Fault Tolerance for PetaScale Systems: Current Knowledge, Challenges and Opportunities

Franck Cappello
INRIA
Email: fci@lri.fr

Abstract

The emergence of PetaScale systems reinvigorates the community interest about how to manage failures in such systems and ensure that large applications successfully complete. This talk starts by addressing the question of failure rate and trend in large systems, like the ones we find at the top of the top500. Where the failures come from and why we should pay more attention to them than in the past? A review of existing techniques for fault tolerance will be presented: rollback-recovery, failure prediction and proactive migration. We observe that rollback recovery has been deeply studied in the past years, resulting in a lot of optimizations; but is this enough to solve the challenge of fault tolerance raised by Petascale systems? What is the actual knowledge about failure prediction? Could we use it for proactive process migration and if so what benefit could we expect? Unfortunately, despite their high degree of optimization, existing approaches do not fit well with the challenging evolutions of large-scale systems. Thus, through this review of existing solutions and the presentation of the latest research results, we will list a set of open issues. Most of existing key mechanisms for fault tolerance come from the distributed system theory and the Chandy-Lamport algorithm for the determination of consistent global states. We should probably continue to optimize them, like by adding hardware dedicated to fault tolerance. Beside, there is room and even a need for new approaches. Opportunities may come from different origins, such as: 1) other fault tolerance approaches that consider failures as normal events in the system and 2) new algorithmic approaches, inherently fault tolerant. We will sketch some of these opportunities and their associated limitations.

Franck Cappello holds a Senior Researcher position at INRIA. He leads the Grand-Large project at INRIA, focusing on High Performance issues in Large Scale Distributed Systems. He has initiated the XtremWeb (Desktop Grid) and MPICH-V (Fault tolerant MPI) projects. He was the director of the Grid5000 project, a nation wide computer science platform for research in large scale distributed systems. He is the scientific director of ALADDIN/Grid5000, the new 4 years INRIA project aiming to sustain the Grid5000 infrastructure and to open it to researches in Cloud Computing, Service Infrastructures and the Future Internet. He has authored papers in the domains of High Performance Programming, Desktop Grids, Grids and Fault tolerant MPI. He has contributed to more than 50 Program Committees. He is editorial board member of the international Journal on Grid Computing, Journal of Grid and Utility Computing and Journal of Cluster Computing. He is a steering committee member of IEEE HPDC and IEEE/ACM CCGRID. He is the General co-Chair of IEEE APSCC 2008, Workshop co-chair for IEEE CCGRID'2008, Program co-Chair of IEEE CCGRID'2009 and was the General Chair of IEEE HPDC'2006.