

More Bang for Your Buck: Improved use of GPU Nodes for GROMACS 2018

Carsten Kutzner^{id},^{*[a]} Szilárd Páll,^[b] Martin Fechner,^[a] Ansgar Esztermann,^[a]
Bert L. de Groot,^[a] and Helmut Grubmüller^[a]

We identify hardware that is optimal to produce molecular dynamics (MD) trajectories on Linux compute clusters with the GROMACS 2018 simulation package. Therefore, we benchmark the GROMACS performance on a diverse set of compute nodes and relate it to the costs of the nodes, which may include their lifetime costs for energy and cooling. In agreement with our earlier investigation using GROMACS 4.6 on hardware of 2014, the performance to price ratio of consumer GPU nodes is considerably higher than that of CPU nodes. However, with GROMACS 2018, the optimal CPU to

GPU processing power balance has shifted even more toward the GPU. Hence, nodes optimized for GROMACS 2018 and later versions enable a significantly higher performance to price ratio than nodes optimized for older GROMACS versions. Moreover, the shift toward GPU processing allows to cheaply upgrade old nodes with recent GPUs, yielding essentially the same performance as comparable brand-new hardware. © 2019 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.26011

Introduction

Molecular dynamics (MD) simulation is a well-established computational tool to investigate and understand biomolecular function in atomic detail from a physics perspective. A simulation system of a solvated molecule can comprise thousands to millions of atoms, depending on whether it is a small protein or a large complex like a ribosome^[1] or a viral shell.^[2] To derive the time evolution of atomic movements on biologically relevant time scales, millions of time steps need to be computed. For statistically significant results, this process is usually repeated many times with varying starting conditions. Consequently, the investigation of a single biomolecular system can easily occupy a number of modern compute nodes for weeks, whereas all simulation projects of a typical research group performing MD calculations will keep a medium-sized compute cluster running nonstop.

Whether the necessary cluster hardware is purchased by the department that uses it or the services of a high performance computing (HPC) center are used, eventually someone has to decide on what to buy. This decision is not straightforward as the available hardware is quite diverse. What node specifications should be used? Should they rather have many weak compute cores or fewer strong ones? Are multisolet nodes better than single-socket nodes? How many CPU cores are needed per GPU and what GPU type is optimal? What about memory and interconnect?

All-rounder cluster nodes designed for many diverse software applications usually contain top-end CPUs, GPUs with high double precision floating point performance, lots of memory, and an expensive interconnect. The result of meeting all these needs at once is a very low ratio of computation performance to node price for each individual application. Our approach is completely opposite: maximized cost-efficiency by specialization. We focus on a particular application, namely MD, and look for hardware that

yields the highest simulation throughput for a fixed budget, measured in total length of produced trajectory over its lifetime.

The set of available MD codes for biomolecular simulations is diverse and includes, among others, ACEMD,^[3] Amber,^[4] CHARMM,^[5] Desmond,^[6] LAMMPS,^[7] NAMD,^[8] OpenMM,^[9] and GROMACS.^[10] We use GROMACS, because it is one of the fastest MD engines available, widely-used, and freely available.

Our basic question is: Given a fixed budget, how can we produce as much MD trajectory as possible? Accordingly, we measure simulation performances for representative biomolecular MD systems and determine the corresponding total hardware price. We do not aim at a comprehensive evaluation of currently available hardware, we merely aim at uncovering hardware that has an exceptional performance to price (P/P) ratio, which is the efficiency metric used in this study, for version 2018 of the GROMACS MD code.

As our study prioritizes the efficiency and total throughput of generating trajectories (assuming plenty of concurrent simulations), we do not consider use-cases where generating an individual trajectory as fast as possible is preferred. Whereas the latter can be important, for example, in exploratory studies, faster simulations require strong scaling, which always comes at a cost due to the inherent trade-off between simulation rate and

[a] C. Kutzner, M. Fechner, A. Esztermann, B. L. de Groot, H. Grubmüller
*Theoretical and Computational Biophysics, Max Planck Institute for
Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany*
E-mail: ckutzne@gwdg.de

[b] S. Páll
*Center for High Performance Computing, KTH Royal Institute of Technology,
10044, Stockholm, Sweden*

Contract Grant sponsor: Deutsche Forschungsgemeinschaft; Contract
Grant number: SPP 1648; Contract Grant sponsor: European Union;
Contract Grant number: H2020-EINFRA-2015-1-675728

© 2019 Wiley Periodicals, Inc.

parallel efficiency. At the same time, running a large number of independent or weakly coupled simulations is a widely used and increasingly important use case. Thanks to continuous advances in simulation performance, many well-optimized MD engines like that of GROMACS have useful trajectory generation rates even without large-scale parallelization. Additionally, problems that previously required the generation of a few long trajectories are nowadays often addressed by ensemble methods that instead need a large number of shorter trajectories.^[11–15] Hence, large compute resources with a single fast interconnect are often not a must. Instead, the ideal hardware in these instances is one with fast “islands” and a more modest interconnect between these. In fact, on such machines, results can in some cases be obtained as fast or faster, and crucially more cost-effectively. In our case, considering current hardware limitations and software characteristics, for ultimate efficiency reasons, the fast “islands” are represented by a set of CPU cores and a single GPU.

In addition to the P/P ratio, when evaluating systems, we take into account the two following criteria: (1) energy consumption, as it is one of the largest contributors to trajectory production costs and (2) rack space, which is limited in any server room. We cover energy consumption in a separate section of the paper, whereas space requirements are implicitly taken into account by our hardware preselection. We require an average packing density of at least one GPU per height unit U of rack space for server nodes to be considered: a 4 U server with 4 GPUs meets the criterion.

In an earlier investigation using GROMACS 4.6 with 2014 hardware,^[16] we showed that the simulation throughput of an optimized cluster is typically two to three times larger than that of a conventional cluster. Since 2014, hardware has been evolving continuously and fundamental algorithmic enhancements have been made. Therefore, using the two exact same MD test systems, we provide this update to our original investigation and point the reader to the current hardware yielding the *best bang for your buck* with GROMACS 2018.^[10,17]

We focus on hardware evaluation and not on how GROMACS performance can be optimized, as that has already been extensively discussed.^[16] Most of the performance advice given in our original paper are still valid if not stated otherwise. Particular remarks for a specific GROMACS version are available in the online user guide at <http://manual.gromacs.org/> in the section called *Getting good performance from mdrun*.

GROMACS load distribution schemes

GROMACS uses various mechanisms to parallelize compute work over available resources so that it can optimally benefit from the

hardware’s capabilities.^[10,18] Processes sharing the computation of an MD system (these processes are called ranks) communicate via the message passing interface (MPI) library, while each rank can consist of several OpenMP threads. Each rank can optionally have its short-range part of the Coulomb and van der Waals interactions (the *pair interactions*) calculated on a GPU, if present; this process is called offloading and is illustrated in Figure 5. The long-range part of Coulomb interactions is computed with the particle mesh Ewald (PME) method,^[19] which can be performed on a subset of the available ranks for improved parallel efficiency. Alternatively, from version 2018 on, PME can also be offloaded to a GPU. On the lowest level of the parallelization hierarchy, SIMD (single instruction multiple data) parallelism is exploited in almost all performance-sensitive code parts.

Summary of the 2014 investigation

To illustrate the advancements, both on the implementation side as well as on the hardware side, over the past 5 years, we summarize the main points of our original investigation in the following.

In our original investigation,^[16] we determined hardware prices and GROMACS 4.6 performances for over 50 different node configurations built from 12 CPU and 13 GPU models. In particular, we compared consumer GPUs with professional GPUs. Professional GPUs like NVIDIA’s Tesla and Quadro models are typically used for computer-aided design, computer-generated imagery, and in HPC. Consumer GPUs, like the GeForce series, are mainly used for gaming. They are much cheaper (up to about 1000 € net compared to several thousand Euro for the professional cards) and lack some of the features the professional models offer, for example, regarding memory and double precision floating-point performance.

Our two main benchmark systems (that we continue to use in the present study) were an 80 k atom membrane protein embedded in a lipid bilayer surrounded by water and ions (MEM) and a 2 M atom ribosome system (RIB) solvated in water and ions, see Table 1 for specification details. On each hardware configuration, we determined the fastest settings for running a single simulation of each system by performing a scan of the parameters controlling the parallelization settings, that is, the number of MPI ranks, the number of OpenMP threads, and the number of separate PME ranks.

We concluded from our investigation (Fig. 1) that single CPU-only nodes and nodes with professional GPUs have a comparably low P/P ratio, whereas consumer GPUs improve the P/P ratio by a factor of 2–3. Adding the first consumer GPU to a node yields the largest increase of the P/P ratio, whereas adding too much GPU power can also lower the P/P ratio (for instance, compare the 2x

Table 1. Specifications of the MD benchmarks. Our principal benchmarks are done with atomistic MEM and RIB systems, whereas for comparison, we also benchmark two coarse grain systems (VES and BIG 0) using the Martini force field.

MD system	MEM ^[20]	RIB ^[1]	VES ^[21]	BIG
Number of particles	81,743	2,136,412	72,076	2,094,812
System size (nm)	10.8 × 10.2 × 9.6	31.2 ³	22.2 × 20.9 × 18.4	142.4 × 142.4 × 11.3
Time step (fs)	2	4	30	20
Cutoff radii (nm)	1.0	1.0	1.1	1.1
PME mesh spacing (nm)	0.12	0.135	–	–

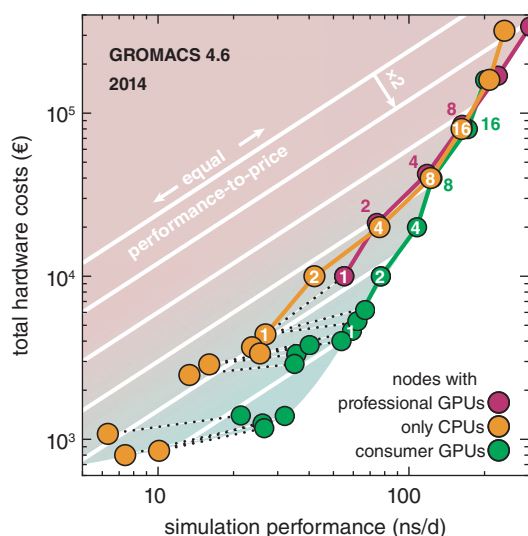


Figure 1. Summary of our original investigation testing GROMACS 4.6 on nodes built from hardware available in 2014.^[16] Hardware costs versus MEM benchmark performance (circles) for three classes of node types: CPU-only nodes (orange), nodes with professional Tesla GPUs (purple), and nodes with consumer GeForce GPUs (green). Dotted lines connect GPU nodes with their CPU counterparts. Circles connected by colored lines denote a cluster built from that node type, numbers in the circles denote how many of these nodes took part in the benchmark. The white lines are isolines of equal P/P ratio. Moving down from one isoline to the next increases the P/P ratio by a factor of two (red shaded area = low P/P ratio, green shaded area = high P/P ratio). [Color figure can be viewed at wileyonlinelibrary.com]

E5-2680v2 nodes with 2 and 4 GTX980 GPUs in Fig. 6 of the original publication^[16]).

