NVIDIA Tesla® K20-K20X GPU Accelerators - Benchmarks

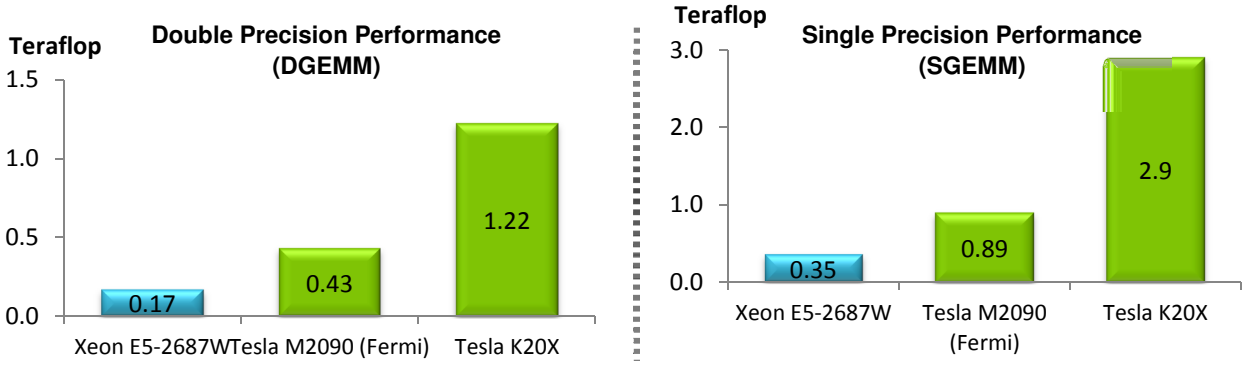
Application Performance Technical Brief

NVIDIA® changed the high performance computing (HPC) landscape by introducing its Fermi-based GPUs that delivered an impressive leap in performance and energy-efficiency while offering a parallel programming model, CUDA ®, as an extension to industry standard languages like C, C++, and Fortran.

Now NVIDIA has introduced the new Tesla® K20-K20X GPU Accelerators that further advance the HPC industry. Built on the revolutionary Kepler™ architecture, these accelerators redefine the standard for energy-efficient computing and feature innovative technologies like SMX, Hyper-Q, and Dynamic Parallelism to boost application performance by up to 10x.

SMX: 3x More Performance Per Watt

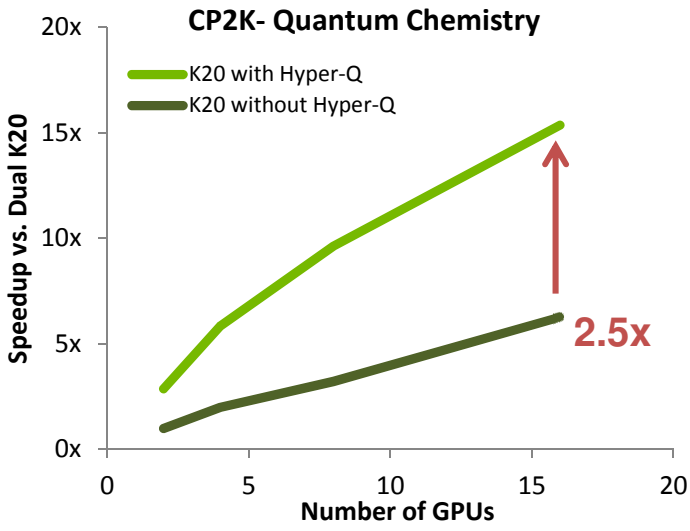
The new SMX (Next Generation Streaming Multiprocessor) is an architectural innovation designed from the ground-up to deliver high efficiency performance. With SMX at its core, Tesla K20/K20X accelerators deliver the industry's highest single and double precision performance- 3.95 teraflops and 1.31 teraflops respectively for Tesla K20X- at an unprecedented 93% computational efficiency.



