

<http://sploid.gizmodo.com/watching-uranium-emit-radiation-inside-a-cloud-chamber-1689997373/>



Uranium emitting radiation in a cloud chamber

# Data-Driven Decision Making in Resilience

Nathan DeBardeleben, Ph.D.  
Los Alamos National Laboratory  
High Performance Computing  
Ultrascale Systems Research Center Lead

# A Controversial Claim

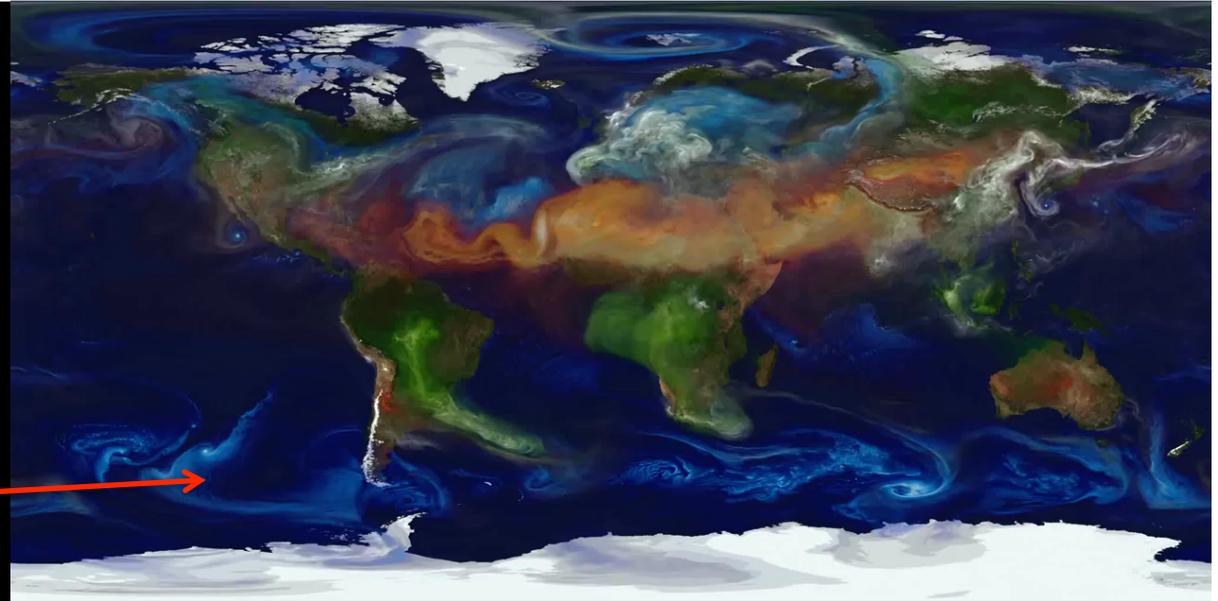


Jose-Luis Olivares/MIT

- We will never build a **reliable** exascale computer out of commodity parts
  - Performance
  - Power
  - Reliability?

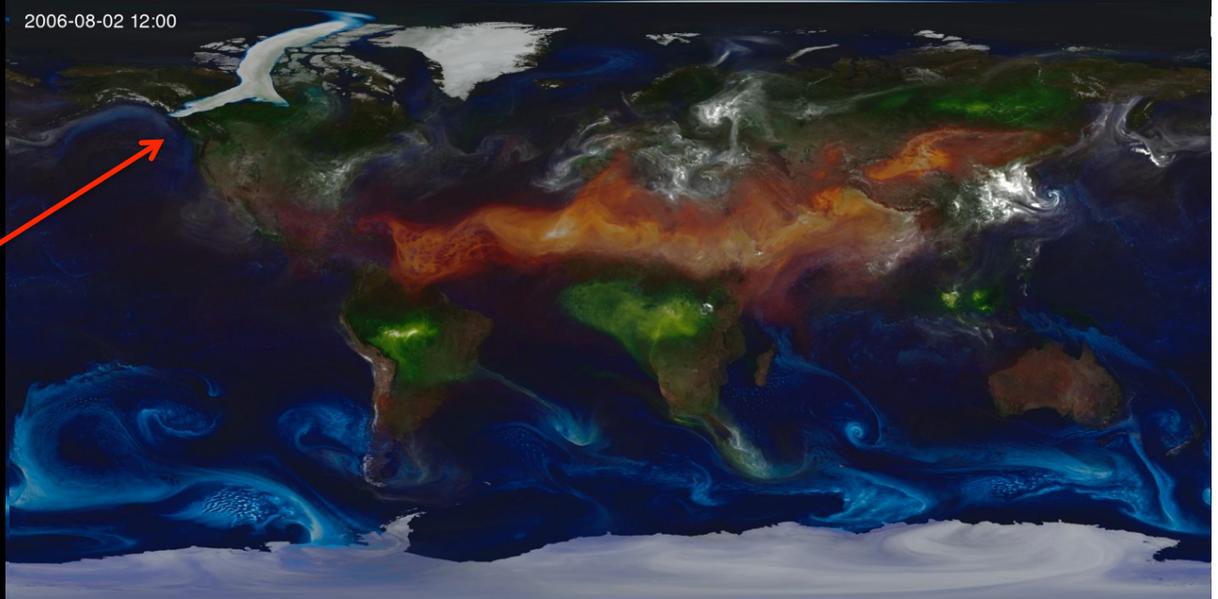
# NASA, SDC in Sea Salt Aerosol Calculation

- Sea Salt aerosol is plotted from white to blue (mostly observable in south of map)
- SDC requiring expert knowledge to identify

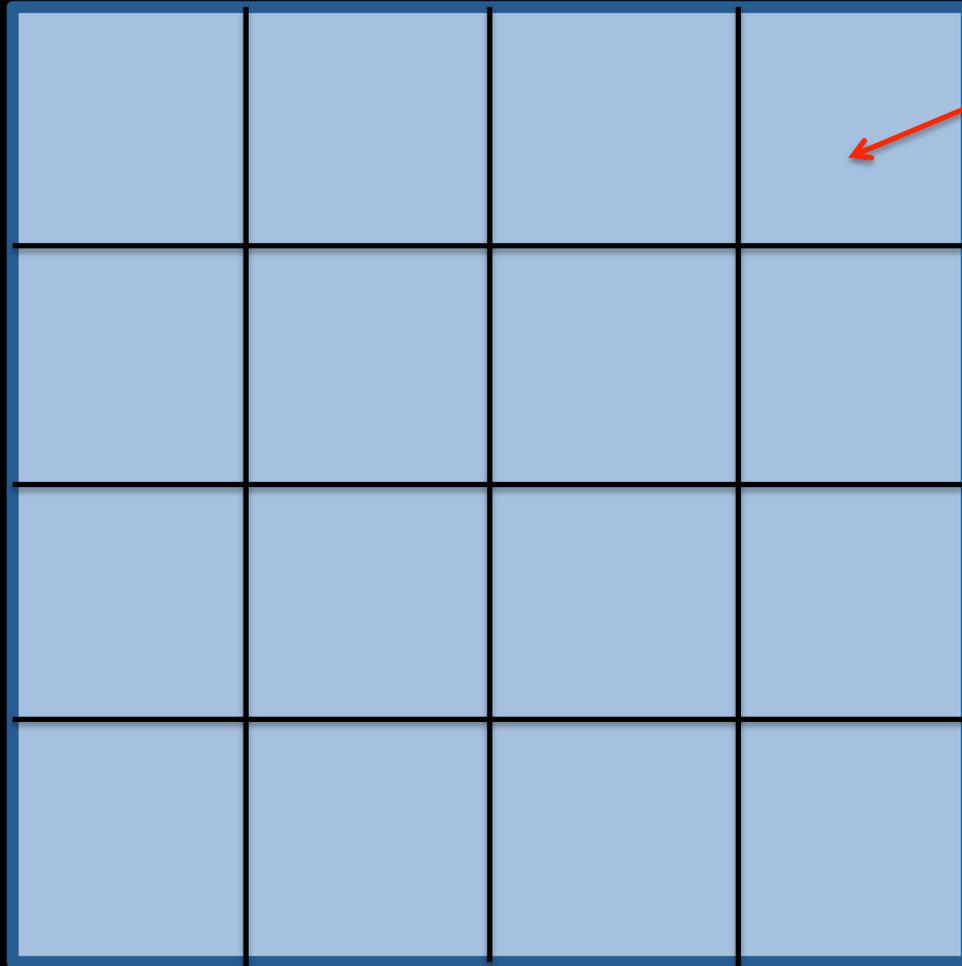


[https://www.youtube.com/watch?v=s\\_DXtWY5ovU#t=81](https://www.youtube.com/watch?v=s_DXtWY5ovU#t=81)

