Adaptive SoCs for Smart Autonomous Systems

Abderazek Ben Abdallah
University of Aizu
benab@u-aizu.ac.jp
http://www.u-aizu.ac.jp/~benab/
Aizu University

Japan
Major JR Lines
- Major rail line
- Shinkansen line
- Station with Japan Rail Pass exchange office

Aizu City / Aizu University
What is a System-on-Chip (SoC)?

- A system-on-chip (SoC) is an integrated circuit (IC) that integrates all components of a computer or other electronic system into a single chip.

- It may contain analog, digital, mixed-signal, and often radio-frequency functions.
What is an Adaptive SoC (ASoC)?

- **Compute** (High-performance, Real-Time)
- **Adapt** (Needs to run in dynamic environments, where physical context, network topologies, and workloads are always changing)
- **Learn** (No programming!)
Why now & why we should care?

1. Huge **progress** in IC technology
2. Emergence of IoT, **Embedded/Ubiquitous/Pervasive** applications.
3. Large bandwidth, low power requirements.
4. Data become more **knowledge intensive** (unstructured).
5. Current approaches do not efficiently deal with such **unstructured data**.
Where can We Use SoCs/ASoCs?

- Hundreds of cores integrated on chip (数百のコアがチップ上に集積)
  - Technology scaling (技術の進歩)
  - 3D integration (3D集積)
  - Many examples (多くの実例)
    - STMicro P2012/STHORM
    - picoChip
    - Tilera Tile GX, Tile Pro
    - Intel Polari
    - ...

- Complex apps, strict/strictive constraints (複雑なアプリケーション、厳しい制約)
  - Massively parallel applications (超並列アプリケーション)
    - Big-data Analysis, Pattern Recognition, Deep Learning
SoC Design Challenges

The performance of ICs is limited by the **Communication Network** rather than the **Computation Logics**.
Given the huge aggregate data rates within multi-core (>100TB/s in 2017 (ITRS)), continued scaling is bounded by the energy requirements of Inter-core communication to a maximum of 1pJ/bit/mm.
ASL Approach I

Scalable Packet - Switched Network on-Chip

Diagram:

- PE
- NI
- RX
- TX

December 19, 2016
benab@aizu.ac.jp
Why now & why we should care?

1. Huge **progress in IC technology**
2. Immergence of **IoT**, **Embedded/Ubiquitous/Pervasive applications.**
3. Large bandwidth, low power requirements
4. Data become more **knowledge intensive** (unstructured).
5. Current approaches do not efficiently deal with such **unstructured data.**
### Scalable Packet-Switched Network-on-Chip (NoC)

#### Packet

<table>
<thead>
<tr>
<th>Tail flit</th>
<th>Body flit</th>
<th>Head flit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ending flit</td>
<td>Carried Payload</td>
<td>Flit information</td>
</tr>
</tbody>
</table>

#### Router Organization

- **PE**: Processing Element
- **NI**: Network interface
- **R**: Router
- **RX**: Receive
- **TX**: Transmit
- **VC**: Virtual Channel
- **Crossbar**: 5x5 Crossbar

#### Multi-hop Communications


Every flit @ Every Router.

- **R**: Router
- **NI**: Network interface
- **PE**: Processing Element
Scalable Packet-Switched Network-on-Chip (NoC)

Router addressed NM (in decimal)

Wire length reduction

Lateral link (1mm ~ 4mm)
Vertical link (10 μm ~ 200 μm)

Footprint reduction

3D- Network-on-Chip architecture

a

b

a/2

b/2

c

c = die thickness (0.6 mm) + inter-die distance
Scalable Packet-Switched Network-on-Chip (NoC)

2-Dimensional NoC

Layer 1

Layer 2

Layer 3

Router addressed NM (in decimal)

Wire length reduction

2D

3D

Lateral link: (1mm ~ 4mm)
Vertical link: (10 μm ~ 200 μm)

3D- Network-on-Chip architecture

Footprint reduction

2D

3D

a

b

b/2

a/2

c = die thickness (0.6 mm) + inter-die distance

December 19, 2016

benab@u-aizu.ac.jp
Scalable Packet Switched Communication Network

**Diameter reduction**
The number of hops that a flit traverses in the longest possible minimal path between a (source, destination) pair.

