THE ULTIMATE GUIDE TO AI DATA CENTER INFRASTRUCTURE SECURITY_

THE AI ARMS RACE_

We are in the midst of a global AI arms race. The leaders will steer global industry and society for the coming century. AI data centers are critical infrastructure and will be targeted with the full power and sophistication of nation state cyberattackers.

AI data center expansions are on an exponential growth curve fueled by this international competition for AI supremacy. The $500 billion Stargate Project, CoreWeave's IPO, and the tens of billions of dollars pledged by NVIDIA, OpenAI and others to the Humain project to build AI infrastructure in the Middle East are just a few examples of how AI is reshaping the global data center landscape. While most of the headlines focus on the scale and investment figures driving the AI boom, the foundational security risks threatening these projects are at increased risk as a new attack surface.

The AI Infrastructure Threat Landscape

The landscape of attacks targeting AI infrastructure is vast and continuously evolving, with numerous threat actors exploiting both AI software and hardware stacks. These include nation-state-sponsored APT groups, ransomware operators, and espionage entities. The consequences of these attacks can range from compromising the integrity of AI models to using AI infrastructure as a vector for further infiltration into victim organization. Some of the commonly observed attack categories and examples against AI infrastructure include:

- Attacks on GPU and other AI hardware components
- Device remote exploitation (for data theft, remote control / hijacking, lateral IT infection) by APT5, APT41, APT28, APT29, APT33, APT34, APT39, UNC2630, FIN11, TA505, UNC2447, Keksec, Foxkitten, UNC3524 QUIETEXIT
- Destructive attacks like VPNFilter, Cyclops Blink, iLOBleed implant, wipers, ViaSat satellite modems hack
- Persistent compromise and implants affecting critical server and network infrastructure equipment, such as Volt Typhoon, Cisco Line Dancer and Line Runner, Palo Alto Networks GlobalProtect UPSTYLE backdoor, Black Lotus, APT41 Speculous, Cisco ROMMON and SYNful Knock, F5 BIG-IP CVE-2022-1388 exploits, Citrix Bleed, Vigor, LoJax, MosaicRegressor, MoonBounce, ESPEcter and FinSpy.
- Counterfeit devices such as counterfeit Cisco equipment
- Supply chain breaches via firmware and software updates (SolarWinds SUNBURST, ShadowHammer, MSI ransomware), firmware build / signing infrastructure, or device interjections (Cisco)
- Attacks targeting models, like malicious AI ML models

Core Risk Categories in AI Infrastructure

AI Infrastructure Risks for Data Centers

From AI Servers, to GPUs, to network infrastructure and other foundational tech, increased security and continuous protection is needed. While the OWASP Top 10 Risks for LLMs and other research have heavily focused on AI software and model risks, the security of the hardware, components, and supply chain have been largely ignored. That said, NIST SP 800-223 has identified key cyber risks related to infrastructure for High Performance Compute (HPC) which includes AI data centers. Risks include:

- **Attacks on critical hardware components** and software manipulation to gain unauthorized access.
- **Rapidly changing infrastructure for HPC firmware and hardware components**, increasing supply chain risk
- **Compute Node Sanitization challenges**, such as validating firmware between task runs on shared compute infrastructure.

The enormous financial and reputational stakes mean that AI systems architects who are working on the next generation of AI infrastructure face massive challenges in assuring the security of their systems and supply chains.

Five significant security risk areas that must be addressed by AI system architects and executives at AI data centers include:

1. **AI Application Security**
2. **Model Security and Trust**
3. **Cloud Delivery Risks**
4. **Software Supply Chain Risks**
5. **Hardware, Firmware, and Data Center Network Infrastructure**

