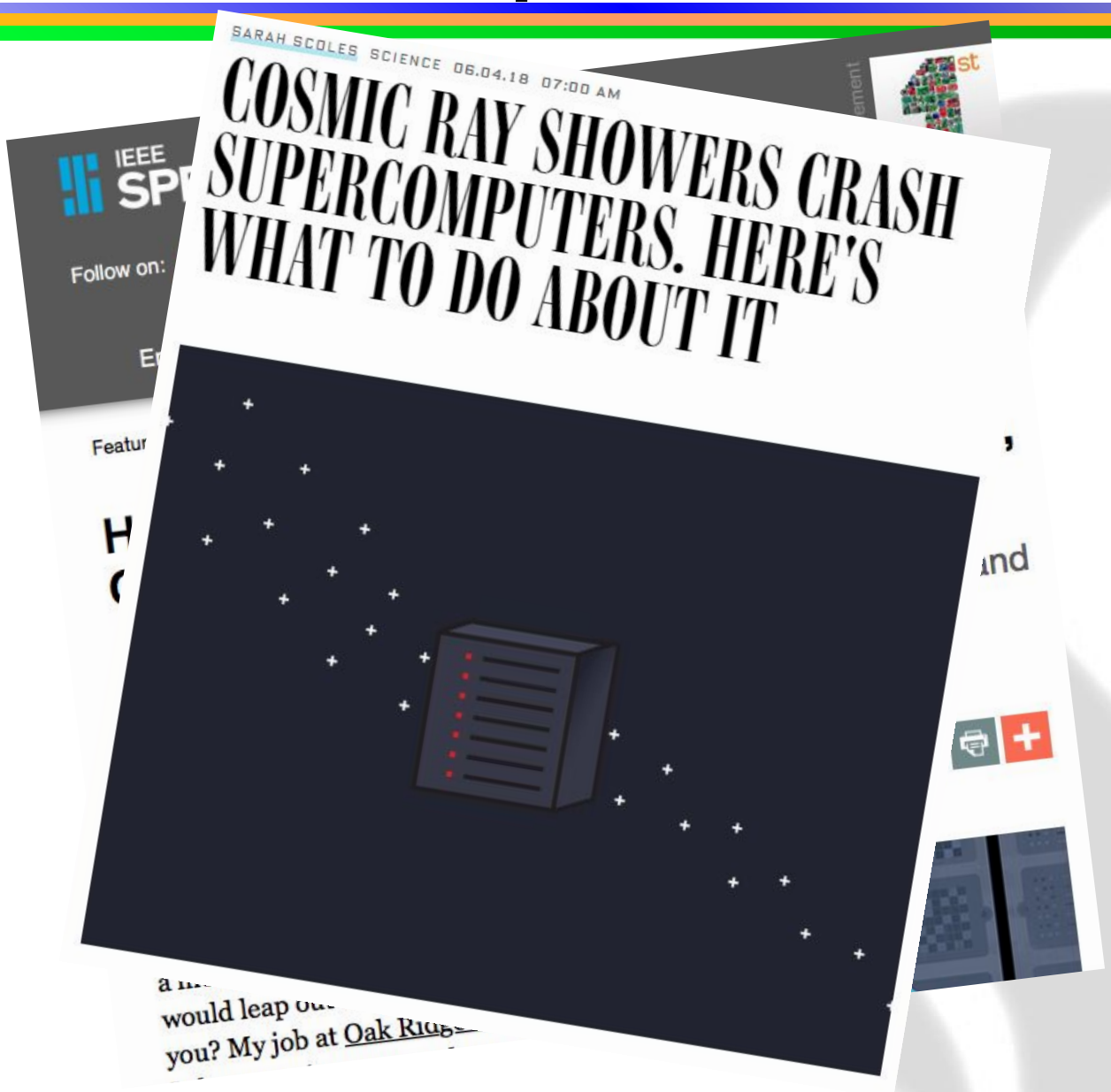
How To Kill A Supercomputer: Dirty Power, Cosmic Rays, and Bad Solder

Will future exascale supercomputers be able to withstand the steady onslaught of routine faults?

By Al Geist
Posted 23 Feb 2016 | 16:00 GMT

As a child, were you ever afraid that a monster lurking in your bedroom would leap out of the dark and get you? My job at Oak Ridge National

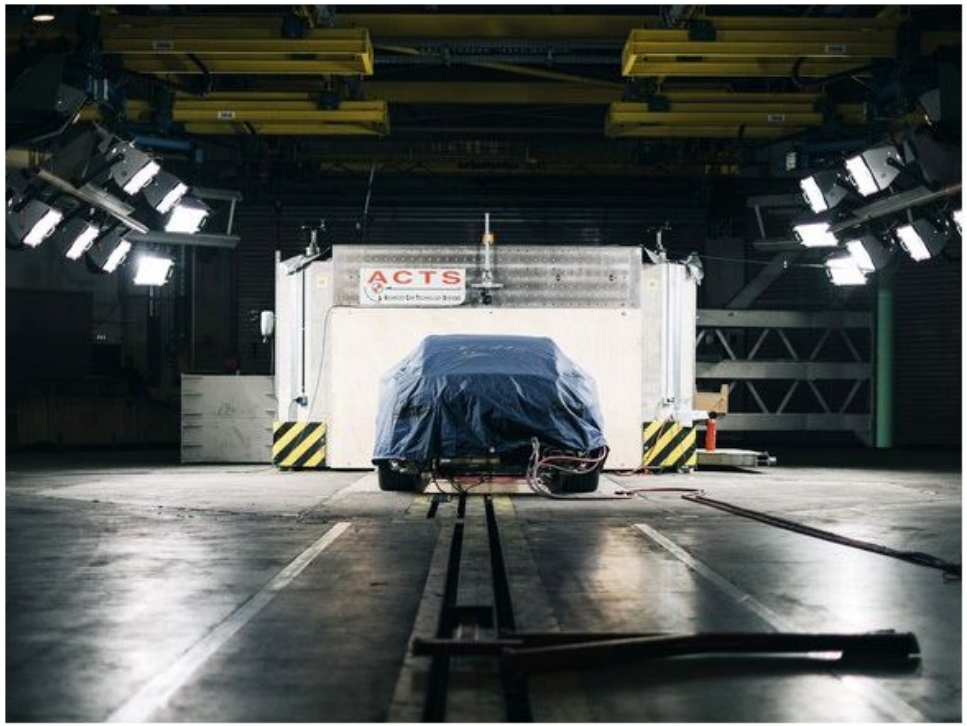
The Importance of Reliability



The Importance of Reliability

AARIAN MARSHALL TRANSPORTATION 03.31.18 07:00 AM

THE UBER CRASH WON'T BE THE LAST SHOCKING SELF-DRIVING DEATH



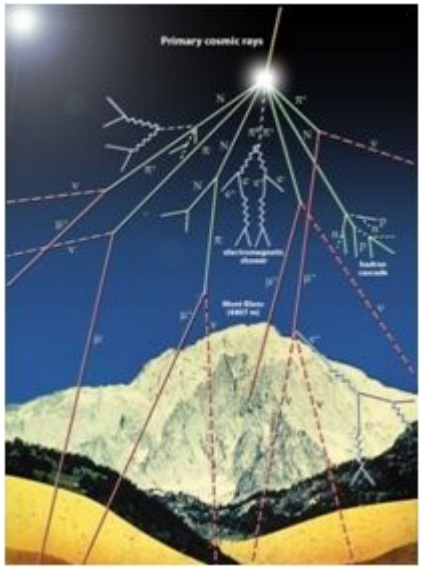
The Importance of Reliability

AARIAN MARSHALL TRANSPORTATION 03.31.18 07:00 AM

THE UBER CRASH WON'T BE THE LAST SHOCKING SELF-DRIVING DEATH

Reliability is a critical issue!

Terrestrial Radiation Environment



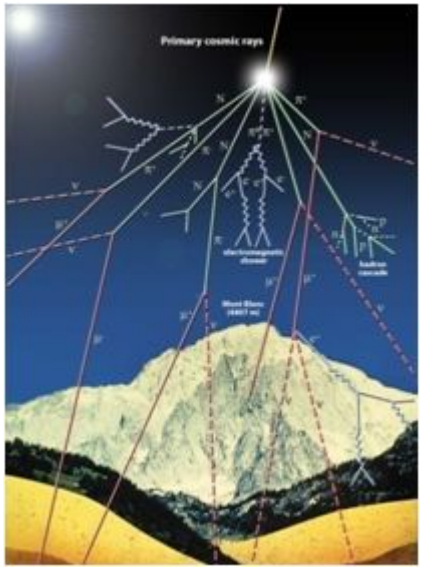
Galactic Cosmic rays interact with atmosphere and produce a shower of energetic particles:

- Muons
- Pions
- Protons
- Gamma Rays
- Neutrons

13 n/(cm^2 x h) @sea level*

*JEDEC JESD89A Standard

Terrestrial Radiation Environment



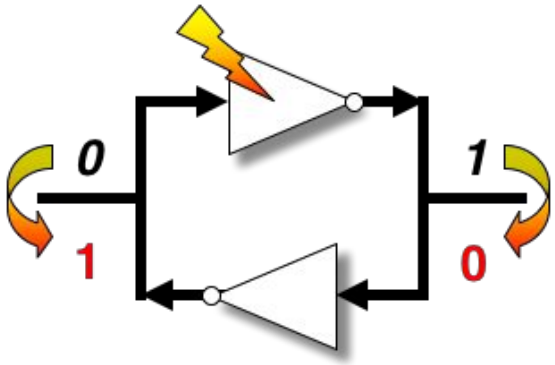
Galactic Cosmic rays interact with atmosphere and produce a shower of energetic particles:

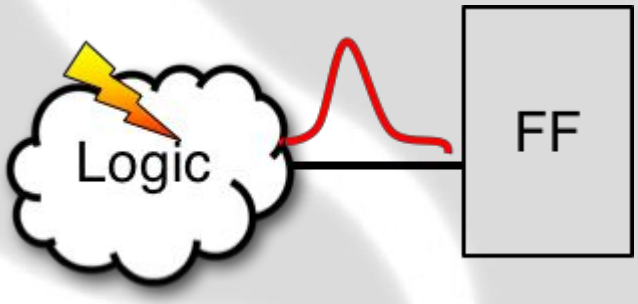
- Muons
- Pions
- Protons
- Gamma Rays
- Neutrons

13 n/(cm^2 x h) @sea level*

*JEDEC JESD89A Standard

Soft Errors: the device is not permanently damaged, but the particle may generate **bit flip(s)** in memory or **logic error(s)**





Silent Data Corruption vs Crash

Soft Errors in:
-data cache
-register files
-logic gates (ALU)
-scheduler

} Silent Data Corruption

Silent Data Corruption vs Crash

Soft Errors in:
-data cache
-register files
-logic gates (ALU)
-scheduler

Silent Data Corruption

Soft Errors in:
-instruction cache
-scheduler / dispatcher
-PCI-e bus controller

DUE (Crash)

Outline

- Radiation Effects Introduction
- **Experimental Methodology**
- Reliability of Mixed-Precision Architectures
  - FPGAs
  - x86
  - GPUs
- Conclusions

Experimental Methodologies (Devices)

FPGA: resources utilization is tailored by the user

higher precision => higher area

Experimental Methodologies (Devices)

FPGA: resources utilization is tailored by the user

higher precision => higher area

X86 (Xeon Phi): the Vector Processing Unit (VPU) executes operations in 64 or 32 bits, on the same HW.

Compiler decides how to use the VPUs

Experimental Methodologies (Devices)

FPGA: resources utilization is tailored by the user

higher precision => higher area

X86 (Xeon Phi): the Vector Processing Unit (VPU) executes operations in 64 or 32 bits, on the same HW.

Compiler decides how to use the VPUs

GPU (Volta V100): dedicated HW for double and single/half precision

1 double operation, 1 single or 2 half operations

Experimental Methodologies

Fault Injection

Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

[PVF/AVF]
Program Vulnerability Factor
Architectural Vulnerability Factor

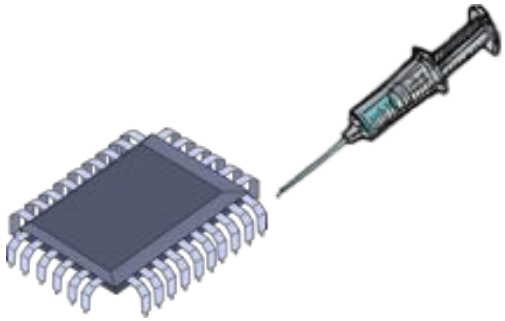
Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

[PVF/AVF]
Program Vulnerability Factor
Architectural Vulnerability Factor



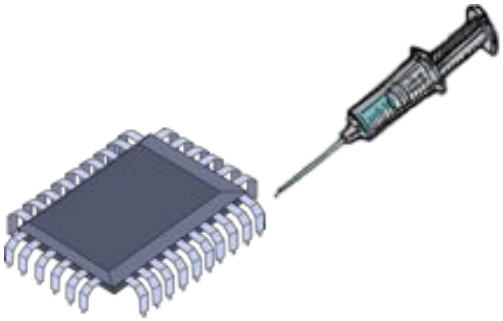
Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

[PVF/AVF]
Program Vulnerability Factor
Architectural Vulnerability Factor



Neutron Beam

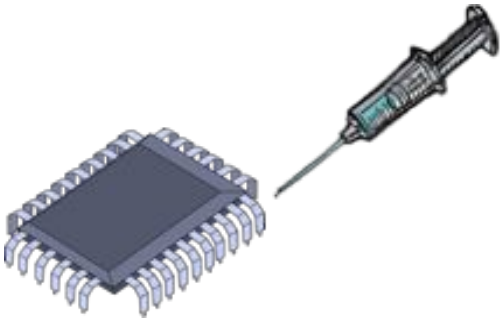
Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

[PVF/AVF]
Program Vulnerability Factor
Architectural Vulnerability Factor



Neutron Beam

We irradiate the devices while running benchmarks.

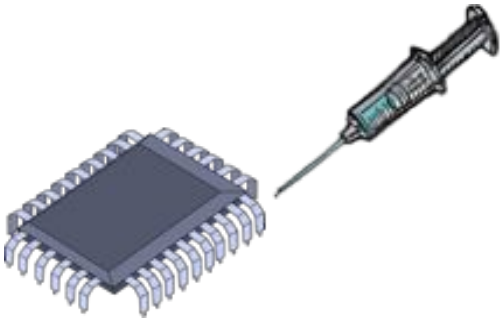
Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

[PVF/AVF]
Program Vulnerability Factor
Architectural Vulnerability Factor



Neutron Beam

We irradiate the devices while running benchmarks.

Realistic estimation of the error rate of a given device executing a given application.

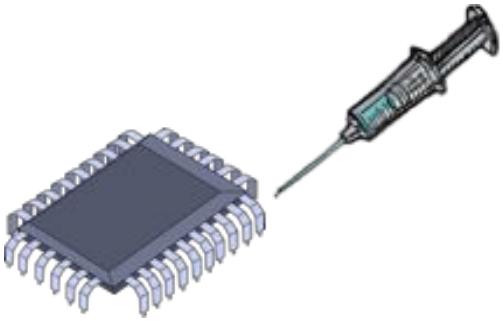
Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

[PVF/AVF]
Program Vulnerability Factor
Architectural Vulnerability Factor



Neutron Beam

We irradiate the devices while running benchmarks.

Realistic estimation of the error rate of a given device executing a given application.

