Arvon: A Heterogeneous SiP Integrating a 14nm FPGA and Two 22nm 1.8TFLOPS/W DSPs with 1.7Tbps/mm^2 AIB 2.0 Interface to Provide Versatile Workload Acceleration

Wei Tang [1*], Sung-Gun Cho [2*], Tim Tri Hoang [2*], Jacob Botimer [1], Wei Qiang Zhu [2], Ching-Chi Chang [2], Cheng-Hsun Lu [1]
Junkang Zhu [1], Yaoyu Tao [1], Tianyu Wei [1], Naomi Kavi Motwani [1], Mani Yalamanchi [2], Ramya Yarlagadda [2], Sirisha Kale [2],
Mark Flanigan [2] , Allen Chan [2], Thungoc Tran [2], Sergey Shumarayev [2], Zhengya Zhang [1]
[1] University of Michigan; [2] Intel corporation; [*] Equal contribution

Abstract

Arvon is a heterogeneous system in a package (SiP) that integrates a 14nm FPGA chiplet with two dense and efficient 22nm DSP chiplets through Embedded Multi-die Interconnect Bridges (EMIBs) as illustrated in Fig. 1. The chiplets communicate via a 1.536Tbps Advanced Interface Bus (AIB) 1.0 interface and a 7.68Tbps AIB 2.0 interface. We demonstrate the first-ever AIB 2.0 I/O prototype using 36µm-pitch microbumps, achieving 4Gbps/pin at 0.10pJ/b (0.46pJ/b including adapter), and a bandwidth density of 1.024Tbps/mm-shoreline and 1.705Tbps/mm^2-area. Arvon is programmable, supporting workloads from neural network (NN) to communication processing (comm) and providing a peak performance of 4.14TFLOPS (FP16, half-precision floating-point) by each DSP chiplet at 1.8TFLOPS/W. A compilation flow is developed to map workloads across FPGA and DSPs to optimize performance and utilization.

Arvon SiP for Versatile Workload Acceleration

Inside Arvon SiP, the host FPGA is connected to DSP1 by EMIB through an AIB 1.0 interface, and DSP1 is connected to DSP2 (a rotated version of the DSP chiplet) by EMIB through an AIB 2.0 interface [1] (Fig. 2). Arvon's primary operation modes (Mode 1 and 3 in Fig. 2) are to have DSPs provide offload for common kernel functions including matrix-matrix multiplication (MMM) and 2D convolution (conv) that are dominant in NN and comm workloads. To complement the kernel functions and support complete workloads, the FPGA's programmable logic is utilized for data arrangements (e.g., transpose and shuffle) and special functions. One realization of the host on FPGA comes in the form of an instruction-based processor (Fig. 2), consisting of instruction memory, data memories for inputs, outputs, and weights, direct memory access (DMA) unit to coordinate data transfer with the DSP chiplets. Instructions govern the DSP configuration in runtime, the data flow between data memories and DSPs, and pre- and post-DSP processing. Arvon's secondary operation mode is to combine DSP1 and DSP2 (Mode 2) to extend the compute capacity. Alternatively, DSP2 can be replaced by a frontend (FE) chiplet, e.g., an optical tile or an ADC tile to implement an entire communication chain from FE to DSP and FPGA, or the reverse.

DSP Chiplet Design for Flexibility, Utilization and Efficiency

A DSP chiplet consists of an AIB 1.0 interface, an AIB 2.0 interface, three DSP clusters, as well as a ring PLL to generate clock, and GPIOs for global configuration (Fig. 2). A DSP cluster contains a flexible vector engine, a bypass buffer, a rotation block to support data framing, and an AXI-compatible bus to abstract the AIB I/Os. Each DSP cluster provides up to 1.38TFLOPS at 675MHz. The core of a DSP cluster is a vector engine made of 4 instances of a 2D systolic array [2], each consisting of 256 processing elements (PEs), and the entire vector engine provides 1,024 PEs in total to support MMM and conv in FP16. Configured by instructions, the vector engine allows inputs to be streamed in for continuous computation. The vector engine provides mapping flexibility: 1) the 4 systolic arrays can be mapped independently; and 2) the 256 PEs in each array can be configured in 32-PE units, supporting 1 to 8 independent workloads. A compilation flow (Fig. 3) is developed to map workloads. The source code is compiled into configurations of conv (filter and input sizes) and/or MMM (matrix dimensions) kernels, and intermediate processing between kernels, considering factors including utilization, data reuse, and end-to-end latency. We demonstrate workload mapping for DNNs, image filtering, and MIMO detection (Fig. 3).

