System Design of Kepler Based HPC Solutions

Saeed Iqbal, Shawn Gao and Kevin Tubbs

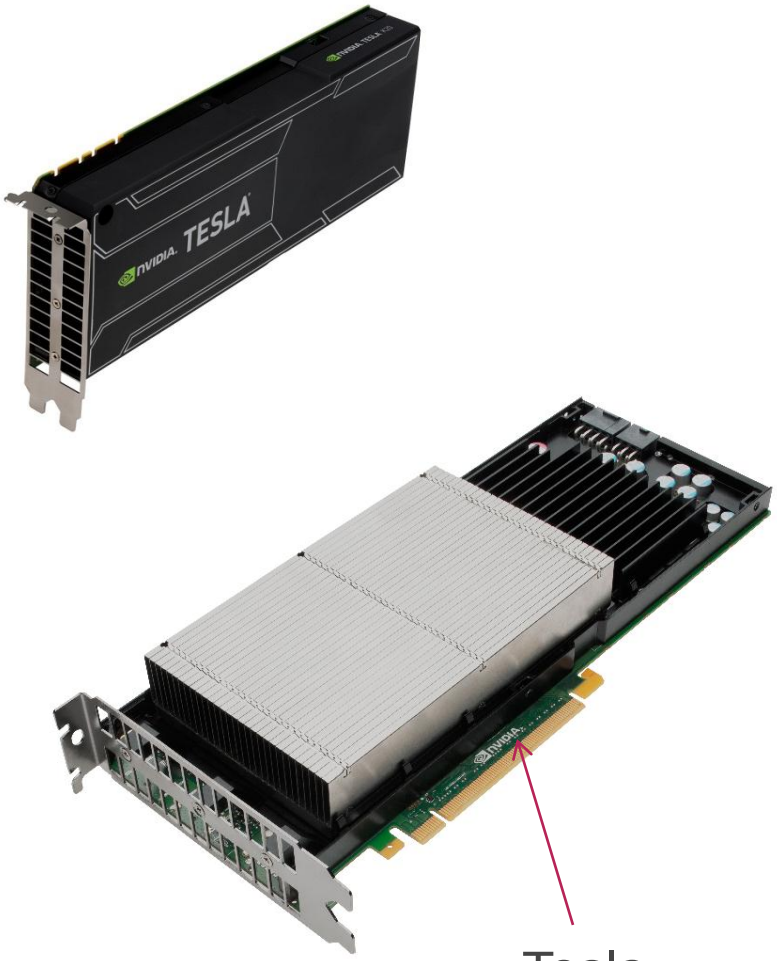
HPC Global Solutions Engineering.

Introduction

The System Level View

K20 GPU is a powerful parallel processor!

- K20 has high raw compute power!
  - **6.3 – 6.5 X CPUs** (Theoretical Peak)
- Compare to M2090
  - Kepler (GK110) architecture
  - Transistors 7.1B **2.4X**
  - Cores 2496 **4.8X**
  - Memory 5GB 0.8X
  - Mem-BW 208GB/s 1.17X
  - Power 225W **0.9X**
  - SP 3.52 TFLOPs (2.7X)
  - **DP 1.17 TFLOPs (1.7X)**



Tesla K20 GPU

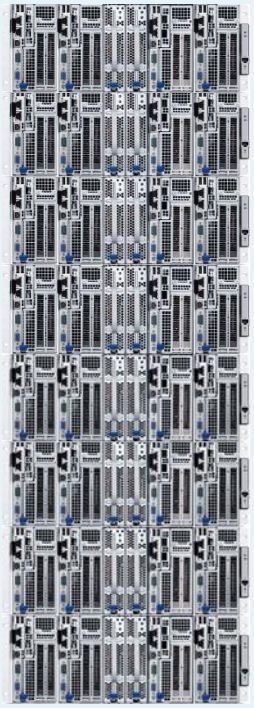
- Challenge:
  - **How to "Realize" & "Extract" maximum performance ?**

K20 GPU is a powerful parallel processor! Yes, how to get the most out of it?

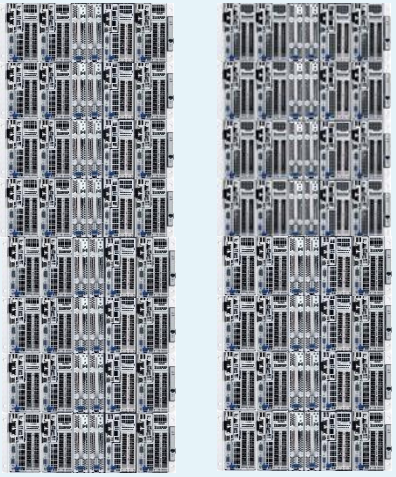
GPUs →

Servers →

Clusters →

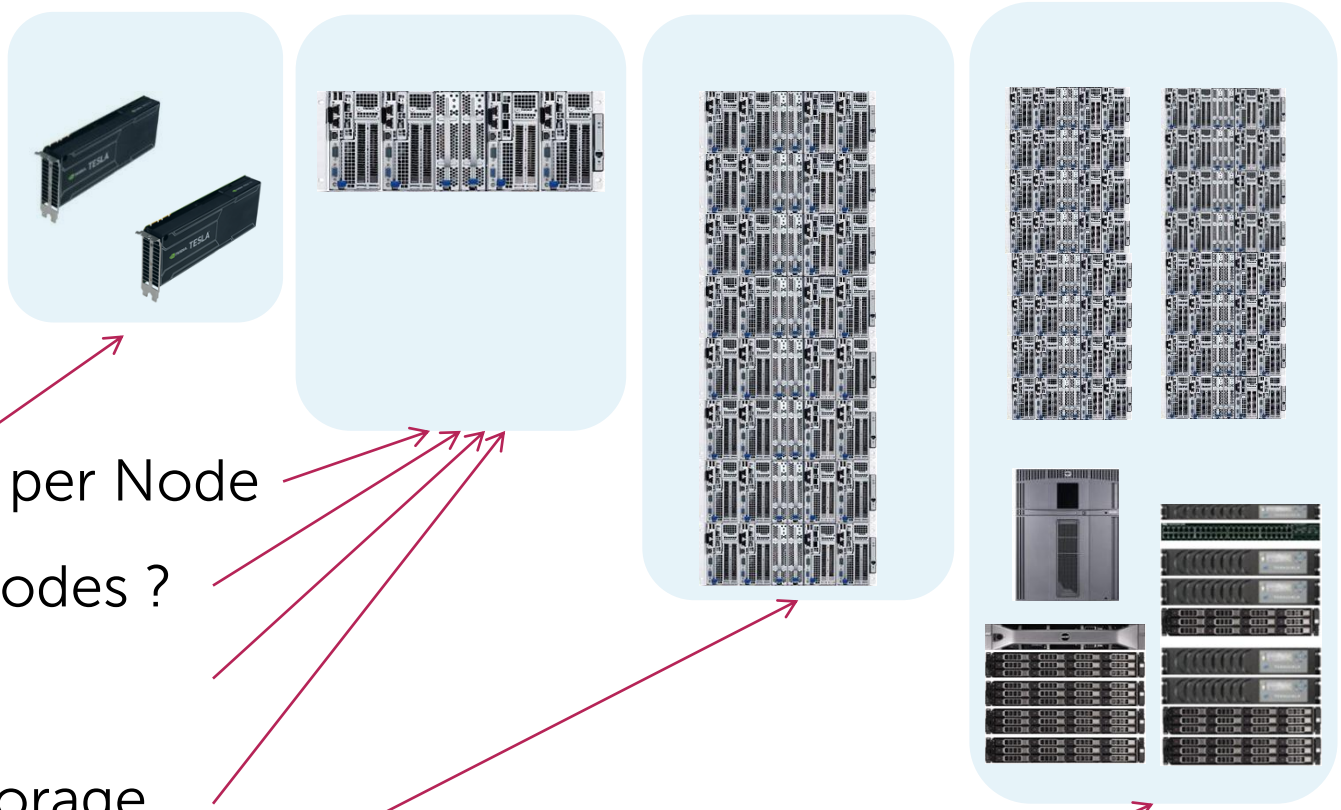


Solutions



- "Performance Chain"
- "Balance"
  - Eliminate bottle necks
- Maximize Return from Investment

Some Key Issues: System Level Design



- Number of GPUs per Node
- Dedicated GPU nodes ?
- GPUs per node
- Local memory/storage
- Power Budget per Rack
- Number of Nodes with GPUs for best ROI

Servers

Two Server Form Factor Options Ready for K20 GPUs

- PowerEdge C8220X
  - "Shared Infrastructure"
  - 4U
  - Higher GPU & CPU Density
  - Higher Configurability
- PowerEdge R720
  - "Conventional Rack Server"
  - 2U
  - Higher memory per node (768GB)
  - Higher storage per node (24TB)

