tom's HARDWARE

Sign in

TRENDING Back to School Deals The death of Win 11 SE Where to Buy Switch 2 MI350X and MI355X AI GPUs

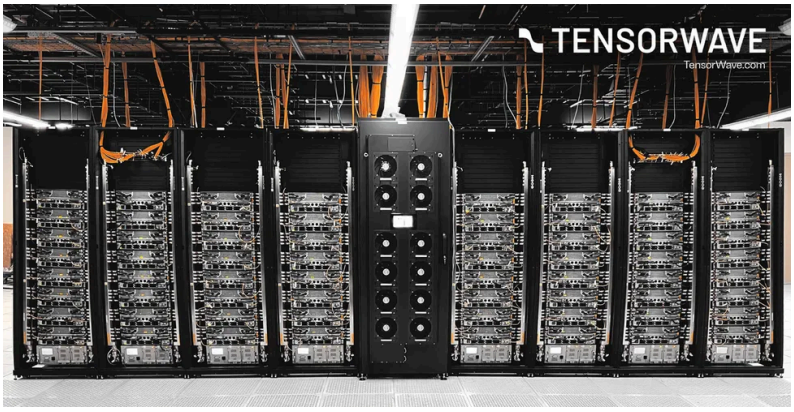
TensorWave just deployed the largest AMD GPU training cluster in North America — features 8,192 MI325X AI accelerators tamed by direct liquid-cooling

News By Hassam Nasir published July 14, 2025

Infinity Fabric, infinite confidence.

Comments (3)

When you purchase through links on our site, we may earn an affiliate commission. Here's how it works.



(Image credit: TensorWave)

AI infrastructure company TensorWave has revealed the deployment of a massive 8,192-GPU cluster powered by AMD's latest Instinct MI325X

ADVERTISEMENT

Ex.1068 / IPR2025-00916 / Page 1 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.

accelerators, claiming to be the largest AMD-based AI training installation in North America to date. The system also features direct liquid cooling, making it the first public deployment of its kind at this scale. Announced on X, the company showcased photos of the cluster's high-density racks with bright orange cooling loops, confirming the system is now fully operational.

8,192 liquid-cooled MI325X GPUs.The largest AMD GPU training cluster in North America.Built by TensorWave. Ready for what's next pic.twitter.com/RlFY4v2JDuJuly 12, 2025

See more

The AMD Instinct MI325X, officially launched late last year, was the company's most aggressive attempt to challenge NVIDIA in the AI accelerator space—up until it was succeeded by MI350X and MI355X last month. Regardless, each MI325X unit features 256GB of HBM3e memory, enabling 6TB/s of bandwidth, along with 2.6 PFLOPS of FP8 compute, thanks to its chiplet design with 19,456 stream processors clocked up to 2.10GHz.

The GPU confidently stands its ground against Nvidia's H200 while being a lot cheaper, but you pay that cost elsewhere in the form of an 8-GPU cluster limitation compared to the Green Team's 72. That's one of the primary reasons it didn't quite take off and precisely what makes TensorWave's approach so interesting. Instead of trying to compete with scale per node, TensorWave focused on thermal headroom and density per rack. The entire cluster is built around a proprietary direct-to-chip liquid cooling loop, using bright orange (sometimes yellow?) tubing to circulate coolant through cold plates mounted directly on each MI325X.

ADVERTISEMENT

ADVERTISEMENT

LATEST VIDEOS FROM TOM'S HARDWARE



YOU MAY LIKE



Pegatron preps 1,177 PFLOP AI rack with 128 AMD MI350X GPUs

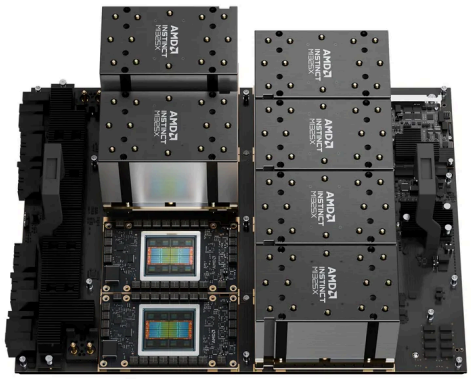


AMD's Instinct MI355X accelerator will reportedly consume 1,400 watts



[UPDATED] AMD says Instinct MI400X GPU is 10X faster than MI300X, will power Helios rack-scale system with EPYC 'Venice' CPUs

At 1,000 watts per GPU, running even a fraction of this hardware takes serious engineering. Thankfully, there are no 16-pin power connectors in sight. Anyhow, a total of 8,192 GPUs will produce over 2 petabytes/s of aggregate memory bandwidth and an estimated 21 exaFLOPS of FP8 throughput, though, as always, sustained performance depends heavily on model parallelism (splitting the AI model across GPUs) and interconnect design. TensorWave's business model is cloud capacity for rent, so the real challenge of scaling models falls to the tenants themselves.



