I/ONX

Scaling AI with Heterogeneous Compute

June, 2025
v1.7

Heterogeneous Compute:

Sustainable Scaling for AI

Author: Justyn Hornor

Artificial Intelligence as a field is still in its infancy, and the hardware being used to build these products is far from ready to scale to meet consumer demand. While the GPU has given the field a jumpstart into proving out how AI can solve real-world problems, we propose a far more sustainable approach to scaling up to meet the market demand: tailored heterogeneous compute for AI training, inference, and other high performance compute workloads.

Compute platforms with the right processor mix can dramatically reduce power consumption, increase speed, reduce physical footprint, and create entirely new ways to train AI products. The I/ONX approach uses state-of-the-art, enterprise-grade processors from a range of fabricators in a fully connected, rack-scale solution — ready for the largest workloads at scale. These rack-scale solutions can be tailored for specific applications.

We provide a working definition of heterogeneous compute, some historical context, direct benefits of power consumption savings, immediate challenges, and current state of the I/ONX platform.

Version 1.7 updates:

- Updated Current State
- Noted 3rd Party Verification via Slalom

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

Introduction

In the transportation and logistics industry, it takes a number of types of vehicles and infrastructure to efficiently transport goods at a global scale. Cranes, cargo ships, trains, semi-trucks, planes, warehouses, and box vans — to name a few — are necessary. A similar analogy can be found on a football field: it takes a quarterback, linemen, running backs, receivers, and more to move the ball down the field.

These are examples of heterogeneous systems where the right vehicle (or player) is assigned the correct task to most efficiently accomplish the broader goals of that system. Logistics firms would never use box trucks for transporting goods across the continent when trains or planes would be far more efficient. And you would not put a wide receiver at center and expect the play to end well.

In the context of computing, we have had heterogeneous platforms for decades. I/ONX has applied these principles to the specific target of training artificial intelligence systems with high-powered processors in large systems that enable those processors to efficiently communicate with each other. Like the logistics and sports team analogies, the right processor for the right algorithms produces a far more efficient and effective system. The whole is greater than the sum of its parts.

In this paper we discuss several classes of processors: CPU, GPU, ASIC, and FPGA. In reality, the I/ONX platform leverages many other processor types, but we group these into the four classes for the purposes of explaining their more general use cases. Processors like the TPU, VPU, DPU, and others can be grouped as forms of ASICs, for example.

Defining Heterogeneous Compute

Briefly, heterogeneous compute (HC) is any system using more than one class of processors to solve a problem. This is not a new approach, but the proposed scale and types of processors in the I/ONX platform is novel. The most common HC you will find in the wild is the CPU and GPU running gaming or computational-heavy workloads. One of the defining characteristics of an HC platform is that it is built to solve specific problem sets.

The closest analogy to the I/ONX platform is Apple's System-on-Chip (SoC) architecture in their M-class systems. These combine multiple processors to support the typical workloads on Mac computers - CPU for applications, GPU for graphics, and NPU for AI workloads. This is an over-simplification of the design, but provides some context for one application of HC where several classes of processors are fully connected for best-in-class general compute performance (Clover, 2024).

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

Outside of gaming and generalized compute, other forms of HC exist in the context of edge devices (Samsung Exynos, 2024), networking (Intel Manitoba, 2003), numerical processing with the Cydrome Cydra-5 (Schlansker, 2005), and others. These are purpose-built platforms designed to solve specific problems.

No such HC architecture exists today for AI training — until now. I/ONX has built the first-of-its-kind HC platform specifically to enable sustainable AI training at massive scale. While a few firms have proposed simpler versions of HC in mid-2025, I/ONX is the first to demonstrate the ability to scale across all four key processor classes.

Benefits of HC for AI Training

Artificial intelligence is an umbrella term for a range of domains where various algorithms are applied to create automated decision-making. The key domains are (Keserer, 2024):

- Machine learning
- Deep learning
- Robotics
- Neural networks
- Natural language processing

Each of these domains require significantly different algorithms and datasets for training their respective models. Yet the field at large continues to use a rudimentary approach to running these algorithms at scale with CPU- and GPU-based platforms. To solve the problem of scale, data centers network many CPU and GPU systems together, further compounding the problem of energy consumption while creating new problems, such as heat generation (Burton, 2024).