[FIT]
Failure in Time

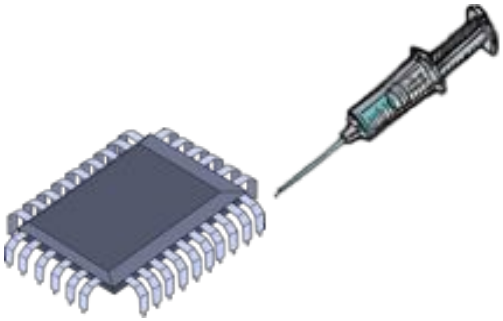
Experimental Methodologies

Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

[PVF/AVF]
Program Vulnerability Factor
Architectural Vulnerability Factor



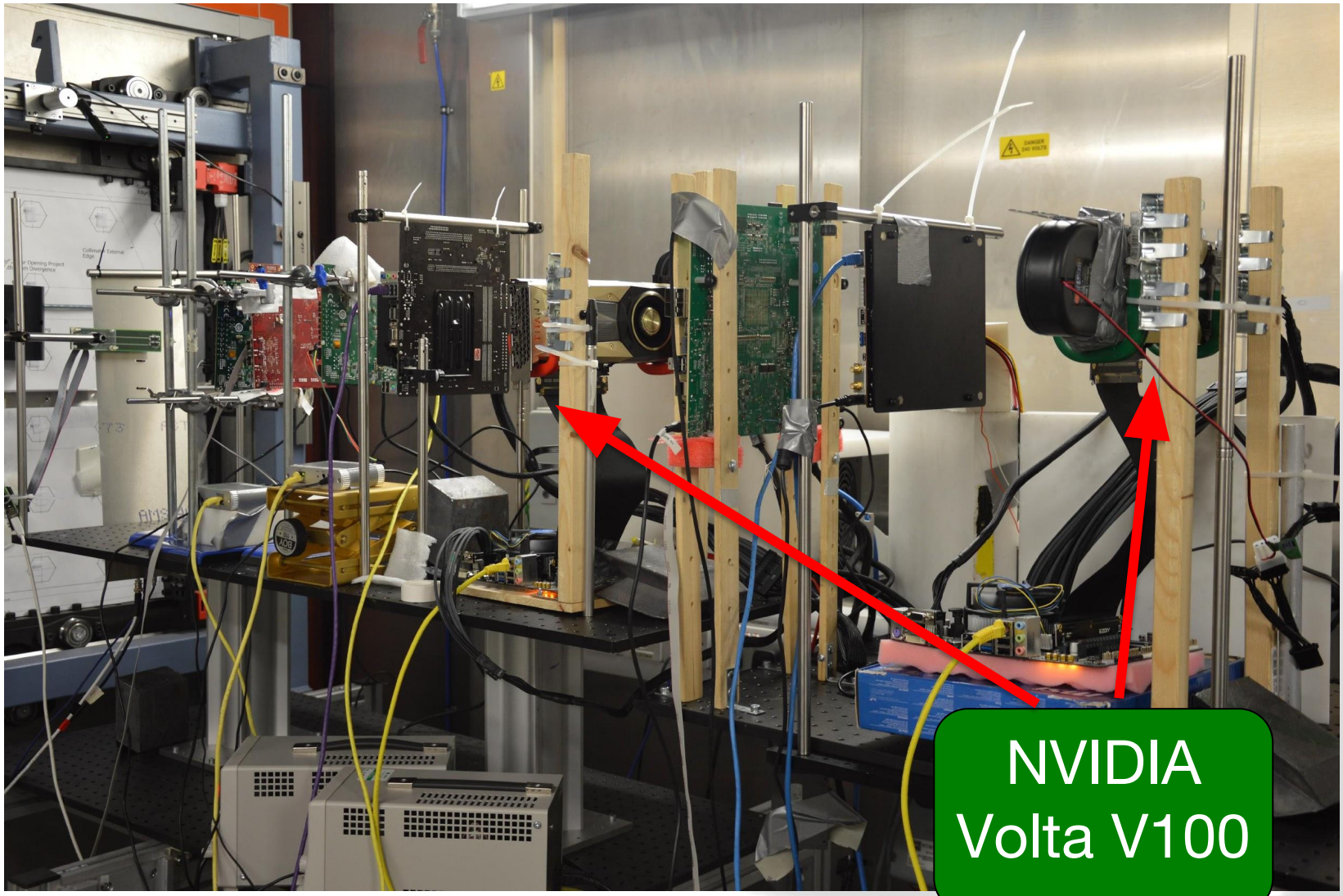
Neutron Beam

We irradiate the devices while running benchmarks.

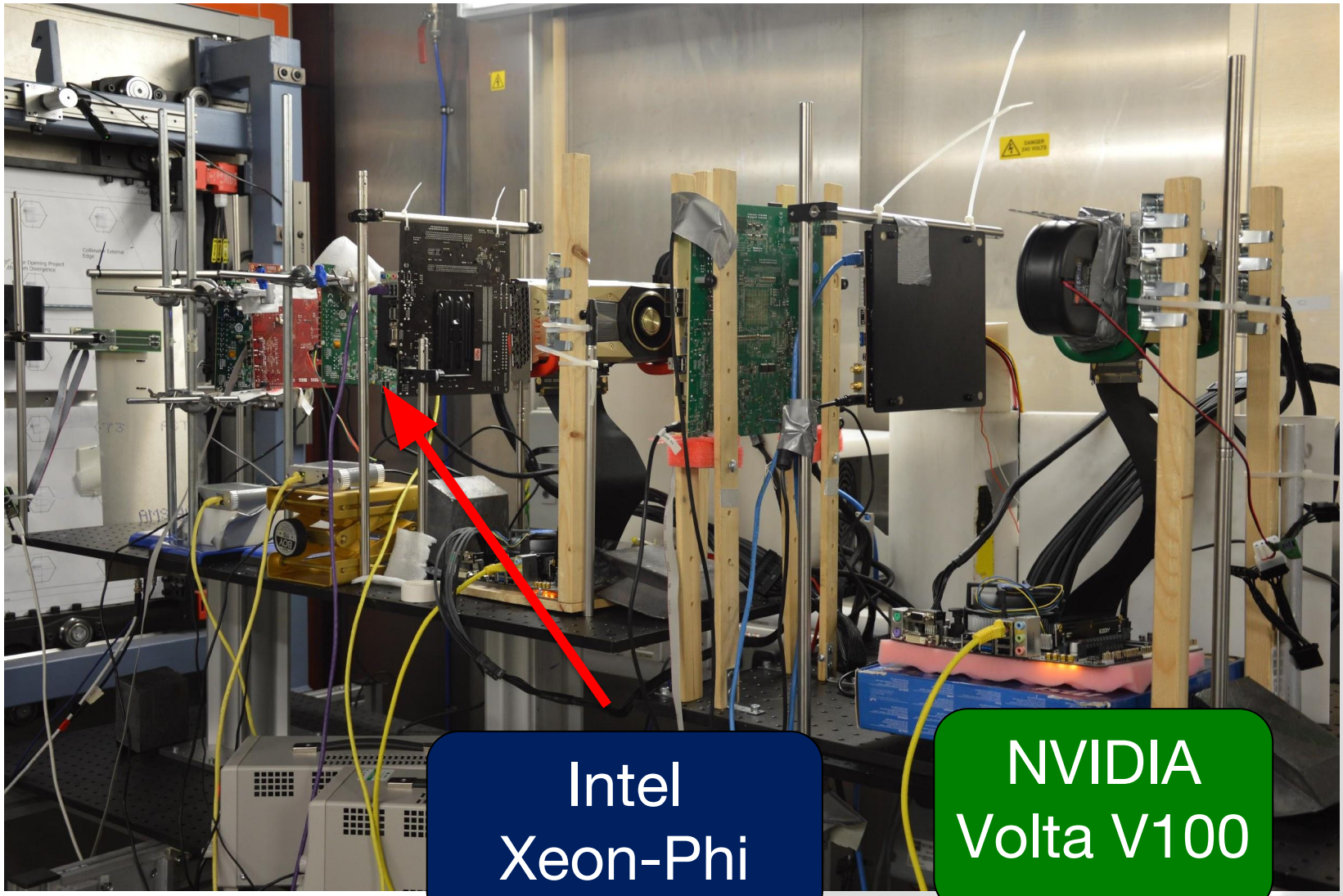
Realistic estimation of the error rate of a given device executing a given application.

[FIT]
Failure in Time

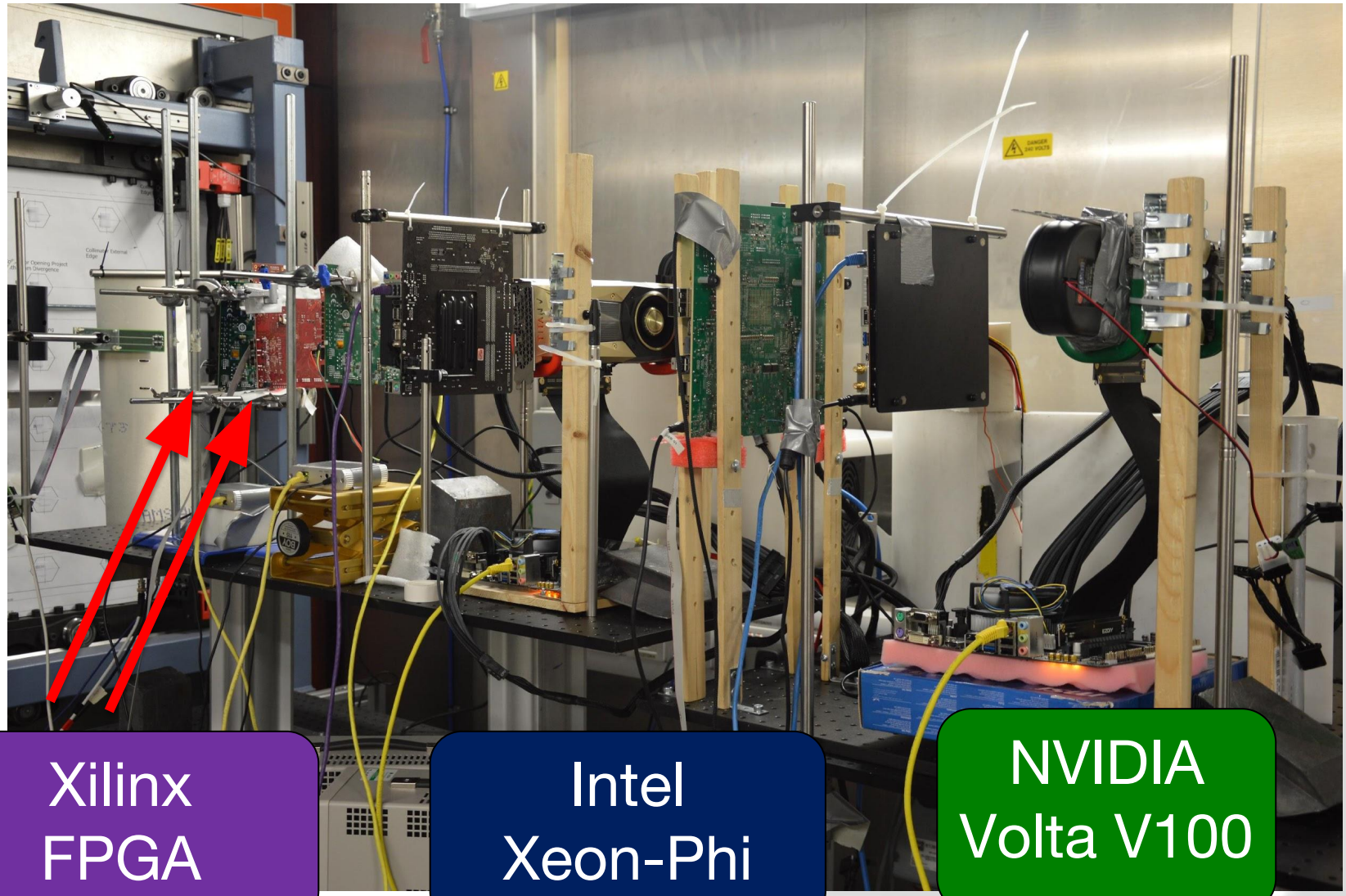
Radiation Experiments @ChipIR



Radiation Experiments @ChipIR



Radiation Experiments @ChipIR



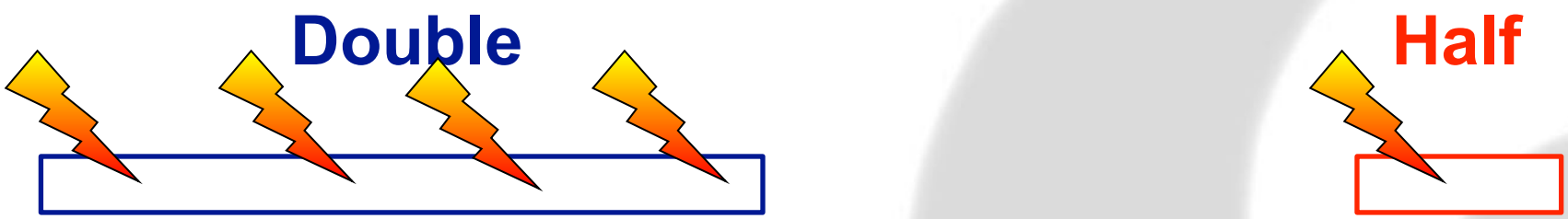
Outline

- Radiation Effects Introduction
- Experimental Methodology
- **Reliability of Mixed-Precision Architectures**
  - FPGAs
  - x86
  - GPUs
- Conclusions

Mixed-Precision Reliability

Double has a much larger area than **Half**

Memory: 4x **Functional Units: ~16x**

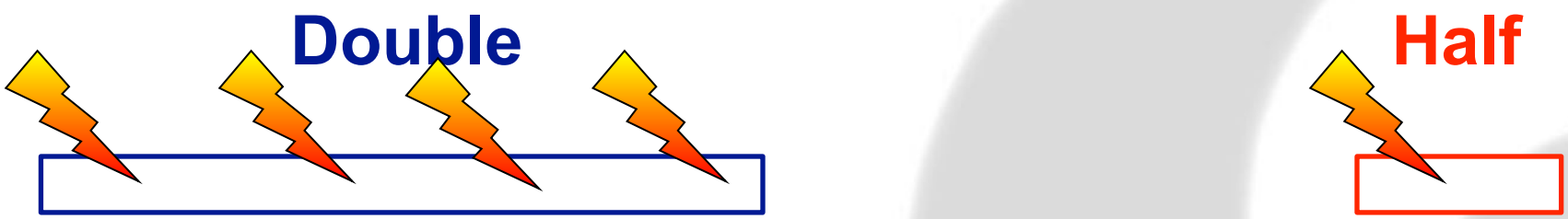


It is much more likely for a double value/operation to be corrupted than a half value/operation

Mixed-Precision Reliability

Double has a much larger area than **Half**

Memory: 4x **Functional Units: ~16x**



It is much more likely for a double value/operation to be corrupted than a half value/operation

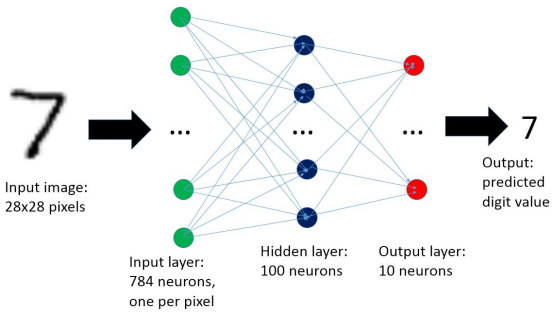
However, a fault in a double value is much less critical than a fault in half

Double 52/64 (81%) bits are mantissa
Half 10/16 (60%) bits are mantissa

Outline

- Radiation Effects Introduction
- Experimental Methodology
- **Reliability of Mixed-Precision Architectures**
  - **FPGAs**
  - x86
  - GPUs
- Conclusions

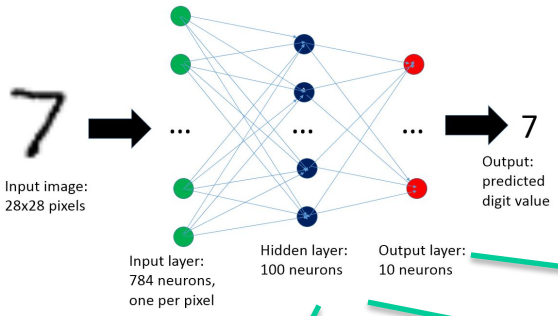
MNIST



Matrix Multiplication (128x128)

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 10 & 8 \end{bmatrix}$$

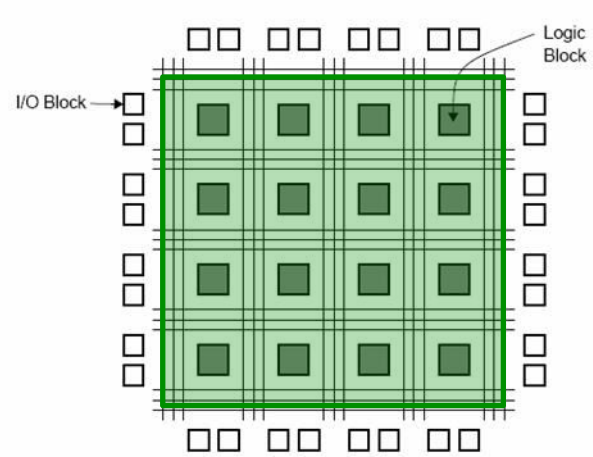
MNIST



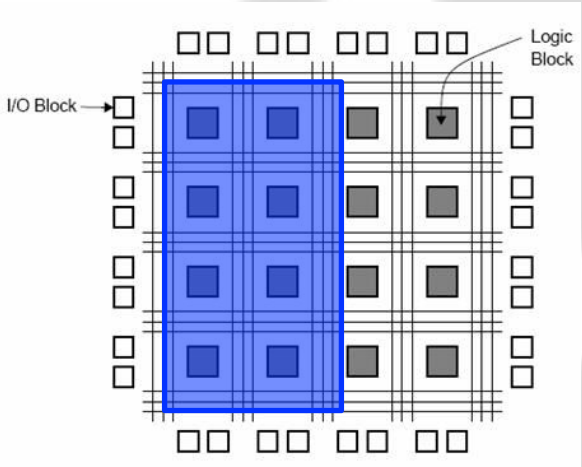
Matrix Multiplication (128x128)

