UT-BATTELLE

The ExaChallenge Symposium:

Fault Tolerance Session

Christian Engelmann

Computer Science Research Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory, USA

Exascale Resilience Workshops in 2012

- **21-24. 2. 2012: Inter-Agency Workshop on Resilience at Extreme Scale, Catonsville, MD, USA**
- **6. 6. 2012: DOE Workshop on Fault Management for Exascale Systems, Baltimore, MD, USA.**
- **4-11. 8. 2012: ICiS Summer Workshop on Addressing Failures in Exascale Computing, Park City, UT, USA.**

- ***15. 11. 2012: Resilience BoF at SC - 12:15PM - 1:15PM:***
  - ***Overview presentations from the previous workshops***

Mission Critical Needs in Resilience

- US Department of Energy's mission focus on energy, national security and science
  - Materials aging in extreme environments
  - Simulations for national security
  - Advanced reactor simulations
  - Climate simulations
- Needs in resilience
  - Accuracy: Validation with bit-level reproducibility
  - Long run times: Weeks to months
  - Extreme-scale: Millions-to-billions of threads
- Trends in reliability
  - Less reliable components (reliance on COTS due to costs)
  - More components (end of frequency scaling ➔ node scaling)

C. Engelmann. Fault Tolerance Session. The ExaChallenge Symposium, Oct. 16-17, 2012.

Problem #1: Trends in Single Processor/ Memory Soft Error Vulnerability

	45 nm		11 nm		Power	Area
	Detected FIT	Undetected FIT	Detected FIT	Undetected FIT	Over-head %	Over-head %
No Additional Protection	**10**	**10-100**	**5,000**	**250-1,500**		
Better ECC	**10**	**10-100**	**250**	**1,400**	**< 5**	**~ 1**
DECTED ECC, Hardened latches & logic	**10-100**	**1-5**	**250-1,500**	**5-50**	**< 25**	**~ 20**

These numbers are rough estimates of soft errors induced by particle strikes based on growing device density with shrinking nanometer technology (not considering the impact of near threshold voltage).

Source: *Mattan Erez (UT Austin), Pradip Bose (IBM), Subhasish Mitra (Stanford), Dean Liberty (AMD), Paul Coteus (IBM). Addressing Failures in Exascale Computing Workshop, Park City, 2012. NOT PUBLISHED.*

Problem #1: Trends in Single Processor/ Memory Soft Error Vulnerability

	45 nm		11 nm		Power	Area
	Detected MTTF h	Undetected MTTF h	Detected MTTF h	Undetected MTTF h	Over-head %	Over-head %
No Additional Protection	**100M**	**10M-100M**	**200k**	**670k-4M**		
Better ECC	**100M**	**10M-100M**	**4M**	**714k**	**< 5**	**~ 1**
DECTED ECC, Hardened latches & logic	**10M-100M**	**200M-1B**	**670k-4M**	**20-200M**	**< 25**	**~ 20**

These numbers are rough estimates of soft errors induced by particle strikes based on growing device density with shrinking nanometer technology (not considering the impact of near threshold voltage).

Source: *Mattan Erez (UT Austin), Pradip Bose (IBM), Subhasish Mitra (Stanford), Dean Liberty (AMD), Paul Coteus (IBM). Addressing Failures in Exascale Computing Workshop, Park City, 2012. NOT PUBLISHED.*

