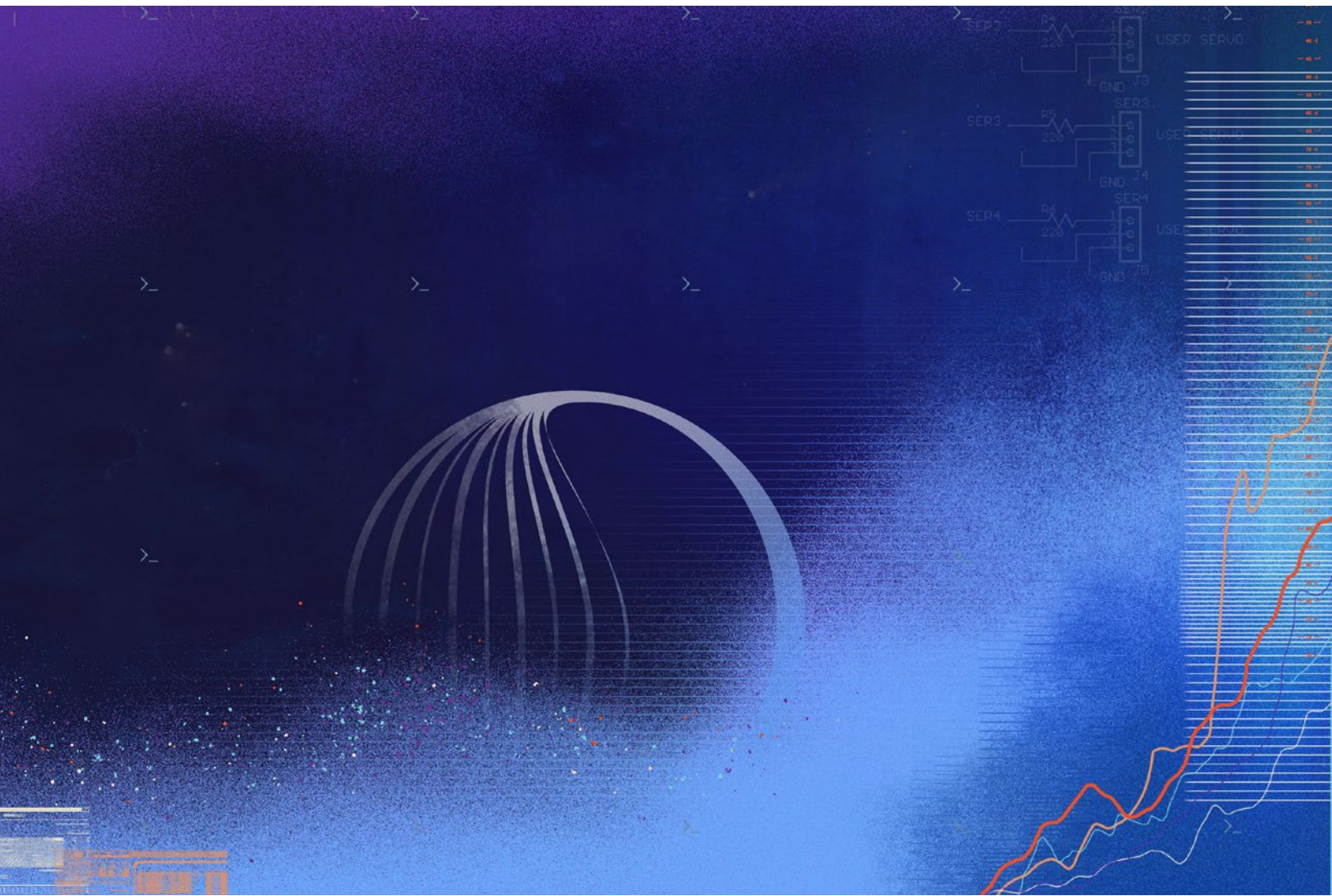




Scaling Agentic AI with Arm AGI CPU: 2x More Work per Rack Over Leading x86 Processor

An independent analysis of the Arm
AGI CPU performance running modern
enterprise and agentic workloads

AUTHOR

Russ Fellows

VP, Labs | Signal65

OCTOBER 2026

IN PARTNERSHIP WITH

arm

Executive Summary

Agentic AI is fundamentally changing the role of the CPU in modern data centers. Beyond model inference, AI agents continuously execute tools, access data, invoke services, and operate within isolated environments. As these workloads scale, CPU performance is increasingly measured not only by how many agents can run simultaneously, but also by how much useful work those agents can complete. With power and space constraining AI infrastructure growth, the rack is clearly a critical unit of performance, where the ability to complete more agentic tasks within a fixed power envelope becomes a key measure of infrastructure efficiency.

