NETINT | AMPERE®

Quadra Video Server
Ampere® Edition

# The problems with video transcoding started way back:

Ultimately, the pandemic isolated everyone forcing the need to stream:

- Video content
- Social media
- Video conferencing
- Remote working
- Education
- E-commerce shopping
- Gaming
- Health and wellness

Over time, this schema become inefficient and inflated both CAPEX and OPEX.
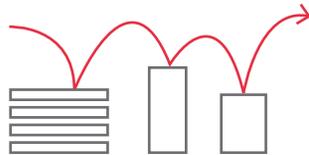
Transcoding functions

historically, have been siloed in standalone servers with ABR packaging and other production tasks performed elsewhere.

In a rush to meet demand, companies pushed their platforms to the lowest hanging clouds.
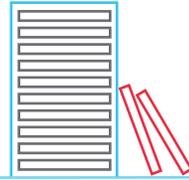
Today, most data centers are full of bulky and power hungry hardware.

The costs are unsustainable.

Now what?

The streaming industry has reached a maturity that no longer tolerates gross power consumption and profit erosion previous accepted with CPU-based video processing. This inflection point is forcing video streaming providers to do more with less in order to remain competitive in this explosive-growth industry.

Re-Introducing ASIC technology

ASIC technology has been at the forefront of innovation since traditional TV broadcasting. Here's how the renaissance of ASICs have revolutionized today's video streaming workflow:

- Efficient encoding and decoding
- Ultra-low power consumption
- Reduced latency in high-quality real-time
- CDN optimization in acceleration and scalability
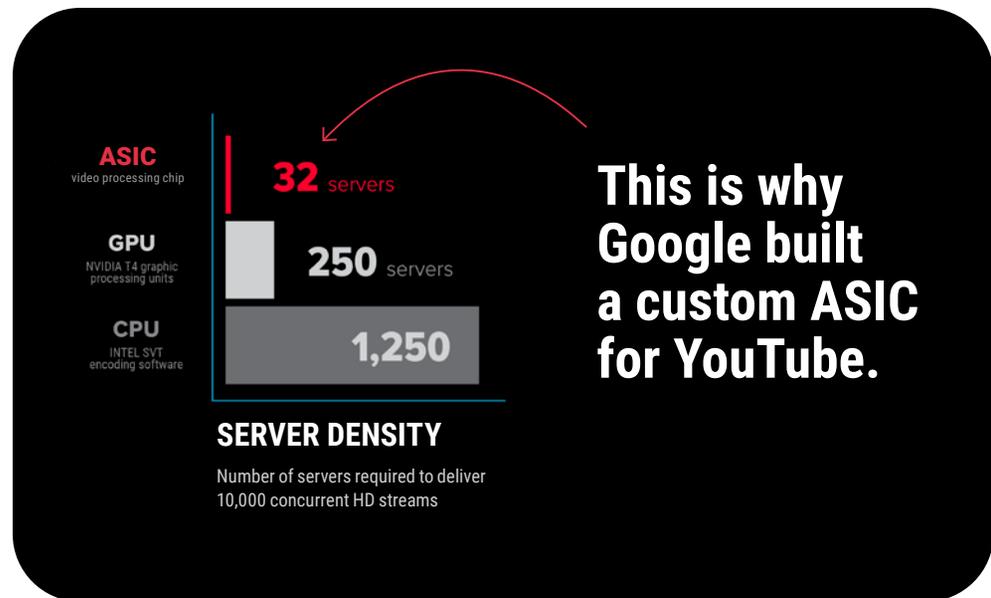- Robust security and user privacy

Industry migration is underway to capitalize on ASICs.

> "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)

ASIC
video processing chip

32 servers

GPU
NVIDIA T4 graphic
processing units

250 servers

CPU
INTEL SVT
encoding software

1,250

SERVER DENSITY

Number of servers required to deliver
10,000 concurrent HD streams

This is why Google built a custom ASIC for YouTube.

Industry titans already pivoted into ASICs

Both Google and Meta built proprietary transcoding chips to reduce costs, reduce footprint and increase density for their streaming platforms only.

For everyone else, we built one for you.

We built one better:
Smart VPUs with AI

We call ours Video Processing Units because they only process video. They're smart because the added AI will supercharge your productivity 10X. Lastly, our VPUs are based on ASIC technology and available worldwide.

NETINT Quadra Smart VPU
ASIC chip, 2nd Gen G5

Built for Speed
Tasking VPUs with only video transcoding, they process, scale, and overlay streams faster than any other transcoder.

