Arvon: A Heterogeneous System-in-Package Integrating FPGA and DSP Chiplets for Versatile Workload Acceleration

Wei Tang, *Member, IEEE*, Sung-Gun Cho, Tim Tri Hoang, Jacob Botimer, *Member, IEEE*, Wei Qiang Zhu, Ching-Chi Chang, Cheng-Hsun Lu, *Student Member, IEEE*, Junkang Zhu, *Graduate Student Member, IEEE*, Yaoyu Tao, *Member, IEEE*, Tianyu Wei, *Graduate Student Member, IEEE*, Naomi Kavi Motwani, Mani Yalamanchi, Ramya Yarlagadda, Sirisha Rani Kale, Mark Flanigan, Allen Chan, Thungoc Tran, Sergey Shumarayev, and Zhengya Zhang, *Senior Member, IEEE*

***Abstract*— Integrating heterogeneous chiplets in a package presents a promising and cost-effective approach to constructing scalable and flexible systems for accelerating a wide range of workloads. We introduce Arvon that integrates a 14-nm FPGA chiplet with two efficient and densely packed 22-nm DSP chiplets using embedded multidie interconnect bridges (EMIBs). The chiplets are interconnected via a 1.536-Tb/s advanced interface bus (AIB) 1.0 interface and a 7.68-Tb/s AIB 2.0 interface. Arvon is programmable, supporting various workloads from neural network (NN) to communication signal processing. Each DSP chiplet delivers a peak performance of 4.14 TFLOPS in half-precision floating-point while maintaining a power efficiency of 1.8 TFLOPS/W. A compilation procedure is developed to map workloads across the FPGA and DSPs to optimize performance and utilization. Our AIB 2.0 interface implementation using 36-μm-pitch microbumps achieves a data transfer rate of 4 Gb/s/pin, with an energy efficiency of 0.10–0.46 pJ/b including the adapter. The bandwidth density reaches 1.024 Tb/s/mm of shoreline and 1.705 Tb/s/mm^2 of area.**

***Index Terms*— Advanced interface bus (AIB), chiplet, heterogeneous integration, system in package (SiP).**

I. Introduction

DSP workloads such as machine learning, robotics, and 5G/6G communications are progressing at a rapid pace [1], [2], [3], [4]. Notably, these workloads are growing to be more dynamic as well as diverse. A flexible and compute-efficient hardware system serves these needs well. Such a system can integrate multiple components including a CPU or FPGA and DSP accelerators, where the CPU or FPGA provide the needed flexibility and DSP accelerators provide efficient kernel acceleration.

One way to implement such a system is to design and fabricate a monolithic system-on-chip (SoC). Building a large SoC is time-consuming and costly. As a cost-effective alternative, heterogeneous system-in-package (SiP) employing 2.5-D or 3-D integration of chiplets offers a promising path toward constructing large-scale systems to deliver a performance comparable to monolithic integration, but without the high cost, risks, and effort associated with monolithic integration.

An SiP consists of interconnected components called chiplets. Each chiplet embodies a functional module that can be fabricated in the most suitable technology node to gain the best performance and efficiency. Since each chiplet is more compact in size and dedicated in function, the design complexity is reduced and the yield is increased. By selecting known good dies (KGD) to assemble the SiP, the system yield can be improved. In an envisioned future chiplet ecosystem, proven chiplets can be sourced from various vendors and reused in constructing diverse systems, removing the challenges and obstacles in the rapid construction of novel systems.

An SiP solution for a versatile accelerator is shown in Fig. 1, comprising an FPGA chiplet, a DSP accelerator chiplet, and potentially an extension chiplet such as an ADC or an optical transceiver. A spectrum of dynamic DSP workloads, from machine learning to communication signal processing, can be conveniently mapped to such a heterogeneous SiP. The FPGA chiplet contributes adaptivity, the DSP chiplet contributes computational capacity at high efficiency, while an extension chiplet offers connectivity to front-end (FE) components like sensors and wireless or optical interfaces. Within an SiP, the die-to-die interface between chiplets plays a critical role. The interface needs to provide a high data transfer bandwidth between the chiplets to match the performance of a monolithic SoC while keeping the energy per bit sufficiently low to remain a competitive solution.

Recent research has showcased the integration of chiplets in SiPs featuring high-bandwidth and efficient die-to-die interfaces [5], [6], [7], [8], [9], [10], [11]. In [5], two duo-Arm

Manuscript received 26 August 2023; revised 8 November 2023; accepted 3 December 2023. Date of publication 27 December 2023; date of current version 28 March 2024. This article was approved by Associate Editor Mototsugu Hamada. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) Program, The Office of Naval Research (ONR), under Grant N00014-17-1-2992; and in part by the ACE Center for Evolvable Computing and the Center for Ubiquitous Connectivity (CUbiC), sponsored by Semiconductor Research Corporation (SRC) and DARPA under the JUMP 2.0 Program. *(Wei Tang, Sung-Gun Cho, and Tim Tri Hoang contributed equally to this work.) (Corresponding author: Wei Tang.)*

Wei Tang, Cheng-Hsun Lu, Junkang Zhu, Yaoyu Tao, Tianyu Wei, Naomi Kavi Motwani, and Zhengya Zhang are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: weitang@umich.edu).

Sung-Gun Cho, Tim Tri Hoang, Wei Qiang Zhu, Ching-Chi Chang, Mani Yalamanchi, Ramya Yarlagadda, Sirisha Rani Kale, Mark Flanigan, Allen Chan, Thungoc Tran, and Sergey Shumarayev are with Intel Corporation, Santa Clara, CA 95054 USA.

Jacob Botimer is with Memryx, Ann Arbor, MI 48109 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2023.3343457.

Digital Object Identifier 10.1109/JSSC.2023.3343457

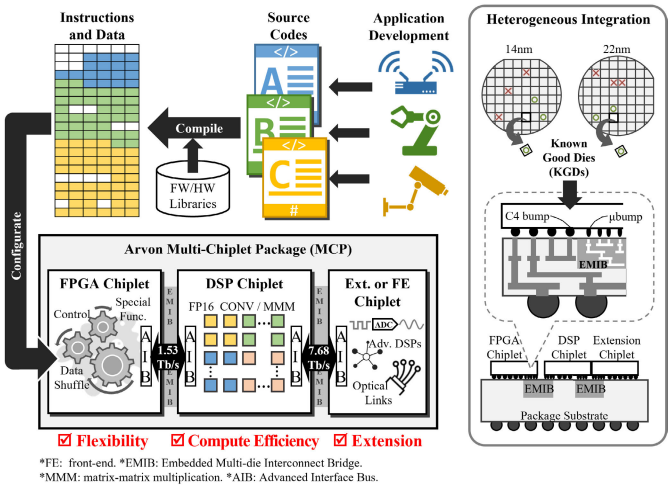


Fig. 1. Arvon SiP heterogeneously integrating FPGA, DSP, and FE chiplets for flexible workload mapping.

