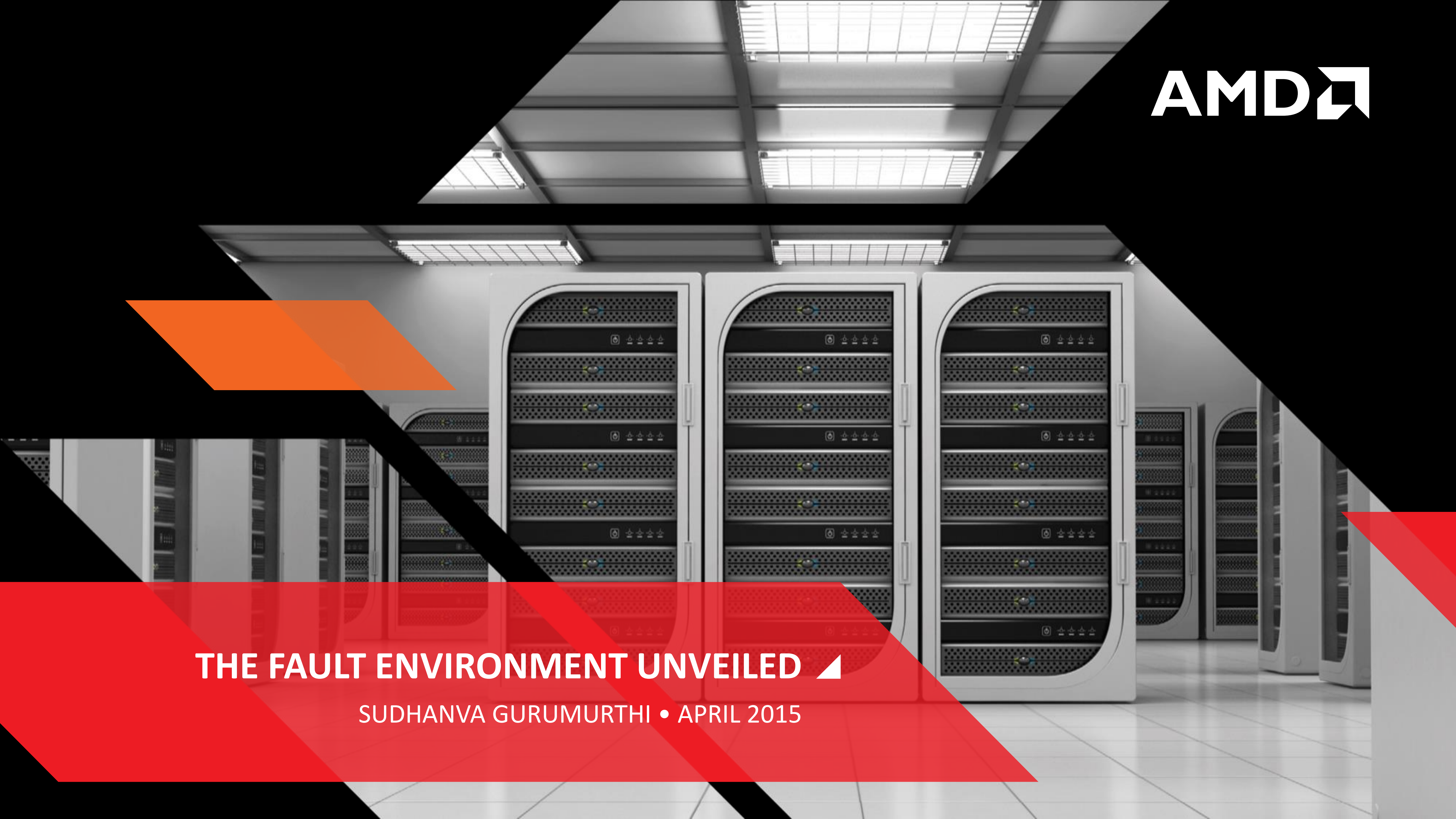


THE FAULT ENVIRONMENT UNVEILED

SUDHANVA GURUMURTHI • APRIL 2015

THE RELIABILITY LANDSCAPE

- Many sources of unreliability in silicon:
  - Particle-induced Transient Faults
  - Permanent Faults
  - Aging
  - Voltage Noise
  - Increased Variability
- Many emerging technologies have reliability problems
- Large systems with high component counts

THE RELIABILITY LANDSCAPE

- Many sources of unreliability in silicon:
  - Particle-induced Transient Faults
  - Permanent Faults
  - Aging
  - Voltage Noise
  - Increased Variability

- Many emerging technologies have reliability problems

- Large systems with high component counts

Need a deep understanding of faults

STUDY FAULTS FROM REAL SYSTEMS

- Many insights can be gained from field data analyses

- Field studies are beneficial for hardware designers, system integrators, and the operators of the system

- This talk:
  - A look into faults and failures from field studies of supercomputers and other large data centers
  - Implications for resiliency and reliability at scale

ANALYSIS OF FAULTS IN SUPERCOMPUTERS

[SRIDHARAN ET AL., ASPLOS’15] [DEBARDELEBEN ET AL., SELSE’14] [SRIDHARAN ET AL., SC’13]

- Collaboration between AMD and the US Department of Energy National Labs

- **Jaguar system at Oak Ridge National Lab**
  - 18,688x 2-socket 8-core AMD Opteron™ processor nodes
  - 8 DDR-2 DIMMs per node, chipkill ECC
  - 11 months of data
- **Cielo system at Los Alamos National Lab**
  - 8518x 2-socket 12-core AMD Opteron™ processor nodes
  - 8 DDR-3 DIMMs per node, chipkill ECC
  - 16 months of data
- **Hopper system at NERSC / Lawrence Berkeley National Labs**
  - 6000x 2-socket 12-core AMD Opteron™ processor nodes
  - 8 DDR-3 DIMMs per node, chipkill ECC
  - 18 months of data
- **Sufficient data to draw statistically significant conclusions**
  - 500M CPU socket-hours in aggregate
  - 40B DRAM device-hours in aggregate

