

UNCONVENTIONAL ARCHITECTURES for High-Throughput Sciences

As science progresses, the need intensifies for scientific discoveries that will help solve our nation's most pressing challenges, such as infrastructure protection, global climate change, and alternative energy solutions. Thanks to advancements over the last decade in high-throughput technologies for scientific instrumentation and increased resolution in scientific simulations, the discovery process has steadily accelerated. However, these innovative technologies have also resulted in unprecedented volumes of scientific data and increases in the rate at which data are being produced.

Science is a Data-Driven Process

Science laboratories and sophisticated simulations are producing data of increasing volumes and complexities, and this poses significant challenges to current data infrastructures as terabytes to petabytes of data must be processed and analyzed. Traditional computing platforms, originally designed to support model-driven applications, are unable to meet the demands of the data-intensive scientific applications.

Pacific Northwest National Laboratory (PNNL) research goes beyond “traditional supercomputing” applications to address emerging problems that need scalable, real-time solutions. The results are new, unconventional architectures for data-intensive applications specifically designed to process the deluge of scientific data.

Too Much, Too Fast

The current challenges stem from differences in the nature of high-throughput science applications versus traditional applications that are based on mathematical models. Unlike traditional applications for high-performance computers, the two main activities that data-intensive applications are designed to address are: massive data mining, or very large scale data analysis; and data streaming—processing large amounts of data streaming from a data source in real time. These activities can quickly overwhelm the capabilities of traditional scientific computing platforms and

processing techniques, which were designed for applications that have well-defined, structured, and localized data accesses. Another issue with current applications is that the way data are explored is irregular, and irregular access to data is bad for cache-based architectures and clusters.

In the traditional architectures, the increasing performance differential between the capabilities of memory subsystems and microprocessors has caused a large class of applications to become memory-bound, that is, their performance is determined mainly by the speed at which the memory subsystem can deliver data words to the microprocessor. Several hardware and software mechanisms have been proposed over the years to increase the performance of such applications by reducing the exposed stall times seen by the microprocessor. Most mainstream microprocessors utilize a cache hierarchy, whereby small sections of high-speed memory hold data which have been recently fetched from the main memory. Cache mechanisms are highly effective for applications that exhibit good temporal and spatial locality. Unfortunately, many data-intensive applications that have irregular data access patterns do not belong to that category.

New Approaches Through New Architectures

The use of hybrid high-performance computing (HPC) systems for processing and analyzing streaming data is a novel application that can pro-

Pacific Northwest National Laboratory (PNNL) research goes beyond “traditional supercomputing” applications to address emerging problems that need scalable, real-time solutions.

