

Future Building Blocks for Parallel Architectures

Ulrich Bruening
University of Mannheim

Wolfgang Giloi
Fraunhofer Institute for Computer Architecture and Software Technology

Early parallel architectures were shared memory systems (UMA, NUMA), which had the disadvantage of the shared memory bottleneck that limited the scalability of the system. In contrast, distributed memory architectures with message passing (NORMAs) provided any desired scalability; however, at the cost of a substantial communication latency. The latency could be reduced by custom communication hardware (examples: SUPRENUM, MANNA) yet since there was still a software routine involved, the remaining latency was in the order of microseconds. Therefore, and because of the simpler programming model of shared memory, it became the trend of the nineties to return to UMAs and NUMAs, employing powerful communication hardware to minimize the remote memory access time. This approach required a complex, highly expensive custom chip set. State-of-the art is to augment the processor hardware on the chip by facilities such as memory controllers, caches, cache-coherent multiple links with switches for NUMA support. This is called "glue-less NUMA," an example being the AMD Opteron. Since one wants to obtain the performance advantages of cache hierarchies with their need for cache coherence mechanisms, the complexity of the additional communication hardware increases exponentially with the number of nodes, thus limiting the scalability severely. Consequently, the demand for scalable, massively parallel systems built with cost-effective COTS (commercial off the shelf) hardware has revived the NORMA architecture, realized with very fast communication devices that are connected to a high-performance IO interface (PCI-X, PCI Express, Hypertransport, ...). Examples for such devices are Myrinet, Infiniband, Quadrix QsNet, or Atoll, which all feature crossbars as interconnects. However, because of the "distance" of such a device from the processor, communication latency is still significantly higher than the local memory access time. This calls for moving as a next step the communication facilities on the chip as multithreading devices, with the interconnect crossbar directly on the die. This will eliminate the multiplexing of the device and the bottleneck of the processor bus. The ever increasing number of transistors on the chip will readily allow such a move. The instruction set of such a processor will be extended by send, receives, remote loads, remote stores, hence providing all the flavors of communication types as machine instruction with a start-up latency of one processor clock tick. The speed-up obtainable with such building blocks will be quantified.