



Whisper:
**A Practical Application for
AI in Livestreaming**

NAB 2024

Company Overview

2015

Company founded to develop next-gen VPU



Vancouver, HQ R&D
Toronto, R&D

160+

Senior Engineers from top-tier IC design companies

20

Average years experience for our architects



More info at netint.com/go

NETINT
2024 Winner

TECH
EMMYS

Design & Deployment
of Efficient Hardware Video
Accelerators for Cloud



Video Processing Units have a single purpose: **Process video.**



Total Transcoding Costs

Our customer executed a remarkable \$8.6 MIL reduction (-86%) in total costs after converting their high-volume CDN streaming platform from CPU-based to ASIC-based Smart VPUs.



More info at netint.com/go

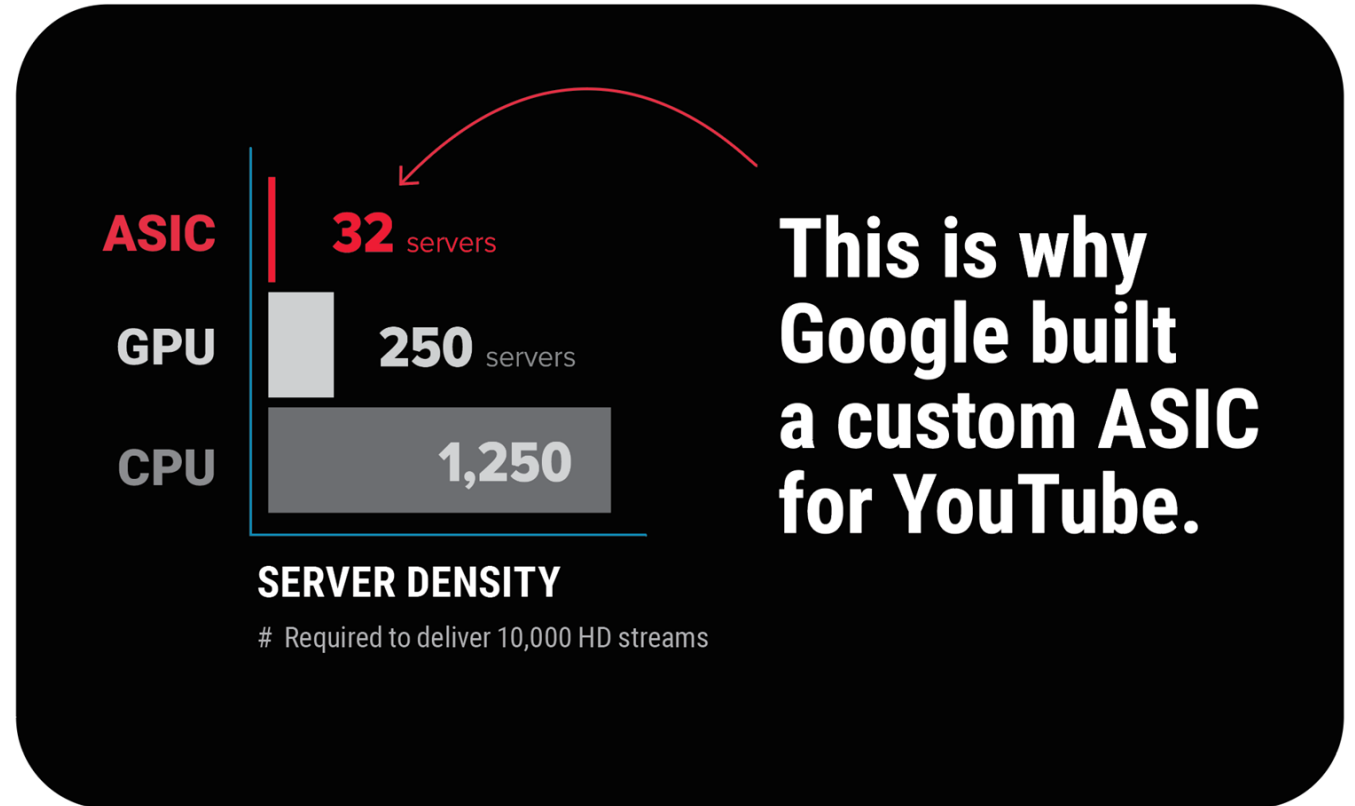


“

“There are two types of companies in the video business.

Those that are using video processing ASICs ...and those that will.”

David Ronca
Video Encoding Expert, Meta
(Formerly from Netflix)



Titans already pivoted to ASICs

For everyone else, we built one for you.

