NETINT

Buyer's Guide

NETINT VPU Product Selection

NETINT Buyer's Guide

This guide is designed to help you choose the best NETINT product for your current and future needs.

As an overview, note that all NETINT hardware products run the same basic software controlled via FFmpeg and GStreamer patches or an SDK. This includes load balancing of all encoding resources in a server. In addition, both generations are similar in terms of latency and HDR support.

NETINT

1 Which Architecture: Codensity G4 (Logan) or Codensity G5 (Quadra)?

Tables 1 and 2 show the similarities and differences between Codensity G4 ASIC-based products (the T408 and T432) and Codensity G5-based products (all Quadra units).

Both architectures are available in either the U.2 or AIC form factor, the latter all half-height half-length (HHHL) configurations.

From a codec perspective, the main difference is that G5 products support AV1 output. In terms of throughput, the G5-based products offer four times the throughput per ASIC, but cost roughly four times more than G4, so the cost per output stream is similar. G5 power consumption is roughly 3x higher per ASIC than G4, but the throughput is 4x, so power consumption per stream is actually lower.

ASIC	Cost per ASIC	Form Factor	Codec Support						Power	Throughput
			H.264	HEVC	JPEG	YUV	VP9	AV1		
G4 - Logan	~ $400	U.2/AIC	Dec/ Enc	Dec/ Enc	no	Dec/ Enc	no	no	~7 watts	Up to 4K60p
G5 - Quadra	~ $1,500	U.2/AIC	Dec/ Enc	Dec/ Enc	Dec/ Enc	Dec/ Enc	Dec	Enc	~20 watts	Up to 8K60p

Table 1. Codec support, throughput, and power consumption.

Table 2 covers other hardware features. From an encoding perspective, G5-based products enable tuning of quality and throughput to match your applications, while quality and throughput are fixed for G4-based products. The G5's quality ceiling is higher than the G4, at the cost of throughput, and the quality floor is lower, with much higher throughput.

G5-based products are much more capable hardware-wise, performing scaling, overlay, and audio compression on board and offering AI processing of between 15 - 18 TOPS per G5 implementation. In contrast, G4-based products scale, overlay, and encode audio via the host CPU and offer no AI processing. You can read about Quadra's AI offering here.

Peer-to-peer DMA is a feature that allows G5-based products to communicate directly with some specific GPUs, which is particularly useful in cloud gaming. This is only available on G5-based products. You can read more about peer-to-peer DMA here.

Note that G4 and G5-based modules can co-exist on the same server, so you can add G5 modules to a server with G4 modules already installed and vice versa.

	Hardware Features					
ASIC	**Configurable Quality/ Throughput**	**Scaling**	**Overlay**	**Audio**	**AI**	**Peer-to-peer DMA**
G4	No	CPU	CPU	CPU	No	No
G5	Yes	Yes	Yes	Yes	15-18 TOPS	Yes

Table 2 Other hardware functionality.

Observations:

All project requirements are unique, but the data in Tables 1 and 2 suggest the products are similar on a cost-per-stream basis, with Quadra slightly more efficient on a watts/stream basis.

Choose G4-based products for:

- The absolute lowest overall cost
- Compatibility with existing G4-based encoding stacks
- Interactive same resolution-in/out productions

Choose G5-based products for:

- AV1 output
- AI integration
- Applications that need quality/throughput tuning
- Applications that involve scaling and overlay
- Maximum throughput from a single server
- Cloud gaming

2 Which G4-based Product?

This section discusses your G4-based modules shown in Figure 1, with the U.2-based T408 on the left and AIC-form factor T432 on the right. These products are designated as Transcoders since this is their primary hardware-based function.

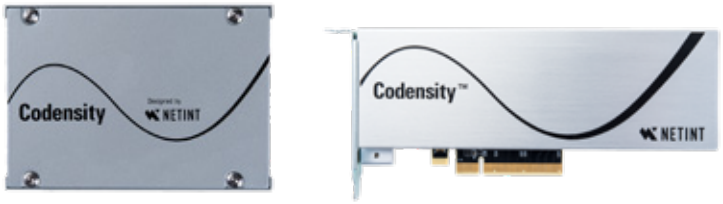

Figure 1. The NETINT T408 in the left, T432 in the right.

Table 3 identifies the key differences between NETINT's two G4-based modules, the T408, which includes a single G4 ASIC in a U.2 form factor, and the T432, which includes four G4 ASICS in an AIC half-height half-length configuration.

Product	Form Factor	Number of ASICs	Cost in Low Quantities	Power	Throughput
T408	U.2	1	$300	~7 watts	Up to 2x 4Kp30
T432	AIC HHHL	4	$1,200	~27 watts	Up to 2x 4Kp30

Table 3. NETINT's two G4-based modules.

Observations:

- The U.2-based T408 offers the best available density for packing units into 1RU servers.
- The AIC-based T432 is the best option for computers without U.2 connections and for maximum density in a single-device installation.

3 Which G5-based Product?

Figure 2 identifies the three Quadra G5-based modules, with the U.2-based T1U (cross hatched) in the back, the AIC-based T1A in the middle (silver), and the AIC-based T2A (red) in the front. These modules are designated Smart VPUs, Video Processing Units, because they're powered by AI and their hardware functionality extends far beyond simple transcoding.

Figure 2. Quadra T1U back, T1A middle, and T2A front.

Table 4 identifies the key differences between NETINT's three G5-based products:

- The T1U includes a single G5 ASIC in a U.2 form factor.
- The T1A includes a single G5 ASIC in an AIC half-height half-length configuration.
- The T2A includes two G5 ASICs in an AIC half-height half-length configuration.

Product	Form Factor	# of ASICs	Cost in Low Quantities	Power watts	AI Process	Throughput
Quadra T1U	U.2	1	$1,500	17 w	15 TOPS	32x 1080p30 8x 4Kp30 2x 8Kp30
Quadra T1A	AIC HHHL	1	$1,500	20 w	18 TOPS	32x 1080p30 8x 4Kp30 2x 8Kp30
Quadra T2A	AIC HHHL	2	$2,750	40 w	36 TOPS	64x 1080p30 16x 4Kp30 4x 8Kp30

Table 4. NETINT's three G5-based products.

Observations

- The U.2-based Quadra T1U offers the best density for installing in a 1RU server
- The Quadra T2A offers the best density for AIC-based installation and is ideal for cloud-gaming servers that need peer-to-peer DMA communication with GPUs
- The AIC-based Quadar T1A is the most affordable AIC option for installs that don't need maximum density

4 Module or Server?

Logan Video Server contains ten T408 VPUs

Quadra Video Server contains ten Quadra T1U Smart VPUs

NETINT offers two streaming servers that use the same Supermicro server; the Logan Video Server contains ten T408 U.2 modules, while the Quadra Video Server contains ten Quadra T1 modules. In general, we recommend buying individual products for testing and pre-deployment integration.

Once you've decided to roll out the transcoders, the servers offer a turnkey option for fast and simple deployment. All server components, including CPU, RAM, hard drive, OS and software versions, have been extensively tested for compatibility, stability, and performance. Supermicro is a very reputable manufacturer and has partnered with NETINT to offer the servers at very aggressive pricing.

While NETINT's hardware has been successfully deployed in hundreds of different types and brands of servers, hardware issues infrequently arise as they do with all computer peripherals. Buying either server will eliminate this eventuality and will help ensure the fastest possible deployment.

As between the servers, the answer to question "module or server" above should control your selection.

NETINT

For more info about encoding with Smart VPUs, contact us.

sales@netint.com
www.netint.com