ICR The Institute of Cancer Research | Imperial College London

CANCER RESEARCH UK | CONVERGENCE SCIENCE CENTRE

Fast, high-quality PRNG for heterogeneous computing

Marco Barbone, Georgi Gaydadjiev, Alexander Howard, Wayne Luk, George Savvidy, Konstantin Savvidy, Andy Rose, Alexander Tapper

m.barbone19@imperial.ac.uk

Imperial College London

Monte Carlo simulations

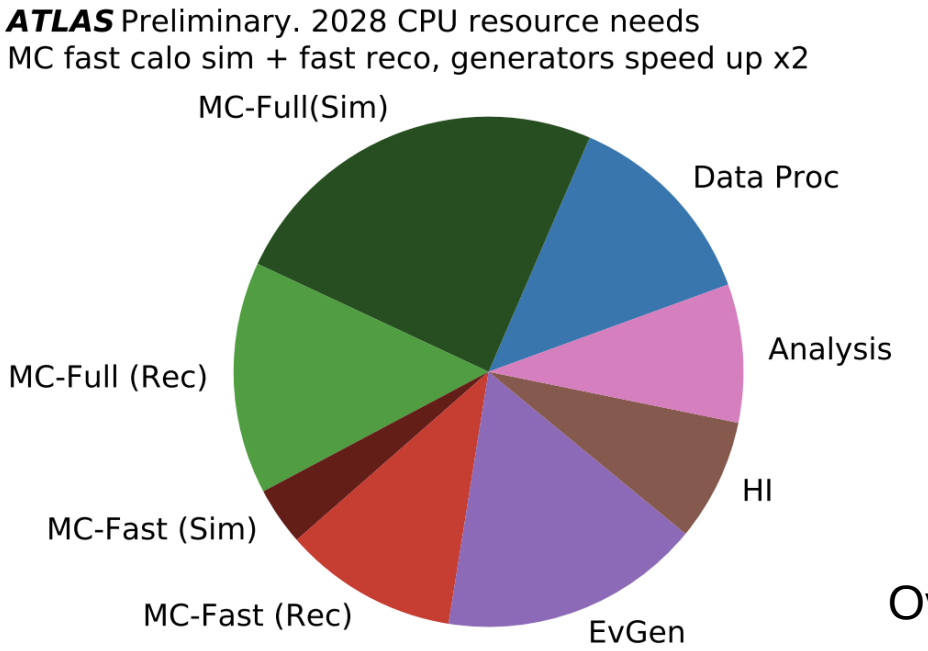


Figure 2: Breakdown of estimated compute workloads in 2028 for ATLAS

Over 50% is required by Monte Carlo related workloads

https://doi.org/10.22323/1.390.0009

Imperial College London

Motivation

GPUs/FPGAs RNGs are:

- Slow and high-quality
- Fast and low-quality
- In the GPU case, closed-source

There is no middle-ground

https://docs.nvidia.com/cuda/curand/index.html

Imperial College London

MixMax RNG

It is a high-quality generator suitable for MC simulations

Quality, speed tunable based on state-size

Even small state size offers high quality numbers

It offers a seeding mechanism that guarantees no collisions between streams

Used in CLHEP & Geant4 (default)

