Predicting Heterogeneity and Serverless Principles of Converged HPC, AI, and Workflows

Pedro Bruel, Sai Rahul Chalamalasetti, Aditya Dhakal, Eitan Frachtenberg, Ninad Hogade, Rolando Pablo Hong Enriquez, Alok Mishra, Dejan Milojicic, Pavana Prakash, and Gourav Rattihalli

Hewlett Packard Labs

Traditional HPC and modern AI computing are converging with workflows as a common paradigm. We predict nine principles of heterogeneity and serverless for this convergence, from high-level programming to low-level hardware.

High Performance Computing (HPC) is increasingly converging with AI, and both are increasingly expressed as workflows. Workflows enable a higher level of abstraction that is easier to develop, (re)use, and operate. Both HPC and AI depend heavily on accelerators, and they both adopt serverless computing. Similarly to workflows, serverless also raises the level of abstraction and simplifies DevOps [1]. In addition, it matches the fine granularity of accelerators in terms of time and size; they intuitively represent a good match in terms of performance and utilization.

When analyzing this convergence, three perspectives with different requirements and benefits exist:

1. End users care about latency or throughput of workflows at scale and ease/convenience of use.
2. Developers care about ease of development, e.g., constructing workflows from existing workloads and making QoS guarantees.
3. Providers (of services infrastructure) primarily care about meeting SLAs for user QoS and maximizing infrastructure utilization.

These three roles intertwine, and individuals could easily play two or even all three roles. For example, an end user of some services can be a provider to other users; a developer can conduct operations.

The principles and approaches we describe strive towards enabling seamless scalability and fluidity for end users; increased productivity of developers; and improved performance efficiency of providers.

To navigate these principles, we organized them in Figure 1, listed in order of description. Each principle enumerates possible benefits. The Figure should be read clockwise, starting with principle 1.

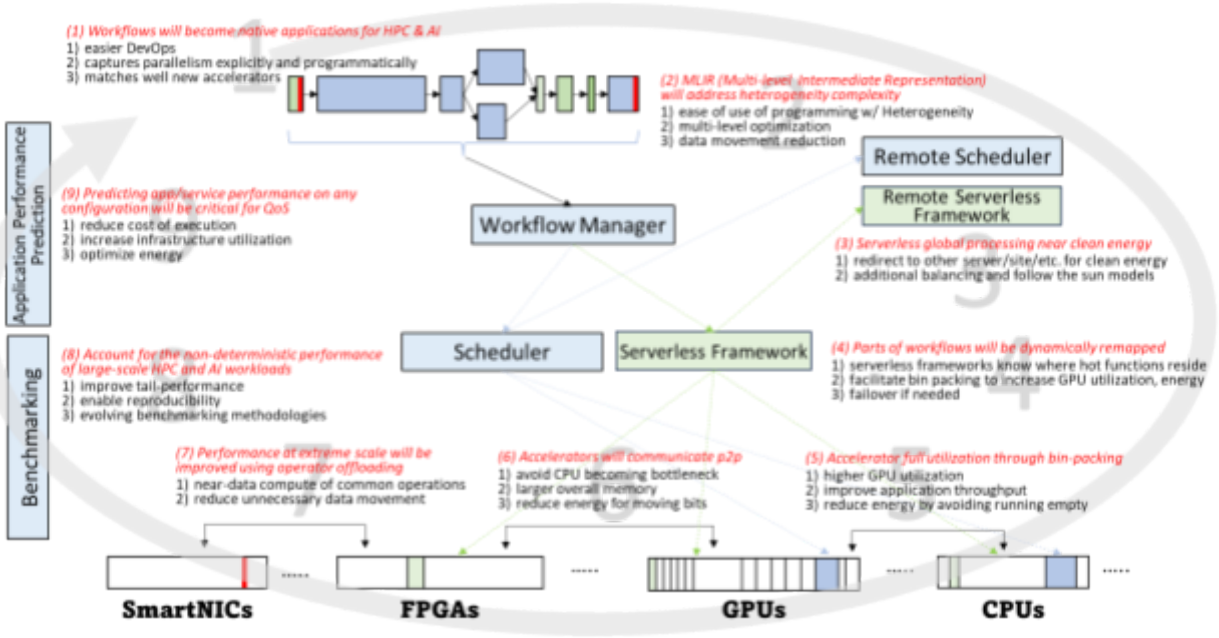

Figure 1. Landscape of Principles

(1) Workflows will become native applications for HPC & AI

With no intention of exploring the history of HPC and while focusing on the software evolution over hardware innovations, we would like to recall simpler times when calculations were performed on isolated, domain-specific problems by homogeneous CPU-based hardware. As the computational problems grew in complexity, they were typically broken into smaller tasks or files. However, it was still feasible then for a single programmer or small team to completely rewrite the code. Today's variety of domain-specific workflows in HPC pushed application scientists and engineers to propose various workflow-management systems with no clear standards or implementation patterns beyond a few workflow specifications and programming languages (Figure 2).

