



Quadra Video Server

Smart VPU | ASIC Video Processing Unit powered by AI

Executive Summary

We created a new category of Smart VPUs (ASIC video processing units) to disrupt the previous limitations of video encoding for streaming delivery platforms. It is unique because it's dense, cost effective and AI-powered making it the ideal technology for platforms to future-proof their services and hyperscale profitably.

Designed as a quickstart solution for high density live video encoding, the Quadra Video Server comprises ten Quadra VPUs in a 1RU chassis that performs the equivalent work of up to 25 dedicated servers running a typical open-source FFmpeg and x264, x265, or SVT-AV1 configuration. The server delivers the lowest TCO of any solution in the market, and is a drop-in replacement for existing CPU and GPU encoding stacks.

The results are profound and transformational.

**Smart VPU's will be the engine
powering all future video
streaming experiences**



Live streaming experiences are seeing rapid adoption

Applications:

- Live events
- Interactive video
- Cloud gaming
- Real-Time video
- Virtual worlds
- 360/VR/AR



The insatiable appetite of video consumers

They want nonstop, never-ending, high-resolution, non-buffering content accessible on any device. Now.

Viewers have developed an addiction to continuous content streaming. Video delivery and entertainment experiences are shifting from file-based to live where low latency and controlling operational costs are paramount.

- Increased public cloud provider costs are stressing businesses
- Live experiences are growing in resolution, color depth, and quality expectations
- Playback is expected on every device using its full capability
- More data centers are needed to handle capacity increases

2021

Social video viewing surpassed Google search traffic

> 1 billion active monthly users on short-form video apps.


65%

Percentage of ALL internet traffic is video streaming, increasing 24% year over year.

40%

Percentage of people 18 to 24 turning to visual-based social media platforms for internet searches.





Why ASICs are needed.

Density is a dirty expensive problem

Global corporations spend 20% of their annual OPEX powering data centers.

Data centers operate 24/7, massively consume energy, and are depleting our planet's resources at an accelerated and unsustainable rate. Today, there are 8,000 data centers globally and their collective consumption is expected to double by 2025.

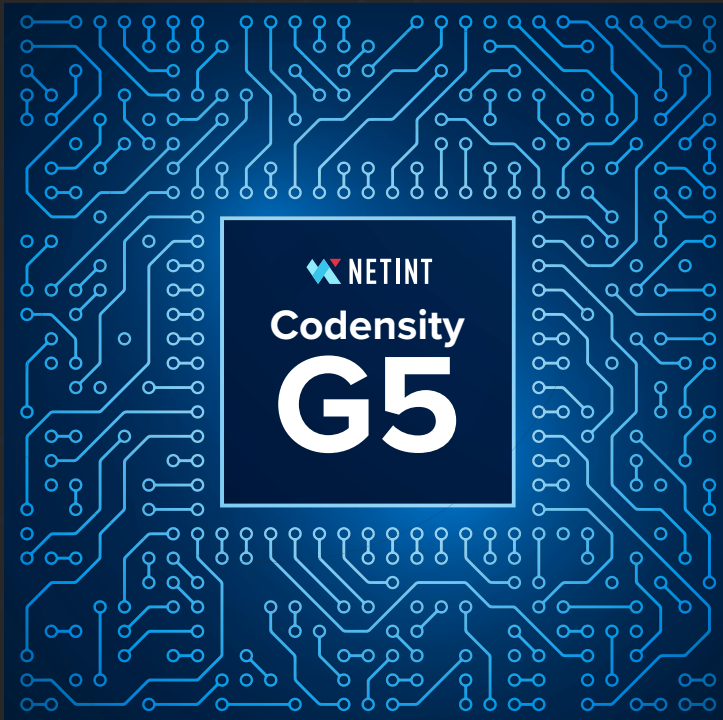
Our solution.

We designed an ASIC to slash the encoding footprint up to 80%

By replacing video encoding software with Smart VPUs (ASIC video processing units), you immediately get:

- 1. Increased encoding capacity using fewer Smart VPU chips**
- 2. Fewer chips require smaller hardware footprint**
- 3. Less hardware consumes less power**

This chain of events saves your bottom line and the planet.

