ECE 574 – Cluster Computing
Lecture 14

Vince Weaver

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

`vincent.weaver@maine.edu`

7 March 2023

Announcements

- HW#6 posted. Tentatively due Friday.
- Discussion on the pain of MPI coding

HW#4 Finally Graded

- Many issues are low-level C issues more than pthread issues
- Trouble splitting up workload
- `calloc()` and pointers

HW#6 More Notes

- Lot of low-level C issues
- MPI gives bad error messages
- Gather is the tricky part
- Be sure to gather into pixels (remember, getting an array of char) not sobel_x (what happens if you gather an array of char on top of a struct? Especially this struct that isn't necessarily followed by the data

HW#6 Cluster fairness

- If your job gets stuck, be nice and kill it (scancel)
- The node isn't currently enforcing times. I could set it up to do so but worried I'd break things
- sbatch scripts I give you have 10 minute timeout, you can lower that if you want to be safer

Reliability in HPC

Good reference is a class I took a long time ago, CS717 at Cornell:

`http://greg.bronevetsky.com/CS717FA2004/Lectures.html`

Sources of Failure

- Software Failure
  - Buggy Code
  - System misconfiguration
- Hardware Failure
  - Loose wires
  - Tin whiskers (lead-free solder)
  - Lightning strike
  - Radiation
  - Moving parts wear out

- Malicious Failure
    - Hacker attack
- Environment issues
    - Fire in datacenter
    - Loss of cooling during heat wave

Types of fault

- Permanent Faults – same input will always result in same failure
- Transient Faults – go away, temporary, harder to figure out

What do we do on faults?

- Detect and recover?
- Just fail?
- Can we still get correct results?

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.
- Bathtub curve

Architectural Vulnerability factor

- Some bit flips matter less
- (branch predictor) others more (caches) some even more (PC)
- Parts of memory that have dead code, unused values

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

Lower-Level Problems

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. Denver more than here. Airplanes. Satellites and space probes are radiation-hardened due to this.
- Smaller devices, 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.

Hardware 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.
- 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.

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 proton. 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

- Memory Errors in Modern Systems
  ASPLOS 2015

- Battling Borked Bits
  IEEE Spectrum December 2015

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

Architectural Vulnerability factor

- Some bit flips matter less
- (branch predictor) others more (caches) some even more (PC)
- Parts of memory that have dead code, unused values