White Paper

An Alternative Approach for Intelligent Acceleration of HPC and AI Workloads

Jaclyn Ludema and Mark Nossokoff
Sponsored by NextSilicon
February 2025

HYPERION RESEARCH OPINION

In the rapidly evolving landscape of technology, high-performance computing (HPC) and artificial intelligence (AI) are becoming increasingly intertwined, redefining how organizations approach data processing and analysis. As industries strive to harness the power of AI for scientific research, there is a growing trend toward integrating AI capabilities within traditional/existing HPC frameworks. This convergence is driven by the desire for more efficient methodologies to tackle complex scientific computing challenges.

Organizations are seeking innovative ways to leverage AI in scientific use cases, recognizing its potential to enhance data analysis, improve predictive modeling, and accelerate discovery processes. However, many users are also contending with the limitations of their existing architectures. The skyrocketing costs associated with new hardware acquisitions and lengthy procurement processes create significant barriers to upgrading systems. This situation can foster a fear-of-missing-out among users, as they worry about the adequacy of their current setups in meeting the demands of cutting-edge research and computational workloads.

As a result, there is a desire within the HPC community for solutions that can maximize the performance and utility of existing applications while minimizing investment into costly application porting and code optimization. Organizations are looking for flexible and adaptable systems that can evolve alongside their workloads, allowing them to extract greater value from their investments. Addressing this need is essential for organizations to remain competitive in an increasingly data-driven world.

In this context, NextSilicon's Maverick-2 offers a promising solution, providing a pathway to intelligent acceleration of HPC and AI workloads. With an innovative architecture that emphasizes flexibility, efficiency, and adaptability, Maverick-2 stands poised to meet the evolving demands of the market, enabling users to harness the full potential of their computational resources without the costs and complexities typically associated with system upgrades. This paper will explore the current challenges in HPC and AI, the potential solutions to these challenges, and their implications for the future of computational research.

SITUATION ANALYSIS

The landscape of HPC and AI is undergoing a significant transformation, driven by the need for more powerful, efficient, and adaptable computing solutions. New breakthroughs are requiring orders of magnitude increases in computing power, data, and electricity. As organizations increasingly rely on these technologies to extract information from vast amounts of data, they face several challenges associated with current architectures and systems.

Challenges and Issues with Current Architectures

Limitations of Existing Architectures

Many existing computing architectures, particularly those based on traditional CPUs and GPUs, are iterations of designs that have evolved over the past eight decades. These systems are primarily rooted in the Von Neumann architecture, which, while groundbreaking eighty years ago, is now reaching its limits in delivering efficient performance at scale for HPC. CPUs offer flexibility but lack throughput, while GPUs excel at parallel processing but struggle with sequential tasks, branching operations, and global memory access. These architectural limitations make it challenging to efficiently handle the complex workloads required by modern AI and HPC applications.

Shift Toward AI-Focused Designs

There is a noticeable shift in the design focus of processors and coprocessors, moving away from purely HPC-oriented architectures toward those optimized for AI workloads, spurred by both financial and technical motivations. This trend is particularly evident in recent processor and accelerator roadmaps, which increasingly prioritize features like reduced precision calculations and specialized matrix operations that primarily benefit AI applications. This transition reflects a growing customer market opportunity for AI and the need for processing a variety of models tailored to AI-specific computations such as parallelism and specialized data handling.

As a result, organizations are seeking hardware that can efficiently manage the increasing data volumes and complexity associated with AI applications. However, this AI-centric focus leaves HPC customers questioning where their unique needs fit in, as they face challenges that differ from those of AI and require solutions that balance high precision, reliability, and sustained performance across a diverse set of applications. The growing emphasis on AI optimization features raises important considerations for HPC users, whose applications often cannot trade numerical precision for increased throughput.

Challenges with Current GPU Ecosystems

Current GPU ecosystems present significant challenges for users. The closed software frameworks tied to many leading GPU products can limit flexibility and innovation for some use cases. These frameworks often require proprietary tools and domain specific languages (DSLs), which can create a software moat that complicates the adoption of alternative solutions. These constraints make it challenging for developers to optimize applications for specific workloads, resulting in inefficiencies that hinder progress and drive up operational costs.

Datacenter Deployment Challenges

Deploying new data centers or upgrading existing ones has become increasingly complex. Data center operators must integrate diverse computing resources with varying specifications and requirements while keeping pace with rapidly evolving technology. The growing prominence of GPUs in AI and HPC

workloads presents particular challenges, as some operators struggle to manage power consumption and cooling systems effectively. For example, the power per rack has leaped by an order of magnitude, from 10 to 15 KW racks to over 150 KW racks. These challenges can directly impact data center capacity and operating costs. In some instances, GPU operational costs may exceed the overall budget of an acquisition. Additionally, the pressure to scale operations quickly while maintaining performance can lead to suboptimal configurations, further exacerbating the inefficiencies of traditional architectures.