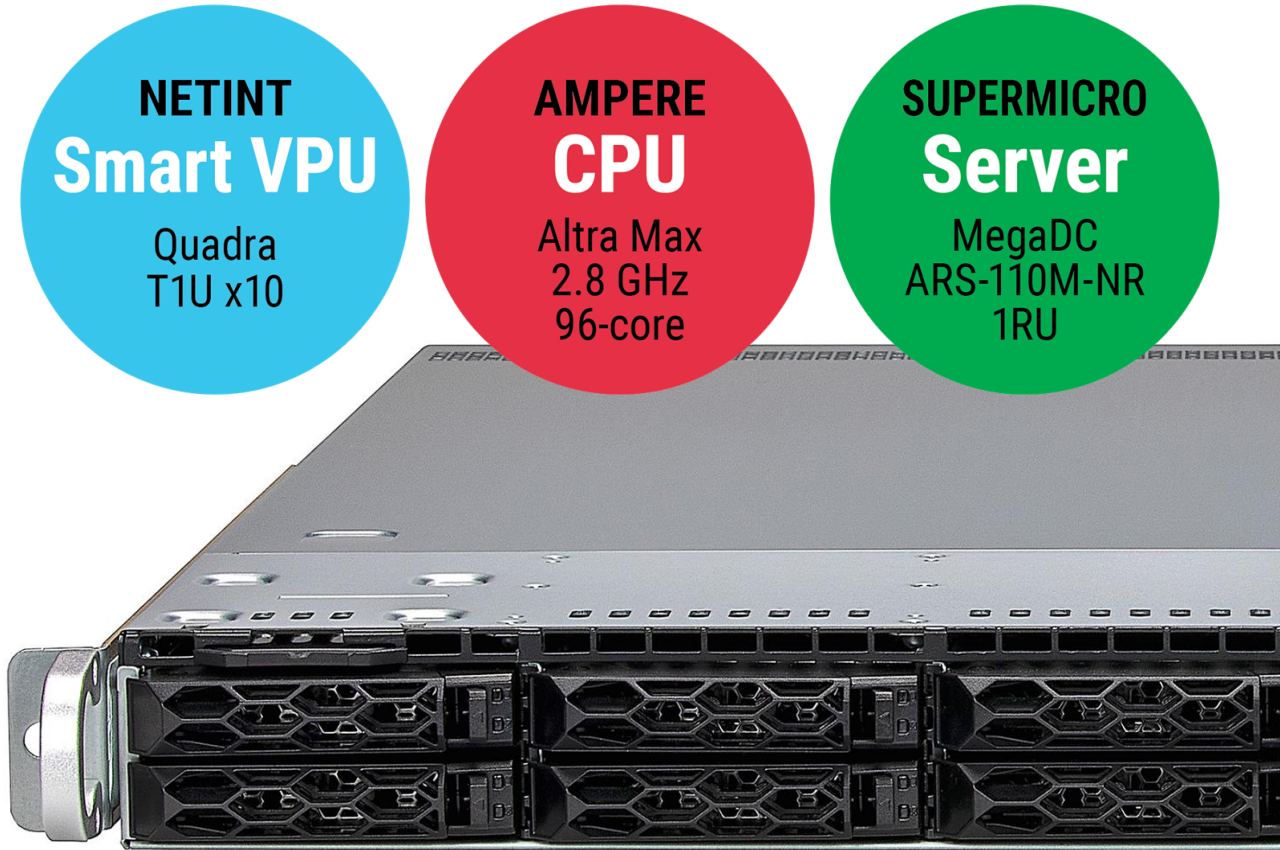
More info at netint.com/go



Quadra Video Server

Ampere Edition

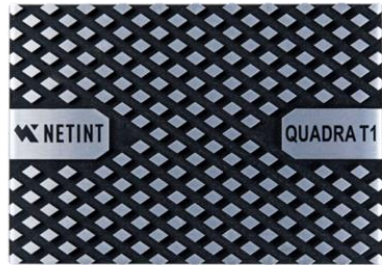
- HEVC, H.264 & AV1 encoding
- Up to 8K resolution, 10-bit HDR
- AI enhanced productivity
- 320 1080p30 live streams



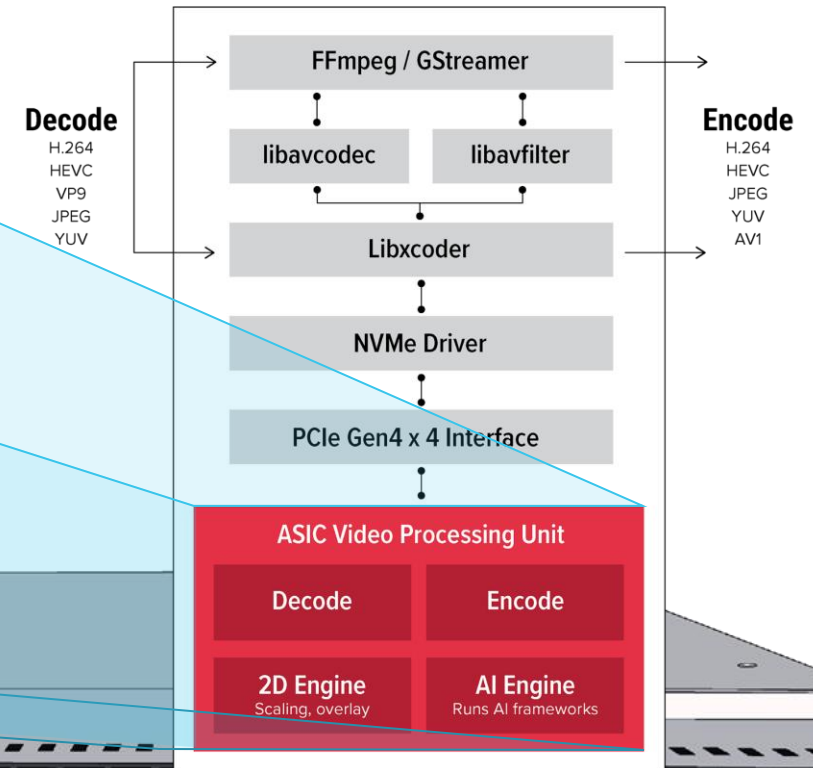
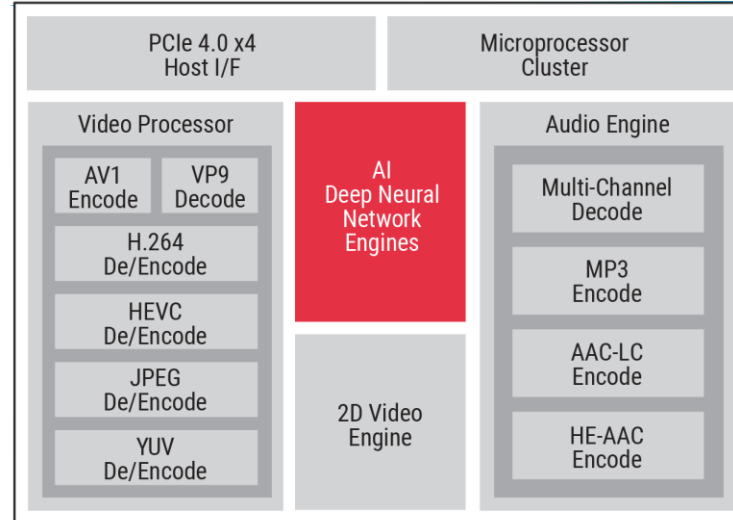
More info at netint.com/go



10x NETINT VPUs per server



Smart VPU
ASIC Video Processing Unit



More info at netint.com/go



Built for High Volume

Significant advantage-

1. Offloads video processing for deinterlacing, decoding, and advanced transcoding.
2. Quadra VPUs boost throughput by 20x while reducing costs -80%.

Enterprise application integration-

- Real-time captioning with OpenAI/Whisper
- Dynamic ABR packaging
- Streaming orchestration
- Content management
- Other AI applications

More info at netint.com/go



Ampere® Altra® Max

Arm Processor



Predictable High Performance

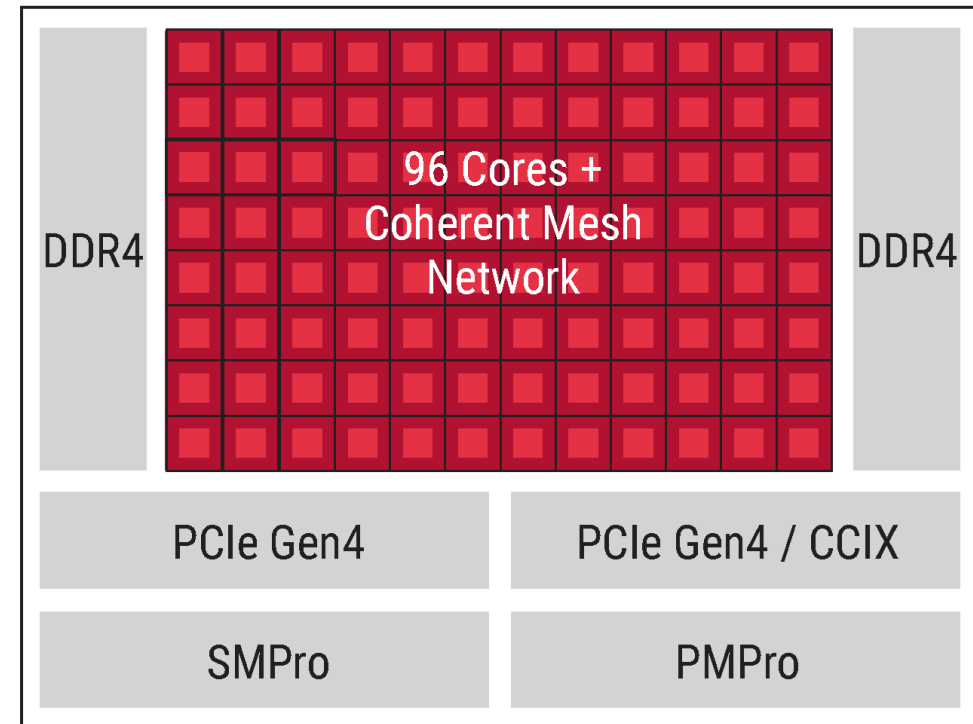
- 96 cores | 128 cores
- Coherent mesh-based interconnect
- High-memory bandwidth and density

