Introduction to AWS Inferentia and Amazon EC2 Inf1 instances

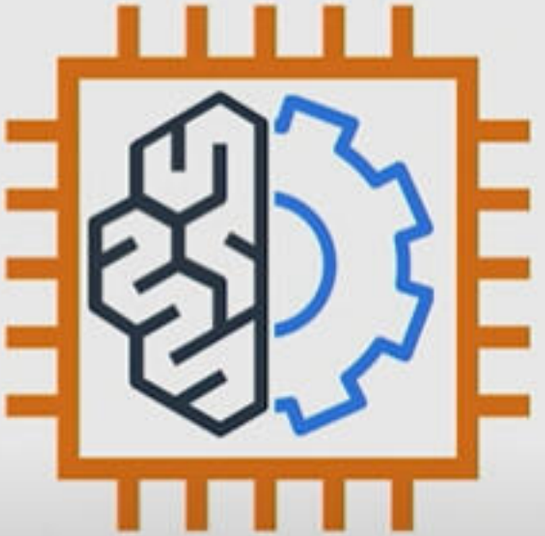
Vibhav Viswanathan

aws training and certification

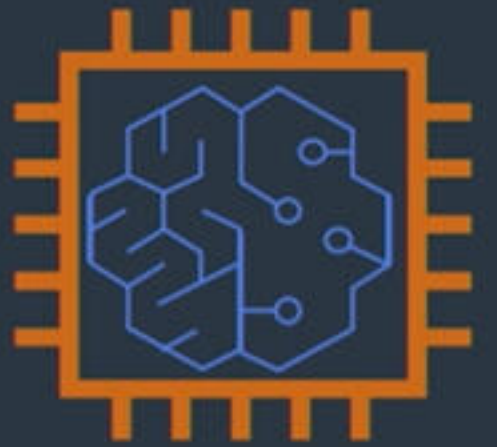
In this video

aws training and certification

- Machine learning use cases and challenges
- AWS Inferentia overview
- Amazon EC2 Inf1 instances introduction
- Benefits
- Pricing and availability



