Dell PowerEdge C4130 Performance with K80 GPUs - HPL

Authors: Saeed Iqbal and Mayura Deshmukh

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 GPUs in a 1U form factor. Also 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.

This blog is about performance characterization of the C4130 on HPL, we present data on performance achieved, power consumption and performance per watt on various system configurations.

The latest HPC focused Tesla series General Purpose Graphic Units (GPU) released from NVIDIA is the Tesla K80, from the HPC prospective *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 a total of 4,992 cores which represent a 73% improvement over K40. The K80 has a total memory of 24GBs which is divided equally between the two internal GPUs; this is a 100% more memory capacity compared to the K40. The memory bandwidth in K80 is improved to 480 GB/s. The rated power consumption of a single K80 is a maximum of 300 watts.

The C4130 offers five configurations “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”, with four GPUs each and the second group of two configurations, “D” and “E”, with two GPUs 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 and the block diagram (Figure 1) below:

Table 1: C4130 Configurations

C4130 Configuration	GPUs	CPUs	Switch Module (SW)	GPU/CPU ratio	Comments
A	4	1	Y	4	Single CPU, optimized for peer to peer communication
B	4	2	Y	2	Dual CPUs, optimized for peer to peer communication
C	4	2	N	2	Dual CPUs, Balanced with four GPUs
D	2	2	N	1	Dual CPUs, Balanced with two GPUs
E	2	1	N	2	Single CPU, Balanced with two GPUs

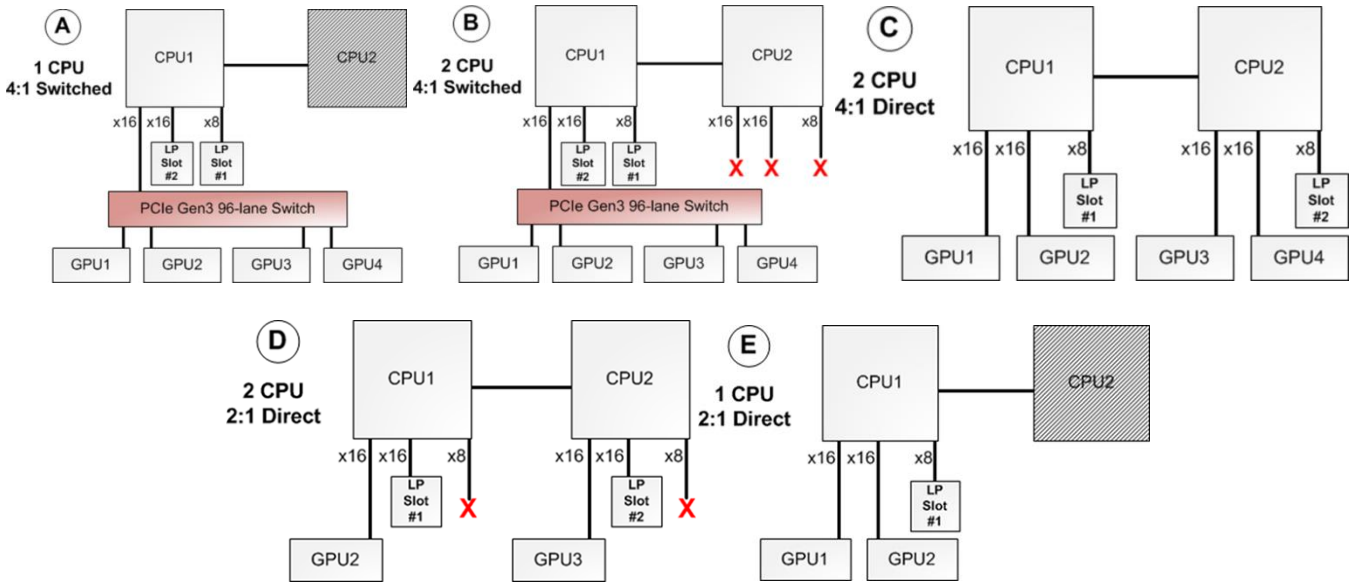

Figure 1: C4130 Configuration Block Diagram

Table 2 gives more information about the hardware configuration and the benchmark details used for the tests. For HPL the problem size used was ~85%-~90% of the system memory.

Table 2: Hardware Configuration and benchmark details

Server	**C4130 Prototype System**
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
HPL	NVIDIA pre compiled HPL 2.1
MPI	Openmpi 1.6.5
MKL	Intel MKL 2015

