Memory

Fuel tank for the processor

Growth in Size

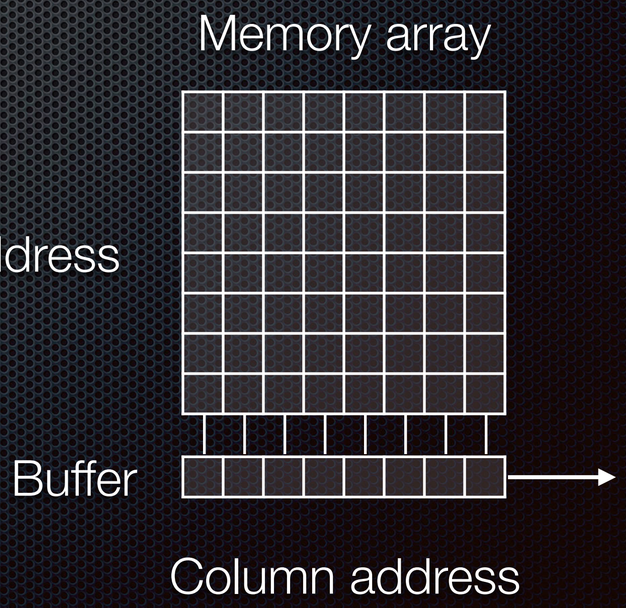
- Most applications are not high performance
  - Even so, growth in features requires more RAM
- High performance applications depend on memory
  - Old mantra: A megabyte per mega-FLOP
  - Often limited by memory capacity — more cores may just mean more cache
- Manufacturers emphasize memory size over speed (latency)
- 64K to 8G = factor 128K times bigger in 32 years

Effects of Size Scaling

- Fewer memory chips for a small system
- More compact large systems
- Reduced cost
- Reduced opportunity for innovation in memory architecture
  - Architecture is almost entirely within the chip
  - Only manufacturers can implement new ideas

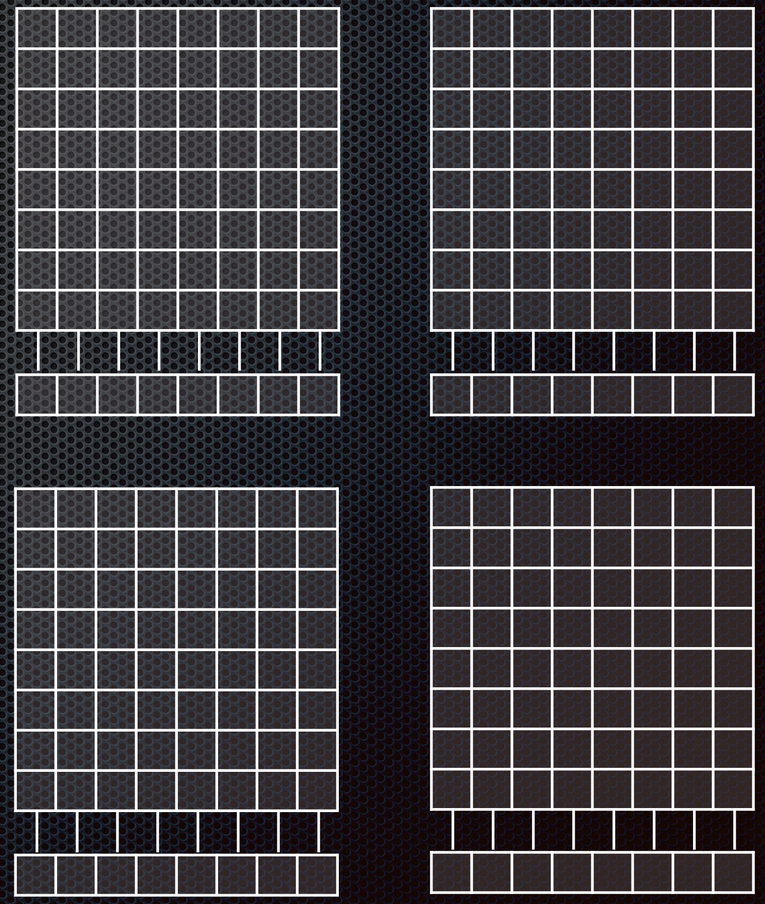
Organization

- Bits arranged in a rectangular array: rows and columns
- Address is divided into row and column portions
- Row address selects a row to read or write
- Row is destructively read out to buffer registers
- Column address selects a register within the buffer
- Write overwrites the register, read copies it out
- Buffer is rewritten to row



