DESRES/TR--2015-01

Desmond/GPU Performance as of October 2015

Michael Bergdorf,[1] Sean Baxter,[1] Charles A. Rendleman,[1] and David E. Shaw[1,2,*]

Desmond/GPU Software Version 1.6.2, October 17, 2015

Summary

Desmond/GPU is a code, written in CUDA C++, that is designed for the execution of molecular dynamics (MD) simulations of biological systems on NVIDIA Graphics Processing Units (GPUs). This paper reports the performance that Desmond/GPU, running on a range of GPU models and configurations, achieves on four biological system benchmarks as of October 2015.

Benchmark Chemical Systems and Simulation Parameters

Performance results were measured for two chemical systems that are commonly used for MD code benchmarking, and two additional systems that are characteristic for free energy perturbation (FEP) simulations. The characteristics of the benchmark systems are summarized in

[1] D. E. Shaw Research, New York, NY 10036, USA.
[2] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA.
[*] To whom correspondence should be addressed: David.Shaw@DEShawResearch.com.

Table 1. The parameters were chosen to optimize Desmond/GPU performance without compromising accuracy [1].

We report simulation rates for two different algorithms for the decomposition of the electrostatic energy into a "near" and a "far" part: the first algorithm is the well-known Particle Mesh Ewald scheme (PME). The second algorithm is "*u*-series" decomposition, a new approximation for electrostatic interactions developed at D. E. Shaw Research [2].

DHFR [3] and ApoA1 [4] were run with a non-bonded interaction cutoff of 9 Å in the NVE ensemble, that is, without temperature or pressure control. In addition to the benchmark time step size of 2.5 fs, we also report the performance of these two systems run using hydrogen mass repartitioning (HMR) [5]. HMR involves increasing the mass of all hydrogen atoms from 1 u to 4 u, while keeping the total mass of water molecules unchanged; in other words, oxygen atoms in water molecules are assigned a mass of 10 u in a HMR system. The repartitioning reduces the frequency of the fastest degrees of freedom and allows us to integrate the system with a time step of 4.0 fs.

Free energy perturbation/replica exchange with solute tempering (FEP/REST) [6] combines enhanced sampling through replica exchange with FEP to accelerate structural reorganization and thus convergence of relative protein-ligand binding affinities. Our two FEP/REST benchmarks, P38C and P38S [7], were run with a non-bonded interaction cutoff of 9 Å in the NVT ensemble using a Nosé-Hoover thermostat. Both consist of 12 replicas (or "windows") each. Every 1.2 simulated picoseconds, replicas are exchanged and energy differences (dE) are computed and written out. Both the p38 complex (P38C) and the p38 inhibitor (P38S) system were run on different numbers of GPUs.

Table 1. Production parameters. The Far ES frequency is the interval at which the "far" electrostatics forces are evaluated.

System	# of atoms	System size (Å^3)	Time Step (fs)	Far ES Frequency	ES algorithm	Grid size
ApoA1	92,224	109 × 109 × 78	2.5	2	PME	128^3
					u-series	64^3
			4.0 (HMR)	2	*u*-series	64^3
DHFR	23,588	62 × 62 × 62	2.5	2	PME	64^3
					u-series	32^3
			4.0 (HMR)	2	*u*-series	32^3
P38C	25,550	80 × 61 × 56	2.0	3	PME	64^3
					u-series	32^3
P38S	2,853	29 × 29 × 36	2.0	3	PME	32^3
					u-series	16^3

Hardware and Operating Environment

We ran the benchmarks on GPUs representing three different architectures: "Kepler" GK104, "Kepler II" GK110/GK110B, and "Maxwell" GM200/GM204/GM206.

The DHFR and ApoA1 benchmarks were run on a variety of host systems: single-processor Dell workstations T3600 (TITAN, GeForce GTX 960, 980); a Super Micro 4027GR 8-GPU server (Tesla K20c, Tesla K40c, GeForce GTX 680, 780 Ti, 980 Ti, TITAN X); and a dual-socket Cirrascale BladeRack-XL 5GVU 8-GPU server with PCIe-Gen2 (GTX 780).

The FEP/REST benchmarks were run on: a dual-socket Super Micro 4 × GTX TITAN X server; and a dual-socket Cirrascale BladeRack-XL 5GVU 8-GPU server with PCIe-Gen2 (GTX 780).

