1 Scalable run-time environments for large-scale parallel applications

Camille Coti ♣ and Franck Cappello ♠ ♦

♣ LIPN, CNRS-UMR7030, Université Paris 13, F-93430 Villetaneuse, France
♠ INRIA Saclay-Île de France, Orsay, F-91893, France
♦ UIUC, T. M. Siebel Center for Computer Sscience, URBANA, IL 61801-2302, USA

References

1. 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.

2. Thara Angskun, George Bosilca, and Jack Dongarra. Binomial graph: A scalable and fault-tolerant logical network topology. In Ivan Stojmenovic, Ruppa K. Thulasiram, Laurence Tianruo Yang, Weijia Jia, Minyi Guo, and Rodrigo Fernandes de Mello, editors, *Proceedings of the 5th International Symposium on Parallel and Distributed Processing and Applications (ISPA 2007)*, volume 4742 of *Lecture Notes in Computer Science*, pages 471–482. Springer, 2007.

3. Thara Angskun, George Bosilca, and Jack Dongarra. Self-healing in binomial graph networks. In Robert Meersman, Zahir Tari, and Pilar Herrero, editors, *OTM Workshops (2)*, volume 4806 of *Lecture Notes in Computer Science*, pages 1032–1041. Springer, 2007.

4. Thara Angskun, Graham E. Fagg, George Bosilca, Jelena Pjesivac-Grbovic, and Jack Dongarra. Self-healing network for scalable fault-tolerant runtime environments. *Future Generation Comp. Syst.*, 26(3):479–485, 2010.

5. Pavan Balaji, Darius Buntinas, David Goodell, William Gropp, Jayesh Krishna, Ewing Lusk, and Rajeev Thakur. Pmi: A scalable parallel process-management interface for extreme-scale systems. In Rainer Keller, Edgar Gabriel, Michael Resch, and Jack Dongarra, editors, *Recent Advances in the Message Passing Interface*, volume 6305 of *Lecture Notes in Computer Science*, pages 31–41. Springer Berlin / Heidelberg, 2010.

6. B. Barrett, J. Squyres, A. Lumsdaine, R. Graham, and G. Bosilca. Analysis of the component architecture overhead in open mpi. In Beniamino Di Martino, Dieter Kranzlmüller, and Jack Dongarra, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, volume 3666 of *Lecture Notes in Computer Science*, pages 175–182. Springer Berlin / Heidelberg, 2005.

7. Dan Bonachea and Jason Duell. Problems with using MPI 1.1 and 2.0 as compilation targets for parallel language implementations. *International Journal of High Performance Computing and Networking (IJHPCN)*, 1(1/2/3):91–99, 2004.

8. 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.

9. George Bosilca, Camille Coti, Thomas Herault, Pierre Lemarinier, and Jack Dongarra. Constructing resiliant communication infrastructure for runtime environments. *Advances in Parallel Computing*, 19:441–451, April 2010. Digital Object Identifier: http://dx.doi.org/10.1016/j.future.2007.02.002.

10. 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.

11. 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.

12. 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.

13. Joshua Bruck, Ching-Tien Ho, Shlomo Kipnis, Eli Upfal, and Derrick Weathersby. Efficient algorithms for all-to-all communications in multiport message-passing systems. *IEEE Transactions on Parallel and Distributed Systems*, 8(11):1143–1156, November 1997.

14. 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.

15. Ralph M. Butler, William D. Gropp, and Ewing L. Lusk. A scalable process-management environment for parallel programs. In Jack Dongarra, Péter Kacsuk, and Norbert Podhorszki, editors, *Recent Advances*

in Parallel Virtual Machine and Message Passing Interface, 7th European PVM/MPI Users' Group Meeting (EuroPVM/MPI'02), volume 1908, pages 168–175. Springer, 2000.

16. Nicolas Capit, Georges Da Costa, Yiannis Georgiou, Guillaume Huard, Cyrille Martin, Grégory Mounié, Pierre Neyron, and Olivier Richard. A batch scheduler with high level components. In *Proceedings of the 5th International Symposium on Cluster Computing and the Grid (CCGRID'05)*, pages 776–783, Cardiff, UK, May 2005. IEEE Computer Society.

17. Franck Cappello, Eddy Caron, Michel Dayde, Frederic Desprez, Yvon Jegou, Pascale Vicat-Blanc Primet, Emmanuel Jeannot, Stephane Lanteri, Julien Leduc, Nouredine Melab, Guillaume Mornet, Benjamin Quetier, and Olivier Richard. Grid'5000: A large scale and highly reconfigurable grid experimental testbed. In *Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing CD (SC—05)*, pages 99–106, Seattle, Washington, USA, November 2005. IEEE/ACM.

18. Ralph H. Castain, Timothy S. Woodall, David J. Daniel, Jeffrey M. Squyres, Brian Barrett, and Graham E. Fagg. The open run-time environment (openRTE): A transparent multi-cluster environment for high-performance computing. In Beniamino Di Martino, Dieter Kranzlmüller, and Jack Dongarra, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface, 12th European PVM/MPI Users' Group Meeting*, volume 3666 of *Lecture Notes in Computer Science*, pages 225–232, Sorrento, Italy, September 2005. Springer.

