Microarchitecture Level Reliability Comparison of modern GPU Designs: first findings

VALLERO Alessandro*, TSELONIS Sotiris+, GIZOPOULOS Dimitris+, DI CARLO Stefano*
* Politecnico di Torino, + University of Athens

ABSTRACT

State-of-the-art GPU chips are designed to deliver extreme throughput for graphics as well as for data-parallel general purpose computing workloads (GPGPU computing).

Unlike graphics computing, GPGPU computing requires highly reliable operation. The performance-oriented design of GPUs requires the vulnerability of GPU workloads to soft-errors to be jointly evaluated with the performance of GPU chips.

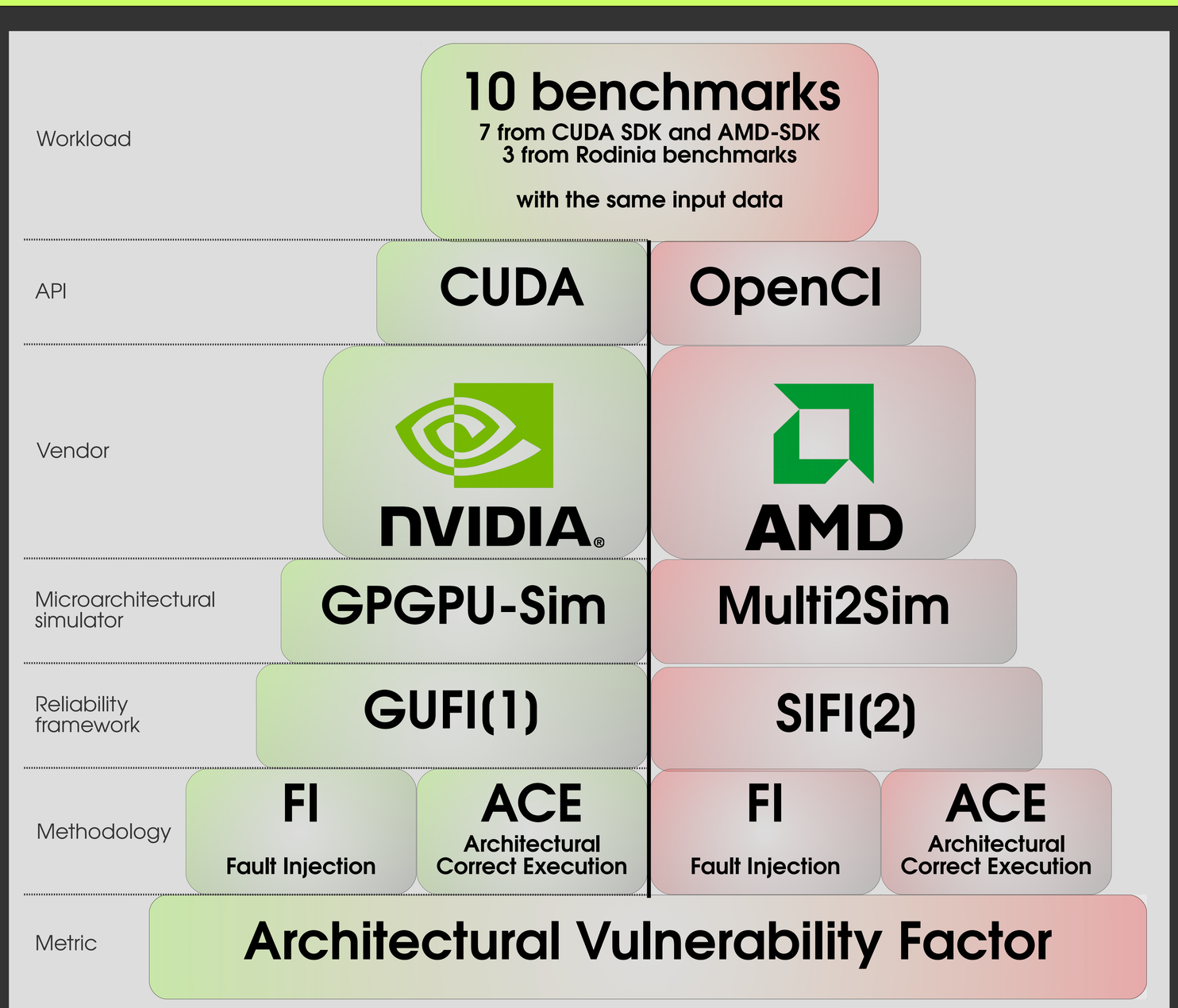
We present the preliminary results of an extensive study aiming at the evaluation of the reliability of four GPU architectures and corresponding chips in correlation with the performance.

INTRODUCTION

Recently, the research community has started tackling the challenging problem of characterizing the reliability of GPGPU based systems, i.e., their vulnerability to soft- and hard-errors. This challenging problem requires the development of accurate and fast reliability assessment techniques to deal with the delicate trade-off between analysis time and accuracy of the reported measurements and able to provide results able to guide system designers in the choice and development of efficient error resilience mechanisms.