IDENTIFYING FAULT TYPES IN THE FIELD

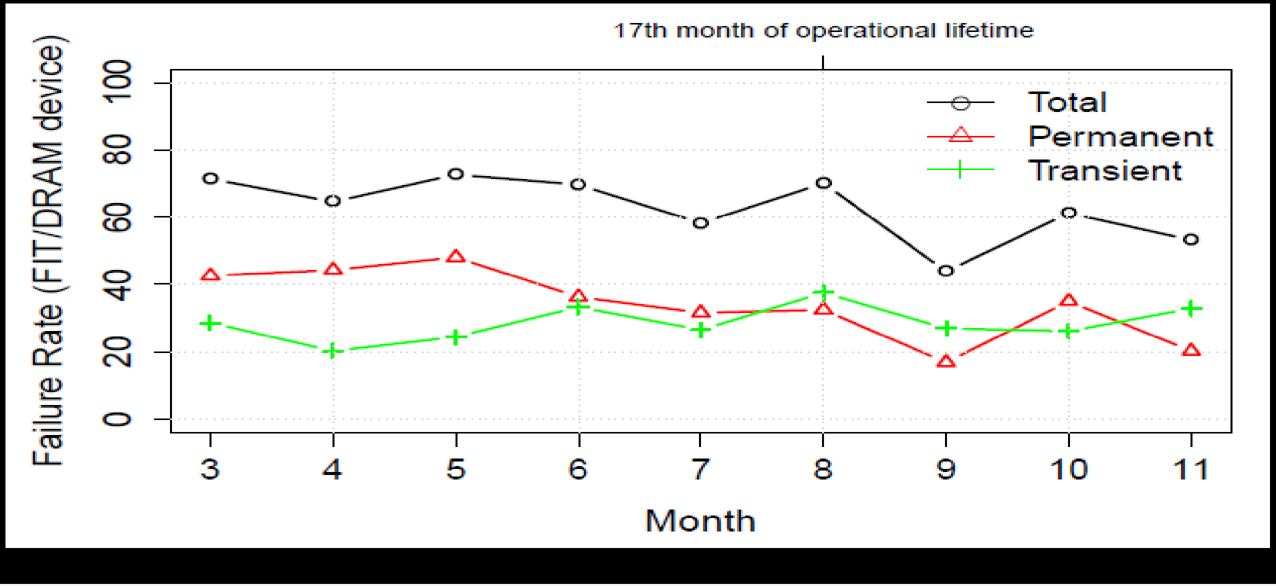
◢ **Data collection**

- Hardware logs errors in hardware error registers
- OS periodically samples error registers and logs corrected errors to the console
- Console log is a *sample* of all errors that occurred in the system
- Can infer fault type based on error log characteristics

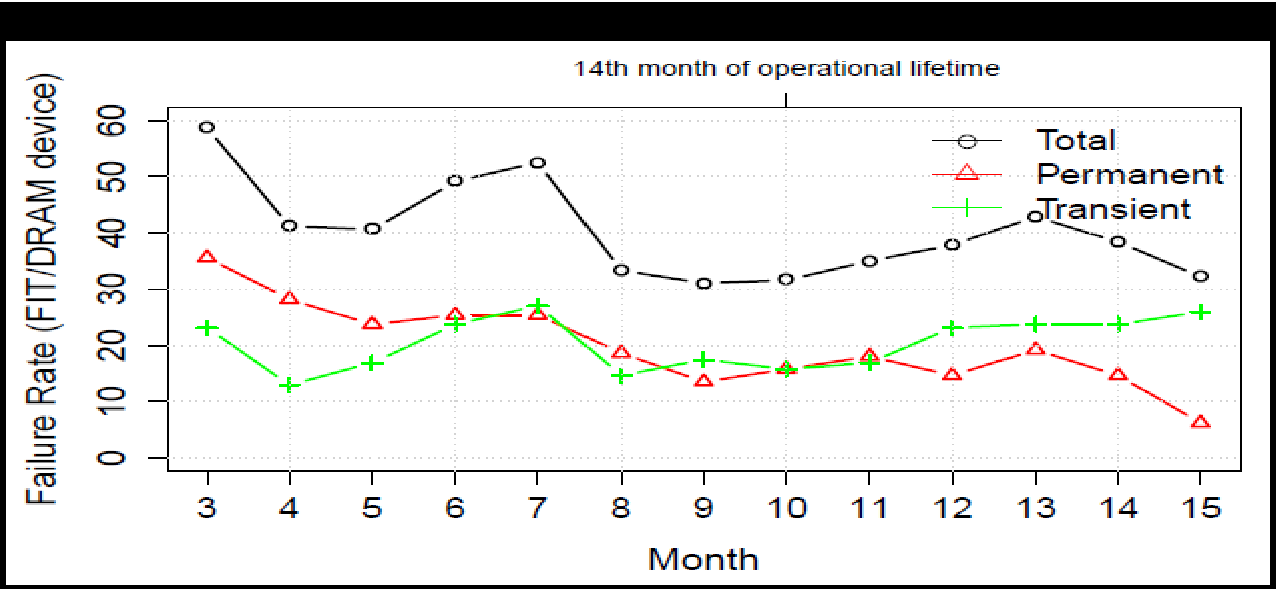
◢ **Scrubber (L2 and L3 caches, DRAM)**

- Periodically reads each memory location, corrects any errors found, writes corrected data back to memory
- Errors in multiple scrub intervals ➔ permanent fault
- Errors in one interval ➔ transient fault (bound)

DRAM: FAULT RATE AND FAULT TYPES



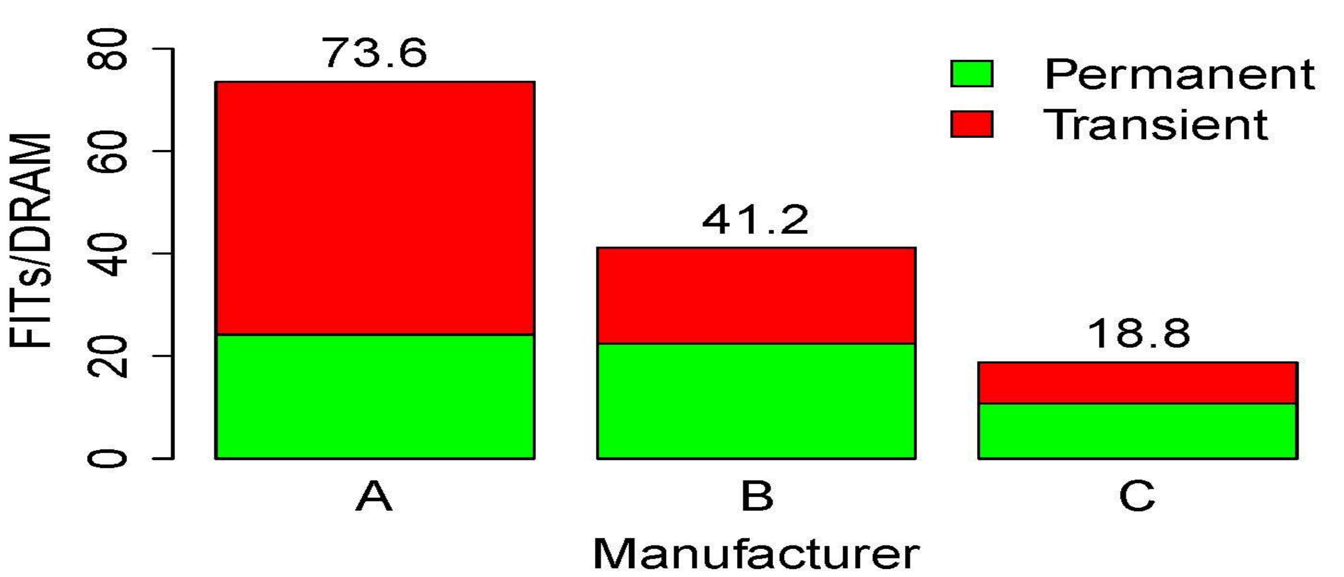
Jaguar (DDR-2)



Cielo (DDR-3)

