

MM5 on Future SGI Platforms

Wesley B. Jones, Ph.D.
SGI
Email: Wesley@sgi.com

June 21, 2000

1 Introduction

SGI™ has a history of working with the mesoscale prediction group at NCAR. We work together to ensure that the full MM5 weather modeling system runs on SGI servers. SGI helps to ensure that the MM5 model¹ runs efficiently in parallel allowing the highest resolution, the largest sized domains and greatest levels of nested runs to be made on SGI servers. SGI is also committed to helping the MM5 user community resolve problems on SGI systems when they occur. This next year SGI will be releasing several new sets of servers, all of which will run MM5 and all will be appropriate for different types of MM5 users and applications of MM5. The SGI 3000 Series servers based on SGI ccNUMA (cache-coherent non-uniform memory access) technology (NUMAlink™), the MIPS™ chip, the IRIX® operating system, and the SGI MIPSpro™ compilers will continue to provide the highest computing power in a cache-coherent single system image (ccSSI). These technologies improve the ease of use of high performance computers and improve the development and application of technical and scientific computational models. Clusters of SGI servers and partitioned systems with the SGI Advanced Cluster Environment, ACE, will provide an improved hardware price/performance ratio and higher resiliency. SGI 1000 Series servers based on commodity technology, the Intel® IA-32 architecture (e.g. Pentium® III), Linux™ operating system, SGI ProPack for Linux™, and the SGI ACE for Linux provide a complete IA32 cluster solution. SGI IA-64 servers are still in the future, but the MM5 model was the first significant FORTRAN application run on an IA64 architecture with Trillian, Linux for IA-64. SGI will have IA-64 servers based on both commodity technology and based on the SGI ccNUMA technology, NUMAlink.

2 SGI Scalable Computing Solution Set

The MM5 weather modeling system includes set of complex applications for the production of a weather

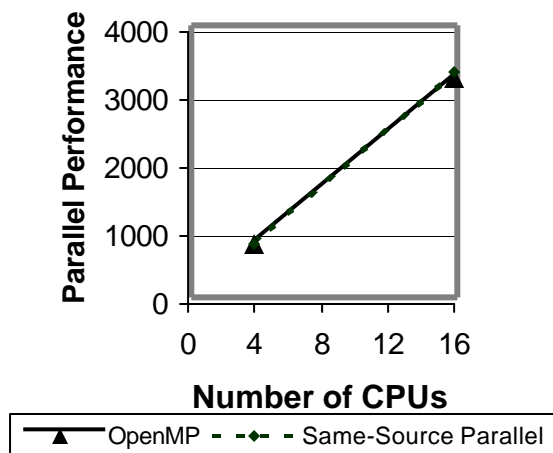
forecast. The most well known and compute intensive application is the fifth-generation Penn State/NCAR mesoscale model (MM5)¹, the numerical weather prediction (NWP) component of the MM5 weather modeling system. The dynamic range of computer power required for different applications of MM5 spans 5 to 6 orders of magnitude². To support this large dynamic range of required compute power, SGI offers a set of scalable computing solutions.

The highest end computing environments are composed of clusters of very large ccNUMA-MIPS-IRIX systems, for example the clusters of 128 CPU SGI 2800 systems at Los Alamos and the 512 CPU SGI 2800 at NASA Ames. These SGI 2000 Series systems will soon be replaced by the SGI 3000 Series systems. The mid range would normally be occupied by the SGI 2200, 2400 and 2800 systems that include 2 to 512 CPUs in a cache-coherent single system image (ccSSI) and allow the standard version of the MM5 model to run in a highly scalable manner using the OpenMP API.³ Because the MM5 model includes the same source parallel version⁴ for support of message passing, mid range computing solutions also include clusters of SGI 2100 systems. The low end of the dynamic range is satisfied by the SGI 2100 Series, Origin200, and clusters of SGI 1000 Series systems depending on the needs and expertise of the user. In the future SGI's scalable computing solution set will include systems based on the Intel IA-64 architecture, starting with the Itanium® processor, and will include both commodity based clusters and NUMAlink based ccNUMA systems both using Trillian, the Linux operating system for IA-64.

