NetFlex: A 22nm Multi-Chiplet Perception Accelerator in High-Density Fan-Out Wafer-Level Packaging

Teyuh Chou [1], Wei Tang [1], Mihai D. Rotaru [2], Chester Liu [1], Rahul Dutta [2], Sharon Lim Pei Siang [2], David Ho Soon Wee [2], Surya Bhattacharya [2], Zhengya Zhang [1]

[1] University of Michigan, Ann Arbor, MI, USA; [2] Institute of Microelectronics, A*STAR, Singapore

2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) | 978-1-6654-9772-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/VLSITechnologyandCir46769.2022.9830249

Abstract

NetFlex is a multi-chiplet package (MCP) for CNN-based perception acceleration. With a balanced parallelism for mapping and a flexible scheduling, the NetFlex chiplet supports convolution, deconvolution and fully connected layers of different shapes, sizes and strides at high utilization. NetFlex adopts depth-first stream processing and an efficient streaming interface in a multi-chiplet daisy chain over Advanced Interface Bus. A 22nm NetFlex chiplet was fabricated and measured to achieve 2.14TOPS/W (16b OP) at a nominal voltage of 0.89V and 492.3MHz. A four-chiplet NetFlex MCP was built in a high-density fan-out wafer-level packaging to demonstrate 428FPS for depth estimation and 7723FPS for pose estimation.

Introduction

Neural network (NN) model size and complexity growths are outpacing NN chip upgrades. Making monolithic chips to keep up with the model evolution is costly and challenging. Instead, modular chiplets can be designed and reused to construct a variety of multi-chip packages (MCP) to address different NN models and tasks, as demonstrated by Nvidia's DNN MCP [1]. The future success of the chiplet approach depends on further developments of chiplets that can be efficiently reused at high utilization, standard and high-bandwidth chiplet interfaces, and high-density packaging.

We apply the chiplet approach to the design of a CNN-based perception accelerator (Fig. 1). A 22nm chiplet named NetFlex is designed to efficiently support convolution (conv), deconvolution (deconv) and fully connected (FC) layers. The chiplet adopts the open and standard Advanced Interface Bus (AIB) interface [2], [3], providing lightweight, AXI-compatible streaming at up to 640Gb/s. Four NetFlex chiplets are integrated in an MCP using high-density fan-out wafer level packaging (HD-FOWLP) [4]. The flexible chiplet, the lightweight standard interface, the low-latency streaming, and the accessible HD-FOWLP substrate allow the NetFlex design to be scaled up to a larger size and scaled out to support other NN applications.

Chiplet Design for Flexibility, Utilization and Efficiency

The NetFlex chiplet includes an INT16 NN core, a dataflow control, an AIB interface, a digitally controlled oscillator for clock generation, and an UART interface for initialization (Fig. 2). The base of the NN core follows [5], and it is parallelized in a balanced way: 8× along XY (mapped to 8 PEs in a row), 16× along C (mapped to 16 PE rows in a sheet), and 8× along K (mapped to 8 PE sheets). In computing conv, the input activations (IA) are broadcast to the 8 PE sheets, and each PE sheet caches weights of an output channel (K). Within a PE sheet, the weights of 16 input channels (C) are each sent to a PE row. A PE row computes 8 MACs every cycle and the inputs undergo X-shifts followed by Y-shifts to compute 2D conv (Fig. 3a). The partial sums are collected along columns of a PE sheet for reduction along C.

The NetFlex chiplet is designed for reuse, so the base NN core is extended to support different layer shapes, conv sizes, strides, deconv and FC with high utilization and efficiency. The core's balanced parallelism provides a well-sized IA unit block of $X \times Y \times C = 8 \times 1 \times 16$ and a weight unit block of $R \times S \times C \times K = 1 \times 1 \times 16 \times 8$ for dividing most common NN layers to obtain a high mapping utilization. Temporal scheduling is adopted for supporting different conv sizes while maintaining a high utilization. Strides larger than 1 are supported by PE gating. The high utilization (Fig. 4a, b) leads to a lower latency. In computing deconv, all-zero rows in IA are removed by skipping the associated temporal processing steps, and element-wise zeros in the remaining rows are squeezed out and PEs are gated (Fig. 3b) for efficiency.