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 10 & 8 \end{bmatrix}$$

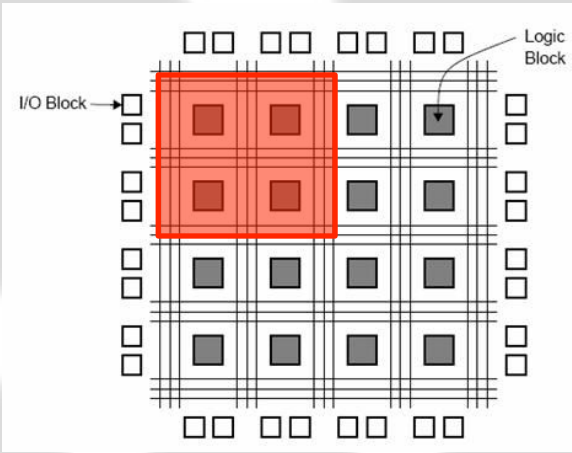
Double



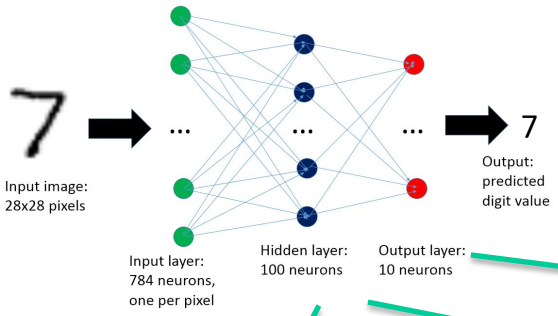
Single



Half



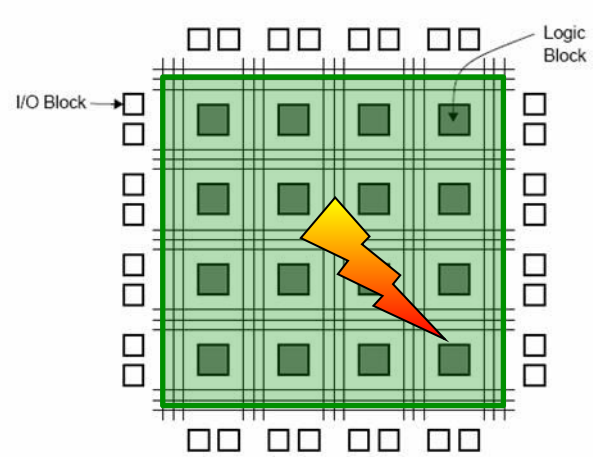
MNIST



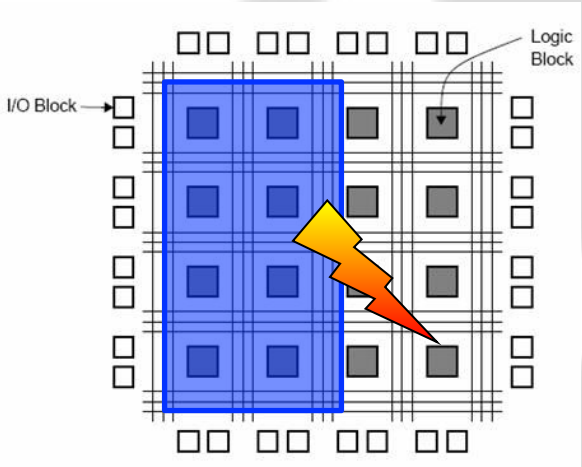
Matrix Multiplication (128x128)

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 10 & 8 \end{bmatrix}$$

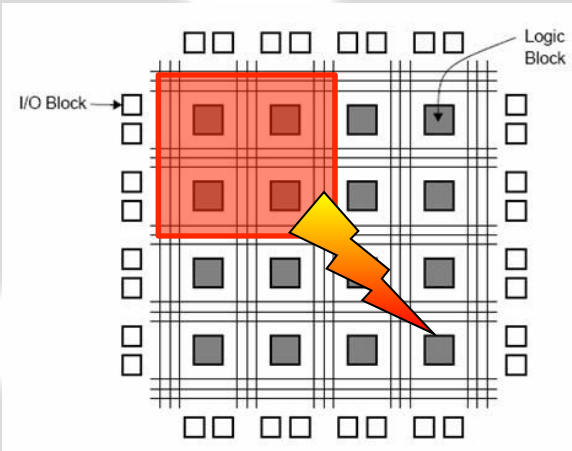
Double



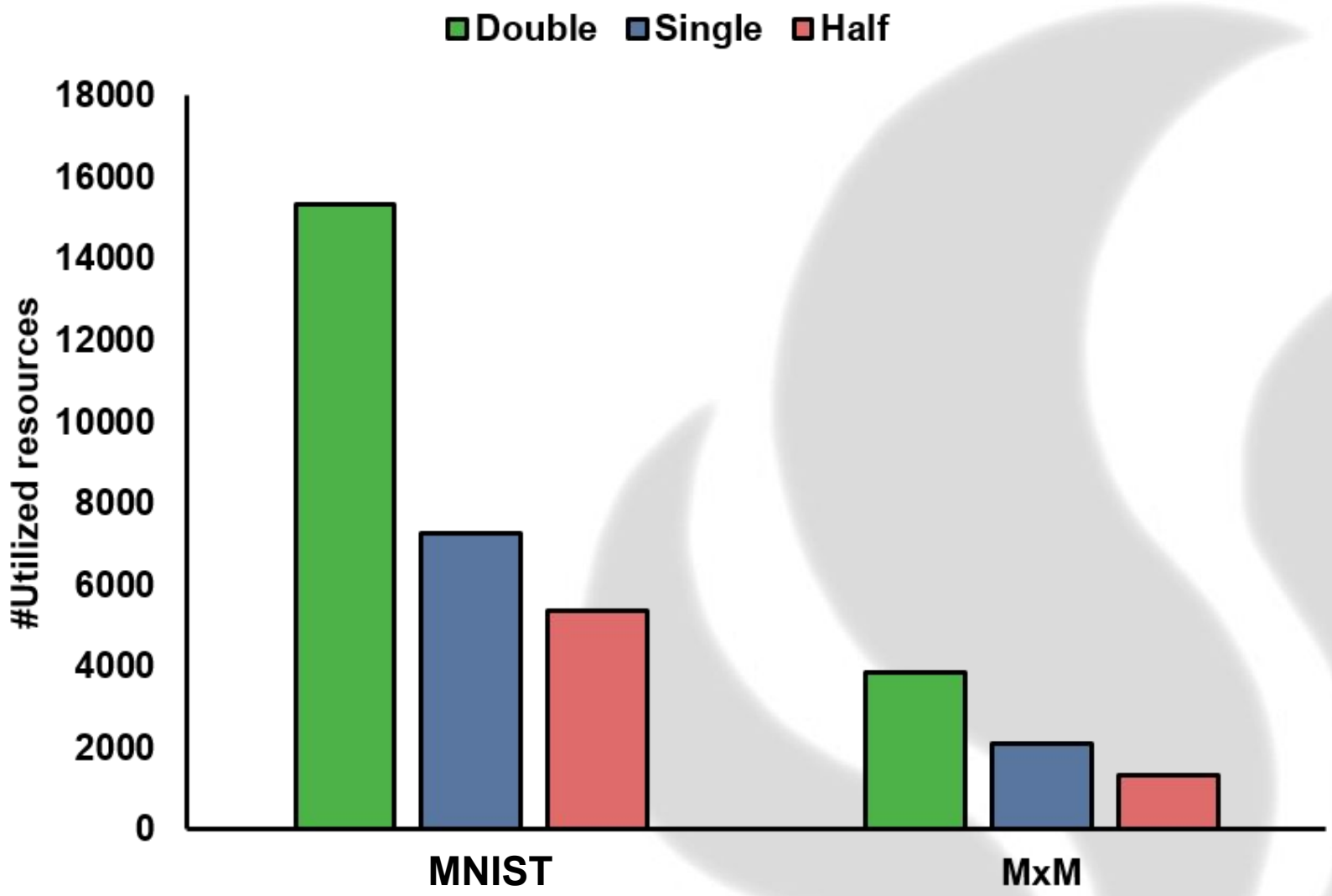
Single



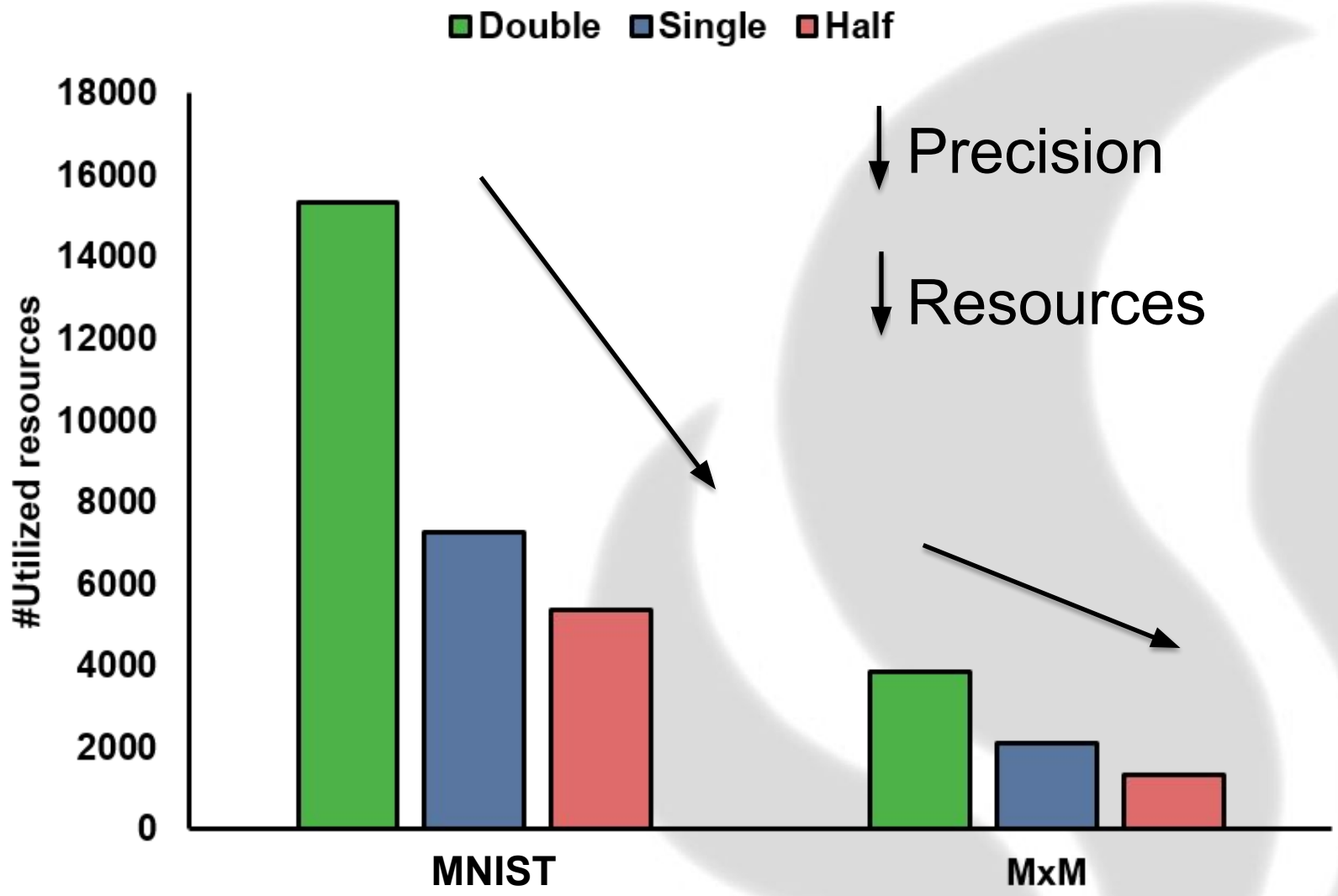
Half



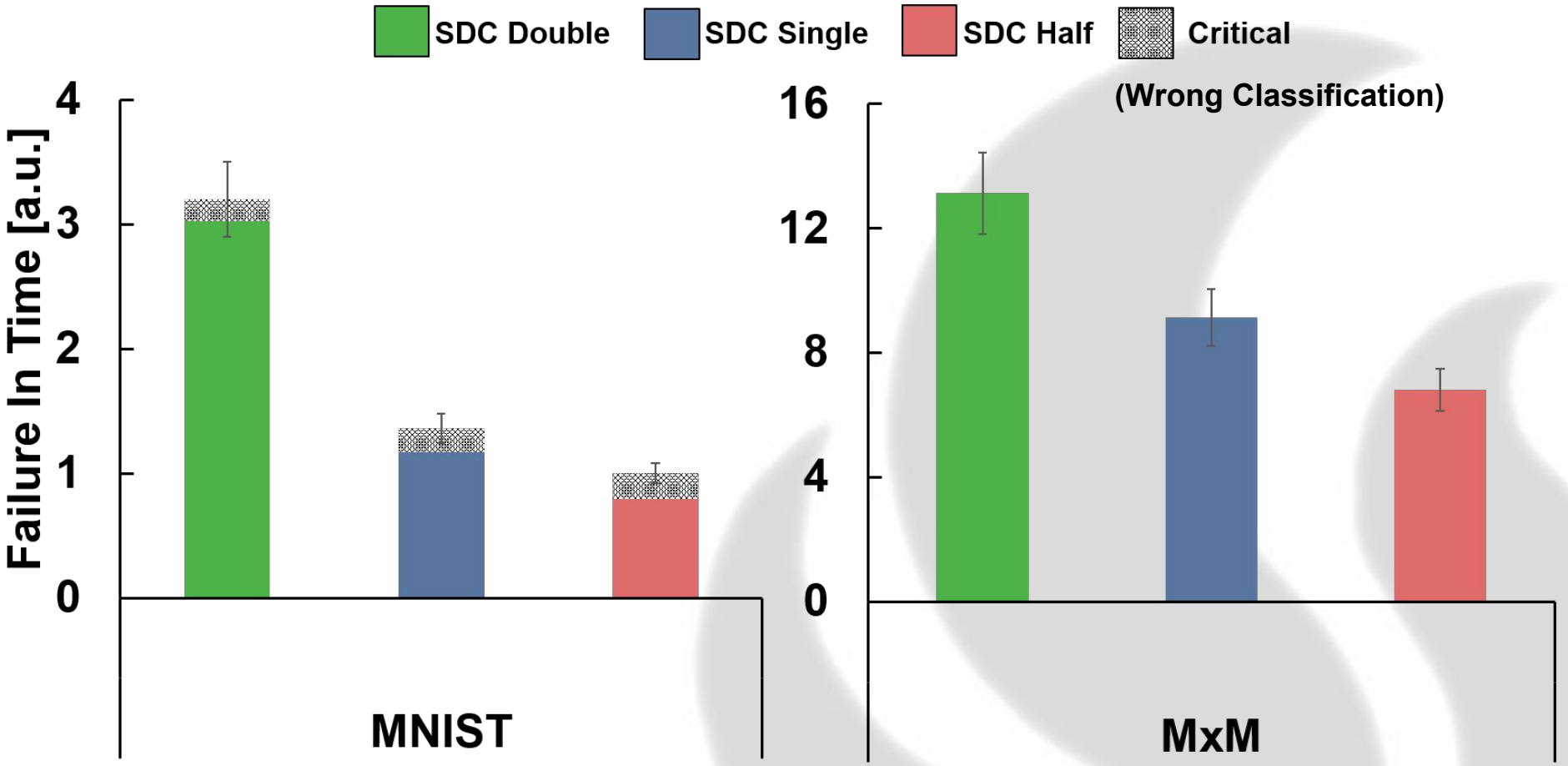
FPGA


Double
Single
Half
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
#Utilized resources
MNIST
MxM

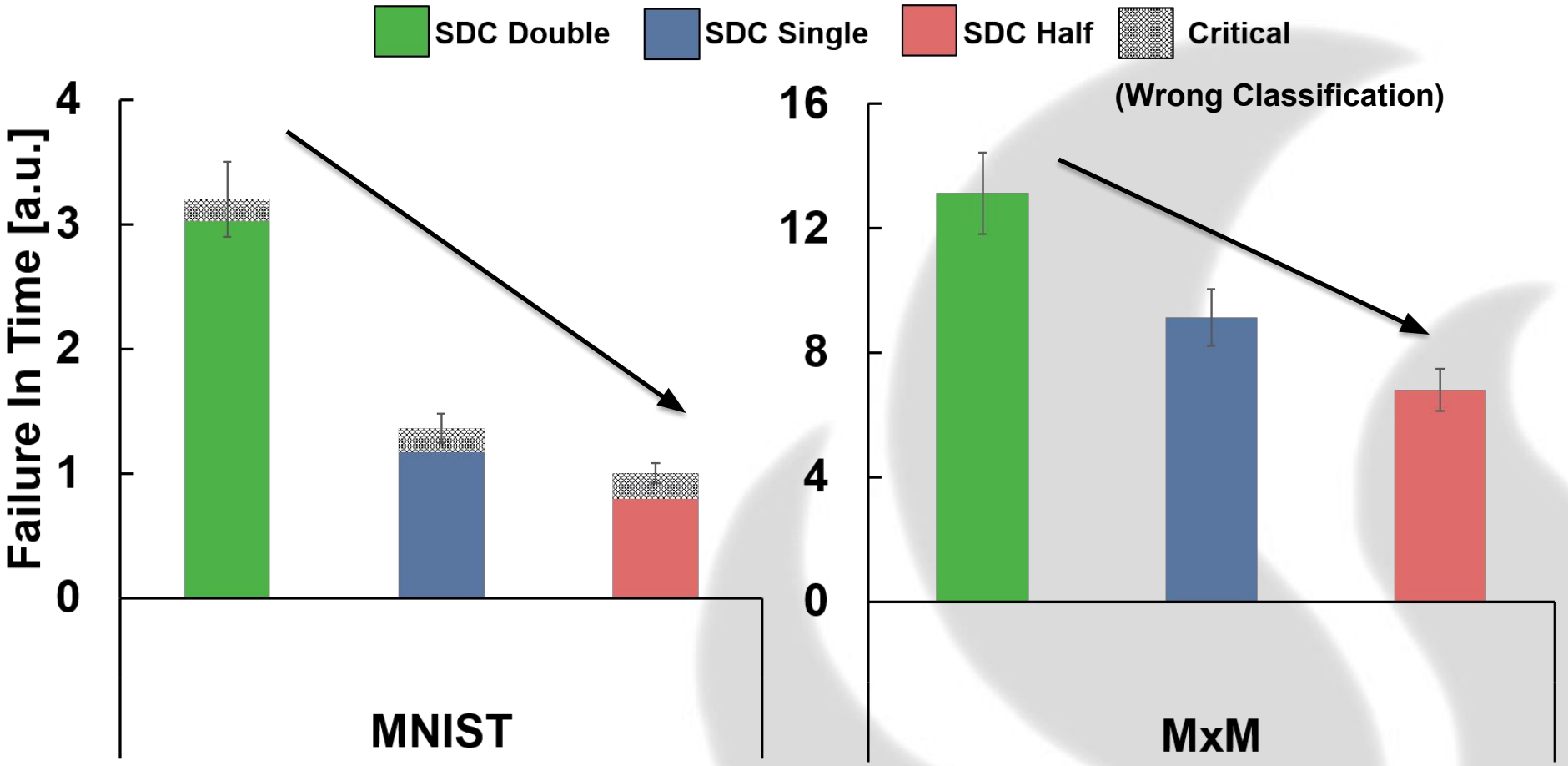
FPGA


Double
Single
Half
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
#Utilized resources
Precision
Resources
MNIST
MxM