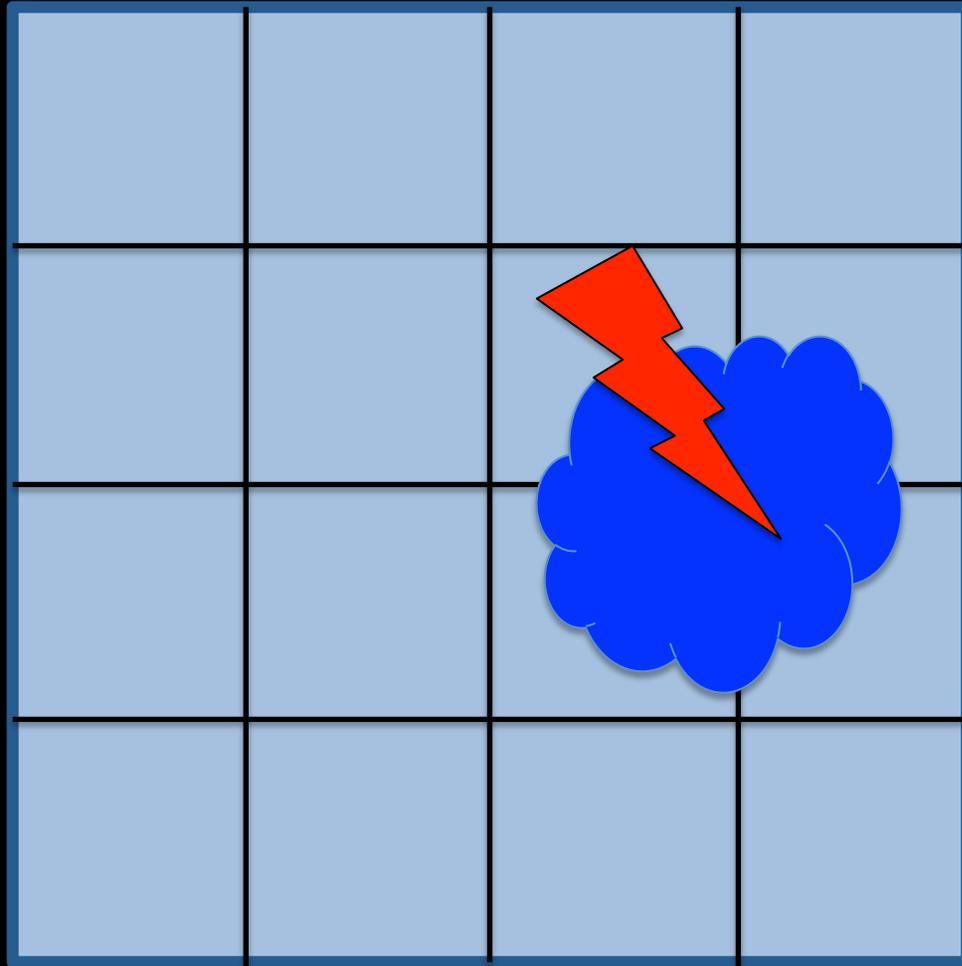
2006-08-02 12:00

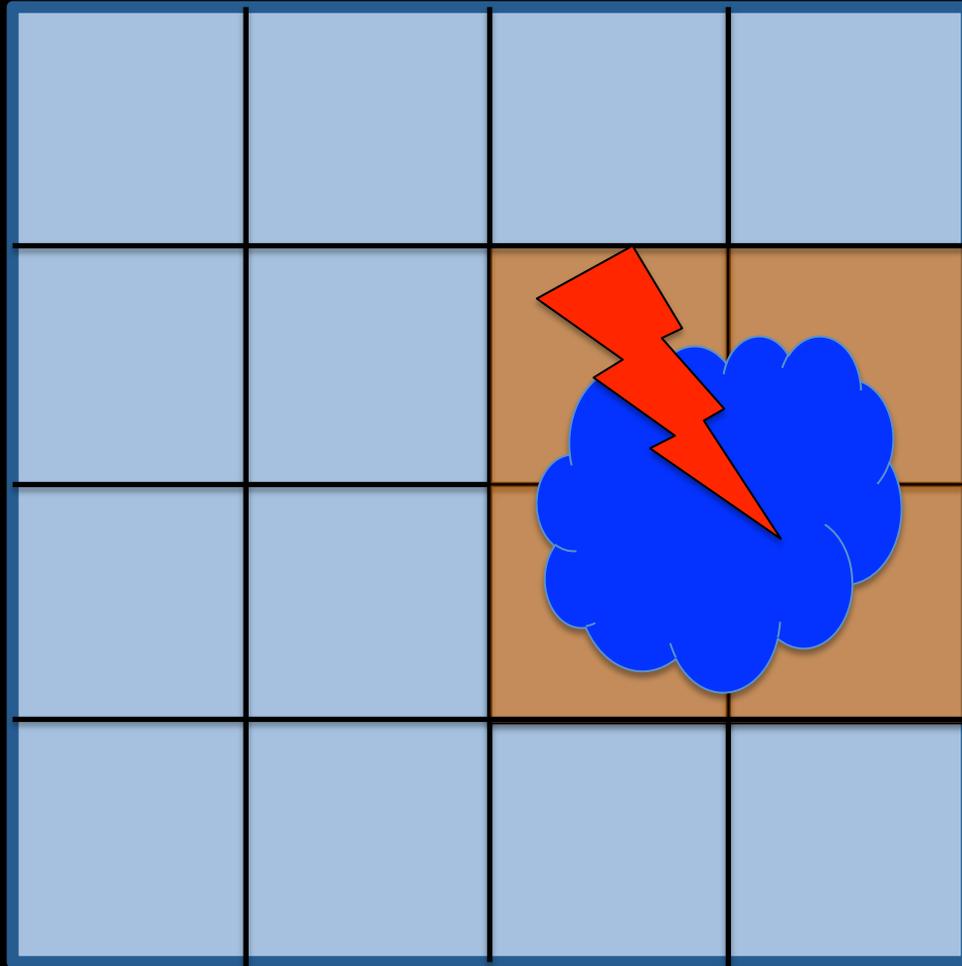


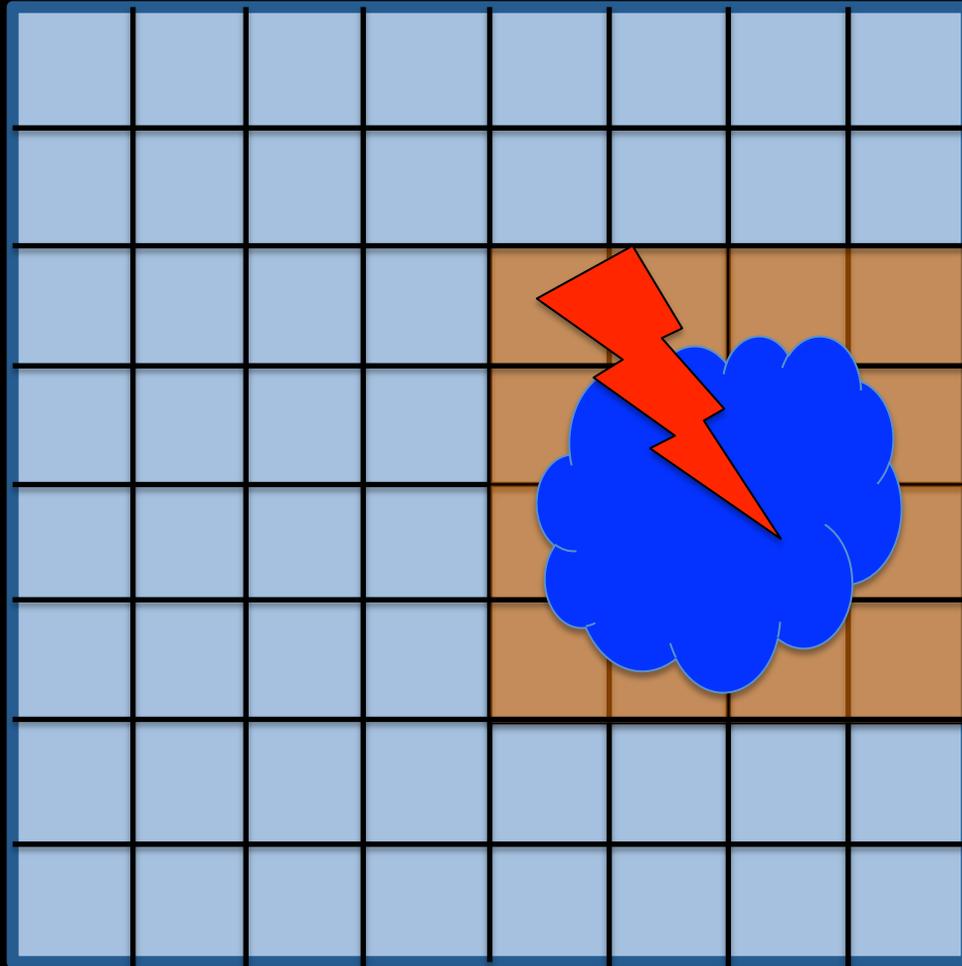
*Courtesy Tyler Simon, DOD, from collaboration with NASA while at UMD*