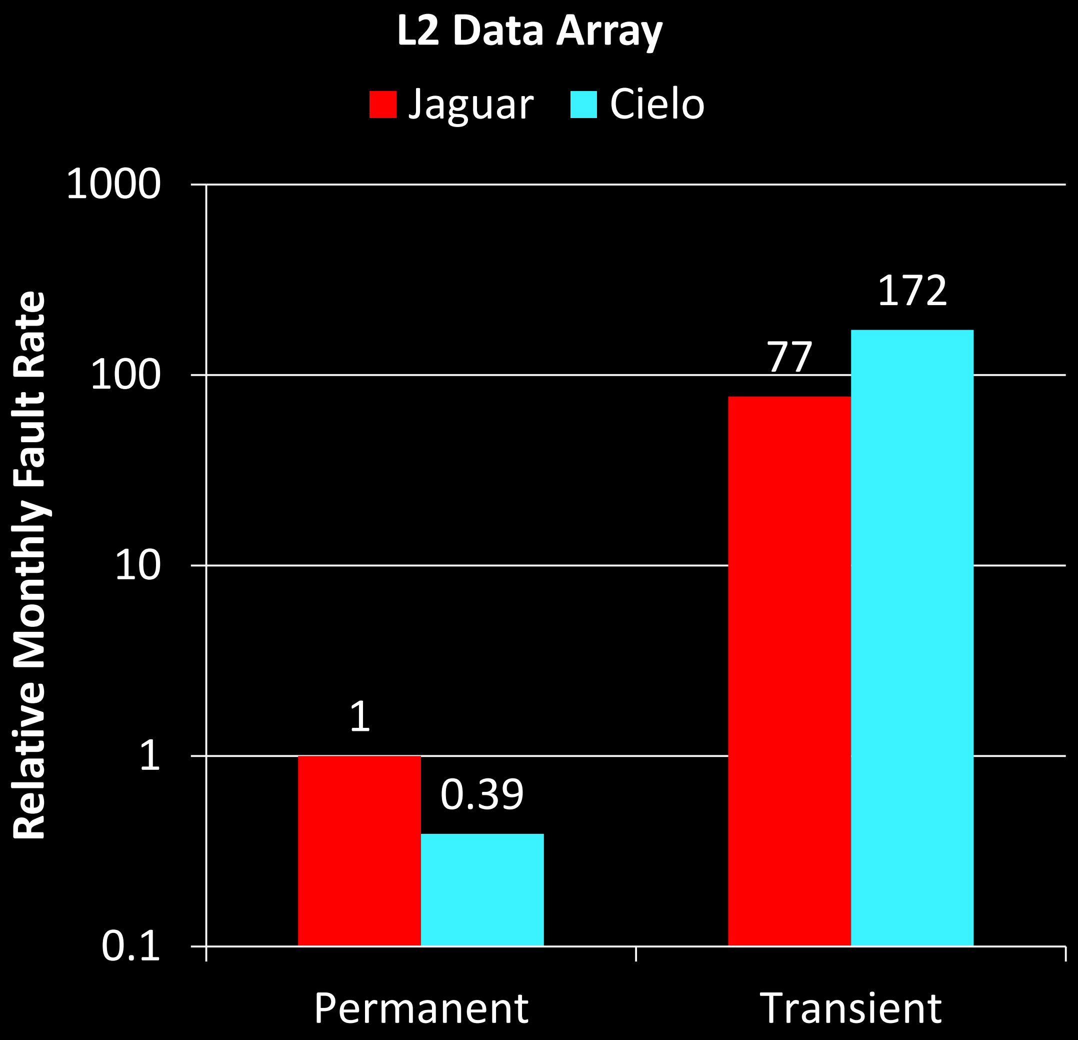
DRAM FAULT MODES

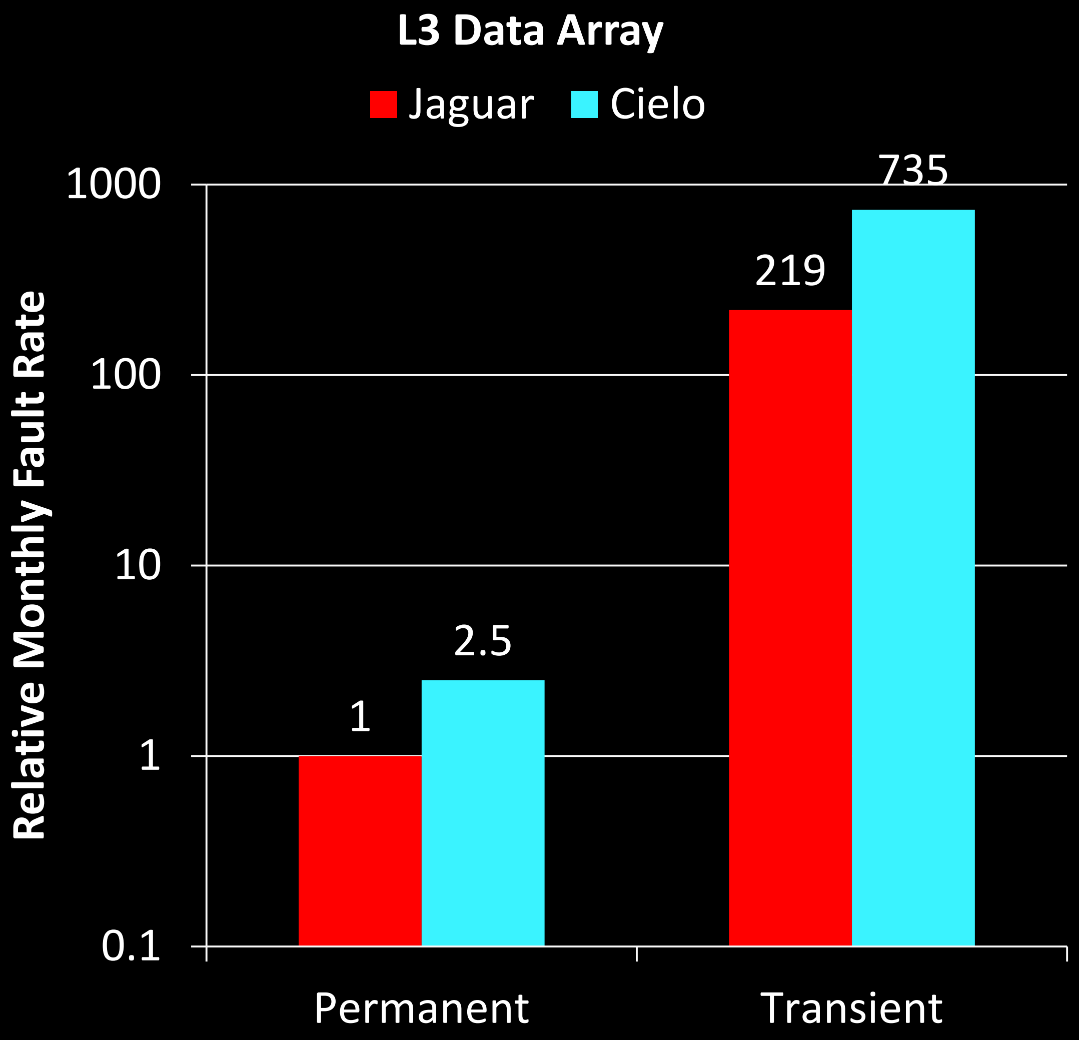
Fault Mode	Vendor A	Vendor B	Vendor C
Single-bit	64.6%	69.5%	58.4%
Single-word	0%	0.3%	0%
Single-column	8.7%	8.8%	11.9%
Single-row	12.2%	10.6%	14.9%
Single-bank	13.5%	7.8%	9.9%
Multiple-bank	1.3%	0.7%	2.0%
Multiple-rank	1.3%	3.0%	3.0%



