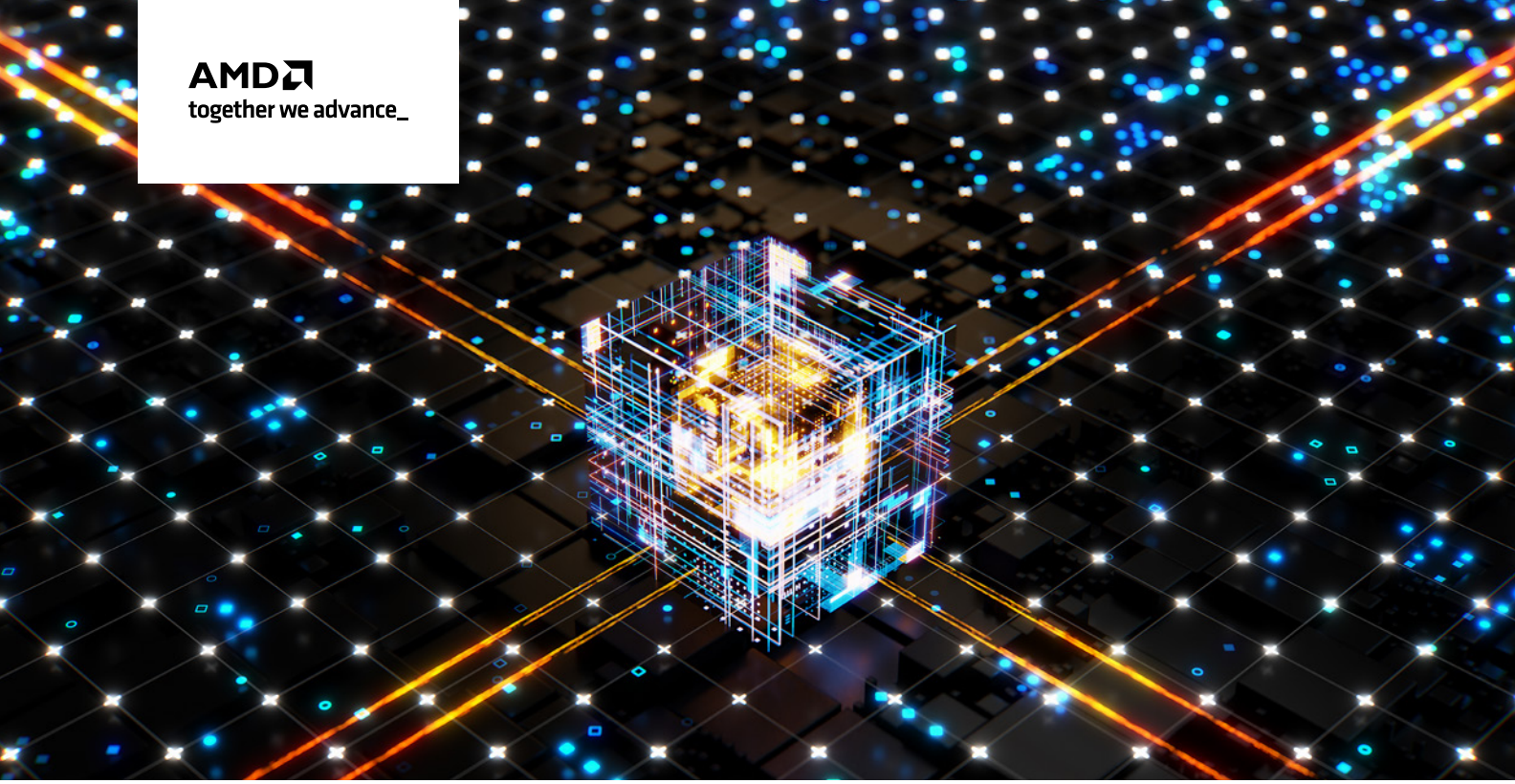

WHITE PAPER

Protecting the AI-ready data center

AI doesn't happen in one place. Workloads are constantly on the move, touching myriad systems both inside the data center and beyond. Data shifts across multiple storage arrays and compute nodes, over networks, to the edge and back again.

Every movement and resting point is a potential opportunity for compromise and, as AI systems grow more capable, so can the threats that target them. Modern AI workloads demand massive data throughput, distributed processing, and complex model architectures, all of which create new surfaces for adversaries to probe. From training data poisoning to model inversion attacks, adversaries don't exploit weaknesses just in software—but also hardware and infrastructure.

Data center infrastructure is itself an attack surface. The problem is that many enterprises still rely primarily on locking down the perimeter. With AI, however, workloads are increasingly distributed and dynamic, spanning dozens of different environments—and rendering many older assumptions obsolete.

We often talk about AI as a workload—or workloads. But for adversaries, it's something more—a new attack surface to exploit. On the other hand, while AI introduces new infrastructure risks and attack vectors, it also offers new ways to guard against them. From anomaly detection to policy enforcement, AI can itself help identify and respond to threats.