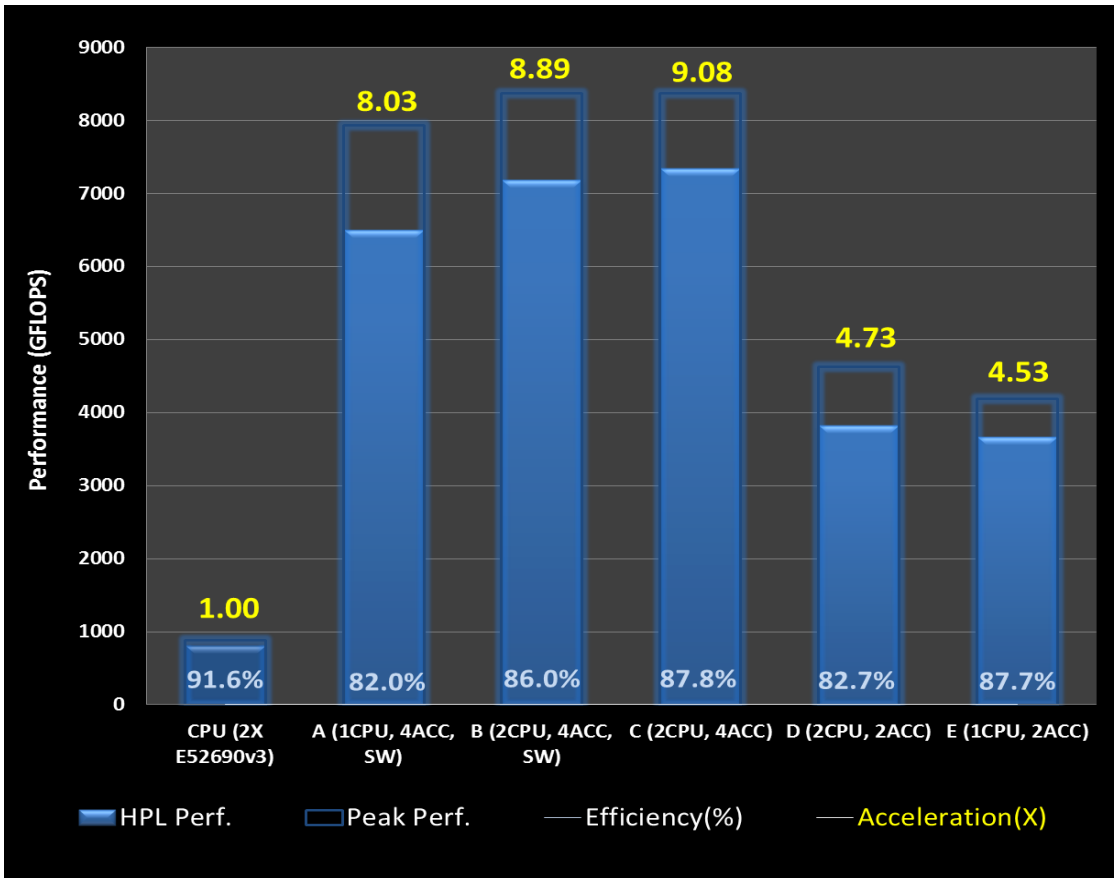

Figure 2: HPL performance, efficiency and acceleration on the five C4130 configurations.

Figure 2 shows the HPL performance characterization of PowerEdge C4130. Configurations "A", "B" and "C" are four GPU 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. The difference can be explained due to one less CPU with configuration "E".

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 GPUs. The HPL efficiency is significantly higher on K80 (low to upper 80s) compared to previous generation of GPUs.

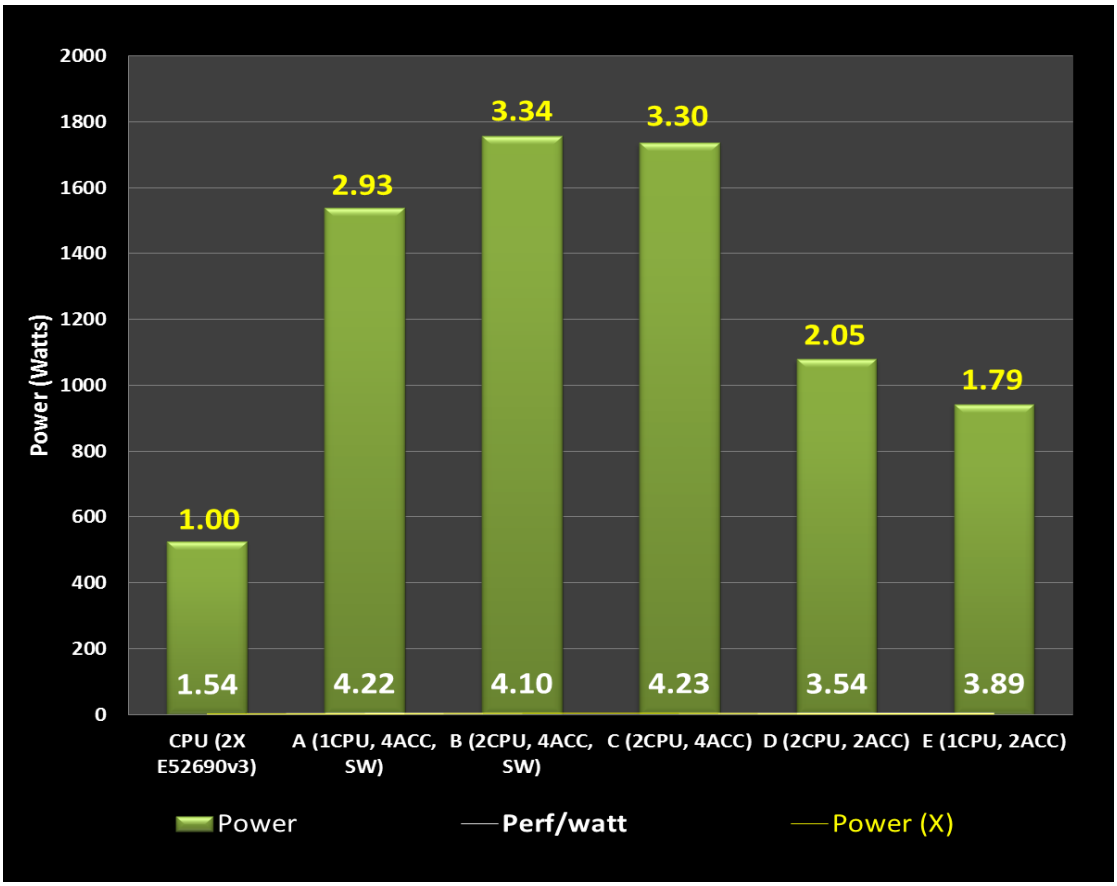

Figure 3: Total power and performance/watt on the five C4130 configurations.

Figure 3 shows the power consumption data for the HPL runs in Figure 2. In general, GPUs can consume substantial power when loaded with 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.

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.