Hard Data on Soft Errors

A Global-scale Survey of GPGPU Memory Soft Error Rates

Imran Haque
Department of Computer Science
Stanford University

http://cs.stanford.edu/people/ihaque
http://folding.stanford.edu
ihaque@cs.stanford.edu

Folding@home distributed computing

Resilience Summit @ LACSS 13 Oct 2010

Motivation

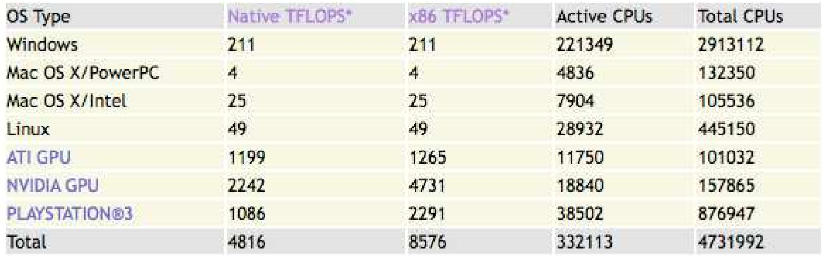
- GPUs originate in **error-insensitive** consumer graphics
- Neither ECC nor parity on most* graphics memory

- **How suitable is the installed base of consumer GPUs** (and consumer GPU-derived professional hardware!) **for *error-sensitive* general purpose computing?**

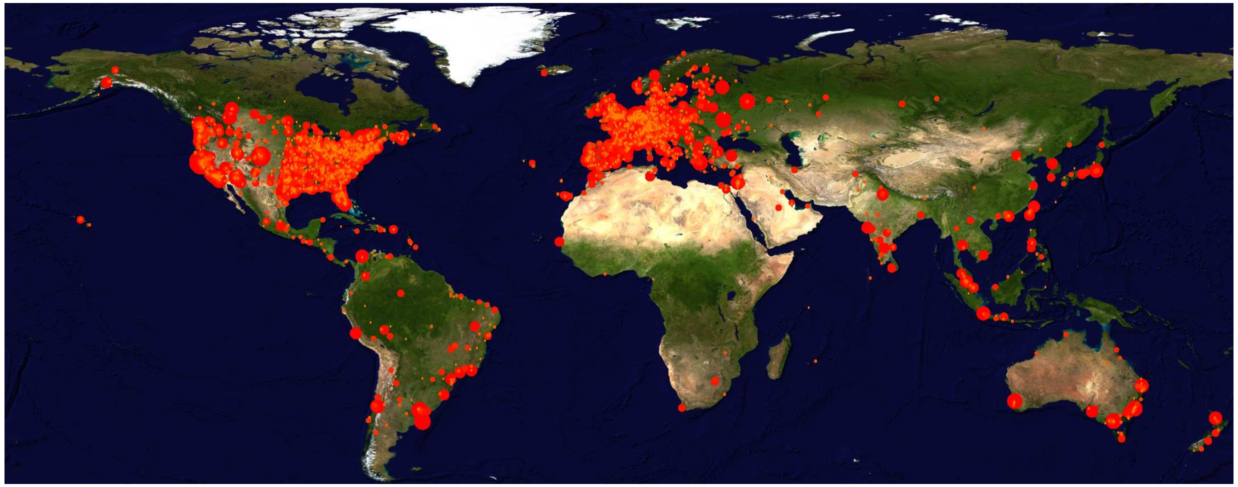
*** of which, more later**

Motivation

CUDA-Enabled Package
Folding@home (molecular dynamics)
OpenMM (molecular dynamics)
PAPER (3-D chemical similarity)
SIML (1-D chemical similarity)



OS Type	Native TFLOPS*	x86 TFLOPS*	Active CPUs	Total CPUs
Windows	211	211	221349	2913112
Mac OS X/PowerPC	4	4	4836	132350
Mac OS X/Intel	25	25	7904	105536
Linux	49	49	28932	445150
ATI GPU	1199	1265	11750	101032
NVIDIA GPU	2242	4731	18840	157865
PLAYSTATION®3	1086	2291	38502	876947
Total	4816	8576	332113	4731992



We've written a lot of GPU-enabled software,
and we run it on a lot of GPUs.

