Making HPC Systems Resilient with Parallel Objects

Esteban Meneses[1][2], Laxmikant V. Kalé[3]

[1]Advanced Computing Laboratory, Costa Rica High Technology Center

[2]School of Computing, Costa Rica Institute of Technology

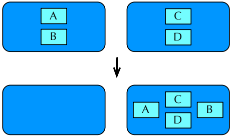
[3]Department of Computer Science, University of Illinois at Urbana-Champaign

SIAM Conference on Parallel Processing for Scientific Computing
SIAM-PP18

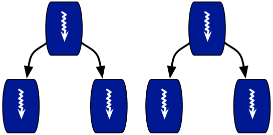
Resilience Techniques

A taxonomy

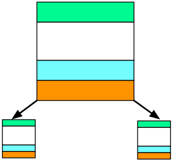
Prevention



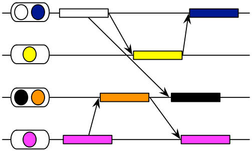
Detection



Containment



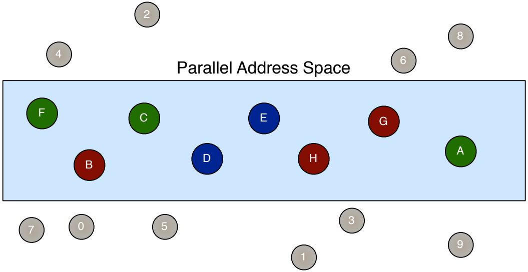
Recovery

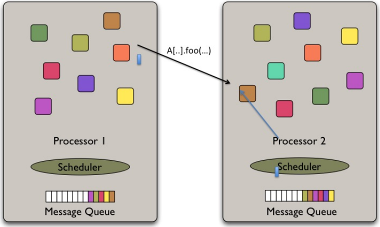


Esteban Meneses, Xiang Ni, Gengbin Zheng, Celso Mendes, Laxmikant Kalé. **Using Migratable Objects to Enhance Fault Tolerance Schemes in Supercomputers**. IEEE Transactions on Parallel and Distributed Systems (TPDS), 2015.

Parallel Objects

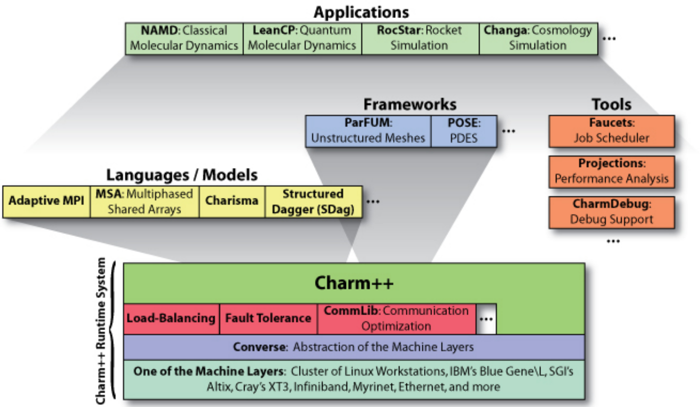
The Charm++ programming model





Introspective Runtime System

The Charm++ RTS



Applications

NAMD: Classical Molecular Dynamics | LeanCP: Quantum Molecular Dynamics | RocStar: Rocket Simulation | Changa: Cosmology Simulation | ...

Frameworks

ParFUM: Unstructured Meshes | POSE: PDES | ...

Tools

Faucets: Job Scheduler

Projections: Performance Analysis

CharmDebug: Debug Support

...

Languages / Models

Adaptive MPI | MSA: Multiphased Shared Arrays | Charisma | Structured Dagger (SDag) | ...

Charm++ Runtime System

Charm++

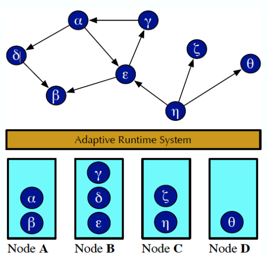
Load-Balancing | Fault Tolerance | CommLib: Communication Optimization | ...

