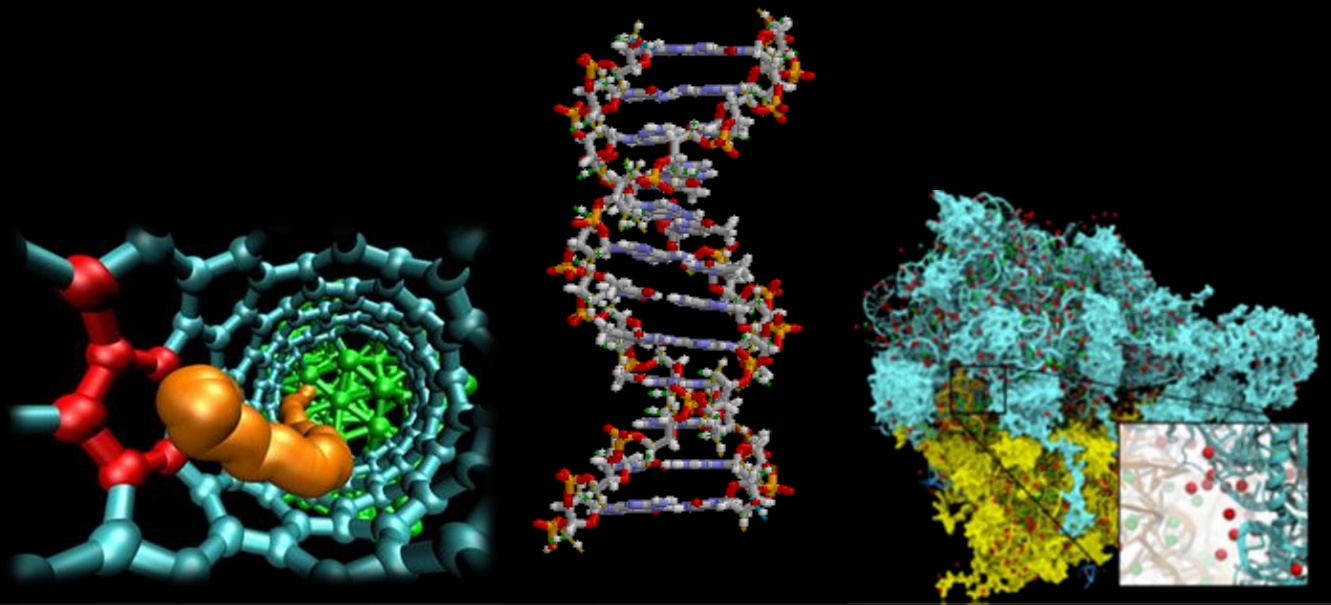
NVIDIA Computational Chemistry & Biology



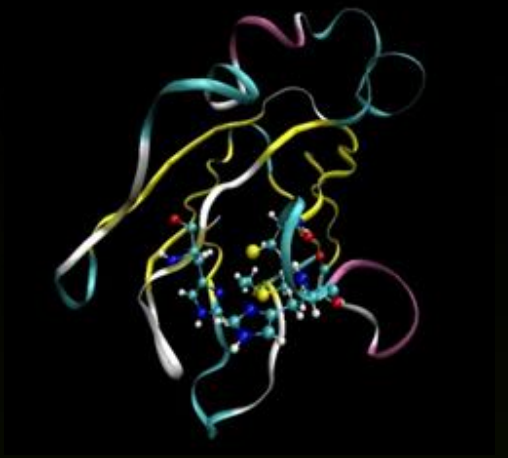
Mark Berger
Senior Alliance Manager
Life and Material Sciences
mberger@nvidia.com

Updated: June 16, 2015

Overview of Life & Material Accelerated Apps

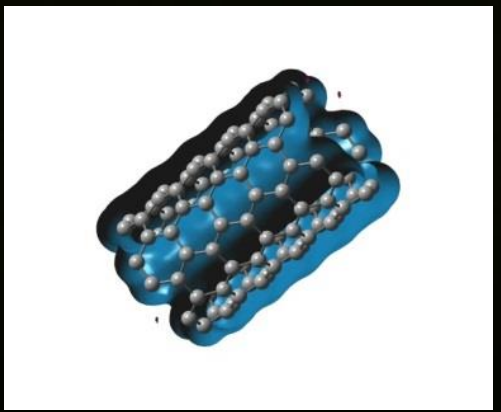
MD: All key codes are GPU-accelerated

- ACEMD*, AMBER (PMEMD)*, BAND, CHARMM, DESMOND, ESPResso, Folding@Home, GPUgrid.net, GROMACS, HALMD, HOOMD-Blue*, LAMMPS, Lattice Microbes*, mdcore, NAMD, OpenMM, SOP-GPU*
- Great multi-GPU performance!
- Focus: on dense (up to 16) GPU nodes & large # of GPU nodes



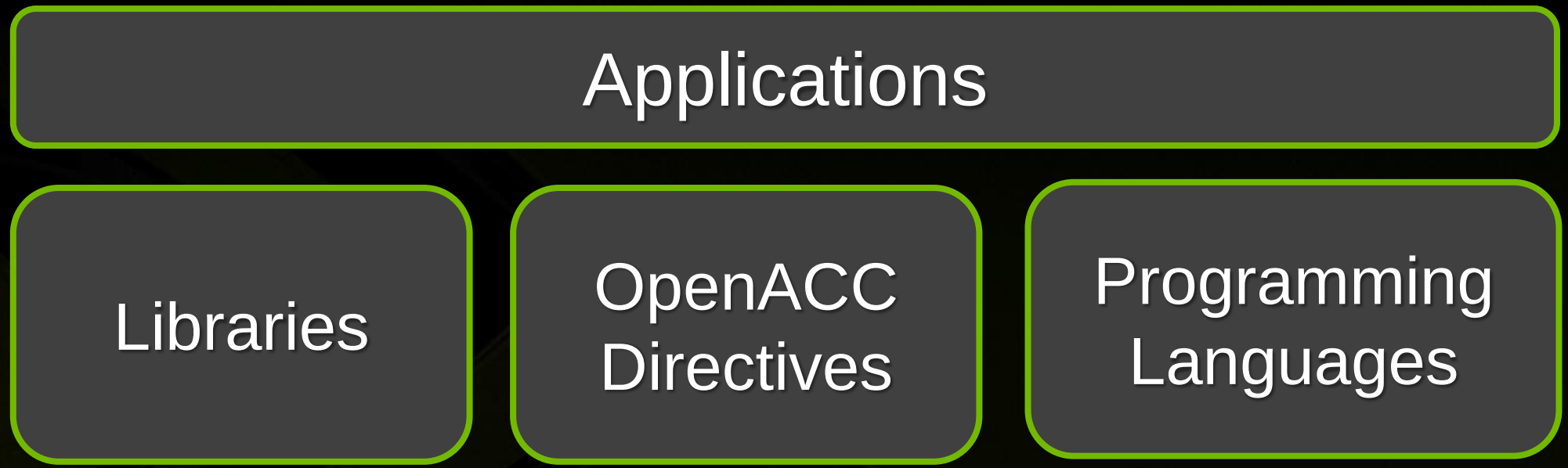
QC: All key codes are ported or optimizing:

- GPU-accelerated and available today:
  - ABINIT, ACES III, ADF, BigDFT, CP2K, GAMESS, Quantum Espresso/PWscf, MOLCAS, MOPAC2012, NWChem, OCTOPUS*, QUICK, Q-Chem, TeraChem*
- Active GPU acceleration projects:
  - CASTEP, CPMD, GAMESS, Gaussian, NWChem, ONETEP, Quantum Supercharger Library*, VASP & more
- Focus: on using GPU-accelerated math libraries, OpenACC directives



green* = application where all the workload is on GPU

3 Ways to Accelerate Applications

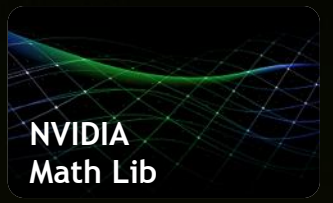


GPU Accelerated Libraries

"Drop-in" Acceleration for your Applications

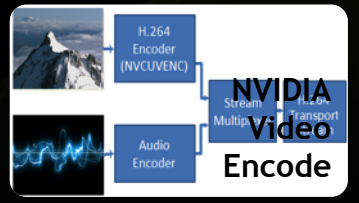
Linear Algebra
FFT, BLAS,
SPARSE, Matrix

