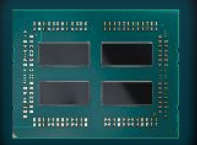
對抗硬體通膨，AMD助力節省基礎建設持有成本

AMD
together we advance_

教務系統基礎建設
節省解決方案

The AMD EPYC™ CPU Journey – Current Portfolio

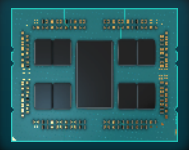
1st Gen AMD EPYC™



General Purpose

7001

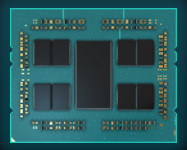
2nd Gen AMD EPYC™



General Purpose

7002

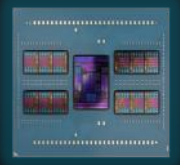
3rd Gen AMD EPYC™



General Purpose, Technical Computing

7003

4th Gen AMD EPYC™

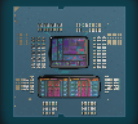


General Purpose, Technical Computing, Cloud

9004

Intelligent Edge

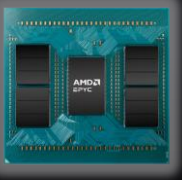
8004



SMB & Hosters

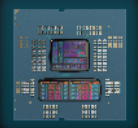
4004

5th Gen AMD EPYC™



General Purpose, Cloud

9005



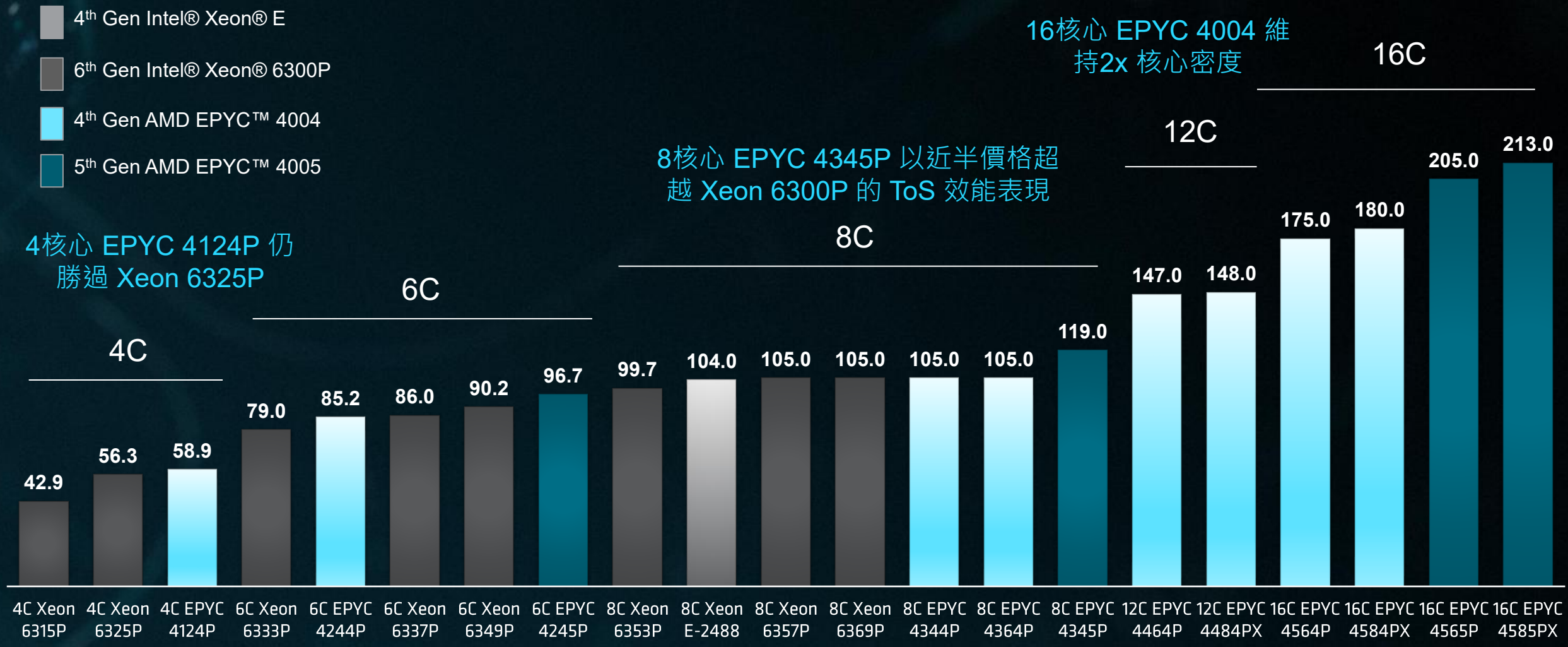
SMB & Hosters

4005

2017 → 2025

AMD EPYC™ 4004 and 4005 vs. Intel® 6300P

1P Server Ranked SPECrate®2017_int_base (max scores compared)

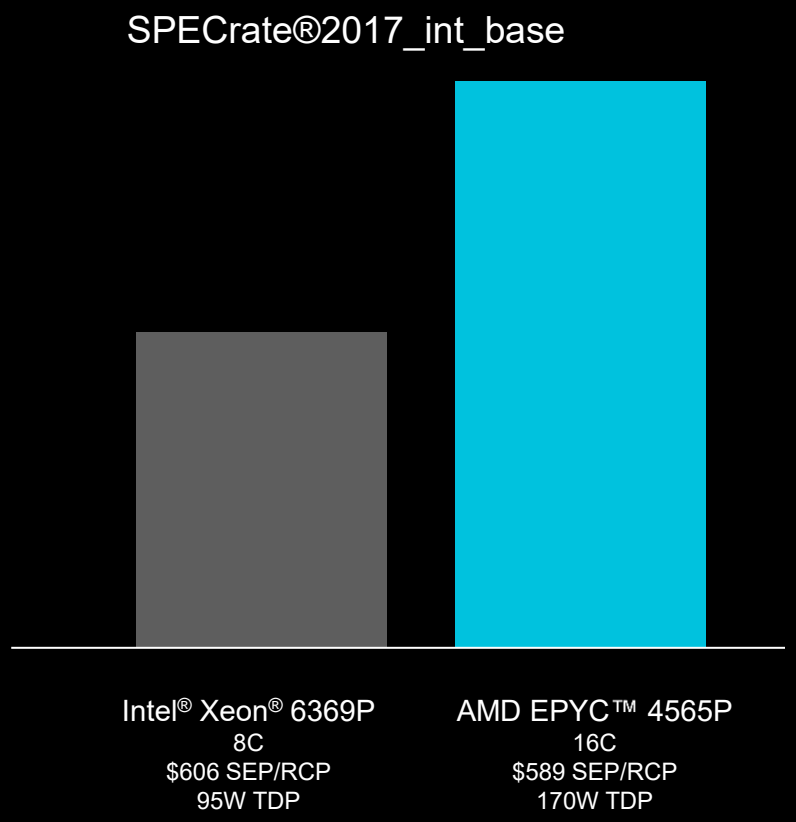


As of 5/13/2025. See endnote E4K-028A

AMD
together we advance_

Throughput Leadership

SPECrate®2017_int_base



1.95X 效能

2.01X 效能 / CPU 成本$

1.98X 效能 / 預估的系統成本 $

1.42X 效能 / 預估的系統電力 W

As of 5/13/2025. See endnote E4K-028

OPTIMIZE LICENSING

with AMD EPYC™ 4000 Series CPUs

Windows Server® 2025

- Standard Edition ~$900*
- Datacenter Edition ~$3,900*

*Additional costs for Client Access Licenses (CAL)