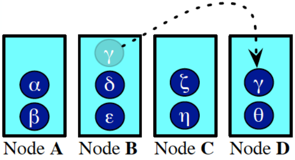
Converse: Abstraction of the Machine Layers

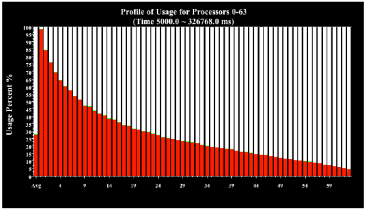
One of the Machine Layers: Cluster of Linux Workstations, IBM's Blue Gene\L, SGI's Altix, Cray's XT3, Infiniband, Myrinet, Ethernet, and more

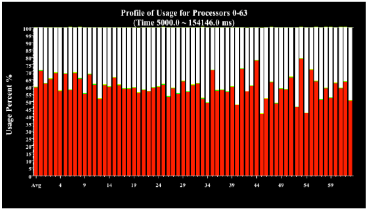
Object Migratability

Load balancing framework



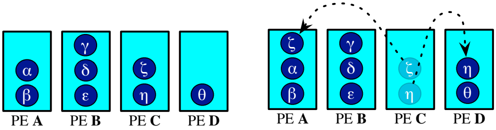


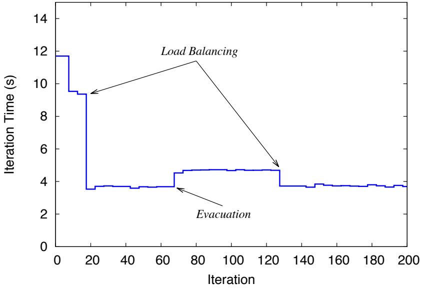




Proactive Fault Tolerance

Evacuation of a faulty node

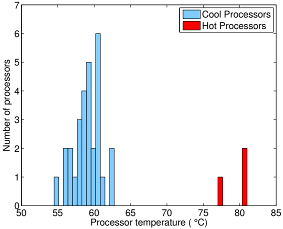


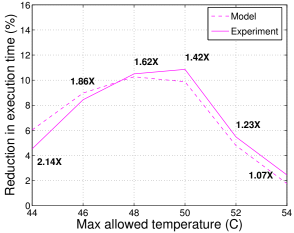


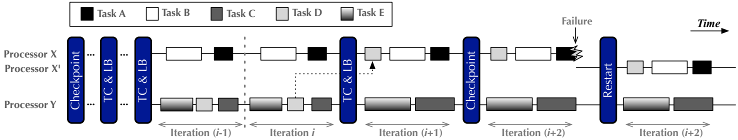
Sayantan Chakravorty, Celso Mendes, and Laxmikant Kalé. **Proactive Fault Tolerance in MPI Applications via Task Migration**. IEEE International Conference on High Performance Computing (HiPC), 2006.

Controllable Resilience

Restraining processor temperature to reduce failure frequency



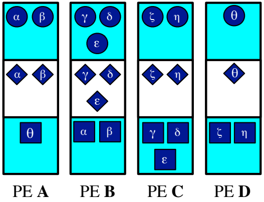


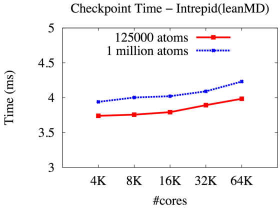


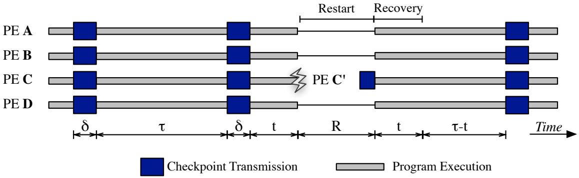
Osman Sarood, Esteban Meneses, and Laxmikant Kalé. **A Cool Way of Improving the Reliability of HPC Machines**. International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2013.

