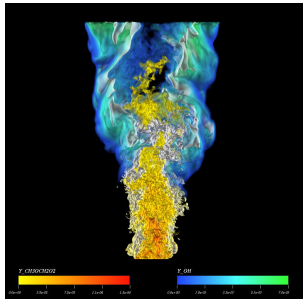
SAND2017-12460C



The Unexpected Virtue of Almost: Exploiting MPI Collective Operations to Approximately Coordinate Checkpoints

Scott Levy, Kurt Ferreira, and Patrick Widener
Center for Computing Research
Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

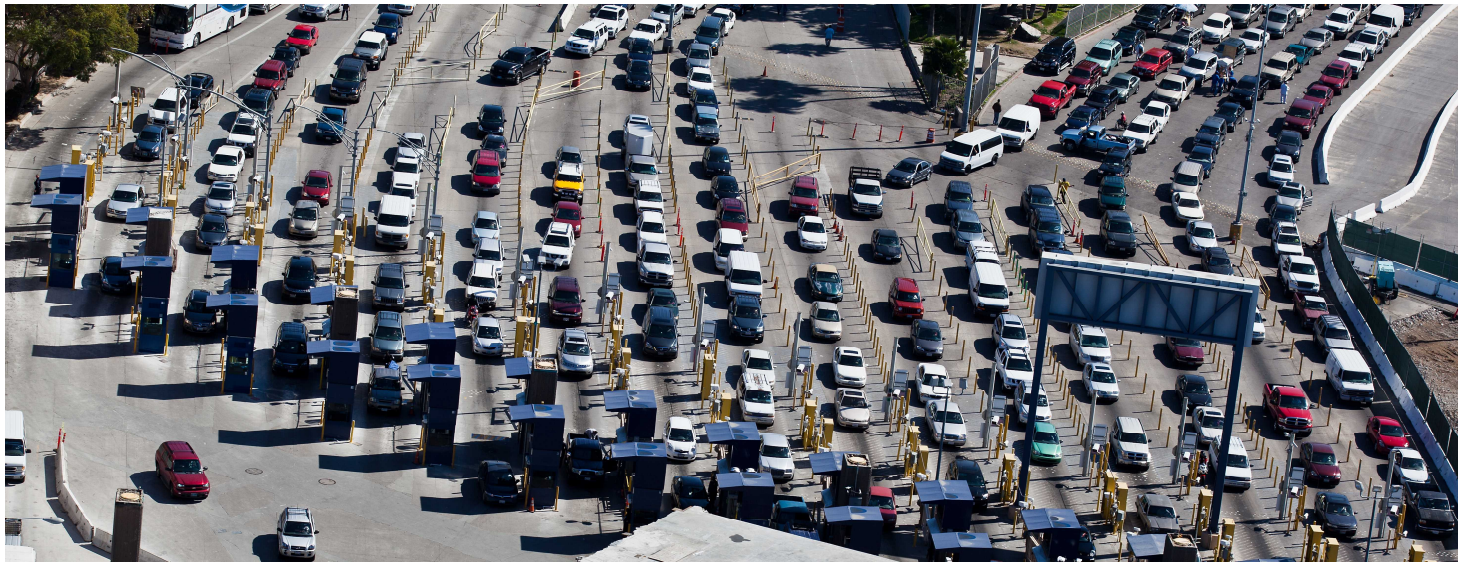
CCR Center for Computing Research

Sandia National Laboratories is a multi mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

