A building block computing system for AI applications

Keio University
Hideharu Amano
Background

• Rapidly Increasing Requirements of Embedded Systems
  – Smart Phones, Tablets, Car electronics, Robotics, IoTs, etc.
  – High Performance, Low Energy Consumption, Low Cost....

• Rapidly Increasing NRE (Non-Recurring Engineering) cost of advanced LSIs.
  – Complicated masks
  – Design/Verification

• SoC (System-on-Chip) is difficult to be implemented for each target.

Building Block Computing Systems
Building Block Computing

1 step: Various Combination can be done after chip fabrication
2 step: Chips can be replaced by users → Field Stackable

Key Technique: Inductive Coupling Through Chip Interface (TCI)
Today’s talk

• What is TCI?
• The first prototype: Cube-1
• Intellectual Property of TCI
• Plug-and-play inter-chip network
• A heterogeneous AI system: Cube-2
### 3D IC technology for going vertical

<table>
<thead>
<tr>
<th>Two chips (face-to-face)</th>
<th>Wired</th>
<th>Wireless</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microbump</td>
<td></td>
<td>Flexibility</td>
</tr>
<tr>
<td>Through silicon via</td>
<td></td>
<td>Capacitive coupling</td>
</tr>
<tr>
<td>More than three chips</td>
<td></td>
<td>Inductive coupling</td>
</tr>
</tbody>
</table>

- **3D IC technology**
  - **Wired**
    - Microbump
  - **Wireless**
    - Flexibility
    - Capacitive coupling
    - Inductive coupling

- **Scalability**
Block diagram of scalable 3D NoC using inductive-coupling ThruChip Interface (TCI).

35bit Packet Structure

<table>
<thead>
<tr>
<th>35</th>
<th>33</th>
<th>32</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Header</td>
<td>Payload</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transfer Type:
- Command/Address
- Single Packet Data
- Burst Multi-Packet Data

CPU Core
Network Interface
Data Link
Clock Link
Host

Accelerator Core
Router
Packet En/Decode
Network Interface
3D Interconnect
TCI

Uplink
Downlink

Accelerator 1
Accelerator 2
Host CPU
Benefits of Inductive Coupling TCI

• More than 8Gbps speed is achieved within 10mW power dissipation.
• The error rate is less than $10^{-9}$.
  – No error correction circuits are required.
• Coils are built by metal wires of common CMOS.
  – No additional process is required.
• No ESD (Electro-Static Discharge) protection devices are required unlike TSV.
• Stacking cost is much lower than that for TSV.
  – Just grind chips to 40-80um thickness and stack them as in a common interposer inside the package.
• First commercial products with TCI will be shipped this year from PEZY.
Problems

• The footprint of a coil is much larger than a via for TSV.
  – Chip does not scale in the vertical direction.
  – Footprint of both coils and TSVs is increased in the advanced technology.
  – Digital circuits can be implemented in coils.
  \rightarrow The real footprint is not larger than that of TSV.

• Power and system clocks are supplied by bonding wires.
  – Now, the power supply by inductive coupling is tried.
  – It may take some time….

• Heat dissipation cannot be done by TCI.
  – The bonding agent used in the current prototype well transfers heat.
  – But heat dissipation more than 1W chips is a challenge.
Today’s talk

• What is TCI?
• The first prototype: Cube-1
• Intellectual Property of TCI
• Plug-and-play inter-chip network
• A heterogeneous AI system: Cube-2
The first prototype Cube-1
Built in 2013. To demonstrate that the TCI can be used in a real system.

The number of accelerators can be changed.

Ring based packet switching network is formed
Microphotograph of stacked test chips.

Host CPU

Accelerator 1

Accelerator 2

Accelerator 3

Host CPU + Accelerator x3 Chip Stack

Fabricated in 65nm CMOS

Host CPU Chip

Accelerator Chip

Microphotograph of stacked test chips.
# Cube-1 vs. Cube-2

<table>
<thead>
<tr>
<th></th>
<th>Cube-1</th>
<th>Cube-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>Fujitsu e-shuttle 65nm 12metal</td>
<td>Renesas SOTB 65nm 7metal</td>
</tr>
<tr>
<td>Area</td>
<td>2.1mm X 4.2mm</td>
<td>3mm X 6mm</td>
</tr>
<tr>
<td>TCI</td>
<td>3Gb/s 240um</td>
<td>2.5Gb/s 240um</td>
</tr>
<tr>
<td>CPU Cache TLB</td>
<td>MIPS R3000 4KW 2way sep. 16-entry shared</td>
<td>MIPS R3000 4KW 2way sep. 16-entry shared</td>
</tr>
<tr>
<td>CMA</td>
<td>CMA-Cube 8x8</td>
<td>CC-SOTB 12x8</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>0.6-1.2V</td>
<td>0.3-1.2V</td>
</tr>
<tr>
<td>Target Frequency</td>
<td>50-100MHz</td>
<td>50-100MHz</td>
</tr>
<tr>
<td>Chip Thickness</td>
<td>40-80um</td>
<td>80um</td>
</tr>
</tbody>
</table>
Evaluation Results from a 2-chip stack system

