2.4 "Zeppelin": An SoC for Multichip Architectures

Noah Beck[1], Sean White[1], Milam Paraschou[2], Samuel Naffziger[2]

[1]AMD, Boxborough, MA
[2]AMD, Fort Collins, CO

Codenamed "Zeppelin", AMD's next-generation System-on-a-Chip (SoC) was designed for use in multiple products and packages in multiple markets, including server, mainstream PC desktop, and high-end desktop. Utilizing GLOBALFOUNDRIES' 14nm LPP FinFET process technology, the "Zeppelin" SoC has over 4.8B transistors. It contains high-performance AMD x86 cores codenamed "Zen" [1][2], caches, memory controllers, PCIe®, SATA, and other IO controllers, and integrated x86 southbridge chipset capabilities. All these functions are connected on the SoC and between multichip packages and multi-socket systems by AMD Infinity Fabric.

The "Zeppelin" SoC was architected with leadership server capabilities as the top priority, but retained features in the single "Zen"-based SoC to support other complementary markets as well:

- Server market: 4-chip SP3 package with 8 DDR4 channels and 128 PCIe® Gen3 lanes, scalable to 2-socket systems with coherent interconnect.
- Client market: Single-chip AM4 package with 2 DDR4 channels and 24 PCIe® Gen3 lanes, platform-compatible with other AMD SoCs.
- High-end desktop market: 2-chip sTR4 package with 4 DDR4 channels and 64 PCIe® Gen3 lanes.

Within the Infinity Fabric (IF), the Scalable Data Fabric (SDF) plane is designed to provide the coherent data transport between cores using the cache-coherent master (CCM), memory using the unified memory controller (UMC), and IO using the IO master/slave (IOMS) (Fig. 2.4.2). The chip-to-chip communication method in the SDF is key to this multichip package approach, enabled by the Coherent AMD Socket Extender (CAKE) component of the SDF. CAKE is designed to take requests/responses from the local chip's SDF, and encode them into flits at 128b per clock cycle. CAKE is bidirectional, also decoding flits each cycle. The flits are suitable for transmission over any serializer/deserializer (SerDes) interface to another chip within the system. To eliminate clock-domain-crossing latency, the clock of all the SDF components, including the CAKE component in "Zeppelin", run at the system DRAM's MEMCLK frequency.

Two different SerDes types are used with CAKE in "Zeppelin": one for Infinity Fabric on-package (IFOP) traffic and one for Infinity Fabric inter-socket (IFIS) traffic. The IFOP SerDes is designed for minimum power across short in-package trace lengths, while the IFIS SerDes is needed for communication across longer socket-to-socket trace lengths.

A custom IFOP SerDes design was created achieving a power efficiency of 2pJ/b. Key elements in achieving the power target:

- 32b of low-swing, single-ended data with differential clock consuming ~50% the power of an equivalent differential driver.
- Zero power driver state during logic-0 transmission due to transmit/receive impedance termination to ground while the driver pullup is disabled, also applied during link idle.
- Data-bit inversion encoding optimizes bit patterns transmitted to take advantage of the low-power logic-0 state (Fig. 2.4.1) saving 10% average power per bit.

To support the high reliability required by server systems, CRC is transmitted along with every cycle of data. The IFOP SerDes transmission bitrate is 4 transfers per CAKE clock. The bandwidth of an IFOP link is overprovisioned by about a factor of two relative to DDR4 channel bandwidth for mixed read/write traffic to provide robust multi-chip performance scaling.

The SerDes used for IFIS also supports PCIe® and SATA protocols. To align the package pinout with standard PCIe® device lane counts, the IFIS SerDes transmits and receives 16 differential data lanes at roughly 11pJ/b. The IFIS SerDes interface runs at 8 transfers per CAKE clock. Due to the 16b data width and in-band CRC overhead, the bandwidth of an IFIS link has 8/9 of the bandwidth of an IFOP link.