Source: https://www.mychoicesoftware.com/collections/windows-server-products as of 21Apr2025

- Windows Server® Datacenter and Standard edition 採用核心為基礎的授權模式，基本價格已包含16個核心
- AMD EPYC 4000 系列處理器提供多達 16 個核心，讓您能最佳化 Windows Server 授權，同時提供額外的擴充空間以配合業務成長。
- Intel E-2400 與 Intel 6300 CPU 系列僅提供最多 8 個核心。
- 聽取更多第三方意見 – YouTube Link

ODM: EPYC™ 4004/5 (AM5) Portfolio

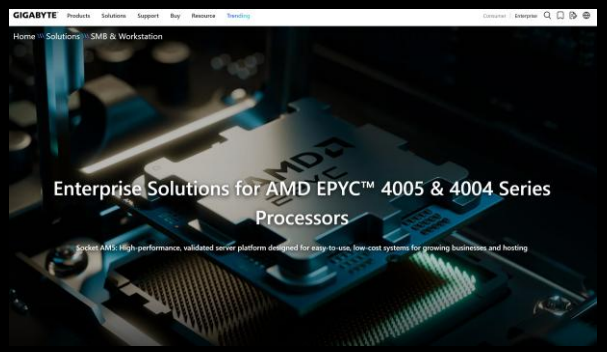
Summary Slide

ASRock Rack

Website Link

- Server:
  - 1U2N2G-AM5/2T
  - 1U2S-B650
  - 1U4L2E-B650 RPSU
  - 1U4LW-B650/2L2T
  - 1U4LW-B650/2L2T RPSU
  - 2U1G-B650
  - 2U1G-B650/EVAC
  - 2U1G-B650/AQUA
- Multi-node:
  - 4U18N-AM5/2T
- Motherboard:
  - EPYC4000D4U
  - AM5D4ID3-2T/X710
  - B650D4U3
  - B650D4U3-2Q/BCM
  - B650D4U3-2L2Q/BCM
  - AM5D4ID2
  - AM5D4ID-2L+/BCM
  - AM5D4ID-2T/BCM

Gigabyte



Website Link

- Server:
  - R113-C10
  - R133-C10/11/13
  - E133-C10
  - R123-C00
- Motherboard:
  - MC13-LE0/LE1
- Tower:
  - W332-Z00

MSI

Website Link

- Server:
  - S1102-01 / 02
- Motherboards:
  - D3050
  - D3051
  - D3052

MiTAC

Website Link

- Server:
  - B8016 (1U)
- Multi-node:
  - HG68-B8016 (6U)
- Motherboards:
  - Tomcat CX S8016

EPYC™ 4004 / 4005 Portfolio

AM5 socket

Single Socket Products

Rack

Product	DIMMs	TDP(W)
1U / 1P / Up to 4 internal HDD / 1 GPU	4	170
2U / 1P / 8 external HDD / 1 GPU	4	170

Tower

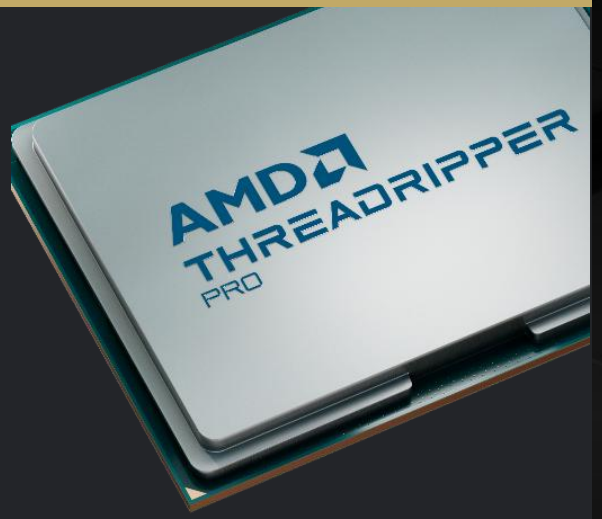
Product	DIMMs	TDP(W)
1P/ 4HDD/ 2 GPU	4	170

3U

Product	DIMMs	TDP(W)
MicroCloud 8 Node 3U / 1P / 2HDD / 1 LP GPU	4	170
MicroCloud 10 Node 3U / 1P/ 2HDD / 1 LP GPU	4	170
MicroCloud 5 Node 3U / 1P / 2HDD / 1 FH/FL GPU	4	170

