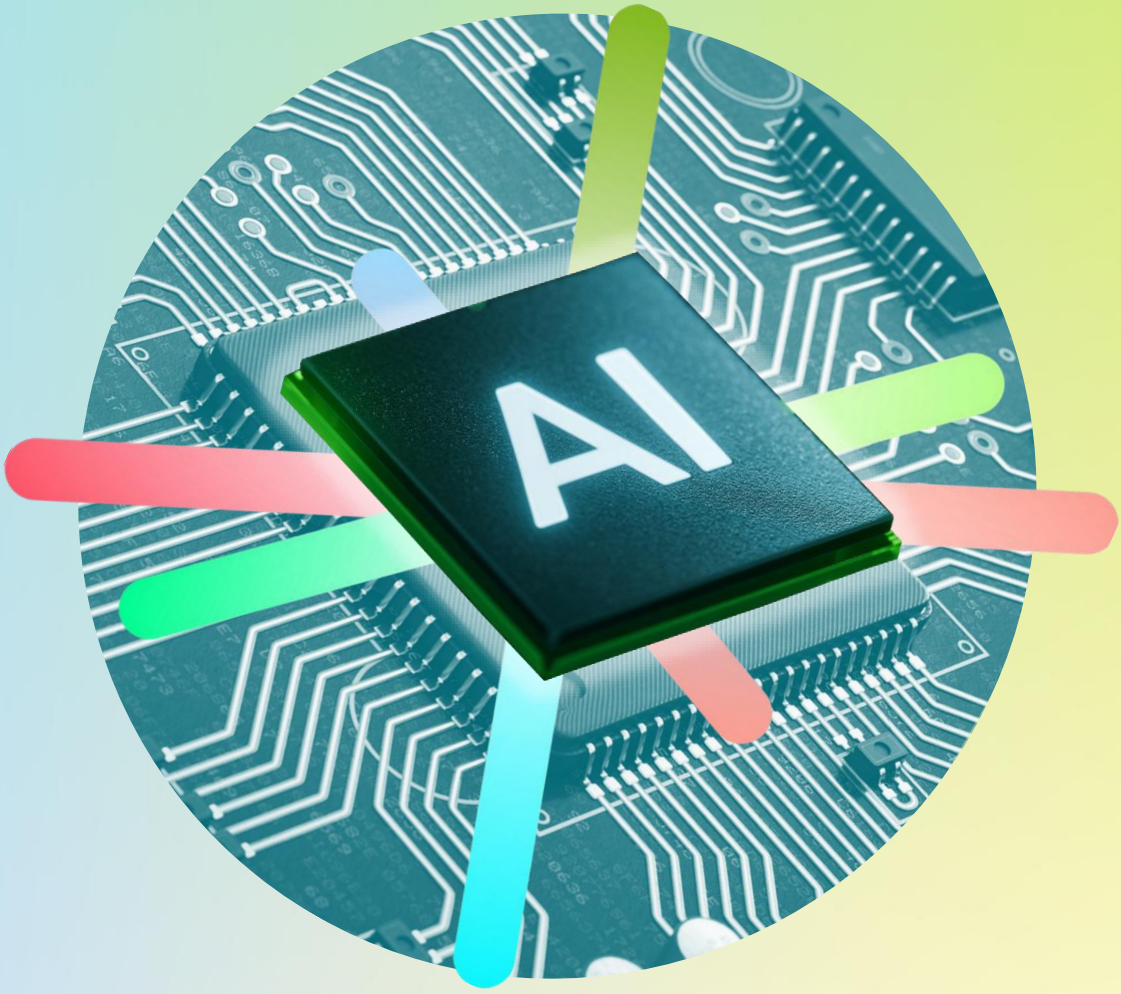
Deloitte. Red Hat intel



Sovereign AI-Ready Private Cloud

Leveraging the Next-Gen Infrastructure

CONTENTS

1. Introduction ***3***

2. Design Principles ***5***

3. Unlocking the Value of Sovereign AI-Ready ***10***

4. The Solution Stack ***14***

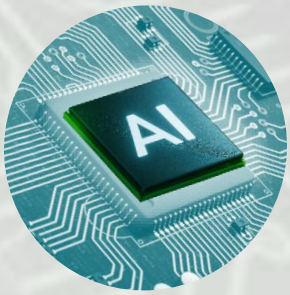
5. Delivering Sovereign AI-Ready ***22***

6. Final Remarks ***25***

Contacts ***26***

© 2026 Deloitte Technology & Transformation, S.L.U..

1. Introduction



The choices for building enterprise infrastructure are now deeply intertwined with the industrialization of AI.
As AI moves decisively from proof of concept to production, organizations face a fundamental reality: the cloud-first strategies that enabled early experimentation are not designed for the economics, performance demands, sovereignty and governance requirements of staggering and continuous AI inference at scale.

AI is becoming a persistent, business-critical capability embedded directly into core operations, decision-making, and customer interactions.

While public cloud platforms remain essential for elasticity, innovation, and rapid experimentation, many enterprises are discovering that **AI at scale introduces structural tensions with unpredictable and escalating costs**, sovereignty constraints, latency limitations, and reduced operational control. These pressures are forcing CIOs and CTOs to re-examine where AI workloads should run, and under what conditions.

At the same time, the underlying **compute architecture** is undergoing a **profound shift.**

AI workloads are accelerating the move from homogeneous, general-purpose infrastructure toward **heterogeneous environments** that combine CPUs, GPUs, and emerging accelerators. Importantly, not all AI workloads require stringent dedicated hardware. A substantial share of enterprise AI capabilities can run efficiently on modern, AI-optimized CPUs, provided the platform is designed for **intelligent orchestration, data sovereignty and performance efficiency**.

The strategic question is no longer “cloud or on-premises,” but how to deliberately **right-size compute across environments** based on cost, performance, sovereignty, and resilience.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

Regulatory mandates, geopolitical uncertainty, and heightened concern over intellectual property (IP) are compelling organizations to regain control over where data resides, how it is processed, and who operates and governs the underlying infrastructure.

For mission-critical and regulated workloads, multi-tenant public cloud models are often subject to extraterritorial jurisdictions and opaque operational boundaries. Sovereignty is no longer a compliance addition; it has become a strategic prerequisite for scaling AI responsibly and sustainably.

In this context, the Sovereign AI-Ready Private Cloud approach represents a foundational pillar of the AI-powered economy. Rather than positioning private cloud as an alternative to public cloud, this model establishes it as a sovereign anchor within an AI-ready hybrid architecture.

By combining cloud-native platforms, AI-optimized infrastructure, and open hybrid orchestration, organizations can run high-volume AI workloads in private cloud with greater cost discipline, lower latency, and stronger governance, while retaining the flexibility to leverage public cloud services where they deliver clear value.

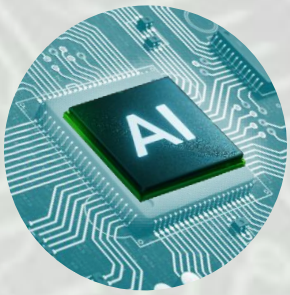
Organizations that invest in a sovereign, and inference-optimized private cloud foundation that is powered by open platforms, interoperability and purpose-built silicon, will be better positioned to control costs, meet regulatory expectations, and scale Hybrid AI with confidence.

*This white paper presents a joint viewpoint from **Deloitte, Intel, and Red Hat** on how enterprises can build a **resilient, future-ready private infrastructure foundation** for the computation renaissance ahead.*

*It outlines the architectural principles required to support AI at scale. It also examines how advances in compute architectures combined with hybrid orchestration platforms, enable a **sovereign, disciplined and economical deployment of AI**.*

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

2. Design Principles



The Sovereign AI-Ready Private Cloud is grounded in a clear and enforceable set of design principles that translate strategic intent into architectural and operational verifiability.

Infrastructure AI-readiness, software lifecycle automation, and governance controls are enforced by five principles for AI workloads to be deployed securely, cost-effectively, and at scale.