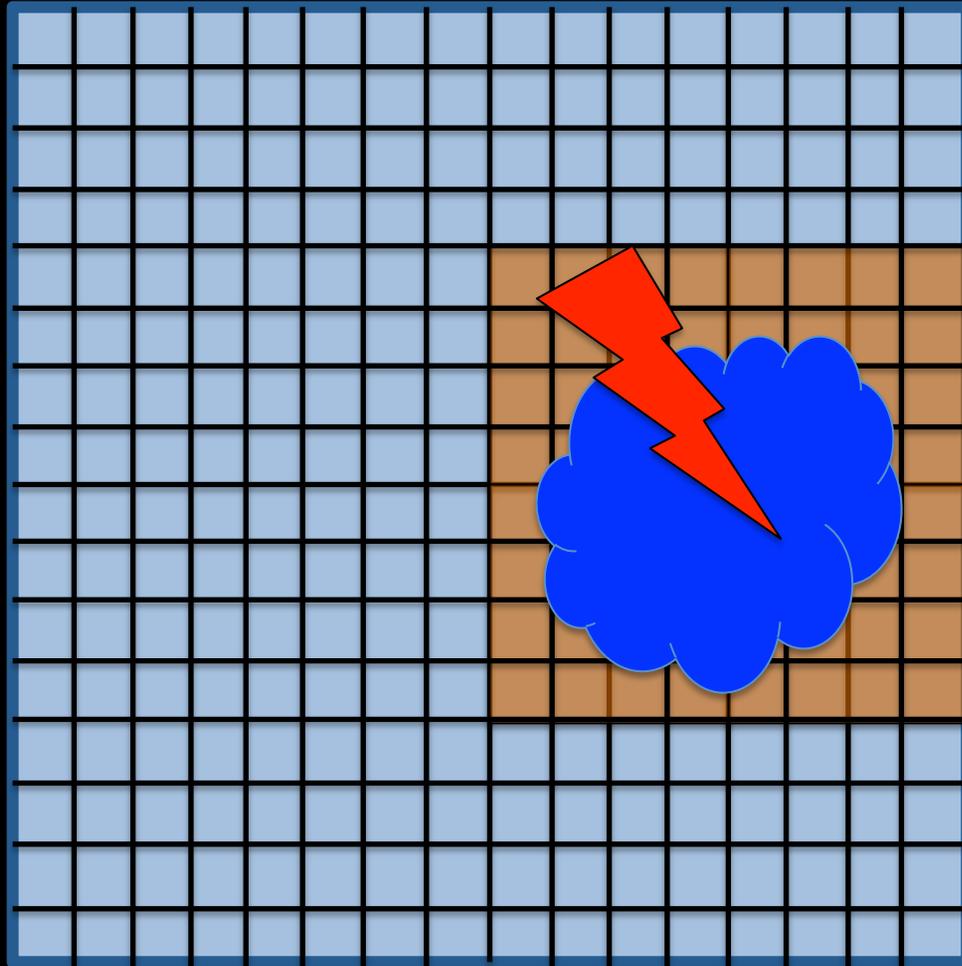


Memory Cell









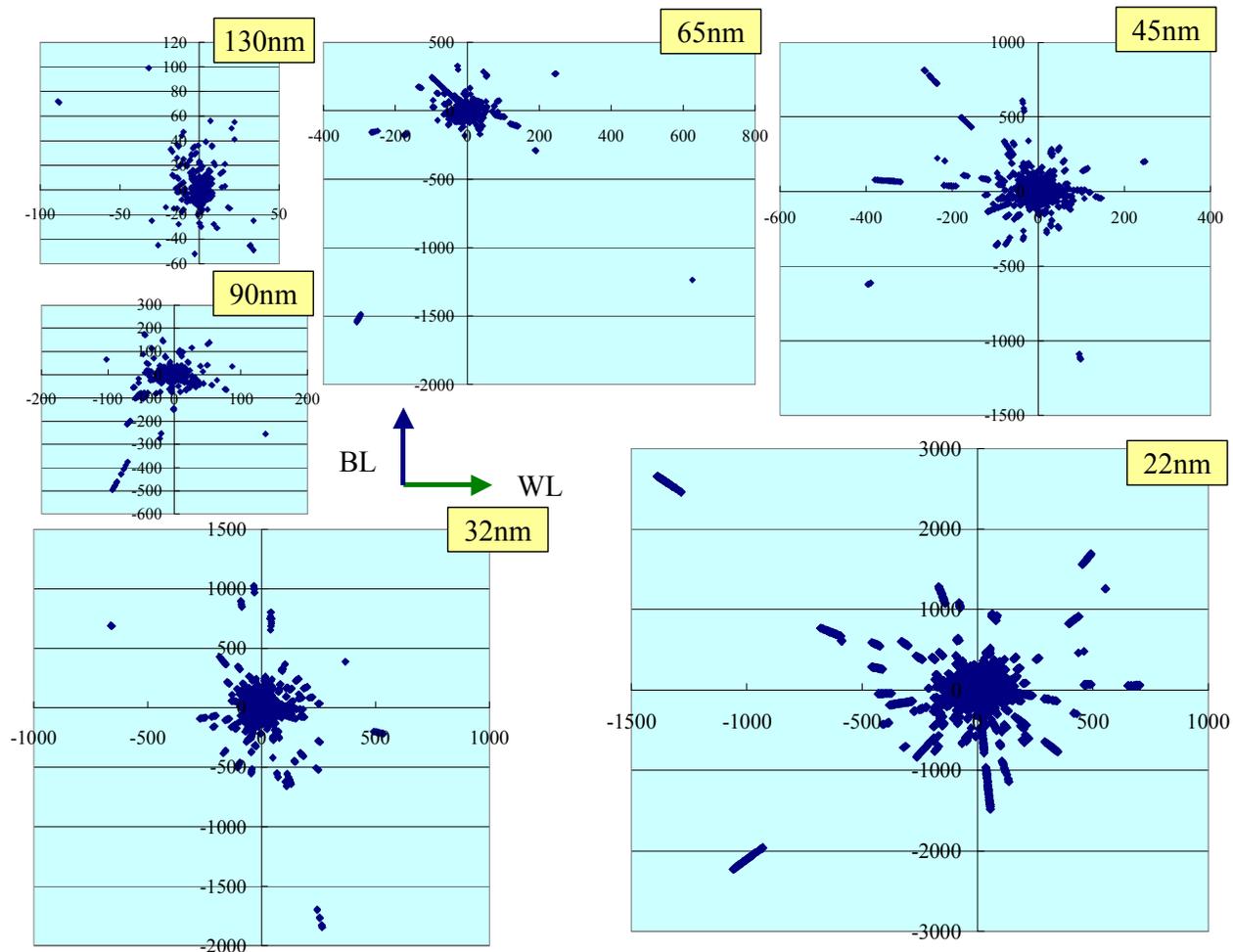
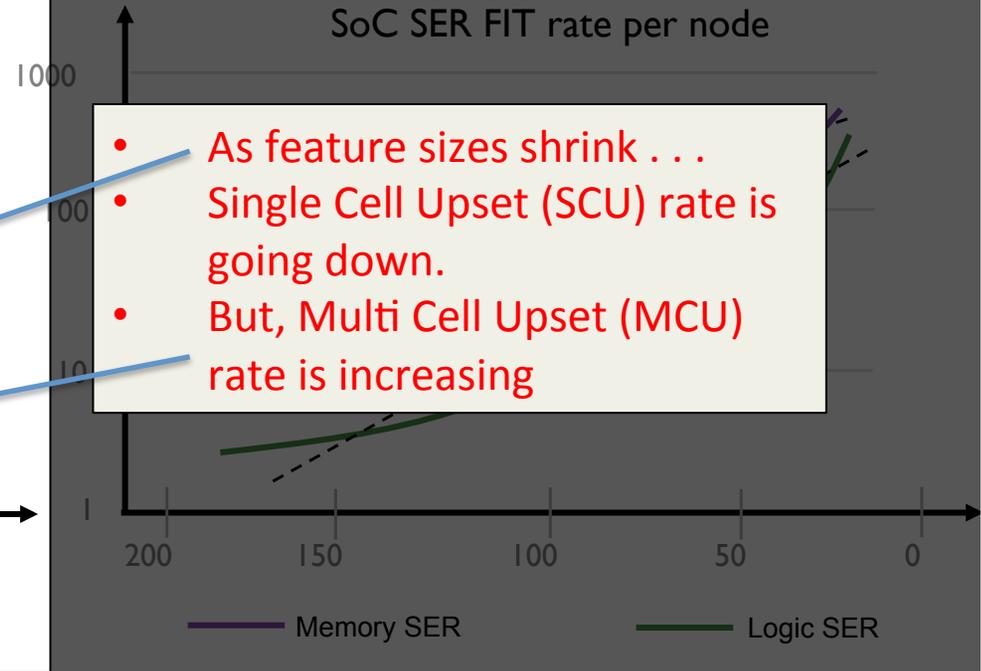
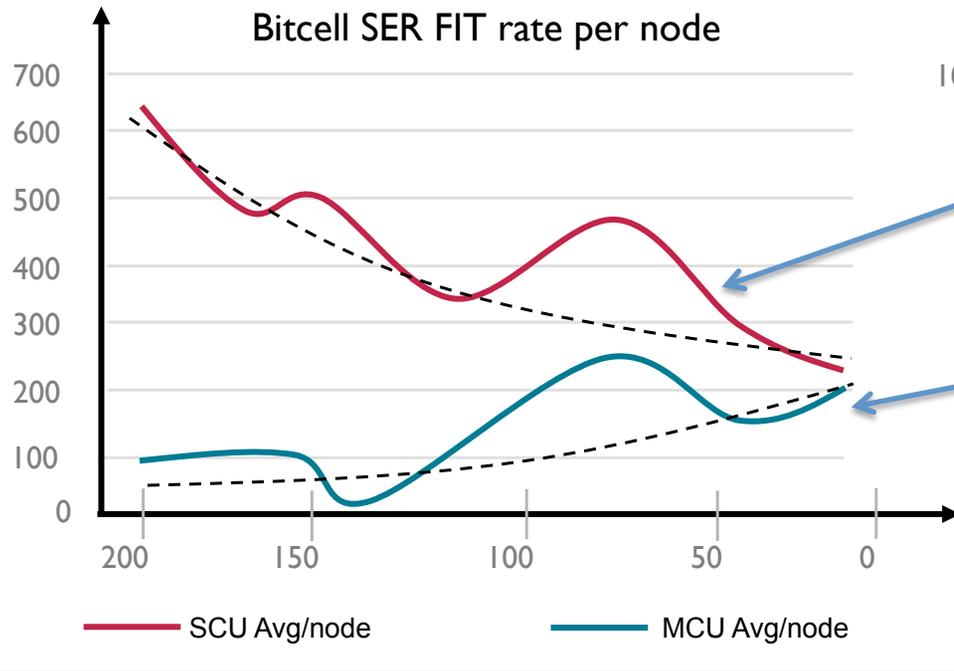


Fig. 7 Failed Bit Maps for Each Generation with CB Pattern.

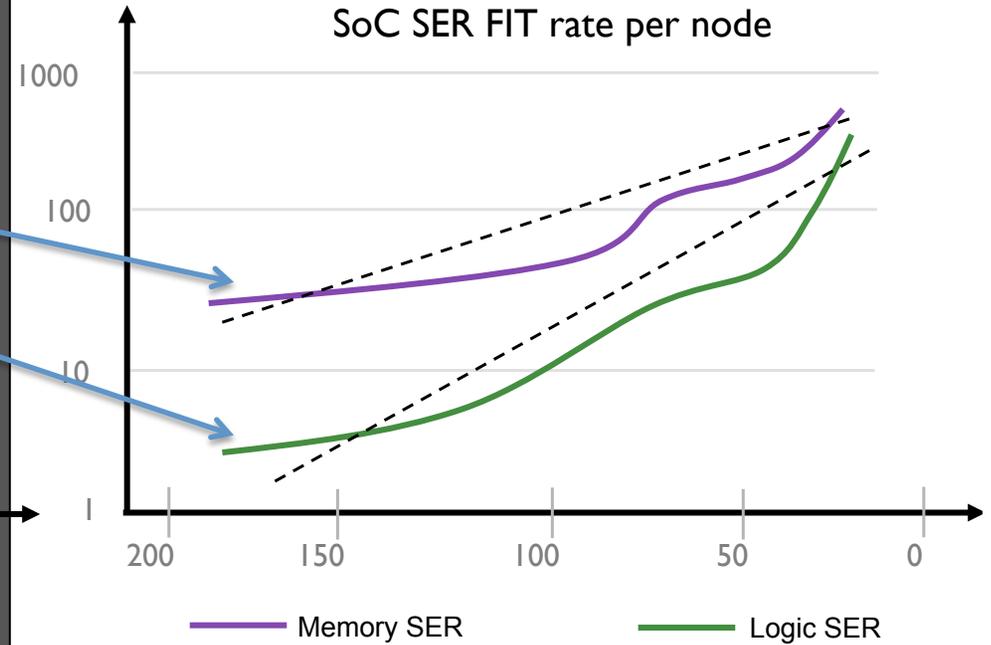
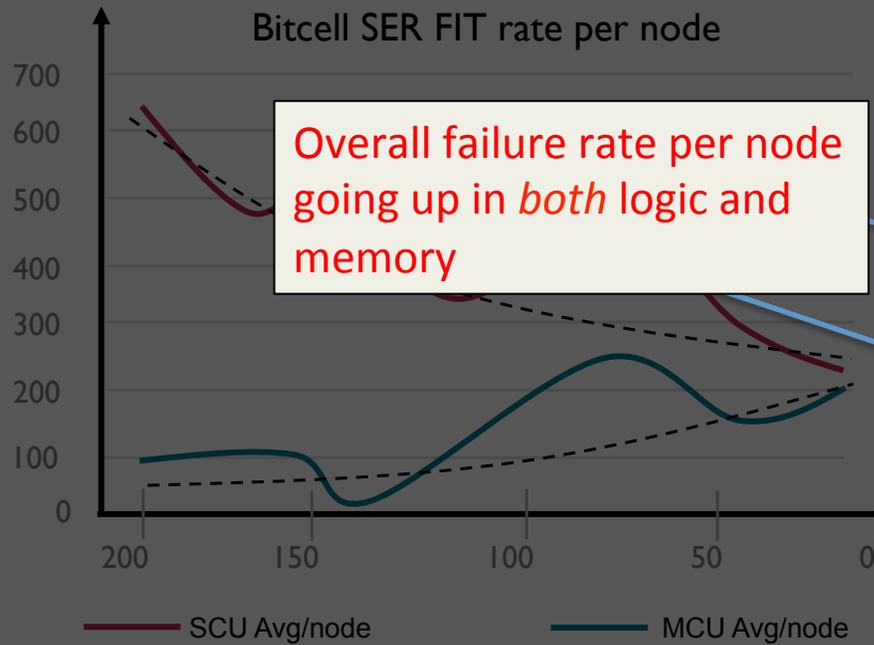
*Scaling Effects on Neutron-Induced Soft Error in SRAMs Down to 22nm Process – Ibe-san, et. al., Hitachi*