Built for Volume
VPUs shift substantial computational loads to the cloud by specializing in a singular processing task so it pushes 10X more streams to the cloud.

Eats Less Power
VPUs have a tiny appetite for power, drawing <7W, and uses 80% less energy than CPU-processing.

Boosted with AI
AI inferencing accelerates accuracy and productivity with key object filtering and facial recognition.

NETINT
2024 Winner

TECH EMMYS®

Design & Deployment of Efficient Hardware Video Accelerators for Cloud

We feel it's worth mentioning, co-winners for this same category were Google, Meta and AMD.

https://theemmys.tv/tech-75th-award-recipients/

Finally, a server with monster throughput has a tiny appetite for power

The Quadra Video Server Ampere Edition

Industry leaders united to consolidate your video transcoding process, accelerate core functionality and integrate multiple production processes into this high-performance server.

- **HEVC, H.264 & AV1 encoding**
- **HEVC, H.264, & VP9 decoding**
- **Up to 8K resolution, 10-bit HDR**
- **AI enhanced productivity**

NETINT
Smart VPU
Quadra T1Us

AMPERE
CPU
2.8 GHz 96-core Altra Max M96-28

SUPERMICRO
Server
MegaDC ARS-110M-NR 1RU

Built for pushing streams to the cloud

320

simultaneous live 1080p30 encoding
sessions at ultra-low latency
in a compact 1RU server

When Quadra Video Processing Units (VPUs) handle complex video encoding and decoding tasks, it leads to significant advantages. Firstly, this approach offloads the workload from the Ampere 96-core CPU, enabling it to focus on other operations such as deinterlacing, and AV1/MPEG-2 decoding. Secondly, leveraging Quadra VPUs for these tasks results in a synergy that boosts throughput by tenfold while reducing costs -80%. This dual benefit represents a highly efficient and cost-effective solution.

Enterprise application integration

The Quadra Video Server Ampere Edition empowers video engineers to run additional publishing-related applications such as:

- Real-time captioning and translating with OpenAI/Whisper
- Dynamic ABR packaging
- Streaming orchestration and
- Content management

Ampere® Altra® Max
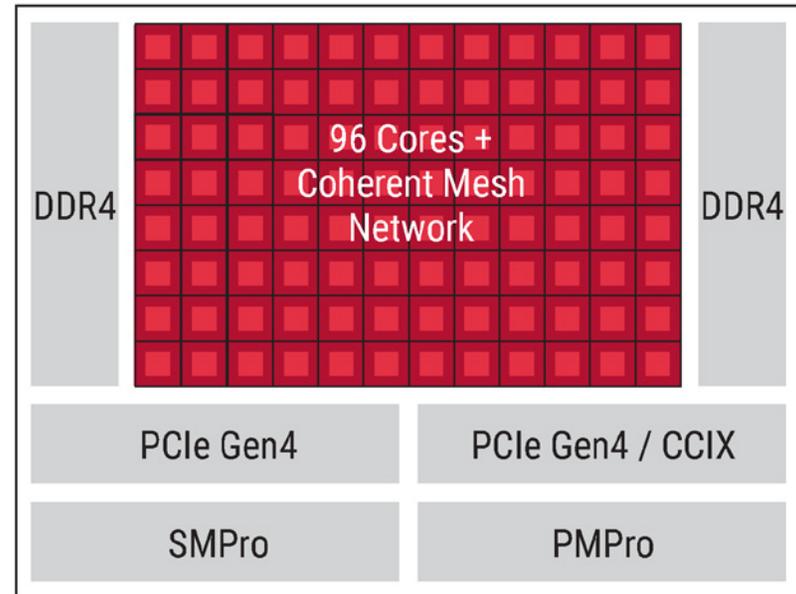## Processor

Predictable High Performance

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

High Scalability

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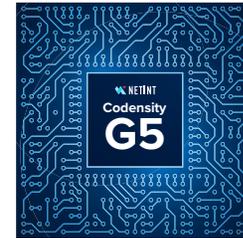- Cache-coherent multi-socket support
- Flexible I/O connectivity