This work aims at:

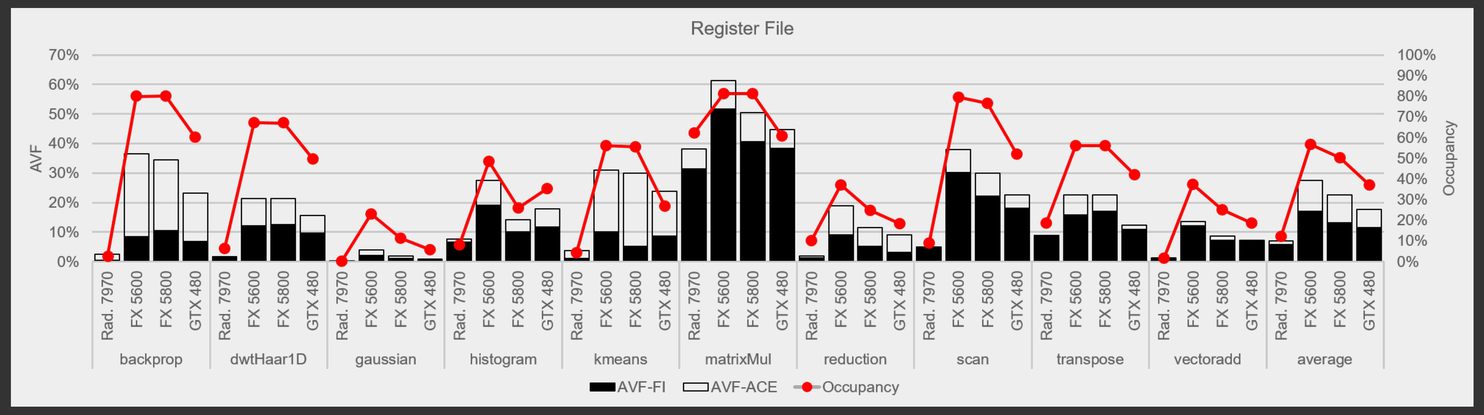
- Showing preliminary results of an extensive study aimed at evaluating the hardware and software features that influence the reliability of GPGPU chips in the presence of soft-errors.
- Comparing reliability and performance of several GPUs from different vendors, architectures, programming model and computational power.
- Evaluating different methodologies for reliability assessment to identify trade-off between analysis time and accuracy of results
- Introducing a metric to jointly analyze reliability and performance.

