LA-UR-09-02403

Approved for public release;
distribution is unlimited.

Title: The PetaFlops Router: Harnessing FPGAs and Accelerators for High Performance Computing

Author(s):
Zachary K. Baker, CCS-1
Justin L. Tripp, CCS-1
Gerd Kunde, P-25
Rajan Gupta, T-2
Matthew Stettler, ISR-3
Mark Dunham, ISR-DO

Intended for:
2009 Symposium on Application Accelerators in High Performance Computing (SAAHPC'09)
July 27, 2009
Urbana, IL

Los Alamos
NATIONAL LABORATORY
EST. 1943

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (7/06)

The PetaFlops Router: Harnessing FPGAs and Accelerators for High Performance Computing

Z.K. Baker, Justin L. Tripp, Gerd Kunde, Rajan Gupta, others
Los Alamos National Laboratory
Los Alamos, NM 87545
Email: zbaker@lanl.gov

Abstract

This paper introduces the PetaFlops Router, a new approach to computing wherein network architecture and compute decisions are customized for a particular application. New Field-Programmable Gate Array (FPGA) and router technologies including multi-gigabit transceivers and application specific blocks can provide vastly improved performance. The PetaFlops Router provides greatly improved data transfer rates, computational efficiency, and programmability compared with applica-tion-specific hardware. The system will be validated and its performance measured for Threat Re-duction and High Energy Physics applications.

1 Introduction

Achieving high performance on modern computers is difficult without detailed architectural knowledge. The goals of computer designers and end-user scientists are often at odds. A computer designer is happy to tie together a variety of high-end processors, each with its own strengths. Using these systems, without an efficient programming paradigm, requires intimate knowledge of the components, which makes it difficult for anyone other than an expert to achieve maximum performance. The end user scientist would prefer to not have to learn these system details due to time constraints and the cost of learning through mistakes when one if forced to consider the ramifications of every programming decision. A scientist's time is better spent doing science, not struggling with low-level programming details. The PetaFlops Router provides the performance of highly parallel, custom hardware, while allowing the user to think in terms of easy-to-use, high-level concepts.

The petaflops router is based on Field Programmable Gate Arrays, a reconfigurable integrated circuit technology. A designer can provide the performance expected of custom-hardware, but can also reconfigure the device to be optimized for many different classes of applications. Not only can the system be customized through reconfiguration of the intra-chip logic, but also by rerouting the flow of information through the inter-chip network. In short, the entire system can be adapted to match a user's application.

Our approach builds up libraries of components for a particular domain of computation, and then efficiently connects them through the on-node switch structure. A user interested in building a signal-processing application might want a collection of FFT blocks, CORDIC cores (for trigonometric functions and exponentials), and matrix and vector units. An image-processing user might want some of the same cores, plus windowing operators and other support for 2-D data. These cores would be available as firmware libraries for the FPGA, built to natively interface with the system. Memory load and stores, as well as data movement between devices and co-processors implementing some other advanced functionality are all controlled within the same framework.

This framework is built on a language that would be familiar to anyone with a background in high-level parallel programming. However, instead of thinking in terms of explicit scheduling and element-by-element movement and computation, computation is scheduled when its data is ready, and data is moved as appropriate for its scheduled computation. In this way, the user can program operations in a way that makes sense for a scientist, and the hardware can handle the underlying issues of timing and operator scheduling. Unlike classic notions of a vector machine, the system can have dozens of functions operating on a stream at one time, with those functions being highly complex (like an FFT or layered matrix multiply). This allows vast parallelism whereas a traditional micro-processor can only execute few multiplies or adds in a given cycle.

The system is not limited to FPGA-based computation. The PetaFlops Router is capable of routing not just to other FPGAs, but PCI Express connected accelerators. The multi-gigabit transceivers on an FPGA can operate in a variety of

modes to connect to a variety of serial standards. Serial has become the approach of choice for high speed point-to-point connections because it is simpler to synchronize a single data line to a clock than dozens of data lines to each other as well as the clock. Only having to deal with a few signal lines make design and reliability much easier. FPGAs can include dozens of these serial transceivers, allowing high aggregate bandwidth. This is useful as modern graphics processing accelerators and coprocessors are commonly connected to their hosts via PCI Express. PCI Express is a relatively new standard wherein peripherals are connected via serial lines. This is useful to the PetaFlops Router largely because it provides a standard, high speed interface that does not require a large number of pins.

