Developing Synthetic Applications Benchmarks on Composable Cyberinfrastructure: A Study of Scaling Molecular Dynamics Applications on GPUs

Richard Lawrence
rarensu@tamu.edu
Texas A&M University
High Performance Research Computing
College Station, TX, USA

Dhruva K. Chakravorty
chakravorty@tamu.edu
Texas A&M University
High Performance Research Computing
College Station, TX, USA

Francis Dang
francis@tamu.edu
Texas A&M University
High Performance Research Computing
College Station, TX, USA

Lisa M. Perez
perez@tamu.edu
Texas A&M University
High Performance Research Computing
College Station, TX, USA

Wesley Brashear
wbrashear@tamu.edu
Texas A&M University
High Performance Research Computing
College Station, TX, USA

Zhenhua He
happidence1@tamu.edu
Texas A&M University
High Performance Research Computing
College Station, TX, USA

Honggao Liu
honggao@tamu.edu
Texas A&M University
High Performance Research Computing
College Station, TX, USA

ABSTRACT

The potential for infinite scaling, improved performance, and better sharing of computing resources motivates researchers to adapt to composable infrastructures. Measuring performance on these systems requires an assessment of how the composed configuration itself affects performance which goes beyond traditional scaling approaches. We emphasize the subtle relationship between the nature of the calculation and the configuration of the composable infrastructure. New application benchmarking strategies are explored to inform the optimal configurations and best computing practices for composable systems. Realistic molecular dynamics research workflows are employed as benchmarks for composed GPU systems to develop a benchmarking strategy that yields recognizable and informative results. We employ the practices on a realistic case for a molecular dynamics research workflow on a composed GPU system. We discuss the identification of computational bottlenecks and establish a need for new benchmark performance suites that can help researchers articulate optimum compositions for composable infrastructure.

PEARC '23, July 23–27, 2023, Portland, OR, USA

ACM ISBN 978-1-4503-9985-2/23/07…$15.00
https://doi.org/10.1145/3569951.3597556

CCS CONCEPTS

• **Computing methodologies** → *Simulation support systems*; • **Hardware** → **Emerging architectures**.

KEYWORDS

High Performance Computing, Accelerators, GPUs, Composable System, Resource Dis-aggregation, Benchmarking, Molecular Dynamics, Cyberinfrastructure

ACM Reference Format:
Richard Lawrence, Dhruva K. Chakravorty, Francis Dang, Lisa M. Perez, Wesley Brashear, Zhenhua He, and Honggao Liu. 2023. Developing Synthetic Applications Benchmarks on Composable Cyberinfrastructure: A Study of Scaling Molecular Dynamics Applications on GPUs. In *Practice and Experience in Advanced Research Computing (PEARC '23), July 23–27, 2023, Portland, OR, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3569951.3597556

1 INTRODUCTION

Emerging technologies are transforming the architecture of traditional high performance computing (HPC) systems. Whereas the traditional model of HPC clusters comprise static distributions of computing resources, many newer systems are migrating towards dynamically composable infrastructure with disaggregated resources that can be configured based on user requirements. In a traditional compute cluster, simulations using a high GPU count can only be achieved by utilizing multiple host nodes. Communication between nodes presents an additional potential bottleneck,

which is eliminated by the composed node configuration. Hardware accelerators such as GPUs enable massive parallelization of structured computational work, but the fixed size of the accelerator itself forces the researcher to distribute across multiple accelerators to enable scaling up to solve the largest problems. This introduces additional layers of complexity; considerations include: limitations on problem size for a single accelerator, communication between accelerators, and a strategy of breaking the problem down for distribution. Composable infrastructure offers researchers a unique opportunity to choose the combination of node resources, including the number of GPU accelerators per node, that best meets the needs of their problem.

Scientific workflows span a complex performance landscape making it hard for researchers to realize the optimal composed configurations. In order to inform researchers how to best use composable cluster resources, it is important to choose realistic applications benchmarks that determine if a given configuration is working well for a given research workflow. However, existing benchmarks [6][7][8] characterize one hardware element of a system and do not capture the important effects that arise when using multiple accelerators in a reconfigurable cluster environment. In particular, performance depends on both details of the composable infrastructure and details of the computational workflow.

Previous and concurrent research that has been conducted on novel infrastructures follows a similar strategy, using machine learning workflows [1][2][4] and mesh solvers [3] as example research workflows.