Arm is one of the key companies at the forefront of this transition. Leading cloud providers are increasingly building AI infrastructure around Arm-based CPUs, from AWS Graviton, Google Axion, and Microsoft Azure Cobalt, to NVIDIA Vera. Building on this foundation, the AGI CPU is Arm's first production silicon, purpose-built for agentic AI, combining high-performance Neoverse V3 cores, substantial memory bandwidth per core and high-speed I/O with rack-level density.

To evaluate these capabilities, we worked with Arm to assess AGI CPU performance against AMD's 5th Gen EPYC 9755 across seven CPU workloads spanning agent execution, tooling, retrieval and model-serving support. The study combines widely used infrastructure components with emerging agentic benchmarks, including Terminal-Bench 2 and SWE-bench, to examine both individual processor performance and the benefits of scaling within a fixed rack power envelope. This analysis provides enterprise technology leaders, architects, and IT decision-makers with evidence to consider when selecting server architectures. The main findings are:

- **The Arm AGI CPU delivers 2x the work per rack.** Across the seven workloads, the AGI CPU delivers an average of 1.99x the modeled performance per rack of the EPYC 9755 with individual workloads ranging from 1.43x to 3.19x.
- **Rack-level density is the key driver.** Arm AGI's focus on efficiency and scalability enables compute density at the rack level. Arm's rated-power model places 21 to 23 AGI CPU servers in a 24 kW rack against 13 EPYC 9755 servers: 5,712 to 5,888 physical cores against 3,328. The density factor is a ratio of server power, so it holds at any rack power level.
- **Higher rack-level performance across agentic workloads.** Arm AGI CPU leads across all three agentic performance categories in modeled rack-level performance, averaging 1.7x for tooling, 2.0x for end-to-end agentic workflows, and 2.2x for reasoning/planning, while operating at a lower rated server power than AMD EPYC 9755 (950 to 1,050 W vs. 1,700 W per two-socket server).
- **Lower sandbox creation latency.** Against a two-socket AMD EPYC 9965, the AGI CPU created sandboxes faster on all three sandbox runtimes reported in Sandbench (runc, Firecracker and gVisor), taking 45% less time for cold starts, 27% less for warm starts and 41% less in the worst case under 4x oversubscription, reducing the overhead of agent environments.



For organizations, that means about half the racks for the same CPU work on web serving, agent replay and model-serving overhead (0.50 to 0.53 racks per EPYC 9755 rack), and 30% to 35% fewer racks for database and vector search. Equivalently, it means about twice the agentic CPU work completed per rack in the same power footprint.

Rack results are modeled from Arm's rated server power and the measured per-chip results; Appendix A details the model. These results reflect early AGI CPU systems; Signal65 plans follow-on testing on final production hardware across a broader range of workloads, including Signal65 PINNACLE CPU-centric benchmarks, and additional competitive comparisons.

Measuring What Matters: Agentic Work per Rack

Agentic workloads introduce a distinct performance challenge. An individual agent may execute a sequence of dependent operations, from initializing a sandbox and running tools to retrieving data, executing code and verifying results. The latency of each step affects how quickly the agent progresses. At scale, thousands of these operations compete for shared compute, memory and I/O resources, making sustained throughput just as important as individual task execution speed.

More cores or hardware threads do not automatically translate into more completed work. Performance depends on how efficiently processors execute each task, how workloads scale under contention, and whether memory bandwidth, I/O capacity or runtime overhead becomes a bottleneck. These factors can have a significant impact on the throughput delivered by a fully populated server or rack.

Rack-level performance introduces another consideration: the power consumed by the complete system, including processors, memory, networking, storage and platform management. The most effective infrastructure is not necessarily the one with the highest-performing individual processor, but the one that delivers the greatest sustained workload throughput within the available power budget.

