A 256Gb/s/mm-shoreline AIB-Compatible 16nm FinFET CMOS Chiplet for 2.5D Integration with Stratix 10 FPGA on EMIB and Tiling on Silicon Interposer

Chester Liu, Jacob Botimer, Zhengya Zhang

University of Michigan, Ann Arbor, MI

Abstract

This work presents an Advanced Interface Bus (AIB)-compatible microcontroller unit (MCU) chiplet in 16nm FinFET CMOS. The MCU chiplet consists of three AIB channels, each providing 20 Tx and Rx pairs to support 80Gb/s/channel over 55µm-pitch microbumps. Two multi-chip modules (MCM) were constructed, one made of two MCU chiplets integrated on a 180nm passive silicon interposer, and the other made by pairing an MCU chiplet with a Stratix 10 FPGA over an Embedded Multi-die Interconnect Bridge (EMIB). The AIB interface provides 256Gb/s/mm-shoreline, consuming 0.83pJ/b at 0.9V and 1GHz. The two MCMs demonstrate the ease and versatility of modular 2.5D integration.

Introduction

The 2.5D integration of chiplets offers a promising path towards constructing large-scale systems to deliver a performance comparable to single-chip integration, but without the high cost, risks and effort associated with monolithic integration. A dual-SoC-chiplet was demonstrated on CoWoS using an 8Gb/s/pin low-voltage in-package-interconnect [1], and a 36-chiplet DNN accelerator was demonstrated on an organic substrate using a 25Gb/s/pin ground-referenced signaling [2][3]. These results show the flexibility of 2.5D integration at competitive performance and efficiency. However, a low-cost standard interface and standard-conforming chiplets are needed to create an open chiplet ecosystem.

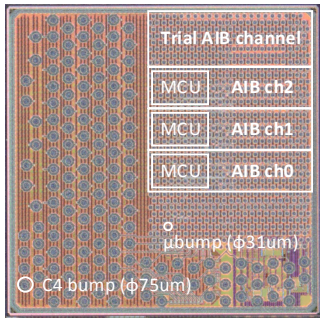
This work adopts the open-source Advanced Interface Bus (AIB) GEN1 [4] as the chiplet-to-chiplet interface standard. AIB transfers parallel digital data with forwarded source clock over a short distance (~3mm) between chiplets at 2Gb/s/wire. The relatively low-speed, short-reach AIB interface follows standard digital design, with a lower complexity, cost, and a better portability than high-speed serial interfaces. The simplicity of the AIB interface leads to over 10× shorter latency than a serial interface [4] and yet a competitive energy efficiency. Supported by advanced packaging technology including microbumps (µbumps) and silicon interposer, the AIB interface rivals the competing serial interfaces in terms of silicon area, bandwidth density and BER (Fig. 1).

A 16nm FinFET microcontroller unit (MCU) chiplet was designed and fabricated with AIB GEN1 interface over µbumps of 55µm pitch to achieve 2Gb/s/pin, a 4ns latency and a data bandwidth density of 256Gb/s/mm-shoreline. A pair of MCU chiplets were integrated on a low-cost, 3-layer, 180nm passive silicon interposer to demonstrate chiplet tiling. The MCU chiplet was also integrated with a Stratix 10 FPGA over Embedded Multi-die Interconnect Bridge (EMIB) [5] to demonstrate heterogenous integration of chiplets in different process technologies.

AIB I/O and Adaptor Design

An AIB I/O cell consists of an SDR-to-DDR serializer, a DDR-to-SDR de-serializer, and a driver. For the Tx, 2 bits of SDR are serialized to 1b DDR and then sent to the driver (Fig. 2). The Tx is clocked at 1GHz. The driver is a tri-state buffer, and we sized the buffer to meet AIB compliance eye mask at the default driving strength with EMIB's extracted RC wire load model. For the Rx, the 1b DDR input received from the µbump is first buffered before being de-serialized to a 2b SDR output (Fig. 3). An AIB I/O cell occupies 203.2µm^2 (the driver occupies 66µm^2) in a 16nm process, smaller than a 961µm^2 µbump pad to allow all I/O cells to be placed directly underneath, or in close proximity to µbumps to minimize the skews between pins.

Data width conversion and clock-domain crossing between the core and the I/O are done by the AIB adapter. In the prototype design, the core is clocked at 500MHz, and both the core clock and the I/O Tx clock are derived from the same source. For the Tx, a width conversion FIFO converts 4b core data to 2b SDR I/O data. For the Rx, the DDR data sampling clock is derived from the Tx clock by a tunable delay. To sample the DDR data at the center of the eye, the tunable delay is set to ¼ clock period. The sampled data are pushed to a phase-compensation FIFO that bridges the sampling clock and the core clock domain. The core pops data from the other end of the FIFO at the core clock rate.



Chiplet Implementation

A prototype MCU chiplet was implemented in 16nm FinFET CMOS. The chiplet integrates three standard AIB channels with 20 Tx and Rx pairs and one trial channel with 4 Tx and Rx pairs for characterization. Each AIB channel contains 96 signal and 42 power/ground µbumps, occupying 312.5µm × 1246.5µm.

