A UCIe Optical I/O Retimer Chiplet for AI Scale-up Fabrics

Pavan Bhargava, Sidney Buchbinder, Đorđe Gluhović, Kaisarbek Omirzakhov, Ayan Biswas, Daniel Jeong, Clay Pearson, Austin Katzin, Eric Jan, Derek Van Orden, Hayk Gevorgyan, Anatoly Khilo, Neil Sapra, Woorham Bae, Erik Anderson, Sung-Jin Kim, Vipin Jasoria, Jie Ding, Norman Chan, John Fini, Chen Sun
Ayar Labs, Inc., 695 River Oaks Pkwy, San Jose, CA 95134
Author e-mail address: pavan@ayarlabs.com

Abstract: We demonstrate a UCIe Optical I/O Retimer for scale-up AI fabrics. It achieves 1.024Tbps of bidirectional bandwidth per optical port from one package to another, delivering an aggregate bandwidth of 8.192Tbps.

The performance and profitability of key AI applications scale directly with the size of the compute domain that is interconnected with high-bandwidth scale-up fabric, which is currently confined to a single rack. This constraint is due to the limited reach of the electrical links used in these fabrics and the inability of traditional pluggable interconnects to meet the bandwidth density, energy-efficiency, cost, and reliability metrics required for this high-bandwidth domain. In order to continue scaling the performance and profitability of emerging AI applications, the scale-up domain must be expanded into a multi-rack system. This breakthrough necessitates a new I/O technology that can match and exceed the energy efficiency of copper interconnects while providing longer reach, higher bandwidth density, and lower latency.

In this work, we demonstrate a UCIe Optical I/O Retimer – a new I/O technology that enables package-to-package, multi-rack system connectivity for scale-up AI fabrics (Fig. 1a) – by combining the energy-efficiency and high-bandwidth density of the electrical UCIe interface and wavelength-division multiplexed (WDM) Optical I/O. This work builds upon [2] and [3], demonstrating the capability for micro-ring based WDM systems to double in aggregate bandwidth approximately every 2 years. The architecture follows the UCIe Retimer [1] specifications:

1. An SoC on *Package 0* (*UCIe Die 0*) connects via UCIe to a UCIe Retimer on *Package 0* (*UCIe Retimer 0*)
2. *UCIe Retimer 0* connects to UCIe Retimer on *Package 1* (*UCIe Retimer 1*) using *Off-Package Interconnect*
3. *UCIe Retimer 1* is connected via UCIe to a UCIe Die on *Package 1* (*UCIe Die 1*).
4. The end-to-end protocol is defined by the SoC ASIC. *UCIe Retimer 0/1* implement a retimer protocol
5. Optical I/O is used as the Off-Package Interconnect to bridge the UCIe protocol between two SoC sockets.

