numascale

IT Center | RWTH AACHEN UNIVERSITY

Numascale

Numascale is a European SME specialising in interconnects for high performance and enterprise computing. The differentiator for Numascale's interconnect are the shared memory and cache coherency mechanisms. These features allow programs to access any memory location and any memory mapped I/O device in a multiprocessor system with a high degree of efficiency. It provides scalable systems with a unified programming model that stays the same from the small multi-core machines used in laptops and desktops to the largest imaginable single system image machines that may contain thousands of processors. The architecture is commonly classified as ccNUMA or NUMA but the interconnect system can alternatively be used as a low latency clustering interconnect. The technology comes as an add on card to standard servers providing a large shared memory at cluster price.

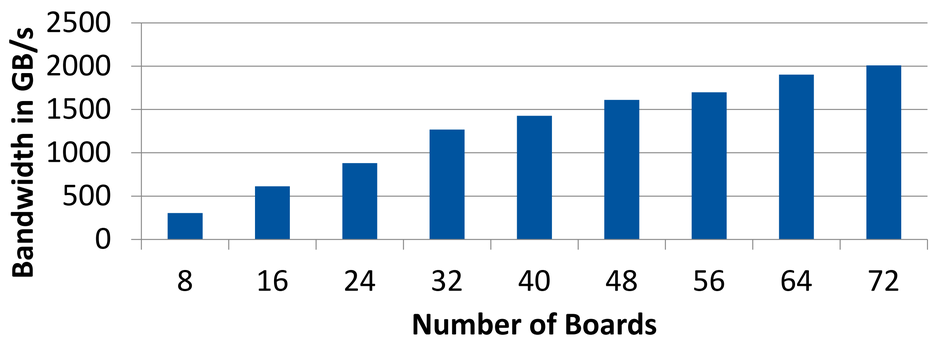
The Oslo System

The University of Oslo's Center for Information Technology, USIT feature this Numascale installation. The systems is a PRACE prototype and has been used for some of the work between Numascale and RWTH Aachen University.

- 72 IBM x3755 M3 nodes
- 144 AMD 6174 CPUs
- 1728 Cores
- 4.6 TB Memory
- 3D Torus Interconnect (3x6x4)

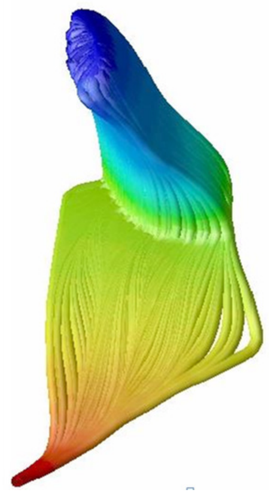
Memory Bandwidth*

- Memory bandwidth is essential for many scientific applications
- Every processor has its own memory and adds to the total memory bandwidth
- Stream [1] is a standard benchmark to evaluate the memory bandwidth of a system



TrajSearch

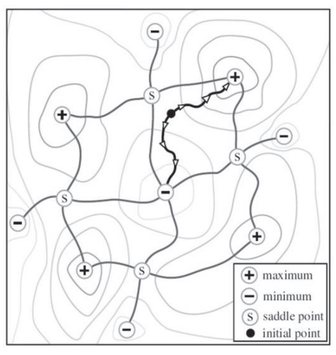
TrajSearch is a code to investigate turbulences which occur during combustion. It is a post-processing code for dissipation element analysis developed by Peters and Wang ([2]). It decomposes a highly resolved three dimensional turbulent flow field obtained by Direct Numerical Simulation (DNS) into non-arbitrary, space-filling and non-overlapping geometrical elements called 'dissipation elements'. Starting from every grid point in the direction of ascending and descending gradient of an underlying diffusion controlled scalar field, a local maximum, respectively minimum point is found. A dissipation element is defined as a volume from which all trajectories reach the same minimum and maximum point. The dissipation element analysis provides a deeper understanding of turbulence and can be employed to reconstruct important statistical properties as has been shown by Gampert and Göbbert ([3]).