Multiple Banks

- A large RAM will be divided into multiple banks
  - Partly for row access speed, partly for yield
  - Each bank has its own buffers, read/write logic
  - Address is further divided into bank portion
- Access is really to buffers
- Opportunity for increased performance



Buffer Access

- Once a row is in a buffer, sequential accesses can be to the buffer without extra row access cycles
- Adding a layer of buffering allows a new row access to start while a prior row is still being read
- Multiple bank buffers enable different pages to be open for fast access at the same time
- Internally, much higher bandwidth is available
  - But manufacturers won’t allow us to use it

DRAM Performance

- Access time has decreased about 6X in 32 years (180ns to 30ns cycle) vs 36X increase in clock rate
- Bandwidth has improved more
  - Double data rate
  - Open pages, fast page mode, synchronous transfer
- Still slower than rate of CPU performance improvement

DDR

- Transfer data on rising + falling clock (double data rate)
- DDR 2.5 volts, 200MHz
- DDR2 1.8 volts, 400MHz
- DDR3 1.5 volts, 800MHz
- DDR4 1 to 1.2 volts, 1600MHz
- GDDR specialized for GPU, 32-bit bus, faster clock

Flash Memory

- Will be covered in a later class in more detail
- Slower and cheaper than DRAM, faster and more expensive than disk
- Nonvolatile (though not permanent), thus lower power
- Limited life (100K cycles)
- Erase blocks to 1s, write pages with 0s
  - Slower to erase than read or write
- MLC stores 2 bits per site, but is slower, has shorter life

Bill Wulf and Sally McKee

Hitting the Memory Wall, SigArch News, 1995

Relative Rates of Technology

- Processor and memory speed growing exponentially
- Different exponents lead to exponential divergence
- Implies future problems in improving system performance

Simple Model

- $t_{avg}=p \times t_c + (1 - p) \times t_m$
- Cache time (t_c) is 1-cycle, all instructions 1 cycle
- No conflict or capacity misses, only compulsory (initial loads), so p (probability of hit) is high and (1 - p) is small
- If t_{avg} grows to equal time between memory accesses, processing time is dominated by memory access time

When will we hit the wall?

- Assume 7% per year increase in DRAM speed
- Processor performance grows 80% per year
- Less than 1% compulsory miss rate
- Should hit in less than a decade (before 2005)