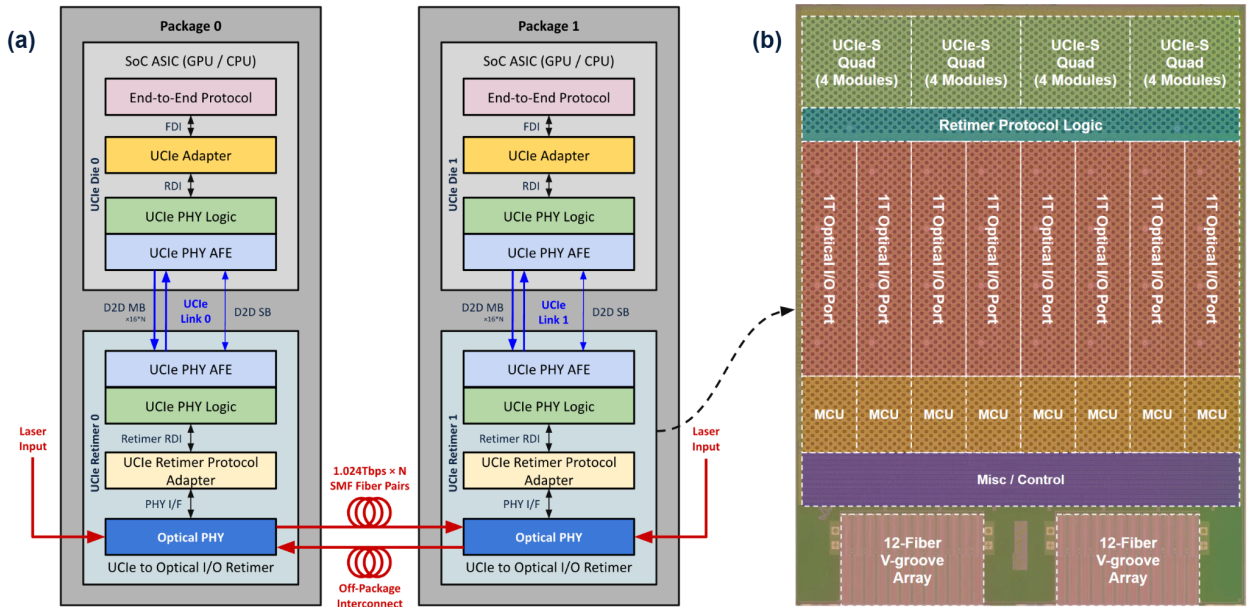

Fig. 1: (a) UCIe Optical I/O retimer system block diagram for a scale-up AI use case. (b) 8T UCIe to Optical I/O Retimer die photo

We implement the UCIe Retimer dies as an 8Tbps Optical I/O chiplet (Fig. 1b). The chiplet consists of eight optical I/O ports at 1.024Tbps per duplex single mode fiber (SMF) fiber pair and 16 ×16 UCIe Standard Package (UCIe-S)

modules (or 8 ×32 stacked modules) at 512Gbps per module grouped into four module quads. The chiplet incorporates UCIe retimer protocol logic and maps UCIe traffic to optical ports. The chiplet integrates all circuits and optical devices, including those needed to address polarization diversity over SMF. The external multi-wavelength, multi-port WDM laser source (ELS) supplies the optical power needed for all optical ports in each retimer.

The retimer chiplet has a 9.4mm beachfront, limited by the 1.143mm pitch of a stacked UCIe-S module [1] and minimum edge keepouts. The chiplet is fabricated in GlobalFoundries Fotonix™ process technology and bumped with copper pillars at 123µm pitch over the UCIe-S PHY and 128µm pitch elsewhere across the chip. The 16 UCIe-S x16 modules are implemented in a stacked UCIe-S module configuration, with each module running at 16 GT/s across 16Tx/16Rx lanes. The protocol logic implements lane deskewing, UCIe multi-module PHY logic (MMPL), clock PPM compensation, and firmware-defined management message transport which can operate over the optical link. Consistent with UCIe retimer specifications, the protocol logic is agnostic to the mainband traffic and is compatible with end-to-end error correction used by the end-to-end protocol (such as CRC/retry or FEC). The unidirectional end-to-end latency from *UCIe Link 0* to *UCIe Link 1* is approximately 25ns[1].

