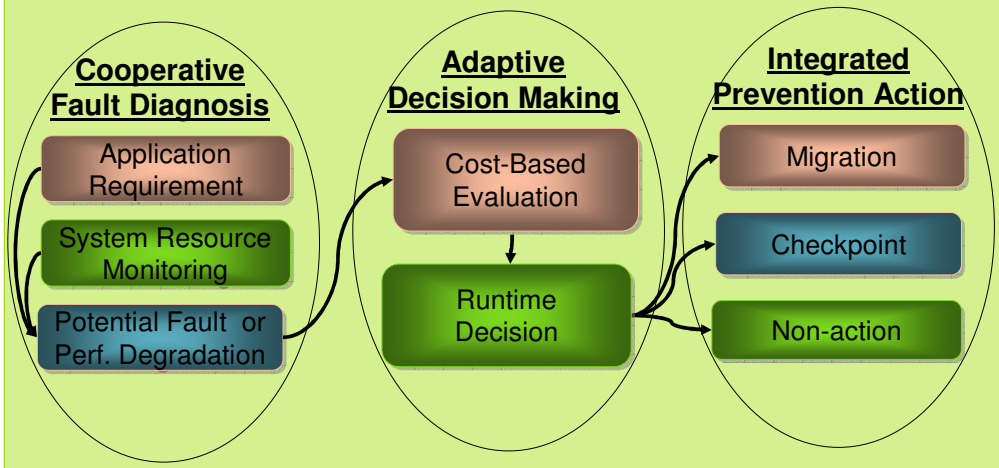
FT-Pro: Adaptive Fault Management for High Performance Computing

Motivations:

- Ever-increasing system size and complexity
- Long-running computational applications
- Progress in failure prediction via HW sensors and SW detection techniques

Project Summary:

- "Intelligently" perform process migration or checkpointing based on the cooperative diagnosis between application and system
- To improve the performance and reliability of HPC applications despite failures (e.g. HW/SW failures or performance degradation)



Analytical Modeling:

- Derive the performance expression for expected execution time:

$$E(T) = T + \frac{c_f F(T)}{\overline{F(T)}} + \frac{\int_0^T t dF(t)}{\overline{F(T)}}$$

$$E_a(T) = G(T) + \int_0^T E_a(T) f(T-t)dt, G(T) = \int_0^T (t + c_f + c_r(t - L_{cp}(t)) + c_{cp}\lfloor N(t)\rfloor)dF(t) + (T + c_{cp}\lfloor N(t)\rfloor)\overline{F(T)}$$

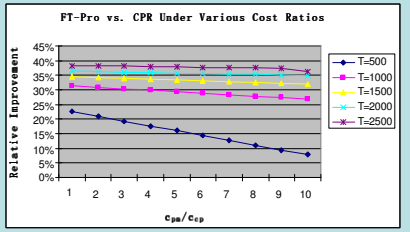
$$E_{fp}(T) = \sum_{j=1}^{K}(E_a(T_{pm}(j) - T_{pm}(j-1) + c_{pm}) + E_a(T - T_{pm}(K)), T_{pm}(j) \text{ the currence time of the jth migration}$$

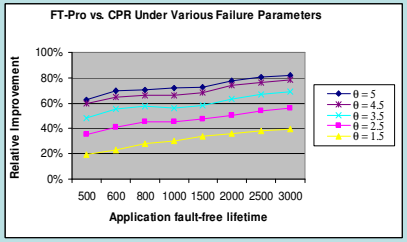
- Prove the optimal migration placement in IFR systems

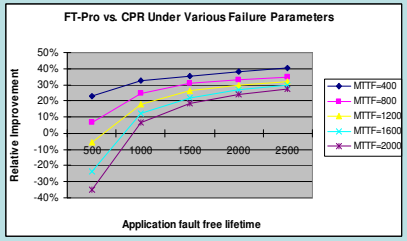
Experiments:

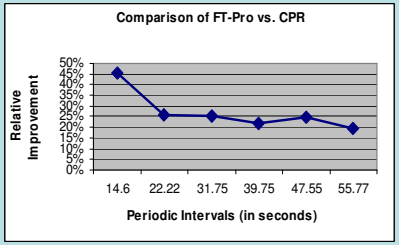
- Weibull failure distributions
- FT-Pro vs. Checkpointing (CPR) under various conditions:
  - Different failure parameters
  - Different cost ratios
- Relative Improvement = $\frac{(\cos t_{CPR} - \cos t_{FT-Pro})}{\cos t_{CPR}}$

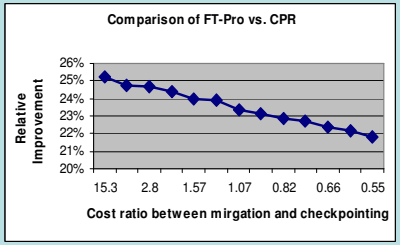
Fault-Free Exec. Time	Aperiodic Checkpointing	FT-Pro
100	100.36	101.594
200	202.37	203.204
500	512.205	509.852
600	616.532	612.698
800	826.794	819.299
1000	1038.88	1027.06
1500	1576.82	1550.77
2000	2125.84	2080.45
2500	2686.77	2615.64
3000	3260.1	3156.36











Contact:

Zhiling Lan and Yawei Li
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616

Email: {lan,liyawei}@iit.edu