

The \$246 Billion CPU Opportunity's Winning Chip Has Not Been Designed Yet

Microsoft Azure's Fleet Data Defines the Agentic CPU. NVIDIA Vera Is Closest, and Intel, Arm, and AMD Each Have a Piece Others Lack

Analyst(s): Brendan Burke

Publication Date: October 1, 2026

Document #: AIOBB202609

What You Need to Know

- **Agents Turn the CPU From Bystander Into Bottleneck:** The market has been asking which server CPU is built for AI agents. A chatbot request makes one model call, so the CPU mostly waits on the GPU. An agent loops through reasoning, tool calls, code execution, API requests, and verification, and most of those steps run on the CPU. Futurum's February report, "[Can the CPU Market Meet Agentic AI Demand?](#)" predicted this shift. This report scores the four leading server CPU architectures of 2026, NVIDIA Vera, Intel Diamond Rapids, Arm AGI CPU, and AMD Venice, against what the largest agent operators have since measured.
- **Microsoft Put the CPU on the Agent's Critical Path:** Two Microsoft Azure studies are the first production measurements of agentic AI on real hardware. In a 24-hour fleet trace, tool execution matched or exceeded LLM inference time for more than 27% of requests. Across 13.5 million GitHub Copilot sessions, LLM and tool calls ran 1:1, and 92% of tool wall-clock time sat on the critical path.
- **The Data Yields Five CPU Parameters:** Azure's cores ran at an IPC of 1.2 to 1.6 because sandboxes sharing a core evict one another's cache lines and branch history, and context switches rose 9x, from 71 to 660 per second, as concurrency grew from 1 to 32 agents while utilization stayed low. Those two failures define what an agentic CPU must be measured on. The parameters Microsoft requests for agentic CPUs are loaded per-core latency, isolation of cache and predictor state under multiplexing, separate core classes for the control plane and the tool runners, memory capacity per core for KV cache offload, and hardware-assisted scheduling. Core count fixes none of them.
- **Vera Is Best Aligned, but Every Vendor Holds a Piece:** NVIDIA chose a big core over a fast clock, partitioned threads for SLA determinism, and one compute die with no NUMA, which answers the runner-role evidence directly and is the only frontier design validated by a third party to date. Arm chose memory latency and bandwidth

partitioning. Intel chose uniform latency and priority cores. AMD chose thread density and a portfolio. No one chose agent-level partitioning or a scheduling engine.

- **The \$246 Billion Winner Has Not Been Built Yet:** Futurum models the 2030 server CPU market at \$245.9 billion, with standalone AI CPUs at \$164.7 billion, or 67%. Catching up is a matter of choices, not process nodes. NVIDIA needs capacity and a role split inside the socket, Intel and AMD need to expose priority-core and QoS machinery per agent, Arm needs a stronger thread, and all four need the scheduling offload Microsoft asked for.

The Futurum View

Our view is that AI agents change what the CPU does and which CPU will win. In the chatbot era, the CPU was a traffic cop, feeding a GPU that generated tokens and otherwise waiting. An agent takes those tokens and does work with them. It runs generated code in a sandbox, calls APIs and databases, browses, compiles, tests, and checks its own output, then loops, and every one of those steps runs on the CPU while the GPU waits. Futurum's February 2026 report, [Can the CPU Market Meet Agentic AI Demand?](#), argued that this shift would pull CPU-to-GPU ratios back toward 1:1 and strain server CPU supply. Seven months on, the largest agent operators have measured it, and the question moves from how many CPUs agents need to which CPU an agent needs.

There is no such thing as an agentic CPU today. The vendors using the label each offer a partial answer to a problem the customers have now measured. NVIDIA says agents need the fastest single thread at scale. AMD says one profile does not fit and sells a six-SKU portfolio. Arm says memory latency per core inside 300 W. Intel promotes 1.28 GB of cache. Madhu Rangarajan opened AMD's Advancing AI conference session on agentic CPUs by asking whether any such thing as an agentic CPU exists and argued the industry must align on what matters at each pipeline step before benchmarks can exist.