FPGA - FIT rate

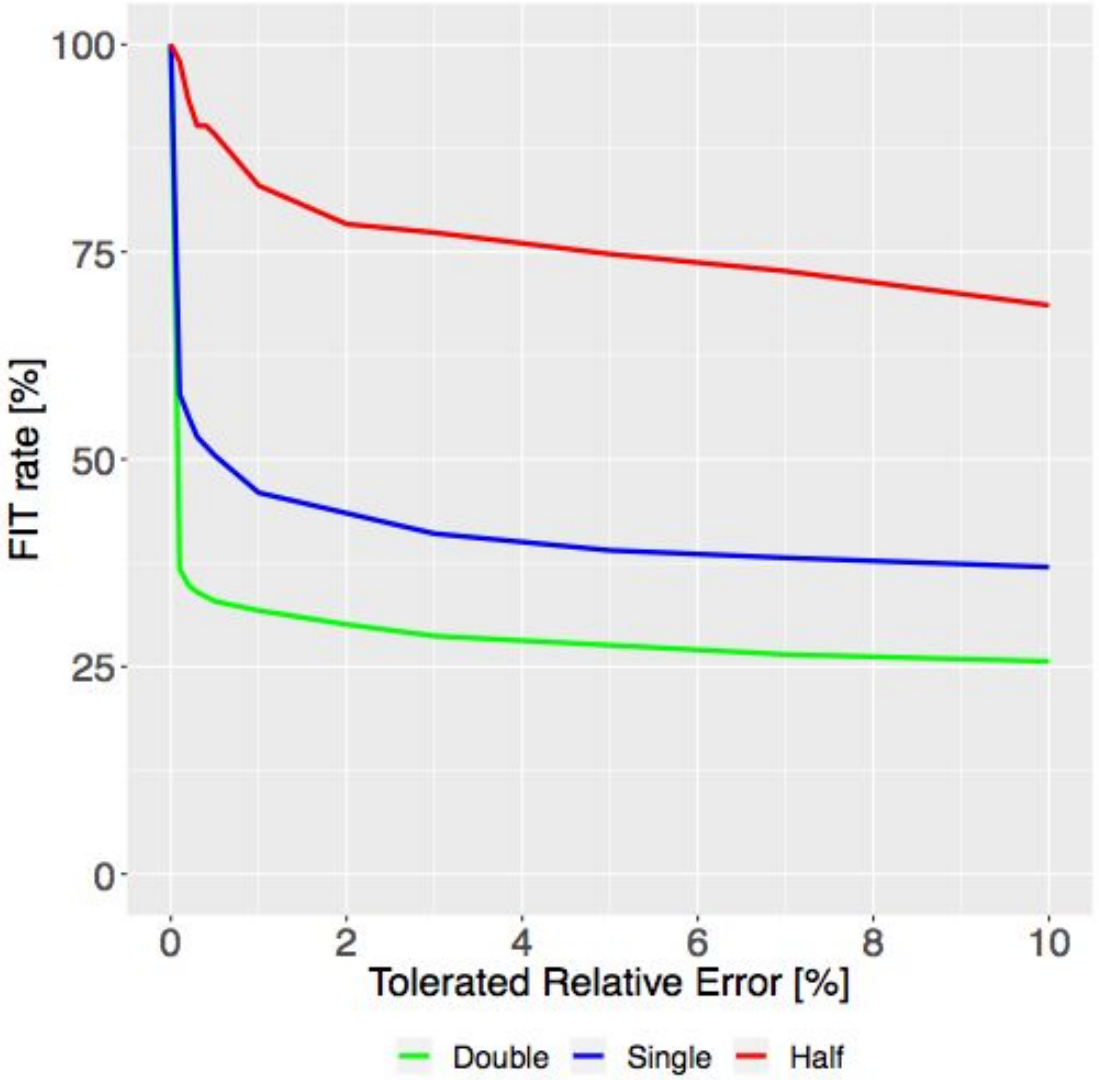


FPGA - FIT rate

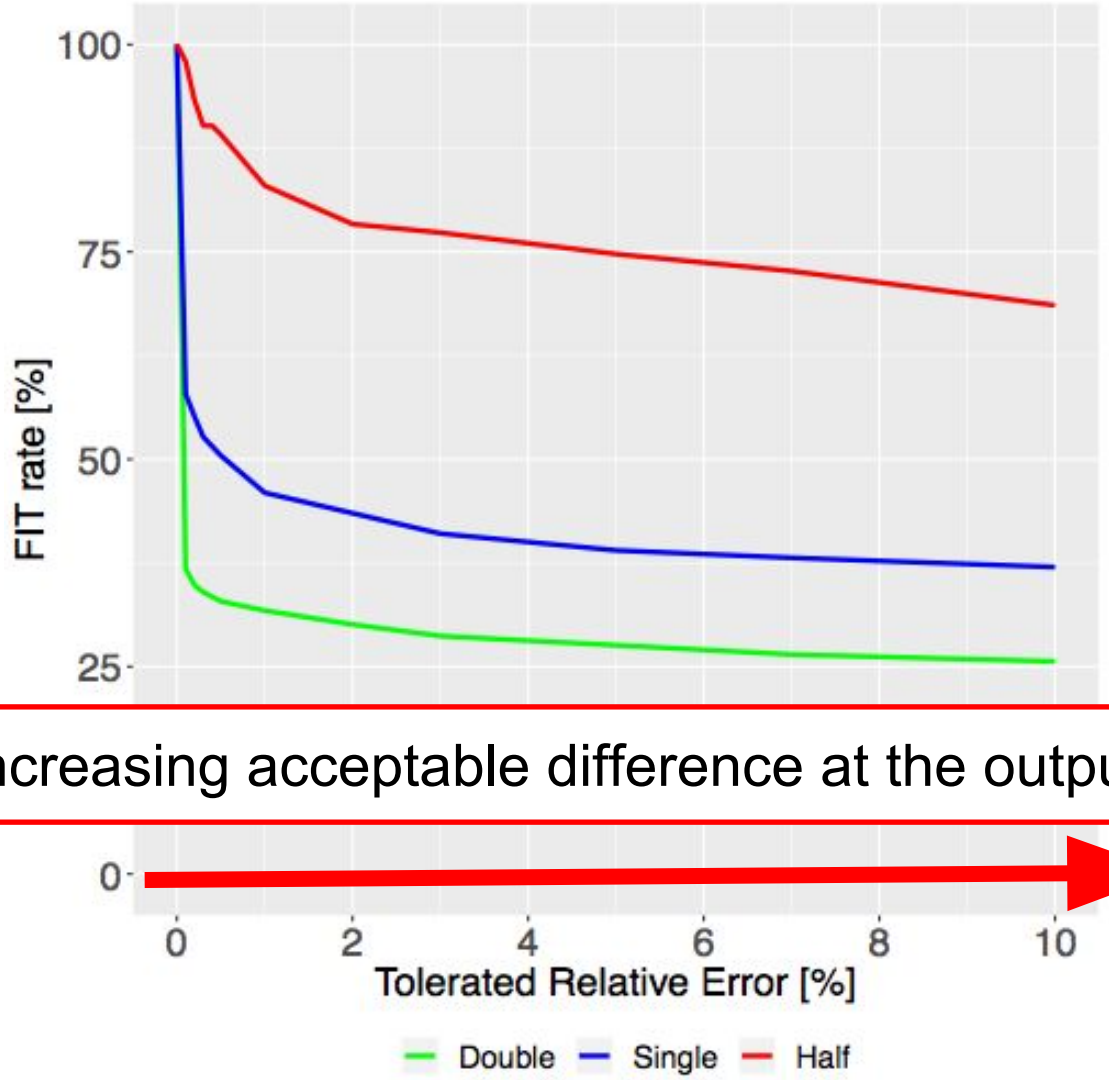


As we lower the precision, we also decrease circuit area, ultimately reducing the overall error rate [FIT].

FPGA - error criticality



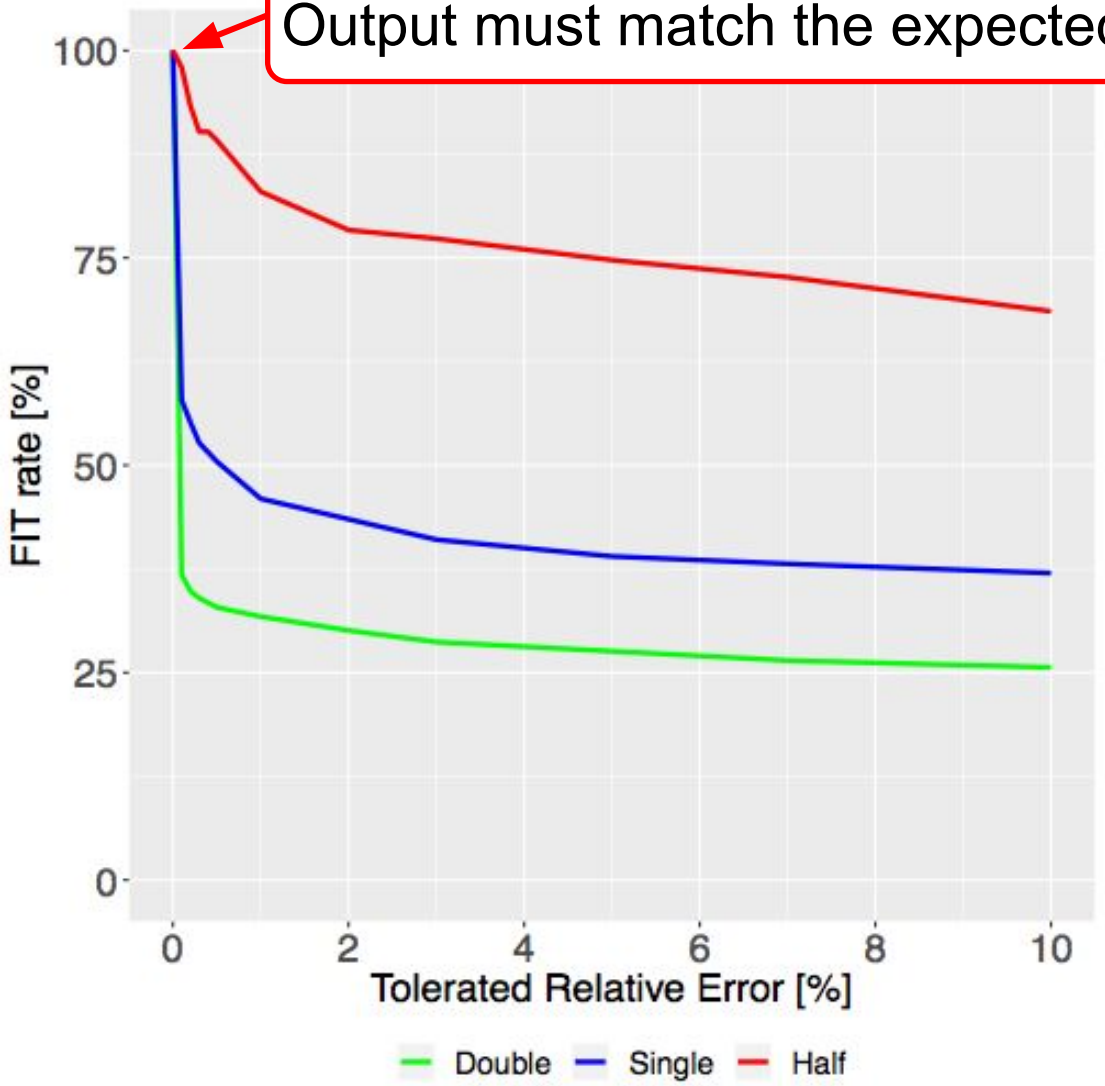
FPGA - error criticality



Increasing acceptable difference at the output

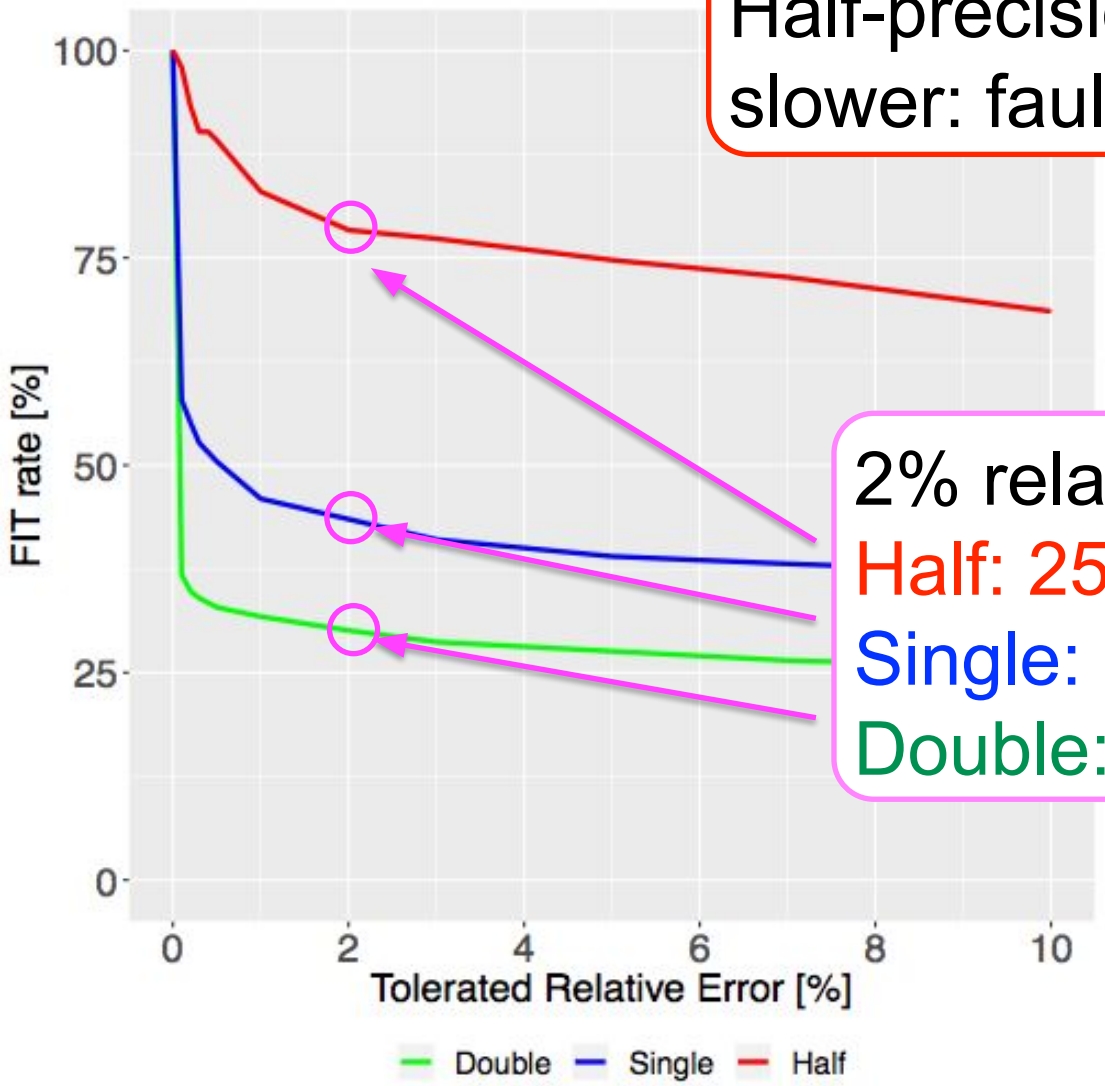
FPGA - error criticality

Output must match the expected output (0% tolerance)



