

Next Generation Intel® Microarchitecture Nehalem

Paul G. Howard, Ph.D.
Chief Scientist, Microway, Inc.

Copyright 2009 by Microway, Inc.

Intel® usually introduces a new processor every year, alternating between a new microarchitecture and a new process technology, including a die shrink of the current microarchitecture. Nehalem¹ is the latest microarchitecture, using the same 45 nm process used in later versions of the Core microarchitecture (Penryn), but with a brand new design.

Some of the more important new features of Nehalem are:

- The new Intel Quick Path Interconnect, replacing part of the front side bus
- An integrated memory controller manages triple-channel DDR3 memory, replacing the rest of the front side bus
- Restructured cache memory, including the introduction of a large inclusive L3 cache
- Integrated power management, including dynamic overclocking
- The return of Hyper-Threading as an integral part of the design
- Some new SSE instructions for string processing, pattern recognition, and networking
- Modular design, allowing a wide variety of chip and package configurations.

Many of the new features are transparent to users. The strength of Nehalem is that it can run existing code faster and with much more efficient power usage than the Core microarchitecture that preceded it.

Road map

Intel refers to their processor development model as a “tick-tock” strategy, the tick being the migration to a new process (typically a die shrink) and the tock being the new microarchitecture, with about one year between successive ticks and tocks. The rationale is that keeping the logical design separate from the physical process will help isolate any issues that arise. The previous microarchitecture, named Core, included chips manufactured first on a 65 nm process² and later on a 45 nm process³. The Nehalem microarchitecture includes Nehalem⁴ (45 nm, the tock), now available, and Westmere (32 nm, the next tick), planned for the 4th quarter of 2009. The next Intel microarchitecture will include Sandy Bridge (32 nm, planned for 2010) and Ivy Bridge (22 nm in 2011). After that will come Haswell (22 nm in 2012).

Nehalem microprocessors

Intel has based a number of different microprocessors on the Nehalem microarchitecture. The Xeon 5500 sequence, Nehalem EP (Efficient Performance)⁵, intended for the server market, includes more than a

1 Named after the Nehalem River in Oregon.

2 Core-based processors manufactured on the 65 nm process included processors code-named Merom for laptops, Conroe and Kentsfield for desktops, and Woodcrest (Xeon 5100 series), Clovertown (Xeon 5300 series), and Tigerton (Xeon 7200 and Xeon 7300 series) for servers.

3 Core-based processors manufactured on the 45 nm process included processors code-named Penryn for laptops, Wolfdale and Yorkfield for desktops, and Harpertown (Xeon 5400 series) and Dunnington (Xeon 7400 series) for servers. The 45 nm Core microarchitecture family is sometimes referred to as the Penryn family.

4 The terminology is somewhat confusing. The term “Nehalem” refers to the microarchitecture, to the 45 nm process family within the Nehalem microarchitecture, and sometimes to the Nehalem EP (Xeon 5500) sequence. Also to the river.

5 Code name “Gainestown”

dozen different microprocessors, differing in clock frequency multiplier, power consumption, memory speed, and internal bandwidth. These are 64-bit quad-core processors with Intel Quick Path Interconnect, three-level cache, triple-channel DDR3 ECC memory, and on-board power management. Nehalem EP for 2-socket servers was launched by Intel in late March 2009. The following table gives the specs for the higher-end Xeon 5500-sequence processors of most interest for HPC.

	Clock speed (GHz)	Clock Multiplier	TDP⁶ (W)	Memory speed (MHz)	QPI Speed (GT/s)	L2 cache (KB)	L3 cache (MB)
Xeon W5580	3.20	24	130	1333	2 x 6.4	4 x 256	8
Xeon X5570	2.93	22	95	1333	2 x 6.4	4 x 256	8
Xeon X5560	2.80	21	95	1333	2 x 6.4	4 x 256	8
Xeon X5550	2.66	20	95	1333	2 x 6.4	4 x 256	8
Xeon E5540	2.53	19	80	1066	2 x 5.86	4 x 256	8
Xeon E5530	2.40	18	80	1066	2 x 5.86	4 x 256	8
Xeon E5520	2.26	17	80	1066	2 x 5.86	4 x 256	8

Other lower-performance “value” chips in the Xeon 5500 sequence include the Xeon E5506, E5504, E5502, L5518, L5506, and L5508. These processors have slower clocks, smaller L3 cache, slower QPI speeds, and slower memory, and do not have Hyper-Threading or Turbo Boost.

