Resiliency for Reliability – Myths and Truths

Shekhar Borkar
Intel Corp.
Salishan Conference
April 28, 2015

This research was, in part, funded by the U.S. Government, DOE and DARPA (UHPC, FF1, X-Stack, FF2, CREST). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

Outline

- Resiliency defined
- Faults, errors, and effects
- Soft-errors
- Permanent faults
- Resiliency framework
- Summary

Resiliency

Definition:
Asymptotically provide the reliability of a tri-modular redundancy scheme with only 10% energy and HW cost

State of the art:

Technique	Coverage
Parity, ECC	Memory only, Soft Errors, Erratic bits
RAZOR	State machines, SER, temporal variations
Residue logic	Static logic only, permanent faults
Redundant execution	Memory, RF, SER only
…	…

$\sum$ Cost > Cost(tri-modular redundancy) ?

Resiliency is NOT:
A solution to error prone shabby engineering!

Resiliency —Three steps

❶ Understand faults

Different types of faults

Frequency of occurrence, probability, and time to error

Behavior now, and in the future

❷ Understand impact of faults

Errors caused by the faults (observe)

Diagnose and pinpoint the fault location

Recover from the error, correct the fault

Impact on system performance, energy,…

❸ Unified resiliency framework

Common, serves all types of faults

Cost (Resiliency) << Cost(TMR)

Understanding Faults

Types of Fault	Examples, Effect	Action
Permanent faults	Fan, power supply, shorts and opens	Sensors for detection Node down
Gradual spatial faults (Process variations)	Variations in frequency Exacerbated at NTV	Design out Costs perf & energy
Gradual temporal faults (temp variation with load)	Temperature increase causing frequency loss	Design out Costs perf & energy
Intermittent faults	Data corruption by noise, Soft errors, control loss Not reproducible	Creative accounting
Slow degradation (Aging Faults)	Frequency loss Erratic bits in memory	Design out Costs perf & energy

1. **Probability of fault (lower is better), and**
2. **Time to error from fault (larger is better)**

Source: John Daly, David Mountain's NSA Resiliency WS 02/2012

Probability of Faults & Time to Error

Fault	Probability	T to Error	Action
Fans	High	Medium	Node down
Power Supply	High	Medium	Node down
CPU / SRAM	Very Low	Small	Node down
DRAM	Medium	Large	Reconfiguration
Solder Joints	Med-High	Small	Node down
Sockets	Med-High	Small	Node down
Disks	Mid to High	Large	Reconfiguration
NAND/PCM	Low-Mid	Large	Reconfiguration
Soft Errors	Low	Small	Clever accounting

Source: John Daly, David Mountain's NSA Resiliency WS 02/2012

Deeply Scaled Technologies

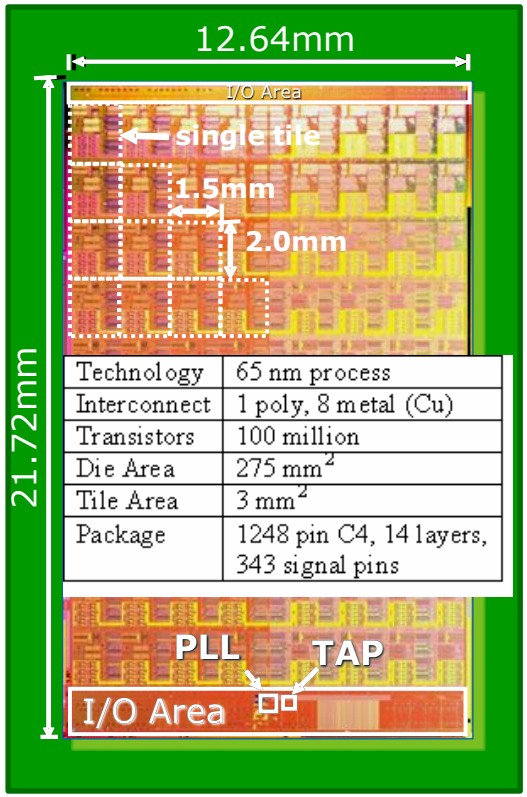
- **Myth:**

Failure rate will increase with deep scaling

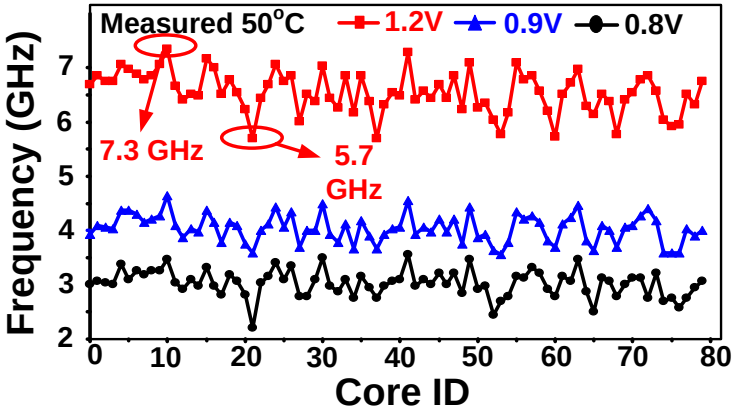
- **Truth: (near future, thermionic devices, CMOS…)**
  - Scaling will continue with acceptable failure rate
  - But compromising performance and energy
  - If the system level resiliency allows increased failures…
  - Then the technology can be aggressive
  - Benefits performance and energy
- **Beyond CMOS? (far future)**
  - Probabilistic?

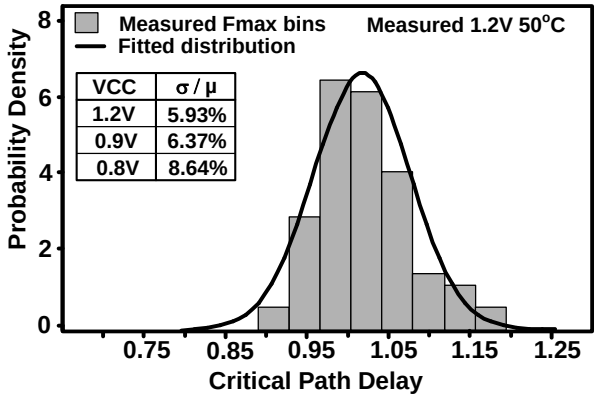
Process Variations—Spatial, Gradual Faults

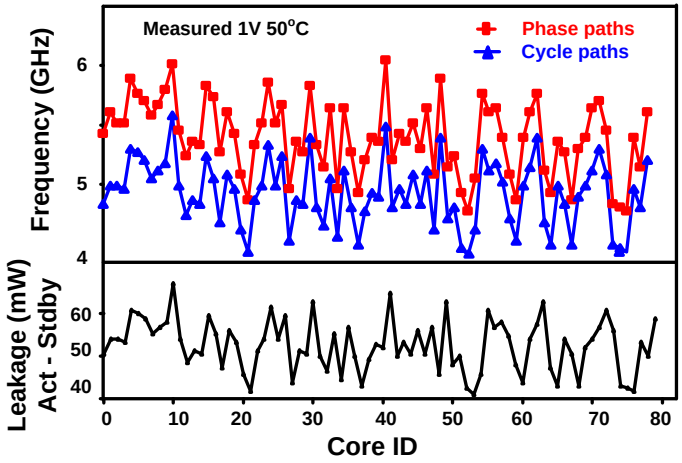
80-core research testchip

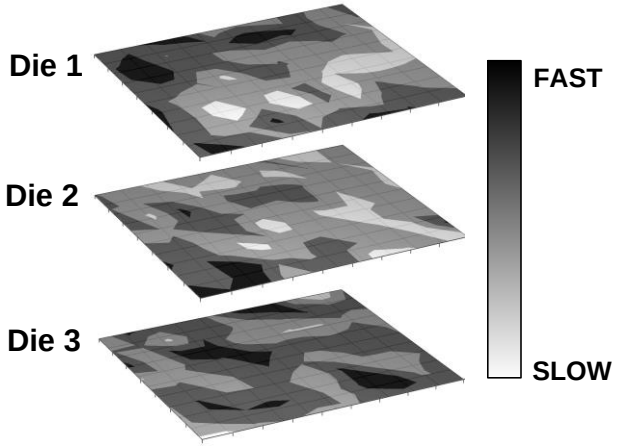


Technology	65 nm process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 million
Die Area	275 mm^2
Tile Area	3 mm^2
Package	1248 pin C4, 14 layers, 343 signal pins