Underserved Customer Requirements

As HPC and AI workloads converge, particularly in scientific research and national laboratory settings, users are finding that their needs are not being met by existing solutions. There is an urgent demand for systems that facilitate this convergence, allowing for faster and more accurate scientific outcomes.

Customers are seeking easy-to-adopt software that minimizes the learning curve and enables quick integration into their workflows. Additionally, there is a growing desire to avoid vendor lock-in, which can stifle innovation and limit options. Users need infrastructure that is not only powerful but also flexible enough to adapt to both current and future applications. As seen with the rapid evolution of approaches -- from AI neural networks to transformer-based models -- customers are increasingly interested in infrastructure that is future-proofed for these types of innovations.

Value of Addressing These Challenges

By rethinking traditional modeling and simulation workloads through innovative architectures, organizations can unlock substantial performance improvements. Enhanced efficiency in processing workloads will not only lead to more rapid scientific discoveries but also provide a competitive edge across various sectors.

The current HPC and AI landscape has a pressing need for solutions that can overcome the limitations of existing architectures. As organizations grapple with the constraints of traditional systems, the demand for flexible, efficient, and adaptable computing resources grows stronger. By addressing these challenges, the potential for transformative advancements in both HPC and AI becomes increasingly attainable.

POSSIBLE SOLUTIONS

To address the pressing challenges faced by the HPC and AI markets, a variety of innovative solutions should be explored. These approaches aim to rethink existing architectures and methodologies, paving the way for more efficient and effective computing systems that can meet the diverse needs of modern workloads. By focusing on flexibility, specialization, and adaptability, these solutions seek to empower organizations to fully leverage their computational resources and enhance their research and operational capabilities.

- Installing Multiple, Smaller Application-Focused Architectures: Developing a range of smaller, specialized systems tailored to specific application needs can significantly enhance performance. These systems can optimize resource allocation for particular tasks, reducing overhead and improving efficiency.
- Processor and GPU Redesign: Creating processors and GPUs that focus specifically on certain traditional modeling and simulation workloads can bridge the gap between HPC and AI

requirements. These types of redesigns can enable faster execution of critical algorithms and improve the overall throughput of scientific computations.

- **Flexible System Designs:** Designing systems that can adapt to various HPC workloads, using scalable and reconfigurable architectures, is crucial. This flexibility allows organizations to dynamically allocate resources based on current demands, ensuring that they can respond quickly to changing computational needs without incurring excessive costs.
- **Alternative Approaches:** Exploring other innovative designs, such as neuromorphic computing or quantum computing, could further enhance computational capabilities. These approaches can provide new paradigms for processing data, potentially delivering breakthroughs that traditional architectures cannot achieve on certain application types.

These proposed solutions highlight a critical shift in computing strategies, emphasizing the need to balance specialization with adaptability to meet the complex demands of modern workloads. The following summary outlines how these approaches aim to address challenges in HPC and AI, while considering their broader implications for the future of computing.

Application-focused architectures enable organizations to achieve significant performance gains by tailoring resources to specific workloads. This approach can optimize efficiency without extensive system overhauls but may limit versatility for broader use cases. Redesigning processors and GPUs to cater to specific workloads addresses the limitations of current architectures; however, this approach incurs high R&D costs and reduces general-purpose flexibility. Transitioning from general-purpose designs to fixed-use architectures could hinder adaptation to future workloads.

Flexible system designs, in contrast, prioritize the increasing need for adaptability by allowing dynamic reconfiguration to meet evolving computational demands. This offers scalability as organizations face evolving demands and keep pace with the needs of ever-changing workloads that are becoming a critical factor in maintaining competitiveness. Finally, exploring alternative computing paradigms opens up exciting possibilities for the future, potentially revolutionizing how complex calculations are performed and paving the way for next-generation applications.

Collectively, these solutions present a comprehensive set of strategies for overcoming the limitations of existing architectures and meeting the growing demands for efficient performance and adaptability of the HPC and AI markets. Organizations must carefully assess their priorities to determine the right balance of these options to deliver computing solutions best suited to their needs.

THE NEXTSILICON APPROACH

Founded in 2018 in Tel Aviv, Israel, by Elad Raz and a team of industry veterans, NextSilicon has a vision to redefine the future of HPC for science and engineering. NextSilicon's solutions leverage intelligent, software-defined processing to address the limitations of traditional computing architectures, particularly those of current CPUs and GPUs. Below is an overview of the NextSilicon approach, detailing its current solutions, technical innovations, and success stories.

Maverick-2 – Powered by the Intelligent Compute Architecture (ICA)

Maverick-2 ICA is NextSilicon's second generation accelerator based on its Intelligent Compute Architecture (ICA). It is designed to provide adaptable and scalable processing capabilities across the computing spectrum. This approach allows organizations to maximize the potential of their workloads while minimizing the complexities typically associated with transitioning to new architectures.