C8000+C8220X

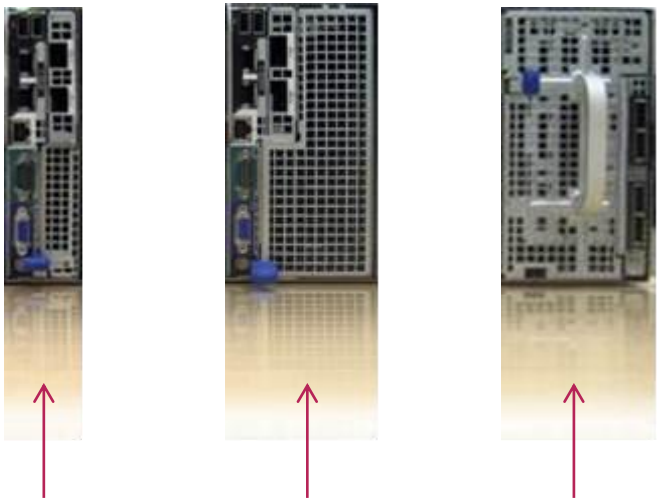
3D View

R720

3D View

The C8000 Series: CPU, CPU+GPU Sleds

Based on the “Shared Infrastructure” design



C8220 C8220X C8220XD

C8000

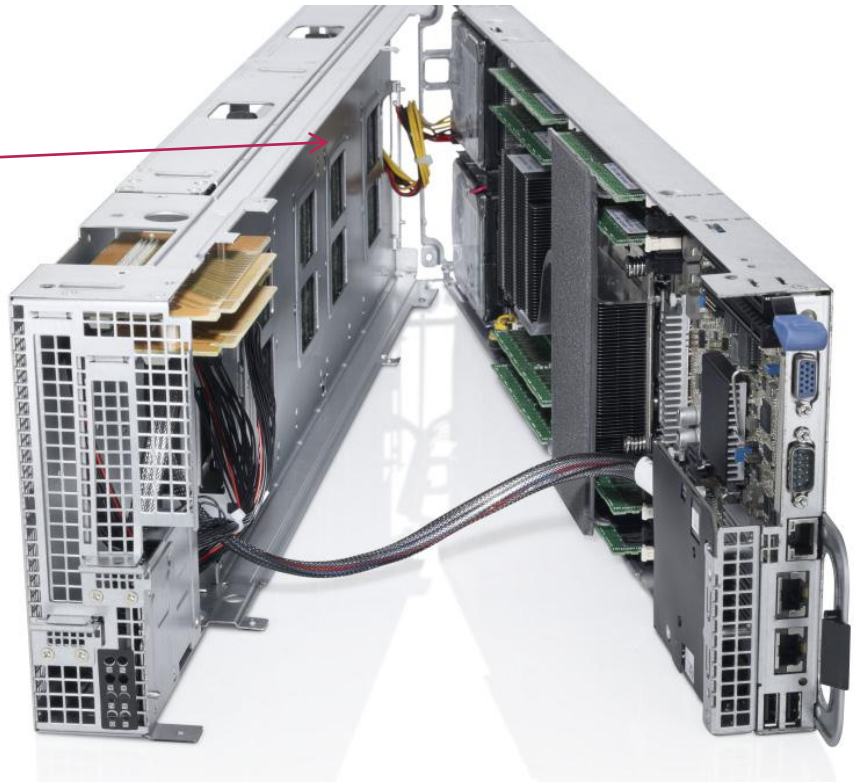
- C8220 (single wide, compute sled)
- C8220X (double wide, compute sled)
- C8220XD (double wide, storage sled)

As demands change it can be reconfigured or scaled out extending the life and value of IT infrastructure investments

Server Details: PowerEdge C8220X

Each C8220X has:

- Up to 2 K20 GPUs
- Two E5-2600 CPUs
- 256GB of memory
- Combine sleds
  - **4 C8220X Sleds in one C8000**
- 8 GPUs in 4U space
  - **2 GPU/U Density**



PowerEdge
C8220X Compute
GPU Sled

Server Details: PowerEdge R720

Each PE R720 has:

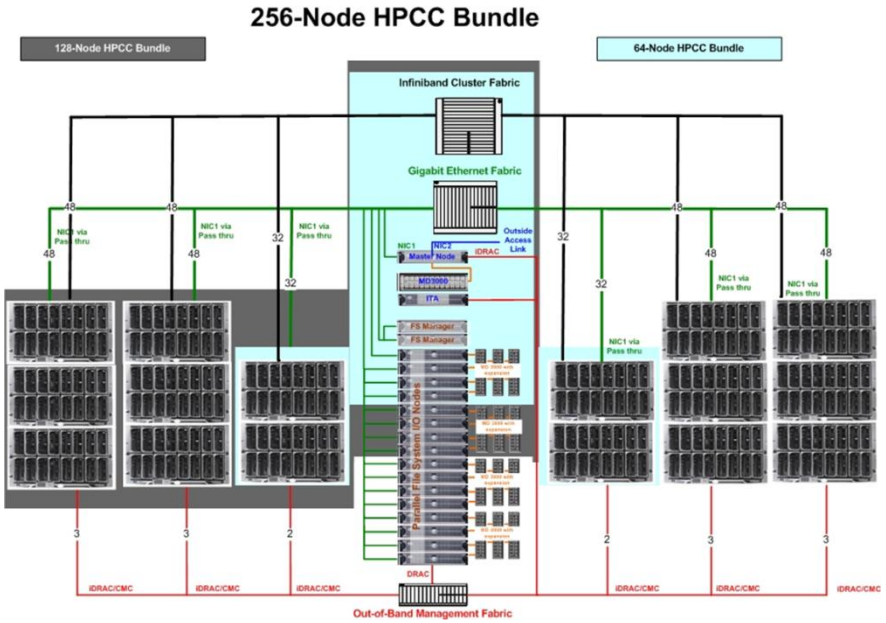
- Up to 2 K20 GPUs
- Two E5-2600 CPUs
- 768GB of memory
  - 24 X 32G DIMM
- 24TB local storage
  - 16 X 2.5TB Drives
- 2 GPUs in 2U
  - **1 GPU/U density**

PowerEdge R720

Solutions

“Dell HPC Solutions” Mean “Value” for you

- Solutions “Goal” is to provide “value”
  - Enables you to focus on you “science”
  - Brings your HW up to speed *FAST*
- Engineering Rigor
  - Performance Envelop
  - Measure Total Power Consumption,
  - Expected Power efficiency
- Best practices
  - HPC Advisor
  - Whitepaper Publications,
  - Public Results



“Dell Solutions” Mean “Value” for you Big part of it the Engineering

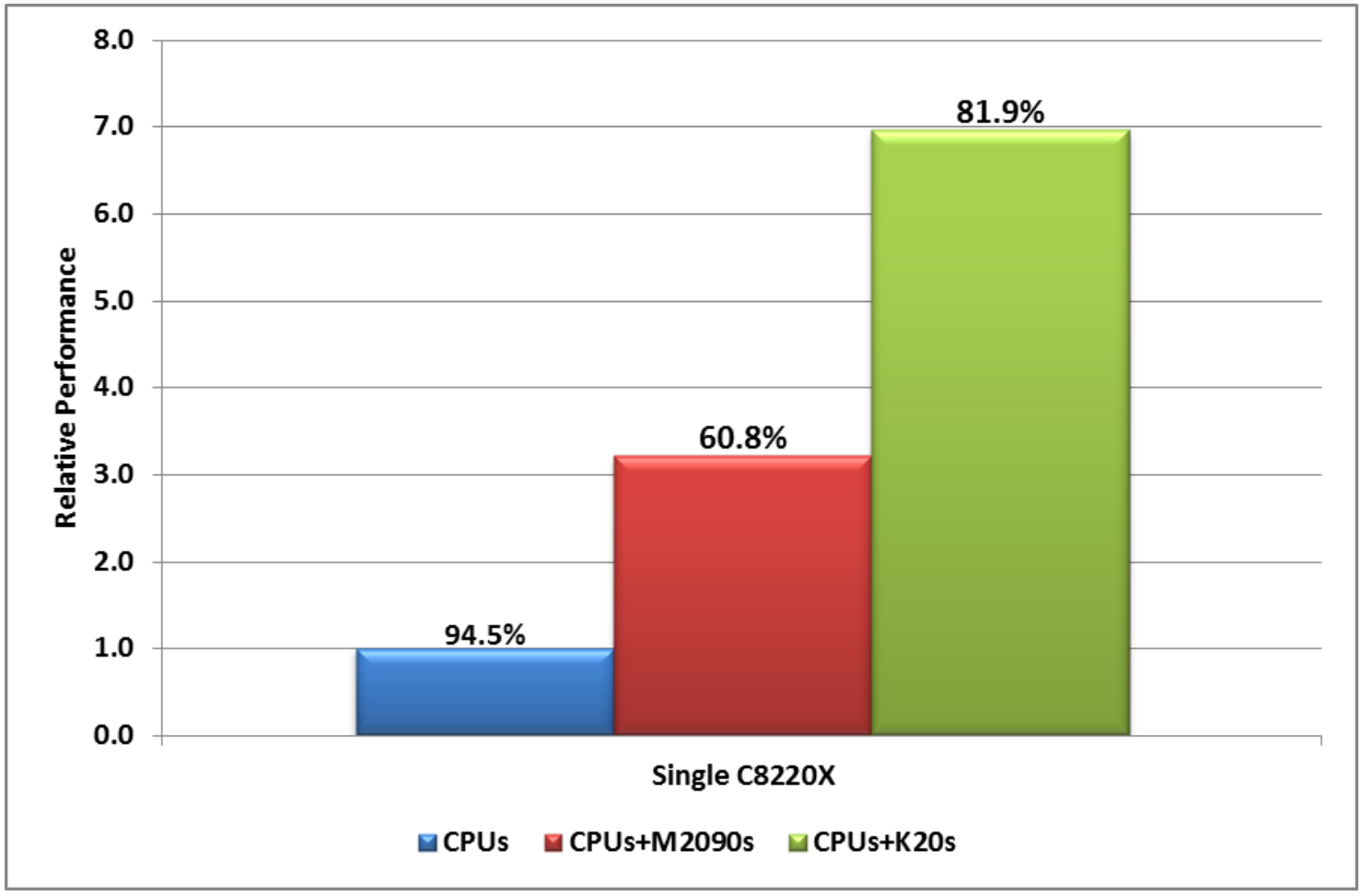
- Tests suite includes
  - Node level Performance
  - Cluster level Performance
- Power
  - Total Measured System Power Consumption
  - Performance/watt studies for efficient configurations
- System level
  - Host-to-device, Device-to-host, Device-to-Device
  - Memory subsytem
- Applications level : Benchmarks and Applications
  - HPL, NAMD, NPB, ANSYS

