Exceptional service in the national interest

Sandia National Laboratories

Revisiting Checkpointing for Exascale-Class Systems

Kurt B. Ferreira
1423 -- Scalable System Software
SNL

April 29, 2015

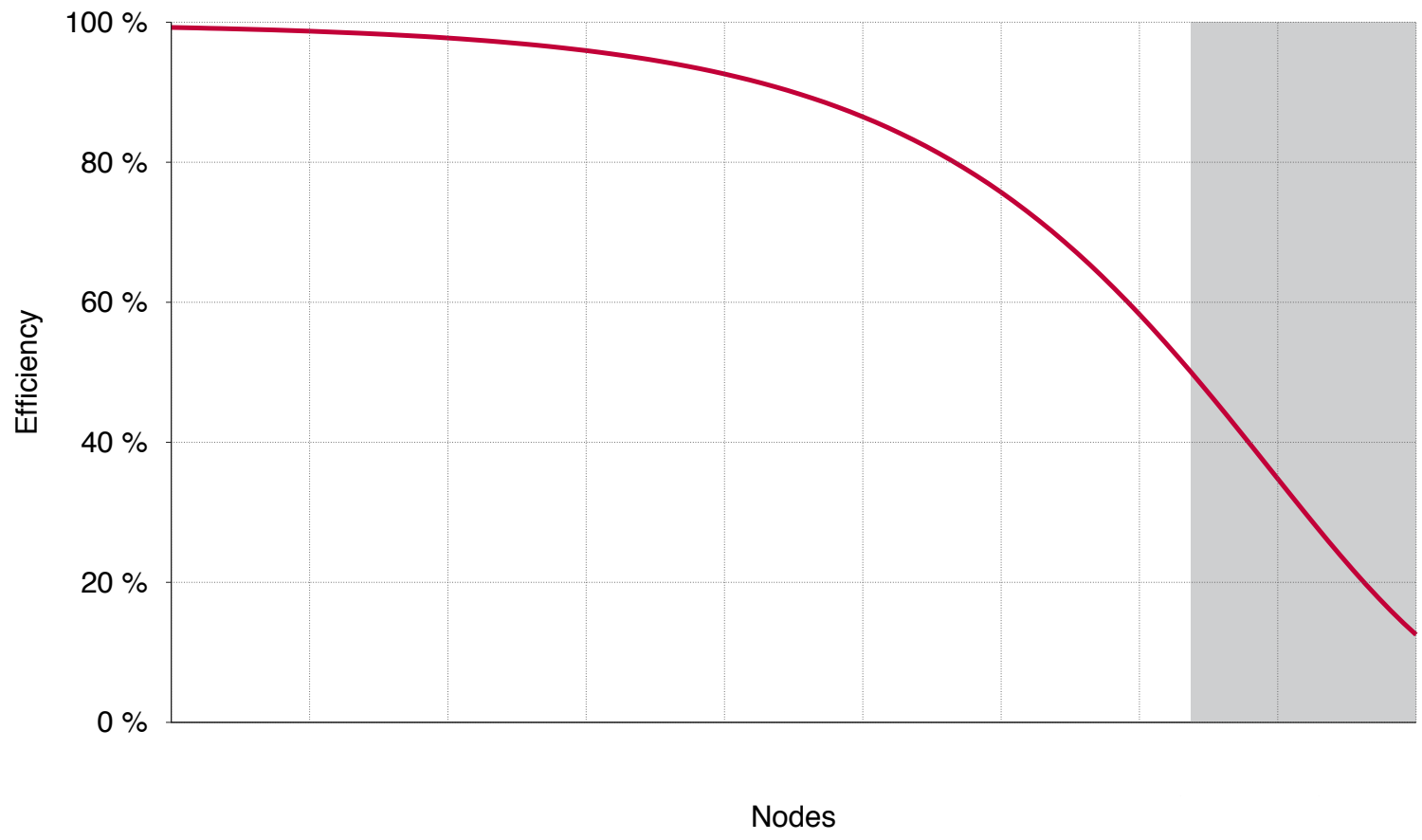
SAND 2015-3382 C

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Checkpointing is Dead !!!



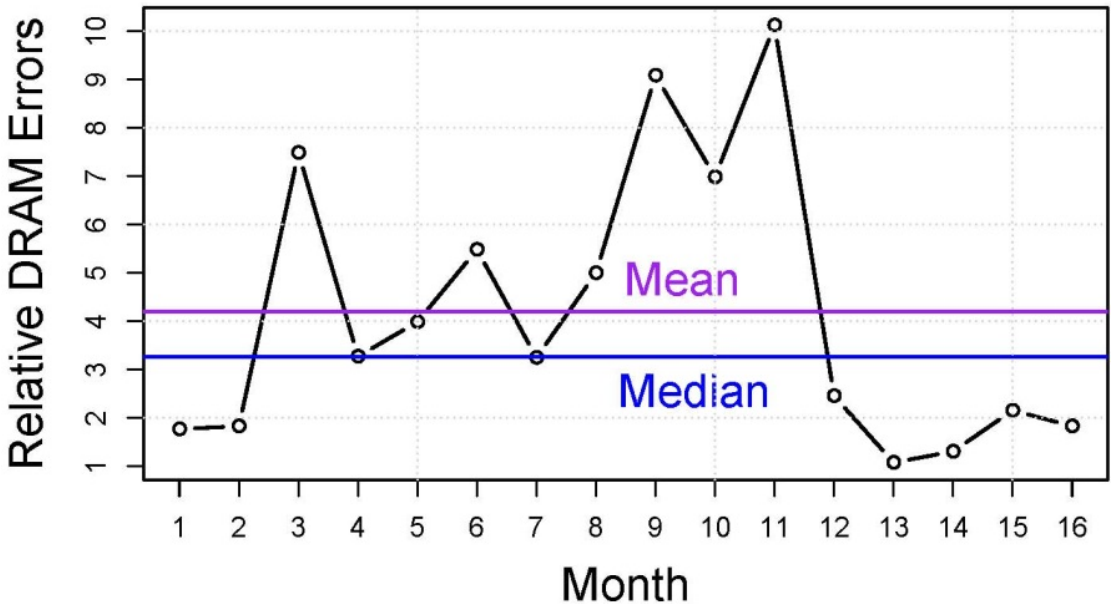
… but, there are flaws in the last plot

- Does not account for state-of-the-art in checkpointing:
  - **Storage Optimizations**: Burst-buffers, multi-level checkpointing (SCR), etc
  - **Checkpoint optimizations**: checkpoint compression, incremental checkpointing, etc
  - **Decreased fault rates**: SEC-DED Vs. Chipkill ECC DRAM protection
- Based on reliability analysis that is flawed: counts errors, **not faults** [1]
  - A single fault generates an arbitrary number of errors (0 -> infinite)
  - Errors are more indicative of the **OS's polling frequency** or **application access pattern** than of hardware health
  - Errors overemphasize the effects of **permanent faults**, which can lead to thousands or millions of errors

[1] Sridharan et al, **Memory Errors in Modern Systems: The Good, The Bad, and The Ugly** *In: 20th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)* ACM

Incorrect Methodology => Incorrect Conclusions on Reliability

- Example:
  - Hopper's DRAM **error rate** was 4x greater than Cielo's ← Cielo is ~~more reliable~~
  - **Reality:** Hopper's DRAM **fault rate** was 37% lower than Cielo's



Revisiting Checkpoint/Restart: Take Aways

Sandia National Laboratories

What do we need from hardware to keep checkpoint/restart viable?

- Fast IO a requirement (burst buffers, etc)
- Checkpoint/restart can be a viable option if per-device (per-MB) reliability remains similar to observed today.
- Checkpoint/restart efficiency can remain above 80% if DRAM reliably drops by 20X.
- At current failure rates, improving DRAM reliability has greater benefit than SRAM.
- Improvements to both DRAM (double chipkill) and SRAM (SEC-DED ECC) reliability, can lead to checkpoint efficiencies similar to observed today.