Parallelizing a simulation over many nodes to increase the performance leads to a dramatic decrease of the P/P ratio (Fig. 1, top right corner). For projects where the total amount of sampling is more important than the length of the individual

trajectories, it is therefore advisable to run many single-node simulations instead of a few multinode ones.

Hardware and software developments 2014–2018

Hardware developments. Over the past 5 years, GPU compute power has significantly increased (compare Table 2 and Fig. 2). The recent NVIDIA *Turing* architecture GPUs (olive bars) are two to three generations ahead of the *Kepler* and *Maxwell* architectures (black bars) we tested in 2014 and have improved single precision (SP) floating point performance by more than three-fold in this period. This was enabled by a leap in semiconductor manufacturing technology, shrinking transistors from the 28 nm process used in *Kepler* and *Maxwell* to 12 nm on *Volta* and *Turing*, and increasing transistor count more than fivefold. In contrast, during the same period, CPU manufacturing has taken a more modest step forward from 22 to 14 nm. However, effective MD application performance of GPUs has in some cases improved even more than what raw floating point performance would suggest, thanks to microarchitectural improvements making GPUs more efficient at and therefore better suited for general purpose compute. As an example, consider the performance of the compute-bound nonbonded pair interaction kernel (Fig. 3, top panel). While the throughput increase between earlier GPU generations, like the Tesla K40 to Quadro M6000 to Quadro P6000, tracked the FLOP rate increase quite well (approximately 1.9× for both), the Tesla V100 shows 1.7× improvement for only 1.1× SP FLOP rate advantage (see purple bar on Figs. 2 and 3). Unlike the aforementioned professional GPUs with similar maximum power ratings, comparing consumer GPUs is less straightforward. However, a similar pattern is still well illustrated when contrasting the *Pascal* generation GTX 1080Ti with the *Turing* RTX 2080 GPU. Despite the approximately 10% lower FLOP rate as well as 10%

Table 2. Technical specifications of GPU models used in this study. Frequency and SP FLOP for professional NVIDIA GPUs are based on the default and maximum “application clocks,” while for the consumer NVIDIA and AMD GPUs are based on the published base and boost clocks. Note that the NVIDIA GeForce GPUs will often operate at even higher clocks under compute workloads than those indicated by the boost clock (even without factory overclocking), especially in wellcooled environments. For GPUs that were available in 2018, the last column lists the approximate net price in 2018. Note that only the NVIDIA cards can execute CUDA code, whereas on AMD cards OpenCL can be used.

Model	Manufacturer	Architecture	Compute units	Base–boost clock (MHz)	SP TFLOPS	≈ price (€ net)
NVIDIA consumer GPUs:						
GTX 680	NVIDIA	Kepler	1536	1006–1058	3.1–3.3	–
GTX 980	NVIDIA	Maxwell	2048	1126–1216	4.6–5	–
GTX 1070	NVIDIA	Pascal	1920	1506–1683	5.8–6.5	310
GTX 1070Ti	NVIDIA	Pascal	2432	1607–1683	7.8–8.2	375
GTX 1080	NVIDIA	Pascal	2560	1607–1733	8.2–8.9	420
GTX 1080Ti	NVIDIA	Pascal	3584	1480–1582	10.6–11.3	610
RTX 2070	NVIDIA	Turing	2304	1410–1710	6.5–7.9	450
RTX 2080	NVIDIA	Turing	2944	1515–1710	8.9–10.1	640
RTX 2080Ti	NVIDIA	Turing	4352	1350–1545	11.8–13.4	1050
AMD GPUs:						
Radeon Vega 64	AMD	Vega	4096	1247–1546	10.2–12.7	390
Radeon Vega FE	AMD	Vega	4096	1382–1600	11.3–13.1	850
NVIDIA professional GPUs:						
Tesla K40c	NVIDIA	Kepler	2880	745–875	4.3–5	–
Tesla K80	NVIDIA	Kepler	4992	562–875	5.6–8.7	–
Quadro M6000	NVIDIA	Maxwell	3072	988–1152	6.1–7.1	–
Quadro GP100	NVIDIA	Pascal	3584	1303–1556	9.3–11.1	–
Quadro P6000	NVIDIA	Pascal	3840	1506–1657	11.6–12.7	4600
Tesla V100	NVIDIA	Volta	5120	1275–1380	13.6–14.1	8000

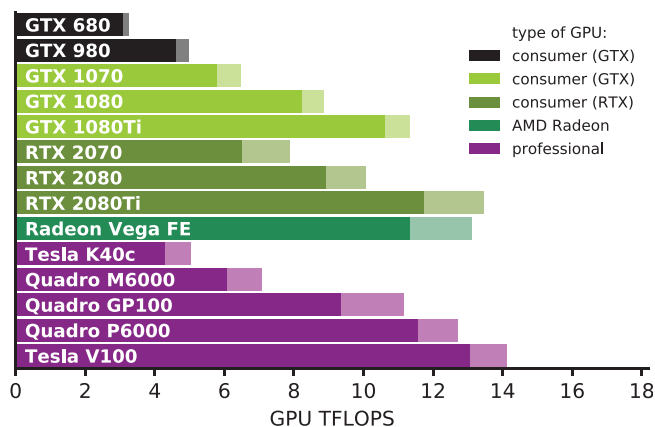


Figure 2. Raw SP floating-point performance of selected GPU models as computed from cores and clock rate (shaded area depicts FLOPS when running at boost/maximum application clock rate). Consumer GPUs that were part of the 2014 investigation are depicted in black, recent consumer GPUs in shades of green, Quadro and Tesla professional GPUs in purple. [Color figure can be viewed at wileyonlinelibrary.com]

lower maximum power rating, the 2080 is 40% and 29% faster in nonbonded and PME computation, respectively. Across two generations of GPUs, we observe up to 6x and 4x performance improvement for the GPU-offloaded GROMACS compute kernels,

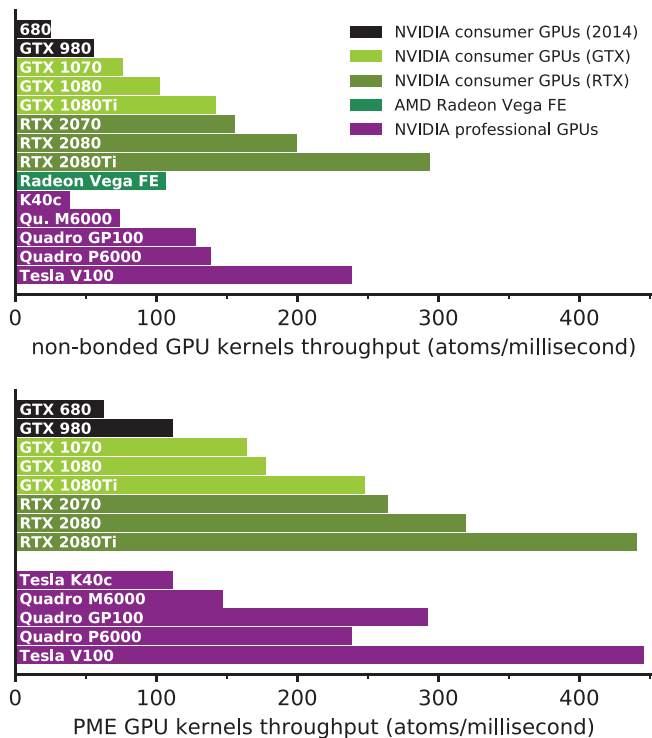


Figure 3. Throughput of the GPU-offloaded computation: short-range nonbonded interactions (top panel) and PME long-range electrostatics (bottom panel) across GPU devices representing the various hardware generations and categories in this study. Coloring matches that of Figure 2. Throughput of computation is expressed as atoms per millisecond to aid comparing to the raw FLOPS. Measurements were done by profiling kernel execution (with concurrency disabled) of a run with a TIP3P water box of 384,000 atoms (cutoff radii 1.0 nm, PME mesh spacing 0.125 nm, time step 2 fs), chosen to allow comparing different GPUs with different scaling behavior each at peak throughput for both kernels. [Color figure can be viewed at wileyonlinelibrary.com]

the nonbonded pair interactions and PME electrostatics, respectively. In contrast, while the theoretical FLOP rate of CPUs has increased by a similar rate as that of GPUs, microarchitectural improvements like the wider AVX512 SIMD instruction sets translated into an only relatively modest gain in application performance, even in thoroughly SIMD-optimized codes like GROMACS.

This confluence of GPU manufacturing and architectural improvements has opened up a significant performance gap between CPUs and GPUs, in particular for compute-bound applications like MD, that do not heavily rely on double-precision floating point arithmetic. Additionally, the performance per Watt advantage and the affordability of high performance consumer GPUs, thanks to the competitive computer gaming industry, have further strengthened the role of GPUs in the MD community. The application performance improvements on GPUs have also led to a shift in typical hardware balance, important in applications that use offload-based heterogeneous parallelization, which motivated developments toward further GPU offload capabilities of the GROMACS MD engine.

