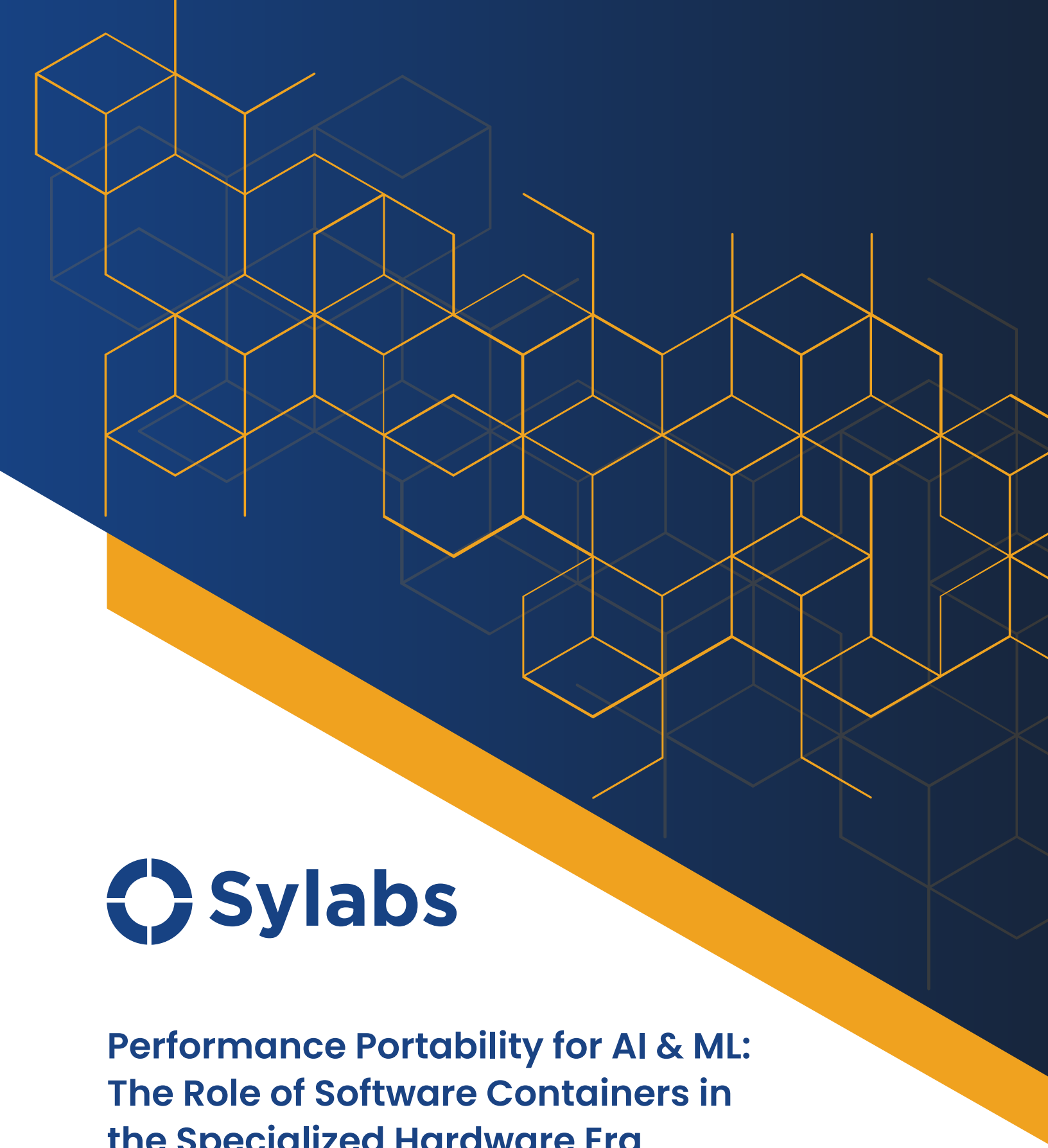

Sylabs

Performance Portability for AI & ML: The Role of Software Containers in the Specialized Hardware Era