Indicates hyperlink to product page

AI 電腦教室基礎建設
節省解決方案



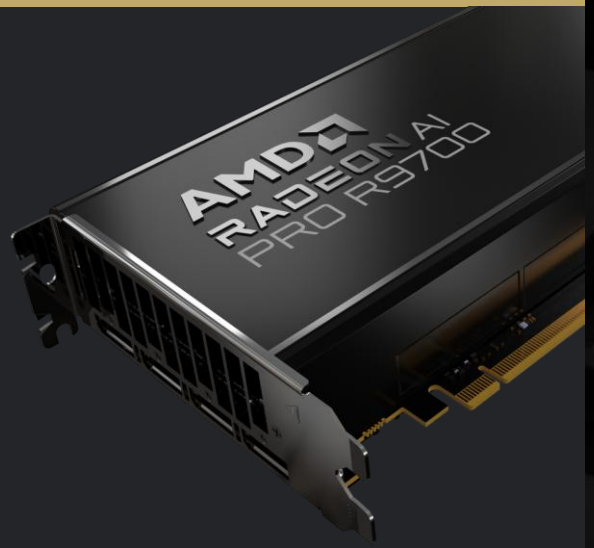
AMD Ryzen™ Threadripper™

9000 Series

PROCESSORS

AMD AI Workstation Technologies

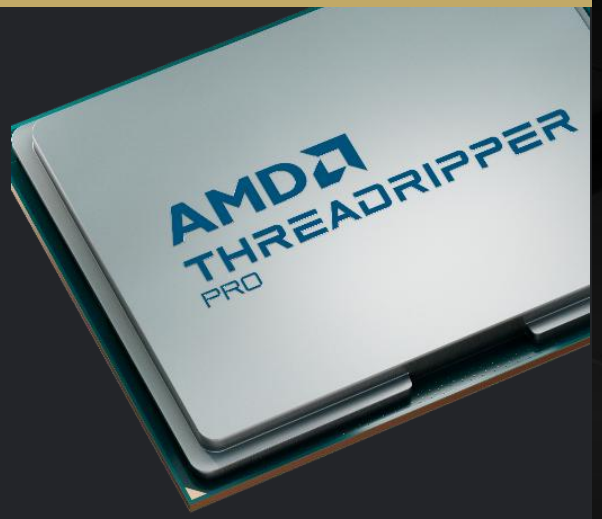
Powering Next Generation Innovation



AMD Radeon™

9000 Series

GPUs



AMD Ryzen™ Threadripper™

9000 Series

PROCESSORS

AMD AI Workstation Technologies

Powering Next Generation Innovation

AMD Radeon™

9000 Series

GPUs

AMD Ryzen™ Threadripper™ PRO 9995WX

vs. Xeon W9-3595X

AMD
together we advance_

Get the most out of your GPU for AI

up to
+49%

DeepSeek R1 32B
CPU + GPU, context-based prompting
Tok/sec

See endnotes SHP-06

AMD Ryzen™ Threadripper™ PRO 9995WX

vs. Xeon W9-3595X

AMD
together we advance_

Get the most out of your GPU for AI

up to
+34%

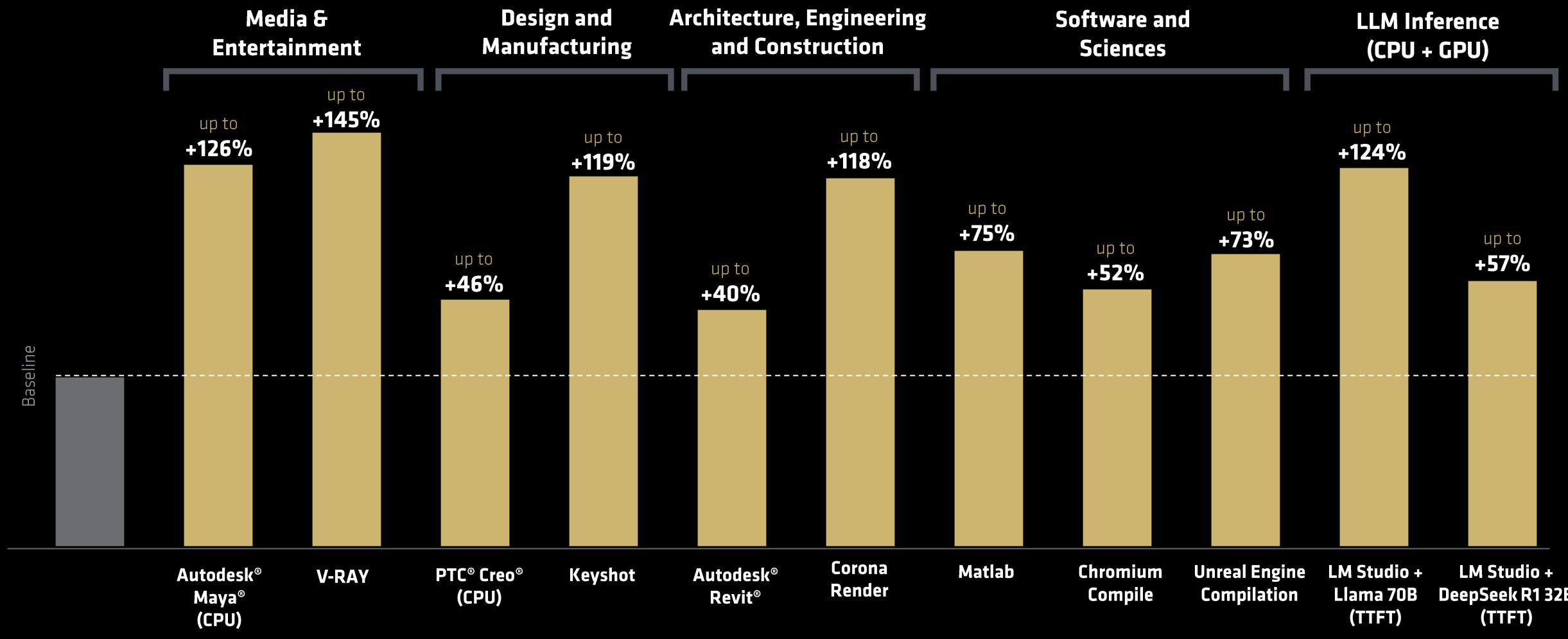
ComfyUI + FLUX.1 [schnell] diffusion model
CPU + GPU, text-to-image (2176x1024), elapsed time