AIB Interface for Bandwidth Density and Energy Efficiency

Both AIB 1.0 and AIB 2.0 [1] are source-synchronous, short-reach, low-latency, and parallel I/O interfaces. 24 channels of AIB 1.0/2.0 are instantiated on the west/east side of the DSP chiplet. One AIB 1.0 channel has 62 pins (1 TX clock, 1 RX clock, 20 TX data, 20 RX data, and other controls), using full-rail signaling. Each data pin is source-synchronized and supports up to 2Gbps (DDR at 1GHz clock). The AIB 1.0 interface provides an aggregate bandwidth of up to 1.92Tbps. Compared to AIB 1.0, AIB 2.0 (Fig. 4) doubles both the data rate per pin and the number of data pins per channel and improves the energy efficiency by low-swing signaling. An automated RX clock phase tuning is employed to find the optimal RX sampling point: the TX sends a PRBS sequence; the phase of the received clock is adjusted by a configurable delay line to sample the RX data; the RX checks for error and detects the eye boundaries; and the optimal delay is set to sample at the estimated midpoint of the eye. The AIB 2.0 interface provides an aggregate bandwidth of up to 7.68Tbps.

A unified I/O cell (Fig. 5) is designed to support both AIB 1.0 (full-rail) and AIB 2.0 (full-rail and 0.4V swing). At TX, a switchable PMOS/NMOS driver is employed for the full-rail/low-swing pull-up. Up to four drivers can be wired together to support tunable driving strength to accommodate channel variation and trade between I/O speed and power. At RX, a standard-cell buffer is employed for a full-rail input or a regenerative comparator for a low-swing input. The comparator is an optimized StrongARM latch [3] to minimize the mean offset to 4.1mV without calibration (Fig.5). PMOS is used to achieve an improved detection of low-swing inputs. With a simple reference voltage generator, the comparator can reliably sense an input down to 0.38V at 2GHz DDR. Since low-skew clock distribution is essential for high-speed parallel I/Os, we adopt a global balanced H-tree plus a local clock mesh to limit the worst clock skew to 8ps.

Measurement Results

The DSP chiplet was fabricated in a 22nm FinFET technology, occupying 32.3mm^2 (Fig. 6). To construct Arvon, an FPGA chiplet and two DSP chiplets were co-packaged and interconnected via two 10-layer EMIBs, using 36µm-pitch µbumps with an average wire length of 1.5mm/0.85mm for the AIB 1.0/2.0 side. Measured at room temperature, a 0.85V I/O voltage, and an 800MHz clock (limited in this design by the FPGA), the AIB 1.0 I/O consumes 0.44pJ/b, and the transfer latency is kept at 3.75ns. Measured at room temperature, a 0.4V I/O voltage, and a 2GHz clock, the AIB 2.0 I/O consumes 0.10pJ/b (0.46pJ/b including adapter) with a transfer latency of 1.5ns.

Arvon's AIB I/O interfaces are compared to the state-of-the-art SiP I/O interfaces in Fig. 7. Similar to the AIB interfaces, SNR-10 [4] (16nm, full-rail, 10µm-pitch), 3D-Plug [5] (28nm, full-rail, 20µm-pitch), and LIPINCON [6] (7nm, 0.3V swing, 40µm-pitch) are parallel I/O interfaces. Among them, LIPINCON provides the highest data rate of 8Gbps/pin and the lowest I/O energy of 0.073pJ/b; 3D-Plug provides the highest bandwidth density of 900Gbps/mm-shoreline; and SNR-10 has the smallest I/O size of 137µm^2. GRS [7] (16nm, 150µm-pitch) is a high-speed serial I/O interface with 25Gbps/pin at 1.17pJ/b. Our AIB 2.0 prototype (22nm, 0.4V swing, 36µm-pitch) demonstrates a competitive I/O energy of 0.10pJ/b (0.46pJ/b with adapter), and the highest bandwidth density of 1.0Tbps/mm-shoreline and 1.7Tbps/mm^2-area (Fig. 8). Arvon's heterogeneous SiP architecture provides flexibility, performance, and efficiency for NN and comm workloads.