Table of Contents

Introduction: The Growing Portability Challenge

The computing landscape is experiencing a significant transformation. Driven by rapid advancements in artificial intelligence (AI), machine learning (ML), and the burgeoning field of quantum computing, there is an ever-growing demand for faster, more efficient hardware. As the industry grapples with the impending limitations of Moore's Law (the observation that the number of transistors on a microchip doubles approximately every two years) and experience the waning benefits of Dennard Scaling (the principle suggesting that power consumption per transistor decreases as transistor density increases),[1] both emerging and established hardware vendors are adapting. They're increasingly focusing on designing specialized CPUs, GPUs, accelerators, and interconnects to address today's most challenging computational tasks. While this evolution promises to harness the true potential of software, it also introduces a challenging obstacle: software portability. At its core, performance portability is about ensuring that software maintains efficient performance across diverse computing platforms.

Software containers have emerged as a critical tool in navigating this challenge. Already widely used by scientists, researchers, academics, and IT professionals across many sectors, containers provide a reliable, efficient way to capture and bundle applications and their dependencies, facilitating their use across disparate computing environments. However, as more and more non-technical users turn to containers, container developers must grapple with the dual challenge of delivering performance-optimized solutions that are both secure and user friendly. The emphasis on security is especially important, especially given potential vulnerabilities in container workflows, related to user privileges and other factors, which much of the industry has been slow to address.

In this technical brief, we explore:

- The driving forces behind the pivot from generalized to specialized hardware, and its implications for the tech ecosystem
- The definition and increasing significance of performance portability, especially in the realms of AI and ML
- The proliferation of software containers as an important tool for bridging gaps between computing systems
- The Open Container Initiative (OCI) and its importance in helping to drive standardization and ensure interoperability among container solutions
- Singularity containers and what sets them apart in terms of support for performance portability initiatives

The Shift to Specialized Hardware

While the AI-driven revolution in chip making is not new, it has accelerated dramatically, driven by a constant need to continually increase the efficiency and speed of training and inferencing. The push for more specialized solutions for AI and ML gained significant momentum in the 2015 to 2016 timeframe when startups like Cerebras and Habana Labs were founded to build specialized processors from the ground up.[2] Around the same time, Google announced its custom Tensor Processing Units (TPUs) for machine learning,[3] followed by the launch of the Inferentia chip from Amazon Web Services (AWS) a few years later in 2019.[4]

While startups and tech giants were shaping one side of the landscape, industry powerhouses Intel and AMD were crafting their strategies to challenge Nvidia's stronghold in the GPU market. Their approach was twofold: strategic acquisitions and innovative in-house developments. Intel's 2019 acquisition of Habana Labs reflects its intent to cater to diverse AI requirements.[5] Further, with Intel's release of an AI-specific chip slated for 2025, the competition is heating up.[6] Not to be outdone, AMD has unveiled plans to release its advanced MI300X GPU, optimized for AI, by the close of 2023.[7]

Yet, it's not just about these familiar faces. Beyond the more well-known players, there are also a host of startups, such as Cerebras, SambaNova Systems, Graphcore, and Tenstorrent, that are also hard at work on innovative and potentially disruptive new solutions.

In a world where Nvidia GPUs, both from a hardware and software perspective, have long reigned supreme for AI tasks, and their demand often outstrips supply, these burgeoning alternatives aren't just welcome; they're essential. Yet, with this growing diversity in hardware options comes a set of unique challenges. With the influx of chips, we're also seeing a proliferation of architectures. Each of these architectures carries its own performance nuances, making it increasingly critical to understand and address the performance portability challenges that will arise for AI and ML applications.

