April 2011

WHITE PAPER

CHOOSING THE RIGHT CPU CONFIGURATION FOR DELL PRECISION WORKSTATIONS

Mitch Markow, Director of Core Technologies

Selecting the right processors is a key consideration when optimizing Dell Precision™ workstations for your business needs. Although memory capacity and graphics hardware can be more important for certain workloads, processor performance often determines how quickly work gets done.

Through partnerships with industry leaders, Dell offers a variety of central processing unit (CPU) configuration choices for tower, rack-mount and mobile workstations. Many Dell Precision workstations models use the same Intel® Xeon® processors found in high-performance servers, for improved reliability and performance over desktop-class systems.

Considerations when selecting a CPU platform

The first questions that typically come to mind when selecting a CPU are performance and cost. Technical considerations include 1 socket (1S) or 2 socket (2S), number of cores, processor frequency and number of memory channels. While these are indeed important considerations, they cannot be answered outside of a business context.

When choosing a processor, it's necessary to first determine what's important to your business, because the processor requirements depend on what you're trying to do. Understanding the degree to which your applications and workload patterns are dependent on processor performance is a key prerequisite in selecting the right processor.

1 Socket vs. 2 Socket in desktop (DT) systems

The major benefit of a 2 socket over a 1 socket is the number of cores and memory capacity/bandwidth available to one or several concurrent applications. The 2S design has the capability to add a second processor through a coherent bus called the Quick Path Interconnect in Intel's designs. The speed of this bus (usually measured in GT/s) is very important, since it is the key factor in the determining the bandwidth of the interconnect and latency between the processors and the associated memory, which is critical in a NUMA (non-uniform memory access) configured system.

Make sure your application can scale with more cores or the improved memory bandwidth and footprint, because there is a cost and power penalty associated with 2S designs.

Core count, SMT and application threading

The number of cores you need depends on two major factors:

1. How parallel your workload/code is (again, understanding that your workload is critical)
2. How many jobs you would like to run concurrently

In general, if your workload is highly parallel, performance scales well with physical and logical cores via what Intel calls Hyper-Threading Technology (HTT). Here's how it works: Some processors are capable of simultaneous multithreading (SMT), which means they are able to issue multiple instructions from multiple threads in one cycle. HTT provides a basic two-thread SMT engine that enables two threads to run concurrently on a single core.

Note that on some single- and dual-threaded workloads, SMT actually causes performance to degrade slightly due to the overhead of the algorithm. Dell recommends turning off SMT in these cases.

As core count per processor continues to rise and technologies like HTT are introduced, Dell Precision customers running AutoCAD or other design applications that do not scale with higher core counts may not require a dual-socket configuration. For these customers, Dell offers the Precision T1500 (up to four cores) and the T3500 (up to six cores) one socket workstations.

Understanding how well threaded or parallel an application has been written is necessary to determine the value of core count and SMT. Note the Dell Workstation Advisor was developed to help customers match processors (and other platform components) to many of today's top applications. The Dell ISV's can also provide valuable information on their workload hardware and software sensitivities.

Processor frequency and Intel Turbo Boost

Processor frequency is still a dominant factor in performance for many workstation applications, especially for data-intensive, single- and dual-threaded workloads where performance does not improve significantly with the higher core count processors. A frequently cited dilemma with newer processors is that the frequency increase has been tapering off as the core counts have increased.

Processor vendors have responded by offering higher frequency, lower core count processors and new modes of operation. Intel also introduced its Intel Turbo Boost technology, which enables an algorithm to clock the processor above the published frequency, as long as the thermals and power are below certain set points. Pay attention to the turbo frequencies, because when selecting processors, the frequencies increase as you go up the stack.

Remember, however, that while higher processor frequency typically means faster time to answer, bottlenecks can occur elsewhere in the system.

Cache size matters

Larger cache size is another key feature to consider when selecting a processor. Processor cache is the on-die memory in the processor complex and can be accessed much faster than system memory (DRAM), and in most cases larger cache provides a nice performance bump for workstation workloads.

Cache is part of the memory hierarchy that is fastest and closest to the cores. There are typically three levels of cache (L1, L2 and L3) in workstation-class processors. The processor vendors spend significant research effort optimizing the size and data movement between the three levels and main memory. Typically, L1 cache is smallest and fastest, and L3 is the largest and slowest of the three; however, the L3 cache is still 6.6 times faster than DRAM.

Likewise, for workstations apps that perform better with a larger amount of system memory, it's important to select a processor that has enough channels to accommodate a larger memory footprint with lower latency. See Bill Sauber's white paper on memory in this series.