Acknowledgements

This work was supported in part by DARPACHIPS program and ONR Grant N00014-17-1-2992. 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] AIB Specifications, https://github.com/chipsalliance/AIB-specification.
[2] S.-G. Cho, et al., VLSI 2021.
[3] B. Razavi, MSSC 2015.
[4] U. Rathore, et al., ISSCC 2022.
[5] P. Vivet, et al., ISSCC 2021.
[6] M. Lin, et al., ISSCC 2010.
[7] J. W. Poulton, ISSCC 2019.

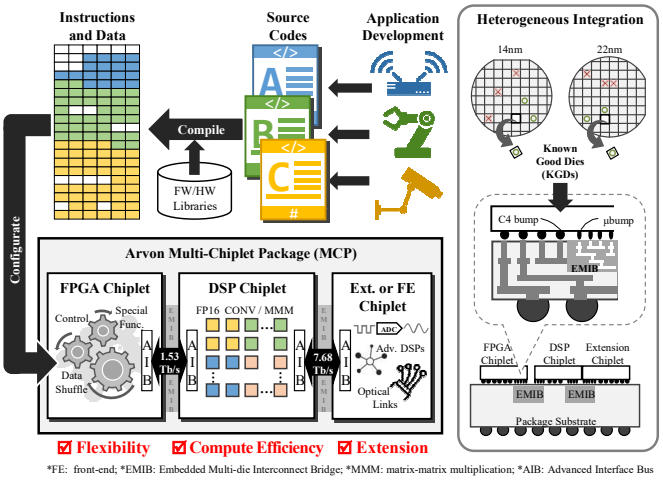


Fig. 1. Arvon SiP heterogeneously integrating FPGA, DSP, FE chiplets for flexible workload mapping.

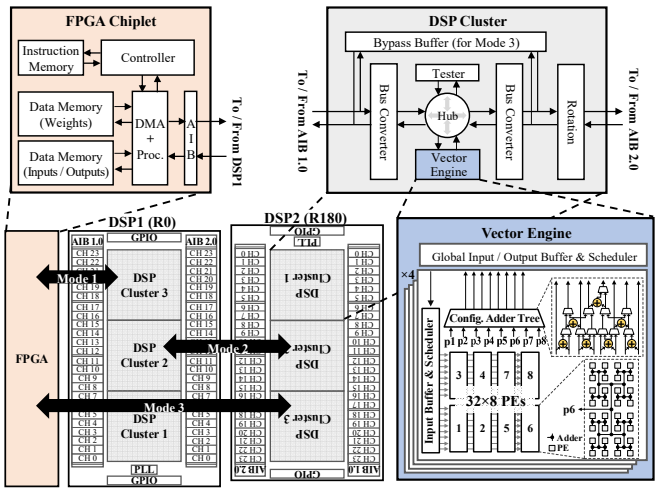
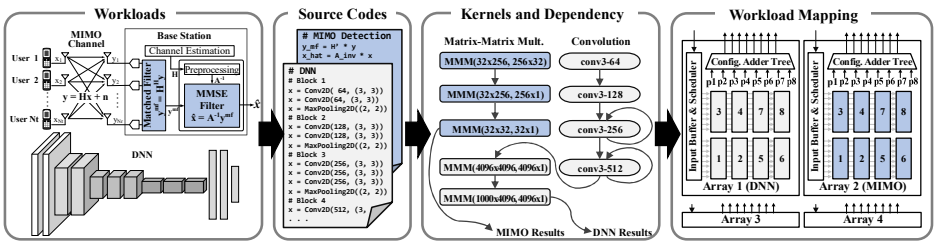


Fig. 2. Data flow modes of Arvon SiP, instruction-based FPGA architecture, DSP cluster, and 2D systolic-based vector engine.



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	416×416	117.3 frame/s	81%
	LeNet	32×32	143.6K frame/s	65%
128x16 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.

Fig. 3. Compilation flow for workload mapping and mapping results for DNN, MIMO communication, and image filtering.

