

The background features a dark blue gradient with large, overlapping, curved shapes in shades of orange, pink, and purple on the right side, creating a modern, abstract design.

AWS re:Invent

NOV. 27 – DEC. 1, 2023 | LAS VEGAS, NV

ARC323-R

Architectural patterns for low-latency applications on AWS

Diego Soares

Principal Solutions Architect
AWS

Yahav Biran

Principal Solutions Architect
AWS



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- 01 Latency level setting
- 02 Use cases where latency matters
- 03 How compute influences latency
- 04 How network influences latency
- 05 How application patterns influence latency

Latency spectrum

LESS LATENCY SENSITIVE



15 minutes to
~350 milliseconds

~To the sun and back
149 million km

LATENCY SENSITIVE



Screen data to
~300 milliseconds

~2x around the earth
71,347 km

LOW LATENCY



Single digit
milliseconds

NYC to US-EAST-1
367 km

ULTRA LOW LATENCY



Nanosecond

1 foot
20 cm

AWS wherever you need it

Same infrastructure, services, APIs, and tools for a consistent experience

Regions



AWS Global
Infrastructure

Metro areas and telco networks



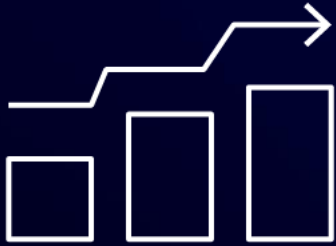
Amazon CloudFront
AWS Local Zones
AWS Wavelength
AWS Telco Network Builder

On premises

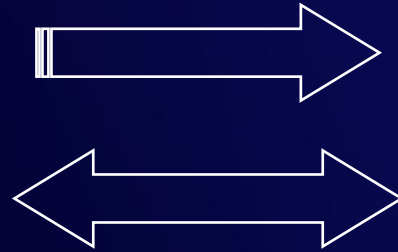


AWS Outposts
Amazon ECS/Amazon EKS
Anywhere
AWS Private 5G
Integrated Private
Wireless on AWS

How to measure latency?



Average
vs.
percentile



One-way
vs.
two-way



Alert-based
vs.
continuously

Use cases where latency matters



Trading applications



Online gaming

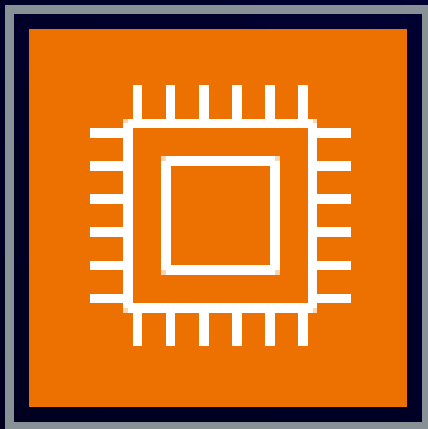


HPC/AI/ML



Live streaming

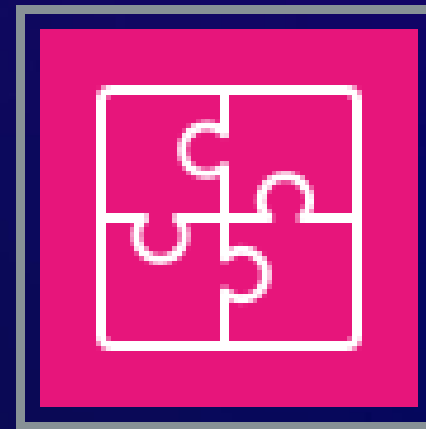
Latency considerations



Compute



Network



**Deployment
patterns**

Questions?



Thank you!