MemtestG80 + MemtestCL

- Custom software, based on Memtest86 for x86 PCs
- Open source (LGPL), available at https://simtk.org/home/memtest

- Variety of test patterns:
  - Constant (ones, zeros, random)
  - Walking ones and zeros (8-bit, 32-bit)
  - Random words (on-GPU parallel PRNG)
  - Modulo-20 pattern sensitivity
  - Novel iterated-LCG integer logic tests
  - Bit fade

MemtestG80 – Validation

- Negative control – verify that it doesn’t throw spurious errors in “known-good” situations
  - Known-good PSUs, machines located in air-conditioned environments.

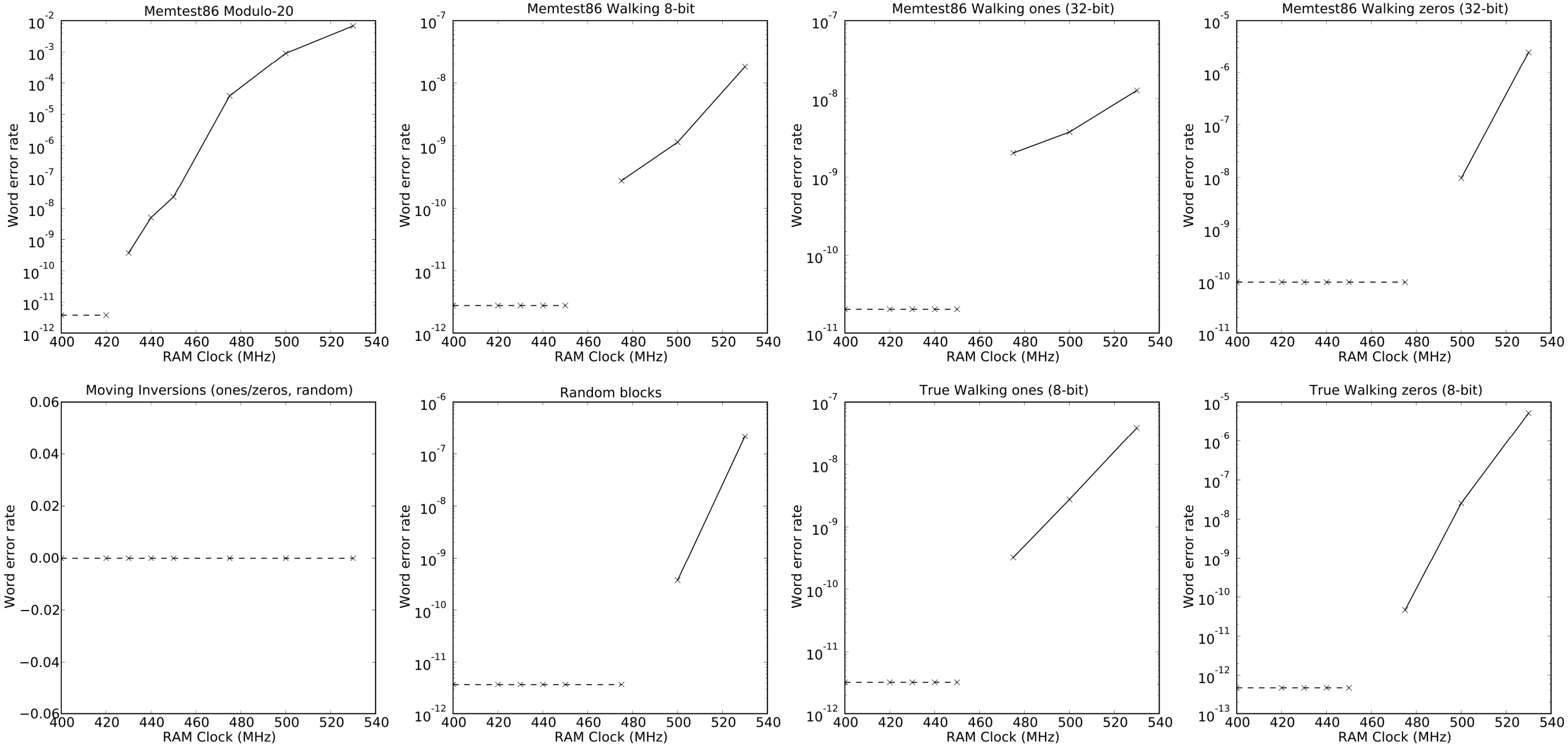
- 93,000 iterations on 700 MiB on GeForce 8800GTX
- >180,000 iters on 320MiB on each of 8 x Tesla C870