What We Tested

Agents combine model inference with tool execution, data retrieval, orchestration and application services, each placing different demands on the CPU. As outlined in Arm's companion whitepaper, [How to Define CPU and Software Stack Performance for Agentic AI](#), evaluating these systems requires two complementary approaches: component-level benchmarks that isolate critical CPU functions and agentic workflow benchmarks that capture how these functions contribute to task execution. The whitepaper further explores how these workloads translate into real-world agentic AI use cases, providing a framework for understanding and evaluating their performance requirements.

Signal65 evaluated seven workloads spanning tooling, agentic workflows, and reasoning/planning — plus Sandbench, which measures sandbox-creation overhead for agent execution:

Table 1. Workloads tested.

Category	Workload	What it represents	Metric
Tooling	MySQL	Transactional database behind agent tools	Transactions/s
Tooling	NGINX	Web and API front end	Requests/s
Orchestration	Sandbench	The overhead of creating isolated sandbox environments for agent tasks, including startup under concurrent demand.	Latency and creation time
Reasoning / Planning	Milvus	Vector search for retrieval	Queries/s
Reasoning / Planning	vLLM CPU overhead	CPU serving overhead with model compute bypassed	Requests/s
Reasoning / Planning	fastText	CPU word-representation training	Words/s
Agentic Workflows	Terminal-Bench 2	72 agent task types in sandboxed containers, 128 copies at once	Task completion time
Agentic Workflows	SWE-bench (replay)	Software-engineering agent workflow, 259 tasks	Tasks/min

Together, these workloads provide a broad assessment of the CPU work that underpins modern agentic AI, from individual services to the execution of multi-step tasks. This combination helps establish not only how the processors perform on specific components, but how those capabilities can translate into more agentic work completed at scale.

The seven agentic component and workflow benchmarks form the basis of the per-chip and modeled per-rack comparisons.

The Arm AGI CPU was evaluated across two SKUs due to system availability during testing, with results compared against AMD EPYC 9755 using workload-specific benchmark settings.¹ Sandbench was evaluated separately against AMD EPYC 9965.²

¹ Arm AGI CPU SKU 1 (128 cores, 2.8 GHz, DDR5-7200) was used for Terminal-Bench and SWE-bench. Arm AGI CPU SKU 2 (136 cores, 3.2 GHz, DDR5-8000) was used for the remaining workloads. Two SKUs were used because of system availability during testing; each workload's rack result uses the rack factor for the SKU it ran on. The AMD EPYC 9755 is a 128-core Turin-family processor. AMD EPYC 9755 results use SMT on, as AMD positions the processor, except vLLM, where AMD publishes results with SMT off.

² Sandbench was evaluated separately against a two-socket AMD EPYC 9965 system.

Results

Rack Density: Accounting for the Complete System

Rack-level performance depends on more than processor power. We need to account for memory, networking interfaces, storage, platform management, power delivery and system cooling. We need to also reserve rack power for supporting infrastructure, including network switches and power distribution equipment.

Because each AGI CPU server draws less rated power, more of them fit in a fixed rack budget. At 24 kW, Arm's model places 23 servers at SKU 1 and 21 at SKU 2, against 13 EPYC 9755 servers.