Diego Soares

 /diegogsoares

Yahav Biran

 /ybiran



Please complete the session
survey in the mobile app

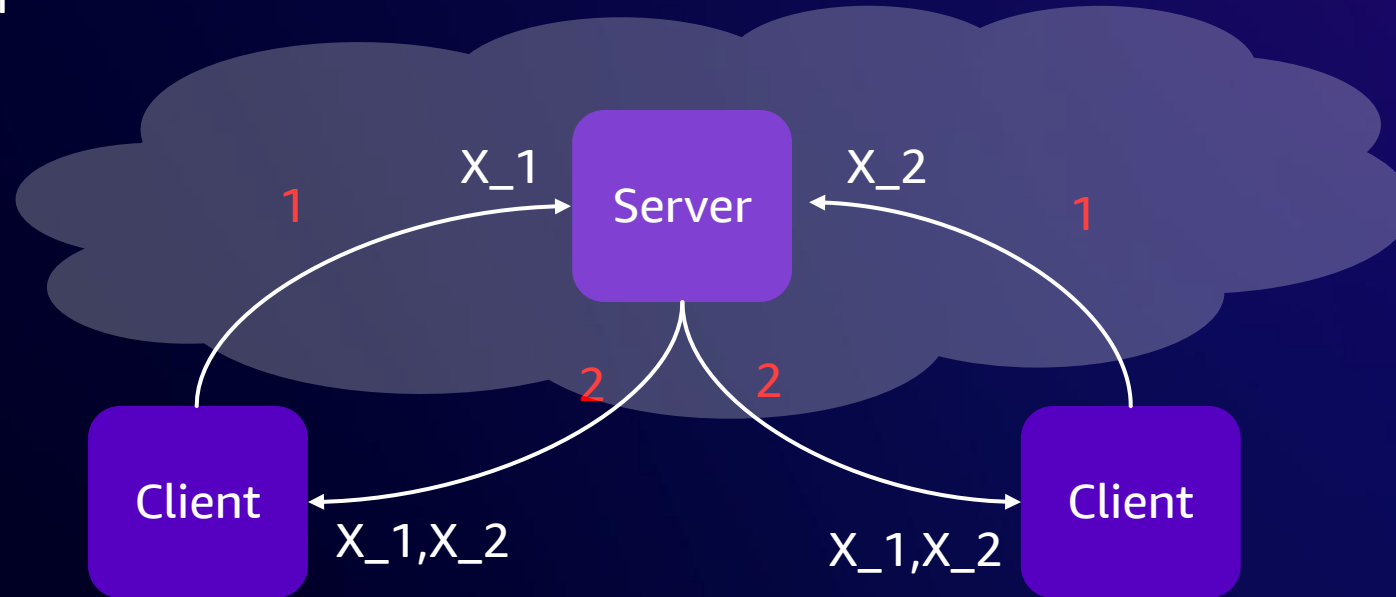
Appendix

Gaming applications



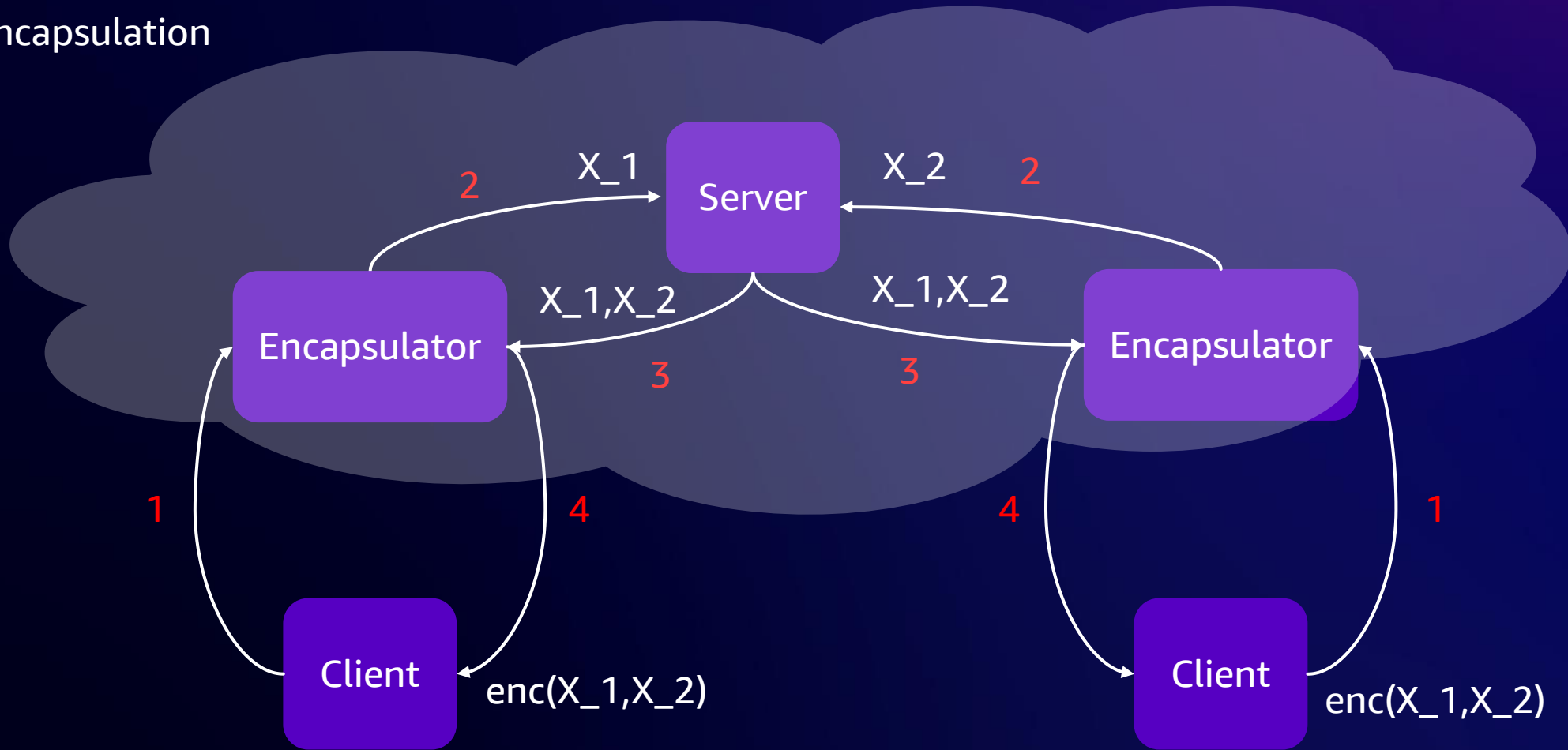
Game server deployment patterns

Network multiplexer



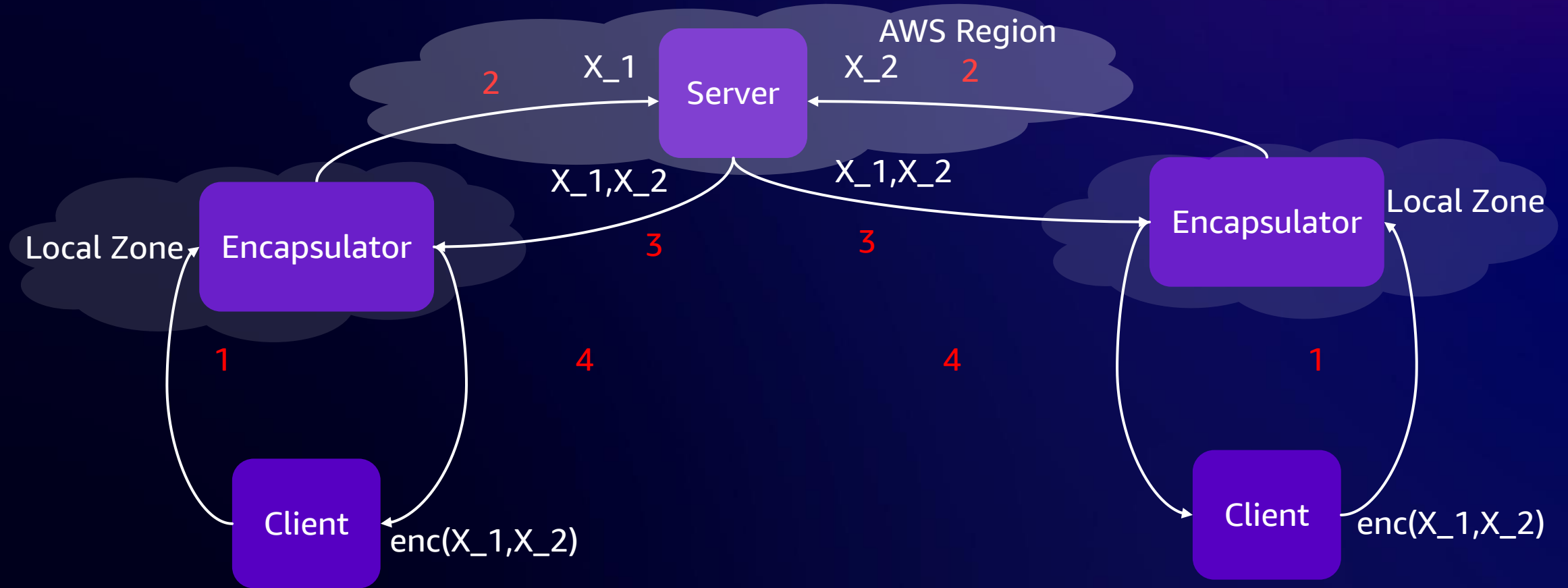