Design principles of Sovereign AI-Ready

Transversal principle

Digital Sovereignty

Ensure the autonomy and control of the different layers of the solution.

- **Openness** – Based on open standards and open-source technologies to ensure flexibility and interoperability
- **Consistency –** Unified management and policy enforcement across all environments
- **Performance –** Hardware and software co-engineered to maximize workload efficiency
- **Security –** Defence–in–Depth approach from silicon to application
- **Sustainability –** Optimized for energy efficiency and resource utilization

These principles define how the environment is designed, governed, and operated to support ***industrial-scale AI inference, regulated data, and hybrid workloads****—while remaining open, compliant, and future-ready.*

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

Digital Sovereignty

Transversal principle

Digital sovereignty is the overarching principle that shapes and constrains all others. It is the ability of an organization to retain independent, enforceable control over its digital assets: data, workloads, AI models, operational processes, and the underlying technology stack, across their entire lifecycle and across all deployment environments.

This control is essential to:

- Comply with industry- and regional-specific regulations and data residency requirements
- Reduce exposure to geopolitical, legal, and supply-chain risks
- Protect intellectual property and proprietary algorithms over the long term

As a transversal principle, digital sovereignty requires that every layer of the cloud—hardware, platform, data services, AI pipelines, and operations—can be governed under clearly defined, locally enforceable rules.

It also demands **auditability and transparency:** the ability to demonstrate where data resides, how it is processed, who accessed it, and how decisions are made across the stack.

Yet, data sovereignty alone is not sufficient. Even when global cloud service providers offer enhanced sovereign controls, they can remain subject to the legal obligations of their home jurisdictions. A Sovereign AI-Ready Private Cloud therefore must also support deployment on server processors and sovereign infrastructure that can be operated by local CSPs.

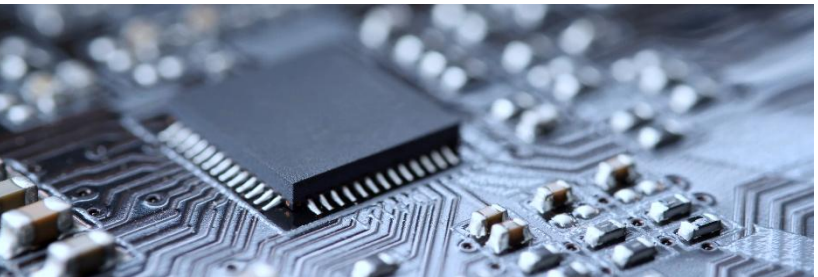
As discussed later, this requires **architectural and operational consistency** so that workloads can be moved between these environments without losing security or governance guarantees. This sovereign infrastructure gives organizations the ability to select a local CSP if necessary.

The proposed architecture inherently accommodates **multi-cloud and hybrid-cloud environments**, ensuring seamless interoperability across deployment models. Local CSPs are no exception in this regard.

The critical requirement is that these local CSPs enable access to and utilization of Intel's key accelerator technologies essential for digital sovereignty (e.g. Intel® Advanced Matrix Extensions (Intel® AMX), Intel® Trust Domain Extensions (Intel® TDX), Intel® QuickAssist Technology (Intel® QAT)).

To guarantee this capability, local CSP offerings should be built on the latest generation of **Intel® Xeon® 6 processors**, with these technologies fully activated and supported. Once these foundations are ensured, the modular framework architecture can integrate local CSPs in the same way it does with global providers, maintaining architectural consistency, performance efficiency, and compliance with sovereignty objectives.

Digital sovereignty extends beyond data sovereignty and technical control into enforceable legal control. Organizations are not locked into a single provider or jurisdiction and can dynamically align workload placement based on risk tolerance, regulatory posture, and business requirements.



 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

Openness

Freedom of choice

Openness is fundamental to sovereignty. A sovereign architecture cannot rely on opaque dependencies, proprietary interfaces, or single-vendor control without introducing structural risk.

This principle ensures freedom of choice over time by mandating:

- **Open standards** and open-source–based platforms
- Modular, **composable architectures**
- Transparent, documented **APIs and data formats**

Such openness allows components to be replaced, upgraded, or augmented for **reversibility and resiliency** without re-architecting the entire system. It enables organizations to introduce new cloud and AI technologies, providers, or deployment models—on-premises, edge, or sovereign cloud extensions—while preserving architectural integrity and reducing governance overhead to leverage hybrid AI.

Consistency

Unified management

Consistency is also key as a principle to leverage hybrid AI. Without it, AI stacks and environments fragment into operational silos, each with its own **tools, control planes and security postures**—rendering seamless governance impossible and sovereignty unenforceable at scale.

This requires a single, coherent **management and policy framework** that spans across on-premises infrastructure. Policies governing data residency, access control, workload placement, security posture, and lifecycle management must work seamlessly and enforced consistently everywhere workloads run.

Sovereign AI-Ready private cloud provides a common **orchestration, automation, and policy layer across different environments**, while hardware-based infrastructure complement with the hardware-level capabilities—such as security features, performance characteristics, and workload isolation—to be exposed and governed in a uniform way. Together, they enable organizations to **manage diverse compute resources as a single, sovereign estate** rather than a collection of disconnected environments.



As outlined next in the stack solution section, architectural and operational consistency reduces complexity and enables auditable enforcement of sovereignty, compliance, and performance.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..



Security

Architected by design

Security cannot be bolted on at the perimeter or delegated to a single control layer. It must be **engineered end to end**—spanning silicon, firmware, platform, and applications—to ensure that trust in the private cloud is continuous, verifiable, and resilient against evolving threats.

This principle mandates a **defense-in-depth security architecture** that begins at the hardware level and extends through the cloud-native platform to the AI and application layer. At the hardware-level, capabilities provide foundational trust through **hardware-rooted security**, including secure boot, firmware integrity, workload isolation, and **confidential computing technologies that protect data not only at rest and in transit, but also in use.** These capabilities establish a trusted execution baseline that higher layers can rely on.

Sovereign AI-Ready private clouds build on this foundation by **enforcing consistent security controls across the runtime environment**, identity and access management, network segmentation, policy-driven isolation, and continuous vulnerability management—across on-premises, edge, and hybrid deployments.

Security policies are embedded directly, ensuring they are applied uniformly as workloads move and scale.

By integrating hardware-level trust with cloud-native security and governance, organizations can demonstrate—not just assume—security and compliance. This approach enables sovereign AI workloads to **operate with confidence in regulated and mission-critical contexts**, even in the presence of insider threats, compromised components, or hostile environments.

Security, in this model, becomes a measurable and enforceable property, not a point solution or overlooked capability due to the implicit physical control of the private infrastructure.

Performance

Workload efficiency

Performance is not defined by peak benchmarks alone, but by how efficiently workloads consume resources under real, continuous operating conditions. As AI inference becomes persistent and business-critical, performance must be engineered across the full stack rather than optimized in isolated layers.

This principle requires tight **co-engineering between hardware and software** to ensure that each workload runs on the most appropriate compute resources with minimal waste.

The processor portfolio—ranging from AI-optimized CPUs to GPUs and accelerators—must provide the **architectural flexibility** to match performance characteristics to workload needs, avoiding unnecessary dependence on specialized hardware where it does not deliver proportional value. Features such as advanced vector extensions, memory bandwidth optimization, and workload isolation must enable high-throughput, low-latency inference at scale.

Private Cloud must translate these hardware capabilities into operational efficiency through the efficient **orchestration of workloads and resource awareness**. By aligning workload requirements with underlying hardware characteristics, the solution must ensure predictable performance, high utilization, and consistent service levels across hybrid environments.

The result is performance with discipline: organizations can meet demanding latency and throughput requirements while controlling costs and energy consumption.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..