FPGA - error criticality

Half-precision FIT reduction is slower: faults are more critical



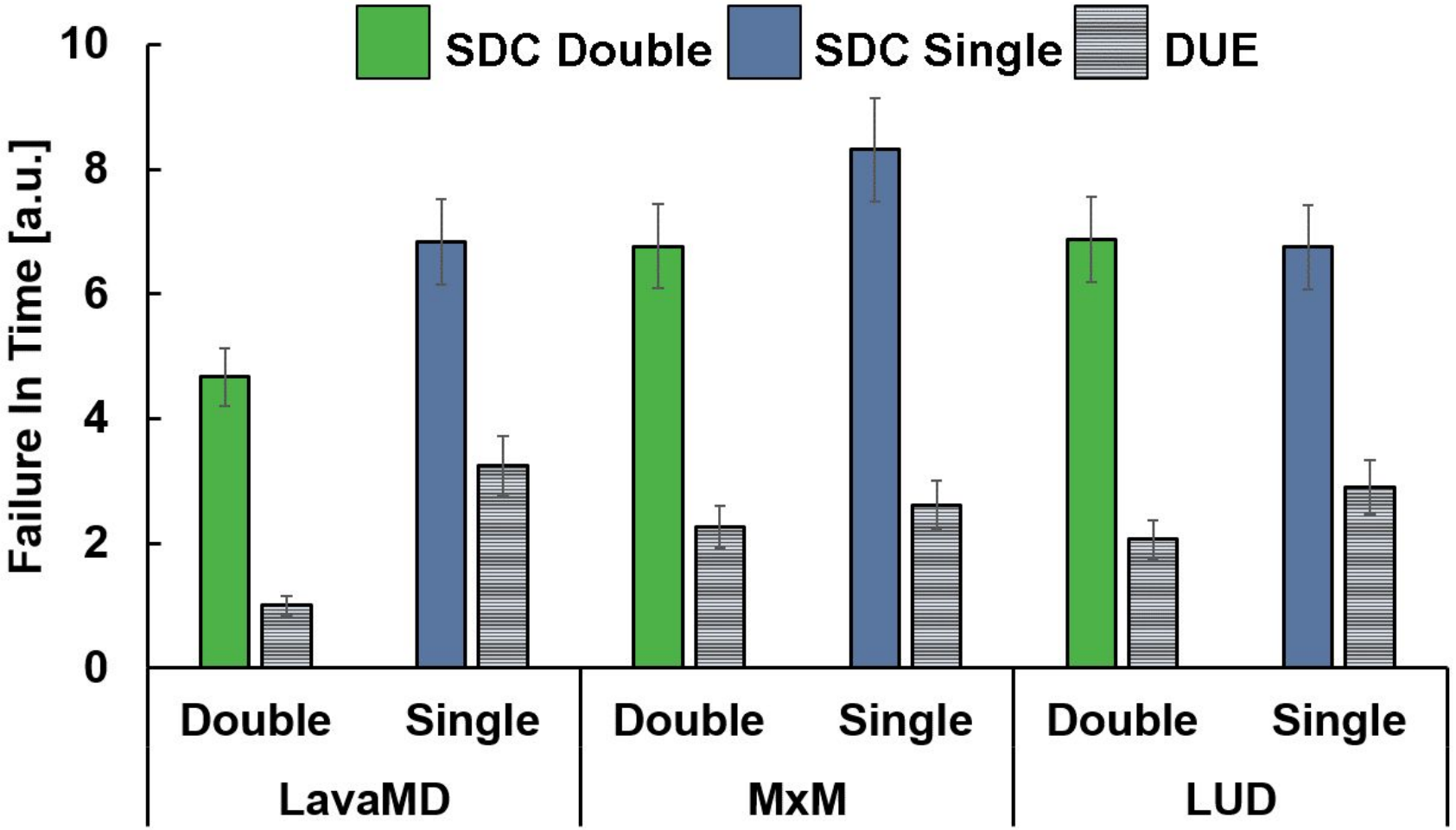
2% relative error tolerated:
Half: 25% FIT reduction
Single: 55% FIT reduction
Double: 72% FIT reduction

Outline

- Radiation Effects Introduction
- Experimental Methodology
- **Reliability of Mixed-Precision Architectures**
  - FPGAs
  - **x86**
  - GPUs
- Conclusions

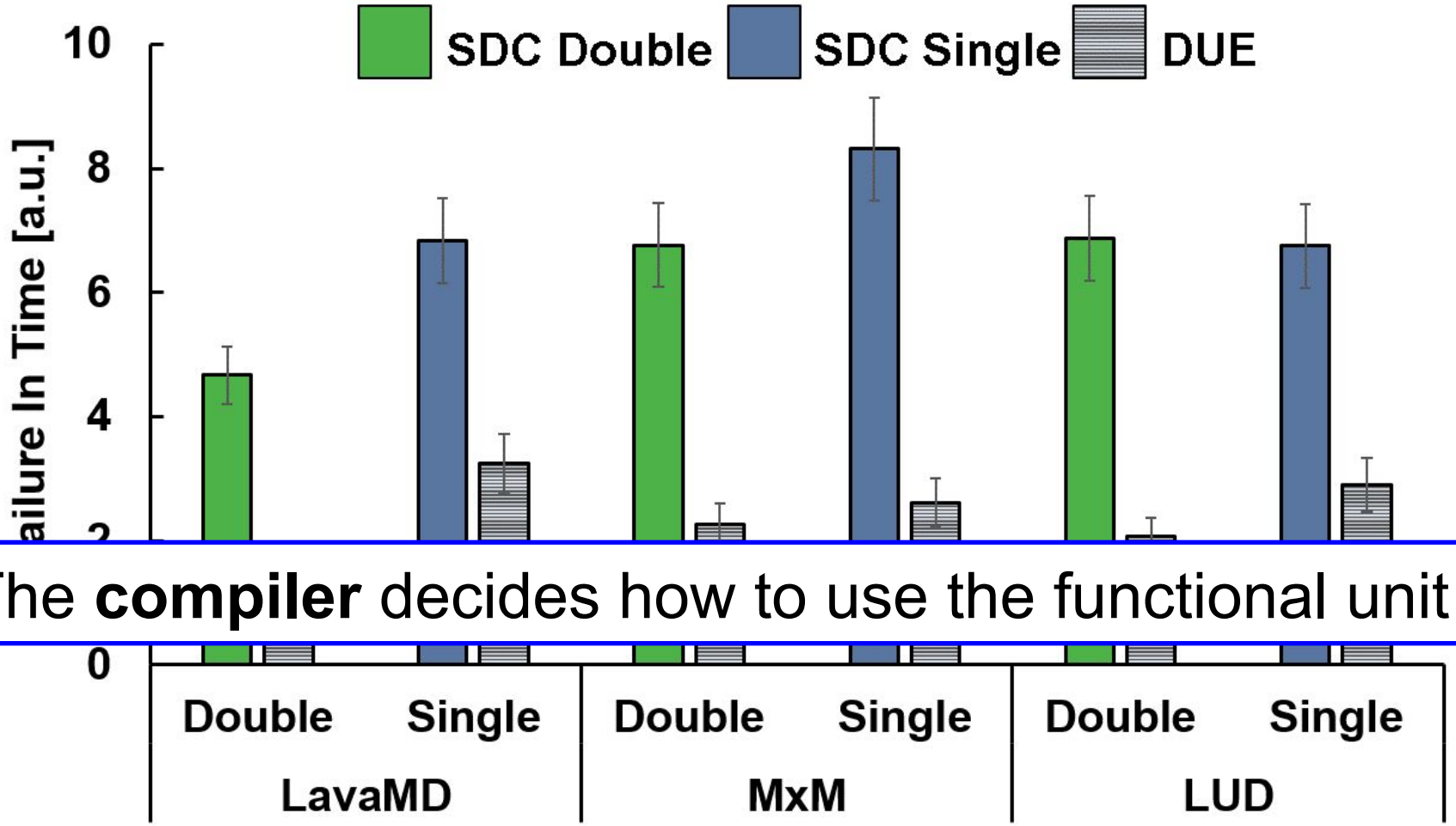
Xeon Phi - Beam

One-size fits all: Xeon Phi does not have dedicated HW for double/single



Xeon Phi - Beam

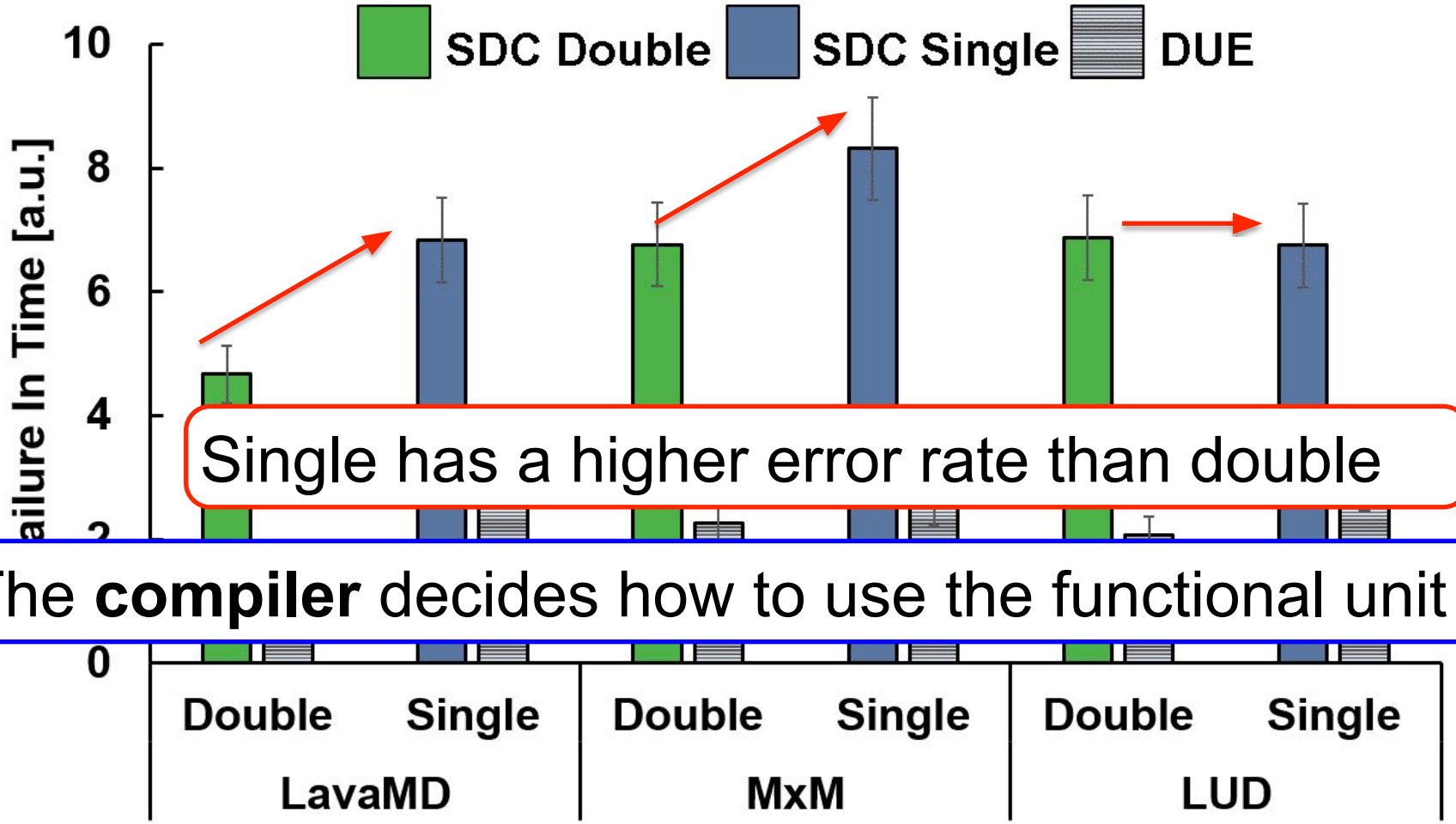
One-size fits all: Xeon Phi does not have dedicated HW for double/single



