Game-changing Extreme GPU computing with The Dell PowerEdge C4130

A Dell Technical White Paper

This white paper describes the system architecture and performance characterization of the PowerEdge C4130. The C4130 offers a highly configurable system design and ultra-high accelerator/coprocessor density. The combination of these factors have resulted in substantial performance improvement in for a number of industry standard HPC benchmarks and applications.

Saeed Iqbal, Mayura Deshmukh & Nishanth Dandapanthula

Dell HPC Engineering

May 2015 | Version 1.3

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

Contents

Figures

1. Introduction

GPU computing is now established and widespread in the HPC community. There is an ever-increasing demand for compute power. This demand has pushed server designs towards higher hardware accelerator density. However, most such designs have a standard system configuration, which may not be optimal for maximum performance across all application classes. The latest high-density design from Dell, the PowerEdge C4130, offers up to four GPU boards in a 1U form factor. In addition, the uniqueness of PowerEdge C4130 is that it offers a configurable system design, potentially making it a better fit, for the wider variety of extreme HPC applications.

In this paper, we describe the new Dell server, C4130; specifically designed for GPU Computing, and demonstrate how a wide class of applications has seen multifold increase in performance on the C4130.

2. The PowerEdge C4130

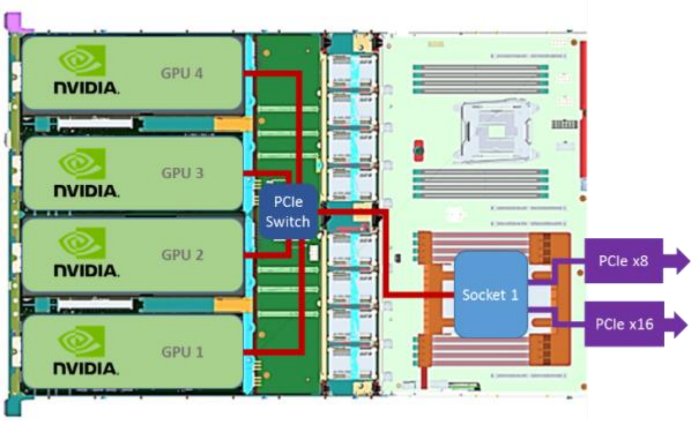
The PowerEdge C4130 is the latest addition to Dell's portfolio to build HPC solutions. The C4130 incorporates several innovations in its system design to enable extreme GPU computing.

Figure 1: The PowerEdge C4130 server

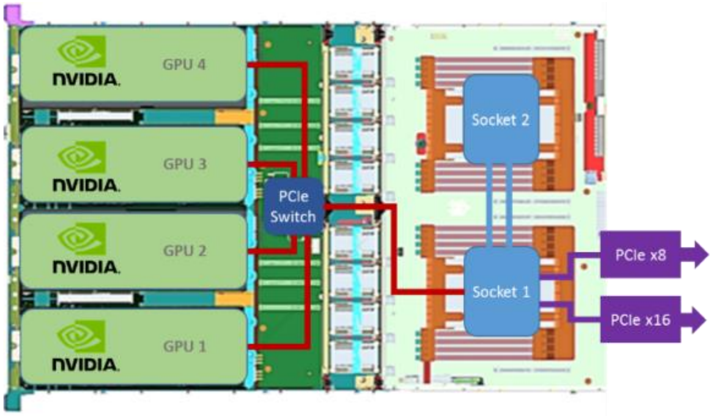
Unique system configurability

The C4130 offers five carefully selected system configurations. The motivation behind offering this configuration is to offer a system design that will best fit the application requirement. This offers users a unique opportunity to match the system design with their application characteristics. The configurations differ in the manner in which accelerators connect to the CPUs, resulting in different accelerator to CPU ratios. They vary from balanced to clustered configurations.

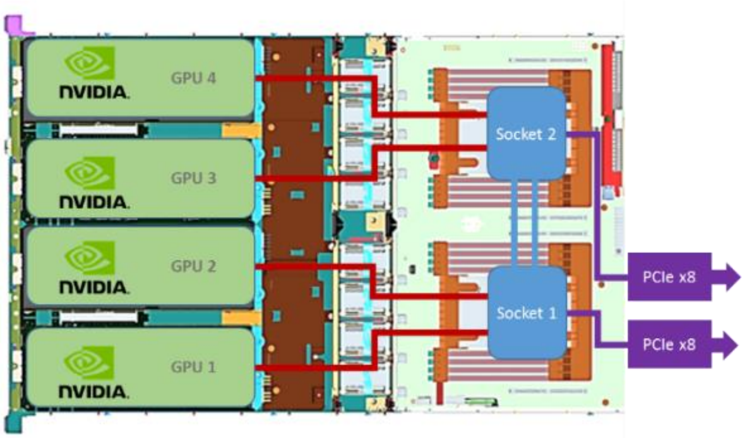
The goal of this white paper is to introduce available server configurations. The C4130 offers five configurations shown in *Figure 2* and *3, A* through *E*. Since GPUs provide the bulk of compute horsepower, the configurations can be divided into groups based on expected performance. The first group of three configurations, *A*, *B* and *C*, has four K80s each. The second group of two configurations, *D* and *E*, has two K80s each. The first two quad GPU configurations have an internal PCIe switch module. The details of the various configurations are shown in the *Table 1*.



Configuration A

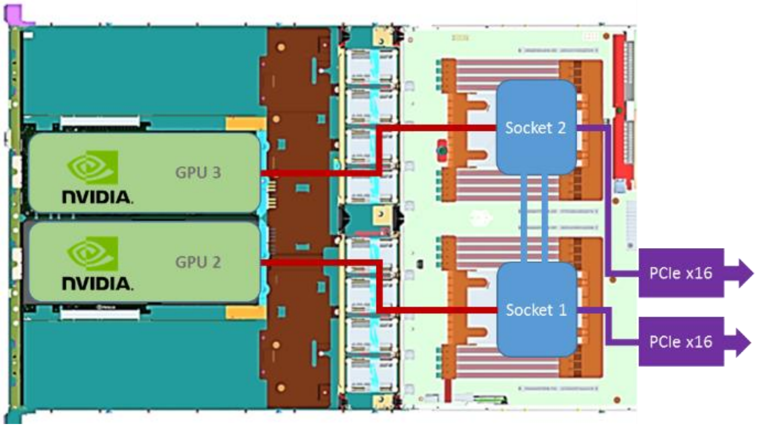


Configuration B

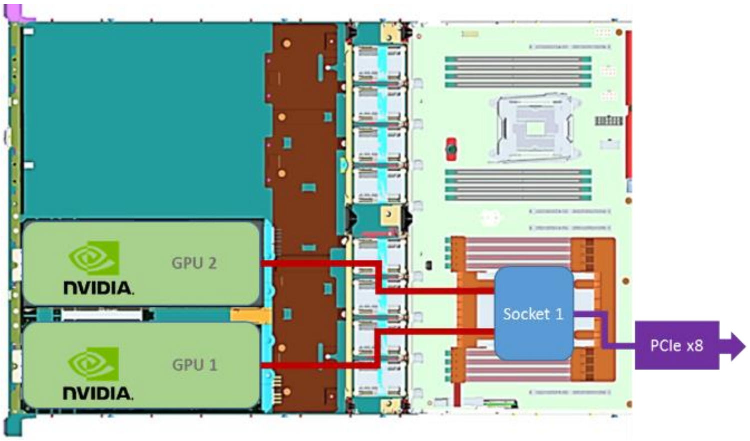


Configuration C

Figure 2: Block diagram of the four GPU board configurations available on the C4130



Configuration D



Configuration E

Figure 3: Block diagram of the two GPU board configurations available on the C4130

Table 1: Characteristics of the C4130 configurations

C4130 Configuration	GPU Boards	CPUs	Switch Module	GPU:CPU ratio	Comments
A	4	1	Y	8:1	Single CPU, optimized for peer to peer communication
B	4	2	Y	8:2	Dual CPUs, optimized for peer to peer communication
C	4	2	N	8:2	Dual CPUs, balanced with four GPU boards
D	2	2	N	4:2	Dual CPUs, balanced with two GPU boards
E	2	1	N	4:2	Single CPU, balanced with two GPU boards

Ultra-high density

The C4130 enables dense GPU computing with up to four accelerators/coprocessors per U. Most current servers in the market offer densities of one or two accelerators/processors per U. The high density combined with configurability proves to be a powerful combination.

