CURTISS-WRIGHT

VPX3-4936

NVIDIA® Ampere® GPU and Configurable Gen4 PCIe Switch with Variants Developed In Alignment with the SOSA™ Technical Standard

Based on the Ampere embedded GA104 GPU (RTX-4500E), the VPX3-4936 is a 3U VPX powerhouse for intense processing and artificial intelligence (AI).

Designed and manufactured by WOLF Advanced Technology, the VPX3-4936's Ampere GPU contains 5888 CUDA cores, 184 Tensor cores dedicated to AI accelerated compute, and 46 Ray Tracing (RT) cores for superior rendering speeds. Designed to work in conjunction with Tensor RT™, CUDA, and cuDNN, the Ampere tensor cores add support for new date types as well as a sparsity feature. The next generation of Tensor Cores and RT cores in the Ampere provides twice the throughput, while new architecture of the streaming multiprocessors can power the CUDA cores to an increase of two to four times of throughput depending on the workload. The combination of the 8 nm manufacturing process and the Ampere architectural innovations yield significant power and processing improvements with up to 154 GFLOPS/W. This is twice the performance per slot compared to the previous Turing® generation's 86 GFLOPS/W.

Moving data quickly and efficiently is important with high-speed GPUs. The VPX3-4936 uses 256-bit GDDR6 memory with ECC. The 512 GB/s maximum memory bandwidth is the highest memory bandwidth of any embedded GPU. The Ampere is also the first NVIDIA GPU to support PCIe Gen4, which provides double the bandwidth of PCIe Gen3. Incorporating a PCIe Gen 4 switch, this module is configurable for compatibility with various OpenVPX™ slot profiles, including support for Non-Transparent Bridge (NTB) and daisy chain options.

The VPX-4936 variants support the SOSA SLT3-PAY-1F1U1S1S1U1U2F1H-14.6.11-0, the SOSA payload legacy slot profile 14.2.3, as well as the original OpenVPX configuration. While pin-compatible with Curtiss-Wright's VPX3-4935s, the VPX3-4936 enables system designers to boost GPGPU capabilities without increasing size, weight, and power (SWaP).

WOLF ADVANCED TECHNOLOGY

Key Features

- **NVIDIA GA104 (RTX4500E)** delivering 17.66 TFLOPS FP32 peak performance
- **Capable of 154 GFLOPS/W**
- **5888 CUDA® Cores for parallel processing**
- **184 Tensor cores** for dedicated AI Accelerated compute
- **46 RT Cores** for superior rendering speed
- **16 GB GRRD6** 256-bit memory
- **512 GB/S** Max memory bandwidth
- **Configurable PCIe® Gen4 x16 switch**
- **Operating power** configurable hard cap: 40-180W
- **SOSA and legacy OpenVPX variants**
- **4 simultaneous video outputs,** supporting DP, DVI, HDMI

Applications

- **ISR and EW** applications requiring the highest performance processing
- **SWaP-constrained deep learning** inference needing 2x-4x more throughput than the previous generation
- **High-performance RADAR, SIGINT, EO/IR,** sensor fusion, and autonomous platforms

©2022 Curtiss-Wright - All rights reserved. Specifications are subject to change without notice. All trademarks are property of their respective owners I D499.1022. This document was reviewed on 2022.10.07 and does not contain technical data.