Software developments. On the software side, there have been continuous improvements throughout the four major GROMACS releases between our previous study and the present work: 5.0 (June 2014), 5.1 (August 2015), 2016 (August 2016), and 2018 (January 2018). These releases yielded advances in algorithms (improved pair interaction buffer estimates), SIMD parallelization (in PME, improved bonded, constraints, and new update kernels), and kernel optimizations (improved force accumulation), as well as widespread multithreading optimizations (like sparse summation for per-thread bonded force outputs). Efforts in designing code to increase on-node parallelism (wider SIMD units, higher core counts, and more accelerator-heavy compute nodes) both aim at fueling performance improvements throughout the years making better use of existing hardware but also at preparing the code for hardware evolution, an investment that promises future benefits.

Two significant improvements allowed the 2018 release to take further leaps in performance. First, the *dual pair list* extension of the cluster pair algorithm^[22] was developed with two goals: reducing the computational cost of pair search (and domain decomposition) and facilitating optimal simulation performance without manual parameter tuning. The dual pair list algorithm enables retaining the pair list far longer, while avoiding nonbonded computation overheads. This is achieved by building a larger pair list less frequently (every 100–200 MD steps) using a suitably longer buffered interaction cutoff. By using a frequent *pruning* step based on a short buffered cutoff, a smaller pair list with a lifetime of typically 5–15 MD steps (value auto-tuned at runtime) is obtained, which is then used in evaluating pair interactions. Less frequent pair search significantly reduces the cost of this expensive computation whereas the list pruning avoids introducing overheads of extra pair interaction evaluations due to a large buffer. The added benefit is that tuning the search frequency to balance these two costs for optimal performance is no longer needed.

The second major improvement is that, while GROMACS versions 4.6, 5.x, and 2016 could offload only the short-range part of Coulomb and van der Waals interactions to the GPU, in the 2018 release the PME mesh part can also be offloaded to CUDA-enabled devices

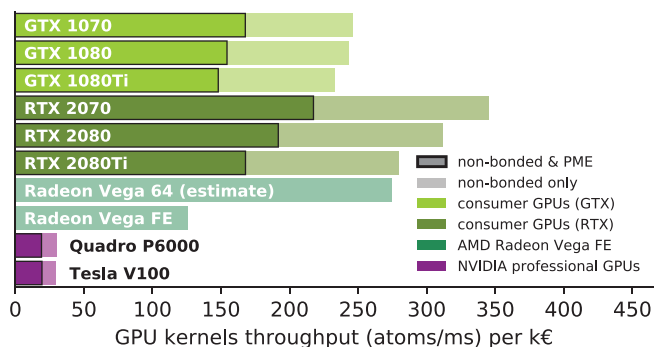


Figure 4. Proxy metric for P/P ratio of selected GPU models, computed as GROMACS 2018 GPU kernel performance divided by price given in Table 2. Light/shaded bars of each row show P/P derived for timings of nonbonded kernels only, dark bars show P/P derived from total timings of nonbonded and PME kernels. The P/P of the AMD Vega 64 is estimated from the measured throughput of AMD Vega FE but the price of Vega 64. [Color figure can be viewed at wileyonlinelibrary.com]

(Fig. 5). By enabling offload of more work, the computational balance within GROMACS can be shifted such that it exploits the shift in hardware balance in the recent years. This improvement makes it possible to efficiently utilize nodes with more and stronger GPUs and it enables significantly higher P/P ratios than prior versions on

recent hardware. At the same time, sticking to the offload-based parallelization approach with improvements focused on both CPU and GPU is still important and has two major benefits: (1) It makes sure that nearly all of the versatile GROMACS features remain supported with GPU acceleration (as opposed to limiting use of GPUs to only the subset of common features ported to GPUs). (2) Additionally, offload allows minimizing the risk of vendor lock-in, which is not negligible when the hardware of the dominant manufacturer can only be employed (with meaningful performance) using proprietary tools and nonstandards-based tools and programming models.

As PME offloading with OpenCL will only be supported starting from the 2019 release, we did not include AMD GPUs in our benchmark study yet. However, we expect that recent AMD GPUs will be competitive against similarly priced NVIDIA consumer GPUs, as suggested by both the nonbonded kernel performance (see Fig. 3) and by the proxy P/P ratio (Fig. 4).

The offloading approach works best at a balanced GPU/CPU compute power ratio, that is, if the ratio is adapted to the typical requirements of GROMACS simulations (Fig. 5). With our benchmarks and our hardware choices, we aim to determine this optimum. The more compute work is offloaded, the more this balance is shifted toward the GPU, which enables higher P/P ratios

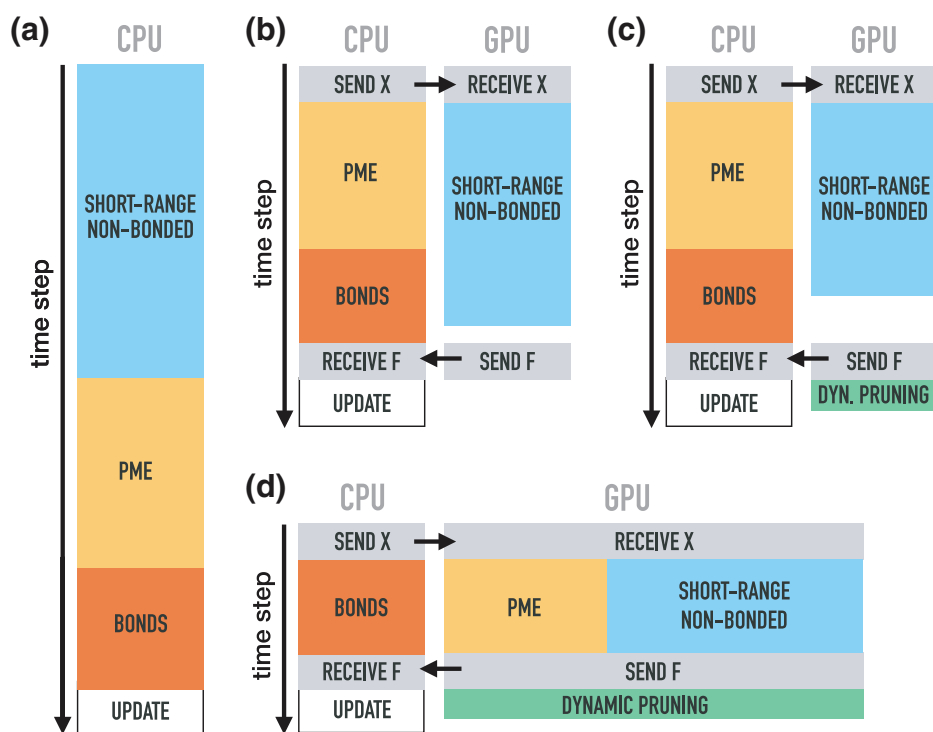


Figure 5. Comparison of different offloading schemes employed by GROMACS 4.6, 5.x, and 2016 (B) and GROMACS 2018 (C, D). Differently colored boxes represent the different main force computation parts of a typical MD step, whereas gray boxes represent the CPU-GPU data transfers. By offloading compute-intensive parts of the MD step and using algorithms to optimize concurrent CPU-GPU execution, the wall-time required by a time step (black vertical arrow) is decreased. A. Without GPUs, the short-range nonbonded interactions (blue), PME (orange), and the bonded interactions (red) are computed on the CPU. Once all forces have been derived, the atomic positions are updated (white). B. Since version 4.6, GPU(s) can compute the nonbonded forces, while the CPU(s) do PME and the bonded forces. As a result, the wall clock time per MD step is significantly shortened, at the small expense of having to communicate positions \mathbf{x} and forces \mathbf{F} between CPU and GPU (gray). C. Version 2018 introduced the dual pair list algorithm, which i) reduces the number of short-range nonbonded interactions that are calculated, and ii) reduces the frequency of doing pair search on the CPU (not shown here). There is no computational overhead added, as the dynamic list pruning (green box) happens on the GPU(s), while the CPU updates the positions (white). D. Since version 2018, also PME can be computed on a GPU, further reducing the wall clock time per MD step, if enough GPU processing power is available. [Color figure can be viewed at wileyonlinelibrary.com]

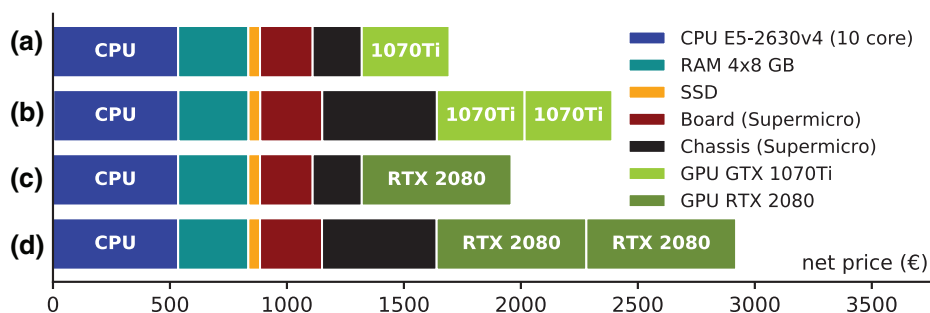


Figure 6. Breakdown of node costs into individual components for four exemplary node types. Costs for CPU, RAM, and SSD disk are identical for nodes A–D, however, with two GPUs (B, D) a chassis with a stronger power supply has to be used, making the node significantly more expensive. [Color figure can be viewed at wileyonlinelibrary.com]

if GPUs are cheap. The switch to GROMACS 2018 shifted the optimal CPU/GPU balance significantly toward the GPU, as shown in the following.

Assembling optimal GPU nodes for GROMACS 2018. Choosing the right hardware components to build a node with a competitive P/P ratio is a puzzle on its own (Fig. 6). Let us for a moment focus on the proxy P/P ratio of the GPUs only (Fig. 4). Considering raw GROMACS GPU kernel throughput, of the *Pascal* architecture GPUs, the 1080 offers the highest P/P ratio, whereas of the *Turing* GPUs, the 2070 performs best. However, considering GPUs with optimal P/P ratio only is not always be the best solution, as a node is often more expensive the more GPUs it can accommodate (see Fig. 6 for an example). As a result, to optimize for the combined P/P ratio of a whole node, it is typically better to choose a consumer GPU with a lower P/P ratio but higher performance.

