DESIGNING RESEARCH COMPUTING SOLUTIONS FOR THE CERN/ATLAS PROGRAM

By Jacob Liberman
Walker Stemple

The ATLAS experiment at the European Organization for Nuclear Research (CERN) Large Hadron Collider represents a major effort to uncover the elementary building blocks of matter and energy. This article describes extensive optimization tests and best-practice recommendations for configuring 11th-generation Dell™ PowerEdge™ servers to help process the enormous amount of data expected from this experiment.

The European Organization for Nuclear Research (CERN) hosts the Large Hadron Collider (LHC)—the world's largest particle collider—on the border between France and Switzerland near Geneva. The ATLAS experiment is the largest of six LHC experiments utilizing proton-proton collisions to uncover the elementary building blocks of matter and energy. Hadrons, or protons, are accelerated to close to the speed of light by superconducting magnets and then carefully steered into an identical oncoming beam within the ATLAS detector. Researchers hope to discover new particles hypothesized to exist, like the Higgs boson, to help complete the standard model of particle physics.

This pure science experiment creates unprecedented challenges in data storage and computing. With a full trigger data rate of 780 MB/sec, ATLAS is expected to generate up to 15 PB of raw detector data per year. The research institutions responsible for organizing and examining the data are organized into a three-tiered worldwide computing grid known as the Worldwide LHC Computing Grid (WLCG). The tier-0 CERN analysis facility selects only the most interesting events at a rate of 320 MB/sec. This data is sent for redundant storage and processing at 10 tier-1 facilities worldwide. The server computational requirement for the processing of this data is estimated to be 1.7 million SI2K, where SI2K is the Standard Performance Evaluation Council (SPEC) SPECint2000 score from the CPU2000 benchmark—which amounts to the computational capability of thousands of servers.

This prestigious scientific endeavor is supported by thousands of scientists at more than 200 institutions and universities in close to 60 countries worldwide. Dell is committed to helping these researchers meet their ATLAS computing challenges by providing advanced computational, interconnect, and storage technologies through key partnerships with industry leaders. A specialized Dell team of industry experts is dedicated to working with researchers and industry partners to advance the goals of the ATLAS experiment. This article describes the collaboration between Dell and CERN researchers to design computing solutions that can process the enormous amount of data generated by the LHC. It includes results from extensive performance optimization tests and shares best-practice recommendations for configuring 11th-generation Dell PowerEdge servers for use as part of the ATLAS experiment.

HEP-SPEC BENCHMARK TESTS

ATLAS researchers developed the High-Energy Physics (HEP)–SPEC benchmark suite to standardize server performance evaluation for their experiments. HEP-SPEC is based on the standard SPEC CPU2006 benchmarking suite. CPU2006 is widely used to rate processor performance for two reasons. First, because it comprises real application benchmarks, CPU2006 results can accurately reflect performance across a broad range of server workloads. Second, the benchmark run rules allow vendors to implement BIOS and compiler optimizations that showcase their systems' peak performance.

HEP-SPEC uses the all_cpp subset of CPU2006 with specific compiler versions and optimization flags. Because HEP-SPEC performance correlates strongly with ATLAS application performance, it can provide a consistent and repeatable benchmark to describe experiment requirements and existing resources. The all_cpp subset was selected over the base benchmark because it matches the scaling behavior and floating-point-to-integer ratio of the production physics codes. Furthermore, HEP-SPEC is typically compiled in 32-bit mode using the open source GNU Compiler Collection (GCC), helping ensure backward compatibility with sites that lack 64-bit resources or commercial compilers.

In May 2009, the Dell HPC engineering team studied HEP-SPEC performance in support of the Dell CERN/ATLAS program. This study had two goals. The first goal was to understand how server subsystem performance—such as processor capabilities or memory bandwidth—affects HEP-SPEC performance, to help the Dell team make informed recommendations on how CERN/ATLAS researchers can maximize their research investment. The second goal was to understand how the BIOS features in 11th-generation Dell PowerEdge servers affect HEP-SPEC performance. These servers introduced multiple BIOS features designed to boost performance and energy efficiency, and understanding how these features affect HEP-SPEC performance helps the Dell team recommend optimal BIOS settings that can accelerate research and decrease time to solution.