As individual applications become workflow-friendly by embedding CPU/GPU programming in single executables, it's becoming clearer that these partial solutions are unlikely to stay relevant, partially because the end of the Dennard-scaling era is leading to accelerator diversification. Creating truly native solutions for HPC will likely involve not only methodologies from multiple domains of knowledge at once, but also: (1) distributing computing tasks efficiently between emerging accelerators, (2) dynamically redeploying at scale, and (3) generating/testing workflows interactively using low-code programming models and simulations. The need for human experts to achieve next-level performance on these increasingly complex computational workflows will be complemented by machine learning and AI. These techniques will not only be

used within the workflows themselves but also as part of the HPC infrastructure. Future HPC is evolving towards harmoniously engineered workflows whose complexity can be abstracted while still performing optimally under flexible conditions.

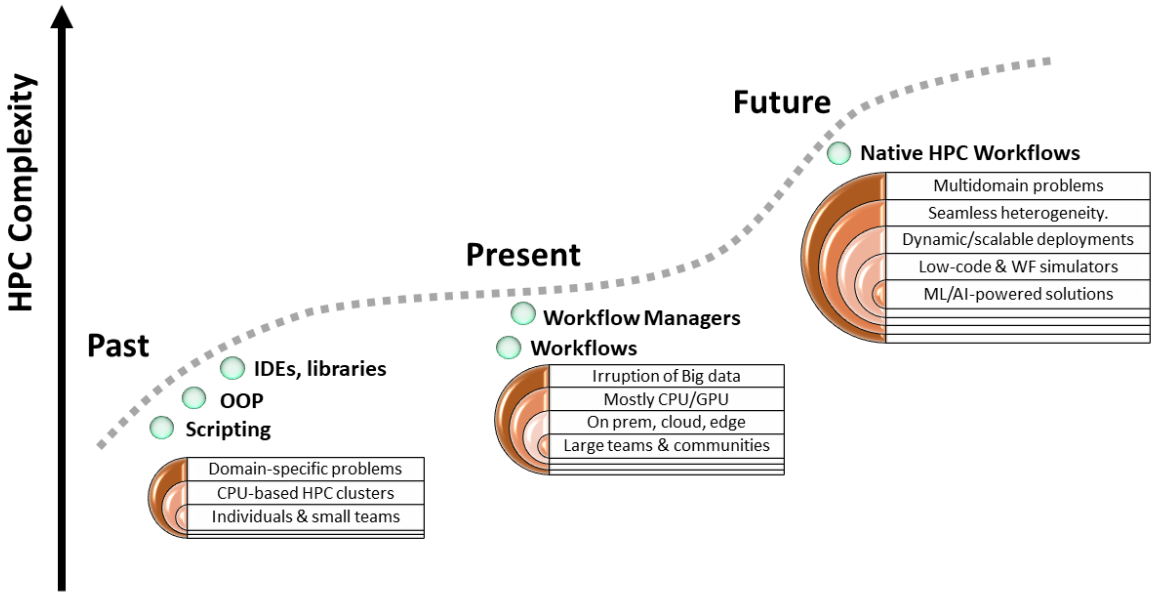

Figure 2. Evolution of Workflows

Optimally deploying these heterogeneous codes will require attention to how these codes are represented, as discussed next.

(2) MLIR will Address Heterogeneity Complexity

As applications become more workflow-centric, it becomes increasingly challenging for programming languages and compilers to ensure efficient task execution, unified representation, interoperability, and good abstraction. Multi-Level Intermediate Representation (MLIR) [2] is an open-source project initiated by LLVM to develop a new intermediate representation for compilers. It addresses some of the limitations of compilers that use traditional IRs (like LLVM-IR) by providing a more expressive, flexible, and efficient way to represent program structures in the front end. It also enables efficient compilation, optimization, and interoperability across diverse programming languages, domains, and hardware platforms, fostering innovation and collaboration in compiler design, AI, HPC, and more.

MLIR offers several advantages to HPC, AI, and workflow optimization. Its robust optimization capabilities, including high-level optimizations like data layout transformation, alongside low-level optimizations such as loop unrolling and vectorization, allow performance tuning of HPC applications. Its hardware-agnostic representation allows workflows to adapt effortlessly to heterogeneous architectures, like CPUs, GPUs, FPGAs, or future Quantum Computing accelerators, enabling customized optimization for various HPC configurations. It provides a

unified representation for complex workflows, including HPC and AI components, ensuring easy integration and optimization across multiple computing workloads in a single format. Its framework-specific dialects, which support TensorFlow, PyTorch, etc., ensure seamless model translation and integration across various components of AI processes, providing compatibility and ease of implementation.

Moreover, the flexibility and dynamic compilation capabilities of MLIR aid in sustainable workflow scheduling (discussed next) and enable dynamic deployment of HPC/AI workflows (principles #3 and #4), enabling real-time optimizations and efficient scaling of individual workloads.