Numerical & Math
RAND, Statistics



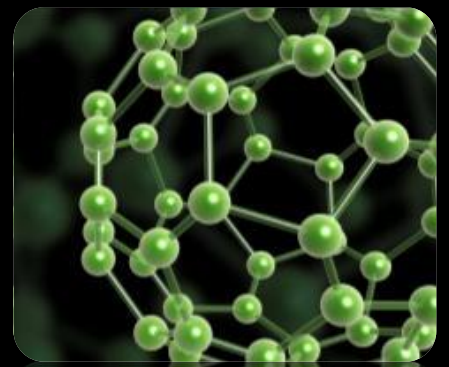
Data Struct. & AI
Sort, Scan, Zero Sum

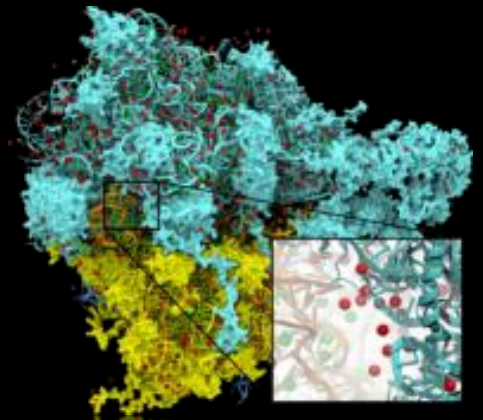
Visual Processing
Image & Video



TESLA

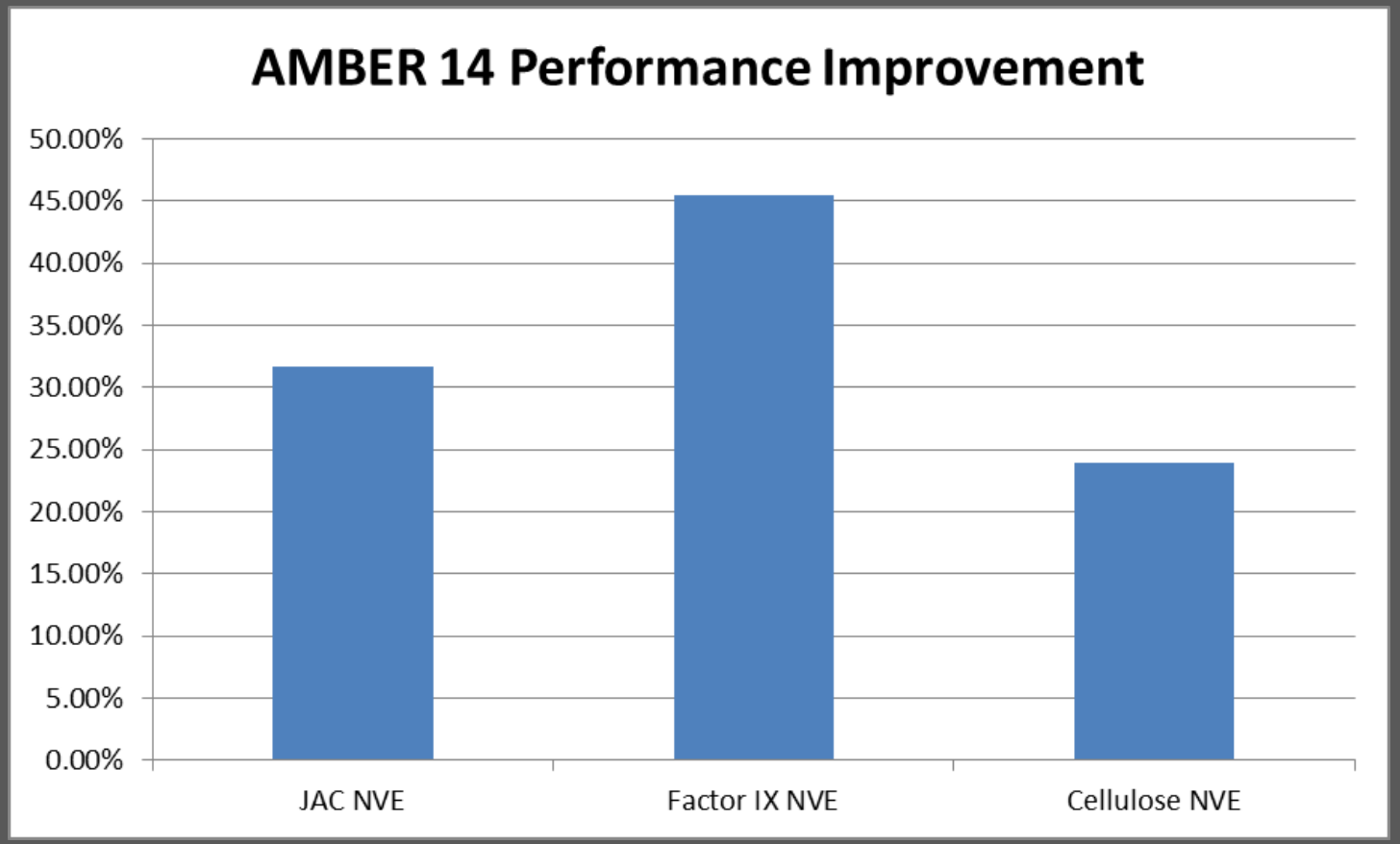
Molecular Dynamics Module





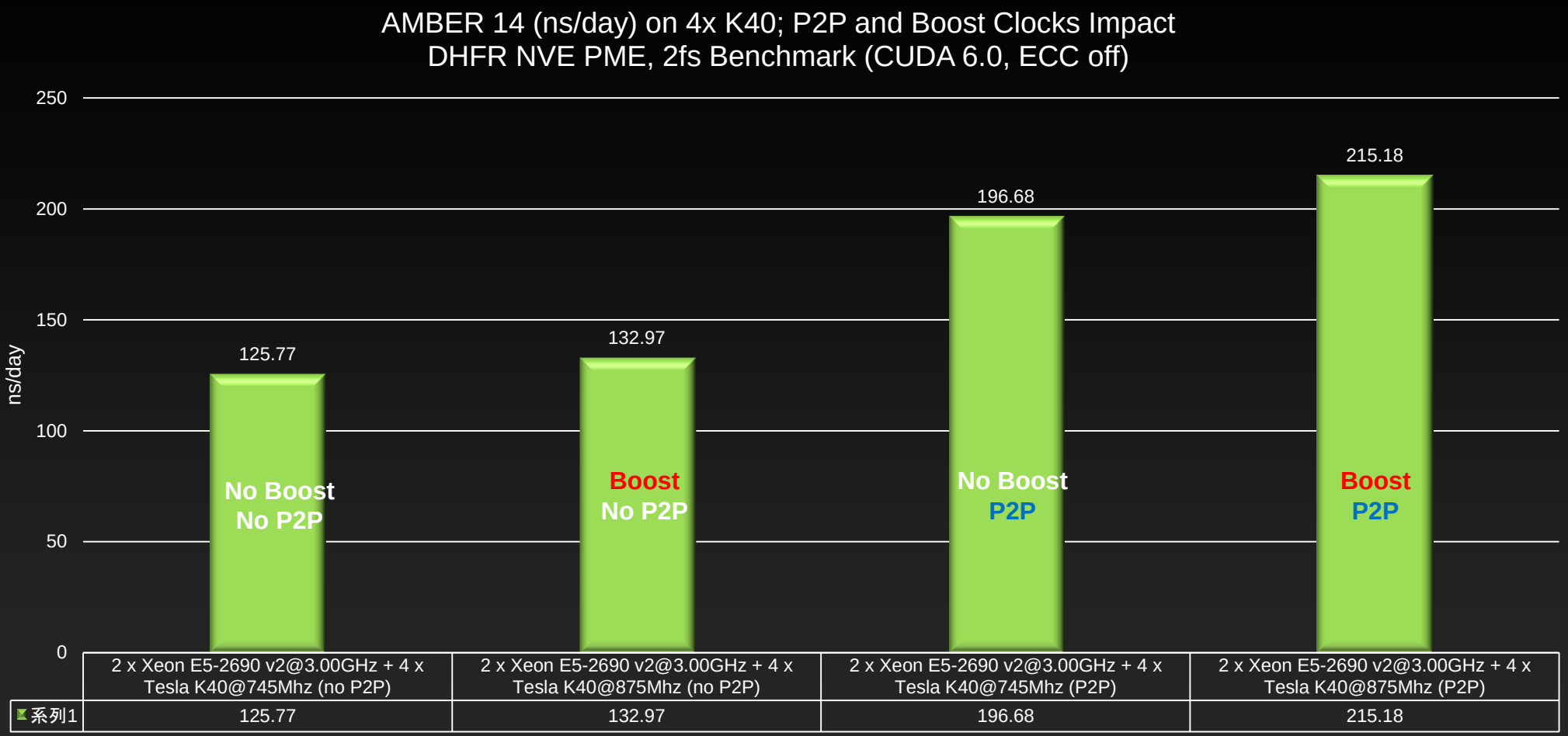
AMBER 14

Compared to AMBER 12

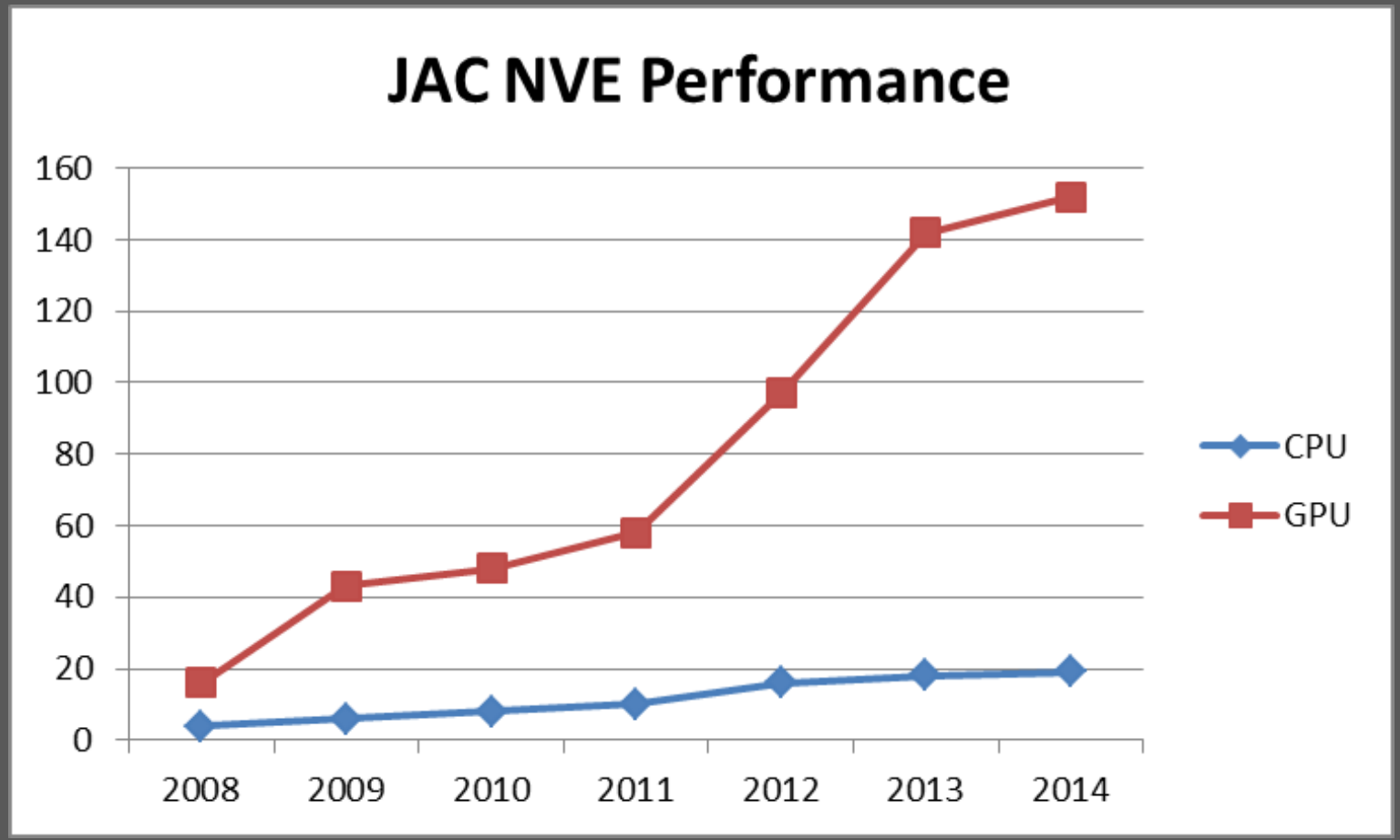


Courtesy of Scott Le Grand From GTC 2014 presentation

AMBER 14; large P2P and small Boost Clocks impacts

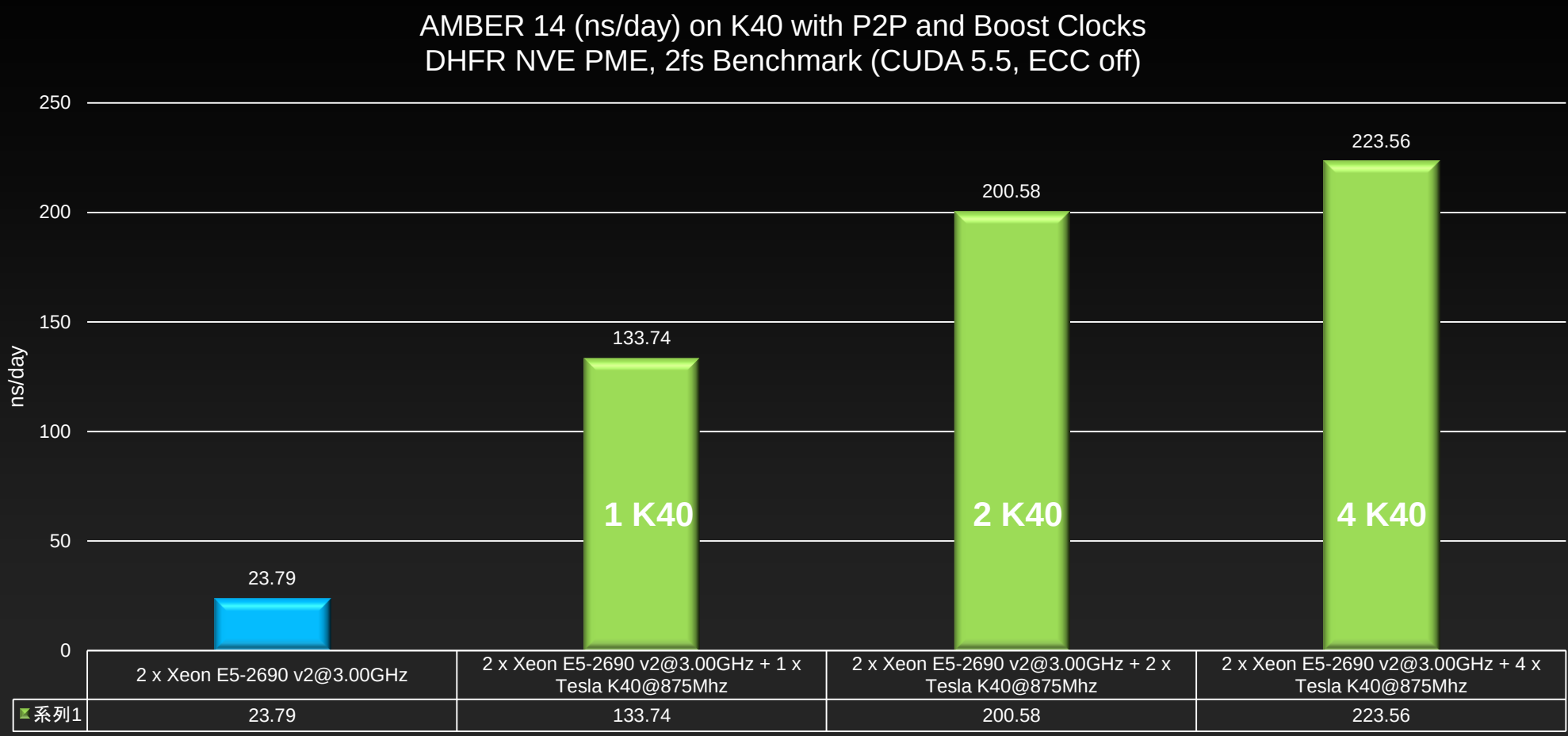


AMBER Performance

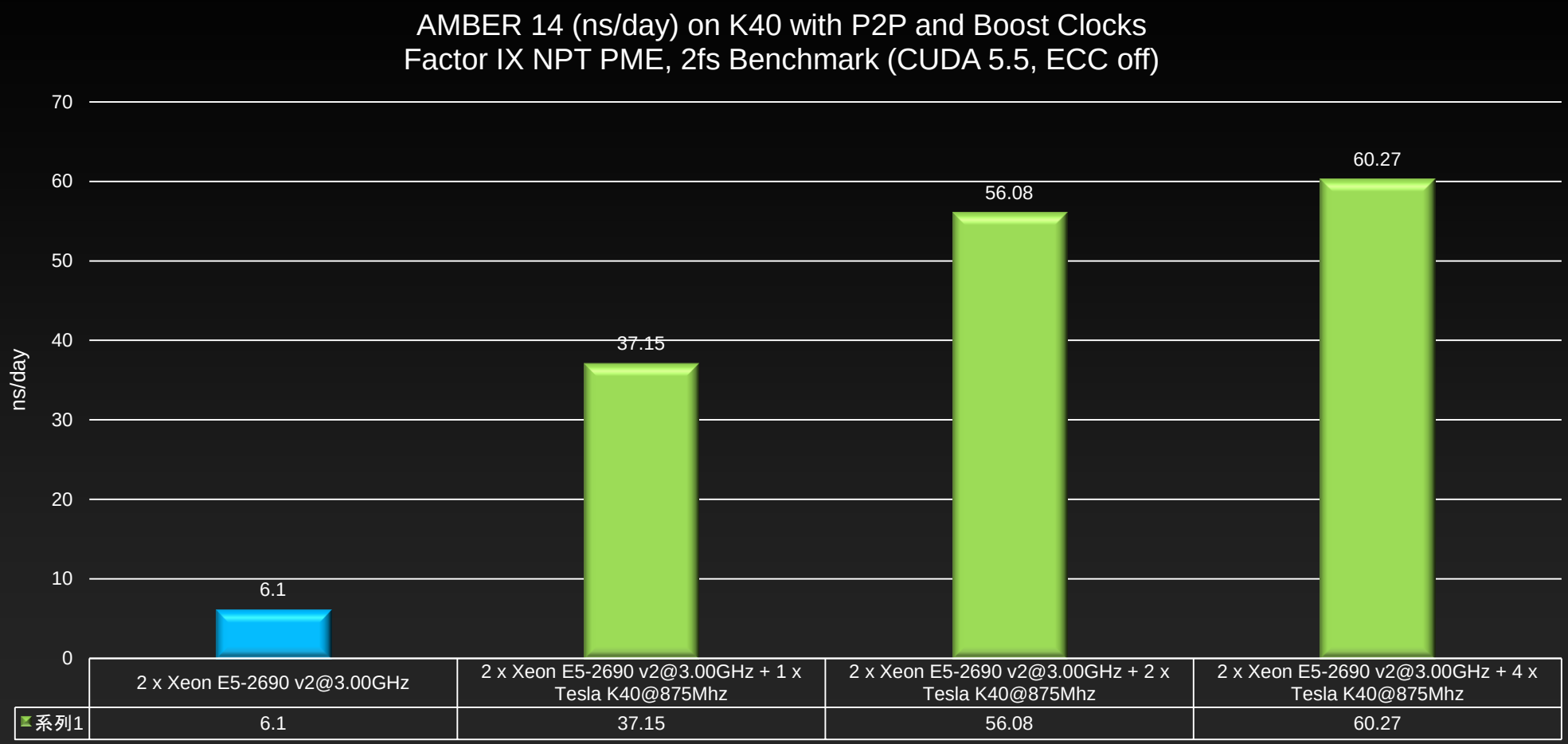


Courtesy of Scott Le Grand From GTC 2014 presentation

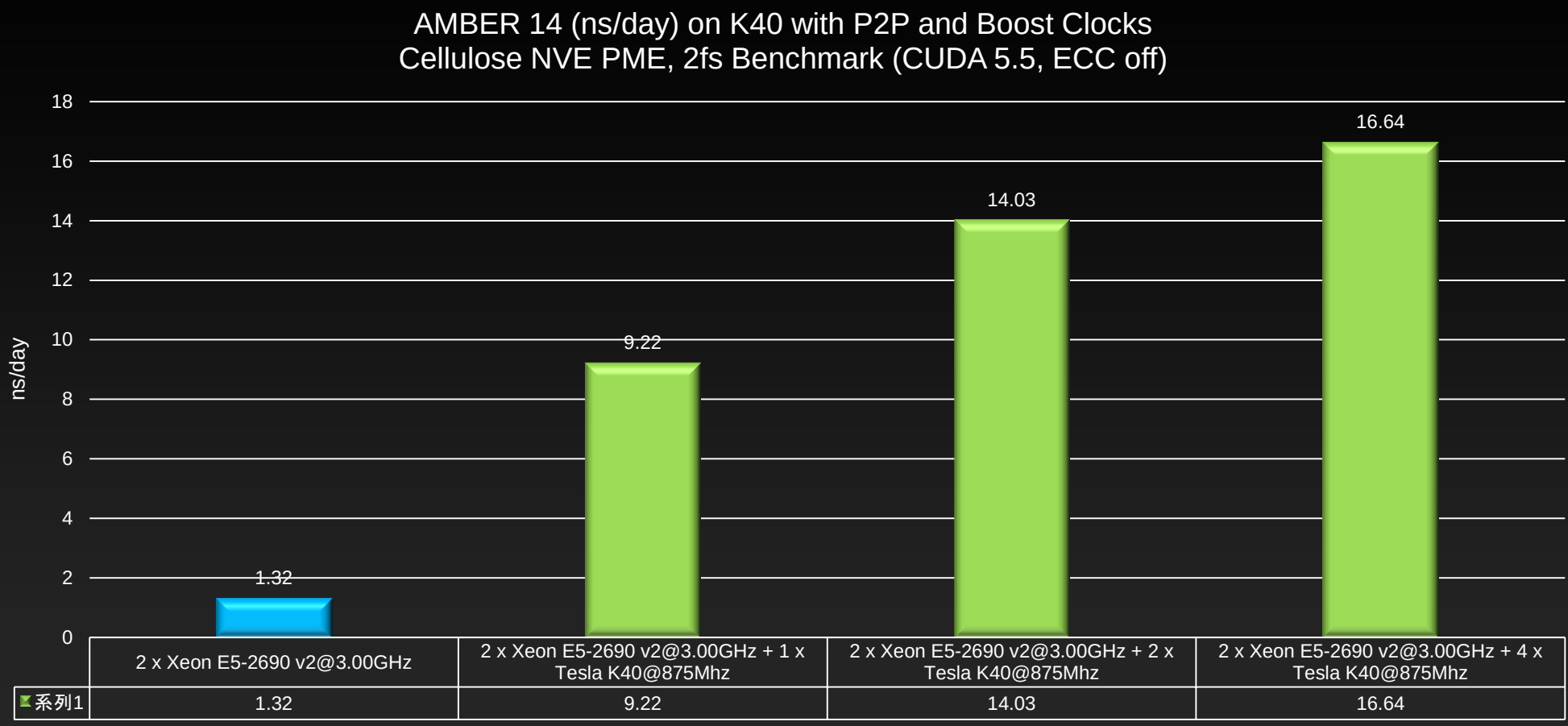
AMBER 14 with P2P, Higher Density Nodes



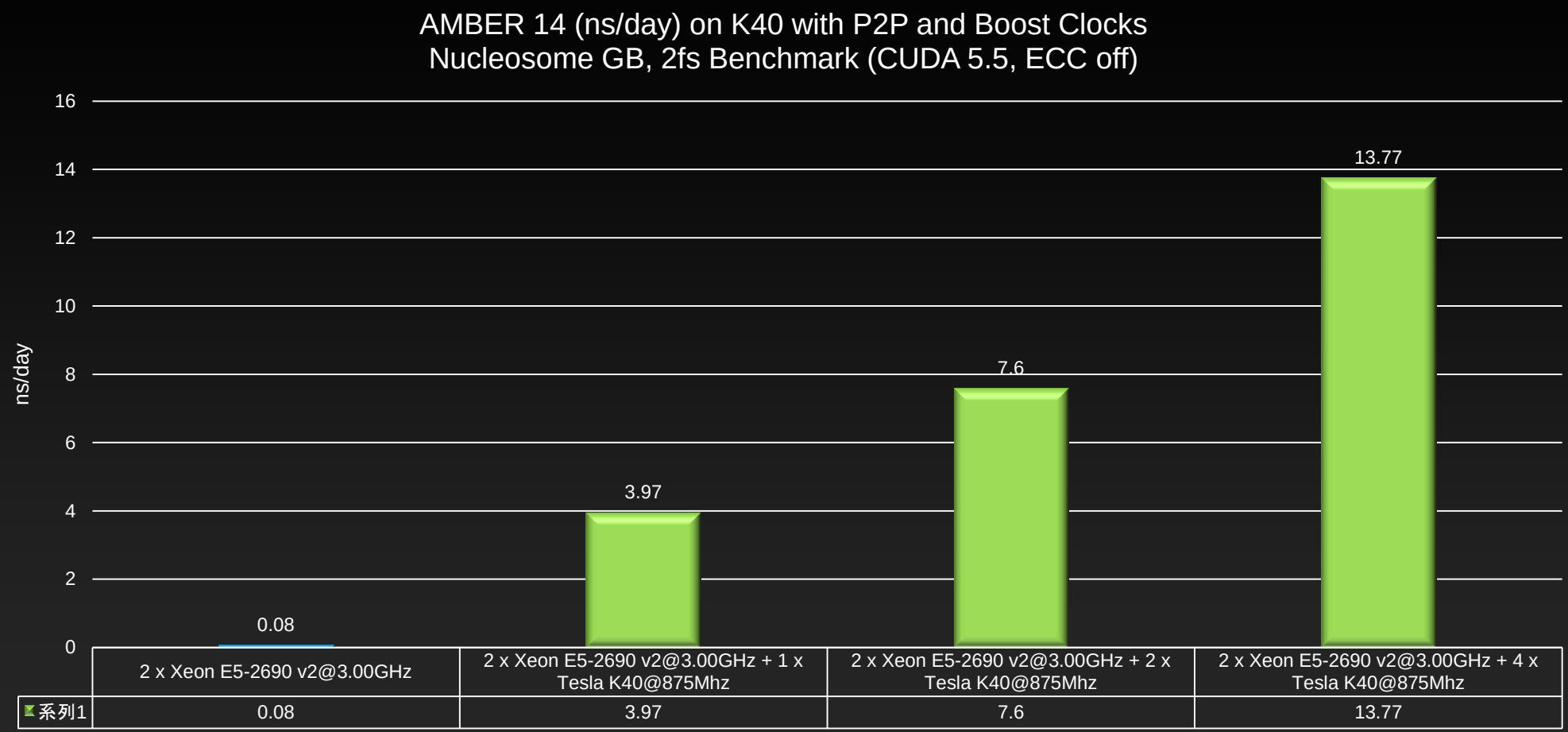
AMBER 14 and K40 with P2P, fastest GPU yet!



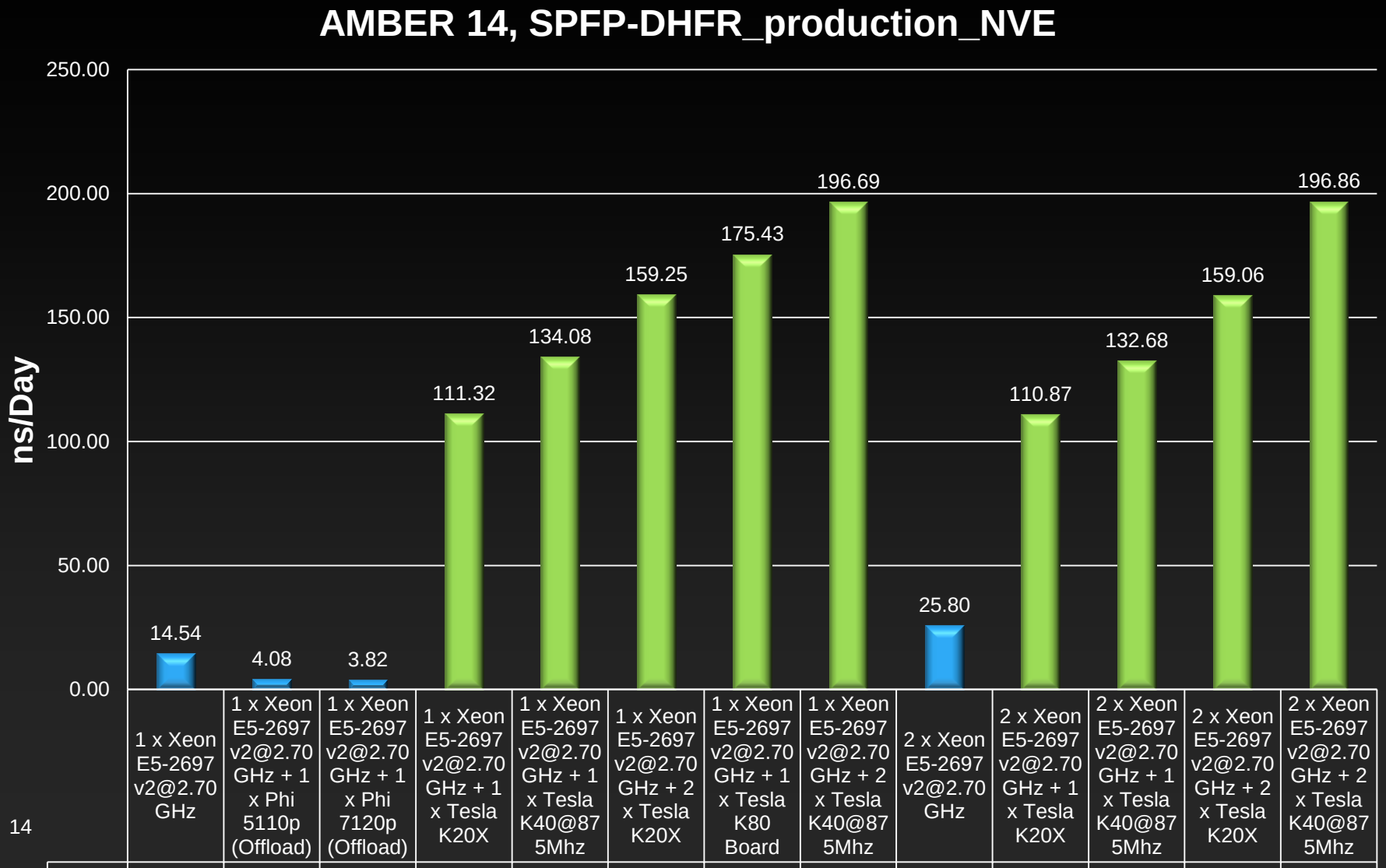
AMBER 14 and K40 with P2P, fastest GPU yet!



AMBER 14 and K40 with P2P, fastest GPU yet!



Kepler - Our Fastest Family of GPUs Yet

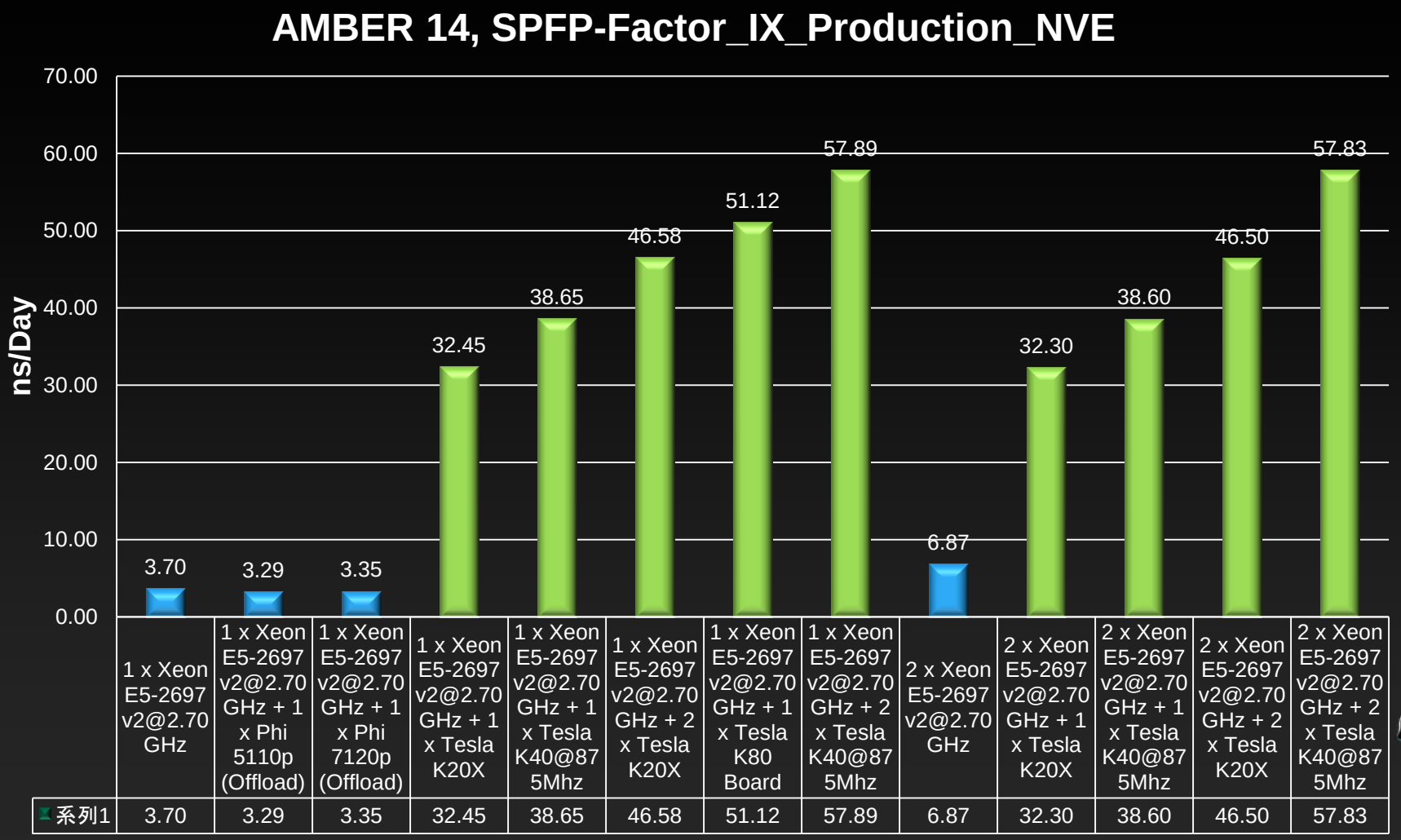


Running AMBER 14

The blue node contains Dual E5-2697 CPUs (12 Cores per CPU).

The green nodes contain Dual E5-2697 CPUs (12 Cores per CPU) and either 1x or 2x NVIDIA K20X, K40 or K80 for the GPU

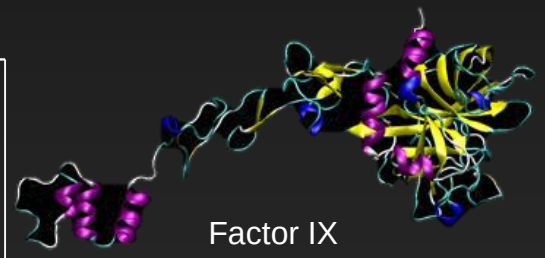
Kepler - Our Fastest Family of GPUs Yet



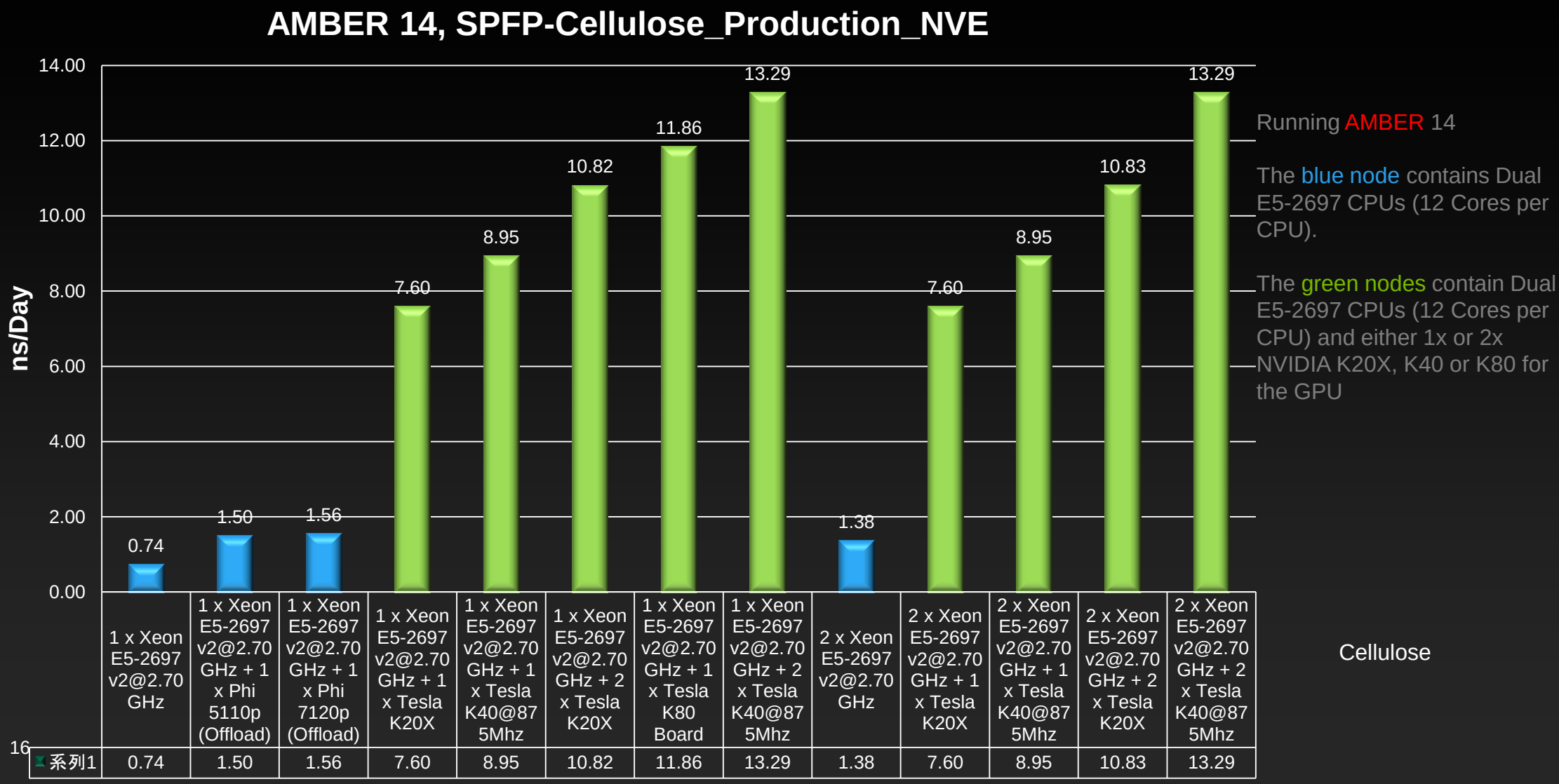
Running AMBER 14

The blue node contains Dual E5-2697 CPUs (12 Cores per CPU).

The green nodes contain Dual E5-2697 CPUs (12 Cores per CPU) and either 1x or 2x NVIDIA K20X, K40 or K80 for the GPU



Kepler - Our Fastest Family of GPUs Yet

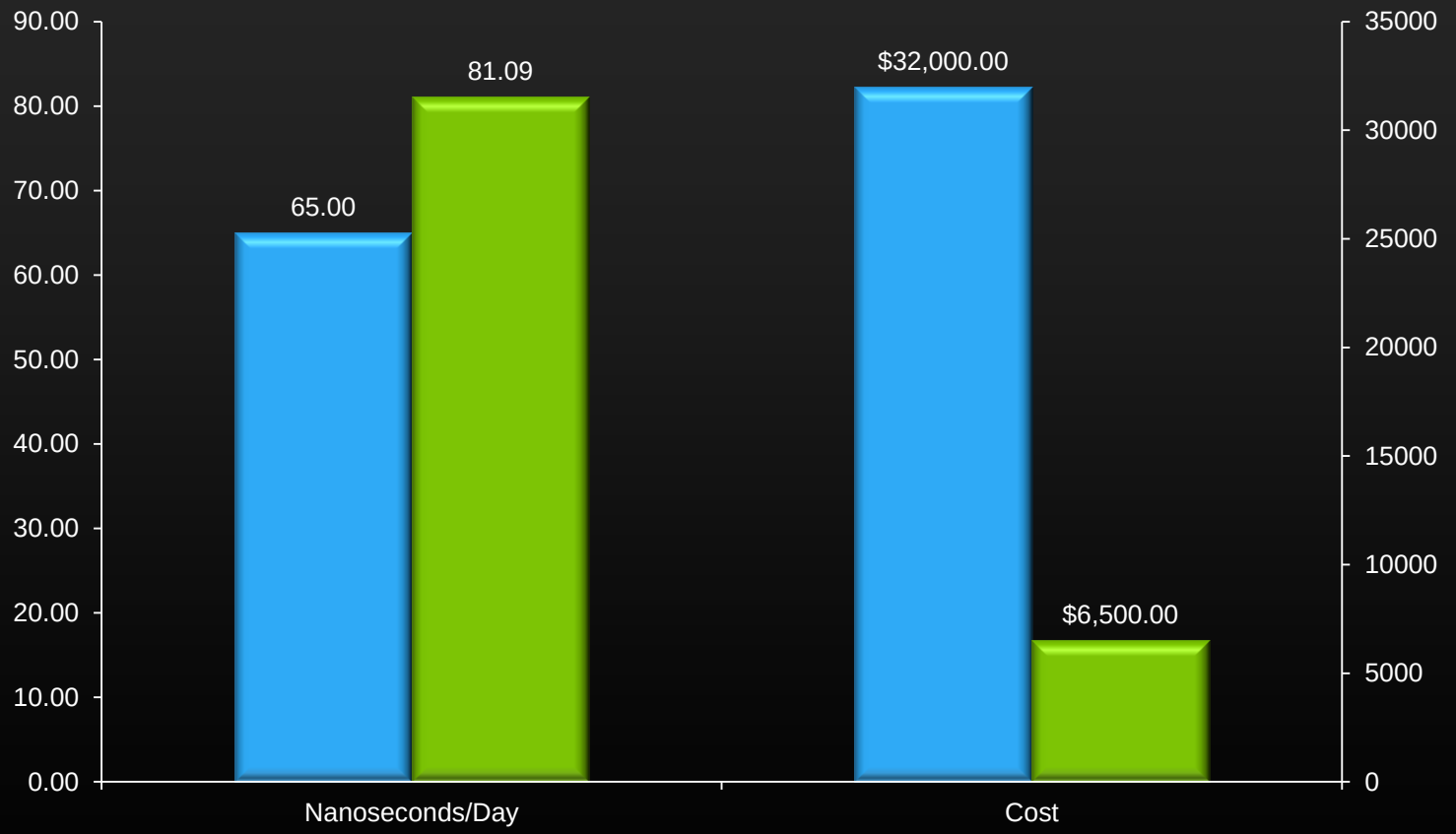


Running AMBER 14

The blue node contains Dual E5-2697 CPUs (12 Cores per CPU).

The green nodes contain Dual E5-2697 CPUs (12 Cores per CPU) and either 1x or 2x NVIDIA K20X, K40 or K80 for the GPU

Replace 8 Nodes with 1 K20 GPU

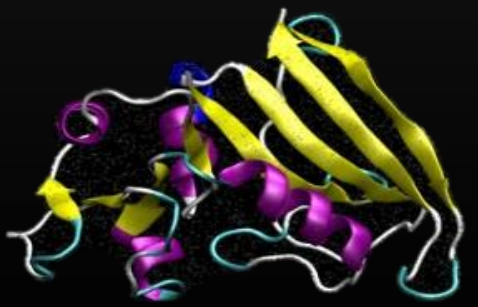


Running AMBER 12 GPU Support Revision 12.1 SPFP with CUDA 4.2.9 ECC Off

The eight (8) blue nodes each contain 2x Intel E5-2687W CPUs (8 Cores per CPU)

Each green node contains 2x Intel E5-2687W CPUs (8 Cores per CPU) plus 1x NVIDIA K20 GPU

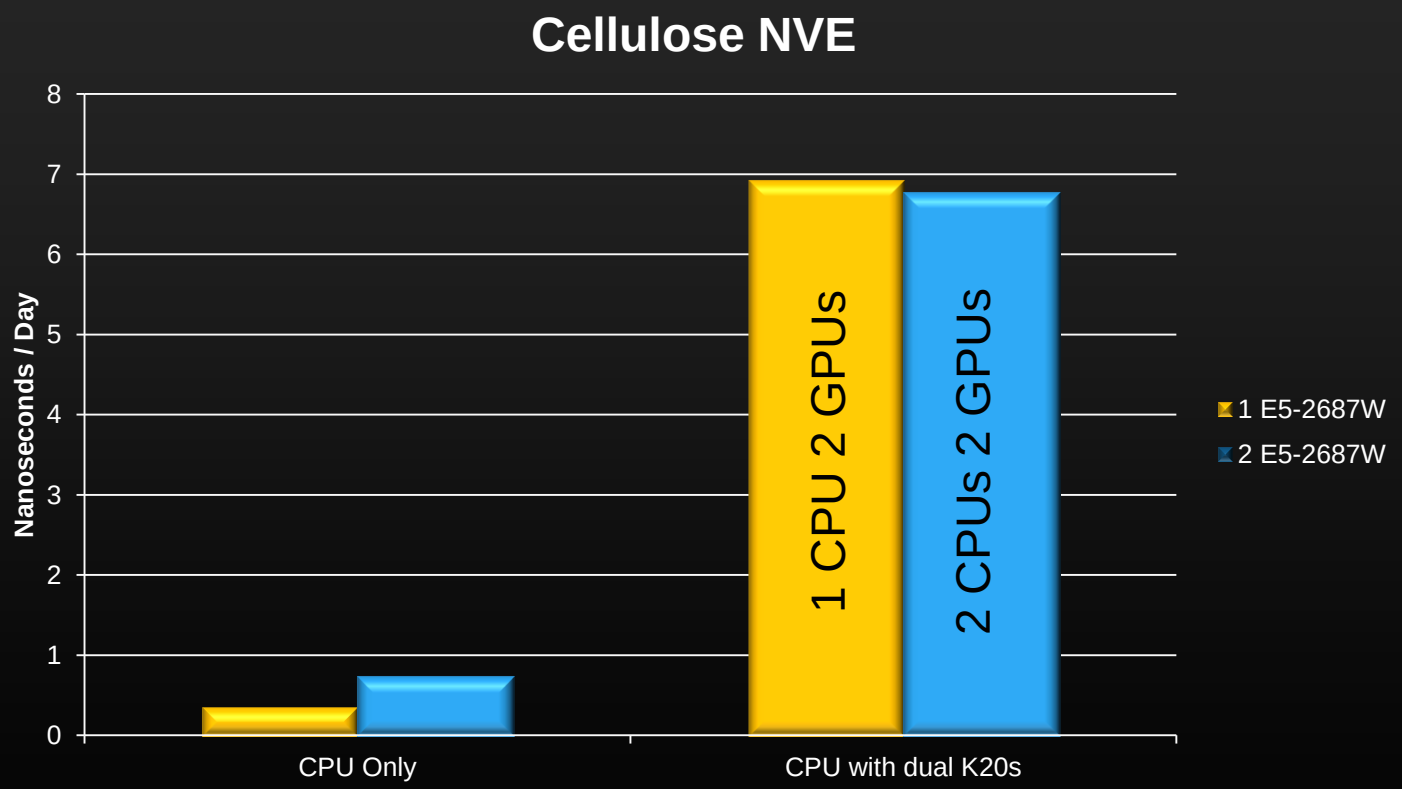
Note: Typical CPU and GPU node pricing used. Pricing may vary depending on node configuration. Contact your preferred HW vendor for actual pricing.



DHFR

Cut down simulation costs to ¼ and gain higher performance

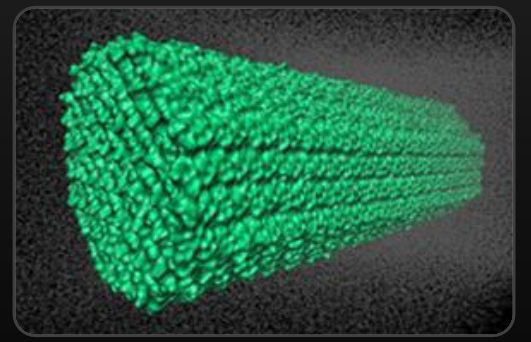
Extra CPUs decrease Performance



Running AMBER 12 GPU Support Revision 12.1

The orange bars contains one E5-2687W CPUs (8 Cores per CPU).

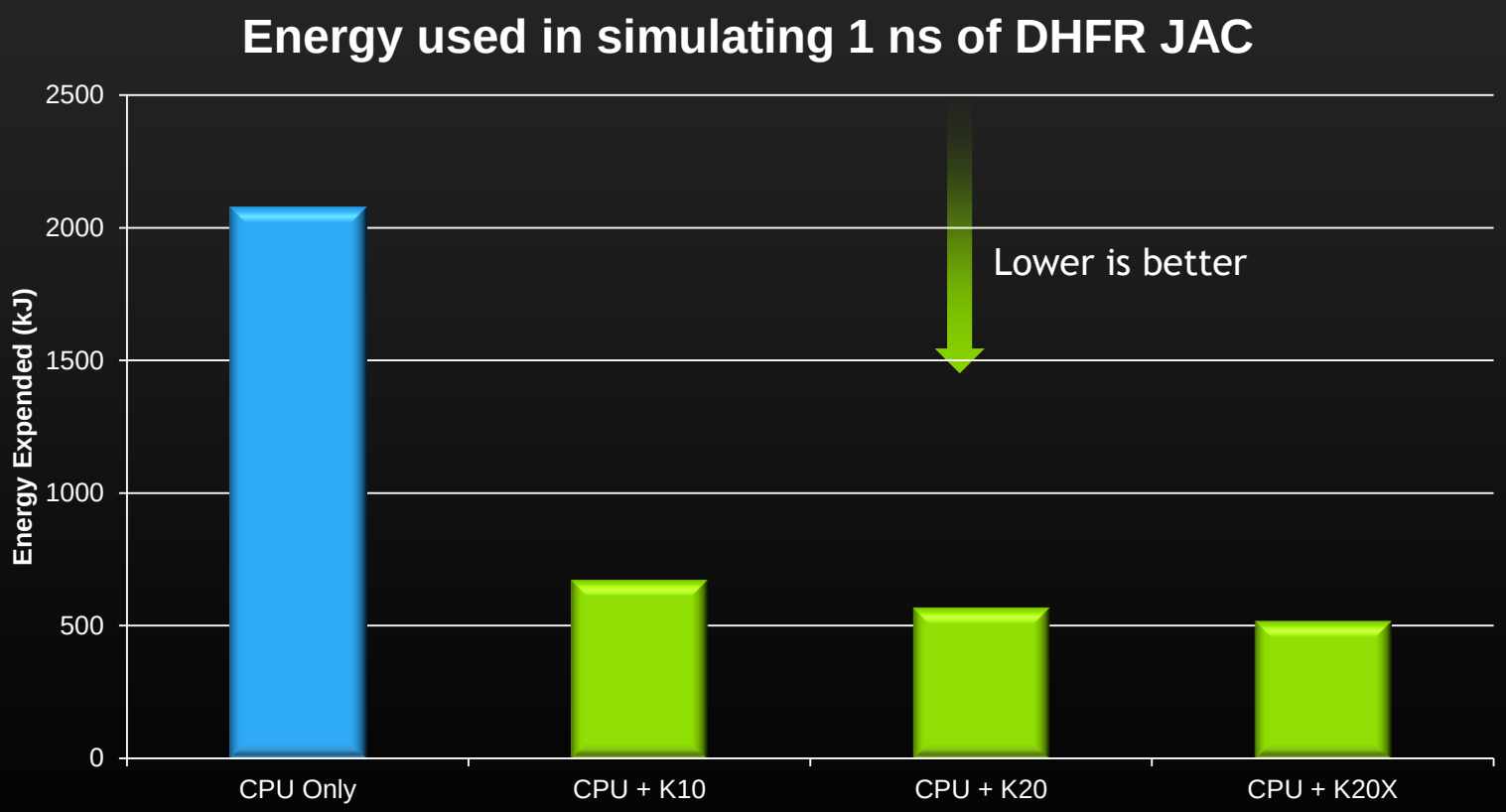
The blue bars contain Dual E5-2687W CPUs (8 Cores per CPU)



Cellulose

When used with GPUs, dual CPU sockets perform worse than single CPU sockets.

Kepler - Greener Science



Running AMBER 12 GPU Support Revision 12.1

The blue node contains Dual E5-2687W CPUs (150W each, 8 Cores per CPU).

