Introducing the Alveo UL3524 FPGA Accelerator

Hamid Salehi

Director of Product Marketing, Data Center

Adaptive & Embedded Computing Group, AMD

Alveo™ UL3524 Adaptable Accelerator

FPGA Accelerator for Ultra-Low Latency Trading

7X Lower Latency vs. Previous Gen[1]

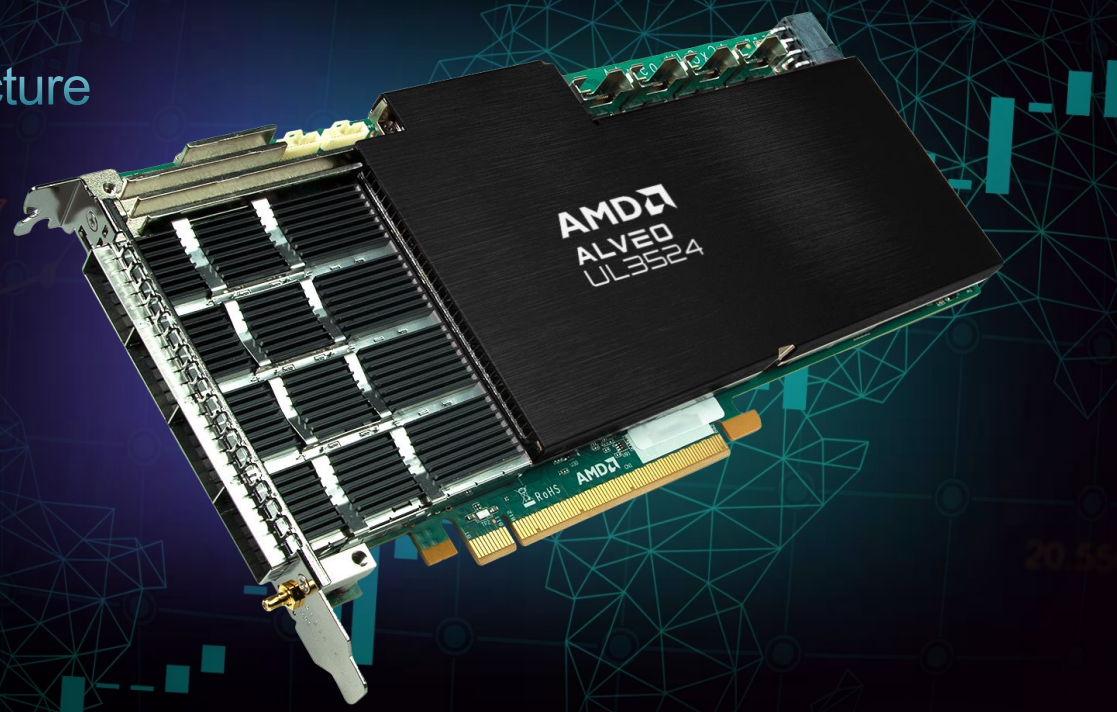
Purpose-Built FPGA with new transceiver architecture

HW Flexibility and AI-Enabled Strategies

Vivado™ flow and opensource low latency AI

Diverse ULL Applications

Trading, Pre-Trade Risk, Market Data Delivery



In Volume Production

1: Based on simulation comparison between Virtex™ UltraScale+™ ultra-low latency GTF transceivers and GTY transceivers
Not a STAC benchmark

AMD

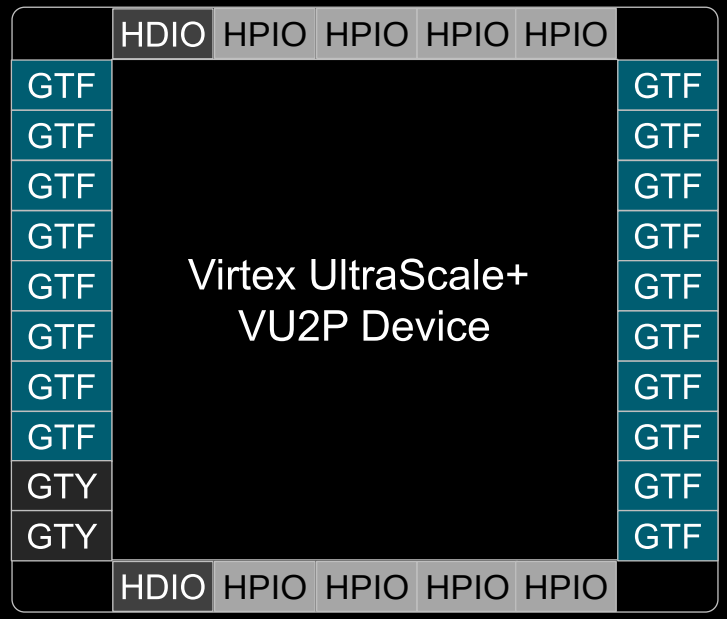
Based on New Virtex UltraScale+ VU2P FPGA

