Microsoft® Research
Faculty Summit 2010

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

Ed Nightingale, John Douceur and Vince Orgovan

Microsoft Corporation

A bit of background – fun while interviewing

- 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

Terminology

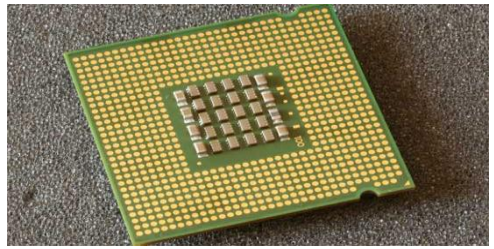
- Failure vs. fault
  - A failure is an incident, while a fault is a condition (defect)
- A failure may be recurring or non-recurring.
- Faults can be out into one of three categories
  - Permanent faults
    - Durable defects (burned out chip)
  - Intermittent faults
    - Fault that persists, causing 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

Data sets

- 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

Failure	Min TACT	Pr[1st failure]	Pr[2nd fail | 1 fail]	Pr[3rd | 2 fails]
CPU (MCE)	5 days	1 in 330	1 in 3.3	1 in 1.8
CPU (MCE)	30 days	1 in 190	1 in 2.9	1 in 1.7
Memory DRAM 1-bit	5 days	1 in 2700	1 in 9.0	1 in 2.2
Memory DRAM 1-bit	30 days	1 in 1700	1 in 12	1 in 2.0
Disk subsystem	5 days	1 in 470	1 in 3.4	1 in 1.9
Disk subsystem	30 days	1 in 270	1 in 3.5	1 in 1.7

- 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%

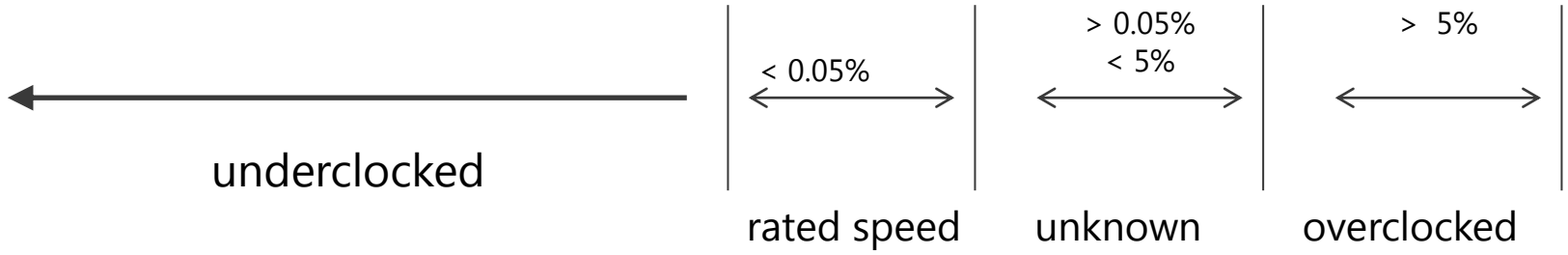
1-bit DRAM fault: Spatial locality analysis

- 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.

Outline

- Effect of machine class
- Effect of machine characteristics
- Temporal Analysis

Overclocking primer



- CPU passes tests and ‘rated’ at a certain speed
  - CPU actually runs within some delta of rated speed: 1995 MHz

Effect of overclocking

	CPU Vendor A		CPU Vendor B	
	No OC	OC	No OC	OC
Pr[1st]	1 in 400	1 in 21	1 in 390	1 in 86
Pr[2nd|1]	1 in 3.9	1 in 2.4	1 in 2.9	1 in 3.5
Pr[3rd | 2]	1 in 1.9	1 in 2.1	1 in 1.5	1 in 1.3

Failure type	No OC	OC
DRAM 1-bit flip	1 in 2800	1 in 560
Disk subsystem	1 in 480	1 in 430

Overclocking greatly increases probability of failure

Effect of underclocking

Failure type	Underclocked	Rated
CPU (MCE)	1 in 460	1 in 330
DRAM 1-bit	1 in 3600	1 in 2000
Disk subsystem	1 in 560	1 in 380