- **No errors ever detected.**

MemtestG80 – Validation

- Positive control – verify that it does throw errors in situations that generate errors
- Overclocking generates memory errors (violation of timing constraints; loss of signal integrity)
- Tested GeForce 9500GT (memory clock = 400MHz) at 400, 420, 430, 440, 450, 475, 500, 530 MHz
  - 20 iterations for each frequency (only 10 @ 530MHz)
  - Cooled down and reset to 400MHz between tests

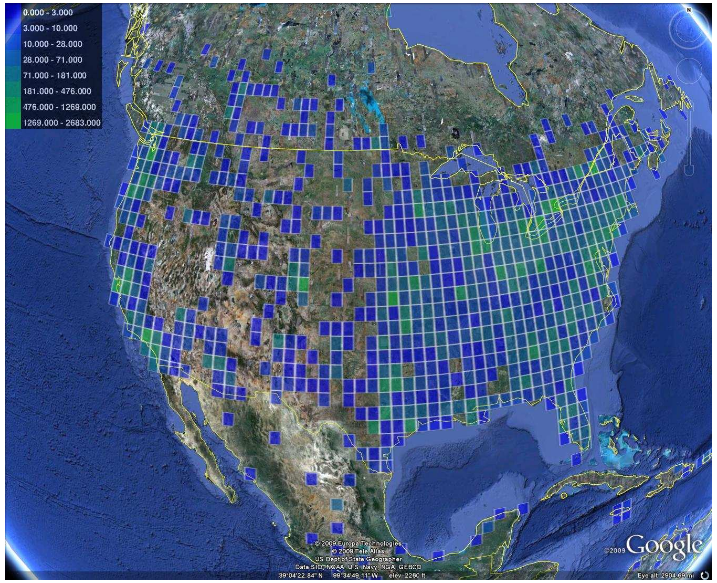
MemtestG80 – Validation

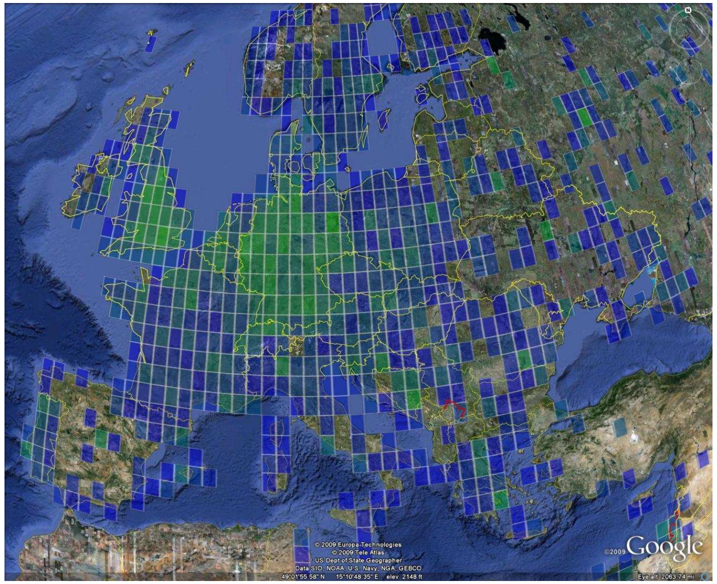


Positive control displays pattern sensitivity of memory tests

Methodology – Folding@home

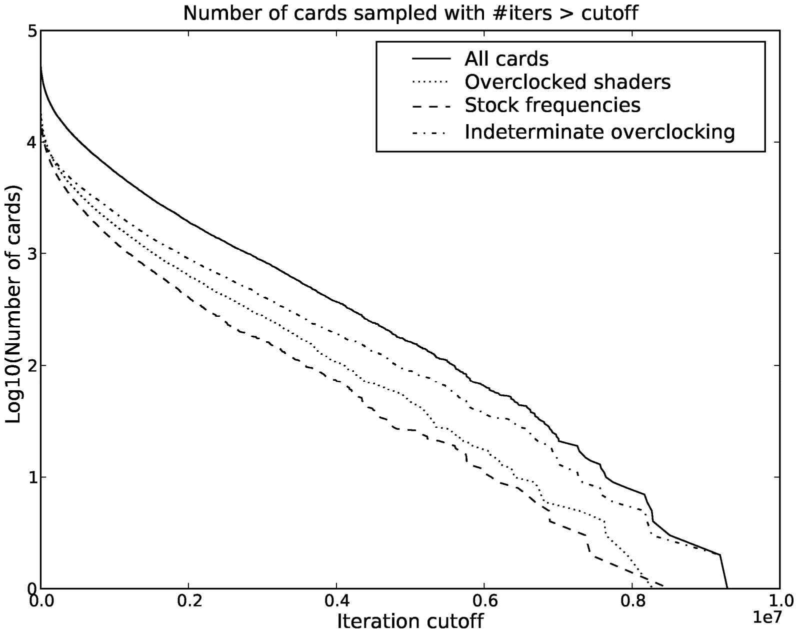
- Expect a low error rate and environment sensitivity, so must sample *many* cards in diverse environments