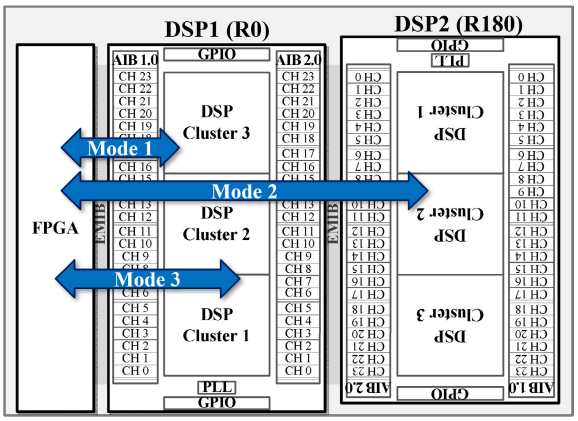


Fig. 2. Data flow modes supported in Arvon SiP: in Mode 1 and 2, the FPGA is connected to one of the DSPs, and in Mode 3, the FPGA is connected to both DSPs.

core chiplets are integrated on chip-on-wafer-on-substrate (CoWoS) with an 8-Gb/s/pin Low-voltage-In-Package-INterCONnect (LIPINCON) interface. In [6], 36 DNN accelerator chiplets are integrated on an organic substrate using a 25-Gb/s/pin ground-referenced signaling (GRS) interface [7]. In [8] and [9], four run-time reconfigurable universal digital signal processors (UDSP) are integrated on silicon interconnect fabric (Si-IF) interposer with a 1.1-Gb/s/pin SNR-10 interface. IntAct [10] integrates six 16-core chiplets on an active silicon interposer with a 1.2-Gb/s/pin 3-D-Plug interface. These results exemplify homogeneous integration, involving the tiling of multiple instances of a modular chiplet to increase the scale of computational systems.

In Arvon, we demonstrate the heterogeneous integration of different types of chiplets to construct a versatile accelerator for DSP workloads. Arvon consists of a 14-nm FPGA chiplet and two 22-nm DSP chiplets integrated through embedded multidie interconnect bridge (EMIB) technology [12], [13]. We prototyped both the first- and second-generation open advanced interface bus (AIB) die-to-die interfaces, known as AIB 1.0 and AIB 2.0, respectively, for connecting the chiplets. The results are demonstrated in an SiP that is capable of effectively accelerating a variety of machine-learning and communication DSP workloads while maintaining substantial hardware utilization. This work also showcases the AIB 2.0 interface that achieves a high bandwidth density of 1-Tb/s/mm shoreline and 1.7-Tb/s/mm^2 area at an energy efficiency of 0.1 pJ/b.

The rest of this article is organized as follows. In Section II, an overview of Arvon SiP is presented. In Section III, we elaborate on the AIB interface design, encompassing the physical (PHY) I/O, clock distribution, and bus adaptation. Section IV delves into the details of the DSP chiplet and its vector engine design. The mapping of various workloads is discussed in Section V. Silicon measurements and system evaluations are presented in Section VI. Finally, Section VII concludes this article.

II. Arvon System Overview

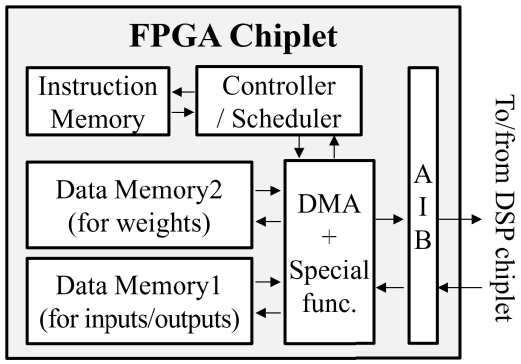
The Arvon system overview is presented in Fig. 2. The system comprises an FPGA chiplet and two instances of a DSP chiplet, named DSP1 and DSP2. DSP2 is a physically rotated version of DSP1. The FPGA is connected to DSP1 using EMIB via an AIB 1.0 interface, and DSP1 is connected to DSP2 using EMIB via an AIB 2.0 interface. Arvon provides three operation modes, as depicted in Fig. 2. In Mode 1 and Mode 2, the FPGA is connected to DSP1 and DSP2, respectively, to offload common computational kernels to the DSPs. The common kernels include matrix–matrix multiplication (MMM) and 2-D convolution (conv) that are essential in neural network (NN) and communication workloads. In Mode 3, DSP1 and DSP2 are combined to augment the computational capacity. DSP2 can also be replaced by an FE chiplet, for example, an optical tile or an ADC tile, to realize a complete communication or sensing system.

A. DSP Chiplet

The DSP chiplet provides offloading and acceleration of compute-intensive workloads. The design of the DSP chiplet is illustrated in Fig. 2. Die-to-die interfaces are placed on both edges of the DSP chiplet. On the west side, there are 24 channels of AIB 1.0 interface that offer a bandwidth of 1.536 Tb/s for communicating with the FPGA. On the east side, there are 24 channels of AIB 2.0 interface that offer a bandwidth of 7.68 Tb/s for communicating with the other DSP. The chiplet contains three DSP clusters, with each cluster offering 1024 16-bit half-precision floating-point processing elements (PEs). Each cluster utilizes up to eight channels of AIB 1.0 interface and up to eight channels of AIB 2.0 interface for its I/O. A low-jitter ring PLL is used to generate clocks for the DSP clusters as well as the AIB 1.0 and AIB 2.0 interfaces. Two rows of GPIOs, along the top and the bottom of the chiplet, facilitate global configuration and debugging.

B. FPGA Host Chiplet

The FPGA plays a crucial role in enabling Arvon's flexibility. The FPGA's programmable logic is utilized to support



*DMA: direct memory access.

Fig. 3. Example of a host FPGA implementation.

various tasks such as data arrangements like transpose and shuffle operations for the DSPs. Additionally, the FPGA can be utilized to provide special functions that are not available on the DSPs, allowing it to fulfill complete processing requirements.

In Arvon, the FPGA assumes the role of the host, taking the form of an instruction-based processor as illustrated in Fig. 3. A simple host processor is equipped with an instruction memory, data memories for inputs, outputs, and weights, and a direct memory access (DMA) unit to manage and coordinate data transfers with the DSP chiplets. Instructions are used to configure and reconfigure the DSPs in runtime, direct data flows between the data memories and the DSPs, as well as to conduct pre- and post-DSP processing.

A workload execution begins when the host processor in FPGAs is triggered to read the first instruction from the instruction memory. The instructions contain all the information such as address and data for register access, memory address, bus address, data length for DMA read and write, and the order of execution. Based on the instructions, the host processor generates the AXI bus transactions to access DSP configuration registers that are sent to the DSPs. It also issues DMA commands for reading from or writing to data memories, as well as for sending and receiving data to and from the DSPs. Given the quick processing time of the vector engines in the DSPs, the FPGA implementation, which includes the host processor, is highly utilized to minimize latency and prevent any potential bottlenecks.

III. AIB Die-to-Die Interface

Within the DSP chiplet, the west side incorporates 24 channels of the AIB 1.0 interface [14], while the east side incorporates 24 channels of the AIB 2.0 interface [14]. An AIB channel consists of two layers: the adapter layer and the PHY I/O layer. The adapter layer coordinates data transfer between the DSP core and the PHY layer I/O. It is responsible for framing and synchronizing the data between these two domains. State machines are employed to initiate the AIB link and enable auto-clock phase tuning. This tuning helps identify the data's eye width and center. In AIB 2.0, the adapter also supports the optional data-bus-inversion (DBI), which reduces bus-switching activity and enhances energy efficiency.

TABLE I
Comparison Between AIB 1.0 and AIB 2.0. PHY

	AIB 1.0	AIB 2.0
Number of TX/RX data pins	20	40
Peak data rate per pin	2Gbps @ 1GHz clock	4Gbps @2GHz clock
I/O swing voltage	0.85V (full-rail)	0.4V to 0.85V
Bump map size	323μm×748μm	312μm×416μm

The PHY layer of the AIB interface implements source-synchronous, short-reach, low-latency, and parallel single-ended I/Os. Each AIB 1.0 I/O operates from 1 Mb/s to 2 Gb/s in double data rate (DDR) mode utilizing full-rail signaling. Each AIB 2.0 I/O operates from 1 Mb/s to 4 Gb/s in the DDR mode utilizing a signal swing from 0.4 V to full-rail. A single AIB 1.0 channel consists of 96 pins, which include two pins for the TX clock, two pins for the RX clock, 20 pins for TX data, 20 pins for RX data, and additional pins for sideband controls and redundancy. In contrast, a single AIB 2.0 channel consists of 102 pins, which include two pins for the TX clock, two pins for the RX clock, 40 pins for TX data, 40 pins for RX data, and additional pins for sideband controls and redundancy. AIB 2.0 improves upon AIB 1.0. It doubles both the data rate per pin and the number of data pins per channel, resulting in a fourfold increase in data transfer bandwidth. Additionally, AIB 2.0 improves energy efficiency through the use of low-swing signaling. A comparison between AIB 1.0 and AIB 2.0 is summarized in Table I. Hereafter, we will primarily focus on the design of AIB 2.0. It is worth noting that AIB 1.0 shares similar design structures to AIB 2.0.