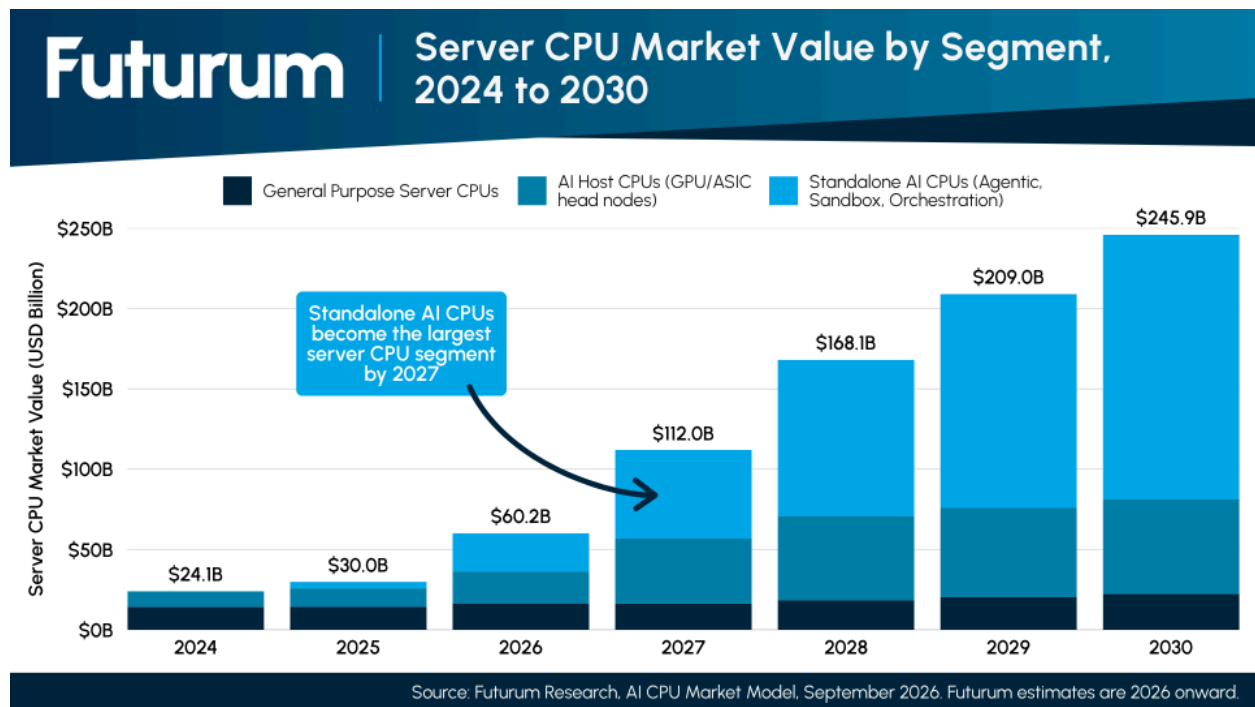
The two largest agent operators have since answered from opposite directions. Microsoft's Azure studies say the host is on the critical path, its cores stall on memory and the front end, its three software roles want three different machines, and coordination cost grows tenfold while utilization stays flat. AWS says roughly 80% of agentic compute is CPU-heavy, then built AgentCore around never paying for it with sessions that bill only for CPU that is working rather than waiting on I/O, cold memory reclaimed, and snapshot-restore that holds cold starts to two seconds at P75 regardless of image size. Microsoft wants a CPU that behaves under sandbox multiplexing. AWS wants one that contains costs while the agent thinks. Signal65, the only independent tester of agentic CPUs, set the priorities of loading every core with its own sandbox, timing the full lifecycle from container creation to teardown, and per-core task completion under full load.

Against that composite, NVIDIA Vera is best aligned because its three defining choices, a big core over a fast clock, statically partitioned threads, and one compute die with no NUMA, were

made for the failure modes Microsoft measured. It is also the most exposed, because its memory and role-split choices assume a Rubin rack around it. Arm, Intel, and AMD each hold a piece Vera lacks, yet none holds the hardware scheduling engine Microsoft prioritizes. The first true agentic CPU will be judged by how little of it that orchestration software has to work around.

Futurum's server CPU model puts standalone AI CPUs, the sockets running sandboxes, orchestration, and tools without an attached accelerator, at \$23.7 billion in 2026 and reaching \$164.7 billion by 2030, when they are 67% of a \$245.9 billion market. The architecture that wins the standalone segment wins the decade.

Figure 1: Server CPU Market Value by Segment, 2024 to 2030 (\$B)



Source: Futurum Research, AI CPU Market Model, September 2026. Futurum estimates are 2026 onward.

Microsoft Settled the Argument About What the CPU Does

Microsoft's August paper [Architectural Implications of Agentic AI Workflows](#) pairs a 24-hour Azure fleet trace with a controlled study and a server prototype called Agora. Host CPU utilization sat near an 11% median during sequential phases and spiked to nearly 100% at stage boundaries. [Agentic Coding in the Wild](#) covers 13.5 million Copilot sessions and 774.7 million tool calls. The median tool call takes 166 ms, but the mean is 16.7 seconds. Containers idle 5.8 seconds inside a turn but 243 seconds across turns, the same bursty profile AWS built AgentCore's runtime around. Two operators share a signal of the needs for agentic servers, including the CPU within them.