Power Efficiency

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

NETINT Codensity G5
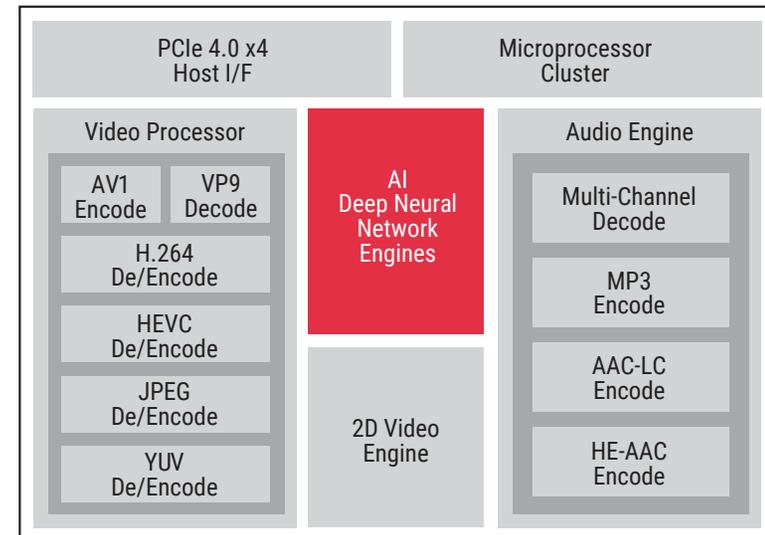
Smart VPU with AI

High Performance Encoding

- 8K UHD Video Encoding
- On-chip AV1, H.264 and HEVC
- Up to 60% reduction in bitrate

Flexible Architecture

- Programmable micro-processor enables firmware optimization
- Pipeline processing improves performance and increases video quality

AI Processing Engines

- Two Deep Neural Network DNN engines onboard at 18 TOPS
- Enables object detection, classification and segmentation for image quality improvement and content-adaptive rate control
- Advanced performance and seamless integration for ROI (region-of-interest) encoding and background replacement

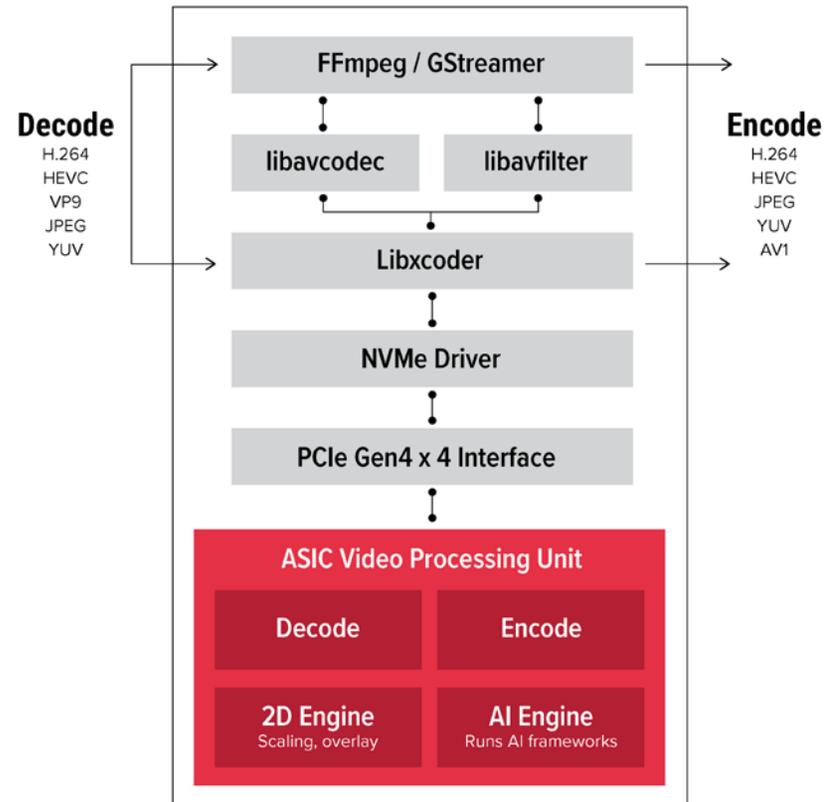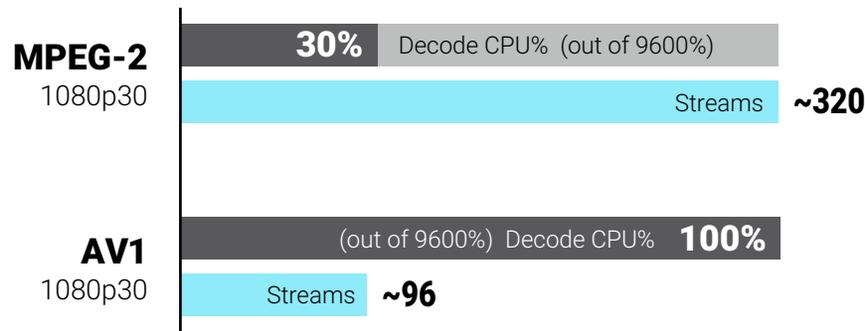| PCIe 4.0 x4 Host I/F		Microprocessor Cluster
Video Processor	**AI Deep Neural Network Engines**	**Audio Engine**
AV1 Encode / VP9 Decode		Multi-Channel Decode
H.264 De/Encode		MP3 Encode
HEVC De/Encode		AAC-LC Encode
JPEG De/Encode	2D Video Engine	HE-AAC Encode
YUV De/Encode		