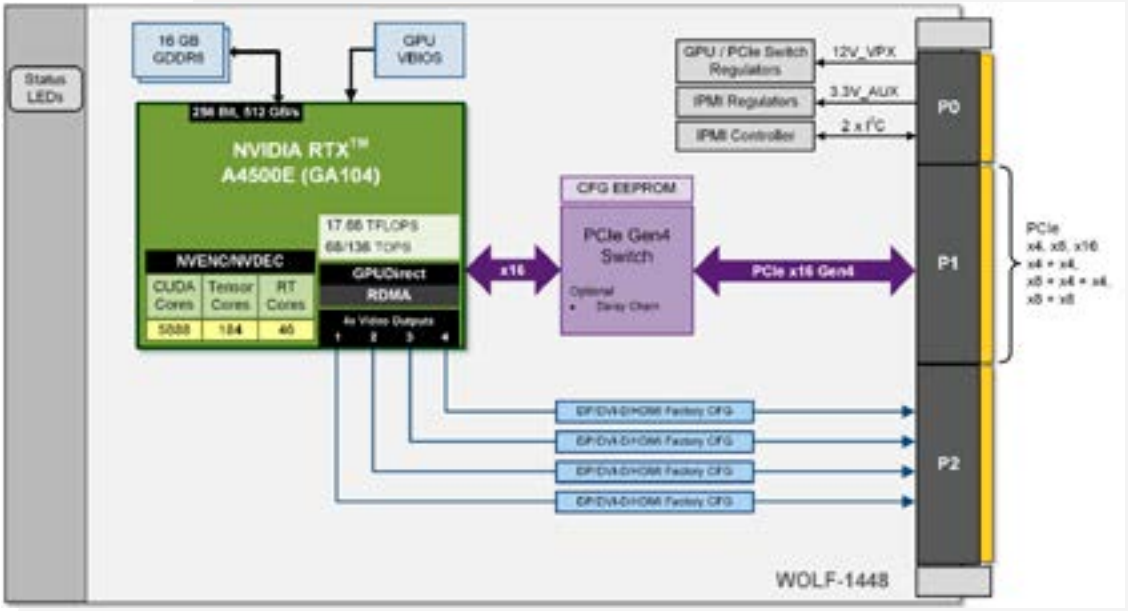

Figure 1: VPX3-4936 standard OpenVPX variant block diagram

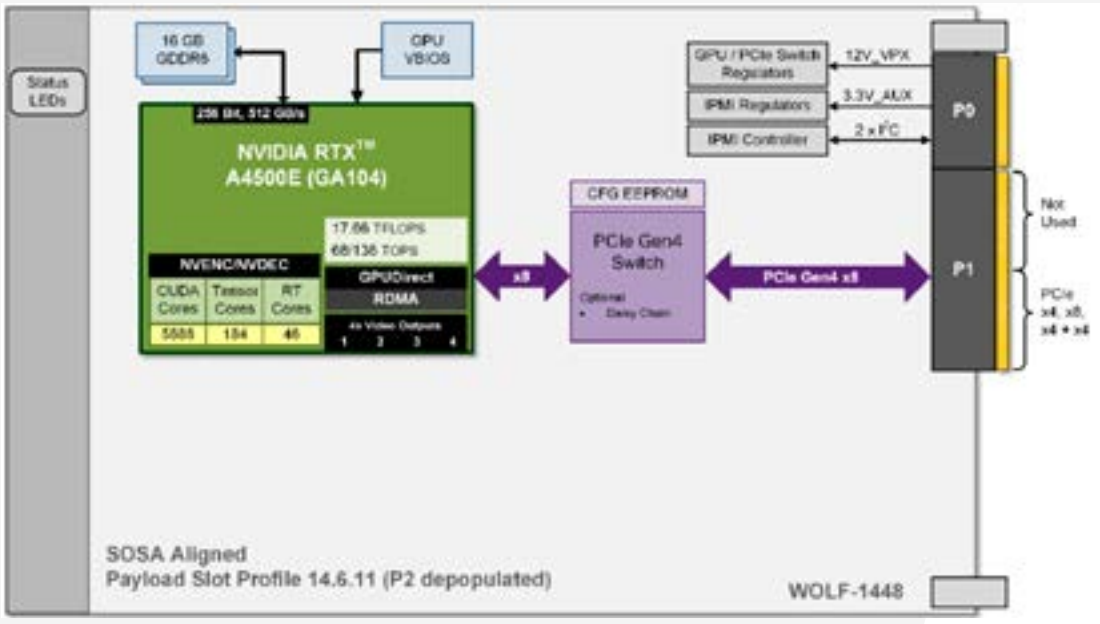

Figure 2: VPX3-4936 SOSA Aligned Variant Block Diagram

©2022 Curtiss-Wright - All rights reserved. Specifications are subject to change without notice. All trademarks are property of their respective owners I D499.1022. This document was reviewed on 2022.10.07 and does not contain technical data.

