Cycles, cells, and platters: An empirical analysis of hardware failures on a million commodity PCs

Ed Nightingale, John Douceur and Vince Orgovan

Microsoft Corporation
Grid/Scientific computing professors
  - DRAM errors are common
  - Notorious non-ECC cluster – 6,000 machines – best 2 out of 3

OS/Architecture Professors
  - You’re crazy!
  - Huge address space + Alpha particles = no failures

Vince Orgovan
  - OCA/ATLAS frequently observes bit flips in the wild
What’s the bottom line?

• **First failure rates are non-trivial.**
  • The probability of crashing once from a CPU, one-bit DRAM, or disk failure is as high as 1 in 190 over an 8 month observation period.

• **Recurrent failures are common.**

• **Recurrent failures happen quickly.**
  • As many as 97% of recurring failures occur within 10 days of the first failure on a machine.

• **CPU speed matters.**
  • Overclocking and underclocking have a large impact of reliability

• **DRAM faults have spatial locality.**
Outline

• Methodology – diagnosis & data sets
• Analyzing the probability of failure
• Effect of machine class
• Effect of machine characteristics
• Temporal Analysis
• Failure vs. fault
  • A failure is an incident, while a fault is a condition (defect)

• A failure may be recurring or non-recurring.

• Faults can be out into one of three categories
  • Permanent faults
    • Durable defects (burned out chip)
  • Intermittent faults
    • Fault that persists, causing 0 or more failures (atomic defect on chip)
  • Transient Faults
    • Instantaneous defect causing a single failure (Alpha particle)
Failure types

- **CPU**
  - Machine-check exception

- **Disk subsystem**
  - Failure during critical OS read

- **DRAM corruption**
  - 1-bit corruption in a kernel-code page
CPU subsystem failure

• CPU throws a machine-check exception (MCE)
  • Internal invariant within CPU is broken and unrecoverable

• Examples:
  • Parity error in ROM
  • parity error in L1 cache
  • error communicating with memory controller
  • bus error, unrecoverable ECC error etc., etc.

• Causes:
  • Manufacturing defect, cracked/stressed motherboard
  • Under-powered power-supply/over-clocking/heat
  • Dust/dirt/grease whatever
Disk subsystem failures

- Failure to read data within critical kernel code
  - Example: Reading from the page file

- Wait! Dump-driver must write to disk
  - Fault eventually disappears
  - Vibration, buggy firmware, disk heisenbug

- Causes:
  - Faulty bus controller, faulty disk controller, buggy firmware
  - Faulty/loose cable, heat, vibrations
  - Faults on platter or disk mechanisms (arm/head/spindle etc)
1-bit DRAM failures

- Mini-dump captures 256 bytes around IP
- ‘diff’ against code kept at Microsoft.
  - If 1 bit differs, mark it as 1-bit corruption
- Only kernel-code pages are compared
  - 30 MB of the address space in Vista
- MMU protects against stray software writes
• OCA (ATLAS)
  • Process mini-dumps submitted by customers
  • No information about absence of failures.
  • Have only some subset of failures for a machine

• RAC
  • Machines anonymously report to Microsoft every 2-4 days.
  • All events reported (absence of failures captured).
  • No minidumps, but result of ATLAS analysis is recorded.
  • Captured a pool of about 1 million machines over 8 months
Outline

- Methodology – diagnosis & data set
- Analyzing the probability of failure
- Effect of machine class
- Effect of machine characteristics
- Temporal Analysis
- A fault-tolerant single-machine OS
### Conditional probability of failure

<table>
<thead>
<tr>
<th>Failure</th>
<th>Min TACT</th>
<th>Pr[1\textsuperscript{st} failure]</th>
<th>Pr[2\textsuperscript{nd} fail</th>
<th>1 fail]</th>
<th>Pr[3\textsuperscript{rd}</th>
<th>2 fails]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (MCE)</td>
<td>5 days</td>
<td>1 in 330</td>
<td>1 in 3.3</td>
<td>1 in 1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPU (MCE)</td>
<td>30 days</td>
<td>1 in 190</td>
<td>1 in 2.9</td>
<td>1 in 1.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory DRAM 1-bit</td>
<td>5 days</td>
<td>1 in 2700</td>
<td>1 in 9.0</td>
<td>1 in 2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory DRAM 1-bit</td>
<td>30 days</td>
<td>1 in 1700</td>
<td>1 in 12</td>
<td>1 in 2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk subsystem</td>
<td>5 days</td>
<td>1 in 470</td>
<td>1 in 3.4</td>
<td>1 in 1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk subsystem</td>
<td>30 days</td>
<td>1 in 270</td>
<td>1 in 3.5</td>
<td>1 in 1.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- When a machine crashes again, it crashes within:
  - CPU subsystem (MCE) 10 days: 84% 30 days: 97%
  - 1-bit DRAM failures 10 days: 97% and 30 days: 100%
  - Disk subsystem 10 days: 86% and 30 days: 99%
• Does spatial locality exist for 1-bit errors?

• Analyzed ~300k 1-bit errors out of ATLAS
  • Of machines that crashed more than once in !NT, 79% crashed at same physical address and same bit flipped.