Sustainability

Energy optimized

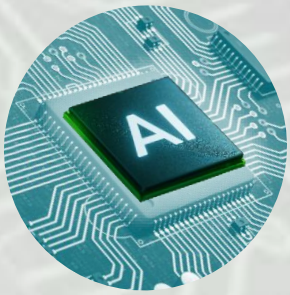
Sustainability is not an abstract ambition; it must be embedded by design in the private infrastructure. As inference scales and becomes continuous, the ability **to deliver performance without proportional increases in energy consumption** or footprint becomes a critical determinant of long-term viability.

Hardware-level capabilities must focus on performance-per-watt, power-aware processor design, and hardware-level telemetry to enable organizations to extract more value from each unit of energy consumed. By matching workloads to the most efficient compute resources (often AI-optimized CPUs rather than over-provisioned accelerators) enterprises can significantly **reduce power draw while maintaining required performance levels**.

Private cloud translates these efficiencies into operational outcomes through **intelligent scheduling, rightsizing, and policy-driven resource management**. Workloads thus can be consolidated, scaled dynamically, and placed where they deliver the best balance of performance and energy efficiency across on-premises and hybrid environments. Unified visibility into resource consumption enables continuous optimization rather than static capacity planning.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

3. Unlocking the Value of Sovereign AI-Ready



AI-ready compute infrastructure, open platforms, and hybrid orchestration work together as an integrated system to industrialize hybrid AI while retaining control over data, workloads, and decision rights.

To realize the value of Next-Generation Sovereign AI-Ready Private Cloud in practice, organizations must focus on four key dimensions to translate architectural intent into real measurable business and operational outcomes.

1

FIRST

Governed Sovereignty & Trust

Sovereignty is realized when control is enforceable across the entire stack. Hardware-rooted trust, confidential computing, and platform-level policy enforcement ensure that data, models, and workloads remain under clearly defined jurisdictional and operational control. Open-source transparency enables auditability which is as important as the control mechanisms to enforce compliance and long-term IP protection, turning sovereignty from a constraint into a strategic enabler.

KEY VALUE DRIVERS

- *Confidential computing.*
- *Open architectures for MLOps.*
- *Orchestrators, enforcement of policies for data residency, security and compliance.*

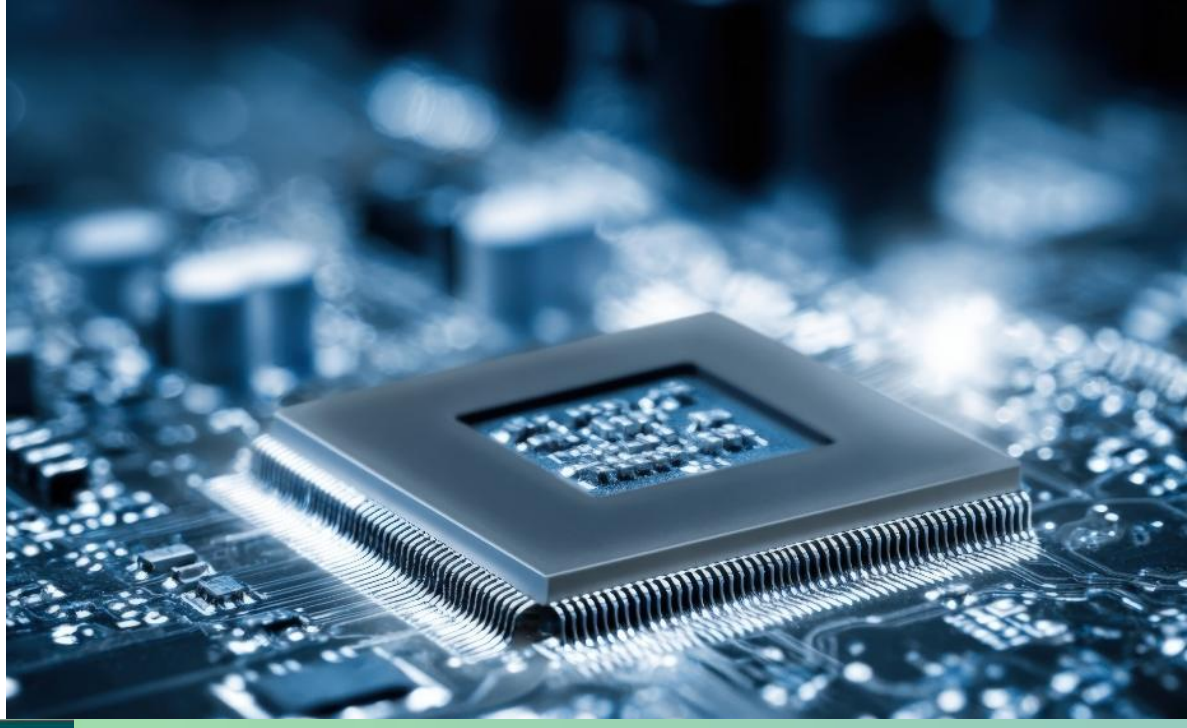
STRATEGIC IMPACT

- *Reduced operational complexity, lower audit risk*
- *AI deployment on sensitive data while meeting regulatory, geopolitical, and IP-protection requirements*

ILLUSTRATIVE KPIS

- *Data Residency Compliance Rate & Policy Enforcement Coverage*
- *Regulatory Risk Exposure Reduction*

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..



2

SECOND

Cost Efficiency & Predictable AI Performance

Value is created when AI workloads run on the right compute for the job. AI-optimized CPUs, combined with accelerators where required, enable high-volume inference with predictable latency and cost discipline.

Private cloud translates hardware capabilities into operational efficiency through intelligent orchestration, workload awareness, and continuous optimization, eliminating over-provisioning and cloud cost volatility.

KEY VALUE DRIVERS

- *AI Accelerators (CPU & GPU based)*
- *AI-powered observability for driving continuous telemetry into intelligent scheduling and load optimization*

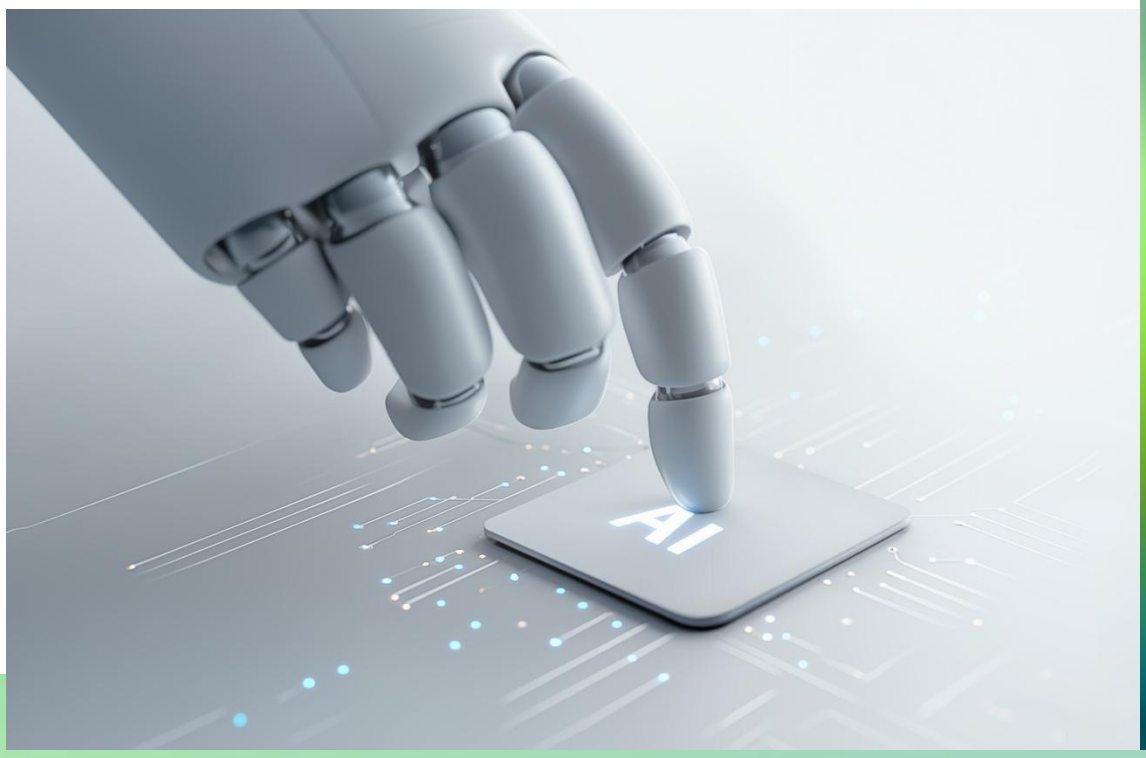
STRATEGIC IMPACT

- *Optimal latency and workload awareness*
- *Predictable costs, and lower TCO for mission-critical AI workloads*

ILLUSTRATIVE KPIS

- *Inference Latency, Throughput per Node*
- *Compute Utilization Rate, Cost per Inference*
- *Cost per Inference, AI Infrastructure Cost Reduction*
- *GPU Overprovisioning Ratio, avoided GPU CAPEX / OPEX*

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..