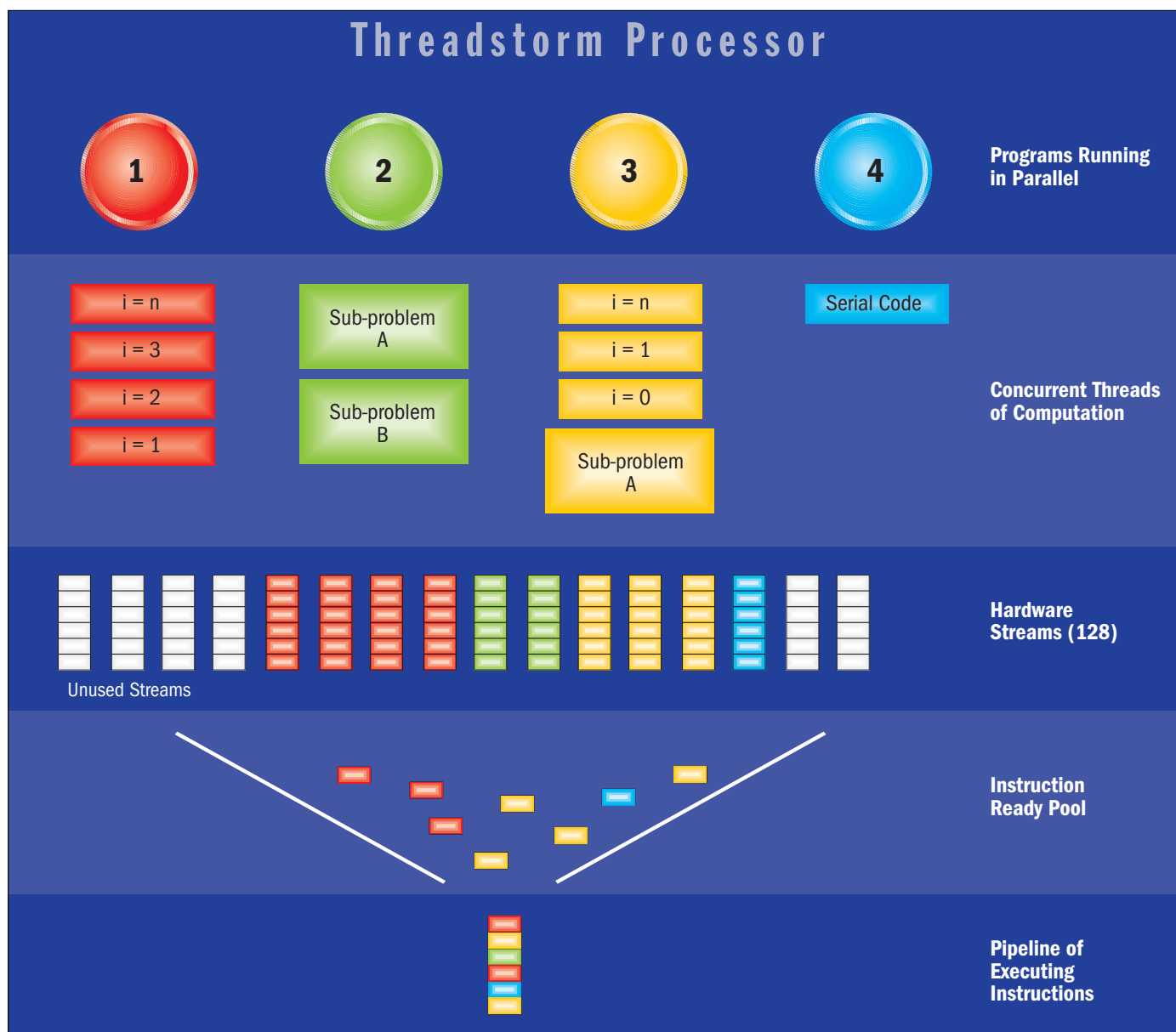


Figure 1. A conceptual illustration of multi-threaded architectures using an example of the Threadstorm processor in the upcoming Cray XMT system. Four example programs are represented here with different forms of parallelism: (1) loop-level parallelism, (2) coarse-grain parallelism represented by two tasks (A and B), (3) a combination of task-level and loop-level parallelism, and (4) a purely sequential code. The parallel code units are translated by the compiler to a pool of threads which then are mapped to the Threadstorm hardware streams. Each of the 128 hardware streams has its own instruction counter, register set, stream status word, and target and trap registers. The processor executes an instruction from a different hardware stream every clock period. Only those streams which have data are placed in the instruction ready pool and can execute while the rest are waiting for the data to be fetched from the memory. As a result, the machine can efficiently execute instructions from all the threads corresponding to the four application codes, while hiding the latency of the memory accesses.

vide the potential for much improved experimental feedback. Dr. Daniel Chavarria, a PNNL scientist, believes the advantages of hybrid HPC architectures over traditional HPC systems can address many of the issues involved with streaming applications. Rather than providing a replacement for traditional architectures, he's using Field Programmable Gate Arrays (FPGAs) to complement them, in order to realistically simulate real-time, online analysis of streaming data coming directly from a mass spectrometer.

With regular microprocessors specific calculations can be easily programmed, whereas for FPGAs the programmability is more complicated. This is an inherent feature of FPGAs because one must describe computation as set of parallel processes. Unlike software processes, they operate at a much lower level. For example, when memory is loaded from a bank on an FPGA system, the system is exposed to memory latency during a number a cycles while data arrive. Programming must account for that in advance.