High Scalability

- Industry leading power/core
- Cache-coherent multi-socket support
- Flexible I/O connectivity

Power Efficiency

- Advanced system, security and power management
- Monolithic die on leading 7nm process
- Leading power/core

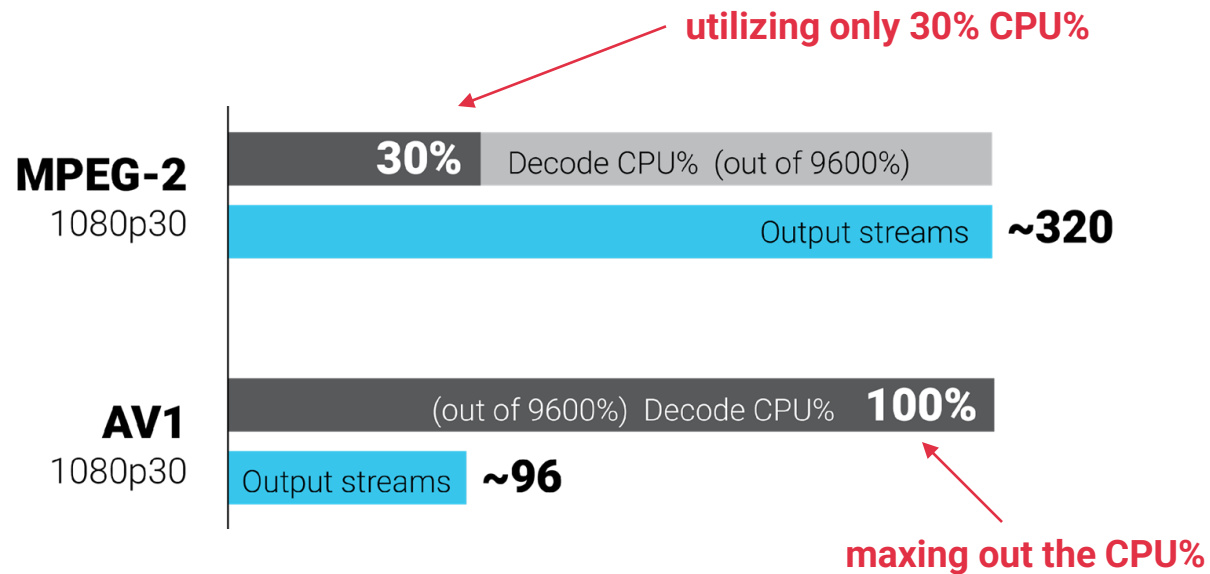


More info at netint.com/go

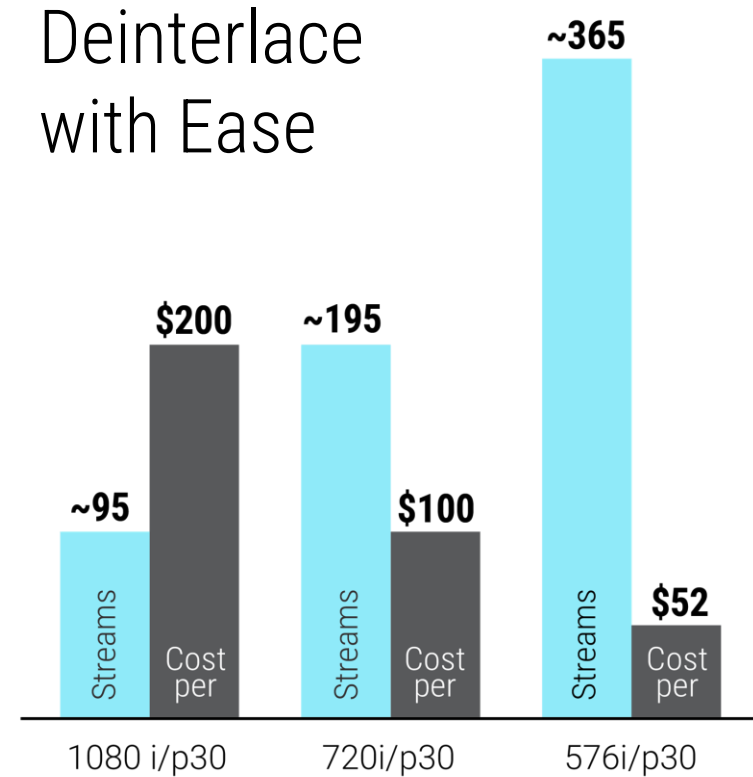


Examples: What 96 Cores Can Do For You

Decode MPEG-2 and AV1
