Chapter 1
Hardware Trends and Implications for Programming Models

Gabriele Jost
The University of Texas at Austin, USA

Alice E. Koniges
Lawrence Berkeley National Laboratory, USA

ABSTRACT

The upcoming years bring new challenges in high-performance computing (HPC) technology. Fundamental changes in the building blocks of HPC hardware are forcing corresponding changes in programming models to effectively use these new architectures. The changes in store for HPC will rival the vector to massively parallel transition that scientific and engineering codes and methodologies endured several years ago. We describe some of the upcoming trends in hardware designs, and suggest ways in which software and programming models will advance accordingly.

BACKGROUND

Exascale computation (i.e., at a rate exceeding 1018 operations per second) by 2018 has been identified as a challenging but attainable for the future of scientific and engineering computation, that will lead to numerous advances in fundamental science. Quoting from "Report on Exascale Computing," (ASCAC (2010)), 'Going to the exascale' will mean a radical change in computing architecture – basically, vastly increasing the levels of parallelism to the point of millions of processors working in tandem – which will force radical changes in how hardware is designed (at minimum, driven by economic limitations on power consumption), in how we go about solving problems (e.g., the application codes), and in how we marry application codes to the underlying hardware (e.g., the compilers, I/O, middleware, and related software tools). On the brink of the

DOI: 10.4018/978-1-61350-116-0.ch001

exascale era, we are already seeing new hardware trends and corresponding advances in programming models. In this chapter we describe some of the recent trends and the programming models that are evolving to fit the new hardware.

HARDWARE TRENDS

The Multi- and Many-Core Era

A new era in computer architecture is emerging as we undergo a paradigm shift to a combination of multi/many core and heterogeneous architectures. This revolution stems from the need to overcome the ultimate end to the continuous run of increasing clock frequency fueled by the Moore's law trend of exponentially increasing number of transistors on a chip. Heat dissipation, physics of scale, and other issues such as the "memory wall" are seriously limiting the increasing gains that were afforded to high performance computing (HPC) systems over the last decades. Multicore architectures (8-12 or more cores) are common, and the future will include many more cores. Typically, a system is referred to as "many-core" if there are more than 32 cores in a shared memory environment often on a single chip. According to the International Exascale Software Project (IESP) roadmap (see Dongarra(2011)) next generation highest end computing systems available in the next decade are likely to have 100 million to 1 billion total cores configured possibly in a multithreaded node architecture consisting of hundreds of cores (many-core) per die and a multithreaded fine-grained concurrency of 10- to 100-way concurrency per core. Hybrid or heterogeneous designs also offer promise in a variety of ways. These include entire machines built on a combination of heterogeneous processors such as Roadrunner at Los Alamos National Laboratory, which combines AMD Opteron and Cell processors (IBM. The Cell processor, originally designed for gaming applications, combines eight SPE's or synergistic processing elements that are useful for highly parallel vector type operations in a package with a controlling Power Processing Element (PPE) that is a conventional microprocessor core. General Purpose GPUs (graphical processor units) are also becoming more evident in high performance computing. These chips, originally designed also for gaming and graphics applications, are becoming more versatile, and can be combined together into systems offering very large flop counts due to the fact that most of their transistors are dedicated to computation. Other designs may use field-programmable gate arrays (FPGA's) either as accelerator co-processors or as building blocks for an entire system. These FPGA's can be tailored by the user or architect to perform well on specific applications after manufacturing, hence the term "programmable".

Multicore processors are not limited to homogenous cores, and many designs include a combination of different fat or thin cores combined together. In this case, a fat core would have possibly more functionality, larger caches, and require more resources. In the future HPC arena, new systems are predicted to contain millions of cores, yet the design and ultimate best composition of these systems is unknown. Studies that predict and help optimize such designs are crucial.

Many experts currently agree that the driving factor in determining next generation architectures is energy consumption. The biggest energy cost is in data movement, especially moving data on and off chip. This can place very strong constraints on memory and interconnect bandwidth. Given these challenges, we might expect a 100-fold increase in parallelism on-chip, and a 10-fold increase in parallelism off chip, without any increase in the clock speed. This is in sharp contrast to the previous era of HPC development, where computers kept getting faster and faster without substantial intervention from software engineering or programming techniques. Starting a couple of years ago, this trend came to a halt, since while the packing density continues to increase (for about 10 years