**CUDA Center of Excellence at Illinois CUDA Achievement Award Submission**

# "Fighting HIV with CUDA Technology from the Desktop to the Petascale"

The first scientific breakthrough achieved with the Blue Waters supercomputer at the University of Illinois was the determination of the structure of the complete HIV capsid in atomic-level detail,[1] a collaborative effort of experimental groups, at the University of Pittsburgh and Vanderbilt University, and the NIH Center for Macromolecular Modeling and Bioinformatics, led by Prof. Klaus Schulten at the University of Illinois. The breakthrough was enabled by the NIH Center's popular and freely available programs NAMD and VMD, both of which incorporate CUDA technology to enable and accelerate the computationally intensive large-scale biomolecular modeling, simulation, and analysis required to perform the 64-million-atom HIV capsid simulation. The process through which the capsid disassembles, releasing its genetic material, is a critical step in HIV infection and a potential target for antiviral drugs. The work was featured on the cover of Nature [1] (Fig. 1) and recognized by an HPCwire Editors' Choice Award for "Best use of HPC in life sciences"[2] at SC13.

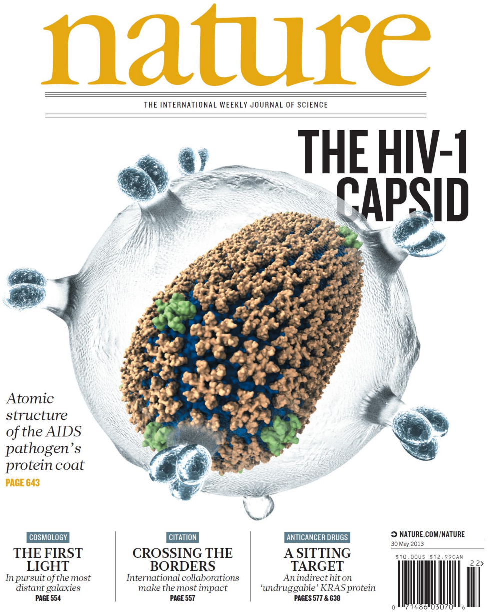
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE HIV-1 CAPSID
Atomic structure of the AIDS pathogen's protein coat
PAGE 643
COSMOLOGY
THE FIRST LIGHT
In pursuit of the most distant galaxies
PAGE 554
CITATION
CROSSING THE BORDERS
International collaborations make the most impact
PAGE 557
ANTICANCER DRUGS
A SITTING TARGET
An indirect hit on 'undruggable' KRAS protein
PAGES 577 & 638
NATURE.COM/NATURE
30 May 2013
$10.00US $12.99CAN
22

Figure 1: The first all-atom structure of the complete HIV virus capsid [1], a breakthrough enabled by CUDA technology in NAMD, VMD, and the Blue Waters supercomputer at Illinois.

## Scientific Achievement

Human immunodeficiency virus (HIV) is classified as a global pandemic by the World Health Organization. Treatments have been developed, but the virus adapts quickly to antiviral drugs such that new compounds need to be developed continuously. HIV infects the human cell by inserting its genes, packaged into a capsid, into the cell's nucleus, and hijacking the cell's machinery. As a virus that infects non-dividing cells, which have their genome protected behind a nuclear membrane, HIV has to take advantage of natural cell responses to induce cooperation of the cell to reach inside the nucleus. Some species of monkeys are immune to HIV by spoiling this cooperation with the capsid, attacking the capsid instead. Likewise, pharmacological interventions seek to attack the HIV capsid to prevent cooperation with the cellular machinery, or simply by breaking the capsid apart. Such interventions, however, require knowledge of the chemical and physical properties of the capsid.

The HIV capsid is large, containing about 1,300 proteins with altogether 4 million atoms. Although the capsid proteins are all identical, they nevertheless arrange themselves into a largely asymmetric structure. The large size and lack of symmetry pose a huge challenge to resolving the chemical structure of the HIV capsid. Recently, a combination of crystallography, nuclear magnetic resonance analysis, and electron microscopy finally resolved the capsid structure, but the key players were two computer programs, NAMD and VMD, that harnessed petascale computing through the NVIDIA GPU accelerated Cray Blue Waters. The structure, shown here, was published in May 2013 [1]. Embedding the structure into a computational model that also included physiological solvent led to a 64-million-atom simulation, the largest such simulation achieved to date. While the solution of the HIV capsid structure was a great achievement, it was just the starting point for the main research agenda, which is developing new antiviral drugs. Computational biologists presently simulate how the capsid exploits cellular proteins to be guided to the cell's nucleus and how antiviral drugs interfere with that process. The currently identified drug candidates, while effective against the capsid, are unfortunately also highly toxic to the patient. From the capsid simulation has emerged an incredibly precise picture of the interactions of the capsid with the cell and drugs, which can help guide future drug development. Such research, however, requires the most extreme computer power available today, which can come about only in partnership with NVIDIA GPUs.

