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

Haryadi S. Gunawi[1], Riza O. Suminto[1], Russell Sears[2], Casey Golliher[2], Swaminathan Sundararaman[3], Xing Lin[4],
Tim Emami[4], Weiguang Sheng[5], Nematollah Bidokhti[5], Caitie McCaffrey[6], Gary Grider[7], Parks M. Fields[7], Kevin Harms[8],
Robert B. Ross[8], Andree Jacobson[9], Robert Ricci[10], Kirk Webb[10], Peter Alvaro[11], H. Birali Runesha[12], Mingzhe Hao[1], Huaicheng Li[1]

[2] PURESTORAGE [3] Parallel/\ [4] NetApp [5] HUAWEI [6] twitter [7] Los Alamos NATIONAL LABORATORY EST.1943 [8] Argonne NATIONAL LABORATORY [9] New Mexico CONSORTIUM [10] THE UNIVERSITY OF UTAH [11] UNIVERSITY OF CALIFORNIA SANTA CRUZ [12] THE UNIVERSITY OF CHICAGO Research Computing Center

Fail-Slow Hardware

Definition: hardware that is still running and functional but in a degraded mode, slower than its expected performance.

Examples:

- Disk throughput drop to 100 KB/s due to vibration.
- SSD operations stall for seconds due to firmware bugs.
- Memory cards can degrade to 25% of normal speed due to loose NVDIMM connection.
- CPUs run in 50% speed due to lack of power.
- NIC performance can collapse to Kbps level due to buffer corruption and retransmission.

Methodology

- Collect **101** reports of fail-slow behaviors from **12** institutions.
- Detailed to hardware types, root causes, symptoms, and impact to high-level software.
- Incident reported range between 2000 to 2017, with only 30 reports predating 2010.
- Each institutions report a unique set of root causes.

The Institutions

Institution	#Nodes	Institution	#Nodes
Company 1	>10,000	Univ. A	300
Company 2	150	Univ. B	>100
Company 3	100	Univ. C	>1,000
Company 4	>1,000	Univ. D	500
Company 5	>10,000	Nat'l Labs X	>1,000
		Nat'l Labs Y	>10,000
		Nat'l Labs Z	>10,000

Table 2: Operational Scale

Observations

Root Causes

Hardware: SSD, disk, memory (Mem), network (Net), and processors (CPU).
Internal root causes: errors(ERR), firmware issues (FW)
External root causes: temperature (TEMP), power (PWR), environment (ENV), and configuration (CONF)

unknown (UNK) implies that the operators cannot pinpoint the root cause, but simply replaced the hardware.

	Hardware types					
Root	SSD	Disk	Mem	Net	CPU	Total
ERR	10	8	9	10	3	40
FW	6	3	0	9	2	20
TEMP	1	3	0	2	5	11
PWR	1	0	1	0	6	8
ENV	3	5	2	4	4	18
CONF	1	1	0	2	3	7
UNK	0	3	1	2	2	8
Total	22	23	13	29	25	112

Table 3: Root causes across hardware types.

