DATE 25
DESIGN, AUTOMATION AND TEST IN EUROPE
THE EUROPEAN EVENT FOR ELECTRONIC SYSTEM DESIGN & TEST
31 MARCH – 2 APRIL 2025
LYON, FRANCE
CENTRE DE CONGRÈS DE LYON

Multi-Mode Borderguard Controllers for Efficient On-Chip Communication in Heterogeneous Digital/Analog Neural Processing Units

Hong Pang[*], Carmine Cappetta[†], Riccardo Massa[‡], Athanasios Vasilopoulos[§], Elena Ferro[*§], Gamze Islamoglu[*], Angelo Garofalo[*¶], Francesco Conti[¶], Luca Benini[*¶], Irem Boybat[§], Thomas Boesch[◊]

[*]*ETH Zurich, Switzerland* [†]*STMicroelectronics, Cornaredo, Italy* [‡]*STMicroelectronics, Agrate, Italy*
[§]*IBM Research - Zurich, Switzerland* [¶]*University of Bologna, Italy* [◊]*STMicroelectronics, Geneva, Switzerland*

Funded by the European Union
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
UKRI Innovate UK
neurSoC
ETH zürich
life.augmented
PULP Parallel Ultra Low Power
IBM
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

A powerful solution for edge AI…

Performs MAC (dot product) operation with high energy efficiency (**> 90%** of overall operations in AI inference)

AIMC

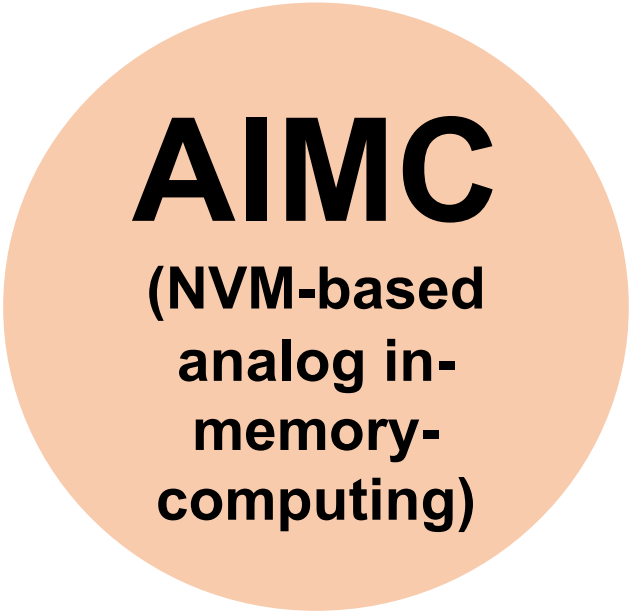
(NVM-based analog in-memory-computing)

Non-volatile memory-based AIMC is particularly attractive for:

- high on-chip weight storage capacity
- high power efficiency

[1] M. Le Gallo et al., “A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference,” Nature Electronics, pp. 1–14, 2023.

We also need other digital units…



Performs MAC (dot product) operation with high energy efficiency (**> 90%** of overall operations in AI inference)

[2] S. Jain et al., ”A Heterogeneous and Programmable Compute-In-Memory Accelerator Architecture for Analog-AI Using Dense 2-D Mesh,” in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 31, no. 1, pp. 114-127, Jan. 2023.

We also need other digital units…

General-Purpose Core
(Accuracy critical OP, activation function not supported by DPU)

CPU

Digital Processing Unit
(Convolutions, pooling, arithmetic, and activation functions)

DPU

Local Storage Unit

SRAM

AIMC
(NVM-based analog in-memory-computing)

Performs MAC (dot product) operation with high energy efficiency
(**> 90%** of overall operations in AI inference)

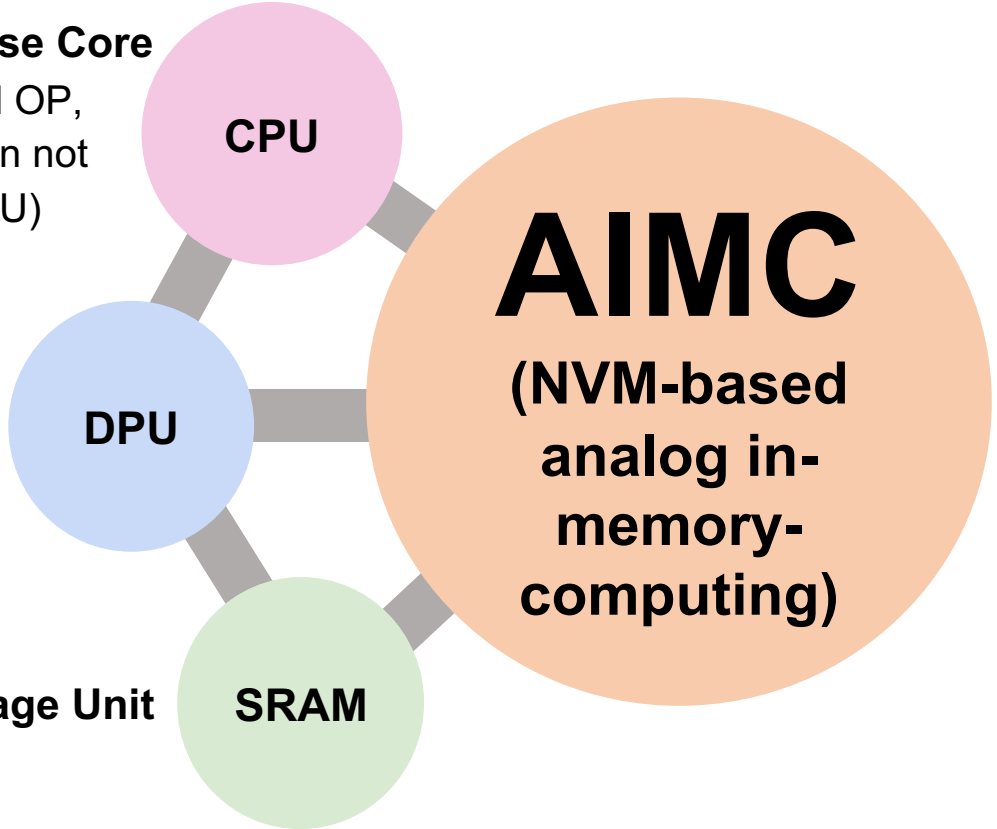
[2] S. Jain et al., "A Heterogeneous and Programmable Compute-In-Memory Accelerator Architecture for Analog-AI Using Dense 2-D Mesh," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 31, no. 1, pp. 114-127, Jan. 2023.