- Ran for ~7 months over 50,000+ NVIDIA GPUs on Folding@home (>840 TB-hr of testing)
- >97% of data tested 64 MiB RAM, k=512 logic LCG

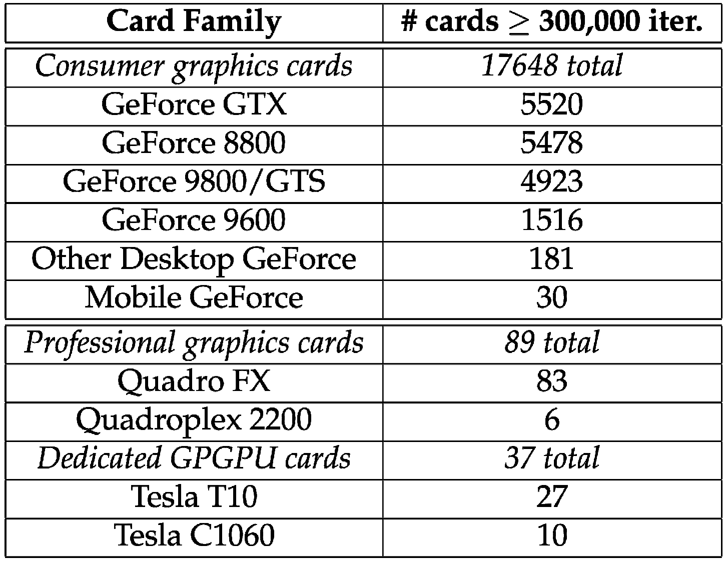



© 2009 Europa Technologies
© 2009 Tele Atlas
US Dept of State Geographer
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
©2009 Google

Methodology – Folding@home

- We achieve good sampling over the NVIDIA consumer product line, and a few pro cards as well.
- Sampled similar numbers of stock and (shader) overclocked boards





Card Family	# cards ≥ 300,000 iter.
Consumer graphics cards	*17648 total*
GeForce GTX	5520
GeForce 8800	5478
GeForce 9800/GTS	4923
GeForce 9600	1516
Other Desktop GeForce	181
Mobile GeForce	30
Professional graphics cards	*89 total*
Quadro FX	83
Quadroplex 2200	6
Dedicated GPGPU cards	*37 total*
Tesla T10	27
Tesla C1060	10

