Meet NETINT: The Startup Selling Datacenter VPUs To ByteDance, Baidu, Tencent, Alibaba, And More

Video encoding ASICs are critical to ad-tech, cloud gaming, video conferencing, and VDI

DYLAN PATEL

AUG 04, 2022 • PAID

6 4 Share

NETINT is one of the most successful semiconductor startups you have never heard of, and they do it in the cutthroat datacenter market. Over 200 billion minutes of video were processed by their silicon in 2021. NETINT was founded in 2015, and they have scaled to 160 engineers across Vancouver, Toronto, and Shanghai. NETINT is the leader in merchant video encoding ASICs.

If this sounds familiar to some of you, Google has a similar silicon effort for their internal workloads called Argos. We wrote about that ASIC in the past and how Google is able to enhance their product capabilities while also replacing the millions of Intel CPUs. Both firms are already using/shipping the 2nd generation of their video encoding ASIC, dubbed VPU (Video Processing Unit).

Before we talk more about NETINT's silicon solution, we want to dive into video content, encoding, and the shifting landscape that is increasing demand for this type of semiconductor in a massive way. This has massive implications for not only semiconductor companies but also major content delivery and ad networks from firms like Google, Amazon, Meta, ByteDance, and Netf1ix.

Video streaming related content is exploding in popularity with it compromising over 80% all internet traffic. In the old days, studios created content, processed, encoded, and distributed content in a few formats. The number of viewers per encoded video was relatively high because viewers had a limited number of TV channels to watch on

demand. These studios could heavily optimize their processing of the content for maximum visual fidelity given the method of delivery whether cable or DVD.

The web has been decentralizing in content creation and more time from consumers is spent on user generated content. Google led this charge with YouTube which is the large source of user generated content with over 700 hours of YouTube videos uploaded every minute by users. All of these videos must be processed and encoded by Google in multiple formats for distribution to users. The processing and encoding involved is critical to ensure that user generated content has no terms of service violations, includes captions, and that videos can be searched for based on what content is contained rather than relying solely on rudimentary tags and titles.

There is a delicate balancing act between optimizing the quality vs the size of a video file. This balancing act is critical as north-south bandwidth out of a data center is very expensive. If the encoding solution produces even 5% smaller files for the same quality of video, then a major content delivery network could save many millions of dollars.

Furthermore, many consumers also have poor internet connections or data caps, so the video must be delivered in the optimal size and resolution for their specific situation. If the correct options of resolution to quality to bandwidth consumption are not provided, they could subconsciously prefer to get their entertainment content from another provider.

Google is not alone in this space. Other firms such as Amazon's Twitch, Meta's Facebook Video and Instagram Reels, and ByteDance's TikTok have also been exploding in user generated content. The amount of watch time per unit of content is significantly lower with user generated content. Furthermore, forms of content which have only 1 user per video stream are also developing at a rapid pace. These use cases include game streaming services, virtual desktops, and other remote interactive computing environments.