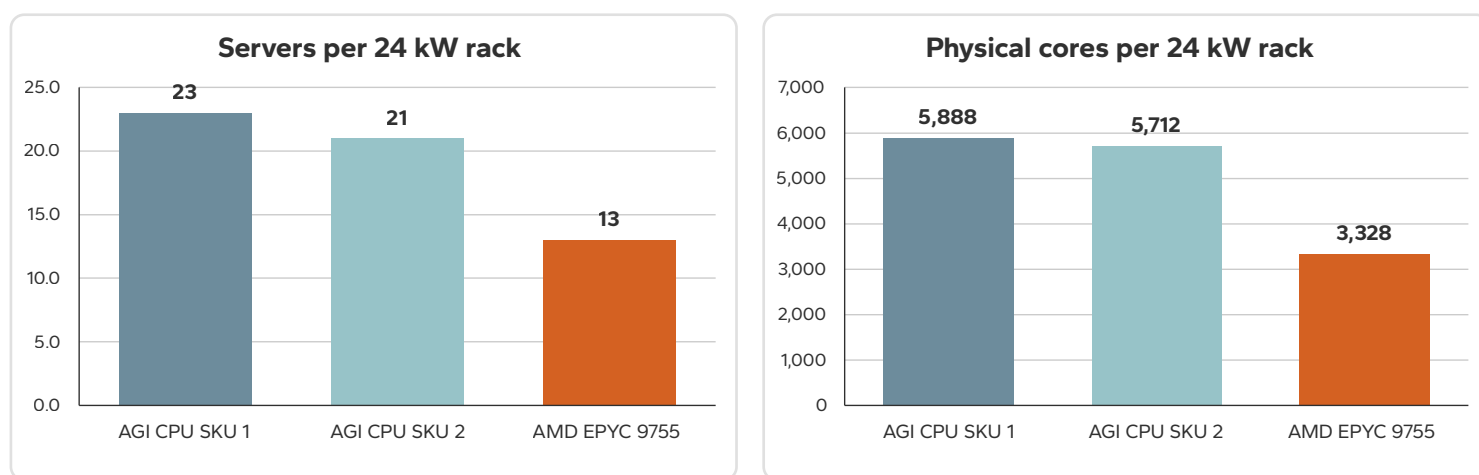


Figure 1: Servers and physical cores per 24 kW rack, from Arm's rated server power model.
The EPYC 9755 rack provides 6,656 hardware threads with SMT enabled.

Performance per Rack

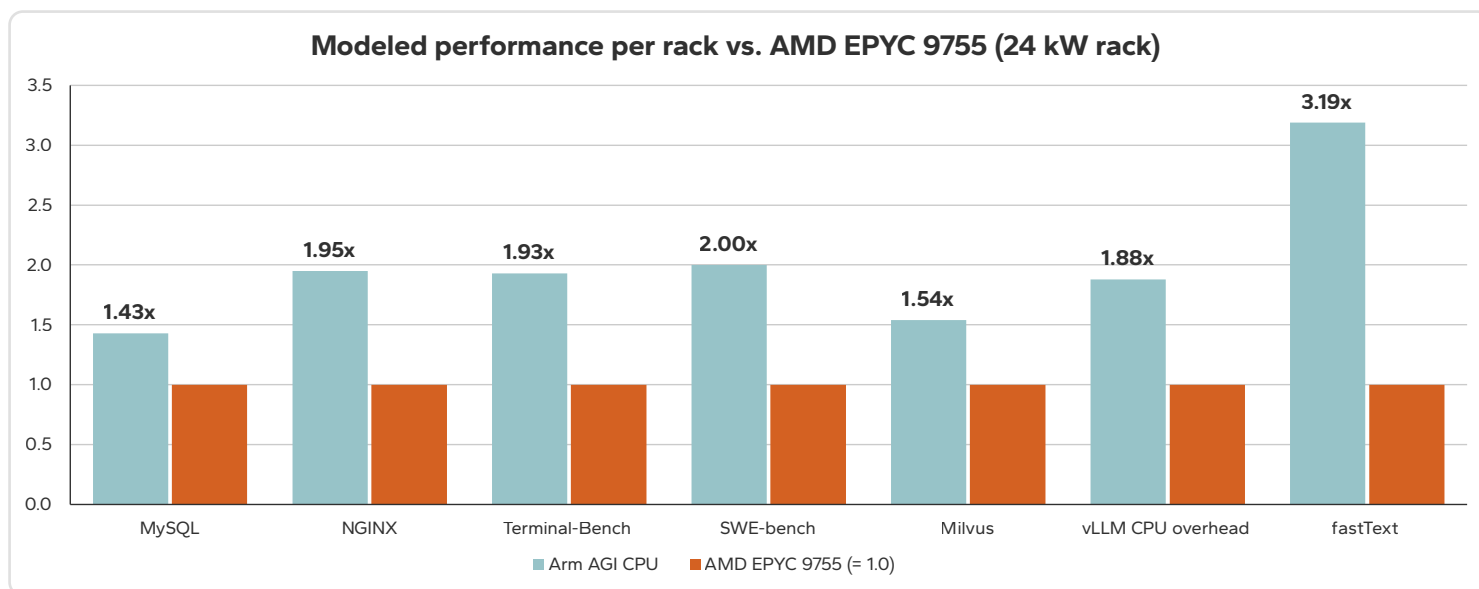


Figure 2: Modeled performance per rack: per-chip result multiplied by the rack density factor. AMD EPYC 9755 = 1.0.