Graphically

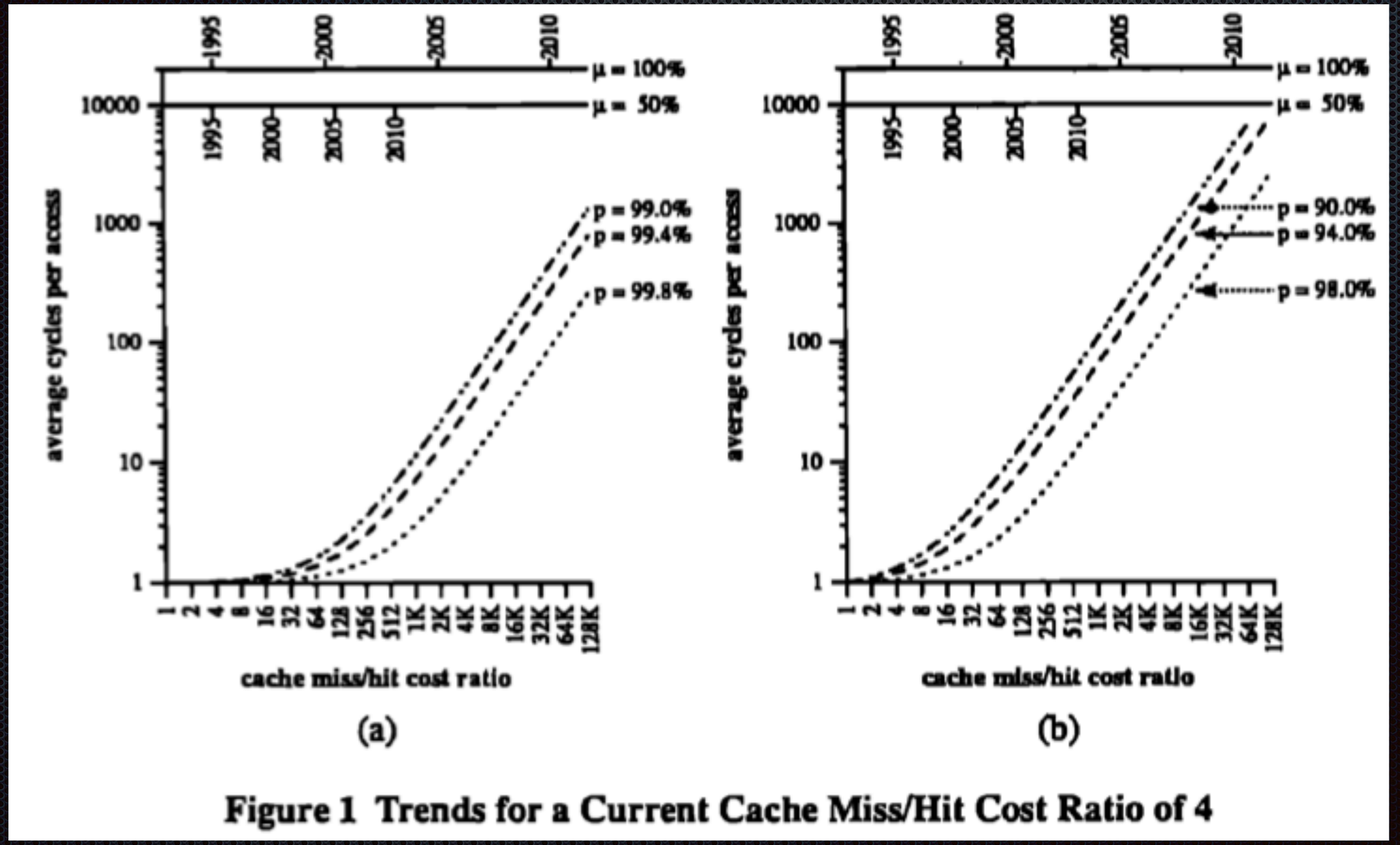
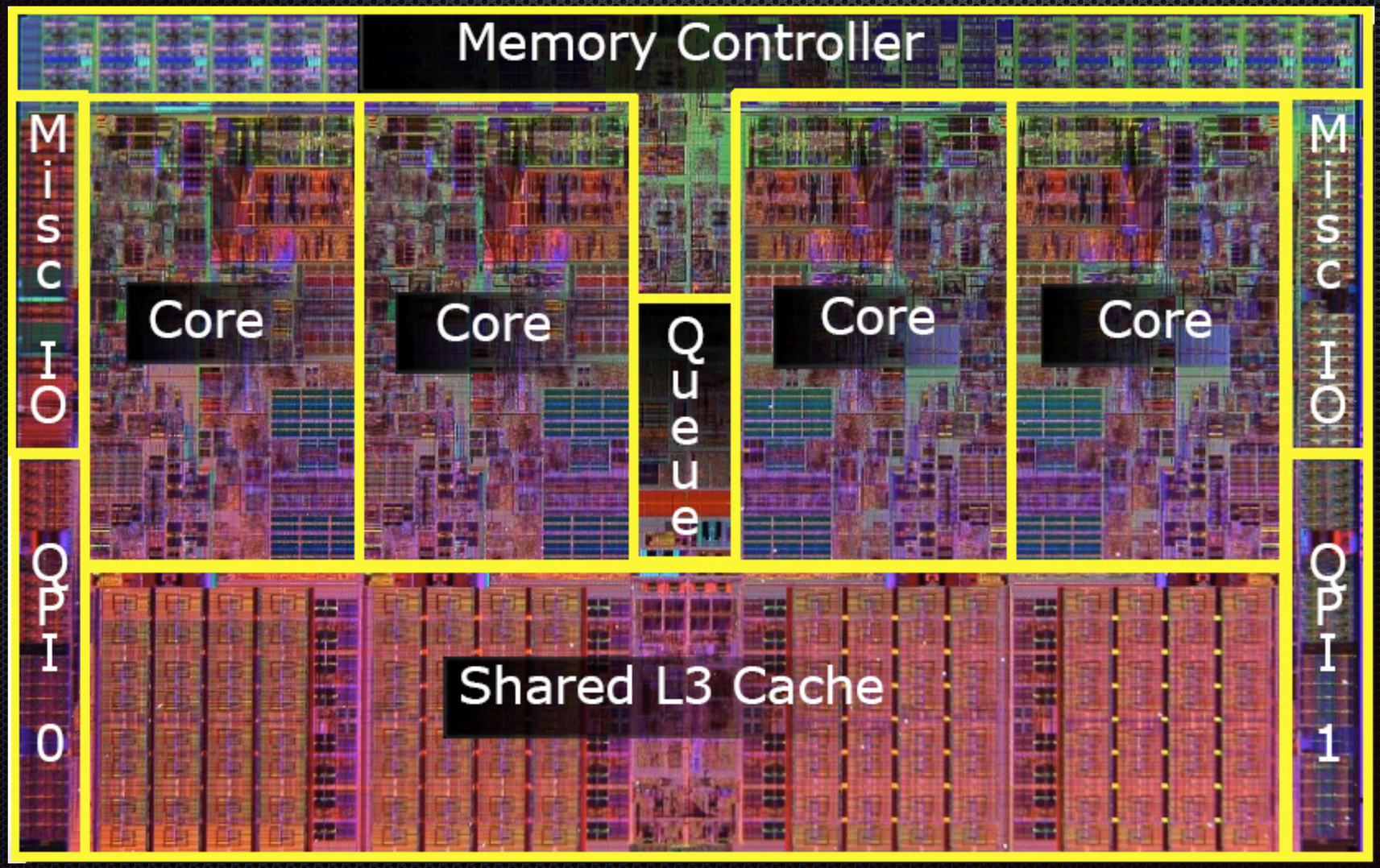


