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]

1 THE UNIVERSITY OF CHICAGO   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

1ˢᵗ anecdote

1ˢᵗ anecdote

"…a *1Gb* **NIC card** *on a machine that suddenly only transmits at 1 kbps,*

1ˢᵗ anecdote

*"…a 1Gb **NIC card** on a machine that suddenly only transmits at 1 kbps,*

this slow machine caused a chain reaction upstream

1ˢᵗ anecdote

*"…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."*

1ˢᵗ anecdote

fail-stop
fail-partial
fail-transient

Slow
hardware!

"…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."

1ˢᵗ anecdote

fail-stop
fail-partial
fail-transient

Slow hardware!

"…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."

Cascading impact!

More anecdotes? **All** hardware?

More anecdotes? **All** hardware?

❑ **Disk** throughput dropped to 100 KB/s due to vibration

More anecdotes? **All** hardware?

❑ **Disk** throughput dropped to 100 KB/s due to vibration

❑ **SSDs** stalled for seconds due to firmware bugs

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

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*

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]

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.

Believe it?

Fail-slow hardware is "not" real. It is rare.

Yes, it's real!

Believe it?

Evidences from …

Institution	#Nodes
Company 1	>10,000
Company 2	150
Company 3	100
Company 4	>1,000
Company 5	>10,000

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

Table 2: **Operational scale.**

Evidences from …

Institution	#Nodes
Company 1	>10,000
Company 2	150
Company 3	100
Company 4	>1,000
Company 5	>10,000

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

Table 2: **Operational scale.**

Fail-slow at **scale**

Data and Methodology

❑ **101 reports**

Data and Methodology

- ❏ **101 reports**
  - ▪ Unformatted text
  - ▪ Written by engineers and operators (who still remember the incidents)
  - ▪ 2000-2017 (mostly after 2010)

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

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)

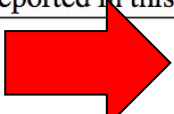| Important Findings and Observations |
|---|
| §3.1 **Varying root causes:** Fail-slow hardware can be induced by internal causes such as firmware bugs or device errors/wear-outs as well as external factors such as configuration, environment, temperature, and power issues. |
| §3.2 **Faults convert 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.**

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.**

Summary of findings

① Varying root causes

- Internal causes: firmware bugs, device errors

① Varying root causes

- Internal causes: firmware bugs, device errors
- External causes: temperature, power, environment, and configuration

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

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

Summary of findings

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

Summary of findings

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

Summary of findings

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)

① **Varying root causes**

① **Varying root causes**

Internal root causes

External root causes

	Hardware types					
Root	**SSD**	**Disk**	**Mem**	**Net**	**CPU**	**Total**
Device errors	10	8	9	10	3	40
Firmware bugs	6	3	0	9	2	20
Temperature	1	3	0	2	5	11
Power	1	0	1	0	6	8
Environment	3	5	2	4	4	18
Configuration	1	1	0	2	3	7
Unknown	0	3	1	2	2	8

① **Varying root causes**

- **Internal**
  - **Device errors/wearouts**

① **Varying root causes**

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

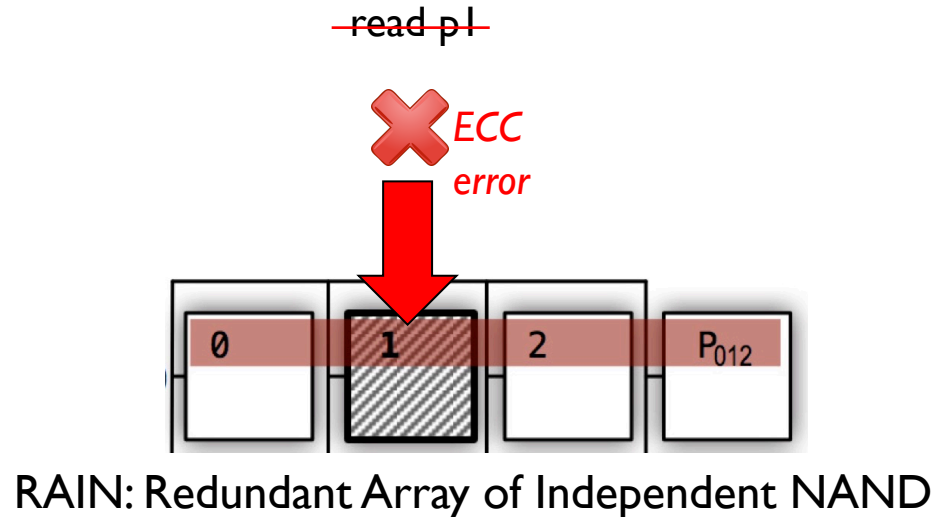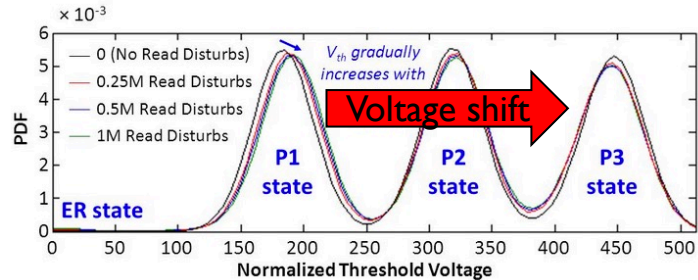# ① Varying root causes

- ## Internal
  - ### Device errors/wearouts
    - Ex: SSD read disturb/retry + page reconstruction → longer latency and more load
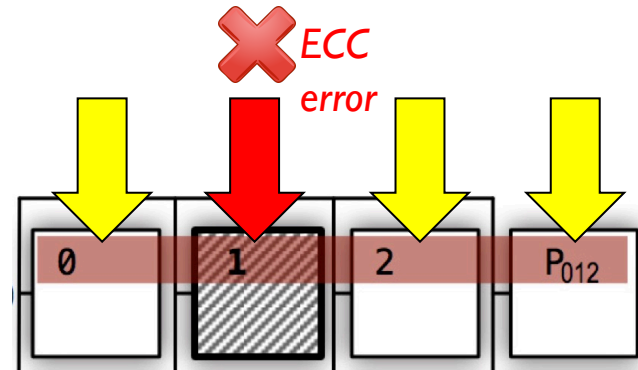
Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/
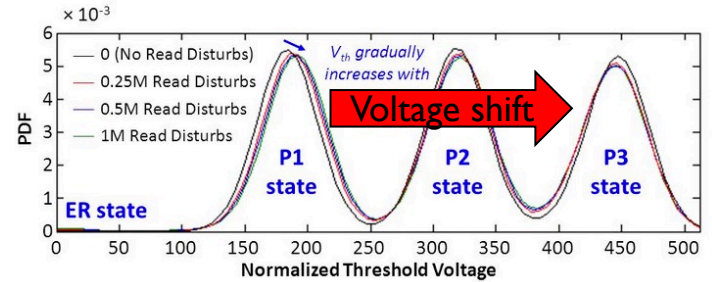
read(page X, Vth=v1)

① **Varying root causes**
- **Internal**
  - **Device errors/wearouts**
    - Ex: SSD read disturb/retry + page reconstruction → longer latency and more load

Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/
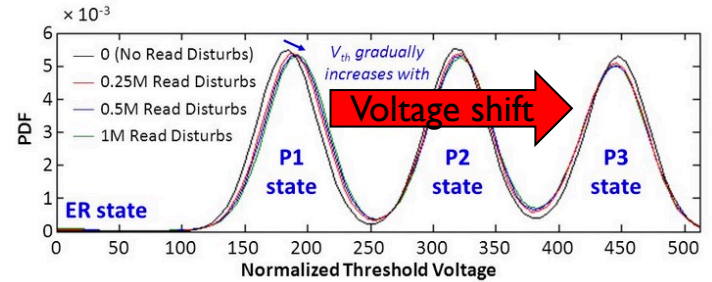
read(page X, **Vth**=v1)❌
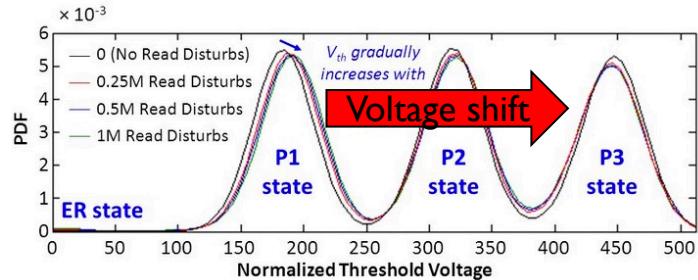read(page X, **Vth**=v2)