@ 1080p



Deinterlace
with Ease



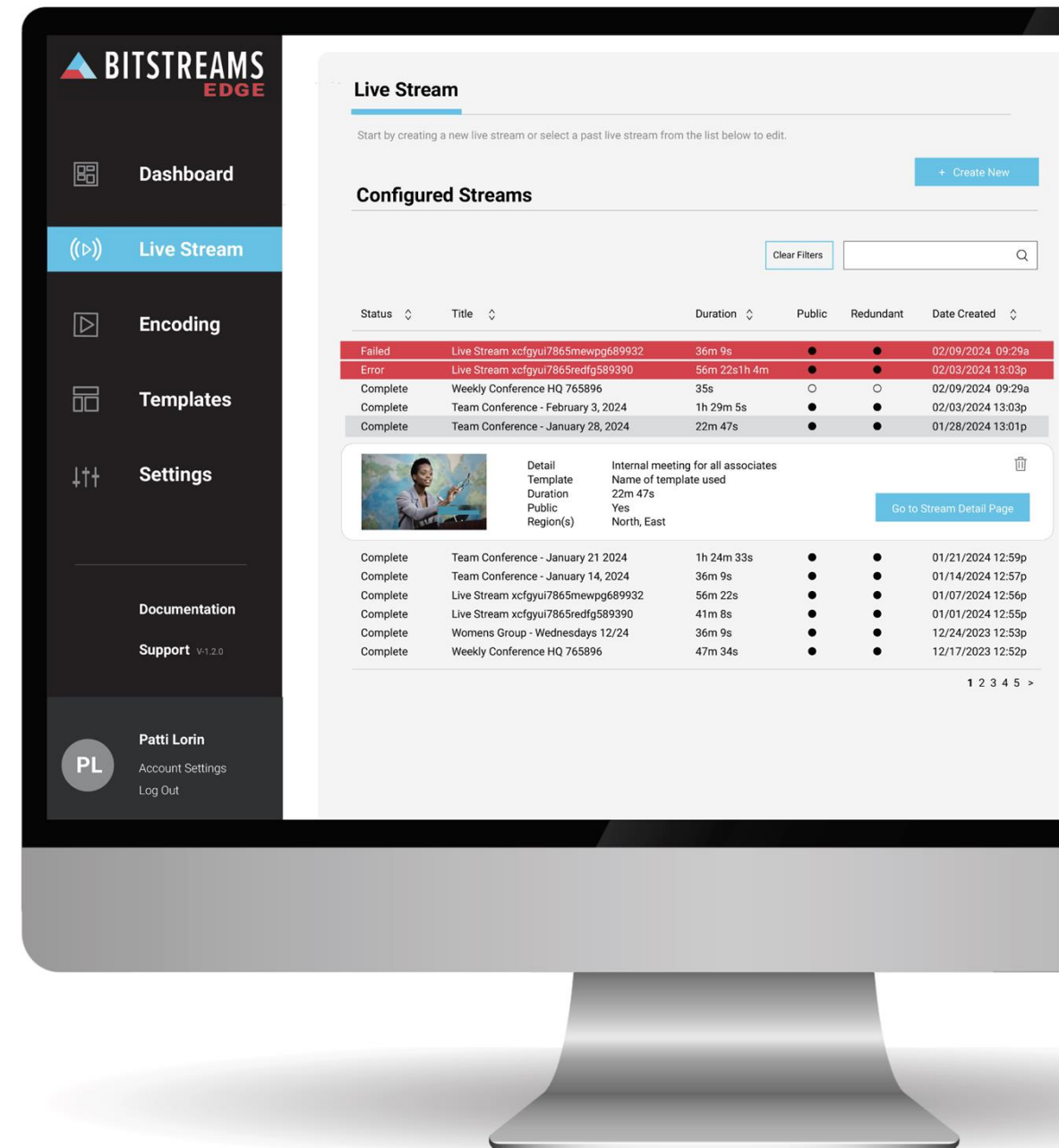
More info at netint.com/go

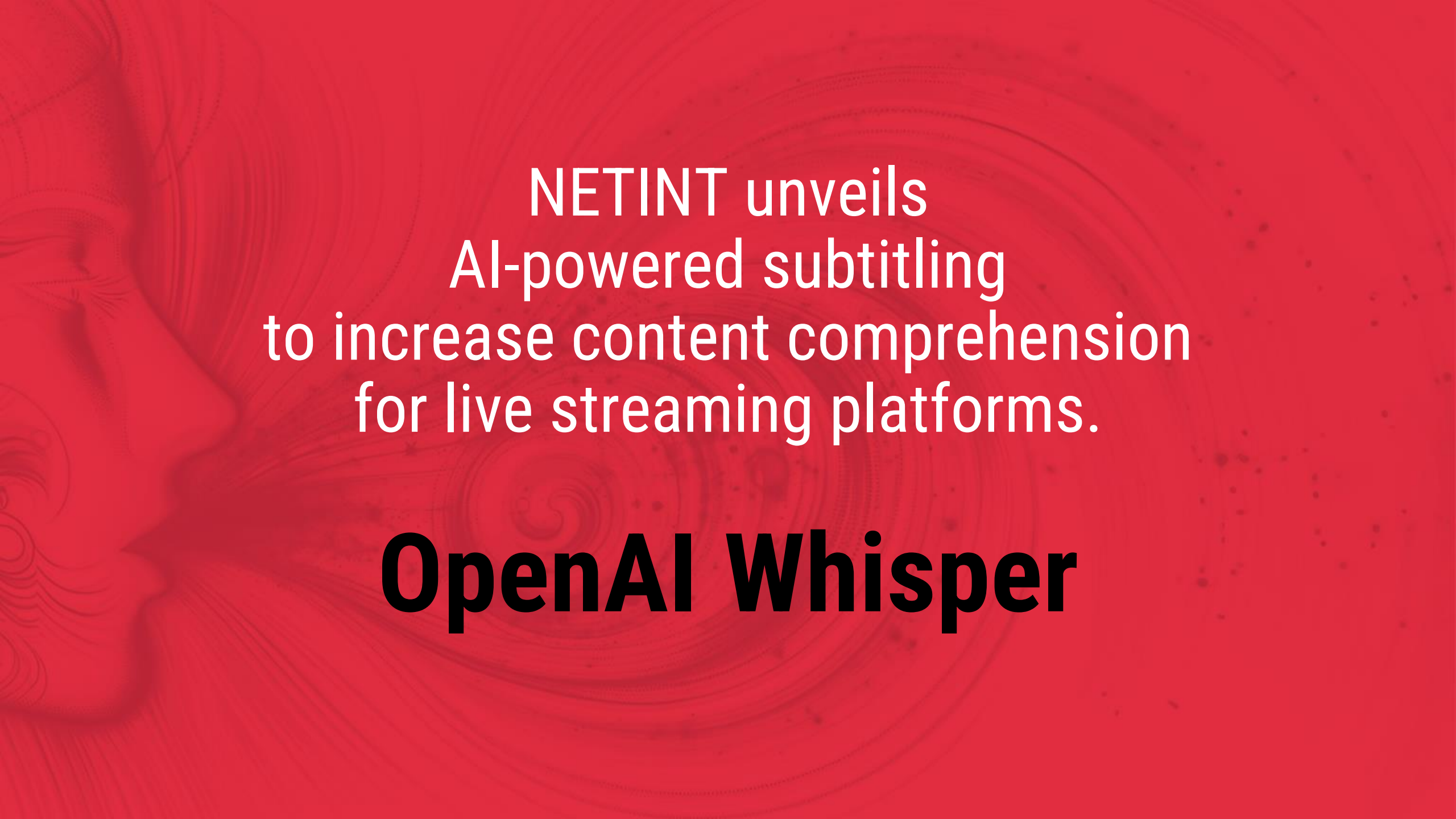
Goodbye FFmpeg. Hello, friendly face.



- Simplified to configure and distribute live streams without touching complicated FFmpeg command lines.
- Preconfigured templates help you manage and monitor your workflow through our web app or an integrated API.

