Hard Errors

Mechanisms and Mitigation

Smruti R. Sarangi

Department of Computer Science
Indian Institute of Technology
New Delhi, India

Outline

1. Introduction
2. Different Types of Hard Errors
   - Failure Modes
   - Combining Failure Rates
   - RAMP-II
3. Prevention and Recovery

Motivation

Example

Let us assume that we want to run a 1024 node server for a bank. The server needs to run 24x7. Moreover, we have the following requirements for financial applications

- Preferably zero down time
- Absolutely no data corruption
- Data security
- Maintainability, Serviceability

Let us focus on zero down time, and no data corruption in this talk.

Motivation-II

Zero Down Time

We would ideally want a computer for any kind of a critical application :banking, insurance, military, aerospace, to be on most of the time. Failures can be catastrophic if computers start failing on planes and space crafts. For critical applications like aerospace, we desire a failure rate as close to 0% as possible. However, for most financial applications, we are fine with 99.999% (5 9s reliability). This translates to about 5 mins per year. If there are 1000 nodes on a server, then the failure rate per node needs to be pretty small.

No Data Corruption

It is possible that because of computer faults, a bit can get inverted. This is a very serious failure mechanism in financial applications. Imagine 10,000Rs becoming 2 crore 10,000 Rs, or the reverse !!!

Definition of Hard Errors

We will discuss *Hard Errors* in this chapter.

Definition (Hard Error)

These errors are caused by defects in the silicon, or in the metalization, in the processor package. They are permanent in nature.

Definition of Hard Errors

We will discuss *Hard Errors* in this chapter.

Definition (Hard Error)

These errors are caused by defects in the silicon, or in the metalization, in the processor package. They are permanent in nature.

There are two kinds of hard errors

- **Extrinsic**
  - Contaminants on the silicon surface
  - Open and short circuits
  - Most of them are detected in the *Burn-In* process
  - They are like birth defects. Most of them are detected in the Burn-In process.

Definition of Hard Errors

We will discuss *Hard Errors* in this chapter.

Definition (Hard Error)

These errors are caused by defects in the silicon, or in the metalization, in the processor package. They are permanent in nature.

There are two kinds of hard errors

- **Extrinsic**
  - Contaminants on the silicon surface
  - Open and short circuits
  - Most of them are detected in the *Burn-In* process
  - They are like birth defects. Most of them are detected in the Burn-In process.
- **Intrinsic**
  - Caused by wear and tear
  - These are age related defects that increase over time

Reliability Challenges

Common failure modes

- Electomigration
  - Over time the structure of a wire changes. Metal atoms tend to amass at one end at one end of the wire.

Reliability Challenges

Common failure modes

- Electomigration
  - Over time the structure of a wire changes. Metal atoms tend to amass at one end at one end of the wire.
- Stress Migration
  - Because of varying amounts of thermal expansion along a wire, electrons migrate to different parts of the wire.

Reliability Challenges

Common failure modes

- Electomigration
  - Over time the structure of a wire changes. Metal atoms tend to amass at one end at one end of the wire.
- Stress Migration
  - Because of varying amounts of thermal expansion along a wire, electrons migrate to different parts of the wire.
- Time Dependent Dielectric Breakdown
  - The gate dielectric in a transistor gradually breaks down over time. It ultimately becomes a short circuit.

Reliability Challenges

Common failure modes

- Electomigration
  - Over time the structure of a wire changes. Metal atoms tend to amass at one end at one end of the wire.
- Stress Migration
  - Because of varying amounts of thermal expansion along a wire, electrons migrate to different parts of the wire.
- Time Dependent Dielectric Breakdown
  - The gate dielectric in a transistor gradually breaks down over time. It ultimately becomes a short circuit.
- Thermal Cycling
  - The solder joints between the package and die interface experience constant changes in temperature. This causes metal fatigue, and ultimate breakage.

