Approaches for Parallel Applications Fault Tolerance

Richard L. Graham

Advanced Computing Laboratory
Los Alamos National Laboratory
Los Alamos, NM 87544, USA
rlgraham@lanl.gov

System component failure - hardware and software, permanent and transient - are an integral part of the life cycle of any computer system. The degree to which a system suffers from these failures depends on factors such as system complexity, system design and implementation, and system size. These errors may lead to catastrophic application failure (termination of an application run with a CPU failure), silent application errors (such as network data corruption), or application hangs (such as when network interface card (NIC) malfunction), all wasting valuable computer time. For certain classes of computer systems, dealing with these failures is a requirement to provide a simulation environment reliable enough to meet end-user needs. Also, the more automated these solutions are, requiring minimal or no end-user intervention, the more likely they are to be used to achieve the required application stability. Dealing with failure, or fault tolerance, while minimizing application performance degradation, is an active research area, with no consensus as to what are optimal solution strategies, or even what failures need to be considered. Errors include items such as transient data transmission errors (dropped or corrupt packets), transient and permanent network failures (NIC), and process failure, to list a few. The current MPI standard addresses a limited number of failure scenarios, with application termination being the default response to failure. While the standard provide a mechanism for users to override this default response, it does not define error codes that provide information on system level failures - hardware or software. None-the-less, these need to be addressed to provide end-users with systems that meet their computing needs. Building on experience gained in the LA-MPI, FT-MPI, and LAM/MPI projects, the Open MPI collaboration has implemented, and is continuing to implement optional solutions that deal with a number of failure scenarios, to decrease the application mean-time-to-failure rate, to acceptable rates. The types of errors currently being dealt with include transient network data transmission errors, transient and permanent NIC failures, and process failure. The talk will discuss fault detection, fault recovery methods, and the degree to which applications need to be modified to benefit from these, if any. In addition, the performance impact of these solutions on several applications will be discussed.

B. Mohr et al. (Eds.): PVM/MPI 2006, LNCS 4192, p. 2, 2006.