Towards a Light-weight Non-blocking Checkpointing System

Kento Sato†1, 6, Adam Moody†2, Kathryn Mohror†2, Todd Gamblin†2,
Bronis R. de Supinski†2, Naoya Maruyama†4, Satoshi Matsuoka†1, 3, 5

†1 Tokyo Institute of Technology
†2 Lawrence Livermore National Laboratory
†3 National Institute of Informatics
†4 AICS
†5 JST CREST
†6 JSPS Research Fellow

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

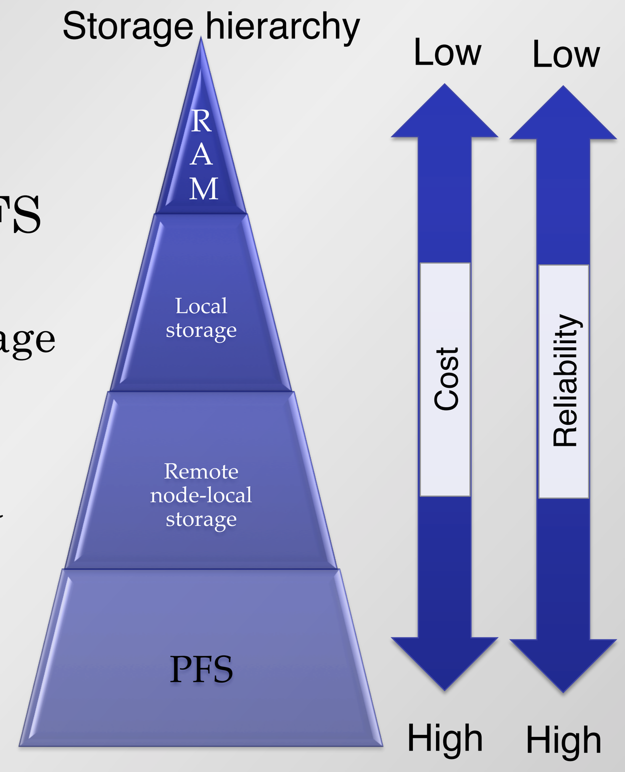
Objective

- Reduce checkpointing overhead & workload to PFS

1. Background

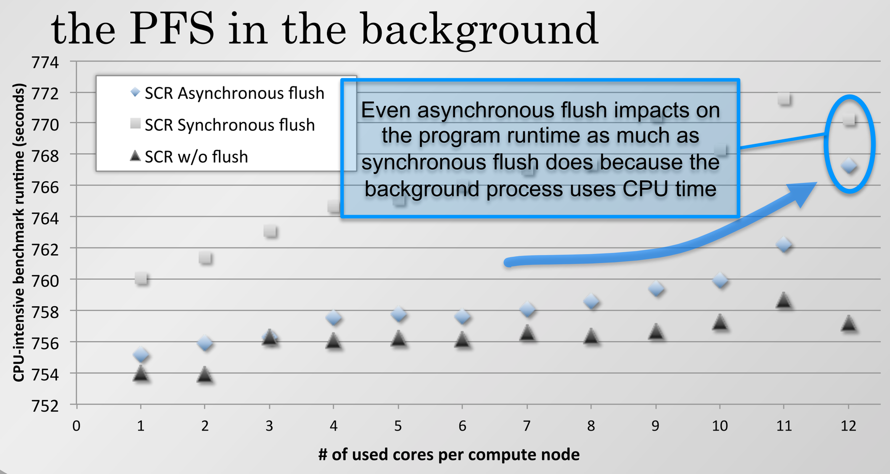
Multi-level checkpoint/restart (MLC)