The green nodes contain Dual E5-2687W CPUs (8 Cores per CPU) and 1x NVIDIA K10, K20, or K20X GPUs (235W each).

Energy Expended = Power x Time

The GPU Accelerated systems use 65-75% less energy

Recommended GPU Node Configuration for AMBER Computational Chemistry

Workstation or Single Node Configuration	
# of CPU sockets	2
Cores per CPU socket	6+ (1 CPU core drives 1 GPU)
CPU speed (Ghz)	2.66+
System memory per node (GB)	16
GPUs	Kepler K20, K40, K80
# of GPUs per CPU socket	1-4
GPU memory preference (GB)	6
GPU to CPU connection	PCIe 3.0 16x or higher
Server storage	2 TB
Network configuration	Infiniband QDR or better

Scale to multiple nodes with same single node configuration

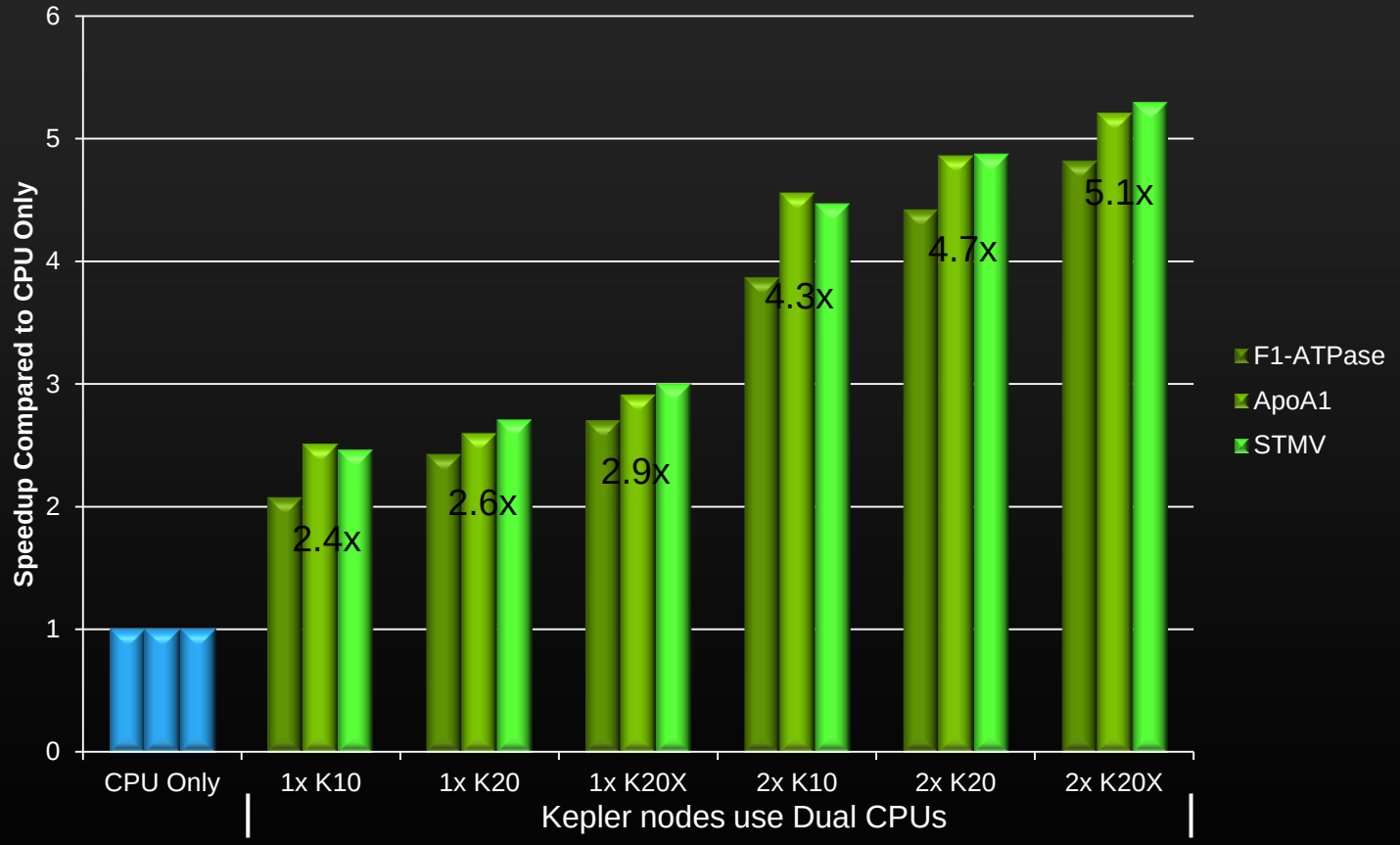
Benefits of GPU AMBER Accelerated Computing

- Faster than CPU only systems in all tests
- Most major compute intensive aspects of classical MD ported
- Large performance boost with marginal price increase
- Energy usage cut by more than half
- GPUs scale well within a node and over multiple nodes
- K20 GPU is our fastest and lowest power high performance GPU yet

Try GPU accelerated AMBER for free – www.nvidia.com/GPUTestDrive

NAMD 2.10

Kepler – Universally Faster



Running NAMD version 2.9

The CPU Only node contains Dual E5-2687W CPUs (8 Cores per CPU).

The Kepler nodes contain Dual E5-2687W CPUs (8 Cores per CPU) and 1 or two NVIDIA K10, K20, or K20X GPUs.

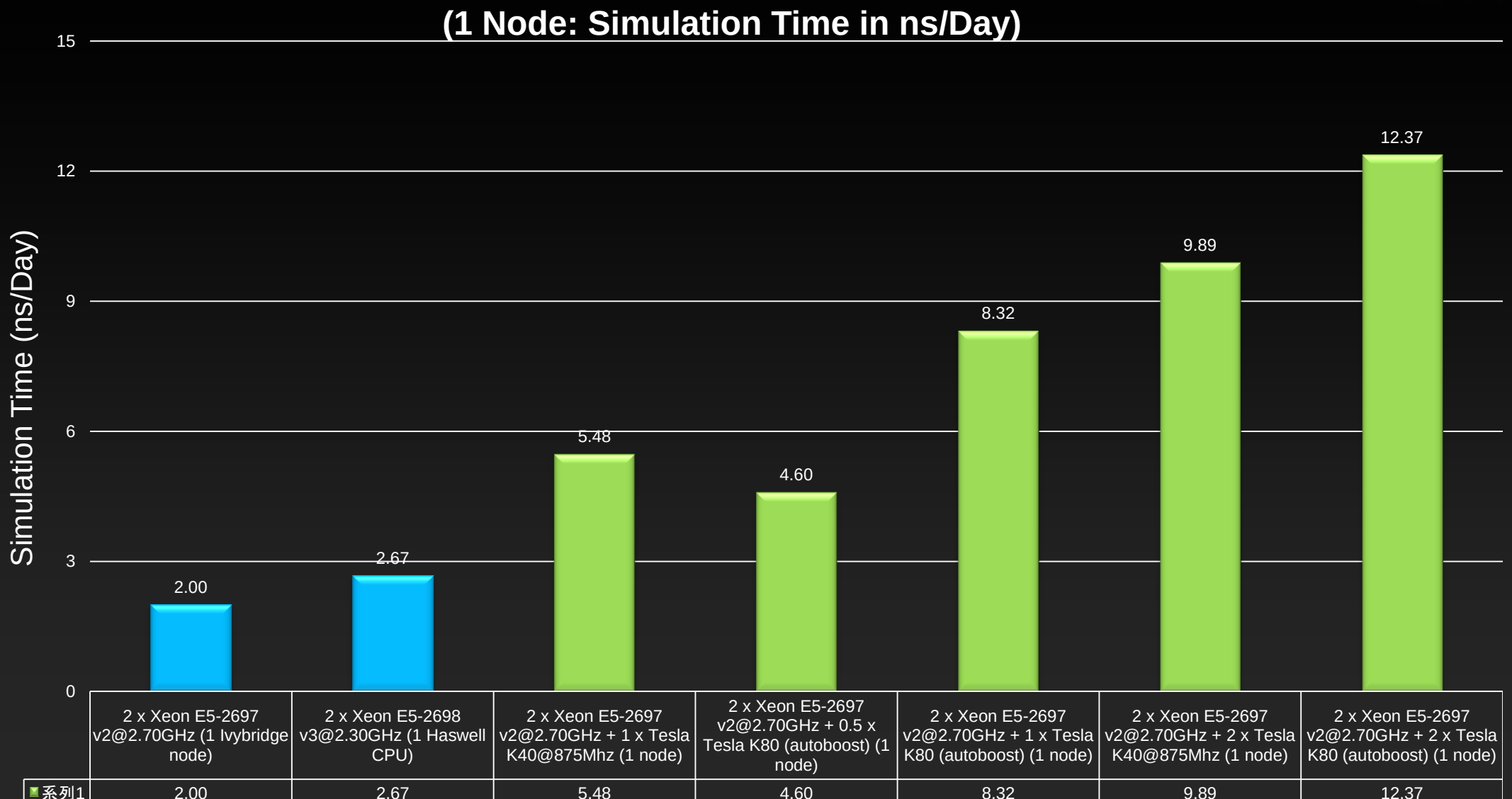
F1-ATPase

The Kepler GPUs accelerate all simulations, up to 5.1x

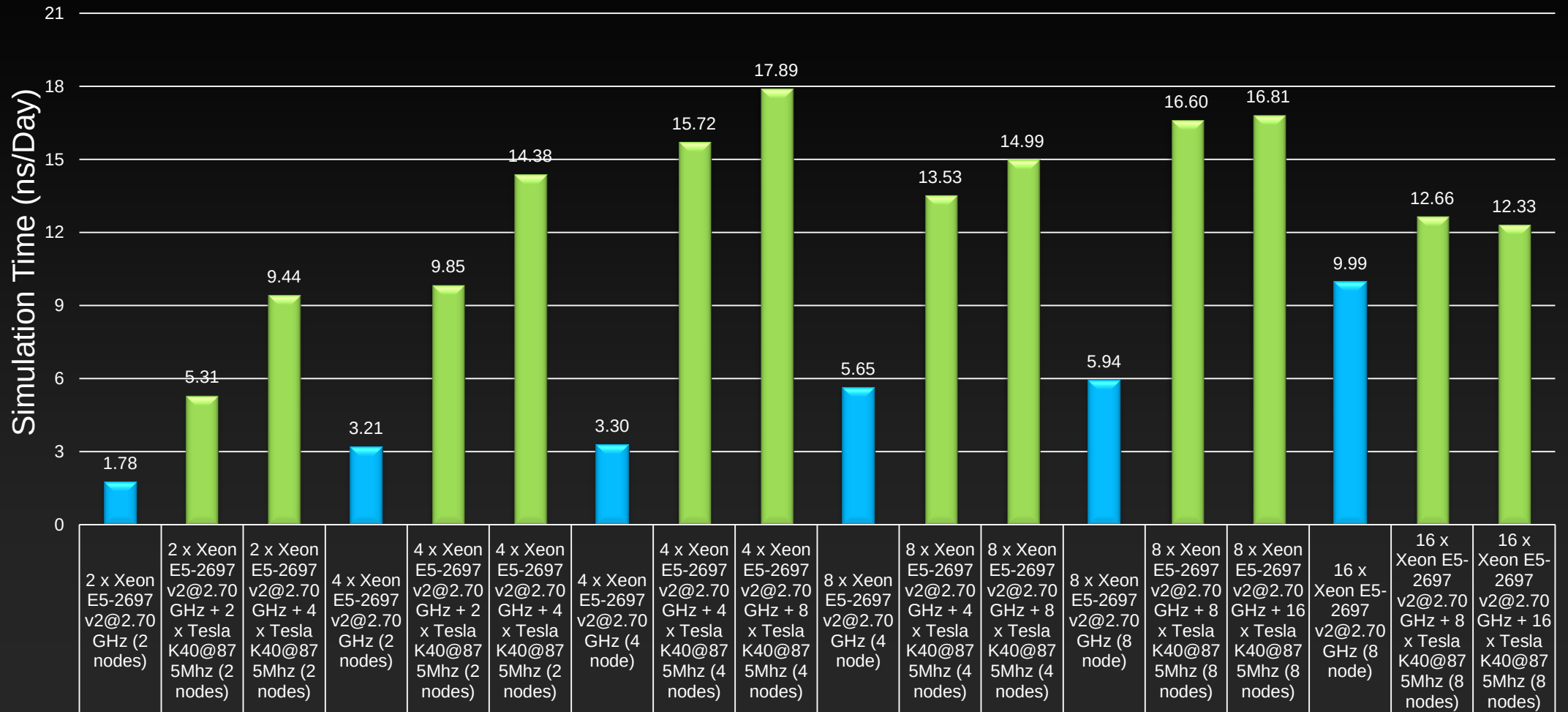
Average acceleration printed in bars

NAMD 2.10
(streaming patch)

NAMD 2.10; ApoA1 on Intel Phi, Tesla K40s and K80s & IVB CPUs

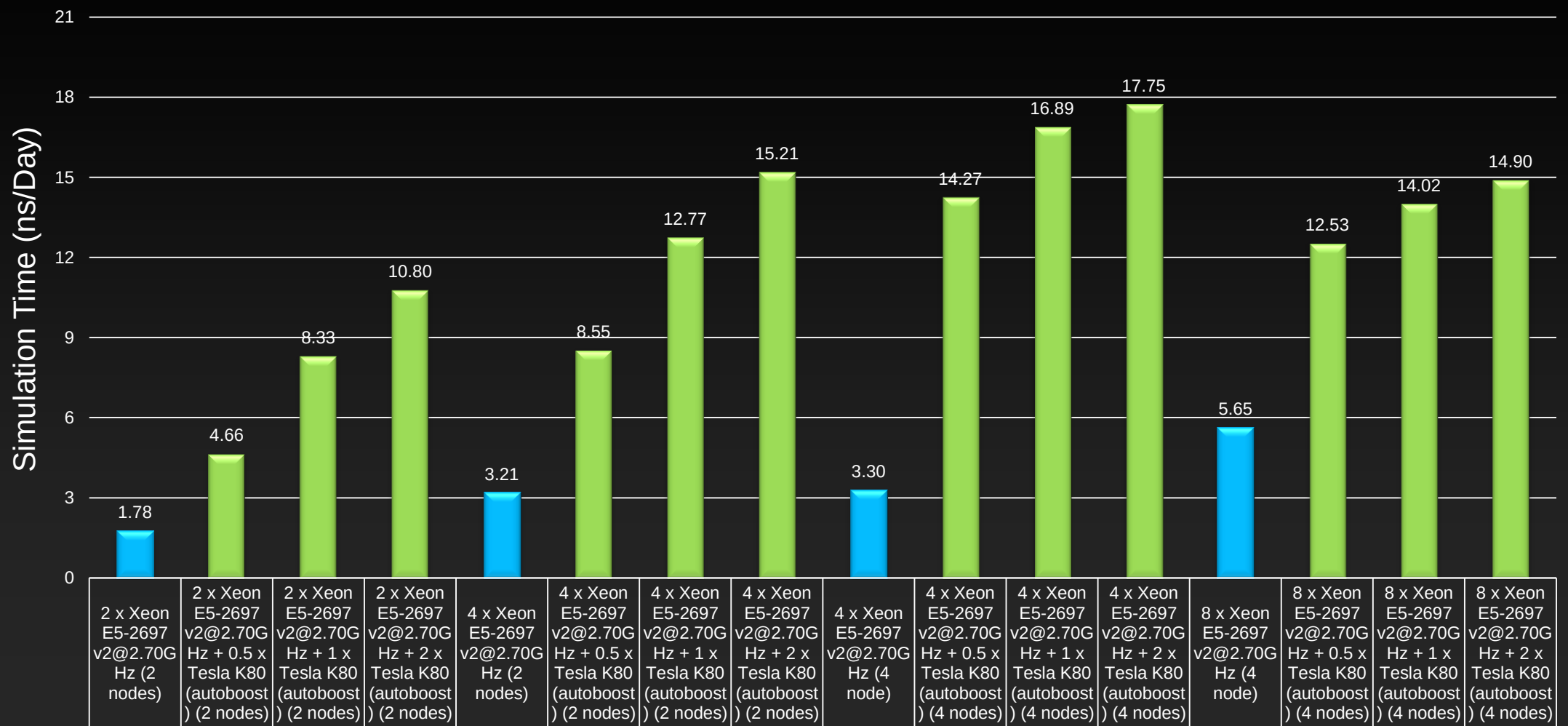


NAMD 2.10; ApoA1 on Tesla K40s & IVB CPUs



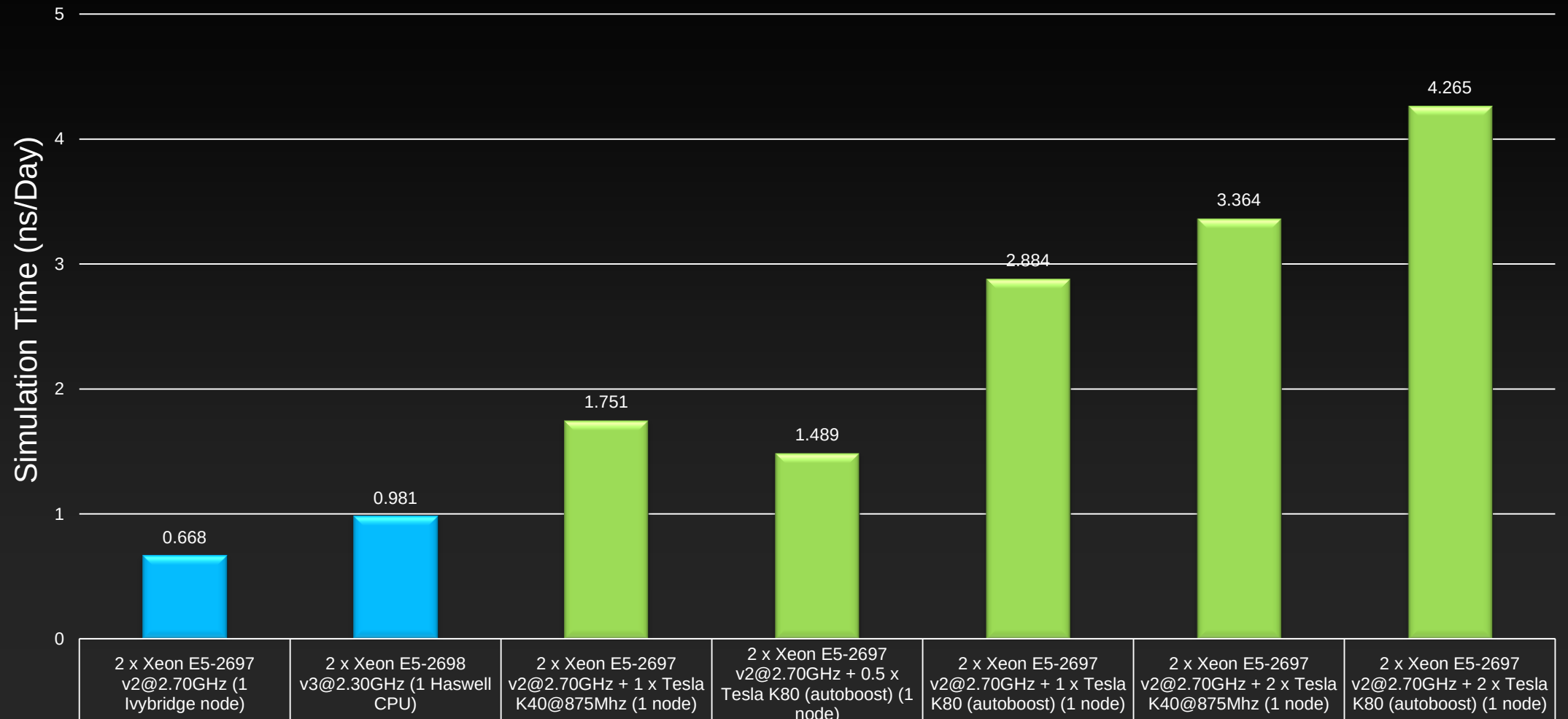
NAMD 2.10; ApoA1 on Tesla K80s & IVB CPUs

(2-4 Nodes: Simulation Time in ns/Day)

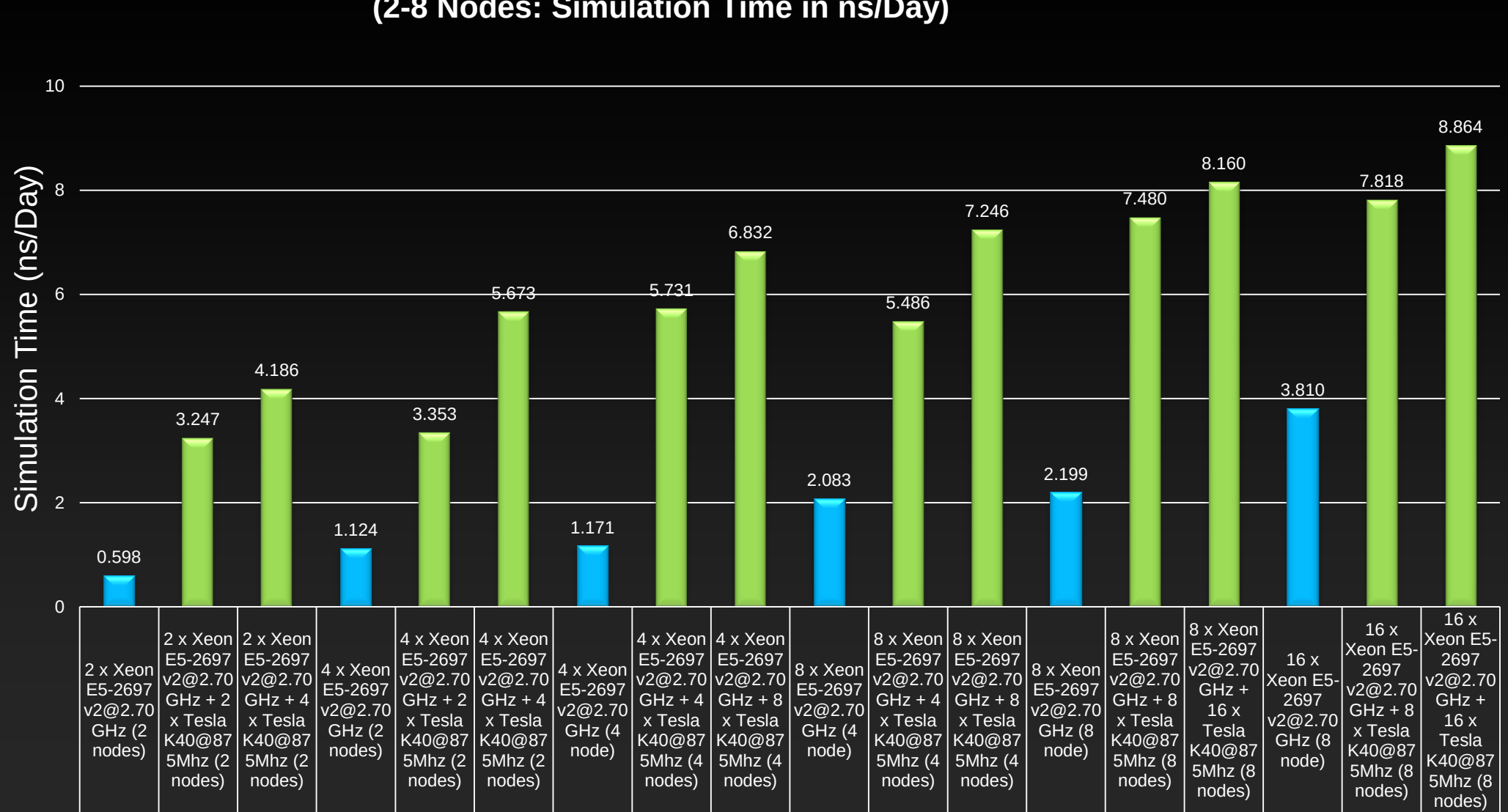


NAMD 2.10; F1-ATPase on Intel Phi, Tesla K40s and K80s & IVB CPUs

(1 Node: Simulation Time in ns/Day)

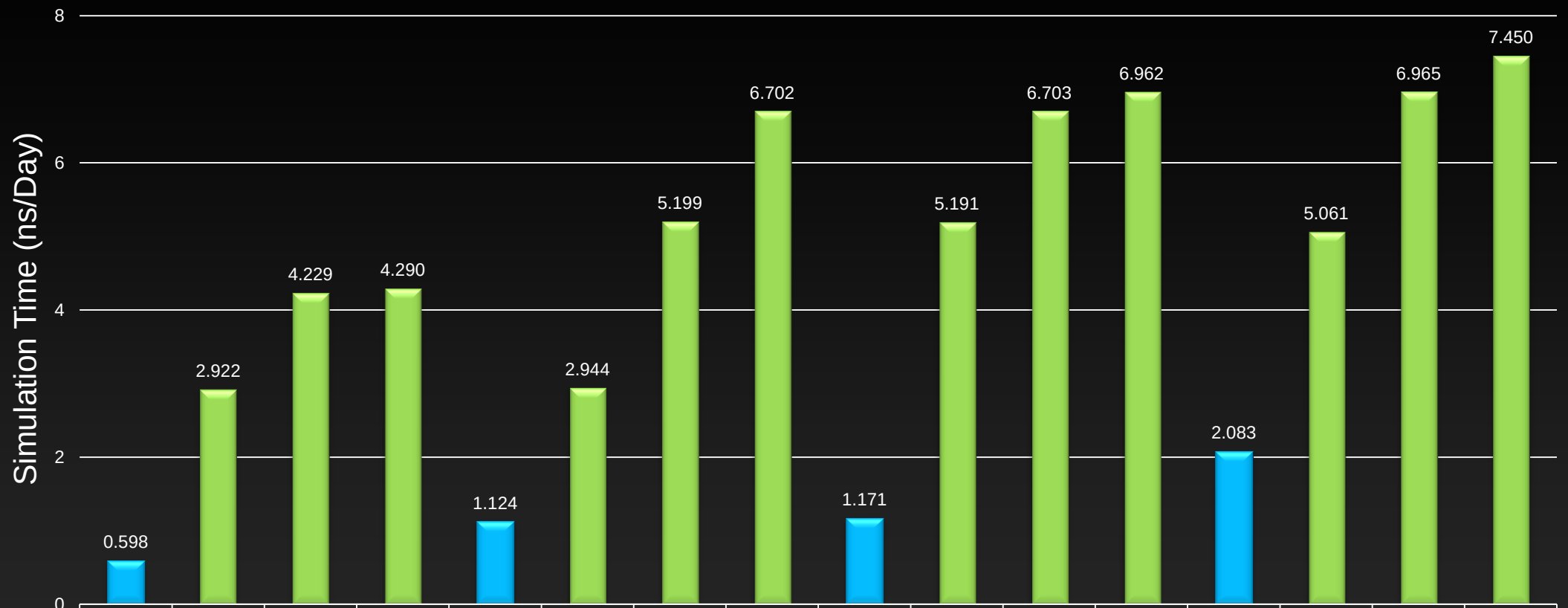


NAMD 2.10; F1-ATPase on Tesla K40s & IVB CPUs



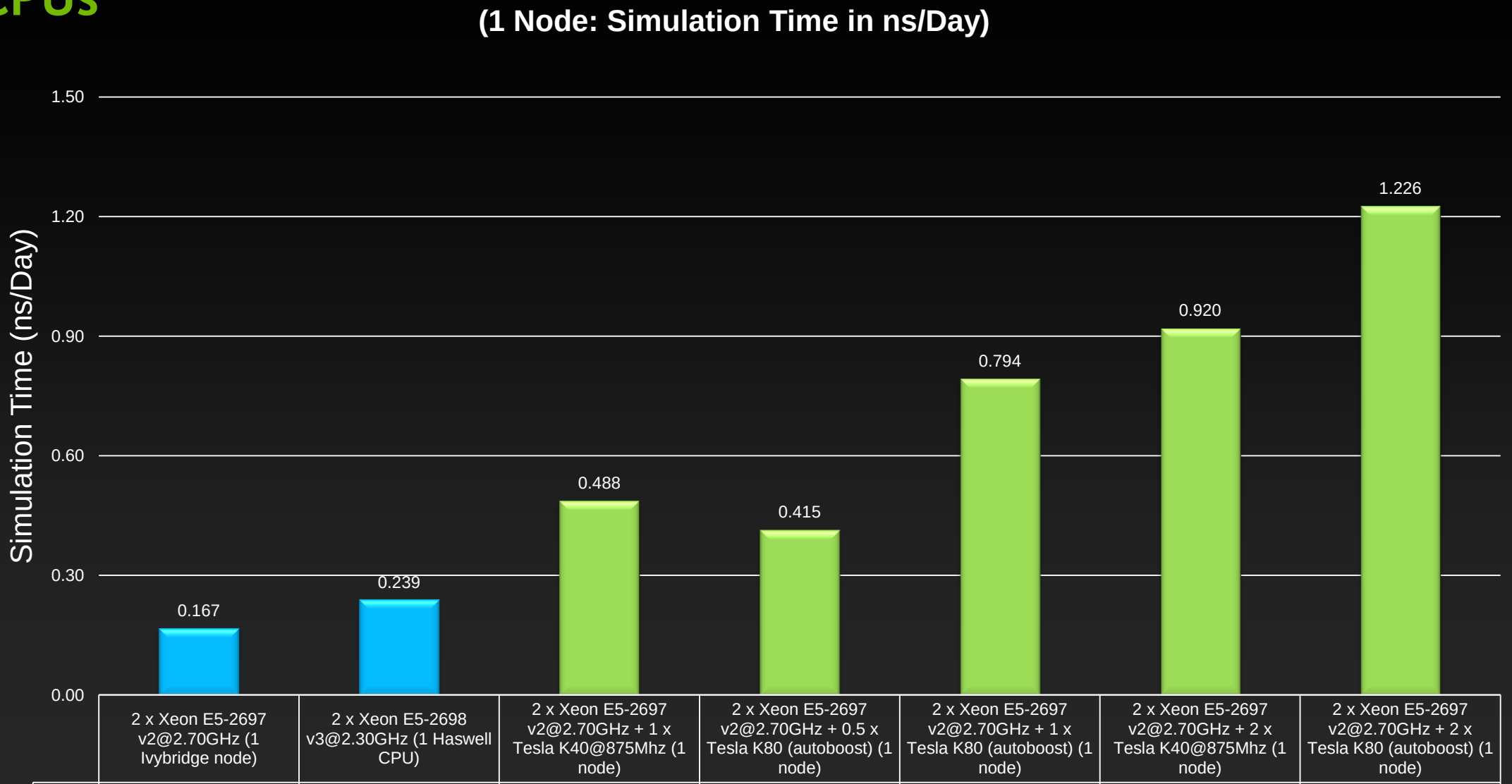
NAMD 2.10; F1-ATPase on Tesla K80s & IVB CPUs