Checkpoint/Restart

Leveraging object migratability



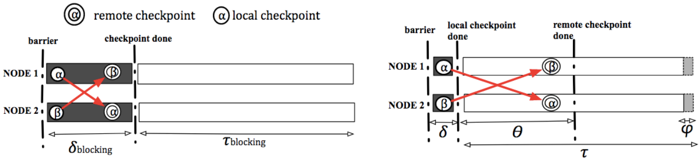




Gengbin Zheng, Xiang Ni, and Laxmikant Kalé. **A Scalable Double In-memory Checkpoint and Restart Scheme towards Exascale**. Workshop on Fault-Tolerance for HPC at Extreme Scale (FTXS), 2012.

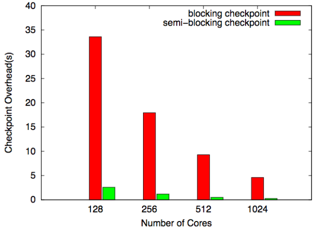
Semi-blocking Checkpoint/Restart

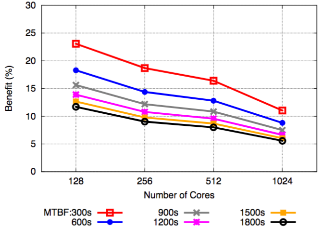
Overlapping checkpoint and communication transmission



(a) Blocking Checkpoint.

(b) Semi-Blocking Checkpoint.

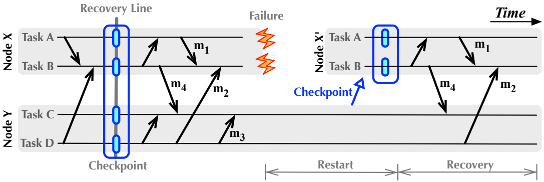


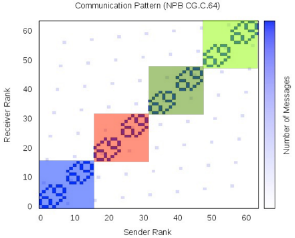


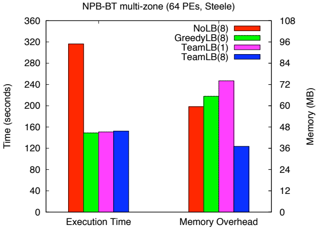
Xiang Ni, Esteban Meneses, and Laxmikant Kalé. **Hiding Checkpoint Overhead in HPC Applications with a Semi-Blocking Algorithm**. IEEE International Conference on Cluster Computing (Cluster), 2012.

Message Logging

Communication is stored and replayed after a failure



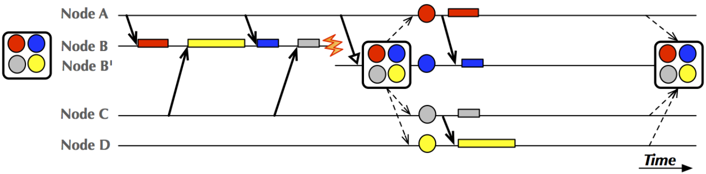


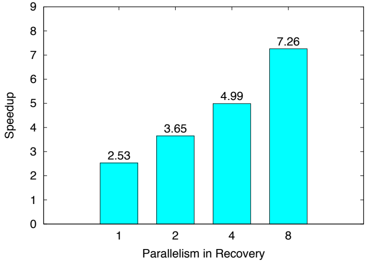


Esteban Meneses, Greg Bronevetsky, and Laxmikant Kalé. **Dynamic Load Balance for Optimized Message Logging in Fault Tolerant HPC Applications**. IEEE International Conference on Cluster Computing (Cluster), 2011.