Simple Integration

Open-source suite of processing tools

Many video processing and transcoding applications developers use FFmpeg and GStreamer, two open-source software libraries offering a vast suite of video processing functions. The Quadra video server includes highly efficient FFmpeg and GStreamer compatible SDKs, allowing operators to apply an FFmpeg or GStreamer patch to complete the integration.

The libavcodec patch on the host server functions between the Quadra NVMe interface and the FFmpeg and GStreamer software layers simplifying integration and enabling fast and efficient performance and capacity upgrades.

What 96 cores can do for you

MPEG-2
1080p30

30% Decode CPU% (out of 9600%)

Streams ~320

AV1
1080p30

(out of 9600%) Decode CPU% 100%

Streams ~96

~365

$200 ~195

~95 $100

$52

Streams | Cost per | Streams | Cost per | Streams | Cost per

1080 i/p30 | 720i/p30 | 576i/p30
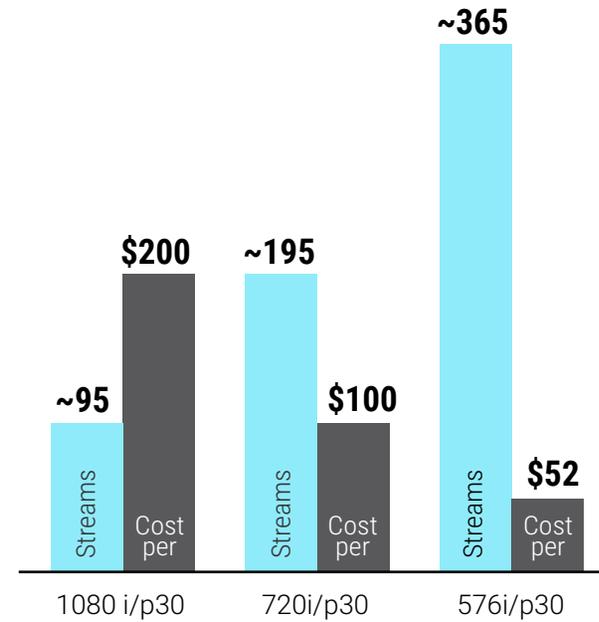
Decode MPEG-2 and AV1

At 1080p30, the Ampere CPU can decode incoming streams:

- over 320 MPEG-2 streams, utilizing only 30% CPU%
- around 100 AV1 streams, maxing out the CPU%
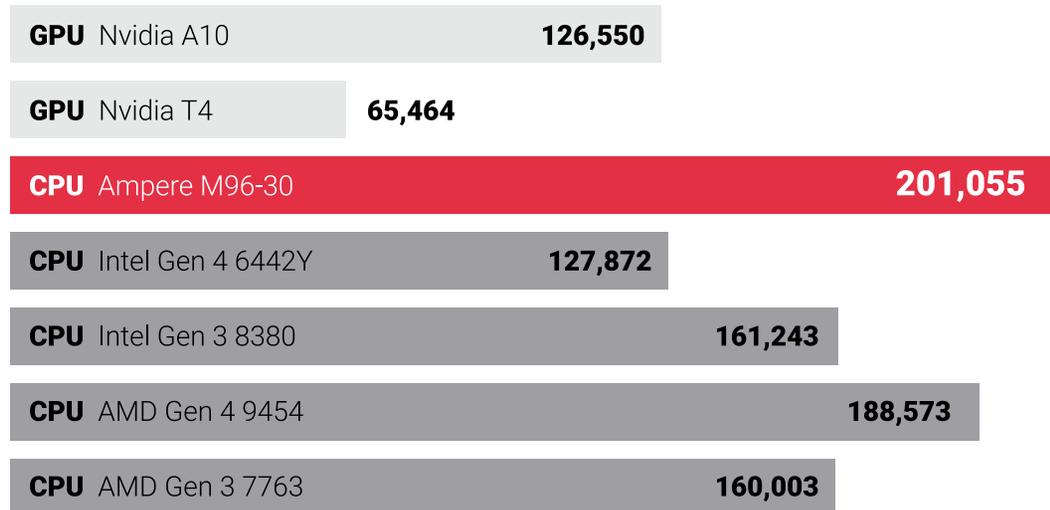
Deinterlace with Ease

The Ampere CPU is ideal for video engineers working with interlaced source content.

