Energy Minimization and Data Movements: Models and Algorithms

Résumé du projet de recherche (Langue 1)

Energy has always been a concern in large scale distributed platforms. Nowadays, it becomes even more critical due to the transition to the next generation extreme scale High Performance Computing platforms (HPC) and the convergence with the notions of Cloud, Big Data and the Internet of Things. An increase of the computing performance of HPC platforms by a factor of 100 is targeted while staying at the current order of magnitude in terms of energy usage. This clearly shows that reaching this target needs a revolution in the way of handling resource management problems. Reducing the energy consumption to obtain radically more Flop/s per watt than today’s systems is the major challenge. Large-scale computing is shifting away from the traditional compute-centric models to a much more data-centric one. This transition is driven by the evolving nature of large-scale distributed computing, no longer dominated by pure computations but by the production of large volumes of data and an increased cost of moving data to the locations where computations are performed. The tasks inside jobs are fragmented between cores and induce heavy communications. There is a huge possible gain in energy here. We propose to study specific allocation policies according to topological constraints of the platform (contiguity of the processors executing a job, clustering jobs into cores/nodes belonging to the same hierarchy, etc). Our goal is to design new allocation methods taking into account such constraints without worsening the performance and leading to a reduced energy consumption.