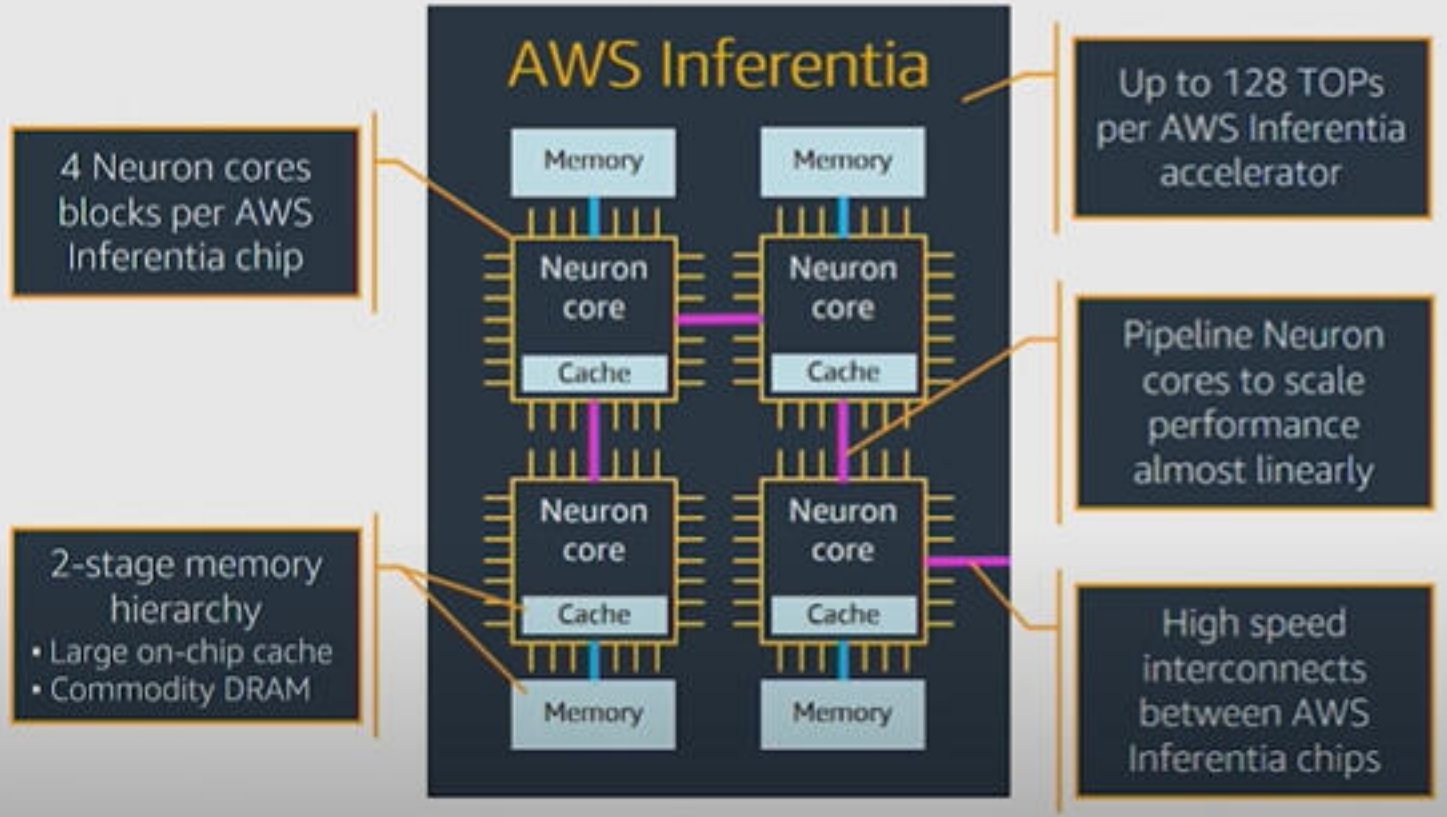
AWS Inferentia



aws training and certification

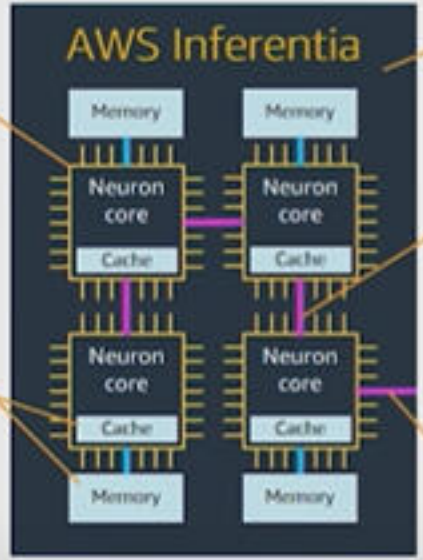
AWS Inferentia accelerators

aws training and certification



AWS Inferentia accelerators

aws training and certification



4 Neuron cores blocks per AWS Inferentia chip

2-stage memory hierarchy
- Large on-chip cache
- Commodity DRAM

Up to 128 TOPs per AWS Inferentia accelerator

Pipeline Neuron cores to scale performance almost linearly

High speed interconnects between AWS Inferentia chips

Data types

- INT8
- BFloat16
- FP16 with mixed precision

ML frameworks

- TensorFlow
- MXNet
- PyTorch

AWS Neuron software

aws training and certification

Build

Neuron Compiler
(NCC)

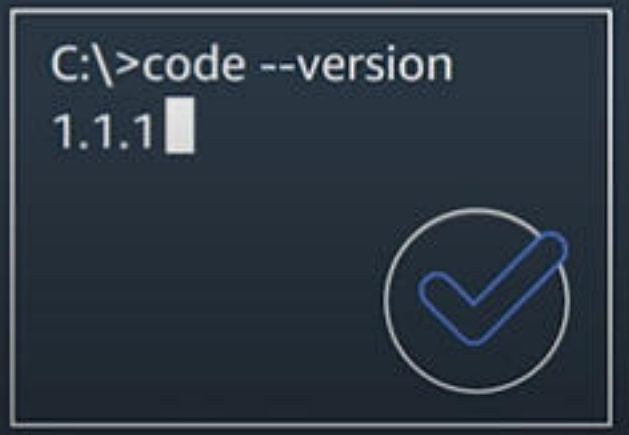
Deploy

Neuron Runtime
(NRT)

Neuron Binary
(NBIN)

Debug/profile

Neuron Debugger/
Profiler



AWS Neuron software

aws training and certification

Build

Neuron Compiler
(NCC)

