www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Outlook: Fault Tolerance in MPI Programs

Janko Strassburg

PATC Parallel Programming Workshop October 2013

With material from W. Gropp, E. Lusk,
Argonne National Laboratory

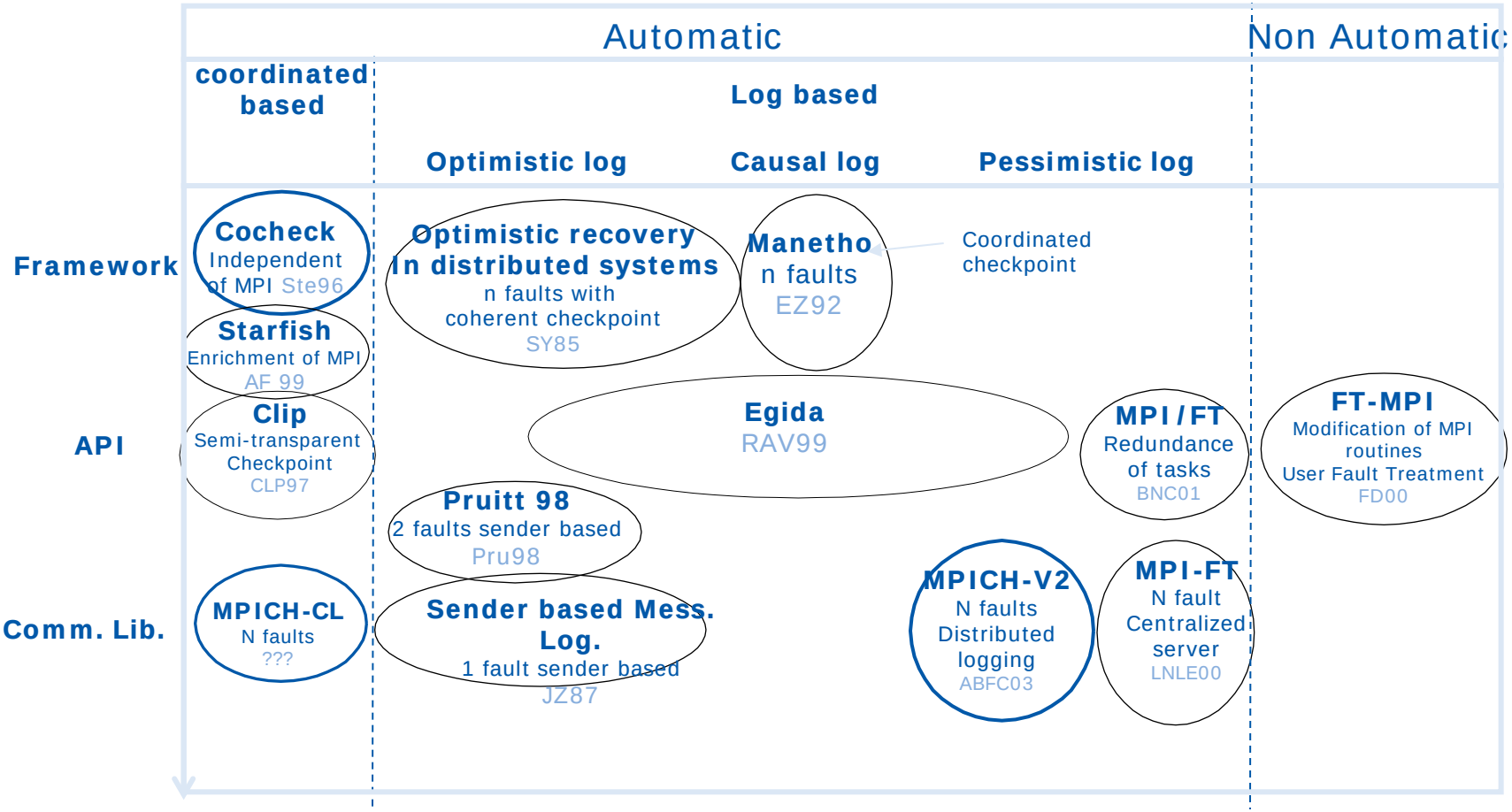
Contents

- Declaration
- Existing FT MPI
- FT & MPI standard
- Write (non-transparent) FT in MPI
- Summary & discussion

Declaration

- Fault tolerance is a property of a program, not of an API specification or an implementation.
- Within certain constraints, MPI can provide a useful context for writing application programs that exhibit significant degrees of fault tolerance.

Current FT MPI



Fault Tolerance & MPI standard

- FT is a property of an MPI program coupled with the MPI implementation.