# SoC soft error trends



Even though per memory bitcell SER sensitivity is decreasing, overall FIT per SoC is increasing

# SoC soft error trends



Even though per memory bitcell SER sensitivity is decreasing, overall FIT per SoC is increasing

DATE 2014

11

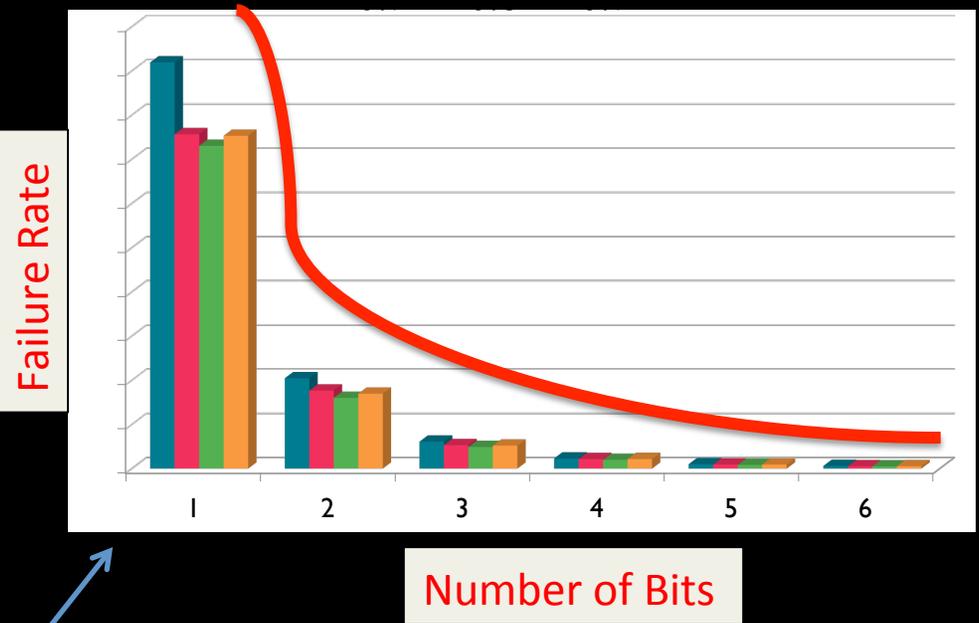
Source: iRoC



Reliability Challenges in Embedded Processors – Vikas Chandra, ARM R&D

Cielo @ LANL  
Hopper @ NERSC  
Titan @ ORNL  
All see similar trends in DRAM.

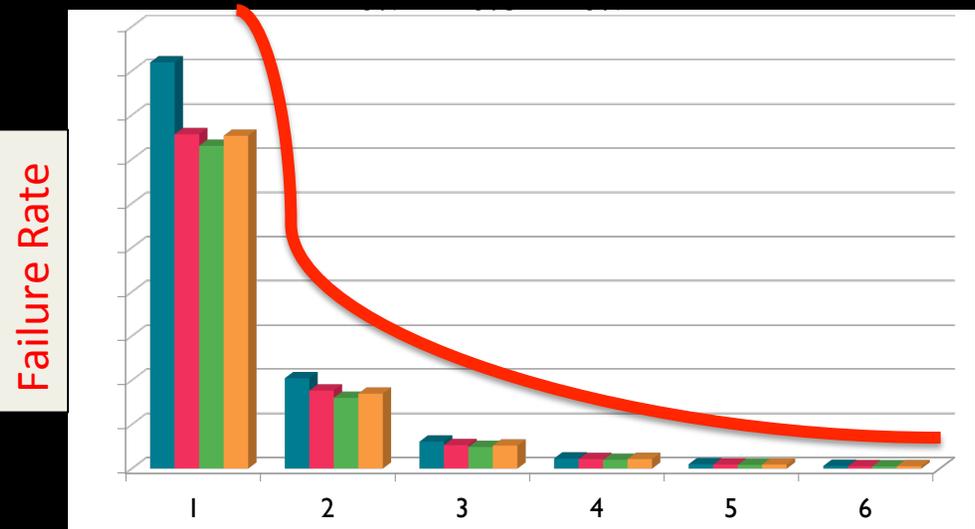
~ SIMILAR



Reliability Challenges in Embedded Processors – Vikas Chandra, ARM R&D

Early GPUs for the HPC market (still in production today at LANL) ~1 multi-bit error / year

~ SIMILAR



Failure Rate

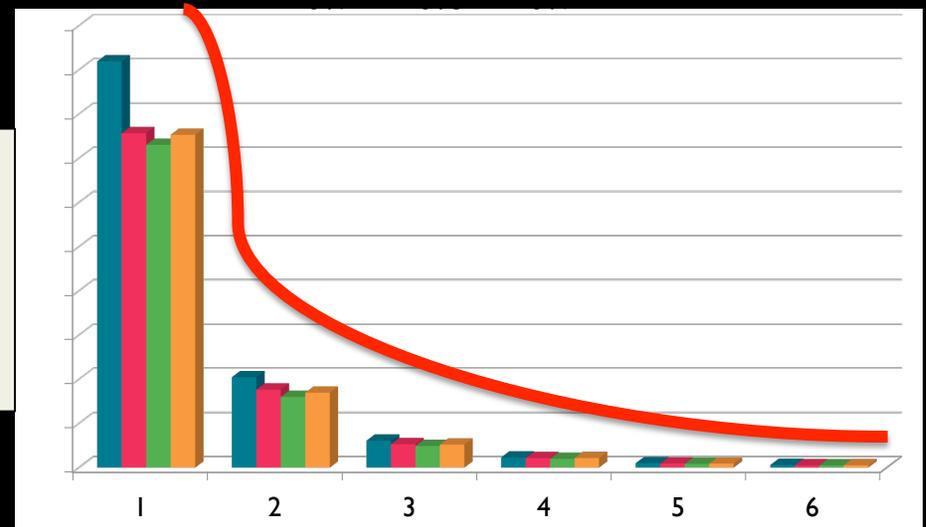
Number of Bits

Reliability Challenges in Embedded Processors – Vikas Chandra, ARM R&D

These early GPUs if deployed on Titan -> 40-50 fail stop errors per day!  
Luckily, the technology has improved!

~ SIMILAR

Failure Rate

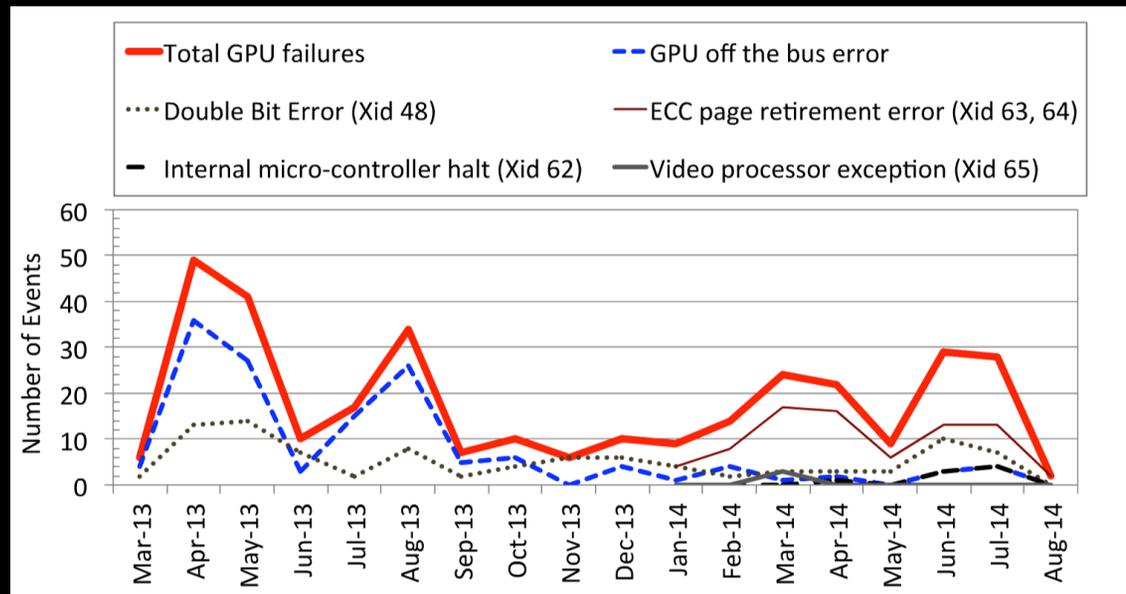


Number of Bits

Reliability Challenges in Embedded Processors – Vikas Chandra, ARM R&D

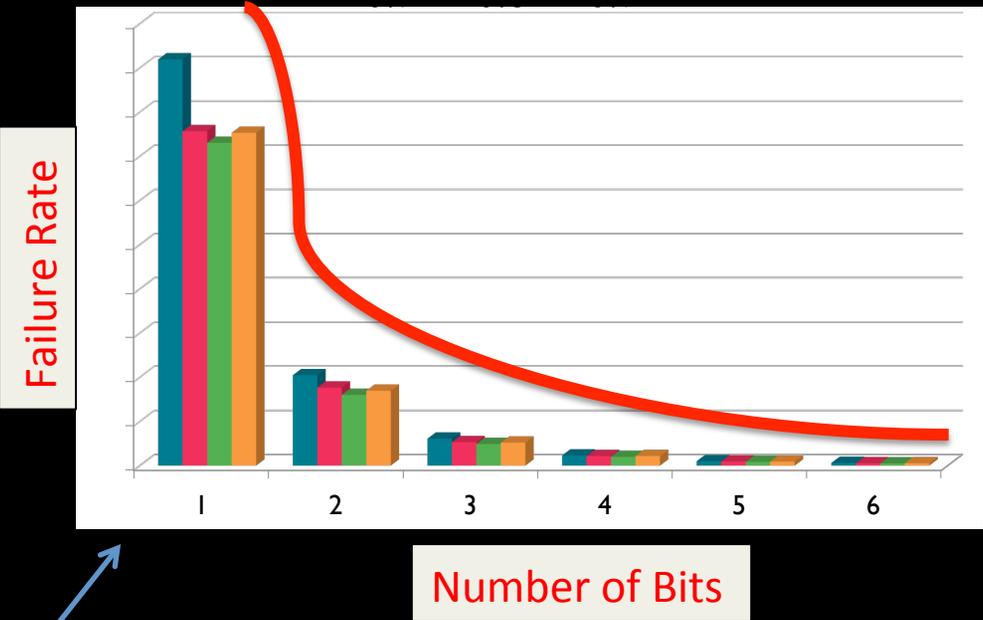
# Titan, K20 GPGPU Errors

- Massive improvements in reliability
- Moonlight -> Titan = 30x improvement
  - 1 error / day -> 1 error / day
- Titan -> Exascale  $\sim$  30x improvement
  - 1 error / day -> 1 error / day?
  - Just GPU errors, what about DRAM, SRAM, MCDRAM, network, storage, . . . ?
  - Already below DOE's 1 / day target!