State of Fault Tolerance

Sandia National Laboratories

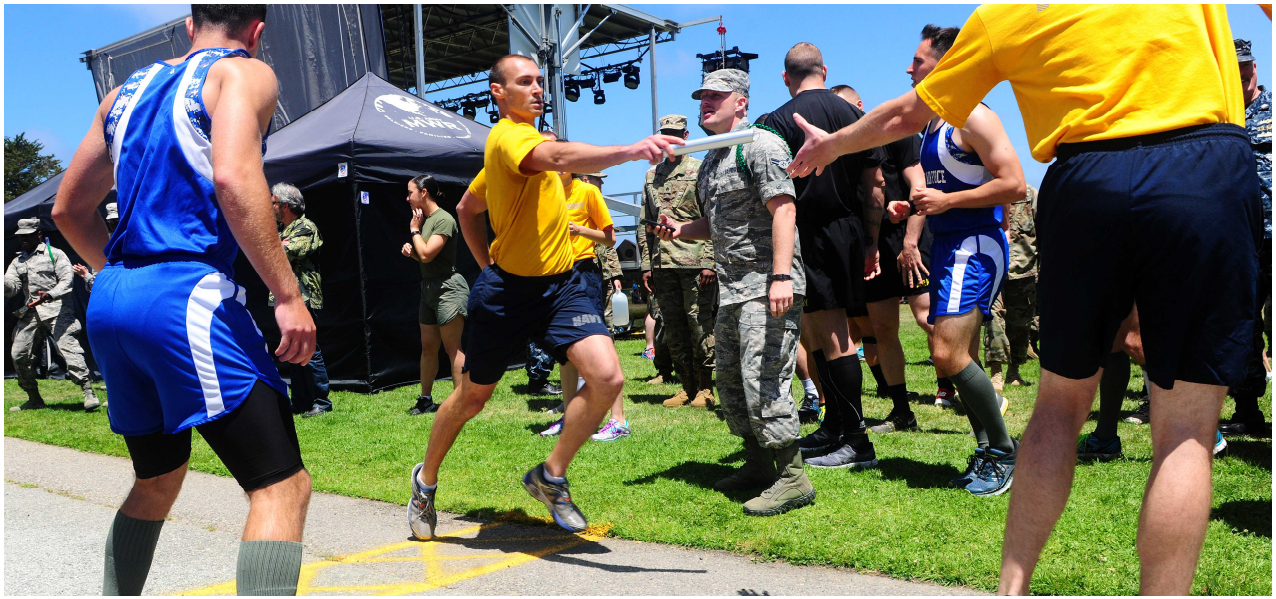
- Coordinated Checkpoint/Restart is dominant
  - every process checkpoints at the same time
  - but...may not scale well due to the costs of coordination and congestion due to contention for storage resources
  - local persistent storage (e.g., burst buffers) may help reduce contention; time will tell