Six Parameters Define the Agentic CPU

Microsoft names the mechanisms required by its server CPU hardware:

- **Loaded Per-Core Latency:** Every tool call sits on the critical path, so the slowest core under full load sets the tail. This is also Signal65's metric.
- **Cache and Branch State Isolation:** The fleet's low IPC and 43% to 47% backend-stall rate come from agents evicting one another's cache lines and predictor history.
- **Role Heterogeneity:** Microsoft splits host work into three software roles. Schedulers dispatch model requests to the GPU, orchestrators route state between agents and enforce workflow dependencies, and runners execute the tools. The first two are latency-critical and steady, needing about 10 cores in total, while the runner pool sits idle between tool bursts and then peaks above 70 cores. One homogeneous core pool serves neither well.
- **Memory Capacity and Bandwidth per Core:** KV cache offload at turn boundaries turns host DRAM into an inference resource.
- **Hardware-Assisted Scheduling:** Every tool call and agent handoff is a short-lived process that the operating system must schedule, switch in, and switch out. As SWE-Agent concurrency rose from one to 32 tasks, involuntary context switches climbed from 71 to 660 per second while CPU utilization stayed low, meaning the host spent a growing share of its cycles coordinating agents rather than running them. Microsoft's authors conclude that general-purpose OS scheduling breaks down at high agent concurrency and call for a dedicated engine that holds agents and their state in hardware queues, so a handoff costs a queue operation rather than a kernel trip.

Table 1 maps each architecture's choices to those parameters and adds power, since stranded capacity sets agent economics.

Table 1: Microsoft-Derived Agentic CPU Parameters Mapped to 2026 Server CPU Architectures

Parameter	Microsoft Evidence	NVIDIA Vera	Intel Diamond Rapids	Arm AGI CPU	AMD Venice
1. Loaded per-core latency for the runner role	Tool time equals or exceeds LLM time in over 27% of fleet requests. run_build and run_command P99 latencies reach hundreds of seconds.	88 Olympus cores, 10-wide decode, 2 MB private L2. NVIDIA claims 100% loaded per-core performance vs 67% for a traditional CPU. Signal65: 1.64x per core vs EPYC 9755.	Up to 256 P-cores, Priority Core Turbo, APX with 32 GPRs. No loaded per-core data disclosed.	136 Neoverse V3 cores at 2.8 GHz typical, up to 3.7 GHz boost. Frequency tuned for balance, not peak.	Zen 6 and Zen 6c. AMD claims 1.2x per-core SPECrate 2026 vs Vera on a 96-core HF SKU.
2. Microarchitectural isolation under agent multiplexing	43% to 47% of pipeline slots lost to backend stalls as agents evict one another's cache and branch state.	Spatial multithreading statically partitions front-end and execution resources between two threads for deterministic rate-1 performance. MPAM partitions SLC capacity and memory bandwidth.	New core idle state retains L2 cache. Enhanced QoS and telemetry. No agent-tagged partitioning of caches or predictors.	MPAM memory-bandwidth limiting and QoS prioritization. 2 MB private L2 per core. No agent-tagged partitioning.	SMT with dynamic sharing, the opposite of Vera's approach. Private L2 plus up to 1 GB L3. No agent-tagged partitioning.
3. Role heterogeneity: control cores vs runner cores	Scheduler plane needs about 10.3 cores, steady. Runner pool sits idle then peaks at 70.1 cores unpartitioned. Authors call for heterogeneous or morphable cores.	Homogeneous 88 cores. Role split delegated outside the socket to BlueField-4 DPU in Vera Rubin.	Homogeneous P-cores. E-cores remain a separate product (Clearwater Forest).	Homogeneous V3 cores. Cortex-M class SCP and LCP processors handle power, not agent scheduling.	Portfolio approach: six SKU families (SP7, SP8, X, HF, Verano, 256c) across Zen 6 and Zen 6c. Not mixed on one socket.
4. Memory bandwidth and capacity per core	Deep-loop users re-prefill 1.1M tokens per miss. Agora stages agent state in host DRAM.	1.2 TB/s LPDDR5X on 1.5 TB SOCAMM (14 GB/s per core). 164 MB SLC. Claims 5x bandwidth per watt vs MRDIMM.	1.6 TB/s from 16 channels of MRDIMM at 12,800 MT/s (about 6.3 GB/s per core). Up to 1.28 GB LLC. CXL 3.0.	About 845 GB/s from 12 channels of DDR5-8800 (6 GB/s per core). Sub-100 ns latency. 6 TB capacity. 128 MB SLC. CXL 3.0.	About 1.6 TB/s from 16 channels of MRDIMM at 12,800 MT/s (about 6.3 GB/s per core). 4 TB per socket. Up to 1 GB L3. CXL 3.1.
5. Hardware-assisted scheduling and context switching	Involuntary context switches rise from 71 to 660 per second as SWE-Agent concurrency grows from 1 to 32. Authors call for a dedicated agent scheduling engine.	None disclosed. 1.8 TB/s NVLink-C2C shortens the CPU-GPU handoff itself.	None disclosed. QAT, DSA, and IAA complexes offload data movement and compression, not scheduling.	None disclosed. GIC interrupt virtualization and SMMU-700 for I/O virtualization.	None disclosed. SDXI offloads data movement and crypto.
6. Power envelope for agents per rack	Host CPU utilization ranges from 6% to 31% median with bursts to 100%. Provisioning for peak strands capacity.	250 W to 450 W configurable. Stable per-core frequency across loaded core count is the design claim.	Not disclosed at Hot Chips. Independent DVFS domains for core, uncore, memory, and I/O.	300 W TDP for 136 cores, TSMC N3P. Rack density under a fixed power envelope is the headline.	600 W default CPU power on 9996. UBPS, UPP, and FAST scale power with utilization.