See endnotes SHP-08

AMD Ryzen™ Threadripper™ PRO 9995WX

vs. Xeon W9-3595X – Max Performance for all Professionals

AMD
together we advance_



See endnotes SHP-03, SHP-05

AMD
together we advance_

AMD Ryzen™ Threadripper
9000 Series
PROCESSORS

AMD AI Workstation Technologies

Powering next generation Inspiration

AMD Radeon™
9000 Series
GPUs

AMD Radeon™ AI PRO R9700

Graphics

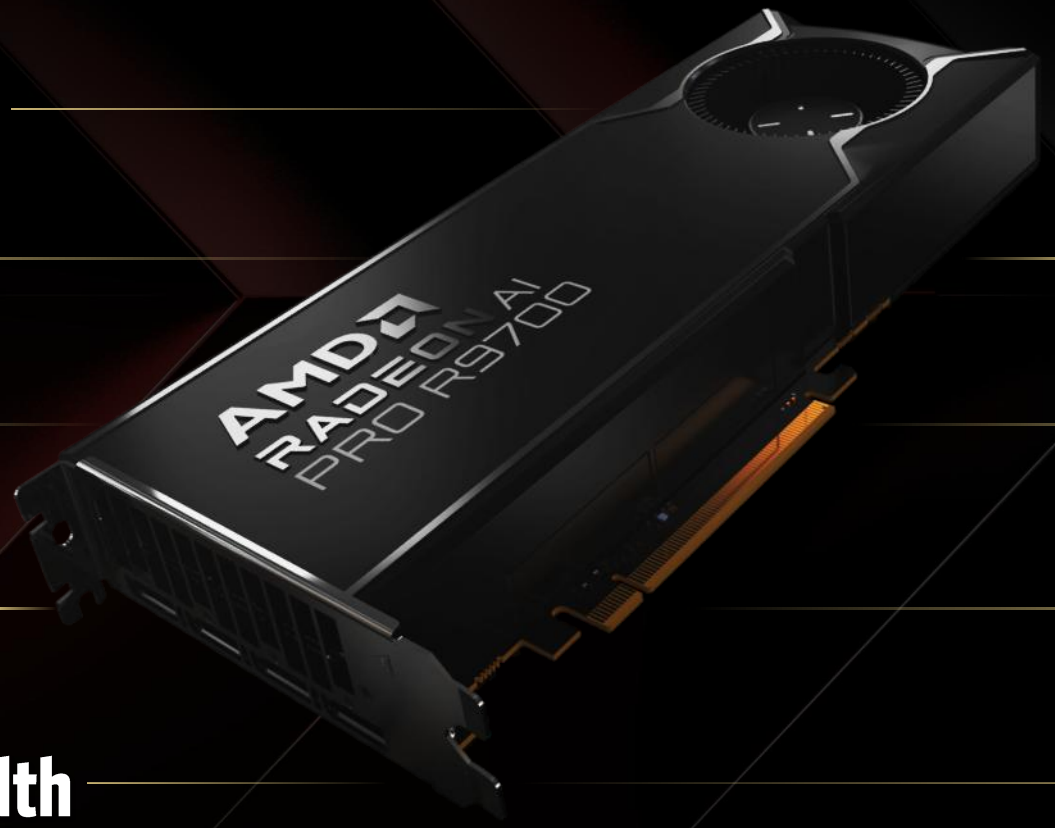
AMD RDNA4™ Architecture

Dual Slot Design

32GB GDDR6

256-bit Memory Interface

640 GB/s Memory Bandwidth



64 Compute Units

128 AI Accelerators

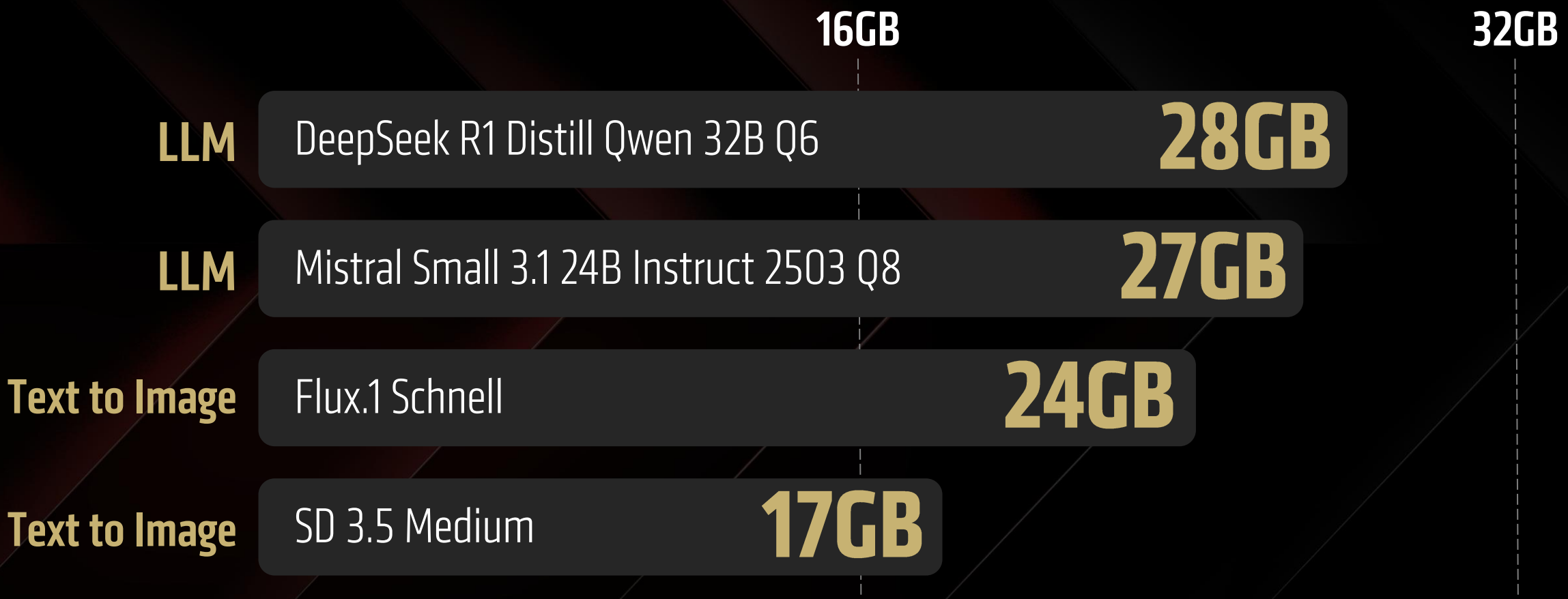
Up to **191 TFLOPS**
FP16 Dense

Up to **1531 TOPS**
INT4 Sparse

300W TDP

Optimal VRAM Buffer for Advanced Local AI

Typical VRAM Usage by Popular Models

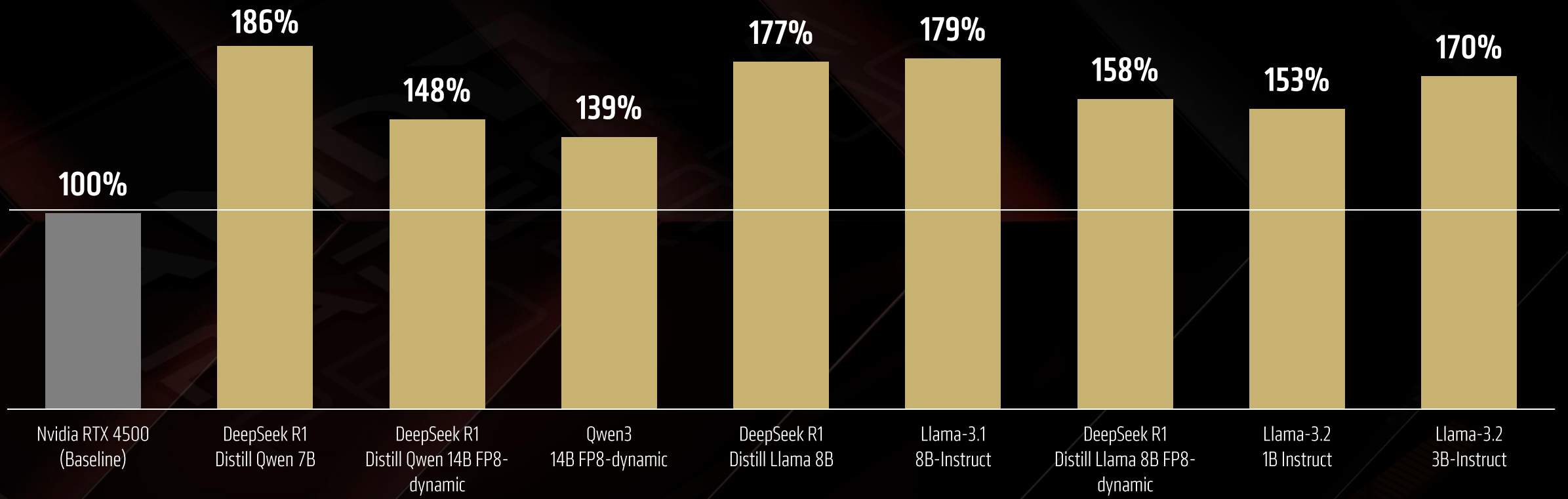


See Endnote RPW-496

Exceptional Value – Single GPU

AMD Radeon™ AI PRO R9700 vs. Nvidia RTX 4500 (Blackwell)

Value (Performance / $), based on Average Tok/Sec



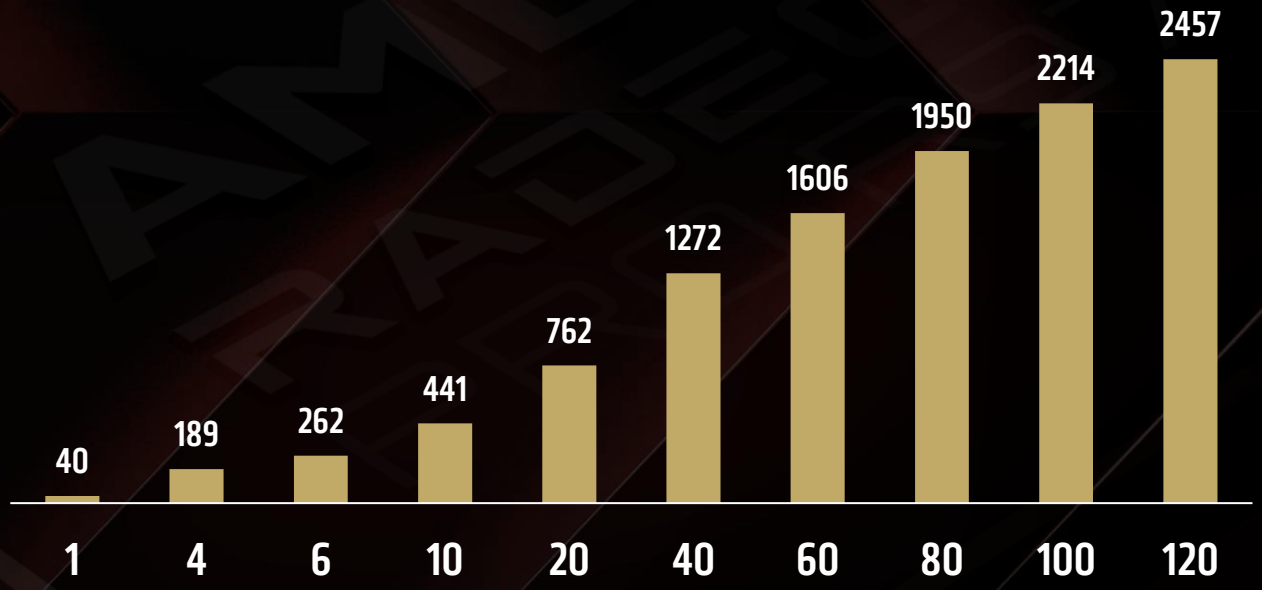
See Endnote RPW-502

Next-Gen Scalability

Multi-GPU PCIe® 5 platform with up to 120 Concurrent Instances

Tokens/sec

(4x) Radeon AI Pro R9700 | DeepSeek-R1-Distill-Qwen-32B



See endnotes RPW-499

WS990T