The **compiler** decides how to use the functional unit

Xeon Phi - Beam

One-size fits all: Xeon Phi does not have dedicated HW for double/single

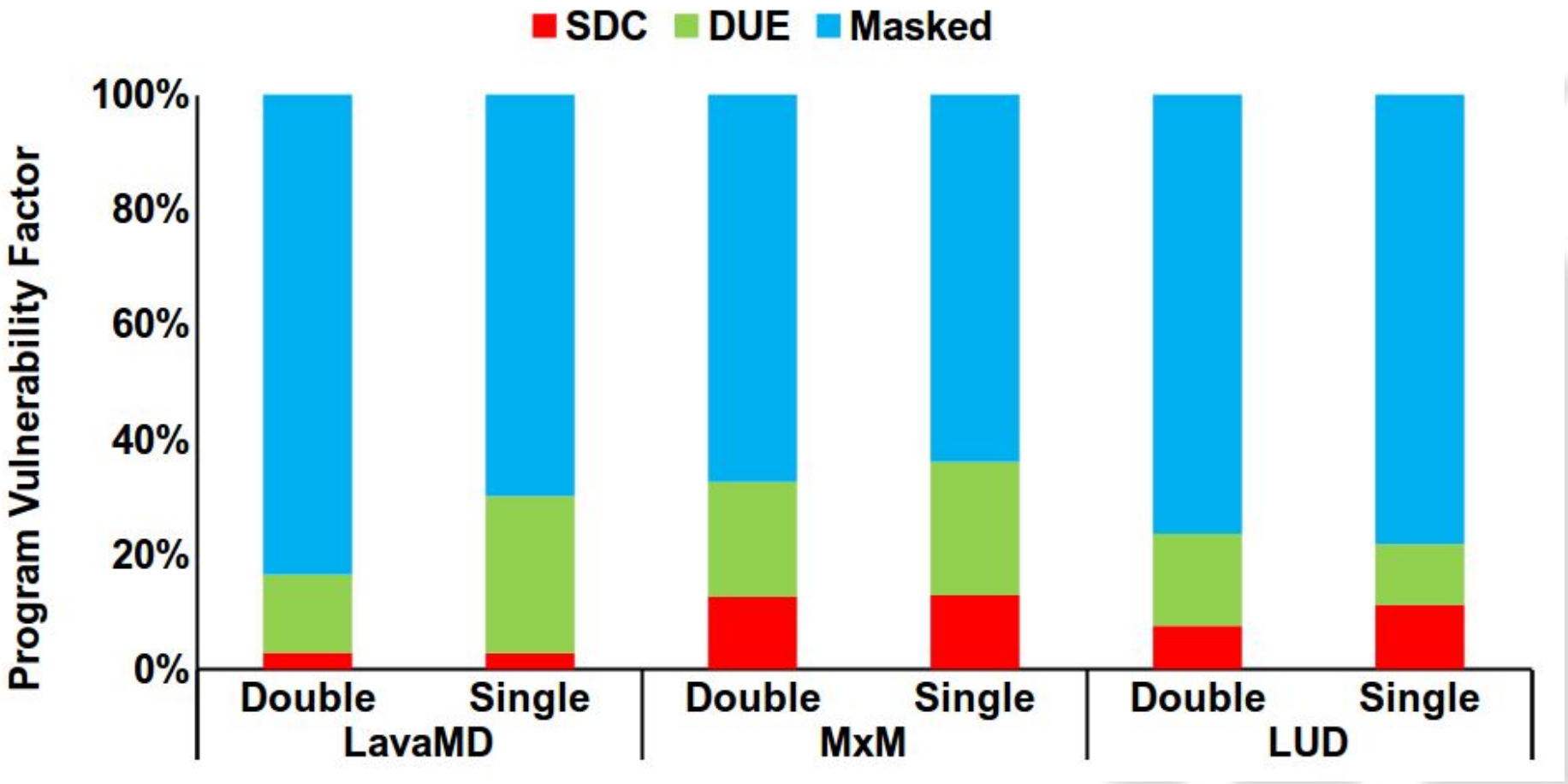


Single has a higher error rate than double

The **compiler** decides how to use the functional unit

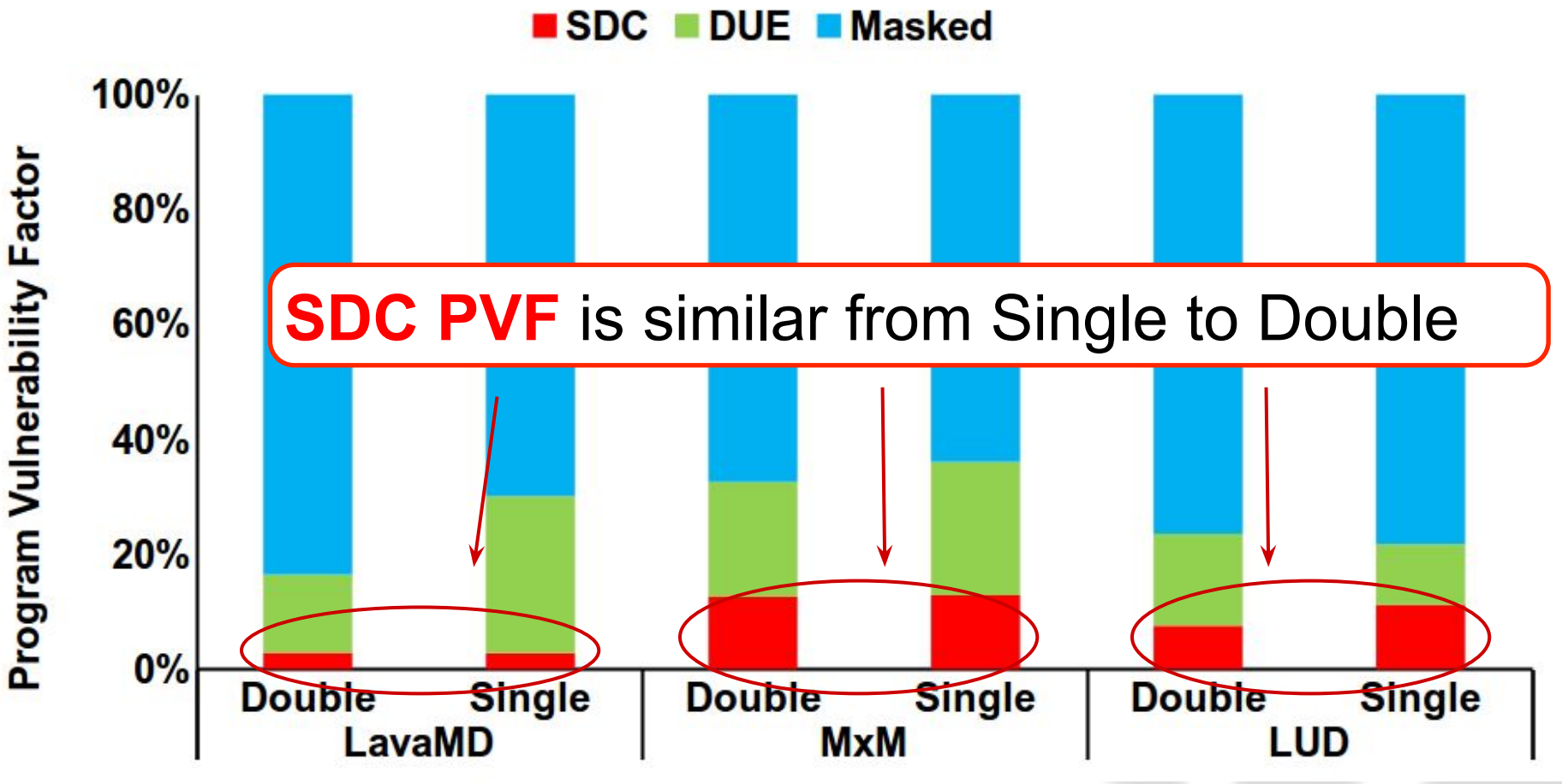
Xeon Phi - Fault Injection

We **inject faults** in variables during execution



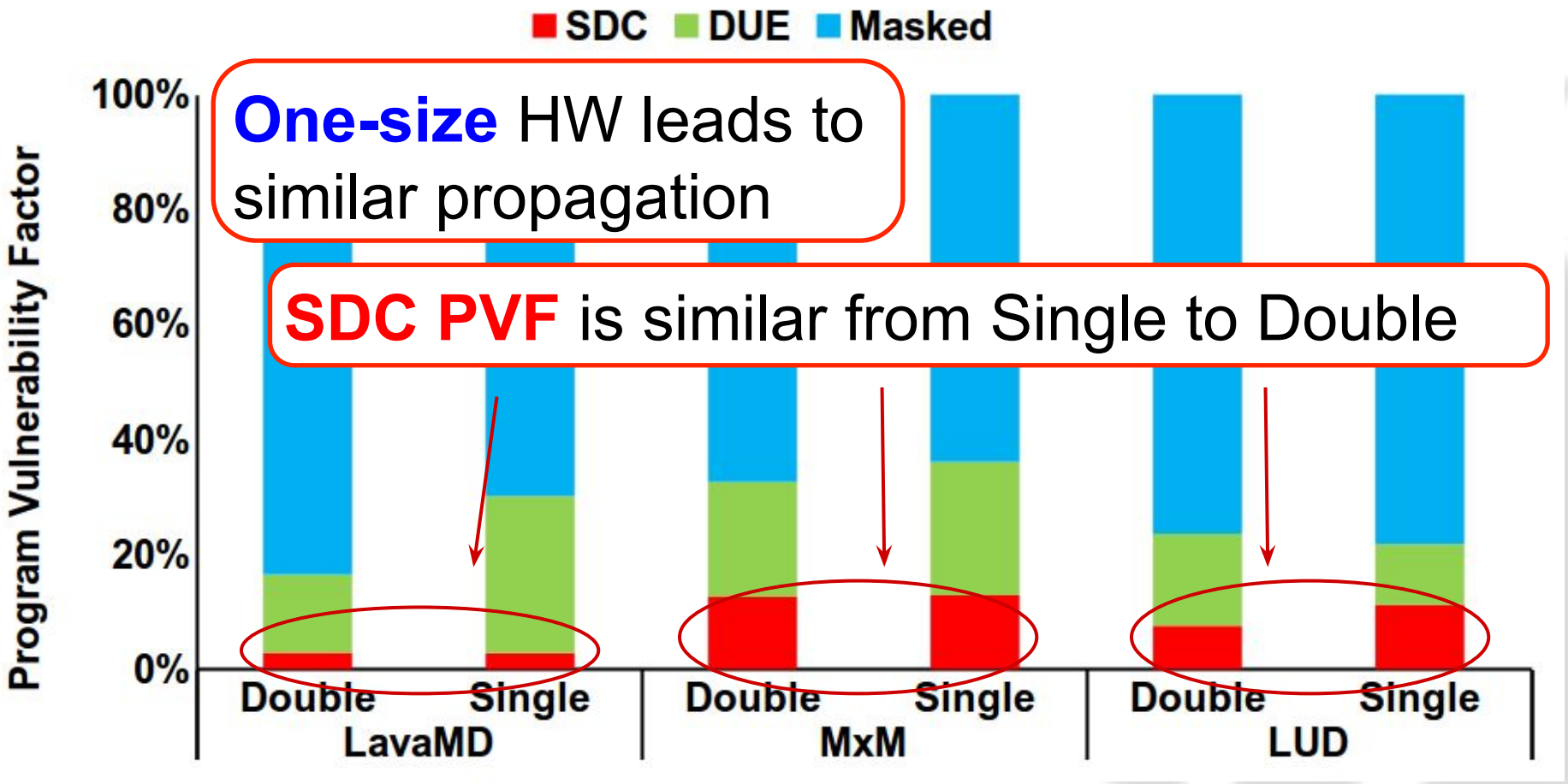
Xeon Phi - Fault Injection

We **inject faults** in variables during execution



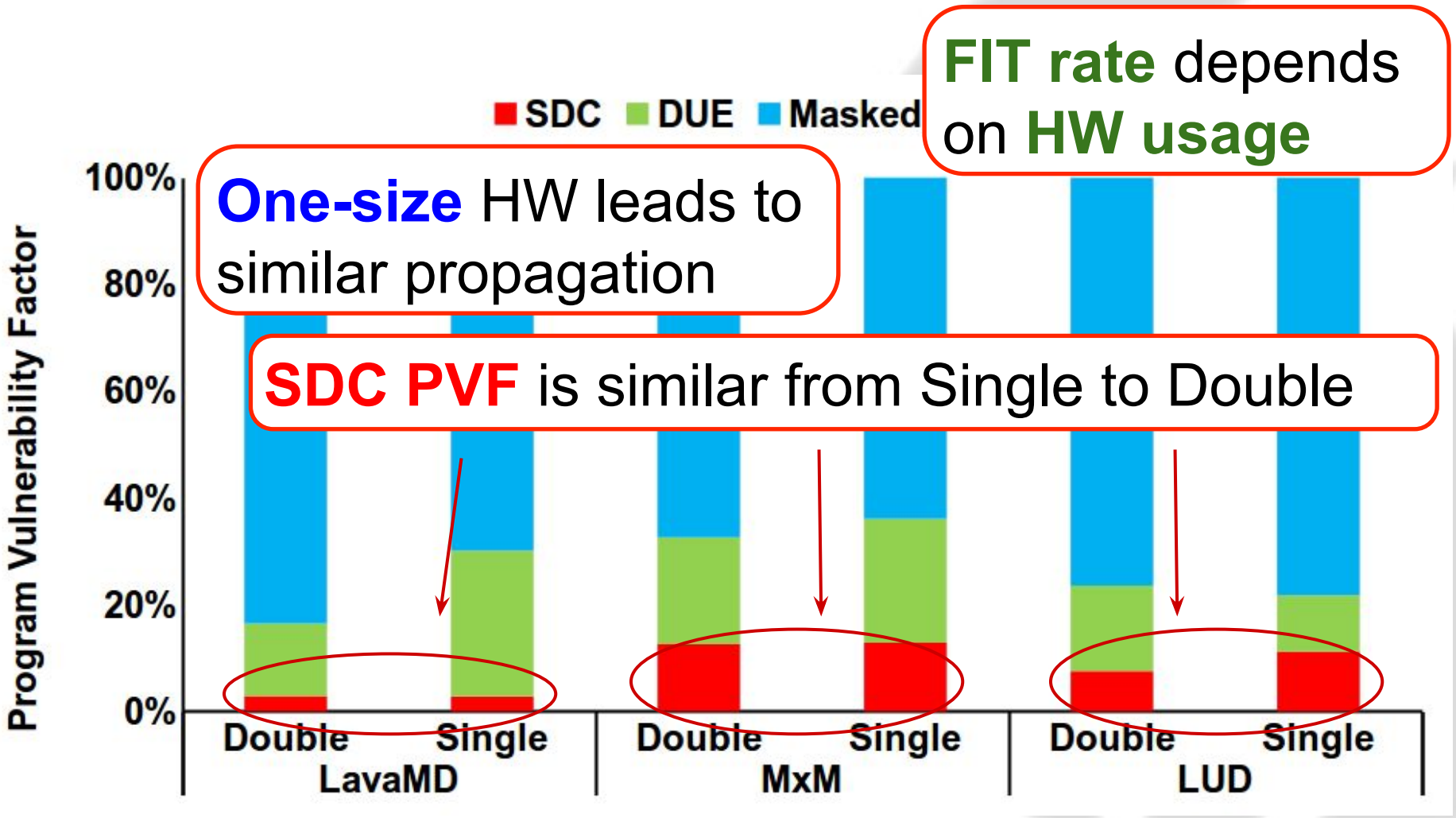
Xeon Phi - Fault Injection

We **inject faults** in variables during execution



Xeon Phi - Fault Injection

We **inject faults** in variables during execution

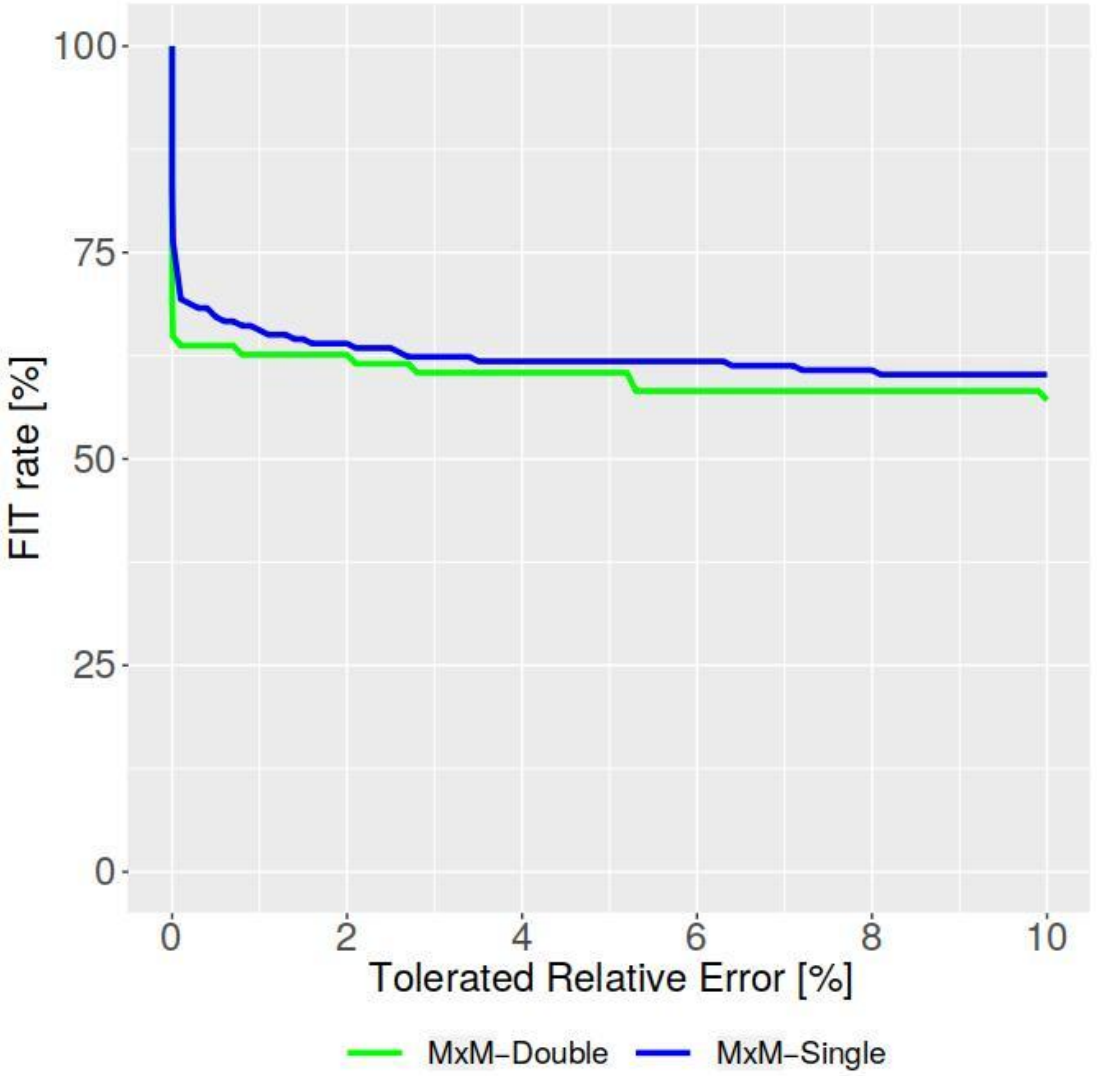


FIT rate depends on **HW usage**

One-size HW leads to similar propagation

SDC PVF is similar from Single to Double

Xeon Phi - Criticality



(more codes are shown in the paper)

Nearly identical error rates.

For MxM, the gap between the error rates is always **less than 3%.**

Outline

- Radiation Effects Introduction
- Experimental Methodology
- **Reliability of Mixed-Precision Architectures**
  - FPGAs
  - x86
  - **GPUs**
- Conclusions

Volta GPU

2,688 cores

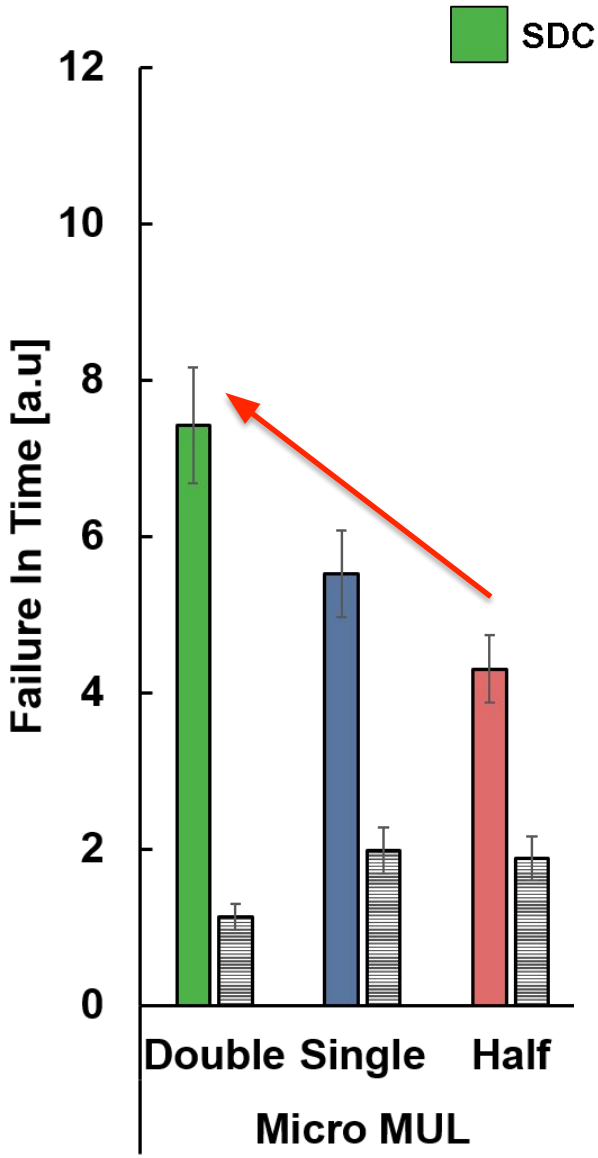
5,376 cores

5,376 cores

Clock cycles depend only on data-type, not on operation. MUL, ADD, FMA have different complexity.

