1 Fault Tolerance Techniques for Distributed, Parallel Applications

Camille Coti
LIPN, CNRS UMR 7030, Université Paris 13, Sorbonne Paris Cité
99, avenue Jean-Baptiste Clément
F-93430 Villetaneuse, FRANCE

References

ABD07. Thara Angskun, George Bosilca, and Jack Dongarra. Binomial graph: A scalable and fault-tolerant logical network topology. In *Parallel and Distributed Processing and Applications*, pages 471–482. Springer, 2007.

ACD+10. Emmanuel Agullo, Camille Coti, Jack Dongarra, Thomas Herault, and Julien Langou. QR factorization of tall and skinny matrices in a grid computing environment. In *24th IEEE International Parallel & Distributed Processing Symposium (IPDPS'10)*, Atlanta, Ga, April 2010.

ACT97. Marcos Kawazoe Aguilera, Wei Chen, and Sam Toueg. Heartbeat: A timeout-free failure detector for quiescent reliable communication. In Marios Mavronicolas and Philippas Tsigas, editors, *Proceedings of the 11th Workshop on Distributed Algorithms (WDAG'97)*, volume 1320 of *Lecture Notes in Computer Science*, pages 126–140. Springer, 1997.

AER+99. L. Alvisi, E. Elnozahy, S. Rao, S.A. Husain, and A. de Mel. An analysis of communication induced checkpointing. In *Fault-Tolerant Computing, 1999. Digest of Papers. Twenty-Ninth Annual International Symposium on*, pages 242–249, June 1999.

AFB+10. Thara Angskun, Graham Fagg, George Bosilca, Jelena Pješivac-Grbović, and Jack Dongarra. Self-healing network for scalable fault-tolerant runtime environments. *Future Generation Computer Systems*, 26(3):479–485, 2010.

AM95. Lorenzo Alvisi and Keith Marzullo. Message logging : Pessimistic, optimistic, and causal. In *Proceedings of the 15th International Conference on Distributed Computing Systems (ICDCS 1995)*, pages 229–236. IEEE CS Press, May-June 1995.

BBC+02. George Bosilca, Aurélien Bouteiller, Franck Cappello, Samir Djilali, Gilles Fédak, Cécile Germain, Thomas Hérault, Pierre Lemarinier, Oleg Lodygensky, Frédéric Magniette, Vincent Néri,

and Anton Selikhov. MPICH-V: Toward a scalable fault tolerant MPI for volatile nodes. In *High Performance Networking and Computing (SC2002)*, Baltimore USA, November 2002. IEEE/ACM.

BCH+03. Aurélien Bouteiller, Franck Cappello, Thomas Hérault, Géraud Krawezik, Pierre Lemarinier, and Frédéric Magniette. MPICH-V2: a fault tolerant MPI for volatile nodes based on pessimistic sender based message logging. In *High Performance Networking and Computing (SC2003)*. Phoenix USA, IEEE/ACM, November 2003.

BCH+05. Aurelien Bouteiller, Boris Collin, Thomas Herault, Pierre Lemarinier, and Franck Cappello. Impact of event logger on causal message logging protocols for fault tolerant MPI. In *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05)*, page 97, Washington, DC, USA, 2005. IEEE Computer Society.

BCH+08. Darius Buntinas, Camille Coti, Thomas Herault, Pierre Lemarinier, Laurence Pilard, Ala Rezmerita, Eric Rodriguez, and Franck Cappello. Blocking vs. non-blocking coordinated checkpointing for large-scale fault tolerant MPI. *Future Generation Computer Systems*, 24 (1):73–84, 2008. Digital Object Identifier: http://dx.doi.org/10.1016/j.future.2007.02.002.

BCH+09. George Bosilca, Camille Coti, Thomas Herault, Pierre Lemarinier, and Jack Dongarra. Constructing resilient communication infrastructure for runtime environments. In *International Conference in Parallel Computing (ParCo2009)*, Lyon, France, September 2009.

BDDL09. George Bosilca, Remi Delmas, Jack Dongarra, and Julien Langou. Algorithm-based fault tolerance applied to high performance computing. *J. Parallel Distrib. Comput.*, 69(4):410–416, 2009.

