OPEN
PROGRAMMABLE
INFRASTRUCTURE
SUMMIT ON DPU/IPUs

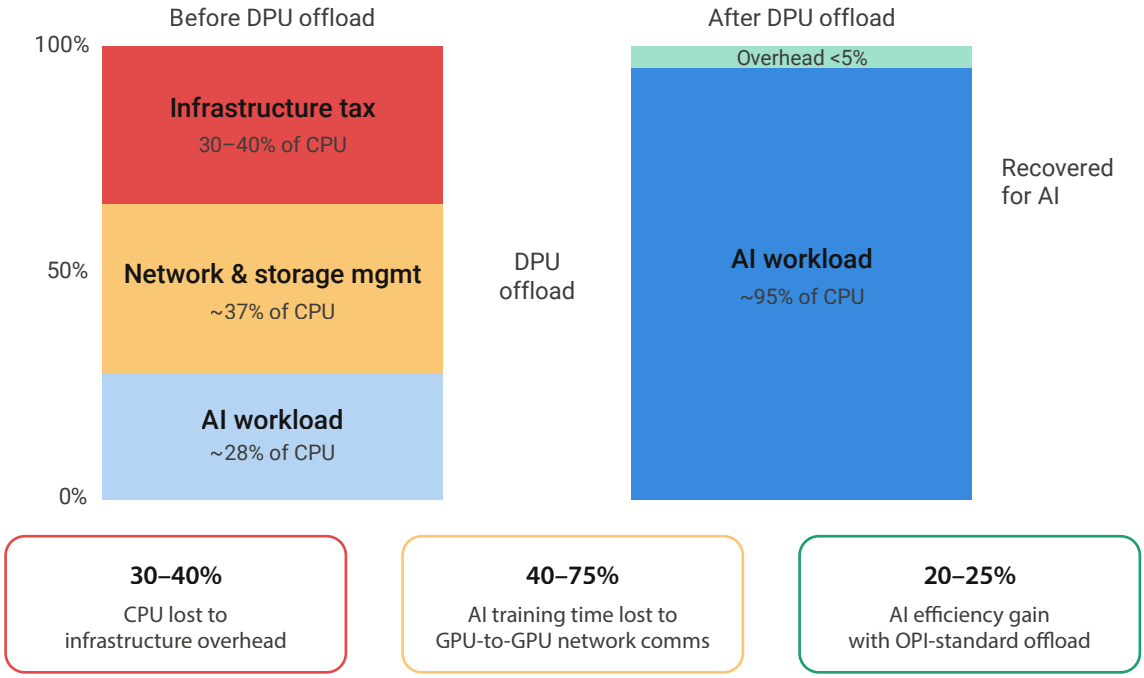
Accelerating the AI Era: The Critical Role of Open Programmable Infrastructure (OPI) in Scaling DPU and IPU Ecosystems

THE LINUX FOUNDATION

Executive Summary

Section 1: Executive Summary and Introduction

The scale and complexity of modern artificial intelligence workloads have fundamentally broken traditional, CPU-centric data center architectures. In conventional deployments, the host CPU is forced to act as the traffic cop for the entire system, managing software-defined networking, storage routing, and stateful security policies. This overhead, widely known as the "Data Tax" or "Infrastructure Tax," consumes between 30% to 40% of available server CPU cycles, according to industry estimates from hyperscaler deployments. For distributed AI clusters—where east-west GPU-to-GPU communication can account for an estimated 40% to 75% of model training time (based on cluster scale and topology)—forcing the host CPU to handle gigascale network processing, starves the AI application of vital compute resources and introduces unacceptable tail latencies.



To eliminate this bottleneck, the enterprise data center has turned to a new hardware paradigm: Data Processing Units (DPUs) and Infrastructure Processing Units (IPUs). Designed as the third pillar of data center computing alongside CPUs and GPUs, these specialized System-on-a-Chip (SoC) devices combine programmable multi-core processors (typically ARM-based), high-speed network interfaces, and dedicated hardware acceleration engines. By fundamentally decoupling infrastructure management from business applications, DPUs and IPUs offload, accelerate, and isolate networking, storage, and security tasks. This architectural shift establishes a strict security air gap between tenant applications and infrastructure controls, recovering the vast majority of the host CPU's capacity for dedicated, revenue-generating AI workloads.