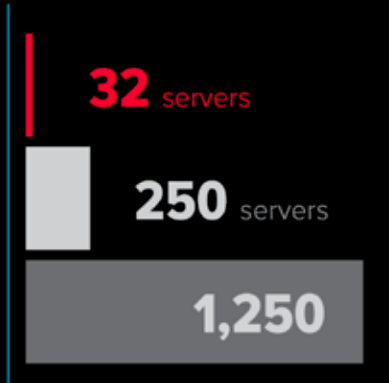


SMART VPU

NETINT video processing units powered by AI

GPU
NVIDIA T4 graphic processing units

CPU
INTEL SVT encoding software



SERVER DENSITY

Servers required to deliver 10,000 concurrent HD streams

SMART VPU

GPU

CPU



ANNUAL OPERATING COST

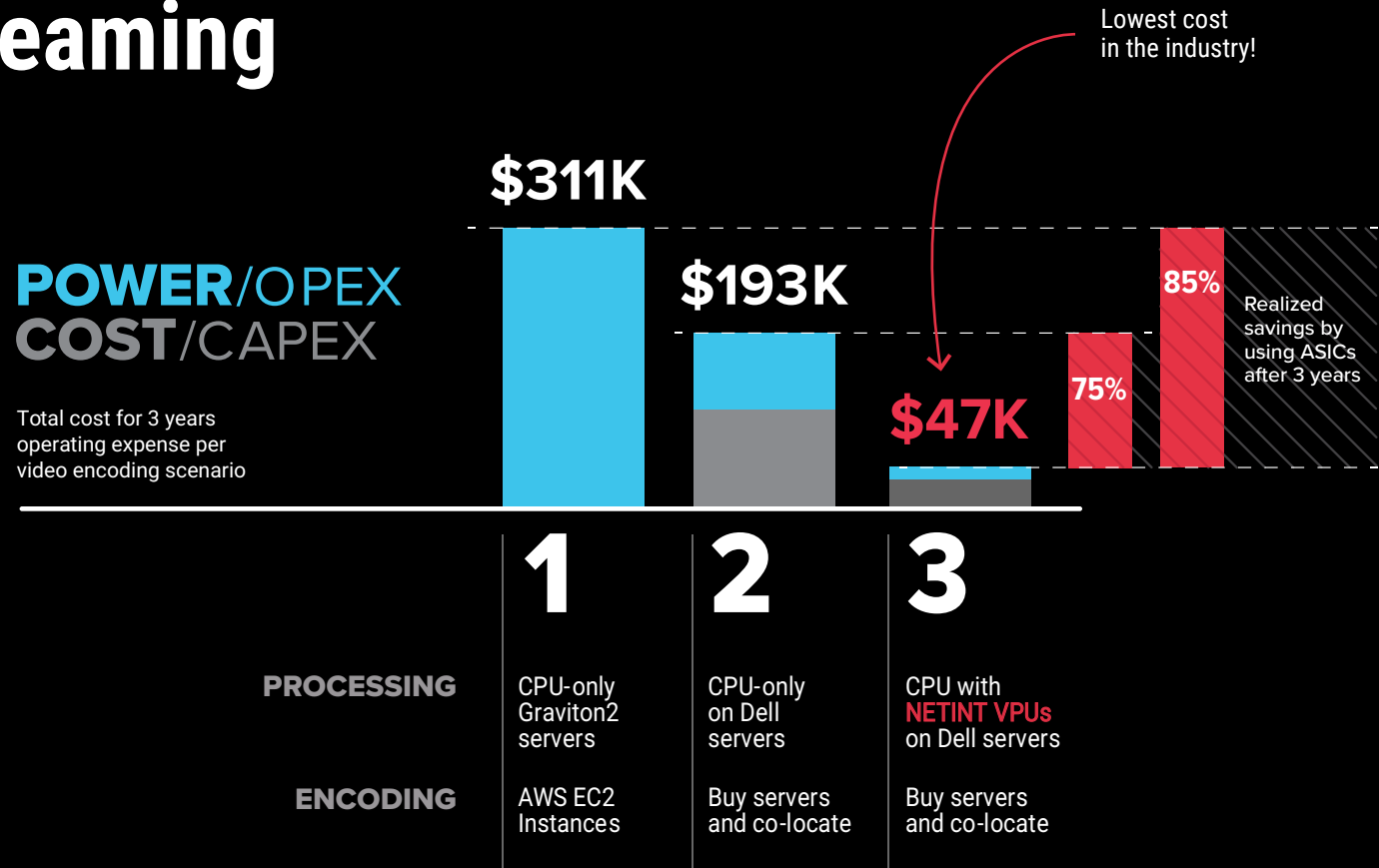
OPEX required to deliver 10,000 concurrent HD streams

