Fail-Slow at Scale
Evidence of Hardware Performance Faults
in Large Production Systems

Haryadi S. Gunawi¹, Riza O. Suminto¹, Russell Sears², Casey Golliher², Swaminathan Sundararaman³, Xing Lin⁴, Tim Emami⁴, Weiguang Sheng⁵, Nematollah Bidokhti⁵, Caitie McCaffrey⁶, Gary Grider⁷, Parks M. Fields⁷, Kevin Harms⁸, Robert B. Ross⁸, Andree Jacobson⁹, Robert Ricci¹⁰, Kirk Webb¹⁰, Peter Alvaro¹¹, H. Birali Runesha¹², Mingzhe Hao¹, Huaicheng Li¹

¹THE UNIVERSITY OF CHICAGO  ²PURE STORAGE  ³PARALLEL  ⁴NETAPP  ⁵HUAWEI  ⁶TWITTER  ⁷LOS ALAMOS NATIONAL LABORATORY  ⁸ARGONNE  ⁹NEW MEXICO CONSORTIUM  ¹⁰THE UNIVERSITY OF UTAH  ¹¹UNIVERSITY OF SANTA CRUZ  ¹²THE UNIVERSITY OF CHICAGO

Research Computing Center
"…a 1Gb NIC card on a machine that suddenly only transmits at 1 kbps, this slow machine caused a chain reaction upstream in such a way that the 100 node cluster began to crawl at a snail's pace."
More anecdotes? All hardware?

- **Disk** throughput dropped to 100 KB/s due to vibration
- **SSDs** stalled for seconds due to firmware bugs
- **Memory cards** degraded to 25% speed due to a loose NVDIMM connection
- **CPUs** ran in 50% speed due to lack of power
Fail-slow Hardware

- Hardware that is still running and functional but in a degraded mode, significantly slower than its expected performance

- In existing literature:
  - “fail-stutter” [Arpaci-Dusseau(s), HotOS ’11]
  - “gray failure” [Huang et al. @ HotOS ’17]
  - “limp mode” [Do et al. @ SoCC ’13, Gunawi et al. @ SoCC ’14, Kasick et al. @ FAST ’10]
  - (But only 8 stories per paper on avg. and mixed with SW issues)
Believe it?

Fail-slow hardware is “not” real.
It is rare.

Let’s write a paper together

Yes, it’s real!
Evidences from ...

<table>
<thead>
<tr>
<th>Institution</th>
<th>#Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company 1</td>
<td>&gt;10,000</td>
</tr>
<tr>
<td>Company 2</td>
<td>150</td>
</tr>
<tr>
<td>Company 3</td>
<td>100</td>
</tr>
<tr>
<td>Company 4</td>
<td>&gt;1,000</td>
</tr>
<tr>
<td>Company 5</td>
<td>&gt;10,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Institution</th>
<th>#Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univ. A</td>
<td>300</td>
</tr>
<tr>
<td>Univ. B</td>
<td>&gt;100</td>
</tr>
<tr>
<td>Univ. C</td>
<td>&gt;1,000</td>
</tr>
<tr>
<td>Univ. D</td>
<td>500</td>
</tr>
<tr>
<td>Nat’l Labs X</td>
<td>&gt;1,000</td>
</tr>
<tr>
<td>Nat’l Labs Y</td>
<td>&gt;10,000</td>
</tr>
<tr>
<td>Nat’l Labs Z</td>
<td>&gt;10,000</td>
</tr>
</tbody>
</table>

Table 2: Operational scale.
Data and Methodology

- **101 reports**
  - Unformatted text
  - Written by engineers and operators (who still remember the incidents)
  - 2000-2017 (mostly after 2010)
  - Limitations and challenges:
    - No hardware-level performance logs [in formatted text]
    - No large-scale statistical analysis

- **Methodology**
  - An institution reports a *unique* set of root causes
    - “A corrupt buffer that slows down the networking card (causing packet loss and retransmission)”
    - Counted as 1 report from the institution (although might have happened many times)
§3.1 Varying root causes: Fail-slow hardware can be induced by internal causes such as firmware bugs or device errors/wear-outs as well as external factors such as configuration, environment, temperature, and power issues.

§3.2 Faults convert from one form to another: Fail-stop, -partial, and -transient faults can convert to fail-slow faults (e.g., the overhead of frequent error masking of corrupt data can lead to performance degradation).

§3.3 Varying symptoms: Fail-slow behavior can exhibit a permanent slowdown, transient slowdown (up-and-down performance), partial slowdown (degradation of sub-components), and transient stop (e.g., occasional reboots).

§3.4 A long chain of root causes: Fail-slow hardware can be induced by a long chain of causes (e.g., a fan stopped working, making other fans run at maximal speeds, causing heavy vibration that degraded the disk performance).

