Amazon SageMaker
Overview
Features
Pricing
FAQs
Developer Resources
Customers

Products / Machine Learning / Amazon SageMaker / …

Amazon SageMaker ML Instance Types

Amazon SageMaker provides a selection of instance types optimized to fit different machine learning (ML) use cases. Instance types comprise varying combinations of CPU, GPU, memory, and networking capacity and give you the flexibility to choose the appropriate mix of resources for building, training, and deploying your ML models. Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

NOTE: Please refer to the actual pricing page for details on the support for each instance type in a particular AWS region and for a particular SageMaker function such as building, processing, training, and deployment of ML models. Not all instances are supported in all AWS regions and/or for all SageMaker functions.

Instance type	Mem	GPU Mem	Network

Amazon SageMaker
Overview
Features ▾
Pricing ▾
FAQs
Developer Resources
Customers

ml.t2.xlarge	4	16	Moderate
ml.t2.2xlarge	8	32	Moderate
ml.t3.medium	2	4	Low to Moderate
ml.t3.large	2	8	Low to Moderate
ml.t3.xlarge	4	16	Low to Moderate
ml.t3.2xlarge	8	32	Low to Moderate
ml.m5.large	2	8	High
ml.m5.xlarge	4	16	High
ml.m5.2xlarge	8	32	High
ml.m5.4xlarge	16	64	High
ml.m5.12xlarge	48	192	10 Gigabit
ml.m5.24xlarge	96	384	25 Gigabit
ml.m4.xlarge	4	16	High

ml.m5d.2xlarge	8	-	32	-	Up to 10 Gbps
ml.m5d.4xlarge	16	-	64	-	Up to 10 Gbps
ml.m5d.8xlarge	32	-	128	-	10 Gbps
ml.m5d.12xlarge	48	-	192	-	10 Gbps
ml.m5d.24xlarge	96	-	384	-	25 Gbps
Memory Optimized - Current Generation					
ml.r5.large	2	-	16	-	Up to 10 Gbps
ml.r5.xlarge	4	-	32	-	Up to 10 Gbps
ml.r5.2xlarge	8	-	64	-	Up to 10 Gbps
ml.r5.4xlarge	16	-	128	-	Up to 10 Gbps
ml.r5.12xlarge	48	-	384	-	10 Gbps
ml.r5.24xlarge	96	-	768	-	25 Gbps

ml.r5d.16xlarge	64	-	512	-	20 Gigabit
ml.r5d.24xlarge	96	-	768	-	25 Gigabit

Compute Optimized – Current Generation

ml.c5.large	2		4		Up to 10 Gbps
ml.c5.xlarge	4	-	8	-	Up to 10 Gbps
ml.c5.2xlarge	8	-	16	-	Up to 10 Gbps
ml.c5.4xlarge	16	-	32	-	Up to 10 Gbps
ml.c5.9xlarge	36	-	72	-	10 Gigabit
ml.c5.18xlarge	72	-	144	-	25 Gigabit
ml.c5d.xlarge	4		8		Up to 10 Gbps
ml.c5d.2xlarge	8		16		Up to 10 Gbps
ml.c5d.4xlarge	16		32		Up to 10 Gbps

ml.c4.8xlarge	36	-	60	-	10 Gigabit
		-		-	
Accelerated Computing – Current Generation					
ml.p3.2xlarge	8	1xV100	61	16	Up to 10 Gbps
ml.p3.8xlarge	32	4xV100	244	64	10 Gigabit
ml.p3.16xlarge	64	8xV100	488	128	25 Gigabit
ml.p3dn.24xlarge	96	8xV100	768	256	100 Gigabit
ml.p2.xlarge	4	1xK80	61	12	High
ml.p2.8xlarge	32	8xK80	488	96	10 Gigabit
ml.p2.16xlarge	64	16xK80	732	192	25 Gigabit
ml.g4dn.xlarge	4	1xT4	16	16	Up to 25 Gbps
ml.g4dn.2xlarge	8	1xT4	32	16	Up to 25 Gbps
ml.g4dn.4xlarge	16	1xT4	64	16	Up to 25 Gbps

ml.inf1.24xlarge	96	192	100 Gbps

Inference Acceleration

Accelerator	Throughput in FP-32 TFLOPS*	Throughput in FP-16 TFLOPS**	Memory
eia2.medium	1	8	2 GB
eia2.large	2	16	4 GB
eia2.xlarge	4	32	8 GB
eia1.medium	1	8	1 GB
eia1.large	2	16	2 GB
eia1.xlarge	4	32	4 GB

* FP-32 TFLOPS = trillion 32-bit floating operations per second

** FP-16 TFLOPS = trillion 16-bit floating operations per second

Amazon SageMaker

Overview

Features

Pricing

FAQs

Developer Resources

Customers

Have more questions?

Contact us

Sign In to the Console

Learn About AWS

What Is AWS?

What Is Cloud Computing?

What Is DevOps?

What Is a Container?

What Is a Data Lake?

AWS Cloud Security

What's New

Resources for AWS

Getting Started

Training and Certification

AWS Solutions Portfolio

Architecture Center

Product and Technical FAQs

Analyst Reports

AWS Partner Network

Developers on AWS

Developer Center

SDKs & Tools

.NET on AWS

Python on AWS

Java on AWS

PHP on AWS

Javascript on AWS

Help

Contact Us

AWS Careers

File a Support Ticket

Knowledge Center

AWS Support Overview

Legal

UK Modern Slavery Statement

Amazon SageMaker

Overview

Features

Pricing

FAQs

Developer Resources

Customers

Language

عربي |

Bahasa Indonesia |

Deutsch |

English |

Español |

Français |

Italiano |

Português |

Tiếng Việt |

Türkçe |

Русский |

ไทย |

日本語 |

한국어 |

中文 (简体) |

中文 (繁體)

Privacy

|

Site Terms