Adressing a Critical Need

Performance portability is the ability of an application to run efficiently across multiple computing architectures, so it is closely related to hardware-agnostic performance and cross-platform efficiency.[8] The endgame is straightforward: simplify the process of transferring data and applications while amplifying their efficiency across platforms. However, the journey is full of challenges. An application optimized for one specific platform may not deliver equivalent performance or efficiency when ported to another without extensive, often costly, optimization efforts.[9] Given the high costs of optimizing applications for new systems, performance portability is especially important in the high performance computing (HPC), AI, and ML arenas. These are areas where specialized hardware is advancing rapidly, partially as a response to challenges presented by the approaching boundaries of Moore's Law for general-purpose chips.

Fortunately, as the need for performance portability has become more urgent, the industry and community have responded on a number of levels.

- **Research & development:** Leading research organizations and think tanks globally, are in the midst of crafting innovative methodologies tailored for performance portability. For example, in the U.S., the Department of Energy has been working on performance portability as part of its Exascale Computing Project (ECP) where it has emphasized creating and enhancing software libraries, programming models and tools to facilitate portability across supercomputing platforms.[10]

- **Programming models:** There's an increased interest in evolving programming models like OpenMP and OpenACC, which offer directives for shared-memory parallelism and accelerators. By providing high-level abstractions, these models can help in writing code that is both efficient and portable across architectures.[11]

- **Software abstraction layers:** Some initiatives are centered around creating software abstraction layers, which act as intermediaries, translating general code into hardware-specific instructions. Libraries like Kokkos and RAJA provide a uniform interface for developers while optimizing performance for various backends.

- **Benchmarking and profiling:** Benchmarking and profiling, which have always been critical aspects of software development in high-performance domains, could play an increasing role in performance portability. The idea is to understand how well a piece of software performs and where its bottlenecks are. Benchmarking can rely on standard tests or custom benchmarks tailored for specific use cases to gauge the performance of applications across different platforms. Profiling tools, like Intel's VTune or Nvidia's Nsight, enable developers to pinpoint inefficiencies deep within software, which is invaluable when moving software between architectures. By regularly benchmarking and profiling, developers can ensure that as they aim for performance portability, they are not sacrificing efficiency.

- **Containerization and virtualization:** Techniques like containerization are being used to package applications with their dependencies, ensuring consistent behavior across different environments. Virtual machines, too, play a role, abstracting the hardware and offering a uniform platform.

In light of these advancements, one thing is clear: the future of performance portability hinges on innovative approaches to application architecture and a firm commitment to agile software development practices, ensuring that as the hardware landscape evolves, software remains nimble and adaptive.

Bridging Gaps with Containers

Of the items mentioned above, containers are a quick win because they can deliver immediate benefits. At its core, a container is a standard unit of software that bundles up code and all its dependencies, ensuring that the application runs quickly and reliably across different computing environments. For example, they can even provide an easy bridge between the development of an application on a laptop and deployment to a favorite cluster. By providing an isolated environment, containers ensure that software can be reliably and securely transferred across different computing platforms. This makes them especially valuable in light of the rise of specialized hardware. Here are five reasons why containers are, and will remain, a key tool for performance portability in the "new normal" of specialized hardware architectures.

- **Uniform application behavior:** One of the primary challenges with specialized hardware is ensuring consistent application behavior across platforms. Among container technologies, Singularity from Sylabs stands out with its use of the Singularity Image Format (SIF)—a portable, container image format designed specifically for performance and security. By encapsulating an application along with its dependencies in the SIF format, Singularity ensures that the application behaves consistently, whether running on a generic CPU, a custom AI accelerator, or an advanced GPU.

- **Optimized resource usage:** Traditional virtualization methods can be resource-heavy. In contrast, containers are lightweight and designed for compatibility and performance, ensuring maximum performance can be squeezed out of the specialized hardware without unnecessary overhead.

- **Agility in deployment:** With the diverse range of specialized hardware entering the market, the ability to quickly deploy, test, and iterate is invaluable. Containers support rapid versioning and, if necessary, easy rollback, allowing teams to address performance anomalies as they adapt to new architectures.

- **Embracing microservices:** Specialized hardware often excels at specific tasks. By adopting a mix of batch and microservices and relying on containerization, organizations can optimize each segment for specific hardware capabilities to ensure that the full potential of a specialized chip is harnessed.