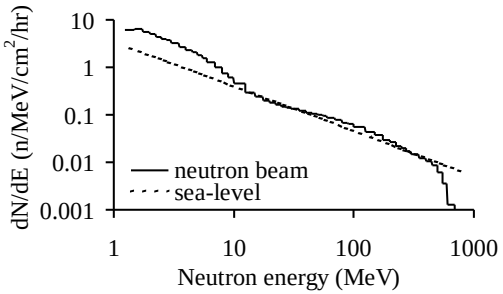




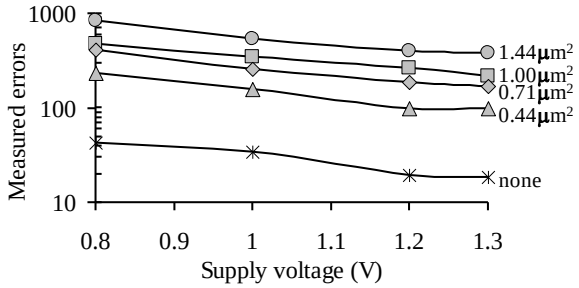
Within-die and die-to-die variation impacts much higher at lower voltages

Resiliency must address spatial, gradual, and temporal faults

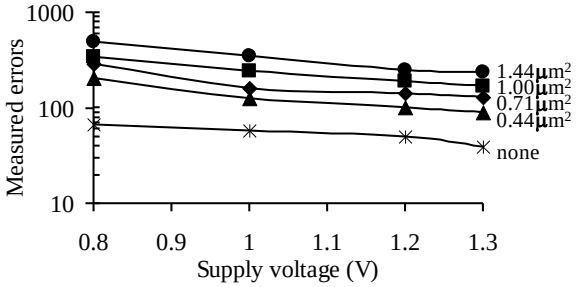
Soft Errors—Intermittent Faults



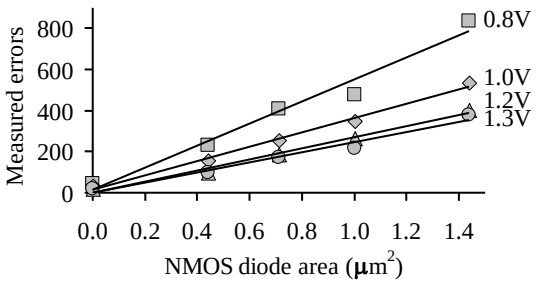
Beam energy spectrum compared to sea level.



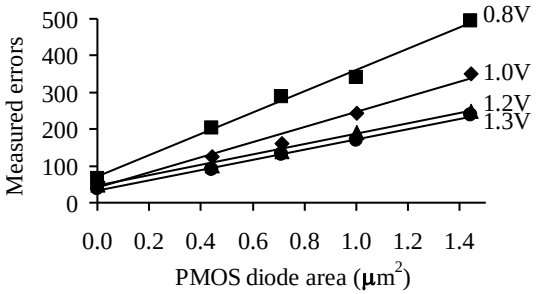
NMOS diode SER dependency on voltage.



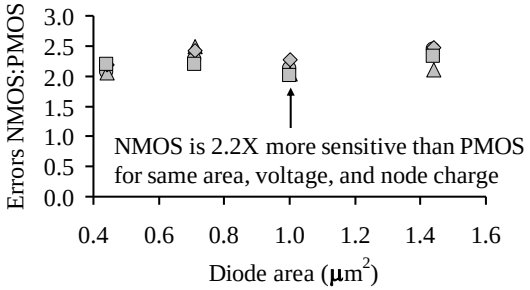
PMOS diode SER dependency on voltage.



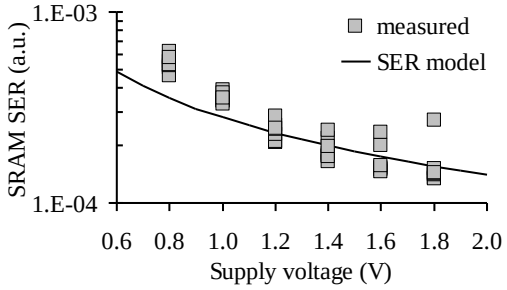
SER dependency on NMOS diode area.



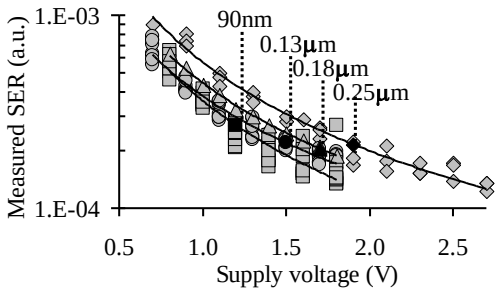
SER dependency on PMOS diode area.



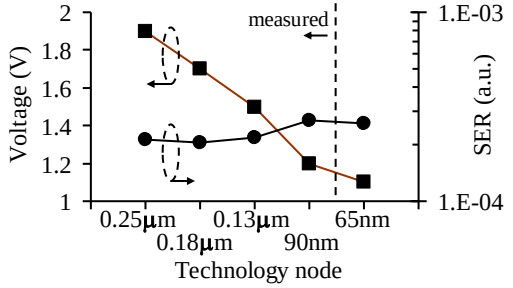
Relative SER of NMOS and PMOS diodes.



Comparison of SRAM SER calculated from a calibrated 90-nm model and measured SER.

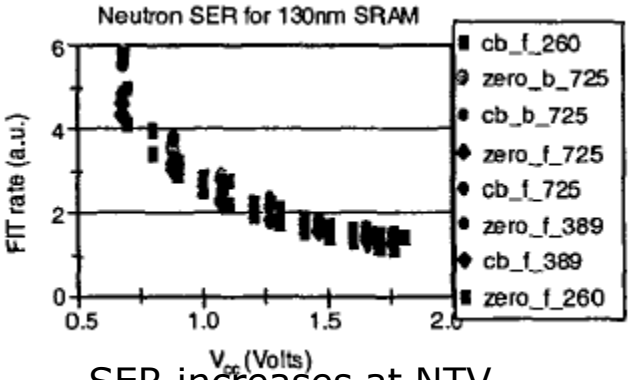


Voltage scaling of neutron SER in SRAM.

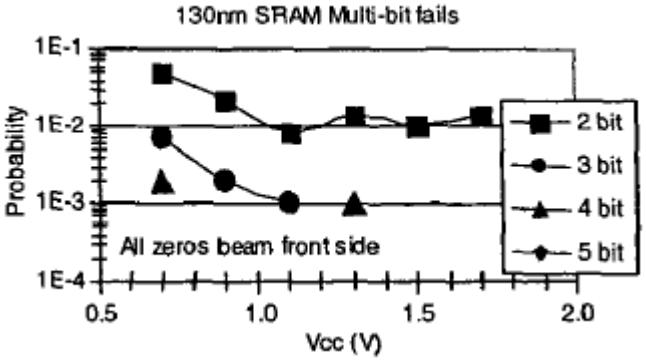


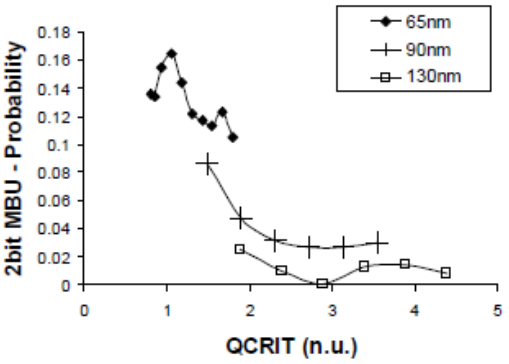
Technology scaling of neutron SER in SRAM.

