AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

CMP333

Scaling network performance on next generation EC2 network optimized instances

Anti Gyori

Senior Product Manager
Amazon

John Pangle

Senior Product Manager
Amazon

Agenda

EC2 Network Optimized Instances

- EC2 journey to network optimized instances
- Why network optimized instances?
- EC2 network optimized offerings

EC2 Networking Innovation

- AWS Scalable Reliable Datagram (SRD) for EFA
- ENA Express for Amazon EC2 instances

2006: EC2 LAUNCHED

Amazon EC2 launched with a single general purpose instance type (M1)

2006

2006: EC2 LAUNCHED

Amazon EC2 launched with a single general purpose instance type (M1)

2008: LAUNCHED

First compute optimized instance type (C1)

2008

2008: LAUNCHED

First compute optimized instance type (C1)

2009: LAUNCHED

First memory optimized instance type (M2)

2009

2009: LAUNCHED

First memory optimized instance type (M2)

2013: LAUNCHED

First Nitro offload card based instance (C3), offload network processes

2013

2013: LAUNCHED

First Nitro offload card based instance (C3), offload network processes

2014: LAUNCHED

First instance collaborating with Annapurna Labs (C4), offload EBS storage

2014

2014: LAUNCHED

First instance collaborating with Annapurna Labs (C4), offload EBS storage

2015:

Amazon acquired Annapurna Labs

2015

2015:

Amazon acquired Annapurna Labs

2017: LAUNCHED

Introduced new hypervisor and full Nitro system (C5), offload remaining control plane and I/O components

2017

2017: LAUNCHED

Introduced new hypervisor and full Nitro system (C5), offload remaining control plane and I/O components

2018: LAUNCHED

First network optimized instance (C5n)

2018

Why?

Enable customers to scale and efficiently run network intensive workloads

Network Intensive Workloads

- Network Virtual Appliances
- Distributed Compute, HPC

Network Intensive Workloads

- Network Virtual Appliances
- CPU-based AI/ML
- Distributed Compute, HPC
- Real-time Comms, 5G UPF
- Data analytics

Network Intensive Workloads

- Network Virtual Appliances
- Distributed Compute, HPC
- Real-time Comms, 5G UPF
- Data analytics
- CPU-based AI/ML
- In-Memory Databases
- High Perf File Systems
- High Density Containers

Amazon EC2 Network Optimized Portfolio

Nitro Innovation For Network Intensive Workloads

Bandwidth

Up to 100Gbps Advanced Networking

Packet Performance

Improved packet performance vs core instances

Performance

Broad instance choice for all network intensive workloads

100Gbps Portfolio

C5N
COMPUTE-OPTIMIZED
(2018)

M5N
GENERAL PURPOSE
(2019)

M5DN
GENERAL PURPOSE
(2019)

R5N
MEMORY-OPTIMIZED
(2019)

R5DN
MEMORY-OPTIMIZED
(2019)

C6GN
COMPUTE-OPTIMIZED
(2020)

What Next?

Higher network bandwidth

- Scale network infrastructure throughput
- Reduce ingestion time from S3, data lakes

Improved packet performance

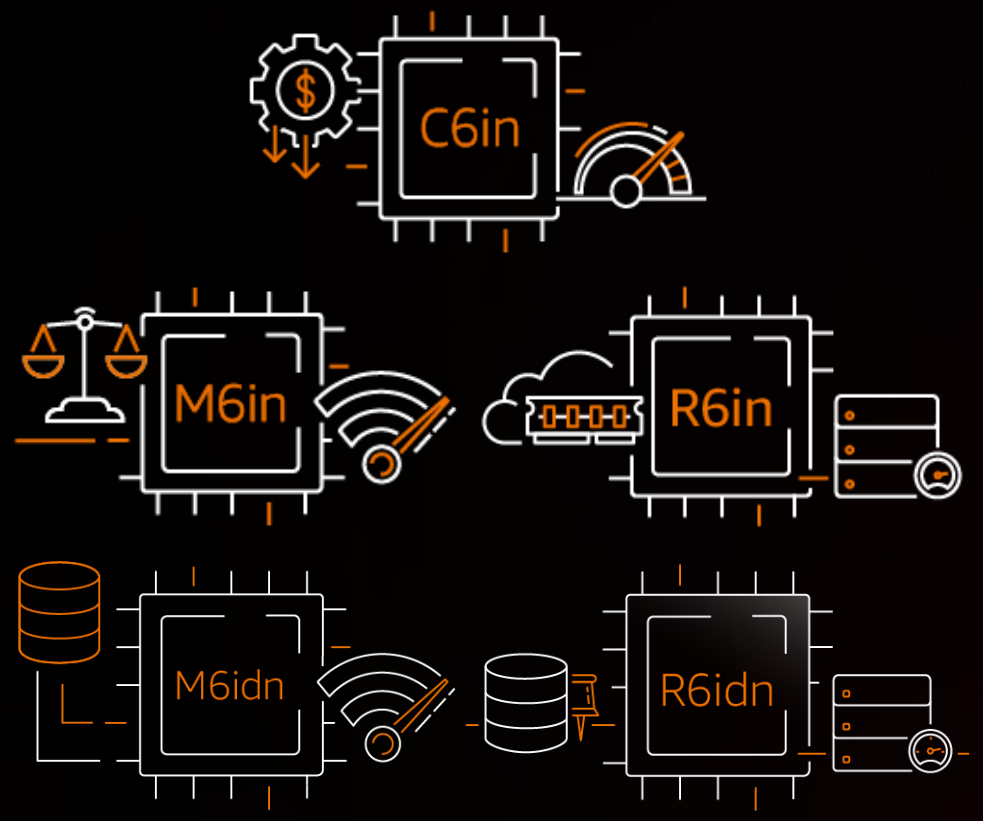
- Higher throughput for smaller packets

High networking ~~OR~~ AND high EBS performance

NEW

6th Gen EC2 Network Optimized Instances

HIGHEST X86-BASED NETWORKING PERFORMANCE WITHIN AMAZON EC2



200Gbps network bandwidth

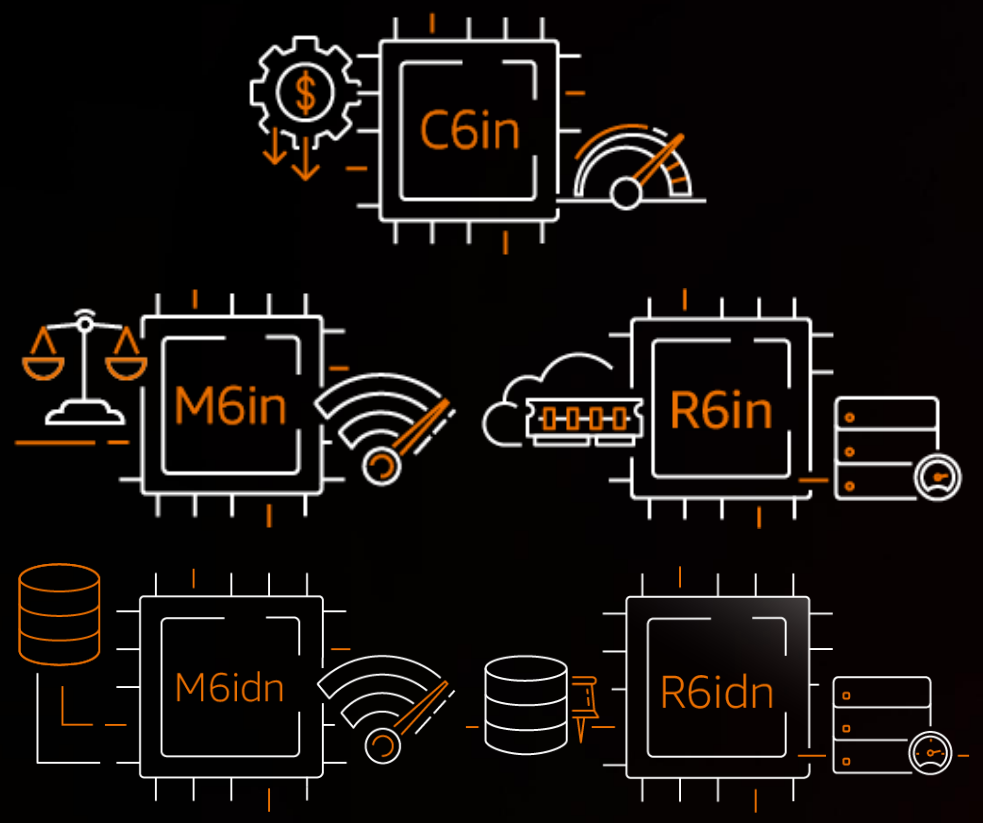
Up to 50Gbps burst network bandwidth

2x higher packet processing performance vs previous gen network optimized instances

NEW

6th Gen EC2 Network Optimized Instances

FIRST AMAZON EC2 INSTANCES TO OFFER HIGH NETWORKING & EBS PERFORMANCE



80Gbps EBS bandwidth, 350K EBS IOPS

Ideal for high performance file systems, databases, big data analytics

2x larger low latency instance storage

Customer Feedback

ARCTIC WOLF

Cybersecurity technology company providing turnkey, managed threat detection and response, risk management, cloud monitoring, and security training and awareness services to organizations.

"With the c6in instance class Arctic Wolf can push networking limits much further than any other instance class. We've already seen a **30% increase in packets per second performance**, and this performance increase has opened possibilities for us that we never had previously."

Customer Feedback

"Intel is excited to see the new C6in network optimized instances introduced in the AWS EC2 line up," said Bob Ghaffari, VP and GM of Intel, NEX.

"C6in, powered by 3rd Gen Intel Xeon Scalable processor, provides new instruction sets for cryptographic and vector processing. These instructions are foundational to delivering the packet processing performance and scaling needed by ISVs that use cloud instances to deliver their networking, security and AI/ML applications.