https://doi.org/10.1051/epjconf/201921402012

Imperial College London

Goal

Accelerate MixMax and compare the performance against state-of-the-art RNGs

Provide a reliable RNG for Monte Carlo simulations

Imperial College London

FPGA-VHDL design

- VHDL-2008
- 62 SLOC
- Clear interface
- Fully pipelined, 100% duty-cycle

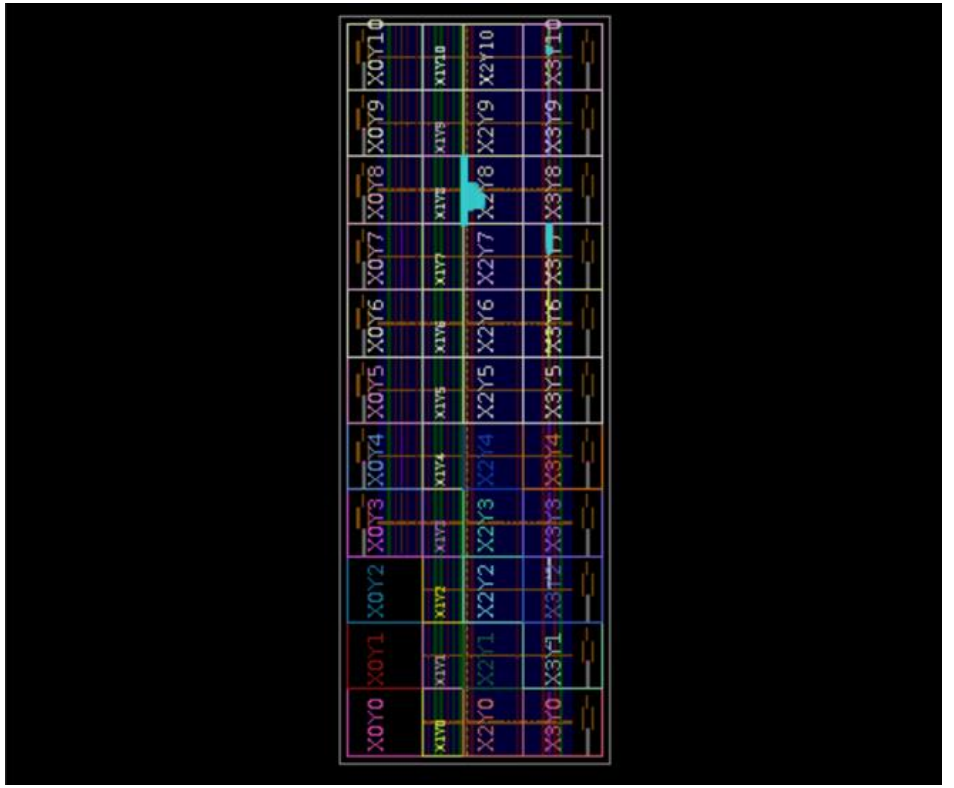
Imperial College London

Experimental Setup

- Xilinx Ultrascale+ FPGA
- Vivado 2020.2
- Default settings

Imperial College London

Performance



- Achieves 300 MHz
- Requires 550 LUTs
  - c.f. 523k available in KU15P, 1.75M available in VU13P
- This is 50% of the Mersenne Twister resources
- or 25% of the Mersenne Twister resources per output bit

https://doi.org/10.1145/272991.272995

Imperial College London

GPU C++/CUDA design

- Each thread contains an RNG, state is not shared
- GPU is seeded with 128 bits, the thread id is concatenated to generate different streams
- Only 128 bits needs to be transferred to the GPU
- The user can change this behavior
- Clear C++ interface

Imperial College London

Experimental setup

- NVIDIA 3090 Ti
- nvcc 11.8
- g++ 9.4.0

Imperial College London

Validation

Compared against the original implementation

Test took ~3 months (server crashed due to blackout....)

Results never diverged

Imperial College London

Performance

Parameters:

Threads	Blocks	Parallelism
128	984	125,952

iteration = generating 125,952 numbers

Algorithm	time (ms)	time/it (ns)	% vs mixmax8	Throughput (GB/s)	
Philox4_32_10	23144	10.78	-7.03	87074	[ref]
MRG32k3a	15203	7.08	-38.93	132555	[ref]
XORWOW	23206	10.81	-6.78	86841	[ref]
MixMaxGPU<240>	55294	25.75	122.11	36446	[ref]
MixMaxGPU<17>	30800	14.34	23.72	65430	[ref]
MixMaxGPU<8>	24895	11.59	0	80949	[ref]

Imperial College London

Against Mersenne twister for GPU (MTGP32)

Parameters

Threads	Blocks	Parallelism
256.00	128.00	32,768

iteration = generating 32768 numbers

Algorithm	time (ms)	time/it (ns)	% vs mixmax8	speedup	Throughput (GB/s)
MTGP32	81161	37.79	1341.84	1	6459
Philox4_32_10	7803	3.63	38.62	10	67190
MRG32k3a	5092	2.37	-9.54	16	102963
XORWOW	7731	3.60	37.34	10	67816
MixMaxGPU<240>	15996	7.45	184.17	5	32776
MixMaxGPU<17>	6075	2.83	7.92	13	86302
MixMaxGPU<8>	5629	2.62	0	14	93140

Conclusion

MixMax is 14 times faster than Mersenne Twister on GPU

25% of the Mersenne Twister resources per output bit on FPGA

Seeding is easier and faster (128 bits, compared to O(KB))

MixMax has the potential to become de-facto the standard RNG on any platform

Suitable for ML & MC due to parallelism, efficiency and quality

Future work: Delivered via CLHEP or standalone repository

ICR The Institute of Cancer Research | Imperial College London

CANCER RESEARCH UK | CONVERGENCE SCIENCE CENTRE

Backup

Imperial College London

What is MixMax?

The MIXMAX generator is a family of pseudorandom number generators (PRNG) and is based on Anosov C-systems (Anosov diffeomorphism) and Kolmogorov K-systems (Kolmogorov automorphism).

https://doi.org/10.1016/j.cpc.2015.06.003

Imperial College London

What is MixMax?

- ~~The MIXMAX generator is a family of pseudorandom number generators (PRNG) and is based on Anosov C-systems (Anosov diffeomorphism) and Kolmogorov K-systems (Kolmogorov automorphism).~~
- It is a very good PRNG
- Used in CLHEP & Geant4 (default)

https://doi.org/10.1051/epjconf/201921402012