Problem #1: Trends in 1,000,000 Processor/ Memory Soft Error Vulnerability

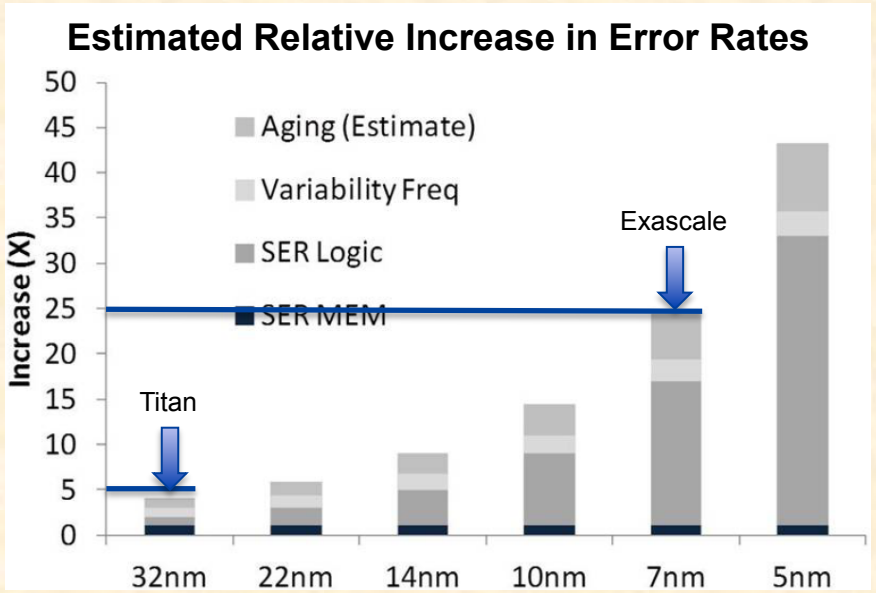
	45 nm		11 nm		Power	Area
	Detected MTTF h	Undetected MTTF h	Detected MTTF h	Undetected MTTF h	Over-head %	Over-head %
No Additional Protection	**100**	**10-100**	**0.2**	**0.67-4**		
Better ECC	**100**	**10-100**	**4**	**0.714**	**< 5**	**~ 1**
DECTED ECC, Hardened latches & logic	**10-100**	**200-1,000**	**0.67-4**	**20-200**	**< 25**	**~ 20**

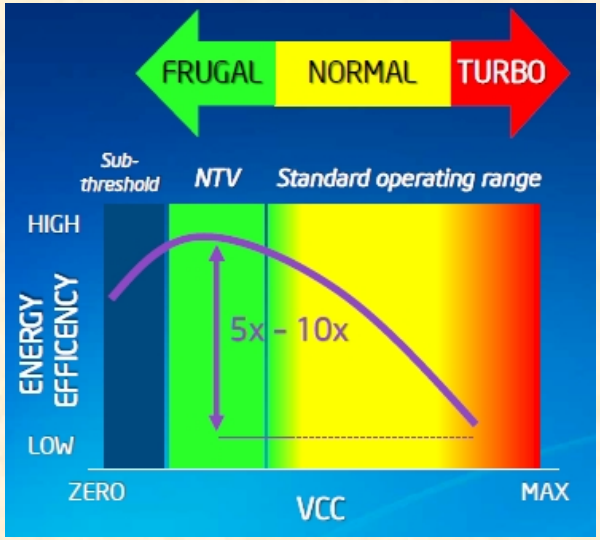
These numbers are rough estimates of soft errors induced by particle strikes based on growing device density with shrinking nanometer technology (not considering the impact of near threshold voltage).

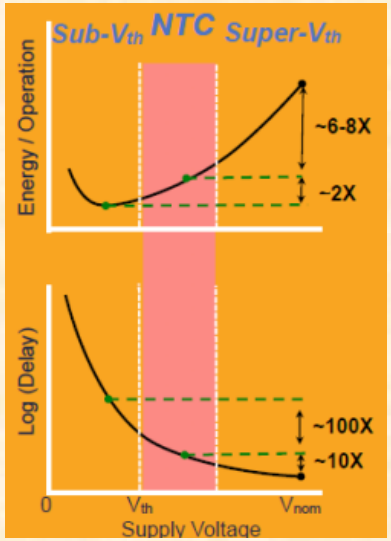
Source: *Mattan Erez (UT Austin), Pradip Bose (IBM), Subhasish Mitra (Stanford), Dean Liberty (AMD), Paul Coteus (IBM). Addressing Failures in Exascale Computing Workshop, Park City, 2012. NOT PUBLISHED.*

Problem #1: Less Reliable Components

- 7-10 nm process technology for exascale in 2020
  - Unknown aging effects
  - Increased variability
  - Increased soft error vulnerability
- Near-threshold voltage to achieve energy efficiency
  - Increased soft error vulnerability
  - Decreased noise immunity