Game server deployment patterns

GPU encapsulation



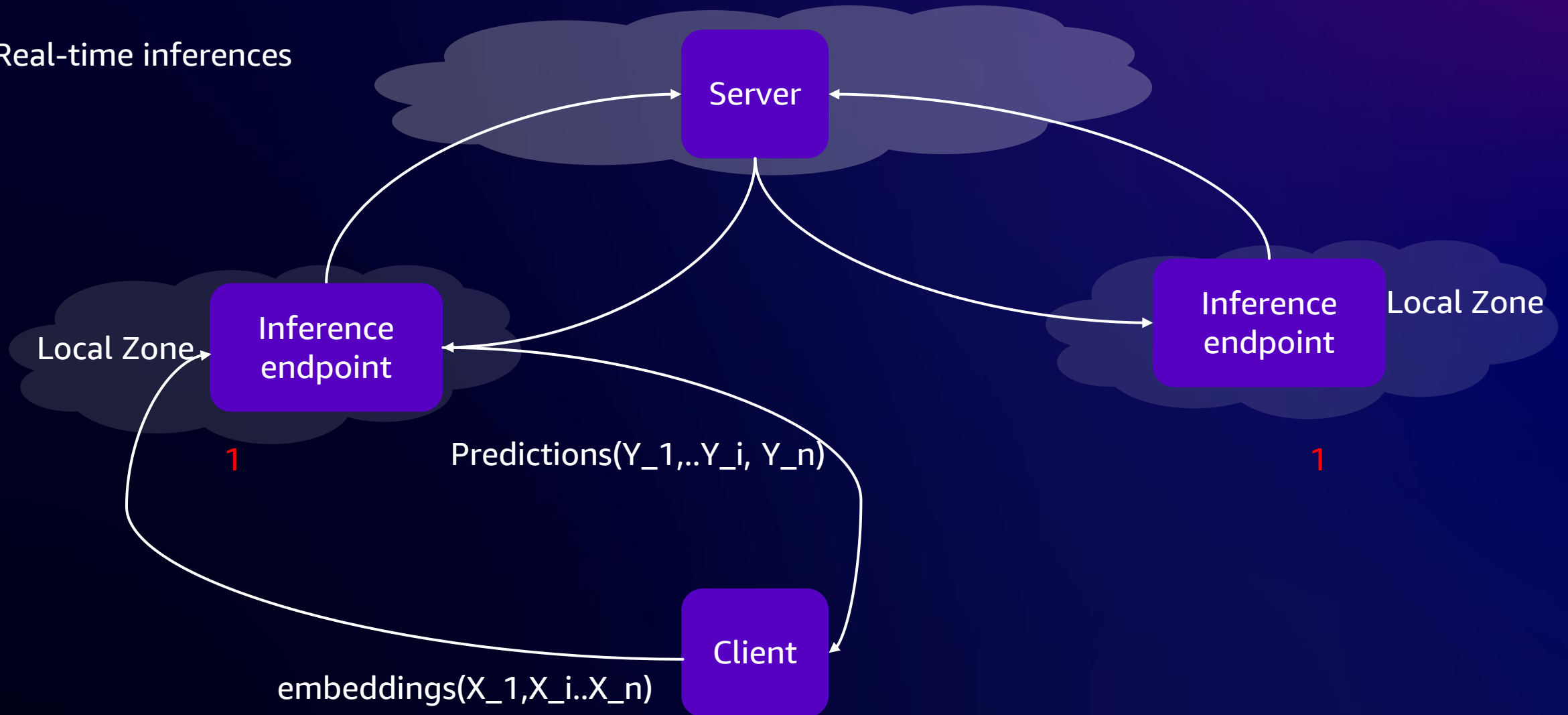
Game server deployment patterns

GPU encapsulation

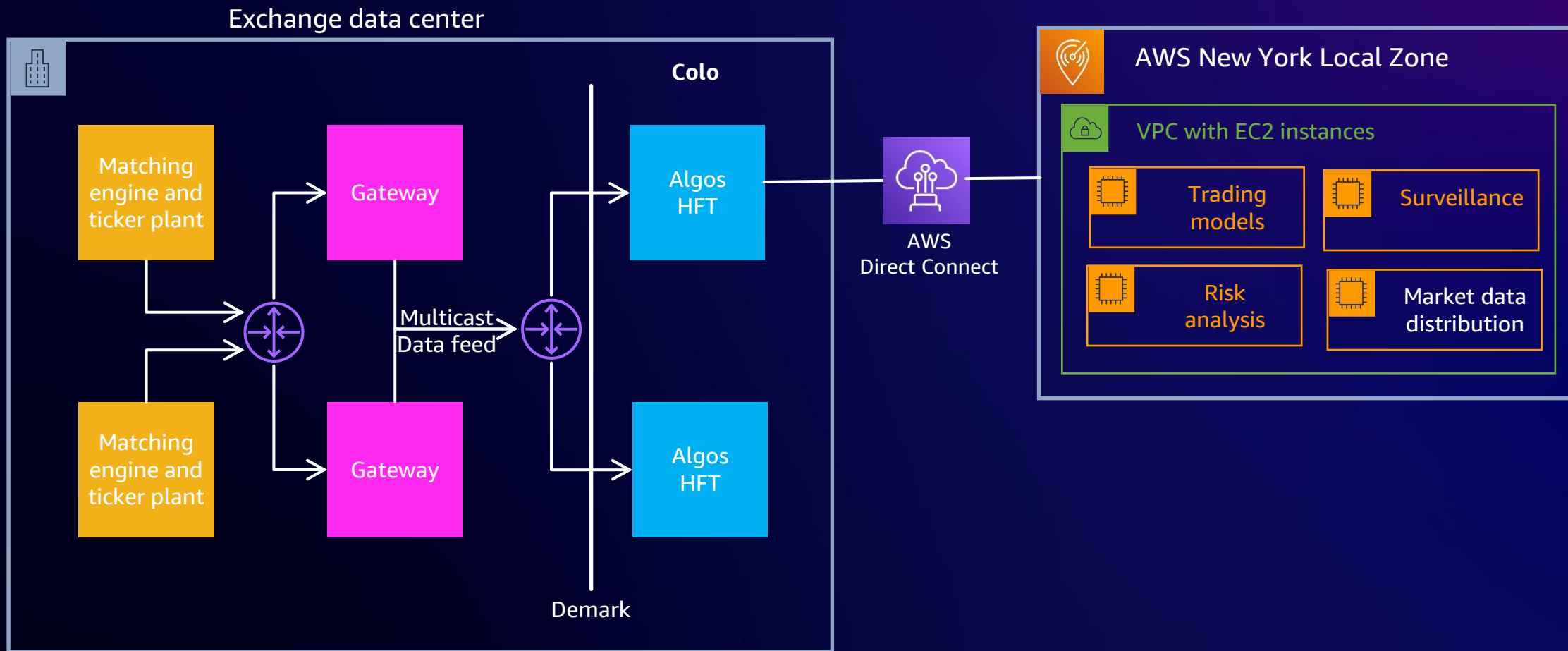


Augmented reality (AR) deployment patterns

Real-time inferences



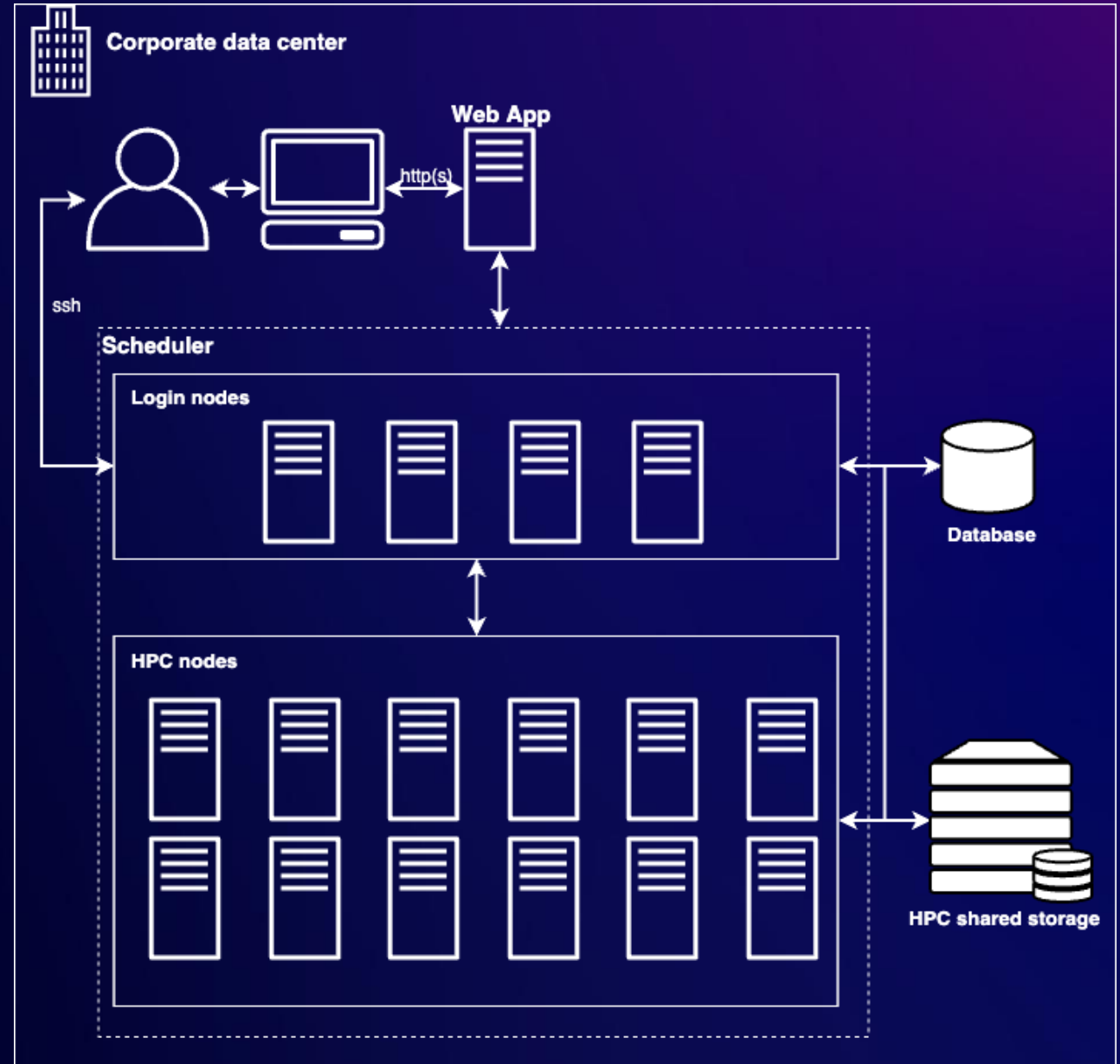
Trading applications