Sources for Failure Data

Cielo at Los Alamos National Lab

Hopper at NERSC / Lawrence Berkeley National

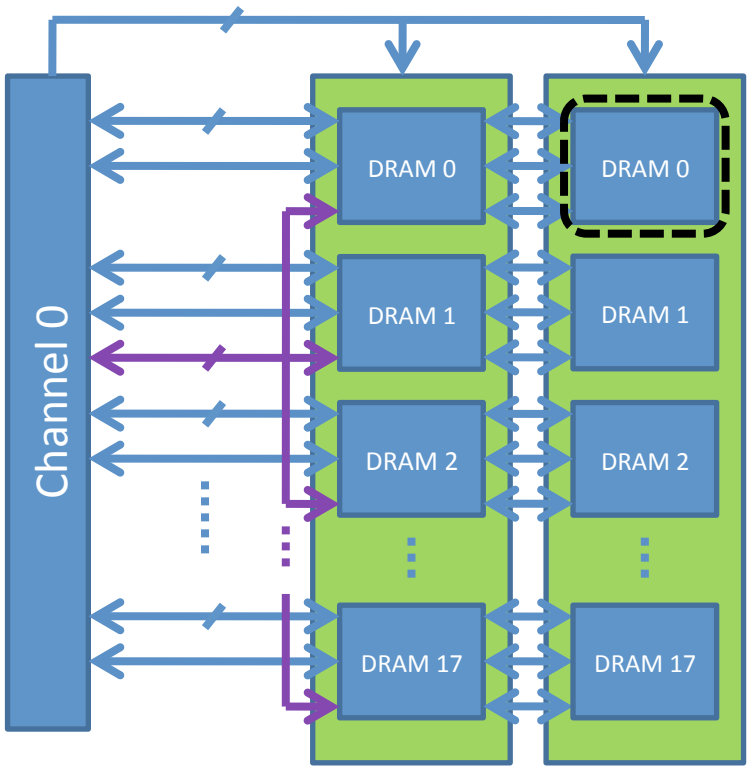
Two Production systems
500M+ CPU socket-hours
40B+ DRAM device-hours

	Cielo	Hopper
Nodes	8568	6000
Sockets / Node	2	2
Cores / Socket	8	12
DIMMs / Socket	4	4
Location	Los Alamos, NM	Oakland, CA
Altitude	7,320 ft	43 ft
DRAM ECC	Chipkill-correct	Chipkill-detect

Exascale Strawman

	Cielo	Exascale
Nodes	8000	50,000 - 100,000
Cores	144,000	1 billion
DRAM Aggregate Capacity	280 TB	32 - 128 PB
DRAM Device Capacity	8 Gbit	8, 16, 32 Gbit
SRAM Capacity / Node	18 MB	150 MB
SRAM Protection	Parity	Parity or SEC-DED ECC
DRAM ECC	Chipkill-correct	Chipkill-correct or double Chipkill correct
Checkpoint Bandwidth	160 GB/sec	10-20 TB/sec

Per-Device/Per MB Reliability



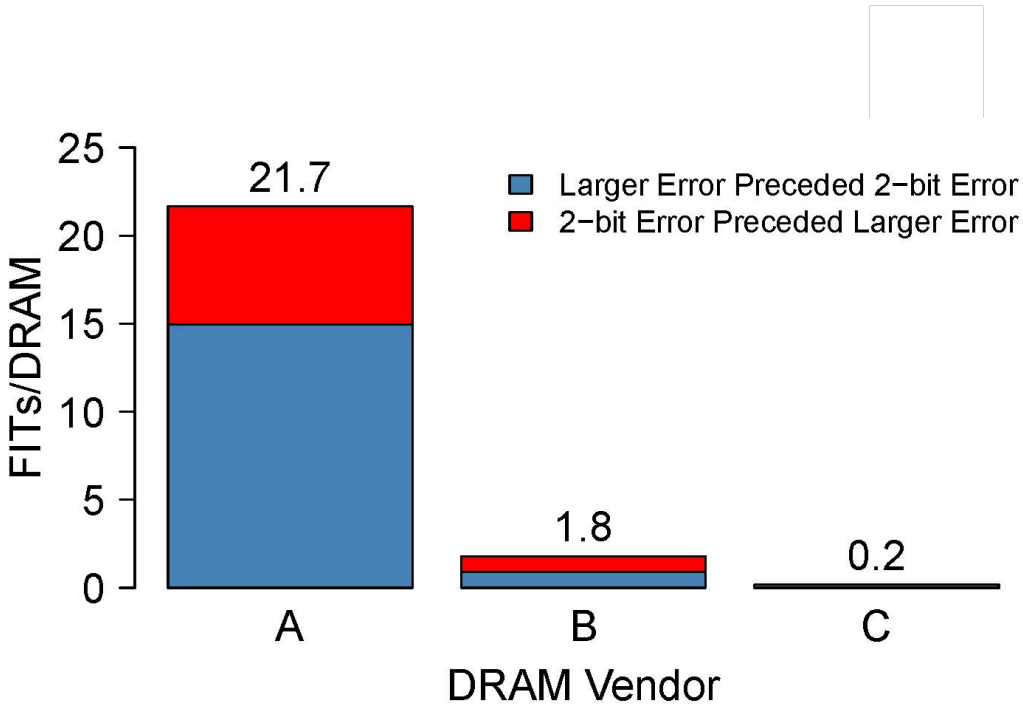
- Unless otherwise stated, we assume:
  - Per-DRAM device reliability @ exascale == Cielo per-device reliability
  - SRAM reliability per MB also remains same as today
  - Therefore, decreased reliability due to dramatic component count increases
  - Historically, vendors have increased per-device reliability while increasing capacity

Failure Rates

$$FIT_{overall} = FIT_{base} + FIT_{DRAM} + FIT_{SRAM}$$

- Base, non-scaling FIT rate (from Cielo)
- FIT rate that scales up
  - DRAM: scales with number of DRAM devices (from Cielo)
  - SRAM: scales per MB of SRAM (Cielo and Hopper)
- For Cielo, base FIT rate dominates
- At scale, DRAM and SRAM dominate depending on configuration