① **Varying root causes**

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

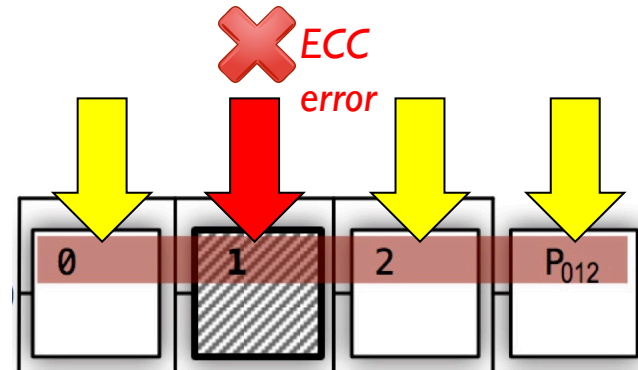Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/
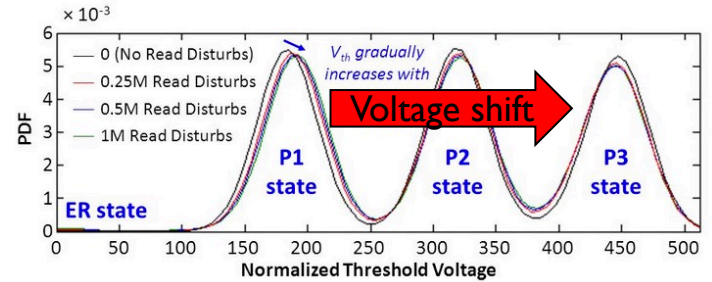
```
read(page X, Vth=v1)✖
read(page X, Vth=v2) ✖
read(page X, Vth=v3)
```
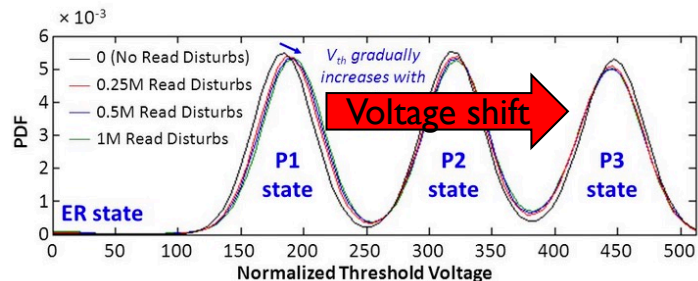
① **Varying root causes**

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

Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/

```
read(page X, Vth=v1)
read(page X, Vth=v2)
read(page X, Vth=v3)
read(page X, Vth=v4)
```

① **Varying root causes**
- ## **Internal**
  - ### **Device errors/wearouts**
    - Ex: SSD read disturb/retry + page reconstruction → longer latency and more load

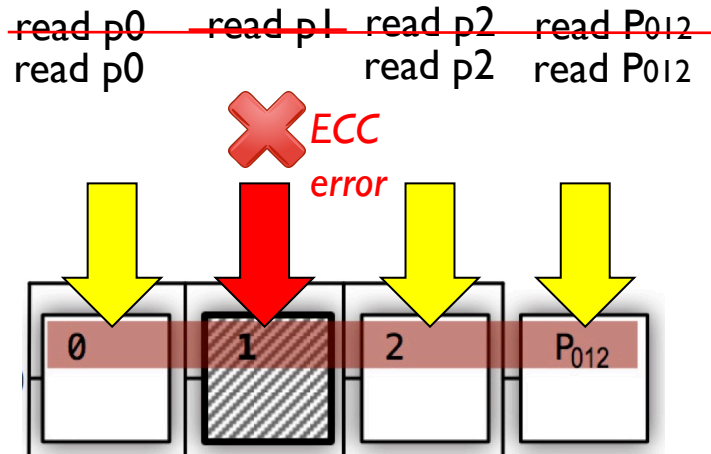Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/
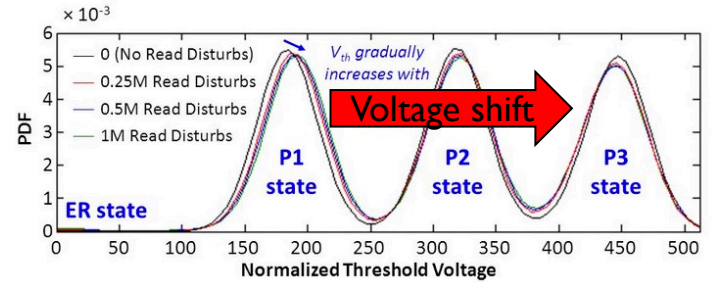
```
read(page X, Vth=v1) ✖
read(page X, Vth=v2) ✖
read(page X, Vth=v3) ✖
read(page X, Vth=v4) ✔
```
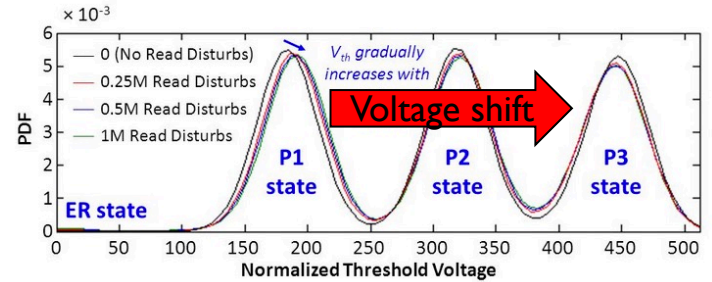4x slower!

① **Varying root causes**

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

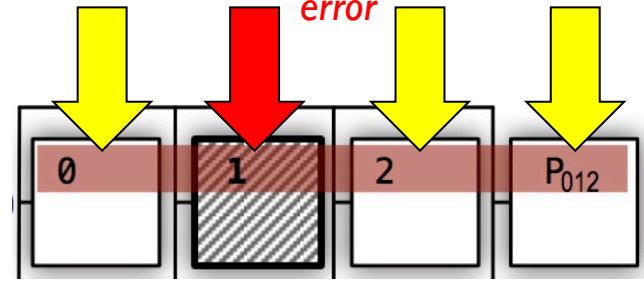Read Disturb Effect on V_{th} Distribution

$\times 10^{-3}$

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

V_{th} gradually increases with

Voltage shift

ER state   P1 state   P2 state   P3 state

PDF

Normalized Threshold Voltage

Picture from http://slideplayer.com/slide/10095910/

```
read(page X, Vth=v1) ✗
read(page X, Vth=v2) ✗
read(page X, Vth=v3) ✗
read(page X, Vth=v4) ✓
```

4x slower!

0   1   2 P_{012}

RAIN: Redundant Array of Independent NAND

① Varying root causes

- **Internal**
  - **Device errors/wearouts**
    - Ex: SSD read disturb/retry + page reconstruction → longer latency and more load
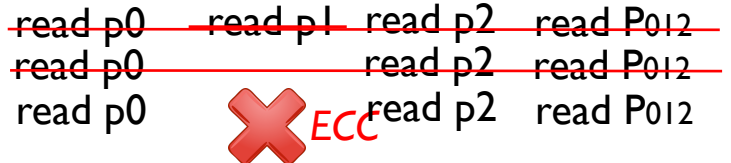
Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/
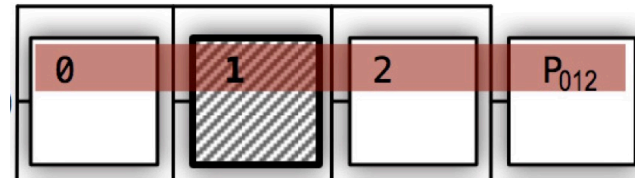
```
read(page X, Vth=v1)  ✗
read(page X, Vth=v2)  ✗✗
read(page X, Vth=v3)  ✗✗
read(page X, Vth=v4)  ✓
```

4x slower!

read p1

RAIN: Redundant Array of Independent NAND

① Varying root causes

- ## Internal
  - ### Device errors/wearouts
    - Ex: SSD read disturb/retry + page reconstruction → longer latency and more load

Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/

```
read(page X, Vth=v1) ✗
read(page X, Vth=v2) ✗
read(page X, Vth=v3) ✗
read(page X, Vth=v4) ✓
```

4x slower!

read p1

✗ ECC error

RAIN: Redundant Array of Independent NAND

① **Varying root causes**

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

Read Disturb Effect on V_{th} Distribution

Voltage shift

Picture from http://slideplayer.com/slide/10095910/

```
read(page X, Vth=v1)  ❌
read(page X, Vth=v2)  ❌❌
read(page X, Vth=v3)  ❌❌
read(page X, Vth=v4)  ✅
```

4x slower!

$$1 = XOR\ (0,\ 2,\ P_{012})$$

read p0   ~~read p1~~   read p2   read P_{012}

ECC error

| 0 | 1 | 2 | P_{012} |

RAIN: Redundant Array of Independent NAND

① **Varying root causes**

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

Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/

```
read(page X, Vth=v1) ✖
read(page X, Vth=v2) ✖
read(page X, Vth=v3) ✖
read(page X, Vth=v4) ✔
```
4x slower!

$$1 = XOR (0, 2, P_{012})$$

read p0   ~~read p1~~   read p2   read P_{012}

ECC error

RAIN: Redundant Array of Independent NAND

① Varying root causes

- ## Internal

  - ### Device errors/wearouts

    - Ex: SSD read disturb/retry + page reconstruction → longer latency and more load

Read Disturb Effect on V_{th} Distribution

V_{th} gradually increases with

Voltage shift

Picture from http://slideplayer.com/slide/10095910/

read(page X, **Vth**=v1) ✗
read(page X, **Vth**=v2) ✗✗
read(page X, **Vth**=v3) ✗✗✗
read(page X, **Vth**=v4) ✓

4x slower!

read retries!

$1 = XOR(0, 2, P_{012})$

~~read p0~~   ~~read p1~~   read p2   read P_{012}
read p0   read p2   read P_{012}

✗ ECC error

| 0 | 1 | 2 | P_{012} |

RAIN: Redundant Array of Independent NAND

① **Varying root causes**

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

Read Disturb Effect on V_{th} Distribution

Picture from http://slideplayer.com/slide/10095910/

read(page X, **Vth**=v1) ✗
read(page X, **Vth**=v2) ✗✗
read(page X, **Vth**=v3) ✗✗✗
read(page X, **Vth**=v4) ✓

4x slower!

read retries!