More info at netint.com/go

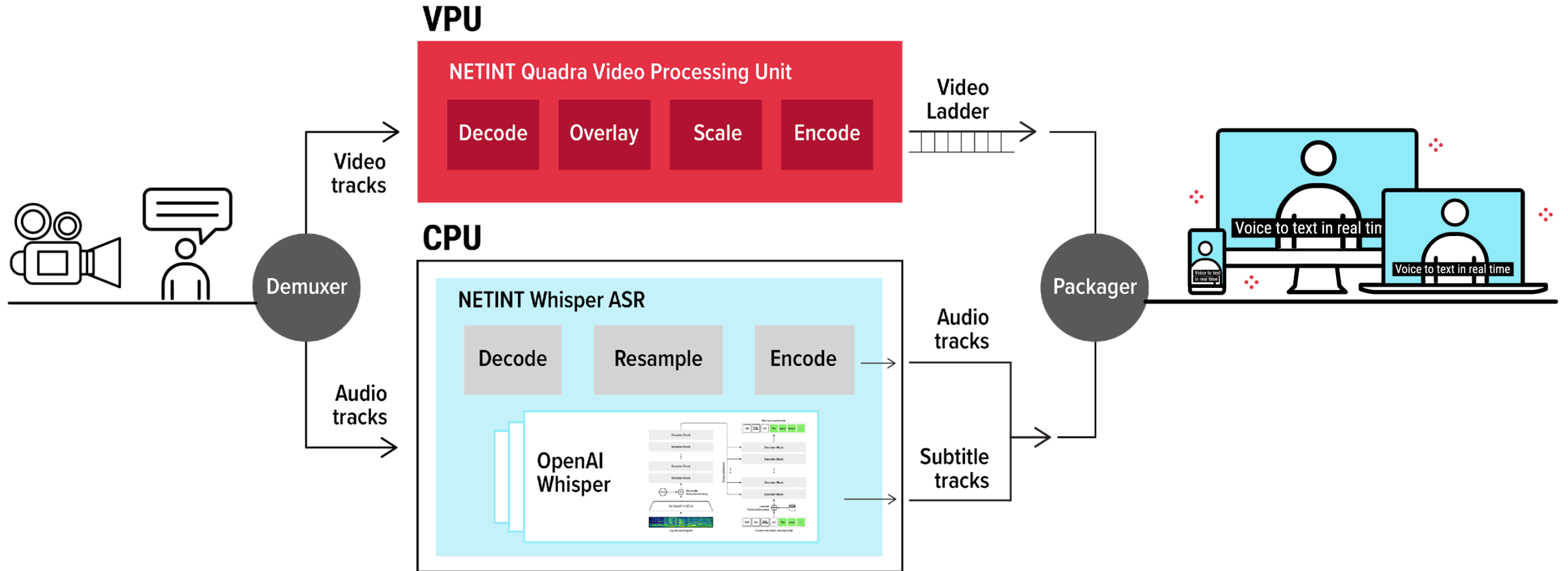




NETINT unveils
AI-powered subtitling
to increase content comprehension
for live streaming platforms.

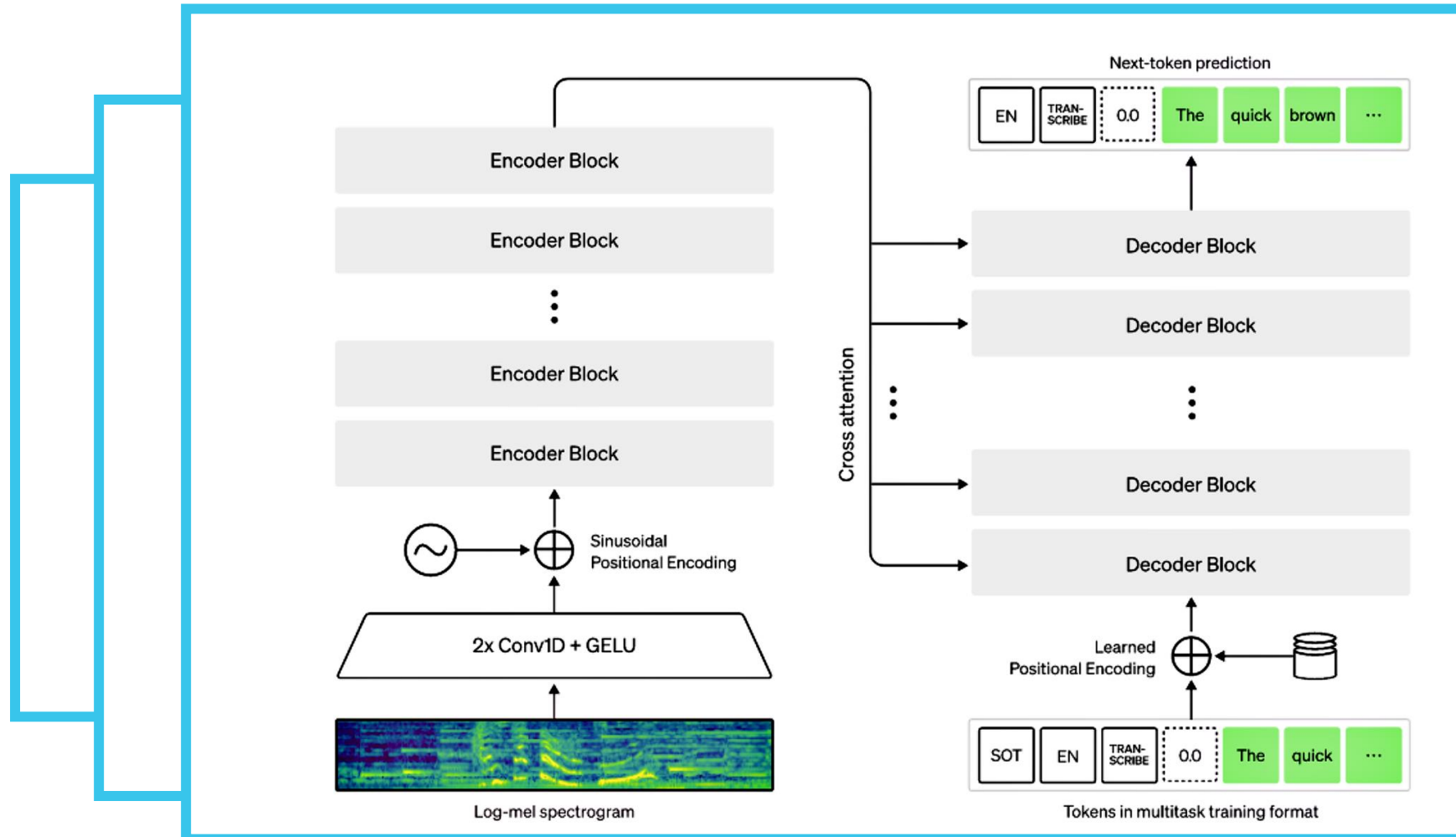
OpenAI Whisper

NETINT Whisper ASR Workflow



More info at netint.com/go

OpenAI Whisper Functional Blocks



More info at netint.com/go

Next level engagement for streaming audiences

OpenAI's whisper feature is the game changer in real-time, text subtitling.

- AI generated captions enhances accessibility and comprehension
- Immediate translations remove language, accent barriers and frustration for non-native speakers
- Live events become more inclusive and interactive engaging wider audiences with dynamic translations