Fail-Slow Symptoms

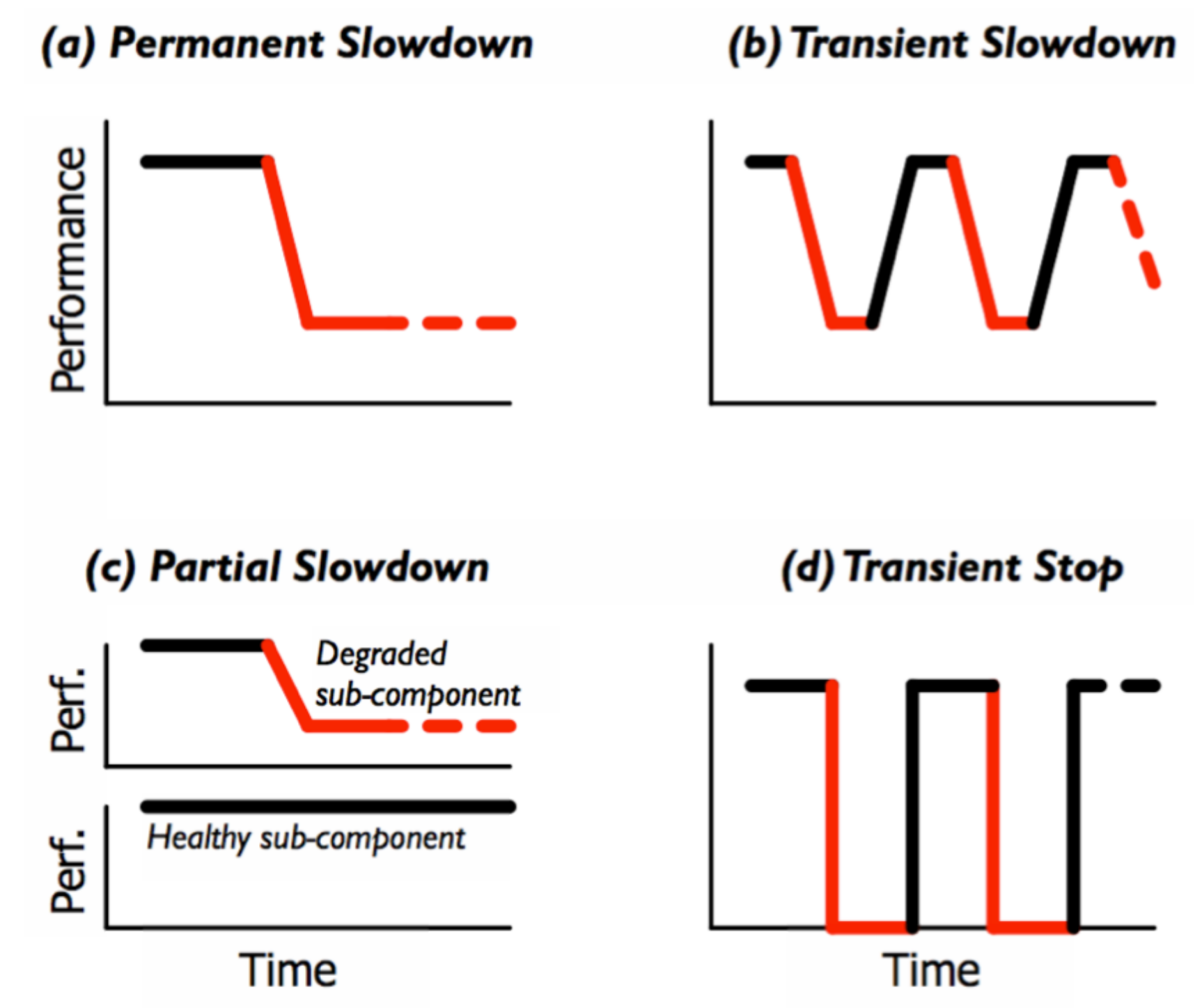

Figure 1: Fail-slow symptoms.

	Symptoms			
HW Type	Perm.	Trans.	Partial	Tr. Stop
SSD	6	7	3	3
Disk	9	4	3	5
Mem	7	1	0	4
Net	21	0	5	2
CPU	10	6	1	3

Table 4: Fail-slow symptoms across hardware types.

	Symptoms			
Root	Perm.	Trans.	Partial	Tr. Stop
ERR	19	8	7	6
FW	11	3	1	4
TEMP	6	2	1	2
PWR	3	2	1	2
ENV	11	3	3	1
CONF	6	1	0	0
UNK	5	1	0	2

Table 5: Fail-slow symptoms across root causes.

Cascading Causes & Impacts

*"… **1Gb NIC** card on a machine that suddenly starts transmitting at **1 Kbps** … [making] the performance of entire workload for a **100 node cluster was crawling at a snail's pace**"*

Rare but Deadly : Long TTD

1% of the cases are detected in minutes, 13% in hours, 13% in days, 11% in weeks, and 17% in months (and unknown time in 45%).

Findings and Suggestions

Internal Root Causes

SSD: Firmware bugs; Read retries with different voltages; RAIN/parity-based read reconstruction; Heavy GC in partially-failing SSD; Broken parallelism by suboptimal wear-leveling; Hot temperature to wear-outs, repeated erases, and reduced space; Write amplification; Not all chips are created equal.

Disk: Firmware bugs; Device errors; Weak heads; and others.

Memory: Device errors; External causes; Unknown causes; SRAM errors.

Network: Firmware bugs; NIC driver bugs; Device errors; External causes; Unknown causes.

Processors: External causes.

External Root Causes

Temperature: Clogged air filter; Cold environment; Broken fans; Improper design/assembly/operation.

Power: Insufficient capacitors; PCU firmware bugs; Fail-partial power supply; Power hungry neighbors; Faulty motherboard sensors.

Environment: Altitude & cosmic events; Loose interconnects; Vibrations; Environment and operating condition mismatch; Unknown causes.

Configuration: Buggy BIOS firmware; Human mistakes.

Important Findings and Observations
§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).
Suggestions
§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

Appeared in the16th USENIX Conference on File and Storage Technologies (FAST), 2018