- Maverick-1, the first generation of NextSilicon's solution, served as a collaborative development platform with leading commercial and research HPC labs starting in 2022. Over the last several years, this joint development paved the way for the innovations introduced in Maverick-2.
- Maverick-2 introduces an innovative, reconfigurable software-defined hardware architecture that dynamically adapts to changing workload demands. NextSilicon's approach leverages intelligent software, allowing developers to bring their own code and quickly achieve performance without major rewrites or lengthy porting processes.

To support developers, Maverick-2 provides a comprehensive SDK that includes automated optimization tools and native support for C/C++, FORTRAN, OpenMP, and Kokkos. Future plans also include integration with CUDA, HIP/ROCm, and popular AI frameworks like PyTorch.

Unlike applications that require porting from DSLs, Maverick-2 users start by compiling and running their existing code on the host system. NextSilicon's management software then identifies the flows within the application that account for the bulk of the runtime. Referred to as the "likely flow", patterns in the application are identified by gathering telemetry from the host such as the number of function calls, control flow patterns, etc. These likely flows are transformed into a compute graph to be executed in silicon —- which NextSilicon calls a "mill core" — that is then replicated many times across the device. During runtime, Maverick-2 utilizes patented algorithms and application telemetry to continuously reconfigure accelerator performance to the specific needs of each application.

Referred to as the "likely flow", patterns in the application are identified by gathering telemetry … then transformed into a compute graph to be executed in silicon —- which NextSilicon calls a "mill core" — that is then replicated many times across the device.

As a result of this approach, the Maverick-2 - powered by its intelligent compute architecture - can achieve exceptional performance through its innovative dataflow design, with internal testing showing up to 4x higher performance per watt compared to traditional GPU solutions. The Maverick-2 architecture employs a distributed memory hierarchy that provides industry-leading random access memory bandwidth, enabling efficient handling of large-scale datasets. This design significantly reduces data movement overhead, a common bottleneck in conventional architectures. The platform's dynamic resource allocation system continuously monitors workload characteristics and adjusts computational resources in real-time, ensuring optimal utilization across different application types. This adaptability is particularly valuable for organizations running mixed HPC and AI workloads, since it eliminates the need to separate specialized hardware for different tasks.

Technical Innovations

NextSilicon's Maverick-2 intelligent compute accelerator incorporates two key technical innovations:

1. The automatic optimization engine provides real-time workload analysis and optimization, dynamic resource allocation based on computational demands, intelligent power management and thermal optimization, and automated workload partitioning and scheduling. Leveraging knowledge of the runtime workload ensures optimal use of silicon area, enhancing both peak flops and energy efficiency beyond what traditional architectures can deliver.

2. The mill core's architecture enables parallel processing of multiple data streams, advanced pipeline management, efficient handling of irregular computational patterns, and reduced memory access latency through intelligent prefetching.

The Maverick-2 architecture supports delivering optimal computational performance by intelligently adapting to workload demands, maximizing silicon efficiency, and reducing potential bottlenecks. Additionally, it allows for software-based performance upgrades to both its intelligent algorithms and developer toolchain, ensuring a future-proof and adaptable solution.

Success Stories

NextSilicon has established collaborations with prominent institutions, including the US Department of Energy Sandia National Laboratory. The partnership with Sandia National Laboratory is particularly significant, as it aims to enhance the National Nuclear Security Administration's Advanced Simulation and Computing program.

The Maverick-2 will allow for dynamic reconfiguration based on workload demands, improving performance and power efficiency in real time. This innovative approach is set to strengthen the capabilities of Sandia and its associated laboratories, and positions NextSilicon to take on future large-scale production systems.

FUTURE OUTLOOK

The HPC-AI market is experiencing significant growth, driven by increasing demand for energy-efficient computing solutions, rising adoption of AI in scientific research, the growing need for flexible, scalable computing architectures, and increased emphasis on sustainable computing solutions. NextSilicon is positioning itself to capture a portion of this expanding market through its innovative solutions and strategic partnerships.

The adoption of NextSilicon's technology has the potential to drive significant changes across multiple sectors. In research and academia, this technology is designed to accelerate scientific discovery through improved computational capabilities, enhancing the ability to process and analyze large datasets, reducing barriers to entry for complex computational research, and facilitating collaboration through standardized tools and platforms.

Commercial applications can benefit from a reduced total cost of ownership for AI and high-performance computing while improving energy efficiency and sustainability. This can enhance an organization's competitive advantage through faster time-to-solution, and greater flexibility in resource allocation.

The environmental and societal impact of these advancements could be significant in many areas. The improved energy efficiency can contribute to reduced carbon footprints in computing facilities, while the enhanced computational capabilities can accelerate climate change research and modeling. Additionally, these technological improvements can support advanced medical research and drug discovery efforts, leading to improved scientific understanding through enhanced simulation capabilities.

The convergence of HPC and AI through solutions like Maverick-2 represents a potential fundamental shift in computational capabilities. As organizations continue to seek more efficient and powerful computing solutions, NextSilicon's innovative approach positions them at the forefront of this

transformation. The company's commitment to advancing both hardware and software capabilities, combined with its focus on practical implementation and user-friendly tools, suggests a promising future for both the technology and its applications across various industries.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