- Four lever of “survive”
  - Automatically recovers (MPICH)
  - Error notification (FT-MPI)
  - Failure can be ignore (Manager/worker)
  - Restart from checkpoint (CoCheck etc)

Fault Tolerance & MPI standard

- MPI Standard does mention about the FT.
  - Require to implement reliable communication
  - Built in or user defined error handlers
  - Predefined error

Writing FT App in MPI

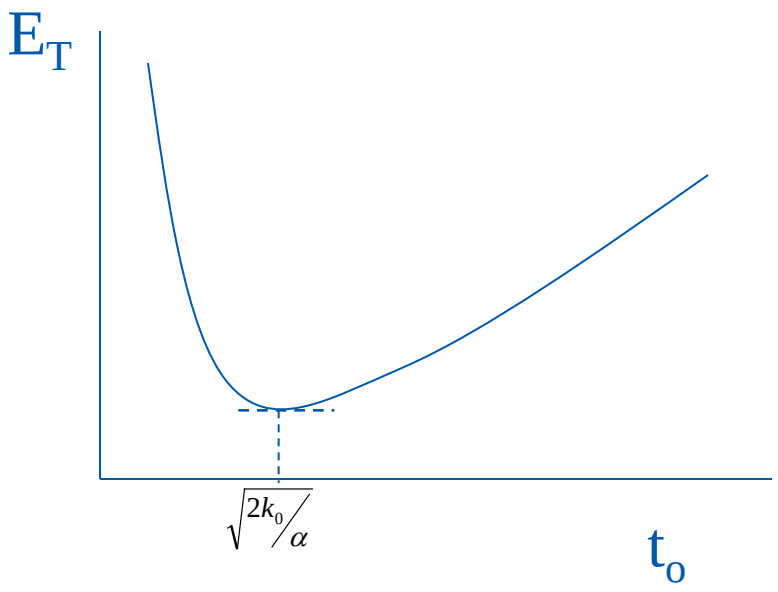
- Basic approach
  - Checkpointing & roll back
    - System directed
    - User directed
  - Redundancy & vote
- Approach technique
  - MPI
  - Modify / Extend MPI

- $E_T = T(1 + k_0/t_0 + a(k_1 + t_0/2))$
- $0 = dE_T/dt_0 = -k_0/t_0^2 + a/2$

$$t_0 = \sqrt{2k_0/\alpha}$$

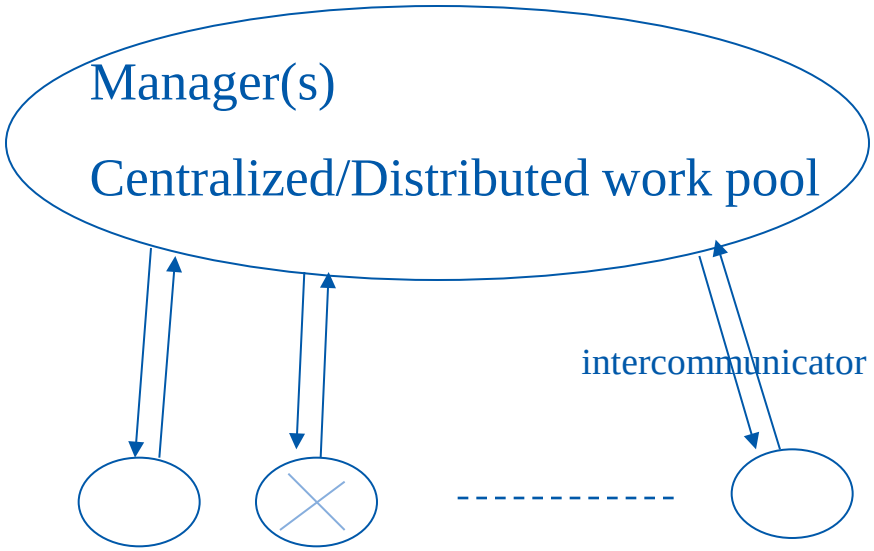
$$E_T = T(1 + \alpha k_1 + \sqrt{2\alpha k_0})$$

Additional cost



Use intercommunicators

Manager/Worker Model



Worker processors

The intermediate status of the computing is stored at the manager party.

Modify/Extend MPI

- Modify MPI Semantics
  - Break the constrain of the MPI semantics
  - Provider the programmer more error information and error handling methods
- Extending MPI
  - Define extensions to MPI (MPE_XXX)
  - Encapsulate the MPI procedures

Summary

- MPI Standard provides in the way of support for writing fault-tolerant programs.
- Many approach could be used to write the “nontransparent” FT MPI program.