Hyper-Q: Easy Speed-up for Legacy MPI Codes

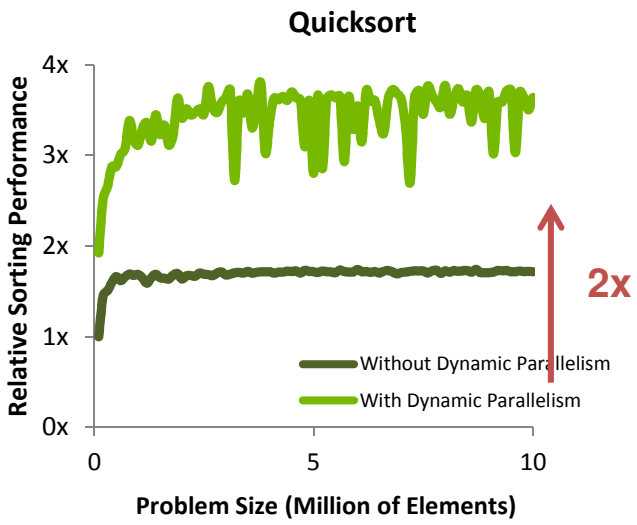
Many legacy MPI codes don't generate enough work to fully occupy the GPU because they were written for CPU cores. Rather than requiring developers to refactor their codes to put more workload per MPI process, the Hyper-Q feature reduces efforts considerably because developers can now throw up to 32 MPI processes with small- and medium-sized workloads at a shared GPU.

To illustrate the power of Hyper-Q, we picked a traditionally difficult code for GPUs called CP2K, a popular MPI-based quantum chemistry code, showing more than 2x performance improvement when 16 MPI ranks are run on the shared GPU.



Dynamic Parallelism: Simplifying Parallel Programming

Dynamic Parallelism allows the GPU to operate more autonomously from the CPU by generating new work for itself at run-time, from inside a kernel. The concept is simple, but the impact is powerful: it can make programming easier, particularly for algorithms traditionally considered difficult such as divide-and-conquer problems.



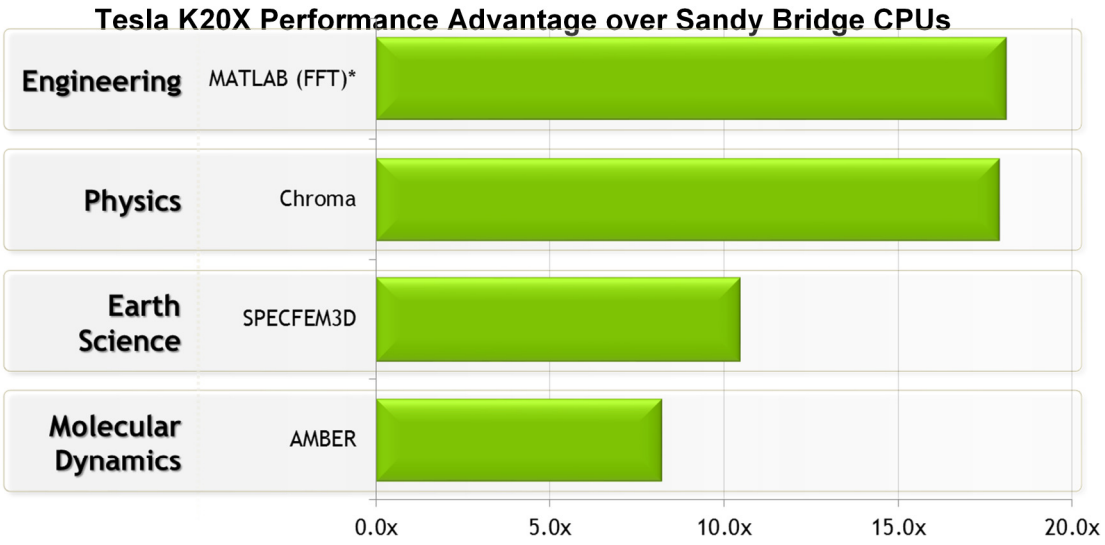
To showcase its potential, we used Dynamic Parallelism on Quicksort, a well-known algorithm for sorting methods, to reduce the lines of CUDA code in half while improving performance by 2x.

