

Characterizing memory system performance for local and remote accesses in high end SMPs, low end SMPs and clusters of SMPs

Ch. Kurmann and T. Stricker

Laboratory for Computer Systems, ETHZ - Swiss Institute of Technology,
CH-8092 Zuerich, SWITZERLAND
{kurmann, tomstr}@inf.ethz.ch

Symmetric Multi-Processors (SMPs) with coherent shared memory have grown into the large market of high performance workstations and high end servers, usable for engineering and database applications. Numerous researchers and technologists have suggested that low end SMP designed by PC manufacturers will become the universal compute nodes for loosely coupled distributed systems and for strongly coupled massively parallel processing systems (MPPs). However up to this date, high end systems and low end systems still differ significantly. High end designs include a carefully engineered memory system with support for any additional data streams caused by memory intensive computations or by the inter-node communication. Low end designs feature extremely low cost, but much weaker memory systems. The questions of our interests are: How significant are the differences in memory system performance? Do those cheaper low end memory systems pose serious limitation for scientific applications and databases? If so, are they only a problem for the local computation or for inter-node communication in distributed systems?

ETH Project CoPs - Building a Cluster of PCs based on multiprocessor Intel Pentium Pro nodes.

Like other research groups in many universities and national laboratories our computer architecture group is currently engaged in designing and building a cluster of PCs from off-the-shelf SMPs linked by a commercial Gigabit interconnect (e.g. Dolphin SCI, Myricom Myrinet or Gigabit Ethernet). As computer scientists we are not trying to deploy some cheap GigaFlops to our colleagues in computational chemistry or physics, but are most interested in understanding the performance of the different SMPs and uni-processors used as compute nodes and our aim is to improve the system software used in such clusters.

Characterizing SMP Memory Systems Performance

As a first step we developed a novel memory system microbenchmark that is independent from naming and coherence issues (i.e. CC SMP, CC NUMA, NON-CC NUMA or Msg Passing) and is therefore capable to characterize the memory system performance for both *local* and *remote* memory, regardless of the underlying architecture. Unlike the simplistic McCalpin loops [2] our test captures more aspects of the memory hierarchy, in particular its behavior with temporal- and spatial locality (varying working set and stride) [4]. The same method is also used to measure remote accesses for machines with partial or without support for coherent shared memory [3].

In the talk we give a detailed picture of the memory systems in low end SMPs like Dual Pentium Class PCs and compare those to high end systems like a DEC 8400, an SGI Onyx 10k, a Sun Enterprise Server Figures 1 and 2 show a sample of the performance data we collected. An important issue is whether the memory system of low cost (internal bus based) SMPs can fully sustain multiprocessing. Figures 3 and 4 try to answer this question by comparing single processor usage to twin/quad processor usage.

Local memory system performance is not only highly important to computational efficiency in applications with large datasets, but it is also the key to fully sustainable inter-node communication. While high end systems can afford memory systems with special hooks for inter-node communication at Gigabit/s speeds, low end systems must rely entirely on standard I/O interfaces (i.e. a PCI bus) for economic reasons. For remote memory accesses we obtained measurements from a small testbed of a few PCs connected by several fast networking technologies. We compare some data from an SGI Cray T3E (non CC-NUMA) or an SGI Origin/Infinite Reality (CC-NUMA), see Figures 5 and 6.

Conclusions

From the memory system characterizations presented we can conclude that only the good memory systems of high end SMPs scale with the number of processors. In low end SMPs, the benefits of symmetric multiprocessing ends abruptly as the working set exceeds the L2 or L3 caches that are located close to the microprocessors.

For the remote systems performance we notice excellent performance on the high end MPP nodes (SGI T3E or Origin). For some copy operations the throughput of remote memory accesses is even higher than for local copies because multiple memory banks are involved. In the low end systems interconnected by Gigabit networks the performance can peak near the bandwidth limits of the PCI-Bus, but only for simple transfer modes like remote store (push) of contiguous blocks of data. For strided data or for remote loads of single words the performance in low end SMPs collapses. We expect that MPP systems built from those low end SMP will get into severe difficulties with applications that require large data sets and complex, dense communication patterns.

Figures/Bibliography, see next page.

1 Figures and bibliography

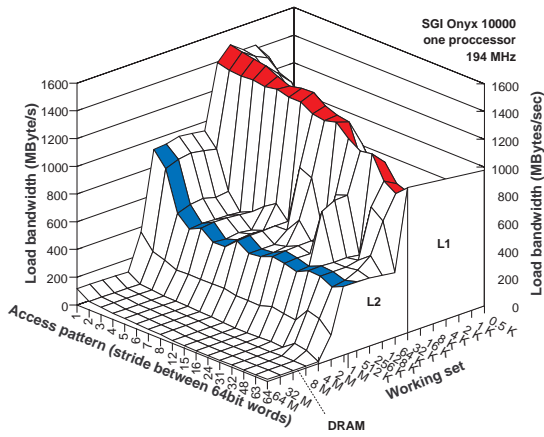


Figure 1: Bandwidth of loads for different access patterns (strides) and different working sets on an SGI Onyx 10000. One processor runs the memory benchmark while the other three processors are idle.

