SAND2013-8662C

Resilience at Exascale – Beyond Checkpoint/Restart

Researchers at Sandia National Laboratories are exploring extreme-scale resilience via methods other than classical checkpoint/restart. By stripping away the assumptions of a static, reliable machine, researchers are developing new, scalable paradigms capable of handling soft, hard, and silent errors, exploiting application-level knowledge and dynamic run-time and system resources.

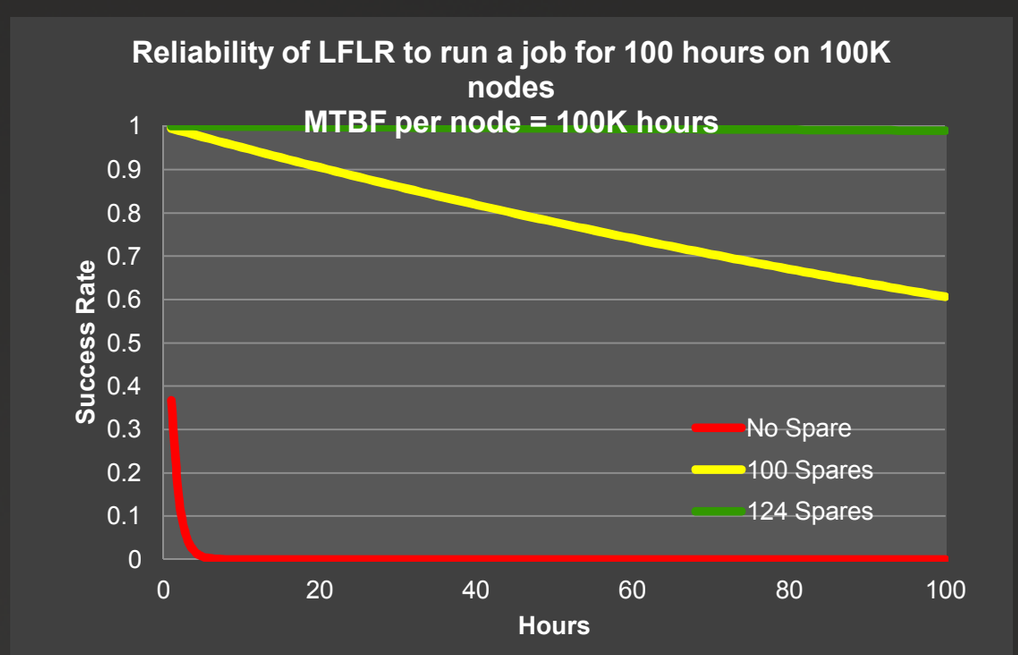
Local Failure Local Recovery Resilience Model

POC: Mike Heroux

Future flagship HPC systems are anticipated to have very frequent node failures (MTBF < 1h). The existing practice of checkpoint-restart, which requires killing and restarting all processes even for single-node failure, is not suitable for such systems due to the poor scaling of global file systems. We are investigating the Local Failure Local Recovery (LFLR) model to address frequent single-node failures.

LFLR Model:

- Adds a small number of spare nodes for program execution
- Uses persistent storage (RAM in neighbor/spare nodes, SSD, etc.)
- Allows scalable recovery for node/data loss



The analytical model shows the probability of a job successfully executing for 100 hours on 100K nodes when the per-node mean time between failure is 100K hours. While the job is nearly guaranteed to have failed after only a few hours when no spares are used, the probability of success is extremely high with only 124 spares.

Weak Scalability of miniFE on IB Cluster using LFLR

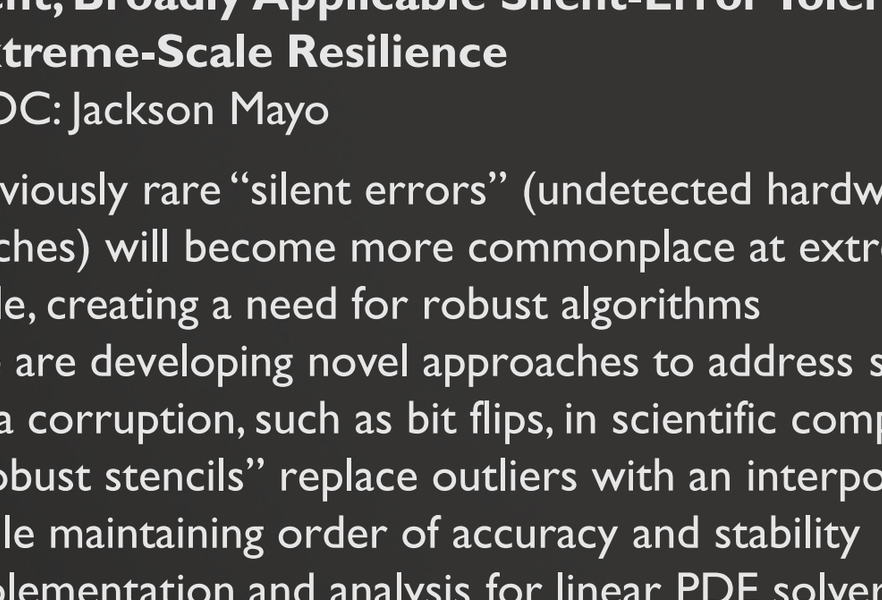
# of Processes	Recovery Time	All Redundant Store	All Execution
1024	0.008	0.015	37.18
2048	0.009	0.014	52.37
4096	0.009	0.014	69.15

With only a single node failure during execution, LFLR adds only a very small overhead to the execution time of miniFE.