§3.4 Cascading impacts: A fail-slow hardware can collapse the entire cluster performance; for example, a degraded NIC made many jobs lock task slots/containers in healthy machines, hence new jobs cannot find enough free slots.

§3.5 Rare but deadly (long time to detect): It can take hours to months to pinpoint and isolate a fail-slow hardware due to many reasons (e.g., no full-stack visibility, environment conditions, cascading root causes and impacts).

§6.1 To vendors: When error masking becomes more frequent (e.g., due to increasing internal faults), more explicit signals should be thrown, rather than running with a high overhead. Device-level performance statistics should be collected and reported (e.g., via S.M.A.R.T) to facilitate further studies.

§6.2 To operators: 39% root causes are external factors, thus troubleshooting fail-slow hardware must be done online. Due to the cascading root causes and impacts, full-stack monitoring is needed. Fail-slow root causes and impacts exhibit some correlation, thus statistical correlation techniques may be useful (with full-stack monitoring).

§6.3 To systems designers: While software systems are effective in handling fail-stop (binary) model, more research is needed to tolerate fail-slow (non-binary) behavior. System architects, designers and developers can fault-inject their systems with all the root causes reported in this paper to evaluate the robustness of their systems.

Table 1: Summary of our findings and suggestions.
① Varying root causes
- Internal causes: firmware bugs, device errors
- External causes: temperature, power, environment, and configuration

② Faults convert
- Fail-stop, -partial, -transient → fail-slow

③ Varying symptoms
- Permanent, transient, and partial slowdown, and transient stop

④ Cascading nature
- Cascading root causes
- Cascading impacts

⑤ Rare but deadly
- Long time to detect (hours to months)
# Varying root causes

<table>
<thead>
<tr>
<th>Root</th>
<th>Hardware types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSD</td>
</tr>
<tr>
<td>Device errors</td>
<td>10</td>
</tr>
<tr>
<td>Firmware bugs</td>
<td>6</td>
</tr>
<tr>
<td>Temperature</td>
<td>1</td>
</tr>
<tr>
<td>Power</td>
<td>1</td>
</tr>
<tr>
<td>Environment</td>
<td>3</td>
</tr>
<tr>
<td>Configuration</td>
<td>1</td>
</tr>
<tr>
<td>Unknown</td>
<td>0</td>
</tr>
</tbody>
</table>
Varying root causes

- **Internal**
  - **Device errors/wearouts**
    - Ex: SSD read disturb/retry + page reconstruction $\rightarrow$ longer latency and more load

**Read Disturb Effect on $V_{\text{th}}$ Distribution**

- $V_{\text{th}}$ gradually increases with wearout.
- ER state
- Normed Threshold Voltage

- 0 (No Read Disturb)
- 0.25M Read Disturb
- 0.5M Read Disturb
- 1M Read Disturb

$V_{\text{th}}$ = $v_1, v_2, v_3, v_4$

$1 = \text{XOR (0, 2, P012)}$

read(page X, $V_{\text{th}}=v_1$)$ \times \times \times$
read(page X, $V_{\text{th}}=v_2$)$ \times \times \times$
read(page X, $V_{\text{th}}=v_3$)$ \times \times \times$
read(page X, $V_{\text{th}}=v_4$)$ \checkmark$

4x slower!

**Voltage shift**

**RAIN: Redundant Array of Independent NAND**

- read p0
- read p1
- read p2
- read P012

$\times \times \times$ read retries!

$\times \text{ECC error}$

Picture from http://slideplayer.com/slide/10095910/
Varying root causes

- Internal
  - Device errors
  - Firmware bugs
    - [No details, proprietary component]
    - SSD firmware bugs throttled $\mu$s to ms read performance
    - Another example: 840 EVO firmware bugs [2014]


1. Varying root causes

- **Internal** Device errors and firmware bugs [More details in paper]

<table>
<thead>
<tr>
<th>SSD</th>
<th>Disk</th>
<th>Memory</th>
<th>Network</th>
<th>Processors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firmware bugs (us to ms read performance, internal metadata writes triggering assertion); Read retries with different voltages; RAIN/parity-based read reconstruction; Heavy GC in partially-failing SSD (not all chips are created equal); Broken parallelism by suboptimal wear-leveling; Hot temperature to wear-outs, repeated erases, and reduced space; Write amplification.</td>
<td>Firmware bugs (jitters, occasional timeouts, read retries, read-after-write mode); Device wearouts (disabling bad platters); Weak heads (gunk/dust accumulates between disk heads and platters); and other external factors such as temperature and vibration.</td>
<td>Address errors causing expensive ECC checks and repairs; Reduced space causing more cache hits; Loose NVDIMM connection; SRAM control-path errors causing recurrent reboots (transient stop).</td>
<td>Firmware bugs (buggy routing algorithm, multicast bad performance); NIC driver bugs; buggy switch-NIC auto-negotiation; Starving from electrons (bad design specification); bad VSCEL laser; Bitflips in device buffer; Loss packets cause TCP retries and collapse.</td>
<td>Buggy BIOS firmware down-clocking CPUs; Other external causes such as hot temperature and lack of power.</td>
</tr>
</tbody>
</table>
1 Varying root causes