More info at netint.com/go





**Whisper is ready
...and action!**

The Time is Now

Global Challenges



Ireland



Amsterdam



Singapore



Frankfurt

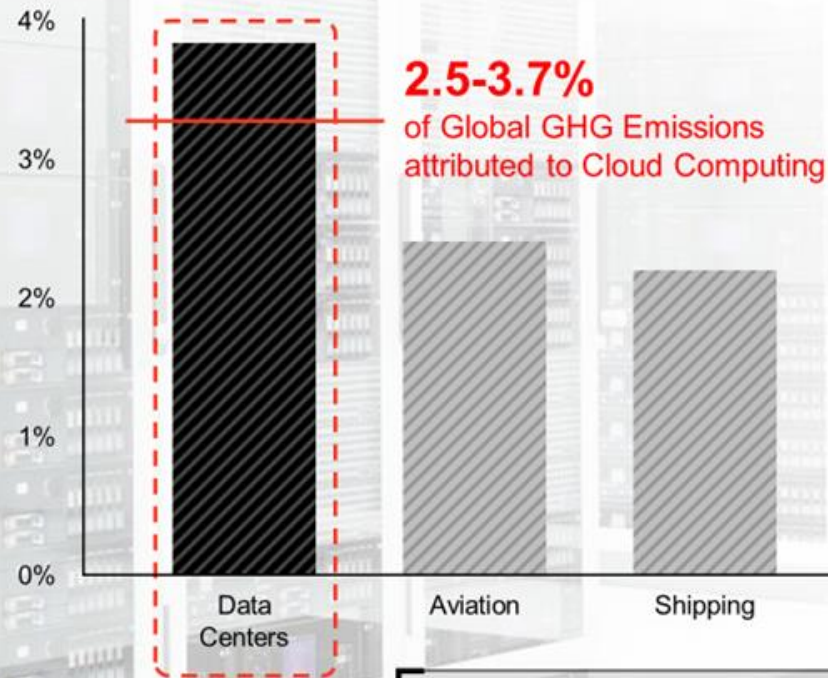


United Kingdom



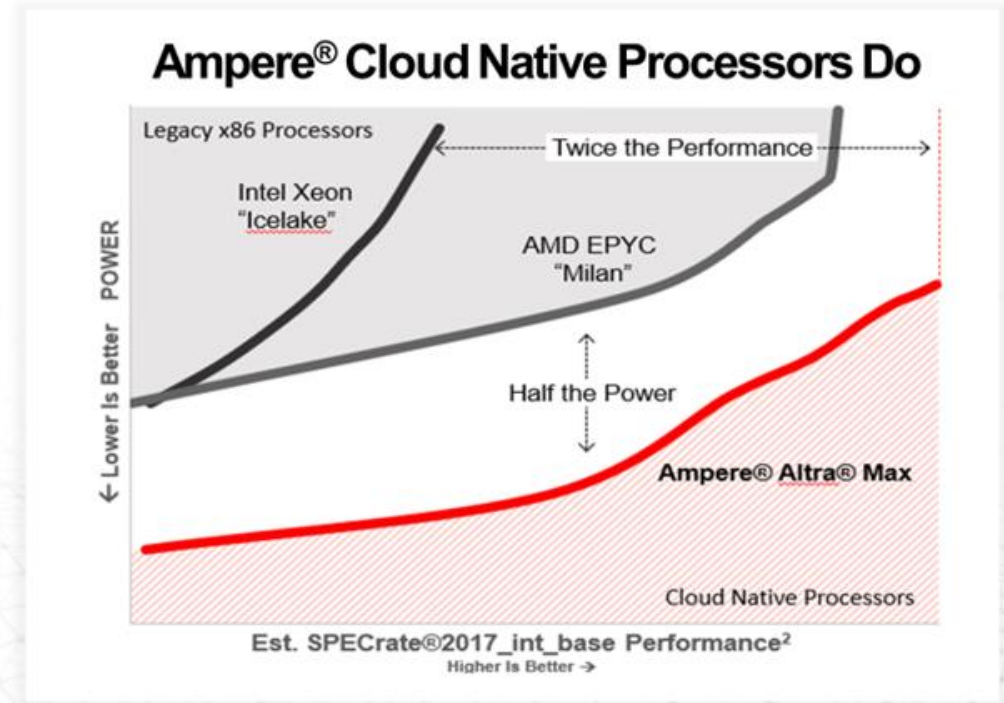
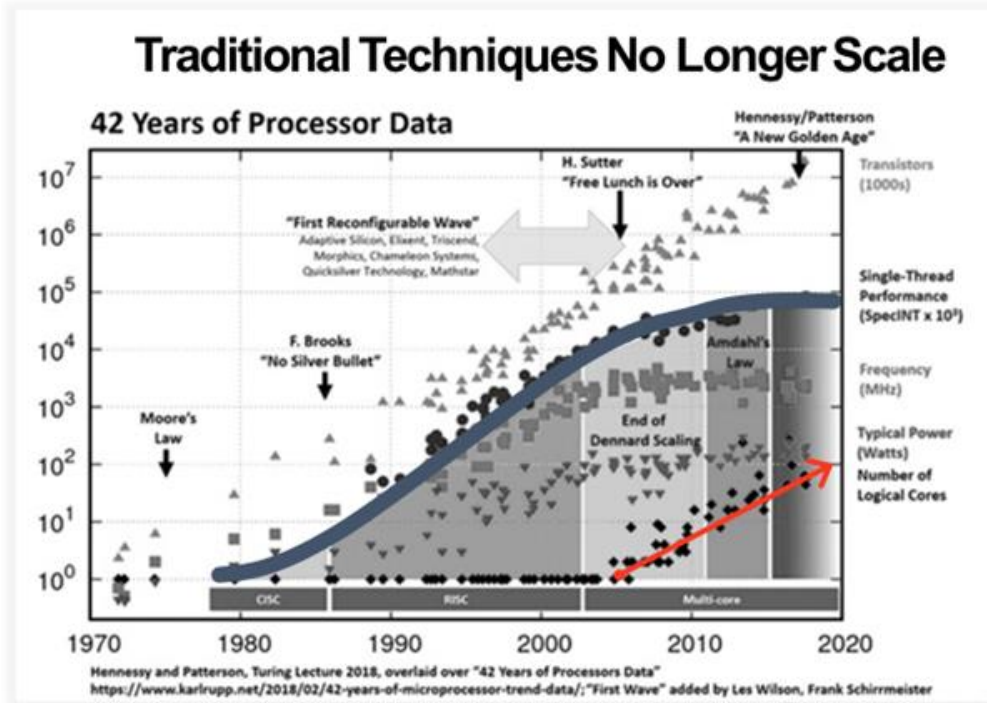
US (VA & AZ)

Energy Costs & Moratoriums
on the rise...



Global Imperative to Increase Data Center Efficiency

Ampere® Cloud Native Processors: Reducing Power by up to 50%



Turbo Frequency
Hyperthreading
Scale Up Accelerators

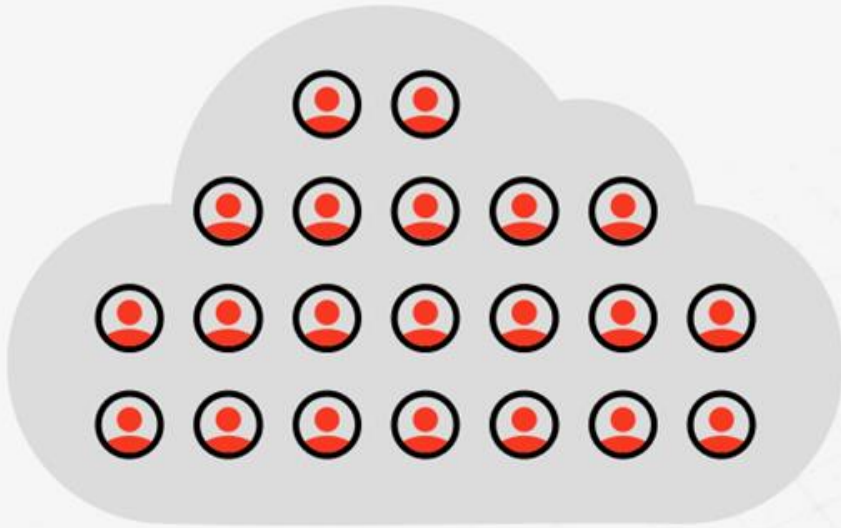
Paradigm Shift

Power Optimized, Consistent Performance
Linear Core Scaling
High Performance, General-Purpose Cores

¹ Full details available at <https://amperecomputing.com/home/efficiency-footnotes>