$1 = XOR (0, 2, P_{012})$

~~read p0~~   ~~read p1~~   read p2   read P_{012}
~~read p0~~   read p2   read P_{012}
read p0   ✗ *ECC error*   read p2   read P_{012}

| 0 | 1 | 2 | P_{012} |

RAIN: Redundant Array of Independent NAND

① **Varying root causes**

- **Internal**
  - Device errors
  - **Firmware bugs**
    - [No details, proprietary component]

① **Varying root causes**

- **Internal**
  - Device errors
  - **Firmware bugs**
    - [No details, proprietary component]
    - SSD firmware bugs throttled **µs** to **ms** read performance

① Varying root causes

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

https://www.anandtech.com/show/8550/samsung-acknowledges-the-ssd-840-evo-read-performance-bug-fix-is-on-the-way
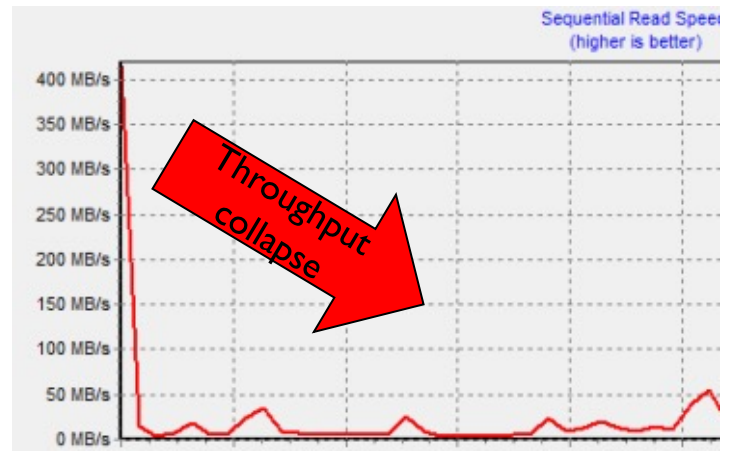
① Varying root causes

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

SSD	Disk	Memory	Network	Processors
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.	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.	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).	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.	Buggy BIOS firmware down-clocking CPUs; Other external causes such as hot temperature and lack of power.

① **Varying root causes**

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

① **Varying root causes**

- Internal [Device errors, firmware bugs]
- **External**
  - **Temperature**
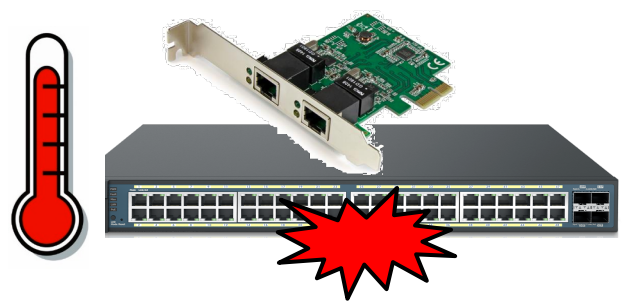
Hot temperature
→ Corrupt packets
→ Heavy TCP retransmission

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

① **Varying root causes**

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

Cold-air-under-the-floor system

Hot temperature
→ Corrupt packets
→ Heavy TCP retransmission

Faster SSD wearouts,
bad Vth →
more read retries

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

① **Varying root causes**

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

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

Hot temperature
→ Corrupt packets
→ Heavy TCP retransmission

Faster SSD wearouts, bad Vth → more read retries

Cold-air-under-the-floor system

① **Varying root causes**

- Internal [Device errors, firmware bugs]
- **External**
  - Temperature
  - **Power**
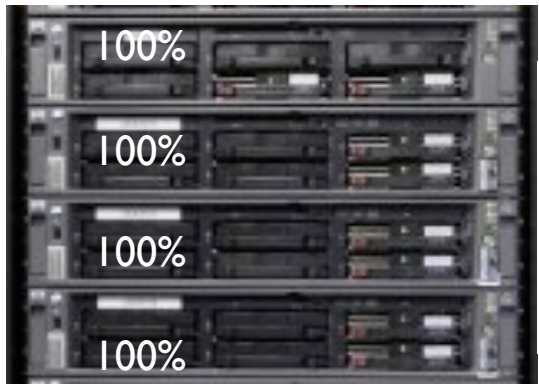
4 machines, 2 power supplies

① **Varying root causes**

- Internal [Device errors, firmware bugs]
- **External**
  - Temperature
  - **Power**
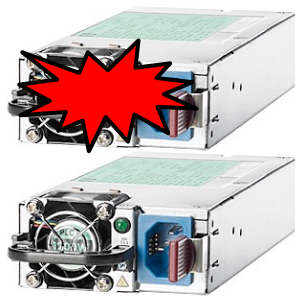
4 machines, 2 power supplies

① Varying root causes

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

4 machines, 2 power supplies

1 dead power → 50% CPU speed

① **Varying root causes**

- Internal [Device errors, firmware bugs]
- **External**
  - Temperature
  - **Power**
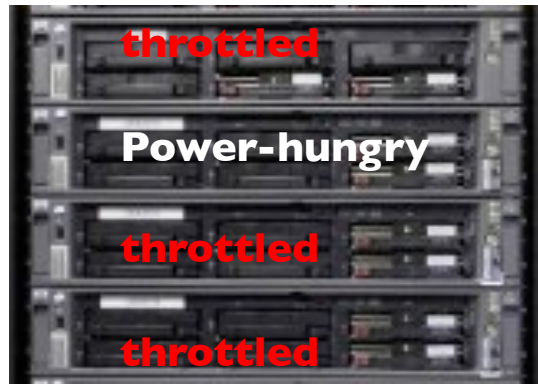
4 machines, 2 power supplies

1 dead power → 50% CPU speed

Power-hungry applications →
throttling neighboring CPUs

① **Varying root causes**

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

  - **Configuration**

① **Varying root causes**

- Internal [Device errors, firmware bugs]
- **External**
  - Temperature
  - Power
  - **Environment**
    - Altitude, pinched cables, etc.
  - **Configuration**

① **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
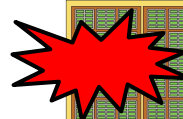   - Initialization code disabled processor cache

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

② **Faults convert**
  – **Fail-transient → fail-slow**
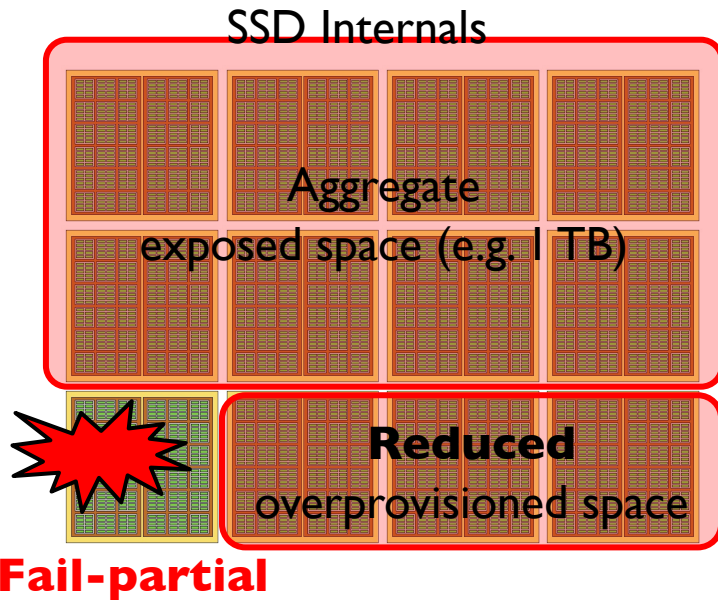
① Varying root causes Device errors, firmware, temperature, power, environment, configuration

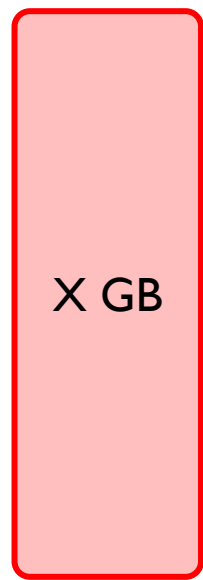② **Faults convert**
  – **Fail-transient → fail-slow**
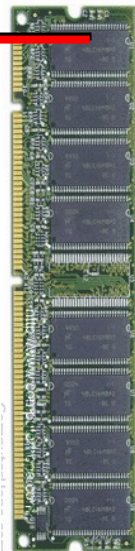
read ━━━

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

② **Faults convert**
  – **Fail-transient → fail-slow**

read ▬▬▬ Bit flips →
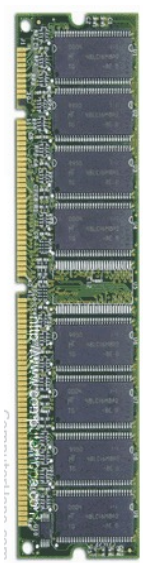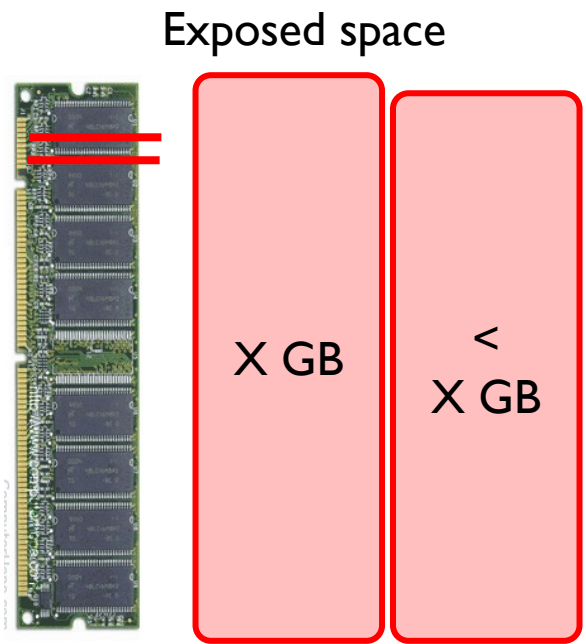ECC repair
(error masking)   Okay if rare