(Image credit: AMD)

This installation follows TensorWave's $100 million Series A round from May, led by AMD Ventures and Magnetar. Unlike most cloud vendors that build primarily around NVIDIA hardware, TensorWave is going all-in on AMD, not just for pricing flexibility, but because they believe ROCm has matured enough for full-scale model training. Of course, NVIDIA still dominates the landscape. Its B100 and H200 accelerators are everywhere, from AWS to CoreWeave, and the entire AI boom seems to be held up by them, but this development shows positive signs for AMD's foothold in the AI sector.

As such, TensorWave's deployment isn't a one-off. According to the team, this is phase one of a much larger rollout, with plans to integrate AMD's MI350X later this year. That chip, based on CDNA 4, introduces FP4 and FP6 precision support, higher bandwidth ceilings, and more power-hungry designs that could push TDPs up to 1,400W per chip—something that couldn't be tamed by air alone, so TensorWave already seems to be on the right track. The ROCm stack still has work to do, but with 8,192 MI325X GPUs already humming away under liquid cooling, AMD finally has the scale to prove it belongs in the same conversation.

Follow Tom's Hardware on Google News to get our up-to-date news, analysis, and reviews in your feeds. Make sure to click the Follow button.

Ex.1068 / IPR2025-00916 / Page 3 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.

Stay On the Cutting Edge: Get the Tom's Hardware Newsletter

Get Tom's Hardware's best news and in-depth reviews, straight to your inbox.

Your Email Address

SIGN ME UP

Contact me with news and offers from other Future brands

Receive email from us on behalf of our trusted partners or sponsors

By submitting your information you agree to the Terms & Conditions and Privacy Policy and are aged 16 or over.

SEE MORE GPUS NEWS ▶

TOPICS

AMD

SEE ALL COMMENTS (3)



Hassam Nasir Contributing Writer

Hassam Nasir is a die-hard hardware enthusiast with years of experience as a tech editor and writer, focusing on detailed CPU comparisons and general hardware news. When he's not working, you'll find him bending tubes for his ever-evolving custom water-loop gaming rig or benchmarking the latest CPUs and GPUs just for fun.

READ MORE

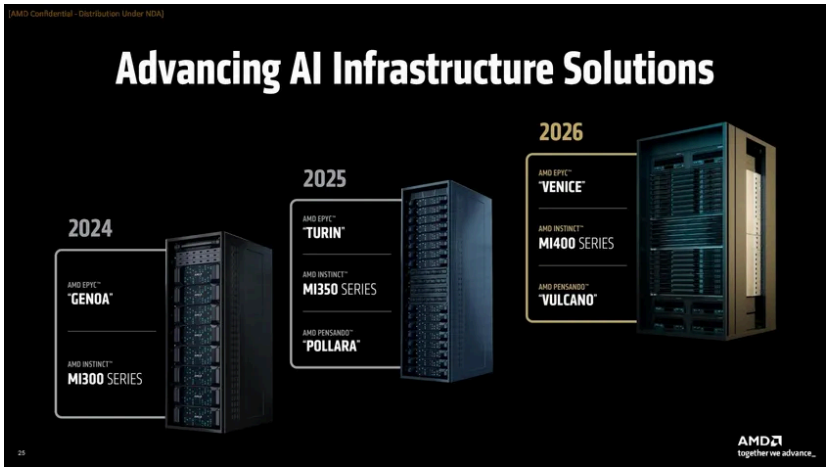


Pegatron preps 1,177 PFLOP AI rack with 128 AMD MI350X GPUs



AMD's Instinct MI355X accelerator will reportedly consume 1,400 watts

Ex.1068 / IPR2025-00916 / Page 5 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.



