Adapt or Become Extinct! The Case for a Unified Framework for Deployment-Time Optimization

Adapt or Become Extinct! The Case for a Unified Framework for Deployment-Time Optimization

The High-Performance Computing ecosystem consists of a large variety of execution platforms that demonstrate a wide diversity in hardware characteristics such as CPU architecture, memory organization, interconnection network, accelerators, etc. This environment also presents a number of hard boundaries (walls) for applications which limit software development (parallel programming wall), performance (memory wall, communication wall) and viability (power wall). The only way to survive in such a demanding environment is by adaptation. In this paper we discuss how dynamic information collected during the execution of an application can be utilized to adapt the execution context and may lead to performance gains beyond those provided by static information and compile-time adaptation. We consider specialization based on dynamic information like user input, architectural characteristics such as the memory hierarchy organization, and the execution profile of the application as obtained from the execution platform's performance monitoring units. One of the challenges of future execution platforms is to allow the seamless integration of these various kinds of information with information obtained from static analysis (either during ahead-of-time or just-in-time) compilation. We extend the notion of information-driven adaptation and outline the architecture of an infrastructure designed to enable information flow and adaptation throughout the life-cycle of an application.