Other Results from Literature



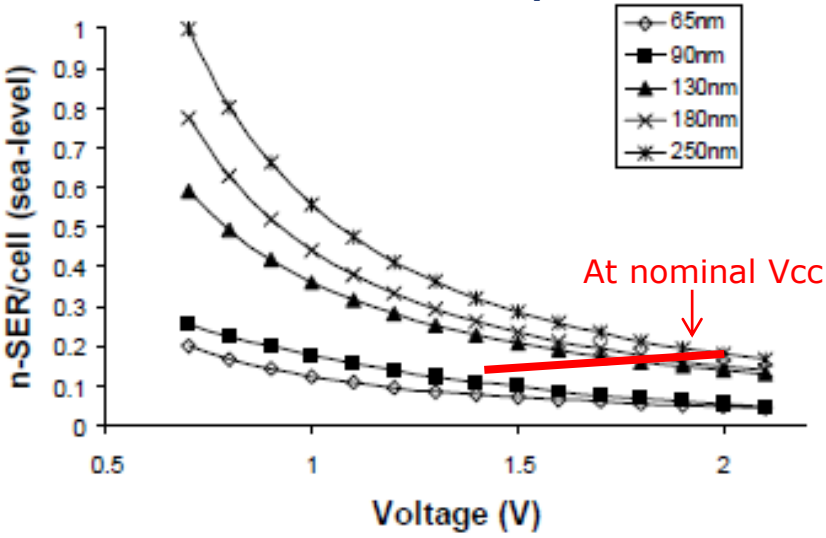
SER increases at NTV





Multi-bit failures become worse at NTV

SER/SRAM bit reduces with scaling



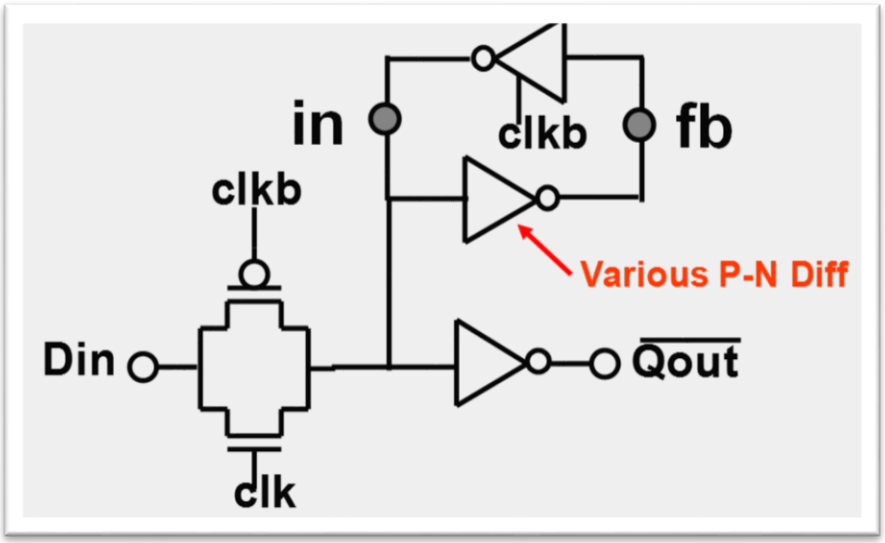
Ratio is fairly constant

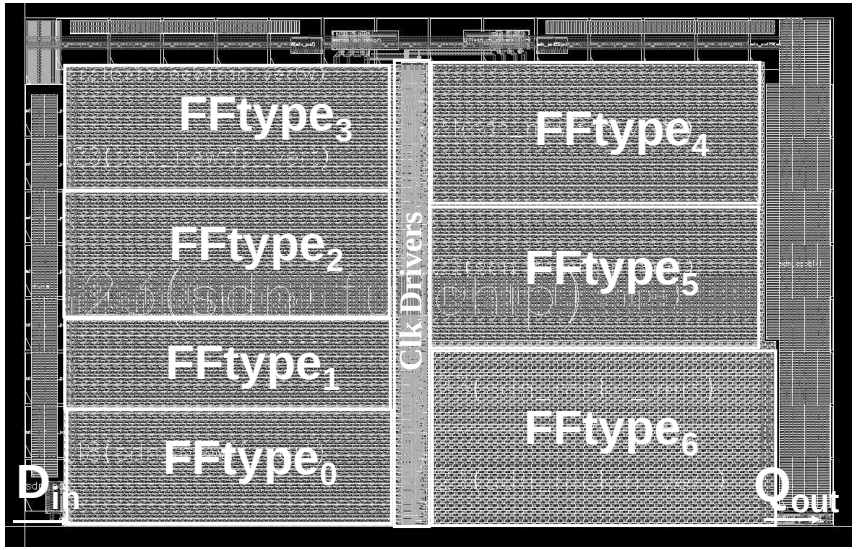
$$SER(QCRIT) \propto A_{diff} \exp(-QCRIT / QCOLL)$$

Diffusion area reduces with scaling

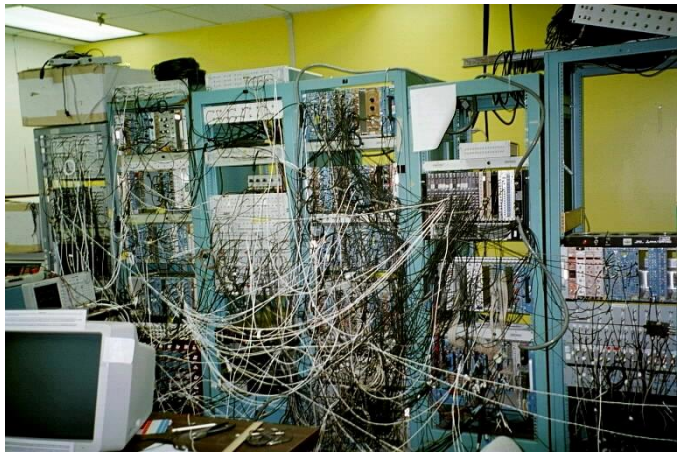
1. SER/bit may reduce with scaling, but system level SER will continue to get worse
2. SER sensitivity to reduce supply voltage (NTV) needs better understanding
3. Multi-bit errors will become worse and need attention

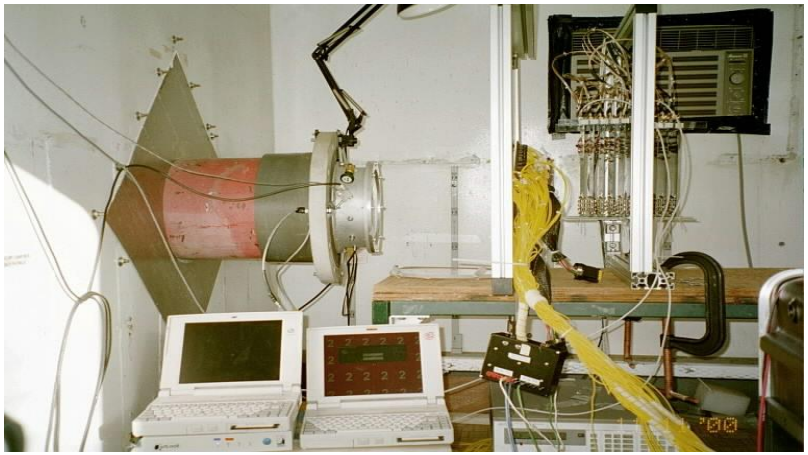
Experiments (180, 130, 90nm)