• Alpha particle unlikely to strike same transistor.
  • Seeing hardware defects in the wild.
  • ECC not coming any time soon.
  • Unreliable hardware is a reality software must address.
• Effect of machine class
• Effect of machine characteristics
• Temporal Analysis
• CPU passes tests and ‘rated’ at a certain speed
  • CPU actually runs within some delta of rated speed: 1995 MHz
<table>
<thead>
<tr>
<th>Failure type</th>
<th>No OC</th>
<th>OC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM 1-bit flip</td>
<td>1 in 2800</td>
<td>1 in 560</td>
</tr>
<tr>
<td>Disk subsystem</td>
<td>1 in 480</td>
<td>1 in 430</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Failure type</th>
<th>CPU Vendor A</th>
<th>CPU Vendor B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr[1\text{st}]</td>
<td>No OC: 1 in 400, OC: 1 in 21</td>
<td>No OC: 1 in 390, OC: 1 in 86</td>
</tr>
<tr>
<td>Pr[2\text{nd}</td>
<td>No OC: 1 in 3.9, OC: 1 in 2.4</td>
<td>No OC: 1 in 2.9, OC: 1 in 3.5</td>
</tr>
<tr>
<td>Pr[3\text{rd}</td>
<td>No OC: 1 in 1.9, OC: 1 in 2.1</td>
<td>No OC: 1 in 1.5, OC: 1 in 1.3</td>
</tr>
</tbody>
</table>

Overclocking greatly increases probability of failure
Effect of underclocking

Underclocked machines up to 80% less likely to crash
- Machines see benefit when underclocked by as little as 1%

<table>
<thead>
<tr>
<th>Failure type</th>
<th>Underclocked</th>
<th>Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (MCE)</td>
<td>1 in 460</td>
<td>1 in 330</td>
</tr>
<tr>
<td>DRAM 1-bit</td>
<td>1 in 3600</td>
<td>1 in 2000</td>
</tr>
<tr>
<td>Disk subsystem</td>
<td>1 in 560</td>
<td>1 in 380</td>
</tr>
</tbody>
</table>
White box vs. Brand name

<table>
<thead>
<tr>
<th>Failure type</th>
<th>Brand name</th>
<th>White box</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (MCE)</td>
<td>1 in 470</td>
<td>1 in 230</td>
</tr>
<tr>
<td>DRAM 1-bit</td>
<td>1 in 3400</td>
<td>1 in 1300</td>
</tr>
<tr>
<td>Disk subsystem</td>
<td>1 in 430</td>
<td>1 in 390</td>
</tr>
</tbody>
</table>

- Brand name if OEM in top 20 by sales volume world wide
- Brand name more reliable across board
  - Least pronounced for disk subsystem faults
Surprise! Laptops more reliable than desktops
- Laptop components designed to be rugged, desktop are not.

<table>
<thead>
<tr>
<th>Failure type</th>
<th>Desktops (1 in x)</th>
<th>Laptops (1 in x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (MCE)</td>
<td>1 in 470</td>
<td>1 in 510</td>
</tr>
<tr>
<td>DRAM 1-bit</td>
<td>1 in 3400</td>
<td>1 in 5100</td>
</tr>
<tr>
<td>Disk subsystem</td>
<td>1 in 430</td>
<td>1 in 590</td>
</tr>
</tbody>
</table>
Effect of machine characteristics

Temporal Analysis
Faster CPUs are more likely to fail...
- But TACT does not normalize for the speed of the CPU
Effect of machine speed (2)

- All CPUs equal probability of failure per CPU cycle.
  - For a given time period, faster CPUs will fail more often
  - Buy the slowest CPU for your given workload
  - Slow CPUs for improved reliability in addition to power savings
Effect of speed ratio (OC/UC)

- CPU failures dramatically impacted as overclocking ratio increases
- Overclocking does not have a large effect on disk failures
• Younger CPUs more likely to fail.
• Older disks more likely to fail.
• Methodology – diagnosis & data sets
• Analyzing the probability of failure
• Effect of machine class
• Effect of machine characteristics
• Temporal Analysis
• By count of failures, recurring > non-recurring

• By count of machines, recurring < non-recurring
  • CPU subsystem: 30% of failing machines show recurrence
  • Disk subsystem: 29% of failing machines show recurrence
  • DRAM (1-bit): 15% of failing machines show recurrence

• However, non-recurrence does not imply transience
  • Intermittent fault might manifest only one failure while under observation
  • Might be other failures before or after observation period
  • For many machines, our observation period is very short
Temporal analysis

• Analytical model of observed failure recurrence time
• Analytical model of observation period
• Calculate the probability that intermittent fault will manifest exactly one failure while under observation
  • CPU subsystem: 24%
  • Disk subsystem: 25%
  • DRAM (1-bit): 20%
• Estimate likelihood of intermittent fault
  • CPU subsystem: 39% of faulty machines are intermittent
  • Disk subsystem: 39% of faulty machines are intermittent
  • DRAM (1-bit): 19% of faulty machines are intermittent
• Methodology – diagnosis & data sets
• Analyzing the probability of failure
• Effect of machine class
• Effect of machine characteristics
• Temporal Analysis
First failure rates are non-trivial.
  • The probability of crashing once from a CPU, one-bit DRAM, or disk failure is as high as 1 in 190 over an 8 month observation period.

Recurrent failures are common.
  • Machines that have crashed once from a hardware failure are up to two orders of magnitude more likely to crash a second time. Intermittent faults make up a significant portion of observed faults. Between 20% and 40% of machines have faults that are intermittent rather than transient.

Recurrent failures happen quickly.
  • As many as 97% of recurring failures occur within 10 days of the first failure on a machine.

CPU speed matters.
  • Overclocking significantly degrades the reliability of a machine, and CPUs that are slightly underclocked are more reliable than those running at their rated speed. Even without overclocking, faster CPUs become faulty more rapidly than slower CPUs.

DRAM faults have spatial locality.
  • Our analysis demonstrates that almost 80% of machines that crashed more than once from a 1-bit DRAM failure had a recurrence at the same physical address as a prior failure.

Configuration matters.
  • Brand name desktop machines are more reliable than white box desktops, but brand name laptops are more reliable than brand name desktops. Machines with more DRAM will suffer more one-bit and CPU errors, but fewer disk failures.