An AIB channel includes an OpenRISC MCU for testing, and it can be configured as either master or slave. In the master mode, the MCU reads from and writes to the AIB channel via its on-chip bus. In the slave mode, the MCU's SRAM can be read by and written to by the AIB channel. To facilitate automated testing, a loop mode is created to allow a BIST controller to send and receive data between a pair of chiplets in a loop over the AIB channels for verification.

A lightweight streaming protocol is designed in this work for the inter-chiplet communication. Round-trip latency between chiplets is 8 AIB clock cycles (8ns), including 2 cycles for the AIB I/O and 2 cycles for the AIB adaptor on both chiplets. The buffer size for the inter-chiplet communication is determined by the protocol and the round-trip transfer latency.

2.5D Homogeneous and Heterogeneous Integration

Two 2.5D multi-chip modules (MCM) were designed for demonstration. The first MCM consists of two 16nm MCU chiplets paired on a silicon interposer (Fig. 4), where one chiplet acts as the master and the other as the slave. The 180nm passive silicon substrate was fabricated with only 3 BEOL metal layers and no FEOL. The top thick metal layer is reserved for power and ground routing. Signals are routed on the lower 2 metal layers with 2.5µm width and 4.7µm pitch. The two chiplets are spaced 1mm apart, and all the AIB signal routes are kept at 2mm. The MCM is stress-tested to measure the transfer bandwidth and efficiency. At room temperature and operating at 0.9V, 1GHz AIB clock and 500MHz core clock, the chiplet-chiplet bandwidth reaches 80Gb/s/channel, or 256Gb/s/mm of chiplet shoreline, at an energy of 0.83pJ/b.

The second MCM consists of a 16nm MCU chiplet packaged with an Intel Stratix 10 FPGA (Fig. 5). The AIB signals are routed on EMIB with 2µm width and 2µm spacing. A test case is created where the MCU is configured as the master and the FPGA acts as the slave: the MCU continuously offloads CRC32 computations to the FPGA by sending blocks of data to the FPGA via AIB, and accepting the outputs from the FPGA via AIB. This test case shows 40× speedup can be achieved when the computation is offloaded to the FPGA via an optimized AIB interface compared to running the same computation on the MCU.

Compared to the state-of-the-art MCM interfaces using a 0.3V swing (Fig. 6), our synthesizable full-swing AIB-compatible interface demonstrates competitive energy efficiency and bandwidth density, while taking the shortest latency. The two fully functional MCMs show a plug-and-play approach towards the rapid construction of low-cost, modular integrated systems using an open standard interface and standard-conforming chiplets.

Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency CHIPS program and the Office of Naval Research under grant N00014-17-1-2992. The authors would like to thank Tim Hoang, Allen Chan, Thungoc Tran, David Kehlet, Sergey Shumarayev, Arnab Sarkar and Upendra Sheth from Intel for advice and assistance.

References:

[1] M.-S. Lin *et al*., VLSI Circuits, 2019.

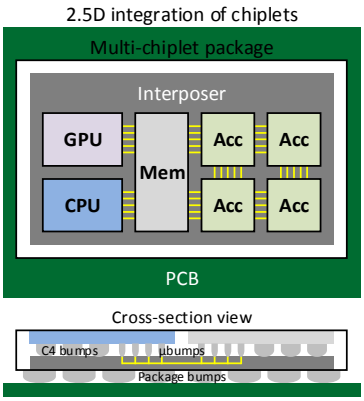
[2] B. Zimmer, *et al*., VLSI Circuits, 2019.

[3] J. W. Poulton, *et al*., JSSC, 2019.

[4] D. Kehlet, "Accelerating Innovation Through A Standard Chiplet Interface: The Advanced Interface Bus (AIB)", Intel White Paper.

[5] D. Greenhill, *et al*., ISSCC, 2017.

2021 IEEE Custom Integrated Circuits Conference (CICC) | 978-1-7281-7581-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/CICC51472.2021.9431555



Complex serial I/O design	
Area	>10000um^2/b
BW density	<500Gb/s/mm^2
Latency	>10ns
BER	~10^{-15}

Simple parallel I/O design	
Area	<500um^2/b
BW density	>500Gb/s/mm^2
Latency	<5ns
BER	~10^{-20}

Fig. 1. 2.5D integration provides the flexibility in constructing large-scale systems with chiplets implemented in different technologies. With advanced packaging technology, a simple parallel data interface rivals a complex serial data interface.