This is why Google built a custom chip for YouTube

For everyone else who isn't Google, we did the heavy lifting for you.

We developed commercial-ready Smart VPU cards for easy drop-in replacement and immediate deployment.

The real cost of live streaming



Test assumptions:

- Servers run 100 concurrent five-rung encoding ladders
- x264 very fast preset used for CPU-only processing

Quadra Video Server

Smart VPU | Codensity Quadra G5

Ultra-high density, low cost
and powered by AI

Built on the Supermicro 1114S-WN10RT server platform, server contains ten Quadra T1U VPUs.

- **HEVC, H.264 and AV1 video encoding**
- **HEVC, H.264, and VP9 video decoding**
- **Up to 8K resolution**
- **10-bit HDR**

Ultra-low latency encoding of up to 320 broadcast quality 1080p30 streams in a compact 1RU form factor. Massive transcoding capacity enables breakthrough reductions of up to 90-95% in OPEX and CAPEX costs compared to software-based encoding systems.

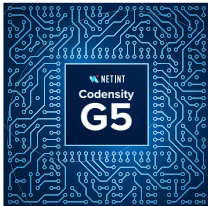
Performance results in this brochure are for the NETINT Quadra Video Server powered by an AMD EPYC™ 7543P (32-core) CPU. For encoding workloads with different encoding demands, the server is also available with the AMD EPYC 7232P (8-core) and 7713P (64-core) CPUs.



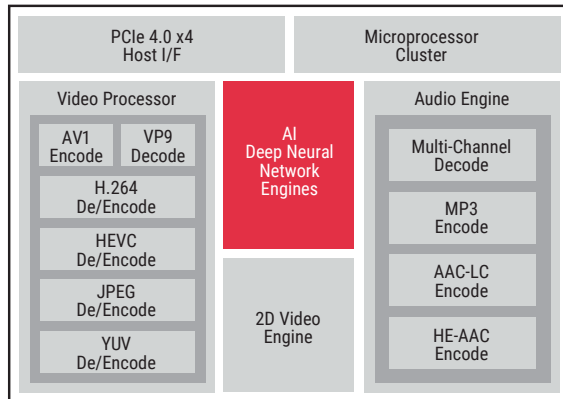
Codensity G5 Smart VPU

ASIC Video Processing Unit, powered by AI

The Codensity G5 architecture uniquely combines on-chip AV1, H.264 and HEVC video encoding and AI processing engines to deliver scalability for metaverse, live streaming, and interactive applications.



The core of NETINT's Codensity technology is an in-house built ASIC that increases encoding density compared to CPU-based software encoding. This density increase expands the number of channels that can be encoded without increasing the rack footprint. This reduces power and HVAC costs to deliver a lower TCO without sacrificing video quality or latency.



8K UHD Video Encoding

The Codensity G5 ASIC enables up to 8K video transcoding using the HEVC and H.264 codecs (AV1 is limited to 4K). Advanced codecs like AV1 and HEVC deliver superior quality to H.264 with up to a 60% reduction in bitrate, but when produced by CPU-only encoders, can require up to 10x the processing power, limiting throughput severely. HEVC and AV1 output with the Codensity G5 ASIC should be similar to H.264, making 4K and 8K live resolutions affordable and scalable for the first time.