The problem of power consumption cannot be understated. Scaling these rudimentary systems up while the fabricators focus on higher-wattage chipsets is creating an exponential, catastrophic crisis in communities. Beyond massive power requirements, water for cooling is another critical resource extracted from communities for operating data centers. These facilities are expected to be one of the largest sources of concentrated power consumption globally in the near future with recent trends in AI further exacerbating these issues (Hoosain, Paul, et al, 2022) (Judge, 2023).

While GPUs are incredible processors that enable parallel processing for simple operations, many different processor classes can operate other necessary algorithms far more efficiently. In the below table we describe several common algorithmic domains and various classes of processors in the I/ONX stack:

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

APPLICATION	CPU	GPU	ASIC	FPGA
General Purpose Computing	Versatile, handles a wide range of tasks efficiently but not specialized	Not optimal for serial tasks, better for data parallelism	🟡 **Not versatile, designed for specific tasks only**	Moderately versatile, can be reconfigured for specific tasks but less efficient than ASICs
Parallel Processing Tasks	Less efficient due to limited cores and parallelism	Highly efficient due to many cores optimized for parallel processing	🟡 **Not typically used, lacks flexibility for varied parallel tasks**	Efficient for parallel processing but usually less so than GPUs due to lower core count
Deep Learning Inference	🟡 **Moderately efficient but less so than GPU or ASICs for large models**	Very efficient, commonly used for training and inference in deep learning	Extremely efficient when designed for specific models or inference tasks	Efficient if optimized, but usually slower than ASICs or GPUs for deep learning inference
Cryptographic Hashing	Moderate efficiency, suitable for small-scale tasks	🟡 **Less efficient, not specialized for hashing tasks**	Most efficient, often used for cryptocurrency mining	Highly efficient for cryptographic tasks when specifically configured
Scientific Simulations	🟡 **Moderately efficient but limited by fewer cores compared to GPUs**	Highly efficient for tasks involving large-scale parallel computation	Highly efficient only if designed specifically for the task	Efficient for scientific simulations, especially when customized for specific algorithms
Matrix Multiplications	Less efficient due to lower parallelism	Extremely efficient due to parallel processing power	Can be extremely efficient if customized for matrix operations	🟡 **Moderate efficiency, can be customized for matrix operations but less efficient than GPUs or ASICs**
Low-Latency Tasks	Highly efficient, suited for low-latency, serial tasks	🟡 **Less efficient due to high latency in switching tasks**	🟡 **Not efficient for tasks requiring flexibility or frequent changes**	Moderately efficient, good for tasks that benefit from reconfigurability and low latency
High Throughput Tasks	🟡 **Moderate throughput, general-purpose processing**	High throughput, suitable for parallelizable tasks	Highest throughput but only for designed tasks	High throughput if reconfigured for specific tasks but generally lower than ASICs for fixed functions

Table 1: Application and suitability based on processor class. Highlighted cells indicate a use case where the processor performs poorly relative to other processor types.

As can be seen, no one processor solves all common algorithmic problems efficiently. This is where HC comes into focus as a platform with all the requisite processors needed to support a wide range of AI-based training problems while also providing best-in-class performance.

In our own testing, I/ONX has found three distinct benefits to an HC platform configured for AI training:

- Reduced power consumption
- Improved speed
- Advanced adaptability

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

Before addressing each, research from academia supports our findings in simpler use cases. Kuan-Chieh and Hung-Wei (2023) reported a 2x improvement in power consumption and 2x improvement in speed when combining a GPU with a TPU (Tensor Processing Unit, a form of ASIC). This approach, called SHMT (Simultaneous Heterogeneous Multi-Threading), leans heavily on software for effectively breaking down an algorithm and operating the required computations on the processor suited to the task. The major differences between the SHMT approach and I/ONX are the size of processors and that I/ONX adds additional processor types to support more advanced use cases required by the data science community.

In December 2024, AWS announced a new class of compute instances: Trainium 2 and Trainium 2 UltraServer (Barr, 2024). The I/ONX platform can be configured to match the specifications of this new class in a single-rack configuration for on-prem AI training workloads.

Reduced Power Consumption

The most power-hungry processors in the typical AI training stack are going to be GPUs. These processors are incredibly efficient for common AI training tasks, such as matrix multiplication. Yet many other types of algorithms are necessary where other processor classes are far more efficient than a GPU (See Table 1).

When an algorithm is offloaded to a processor that consumes less power, the overall system power drops. For example, given that a GPU requires 1,000 watts and an ASIC requires 150 watts, an algorithm that is split up to operate half the functionality on the GPU and the other half on the ASIC will reduce the overall power consumption to 575 watts. This is overly simplistic, but it explains how such a dramatic reduction in power consumption works in a heterogeneous-compute architecture.