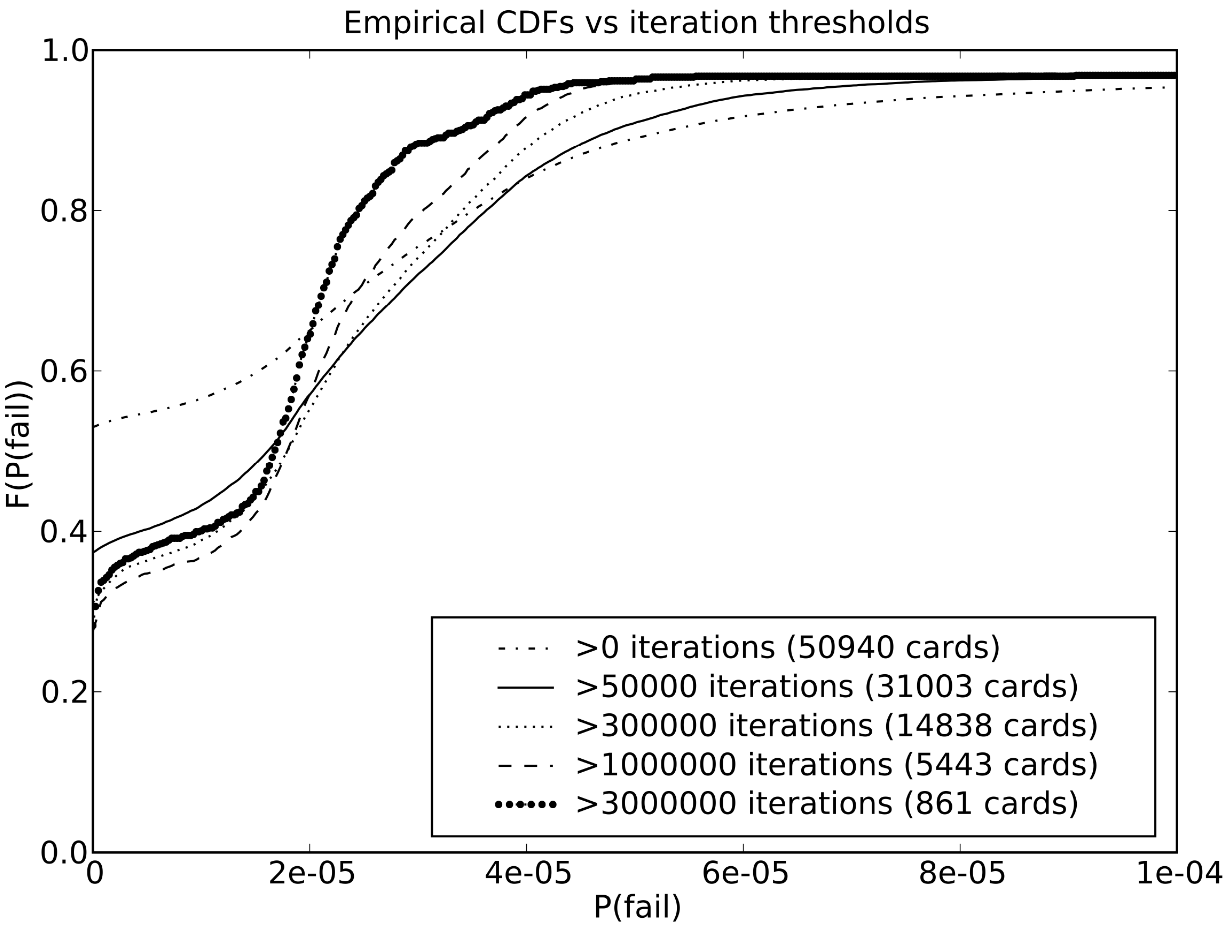
Results

- We call a failure if any test in a MemtestG80 iteration failed (ignore exact WER)

- Model: each card has its own probability of error (test failure) = P_f. Cards are drawn iid from an underlying distribution $\mathbf{P}(P_f)$