A. AIB 2.0 Adapter

An AIB adapter manages the data transfer between the DSP core and the PHY I/O layer. The data path includes serializers at the TX end and deserializers at the RX end. Fig. 4 illustrates an example of data transfer. In Chiplet 1, an AIB 2.0 TX channel gathers four 80-bit-wide data streams at a time from the DSP core, which is clocked at 500 MHz. The serializers, implemented using two-level 2:1 multiplexers, convert the parallel data streams into a single 80-bit-wide data stream for transmission. Following the optional DBI, the high 40-bit and the low 40-bit segments of the 80-bit data are sent to the data0 and data1 pins of the 40 TX I/O cells. Each of the 40 TX I/O cells transfers 2 bits at a time at a rate of 2 GHz in DDR mode, resulting in an effective transmission speed of 4 Gb/s. The differential 2-GHz TX clock is forwarded to Chiplet 2 along with the data. In Chiplet 2, one AIB 2.0 RX channel is responsible for receiving 80-bit-wide data from the 40 RX I/O cells. The data are sampled at a rate of 2 GHz in the DDR mode. The received data is then passed through deserializers, implemented using two-level 1:2 demultiplexers, recovering four streams of 80-bit-wide data. The phase of the forwarded clock from TX is adjusted using a delay line, serving as the sampling clock for the RX I/O cells.

1) Automated Clock Phase Tuning: During the initialization phase of the link, the RX clock phase is adjusted to sample the RX data at the optimal point. The adapter incorporates an automated RX clock phase tuning mechanism. The TX transmits known pseudorandom binary sequence (PRBS) patterns, while

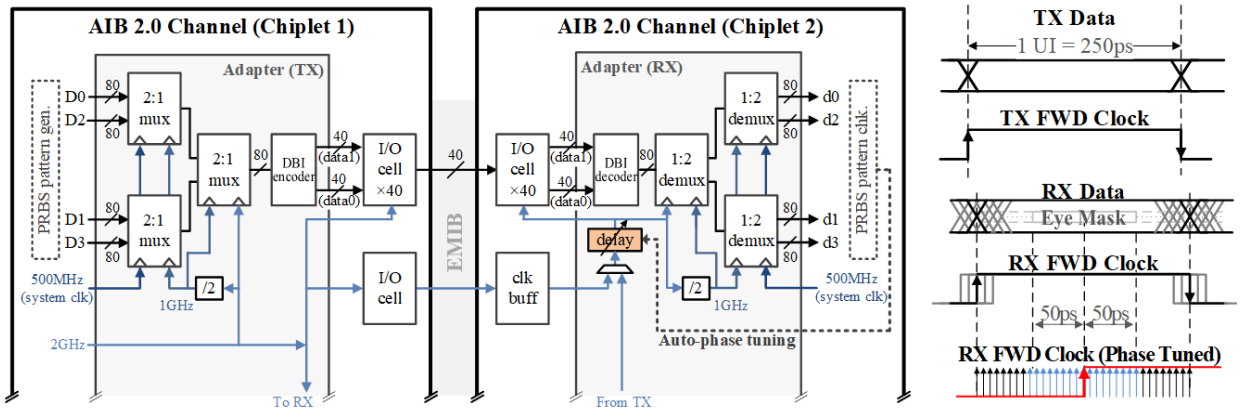


Fig. 4. AIB2.0 channel top-level diagram and automated clock phase tuning.

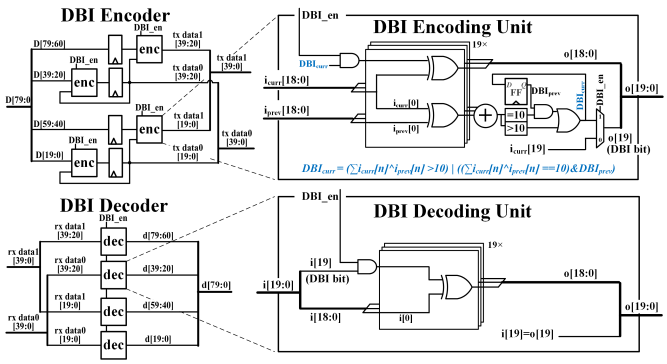


Fig. 5. 1:19-ratio DBI encoder (top) and decoder (bottom).

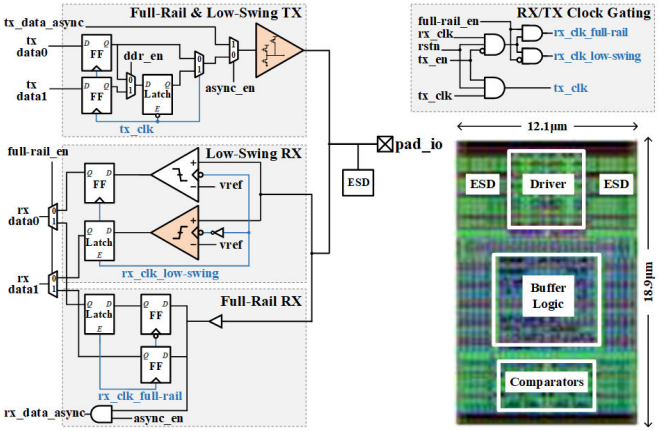


Fig. 6. Schematic and layout of one unified AIB I/O cell.

the RX monitors for errors by sweeping the delay of the received clock signal from the TX using a configurable delay line. By monitoring for errors in the received PRBS patterns, the RX can estimate the boundaries of the eye diagram. The goal is to set the delay, and consequently the sampling point, at the estimated midpoint of the eye.

2) Data Bus Inversion: AIB 2.0 supports DBI, which effectively reduces the transition and the simultaneous switching output (SSO) noise in the single-ended and source-synchronous interface. Fig. 5 illustrates a 1:19-ratio DBI encoder and decoder. In the TX, 80-bit data are encoded by four parallel DBI encoding units. Each unit takes 19-bit data bus values (denoted by $i_{curr}[18:0]$ in Fig. 5) and counts the number of bits that transition from the previously encoded data ($i_{prev}[18:0]$). The DBI encoding units invert these bits and assign a HIGH value to the DBI bit if the count exceeds 10. If the count is equal to 10 and the preceding DBI bit is already HIGH, the DBI bit remains HIGH. If neither of these conditions is met, the data remains unaltered, and the DBI bit is set to LOW. The DBI bit is then combined with the encoded 19-bit data, packed at the MSB into a 20-bit TX data, and sent to 20 I/O cells. In the RX, four parallel DBI decoding units are employed. Each unit inverts the received 19-bit data bits if the DBI bit (the MSB of the received 20-bit data block) is HIGH while leaving the data unaltered if the DBI bit is LOW.