Improved Speed

The same principle applies when a given processor is able to operate an algorithm more efficiently. ASICs will always outperform other processor types when running the algorithm they were designed to operate. If that ASIC is twice as fast, then the overall system gets that speed improvement when that processor is used.

The Tensor Processing Unit (TPU) is a form of ASIC that has been proven to be far more efficient at specific AI inference tasks, such as convolutional neural networks (CNN) (Jouppi, Young, Patil, et al, 2017). When applied to the very specific use cases, the TPU was able to outperform the GPU more than 20x for speed while also reducing power consumption (Jouppi, et al, 2017).

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

But because the TPU is purpose-built to solve a specific algorithm, it cannot be used for other applications. As processor manufacturers bring new products to market, you must be aware of the specific benefits, as many will claim they solve all the problems in the AI-training field.

Advanced Adaptability

A further benefit of an HC platform is the advanced adaptation that can be applied during a given workload to improve on customer-defined outcomes. For example, if an AI model must have extremely precise inference, the algorithms can be tuned to use higher precision (for instance, Floating Point 32 will be considerably higher precision than Floating Point 8), and the workload can be modified to operate on the most appropriate processors.

Should the customer need a low-power solution, ASICS and FPGA processors can be used more heavily than GPUs. While this may slow down the speed, the requirement for the power consumption is met without the need for downgrading the entire system. Customers with variable power are excellent use cases where the system may need to adapt in real time.

Another example includes being able to incorporate novel processors into the system. FPGAs are commonplace in industries where connectivity to sensors and other devices are required. As quantum computing becomes commercially available at scale, the I/ONX platform can easily be adapted to interface with these quantum systems where the I/ONX stack handles data preprocessing and postprocessing.

Moving Toward Heterogeneous Compute for AI Training

The GPU has enabled data scientists to prove artificial intelligence products can be built at scale, and for that we should always be grateful. The AI boom was built on this processor, but the industry has misguidedly continued to attempt to scale on the rudimentary CPU and GPU stack.

Why? There are at least three reasons.

The first reason is that, simply put, all of the processors in Table 1 are manufactured by a different company. They have no incentive to sell a product with processors from their direct competitors. Each is trying to sell their respective processors as the best-in-class for AI training (or other workloads) and they focus their development efforts on making their processors faster and more powerful.

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

The second, less obvious reason is software. NVIDIA has a powerful software ecosystem called CUDA which is supported by common data science tools. AMD has ROCm, but the ecosystem is still early-stage, yet growing quickly. Software is the biggest challenge facing adoption of any HC platform, as each processor requires special compilation processes to run efficiently. While PyTorch has created opportunities for the portability of CUDA to other processors, the challenges are still significant for organizations with legacy code (Patel, 2023).

The third reason: it's really f*&#ing hard. I/ONX assembled a team from six specific domains of expertise ranging from embedded systems, hardware design, software engineering, data science, data-center operations, and product development to attack the problem. We have run countless experiments over several years on various hardware stacks using different approaches with software to get to where we are today with a fully functional, data-center-scale solution.

In sum, it's easier to quickly spin up scaled GPUs than it is to step back and architect a sustainable AI training hardware platform. I/ONX has taken that step back and has delivered a state-of-the-art, scalable, sustainable platform ready for the next generation of AI training.

Understanding the Business Drivers

The two primary customers of the I/ONX platform are:

1. AI Model Builders
    a. Academic institutions
    b. Research facilities
    c. Businesses with AI products
    d. Governments
2. Data Centers targeting demand for AI training compute
    a. Hyperscalers
    b. Edge

AI Model Builders

The first customer group is fairly self-explanatory. They need to run training workloads for an AI model but may not have access to state-of-the-art hardware, or they may have clear goals of operating compute resources in a sustainable manner.

Demand for high-performance compute wildly outstrips supply. Worse, only the largest organizations have been able to afford to buy up most of the current and near-term future supply (Gilbert, 2024). Elon Musk alone bought 100,000 of NVIDIA's

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

H100 processors with other large companies in similar purchase order sizes (Buntz, 2024).

With supply and power limitations, data scientists have had to resort to writing algorithms that reduce the need for GPUs at the expense of higher-quality products. In a recent paper, scientists from the University of California, Santa Cruz, Soochow University, and University of California, Davis partnered up to train a large language model without the need for matrix multiplication — a key reason for needing GPUs (Zhu, et al, 2024). While a powerful use case in ternary operators at scale, this is evidence of a community of researchers looking for solutions to the supply and power-consumption problems.

