EIZO Rugged Solutions and NVIDIA Turing Push the Boundaries of Rugged AI

CONTENTS

1 Introduction....1
2 Real-time Ray Tracing....1
3 Deep Learning and AI....3
4 Video Processing and Encoding....3
5 Cutting Edge Memory Enhancements....4
6 Turing Availability in the Defense Industry....4

White Paper
Revision 2
August 2020
By Nick Whitlock

1. Introduction - AI and GPUs Now Critical to Defense

Artificial Intelligence (AI) as a field in computing has been around for nearly three-quarters of a century, and yet AI only recently has made significant impacts into industries. Google pioneered the use of neural networks back in 2012 through its "cat paper," Building High-Level Features Using Large-Scale Unsupervised Learning – so-called because its neural network was so good at detecting faces that it was not only able to detect human faces, but cat faces as well [i, ii]. The term, Deep Learning (DL), is used to describe the process of training artificial intelligence using neural networks.

Thanks to the computational breakthroughs in GPU technology, a global arms race is now ramping up to utilize Deep Learning for military purposes. In the United States, the Pentagon has spent nearly $70 million on Project Maven – a project to use Artificial Intelligence in drones to recognize people and objects and determine friend from foe . These drones would then fly completely autonomously, scouting for hostile forces and civilians, and returning back to base to recharge. In 2019, the Department of Defense issued a Request for Information (RFI) for a swarm of drones designed to launch from air and sea, autonomously follow predetermined routes, and stream video information back to a remote human operator for the purposes of reconnaissance and search and rescue operations . The Chinese government has unveiled its plan to invest in artificial intelligence, hoping to develop a $150 billion industry by 2030. South Korea has pledged $1 billion, and Canada has pledged $125 million . It is clear that the future of the military involves AI, and therefore all future video processing and encoding platforms should consider the implementation of accelerated AI capabilities. We believe that the future of military artificial intelligence lies with one of the progenitors of the AI "big bang" – NVIDIA.

2. Real-Time Ray Tracing

Casual gamers may recognize the NVIDIA name from their work in 3D graphics. The primary defining feature of NVIDIA's RTX-enabled cards (and the Turing architecture in general) is real-time ray tracing (RT). In the past, graphics pipelines were accustomed to a specular workflow, which uses tricks such as pre-baking the shadows cast by world lighting onto textures, increasing the brightness of specific triangles in a scene, and drawing simple blob shadows as a black texture. This was necessary, since early dedicated 3D graphics cards were limited in their computational power and were not able to perform as intensive computations as today's GPUs can do . More recently, the games industry has been moving towards a concept known as Physically Based Rendering (PBR), where light and textures attempt to follow the laws of nature to the exact letter. This includes such minutia as the Fresnel Effect, where reflections of objects on water appear to be blurrier at more acute angles, and clearer when viewing them at obtuse angles. [vi]

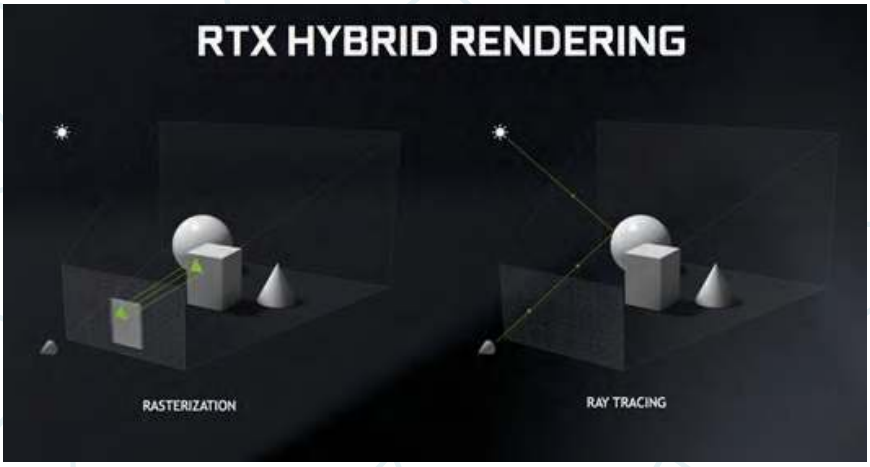


Figure 1 : A demonstration of rasterization vs ray tracing.
(Image Credit: NVIDIA)

Real-time raytracing is considered the holy grail of physically-based rendering. Instead of using a rasterization approach for rendering triangles on the screen, each pixel on the screen has an individual trace that bounces off surfaces that it hits, being lit by light sources in the scene. The jump from rasterization to raytracing allows for more realism in the picture since the exact scientific properties of photons can be taken into account instead of approximating them.