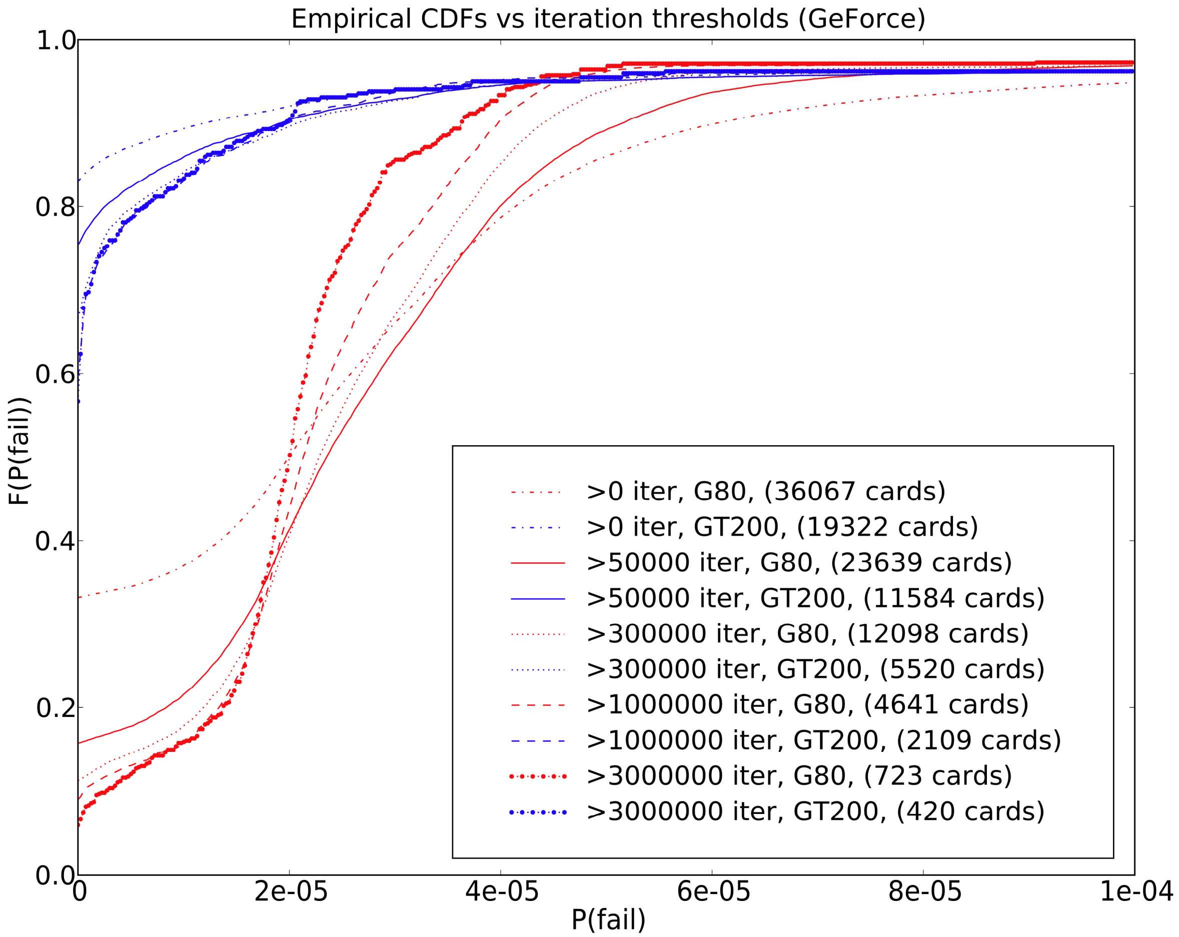
- What is the distribution of failure probabilities?

Results



Population of failing cards has a mode around $P_f = 2x10^{-5}$ = ~4 failures/week