專業型工作站由 AMD RYZEN™ Threadripper™ PRO 7000 WX 系列處理器驅動，配備 DDR5 ECC RDIMM 記憶體、雙 10 Gb 網路和最多 4 張 NVIDIA® Quadro RTX™ 顯示卡。支援最多四個 PCIe® 5.0 M.2 插槽、四個 SATA 6Gb/s 連接埠、40Gbps 和 20Gbps USB Type-C、高達 2000W 鈦金級電源供應器，以及 ASUS Control Center Express。

GAI4G-R9700

This model is for business customer only, and not be sold in retail channel.

- 4U Rackmount / Tower with 1, 80-PLUS Platinum, 2500W ATX 3.1 PSU
- 4 x ASRock AMD Radeon™ AI PRO R9700 Creator 32GB
- Supports AMD Ryzen™ Threadripper™ PRO 9000/7000 WX-Series Processors
- 8 x DDR5 DIMMs
  Supports Eight Channel ECC Registered memory, up to 6400 MT/s
- 16 x DisplayPort 2.1a (from 4 Graphics Cards)
  1 x DisplayPort 1.1a (from BMC)
- 2 x 2.5" Drive Bays
  1 x M.2 M-key (PCIe 5.0 x4)
  1 x M.2 M-key (PCIe 4.0 x4)
- 2 x USB 4 Type-C (Rear)
  4 x USB 3.2 Gen2 Type-A (Rear)
  4 x USB 3.2 Gen1 Type-A (2 x Rear, 2 x Front)
- 2 x Intel® X710 10Gb LAN
  1 x Dedicated IPMI LAN

AMD
together we advance_

AMD Ryzen™ AI Max PRO Series Processors

為小型工作站重新定義效能

ZEN 5	AMD RDNA 3.5	AMD XDNA 2	
桌機等級 CPU 核心數	獨顯等級 整合式顯示卡	新型節能型 AI 運算單元 (NPU)	統一記憶體架構
Up to **16 Cores** and 32 threads	Up to **40 CUs** 80 AI accelerators ISV Certifications	Up to **50+ TOPS** Peak AI Performance*	Up to **128/96GB** Unified/GPU Memory