The Core i7⁷ processors, intended for desktop use, use non-ECC (i.e., non-error-correcting) memory. Non-ECC memory a) is slightly faster than ECC memory, by less than 5 percent; b) consumes slightly less power, by less than 10 percent; c) costs less, because ECC requires extra RAM chips on the DIMMs; *but* d) is less reliable: a few memory errors per year, mainly due to background cosmic radiation. Quad-core Core i7 processors include Core i7-920, Core i7-940, Core i7-950, Core i7-965 Extreme, and Core i7-975 Extreme. Core i7 processors were first released in November 2008, and the Core i7-950 and Core i7-975 Extreme were announced in early June 2009. The Xeon W3500 series are single-socket server processors, announced in late March 2009. They support ECC memory and can't be overclocked, but are otherwise essentially the same as Core i7.

Intel QuickPath Interconnect

The new Intel QPI (QuickPath Interconnect) takes over the CPU-to-CPU and CPU-to-I/O hub functions of the front side bus. QPI is a 20-bit, direct connect, socket-to-socket design, capable of 6.4 GT/s (billion transfers per second), or 12.8 GB/s in both directions. Because of Nehalem's modular design, multicore processors can be built with more QPI links, giving some degree of scalability for between-core communication.

Memory

With the IMC (integrated memory controller) in Nehalem, Intel has eliminated CPU-to-memory part of the front side bus bottleneck of previous microarchitectures. In Nehalem, each socket has three channels to DDR3 SDRAM, with one or two 8GB DIMMs per channel. Three DDR3-1333 channels gives 32 GB/s theoretical bandwidth (3 channels x 1.333 GT/s x 8 bytes/transfer). Why 3 channels? Two channels don't give enough bandwidth; four channels require too many data pins on the package. Why DDR3 memory? It uses 30 percent less power than DDR2 and has a wider prefetch buffer.

⁶ Thermal design power, the amount of power that the system must dissipate

⁷ Code name “Bloomfield”

The IMC also reduces memory latency by removing one hop (CPU to front side bus) from CPU-to-memory transfers. In conjunction with the QPI, each core is only one additional CPU-to-CPU hop away from another core's memory.

Cache

Nehalem features a new 3-level cache design, all cache being on the die. Each core has its own 32KB 4-way set associative⁸ L1 instruction cache and its own 32KB 4-way set associative L1 data cache, each with 4 clocks latency (up from 3 clocks in Core). Each core has a 256KB 8-way set associative L2 cache, with 11 clocks latency (down from 15). Nehalem's L2 cache is fast but a bit small because of the use of 8T (8 transistors per cell) SRAM, which uses less power but (obviously) uses more transistors than the traditional 6T SRAM⁹.

The big change is the introduction of a large 2 to 3 MB per core shared 16-way set associative L3 cache. The L3 cache is *inclusive*; that is, for each core, the L3 cache mirrors the L1 and L2 caches. This means that after L1 and L2 misses, a core can usually effectively snoop the other cores (search them to see if their L1 or L2 caches hold the required data) without affecting the other cores. The cost of inclusivity is that a substantial part of the L3 cache is used just for the mirroring of lower-level caches. L3 latency is about 40 clocks, or about 15 ns at 2.66 GHz. By contrast, main memory latency is over 100 clocks.

Nehalem also introduces a second-level TLB (translation lookaside buffer), a processor cache used to translate virtual memory addresses to physical addresses.

Power management

During the entire Nehalem design process, Intel has required that each new feature's power requirements must be accompanied by an equal or greater performance increase, based on quantifiable metrics. The result is greater power efficiency at any power envelope.

To provide precise control over power usage, Intel has introduced an extremely flexible and efficient power management system: the PCU (Power Control Unit), a programmable microcontroller. The PCU monitors CPU load, heat, and power draw, and allows per-core power control. An idle core can be put into a reduced power state, or even turned off, independent of all other cores. In addition, Turbo Boost Technology provides a degree of dynamic overclocking. If some cores are in reduced power states and the temperature is low enough, the clock speed multiplier of other, more active, cores may be increased by one or even two steps, allowing the CPU speed to increase by 133 MHz or 266 MHz.

Because the PCU is built onto the die, it does not consume system resources. In fact, power control is transparent to the operating system.

Hyper-Threading and architecture enhancements

With Nehalem, Intel reintroduces Hyper-Threading, their implementation of SMT (Simultaneous Multi-Threading). As with Hyper-Threading on the Pentium 4, the idea is that some parts of the processor core

⁸ Set associative cache involves a tradeoff between fully associative cache, which is fast but requires complex hardware logic to search many slots at once, and direct mapped cache, which is simpler to implement but results in more collisions and hence more cache misses. The more “ways” in a set-associative cache, the closer you are to fully associative.