- 72 state-of-the-art, ultra-low latency GTF transceivers, purpose-built for ULL trading
- Monolithic device: single SLR (Super-Logic Region), 787K LUTs of FPGA fabric
- High performance: 644MHz FPGA fabric clock,16-bit at 10G, 40-bit at 25G

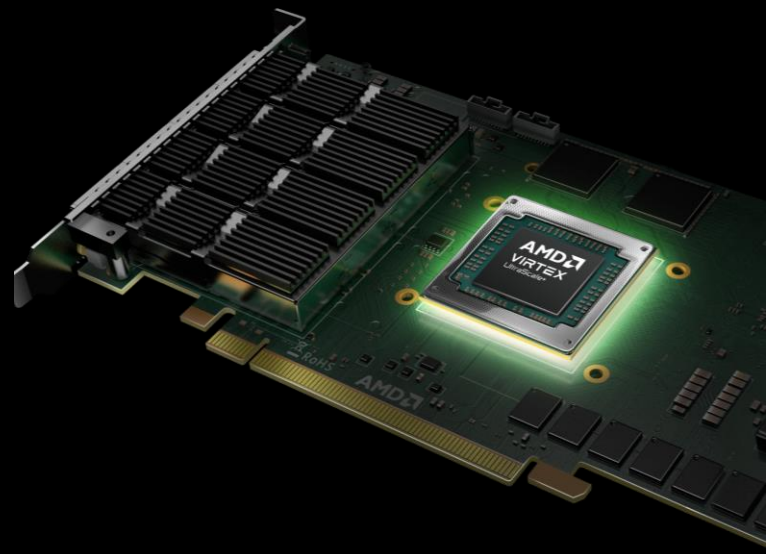
Device Specifications

FEATURE	SPECIFICATION
LUTs	787K
Embedded RAM	156Mb
DSP	1,680
GTF (28G) ULL Transceivers	72
GTY (32G) Transceivers	8
HPIO	520
PCIe® Gen4 x8	1
PCIe® Gen3 x8	1
Package	47.5 x 47.5mm