Optimization of SerDes placement in the SoC floorplan required careful consideration of the pinout and routing challenges of the 2-socket-capable 4-chip SP3 package (Fig. 2.4.2). The package pinout is dominated by eight DDR4 channels (4 per side) and eight 16-lane high-speed SerDes links (4 each top and bottom). To support routing the DDR4 channels in the package, the chips on the left side were rotated 180° compared to the chips on the right. To support common platform configurations, the SerDes providing IFIS links to a second socket were routed to the top of the package with each chip providing one such link. The remaining SerDes links, supporting PCIe®/SATA, were routed to the bottom of the package.

While only one 16-lane SerDes per chip needs to support IF traffic at a time, muxing the IFIS capability to SerDes placed at opposing corners of the chip allowed half of the DDR4 routes to coexist on the same two package signal routing layers as the IFIS/PCIe®/SATA SerDes routes (Fig. 2.4.3). A second PCIe® controller on the chip enabled rotated chips to support PCIe® to the bottom of the package, and also enabled support for the I/O-heavy single-socket SP3 option – up to 128 lanes of PCIe®.

Three IFOP SerDes from each chip were required to support full connectivity in the 4-chip package. However, four SerDes were ultimately placed in order to keep all IFOP SerDes package routes restricted to two package layers, and to allow those two package layers to also provide half of the DDR4 package routes. Two IFOP SerDes are placed on the side opposite DDR4 and used by all chips. The DDR4 PHY itself is placed in between two additional IFOP SerDes. One of these two SerDes is unused and clock gated on each chip. The resulting IO and core complex locations are overlaid on the die image in Fig. 2.4.4. The total SoC die area is 213mm^2. Having met the server 4-chip package routing challenges, a 2-chip sTR4 package (Fig. 2.4.5) was created with half of the high-speed IO pins, and a single-chip AM4 package was created to drop into previously existing AM4 platforms.

The IFOP SerDes and digital logic, such as CAKE, that support the 4-chip architecture of SP3, add area which adds to cost; the total silicon area in SP3 is 852mm^2. Creating a monolithic 32-core die without the multichip support would only save about 10% of the area, resulting in a 777mm^2 die [3]. AMD projects that the large die would cost ~40% more to manufacture and test than four small chips. Adding to the cost benefits, the multichip design provides ~20% higher full 32-core yield than would the single-chip version. To make only the top-of-stack 32-core parts, the cost for the large die jumps to 70% more than the cost of the 4 small chips. A very high-yielding multichip assembly process is required, or the improved silicon yields are lost at the package level. AMD internal data has demonstrated success at achieving assembly yields that have a negligible impact on overall cost. In order to ensure that chips with similar maximum frequency capabilities can be matched to each other for assembly into the same package, on-die frequency sensors containing representative critical-path logic are consulted before chips are selected for assembly into packages [4].

The SP3 package delivers power with ±25mV accuracy to all cores, as seen in Fig. 2.4.6, which shows the measured variation in core voltage across the SP3 package, chips and cores running a maximum power pattern at 2.5GHz. On-die per-core low-drop-out (LDO) voltage regulators reduce voltage to faster cores to save power. Idle cores are power-gated for maximum power savings.

The design choices made in the "Zeppelin" SoC definition enabled a wide variety of products for both legacy and new platforms, ranging from single-chip up to 8-chip in the largest 2-socket configuration. The careful balance between compute cores, memory and IO capability on the base design, with the high-bandwidth fabric provides scalable performance across these products (Fig. 2.4.7). As Moore's Law slows in its ability to deliver more transistors per area, multichip architectures such as "Zeppelin" are necessary to provide continued increases in functionality that can be delivered to a single package.

References:

[1] T. Singh, et al., "The Next-Generation High-Performance x86 Core: Zen," *ISSCC*, pp. 52-53, 2017.
[2] M. Clark, "A New x86 Core Architecture for the Next Generation of Computing," *Hot Chips*, 2016.
[3] K. Lepak, et al., "The Next Generation AMD Enterprise Server Product Architecture," *Hot Chips*, 2017.
[4] S. Sundaram, et al., "Bristol Ridge: A 28-nm x86 Performance-Enhanced Microprocessor Through System Power Management," *IEEE JSSC*, vol. 52, no. 1, pp. 89-97, 2017.