⁹ L1 cache also uses 8T SRAM.

are duplicated, namely the processor registers, the return stack, and the large-page ITLB (instruction translation lookaside buffer); the remaining processor resources are shared, some statically and some competitively. Because there are two sets of registers, two threads can be active in the core at a time. To oversimplify a little bit, a scheduler passes instructions from the two threads to the appropriate execution units, keeping them busy even when one thread is stalled. The details are a little more complicated: since Nehalem is a 4-issue super-scalar out-of-order processor, if there is only one thread, the scheduler tries to issue 4 instructions to the execution units per clock from the currently ready set of instructions from that thread. With Hyper-Threading, there are two threads contributing to the mix of instructions ready for issue, and hence more chance that the scheduler can actually issue 4 of them.

The effect of Hyper-Threading will be to improve performance of highly threaded applications, especially those that are not competing for processor resources. It becomes more important for developers to use multithreading optimization tools.

Nehalem incorporates a number of low level changes to take full advantage of the Hyper-Threading capability. The instruction ROB (reorder buffer) is expanded to 128 entries, statically partitioned between the two threads to prevent starvation. The reservation station is expanded to 36 entries, competitively shared between the threads to handle changing workloads. The number of load buffers and store buffers are also expanded; these data buffers are statically partitioned between the threads.

In addition, Nehalem includes a LSD (loop stream detector) and instruction cache after the instruction fetch and decode stages, to reduce latency and power consumption by eliminating those stages during loop execution.

Instruction set

Nehalem's instruction set features seven new SSE¹⁰ instructions, collectively called SSE 4.2. Five of the new instructions are for string comparisons, useful for text processing, especially XML processing. The others, referred to as “application targeted accelerators”, are a CRC32 accumulate instruction that can greatly speed up certain low-level network and storage protocols, and a population count instruction (counts the number of 1-bits in the source word), useful in pattern matching applications such as genome mining and handwriting recognition.

Modularity

Nehalem is designed from the start to be modular. In particular, there are two separate sections, the “core” and the “uncore”. The core contains 1, 2, 4, or 8 processor cores and their associated L1 and L2 caches; the uncore contains everything else, including the shared L3 cache, the QPI links, the integrated memory controller, the power control unit, the clock, and (eventually) the integrated graphics processor. The core and uncore use different multipliers of the 133 MHz base clock, so the uncore can run at a lower frequency and use less power, since uncore clock speed has less effect on application performance than core clock speed.

The modular design allows Intel to manufacture processors for specific markets. The number of cores in the “core” section, the number of channels on the integrated memory controller, the type of memory supported, the number of links in the QuickPath interface, the L3 cache size, enabling or disabling of Hyper-Threading, and other features can all be adjusted depending on the performance, power, and

¹⁰ SSE (Streaming SIMD Extensions) instructions allow treating a 128-bit word as 4 single-precision numbers, 2 double-precision numbers, 4 32-bit integers, 8 16-bit short integers, or 16 bytes. The instructions operate on all the sub-entities at once, in effect giving word-level parallelism. Nehalem supports all earlier SSE instructions and registers in addition to the new SSE 4.2 instructions.

reliability requirements of any target market, without any fundamental redesign.

Physical characteristics

The Nehalem family of microprocessors is fabricated using the 45 nm¹¹ high- κ metal-gate¹² process introduced in the Penryn family of the Core microarchitecture. All cores are on a single die. The quad-core processors contain 731 million transistors, many of them used for the 8MB L3 cache¹³. The die area is 263 sq mm, and the size of the physical package is 45x42.5 mm.

Nehalem is packaged in a 1366-pin LGA (land grid array) interface. In the LGA interface, there are no pins on the chip, only gold-plated copper pads that touch pins in an LGA1366 socket on the motherboard. This design allows higher pin densities, hence more power contacts, leading to a more stable power supply. Intel has used LGA sockets since 2004, and for Xeon processors since 2006.

Chipsets and motherboards

Dual-socket motherboards for Nehalem use the Intel 5520 or 5500 Chipset¹⁴, which communicates with the processor using two 20-lane bidirectional QuickPath Interconnect interfaces. The northbridge chip is the Intel 5520 or 5500 Chipset IOH (I/O hub). The Intel 5520 Chipset IOH has 36 PCI Express Gen 2 lanes for graphics (two x16 links and one x4 link, which can be reconfigured). The x16 links each support up to 8 GB/s/direction. The Intel 5500 Chipset IOH has 24 PCI Express Gen 2 lanes (one x16 link and two x4 links, also reconfigurable). The northbridge lacks the usual memory interface (front side bus), since that function is assumed by the integrated memory controller in the Nehalem microarchitecture. Communication to the southbridge I/O controller hub is through one x4 ESI (Enterprise South Bridge Interface) link interface that supports up to 2.5 GB/s transfer rate.