Accelerator-friendly layout

The C4130 is a purpose-built server with an accelerator-friendly thermal design, shown in *Figure 4*. The accelerators/coprocessors are loaded in the front of the system. This layout allows incoming cold air to reach the accelerator/coprocessors before it reaches the CPUs. This layout keeps the four accelerators/coprocessors operating at optimal performance. It also helps increase the reliability of the accelerators/coprocessors.

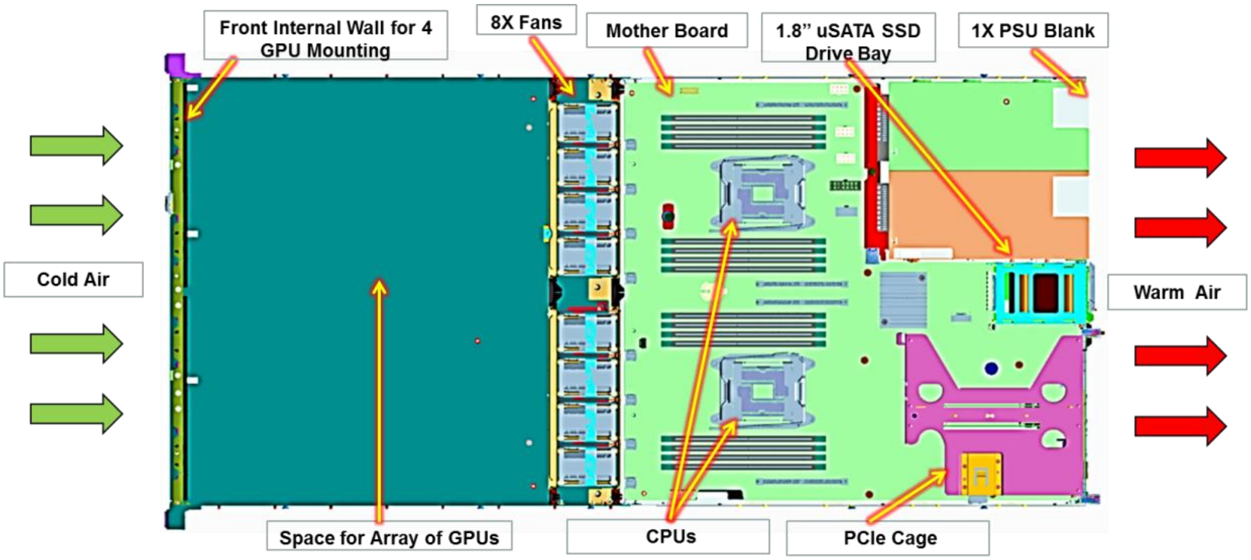


Figure 4: Unique placement of GPUs. Front loading the C4130 with GPU enables cold air to reach the GPUs first, and keeps GPUs at optimal performance level.

Advanced systems management

The C4130 supports the Dell iDRAC 8 for systems management. The Dell Remote Access Controller (iDRAC) with Lifecycle Controller technology, server deployment, configuration and updates are streamlined across the OpenManage portfolio and through integration with third-party management solutions.

3. The Tesla K80 GPU accelerator board

The latest HPC-focused Tesla series General Purpose Graphic Units (GPU) released from NVIDIA is the Tesla K80. From an HPC perspective, the most important improvement is the 1.87 TFLOPs (double precision) compute capacity, which is about 30% more than K40, the previous Tesla card. The K80 auto-boost feature automatically provides additional performance if additional power head room is available. The internal GPUs are based on the GK210 architecture and have 4,992 cores, which represent a 73% improvement over K40. The K80 accelerator board has a total memory of 24GBs, which is equally divided between the two internal GPUs; this is a 100% more memory capacity compared to the K40. The memory bandwidth in K80 has improved to 480 GB/s. The rated power consumption of a single K80 is a maximum of 300 watts.

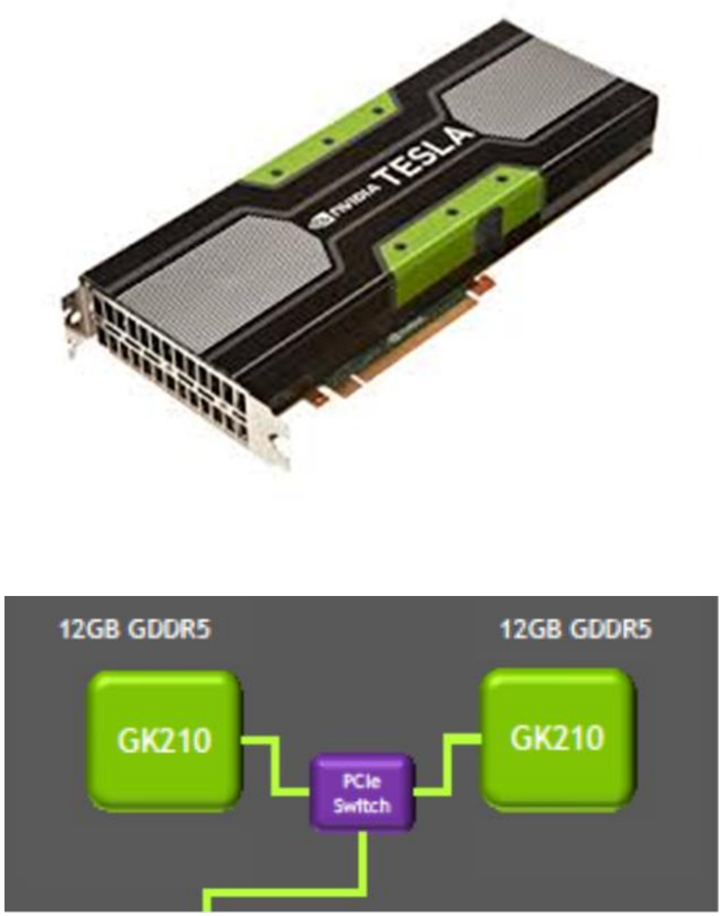


Figure 5: The K80 GPU board with two internal GPUs connected via a PCIe switch

Combining K80s with the latest high GPU density design from Dell, the PowerEdge C4130 provides an extra-ordinarily powerful compute node. The C4130 can have up to four K40 or K80 GPU boards in a 1U form factor. In addition, it is available in several workload-specific configurations, potentially making it a better fit for specific HPC codes.

4. Performance characterization

4.1 Bandwidths between CPU to GPU

We measure the host-to-device (H2D) and device-to-host (D2H) bandwidth of the five C4130 configurations. *Figure 6* and *7* show the measured bandwidths. Two CPUs and eight GPUs (two internal GPUs per K80 board) yield 16 CPU-to-GPU combinations. The CPU (host) and GPU (device) bandwidth measurements for each configuration is shown in figures below. The Host to Device (H2D) and Device to Host measurements are about 12000 MB/s, which are the state-of-art achievable in Gen PCIe, with a peak of 15754 MB/s. It is noteworthy that the measurement is consistent and does not vary significantly with configurations.

