A 0.11 pJ/Op, 0.32-128 TOPS, Scalable Multi-Chip-Module-based Deep Neural Network Accelerator with Ground-Reference Signaling in 16nm

Brian Zimmer[1], Rangharajan Venkatesan[1], Yakun Sophia Shao[1], Jason Clemons[3], Matthew Fojtik[2], Nan Jiang[5], Ben Keller[1], Alicia Klinefelter[2], Nathaniel Pinckney[3], Priyanka Raina[1], Stephen G. Tell[2], Yanqing Zhang[1], William J. Dally[1,6], Joel S. Emer[4,7], C. Thomas Gray[2], Stephen W. Keckler[3], Brucek Khailany[3]

[1]NVIDIA, Santa Clara, CA; [2]NVIDIA, Durham, NC; [3]NVIDIA, Austin, TX; [4]NVIDIA, Westford, MA; [5]NVIDIA, St. Louis, MO; [6]Stanford University, Stanford, CA; [7]MIT, Cambridge, MA. Email: bzimmer@nvidia.com

Abstract

This work presents a scalable deep neural network (DNN) accelerator consisting of 36 chips connected in a mesh network on a multi-chip-module (MCM) using ground-referenced signaling (GRS). While previous accelerators fabricated on a single monolithic die are limited to specific network sizes, the proposed architecture enables flexible scaling for efficient inference on a wide range of DNNs, from mobile to data center domains. The 16nm prototype achieves 1.29 TOPS/mm^2, 0.11 pJ/op energy efficiency, 4.01 TOPS peak performance for a 1-chip system, and 127.8 peak TOPS and 2615 images/s ResNet-50 inference for a 36-chip system.

Introduction

Deep neural networks (DNNs) have diverse performance, accuracy, and power targets. Building a dedicated accelerator for each target is often prohibitive due to high design and manufacturing costs. We propose using low-energy, high-bandwidth chip-to-chip communication links to assemble systems with various compute capacities from a single inference accelerator chip. Flexible and efficient scalability of the processing elements (PE) is supported via a distributed tile-based architecture connected by packet-switched network-on-chip (NoC) and network-on-package (NoP) routers.

Scalable Neural Network Accelerator

Figure 1 shows the same die packaged into 1-chip, 4-chip, and 36-chip systems on an organic substrate. DNN weights are tiled statically across the on-chip memories, supporting storage for 0.5-18 million 8b weights and peak performance of 4-128 TOPS for 1-36 chip systems. Ground-referenced signaling (GRS) transceivers [1] are used for inter-chip communication. The scalability is not inherently limited by package size, as these transceivers can also communicate short distances on a PCB between multiple packages.

Figure 2 shows the architecture of the proposed inference accelerator. It integrates 36 chips on a package, connected via a mesh topology. Each die contains a NoP router, a global buffer (GB), a RISC-V control processor, and 16 PEs connected via a mesh NoC with 64-bit wide routers. The NoP router transfers packets between the NoC and neighboring chips via GRS transceivers. The GB contains a 64KB unified buffer that acts as second-level storage for activations. The PEs execute convolutional layers, fully-connected layers, and post-processing functions like bias addition, ReLU, and pooling. Each PE includes an 8KB input buffer, a 32KB weight buffer, and a 3KB accumulation buffer. A large weight buffer enables persistent weight storage to minimize DRAM traffic. Input activations are reused spatially across 8 parallel lanes of vector multiply-accumulate (MAC) units, and weights are reused temporally. Each vector MAC performs an 8-way dot product

and accumulates the partial sum into the accumulation buffer every cycle. An 8b fixed-point precision for weights and activations and 24b accumulation precision were chosen to optimize energy efficiency without incurring loss of accuracy.