19. T.D. Chandra, V. Hadzilacos, S. Toueg, and B. Charron-Bost. Impossibility of group membership in asynchronous systems. In *Proceedings of the 15th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, pages 322–330, May 1996.

20. 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.

21. Zizhong Chen, Graham E. Fagg, Edgar Gabriel, Julien Langou, Thara Angskun, George Bosilca, and Jack Dongarra. Fault tolerant high performance computing by a coding approach. In Keshav Pingali, Katherine A. Yelick, and Andrew S. Grimshaw, editors, *Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP 2005, June 15-17, 2005, Chicago, IL, USA*, pages 213–223. ACM, 2005.

22. C. Cristian Coarfa, Yuri Dotsenko, John Mellor-Crummey, Daniel Chavarria-Miranda, Francois Cantonnet, Tarek El-Ghazawi, Ashrujit Mo-

hanti, and Yiyi Yao. An evaluation of global address space languages: Co-array fortran and unified parallel. June 2005.

23. UPC Consortium. UPC Language Specifications, v1.2. Technical Report LBNL-59208, Lawrence Berkeley National, 2005.

24. Camille Coti. *Environnements d'exécution pour applications parallèles communiquant par passage de messages pour les systèmes à grande échelle et les grilles calcul.* PhD thesis, Université Paris Sud-XI, November 2009.

25. Camille Coti, Thomas Herault, Sylvain Peyronnet, Ala Rezmerita, and Franck Cappello. Grid services for MPI. In Thierry Priol et al, editor, *Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid'08)*, pages 417–424, Lyon, France, May 2008. ACM/IEEE.

26. Leonardo Dagum and Ramesh Menon. OpenMP: an industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng.*, 5:46–55, January 1998.

27. Graham E. Fagg and Jack Dongarra. FT-MPI: Fault tolerant MPI, supporting dynamic applications in a dynamic world, 2000.

28. Graham E. Fagg and Jack J. Dongarra. HARNESS fault tolerant MPI design, usage and performance issues. *Future Generation Computer Systems*, 18(8):1127–1142, October 2002.

29. 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.

30. Edgar Gabriel, Graham E. Fagg, George Bosilca, Thara Angskun, Jack J. Dongarra, Jeffrey M. Squyres, Vishal Sahay, Prabhanjan Kambadur, Brian Barrett, Andrew Lumsdaine, Ralph H. Castain, David J. Daniel, Richard L. Graham, and Timothy S. Woodall. Open MPI: Goals, concept, and design of a next generation MPI implementation. In *Recent Advances in Parallel Virtual Machine and Message Passing Interface, 11th European PVM/MPI Users' Group Meeting (EuroPVM/MPI'04)*, pages 97–104, Budapest, Hungary, September 2004.

31. Al Geist, William D. Gropp, Steven Huss-Lederman, Andrew Lumsdaine, Ewing L. Lusk, William Saphir, Anthony Skjellum, and Marc Snir. MPI-2: Extending the message-passing interface. In Luc Bougé, Pierre Fraigniaud, Anne Mignotte, and Yves Robert, editors, *1st European Conference on Parallel and Distributed Computing (EuroPar'96)*, volume 1123 of *Lecture Notes in Computer Science*, pages 128–135. Springer, 1996.

32. William Gropp. Mpich2: A new start for mpi implementations. In Dieter Kranzlmller, Jens Volkert, Peter Kacsuk, and Jack Dongarra, editors,

Recent Advances in Parallel Virtual Machine and Message Passing Interface, volume 2474 of *Lecture Notes in Computer Science*, pages 37–42. Springer Berlin / Heidelberg, 2002.

33. William D. Gropp and Ewing L. Lusk. A high-performance MPI implementation on a shared-memory vector supercomputer. *Parallel Computing*, 22(11):1513–1526, January 1997.

34. Amina Guermouche, Thomas Ropars, Elisabeth Brunet, Marc Snir, and Franck Cappello. Uncoordinated checkpointing without domino effect for send-deterministic message passing applications. May 2011.

35. P. Hilfinger, D. Bonachea, K. Datta, D. Gay, S. Graham, B. Liblit, G. Pike, J. Su, and K. Yelick. Titanium language reference manual. Technical Report UCB/CSD-2005-15, U.C. Berkeley, 2005.

36. Morris A. Jette, Andy B. Yoo, and Mark Grondona. SLURM: Simple linux utility for resource management. In Dror G. Feitelson, Larry Rudolph, and Uwe Schwiegelshohn, editors, *Proceedings of the 9th International Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP'03)*, volume 2862, pages 44–60. Springer-Verlag, 2003.

37. 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.

38. Raymond Namyst. *Contribution á la conception de supports exécutifs multithreads performants*. Habilitation á diriger des recherches, Université Claude Bernard de Lyon, pour des travaux effectués á l'école normale supérieure de Lyon, DEC 2001.

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

40. Richard D. Schlichting and Fred B. Schneider. Fail stop processors: An approach to designing fault-tolerant computing systems. *ACM Transactions on Computer Systems*, 1:222–238, 1983.

41. Gerard Tel. *Introduction to Distributed Algorithms*. Cambridge University Press, 1994.