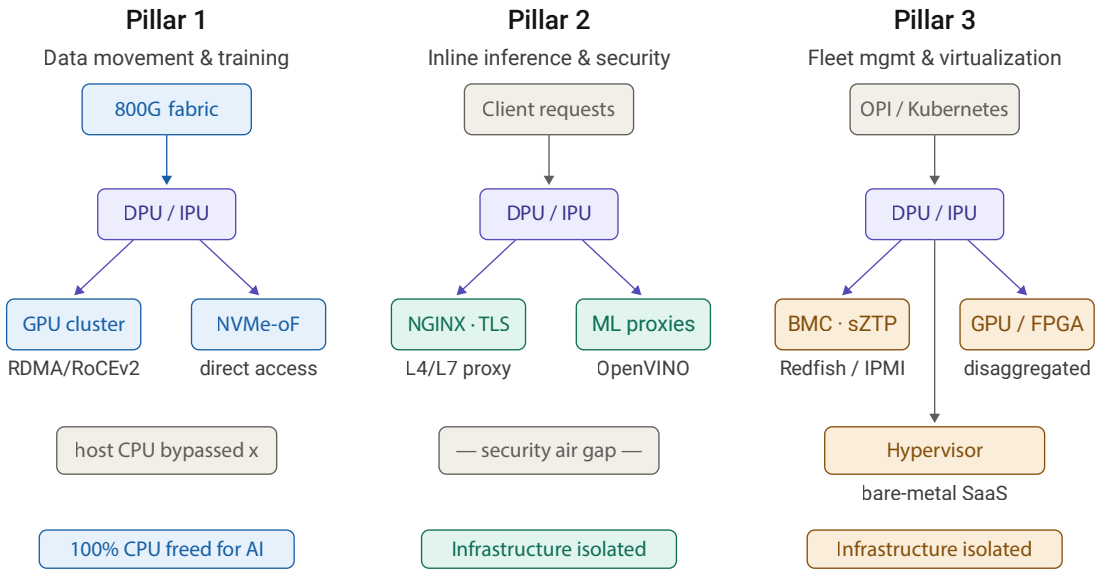
The Core Problem: The Fragmentation Crisis

While the silicon capabilities of modern DPUs and IPUs are undeniably transformative, their deployment at scale has introduced a critical operational crisis: The Fragmentation Problem. The current accelerator market is heavily fractured, with each major hardware vendor relying on bespoke, proprietary Software Development Kits (SDKs) and closed toolchains to manage their devices.

For example, programming an NVIDIA BlueField relies heavily on the DOCA framework, AMD Pensando utilizes its own SSDK, and Marvell relies on a tightly controlled OCTEON SDK. Consequently, enterprise IT architects attempting to build dual-vendor or multi-vendor supply chains are met with a mixed-fleet management nightmare. Without standardized Application Programming Interfaces (APIs), engineering teams are forced to rewrite applications, duplicate validation efforts, and maintain entirely separate provisioning processes for every distinct hardware architecture. Because every DPU today acts as a siloed appliance operating with its own drivers and programming models, establishing consistent behavior and telemetry across a mixed fleet is nearly impossible at scale.

To scale AI infrastructure without crippling vendor lock-in, the industry requires a unified abstraction layer. This is the foundational mission of the Open Programmable Infrastructure (OPI) Project. Operating under the Linux Foundation, the OPI Project is building a community-driven, standards-based open ecosystem for DPUs and IPUs. By defining vendor-neutral APIs, common behavioral models, and unified lifecycle management frameworks, OPI acts as the shared integration layer—essentially the “operating system” for DPUs.

This whitepaper explores how standardizing the DPU and IPU software stack through the OPI framework is the mandatory next step for enterprises aiming to build resilient, portable, and scalable AI infrastructure.



Section 2: The Architecture of Offloading & The Vendor Landscape

To achieve the performance thresholds required by modern artificial intelligence, data centers must transition from traditional compute-centric models to data-centric architectures. Data Processing Units (DPUs) and Infrastructure Processing Units (IPUs) drive this transition as highly integrated Systems-on-a-Chip (SoCs). They combine industry-standard multi-core processors (typically Arm Neoverse or Cortex), high-speed networking interfaces, and highly programmable hardware acceleration engines. These devices are fundamentally designed to absorb four primary infrastructure pillars:

- **Networking and RDMA for GPU-to-GPU Communication:** AI training relies on synchronous collective operations (such as All-Reduce), making it highly sensitive to tail latency. DPUs execute Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCEv2) or InfiniBand directly in hardware, enabling zero-copy data transfers between GPU memories while completely bypassing the host CPU and OS network stack. Advanced DPUs also implement hardware-based adaptive routing and path-aware congestion control (such as DCQCN or proprietary equivalents) to manage the massive, bursty “elephant flows” characteristic of distributed AI.