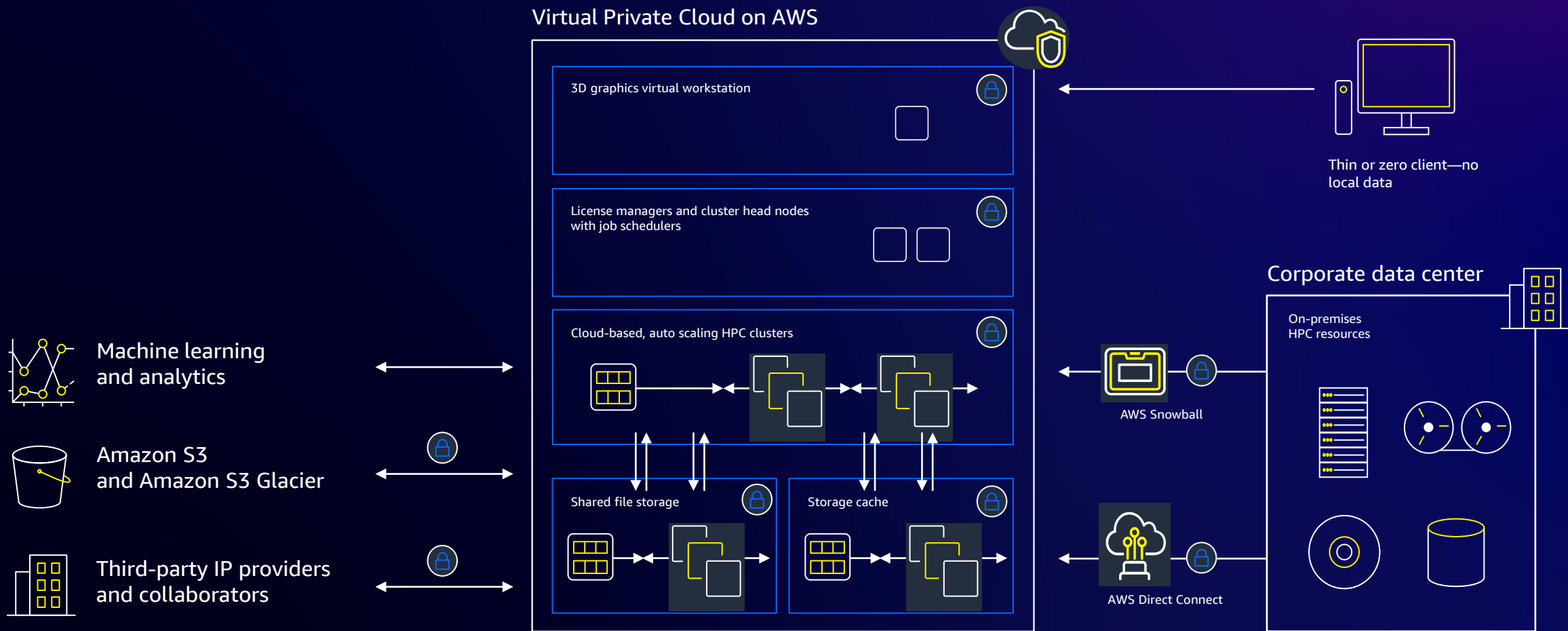
HPC



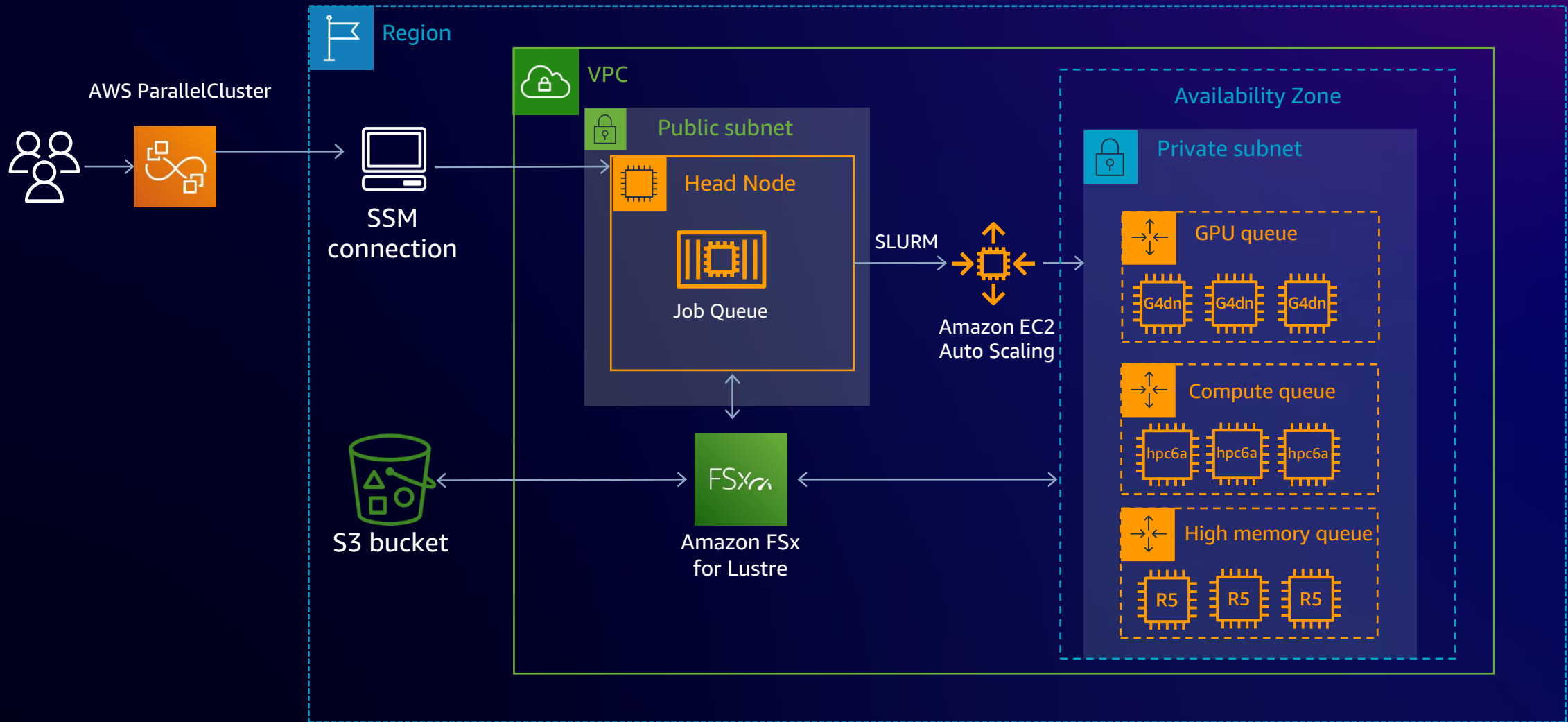
Traditional HPC architecture



High performance computing (HPC) on AWS



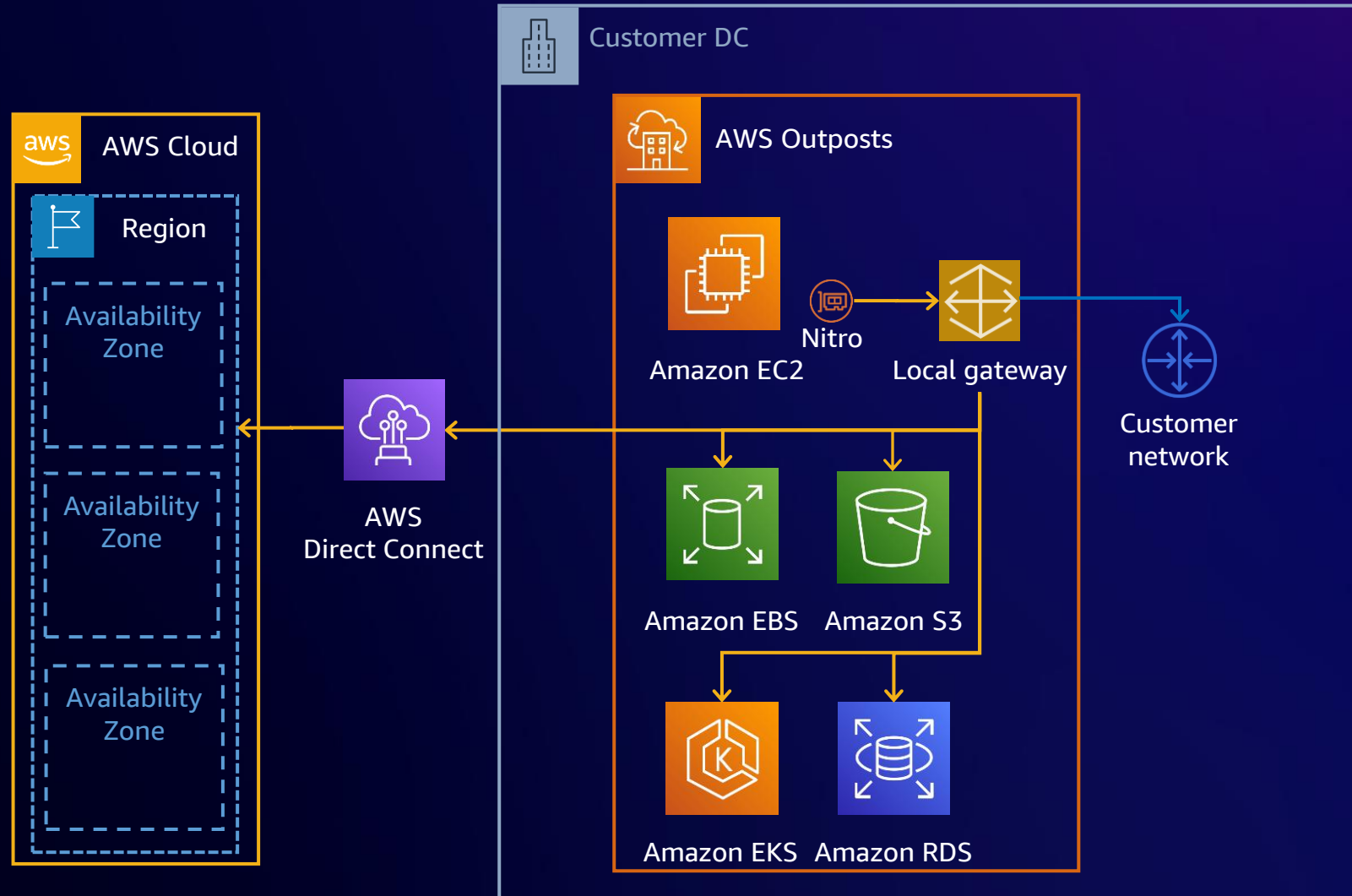
HPC reference architecture



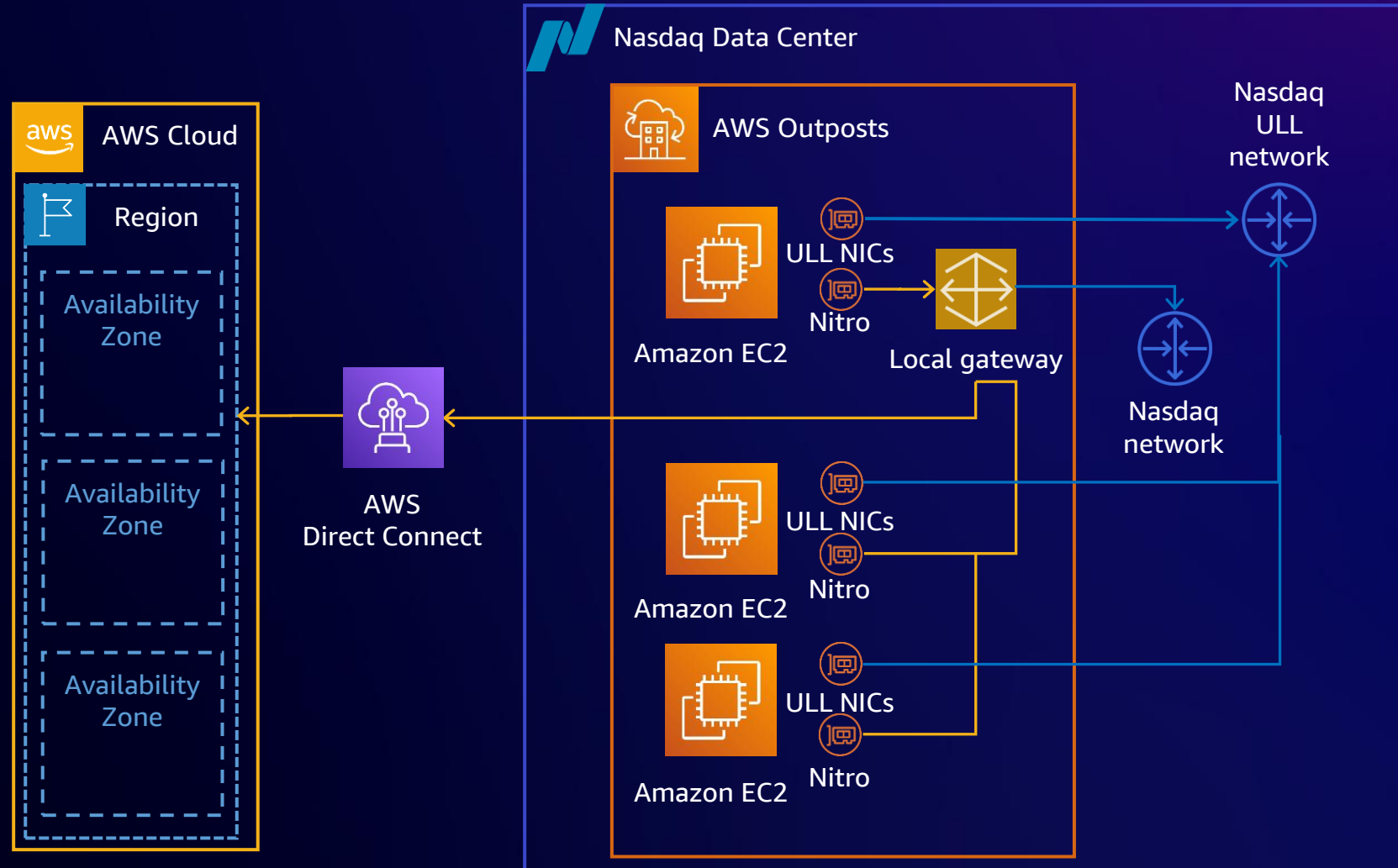
AWS Outposts



Outposts local connectivity



Nasdaq Outposts deployment



