Google Cloud AMD streaem

How Streaem uses AMD CPU-powered T2D VMs for better latency, cost-benefit, and scaling

When Streaem, a Netherlands-based retail media platform, set out to drive long-term growth for brands and retailers, the company realized that its mission hinged on the ability to run its platform on technology that offered the lowest controlled latency it could achieve.

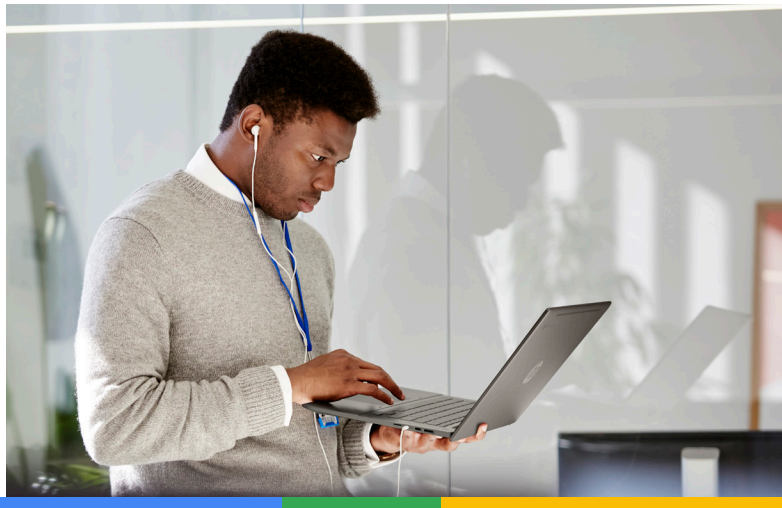
Streaem operates by enabling brands to run full-funnel campaigns on retailers' channels. What does that mean in practice? Consider the process of shopping for clothing online as an example. Once a user types in the word "jumper" into an online retailer's search bar, Streaem populates search results with ads from other clothing manufacturers at the top of the retailer's search results. It's very similar to how Google search ads currently work. Streaem's self-service platform facilitates this process by working with brands to set up campaigns and place bids on a retailer's ad space within that retailer's search results. Streaem's system then utilizes an auction-based process to display a brand's ad in a pre-defined slot triggered by, for instance, a user's search.

In order to quickly and reliably deliver these ads, Streaem knew it needed a fast, reliable virtual machine (VM) to run its workloads. This is why the company selected Google Cloud T2D VMs, powered by AMD EPYC CPUs to power a users search.

Google Cloud T2D VMs powered by 3rd Gen AMD EPYC processors deliver better cost-performance

Google Cloud's Tau VM family has expanded Compute Engine's capabilities in a variety of capacities, most notably for scale-out workloads. The first iteration of this family is the Google Cloud T2D powered by AMD EPYC processors.

When put to the test by Phoronix, a firm that provides Linux hardware review and benchmarking, its founder, Michael Larabel, said, "Across a wide range of tests carried out, the Google Cloud T2D VMs consistently showed great value and performance-per-dollar."[1] On average, across a variety of tests (including image processing, database, video codec, compile time, compression, and cryptography), they achieved 52% higher performance for 8 vCPU VMs and 47% higher performance for 32 vCPU VMs — all while negating the need to migrate away from preexisting architecture (such as the x86).



[1] Tau T2D VMs now in GA: Independent testing validates market-leading price-performance

Google Cloud AMD streaem

How Streaem uses T2D VMs to power reliable, low controlled latency ad placement

What factors influenced Streaem's decision to choose the T2D VMs as its solution, and how did this choice effectively address the company's business challenges? In the selection process, Streaem focused on three main criteria: controlled latency, cost efficiency, and resilient scaling.

1. Controlled latency: Streaem gets hundreds of millions of requests each day. And when it comes to latency, each percentage point of lost traffic represents millions of unanswered requests.

But average latency isn't what Streaem CTO Garry Turkington is most interested in. The company needed *controlled* latency — in other words, reliable and consistent performance, even at the 99.9th percentile mark.

In their latency tests on the T2D VMs, Streaem conducted one hour runs of single instances on collaborating services. This consisted of running an increasing number of sessions, each with just under 5 requests, yielding a request rate of roughly "x" number of sessions per 5 requests.

In this test, Turkington hoped the T2D VMs would help Streaem meet their SLA: 99% of requests serviced in under 40 milliseconds. The T2D VMs delivered. Streaem's goal was to achieve their SLA at a workload of 80 sessions. The VMs exceeded that goal, maintaining a latency of under 40 milliseconds until reaching 180 sessions. This was 900 requests processed per second, per service.

"These are very high numbers, and we're still well above 99%." said Turkington, "Through testing, T2D proved to be the best, both in terms of the control of that latency, but also, it was much more predictable in terms of responsiveness."

Streaem T2D benchmarking results

Sessions/sec	40	60	**80**	100	120	140
Requests/sec	191	286	**381**	477	572	668
Mean response (ms)	8	7	**7**	7	8	8
Minimum response(ms)	5	5	**5**	4	4	4
95% response (ms)	9	8	**8**	8	14	12
99% response (ms)	10	9	**11**	10	19	17
99.9% response (ms)	12	11	**12**	13	32	28
99.99% response (ms)	23	32	**33**	31	73	179
Maximum response (ms)	303	1030	**1015**	394	1041	1033
Standard deviation (ms)	1	2	**1**	1	3	3
Total number of requests	756900	1134900	**1512900**	1890900	2268900	2646900
Number of requests <= 40ms	756865	1134801	**1512867**	1890753	2267969	2645508
Number of requests > 40ms <= 120ms	28	44	**20**	112	789	918
Number of requests > 120ms	7	55	**13**	35	142	474

Data provided by Streaem

Google Cloud AMD streaem

Turkington further noted that one of the main benefits of controlled latency is that it helps to develop strong relationships with the online retail sites that are hosting these ads. With Streaem having more control over latency, the company can predictably guarantee consistent fast page-load times for the online retailer hosting the ads. This consistency creates trust with the online retailers, which are focused on creating a positive experience for consumers.

"Every millisecond we burn could add to page load time. So, we're focused on making very performant, consistent low-latency decisions," said Turkington.

2. Cost efficiency: When identifying the best VMs for its technology, Streaem did take cost into consideration. "We are a startup; we do not have infinite money," noted Turkington. After confirming that Google Cloud T2D VMs met performance requirements, the Streaem team also determined that it was priced competitively, especially when compared with other VM options.

3. Resilient scaling: Streaem is, at its core, multiple services all being deployed in order for the technology to operate. Streaem uses Google Kubernetes Engine (GKE) to deploy these services in a resilient fashion across multiple zones in Google Cloud regions. By allowing GKE to run Streaem's Kubernetes clusters, the company feels confident that its infrastructure is capable of scaling as needed, which allows the team to focus on adding value for its customers in the application.

> Every millisecond we burn could add to page load time. So, we're focused on making very performant, consistent low-latency decisions,
>
> – *Garry Turkington, Streaem CTO*

What's next for Streaem?

Given Streaem's growth plans in the next year, the company could see throughput on the platform increase substantially. "We're now live in the Netherlands, but we are looking to expand into the broader EMEA region and beyond," said Turkington. And with the ability to scale traffic, both on the existing nodes as well as scaling out to additional nodes, Streaem feels certain the Google Cloud T2D VMs powered by AMD 3rd Gen EPYC processors have the controlled latency and cost-performance to provide a quality experience for online retailers and their customers.

To learn more about Google Cloud T2D VMs, visit the machine families resource and comparison guide.