Outline

1. Introduction
2. Different Types of Hard Errors
   - Failure Modes
   - Combining Failure Rates
   - RAMP-II
3. Prevention and Recovery

Electromigration

Conducting electors in aluminum or copper interconnects transfer some of their momentum to surrounding metal atoms. These metal atoms slowly drift along with the electrons. Over time, atoms from one end drift and accumulate at the other end. The depletion sites exhibit increased resistance. Finally, the wire ceases to conduct.

$$MTTF_{em} \propto (J - J_{crit})^{-n} e^{\frac{E_{aEM}}{kT}}$$

- n and E_{aEM} are constants that depend on the metal type. They are close to 1.
- J is the current density. J_{crit} is typically 2 orders of magnitude smaller than J.

Electromigration II

$$J = \frac{CV_{dd}}{WH} * f * p$$

- C: capacitance, V_{dd}: supply voltage, W: width
- H: height, f: frequency, p: switching probability

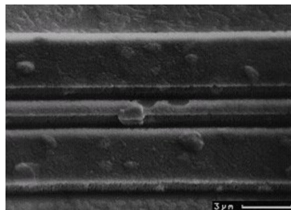


Figure 1: Electromigration (courtesy: Philip Koopman, CMU)

Stress Migration

This is a phenomenon similar to electromigration. Atoms migrate because of a variable amount of thermal expansion. This thermomechanical stress gives wires and contacts irregular shapes. Much of the reasons behind stress migration are not well understood.

$$MTTF_{sm} \propto |T_0 - T|^{-n} e^{\frac{E_{aSM}}{kT}}$$

- T_0: Temperature at which the metal was originally deposited (500K)
- T: Operating temperature, E_{aSM} and n are material dependent constants
- $E_{aSM} = 2.5$ and $n = 0.9$ for copper interconnects

Time Dependent Dielectric Breakdown

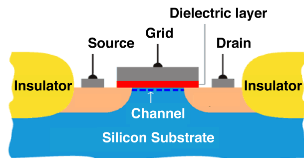


Figure 2: MOS Transistor (courtesy Gian-Marco Rignanese)

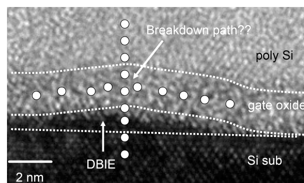


Figure 3: Dielectric Breakdown (courtesy Kin-Leong Pey)

Time Dependent Dielectric Breakdown - II

The gate dielectric is typically very thin (about 2 nm). Over time it wears down and current passes through it. It forms a conducting path. It is hyper-exponentially dependent on temperature.

$$MTTF_{tddb} \propto \frac{1}{V}^{(a-bT)} e^{\frac{X+\frac{Y}{T}+ZT}{kT}}$$

Thermal Cycling - Metal Fatigue

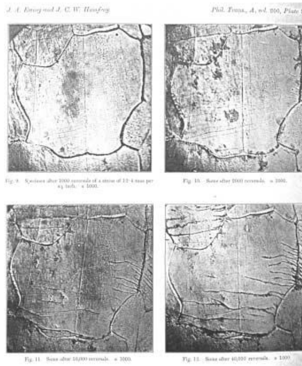


Figure 4: Metal Fatigue (wikipedia)

Thermal Cycling-II

- Cycles of expansion and contraction give rise to metal fatigue especially at the I/O contacts.
- Over time, cracks begin to form
- There are two kinds of thermal cycles
  - Large cycles : Powering up/down the processor
  - Small cycles: Periods of activity/inactivity while executing a workload.

Coffin Manson Equation

$$MTTF_{tc} \propto \left(\frac{1}{T - T_{ambient}} \right)^{q}$$

- q = 2.35 in the reference.

Outline

1. Introduction
2. Different Types of Hard Errors
   - Failure Modes
   - Combining Failure Rates
   - RAMP-II
3. Prevention and Recovery

Sum of Failure Rates Model

Assumptions

- The processor is a series failure system. One failure makes the entire processor fail.
- Every failure mechanism has an exponential distribution.

Definition

Failure Rate $h(t)$: It is the conditional probability that a component will fail between t and $(t + \delta t)$, given the fact that it has survived till t. $h(t) = \lambda$, means that the application has a constant failure rate.