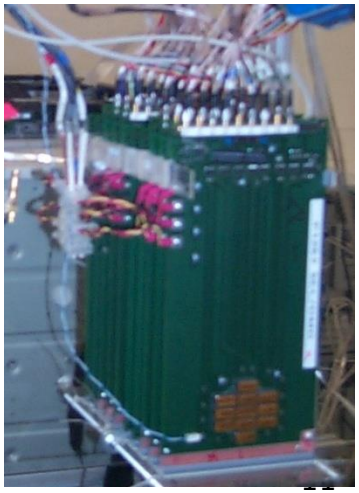




130nm: 8490 FF * 22 dies * 10 boards = 1.87million

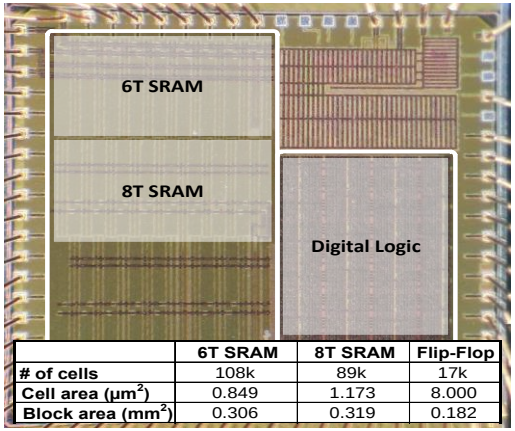




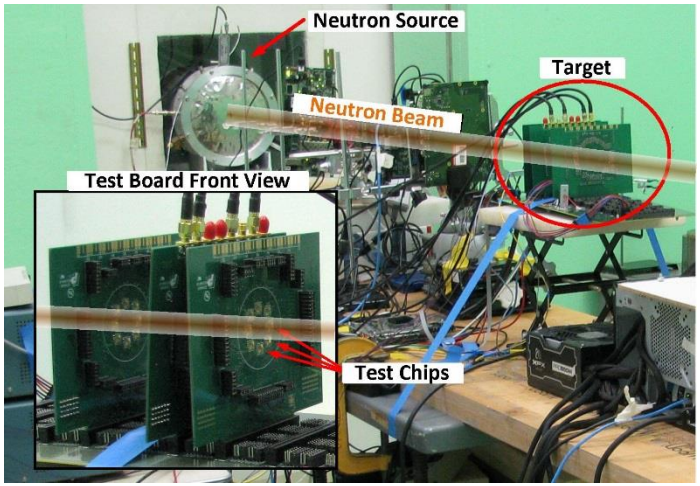


Recent 65 nm Experiments

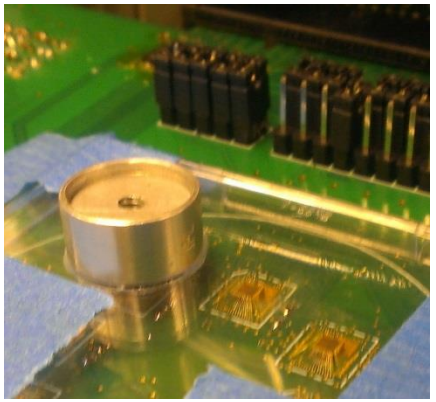
Tapeout: May-2013
Debug: Aug-2013
Los Alamos: Sep-2013
OSU Nuclear-Eng: Nov-2013



	6T SRAM	8T SRAM	Flip-Flop
# of cells	108k	89k	17k
Cell area (μm^2)	0.849	1.173	8.000
Block area (mm^2)	0.306	0.319	0.182



Neutron Beam (memory)

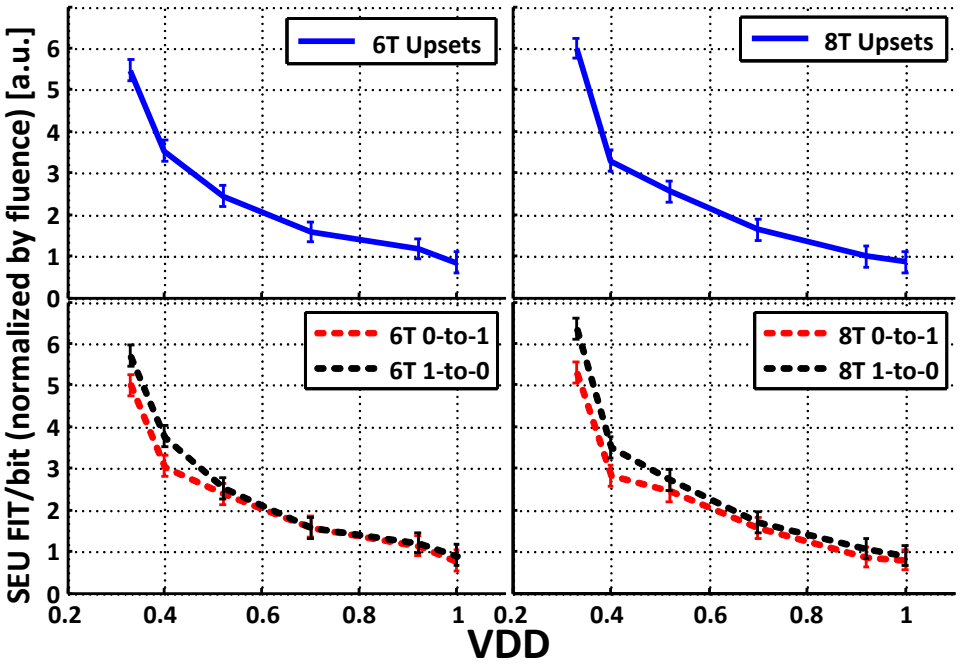


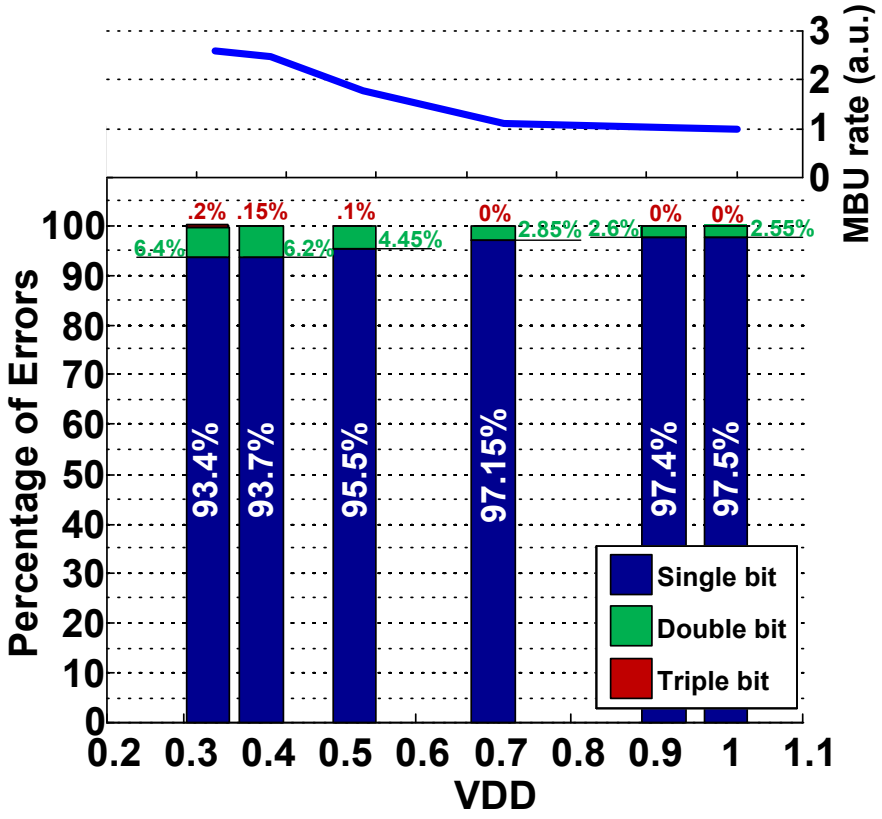
Alpha source (logic)

R. Pawlowski et al, "Characterization of Radiation-Induced SRAM and Logic Soft Errors from 0.33V to 1.0V in 65nm CMOS", CICC, 2014

Acknowledgement: DARPA funded CREST project, Oregon State University, Prof Patrick Chiang, Robert Pawlowski, Joe Crop, and LANL (Nathan et al).