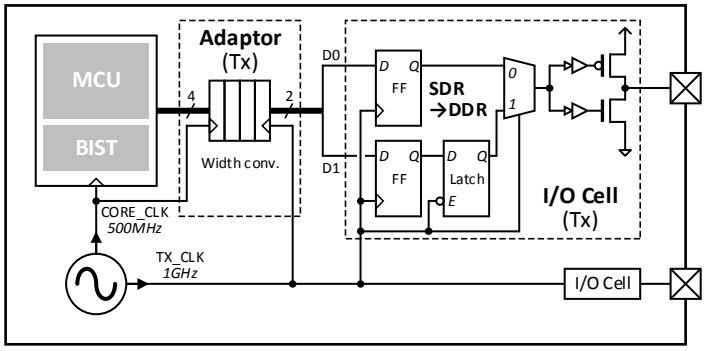


Fig. 2. Chiplet Tx hardware architecture. The adaptor converts 4b-wide 500MHz data from the MCU to 2b-wide 1GHz SDR. The 2b-wide 1GHz SDR data are serialized to 1GHz DDR and sent to the IO driver for transmission. The clock is transmitted alongside.

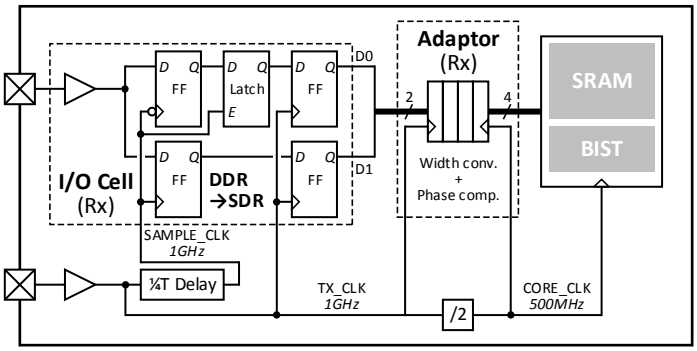


Fig. 3. Chiplet Rx hardware architecture. The adaptor converts received data width and compensates the phase difference between the received Tx clock and the reference core clock generated from the Tx clock.

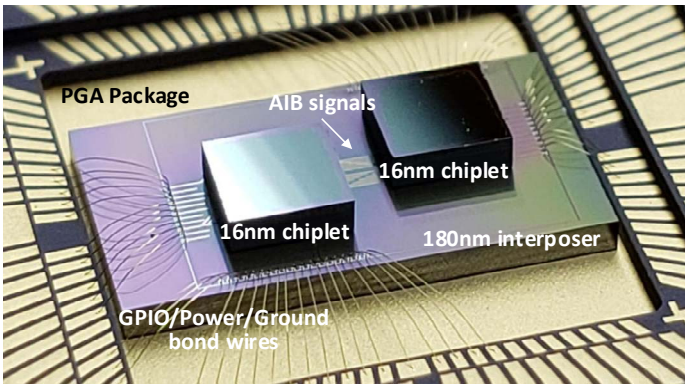


Fig. 4. Two 16nm chiplets are spaced 1mm apart and assembled on a 180nm passive silicon interposer. The MCM is placed in a PGA package for testing.

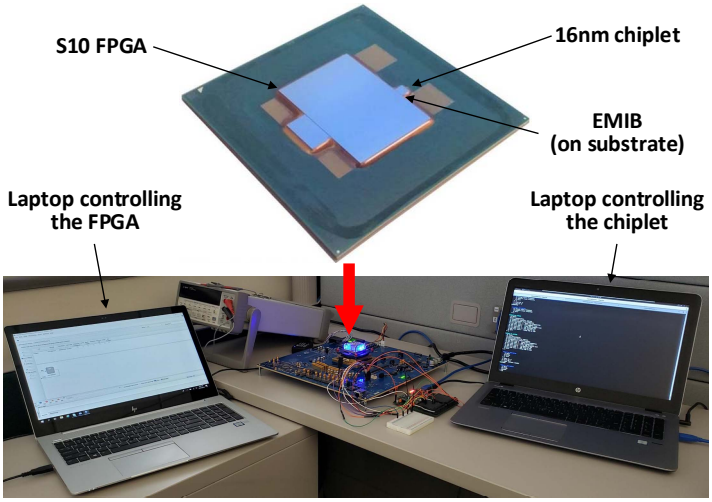


Fig. 5. A 16nm chiplet is integrated with an Intel Stratix 10 FPGA via EMIB on the package substrate.

	This Work	LIPINCON [1]	GRS [3]
Technology	**16nm FinFET**	7nm FinFET	16nm FinFET
Voltage swing	**0.9V**	0.3V	0.3V
Bump pitch	**55um**	40um	140um
Chiplet carrier	**Silicon interposer 3-layer / EMIB 4-layer**	CoWoS 15-layer	Organic substrate 12-layer
Reach	**2mm**	500um	80mm
I/O size	**203.2um^2/b**	500um^2/b	10175um^2/b
Data rate per pin	**2Gb/s**	8Gb/s	25Gb/s
Energy efficiency	**0.83pJ/b**	0.56pJ/b	1.17pJ/b
Shoreline BW density	**256Gb/s/mm**	1.6Tb/s/mm	354Gb/s/mm
Area BW density	**614.4Gb/s/mm^2**	1.6Tb/s/mm^2	516Gb/s/mm^2
Latency	**4ns**	-	<20ns

Fig. 6. Comparison with state-of-the-art MCM interfaces.