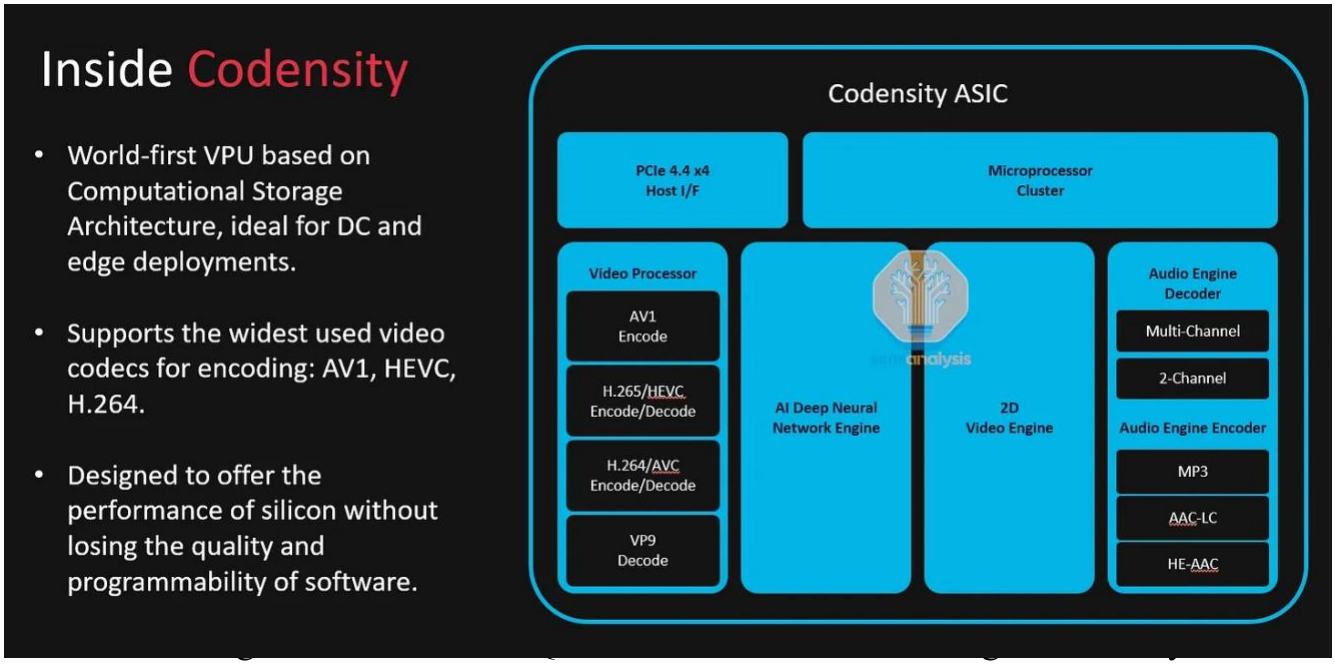
Even when video content comes from a more traditional studios and has a limited number of streaming options, the content delivery network may still need to do high volumes of additional processing and encoding on the f1y. One such example is advertisement supported content. Many paid streaming services including Netf1ix are moving to offering ad supported content.

Ads can come in a variety of different resolutions, sizes, and bitrates which may not match the format of the content. The companies delivering the ads may own their own network like Google's AdSense, but most companies must rely on 3rd party clearing houses. Advertisers pay more for higher impression and conversion ratios. For advertisements to have the maximum effectiveness, they must not break the user's immersion. Keeping immersion includes matching the resolution and quality of the content the user came for. Every ad must be processed further by the content network or else the effectiveness and ultimately revenue from them could fall.

Ads and ad blocking methods are also in a constant arms race, and free content networks such as YouTube and Twitch have suffered untold amounts of lost revenue because of ad blockers. While the arms race has been dynamic, generally ad blockers have won, and the only full proof solution to beating ad blockers is to encode the video content with the ads.

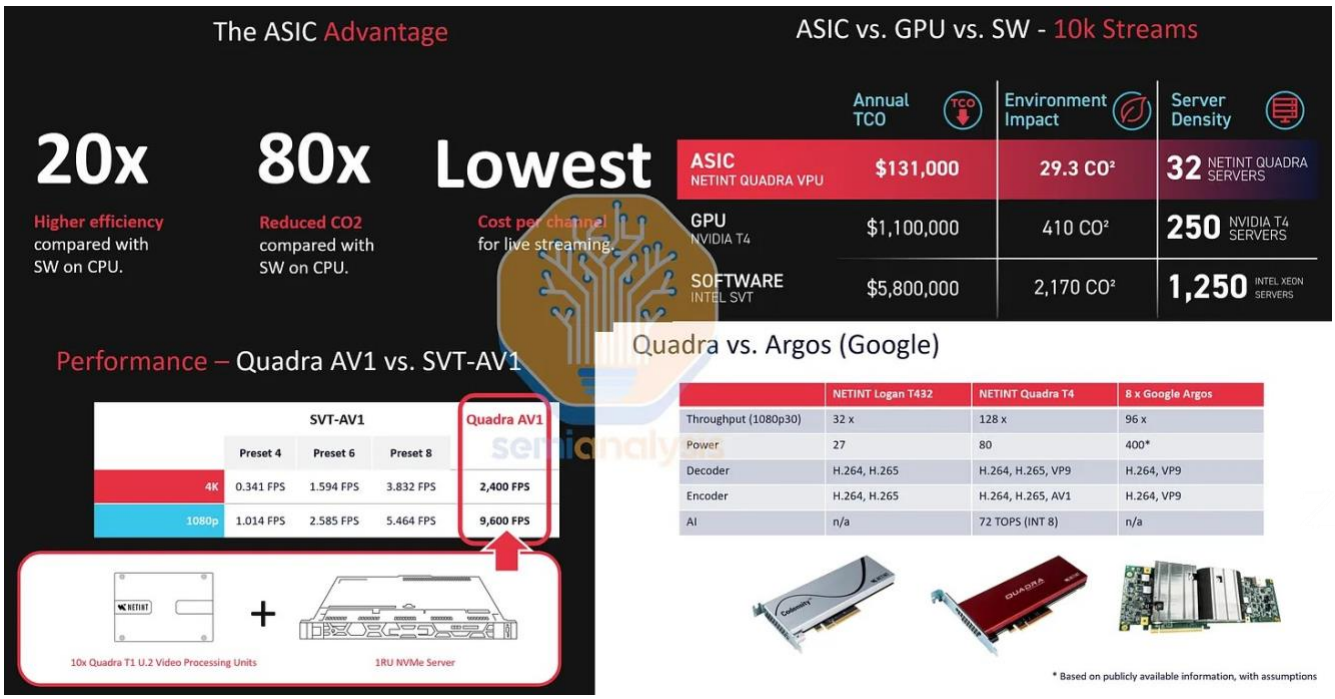
The on-demand ad market is highly dynamic and includes time-based and user-based targeting. To maximize ad revenue and effectiveness, content networks should splice the content stream and encode it with a personalized ad that was sold in the demand auction for that specific user's advertising ID. This is computationally expensive processing step that cannot be achieved economically with CPUs and GPUs in a cost-effective manner.

