Custom Systems for AI Acceleration

PURPOSE BUILT PERFORMANCE

I/ONX

Justyn Hornor

I/ONX HPC

TABLE OF CONTENTS

3RD PARTY VALIDATION: EXECUTIVE SUMMARY

Bold claims require strong evidence. We invited a team from Slalom to validate our revolutionary, rack-scale system for AI and other high performance computing use cases. The research the Slalom team conducted was aimed at confirming our platform is standalone, operates standard Pythonic workloads, and operates at or above state-of-the-art speeds. Additional validations include I/ONX's advanced Rust-based kernel fusion capabilities along with our unique ability to orchestrate multiple processor classes in a single rack: CPU, GPU, ASIC, and FPGA.

This summary document is a companion to the full report to provide decision-makers with a high-level understanding of the results of the Slalom team's analysis while orienting you to the capabilities I/ONX can bring to your organization.

OUR PHILOSOPHY: A NEW WAY TO THINK ABOUT AI PERFORMANCE

At I/ONX, we understand that unlocking AI's full potential requires a fundamental rethinking of infrastructure and performance — and we're building the solutions to make it possible. Through next-generation infrastructure engineering, we empower organizations to tackle their most demanding high performance computing challenges with greater speed, scalability, and operational efficiency.

THE NEXT GEN STACK: TOOLS REDEFINING AI INFRASTRUCTURE

The dominant model for scaling AI — simply adding more GPUs — is no longer viable: operationally, financially, or sustainably.

While GPUs have fueled major breakthroughs, they now drive runaway power consumption, excessive heat generation, vendor lock-in, and unsustainable infrastructure demands — straining both budgets and growth.

At I/ONX, we are fundamentally redefining the infrastructure model to meet the demands of both today and tomorrow:

- **Heterogeneous Computing**: Seamlessly integrating CPUs, GPUs, FPGAs, and ASIC processors into a cohesive, composable system optimized for AI workloads.
- **Streamlined Operations**: Consolidating diverse architectures into a single, flexible stack — dramatically lowering total cost of ownership.
- **Seamless AI Scaling**: Enabling intelligent workload movement across processor types to maximize efficiency, performance, and adaptability.
- **Energy Efficiency Gains**: Slashing power usage and operational costs while aligning with aggressive sustainability targets.
- **Open, Flexible Architecture**: Designed to eliminate vendor lock-in, the I/ONX stack supports open-source frameworks and diverse chipsets.

By moving beyond the GPU-centric paradigm, I/ONX offers organizations a decisive financial and operational edge:

- **Maximize ROI**: Scale intelligently while eliminating wasted infrastructure spending.
- **Lower Energy and Cooling Costs**: Significantly reduce one of the largest operational expenses in AI environments.
- **Increase Operational Efficiency**: Consolidate and simplify diverse compute resources within a unified, streamlined system.
- **Accelerate Time-to-Value**: Enable faster deployment, dynamic scaling, and quicker realization of AI-driven business outcomes.
- **Future-Proof Investments**: Build flexible, sustainable infrastructure ready to adapt to evolving technologies and AI innovations.

PERFORMANCE LEVERS: WHERE WE DRIVE THE MOST GAIN

At I/ONX, we focus on real-world impact—empowering enterprises to run AI workloads with greater speed, efficiency, and control. Our validation work highlights key levers that translate into business value:

- **On-Premise Execution**: Inference runs were confirmed to execute locally with no reliance on cloud compute. This reduces latency, enhances data security, and cuts down cloud-related operational costs.

- **Thermal Resilience**: Under sustained use, our system remained within optimal temperature thresholds. This reduces hardware wear, enables consistent performance, and lowers cooling requirements—directly contributing to cost savings and reliability.
- **Scalable Efficiency**: Performance scaled effectively across the range of tests we performed.
- **Versatile Model Support**: Our platform supports both visual and speech models out-of-the-box. Whether generating images or transcribing audio, organizations can rely on a unified system to handle diverse AI tasks.
-

SYSTEM SPECS: BUILT FOR SCALE AND SPEED

The I/ONX platform is engineered to meet the evolving infrastructure needs of AI-driven organizations. Its architecture supports demanding workloads while delivering operational simplicity and financial efficiency:

- **Enterprise-Grade Acceleration**: The specific configuration tested is powered by 16 AMD MI300X accelerators, logically grouped and physically arranged for optimal performance and thermal management.
- **Framework Flexibility**: Fully compatible with PyTorch and ROCm libraries, the system avoids vendor lock-in and supports modern AI models with minimal adaptation.
- **Thermal-Aware Design**: Both on-device sensors and thermal imaging validated efficient cooling—enabling dense deployments without risk of thermal throttling.
- **Optimized Software Stack**: Tuned for peak performance with dynamic kernel fusion and auto-tuning, maximizing throughput across varied batch sizes.
- **Data Sovereignty Built-In**: Inference occurs locally with no external API calls, ensuring full control over data handling and compliance.

PROOF IN PRACTICE: RESEARCH, BENCHMARKS, AND VALIDATION

Our system was validated through rigorous testing across three real-world scenarios, demonstrating consistent results that matter to enterprise buyers:

- **Operational Independence**: During image generation with Stable Diffusion, network traffic logs confirmed self-contained execution—eliminating external dependencies and reinforcing security.
- **Thermal and Power Stability**: Thermal imaging captured localized heat generation with no spread to inactive GPUs. Power draw aligned precisely with GPU activation, indicating energy-efficient operation.
- **Performance Scaling**: Across batch sizes and GPU counts, latency trends confirmed predictable improvements—ideal for scaling production workloads intelligently.
- **Multimodal Compatibility**: The Whisper speech model compiled and ran flawlessly, showcasing support for diverse AI workloads from image to audio.
- **High Token Output**: Whisper achieved 3,507 tokens/second on a single GPU, underscoring its readiness for enterprise-grade conversational AI and transcription services.

From thermal stability to throughput metrics, these findings validate the I/ONX system as a purpose-built platform for sustainable, scalable AI infrastructure. The result: faster outcomes, lower costs, and a smarter path to AI at scale.

VISION ROADMAP

Our ongoing vision centers on continuous innovation and validation in critical technology areas:

- **Rust-based Kernel Fusion Validation**: Current preliminary validations indicate significant performance improvements through advanced Rust-based kernel fusion techniques. Further comprehensive tests will confirm broader applicability and benefits.
- **Heterogeneous Computing Validation**: Initial tests underscore the capability and advantages of seamlessly moving AI workloads across diverse processing elements (CPUs, GPUs, FPGAs, ASIC).

Future validation plans include expanded scenario testing and real-world workload integration to ensure sustained and repeatable performance gains.

EXPLORE WHAT I/ONX CAN DO FOR YOU

Embrace the future of AI infrastructure with I/ONX. Discover how our innovative approach, validated by rigorous testing and real-world scenarios, can dramatically transform your AI operational strategy. Connect with our experts today to explore a comprehensive demonstration and learn firsthand how our technology delivers unparalleled efficiency, sustainability, and performance.