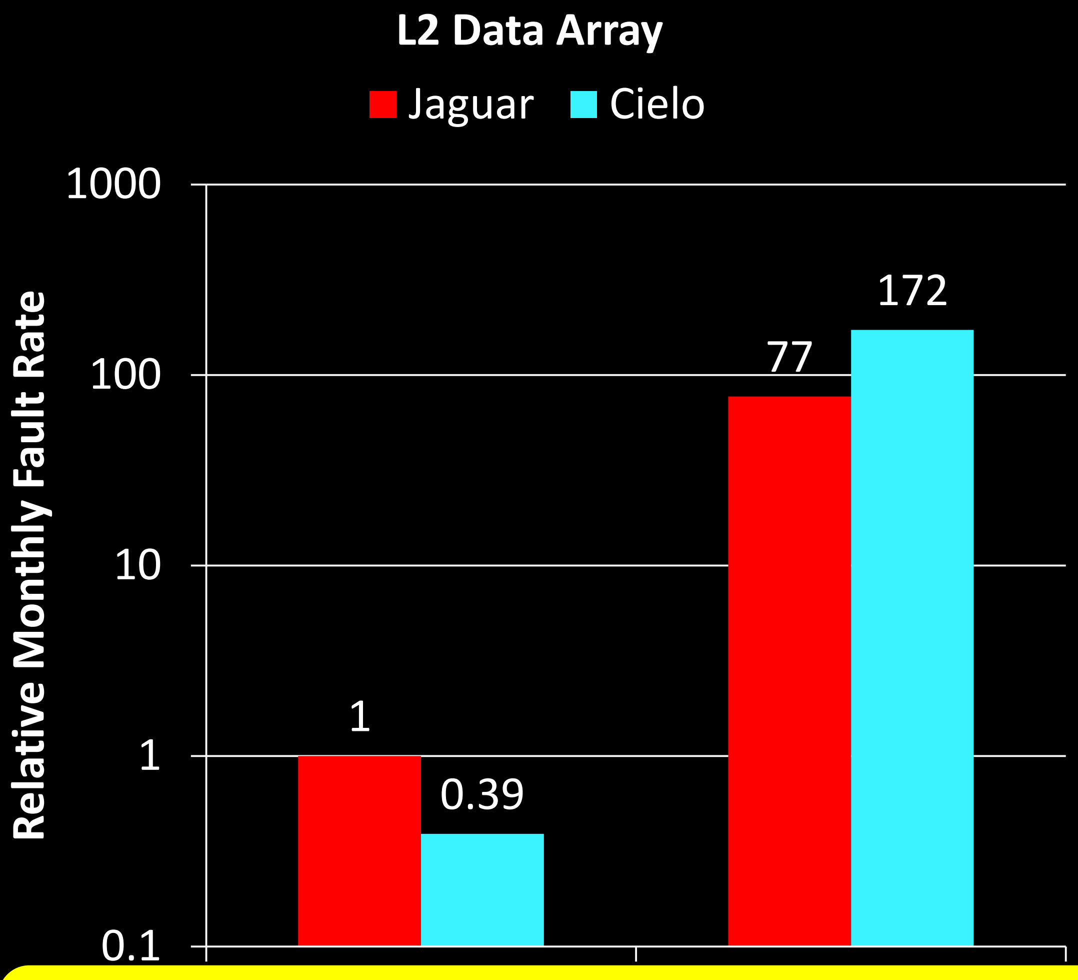
Overall fault rate per vendor

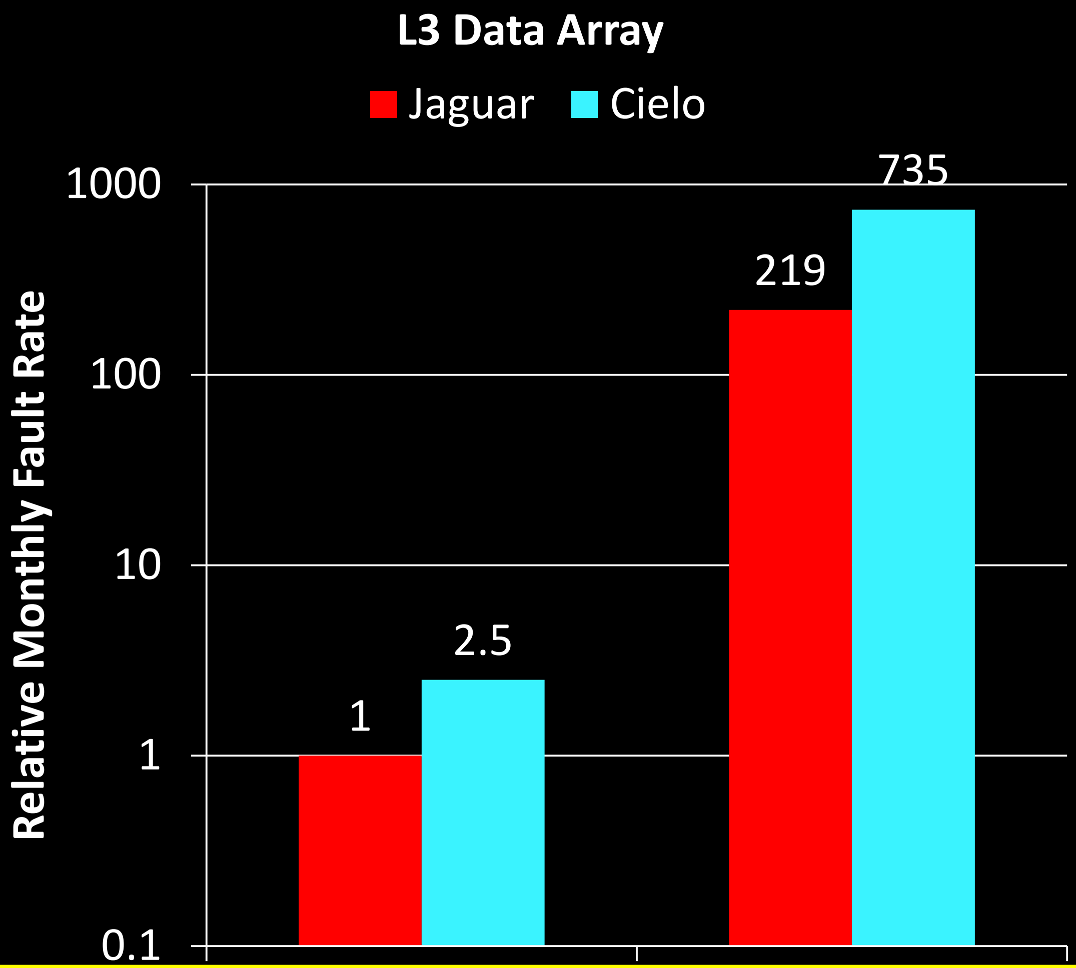
SRAM FAULTS


L2 Data Array
Jaguar
Cielo
Relative Monthly Fault Rate
1000
100
10
1
0.1
1
0.39
77
172
Permanent
Transient


L3 Data Array
Jaguar
Cielo
Relative Monthly Fault Rate
1000
100
10
1
0.1
1
2.5
219
735
Permanent
Transient