Execution Time (msec)

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Gray</th>
<th>Sepia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geyser Only</td>
<td>2-chip stack</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Energy (μJ)

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Gray</th>
<th>Sepia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geyser Only</td>
<td>2-chip stack</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Energy Reduction
Evaluation Results

- Using three accelerators.
- The target image block: 128 x 96 pixel
Measured PER and TCI power dissipation dependence on supply voltage.
Demonstration system.
Cube-1 was a good starting point

Please check Micro 2013.11/12

http://www.am.ics.keio.ac.jp/kaken_s

Demonstration in HotChips2014, Coolchips2015, and FPL2015.
The problems on Cube-1

• Not stable with three chips
  – Interference between vertical inductors which must be placed for the ring network.

• OK. It’s an interesting system. But what for?
  – Only two chip types were provided.

• For development of various types of chips:
  – Easy design.
  – Automatic performance tuning.
  – Thermal problem on stacking.
  – Practical power supply.

  Comprehensive Researches are needed.
What is the next step?

- **Intellectual Property of the TCI**
  - It must be embedded in the LSI design flow.
  - Higher level layers are needed.

- **A new plug-on network based on the IP**

- **Automatic performance tuning**
  - Software approach
  - Hardware approach

- **Thermal dissipation problems**

The new target systems Cube-2 for the AI application
Today’s talk

• What is TCI?
• The first prototype: Cube-1
• Intellectual Property of TCI
• Plug-and-play inter-chip network
• A heterogeneous AI system: Cube-2
TCI-IP Physical layer

Physical layer = Inductors + Tx/Rx + SERDES
TCI physical layer

Can be used for the both direction
Serial transfer synchronized with Txclk
Fully Depleted Silicon on Insulator (FD-SOI): SOTB MOSFET

- Dopant less Structure (low variability)
- Formed on an ultra thin BOX layer (about 10nm)

High efficiency of leakage reduction with body bias control
Body Bias Control of SOTB

\[ V_{BN} < V_S \) (Reverse body bias: RBB)
\[ V_{BP} > V_S \]
- Low leakage current
- Low speed

\[ V_{BN} > V_S \) (Forward body bias: FBB)
\[ V_{BP} < V_S \]
- High speed
- Large leakage current

It is important to decide optimal voltages of body bias

See [Okuhara IEEE TVLSI 2017]
TCI IP
80μm thickness
36bit/50MHz with normal bias.
Including SER/DES circuits.
Half-duplex data transfer.
CC SOTB Layout

Renesas Electronics 65nm SOTB 7Metal layers
6mm X 3mm

TCI IP × 4sets (Up/Down)
Three layers for the TCI

• Full duplex bi-directional channel is formed with two physical IPs.
• The flow control is embedded into the link controller.
• Various networks can be easily formed.

- **Router**
  - Verilog Soft core. Parameterized router. 3-stage pipeline.
- **Link Controller**
  - Verilog Soft core. 8-Virtual Channels. Flow control. 17flits packet buffer
- **TCI transceiver**
- **TCI receiver**
  - Physical layout. SERDES is embedded.

Open in the VDEC web site
Today’s talk

• What is TCI?
• The first prototype: Cube-1
• Intellectual Property of TCI
• Plug-and-play inter-chip network
• A heterogeneous AI system: Cube-2
The ring network on Cube-1

8x8 PE Array

Bonding wire

Network IF

μ-Controller
The escalator network

- Just by stacking chips with three layers IP, any numbers of chips can be stacked.
- The simplest 3-port routers are used.
- Vertical coils never interfere with each other.
The spacer chips

Drops

Spacer chips
Ring vs. Escalator

![Graph showing the comparison between Ring and Escalator in terms of average latency versus injection rate. The graph indicates that Ring outperforms Escalator at lower injection rates, while Escalator performs better at higher injection rates. The graph also distinguishes Escalator with and without Virtual Channels (VC).]
The overhead of flow-control
2 or 17-flit packets

→ The overhead can be ignored
Today’s talk

• What is TCI?
• The first prototype: Cube-1
• Intellectual Property of TCI
• Plug-and-play inter-chip network
• A heterogeneous AI system: Cube-2
A heterogeneous multicore system: Cube-2

Image recognition for robotics, cars or other edge systems.