(3) Serverless global processing near clean energy

In the drive toward a more sustainable digital realm, the potential of workflow scheduling emerges as a transformative force. Through the decomposition of workflows into smaller, manageable functions, an opportunity arises to strategically deploy these functions based on geographical and environmental considerations. Geo-distributed computational resources differ in their energy sourcing. Some benefit from renewable energy, while others draw power from carbon-intensive fossil fuels; some have faster computing capabilities, and vice versa. By leveraging workflow scheduling, functions can be located based on utilization, operating cost, and critically, environmental sustainability [3, 4]. This inherent flexibility allows functions within workflows to run in varied geographical regions, aiming to either minimize carbon emissions or optimize the total runtime of a workflow.

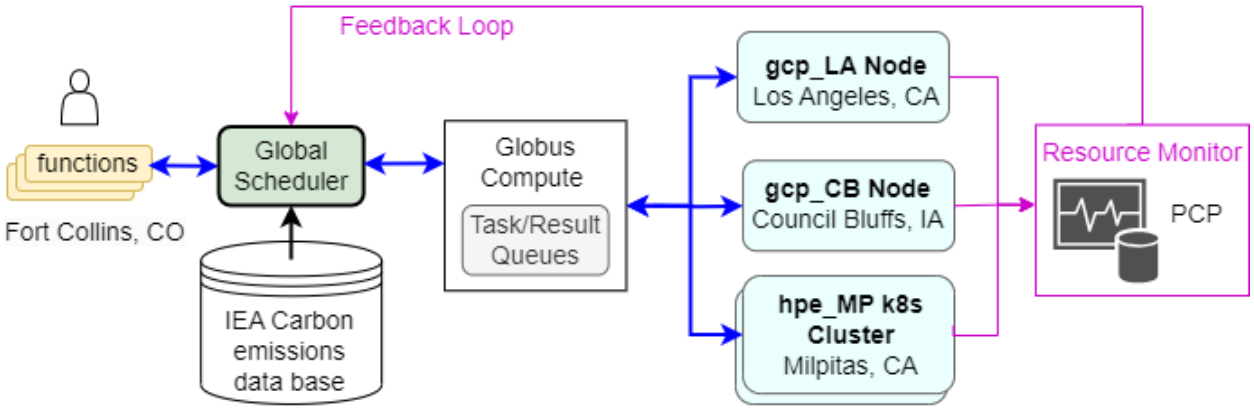

Figure 3: Carbon-aware scheduling of functions within a workflow

To validate this idea, we proposed a framework (shown in Figure 3) and conducted a series of experiments. We used Globus Compute [5] as the backbone of our framework to distribute functions across various geo-distributed compute resources. Our experimental setup encompassed a variety of computational resources: a Kubernetes cluster at HPE in Milpitas, CA, and two Google Cloud Platform servers, one in Los Angeles, CA, and another in Council Bluffs, IA. Each location had its distinct energy profile, influencing the function deployment

strategies. The functions originated in HPE's office in Fort Collins, CO. The global workflow scheduler was interfaced with Performance Co-Pilot (PCP) to monitor crucial metrics such as power consumption. Our experimental focus was the implementation of a carbon-aware scheduling policy aimed at minimizing carbon emissions in function deployment. Utilizing the proposed policy, the system was able to execute more functions (utilize more hardware) with lower carbon emissions.

Building upon the operational advantages of geo-distributed scheduling, we next explore dynamic redeployment, harnessing heterogeneous hardware to further optimize our global workflow framework.

(4) Parts of workflows will be dynamically deployed

With the increase in HPC and AI in computational research, the need to alleviate the bottleneck of statically deployed workflows grows urgent. Traditionally, workflows have been hosted on a fixed number of machines, resulting in resource underutilization. The growth of cloud-based virtual machines and bare-metal nodes enabled a game-changing solution: dynamic (re-)deployment of workflow components. This strategy ensures that specified components of a workflow operate on specialized hardware, maximizing utilization and decreasing workflow execution runtime [6]. These components can be dynamically re-deployed on specific hardware accelerators when they are available.

To validate this dynamic redeployment model, we used the GROMACS Lysozyme workflow [7]. We tried three different execution methods:

1. Serial Monolithic Execution (**monolith**): The traditional approach, executing the workflow as serial, monolithic tasks.
2. Decomposed but Heterogeneity-Oblivious Execution (**decomposed_NHHA**): The workflow was decomposed into its constituent functions and executed on a serverless platform without specific hardware considerations.
3. Decomposed and Heterogeneous-Hardware-Aware Execution (**decomposed_HHA**): Enhancing the second method, this technique dynamically redeployed the decomposed functions, assigning intensive tasks to specialized nodes with GPUs (when available).

Figure 4 details these execution techniques. Monolith and decomposed_NHHA are inefficient since they don’t utilize the specialized hardware, thus taking longer to run. While decomposed_NHHA has extended execution times due to container cold starts, the decomposed_HHA method significantly reduces runtime.