CPU architectures ranged from Sandy Bridge to Haswell and affected simulation performance very weakly, since in Desmond/GPU the CPU is used only to dispatch work to the GPUs. All host systems ran under CentOS 6.7. Desmond/GPU was compiled using CUDA 7.0.28 and

GCC 4.7.2. All simulations used NVIDIA driver version 346.47 or 346.96. Note that performance may vary depending on the driver version. In general, we chose the oldest driver that had all necessary bug fixes and supported our CUDA version. Note also that in our experience GeForce cards require an extensive testing, selection, and burn-in period before a stable and reliable set is deemed suitable for use in a production setting.

Results

In this section, Desmond/GPU simulation rates are reported in units of simulated nanoseconds per day. Note that the version of Desmond/GPU used in this report does not support the use of more than one GPU in a single MD simulation, though replica exchange simulations can use more than one GPU. This change in the software allowed for large improvements in performance and increased simulation system sizes. For a comparison of Desmond performance on CPUs versus GPUs, and for comparative performance of the current version of Desmond/GPU with the previous version, we refer the reader to the preceding performance study [8].

Table 3 and Figure 1 report the performance of Desmond/GPU for DHFR and ApoA1 for different GPUs. Table 4 and Figure 2 report the performance for the two representative FEP/REST systems described above.

Table 2. DHFR/ApoA1 benchmarks. Simulation rates are given in simulated nanoseconds per day for the three Desmond configurations tested.

GPU	Configuration	DHFR Rate (ns/day)	ApoA1 Rate (ns/day)
GeForce GTX 680	PME	158.9	30.1
	u-series	187.0	35.9
	u-series HMR	276.0	54.8
GeForce GTX 780	PME	254.8	59.5
	u-series	289.6	70.3
	u-series HMR	429.2	106.5
GeForce GTX 780 Ti	PME	281.5	65.4
	u-series	322.4	77.6
	u-series HMR	478.3	117.4
GeForce GTX 960	PME	155.9	27.1
	u-series	195.7	36.1
	u-series HMR	290.8	54.8
GeForce GTX 980	PME	289.3	56.6
	u-series	341.3	73.2
	u-series HMR	501.4	111.5
GeForce GTX 980 Ti	PME	342.5	80.4
	u-series	382.4	100.6
	u-series HMR	567.0	152.5
GeForce GTX TITAN	PME	263.9	60.6
	u-series	301.3	73.3
	u-series HMR	448.2	110.9
GeForce GTX TITAN X	PME	339.5	78.7
	u-series	383.3	97.6
	u-series HMR	567.7	147.9
Tesla K20c	PME	182.5	39.4
	u-series	212.6	49.4
	u-series HMR	316.7	74.8
Tesla K40c	PME	219.3	51.1
	u-series	250.7	61.6
	u-series HMR	369.2	93.0

Figure 1. Desmond 3.6 and Desmond/GPU DHFR/ApoA1 benchmarks. A: the performance of Desmond/GPU running the ApoA1 (dark green) and the DHFR (light green) benchmark with the two different Far ES algorithms (see Table 2). The darkened extensions indicate improvement in performance that can be obtained by running the code using *u*-series. B: Desmond/GPU performance using HMR for the same benchmarks for different GPU models in order of increasing DHFR performance (see Table 2 for details).

