



## PRODUCT BRIEF

# Arm AGI CPU

## KEY FEATURES AND BENEFITS

- **AGENTIC-TUNED PERFORMANCE**  
Up to 136 Arm Neoverse V3 cores, with class-leading 6GB/s per-core memory bandwidth at sub-100ns latency.
- **RACK-SCALE ARCHITECTURE**  
A 300W TDP enables deployment of up to 8,160 cores per standard 36kW air-cooled rack.
- **UNMATCHED COMPUTE DENSITY**  
More than 2x the performance per rack of comparable x86-based deployments.<sup>1</sup>

## INTRODUCTION

The Arm® AGI CPU introduces a new class of production-ready silicon designed to power the next generation of AI-native data centers. Built on the Arm Neoverse™ platform, the AGI CPU is purpose-built to deliver high-density, energy-efficient compute, optimized for rack-scale deployments. It is the ideal processor to handle the expanding role of CPUs in AI infrastructure: managing continuous inference workloads, coordinating compute across heterogeneous systems, orchestrating AI agents, and enabling massively parallel fan-out tasks. It is also a perfect solution for customers already running workloads on Arm in the public cloud today, providing a path to high-density Arm performance and efficiency in their own on-premise or co-located data center deployments.

<sup>1</sup> Based on estimates.

---

## HIGHLIGHTS

### **High-performance CPU**

Built on the Armv9.2 instruction set architecture (ISA), the Arm AGI CPU integrates up to 136 high-performance Arm Neoverse V3 cores, each equipped with dual 128-bit SVE2 (Scalable Vector Extension 2) units. These enable advanced AI and ML acceleration with support for bfloat16 and INT8 MMLA instructions, operating at up to 3.2GHz, all-core clock speed with boost speeds up to 3.7GHz.

### **High-memory bandwidth**

Up to 6GB/s per-core memory bandwidth to support AI workloads that require high data throughput, reducing memory bottlenecks and improving system-level performance for AI and cloud workloads.

### **Advanced I/O and accelerator connectivity**

Enables large-scale heterogeneous compute with 96 lanes of PCIe Gen6, support for CXL 3.0, and AMBA CHI Extension Link.

### **Enterprise security architecture**

Securing modern cloud and AI infrastructure with hardware-level security capabilities for multi-tenant environments:

- Runtime Security Engine (RSE)
- Pointer Authentication
- Branch Target Identification protections

## BETTER RACK-LEVEL PERFORMANCE AND EFFICIENCY FOR AI DATA CENTERS

Arm AGI CPU establishes a new silicon foundation for the next generation of intelligent infrastructure. Every aspect of the architecture, from core density to memory bandwidth and I/O connectivity, is optimized to maximize usable compute at rack scale while operating within available power envelopes. The result is a high-density, energy-efficient compute foundation optimized for AI inference, agent orchestration, and cloud-native services.

Specs	Arm AGI CPU 136C (max core count)	Arm AGI CPU 128C (TCO optimized)	Arm AGI CPU 64C (max mem/core)
SKU	SP113012	SP113012S	SP113012A
Processing Cores	136 Neoverse V3 2x 128 SVE 2MB/core L2	128 Neoverse V3 2x 128 SVE 2MB/core L2	64 Neoverse V3 2x 128 SVE 2MB/core L2
CPU Architecture	Armv9.2 bfloat16 and INT8 AI instructions	Armv9.2 bfloat16 and INT8 AI instructions	Armv9.2 bfloat16 and INT8 AI instructions
System-Level Cache	128MB	128MB	128MB
Max Frequency	3.5GHz	3.5GHz	3.7GHz
Base TDP*	300W	300W	300W
RDIMM Memory	12x DDR5 up to 8800MT/s	12x DDR5 up to 8800MT/s	12x DDR5 up to 8800MT/s
Memory Throughput/core	6GB/s per core	6.3GB/s per core	13GB/s per core
PCIe/IO	96x lanes PCIe Gen6 CXL 3.0 Type 3	96x lanes PCIe Gen6 CXL 3.0 Type 3	96x lanes PCIe Gen6 CXL 3.0 Type 3
PCIe Control Lanes	6x 1 Gen4	6x 1 Gen4	6x 1 Gen4
2-Socket Support	Yes	Yes	Yes
2 DIMMS per channel	Yes	Yes	Yes

\*Represents a preset TDP value within the configurable TDP range

---

## ACCELERATING DEPLOYMENT WITH REFERENCE PLATFORMS

Arm provides an OCP-compliant 1OU modular reference server platform. This serves as a catalyst for customers moving from evaluation to deployment. By providing a verified baseline, the reference server enables rapid validation of software stacks and infrastructure designs. Built around the AGI CPU, the design enables faster time to market for our clients.

1U2N	Arm AGI CPU 1OU Dual Node Reference Server	21" DC_MHS 1U air-cooled
2U2P	Arm AGI CPU 2U2P Reference Server	19" air-cooled