Figure 1 Trends for a Current Cache Miss/Hit Cost Ratio of 4

What happened?

- Memory access is now about 120 cycles (speed growth rose to 10%/year)
- Parallel memory channels satisfy multiple requests
- Cycle time plateaus (falls off to 60%/year then 22%), but instruction level parallelism grows
- More cores with shared memory (some compulsory misses, e.g., common code, resolve multiple accesses)
- Cache grows in size, goes on chip, and adds prefetch

Caches are Working Set Size



Workloads

- Common desktop workloads don’t suffer many compulsory misses once code is cached, users tolerate pauses
- Many server workloads also have high cache residency of code and data
- High performance applications generate much of their data in place
- ILP benefits from bandwidth because of spatial locality
- What workloads hit the wall?

Discussion

Vilas Sridharan, et. al.

Memory Errors in Modern Systems, The Good, The Bad, and The Ugly, ISCA 2015

Supercomputer Testbeds

- Hopper, 6000 nodes, L. Berkeley Lab
- Cielo, 8500 nodes, Los Alamos Lab
- AMD Opterons with DDR3
- 45B DRAM device hours of data

The Good

- Adding parity checks to command and address portions of the DDR interface significantly improves reliability and error detection
- Except for cosmic ray faults, SRAM is reliable
- Most SRAM errors are single-bit, which can be easily fixed with known engineering techniques

The Bad

- DRAM is more susceptible to cosmic ray faults than suspected, especially at higher altitudes
- Future DRAM technology needs to do more to address resiliency issues

The Ugly

- Existing methodologies (counting errors instead of faults) are inaccurate
- Common ECC approaches can result in undetected errors at a rate of 20 FIT per DRAM device
  - FIT is failure in time per billion device hours