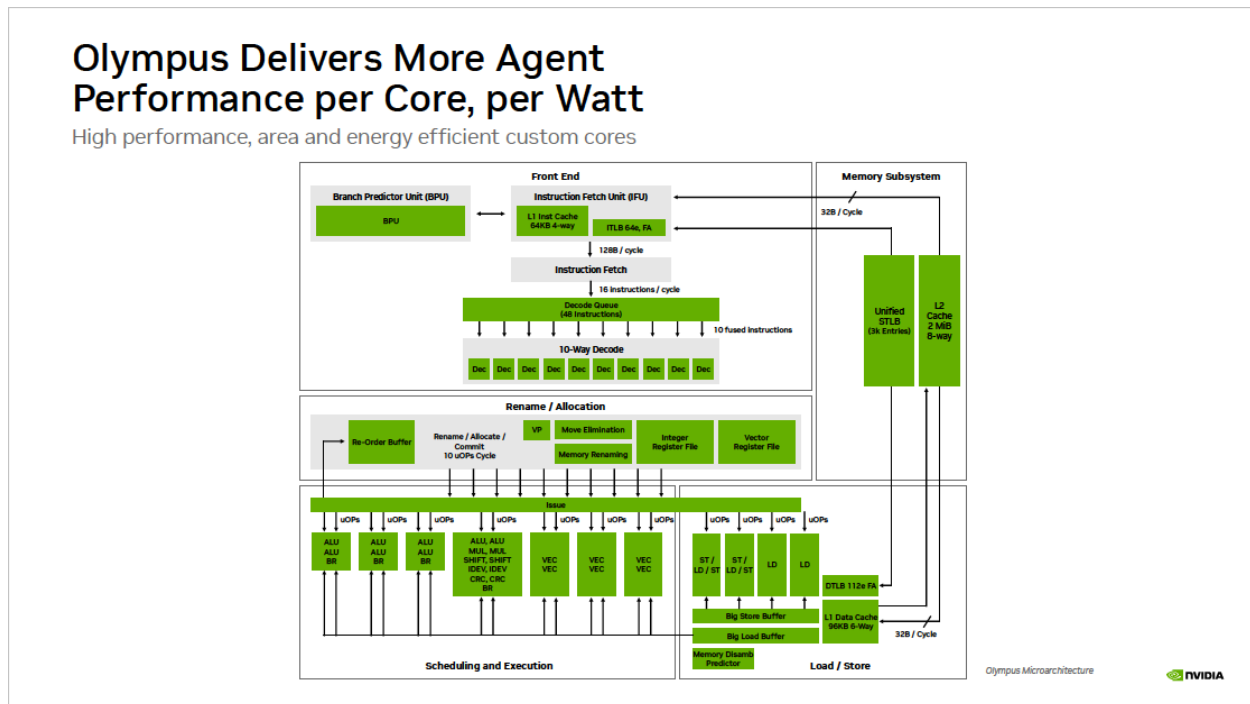
Source: UT Austin and Microsoft Azure (arXiv 2608.04458), Microsoft Azure Research (arXiv 2608.00101), NVIDIA, Intel, and Arm Hot Chips 2026 presentations, AMD EPYC 9006 whitepaper (September 2026), Signal65 (September 15, 2026), Futurum analysis. Vendor performance figures are vendor claims unless attributed to Signal65.