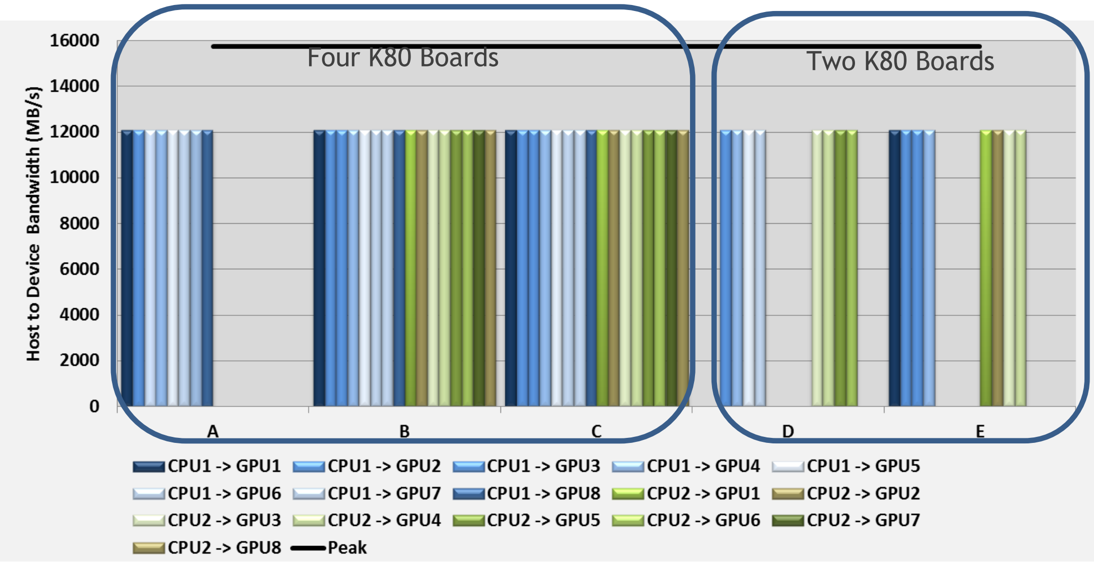


Figure 6: Measured host-to-device (H2D) bandwidths on various configurations

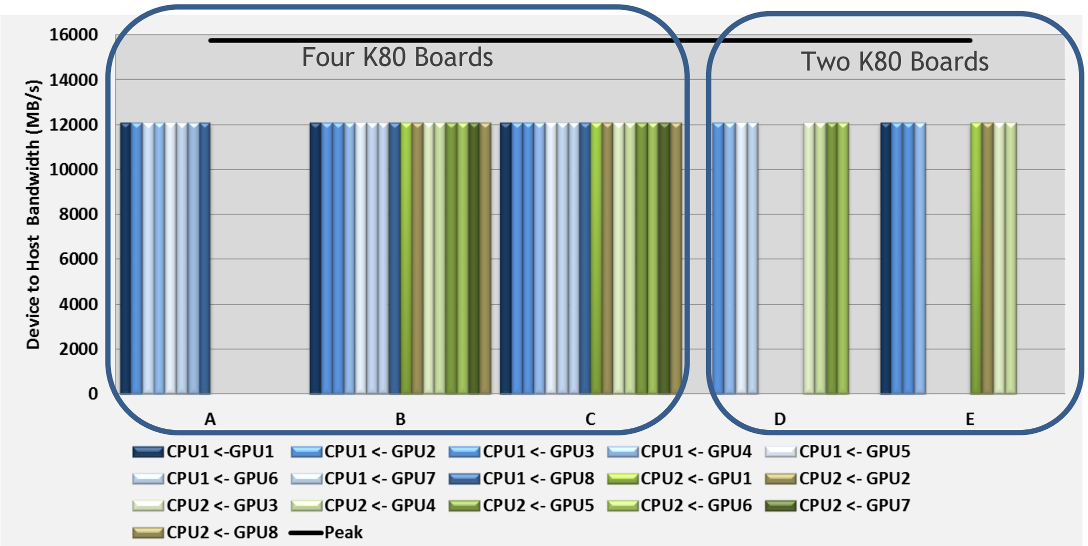


Figure 7: Measured device-to-host (D2H) bandwidth on various configurations

4.2 Accelerating high performance Linpack (HPL)

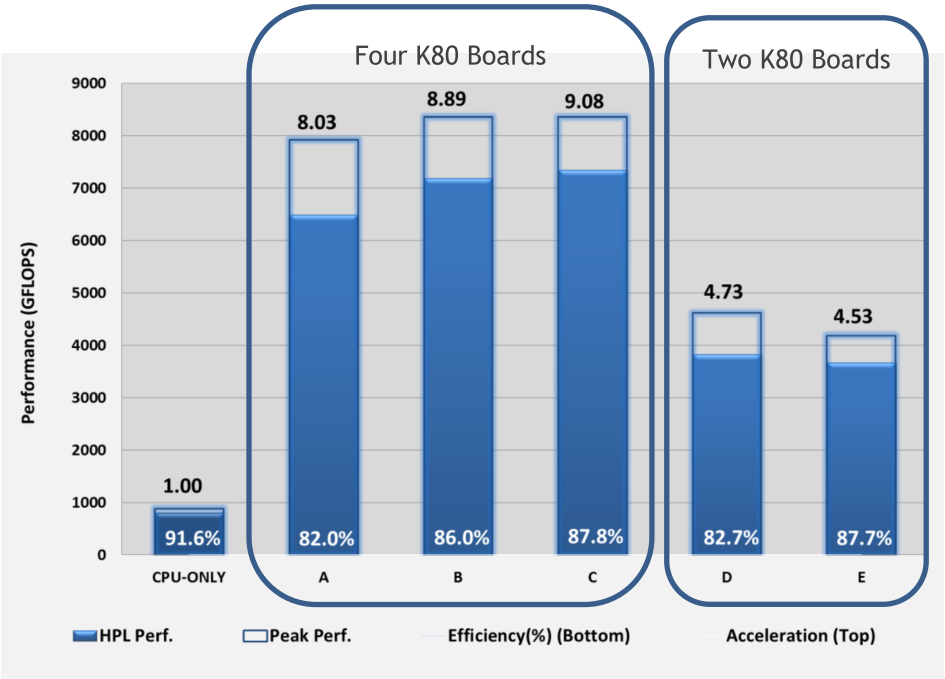


Figure 8: HPL performance, efficiency and acceleration compared to CPU-only

In this section, we evaluate the performance of C4130 with up to four K80 GPU boards on HPL. Given the importance of HPL in comparing HPC computing systems, this section shows key performance characterization data for the C4130. The performance achieved, HPL acceleration, HPL efficiency, power consumption and performance per watt on various system configurations are measured.

Figure 8 shows the HPL performance characterization. Configurations *A*, *B* and *C* are four K80 configurations with performance from **6.5 to 7.3 TFLOPS**. The difference from *A* to *B* is due to the extra CPU in configurations *B*. Overall; the *C* configuration has the highest performance of 7.3 TFLOPS. The difference from *B* to *C* is due to different GPU to CPU ratios; both have the same number of compute resources. Configuration *C* is balanced with two GPUs per CPU while *B* has the all four GPU attached to a single CPU. On the two GPU configurations, ***D* is higher with 3.8 TFLOPS and *E* with 3.6 TFLOPS** — one less CPU in configuration E explains the difference.