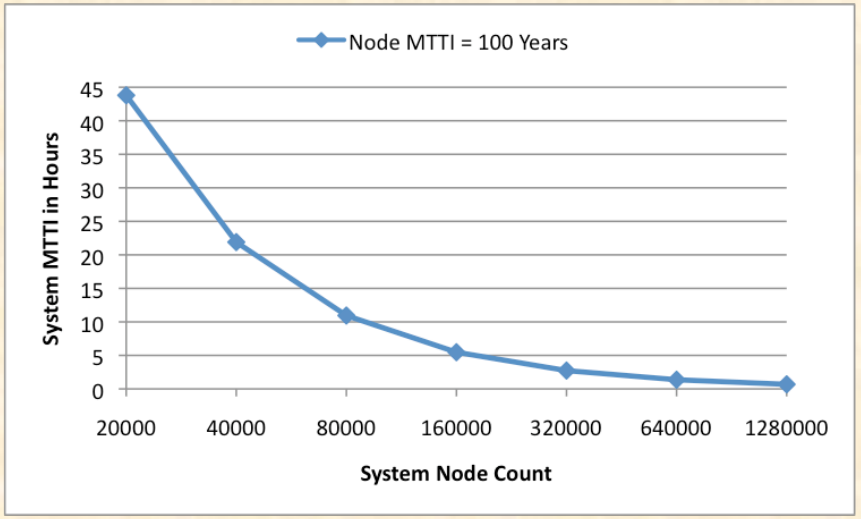


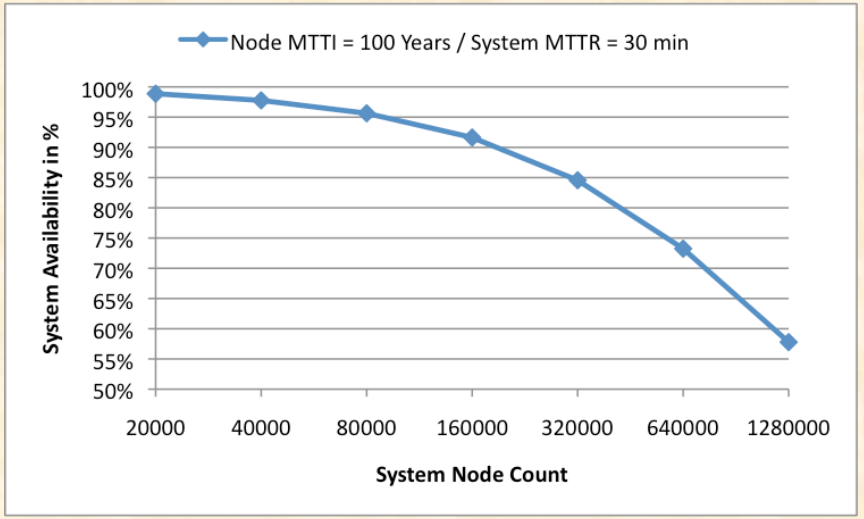


C. Engelmann. Fault Tolerance Session. The ExaChallenge Symposium, Oct. 16-17, 2012.

Problem #2: More Components

- System MTTI decreases as component count increases
- MTTI stayed roughly the same for the last 10 years
  - ~40x more cores/nodes
- Node reliability has been improved accordingly
  - Removed disks and fans
  - Improved system software
- No room to improve node reliability (see Problem #1)





HPC Resilience Solutions

- Application-level checkpoint/restart to a parallel file system is the current standard
- Advanced resilience solutions
  - System-level and incremental/differential checkpoint/restart
  - Checkpoint/restart to memory/SSDs in neighbor or I/O nodes
  - Uncoordinated checkpoint/restart with message logging
  - Fault tolerant MPI and algorithm-based fault tolerance
  - Proactive fault tolerance (migration-based fault avoidance)
  - Rejuvenation (reboot/refresh to clear latent errors)
  - Process and data-level redundancy (DMR and software ECC)
- Only system-level checkpoint/restart is used in production
- None of the other solutions are even close to production

HPC Resilience Gaps

- Local checkpointing techniques are required in the short term to improve application-level checkpoint/restart
- Advanced resilience techniques can be deployed in the long term to mitigate risks (based on costs)
- The resilience problem is still not well understood
  - Errors types, rates, fault root causes, and propagation
  - For current and future systems
  - Cost trade-offs: Power, resilience, performance, deployment
  - Cross-layer fault models and interfaces
  - Standard test suite and metrics to stress resilience solutions and compare them fairly

Discussion

- What fault types and frequencies should be expected?
- Can evolutionary fault tolerance approaches provide resilience, or are more revolutionary concepts needed?
- Which layer (hardware, OS, runtime, and/or application) is responsible for assuring resilience?
- What are the performance, resilience, power, and deployment cost trade-offs at exascale?
- Do we need standards for HPC resilience terms, metrics, methods, and APIs?
- Does the local OS need to be resilient (in addition to the global OS) ?