Layer	Risk Examples	Controls and Tools
AI Application Security	Jailbreaking, prompt injection, hallucinated output, "slopsquatting" on hallucinated code package names.	Red teaming, output validation, and keeping a human in the loop before using GenAI outputs for business critical functions are all key controls. Selecting AI service providers, models, and securing the entire AI stack is critical for trusting and securing the application layer. Using App Security standards like OWASP and maintaining SBOMs for all applications can support AI application security and trust.
Model Security and Trust	Data poisoning, leakage of proprietary training data, and faulty decision making based on faulty models.	Protecting training data and assuring that model weights are not poisoned is a foundational requirement for securing and trusting Generative AI. Companies like Protect AI, HuggingFace, and others strive to provide access to trusted model weights and to scan existing models for malicious or vulnerable traits.
Cloud Delivery Risks	Unintentional cross-tenant data leakage of model weights or training data. Data poisoning due to unintentional cross-tenant access. Vulnerabilities such as those in NVIDIA Container Kit allowing malicious data theft in cloud and container scenarios.	Much of AI model training and inference will be delivered by third party neocloud data centers or hyperscalers. Assuring the security of these data and services both at the data center and in transit between data center and end customer is crucial. AI Data Center providers must clarify the shared responsibility model of cloud security with their customers.

Software Supply Chain Risks	Backdoors introduced in open source packages, Linux distributions, and other foundational software components that underpin much of enterprise IT infrastructure.	Third party risk management, Software Bill of Materials, Hardware and Firmware Bill of Materials all contribute to mitigating software supply chain risks. Vendors such as ChainGuard and Endor are making strides in providing secure, hardened versions of widely used Open Source software so that enterprises can avoid accepting unknown risk from the FOSS supply chain.
Hardware, Firmware, and Data Center Network Infrastructure	Firmware backdoors, bootloaders, memory leaks, and incomplete device cleansing between AI data center tenants can lead to model leakage, intellectual property loss, and poisoned models being used for business critical decisions.	AI data center providers and enterprises who run their own AI model training should rigorously vet all of the GPUs and AI hardware they use, as well as the ancillary networking infrastructure, including routers, switches, firewalls, and load balancers. Eclypsium monitors AI data center infrastructure down to the hardware and firmware level to assure integrity and detect indicators of compromise.

Cheat Sheet: Self-Assessment for AI Data Center Providers

Here are sample assessment questions that AI data centers should ask themselves for each security measure:

Trusted Computing for AI Accelerators

Hardware Security:

- Do we implement hardware-based attestation for all AI accelerators (GPUs, TPUs, custom chips) to verify their integrity before workload execution?
- Have we established secure boot processes for AI hardware that prevent unauthorized firmware or driver modifications?
- Do we use hardware security modules (HSMs) or trusted platform modules (TPMs) to protect cryptographic keys used in AI workload processing?

Runtime Protection:

- Can we guarantee that AI model weights and training data remain encrypted even during computation on accelerators?
- Do we have mechanisms to detect and respond to hardware tampering or side-channel attacks on AI chips?
- Are we using confidential computing technologies (like secure enclaves) for sensitive AI workloads?

Cheat Sheet (Continued)

Network and Tenant Isolation Guarantees

Multi-Tenancy Security:

- How do we ensure complete isolation between different customers' AI training runs and inference workloads?
- Can we guarantee that one tenant's model data, gradients, or intermediate computations are never accessible to another tenant?
- Do we have network segmentation that prevents cross-tenant data leakage during distributed training across multiple nodes?

Data Flow Control:

- Are all inter-node communications for distributed AI workloads encrypted and authenticated?
- Do we have microsegmentation policies that restrict network access based on the specific AI workload requirements?
- Can we provide cryptographic proof of tenant isolation to customers upon request?

Innovation in Operational and Physical Security for Data Centers

AI-Aware Physical Security:

- Have we implemented biometric access controls and continuous monitoring for areas housing high-value AI training infrastructure?
- Do our security protocols account for the unique risks of AI model theft, including protection against sophisticated espionage attempts?
- Are we using AI-powered security systems for anomaly detection in physical access patterns and environmental monitoring?

Operational Excellence:

- Do we have incident response procedures specifically designed for AI workload compromises or data exfiltration attempts?
- Are our staff trained on AI security risks and cleared to appropriate levels for handling sensitive AI assets?
- Do we maintain air-gapped backup systems for critical AI infrastructure and model checkpoints?

AI-Specific Audit and Compliance Programs

Specialized Auditing:

- Do we conduct regular audits specifically focused on AI model security, including verification that customer models haven't been compromised or leaked?
- Are our audit trails comprehensive enough to track the complete lifecycle of AI training data and model artifacts?
- Do we have third-party security assessments that specifically evaluate our AI infrastructure against emerging threats?