Compared to C5n instance, C6in instances deliver **1.6x IPsec performance boost** with new crypto instruction sets and software libraries, **7.3x performance boost in AI inferencing** with oneDNN and Intel Neural Compressor software, **2.8x connections per second for Nginx/Https**, **3x boost in threat prevention using Hyperscan software**."

Nitro Innovation

SECURITY

Enhanced security that continuously monitors, protects, and verifies instance hardware and firmware

PERFORMANCE

Better performance across CPU, networking, and storage

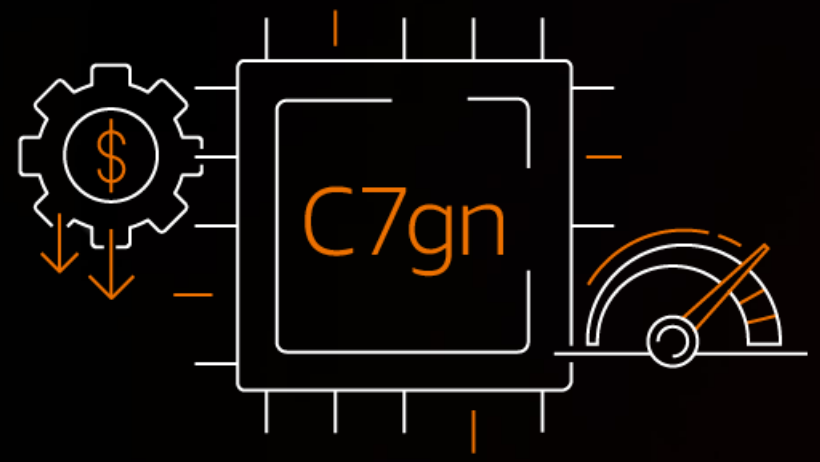
NITRO + GRAVITON

End-to-end solutions enable leadership performance and price-performance

NEW

Amazon EC2 C7gn Instances

FIRST AMAZON EC2 GRAVITON INSTANCE TO SUPPORT 200GBPS NETWORKING



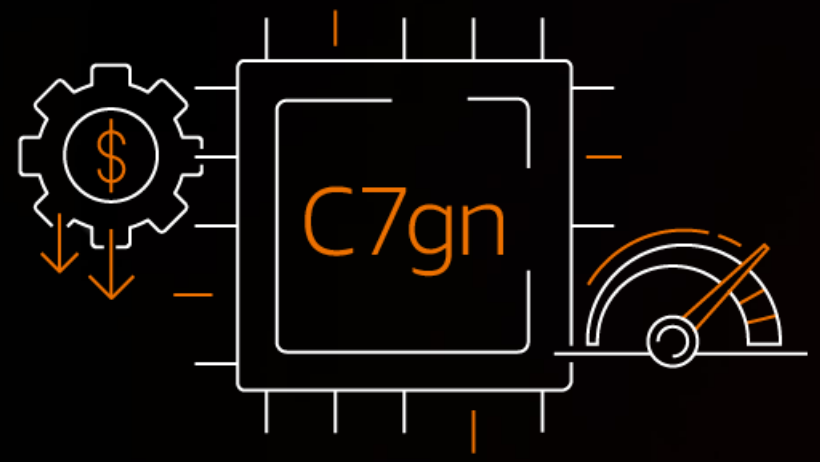
200Gbps network bandwidth (2x vs C6gn)

Up to 50Gbps burst network bandwidth

NEW

Amazon EC2 C7gn Instances

HIGHEST PACKET PERFORMANCE ACROSS EC2 NETWORK OPTIMIZED INSTANCES



200Gbps network bandwidth (2x vs C6gn)

Up to 50Gbps burst network bandwidth

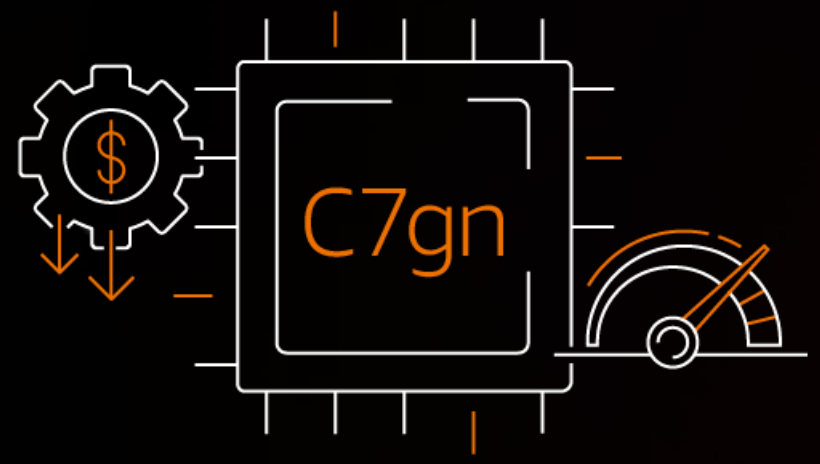
Over 50% higher packet performance vs C6gn

2x higher PPS/vCPU vs C6in

NEW

Amazon EC2 C7gn Instances

HIGHEST NETWORK PERFORMANCE ACROSS EC2 NETWORK OPTIMIZED INSTANCES



200Gbps network bandwidth (2x vs C6gn)

Up to 50Gbps burst network bandwidth

Over 50% higher packet performance vs C6gn

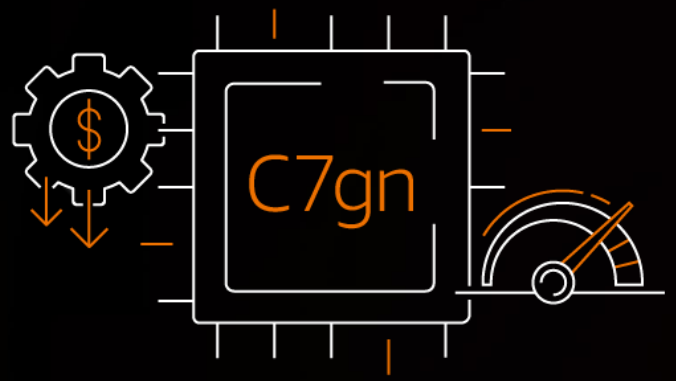
2x higher PPS/vCPU vs C6in

Highest aggregate and per vCPU network performance

NEW

Amazon EC2 C7gn: Nitro v5 Card

C7GN: FIRST EC2 INSTANCE FEATURING NEXT GEN NITRO V5 CARD



Nitro v5 card:

Latest generation interfaces (PCIe Gen5, DDR5)

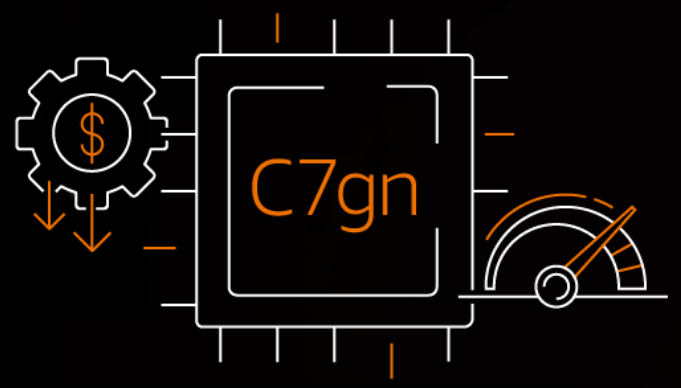
30% lower latency (Nitro v5 chip)

40% better performance/Watt (Nitro v5 chip)

NEW

Amazon EC2 C7gn: Nitro v5 Card

C7GN: POWERED BY LATEST GRAVITON3E CPU



Nitro v5 card:

Latest generation interfaces (PCIe Gen5, DDR5)

30% lower latency (Nitro v5 chip)

40% better performance/Watt (Nitro v5 chip)

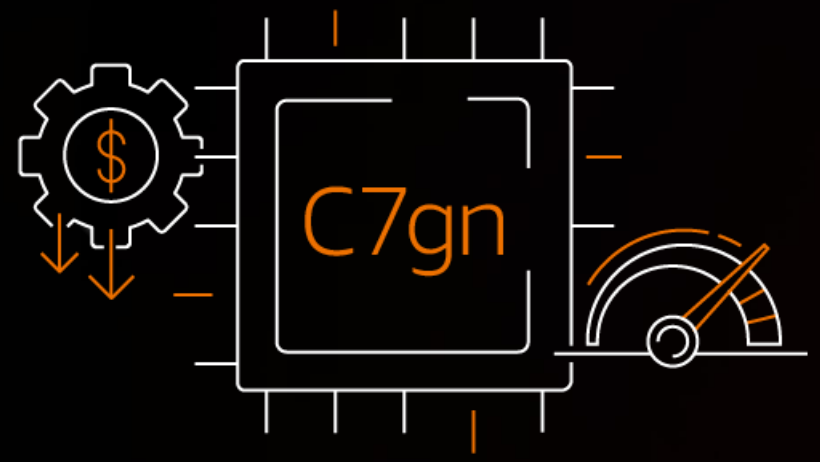
Graviton3E CPU:

Up to 35% higher vector instruction performance vs existing Graviton3 instances

NEW

Amazon EC2 C7gn: Call To Action

APPLY FOR PREVIEW



C7gn available for preview today. Be among the first to experience all this performance!