- **Integrated management tools:** Many leading container options include tools optimized for monitoring, verifying container reproducibility, and secure multi-tenancy, which are essential when deploying across specialized hardware platformsverifying container reproducibility, and secure multi-tenancy, which are essential when deploying across specialized hardware platforms.

Ultimately, by ensuring consistency, optimizing resources, and providing a collaborative platform, containers help organizations navigate the complexities and harness the power of evolving hardware. The challenge for technology developers is to continually make life easier for those building and using containers for everything from performance portability to ensuring robust security measures, seamless interoperability across platforms, and accommodating the ever-evolving requirements of modern computational workloads.

Open Industry Standards: The Power of OCI

In a landscape where software portability is crucial, the same rings true for container interoperability. Recognizing this imperative, industry leaders established the Open Container Initiative (OCI). Designed with an open governance structure, the OCI's mission is to develop open industry standards for container formats and runtimes. The goal is to ensure that containers behave consistently, irrespective of the environment or platform they're deployed on. As the hardware arena becomes increasingly diverse, such standardization isn't just beneficial—it's indispensable for laying the groundwork for more seamless performance portability.

Sylabs and Singularity: Leading the Charge for Performance-Portability-Optimized Containers

At Sylabs, our mission is to make HPC accessible to researchers, scientists, and engineers by providing the most advanced container technologies for performance-intensive applications. From this mission, Sylabs, along with our open-source community, has built a container platform that is ideally suited for performance portability in the AI era and beyond.

Singularity was first developed as an open source project at Lawrence Berkeley National Laboratory to run complex applications on HPC clusters in a simple, portable, and reproducible way and quickly became popular at other HPC sites, academic sites, and beyond.

Central to Sylabs Singularity container technology is the Singularity Image Format (SIF). This format empowers users to easily share and distribute Singularity container files across diverse computing environments without breaking. SIF images enforce a unique security model that enables untrusted users to safely run untrusted containers on multi-tenant resources. Given its utility, the broader industry has adopted the SIF format, which today includes integrations from Red Hat (Podman), and Apptainer. Anchore's Syft (software bill of materials [SBOM]) and Grype (vulnerability scanning for containers) open source tools also support SIF.

Today, the portfolio of Sylabs Singularity tools, including SingularityCE, SingularityPRO, and Singularity Enterprise, is the predominant container technology in shared HPC environments, which speaks to its value in terms of portability, performance, security, and more. For example, leading research organizations turn to Singularity for its:

- **Strong security model:** Singularity uses a unique security model to mitigate privilege escalation risks.
- **All-in-one application environment:** Singularity provides a platform to capture a complete application environment into a single file.
- **Universal support:** Singularity supports all the major distributions, as well as multiple architectures.

But the innovation doesn't stop there. At its core, Sylabs is committed to enhancing container workflows. The company's forward-looking approach encompasses the integration of features like SBOM and crucial extensions for improved compatibility with OCI-based containers, especially within microservices environments utilizing Docker, Podman, and Kubernetes.

Overall, Sylabs' holistic approach ensures Singularity's versatility across complex computing ecosystems.

Seamless OCI Interoperability with Singularity

When it comes to OCI compatibility, most other container providers provide workarounds rather than seamless interoperability between OCI containers regardless of a container's origin. Singularity can use an OCI image as a source to run or create a container image, which is then used to deploy across compute resources, whether on-premise, in the cloud or in hybrid computing environments. This conversion process facilitates a fluid integration, allowing users to consistently integrate technologies while preserving their investments.

Singularity also delivers SIF Encapsulated OCI Images Support, a feature that enables users to seamlessly integrate OCI data and configurations into SIF containers, minimizing the need for complex tooling while optimizing security and performance advantages from SIF.