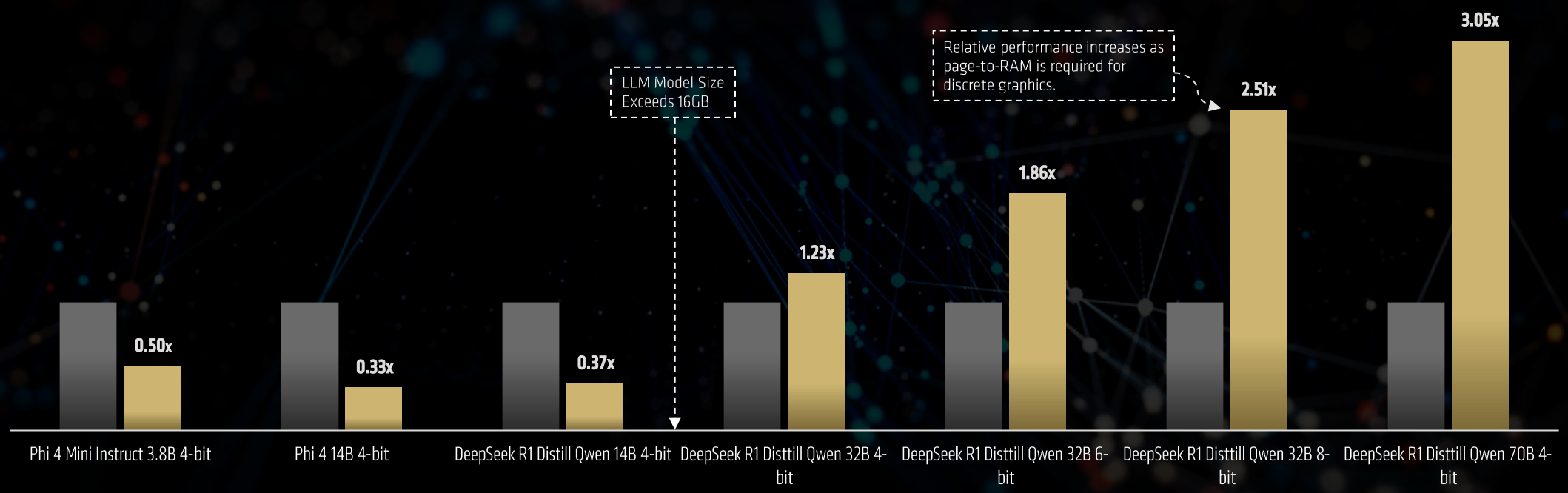
Scalable Solution to Power The Next Generation of Copilot+ PC Workstations

* See endnote GD-243

CU = Compute Units

AMD RYZEN™ AI 9 MAX 395+ PROCESSOR 128GB

LM STUDIO 0.3.11 – *TOKENS PER SECOND* – LARGE LLMS AND QUANTIZATIONS REQUIRE LARGE VRAM



Testing as of March 2025 by AMD. All tests conducted on LM Studio 0.3.11. Llama.cpp runtime 1.18. Tokens/s using prompt "How long would it take for a ball dropped from 10 meter height to hit the ground?" Models tested: Phi 4 Mini Instruct Q4 K M, Phi 4 Q4 K M, DeepSeek R1 Distill Qwen 14b Q4 K M, DeepSeek R1 Distill Qwen 32b Q4 K M, DeepSeek R1 Distill Qwen 32b Q6, DeepSeek R1 Distill Qwen 32b Q8 and DeepSeek R1 Distill Llama 70b Q4 K M AMD Ryzen™ AI MAX+ 395 CRB 55W with 128GB 8000 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.3.1 WHQL. NVIDIA GeForce RTX 5080 16GB with a Ryzen 7 9800 X3D and 32GB 6000 MT/s memory, Windows 11 Pro 24H2 and GeForce 572.47 drivers. Performance may vary. SH0-28.

Available now in LM Studio

AMD RYZEN AI MAX+

Industry leading Windows support

For LLM deployment in thin and light

Meta Llama 4 Scout 109B (17B active)
Mistral Large 123B
DeepSeek R1 Distill Llama 70B
Nemotron 70B
Llama 3.3 70B
Qwen3 30B A3B
Qwen3 32B
DeepSeek R1 Distill Qwen 32B
Google Gemma 3 27B
Google Gemma 3 12B
Mistral Nemo 12B
Devstral 24B
Microsoft Phi 4 Reasoning Plus 14B
DeepSeek R1 Distill Qwen 8B 0528
Meta Llama 7B
IBM Granite Vision 3.2 2B
Microsoft Phi 4 Mini 3B

4-bit to 16-bit quantization sizes in llama.cpp

Available now in LM Studio

Run MCP-sized context length

MCP tool calls require large context sizes

4096 Context Length

LM Studio Defaults

Not enough for typical MCP use-cases.

Parsing AMD.com's landing page with a Microsoft Playwright navigate_browser tool call returns 9358 tokens as it tokenizes the entire contents of the page and will not fit inside the default context length of 4096.

32,000 Context Length

Flash Attention: ON, KV Cache Q8

Fast and agile: Google Gemma 3n E4b

3 tool calls with a similar token length in-context would require a context size of 28,074.

200,000 Context Length

Flash Attention: ON, KV Cache Q8

Power User: Llama 4 Scout

The 96 GB Variable Graphics Memory on AMD Ryzen™ AI MAX+ 128 GB is sufficient to hold up to 21 such tool calls in-context.

AMD Ryzen™ AI Max+ 395 128GB and AMD Software: Adrenalin Edition™ 25.8.1 WHQL

AMD × LM Studio
together we advance_

Competitive Comparison

HP Z2 Mini G1a vs Nvidia DGX Spark

AMD × hp
together we advance_

	CPU	GPU	NPU	Memory	Memory Bandwidth	Networking I/O	OS Support
HP Z2 Mini G1a	"Strix Halo" 16 "Zen 5" Cores / 32 threads (x86)	RDNA 3.5 (40 CUs)	YES (50 TOPS)	128GB LPDDR5x (UMA)	256GB/s	2.5Gbps Ethernet Dual USB4 (2x 40Gbps)	ubuntu Windows 11
NV DGX Spark	"Grace Blackwell GB10" 10p+10e Cores (ARM)	Blackwell	NO	128GB LPDDR5x (UMA)	273GB/s	10Gbps Ethernet NV ConnectX (2x 200Gbps)	NV DGX OS 7

*Based on MSRPs on nvidia.com and hp.com as of October 2025.