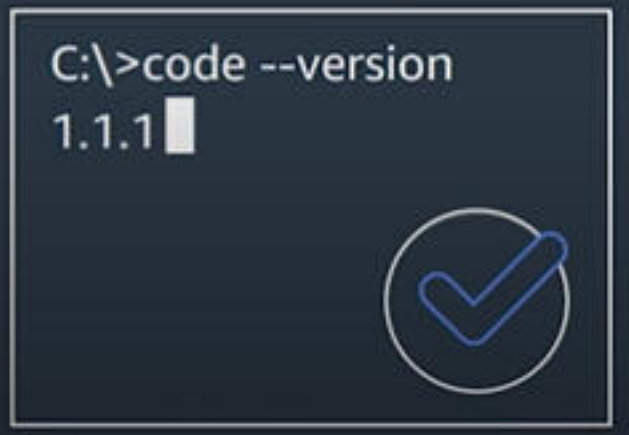
Deploy

Neuron Runtime
(NRT)

Neuron Binary
(NBIN)

Debug/profile

Neuron Debugger/
Profiler



AWS Neuron software

aws training and certification

Build

Deploy

Debug/profile

Neuron Runtime (NRT)

Neuron Debugger/ Profiler

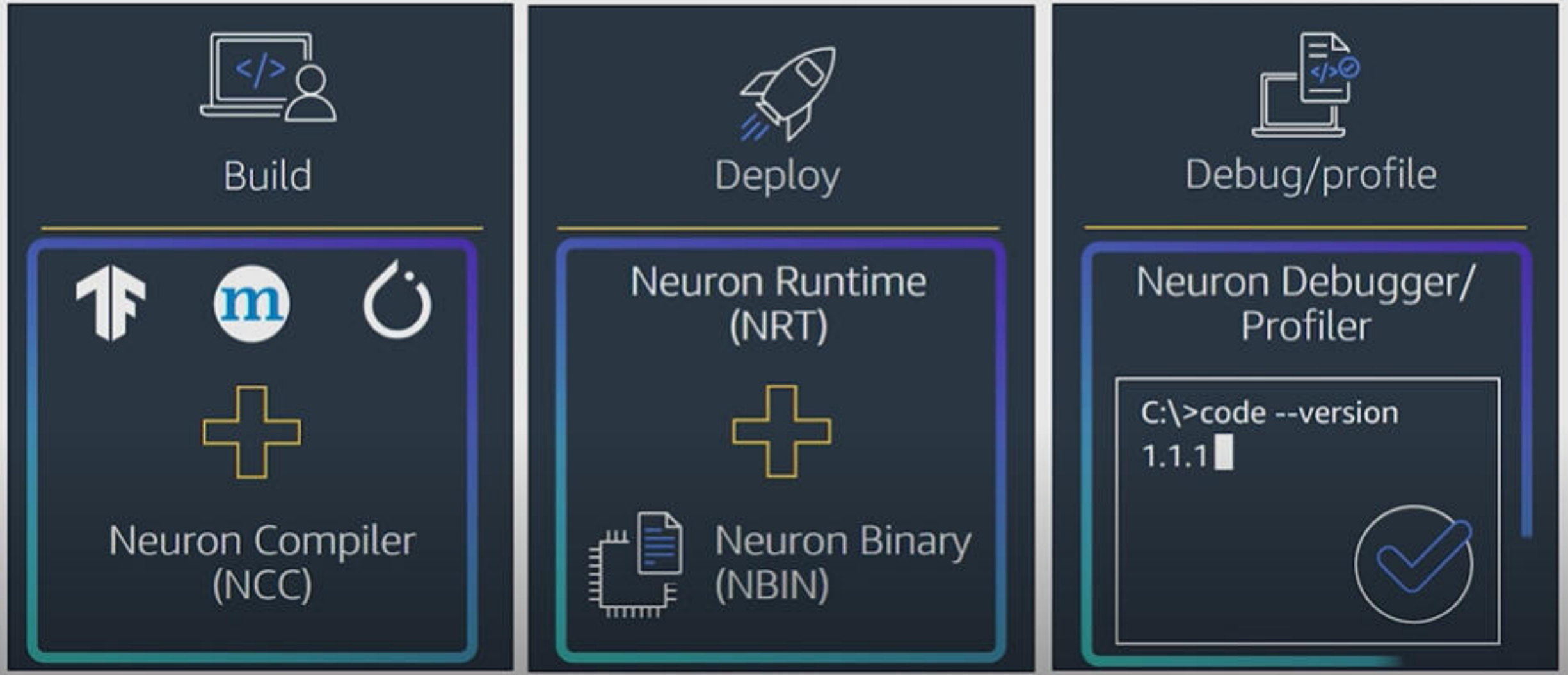
```
C:\>code --version
1.1.1
```

Neuron Compiler (NCC)