NVIDIA Chose a Big Core, Partitioned Threads, and No NUMA

Every Vera choice traces to one premise: the agent loop stalls on the slowest thread, so the CPU must hold single-thread performance while every core is busy. At Hot Chips, Olympus architect Polychronis Xekalakis said the team went for IPC over frequency, spent area on a large branch predictor for big footprint code and a graph prefetcher for pointer chasing, and then rejected conventional SMT for spatial multithreading, giving each thread its own front-end and scheduling resources so a noisy neighbor cannot break an SLA. That answers Microsoft's isolation parameter at the core and the loaded per-core parameter in the only third-party test so far.

NVIDIA put all 88 cores on one compute die with no NUMA to avoid the interposer tax on core-to-L3 traffic, and split memory onto four LPDDR dies whose penalty is paid only on an L2 miss. Asked what bottlenecks CPU-GPU agent traffic, chief CPU Architecture lead Jonathon Evans said mostly bandwidth, which is why the design spends on NVLink-C2C rather than DRAM capacity. As a result, LPDDR trades capacity, and in AMD's telling, RDIMM-class error correction, for bandwidth per watt, which limits KV cache offload to host memory. The homogeneous core pool leaves Microsoft's role split to a Bluefield DPU. Nothing offloads scheduling. To align with customer demand, NVIDIA needs a capacity tier behind SOCAMM, a control-plane core class inside the socket, and a hardware queue for agent context.

Figure 2: Olympus Delivers More Agent Performance per Core, per Watt (NVIDIA)



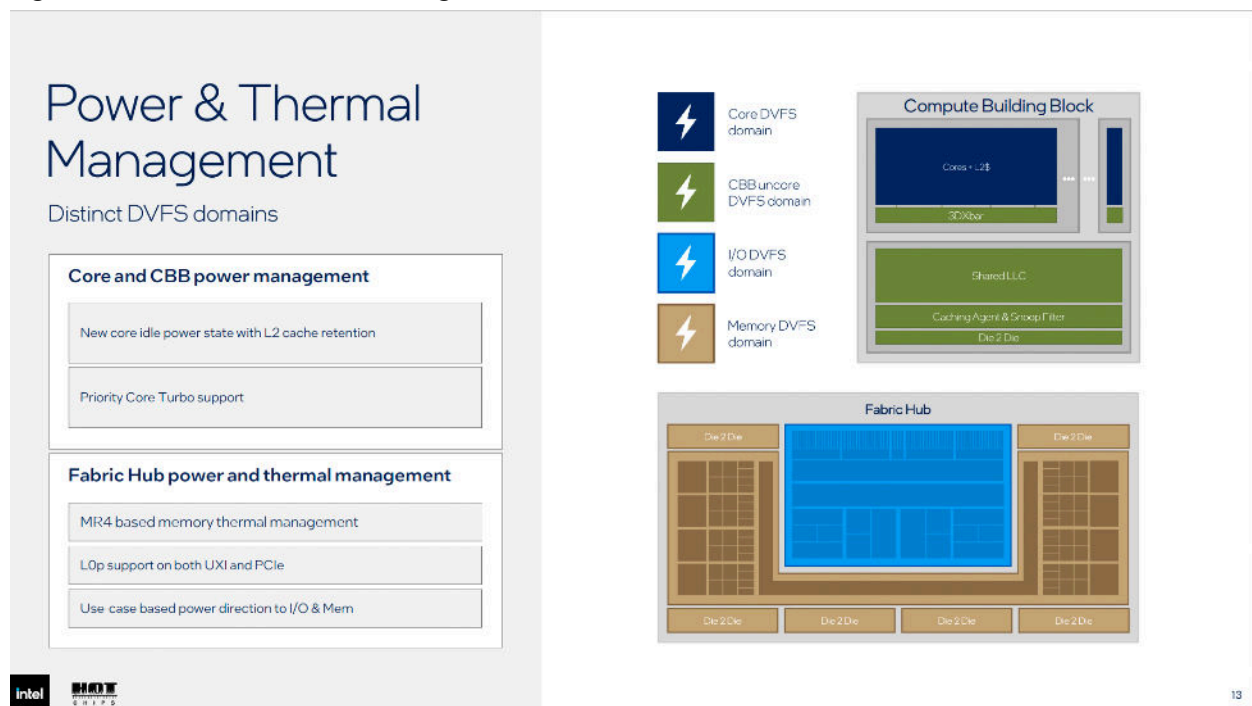
Source: NVIDIA

Intel Chose Uniform Latency and Priority Cores

At Hot Chips, Krishnakanth Sistla organized Diamond Rapids around uniform low-variability latency to memory from any core. Every compute building block connects directly to every fabric hub and the shared LLC sits on the base tile so 1.28 GB is reachable socket-wide. For agents that migrate across cores, uniformity is a virtue, and the L2-retaining idle state protects a stalled agent's working set.