- **Storage Disaggregation via NVMe-oF:** DPUs emulate local NVMe virtual functions (VFs) to the host system while routing the actual storage traffic over the network via NVMe over Fabrics (NVMe-oF) or NVMe/TCP. This enables bare-metal hosts to access remote distributed flash arrays with near-local SSD latency, while the DPU’s hardware accelerators concurrently handle inline data compression, deduplication, and erasure coding.

- **Virtualization and OVS Offload:** In cloud-native deployments, software-based virtual switches like Open vSwitch (OVS) consume significant CPU cycles. DPUs offload the OVS datapath, packet classification, flow matching, and tunnel encapsulation/decapsulation (e.g., VXLAN, GENEVE) directly into programmable hardware pipelines.

- **Security and the Zero-Trust Edge:** By physically isolating infrastructure controls from the host CPU, DPUs establish a strict security “air gap”. They enforce zero-trust security at the edge via hardware roots of trust, inline cryptography (IPSec, AES-XTS, TLS), and distributed stateful L4 firewalls, preventing lateral movement even if the tenant’s application environment is compromised.

The Diverse Vendor Landscape

The silicon driving this architectural shift is both powerful and highly diverse, with multiple vendors engineering distinct approaches to infrastructure offloading:

- **NVIDIA** dominates the tightly-integrated AI ecosystem with its BlueField series, featuring Arm cores paired with custom datapath accelerators and deep integration into the NVIDIA Collective Communications Library (NCCL).
- **Intel** focuses on cloud-native separation with its IPU E2100 series, employing a Nitro-inspired architecture with dedicated Lookaside Cryptography and Compression Engines.
- **AMD Pensando** (Salina and Pollara series) heavily targets hyperscaler deployments utilizing a highly flexible P4-programmable pipeline, with its upcoming Pollara 400 purpose-built as an Ultra Ethernet Consortium (UEC) ready AI NIC.
- **Marvell** targets extreme compute density with the OCTEON 10, utilizing a 5nm process and incorporating dedicated hardware ML/AI accelerators that boost inline inference performance up to 100x (according to Marvell's internal performance studies) over software.
- **MangoBoost** provides highly specialized DPU architectures focused on accelerating NVMe/TCP storage protocols, maximizing GPU utilization for distributed AI training workloads.
- **Xsight Labs** disrupts the high-bandwidth sector with its 800G E1-SoC DPU, delivering full control plane and datapath programmability for edge and cloud environments.

The OPI Tie-In: The Pain Points of Hardware Diversity

While this diverse hardware landscape fuels rapid silicon innovation, it has inadvertently triggered a severe fragmentation crisis at the software layer. Currently, every DPU and IPU on the market operates as a siloed appliance, requiring its own proprietary operating environment, drivers, and software development kit. To deploy a NVIDIA BlueField, engineers must utilize the proprietary DOCA framework; AMD Pensando requires its SSDK with specific P4 toolchains; and Marvell mandates its restricted OCTEON SDK.

For enterprise IT architects, this lack of standardized Application Programming Interfaces (APIs) makes deploying a multi-vendor supply model nearly impossible. Integrating these devices into modern, Kubernetes-driven AI environments requires deep expertise across the specific vendor's software stack to configure Container Network Interfaces (CNIs), persistent storage plugins, and hardware operators. Managing a mixed fleet means operational silos are unavoidable—teams must duplicate validation efforts, rewrite infrastructure-as-code automation, and maintain entirely separate provisioning pipelines for each hardware type.

Without a unifying framework, the operational complexity of consuming proprietary APIs threatens to choke the scalability of cloud-native AI infrastructure. Overcoming this friction is the exact mandate of the Open Programmable Infrastructure (OPI) Project, which seeks to abstract these silicon differences and provide a unified, vendor-agnostic integration layer for the data center.