We test micro-benchmarks to investigate GPU reliability (realistic codes are shown in the paper)

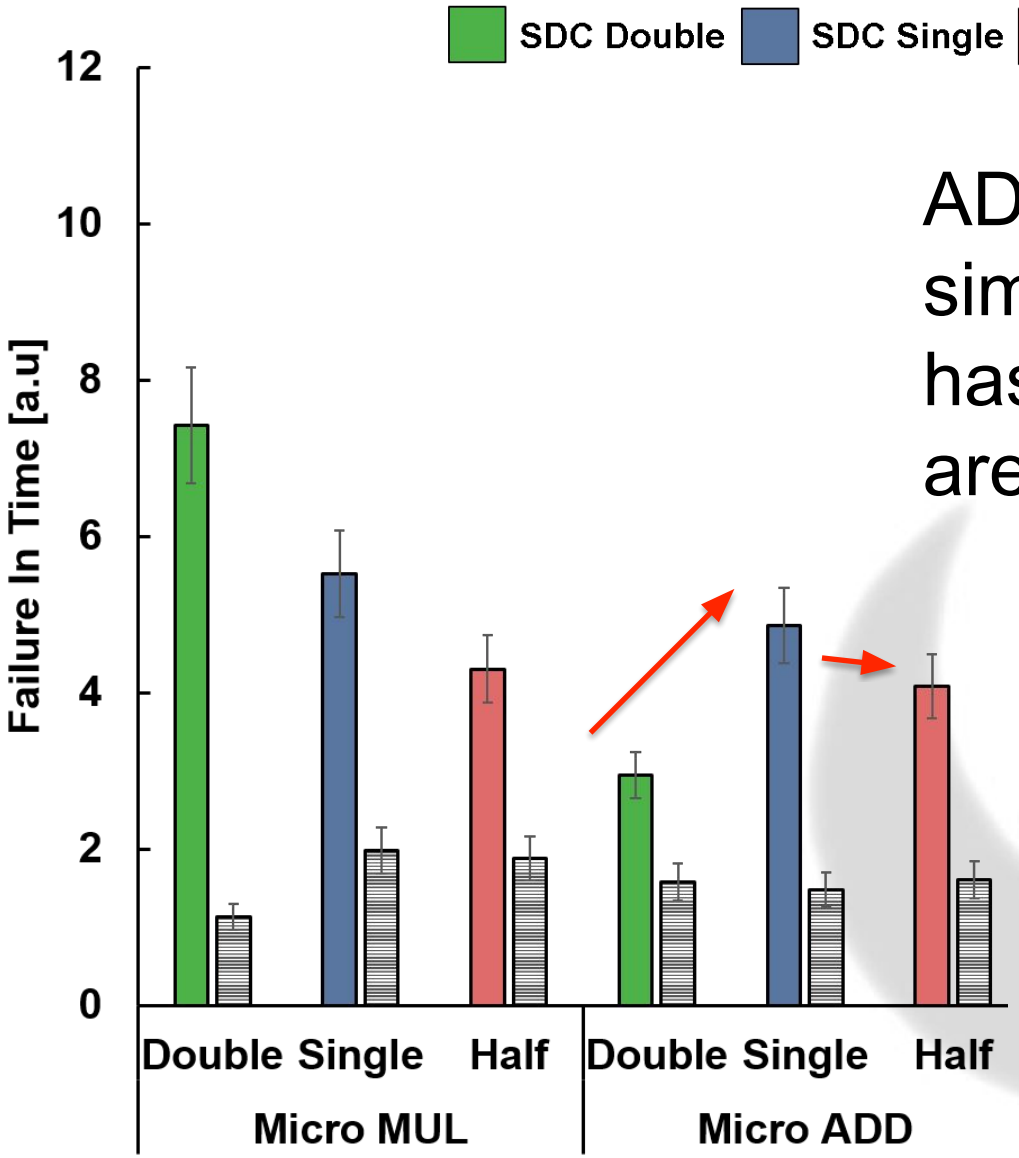
Volta GPU - FIT rates



Double has higher FIT rate even if there are fewer cores than **Single** and **Half** (**2,688** vs **5,376**)

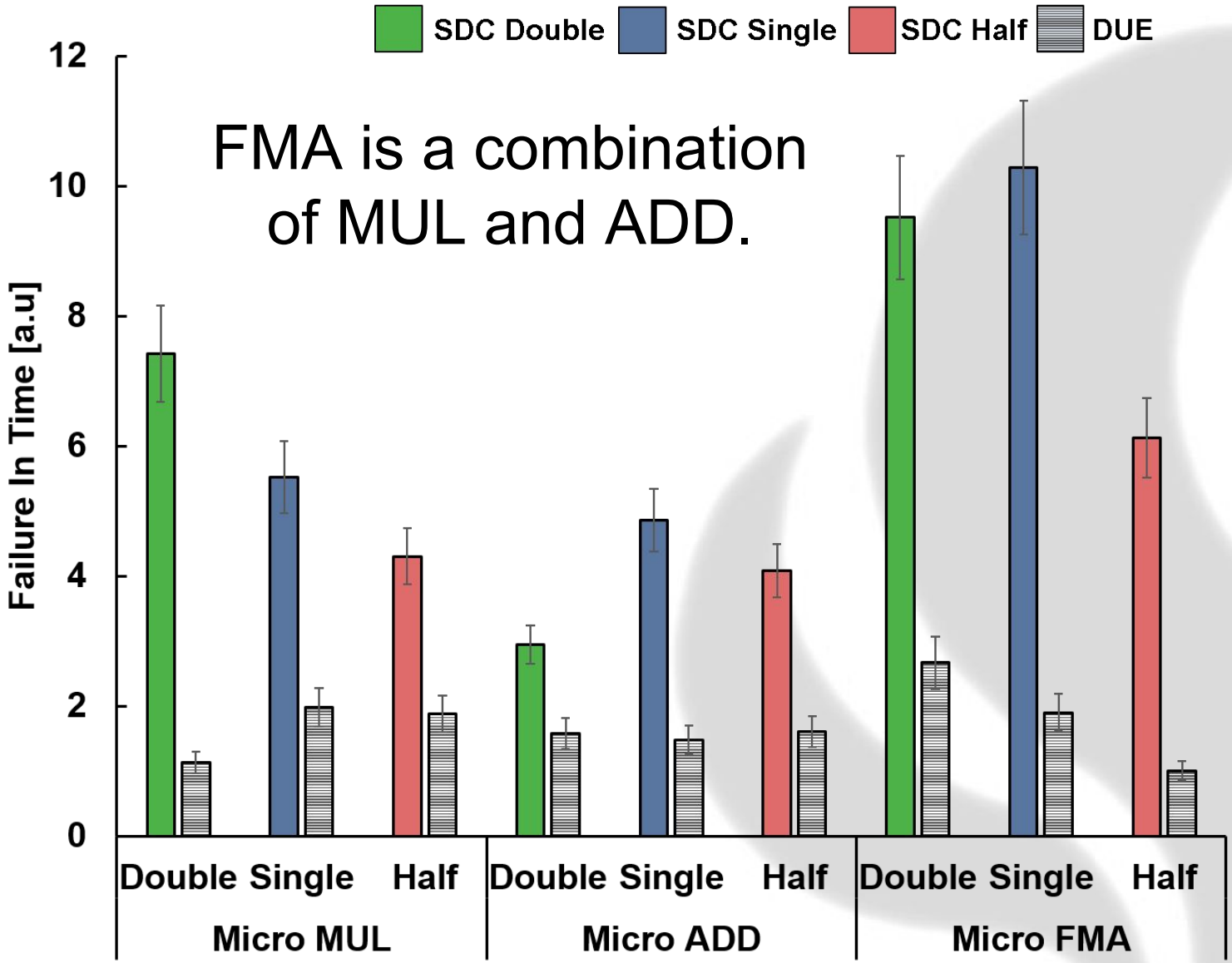
MUL hardware complexity (and, then, error rate) increases a lot when precision is increased

Volta GPU - FIT rates



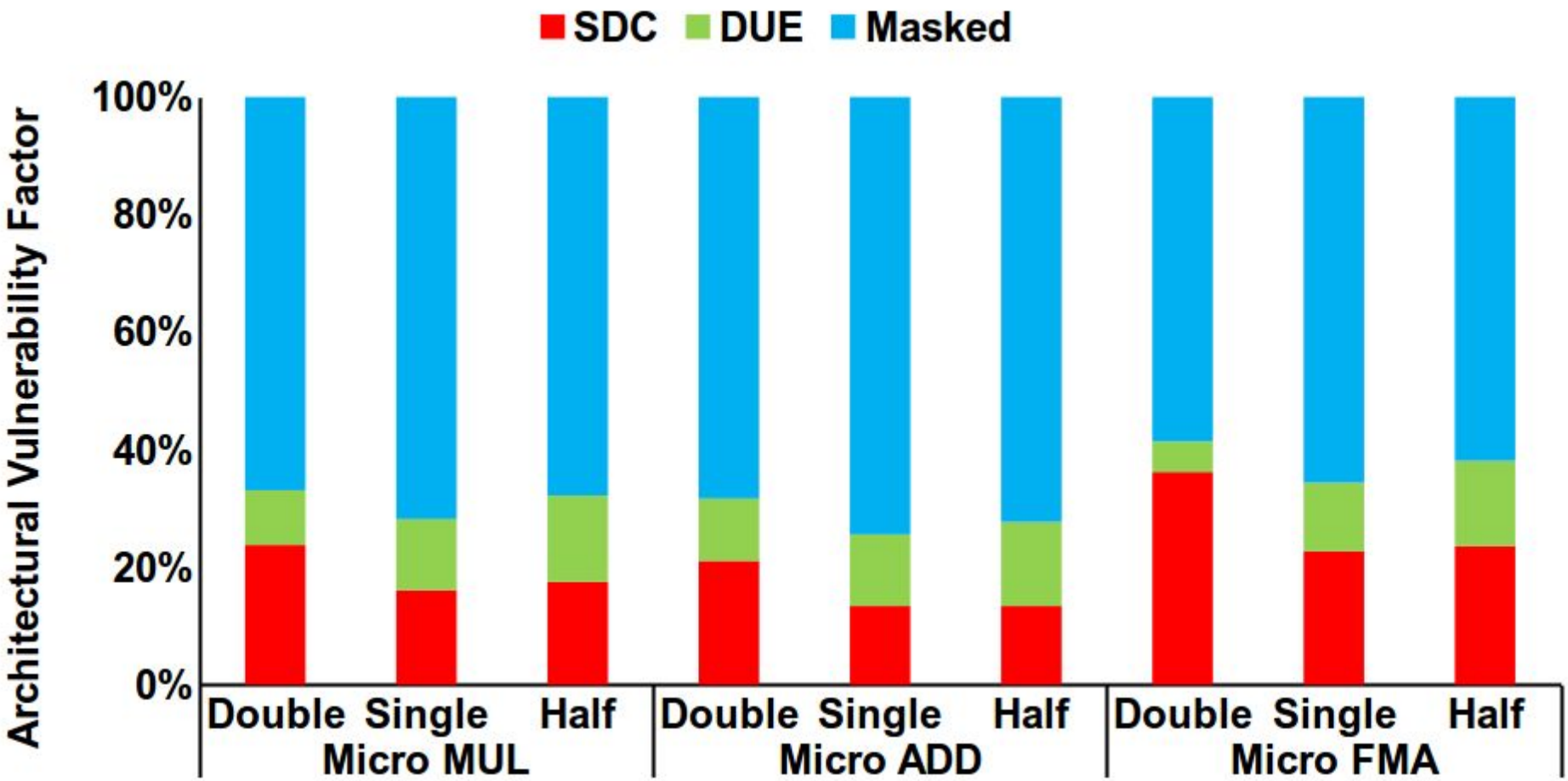
ADD hardware is much simpler than MUL. **Double** has lower FIT rate as there are fewer cores.

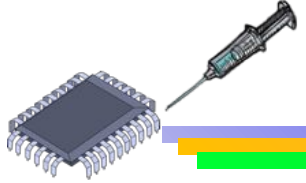
Volta GPU - FIT rates



GPU - fault injection

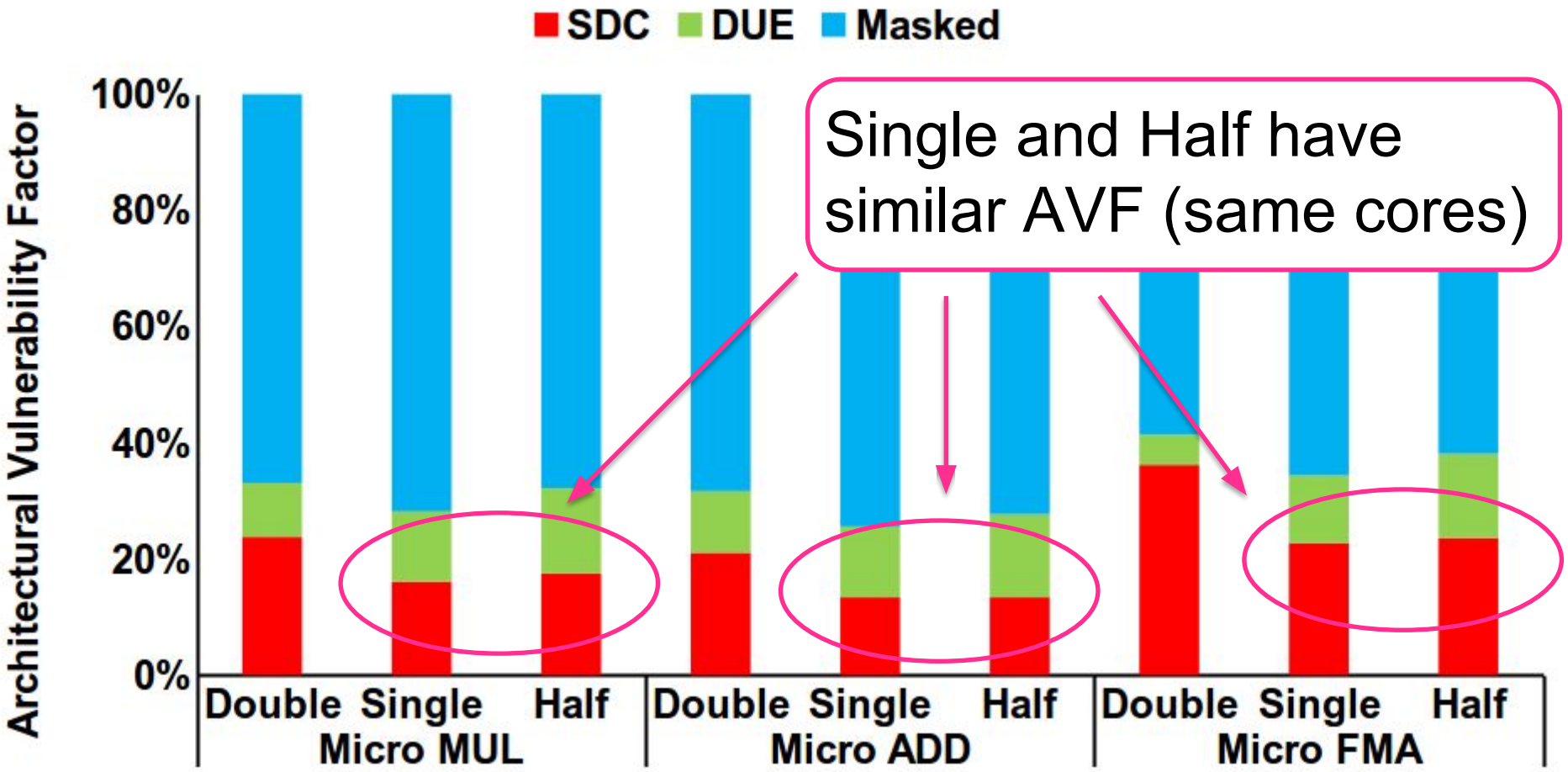
We inject errors in registers (lowest possible level, yet)