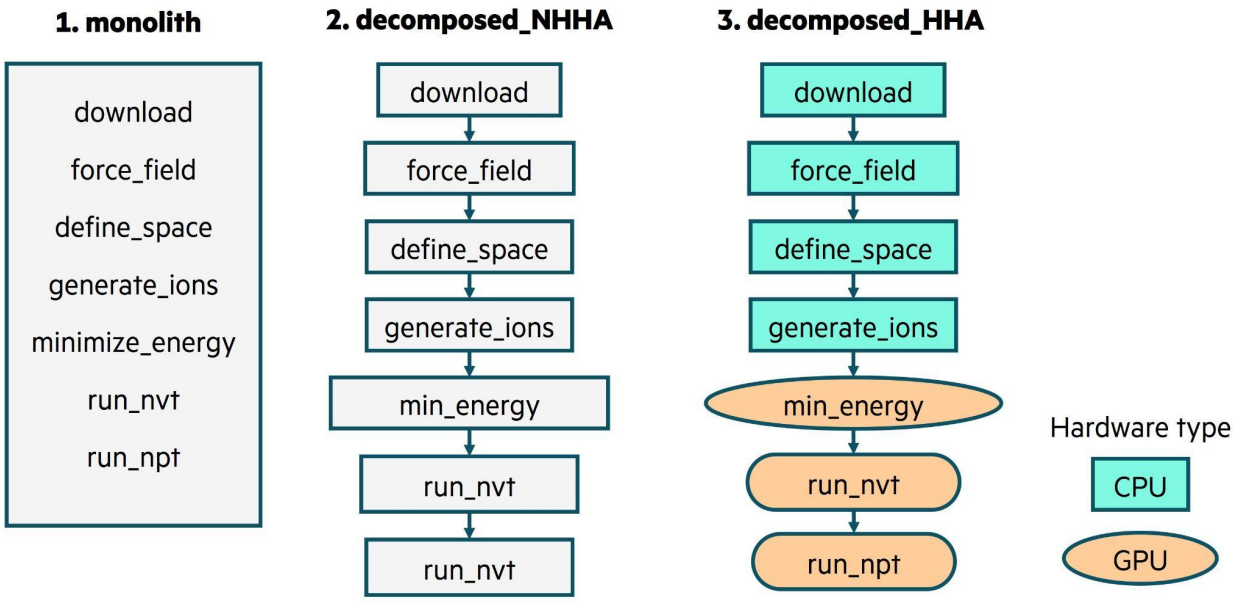

Figure 4: Three distinct execution techniques for GROMACS Lysozyme workflow

Dynamically redeploying workflow components optimizes task execution on suitable hardware, ensuring enhanced cluster utilization and minimized runtime. This shift towards dynamic deployment signifies a future of optimized resource allocation in computational research.

There is a huge demand for GPUs nowadays, shifting importance from user workload execution to maximizing GPU utilization, which leads us to the next principle.

(5) GPUs will be fully utilized through bin-packing

A workflow task might not saturate the entire GPU, so exploiting accelerator granularity could be increasingly important for HPC [8, 9]. To motivate finer accelerator granularity, we present an experiment where we ran the same nano-LAMMPS workflow with different GPU partition sizes. We picked a kernel that ran over 6000 times during the workflow execution and plotted its runtime with varying amounts of GPU compute (GPU%) in Figure 5(a). We can see the runtime of some runs (0-1000) improve when the GPU% gets higher, although not as much between kernels running at 50% GPU and 100% GPU. However, the runtime of almost all kernels hovers around 5 microseconds and does not change regardless of the GPU %. These measurements show that the kernels in these workflows do not require 100% GPU, and often 20% GPU suffices. Packing the GPU with multiple workflows, each getting a certain GPU% could increase the GPU throughput.