Understanding GPU Errors on Large-scale HPC Systems and the Implications for System Design and Operations, Devesh Tiwari, et. al., HPCA 2015

# HPC? Scale Scale Scale!



*Reliability Challenges in Embedded Processors – Vikas Chandra, ARM R&D*

- MCDRAM (3D stacked memory)
- An opportunity to study new technology on Trinity

Emphasis added

The results of the MTTF evaluation are shown in Fig. 10. Here, MTTF is less than 100 hours when the system has more than 100,000 memory modules. This points to a serious reliability problem for the next generation of HPC systems because the next generation of HPC systems will have more than 100,000 memory modules. The sensitivity analysis

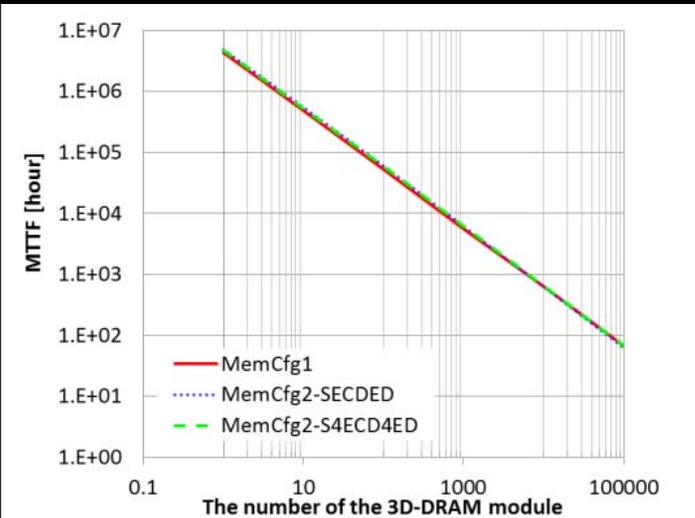
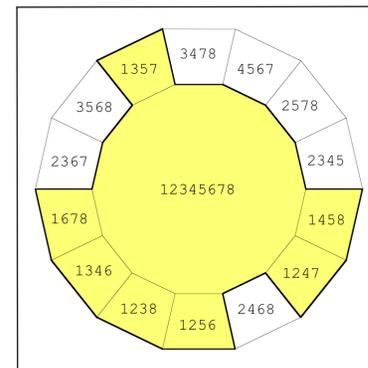


Figure 10: MTTF Evaluation Results

*Analytical Reliability Model of Die-Stacked DRAM Protected by Error Control Code and TSV Fault Tolerant Coding Technique – Matsumura, et. al., Hitachi, SASIMI 2015*

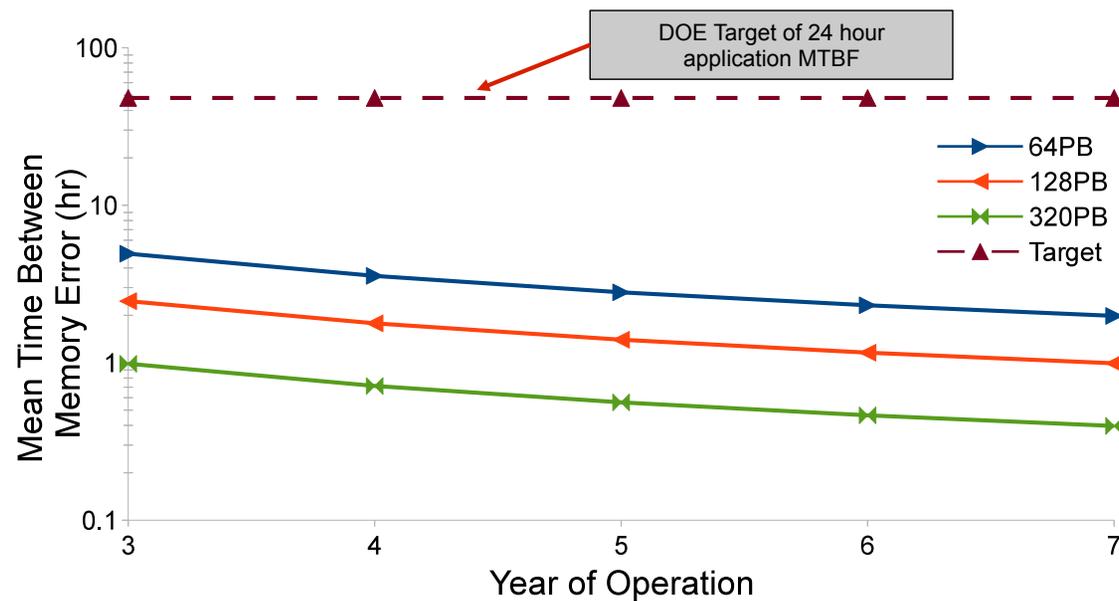
# What Level of Chipkill is Required?

- We clearly need more than double bit error protection at our scale
- Most DOE systems deploy some level of Chipkill error correction
- What if we build an exascale system out of single chipkill correct (like Cielo & Titan)?
  - Results of analytical model, Monte Carlo simulation (250+ million simulations), all fed with failure data from Jaguar



# Single Chipkill Correct is not enough

- Conclusion: single chipkill correct not sufficient

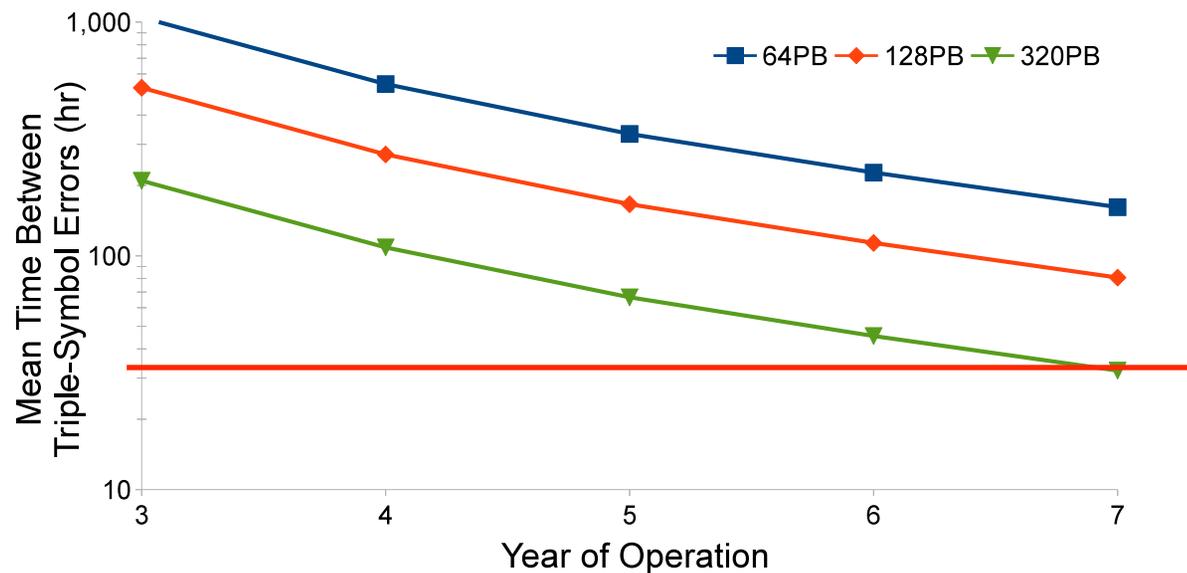


*Nothing limits you like not knowing your limits - Tom Hayes*



# Double Chipkill Correct Should be Good Enough for DOE @ Exascale

- **Conclusion:** double-chipkill correct sufficient for DUE target but if we reduce to double-chipkill *detect*, possibly not enough for SDC target

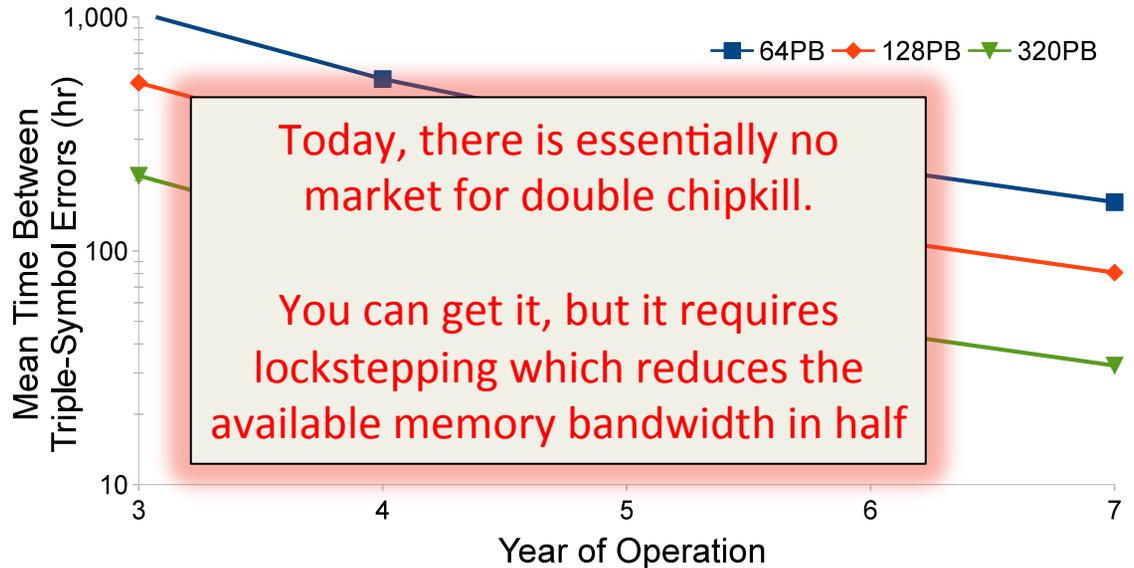


*Nothing limits you like not knowing your limits - Tom Hayes*



# Double Chipkill Correct Should be Good Enough for DOE @ Exascale

- Conclusion: double-chipkill correct sufficient for DUE target but if we reduce to double-chipkill **detect**, possibly not enough for SDC target