Compliance Framework:

- Have we developed compliance programs that address AI-specific regulations and standards (such as AI safety requirements, model governance, and algorithmic accountability)?
- Can we demonstrate compliance with data protection regulations when handling training datasets that may contain personal information?
- Do we maintain documentation and controls that would satisfy auditors examining our AI security posture and customer data protection measures?

The Complexity of Securing AI Infrastructure

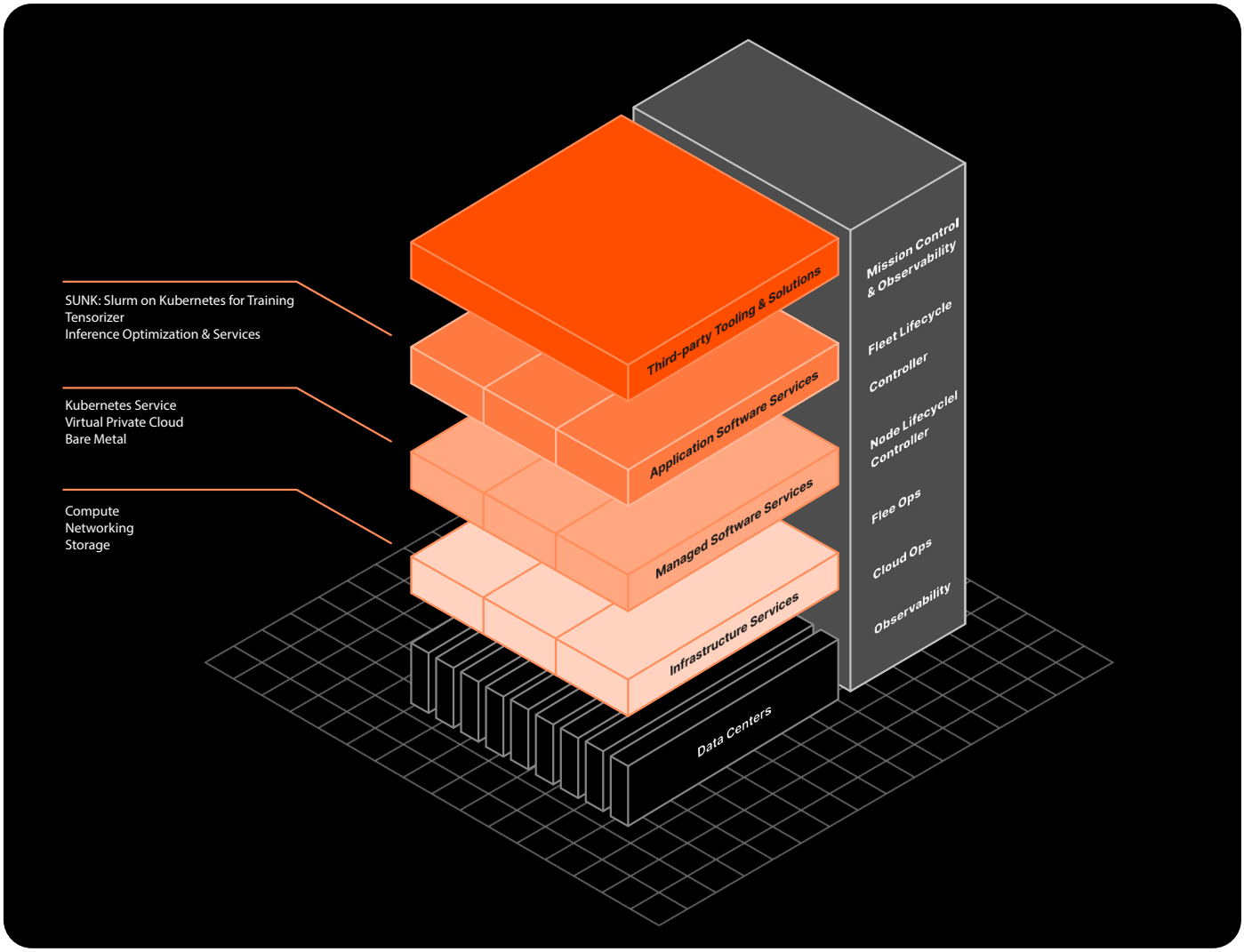
NVIDIA founder Jensen Huang recently described a single AI rack as containing 600,000 components, two miles of cable, and weighing 3,000 pounds. An AI data center will contain many such racks, networked together with routers, switches, protected by firewalls, and otherwise deeply entwined with technology that contains a large number of vulnerabilities and is being targeted by attackers and vulnerabilities. This complexity, and deep interconnection with more traditional computing and networking gear, introduces risk.

