

Building Trusted AI Agents from the Silicon Root of Trust



Building Trusted AI Agents from the Silicon Root of Trust

From agent isolation to PUF-based Root of Trust across the heterogeneous AI data center

October 2026

Executive perspective

Agentic AI changes the security problem. A generative AI chatbot produces content in response to a prompt, and a person decides what to do with it. An agent acts on that output itself: it runs persistently, holds delegated credentials, browses, executes code, calls connected services and coordinates sub-agents on the user's behalf. Meta's Muse is a concrete example of this transition. Meta gives each Muse user a dedicated Linux cloud computer and deliberately separates the agent runtime from security-sensitive services, credentials and external actions [1].

This software architecture makes a larger point visible: an agent can only inherit as much trust as the infrastructure underneath it. Isolation, permission checks and credential brokers are essential, but each of those controls eventually depends on the integrity and identity of the platform that executes them.

In parallel, at the infrastructure layer, Meta and Arm are jointly developing a new CPU family for AI data centers. Meta is lead partner and co-developer of Arm AGI CPU, which is designed to operate alongside Meta's MTIA accelerators and increase performance density in AI-optimized data centers [2]. Arm explicitly positions AGI CPU for agentic-AI infrastructure [3].

As agentic AI becomes part of large-scale cloud infrastructure, the security discussion naturally extends from the agent and its secure compute environment down to the CPU and ultimately to the silicon Root of Trust. The same principle then expands across the many security-sensitive chips that make up an AI data center.

1. How agentic AI changes the trust model

Meta describes Muse as a personal agent that can work in the background, launch swarms of sub-agents, build tools and interact with connected services. Each user and Muse share a dedicated cloud VM with CPU, memory, storage and a browser. The VM is the system of record for the user's Muse environment, while limited data leaves the VM when required for inference and telemetry [1].

Meta's security design is notable because it does not assume that an intelligent agent should be fully trusted simply because it is operating for the user. Instead, Meta describes the correct mental model as two isolated security domains on one box. The Hatch daemon (Meta's core agentic harness), workspace and executable tools operate inside a systemd-nspawn runtime container. Root inside that runtime cell maps to an unprivileged host user; the environment receives its own filesystem and virtual network interface, filtered system calls and restricted Linux capabilities [1].

Security-sensitive services sit outside the runtime cell. A separate host-side Sentinel agent is the sole authority for network egress and connector actions. Code inside the runtime cell only ever sees surrogate tokens; after a request is authorized, Sentinel swaps in the real credential at the network boundary [1]. These software

boundaries are necessary but not self-supporting. Sentinel, the credential service and the runtime cell share one host and one kernel, so their separation rests on the integrity of the kernel, hypervisor and firmware beneath them.

Meta states that the launch Secure VM “does not prevent Meta from accessing data when necessary to support, secure or operate the service.” However, its planned Muse Confidential VM will encrypt the entire VM, including a person's data and conversations, using a key only the user holds [1][12]. That step moves the trust question below the software layer: it depends on keys protected in hardware and on attestation that lets a remote party verify what is running before secrets are released.

2. Trusted agents require trusted compute

When an agent can access email, files, browsers, enterprise applications and external APIs, platform identity becomes more consequential. The system needs confidence not only in what the agent intends to do, but also in what platform is executing the workload, what firmware it booted, whether critical software has been measured, and whether secrets are protected from less-trusted components.

This creates a hierarchy of assurance. At the top are user intent and agent permissions. Beneath them are VM isolation, credential handling and network policy. Beneath those controls are operating-system and firmware integrity. At the bottom sits a hardware trust anchor that must be available before higher software layers can be considered trustworthy.

Three distinct properties are at stake. **Identity** establishes which physical device is running the workload. **Integrity** establishes what it booted: secure boot refuses unsigned code, while measured boot records what actually ran so that another party can evaluate it. **Confidentiality** ensures secrets are released only to an environment in that verified state.

