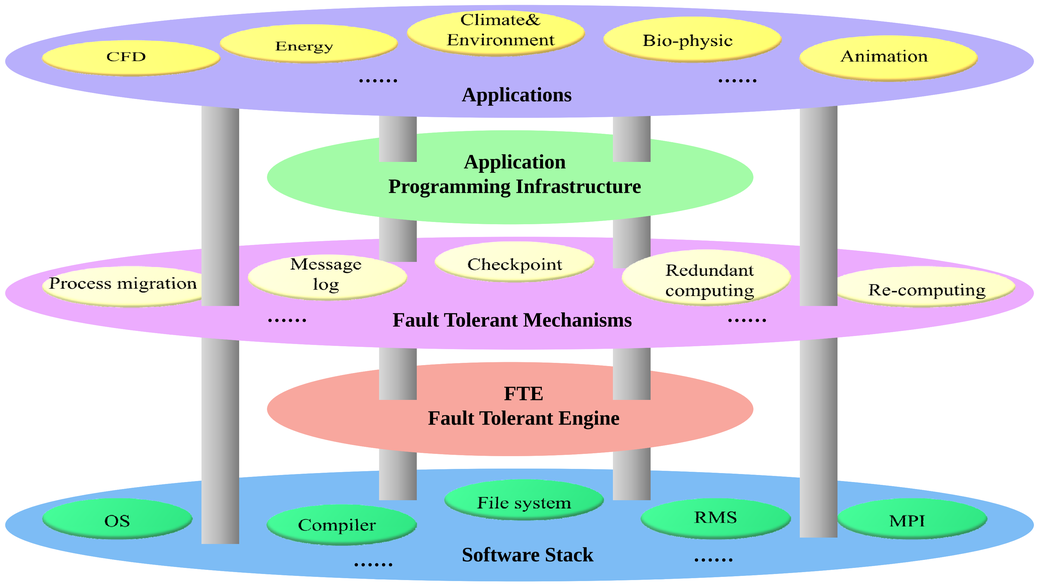
# Resilience Computing Framework For Large-scale High Performance Computing System

Prof. Yutong Lu (ytlu@nudt.edu.cn)
Lab of System Software
Department of Computer Science
National University of Defense Technology

## Abstract

Fault tolerance becomes a more and more important issue for HPC systems, especially for the post-petascale and future Exascale computing. We propose an intelligent resilience computing framework，based on the fault tolerant engine, involved in the full system software stack. In addition, this framework is consisted of fault detection and dealing, cooperated with multiple data recovery mechanisms, to realize the scalable resilience computing for various longtime-running large parallel applications. We demonstrate a fault resilience MPI system (NR-MPI) based on this framework，which supports recovery of corrupted communicator and non-blocking collective operations. Using our framework, NR-MPI can also be integrated with application programming infrastructures to implement user transparent resilience computing when encounter faults in the system.

## Resilience Computing Framework

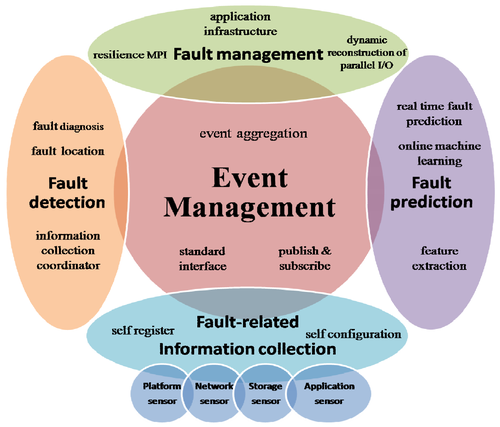



- Capability to support post-petaflops and Exascale computing
- Collaboration with whole system software stack
- Coherent fault detection
- Coordinate fault tolerant decision
- Cooperation of multiple fault recovery mechanics
- Combination of proactive and reactive strategies
- Customizable fault detection, prediction and recovery approaches
- Support various parallel models

## Research Background

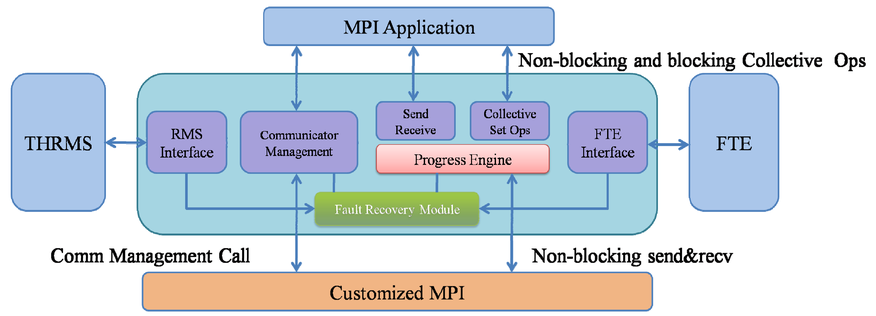
- Development of full system software stack for domestic YH/TH HPC
  - Kernel, Compiler, GLEX, THMPI, RMS, HAPF, HPVZ
  - Optimization focusing on performance, scalability and reliability
  - Large scale applications on complex science and engineering
- Research on fault tolerance
  - Compiler-guided error detection & Error propagation in parallel program
  - Fault tolerant parallel algorithm
  - System level checkpoint/restart for MPI & OpenMP applications

## Fault Tolerant Engine




- Fault related information collection based on sensors
- Fault detection based on rules
- Fault prediction based on machine learning
- Event management based on publishing and subscribing
- Scalable and low-overhead fault-related information sharing and distribution
- Coherent fault detection and fault location
- Online learning and real time fault prediction

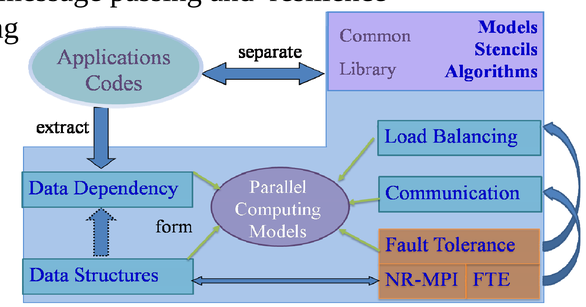
## NR-MPI




- Integrated with existing MPI (MPICH/OpenMPI)
- Relay on FTE providing fault information
- Reconstruction of broken communicator
- Recovery of ranks' connections
- Provide non-blocking fault tolerant collective operations
- Provide flexible data recovery mechanisms, CR/Redundant computing/APP

## Application transparent Resilience Computing

- Domain-specific Application programming Infrastructure
  - Domain scientists: focus on physical problem, algorithm, parameters
  - App-Infrastructure: handles parallelization, message passing and resilience
- Application transparent Resilience Computing
  - Now for Jasmin
- Potential fault tolerance pillar for multiple domain-specific application Infrastructures
  - Future for others
    --OpenFoam --mpiBlast
    --Nwchem --NAMD
    --Matlab etc.




## Open Issues

- New Fault tolerant programming model
- Light-weight fault resilience mechanisms
- Domain-specific application validation