https://pages.awscloud.com/C7gn-Preview.html

Nitro Innovation

Nitro Card

Local NVMe storage

Amazon EBS

Networking, monitoring, and security

Nitro Security Chip

Integrated into motherboard

Protects hardware resources

Nitro Hypervisor

Lightweight hypervisor

Memory and CPU allocation

Bare metal-like performance

Nitro Enclaves

Isolated environments for highly sensitive data

Nitro SSD

60% lower I/O latency

Firmware Upgrades w/o Interruption

Encryption at rest

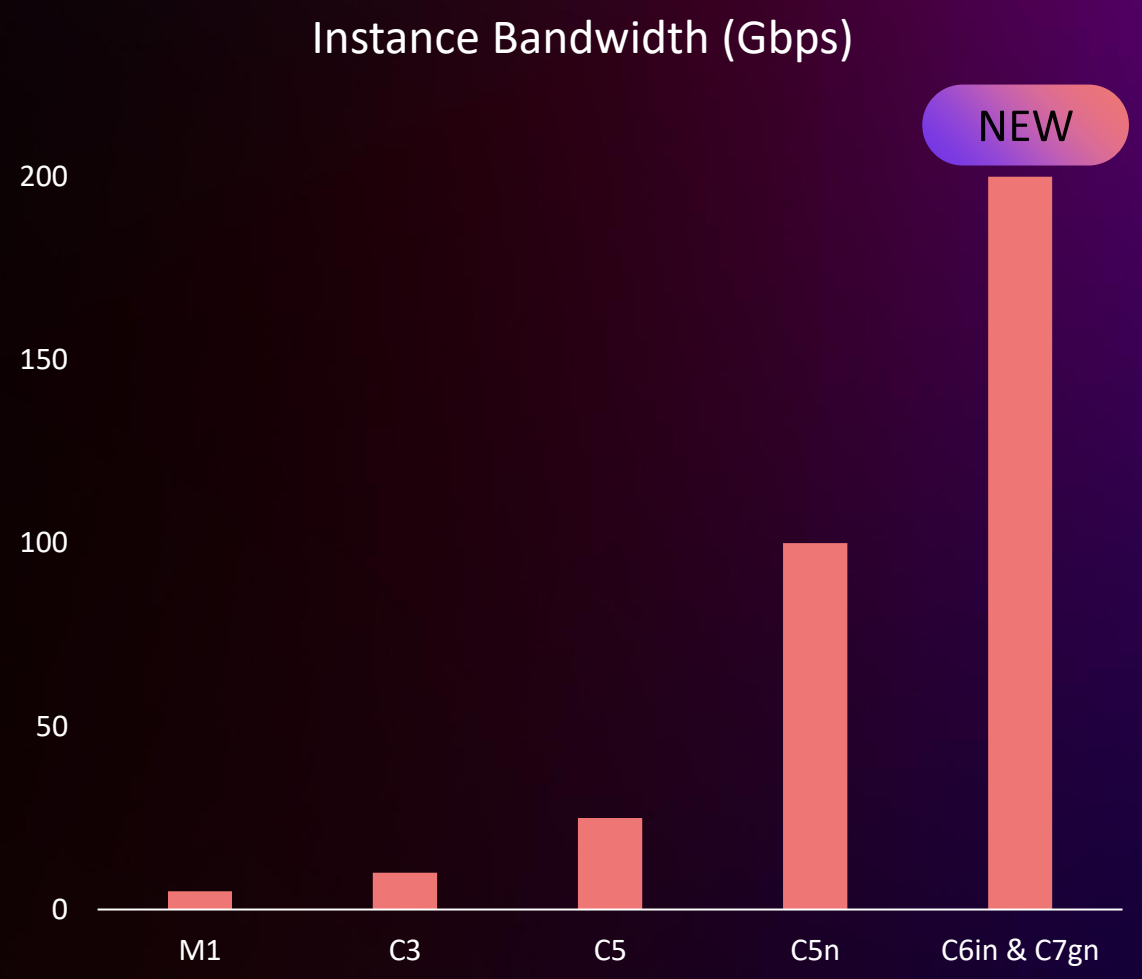
Nitro TPM

TPM 2.0 specification

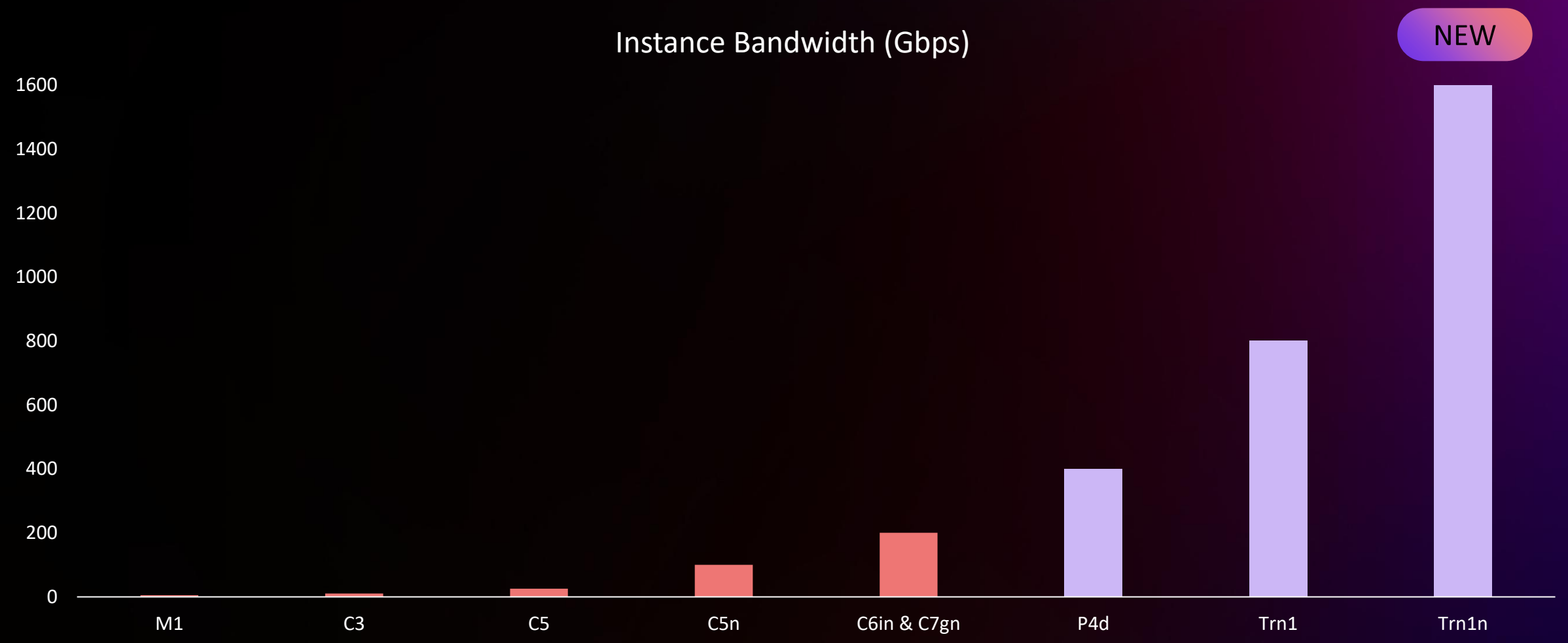
Cryptographic attestation of instances integrity

Elastic Network Adapter

- Offload network functions from Instance
  - Free up server for your applications
- Encapsulation, security groups, routing
- Enabled Enhanced Networking
  - PPS optimized
  - Low Latency
  - SRIOV
  - 200 Gbps bandwidth for network optimized instances using multiple network cards