Flexible Architecture

The Codensity G5 is built on a programmable microprocessor architecture to optimize the firmware and pipeline processing for improved performance and increased video quality. This counters a criticism that silicon-based encoders lack upgrade flexibility.

AI Engine

Two Deep Neural Network engines capable of up to 18 TOPS (trillion operations per second) enable object detection, classification, and segmentation to provide additional data to the encoding engine for image quality improvement and content-adaptive rate control for advanced performance and functionality. Seamlessly integrated for region-of-interest (ROI) encoding and background replacement. Additional features to be released.

Designed for the Cloud

High-density live UHD transcoding

The NETINT Quadra VPU takes full advantage of the video processing capability inside the Codensity G5 ASIC to support H.264, HEVC, and AV1 HEVC live encode functionality of up to 8K UHD video in SDR and HDR with HDR10 and other popular high dynamic range standards. By offloading complex encode and decode processing to the Codensity G5 ASIC, the Quadra VPU minimizes host CPU utilization. The result is a significant improvement in real-time transcoding density compared to any software or GPU-based transcoding solution.

Every NETINT Quadra Video Server installed in a data center would replace as many as 25 software-based video encoding servers.
(See appendices for details).

High power efficiency

Each NETINT Quadra U.2 module consumes only 20W of power at full load. This makes the Quadra Video Server, the most energy efficient video transcoder available.

Enterprise NVMe integration

Deployed in a U.2 form factor (and also available in HHHL AIC form factor), Quadra offers a simple upgrade path from CPU-based software to ASIC video encoding on any enterprise-class server.