Low-Latency Stream Processing and Chiplet Integration

As the chiplet size and memory are limited, NetFlex adopts depth-first processing (Fig. 5a) in conjunction with streaming between chiplets to timely consume activations to reduce activation memory and latency. The NetFlex chiplet employs line buffers [6], and as soon as a minimum number of lines of IA (e.g., 3 rows for 3×3 conv) are available, the processing can kickstart. To enable seamless streaming, the orders of output activations (OA) production and IA consumption are aligned without costly data rearrangement: a chiplet produces OA along output channel (K) first, then XY; and the next chiplet consumes IA along input channel (C) first, then XY (Fig. 5b). Compared to layer-by-layer, depth-first reduces the latency by over 2.7× (Fig. 4c).

Four NetFlex chiplets are connected in a daisy chain (Fig. 6a) in a prototype to enable stream processing via AIB. Absent of routers, a daisy chain is a lightweight alternative to a mesh [1]. By linking only pairs of chiplets, the wires are kept short, allowing a high I/O bandwidth and energy efficiency. An NetFlex chiplet uses 8 AIB channels (Fig. 6b) over 55µm-pitch µbumps, each channel providing up to 80Gb/s with a 4ns transfer latency. The NetFlex MCP is built on a 5-layer HD-FOWLP (Fig. 7a), where two layers are used for AIB routing with a 2µm width and a 2µm minimum spacing, and two layers for power delivery [4]. The wire lengths are equalized within a channel and kept to 4.4-5.8mm to meet the skew and frequency requirement. On top of AIB, we define an AXI-compatible bus interface to handle packing/unpacking of data to/from the AIB channels and provide a burst mode to efficiently utilize the bandwidth for streaming. Each AIB channel can be flexibly configured as leader or follower to adjust the Tx/Rx bandwidth. Additional modes including forwarding via a relay chiplet and bypassing a chiplet are added for flexibility.

Chip and MCP Measurement Results

The NetFlex chiplet was fabricated in an Intel 22nm FinFET Low Power technology (Fig. 7b). The processing part occupies 7.8mm^2 and the AIB I/Os take 3.3mm^2. The chiplet is measured to consume 499.8mW with a supply of 0.89V and a clock frequency of 492.3MHz in room temperature. The measurements translate to 2.14TOPS/W (16b OP) and 0.14TOPS/mm^2. A peak efficiency of 7.19TOPS/W (16b OP) is measured at 0.6V at 190.3MHz (Fig. 7c). The NetFlex MCP measures 13.5mm×13.5mm. The HD-FOWLP MCP is molded to form a BGA package, and assembled on PCB (Fig. 7d).

An end-to-end CNN-based perception model, SFM Learner [7], is mapped to the NetFlex MCP. The MCP provides 428FPS for depth estimation and 7723FPS for pose estimation using the KITTI dataset (416×128 frame size as in [7]). Compared to recent visual SLAM and DNN-SLAM approaches [5], [8] (Table I), NetFlex provides better adaptation and scalability, but the CNN-based approach requires more computation and storage. Compared to the state-of-the-art multi-chiplet [1] and multi-chip [9] NN accelerators (Table II), NetFlex provides a competitive efficiency due to the optimized utilization and the efficient streaming. NetFlex's standard interface and HD-FOWLP substrate will contribute to accessible chiplet and MCP designs.

Acknowledgements

The UM effort is supported in part by DARPA CHIPS and ONR under grant N00014-17-1-2992 and DARPA HI3. We thank Intel for the chiplet fabrication and bumping, S.-G. Cho, F. Sheikh, M. Flanigan, D. Kehlet, and S. Shumarayev from Intel for advice and assistance.

References

[1] B. Zimmer, et al., VLSI 2019.
[2] D. Greenhill, et al., ISSCC 2021.
[3] C. Liu, et al., CICC 2021.
[4] M. Rotaru, et al., ECTC 2021.
[5] Z. Li, et al., ISSCC 2019.
[6] K. Goetschalckx, VLSI 2021.
[7] T. Zhou, et al., CVPR 2017.
[8] A. Suleiman, et al., VLSI 2018.
[9] M. Giordano, et al., VLSI 2021.