K20X Benchmarks by Application

Accelerating Key Scientific Applications by up to 10x

Today, hundreds of applications take advantage of GPU acceleration, spanning all scientific disciplines and engineering domains, and the number of applications continues to grow. In the past year alone, the number of CUDA-accelerated applications has grown by over 60%.

When Tesla K20 GPU Accelerators are added to servers with Sandy Bridge CPUs, CUDA-enabled applications are typically accelerated up to 10x. The K20X benchmark below shows single node performance of leading applications in various science domains.

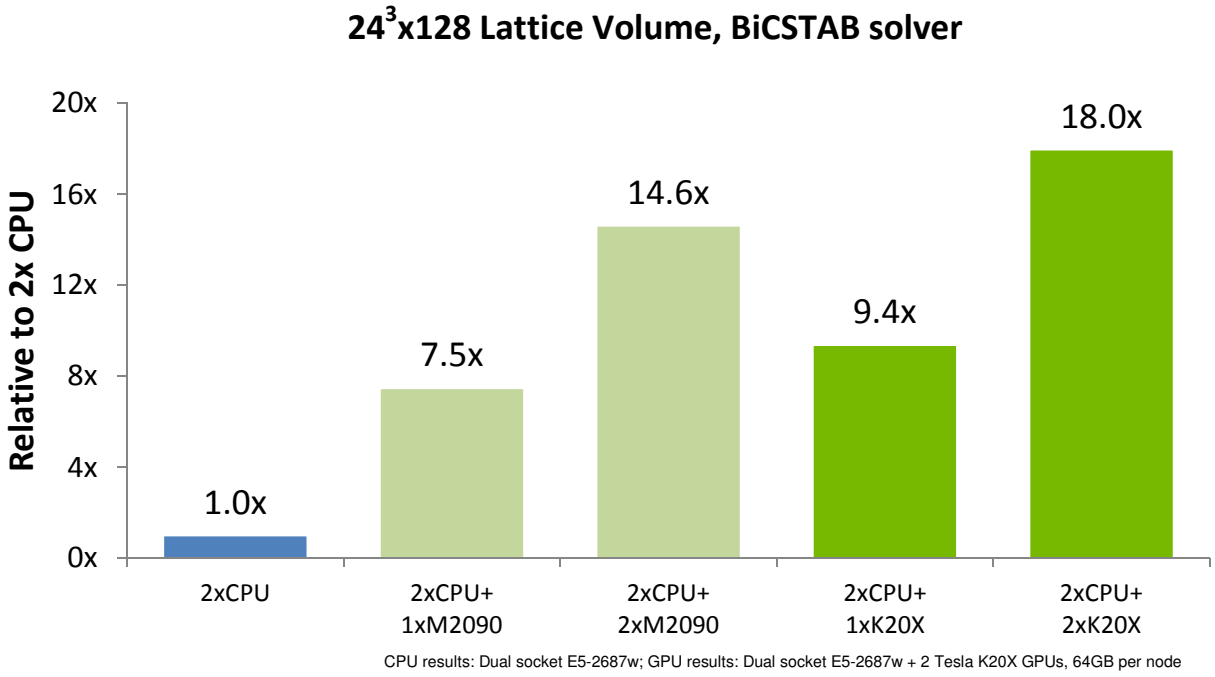


CPU results: Dual socket E5-2687w; GPU results: Dual socket E5-2687w + 2 Tesla K20X GPUs
*MATLAB results comparing one i7-2600K CPU vs. Tesla K20 GPU

Here's a closer look at three of the applications listed above.