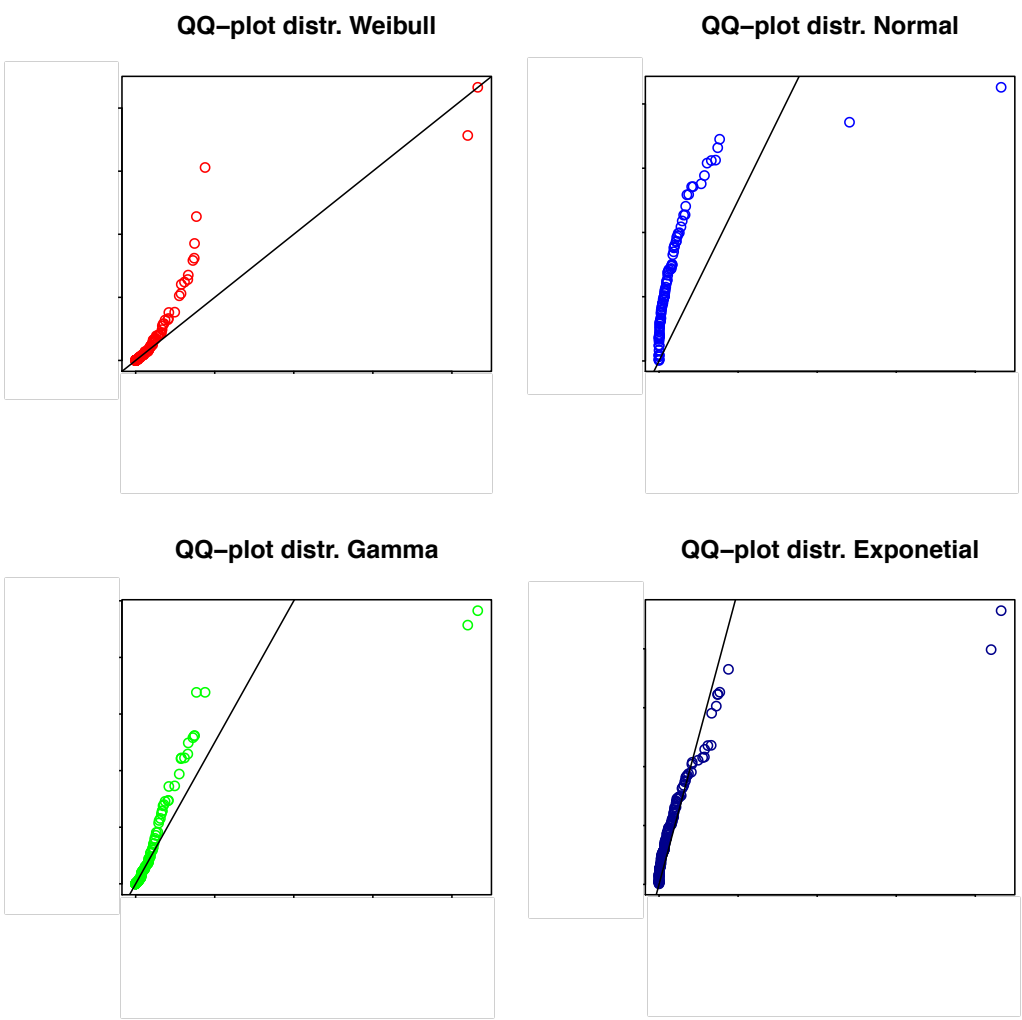
Assumption: SDC not a major concern



- All bets off with checkpointing if veracity a concern
- Straight forward to protect regular storage arrays (DRAM, SRAM)
  - More ECC bits
  - Bit interleaving
- Current cost: ~20% additional circuits, ~25% additional power
- Protecting logic (ALUs) is different story, and ignored in this talk

Sridharan et al, **Memory Errors in Modern Systems: The Good, The Bad, and The Ugly** *In: 20th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)* ACM

Faults are Exponentially Distributed



- Previous work suggested Weibull as a better fit for distribution
- Though a few significant outliers exist, Exponential appears to be the better fit
- We are investigating this difference

Time to Solution due to Daly[2]

$$T_w = M_{sys} * e^{\frac{R}{M_{sys}}} \left(e^{\frac{\tau+\delta}{M_{sys}}} - 1 \right) \frac{T_s}{\tau}$$

T_w: Time to solution with failures
T_s: Time to solution without failures
M_{sys}: System mean time between failures
R: Restart time
τ: Checkpoint frequency
δ: Checkpoint commit time

Workload Characteristics

- Time to solution without failures (T_s): 168 hours

- Per-process checkpoint sizes
  - Application-level checkpoints
  - Collected from five DOE production applications
  - Per-process sizes of minimum and maximum vary by order of magnitude

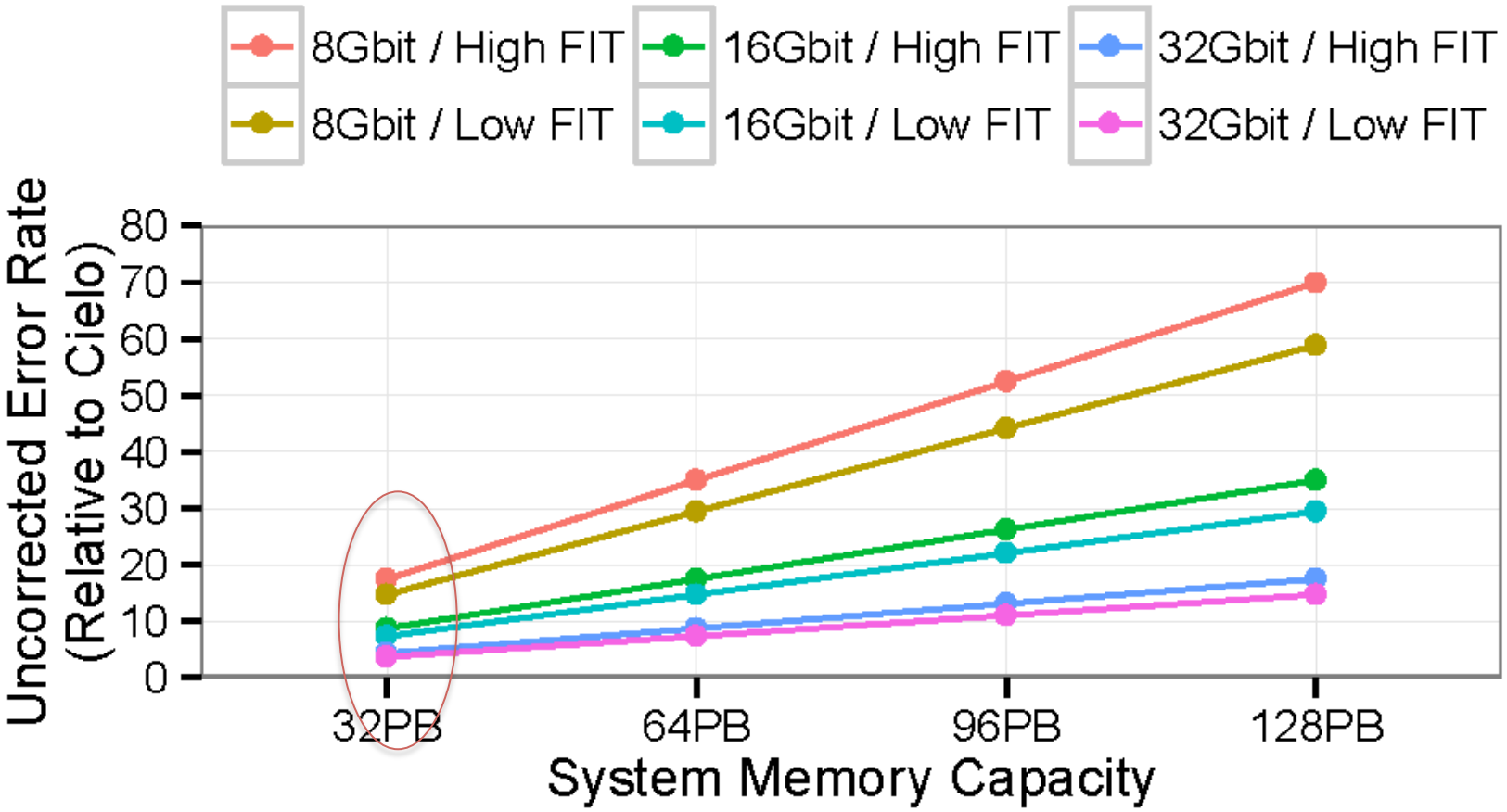
- Checkpoint Bandwidths: Cielo (160GB/sec), Trinity (~1TB/sec), and an Exascale strawman(10-20TB/sec)