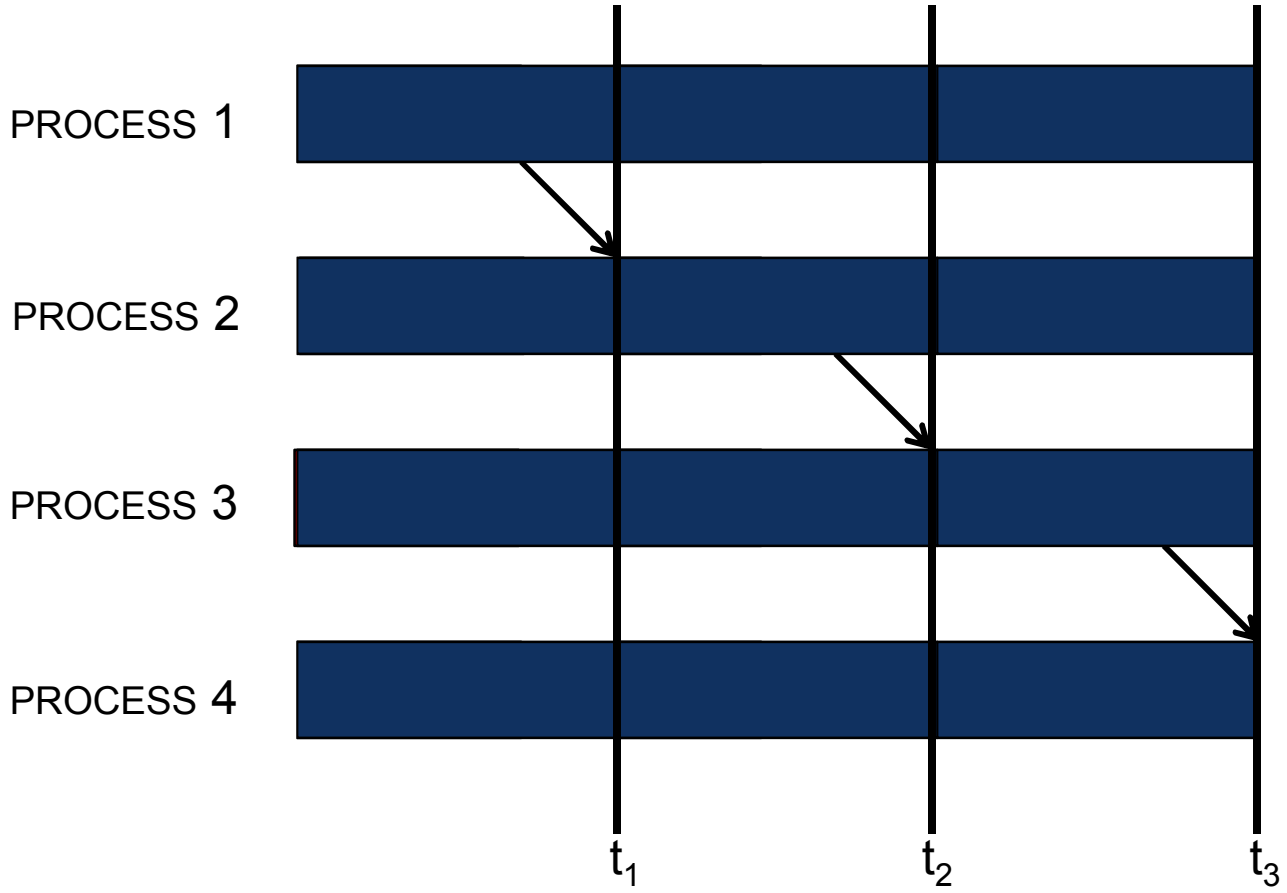
State of Fault Tolerance (cont'd)

Sandia National Laboratories

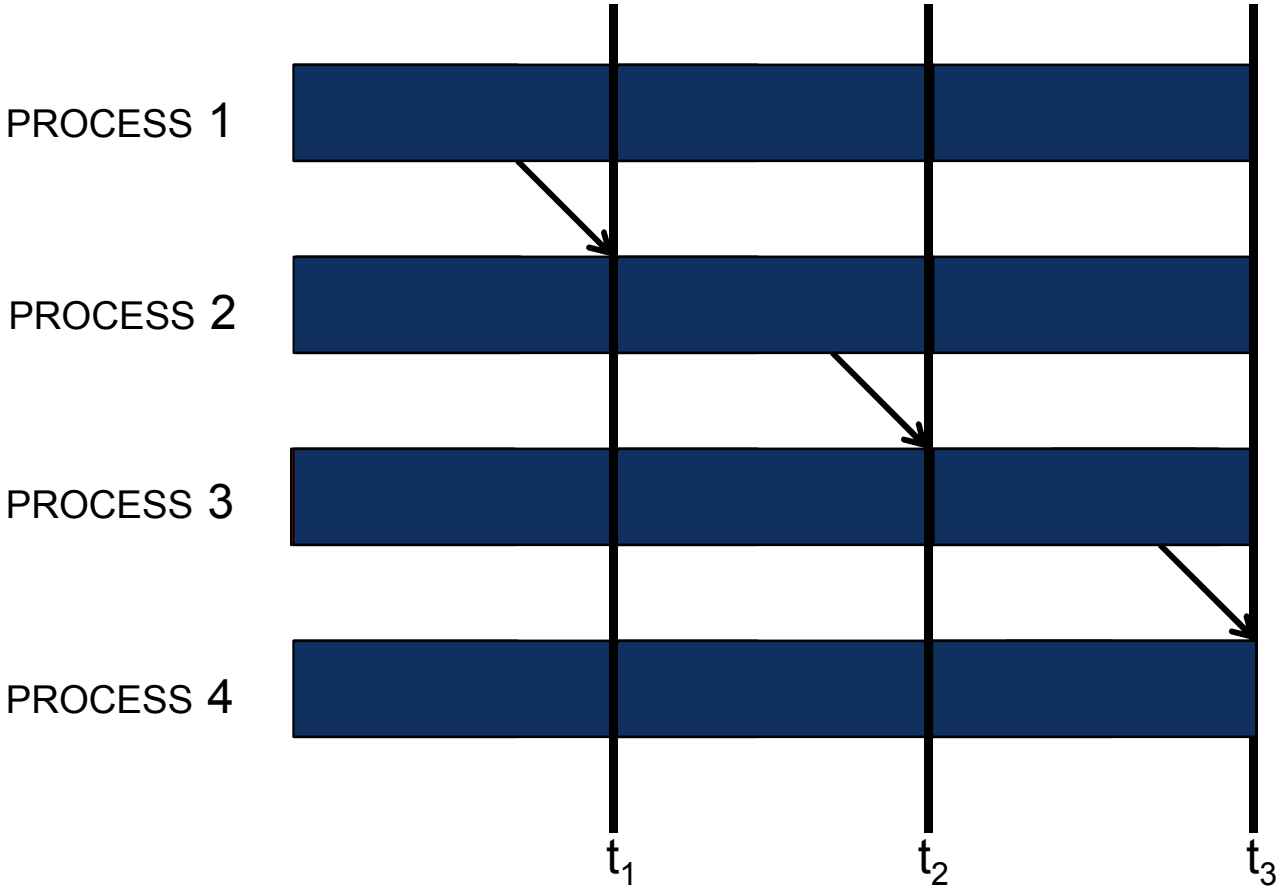
- Uncoordinated Checkpoint/Restart
  - eliminates the requirement of inter-process coordination
  - additional mechanisms (e.g., message logging) are needed to ensure that checkpoints represent a consistent machine state
  - but...may not scale well because checkpointing delays may propagate along communication dependencies



