An Extended Model of HPCG Performance for ccNUMA Architectures

A.V. Levchenko, E.P. Petukhov

Supercomputer Center “Polytechnic”, SPbPU

In this paper, we present an extended performance model that will provide insight into performance prediction of ccNUMA architectures. The High Performance Conjugate Gradients (HPCG) Benchmark was used for creating a workload with the low ratio of computations to data access that is representative to the major communication and computational patterns. The actual allowed optimizations [1], generalized experience of the early optimizations [2] and general-purpose performance model [3] of the HPCG Benchmark are known.

The rich memory hierarchy of ccNUMA architecture can lead to hybrid MPI/OpenMP performance degradation due to the reduction operations, memory hot-spotting and mismatch of data access model and actual distribution of data in memory. A scalable directory based cache coherence protocol along with using L4 cache for the acceleration of remote memory transactions have a complex multilevel structure of latency. This combination of strong NUMA effects renders undetermined influence on memory bandwidth. Thereby, the hypothetical performance prediction of ccNUMA systems appears unconvincing. A HPCG-based extended model will allow to measure reliably the performance of current and future ccNUMA systems and to compare it with the results of other problem oriented architectures.

Our specific contributions of this paper are (1) to extend the basic model [3] for the resource of available global shared memory of *macronode(s)* with terabytes of RAM running in a single operating system image mode and (2) to provide a set of characteristics that affect ccNUMA system performance. The model is based on measuring performance time of HPCG computing kernels (the largest of them is *SimGS*) and communication operations. In contrast to the original model [3], the first things studied are features of hybrid HPCG implementation; the significance of communication overhead increases for the ccNUMA substantially in proportion to the aggregation of macronode presumed by single-node system concept. In particular, the *MPI_Allreduce* operation performance time model has been redefined with regard for groups of NUMA domain processes. The model takes account of the number of started threads and does not allow the OpenMP dynamic balancing feature that causes an application crash in a number of cases. The OpenMP global reduction operations are the major NUMA-related bottleneck in the process of conjugate gradients method implementation. Besides, threads idleness time during *fork/join* operations is taken into account. Thereby, total performance is proportional to the memory bandwidth of a macronode with regard for communication overhead.

We ran our model validation experiments on a global shared memory system, installed at Supercomputer Center “Polytechnic”, SPbPU. This target system is essential for scientific in-memory computing. Based on a global cache and memory coherency model, the ccNUMA architecture allows to load the multiple nodes as a single image operating system environment with 12Tb of RAM, 3072 cores/1536 FPU. For a more in-depth study of NUMA-related challenges, we performed our early-stage experiments with hybrid HPCG running on macronodes from 188Gb of RAM (48 cores) with aggregation of macronode memory to 3Tb of RAM (768 cores) and with subsequent integration into a single macronode with up to ≈12Tb of RAM (3072 cores) at the final stage. Optimized *libgomp* supports more than 1024 threads, a stack is allocated to each thread. Generation of instructions to prefetch memory is used for increasing performance of loops that access large arrays. Load is balanced for improving efficiency of OpenMP application, distributing threads through all accessible NUMA nodes, using more FPU and reducing load on the memory interface and L3 cache.

Figure 1 compares our predictions with the actual measured results of hybrid HPCG on

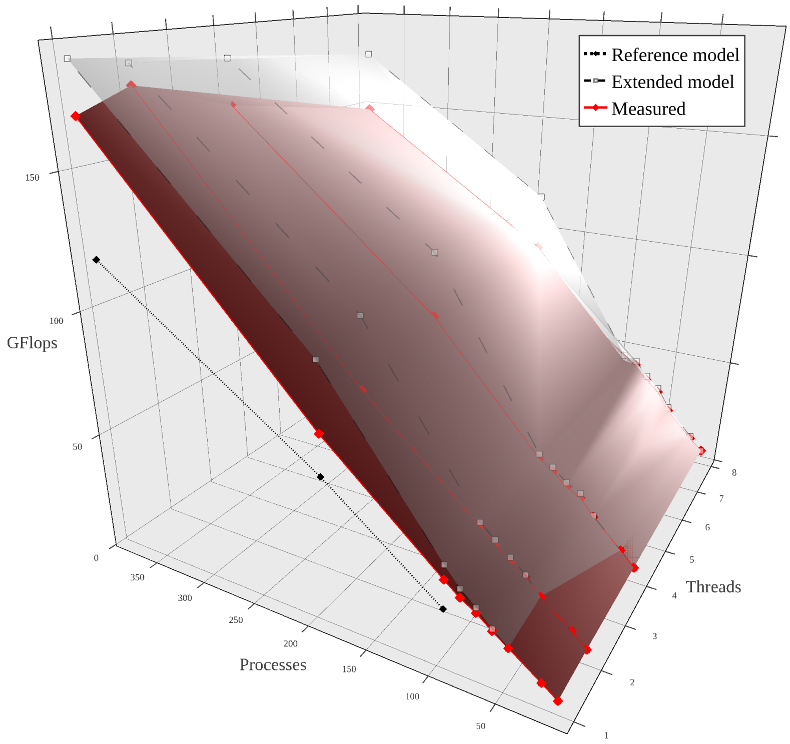

Figure 1. Modeled vs. measured HPCG results on the target macronode with 12Tb of RAM

the macronode with 12Tb of RAM. In contrast to the results of work [3], the hybrid HPCG scales non-linearly; non-uniformity of the system results in surface separation whose causes will be studied. The most significant NUMA-related characteristics are latency for frequent cache coherence traffic and remote-read overhead. Thus, the preliminary model demonstrates relative predictability of HPCG performance for macronode with up to 12Tb of RAM.

Future work will include the refinement of the cache locality model with the help of the novel HPCG optimization technique proposed in the paper [2], namely coloring along two areas XY at a time in *SymGS*.

References

1. Dongarra J., Heroux M. A., Luszczek P. High-Performance Conjugate-Gradient Benchmark: a New Metric for Ranking High-Performance Computing Systems // International Journal of High Performance Computing Applications. 2016. Vol. 30, No. 1. P. 3–10. URL: http://dx.doi.org/10.1177/1094342015593158

2. Agarkov A., Semenov A., Simonov A. Optimized Implementation of HPCG Benchmark on Supercomputer with “Angara” Interconnect // Proceedings of the 1st Russian Conference on Supercomputing — Supercomputing Days 2015, Moscow, Russia, September 28-29, 2015 / Ed. by Vladimir Voevodin. Moscow, Russia : Research Computing Center, Moscow State University, 2015. P. 294–302. URL: http://2015.russianscdays.org/files/pdf/294.pdf

3. Marjanović V., Gracia J., Glass C. W. Performance Modeling of the HPCG Benchmark // High Performance Computing Systems. Performance Modeling, Benchmarking, and Simulation: 5th International Workshop, PMBS 2014, New Orleans, LA, USA, November 16, 2014. Revised Selected Papers / Ed. by Stephen A. Jarvis, Steven A. Wright, Simon D. Hammond. Cham : Springer International Publishing, 2015. P. 172–192. ISBN: 978-3-319-17248-4. URL: http://dx.doi.org/10.1007/978-3-319-17248-4_9