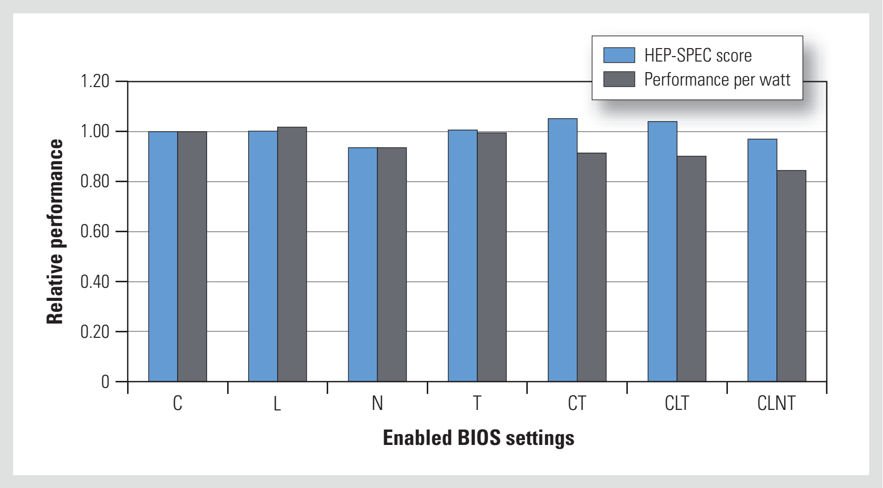

Figure 1. *HEP-SPEC performance and efficiency relative to a system with all BIOS settings disabled*

TEST RESULTS: PERFORMANCE AND ENERGY EFFICIENCY

The 11th-generation Dell PowerEdge server family is based on the Intel® Xeon® processor 5500 series microarchitecture, and introduced architectural enhancements and BIOS features designed to increase performance and enhance energy efficiency. The Dell HPC engineering team evaluated the impact of the following BIOS settings on HEP-SPEC performance in 11th-generation PowerEdge servers in a variety of configurations:

- **C-states:** Allows the system BIOS to throttle power to individual processor cores based on need, which can enhance energy efficiency
- **Logical processor (formerly called Intel Hyper-Threading Technology):** Improves thread-level parallelism by sharing the same physical core between multiple threads, which can increase performance for some codes
- **Node interleaving:** Creates uniform memory access speed by interleaving memory across both processor sockets, which can help increase performance for codes that require a large global memory address space
- **Turbo mode:** Increases processor clock rate by 1–3 increments of 133 MHz if there is available system power and heat headroom
- **Power management profile:** Allows the OS or system BIOS to control power to the processor sockets, memory, and fans, which can help increase performance or improve energy efficiency; all servers in this study used the Max Performance profile[1]

Figure 1 shows the normalized HEP-SPEC performance and performance per watt of different combinations of BIOS settings in a PowerEdge R610 server relative to the same server with all settings disabled; "C" represents the C-states setting, "L" represents the logical processor setting, "N" represents the node interleaving setting, and "T" represents the turbo mode setting. Performance per watt was calculated by dividing HEP-SPEC performance by average power consumed (in watts) during the run. A higher score is better for both measures.

As these results show, enabling both the C-states and turbo mode settings increased HEP-SPEC performance by 5 percent compared with having all settings disabled; enabling these settings individually, however, did not increase performance. The combination of the C-states and turbo mode settings also reduced energy efficiency by approximately 10 percent—meaning that the performance gain is outweighed by a

[1] For more information on these settings and how they can affect HPC performance, see "Optimal BIOS Settings for High Performance Computing with PowerEdge 11G Servers," by Jacob Liberman and Garima Kochhar, Dell Product Group, July 2009, i.dell.com/sites/content/business/solutions/whitepapers/en/documents/HPC_Dell_11G_BIOS_Options.pdf.