- Application Efficiency

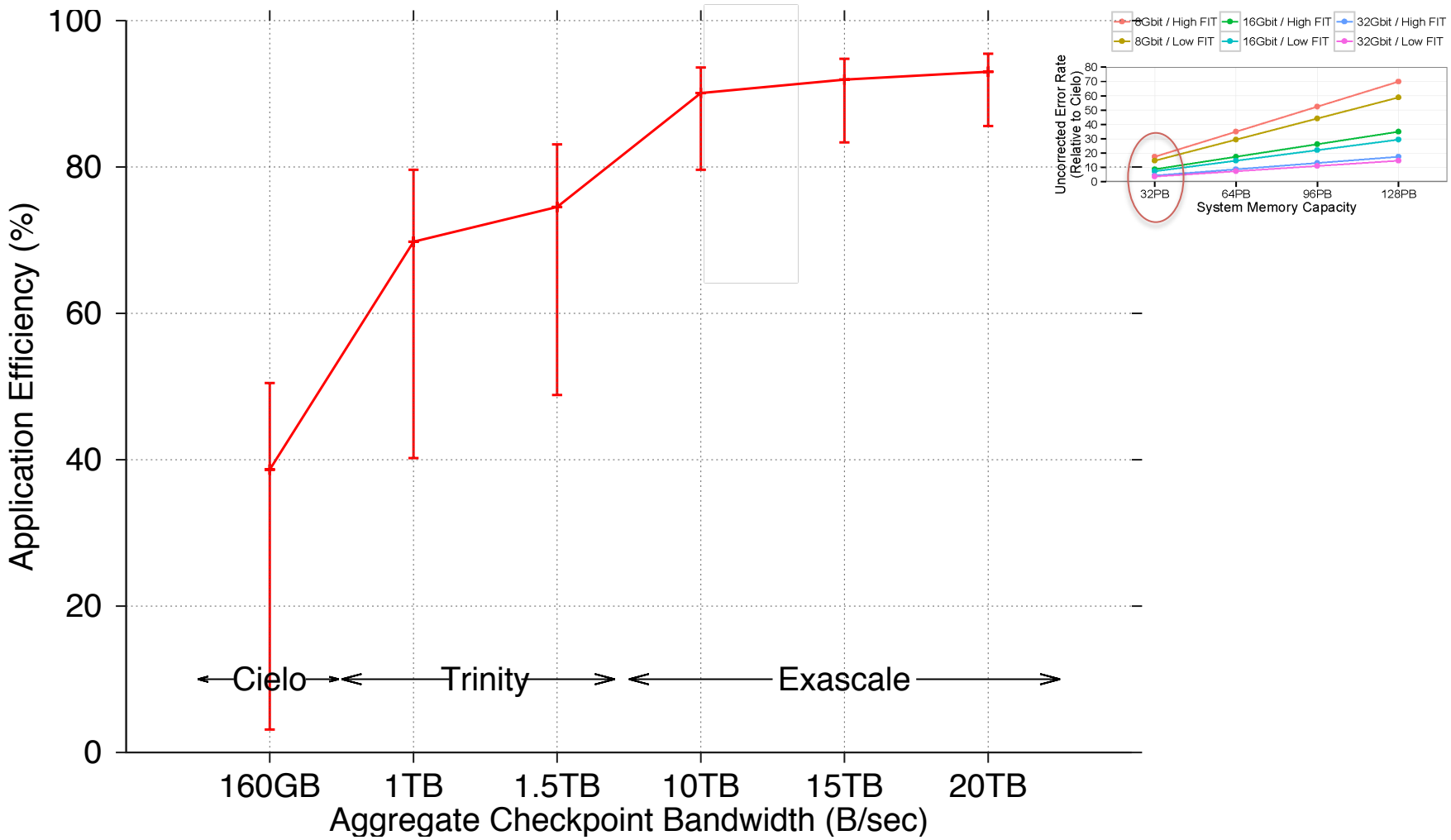
$$Efficiency = \frac{T_s}{T_w} * 100$$

Checkpointing at Exascale with Current FIT Rates

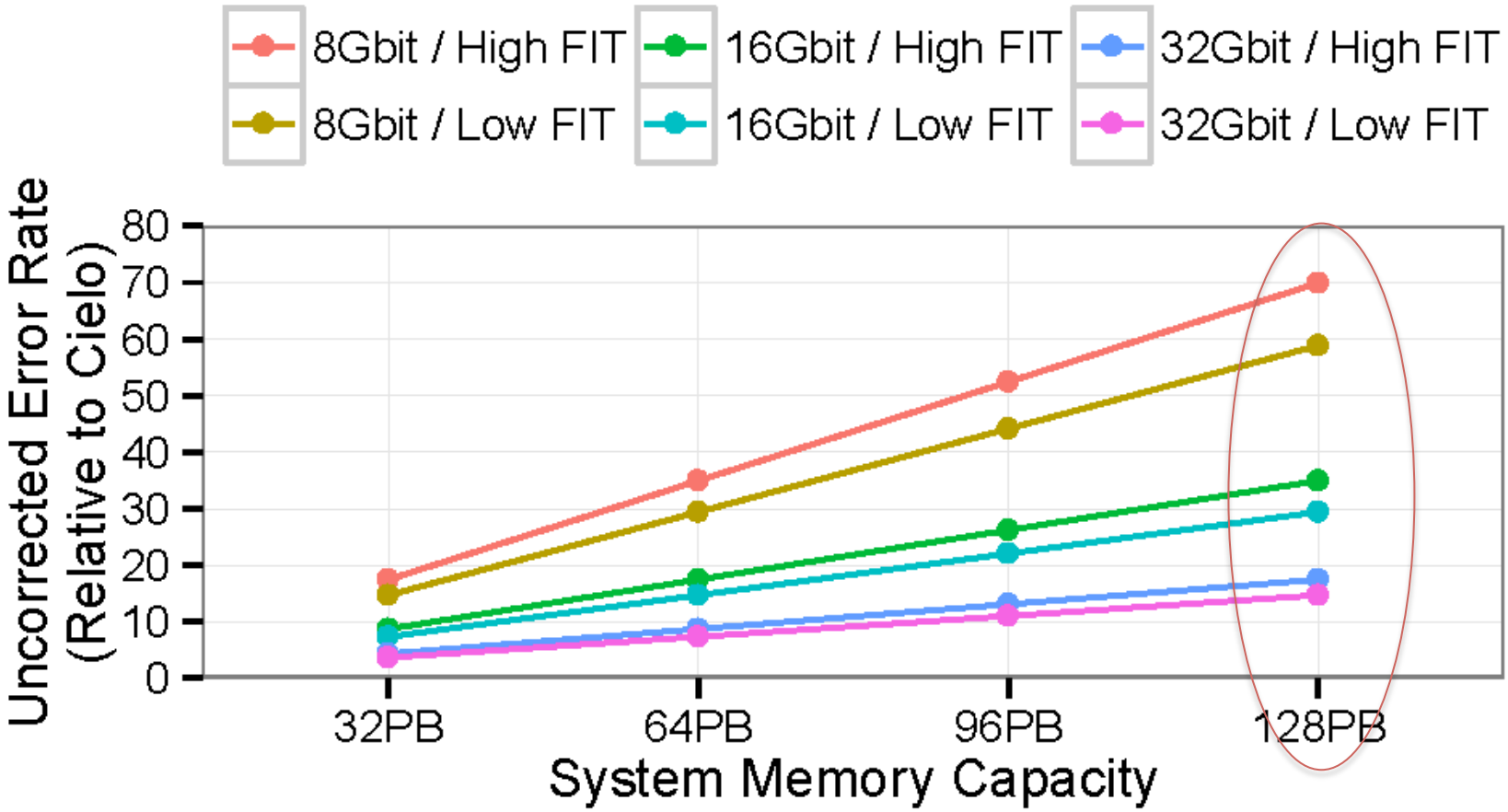
Uncorrected DRAM Error Rate for Strawman



