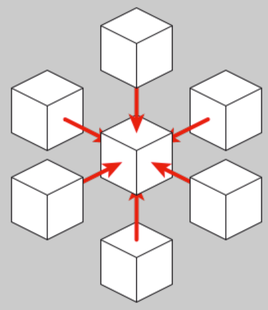
SAND2014-16079D

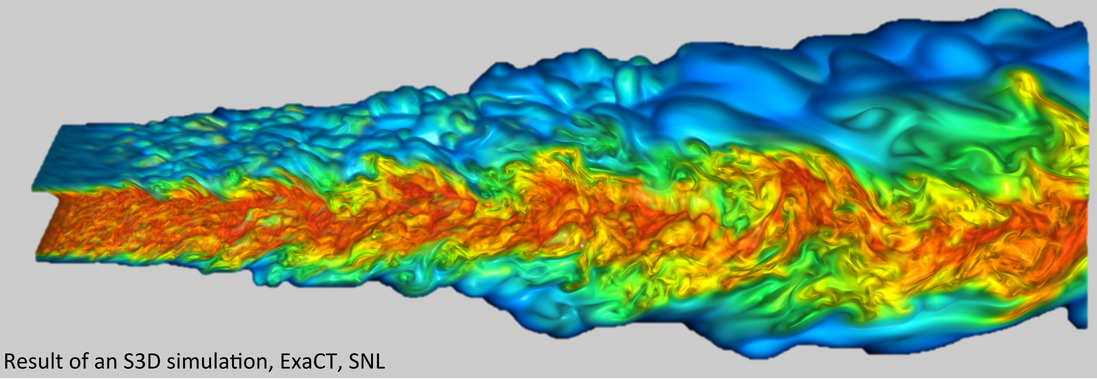
Exceptional service in the national interest

Sandia National Laboratories

Titan Cray XK7 supercomputer, ORNL



6-point 3D von Neumann stencil



Result of an S3D simulation, ExaCT, SNL

Tolerating Hard Failures for Stencil-based Applications on HPC environments

Marc Gamell*, Keita Teranishi+

* Rutgers University, PhD Computer Engineering, est. 2016 *mgamell@{cac.rutgers.edu,sandia.gov}*
+ Org 8953, Scalable Modeling & Analysis *knteran@sandia.edu*

Goal

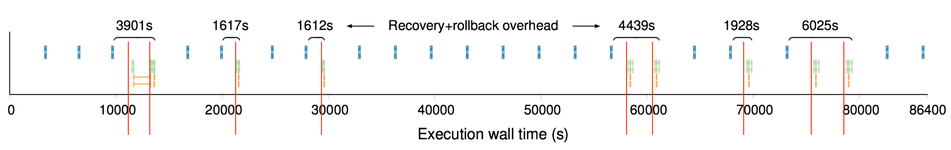
- To design **local recovery algorithm for Stencil-based applications**.
- To evaluate its performance on distributed Partial Differential Equation numerical solvers:
  - One-dimensional kernel
  - Three-dimensional kernel
  - S3D combustion simulator developed at Sandia (ExaCT)

Motivation

- Exascale will have **low reliability** due to large number of cores (10-year processor MTBF, 1M processors = 5.25 mins system MTBF)
- An important failure class is process/node failures.
- As suggested by recent studies, the abstraction of a failure-free machine will not be sustainable at extreme scales
- While other models (e.g. task-DAG) can be designed with resilience features, SPMD and MP are not designed to handle process failures
- Local Failure Local Recovery model tries to address this for SPMD.

Current approach (coordinated, PFS-based Checkpoint/Restart), unfeasible towards exascale
Application-aware resilience required

Current approach: Global Checkpoint and Offline Restart using state-of-art I/O



S3D production runs on Titan Cray XK7 (125k cores)
9 process/node failures over 24 hours
Failures are promoted to **job failures**
Checkpoint (5.2 MB/core) stored in the PFS

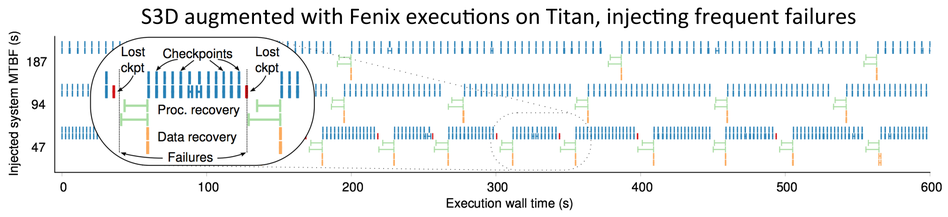
	Average	Total
Checkpoint data	**55 s**	1.72 %
Restarting processes	**470 s**	5.67 %
Loading checkpoint	**44 s**	1.38 %
Rollback overhead	**1654 s**	22.63 %
Total overhead due to fault tolerance		**31.40 %**