[UPDATED] AMD says Instinct MI400X GPU is 10X faster than MI300X, will power Helios rack-scale system with EPYC 'Venice' CPUs

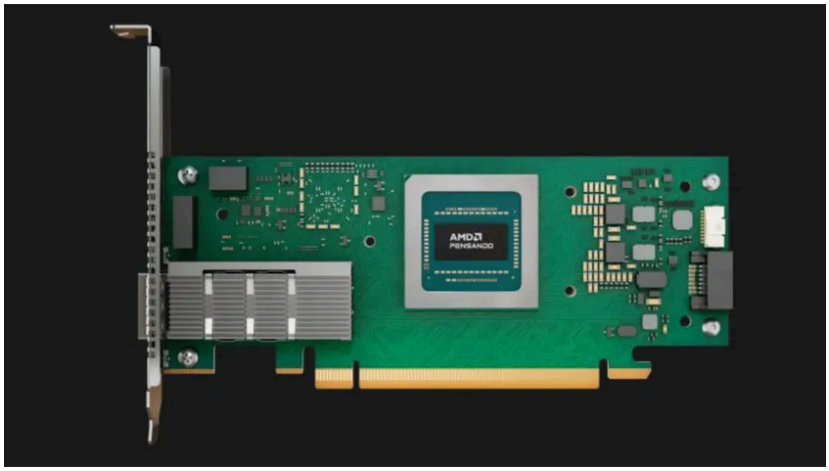
Ex.1068 / IPR2025-00916 / Page 6 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.



AMD preps rack-scale Instinct MI450X IF128 for 2026 release



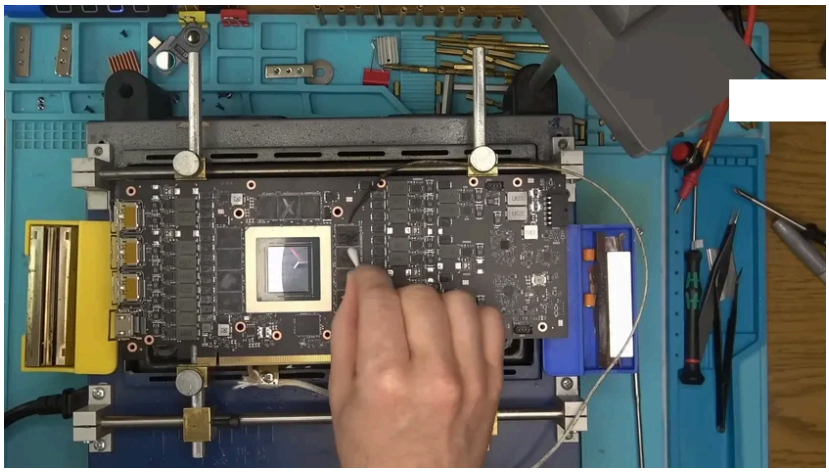
Nvidia's newest top-tier AI supercomputers deployed for the first time



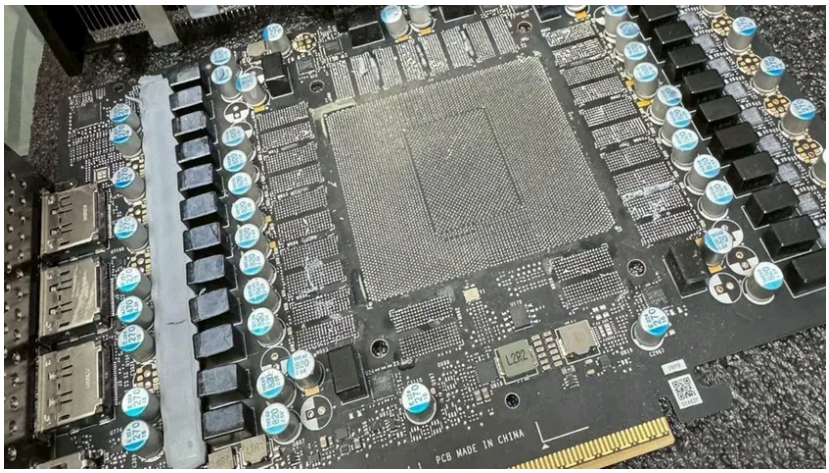
AMD deploys its first Ultra Ethernet ready network card — Pensando Pollara provides up to 400 Gbps performance

LATEST IN GPUS

Ex.1068 / IPR2025-00916 / Page 8 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.



Dead RTX 4090 with 48GB VRAM gets cooked beyond salvation



Retailer ships RTX 5090 with missing GPU and memory chips to customer

Ex.1068 / IPR2025-00916 / Page 9 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.