This paper looks at the strategies IT leaders can use to tackle the unique security challenges that come with modernizing a data center for AI workloads.

Produced by TechTarget

1. Rethinking data center security in the AI era

AI hasn't just redefined data center performance requirements—it has reshaped how infrastructure needs to be secured. While traditional data architecture focused on securing applications, users, and the perimeter, AI workloads have introduced new complexities: They don't just run on the infrastructure—they change it. AI workloads are dynamic, distributed, and highly data-intensive, making them inherently more vulnerable to novel and emerging attack vectors.

As AI becomes more deeply embedded into everyday business operations, adversaries have started to target not just endpoints and networks, but AI systems themselves. Examples include model inversion attacks, training data poisoning, and exploiting underlying hardware configurations. These threats, while still in their infancy, are far from hypothetical. Recent years have seen an uptick in attacks targeting AI data pipelines, inference APIs, and shared compute environments.

Legacy security frameworks simply weren't designed for the scale of automation and data movement that AI deployments involve. Protecting these systems requires a major rethink of how enterprises protect their AI-ready data centers—from the silicon to the orchestration layer. Building a robust foundation starts with understanding the unique security challenges that AI introduces before implementing an adaptive, zero-trust framework that aligns with how AI workloads operate.

Key AI data center security challenges

- Monitoring systems and data at scale: AI systems consume and generate massive amounts of data and often lead to the proliferation of disparate tools that create more noise than clarity.
- Increasingly sophisticated AI-specific attacks: Many enterprises are not adequately prepared for threats like model inversion, adversarial prompting, or inference-time attacks.
- Traditional, perimeter-oriented security toolchains—designed for static VMs and monolithic applications—struggle to provide continuous visibility and policy enforcement for ephemeral, containerized AI workloads that span on-premises and multiple public clouds. The result: critical blind spots.

Step 1: Build toward zero trust

Zero trust assumes no implicit trust. Instead, every workload, device, and user must be authenticated and continuously verified. This is especially important in AI-specific use cases, where trust boundaries are fluid and workloads shift constantly among compute nodes, storage environments, and edge locations. Enterprises can enforce zero trust with tools such as micro-segmentation, identity-aware access controls, and encrypted communication paths across all data pipelines.

Step 2: Bake security protections into AI workflows

No longer can security be treated as a separate layer added post-deployment. Instead, it needs to be fully embedded throughout the AI lifecycle—particularly during data ingestion, model training, and inference. Best practices include input validation and output monitoring to prevent adversarial inputs; data lineage tracking to detect misuse; and helping secure MLOps pipelines with automated security checks. Embedding these checks into CI/CD pipelines and MLOps platforms can ensure early threat detection.

Step 3: Invest in AI-ready threat detection

Traditional intrusion detection and prevention (IDS/IPS) systems aren't designed to detect AI-specific anomalies, so new classes of tools have emerged to recognize the often subtle patterns of adversarial behavior or data leakage in AI workloads. One of the ironies of AI is that, despite the unique risks it introduces, it's also an extremely powerful tool for detecting AI-specific threats—such as carrying out behavioral telemetry and attestation to verify security feature integrity from the hardware layer up.

2. Safeguarding data in motion, at rest, and in use

Perimeter defenses and encryption of data at rest are still important parts of any comprehensive security strategy, but they are no longer sufficient alone. With AI workloads constantly ingesting, analyzing, and generating huge amounts of data across diverse infrastructure layers, data center operators must think beyond static protections. After all, every stage of the data journey—whether in motion, at rest, or in active use—represents a potential opportunity for compromise.

The number of potential single points of failure is immense. This is especially true for memory channels and I/O layers, which play an increasingly critical role yet are commonly overlooked in existing security strategies. As data flows across CPUs, GPUs, accelerators, storage arrays, and edge devices, even a momentary lapse in protection can leave sensitive training data or inference results exposed.

To protect against these threats, data security must start at the hardware level. Confidential computing features—such as AMD Secure Encrypted Virtualization (SEV) or hardware-level security features like trusted platform modules (TPMs)—enforce protections at the system's core. This approach helps ensure that, even during execution, AI models and data remain shielded from unauthorized access, thus bringing trust into the fabric of the infrastructure itself.

Key AI data movement challenges

- Data is more exposed when it's active: While encryption of data at rest is standard, or when it's in motion between remote environments, it's often vulnerable during real-time processing.
- I/O channels become a soft underbelly: With high-throughput demands and insufficient access controls, interconnects between compute and storage can become targets for adversaries.
- Multitenancy multiplies the risk: Hybrid and cloud environments often share resources across tenants. Without strong isolation primitives like hardware enclaves, one tenant's misconfiguration or compromise can leak data or hijack resources belonging to another, making it harder to enforce consistent security standards throughout the AI lifecycle.