**NETINT's Quadra Video Server
hosts ten Quadra VPUs supporting up
to 320 simultaneous live 1080p30
encoding sessions.**

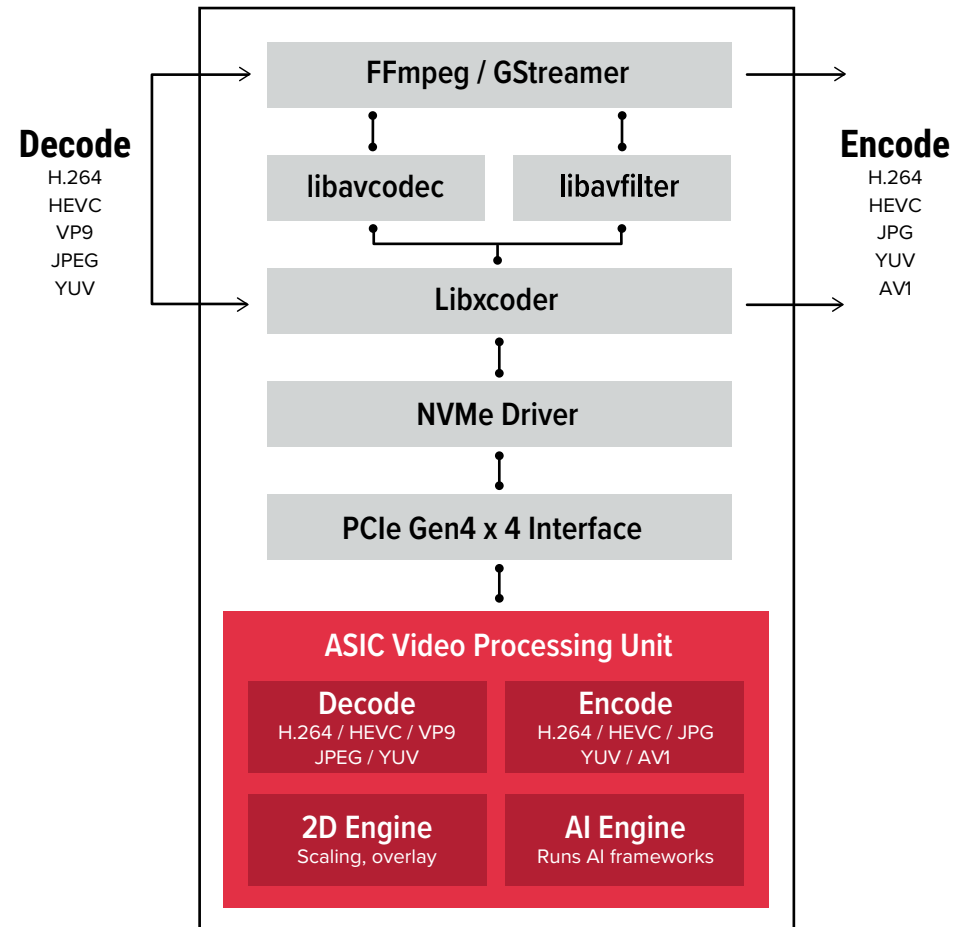


Simple Integration

Open-source suite of processing tools.

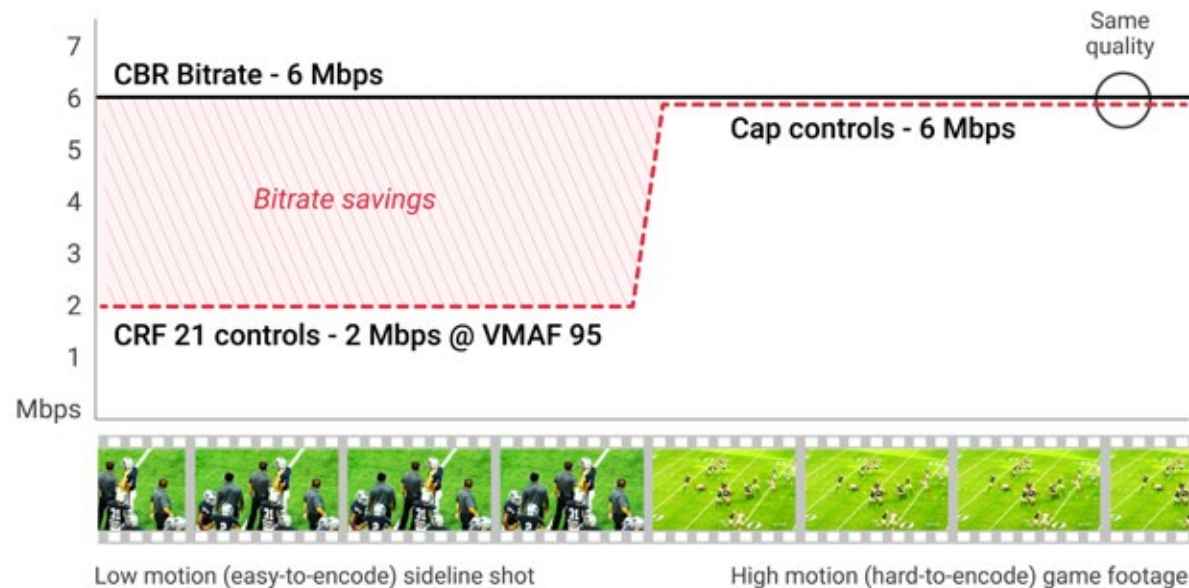
Many video processing and transcoding applications developers use FFmpeg and GStreamer, two open-source software libraries offering a vast suite of video processing functions. The Quadra video server includes highly efficient FFmpeg and GStreamer compatible SDKs, allowing operators to apply an FFmpeg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