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

② **Faults convert**

– **Fail-transient → fail-slow**

read —
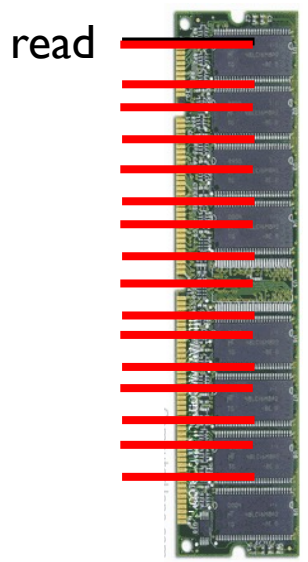
Bit flips →
ECC repair
(error masking)

Okay if rare

But, **frequent** errors
→ frequent error-masking/repair
→ repair latency becomes the *common* case

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

② **Faults convert**
  – Fail-transient → fail-slow
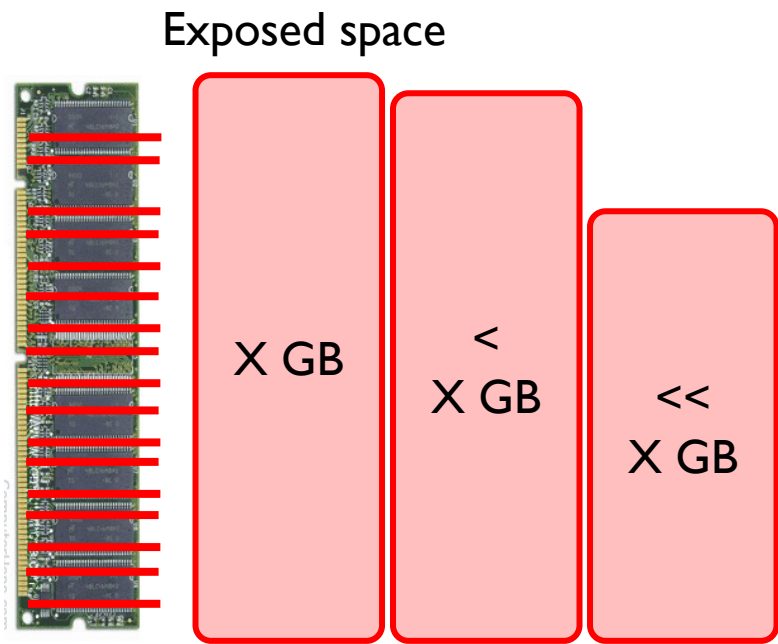 – **Fail-partial → fail-slow**

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

② **Faults convert**
  – Fail-transient → fail-slow
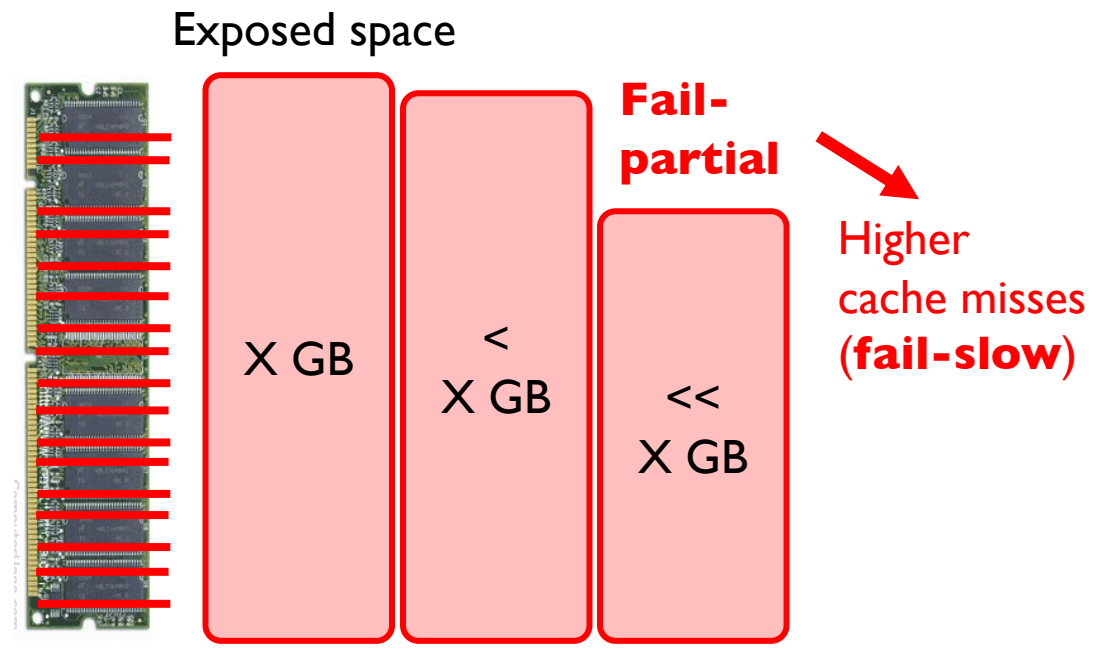 – **Fail-partial → fail-slow**

SSD Internals

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

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

SSD Internals

Aggregate
exposed space (e.g. 1 TB)

Overprovisioned space

Picture from https://flashdba.com/2014/06/20/understanding-
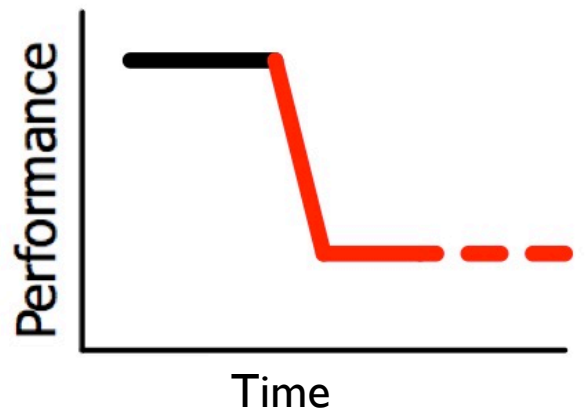flash-blocks-pages-and-program-erases/

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

② **Faults convert**

  – Fail-transient → fail-slow

  – **Fail-partial → fail-slow**

SSD Internals

Aggregate
exposed space (e.g. 1 TB)

"Not all chips are created equal"
(some chips die faster)

Reduced
overprovisioned space

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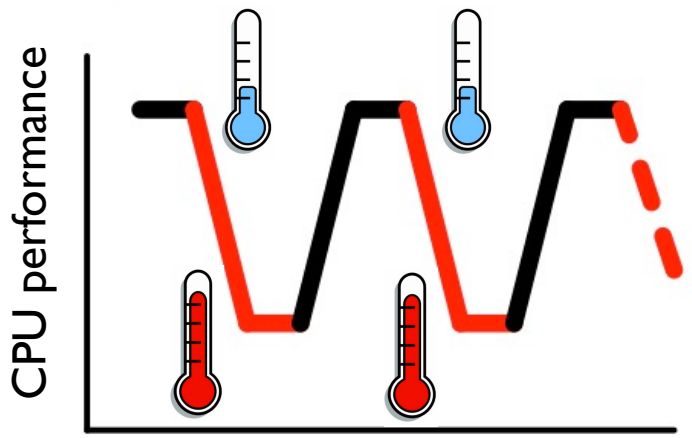
① Varying root causes <small>Device errors, firmware, temperature, power, environment, configuration</small>

② **Faults convert**

- – Fail-transient → fail-slow
- – **Fail-partial → fail-slow**

SSD Internals

"Not all chips are created equal"
(some chips die faster)

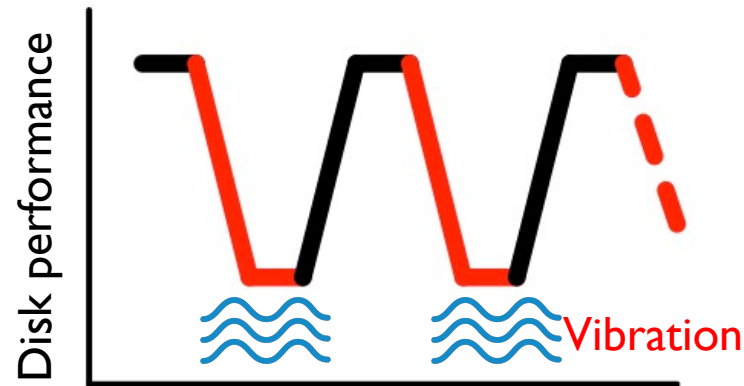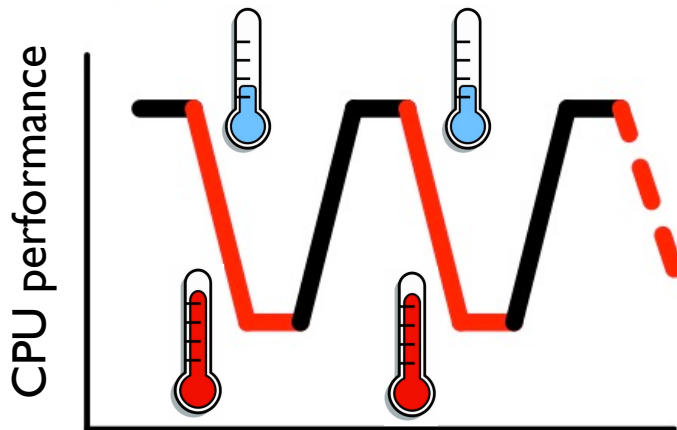→ Reduced overprovisioned space
→ *More frequent* GCs → **Slow** SSD