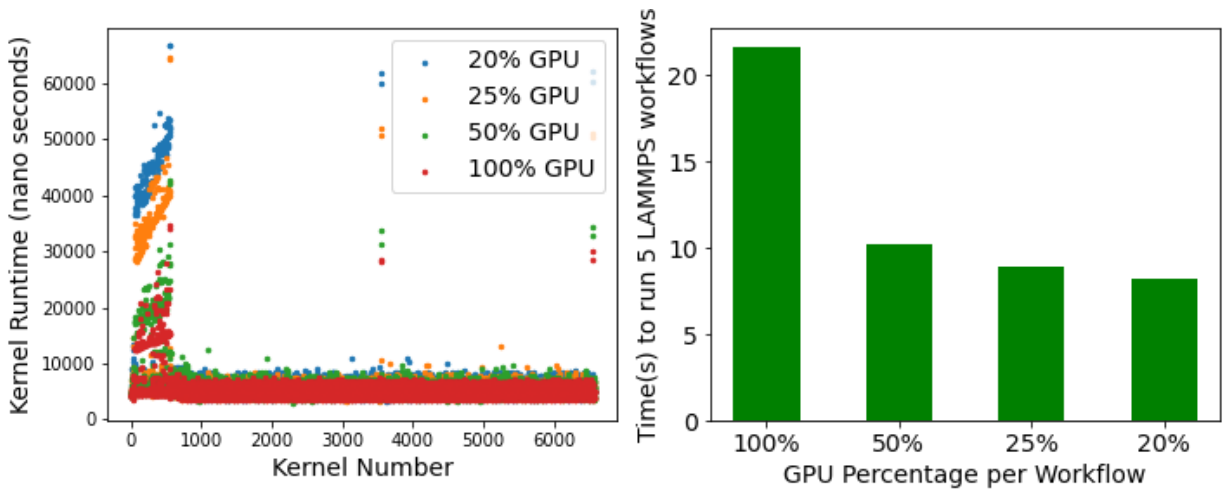

Figure 5 (a) Kernel runtime of workflow across different GPU% (b) Completion time of 5 workflows running concurrently

In another experiment, we look at the throughput of bin-packed GPU compared to where GPU is not multiplexed. In Figure 5(b), we present the time to run 5 LAMMPS workflows. Giving each workflow its own “bin” with 20% GPU completes running all workflows 60% faster than running each workflow individually with 100% GPU. This reduction in makespan stems from the increased throughput due to partitioning the GPU and running workflows concurrently.

Data transfer across computing elements (CPUs, GPUs, FPGAs, SmartNICs, etc.) is the slowest part of such workflows. Therefore, optimizing this communication is imperative. The next principle discusses optimizing performance with peer-to-peer (p2p) communications.

(6) Accelerators will communicate p2p

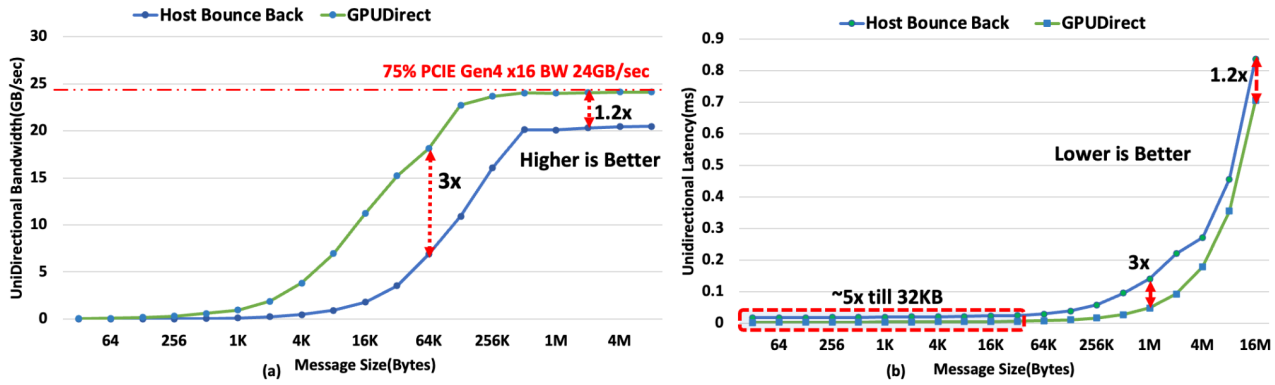

Figure 6. (a) Unidirectional inter-node GPU to GPU BW (b) Unidirectional inter-node GPU to GPU Latency

When accelerators consume the majority of the application's computation it is necessary to enable faster data movement to them, which currently uses the PCIe interface. However, the number of PCIe lanes at the CPU socket level limits the number of accelerators at a node level. Although a dual-socket-based server allows more accelerators per server, the NUMA-node connectivity interface for communication between accelerators across sockets can become a bottleneck. This is where p2p access to accelerators can help.

To illustrate this technique, we measured transfers from Mellanox InfiniBand (IB) 200Gbps NIC to Nvidia A100 GPU using GPUDirect [10], which uses PCIe p2p transfers. We compared it to a host bounce-back (host buffer as an intermediate data copy). We used OpenMPI 4.1.1 which supports GPUDirect transfer and point-to-point OSU latency and bandwidth benchmarks for the analysis [11].

Figure 6(a) shows that GPU-to-GPU device communication across nodes through IB NIC using GPUDirect/p2p exhibits 1.2x--3x higher bandwidth than host bounce-bank. Also, GPUDirect/p2p GPU device buffer transfer could saturate the PCIe interface to the practical bandwidth limit of 24GB/sec (75% of the theoretical PCIe Gen4 x16 bandwidth of 32GB/sec). In addition, GPUDirect/p2p (Figure 6(b)) also exhibits 1.2x--5x lower latency for message sizes below 8MB, but for message sizes larger than 16MB the latency of GPUDirect/p2p transfers converges to host bounce-back transfer latency.

With machine-learning training workloads that rely on GPU-to-GPU communication, higher p2p transfer bandwidth allows faster training speeds. For HPC workloads, most of the GPU-to-GPU communication would use small message sizes, so lower latency using p2p transfer will reduce overall application execution time.

In addition to GPU partitioning and p2p optimizations, the next principle introduces operator offloading as another important performance optimization.