For an agent platform, the party that needs this evidence is usually remote: a key service deciding whether to unlock a user's confidential VM, or a fleet controller deciding whether a server may host agent workloads. Remote attestation provides it. The device signs a report of its measurements with a key bound to its hardware identity, and a verifier checks that report against expected values before releasing secrets [14]. Without a hardware-bound key, the report proves nothing, since compromised software could forge it.

The trust boundary also extends beyond one chip. Meta notes that limited data leaves the Muse VM for inference [1], so prompts and context reach accelerators, network interfaces and storage. Each device must be able to attest its own identity and firmware state, typically through protocols such as DMTF SPDm [15], before it joins the trusted path. Trust must also hold over time: firmware must not be rolled back to vulnerable versions, debug access must be locked in production, and compromised keys must be revocable. All of this depends on persistent, tamper-resistant state on the chip itself.

From Trusted AI Agent to Silicon Root of Trust

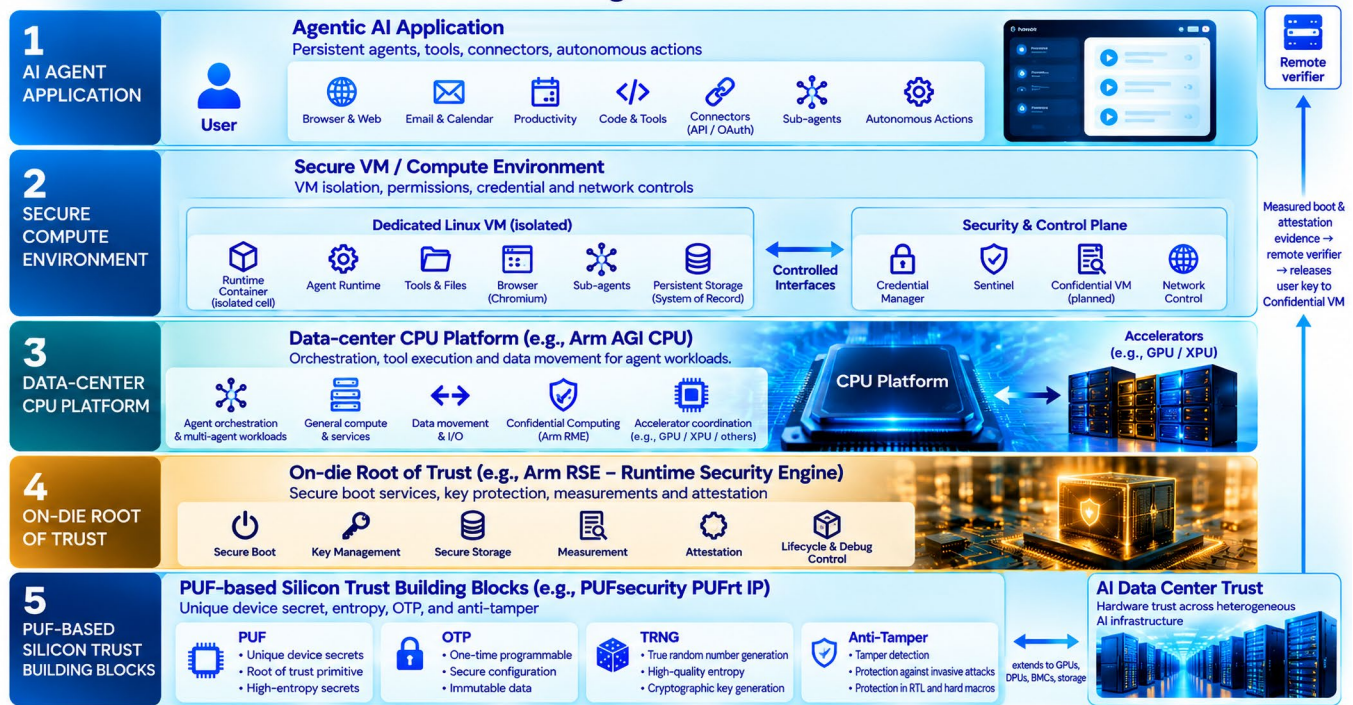


Figure 1. The silicon-to-agent trust stack: from trusted AI agent to silicon Root of Trust. Illustrative architecture.