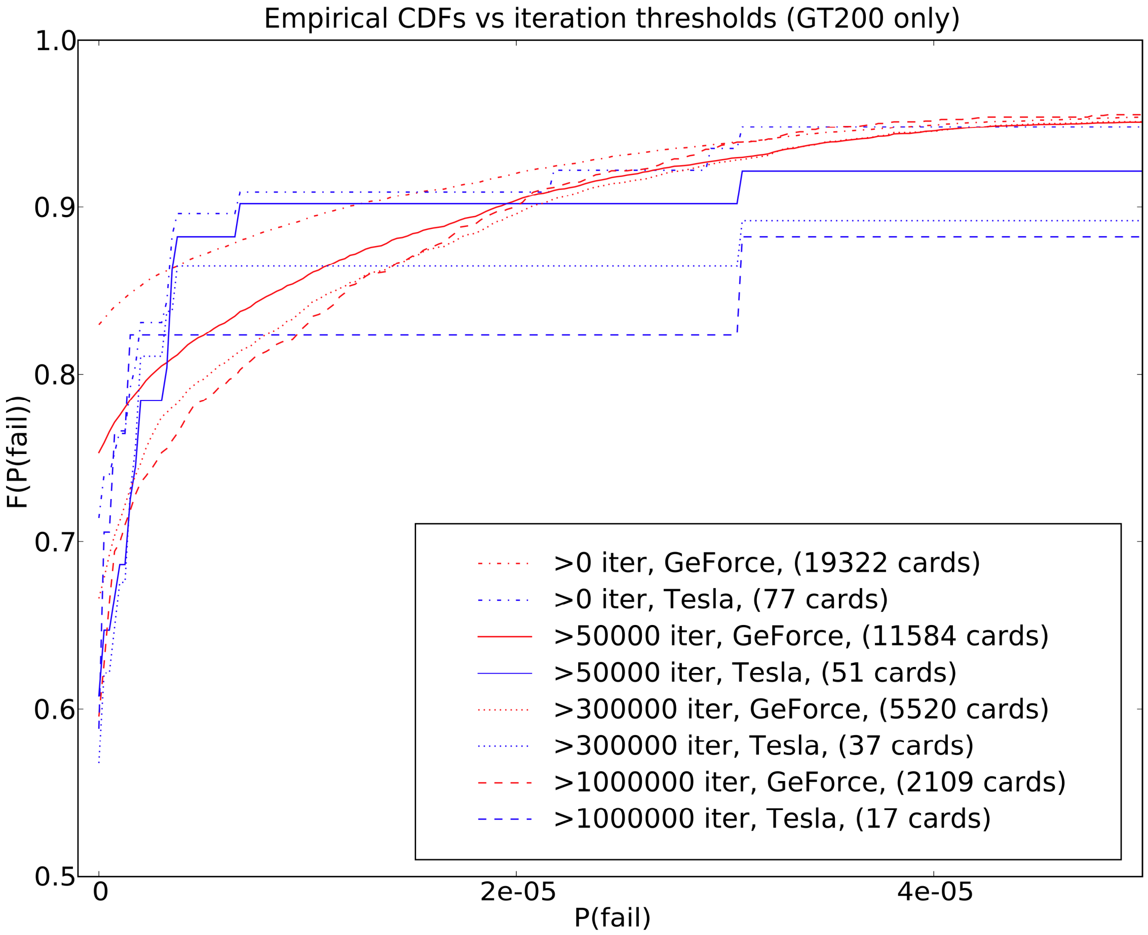
Analysis – Breakdown by Architecture



GT200 has typical P_f = 2.2x10^{-6} (one-tenth of G80!)

Both archs. show monotonic decline in zero-error populations.