So far, security discussions around AI have focused largely on models, data, and applications, leaving out the much larger, more complex and foundational risk area of compute hardware and network infrastructure. However, the relationship between hardware security and model security was a key discussion in a recent OpenAI post Reimagining secure infrastructure for advanced AI, which noted that "AI is the most strategic and sought after technology of our time. It is pursued with vigor by sophisticated cyber actors with strategic aims."

Much of AI data center infrastructure is similar to traditional data centers, but there are key differences that make it both more valuable as a target for cyberattackers. The extreme demand for performance in AI workloads means that bare metal, as opposed to virtualized infrastructure, is often preferable. Companies using bare metal infrastructure have greater access to the hardware capabilities, which also creates more potential for them to introduce risky backdoors or settings to the bare metal resources. Eclypsium researchers discovered a real world instance of bare metal resources in IBM Cloud being compromised in this way.

While hardware-level root of trust capabilities, confidential computing, and trusted execution environments (TEEs), have come a long way in CPU-land, they are still relatively new in the world of GPUs and TPUs for AI acceleration. The OpenAI blog referenced above specifically noted that: "As promising as confidential computing for GPUs is, the technology is still nascent. Investment in both hardware and software is required to unlock the scale and performance necessary for many large language models and use-cases. Additionally, confidential computing technologies on CPUs have had their share of vulnerabilities, and we cannot expect GPU equivalents to be flawless. Its success is far from given, which is why now is the time to invest and iterate so we can one day realize its potential."

To provide a sense of the numbers of moving parts that make up AI infrastructure, here is a non-exhaustive diagram followed by a list of elements and considerations that are all standard parts of the AI infrastructure, and are parts of the attack surface.



Hardware: GPUs (NVIDIA, AMD, Intel), TPUs (Google TPUs), FPGAs (Intel, Xilinx), ASICs (Application-Specific Integrated Circuits), CPUs, AI Edge Devices (NPU chips, custom AI SoCs), Memory (VRAM, HBM, GDDR), Interconnects (NVLink, PCIe, InfiniBand).

Firmware: GPU Firmware, FPGA Firmware, ASIC Microcode, BIOS/UEFI Firmware, Bootloaders, Security Modules (TPM, DICE, RoT).

Driver Software: GPU Drivers (CUDA, ROCm, DirectML), FPGA Drivers, ASIC Drivers, Kernel Modules, Hypervisor Drivers (vGPU, SR-IOV).

ML Frameworks: Training Frameworks (TensorFlow, PyTorch, JAX, MXNet), Inference Frameworks (ONNX, TensorRT, OpenVINO), Compiler Toolchains (XLA, TVM), API Libraries (CUDA, ROCm, SYCL), Middleware (Horovod, Ray).

Pre-Trained Models: Model Weights (Transformer Models, CNNs), Model Files (ONNX, SavedModel), Transfer Learning Modules, Model Repositories (Hugging Face, Model Zoo).

Security Controls: Secure Boot, Firmware Attestation, Isolation Techniques (VMs, Containers, TEEs),

Cryptographic Modules (TPM, HSMs), Secure Communication Protocols (TLS, IPsec).

Data Handling: Data Pipelines, Data Encryption (At-rest, In-transit), Data Access Control (RBAC, ABAC), Data Provenance (Tracking and ensuring data integrity).

Virtualization & Orchestration: Virtual Machines (VMs), Containers (Docker, Kubernetes), Hypervisors (VMware, KVM), Resource Managers (Kubernetes, Slurm).

Monitoring & Logging: GPU Monitoring Tools (nvidia-smi, ROCm SMI), Logs (System, Firmware, Network), Security Monitoring Platforms (SIEM, XDR).

Networking & Interconnects: NVLink, PCIe, InfiniBand, Ethernet, Data Center Network Fabrics, Secure Communication Protocols (TLS, etc).