3. Arm AGI CPU brings the CPU back to the center of AI infrastructure

AI infrastructure is often described primarily in terms of accelerators. Yet agentic workloads increase the amount of work performed outside the matrix-multiplication engine: orchestration, tool execution, data movement, networking, storage access, service coordination, security and management all remain CPU-intensive.

Meta and Arm announced in March 2026 that they will co-develop multiple generations of CPUs for AI workloads and general-purpose computing. Meta is the lead partner and co-developer for the first Arm AGI CPU. Meta states that it will deploy AGI CPU alongside its custom MTIA silicon; Meta also plans to release board and rack designs for the CPU through the Open Compute Project [2].

Arm describes AGI CPU as its first Arm-designed data-center CPU and specifically frames it around agentic AI infrastructure. Arm also claims more than twice the performance per rack versus the x86 platforms used in its comparison, emphasizing the importance of compute density at data-center scale [3].

Muse and Arm AGI CPU reflect the same shift at two layers of the stack. At the application layer, Meta is deploying persistent agents that require strong isolation; at the infrastructure layer, it is co-designing CPUs with Arm for the orchestration, data movement and system services that agentic workloads generate [1][2][3].

4. PUFsecurity and Arm: a multi-year path in hardware-rooted trust

PUFsecurity has worked with Arm on hardware-rooted trust for several years. In 2022, Arm selected PUFrt for the secure subsystem in its reference implementation of the Armv9 Confidential Compute Architecture [11].

eMemory has since joined Arm Total Design and introduced PUFrt as the hardware Root of Trust for the Runtime Security Engine (RSE) in Arm Neoverse Compute Subsystems (CSS). Arm defines RSE as a secure Root of Trust

responsible for functions including secure firmware loading, protected cryptographic keys, measurements and attestation [4].

As data-center CPUs are redesigned for agentic workloads, as Arm AGI CPU illustrates, the requirement for a hardware foundation that establishes identity, protects secrets and anchors the chain of trust carries forward.