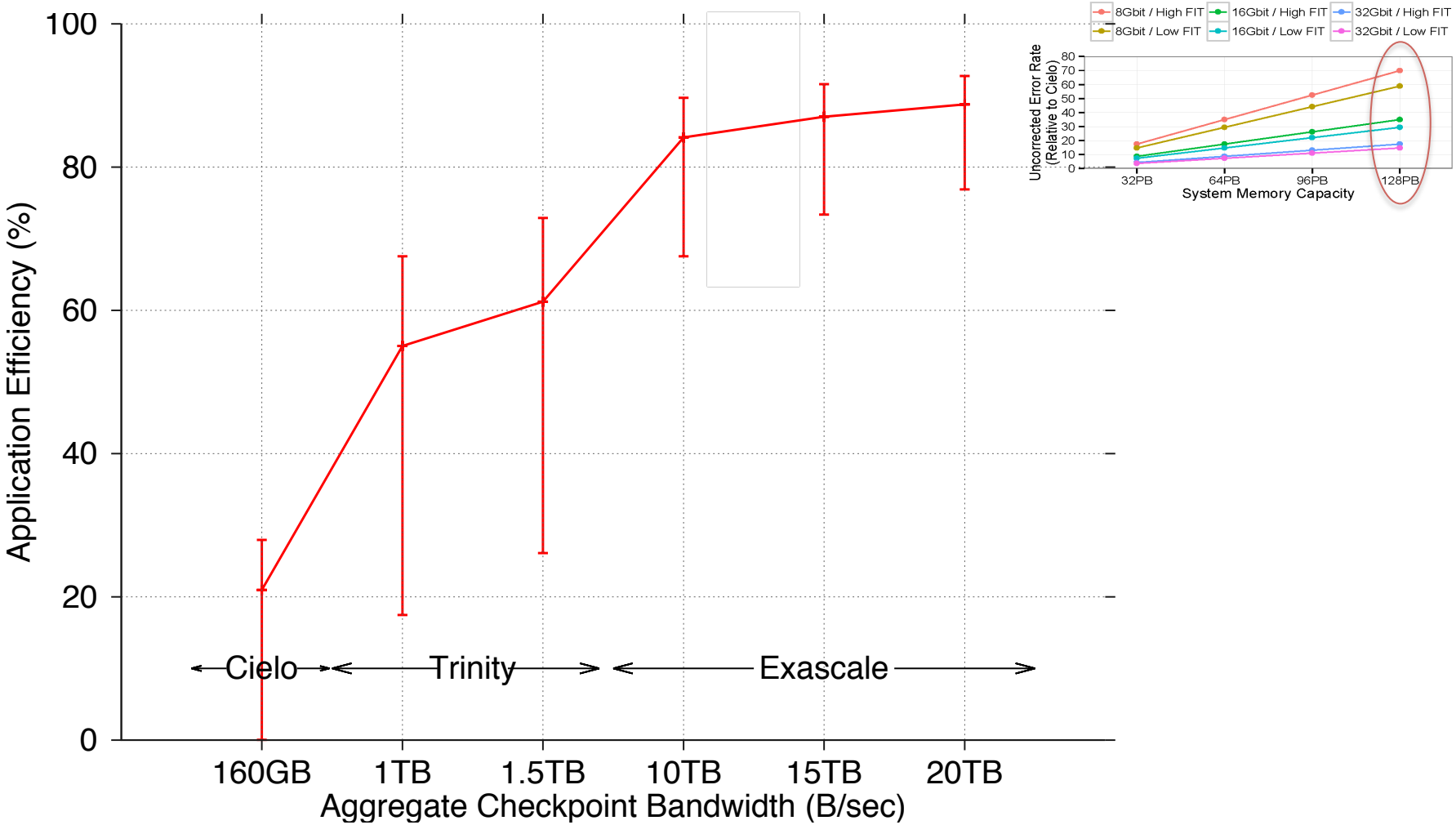
Efficiency @ 32PB >= 90%



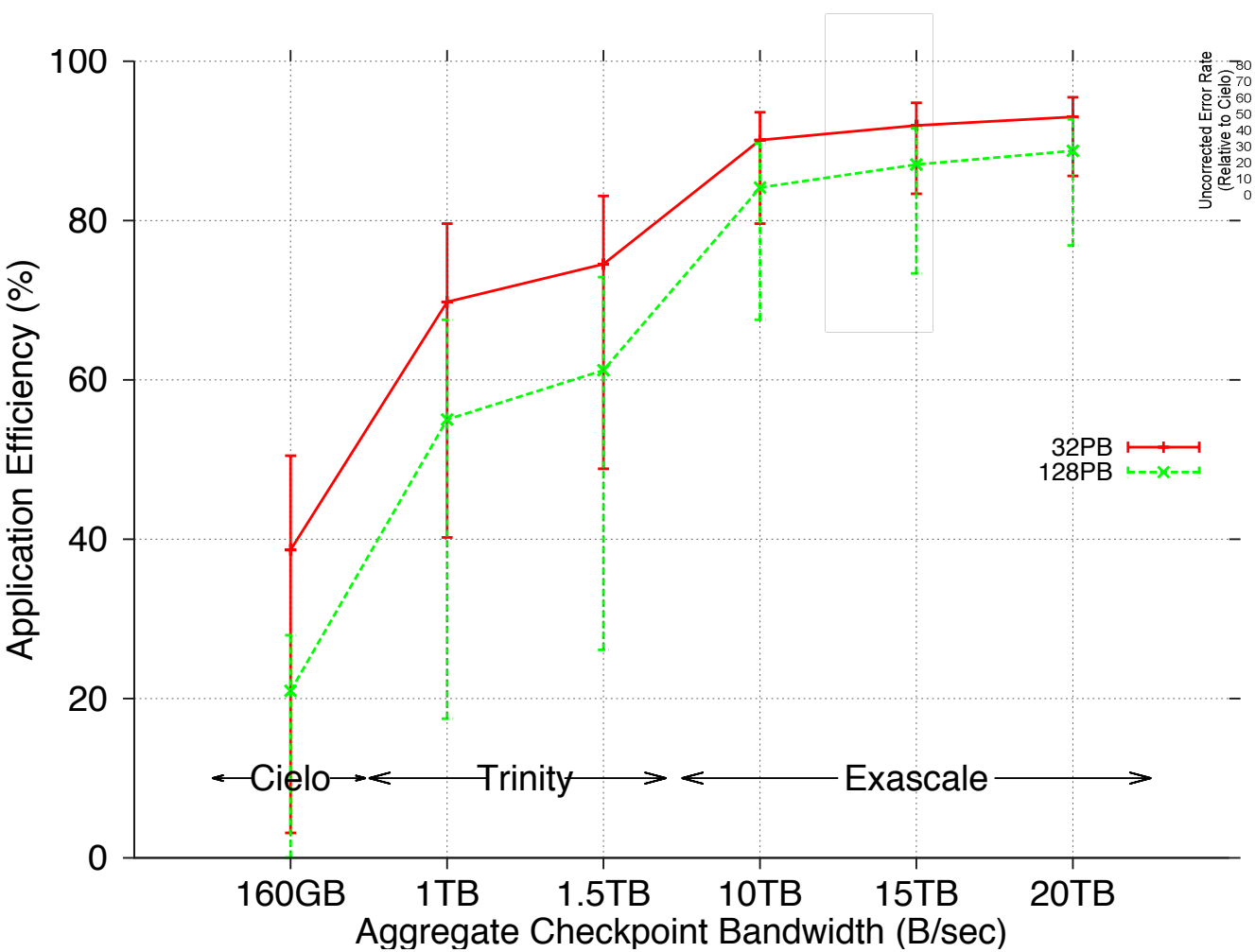
Uncorrected DRAM Error Rate

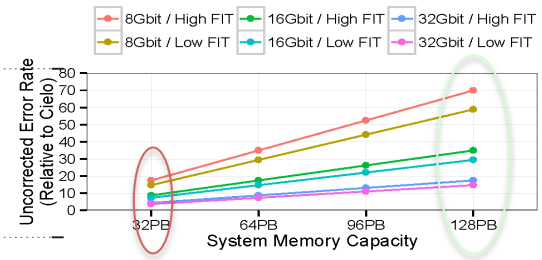


Efficiency @ 128PB > 80%