 978-1-5090-4940-0/18/$31.00 ©2018 IEEE

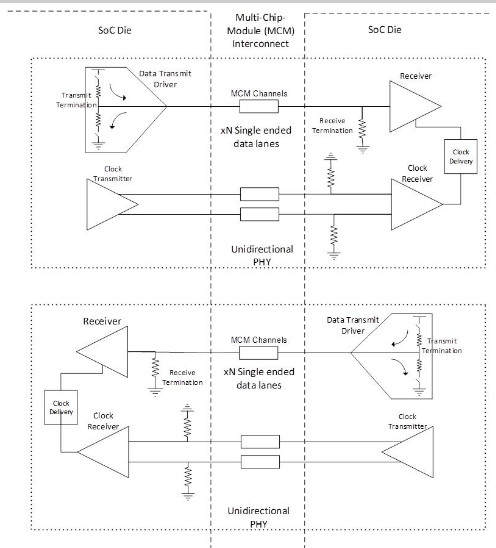


Figure 2.4.1: Infinity Fabric on-package SerDes link circuit diagram.

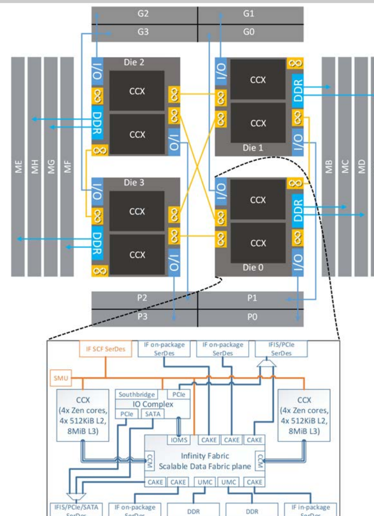


Figure 2.4.2: SP3 package pinout with high-speed connectivity and die architectural detail.

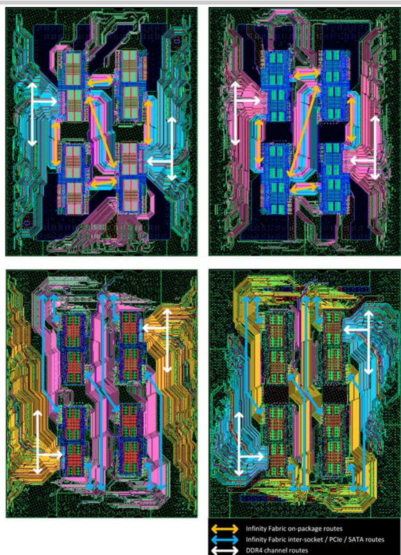


Figure 2.4.3: Four signal routing layers of the SP3 package.

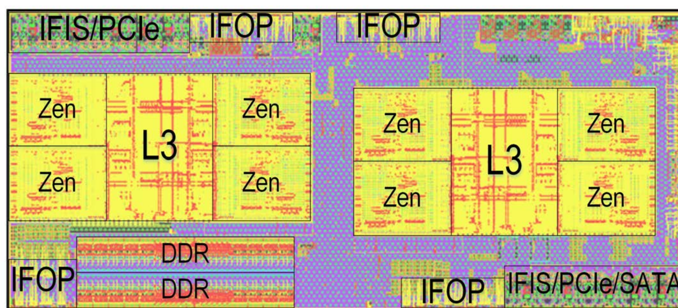


Figure 2.4.4: Die image.