Compared to a CPU-only performance, an acceleration of 9X is obtained by using four K80 and an acceleration of 4.7X with two K80 boards. The HPL efficiency is significantly higher on K80 (low to upper 80s) compared to previous generation of GPUs.

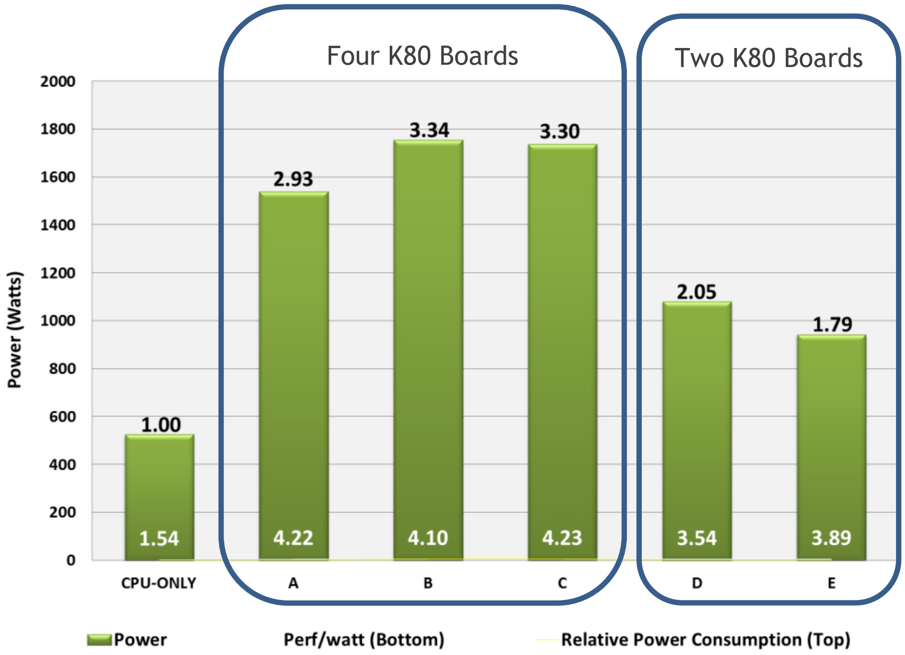


Figure 9: HPL Power, performance/watt and power consumption compared to CPU-only

Figure 9 shows the power consumption data for the HPL runs. In general, GPUs can consume substantial power on compute intensive workloads. As shown above, the power consumption of configurations A, B and C is significantly higher (2.9X to 3.3X), compared to CPU-only runs; this is due to the four K80 GPUs. Power consumption of D and E is lower (1.8X to 2.0X compared to CPU-only runs).

The power efficiency, i.e. the useful work delivered for every watt of power consumed, is in the 4+ GFLOPS/W range for quad GPU configurations and about 1.8X to 2X range for dual GPU configurations. Configuration *C* offers the highest Performance per watt at about 4.23 GFLOPS/W.

Compared to the CPU-only performance per watt of just 1.5 GFLOPS/w, the quad GPU configurations show a 2.7X and dual GPU configurations show a 2.3X improvement in the overall performance/watt.

4.3 Accelerating Molecular Dynamics with NAMD

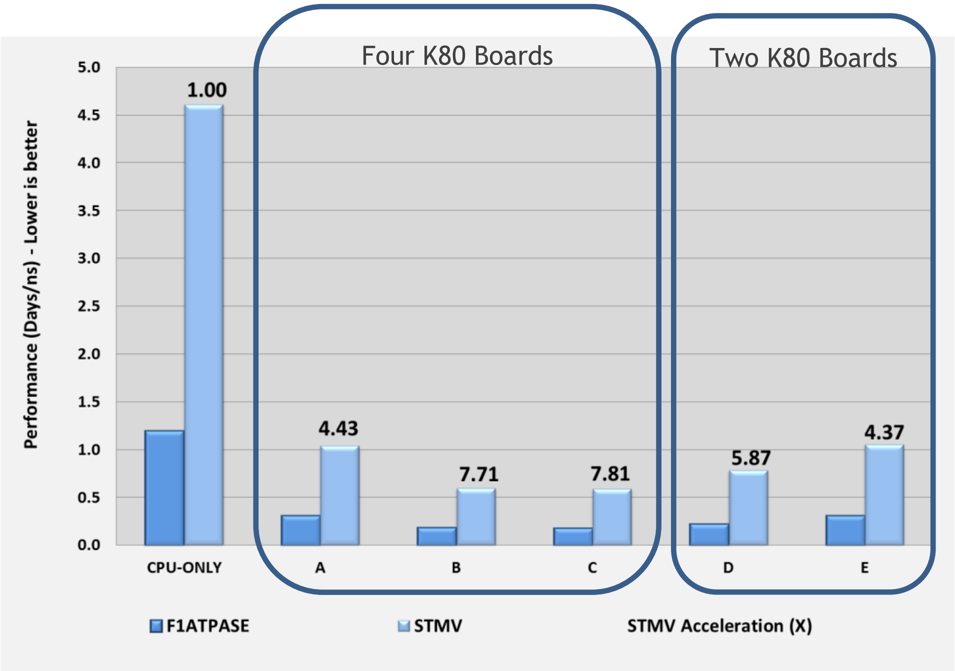


Figure 10: NAMD performance and acceleration compared to CPU-only

The advent of hardware accelerators has influenced Molecular Dynamics by reducing the time to results and therefore providing a tremendous boost in simulation capacity. LAMMPS and GROMACS are two open source Molecular Dynamics (MD) applications that can take advantage of these hardware accelerators, and have seen a tremendous boost in simulation capacity. NAMD is a freely available, feature-rich, GPU-enabled molecular dynamics simulator.

In this section, we evaluate NAMD performance improvement wit, on two proteins F1ATPASE and STMV. These proteins consist of 327K and 1066K atoms respectively. The performance measure is in “days/ns”, which shows the number of days required to simulate one nanosecond of real-time.

Figure 10 shows the performance of NAMD on the PowerEdge C4130. Configurations *A*, *B* and *C* are the four GPU configurations. However the acceleration on NAMD also seems to be sensitive to number of CPUs, e.g., there is a significant difference in the acceleration between *A* and *B*. ***B*** has an additional CPU compared to *A*. Among the three 4 GPU configurations the current version of NAMD performs best on configuration *C*. The difference in two highest performing configurations *C* and *B* is the manner in which GPUs connect to the CPUs. The balanced configuration *C* has two GPUs attached to two CPUs resulting in 7.8X acceleration over the CPU-only runs. The same four GPUs attached via a switch module to a single CPU results in about 7.7X acceleration. On the dual GPU configurations, *D* performs better with 5.9X acceleration compared to 4.4X in configuration *E*. Configuration *D* performs better, which is in line with the assumption that a 2nd CPU is helpful for NAMD, as we observed *B* & *C* (the dual CPU quad GPU configurations) performing best among the four GPU configurations.