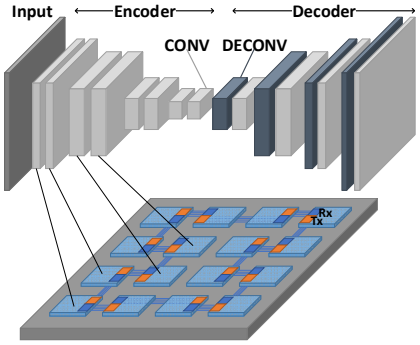

Fig. 1. Illustration of NN mapping on multi-chiplet system.

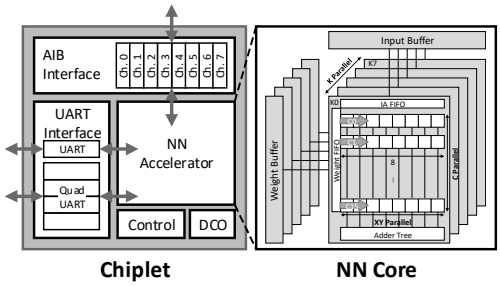

Fig. 2. The NetFlex chiplet block diagram and the NN core design.

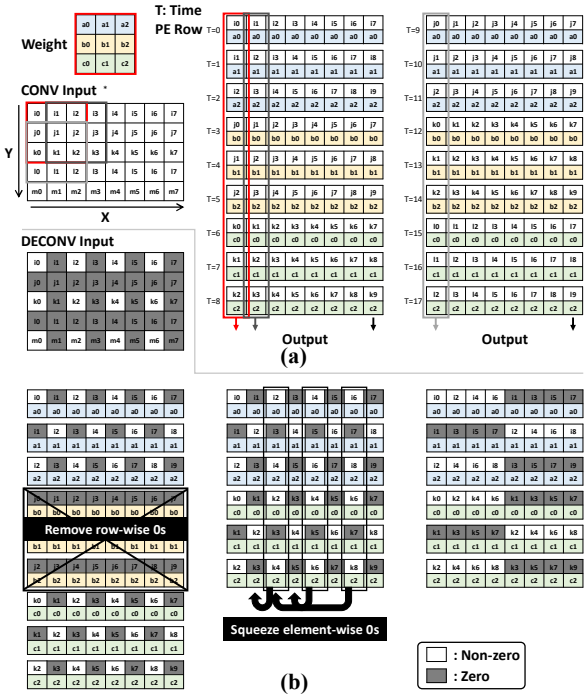

Fig. 3. (a) A 3×3 convolution operation example with stride of 1. (b) A 3×3 deconvolution example.

TABLE I: COMPARISON OF NETFLEX CHIPLET WITH PRIOR PERCEPTION ACCELERATORS

	VLSI 2018 [8]	ISSCC 2019 [5]	This Work
Type	Harris feature and visual SLAM	CNN feature and visual SLAM	End-to-end CNN
Technology	65nm	28nm	22nm
Area	20mm^2	10.92mm^2	7.8mm^2
Memory Size	854kB	1126kB	2492kB
Voltage	1V	0.63-0.9V	0.6-0.89V
Frequency	62.5MHz 83.3MHz	90-215MHz	190.3-492.3MHz
Core Power	24mW [a]	61.75-243.6mW	57.6-499.8mW
Performance (TOPS)	0.011-0.059 (INT) 0.001-0.006 (FP64) [b]	0.329-0.879 (INT8, INT32) [c]	0.41-1.07 (INT16)
Energy Efficiency (TOPS/W)	0.42-2.46 (INT) 0.042-0.25 (FP64) [b]	3.6-5.34 (INT8, INT32) [c]	2.14-7.19 (INT16)
Dataset	EuRoC	KITTI	KITTI
Image Size	752x480	640x480	416x128 [d]
Throughput	90FPS	80FPS	Depth net: 108FPS (1 chiplet) Pose net: 2001FPS (1 chiplet)

a: IMU power excluded; b: VFE precision: N/A, BE: double precision; c: CNN precision 8b and BA precision 32b; d: Image crop used by [7]

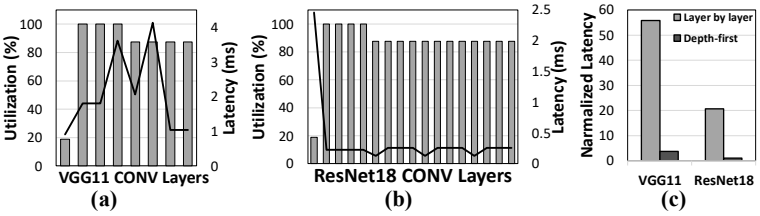

Fig. 4. Utilization and latency of (a) VGG11 and (b) ResNet18 CONV layers. (c) Network latency reduction of layer by layer and depth-first approach.