Underclocked machines up to 80% less likely to crash

- Machines see benefit when underclocked by as little as 1%

White box vs. Brand name

Failure type	Brand name	White box
CPU (MCE)	1 in 470	1 in 230
DRAM 1-bit	1 in 3400	1 in 1300
Disk subsystem	1 in 430	1 in 390

- Brand name if OEM in top 20 by sales volume world wide
- Brand name more reliable across board
  - Least pronounced for disk subsystem faults

Destkop vs. Laptop

Failure type	Desktops	Laptops
CPU (MCE)	1 in 470	1 in 510
DRAM 1-bit	1 in 3400	1 in 5100
Disk subsystem	1 in 430	1 in 590

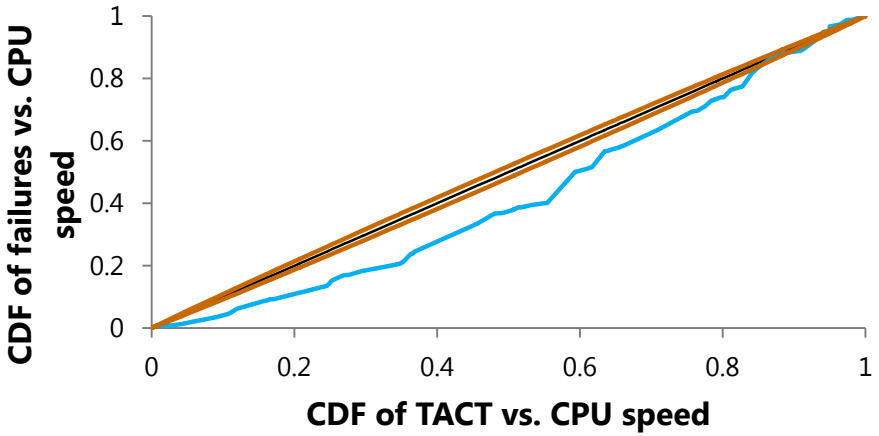
- Surprise! Laptops more reliable than desktops
  - Laptop components designed to be rugged, desktop are not.