We also need other digital units…

General-Purpose Core
(Accuracy critical OP, activation function not supported by DPU)

Digital Processing Unit
(Convolutions, pooling, arithmetic, and activation functions)

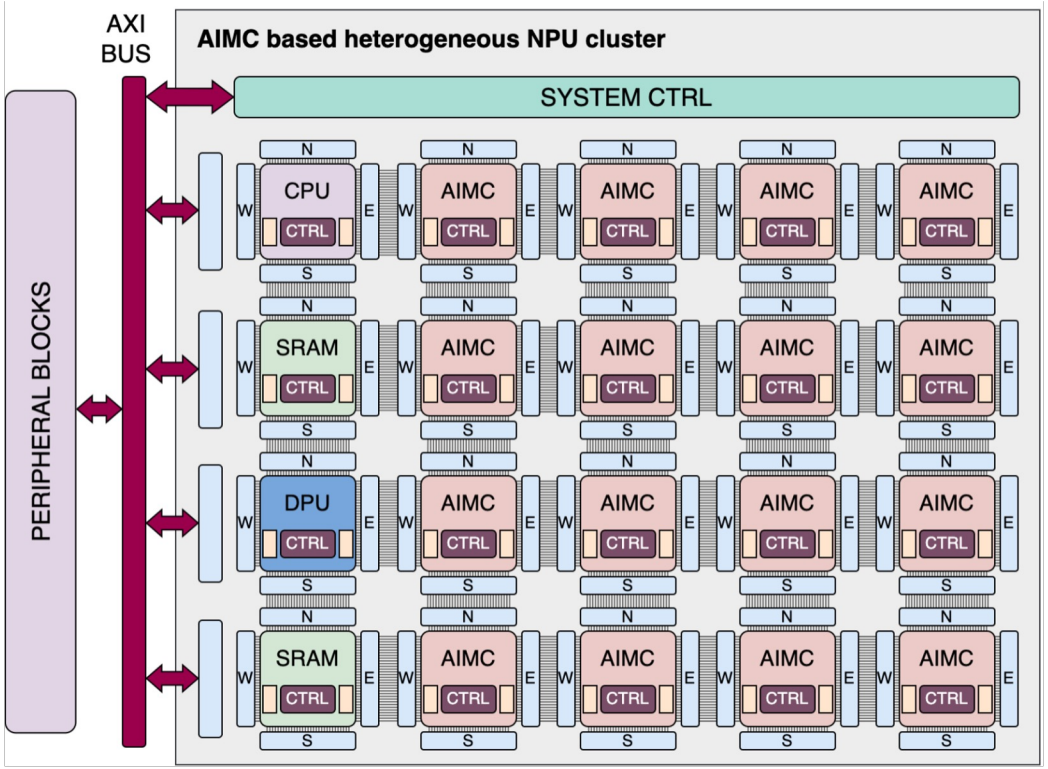
Local Storage Unit



Performs MAC (dot product) operation with high energy efficiency
(**> 90%** of overall operations in AI inference)