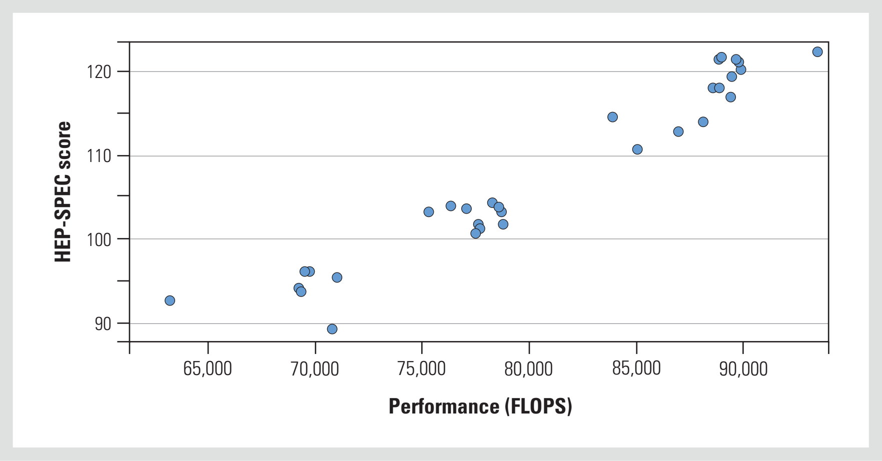

Figure 2. *Plotted relationship between FLOPS and HEP-SPEC score*

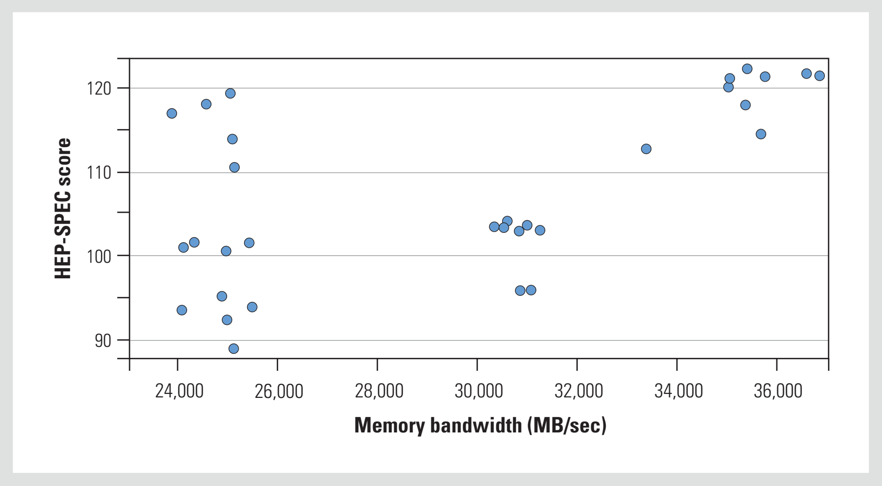

Figure 3. *Plotted relationship between memory bandwidth and HEP-SPEC score*

proportionate increase in power consumption. The other tested settings had no impact on HEP-SPEC performance except for node interleaving, which reduced performance even when the C-states and turbo mode settings were also enabled. Enabling the logical processor setting can help increase HEP-SPEC results when running 16 threads, but HEP-SPEC run rules often stipulate that this setting be disabled because it has traditionally not been used with CERN/ATLAS production codes.

After having identified the optimal BIOS settings, the Dell team next focused on identifying system components that could help maximize performance and energy efficiency. CERN/ATLAS research computing resources fall into two categories: those that are dedicated exclusively to processing ATLAS data, and those that conduct general-purpose research computing in addition to ATLAS data processing. During this second phase of the study, the Dell team evaluated HEP-SPEC performance across approximately 70 processor and memory combinations.

The results reveal several important facts about HEP-SPEC. First, HEP-SPEC is a processor-bound workload, meaning that faster processors translate into faster time to results. Figure 2 shows the strong correlation between floating-point operations per second (FLOPS)—a measure of the rate at which processors can solve floating-point calculations—and HEP-SPEC performance in the tested servers, which increases almost linearly with processor capability. In fact, the statistical correlation between these two metrics is 97 percent, meaning that the HEP-SPEC result can be predicted with 97 percent accuracy based on the measured FLOPS result alone. Figure 3, in contrast, plots the relationship between memory bandwidth and HEP-SPEC performance. The statistical correlation between these two metrics is much weaker than the correlation shown in Figure 2, indicating that increased memory performance does not translate directly into an increased HEP-SPEC score. Memory bandwidth and floating-point performance were measured with the STREAM and Double Precision General Matrix Multiply (DGEMM) benchmarks, respectively.

Several important design recommendations can be drawn from these measurements. First, for dedicated HEP-SPEC computing resources, faster processors accelerate data processing more than faster memory. The 11th-generation PowerEdge server family supports dual in-line memory modules (DIMMs) at speeds of 1,066 MHz and 1,333 MHz. The HEP-SPEC performance difference between DIMM speeds is less than 3 percent; therefore, if the cost difference between DIMM speeds is greater than 3 percent, ATLAS researchers generally would benefit more from upgrading their processors than from buying faster DIMMs.

Figure 4 shows the normalized HEP-SPEC performance of a PowerEdge R610 server with different registered DIMM (RDIMM) and unbuffered DIMM (UDIMM) configurations relative to the same server with six 4 GB UDIMMs at 1,066 MHz. The performance differences between DIMM types, speeds, and count per channel are within 3 percent; however, the power consumption differs by up to 12 percent. Therefore, the baseline configuration of six 4 GB UDIMMs at 1,066 MHz can provide an optimal fit for HEP-SPEC in terms of performance and energy efficiency.