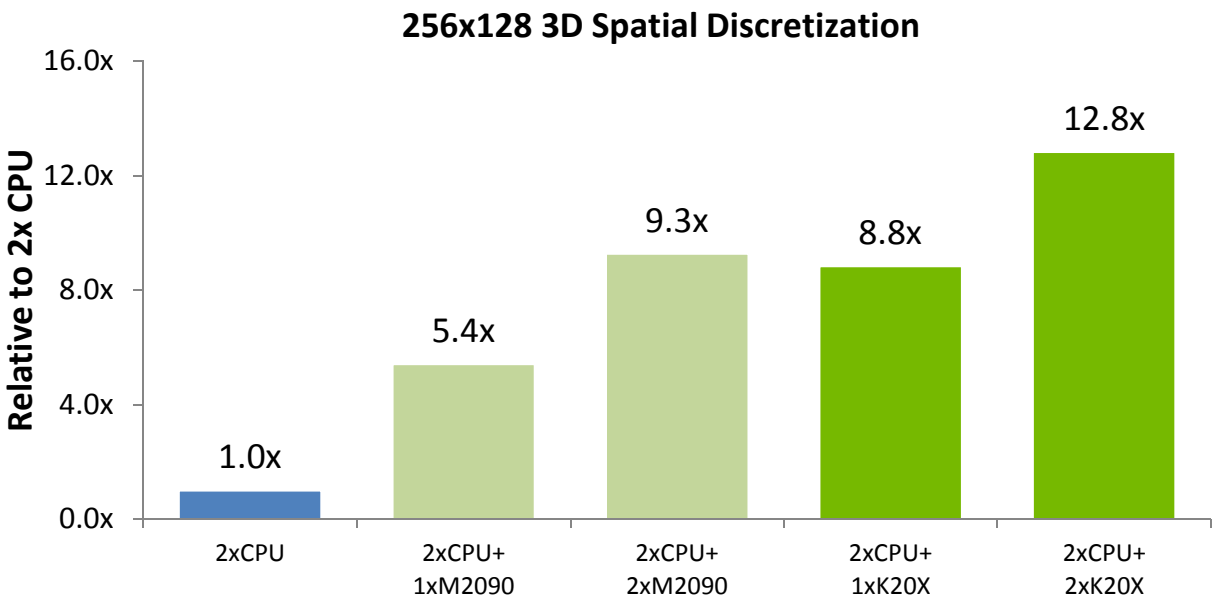
Chroma: High Energy & Nuclear Physics

Chroma is used by scientists to test alternatives to the Standard Model of physics in order to develop a deeper understanding of the fundamental properties of matter and energy. The benchmark below show significant performance boost when adding one or two Tesla K20X GPU Accelerators to dual socket CPUs.