65 nm SRAM Results





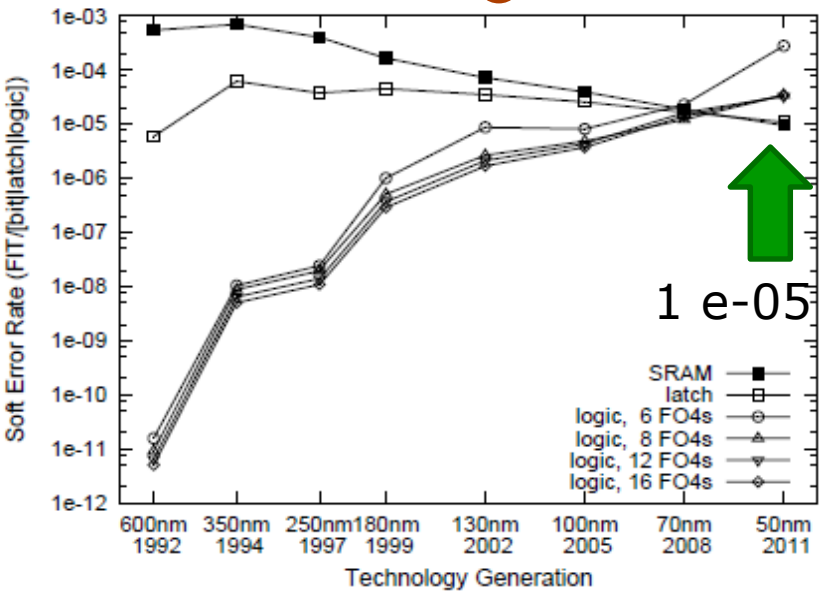
NTV exacerbates SRAM SER, multi-bit errors increase

Acknowledgement: DARPA funded CREST project, Oregon State University, Prof Patrick Chiang, Robert Pawlowski, Joe Crop and LANL (Nathan et al).

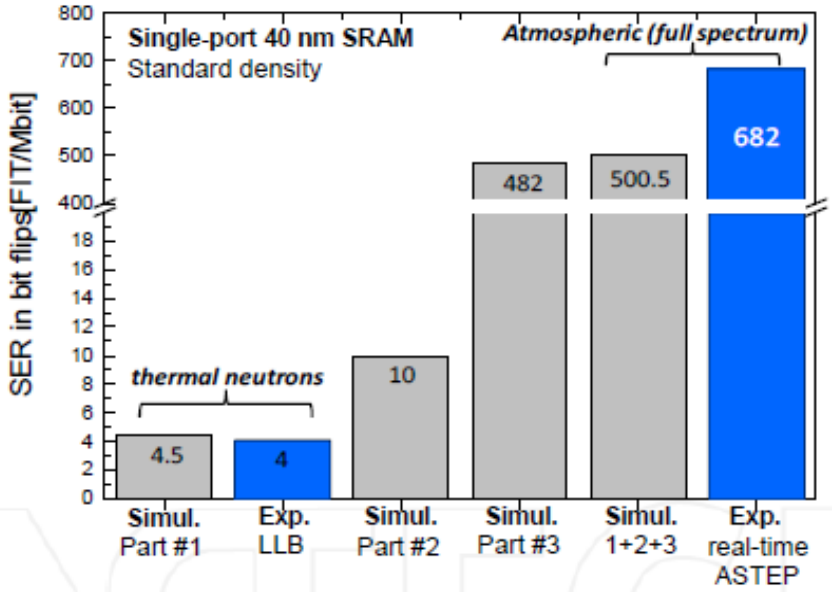
Soft Error FIT Rate (Neutrons)

1 e-03

❶



❷



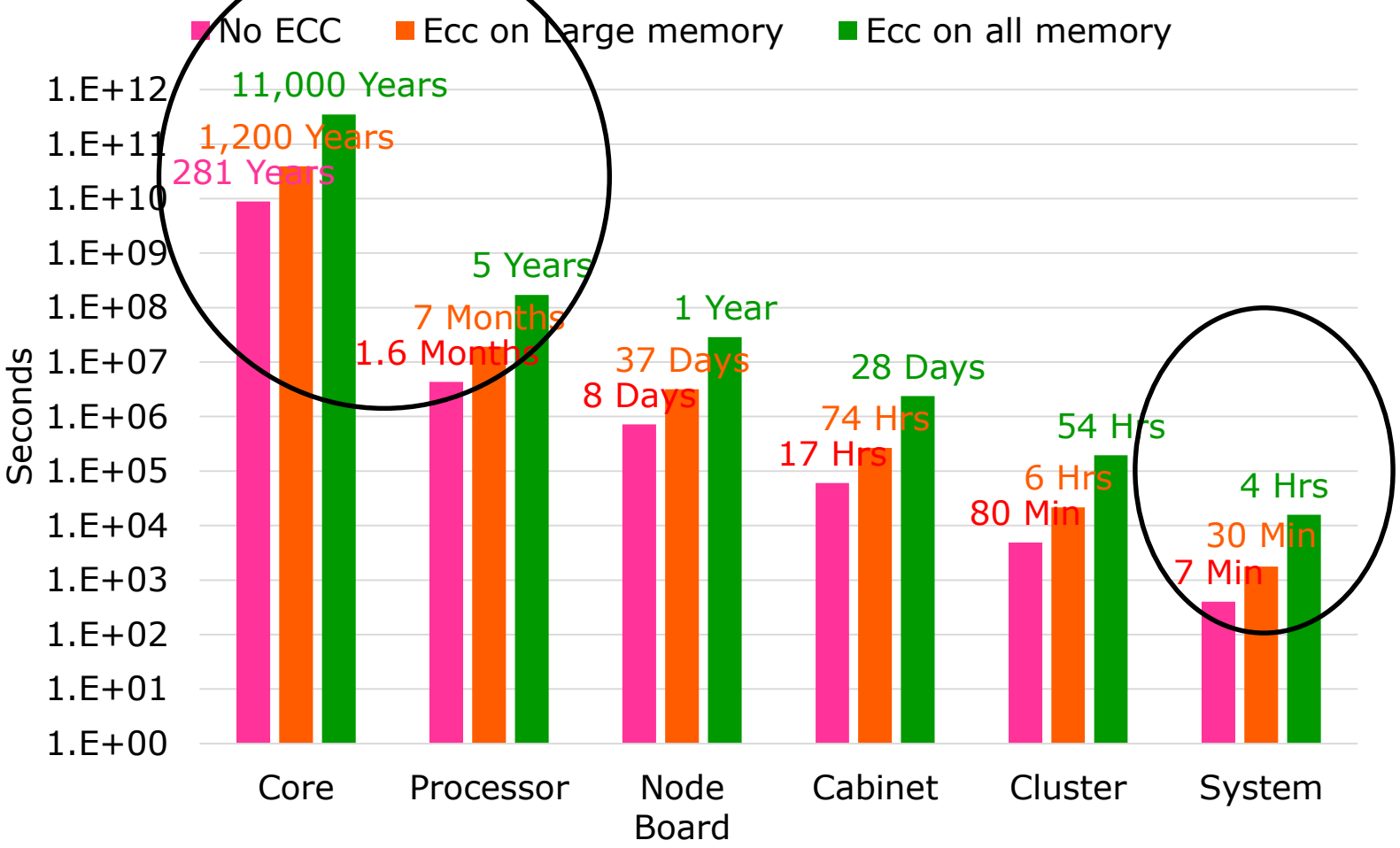
❸ Xilinx published data suggests 9e-05 (N) + 4.5e-05 (α) = 1.3e-04
http://www.xilinx.com/support/documentation/user_guides/ug116.pdf on page 28