(7) Operator offloading will enable performance at an extreme scale

In addition to hardware accelerators, distributed HPC and AI workflows rely on industry-standard libraries such as the Message Passing Interface (MPI) to effectively distribute, perform, and synchronize computation across interconnected machines. Common distributed operations, such as synchronizing data buffers in multiple machines via an aggregation operation, are encapsulated in MPI *collective communication routines*, or collectives. The operation *AllReduce* collective in MPI parlance is a fundamental operation of AI training workloads and also appears in many HPC workloads [12]. Driving these critical communication operations using the CPU or an accelerator such as the GPU consumes valuable computation and memory bandwidth, slowing down application performance [13]. Other than MPI collectives, operators such as data sorting, filtering, and encrypting/decrypting [14] are also ubiquitous in HPC and AI workflows. Since all of these critical and high-frequency operators depend on network communication, freeing CPU and GPU bandwidth by offloading operators to network

hardware such as Network Interface Controllers (NICs) and switches presents a great opportunity to improve the performance of HPC and AI workflows [15].

When pushing AI and HPC workflows toward extreme scales, for example, systems with tens or hundreds of thousands of GPUs, it becomes extremely hard to hide communication operations behind computation operations. Figure 7 presents the results of analytical simulations and modeling of distributed training of Convolutional Neural Networks (CNNs) and Large Language Models (LLMs) in extreme-scale systems. The figure highlights the large impact that driving communication has on GPU bandwidth, especially when increasing the message size of MPI collectives. We modeled analytically and simulated the use of AllReduce operations in AI workflows without NIC offloading (shown in green lines), with NIC offloading (shown in blue lines), and with NIC offloading plus full gradient caching (red lines). This last scenario is impractical in real hardware since the necessary cache size would be too expensive, but it can serve as a basis for comparison for increasing NIC cache sizes from the feasible scenario shown on the blue lines. We simulated real AI training workloads, from ResNet200 to GPT-4, highlighting the total AllReduce size involved in training each neural network. These simulations and models demonstrate that, after saturation, driving communication during training would leave only around 30% of GPU memory bandwidth free for computation, while offloading the AllReduce to a capable NIC would leave up to 87% GPU memory bandwidth free, which represents an expressive amount of computing power at extreme scales.

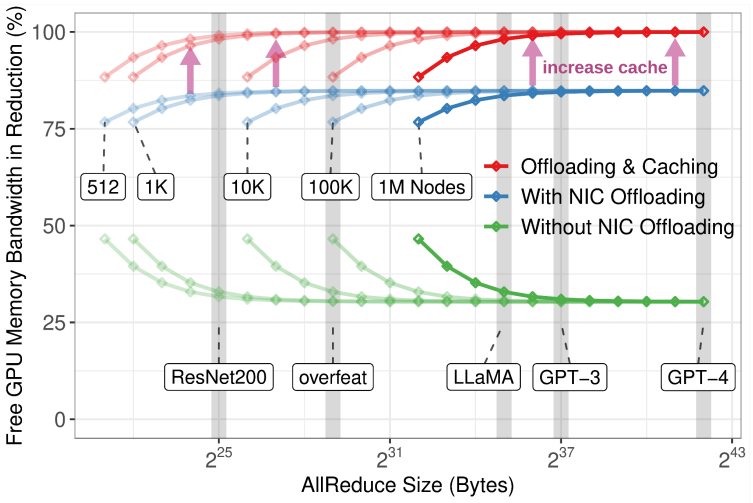

Figure 7: Analytical simulations and modeling of distributed training of CNNs and LLMs in extreme scale systems, highlighting the large impact that driving communication has on GPU bandwidth.

The seven optimization mechanisms and operation principles described so far emphasize heterogeneity in hardware in software, which complicates performance evaluation, as discussed next.

(8) Account for the non-deterministic performance of large-scale HPC and AI workloads

The combination of large-scale and heterogeneous software, middleware, and hardware means that every time we measure system performance we could be getting a different result (as shown for illustration in Figure 8). Some variability could be reduced or controlled, but likely not all of it. If the observed performance differences are relatively large and unpredictable, this non-deterministic behavior obfuscates the actual performance of the underlying system. It therefore becomes increasingly harder to answer critical business questions such as: does system A perform better than system B? What is the cost/performance of a system? Did its performance regress or improve?

The key to answering such questions is to handle performance like every other non-deterministic factor using statistical tools for distributions, similar to the research tools used in social and medical sciences. These tools can range from simple hypothesis testing and quantification of uncertainty to more advanced topics such as divergence metrics and causality analysis. Although these tools carry an implicit penalty, both in additional work and additional expertise, they also carry the promise of better performance reproducibility, correct interpretations, and actionable insights.