Across the seven workloads, modeled performance per rack averages just about 2x. What is notable is the uplift on the emerging agentic benchmarks: on SWE-bench, the AGI CPU completes 2.00x the tasks per rack of the AMD EPYC 9755; on Terminal-Bench 2, which measures how quickly batches of agent tasks complete, it finishes 1.09x faster per chip, or 1.93x modeled per rack.

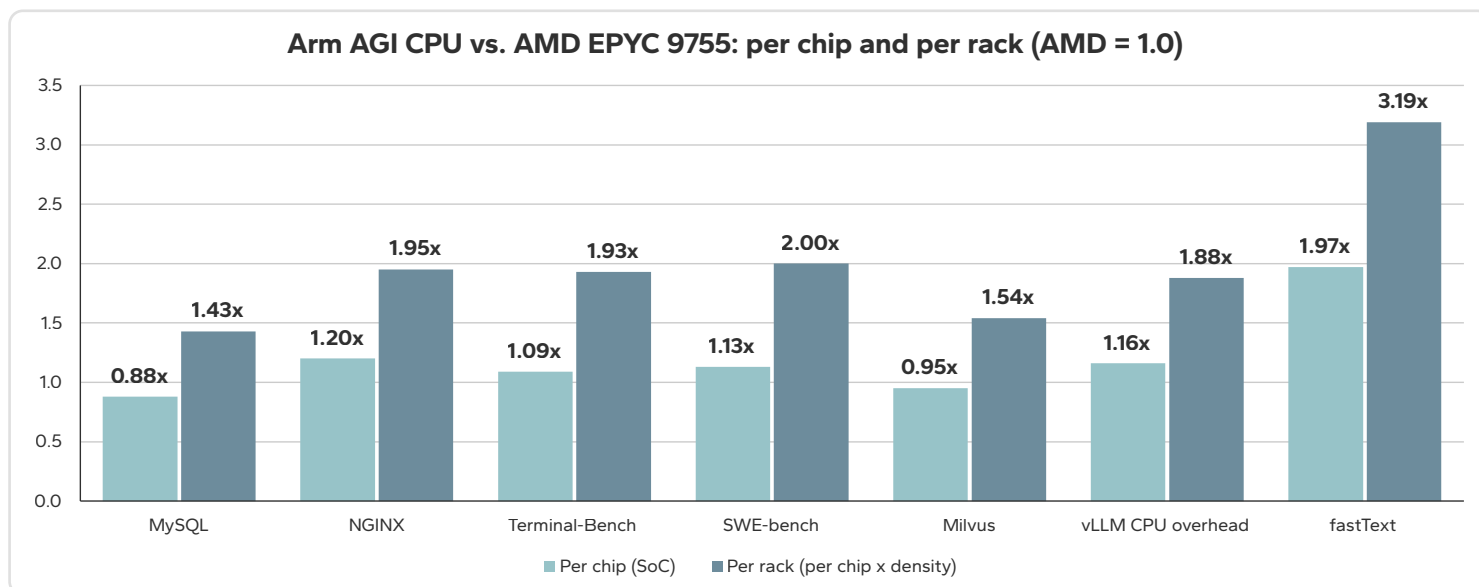


Figure 3: Per-chip results compared with per-rack results. Most of the per-rack advantage comes from density.

Arm AGI CPU delivers higher per-chip performance on five of seven workloads while operating at lower rated server power than AMD EPYC 9755 (950 to 1,050 W vs. 1,700 W per two-socket server) — a combination that compounds at rack scale. AGI CPU provides on average a 20% advantage over AMD EPYC 9755. However, taking the efficiency and scalability at the rack-level into account (details in Appendix), AGI CPU delivers about a 2x modeled performance advantage per rack.