Assume: FIT Rate for SRAM & FF ~ 1e-04 with 10X uncertainty

❶ P. Shivakumar et. al, "Modeling the Effect of Technology Trends on the Soft Error Rate of Combinational Logic", Proceedings of the 2002 International Conference on Dependable Systems and Networks

❷ Numerical Analysis and Scientific Computing, "Numerical Simulation - From Theory to Industry", book edited by Mykhaylo Andriychuk, ISBN 978-953-51-0749-1, Published: September 19, 2012

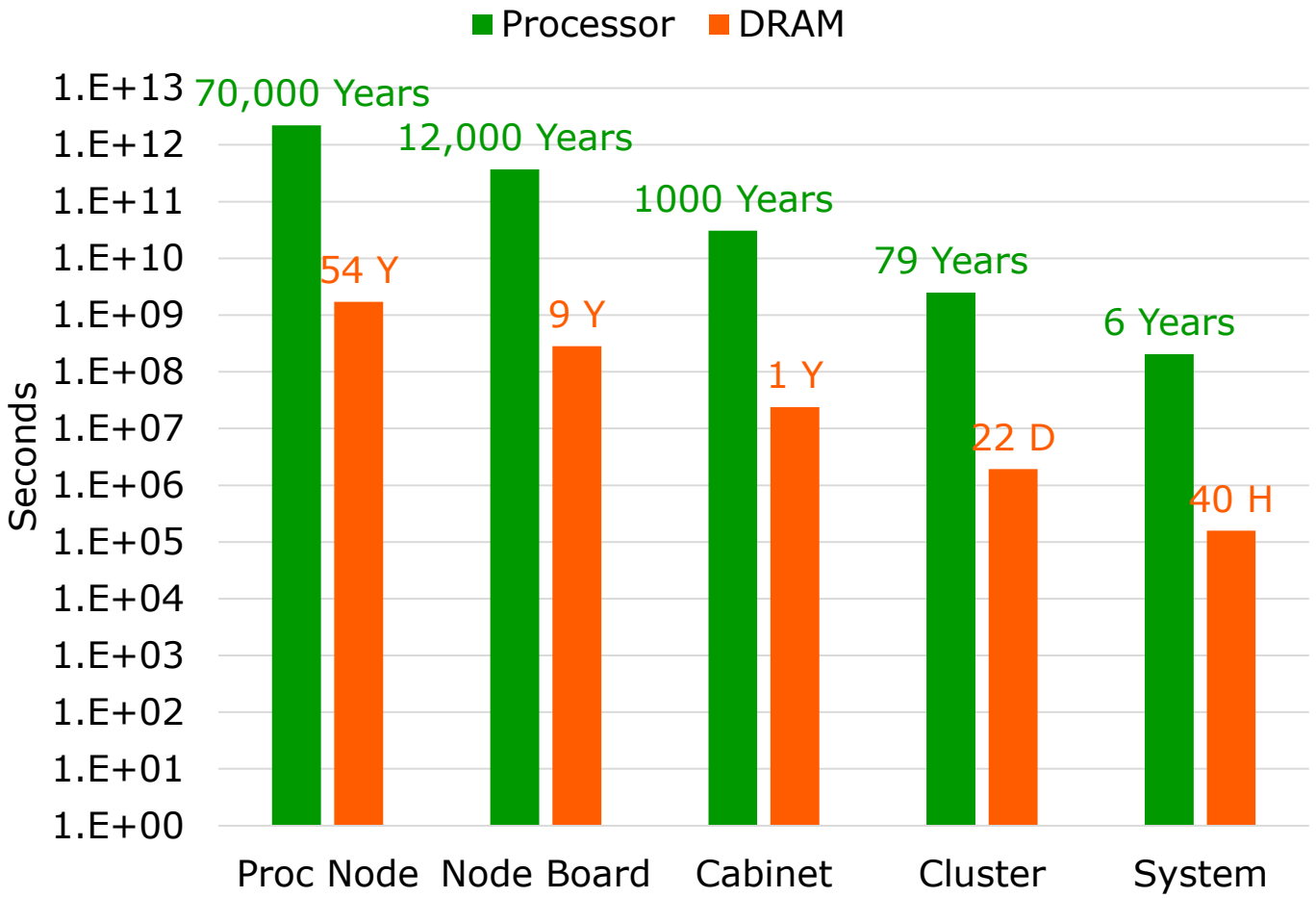
Mean Time to Soft Error (Exascale)



Myth: Soft Errors are frequent

Truth: Not if they are confined

Permanent Failure Rate (VLSI Chips)



VLSI Chips are highly reliable; DRAMs more fragile
Both are VLSI, so why the difference?

V. Sridharan et. al, "A Study of DRAM Failures in the Field", SC12

Resiliency Framework Assumptions

Faults occur (relatively) infrequently, cause errors (observable)

- Diagnosis & corrective actions, do not impact performance and energy (much)

Only one fault occurs at any time in the confined area

Time to service an error or diagnose a fault is small

- Mean time to a fault is much larger than the time it takes to service a fault; assumes convergence

Fault isolation, confinement, reconfiguration, recovery and adaptation—all done in the system software (R-manager)

All levels in the stack, from Applications down to Circuits need to participate

- Error detection in hardware. Diagnosis, recovery using software

Reactive and Proactive Majors

Reactive major

- Detect error in hardware
- Resiliency manager (system SW) notified
- Isolate the fault (where did it happen?)
- Confine the fault (it does not impact other HW)
- Recover, reconfigure if necessary, and adapt

Proactive major

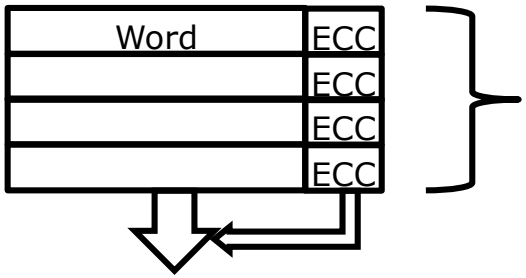
- Continually test the hardware (once a day, week?)
- When energy is available and not in performance critical path
- Detect marginalities, reconfigure hardware as necessary

Hierarchical, incremental check-pointing for recovery

- Check-pointing and recovery scheme determined by mean time to fault

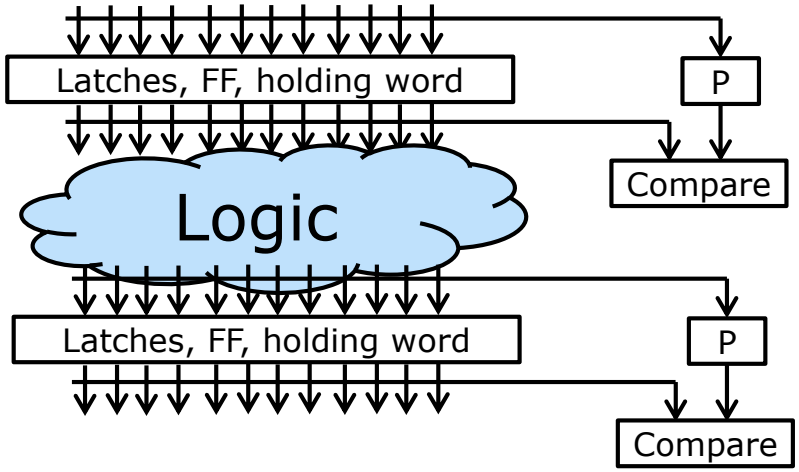
Simple Detection Hardware

❶ Parity/ECC covered memory with notification



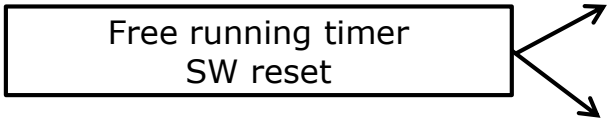
Notify where the error was detected, statistics made available