Efficient Multi-Chip Inference

Figure 3 shows an example multicast operation where input activations are sent from one chip's GB to other chips' PEs across the NoC and NoP. The IO region of the die comprises 8 chip-to-chip GRS transceiver macros, where 4 are configured as receivers (RX) and 4 as transmitters (TX). Each transceiver macro has 4 data lanes and a clock lane with configurable speed from 11Gbps/pin to 25Gbps/pin, consumes 0.82-1.75 pJ/bit, and occupies 0.26mm^2 for a total peak chip bandwidth of 800Gbps and peak bandwidth density of 384Gbps/mm^2. Compared to previous MCM interconnect [2], GRS has about 3.5× higher bandwidth per chip area and lower energy per bit. Measured results show an eye opening of 0.7UI at 25Gbps.

Figure 4 shows how parallelism and reuse in DNNs are leveraged through the tiling of computation spatially across different chips and PEs to maximize MAC utilization while minimizing communication power. The RISC-V control core configures communication between PEs and GBs via software-managed registers. At the package level: 1) weights are split between different chips along input channel (C) and output channel (K) dimensions; 2) input activations are multicast along rows of chips with matching C; and 3) output activations are reduced along columns of chips. At the chip level, weights are tiled similarly along the C and K dimensions while input activations are multicast via the NoC. Other tiling options are also supported for layers with fewer input or output channels. At the PE level, weights are loaded once from the weight buffer while traversing the input images (weight stationary dataflow).

Measurement Results

The 6mm^2 inference accelerator (Figure 5) is fabricated in a TSMC 16nm FinFET process. Figure 6 reports 36-chip core energy efficiency for different voltages. Figure 7 presents the measured performance of a 1-chip or 4-chip automotive-scale system running DriveNet [3], an end-to-end DNN framework for self-driving cars. Power and performance measurements begin after the weights have been loaded into each PE's weight SRAM and the inputs have been loaded into the GBs. Figure 8 demonstrates the architecture's scalability with measured performance of a 36-chip datacenter-scale system running each layer of ResNet-50 [4]. The system achieves 2615 images/s at a nominal voltage of 0.85V.

Figure 9 compares the 1-chip, 4-chip, and 36-chip system to prior work. The core (PE, GB, NoP, and RISC-V) power is reported separately from IO (GRS, JTAG, and GPIO) to allow comparison to prior work, which generally reports core power.

Conclusion

The presented 1-chip system achieves 16×-99× higher area efficiency (1.29 TOPS/mm^2), 1.7×-10× better energy efficiency (0.11 pJ/op), and 5.8×-40× higher peak performance

(4.01 TOPS) than prior work at 8b precision [5], [6], [7], and the 36-chip MCM system enables efficient scaling to 185×-1280× higher peak performance (128 TOPS). The scalable architecture enables efficient execution of diverse DNNs.