1 SLR (Monolithic Device)



Powering the Alveo™ UL3524



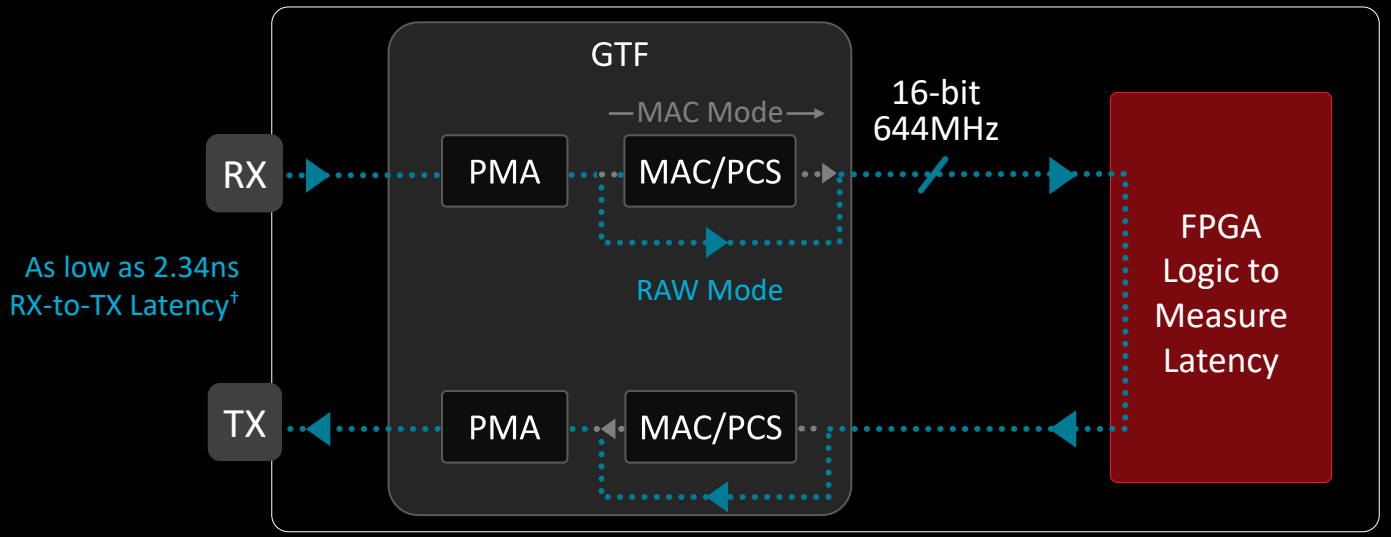
AMD

Not STAC benchmarks

Ultra-Low Latency GTF Transceiver Architecture

- Two Modes for ULL GTF transceiver:
  1) **RAW Mode** (bypasses PCS/MAC): 2.34ns latency††
  2) **MAC Mode** (includes PCS/MAC): 5.44ns latency †††
- Reference design available for benchmarking

GTF Loopback Benchmark Design



Latency Performance at 10.3Gb/s

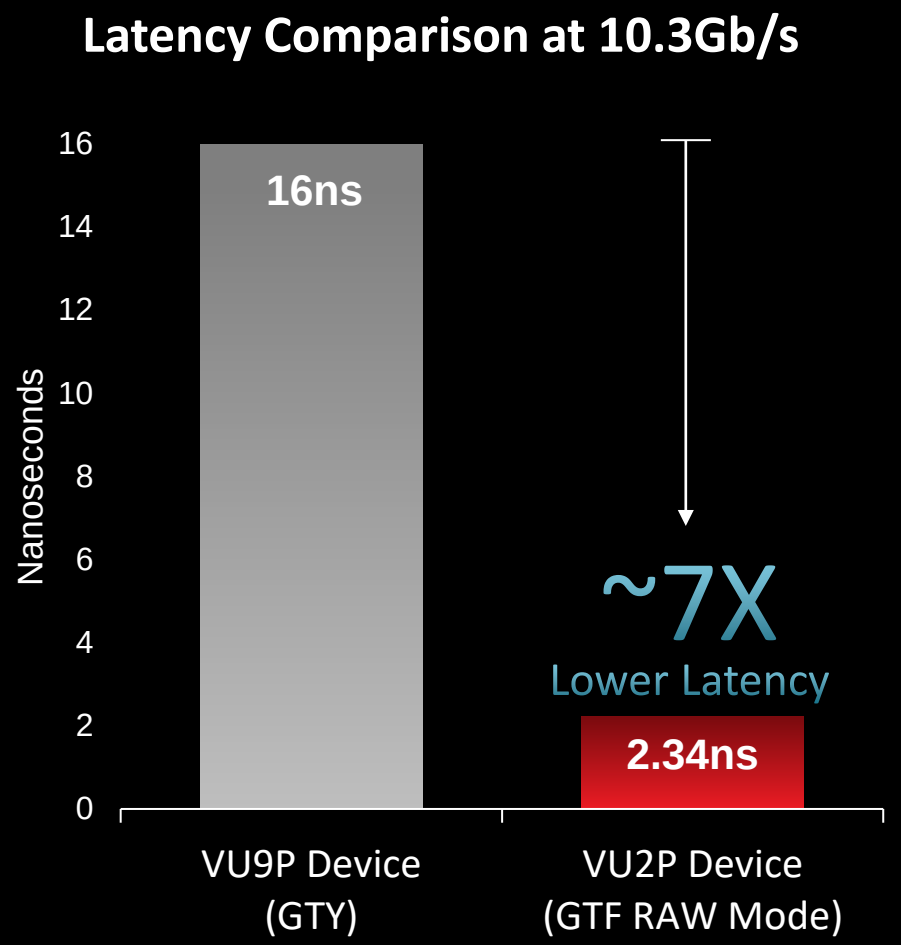
RAW Mode (bypass MAC/PCS)	2.34ns††
MAC Mode	5.44ns†††

Note: Latency measurement does not include protocol overhead, protocol framing, programmable logic (PL) latency, TX PL interface setup time, RX PL interface clock-to-out, package flight time, and other sources of latency. See RAW Mode †† and MAC Mode ††† endnotes for more details.