Perfectly Coordinated C/R



Completely Uncoordinated C/R



Approximately Coordinated C/R

Sandia National Laboratories

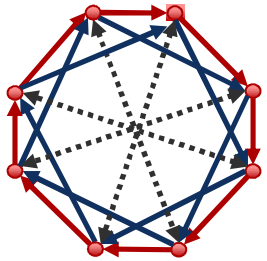
In this paper, we seek to answer the question:

Do collective operations provide enough synchronization to mitigate the performance impact of Uncoordinated C/R?

Collective Algorithms

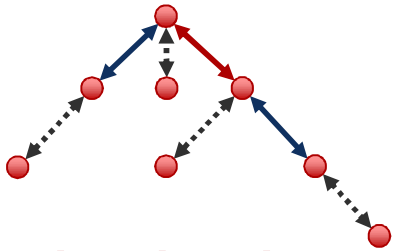
- Different collective algorithms have different effects on inter-process synchronization
  - Dissemination

    (e.g., to implement `MPI_Allreduce`)

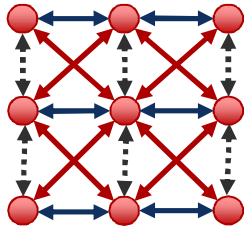


  - Binomial tree dispersal/aggregation

    (e.g., to implement `MPI_Bcast`/`MPI_Reduce`)



  - Stencil communication(e.g., to implement `MPI_Neighbor_alltoall`)



- We don't currently distinguish among collective operations

Why Extend MPI?

Sandia National Laboratories

- The application programmer can (in most cases) ensure that checkpoints occur after a collective operation
- *However*, ensuring that checkpoints are taken at the end of an appropriate interval requires additional software infrastructure
- MPI is well-positioned to ensure that checkpoints are taken after a collective operation
- Existing research on extending MPI to support checkpointing provides guidance and shows that the basic premise is sound

Experimental Approach: Simulator

Sandia National Laboratories

- Results collected using LogGOPSim (Hoefler et al., HPDC 2010; *see also* Levy et al., PMBS 2013), a discrete-event simulator for MPI programs
- Simulates workload execution based on traces of MPI operations collected using MPI Profiling interface
- Time between MPI operations is modeled as computation
- Very simple network model: fully-connected network, LogGOPS network model is used to determine the time required to send messages between any two processes
- Simulator was modified to support checkpoint/restart, including an option to force checkpoints to occur after a collective operation