Two choices matter more than the cache. Priority Core Turbo lets software name the cores that feed a GPU and hold them at top frequency, the closest any 2026 socket comes to Microsoft's role-matched cores, and a hardware power manager senses how latency-sensitive the cores are and retunes fabric and memory clocks. Missing is the runner half. Intel showed no loaded per-core data, which could be a concern as its accelerator complexes move data rather than schedule agents. To align, Intel should extend Priority Core Turbo and its QoS lineage into agent-tagged LLC and predictor partitioning, and publish per-core results under full sandbox load.

Figure 3: Power & Thermal Management (Intel)



Source: Intel

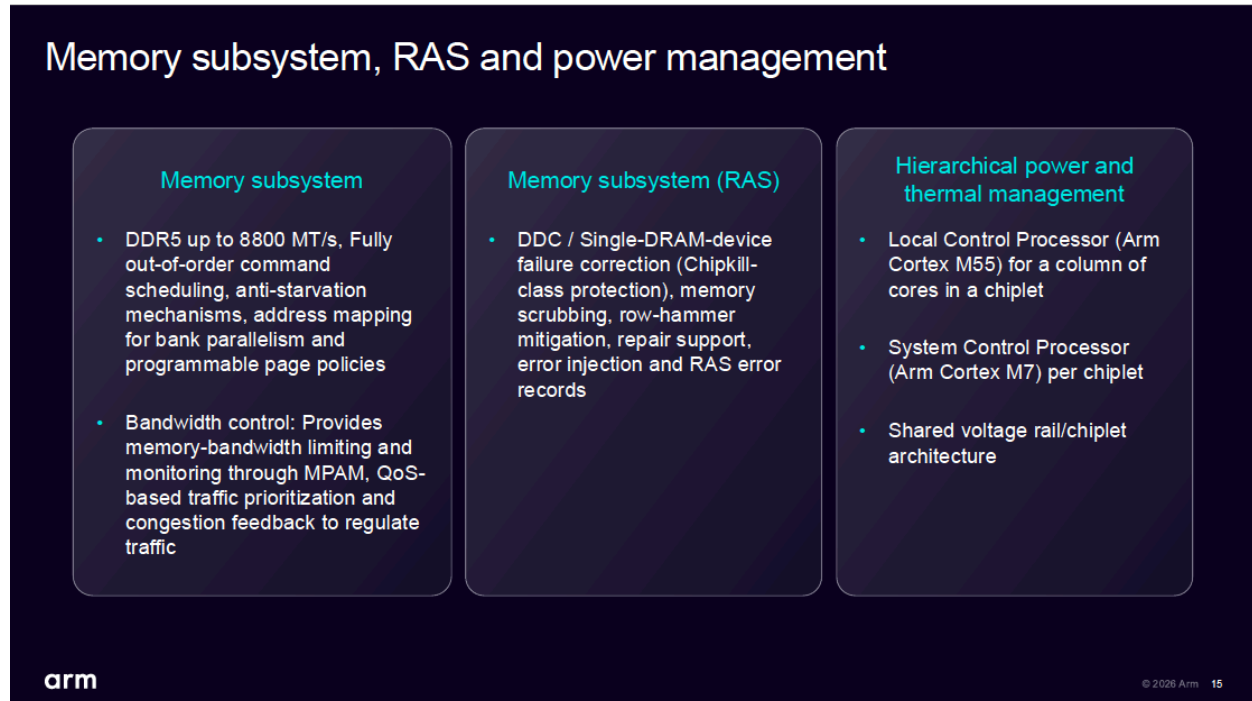
Arm Chose Latency, Partitioning, and Balance Over Peak

At Hot Chips, Deepak Goel described the AGI CPU as a set of deliberate refusals. Arm chose V3 over N3 cores to guarantee a minimum single-thread target, then declined to spend gates on frequency, setting the sweet spot at 2.8 to 3.2 GHz. It made the die-to-die link faster than memory so an interleaved NUMA1 mode works and added NUMA2 and NUMA4 modes that split the socket into two or four memory domains. Arm made fabric congestion feedback

sub-100 ns load-to-use, matched I/O bandwidth to memory bandwidth as a design rule, and exposed MPAM bandwidth partitioning and fabric QoS.

Asked what was done for agentic AI, Goel named memory latency, instruction-level parallelism, and I/O bandwidth. That answers Microsoft's memory parameter squarely, and MPAM, which NVIDIA also ships on Vera, plus Arm's NUMA modes are the only levers cloud customers have to limit an agent's contention domain.