Not STAC benchmarks

Breakthrough Architecture over Previous Generation

- **SerDes Architectural Optimizations**
  - Native support for required bit-widths
  - Integrates only the features required for target applications
- **Retains key features of UltraScale+™ architecture**
  - Same high speed clock topology & low jitter
  - Same equalization architecture (for signal integrity)
- **End-result is breakthrough performance**
  - ~7X lower latency vs. GTY transceivers at 10.3Gb/s[1]
  - 7.6X lower latency vs. GTY transceivers at 25.8Gb/s[1]



1: Based on simulation comparison between Virtex™ UltraScale+™ GTY transceivers and ultra-low latency GTF transceivers

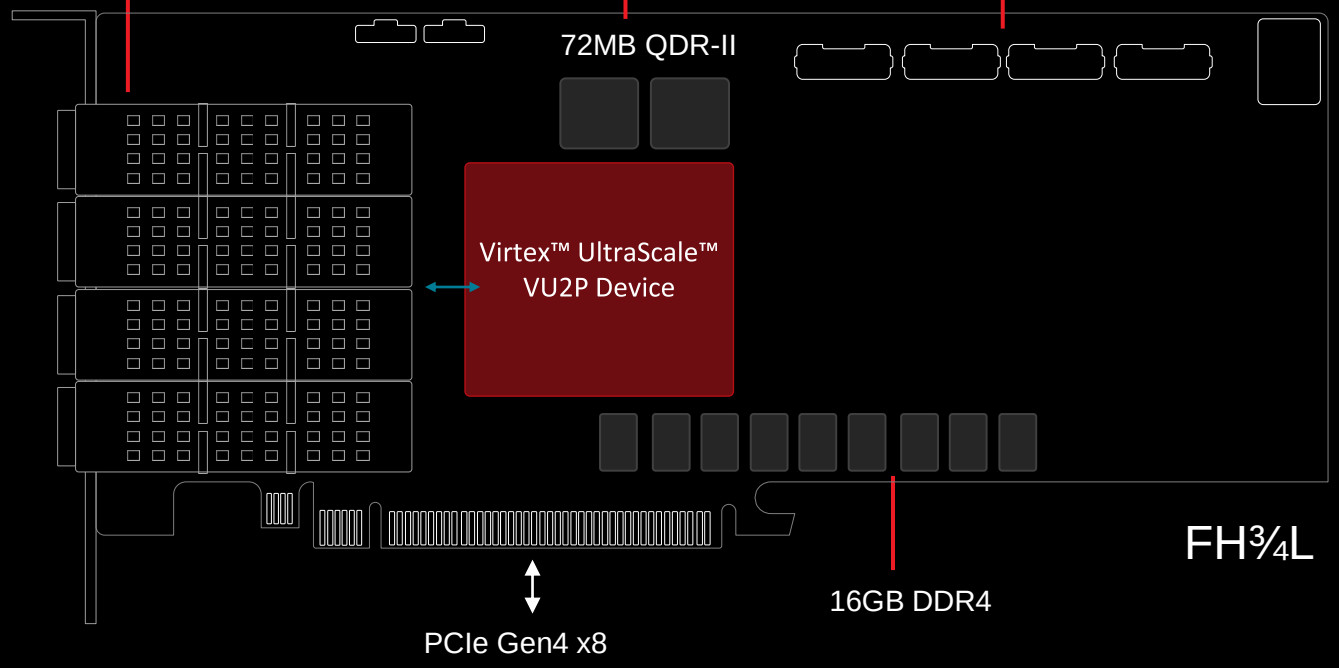
AMD

Alveo™ UL3524 Card Specifications

4x25G Networking
- QSFP-DD
- 32x 10/25G Ports

On-Board Memory
- 16GB DDR
- 72MB QDR-II

Expansion Ports
- Connect Multiple Cards

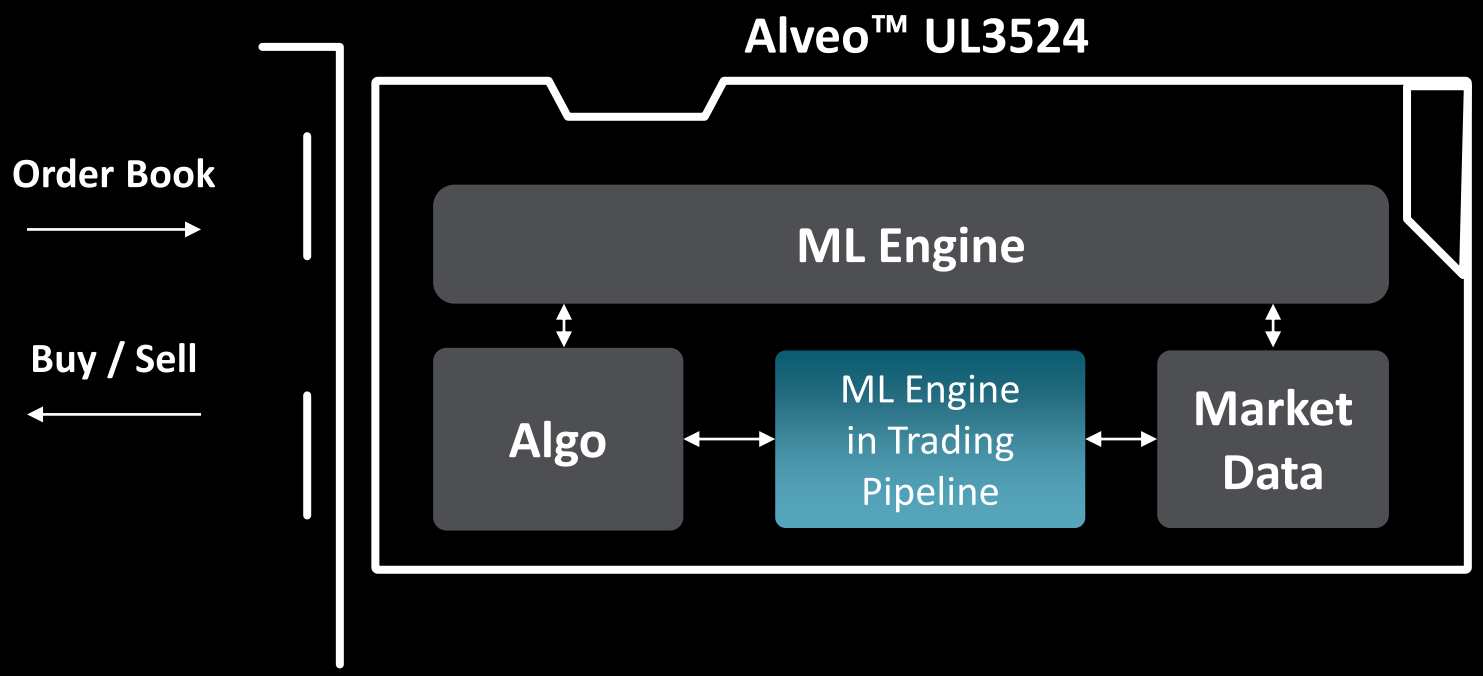


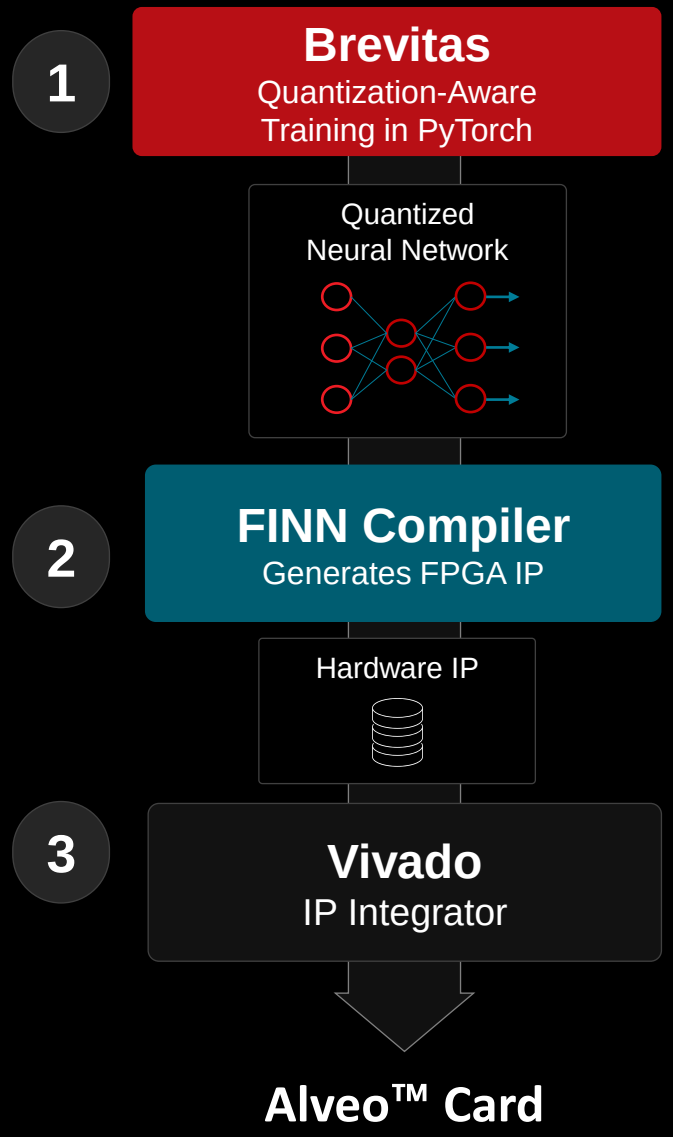
FEATURES	SPECIFICATION
On-Board Memory	• 16GB DDR4-2666 • 72MB QDR II+, 550Mhz
Network Interface	4x QSFP DD (32x 10/25G Ports)
Expansion Ports	• 4 ARF6 32x10/25G ports • Connects multiple cards • 2 Pico-Clasp connectors for sideband
Form Factor	Full-height, ¾ Length (FH¾L) Single Slot
PCIe® Interface	PCIe® Gen4 x8
Power	• 180W Electrical • 125W TDP • Passive cooling
Product SKU	A-UL3524-P16G-PQ-G

