



Amazon GameLift Streams pricing

Overview

With Amazon GameLift Streams you incur two types of charges:

Stream capacity: Stream capacity represents the maximum number of concurrent streams that can be active at a time. Stream capacity charges cover the compute resources that are available to run stream sessions. Stream capacity charges are on a per second basis. You are charged for stream capacity per second, regardless of whether that allocated capacity is actively streaming or idle. Billing for Amazon GameLift Streams begins when stream capacity is allocated in a stream group (compute resources that are allocated to stream your game) and continues as long as that stream group's allocated capacity is maintained. Billing ends when you reduce the allocated capacity in your stream group to zero, or delete the stream group entirely.

Content storage: These charges cover the cost to store your game or application and other data. Storage prices are based on a per-GB, per-month basis.

Amazon GameLift Streams is available in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Tokyo), Europe (Frankfurt), and Europe (Ireland).

The AWS Free Tier is not available for Amazon GameLift Streams.

Pricing Details

Stream capacity cost overview

You are billed for "ready to use" compute resources that the service deploys on your behalf, which is called allocated capacity. Allocated capacity includes all capacity that's available to you, regardless of whether games are being actively streamed to players.

You request stream capacity at the stream-group level. In a stream group, you specify the stream class, which is the type of compute resource you want to use (e.g. stream class with NVIDIA's T4 GPU, 8 vCPUs and Microsoft Windows runtime). Stream capacity is how many instances of that compute resource you want to provision. For each stream group, you'll be billed per allocated capacity count per second.

Your billable usage starts once your first stream group has active capacity. The default settings for new stream groups include one unit of active capacity. There are three key capacity settings available:

- Minimum (always-on) capacity:** A base level of streaming capacity that is always allocated at all times and ready for immediate use. You are billed for this capacity continuously, whether used or idle.
- Maximum capacity:** The total capacity ceiling for a stream group. This includes minimum capacity plus any additional capacity that can be automatically allocated based on demand. New capacity may take a few minutes to become available.
- Target idle (pre-warmed) capacity:** A buffer of pre-allocated capacity maintained beyond your current usage to handle incoming requests quickly. You are billed for this intentionally idle capacity while it's being held.

GameLift Streams first tries to fulfill new session requests using idle capacity which is already allocated to you. If this causes the amount of idle capacity to drop below your target idle capacity, then new capacity is allocated asynchronously. If no idle capacity is available, the request is paused while new capacity is allocated on demand, up to the maximum capacity for the stream group. If the maximum is reached and there is still no idle capacity available, the session request will wait for an existing session to terminate and free capacity.

When sessions terminate, the corresponding capacity is marked as idle. If there is more idle capacity than the target idle value, the excess capacity will be deallocated and returned to the service after a brief delay. The service will not deallocate idle capacity if that would drop your capacity level below the configured minimum (which could be zero).

Your billing rate depends on AWS Region where your stream groups are deployed and the Stream class selected for each stream group. For additional details, see the billing FAQs.

Minimum (always-on) capacity

Minimum capacity represents your base level of streaming resources that remain allocated to you at all times. This capacity is ideal for maintaining consistent, immediate availability for your core player base. You are billed for this capacity continuously since it's available exclusively for your use. You can adjust this setting to balance availability against ongoing costs.

The service maintains this minimum level at all times, ensuring your streams can start within seconds for this portion of your capacity. This makes it ideal for workloads requiring predictable, instant availability.

Maximum capacity

Maximum capacity represents the upper limit of resources that can be allocated to your stream group, including minimum (always-on) capacity, target idle capacity, and any additional on-demand allocations. When demand increases beyond your currently allocated capacity, the service can provision new resources up to this maximum limit, typically taking a few minutes to become available. This capacity is released when no longer needed, providing cost efficiency. The maximum capacity helps you control costs and manage quotas while ensuring sufficient scalability for peak demands.

Region: Europe (London)					
Stream Class	GPU	vCPU	VRAM	RAM	Price per hour
gen6n_pro_win2022	NVIDIA L4 Tensor Core GPU	16	24	64	\$3.36
gen6n_ultra_win2022	NVIDIA L4 Tensor Core GPU	8	24	32	\$2.05
gen6n_high	NVIDIA L4 Tensor Core GPU	4	12	16	\$0.81
gen6n_pro	NVIDIA L4 Tensor Core GPU	16	24	64	\$2.33
gen6n_ultra	NVIDIA L4 Tensor Core GPU	8	24	32	\$1.72
gen5n_win2022	NVIDIA A10G Tensor GPU	8	24	32	\$2.16
gen5n_high	NVIDIA A10G Tensor GPU	4	12	16	\$0.94
gen5n_ultra	NVIDIA A10G Tensor GPU	8	24	32	\$1.77
gen4n_win2022	NVIDIA T4 Tensor GPU	8	16	32	\$1.70
gen4n_high	NVIDIA T4 Tensor GPU	4	8	16	\$0.63
gen4n_ultra	NVIDIA T4 Tensor GPU	8	16	32	\$1.21
gen6n_medium	NVIDIA L4 Tensor Core GPU	2	6	8	\$0.39
gen6n_small	NVIDIA L4 Tensor Core GPU	1	2	5	\$0.17

Region: US East (Ohio)	
Storage Class	Price
GameDataStorageHours	\$0.03 per GB-Month

Pricing examples

Playtesting with minimum (always-on) capacity

20 concurrent players testing a new game for 8 hours a day, no wait time

This example illustrates the scenario of testing a new game before launch. This requires capacity for 20 concurrent players on a gen6n_ultra_win2022 stream class in Oregon (us-west-2). The customer chooses a Microsoft Windows based stream class to ensure the playing runtime environment matches the development runtime. Since the scale and usage is predictable in this scenario, minimum (always-on) streaming capacity can be used to meet the concurrent stream needs without any start time delay. The game only requires streaming capacity during business hours for 8 hours a day, so the customer will de-provision the minimum (always-on) streaming capacity when it is not being utilized.