Sandbench: Sandbox Creation Latency

Agents that run code need isolated sandboxes, created and destroyed constantly. Running in a sandbox incurs overhead in terms of time it takes to start a sandbox (cold, warm, under load/ over subscribed). Lowering this 'tax' leads to faster agent execution.³ This comparison uses a single-socket 128C AGI CPU SKU 1 system and a two-socket AMD EPYC 9965 with SMT enabled, which differ from the systems used for the rack results. Results cover runc, Firecracker and gVisor.

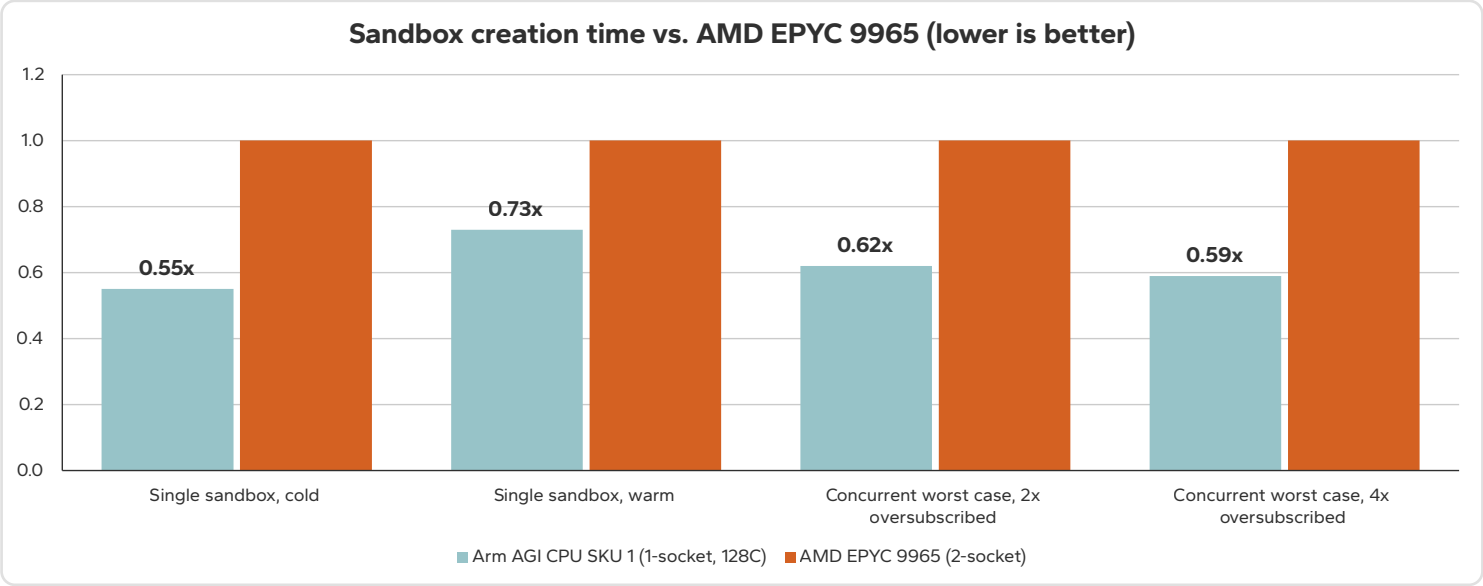


Figure 4: Sandbox creation time, Arm AGI CPU SKU 1 (1-socket, 128C) vs. AMD EPYC 9965 (2-socket) = 1.0; lower is better. Geometric mean of runc, Firecracker and gVisor; concurrent worst case is the slowest successful start of 32 (2x) or 64 (4x) starts.

Lower Latency: the AGI CPU was faster on all three sandbox runtimes reported (runc, Firecracker and gVisor), for cold and warm starts and under concurrency.

³ This comparison uses a two-socket AMD EPYC 9965 and a single-socket 128C AGI CPU SKU 1 system, so it is reported separately from the rack results.

Key Takeaways: Scaling Agentic AI with Arm AGI CPU

Rack Capacity: Fewer Racks for the Same Work

Across the tested workloads, the modeled results translate to the same work from fewer racks — about half as many for four of the seven workloads, 0.65 and 0.70 of a rack for Milvus and MySQL, and under a third for fastText. For power- or space-constrained deployments, that means more work within the same footprint or lower infrastructure requirements for the same workload.