A Real-World Video Heavy Web Service @ Scale

Analyzing a Modest Web Workload
Under Service Level Constraint



1.3 Million Requests Every Second

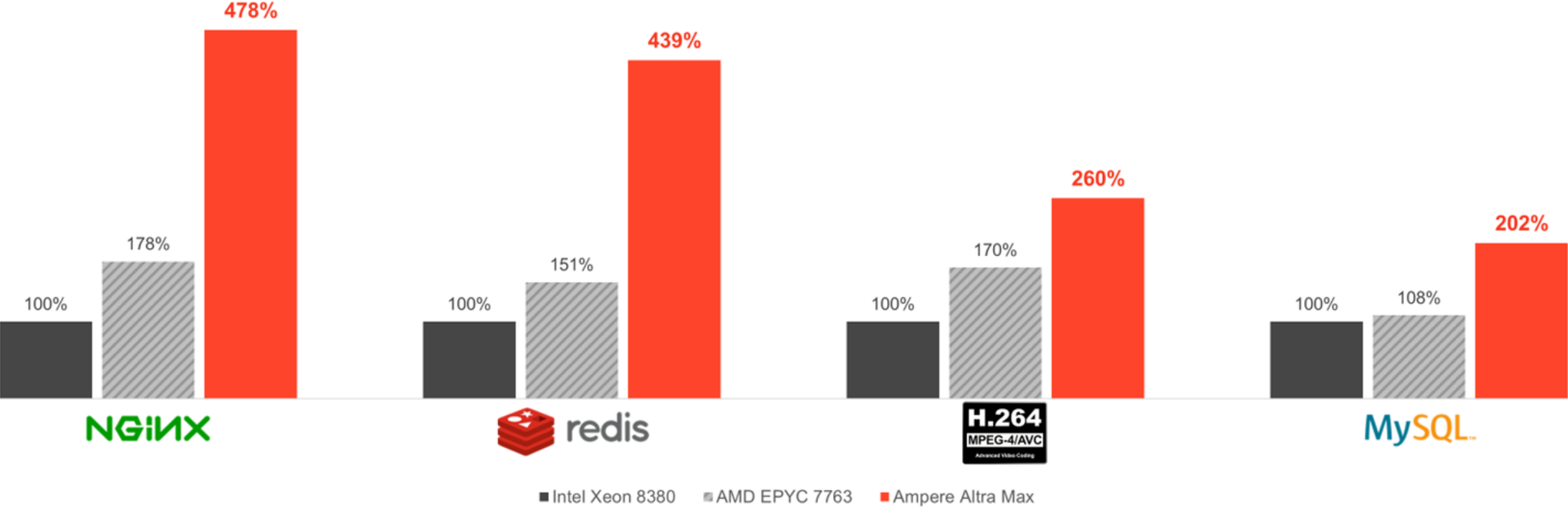


How many racks are
required?



Model Video Rich Web Service Components

Performance/W for Major Web Service Application Layers



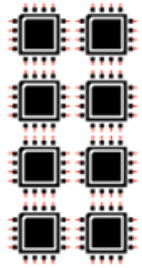
Unprecedented Energy Efficiency Demonstrates the Value

Leadership Today in Rack Efficiency

2.3x

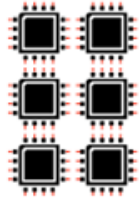
Greater performance / rack¹

81 CPUs
Required



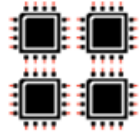
INTEL
XEON 3rd Gen

49 CPUs
Required



AMD
EPYC 3rd Gen

35 CPUs
Required



AMPERE
Altra[®] Max

1.3 Million Requests Every Second

Use

2.8x

Less Power

34.9kW
Used



INTEL
XEON 3rd Gen

22.7kW
Used



AMD
EPYC 3rd Gen

12.7kW
Used



AMPERE
Altra[®] Max

1.3 Million Requests Every Second

Use

1/3

the Rack Space

3 Racks
Required



INTEL
XEON 3rd Gen

2 Racks
Required



AMD
EPYC 3rd Gen

1 Rack
Required



AMPERE
Altra[®] Max

1.3 Million Requests Every Second

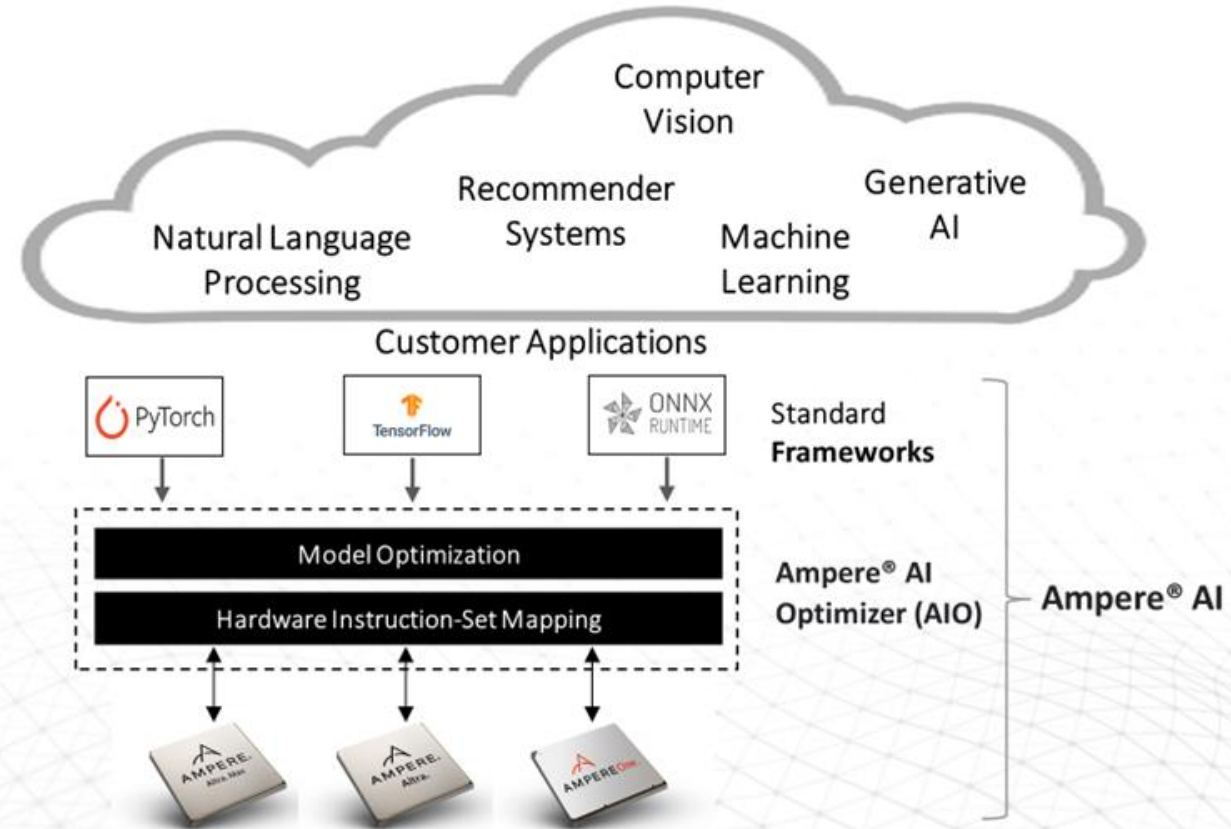
Save Money. Save Power. Save Space. Do More with Less.