Energy & Cooling Systems: Power Supply Units (PSUs), Liquid Cooling, Air Cooling, AI-Specific Cooling Systems, Power Management Tools.

Deployment & Management Tools: Infrastructure-as-Code Tools (Terraform, Ansible), Model Deployment Tools (ONNX Runtime, Triton), CI/CD Pipelines, Configuration Management Tools.

The Lifecycle of AI Infrastructure: Onboarding, Production Use, Recycling, Decommissioning

Technology Validation and Onboarding

AI data center operators can reduce risk and maintain compliance with security standards during the onboarding phase for AI infrastructure by scanning the hardware, firmware, and components of every GPU and component before deploying them. The benefits are:

- Complete inventory of firmware and driver versions
- Identify known vulnerabilities
- Detect counterfeits

Production Use

Legacy network infrastructure, including routers, switches, or baseboard management controllers (BMCs), remains a significant security blind spot for operators that lack the tools or telemetry to detect threats at the firmware level. These devices are essential to AI data center operations but are often excluded from traditional endpoint detection and response (EDR) and security information and event management (SIEM) systems. Forward thinking AI data center operators should deploy technologies that are able to baseline and continuously monitor the firmware and integrity of networking infrastructure in order to keep pace with increasingly sophisticated firmware-based attack techniques.

Recycling

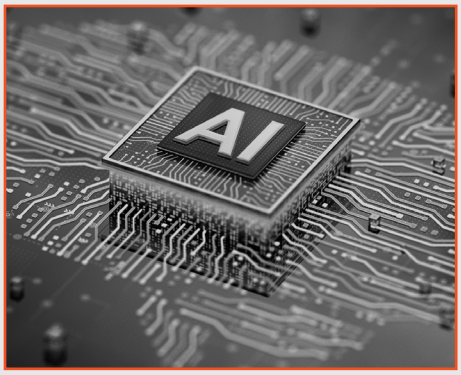
It is crucial for AI data center operators to verify the integrity of their GPU resources between customers. This includes the delivery of actionable summaries of vulnerabilities and integrity failures for individual GPUs, or all GPUs throughout a data center, thereby ensure secure firmware and configuration of shared AI resources before releasing to subsequent customers.

Secure Decommissioning

The average lifespan of a GPU in a server environment, especially in data centers, is typically between 3-5 years. When AI data center operators are preparing GPU Servers for Resale, it is critical to ensure that no sensitive data remains on this highly valuable equipment.

Case Study: Large Neocloud Provider

Eclypsium already delivers supply chain security for high performance computing datacenters by continuously monitoring firmware and components of hard-to-monitor network devices. Eclypsium continuously baselines and checks the integrity of firmware against commonly attacked routers, switches, load balancers, servers, and other network gear. Eclypsium also detects both known and unknown vulnerabilities as well as active attacks and indicators of compromise against these devices.



With industry-leading innovation, close ties to industry titans such as NVIDIA, Microsoft and OpenAI, and meteoric growth across the US and Europe, this organization has quickly become one of hottest cloud platform providers in the industry.

But this rise naturally comes with challenges. The company's services rely on a rapidly growing fleet of highly specialized server hardware and AI GPUs, and it is up to the security team to ensure the integrity and security posture of the hardware and components at the heart of the platform. They also wanted to be certain that, when AI compute resources are shifted to a new customer or training run, the hardware and firmware are secure and integrity guaranteed to the highest degree possible.

By partnering with Eclypsium, this organization was able to harness a turn-key solution for the assessment and ongoing monitoring of their servers and internal components such as UEFI firmware, NVIDIA GPUs, Intel CPUs, and more. This enabled their team to proactively verify technology supply chains while ensuring the highest levels of security for their customers.