(2-4 Nodes: Simulation Time in ns/Day)

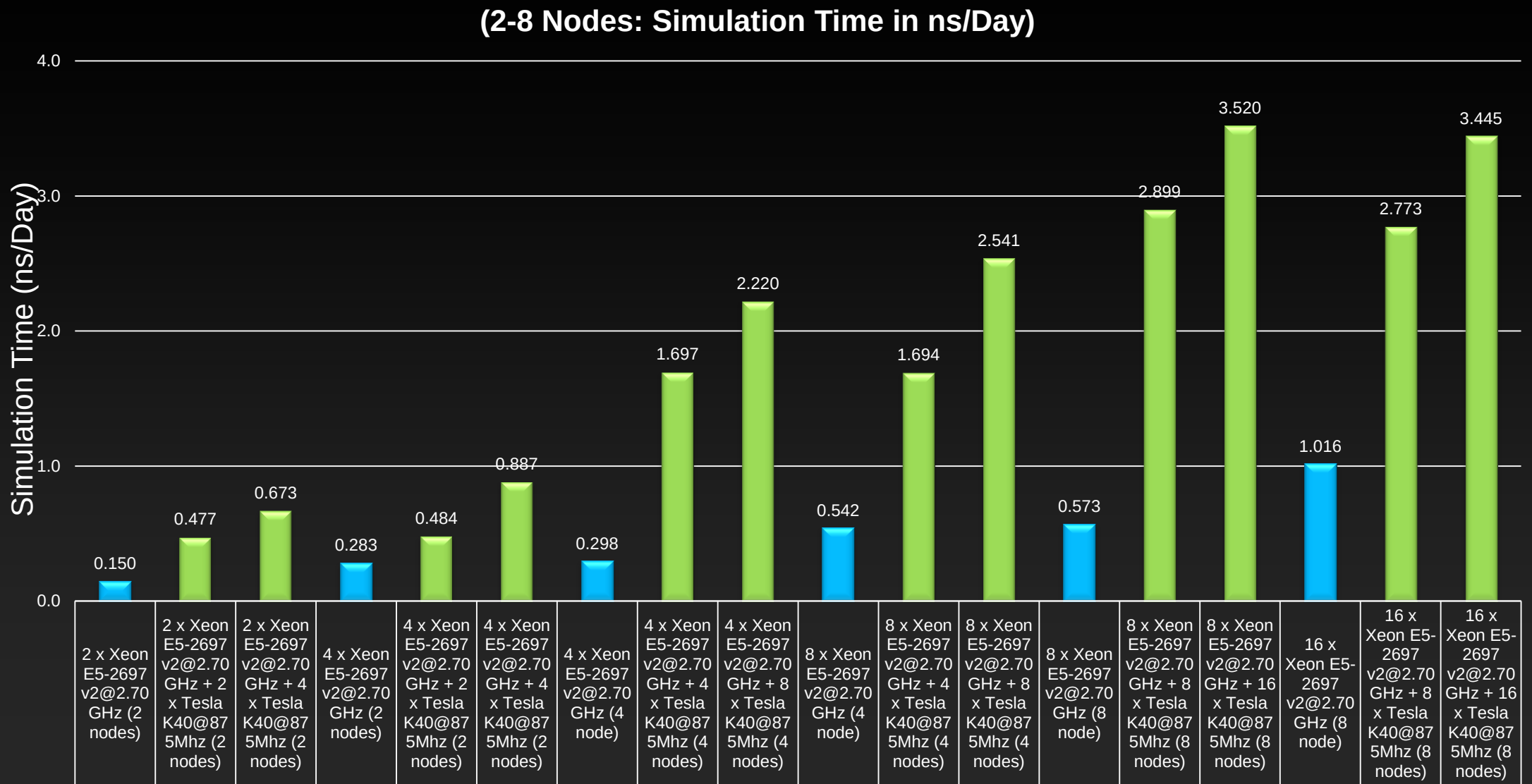


	2 x Xeon E5-2697 v2@2.70GHz (2 nodes)	2 x Xeon E5-2697 v2@2.70GHz + 0.5 x Tesla K80 (autoboost) (2 nodes)	2 x Xeon E5-2697 v2@2.70GHz + 1 x Tesla K80 (autoboost) (2 nodes)	2 x Xeon E5-2697 v2@2.70GHz + 2 x Tesla K80 (autoboost) (2 nodes)	4 x Xeon E5-2697 v2@2.70GHz (2 nodes)	4 x Xeon E5-2697 v2@2.70GHz + 0.5 x Tesla K80 (autoboost) (2 nodes)	4 x Xeon E5-2697 v2@2.70GHz + 1 x Tesla K80 (autoboost) (2 nodes)	4 x Xeon E5-2697 v2@2.70GHz + 2 x Tesla K80 (autoboost) (2 nodes)	4 x Xeon E5-2697 v2@2.70GHz (4 node)	4 x Xeon E5-2697 v2@2.70GHz + 0.5 x Tesla K80 (autoboost) (4 nodes)	4 x Xeon E5-2697 v2@2.70GHz + 1 x Tesla K80 (autoboost) (4 nodes)	4 x Xeon E5-2697 v2@2.70GHz + 2 x Tesla K80 (autoboost) (4 nodes)	8 x Xeon E5-2697 v2@2.70GHz (4 node)	8 x Xeon E5-2697 v2@2.70GHz + 0.5 x Tesla K80 (autoboost) (4 nodes)	8 x Xeon E5-2697 v2@2.70GHz + 1 x Tesla K80 (autoboost) (4 nodes)	8 x Xeon E5-2697 v2@2.70GHz + 2 x Tesla K80 (autoboost) (4 nodes)
系列1	0.598	2.922	4.229	4.290	1.124	2.944	5.199	6.702	1.171	5.191	6.703	6.962	2.083	5.061	6.965	7.450

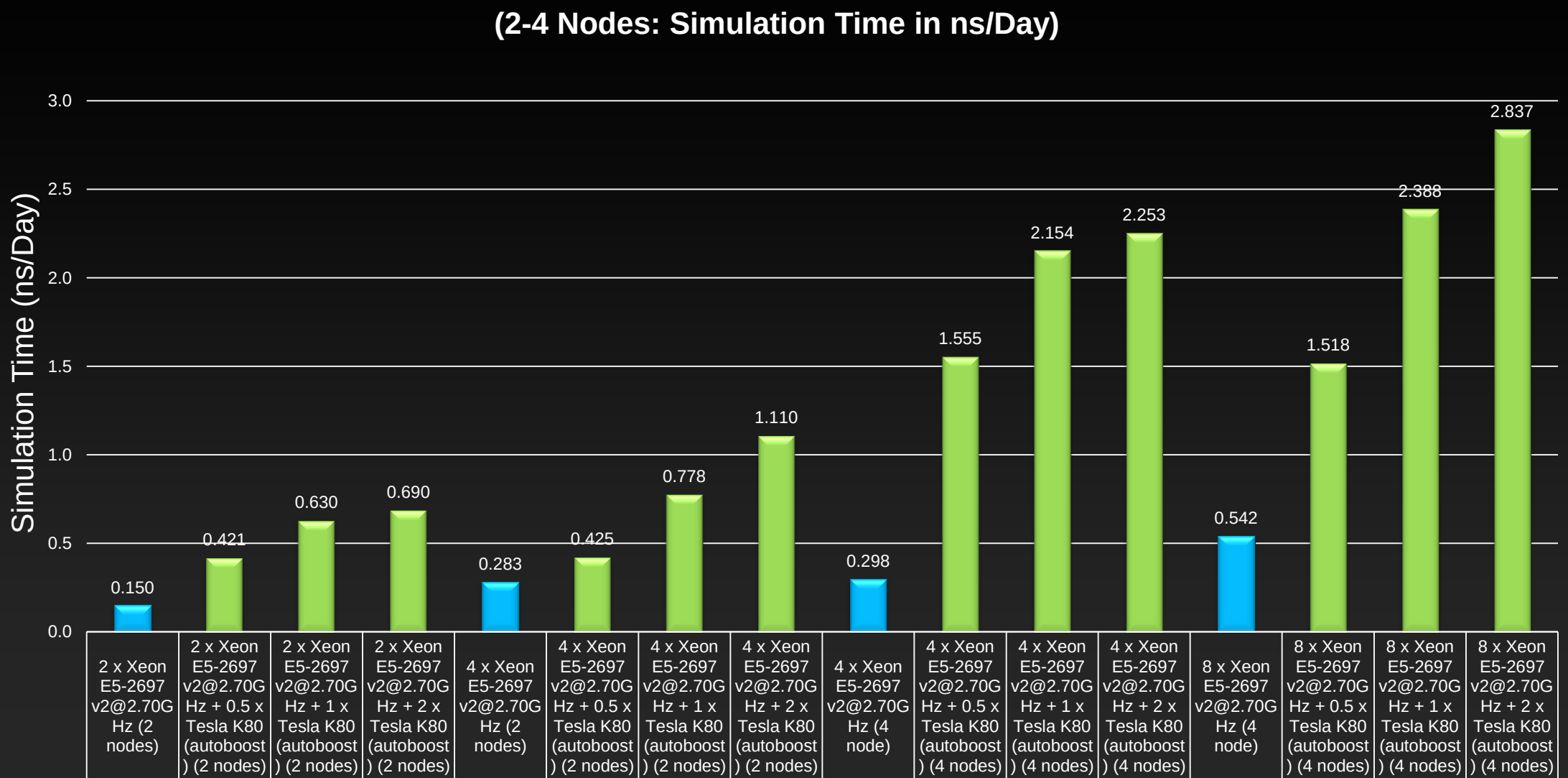
NAMD 2.10; STMV on Intel Phi, Tesla K40s and K80s & IVB CPUs



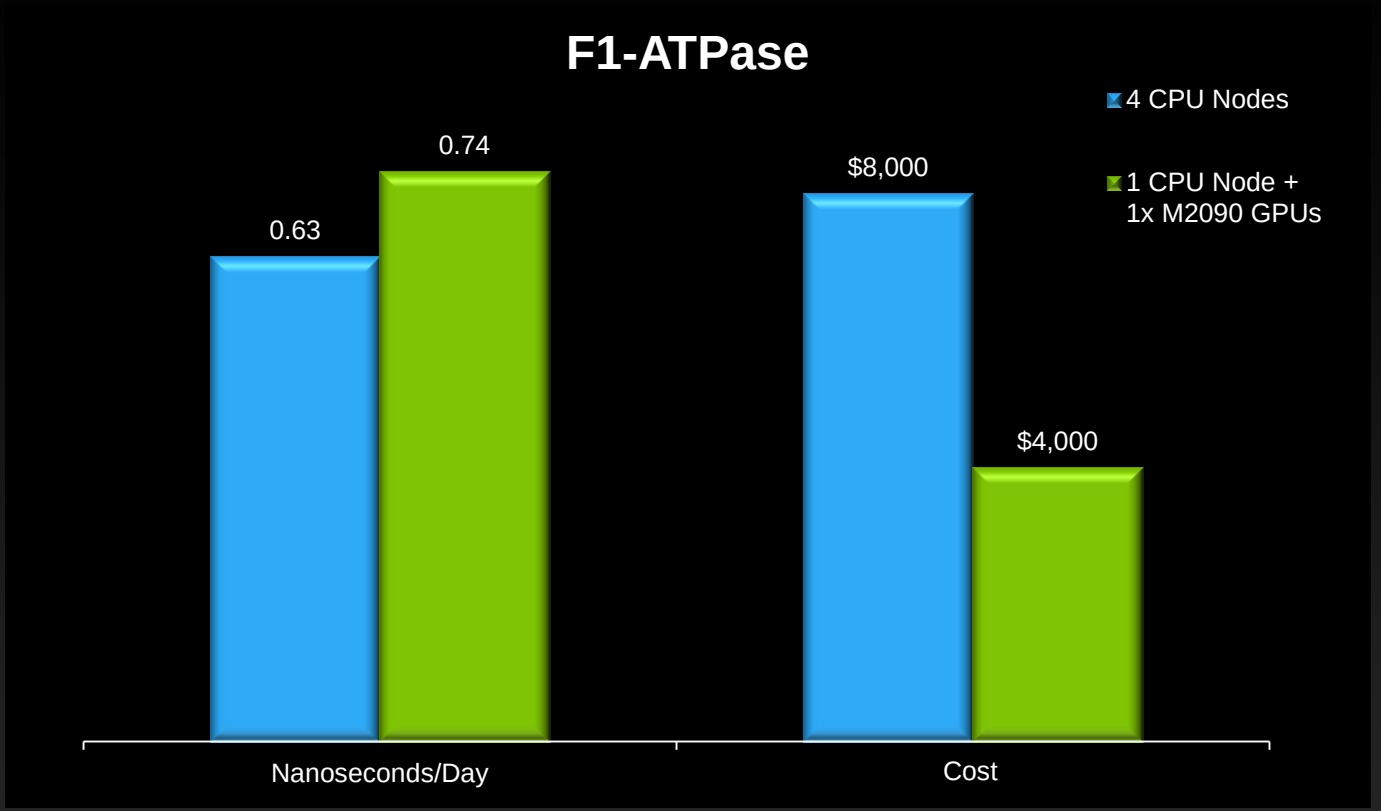
NAMD 2.10; STMV on Tesla K40s & IVB CPUs



NAMD 2.10; STMV on Tesla K80s & IVB CPUs



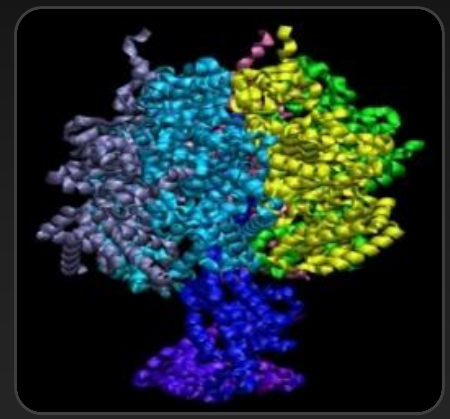
Replace 3 Nodes with 1 2090 GPU



Running NAMD version 2.9
Each blue node contains 2x Intel Xeon X5550 CPUs (4 Cores per CPU).

The green node contains 2x Intel Xeon X5550 CPUs (4 Cores per CPU) and 1x NVIDIA M2090 GPU

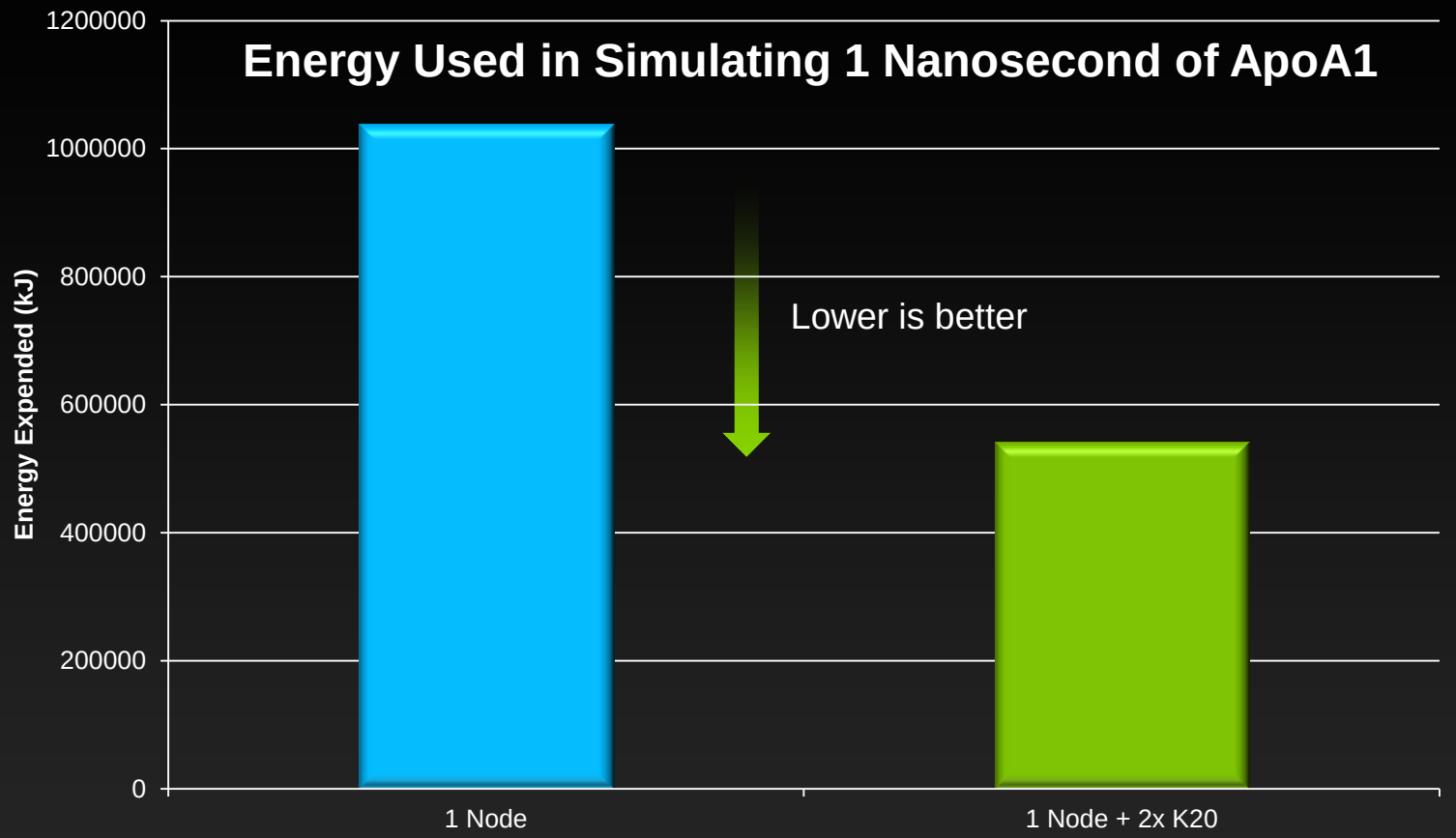
Note: Typical CPU and GPU node pricing used. Pricing may vary depending on node configuration. Contact your preferred HW vendor for actual pricing.



F1-ATPase

Speedup of 1.2x for 50% the cost

K20 - Greener: Twice The Science Per Watt



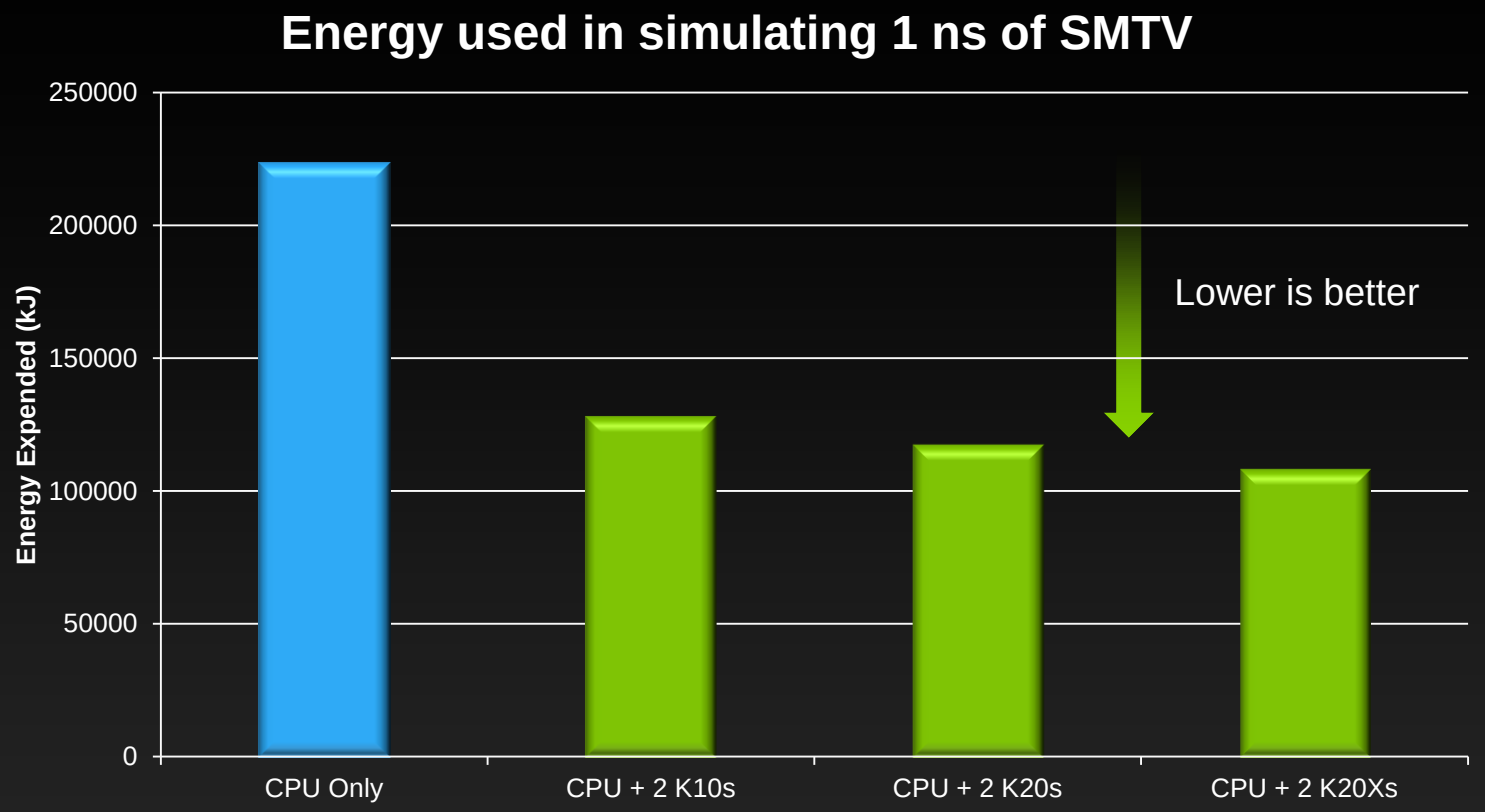
Cut down energy usage by ½ with GPUs

Running NAMD version 2.9
Each blue node contains Dual E5-2687W CPUs (95W, 4 Cores per CPU).

Each green node contains 2x Intel Xeon X5550 CPUs (95W, 4 Cores per CPU) and 2x NVIDIA K20 GPUs (225W per GPU)

Energy Expended
= Power x Time

Kepler - Greener: Twice The Science/Joule



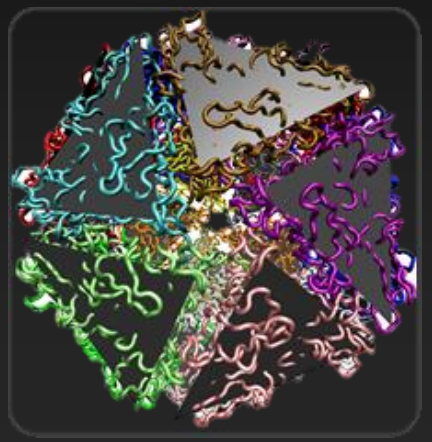
Cut down energy usage by ½ with GPUs

Running NAMD version 2.9

The blue node contains Dual E5-2687W CPUs (150W each, 8 Cores per CPU).

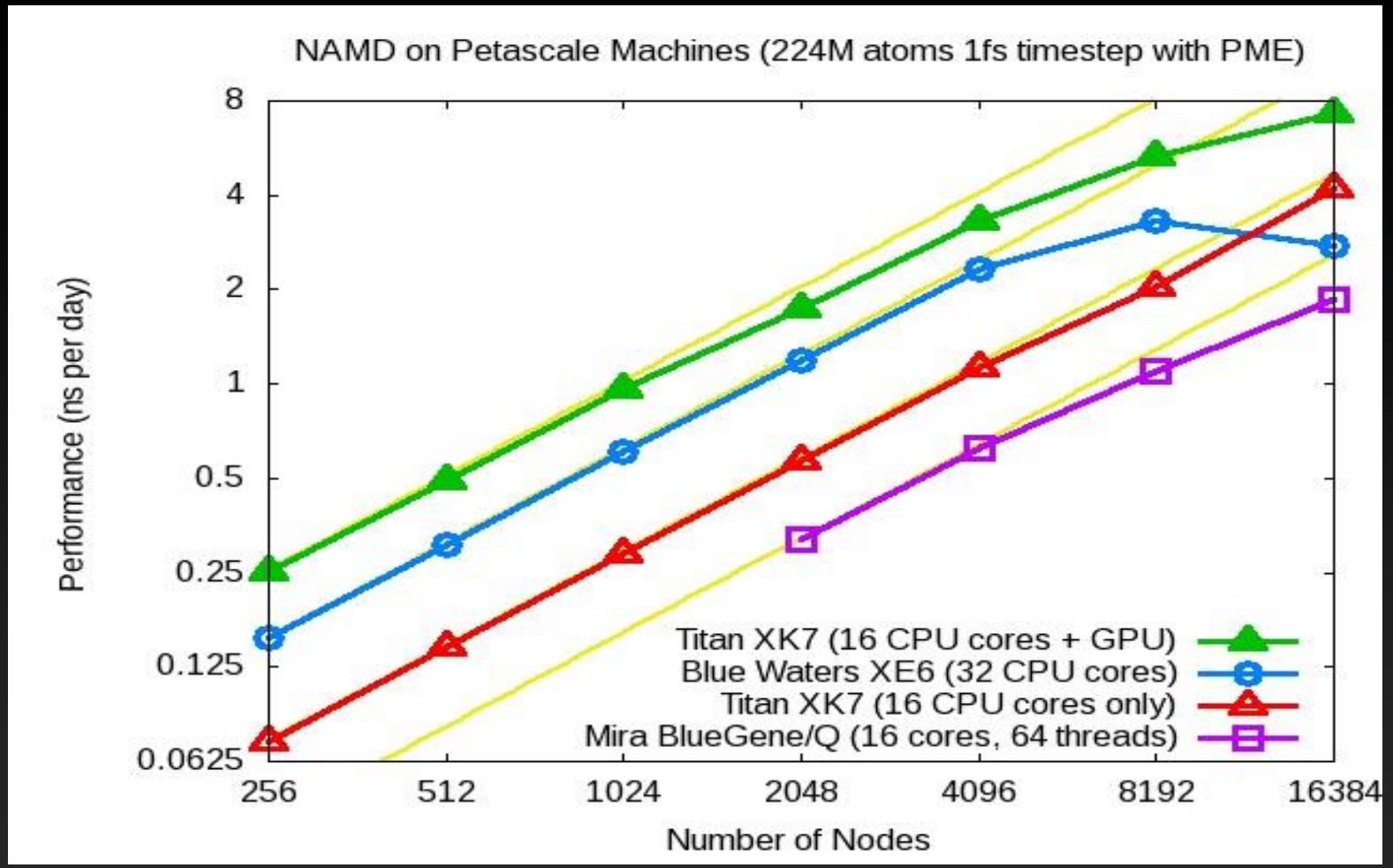
The green nodes contain Dual E5-2687W CPUs (8 Cores per CPU) and 2x NVIDIA K10, K20, or K20X GPUs (235W each).