Co-processors are needed because of the need for certain operators that turn out to be very expensive to implement on FPGAs. FPGAs can cheaply implement boolean, byte, and fixed-point operators, making them well-suited for signal processing, encryption, and other highly parallel kernels such as genome sequencing. However, floating point, and in particular double-precision floating point, are resource intensive and reduce the amount of parallelism achievable simply because the large operators do not pack well into limited hardware area. While FPGAs have moved beyond 'glue logic', they still retain the volume of input/output pins required for providing high bandwidth connections to memory interfaces; these pins can be used to connect an FPGA to a floating-point accelerator or any other custom ASIC to provide application-specific co-processing.

The hardware system discussed here is quite versatile. Expert FPGA designers are used in the first phase, but are not the development bottleneck. FPGAs are not the focus of the work, but an enabling technology that allows the efficient connection of the disparate pieces. In the past, 'glue logic' meant that FPGAs served in lieu of a few gates and provided the functionality of a few small-scaled integration parts. Here, FPGAs serve as a "Smart Glue", not only connecting pieces together, but making decisions about how the computation will move forward through a hybrid data-flow architecture.

The idea of organizing computation in a data-flow style architecture is not new. Likewise, organizing heterogeneous computation through the use of data-flow has been explored before. However, the use of FPGAs to organize and control the movement of data, as well as providing certain custom computational blocks, is new as it can provide a mechanism to harness the power of FPGAs without requiring expert FPGA designers for the lifetime of a program. The FPGA can provide lower latency for determining the next phase of computation, as well as having the ability to connect natively to high-speed serial data links. Several programs at LANL are already exploring multigigabit serial links for FPGA and CPU interconnect, including the Joint Architectural Standard (JAS) project, the AngelFire SWAP (Size, Weight, and Power) Reduction effort, and the CERN LHC trigger mechanism.

Thus far we have discussed the base hardware, namely, FPGAs and attached co-processors. These pieces are connected together via a highly capable network switch that operates on high-speed serial interconnect. The basis of the network is a crossbar switch, a commodity device but revolutionary when applied at a node level. That is, while crossbars have been used historically to connect together groups of nodes, our placement of a crossbar on every node allows a significant extension of capability. The power of the crossbar is that it allows provides a set of n2 connections to n input ports and n output ports. Thus, any input can be connected to any of the outputs. Not only can it provide an independent path for each of the inputs to each of the outputs, it can also efficiently perform one-to-all communication by connecting an input to all of the outputs. This provides high levels of flexibility and connectivity. A single crossbar can connect up to 32 nodes, which is useful but not scalable. By putting a crossbar on every processing node, essentially any network topography desired can be built from a collection of nodes. The crossbars can be modified at runtime to act as network switches, or can be setup at application setup to build a particular network topology.

An important aspect of the ability of the system to adapt the system to the particular application. The PetaFlops Router takes application topology into account, adapting algorithm dynamics and knowledge of moment-to-moment system state to optimize performance. This allows the system to be designed for a particular class of applications, either by adjusting the crossbar connections or by adjustments to physical connections. The flexibility provided by the crossbars makes the system ideal for a wide range of scales, from a small embedded system to large HPC installations. Each node can be connected to a host, to other nodes directly, or through a backplane to additional crossbars and other racks.

2 SmartGlue

A key advance in our architecture is that SoPCs serve as "Smart Glue," not only connecting pieces together but also making decisions about how the computation will move forward through a hybrid data-flow architecture. The Smart Glue is far more capable that just moving data between computational units, although providing scheduling and routing decisions will be a large part of both the development and the contribution of the system. The Smart Glue has the potential to be both self-optimizing and error-tolerant. Link failures or simple congestion (not necessarily network, but computational or memory bandwidth) can make the flexibility of an FPGAcontrolled routing fabric a

"powerful feature." Having multiple routes to get data in and out transparently to the user is a very powerful feature.