Server Performance & Sensitivity to K20 parameters

System Configuration:
K20 on C8220X and R720 (change of 225W)

Component	Value
Server	PowerEdge C8220X
Processor	Dual Intel Xeon E5-2670 @ 2.6GHz
Memory	128GB @ 1600MHz
InfiniBand	Mellanox Connect-X 3 FDR
GPU	NVIDIA Tesla K20
Rated GPU Power	225W
CUDA Version	5.0 (304.51)
OS	RHEL 6.3

C8000+C8220X: Single Node HPL Relative Performance

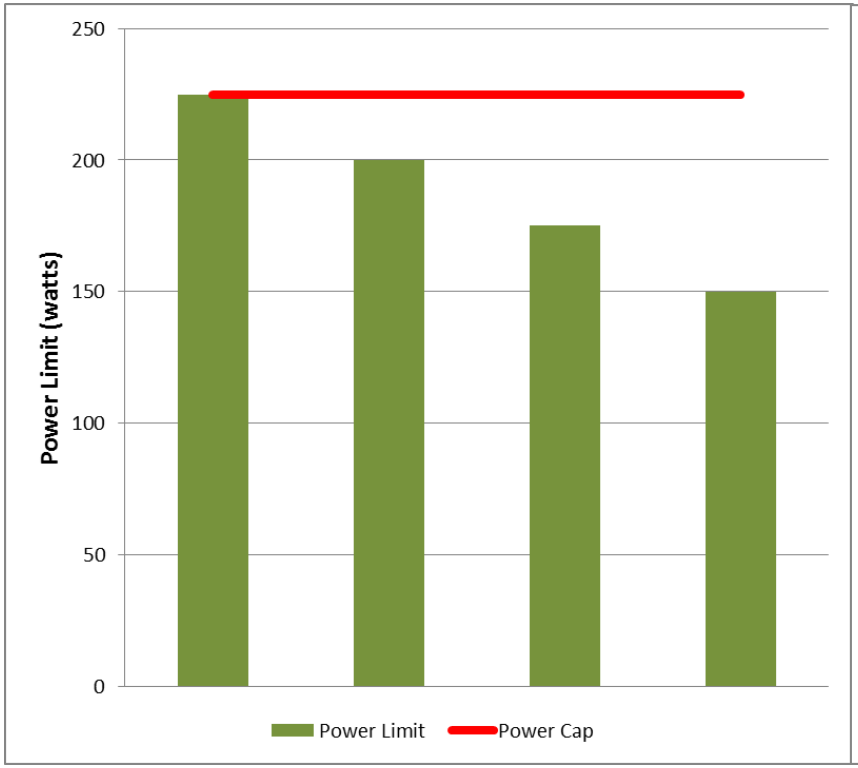


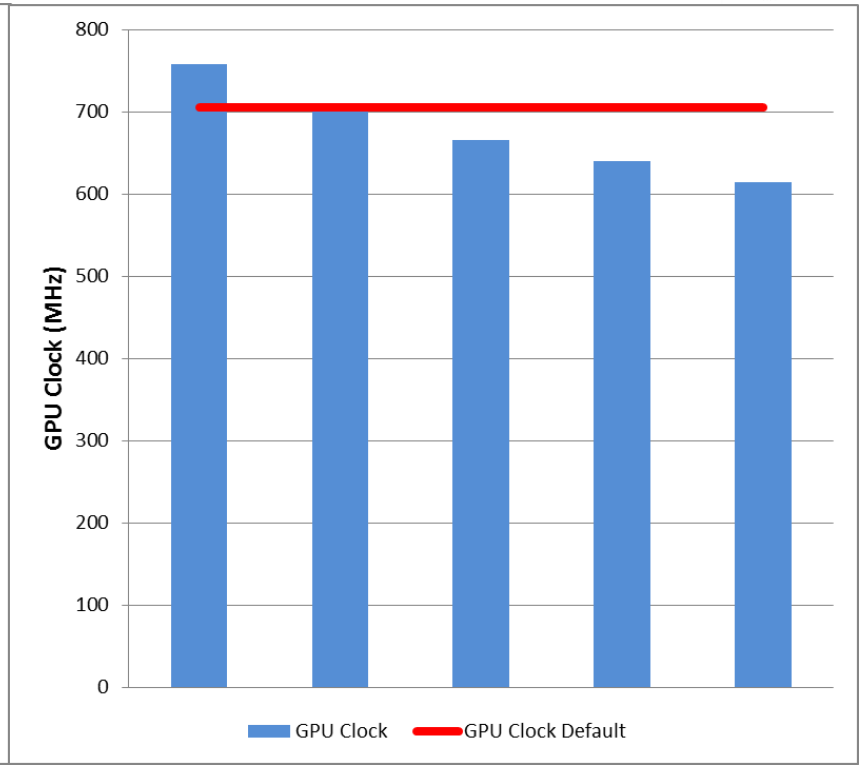
K20 is up to 6.9X the CPUs and twice as fast as M2090s !

1-node PEC8220X, Dual E5-2670@2.6GHz, 128GB 1600MHz Memory; C8000 has 1 C8220X sled, CUDA 5.0.24 (304.54)

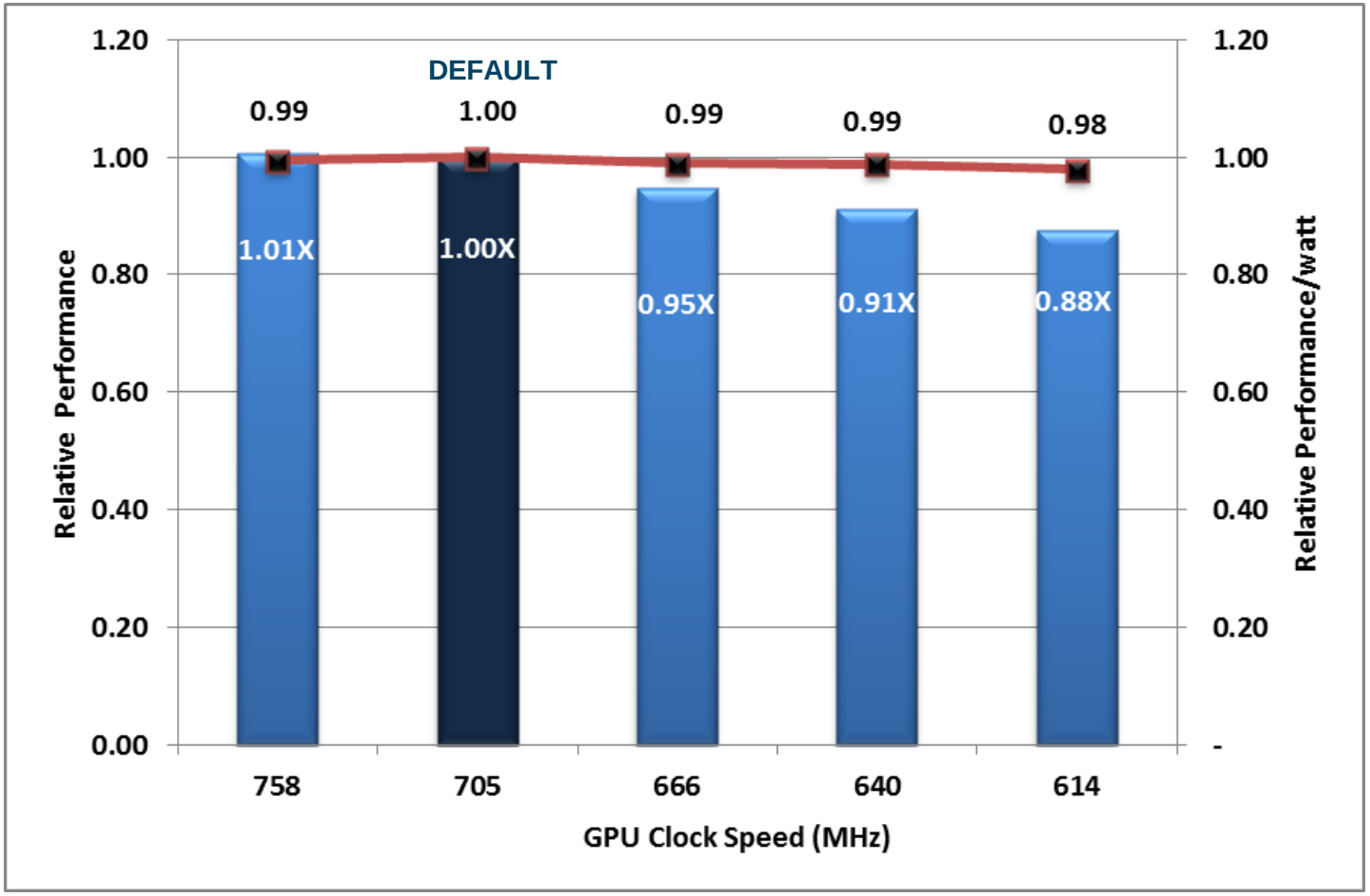
Changing “Power limit” and “GPU Clock Speed”