Experimental Approach : Resilience

- Start with failure-free operation
- Optimal checkpoint interval for Uncoordinated C/R is unknown
- To begin our exploration, failure-free execution with:
  - checkpoint commit time (δ) = 1 second
  - checkpoint interval (τ) = 2 minutes
- Corresponds to the optimal Coordinated C/R interval for a system with an MTBF of 2 hours
- Overhead

minimum

$$overhead_{min} = \frac{\delta}{2\tau + \delta} = 0.41\%$$

maximum

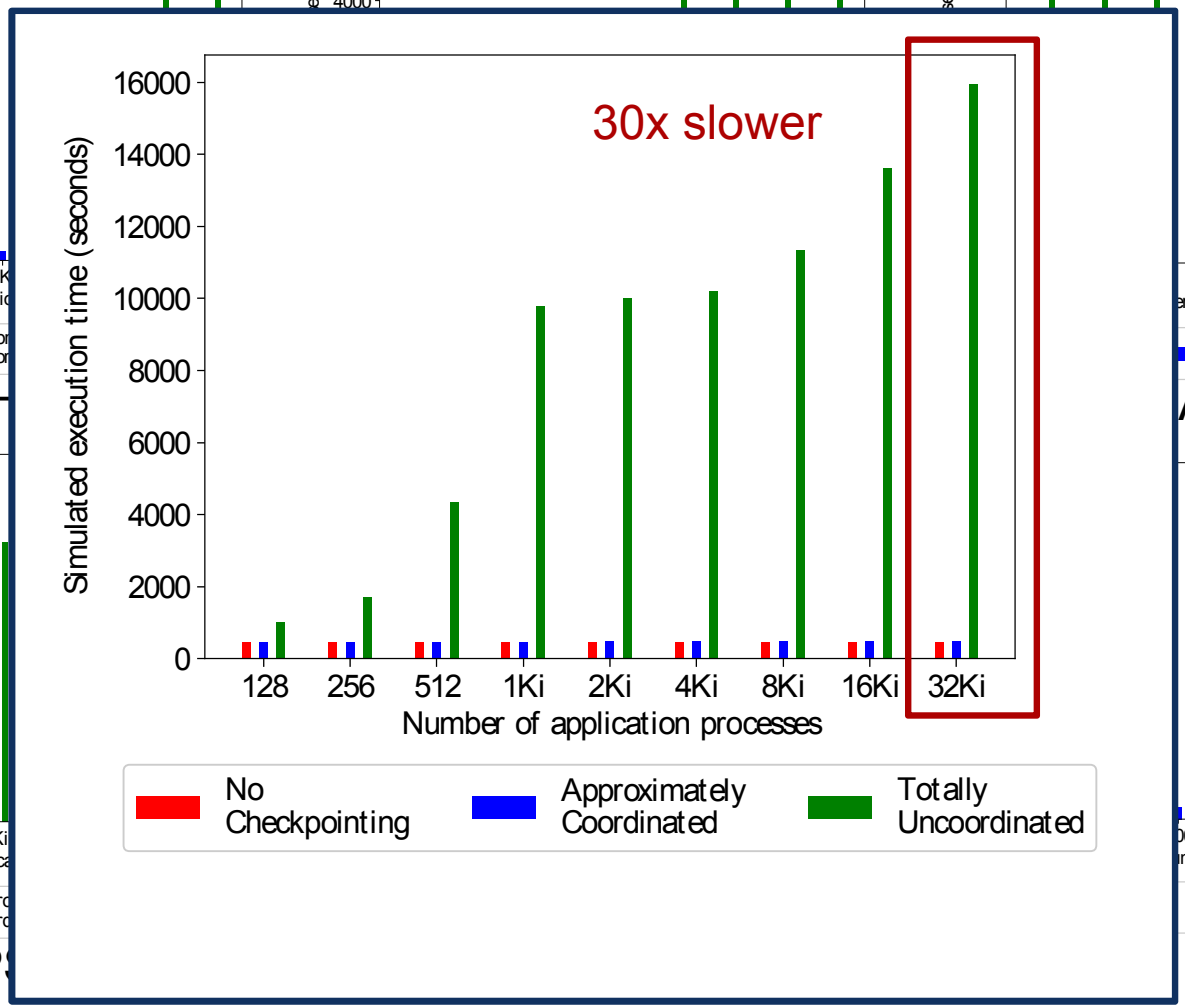
$$overhead_{max} = \frac{\delta}{\tau + \delta} = 0.83\%$$