Thermal Design Point (TDP), average power and idle power

Processors typically are offered in several power levels (150W, 130W, 95W, 35W, etc.) called Thermal Design Points (TDP), which is the maximum power that will be dissipated by that model of CPU and used to size the thermal solution. Typically, the highest frequency processors also have the highest TDP in a given processor family and should be used where maximum performance is required. If maximum performance/watt is desired for a desktop workstation, then the faster processors in the 80-95W range should be evaluated for a given workload. Note that TDP is not an accurate proxy for idle or average power.

It's important to understand that the average power used by the CPU can be significantly less then TDP and depends on how CPU intensive the workload is on a running basis. Clearly understanding your workload and sizing your CPU (cores, cache, memory, frequency, power, etc.) accordingly will allow you to optimize your hardware.

Note that the CPU vendors continue to drive the average and idle power down through clock and power gating, as well as by adopting more aggressive low-power states, especially on mobile processors where battery life is critical.

Data integrity and Intel Core™ vs. Intel Xeon

Intel Xeon processors are specifically designed for applications where data integrity with high performance and reliability is a priority. The most common additional feature of a Xeon CPU is ECC (error correcting code) in the memory controller. ECC allows data that is being read or transmitted to be checked for errors and, when necessary, corrected on the fly.

Selecting processors with ECC is important for data integrity to reduce the probability of data corruption for sensitive workloads. This is another differentiating factor between workstation-class machines and those designed for less critical computing tasks. Xeon processors also have more RAS (reliability, availability, serviceability) features, which allow for the processor to better survive an error in the CPU/memory complex.

Over-clocking

Over-clocking is cool, but can be dangerous system reliability or data errors are of concern. Over-clocking is available in some processors (not Xeon) and allows users to up the CPU clock and memory frequencies for improved performance. However, the over-clocked frequencies are outside the reliability limits set by the processor vendors and almost always negate the warranty once a processor has been over-clocked. Due to the reliability concerns and possible loss of data, we do not recommend over-clocking in the workstation space.

Consider total cost of ownership

When selecting a CPU configuration, customers should consider total cost of ownership (TCO), which involves asking the question, 'What's good enough for what we're doing?"

While some customers may require a workstation designed for maximum scalability and performance on large data sets and complex multi-threaded applications, others may find that a one-socket system with ECC will meet all of their needs at a lower price point.

Performance workload sensitivities and analysis

Let's walk through a typical comparison of processor performance across the different offerings of the Intel Nehalem family mobile processor line and one older generation Penryn mobile processor with the front side bus architecture. The Nehalem mobile architecture integrated the Northbridge (memory controller) and high speed PCIe links into the processor, alleviating the memory bandwidth bottleneck of the older FSB. Note that we used similar memory and system configurations for this example.

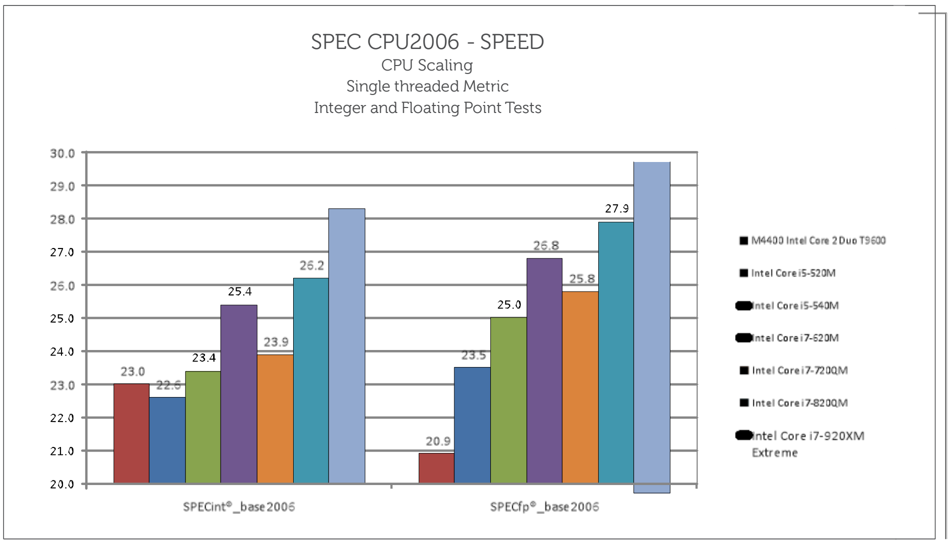
We will use the SPEC CPU 2006 (http://www.spec.org/cpu2006) benchmark for the analysis. The CPU 2006 benchmark is has an integer test suite (CINT) and a floating point test suite (CFP), where each suite is made up of several tests and can be found at the SPEC website.

CPU 2006 can be run in two thread configurations: single-threaded (spped) or multi-threaded (rate). The first step is to compare the key features of the processors, including cores, frequencies, cache and architecture (FSB, memory channels, etc.)