Energy Expended
= Power x Time

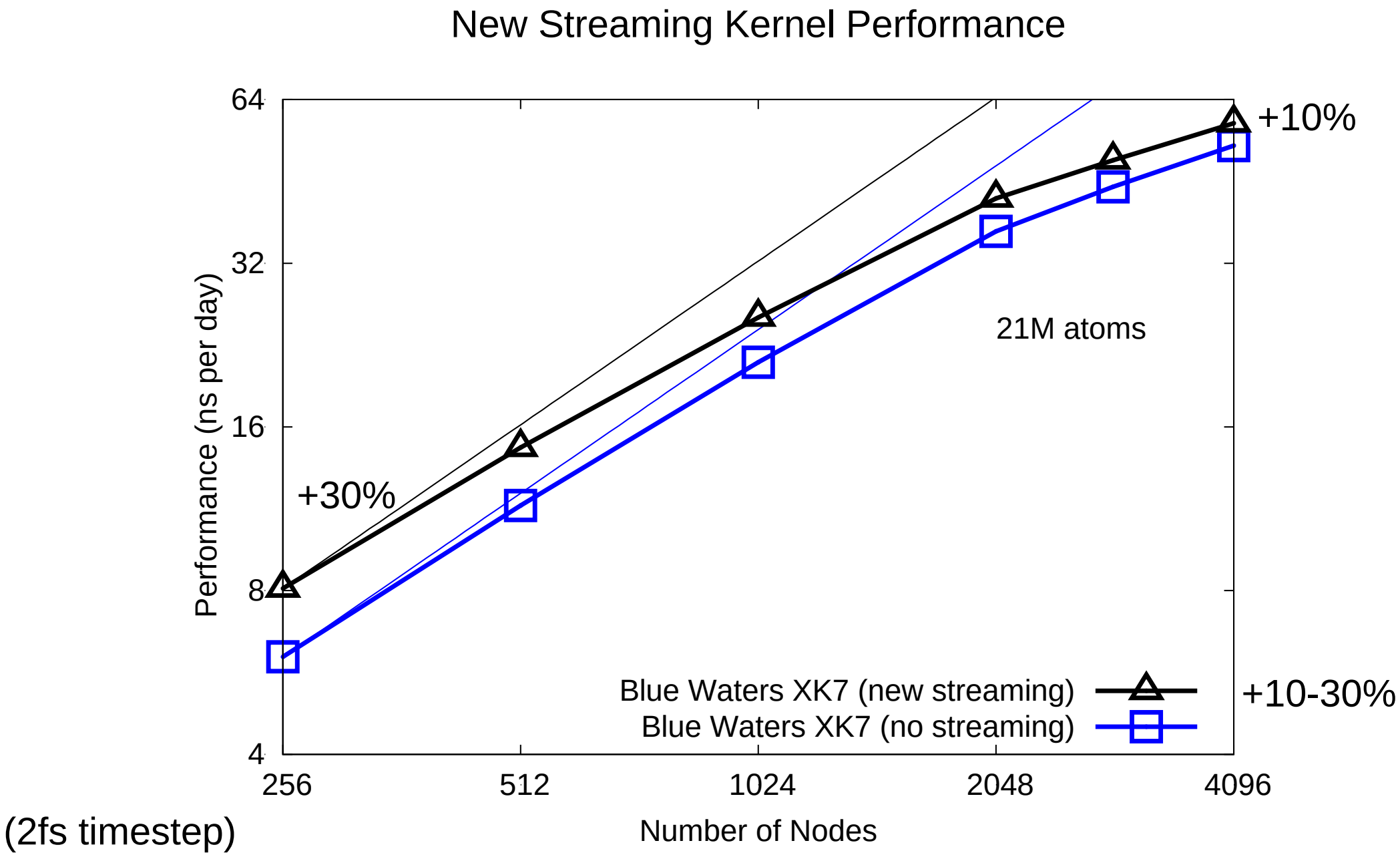


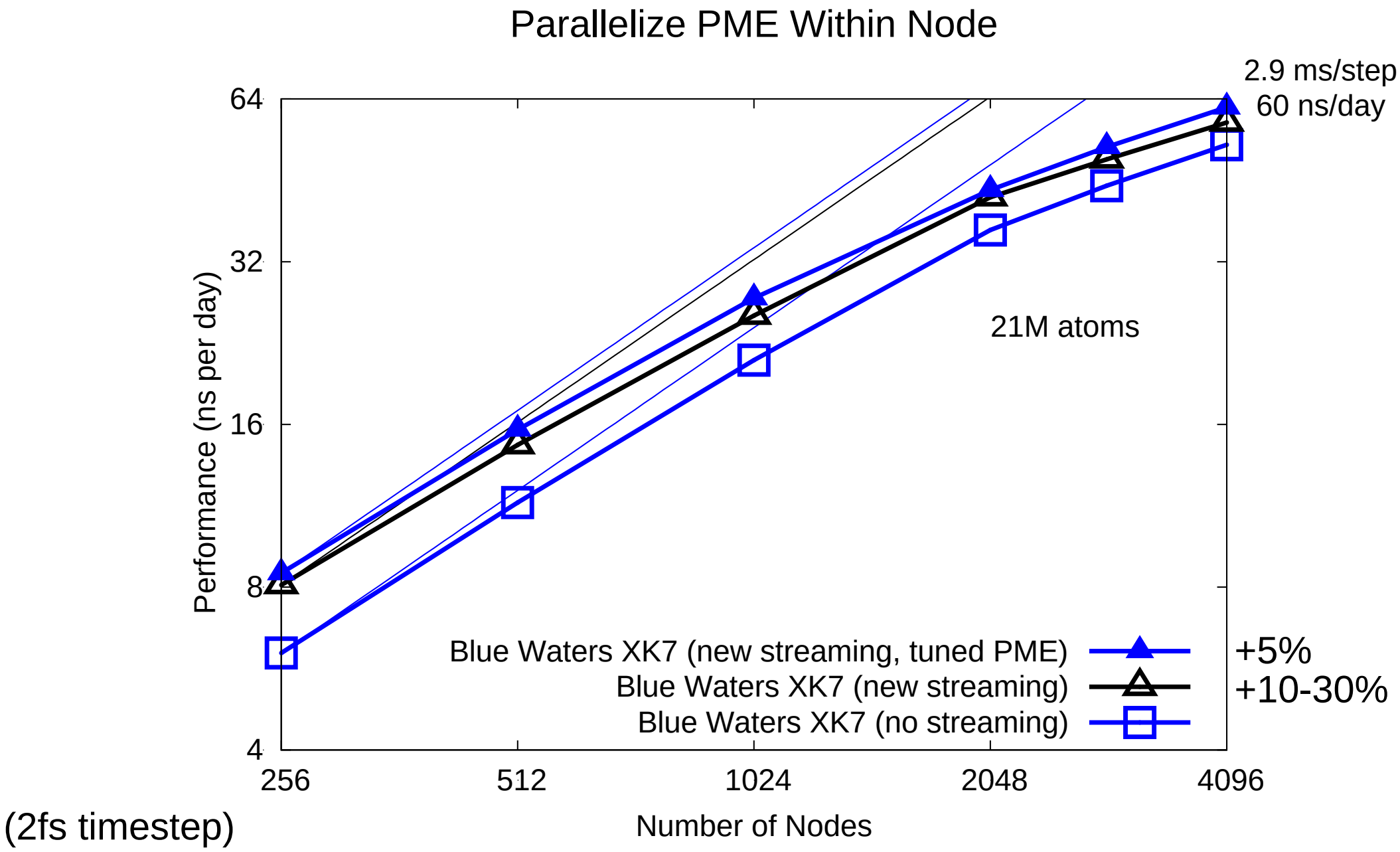
Satellite Tobacco Mosaic Virus

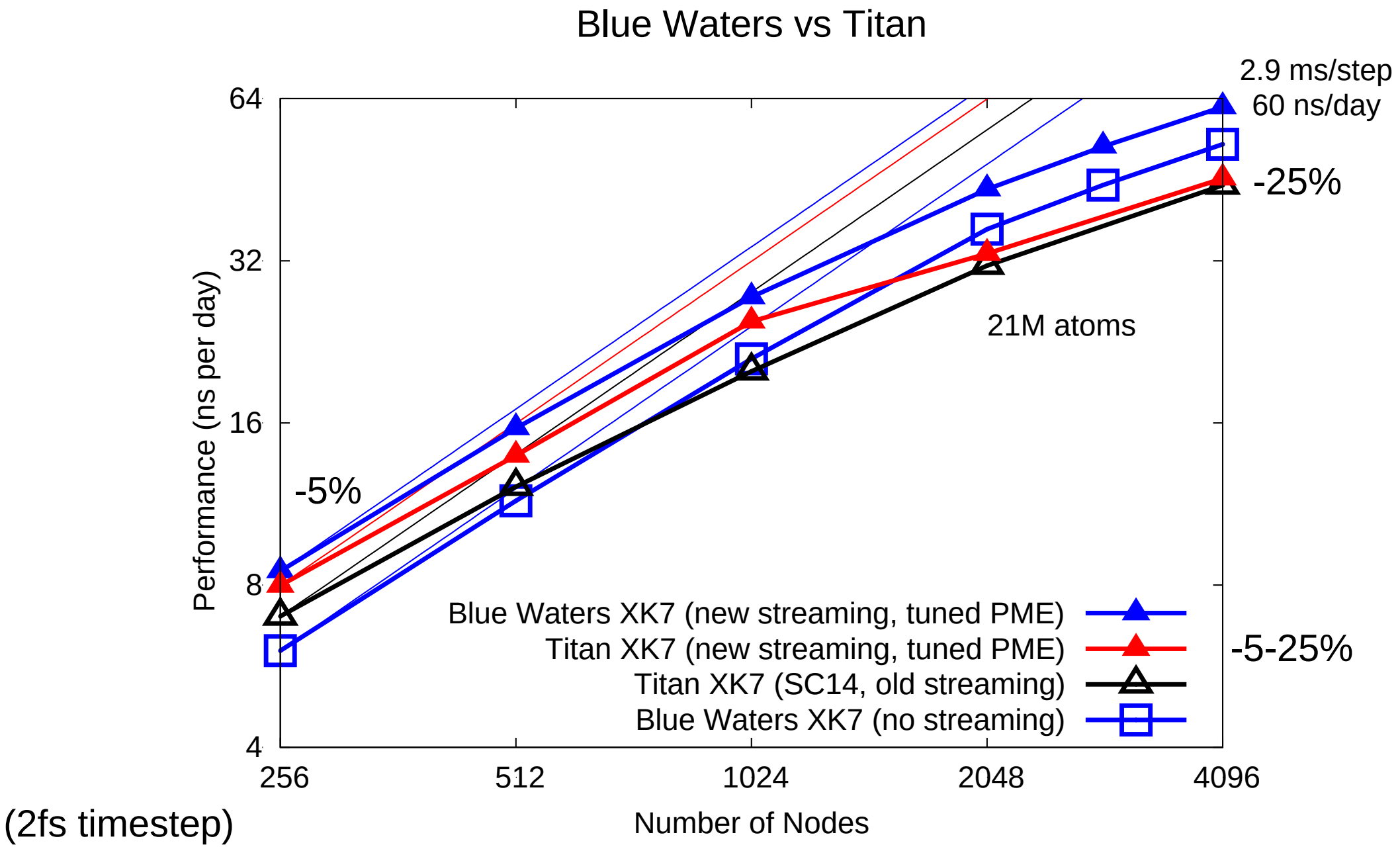
NAMD CPU and GPU Scaling



The next 8 slides are courtesy of Jim Phillips @ Beckman Institute UIUC

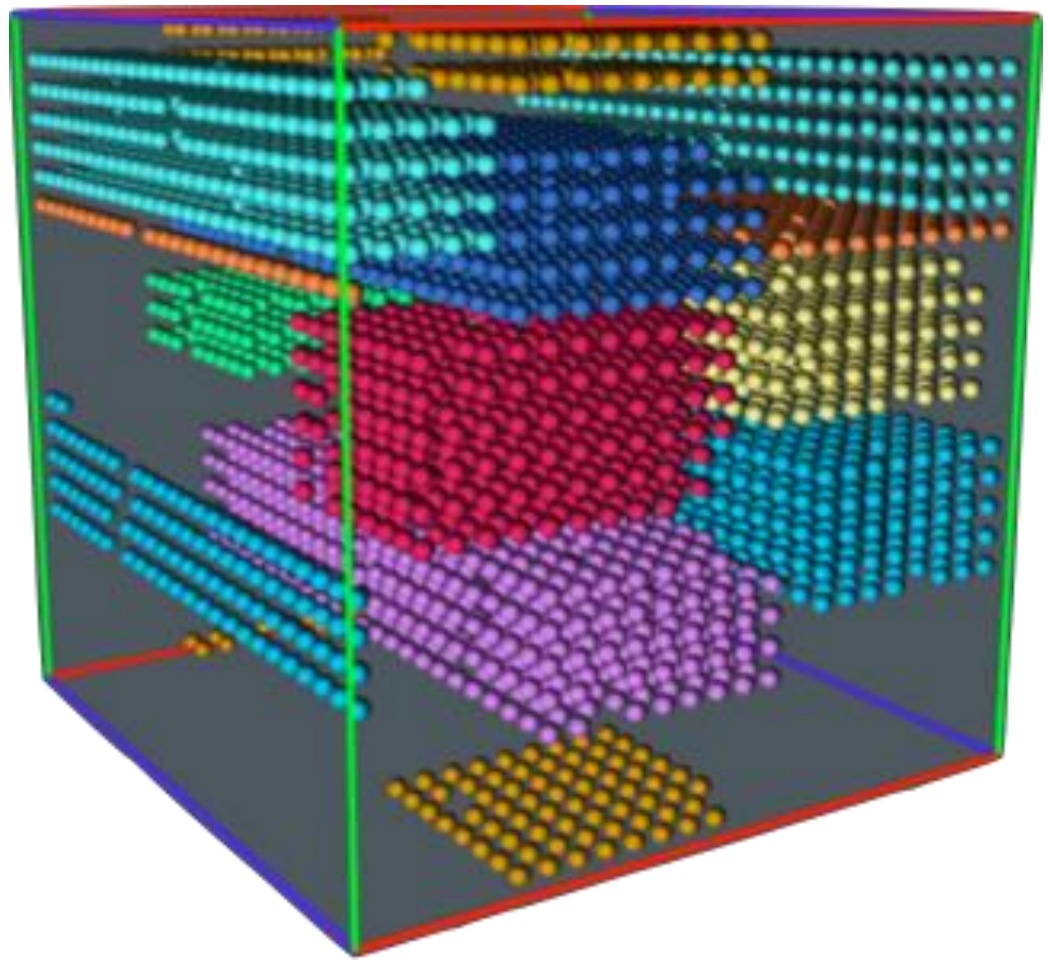

New Streaming Kernel Performance
64
32
16
8
4
Performance (ns per day)
256
512
1024
2048
4096
Number of Nodes
21M atoms
+30%
+10%
+10-30%
Blue Waters XK7 (new streaming)
Blue Waters XK7 (no streaming)
(2fs timestep)

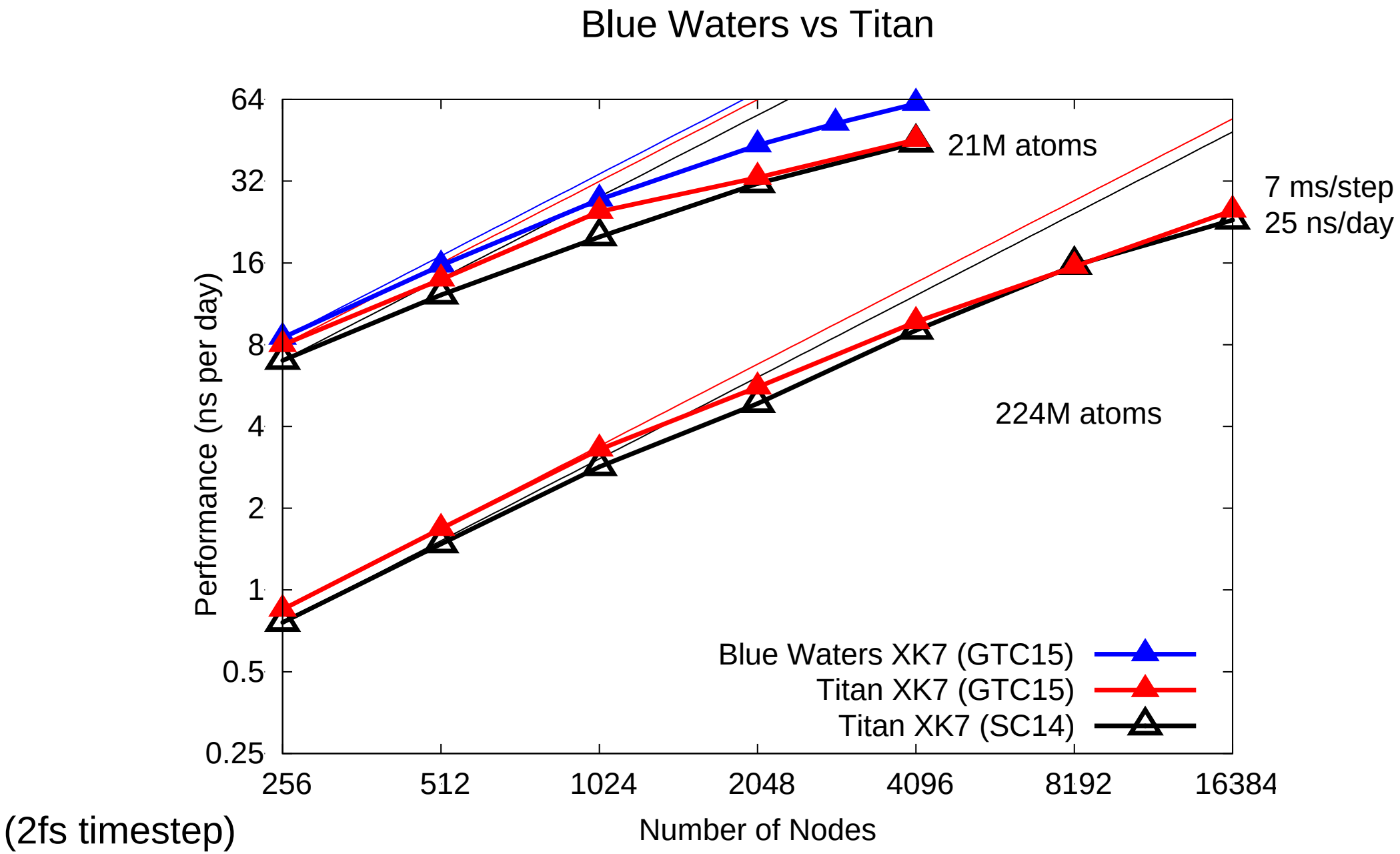

Parallelize PME Within Node
2.9 ms/step
60 ns/day
64
32
16
8
4
Performance (ns per day)
21M atoms
Blue Waters XK7 (new streaming, tuned PME)
Blue Waters XK7 (new streaming)
Blue Waters XK7 (no streaming)
+5%
+10-30%
256
512
1024
2048
4096
Number of Nodes
(2fs timestep)


Blue Waters vs Titan
2.9 ms/step
60 ns/day
-25%
-5%
-5-25%
21M atoms
Performance (ns per day)
64
32
16
8
4
256
512
1024
2048
4096
Number of Nodes
Blue Waters XK7 (new streaming, tuned PME)
Titan XK7 (new streaming, tuned PME)
Titan XK7 (SC14, old streaming)
Blue Waters XK7 (no streaming)
(2fs timestep)

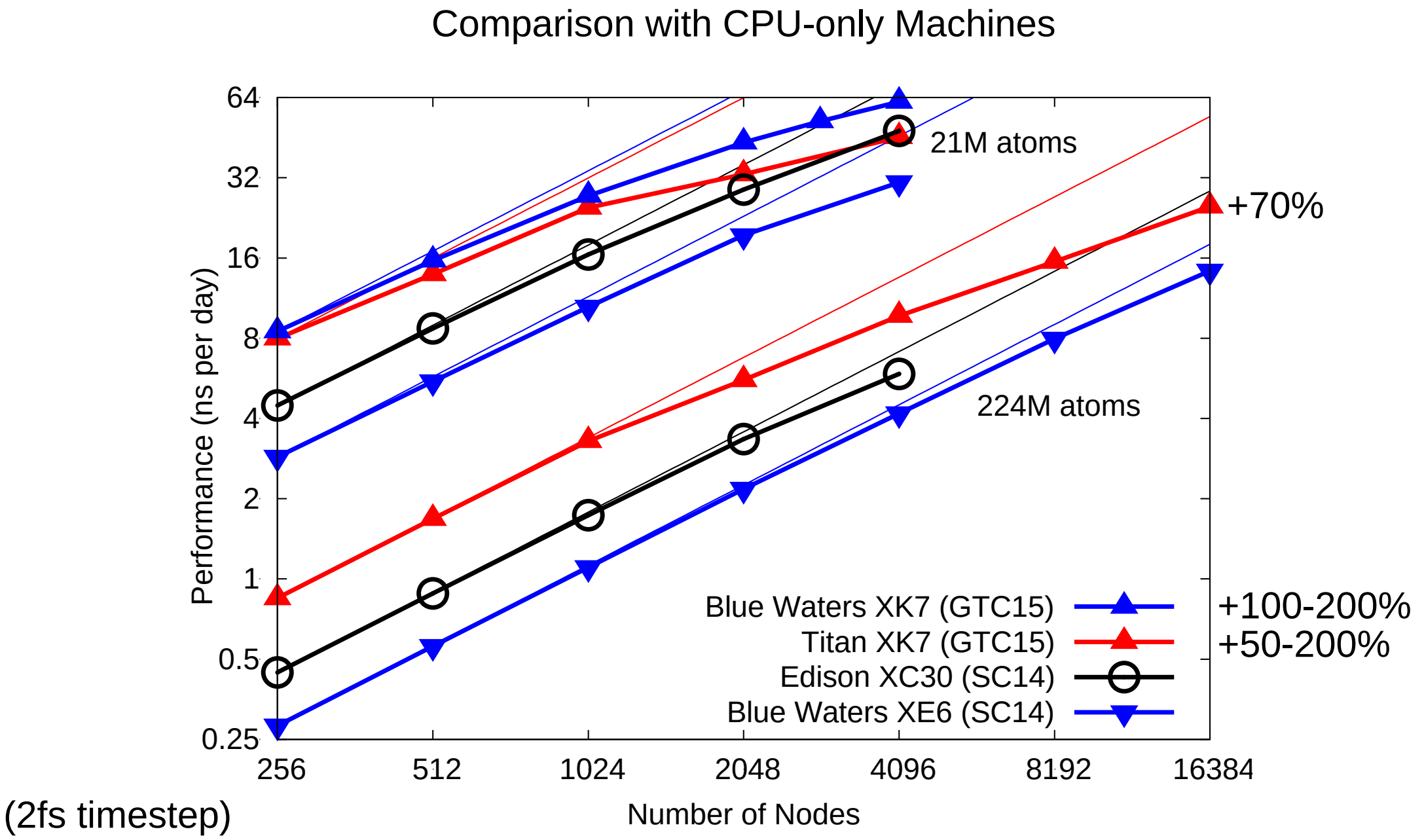
Topology-Aware Scheduling on Blue Waters

- Map jobs to convex sets to avoid network interference
- NCSA, Cray, Adaptive
- Just enabled January 13
- Most likely explanation for Blue Waters performance advantage over Titan
- See Enos *et al.*, CUG 2014





(2fs timestep)


Comparison with CPU-only Machines
Performance (ns per day)
64
32
16
8
4
2
1
0.5
0.25
256
512
1024
2048
4096
8192
16384
Number of Nodes
21M atoms
224M atoms
+70%
+100-200%
+50-200%
Blue Waters XK7 (GTC15)
Titan XK7 (GTC15)
Edison XC30 (SC14)
Blue Waters XE6 (SC14)
(2fs timestep)

Recommended GPU Node Configuration for NAMD Computational Chemistry

Workstation or Single Node Configuration	
# of CPU sockets	2
Cores per CPU socket	6+
CPU speed (Ghz)	2.66+
System memory per socket (GB)	32
GPUs	Kepler K20, K40, K80
# of GPUs per CPU socket	1-2
GPU memory preference (GB)	6-12
GPU to CPU connection	PCIe 3.0 or higher
Server storage	500 GB or higher
Network configuration	Gemini, InfiniBand

Scale to multiple nodes with same single node configuration

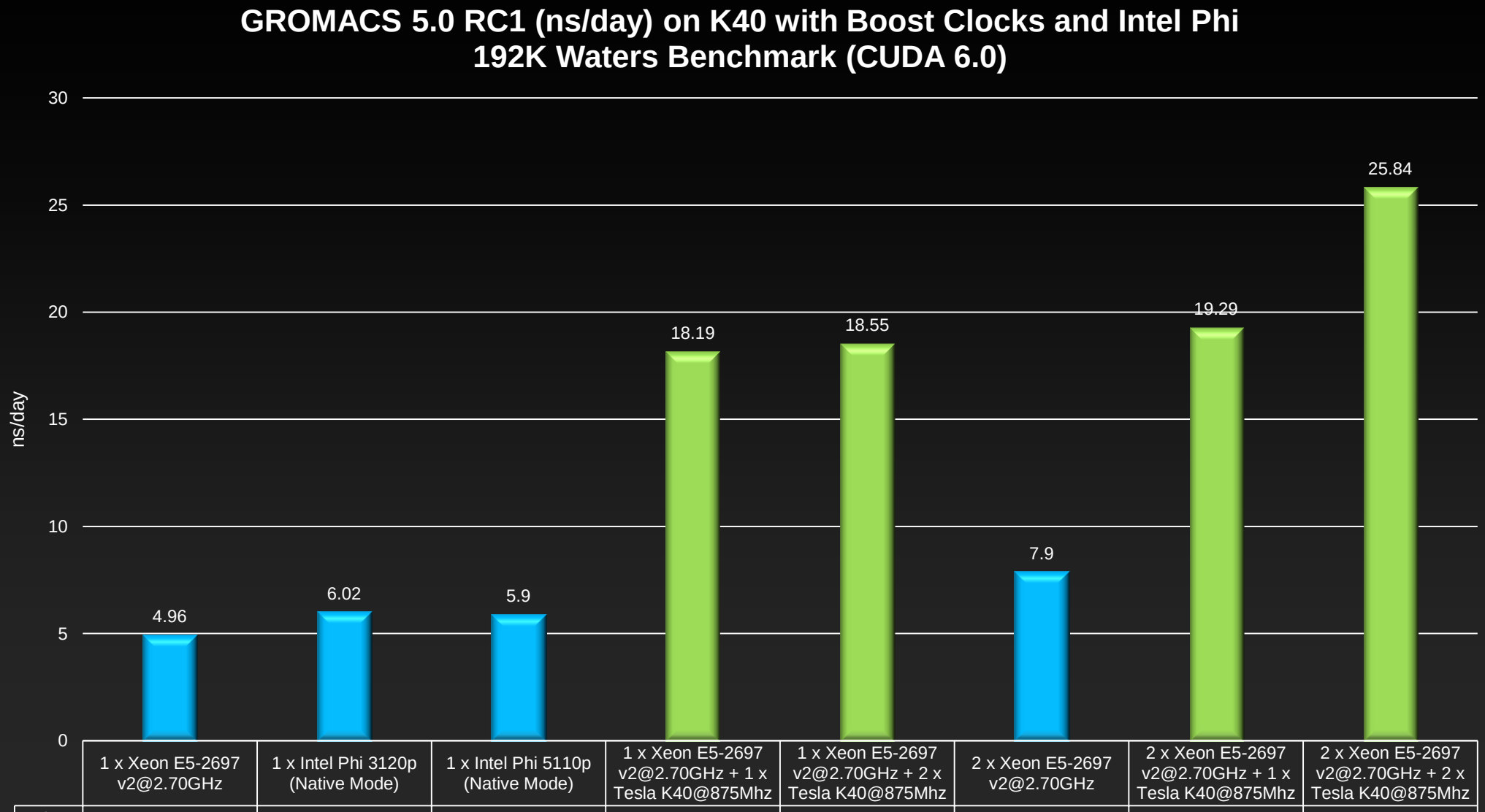
Summary/Conclusions
Benefits of GPU Accelerated Computing

- Faster than CPU only systems in all tests
- Large performance boost with small marginal price increase
- Energy usage cut in half
- GPUs scale very well within a node and over multiple nodes
- Tesla K20 GPU is our fastest and lowest power high performance GPU to date

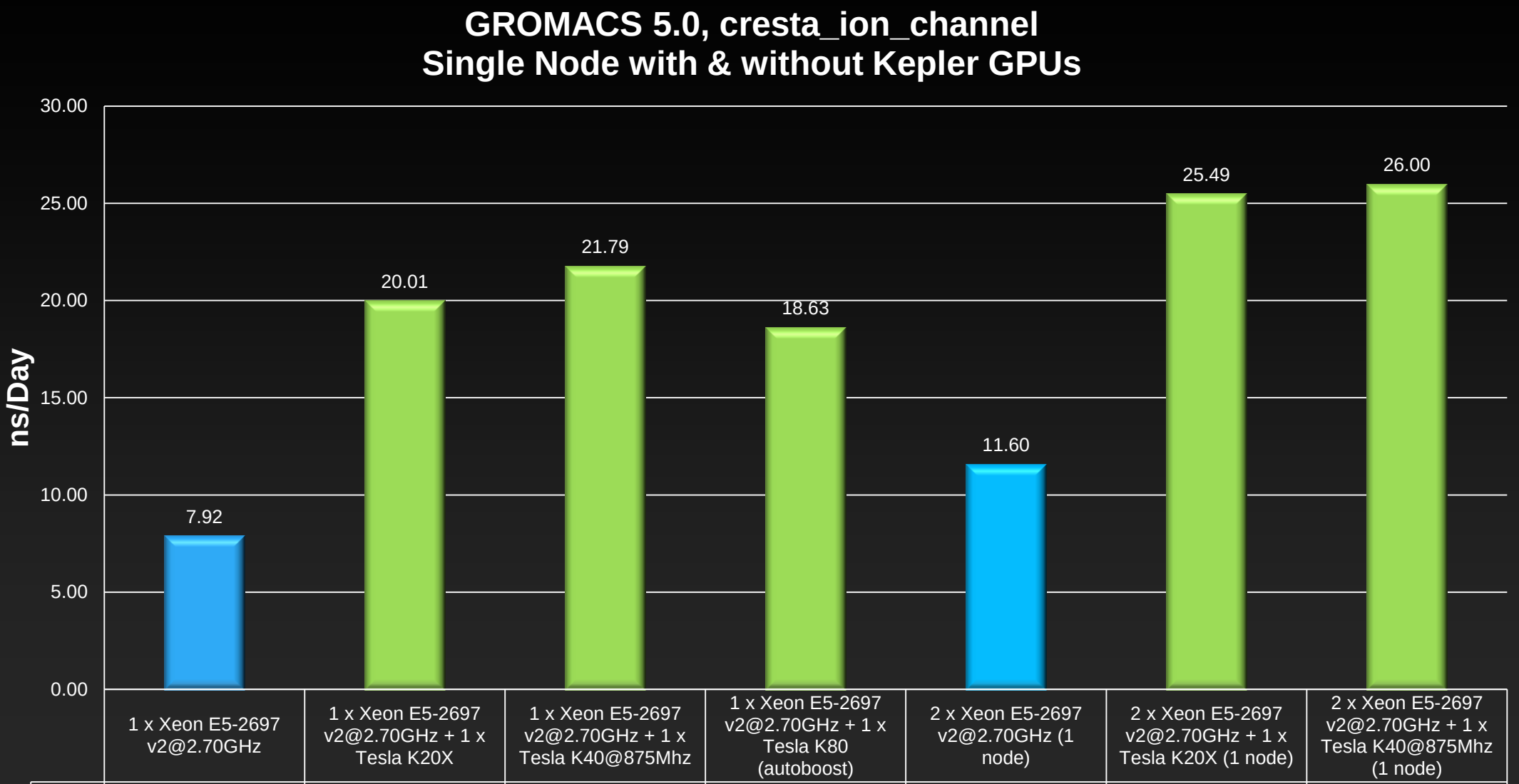
Try GPU accelerated NAMD for free – www.nvidia.com/GPUTestDrive

GROMACS 5.0

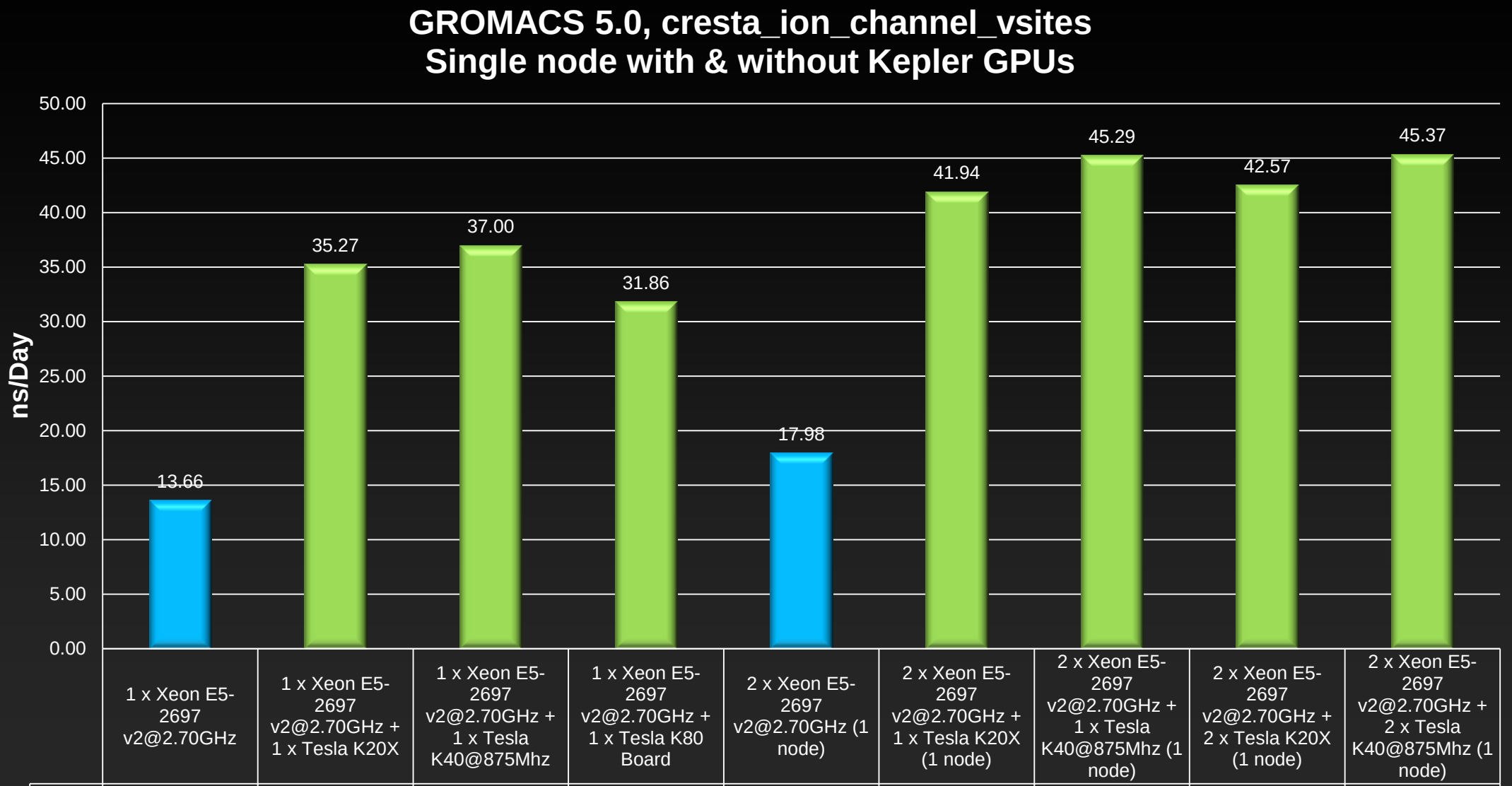
GROMACS 5.0: Phi vs. Kepler K40 fastest GPU!