Outline

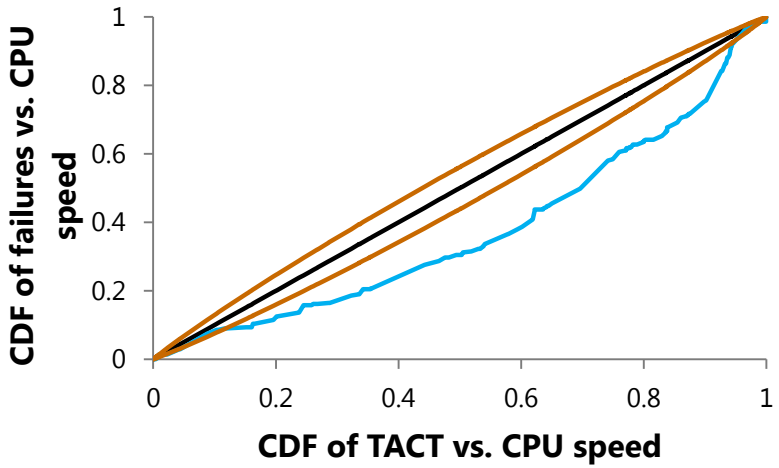
y

- Effect of machine characteristics
- Temporal Analysis

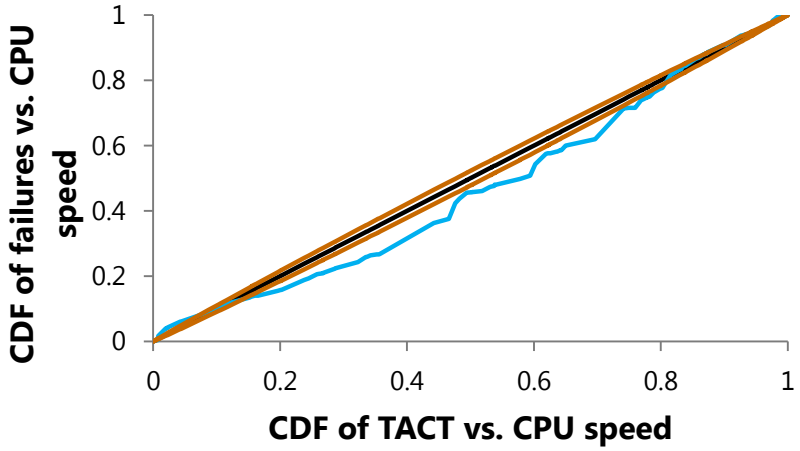
Effect of machine speed



CPU Failures vs. TACT



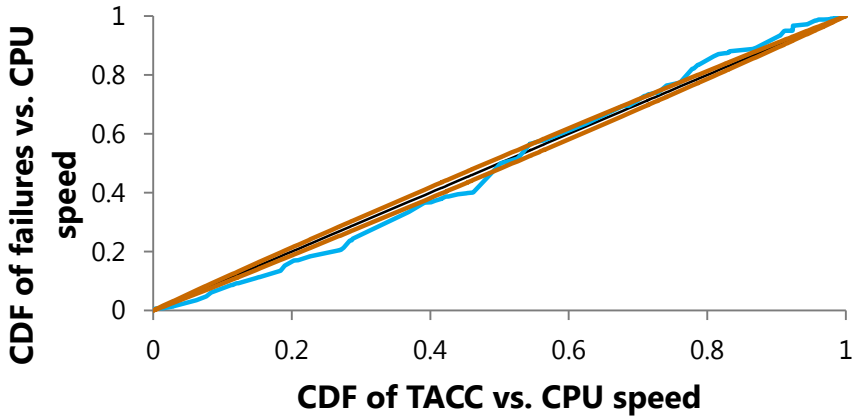
DRAM Failures vs. TACT



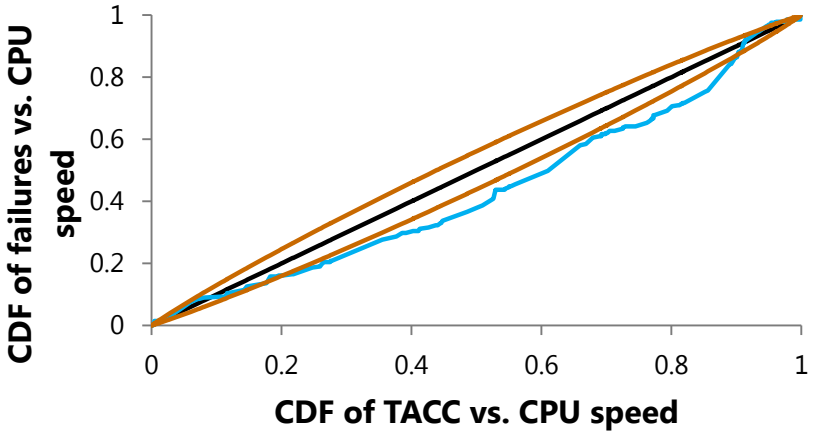
Disk Failures vs. TACT

- Faster CPUs are more likely to fail…
  - But TACT does not normalize for the speed of the CPU

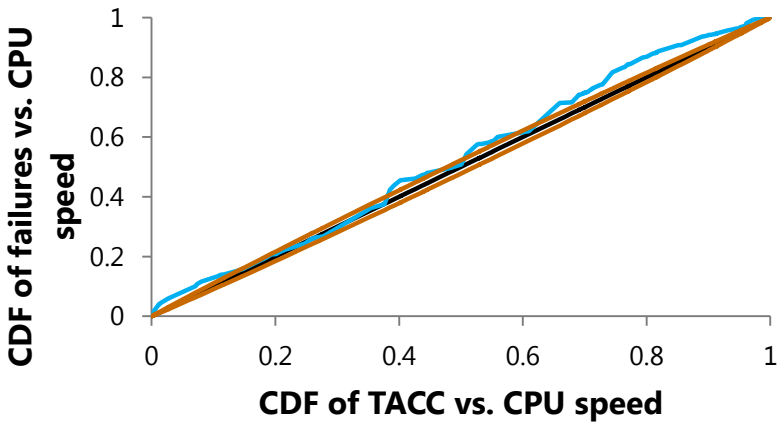
Effect of machine speed (2)