Step 1: Encrypt memory, not just storage

AI workloads often operate in memory rather than on disk, making memory encryption essential for protecting sensitive data during processing. AI-ready CPUs and accelerators can mitigate these risks with hardware-level encryption and secure memory partitioning to isolate sensitive workloads into secure memory spaces that are inaccessible to unauthorized users. These processors can also provide rapid encryption and decryption—making them vital in latency-sensitive workloads like inference.

Step 2: Establish end-to-end trust

Confidential computing enables workloads to run in isolated execution environments known as trusted execution environments (TEEs). These shield data and code, thus protecting training data and model logic. They also help ensure verifiable trust with attestation to confirm the workloads are running in security-enabled environments. Moreover, the principle of least privilege (PoLP)—where every module has access only to resources and data it needs for its purpose—becomes enforceable at the hardware level.

Step 3: Protect the pipeline, not just the perimeter

AI workloads are rarely self-contained. They're pipelines that stretch across multiple infrastructure components both inside and outside the data center. For instance, a fraud detection system may ingest transaction data from an edge location, transform and evaluate it in real time using inferencing models, and pass outcomes to automated decision-making systems. That's why data needs to be encrypted across all internal services, API calls, and control planes—as well as external connections.

3. Building an integrated security architecture for AI

AI workloads are fast-moving, multi-layered, and highly automated. Moreover, the technology itself is advancing at a pace that's unmatched throughout the history of computing. These characteristics have created an environment where security features need to be built in by design and default across every layer of the system.

Until recently—and especially during the early days of AI experimentation and adoption—enterprises often made do with fragmented, ad-hoc security measures. But these can create blind spots and lead to operational complexity by introducing inconsistencies in how policies are applied. Instead, enterprises need a cohesive and adaptable security architecture that's designed specifically for the dynamic nature of AI workloads.

Continuous systemwide integration is the most viable approach. By bringing together observability, policy enforcement, and attestation across every layer of the tech stack, data center teams can build environments where AI workloads remain protected—without compromising on service availability, or agility, while minimizing the impact on performance.

Key challenges in AI security integration

- Fragmented toolchains responsiveness: Disparate systems often fail to communicate, resulting in siloed security practices that hinder end-to-end threat detection and could impact response time.
- Tech sprawl breeds complexity and risk: It stands to reason that the more standalone tools in use, the harder it is to consistently enforce security policies, leading to blind spots and reduced threat detection.
- Keeping pace with change: With new AI models being released and updated frequently, there's an increased risk of configuration drift that security teams can't stay ahead of.

Step 1: Build security in from the silicon up

Designing and integrating security frameworks from the silicon up helps ensure that every layer above inherits and validates security properties. For instance, secure boot processes can validate firmware and OS integrity before launch, while hardware-enforced isolation protects memory spaces and boundaries between tenants, applications, and users—including for virtual machines or containers. These features lay the groundwork for attestation and security telemetry, thus helping protect the whole stack.

Step 2: Centralize observability and control

Distributed workloads require distributed awareness. Yet without centralized observability and policy enforcement, enterprises risk blind spots and inconsistent policy enforcement. Modern observability tools provide systemwide telemetry collection to capture deep insights into performance, access, and behavior across all layers of the tech stack. Augmented with real-time analytics and correlation, these tools can allow you to identify threats quickly and home in on the root cause of any vulnerabilities.

Step 3: Verify everything with continuous attestation

Reinforcing the principles introduced earlier, zero trust is not a one-time setup—it requires continuous verification. This is vital given the dynamic nature of AI workloads: It helps prevent configuration drift and can ensure that inference engines aren't executing on containers that are outdated or may have been tampered with. Continuous attestation, combined with confidential computing, enables security features that are enforced in real time at every layer of the tech stack.

Choosing the right AI infrastructure partner

Modernizing the data center for AI isn't just about scaling compute—it's about reengineering trust across every layer of the tech stack. From adopting zero-trust principles to safeguarding data in motion, to embedding observability and enforcing continuous attestation, protecting AI workloads requires a planned architecturewide approach. As AI continues to advance, becoming more capable and more embedded into business operations, enterprises that treat security features as a dynamic, infrastructure-aware capability will be far better positioned to innovate safely and over the long term.

AMD helps enterprises design and deploy AI-ready data center infrastructure that doesn't compromise on performance or protections. With a portfolio that includes powerful CPUs, GPUs, and AI accelerators; hardware-layer security features like AMD Secure Encrypted Virtualization (SEV); and robust support for confidential computing, AMD gives you the building blocks needed to shape scalable and secure AI adoption. Whether you're modernizing an existing data center or building one from the ground up, AMD offers the tools and expertise you need to support AI workloads at every stage of the lifecycle.

Learn more about AMD's cutting-edge AI solutions