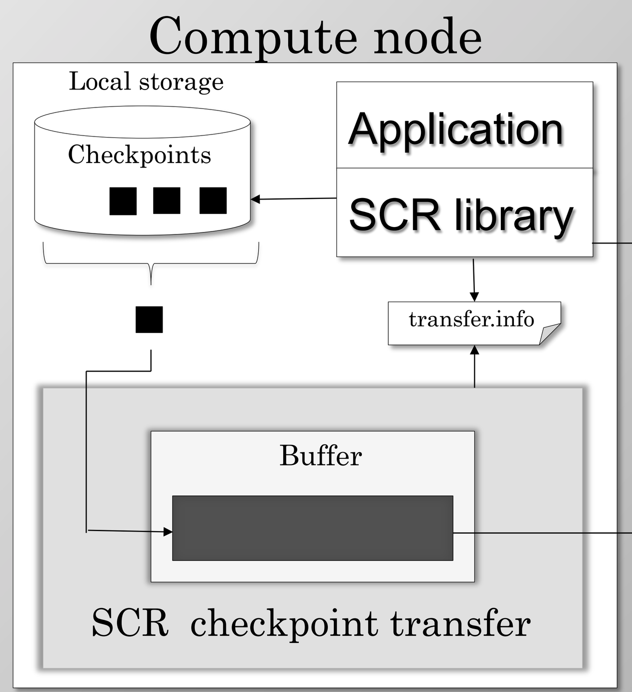
- Promising approach to address the problem
  - Uses multiple storage levels
  - Writes checkpoints to
    - Inexpensive local storage frequently
    - Reliable, but expensive PFS less frequently
- Even with MLC, some checkpoints to the PFS are required to survive multi-node failures
  - e.g. 1) Rack level failure every 12 days on average in TSUBAME2.0
  - e.g. 2) 15% of production application runs on Coastal, Hera and Atlas required to restart from a checkpoint in the PFS
- Problems in MLC
  - High PFS checkpoint cost
  - Failure due to heavy load on the PFS



Checkpoint to PFS with the SCR library

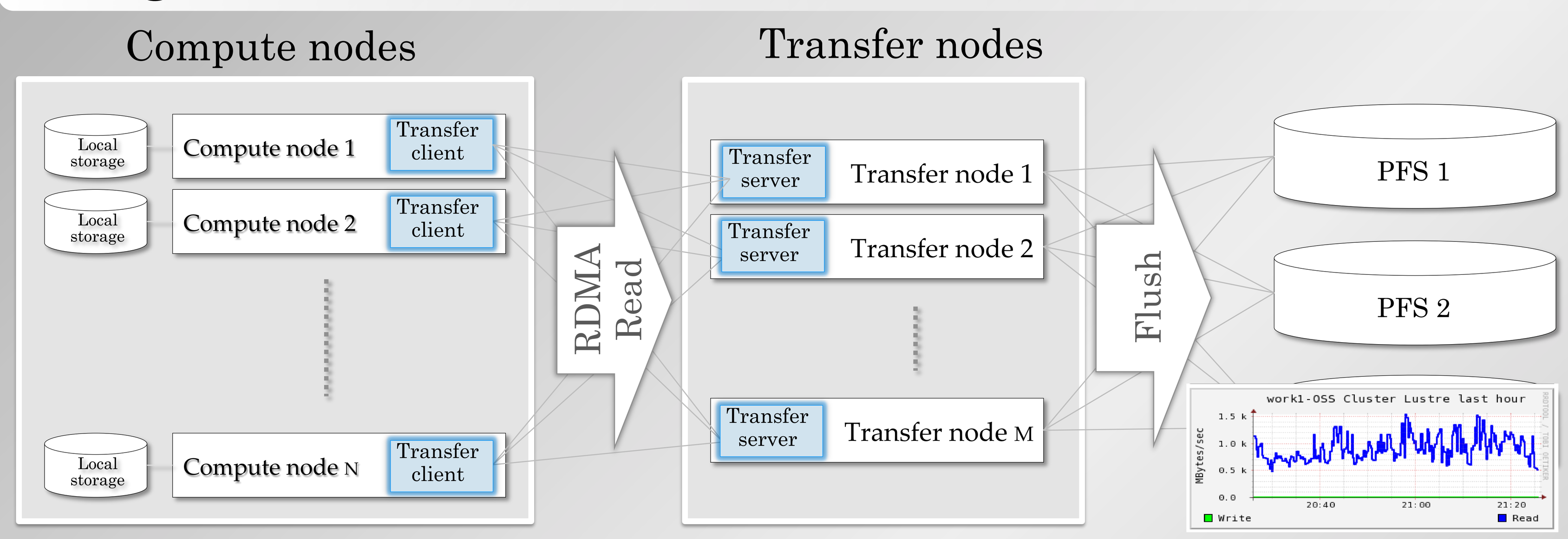
- Blocking checkpoint
  - Blocks the application until the flush has completed
- Non-blocking checkpoint
  - Another process flushes the checkpoint to the PFS in the background