Region anatomy

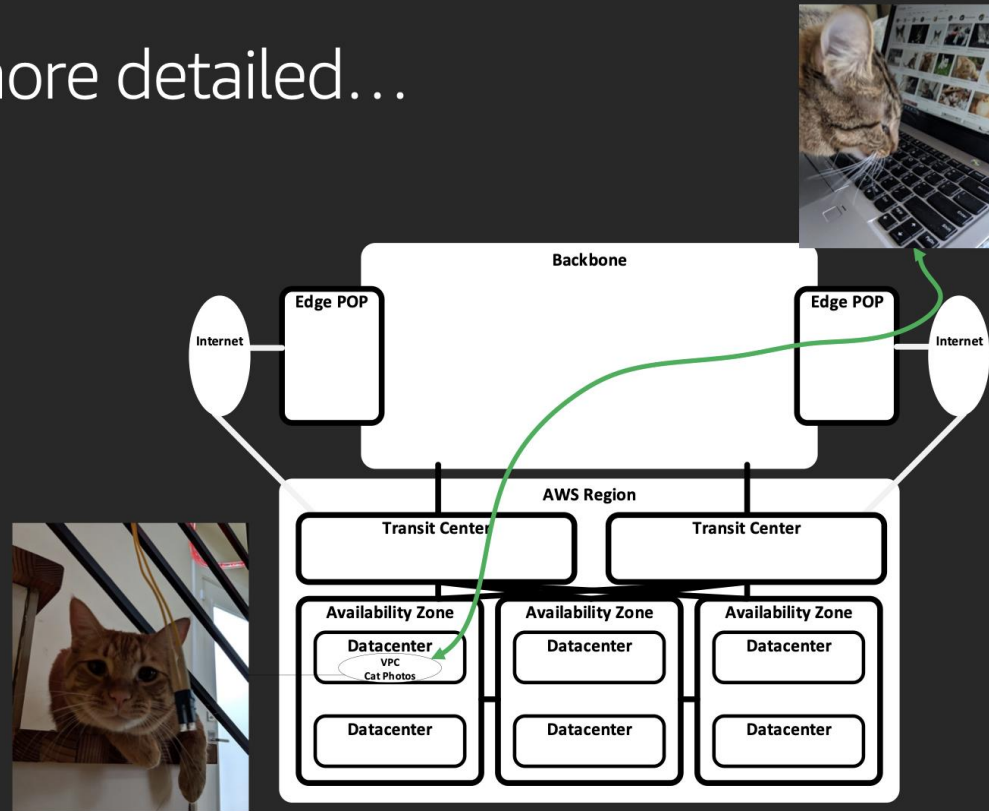


AWS global infrastructure



AWS Region networking

A bit more detailed...



aws
re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.



AWS network 10,000 foot view

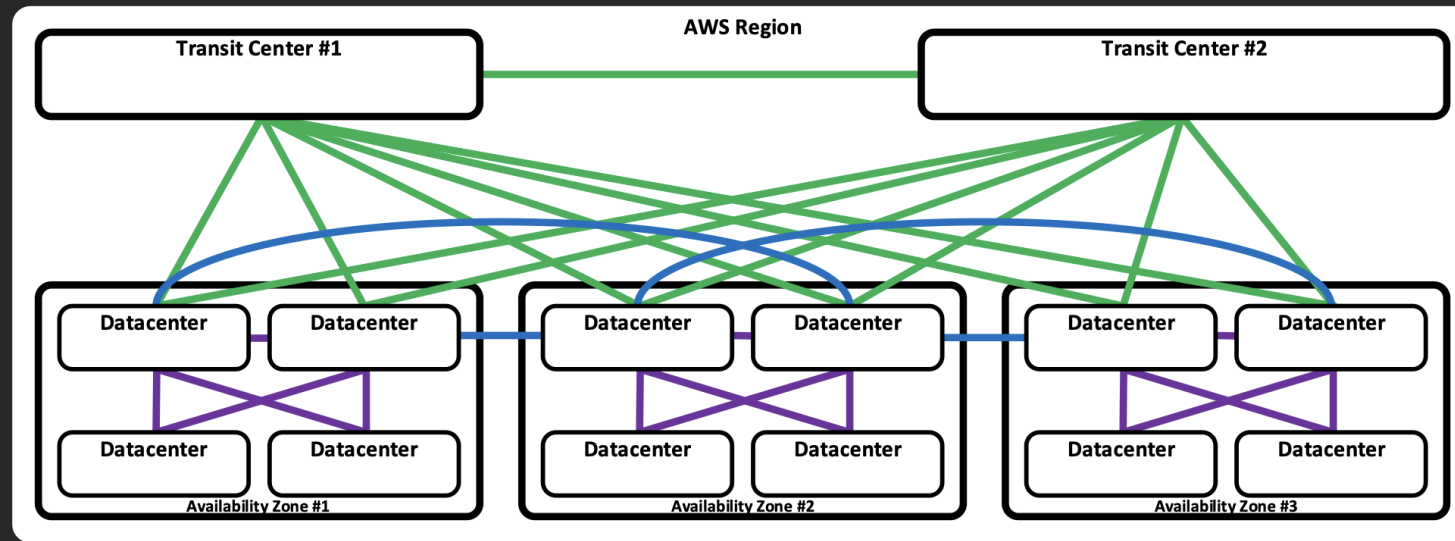


© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Source: https://www.youtube.com/watch?v=UObQZ3R9_4c

AWS Region networking

Availability Zones



aws
re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Region network

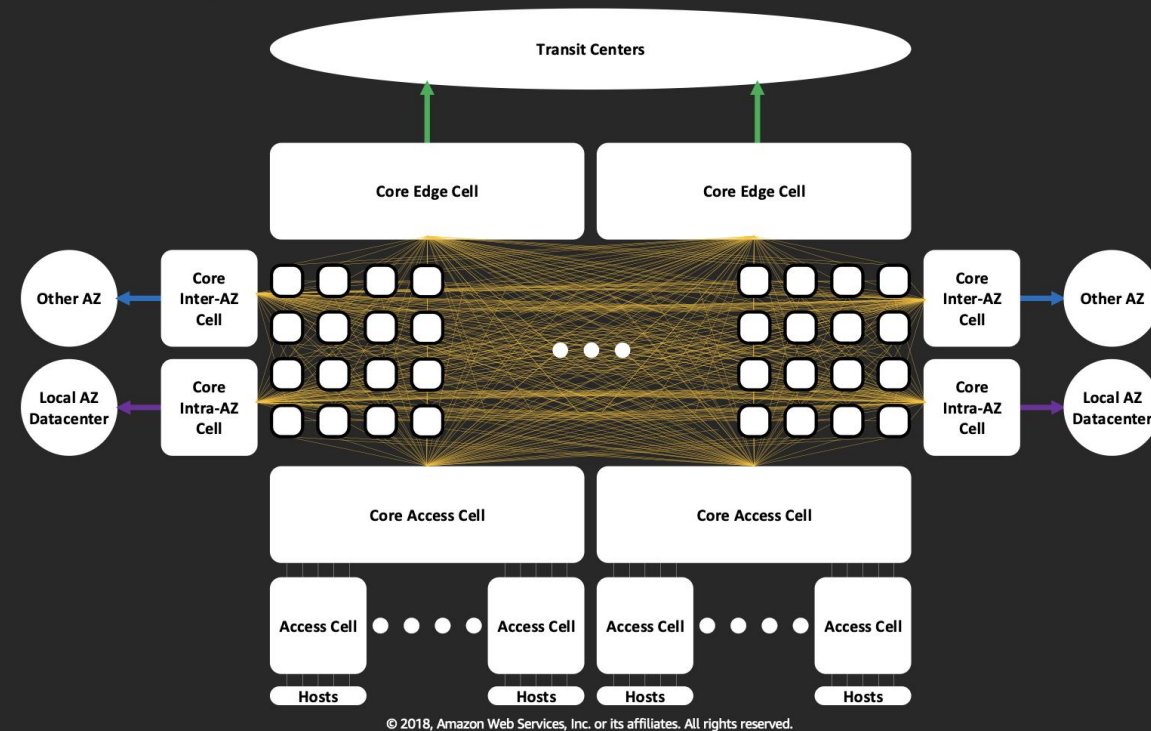


© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Source: https://www.youtube.com/watch?v=UObQZ3R9_4c