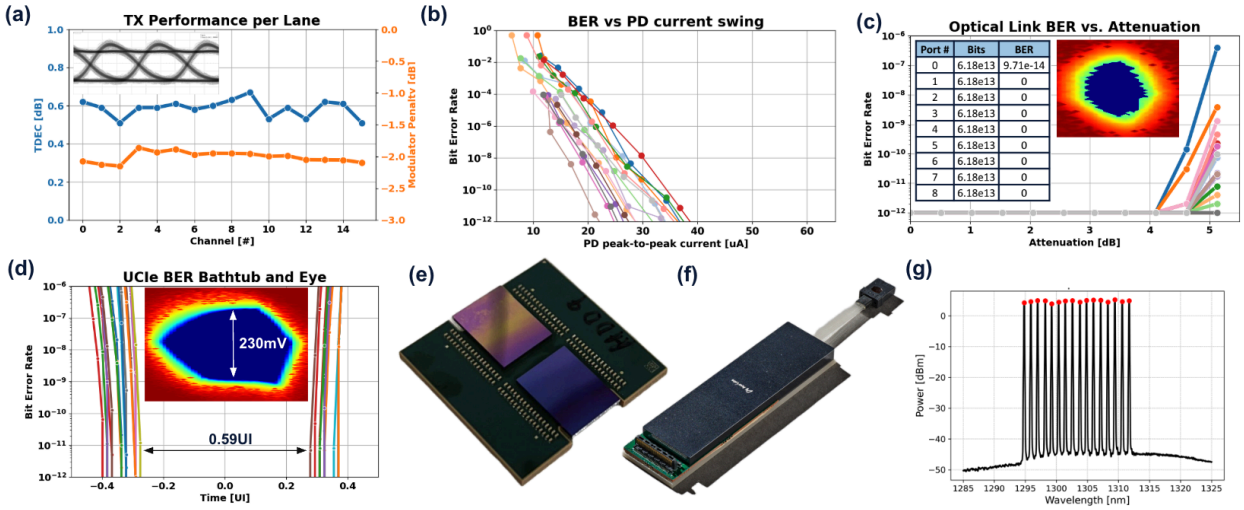

Fig. 2: Characterization results for a 16λ optical port PRBS7 transmitter: optical eye diagram, TDEC and modulator penalty (a), receiver BER vs. input photocurrent sensitivity (b), measured 16λ link BER vs. added optical attenuation between Tx/Rx, BER across ports and receiver eye scan (c). Aggregate UCIe module BER bathtub and eye diagram (d), Dual-die package test platform used for all measurements (e), 16λ 8-port ELS module (f) and ELS output spectra (g).

The optical port architecture is derived from that of [2], enhanced for a tighter channel spacing of 200GHz aligned to the CW-WDM MSA, doubled channel count, and data-path polarization diversity over SMF fiber. Fig. 2 shows the measured characteristics of individual components[2]. Microring modulators and driver circuits are used to modulate bits on each of the 16λ at 32Gbps NRZ (Fig. 2a). Microring-based filters and receiver circuits perform the 16λ demultiplexing and resolve the modulated optical signal back into digital data. The sensitivity of each receiver lane is shown in Fig. 2b, where all lanes meet $<10^{-12}$ BER at less than $<40\mu A_{pp}$ input current. When operating as a full 16λ × 32Gbps link, firmware running on embedded microcontrollers work in conjunction with on-chip feedback circuits to initialize and maintain microring locking to the incoming laser grid. BER characterization of a full 16λ × 32Gbps link with 4dBm/λ laser power and added optical attenuation (Fig. 2c) shows >4dB link margin at $<10^{-12}$ BER, closing links at 1.024Tb/s per duplex SMF pair. All 8 ports of the retimer achieve a similar 16λ link BER.

For UCIe and system characterization, we construct dual-die packages (DDPs) with two dies co-packaged together on an 8-2-8 organic substrate (Fig. 2e). The two chips on the DDP have an edge-to-edge separation of 3mm. Aggregate (all UCIe lanes in a module) eye diagrams taken *in-situ* across the UCIe interface indicate >0.5UI eye opening (Fig. 2d) and are error-free for $>10^{16}$ bits in longer running tests, indicating significant link margin.

[1]Includes the latency through UCIe Rx, Retimer Protocol, Optical Tx, Optical Rx, Retimer Protocol, UCIe Tx. Optical time of flight through fiber is not included, but is estimated at an additional ~5ns per meter of fiber.
[2]Where applicable, the pseudo-random bit sequences (PRBS) used for Tx/Rx characterization are generated/checked by on-chip built-in self-test (BIST) circuits.

The 16λ 8-port ELS (Fig. 2f) produces 128λ aggregated across 8 fiber ports (16λ/port). The laser combines a 16λ DFB laser array with a passive planar lightwave circuit (PLC) that performs the multiplexing and splitting to an 8-fiber output array, with each output fiber containing all 16λ.