5 reasons to consider Singularity

- **Performance portability:** Singularity is designed with diverse computing environments in mind, providing easy access to GPUs and Infiniband. It ensures that applications consistently run at peak efficiency, irrespective of the hardware they're deployed on. This capability is especially vital for AI and ML workloads where the choice of hardware can profoundly influence performance.
- **Interoperability with OCI standards:** While Singularity offers a unique value proposition, Sylabs and the Singularity community are closely attuned to industry standards. Its compatibility with OCI specifications ensures it integrates seamlessly with other container tools and orchestration platforms.
- **Built for multi-tenant environments:** Singularity is engineered to function efficiently in multi-tenant settings where multiple users or teams share computing resources. Its robust security and isolation features ensure that each tenant's applications are compartmentalized, reducing the risk of unauthorized access or interference. This makes Singularity a scalable solution for enterprises or research institutions that require both high performance and rigorous security.
- **Optimized for HPC and compute-intensive tasks:** Unlike many other container solutions designed primarily for microservices and web applications, Singularity is optimized for HPC. Its design principles take into account the specific needs of HPC, making it a preferred choice for compute-intensive applications. Moreover, it provides a proven bridge between merging cloud-native and HPC workloads.
- **Secure by design:** Singularity employs an innovative security model that prioritizes isolating sensitive resources from non-privileged users in shared environments.

A More Efficent Way Forward

As the hardware landscape continues its rapid diversification, ensuring software portability will be pivotal in optimizing AI and ML workloads across specialized hardware and emerging architectures. Achieving peak performance portability requires a multifaceted approach, but container solutions, particularly those like Sylabs Singularity, offer a robust means of addressing the complexity. Looking ahead, the significance of container technology will undoubtedly amplify, as developers delve deeper into refining and innovating these platforms to meet ever-evolving needs.

About Sylabs

Sylabs is the global leader in providing professional tools and services for high performance container runtime technology. Sylabs makes high performance computing more accessible to researchers, scientists, and engineers using Singularity, the most advanced open source container runtime technology for performance-intensive applications and environments. As the most active contributor to the Singularity ecosystem, through both the community edition and Sylabs' enterprise-supported and professional implementations, Sylabs is dedicated to enabling cutting-edge research and facilitating rapid scientific discovery to solve some of humanity's greatest challenges.

For more information about Singularity runtime technology, including SingularityCE (Community Edition), Singularity Container Services, SingularityPRO, and Singularity Enterprise, visit https://www.sylabs.io.

Sources:

[1] Moore's Law Is Dead. Where Is Energy Saving Heading in the Electronic Information Industry?, Light Reading, 2022.

[2] Specialized Hardware for AI: Rethinking Assumptions and Implications for the Future, Gradient Flow, February 2023.

[3] Tensor Processing Units: Both History and Applications, Medium, April 2019.

[4] A first look at AWS Inferentia, Medium, December 2019.

[5] Intel Acquires Artificial Intelligence Chip Maker Habana Labs, Business Wire, December 2019.

[6] Intel gives details on future AI chips as it shifts strategy, Reuters, May 2023.

[7] AMD reveals new A.I. chip to challenge Nvidia's dominance, CNBC, June 2023.

[8]Beyond Speed: Why Performance Portability is Increasingly Crucial for HPC and Enterprise Evolution, EETimes, September 2023.

[9] Performance Portability, ScienceDirect, 2023.

[10] Performance Portability in the Exascale Computing Project: Exploration Through a Panel Series, IEEE Computer Society, 2021.

[11] Ami Marowka. 2022. On the Performance Portability of OpenACC, OpenMP, Kokkos and RAJA. In International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia2022), June 03–05, 2018, Virtual Event, Japan. ACM, New York, NY, USA 12 Pages. https://doi.org/10.1145/3492805.3492806 (https://doi.org/10.1145/3492805.3492806)

Sylabs

Deploying Performance Intensive Workloads Easily and Securely

Contact us today

For help choosing the best container solution for your application, contact Sylabs today to chat with one of our Solution Specialists!

Phone: (530) 428-5606 | Email: info@sylabs.io