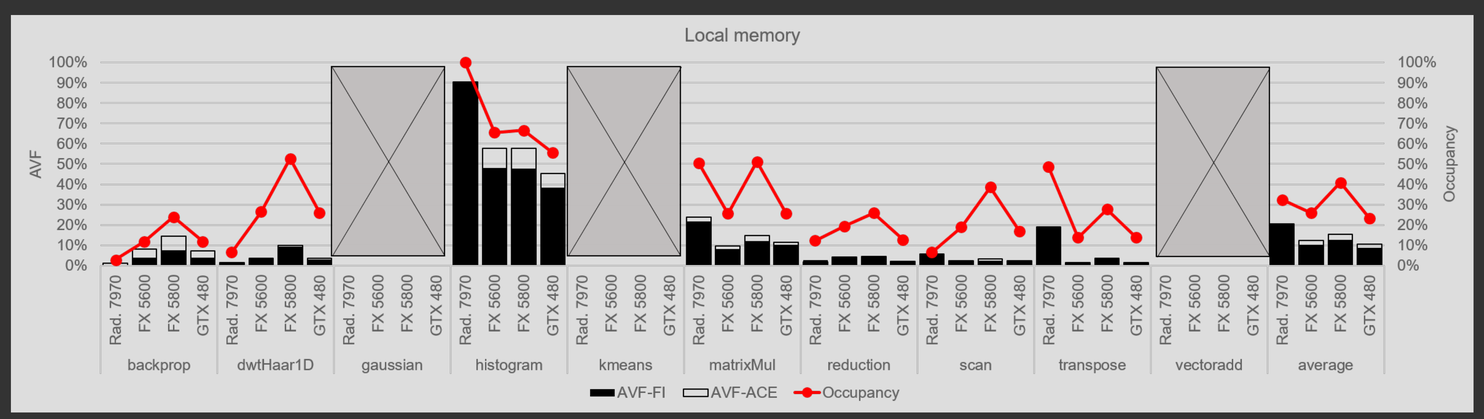


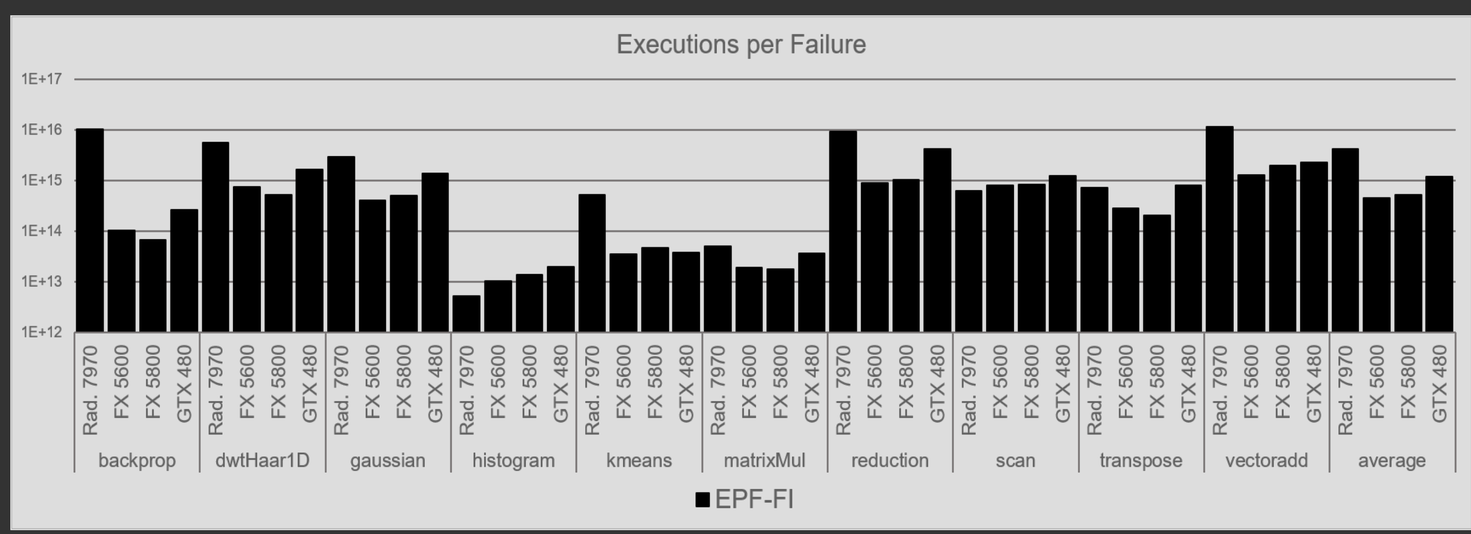
THE GPU COMPARISON

We evaluated reliability and performance of the most important GPU families of different vendors, microarchitectures, ISAs, computational models using the same set of benchmarks.
Concerning the reliability analysis, we developed a framework to perform fault injection campaigns and ACE analysis for the selected GPUs, targeting the general purpose register file and the local memory. We computed AVF for these memory arrays aiming at correlating it with their size and occupancy alongside the execution scheduling.

Chip name		Quadro™ FX 5600	Quadro™ FX5800	Geforce™ GTX 480	HD Radeon™ 7970
Architecture		G80	GT 200	Fermi	Southern Islands
Frequency		337.5 MHz	325 MHz	700 MHz	925MHz
Technology		90 nm	55 nm	40 nm	28 nm
Register File		32KB	64KB	128KB	256KB
Local Memory		16KB	16KB	48KB	64KB
SIMD Units		1	1	2	4
Max	#work-groups	8	8	8	40
	#wavefronts	24	32	48	40
	#work-items	768	1024	1536	1840







EXECUTIONS PER FAILURE

We introduced Executions Per Failure (EPF), a new metric to evaluate reliability and performance jointly:

$EPF = \frac{EIT}{FIT_{GPU}}$,where EIT is the number of executions in 10^9 hours, while FIT_{GPU} is the Failures In Time of the GPU and it is computed as:

$$FIT_{GPU} = AVF_{RF} \times \lambda_{tech} \times nBits_{RF} + AVF_{LM} \times \lambda_{tech} \times nBit_{LM}$$

where λ_{tech} is the raw FIT per bit of technology obtained from (3).

PRELIMINARY RESULTS

Results show that the AVF can have significant variations moving from one application to another but also variations can be observed for the same application executed on different GPUs. Red lines reporting the occupancy of the considered memory structures show a strong correlation of the AVF with this parameter.
It is interesting to note that while for the register file the ACE analysis significantly overestimates vulnerability compared to FI, the same technique is very accurate (very close to FI) for the local memory, suggesting that for this structure ACE analysis can be used without significant loss of accuracy.
Larger EPF numbers show a larger number of executions between failures and different protection mechanisms can deliver different improvements in the FIT rates and can also have different impact on performance. Combining performance and reliability measurements in the EPF metric delivers a broader view for decision-making.

REFERENCES

(1)S. Tselonis et al., "GUFI: A framework for GPUs reliability assessment," ISPASS, 2016

(2) A. Vallero et al., "SIFI: AMD Southern Island GPU Microarchitectural Level Fault Injector ", IOLTS 2017

(3) E. Ibe et. Al, "Impact of Scaling on Neutron-Induced Soft Error in SRAMs From a 250 nm to a 22 nm Design Rule," in IEEE Transactions on Electron Devices, 2010

CONTACTS

ALESSANDRO VALLERO
alessandro.vallero@polito.it

SOTIRIS TSELONIS
tseloniss@di.uoa.gr

DIMITRIS GIZOPOULOS
dgizop@di.uoa.gr

STEFANO DI CARLO
stefano.dicarlo@polito.it