**Packet energy reduction**

\[ E_{packet} = n \sum_{j=1}^{h} E_{hop,j} \]

\[ E_{hop} = E_{switch} + E_{wire} \]

Router addressed **NM** (in decimal)
Scalable Packet Switched Communication Network

Router addressed \textbf{NM} (in decimal)

\textbf{Diameter reduction}
The number of hops that a flit traverses in the longest possible minimal path between a (source, destination) pair.

\[ L_p = L_{\text{sender}} + L_{\text{transport}} + L_{\text{receiver}} \]

Tightly dependent on the latency overhead of each hop
The current out-port is read from the flit and the next-node address is computed.

2- The three possible directions are calculated: North, East, and Up.

3- Verify the link status of the three directions, and eliminate the faulty path: East.

4- Calculating the diversity value and select the highest one:
   North=3 (North, east, and up); Up=2 (North and east)
1- The current out-port is read from the flit and the next-node address is computed.
2- The three possible directions are calculated: North, East, and Up.
3- Verify the link status of the three directions, and eliminate the faulty path: East.
4- Calculating the diversity value and select the highest one: North=3 (North, east, and up); Up=2 (North and east).
3D-FETO: Fault Tolerant Highly-reliable 3D-NoC

Table 1: Simulation configurations.

<table>
<thead>
<tr>
<th>Parameter/System</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Size (z x y x x)</td>
<td></td>
</tr>
<tr>
<td>Matrix</td>
<td>3 x 6 x 5</td>
</tr>
<tr>
<td>Transpose</td>
<td>4 x 4 x 4</td>
</tr>
<tr>
<td>Uniform</td>
<td>4 x 4 x 4</td>
</tr>
<tr>
<td>Hotspot 10%</td>
<td>4 x 4 x 4</td>
</tr>
<tr>
<td>H.264</td>
<td>5 x 3 x 3</td>
</tr>
<tr>
<td>VOPD</td>
<td>2 x 2 x 3</td>
</tr>
<tr>
<td>MWD</td>
<td>3 x 2 x 2</td>
</tr>
<tr>
<td>PIP</td>
<td>2 x 2 x 2</td>
</tr>
<tr>
<td>Node's Delivered Packets per transmission session</td>
<td></td>
</tr>
<tr>
<td>Matrix</td>
<td>10</td>
</tr>
<tr>
<td>Transpose</td>
<td>10</td>
</tr>
<tr>
<td>Uniform</td>
<td>128</td>
</tr>
<tr>
<td>Hotspot 10%</td>
<td>128</td>
</tr>
<tr>
<td>Stream's Delivered Packets per transmission session</td>
<td></td>
</tr>
<tr>
<td>H.264</td>
<td>3,494</td>
</tr>
<tr>
<td>VOPD</td>
<td>1,120</td>
</tr>
<tr>
<td>MWD</td>
<td>512</td>
</tr>
<tr>
<td>PIP</td>
<td></td>
</tr>
<tr>
<td>Packet Size</td>
<td></td>
</tr>
<tr>
<td>Hotspot 10%</td>
<td>10 flits + 10%</td>
</tr>
<tr>
<td>Others</td>
<td>10 flits</td>
</tr>
<tr>
<td>Fills Size</td>
<td>44 bits</td>
</tr>
<tr>
<td>Header Size</td>
<td>14 bits</td>
</tr>
<tr>
<td>Payload Bit</td>
<td></td>
</tr>
<tr>
<td>Baseline, 3D-FETO</td>
<td>30 bits</td>
</tr>
<tr>
<td>Soft Error Tolerance, 3D-FETO</td>
<td>18 bits</td>
</tr>
<tr>
<td>Parity Bit</td>
<td></td>
</tr>
<tr>
<td>Baseline, 3D-FETO</td>
<td>0 bits</td>
</tr>
<tr>
<td>Soft Error Tolerance, 3D-FETO</td>
<td>12 bits</td>
</tr>
<tr>
<td>Buffer Depth</td>
<td>4</td>
</tr>
<tr>
<td>Switching</td>
<td>Wormhole-mode</td>
</tr>
<tr>
<td>Flow control</td>
<td>Stop &amp; Go</td>
</tr>
<tr>
<td>Routing</td>
<td>LAFT</td>
</tr>
</tbody>
</table>

Table 3: Successful arrival-rate comparison results for a 5x5x4 system configuration under Transpose traffic.