Section 3: The OPI Project Framework

The transition from legacy SmartNICs to fully independent DPUs and IPUs is fundamentally hindered by closed, proprietary ecosystems. To solve this fragmentation crisis, the Linux Foundation formed the Open Programmable Infrastructure (OPI) Project, an initiative designed to foster a community-driven, standards-based open ecosystem for next-generation hardware architectures. Within the industry, OPI is increasingly recognized as a unifying, standardized abstraction layer for DPUs. While it does not replace the native Linux distribution running on the hardware, OPI provides the essential common control and integration layer required to make heterogeneous accelerators universally usable at scale.

To achieve universal compatibility, the OPI architecture relies on a bifurcated abstraction model that shields the orchestration layer from underlying hardware complexities:

Southbound Abstraction: On the lower level, OPI connects downward to diverse silicon, firmware, and vendor-specific hardware capabilities. It translates standardized, vendor-neutral commands into silicon-specific instructions through dedicated implementation bridges. For example, the OPI SPDK Bridge (utilizing gRPC to SPDK JSON-RPC) for storage and the OPI EVPN Bridge for networking demonstrate how a common API can be realized across diverse hardware stacks without modifying the core logic. This abstraction ensures that the DPU's domain-specific hardware acceleration can be leveraged uniformly, regardless of the underlying silicon provider.

Northbound Integration: On the upper level, OPI exposes stable, cloud-native interfaces directly to orchestration platforms like Kubernetes or custom control planes. This northbound connectivity allows orchestration systems to treat DPUs and IPUs as well-defined, schedulable infrastructure resources. By integrating via Operators and Custom Resource Definitions (CRDs), Kubernetes can seamlessly schedule workloads that span both CPU and DPU capabilities across multi-tenant environments.

By defining shared APIs and behavioral models across the critical domains of networking, storage, and security, OPI significantly mitigates hardware vendor lock-in. This standardization simplifies the APIs within the application, enabling highly portable software while still delivering the full hardware-accelerated benefits of software-defined networking (SDN) overlays, remote storage access, and inline encryption. Furthermore, moving these standardized controls to the DPU creates a rigid security air gap. Virtualization control, distributed firewalls, and cryptographic certificates are kept strictly on the DPU, making it impossible for malware on the host node to compromise the cloud's underlying infrastructure or other tenant applications.

Finally, managing a heterogeneous fleet of thousands of DPUs in production requires rigorous operational standardization. OPI tackles the critical challenge of fleet management at scale by defining vendor-agnostic workflows for device discovery, secure zero-touch provisioning (sZTP), secure boot protocols, and firmware updates. By integrating DPU lifecycle and telemetry data (such as

metrics, logs, and tracing via OpenTelemetry) directly into existing enterprise automation stacks, OPI ensures that architects can seamlessly deploy, monitor, and update a multi-vendor accelerator fleet from a single, unified control plane. Ultimately, the OPI framework transforms DPUs from bespoke, siloed appliances into first-class, standardized infrastructure elements.

Infrastructure Control: Virtualization and Fleet Management at Scale

Beyond moving massive datasets and hosting inline security proxies, DPUs and IPUs serve a foundational third role: they act as independent control hubs for the entire server fleet. Historically, infrastructure management tasks—like booting servers, installing operating systems, and deploying hypervisors—were deeply intertwined with the host CPU. By leveraging the Open Programmable Infrastructure (OPI) framework, organizations can completely decouple these controls, transforming the DPU into an autonomous management endpoint.

- **Virtualization and Bare-Metal-as-a-Service:** DPUs enable true hardware-level multi-tenancy and virtualization offload. Because the DPU possesses its own processor and boots a general-purpose operating system, it can independently run hypervisors (such as VMware ESXi) or container platforms isolated from the host. This allows cloud and enterprise operators to rent out 100% of the host CPU and GPU to tenants as a bare-metal instance without sacrificing control over the network or storage. By enforcing virtualization on the DPU, OPI creates a strict "security air gap"—meaning even if a tenant completely compromises the host operating system, they cannot access the hypervisor or the cloud provider's underlying management plane.