Processor	Cores	Cache MB	Base Freq GHz	Max Freq w/ Turbo	Architecture	Comment
Core 2 T9600	2	6	2.8	2.8	Penryn	FSB based architecture
Core i5-520M	2	3	2.4	2.93	Nehalem	2C with small cache
Core i5-540M	2	3	2.53	3.06	Nehalem	2C with small cache
Core i7-620M	2	4	2.66	3.33	Nehalem	2C with larger cache
Core i7-720M	4	6	1.6	2.80	Nehalem	4C with smaller cache
Core i7-820M	4	8	1.73	3.06	Nehalem	4C with larger cache
Core i7-920xM	4	8	2.0	3.20	Nehalem	4C with larger cache

Key feature observations:

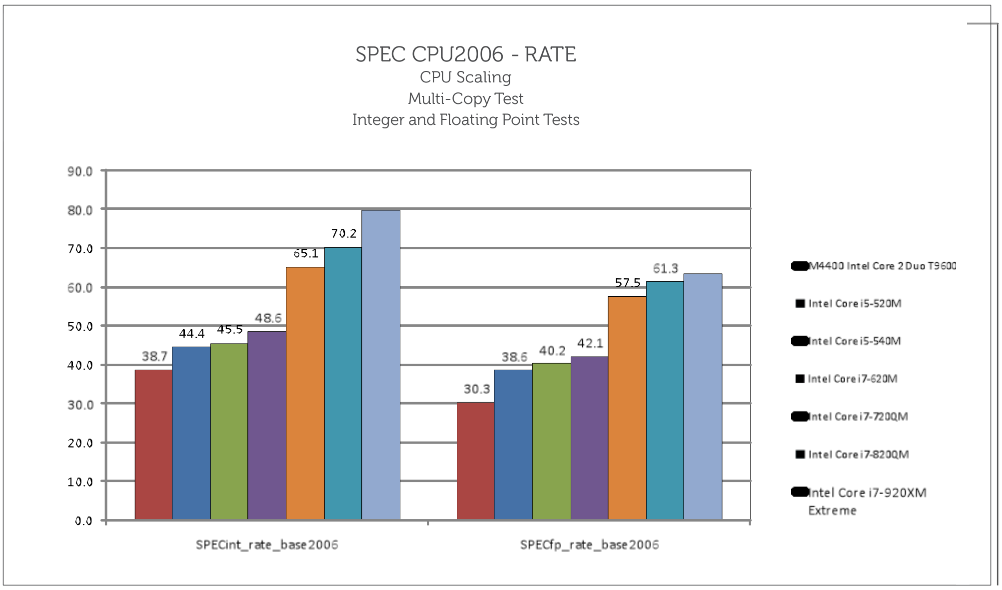
1. The Penryn T9600 has the highest base frequency at 2.8 GHz, but does not have the turbo feature and has the older processor architecture with a front side bus.
2. The Core i5 and i7 series are all based on the Nehalem architecture. Intel differentiates the models using frequency, number of cores and cache.
3. The 520M has half the cache and lower base frequency as the older T9600, but has a higher max frequency at 2.93 GHz.
4. Note that the entry 4C (720M) has the lowest base frequency of all the processors in the chart—not a good pick if your workload is frequency sensitive and single threaded.
5. The Core i7-920XM Extreme has the most cores, most cache and highest max frequency, but also carries the highest price.



We ran the CPU 2006 benchmark on two Dell Precision mobile workstations, one with the Nehalem-based processors and one with the Penryn-based processor. The results and color commentary are consolidated into a CPU 2006 Base chart and a CPU 2006 Rate chart below. The data is further segmented into SPECint (integer) results and SPECfp (floating point) results.

The SPEC CPU speed metric roughly correlates to lightly threaded or single-threaded applications. Examples are CAD applications such as CATIA, NX, SolidWorks, Solid Edge, AutoCAD and other applications. It also correlates to DCC editing and software development. Note that base and turbo frequency are more important in this workload than number of cores.

The 2C Core i7-620M shows a significant gain from the 2C Core i5-540M due to its larger cache and its 2 bin jump in max frequency; note that it also beats the slower 4C 720QM. The same pattern occurs for 720QM -> 820QM (quad-core).



The SPEC CPU rate metric roughly correlates with highly threaded applications like simulation and analysis software such as Nastran, ANSYS and other applications such as DCC rendering and trans-coding applications. The numbers of cores has a higher impact on the performance of this workload than higher CPU frequency. Note that Intel Hyper-Threading was on for all tests. As expected, the quad-core processors outperformed the dual-core processors across the board.

Customers are encouraged to review benchmark testing results from Standard Performance Evaluation Corporation (SPEC) while assessing their CPU requirements for Dell Precision workstations: www.spec.org