B. AIB 2.0 I/O

The schematic and layout of a compact unified AIB 2.0 I/O cell are depicted in Fig. 6. To meet the target bump pitch of 36 μm, the layout of the unified I/O cell has been optimized with each cell connected under the corresponding microbump to ensure that the layout fits within the specified bump pitch. The unified I/O cell supports multiple modes. First, the transfer direction can be set to either TX or RX mode. This capability facilitates redundancy repair and flexible interconnection between chiplets. In the TX mode, the clocks for the RX components are gated to conserve power, while in the RX mode, the clocks for the TX components are gated. Second, the I/O signal swing can be set to full-rail for AIB 1.0 and AIB 2.0, and lower swing down to 0.4 V for AIB 2.0. Third, the transfer mode can be set to the single data rate (SDR) mode or the DDR mode. In the DDR mode, data0 and data1 are serialized, with data1 being delayed by half a clock cycle. Consequently, data0 is transmitted to the driver at the positive edge of the TX clock, while data1 is sent at the negative edge of the TX clock. This process is mirrored in the RX for data deserialization. The SDR mode employs only data0, which is sent to the driver at the positive edge of the TX clock. Finally, the I/O cell can be set to operate in the asynchronous mode for the clock and other sideband signals.

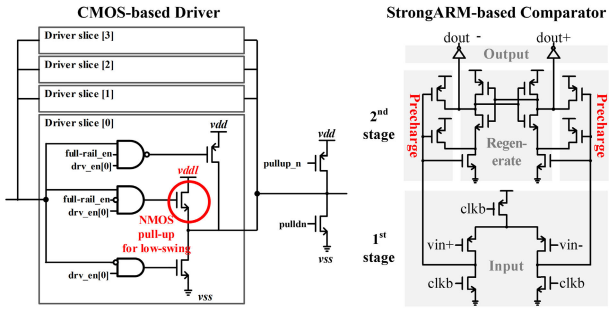


Fig. 7. Schematic of the CMOS-based TX driver (left) and strongARM-based RX (right).

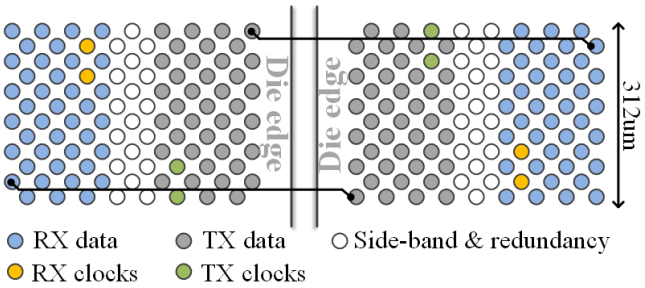


Fig. 8. Bump map of an AIB 2.0 channel.

1) TX Driver: The TX driver is depicted in Fig. 7 (left). It utilizes a segmented driver design with four slices. The design allows for the wiring together of up to four slices to achieve tunable driving strength, enabling the accommodation of channel variation and the balancing of I/O speed and power tradeoffs. Each slice consists of an NMOS for pull-down and a switchable PMOS/NMOS driver for the full-rail or low-swing pull-up. In the low-swing mode, the NMOS pull-up is overdriven to achieve a balanced driving strength with the pull-down. A power-on initial value can be set by configuring a weak pull-up and pull-down.

2) RX Buffer: For the RX, a standard-cell buffer is used for full-rail inputs, and a regenerative comparator is employed for low-swing inputs. The comparator, illustrated in Fig. 7 (right), is an optimized version of the StrongARM latch [15], [16], which reduces the mean offset to 4.1 mV without requiring calibration. PMOS is utilized to enhance the detection of low-swing inputs. The design employs a simple reference voltage generator. The comparator can reliably detect inputs as low as 0.38 V at 2-GHz DDR.

3) Bump Map: Fig. 8 illustrates a 12 × 17 bump map of an AIB 2.0 channel. This channel consists of 40 pins for TX data, 40 pins for RX data, two pins for TX forwarded clocks, two pins for RX forwarded clocks, and 18 pins for sideband and redundancy purposes. The design of the TX and RX bumps is symmetric, enabling equal-length wiring of each TX–RX pair on EMIB. With 80 data pins operating at a data rate of 4 Gb/s/pin, one AIB 2.0 channel offers a total bandwidth of 320 Gb/s. With a microbump pitch of 36 μm and a channel shoreline width of 312.08 μm, the design achieves a bandwidth density of 1024 Gb/s/mm of shoreline.

C. Clock Distribution

For high-speed parallel I/O interfaces like AIB, it is crucial to have low-skew clock distribution to ensure that all the data pins in a given channel are properly phase-aligned. As depicted in Fig. 9, we utilize a two-layer clock distribution in each AIB channel. The upper layer consists of a balanced H-tree that spans the entire channel, while the lower layer is formed by a local clock mesh. This two-layer design restricts the depth of the H-tree, ensuring a better balance of the branches. Moreover, the local clock mesh provides more consistent clock sinks without a substantial power drain. Consequently, the clock network manages to keep the worst clock skew to 8 ps. Both the H-tree and the mesh clock networks were created and assessed using the multisource clock tree synthesis (MSCTS) flow of the IC Compiler II.

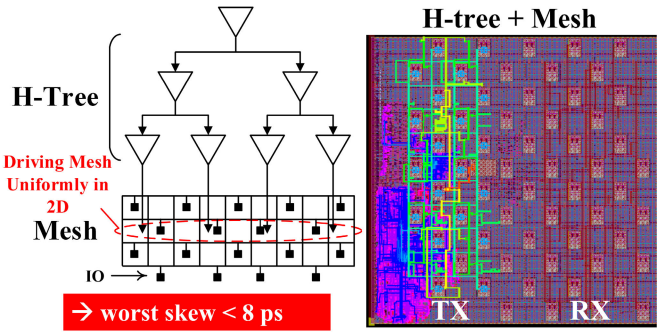


Fig. 9. Two-level clock distribution.

IV. DSP Cluster

Each DSP cluster shown in Fig. 10 comprises a flexible vector engine, a bypass buffer, a rotation block for data framing, two AXI-compatible bus converters for packing and unpacking data between multiple AIB channels, and an AXI-compatible system bus. Additionally, a bus hub is included to establish connections between the vector engine and either the tester, the AIB 1.0 interface, or the AIB 2.0 interface. The bypass buffer supports Arvon's Mode 2 operation, which enables a direct connection between the FPGA and DSP2, bypassing DSP1. This connection allows AIB 1.0 transactions from the FPGA to be directly forwarded to AIB 2.0 transactions with DSP2. The rotation block reverses the channel index ordering of the AIB interfaces. For example, when connecting DSP1 to DSP2, the rotated version of DSP1, channels 1–8 of DSP1 are connected to channels 24–17 of DSP2, requiring DSP2's rotation block to reverse the connection order.

A. Vector Engine

The central component of a DSP cluster is the vector engine, which consists of four instances of a 2-D systolic array [17]. Each systolic array comprises 256 PEs, with each PE performing multiplication in half-precision floating-point format (FP16). The 256 PEs are organized into eight units, with each unit containing 32 PEs. The sum results of each 32-PE unit are then inputted into a configurable adder tree. The configurable adder tree can flexibly support various workload mapping by selecting which of eight units to be summed together. This design offers a shorter partial-sum accumulation path and enables a higher utilization through

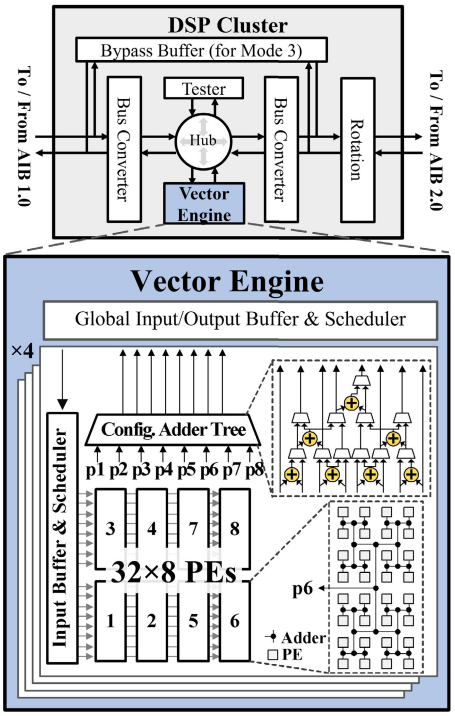


Fig. 10. DSP cluster (top) and vector engine (bottom).

concurrent workloads, distinguishing it from a classic systolic array. The entire vector engine provides 1024 PEs in total to support MMM and conv. A global I/O buffer and scheduler are implemented to distribute inputs to the PE arrays using either multicast or round-robin arbitration techniques.