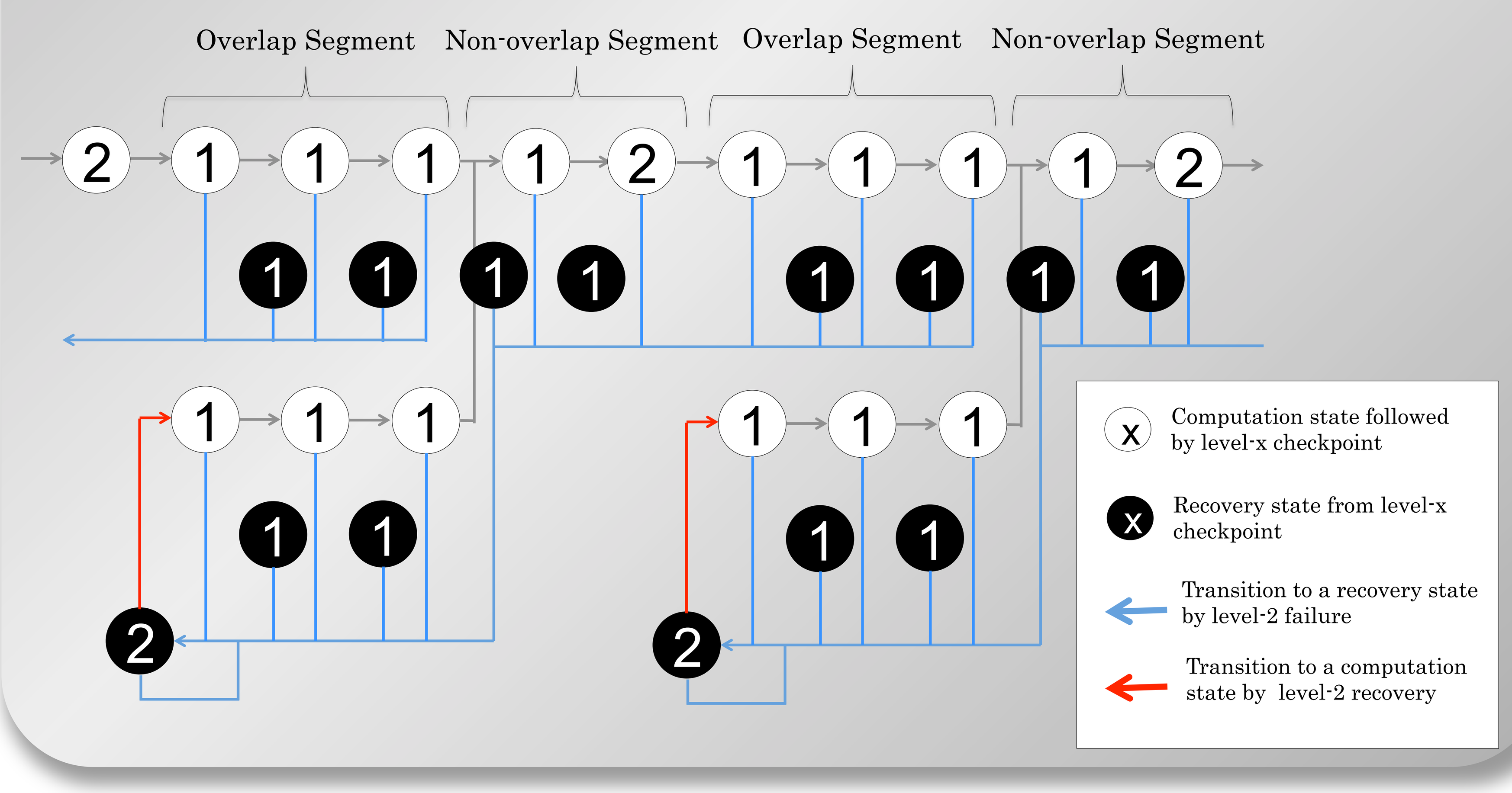


2. Non-blocking checkpointing system

Design



Modeling (Level 2 failures and recoveries)