THIRD

Hybrid Agility & Operational Resilience

Sovereign Private Cloud does not replace public cloud; it anchors hybrid operations. Consistent APIs, common orchestration, and unified management enable workloads and data to move across private, edge, and public environments in a governed way.

This reversibility strengthens resilience, reduces dependency risk, and allows enterprises to adapt to changing regulatory, economic, or business conditions without re-architecting.

KEY VALUE DRIVERS

- *Open, industry-standard AI frameworks and APIs*
- *End-to-end observability across infrastructure, platforms, and AI workloads*
- *Consistent CI/CD pipelines across hybrid environments*
- *ITSM integration and workload orchestration powered by automation to drive AIOps*

STRATEGIC IMPACT

- *Higher availability*
- *Faster incident response and root cause analysis*

ILLUSTRATIVE KPIS

- *Workload Portability Rate*
- *Mean Time to Redeploy across environments*
- *Recovery Time Objective (RTO), Availability of AI Services*
- *Downtime Cost Avoidance, Avoided Migration Cost*

3

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

FOURTH
Sustainability & Long-Term Viability

The combination of right-sized compute, high utilization, and energy-efficient hardware improves both total cost of ownership and environmental performance. Open platforms avoid lock-in and preserve future flexibility, while consumption-based operating models align spend with actual business value. Sustainability becomes a built-in outcome, not an afterthought.

KEY VALUE DRIVERS

- *AI-optimized CPUs - GPUs and accelerators for hardware-level energy efficiency (vector extensions, etc.)*
- *Open platforms for MLOPs with modular architectures for reusability*
- *Workload scheduling and resource governance for optimal utilization*

STRATEGIC IMPACT

- *Lower energy consumption and reduced carbon footprint per AI workload*
- *Improved capital efficiency and extended infrastructure lifespan*

ILLUSTRATIVE KPIS

- *Performance per Watt (inferences/sec/W)*
- *Carbon Intensity per Inference (kg CO_2e)*
- *Token Efficiency (Tokens per second/ throughput, Watts per token/second)*

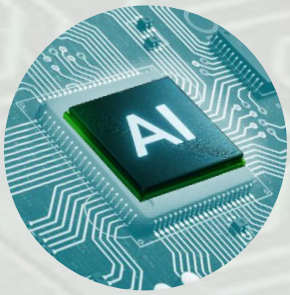
Together, these drivers define how organizations capture measurable value from a Sovereign Private Cloud.

*By combining **Intel's purpose-built silicon and security capabilities with Red Hat's open, hybrid cloud platforms**, enterprises can move beyond fragmented infrastructure decisions and establish a **scalable, governable foundation for AI**—one that balances innovation with control, and performance with digital trust.*

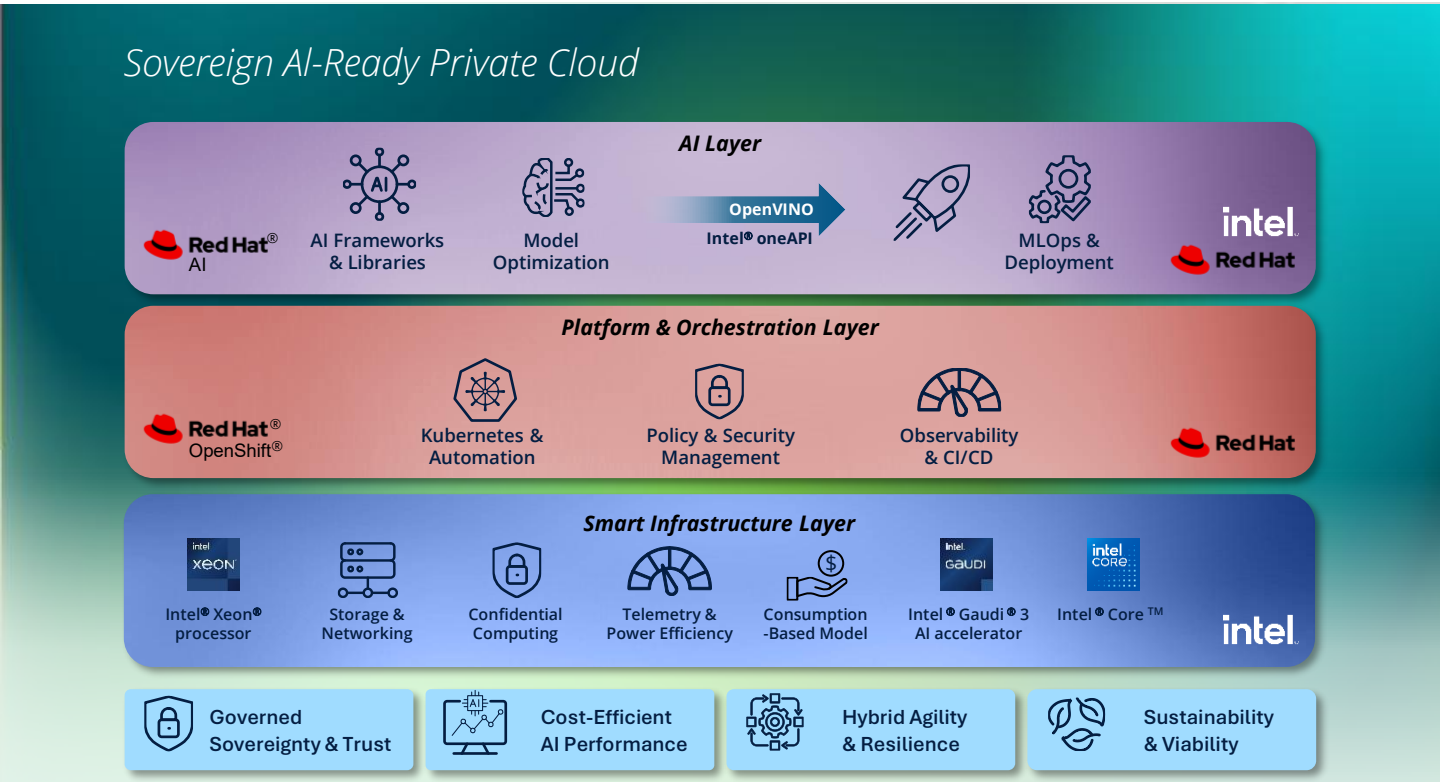


 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

4. The Solution Stack



Organizations can leverage a seamless blueprint that unifies compute, orchestration, and AI intelligence into a governed platform for scalable and sovereign AI for the Enterprise.

