

NVIDIA TURING 22 TFLOPS, 768 TENSOR CORES,

PRELIMINARY INFORMATION

KEY FEATURES

- Two NVIDIA Quadro Turing TU104 with 22 TFLOPS, 6144 CUDA Cores, 768 Tensor Cores, 96 RT Cores
- 32 GB GDDR6 256-bit memory
- WOLF FGX provides support for additional video formats such as SDI and CVBS
- Module power: 100-300W

ADDITIONAL FEATURES

- Eight DisplayPort 1.4 digital video outputs:
 - High Dynamic Range (HDR) video
 - 4K at 120Hz, 8K at 60Hz; 10-bit color depth
 - HDMI and DVI options
- Four SDI outputs, four CVBS outputs
- Turing GPGPU parallel processing:
 - 6144 CUDA® cores
 - CUDA Toolkit 10, CUDA Compute capability 7.5
 - OpenCL™ 1.2, DirectX® 12, OpenGL 4.6, OpenGL ES 3.2, Vulkan™ 1.0
- 768 Turing Cores for advanced AI inference
- 32 GB GDDR6 memory provides twice the bandwidth of the previous generation of GDDR5 memory, up to 448 GB/s to each GPU
- HEVC (H.265) and AVC (H.264) Turing NVENC and NVDEC hardware acceleration with up to 8K encode resolution and B-frame support
- PCI Express x16
- Daisy Chain and NTB Options Available
- NVLink® 2.0 high-speed GPU interconnect option; up to 50GB/s of bidirectional bandwidth between the dual GPUs on the module
- Windows and Linux drivers

SPECIFICATIONS

- High level of ruggedization:
 - Rugged Conduction-cooled or Air-cooled
 - Operating temperature: -40° to +85°C
 - Vibration (sine wave): 10G peak, 5 - 2000Hz
 - Shock: 30G peak for air-cooled, 40G peak for conduction-cooled
- Dimensions: 160mm x 233mm x 25.4mm
- Weight: To be Announced
- ANSI/VITA 48, 65 (VPX-REDI, OpenVPX)

OVERVIEW

The VPX6U-RTX5000E-DUAL-CV module includes two NVIDIA® Quadro® Turing™ RTX5000 embedded GPUs and a WOLF FGX in a rugged 6U VPX module. The RTX5000 includes CUDA cores for parallel processing, Tensor cores for dedicated AI inference and ray tracing cores for superior rendering speeds. The WOLF FGX provides video conversion to formats which are not native to the GPU, such as SDI and analog formats.