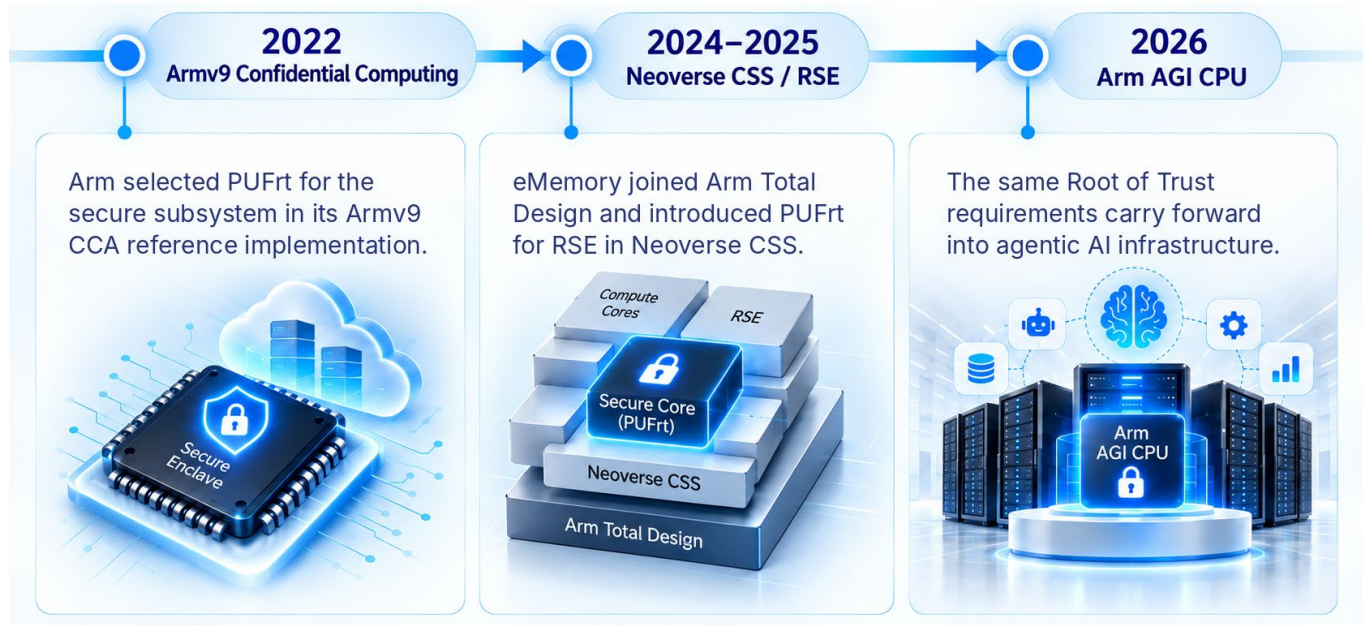


Figure 2. PUFsecurity's work with Arm on hardware-rooted trust (2022–2025), and similar Root of Trust requirements carry into the agentic AI infrastructure (2026).

5. PUFrt: a PUF-based hardware Root of Trust

PUFrt is PUFsecurity's hardware Root of Trust IP. PUFsecurity describes the product as integrating a 1024-bit physical unclonable function, a true random number generator, secure OTP and anti-tamper mechanisms. Four 256-bit PUF fingerprints can be used as unique identifiers or root-key seeds, while secure storage protects customer keys and sensitive information [5][6].

A physical unclonable function (PUF) derives a device-unique secret from random physical variation in the silicon. That secret serves as the root from which a chip's identity and key material are generated, so every key hierarchy and authentication mechanism built above it is bound to one specific piece of silicon.

In a typical PUF-anchored design, the Root of Trust reconstructs a unique device secret from the PUF at power-on rather than reading a stored key. From that secret it derives a key hierarchy, including keys that protect stored data and a device identity key. Before passing control to each firmware stage, the Root of Trust measures that stage and binds the measurement into the keys derived for it, the layered pattern standardized in the Trusted Computing Group's DICE architecture [13]. The result is an attestation identity tied both to a specific chip and to the firmware it actually booted, which a remote verifier can check before releasing secrets such as the user key of a confidential VM.

Because the root secret is regenerated from the silicon rather than programmed and stored, there is no root key at rest in non-volatile memory for an attacker to read out, and no manufacturing step in which a root key must be injected and could be exposed. Additionally, PUFrt incorporates a TRNG and secure OTP to provide functions

that include secure boot, secure debug, entropy generation, and key provisioning [5]. The architecture is modular enough to operate as a lightweight security anchor or as part of a more complete security subsystem [6].

For agentic-AI infrastructure, the significance is functional: PUFrt gives a chip a hardware-rooted identity and a protected source of secrets from which the rest of the security chain can begin.



Figure 3. NeoPUF, NeoFuse OTP, and TRNG are complementary foundations for a hardware Root of Trust.

6. Chiplets multiply the Roots of Trust

Arm AGI CPU is built from two identical CSS V3 compute chiplets joined by UCIe die-to-die links. Each chiplet contains its own System Control Subsystem, including a Runtime Security Engine (RSE) that Arm identifies as the Root of Trust, and the processor cores support the Realm Management Extension (RME) underlying Arm's Confidential Compute Architecture [16]. A single package therefore holds two Roots of Trust, and a two-socket server four.

Multiple on-package Roots of Trust raise questions a monolithic SoC never faces. Each die must hold its own identity, so the package can confirm that the die across the link is genuine. Boot and measurement must be coordinated so the package attests as one system, and secrets must remain protected as they cross the die-to-die link.

These questions become harder as chiplets become heterogeneous. Arm led the Foundational Chiplet System Architecture (FCSA), a vendor-neutral specification derived from Arm's Chiplet System Architecture and contributed to OCP, which defines how chiplets from different vendors discover each other, boot together and secure the system [17]. In a multi-vendor package, trust cannot depend on any single supplier's key provisioning.