CPU Failures vs. TACC



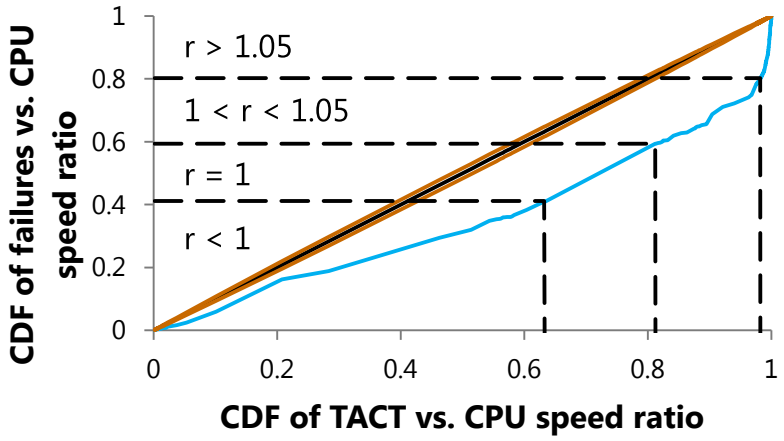
DRAM Failures vs. TACC



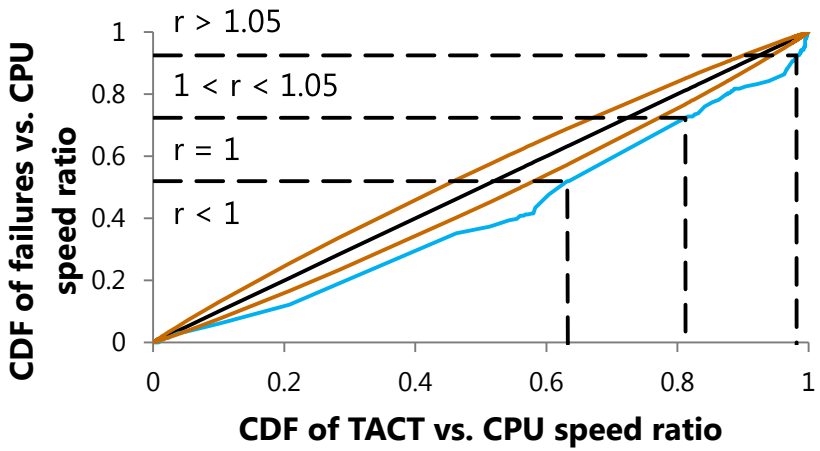
Disk Failures vs. TACC

- 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

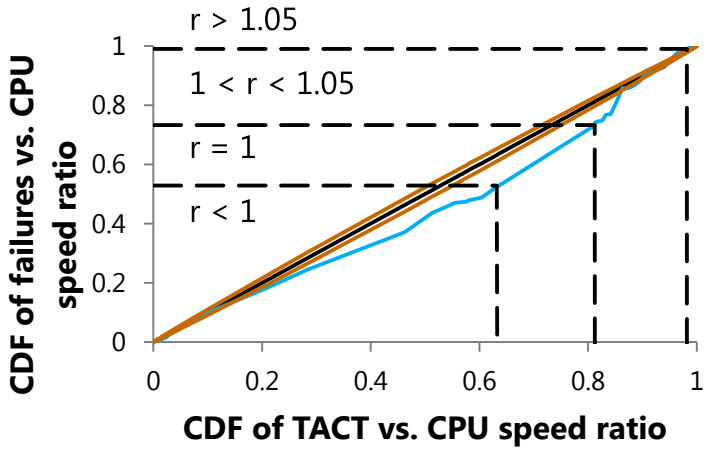
Effect of speed ratio (OC/UC)