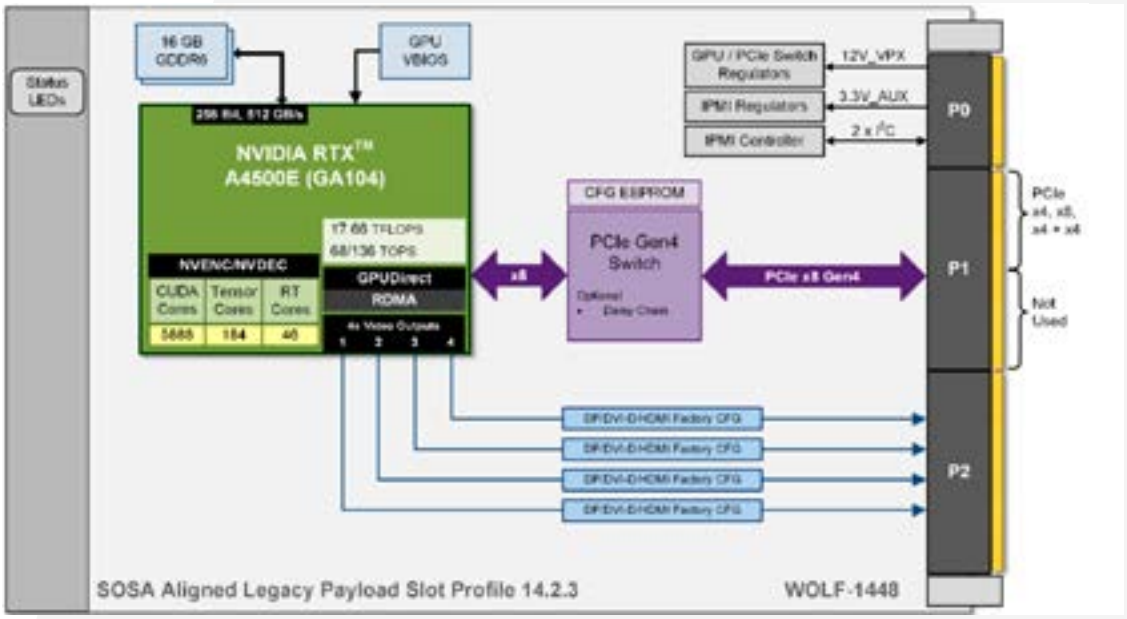

Figure 3: VPX3-4936 SOSA Aligned Variant Legacy Block Diagram

NVIDIA Ampere Streaming Multiprocessor (SM)

In the Ampere, each of the Streaming multiprocessors contains CUDA cores for FP and INT operations, Tensor Cores for AI, Ray Tracing (RT) cores for rendering, Texture Units, a register file, and L1/Shared Memory. The NVIDIA Ampere architecture is a major step forward in performance and efficiency over the previous Turing generation, including more flexible concurrent execution of floating point and integer streams. The Turing generation's SM had two primary data paths, one for FP32 processing while the other only processed integer operations. The Ampere architecture still supports two data paths with one dedicated to processing FP32 data while the other one can execute either FP32 or INT32 operations. NVIDIA estimates that 70% of all GPU applications uses only FP32. With this new architecture, those applications will have access to all the CUDA cores by using both data paths.

NVIDIA Tensor Cores for Artificial Intelligence and HPEC

Designed to speed up the tensor/matrix computations used for deep learning neural network training and inference operations, Tensor cores first became available in Volta GPUs that were not available for the embedded space. Turing GPUs (2nd generation Tensor cores) were enhanced for inferencing add INT8 and INT4 precision modes for workloads that tolerate quantization and do not require high precision. The third generation Tensor cores in the Ampere architecture support new features and datatypes that improve performance, efficiency, and programming flexibility. Examples include a new sparsity feature as well as TF32 and BF32 data types. Depending on the workload, these improvements provide two to four times more throughput than the previous generation.

NVIDIA provides CUDA-X AI and CUDA-X HPEC libraries which are designed to work with NVIDIA GPUs with Tensor cores to accelerate development of applications for AI and HPEC.

Hardware Accelerated Video Encode/Decode

The Ampere GPU includes the NVENC video encode (version 7.2) and NVDEC decode (version 5) hardware acceleration engine. Using the Ampere GPU for video encoding provides an efficient method to achieve real time 8K and 4K encoding without burdening the system CPU. The Ampere decoding engine includes support for several codecs, including AV1 hardware. The NVIDIA Video Codec SDK provides a complete set of APIs, samples and documentation for hardware accelerated video encode and decode.

©2022 Curtiss-Wright - All rights reserved. Specifications are subject to change without notice. All trademarks are property of their respective owners I D499.1022. This document was reviewed on 2022.10.07 and does not contain technical data.