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

② **Faults convert**

  – Fail-transient → fail-slow

  – **Fail-partial → fail-slow**

SSD Internals

Aggregate
exposed space (e.g. 1 TB)

"Not all chips are created equal"
(some chips die faster)

Reduced
overprovisioned space

Fail-partial

→ Reduced overprovisioned space
→ *More frequent* GCs → **Slow** SSD

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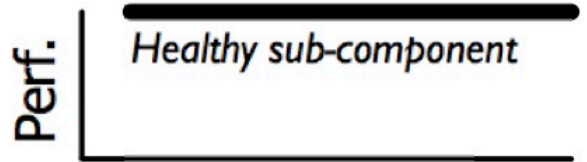
① Varying root causes Device errors, firmware, temperature, power, environment, configuration

② **Faults convert**

– Fail-transient → fail-slow

– **Fail-partial → fail-slow**

Exposed space

X GB
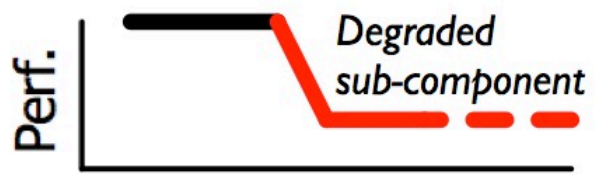
① Varying root causes Device errors, firmware, temperature, power, environment, configuration

② **Faults convert**

- Fail-transient → fail-slow
- **Fail-partial → fail-slow**

Exposed space

Custom memory chips that mask (hide) bad addresses

X GB

< X GB
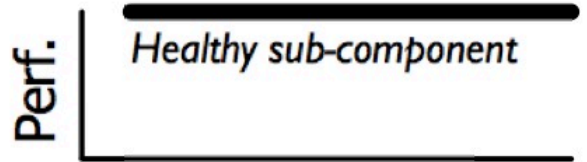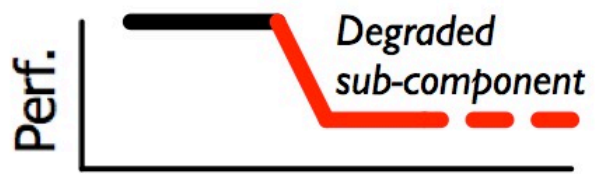
① Varying root causes Device errors, firmware, temperature, power, environment, configuration

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

Custom memory
chips that mask (hide)
bad addresses

Exposed space

X GB

<
X GB

<<
X GB

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

② **Faults convert**

- Fail-transient → fail-slow
- **Fail-partial → fail-slow**

Exposed space

Custom memory chips that mask (hide) bad addresses

X GB

< X GB

<< X GB

Fail-partial

Higher cache misses (**fail-slow**)

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

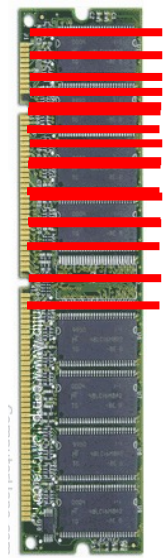② Faults convert Fail-stop, -transient, -partial → fail-slow

③ **Varying symptoms**

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

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

③**Varying symptoms**
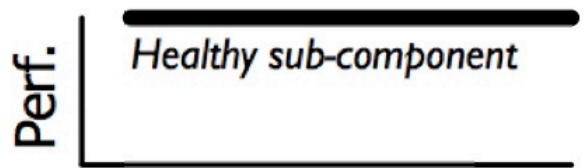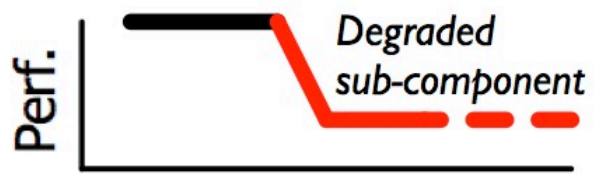   – **Permanent slowdown**

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**

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

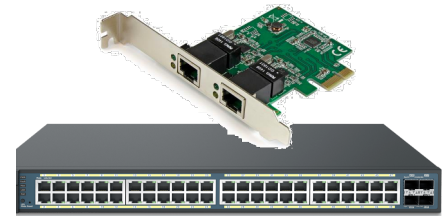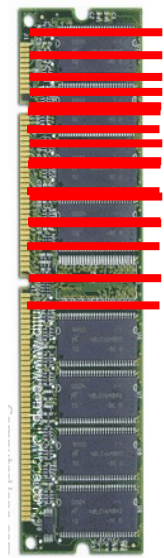② **Faults convert** Fail-stop, -transient, -partial → fail-slow

③ **Varying symptoms**

  – Permanent slowdown

  – **Transient slowdown**

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**

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
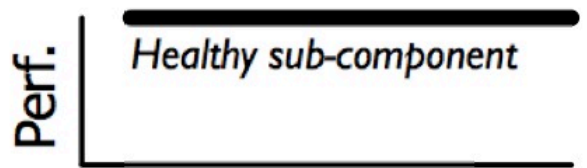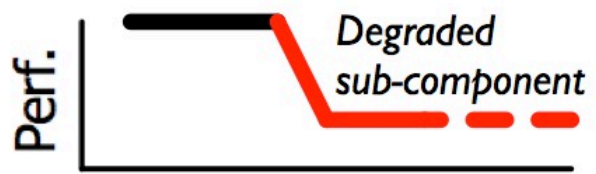 – **Partial slowdown**

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**

Slow reads (ECC repairs)

Fast reads

Degraded sub-component

Healthy sub-component

THE UNIVERSITY OF **CHICAGO**

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

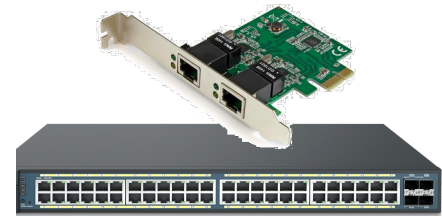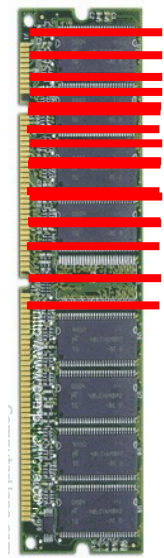② Faults convert Fail-stop, -transient, -partial → fail-slow

③ **Varying symptoms**

– Permanent slowdown
– Transient slowdown
– **Partial slowdown**

Degraded sub-component

Healthy sub-component

Slow reads (ECC repairs)

Fast reads

Small packets (fast)

>1500-byte packets (very slow)

THE UNIVERSITY OF CHICAGO

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**

Degraded sub-component

Healthy sub-component

Slow reads (ECC repairs)

Fast reads

Small packets (fast)
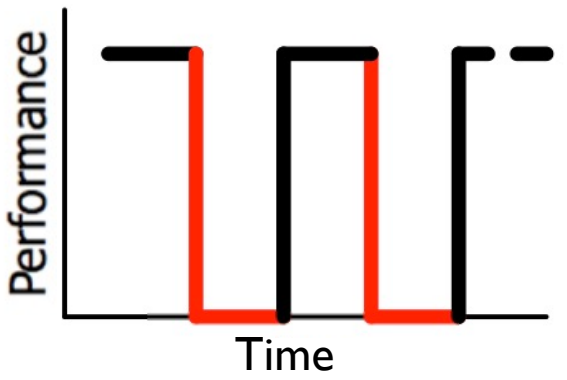
>1500-byte packets (very slow)

[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**

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
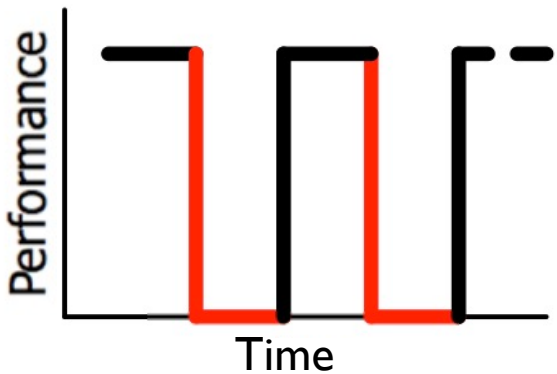
② Faults convert Fail-stop, -transient, -partial → fail-slow

③ **Varying symptoms**

- Permanent slowdown
- Transient slowdown
- Partial slowdown
- **Transient stop**