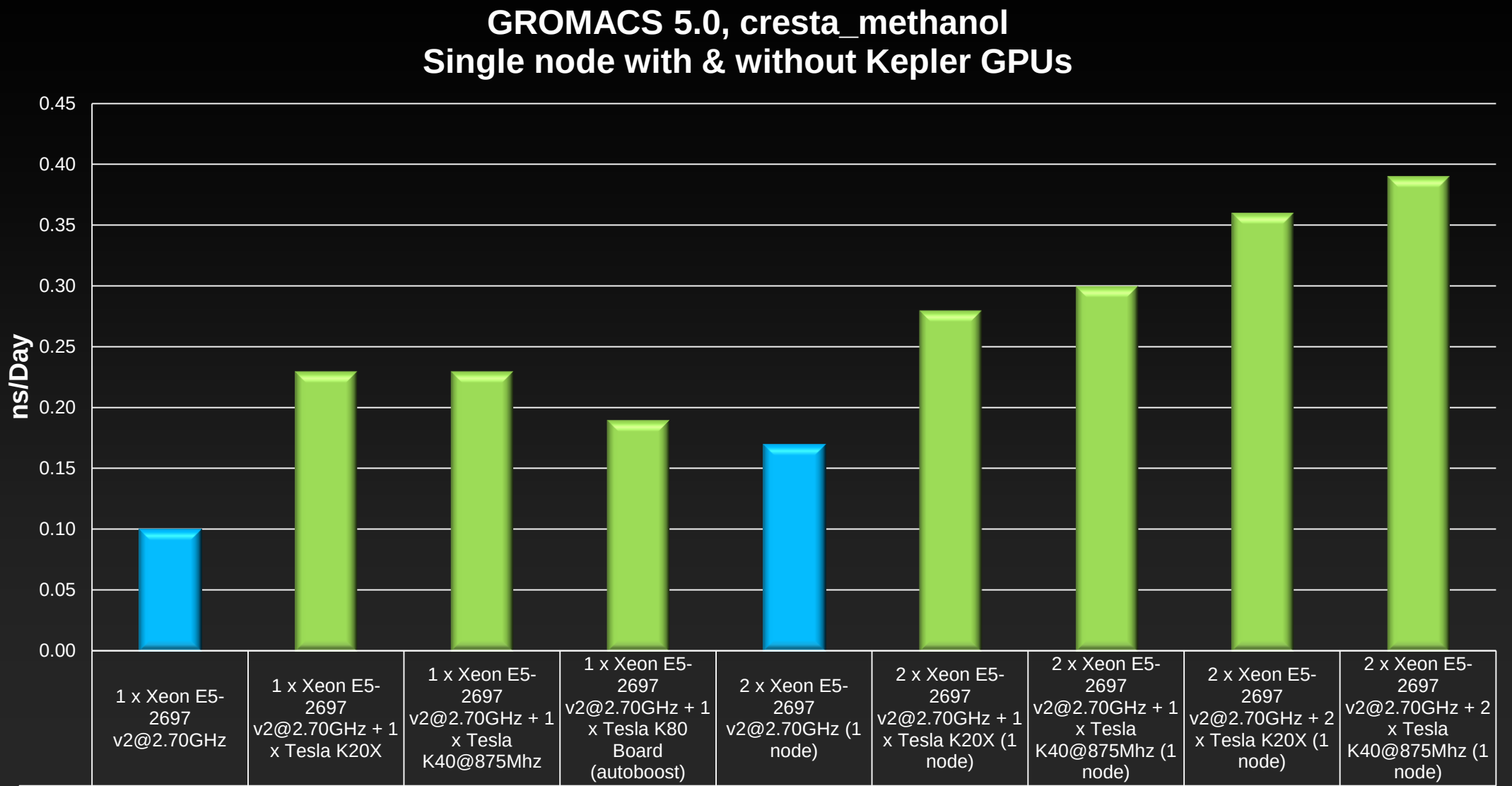
GROMACS 5.0 & Fastest Kepler GPUs yet!



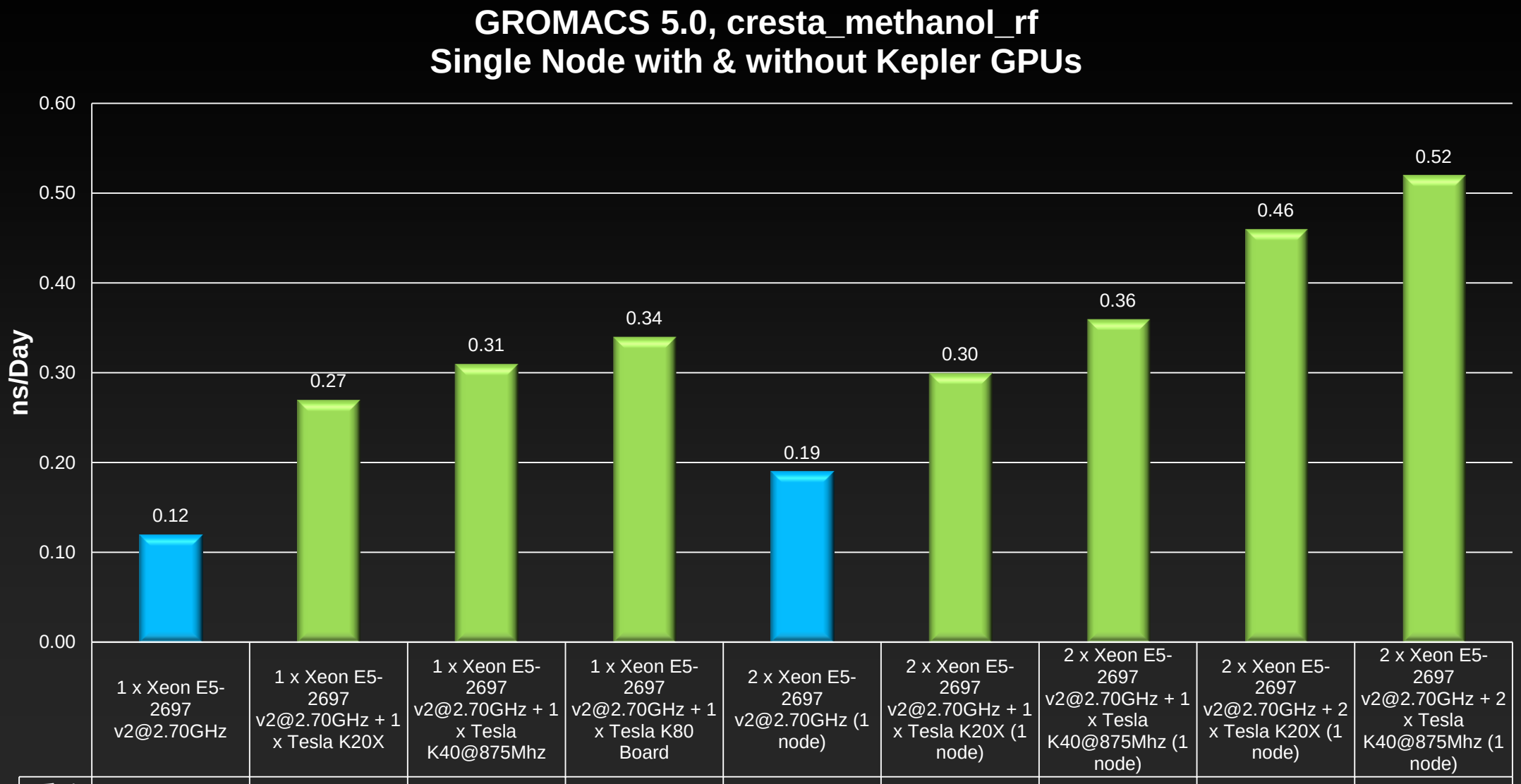
GROMACS 5.0 & Fastest Kepler GPUs yet!



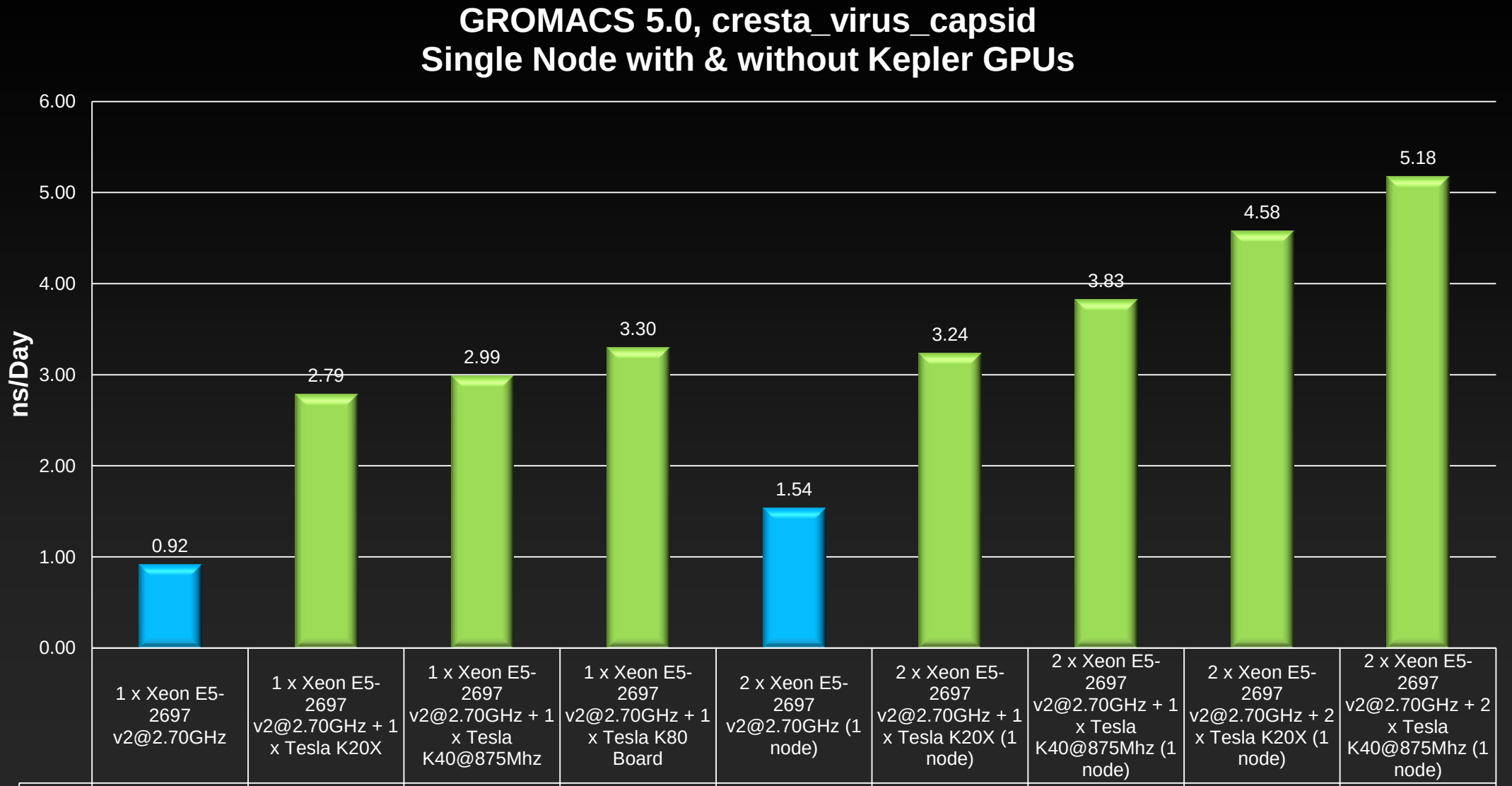
GROMACS 5.0 & Fastest Kepler GPUs yet!



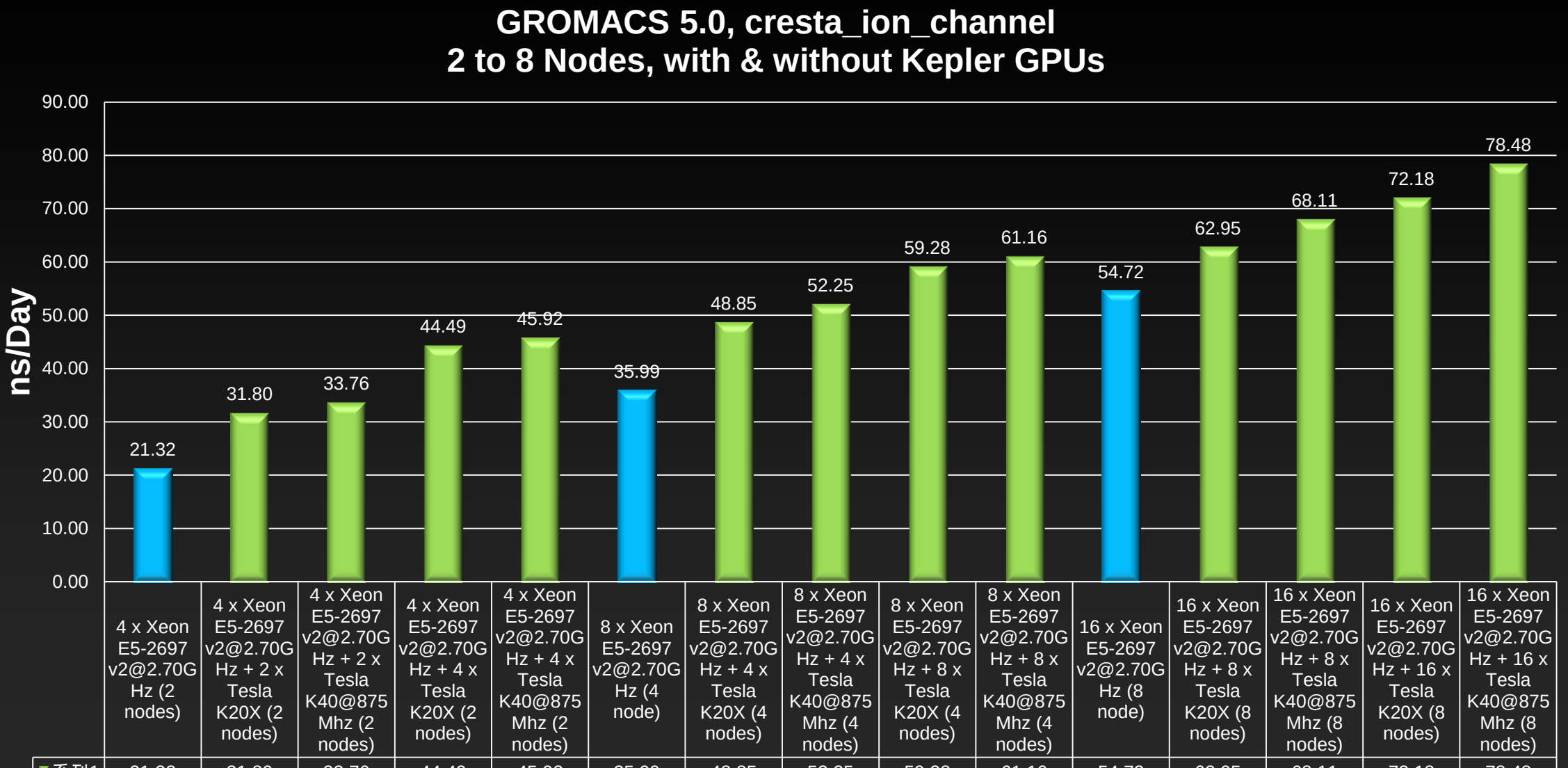
GROMACS 5.0 & Fastest Kepler GPUs yet!



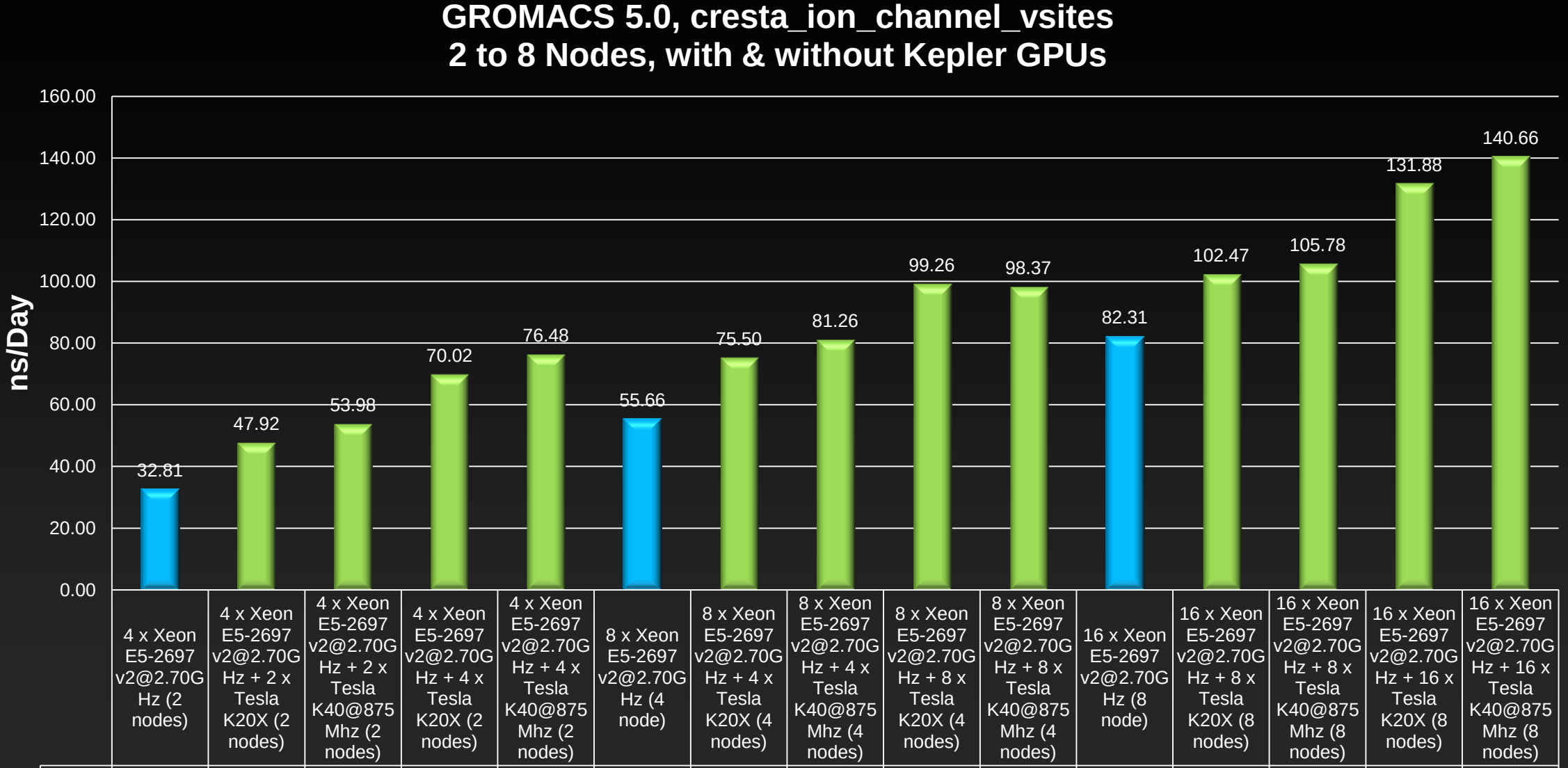
GROMACS 5.0 & Fastest Kepler GPUs yet!



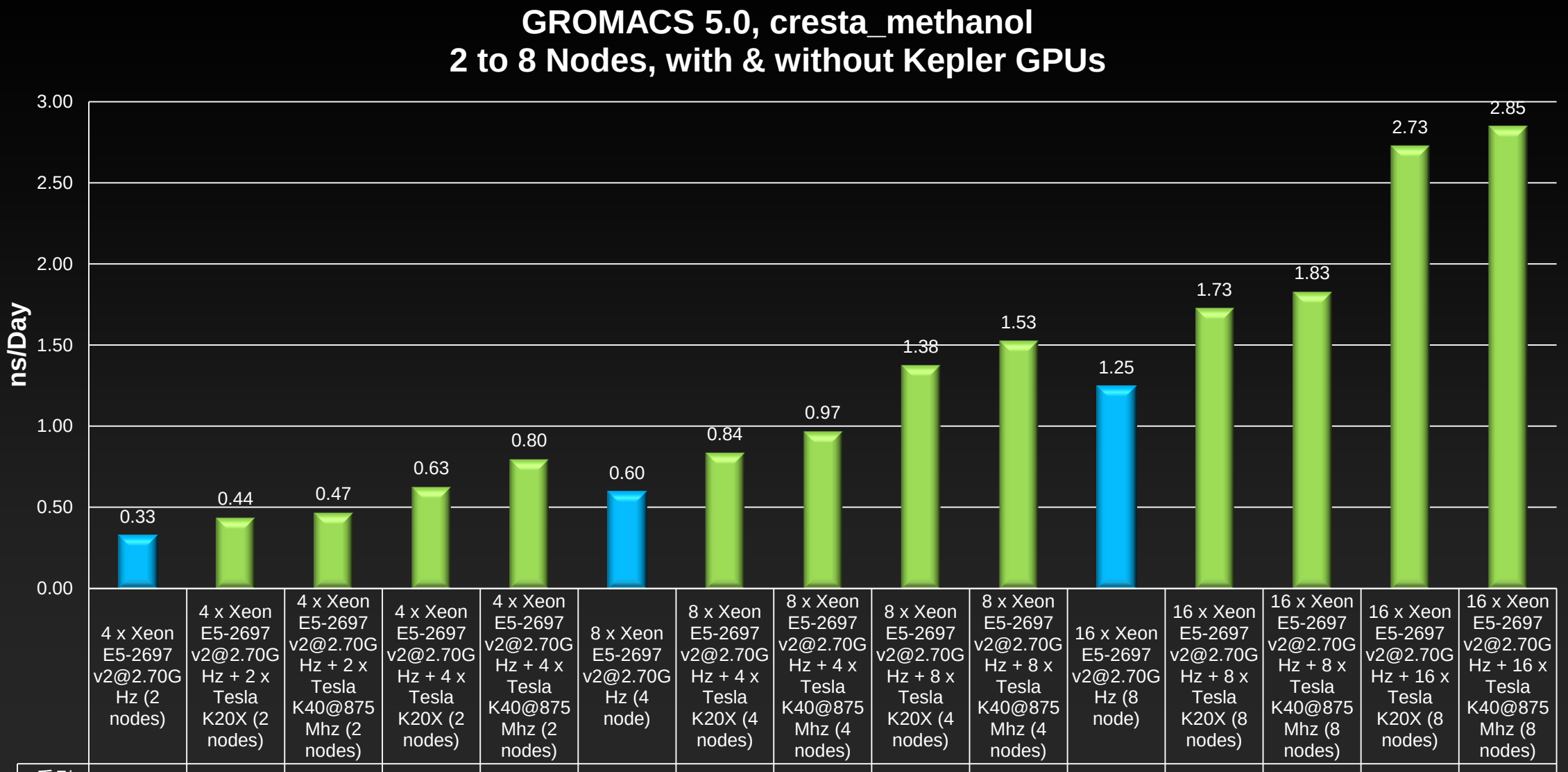
GROMACS 5.0 & Fastest Kepler GPUs yet!



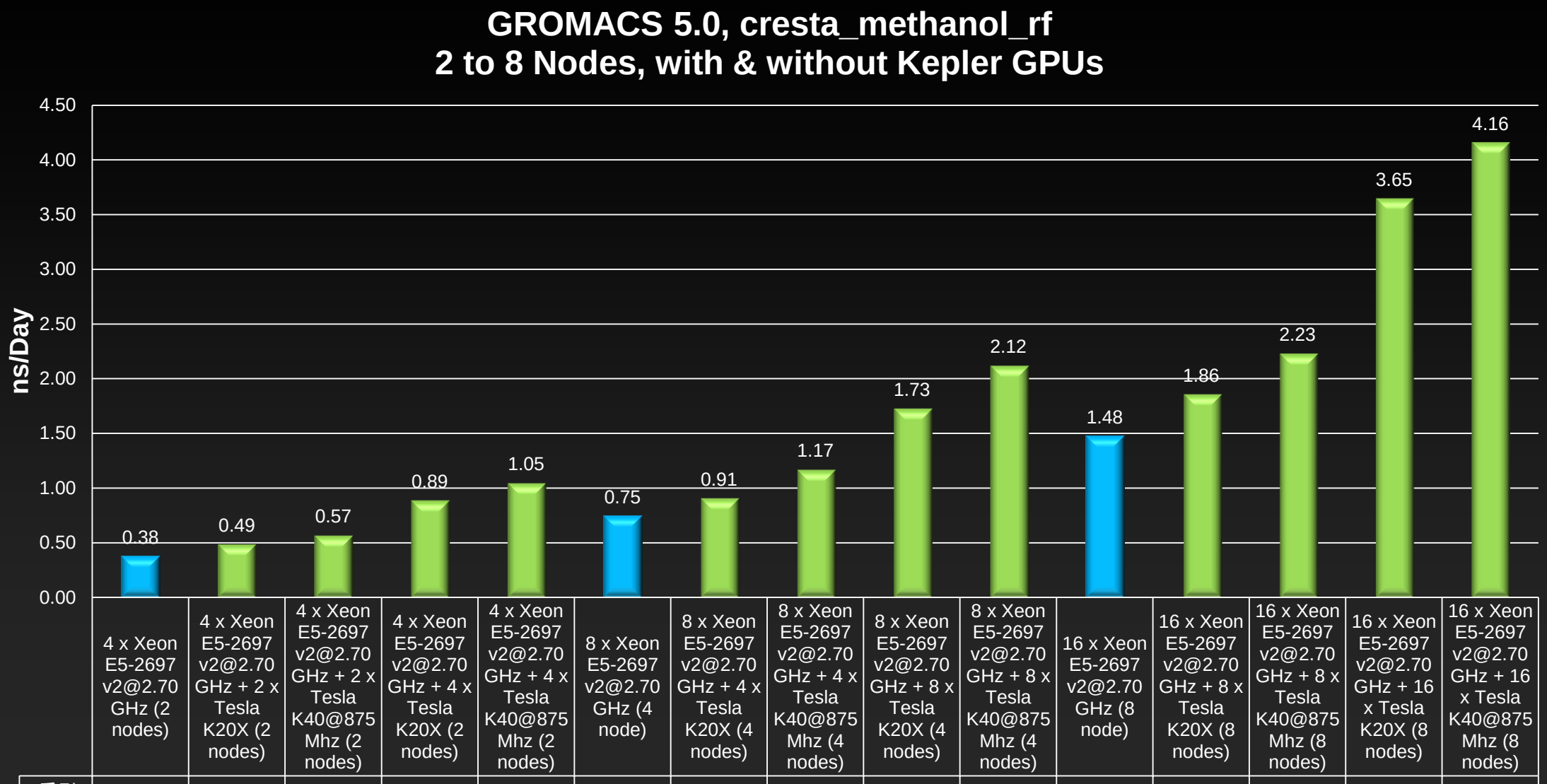
GROMACS 5.0 & Fastest Kepler GPUs yet!



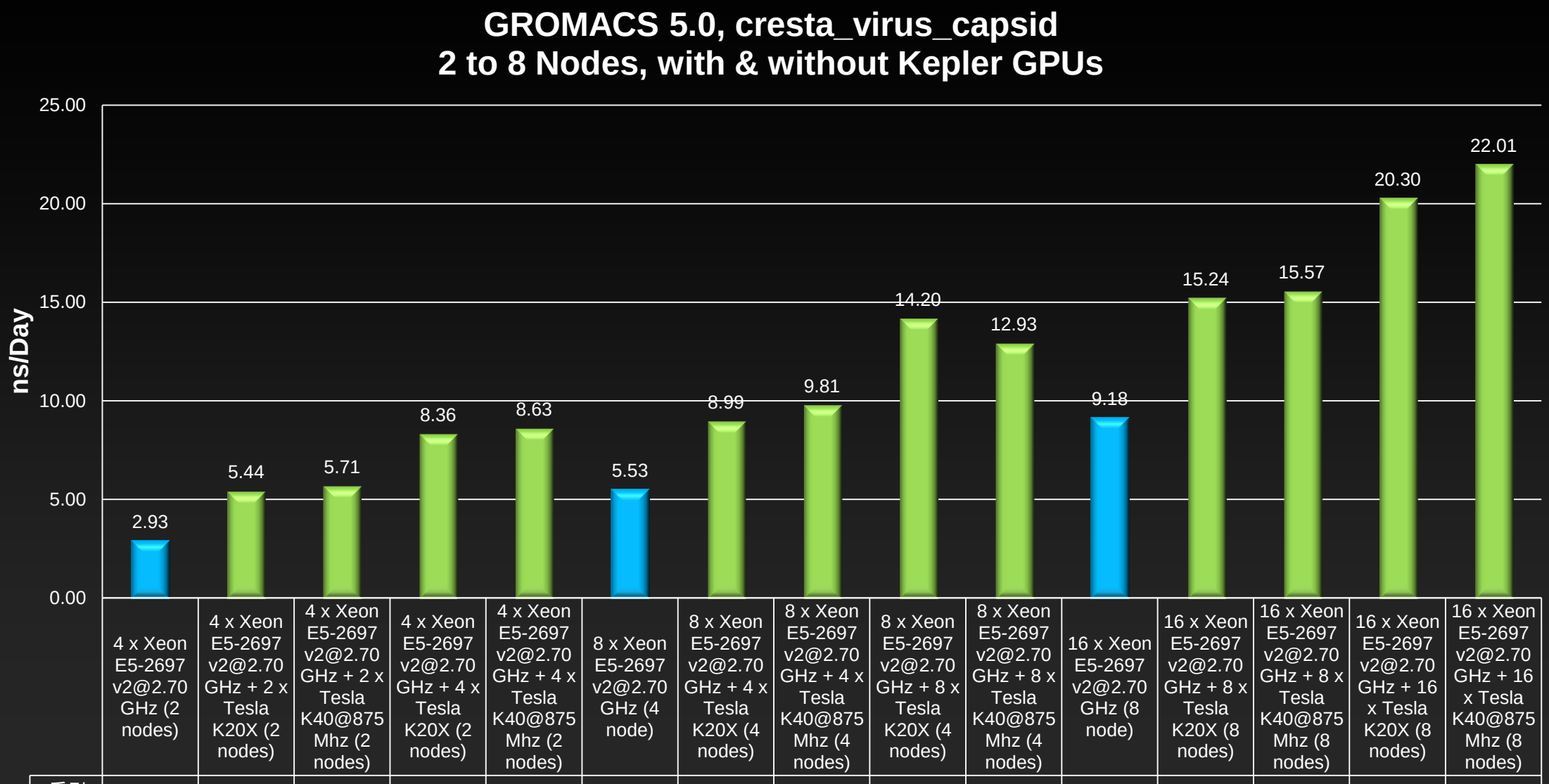
GROMACS 5.0 & Fastest Kepler GPUs yet!