The economics of the service provider are key to scaling capabilities of these content networks. Google needed specialized silicon to implement VP9 across YouTube, otherwise it would have taken over 10 million CPUs. This is where Google's homegrown solution and NETINT's VPU solutions come into play. Google will only serve their own use cases, but NETINT has been working just as long if not longer on their products. They also serve a much larger market of firms. As such, NETINT has racked up a number of major wins including major names such as ByteDance, Baidu, Tencent, Alibaba, Kuaishou, and a similar sized US based global platform.



2-year cycle and the 3rd generation Mt. Augusta will be shipping in 2024/2025. The Quarda ASIC has 4 lanes of PCIe 4.0, attached DRAM, and the hardware ASICs for video decoding/encoding, audio encoding/decoding, AI inference, RISC CPUs, and a 2D video engine. The biggest standout capabilities are AV1 and H265/HEVC. NETINT's comparisons are generally using H265 and AV1 as those are the codecs where the TCO advantage is larger.

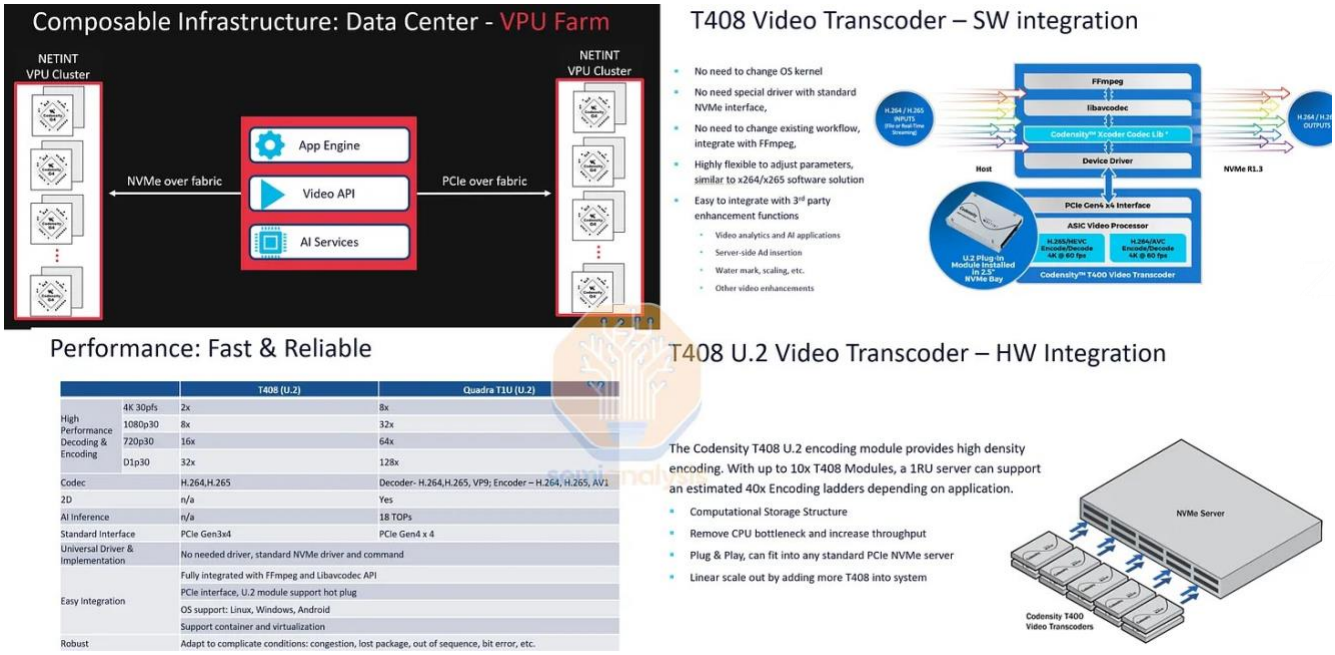
In general ASICs need to provide an order of magnitude better capability in their target workload to be successful. While their chip would provide a meaningful advantage in H264, it truly shines when HEVC or AV1 is utilized. The amount of computation required for these codecs is significantly higher than that of H.264.