References

[1] J. Wilson, *et al., ISSCC*, 2018, pp. 276-278. [2] N. Beck, *et al., ISSCC*, 2018, pp. 40-42. [3] M. Bojarski, *et al.*, "End to End Learning for Self-Driving Cars," *arXiv*, 2016. [4] K. He, *et al.,* "Deep Residual Learning for Image Recognition," *arXiv,* 2015. [5] B. Moons, *et al*.," *ISSCC*, 2017, pp. 246-247. [6] Z. Yuan, *et al.*, *VLSI Circuits*, 2018, pp 33-34. [7] J. Lee, *et al.*, *ISSCC*, 2018, pp. 218-219.
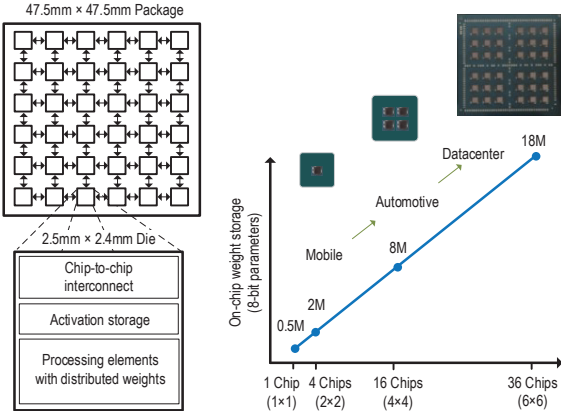
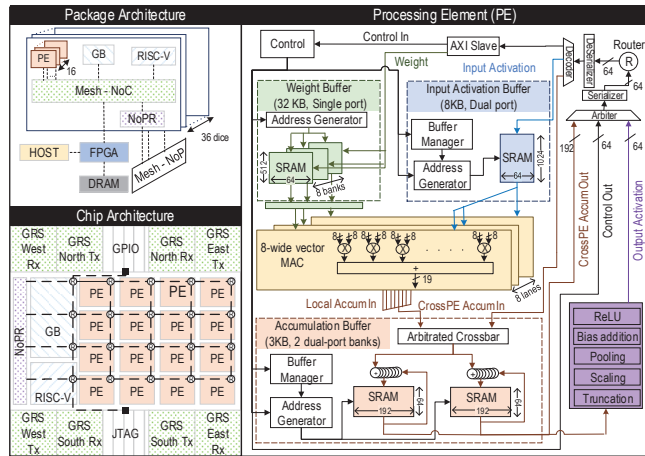Fig. 1: System overview of the proposed MCM-based DL accelerator.
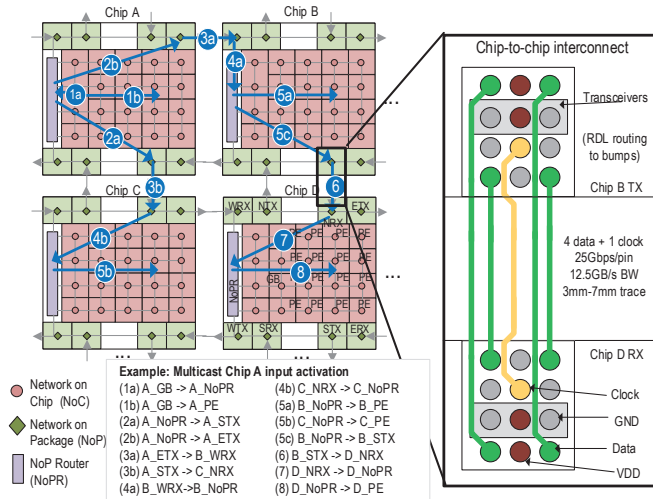
Fig. 2: Package, chip, and PE architecture.

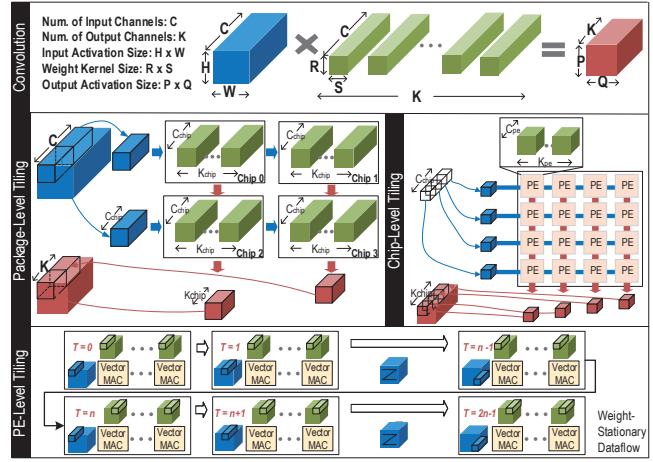Fig. 3: Data movement across chips via NoC and NoP.

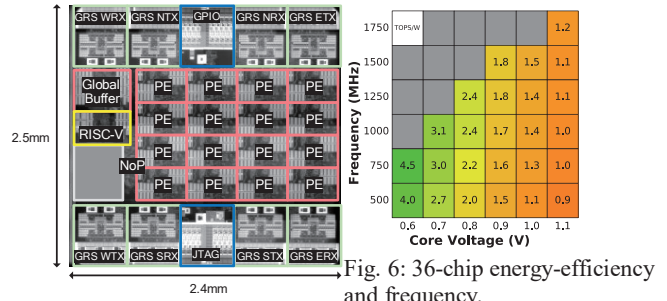Fig. 4: Application tiling at Package, Chip, and PE level.

Fig. 5: Die micrograph.