In addition to the MM5 model, there are a handful of auxiliary applications for creating initial conditions, post-processing diagnostics and visualization. This set of applications is primarily composed of applications written in FORTRAN90 that must be built by the end user before execution. To enable support and development of the full MM5 weather modeling system, SGI has made available a 4 CPU SGI 1000 Series system on site at NCAR. SGI will also make available a multiprocessor SGI IA-64

system and work with the MM5 support and development team. This is to ensure that when SGI IA-64 servers are available, the complete MM5 weather modeling system will run on the SGI IA-64 architecture using the recently open sourced SGI PRO64™ FORTRAN90 and C compilers for the IA-64 architecture. Current and future SGI MIPS-IRIX servers will be binary compatible with current servers and run the same set of SGI MIPSpro FORTRAN90 and C compilers.

**MM5v3 Performance on
16 CPU 400 MHz SGI 3000 Series**



3 MM5 on SGI 3000 Series (ccNUMA-MIPS-IRIX)

The SGI 3000 Series will soon be available. The SGI 3000 series is a ccNUMA computer with the MIPS R12000™ processor running the IRIX operating system. From a developer, user and administrative perspective, it is nearly identical to the SGI 2000 Series (i.e. the Origin2000). It runs the same operating system, uses the same compilers, and is binary compatible with the SGI 2000 Series. From an architectural perspective it is faster, wider, and more scalable. It has twice the main memory bandwidth per CPU when compared to the SGI 2000 Series and four times the main memory bandwidth per c-brick. Each c-brick contains 4 CPUs as opposed to the SGI 2000 Series which contains 2 CPUs. The NUMAlink interconnect has a lower latency and twice the bandwidth per link compared with the CRAYlink™ of the SGI 2000 Series. The following table includes a plot of the performance of the MM5v3 model run with the OpenMP version and the same-source parallel version.

4 MM5 on SGI 1000 Series (IA-32 – Linux – Intel)

The SGI 1000 Series is a commodity technology based IA-32 server. The SGI 1200 has 2 Intel Pentium III processors and runs the Linux operating system with the SGI ProPack for Linux and SGI ACE (Advanced Cluster Environment) for Linux. We have an SGI 1400 on site at NCAR to work on issues related to enabling the whole MM5 modeling system to run on SGI 1000 series in both shared-memory, OpenMP, mode and on clusters of SGI 1000 Series computers. This work includes helping to update the MM5 model and auxiliary programs to support the combination of software packages required to run MM5, and to identify, report, and get fixed bugs in the Portland Group Compilers that the MM5 auxiliary applications expose. The general MM5 user community will benefit from the reduced exposure to these compile and run time environment problems. It also include understanding how the model runs under the Linux operating system in a cache-coherent single system image (ccSSI) and a cluster environment.

5 MM5 on SGI IA-64

Intel will be releasing 64-bit chips based on the IA-64 architecture. The first chip announced by Intel is the Intel Itanium® processor. SGI will be producing systems based on this chip with systems that are based completely on commodity technology for small servers and clusters and with systems based on SGI ccNUMA technology, NUMAlink, for large servers, partitioned systems and clusters of large servers. The systems will run the Linux operating system and use the SGI PRO64 Compiler suite.

SGI has signed up to port the basic components of the MM5 weather modeling system. As part of this plan a beta IA-64 server will be made available onsite at NCAR for porting the MM5 weather modeling system. This is to ensure that when these systems are generally available, the MM5 weather modeling system will run on the IA-64 architecture using the SGI PRO64 compilers and the end users will not have to wait for the port to be completed. In January 2000 SGI already had the MM5 model running on alpha IA-64 hardware using the SGI PRO64 FORTRAN90 compiler.