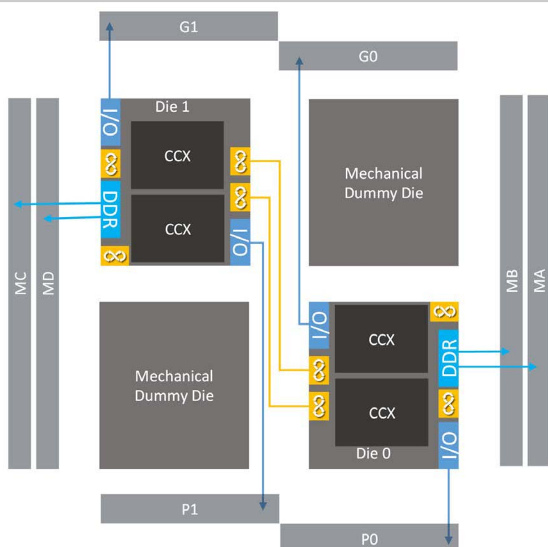


Figure 2.4.5: sTR4 high-end desktop package.

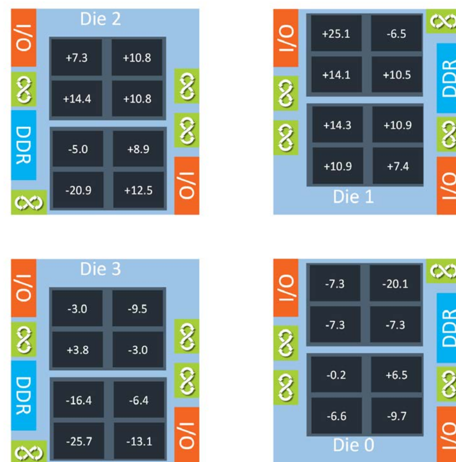
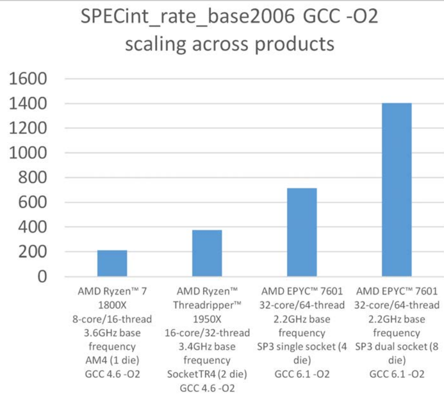


Figure 2.4.6: Measured variation in core voltage across SP3 package core locations.



AMD Ryzen™ 7 1800X CPU scored 211, using estimated scores based on testing performed in AMD Internal Labs as of 30 March 2017. System config: Ryzen™ 7 1800X: AMD Myrtle-SM with 95W R7 1800X, 32GB DDR4-2667 RAM, Crucial CT256M550SSD, Ubuntu 15.10, GCC –O2 v4.6 compiler suite.
AMD Ryzen™ Threadripper™ 1950X CPU scored 375, using estimated scores based on testing performed in AMD Internal Labs as of 7 September 2017. System config: Ryzen™ Threadripper™ 1950X: AMD Whitehaven-DAP with 180W TR 1950X, 64GB DDR4-2667 RAM, CT256M4SSD disk, Ubuntu 15.10, GCC –O2 v4.6 compiler suite.
AMD EPYC™ 7601 CPU scored 702 in a 1-socket using estimated scores based on internal AMD testing as of 6 June 2017. 1 x EPYC™ 7601 CPU in HPE Cloudline CL3150, Ubuntu 16.04, GCC -O2 v6.3 compiler suite, 256 GB (8 x 32 GB 2Rx4 PC4-2666) memory, 1 x 500 GB SSD
AMD EPYC™ 7601 scored 1390 in a 2-socket system using estimated scores based on internal AMD testing as of 6 June 2017. 2 x EPYC™ 7601 CPU in Supermicro AS-1123US-TR4, Ubuntu 16.04, GCC -O2 v6.3 compiler suite, 512 GB (16 x 32GB 2Rx4 PC4-2666 running at 2400) memory, 1 x 500 GB SSD.

Figure 2.4.7: SPECint_rate_base2006 GCC -O2 scaling.

978-1-5090-4940-0/18/$31.00 ©2018 IEEE