SPECFEM3D: Earth Sciences

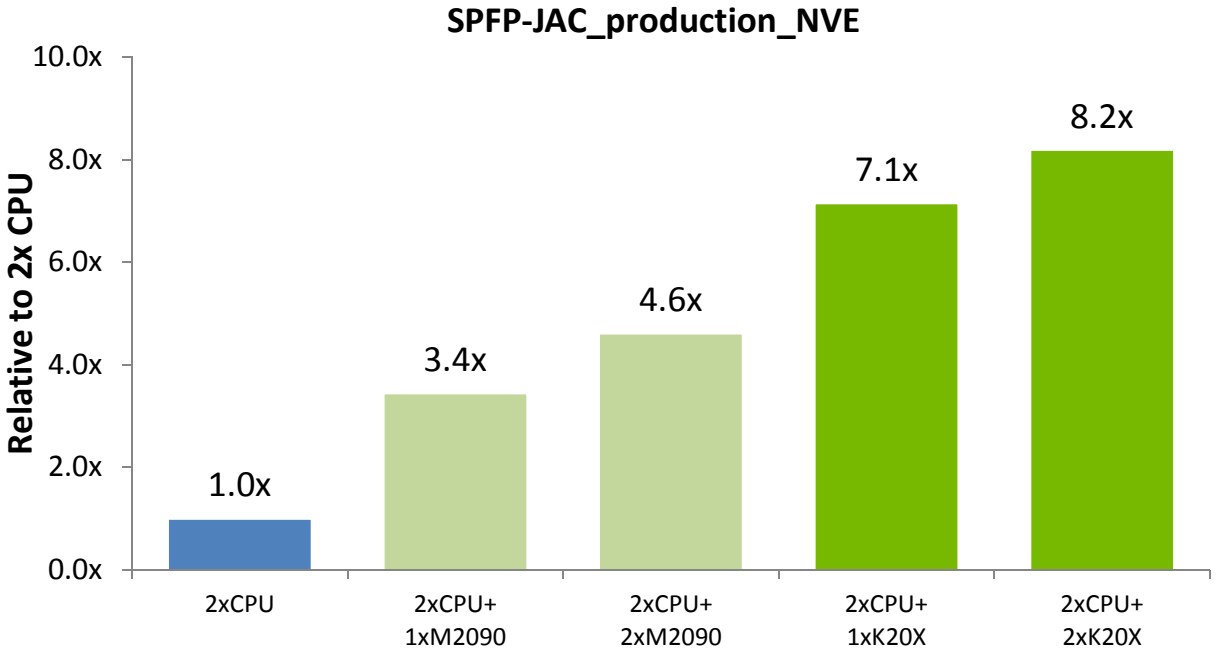
Researchers use SPECFEM3D for a deeper understanding of the phenomena that shape seismic activity, thereby allowing engineers to assess potential hazards and to build structural countermeasures. This code was a Gordon Bell winner in 2008 prior to the GPU work, so it was a highly tuned code even before the recent work to accelerate on GPUs.



CPU results: Dual socket E5-2687w; GPU results: Dual socket E5-2687w + 2 Tesla K20X GPUs, 64GB per node

AMBER: Molecular Dynamics

Molecular dynamics (MD) allows the study of biological and chemical systems at the atomistic level on timescales from femtoseconds to milliseconds. Numerous software packages exist for conducting MD simulations of which one of the most widely used is AMBER. With the Tesla K20X GPU Accelerators, acceleration by up to 80% is observed on AMBER, over compared to Tesla M2090 GPUs.



CPU results: Dual socket E5-2687w; GPU results: Dual socket E5-2687w + 2 Tesla K20X GPUs, 64GB per node