BGL01. Ralph Butler, William D. Gropp, and Ewing L. Lusk. Components and interfaces of a process management system for parallel programs. volume 27, pages 1417–1429, 2001.

BHK+06. Aurélien Bouteiller, Thomas Herault, Géraud Krawezik, Pierre Lemarinier, and Franck Cappello. MPICH-V project: A multiprotocol automatic fault-tolerant mpi. *International Journal of High Performance Computing Applications*, 20(3):319–333, 2006.

BLKC03. Aurélien Bouteiller, Pierre Lemarinier, Géraud Krawezik, and Franck Cappello. Coordinated checkpoint versus message log for fault tolerant MPI. In *IEEE International Conference on Cluster Computing (Cluster 2003)*. IEEE CS Press, December 2003.

CHL+06. Camille Coti, Thomas Herault, Pierre Lemarinier, Laurence Pilard, Ala Rezmerita, Eric Rodriguez, and Franck Cappello. Blocking vs. non-blocking coordinated checkpointing for large-scale fault tolerant MPI. In ACM/IEEE, editor, *Proceedings of the International Conference for High Performance Networking Computing, Networking, Storage and Analysis (SC—06)*, page electronic, Tampa, USA, November 2006.

CL85. K. Mani Chandy and Leslie Lamport. Distributed snapshots : Determining global states of distributed systems. In *Transactions on Computer Systems*, volume 3(1), pages 63–75. ACM, February 1985.

Cot15. Camille Coti. Exploiting redundant computation in communication-avoiding algorithms for algorithm-based fault tolerance. *CoRR*, abs/1511.00212, November 2015.

DGG10. Simplice Donfack, Laura Grigori, and Alok Kumar Gupta. Adapting communication-avoiding lu and qr factorizations to multicore architectures. In *Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on*, pages 1–10. IEEE, 2010.

DGHL08. James Demmel, Laura Grigori, Mark Hoemmen, and Julien Langou. Communication-avoiding parallel and sequential QR factorizations. *CoRR*, abs/0806.2159, 2008.

EAWJ96. Elmootazbellah Elnozahy, Lorenzo Alvisi, Yi-Min Wang, and David B. Johnson. A survey of rollback-recovery protocols in message passing systems. Technical Report CMU-CS-96-181, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, October 1996.

EAWJ02. Elmootazbellah Elnozahy, Lorenzo Alvisi, Yi-Min Wang, and David B. Johnson. A survey of rollback-recovery protocols in message-passing systems. *ACM Computing Surveys (CSUR)*, 34(3):375 – 408, september 2002.

FD00. GrahamE. Fagg and JackJ. Dongarra. Ft-mpi: Fault tolerant mpi, supporting dynamic applications in a dynamic world. In Jack Dongarra, Peter Kacsuk, and Norbert Podhorszki, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, volume 1908 of *Lecture Notes in Computer Science*, pages 346–353. Springer Berlin Heidelberg, 2000.

FGB+03. Graham E. Fagg, Edgar Gabriel, George Bosilca, Thara Angskun, Zhizhong Chen, Jelena Pjesivac-Grbovic, Kevin London, and Jack J. Dongarra. Extending the MPI specification for process

fault tolerance on high performance computing systems. In *In Proceeding of International Supercomputer Conference (ICS*, 2003.

FGC+05. Graham E Fagg, Edgar Gabriel, Zizhong Chen, Thara Angskun, George Bosilca, Jelena Pjesivac-Grbovic, and Jack J Dongarra. Process fault tolerance: Semantics, design and applications for high performance computing. *International Journal of High Performance Computing Applications*, 19(4):465–477, 2005.

FLP+10. Daniel Ford, François Labelle, Florentina I Popovici, Murray Stokely, Van-Anh Truong, Luiz Barroso, Carrie Grimes, and Sean Quinlan. Availability in globally distributed storage systems. In *OSDI*, pages 61–74, 2010.

For94. Message Passing Interface Forum. MPI: A message-passing interface standard. Technical Report UT-CS-94-230, Department of Computer Science, University of Tennessee, April 1994. Tue, 22 May 101 17:44:55 GMT.

GS97. Rachid Guerraoui and André Schiper. Software-based replication for fault tolerance. *Computer*, (4):68–74, 1997.