PUF and OTP address both cases. A PUF gives every chiplet a unique device secret without a key-injection step at each supplier. eMemory's NeoFuse anti-fuse OTP, available from 0.15 μm down to 3 nm FinFET [7], and with 2nm GAA in development, holds the persistent state each die needs: lifecycle status, revocation data and anti-rollback counters.

7. The opportunity beyond the host CPU, by Caliptra

A modern AI server contains far more security-sensitive silicon than the host CPU. Accelerators, DPUs, NICs, BMCs, storage controllers, retimers and other devices may all execute firmware, store configuration, communicate across privileged fabrics or handle valuable data. Compromising one component can undermine confidence in the larger system.

OCP's platform-security model requires every device to have a Root of Trust that verifies firmware at boot, maintains authenticity through updates and supports recovery, with each device reporting its integrity to the platform through attestation [10]. Caliptra moves that Root of Trust inside the silicon: instead of a separate chip beside the processor, it is a block integrated into the CPU, GPU or DPU itself, serving in version 1.x as a Root of Trust for Measurement and adding update and recovery in 2.0 [8][9].

Each class of device uses its Root of Trust differently. In CPUs it measures and attests platform boot and identity; in accelerators it establishes authenticity and firmware integrity before the device joins a trusted compute pool; in DPUs and SmartNICs it protects the identities and keys used in privileged data paths; in BMCs it anchors the management plane; in storage it protects device identity and media encryption keys. The implementations differ, but the primitives do not. Caliptra is anchored by a unique device secret generated from on-chip entropy, kept in one-time-programmable fuses and expanded into a DICE identity chain for attestation [8].

Arm RSE and Caliptra have different architectures, but both reflect the movement of foundational trust into the SoC itself. Whatever architecture a chip adopts, open or proprietary, it needs the same primitives: a unique device root secret, a high-quality entropy source and protected non-volatile memory. PUFrt supplies all three, with secure storage built on eMemory's NeoFuse OTP [6].