Configured by instructions, the vector engine facilitates the streaming of inputs for continuous computation. The vector engine also offers a high degree of mapping flexibility. First, the four systolic arrays can be independently mapped. Additionally, the 256 PEs within each array can be configured in units of 32 PEs, accommodating from 1 to 8 independent workloads.

B. System Bus and Bus Converter

The AIB connectivity is abstracted by an AXI-compatible point-to-point system bus. The bus converter handles the packing and unpacking of data across multiple AIB channels. It also supports a burst mode to maximize bandwidth utilization for streaming. The channels and signals of the system bus are illustrated in Fig. 11. The system bus comprises four channels: a read command channel, a write command channel, a read data channel, and a write data channel. A master issues a write/read command with a 32-bit address and a 6-bit burst length, alongside 512-bit write data and a write command. In response to a read command, a slave sends 512-bit read data back to the master. The conversion between the system bus and AIB channels is done by the bus converter. We design the bus converter with a header-based streaming approach to achieve high bandwidth and low latency. Up to eight AIB channels can be utilized by a vector engine to ensure optimal utilization. Each AIB channel can be flexibly configured as either master or slave, allowing adjustment of the TX/RX bandwidth as needed.

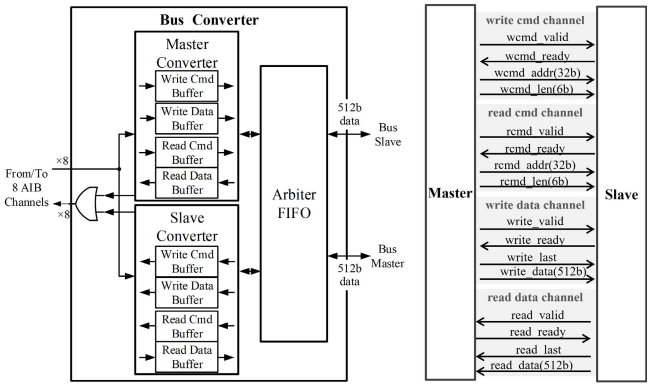


Fig. 11. AXI-compatible system bus: the bus converter (left) and the bus interface channels and signals (right).

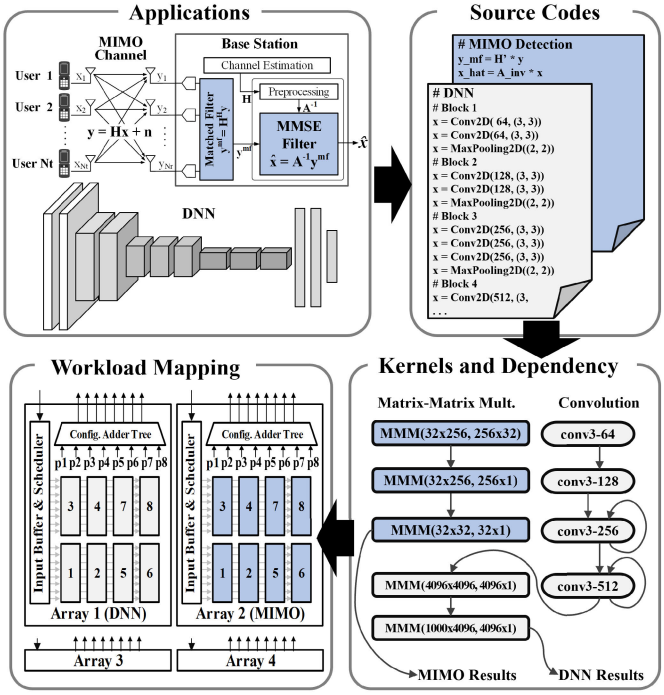


Fig. 12. Illustration of compilation flow for workload mapping.

V. Workload Mapping

Designed as a versatile computational platform, Arvon supports diverse computations of varying sizes that can dynamically change during runtime. To ensure efficient processing, it is essential to establish a systematic approach for mapping workloads to optimal hardware configurations and data arrangements.

To achieve this objective, a compilation procedure has been developed, as illustrated in Fig. 12. A workload is first segmented into parts, namely those utilizing conv or MMM kernels, or both, which can be directly mapped to the Arvon DSP through appropriate configurations. Additionally, some parts represent intermediate steps between these computational kernels, which can be executed by the FPGA host. Specifically, the conv configuration is formulated based on the sizes of the filter and input ($R \times S \times C$), while the MMM configuration is formulated based on the matrix dimensions. The conv and

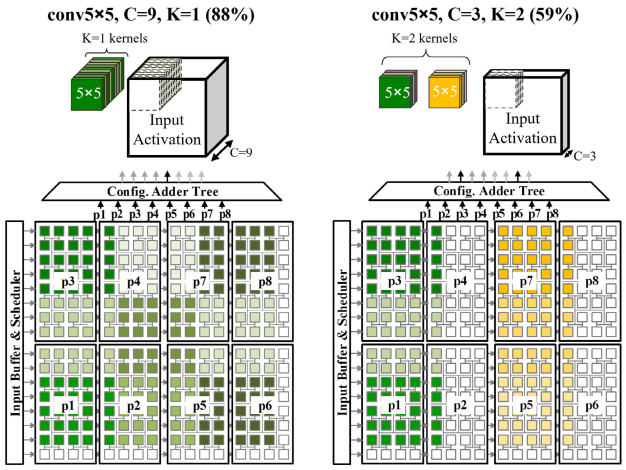


Fig. 13. Examples of mapping conv of different kernel sizes.

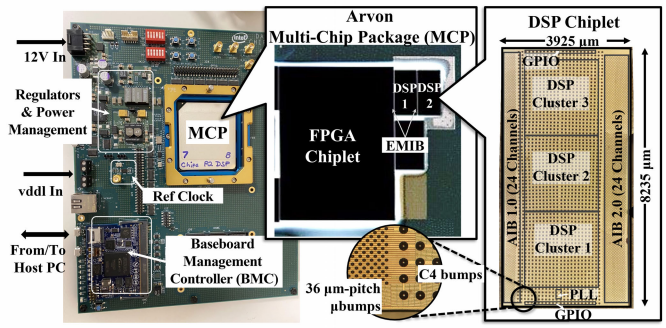


Fig. 14. Test setup, Arvon multichiplet package, and DSP chiplet microphotograph.

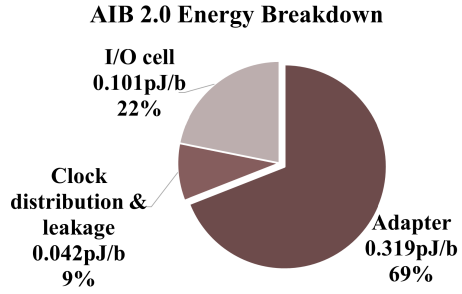


Fig. 15. Energy breakdown of the AIB 2.0 interface.

MMM kernels are then scheduled and allocated to the vector engines of the Arvon DSP by specifying instructions and arranging memory data accordingly. This allocation takes into account factors such as utilization, data reuse, and end-to-end latency.

The vector engine follows a weight-stationary scheme, so the weights of kernels are assigned to PEs. To map MMM to the vector engine [17], each row of a weight matrix is assigned to PEs, effectively distributing the 1-D vector across the 2-D array. Rows with the same weight matrix can be allocated to the same set of PEs. In a multitenant scenario involving multiple kernels, the rows of different weight matrices can be assigned to different partitions, denoted as p1–p8 in Figs. 12 and 13. The partition outputs are directed to their corresponding inputs of the configurable adder tree, ensuring separate sums are computed as outputs.

