A Scalable Multi-Chiplet Deep Learning Accelerator with Hub-Side 2.5D Heterogeneous Integration

Zhanhong Tan¹, Yifu Wu², Yannian Zhang², Haobing Shi², Wuke Zhang², Kaisheng Ma¹

¹Tsinghua University,
²Polar Bear Tech
Abstract

With the slowdown of Moore’s law, the scenario diversity of specialized computing, and the rapid development of application algorithms, an efficient chip design requires modularization, flexibility, and scalability. In this study, we propose a Chiplet-based deep learning accelerator prototype that contains one HUB Chiplet and six extended SIDE Chiplets integrated on an RDL layer for the 2.5D package. The SIDE and the HUB contain one and four AI cores, respectively.

Given that our Chiplet-system targets diverse scenarios via scalable connected SIDE Chiplets, we need to handle three challenges: a) devise a flexible architecture design supporting diverse shapes, b) search for a workload mapping with low die-to-die communication, and c) adopt a high-bandwidth die-to-die interface to maintain efficient data transfer.

This study proposes a flexible neural core (FNC) featuring dynamic bit-width computing and flexible parallelism. Next, we use a hierarchy-based mapping scheme to decouple different parallelism levels and help analyze the communication. A 12Gbps D2D interface is introduced to achieve 192Gb/s bandwidth per D2D port with 1.04pJ/bit efficiency and 55μm bump pitch.

The proposed seven-Chiplet accelerator achieves a peak performance of 10/20/40 TOPS for INT16/8/4. When enabling 0~6 SIDE Chiplets, the system power ranges from 4.5W to 12W. The power efficiency of the FNC is 2.02TOPS/W while that of the overall system is 1.67TOPS/W.
Decouple a monolithic SoC into Chiplets
- Better die yield
- Scalability for diverse scenarios
  \[ M \times \text{Chiplet}_1 + N \times \text{Chiplet}_2 + K \times \text{Chiplet}_3 \]
- Rapid development pace to deliver new products
Decouple a monolithic SoC into Chiplets
- Better die yield
- Scalability for diverse scenarios
- Rapid development pace to deliver new products

Challenges

**Challenge-1**
Flexible architecture design supporting diverse shapes

**Scenario 1: Mobile**
Low-Res, small filter

**Scenario 2: Surveillance**
High-Res, large batch

**Scenario 3: Automotive**
High-Res, large channel

**Challenge-2**
Efficient workload mapping to optimize Die-to-Die communication

**Challenge-3**
High-bandwidth D2D and high-density package
## Flexible Neural Core (FNC)
- Reconfigurable architecture for the shape diversity

## Mapping dataflow
- Die-to-Die communication-aware workload generator

## Interconnection
- High-bandwidth Die-to-Die based on 2.5D package
- Efficient chiplet routing unit (CLRU)
Flexible Neural Core

- **Flexible Interconnect**
  - Arbitrary tile-based workload assignment to 8 cores via a configurable interconnect fabric

- **MAC array with Dynamic Bit-width**
  - Support INT4/8/16 with full bandwidth utilization of a tensor core

- **Configurable Weight Buffer**
  - Configured for NUMA/UMA mode for arbitrary workload assignment
Flexible Neural Core

The MAC Pair Supporting for Dynamic Bit-width

- Support three quantization modes
  - 8b-Activation × 4b-Weight
  - 8b-Activation × 8b-Weight
  - 16b-Activation × 8b-Weight

- Each INT-8 MAC-Pair has eight 4×4 multipliers for mode reuse

- In three modes, the bandwidth and compute resources of one MAC-pair are fully utilized
Flexible Neural Core

- **Flex-Interconnect and Configurable Weight Buffer**
  - Support diverse eight-PE compositions
    - 8-tile mode: share weights across 8 PEs for independent output in height/width
    - 4-tile mode: share weights across 4 PEs and 2 4-PE groups process 2 chunks of output channels
    - 2-tile mode: 4 2-PE groups for 4 chunks of output channels
    - 1-tile mode: 8 PEs for 8 chunks of output channels