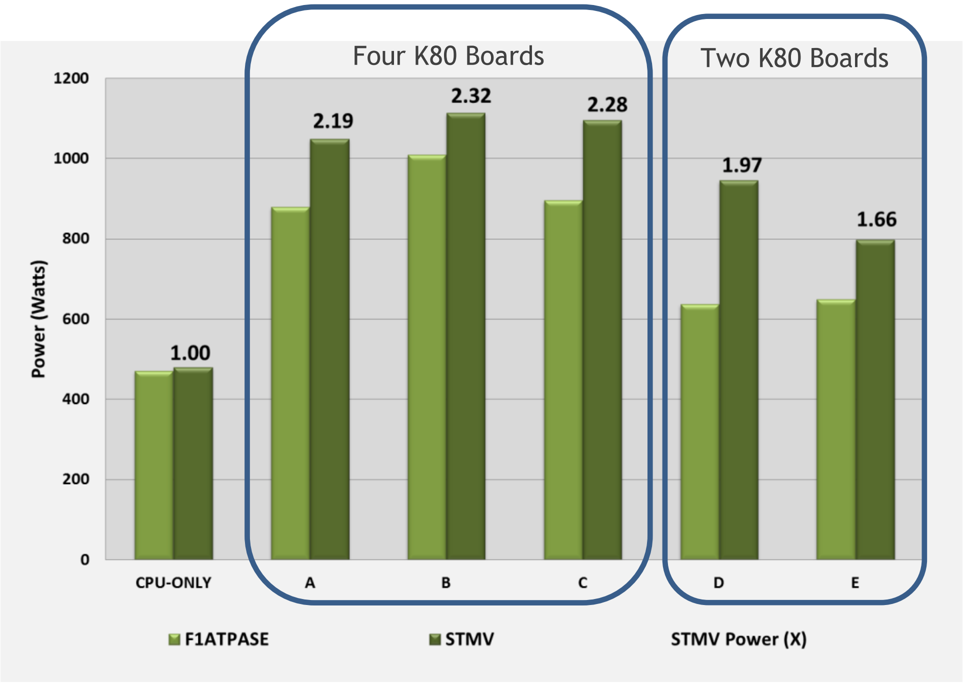


Figure 11: NAMD power consumption and relative power consumption compared to CPU-only

The power consumption and relative power consumption shown in *Figure 11* for GPU configurations is about 2.1X to 2.3X resulting in accelerations from 4.4X to 7.8X. From performance per watt perspective (an acceleration of 7.8X for 2.3X more power), configuration *C* does the best. The power consumption of *D* is higher than *E* due to the additional CPU and improved utilization of GPUs resulting in better acceleration.

As shown in above the K80 GPUs can substantially accelerate NAMD in a power-efficient manner. The balanced configurations seem to do better with NAMD. Configurations *C* and *D* are best for NAMD, but the particular choice depends on required GPU density/U.

4.4 Accelerating Molecular Dynamics with LAMMPS

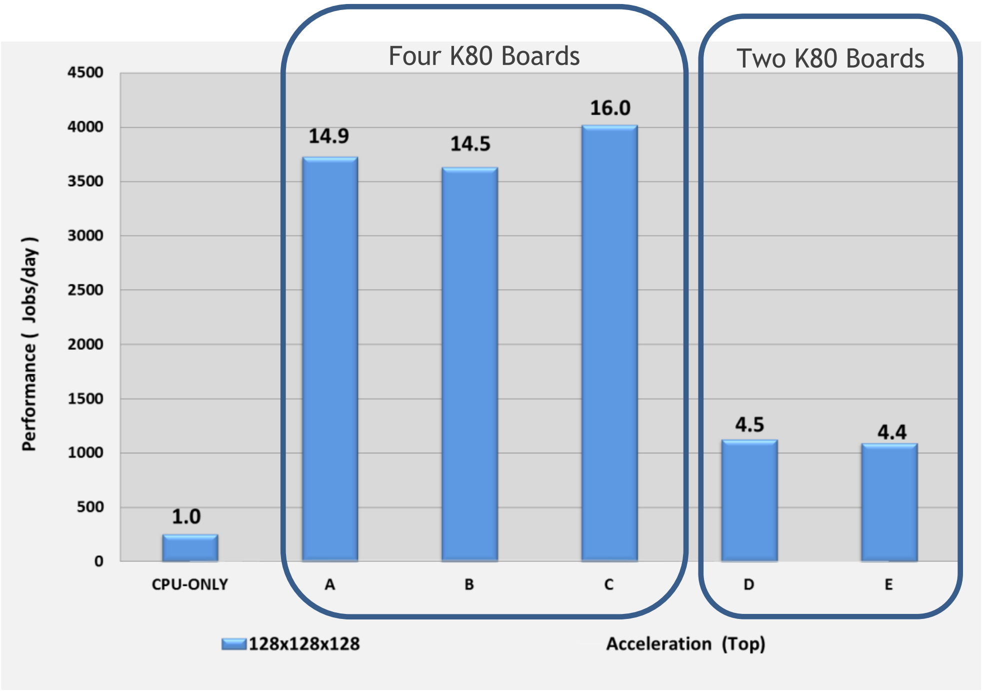


Figure 12: LAMMPS performance and acceleration compared to CPU-only

In this section, we evaluate the performance of second common molecular dynamics code, LAMMPS. LAMMPS stands for "Large-scale Atomic/Molecular Massively Parallel Simulator." LAMMPS is used to model solid-state materials and soft matter. The performance measure is in "Jobs/day" for LAMMPS, and a higher score is better. The benchmark ran on LAMMPS LJ (Lennard-Jones liquid benchmark), it has 8388608 atoms for 1000 steps.

Figure 12 compares the performance of LAMMPS on the five C4130 configurations mentioned in previous sections. As a reference, we also compare to the CPU-only runtimes to quantify the acceleration offered by various configurations on the K80 GPU boards. Configurations *A* and *B* are the two four K80 switched configurations, with the only difference being that B has an extra CPU. Since, at this time LAMMPS just uses the GPU cores for actual compute intensive calculations, the extra CPU does not increase the performance substantially. Configuration *C* is the balanced four-GPU non-switched configuration. Configuration *C* performs better than *A* and *B*. This is partly due to the PCIe switch in configurations *A* and *B* that introduces one extra hop during communications, increasing the latency when compared to *C*.

Configurations *D* and *E* both have two K80s. Configuration *D* performs slightly better than *E* and this is due to the balanced nature of *D*. As mentioned previously, LAMMPS cannot use the extra CPU in *D*.

An interesting observation here is that when moving from two K80s to four K80s (i.e. comparing *D* and *C* configurations) the performance almost quadruples. This shows that for each extra K80 added (2 GPUs per K80) the performance doubles. This can be partially attributed to the size of the dataset used.

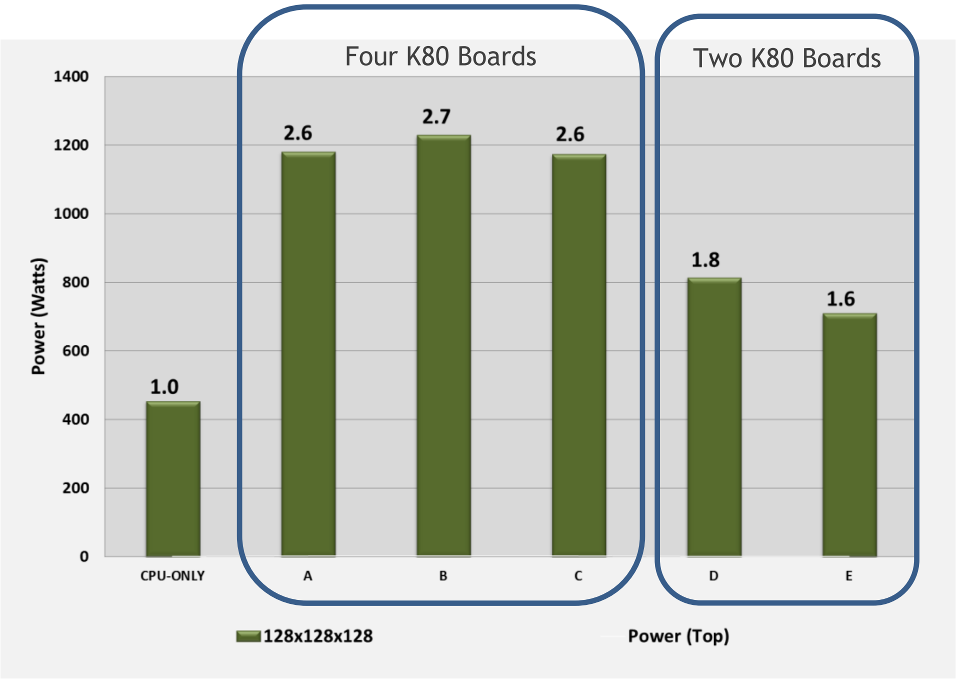


Figure 13: LAMMPS power and relative power consumption compared to CPU-only

Figure 13 shows the power consumption of LAMMPS. The maximum power consumption is in configuration B, but the difference between configurations B and A is small — about 100 watts, implying that the extra CPU is B is not loaded. In case of LAMMPS, the order of power consumption is as follows. B > A >= C > D > E.