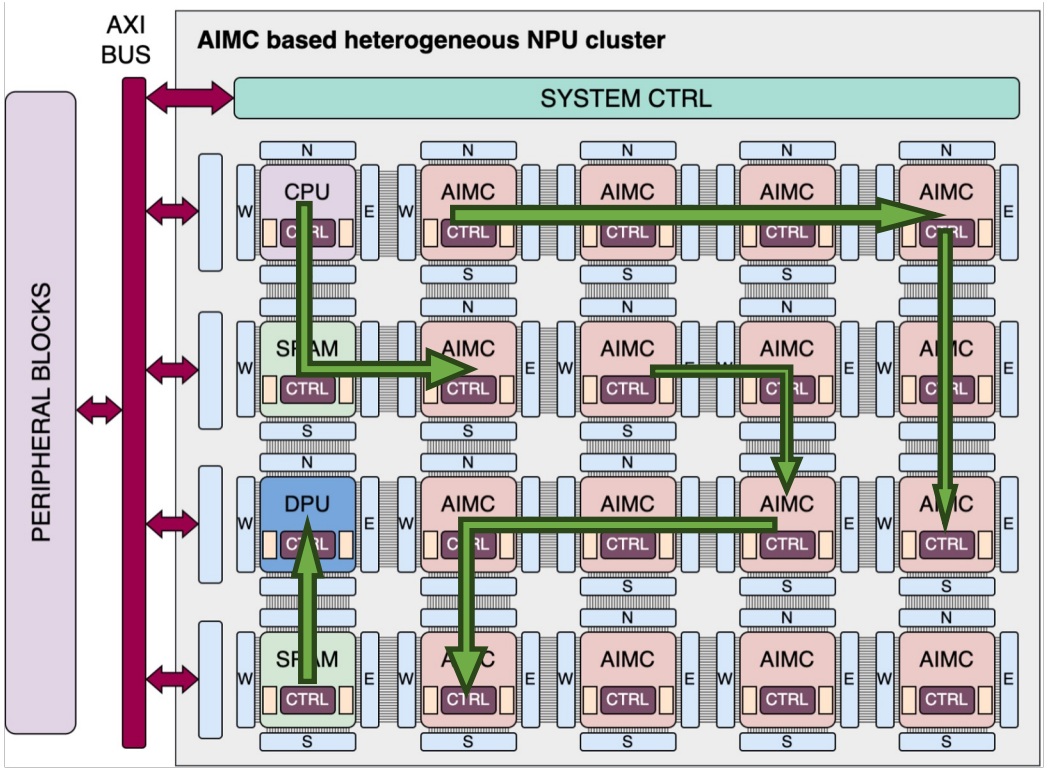
[2] S. Jain et al., "A Heterogeneous and Programmable Compute-In-Memory Accelerator Architecture for Analog-AI Using Dense 2-D Mesh," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 31, no. 1, pp. 114-127, Jan. 2023.

What is Missing… Communication!



Can be **arbitrary** topology, we take mesh as an example

What is Missing… Communication!

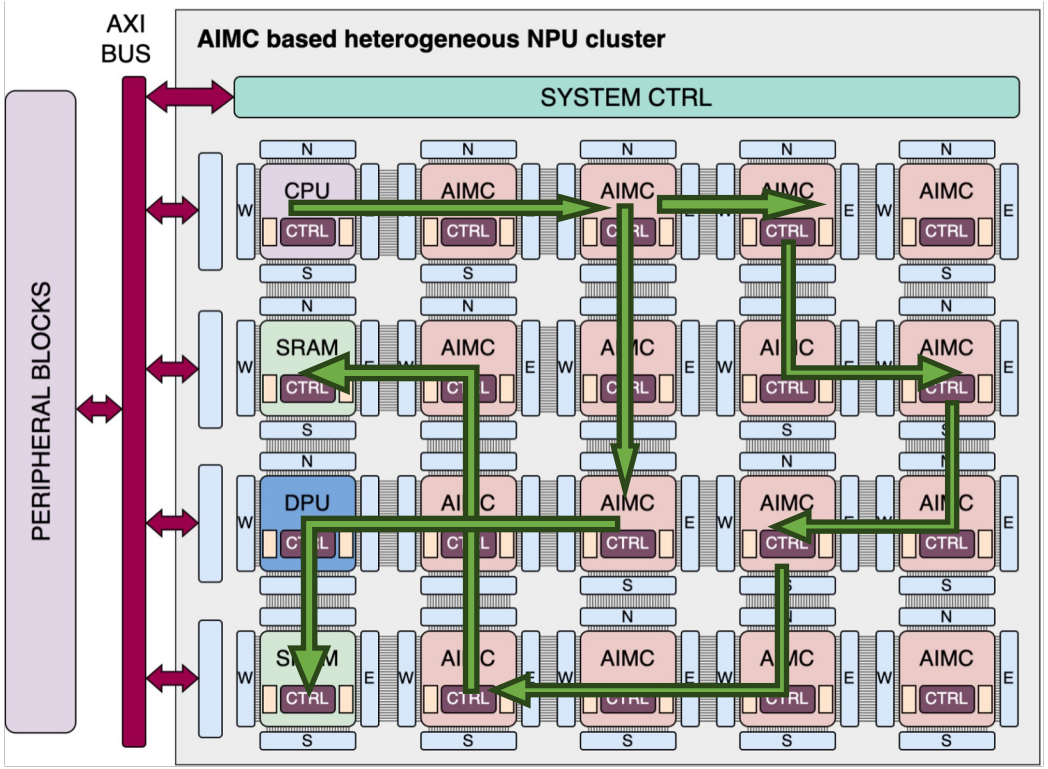


Application 1 Data Flow

Can be **arbitrary** topology, we take mesh as an example

Different AI application mapping leads to various transmission patterns

What is Missing… Communication!