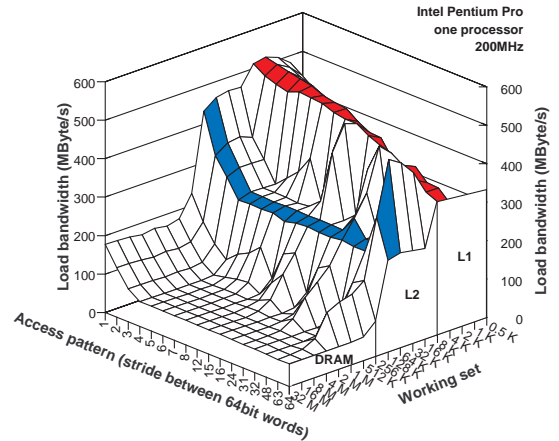


Figure 2: Load bandwidth for different access patterns (strides) and different working sets on a Dual Pentium Pro PC. One processor runs the benchmark while the second is idle.

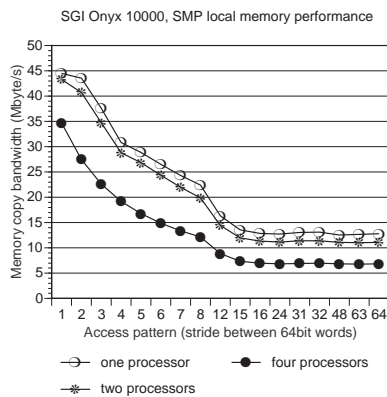


Figure 3: Measured performance of the local memory system of an SGI Onyx 10000 for large transfers to a shared memory segment, with either one, two and four processors. The transfer is done by strided loads and contiguous stores.

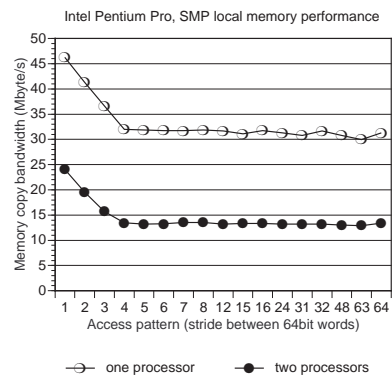


Figure 4: Measured performance of the local memory system of a Pentium Pro PC for large transfers to a shared memory segment, with either one or two processors. The transfer is done by strided loads and contiguous stores.

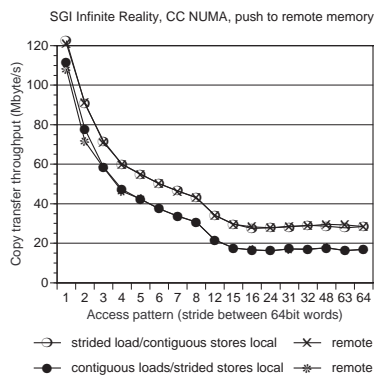


Figure 5: Measured performance for large transfers in the memory system local on a node and remote to a second node for an SGI Origin Infinite Reality. Results for either contiguous or strided stores.

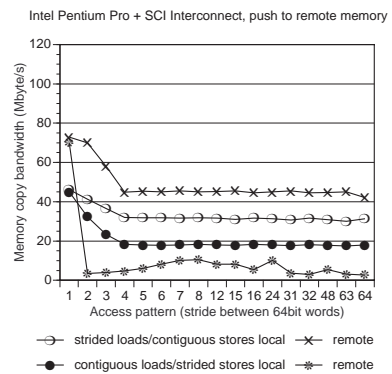


Figure 6: Measured performance of large transfers in the local memory system of a Pentium Pro PC (440FX Chipset) and to a second system connected by an SCI-Interconnect via the PCI-Bus. Results with either contiguous or strided stores.

References

- [1] INTEL Corporation. *INTEL 440 FX PCISSET*, 1996.
- [2] John D. McCalpin. Sustainable memory bandwidth in current high performance computers. Technical report, 1995.
- [3] T.Gross T.Stricker. Optimizing memory system performance for communication in parallel computers. In *Proceedings of the 22nd International Symposium on Computer Architecture (ISCA22)*, 1995.
- [4] T.Gross T.Stricker. Global address space, non-uniform bandwidth: A memory system performance characterization of parallel systems. In *Proceedings of the ACM conference on High Performance Computer Architecture (HPCA3)*, 1997.