Towards exascale, **O(1)** process/node **failure per minute**

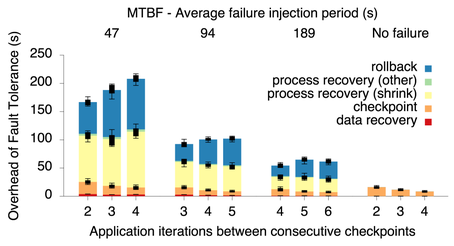
- **Checkpoint frequency** has to be dramatically **increased**
- Current **checkpoint cost, O(1) minute, is unfeasible**
  - → **in-memory, application-specific, high-frequency checkpointing**
- Recovery cost must be reduced
- Node failures cannot be propagated
  - → **local online recovery**

Online Rollback Global Recovery (SC'14)

- Online recovery allows the usage of in-memory checkpointing, O(1s).



- Efficient recovery from high frequency node failures, as exascale compels.
- **With failures injected every 189, 94 and 47 seconds,** the total job run-time penalty is as low as **10%, 15% and 31%, respectively**.
- This can dramatically improve by optimizing ULFM comm_shrink.



Online Roll-Forward Local Recovery

- Towards extreme-scale, failures need to be contained: the effect of a failure should be proportional to its size, not to the system size!
- Upon process/node failure, only the failed ranks have to rollback. While close neighbors need to wait for respawned ranks to catch up, distant neighbors may not be delayed by failures (ideal scenario for scalability!)
- Need to model the delay and performance loss caused by a failure.
- Recovery mechanism leverages existing ghost point mechanism.
  - Partial data redundancy is already needed by the PDE operation
  - Fault Tolerance mechanism can leverage it!
- Need to evaluate how different buffer sizes affect performance, depending on the MTBF

Implementation:
Failure detection and survival uses MPI+ULFM

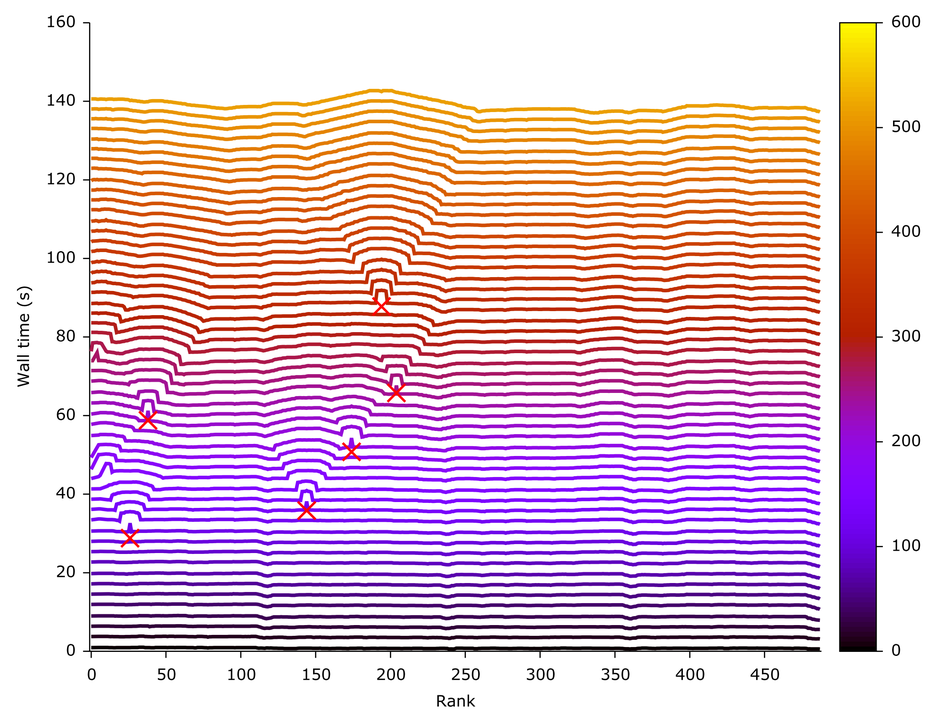
- MPI implementations consider a process loss a fatal error.
- ULFM is an extension proposed to the MPI-Forum to allow Run-Through Stabilization.

Spare rank allocation

- As MPI_Spawn is not implemented on Cray systems, we pre-allocate a set of spare ranks.

Failure recovery:

- Upon failure, whichever neighbor detects the failure first, notifies the recovery coordinator, who will assign spare resources and manage its connectivity with the neighbors
- No global communicator fix is required, which is the major source of overhead with MPI+ULFM.



Conclusions

- By leveraging application knowledge, more efficient failure handling mechanisms can be designed.
- Stencil-based schemes provide a good communication pattern to tolerate failures without involving global communicators.
- Work in progress demonstrates better performance and scalability with local recovery.

Future work

- Checkpoint-free local recovery
- Neighbor-assisted local recovery

12 15 18
1010101
TERA PETA EXASCALE & BEYOND

U.S. DEPARTMENT OF ENERGY
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. 142964