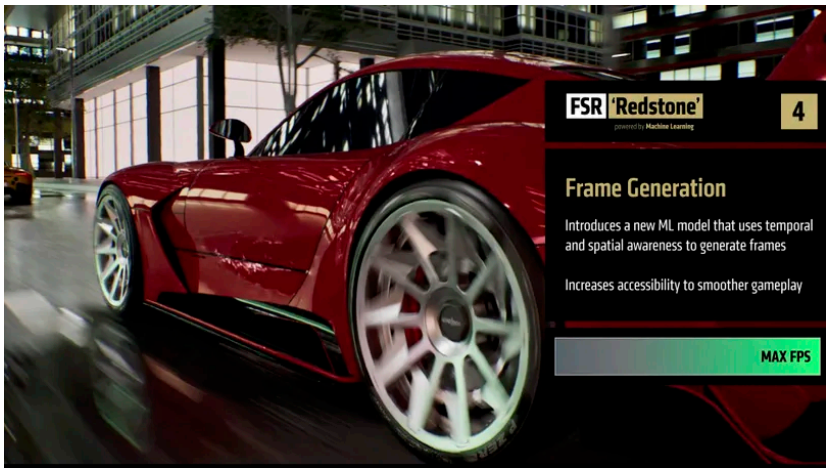


Battlefield 6 keeps crashing on Intel Arc GPUs

Ex.1068 / IPR2025-00916 / Page 10 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.



AMD quietly launches Radeon RX 7400 8GB budget gaming GPU — RDNA 3 goes another rung lower on the ladder



AMD FSR 3.1.4 update improves image quality, eases future developer integrations

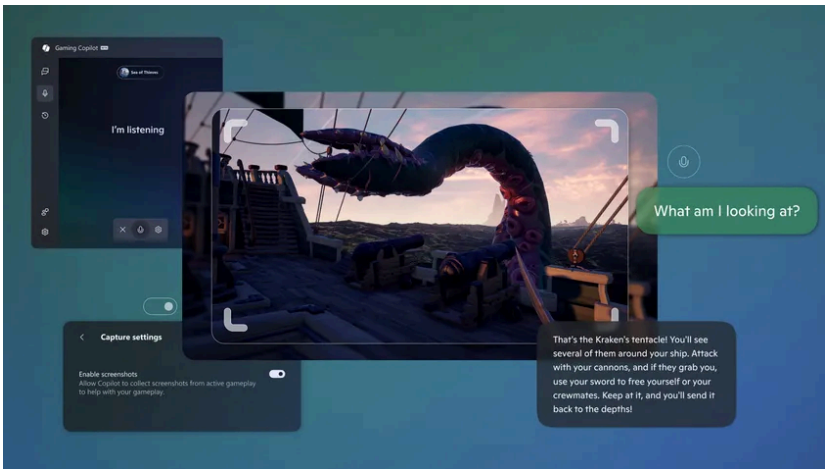


The best RTX 5070 GPU deal yet, grab a record-low MSI Ventus 2X OC for a mere $529

LATEST IN NEWS

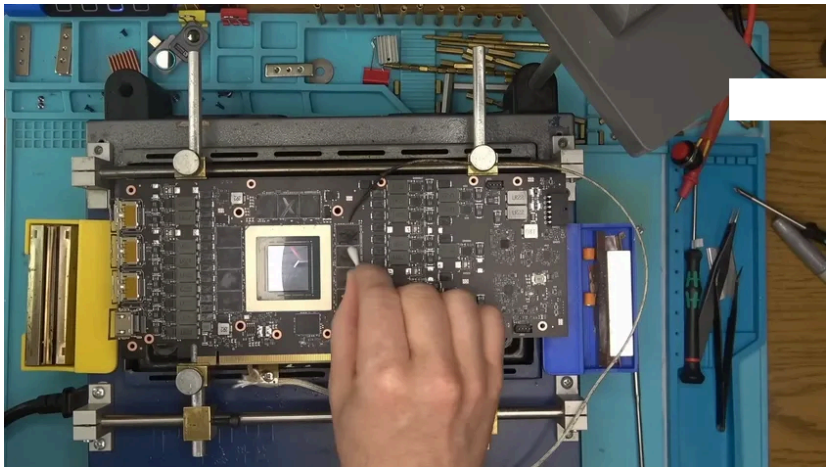


New local AI integration into Firefox spurs complaints of 'CPU going nuts'