Overall, the performance of LAMMPS is substantially improved. The total number of GPUs seems to determine the total acceleration and power consumption. We see up to 16X improvement with only a 2.6X more power consumption. The comparisons in this case are with a dual CPU only configuration. Obviously, there are a lot other factors can come into play when scaling these results to multiple nodes — GPU direct, interconnect, size of the dataset/simulation are just a few of these.

4.5 Accelerating Molecular Dynamics with GROMACS

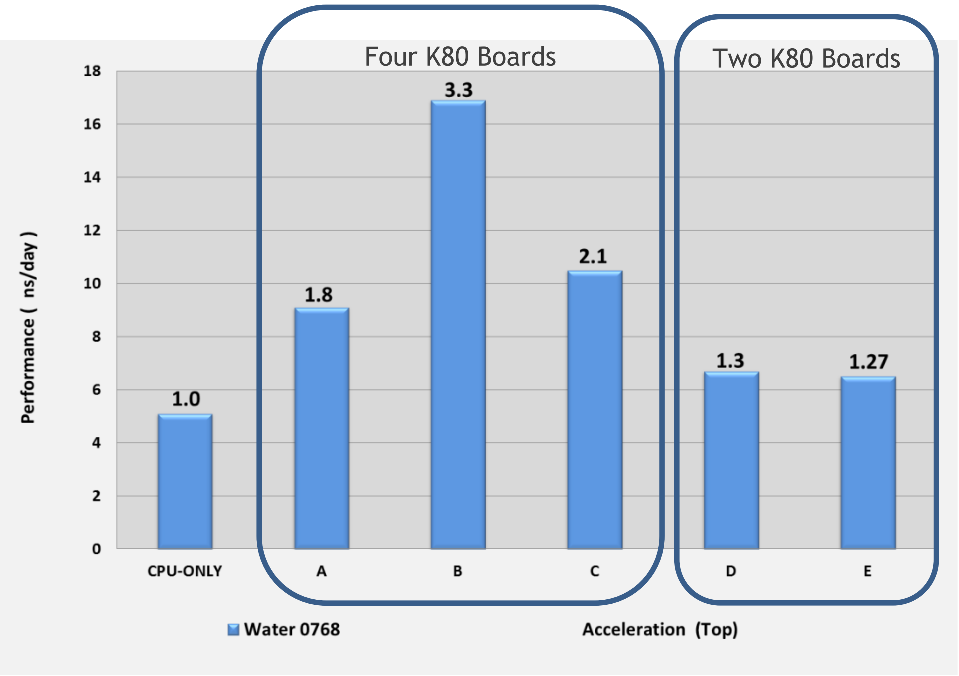


Figure 14: GROMACS performance and acceleration compared to CPU-only

In this section, we evaluate the performance of the third molecular dynamics application. GROMACS is short for “Groningen Machine for Chemical Simulations.” The primary usage for GROMACS is simulations for biochemical molecules (bonded interactions). Because of its efficiency in calculating non-bonded interactions (atoms not linked by covalent bonds), the user base is expanding to non-biological systems. GROMACS performance quantified as “ns/day” (inverse of the number of days required to simulate 1 nanosecond of real-time), the higher score is better. The benchmark ran in GROMACS is the “Water 0768.”

Figure 14 shows the performance of GROMACS on the five C4130 configurations, and the CPU only configuration. Among the quad CPU configurations *A*, *B* and *C*, *B* performs the best. In addition to the four GPUs attached to CPU1, GROMACS also uses the second CPU2, making *B* the best performing configuration. It seems GROMACS benefits from the second CPU as well as the switch. It is likely that application has substantial GPU to GPU communication. Configuration *C* outperforms *A*. This is due to the balanced nature of *C*, and additional latency introduced due to the PCIe switch in configuration *A*. In the dual GPU configurations *D* and *E*, *D*, which is the balanced of the two, slightly outperforms *E*.

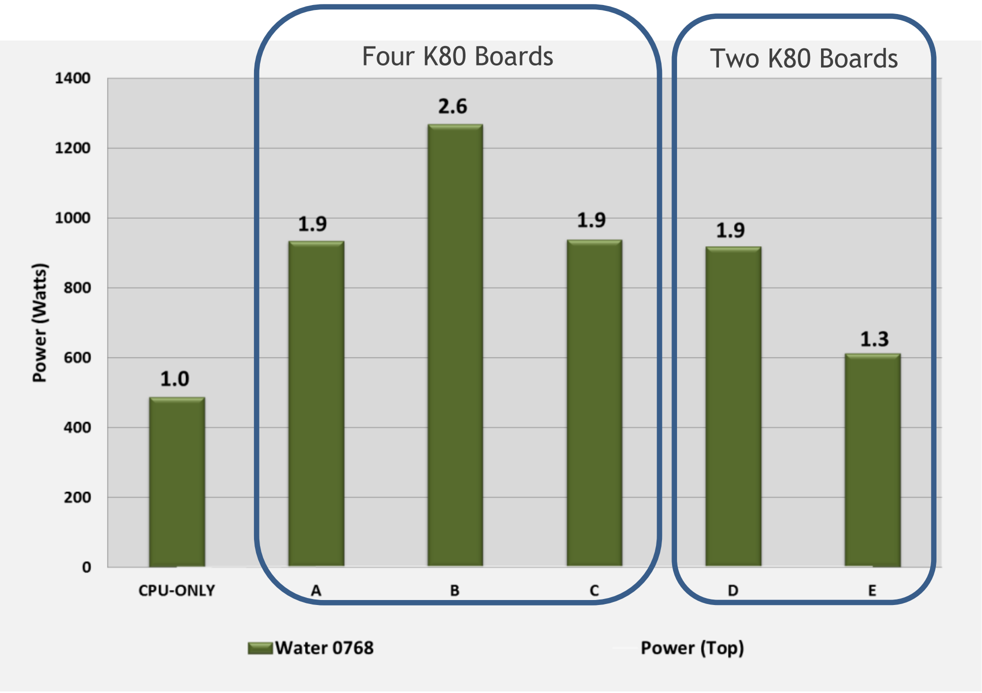


Figure 15: GROMACS power and relative power consumption compared to CPU-only

Power consumption is another critical factor to consider when using performance-optimized servers as dense as the Dell PowerEdge C4130 with 4 x 300 Watt accelerators. *Figure 15* answers questions about how much power these platforms consume.