2 METHODS

2.1 Molecular Dynamics with LAMMPS

LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) is a classical molecular dynamics software suite that has potentials for different many-body interactions. LAMMPS supports multiple frameworks for scaling onto cluster resources, including MPI, CUDA, and Kokkos and can orchestrate billions of atoms in memory. This popular simulation software suite is used on all major national computing platforms and therefore provides a benchmark that is of use to researchers.

The simulations performed in this work used a reproducible build of the LAMMPS physics engine published in the NVIDIA Container Repository [5][7]. Three established LAMMPS benchmark problems were chosen [8] to simulate individual atoms in different systems that explore the scaling of bonding and non-bonding force calculations. First, the Lennard-Jones (LJ) system is an unordered liquid of atoms interacting at short range. The short-range pair-wise interaction implies that minimal communication is needed. Next, the Embedded Atom Model (EAM) system is a lattice of metal atoms interacting with both neighbors and the long-range electric field. The long-range interaction requires additional operations and some communication between distant regions. Finally, the Rhodopsin system is a large protein molecule embedded within a lipid bilayer and surrounded by water and common ions. The bonding forces involve 3 or more atoms, greatly increasing the number of operations and need for communication.

Table 1: Replication factors and corresponding approximate atom count.

1x	2x	4x	8x	16x	32x
32 k	64 k	128 k	256 k	512 k	1 M

64x	128x	256x	1024x	2048x	4096x
2 M	4 M	8 M	33 M	66 M	131 M

Table 2: Single A100 GPU Scaling Study for three LAMMPS systems to determine the optimal number of atoms to effectively utilize a single GPU. The best performance found among replication factors for each system is highlighted red, reported in million atom-timesteps/s.

	4x	8x	16x	32x	64x	128x	256x	1024x
LJ		411	487	526	543	523	520	527
EAM		125	155	172	181	190	189	188
Rh.	7.6	8.6	9.0	7.2	6.6	6.3	6.3	6.3

2.2 Single-GPU Performance

Benchmarks were performed on NVIDIA A100 GPUs with 40GB of internal memory on the Grace and FASTER clusters hosted by Texas A&M High Performance Research Computing.

Data were collected over 3 independent runs. Each simulation was performed for 10,000 steps with a time step duration specific to each problem. Each run has 10,000 time steps in order to average out the warming up effects, which occur before time step 100. The measure of performance for a LAMMPS benchmark is given in units atom-timesteps/s, or alternatively, million atom-timesteps/s.

Each atomic system is described in terms of a unit cell of 32,000 atoms. Scaling the problem to larger system sizes is important because a realistic workflow puts a heavy load on the resources. In order to scale the problem up, the unit cells are replicated in 3D to fill a larger volume with even more atoms as shown in Figure 1. This strategy is needed to produce a benchmark that will reveal the correct computational bottlenecks. The replication factors are varied from 1 to 4096 as shown in Table 1.

Before benchmarking multiple GPUs, we establish an expectation of performance by finding the performance of a single GPU. In this section, scaling tests and observations are conducted to ensure that the GPU is optimally utilized by providing a large chunk of computational work to the GPU. This approach ensures that the GPU can perform work independently for some time, reducing the fraction of time spent in communication with the node. For the case of LAMMPS, this translates to having a large number of atoms; The results shown in Table 2 point to a minimum optimal load on a single GPU of 64x32000 atoms of LJ, 128x32000 atoms of EAM, or 16x32000 atoms of Rhodopsin.

2.3 Multi-GPU Performance

The Grace cluster is a traditional cluster in which each node is connected to a fixed number of accelerators. The FASTER cluster is

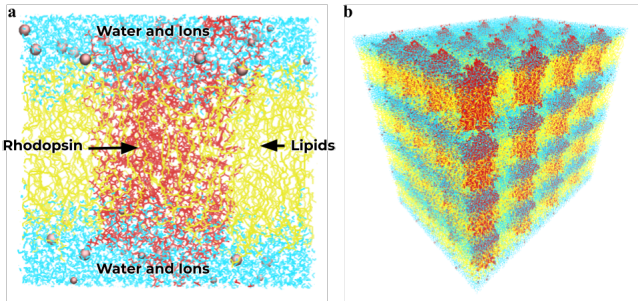

Figure 1: Visualization of Rhodopsin in lipid bilayer. (a) Single system. (b) System replicated in a 4x4x4 array, total replication factor 64. Note: this replicated system is not physically meaningful; it is merely an artificial workload for benchmarking.