Experimental Approach : Workloads

Sandia National Laboratories

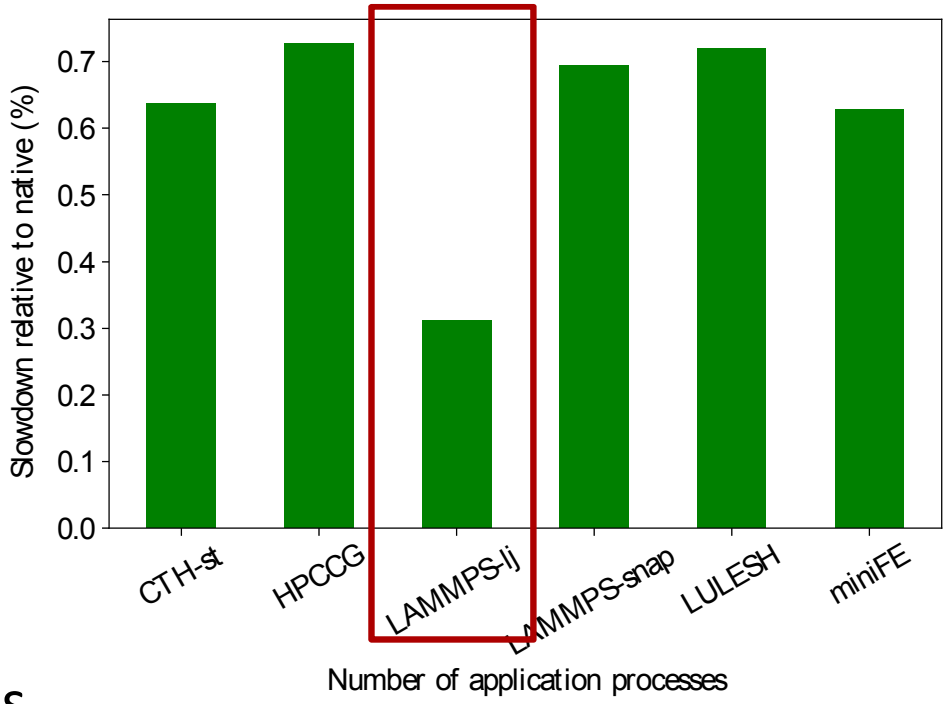
- Examined six workloads
  - LAMMPS : molecular dynamics simulation from Sandia National Laboratories. We used the LAMMPS SNAP and Lennard-Jones (LJ) potentials.
  - CTH : application from Sandia that models complex problems that are characterized by large deformations or strong shocks
  - HPCCG : conjugate gradient solver from the Mantevo suite of mini-applications
  - LULESH : proxy application that represents behavior typical of hydrocodes
  - miniFE : proxy app that captures the key behaviors of unstructured implicit finite element codes

