Introduction: This document provides an overview of the performance features of the Intel® Nehalem processor.

Overview of Nehalem
Nehalem is the newest micro-architecture from Intel. Nehalem is designed to be modular and scalable. To enable this, Intel divided the chip into 2 parts the core and the un-core. The core contains the execution units and associated logic, and L1 and L2 caches. It is designed to be easily replicated for dual core, quad core, and beyond. The un-core is the shared L3 cache, memory controller, and I/O. This new architecture will allow the same processor core to be used in everything from laptops to servers. The number of cores, and the capacity of the un-core can be scaled for the needs of a particular market.

Performance Features of the “Core”
Nehalem has updated the Core microarchitecture with a number of new features to improve performance:

- More macofusion instructions to improve through put and lower execution latency.
- The Loop Stream Detector (LSD) has been move to after instruction decode so the decode logic can be disabled while the LSD replays micro-ops for each loop iteration. It has also been expanded in size to improve performance by allowing larger loops to be detected.
- Faster synchronization primitives to improve scalability of multi-threaded applications. This also allowed Intel to re-introduce Hyper-Thread.
- Hyper-Thread allows 2 threads to be run at the same time per core. HT allows the core to hide the latency of a single thread thus enabling more effective use of the wide execution engine. Figure 1 shows the performance improvement in several application benchmarks when Hyper-Threading is enabled.
- Turbo Mode allows the processor to opportunistically improve performance by raising the frequency of the processor when there is thermal and electrical headroom. The magnitude of the frequency improvement is greater as the number of active cores decreases.

Performance Features of the “Un-Core”
In addition to microarchitecture performance features, Nehalem also radically changed the way memory and I/O are accessed. The Nehalem architecture eliminated the front side bus for accessing memory and I/O and moved to one where each CPU package includes an integrated memory controller and one or more high speed serial links known as Quick Path Interconnect (QPI) to access other CPU packages and I/O. The result is a massive improvement in memory and I/O bandwidth. Figure 2 shows the improvement in memory bandwidth as measured by the STREAM benchmark.
Intel also revamped the processor cache implementation. The new cache hierarchy is known as Smart Cache. The Smart Cache implementation on Nehalem has 2 small caches included in each processor core to improve performance and scalability. A shared 3rd level cache is included in the “un-core”. This allows the cache to vary in size with the number of cores, and can easily be increased in size with future implementations. The L3 cache has an inclusive policy – everything in each core’s L1/L2 cache must be present in the L3 cache. This keeps snoop traffic constant as the core count increases, and minimizes the effective cache latency by eliminating cross-core snoops in the common case.

Conclusion

The Nehalem architecture provides a high performance, scalable platform to address the needs of many markets. The new design eliminates many of the bottlenecks present in the previous Front Side Bus based architecture, and allows more effective use of the execution resources available in the Intel Core architecture. The new architecture will allow many applications to see a significant improvement in performance over the previous generation.