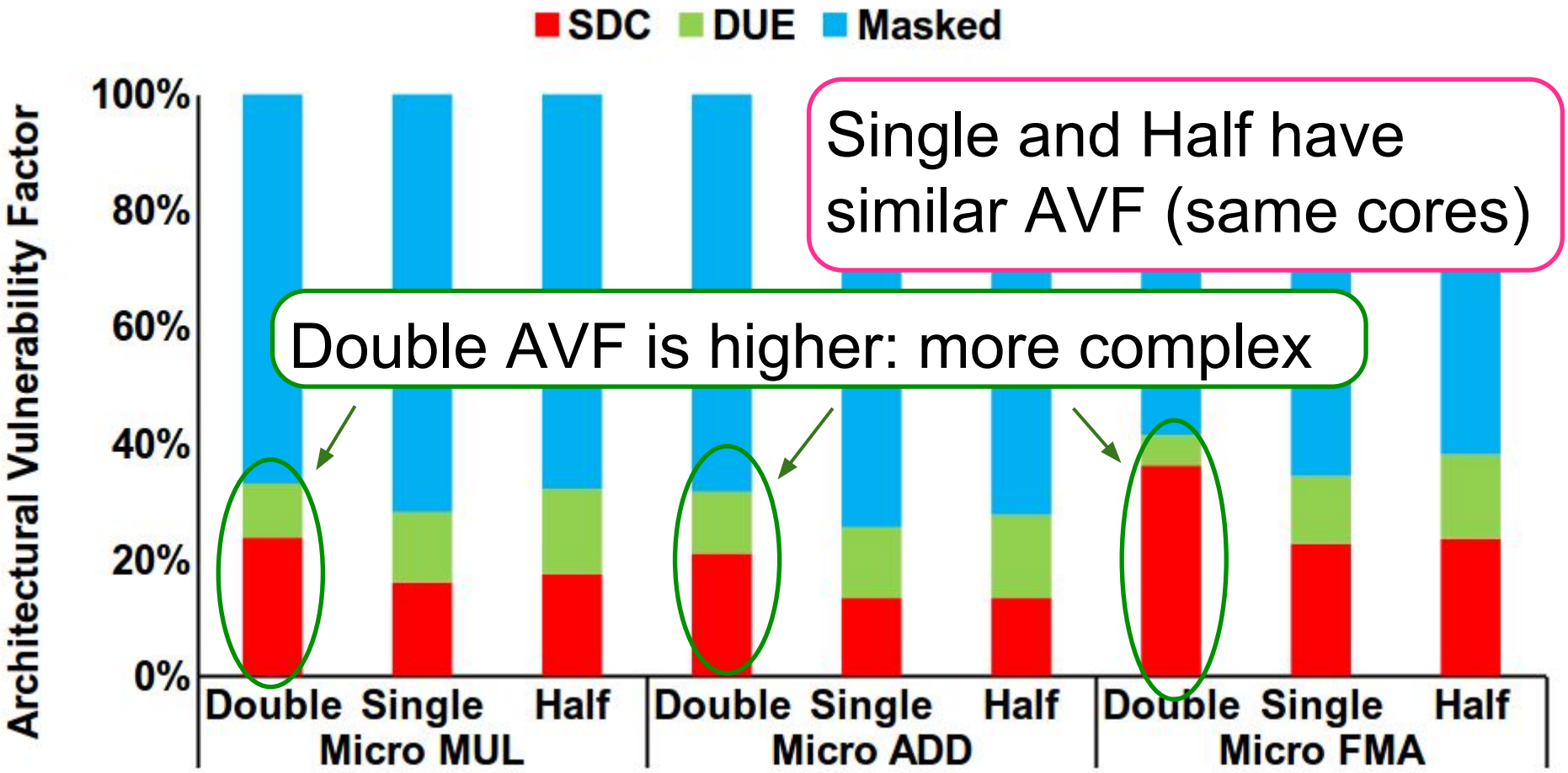
GPU - fault injection

We inject errors in registers (lowest possible level, yet)

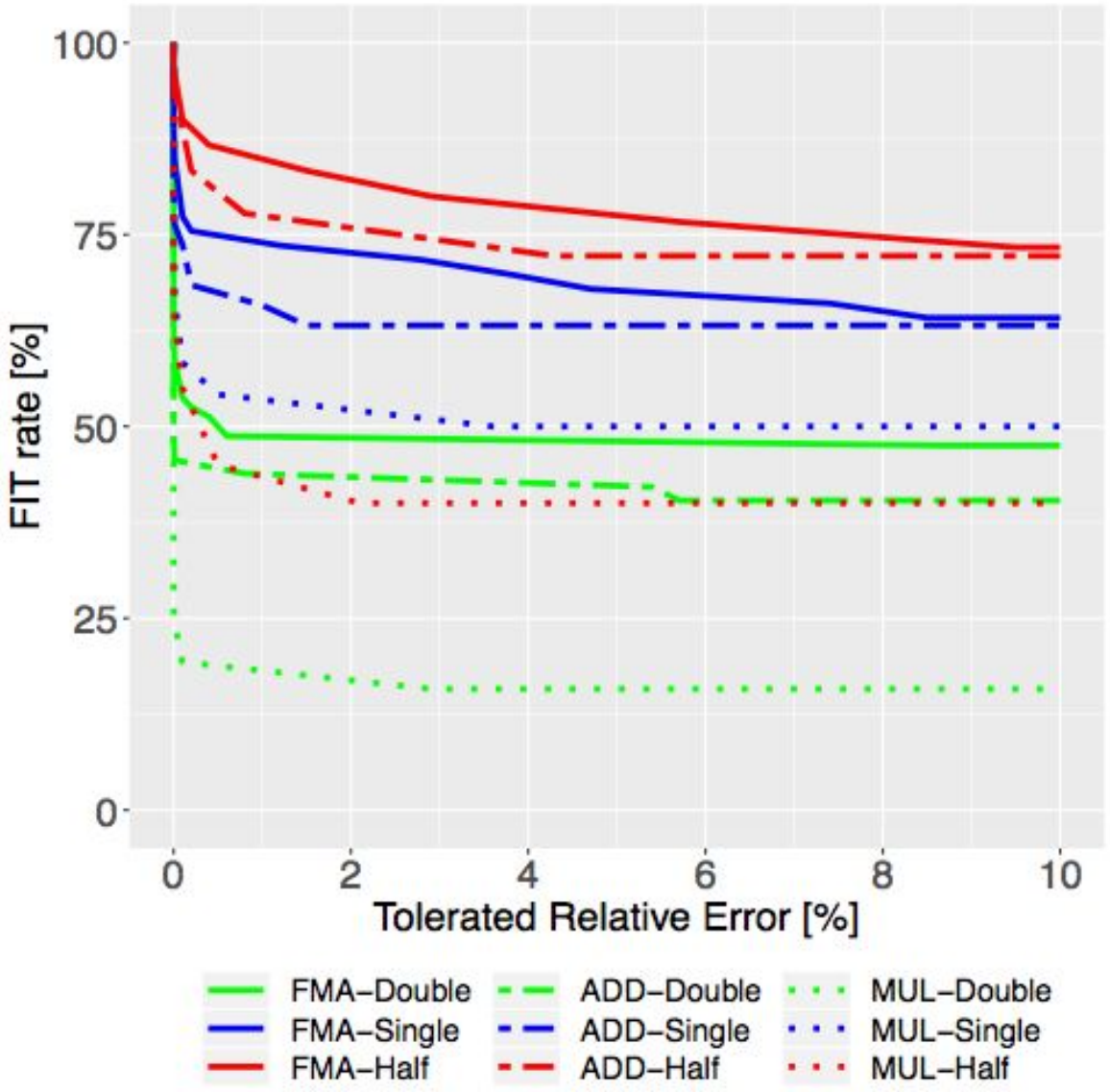


GPU - fault injection

We inject errors in registers (lowest possible level, yet)



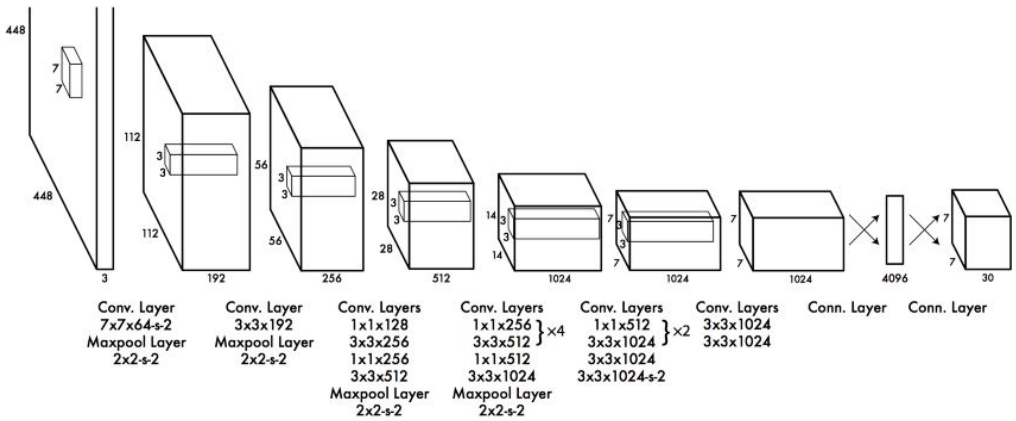
GPU - Error Criticality

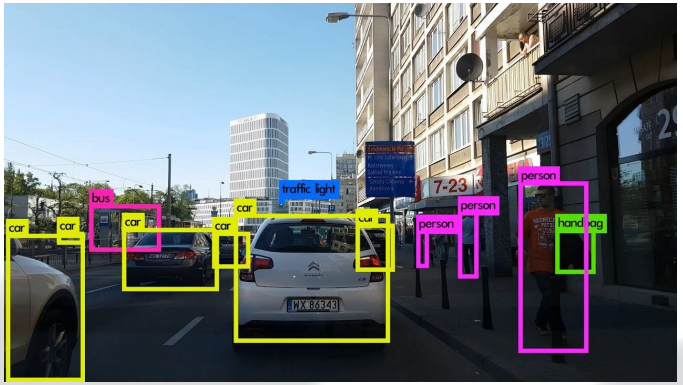


Precision

Error Criticality

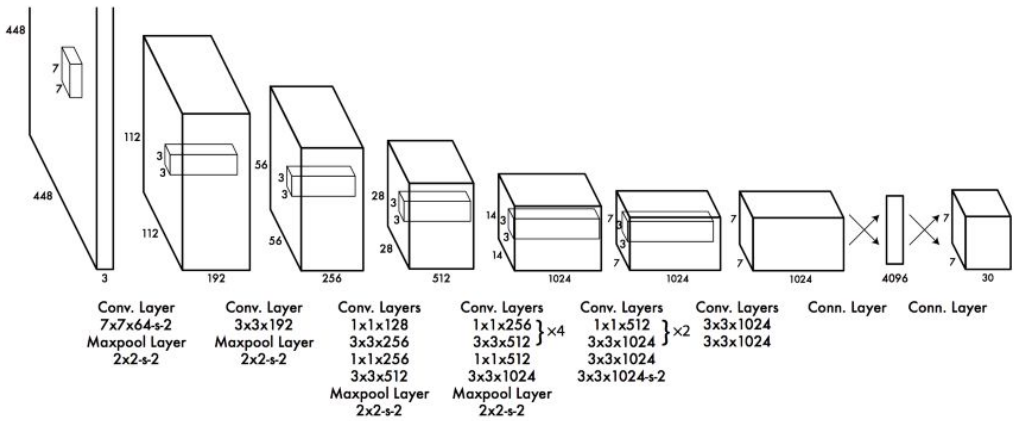
GPU - Neural Networks

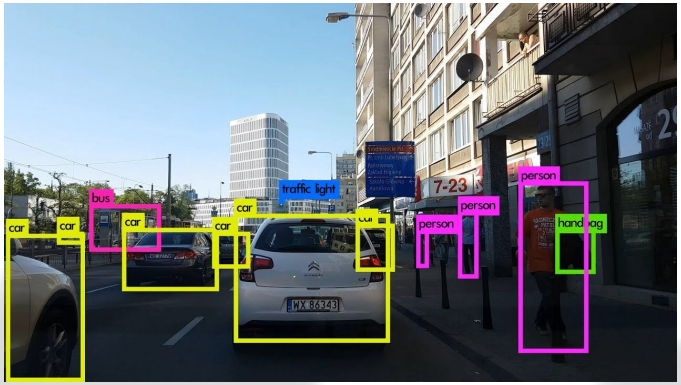




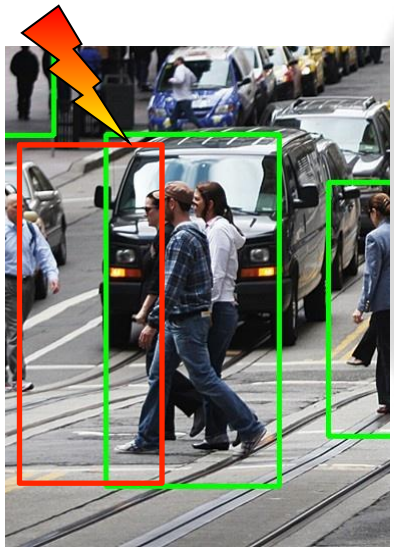
We tested YOLOv3 implemented in Double, Single, Half precision

GPU - Neural Networks





We tested YOLOv3 implemented in Double, Single, Half precision

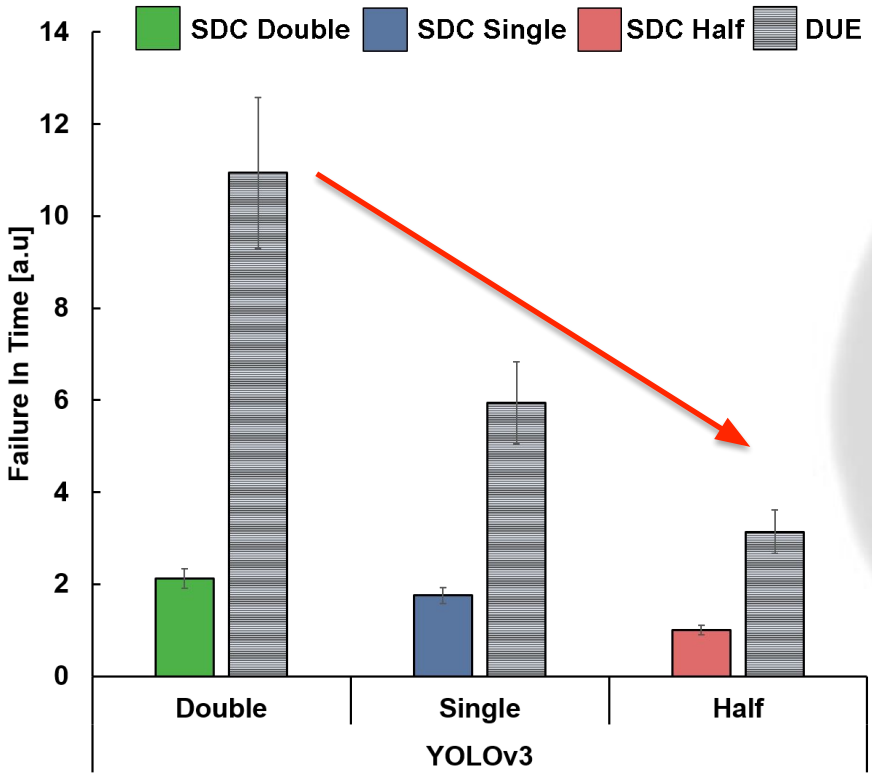


We consider as **critical** faults that significantly modify detection/classification

GPU - Neural Networks

↓ FIT rate

↓ Precision

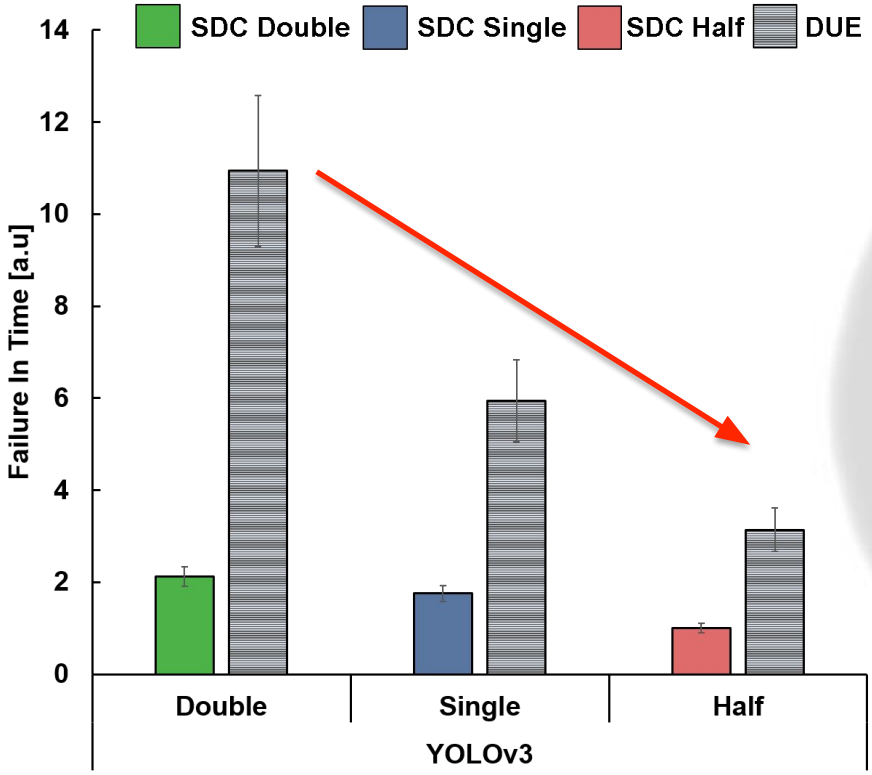


GPU - Neural Networks

↓ Precision

↓ FIT rate

↑ Error criticality



Outline

- Radiation Effects Introduction
- Experimental Methodology
- Reliability of Mixed-Precision Architectures
  - FPGAs
  - x86
  - GPUs
- **Conclusions**

Conclusions and Future Work

-Mixed precision architectures significantly improve performance

-Reducing precision impacts the code error rate in a non-obvious way

-Low precision can improve reliability: more data can be processed before experiencing an error (details in the paper)

-Future work: duplication using mixed-precision to detect critical faults with low-overhead


BEAM ON
RADHI

Acknowledgments

Lucas Weigel
Lucas Klein
Pedro Pimenta
Philippe Navaux
Luigi Carro

Nathan DeBardeleben
Sean Blanchard
Heather Quinn
Thomas Fairbanks
Steve Wender

Chris Frost
Carlo Cazzaniga

Timothy Tsai
Siva Hari
Michael Sullivan
Steve Keckler

Matteo Sonza Reorda
Luca Sterpone

John Brunhaver

Reliability Evaluation of Mixed-Precision Architectures

Fernando F. Santos, Caio Lunardi, Daniel Oliveira, Fabiano Libano, and Paolo Rech

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf
INSTITUTO DE INFORMÁTICA
UFRGS