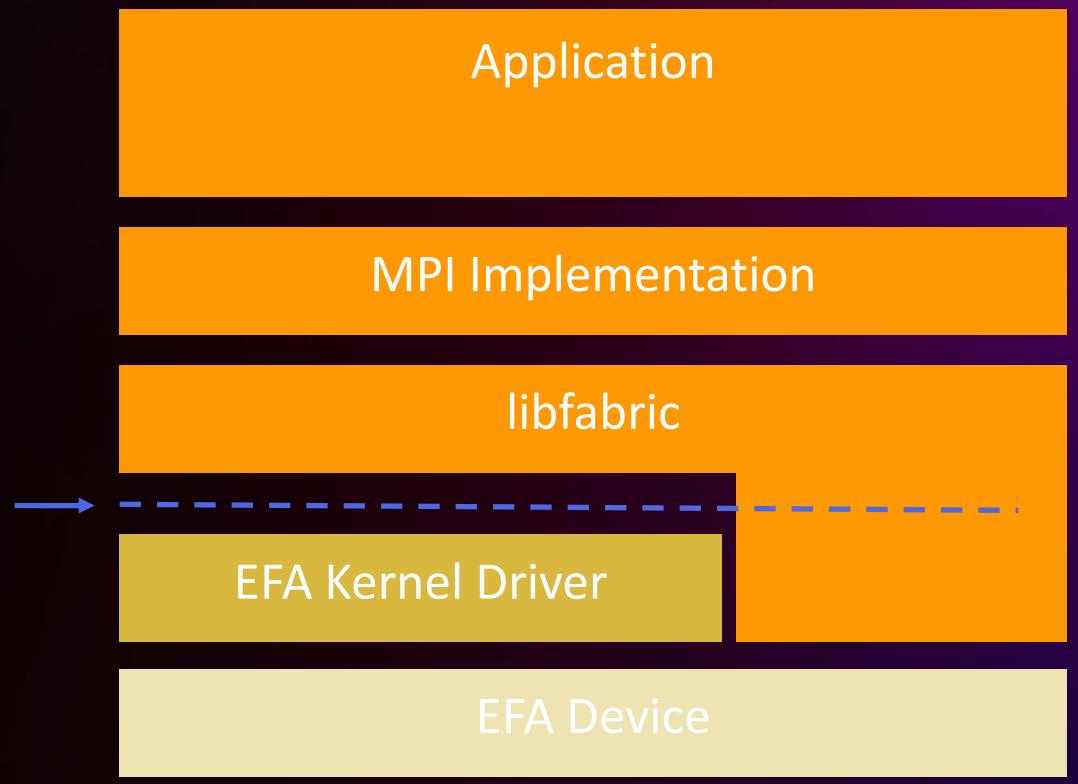


Accelerated Computing Instances



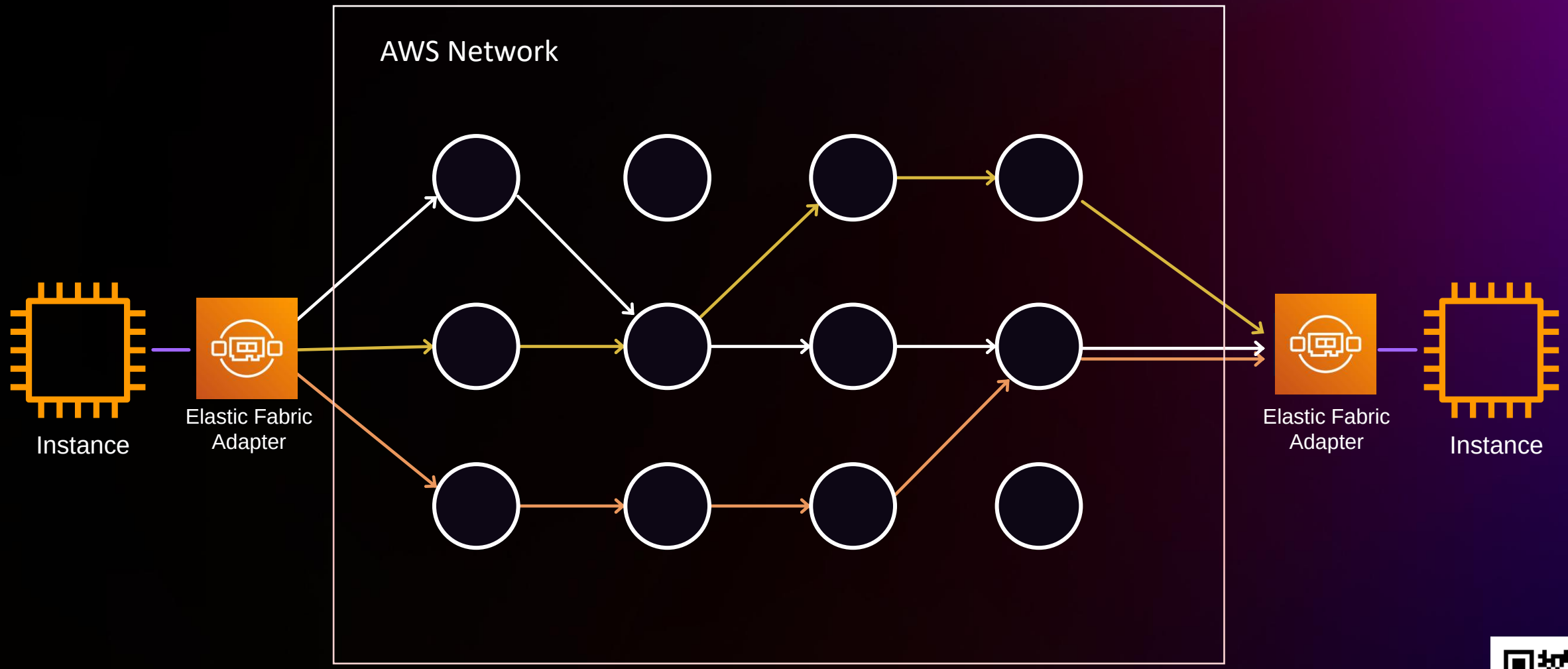
Elastic Fabric Adapter

- Machine Learning, HPC Applications
  - High Bandwidth, Low Latency
  - Distributed Workloads
- NCCL/MPI middleware Communication
- LibFabric
- Kernel Bypass
- EFA Interface
  - Custom built protocol Scalable Reliable Datagram (SRD)



Scalable Reliable Datagram (SRD)

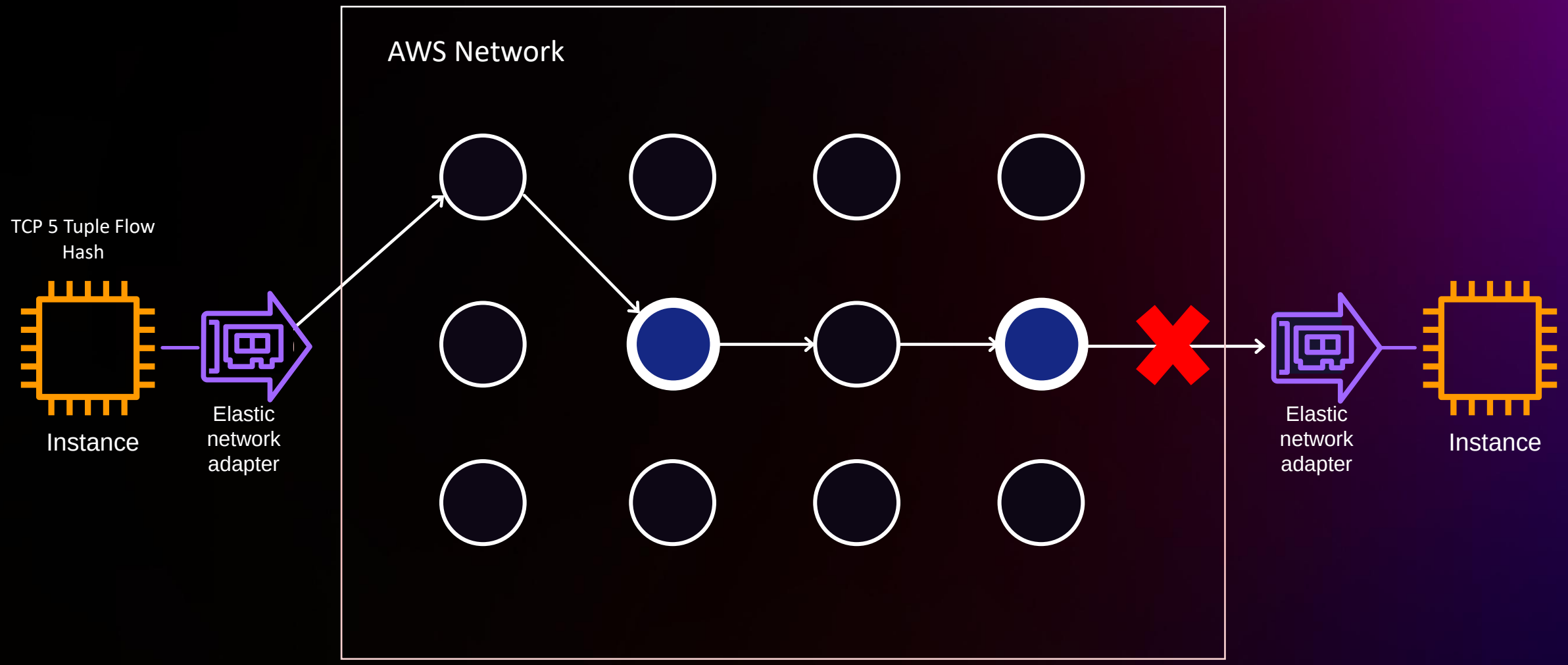
LOW LATENCY FOR NETWORK-INTENSIVE APPLICATIONS



ENA Networking Opportunity

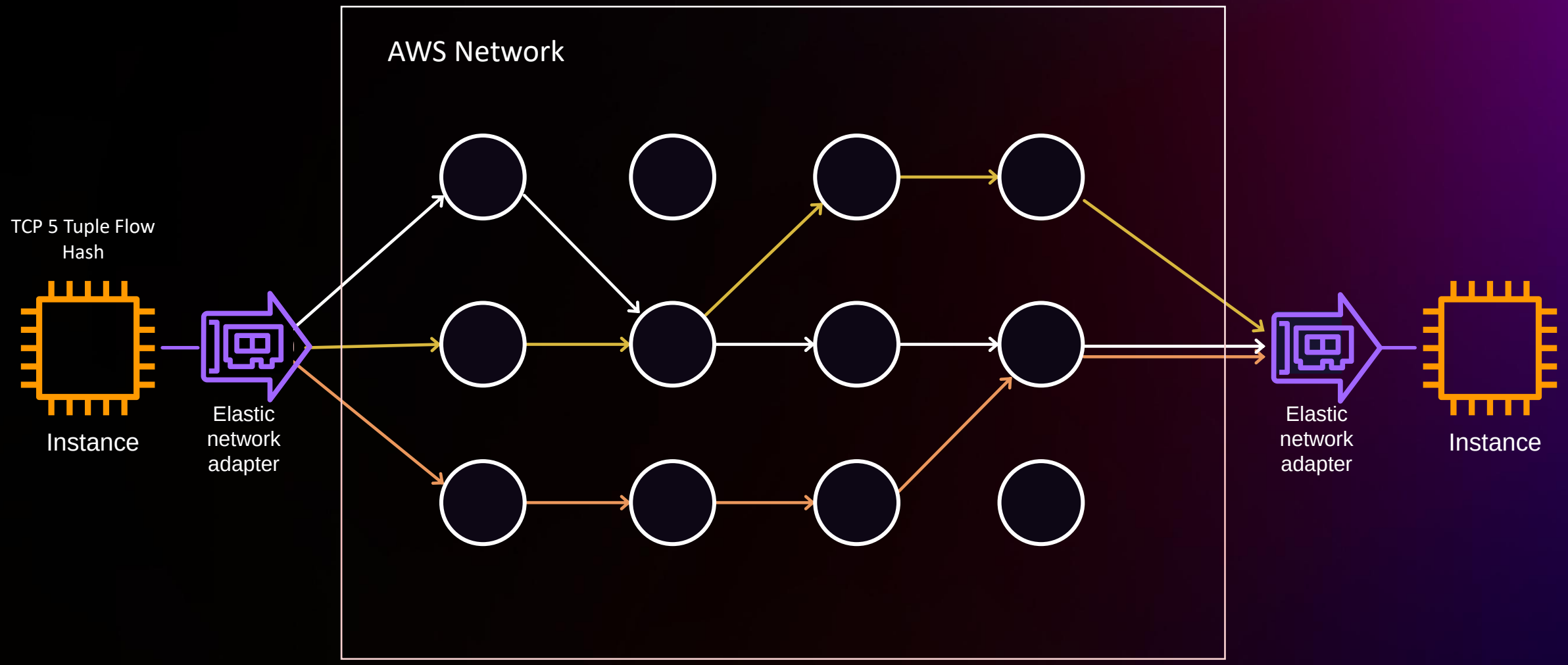
- How can we leverage SRD for general purpose apps?
  - It provides the latency and throughput
  - Its retransmits quicker than TCP, hiding network inefficiencies
- BUT
  - Can it replace TCP/UDP again?
  - How does it handle packet delivery?
  - How do customers manage it?