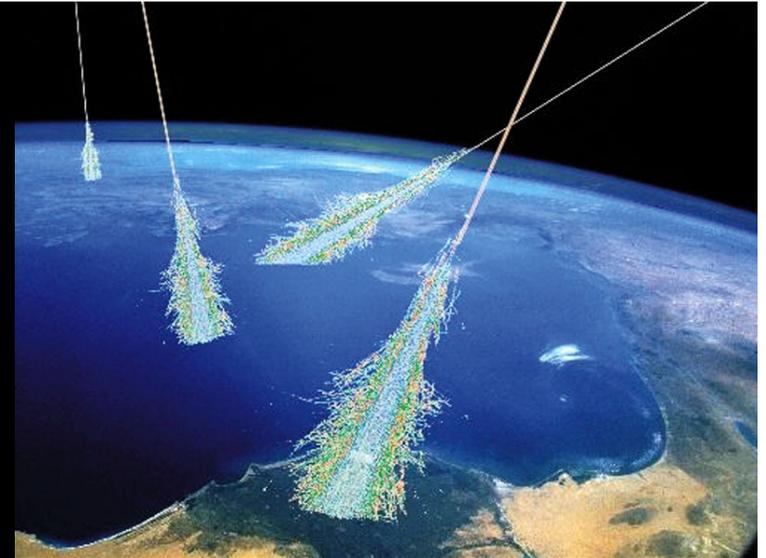
LA-UR-13-29085

Analyzing Reliability of Memory Sub-systems with Double-chipkill Detect/Correct, Xun Jian, et. al., PRDC 2013

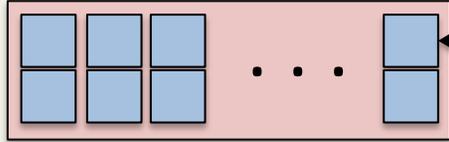
*Nothing limits you like not knowing your limits - Tom Hayes*

## Correctable Errors – Operator Data

- Cielo – **~0.3** correctable errors / min
- Hopper – **~1** correctable error / min
- Titan – **~1.4** correctable errors / min
- BlueWaters (DSN2014) – **~4.2** correctable errors / min

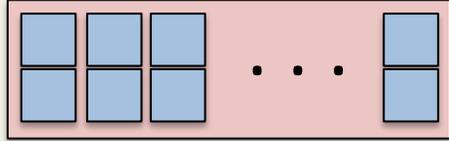


**DIMM 1** *Titan DIMMs have DRAM devices on both side*



**DRAM DEVICE**  
*EX: Titan has 36 DRAM devices / DIMM (18 / side)*

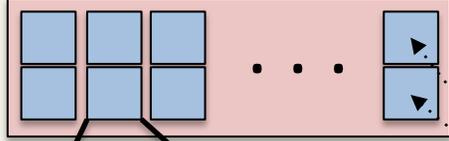
**DIMM 2**



**Node, Channel**  
*determines which DIMM*

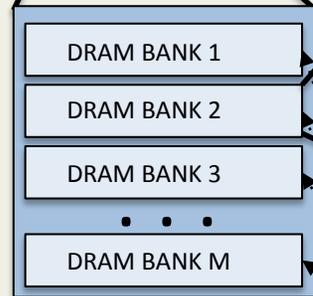
• • • *EX: Titan has 4 DIMMs / host*

**DIMM N**



**ChipSelect, Lane**  
*determines which DRAM device*

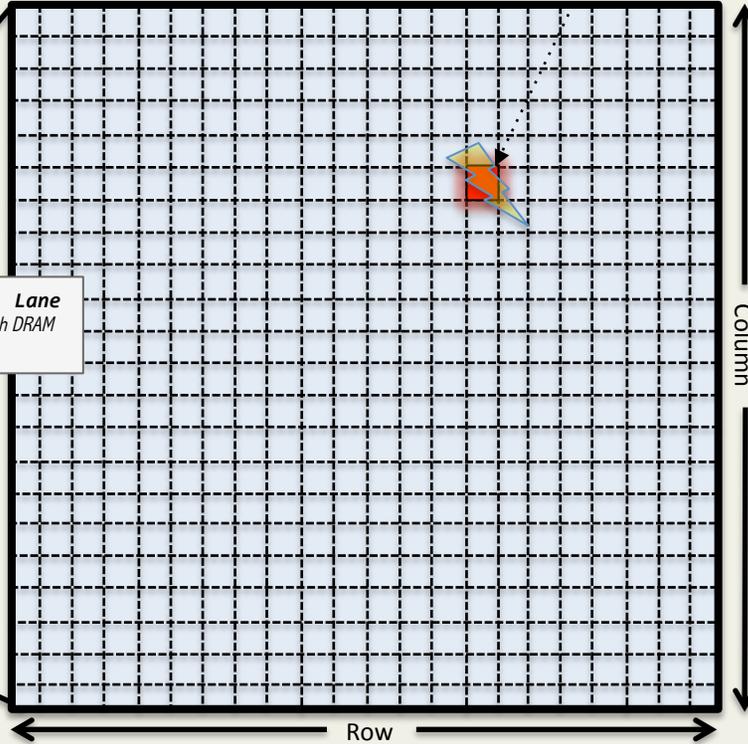
**DRAM DEVICE**



*EX: Titan has 8 DRAM banks / DRAM device*

**Bank**  
*determines which DRAM Bank*

**DRAM BANK**



**host**  
*determines which host / motherboard*

**Bit**  
*Refers to which DQ pins were signaled. On Titan, there are 4 pins (this is known as x4 or "by four") (0,1,2,3) and all combinations (0,12,023,etc) are valid. # of integers = number of bits involved in the error*

**Row and Column**  
*determines location of error on a DRAM bank*

# Strong Collaboration is Beneficial to Both Parties

- We can benefit from collaborating with hardware vendors to better understand their hardware in our systems
- Many hardware errors cannot be decoded properly without vendor assistance
- Collaboration with AMD – Sudhanva Gurumurthi and Vilas Sridharan since 2012
  - Many DOE collaborators
  - Jaguar @ ORNL
  - Cielo @ LANL - ~24billion DRAM hours – years of system data
  - Hopper @ NERSC - ~22billion DRAM hours
  - Titan @ ORNL (ongoing)
- I share externally only a fraction of what we discover
- DOE is good at protecting data and analysis
- **Now for a sampling of outcomes of collaboration . . .**

John Shalf

Kurt Ferreira, Jon Stearley @ SNL

Christian Engelmann, Devesh Tiwari @ ORNL

# Most of the system **NEVER** experiences a DRAM error

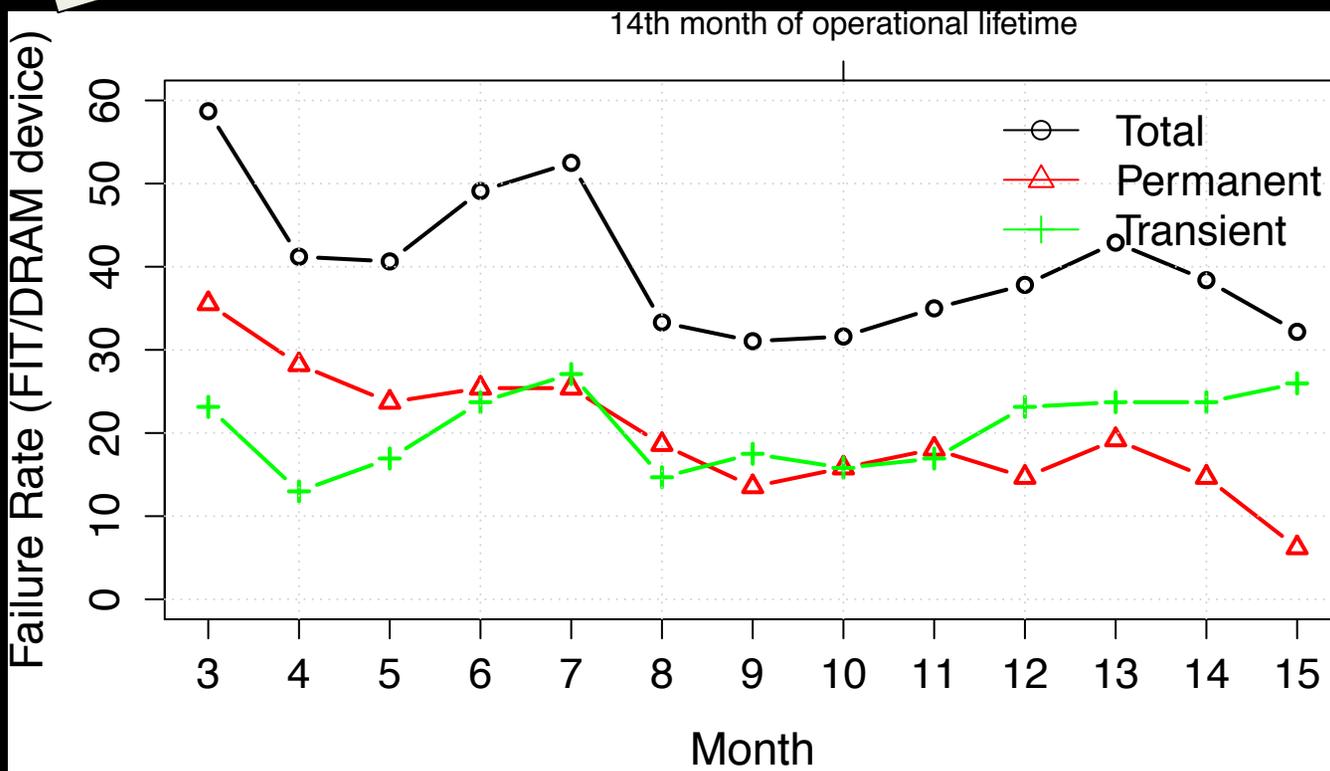
- Yet because of our scale, error rates are still a problem
- Memory with permanent damage needs to be removed promptly