S. NEELI, PNNL, WITH PERMISSION OF ORNL

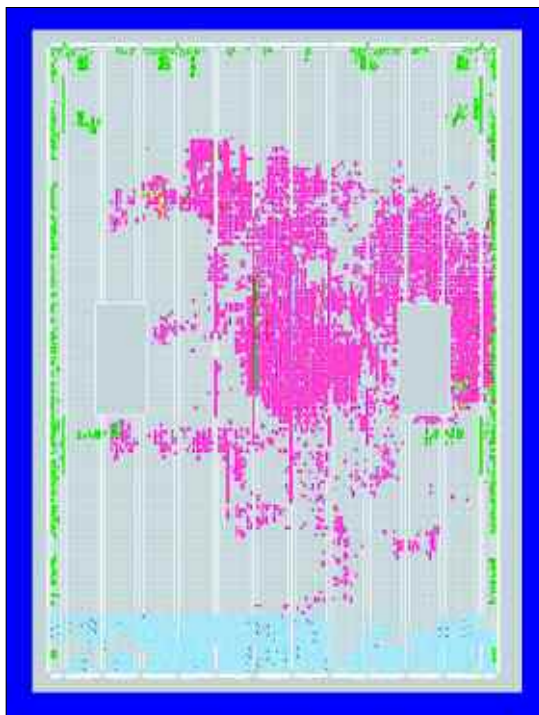


Figure 2. The preliminary VirtexII-Pro FPGA floorplan for a design that focuses on streaming processing of mass spectrometry data. An FPGA floorplan is an image of the occupied cells in the FPGA with colors to indicate cells used for the application (pink) and cells used for system interfaces (green and light blue). The latter are physically on the borders to be proximal to the FPGA's physical connection to the rest of the system.

Once the programming is complete, the execution of computation will be fully deterministic (in the case of streaming data at a fixed rate) and the advantage of FPGAs can be realized. If the rate is variable, the programming must take into account the highest rate.

Dr. Chavarria's working prototype is based on general-purpose microprocessors and FPGAs. The FPGA components themselves are, by default, real-time devices. When loaded with appropriate configurations, FPGA components will operate with guaranteed performance bounds, so they are ideal for handling real-time streaming tasks.

Dr. Chavarria saw a prime application in proteomics research at PNNL, where scientists were capturing mass spectrometry data at the rate of tens of megabytes per second. Scientists had to choose between losing data, because the data streamed in faster than they could be stored, or reducing the rate at which the data streamed, which prevented the use of more sophisticated instruments (sidebar "Boosting Proteomics Platforms to the Next Level," p50).

FPGAs are a huge help to high-throughput scientific instruments because they enable stream-

ing processing for instruments that have much higher throughput rates (figure 1, p47). Furthermore, FPGAs can attach directly to the instrument. That means data reduction can be done as the data are being collected, potentially reducing the amount of data that must be stored. FPGA-based processing can lead to scientists' ability to incorporate real-time experimental feedback.

Dr. Chavarria's research is contributing to a new paradigm for high-throughput experimental instruments by enabling faster, more efficient processing of a richer data reservoir. Combined with Dr. Dick Smith's new proteomics capability (sidebar "Boosting Proteomics Platforms to the Next Level," p50), these innovative approaches could one day lead to near-real-time diagnostics of organisms within communities.

Multi-threaded Architectures

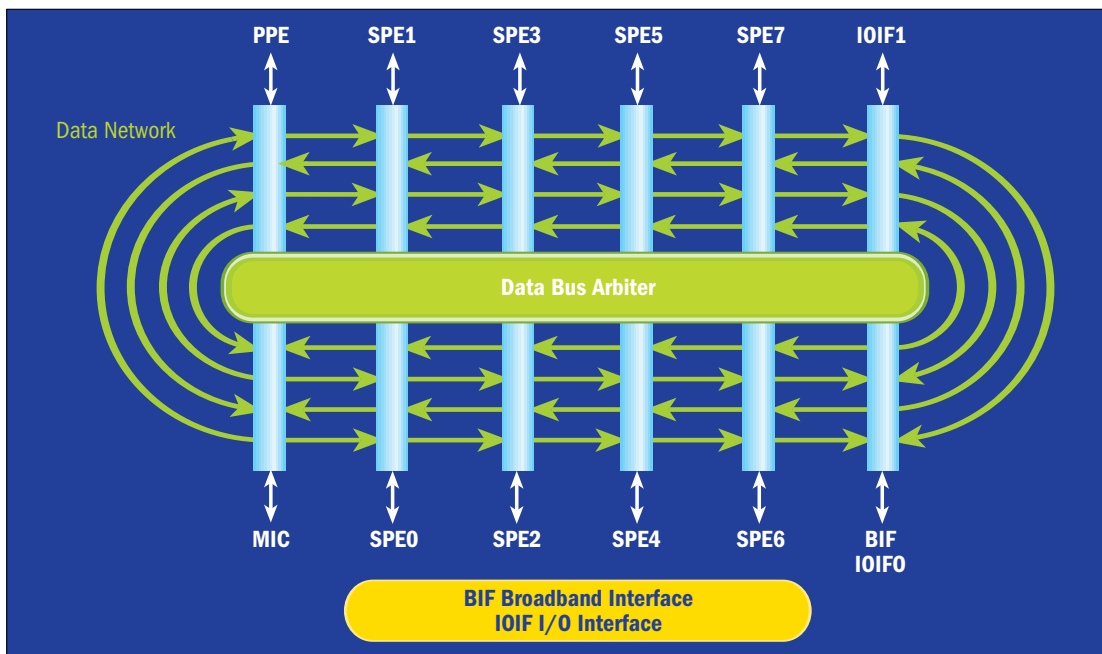
In recent years, there has been a renewed interest in shared memory architectures with hardware multi-threading. Multiple threads can use hardware resources more efficiently by pre-empting threads that are waiting for data and running threads that have data available. Systems such as the Cray MTA-2, Eldorado (XMT), or Sun Niagara provide a natural latency hiding capability and thus can efficiently execute applications with irregular memory references.