Example

The exponential distribution $f(x) = \lambda e^{-\lambda x}$, has a constant failure rate, λ.

Sum of Failure Rates Model - II

Definition

FIT: Failure-in-time. The number of failures in a billion hours.

$$MTTF_p = \frac{1}{\lambda_p} = \frac{10^9}{FIT} = \frac{1}{\sum_i \sum_j \lambda_{ij}}$$

- λ_{ij} is the failure rate of i_{th} structure due to the j_{th} failure mechanism.

Mathematical Trivia

Theorem

The failure rate for the exponential distribution is λ.

Proof.

Let the time of failure be t_f. Let the failure rate be $h(t)$. The exponential equation for the pdf of the failure is $f(t) = \lambda e^{-\lambda t}$. We have:

$$
\begin{aligned}
h(t) &= P(t \leq t_f \leq t + \delta t | t_f < t) \\
&= \frac{P((t \leq t_f \leq t + \delta t) \wedge (t_f < t))}{P(t_f < t)} \\
&= \frac{f(t)}{1 - \int_0^t f(x)} \\
&= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} \\
&= \lambda
\end{aligned}
\tag{1}
$$

□

Mathematical Trivia - II

Theorem

The MTTF for the exponential distribution is $\frac{1}{\lambda}$.

Proof.

The exponential equation for the pdf of the failure is $f(t) = \lambda e^{-\lambda t}$.
We have:

$$\begin{aligned} MTTF &= \int_0^{\infty} tf(t)dt \\ &= \int_0^{\infty} t\lambda e^{-\lambda t} dt \\ &= \frac{1}{\lambda}\int_0^{\infty} xe^{-x}dx \quad (x = \lambda t) \\ &= \frac{1}{\lambda}\left((-xe^{-x}) \mid_0^{\infty} + \int_0^{\infty} e^{-x}dx\right) \\ &= \frac{1}{\lambda} \end{aligned} \tag{2}$$

RAMP Model

Assumptions

- Processors are designed to have an MTTF of 30 years.
- The FIT value is therefore 4000.
- Worst case operating conditions will lead to a failure rate of 30 years.
- The total failure rate is distributed evenly across the four failure modes.

Let us now try to find the constants in the equations.

- There are four parameters considered : T, V, f, and p(activity factor)
- Let there values for the worst case be : T_{qual}, V_{qual}, f_{qual}, and p_{qual}.
- V_{qual} is the highest supply voltage (about 1.2 V), f_{qual} is the highest frequency
- p_{qual} is 1. T_{qual} is dependent on the manufacturing specs. Typically 393K (120 °C).

Ramp Model Contd...

- T_{qual} is typically provided by the manufacturer.
- Most references assume a value of 120 °C.
- The RAMP model computes the values of the proportionality constants for the four failure modes
- The RAMP paper evaluates the hard error FIT rates for different configurations.

Outline

1. Introduction
2. Different Types of Hard Errors
   - Failure Modes
   - Combining Failure Rates
   - RAMP-II
3. Prevention and Recovery

Normal Distribution

Central Limit Theorem

Let $(x_1, \ldots, x_n)$ be a sequence of numbers that are identically and independently distributed(i.i.d). The sum of the sequence for large values of n, is normally distributed.

Normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ is the mean, σ is its variance
- The normal distribution is thus a very fundamental distribution.
- Most natural phenomena are modeled by this distribution, or by variants of it.

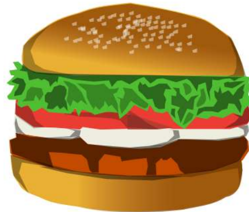
Log-Normal Distribution

- The exponential distribution has a constant failure rate.
  - Most failure rates increase over time
  - This is not an accurate assumption
- Such kind of failures are typically model by log-normal distributions
  - x has a log-normal distribution, if $ln(x)$ has a normal distribution

Log-Normal distribution:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}} \quad , x > 0$$

Food for Thought



Question

Let us consider a failure process, in which the amount of degradation between the time instants, t_n and t_{n-1} is proportional to the degradation at t_{n-1}. If the degradation at t_n is x_n, we have: $x_n - x_{n-1} = \alpha_n x_{n-1}$. Let us further assume that α_n has an i.i.d distribution. Prove that x_n has a log-normal distribution.