AMD

AI-Enabled Trading Strategies: Brevitas and FINN Framework

- **AMD Opensource Projects for Streaming AI accelerators**
  1) Brevitas PyTorch Library for neural network quantization / optimization
  2) FINN compiler generates FPGA IP from neural network
  3) System integration with Vivado™ IP Integrator
- **Combine with ULL transceivers for fast trade execution**
  - Example design available on GitHub





Learn More

Web

www.amd.com/ul3524

GitHub

Benchmark Design and AI Example

Talk to an Expert

ull-fintech@amd.com

Now Shipping and in Production

AMD


AMD

Disclaimer & Attribution

Timelines, roadmaps, and/or product release dates shown in these slides are plans only and subject to change.

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD’s products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale.

Endnotes

†Based on simulation comparison between Virtex™ UltraScale+™ GTY transceivers and ultra-low latency GTF transceivers

††Testing conducted by AMD Performance Labs as of 8/16/23 on the Alveo UL3524 accelerator card, using Vivado™ Design Suite 2023.1 and running on Vivado Lab (Hardware Manager) 2023.1. Based on the GTF Latency Benchmark Design configured to enable GTF transceivers in internal near-end loopback mode. GTF TX and RX clocks operate at same frequency of ~644MHz with a 180 degrees phase shift. GTF Latency Benchmark Design measures latency in hardware by latching value of a single free running counter. Latency is measured as the difference between when TX data is latched at the GTF transceiver and when TX data is latched at the GTF receiver prior to routing back into the FPGA fabric. Latency measurement does not include protocol overhead, protocol framing, programmable logic (PL) latency, TX PL interface setup time, RX PL interface clock-to-out, package flight time, and other sources of latency. Benchmark test was run 1,000 times with 250 frames per test. Cited measurement result is based on GTF transceiver “RAW Mode”, where PCS (physical medium attachment) of the transceiver passes data ‘as-is’ to FPGA fabric. Latency measurement is consistent across all test runs for this configuration. System manufacturers may vary configurations, yielding different results. ALV-10

†††Testing conducted by AMD Performance Labs as of 8/16/23 on the Alveo UL3524 accelerator card, using Vivado Design Suite 2023.1 and running on Vivado Lab (Hardware Manager) 2023.1. Based on the GTF Latency Benchmark Design configured to enable GTF transceivers in internal near-end loopback mode. GTF TX and RX clocks operate at same frequency of ~644MHz with a 180 degrees phase shift. GTF Latency Benchmark Design measures latency in hardware by latching value of a single free running counter. Latency is measured as the difference between when TX data is latched at the GTF transceiver and when TX data is latched at the GTF receiver prior to routing back into the FPGA fabric. Latency measurement does not include protocol overhead, protocol framing, programmable logic (PL) latency, TX PL interface setup time, RX PL interface clock-to-out, package flight time, and other sources of latency. Benchmark test was run 1,000 times with 250 frames per test. Cited measurement result is based on GTF transceiver “MAC Mode”, where PCS logic is used to encode and decode data according to a standard protocol, such as 10 Gigabit Ethernet2. 5.44ns represents lowest latency measured, with average latency 6.463ns across 1,000 test runs. System manufacturers may vary configurations, yielding different results. ALV-11