3. Evaluation

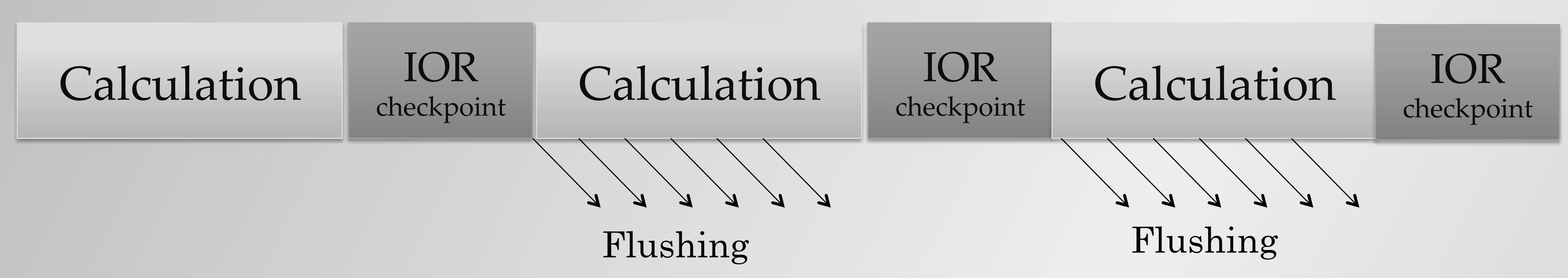
CPU-intensive application case

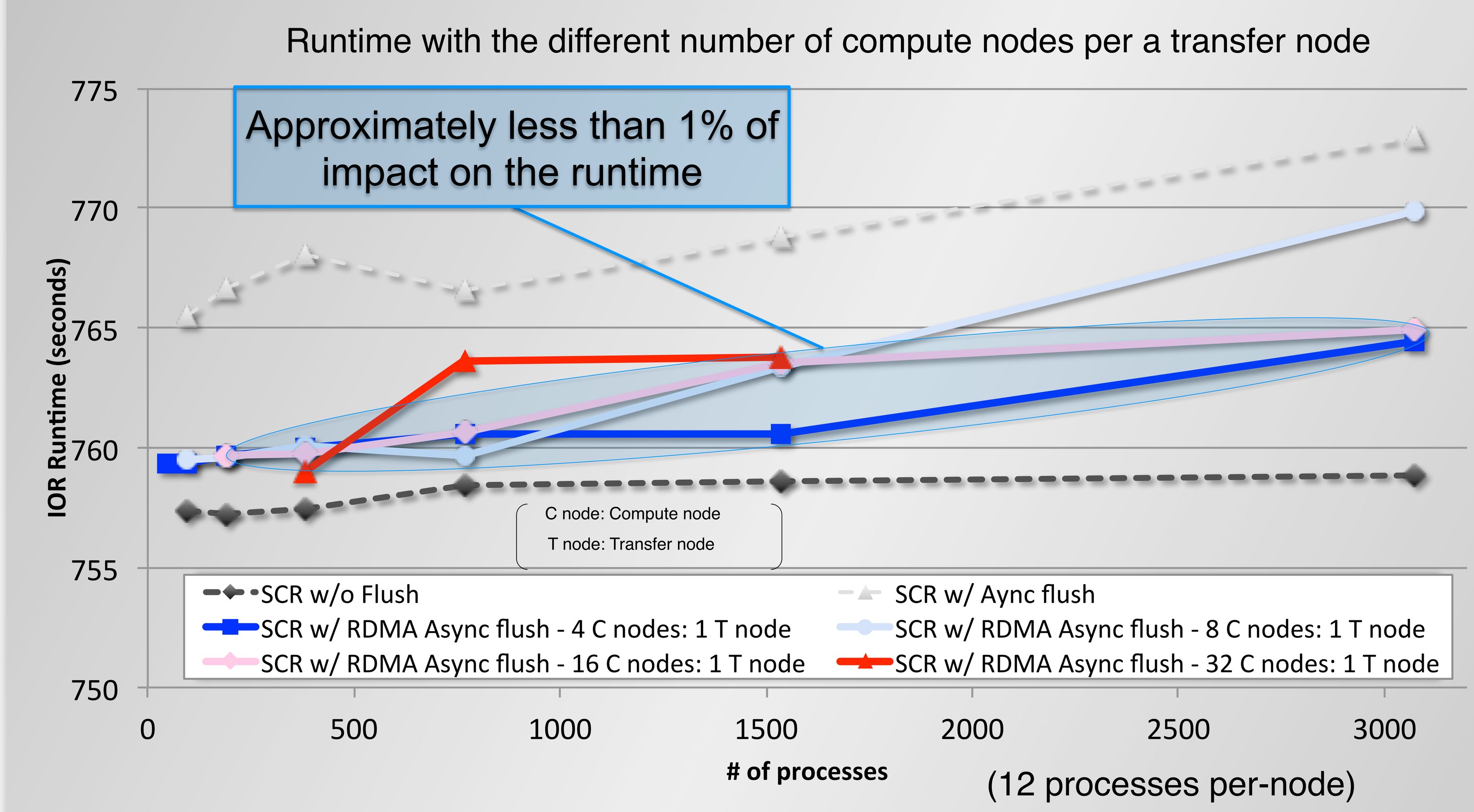
- Purpose
  - To examine that the impact on CPU-intensive applications with the non-blocking checkpointing system
- Benchmark: IOR + CPU-intensive loop
- Evaluation environment: Sierra cluster at LLNL