But performing raytracing in real-time is an extraordinary jump since previously ray tracing required massive rendering farms accessible to only companies like Pixar and DreamWorks. There is some level of randomness associated with real-time raytracing however, since not all of the lights and pixels can be captured in the ~16 milliseconds required to render just one frame of a 60 frames-per-second game, and this results in noise in the image.

Figure 2: Quake 2 ray-traced; very noisy.
(Image Credit: Kotaku Australia and Edd Bidulph)

Reducing the noise that ray tracing brings is a complex process that goes beyond the scope of this white paper[2]. However, it is important to note that the de-noising process typically involves artificial intelligence, and that is where the Turing architecture's Tensor cores become highly valuable to the pipeline. Tensor cores are highly specialized GPU hardware components that are designed to tackle complex AI problems.

In games, this involves both de-noising a raytraced picture and antialiasing. NVIDIA has recently introduced DLSS, or Deep Learning Super Sampling, which applies a trained neural network to the image to reduce jagged edges [vii].

In addition to performing DLSS anti-aliasing and assisting with ray tracing de-noising in real-time, they are also capable of performing high-speed Dense Optical Flow. Calculating optical flow involves measuring the velocity of either edges (Sparse Optical Flow) or individual pixels in the image (Dense Optical Flow) from two input frames using algorithms such as Lucas-Kanade[3]. Sparse Optical Flow is much less computationally expensive than Dense Optical Flow; the previous EIZO Rugged Solutions Condor cards using the Pascal architecture were capable of calculating Sparse Optical Flow and displaying results at 60 frames per second, but under dense optical flow they would audibly whine from the computational strain and could only output a few frames per second even under completely ideal circumstances [4].

Figure 3: Left: Sparse Optical Flow; Right: Dense Optical Flow (Image Credit: NVIDIA)

However, the specialized hardware in the Turing architecture is capable of performing dense optical flow at 30 frames per second when tested in the same environment, which is a significant improvement from the Pascal architecture. This is driven through the NVIDIA Optical Flow SDK, which uses NVIDIA's own optical flow algorithms. [viii] The NVIDIA Optical Flow SDK has been integrated into the popular OpenCV computer vision library.

3. Deep Learning and AI

NVIDIA's AI stack is referred to as CUDA-X [ix] , and comprises several software SDKs that optimize neural network training and inference. The Tensor cores in the Turing architecture are capable of being used for training arbitrary neural networks as well as inferencing them. During the training phase of a neural network, labeled data is given to the network to provide a general idea of how to solve a problem. Training neural networks can take days or even weeks [5], as millions of images or videos are used as input to help generalize a pattern that the computer can understand. This training process is typically done on a GPU server to increase throughput. NVIDIA provides a robust number of GPU-accelerated solutions for neural network training such as DALI (Data Loading Library), cuDNN (CUDA Deep Neural Network), NCCL (NVIDIA Collective Communications Library), and NeMo (NVIDIA Neural Modules), and has been recognized as an industry leader for neural network training hardware. The Turing architecture offers up to twenty-four times the training throughput and three times the performance of the Pascal architecture, and forty-seven times the performance of a CPU-powered solution [x].

Once training has been completed, the neural network is ready to be deployed in the field and executed on an embedded GPU payload. Neural networks for object classification prove difficult to run in real-time, with single-shot detection (SSD) being seen as the optimal model to be run in real-time for its balance of accuracy and speed [xi] . While single-shot detection can be run in real-time at 720p at 30 frames per second on the Pascal architecture [6] , it is not as accurate as two-shot detection algorithms such as faster R-CNN. Our tests on the Turing architecture have revealed that single shot detection can be run at 60 frames per second in Full HD 1080p, and faster R-CNN performs twice as well.

Both of these benchmarks are propped up in part by NVIDIA's TensorRT hardware. TensorRT takes a compatible AI model (such as a Caffe, TensorFlow, or ONNX model) and optimizes it for the target hardware, often removing redundant calculations from the model in the process [xii] . It is compatible with both the Pascal and the Turing architecture, although it performs much better on the Turing architecture, as it can replace some operations in the model with new ones that are more optimized for running on Tensor cores. The process of converting a slower, CPU-bound model to an optimized TensorRT model does take some time, and the converted model is not transferable to other hardware, but the result can be cached, resulting in the model being quickly loaded.