HLAD09. B. Hadri, H. Ltaief, E. Agullo, and J. Dongarra. Tall and skinny qr matrix factorization using tile algorithms on multicore architectures. Technical report, Innovative Computing Laboratory, University of Tennessee, September 2009.

HML09. Joshua Hursey, Timothy I Mattox, and Andrew Lumsdaine. Interconnect agnostic checkpoint/restart in open mpi. In *Proceedings of the 18th ACM international symposium on High performance distributed computing*, pages 49–58. ACM, 2009.

HMR97. Jean-Michel Hélary, Achour Mostefaoui, and Michel Raynal. Virtual precedence in asynchronous systems: Concept and applications. In Marios Mavronicolas and Philippas Tsigas, editors, *Distributed Algorithms*, volume 1320 of *Lecture Notes in Computer Science*, pages 170–184. Springer Berlin Heidelberg, 1997.

HMR99. Jean-Michel Hélary, Achour Mostefaoui, and Michel Raynal. Communication-induced determination of consistent snapshots. *IEEE Transactions on Parallel and Distributed Systems*, 10(9):865–877, 1999.

JD03. Eric Roman Jason Duell, Paul Hargrove. The design and implementation of berkeley lab's linux checkpoint/restart. Technical Report publication LBNL-54941, Berkeley Lab, 2003.

KBJ02. James Stevens Klecka, William F Bruckert, and Robert L Jardine. Error self-checking and recovery using lock-step processor pair architecture, May 21 2002. US Patent 6,393,582.

Lam78. Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7):558–565, 1978.

Lan10. Julien Langou. Computing the R of the QR factorization of tall and skinny matrices using MPI_Reduce. *arXiv preprint arXiv:1002.4250*, 2010.

LBH+04. Pierre Lemarinier, Aurélien Bouteiller, Thomas Herault, Géraud Krawezik, and Franck Cappello. Improved message logging versus improved coordinated checkpointing for fault tolerant MPI. In *IEEE International Conference on Cluster Computing (Cluster 2004)*. IEEE CS Press, 2004.

LTBL97. Michael J. Litzkow, Todd Tannenbaum, Jim Basney, and Miron Livny. Checkpoint and migration of UNIX processes in the condor distributed processing system. Technical Report 1346, University of Wisconsin-Madison, 1997.

LV62. Robert E Lyons and Wouter Vanderkulk. The use of triple-modular redundancy to improve computer reliability. *IBM Journal of Research and Development*, 6(2):200–209, 1962.

PBKL95. James S. Plank, Micah Beck, Gerry Kingsley, and Kai Li. Libckpt: Transparent checkpointing under unix. In *USENIX Winter*, pages 213–224, 1995.

PKD95. James S Plank, Youngbae Kim, and Jack J Dongarra. Algorithm-based diskless checkpointing for fault tolerant matrix operations. In *Fault-Tolerant Computing, 1995. FTCS-25. Digest of Papers., Twenty-Fifth International Symposium on*, pages 351–360. IEEE, 1995.

PLP98. James S Plank, Kai Li, and Michael Puening. Diskless checkpointing. *Parallel and Distributed Systems, IEEE Transactions on*, 9(10):972–986, 1998.

RdLM06. Daniel A. Reed, Charng da Lu, and Celso L. Mendes. Reliability challenges in large systems. *Future Generation Computer Systems*, 22(3):293 – 302, 2006.

SSB+05. Sriram Sankaran, Jeffrey M Squyres, Brian Barrett, Vishal Sahay, Andrew Lumsdaine, Jason Duell, Paul Hargrove, and Eric Roman. The lam/mpi checkpoint/restart framework: System-initiated checkpointing. *International Journal of High Performance Computing Applications*, 19(4):479–493, 2005.

WSVM13. Meg Walraed-Sullivan, Amin Vahdat, and Keith Marzullo. Aspen trees: balancing data center fault tolerance, scalability and cost. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, pages 85–96. ACM, 2013.

ZSK04. Gengbin Zheng, Lixia Shi, and Laxmikant V Kalé. Ftc-charm++: an in-memory checkpoint-based fault tolerant runtime for charm++ and mpi. In *Cluster Computing, 2004 IEEE International Conference on*, pages 93–103. IEEE, 2004.