AMD Ryzen™ AI Max PRO Series Processors

Product Stack

		CPU					GPU	
AMD Ryzen™ AI **Max+ PRO**	395	16 cores 32 threads	up to 5.1 GHz max boost	80 MB cache	up to 50+ TOPS NPU* Copilot+PC	55W 120W	AMD Radeon™ 8060S Graphics incl. Radeon PRO driver	**40 CUs (80 AI accelerators)**
AMD Ryzen™ AI **Max PRO**	390	12 cores 24 threads	up to 5.0 GHz max boost	76 MB cache			AMD Radeon™ 8050S Graphics incl. Radeon PRO driver	**32 CUs (64 AI accelerators)**
AMD Ryzen™ AI **Max PRO**	385	8 cores 16 threads	up to 5.0 GHz max boost	40 MB cache			AMD Radeon™ 8050S Graphics incl. Radeon PRO driver	**32 CUs (64 AI accelerators)**
AMD Ryzen™ AI **Max PRO**	380	6 cores 12 threads	up to 4.9 GHz max boost**	22 MB cache			AMD Radeon™ 8040S Graphics incl. Radeon PRO driver	**16 CUs (32 AI accelerators)**

* See endnote GD-243

** See endnote GD-150

AMD
together we advance_

Competitive Comparison

Performance Results – LLM Inference

AMD × hp
together we advance_

Strong Tokens/s performance

TTFT <1 sec (8 out of 9 tests)

Workload	App	Model	HP Z2 Mini G1a 128GB, 96GB VRAM Ubuntu 24.04 LTS ROCm 7.9 Preview		DGX Spark Ubuntu 24.04, 580.95 .0 CUDA13	
			tokens/sec	TTFT (s)	tokens/sec	TTFT
LLMs	LM Studio (llama.cpp)	gpt-oss-20b-GGUF	65.65	0.15	54.97	0.09
		gpt-oss-120b-GGUF	45.83	0.65	38.69	0.72
		Qwen3-32B-GGUF Q4_K_M	10.08	0.29	8.55	0.29
		Qwen3-32B-GGUF Q4_K_M (large prompt)	8.67	23.16	8.55	7.48
		gemma-3-27b Q4_0	12.36	0.18	11.85	0.24
		qwen3-coder-30b Q4_K_M	68.85	0.14	60.66	0.76
		qwen3-coder-30b Q8	42.85	0.14	39.83	0.35
		DeepSeek R1 Distill Llama 70B GGUF Q4_K_M	4.67	0.70	4.07	0.50
		GLM 4.5 Air Q4 K M	18.34	0.48	16.17	0.26

LM Studio Prompts
Prompt 1: how long will it take for a ball dropped from a 10 m height to hit the ground?
Prompt 2: Summarize the following in exactly five lines: https://www.folger.edu/explore/shakespeares-works/romeo-and-juliet/read/1/1/
Prompt 3: Create a browser-based OS using html, js and css

AMD × hp
together we advance_

AMD Ryzen™ AI Max PRO Series Processors Powering Comprehensive AI Workstation Solutions

Choice of Operating System

Support for Windows® and Linux empowers AI developers with a more flexible solution