Figure 4: Memory Subsystem, RAS, and Power Management (Arm)



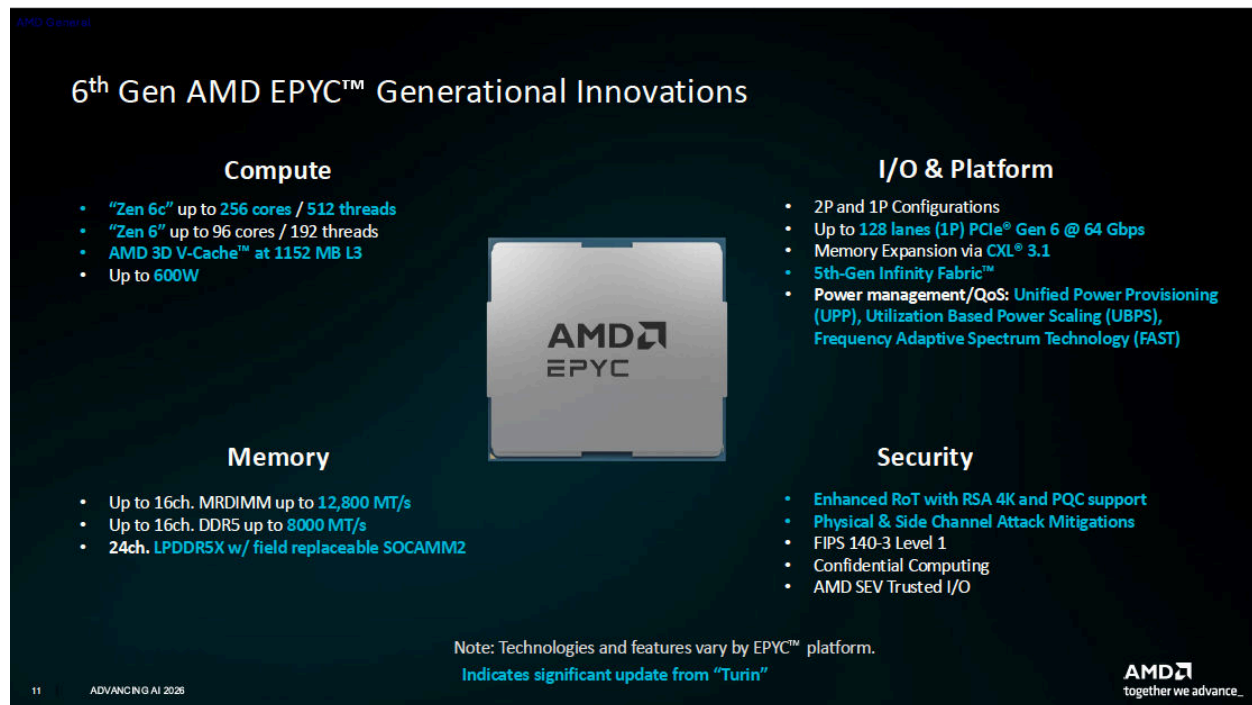
Source: Arm

Missing is the thread itself. A frequency ceiling chosen for perf per watt caps loaded per-core latency, and homogeneous chiplets offer no control-plane core class. To align, Arm needs a higher-performance core option in the same CSS and should extend MPAM from bandwidth into cache and predictor state.

AMD Chose a Portfolio and Thread Density

AMD's answer to heterogeneity is to sell it as SKUs: Venice HF on the GPU host, a 256-core Zen 6c part in the sandbox tier, SP8 in the enterprise. We believe the sandbox tier will need a faster cadence than AMD's two-year rhythm to keep pace with market needs and now there are more SKUs to update. Within a socket, AMD chose SMT because sandboxed tools wait on memory, storage, and network, the opposite bet from Vera's static partitioning, and added FAST to give critical-path cores priority frequency, unified power provisioning across CPU and DIMMs, and direct cache injection from the network. Those answer Microsoft's role and power parameters, and FAST resembles Intel's Priority Core Turbo.

Figure 5: 6th Gen AMD EPYC Generational Innovations



Source: AMD

Missing is evidence and isolation. Dynamic SMT is the mechanism Microsoft measured polluting caches and predictors. A portfolio also cannot follow a role mix that shifts inside one workflow. To align, AMD should publish Signal65-style loaded per-core results, expose FAST and QoS per agent rather than per VM, and decide whether the sandbox part needs a partitioned thread mode.