The performance comparisons are very impressive. T432 is their first VPU product with 4 Logan chips. Quadra T4 is the 2nd generation VPU that is available in 1, 2, and 4 chip versions. Using the HEVC codec, NETINT crushes Nvidia's last generation T4 (there are newer Ampere based GPUs) and Intel's Skylake/Cascade Lake servers. The density and power consumption that can be achieved with a video ASIC is unmatched compared to CPUs and GPUs. The comparison using AV1 is even more powerful.

NETINT even compares to Google Argos. That one is a bit unfair given we know the 2nd generation is in production. With that said, the power consumption and throughput comparison are only with stated specs, and real-world comparisons will likely vary.

Google has not detailed that chip, so it is the best possible comparison NETINT could have done. The comparison is also pointless as Google is not looking to sell their VPU externally.



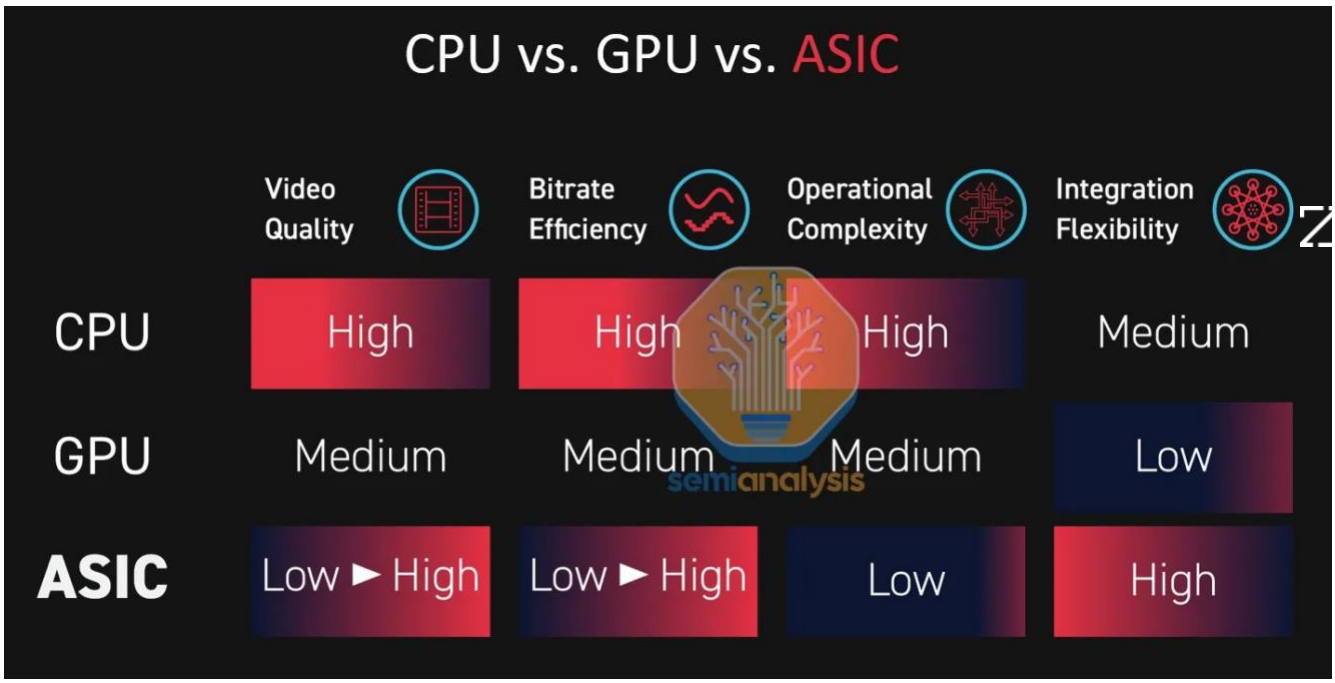
		T408 (U.2)	Quadra T1U (U.2)
High Performance Decoding & Encoding	4K 30pfs	2x	8x
	1080p30	8x	32x
	720p30	16x	64x
	D1p30	32x	128x
Codec		H.264,H.265	Decoder- H.264,H.265, VP9; Encoder – H.264, H.265, AV1
2D		n/a	Yes
AI Inference		n/a	18 TOPs
Standard Interface		PCIe Gen3x4	PCIe Gen4 x 4
Universal Driver & Implementation		No needed driver, standard NVMe driver and command	
Easy Integration		Fully integrated with FFmpeg and Libavcodec API	
		PCIe interface, U.2 module support hot plug	
		OS support: Linux, Windows, Android	
		Support container and virtualization	
Robust		Adapt to complicate conditions: congestion, lost package, out of sequence, bit error, etc.	