- **Standardizing Fleet Lifecycle Management:** Managing a mixed-vendor fleet of thousands of DPUs requires rigorous automation. OPI's Provisioning and Lifecycle working groups provide standard abstractions for the entire device lifecycle. Instead of using fragmented vendor tools, operators can rely on OPI's standardized Secure Zero Touch Provisioning (sZTP), automated discovery, and unified firmware updates. OPI integrates deeply with out-of-band management (OOBM) technologies, allowing orchestrators to communicate securely with the DPU's Baseboard Management Controller (BMC) via Redfish or IPMI protocols. This means infrastructure teams can power-cycle servers, monitor hardware sensors, and stream OpenTelemetry metrics entirely outside the data path.

- **Resource Disaggregation and Pooling:** Finally, DPUs serve as the bridge to completely disaggregated architectures. By moving infrastructure controls onto the DPU, resources that were previously trapped inside individual servers—like GPUs, FPGAs, and NVMe SSDs—can be abstracted and presented to the network. OPI's standard APIs allow orchestrators to dynamically pool these heterogeneous resources, granting an application on one server direct, low-latency access to an AI accelerator on another. This drastically improves data center utilization and lowers the Total Cost of Ownership (TCO) for expensive AI hardware.

Section 4: Accelerating AI Workloads with OPI

As enterprise focus expands from foundational model training to large-scale AI inference, the volume of east-west data movement across the network has become a primary bottleneck. Distributed AI operations—whether multi-node GPU training or clustered inference—require lossless, ultra-low-latency communication architectures leveraging Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCEv2).

Historically, programming these hardware-accelerated data flows required deep integration with proprietary, vendor-specific toolchains. The Open Programmable Infrastructure (OPI) Project resolves this by standardizing the APIs and behavioral models used to manage these network offloads. By utilizing OPI's standardized plugins and APIs to manage massive AI data movement and infrastructure optimization, enterprises demonstrate highly measurable efficiency gains, with industry benchmarks suggesting they can run AI workloads 20% to 25% more efficiently. This standardized offloading shrinks the compute footprint of traditional networking and virtualization tasks, freeing up rack space, cooling, and power specifically for high-density AI hardware.

Validated Use Case: Confidential AI Inference Security

While standardizing GPU-to-GPU network flows is critical for training, securing AI inference pipelines presents an equally complex challenge. Hosting security proxies and AI models on the same host CPU creates shared vulnerabilities and forces the infrastructure software to compete with the AI application for CPU cycles, memory, and bandwidth.

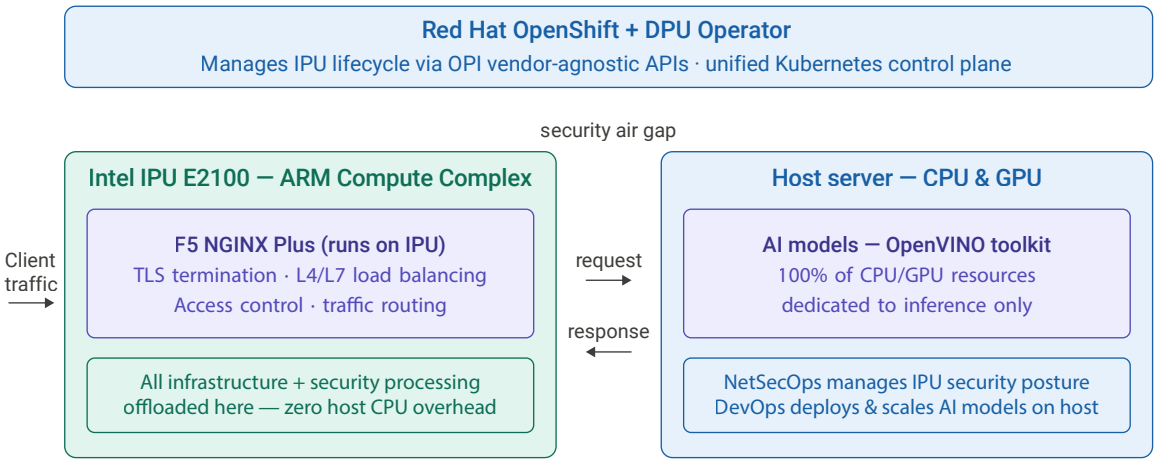
To solve this, the OPI community has validated a comprehensive reference architecture for confidential AI inference. This solution leverages the **Intel Infrastructure Processing Unit (IPU) E2100**, **Red Hat OpenShift**, **F5 NGINX Plus**, and the **OpenVINO toolkit** to create a highly secure, hardware-accelerated inference pipeline.