The southbridge chip, the ICH10¹⁵ (I/O controller hub), has 12 USB 2.0 ports (480 MB/s each), 6 PCI Express 1.0 lanes (500 MB/s each), one x4 link (reconfigurable) and two x1 links, Gigabit networking support, HD audio, and 6 SATA channels (3 GB/s each).

Single-socket motherboards for Nehalem (for example, for Xeon W3500 series or Core i7 processors) use Intel's X58 Express Chipset¹⁶. The northbridge chip is the X58 IOH (I/O hub), which like the Intel 5520 Chipset IOH, has 36 PCI Express lanes. Communication between the northbridge and the southbridge is through 2 GB/s DMI (direct media interface). The southbridge chip is the same ICH10 used in the Intel 5520 Chipset.

Nehalem EX coming in late 2009

Nehalem EX¹⁷ (Expandable), the basis for the upcoming Xeon 7500 sequence of microprocessors, is the 8-core/16-thread version of Nehalem for 4- or 8-socket servers. It was first demonstrated in May 2009. Intel says production should begin in the second half of 2009, and the processors will be available in systems in 2010. The 8-socket version will allow 128 simultaneous threads. The design will support up to

11 45 nm is the nominal minimum feature size, or critical dimension, of designs using this process.

12 The high- κ metal-gate process replaces silicon dioxide with a hafnium-based gate dielectric material (not identified by Intel) that has a high dielectric constant κ . The purpose is to reduce leakage currents and improve reliability. The effect is to allow further miniaturization beyond the limits imposed by the properties of silicon dioxide. Hafnium is a “transition metal”, like titanium and especially like zirconium.

13 The SRAM memory in Nehalem's L3 cache uses 6 transistors per cell.

14 Code name “Tylersburg”

15 Or ICH10R with RAID support.

16 Also with code name “Tylersburg”

17 Code name “Beckton”

16 DIMMs per socket.

Nehalem EX will include Intel Hyper-Threading, Intel Turbo Boost, a 24 MB shared L3 cache, an integrated UDDR3 memory controller, and 4 high-bandwidth QPI links. The processors will contain 2.3 billion transistors. Nehalem EX will include MCA (machine check architecture) recovery, an RAS (reliability, availability, and serviceability) feature currently found in Intel's Itanium processors. MCA error correction will allow recovery from previously fatal CPU, memory, and I/O errors.

Westmere coming in 2010

Westmere is the next tick in the Intel's tick-tock microarchitecture, a die-shrink of the Nehalem processor core to a 32 nm process. The integrated memory controller and the new integrated graphics controller will remain on the 45 nm process, as a separate chip in the same package. The processor will have 6 cores, allowing 12 threads. Westmere will include 7 new instructions for AES encryption and decryption, speeding up some operations by a factor of 3. Westmere is scheduled to appear first as a high-end desktop unit, then in other configurations including servers in early 2010.

The bottom line for the user

Intel has developed a remarkable new microarchitecture in Nehalem. They have made a large number of incremental improvements, fixed some issues (like the front side bus bottleneck) that have arisen through the years, and introduced a few application-specific features like the new SSE 4.2 instructions. The improvements work together, with symmetric multithreading – Hyper-Threading – as a common basis. The improvements are balanced – no feature has advanced beyond the others to produce a new bottleneck situation. The design is extremely flexible, so the balance should remain, even after the migration to a 32 nm process. The design is mostly scalable, to more cores and more memory. And Intel achieved all of this with no increase in power consumption.

Where will you see improvements in your applications? The most useful changes from a user's point of view are the large L3 cache, the integrated memory controller providing increased memory bandwidth and reduced latency, and Hyper-Threading. Integrated power management produces indirect benefits as well. Specifically, if your multi-threaded application is well designed, Hyper-Threading should give 10 to 40 percent more performance by keeping all the cores busy. If your application is not threaded or poorly designed, the power management features should provide the same performance, or even slightly improved performance because of Turbo Boost, with lower power consumption.

References

Information in this white paper was gathered from a variety of online sources, including articles and white papers by Intel (intel.com), articles in Wikipedia (wikipedia.com), and articles at Real World Technology (realworldtech.com), X-bit Laboratories (xbitlabs.com), Register Hardware (reghardware.co.uk), Ars Technica (arstechnica.com), AnandTech (anandtech.com), bit-tech.net, and PC Perspective (pcper.com)

Especially helpful were:

“Inside Nehalem: Intel's Future Processor and System” by David Kanter at realworldtech.com, where much more low level detail is available;

“First Look at Nehalem Microarchitecture” by Ilya Gavrichenkov at xbitlabs.com.