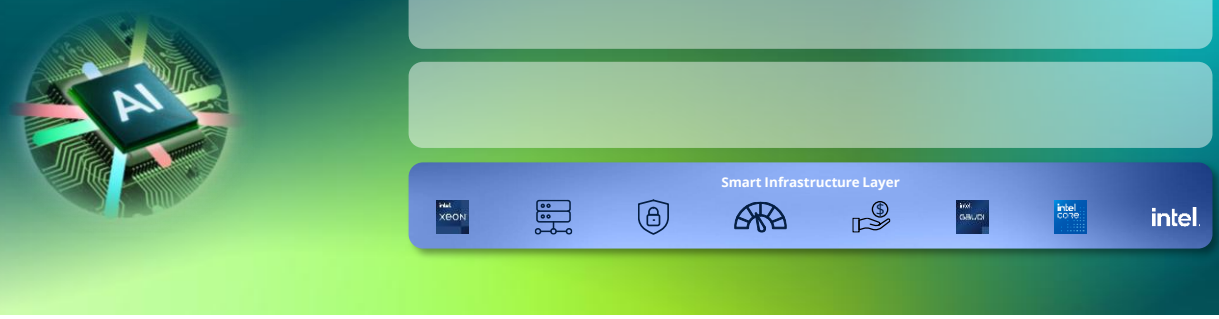


The Sovereign AI-Ready Private Cloud is structured around three integrated layers built on Intel-optimized hardware and Red Hat open platforms: **Smart Infrastructure, Platform & Orchestration, and AI**

The full strategic value lies in its combination: **infrastructure designed for consumption-based private cloud**, along with a hybrid cloud-native platform that abstracts and automates the infrastructure, and an AI layer that leverages it with open frameworks to capture the benefits of hardware acceleration.

This multi-layered blueprint means that the forementioned value drivers are not addressed by isolated features but **enforced end-to-end across the stack**. By design, this stack transforms discrete technologies into a coherent, sovereign, AI-ready private cloud that **supports scalable AI workloads** with predictability, governance, and resilience.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..



Smart Infrastructure Layer

The elastic & sovereign compute backbone

It forms the sovereign, programmable backbone of the cloud. It is designed for **consumption-based operation** (elastic, metered, and fully under organizational control) rather than traditional, static on-premises infrastructure.

Built on **Intel Xeon** processors, along with modular storage and networking components, it is dimensioned for data-intensive and AI workloads while maintaining **jurisdictional control and security**. Intel's power-efficient processors, telemetry, and AI acceleration features ensure predictable performance and energy efficiency.

This layer is intentionally modular: multiple infrastructure vendors and components can be integrated, provided they comply with sovereignty, openness, and operational standards. From the perspective of upper layers, all resources are abstracted into a **single sovereign resource pool** that behaves like internal IaaS: compute, storage, and network are provisioned on-demand, scaled elastically, and consumed based on actual usage.

For instance, to optimize AI pipelines, organizations can turn to the latest Intel Xeon processors with **Intel AMX**, a built-in AI accelerator that is designed to balance inference, the most prominent use case for a CPU in AI applications, with more capabilities for training.

intel

- ***Intel® Advanced Matrix Extensions (Intel® AMX)***
- ***Intel® Advanced Vector Extensions 512 (Intel® AVX-512)***
- ***Intel® Software Guard Extensions (Intel® SGX)***
- ***Intel® Trust Domain Extensions (Intel® TDX)***
- ***Attestation Service***

More information about Intel products

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

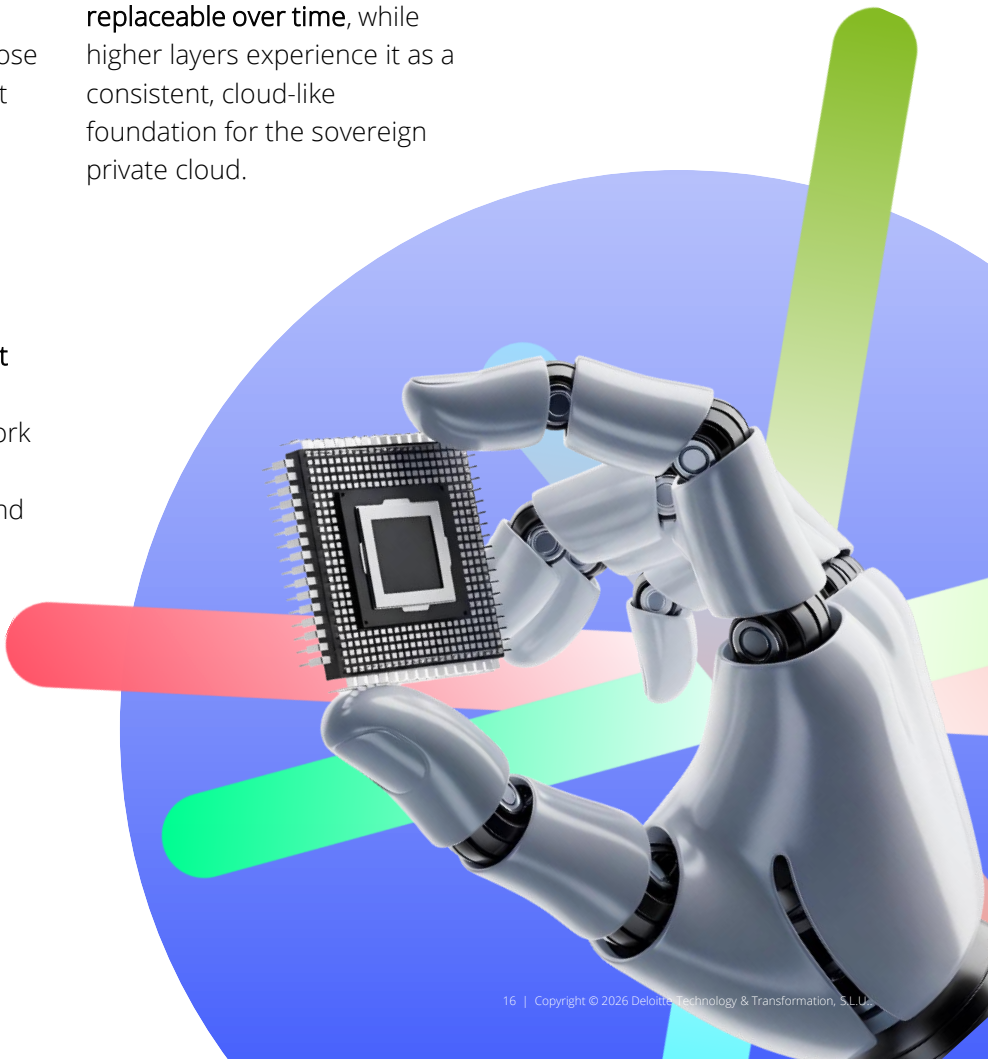
Benefiting from Private IaaS

- The **Private IaaS** capability of Sovereign AI-Ready Private Cloud can be realized through modern consumption-based infrastructure models that deliver **enterprise data center resources as a service**. These models enable organizations to operate on-premises or sovereign-hosted infrastructure using pay-per-use economics, aligning infrastructure costs directly with actual workload consumption.
- Such approaches support **elastic provisioning of compute, storage, and networking resources** while providing real-time visibility into utilization and cost. By shifting from CAPEx capacity planning to OPEx consumption models, enterprises can improve financial flexibility, reduce overprovisioning, and optimize resource efficiency.
- This **IaaS operating model is designed to integrate with open orchestration platforms and hybrid cloud architectures**, allowing private sovereign environments to behave with the operational simplicity, governance control, and scalability characteristics of cloud services while maintaining jurisdictional and security requirements.

These private IaaS solutions can be incorporated seamlessly with Intel as building blocks beneath the platform, provided they expose the necessary APIs and meet sovereignty and security requirements.

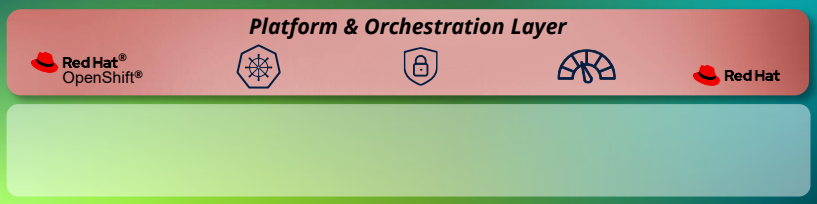
From the upper layers' perspective, all of them are abstracted into a single sovereign resource pool that behaves like internal IaaS: compute, storage and network can be provisioned on demand, scaled elastically and billed according to actual usage instead of large upfront CAPEX.