The weight mapping for conv is similar to that of multitenant MMMs, as there may be multiple conv kernels involved. Fig. 13 illustrates the mapping of two examples of conv operations. Each kernel has a size of $R \times S \times C$ and is unfolded to the 2-D PE array by knitting the slices of the third dimension in 2-D. The 3-D input activation elements under a sliding conv window are also unfolded accordingly onto the 2-D PE array. The input activation is kept inside the PE array for horizontal and/or vertical reuse via systolic data forwarding between neighboring PEs. For a single-kernel conv, such as the first example in Fig. 13, mapping can be done irrespective of the partition boundaries, resulting in an efficient utilization. However, when there are multiple kernels, for example, in the second example in Fig. 13, each kernel needs to be aligned to the boundaries of partitions, resulting in a lower utilization.

VI. Chip Measurement and Comparison

The DSP chiplet was fabricated in a 22-nm FinFET technology, which occupies an area of 32.3 mm^2 as depicted in Fig. 14. To construct Arvon SiP, a 14-nm FPGA chiplet and two DSP chiplets were copackaged and interconnected via two ten-layer EMIBs, using 36-μm-pitch microbumps. For the AIB 1.0 side, the average wire length is 1.5 mm, while for the AIB 2.0 side, the average wire length is 0.85 mm.

At room temperature and a core voltage of 0.85 V, each DSP cluster operates at a maximum frequency of 675 MHz and consumes 0.76 W. With this configuration, the peak performance of the DSP chiplet is 4.14 TFLOPS, and it achieves a power efficiency of 1.8 TFLOPS/W. With a 0.85-V I/O voltage and an 800-MHz clock (limited by the FPGA clock frequency in this design), the AIB 1.0 I/O consumes 0.44 pJ/b, or 0.85 pJ/b including the adapter, with a transfer latency of 3.75 ns. At room temperature, with a 0.4-V I/O voltage and a 2-GHz clock, the AIB 2.0 I/O consumes 0.10 pJ per bit, or 0.46 pJ/b including the adapter, and achieves a transfer latency of 1.5 ns. The energy breakdown of the AIB 2.0 interface is shown in Fig. 15. The adapter contributes the majority of energy consumption, using 0.32 pJ/b, approximately 69% of the total energy. On the other hand, the I/O cells consume only 0.10 pJ/b, approximately 22% of the total energy. The lower I/O cell energy consumption is made possible by the utilization of a low-signal swing of 0.4 V.

Arvon's AIB I/O interfaces are compared to the state-of-the-art SiP's I/O interfaces in Table II. Similar to the AIB interfaces, SNR-10 [8], 3-D-Plug [10], and LIPINCON [5] are also parallel I/O interfaces. Among them, LIPINCON provides the highest data rate of 8 Gb/s/pin and the lowest I/O energy consumption of 0.073 pJ/b with a 0.3-V signal swing; 3-D-Plug offers the highest bandwidth density of 900-Gb/s/mm shoreline; and SNR-10 demonstrates the smallest I/O size of 137 μm^2. GRS [7] is a high-speed serial I/O interface that provides 25 Gb/s/pin at an energy efficiency of 1.17 pJ/b. Our AIB 2.0 prototype presents a compelling solution with a competitive I/O energy consumption of 0.10 pJ/b, or 0.46 pJ/b when including the adapter. It also achieves the highest bandwidth density of 1.0-Tb/s/mm shoreline and 1.7-Tb/s/mm^2 area, as detailed in Table II. Fig. 16

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART SIP I/O INTERFACE

	Nagi JSSC23 [9]	Vivet JSSC21 [10]	Lin JSSC20 [5]	Poulton JSSC18 [7]	**This work**	
Interface	SNR-10	3D-Plug[(a)]	LIPINCON	GRS	**AIB1.0**	**AIB2.0**
Substrate	2-layer SI-IF	65nm Silicon Interposer	15-layer CoWoS	12-layer Organic Interposer	**10-layer EMIB**	
Reach (mm)	0.35	1.5 - 1.8	0.5	80	**0.85 - 1.5**	
Bump Pitch (um)	10	20	40	150	**36**	
Technology	16nm	28nm	7nm	16nm	**22nm**	
I/O Size (um2/pin)	137	-	500	10,175	**229**	**229**
I/O Data Rate (Gbps/pin)	1.1	1.25	8	25	**1.6**	**4**
I/O Swing Voltage (V)	0.8	1.0	0.3	0.3	**0.85**	**0.4**
Interface Energy Efficiency (pJ/b)	0.38	0.75	0.56	1.17	**0.85**	**0.46**
I/O Energy Efficiency (pJ/b)	<0.38[(b)]	<0.75[(b)]	0.073	0.55[(c)]	**0.44**	**0.10**
Shoreline Bandwidth Density (Gbps/mm)	297	*900*	672	354	**205**	**1,024**
Area Bandwidth Density (Gbps/mm2)	803	500	1600	516	**574**	**1,705**
Delays (ns)	2.8	7.2	5.5	-	**3.75[(d)]**	**1.5[(d)]**

[(a)] 3D-Plug synchronous version for passive link and short reach.
[(b)] I/O energy is conservatively estimated to be less than the reported interface energy.
[(c)] TX + RX derived from the reported breakdown.
[(d)] I/O TX to RX delay.

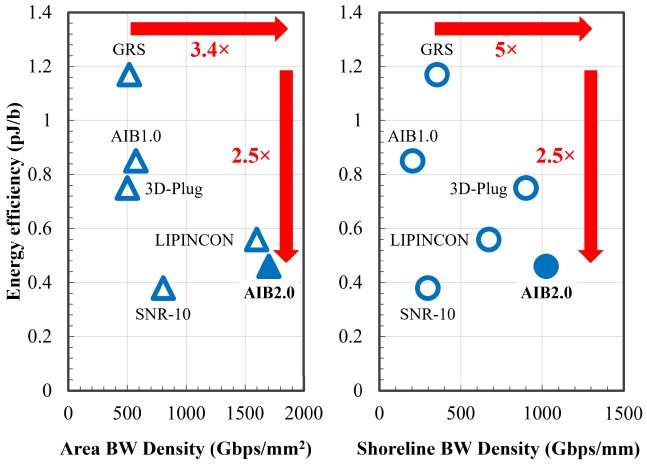


Fig. 16. Energy versus area bandwidth density (left) and shoreline bandwidth density (right).

TABLE III
WORKLOAD MAPPING RESULTS

Domain	Workload	Frame Size	Throughput [(a)]	Utilization
DNN [(b)]	AlexNet	227×227	178.0 frame/s	61%
	VGG-16	227×227	59.7 frame/s	87%
	Tiny-YOLO	41×416	117.3 frame/s	81%
	LeNet	32×32	143.6K frame/s	65%
128×16 MIMO Detection	MMSE Filtering	N/A	14.4GS/s [(c)]	100%
	Matched Filtering	N/A	2.4 GS/s [(c)]	100%
Image Filtering	16 5×5 Filters	1280×720	448.6 frame/s	59%
	16 3×3 Filters	1280×720	807.8 frame/s	42%

[(a)] At 400MHz clock frequency.
[(b)] Softmax and pooling are not included in all DNN workload.
[(c)] Giga QAM symbols per second.

compares the energy efficiency, area bandwidth density, and shoreline bandwidth density for the die-to-die interfaces. The AIB 2.0 interface outperforms the rest, with improvements of 2.5 times, 3.4 times, and 5 times in energy efficiency, area bandwidth density, and shoreline bandwidth density, respectively, over the GRS interface.