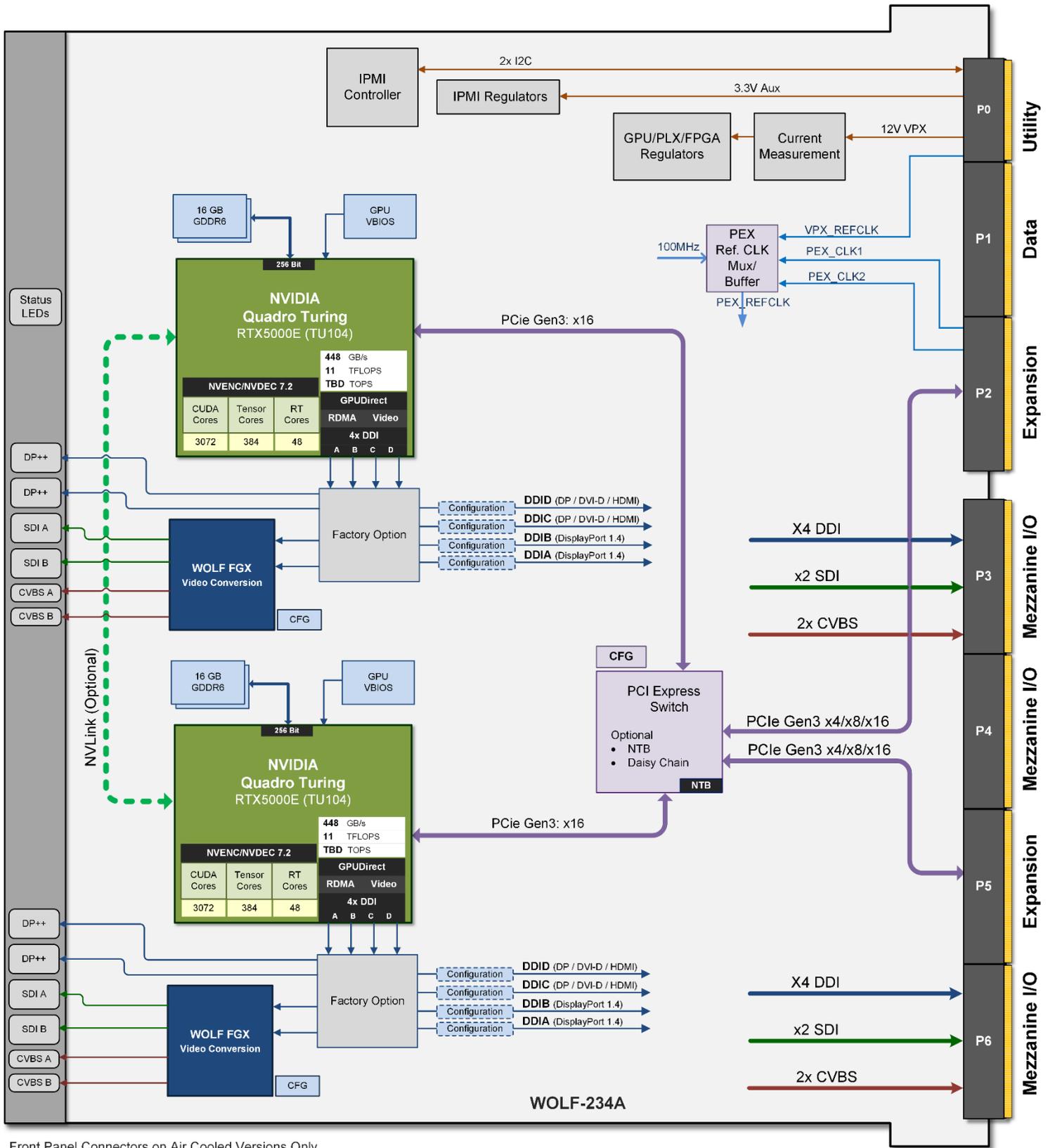
Getting data into and out of the module is an important consideration for this high-speed GPU. Support for GDDR6 memory provides twice the bandwidth of the previous generation's GDDR5 memory.

The Turing GPU with its Tensor cores provides this module with the underlying architecture required for an efficient AI inference engine. Intended to work in conjunction with TensorRT, CUDA and CuDNN, the Turing Tensor Core design adds INT8 and INT4 matrix operations, while continuing to support FP16 for higher precision workloads.

Unlocking the best performance requires the best cooling capability. WOLF's advanced cooling technology is designed to move heat using a low weight, high efficiency pipeline from the GPU die to the wedgelocks.



This datasheet is preliminary and is subject to change



Front Panel Connectors on Air Cooled Versions Only

This datasheet is preliminary and is subject to change

PRELIMINARY INFORMATION

NVIDIA TURING STREAMING MULTIPROCESSOR (SM)

The NVIDIA Turing architecture provides a 50% improvement in delivered performance per CUDA core compared to the Pascal generation. This is due to the new Turing Streaming Multiprocessor's independent integer datapath, allowing execution of concurrent integer and floating-point instructions, and the redesigned memory path which provides two times the bandwidth and more than two times the capacity for common workloads.

NVIDIA TENSOR CORES FOR ARTIFICIAL INTELLIGENCE AND HPC

Tensor Cores are designed to speed up the tensor / matrix computations used for deep learning neural network training and inferencing operations. Turing GPUs include a new version of the Tensor Core design that has been enhanced for inferencing. Turing Tensor Cores add new INT8 and INT4 precision modes for inferencing workloads that can tolerate quantization and don't require FP16 precision.