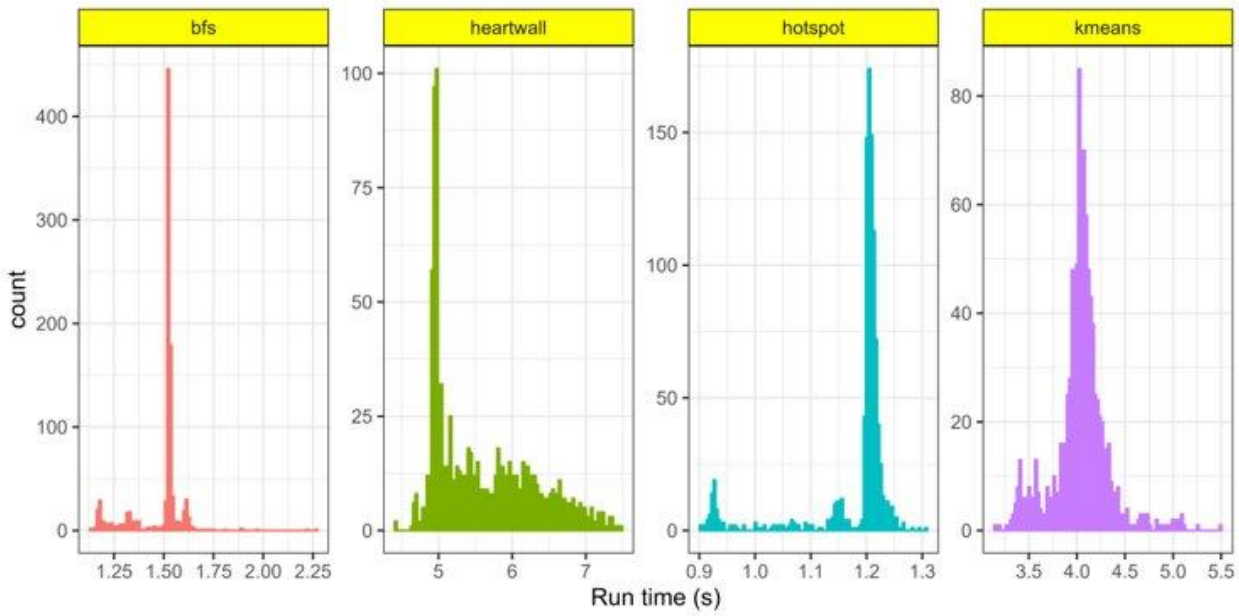

Figure 8: Performance histograms for four example applications from the Rodinia HPC benchmark suite, run 1,000 times each on a single machine with no interference. These applications represent, in clockwise order, distributions that are approximately near-constant, right-tailed, symmetric, and left-tailed.

(9) Predicting app/service performance on any configuration will be critical

The nondeterministic behavior of complex system performance brings to light the need for accurate estimates of performance and its distribution. For example, performance models of key applications have always been critical in the design and procurement of future computer systems [16], but these models often assume a homogeneous workload and architecture. As another example, when making scheduling decisions on a shared cluster, understanding the expected tail and outliers of the performance distribution of an application can impact the timing of its scheduling, to maintain service-level agreements for other jobs.

Together with collaborator Izzat el Hajj at the American University of Beirut, we have been developing performance prediction mechanisms based on a machine-learning model trained on low-level performance metrics. This model has been successfully demonstrated in tasks such as predicting the performance of known applications on new hardware configurations [17], which can be applied to the problem of selecting new hardware without benchmarking on all available choices. These techniques were also successful in predicting when applications are near the end of their execution, as a very useful prediction for supercomputer and mission-critical schedulers.

Summary

We presented nine principles of convergence of HPC, AI, and workflows. These end-to-end principles cover workflows, through middleware, to hardware. Nevertheless, there are many other missing aspects where this convergence can apply [18, 19, 20]. We did not even touch on nonfunctional aspects, such as security, reliability, scale, availability, etc. Each of these represents a considerable challenge but also an opportunity for improved usability, development, and delivery of converged HPC, AI, and workflows.

REFERENCES

1. Leonardo Leite, Carla Rocha, Fabio Kon, Dejan Milojicic, Paulo Meirelles, “A survey of DevOps concepts and challenges,” ACM Computing Surveys (CSUR), Vol. 52, Issue 6, Nov 2019, pp 1-35.
2. Lattner, Chris, et al. "MLIR: Scaling compiler infrastructure for domain-specific computation." 2021 IEEE/ACM Proceedings of CGO. IEEE, 2021.
3. S Qi, D Milojicic, C Bash, S Pasricha, “SHIELD: Sustainable Hybrid Evolutionary Learning Framework for Carbon, Wastewater, and Energy-Aware Data Center Management,” Proceedings of IEEE IGSC, best paper, 2023.
4. C Bash, N Hogade, D Milojicic, G Rattihalli, CD Patel, “Sustainability: Fundamentals-based approach to paying it forward,” IEEE Computer 56 (1), 125-132 2023.
5. “Computing with Globus,” https://www.globus.org/compute, accessed: 11-07-2023.
6. Gourav Rattihalli, Ninad Hogade, Aditya Dhakal, Eitan Frachtenberg, Rolando Pablo Hong Enriquez, Pedro Bruel, Alok Mishra, Dejan Milojicic, “Fine-Grained Heterogeneous Execution Framework with Energy Aware Scheduling,” Proceedings of IEEE CLOUD. 2023, pp 35-44.
7. Lysozyme in water,” http://www.mdtutorials.com/gmx/lysozyme/, accessed: 11-07-2023.