[1] http://news.illinois.edu/news/13/0529HIVcapsid_KlausSchulten.html

[2] http://www.ncsa.illinois.edu/news/story/ncsa_receives_honors_in_2013_hpcwire_readers_and_editors_choice_awards

# Impact of CUDA Technology

The work leading to the successful HIV simulation on Blue Waters began seven years ago, with the first incorporation of CUDA-based GPU-acceleration into NAMD and VMD [2]. Research conducted from early 2007 until the present helped demonstrate to the molecular modeling and high-performance computing fields that GPU-accelerated supercomputers could help achieve new scientific breakthroughs. The CUDA Center demonstrated the benefits GPUs bring to both performance and energy efficiency of molecular modeling tasks [3] helping to motivate the incorporation of GPUs in the fastest supercomputers in the world, including Titan at Oak Ridge National Laboratory, and Blue Waters at the National Center for Supercomputing Applications at the University of Illinois. The effective use of GPUs for simulations such as HIV has required an intensive ongoing development effort involving contributions by many researchers at the CUDA Center, and by the engineers at NVIDIA.

**Molecular Dynamics Simulation** Preparations to use petascale machines such as Blue Waters began in 2006 with NAMD [4] being selected as a target application for the NSF “track 1” petascale computing environment solicitation that would eventually fund Blue Waters. The solicitation called for a benchmark simulation of 100 million atoms, an immense system given that the Center had only that same year published its first million-atom simulation [5], itself a collaborative project with experimentalists that characterized in full atomic detail the structure of the tiny satellite tobacco mosaic virus (STMV).

2006 also marked the NIH Center’s first exposure to CUDA technology at an SC2006 workshop, with an initial skepticism alleviated by the first CUDA programming course at Illinois in the spring of 2007 by now CUDA Center of Excellence director Wen-Mei Hwu and then NVIDIA Chief Scientist David Kirk. The NIH Center then quickly published in 2007 the first paper demonstrating the promise of CUDA technology for a variety of molecular modeling tasks [2], including electrostatic calculations on the million-atom STMV simulation, and presented CUDA tutorial lectures at SC07, ISC08, SC08, IEEE Cluster 2009, and SC09. Adaptation of NAMD to run on GPU clusters was reported at SC08 [6].

Recent NAMD development has focused on 100-million-atom simulations, driven by the NSF as a acceptance test for Blue Waters, and currently enabling a variety of simulations on GPU-accelerated supercomputers such as ORNL Titan (Fig. 2). NAMD relies on a direct interface between the Charm++[3] messaging layer underlying NAMD and the Cray Gemini network hardware, as reported at SC12 [7], demonstrating scaling to all 300,000 cores of the Titan machine (before GPUs were available on all nodes). Scalability has been improved by minimizing communication between nodes and further by aligning the NAMD parallel decomposition to the 3-dimensional torus topology of the Cray network.

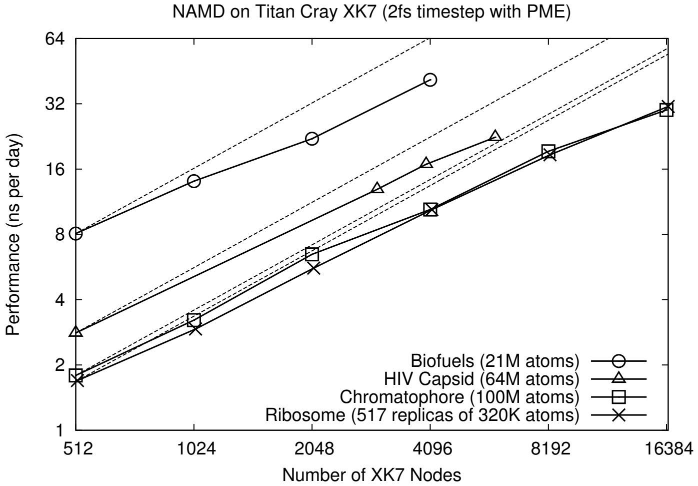

Figure 2: NAMD strong scaling with Tesla K20X GPUs on ORNL Titan. Parallel efficiencies on 4096 XK7 nodes vs. 512 nodes are 64% for biofuels, 77% for HIV, 73% for chromatophore, and 77% for ribosome.