DRAM Vendor and Altitude

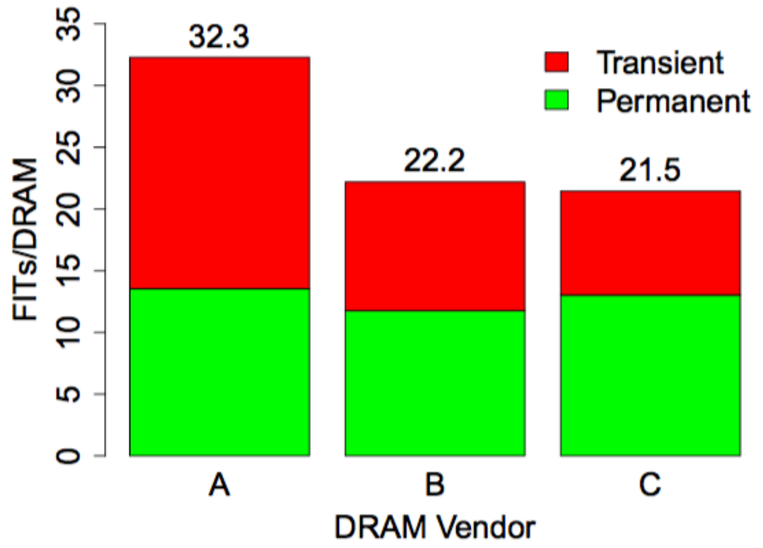


Figure 4. Fault rate per DRAM vendor. Hopper sees substantial variation in fault rates by DRAM vendor.

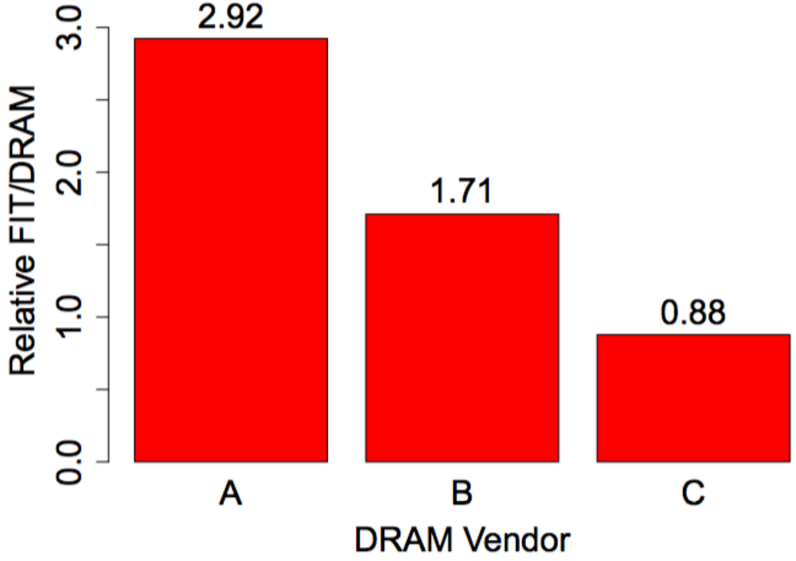


Figure 5. Cielo DRAM transient fault rates relative to Hopper (Hopper == 1.0). Vendor A shows a significantly higher transient fault rate in Cielo, likely attributable to altitude.

Vendor A has the smallest number of installed modules

Verification of Accelerated Testing Results

- When you have access to neutron and alpha particle beams, you can shoot them at processor chips to estimate fault rates
- Predictions were a little low compared with fielded system data