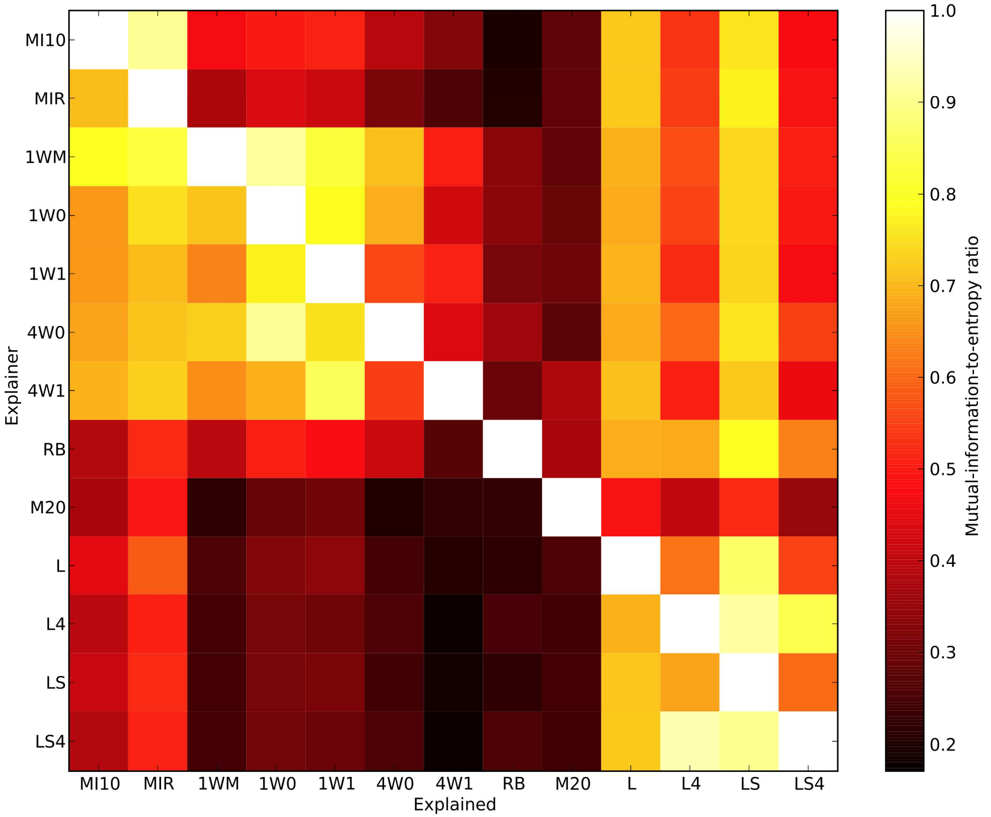
Analysis – GeForce vs Tesla



Tesla traces are rougher from poorer sampling, but appear to represent same error distribution as GeForce data.

Analysis – Test Mutual Information

- Consider mutual information between tests as a nonlinear covariance measure.
- Mod-20 test is unique
- Random blocks test is a good logic workout
- Logic tests measure a failure mode distinct from memory tests



What about Fermi?

- NVIDIA’s new Fermi (GF100) architecture adds SECDED ECC (disabled in consumer GeForce line), GDDR5 memory bus ECC, and L1/L2 caches

- **Does Fermi redesign affect architectural vulnerability (error rate or error type)?**
  - G80/GT200 typically failed on Mod-20 test first

- FAH test does not run (yet) on Fermi; used standalone MemtestG80 w/reporting capabilities
  - **In-house: 1 GeForce GTX 480, 1 Tesla C2050**
  - **Public: 44 GeForce GTX 470, 43 GeForce GTX 480**

Results – Fermi

- **Tesla**: no app-level errors seen, at least one double-bit error reported by ECC

- **GeForce**: most cards exhibited memory errors – observed in-house $\mathbf{P_f = 1.6 \times 10^{-5}}$
  - Non-overclocked cards vulnerable to 8-bit walking zeros
  - RAM-overclocked first failed 8- or 32-bit walking zeros
  - Core/shader-overclocked failed random blocks

- Very different vulnerabilities than G80/GT200 – but problems still exist!

What about AMD…and the CPU?

- RV700 and Evergreen both have GDDR5 (GDDR3 on low-end models) and L1/L2 hierarchy

- No current OpenCL cores on FAH; used volunteer submissions from standalone MemtestCL
  - **In-house:**
    - **Radeon 4870 (RV770); Radeon 5870 (Cypress)**
  - **Public:**
    - **RV700: 2 RV710, 15 RV730, 88 RV770**
    - **Evergreen: 1 Cedar, 6 Redwood, 50 Juniper, 103 Cypress**
    - **CPUs: 16 Core i7, 11 Core 2, 17 Phenom/Athlon II**

Results – AMD+CPU

- **CPU**: no errors seen
- **RV770**: typically fail random blocks/mod-20 – around **P_f = 7 x 10^{-4}**
- **Cypress**: almost all cards eventually fail random blocks – around **P_f = 4 x 10^{-4}**
- **BUT**: error patterns (#bits failed/iteration) are suspicious – currently working with AMD to see if it’s a software (MemtestCL or CL runtime) problem.

Acknowledgments

- Pande lab, Stanford University
- Simbios (NIH Roadmap GM072970)
- NVIDIA
- AMD
- **Folding@home donors**

Summary

- Wrote MemtestG80 to test for GPU memory errors.
- Verified proper operation of MemtestG80 with negative and positive control tests.
- Ran MemtestG80 on over 50,000 GPUs, 840+ TB-hr

- 2/3 of tested GPUs exhibit pattern-sensitive soft errors
- Architecture makes a difference: GT200 is much more reliable than G80; GF100 introduces a new set of vulnerabilities; AMD is yet another story.

- GT200 Tesla cards on FAH performed similarly to GeForces (but GF100 ECC seems to make a difference on Tesla C20xx)

Conclusions

- Sufficiently high hard error rate (2%) that explicit testing is warranted.
- Some form of ECC appears to be crucial for reliable GPGPU computation.

https://simtk.org/home/memtest

ihaque@cs.stanford.edu