4. Video Processing and Encoding

Another NVIDIA software product that is still in development is known as NVIDIA NGX. Exclusive to the Turing architecture, NGX will offer features such as image in-painting (where removed regions of a photo are filled in) and image super-sampling. It does this in part by running inference on specialized neural networks in the driver that are highly optimized for the NVIDIA hardware. According to NVIDIA, NGX will also offer the ability to pass in raw video frames and increase the resolution by two to four times in each dimension, with no loss in visual fidelity. The frames can also be passed into the onboard NVENC cores to rapidly encode the video to H.264 or H.265 (HEVC). NVIDIA NGX is currently available through an early access program, and only a minor evaluation of the beta release has been done by EIZO Rugged Solutions. Initial results look promising.

While the Condor cards are an excellent choice for feature identification and classification of video, many Condor cards additionally provide dedicated video inputs to be utilized by the mating GPU. The ability to encode video for later analysis or for streaming to a central hub is paramount in these payloads. The onboard NVIDIA GPU makes it easy to both encode and decode video in H.264 and H.265 (HEVC) format through the NVENC (encoding) and NVDEC (decoding) cores.

6 Condor GR4 using TensorRT; MobileNet SSD

While the Condor cards are an excellent choice for feature identification and classification of video, many Condor cards additionally provide dedicated video inputs to be utilized by the mating GPU. The ability to encode video for later analysis or for streaming to a central hub is paramount in these payloads. The onboard NVIDIA GPU makes it easy to both encode and decode video in H.264 and H.265 (HEVC) format through the NVENC (encoding) and NVDEC (decoding) cores.

5. Cutting Edge Memory Enhancements

Lastly, the Turing GPUs are the first such GPUs to support GDDR6 DRAM. GDDR6 memory was first introduced in 2018 and offers significant power consumption and bandwidth improvements over GDDR5 [xiii]. Below is a table outlining the key differences:

Statistic	*Pascal GDDR5*	*Turing GDDR6*
Bandwidth	288 GB/s	616 GB/s
Bus Width	256-bit	352 to 384-bit
Power Draw	1.5 V	1.3 V

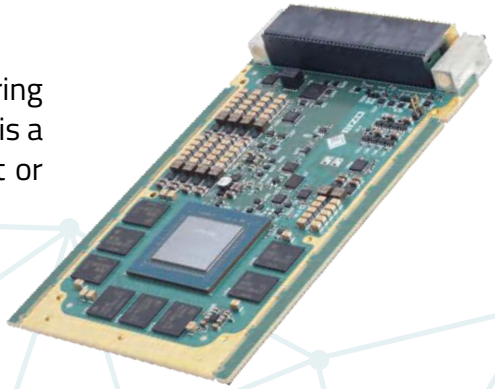
To use GDDR6 DRAM, the Pascal architecture had to be redesigned from the ground-up by NVIDIA, but in the end, it resulted in higher speed and reduced noise, as well as 20% improved power efficiency [xiii]. The benefits of a larger and faster memory bus are most noticeable in GPGPU/CUDA operations, where large amounts of data are being processed at a rapid pace. This improvement in memory throughput is critical for customers processing large datasets on the GPU.

6. Turing Availability in the Defense Industry

As video-oriented products, the Condor line of hardware is most often used in military reconnaissance and search-and-rescue operations. Our products can be found on naval vessels for processing video data, as well as in the various displays on aircraft (jets, gunships, helicopters, etc.). The VPX-enabled Condor cards in particular are designed to handle the kinds of extreme heat, shock, salt, and vibration that those types of jobs require. EIZO is in a unique position to deliver a product that can meet those needs, while not compromising on GPU performance/power, video quality, and latency.

The Turing architecture offers considerable upgrades from Pascal, including ray tracing for graphical applications, Tensor cores for specialized AI applications, an improved NVENC video encoder, and NVDEC video decoder, and DRAM upgraded from GDDR5 to GDDR6 for increased bandwidth and lower power draw. The decision to upgrade from the Pascal to the Turing architecture is an important one and one which we are confident that our customers will make great use of.

EIZO Rugged Solutions is now offering embedded RTX-enabled Turing GPUs as part of the Condor RTX series. The first Condor RTX product is a 3U VPX RTX5000 GPU Output card, offering either four DisplayPort or DVI outputs that can be customized to suit our clients' needs.