Data Centers

Outside of individual organizations, data centers represent significant buying power for high-performance computing, but they will only buy once demand has been demonstrated by the first group of customers.

Once that demand has been established, a key finding from I/ONX customer interviews is that data centers will consume all available power. If the I/ONX platform can reduce power consumption by 50%, these data centers will simply buy more hardware. The next constraint becomes physical space in the facility — which is addressed by radically compressing the footprint of high-performance compute using HC. Therefore, we have found that reducing power consumption and space — even while maintaining current state-of-the-art speeds — will result in data centers purchasing more hardware.

There are other critical factors beyond power and space. Cooling alone can contribute significantly to additional power consumption. Liquid immersion cooling is an option but requires completely retrofitting building infrastructure to support the weight — or building a new facility altogether.

For the edge data center use cases, these facilities can be as small as a few racks of hardware but can be several thousand square feet. These facilities serve organizations where on-premise solutions are not available but the data must be kept as close to the customer as possible. HC delivers for these customers with lower power consumption, higher heat dissipation, and reduced physical footprint, which are critical for providing high-performance compute in a smaller physical space.

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

Supply Chains

Most of the GPU-based processors are being fabricated at Taiwan Semiconductor Manufacturing Company (TSMC), but a number of companies are expanding their fabrication to U.S. soil. I/ONX has made a strategic choice to build our platform using processors from those companies who are shifting to the U.S.

The GPUs in the I/ONX stack will be 100% U.S. based between late 2025 into 2026 as their fabrication comes online, while other processor classes are either already made in the U.S.A. or will be around the same time as GPU manufacturing.

In all cases, we work with suppliers capable of scaling up into the tens of thousands or more for their chipsets. Early stage funding for I/ONX is focusing on locking down the supply chain with minimum order quantities with these vendors.

Software for Heterogeneous Compute

I/ONX primarily builds software using Rust, a memory safe, highly performant software language that has many benefits over Python for AI training (Gadaleta, 2024). Further, each processor in our stack requires custom development for specific algorithms to operate efficiently, as well as networking data flow between processors.

These additional processors — ASICs and FPGAs — are extremely difficult to program for and require specialized engineering talent to use them efficiently. The I/ONX software platform being developed obfuscates that complexity and enables data scientists to focus on their datasets and algorithmic architectures.

Beyond the specific data science toolkits, we also manage orchestration across multiple (even hundreds of) servers operating the same workloads.

Current State: 25Q2

As we wrap up 25Q2, I/ONX is coming out of stealth mode after completing 3rd party verifications with Slalom. The results of the Slalom reports, available upon request, demonstrate:

1. The system tested is fully standalone - no outside networking traffic proves the workloads being tested were in fact run on the rack-scale system (ie - we didn't cheat).
2. Two complex models, Whisper V3 Large and Stable Diffusion 1.5, ran with standard Python on the GPUs at greater than SOTA speeds - demonstrating out-of-the-box usability for customers running standard AI workloads and immediate benefits.

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

3. Rust-based Kernel Fusion - I/ONX has the ability to deploy fused kernels to any processor type with a compiler.
4. Heterogeneous workloads - a fully operationalized ResNet-50 example was validated using CPUs, GPUs, ASICs, and FPGAs, which is a world's first.

Future-Proof Architecture

The I/ONX hardware platform has been architected from the ground up to be future-proof. As new processors become available, our platform can be easily upgraded without needing to be replaced.

For example, if a new FPGA solution comes to market, the existing FPGAs in a rack can be easily pulled and replaced with the latest state-of-the-art without the full rack needing to be cycled out of the infrastructure. The same holds true for other processors in our stack, and entirely new processor types can be integrated into an existing installation.

Quantum computing is coming, and I/ONX has partnered with several organizations to ensure we are at the forefront of integrating into these processor classes. High-performance compute will be necessary to interface with quantum computing both as a data preprocessor and postprocessor. We have successfully conducted experiments on existing quantum processors.

Our roadmap extends our custom PCB designs for FPGAs to fully incorporate all necessary processors together on a single board enabling up to 192 FPGAs on a single motherboard. This approach further reduces energy consumption, makes the platform more flexible, and enables smaller product configurations for workstation deployments.