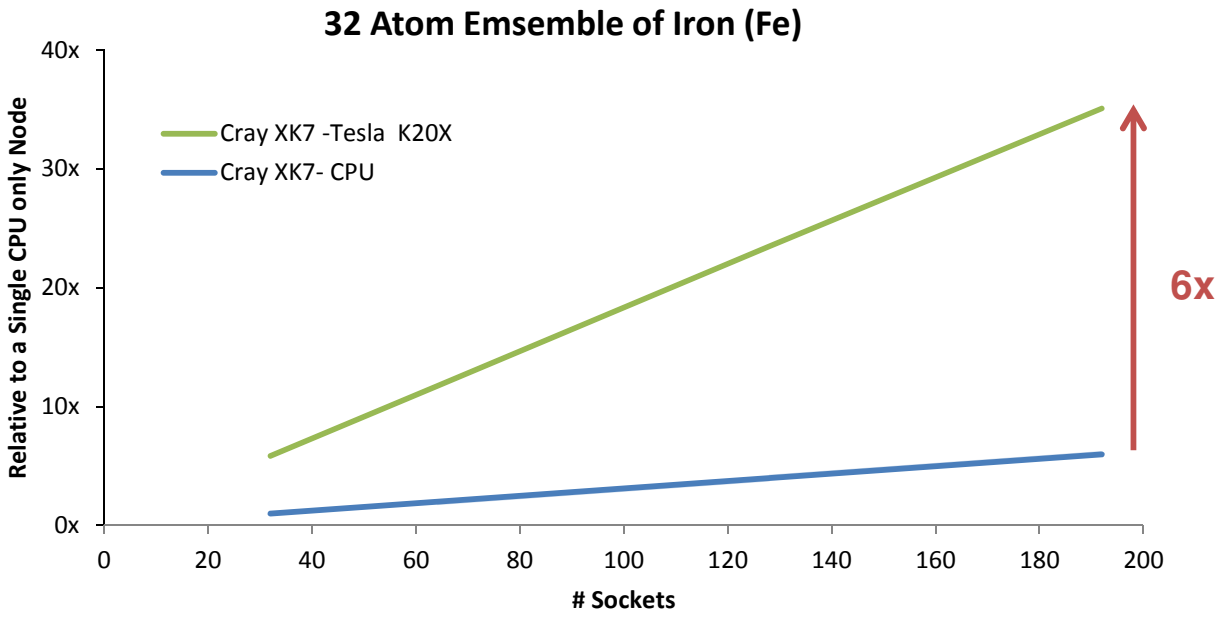
Large Cluster Scaling for GPU-Accelerated Applications

Large clusters around the world are increasingly deploying GPUs to accelerate their workload. In the Top500 list, from June 2011 to June 2012, number of GPU-accelerated systems grew by over 400%.

Whether an application is GPU-accelerated or not, designing code to scale well across many nodes has never been an easy task. However, many GPU-accelerated applications are scaling well as developers work on extracting more parallelism by allocating more parallel work within a node and use MPI to distribute GPU workloads.