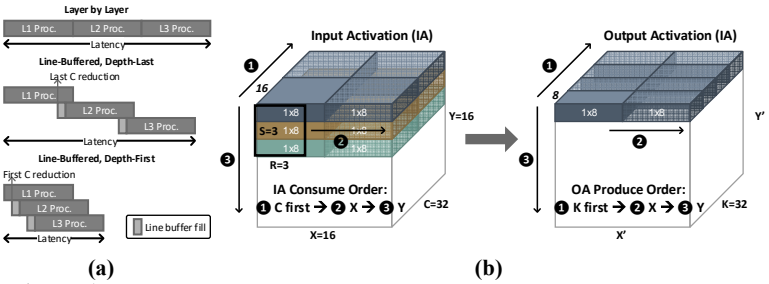

Fig. 5. Line Buffered depth-first (a) pipeline and (b) process scheduling.

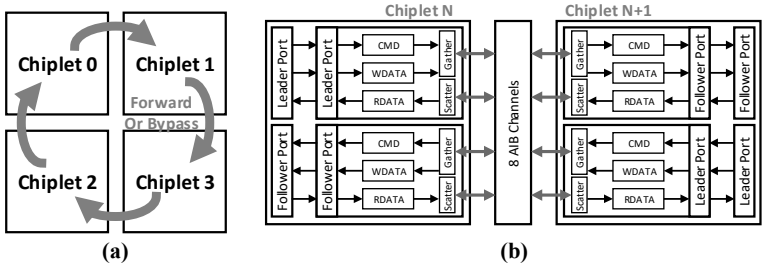

Fig. 6. (a) The daisy chain topology of chiplets with forward and bypass mode. (b) AIB adaptor with reconfigurable channels for leader or follower mode.

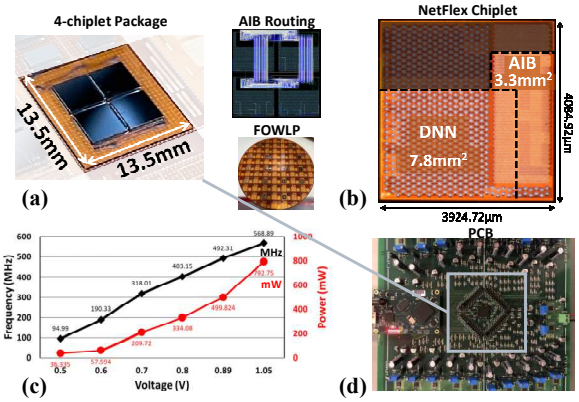

Fig. 7. (a) 4-chiplet package photo, FOWLP photo, and AIB routing. (b) Die photo of NetFlex chiplet. (c) Frequency and power consumption with voltage scaling. (d) PCB photo of the system,

TABLE II: COMPARISON WITH PRIOR MCP WORKS

	VLSI 2019 [1]	VLSI 2021 [9]	This Work
Technology	16nm	40nm	22nm
Area	6mm^2 Core: 3.1mm^2	29.2mm^2	11.1mm^2 Core: 7.8mm^2
Memory Size	752KB SRAM	0.5MB SRAM 2MB RRAM	2492kB SRAM
Voltage	0.42-1.2V	1.1V	0.6-0.89V
Frequency	161-2001MHz	200MHz	190.3-492.3MHz
Power	30-4160mW	126mW	Core: 57.6-499.8mW
Performance (TOPS)	0.32-4.01 (INT8)	0.92 (INT8, FP16)	0.41-1.07 (INT16)
Energy Efficiency (TOPS/W)	0.96-9.5 (INT8)	2.2 (INT8, FP16)	2.14-7.19 (INT16) 1.75-2.44 (INT16) w/ AIB [a]
Package	Organic	PCB	HD-FOWLP
I/O	GRS	C2C links	AIB
I/O Energy	0.82-1.75pJ/b	77pJ/b	3.07 pJ/b (I/O and bus interface) [a]

a: Measured at 713.5MHz DDR for I/O and 356.7MHz for bus interface