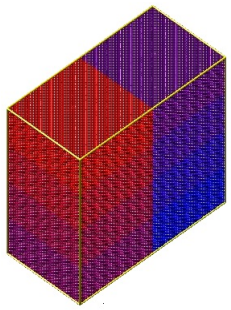

Figure 2: The partitioning for LJ system on 8 GPUs. LAMMPS created eight rectangular partitions, colored red to blue, arranged in two stacks of four. Each partition gets assigned to a different GPU. Communication between neighboring partitions could be a mechanism for the topology of the infrastructure to come into play.

a composable cluster that uses a software-defined PCIe-Gen4 Liqid fabric; within a server rack to compose nodes in configurations that extend beyond 16 accelerators per node [2].

In the Grace cluster, two A100 GPUs are connected directly to a node's PCIe-Gen4 slots. In the Liqid fabric of the FASTER cluster, multiple nodes and A100 GPUs are connected by longer PCIe cables to PCIe switches. Multiple PCIe switches are connected within a rack to form a web-like topology. Traditional benchmarks that measure connectivity between pairs of devices, such as NVIDIA Peer-to-Peer Bandwidth/Latency Test, are not aware of the topology, and report throughput and latency to be similar in both clusters.

Comparison of multi-node workflows on traditional and composed nodes, shown in Figure 3, provides evidence that the composed configuration reduces communication overhead and computation time by showing that concentrating multiple GPUs onto a single composed node improves performance. Given a fixed number of GPUs, those GPUs could be composed into nodes of varying size, either all on one node or across multiple nodes that are used together corresponding to a traditional system configuration. Although the traditional infrastructure and the composable infrastructure are comparable for multi-GPU workflows, a communication bottleneck occurs when using multiple nodes, which is eliminated by using a single composed node. Figure 2 visualizes how LAMMPS performs the division of work among eight GPUs for these types of workflows.

2.4 Description of Scaling Tests

In order to test the composable GPU infrastructure of the FASTER system, problems are distributed across multiple GPUs on composed nodes. Scaling tests are performed using each of three systems, replicated to various sizes, based on the expectation of proportional scaling relative to a single A100 GPU.

3 RESULTS AND DISCUSSIONS

3.1 Strong Scaling Across Multiple GPUs

Strong-scaling is the ability of a parallel computing system to efficiently solve a larger problem by increasing the number of processing units. Results are shown in Figure 4. All of these systems scale positively, but not proportionally. Some loss is associated with increasing GPU count. This is because the transition from one GPU to two or more GPUs introduces a communication delay mechanism that prevents perfect scaling relative to one GPU. The maximum performance achieved at high GPU count is around 80% of ideal for LJ, 70% of ideal for EAM, and 60% of ideal for Rhodopsin. The smallest system sizes scale noticeably worse than the larger system sizes. In addition, an unusual non-monotonic zig-zag pattern appears in the scaling for the Rhodopsin case at high GPU count. This pattern was not found to be correlated with the identity of the host node.

3.2 Weak Scaling Across Multiple GPUs

Weak-scaling refers to the ability of a parallel computing system to efficiently solve a larger problem by increasing the number of processing units and the size of the problem proportionally, while keeping the workload per GPU constant. Results are shown in Figure 5. All of these systems scale negatively, which shows that an increasing number of GPUs has an increasing communication overhead. The smallest system sizes of LJ and EAM are well below the others, which indicates that as expected, a minimum ratio of atoms per GPU is required to expect good performance. This reduces the relative fraction of the cost due to communication between the node and the GPUs. However, in the LJ and EAM cases, the minimum number of atoms per GPU at high GPU count is higher than the optimal result for the single-GPU case in Table 2; a multi-GPU run benefits from a greater number of atoms per GPU than a single GPU run.

