PVM Grids to Self-Assembling Virtual Machines*

Al Geist
Oak Ridge National Laboratory
PO Box 2008
Oak Ridge, TN 37831-6016
gst@ornl.gov
http://www.csm.ornl.gov/geist

Abstract. Oak Ridge National Laboratory (ORNL) leads two of the five big Genomes-to-Life projects funded in the USA. As a part of these projects researchers at ORNL have been using PVM to build a computational biology grid that spans the USA. This talk will describe this effort, how it is built, and the unique features in PVM that led the researchers to choose PVM as their framework. The computations such as parallel BLAST are run on individual supercomputers or clusters within this P&P grid and are themselves written in PVM to exploit PVM's fault tolerant capabilities.

We will then describe our recent progress in building an even more adaptable distributed virtual machines package called Harness. The Harness project includes research on a scalable, self-adapting cote called H2O, and research on fault tolerant MPI. Harness software framework provides parallel software "plug-ins" that adapts the run-time system to changing application needs in real time. This past year, we have demonstrated Harness' ability to self-assemble into a virtual machine specifically tailored for particular applications.

Finally, we will describe DOE's plan to create a National Leadership Computing Facility, which will house a 100 TF Cray X2 system, and a Cray Red Storm at ORNL, and an IBM Blue Gene system at Argonne National Laboratory (ANL). We will describe the scientific missions of this facility and the new concept of "computational end stations" being pioneered by the Facility.

1 Genomics Grid built on PVM

The United States Department of Energy (DOE) has embarked on an ambitious computational biology program called Genomes-to-Life[1]. The program is using DNA sequences from microbes and higher organisms, for systematically tackling questions about the molecular machines and regulatory pathways of living systems. Advanced technological and computational resources are being employed to identify and understand the underlying mechanisms that enable living organisms to develop and survive under a wide variety of environmental conditions.

ORNL is a leader in two of the five Genomes-to-Life centers. As part of this effort ONRL is building a Genomics Computational Grid across the U.S. connecting ORNL, ANL, Pacific Northwest National Laboratory, and Lawrence Berkeley National Laboratory. The software being deployed is called The Genome Channel[2]. The Genome Channel is a

*The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

computational biology workbench that allows biologists to run a wide-range of genome analysis and comparison studies transparently on resources at the grid sites. Genome Channel is built on top of the PVM software. When a request comes in to the Genome Channel, PVM is used to track the request, create a parallel virtual machine combining database servers, Linux clusters, and supercomputer nodes tailored to the nature of the request, spawning the appropriate analysis code on the virtual machine, and then returning the results.

The creators of this Genomics Grid require that their system be available 24/7 and that analyses that are running when a failure occurs are reconfigured around the problem and automatically restarted. PVM's dynamic programming model and fail tolerance features are ideal for this use. The Genome Channel has been cited in "Science" and used by thousands of researchers from around the world.

2 Harness: self-assembling virtual machine

Harness[3] is the code name for the next generation heterogeneous distributed computing package being developed by the PVM team at ORNL, the University of Tennessee (UT), and Emory University. The basic idea behind Harness is to allow users to dynamically customize, adapt, and extend a virtual machine's features to more closely match the needs of their applications and to optimize the virtual machine for the underlying computer resources, for example, taking advantage of a high-speed I/O. As part of the Harness project, UT is developing a fault tolerant MPI called FT-MPI. Emory has taken the lead in the architectural design of Harness and development of the H2O core[4].

Harness was envisioned as a research platform for investigating the concepts of parallel plug-ins, distributed peer-to-peer control, and merging and splitting of multiple virtual machines. The parallel plug-in concept provides a way for a heterogeneous distributed machine to take on a new capability, or replace an existing capability with a new method across the entire virtual machine. Parallel plug-ins is also the means for a Harness virtual machine to self-assemble. The peer-to-peer control eliminates all single points of failure and even multiple points of failure. The merging and splitting of Harness virtual machines provides a means of self healing.

The project has made good progress this past year and we have demonstrated the capability for a self-assembling virtual machine with capabilities tuned to the needs of a particular chemistry application. The second part of this talk will describe the latest Harness progress and results.

3 DOE National Leadership Computing Facility

In May 2004, it was announced that ORNL had been chosen to provide the USA's most powerful open resource for capability computing, and we propose a sustainable path that will maintain and extend the national leadership for the DOE Office of Science in this critical area.

The effort is called the National Leadership Computing Facility (NLCF) and engages a world-class team of partners from national laboratories, research institutions, computing centers, universities, and vendors to take a dramatic step forward to field a new capability for high-end science. Our team offers the Office of Science an aggressive deployment plan, using technology designed to maximize the performance of scientific applications, and a means of engaging the scientific and engineering community.

The NLCF will immediately double the capability of the existing Cray X-1 at ORNL and further upgrade it to a 20TF Cray X1e in 2004. The NLCF will maintain national leadership in scientific computing by installing a 100TF Cray X2 in 2006. We will simultaneously conduct an in-depth exploration of alternative technologies for next-generation leadership-class computers by deploying a 20TF Cray Red Storm at ORNL and a 50TF IBM BlueGene/L at RNL. These efforts will set the stage for deployment of a machine capable of 100TF sustained performance (300TF peak) by 2007.

NLCF has a comparably ambitious approach to achieving a high level of scientific productivity. The NLCF computing system will be a unique world-class research resource, similar to other large-scale experimental facilities constructed and operated around the world. At these facilities, scientists and engineers make use of “end stations” -best-in-class instruments supported by instrument specialists that enable the most effective use of the unique capabilities of the facilities. In similar fashion, the NLCF will have Computational End Stations (CESs) that offer access to best-in-class scientific application codes and world-class computational specialists. The CESs will engage multi-institutional, multidisciplinary teams and undertaking scientific and engineering problems that can only be solved on the NLCF computers and who are willing to enhance the capabilities of the NLCF and contribute to its effective operation. All CESs will be selected through a competitive peer-review process. It is envisioned that there will be computational end stations in climate, fusion, astrophysics, nanoscience, chemistry, and biology s these offer great potential for breakthrough science in the near term.

The last part of this talk describes how the NLCF will bring together world-class researchers; a proven, aggressive, and sustainable hardware path; an experienced operational team; a strategy for delivering true capability computing; and modern computing facilities connected to the national infrastructure through state-of-the-art networking to deliver breakthrough science. Combining these resources and building on expertise and resources of the partnership, the NLCF will enable scientific computation at an unprecedented scale.

4 References

1. G. Heffelfinger et al., “Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling,” OMICS Journal of Integrative Biology, November 2002, www.genomes-to-life.org.
2. M. Land et al., “Genome Channel,” http://compbio.ornl.gov/channel.
3. G. A. Geist et al., “Harness,” www.csm.ornl.gov/harness.
4. D. Kurzyniec et al., “Towards Self-Organizing Distributed Computing Framwork: The H2O Approach,” International Journal of High Performance Computing, 2003.