- **Internal** [Device errors, firmware bugs]
- **External**
  - **Temperature**

**Hot temperature**
- Corrupt packets
- Heavy TCP retransmission

**Faster SSD wearouts, bad Vth**
- more read retries

**Cold-air-under-the-floor system**

- **Slower disk performance** at bottom of the rack (read-after-write mode)
① Varying root causes

- Internal [Device errors, firmware bugs]
- External
  - Temperature
  - Power

4 machines, 2 power supplies
1 dead power → 50% CPU speed

Power-hungry applications → throttling neighboring CPUs
Varying root causes

- **Internal** [Device errors, firmware bugs]
- **External**
  - Temperature
  - Power
  - **Environment**
    - Altitude, pinched cables, etc.
  - **Configuration**
    - A BIOS incorrectly downclocking CPUs of new machines
    - Initialization code disabled processor cache
1 Varying root causes Device errors, firmware, temperature, power, environment, configuration

2 Faults convert

- Fail-transient $\rightarrow$ fail-slow

Bit flips $\rightarrow$ ECC repair
(error masking)

Okay if rare

But, frequent errors
$\rightarrow$ frequent error-masking/repair
$\rightarrow$ repair latency becomes the common case
① Varying root causes
Device errors, firmware, temperature, power, environment, configuration

② Faults convert
- Fail-transient $\rightarrow$ fail-slow
- Fail-partial $\rightarrow$ fail-slow

“Not all chips are created equal” (some chips die faster)

$\rightarrow$ Reduced overprovisioned space
$\rightarrow$ More frequent GCs $\rightarrow$ Slow SSD

Picture from https://flashdba.com/2014/06/20/understanding-flash-blocks-pages-and-program-erases/
① Varying root causes  Device errors, firmware, temperature, power, environment, configuration

② Faults convert
  - Fail-transient $\rightarrow$ fail-slow
  - Fail-partial $\rightarrow$ fail-slow

Custom memory chips that mask (hide) bad addresses

Exposed space

- X GB
- < X GB
- << X GB

Higher cache misses (fail-slow)
① Varying root causes  Device errors, firmware, temperature, power, environment, configuration

② Faults convert  Fail-stop, -transient, -partial → fail-slow

③ Varying symptoms
   - Permanent slowdown
1. **Varying root causes** Device errors, firmware, temperature, power, environment, configuration

2. **Faults convert** Fail-stop, -transient, -partial → fail-slow

3. **Varying symptoms**
   - Permanent slowdown
   - **Transient slowdown**

![Diagrams showing varying performance and temperature](image)
1. Varying root causes: Device errors, firmware, temperature, power, environment, configuration

2. Faults convert: Fail-stop, -transient, -partial → fail-slow

3. Varying symptoms:
   - Permanent slowdown
   - Transient slowdown
   - Partial slowdown

    ![Diagram showing performance degradation and sub-components](image)

- Slow reads (ECC repairs)
- Fast reads
- Small packets (fast)
- >1500-byte packets (very slow)
- [Buggy firmware/config related to jumbo frames]
① **Varying root causes**  Device errors, firmware, temperature, power, environment, configuration

② **Faults convert**  Fail-stop, -transient, -partial → fail-slow

③ **Varying symptoms**
- Permanent slowdown
- Transient slowdown
- Partial slowdown
- **Transient stop**

A bad batch of SSDs "disappeared" and then reappeared

A firmware bug triggered hardware assertion failure

Uncorrectable bit flips in SRAM control paths

---

**Performance**

**Time**

**Host Bus Adapter**

recurrent resets
① **Varying root causes**  
Device errors, firmware, temperature, power, environment, configuration

② **Faults convert**  
Fail-stop, -transient, -partial → fail-slow

③ **Varying symptoms**  
Permanent, transient, partial slowdown and transient stop

④ **Cascading nature**

- **Cascading root causes**

Fans
- normal speed
- One died

Other fans
- maximum speed
- Noise and vibration

Disk
- throughput collapses to  
  KB/s

Bad disks? No!
1. **Varying root causes**  
   Device errors, firmware, temperature, power, environment, configuration

2. **Faults convert**  
   Fail-stop, -transient, -partial → fail-slow

3. **Varying symptoms**  
   Permanent, transient, partial slowdown and transient stop

