Title: A New Architecture Revolution for Supercomputing

Author(s): CCS: Z. Baker <zbaker@lanl.gov>, A. McPherson, J. Tripp;
          ISR: P. Graham <grahamp@lanl.gov>, M. Dunham, M.
          Stettler
          P: G. Kunde <g.j.kunde@lanl.gov>, A. Klein;
          T: R. Gupta <rg@lanl.gov>;
          HPC: Jeff Inman <jti@lanl.gov>

Intended for: Engineering Grand Challenge Proposal
Technical Information Release Continuation Page

“A New Architecture Revolution for Supercomputing”

For: Engineering Grand Challenge Proposal

Continuation of Authors:

<table>
<thead>
<tr>
<th>Author</th>
<th>Name</th>
<th>Z#</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dunham</td>
<td>Mark E.</td>
<td>101035</td>
<td>ISR-DO</td>
</tr>
<tr>
<td>Stettler</td>
<td>Matthew W.</td>
<td>107603</td>
<td>ISR-3</td>
</tr>
<tr>
<td>Kunde</td>
<td>Gerd J.</td>
<td>146421</td>
<td>P-25</td>
</tr>
<tr>
<td>Klein</td>
<td>Andreas</td>
<td>187935</td>
<td>P-23</td>
</tr>
<tr>
<td>Gupta</td>
<td>Rajan</td>
<td>099533</td>
<td>T-08</td>
</tr>
<tr>
<td>Inman</td>
<td>Jeffrey T.</td>
<td>144219</td>
<td>HPC-1</td>
</tr>
</tbody>
</table>
A New Architecture Revolution for Supercomputing

CCS: Z. Baker <zbaker@lanl.gov>, A. McPherson, J. Tripp;
ISR: P. Graham <grahamp@lanl.gov>, M. Dunham, M. Stettler
P: G. Kunde <g.j.kunde@lanl.gov>, A. Kleini
T: R. Gupta <rg@lanl.gov>; HPC: Jeff Inman <jti@lanl.gov>

A. Grand Challenge Statement

This trans-LANL team questions the established catechism for High Performance Computing; that “knowledge of the machine” must be concealed from scientific users and programming staff. LANL’s Roadrunner HPC architecture demonstrates that more complex co-processor approaches produce large gains in computational efficiency, albeit at the expense of non-standard programming models. Similar gains in efficiency of up to 1000X over COTS microprocessors are achieved by LANL’s ISR Division, in Reconfigurable Computing systems for space, airborne, and ground applications. Our core tenet is that the next great leap in computing requires redesign of the compiler models extant since the dawn of computer science.

Further, none of these existing architectures directly address I/O binding---the principal efficiency limit in modern HPC. Here we propose to break the I/O binding by rebuilding the computational fabric within modern serial network architectures; a scheme we call PetaFlops Router. Such work is underway at LANL for the CERN/LHC/CMS detector trigger system, with aggregate input dataflow rate of 300 Gigabits/second, which is then processed in real time by a network of 1000 each Xilinx System-on-Chip (SoPC) devices. These devices use high-speed serial I/O as their principal communications pathways, as do an increasing number of digital signal processors (DSP), memory devices, and GPU processors. Linking them with modern crosspoint switch technology, we can create a true router architecture where all major HPC components are richly supplied with low-latency shared cache storage memory.

In addition to new router fabric as the data backbone, we also propose the new concept of using Xilinx SoPC as “SmartGlue” routing hubs, connecting standard GPU, accelerators, and other devices as co-processor elements to the router. In contrast to a microprocessor with one I/O bus, the Xilinx SoPC have over 1000 I/O connections which can be used to build adaptive high bandwidth connections to the co-processors. This same I/O capability enables a third key piece of the architecture: low latency and large shared memory caches. This combination of large, fast memories and high bandwidth through PetaFlops Router network enables a very efficient shared memory architecture. Therefore a key differentiator of this EGC proposal versus the existing IS&T thrust is our focus on revolutionary approaches to machine architecture, which are neither bounded nor limited by existing programming models.

As such, the programming model to be used for such a machine is a key part of our proposed research, since almost every aspect of the hardware is reconfigurable. In essence the entire supercomputer can now adapt itself to the problem at hand, increasing the programming dimensionality dramatically. This points to a need for a large effort in compiler level development of programming tools, which is already recognized in the IS&T Grand Challenge, but within a conventional server computer context. Although expert machine-level programmers will be needed at the inception of the PetaFlops Router architecture, our endpoint goal is to allow the scientist user to define the operations in a coarse-grain data-flow of composable blocks, but he will not explicitly define the timing and connections to those blocks. The classic MIT token architecture for data-flow computing can be leveraged. In our application, products of a computational block are tagged with a ‘token’ that determines what the next computation will be; tokens are indexed into a lookup table that details what other data inputs are required, what computation will be done, and where the data target will be. This allows the system to determine where and when computation occurs, rather than requiring explicit definition. We further
propose Domain Specific Languages (DSLs) to built applications on TeraOp system, including the attached accelerators. The pervasive use of DSLs can provide significant gains in the productivity and creativity of our scientists, portability of applications, and greatly enhanced application performance. Increasing productivity of scientists is the prime goal of this effort.