NBTI (Not in course ...)

- RAMP-II adds a fifth failure mode called NBTI
  - NBTI : Negative bias temperature instability

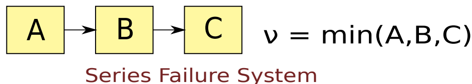
Definition

NBTI: When the p-FET gate is biased negative w.r.t to the drain and source, then positive charge tends to accumulate on the gate oxide. This accumulation causes the threshold voltage of the transistor to increase, ultimately leading to its failure.

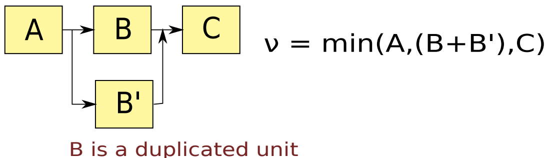
$$MTTF \propto \left[\left(ln \left(\frac{A}{1 + 2e^{\frac{B}{kT}}} \right) - ln \left(\frac{A}{1 + 2e^{\frac{B}{kT}}} - C \right) \right) * \frac{T}{e^{-\frac{D}{kT}}} \right]^{\frac{1}{\beta}}$$

Duplication

- Duplication : Maintain multiple copies of a functional unit
  - Example : Two adders, or two versions of the fetch unit
  - Keep one unit permanently powered off
  - When the first unit fails, switch to the second one
  - Negative Point: We are increasing the area of the chip in the process



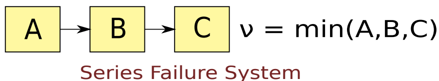
Series Failure System



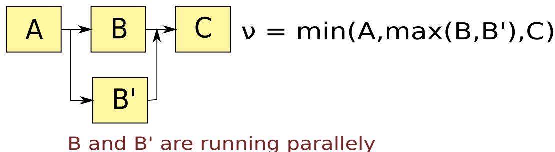
B is a duplicated unit

Graceful Degradation

- Degradation : Gradually reduce the functionality of a unit
  - Example : We can gradually decrease the size of buffers like the branch predictor. This will decrease performance but not correctness.
  - The instruction queue can be segmented. We can gradually decrease the number of segments.
  - Positive Point: Very little area penalty

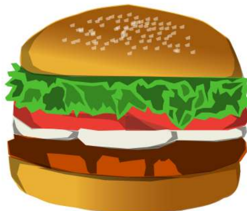


Series Failure System



B and B' are running parallelly

Food for Thought



Theorem

The distribution of the minimum of n exponential random variables with rates, $\lambda_1, \ldots, \lambda_n$, is another exponential variable with rate, $(\lambda_1 + \ldots + \lambda_n)$.

Handling Faults in Memories

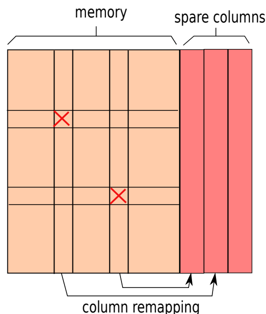


Figure 5: Remapping columns for a memory bank

- Perform a BIST (Built-in self test) to find faulty memory cells
- Change the decoder logic to map the columns with faulty cells to spare columns
- Can be done right after manufacturing, or can be done online

The Case for Lifetime Reliability-Aware Microprocessors, Jayanth Srinivasan, Sarita V. Adve, Pradip Bose, and Jude A. Rivers, Proceedings of 31st International Symposium on Computer Architecture (ISCA '04) June 2004.
http://rsim.cs.illinois.edu/~srinivsn/Pubs/isca04.pdf

Jayanth Srinivasan, Sarita V. Adve, Pradip Bose, Jude A. Rivers: Exploiting Structural Duplication for Lifetime Reliability Enhancement. ISCA 2005: 520-531
http://rsim.cs.illinois.edu/~srinivsn/Pubs/isca05.pdf