Performance Throughput: More Agentic Work Completed

Agentic infrastructure performance is ultimately about work completed, not simply the number of agents or hardware threads running. On SWE-bench replay and Terminal-Bench 2, Arm AGI CPU delivers roughly 2x modeled rack-level performance versus AMD EPYC 9755. That translates to more CPU-side agent work completed per rack for coding agents and other tool-intensive workloads at scale.

Software Readiness: Easy to Verify

Every workload in this study ran natively on the Arm AGI CPU, including 72 agent task types in native Arm64 containers. The web servers, databases, vector search, model serving and sandbox runtimes used here all run on Arm today. Organizations can confirm the fit for their own services with a focused pilot before committing.

On Premises or Cloud

The rack-level advantage matters across deployment models. In on-premises and colocation environments, higher density means more agentic work within a fixed power and space envelope. In the cloud, the same efficiency can translate into better infrastructure economics and price-performance, which providers and customers can evaluate against their own workloads.



Conclusion

As agentic AI moves from experimentation to production-scale deployment, infrastructure performance will increasingly be defined by how much useful work agents complete, rather than how many agents or hardware threads a system can support. This shift places greater importance on the CPU's ability to execute tools, manage isolated environments and support the services that turn AI reasoning into action, all within increasingly constrained data center power budgets.

Signal65's analysis demonstrates that Arm AGI CPU combines strong processor-level performance with a significant rack-scale efficiency advantage, delivering nearly 2x modeled CPU workload performance per rack compared with AMD EPYC 9755. The results across agentic execution benchmarks and supporting infrastructure workloads reinforce the importance of evaluating CPU platforms at the system level, where performance, power efficiency and compute density collectively determine infrastructure productivity.

For hyperscalers, AI cloud providers and enterprises scaling agentic workloads, these findings make a compelling case for Arm AGI CPU as a foundation for the next generation of AI infrastructure, where the defining measure of performance is the work completed within a given power and physical footprint. Signal65 plans to build on these findings with testing on final production systems, expanded Signal65 PINNACLE CPU-centric workloads and a wider set of competitive configurations.

Testing and Verification

Arm ran all benchmarks at its facilities. Signal65 observed the test process remotely and watched the runs as they executed. Signal65 analyzed run files, including logs, result files, configuration records and scripts. Signal65 parsed these results directly from the source files, computed each data point, and provided a consolidated summary.

- **Independent analysis of raw data.** Results come from the original run files. Each per-run value is traceable to its source file.
- **Independent review.** A separate review team rebuilt the calculations with its own code and reproduced all seven results within 0.5%.
- **Rated power.** Rack results use rated server power model: 950 W and 1,050 W per two-socket AGI CPU server at SKU #1 and SKU #2, 1,700 W per two-socket EPYC 9755 server, and a 24 kW rack.
- **Result Summary.** Each per-chip result is the Arm AGI CPU result divided by the EPYC 9755 result, values above 1.0 favor Arm. Six workloads measure throughput; Terminal-Bench measures completion time (AMD time / Arm time). Per rack is the per-chip ratio multiplied by the rack density factor.

Appendix A: Results and Calculations

Table 2. Per-workload results. Per rack = per chip x rack factor.

Workload	Category	SKU	EPYC 9755 SMT	Per chip	Rack factor	Per rack
MySQL	Tooling	2	On	0.88	1.62	1.43
NGINX	Tooling	2	On	1.20	1.62	1.95
Terminal-Bench	Agentic Workflows	1	On	1.09	1.77	1.93
SWE-bench	Agentic Workflows	1	On	1.13	1.77	2.00
Milvus	Reasoning / Planning	2	On	0.95	1.62	1.54
vLLM CPU overhead	Reasoning / Planning	2	Off	1.16	1.62	1.88
fastText	Reasoning / Planning	2	On	1.97	1.62	3.19
Average (arithmetic)				1.20		1.99

Table 3. Arm's rated server-power model (CPU power from TDP estimates; SKU 2 updated from silicon measurements).