❷ Parity covered datapath (not just data, but any ensemble of bits)



Notify error and location

❸ Watch-dog timers (state-machine hangs)

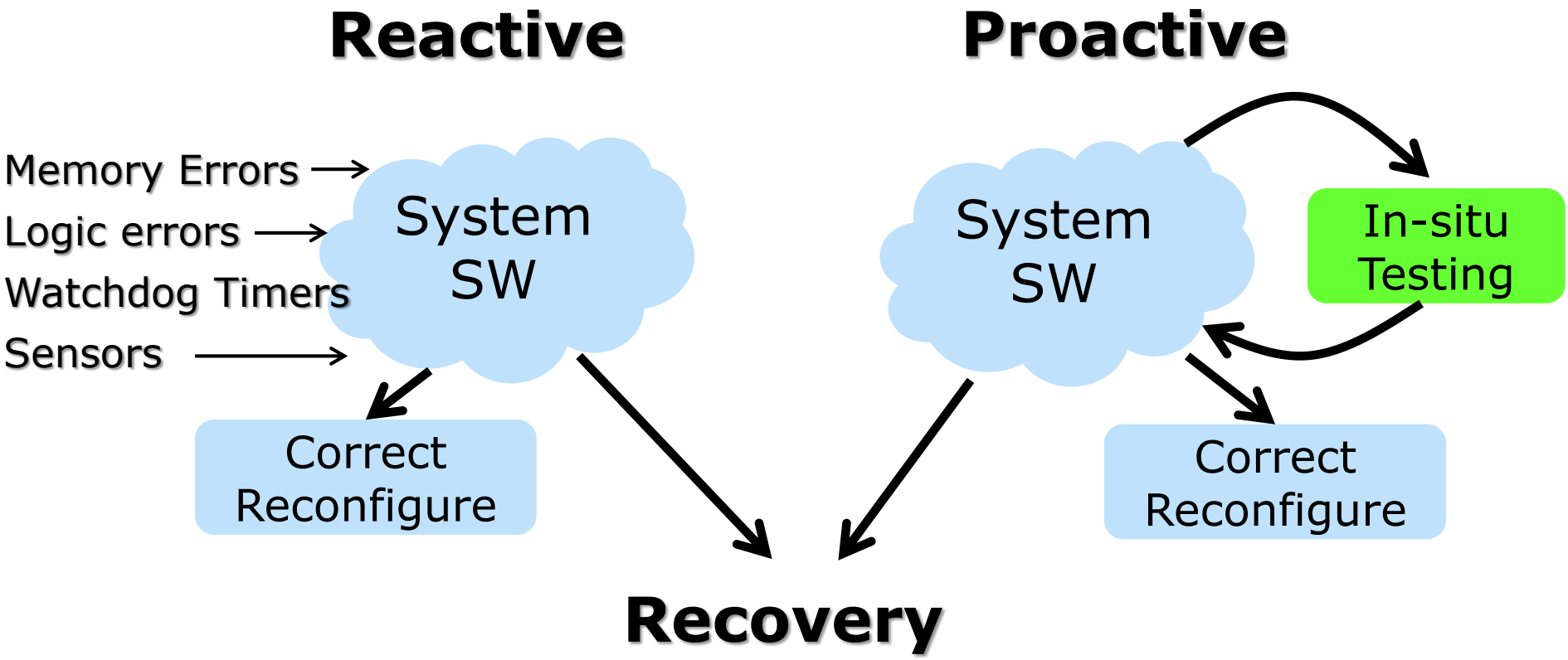


Notify neighbors upon overflow

❹ Sensors everywhere—Fans, Power supplies,…

Cost ~3% die-area & power

SW for Diagnosis and Recovery



Strategy depends on mean time to fault (T)
For large T, traditional check-pointing may be good enough
For small T, incremental, hierarchical check-pointing

❶ System SW, ❷ Test, ❸ Recovery

Proposed Check-pointing & Recovery

Confinement, state-store, and recovery based on:

1. **Type of fault,**
2. **Probability of fault, and**
3. **Time to error**

Fault	Probability	T to Error
Fans	High	Medium
Power Supply	High	Medium
CPU / SRAM	Very Low	Small
DRAM	Medium	Large
Solder Joints	Med-High	Small
Sockets	Med-High	Small
Disks	Mid to High	Large
NAND/PCM	Low-Mid	Large
Soft Errors	Low	Small

Small T to error

Smaller confinement (Core level)

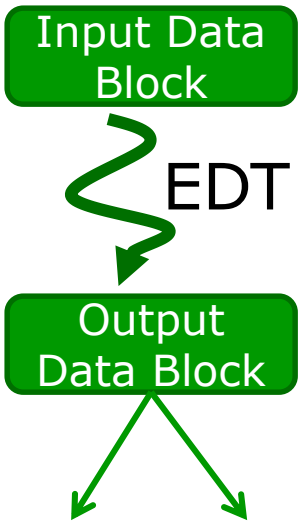
Reactive measure:

Detect in HW

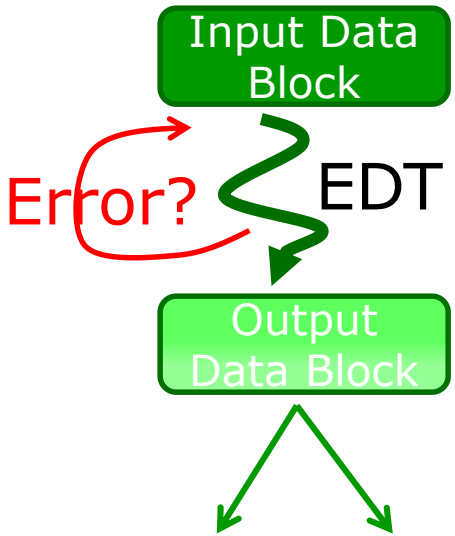
Harmonize with system SW (Exec Model) to recover

Harmonizing with Execution Model

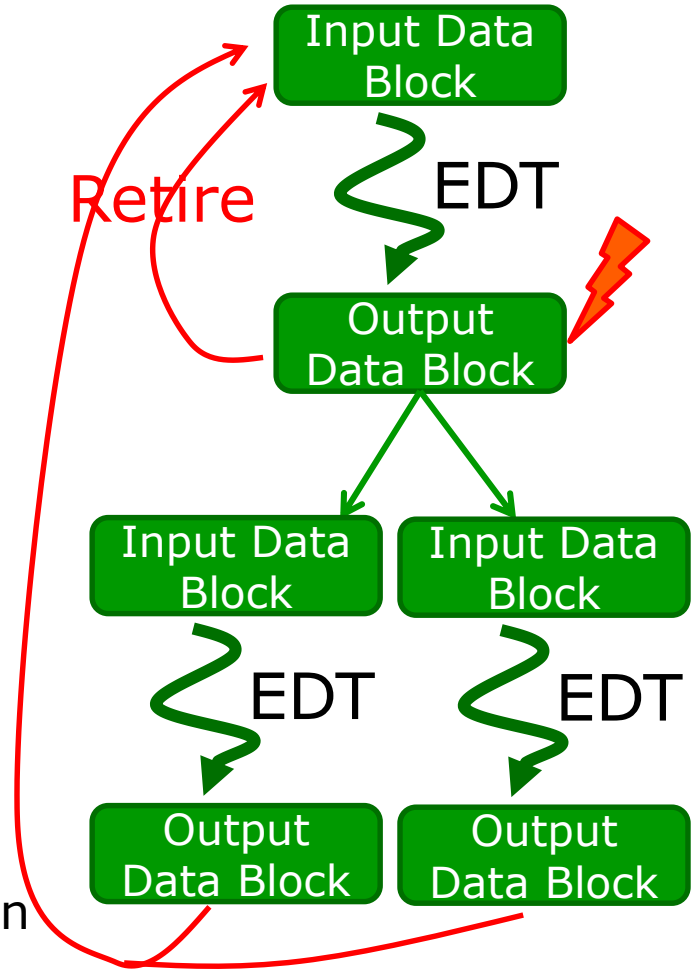
Event Driven Tasks (EDT)