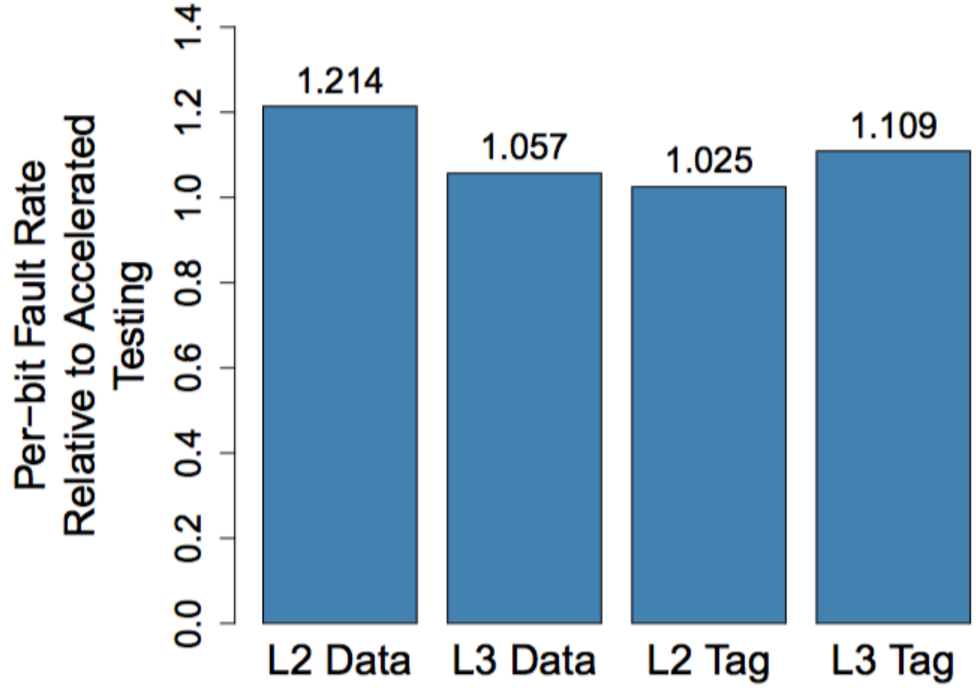


Figure 8. Rate of SRAM faults in Hopper compared to accelerated testing.

Faults are a stronger measure than errors

- Prior work has recorded errors logged in a hardware register, but the OS infrequently polls the register
- Multiple non-hardware mechanisms affect the reported error rate
- Faults indicate raw device reliability

SECDED vs Chipkill

- Chipkill is an IBM trademark, other vendors have different terms
- Similar idea to RAID: Scatter bits across chips using an advanced error correcting code
- Enables recovery even when an entire chip fails
- Fault patterns encountered indicate a high rate of errors that traditional SECDED codes can’t even detect

On-Chip SRAM needs more than parity protection

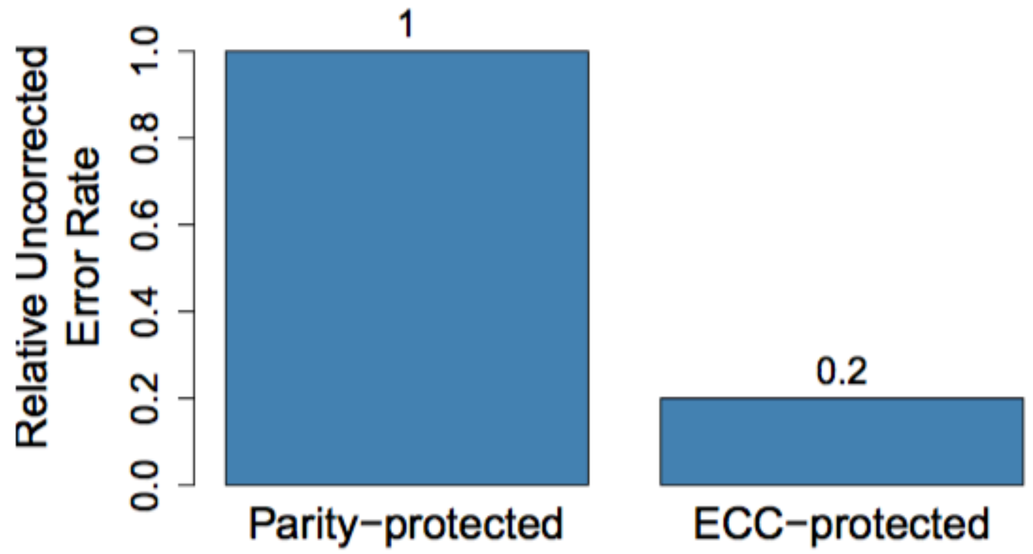


Figure 12. Rate of SRAM uncorrected errors in Cielo from parity- and ECC-protected structures.

Discussion