- On a K20, power consumption and clock rate can be adjusted:
  - nvidia-smi -ac --application-clocks=<memory, graphics>
  - nvidia-smi -pl --power-limit=<limit>





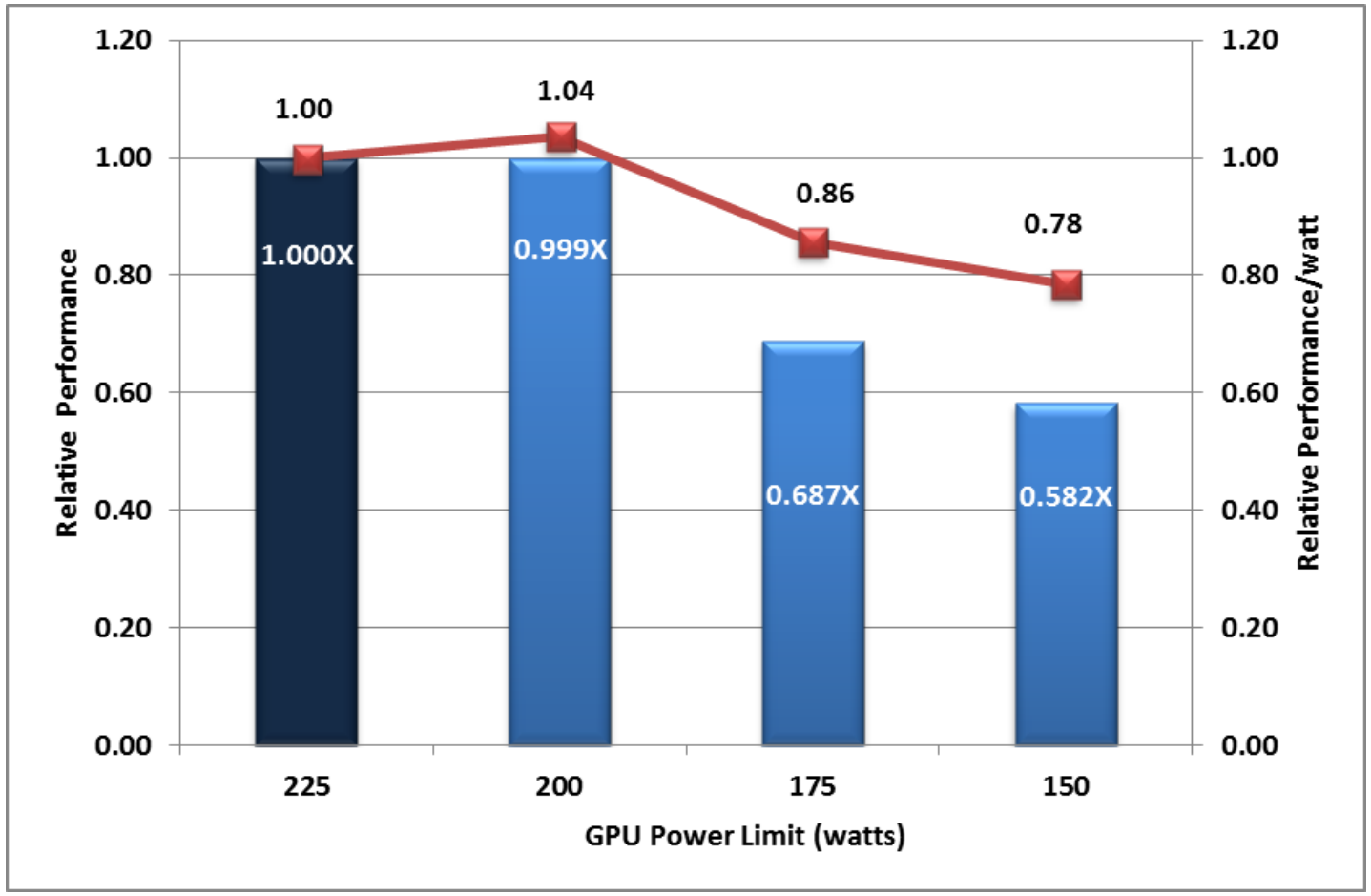
C8000+C8220X: Single Node Performance Sensitivity to GPU clock



N=109312
NB=1024

1-node PEC8220X, Dual E5-2670@2.6GHz, 128GB 1600MHz Memory; C8000 has 1 C8220X sled, CUDA 5.0.24 (304.54)

C8000+C8220X: Single Node Performance Sensitivity to Power Limit



N=109312
NB=1024

HPL Perf. did not degrade significantly at 200W! **save 36.4 Watts**

1-node PEC8220X, Dual E5-2670@2.6GHz, 128GB 1600MHz Memory; C8000 has 1 C8220X sled, CUDA 5.0.24 (304.54)

Changing Power limit and Clock rate

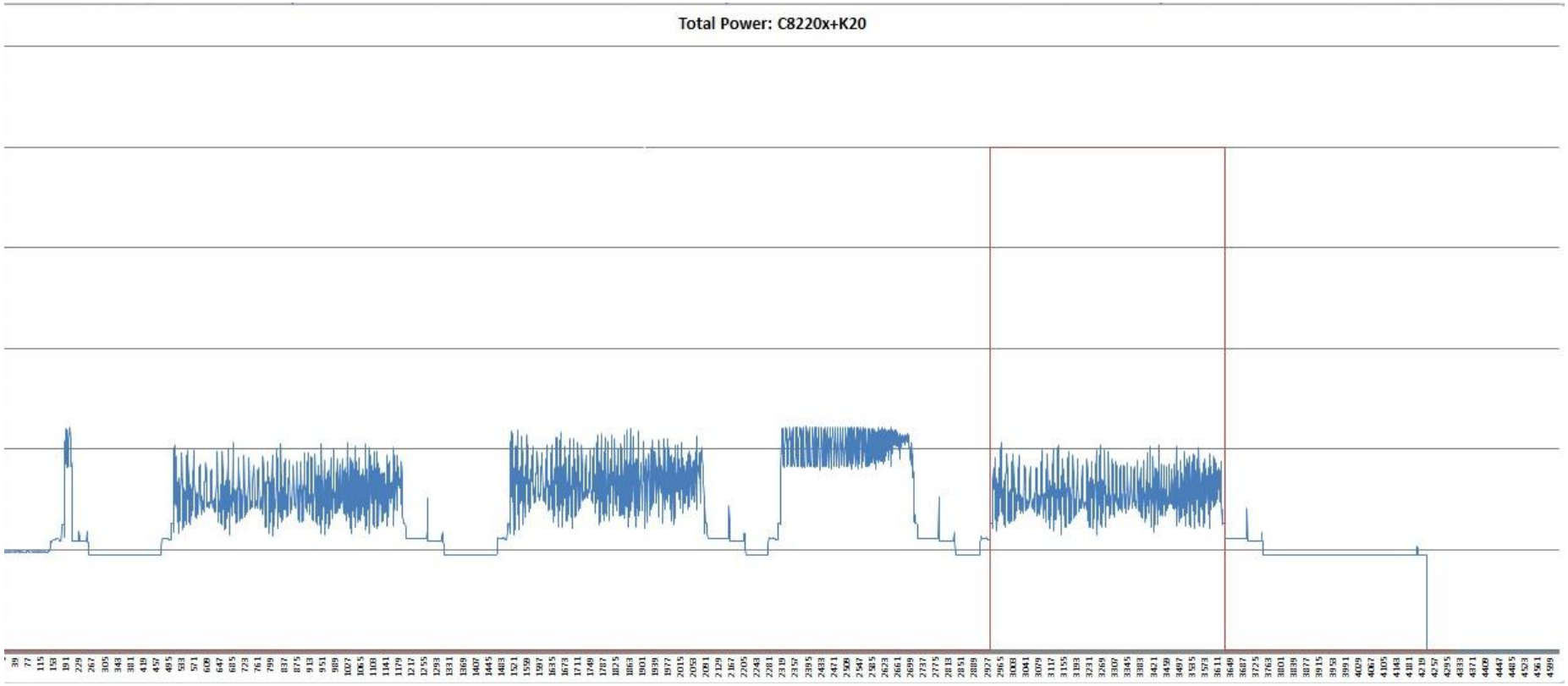
- Performance is more sensitive to “power limit” than “clock”
- Useful for Power Budget limited Cluster design
- Overclock does improve performance but only slightly.
- Useful for “Power Aware” GPU based HPC solution design.