¹ Full details available at <https://amperecomputing.com/home/efficiency-footnotes>

The Ampere GPU-Free AI Inference Class of Server

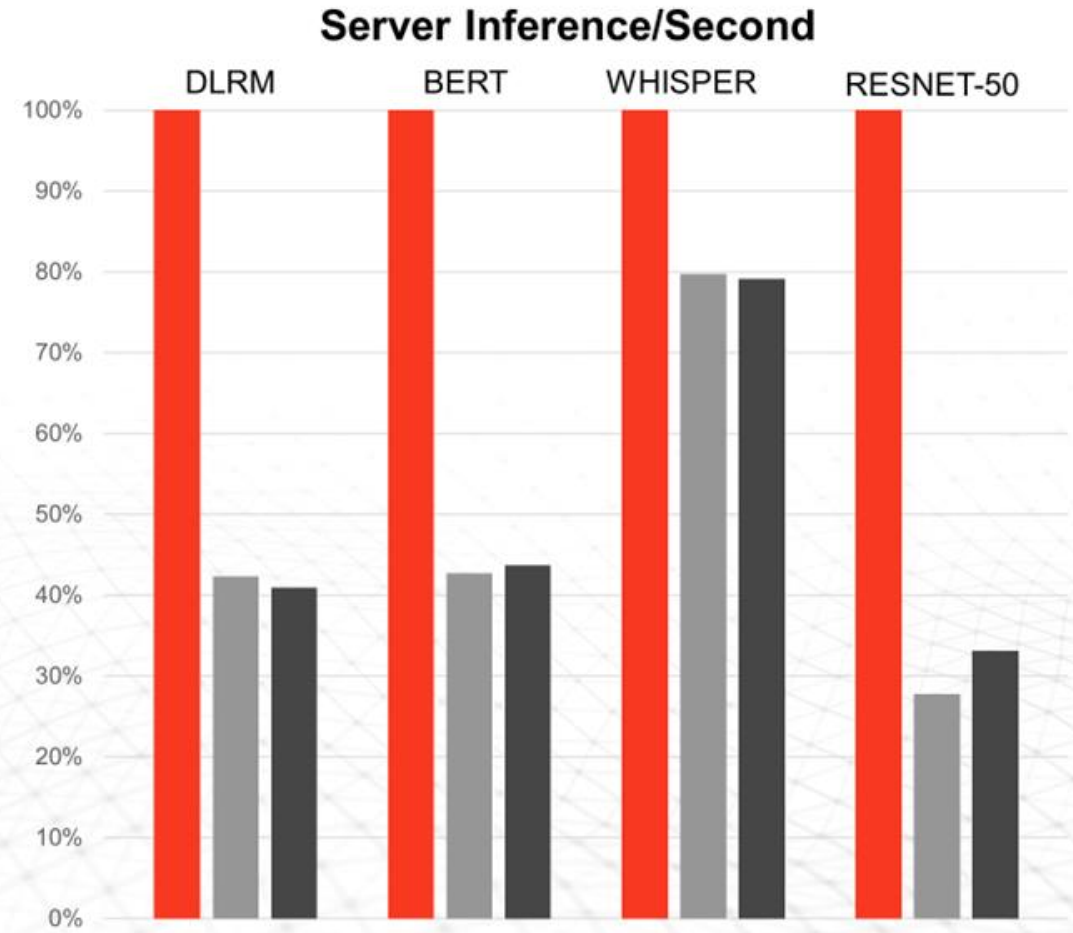
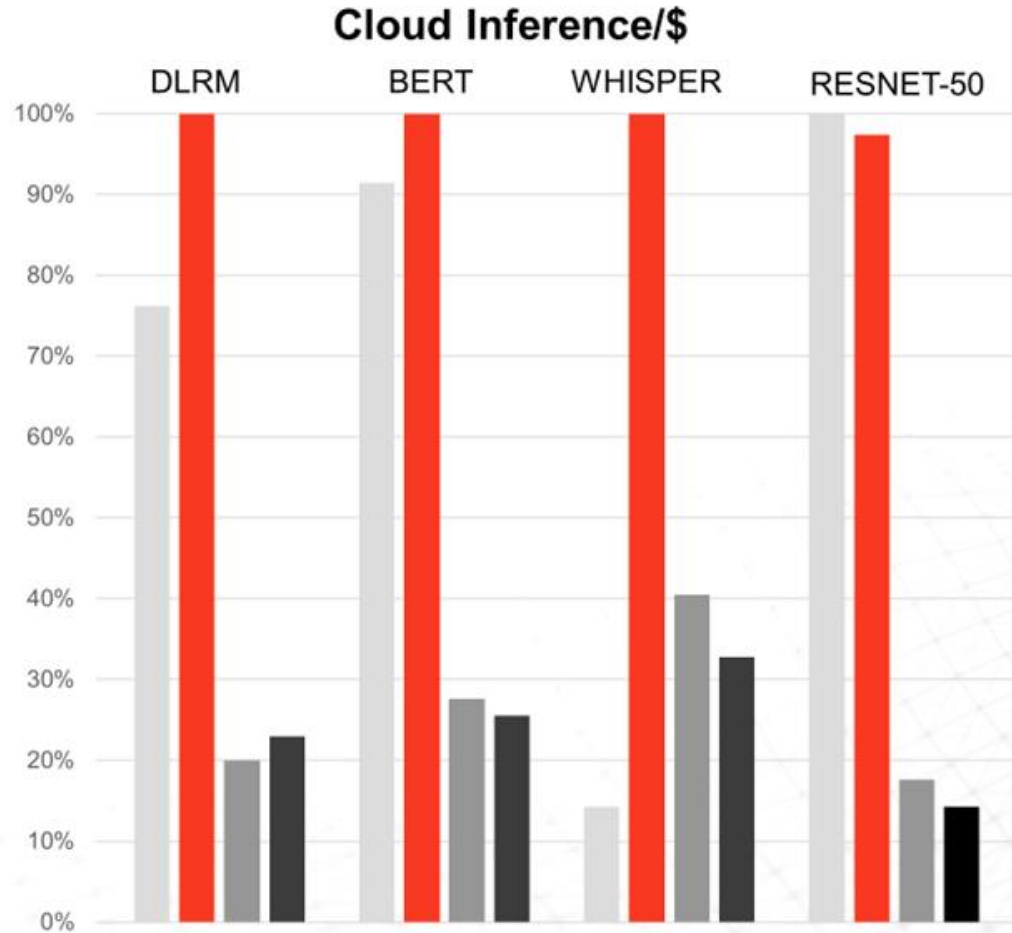


Optimized for AI Inference



***Open Frameworks, Acceleration,
Disruptive Performance***

Ampere® AI: GPU Free Leadership Inference Performance



- OCI A1 w/Ampere® Altra®
- AWS G5.16Large w/Nvidia A10 GPU
- AWS M6i.16xlarge w/Intel Xeon Ice Lake
- AWS M7g.16xlarge w/AWS Graviton 3

- Ampere® Altra® Max M128-30
- Intel Xeon 8380
- AMD EPYC 7763

Thank You



The New Quadra Video
Server

Ampere Edition

NETINT
Smart VPU

Quadra
T1U x10

AMPERE
CPU

Altra Max
2.8 GHz
96-core

SUPERMICRO
Server

MegaDC
ARS-110M-NR
1RU