Specifications

Form factor

- 3U OpenVPX
- SOSA
  - SLT3-PAY-1F1U1S1S1U1U2F1H-14.6.11-0
  - P2 depopulated
  - PCIe support P1B
  - PCIe Configurations: x4, x8, x4+x4
- SOSA Aligned Legacy Payload Slot
  - Slot Profile: 14.2.3
  - P2 populated with support for 4 video outs
  - PCIe support P1B
  - PCIe Configurations: x4, x8, x4+x4
- OpenVPX Legacy
  - P2 populated with support for 4 video outs
  - PCIe support P1
  - PCIe Configurations: x4, x8, x16, x4+x4, x8+x4+x4, x8+x8

GPU NVIDIA Ampere GA104 (RTXA45000E)

- 5888 CUDA Core, up to 17.66 TFLOPS of FP32 peak performance
- 184 Gen 3 Tensor Cores - 68 dense/134 sparse TOPS
- 48 Ray-Tracing (RT) Gen 2 Cores
- 46 Streaming Multiprocessors
- 16 GB GDD6
- Max Memory bandwidth: 512 GB/s
- Memory Width: 256-bit
- ECC Memory

PCIe

- GPU with PCIe x16 Gen4 interface
- Configurable PCIe Gen 4 switch x4, x8 or x16
- Daisy chain option supported
- NTB options supported

Hardware Accelerated Video Encode/Decode

- NVENC - 7th Gen with HEVC B-Frame Support
- NVDEC – 5th Gen with AVI
- Up to 8k support

Video Outputs

- 4x independent simultaneous video outputs
- Supported outputs: Display Port, DVI, DHMI
- Support for High Dynamic Range (HDR) video
- Display port 1.4a
  - 4K @240 Hz
  - 8K @ 60 Hz
- HDMI 2.1
  - 4K @240 Hz
  - 8K @ 60 Hz

Power

- Primary Power: Vs1 (12 V)
- 3.3V_AUX power
- Configurable GPU hard cap: 40-150W (Preliminary)

Environmental

- Rugged conduction-cooled
- -40°C to 85°C operating temperature
- Other environmental specifications are per WOLF Advanced Technology
- Humiseal 1B73 Conformal coating

Software Support

- Linux NVIDIA drivers
- CUDA Toolkit 11
- CUDA Compute version 8.6
- DirectX®12 Ultimate
- OpenCL™ 3.0
- OpenGL 4.6
- OpenGL ES 3.2
- Vulkan 1.0

©2022 Curtiss-Wright - All rights reserved. Specifications are subject to change without notice. All trademarks are property of their respective owners I D499.1022. This document was reviewed on 2022.10.07 and does not contain technical data.

VPX3-4936 Ordering information

Part Number	Variants
VPX3-4936-C141-001	• 3U OpenVPX module with GA104 GPU (RTX-4500E) • 5088 CUDA Cores, up to 17.66 TFLOPS peak theoretical • 184 Tensor Cores, 46 RT Cores • 16 GB GDDR6, 512 GB/Sec max bandwidth • Conduction-Cooled, “1.0” pitch, temperature range (-40 - 85°C) • PCIe Gen 4 x16 with switch • OpenVPX profile, default PCIe x16 configuration on P1 • Configurable power 50-150W • +12V, 4 display outputs: 3xDP, 1xDVI on P2
VPX3-4936-C141-101	• 3U OpenVPX module with GA104 GPU (RTX-4500E) • SOSA legacy payload Slot Profile: 14.2.3 • 5088 CUDA Cores, up to 17.66 TFLOPS peak theoretical • 184 Tensor Cores, 46 RT Cores • 16 GB GDDR6, 512 GB/Sec max bandwidth • Conduction-Cooled, “1.0” pitch, temperature range (-40 - 85°C) • PCIe Gen 4 x16 with switch • Default PCIe x8 on wafer 9-16 on P1 • Configurable power 50-150W • +12V, 4 display outputs: 3xDP, 1xDVI on P2
VPX3-4936-C241-102	• 3U OpenVPX module with GA104 GPU (RTX-4500E) • SOSA Profile: SLT3-PAY-1F1U1S1S1U1U2F1H-14.6.11-0 • 5088 CUDA Cores, up to 17.66 TFLOPS peak theoretical • 184 Tensor Cores, 46 RT Cores • 16 GB GDDR6, 512 GB/Sec max bandwidth • Conduction-Cooled, “1.0” pitch, temperature range (-40 - 85°C) • PCIe Gen 4 x16 with switch • Default PCIe x8 on wafer 9-16 on P1 • Configurable power 50-150W • +12V, No video out

©2022 Curtiss-Wright - All rights reserved. Specifications are subject to change without notice. All trademarks are property of their respective owners I D499.1022. This document was reviewed on 2022.10.07 and does not contain technical data.