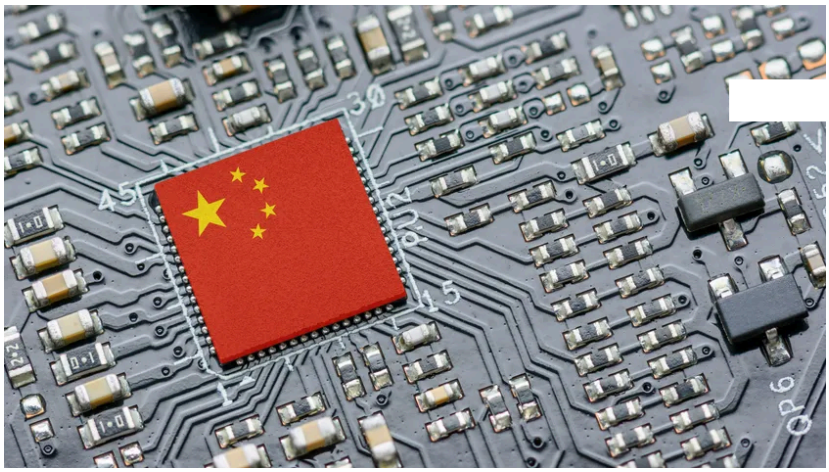


Microsoft brings AI-powered assistance to gaming

Ex.1068 / IPR2025-00916 / Page 13 of 17
Intel Corporation et al v. Advanced Cluster Systems, Inc.



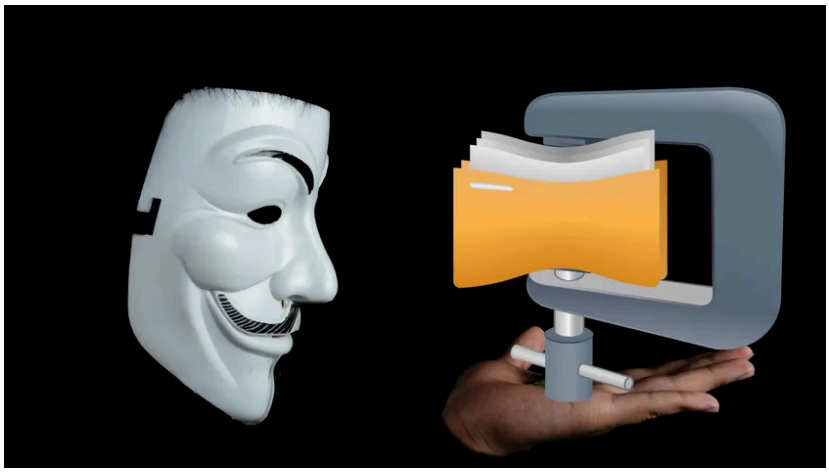
Dead RTX 4090 with 48GB VRAM gets cooked beyond salvation



China state media says Nvidia H20 GPUs are unsafe and outdated, urges Chinese companies to avoid them



Linus Torvalds calls RISC-V code from Google engineer 'garbage' and that it 'makes the world actively a worse place to live'



Newly discovered WinRAR exploit linked to Russian hacking group, can plant backdoor malware

3 COMMENTS COMMENT FROM THE FORUMS ▶

jp7189

This is a great first step, but there are still missing pieces. We need legions of young people just graduating from school that have

been tinkering with rocm on their gaming rigs and are willing to put in the extra effort to save a buck. Maybe if amd enables the full rocm stack on everything they make with a GPU and can keep it going for a few years.. then maybe they'll have something. They are working towards it, but they have a ways to go.

The problem is, I dont think they have the bank roll to keep it going that long. As it stands, AMD ventures is paying tensor wave to buy gpus from AMD to show design wins and revenue on the AMD datacenter side. That's a very expensive shell game. The snake can only eat its tail for so long.

REPLY ▶

thisisaname

Just leaves the big question how does it compare with Nvidia's offering?

REPLY ▶

DS426

> ***jp7189*** **said:**
>
> This is a great first step, but there are still missing pieces...

AMD knows and pretty much figures hardware and software won't be directly comparable to nVidia's until MI450X and ROCm of that time in 2026. That said, yeah, kids and adults alike need to be tinkering with this stuff now. The whole ecosystem has changed rapidly enough that anecdote of professors

AMD Ventures doesn't have to keep the shell games going for long -- it's just getting the ground work laid and confidence built in startups like TensorWave.

REPLY ▶

VIEW ALL 3 COMMENTS ▶

Tom's Hardware is part of Future US Inc, an international media group and leading digital publisher. **Visit our corporate site**.

Terms and conditions | Contact Future's experts | Privacy policy
Cookies policy | Accessibility Statement | Advertise with us
About us | Coupons | Careers
Do not sell or share my personal information

© Future US, Inc. Full 7th Floor, 130 West 42nd Street, New York, NY 10036.