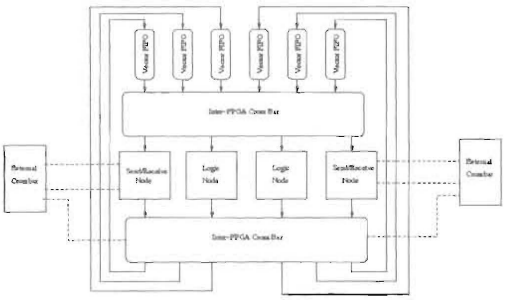

Figure 1. General internal crossbar architecture to connect functional units

The SmartGlue system is based around input/output crossbars and a series of FIFO registers (Figure 1. The First-In First-Out (FIFO) registers are essentially a vector register with a read and write port, such that the system can simultaneously read and write to the same register. The FIFO handles itself to prevent read-before-write hazards that could introduce corrupted data into the system. The crossbar is a network component that allows connection from any one of the inputs to any number of outputs. This allows data movement from computational components to vector registers (also allowing replication of registers for computations where output products are used multiple times).

This foundation connects together the components that actually do computation, as well as link the system to other coprocessors. The system can work like a classic vector processor in some cases, namely, when the size of the data vectors is smaller than the size of the FIFO register. However, breaking long vectors up into blocks is not always necessary; when there are sufficient computational units and data path to move the data between units, it is possible to start streaming data into the next computation unit before it has fully populated the transfer register. This is an exciting capability, because it allows parallelism to be achieved without the programmer having to make explicit schedules.

Each chip has potentially hundreds of computation units and several hundred 512x32 data FIFOs. These handle computation and data movement within one FPGA. Between FPGAs, the system becomes more complex as data bandwidth outside of the chip is necessarily less than on-chip band-width. Communication between FPGAs is handled by external crossbars that are operationally similar to the on-chip crossbars. The external crossbars are a major component of the PetaFlops Router hardware effort.

The control of the crossbars, and thus the control of the data movement and computational units, is achieved through the use of a novel implementation of Tomosulo's algorithm distributed between multiple chips. We are currently in the process of developing control heuristics to manage data movement and computation between chips, as well as allocating resources across the system.

3 High Energy Physics

One of the key applications of interest is simulations of Lattice QCD, a fundamental theory of nature and an area in which LANL has played a leadership role over the last 20 years. Quantum Chromodynamics (QCD) is a mathematical formulation of the interactions between quarks and gluons which bind together to make up neutrons and protons. The goal is to be able to predict the properties of neutrons, protons and other nuclear matter from first principal calculations. Using large-scale simulations of this theory, we today know how to calculate a number of properties of QCD (masses, decay properties and the equation of state).

The challenges are that current experimental determinations of many of these quantities are more precise than can be calculated from theory with current computing power, and at the same time simulations of additional phenomena are needed to interpret the results of other experiments. Precision test of the theory and search for new physics in either approach requires reduction in statistical and systematic errors in the theoretical calculations by a factor of 5.

Improvements in precision have come from both improved formulations of the theory and from access to more powerful state-of-the-art supercomputers. It became clear very early to the practitioners that the scale at which precise results, with control over both statistical and systematic errors will be achieved, is the petaflop. As a result the community has worked with, and helped develop, almost all parallel supercomputers.

The teams ultimate physics goal is to study the quark gluon plasma to be produced in the planned heavy ion experiments at the highest energies at the Large Hadron Collider at CERN. In particular, it is crucial to obtain the Equation of State (EoS) of the Quark Gluon Plasma over the temperature range (150-700 MeV) and the transition temperature from hadronic matter to quark gluon plasma to provide crucial guidance in the interpretation of experimental measurements in lead-lead collision at CERN/LHC.

4. Conclusion

In this paper, we present the PetaFlops Router, a new approach to computing wherein network architecture and compute decisions are customized for a particular application. By integrating Field-Programmable Gate Array (FPGA) and router technologies including multi-gigabit transceivers and application specific blocks, the system provides vastly improved performance.