Cluster Performance & Scalability

Cluster Performance

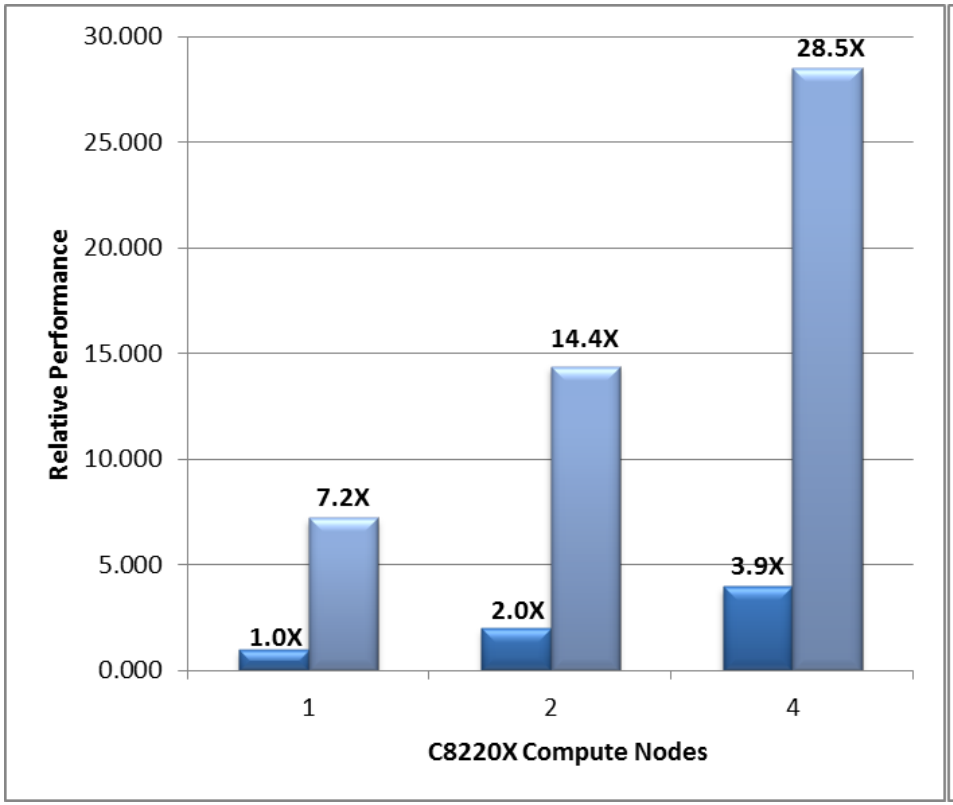
- Performance Scalability
  - Number of nodes
  - Number of GPUs
- Total Solution Power Measured
- Standard un-tuned benchmarks
  - HPL, NAMD, NPB

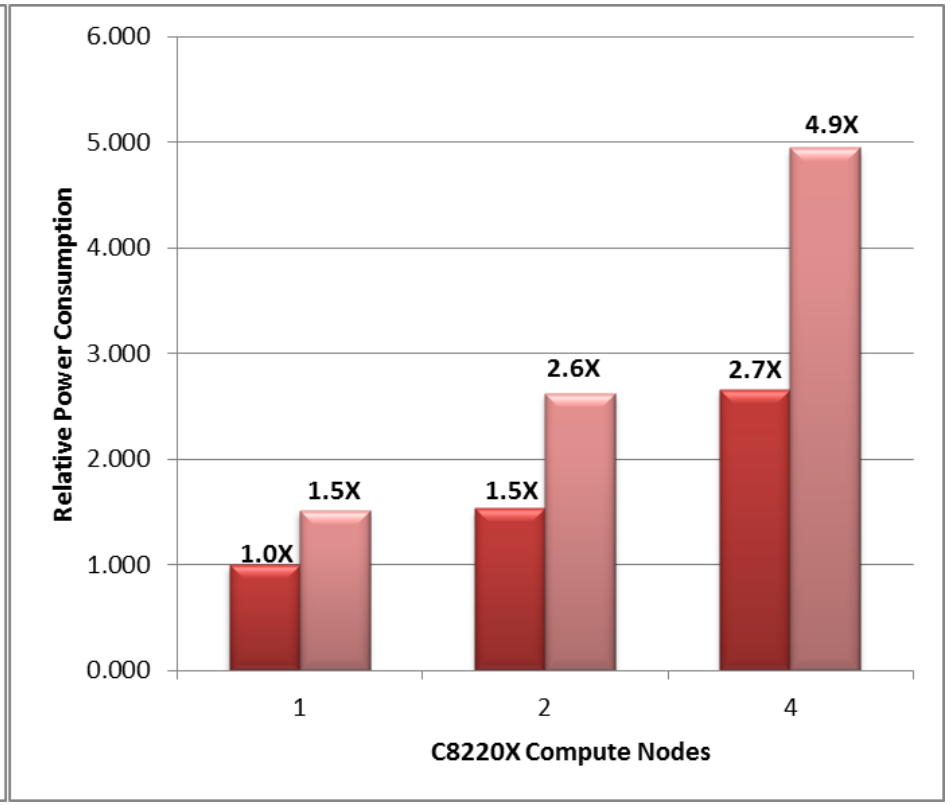
Cluster Power Consumption



Total Solution Power is Measured

C8000+C8220X: Scalability
HPL Perf. & Power (K20 vs. CPU)