SRAM FAULTS

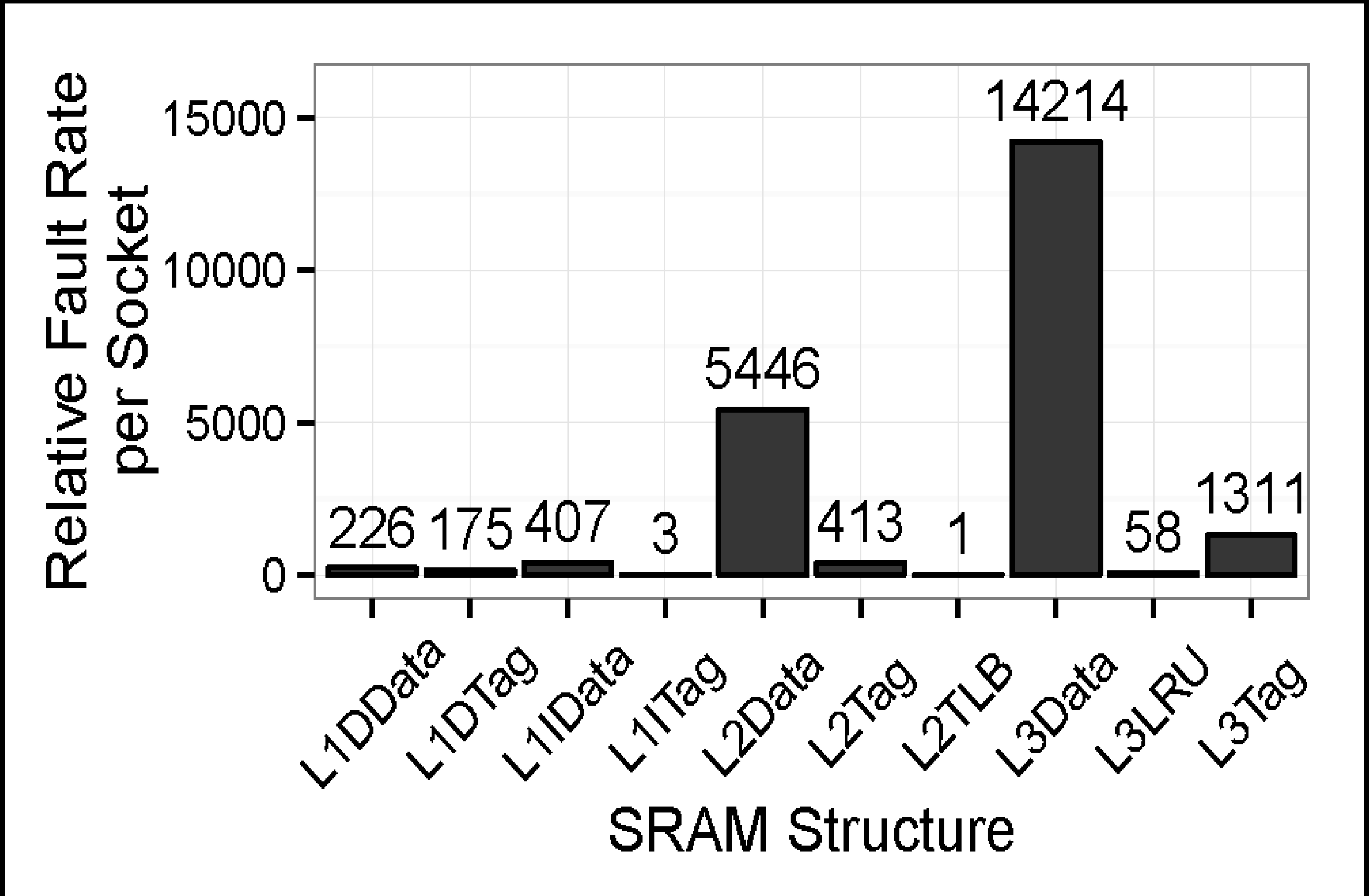




Most SRAM faults are transient, especially in mature process technologies

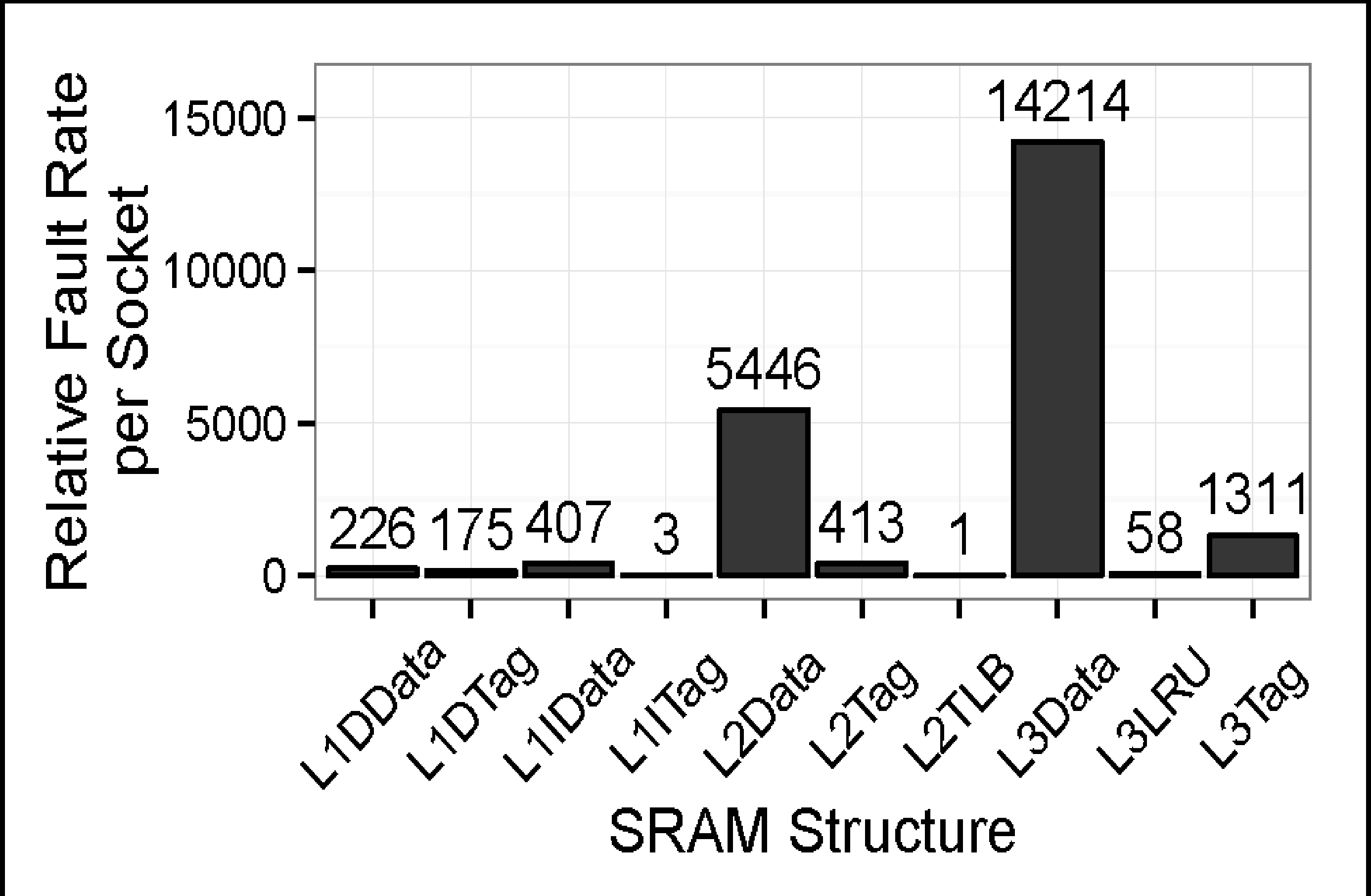
MANY SRAM STRUCTURES EXPERIENCE FAULTS

- Most faults are in L2 and L3 caches
  - Largest on-chip structures
- Even small structures see faults
  - TLB, tag arrays
- Exascale systems will:
  - Have 4-5x the number of compute sockets
  - Have much more SRAM per socket