Apart from the Gold6148F × 2 node with two Tesla V100's (Table 3 bottom, and Fig. 9 top), we did not build and benchmark any nodes with professional GPUs for two reasons: (i) A Tesla GPU already costs more than most of the tested nodes including their consumer GPU(s). (ii) For MD simulations with GROMACS, the added benefit of using professional GPUs is marginal (ECC reliability and warranty) and consumer models with comparable application performance generally exist.

Methods

For the main part of this study, the same two benchmark input files as in our 2014 investigation were used (Table 1), to facilitate the comparison between new and old hardware and software.

Software environment

Benchmarks done for evaluating GROMACS developments used the latest release of each version. All other benchmarks have been performed using GROMACS 2018, with AVX2_256 SIMD instructions switched on at compile time for Intel CPUs and AVX2_128 SIMD instructions for AMD CPUs (with the exception of Table 4, where AVX_256 was used reflecting the hardware capabilities). Additionally, as version 4.6 did not have SIMD kernel support for the AVX2_256 instruction set, here we used the AVX_256 build with adding the `-mavx2 -mfma` compiler optimization flags to allow AVX2 instructions to be generated.

On nodes with no or a single GPU, GROMACS' built-in thread-MPI library was used, whereas on multi-GPU nodes Intel MPI 2017 was used. OpenMP support was always enabled.

Additionally, GROMACS was linked against the portable hardware locality library `hwloc`^[23] version 1.11. On nodes without GPUs, the FFT needed by PME is calculated on the CPU. Therefore, we employed FFTW 3.3.7,^[24] compiled with GCC 4.8 using the options `--enable-sse2 --enable-avx --enable-avx2` that are recommended for best GROMACS performance. On GPU nodes, CUDA cuFFT was automatically used with PME offloading.

All hardware was tested in the same software environment by booting the nodes from a common image with Scientific Linux 7.4 as operating system (except the Gold6148/V100 nodes, which ran SLES 12p3, and the GROMACS evaluation benchmarks, which ran on nodes with Ubuntu server 16.04.)

GROMACS was compiled either with GCC 5.4 and CUDA 8.0 or with GCC 6.4 and CUDA 9.1 for the main study, while in the GROMACS evaluation section GCC 7.3 and CUDA 10 was used.

Impact of compiler choice and CUDA version

To assess the impact of the chosen GCC/CUDA combination on the measured performances, we ran MEM benchmarks on identical hardware, but with both CUDA/GCC combinations. On a node with two E5-2670v2 CPUs plus two GTX 1080Ti GPUs, our MEM benchmark runs consistently faster with GCC6.4/CUDA9.1 over GCC 5.4/CUDA8, by 3.5% (average over 10 runs). On an E3-1270v2 CPU with GTX 1070, the factor is about 4%, whereas on a workstation with E5-1650v4 and GTX 980 it is about 1.5%. To correct for this effect when comparing hardware, the performances measured with the older GCC/CUDA combination have been multiplied with the factor 1.025.

The performance difference between using CUDA 9.1 and CUDA 10.0 was determined in a similar manner and as it turned out to be less than 0.5%, we did not correct for this small effect.

Benchmarking GROMACS performance evolution

For GROMACS evaluation benchmarking (Figs. 7 and 8), data collection protocols tailored for characterizing performance of the various code versions were used. All runs were carried out using a GPU attached to the PCI bus of the CPU employed (or the first CPU where the master thread of the run was located when both CPUs were used in Fig. 8). Two CPU threads per core were used, profiting from HyperThreading with threads pinned.

For the evaluation of performance as a function of CPU cores per GPU (Fig. 8), we would ideally use CPUs that only differ in the number of cores and are identical otherwise. We mimicked such a scenario with a single CPU model by using just a part of

Table 3. Single-node performances (average over two runs) for GROMACS 2018 and corresponding P/P ratios. On nodes with N GPUs, the aggregate performance of N simulations is reported (except 4R where four simulations are run).

U	processor AMD/Intel	Sockets × cores	Clock (GHz)	Mounted GPUs	MEM (ns/d)	RIB (ns/d)	≈price (€ net)	P/P MEM (ns/d/ 122 €)	P/P RIB (ns/d/ 1530 €)
1	E3-1270v5 ^[a]	1 × 4	3.6	1070	62.96	3.14	1300	5.78	3.61
1	E3-1270v5 ^[a]	1 × 4	3.6	1070Ti	70.05	3.23	1395	6.13	3.54
1	E3-1270v5 ^[a]	1 × 4	3.6	1080	71.16	3.1	1440	6.03	3.29
1	E3-1270v5 ^[a]	1 × 4	3.6	1080Ti	84.63	3.8	1630	6.33	3.57
1	E3-1240v6 ^[b]	1 × 4	3.7	–	11.7	0.69	1040	1.37	1.02
1	E3-1240v6 ^[a]	1 × 4	3.7	1070	68.72	3.48	1350	6.21	3.94
1	E3-1240v6 ^[a]	1 × 4	3.7	1080	78.04	3.48	1460	6.52	3.65
1	E3-1240v6 ^[a]	1 × 4	3.7	1080Ti	95.45	4.38	1650	7.06	4.06
1	E3-1240v6 ^[b]	1 × 4	3.7	2080	110.69	4.85	1680	8.04	4.42
1	E3-1240v6 ^[b]	1 × 4	3.7	2080Ti	133.26	5.77	2090	7.78	4.22
1	Core i7-6700K ^[a]	1 × 4	4.0	1070	67.69	3.28	1350	6.12	3.72
1	Core i7-6700K ^[a]	1 × 4	4.0	1080Ti	92.52	4.03	1650	6.84	3.74
1	Silver 4110 ^[a]	1 × 8	2.1	1080Ti	97.63	5.27	1910	6.24	4.22
1	Silver 4110 ^[a]	1 × 8	2.1	1080 × 2	131.17	6.98	2840	5.63	3.76
1	Silver 4110 ^[a]	1 × 8	2.1	1080Ti × 2	155.66	8.84	3220	5.9	4.2
1	E5-2630v4 ^[a]	1 × 10	2.2	1070	73.61	3.84	1630	5.51	3.6
1	E5-2630v4 ^[a]	1 × 10	2.2	1070Ti	82.41	4.02	1695	5.93	3.63
1	E5-2630v4 ^[a]	1 × 10	2.2	1080	83.42	3.83	1740	5.85	3.37
1	E5-2630v4 ^[a]	1 × 10	2.2	1080Ti	104.01	4.99	1930	6.57	3.96
1	E5-2630v4 ^[b]	1 × 10	2.2	2080	115.41	5.88	1960	7.18	4.59
1	E5-2630v4 ^[b]	1 × 10	2.2	2080Ti	146.28	7.27	2370	7.53	4.69
1	E5-2630v4 ^[a]	1 × 10	2.2	1080Ti × 2	179.51	8.58	2860	7.66	4.59
1	E5-2630v4 ^[b]	1 × 10	2.2	2080 × 2	201.33	10.13	2920	8.41	5.31
1	Silver 4114 ^[a]	1 × 10	2.2	1070Ti	81.31	4.49	1855	5.35	3.7
1	Silver 4114 ^[a]	1 × 10	2.2	1080	82.89	4.34	1900	5.32	3.49
1	Silver 4114 ^[a]	1 × 10	2.2	1080Ti	103.9	5.57	2090	6.06	4.08
1	Silver 4114 ^[b]	1 × 10	2.2	2080Ti	147.29	7.85	2530	7.1	4.75
1	Silver 4114 ^[a]	1 × 10	2.2	1080 × 2	142.7	7.29	3020	5.76	3.69
1	Silver 4114 ^[a]	1 × 10	2.2	1080Ti × 2	165.19	9.58	3400	6.07	4.31
D	Ryzen 1950X ^[b]	1 × 16	3.4	1080Ti	94.9	5.01	2440	4.75	3.14
D	Ryzen 1950X ^[b]	1 × 16	3.4	2080	106.08	5.56	2470	5.24	3.44
D	Ryzen 1950X ^[b]	1 × 16	3.4	1080Ti × 2	172.62	9.16	3050	6.9	4.6
D	Ryzen 1950X ^[b]	1 × 16	3.4	2080 × 2	196.73	10.08	3110	7.72	4.96
D	Ryzen 1950X ^[b]	1 × 16	3.4	2080 × 3	267.51	12.79	3750	8.7	5.22
D	Ryzen 1950X ^[b]	1 × 16	3.4	2080 × 4	332.83	14.07	4390	9.25	4.9
1	Epyc 7401P ^[b]	1 × 24	2.0	–	28.71	2.28	3500	1	1
1	Epyc 7401P ^[b]	1 × 24	2.0	1080Ti × 2	191.66	9.49	4720	4.95	3.08
1	Epyc 7401P ^[b]	1 × 24	2.0	1080Ti × 4	369.06	16.96	5940	7.58	4.37
2	Gold6148F × 2 ^[c]	2 × 20	2.4	V100 × 2	300.76	19.95	23,200	1.58	1.32
2	Gold6148F × 2 ^[c]	2 × 20	2.4	V100 × 2 (4R)	393.32	27.27	23,200	2.07	1.8

U = rack space requirements in units per node, D for desktop chassis. P/P ratios were normalized such that values ≥1 result.
[a] Using CUDA 8.0 + GCC 5.4 + Intel MPI 2017.
[b] Using CUDA 9.1 + GCC 6.4 + Intel MPI 2017.
[c] Using CUDA 10.0 + GCC 6.4 + Intel MPI 2018.

its available cores. However, as modern CPUs, when only partially utilized, can increase the clock frequency of the busy cores, the comparison would be unfair. We therefore made sure that the cores not used by our benchmark were kept busy with a concurrent CPU-only GROMACS run (using the same input system as the benchmark), so that approximately the same clock frequency is used independent of how many cores the benchmark runs on.* GROMACS evaluation benchmarks were repeated three times, with averages shown.