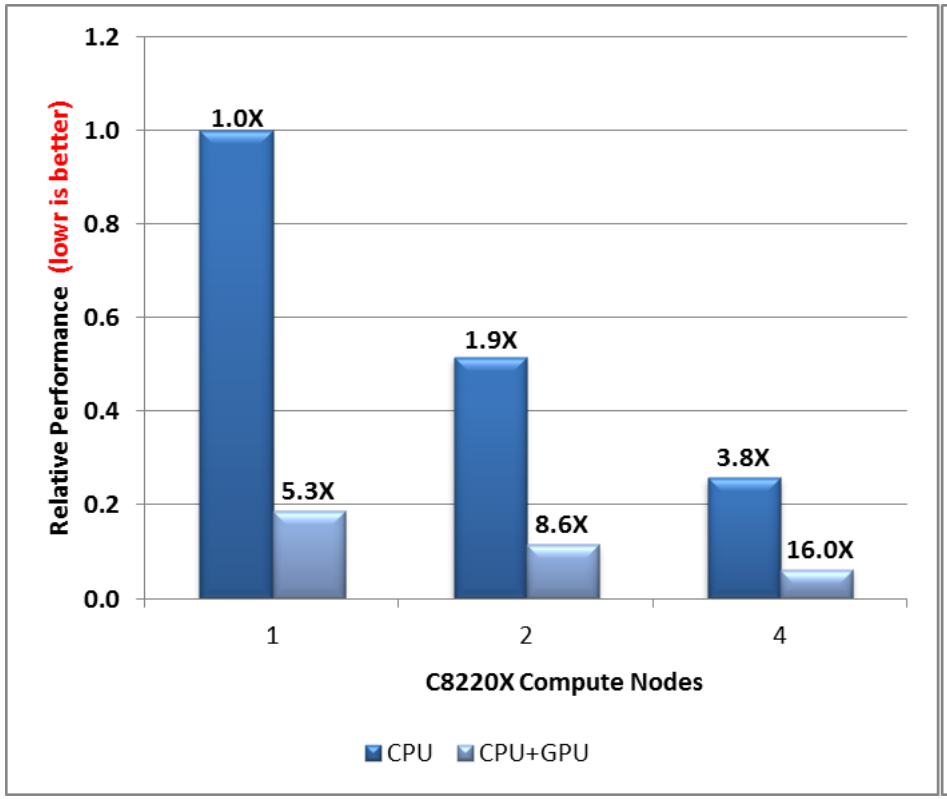


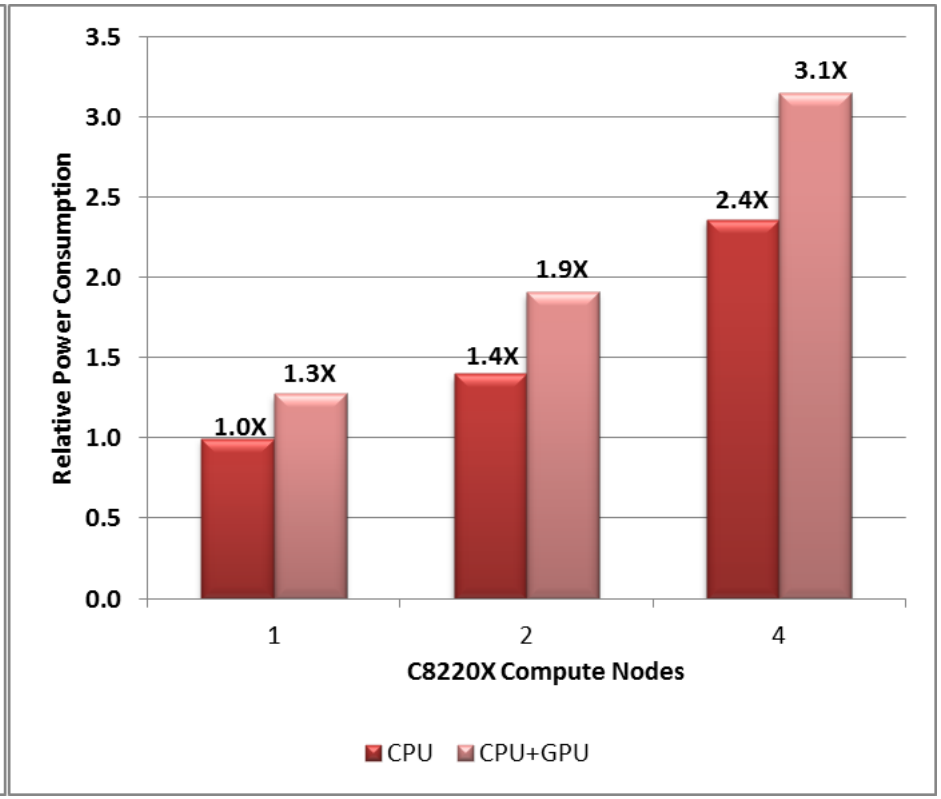


Acceleration is up to 7X for less than 2X power compared to CPUs

1-node PEC8220X, Dual E5-2670@2.6GHz, 128GB 1600MHz Memory; C8000 has 1 C8220X sled, CUDA 5.0.24 (304.54)

C8000+C8220X: Scalability
NAMD Perf. & Power (K20 vs. CPU)





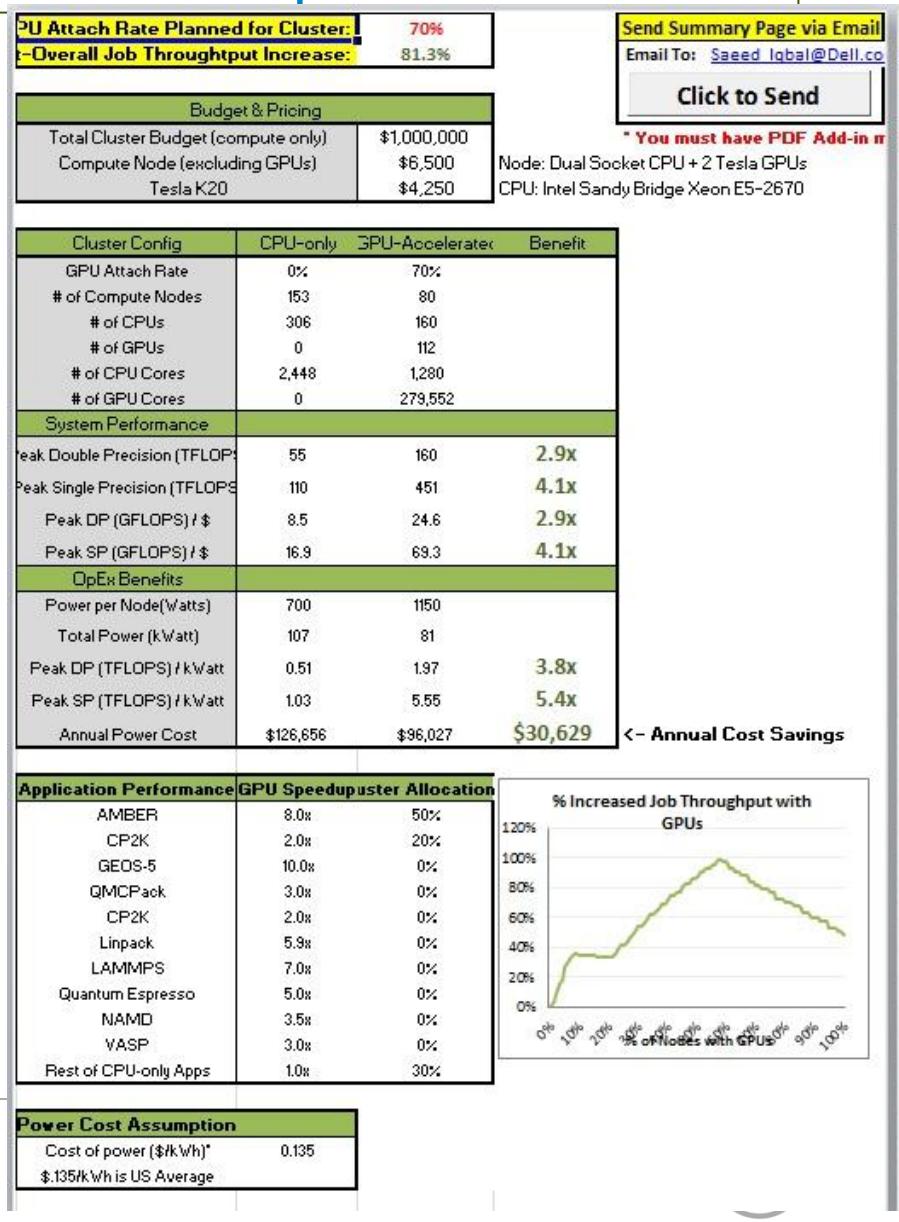
Acceleration is up to 5X for less than 1.4X power compared to CPUs

1-node PEC8220X, Dual E5-2670@2.6GHz, 128GB 1600MHz Memory; C8000 has 1 C8220X sled, CUDA 5.0.24 (304.54)

Solution

How many GPUs ?
NVIDIA has a tool to answer the question

- Current Customers may be running GPU enabled apps on CPU only systems.
- New Customers help maximizing return on investment.
- Inputs
  - Fixed Budget
  - Should I buy
    - CPU-only node
    - CPU+GPU node
  - Maximize throughput
    - Future change in application mix
    - Power Savings due to GPUs
- Please contact Dell Sales Contacts (Tool is not publically available.)



PU Attach Rate Planned for Cluster: 70%
-Overall Job Throughput Increase: 81.3%

Send Summary Page via Email
Email To: Saeed_Iqbal@Dell.co
Click to Send
* You must have PDF Add-in

Budget & Pricing	
Total Cluster Budget (compute only)	$1,000,000
Compute Node (excluding GPUs)	$6,500
Tesla K20	$4,250