MANY SRAM STRUCTURES EXPERIENCE FAULTS

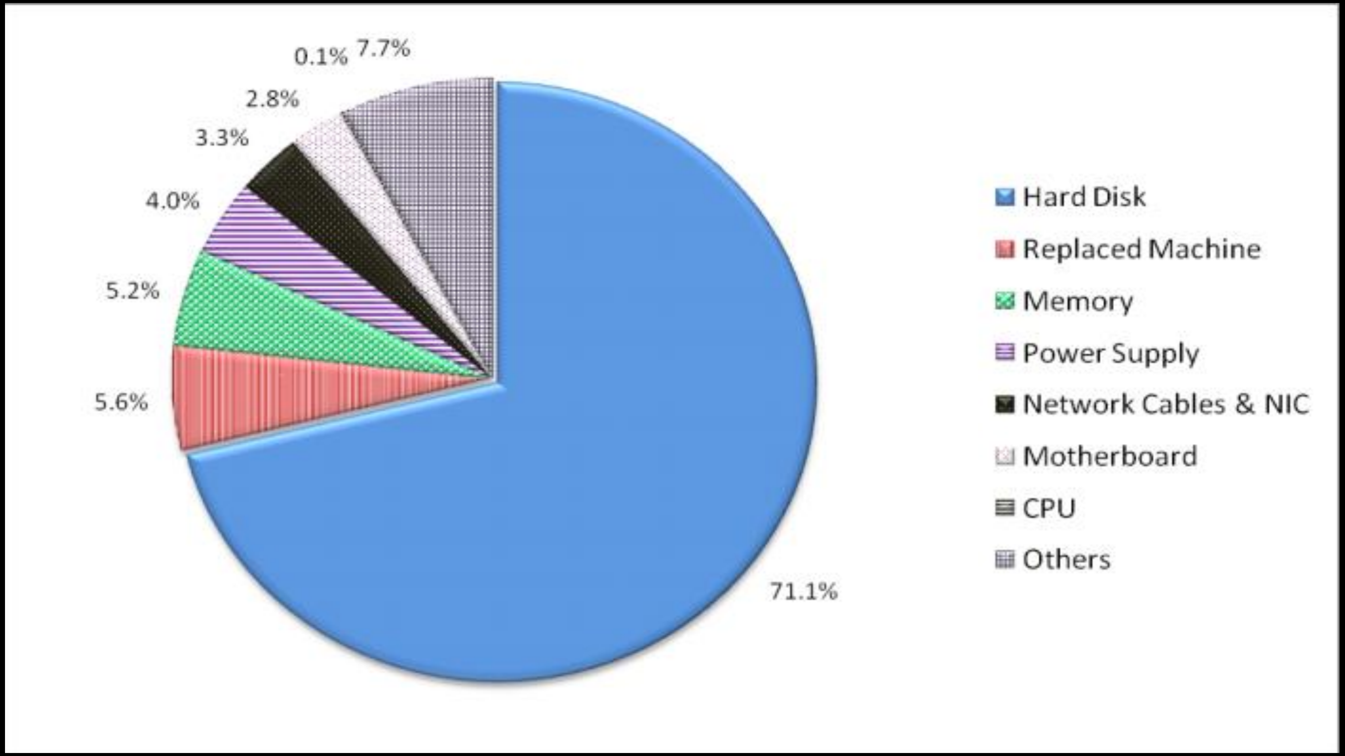
- Most faults are in L2 and L3 caches
  - Largest on-chip structures
- Even small structures see faults
  - TLB, tag arrays
- Exascale systems will:
  - Have 4-5x the number of compute sockets
  - Have much more SRAM per socket



Exascale systems will experience more faults!