Each bank has 16 sub-banks for 16 columns of MACs
Dynamic Workload Parallelism

- **Critical Position for the “X” buffer:** the inner-most loop related to the index of the X-buffer data (decide the data size on-core)

- **Reuse Region for the “X” buffer:** indicate the reuse efficiency when caching the data in inner loops decided by critical position

- **Search for an optimized loop range with the highest memory utilization (the largest data size that can be buffered on-core) and reuse efficiency for each buffer**
### Chiplet Interconnection and Package

#### High-Bandwidth D2D Interface

- **Bandwidth per D2D**: RX: 192Gb/s, TX: 192Gb/s
- **Data width per lane**: 8bit
- **Data Rate**: 12Gbps
- **Bump Pitch**: 55μm
- **Power**: 1.04pJ/bit
- **Area**: 2.2 × 0.5mm

#### Chiplet Router Unit (CLRU)

- **Config Reg**
- **AXI-M**
- **AXI-S**
- **4K Boundary Processing**
- **Data Parser**

- **Four FIFO queues to deal with burst transfer**
- **Data parser: support the data request from another Chiplet (access memory / other CLRU)**

> The head packages indicate the transfer mode
Chiplet Interconnection and Package

- **Non-conflict IO layout** in the HUB Chiplet to improve the fan-out efficiency
- 2.5D integration with a **high-density 65nm RDL** layer providing 55μm bump pitch
- The RDL layer contributes to a simpler 8-layer substrate of 3-2-3
A Scalable Multi-Chiplet Deep Learning Accelerator with Hub-Side 2.5D Heterogeneous Integration

Software Stack

Pre-Integrated Solutions
- Model Zoo
- NN-Profiler
- NN APIs

Rule Checker

Application Enablement

Supported Frameworks
- PyTorch
- TensorFlow
- PaddlePaddle
- ONNX

Model Optimization (quantization, pruning, op-tuning)

Full-Stack

Graph Optimization
- IR Gen.

Scheduler
- NN Lib.
- Code Gen.

Dev-AIDS
- Acc. Analyzer
- Func/Cyc-Simulator
- Testbench
- Design Explorer

Chip Enablement

Latency-Constraint Optimization
- Core-level task scheduling

Sample-per-Sec Optimization
- Batch-level Parallelism

Query-Per-Sec Optimization
- Workload pipeline for high throughput

Response Time Optimization
- Multi-level parallelism for high utilization
### Evaluation

#### Evaluations on one Flexible Neural Core

<table>
<thead>
<tr>
<th>Model</th>
<th>Runtime (ms)</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLabv3+</td>
<td>59.6ms</td>
<td>16.8fps</td>
</tr>
<tr>
<td>SqueezeNet</td>
<td>1.5ms</td>
<td>666.7fps</td>
</tr>
<tr>
<td>ResNet101</td>
<td>15.1ms</td>
<td>66.2fps</td>
</tr>
<tr>
<td>ResNet50</td>
<td>8.7ms</td>
<td>114.9fps</td>
</tr>
<tr>
<td>YOLOv3</td>
<td>119.3ms</td>
<td>8.4fps</td>
</tr>
<tr>
<td>YOLOv5s</td>
<td>13.1ms</td>
<td>76.3fps</td>
</tr>
<tr>
<td>YOLOv5n</td>
<td>6.2ms</td>
<td>161.3fps</td>
</tr>
<tr>
<td>DrivableNet</td>
<td>89.7ms</td>
<td>11.1fps</td>
</tr>
</tbody>
</table>

#### Evaluation on computing-bound workload

![Graph showing speedup by 9.95 with 10 cores]

#### Evaluation on IO-bound workload

![Graph showing speedup by 6.25 with 10 cores]
System Board and Demo

- System PCIe-based Board
  - 2GB GDDR6
  - AI Chip
  - Host Communication for offloading workloads
  - For the cooling module

- Demo for running concurrent 4 models
  - Model 1: VGG-16
  - Model 2: YOLO v5s
  - Model 3: ResNet50
  - Model 4: DeepLab v3+
### Chip Summary

