1. Background

**Multi-level checkpoint/restart (MLC)**
- Promising approach to address the problem
  - Uses multiple storage levels
  - Inexpensive local storage frequently
- Even with MLC, some checkpoints to the PFS are required to survive multi-node failures
  - e.g. 1) Rack level failure every 12 days on average in TSUBAME2.0
  - e.g. 2) 15% of production applications runs on Coastal, Hera and Atlas required to restart from a checkpoint in the PFS

**Checkpoint to PFS with the SCR library**
- Blocking checkpoint
  - Blocks the application until the flush has completed
- Non-blocking checkpoint
  - Another process flushes the checkpoint to the PFS in the background

2. Non-blocking checkpointing system

**Failure rates in HPC systems**
- Overall failure rate is increasing
  - e.g.) TSUBAME2.0@Tokyo Tech
    - About 962 node failures (Period: Nov, 2010 ~ April, 2012)
- In exascale systems, MTTI is projected to shrink to a few minutes
- Reliability of HPC systems is becoming more important for post-peta/exascale systems
- Checkpoint/Restart techniques are widely used in HPC systems

**Problems in Checkpoint/Restart**
- Checkpointing overhead to parallel file system (PFS)
  - 50GB checkpoint x 1408 thin nodes on TSUBAME2.0, Lustre (20GB/s) => About 5 hours for a checkpoint
- Huge workload by a large number of concurrent checkpoints

**Objective**
- Reduce checkpointing overhead & workload to PFS

3. Evaluation

**CPU-intensive application case**
- **Purpose**
  - To examine the impact on CPU-intensive applications with the non-blocking checkpointing system
- **Benchmark** IOR + CPU-intensive loop
- **Evaluation environment:** Sierra cluster at LLNL

<table>
<thead>
<tr>
<th>Sierra cluster</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.8 GHz 6-core Intel Xeon 5660 processor x 2 (= 12 cores)</td>
</tr>
<tr>
<td>Memory</td>
<td>128GB</td>
</tr>
<tr>
<td>Network</td>
<td>Infiniband QLogic 1322 QDR InfiniBand (4× = 32 GB/s)</td>
</tr>
<tr>
<td>File system (cache)</td>
<td>RAM 8 (tmp)</td>
</tr>
<tr>
<td>File system (PFS)</td>
<td>Lustre (block/striped throughput: 30 GB/s)</td>
</tr>
</tbody>
</table>

**Efficiency**
- **Model parameters**
  - Failure rate:
    - L1: 3.3309e-8 (A single node failure: System board, CPU, Memory etc.)
    - L2: 1.0186e-9 (multiple node failure: Shared PSU, Switch etc.)
  - Checkpoint size: 10Gbytes per node
  - PFS throughput: 200GB/s

**Evaluation environment:** Sierra cluster at LLNL

**Runtime with the different number of compute nodes per a transfer node**

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52- 07NA27344. LLNL-POST-561176