This ensures that the Smart Infrastructure layer remains sovereign, efficient and replaceable over time, while higher layers experience it as a consistent, cloud-like foundation for the sovereign private cloud.



 Copyright © 2026 Deloitte Technology & Transformation, S.L.U.





Platform & Orchestration Layer

The operational backbone for sovereign hybrid

It turns the Smart Infrastructure into a cloud-native, developer-friendly environment. Built on **Red Hat OpenShift®** and **Red Hat Enterprise Linux CoreOS**, it provides a consistent, governed platform for deploying, scaling, and managing workloads across private, edge, and hybrid extensions.

Key capabilities for sovereignty include:

- **Standardized orchestration** via **Kubernetes** and open APIs for consistent deployment and governance
- **Infrastructure-as-code** provisioning and lifecycle management of clusters, ensuring repeatable, auditable deployments
- **Integrated observability**: logging, metrics, and tracing across infrastructure, platforms, and workloads
- **Policy enforcement**: security, networking, compliance, and configuration defined once and applied everywhere

This layer ensures operational consistency, reversibility, and hybrid agility, allowing workloads to move seamlessly across environments while maintaining control, auditability, and compliance. **Red Hat Ansible® Automation Platform** and **CI/CD pipelines** enable teams to deliver AI and applications at cloud speed without compromising governance.

By standardizing Kubernetes and open-source components, this layer enforces the principles of openness and flexibility while ensuring that operations teams work with a single control plane rather than a patchwork of tools.

Operationally, the layer offers **end-to-end lifecycle management**: infrastructure is consumed via infrastructure-as-code, clusters are provisioned and upgraded in a controlled manner, and policies for security, networking, observability and configuration are defined once and applied everywhere. Integrated logging, metrics and tracing give full visibility into workloads and platform health, while **GitOps** and **CI/CD pipelines** enable repeatable, auditable deployments that align with sovereignty and compliance requirements.

Red Hat

OpenShift®

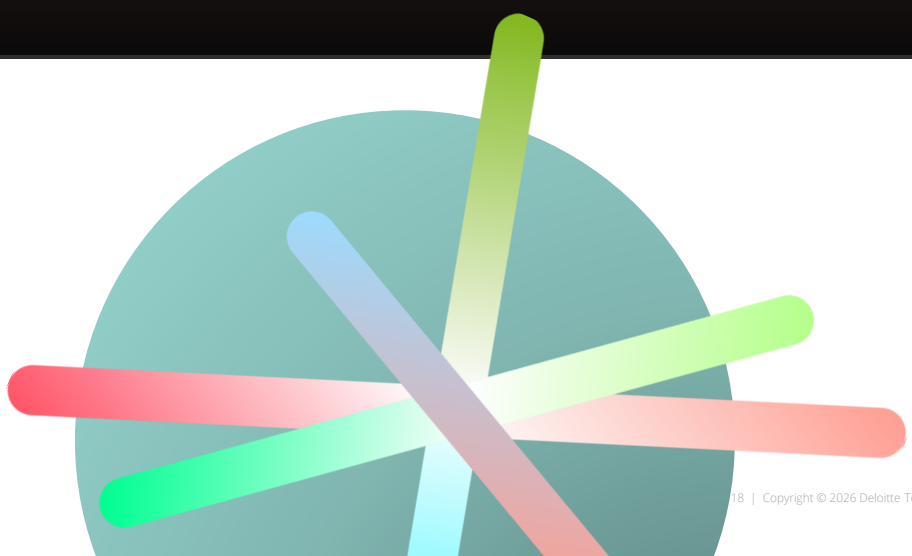
OpenShift® AI

More information about Red Hat products

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U.

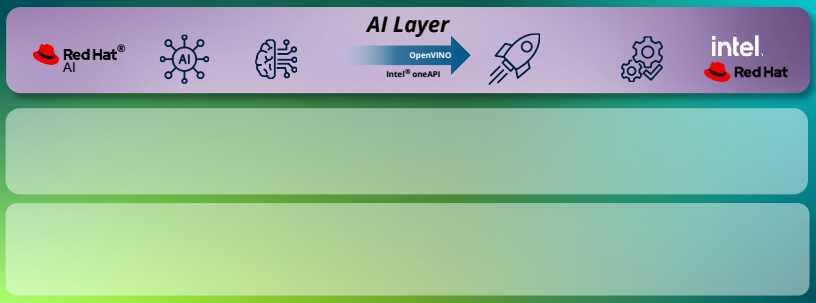
A Robust Hybrid Sovereign Model

- Using **Red Hat Build of Trustee** together with **Intel Trust Domain Extensions (TDX)** and **Intel Trust Authority (ITA)** creates a confidential computing model where workloads run inside hardware-isolated environments and can only access secrets after their integrity is cryptographically verified. Intel TDX protects the virtual machine or container memory from the hypervisor and other tenants, while Trustee organizes the remote attestation process and Intel Trust Authority provides **remote verification of attestation evidence proving the hardware and platform are genuine and correctly configured**.
- Red Hat Build of Trustee acts as the policy gatekeeper that releases encryption keys or sensitive data only when this attestation succeeds. In this context Trustee most likely will be installed in the companies Trusted Compute Base (TCB) as its most secure environment. The TCB is usually deployed in an on-premise installation. The execution of the application can be done in a separate cloud-based installation. The result is a **hybrid cloud implementation where data and secrets remain protected from cloud system administrators** with a zero-trust architecture.
- This architecture supports digital sovereignty requirements because organizations maintain control over when and where sensitive workloads can run and who can access their data, even in shared or foreign cloud environments.
- Within **Red Hat OpenShift**, this model is typically implemented through **Confidential Containers**, where Kubernetes pods run inside confidential VMs executed on a hardware backed confidential enabled CPU like Intel TDX; Trustee manages attestation and secrets, **enabling secure multi-tenant container platforms** that meet strict regulatory and sovereignty constraints while preserving the operational model of cloud-native workloads. Running applications in a confidential environment only needs minor changes to the deployment to be able to be executed inside a confidential container. No application code changes are required to activate a trusted domain enabled by Intel TDX inside a VM running on Red Hat platform. This is a key enterprise advantage– **existing applications can gain confidential computing protection without any refactoring or effort**. As AI/ML workload is getting more common, another requirement is to integrate **Sovereign AI**. With the availability of GPUs in a confidential mode into the architecture, it is possible to run sensitive AI pipelines in the cloud without exposing models or data to the underlying infrastructure or to the cloud provider.



 | Copyright © 2026 Deloitte Technology & Transformation, S.L.U..





AI Layer

The platform for sovereign AI at scale

It converts the sovereign private cloud into a complete, production-grade AI platform focused not only on technical capability, but on commercial viability, **performance efficiency,** and **predictable outcomes**. Its objective is to ensure that **AI workloads deliver measurable business value** while operating within controlled cost, performance, and energy envelopes.

Private Accelerated AI Inferencing

In addition to isolation, Intel platforms enable strong AI inference performance within confidential environments.

Intel AMX provide built-in acceleration for matrix and tensor operations , allowing many enterprise AI workloads to run efficiently on CPUs inside Intel TDX confidential VMs.

This enables organizations to combine **sovereignty, protection of sensitive data and models**, and predictable CPU-based inference economics—without requiring dedicated accelerator infrastructure for a large class of AI workloads.