TCP Flow Hashing

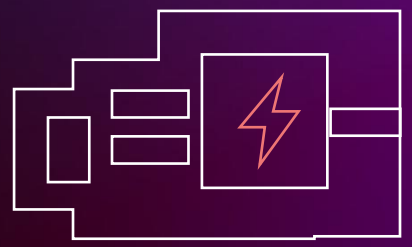


Introducing: ENA Express

NEW



ENA Express Benefits



5x

Single Flow Bandwidth 5 to 25 Gbps

85%

P99.9 Tail Latency/Jitter Reduction

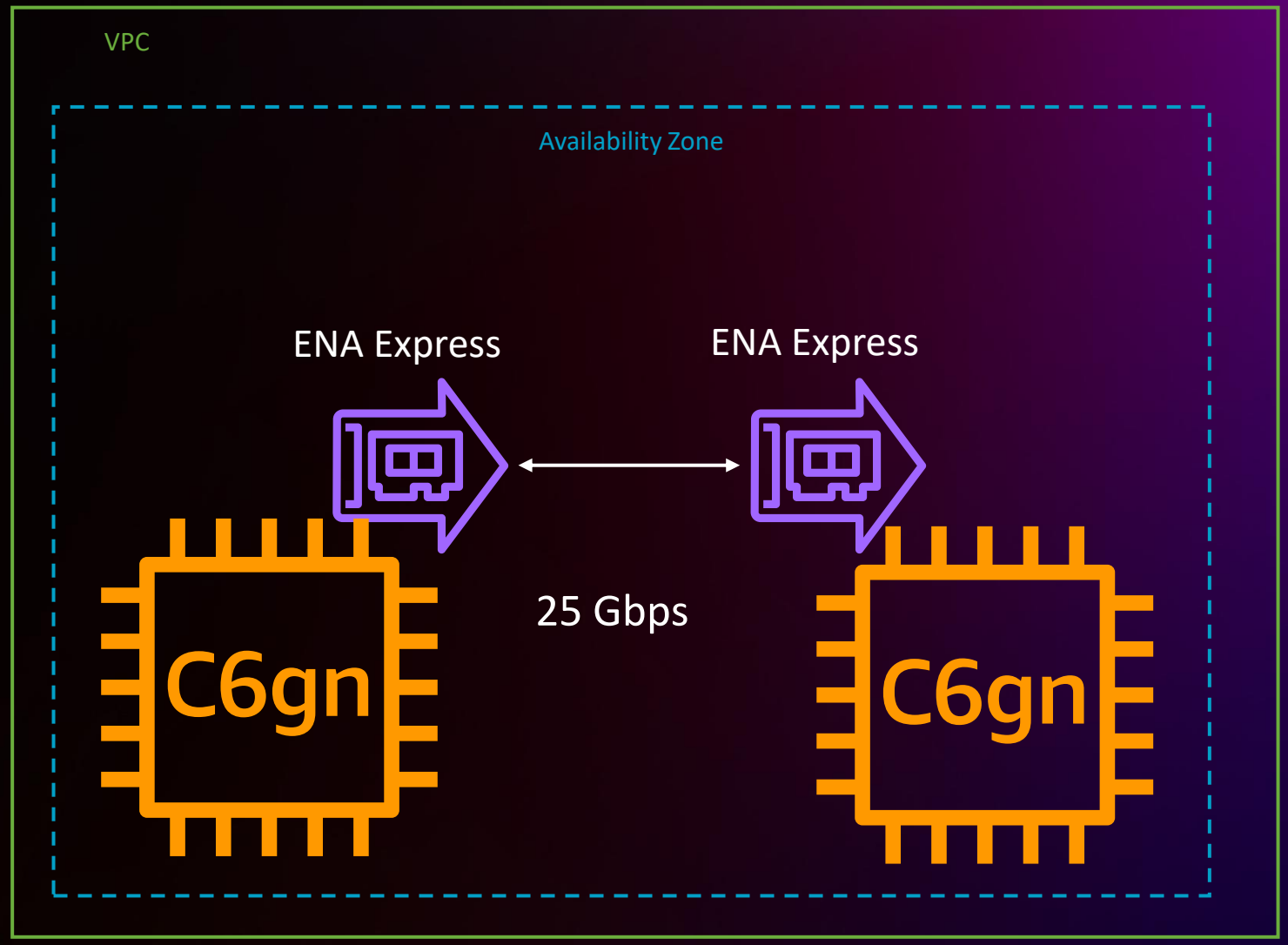
Simple Configuration

Same AZ Support

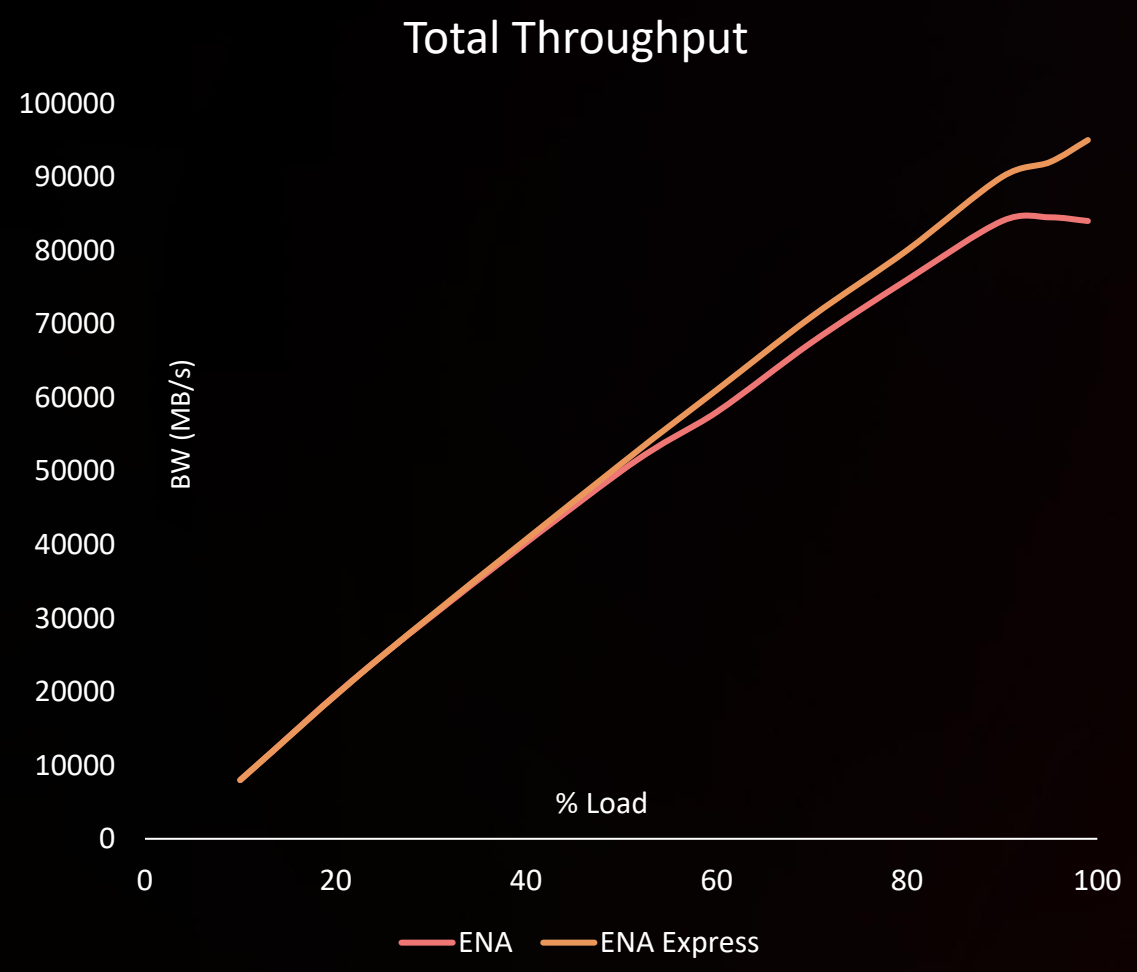
Transparent to TCP/UDP

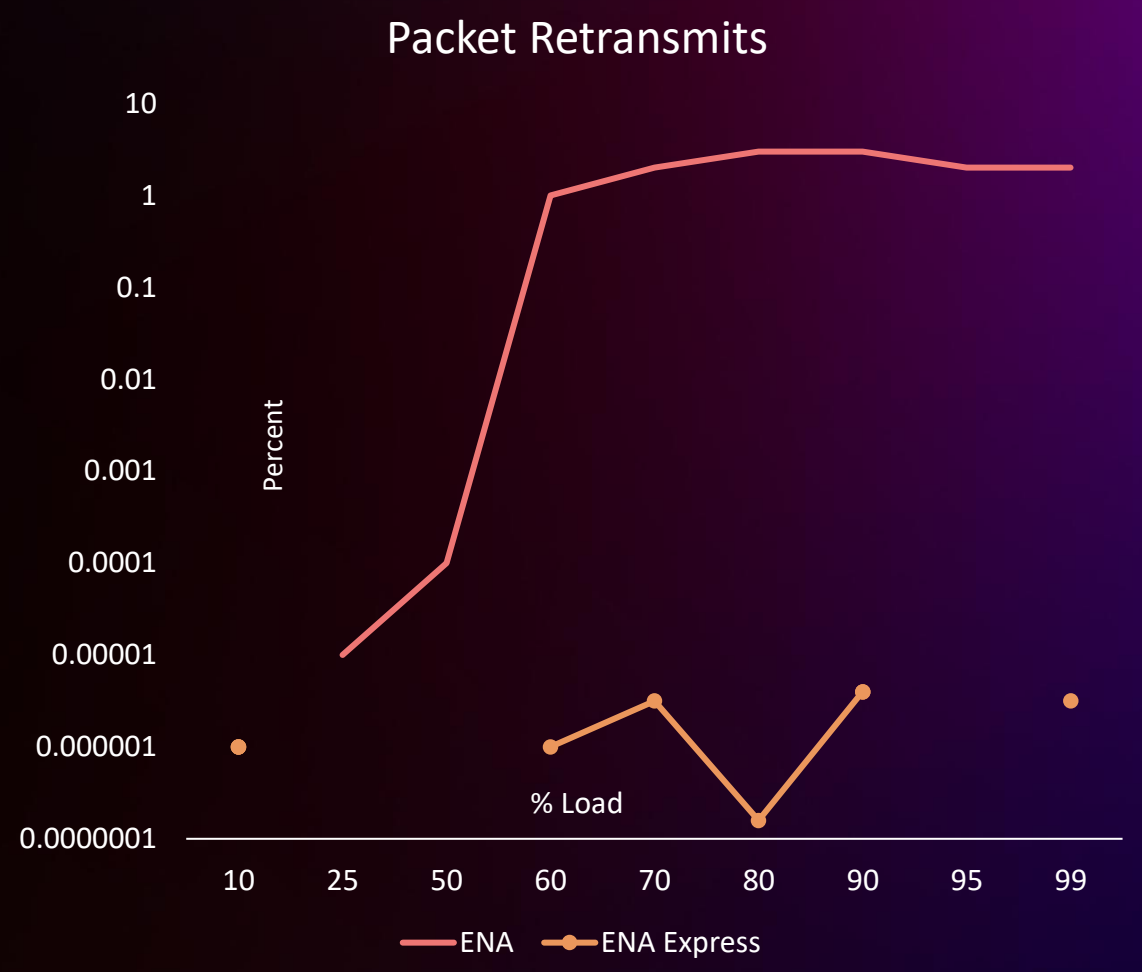
How do you get started?