*Note that with this protocol the last level cache is shared by the two runs collocated on the CPU. Hence, measurements are not equivalent with turning off CPU cores not intended to be used and fixing a constant CPU frequency across all active cores, an alternative which would provide benchmarks for a small number of cores with an unrealistic amount of cache.

Lastly, all compile and runtime settings, other than the ones tested, were left at their default or autotuned values (including pair search frequency and CPU-GPU load balancing).

Measuring the performance of a node

The performance of different node types can be assessed in various ways. Moreover, different benchmark procedures lead to different hardware rankings. Our requirements on the benchmarking procedure and the resulting hardware ranking were (1) The benchmarks should closely mimic the intended hardware use. (2) Aggregation of compute power (e.g. combining hardware components of two nodes into a single node), which may offer price and rack space savings, should not be penalized.

Our motivation for the second requirement is the following: Assume you compare (a) two single-socket nodes with CPU X

Table 4. Single-node performances for GROMACS 2018 as in Table 3 but for upgrading existing nodes with modern GPUs. Here, the P/P ratios have been calculated from the performance increase (i.e., performance with upgraded GPU(s) minus performance with old GPU) and the cost of the GPU(s). All benchmarks use CUDA 9.1, GCC 6.4, and, for the multi-GPU setups, Intel MPI 2017. For optimum performance, PME mesh part was offloaded to the GPU, except where indicated (*).

U	Processor Intel	Sockets × cores	Clock (GHz)	Mounted GPUs	MEM (ns/d)	RIB (ns/d)	≈price (€ net)	P/P MEM (ns/d/ 122 €)	P/P RIB (ns/d/ 1530 €)
Existing node with old GPU:									
1	E3-1270v2	1 × 4	3.5	680	26.9*	1.6	0	-	-
With upgraded GPU:									
1	E3-1270v2	1 × 4	3.5	2080	91.7	4	640	12.4	5.7
Existing node with old GPUs:									
2	E5-2670v2 × 2	2 × 10	2.5	780Ti × 2	104.8*	6.7	0	-	-
With upgraded GPUs:									
2	E5-2670v2 × 2	2 × 10	2.5	1080 × 2	163.4	7.8	840	8.5	1.9
2	E5-2670v2 × 2	2 × 10	2.5	1080Ti × 2	208.4	10.2	1220	10.4	4.3
2	E5-2670v2 × 2	2 × 10	2.5	1080Ti × 4	361.2	17.9	2440	12.8	7
Existing node with old GPUs:									
2	E5-2680v2 × 2	2 × 10	2.8	K20Xm × 2	83.2*	5	0	-	-
With upgraded GPUs:									
2	E5-2680v2 × 2	2 × 10	2.8	1080Ti × 2	212.7	10.1	1220	12.9	6.4
2	E5-2680v2 × 2	2 × 10	2.8	2080 × 2	237.7	11.5	1280	14.7	7.8
2	E5-2680v2 × 2	2 × 10	2.8	2080 × 4	409.6	20.3	2560	15.6	9.2

and GPU Y each, to (b) a dual-socket node with two CPUs of type X and two GPUs of type Y. The aggregate performance of (a) and (b) is expected to be identical, as two independent simulations can always run on (b).

A benchmark protocol matching both requirements is running N simulations on nodes with N GPUs in parallel, each using $1/N$ of the available CPU cores or hardware threads, and reporting the aggregate performance, that is, the sum of the performances of the individual simulations. This protocol can easily be executed by using the GROMACS-multidir command line argument. As in the initial phase, the load balancing mechanisms have not yet found their optimum, we excluded the first n time steps from the performance measurements that were run for a total of n_{tot} time steps. For the MEM and VES benchmarks, we used $n = 15,000$ and $n_{\text{tot}} = 20,000$, whereas for the RIB and BIG benchmarks, we used $n = 8000$ and $n_{\text{tot}} = 10,000$.

On single-socket nodes with one GPU, using a single rank with as many OpenMP threads as available cores (or hardware threads) is usually fastest, as one avoids the overhead introduced by MPI and domain decomposition.^[16,17] Therefore, the single-GPU benchmarks of the old investigation and the present study are comparable, whereas on multi-GPU nodes, the new benchmark protocol is expected to yield a higher aggregate performance than the single-simulation performance that was measured in 2014.

Power consumption measurements

We measured the power consumption of a node with the Linux programs ipmi-sensors version 1.5.7 (<http://www.gnu.org/software/freeipmi/>) and ipmitool version 1.8.17 (<https://github.com/ipmitool/ipmitool>). On some nodes, for example, on the Ryzen workstations that do not support reading out the power draw via ipmi, we used a VOLTcraft Power Monitor Pro multimeter. In all cases, we computed the average of 60 separate power readings with 1 s time difference each. During the

power measurements, the RIB benchmark was running with the same settings that were used to derive the performance.

Coarse grain simulations with Martini

In order to check to what extent the results from the atomistic benchmarks are transferable to coarse grain simulations, we added benchmarks that use the Martini^[25,26] force field. With Martini, the particle density is lower than for all-atom force fields and the electrostatic interactions are usually not calculated via PME. Therefore, it is unclear whether the hardware optimum for coarse grain simulations is the same as for atomistic simulations.

To facilitate the comparison with the atomistic systems, we chose coarse grain systems with a similar number of particles, respectively, see Table 1. The small Martini benchmark system “VES” is a POPE vesicle in water and with a total of 72,076 particles, comparable in size with the MEM atomistic benchmark. It

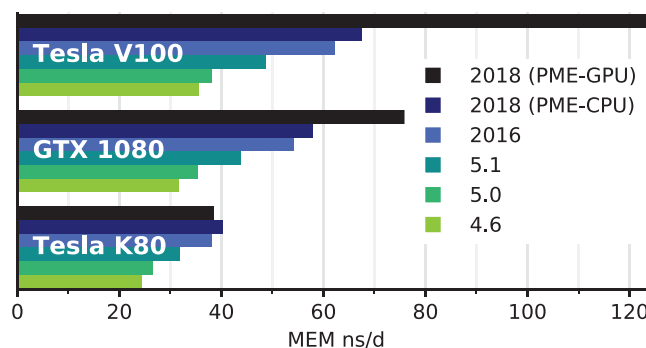


Figure 7. Evolution of the GROMACS performance on GPU nodes for versions 4.6–2018. The short-range nonbonded interactions were offloaded to the GPU in all cases, whereas for version 2018, also the PME mesh contribution can be offloaded (topmost black bars). Tesla V100 and GTX 1080 GPUs were mounted in a node with two Xeon E5-2620v4 processors (2 × 8 cores). The Tesla K80 GPU was mounted in a node with two Xeon E5-2620v3 processors (2 × 6 cores). [Color figure can be viewed at wileyonlinelibrary.com]

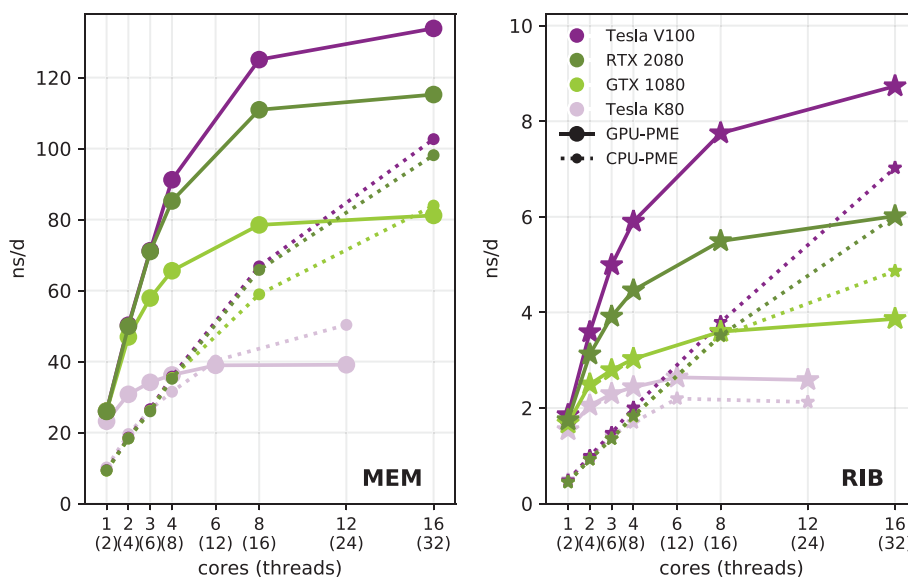


Figure 8. GROMACS 2018 performance as a function of CPU cores used per GPU. The GTX 1080, RTX 2080, and Tesla V100 cards were installed in server nodes along Xeon E5-2620v4 processors (2 × 8 cores), whereas the Tesla K80 was installed in a node with two E5-2620v3 processors (2 × 6 cores). Solid lines illustrate performance with PME offloaded to the GPU, whereas dotted lines with smaller symbols indicate performance with PME computed on the CPU (cores indicated on the horizontal axis). [Color figure can be viewed at wileyonlinelibrary.com]

is one of the example applications available for download at www.cgmartini.nl.^[21] The large Martini benchmark system “BIG” was created from the mammalian plasma membrane example system.^[27] To arrive at the final large benchmark system, a patch of 2 × 2 copies in *x*- and *y*-directions of the original membrane was created with `gmx genconf`, resulting in a final membrane with 2,094,812 particles, thus comparable in size to the RIB atomistic benchmark.

Following the suggestions for Martini simulations with GPUs, we used the *New-RF* set of simulation parameters.^[28,29] As with version 2018, the dual pair list algorithm was introduced (see Fig. 5C), we increased the neighbor searching interval from 20 (as used in the *New-RF* parameter set) to 50 steps (with an inner pair list lifetime of four steps) for improved performance.