Advanced Encoder Feature: Capped CRF

CRF 21 with a Cap of 6 Mbps *versus* 6 Mbps CBR



All Quadra VPU Products

Smart VPU | Codensity Quadra G5

	2nd Generation Smart VPUs			
	Server	Modules		
	G5 Smart VPU Quadra	Smart VPU T1U	Smart VPU T1A	Smart VPU T2A
Performance				
ASIC Codensity chip	G5, T1Us (10x)	G5	G5	G5 (2x)
Price	starting at \$19,000	\$1,500	\$1,500	\$2,750
Form Factor	1RU Server	U.2	AIC, HHHL	AIC, HHHL
Power Consumption	~500W	17W	20W	40W
Real-time Throughput Up to:	320x 1080p30 80x 4Kp30 20x 8Kp30	32x 1080p30 8x 4Kp30 2x 8Kp30	32x 1080p30 8x 4Kp30 2x 8Kp30	64x 1080p30 16x 4Kp30 4x 8Kp30
Latency	8 ms	8 ms	8 ms	8 ms
Encode Codecs	H.264, HEVC, JPEG, YUV, AV1			
Decode Codecs	H.264, HEVC, JPEG, YUV, VP9			
Audio Codecs	MP3, AAC-LC, HE-AAC			
Features				
Artificial Intelligence	150 TOPS	15 TOPS	18 TOPS	36 TOPS
New Capped CRF	●	●	●	●
Flexible GOP	●	●	●	●
Scaling	●	●	●	●
Cropping and Padding	●	●	●	●
Video Overlay	●	●	●	●
YUV / RGB Conversion	●	●	●	●
Configurable throughput	●	●	●	●

● Feature supported on VPU

Artificial Intelligence Specs

AI Deep Neural Network Inference Engines

INT8 Trillion Operations Per Second (TOPS)

- T1U: 15 TOPS
- T1A: 18 TOPS
- T2A: 36 TOPS

AI Deep Learning Frameworks

Support models trained with these major deep learning frameworks:

- Caffe
- Darknet
- Keras
- ONNX
- PyTorch
- TensorFlow
- TensorFlow Lite

Applications for Quadra AI Inference Engine include

- ROI-Encoding
- Scene Detection
- Background Removal
- Video Enhancement
- Facial Recognition
- Object Detection

Deployment Workflow for Pre-trained AI models

AI deep learning models are imported to Quadra VPUs with the NETINT AI Toolkit then processed (Import, quantization, validation and optimization), exported, and executed on Quadra Neural Processing Units (NPUs).



Supported Frameworks

- ONNX
- TFLite
- PyTorch
- Darknet
- TensorFlow
- Keras

Features

- 8 /16 bit quantization
- Hybrid quantization
- Accuracy validation
- Graph optimization
- Pre-processing integration

Features

- Hardware-aware optimization
- Execution graph generator
- Performance profiling
- Python/C Inference API

Specification	Input Size	Performance FPS @ 1 GHz
Yolov5s	640x640	78
Yolov5s	320x320	231
Yolov4-tiny	416x416	276
ResNet 50	224x224	228
MobileNetv2	224x224	1234
FSRCNNx3	360x640	36
DeepLabv3	257x257	452
BiSeNetV1	512x512	51
HrNet	256x192	72

Specifications

- Test hardware: T1A
- Test firmware version: 3.1
- AI capability per G5 ASIC: 18 TOPS
- Datatype for evaluation: INT8
- Batch size: 1
- Performance based on original model without pruning, sparsity or modification
- Quadra supports multiple AI modes (Full, Eco, Off) depending on the power/performance requirement

Quadra T2A Smart VPU

AI powered Video Processing Unit | Codensity G5



Form Factor	AIC (HH HL)
ASIC	2x Codensity G5
Interface	PCIe 4.0 x4x4
Power Consumption (Typ)	40W
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 64x 1080p30, 16x 4Kp30, 4x 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	36 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 T1A Smart VPU

AI powered Video Processing Unit | Codensity G5



Form Factor	AIC (HH HL)
ASIC	1x Codensity G5
Interface	PCIe 4.0 x4
Power Consumption (Typ)	20W
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	18 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 T1U Smart VPU

AI powered Video Processing Unit | 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