System	0	1	2	3
Cielo	90.07%	9.10%	0.75%	0.08%
Jaguar	94.07%	5.48%	0.39%	0.06%

**Table 2: Percentage of hosts with 0, 1, 2, or 3 faulty DRAMs.**

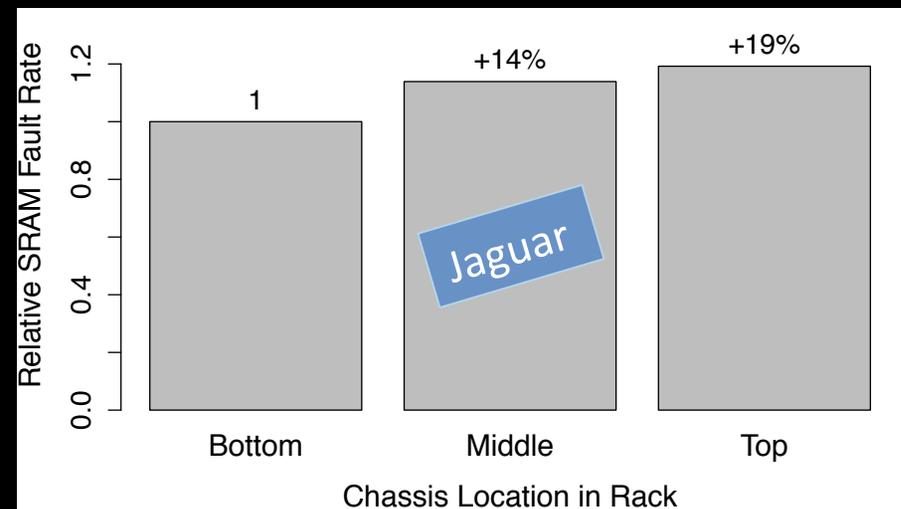
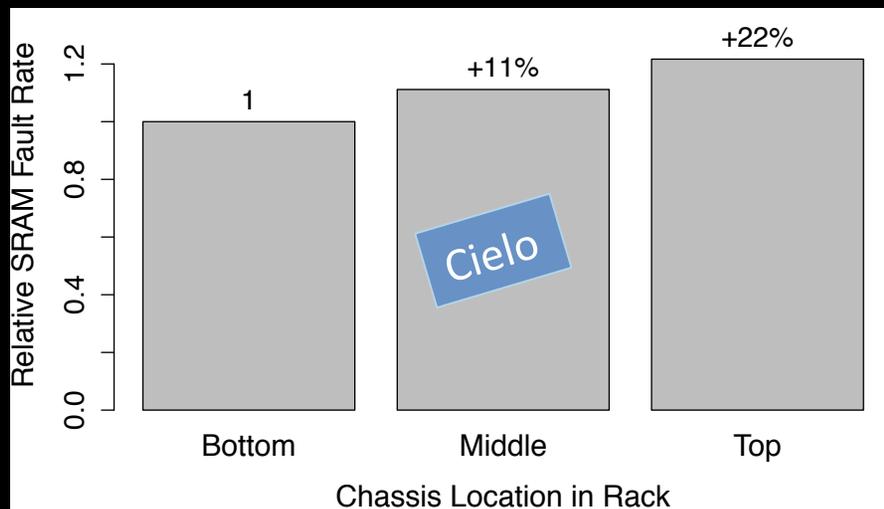
# Over Time, Faulty Hardware is Weeded Out

Cielo



# Schedule Your Jobs at the Bottom of a Rack!

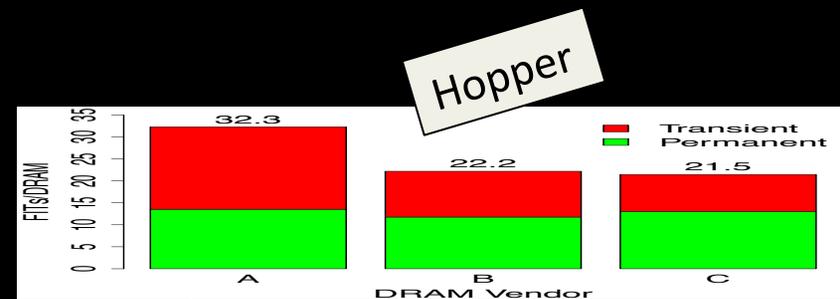
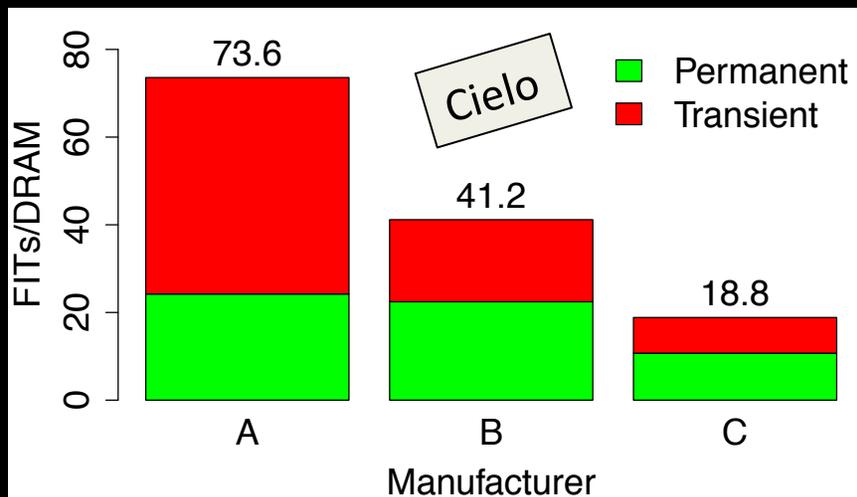
- ~10% increase in SRAM fault rates at each chassis level
- Temperature?
- Cosmic radiation shielding?



*Feng Shui of Supercomputer Memory, Vilas Sridharan, et. al., SC 2013*

# Not all DRAM Vendors are Created Equal

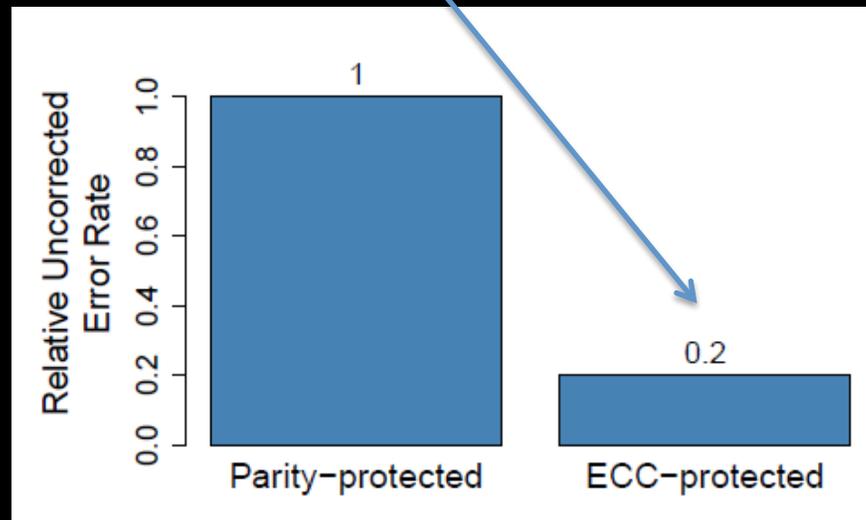
- Study your failure rates by vendor
- Altitude effects are real, but can be mitigated by quality error protection



Squishing to attempt to put on the same scale

# SRAM Uncorrected Errors

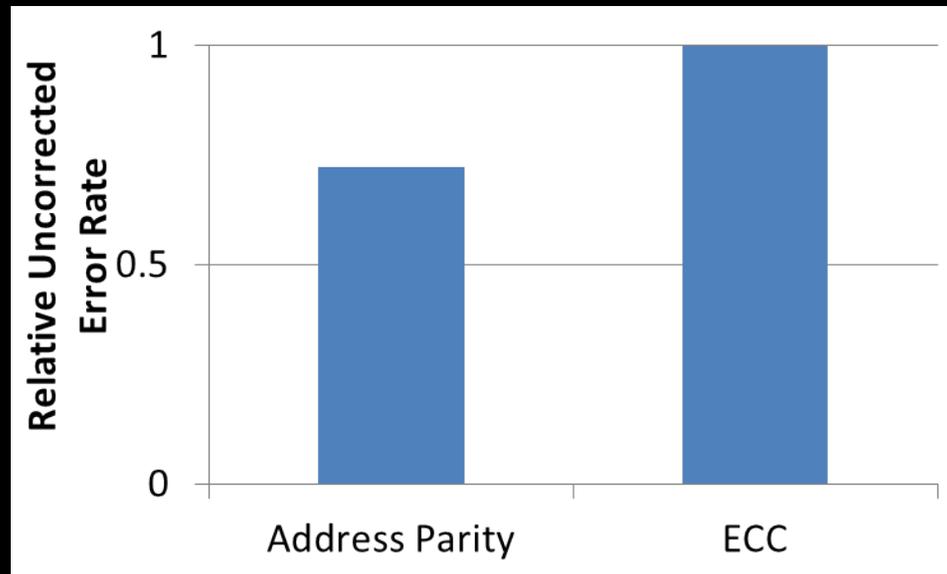
- The majority of SRAM DUEs are in parity-protected structures
- So extending ECC to these structures = win (more on the feasibility of this later)
- But that won't "fix it all"



*Memory Errors in Modern Systems: The Good, the Bad, and the Ugly*, Vilas Sridharan, et. al., ASPLOS 2015