GROMACS shows the order of power consumption is as follows. *B* >> *A* >= *C* > *D* > *E*. In configuration *B* (switched configuration) has an extra CPU more than Configuration *A*, and takes up considerably more power than *A* and *C* when compared LAMMPS scores. This is because GROMACS uses the extra CPU in *B*, while LAMMPS does not. Configuration *A* incurs a slight overhead of the switch and thus takes up slightly more power than *C*. Configuration *D* is a dual GPU/dual CPU configuration and takes up more power than *E*, which is a single CPU/dual GPU configuration

Overall, the acceleration in GROMACS is less than in LAMMPS and NAMD. This is due to the nature and implementation of the underlying code and algorithms. The best result we see in GROMACS is a 3.3X improvement in performance while talking up 2.6X more power. GROMACS seems to be sensitive to the number of GPUs and CPUs and is able to utilize all available compute resources.

5. Performance improvement compared to previous generation of PowerEdge C410X solutions

The compute power of GPU solutions has increased many times over in recent years. The latest GPU-based PowerEdge C4310 solution offers a substantial performance improvement compared to the previous PowerEdge C410X solution. In this section, we compare the relative performance on the C4130 solution to the C410X-based solution. This data will prove useful for users considering switching from the previous external GPU-based C410X solution to the current internal GPU-based C4130 offering. The number of GPUs is the constant in both cases. Table 1 below shows the configuration of two systems.

Table 1. Configurations of current and previous GPU solutions

Server	PowerEdge C4130	PowerEdge C410X/C6100
Processor	1 or 2 x Intel Xeon CPU E5-2690 v3 @ 2.6 GHz (12 core)	2x Intel X5650 @ 2.67 GHz (6 core)
Memory	64GB or 128GB @ 2133MHz	48GB @ 1333MHz
GPU Board	2 or 4 x NVIDIA Tesla K80	2 or 4 x NVIDIA Fermi M2070
Number of internal per GPU Board	K80 has two internal GPUs	M2070 has one internal GPU
GPU Connection to host	Internal	External (via HIC)
GPU Memory	24 GB	6 GB
GPU power	300W	225W
Power supply	2 x 1,600W	2 x 1600W
Operating System	RHEL 6.5, (2.6.32-431.el6.x86_64)	RHEL 5.5, (2.6.18-194.e15)
BIOS options	System profile - max performance	System profile – max performance
	Logical processor - disabled	Logical processor - disabled
CUDA Version and driver	CUDA 6.5 (340.46)	CUDA 4.0
BIOS firmware	1.1.0	1.54.92
HPL	NVIDIA pre-compiled HPL 2.1	NVIDIA pre-compiled HPL 1.1
NAMD	Version 2.9	Version 2.8b1

As show in the table above, there are several advances in the hardware and software components of the solution. The processor core count has doubled from six to twelve. The system memory is now 128GB for two CPUs (64 GB for single CPU). The bulk of the improvement is in the raw compute power of the GPUs. The M2070 is rated at 515 GFLOPS (double precision) and K80 has a rating of 1.87 to 2.91 TFLOPS (double precision), giving a 3.6X to 5.6X improvement over M2070. The GPU memory has increased by four fold from 6GB per GPU to 24 GB per GPU. The system architecture also plays a role. The previous solution had external GPUs. There have been numerous improvements in the application code, with both HPL and NAMD going through major revisions. Given these specifications, we will compare the relative performance of these solutions by **combining all advances in hardware and software components**.

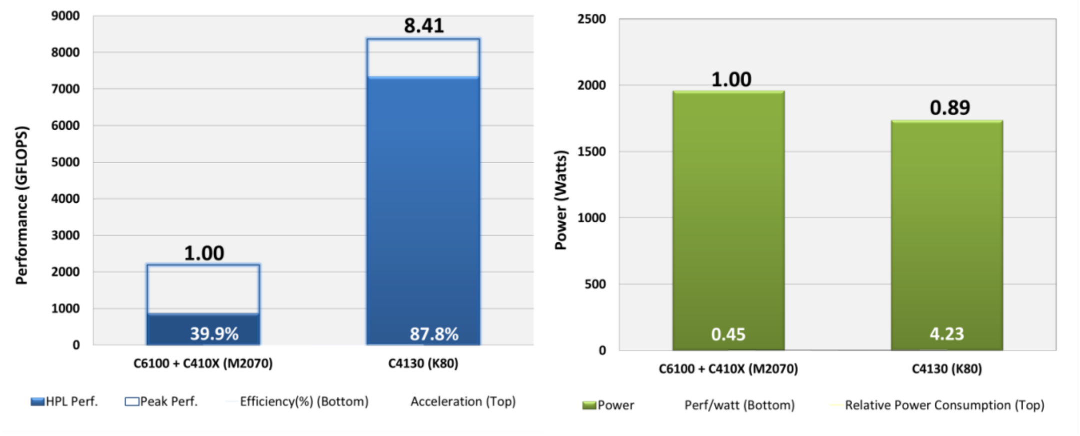


Figure 16: Comparison of HPL performance between C410X and C4130 with four GPU boards

Figure 16 shows the comparison on HPL with four GPU boards. The total peak performance goes from 2.2 TFLOP to 8.4 TFLOP — an improvement of 3.8X. The actual achieved or sustained performance is more complex. First to note is the HPL efficiency has gone from 39.9% to 87.8%. This is mainly due to code enhancements and having internal GPUs. Due the higher efficiency, the effective increase in sustained performance is about 8.4X. On the power side, even with higher rated power of 300W, the actual power consumption is less. This is due to two main factors — improved system architecture with internal GPUs reduces power required by external chassis and various architectural improvements in the GPUs that improve performance per watt. **The net gain in performance per watt is about 10X**.

Similarly, *Figure 17* shows the performance improvement with two GPU boards. The peak performance goes from 1.1 TFLOPS to 4.6 TFLOPS an increase of about 4X. The sustained performance improves by 5.8X this is lower than the previous four GPU case because the HPL efficiency of C6100+C410X is higher with two GPUs (56.3%). The power consumption difference is larger than the four-board case, because with the two board configuration, the C4130 uses less power — about 62% of the previous C410X based solution. **Finally, the total gain in performance per watt is 9.3X for the two GPU case.**

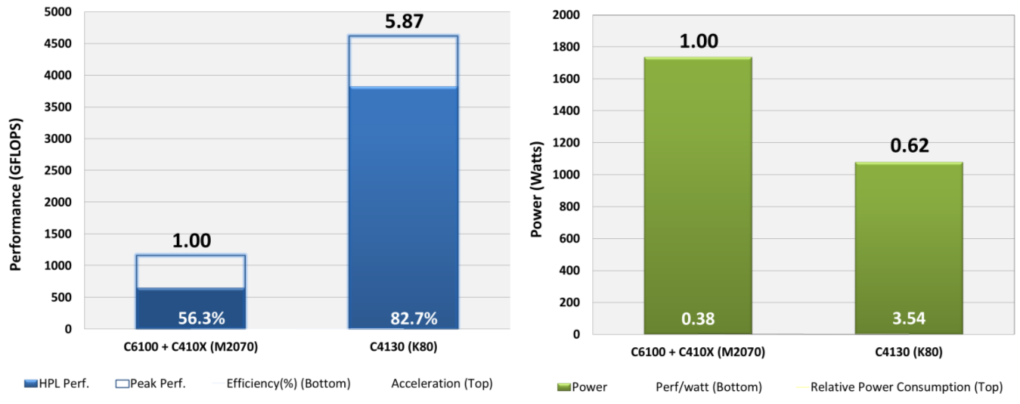


Figure 17: Comparison of HPL performance between C410X and C4130 with two GPU boards