We demonstrate workload mapping for various applications that can utilize Arvon, including DNNs, MIMO signal processing, and image filtering. The workload size, overall throughput, and utilization are summarized in Table III. In addition to the commonly used DNN models, the 128 × 16 MIMO detection workload utilizes 128 receive antennas to detect 16 single-antenna users. The processing involved in this task includes the MMSE filtering operation, which requires the computation of the filter matrix using MMM and subsequent application of the filter matrix using MMM. To carry out these operations, MMM kernels with matrix sizes of 32 × 256, 256 × 32, 32 × 32, and 32 × 1 are required for this workload. These kernels can be efficiently mapped to the PE array with 100% utilization. The image filtering workload involves 16 filters of size 5 × 5 and 16 filters of size 3 × 3. These 2-D filters are applied to image frames of size 1280 × 720. The conv kernels are employed to carry out these operations. However, due to the small filter sizes, the utilization is lower than that of other workloads. The results from these sample workloads demonstrate that Arvon's heterogeneous SiP architecture provides flexibility, performance, and efficiency for NN and comm processing.

VII. CONCLUSION

Arvon is a heterogeneous SiP that integrates an FPGA chiplet and two DSP chiplets using EMIBs. This integration enables Arvon to leverage the FPGA's flexibility as a host while benefiting from the DSPs' high computational performance and efficiency.

The key feature of the SiP is the parallel, short-reach AIB 1.0 and AIB 2.0 interfaces for seamlessly connecting the chiplets. The I/O cells are designed to be compact, predominantly digital, and synthesizable. The cells are flexible

and can support several modes. Additionally, they employ mode-dependent power gating and two-level clock distribution, improving energy efficiency. Our implementation of the low-swing 4-Gb/s AIB 2.0 interface using 36-μm-pitch microbumps demonstrates an energy efficiency of 0.10 pJ/b, or 0.46 pJ/b including the adapter, and a bandwidth density of 1.024-Tb/s/mm shoreline and 1.705-Tb/s/mm^2 area. The interface is abstracted using an AXI-compatible bus protocol, simplifying its use by the host and DSP.

Each DSP chiplet in Arvon follows a low-latency systolic array architecture, featuring 3072 FP16 PEs. The PEs are hierarchically organized into three clusters, with eight 32-PE units per cluster. This granular organization allows for the parallel execution of multiple workloads concurrently. Each DSP chiplet provides a peak performance of 4.14 TFLOPS at a power efficiency of 1.8 TFLOPS/W. We developed a systematic procedure for mapping workloads onto Arvon and demonstrated diverse workloads that can be accelerated by Arvon to achieve competitive performance and utilization.

ACKNOWLEDGMENT

The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

[1] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[2] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[3] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018.

[4] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," 2021, *arXiv:2106.08962*.

[5] M.-S. Lin et al., "A 7-nm 4-GHz arm-core-based CoWoS chiplet design for high-performance computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 4, pp. 956–966, Apr. 2020.

[6] B. Zimmer et al., "A 0.32–128 TOPS, scalable multi-chip-module-based deep neural network inference accelerator with ground-referenced signaling in 16 nm," *IEEE J. Solid-State Circuits*, vol. 55, no. 4, pp. 920–932, Apr. 2020.

[7] J. W. Poulton et al., "A 1.17-pJ/b, 25-Gb/s/pin ground-referenced single-ended serial link for off- and on-package communication using a process- and temperature-adaptive voltage regulator," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 43–54, Jan. 2019.

[8] U. Rathore, S. S. Nagi, S. Iyer, and D. Markovic, "A 16 nm 785 GMACs/J 784-core digital signal processor array with a multilayer switch box interconnect, assembled as a 2 × 2 dielet with 10 μm-pitch inter-dielet I/O for runtime multi-program reconfiguration," in *Proc. IEEE Int. Solid- State Circuits Conf. (ISSCC)*, Feb. 2022, pp. 52–54.

[9] S. S. Nagi, U. Rathore, K. Sahoo, T. Ling, S. S. Iyer, and D. Markovic, "A 16-nm 784-core digital signal processor array, assembled as a 2 × 2 dielet with 10 μm pitch interdielet I/O for runtime multiprogram reconfiguration," *IEEE J. Solid-State Circuits*, vol. 58, no. 1, pp. 111–123, Jan. 2023.

[10] P. Vivet et al., "IntAct: A 96-core processor with six chiplets 3D-stacked on an active interposer with distributed interconnects and integrated power management," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 79–97, Jan. 2021.

[11] W. Tang et al., "Arvon: A heterogeneous SiP integrating a 14 nm FPGA and two 22 nm 1.8 TFLOPS/W DSPs with 1.7 Tbps/mm2 AIB 2.0 interface to provide versatile workload acceleration," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Jun. 2023, pp. 1–2.

[12] R. Mahajan et al., "Embedded multi-die interconnect bridge (EMIB)—A high density, high bandwidth packaging interconnect," in *Proc. IEEE 66th Electron. Compon. Technol. Conf. (ECTC)*, May 2016, pp. 557–565.

[13] G. Duan, Y. Kanaoka, R. McRee, B. Nie, and R. Manepalli, "Die embedding challenges for EMIB advanced packaging technology," in *Proc. IEEE 71st Electron. Compon. Technol. Conf. (ECTC)*, Jun. 2021, pp. 1–7.

[14] *AIB Specifications*. Accessed: Dec. 13, 2023. [Online]. Available: https://github.com/chipsalliance/AIB-specification

[15] B. Razavi, "The StrongARM latch [A circuit for all seasons]," *IEEE Solid StateCircuits Mag.*, vol. 7, no. 2, pp. 12–17, Jun. 2015.

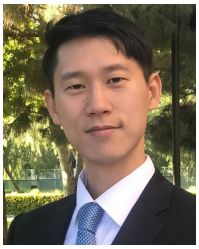
[16] M. Miyahara, Y. Asada, D. Paik, and A. Matsuzawa, "A low-noise self-calibrating dynamic comparator for high-speed ADCs," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Nov. 2008, pp. 269–272.

[17] S.-G. Cho, W. Tang, C. Liu, and Z. Zhang, "PETRA: A 22 nm 6.97 TFLOPS/W AIB-enabled configurable matrix and convolution accelerator integrated with an Intel Stratix 10 FPGA," in *Proc. Symp. VLSI Circuits*, Jun. 2021, pp. 1–2.

[18] *Intel Stratix 10 TX 2800 Specification*. [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/210283/intel-stratix-10-tx-2800-fpga/specifications.html

Wei Tang (Member, IEEE) received the B.S. degree from National Chiao-Tung University, Hsinchu, Taiwan, in 2011, and the M.S. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2019.

He was a Visiting Ph.D. Scholar at Lund University, Lund, Sweden, and a Graduate Research Intern at Intel Labs, Hillsboro, OR, USA. He is currently an Assistant Research Scientist with the Department of Electrical Engineering and Computer Science, University of Michigan. His research interests include high-speed, energy-efficient, and flexible VLSI designs for communications, machine learning, and robotics.



Sung-Gun Cho received the B.S. and M.S. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2010 and 2012, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA, in 2020.

From 2012 to 2015, he was with SK Hynix Inc., Seongnam, South Korea, where he worked on system-on-chip (SoC) design and implementation of error control coding. In 2020, he joined Intel Programmable Solutions Group CTO Office, San Jose, CA, USA, as an SoC Design Engineer to develop chiplets for various applications. His current research interests include energy-efficient, high-performance design for signal processing, error control coding, and machine-learning acceleration.

Tim Tri Hoang received the B.S. degree in EECS from the University of California at Berkeley (UC Berkeley), Berkeley, CA, USA.