The NVIDIA GPUs on Titan and Blue Waters increase NAMD performance but also proportionately increase the degree to which inter-node communication limits parallel scalability. This was addressed by doubling the interpolation order used by the particle-mesh Ewald (PME) long-range electrostatics method, allowing an 8-fold reduction in parallel communication at the expense of an 8-fold increase in the per-atom cost of the calculation; this increase is now being addressed by the development of a new GPU-accelerated PME kernel, targeting the latest NVIDIA Kepler GPUs.

[3] http://charm.cs.illinois.edu/

| Movie Res. | VMD Renderer Mode | Node | Nodes | Execution Time | | |
|---|---|---|---|---|---|---|
| | | | | Scripts & I/O | Geometry and Rendering | Total |
| HD<br>1920 × 1080<br>2M pixels | TachyonL-OptiX<br>GPU ray tracing | XK7 | 64 | 40 s | 655 s | 695 s |
| | | XK7 | 256 | 117 s | 171 s | 288 s |
| | Tachyon<br>CPU ray tracing | XE6 | 256 | 167 s | 1,374 s | 1,541 s |
| | | XE6 | 512 | 224 s | 808 s | 1,032 s |

Table 1: VMD Parallel Movie Rendering Performance: 1,079 Frame HIV-1 Movie [9].

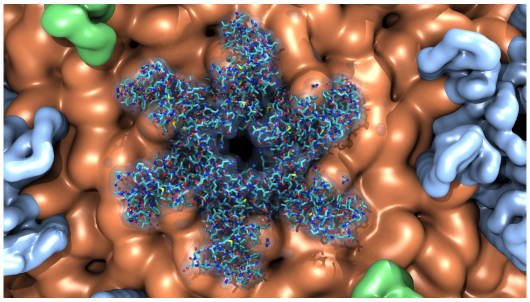

Figure 3: VMD close-up rendering of the 64M-atom atomic structure of the HIV-1 capsid [1] taken from the HIV movie performance test [8, 9].

**Simulation Preparation, Visualization, and Analysis** VMD [10] was one of the first molecular visualization tools to employ GPUs for functions beyond OpenGL rasterization [2]. Led by VMD's broad use of GPUs [2, 11, 12, 13], the molecular modeling field has subsequently seen broad adoption of GPUs for many computationally demanding algorithms. VMD has been adapted to leverage the computational capabilities of GPUs for acceleration of data-parallel algorithms, initially focusing on the computationally most demanding non-graphical algorithms such as calculation of electrostatic potential maps [2, 14, 15] and molecular dynamics simulation trajectory analysis routines [3, 16]. Beyond their use for OpenGL rasterization, VMD makes extensive use of GPUs to enable interactive computational visualizations that would not be possible using conventional multi-core CPUs, including calculation and display of molecular orbitals [17], molecular, cellular, and neuronal surfaces [12, 13, 9, 18], and 3-D cross correlation maps for atomic structures solved by molecular dynamics flexible fitting and similar hybrid structure determination techniques [19]. VMD is currently being ported to power- and performance-constrained mobile devices such as Android smartphones and tablets, and it is expected that recent work evaluating VMD performance and energy efficiency on CUDA-accelerated Tegra 3 development platforms will ultimately benefit all platforms [20].

VMD has recently been adapted to GPU-accelerated petascale supercomputers using a hybrid computational approach that combines MPI, multithreading, and CUDA GPU acceleration. The incorporation of GPUs into petascale systems has created a new opportunity to apply GPUs to large scale simulation analysis tasks, and to the creation of advanced visualizations and movie renderings [3, 8, 9]. VMD incorporates a new CUDA- and OptiX-accelerated "lightweight" implementation of the Tachyon parallel ray tracing engine for cases where high fidelity visualizations are required. The tight internal integration of CUDA, OptiX, and Tachyon into VMD allows rendering of shadows, ambient occlusion lighting, high-quality transparency, and depth-of-field effects that are poorly suited to the OpenGL shading pipeline, and it outperforms even the fastest external photorealistic rendering tools by virtue of direct access to the in-memory VMD molecular scene, thereby eliminating I/O that would otherwise cripple performance (Table 1, Figs. 1, 3). The ability to perform large scale rendering jobs directly on petascale supercomputers allows researchers to avoid the transfer of tens of terabytes of simulation trajectories to other sites for rendering. The parallel analysis and movie rendering infrastructure in VMD allows rapid completion of I/O, analytical computations, computational visualization, and photorealistic movie rendering tasks that could take weeks or months (even with the acceleration provided by GPUs) on a desktop workstation [8, 9, 19].