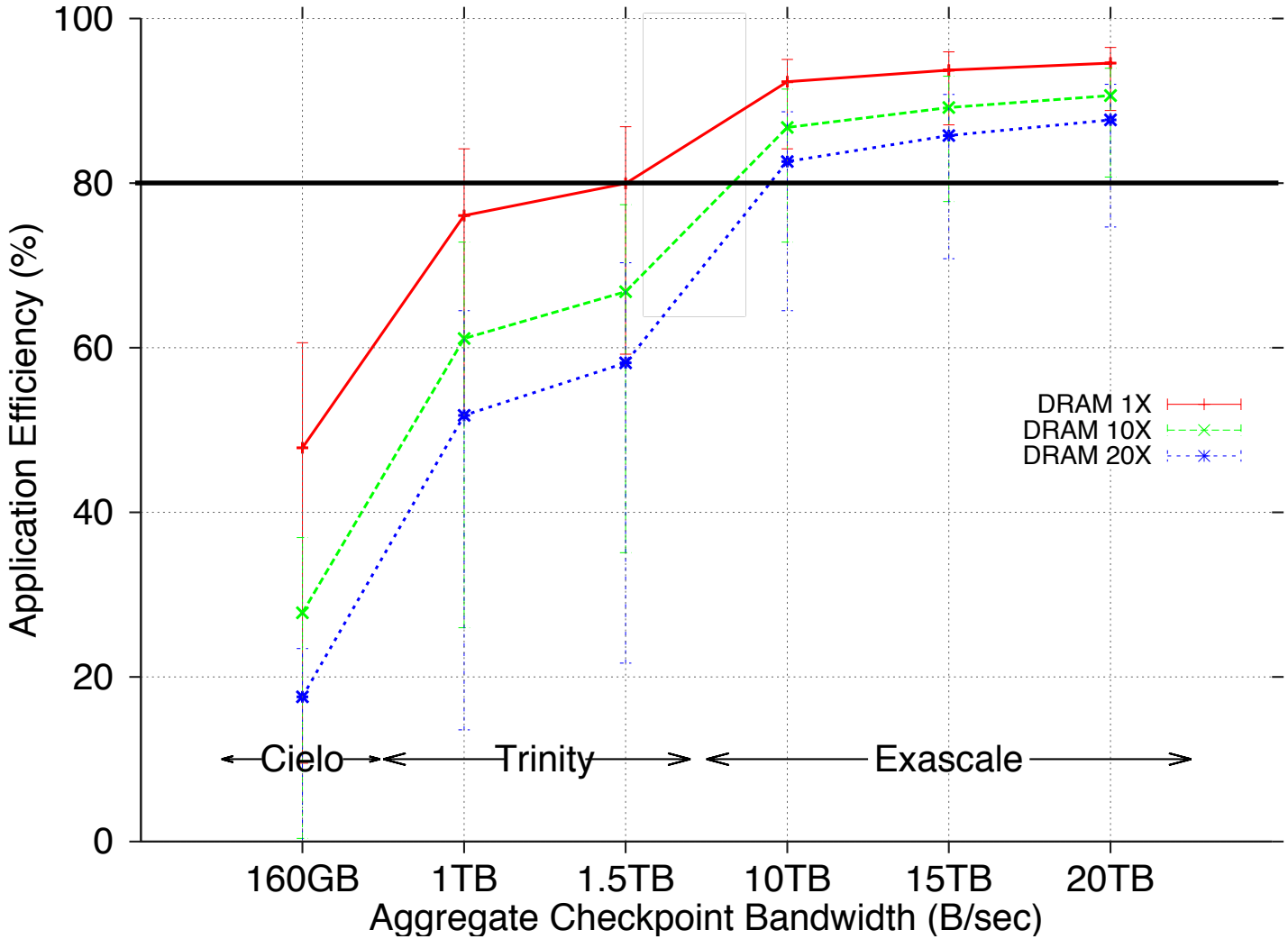
Comparison of Efficiency @ 32PB and 128PB





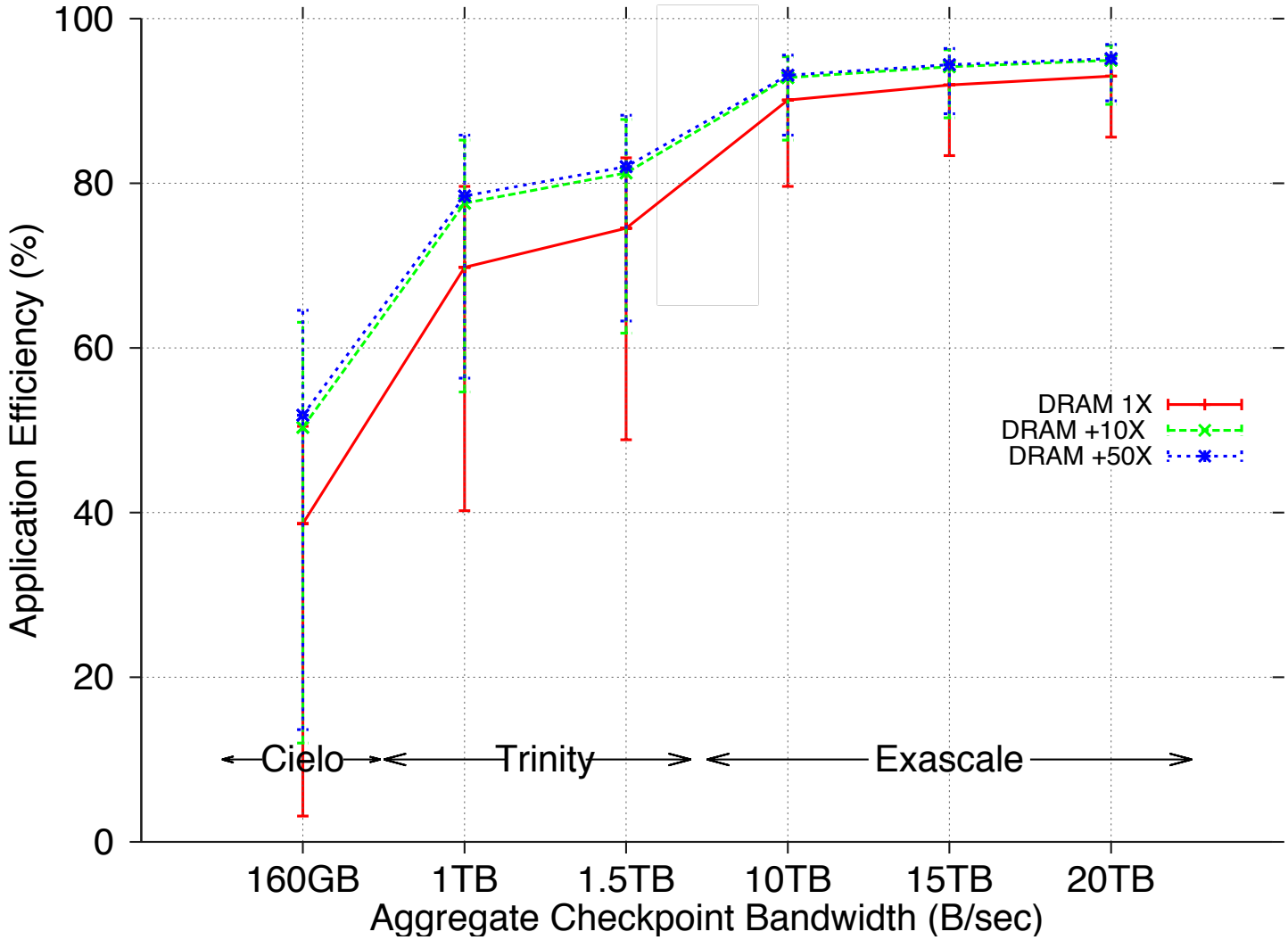
How unreliable can DRAM get and remain 80% Efficient?