Workload Performance



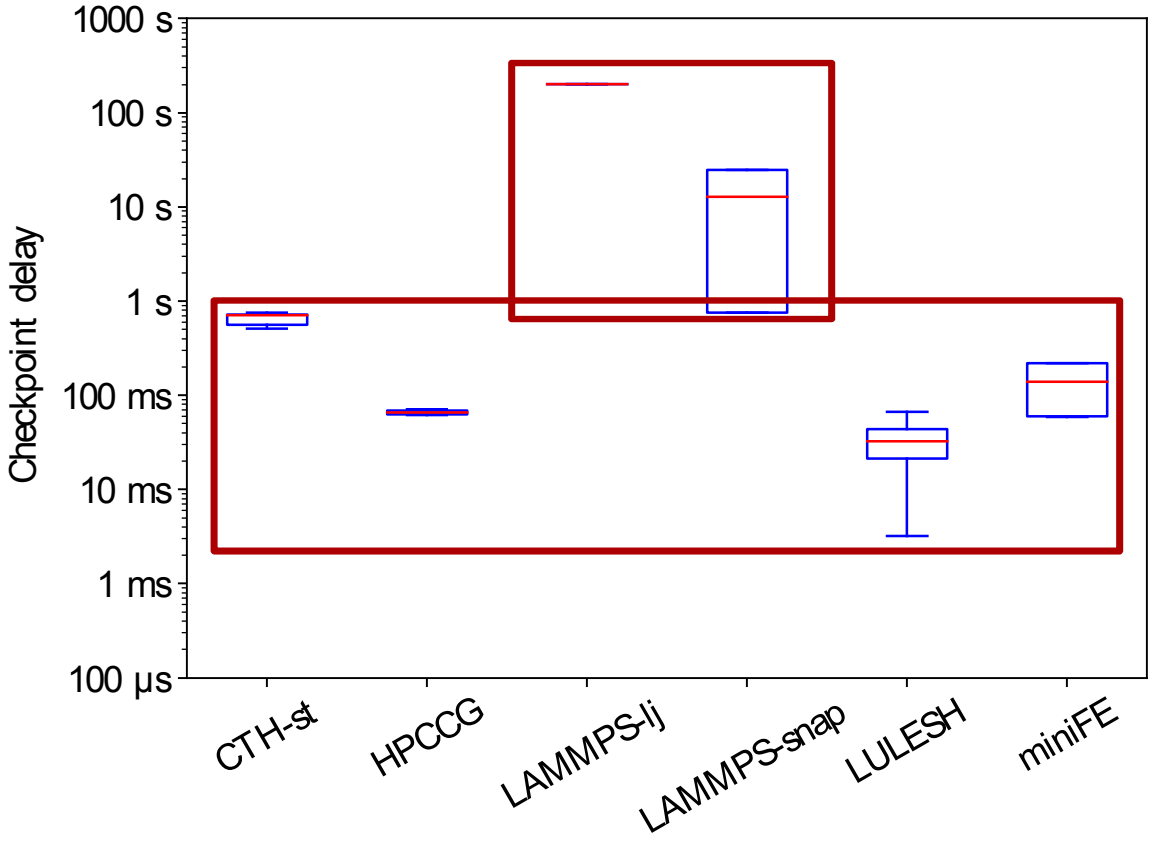
Cost of Approximately Coordinated C/R

- Isolated overhead on 32 Ki processes
- RECALL: overhead due to checkpointing ranges from (0.41% to 0.83%)
- Delays propagating along communication dependencies have a very modest impact on application performance