- Launch two 6th gen Instances
- Configure the network interface:
- If you use jumbo frames – set max MTU to 8900
- Load iperf
- Start sending traffic

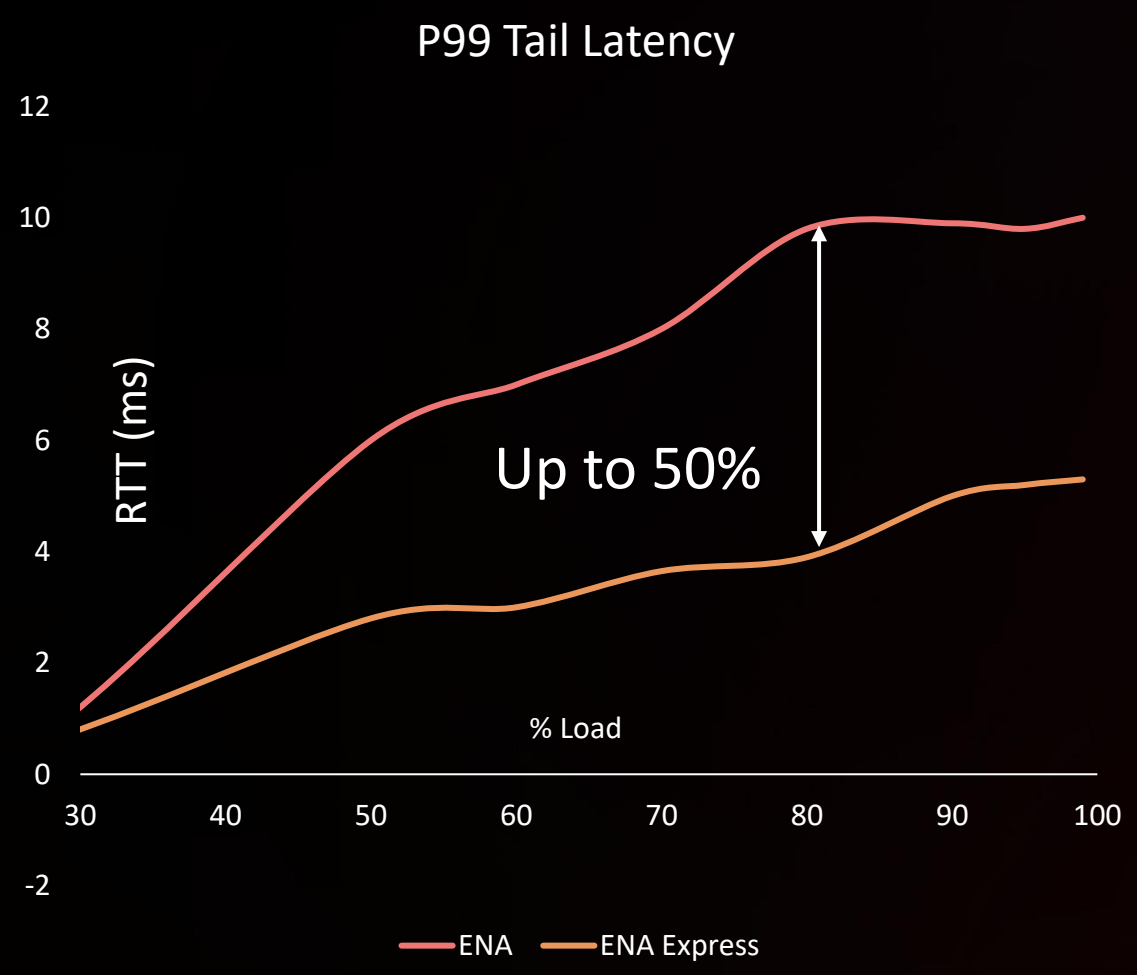


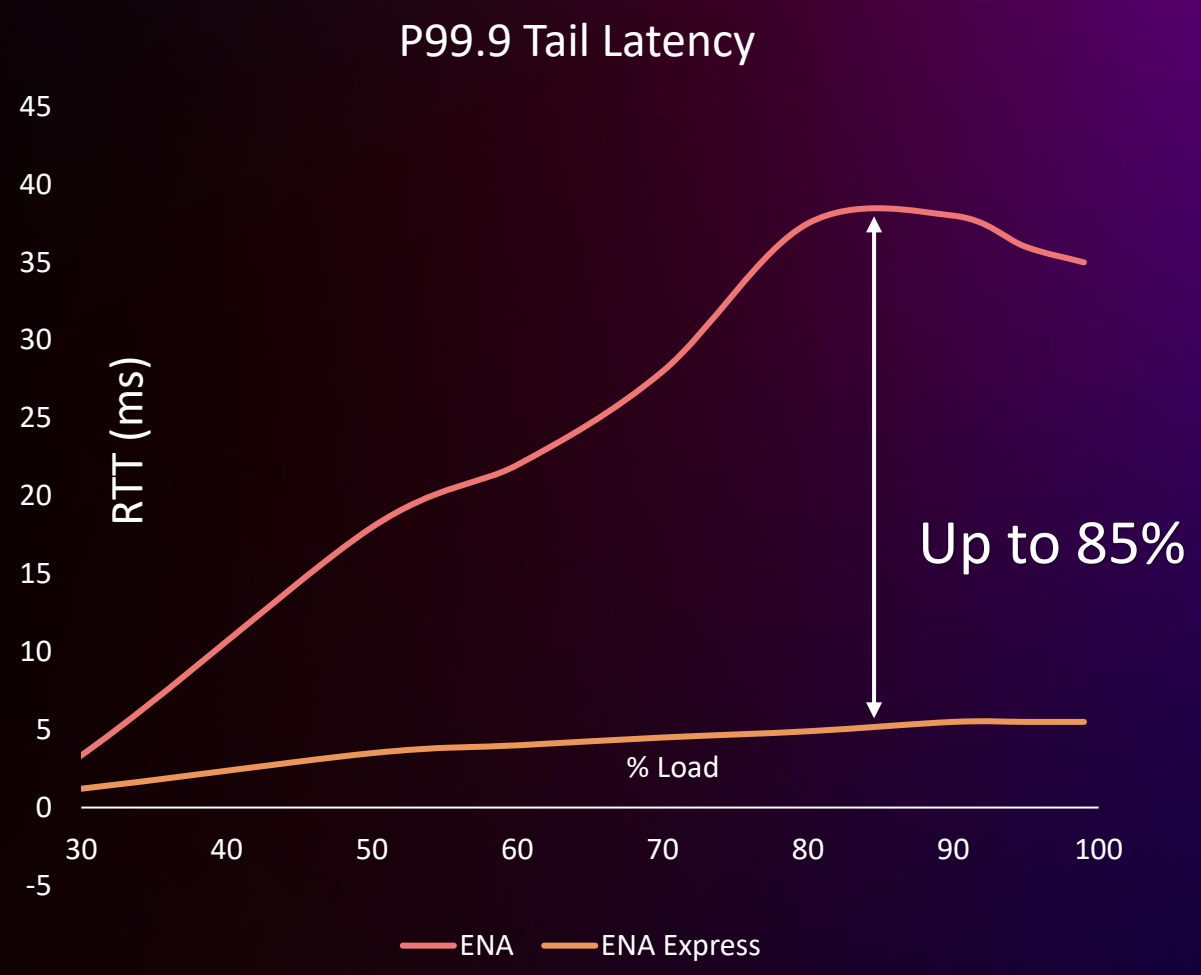
Benchmarks



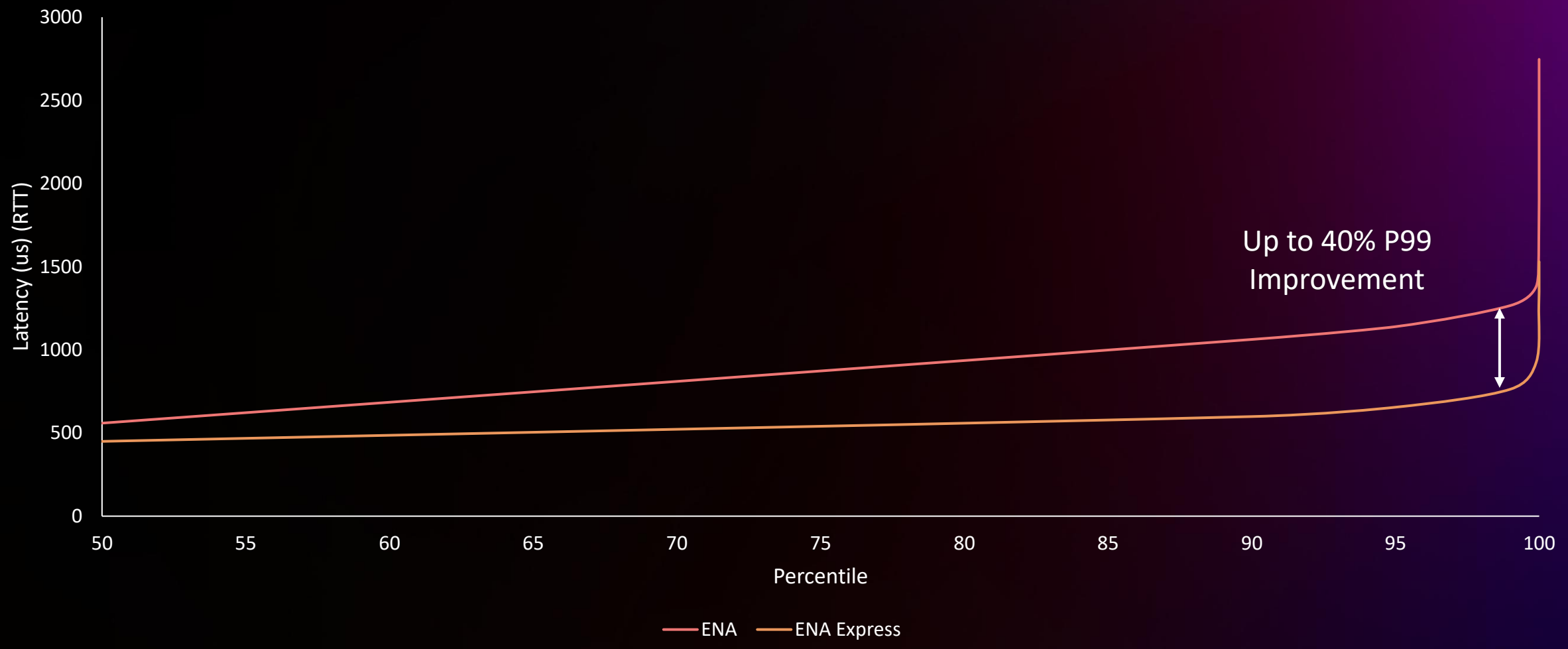


Tail Latency



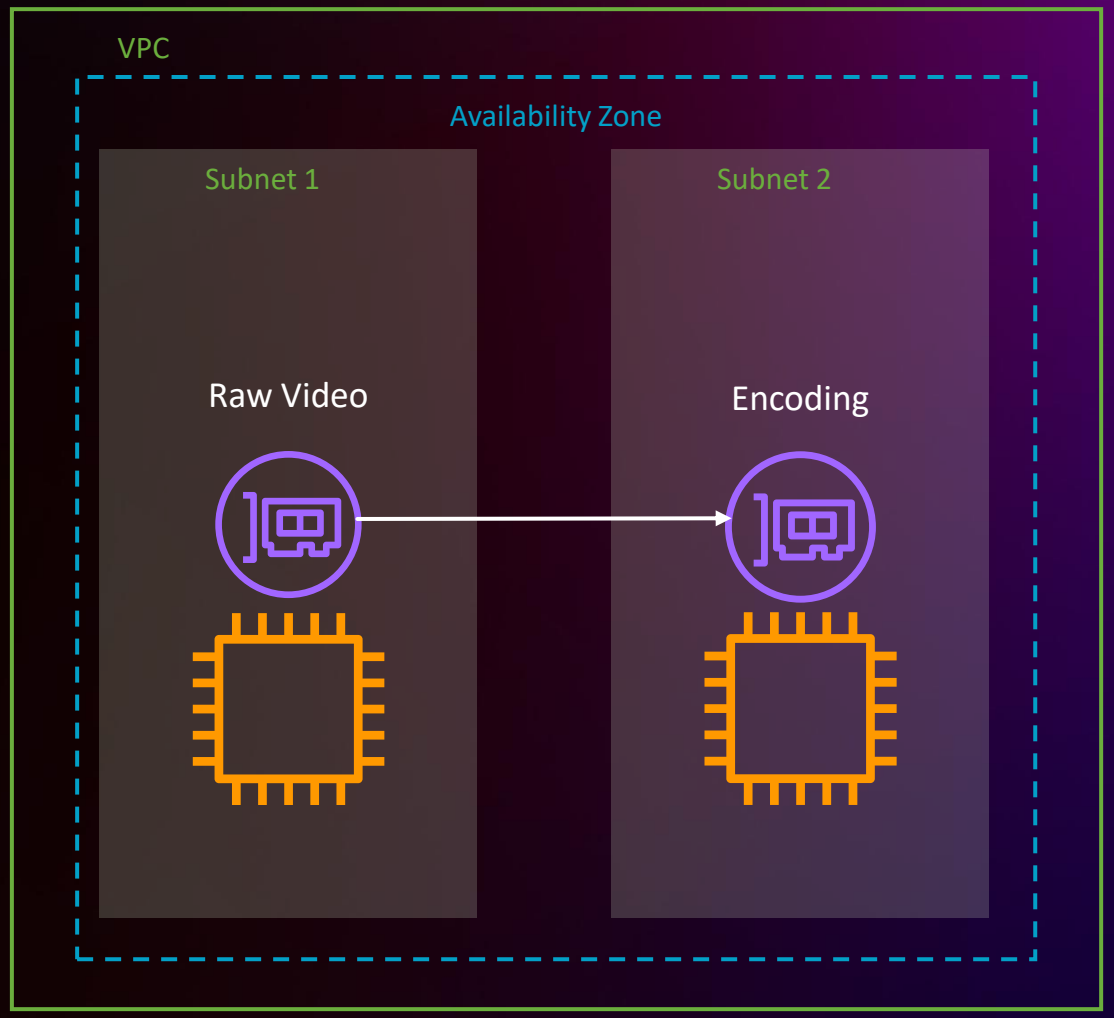


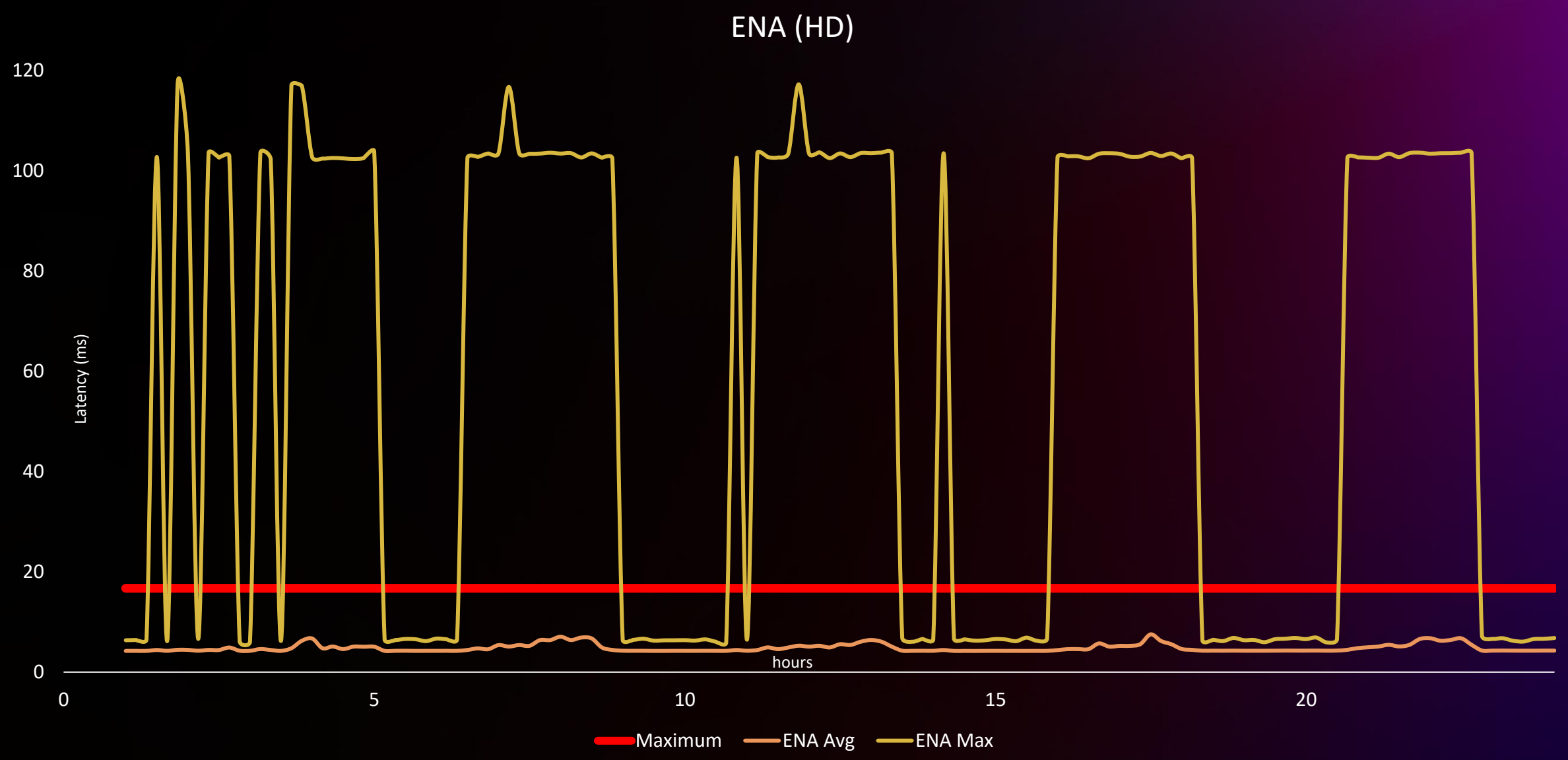
Redis – In Memory Database Reads

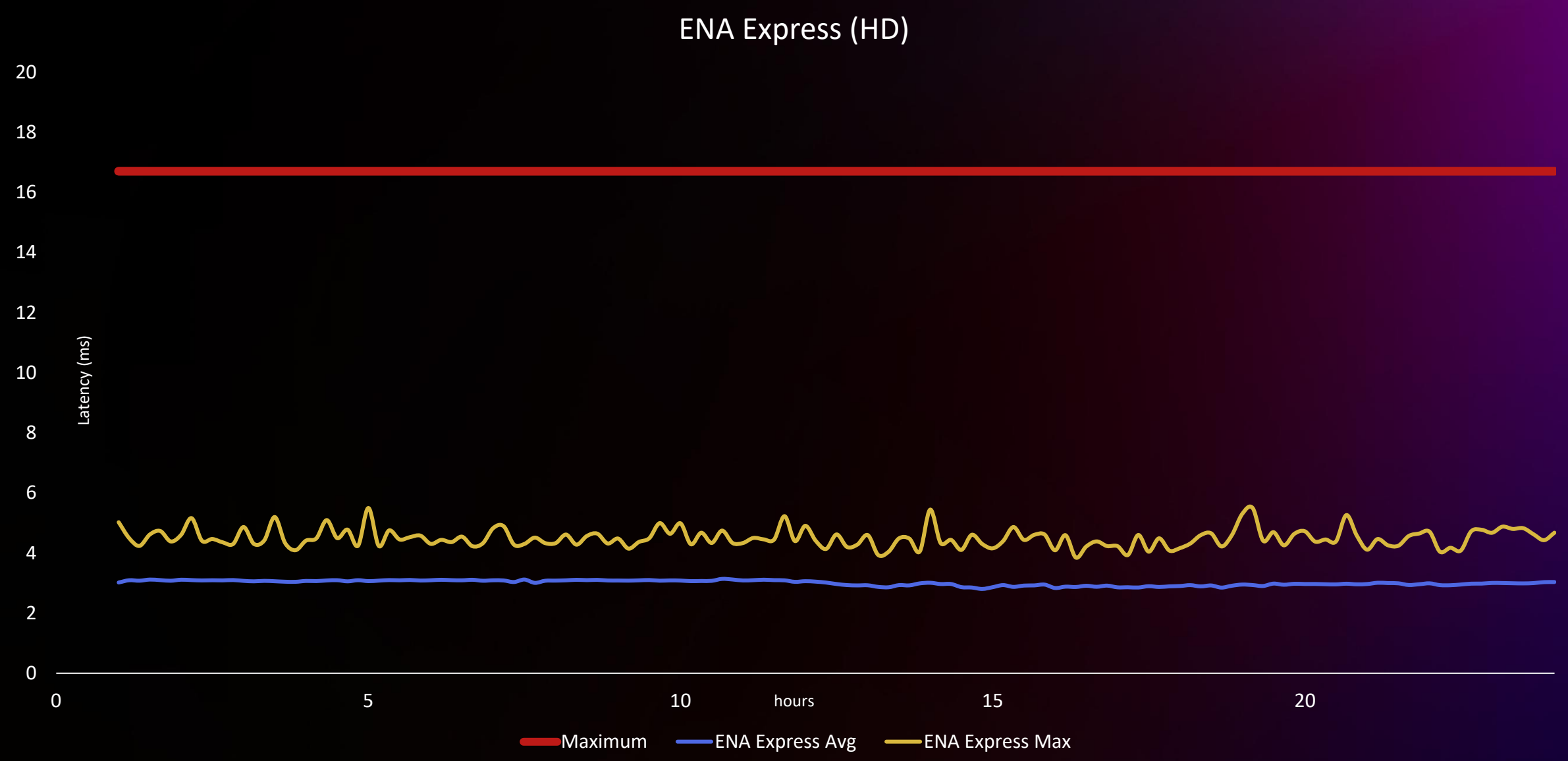


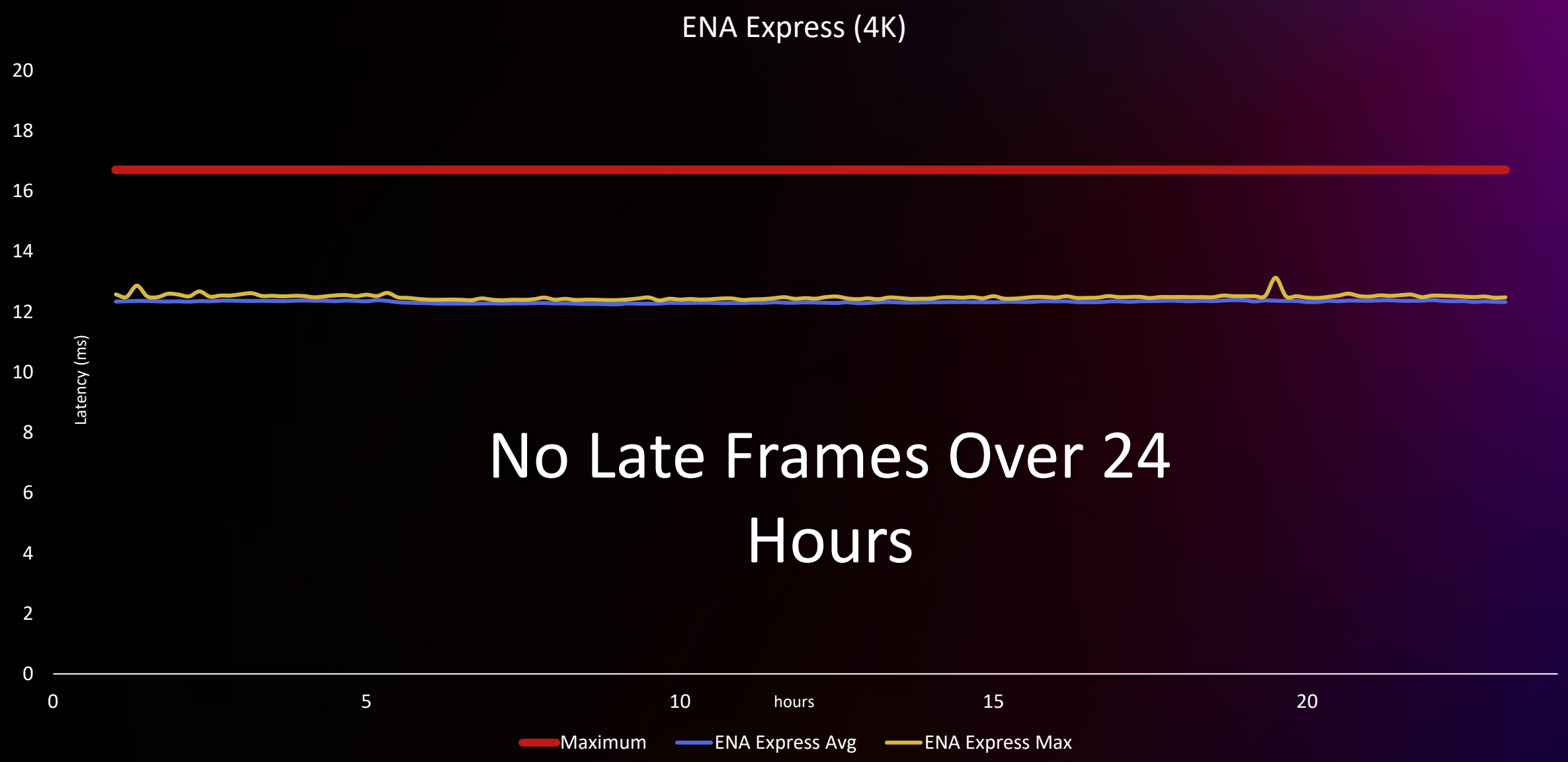
Live Video Encoding

- High Definition (HD) Flows
  - 3 Gbps
- 4k or Ultra HD Flows
  - 12 Gbps
- 60 Frames per second