DESIGN RECOMMENDATIONS

Based on the results of this study, the Dell HPC engineering team recommends

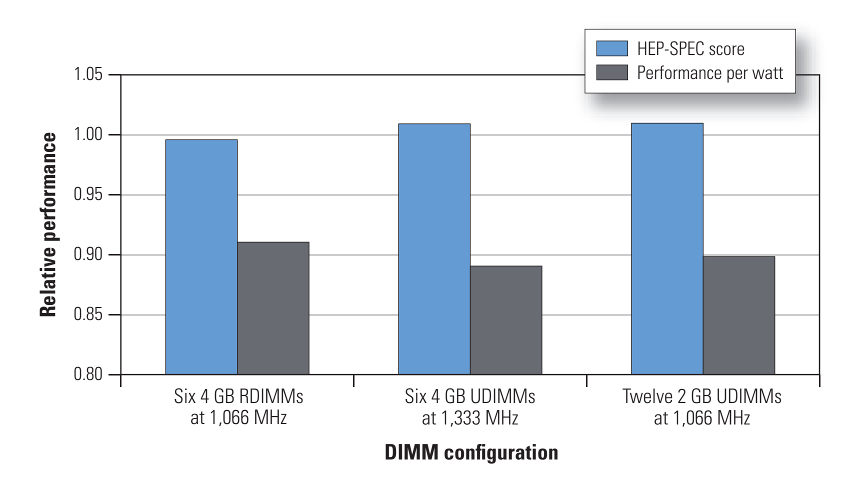

Figure 4. *HEP-SPEC performance and efficiency relative to a system with six 4 GB UDIMMs at 1,066 MHz*

enabling the C-states and turbo mode BIOS settings, disabling the logical processor and node interleaving settings, and using the Max Performance power management profile to help maximize HEP-SPEC performance on 11th-generation Dell PowerEdge servers. These optimized BIOS settings can be configured at the factory before servers are shipped to CERN sites.

Although enabling the logical processor setting did not increase HEP-SPEC performance in these tests, CERN/ATLAS researchers have found that enabling this setting does help increase the performance of their production codes. Therefore, the Dell HPC engineering team also recommends testing whether this feature increases the performance of specific production codes.

The Dell CERN/ATLAS team has also developed several reference architectures based on the results of this study. These configurations are designed to provide optimal configurations to help meet a variety of goals, including energy efficiency, performance, and cost value (see Figure 5). The balanced configuration is designed to be competitive in each of those three categories, while the general-purpose configuration is well suited for general-purpose cluster computing as well as ATLAS data processing.

It is important to keep in mind that HEP-SPEC is not a traditional clustered HPC application, and does not require communication or data coherency across cluster nodes. For this reason, clusters designed to maximize HEP-SPEC throughput may not be ideally suited for clustered HPC applications. Purchasing decisions for general-purpose clusters should not be based solely on HEP-SPEC results; these results should be supplemented with standard cluster benchmarks such as SPEC MPI or HPC Challenge.

From a design standpoint, general-purpose clusters can likely benefit from a low-latency interconnect such as 10 Gigabit Ethernet or InfiniBand, and most clustered HPC applications can also benefit from increased memory bandwidth. Therefore, when designing a general-purpose cluster that occasionally processes HEP-SPEC data, a balanced memory configuration and increased DIMM speeds could help increase performance.

DELL AND ATLAS

ATLAS institutions that are currently utilizing 10th-generation Dell PowerEdge servers with the Intel Xeon processor 5400 series can expect a significant boost in HEP-SPEC performance when switching to 11th-generation PowerEdge servers with the Intel Xeon processor 5500 series. Dell continues to monitor emerging technologies for applicability to high-energy physics experiments like ATLAS. Anticipated directions for future study include evaluating the performance and energy efficiency of production ATLAS codes on 11th-generation PowerEdge servers.

Jacob Liberman is a development engineer in the Scalable Systems Group at Dell.

Walker Stemple is a business development manager on the CERN/ATLAS team at Dell.

	Server model	Processor model	DIMM configuration
Energy efficiency	Dell PowerEdge M610	Intel Xeon L5520	Four 4 GB UDIMMs at 1,066 MHz
Performance	Dell PowerEdge R610	Intel Xeon X5570	Six 4 GB RDIMMs at 1,333 MHz
Balanced	Dell PowerEdge R410	Intel Xeon E5540	Six 4 GB UDIMMs at 1,066 MHz
Value	Dell PowerEdge R410	Intel Xeon E5520	Four 4 GB UDIMMs at 1,066 MHz
General-purpose	Dell PowerEdge M610	Intel Xeon X5550	Six 4 GB UDIMMs at 1,066 MHz

Figure 5. *Recommended CERN/ATLAS reference configurations to help meet different goals*

MORE
ONLINE
DELL.COM/PowerSolutions

QUICK LINK

Dell and ATLAS:
DELL.COM/ATLAS