Data-intensive applications must derive insightful knowledge from massive datasets, where the objectives are not known beforehand (as is the case in many model-driven applications). Analyzing massive amounts of data loaded from secondary storage into the main memory is compute-intensive, and the increasing disparity in performance between the main memory and the microprocessor has caused many applications to become "memory-bound," where performance is determined mainly by the speed at which the memory can deliver information to the microprocessor. Indeed, in many large-scale applications, the processor is idle 80% to 90% of the time while waiting for data from memory.

Enter multi-threaded processors. Such processors tolerate latencies by switching contexts to computation threads with available work (figure 2), making them ideal for applications with high levels of irregular memory access, such as those found in data-intensive applications.

At PNNL, Dr. Jarek Nieplocha is examining two PNNL-developed applications with two very different multithreading approaches—Cray's MTA-2 and Sun's Niagara. The first application is a power system state estimator that computes the state of electrical grid systems, and the second is an application in statistical-based anomaly detection for categorical data that can be used for Inter-

Dr. Chavarria's research is contributing to a new paradigm for high-throughput experimental instruments by enabling faster, more efficient processing of a richer data reservoir.



By improving the execution speed of irregular data-intensive applications, multi-threaded platforms have the potential to realize significantly higher application performance—a boon for big datasets that must be mined to detect the unknown.

ILLUSTRATION: A. TOREY SOURCE: F. PETRINI, PNNL

Figure 3. The element interconnect bus (EIB), the heart of the Cell processor's communication architecture, which enables communication among the power processor element (PPE), the synergistic processing elements (SPE), main system memory, and external I/O. The EIB has separate communication paths for commands (requests to transfer data to or from another element on the bus) and data.

net traffic analysis. Both applications have irregular memory access patterns.

Dr. Nieplocha and his colleagues showed that it is possible to obtain significant speedups on multi-threaded platforms due to their ability to tolerate the unpredictable access patterns.

The group also examined the strengths and weaknesses of the architectures in running the two applications. They found that although both architectures use thread level parallelism (TLP), Niagara's performance drops as the number of threads increases. They concluded that Niagara's architecture needs the backing of large memory bandwidth for those applications that make the best use of TLP.

MTA's processor has the advantage of simplifying the programming model, opening more opportunities for the compiler to extract TLP out of sequentially coded programs.

By improving the execution speed of irregular data-intensive applications, multi-threaded platforms have the potential to realize significantly higher application performance—a boon for big datasets that must be mined to detect the unknown.

IBM Cell

In another unconventional approach, PNNL scientist Dr. Fabrizio Petrini and his colleagues, Dr. Daniele Scarpazza and Dr. Oreste Villa, are implementing a string-scanning algorithm on the IBM Cell Broadband Engine (Cell BE) processor

(figure 3) to achieve more efficient cyber security on a large network.

String matching is an essential component in network security, but as network links become faster and faster, string matching is becoming more difficult to perform in real-time. Traditional processors are not keeping up with the performance demands, and often specialized hardware is not able to compete with commodity hardware in terms of cost effectiveness, reusability and ease of programming.