Sierra cluster

1944 nodes	
23,328 CPU cores	
261.3 TFlops (Peak)	

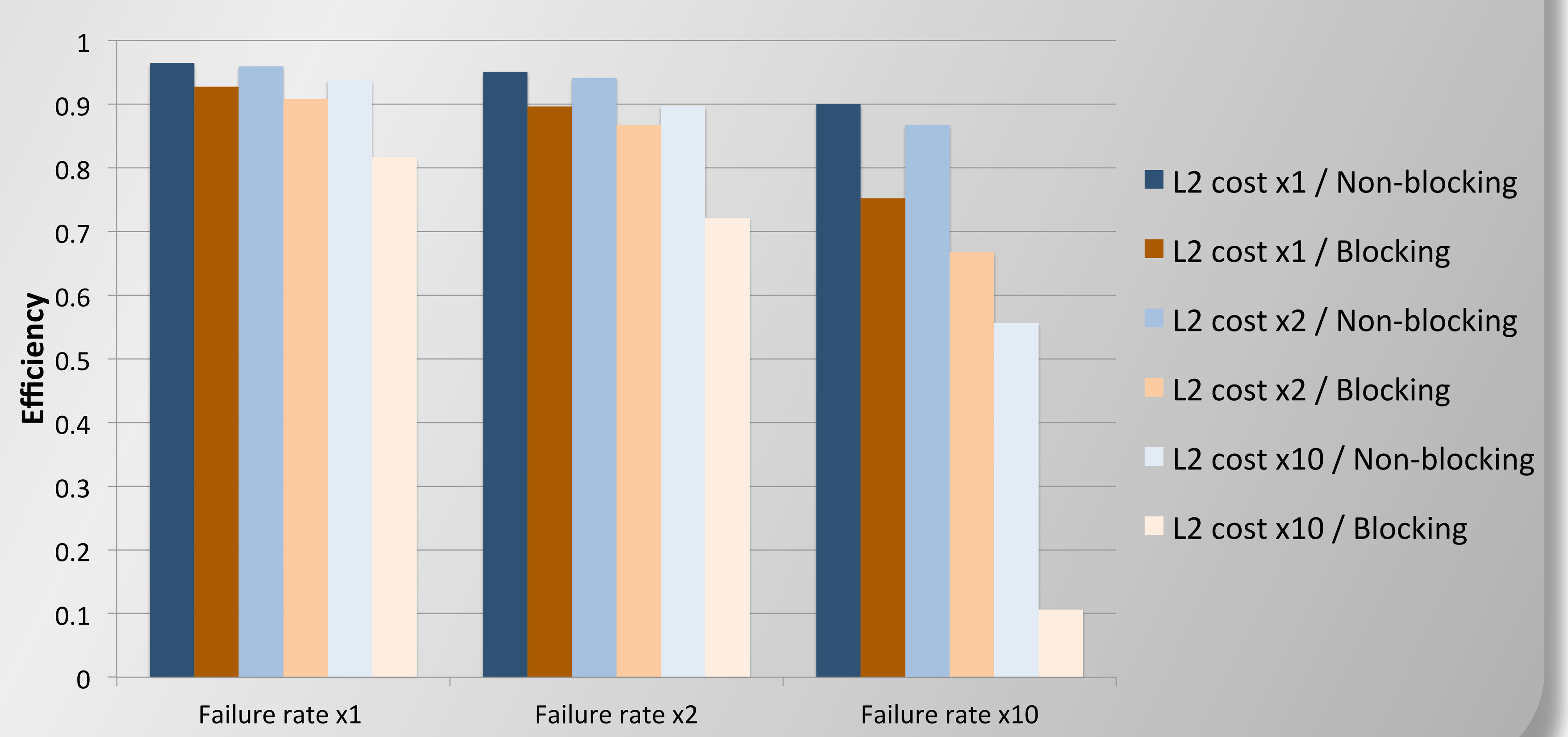
CPU	2.8 GHz 6-core Intel Xeon 5660 processor x 2 (= 12 cores)
Memory	24GB
Network	Qlogic IBA7322 QDR Infiniband 4x (= 32 Gbit/s)
File system (cache)	RAM fs (/tmp)
File system (PFS)	Lustre (lscratchc, theoretical throughput: 30 GB/s)





Efficiency

- Model parameters
  - Failure rate:
    - L1: 3.3308e-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: 20Gbytes/s



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