8. Pfandzelter, T., Dhakal, A., Frachtenberg, E., Chalamalasett, S., Emmot, D., Hogade, N., Hong Enriquez, P.R., Rattihalli, G., Bermbach, D., Milojicic, D., “Kernel-as-a-Service: A Serverless Programming Model for Heterogeneous Hardware Accelerators”, Proceedings of ACM Middleware 2023.
9. Aditya Dhakal, Philipp Raith, Logan Ward, Rolando P. Hong Enriquez, Gourav Rattihalli, Kyle Chard, Ian Foster, and Dejan Milojicic. “Fine-grained accelerator partitioning for Machine Learning and Scientific Computing in Function as a Service Platform”. In Proceedings of ROSS’23 at SC23, November 12–17, Denver, CO, USA. ACM, New York, NY, USA..
10. GPUDirect, “https://developer.nvidia.com/gpudirect”, accessed 11-10-2023.
11. OSU Benchmarks 5.8, “https://mvapich.cse.ohio-state.edu/benchmarks/”, accessed 11-10-2023.
12. Chunduri, Sudheer, et al. "Characterization of MPI usage on a production supercomputer." Proceedings of SC18 IEEE, 2018.
13. Rashidi, Saeed, et al. "Enabling compute-communication overlap in distributed deep learning training platforms." *2021 ACM/IEEE 48th ISCA*. IEEE, 2021.
14. Korolija, Dario, et al. "Farview: Disaggregated Memory with Operator Off-loading for Database Engines." *Proceedings of CIDR 2022*.
15. Torsten Hoefler, “General in-network processing - time is ripe!,” Keynote at the High-performance“Interconnects Forum in HPC China 2020, https://www.youtube.com/watch?v=t6jdjnnIRZs, accessed: 11-07-2023.
16. D. J. Kerbyson, H. J. Alme, A. Hoisie, F. Petrini, H. J. Wasserman, and M. Gittings. 2001. Predictive performance and scalability modeling of a large-scale application. In Proceedings of the 2001 ACM/IEEE SC '01.
17. Amir Nassereldine, Safaa Diab, Mohammed Baydoun, Kenneth Leach, Maxim Alt, Dejan Milojicic, Izzat El Hajj, “Predicting the Performance-Cost Trade-off of Applications Across Multiple Systems”, Proceedings of CCGrid, 2023.
18. Badia, R.M., Foster, I., Milojicic, “More Real Than Real: The Race to Simulate Everything,” IEEE Computer 2023.
19. Dube, N., Faraboschi, P., Milojicic, D., Roweth, D., “Internet of Workflows”, IEEE Internet Computing, Sept/Oct 2021
20. Milojicic, D. Faraboschi, P. Dube, N., Roweth, D., “Future of HPC: Diversifying heterogeneity,” 2021 Design, Automation & Exhibition (DATE), 2021.

Acknowledgments

We would like to thank our many collaborators who graciously contributed to different aspects of this work: Wen-mei Hwu and Deming Chen of UIUC, Gustavo Alonso of ETH, Izzat El Hajj of AUB, Ian Foster of the University of Chicago, Avi Mendelson of Technion, and Sudeep Pasricha of Colorado State. We would also like to thank Alex Qi, Alex Weaver, Hongzheng Tian, Kaiwen Cao, Kanchu Kiran, Liad Gerstman, Philipp Raith, Vijay Thurimella, and Viyom Mittal for contributing to some of the insights.

Authors

PEDRO BRUEL is a Scientist at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact him at bruel@hpe.com.

SAI RAHUL CHALAMALASETTI is a Senior Scientist at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact him at sai-rahul.chalamalasetti@hpe.com.

ADITYA DHAKAL is a Scientist at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact him at aditya.dhakal@hpe.com.

EITAN FRACHTENBERG is a Master Technologist at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact him at eitan.frachtenberg@hpe.com.

NINAD HOGADE is a Scientist at Hewlett Packard Labs, Fort Collins, Colorado, 80528, USA. Contact him at ninad.hogade@hpe.com.

ROLANDO PABLO HONG ENRIQUEZ is a Senior Scientist at Hewlett Packard Labs, Milpitas, California, 95035, UK. Contact him at rhong@hpe.com.

ALOK MISHRA is a Scientist at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact him at alok.mishra@hpe.com.

DEJAN MILOJICIC is an HPE Fellow and VP at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact him at dejan.milojicic@hpe.com.

PAVANA PRAKASH is a Scientist at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact her at prakash@hpe.com.

GOURAV RATTIHALLI is a Scientist at Hewlett Packard Labs, Milpitas, California, 95035, USA. Contact him at gourav.rattihalli@hpe.com.