4 CONCLUSIONS AND FUTURE WORK

The performance profile of molecular dynamics benchmarks for GPUs reveals an unexplained communication bottleneck in a composed Liqid fabric system. Investigating the relationship between the composable infrastructure's topology and the communication bottleneck will reveal more about the benefits and limitations of the composable infrastructure, and provide a guidance for researchers as to how to best use cluster resources. We conjecture that the unusual non-monotonic effect seen for Rhodopsin in Figure 4 is

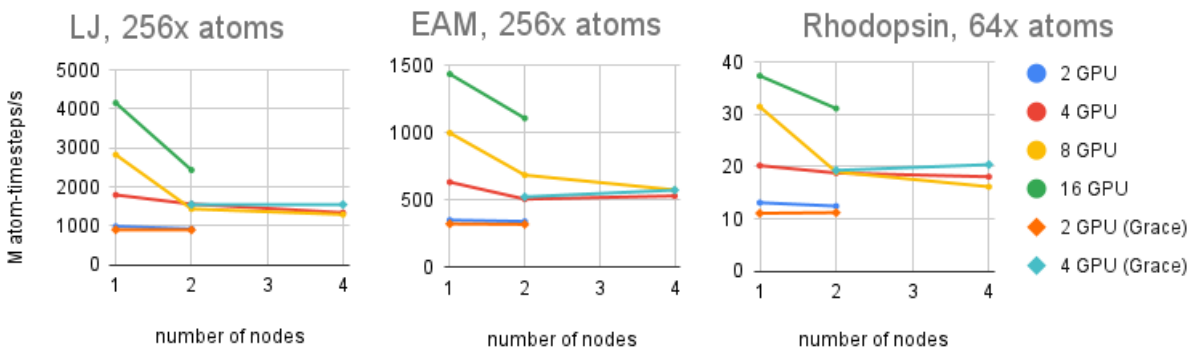

Figure 3: Performance of LAMMPS systems using multiple GPUs on one node or distributed across multiple nodes, using composed nodes on FASTER or traditional nodes on Grace. Performance is greatest when all GPUs are on a single node. The effect is more pronounced with larger numbers of GPUs; for eight or more GPUs the advantage of the single composed node strategy is dramatic. Traditional nodes on Grace perform similarly to composed nodes on FASTER at low GPU counts.

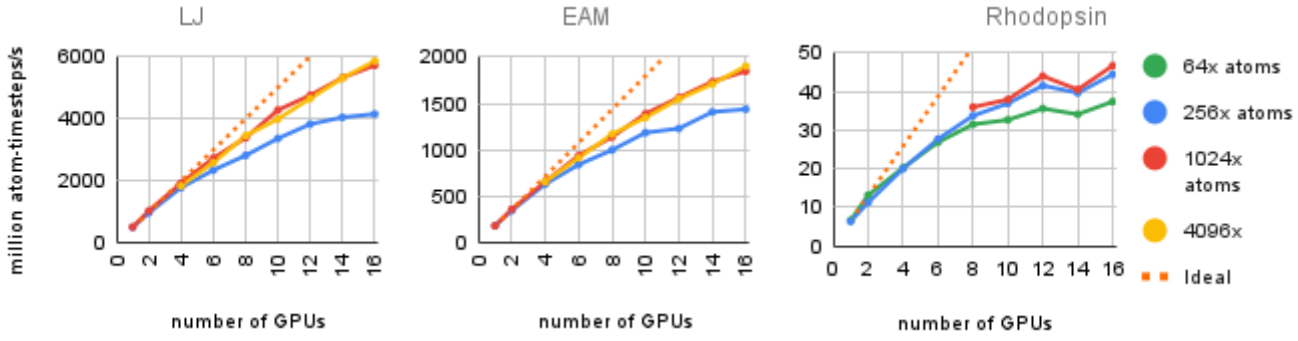

Figure 4: Strong scaling of three LAMMPS systems at three system sizes each on multiple NVIDIA A100 GPUs. The ideal line represents perfect proportional scaling relative to the one-GPU case.