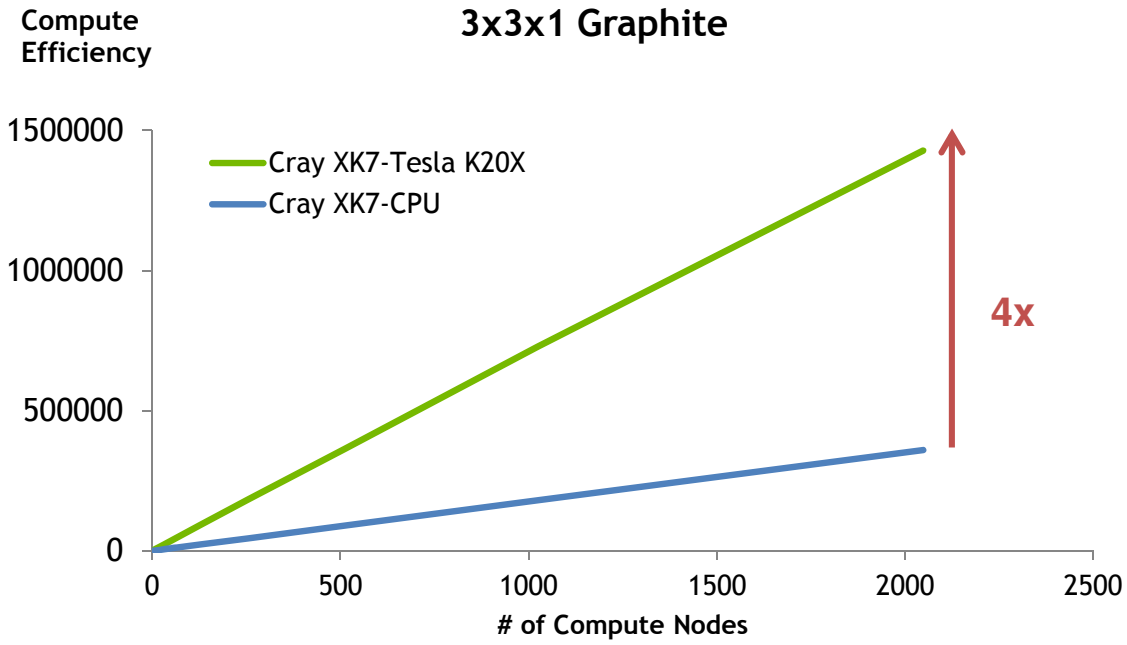
WL-LSMS: Material Science

WL-LSMS simulates the magnetic behavior of materials at the atomic and nanoscale in order to design lightweight, powerful magnetic components for highly efficient electric motors, generators, and magnetic storage devices. WL-LSMS was a Gordon Bell winner in 2009, so it was already a highly-tuned CPU code prior to the recent work to accelerate on GPUs.



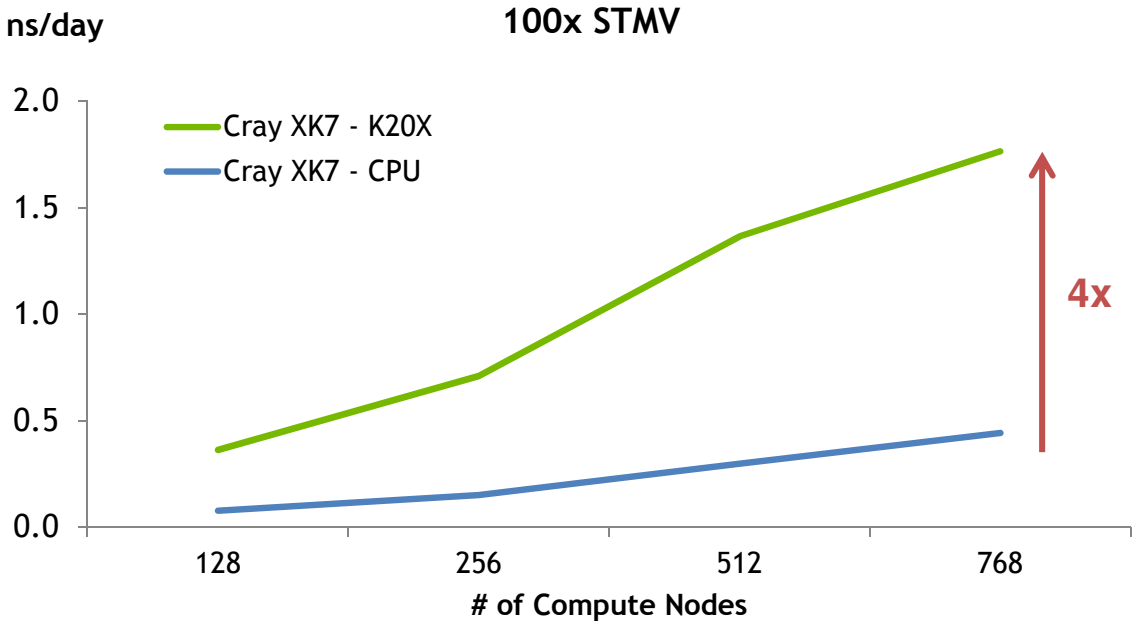
QMCPACK: Material Science

QMCPACK is used by researchers to develop new insights into condensed matter physics, materials science, and chemistry by using more physically accurate methods. For large systems on the order of 200 electrons or more, simulations that traditionally would take upwards of 12 hours or more can now run on the order of 3 hours per simulation.



NAMD: Molecular Dynamics

NAMD (Not (just) Another Molecular Dynamics program) is a free-of-charge molecular dynamics simulation package written using the Charm++ parallel programming model, noted for its parallel efficiency and often used to simulate large systems (millions of atoms). NAMD also scales well across many nodes accelerated with GPUs.



For more information on Tesla K20/K20X GPU Accelerators, visit www.nvidia.com/tesla.

Notice

Copyright

www.nvidia.com