Node: Dual Socket CPU + 2 Tesla GPUs
CPU: Intel Sandy Bridge Xeon E5-2670

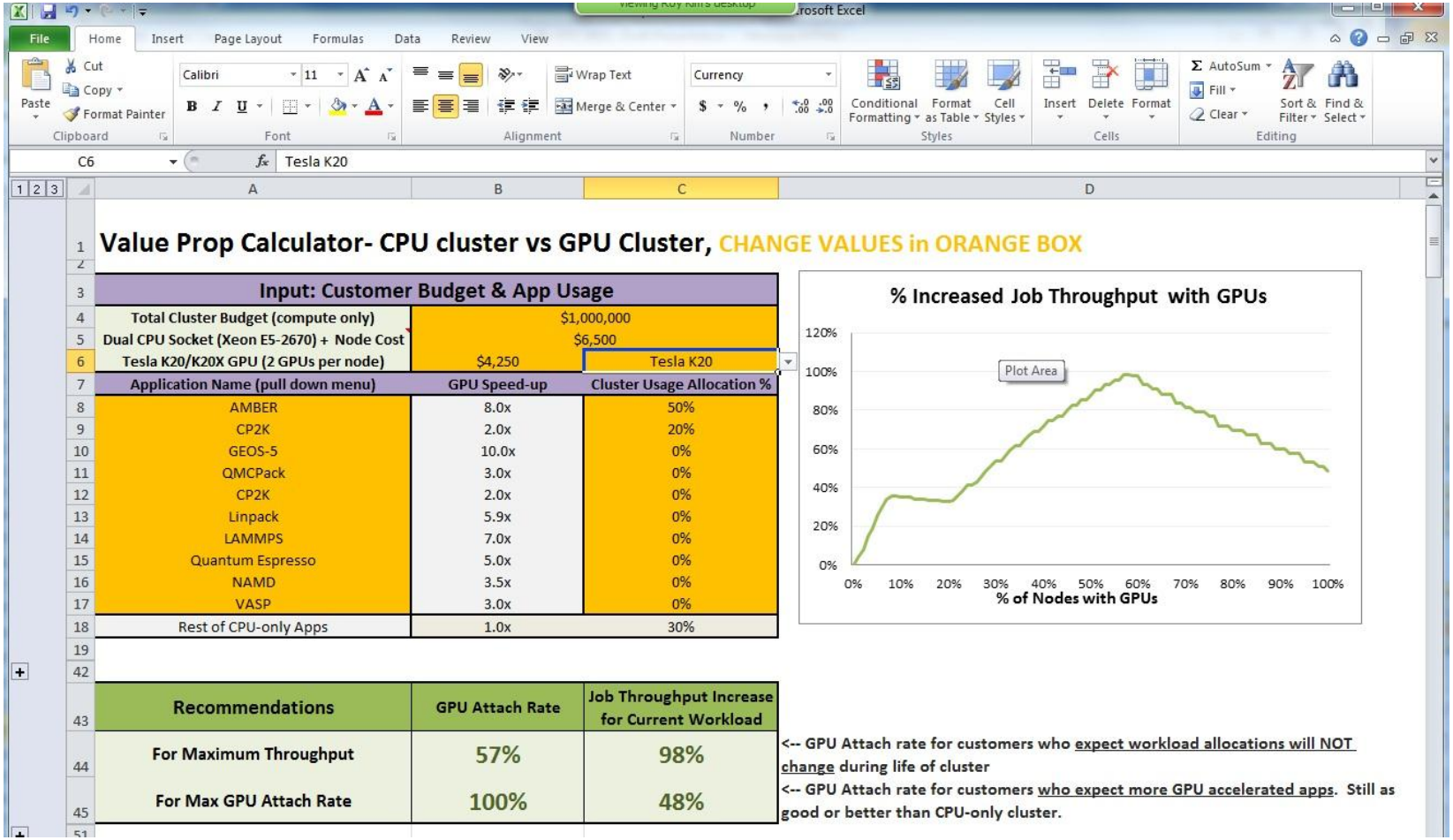
Cluster Config	CPU-only	GPU-Accelerated	Benefit
GPU Attach Rate	0%	70%	
# of Compute Nodes	153	80	
# of CPUs	306	160	
# of GPUs	0	112	
# of CPU Cores	2,448	1,280	
# of GPU Cores	0	279,552	
System Performance			
Peak Double Precision (TFLOPS	55	160	2.9x
Peak Single Precision (TFLOPS	110	451	4.1x
Peak DP (GFLOPS) / $	8.5	24.6	2.9x
Peak SP (GFLOPS) / $	16.9	69.3	4.1x
OpEx Benefits			
Power per Node(Watts)	700	1150	
Total Power (kWatt)	107	81	
Peak DP (TFLOPS) / kWatt	0.51	1.97	3.8x
Peak SP (TFLOPS) / kWatt	1.03	5.55	5.4x
Annual Power Cost	$126,656	$96,027	$30,629

<- Annual Cost Savings

Application Performance	GPU Speedup	uster Allocation
AMBER	8.0x	50%
CP2K	2.0x	20%
GEOS-5	10.0x	0%
QMCPack	3.0x	0%
CP2K	2.0x	0%
Linpack	5.9x	0%
LAMMPS	7.0x	0%
Quantum Espresso	5.0x	0%
NAMD	3.5x	0%
VASP	3.0x	0%
Rest of CPU-only Apps	1.0x	30%

Power Cost Assumption	
Cost of power ($/kWh)*	0.135
$.135/kWh is US Average	

How many GPUs ?
NVIDIA has a tool to answer the question



Value Prop Calculator- CPU cluster vs GPU Cluster, CHANGE VALUES in ORANGE BOX

Input: Customer Budget & App Usage		
Total Cluster Budget (compute only)	$1,000,000	
Dual CPU Socket (Xeon E5-2670) + Node Cost	$6,500	
Tesla K20/K20X GPU (2 GPUs per node)	$4,250	Tesla K20

Application Name (pull down menu)	GPU Speed-up	Cluster Usage Allocation %
AMBER	8.0x	50%
CP2K	2.0x	20%
GEOS-5	10.0x	0%
QMCPack	3.0x	0%
CP2K	2.0x	0%
Linpack	5.9x	0%
LAMMPS	7.0x	0%
Quantum Espresso	5.0x	0%
NAMD	3.5x	0%
VASP	3.0x	0%
Rest of CPU-only Apps	1.0x	30%

Recommendations	GPU Attach Rate	Job Throughput Increase for Current Workload
For Maximum Throughput	57%	98%
For Max GPU Attach Rate	100%	48%

<-- GPU Attach rate for customers who expect workload allocations will NOT change during life of cluster

<-- GPU Attach rate for customers who expect more GPU accelerated apps. Still as good or better than CPU-only cluster.

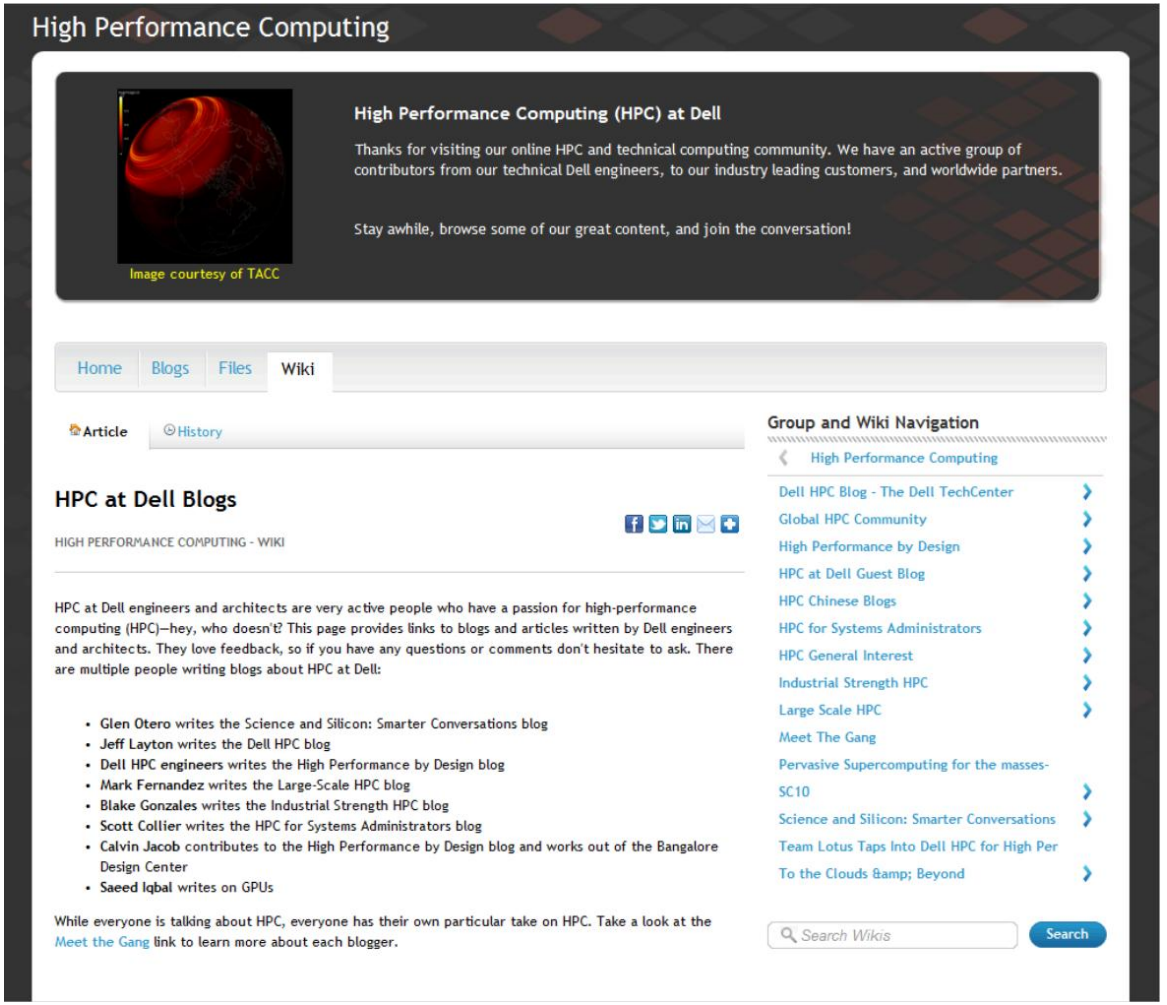
Outputs: **Given a Budget & "Application Mix" Maximize the throughput**

Note: Please contact Dell Sales Contacts (Tool is not publically available).