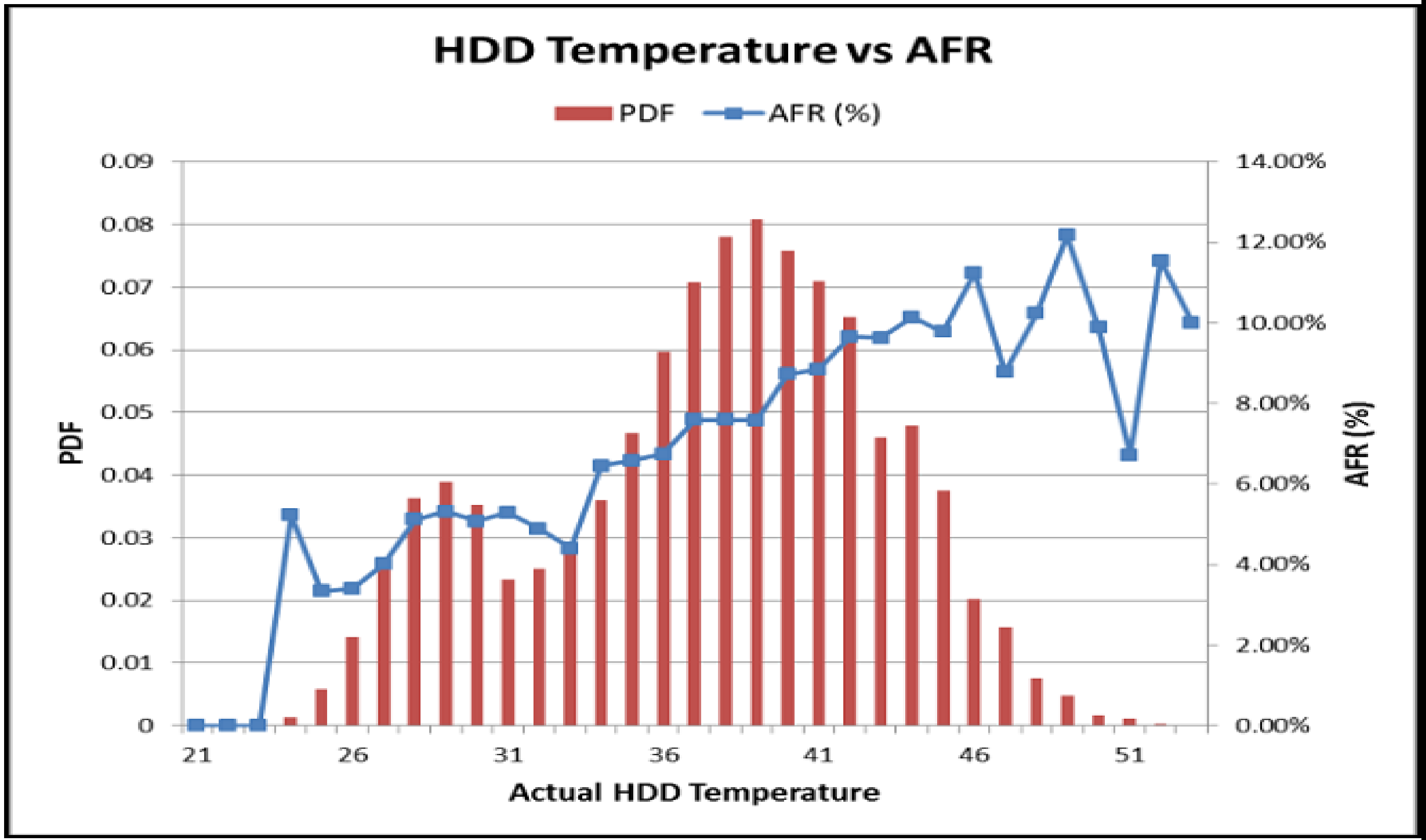
FAILURES IN OTHER NODE COMPONENTS

[SANKAR ET AL., ACM TOS'13]



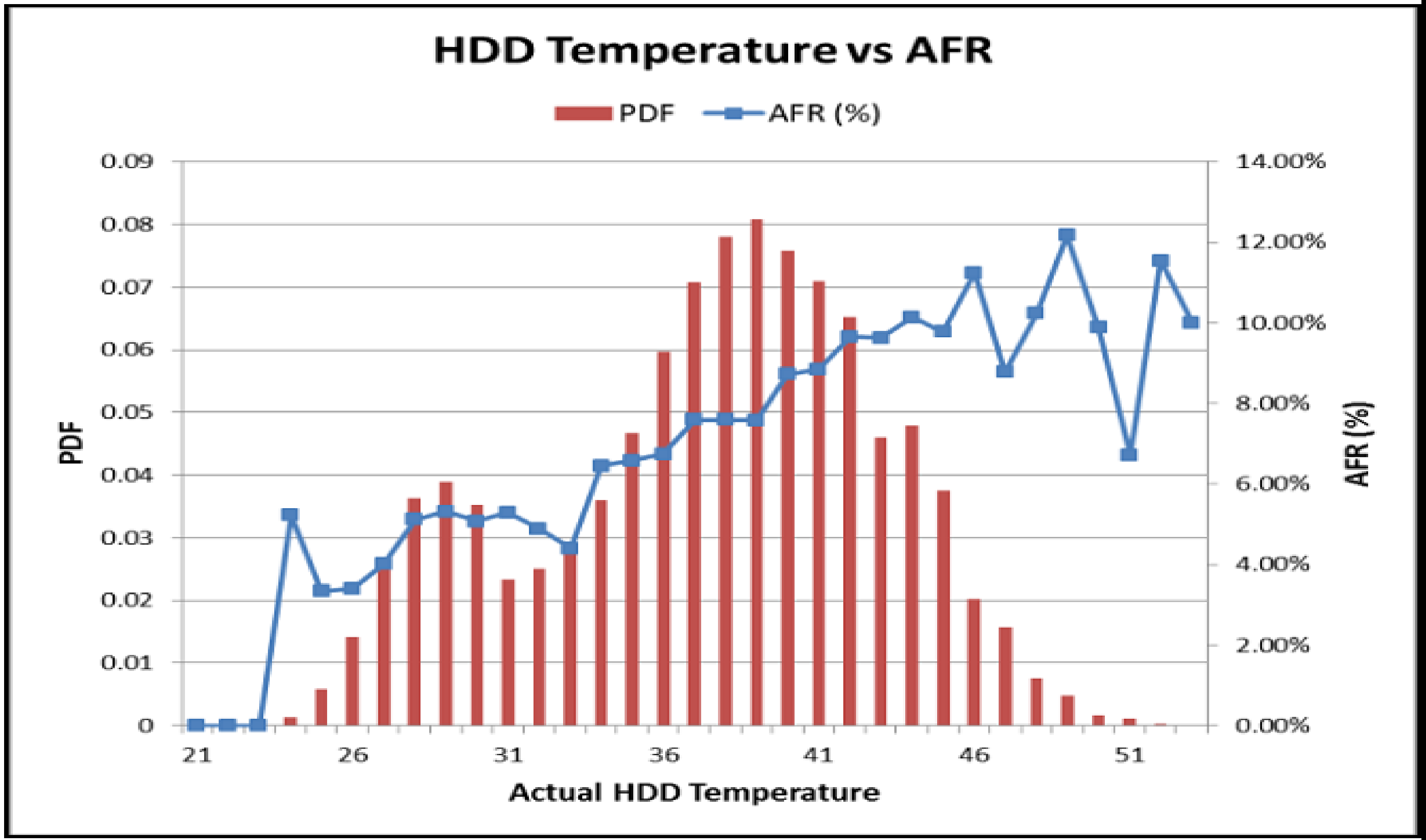
- Work done at the University of Virginia in collaboration with Microsoft
- Hardware component failures observed over two years from data centers with 100,000+ servers

HDD FAILURE RATE HIGHLY CORRELATED WITH TEMPERATURE



- Data contradicts some prior studies on the impact of temperature on HDD failures