Parallel Recovery

Migrate objects to speed up recovery

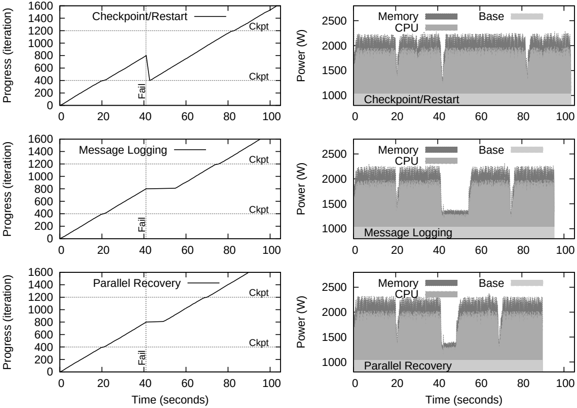




Sayantan Chakravorty and Laxmikant Kalé. **A Fault Tolerance Protocol with Fast Fault Recovery**. IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2007.

Advantages of Message Logging

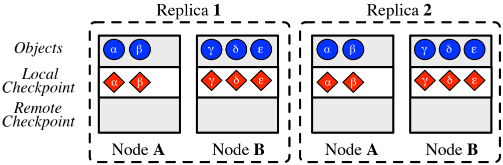
Faster and greener

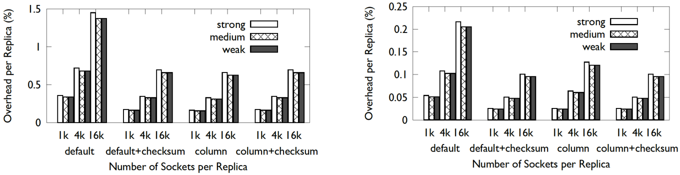


Esteban Meneses, Osman Sarood, and Laxmikant Kalé. **Energy Profile of Rollback-Recovery Strategies in High Performance Computing**. Parallel Computing (ParCo), 2014.

ACR: Automatic Checkpoint/Restart

Soft and hard error protection





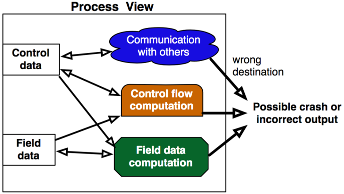
(a) Jacobi3D Charm++

(b) LeanMD

Xiang Ni, Esteban Meneses, Nikhil Jain, and Laxmikant Kalé. **ACR: Automatic Checkpoint/Restart for Soft and Hard Error Protection**. International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2013.

FlipBack

Automatic targeted protection against silent data corruption



Process View

Control data

Communication with others

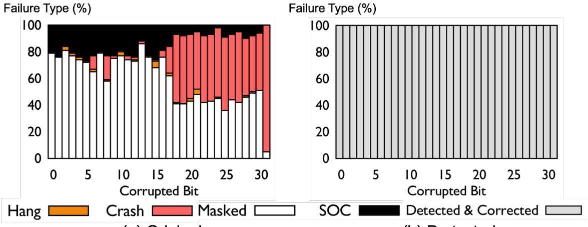
Control flow computation

Field data

Field data computation

wrong destination

Possible crash or incorrect output



(a) Original

(b) Protected

Xiang Ni and Laxmikant Kalé. **FlipBack: Automatic Targeted Protection Against Silent Data Corruption**. International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2016.

Acknowledgments

- Discussion with Dr. Harshitha Menon from Lawrence Livermore National Laboratory.
- Discussion with Dr. Xiang Ni from IBM T.J. Watson Research Center.
- Travel grant from School of Computing at Costa Rica Institute of Technology.
- Travel funds from Costa Rica National High Technology Center.

Concluding Remarks

- Parallel objects provide a fertile ground to **enhance resilience techniques**
- **Adaptivity and introspection** at the core of novel strategies
- Objects are natural **failure containment units**

Thank You!
Q&A

CeNAT
Centro Nacional de Alta Tecnología

TEC
Tecnológico
de Costa Rica