- The cost per player streaming capacity-hour for gen6n_ultra_win2022 stream class in Oregon (us-west-2) is \$1.82.
- Minimum (always-on) Capacity x Provisioned Hours x Stream group hourly rate = Monthly billable stream capacity.
- 20 minimum (always-on) capacity x (8 hours per day x 30 days) x \$1.82= \$8,736.00 monthly billable.

Party Game with minimum (always-on) capacity, maximum capacity and target-idle capacity

20 to 50 concurrent players play the game for 4 Hours per Day. Optimized capacity use with some delays in start time

A modern action role-playing game launches, expecting a minimum of 20, and a max of 50 concurrent players during its first week. To optimize costs, the game developer selects Proton as the runtime environment, running on gen4n_high (NVIDIA) stream class in Oregon (us-west-2). Players engage with the game for an average of 4 hours daily. The developer sets minimum (always-on) streaming capacity to 20, and maximum capacity to 50. The first 20 concurrent player are able to join instantly as minimum (always-on) capacity offer no wait time. For any additional player beyond the 20 (minimum always on capacity) limit, the service will provision a new stream up to a maximum of 50, typically taking a few minutes to become available. This additional provisioned capacity (beyond the minimum capacity) is released back to the service automatically when the session ends, providing cost efficiency. Setting a maximum capacity limits the number session that can run concurrently, help control costs and manage quotas while ensuring scalability for peak demands.

The cost is calculated as Stream Capacity x Provisioned Hours x Stream group hourly rate = Monthly billable stream capacity. Assuming the max concurrency is 35 players for 2 hours in the first week, then you would be charged for the 20 minimum (always on) capacity for 4 hours * 7 days and 15 additional for 2 hours each day for 7 days. This 15 stream capacity get automatically de-provisioned when the session ends.

So the cost for the first week will be:

- The cost per streaming capacity-hour for gen4n_high in Oregon (us-west-2) is \$0.4982.
- Total cost is calculated using: Minimum (always-on) Capacity x Provisioned Hours x Stream group hourly rate = Monthly billable stream capacity.
- Total cost for that day will be [20 always-on * 4 hrs + 15 on-demand * 2 hours] x 7 days x \$0.4982 = \$383.614.

Adding 10 pre-warm sessions (target-idle capacity) on top of existing capacity

After the initial release of a game, the player engagement is at under 10 additional players per minute. To balance between session start up time and cost, the game publisher configures the target-idle capacity to 10 to pre-warm sessions and minimize player wait time, while efficiently scaling down resources during low-usage periods.

- Maximum cost of this month for keeping extra (10 pre-warmed streaming x 4 hours per day x 7 days x \$0.4982) = \$139.496 billable stream capacity.

New Game Launch with Minimum (always-on) scaling

2000 players on a Cloud gaming service with party games

A new ad-supported cloud gaming service will launch a catalog of party games on smart TVs. 2000 concurrent players engage with it during its first month. To optimize costs, the game developer selects Proton as the runtime environment, running on gen4n_high (NVIDIA-based) stream class in Frankfurt (eu-central-1). Players engage with the games for an average of 30 minutes daily. The customer uses minimum (always-on) capacity and leverages the 30-second start time to prepare the player for the game by first playing a 30 second advertisement, then displaying game controls and stepping through gameplay instructions, while automatically provisioning resources for the stream.

- The cost per stream group capacity-hour for gen4n_high (NVIDIA) stream class in Frankfurt (eu-central-1) is \$0.6571.
- Stream Group Capacity x Provisioned Hours x Stream group hourly rate = Monthly billable stream capacity.

Month of launch:

- 2000 players streaming x (0.5 hours per day x 30 days) x \$0.6571 = \$19,713 billable stream capacity.

Next steps

Console

Try Amazon GameLift Streams in the Console

→

Contact us

Connect with an expert to get started today

→

Learn

What Is AWS?
What Is Cloud Computing?
What Is Agentic AI?
Cloud Computing Concepts Hub
AWS Cloud Security
What's New
Blogs
Press Releases

Resources

Getting Started
Training
AWS Trust Center
AWS Solutions Library
Architecture Center
Product and Technical FAQs
Analyst Reports
AWS Partners

Developers

Builder Center
SDKs & Tools
.NET on AWS
Python on AWS
Java on AWS
PHP on AWS
JavaScript on AWS

Help

Contact Us
File a Support Ticket
AWS rePost
Knowledge Center
AWS Support Overview
Get Expert Help
AWS Accessibility
Legal

Back to top