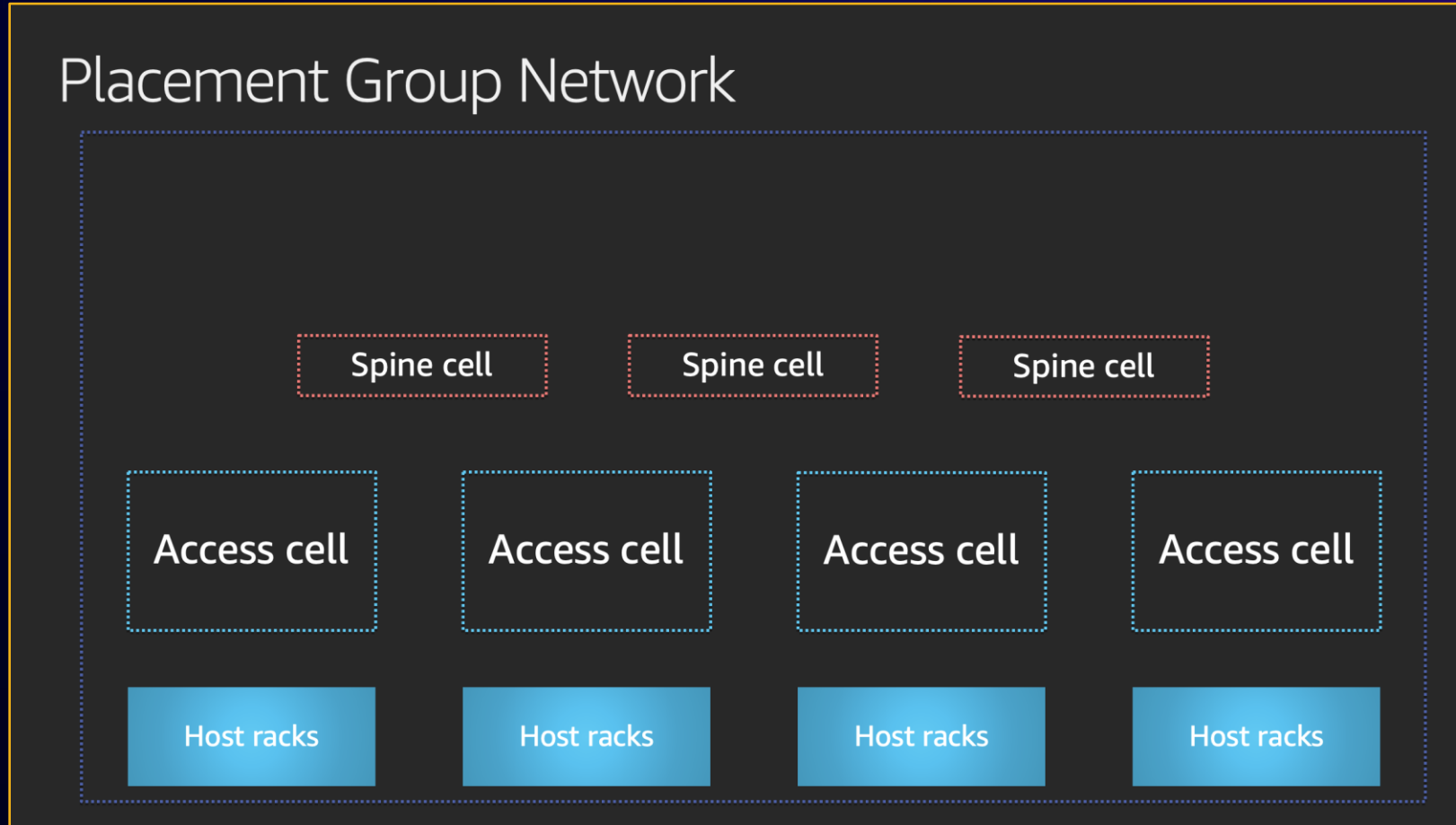
AWS Region networking

Building cellular data center architectures



Availability zone cell networking

AWS Region networking



Placement group

Latency considerations

Compute considerations



Location

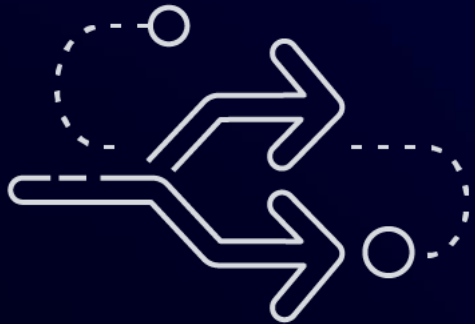


Compute
type



Optimizations

Network considerations



Network
path

TCP gRPC
HTTP
UDP QUIC

Protocols

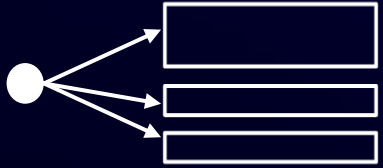


AWS
services

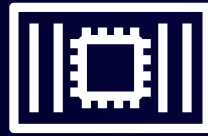


OS tuning

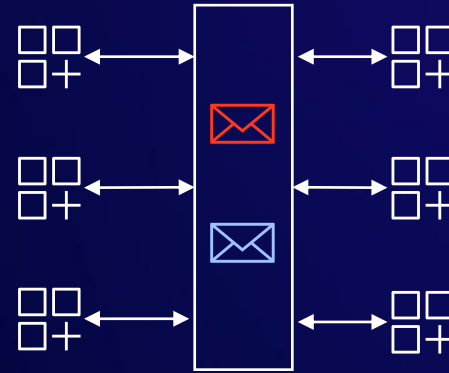
Deployment patterns considerations



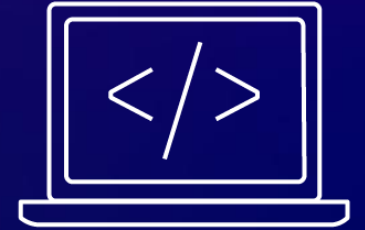
Queueing
systems



Virtualization
type



Messaging
buses



Programming
language

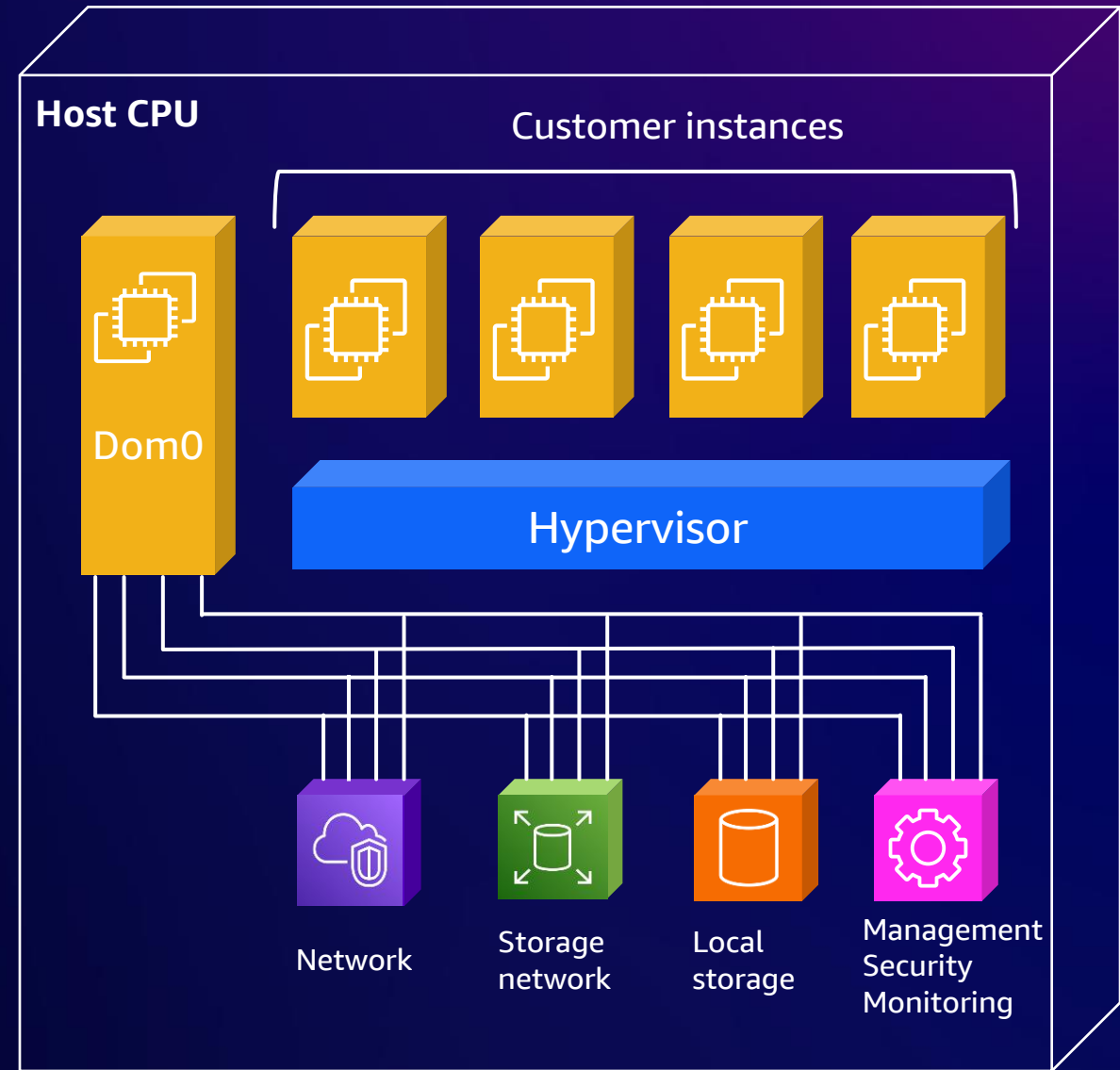
AWS Nitro System



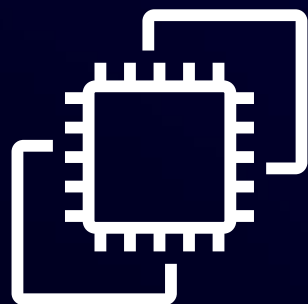
Typical cloud instance

In a typical cloud instance resources are consumed for cloud virtualization

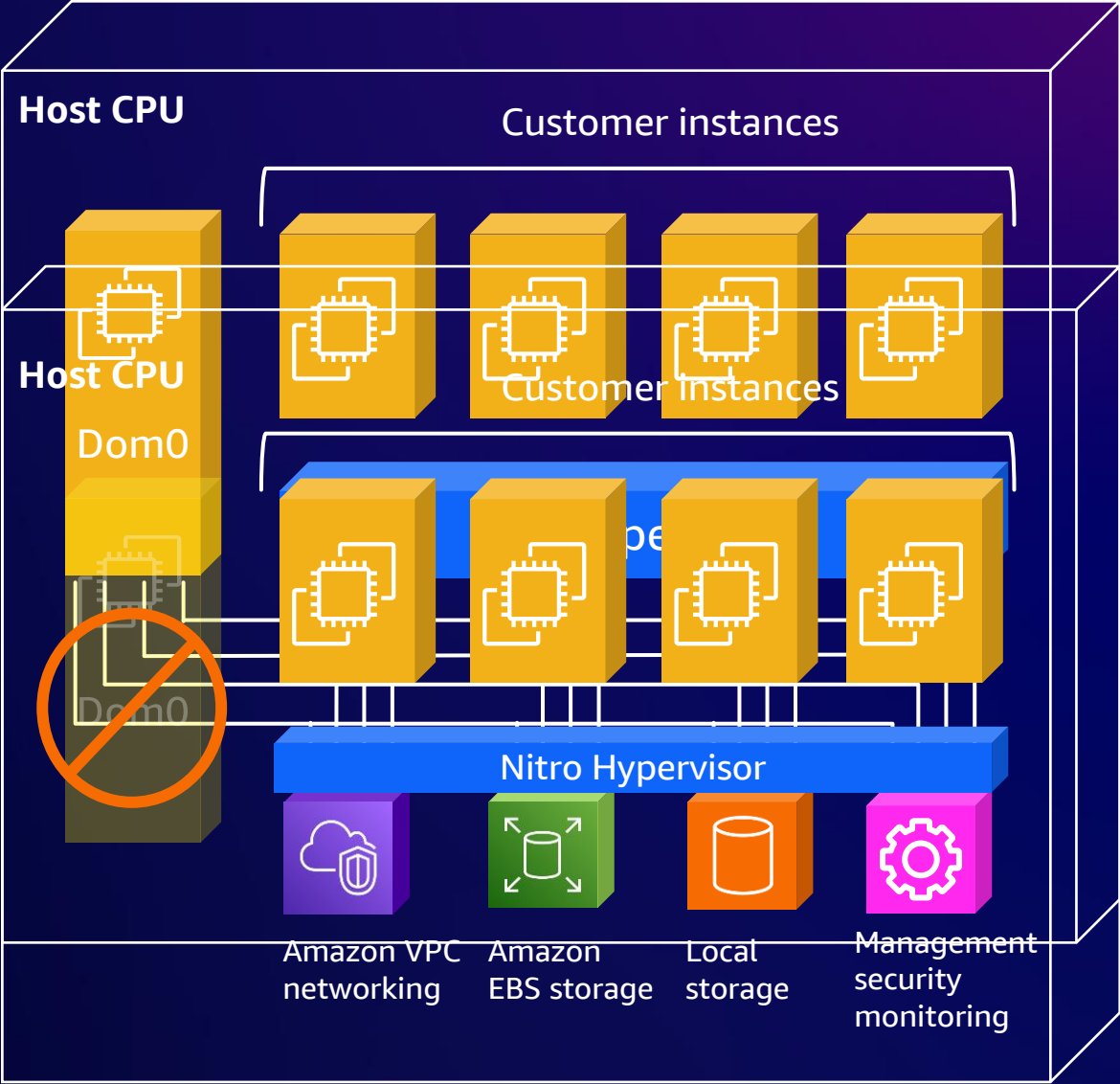
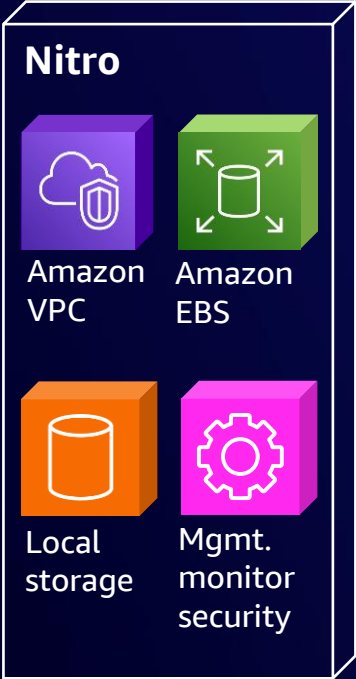
Only a part of the server is available to the customer