<table>
<thead>
<tr>
<th>Algorithm / Fault-rate</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYZ</td>
<td>83%</td>
<td>46%</td>
<td>31%</td>
<td>14%</td>
<td>11%</td>
</tr>
<tr>
<td>Hybrid-XYZ</td>
<td>99%</td>
<td>68%</td>
<td>42%</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>8-RW</td>
<td>93%</td>
<td>82%</td>
<td>62%</td>
<td>44%</td>
<td>36%</td>
</tr>
<tr>
<td>Odd-Even</td>
<td>97%</td>
<td>84%</td>
<td>53%</td>
<td>42%</td>
<td>32%</td>
</tr>
<tr>
<td>Hybrid-Odd-Even</td>
<td>99%</td>
<td>92%</td>
<td>77%</td>
<td>62%</td>
<td>53%</td>
</tr>
<tr>
<td>4N-FIRST</td>
<td>96%</td>
<td>86%</td>
<td>68%</td>
<td>50%</td>
<td>37%</td>
</tr>
<tr>
<td>4NP-FIRST</td>
<td>100%</td>
<td>97%</td>
<td>89%</td>
<td>75%</td>
<td>63%</td>
</tr>
<tr>
<td>LAFT-OASIS</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>98%</td>
</tr>
<tr>
<td>3D-FETO</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>98%</td>
</tr>
</tbody>
</table>

Figure 10: Average packet latency evaluation of the realistic benchmarks.
Chip Implementation

- Technology: 45nm CMOS Process
- Chip Size: 2.205X2.220 (micron)
- Frequency: 0.91 GHZ Confirmed
- Supply voltage: 1.1V
- Power Dissipation: 222.387 uW
- Number of Pins: 557

ANSMOM SoC Implementation

\[ R_y[L] = \sum_{n=0}^{N} y[n] \times y[n-L] \]

- \( R_y \): autocorrelation value
- \( y[n] \): filtered ECG signals
- \( N \): the number of times needed for calculation to get the period
  \[ 0 \leq L < N \]

Figure 2. BANSMOM system architecture.
ASL Approach II

Hybrid Electro-Photonic NoC
Why now & why we should care?

1. Huge progress in IC technology

2. Immergence of IoT, Embedded/Ubiquitous/ Pervasive applications.

3. **Large bandwidth, low power requirements**

4. Data become more knowledge intensive (unstructured).

5. Current approaches do not efficiently deal with such unstructured data.
The Concept

Replace wires with waveguides and electrons with photons!

(Photo: Spectrum 2005, Pinnacle)
Electronics is not Good @ High bit/s Communication.
Hybrid Photo-Electronic NoC

On-chip Electronic Communication
- Buffer, receive and re-transmit at every switch
- Off chip is pin-limited
- Large power/energy

On-Chip Photonic Communication
- Modulate/receive ultra-high bandwidth data stream once per communication.
- Switch routes entire multi-wavelength
- Low power switch fabric, scalable

(1) Reserve the Photonic Path
(2) ACK
(3) Transmit Data on the Photonic Medium
(4) Release the Path

December 19, 2016
benab@u-aizu.ac.jp
(a) 3x3 Mesh-based System, (b) 5x5 non-blocking photonic switch, (c) Unified Title.
Hybrid Photo-Electronic NoC

2D, 3D System Configuration.
Wavelengths are modulated and detected properly.

Rings’ resonance shifted by $\Delta \lambda$ due to temperature variation.

$\mu R_1$ is permanently Off due to process variation.

Photonic Link affected by process and temperature variations.
Model of a 5 Port FTTDOR Switch and a Wavelength Shifting Controller

$$\text{Loss}_{\text{router}} = \text{Loss}_{\text{bend}} \times N_{\text{bend}} + \text{Loss}_{\text{cross}} \times N_{\text{cross}} + \text{Loss}_{\text{MRoff}} \times N_{\text{MRoff}} + \text{Loss}_{\text{MRon}} \times N_{\text{MRon}}$$

$$\text{Delay}_{\text{router}} = \text{Delay}_{\text{bend}} \times N_{\text{bend}} + \text{Delay}_{\text{cross}} \times N_{\text{cross}} + \text{Delay}_{\text{MRoff}} \times N_{\text{MRoff}} + \text{Delay}_{\text{MRon}} \times N_{\text{MRon}}$$

PHENIC’s electronic controller layout in 45 nm process
Fig. 10 Latency results of each system as faults are introduced. Meyer, CANDAR’16
ASL Approach III

Adaptive Neurochip for Spiking Neural Networks

Mimicking the Structure of the Human Brain.
Why now & why we should care?