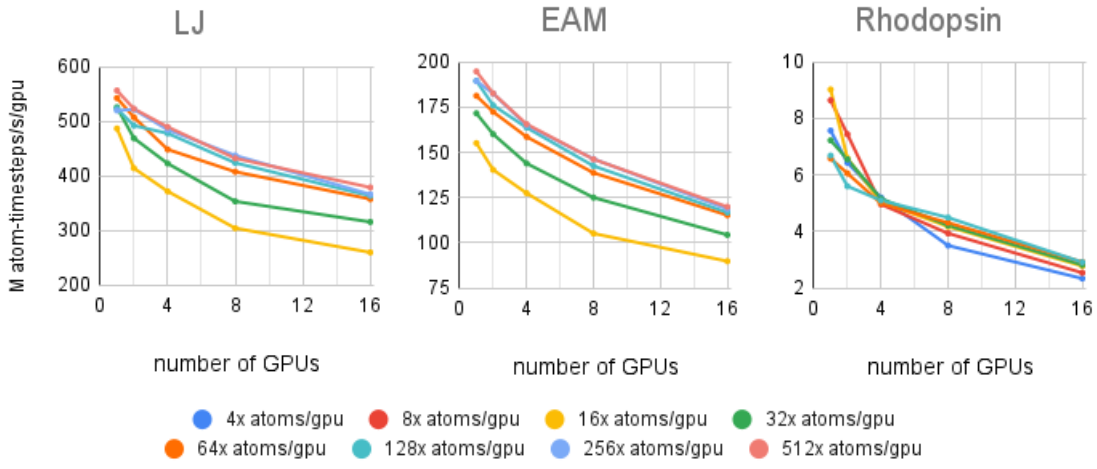

Figure 5: Weak scaling of three LAMMPS systems at three sizes each on multiple NVIDIA A100 GPUs.

related to the complex topology of the Liqid fabric. Additional work is needed to understand the precise mechanism at play; specifically, we will use NVIDIA throughput measurement tools to check the connection bandwidth between different nodes and GPUs within a fabric as the communication load on the fabric varies.

Measuring performance on composed architectures in general requires new benchmarking strategies because effects such as those described in this work are not captured by standard benchmarking tools which focus on individual hardware elements. New performance benchmarking strategies based on our findings will be used to inform the optimal configurations and best computing practices for composable systems.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) award number 2112356 ACES - Accelerating Computing for Emerging Sciences and NSF award number 2019129 FASTER - Fostering Accelerated Scientific Transformations, Education, and Research. Additional thanks to Dr. David Toback, Texas A&M University.

REFERENCES

[1] Kaoutar El Maghraoui, Lorraine M. Herger, Chekuri Choudary, Kim Tran, Todd Deshane, and David Hanson. 2021. Performance Analysis of Deep Learning Workloads on a Composable System. In *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 951–954. https://doi.org/10.1109/IPDPSW52791.2021.00147

[2] Zhenhua He, Aditi Saluja, Richard Lawrence, Dhruva K. Chakravorty, Francis Dang, Lisa M. Perez, and Honggao Liu. 2023. Performance of Distributed Deep Learning Workloads on a Composable Cyberinfrastructure. In *Practice and Experience in Advanced Research Computing* (Portland, OR, USA) *(PEARC '23)*. Association for Computing Machinery, New York, NY, USA, 12 pages. https://doi.org/10.1145/3569951.3603632

[3] Sambit Mishra, Freddie Witherden, Dhruva K. Chakravorty, Lisa M. Perez, and Francis Dang. 2023. Scaling Study of Flow Simulations on Composable Cyberinfrastructure. In *Practice and Experience in Advanced Research Computing* (Portland, OR, USA) *(PEARC '23)*. Association for Computing Machinery, New York, NY, USA, 6 pages. https://doi.org/10.1145/3569951.3597565

[4] Abhinand Nasari, Hieu Le, Richard Lawrence, Zhenhua He, Xin Yang, Mario Krell, Alex Tsyplikhin, Mahidhar Tatineni, Tim Cockerill, Lisa Perez, Dhruva Chakravorty, and Honggao Liu. 2022. Benchmarking the Performance of Accelerators on National Cyberinfrastructure Resources for Artificial Intelligence / Machine Learning Workloads. In *Practice and Experience in Advanced Research Computing* (Boston, MA, USA) *(PEARC '22)*. Association for Computing Machinery, New York, NY, USA, Article 19, 9 pages. https://doi.org/10.1145/3491418.3530772

[5] NVIDIA. 2021. *NVIDIA Container Registry: LAMMPS*. https://catalog.ngc.nvidia.com/orgs/hpc/containers/lammps tag:29Sep2021.

[6] NVIDIA. 2023. *NVIDIA Container Registry: HPC Benchmarks*. https://catalog.ngc.nvidia.com/orgs/nvidia/containers/hpc-benchmarks

[7] NVIDIA. 2023. *NVIDIA HPC Application Performance*. Retrieved April 2023 from https://developer.nvidia.com/hpc-application-performance

[8] Steve Plimpton, Aidan Thompson, Stan Moore, Axel Kohlmeyer, and Richard Berger. 2016. *LAMMPS Benchmarks*. Retrieved April 2023 from https://www.lammps.org/bench.html