He is currently a Principal Engineer at Intel (PSG/CTO Office), San Jose, CA, USA. He is also a Parallel Interface Architect for the Intel CHIPS, PIPES, and Domestic Foundry DARPA programs. He has more than ten years of experience in transceiver technology. Before working with transceivers, he was part of the analog design team responsible for developing PLLs and analog IPs in various FPGA platforms. He has a number of technical articles and holds many issued patents in the areas of programmable circuits, I/O, and analog and mix-signal IPs.

Jacob Botimer (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2016 and 2019, respectively.

He did internships with the Power Electronics Group, Texas Instruments, Dallas, TX, USA, in 2015 and 2016. In 2019, he joined the startup company MemryX, Ann Arbor, where he has been working on computer architecture and circuit design for hardware accelerators. His research interests include in-memory computing and 2.5-D integration.

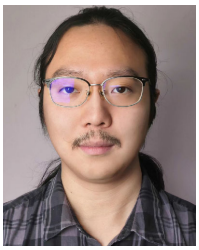
Wei Qiang Zhu, photograph and biography not available at the time of publication.

Ching-Chi Chang received the B.S. degree in electronic engineering from Chang Gung University, Taoyuan, Taiwan, and the M.S. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA.

He was a Memory IO PHY Circuit Design Engineer and a Design Verification Engineer for DDR memory controller and memory subsystem with Altera Corporation, San Jose, CA, USA. He is currently working on advanced chiplet/heterogeneous integration research projects at Intel Programmable Solutions Group CTO Office, San Jose.

Cheng-Hsun Lu (Student Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree from the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Michigan, Ann Arbor, MI, USA.

His current research interests include energy-efficient, high-performance VLSI circuits and systems for machine learning, DSP algorithms, and forward error correction.



Junkang Zhu (Graduate Student Member, IEEE) received the B.S. degree in physics from Nanjing University, Nanjing, China, in 2017, and the M.S. degree in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA, in 2019, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

His current research interests include novel microarchitecture, SoC design, and system integration for machine learning and computer vision applications.

Yaoyu Tao (Member, IEEE) received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, the M.S. degree from Stanford University, Stanford, CA, USA, and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA, all in electrical engineering.

He also held industry research positions at Qualcomm Wireless Research and Development, San Jose, CA, USA, Oracle VLSI Laboratory, Santa Clara, CA, USA, and Texas Instruments Kilby Laboratory, Dallas, TX, USA. He is currently an Assistant Professor with the Institute of Artificial Intelligence, Peking University, Beijing, China, where he is also affiliated with the PKU's School of Integrated Circuits. His research interests include efficient VLSI and hardware system design enabled by emerging devices, such as memristors and Moiré superlattice semiconductors.

Dr. Tao was a recipient of the NSFC Excellent Young Scholarship, the Best Paper Award at the IEEE Global Communication Conference, and the Qualstar Awards at Qualcomm Wireless Research and Development, and was selected as the Young Fellow at the Beijing Academy of Artificial Intelligence.

Tianyu Wei (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Southeast University, Nanjing, China, in 2021. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Michigan, Ann Arbor, MI, USA.

His current research interests include multi-chiplet system integration, system-on-chip design, and mixed-signal design.

Naomi Kavi Motwani, photograph and biography not available at the time of publication.

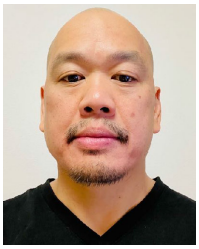
Mani Yalamanchi, photograph and biography not available at the time of publication.

Ramya Yarlagadda, photograph and biography not available at the time of publication.

Sirisha Rani Kale received the bachelor's degree in computer science from Pondicherry University, Kalapet, India, in 2003, and the master's degree in VLSI CAD from Manipal University, Manipal, India, in 2009.

She joined Intel Labs, Hillsboro, OR, USA, in 2004, where she worked on the PHY design of the industry's first 48-iA core Tera-scale prototype. She is primarily responsible for backend design, system-on-chip (SoC) integration, STA, and tape-out verification. Some of the emerging research areas that she contributed to are quantum computing, wireless chip-to-chip communications, neuromorphic computing, and accelerator and RISC processors. Her research interests include VLSI design and VLSI architecture for AI.

Mark Flanigan, photograph and biography not available at the time of publication.



Allen Chan received the B.S. degree in electrical engineering from the University of California at Davis, Davis, CA, USA.

He is currently a Principal Engineer at the Intel Programmable Solutions Group's CTO Office, San Jose, CA, USA. He specializes in the field of electrical engineering ranging from analog/mixed signals integrated circuit design, high-speed chip interfaces, multipackage integration, compute optimization, workload acceleration, digital signal processing, copackaged optics, wireless and wireline communications, machine learning, and artificial intelligence.

Thungoc (Tina) Tran received the B.S. degree in electrical engineering and computer science from the University of California at Berkeley, Berkeley, CA, USA, in 1987.

She holds a position as an SOC Hardware Engineer at the Interconnect Strategy, Intel Programmable Solutions Group (PSG) CTO office, San Jose, CA, USA. Her past experience was in transceiver IP design since 2002 at Altera Corporation, San Jose, now part of Intel. She is interested in integrating and manufacturing IPs with FPGA chiplet in the same platform.

Sergey Shumarayev received the B.S. and M.S. degrees.

He attended Moscow Technical University (signal and systems), Moscow, Russia; the University of California at Berkeley (UC Berkeley) (dual in EECS and material science), Berkeley, CA, USA; and Cornel University (EE), Ithaca, NY, USA. He is currently a Senior Principal Engineer with the Strategy Team, which is part of the Intel Programmable Solutions Group (PSG) CTO Office, San Jose, CA, USA. He is the Principal Investigator for the Intel DARPA CHIPS Team, DARPA PIPES (optical integration), and DARPA HI3 (domestic foundry) programs. He has 25 years of global management, technical, and architecture experience with a focus on programmable I/O, interconnect, mix-signal IP, and SERDES architectures. In his management experience as a Sr. Director at Altera Corporation, San Jose, he was responsible for the overall global custom IP team of 250 engineers that provided all Altera IPs. He drove the development of the first overall heterogeneous multichip assembly (MCA) strategy at Altera, architecting MCA interfaces to disintegrate the monolithic flagship Stratix 10 FPGA family to adopt heterogeneous multifoundry technologies, through EMIB and establishing the Intel PSG tile strategy. His CTO team's latest DARPA-funded CHIPS research resulted in successful technology transfer to new Intel's product line leapfrogging competition. He is a Distinguished PSG Architect and holds more than 215 issued patents in the area of universal programmable interfaces, IO, analog, and mixed-signal IP.

Zhengya Zhang (Senior Member, IEEE) received the B.A.Sc. degree in computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2003, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Berkeley (UC Berkeley), Berkeley, CA, USA, in 2005 and 2009, respectively.

He has been a Faculty Member with the University of Michigan, Ann Arbor, MI, USA, since 2009, where he is currently a Professor with the Department of Electrical Engineering and Computer Science. His research interests include low-power and high-performance VLSI circuits and systems for computing, communications, and signal processing.

Dr. Zhang was a recipient of the University of Michigan College of Engineering Neil Van Eenam Memorial Award in 2019, the Intel Early Career Faculty Award in 2013, the National Science Foundation CAREER Award in 2011, and the David J. Sakrison Memorial Prize from UC Berkeley in 2009. He has been serving on the Technical Program Committee of the IEEE Custom Integrated Circuits Conference (CICC) since 2019. He served on the Technical Program Committee of the IEEE VLSI Symposium on Technology and Circuits from 2019 to 2022. He served as an Associate Editor for the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS from 2015 to 2022, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART I: REGULAR PAPERS from 2013 to 2015, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART II: EXPRESS BRIEFS from 2014 to 2015. He is an IEEE Solid-State Circuits Society Distinguished Lecturer from 2023 to 2024.