Optimizing processes on multicore for speed and latency

Christoph H. Lameter
Graphe Inc.
cl@gentwo.org

Abstract

High Performance Computing is coming to every file server and every desktop machine these days. The processing power of the average server will grow significantly with the introduction of Westwere (24 threads dual socket), Nehalem EX for quad socket (64 threads) and 8 socket machines (128 threads) which will make entirely new applications possible. In the financial market it is then possible to run a complete trading system setup on a single machine as demonstrated by running a NYSE simulation at the IDF conference by Intel. However, the same issues already show up with a smaller effect even on the run of the mill dual quad core systems.

Intel Nehalem processors support NUMA – a technology so far only known from large supercomputers. The NUMA effects are small on todays dual quad core file servers but as the number of processors rises the distances of processors to memory will also increase and put more demands on the operating system and application software to obtain memory that a processor can reach in an efficient manner. Temporal locality issues dominate even within a core because the most effective storage is in the L1 cache that is local to one execution context but unreachable from another. It is vital that techniques originally developed for HPC are used to exploit the full potential that todays hardware provides.

We will discuss Westmere, Nehalem EX, temporal and spatial locality management techniques, managing cpu caches, hyperthreading, latency and performance in Linux.

Proceedings of the Linux Symposium

July 13th–16th, 2010
Ottawa, Ontario
Canada

Conference Organizers

Programme Committee

Proceedings Committee