4. **Cascading nature**
   - Cascading root causes
   - Cascading **impacts** e.g. in Hadoop MapReduce

A **fast** map task  
(read locally)

![Diagram showing slow NIC](image)

**All reducers are slow**  
(“no” stragglers → no Speculative Execution)

Use (lock-up) task **slots** in  
healthy machines for a long time

Eventually **no free** task slots  
→ **Cluster collapse**
① Varying root causes
Device errors, firmware, temperature, power, environment, configuration

② Faults convert
Fail-stop, -transient, -partial → fail-slow

③ Varying symptoms
Permanent, transient, partial slowdown and transient stop

④ Cascading nature
- Cascading root causes
- Cascading impacts

Facebook Hadoop Jobs, 30 nodes

[From PBSE @ SoCC ’17]
1. Varying root causes  Device errors, firmware, temperature, power, environment, configuration

2. Faults convert  Fail-stop, -transient, -partial → fail-slow

3. Varying symptoms  Permanent, transient, partial slowdown and transient stop

4. Cascading nature

5. Rare but deadly
   - 13% detected in hours
   - 13% in days
   - 11% in weeks
   - 17% in months
   - (50% unknown)

Why?
- External causes and cascading nature
  (vibration → slow disk); offline testing passes
- No full-stack monitoring/correlation
  hot temperature → slow CPUs → slow Hadoop → debug Hadoop logs?
- Rare? Ignore?
① Varying root causes  Device errors, firmware, temperature, power, environment, configuration

② Faults convert  Fail-stop, -transient, -partial → fail-slow

③ Varying symptoms  Permanent, transient, partial slowdown and transient stop

④ Cascading nature

⑤ Rare but deadly

Suggestions to vendors, operators, and systems designers

<table>
<thead>
<tr>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>§6.1 To vendors:</strong> When error masking becomes more frequent (<em>e.g.</em>, due to increasing internal faults), more explicit signals should be thrown, rather than running with a high overhead. Device-level performance statistics should be collected and reported (<em>e.g.</em>, via S.M.A.R.T) to facilitate further studies.</td>
</tr>
<tr>
<td><strong>§6.2 To operators:</strong> 39% root causes are external factors, thus troubleshooting fail-slow hardware must be done online. Due to the cascading root causes and impacts, full-stack monitoring is needed. Fail-slow root causes and impacts exhibit some correlation, thus statistical correlation techniques may be useful (with full-stack monitoring).</td>
</tr>
<tr>
<td><strong>§6.3 To systems designers:</strong> While software systems are effective in handling fail-stop (binary) model, more research is needed to tolerate fail-slow (non-binary) behavior. System architects, designers and developers can fault-inject their systems with all the root causes reported in this paper to evaluate the robustness of their systems.</td>
</tr>
</tbody>
</table>
Varying root causes
Device errors, firmware, temperature, power, environment, configuration

Faults convert
Fail-stop, -transient, -partial -> fail-slow

Varying symptoms
Permanent, transient, partial slowdown and transient stop

Cascading nature

Rare but deadly

Conclusion:
Modern, advanced systems
+ Fail-slow hardware

Thank you!

Questions?
EXTRA
Suggestions

- **To vendors:**
  - Make the implicits explicit
    - Frequent error masking → hard errors
  - Record/expose device-level performance statistics

- **To operators:**
  - Online diagnosis
    - (39% root causes are external)
  - Full-stack monitoring
  - Full-stack statistical correlation

- **To systems designers:**
  - Make the implicits explicit
    - Jobs retried “infinite” time
  - Convert fail-slow to fail-stop? (challenging)
  - Fail-slow fault injections
<table>
<thead>
<tr>
<th>HW Type</th>
<th>Perm.</th>
<th>Trans.</th>
<th>Partial</th>
<th>Tr. Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Disk</td>
<td>9</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Mem</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Net</td>
<td>21</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>CPU</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4: **Fail-slow symptoms across hardware types.**

<table>
<thead>
<tr>
<th>Root</th>
<th>Perm.</th>
<th>Trans.</th>
<th>Partial</th>
<th>Tr. Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERR</td>
<td>19</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>FW</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>TEMP</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PWR</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ENV</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>CONF</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UNK</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5: **Fail-slow symptoms across root causes.**
Operators

- Cannot use application bandwidth check (all are affected)

*Facebook Hadoop Jobs, 30 nodes*

- Normal
- w/ 1 limping node

- ~170 jobs/hour
- 1 job/hour !!!

Hadoop, not fully tail/limpware tolerant??
① Varying root causes
Device errors, firmware, temperature, power, environment, configuration

② Faults convert

- Fail-stop $\rightarrow$ fail-slow
  - Fail-stop power $\rightarrow$ fail-slow CPUs
  - Fail-stop disk $\rightarrow$ fail-slow RAID