# References Cited

[1] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497:643–646, 2013.

[2] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, and K. Schulten. Accelerating molecular modeling applications with graphics processors. *J. Comp. Chem.*, 28:2618–2640, 2007.

[3] J. Enos, C. Steffen, J. Fullop, M. Showerman, G. Shi, K. Esler, V. Kindratenko, J. E. Stone, and J. C. Phillips. Quantifying the impact of GPUs on performance and energy efficiency in HPC clusters. In *International Conference on Green Computing*, pages 317–324, 2010.

[4] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comp. Chem.*, 26:1781–1802, 2005.

[5] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14:437–449, 2006.

[6] J. C. Phillips, J. E. Stone, and K. Schulten. Adapting a message-driven parallel application to GPU-accelerated clusters. In *SC '08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*. IEEE Press, Piscataway, NJ, USA, 2008.

[7] Y. Sun, G. Zheng, C. Mei, E. J. Bohm, T. Jones, L. V. Kalé, and J. C. Phillips. Optimizing fine-grained communication in a biomolecular simulation application on Cray XK6. In *Proceedings of the 2012 ACM/IEEE conference on Supercomputing*. IEEE press, Salt Lake City, Utah, 2012.

[8] J. E. Stone, B. Isralewitz, and K. Schulten. Early experiences scaling VMD molecular visualization and analysis jobs on Blue Waters. In *Proceedings of the XSEDE Extreme Scaling Workshop*, 2013.

[9] J. E. Stone, K. L. Vandivort, and K. Schulten. GPU-accelerated molecular visualization on petascale supercomputing platforms. In *Proceedings of the 8th International Workshop on Ultrascale Visualization*, UltraVis '13, pages 6:1–6:8. ACM, New York, NY, USA, 2013. URL `http://doi.acm.org/10.1145/2535571.2535595`.

[10] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

[11] J. E. Stone, D. J. Hardy, I. S. Ufimtsev, and K. Schulten. GPU-accelerated molecular modeling coming of age. *J. Mol. Graph. Model.*, 29:116–125, 2010.

[12] M. Krone, J. E. Stone, T. Ertl, and K. Schulten. Fast visualization of Gaussian density surfaces for molecular dynamics and particle system trajectories. In *EuroVis - Short Papers 2012*, pages 67–71, 2012.

[13] E. Roberts, J. E. Stone, and Z. Luthey-Schulten. Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. *J. Comp. Chem.*, 34:245–255, 2013.

[14] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips. GPU computing. *Proc. IEEE*, 96:879–899, 2008.

[15] D. J. Hardy, J. E. Stone, and K. Schulten. Multilevel summation of electrostatic potentials using graphics processing units. *J. Paral. Comp.*, 35:164–177, 2009.

[16] B. G. Levine, J. E. Stone, and A. Kohlmeyer. Fast analysis of molecular dynamics trajectories with graphics processing units–radial distribution function histogramming. *J. Comp. Phys.*, 230:3556–3569, 2011.

[17] J. E. Stone, J. Saam, D. J. Hardy, K. L. Vandivort, W. W. Hwu, and K. Schulten. High performance computation and interactive display of molecular orbitals on GPUs and multi-core CPUs. In *Proceedings of the 2nd Workshop on General-Purpose Processing on Graphics Processing Units, ACM International Conference Proceeding Series*, volume 383, pages 9–18. ACM, New York, NY, USA, 2009.

[18] E. Cai, P. Ge, S. H. Lee, O. Jeyifous, Y. Wang, Y. Liu, K. M. Wilson, S. J. Lim, M. A. Baird, J. E. Stone, K. Y. Lee, D. G. Fernig, M. W. Davidson, H. J. Chung, K. Schulten, A. M. Smith, W. N. Green, and P. R. Selvin. Stable small quantum dots for synaptic receptor tracking on live neurons, 2014. (Submitted).

[19] J. E. Stone, R. McGreevy, B. Isralewitz, and K. Schulten. GPU-accelerated analysis and visualization of large structures solved by molecular dynamics flexible fitting, 2014. (Submitted).

[20] J. E. Stone, M. J. Hallock, J. C. Phillips, J. R. Peterson, K. L. Vandivort, Z. Luthey-Schulten, and K. Schulten. Evaluation of emerging energy-efficient heterogeneous computing platforms for biomolecular and cellular simulation workloads, 2014. (Submitted).