Non-preemptive completion
Invoke dependent threads



Invoke dependent threads only after error free completion



Implemented in Open Community Runtime (OCR)

Proposed Check-pointing & Recovery

Confinement, state-store, and recovery based on:

1. **Type of fault,**
2. **Probability of fault, and**
3. **Time to error**

Fault	Probability	T to Error
Fans	High	Medium
Power Supply	High	Medium
CPU / SRAM	Very Low	Small
DRAM	Medium	Large
Solder Joints	Med-High	Small
Sockets	Med-High	Small
Disks	Mid to High	Large
NAND/PCM	Low-Mid	Large
Soft Errors	Low	Small

Sensors detect and notify

Larger confinement (Node, socket, board)

Large T to error

Reactive measure:

Store EDT states for re-execution and recovery

Proposed Check-pointing & Recovery

Confinement, state-store, and recovery based on:

1. **Type of fault,**
2. **Probability of fault, and**
3. **Time to error**

Fault	Probability	T to Error
Fans	High	Medium
Power Supply	High	Medium
CPU / SRAM	Very Low	Small
DRAM	Medium	Large
Solder Joints	Med-High	Small
Sockets	Med-High	Small
Disks	Mid to High	Large
NAND/PCM	Low-Mid	Large
Soft Errors	Low	Small

Large T to error

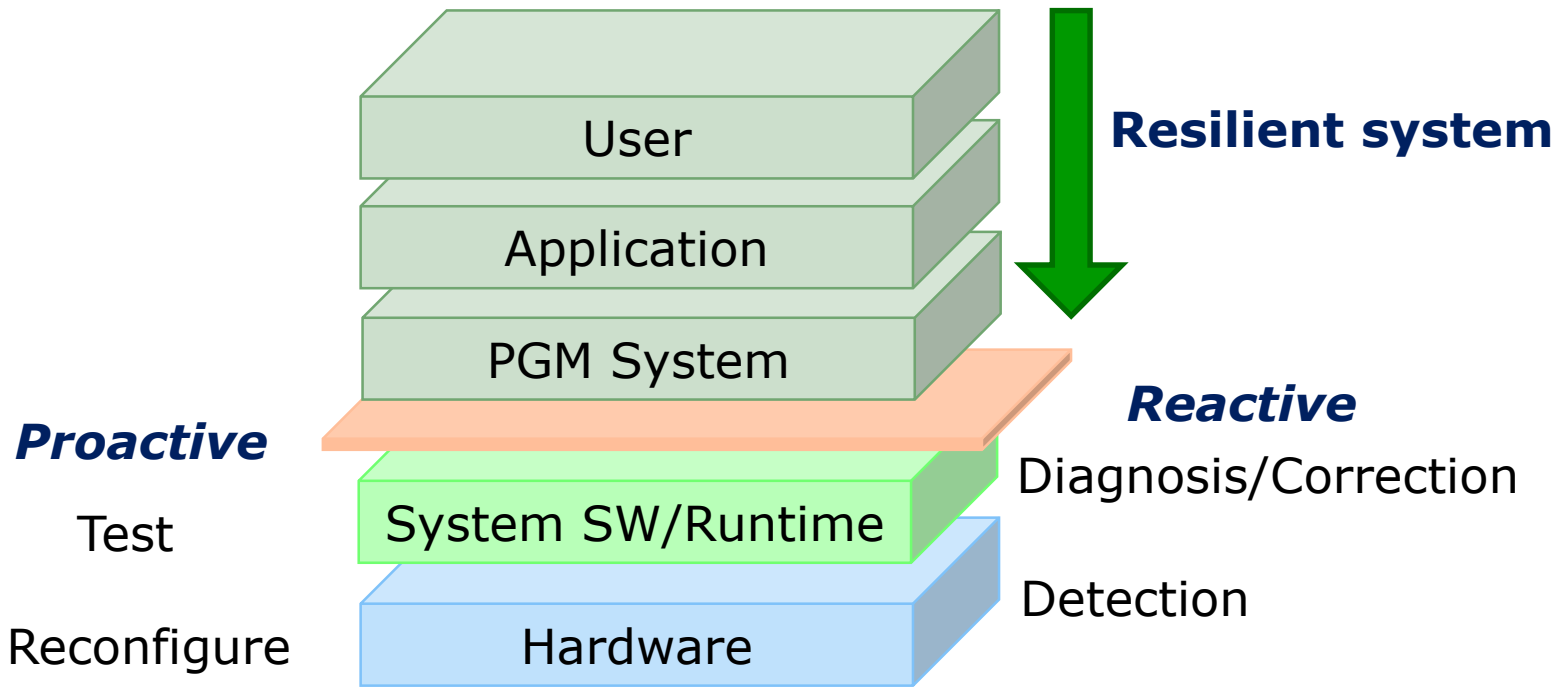
Smaller confinement (Node)

Proactive measure:

Detect marginality

Decommission node to replace component

User Experiences Reliable System



Summary

- **Understand faults**
- **Resiliency framework covering all types of faults**
- **Detection in HW, diagnosis and correction in system SW**
- **Then devise recovery scheme(s) considering all of the above**

References

1. J. Autran, et. al., Real-time Soft-Error testing of 40nm SRAMs, *2012 IEEE International Reliability Physics Symposium (IRPS)*, Page(s): 3C.5.1 - 3C.5.9

2. P. Hazucha, et. al., Measurements and analysis of SER-tolerant latch in a 90-nm dual-Vt CMOS process, *IEEE Journal of Solid-State Circuits,* Volume 39, Issue 9, Sept. 2004 Page(s):1536 – 1543

3. S. Dighe, et. al., Within-Die Variation-Aware Dynamic-Voltage-Frequency-Scaling With Optimal Core Allocation and Thread Hopping for the 80-Core TeraFLOPS Processor, *IEEE Journal of Solid-State Circuits*, Volume: 46, Issue: 1, Jan 2011 Page(s): 184 – 193

4. K. Bowman, et. al., A 45 nm Resilient Microprocessor Core for Dynamic Variation Tolerance, *IEEE Journal of Solid-State Circuits*, Volume: 46, Issue: 1, Jan 2011, Page(s): 194 – 208

5. P. Hazucha, et. al., Neutron Soft Error Rate Measurements in a 90-nm CMOS Process and Scaling Trends in SRAM from 0.25-pm to 90-nm Generation, IEDM 2003

6. J. Maiz, et. al., Characterization of Multi-bit Soft Error events in advanced SRAMs, IEDM 2003

7. N. Seifert, et. al., Radiation-Induced Soft Error Rates of Advanced CMOS Bulk Devices, 44th Annual International Reliability Physics Symposium, 2006

8. P. Shivakumar et. al, "Modeling the Effect of Technology Trends on the Soft Error Rate of Combinational Logic", Proceedings of the 2002 International Conference on Dependable Systems and Networks

9. V. Sridharan et. al, "A Study of DRAM Failures in the Field", SC12

10. V. Sridharan et. al, "Memory Errors in Modern Systems The Good, The Bad, and The Ugly", ASPLOS 15