Results

Evaluation of GROMACS performance developments

As shown in Figure 7, thanks to the algorithmic improvements and optimizations described earlier, across the initial four releases since our previous study, between versions 4.6 and 2016, simulation performance improved by up to 65% on previous-generation hardware (e.g. Tesla K80) and by as much as 75–90% on more recent hardware (e.g. GTX1080 and Tesla V100). Between the 2016 and 2018 versions with PME on the CPU, we measured a 6%–9% performance increase (light and dark blue bars in Fig. 7), which is largely due to the dynamic pruning algorithm. We expect this advantage to grow even larger with future GPU hardware because the faster the GPU-offloaded computation gets, the larger the benefit of this algorithm will be. For the given benchmark systems and server setups, we observe additional PME offload improvements of 35%–84% when using recent GPUs (see black bars in Fig. 7). At the same time, on the legacy hardware setup, offloading to the older generation Tesla K80 leads to slowdown. To better understand this performance change, we explore the performance

characteristics of the heterogeneous PME-offload code in GROMACS 2018 in the following.

Our evaluation benchmarks were carried out on servers representing GPU-dense setups, consisting of Xeon CPUs with rather modest performance combined with fast accelerators. These are traditionally challenging for GROMACS as illustrated by the strong dependence of the performance on the number of cores used (indicated by the dotted lines in Fig. 8). One of the main performance goals of the GROMACS 2018 development was to reach a close to the peak simulation rate of the previous offload scheme (non-bonded only) on balanced hardware, but with only a few CPU cores accompanying the GPU. Of our benchmarked hardware setups, the GTX 1080 (light green curves in Fig. 8) combined with about 12–14 cores of the E5-2620v4 processors could be considered a balanced setup. For the two systems, with only four cores per GPU, the PME-offload feature allows reaching 80% and 90% of the peak with no PME offload, respectively.

Whereas the 2016 release required many fast cores to achieve a good load balance between CPU and GPU, with the 2018 release, in most cases, only 4–6 slower (typical server) cores as the ones in our benchmarked systems are sufficient to reach >80% of the peak simulation performance (e.g., as obtained with all 16 cores of a node here). Workstations typically have a few cores, but these are fast. In contrast, servers often have more but slower cores than workstations. To compare the raw CPU processing power of both node types, we consider the “core-GHz”, that is, the number of cores multiplied with the clock frequency. We determined that 10–15 “core-GHz” is generally sufficient to reach close to peak performance with a mid- to high-end GPU in typical biomolecular simulation workloads like the ones used here. If there is however significant work left for the CPU after offloading the nonbonded and PME computation (e.g. a large amount of bonded interactions or a free energy computation setup), more CPU cores may be required for an optimal balance. Additionally, this balance does of course shift as bigger and faster GPUs become available, like

the Tesla V100 (or the similarly performing GeForce RTX 2080Ti) does in fact need around 8–10 cores equivalent to 16–20 “core-GHz” before the performance curve flattens (purple lines). The increasing size of GPUs however also poses a computational challenge: large devices are difficult to saturate and cannot obtain their peak throughput with common workloads like the MEM benchmark, which is why the advantage of the Tesla V100 over the RTX 2080 is relatively small especially in comparison to the much larger RIB benchmark case (purple and dark green lines on Fig. 8).

An additional benefit of PME offload is that the achievable peak performance also increases and, with fast GPUs, a significant performance increase is achieved (see dark green and purple lines in Fig. 8) that was previously not possible with slower accompanying CPUs. Conversely, however, with CPU-GPU setups ideal for earlier GROMACS versions, PME offload may not improve performance. In particular on legacy GPU hardware, PME offload is often slower what is reflected in the early performance cross-over around 5–6 cores per GPU in the Xeon E5-2620v3 CPUs with Tesla K80 benchmarks (solid and dotted pink lines).

In summary, the 2018 version, thanks to the combination of the dynamic pruning algorithm and PME offload, enables using fewer and/or slower CPU cores in combination with faster GPUs, which is particularly suitable for throughput studies on GPU-dense hardware. With hardware setups close to balanced for the 2016 release, PME offload will not lead to significant performance improvements but the new code capabilities open up the possibility for upgrades with further accelerators. As an example, given an 8-core workstation CPU (like an AMD Ryzen 7 2700), which will be approximately as fast as 10–16 cores in Figure 8, when combined with a GTX 1080, there would be little improvement from PME offload, and even with an RTX 2080, the improvement would be modest (assuming similar workload as ours). However, in such a workstation, adding a second GPU would nearly double the performance.

Which hardware is optimal for MD?

Table 3 and Figure 9 show the results of our current hardware ranking. Overall, the P/P ratio of the examined consumer GPU nodes is about a factor of 3–6 higher compared to their

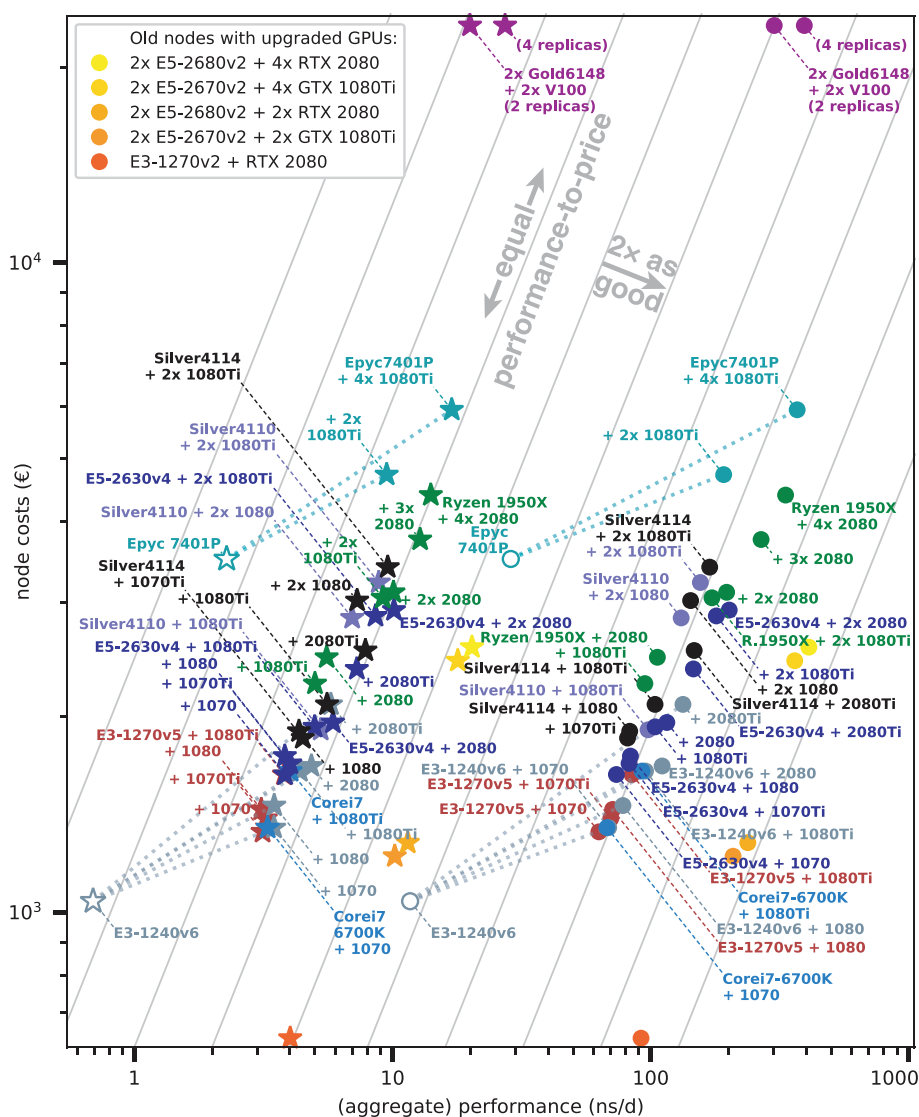


Figure 9. (Aggregate) simulation performance in relation to net node costs. MEM (circles) and RIB (stars) symbols are colored depending on CPU type. Symbols with white fill denote nodes without GPU acceleration; dotted lines connect GPU nodes with their CPU counterparts. Gray: isolines of equal P/P ratio like in Figure 1 with superior configurations to the lower right. Old nodes with upgraded GPUs from Table 4 are shown in yellow-orange colors (legend). [Color figure can be viewed at wileyonlinelibrary.com]

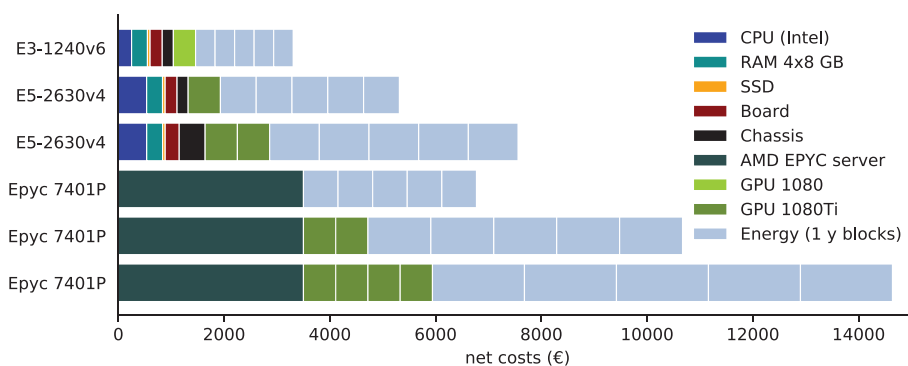


Figure 10. Breakdown of total costs for selected node types, taking into account 0.2 € per kWh for energy and cooling, for a lifetime of 5 years. [Color figure can be viewed at wileyonlinelibrary.com]

counterparts without GPUs. The P/P ratios of new nodes with consumer GPUs are all very similar; most of them are less than a factor of 1.5 apart and thus scatter about one isoline in the log–log plot (Fig. 9). Note that this similarity results from our hardware preselection and does not imply that any node with consumer GPU(s) has a comparable P/P ratio. There are lots of hardware combinations possible that we did not consider because high costs of one or more individual components preclude a competitive P/P ratio from the start.