Neuron Binary (NBIN)

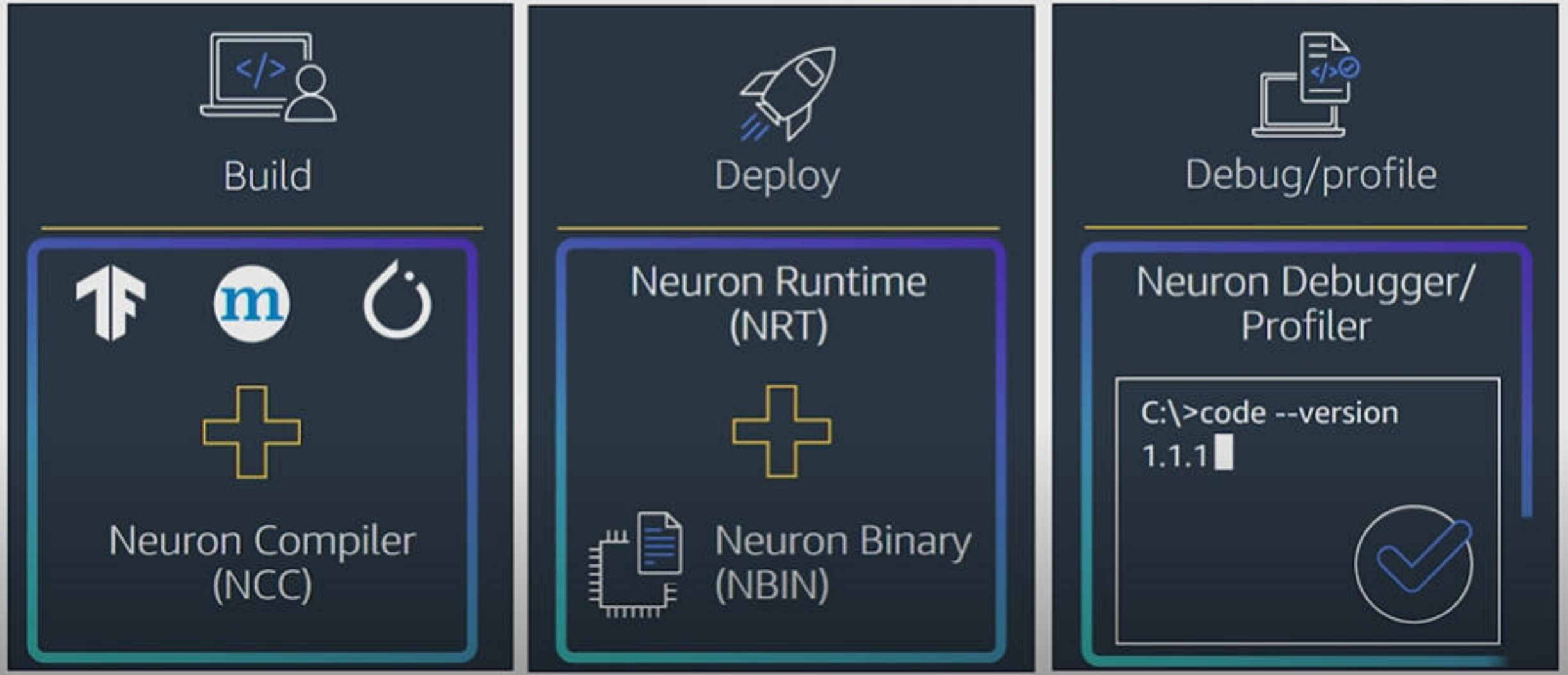
AWS Neuron software

aws training and certification

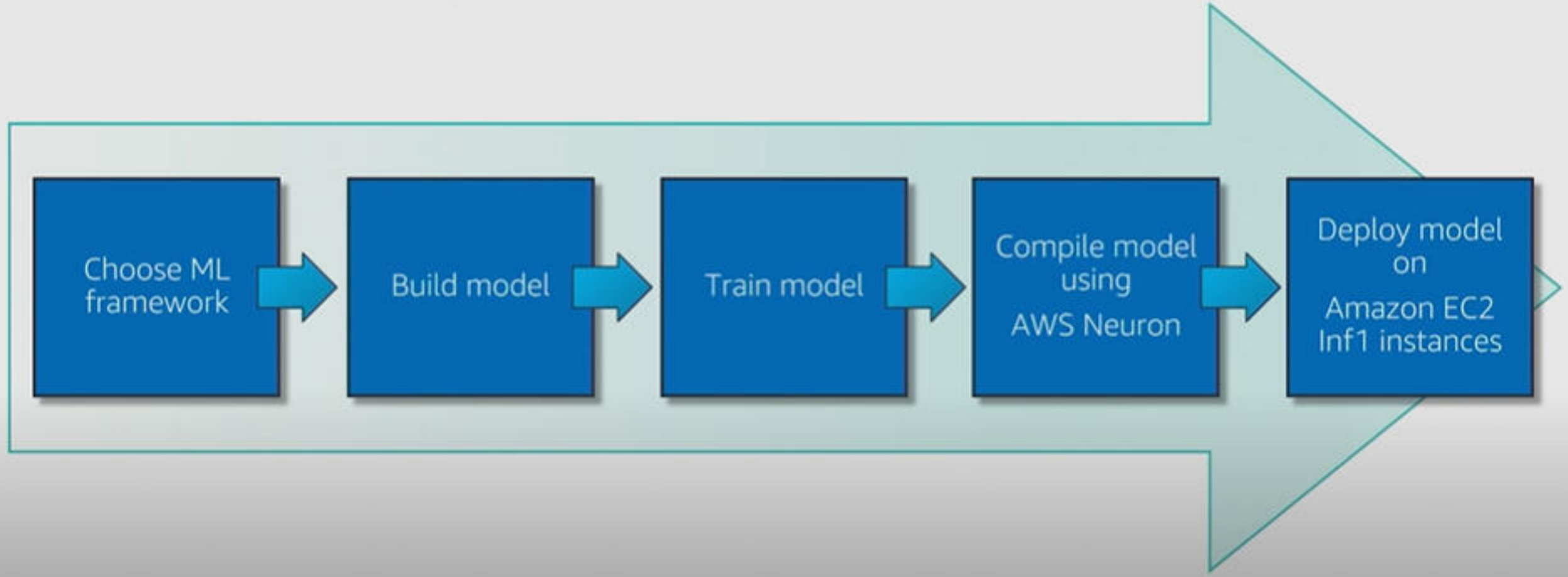


AWS Neuron software

aws training and certification



AWS Neuron developer flow



AWS Neuron highlights

aws training and certification

Smart partitioning

Automatically optimize neural-net compute

Auto FP32 casting

Ingest FP32 trained model, and cast to BFloat16

AWS Neuron Core pipeline

Super low-latency full bandwidth

AWS Neuron Core groups

Concurrently run multiple models

AWS Neuron highlights

aws training and certification

Smart partitioning

Automatically optimize neural-net compute

Auto FP32 casting

Ingest FP32 trained model, and cast to BFloat16

AWS Neuron Core pipeline

Super low-latency full bandwidth

AWS Neuron Core groups

Concurrently run multiple models

AWS Neuron highlights

aws training and certification

Smart partitioning

Automatically optimize neural-net compute

Auto FP32 casting

Ingest FP32 trained model, and cast to BFloat16

AWS Neuron Core pipeline

Super low-latency full bandwidth

AWS Neuron Core groups

Concurrently run multiple models

Amazon EC2 Inf1 instances powered by AWS Inferentia

aws training and certification

Amazon EC2 Inf1 instances use:

- Custom Intel Cascade-Lake processors
- Custom AWS Inferentia ML acceleration chips
- AWS Nitro up to 100-Gbps networking