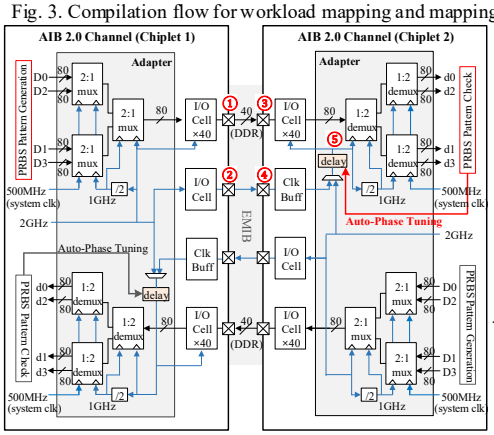


Fig. 4. AIB2.0 channel top-level diagram and automated clock phase tuning.

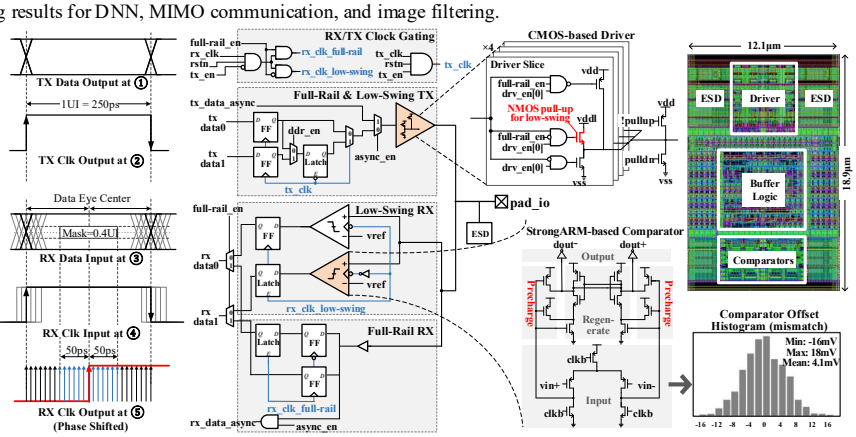
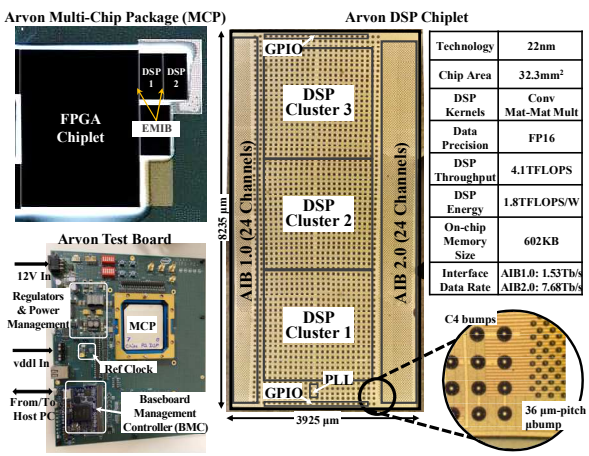


Fig. 5. Schematic and layout of the unified AIB I/O cell for full-rail and low-swing with CMOS-based TX and comparator-based RX, and comparator offset.



Technology	22nm
Chip Area	32.3mm²
DSP Kernels	Conv Mat-Mat Mult
Data Precision	FP16
DSP Throughput	4.1TFLOPS
DSP Energy	1.8TFLOPS/W
On-chip Memory Size	602KB
Interface Data Rate	AIB1.0: 1.53Tb/s AIB2.0: 7.68Tb/s

Fig. 6. Arvon SiP, test board, DSP chiplet photo, and DSP feature table.

	Rathore ISSCC22 [4]	Vivet JSSC21 [5]	Lin JSSC20 [6]	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 (μm)	10	20	40	150	36	
Technology	16nm	28nm	7nm	16nm	22nm	
I/O Size (μm²/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. 7. Comparison table of state-of-the-art SiP interface designs.

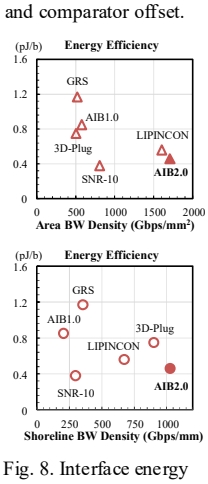


Fig. 8. Interface energy versus bandwidth density.