The cheapest nodes with a good P/P ratio, starting at ≈ 1400 € net, are Intel E3-1270v5, E3-1240v6, or Core i7-6700K CPUs combined with a GeForce 1070(Ti), respectively. The best P/P ratio (with current pricing) is offered by combining a E5-2630v4

or Ryzen 1950X CPU with two (or possibly more) RTX 2080 GPUs starting at 3000 € net. The best aggregate performance for consumer GPU nodes was identified for the AMD Epyc 24-core node combined with four 1080Ti GPUs.

Concerning space requirements, most node types listed in Table 3 fit in one height unit (U) of rack space. One exception is the Ryzen Threadripper 1950X that was available in a desktop chassis only (using ≈ 4 U, if mounted in a rack).

Alternative: Upgrade existing nodes with recent GPUs

An attractive alternative to replacing old GPU nodes with new hardware is to solely exchange the existing GPUs with

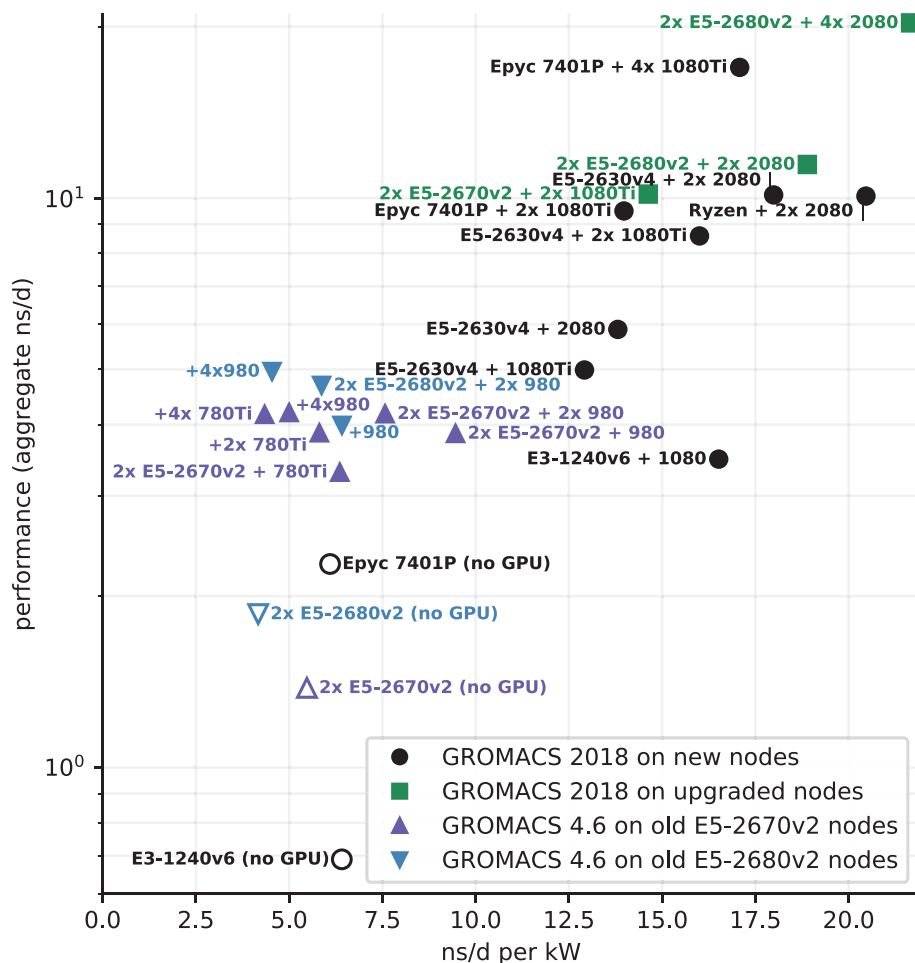


Figure 11. RIB performance of selected node types in relation to their energy efficiency. Nodes without GPUs (white fill) show both low performance as well as low energy efficiency, independent of GROMACS version and CPU generation. Best energy efficiency is recorded for GROMACS 2018 in combination with new GPUs (black- and green-filled symbols). [Color figure can be viewed at wileyonlinelibrary.com]

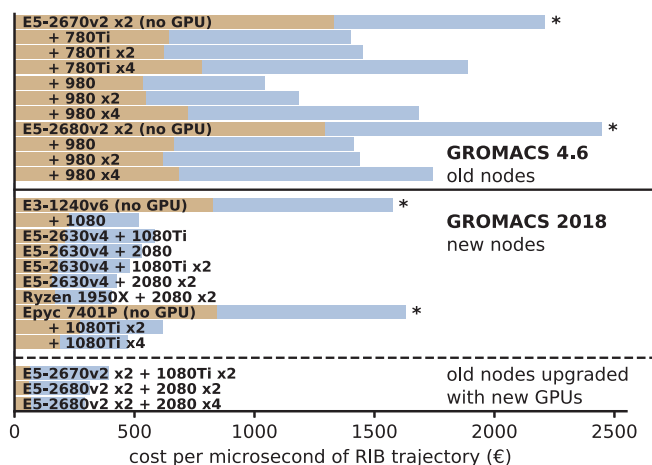


Figure 12. RIB trajectory costs for selected node types assuming 5 years of operation, including costs of 0.2 € per kWh for energy and cooling. The top part shows results from 2014 using GROMACS 4.6,^[16] the lower part depicts results using GROMACS 2018 on recent hardware and on old hardware that was upgraded with new GPUs (lowermost three bars). Nodes without GPUs have the highest trajectory production costs (asterisks). [Color figure can be viewed at wileyonlinelibrary.com]

more powerful recent models. Due to offloading even more interaction types to the GPU, compared to older versions, GROMACS 2018 demands more compute power on the GPU

side, but less on the CPU side. As a result, CPU models from a few years ago often ideally combine with modern GPUs.

For instance, the performance gain for a dual 10-core node with two K20Xm GPUs that was part of the 2014 investigation is a factor of 3.5 for the MEM benchmark for switching the old GPUs with recent RTX 2080 models. Table 4 lists the performance gains for different upgrade scenarios. The top line in each section shows the performance of an old node with GROMACS 2018, whereas the following lines show how performance increases when GPUs are upgraded. Depending on the exact old and new GPU type, one can easily achieve a twofold higher aggregate performance from these nodes for the comparatively small investment of just having to buy the GPUs and not whole nodes. Performance-wise it makes no difference whether four 1080Ti GPUs are combined with a new Epyc 7401 processor (24 cores), or with two old E5-2670v2 processors (2 × 10 cores). The yellow and orange symbols in Fig. 9 show the performance of these nodes in relation to the costs for upgrading them with modern GPUs.

Energy efficiency

To establish the total costs of a node over its lifetime, we determined its power consumption. Assuming a net cost of 0.2 € per kWh for energy and cooling, we calculated the total costs of selected node types as the sum of hardware and energy costs,

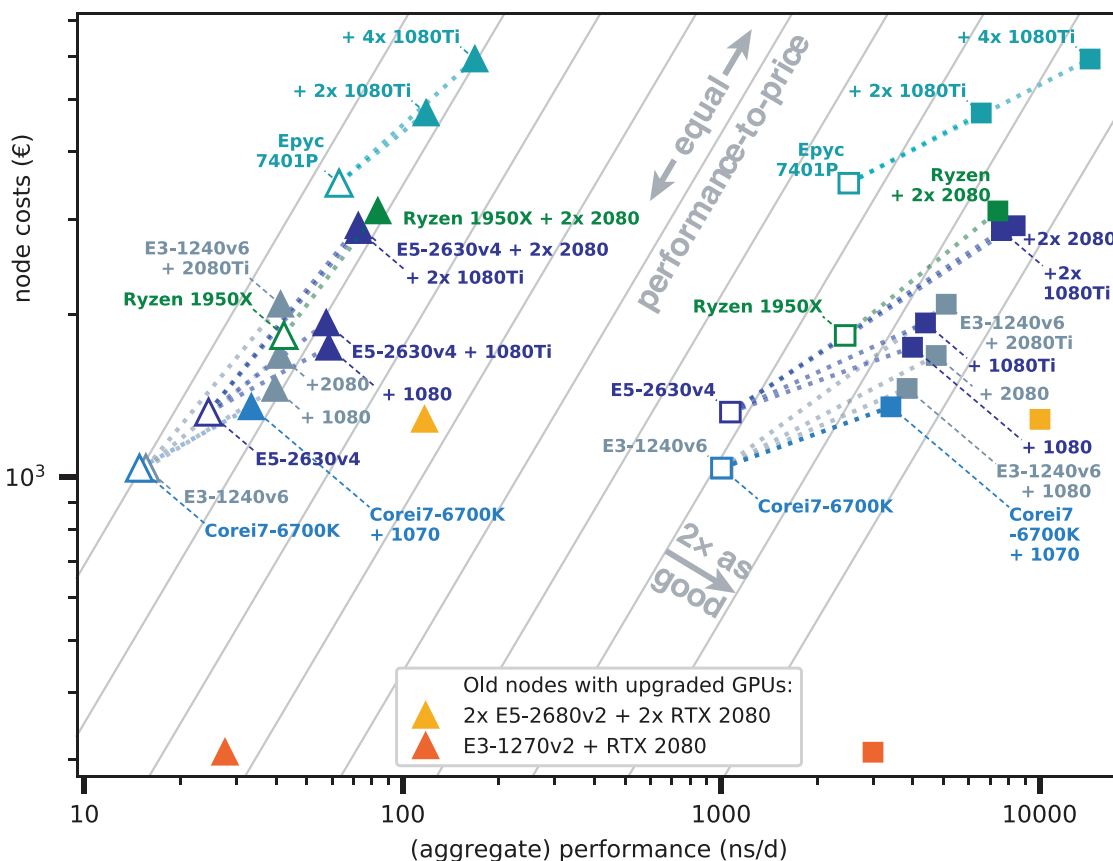


Figure 13. (Aggregate) simulation performance in relation to net node costs as in Figure 9, but for the two coarse grain benchmarks (the vesicle and the big membrane patch) using the Martini force field. VES (squares) and BIG (triangles) symbols are colored depending on CPU type. Symbols with white fill denote nodes without GPU acceleration; dotted lines connect GPU nodes with their CPU counterparts. Reused old nodes with new GPUs are shown in orange colors (legend). [Color figure can be viewed at wileyonlinelibrary.com]

over 5 years of operation (Fig. 10, five separate 1 year blocks given for energy and cooling costs). For the considered node types and an average lifetime of 3–5 years, the costs for hardware and energy are similar. This will, however, not generally be the case, for example, with professional GPUs, the costs for hardware can easily be three times as high.