Efficient, Broadly Applicable Silent-Error Tolerance for Extreme-Scale Resilience

POC: Jackson Mayo

- Previously rare "silent errors" (undetected hardware glitches) will become more commonplace at extreme scale, creating a need for robust algorithms
- We are developing novel approaches to address silent data corruption, such as bit flips, in scientific computing
- "Robust stencils" replace outliers with an interpolation while maintaining order of accuracy and stability
- Implementation and analysis for linear PDE solver with emulated memory bit flips: Error tolerance benefit is expected to increase asymptotically under weak scaling
- Single-processor runs: Robust stencil tolerates ~1000x higher error (SDC) rates than standard stencil
- Approach is being extended to realistic nonlinear PDEs

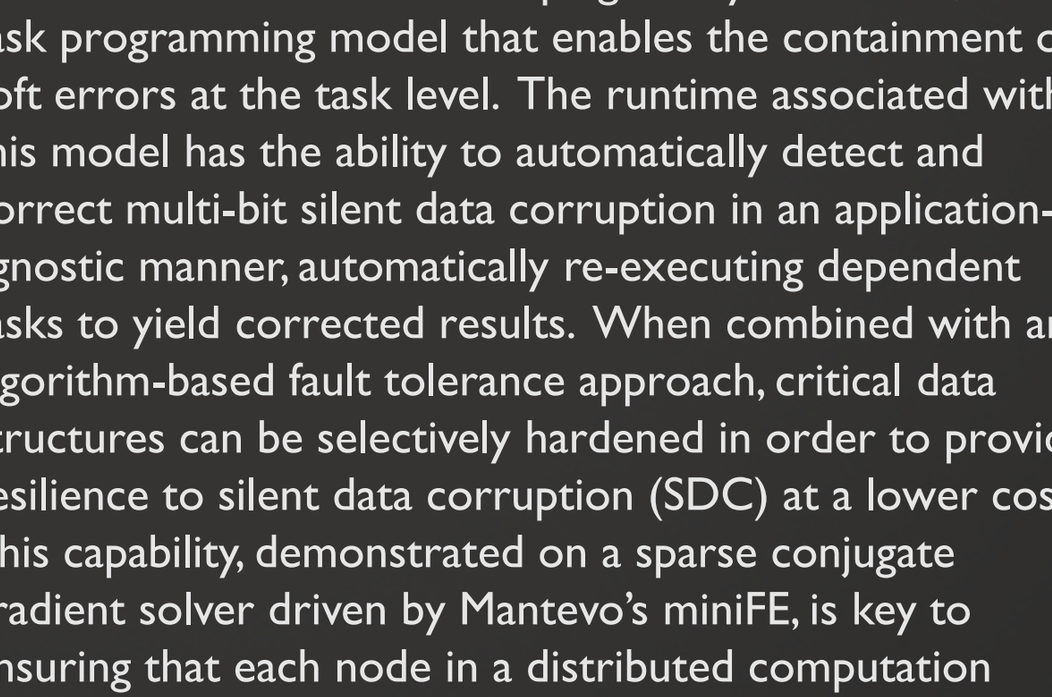


Numerical solution (red points) versus exact solution (gray curve) for a linear advection equation, with various floating-point glitch rates. Small-scale tests more realistic than this demo show the robust stencil tolerates ~1000x higher silent-error rates than the standard Lax–Wendroff stencil.

Scalable, Fault-Tolerant Programming Models

POC: Nicole Slattengren

Sandia researchers are developing an asynchronous, many-task programming model that enables the containment of soft errors at the task level. The runtime associated with this model has the ability to automatically detect and correct multi-bit silent data corruption in an application-agnostic manner, automatically re-executing dependent tasks to yield corrected results. When combined with an algorithm-based fault tolerance approach, critical data structures can be selectively hardened in order to provide resilience to silent data corruption (SDC) at a lower cost. This capability, demonstrated on a sparse conjugate gradient solver driven by Mantevo's miniFE, is key to ensuring that each node in a distributed computation continues marching toward the correct solution in the face of faults, falling back on hard-error recovery techniques as infrequently as possible in order to maximize scalability and performance on extreme-scale systems.



Strong scalability of Sandia's on-node FTPM conjugate gradient port, driven by miniFE, as compared to the Mantevo NUMA-aware OpenMP reference version on a matrix of size 2M with 57M non-zeros. The FTPM code exhibits excellent scalability and competitive performance both without resilience features and with selective resilience.

FUNDING AGENCY: ASC

FUNDING ACKNOWLEDGEMENT: NNSA's Office of Advanced Simulation & Computing, NA-114.

CONTACT: Robert Clay, Sandia National Labs, rlclay@sandia.gov

U.S. DEPARTMENT OF ENERGY