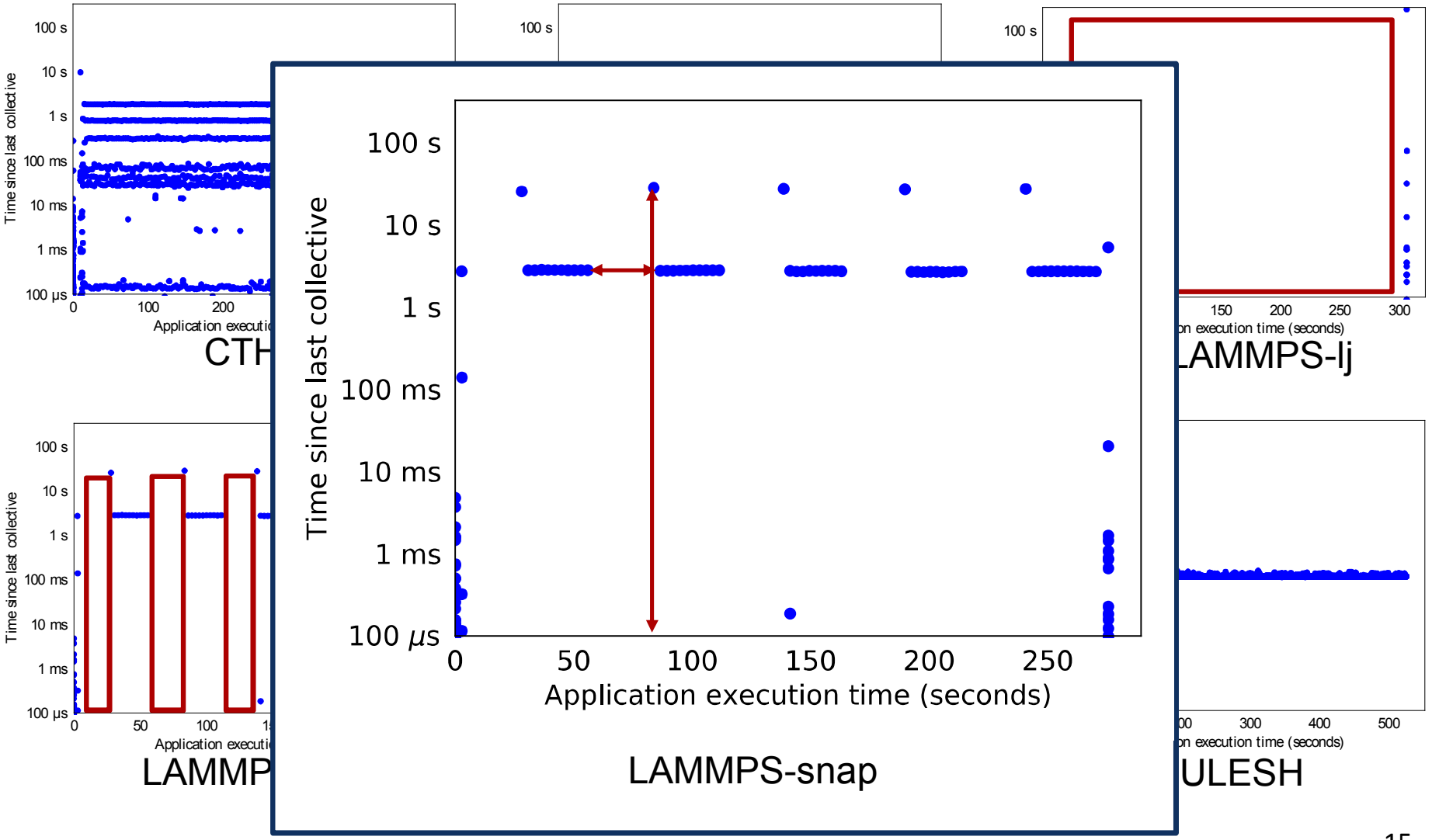


Checkpoint Interval (τ) Fidelity

- How does forcing checkpoints to occur after a collective operation impact the nominal checkpoint interval?



Collective Interarrival Intervals



Discussion

- Failures
  - initial study is on failure-free operation
  - checkpoint interval perturbation is small for several workloads, but for others (e.g., LAMMPS) we need to understand the consequences of altering the checkpoint interval
- Infrequent Collectives
  - for some workloads, the interval between collective operations may be much greater than the checkpoint interval
  - may require additional checkpoints to be inserted by the MPI runtime; we're still working out the details
- Message Logging
  - message logging is required because approximate coordination doesn't guarantee consistent state
  - …but approximate coordination may allow for efficient log purging

Conclusion

Sandia National Laboratories

- Leveraging the synchronization introduced by existing collective operations can significantly reduce the failure-free overhead of Uncoordinated Checkpoint/Restart

- Because some workloads use collective operations infrequently, additional mechanisms are necessary to ensure that the right balance is struck between the overhead of checkpointing and the cost of lost work

- Promising initial results; details still to be worked out

Co-authors

- Kurt B. Ferreira
  Sandia National Laboratories

- Patrick Widener
  Sandia National Laboratories

Questions?

sllevy@sandia.gov
www.sandia.gov/~sllevy

Sandia National Laboratories is a multi mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.