Fig. 6: 36-chip energy-efficiency and frequency.

System	Total Latency (ms/batch)	Throughput (images/s)	Core Energy (mJ/image)
1-die (batch=1 image)	0.568	1761	0.195
4-die (batch=4 images)	0.611	6548	0.228

Fig. 7: Measurement of a 4-chip system running DriveNet at 0.85V.

Layer	Activation Traffic (KB)		Weights (KB)	Latency (µS)	Core Energy (µJ)	GRS Energy (µJ)
	Input	Output				
conv1-pool1	147	784	9.19	77.09	1922.78	300.76
res2[a-c]_branch2b	196	196	36	13.2	486.57	51.82
res2[a-c]_branch2c	196	784	16	13.21	374.08	51.6
res2[b-c]_branch2a	784	196	16	23.63	909.39	92.96
res2a_branch1	196	784	16	13.21	374.08	51.6
res2a_branch2a	196	196	4	8.46	219.44	32.96
res3[a-d]_branch2b	98	98	144	13.53	470.06	53.16
res3[a-d]_branch2c	98	392	64	11.36	420.77	44.59
res3[b-d]_branch2a	392	98	64	11.75	447.3	46.19
res3a_branch1	784	392	128	8.74	373.66	34.46
res3a_branch2a	784	98	32	10.37	366.14	40.67
res4[a-f]_branch2b	49	49	576	19.72	626.83	77.48
res4[a-f]_branch2c	49	196	256	8.38	372.79	33.09
res4[b-f]_branch2a	196	49	256	8.66	390.46	34.17
res4a_branch1	392	196	512	10.98	595.32	43.62
res4a_branch2a	392	49	128	7.21	258.98	28.25
res5[a-c]_branch2b	24.5	24.5	2304	22	742.49	86.6
res5[a-c]_branch2c	24.5	98	1024	10.98	335.12	43.2
res5[b-c]_branch2a	98	24.5	1024	12.25	370.03	48.18
res5a_branch1	196	98	2048	20.43	679.14	80.61
res5a_branch2a	196	24.5	512	11.62	326.16	45.6
Total (batch=2 images)			2615 images/s	16.1 mJ/image	1.95 mJ/image	

Fig. 8: Measurement of 36-chip system running ResNet-50 at 0.85V.

	ISSCC 2017, B. Moons, ENVISION [5]	VLSI 2018, Z. Yuan, STICKER [6]	ISSCC 2018, J.Lee, UNPU [7s]	Proposed*		
				1 Chip	4 Chip (2 × 2)	36 Chip (6 × 6)
Technology	28nm	65nm	65nm	16nm		
Cumulative Core Area	1.87 mm²	7.8 mm²	13 mm²	3.1 mm²	12.4 mm²	111.6 mm²
Cumulative Die Area	unknown	12 mm²	16 mm²	6 mm²	24 mm²	216 mm²
Precision	4b,8b,16b	8b	1-16b	8b		
On-Chip SRAM (MB)	0.14	0.17	0.25	0.625	2.5	22.5
Supply Voltage (V)	1	0.67-1.1	0.63-1.1	0.41-1.2	0.52-1.2	0.52-1.1
Frequency (MHz)	200	200	5-200	161-2001	515-1998	484-1797
Core Power (mW)	165	21-248	3.2-297	30-4160	630-16,420	5,310-106,090
GRS Power† (mW)	n/a	n/a	n/a	n/a	215-220	3,840-4,090
MACs per cycle	512 @8b	256	1,728@8b	1,024	4,096	36,864
Performance (TOPS)	~0.15@8b	0.1	0.69@8b	0.32-4.01	3.93-15.7	32.5-127.8
Core Energy Efficiency (pJ/op)	~1.1@8b	0.96	~0.18@8b	0.105-1.04	0.160-1.05	0.164-8.30
Core Area Efficiency (TOPS/mm^2)	0.08	0.013	0.053	0.10-1.29	0.32-1.27	0.29-1.15

* Measured results reported for 40% density weights and input activations   †11Gbps mode

Fig. 9: Comparison table.