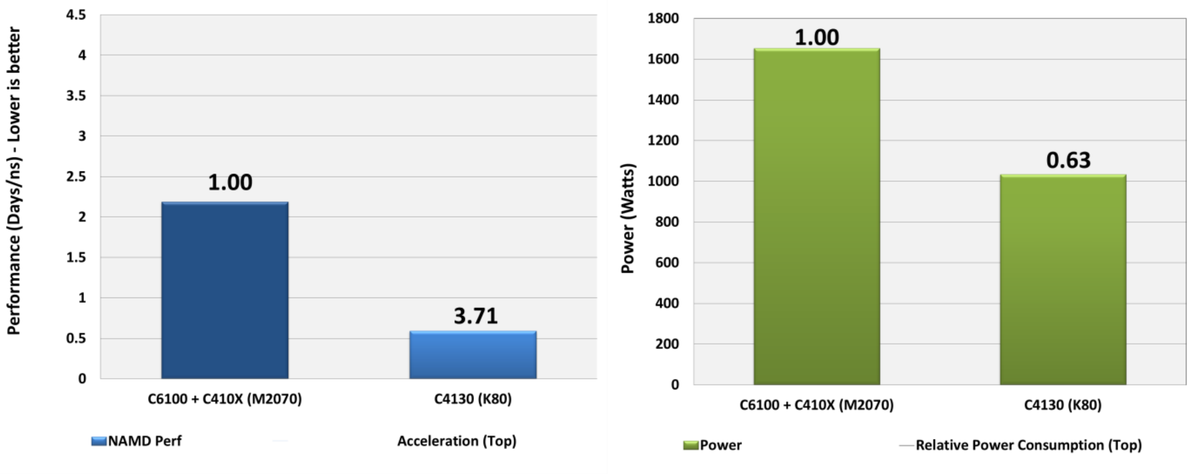


Figure 18: Comparison of NAMD performance between C410X and C4130 with four GPU boards

Now we consider the relative performance on a molecular dynamics application NAMD. *Figure 18* shows the performance of NAMD with four GPU boards. In case of NAMD, lower is better, so we see a 3.7X performance improvement on the STMV benchmark. The power consumption is 63% better in C4130 due the improvements in GPU and system design. **The performance per watt improvement can be estimated as 3.7X x (1/0.63) = 5.8X**. Similarly, *Figure 19* show the performance improvement with just two GPU boards, the performance gain is higher at about 4.3X and power consumption ratio are similar at about 59%. **The relative performance per watt is approximately 4.13X x (1/0.59) = 7X.**

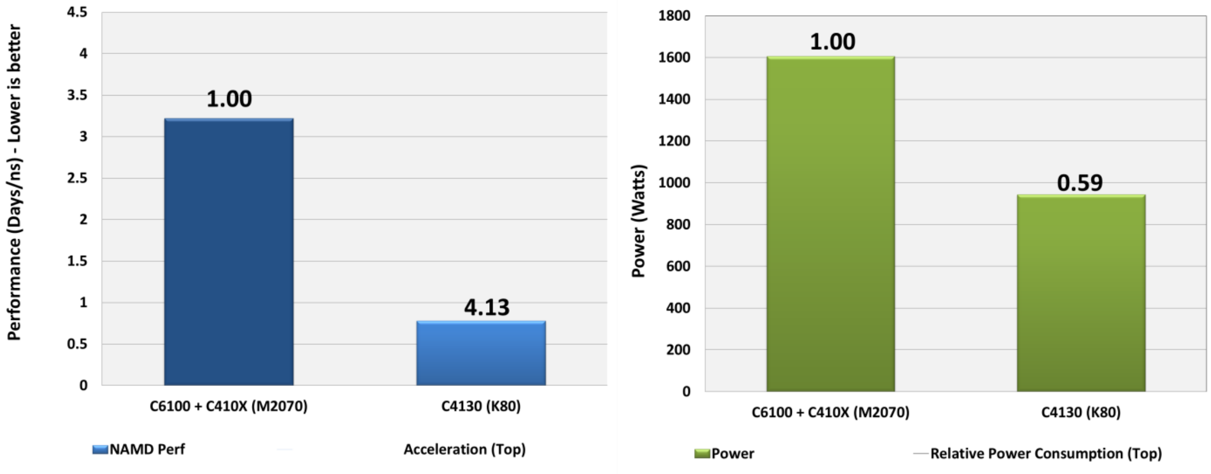


Figure 19: Comparison of NAMD performance between C410X and C4130 with two GPU boards

6. Conclusion

In conclusion, the C4130 meets the current challenges of a high-density, accelerator-enabled compute node. Targeted specifically towards the HPC market, it offers world-class performance and unique configurability options to fit extreme HPC requirements. The following are the main results:

- The various configurations of the C4130 offer a range of CPU to GPU ratios
- C4130 offers both balanced and switched connections between CPU and GPUs
- The unique placement of GPU in the C4130 ensures optimal GPU performance
- C4130 offers a wide range of GPU boards and coprocessors
- On the K80s, the HPL performance is about **9X compared to CPUs** resulting in **4.2 GFLOPS per watt** power efficiency
- Using the K80s can substantially accelerate industry-standard molecular dynamic codes. Observed **acceleration is from 2X to 16X depending on the code** and benchmark used. This improvement is achieved in a power efficient manner, consuming only 2X to 4X more power.
- Compared to the previous generation, the C410X-based GPU solution using M2070s, the C4130 with K80s offers a vastly improved solution. Comparing four GPU boards, the following are the main enhancements:
    - The **performance on HPL is 5X to 8X** better with reduced power consumption, resulting in a **9X to 10X performance per watt improvement**
    - The **performance on NAMD is 3X to 4X** better with reduced power consumption, resulting in **6X to 7X performance per watt improvement**
    - On a **“GPU per U” basis the compute density is improved 2.5X to 3.5X**. Previously the C410X based solution required at least 5U to 7U for 16 GPUs (sixteen M2070); with the latest C4130, users could have access to 16 GPUs (eight K80s) in only 2Us.

Appendix A: Hardware Configuration of the C4130

Server	PowerEdge C4130
Processor	1 or 2 x Intel Xeon CPU E5-2690 v3 @ 2.6 GHz (12 core)
Memory	64GB or 128GB @ 2133MHz
GPU	2 or 4 x NVIDIA K80 (4,992 CUDA cores, base clock 562 MHz, boost clock 875MHz, power 300W)
Power supply	2 x 1,600W
Operating System	RHEL 6.5 – kernel 2.6.32-431.el6.x86_64
BIOS options	System profile - performance
	Logical processor - disabled
	Power supply redundancy policy – Not redundant
	Power supply hot spare - Disabled
CUDA Version and driver	CUDA 6.5 (340.46)
BIOS firmware	1.1.0
iDRAC firmware	2.02.01.01
MPI	Openmpi 1.6.5
MKL	Intel MKL 2015

Appendix B: Applications and Benchmarks

Name	Version
HPL	NVIDIA pre-compiled HPL 2.1
NAMD	2.9
LAMMPS	Feb 1, 2014 stable version with lib/CUDA for GPU acceleration
GROMACS	4.6.6

References

1. Visit http://www.nvidia.com/tesla for more information on GPUs.
2. For general information on Top 500 GPUs, please visit http://www.top500.org
3. For more information on the top Energy efficient supercomputers, please visit http://www.green500.org/
4. For more information on NAMD, please visit http://www.ks.uiuc.edu/Research/namd/
5. NAMD, available: http://www.ks.uiuc.edu/Research/namd/
6. LAMMPS available: http://lammps.sandia.gov/
7. Lennard-Jones liquid benchmark http://lammps.sandia.gov/bench.html#lj
8. GROMACS. Available: http://www.gromacs.org/
9. GROMACS benchmark ftp://ftp.gromacs.org/pub/tmp/water-clean-input.tar.gz