20X Decrease in DRAM Reliability Yields Little Cost



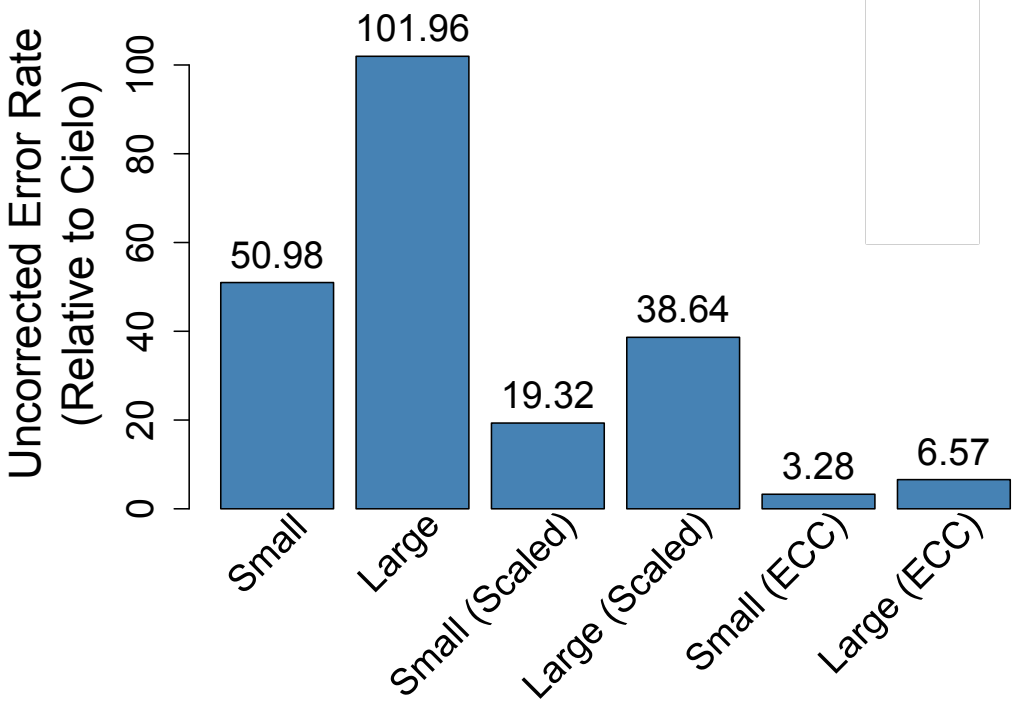
What if DRAM reliability increases?

Greater than 10X Increase, Yields Little Benefit



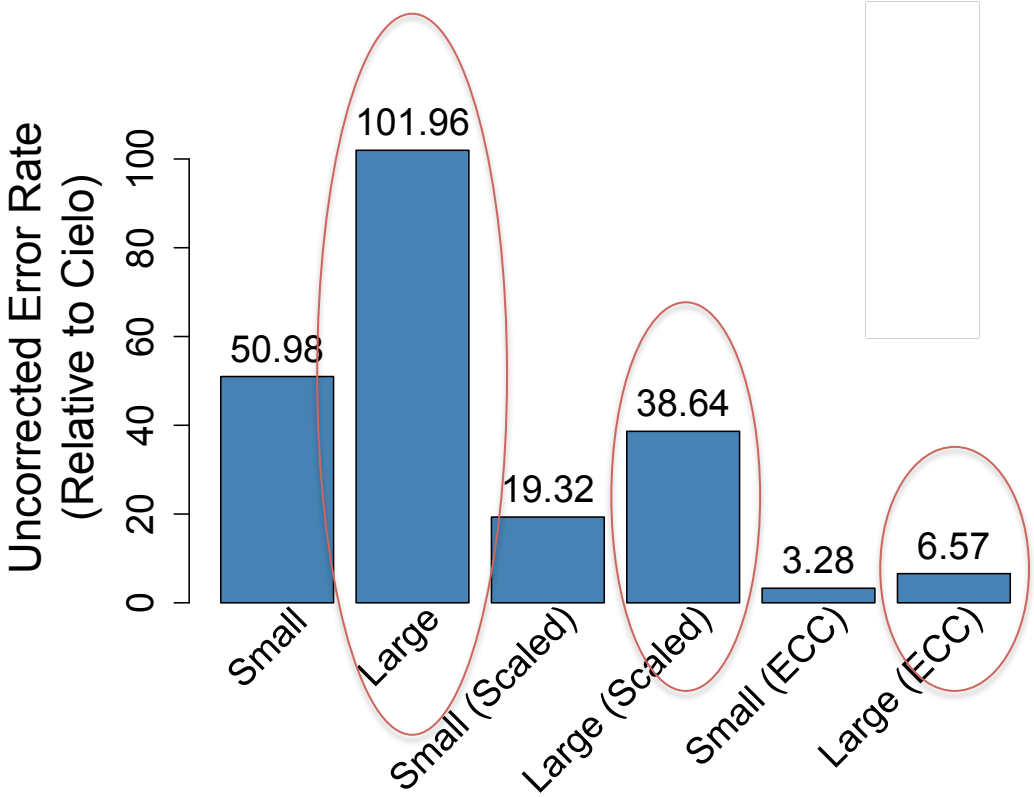
Impact of SRAM Reliability

Uncorrected SRAM Rate for Strawman



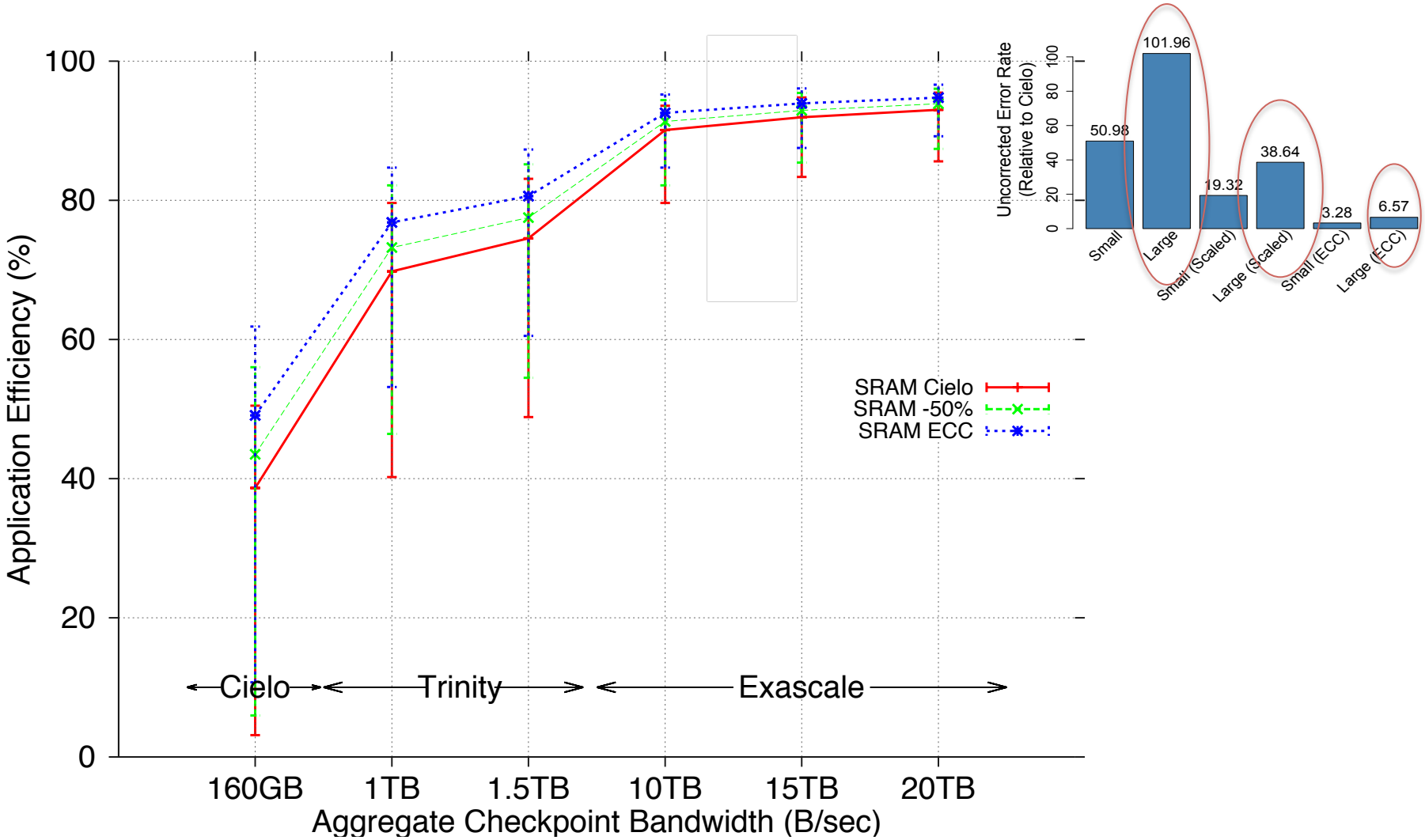
- Node Count
  - **Small**: 50,000
  - **Large**: 100,000
- **Scaled**: SRAM FIT rate decreases 50% over a number of generations [3]
- **ECC**: SEC-DED protection for SRAM