Final Thoughts

Heterogeneous compute is the future of AI. The complexity of the algorithmic workloads — especially at scale — means the field must move past rudimentary processor architectures that are creating unsustainable environmental impacts.

There is no single processor manufacturer that has developed all of the necessary processor classes in-house. This means an efficient system requires a company like I/ONX to put together a platform from multiple vendors — not unlike how Apple has developed their own HC for generalized computing.

I/ONX is the only organization building an HC platform tailored for AI training that can be scaled. The latest version, Symphony, is available as of June, 2025 with early

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

adopter customers already placing orders. With the hardware now firmly completed, we are shifting our focus to the development of software that will enable the data science community to easily transition to our platform.

References:

Barr, J. (2024). Amazon EC2 Trn2 Instances and Trn2 UltraServers for AI/ML training and inference are now available. Retrieved from: https://aws.amazon.com/blogs/aws/amazon-ec2-trn2-instances-and-trn2-ultraservers-for-aiml-training-and-inference-is-now-available/ on December 3, 2024.

Buntz, B. (2024). This week in AI: Musk unveils world's largest AI cluster, OpenAI eyes premium subscriptions. Retrieved from: https://www.rdworldonline.com/this-week-in-ai-musk-unveils-worlds-largest-ai-cluster-openai-eyes-2000-month-subscriptions/ on September 6, 2024

Burton, G. (2024). Squaring the circle: The high-performance computing energy paradox. Retrieved from: https://www.datacenterdynamics.com/en/marketwatch/squaring-the-circle-the-high-performance-computing-energy-paradox/ on September 8, 2024.

Clover, J. (2024). Apple Silicon: The Complete Guide. Retrieved from https://www.macrumors.com/guide/apple-silicon/ on September 9, 2024.

Gadaleta, F. (2024). Could Rust Be The Future Of Ai? Retrieved from: https://www.datasciencetalent.co.uk/could-rust-be-the-future-of-ai-by-francesco-gadaleta/ on September 2, 2024.

Gilbert, B. (2024). NVIDIA's Blockbuster Earnings All About the Supply/Demand Imbalances for AI. Retrieved from: https://www.carsongroup.com/insights/blog/nvidias-blockbuster-earnings-all-about-the-supply-demand-imbalances-for-ai/ on September 10, 2024.

Hoosain MS, Paul BS, Kass S, Ramakrishna S. Tools Towards the Sustainability and Circularity of Data Centers. Circ Econ Sustain. 2023;3(1):173-197. doi: 10.1007/s43615-022-00191-9. Epub 2022 Jul 1. PMID: 35791435; PMCID: PMC9247908.

Intel Manitoba (2003). https://www.gettyimages.com/detail/news-photo/intel-demonstrates-the-all-in-one-smartphone-manitoba-news-photo/524064410

Jouppi, N. P., Young, C., Patil, N., et al. (2017). In-Datacenter Performance Analysis of a Tensor Processing Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA), 1-12.

Judge, P. (2023). AI data centers could use more electricity than the Netherlands by 2027. Retrieved from: https://www.datacenterdynamics.com/en/news/ai-data-centers-could-use-more-electricity-than-the-netherlands-by-2027/ on September 9, 2024.

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.

Keserer, E. (2024). The six main subsets of AI: (Machine learning, NLP, and more). Retrieved from https://www.akkio.com/post/the-five-main-subsets-of-ai-machine-learning-nlp-and-more on September 9, 2024

Kuan-Chieh, H. and Hung-Wei T. (2023). Simultaneous and Heterogenous Multithreading. *56th Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO '23. York, NY, USA: Association for Computing Machinery. pp. 137–152.

Patel, D. (2023).How Nvidia's CUDA Monopoly In Machine Learning Is Breaking - OpenAI Triton And PyTorch 2.0. Retrieved from: https://www.semianalysis.com/p/nvidiaopenaitritonpytorch on September 8, 2024.

Samsung Exynos (2024). https://semiconductor.samsung.com/us/processor/ Retrieved on September 8, 2024.

Schlansker, M. (2011). Cydra 5. In: Padua, D. (eds) Encyclopedia of Parallel Computing. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09766-4_123

Zhu, R.J., Zhang, Y. Sifferman, E., Sheaves, T., Wang, Y., Richmond, D., Zhou, P., Eshraghian, J. (2024). Scalable MatMul-free Language Modeling. Retrieved from: https://arxiv.org/abs/2406.02528 on August 5, 2024.

© 2025 - All Rights Reserved - Cloud Computing Solutions Group, Inc.