  - 1 / 60 = 16.67 ms
- Video Broadcast Requirements
  - Zero frame delays over a 24 hour period





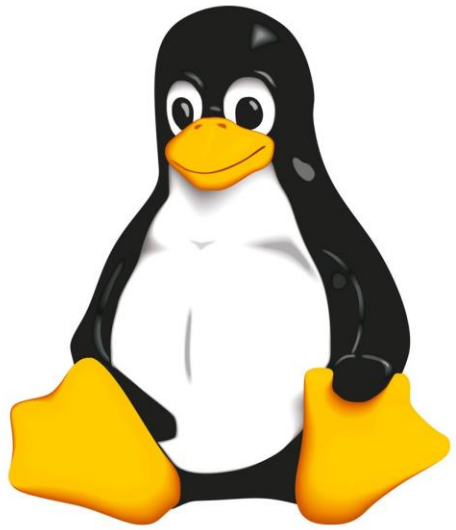




No Late Frames Over 24 Hours

How do you I know if SRD is working Properly?

- Ethtool Monitoring:
  - Counters:
    - Packets Eligible
    - Packets Transmitted
    - Packets Received
  - Resource Tracking
    - SRD Resource Utilization
  - Booleans - SRD_mode



Summary

- Network optimized instances innovation
  - New 200Gbps instances portfolio
  - Intel CPU-based instances generally available
  - Graviton-based C7gn instances in preview – Apply today!

C7gn preview sign-up

ENA Express

- Available in all commercial regions
- Available on C6gn.16xl with more instances coming soon
- Check out our Launch Blog to get started:

ENA Express Blog

Additional Sessions

CMP306-R - **Building apps to isolate & process sensitive data with AWS Nitro Enclaves**

STG307-R - **Amazon EBS: A tech deep dive**

CMP407 **- Elastic Fabric Adapter advanced topics for AI/ML and HPC**

CMP223-L - **Compute innovation to enable any application in the cloud**

NET211-L - **Leaping ahead: The power of cloud network innovation**

Thank you!

Please complete the session survey in the **mobile app**