Extending Confidentiality across Distributed Systems

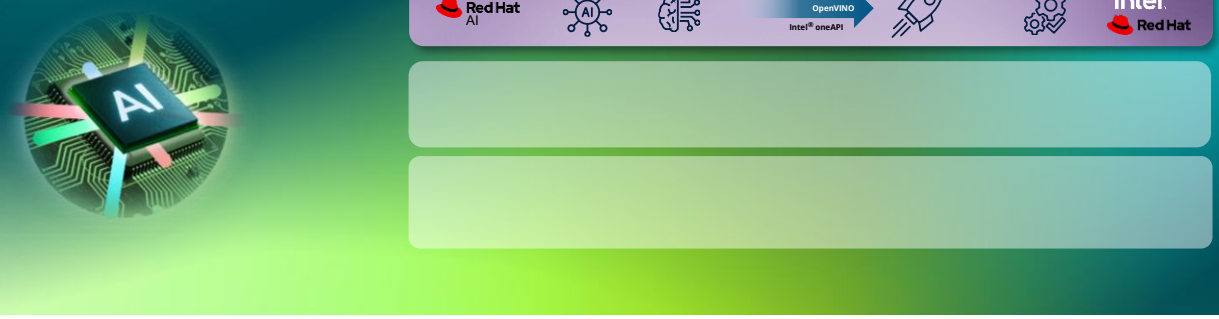
Real-world sovereign systems are rarely confined to a single host. AI inference pipelines often depend on distributed components—accelerators, storage, networking, and specialized services. **Intel® Trust Domain Extensions Connect (Intel® TDX Connect)** extends the confidential computing model by **enabling confidential trust boundaries beyond the CPU** and across attached devices and system components, supporting architectures where **sensitive workloads require end-to-end protection** across the platform

From a sovereignty standpoint, **Intel TDX Connect** helps address:

- **Data sovereignty** by reducing exposure of sensitive data as it moves between compute and connected components
- **Operational** by strengthening isolation across the broader platform, not only inside the VM boundary
- **Assurance** by enabling more comprehensive verification of the environment supporting a confidential workload
- **Technological** by enabling a portable, vendor-independent AI architecture built on open standards and interoperable, open-source ecosystems with end-to-end trusted execution across heterogeneous infrastructures

This capability is particularly relevant for sovereign AI deployments that require **strong controls across the entire inference path**—where confidentiality and integrity must be preserved not only in compute, but throughout the system interactions that surround it.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..



While **Intel TDX Connect** represents a forward-looking capability, elements of this model can already be realized today. The diagram below shows the various options available for Confidential AI on Intel Xeon processors.

A Sovereign Foundation for AI at Scale

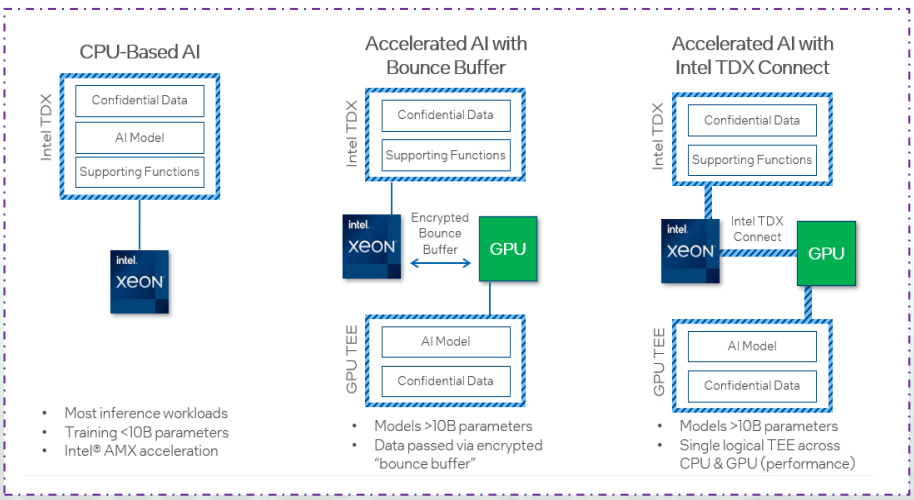
Confidential Computing on Intel technologies significantly strengthen sovereignty by enabling a hardware-rooted, verifiable trust model for sensitive and regulated workloads—particularly those involving valuable data and AI intellectual property.

Intel TDX provides isolated execution for confidential VMs, Intel's attestation capabilities deliver measurable assurance that systems operate as intended, and Intel TDX Connect will support more realistic sovereign architectures by extending confidentiality beyond a single host.

Together, these capabilities help enterprises scale AI with greater control, reduced trust in privileged layers, and stronger evidence for compliance—turning sovereignty from a policy aspiration into an enforceable technical reality.

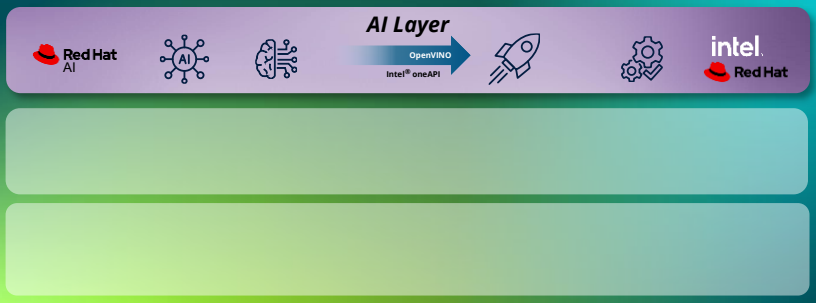
The **AI layer supports both CPU- and GPU-based vLLMs** via open-source frameworks and libraries (such as TensorFlow, PyTorch, Spark, Pandas, and XGBoost and vLLM for high-throughput LLM serving) directly on the Platform & Orchestration layer, enabling data scientists and engineers to use familiar tools while **keeping all data, models, and inference pipelines within the sovereign perimeter**.

With the staggering adoption of LLMs and agentic AI systems, the AI Layer address workloads that involve automated reasoning, multi-step orchestration, and domain-specific intelligence.



 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..





Leveraging Agentic AI

Agentic AI introduces the capability to route tasks dynamically across specialized models, reducing reliance on a single massive general-purpose model and improving efficiency, cost-effectiveness, and SLA adherence.

The key design considerations of the AI-Layer for LLMs include:

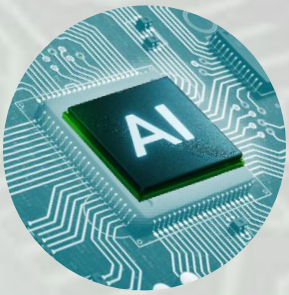
- TCO-Driven Model Placement: Utilizing vLLM's advanced memory management, LLMs with fewer than 13 billion parameters can often run efficiently on CPU-based infrastructure. Leveraging CPUs for these models significantly reduces infrastructure costs while maintaining performance for many enterprise AI tasks.
- Smarter Workflows through Agentic AI: By enabling models to interact, reason, and delegate subtasks autonomously, the AI Layer can implement agentic orchestration patterns.
- Domain-Specific Expertise: Smaller, specialized models can serve as "experts" in their respective domains. The agentic architecture coordinates these models to deliver higher quality outputs with lower token usage and reduced latency, improving both TCO and user experience.
- Resource-Aware Optimization: Monitoring and orchestration tools track model execution, token consumption, throughput, and latency, enabling dynamic adjustment of workloads between CPUs and accelerators to meet SLA requirements while minimizing cost and energy usage

The result is an AI platform powered by open, hardware-aware AI runtimes and toolchains from Intel (oneAPI, OpenVINO, ONNX, etc.) where performance and effectiveness are governed by outcomes, not best-effort aspirations.