NVIDIA provides CUDA-X AI and CUDA-X HPC libraires which are specialized libraires built on top of CUDA. They have been designed to work with NVIDIA Tensor Core GPUs to provide the tools needed to accelerate development of applications for AI and HPC.

HARDWARE ACCELERATED VIDEO ENCODE/DECODE

The RTX5000E chip includes the latest generation video encode/decode hardware acceleration engine (version 7.2). This adds support for HEVC (H.265) 8K encoding at 30 fps and B-Frame support. It also provides up to 25% bitrate savings for HEVC and up to 15% bitrate savings for AVC (H.264). Using the Turing encoding engine for video encoding provides an efficient, high quality method to achieve real time 8K and 4K encoding without burdening the system CPU.

As with previous versions of the engine, NVENC supports CBR and VBR rate control, programmable intra-refresh for error resiliency, and a motion estimation (ME) only mode. The NVIDIA Video Codec SDK provides a complete set of APIs, samples and documentation for hardware accelerated video encode and decode on Windows and Linux.

DESIGNED FOR SYSTEM INTEGRATION

The VPX architecture is diverse, spanning custom backplanes, an ambiguous system specification and differing input and output methodologies. That is precisely why WOLF modules come with factory configuration options to solve system integration challenges.

This WOLF module has been designed to include a WOLF FGX to support additional video output protocols that are not native to the Turing GPU. Through MCOTS services, the number of outputs and the protocols used can be changed to meet the requirements of many different system architectures.

This module has been designed to support VPX REDI (VITA 48) and OpenVPX (VITA 65). Other modules from the same family can be configured for SOSA Aligned slot profiles.

This datasheet is preliminary and is subject to change

PRELIMINARY INFORMATION

ORDERING CODES

The following table defines series of common order codes for the VPX6U-RTX5000E-DUAL-CV module. The asterisks denote characters of the part number that are defined based on common configuration options. Some common configuration options for this module are:

- Display Interfaces
- Conformal Coating Type
- COTS, MCOTS or Locked
- PCI Express Bus P1 Configuration

Ordering Number	Description
6U VPX Turing RTX5000 Single Slot Configurations	
234A23-F***VPX6vA0	6U VPX, Air Cooled, 1", Dual NVIDIA Turing RTX5000, WOLF FGX
234A33-F***VPX6vA0	6U VPX, Conduction Cooled, 1", Dual NVIDIA RTX5000, WOLF FGX
234A23-F***VPX6vA0	6U VPX, Air Cooled, 1", Single NVIDIA Turing RTX5000, WOLF FGX
234A33-F***VPX6vA0	6U VPX, Conduction Cooled, 1", Single NVIDIA RTX5000, WOLF FGX

Contact Sales for the latest Ordering Numbers and available options

MANUFACTURING AND QUALITY

ASSURANCE

WOLF designs modules to pass the following environmental standards:

- MIL-STD-810 (United States Military Standard for Environmental Engineering Considerations and Laboratory Tests)
- MIL-HDBK-217 (Reliability Prediction of Electronic Equipment)
- RTCA DO-160 (Environmental Conditions and Test Procedures for Airborne Equipment) on request

WOLF complies with the following management systems:

- AS9100D: Quality Management System - Requirements for Aviation, Space and Defense Organizations (certified)
- ISO 9001:2015: Quality management systems (certified)
- AS5553: Counterfeit Electronic Parts; Avoidance, Detection, Mitigation, and Disposition (compliant)
- NIST SP 800-171: Protecting Controlled Unclassified Information in Nonfederal Systems (compliant)

Boards are manufactured to meet the following standards:

- IPC-A-610 CLASS 3 (Acceptability of Electronic Assemblies)
- IPC 6012 CLASS 3 (Qualification and Performance Specification for Rigid Printed Boards, Class 3 for High Reliability Electronic Products)
- IPC J-STD-001 (Requirements for Soldered Electrical and Electronic Assemblies)



This datasheet is preliminary and is subject to change