Uncorrected SRAM Rate for Strawman



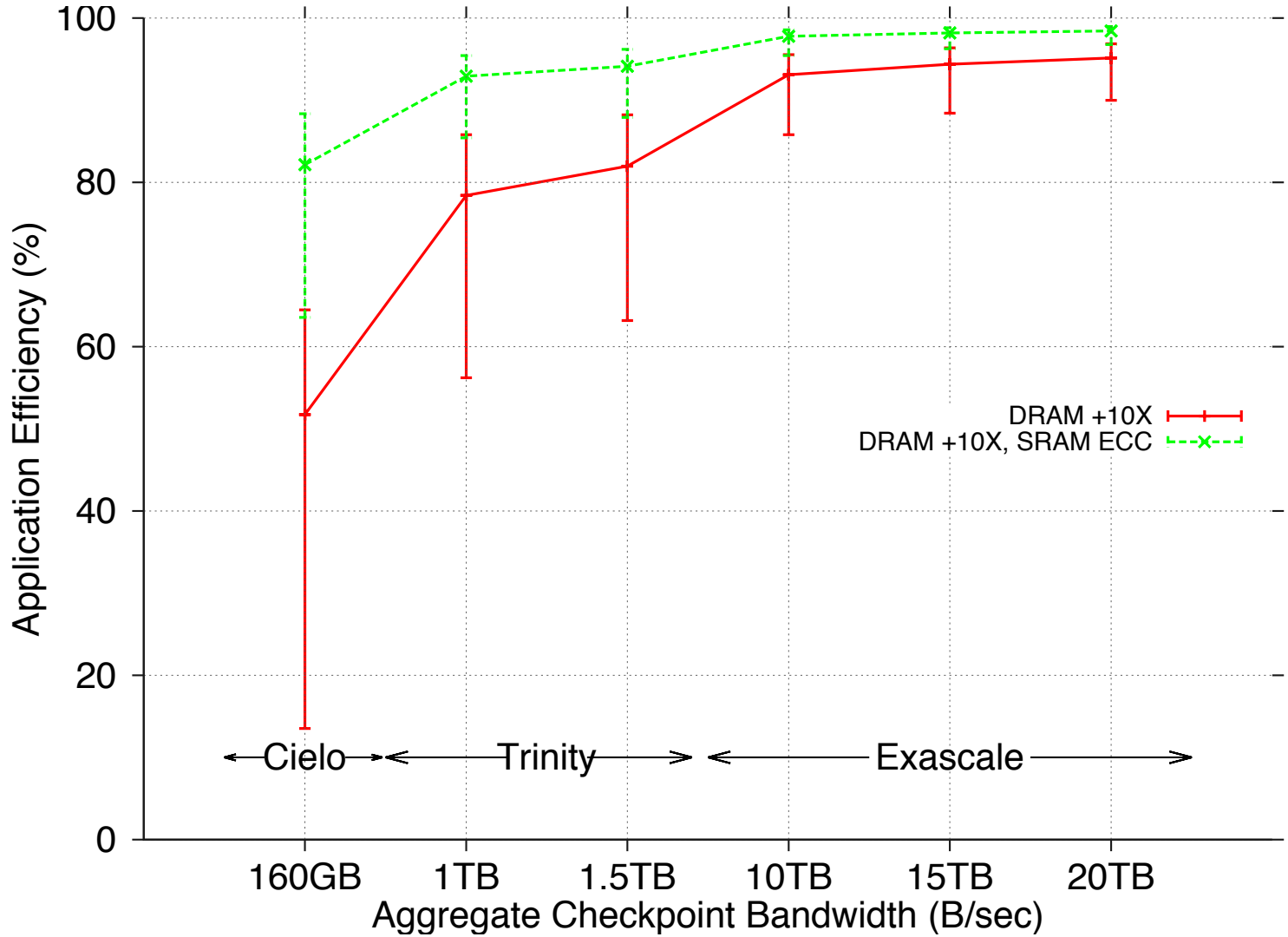
- Node Count
  - **Small**: 50,000
  - **Large**: 100,000
- **Scaled**: SRAM FIT rate decreases 50% over a number of generations [3]
- **ECC**: SEC-DED protection for SRAM

SRAM Reliability Has Little Impact on Efficiency



What if we can increase reliability of both DRAM and SRAM?

SRAM ECC + DRAM double Chipkill ≈ Efficiency Today



Summary

- Importance of using the current state-of-the-art in evaluating the future of checkpoint/restart

- Checkpointing may have a future at exascale, it but comes at a cost:

  - Fast IO a requirement (burst buffers, SCR, etc)
  - We can afford a 20X decrease in DRAM reliability and still remain 80% efficient
  - At current FIT rates, improved DRAM reliability more beneficial than SRAM
  - DRAM double-chipkill correct and SRAM SEC-DED ECC can lead to checkpoint efficiencies similar to those observed on current systems.

Thanks!

Acknowledgements:

- **SNL**: Scott Levy, Jon Stearley, and Patrick Widener
- **UNM:** Dorian Arnold and Patrick Bridges
- **LANL:** Nathan DeBardeleben, Sean Blanchard, and Qiang Guan
- **AMD**: Vilas Sridharan and Sudhanva Gurumurthi
- **NERSC**: John Shalf

References

[1] Vilas Sridharan, Nathan DeBardeleben, Sean Blanchard, Kurt B Ferreira, Jon Stearley, John Shalf, Sudhanva Gurumurth (2015) **Memory Errors in Modern Systems: The Good, The Bad, and The Ugly** *In: 20th International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS)* ACM.

[2] J. T. Daly**, A higher order estimate of the optimum checkpoint interval for restart dumps**, In. Future Generation Computer Systems, 2006.

[3] E. Ibe, H. Taniguchi, Y. Yahagi, K. i. Shimbo, , and T. Toba**. Impact of scaling on neutron-induced soft error in SRAMs from a 250 nm to a 22 nm design rule.** In *IEEE Transactions on Electron Devices*, pages 1527–1538, Jul. 2010.