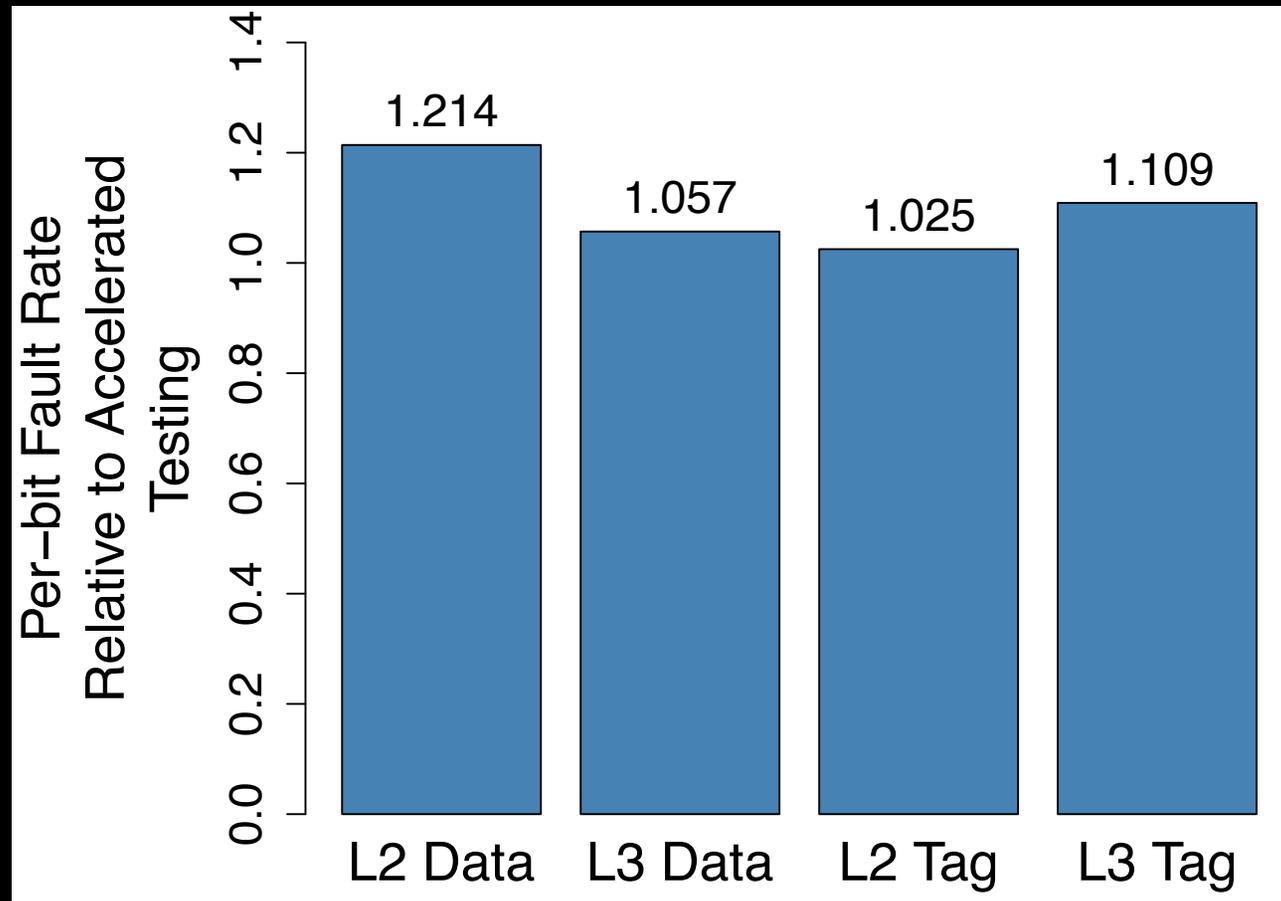
# DDR-3/4 Improvements are Valuable

- Command / Address parity added to DDR-3 to JEDEC specification
- Improves system reliability



*Memory Errors in Modern Systems: The Good, the Bad, and the Ugly,*  
Vilas Sridharan, et. al., ASPLOS 2015

# Accelerated Testing Compares Favorably



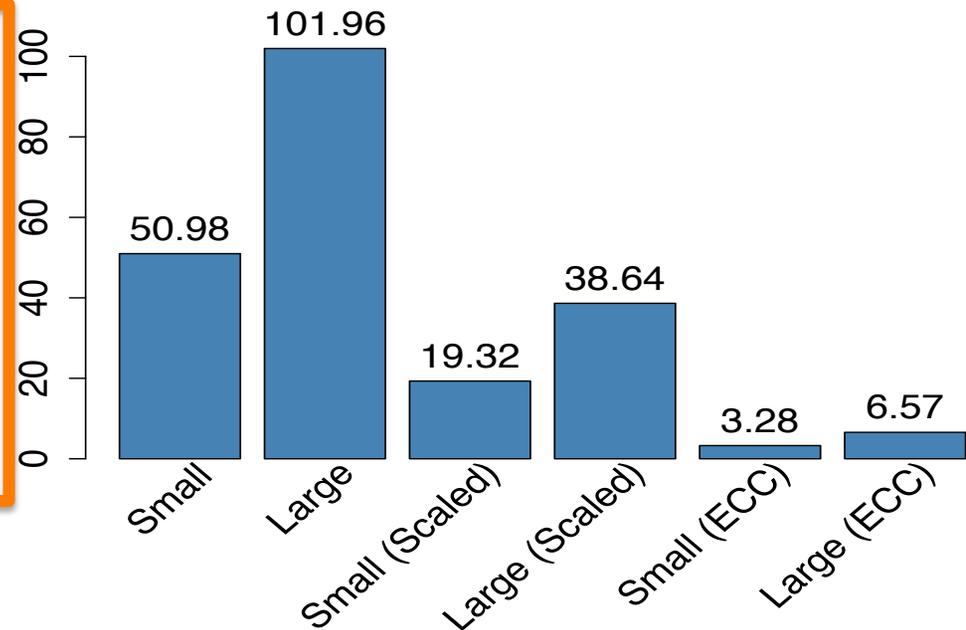
*Memory Errors in Modern Systems: The Good, the Bad, and the Ugly,*  
Vilas Sridharan, et. al., ASPLOS 2015



# Is the Sky Falling?

- Just add more advanced error protection to circuits
- But who will pay for it?
- Does the server and commodity market need this level of protection?
  - Historically, no
- So while we **COULD** solve this, a better question is whether we **WILL**.

Uncorrected Error Rate  
(Relative to Cielo)

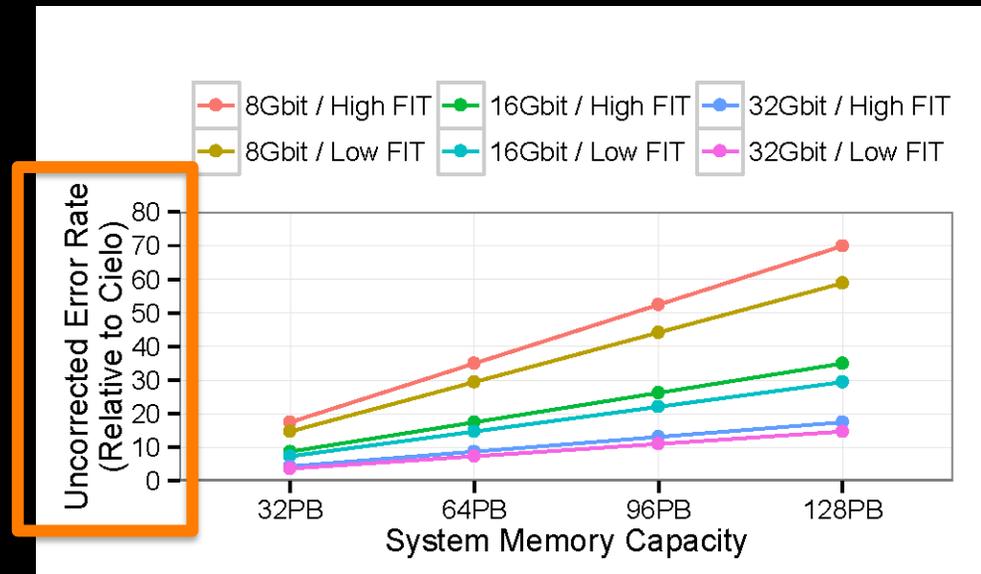


*Memory Errors in Modern Systems: The Good, the Bad, and the Ugly*, Vilas Sridharan, et. al., ASPLOS 2015



# Is the Sky Falling?

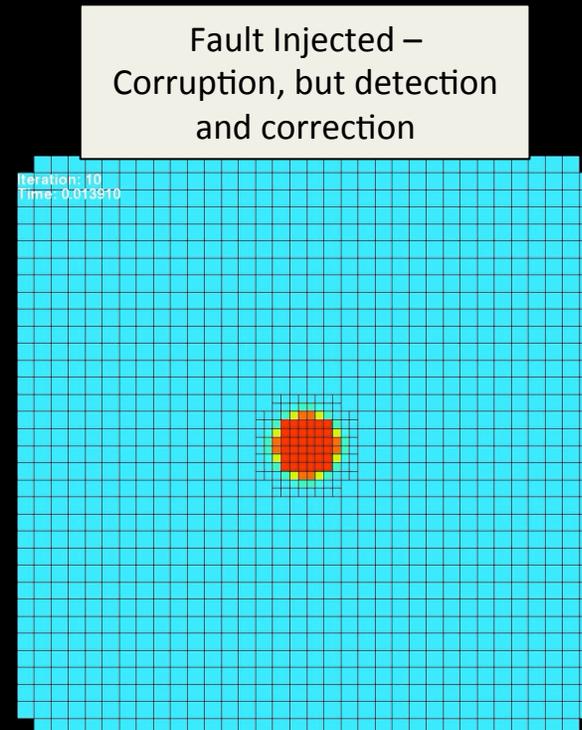
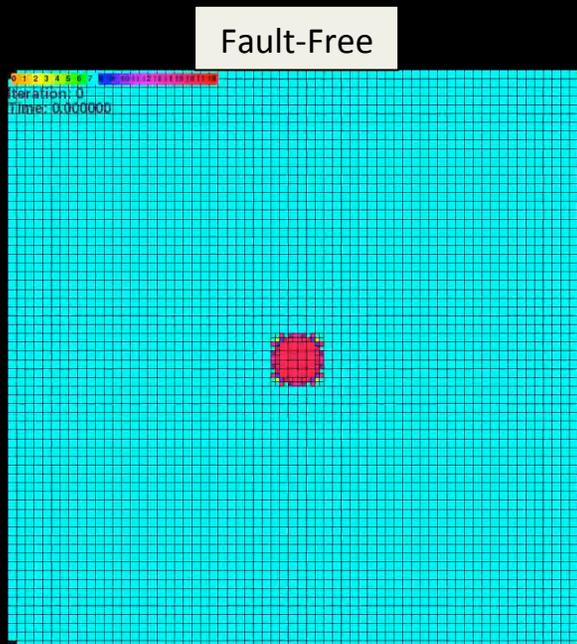
- The story for DRAM is largely the same
- For all prospective system sizes, uncorrectable error rates increase
- Note projection is still ~15x - ~70x (just in DRAM DUE increases from Cielo) at 128PB
- Don't forget to add SRAM DUE



*Memory Errors in Modern Systems: The Good, the Bad, and the Ugly*, Vilas Sridharan, et. al., ASPLOS 2015

# Applications Need Tools to Understand Their Resiliency

- F-SEFI – a VM-based fault injector
- University collaborators: Clemson University, Syracuse University, Coastal Carolina University, University of North Carolina
- CLAMR AMR hydro proxy app (LANL, XCP, Robert Robey)
  - Built detection and correction mechanisms



# A Controversial Claim: Redux



Jose-Luis Olivares/MIT

- We will never build a **reliable** exascale computer out of commodity parts
  - Without understanding our systems
  - Without tools to build resilient applications

# Takeaways

小不忍则乱大谋  
“A small leak will sink a great ship” or  
“Lack of forbearance in small matters  
upsets great plans”

- We *need* double chipkill DRAM for exascale
- We aren't going to get it via cheap market forces
- We aren't going to buy it
- What to do?
  - **Understand** with clarity
    - Research in collaboration with vendors
  - **Implement** with clarity
    - cross-cutting research and work with architects and systems people
    - cross-cutting research and work with code teams and users

# Takeaways

- We can get there from here, but not without understanding where we are, where we are going, and setting signposts along the way



<http://sploid.gizmodo.com/watching-uranium-emit-radiation-inside-a-cloud-chamber-1689997373/>



Uranium emitting radiation in a cloud chamber

# Data-Driven Decision Making in Resilience

Thank you for your attention!

Nathan DeBardeleben, Ph.D.  
Los Alamos National Laboratory  
High Performance Computing  
Ultrascale Systems Research Center Lead