Figure 11 shows the energy efficiency of selected node types in relation to their GROMACS performance. With GROMACS 2018, in addition to their considerably higher simulation performance, GPU nodes deliver more than two times the performance per Watt compared to CPU nodes.

We derive the total costs for producing MD trajectories on different node types by putting the hardware and energy costs in relation to the amount of produced trajectory (Fig. 12). We make three main observations: (1) Trajectory costs are highest on nodes without consumer GPUs (these are marked with asterisks in the Figure). (2) For GROMACS 4.6 on hardware of 2014, trajectory costs on the best GPU nodes are 0.5–0.6 times that of their CPU-only counterparts. (3) With GROMACS 2018 and current hardware, this factor is reduced to about 0.3.

Coarse grain models

The Martini systems were run on a subset of the node types used for the atomistic benchmarks, Figure 13 shows the results. The overall picture is quite similar to the atomistic benchmarks (Fig. 9) but there are some differences as well.

As for the atomistic systems, the P/P ratio is significantly higher for nodes with consumer GPUs than for CPU nodes. However, the gap between CPU nodes and consumer GPU nodes is less pronounced than in the atomistic case. It is about a factor of 2–4 for the small VES system and a factor of 1.5–2 for the BIG membrane in terms of P/P. As in the atomistic case, reused old nodes upgraded with up-to-date consumer GPUs have the best P/P ratios.

Although the workload of a Martini coarse grain MD system is quite different from the workload of an atomistic system (lower particle density, no PME mesh), it turns out that the node types that are optimal for atomistic MD are also very well suited for running coarse grain simulations.

Conclusions for GROMACS 2018

In 2014, we found that nodes with consumer GPUs provide the *best bang for your buck* due to their significantly higher trajectory output per invested Euro compared to nodes without GPUs or nodes with professional GPUs. This applies equally to GROMACS 2018 on current hardware. Moreover, the existing gap has considerably widened: Taking into account raw node prices, today with GROMACS 2018 one can get a factor of three to six times more trajectory on consumer GPU nodes as compared to a factor of two to three in 2014 with GROMACS 4.6. When including costs for energy and cooling, this factor has increased from two to about three.

This marked improvement became possible by offloading also the PME mesh computations to the GPU, in addition to the short-ranged nonbonded interactions. PME offloading moves the optimal hardware balance even more toward consumer GPUs. CPU/GPU combinations identified as optimal in P/P ratio require

about four to eight CPU cores per 1080Ti or 2080; a generally useful rule-of-thumb is that for similar simulation systems as the ones shown here, 10–15 “core-GHz” are sufficient and 15–20 “core-GHz” are also future-proof for upgrades (or better suited for workloads with additional CPU computation).

Additionally, PME offloading offers the possibility to cheaply upgrade GPU nodes once tailored for older GROMACS versions. By keeping everything but exchanging the old GPUs by state-of-the-art models, an optimal CPU/GPU balance can be restored for GROMACS 2018, at the comparatively small investment for GPUs only.

Outlook

Since hardware is continuously evolving and new components (CPUs, GPUs, barebones, boards, etc.) will become available in future, it is likely that configurations with an even higher P/P ratio than identified in this article will appear. Readers who have access to hardware configurations that were not covered here are encouraged to download our CC-licensed benchmark input files from <https://www.mpibpc.mpg.de/grubmueller/bench> to perform their own benchmarks such that we can include these data into updated versions of the Tables.

It is worth noting that the presented results do transfer very well to the GROMACS 2019 code, just released at the time of writing this article. On the performance front, this release has only modest additions with a few notable exceptions only. This release introduces PME offload support using OpenCL, which is particularly useful on AMD GPUs, especially in light of how favorably the (now previous-generation) Radeon Vega GPUs compare to the competition. Their advantage is particularly pronounced when comparing their P/P ratio against Tesla GPUs (see Figs. 3 and 4).

The additional feature of the 2019 release worth noting is the ability to offload (most) bonded interactions with CUDA. However, as GROMACS has highly optimized SIMD kernels for bonded interactions, this feature will have a beneficial performance impact only in cases where either available CPU resources are low or the simulation system contains a significant amount of bonded work. For our benchmarks, that would mean that the cases where just a few cores are paired with fast GPUs would be improved, as the 1–3 core range of the purple and dark green lines in Figure 8 indicate.

Acknowledgments

We thank Petra Kellers for thoroughly reading the manuscript; her suggestions led to numerous improvements. We thank the MPCDF, especially Markus Rampp and Hermann Lederer, for general help and for providing some of the hardware that has been benchmarked. This study was supported by the DFG priority programme Software for Exascale Computing (SPP 1648) and by the BioExcel CoE (www.bioexcel.eu), a project funded by the European Union contract H2020-EINFRA-2015-1-675728, the SSF Infrastructure Fellow programme, and the Swedish e-Science Research Centre (SeRC).

Keywords: molecular dynamics · GPU · parallel computing · energy efficiency · benchmark · GROMACS · computer simulations · CUDA · performance to price · high throughput MD

How to cite this article: C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B. L. de Groot, H. Grubmüller. *J. Comput. Chem.* **2019**, 00, 1–14. DOI: 10.1002/jcc.26011

- [1] L. Bock, C. Blau, G. Schröder, I. Davydov, N. Fischer, H. Stark, M. Rodnina, A. Vaiana, H. Grubmüller, *Nat. Struct. Mol. Biol.* **2013**, 20, 1390.
- [2] M. Zink, H. Grubmüller, *Biophys. J.* **2009**, 96, 1350.
- [3] M. Harvey, G. Giupponi, G. D. Fabritiis, *J. Chem. Theory Comput.* **2009**, 5, 1632.
- [4] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, R. C. Walker, *J. Chem. Theory Comput.* **2013**, 9, 3878.
- [5] B. R. Brooks, C. L. Brooks, III., A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Cafisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009**, 30, 1545.
- [6] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, D. E. Shaw, Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06): Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, Tampa, FL, **2006**, pp. 43-43. <https://ieeexplore.ieee.org/document/4090217>
- [7] W. M. Brown, A. Kohlmeyer, S. J. Plimpton, A. N. Tharrington, *Comput. Phys. Commun.* **2012**, 183, 449.
- [8] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, *J. Comput. Chem.* **2005**, 26, 1781.
- [9] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, V. S. Pande, *J. Chem. Theory Comput.* **2013**, 9, 461.
- [10] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, 1-2, 19.
- [11] M. Lundborg, C. L. Wennberg, A. Narangifard, E. Lindahl, L. Norlén, *J. Control. Release* **2018**, 283, 269.
- [12] J. D. Chodera, F. Noé, *Curr. Opin. Struct. Biol.* **2014**, 25, 135.
- [13] I. Buch, T. Giorgino, G. De Fabritiis, *PNAS* **2011**, 108, 10184.
- [14] N. Plattner, S. Doerr, G. De Fabritiis, F. Noé, *Nat. Chem.* **2017**, 9, 1005.
- [15] S. Chen, R. P. Wiewiora, F. Meng, N. Babault, A. Ma, W. Yu, K. Qian, H. Hu, H. Zou, J. Wang, S. Fan, G. Blum, F. Pittella-Silva, K. A. Beauchamp, W. Tempel, H. Jiang, K. Chen, R. Skene, Y. G. Zheng, P. J. Brown, J. Jin, C. Luo, J. D. Chodera, M. Luo, *bioRxiv* **2018**, 8, e45403.
- [16] C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B. de Groot, H. Grubmüller, *J. Comput. Chem.* **2015**, 36, 1990.
- [17] S. Páll, M. J. Abraham, C. Kutzner, B. Hess, E. Lindahl, In Lect. Notes Comput. Sci. 8759, EASC 2014; S. Markidis, E. Laure, Eds., Springer International Publishing, Switzerland, **2015**, p. 1.
- [18] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, *J. Chem. Theory Comput.* **2008**, 4, 435.
- [19] U. Essmann, L. Perera, M. Berkowitz, T. Darden, H. Lee, *J. Chem. Phys.* **1995**, 103, 8577.
- [20] B. L. de Groot, H. Grubmüller, *Science* **2001**, 294, 2353.
- [21] MARTINI. Coarse grain forcefield for biomolecules—Example application “lipid vesicles”. <http://www.cgmartini.nl/index.php/example-applications2/lipid-vesicles> Accessed May 20, 2019.
- [22] S. Páll, B. Hess, *Comput. Phys. Commun.* **2013**, 184, 2641.
- [23] F. Broquedis, J. Clet-Ortega, S. Moreaud, N. Furmento, B. Goglin, G. Mercier, S. Thibault, R. Namyst, hwloc: A generic framework for managing hardware affinities in HPC applications. Parallel, Distributed and Network-Based Processing (PDP), 2010, 18th Euromicro International Conference on. 2010; pp 180–186.
- [24] M. Frigo, S. G. Johnson, *Proc. IEEE* **2005**, 93, 216.
- [25] S. J. Marrink, A. H. De Vries, A. E. Mark, *J. Phys. Chem. B* **2004**, 108, 750.
- [26] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. De Vries, *J. Phys. Chem. B* **2007**, 111, 7812.
- [27] MARTINI. Coarse grain forcefield for biomolecules—example application “plasma membrane”. <http://www.cgmartini.nl/index.php/example-applications2/lipid-membranes>, Accessed May 5, 2019.
- [28] D. H. De Jong, S. Baoukina, H. I. Ingólfsson, S. J. Marrink, *Comput. Phys. Commun.* **2016**, 199, 1.
- [29] F. Benedetti, C. Loison, *Comput. Phys. Commun.* **2018**, 228, 146.

Received: 4 March 2019

Revised: 29 May 2019

Accepted: 4 June 2019