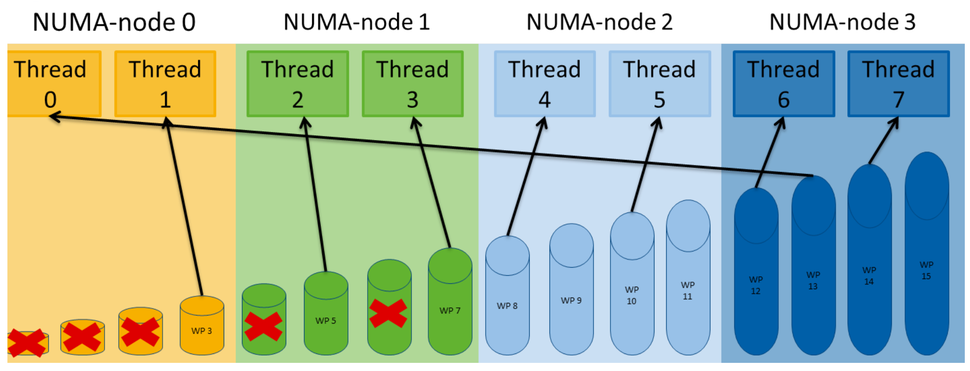
Scaling OpenMP Programs to Thousand Cores on the Numascale Architecture

Optimizations for Large NUMA Systems

- Reduced Synchronization Overhead:
  - local data copies where possible
  - local buffers to synchronize writes
  - multithreaded memory allocator, kmp_malloc
  - thread local random number generator
- NUMA aware memory placement
- Since the length of the trajectories differs, load imbalance occurs. The Adaptive NUMA Scheduler is used to handle load imbalance and maintain as much data locality as possible

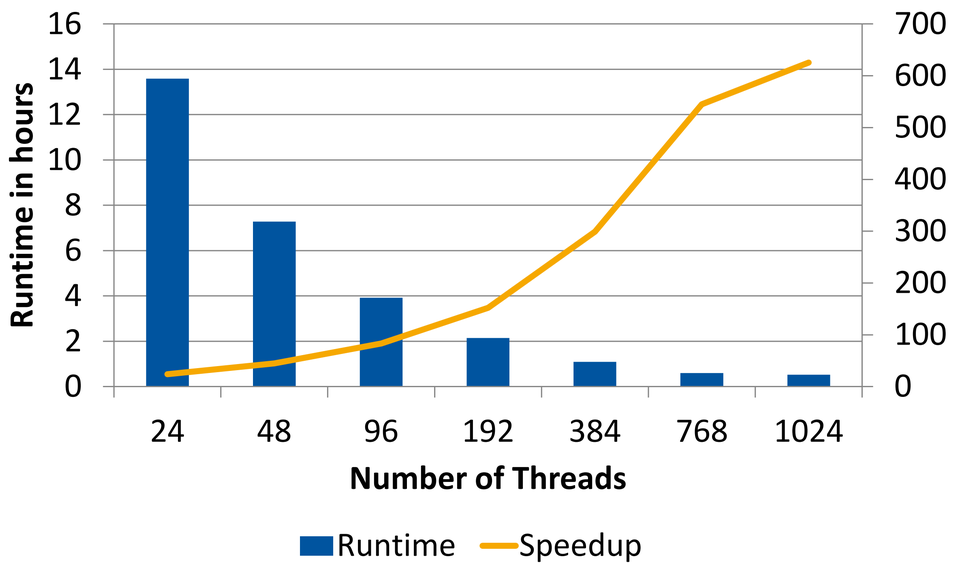


The Adaptive NUMA Scheduler



- Combines the principles of static and dynamic scheduling
- Static layout defines how work should be distributed
- Idle threads take work from the threads with least progress
- Work of foreign threads will get iterations from the highest index backwards to reduce interference

Performance Results*



Runtime and Speedup of Trajsearch with an increasing number of threads.

*Executables were compiled with the Oracle Solaris Studio 12.3 Compiler.

Summary

- The investigated Numascale system combines 1728 cores and 4.5 TB of main memory in a shared memory machine
- The accumulated memory bandwidth increases with the number of nodes to over 2 TB/s
- The application TrajSearch was optimized to run on large NUMA systems by reducing synchronization, optimizing the memory layout and using the Adaptive NUMA Scheduler for work scheduling
- Overall the code delivers a speedup of 625 running with 1024 threads

References

[1] John D. McCalpin. STREAM: *Sustainable memory bandwidth in high performance computers*, http://www.cs.virginia.edu/stream/, 1995.

[2] N. Peters and L. Wang. *Dissipation element analysis of scalar fields in turbulence.* C. R.Mechanique, 334:493–506, 2006.

[3] Markus Gampert, Jens Henrik Göbbert, Michael Gauding, Philip Schäfer, and Norbert Peters. *Extensive strain along gradient trajectories in the turbulent kinetic energy field.* New Journal of Physics, 2011.

Dirk Schmidl, RWTH Aachen University
schmidl@itc.rwth-aachen.de

Atle Vesterkjær, Numascale AS
av@numascale.com