In combination with vLLM—the high-performance LLM serving engine and the MLOps capabilities provided by Red Hat OpenShift AI, organizations can scale AI with predictable cost, enforceable SLAs, and clear alignment between business value and infrastructure consumption, while preserving openness, sovereignty, and freedom from proprietary AI services.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

5. Delivering Sovereign AI-Ready



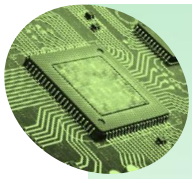
The Sovereign AI-Ready Private Cloud is more than a future-ready technology stack, it is a joint transformation program that brings together sovereignty, AI readiness, and predictable economics. ***Through the combined capabilities of Intel, Red Hat, and Deloitte,*** *organizations can translate complex regulatory, operational, and infrastructure management challenges into enterprise value with a scalable, secure, and AI-ready sovereign private cloud environment.*

ADVISE for a consistent sovereign strategy

Defining a sovereign cloud strategy, including target architecture, workload placement across environments, regulatory alignment, and a clear business case.

The approach is tailored to industry-specific requirements, ensuring that sensitive workloads meet regulatory and operational imperatives.

The outcome is a clear AI-Ready Sovereign roadmap that links sovereignty requirements directly to clear architectural intent and operational choices.



Outcomes

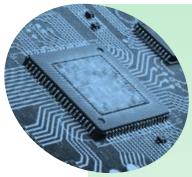
- *Unified Operations & Control Plane*
- *Integrated Observability & AIOps*
- *Security, Sovereignty & Compliance Operations*
- *SLA & Performance Management*
- *Cost, Energy & Carbon Optimization*

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

IMPLEMENT for an effective & sovereign deployment

Delivering a **cohesive, AI-ready platform** built on Intel's infrastructure, Red Hat Enterprise Linux and OpenShift for cloud-native operations, and open AI frameworks.

Workloads are migrated and modernized, AI pipelines are industrialized, and performance- and cost-aware placement is enabled across CPUs, GPUs, and accelerators, all within a sovereign perimeter.

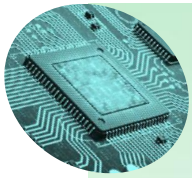


Outcomes

- *Sovereign Infrastructure Foundation (Intel-based)*
- *Cloud-Native Platform Stack (Red Hat OpenShift)*
- *Sovereign Landing Zones*
- *AI & MLOps Enablement (Red Hat OpenShift AI)*
- *SLA-Driven Workload Placement Engine*
- *Migration & Modernization Tooling*

OPERATE for reliable support & optimization

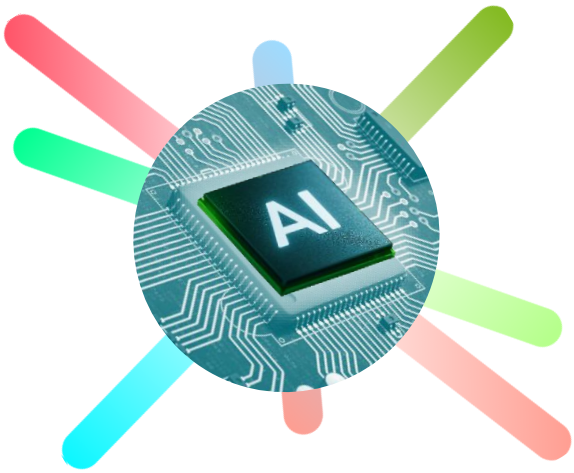
Customers benefit from **continuous operational support**, including integrated observability, security, compliance, and cost management. Policies, observability, and automation (AIOps) ensure **high performance, predictable economics, and auditability**, while enabling rapid adaptation as business or regulatory conditions evolve.



Outcomes

- *Unified Operations & Control Plane*
- *Integrated Observability & AIOps*
- *Security, Sovereignty & Compliance Operations*
- *SLA & Performance Management*
- *Cost, Energy & Carbon Optimization*

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..



Measurable Outcomes of a Sovereign AI-Ready Transformation

The combined capabilities of Intel, Red Hat, and Deloitte create one coherent transformation program, delivering measurable outcomes:

- ***Reduced risk*** through integrated design, testing, and governance to address industry- and region-specific regulations.
- ***Accelerated time to value*** via a single program approach.
- ***Consistent operational excellence*** across AI workloads and hybrid environments.
- ***Future-proof flexibility*** through open standards, modular architecture, and consumption-based infrastructure.

This customer-centric partnership enables organizations to safely navigate a sovereign transformation toward a trusted, end-to-end AI platform. Data, workloads, and AI models remain secure, optimized for performance, and compliant, helping organizations turn enterprise AI into a lasting strategic advantage.

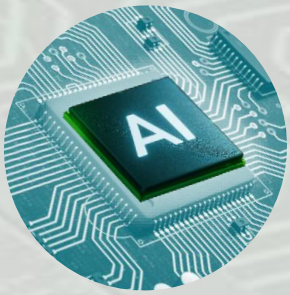
Deloitte.

Red Hat

intel.

 Copyright © 2026 Deloitte Technology & Transformation, S.L.U..

6. Final Remarks



Enterprises that build a sovereign, open, and AI-optimized private cloud foundation will gain the operational control and economic predictability required to scale AI in hybrid environments with confidence across core business processes.

Treat Sovereignty as a Design Principle

Sovereignty is no longer a compliance checkbox. It is a competitive differentiator that enables enforceable control over data, models, and workloads (across jurisdictions and providers). Leaders should assess what levels of jurisdictional, technological, and operational sovereignty are required, ensuring architectural choices preserve interoperability in hybrid environments without loss of integrity and auditability for unified governance.

Engineer for AI Economics

Most enterprise AI value comes from mapping the use cases and business goals with the best patterns. This demands performance-cost ratio optimization.

AI-optimized CPUs must be first-class citizens, with GPUs and accelerators reserved for workloads that truly require them.

SLA-driven workload placement becomes a core architectural function that requires hardware-aware orchestration.

Build a Seamless Control Plane for Hybrid AI

Fragmented stacks erode governance, obscure operational visibility, and drive escalating costs.

AI at scale demands a unified control plane that orchestrates infrastructure, platforms, and AI workloads with common policies, shared observability, and consistent automation across hybrid environments.

Hybrid agility in AI must be designed into the architecture from day one, not assembled later from disconnected clouds and tactical orchestration tools.

Adopt Consumption-Based Infrastructure

Consumption-based infrastructure allows to operate sovereign environments aligning costs directly with real workload consumption.

Combine elastic resource provisioning with real-time visibility into utilization and spend to shift from static CAPEX investments to flexible OPEX models. This approach enables private infrastructure to deliver financial discipline and operational simplicity of cloud services while maintaining full governance and jurisdictional control.



 Copyright © 2026 L

Contacts



Alfons Buxó
Partner
abuxoferrer@deloitte.es



Julio Guijarro
CTO EMEA
jguijarr@redhat.com



Piyush Khurana
GSI Partnership Manager
piyush.khurana@intel.com



Alvaro Martín
Senior Manager
amartindelvalle@deloitte.es



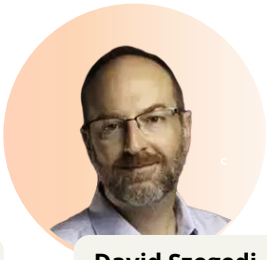
Verónica Rodríguez
ABM Marketing Manager
vrodrigu@redhat.com



Can Yurtseven
AI & Cloud Architect
can.yurtseven@intel.com



Pablo Jarabo
Senior Consultant
pjarabovaldivieso@deloitte.es



David Szegedi
Chief Architect
dszegedi@redhat.com



Paul O'Neill
Senior Director
paul.oneill@intel.com



Stefan Bergstein
Chief Architect
sbergste@redhat.com



Deloitte.

Red Hat

intel.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms or their related entities (collectively, the "Deloitte organization") is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. DTTL and each of its member firms, and their related entities, are legally separate and independent entities.

About Red Hat

Red Hat is an open hybrid cloud technology leader, delivering a consistent, comprehensive foundation for transformative IT and artificial intelligence (AI) applications in the enterprise. As a trusted adviser to the Fortune 500, Red Hat offers cloud, developer, Linux, automation, and application platform technologies, as well as award-winning services.

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

Intel, the Intel logo and other Intel marks contained throughout this document are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

© 2026. For more information, please contact Deloitte, Intel, and Red Hat