Reliability Evaluation of Mixed-Precision Architectures

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

Double (64 bit) [ ] [ ]

Single (32 bit) [ ] [ ]

- Precision
- Area
- Circuit complexity
- Power consumption
- Execution time
Mixed-precision architectures can improve performance, efficiency and reliability.
Approximate computing has shown that some operations in some algorithms can be approximated without affecting significantly the final result.
Mixed-Precision Data and Operations

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

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.

Most of CNNs weights can be represented in half precision.
Reduced precision operations are particularly interesting for **Neural Networks training and execution**.

### Neural Nets accuracy*

<table>
<thead>
<tr>
<th>Network</th>
<th>FP32 Baseline</th>
<th>Mixed precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>56.8%</td>
<td>56.9%</td>
</tr>
<tr>
<td>VGG-D</td>
<td>65.4%</td>
<td>65.4%</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>68.3%</td>
<td>68.4%</td>
</tr>
<tr>
<td>Inception v2</td>
<td>70.0%</td>
<td>70.0%</td>
</tr>
<tr>
<td>Inception v3</td>
<td>73.9%</td>
<td>74.1%</td>
</tr>
<tr>
<td>Resnet 50</td>
<td>75.9%</td>
<td>76.0%</td>
</tr>
<tr>
<td>ResNeXt 50</td>
<td>77.3%</td>
<td>77.5%</td>
</tr>
</tbody>
</table>

*data from nvidia

---

**Mixed-Precision Data and Operations**

Half (16 bit)

Short INT (8 bit)

(1 bit)
Reduced precision operations are particularly interesting for **Neural Networks** training and execution.
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

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 Galst
Posted 23 Feb 2016 | 18:00 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

COSMIC RAY SHOWERS CRASH SUPERCOMPUTERS. HERE'S WHAT TO DO ABOUT IT
THE UBER CRASH WON’T BE THE LAST SHOCKING SELF-DRIVING DEATH
The Importance of Reliability

Reliability is a critical issue!
Galactic Cosmic rays interact with atmosphere and produce a shower of energetic particles:
- Muons
- Pions
- Protons
- Gamma Rays
- Neutrons

13 n/(cm$^2 \times h$) @sea level*

*JEDEC JESD89A Standard
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

**Silent Data Corruption**
- data cache
- register files
- logic gates (ALU)
- scheduler

**DUE (Crash)**
- instruction cache
- scheduler / dispatcher
- PCI-e bus controller
Outline

- Radiation Effects Introduction

- Experimental Methodology

- Reliability of Mixed-Precision Architectures
  - FPGAs
  - x86
  - GPUs

- Conclusions
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 VPU.

**GPU (Volta V100):** dedicated HW for double and single/half precision
1 double operation, 1 single or 2 half operations
Fault Injection
Fault Injection

We purposefully inject faults in registers and variables.
Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.
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
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
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
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.
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.
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
Fault Injection

We purposely inject faults in registers and variables.

Probability of faults propagating to the output.

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.

Failure in Time
Radiation Experiments @ChipIR

NVIDIA Volta V100
Radiation Experiments @ChipIR

Intel Xeon-Phi

NVIDIA Volta V100
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.
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

Input image: 28x28 pixels
Input layer: 784 neurons, one per pixel
Hidden layer: 100 neurons
Output layer: 10 neurons
Output: predicted digit value

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**

Input image: 28x28 pixels

Input layer: 784 neurons, one per pixel

Hidden layer: 100 neurons

Output layer: 10 neurons

Output: predicted digit value

**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**

I/O Block

Logic Block

Reliability Evaluation of Mixed-Precision Architectures – INF, UFRGS
MNIST

Matrix Multiplication (128x128)

\[
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
\end{bmatrix}
\times
\begin{bmatrix}
2 & 0 \\
1 & 2 \\
\end{bmatrix}
=
\begin{bmatrix}
4 & 4 \\
10 & 8 \\
\end{bmatrix}
\]
Reliability Evaluation of Mixed-Precision Architectures – INF, UFRGS

![Graph showing the utilization of resources for different precision levels (Double, Single, Half) for MNIST and MxM operations. The graph indicates a decrease in precision and resources as the precision level decreases.](image)

- **MNIST**
  - Double: 18,000 resources
  - Single: 14,000 resources
  - Half: 10,000 resources

- **MxM**
  - Double: 6,000 resources
  - Single: 4,000 resources
  - Half: 2,000 resources

Arrow indicating a decrease in precision and resources as the precision level decreases.
FPGA - FIT rate

- SDC Double
- SDC Single
- SDC Half
- Critical

(Wrong Classification)

Failure in Time [a.u.]

MNIST

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

The graph illustrates the FIT rate (%) against the tolerated relative error (%) for different precision levels: Double, Single, and Half. The curves show how the FIT rate changes with increasing tolerated relative error for each precision level.
Increasing acceptable difference at the output
FPGA - error criticality

Output must match the expected output (0% tolerance)
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
One-size fits all: Xeon Phi does not have dedicated HW for double/single

![Graph showing failure in time for different operations and precision levels.](Image)
One-size fits all: Xeon Phi does not have dedicated HW for double/single.

The compiler decides how to use the functional unit.
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
We inject faults in variables during execution.
We inject faults in variables during execution.

SDC PVF is similar from Single to Double.
We inject faults in variables during execution.

One-size HW leads to similar propagation.

SDC PVF is similar from Single to Double.
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.
Nearly identical error rates.

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

(more codes are shown in the paper)
Outline

- Radiation Effects Introduction
- Experimental Methodology
- Reliability of Mixed-Precision Architectures
  - FPGAs
  - x86
  - GPUs
- Conclusions
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.
ADD hardware is much simpler than MUL. **Double** has lower FIT rate as there are fewer cores.
FMA is a combination of MUL and ADD.
GPU - fault injection

We inject errors in registers (lowest possible level, yet)
GPU - fault injection

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

Single and Half have similar AVF (same cores)
We inject errors in registers (lowest possible level, yet).

Single and Half have similar AVF (same cores)

Double AVF is higher: more complex.
GPU - Error Criticality

Precision

Error Criticality

![Graph showing the relationship between FIT rate and tolerated relative error for different operations (FMA, ADD, MUL) across different precision levels (Double, Single, Half). The graph indicates how error criticality affects precision.]
We tested YOLOv3 implemented in Double, Single, Half precision.
We tested YOLOv3 implemented in Double, Single, Half precision.

We consider as **critical** faults that significantly modify detection/classification.
GPU - Neural Networks

Precision → FIT rate

![Graph showing failure rate with double, single, and half precision for YOLOv3](image)

- **SDC Double**
- **SDC Single**
- **SDC Half**
- **DUE**

Failure in Time [a.u.]

- Double
- Single
- Half

Reliability Evaluation of Mixed-Precision Architectures – INF, UFRGS
Reliability Evaluation of Mixed-Precision Architectures – INF, UFRGS
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
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