<table>
<thead>
<tr>
<th>Items</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>CMOS 12nm</td>
</tr>
<tr>
<td><strong>Die Area</strong></td>
<td></td>
</tr>
<tr>
<td>HUB Chiplet</td>
<td>8.5mm × 6.8mm = 57.8mm²</td>
</tr>
<tr>
<td>SIDE Chiplet</td>
<td>3.5mm × 2.8mm = 9.8mm²</td>
</tr>
<tr>
<td><strong>Supply Voltage</strong></td>
<td>0.8V ~ 1.2V</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td>100MHz – 1GHz</td>
</tr>
<tr>
<td><strong>Peak Performance</strong></td>
<td></td>
</tr>
<tr>
<td>INT4</td>
<td>40TOPS (8b-A × 4b-W)</td>
</tr>
<tr>
<td>INT8</td>
<td>20TOPS (8b-A × 8b-W)</td>
</tr>
<tr>
<td>INT16</td>
<td>10TOPS (16b-A × 8b-W)</td>
</tr>
<tr>
<td><strong>NPU Core Efficiency</strong></td>
<td>2.02TOPS/W</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>4.5W ~ 12W</td>
</tr>
<tr>
<td><strong>D2D Bandwidth</strong></td>
<td>6×24GB/s for TX/RX</td>
</tr>
<tr>
<td><strong>External Memory Bandwidth</strong></td>
<td>64GB/s (GDDR6)</td>
</tr>
<tr>
<td><strong>Bump Pitch for 2.5D Pkg</strong></td>
<td>55μm</td>
</tr>
</tbody>
</table>
Comparison with Prior Multi-Chiplet Accelerator Works

<table>
<thead>
<tr>
<th></th>
<th>Simba (NVIDIA)</th>
<th>CHIMERA (Stanford)</th>
<th>NetFlex (A*STAR)</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td>2019</td>
<td>2021</td>
<td>2022</td>
<td>2023</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td>16nm</td>
<td>40nm</td>
<td>22nm</td>
<td>12nm</td>
</tr>
<tr>
<td><strong>Area</strong></td>
<td>6mm²</td>
<td>29.2mm²</td>
<td>11.1mm²</td>
<td>HUB: 57.8mm² Side: 9.8mm²</td>
</tr>
<tr>
<td><strong>Memory Size</strong></td>
<td>752KB SRAM</td>
<td>0.5MB SRAM 2MB RRAM</td>
<td>2492KB SRAM</td>
<td>HUB: 1.7MB Side: 439KB</td>
</tr>
<tr>
<td><strong>Voltage</strong></td>
<td>0.42V ~ 1.2V</td>
<td>1.1V</td>
<td>0.6V ~ 0.89V</td>
<td>0.8V ~ 1.2V</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td>161MHz – 2001MHz</td>
<td>200MHz</td>
<td>190.3 – 492.3MHz</td>
<td>600MHz – 1.2GHz</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>30 – 4160mW</td>
<td>126mW</td>
<td>57.6 – 499.8mW</td>
<td>Side: 0.72W Hub: 4.75W</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>0.32 – 4.01 (INT8)</td>
<td>2.2 (INT8, FP16)</td>
<td>0.41 – 1.07 (INT16)</td>
<td>Side Die: 1/2/4 for INT16/8/4, Hub Die: 4/8/16 for INT16/8/4, Total: 10/20/40 for INT16/8/4</td>
</tr>
<tr>
<td><strong>Package</strong></td>
<td>Organic MCM</td>
<td>PCB</td>
<td>HD-FOWLP</td>
<td>2.5D RDL</td>
</tr>
<tr>
<td><strong>D2D I/O</strong></td>
<td>GRS</td>
<td>C2C Links</td>
<td>AIB</td>
<td>12Gbps Parallel Interface</td>
</tr>
<tr>
<td><strong>I/O Energy</strong></td>
<td>0.82 – 1.75pJ/b</td>
<td>77pJ/b</td>
<td>3.07pJ/b</td>
<td>1.04pJ/bit</td>
</tr>
</tbody>
</table>
Thank You

tanzh@mails.tsinghua.edu.cn