In summary, we propose a new solution to “knowledge of the machine” issues now being seen in all supercomputing. A key advance in our architecture is that SoPC 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. This Smart Glue also has the potential to be both self-optimizing and error-tolerant, as the research matures. Link failures or simple congestion (not just network jamming, but actual CPU or memory bandwidth) can be addressed in PetaFlops Router by having a redundant routing fabric. These multiple routes to get data in and out transparently to the user, are a key feature of PetaFlops Router, and are one of its most powerful features. If our work is successful, we forsee rapid uptake of the technology by both supercomputer mainframe developers and also the academic community.

B. Required Engineering Disciplines


C. Relation to Current Science and Engineering Technology

ADTR and LANL have identified military signal and image processing (DSP) as a key growth area. Development of such systems has been a core thrust in ADTR for over 15 years, while the more recent LANL development of adaptive trigger systems at CERN/LHC is recognized by P Division as the path forward for growth of experimental nuclear and high-energy physics. The TeraOp router provides multiple options for implementing applications to the system. For DSP we often use only the SoPC for computing, since they are excellent for bit through 1 to 32-bit integer operations. Other applications such as datamining, cybersecurity, encryption/decryption, and video/image compression all are excellent fits as well. However, by appending various accelerators to the base TeraOp architecture, we can build a system which leverages the best COTS floating point co-processors to advantage as well. By doing this, we can build upon the knowledge in heterogeneous architectures gained from Roadrunner, and extend 10X - 100X higher in computational capability.

One of the key applications of interest is simulations of Lattice QCD. Determining the EoS and other properties of the Quark-Gluon Plasma, requires non-perturbative methods. The most promising and reliable approach is first principal simulations of Lattice QCD, using a 4-D lattice with one temporal and 3 special coordinates, and also using a discretized version of QCD called Lattice QCD. Our first goals are, therefore, precise estimates of the transition temperature and the equation of state as a function of temperature to provide crucial guidance in the phenomenological interpretation of experimental measurements. This advance requires high statistics simulations of Lattice QCD on large lattices, enabled by going beyond the current state-of-the-art traditional computer architectures. There are a number of reasons Lattice QCD is an ideal test problem for new computer architectures: 1) The 4 dimensional space-time (1-time and 3-space) grid is homogeneous so the same operations are done at all space-time points. 2) Communications are local., so nearest neighbor communications in a 4-d torus (or 3-d with time stacked on 3-d space grid) configuration are sufficient. 3) The data needed for computation can be staged, so efficient DMA strategies for communication of data can be developed. 4) The compute to communication ratio is high, typically 10, and the pattern of communications is uniform. 5) Global operations, such as sum of a vector array over all sites, constitute a small fraction of the calculations. 6) Lattice QCD calculations run uninterrupted for months and highly optimized codes provide close to achievable efficiency. Thus Lattice QCD codes provide
very stringent diagnostics of all aspects of both the hardware and software as documented on all parallel computers at LANL since the CM-2, which was installed in 1989.

An HEP experimental example is handling the higher luminosity (factor of 10) of the planned upgrade of LHC to SLHC, where it is clear that a new hardware and software model has to be developed. Traditionally, for the level 1 (L1) trigger, custom built hardware is used, with fixed latency and dedicated algorithms tailored to the specific task; i.e. finding particles with a transverse energy above a certain threshold based on energy deposit in geometrical areas. With the higher luminosity, one has to push part of the more computing intensive tasks of the L2 system into the realm of the L1. Instead of reading out all the pixels from the central tracker, and forming tracks in the L2 stage, it will be required to form tracks already at the first stage, which can then be used in the L1 trigger decision. Such a transition however, will require that part of the processing system has floating point capability, which in turn requires either CPUs or GPUs in addition to FPGAs. Such a system also allows for the needed flexibility in reconfiguring the trigger system without rewiring, i.e. the use of cross point switches on a high speed bus. The reconfigurable microcode in the FPGA and the ability to perform tracking with floating point precision will be key to discovering new physics in the upcoming LHC and SLHC experiments.

**D. Metrics for Success**

1) Hardware metrics for success will be easily defined as 3 iterations of a prototype hardware architecture, with associated basic performance metric tests.

2) Software success will be determined by demonstration of basic automated data-flow connectivity through a small-scale system built on existing hardware platforms. Follow-on goals would include self-optimization and custom hardware as well as physics applications.

3) System metrics will be defined by the degree to which LANL can lead its sponsor communities to acceptance of this architecture, based on solid engineering results and architectural vision. The ultimate goal will be to help redirect the global HPC community to an evolution in how supercomputing platforms are designed.

**E. Relation to LANL Mission**

- HPC is a core laboratory mission and thrust area, supporting multiple programs.
- Embedded supercomputing is now a key growth area for ADTR programs.
- The LHC/CMS TeraOps Router trigger is a direct path for LANL to become a CMS partner.

**F. Relation to Other Grand Challenges & MARIE**

NSF EGC: Engineer the Tools of Scientific Discovery; Reverse Engineer the Brain Themes: Coupled Simulation & Experiment; Enabling & Exploiting heterogeneous computing Other LANL Grand Challenges; Beyond Standard Model at LHC; Ubiquitous Sensing systems & analysis/modeling; Information Science & Technology compilers/languages

**G. Business Case for Investment**

External sponsors have already shown willingness since 1992 to fund research in novel computing hardware for ubiquitous sensing. More recently we are receiving funding directly from CERN to develop the Super LHC CMS primary trigger system architecture. If we can demonstrate fundamental advances in HPC capability for large physics modeling problems as well, then LANL will be poised to capture a larger share of Office of Science funding, and to revitalize NNSA’s ASCI program. Our ultimate goal is to change the course of supercomputing.