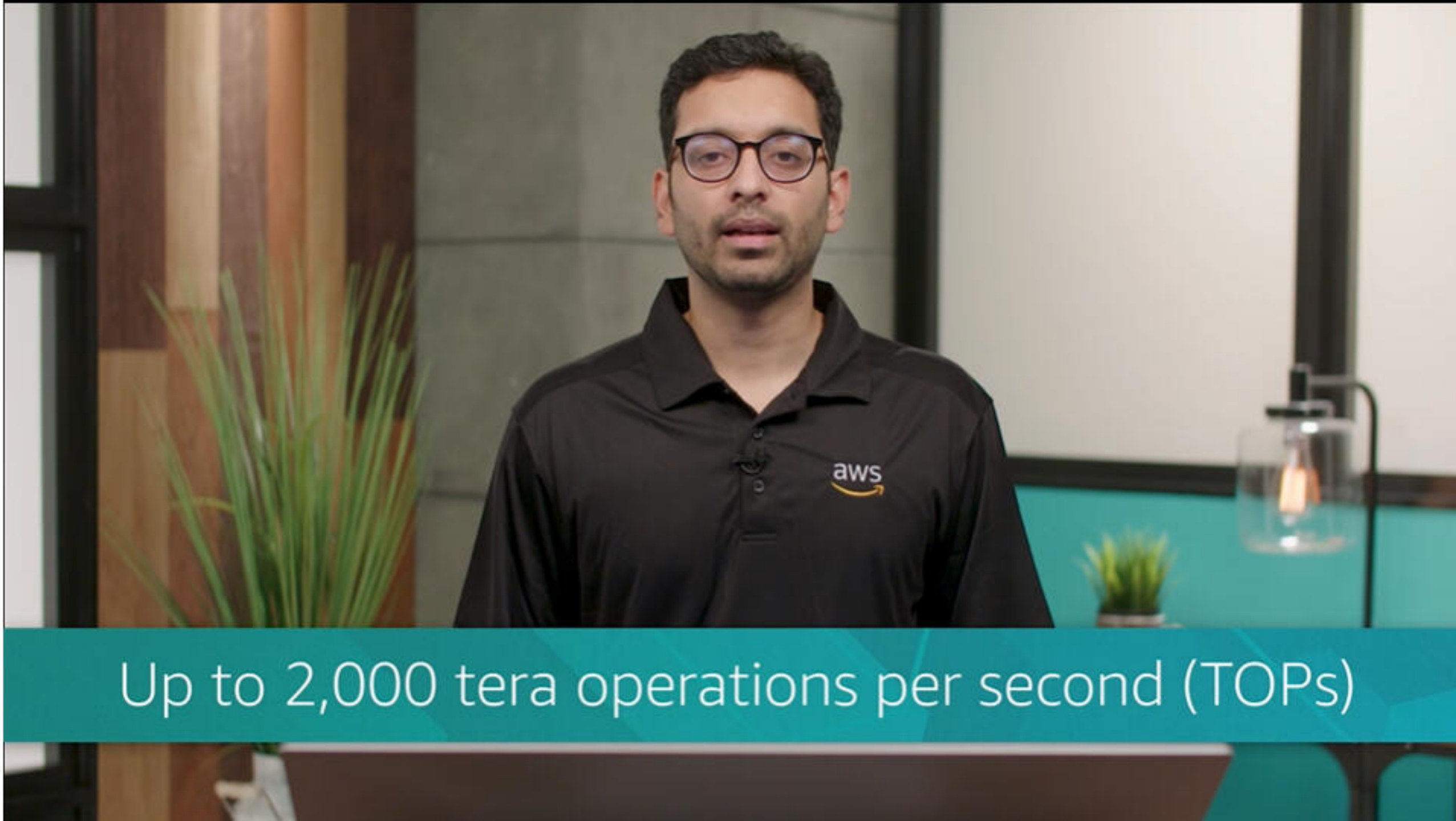
Amazon EC2 Inf1 instance benefits

aws training and certification

Low cost

High performance

Increased agility to meet ML needs


aws
Up to 2,000 tera operations per second (TOPs)

Amazon EC2 Inf1 instance benefits

aws training and certification

Low cost

High performance

Increased agility to meet ML needs

Amazon EC2 Inf1 instance benefits

aws training and certification

Low cost

High performance

Increased agility to meet ML needs

Instance sizes

aws training and certification

Instance size	vCPUs	Memory (GiB)	AWS Inferentia chips	Chip-to-chip interconnect	Storage	Network bandwidth	Amazon EBS bandwidth
Inf1.xlarge	4	8	1	N/A	EBS only	Up to 25 Gbps	Up to 3.5 Gbps
Inf1.2xlarge	8	16	1	N/A	EBS only	Up to 25 Gbps	Up to 3.5 Gbps
Inf1.6xlarge	24	48	4	Yes	EBS only	25 Gbps	3.5 Gbps
Inf1.24xlarge	96	192	16	Yes	EBS only	100 Gbps	14 Gbps