1. Huge progress in IC technology
2. Immergence of IoT, Embedded/Ubiquitous/Pervasive applications.
3. Large bandwidth, low power requirements
4. Data become more **knowledge intensive** (unstructured).
5. Current approaches do not efficiently deal with such **unstructured data**.
A Brief History of Neural Networks

1954
- F-Forward (FFNN)

1982
- Hopfield (HN)

1986
- Boltzmann, Restricted Boltzmann (RBM)

1988
- Radial Bas. Func, Autoencoder (AE)

1990
- Recurrent Neural (RNN)

2004
- Echo State Network (ESN)

2007
- Deep Belief Net (DBN), SAE

2010
- Deconvolutional Network (DN)

2013
- Markov Chain (MC), VAE, ELM

2014
- Generative adversarial net (GAN, GRU, NTM)

2015
- Deep Residual Net (DRN), DCIGN

2015
- Deep Convol. Inverse graphics Net (DCIGN)

Ex.
I: red flower
O: red flower image

References:
[Rosenblatt, 1958]
[Hopfield, 1982]
[Hinton, 1986]
[Broomhead, 1988]
[Hochreiter, 1997]
[Elman, 1990]
Brain Features

- Ten billion \((10^{10})\) neurons
- Neuron switching time \(>10^{-3}\) secs
- On average, each neuron has several thousand connections
- Hundreds of operations per second
- Face Recognition \(~0.1\) secs
- High degree of parallel computation, Distributed representations
- Each neuron is connected to the others through 10000 synapses
- It can learn, reorganize itself
Computer Vs. Brain

• **Computer**
  – Calculation
  – Precision
  – Logic

• **Brain**
  – Pattern Recognition
  – Noise Tolerance
  – Complexity
Computer Vs. Brain

- High Power
- Storage and computation are separated
- Poor at recognition

- Low Power
- Storage and computation are not separated
- Good at recognition
Neuron Organization & Modelling

\[ Y = f \left( \sum_{i \in n} X_i \cdot W_i - \theta \right) \]
Neuro-inspired Adaptive SoC based on Scalable Packet Switched NoC

Structure of one PE

Mimicking the bio-neuron

Decoder LUT
Encoder
Controller

bio-Neuron Structure
Neuro-inspired Adaptive SoC based on Scalable Packet Switched NoC

Structure of one PE

Processing Stages in one PE
Neuro-inspired Adaptive SoC based on Scalable Packet Switched NoC

Neuro-Packet

H P P P P P P P P

O₁ O₂ O₃ O₄ O₅ O₆ O₇ O₈

LUT
Neuro-inspired Adaptive SoC based on Scalable Packet Switched NoC

FF-ANN NoC with 8L/8N

Router  Input layer  Neuro Processor  Hidden layers  Output layer
Neuro-inspired Adaptive SoC based on Scalable Packet Switched NoC

Pattern Recognition
(5-7-3-7-5 neurons Mapping)

NPCA Image Processing
(4-10-1-10-4 neurons Mapping)
Demo 1
ASL Adaptive SoCs

2006
OASIS-1 – Scalable Packet-Switched Network-on-Chip
JASSSTo6, MCSOC12, JPDC14, SUP14

2013
OASIS-2 - Fault-Tolerant Network-on-Chip
MCSOC14, JPDC14, SUP16

2014
BANSMOM - Bio-Chip for Elderly Monitoring
ES2016, ACHRAF-MS1, KIMEZAWA-MS

2015
PHENIC- High-bandwidth Photonic NoC
SUP16, MCSOC15, CANDAR16,

2016
ANSPINN- Adaptive Neuro-inspired Processor and Platform
TR-OASIS-NP-042015
References

http://adaptive.u-aizu.ac.jp/
Welcome

The Adaptive Systems Laboratory (ASL) is affiliated with the Division of Computer Engineering, School of Computer Science and Engineering, The University of Aizu. ASL members include 2 faculty members, 3 Doctor, 3 M.S, and 13 B.S students. ASL has as its mission to develop research on adaptive computing system technologies. In particular, we develop adaptive systems for incremental learning and adaptation in dynamic and harsh environments.

What's New?

- Oct 8-9, 2016: Open Campus (Autumn session)
- Schedule for Matters regarding Graduation Thesis
- August 10, 2016: Schedules for Presentation Meetings of Research Plan Progress Seminar in the 2nd quarter, AY2016.
- June 27-29, 2016: 2nd year Doctor candidate Khaih N. Dang presented part of his research work at the IEEE Int. Conf. on Integrated Circuit Design and Technology (ICICDT 2016), June 27-29, 2016.
Thank you!

Abderazek Ben Abdallah
benab@u-aizu.ac.jp