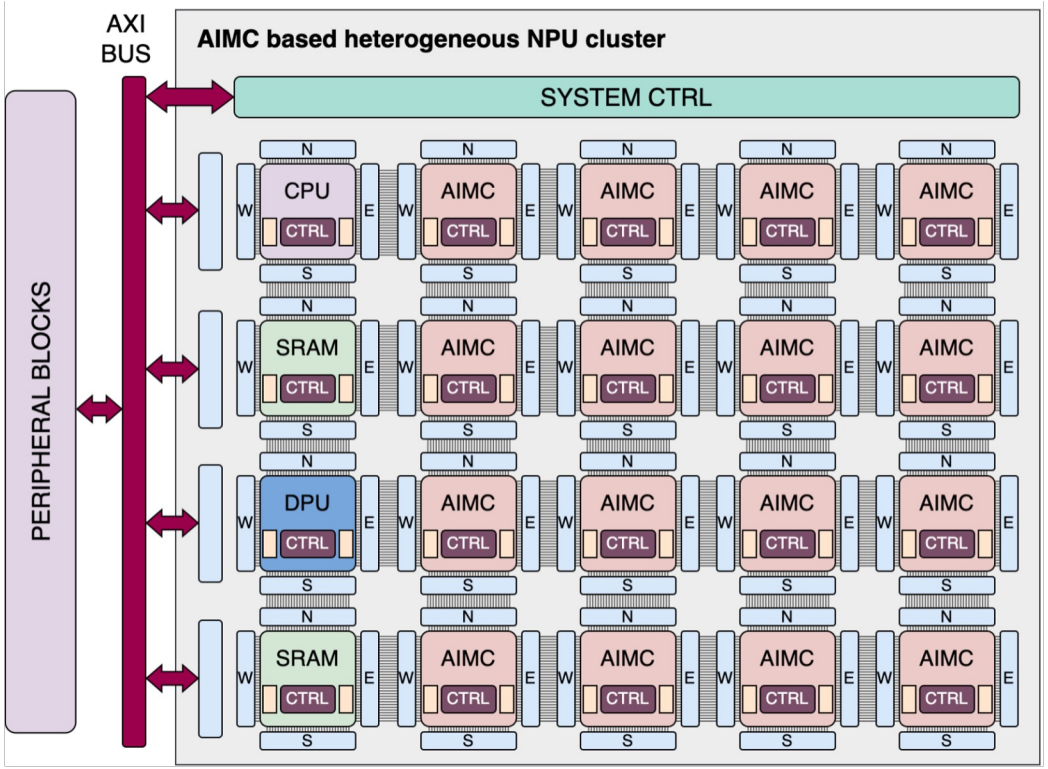
Application 2 Data Flow

Can be **arbitrary** topology, we take mesh as an example

Different AI application mapping leads to various transmission patterns

Data path **reconfiguration**

What is Missing… Communication!



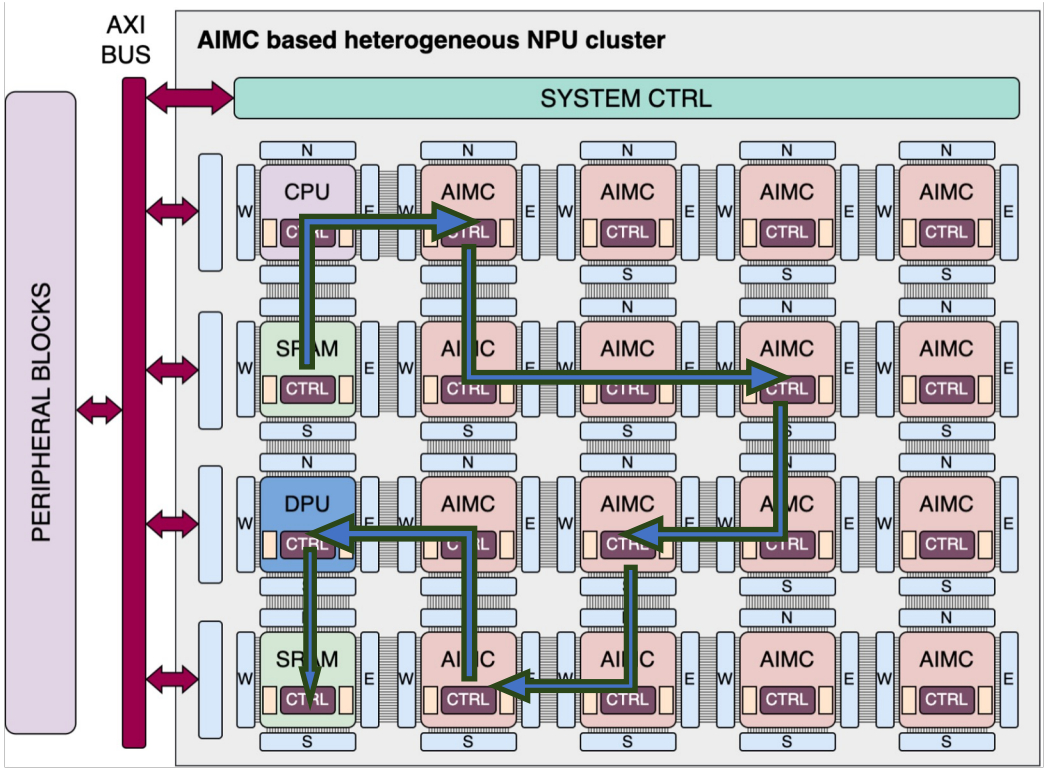
Can be **arbitrary** topology, we take mesh as an example

Different AI application mapping leads to various transmission patterns

Data path **reconfiguration**

Data flow pattern can vary at **different time stages** during a **single** application

What is Missing… Communication!



A single complex application
Time stage 1

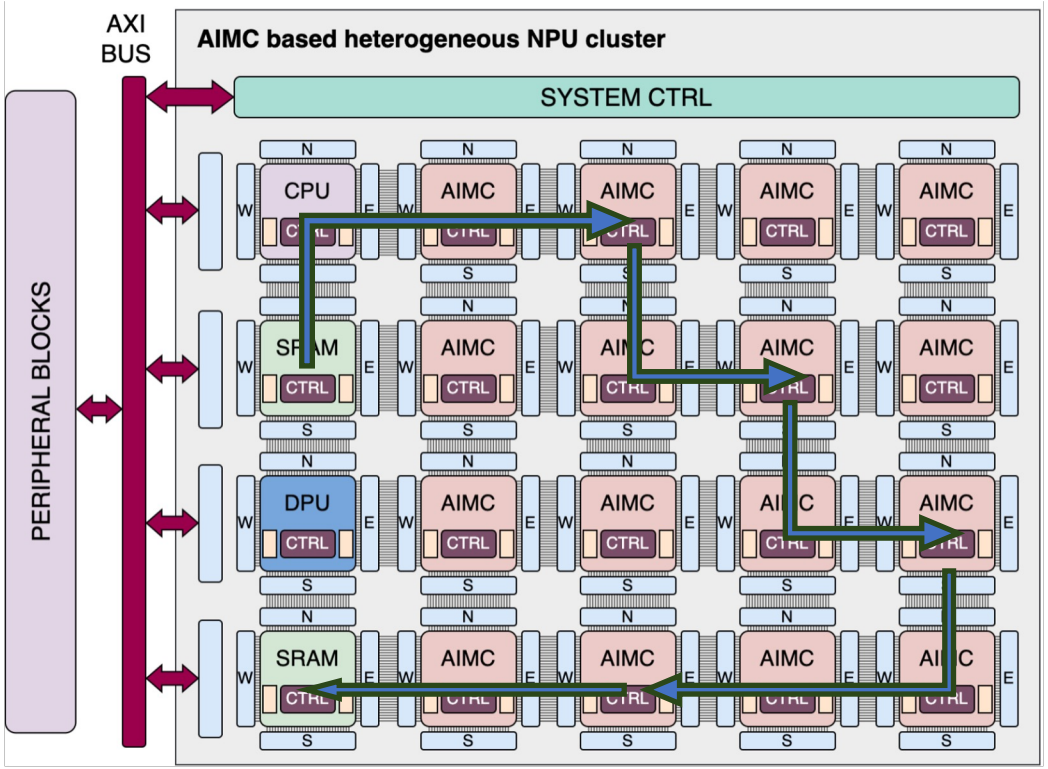
Can be **arbitrary** topology, we take mesh as an example

Different AI application mapping leads to various transmission patterns

Data path **reconfiguration**

Data flow pattern can vary at **different time stages** during a **single** application

What is Missing… Communication!



A single complex application
Time stage 2

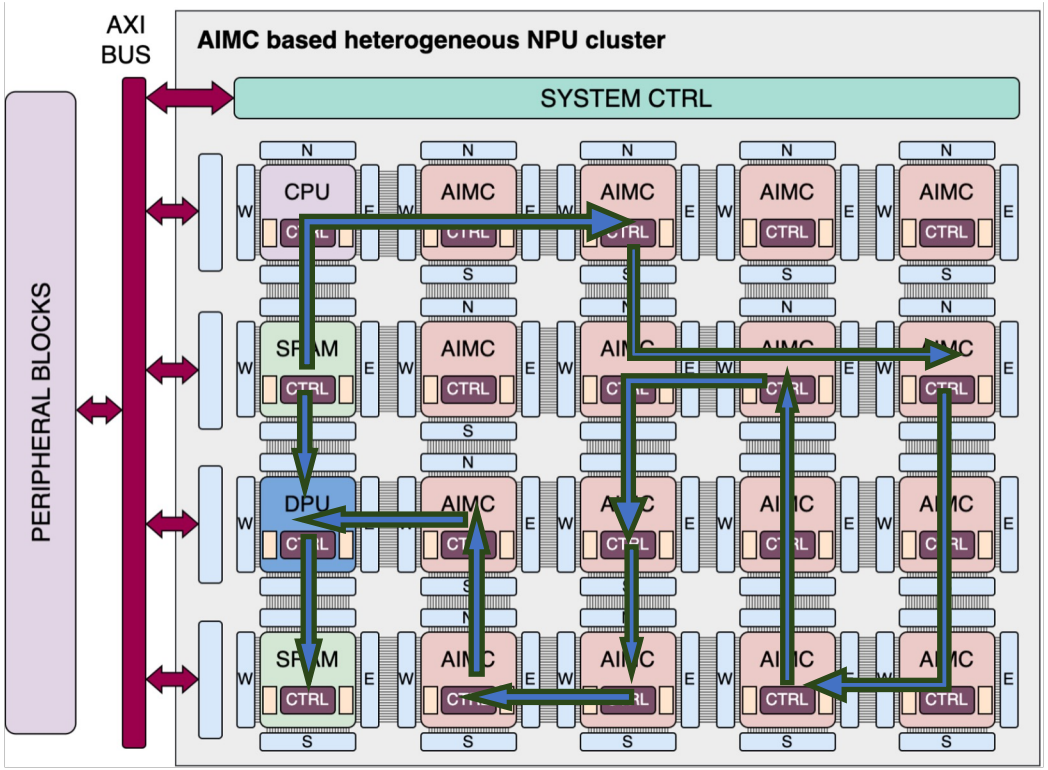
Can be **arbitrary** topology, we take mesh as an example

Different AI application mapping leads to various transmission patterns

Data path **reconfiguration**

Data flow pattern can vary at **different time stages** during a **single** application

What is Missing… Communication!



A single complex application
Time stage 3

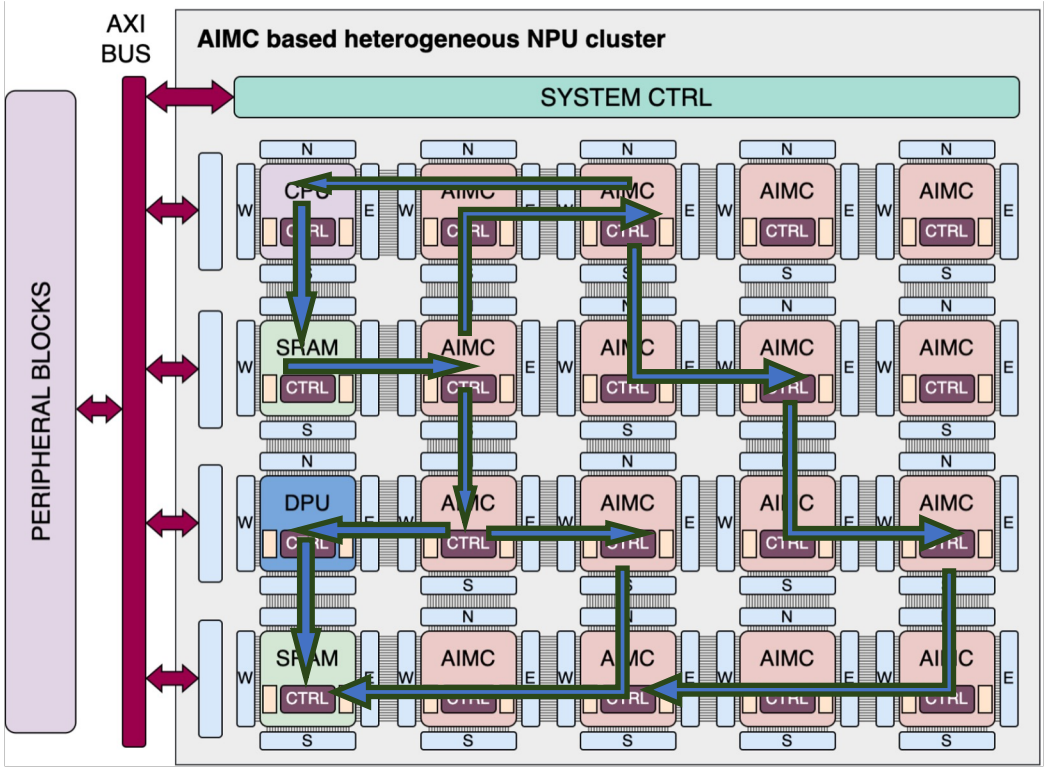
Can be **arbitrary** topology, we take mesh as an example

Different AI application mapping leads to various transmission patterns

Data path **reconfiguration**

Data flow pattern can vary at **different time stages** during a **single** application

What is Missing… Communication!



A single complex application
Time stage 4

Can be **arbitrary** topology, we take mesh as an example

Different AI application mapping[illegible]ds to various transmission patt[illegible]

Data pa[illegible]uration

D[illegible]ttern can vary at **different** tim[illegible]ages during a **single** application

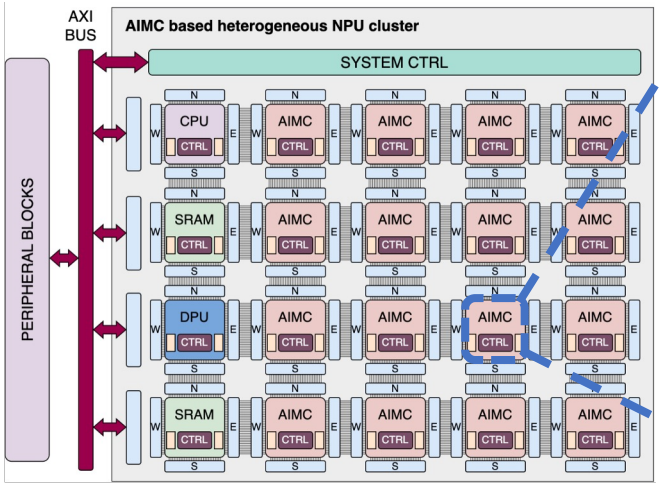
Low Energy Consumption

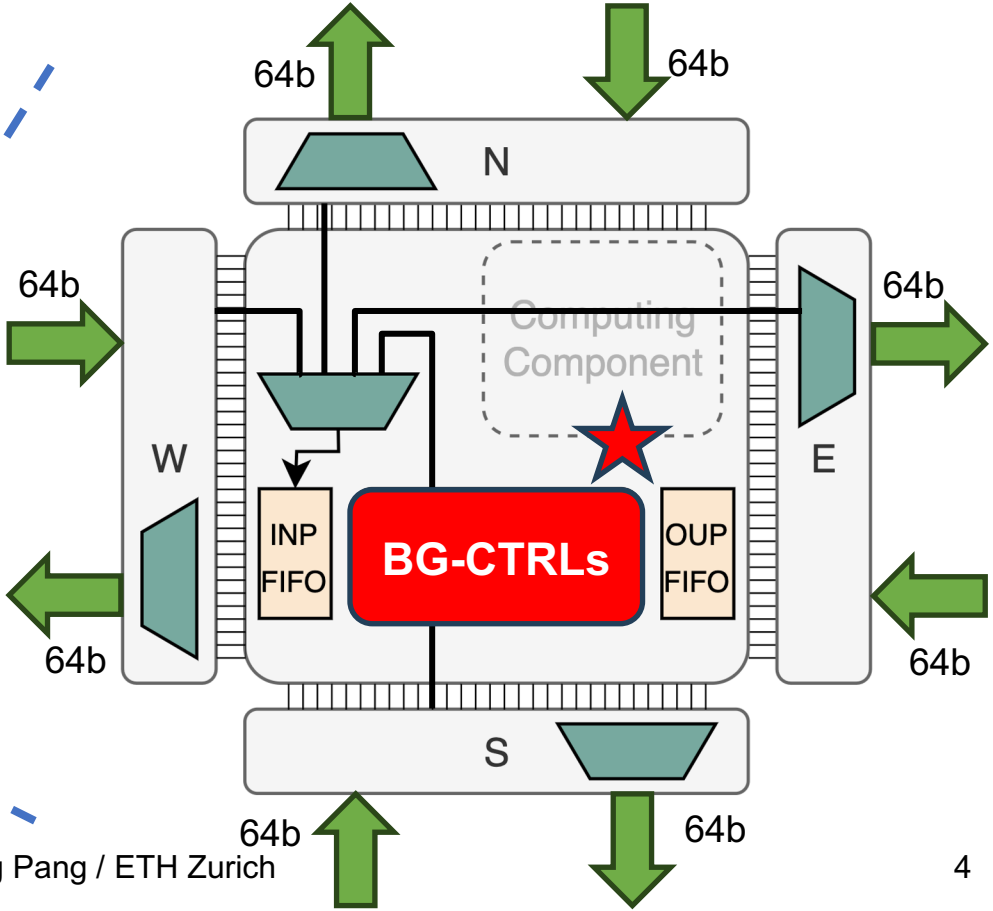
Flexible data path **Switching** during runtime

Our solution is…

BG-CTRL !

NPU Node Architecture





Challenges we have…

Generality

The interconnect can work on various layer mapping and data flow scenarios.

High routing flexibility

Each node can be used for multiple data paths during a **single** application.

Synchronization

All nodes within the cluster should transmit data following dependency.

High Energy Efficiency

Challenges we have…

Generality

High routing flexibility

Fine-grained data path reconfiguration

Synchronization

A global reference timestamp telling when to transmit data

High Energy Efficiency

BG-CTRL solves them!

We solve…

- Generality
- High routing flexibility
- Synchronization
- High Energy Efficiency

by adding…

Programmability

BG-CTRL solves them!

We solve…

- **Generality**
- **High routing flexibility**
- **Synchronization**
- **High Energy Efficiency**

by adding…

Programmability

We design a **specialized instruction set**…

Instruction tells BG-CTRL…

- **When** to transmit data from **where** to **where**
- **Loop** repeated data transmission pattern

This enables…

- **Cycle-wise** data path formation/switching
- **Precise timestamp** of data transmission
- **Less energy consumption** on instr. fetching

BG-CTRL solves them!

Generality

Synchronization

High routing flexibility

High Energy Efficiency

Programmability

High Energy Efficiency

Multiple Operating Modes

BG-CTRL solves them!

Generality

High routing flexibility

Synchronization

High Energy Efficiency

Programmability

High Energy Efficiency

Multiple Operating Modes

Fixed Configuration

Form data path via **configuration registers**

- **Cannot** reconfigure data path
- Data is transmitted whenever available
- **Higher energy efficiency** than program
- Recommended for simple data flow

BG-CTRL solves them!

Generality

High routing flexibility

Synchronization

High Energy Efficiency

Programmability

High Energy Efficiency

Multiple Operating Modes

Time-sliced Configuration

Extend fixed configuration with several **data path sets**, each activated in certain period

- **More flexible** than fixed configuration
- **Higher energy efficiency** than program
- Recommended for data flow with regular path switching

BG-CTRL solves them!

Generality

High routing flexibility

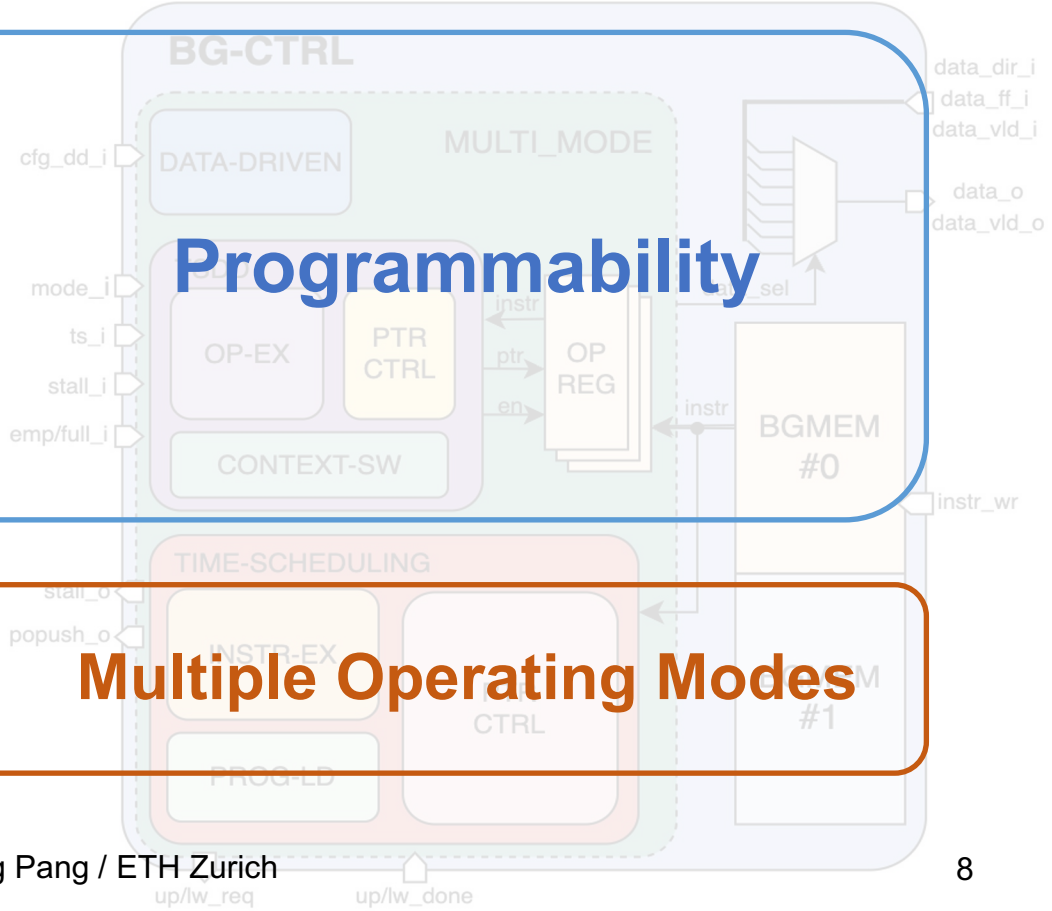
Synchronization

High Energy Efficiency

Programmability

High Energy Efficiency

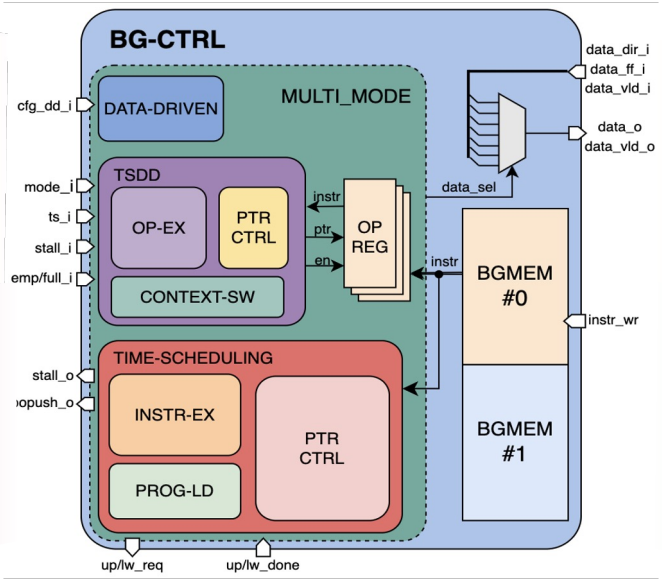
Multiple Operating Modes



STM FD-SOI 28nm for Evaluation

We integrate it into node and build a 3x3 mesh…

640 MHz
@SS Corner



200 kGE
~5% of the node area

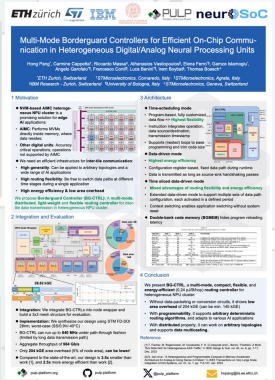
Aggregate throughput
984 Gb/s

0.24-0.41 pJ/B/hop
@typical corner

We present BG-CTRL…

A **multi-mode, flexible**, **compact** and **energy-efficient** (0.24 pJ/B/hop) **routing controller** for heterogeneous NPU cluster, which is planned to be **taped-out** in the near future.

Welcome to the poster for more details! 👉



- Support for **arbitrary deterministic routing algorithms**;
- Capability on **data multicasting & path through**;
- Capability on **context switching**;
- Adaptability to **various topologies**;
- Implementation & Evaluation in **STM FD-SOI 28nm** technology;
- BG-CTRL **integration** into node.