Instance sizes

aws training and certification

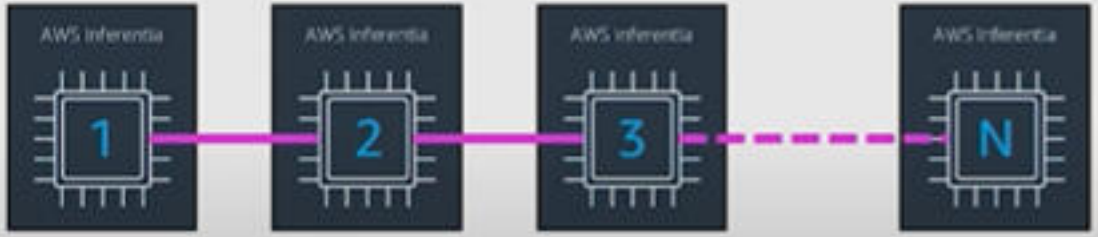
Instance size	vCPUs	Memory (GiB)	AWS Inferentia chips	Chip-to-chip interconnect	Storage	Network bandwidth	Amazon EBS bandwidth
Inf1.xlarge	4	8	1	N/A	EBS only	Up to 25 Gbps	Up to 3.5 Gbps
Inf1.2xlarge	8	16	1	N/A	EBS only	Up to 25 Gbps	Up to 3.5 Gbps
Inf1.6xlarge	24	48	4	Yes	EBS only	25 Gbps	3.5 Gbps
Inf1.24xlarge	96	192	16	Yes	EBS only	100 Gbps	14 Gbps

Instance sizes

aws training and certification

Instance size	vCPUs	Memory (GiB)	AWS Inferentia chips	Chip-to-chip interconnect	Storage	Network bandwidth	Amazon EBS bandwidth
Inf1.xlarge	4	8	1	N/A	EBS only	Up to 25 Gbps	Up to 3.5 Gbps
Inf1.2xlarge	8	16	1	N/A	EBS only	Up to 25 Gbps	Up to 3.5 Gbps
Inf1.6xlarge	24	48	4	Yes	EBS only	25 Gbps	3.5 Gbps
Inf1.24xlarge	96	192	16	Yes	EBS only	100 Gbps	14 Gbps

AWS Inferentia chip-to-chip fast interconnects



AWS integration

aws training and certification

AWS Deep Learning AMIs

AWS Deep Learning Amazon EC2 AMIs

Amazon EKS

Amazon ECS

AWS Deep Learning Containers

Amazon SageMaker

Amazon SageMaker managed ML service

AWS integration

aws training and certification

AWS Deep Learning AMIs

AWS Deep Learning Amazon EC2 AMIs

Amazon EKS

Amazon ECS

AWS Deep Learning Containers

Amazon SageMaker

Amazon SageMaker managed ML service

AWS integration

aws training and certification

AWS Deep Learning AMIs

AWS Deep Learning Amazon EC2 AMIs

Amazon EKS

Amazon ECS

AWS Deep Learning Containers

Amazon SageMaker

Amazon SageMaker managed ML service

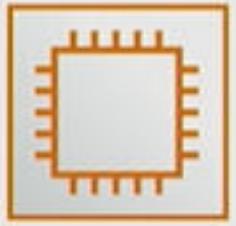
Pricing models and availability

aws training and certification

Pricing

- Supports all Amazon EC2 pricing models

On-Demand instances



Reserved instances

Spot instances

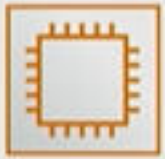
Pricing models and availability

aws training and certification

Pricing

- Supports all Amazon EC2 pricing models

On-Demand instances



Reserved instances

Spot instances

Pricing models and availability

aws training and certification

Pricing

- Supports all Amazon EC2 pricing models

On-Demand instances

Reserved instances

Spot instances

Availability

- Initial launch December 2019
- Deployment to all AWS Regions

aws training and certification

Certificate of Completion

Hem Bahadur Gurung

Has successfully completed

Introduction to AWS Inferentia and Amazon EC2 Inf1 Instances

Director, Training and Certification

15 minutes

Duration

10 September, 2021

Completion Date

aws training and certification

Certificate of Completion

Hem Bahadur Gurung

Has successfully completed

Introduction to AWS Inferentia and Amazon EC2 Inf1 Instances

Director, Training and Certification

15 minutes

Duration

10 September, 2021

Completion Date