Who Is Best Aligned and How the Others Catch Up

Against Microsoft's five parameters and Signal65's test, Vera is best aligned today. It was designed for loaded per-core latency and isolation and it is the only design with independent data. Its weakness is that it was designed as half of a rack. Intel is best aligned on memory uniformity and role priority, Arm on memory latency and partitioning, and AMD on density and power management.

Catching up is a matter of choices already within reach. Intel and AMD own the priority core and QoS machinery that can become role-matched cores and agent-tagged partitioning once it is exposed per agent. Arm owns memory latency inside 300 W and needs the low latency and control plane threads to keep pace. NVIDIA owns the thread and needs capacity and a role split that does not depend on a DPU. None owns the scheduling engine Microsoft's authors rank first. Microsoft's software achievement of 95% harvested throughput with under 3% agent slowdown shows software will keep taking that ground until silicon does.

What to Watch

- **Independent Agentic Benchmarks Arrive in 1H 2027:** Signal65 plans to run its Terminal-Bench 2 replay on AMD and Intel parts. AI Infra Summit's agentic CPU panel agreed on the KPIs: time-to-interactive, tool execution latency, and sandbox density. The first vendor to publish loaded per-core results against them sets the sandbox tier's reference price.
- **Hyperscaler Runtimes Decide What the CPU Is Paid For:** AWS AgentCore bills only for CPU that is working and is adding x86 microVMs and suspend-resume with memory snapshots. Merchant CPUs can win an argument yet may lose the attach unless they land agent-branded instances, as AMD did with Azure HDv2.
- **Memory Technology Becomes the Dividing Line:** Microsoft's turn-boundary offload and AWS's snapshot-restore both make host DRAM an inference resource, favoring Venice's 4 TB and Arm AGI's 6 TB sockets over Vera's 1.5 TB of LPDDR5X. The winner is the architecture with the lowest memory bill per agent session.
- **The CPU-to-Accelerator Ratio Keeps Rising:** Futurum's model has AI host CPUs per accelerator rising from 0.25 in 2021 to 0.36 in 2026 and standalone AI CPUs per accelerator reaching 1.0 by 2029. AMD sees 1:1 moving past 2:1. If Microsoft's agent data generalizes beyond coding agents, the CPU line in AI capex stops being a rounding error.

Other Insights from Futurum

[Can AMD EPYC Extend Its Lead Over Vera and Xeon in the Agentic Data Center?](#)

[NVIDIA Engineers the Agentic Data Center with DSX Software Control and Vera CPU Acceleration](#)

[Arm's \\$15 Billion CPU Opportunity Hinges on Agentic Data Center Design](#)

[COMPUTEX 2026: Are Agentic CPUs Rivals or Complements?](#)

About Us

About the Authors

Brendan Burke is Research Director for Semiconductors, Supply Chain, and Emerging Tech at The Futurum Group. He covers AI accelerators, advanced packaging, memory markets, CPU architecture, foundry dynamics, EDA, and AI infrastructure economics, and he produces analyst reports, Market Coverage News, and research for Futurum clients. He appears as a commentator on Six Five Media and represents Futurum in media, including CNBC Asia and Bloomberg. He can be reached at bburke@futurumgroup.com.

About The Futurum Group

Every day, The Futurum Group's analysts, researchers, and advisors help business leaders worldwide anticipate tectonic shifts in their industries and leverage disruptive innovation. Unlike traditional analysts, The Futurum Group works not only in analysis and research but also takes that insight and knowledge even further, engaging all the way through the go-to-market process.

Futurum Research provides in-depth research and insights on global technology markets using advisory services, custom research reports, strategic consulting engagements, digital events, go-to-market planning, and message testing. It also creates, distributes, and amplifies rich media content that all stakeholders read, watch, and listen to.

See more details on The Futurum Group at futurumgroup.com.

Copyright & Use License

Copyright Notice

Copyright ©2026 by The Futurum Group, LLC. All rights, including that of translation into other languages, are specifically reserved. No part of this publication may be reproduced in any form, stored in a retrieval system, or transmitted by any method or means, electrical, mechanical, photographic, or otherwise, without the express written permission of The Futurum Group futurumgroup.com. United States copyright laws and international treaties protect this publication. Unauthorized distribution or reproduction of this publication, or any portion of it, may result in severe civil and criminal penalties and will be prosecuted to the maximum extent necessary to protect the publisher's rights.

License Notice

This document may be distributed within the licensed organization only.

The following acts are prohibited:

Transmittal to others outside your immediate organization, including partners, resellers, external consultants, etc., in any media format

Posting on a website that is accessible to others outside your immediate organization

The possession or use within an unlicensed organization

Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. The Futurum Group shall have no liability for errors, omissions, or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve their intended results. The opinions expressed herein are subject to change without notice.