A bad batch of SSDs "*disappeared*" and then reappeared

Performance

Time

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*

Performance

Time

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*

Host Bus Adapter
recurrent resets

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

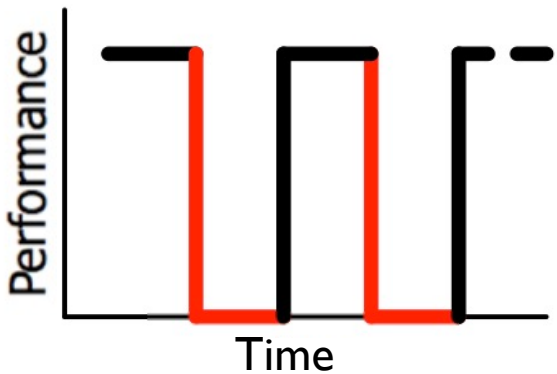② Faults convert Fail-stop, -transient, -partial → fail-slow

③ **Varying symptoms**

– Permanent slowdown

– Transient slowdown

– Partial slowdown

– **Transient stop**

A bad batch of SSDs "*disappeared*" and then reappeared

A firmware bug triggered hardware assertion failure

Host Bus Adapter *recurrent resets*

Uncorrectable bit flips in SRAM control paths

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**

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**

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

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

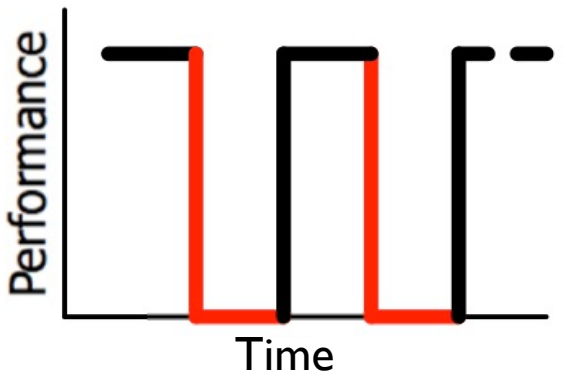③ Varying symptoms Permanent, transient, partial slowdown and transient stop

④ **Cascading nature**

  – Cascading **root causes**

Fans
normal
speed

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

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**

One died

Fans normal speed

Other fans *maximum* speed

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

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

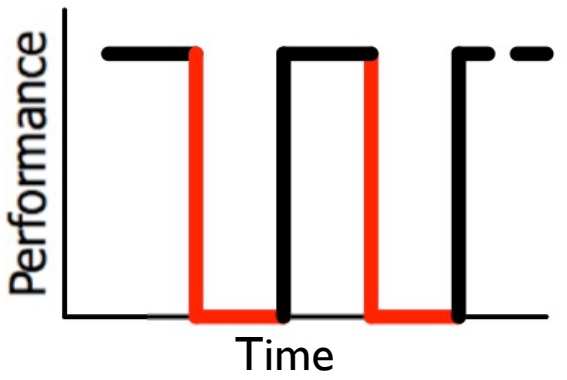③ Varying symptoms Permanent, transient, partial slowdown and transient stop

④ **Cascading nature**

– Cascading **root causes**

Fans normal speed

One died

Other fans *maximum* speed

Noise and vibration

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**

One died

Fans normal speed

Other fans *maximum* speed

Noise and vibration

Disk throughput collapses to **KB/s**

① Varying root causes <small>Device errors, firmware, temperature, power, environment, configuration</small>

② Faults convert <small>Fail-stop, -transient, -partial → fail-slow</small>

③ Varying symptoms <small>Permanent, transient, partial slowdown and transient stop</small>

④ **Cascading nature**

– Cascading **root causes**

Bad disks?
No!

Fans normal speed

One died

Other fans ***maximum*** speed

Noise and vibration

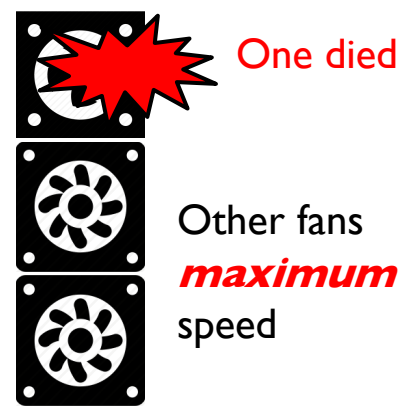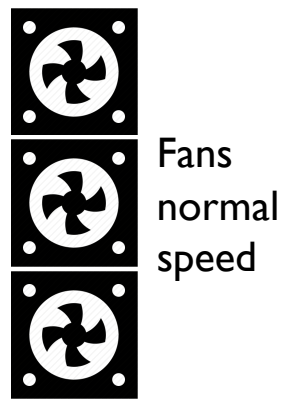Disk throughput collapses to **KB/s**

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
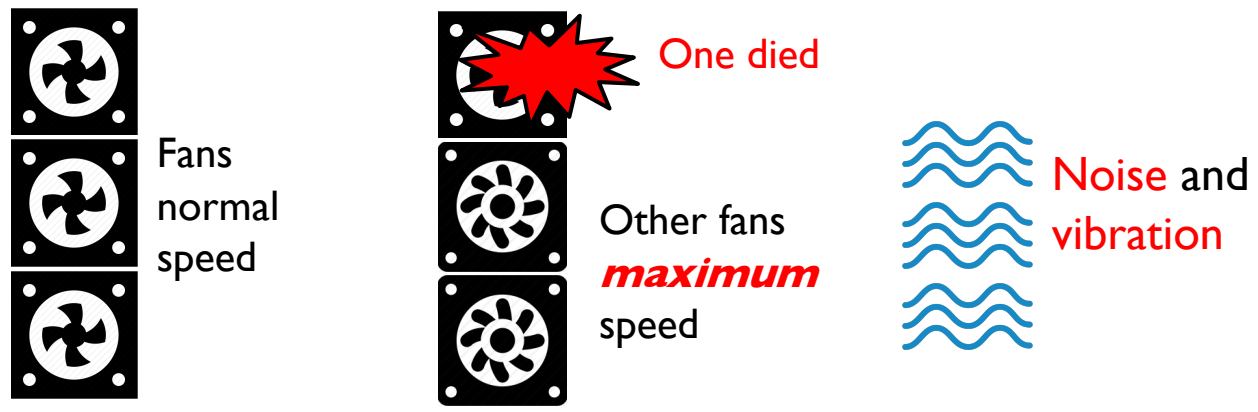
④ **Cascading nature**

– Cascading root causes
– Cascading **impacts** e.g. in Hadoop MapReduce

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

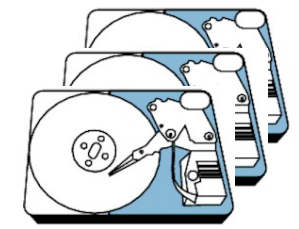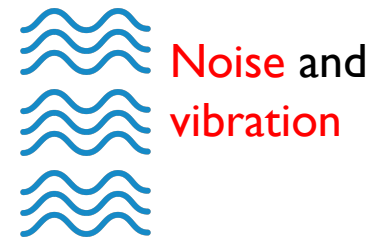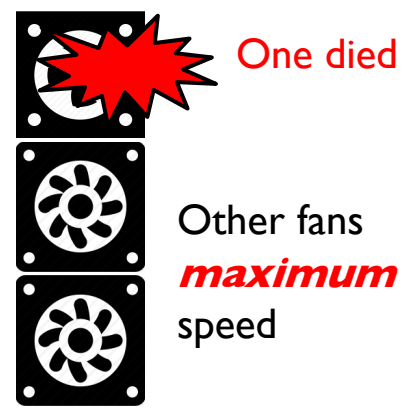*A **fast** map task*
(read locally)

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

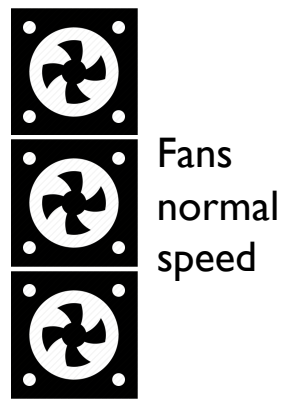– Cascading **impacts** e.g. in Hadoop MapReduce

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)
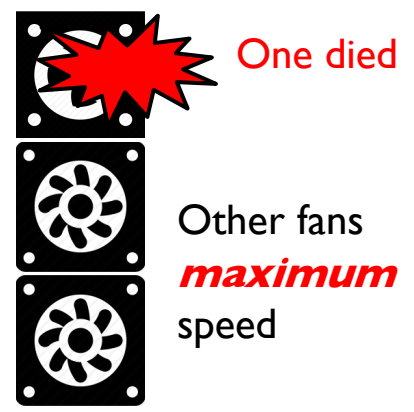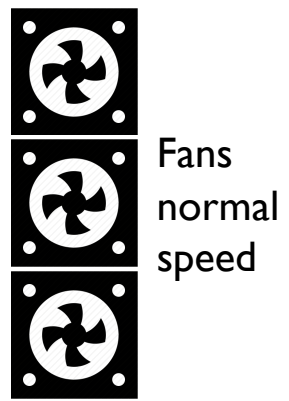

M1

R1   R3

R2

Slow
NIC

Shuffle

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)*

M1

R1   R3

R2

Slow NIC

Shuffle

Slow!!