The basic plan is to port the following applications: terrain, pregrid, regrid, little_r, MM5v3, tovis5d (new version with diags), vis5d. We are working to

ensure issues with Linux are resolved and supported for each of the applications using the Portland Group Compilers. The next step will be to ensure that all of the applications run with the SGI MIPSpro compilers, compiled in 64-bit mode work. The SGI PRO64 compilers are based on the same front end as the SGI MIPSpro compilers. We will also work to remove any latent bugs associated with uninitialized variables, out of bounds array usage, and FORTAN90 compliance using the advanced features of the SGI FORTRAN90 compilers for MIPS processors. Finally, we will work to ensure that any bugs in the SGI FORTRAN90 compilers are fixed that affect MM5 application or workarounds are provided for the applications required to run a full weather forecast based on the MM5 weather modeling system.

6 SGI ccNUMA, Clustered, and Partitioned Systems

In many cases the price of the hardware to build a cluster of thin, fat or wide node systems is less than the cost of the hardware to build a single very fat or very wide node system. An example of this is the hardware price of a 4 node cluster of 8 CPU SGI 2100 MIPS-IRIX systems is significantly less than the cost of a 32 CPU SGI 2400 MIPS-IRIX system. Part of the savings comes from the cost reduction philosophy of the SGI 2100. It has processors which are slightly slower than the latest processor for the SGI 2200 processor. Yields on slightly slower processors are better than the highest frequency processor and thus reduce the cost of the chip. The secondary cache in the SGI 2100 is smaller than in the SGI 2200 to again reduce cost. The network connecting a cluster of SGI 2100 servers is slower than the ccNUMA interconnect of the SGI 2200, so the SGI 2200 may scale better than a cluster, and this is application and data set dependent. For many MM5 forecasts a cluster of SGI 2100 systems connected together via HIPPI is a perfectly acceptable solution and has a similar hardware price/performance to a production IA32 cluster solution with Myrinet networking.

In favor of ccSSI configurations, the SGI ccNUMA systems offer an ease of use and capability that is unmatched by non-ccNUMA systems such as the IBM SP series and other clusters of shared memory computers. The philosophy for purchasing very wide node ccNUMA based clusters comes from a trend in computing where people and software cost much more than hardware, even value added ccNUMA based technology. The most discussed

capability of ccNUMA systems is the ability to run scalable shared memory applications using the OpenMP application programming interface (API). From a developer standpoint this may be the most important feature and may save a developer hundreds of thousands of dollars a year in development cost. However, the advantages extend far beyond the developer perspective. An example is that a person costs on the order of \$250,000 (including overhead) per year to a national lab or a private company. Consider that a value added ccNUMA based technology may save you the cost of one or more people because there is less administrative cost, less developer cost, more productive users, and higher system utilization. It becomes possible to make up a fractional increase in the hardware cost going from a thin node cluster based architecture to a very wide node ccNUMA based cluster. In the most advantageous scenarios the increased productivity and subsequent creativity and insight leads to the solution of problems on ccNUMA based systems that could not be solved on cheaper or larger thin node clusters.

One feature of a cluster is that it may be more resilient to down time than a single ccNUMA system. For environments that need an extremely high availability of systems, a small cluster, as small as two nodes, may be used to achieve both the resiliency of a cluster and the capability of a ccNUMA architecture. In addition, the SGI 3000 series will introduce the ability to partition the cache-coherent single system image into multiple system images allowing users access to the resiliency of a cluster using a partitioned ccNUMA system.

¹ Grell, G.A., J. Dudhia, and D. R. Stauffer, 1995: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Technical Note NCAT/TN-398+STR, 138 pp.

² W. B. Jones, 1999: Scaling up, Scaling down: MM5 on SGI. *Preprint, The Eighth PSU/NCAR Mesoscale Model User' Workshop, 19-21 July, 1999, Boulder Colorado*

³ W. B. Jones, 1997: A Highly Scalable Shared Memory Version of MM5, MM5V2+. *Preprint, The Sixth PSU/NCAR Mesoscale Model User' Workshop, 21-23 July, 1997, Boulder Colorado.*

⁴ John Michalakes, 1998: *The Same-Source Parallel MM5*. In proceeding of the *Second International Workshop on Software Engineering and Code Design in Parallel Meteorological and Oceanographic Applications*. Preprint NASA GSSFC/CP-1998-206860, pp. 129—139.