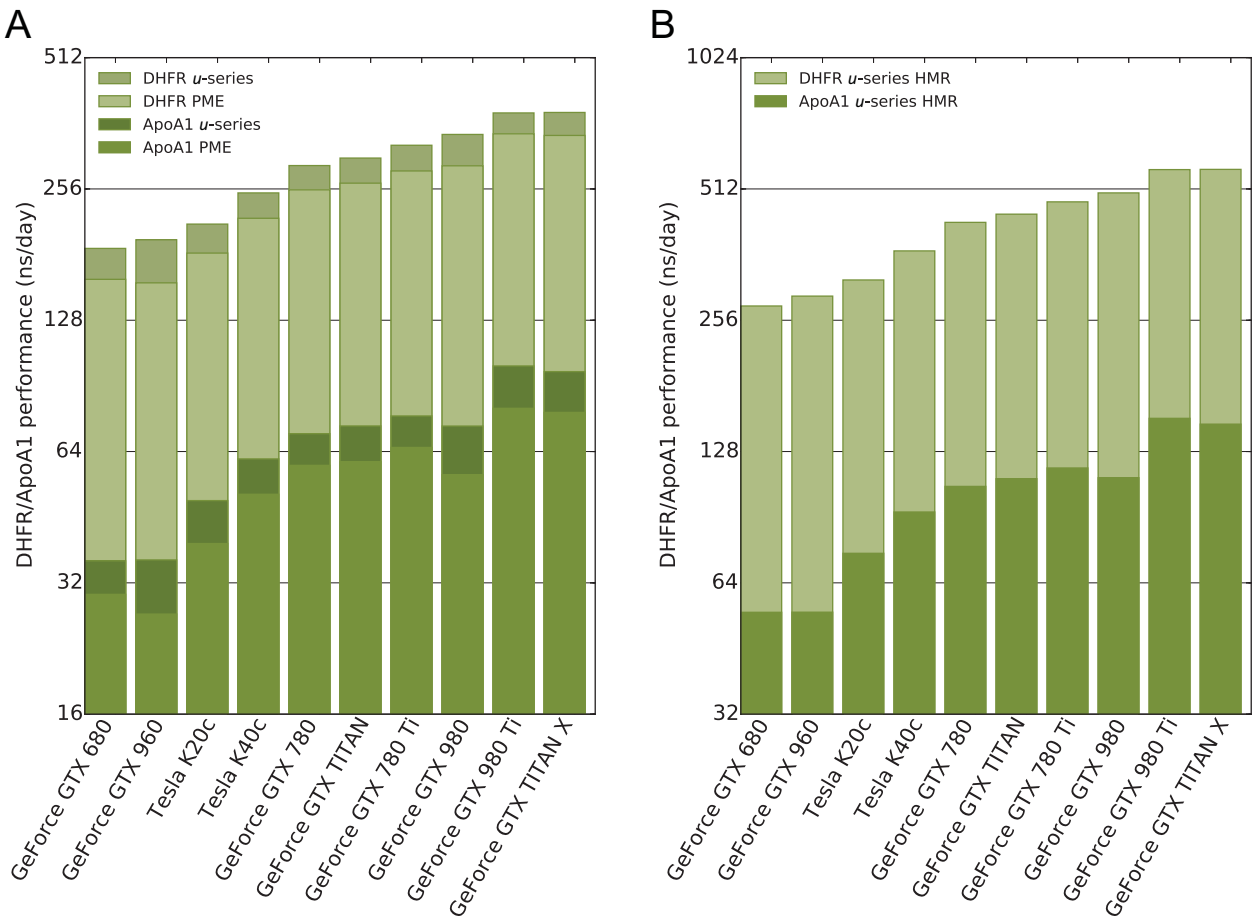
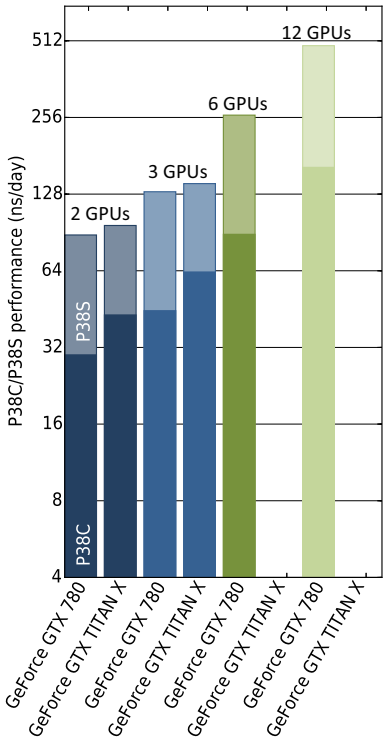


Table 3. The "# of GPUs" column gives the number of GPUs used to simulate the 12 replicas. Simulation rates are given in simulated nanoseconds per day.

GPU	Far ES algorithm	# of GPUs	P38C Rate (ns/day)	P38S Rate (ns/day)
GeForce GTX 780	PME	2	24.8	81.9
		3	36.7	122.5
		6	73.0	246.1
		12	136.2	460.4
	u-series	2	29.9	88.5
		3	44.6	130.8
		6	89.0	261.4
		12	163.0	489.9
GeForce GTX TITAN X	PME	2	35.4	91.2
		3	53.0	134.2
	u-series	2	42.9	96.4
		3	63.3	140.8

Figure 2. FEP/REST benchmark systems. The dark bars give the performance for the complex, P38C, while the light bars show the performance for the smaller p38 inhibitor FEP/REST system, P38S (see Table 3 for details).