Operating under the OPI framework, the deployment functions as follows:

- **Unified Orchestration:** Red Hat OpenShift utilizes the DPU Operator to interface with the Intel IPU via OPI's vendor-agnostic APIs. This allows Kubernetes to seamlessly manage the IPU's lifecycle using standard, cloud-native workflows.
- **Infrastructure Offload:** Instead of running on the host, the F5 NGINX Plus reverse proxy is deployed directly onto the IPU's Arm Compute Complex (ACC). NGINX Plus handles all traffic routing, L4/L7 load balancing, access control, and TLS termination.
- **Dedicated AI Processing:** The proprietary AI models, optimized by the OpenVINO toolkit, run purely on the host server's CPU or attached GPUs.

This architecture establishes a physical "security air gap" between the untrusted external client requests terminating on the IPU and the valuable AI models residing on the host. Because the IPU absorbs almost all infrastructure and security processing, the vast majority of the host CPU's resources are preserved for high-performance OpenVINO inference.

Furthermore, this OPI-standardized deployment restores a critical separation of duties at the hardware level: NetSecOps teams can independently manage the routing and security posture on the IPU, while DevOps teams deploy and scale AI models on the host, eliminating operational friction without sacrificing security.



Section 5: Business Impact & Call to Action

As enterprises and hyperscalers invest billions into artificial intelligence, extracting maximum value from these deployments hinges on a highly optimized, scalable, and flexible underlying architecture. Embracing an open ecosystem for Data Processing Units (DPUs) and Infrastructure Processing Units (IPUs) fundamentally transforms the financial and operational calculus of the modern data center.

The Strategic Business Value of Open Programmable Infrastructure

From a financial perspective, standardizing on the Open Programmable Infrastructure (OPI) framework dramatically lowers Total Cost of Ownership (TCO). By offloading the 30% to 40% “infrastructure tax” from host servers directly onto DPUs and IPUs, organizations immediately reclaim precious CPU cycles. This elimination of overhead ensures that highly expensive GPU clusters are fed data at line rate rather than starved by host bottlenecks. In distributed AI training, where network communication can consume 40% to 75% of job time, OPI-standardized offloading reduces Job Completion Times (JCT) and drives a significantly higher Return on Investment (ROI) on GPU hardware.

Operationally, an open ecosystem eliminates the crippling risk of vendor lock-in. By abstracting proprietary silicon and Software Development Kits (SDKs) behind standard OPI Application Programming Interfaces (APIs), organizations gain the agility to adopt a multi-vendor supply chain. Engineering teams can write cloud-native infrastructure code once and deploy it seamlessly across diverse hardware fleets—whether utilizing NVIDIA, Intel, AMD, or Marvell silicon. Furthermore, OPI’s standardized approaches to secure zero-touch provisioning, fleet telemetry, and lifecycle management drastically reduce the OpEx and validation efforts required to maintain gigascale AI environments.

Ultimately, aligning with OPI future-proofs AI infrastructure. As networks scale to 400G, 800G, and integrate with emerging networking standards like the Ultra Ethernet Consortium (UEC), an open software layer guarantees that your orchestration platforms remain adaptable to next-generation hardware innovations without requiring massive software rewrites.

Call to Action: Shape the Future of AI Infrastructure

The transition to programmable AI data centers is inevitable, but its openness depends on industry collaboration. Relying on fragmented, siloed appliances will only stifle innovation and inflate costs.

The future of programmable AI infrastructure must be open. We invite cloud architects, NetSecOps professionals, and IT leaders to join the Open Programmable Infrastructure (OPI) Project community under the Linux Foundation.

Take action today:

- **Join the Working Groups:** Contribute your expertise to the API & Behavioral Model, Provisioning & Platform Management, Dev Platform/PoC, or Use Case working groups.
- **Contribute to the OPI Lab:** Help validate architectures by sharing hardware, experimental results, and proof-of-concept AI deployments.
- **Demand Open Standards:** Mandate OPI compatibility in your upcoming hardware RFPs to ensure you are adopting vendor-agnostic, OPI-compliant DPUs and IPUs for your next-generation data centers.

Together, we can build the open standards that will power the scalable, secure, and efficient AI factories of tomorrow.