GROMACS 5.0 & Fastest Kepler GPUs yet!



GROMACS 5.0 & Fastest Kepler GPUs yet!

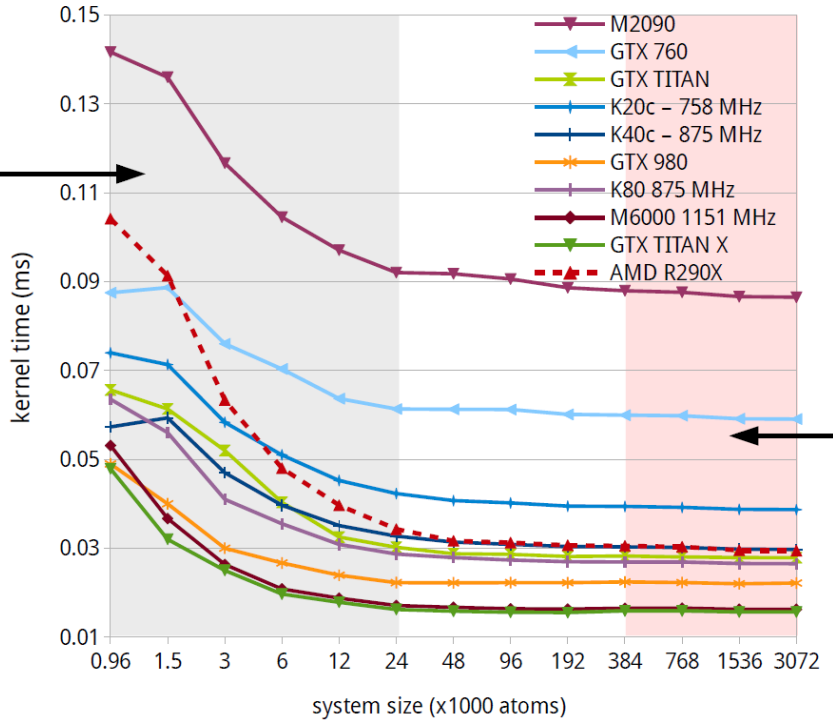


Slides - courtesy of GROMACS Dev Team

Kernel performance and scaling

Strong scaling regime:

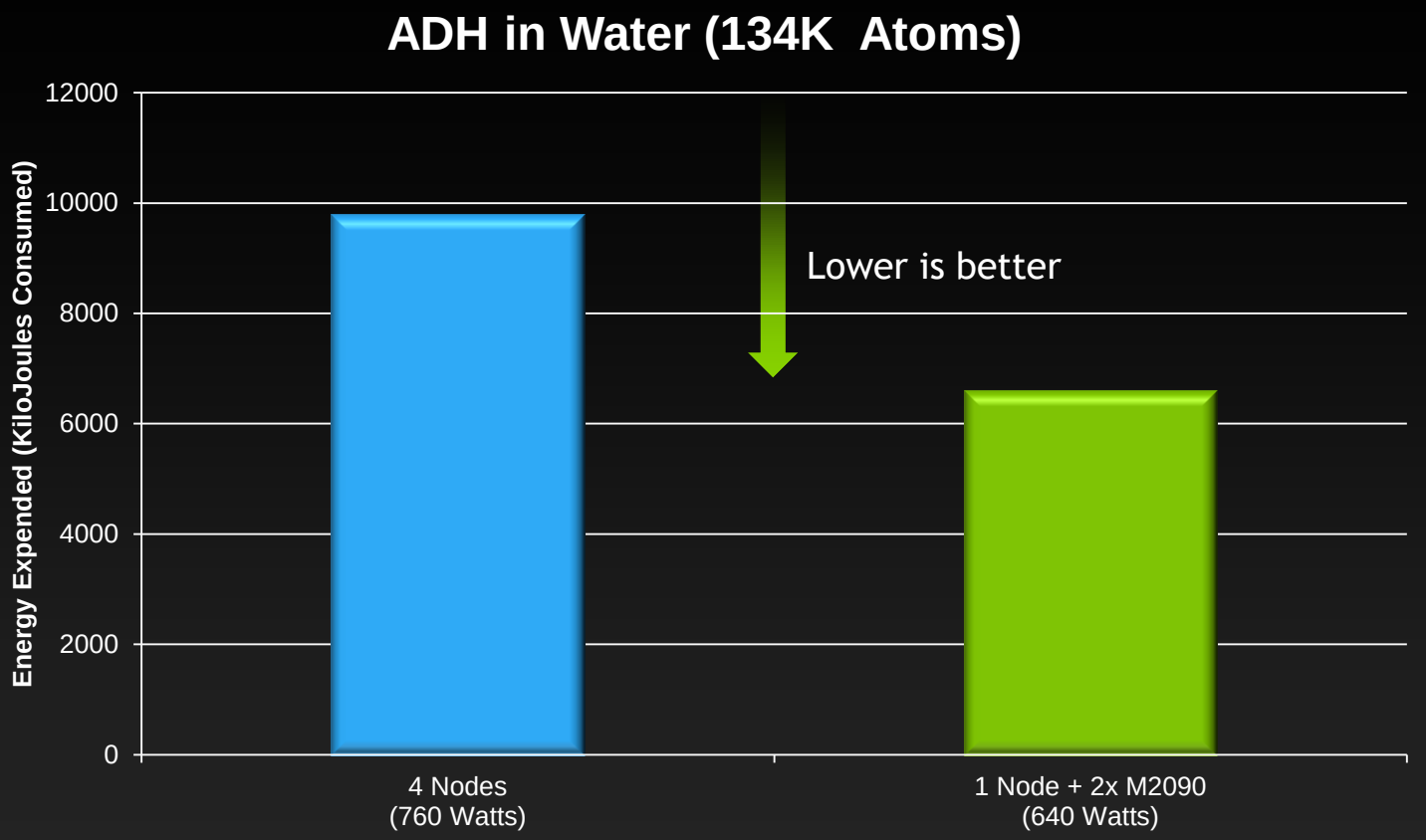
This is where most of our efforts go!



Benchmark "show-off" regime:

This is where the "free lunch" from new hardware comes in full effect

Greener Science



Running GROMACS 4.6 with CUDA 4.1

The blue nodes contain 2x Intel X5550 CPUs (95W TDP, 4 Cores per CPU)

The green node contains 2x Intel X5550 CPUs, 4 Cores per CPU) and 2x NVIDIA M2090s GPUs (225W TDP per GPU)

Energy Expended
= Power x Time

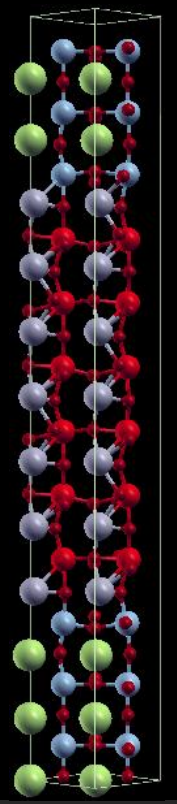
In simulating each nanosecond, the GPU-accelerated system uses 33% less energy

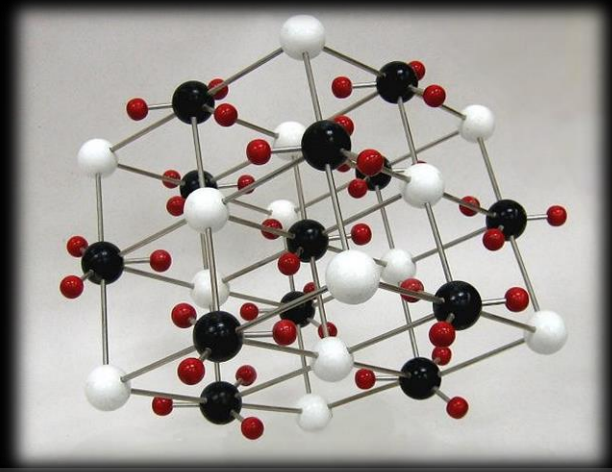
Recommended GPU Node Configuration for GROMACS Computational Chemistry

Workstation or Single Node Configuration	
# of CPU sockets	2
Cores per CPU socket	6+
CPU speed (Ghz)	2.66+
System memory per socket (GB)	32
GPUs	Kepler K20, K40, K80
# of GPUs per CPU socket	1x Kepler GPUs: need fast Sandy Bridge or Ivy Bridge, or high-end AMD Opterons
GPU memory preference (GB)	6
GPU to CPU connection	PCIe 3.0 or higher
Server storage	500 GB or higher
Network configuration	Gemini, InfiniBand

TESLA

Quantum Chemistry Module





GAUSSIAN

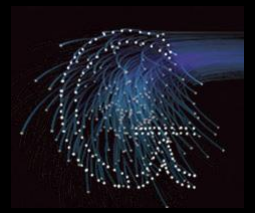
Gaussian

- ACS Fall 2011 press release
  - Joint collaboration between Gaussian, NVDA and PGI for GPU acceleration: http://www.gaussian.com/g_press/nvidia_press.htm
  - No such press release exists for Intel MIC or AMD GPUs
  - Mike Frisch quote from press release:
    - *"Calculations using Gaussian are limited primarily by the available computing resources," said Dr. Michael Frisch, president of Gaussian, Inc. "By coordinating the development of hardware, compiler technology and application software among the three companies, the new application will bring the speed and cost-effectiveness of GPUs to the challenging problems and applications that Gaussian's customers need to address."*

Select Slides from “Enabling Gaussian 09 on GPGPUs” at GTC March 2014

- In 2011 Gaussian, Inc., NVIDIA Corp. and PGI started a long-term project to enable all the performance critical paths of Gaussian on GPGPUs.
  - Ultimate goal is to show significant performance improvement by using accelerators in conjunction with CPUs
  - Initial efforts are directed towards creating an infrastructure that will leverage the current CPU code base and at the same time minimize the additional maintenance effort associated with running on GPUs.
- Current status of this work for Direct Hartree-Fock and triples-correction calculations as applied in for example Coupled Cluster calculations that uses mostly the directives based OpenACC framework.

- Slides & Audio: http://on-demand.gputechconf.com/gtc/2014/video/S4613-enabling-gaussian-09-gpgpus.mp4



PGI

NVIDIA

GPU TECHNOLOGY CONFERENCE

CURRENT STATUS OF THE PROJECT TO ENABLE GAUSSIAN 09 ON GPGPUS

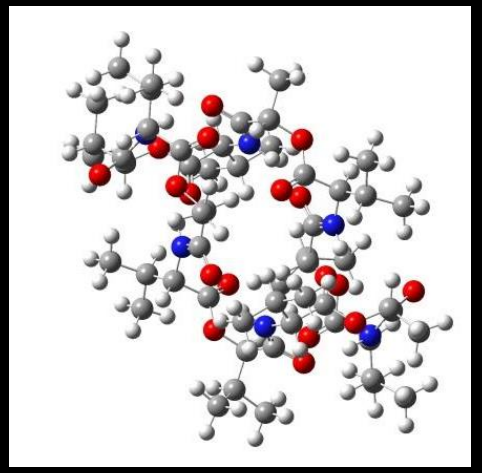
Roberto Gomperts (NVIDIA, Corp.)

Michael Frisch (Gaussian, Inc.)

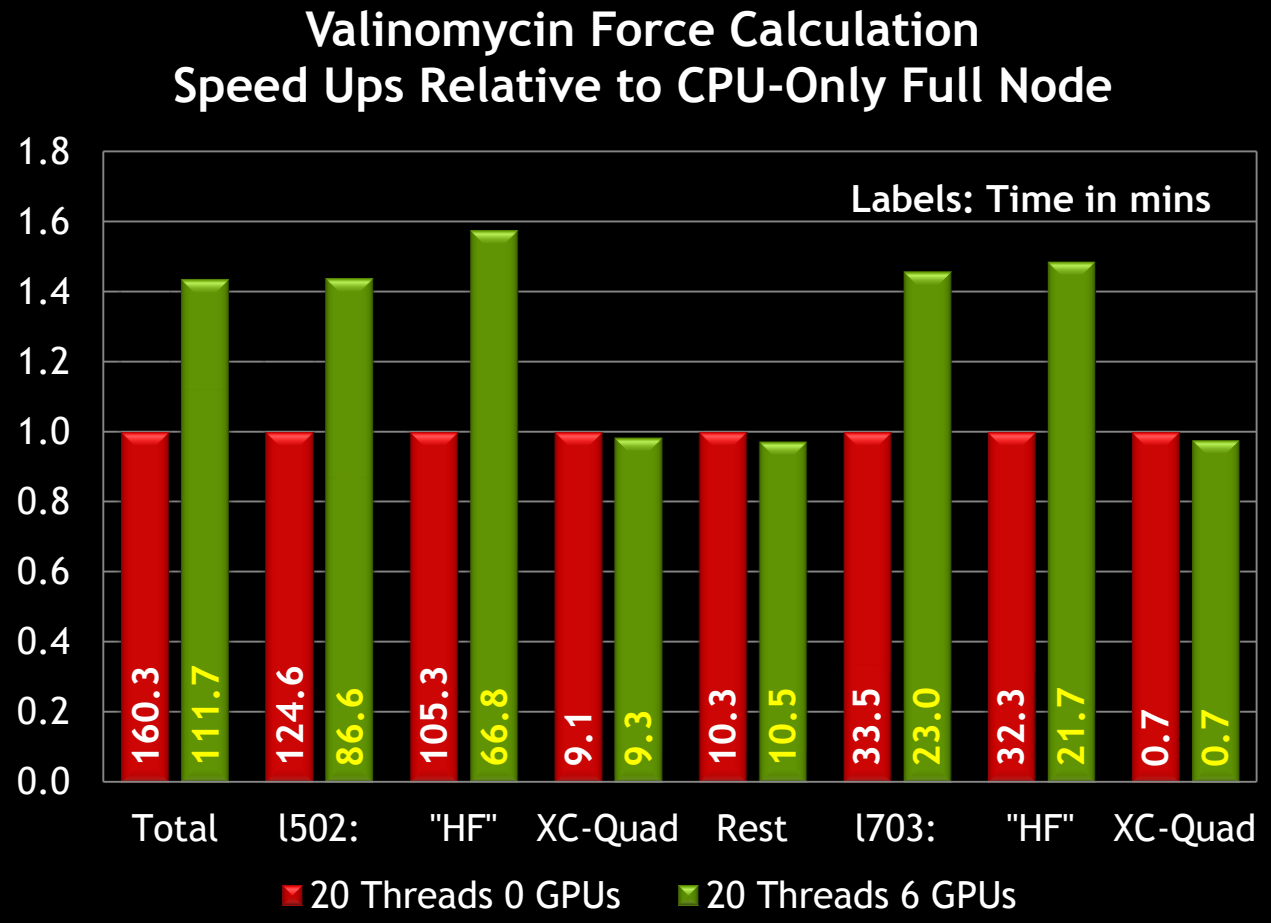
Giovanni Scalmani (Gaussian, Inc.)

Brent Leback (PGI)

EARLY PERFORMANCE RESULTS (DIRECT SCF)



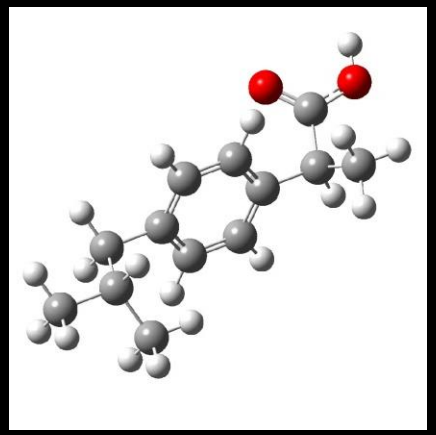
Method	rB3LYP
No. of Atoms	168
Basis Set	6-31G(3df,3p)
No. of Basis Funcs	3 642
No. of Cycles	17



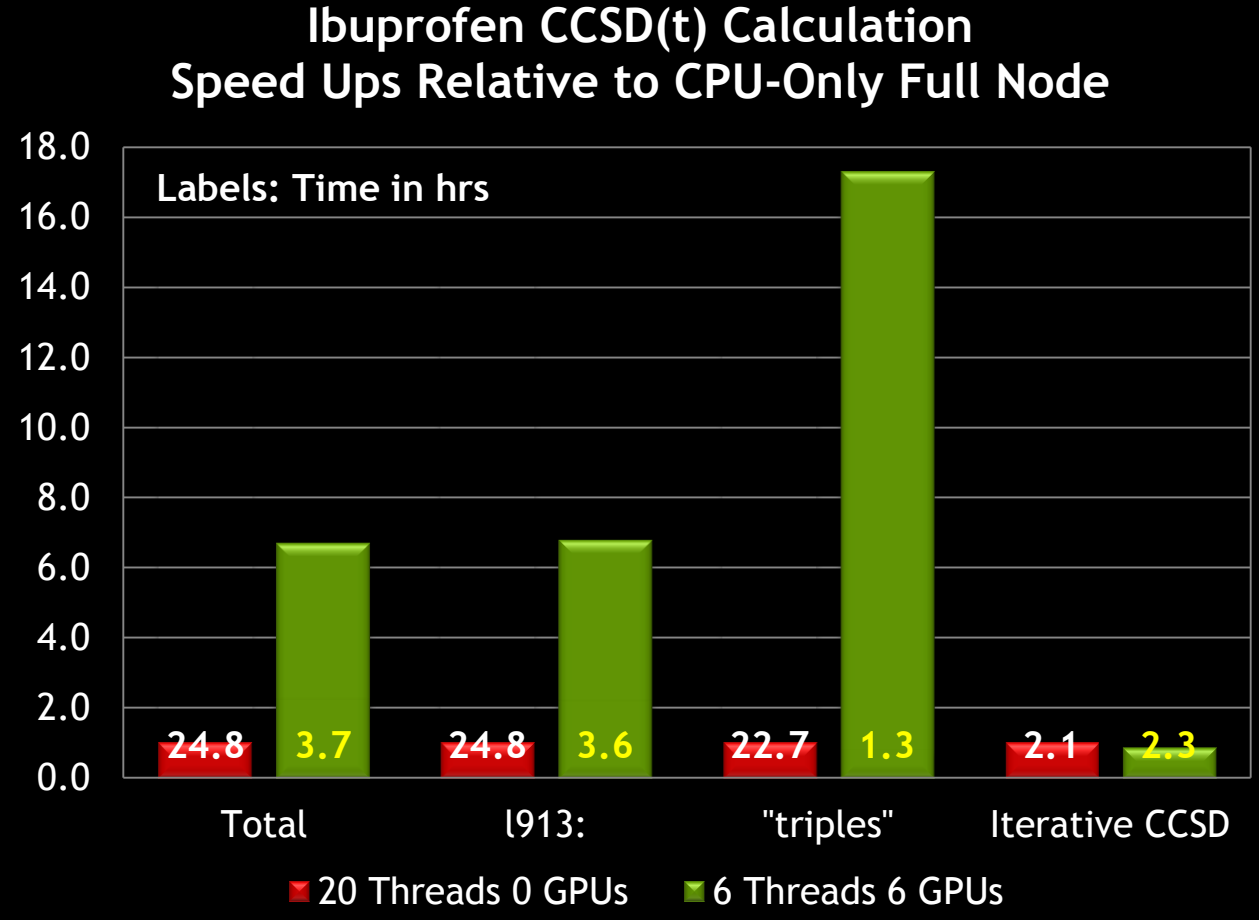
System: 2 Sockets E5-2690 V2 (2x 10 Cores @ 3.0 GHz); 128 GB RAM (DD3-1600); Used 108 GB
GPUs: 6 Tesla K40m (15 SMPs @ 875 MHz); 12 GB Global Memory

GPU TECHNOLOGY CONFERENCE

EARLY PERFORMANCE RESULTS (CCSD(T))



Method	CCSD(t)
No. of Atoms	33
Basis Set	6-31G(d,p)
No. of Basis Funcs	315
No. Occ Orbitals	41
No. Virt Orbitals	259
No. of Cycles	15
No. CCSD iters	16



System: 2 Sockets E5-2690 V2 (2x 10 Cores @ 3.0 GHz); 128 GB RAM (DD3-1600); Used 108 GB
GPUs: 6 Tesla K40m (15 SMPs @ 875 MHz); 12 GB Global Memory

GPU TECHNOLOGY CONFERENCE

CLOSING REMARKS

- Significant progress has been made in creating a framework that keeps an unified code structure for GPU enabled Gaussian
- There is room for performance improvement in the Direct SCF work
- The (t) correction performance looks promising
- Further work: Continue working towards a “product” quality version of Gaussian to be released to customers
  - Continue unification of the code base
  - Tackle non-default paths of the currently enabled code
  - Expand enabling of other Gaussian functionality (2^{nd} Derivatives, XC-quadrature, TDDFT, MP2, etc.)
  - Performance tuning

GPU TECHNOLOGY CONFERENCE

GAMESS

GAMESS Partnership Overview

- **Mark Gordon and Andrey Asadchev, key developers of GAMESS, in collaboration with NVIDIA. Mark Gordon is a recipient of a NVIDIA Professor Partnership Award.**

- **Quantum Chemistry one of major consumers of CPU cycles at national supercomputer centers**

- **NVIDIA developer resources fully allocated to GAMESS code**

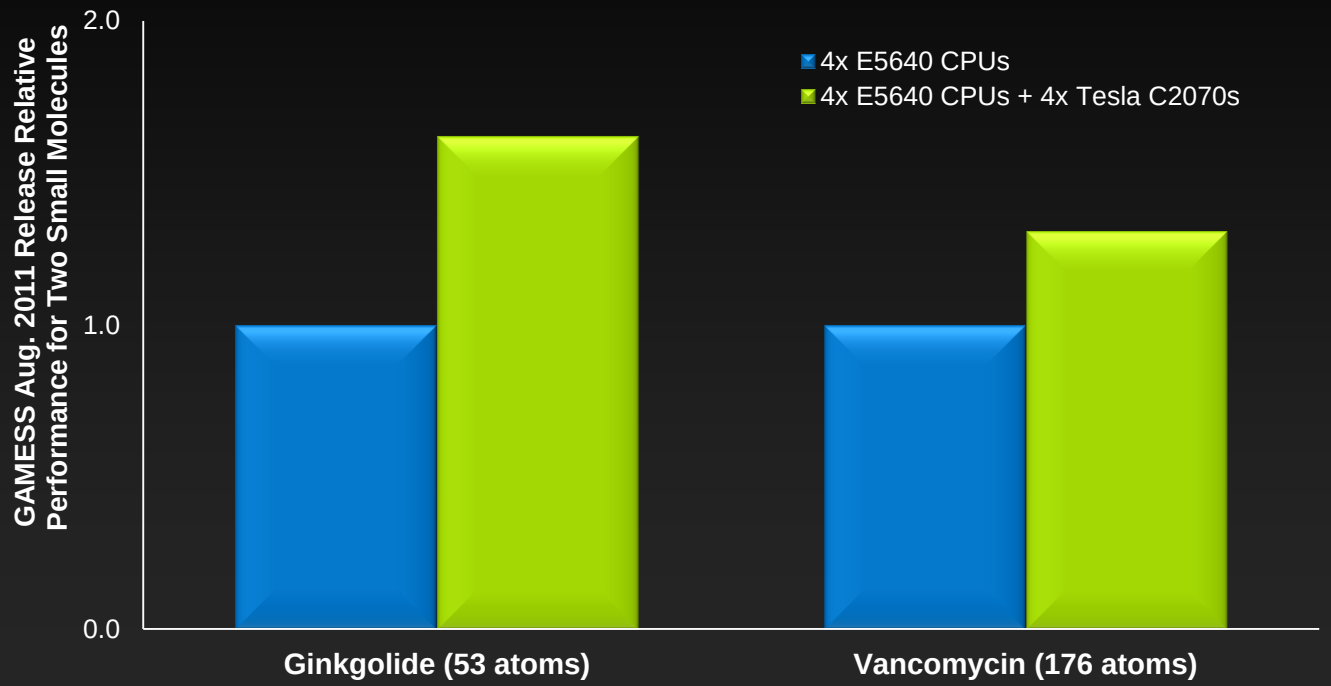
"We like to push the envelope as much as we can in the direction of highly scalable efficient codes. GPU technology seems like a good way to achieve this goal. Also, since we are associated with a DOE Laboratory, energy efficiency is important, and this is another reason to explore quantum chemistry on GPUs."

Prof. Mark Gordon

Distinguished Professor, Department of Chemistry, Iowa State University and Director, Applied Mathematical Sciences Program, AMES Laboratory

GAMESS August 2011 GPU Performance

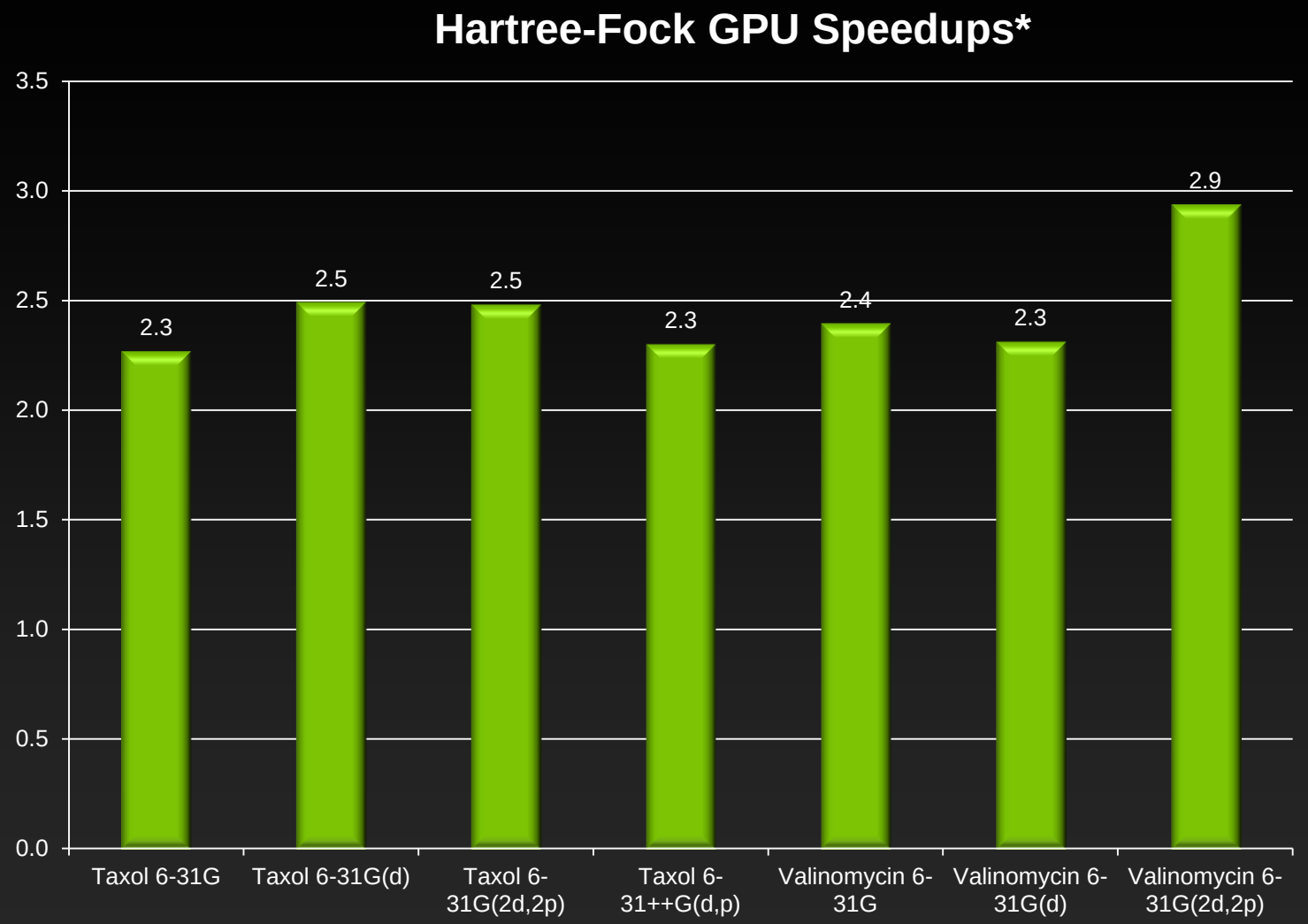
- First GPU supported GAMESS release via "libqc", a library for fast quantum chemistry on multiple NVIDIA GPUs in multiple nodes, with CUDA software
- 2e- AO integrals and their assembly into a closed shell Fock matrix



Upcoming GAMESS Q4 2013 Release

- Multi-nodes with multi-GPUs supported
- Rys Quadrature
- Hartree-Fock
  - 8 CPU cores: 8 CPU cores + M2070 yields 2.3-2.9x speedup. See 2012 publication
- Møller–Plesset perturbation theory (MP2):
  - Paper published
- Coupled Cluster SD(T): CCSD code completed, (T) in progress