Business Needs

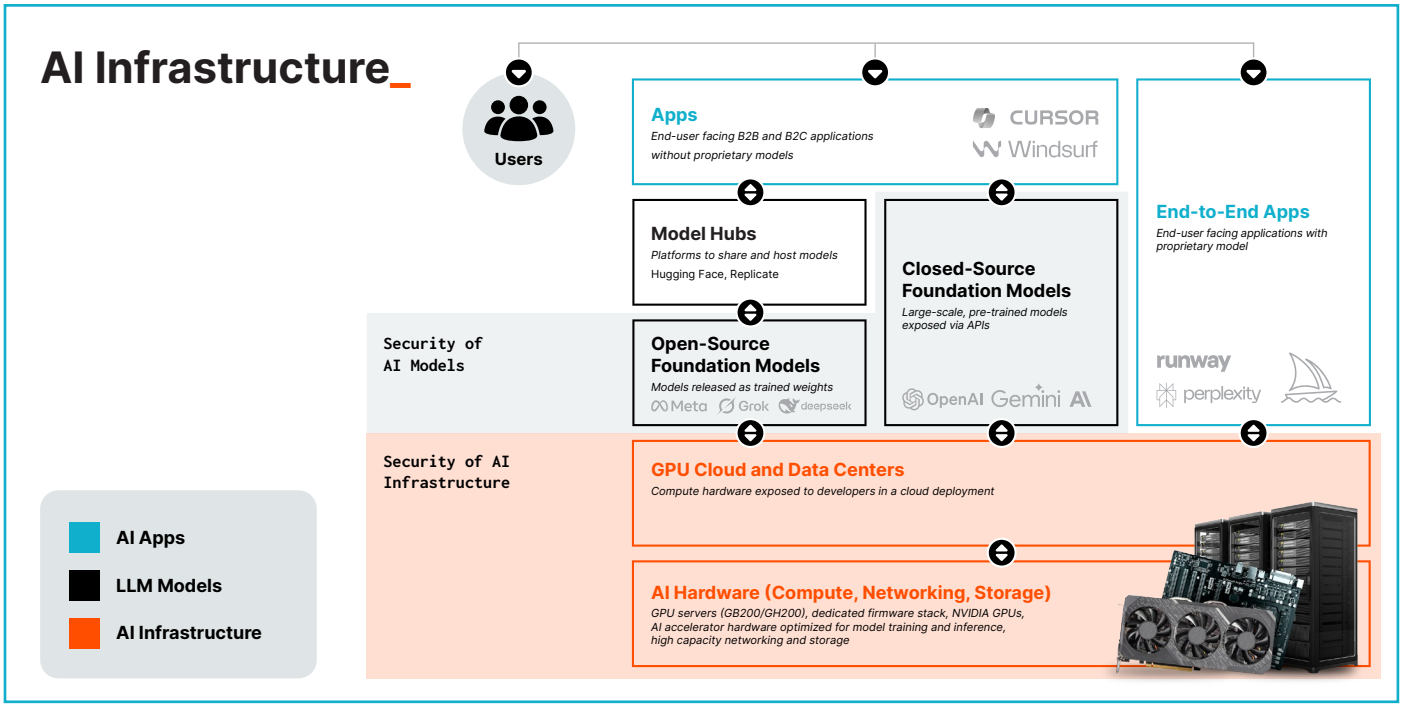
- Verify the integrity of firmware in AI server infrastructure and components and proactively alert on any changes.
- Baseline and re-verify integrity of AI infrastructure between customers.
- Firmware-specific vulnerability management, updating, and configuration management, fulfilling shared responsibility obligation to secure AI infrastructure
- Support fast, efficient rollout while minimizing costs and impact to staff.

Eclypsium Benefits

- Deep vendor-agnostic visibility into AI servers and components including NVIDIA GPUs
- Automated assessments to identify vulnerabilities, threats, or changes to device firmware and proactively verify firmware and supply chain integrity.
- Simple, easy-to-deploy solution at a significantly lower cost than developing custom solutions for each vendor.

How Eclypsium Supports AI Data Center Supply Chain Security

Eclypsium already delivers supply chain security for high performance computing datacenters by continuously monitoring firmware and components of hard-to-monitor network devices. Eclypsium continuously baselines and checks the integrity of firmware against commonly attacked routers, switches, load balancers, servers, and other network gear. Eclypsium also detects both known and unknown vulnerabilities as well as active attacks and indicators of compromise against these devices.



By inventorying assets at the hardware, firmware, and component level, and monitoring for loss of integrity and attacks, Eclypsium radically improves the security posture of High Performance Computing (HPC) data centers, as well as AI data centers and infrastructure that are rapidly becoming foundational to the world's leading enterprises.