The formfactor of the NETINT VPU is very interesting. They offer it as an NVMe device in the U.2 form factor. This is the formfactor a data center would use for an array of NAND based SSDs. This allows NETINT to easily slot their products into a chassis type and form factor that every datacenter already uses. It can also come as a standard PCIe card, but the U.2 form factor is higher density. It is sold through Supermicro and Inspur systems. The NETINT VPU also supports composable server architecture.

The software for the VPU is also highly configurable given there are a large variety of incoming stream sizes and resolutions and a wide variety of output resolutions and quality levels. This is especially critical as more video is outputted from smartphones. The range of input resolutions and constant orientation changes from vertical to horizontal can cause performance issues if the encoder is not built for these use cases. All encoders are not built equal and that can be seen in the GPU world with how much better Nvidia's encoder is versus AMD's.

One example of the configurability is that a NETINT customer has been able to program the ASIC to preferentially encoding certain parts of the scene to higher quality

and the other parts to power quality to maximize the end quality of a stream while minimizing file sizes. The built in AI processing allows backgrounds to be removed, filtering, identifying key areas or people to pay attention to. These techniques have allowed customer to reduced encoding cost by as much as 5x.



In a sense, the VPU is more f1exible than other encoding methods. A CPU could target lower video qualities and bitrate efficiencies, but its performance wouldn't improve enough to be worth it. Furthermore, that CPU would still consume too much power and space so that the TCO wouldn't be worth it. With a GPU, running the driver stack is a complicated mess for some applications. Various versions of Linux or Windows do no work correctly. This sort of software issues has held Intel GPUs back including the cancelled Xe HP tiled GPU architecture which was very optimized and even marketed for the datacenter video market.

NETINT has an impressive group of customers including but not limited to ByteDance, Alibaba, Baidu, Kuaishou, and a similarly massive US global platform. They have created a product that is one of the kind in the merchant silicon market and offers the ability for major content delivery networks and cloud vendors to scale new experiences for video based content such as user generated short form content, cloud gaming, and real time ad insertion.

We will next be covering why we believe many US based companies (outside of Google) are late to using VPUs, including US tech firm adoption going forward. Nvidia and AMD's attitude and plans in this space will also be discussed. We also want to talk about the battle between H265 vs AV1. Lastly there are some important notes about the company's history and their funding that we feel should be discussed and investigated further.

That discussion will all be in the paid subscriber only section below.

Previous Next