# Improving Fault Resilience of High Performance Applications

Yawei Li and Zhiling Lan

Illinois Institute of Technology, Contact: {liyawei,lan}@iit.edu

## The Problem

Reliability is becoming a fundamental challenge to the continuous scaling of HPC

- Failure rates accelerate dramatically as the size and complexity of HPC systems grows, e.g. tens-of-thousands to hundreds-of-thousands of processing components
- The inherent parallel paradigm makes HPC applications more failure-prone, e.g. a single component failure crashes the entire application

A new fault tolerance approach is needed for the present and future HPC

- The conventional checkpointing/recovery approach is not efficient, e.g. rollback, overhead, downtime…
- The emerging proactive approach is not reliable, e.g. false alarms and prediction misses

## Proposed Solution

Fault learning and prediction [1]

- Statistical learning
- Rule-based mining
- Advanced learning

Performance-based adaptive strategy[2]

- Opportunistic SKIP, to reduce unnecessary fault tolerance operations;
- Selective CKP, to reduce potential performance loss caused by unforeseeable failures;
- Preemptive migration, to avoid anticipated failures

System support [3,4]

- System-wide fault-driven resource allocation and rescheduling
- Augment of open source MPI package

## On-going Research

- Online fault prediction
- Coupling of application adaptation with system adaptation
- Integration of different components as an end-to-end package
- Extensive evaluation and validation

## Case Studies

- Integrate adaptive fault management with MPICH-V
- Compare with periodic CPR
- Testbed: IA32 cluster at TG/ANL
- Use an actual failure trace of HPC system

- Production HPC applications
  - (1) NPB benchmarks
  - (2) Cosmology application ENZO
  - (3) Molecular dynamic application GROMACS

- Investigate the following issues:
  - (1) Impact of computing scale
  - (2) Impact of prediction accuracy
  - (3) Impact of spare node allocation

---

**Reference**