GAMESS - New Multithreaded Hybrid CPU/GPU Approach to H-F

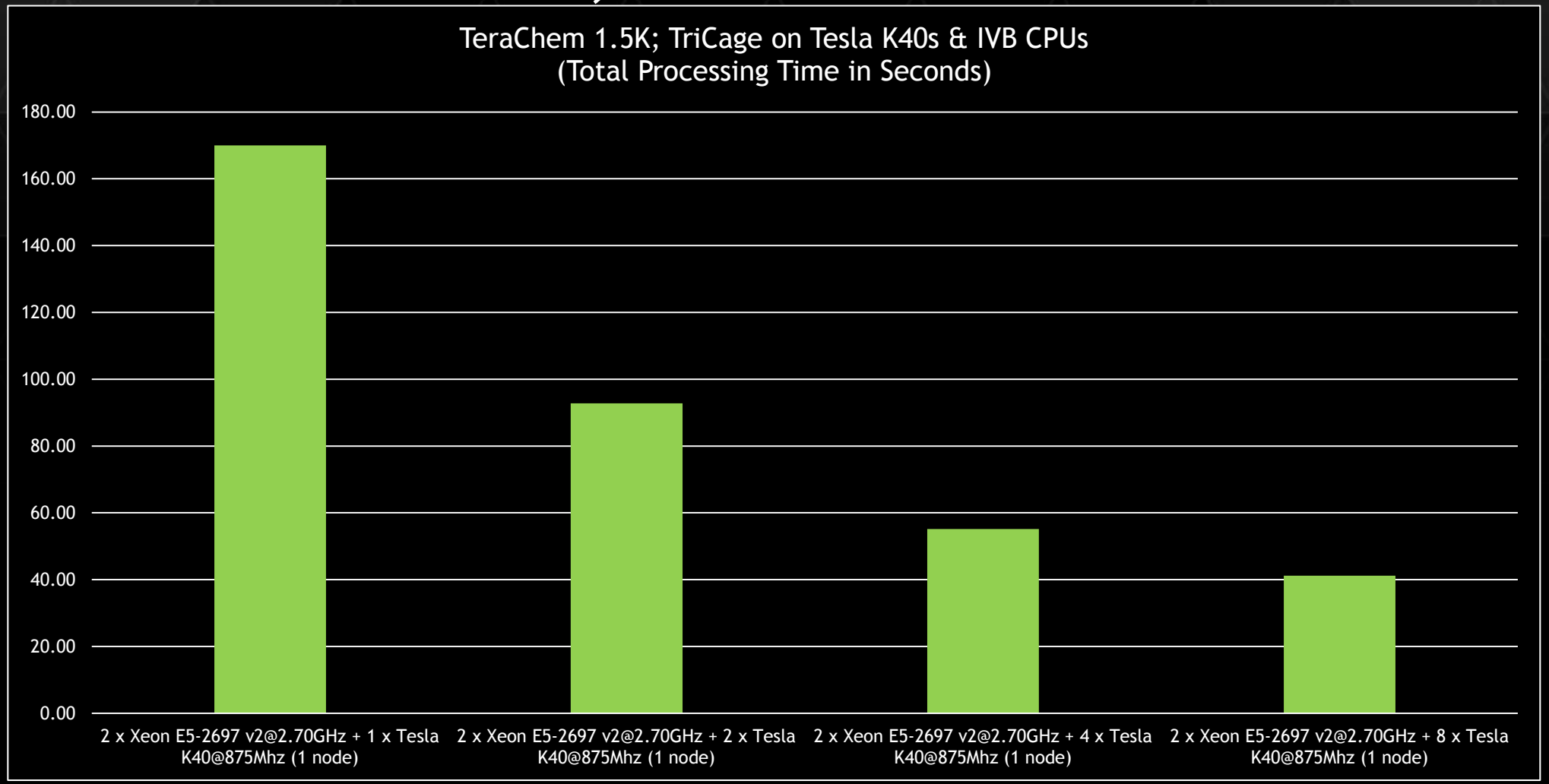


Adding 1x 2070 GPU speeds up computations by 2.3x to 2.9x

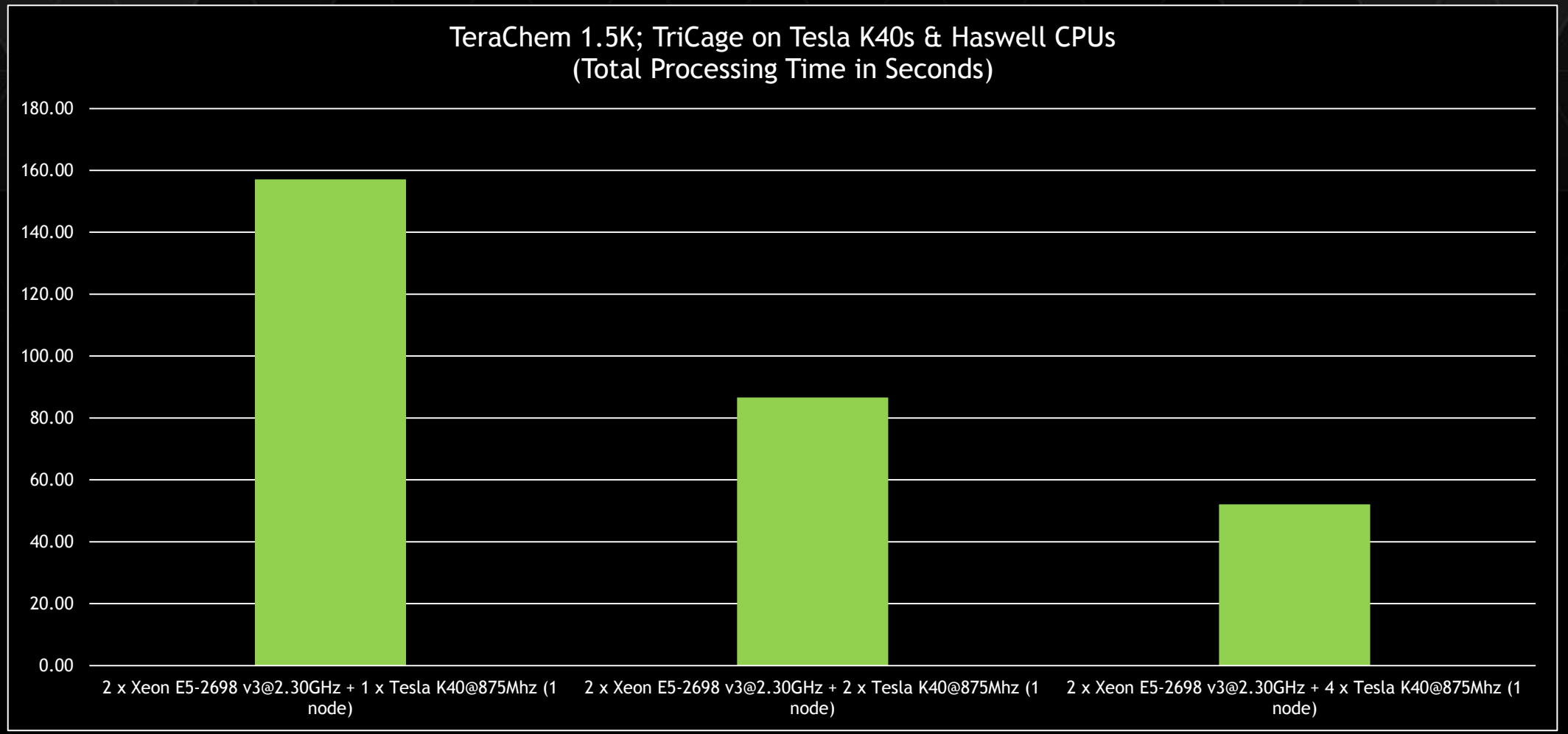
* A. Asadchev, M.S. Gordon, "New Multithreaded Hybrid CPU/GPU Approach to Hartree-Fock," Journal of Chemical Theory and Computation (2012)

TeraChem

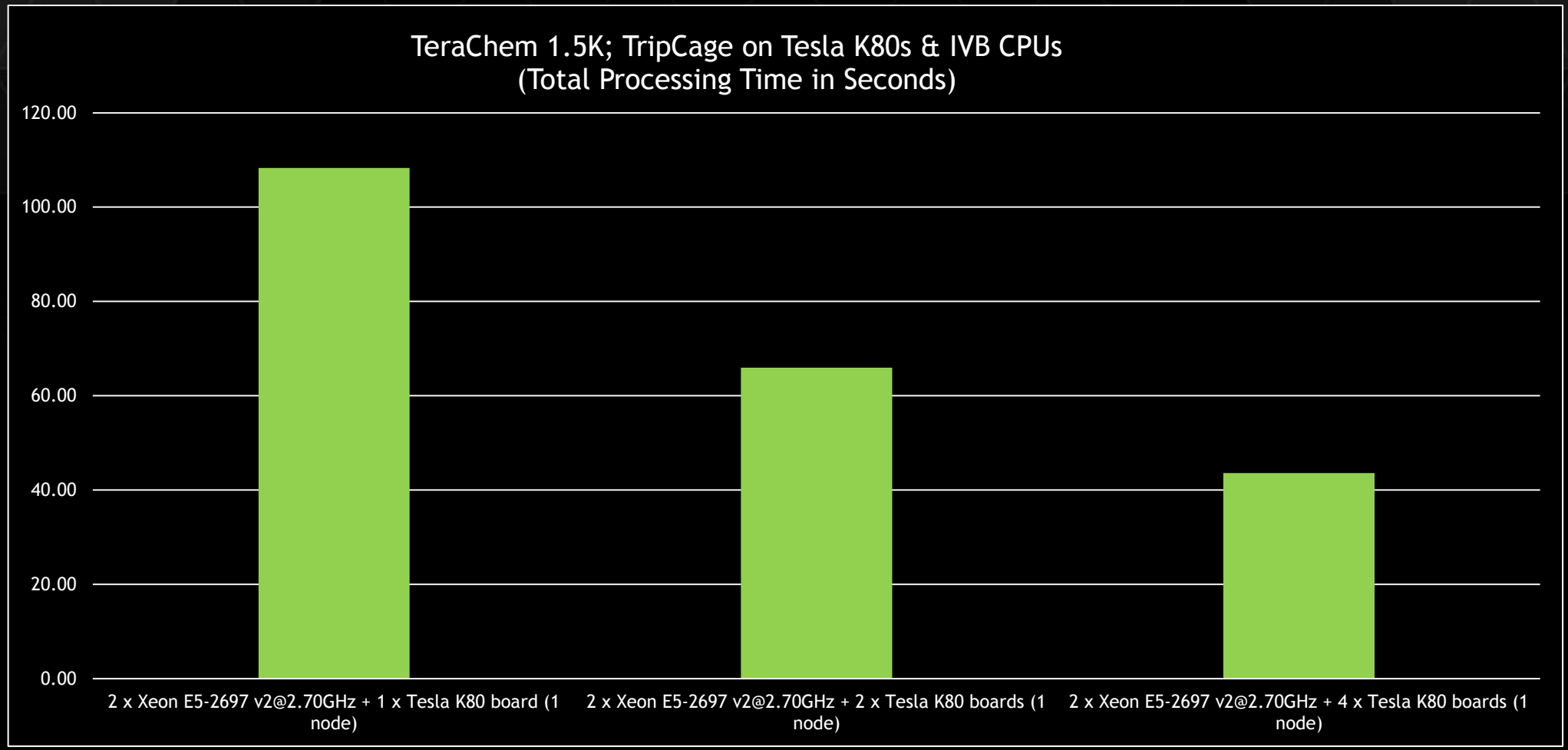
TERACHEM 1.5K; TRPCAGE ON TESLA K40S



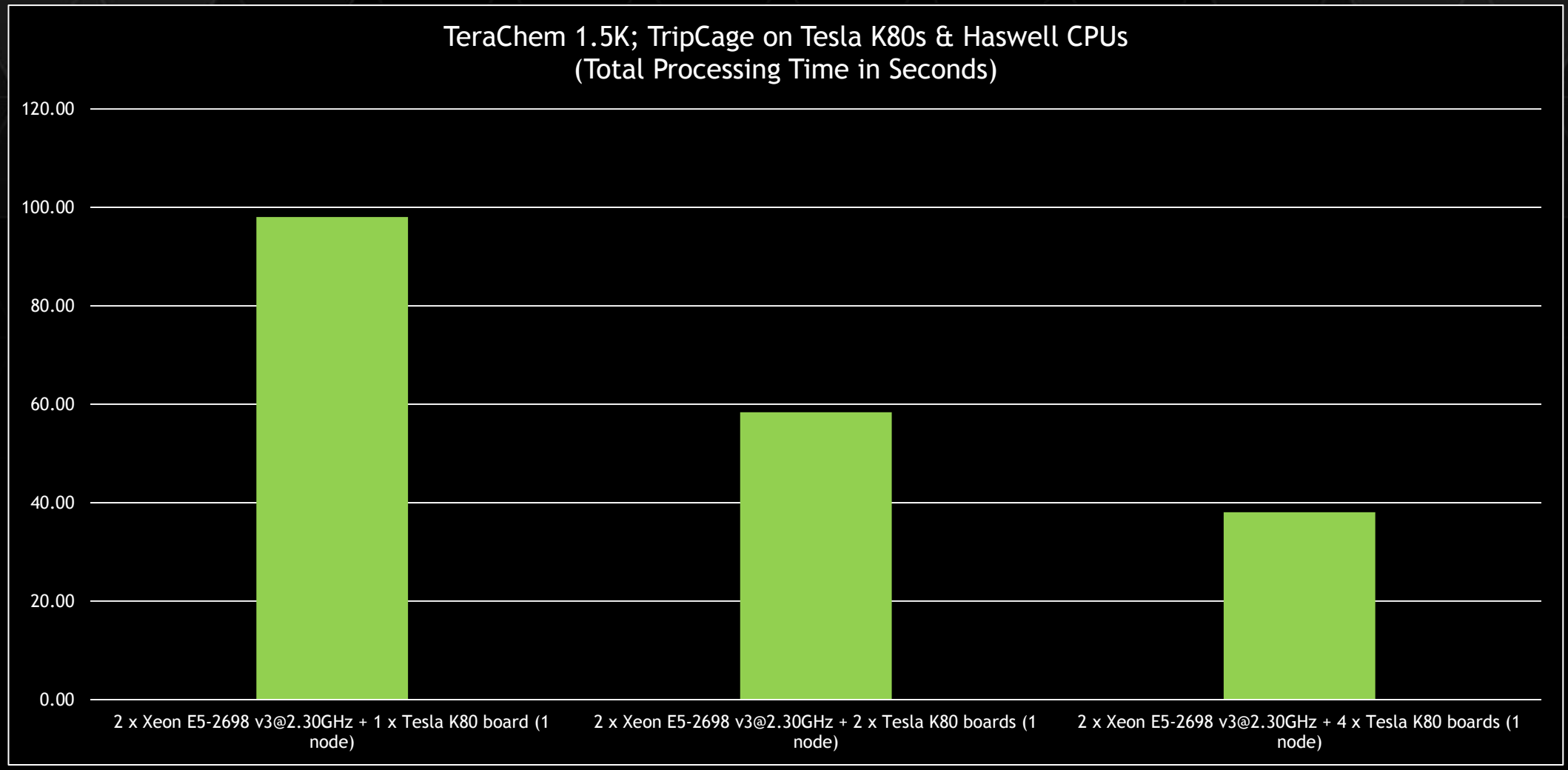
TERACHEM 1.5K; TRPCAGE ON TESLA K40S & HASWELL CPUS



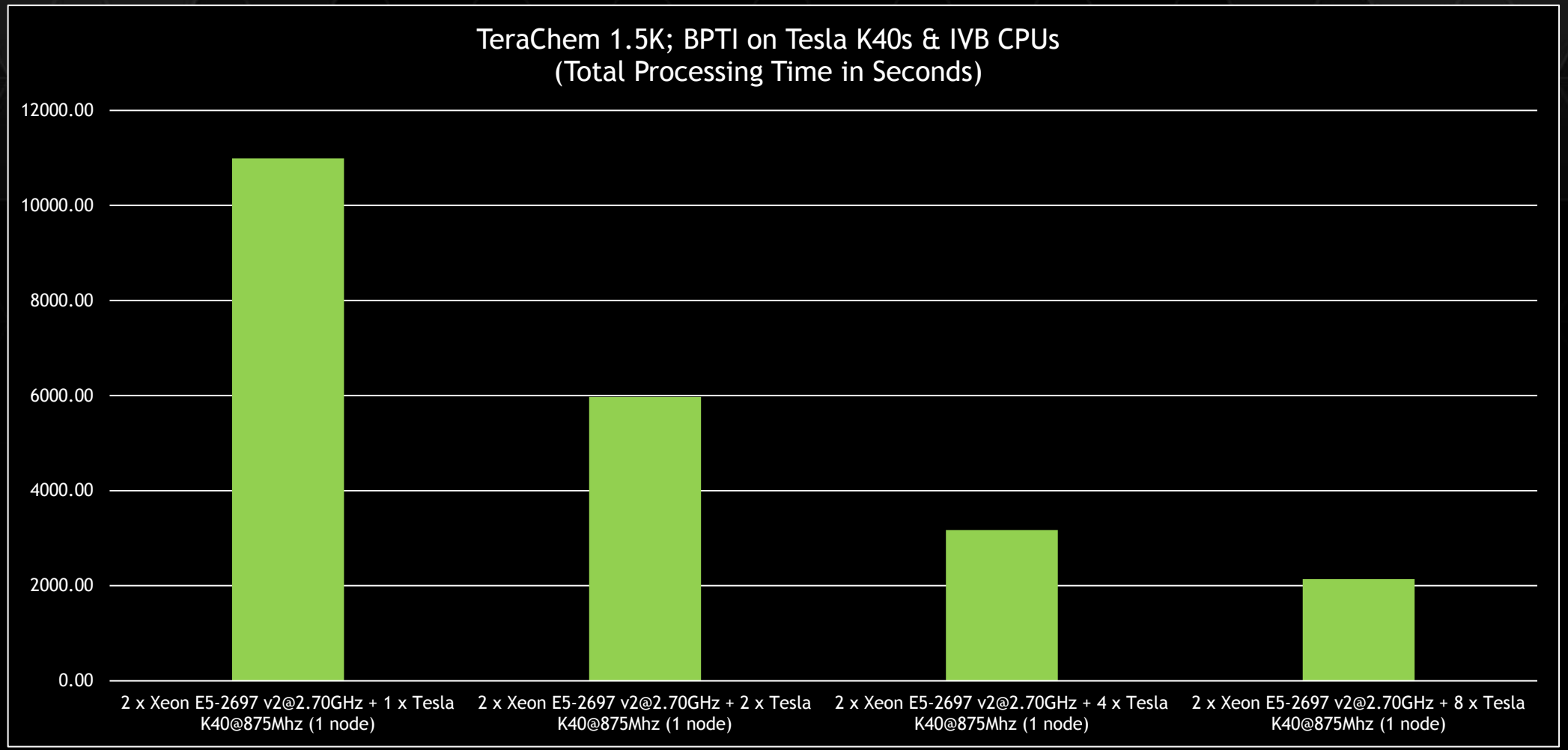
TERACHEM 1.5K; TRPCAGE ON TESLA K80S & IVB CPUS



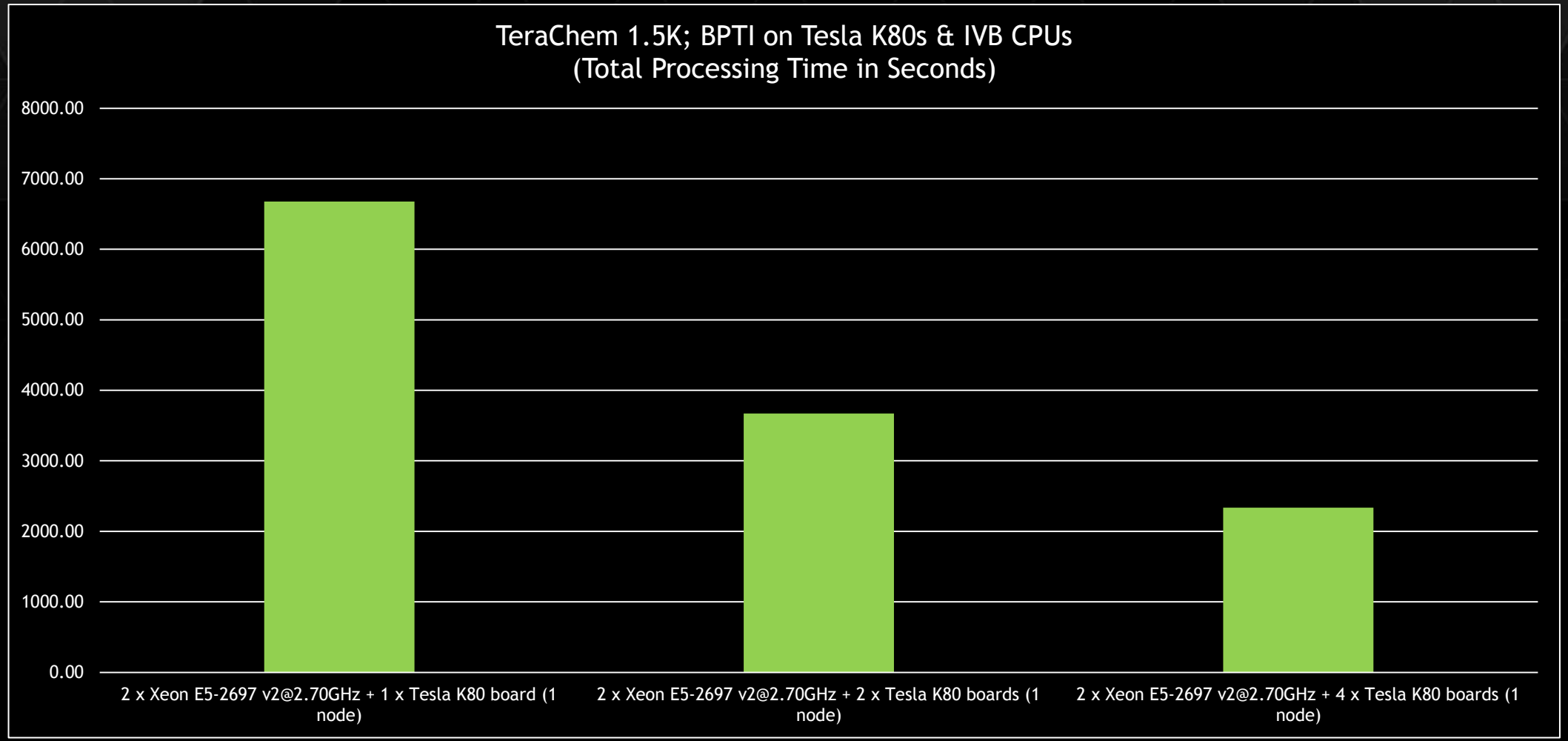
TERACHEM 1.5K; TRIPCAGE ON TESLA K80S & HASWELL CPUS



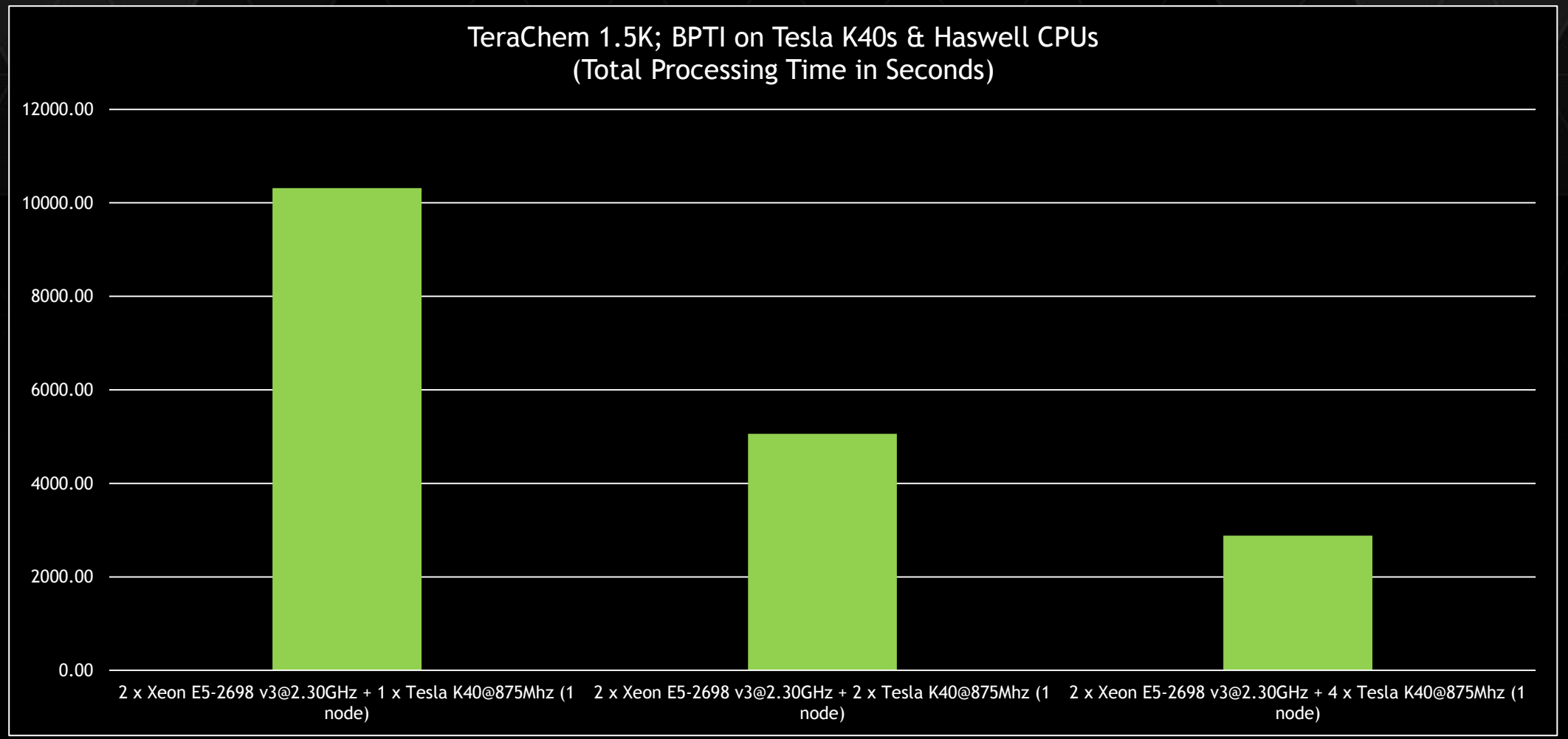
TERACHEM 1.5K; BPTI ON TESLA K40S & IVB CPUS



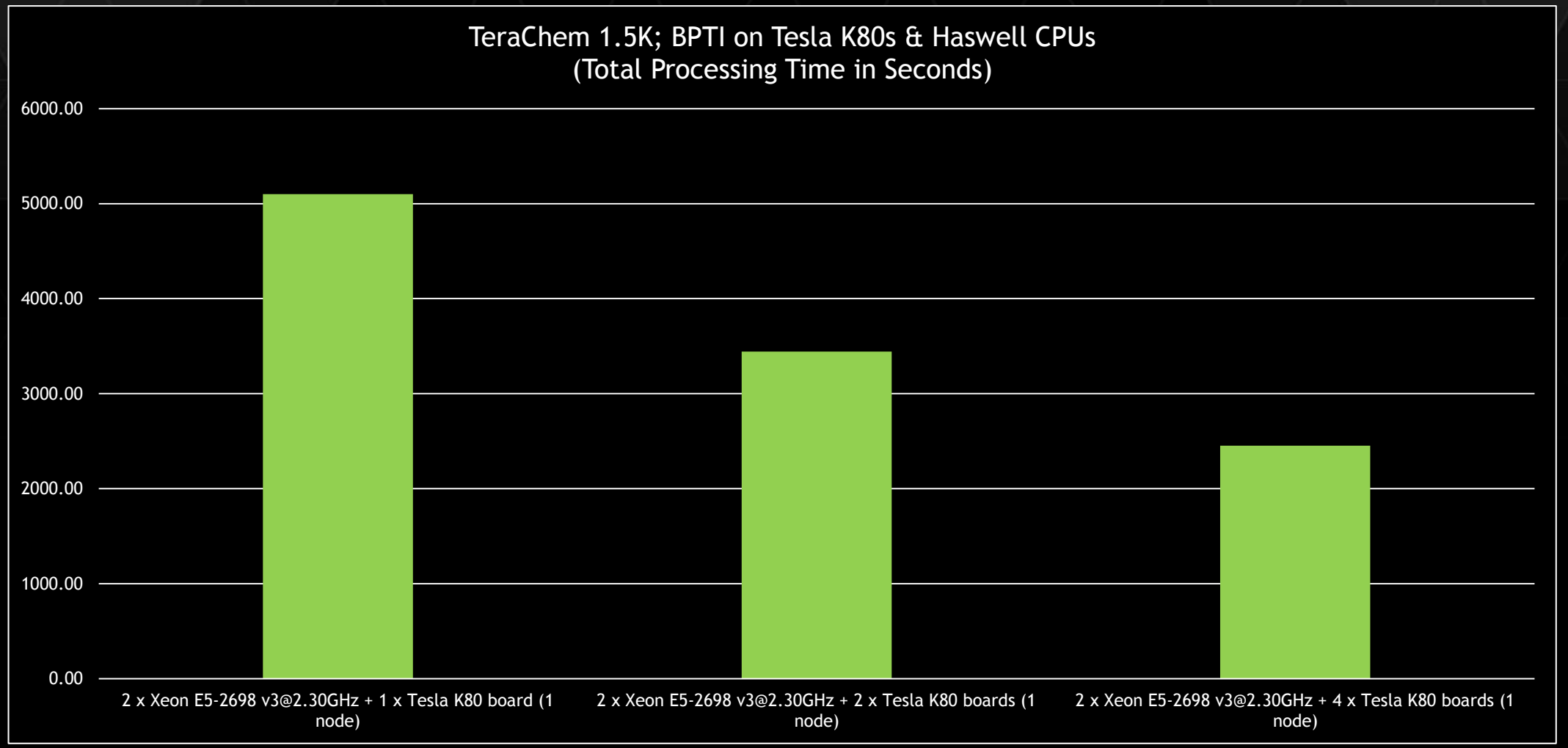
TERACHEM 1.5K; BPTI ON TESLA K80S & IVB CPUS



TERACHEM 1.5K; BPTI ON TESLA K40S & HASWELL CPUS



TERACHEM 1.5K; BPTI ON TESLA K80S & HASWELL CPUS



Benefits of GPU Accelerated Computing

- Faster than CPU only systems in all tests
- Large performance boost with marginal price increase
- Energy usage cut by more than half
- GPUs scale well within a node and over multiple nodes
- K20 GPU is our fastest and lowest power high performance GPU yet

Try GPU accelerated TeraChem for free – www.nvidia.com/GPUTestDrive

Test Drive K80 GPUs!

Experience The Acceleration

www.nvidia.com/GPUTestDrive

- Run Computational Chemistry Apps on Tesla K80 GPUs today
- Try Preconfigured Apps:
  AMBER, NAMD, GROMACS, LAMMPS, Quantum Espresso, TeraChem
  Or Load Your Own
- Sign up for FREE GPU Test Drive on remotely hosted clusters