Summary

Resources

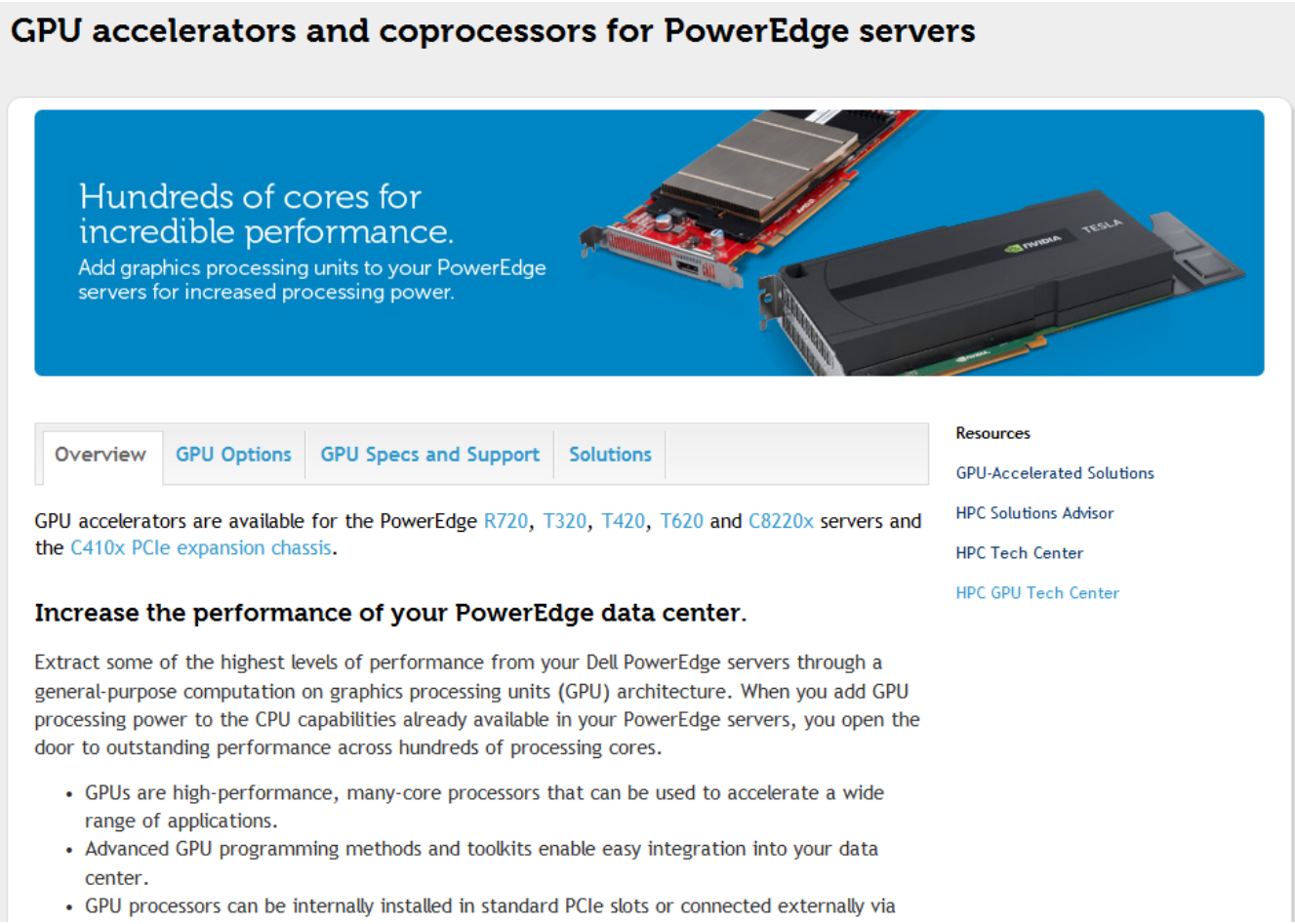
- Blogs
- Whitepapers
- HPC Advisor Online



- www.dell.com/gpu
- www.dell.com/hpc
- www.hpcatdell.com
- www.DellHPCSolutions.com
- http://www.hpcadvisorycouncil.com/best_practices.php

Resources: www.dell.com/gpu

- Overview
- Supported GPUs
- GPU Specs
- GPU Solutions



GPU accelerators and coprocessors for PowerEdge servers

Hundreds of cores for incredible performance.

Add graphics processing units to your PowerEdge servers for increased processing power.

Overview | GPU Options | GPU Specs and Support | Solutions

GPU accelerators are available for the PowerEdge R720, T320, T420, T620 and C8220x servers and the C410x PCIe expansion chassis.

Increase the performance of your PowerEdge data center.

Extract some of the highest levels of performance from your Dell PowerEdge servers through a general-purpose computation on graphics processing units (GPU) architecture. When you add GPU processing power to the CPU capabilities already available in your PowerEdge servers, you open the door to outstanding performance across hundreds of processing cores.

- GPUs are high-performance, many-core processors that can be used to accelerate a wide range of applications.
- Advanced GPU programming methods and toolkits enable easy integration into your data center.
- GPU processors can be internally installed in standard PCIe slots or connected externally via

Resources

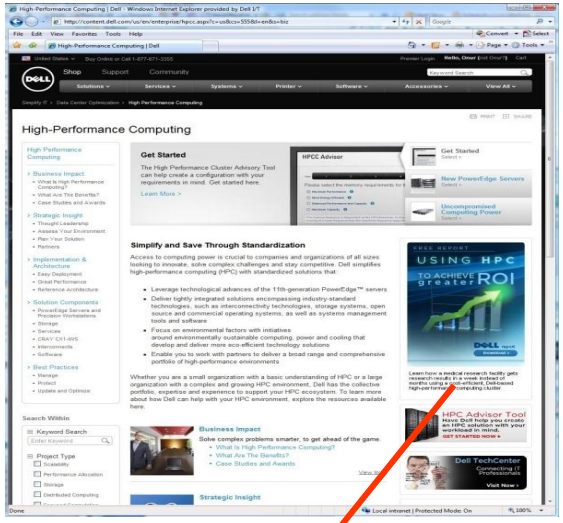
GPU-Accelerated Solutions

HPC Solutions Advisor

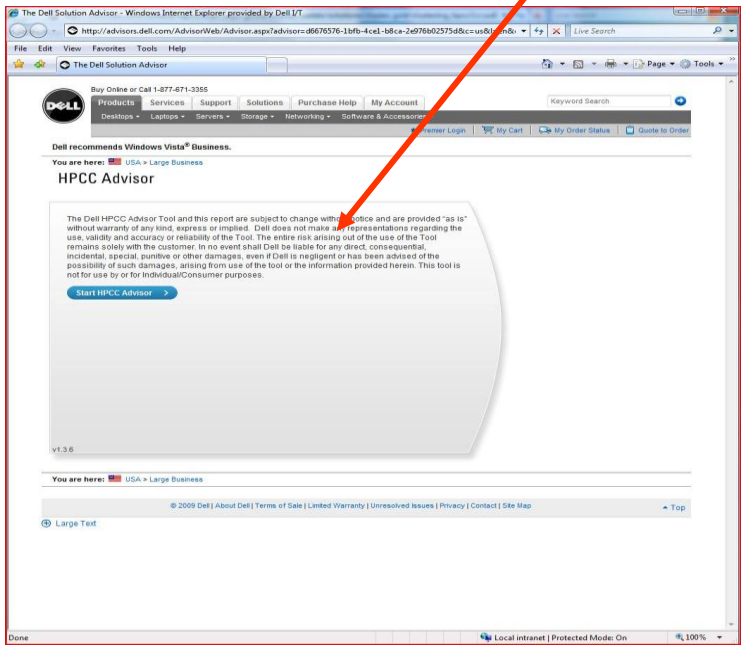
HPC Tech Center

HPC GPU Tech Center

Resources: HPC Advisor - Tool



- Software application that recommends the best fit Dell products and solutions based on customers specific needs
- Goal: Create balanced cluster designs using best practices.
- Example: The HPC Advisor asks user:
  - OS type?
  - Optimize for performance, power or density
  - Desired sustained or theoretical performance
  - Recommends a solution based on this input.



Available on Dell.com http://dell.com/hpc

Summary of the Key Features of HPC Solutions

- **Balanced** (GPUs, Compute, Storage, Networking)
- **Powerful** (1 or 2 GPU/U Density)
- **Adaptable** (workload based configuration)
- **Flexible** (modular components)
- **Scalable** (Modular building blocks)
- **Efficient** (Compared to equivalent CPU only clusters)

Start Small,

Grow and Adapt your HPC solution based on your needs!