	AGI CPU SKU #1	AGI CPU SKU #2	AMD EPYC 9755
CPU TDP (2 sockets)	500 W	600 W	1,000 W
Memory, I/O, chassis and cooling	450 W	450 W	700 W
Memory (included above)	192 W	192 W	336 W
Total server power (rated/modeled)	950 W	1,050 W	1,700 W
Servers per 24 kW rack (Arm model)	23	21	13
Physical cores per rack	5,888	5,712	3,328
Rack factor	1.77 (23/13)	1.62 (21/13)	1.00

Memory is sized at 4 GB per hardware thread, so the SMT-enabled EPYC 9755 server is modeled with about 2 TB of memory (336 W) versus about 1 TB (192 W) for AGI CPU servers. Sizing memory per physical core instead would lower the SKU #2 rack factor from 1.62x to about 1.41x, and the seven-workload average from 1.99x to about 1.81x.

This rack-level model takes into account not just the CPU power, but also power for other server components - memory, NIC, storage, management, cooling.

For calculating the number of servers per rack, the model reserves about 8% of rack power for networking and power distribution; the rest is available for servers. Server counts are rounded to the nearest whole server.

Sandbench Detail

Table 4. Sandbox creation time in ms, AGI CPU / AMD EPYC 9965; lower is better. Concurrent worst case is the slowest successful start of 32 (2x) or 64 (4x) starts.

Sandbox	Cold creation	Warm creation	Concurrent worst case, 2x	Concurrent worst case, 4x
firecracker	803 / 2,763	652 / 1,157	1,522 / 2,114	2,673 / 3,947
gvisor	331 / 436	196 / 228	1,308 / 2,221	2,574 / 4,508
runc	260 / 340	188 / 239	1,213 / 2,127	2,483 / 4,694

Appendix B: Test Configuration and Limitations

Table 5. Systems as recorded in the delivered configuration files.

	Arm AGI CPU	AMD EPYC 9755
Cores	136 (SKU #2) or 128 (SKU #1), one thread per core	128 per socket, 2 threads per core (SMT on); one socket used
Frequency / memory	3.2 GHz, DDR5-8000 (SKU #2); 2.8 GHz, DDR5-7200 (SKU #1)	Up to 4.1 GHz boost; DDR5-6400
Main test environments	Ubuntu 24.04; SWE-bench Ubuntu 26.04. Custom Linux 7.1.3 builds.	Main recorded baseline Ubuntu 24.04, Linux 6.8; configurations vary by test.
Sandbench system	Single-socket 128C AGI CPU SKU 1 system (2.8 GHz, DDR5-7200)	Two-socket AMD EPYC 9965, SMT on

- **Modeled rack results.** Per-rack values combine measured per-chip results with Arm's rated server power.
- **Software differences.** Kernels, operating system versions and some system settings differ between platforms. BIOS and memory settings were verified for the EPYC 9965 Sandbench system only (root sysinfo, Oct 7); they were not available for the EPYC 9755 and Arm systems and were not independently verified.
- **Run counts.** Most workloads have three or five runs per platform; MySQL, Terminal-Bench and Sandbench have one campaign each.
- **Kata Containers.** Sandbench results for Kata Containers are under investigation and are excluded.
- **Scope.** The study did not measure accelerator utilization or cloud instance pricing. Rack results depend on the stated system configurations and Arm's rated server power model.
- **Architectural attribution.** The tests compare complete systems and do not isolate the effects of SMT, cache, memory bandwidth, instruction set or other architectural factors.

Important Information About this Report

CONTRIBUTORS

Russ Fellows

VP, Labs | Signal65

Cameron Moccari

Director of Operations | Signal65

PUBLISHER

Ryan Shrout

President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and lab services to many high-tech companies, including those mentioned in this paper. Research of this document was commissioned by Arm.

ABOUT SIGNAL65

Signal65 exists to be a source of data in a world where technology markets and product landscapes create complex and distorted views of product truth. We strive to provide honest and comprehensive feedback and analysis for our clients in order for them to better understand their own competitive positioning and create optimal opportunities to market and message their devices and services.



IN PARTNERSHIP WITH

arm



CONTACT INFORMATION

Signal65 | signal65.com