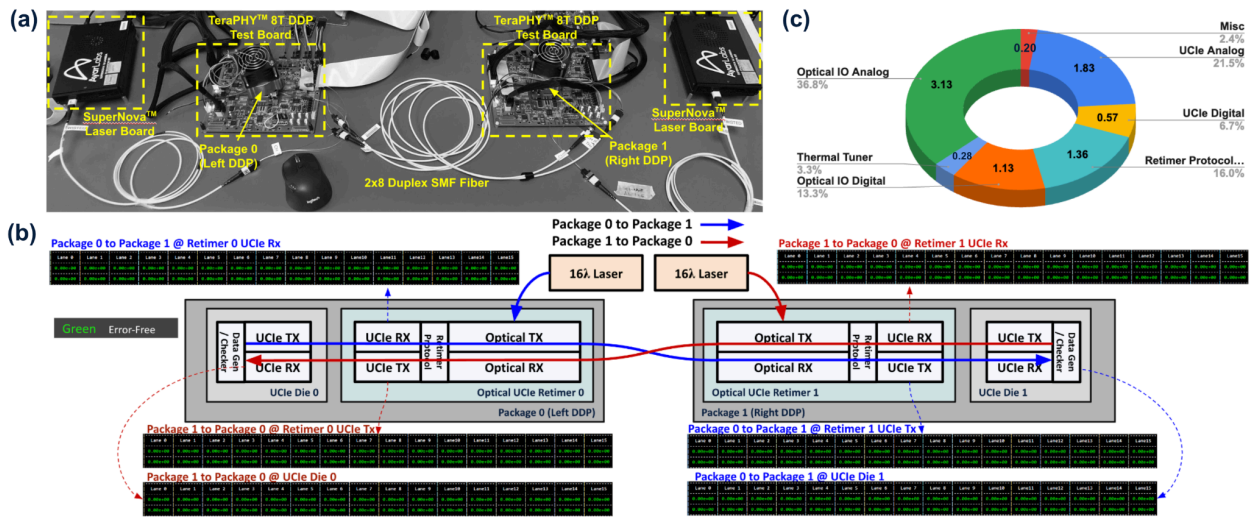

Fig. 3: System demonstration test setup (a). UCIe Optical I/O retimer system demonstration with measured error-free performance annotated at each point in the duplex link package-to-package links (b). Measured chip power breakdown (c)

Full system measurements (Fig. 3) are conducted with package-to-package duplex links using two DDPs; one of two dies in each DDP acts as a *UCIe (Optical I/O) Retimer* and the second die in each DDP acts as a *UCIe Die*, emulating a host SoC generating UCIe traffic in raw streaming mode. In each direction of the duplex link, traffic is generated by the *UCIe Die* in one package and consumed by the *UCIe Die* in the other package, with bits passing across three interfaces (two UCIe links and one optical link).

Fig. 3b shows a demonstration of end-to-end full duplex UCIe traffic over optics in raw streaming mode. The retimer protocol is enabled on the *UCIe Retimer* dies to perform clock PPM compensation across the two packages. BIST circuits in the protocol logic measure BER at three points along the way in each direction, allowing the system to pinpoint bit-errors introduced by each of the three interfaces. In this demonstration, one optical port and two UCIe modules are running in full duplex simultaneously on each *UCIe Retimer* die, achieving 1.024Tbps of bidirectional bandwidth. End-to-end error-free performance is achieved across both the optical and UCIe interfaces.

Measured power consumption is shown in Fig. 3c, with the breakdown of sub-component power estimated from pre-silicon data. The total power of the optical Tx/Rx including all analog, digital, optical control systems is 4.5pJ/b. UCIe interface power of 2.4pJ/b (with optional Rx termination resistor enabled) is higher than the <1pJ/b target for 16GT/s UCIe-S [1], owing to the 1V supply in this process as compared to the 0.7V target typical for process nodes implementing UCIe-S. At an average output power of 3dBm/λ and operating at 32Gbps/λ, the ELS energy-efficiency is 1pJ/b.

In conclusion, we have developed and demonstrated an 8 Tbps UCIe Optical I/O Retimer capable of bridging the UCIe interface across two dies in separate packages to address the needs of AI scale-up applications. To our knowledge, this is the first demonstration of UCIe over an optical off-package interconnect.

Acknowledgements: The authors thank the full Engineering, Operations, and Manufacturing teams at Ayar Labs for all their help in making this work possible.

References

[1] Universal Chiplet Interconnect Express (UCIe) Specification V2.0, https://www.uciexpress.org/specifications
[2] Bhargava, P., *et al* "A 256Gbps Microring-based WDM Transceiver with Error-free Wide Temperature Operation for Co-packaged Optical I/O Chiplets," *IEEE Symposium on VLSI Technology and Circuits*, June 2024.
[3] Wade, M., *et al* "An error-free 1 Tbps WDM optical I/O chiplet and multi-wavelength multi-port laser," in *Optical Fiber Communication Conference (OFC) 2021*, paper F3C.6.