Add real-time captioning and translating to your live streams with AI

International accessibility with OpenAI/Whisper

The 96-core Ampere CPU is ideal for video engineers to run additional publishing-related applications like Whisper AI for live captioning and translating.

(Available Q2 2024)

		Number of inferences per second
GPU	Nvidia A10	126,550
GPU	Nvidia T4	65,464
CPU	Ampere M96-30	201,055
CPU	Intel Gen 4 6442Y	127,872
CPU	Intel Gen 3 8380	161,243
CPU	AMD Gen 4 9454	188,573
CPU	AMD Gen 3 7763	160,003

Number of inferences per second

Quadra T1U
Smart VPU

Video Processing Unit with AI | Codensity G5

Form Factor	U.2
ASIC	1x Codensity G5
Interface	PCIe 4.0 x4
Power Consumption (Typ)	17W
Usage	24/7 Operation
Operation Temperature	0 - 50°C
RoHS Compliance	European Union (EU) ROHS Compliance Directives
Product Health Monitoring	Self-Monitoring, Analysis, and Reporting Technology (SMART) commands Temperature Monitoring and Logging

Video Encoding Standards/Formats	AVC/H.264 Baseline, Main, High, High 10 HEVC/H.264 Main, Main 10 JPG YUV 420 8 bit/10 bit encoding AV1 Main
Video Decoding Standards/Formats	AVC/H.264 Baseline, Main, High, High 10 HEVC/H.265 Main, Main 10 VP9 Profile 0, 2 JPEG YUV 420 8 bit/10 bit decoding
Throughput Capacity	Up to 32x 1080p30, 8x 4Kp30, 2x 8Kp30
Audio Codecs	MP3, AAC-LC, HE-AAC
Level	1 to 6.2 Main Tier
Resolution	32 x 32 to 8192 x 5120
Scan Type	Progressive
Bitrate	64kbit/s to 700Mbit/s
Software Integration	FFmpeg SDKs, GStreamer, LibXcoder API integration

AI Deep Neural Network Engines	15 TOPS AI Assisted Encoding
Region of Interest (ROI)	ROI enables the quality of some regions to be improved at the expense of other regions
Closed Captioning	EIA CEA-708 for H.264 and HEVC encode/decode
High Dynamic Range (HDR)	HDR10, HDR10+, HLG for H.264 & HEVC encode/decode
Low Latency	Sub-frame latency
IDR Insert	Forced IDR frame inserts at any location
Flexible GOP Structure	8 presets plus customizable GOP structure
Video 2D Processing Engine	Crop & Padding/Scaling/Overlay/YUV & RGB Conversion

Quadra Video Server

Ampere Edition

Supermicro | MegaDC ARS-110M-NR

CPU Options	Ampere Altra Max M96-28, 96-cores
Operating System	Ubuntu 22.04.3 LTS
Memory	256GB of DDR4-3200 RDIMM
Storage	400GB M.2 SSD
NVMe Support	10x
PCIe Expansion	Three PCIe 4.0 x16 LP slots, one PCIe 4.0 x16 AIOM slot
Network Options	1 RJ45 Dedicated IPMI LAN port 2x 25Gb SFP28 Ethernet LAN Ports
Power Consumption	~500W
Power Supply	800W Redundant Platinum Level power supplies
Transcoders	10x NETINT Quadra T1Us
Encoding Capacity	Up to 20x 8Kp30, 80 4Kp30 or 320x 1080p30
Codec Support	H.264 - Encode/Decode
	HEVC - Encode/Decode
	JPG - Encode/Decode
	VP9 - Decode
	AV1 - Encode
Software Integration	FFmpeg, GStreamer, NETINT SDK

Physical Dimensions	W: 54.5 mm x H: 40.25 mm x D: 220 mm
Rack Size	1U
Weight	39 lbs (17.69 kg) *(fully loaded with 10 T1U VPUs)*
Environmental	50 degrees F to 95 degrees F Operating Temperature, 8% to 90% Operating Relative Humidity
Power Inputs	750W: 100-127Vac / 50-60Hz
	800W: 200-240Vac / 50-60Hz
	800W: 230-240Vdc / 50-60Hz
Certifications	RoHS Compliant, UL Approved

NETINT
HYPERSCALE PROFITABLY

For more information on NETINT encoding solutions, contact us.

sales@netint.com
netint.com