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

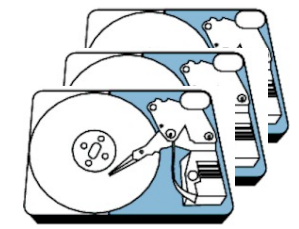② Faults convert Fail-stop, -transient, -partial → fail-slow

③ Varying symptoms Permanent, 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)*

M1

Slow NIC

Shuffle **Slow!!**

R1   **R3**

R2

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

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
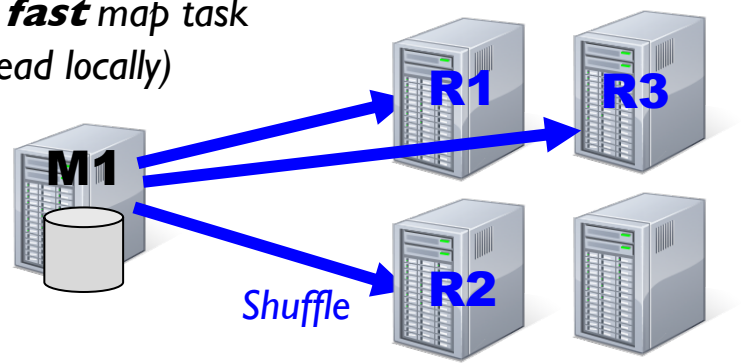
② Faults convert Fail-stop, -transient, -partial → fail-slow

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)

M1

R1   R3

R2

*Slow
NIC*

Shuffle
Slow!!

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

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

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
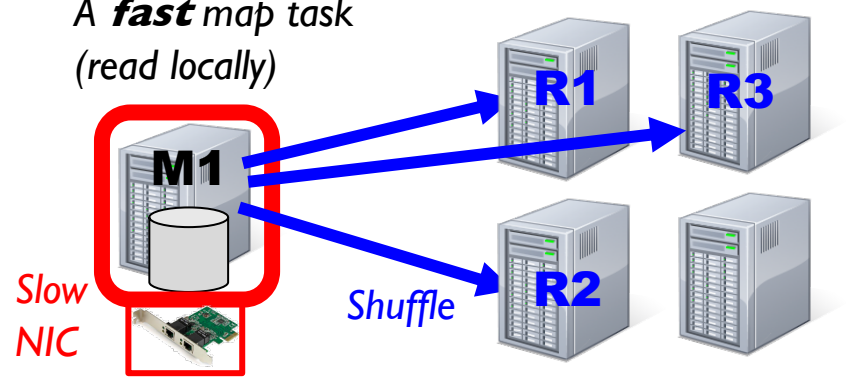
② Faults convert Fail-stop, -transient, -partial → fail-slow

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)

M1

Slow
NIC

Shuffle
Slow!!

R1   R3

R2

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**

THE UNIVERSITY OF
CHICAGO

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
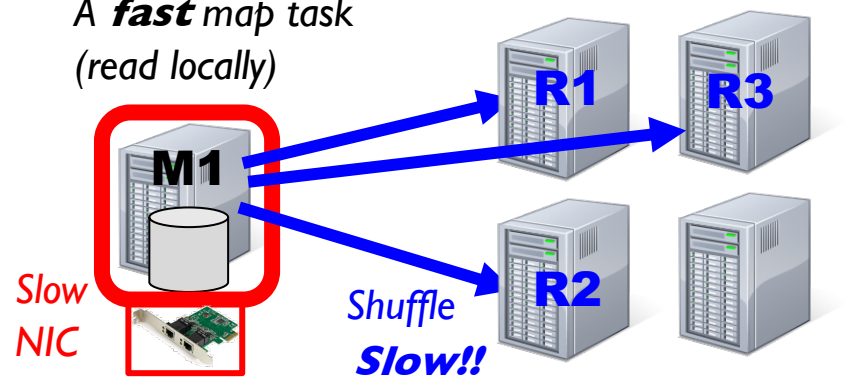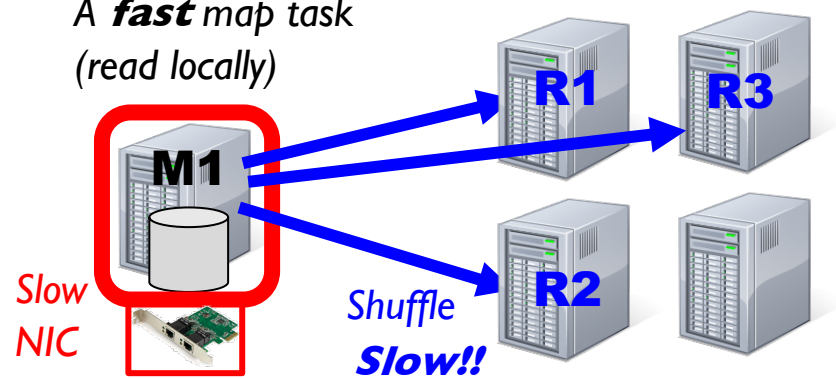
② Faults convert Fail-stop, -transient, -partial → fail-slow

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

④ **Cascading nature**

  – Cascading root causes

  – Cascading **impacts**

Facebook Hadoop Jobs, 30 nodes

~170 jobs/hour

[From PBSE @ SoCC '17]

① Varying root causes <small>Device errors, firmware, temperature, power, environment, configuration</small>

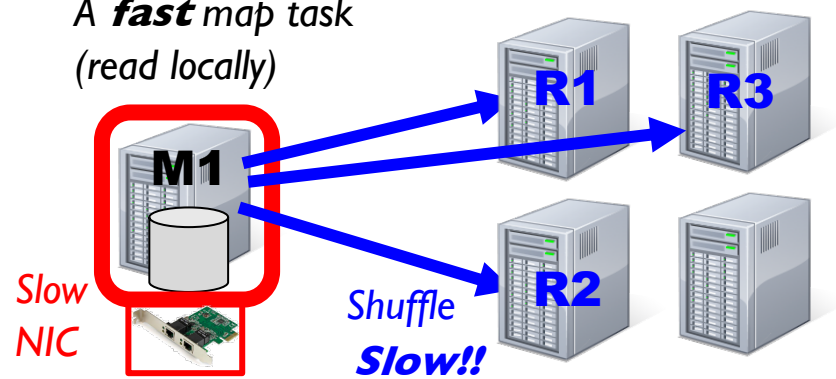② Faults convert <small>Fail-stop, -transient, -partial → fail-slow</small>

③ Varying symptoms <small>Permanent, transient, partial slowdown and transient stop</small>

④ **Cascading nature**

- Cascading root causes
- Cascading **impacts**

Facebook Hadoop Jobs, 30 nodes

~170 jobs/hour

1 job/hour !!!

[From PBSE @ SoCC '17]

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

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

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

④Cascading nature

⑤**Rare but deadly**

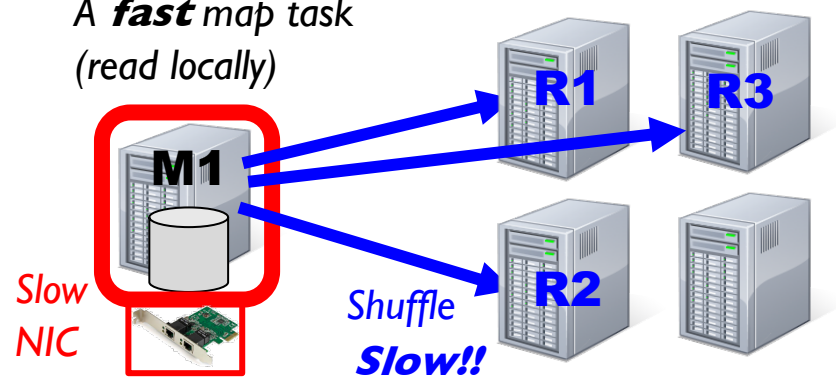① Varying root causes Device errors, firmware, temperature, power, environment, configuration

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)

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
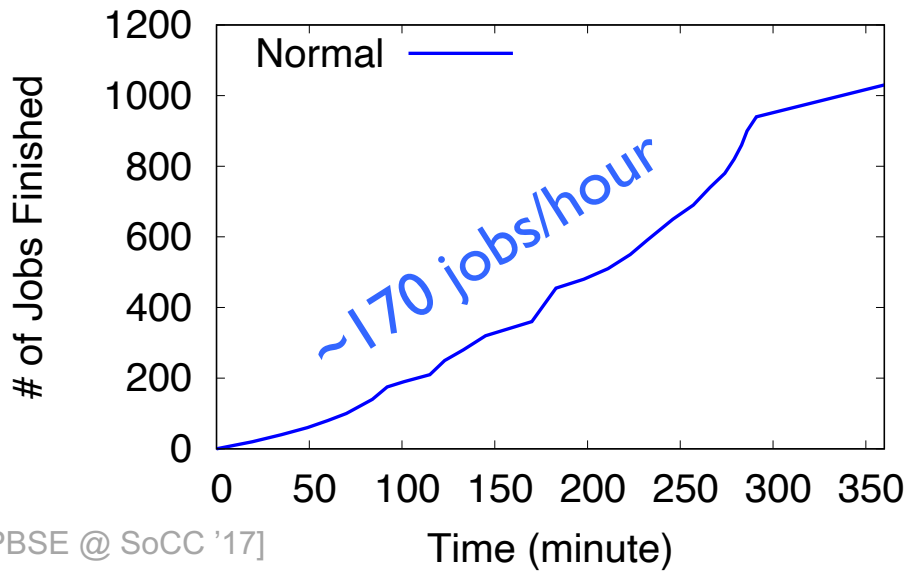
② Faults convert Fail-stop, -transient, -partial → fail-slow