AWS Nitro System



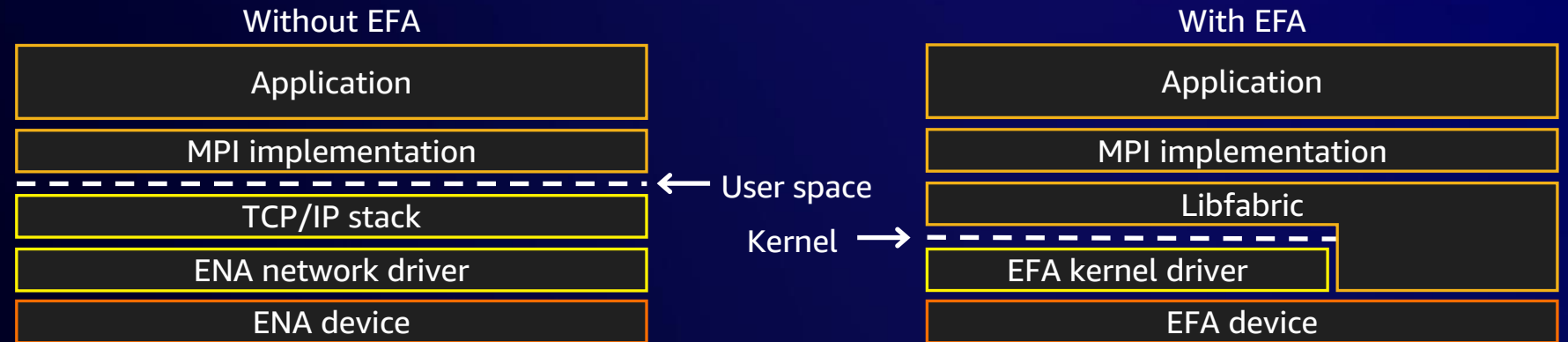
AWS Nitro System



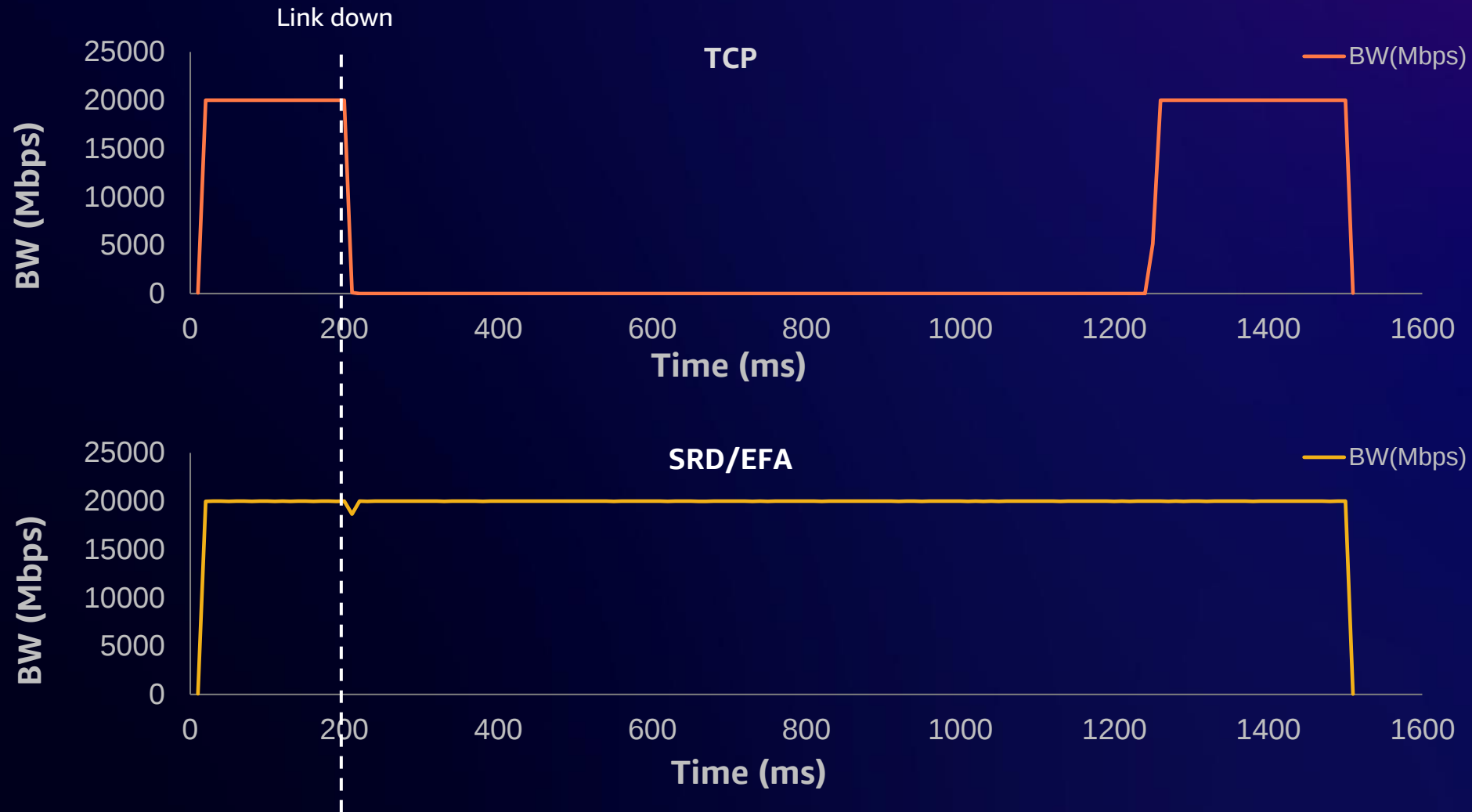
Elastic Fabric Adapter (EFA)



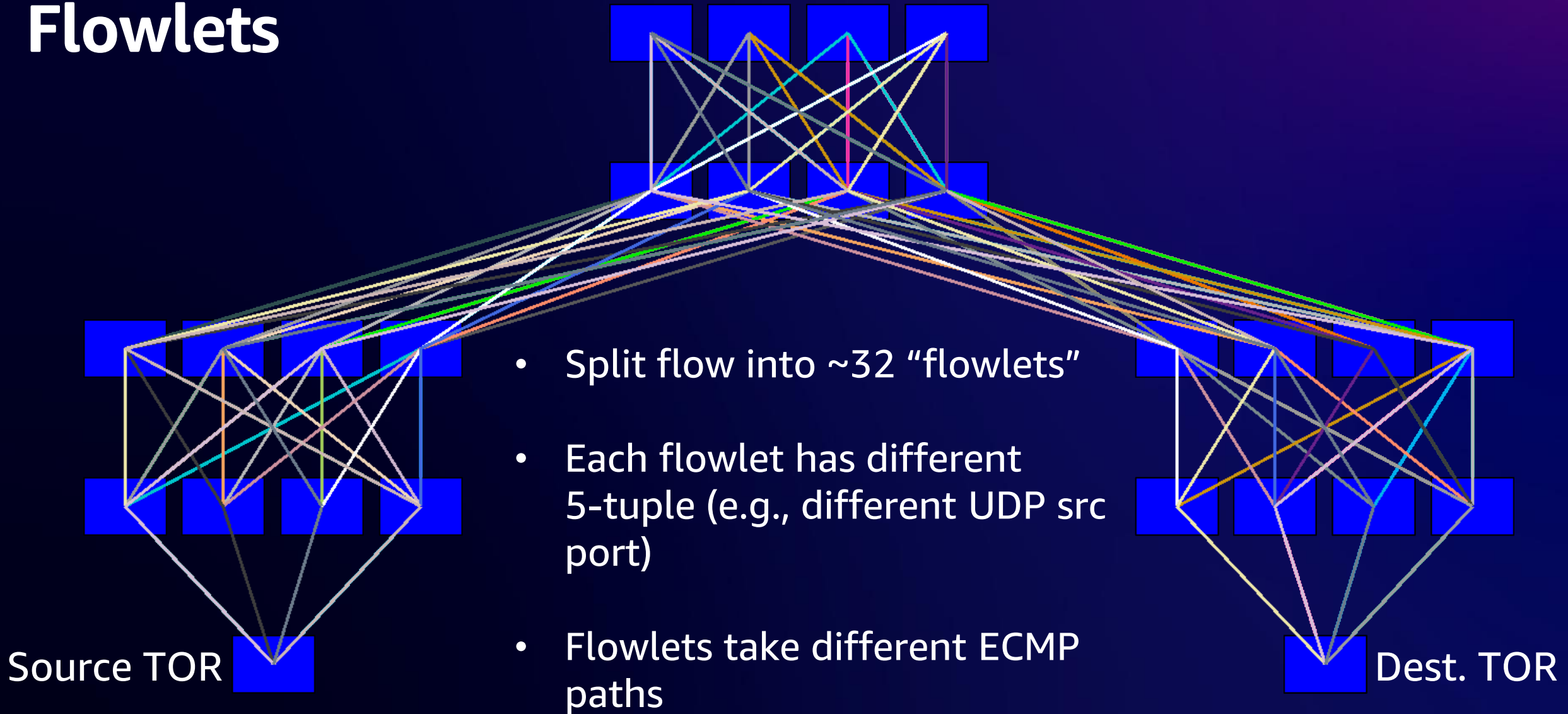
**SRD
protocol**



Graceful error recovery: Throughput



Scalable Reliable Datagram (SRD) protocol – Flowlets

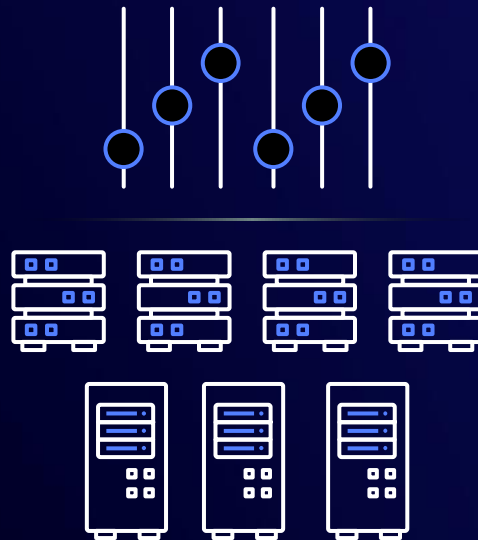


Amazon FSx for Lustre



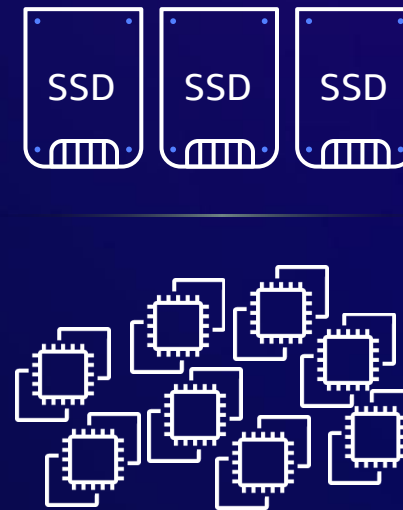
**High and
scalable
performance**

Parallel file system



+100 GiB/s throughput
Millions of IOPS
Consistent submillisecond latencies

SSD-based



Supports concurrent access from
hundreds of thousands of cores