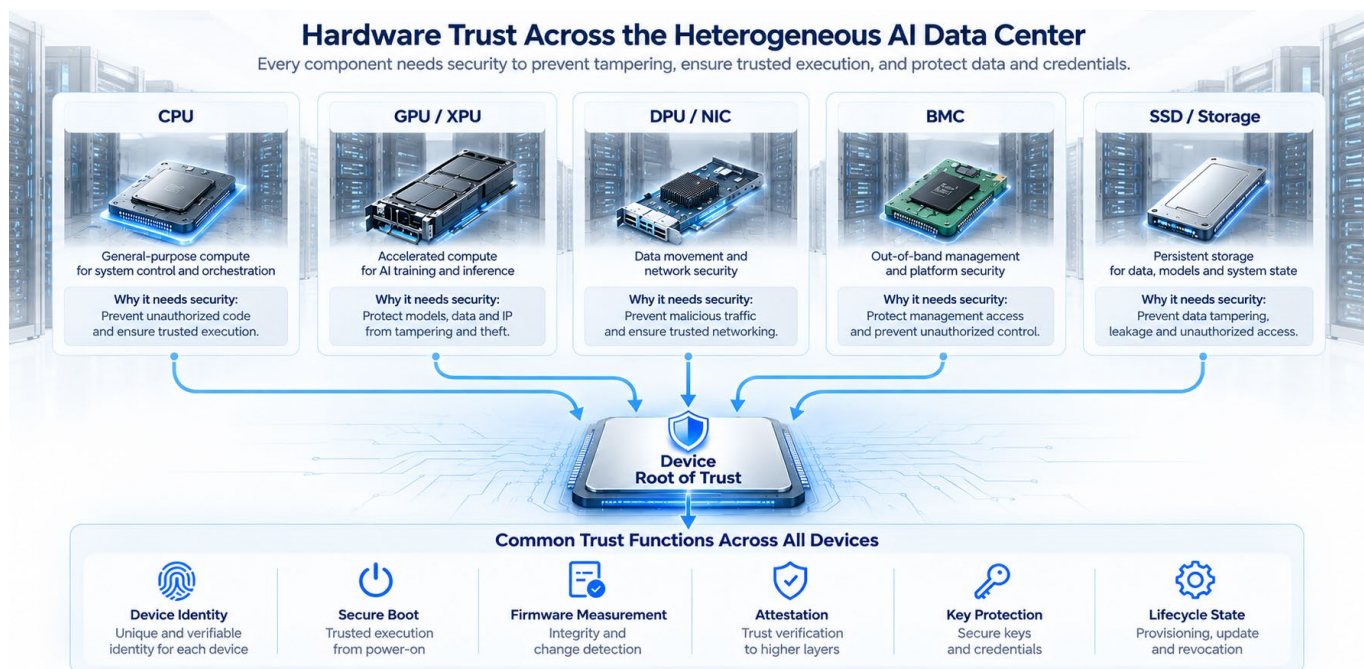


Figure 4. Every class of data-center silicon needs its own hardware Root of Trust.

8. Building trusted AI agents from secured silicon foundations

Agents like Muse change what a compromise costs: a misled agent does not just return a bad answer, it acts with the user's credentials. Meta's design contains that risk with software isolation, and Meta plans a Confidential VM encrypted with a key only the user holds [1][12]. Delivering that guarantee requires keys released only to a platform that can prove what it booted.

That proof begins in silicon, and no longer in one place. Arm AGI CPU carries a Runtime Security Engine on each of its chiplets and supports the Realm Management Extension behind Arm's confidential computing architecture [16]. Accelerators, DPUs, BMCs and storage each need their own Root of Trust under OCP's model [8][10], and FCSA extends the same requirement to multi-vendor chiplet packages [17].

Every one of those Roots of Trust rests on the same foundation: a unique device secret protected at rest, a trustworthy entropy source, and persistent state that cannot be rolled back; this is the layer PUFrt provides [5][6][7]. Trusted agents are built from the silicon upward, one Root of Trust at a time.

References

1. Meta AI Research, “How We Built Safety Into Muse,” Sept. 8, 2026. [Source](#)
2. Meta, “Meta Partners With Arm to Develop New Class of Data Center Silicon,” Mar. 24, 2026. [Source](#)
3. Arm, “Arm expands compute platform to silicon products in historic company first,” Mar. 24, 2026. [Source](#)
4. Arm Neoverse Reference Design Platform, “Runtime Security Engine.” [Source](#)
5. PUFsecurity, “PUFrt - Hardware Root of Trust.” [Source](#)
6. PUFsecurity, “PUFrt Hardware Root of Trust Datasheet.” [Source](#)
7. eMemory, “NeoFuse.” [Source](#)
8. CHIPS Alliance, “Caliptra Specification,” GitHub (chipsalliance/Caliptra, doc/Caliptra.md). [Source](#)
9. Open Compute Project, “Cloud Security: Integrating Trust into Every Chip,” Oct. 17, 2022. [Source](#)
10. Open Compute Project, “OCP Security Announces version 1.0 specs for Root of Trust,” Nov. 9, 2020. [Source](#)
11. PUFsecurity and eMemory, “PUFsecurity and eMemory Launch Next-Gen PUF-based Hardware Root of Trust IP for Future Computing,” Feb. 7, 2022. [Source](#)
12. Meta, “Introducing Muse: The World’s First Personal AI Agent Built for Everyone,” Sept. 8, 2026. [Source](#)
13. Trusted Computing Group, “DICE Attestation Architecture,” Version 1.1. [Source](#)
14. IETF, “Remote ATtestation procedureS (RATS) Architecture,” RFC 9334, Jan. 2023. [Source](#)
15. DMTF, “Security Protocols and Data Models (SPDM),” DSP0274. [Source](#)
16. S. Pradhan and D. Goel, Arm, “Arm AGI: A Disaggregated, Chiplet-Based Server SoC for Scalable Coherency and Memory Bandwidth in the Terabyte/s Era,” Hot Chips 2026, Aug. 24, 2026. [Source](#)
17. Open Compute Project, “OCP Open Chiplet Economy is Leading the Next Wave of AI: Inference,” Feb. 17, 2026. [Source](#)