The Cell BE processor is the first implementation of the Cell Broadband Engine Architecture (CBEA), developed jointly by Sony, Toshiba, and IBM. It includes one POWER processing element (PPE) and eight synergistic processing elements (SPEs). The CBEA is designed to be well-suited for a wide variety of programming models, and allows for partitioning of work between the PPE and the eight SPEs. Dr. Petrini's work has shown that Cell BE can outperform other state-of-the-art processors by approximately an order of magnitude.

Dr. Petrini believes that the Cell is an ideal candidate to tackle modern security needs. He found that two processing elements alone provide sufficient computational power to filter a network link with bit rates in excess of ten gigabits per second.

Dr. Petrini's work has also shown that multiple Cell processors can be flexibly combined to achieve higher performance or to use larger dictionaries. When applied in this domain, the Cell BE is a fast and flexible architecture, in addition to

Boosting Proteomics Platforms to the Next Level

Dr. Chavarria is applying his FPGA research to that of Battelle Fellow Dr. Dick Smith (figure 4), a world-renowned innovator of mass spectrometry technologies for proteomics research. Dr. Chavarria's work focuses on implementing a signal processing application that can handle the deconvolution of data streaming from an advanced ion mobility mass spectrometer (figure 4). The ion mobility instrument convolves the data itself by pulsing ions into its drift tube with a certain preset binary pattern. The digitally sampled output data must then be deconvolved in order to reconstruct the mass spectra. Preliminary results on the hybrid HPC experimental setup indicate that data can be accepted at over one gigabyte per second from a simulated source on the main CPU to its destination on the FPGA. In the actual ion mobility mass spectrometer, data will be captured directly by an analog-to-digital converter attached to a specialized FPGA processing board using a configuration similar to the setup being tested in the hybrid HPC system.



Figure 4. New proteomics platforms can realize even greater potential with the help of FPGAs. Dr. Dick Smith, PNNL, stands with an instrument that incorporates ion-mobility separation with time-of-flight mass spectrometry. The instrument will increase the throughput of proteomics measurements by up to about 100-fold.

Dr. Smith is excited about how Dr. Chavarria's work can help take it to a new level. Says Dr. Smith, "The next-generation proteomics platforms are demanding new computational paradigms to manage the

massive data volumes as well as extract meaning from raw data. The successful development of new capabilities will be enabled by data-intensive computing developments in our laboratory."

being readily available and relatively inexpensive. And since string matching is an essential component of search engine development, computational biology and other science domains, its potential application and benefits could be extensive.

A Blueprint for Progress

More technologies are on the horizon. Many industry vendors are working on many-core processors (for example, Intel's 80-core arch), and Cray has a long-term development effort with a focus on adaptive supercomputing. The Defense Advanced Research Projects Agency (DARPA) program on polymorphic computing is also bound to result in some new designs.

Unless novel disruptive technologies manage to reduce the processor-memory latency gap, large multi-core architectures will invariably exploit multi-threaded latency hiding mechanisms as well. An example is IBM's Cyclops64 with 80 processors on a chip. Each processor has two thread units that share a floating point unit. The chip also hosts SRAM memory that is interconnected to the processors with a fully synchronous crossbar in dancehall fashion. Intel's terascale research investigates chip technologies surpassing 100M transistors: the terascale processor hosts 80 small core tiles that are easily designed, replicated or specialized according to functionality.

The power-aware design allows tiles to be enabled or disabled during program execution to improve power efficiency. A 2D mesh network provides the required interconnectivity between tiles.

Industry vendors know they cannot form the future working in silos. At the International Parallel and Distributed Processing Symposium (IPDPS'07), Luiz DeRose (Cray) and Dr. Nieplocha together organized the first Workshop on Multi-Threaded Architectures and Applications to provide a forum for discussing how scientific applications can exploit multi-threaded architectures and which programming methodologies and software tools are appropriate for this class of systems.

Overall, there is strong support growing in the commercial market for unconventional architectures, evidenced by IBM's confidence in their BlueGene and Cell machines and Cray's vision for all new generations of supercomputers to have heterogeneous architectures ("Supercomputing and Industry," *SciDAC Review*, Spring 2007, p8) As PNNL researchers pave the way with these industry leaders, the resulting technologies will undoubtedly enable faster, more efficient scientific research. ●

Contributors: Written by PNNL scientists, Dr. Jarek Nieplocha, Dr. Andres Marquez, Dr. Fabrizio Petrini, and Dr. Daniel Chavarria