Condor GR5-RTX5000
Embedded Graphics & GPGPU Processing Card

The Condor GR5-RTX5000 card is compatible in parallel with our existing Condor video capture cards and can be used to encode video data or perform CUDA or AI operations on the captured video data. The integrated NVIDIA RTX GPU offers various improvements over the existing output cards, including greatly improved deep learning acceleration, an improved NVENC core for better video encoding, integrated hardware Dense Optical Flow, and an upgraded GDDR6 memory bus for improved bandwidth, throughput, and power efficiency [xiii].

The Condor GR5-RTX5000 meets strict data integrity requirements for mission-critical applications with uncompromised computing accuracy and reliability. The 3072 CUDA® parallel processing cores in the NVIDIA Turing™ architecture offer a multitude of capabilities such as mesh shading, variable rate shading, texture space shading, multi-view rendering, and ultra-high performance GPGPU computing. The GPUDirect® RDMA implementation offers fast data transfer/communication from connected hardware, such as FPGAs, and switches directly into GPU memory, avoiding unnecessary memory copies and CPU overhead resulting in minimal latency. With 384 Tensor cores and 48 RT cores, the Condor GR5-RTX5000 delivers high AI inferencing performance. Multiple precision modes such as FP64, FP32, FP16, INT8, INT4, and INT1, enables up to 32X throughput compared to previous generations and even offers features like AI de-noising.

For more information regarding NVIDIA Turing-based products, visit EIZO's website, https://www.eizorugged.com/products/rugged-graphics-and-video-capture/nvidia/

i. Q. V. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean and A. Y. Ng, "Building High-Level Features Using Large-Scale Unsupervised Learning," in 29th International Conference on Machine Learning, Edinburgh, Scotland, 2012.
ii. Wired, "Google's Artificial Brain Learns to Find Cat Videos," 26 June 2012. [Online]. Available: https://www.wired.com/2012/06/google-x-neural-network/. [Accessed 15 July 2020].
iii. Z. Fryer-Biggs, "Inside the Pentagon's Plan to Win Over Silicon Valley's AI Experts - Wired," 21 December 2018. [Online]. Available: https://www.wired.com/story/inside-the-pentagons-plan-to-win-over-silicon-valleys-ai-experts/. [Accessed 14 July 2020].
iiii. General Services Administration and Federal Systems Integration and Management Center, "Request for Information to All Interested Industry Partners," 26 December 2019. [Online]. Available: https://beta.sam.gov/api/prod/opps/v3/opportunities/resources/files/da7309e2df084e7bb85debb4f64bbafb/download?api_key=undefined&token=. [Accessed 15 July 2020].
iv. C. Metz, "As China Marches Forward on A.I., the White House Is Silent," The New York Times, 12 February 2018.
v. A. Edelsten, "NVIDIA DLSS: Your Questions, Answered," 15 February 2019. [Online]. Available: https://www.nvidia.com/en-us/geforce/news/nvidia-dlss-your-questions-answered/. [Accessed 14 July 2020].
vi. NVIDIA Corporation, "NVIDIA Optical Flow SDK," [Online]. Available: https://developer.nvidia.com/opticalflow-sdk. [Accessed 14 July 2020].
vii. NVIDIA Corporation, "NVIDIA Optical Flow SDK," [Online]. Available: https://developer.nvidia.com/opticalflow-sdk. [Accessed 14 July 2020].
viii. PNY, "NVIDIA Tensor Cores," [Online]. Available: https://www.pny.com/professional/explore-our-products/learn-about-nvidia-quadro/nvidia-tensor-cores. [Accessed 14 July 2020].
viiii. G. Hyams and D. Malowany, "The Battle of Speed vs Accuracy: Single-Shot vs Two-Shot Detection Meta Architecture - Allegro AI," 8 March 2020. [Online]. Available: https://allegro.ai/blog/the-battle-of-speed-accuracy-single-shot-vs-two-shot-detection/. [Accessed 14 July 2020].
xii. NVIDIA Corporation, "NVIDIA TensorRT," [Online]. Available: https://developer.nvidia.com/tensorrt. [Accessed 14 July 2020].
xiii. NVIDIA Corporation, "NVIDIA Turing GPU Architecture Whitepaper".
xiv. T. Wilde, "PC graphic options explained - PCGamer," 13 March 2019. [Online]. Available: https://www.pcgamer.com/pc-graphics-options-explained/2/. [Accessed 14 July 2020].