Desmond/GPU Availability

Desmond/GPU is available without cost from D. E. Shaw Research[3] for non-commercial research use by non-profit institutions, and under commercial license from Schrödinger, LLC[4] for other purposes or parties.

The current release is based on CUDA 7.0 and supports NVIDIA GPUs with compute capabilities 3.0, 3.5, 5.0, and 5.2.

[3] http://www.deshawresearch.com/resources_desmond.html
[4] http://www.schrodinger.com/Desmond

References

1. Edmond Chow, Charles A. Rendleman, Kevin J. Bowers, Ron O. Dror, Douglas H. Hughes, Justin Gullingsrud, Federico D. Sacerdoti, and David E. Shaw, "Desmond Performance on a Cluster of Multicore Processors," D. E. Shaw Research Technical Report DESRES/TR--2008-01, July 2008.

2. Cristian Predescu, Ross A. Lippert, Adam K. Lerer, Brian Towles, J. P. Grossman, Robert M. Dirks, and David E. Shaw, "The *u*-series: A separable and accurate decomposition for electrostatics computation," *In preparation*.

3. MD Benchmarks for Amber, CHARMM and NAMD. http://ambermd.org/amber8.bench2.html

4. James C. Phillips, Gengbin Zheng, Sameer Kumar, and Laxmikant V. Kalé, "NAMD: Biomolecular Simulation on Thousands of Processors," *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing (SC02)*, Baltimore, Maryland, November 16–22, 2002.

5. K. Anton Feenstra, Berk Hess, and Herman J. C. Berendsen, "Improving efficiency of large timescale molecular dynamics simulations of hydrogen-rich systems," *Journal of Computational Chemistry*, vol. 20, 1999, pp. 786–798.

6. Lingle Wang, B. J. Berne, and Richard A. Friesner, "On Achieving High accuracy and Reliability in the Calculation of Relative Protein–Ligand Binding Affinities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 6, 2012, pp. 1937–1942.

7. David M. Goldstein, Michael Soth, Tobias Gabriel, Nolan Dewdney, Andreas Kuglstatter, Humberto Arzeno, Jeffrey Chen, William Bingenheimer, Stacie A. Dalrymple, James Dunn, Robert Farrell, Sandra Frauchiger, JoAnn La Fargue, Manjiri Ghate, Bradford Graves, Ronald J. Hill, Fujun Li, Renee Litman, Brad Loe, Joel McIntosh, Daniel McWeeney, Eva Papp, Jaehyeon Park, Harlan F. Reese, Richard T. Roberts, David Rotstein, Bong San Pablo, Keshab Sarma, Martin Stahl, Man-Ling Sung, Rebecca T. Suttman, Eric B. Sjogren, Yunchou Tan, Alejandra Trejo, Mary Welch, Paul Weller, Brian R. Wong, and Hasim Zecic, "Discovery of 6-(2,4-Difluorophenoxy)-2-[3-hydroxy-1-(2-hydroxyethyl)propylamino]-8-

methyl-8H-pyrido[2,3-d]pyrimidin-7-one (Pamapimod) and 6-(2,4-Difluorophenoxy)-8-methyl-2-(tetrahydro-2H-pyran-4-ylamino)pyrido[2,3-d]pyrimidin-7(8H)-one (R1487) as Orally Bioavailable and Highly Selective Inhibitors of p38α Mitogen-Activated Protein Kinase," *Journal of Medicinal Chemistry*, vol. 54, no. 7, 2011, pp. 2255–2265.

8. Michael Bergdorf, Eric T. Kim, Charles A. Rendleman, and David E. Shaw, "Desmond/GPU Performance as of November 2014" D. E. Shaw Research Technical Report DESRES/TR—2014-01, November 2014.

Appendix

The configuration and structure files used in the simulations reported in this document are available on our website.

Desmond/GPU memory requirements

The memory requirements of Desmond/GPU for intermediately sized systems can be estimated as

$$235 \text{ MB} + 2.7 \times \text{number of particles in thousands},$$

thus, a system with a million atoms will roughly require a GPU with at least 3 GB of RAM.