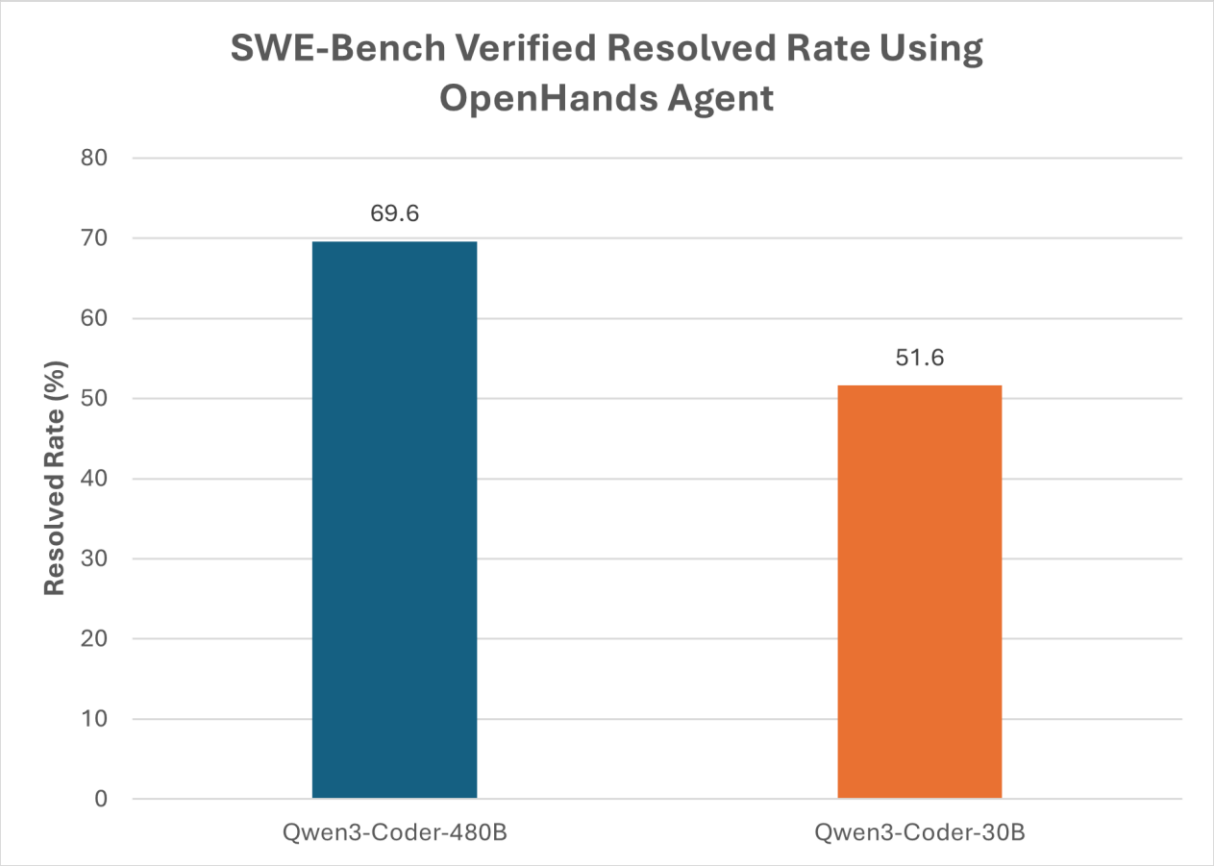
Broad AI workload support

Cutting-edge on-device AI compute capabilities for more industries and workloads

Cost Benefit

Lower cost of entry with numerous config options

CODING AGENT – 單純成本考量



什麼適合在 local 做（省最多）

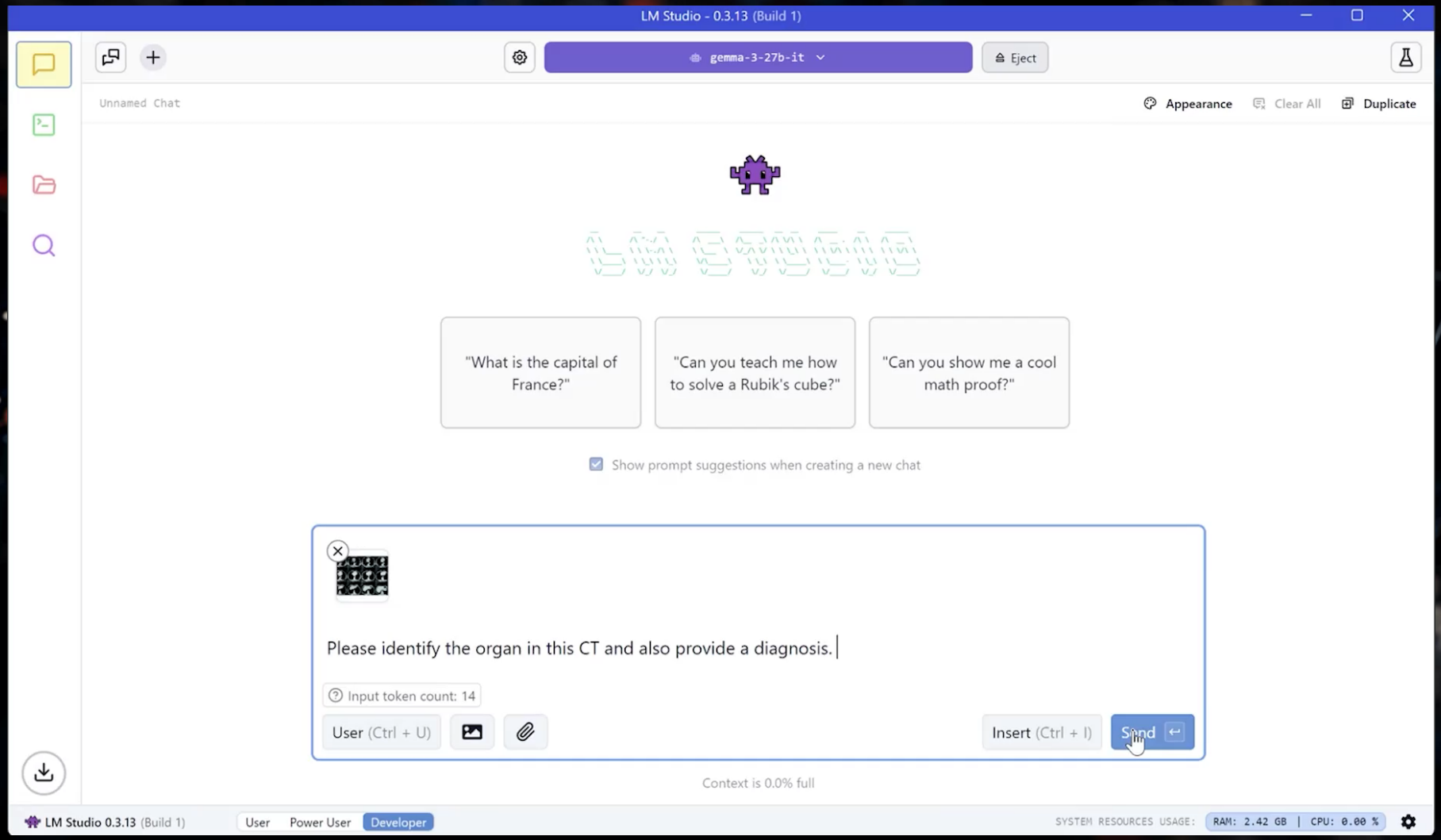
- ❑ 樣板生成：CRUD、API client、DTO/Schema、Terraform/Helm、CI YAML、README、註解
- ❑ 小步重構：命名、抽函式、拆模組、搬檔案、加型別、加 lint
- ❑ 單元測試雛形：先把測試框架、mock、測試案例骨架吐出來
- ❑ 多輪嘗試：你本來要試 10 次 prompt 才能產出雛形的東西 → 都放 local，省的就是「迭代次數的 API 錢」
- ❑ 核心概念：把 token 大頭（大量生成 + 多次改寫）留在 local。

什麼時候再丟到 cloud（花得值得）

- ❑ 很棘手的 debug：跨模組、時序/併發、memory leak、邏輯推理鏈很長
- ❑ 需要高精度定位錯誤：看 stack trace + 多檔案上下文 + 提出最小修復
- ❑ 需要高品質 code review：安全性、邊界條件、性能瓶頸、架構建議
- ❑ 卡在「為什麼」：local 常能寫出「看起來對」但其實有 subtle bug；cloud 用來做最後的「判案」
- ❑ 核心概念：把 cloud 用在「少次數、高價值」的推理與驗證。

https://www.amd.com/en/developer/resources/technical-articles/2025/OpenHands.html
https://www.amd.com/en/developer/resources/technical-articles/2025/ryzen-ai-radeon-llms-with-lemonade.html

FASTEST X86 PROCESSOR FOR GOOGLE GEMMA 3 27b VISION MODEL



See Endnote: SHO-27

https://www.amd.com/en/blogs/2025/amd-ryzen-ai-max-395-processor-breakthrough-ai-.html

AMD Ryzen™ AI Max PRO Series Processors

New Desktop Workstation Designs

"Mini Workstation.
Transformative AI Performance."

Available Now

Dimensions:
3.4"H x 6.6"W x 7.9"D (8.55cm x 16.8cm x 20cm)
Weight: Starting at 5.07lbs / 2.4kg

Integrated power supply

AMD RYZEN AI MAX PRO Series

Rear view

Copilot+PC

Side view

AI Max 300 Box Series

AI BOX-A395

Fanned BOX PC

ASRock Industrial

KEY FEATURES

- AMD Ryzen™ AI Max 300 Series (Strix Halo)
- Supports up to 128GB LPDDR5X-8000 memory
- 2 x USB4, 1 x USB 3.2 Gen2 Type C, 2 x USB 3.2 Gen2, 2 x USB 2.0
- 1 x M.2 Key E (WiFi7 Module)+2 x M.2 Key M
- 1 x Realtek 2.5 Gigabit LAN, 1 x Marvell 10 Gigabit LAN
- Supports Quad display, 2 x HDMI 2.1, 3 x DP 2.1 (2 from USB4, 1 from Type-C)
- 1 x Mic-in, 1 x Line-out
- TPM 2.0 onboard IC
- Support Redundant BIOS
- Supports RAID 0/1
- 1 x 400W FLEX ATX
- 200 x 100 x 232mm (7.87" x 3.94" x 9.13"), Fanned Barebone


AMD
together we advance_