- More than a single chip-stacking.
  - Embedded CPU: GeyserTT
  - Shared memory chip: SMTT

- Accelerator chips
  - CC-SOTB2
  - SNACC
  - KVS

- Various combination of chips.
## Cube-1 vs. Cube-2

<table>
<thead>
<tr>
<th></th>
<th>Cube-1</th>
<th>Cube-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>Fujitsu e-shuttle 65nm 12metal</td>
<td>Renesas SOTB 65nm 7metal</td>
</tr>
<tr>
<td>Area</td>
<td>2.1mm X 4.2mm</td>
<td>3mm X 6mm or 6mm X 6mm</td>
</tr>
<tr>
<td>TCI</td>
<td>3Gb/s 240um</td>
<td>2.5Gb/s 240um</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>0.6-1.2V</td>
<td>0.3-1.2V</td>
</tr>
<tr>
<td>Target Frequency</td>
<td>50-100MHz</td>
<td>50-100MHz</td>
</tr>
<tr>
<td>Chip Thickness</td>
<td>40-80um</td>
<td>80um</td>
</tr>
</tbody>
</table>
Geyser-TT (Twin Tower)

TCI for the twin tower

TCI for a single tower

MIPS R3000 Compatible CPU with three TCI IPs
Geyser-TT

Geyser CPU Core

- Basic CPU Core
- I Cache
- D Cache
- TLB

Geyser CPU Core (MIPS ISA)

Geyser Internal Bus

TCI for Single Tower
- TCI
- Router
- Router I/F

TCI for Twin Tower
- TCI
- Router
- Router I/F

DMAC

External Bus Controller

Geyser External Bus: Main Memory, I/Os
A simple stacking

An escalator network is formed.
Twin-tower with a single core

Geyser TT is used as a bridge of two chip stacks
Bonding wires plan
Shared Memory for twin-tower (SMTT)

- A large memory system is needed in various configuration.
  - A chip with DRAM module is needed.
  - From the limitation of money and power budget, only 250KB static RAM is provided.
  - Ultramemory’s DRAM with TCI will hopefully take place in the future.

- Multiple hosts in multiple towers can exchange the data with the shared memory.

- Synchronization memory is also provided.
SMTTT chip

TCI for the twin tower
Twin-tower with multi cores

GeyserTT
CC
SNACC

SNACC
KVS

Shared Memory
Bonding wires plan
Accelerator family

• CC-SOTB2
  – Coarse Grained Reconfigurable Array (CGRA)
  – Low power image processing

• SNACC
  – Convolutional Neural Network Accelerator

• KVS
  – Key-value store accelerator

• All chips have the same TCI IP.
CC-SOTB2

Variable Pipeline structure

PE array with 12x8 PE

12x8 PE array

Data manipulator for flexible load/store

12x2 interleaved/banked memory
Energy optimization with body bias control

- The previous presentation by Anh-Vu-san [McSoC2017]
- FPL2017 etc.
- More than 800MOPS/mW was achieved.
SNACC: a convolutional neural network accelerator

- SIMD cores
- Dedicated instruction set and ALUs
- Yesterday presentation by Sakamoto-san [McSoC2017]
KVS Chip

- Memory search
- Set/Get operations

[HotInterconnect 2016]
Streaming processing in Cube-2

• Shared memory using cache/DMA transfer
  – All memory modules of accelerators are mapped on the same address space.
  – Memory modules of accelerators are treated as the 2\textsuperscript{nd}-level cache.
  – DMA transfer between accelerators is triggered automatically.

• Streaming processing between chips can be efficiently done.
Execution of accelerators and communication between chips can be overlapped.
The current status

• 3-chip stack is now operational.
• Chips are available on this December.
Special thanks to:

- Prof. Kuroda (Keio Univ.)
- Prof. Kondo & Dr. Sakamoto (Univ. of Tokyo)
- Prof. Matsutani (Keio Univ.)
- Prof. Nakamura (Univ. of Tokyo)
- Prof. Namiki (Tokyo Agriculture & Tech.)
- Prof. Usami (Shibaura Inst. Tech.)

This work has been supported by JSPS Kakenhi S Grant # 25220002
Thank you!

Please visit our web-site:
  – http://www.am.ics.keio.ac.jp/kaken_s

For other projects:
  – http://www.am.ics.keio.ac.jp