ECE 574 – Cluster Computing
Lecture 21

Vince Weaver

https://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

28 October 2026

Announcements

- Midterm: graded, average=83%
- Still working on homeworks
- Don't put off HW#7, tricky

Metrics

- MTBF – mean time before failure
- FIT (failure in Time)
  One failure in billion hours. 1000 years MTBF is 114FIT. Zero error rate is 0FIT but infinite MTBF
  Designers just FIT because additive.
- Nines. Five nines 99.999% uptime (5.25 minutes of downtime a year)
  Four nines, 52 minutes. Six nines 31 seconds.
  Note: avoid nine-fives
- Bathtub curve

Architectural Vulnerability factor

- Some bit flips matter more
  - caches, program counter, registers
- Some matter less
  - branch predictor state
  - Parts of memory that have dead code, unused values
  - Low mantissa bits in floating bit numbers
  - Colors in graphics shown for only a frame

Hardware Errors

Soft errors/Radiation

- Chips so small, that radiation can flip bits. Thermal and Power supply noise too.
- Soft errors – excess charge from radiation. Usually not permanent.
- Sometime called SEU (single event upset)

Radiation

- Neutrons: from cosmic rays, can cause “silicon recoil” Can cause Boron (doped silicon) to fission into Li and alpha.
- Alpha particles: from radioactive decay
- Cosmic rays – higher up you are, more faults Denver 3-5x neutron flux than sea level. Airplanes. Satellites and space probes are radiation-hardened due to this.
- Smaller devices/transistors, more likely can flip bit.

Shielding

- Neutrons: 3 feet concrete reduce flux by 50%
- alpha: sheet of paper can block, but problem comes from radioactivity in chips themselves

Case Studies

- "May and Woods Incident" first widely reported problem. Intel 2107 16k DRAM chips, problem traced to ceramics packaging downstream of Uranium mine.
- "Hera Problem" IBM having problem. ^{210}Po contamination from bottle cleaning equipment.
- "Sun e-cache" Ultra-SPARC-II did not have ECC on cache for performance reasons. High failure rate.

Things you can do for reliable Hardware

Hardware Replication / Redundancy

- Lock step – Have multiple machines / threads running same code in lock-step Check to see if results match. If not match, problem. If replicated a lot, vote, and say most correct is right result.
- RAID – (redundant array of inexpensive disks)
- Memory checksums – caches, busses
- Power conditioning, surge protection, backup generators, UPS
- Hot-swappable redundant hardware

Lower Level (Inside your Computer)

- Replicate units (ALU, etc)
- Replicate threads or important data wires
- CRCs and parity checks on all busses, caches, and memories

Device Level Fixes

- Using doping less susceptible to Boron fission
- Use low-radiation solder
- Silicon-on-Insulator
- Double-gate devices (two gates per transistor)
- Larger transistor sizes
- Circuits that handle glitches better.

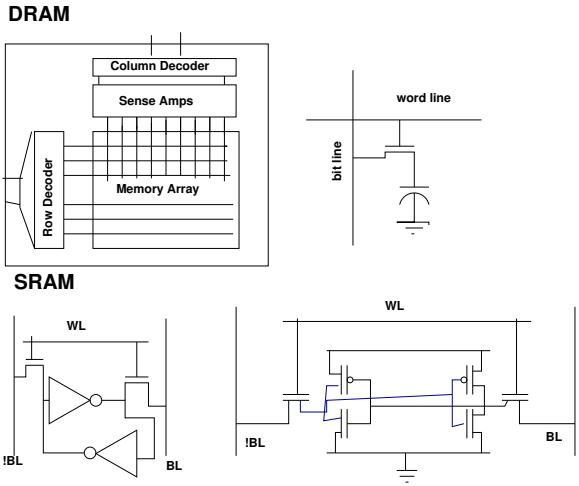
Extreme Testing

- Single event upset characterization of the Pentium MMX and Pentium II microprocessors using proton irradiation", IEEE Transactions on Nuclear Science, 1999.
- Pentium II, took off-shelf chip and irradiated it with protons. Only CPU, rest shielded with lead. Irradiate from bottom to avoid heatsink
- Various errors, freeze to blue screen. no power glitches or "latchup" 85% hangs, 14% cache errors no ALU or FPU errors detected.

Memory Failures

- Major causes of reliability issues

Memory Overview



Memory Reliability Paper

- Memory Errors in Modern Systems
  ASPLOS 2015

Memory Reliability Paper 2

- Battling Borked Bits
  IEEE Spectrum December 2015
- HPC Clusters crashes, 60% hardware cause, most common DRAM
- 12-45% one DRAM error/year (ECC), .2-4% uncorrectable
- Small amount of machines cause the most trouble

Intentional Memory Failures?

- Rowhammer
- DRAM is just holding RAM contents in capacitors, which leak away and need to be constantly refreshed
- Need to refresh every 32 to 64ms
- If you access a memory location a lot, it can also make nearby locations drain faster and make them have bit flips
- Interesting project?

Memory Fixes

- ECC code
- spread bits out. Right now can flip adjacent bits, flip too many can't correct.
- Memory scrubbing: going through and periodically reading all mem to find bit flips.