CPU Failures



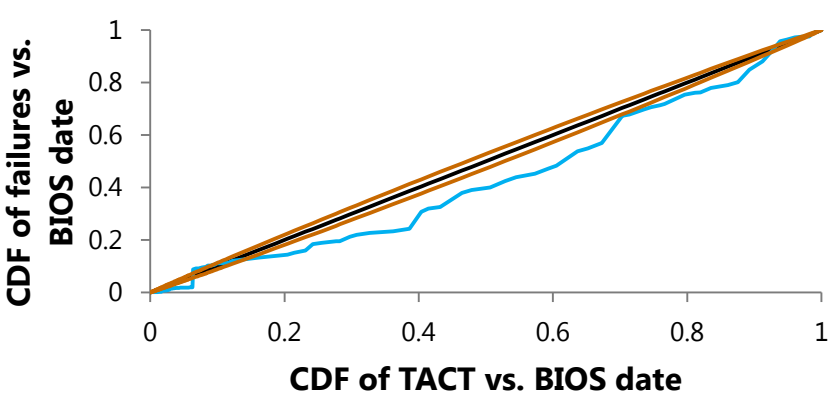
DRAM Failures



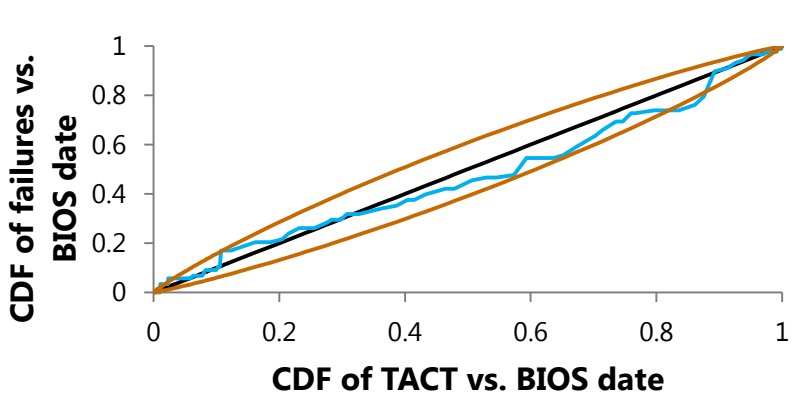
Disk Failures

- CPU failures dramatically impacted as overclocking ratio increases
- Overclocking does not have a large effect on disk failures

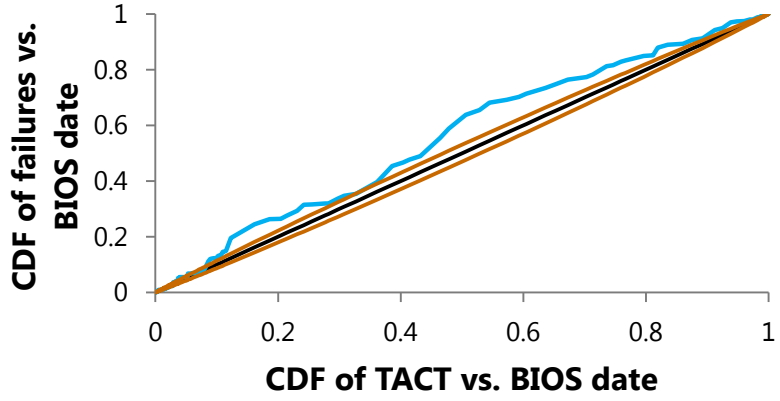
Effect of BIOS date



CPU Failures



DRAM Failures



Disk Failures

- Younger CPUs more likely to fail.
- Older disks more likely to fail.

Outline

- Methodology – diagnosis & data sets
- Analyzing the probability of failure
- Effect of machine class
- Effect of machine characteristics
- Temporal Analysis

Intermittent vs. transient faults

- 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

Outline

- Methodology – diagnosis & data sets
- Analyzing the probability of failure
- Effect of machine class
- Effect of machine characteristics
- Temporal Analysis

Conclusion

- 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.

Microsoft® Research

Faculty Summit 2010