Smart VPU | Codensity Quadra G5



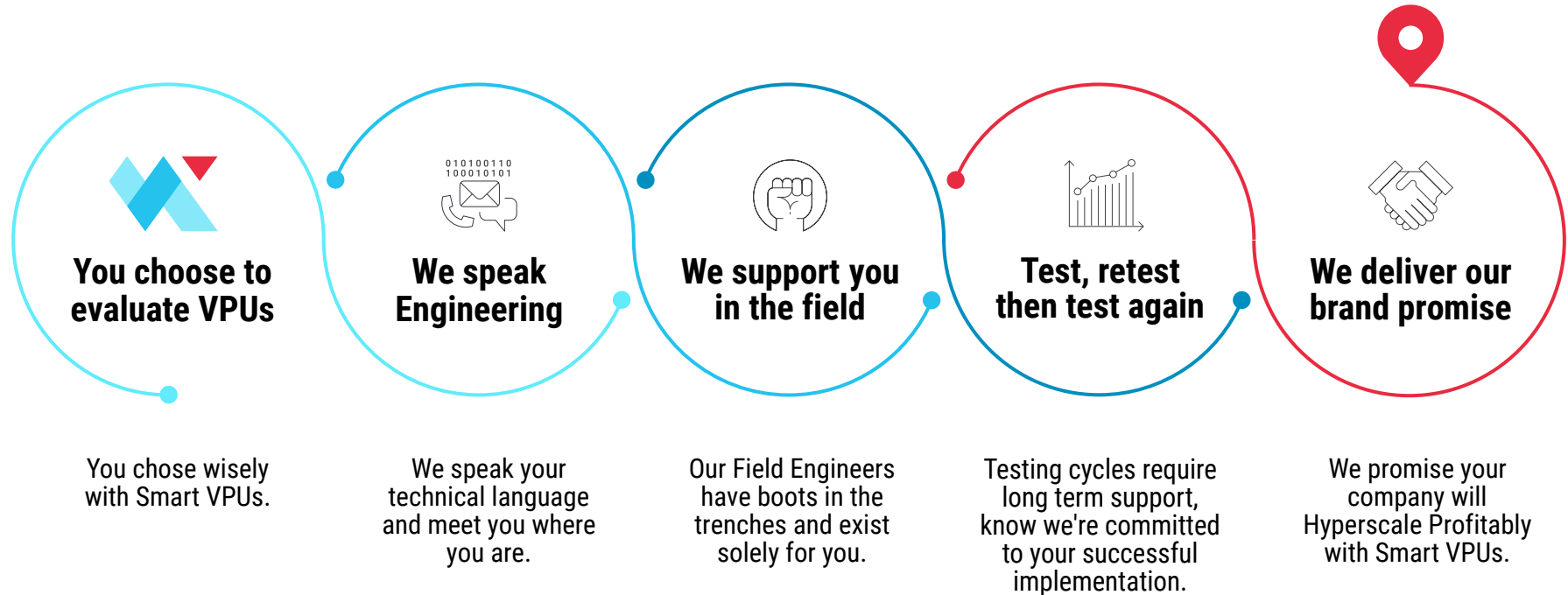
CPU Options	AMD EPYC™ 7232P Server Processor (8-core)
	AMD EPYC 7543P Server Processor (32-core)
	AMD EPYC 7713P Server Processor (64-core)
Operating System	Ubuntu 20.04.05 LTS
Memory	8x 16GB DDR4-3200
Storage	400GB M.2 SSD
NVMe Support	10x
PCIe Expansion	Up to 3x PCIe slots
Network Options	Dual 10GBase-T LAN
Power Consumption	~500W
Power Supply	700W: 100 - 140Vac
	750W: 200 - 240Vac
	750W: 200 - 240Vdc (CCC only)
Transcoders	10x NETINT Quadra T1U
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

Physical Dimensions	W: 17.2" (437mm), H: 1.7" (43mm), D: 23.5" (597mm)
Rack Size	1U
Weight	39 lbs (17.69 kg) <i>(fully loaded with 10 T1U VPUs)</i>
Environmental	50 degrees F to 95 degrees F Operating Temperature, 8% to 90% Operating Relative Humidity
Power Inputs	100 - 140Vac / 8 - 6V / 50-60Hz
	200 - 240Vac / 4.5 - 3.8A / 50-60Hz
	200 - 240Vdc / 4.5 - 3.8A (CCC Only)
Certifications	RoHS Compliant, UL Approved

Your Buying Journey

What to expect when evaluating NETINT

We know the typical sales cycle prospective buyers endure is a 12-18 month process and we're prepared to stand beside you and navigate you through. We're demonstrating our commitment to supporting you by heavily investing in this process so you can realize the value in our product and in working with us.





For more information on NETINT
encoding solutions, contact us.

sales@netint.com

netint.com