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

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
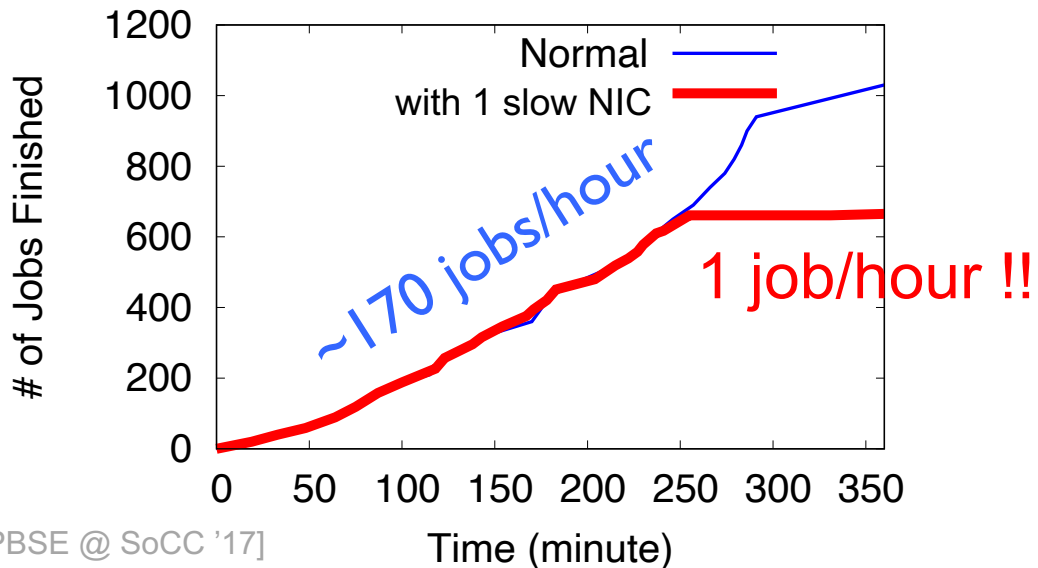
② Faults convert Fail-stop, -transient, -partial → fail-slow

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?*

① Varying root causes <small>Device errors, firmware, temperature, power, environment, configuration</small>

② Faults convert <small>Fail-stop, -transient, -partial → fail-slow</small>

③ Varying symptoms <small>Permanent, transient, partial slowdown and transient stop</small>

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

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.

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

Conclusion:

Modern, advanced systems

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

Conclusion:

Modern, advanced systems

+ Fail-slow hardware

① Varying root causes <small>Device errors, firmware, temperature, power, environment, configuration</small>

② Faults convert <small>Fail-stop, -transient, -partial → fail-slow</small>

③ Varying symptoms <small>Permanent, transient, partial slowdown and transient stop</small>

④ Cascading nature

⑤ 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

HW Type	Symptoms			
	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.**

Root	Symptoms			
	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.**

Operators

❑ Cannot use application bandwidth check (all are affected)
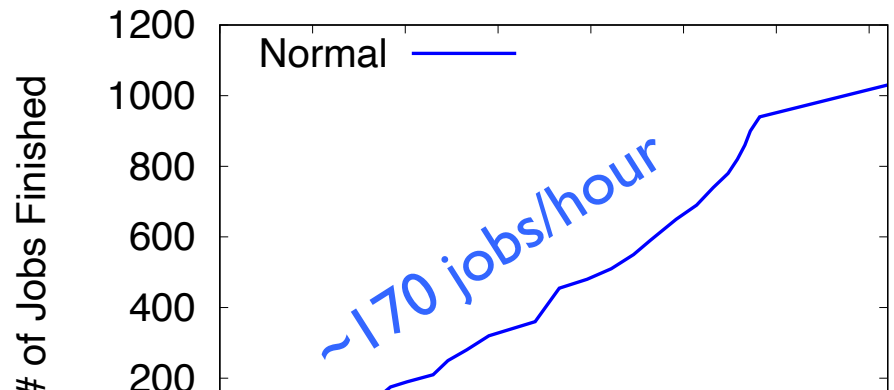
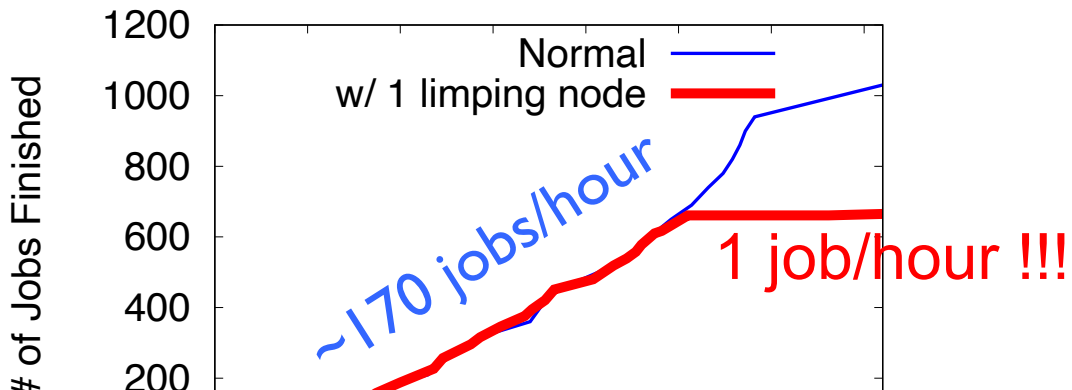**Facebook Hadoop Jobs, 30 nodes**

Operators

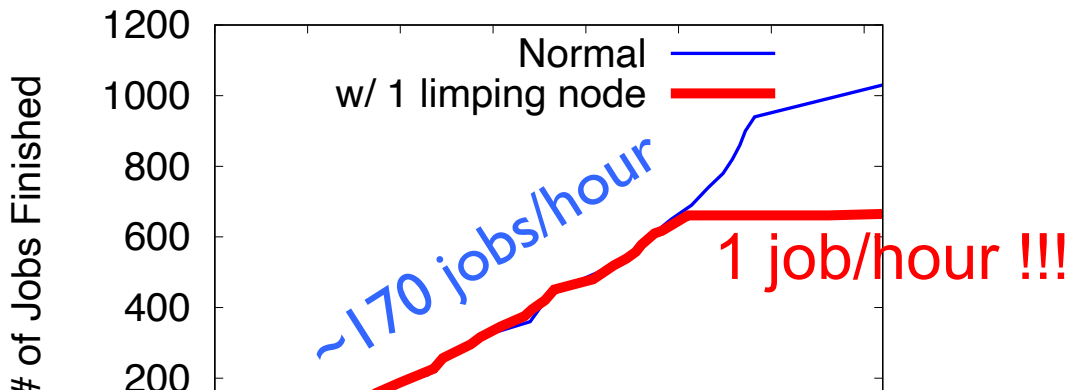❑ Cannot use application bandwidth check (all are affected)

Facebook Hadoop Jobs, 30 nodes

~170 jobs/hour

1 job/hour !!!

Operators

❑ Cannot use application bandwidth check (all are affected)

Hadoop, not fully tail/limpware tolerant??

Facebook Hadoop Jobs, 30 nodes

~170 jobs/hour

1 job/hour !!!

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
② **Faults convert**

① Varying root causes Device errors, firmware, temperature, power, environment, configuration
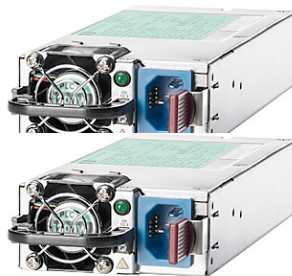
② **Faults convert**
- **Fail-stop → fail-slow**

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

② **Faults convert**
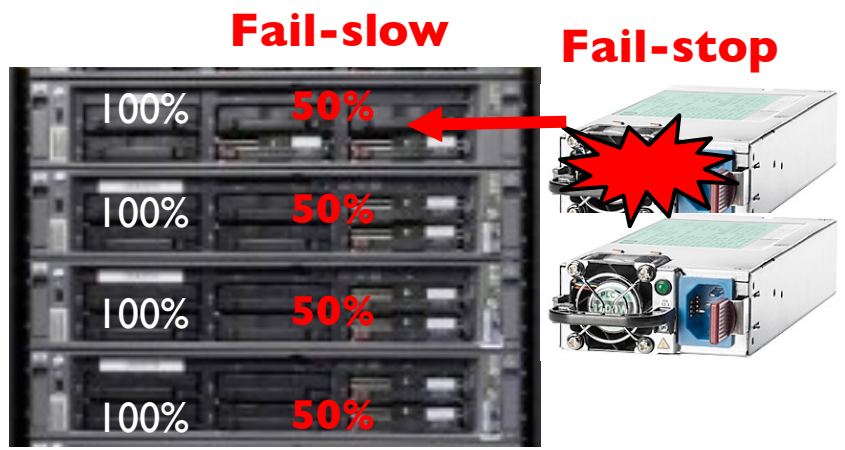
- **Fail-stop → fail-slow**
  - Fail-stop power → fail-slow CPUs

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

② **Faults convert**

- **Fail-stop → fail-slow**
  - Fail-stop power → fail-slow CPUs

Fail-slow   **Fail-stop**

100%   50%
100%   50%
100%   50%
100%   50%

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

② **Faults convert**

- **Fail-stop → fail-slow**
  - Fail-stop power → fail-slow CPUs
  - Fail-stop disk → fail-slow RAID

Fail-slow   **Fail-stop**