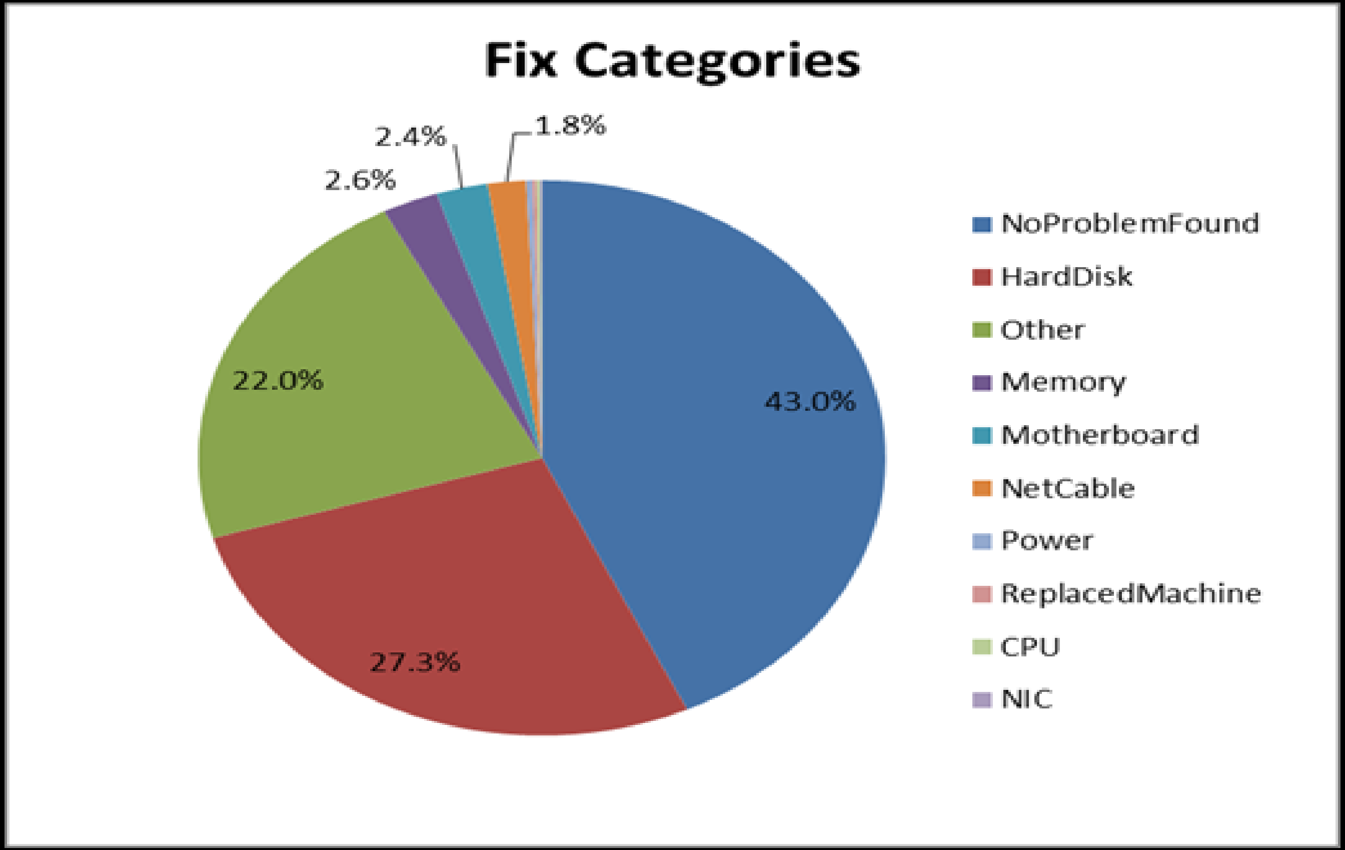
HDD FAILURE RATE HIGHLY CORRELATED WITH TEMPERATURE



Need more field studies!

“NO PROBLEM FOUND” (NPF) FAILURES

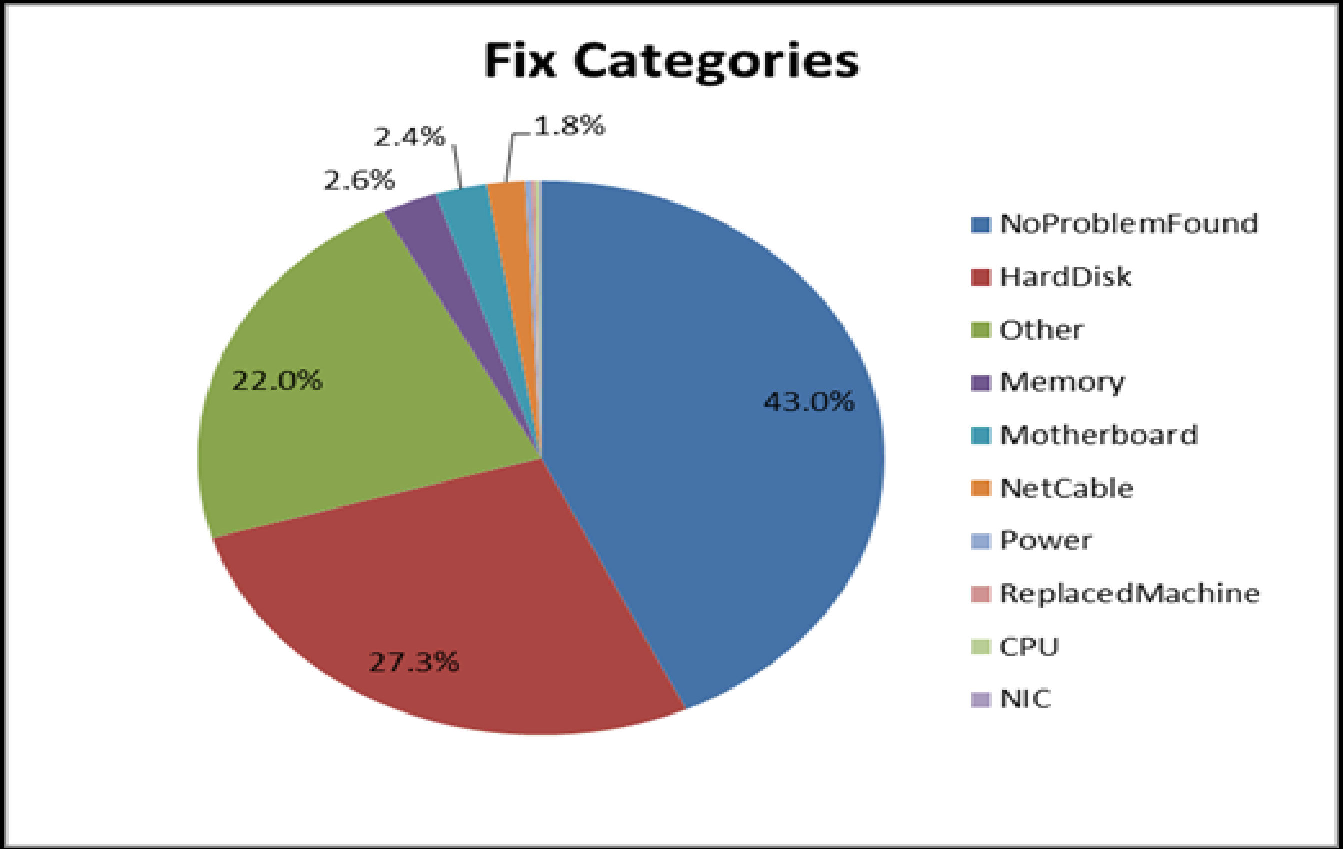
[SANKAR ET AL., IEEE CAL’14]



- Machines report a failure but no hardware failure detected
- Hard power-cycling, reseating HDDs, cables, etc. sometimes fix the problem
- Takes a long time to diagnose and hence affects service availability and quality

"NO PROBLEM FOUND" (NPF) FAILURES

[SANKAR ET AL., IEEE CAL'14]



- Machines report a failure but no hardware failure detected

Need better root cause analysis and failure prediction capabilities

SUMMARY

ANALYZE

COLLABORATE

SHARE

RELATED PUBLICATIONS

- V. Sridharan, N. DeBardeleben, S. Blanchard, K. Ferreira, J. Stearley, J. Shalf, S. Gurumurthi, Memory Errors in Modern Systems: The Good, The Bad, and the Ugly, International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2015.
- N. DeBardeleben, S. Blanchard, V. Sridharan, S. Gurumurthi, J. Stearley, K. Ferreira, Extra Bits on SRAM and DRAM Errors - More Data from the Field, IEEE Workshop on Silicon Errors in Logic - System Effects (SELSE), 2014.
- V. Sridharan, J. Stearley, N. DeBardeleben, S. Blanchard, S. Gurumurthi, Feng Shui of Supercomputer Memory - Positional Effects in DRAM and SRAM Faults, Supercomputing (SC), 2013.
- S. Sankar, S. Gurumurthi, Soft Failures in Large Datacenters, IEEE Computer Architecture Letters (CAL), 2014.
- S. Sankar, M. Shaw, K. Vaid, S. Gurumurthi, Datacenter Scale Evaluation of the Impact of Temperature on Hard Disk Drive Failures, ACM Transactions on Storage (TOS), 2013.

DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION