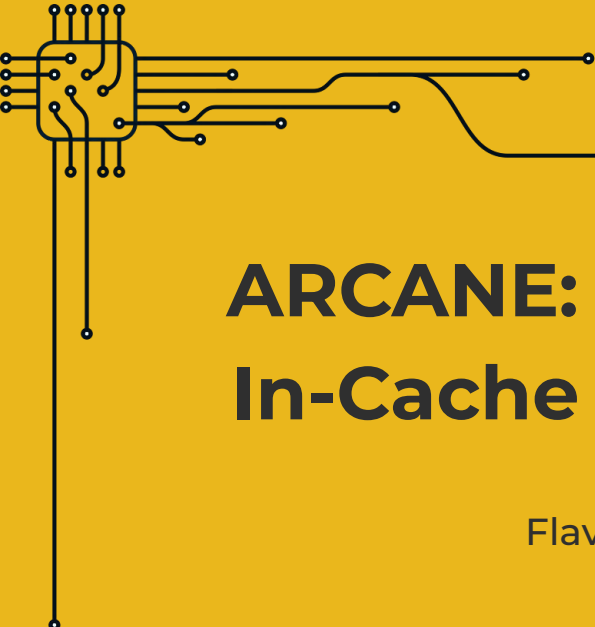




ARCANE: Enabling High-Performance In-Cache Tensor Extensions in RISC-V

Flavia Guella, Vincenzo Petrolo, Michele Caon,
Guido Masera, Maurizio Martina



Politecnico
di Torino

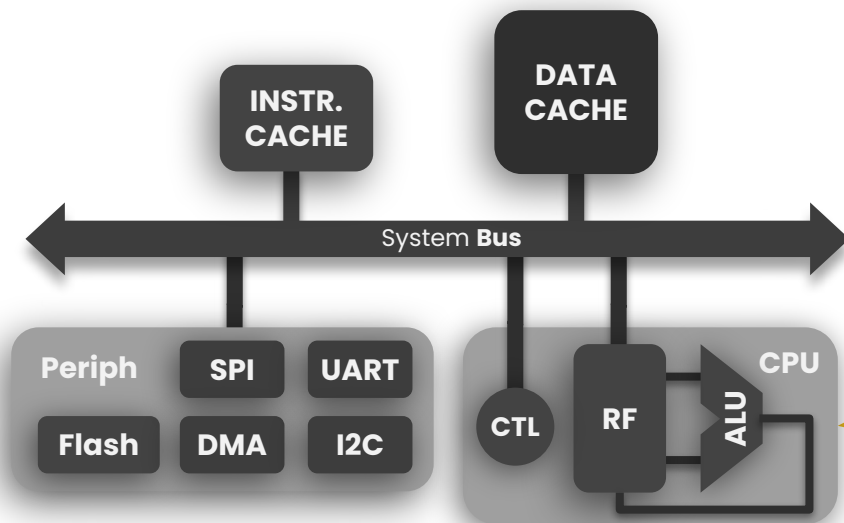


EPFL

RISC-V Summit Europe 2026,
Bologna, Italy



Traditional SoC

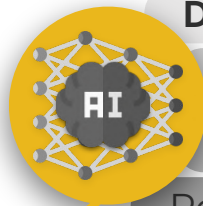


Data-driven tasks

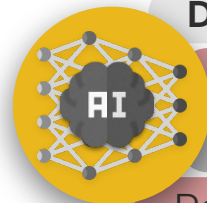
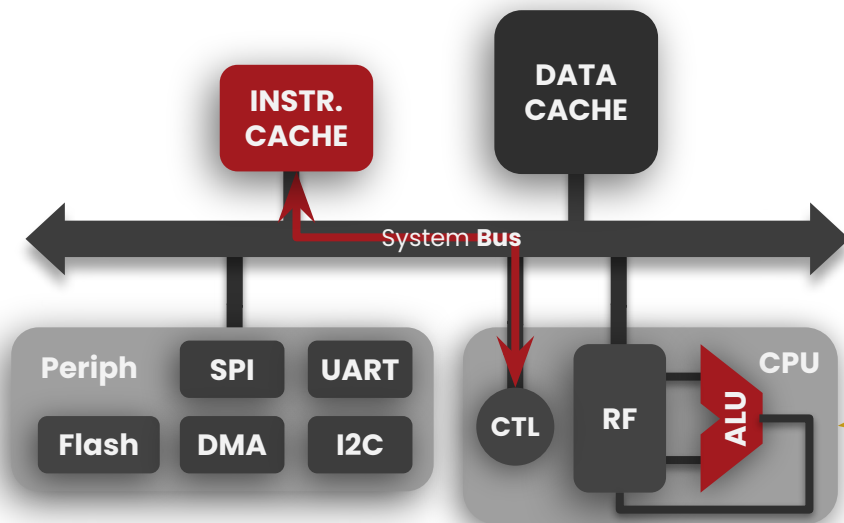
Massive **computation** requirements

Memory **bandwidth** limits

Poor energy **efficiency**



Traditional SoC



Data-driven tasks

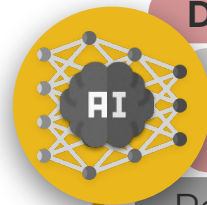
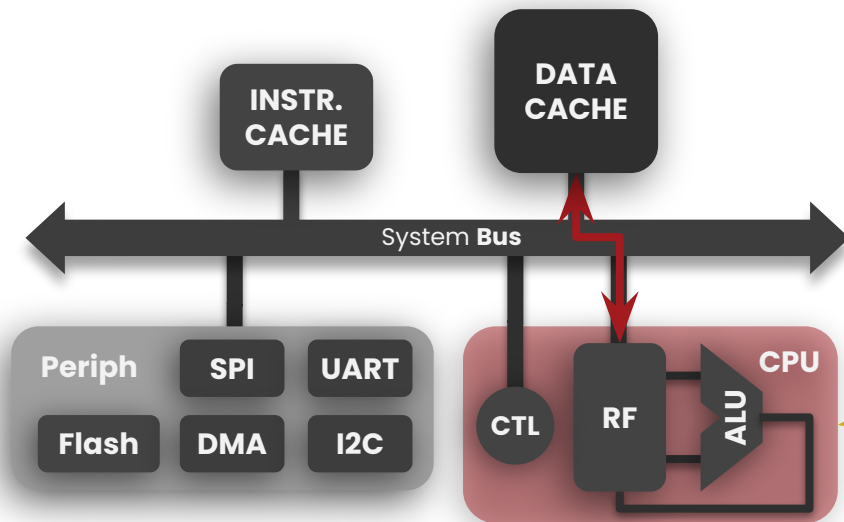
Massive **computation** requirements

Memory **bandwidth** limits

Poor energy **efficiency**



Traditional SoC



Data-driven tasks

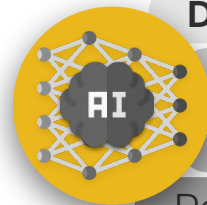
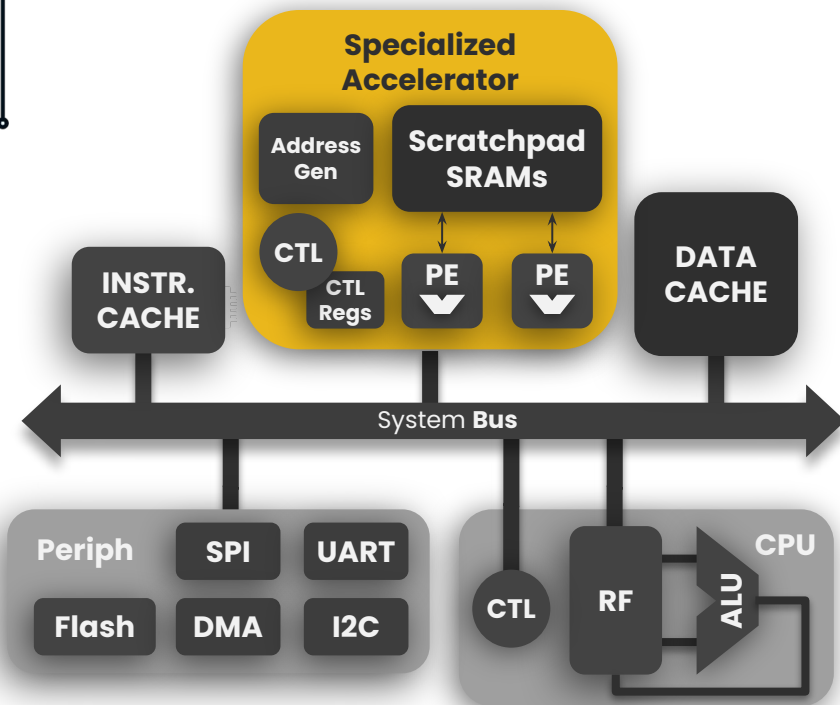
Massive **computation** requirements

Memory **bandwidth** limits

Poor energy **efficiency**



Heterogeneous Systems



Data-driven tasks

Massive **computation** requirements

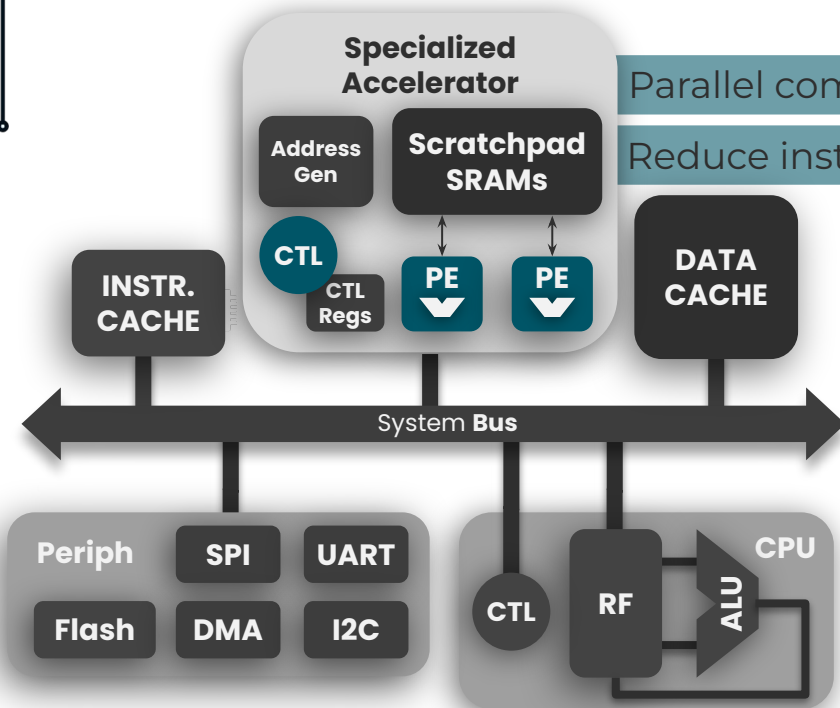
Memory **bandwidth** limits

Poor energy **efficiency**



Heterogeneous Systems

Strengths



Parallel computation

Reduce instruction fetch cost

Data-driven tasks

Massive **computation** requirements

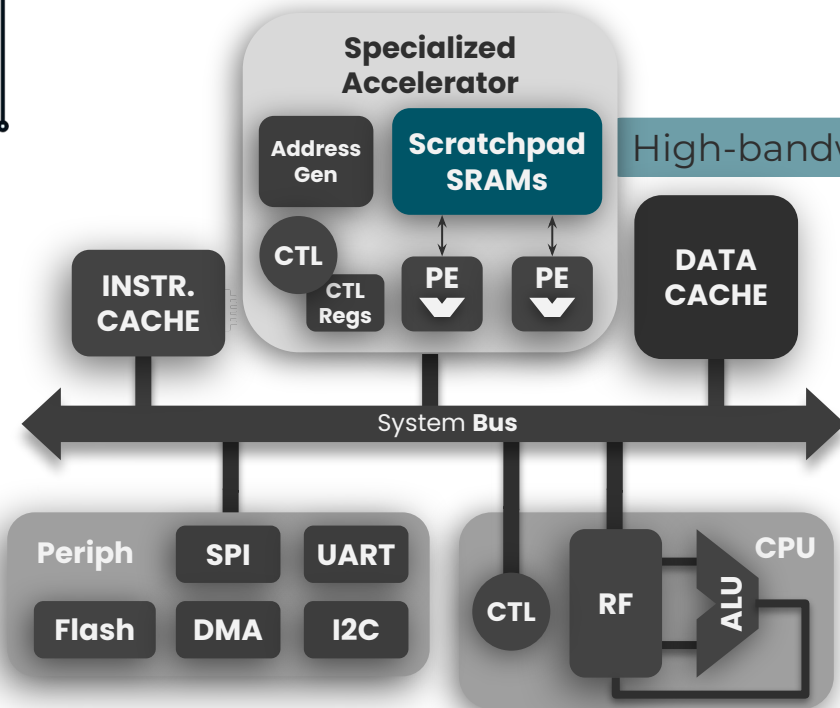
Memory **bandwidth** limits

Poor energy **efficiency**



Heterogeneous Systems

Strengths



Data-driven tasks

Massive **computation** requirements

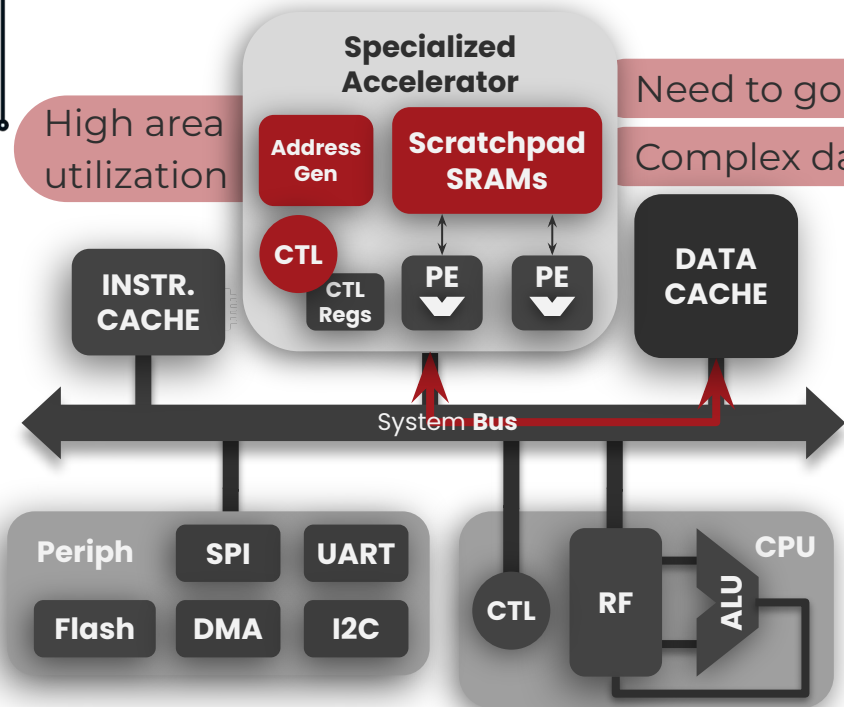
Memory **bandwidth** limits

Poor energy **efficiency**



Heterogeneous Systems

Drawbacks



High area utilization

Need to go through system bus

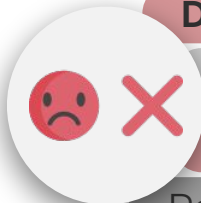
Complex data management

Data-driven tasks

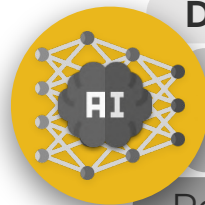
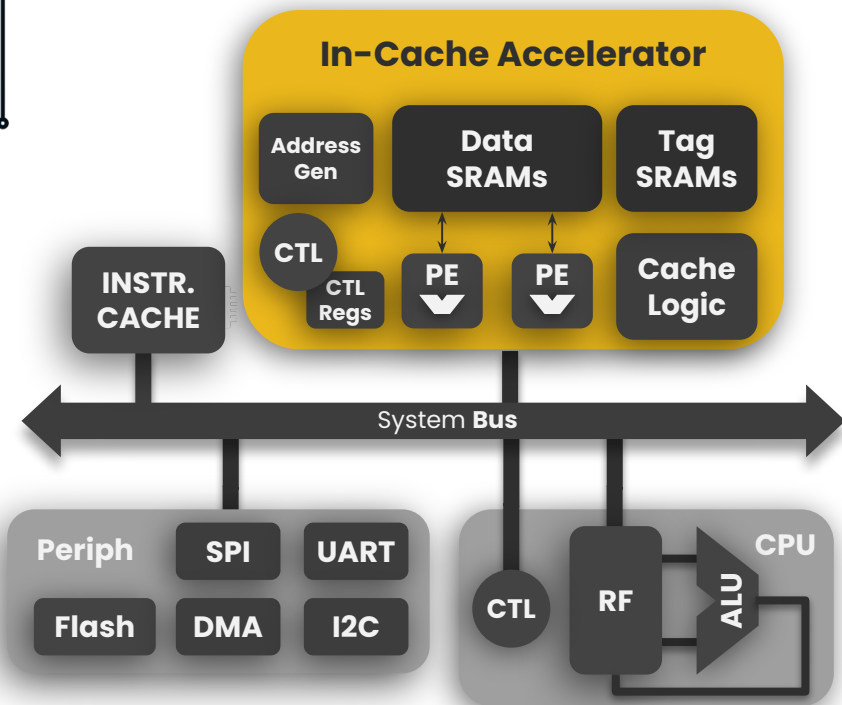
Massive **computation** requirements

Memory **bandwidth** limits

Poor energy **efficiency**



In-Cache Computing



Data-driven tasks

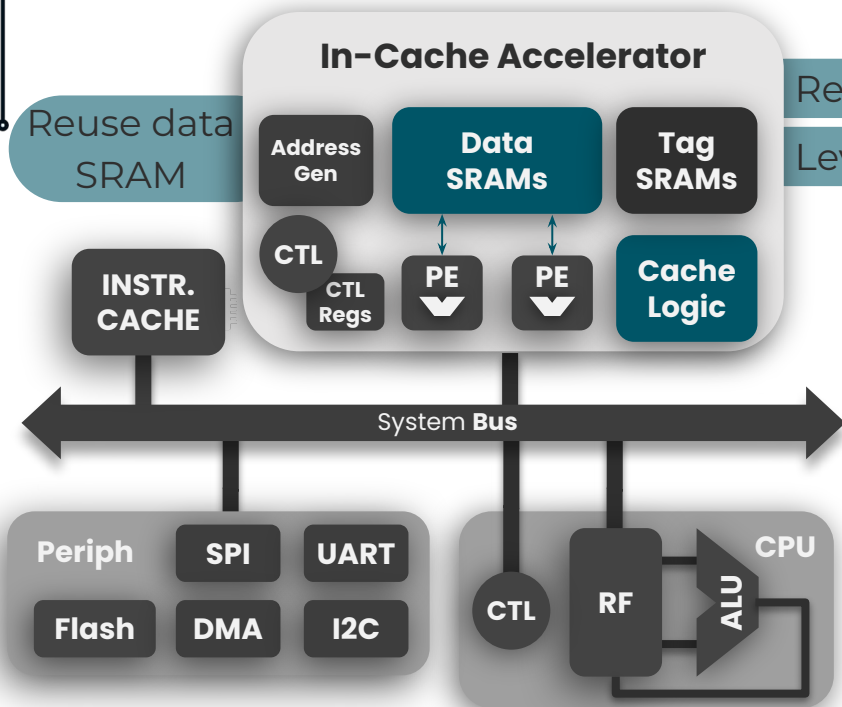
Massive **computation** requirements

Memory **bandwidth** limits

Poor energy **efficiency**



In-Cache Computing



Data-driven tasks

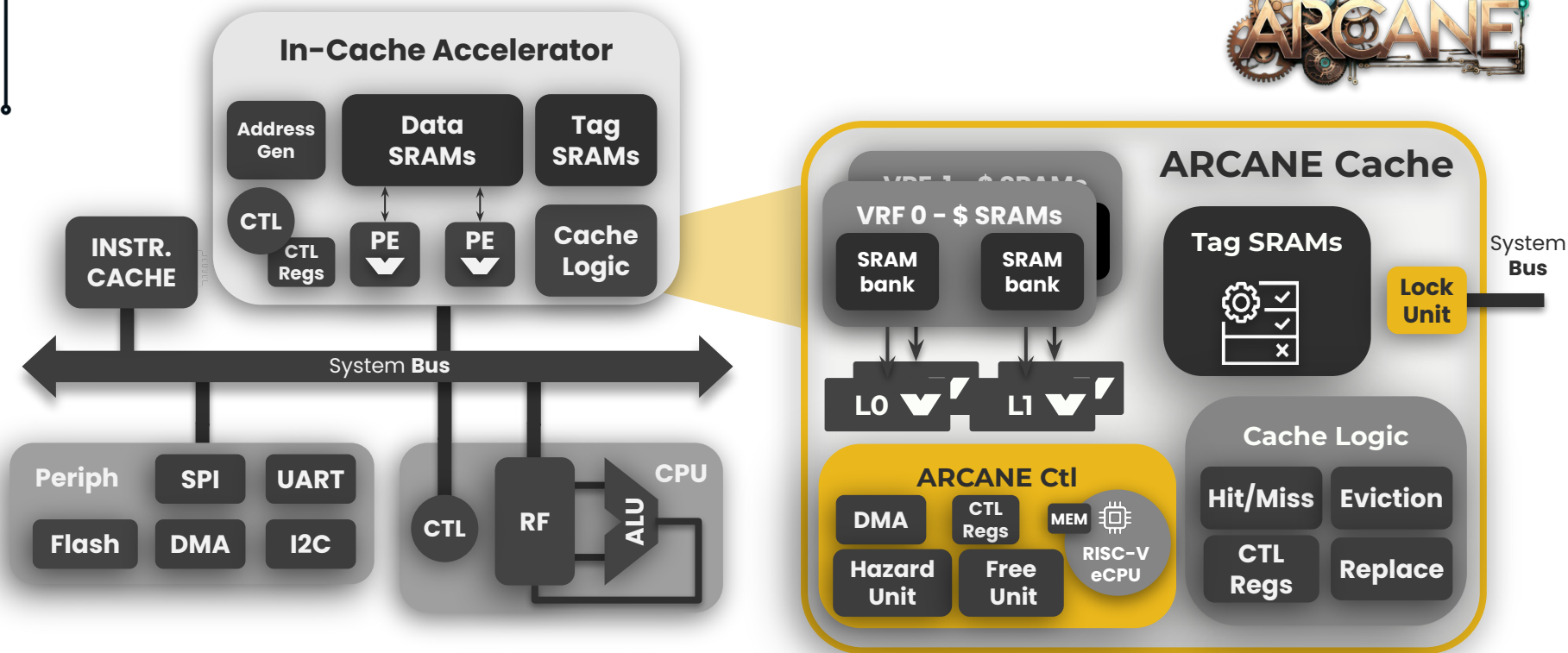
Massive **computation** requirements

Memory **bandwidth** limits

Poor energy **efficiency**



ARCANE: In-Cache Computing



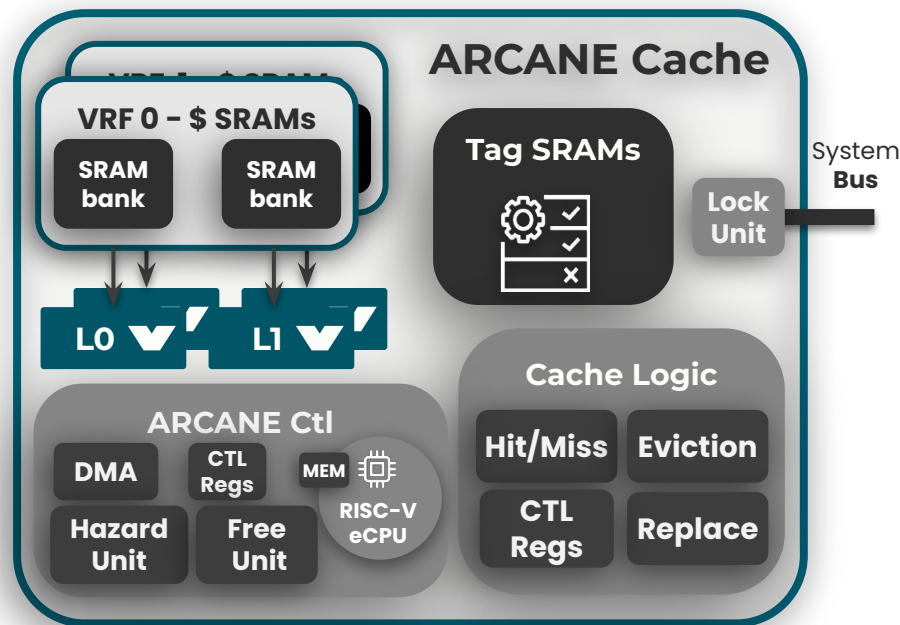
Computing Units



Multi-lane Vector Processing Units

Sliced Vector Register File reuses data-cache SRAMs

Tunable VRF size, #lanes, #VPUs



Cache Logic

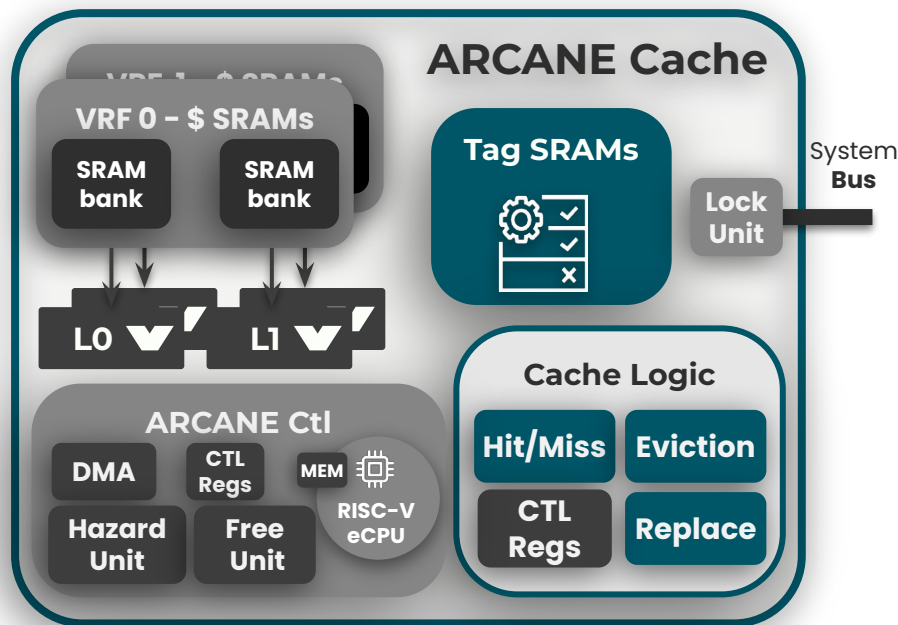


Additional Status Flag



Concurrent computing
and cache mode

Fine-grained selection of usable cache
lines and lines busy computing



Data Management Logic

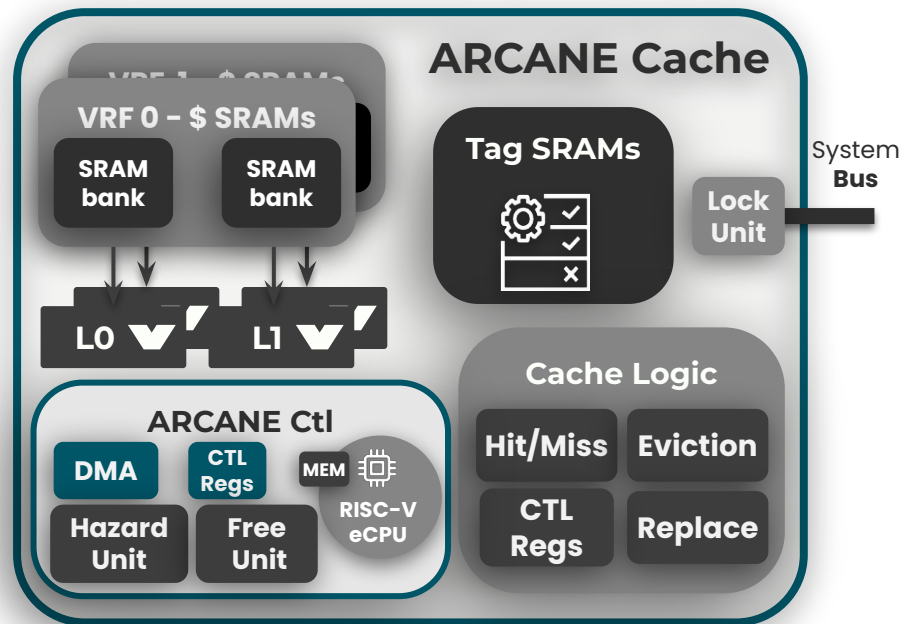


Data Management Logic



Reorganizes data according to VRF layout for desired kernel

Uses cache logic to check for hit on requested data



Data Management and Cache



Data Management Logic

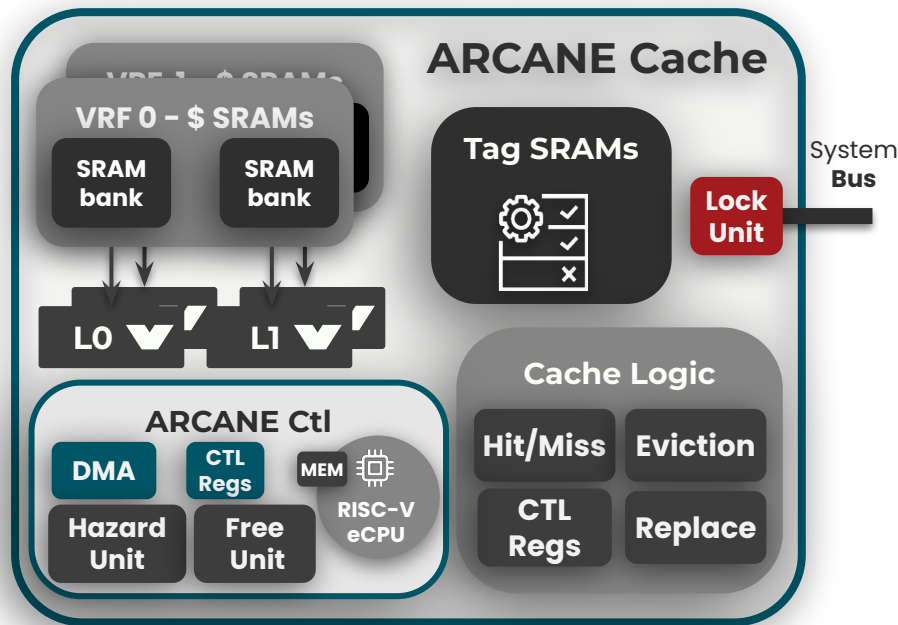


Reorganizes data according to VRF layout for desired kernel

Uses cache logic to check for hit on requested data



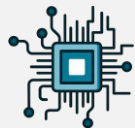
During data movement cache cannot answer system requests



Computing and Control

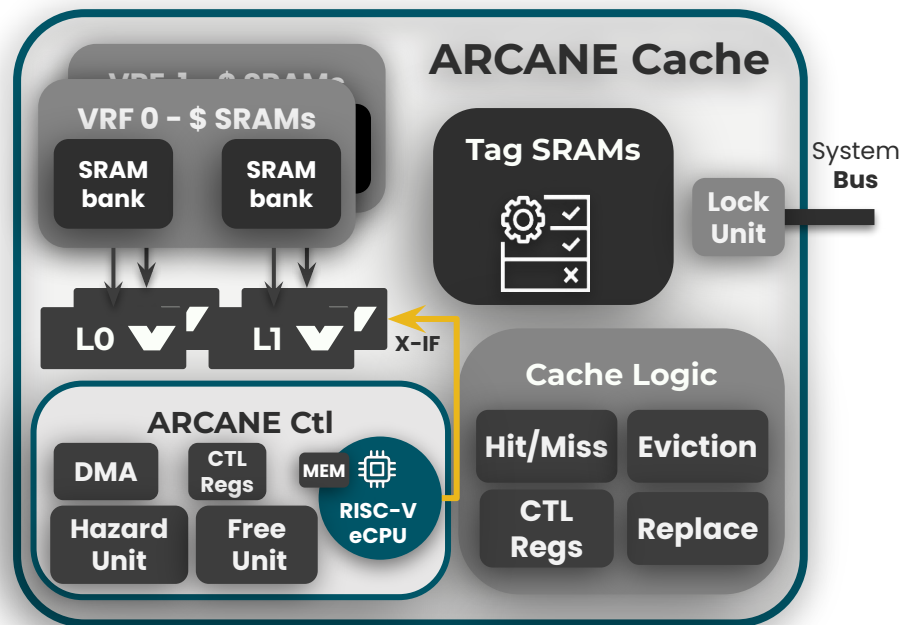


RISC-V Embedded CPU

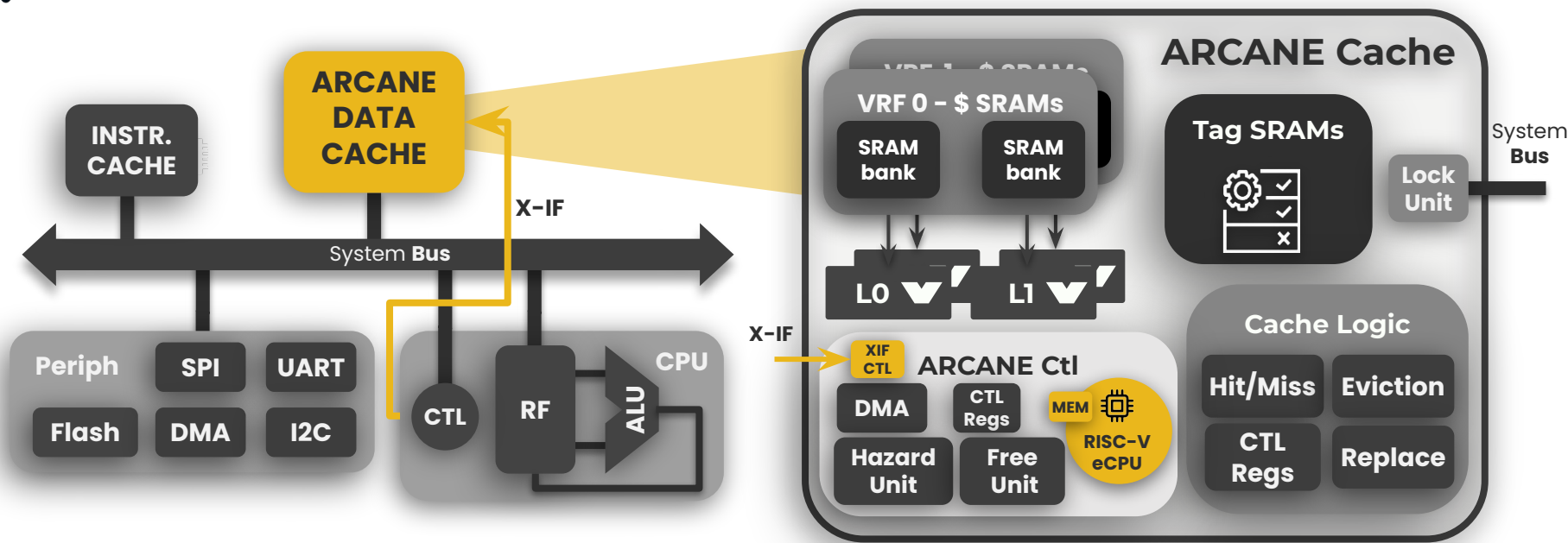


Custom subset of Zve32x (C-V-XIF)

Orchestrates other modules



ARCANE: In-Cache Computing



Programming Model

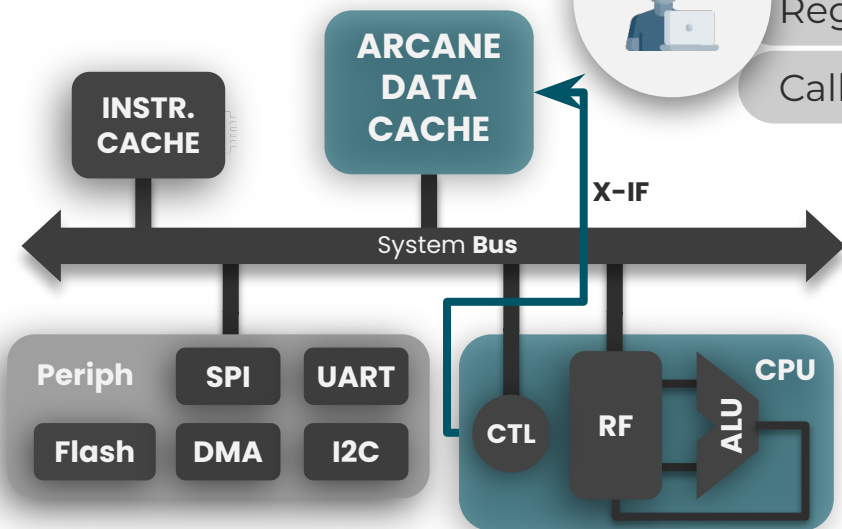


Tensor-based ISA extension

Software-defined ISA

Register 4-dimensional tensors

Call a registered kernel



```
main.c

int main(void) {
    int A[rowsA][colsA] = {...};
    int F[rowsF][colsF] = {...};
    int R[rowsR][colsR];

    // Reservation
    _xmr_w(m0, A, rowsA, colsA);
    _xmr_w(m1, F, rowsF, colsF);
    _xmr_w(m2, R, rowsR, colsR);

    // Matrix Kernel
    _xmk0(m2, m0, m1);
    ...

    // Access to results
    print(R[0][0]);
    ...
}
```

Programming Model



Tensor-based ISA extension

Software-decoded instructions

Auto-tiling

Rich kernel library

ARCANE Runtime

Kernel Decoder

Kernels Library



Kernel Queue



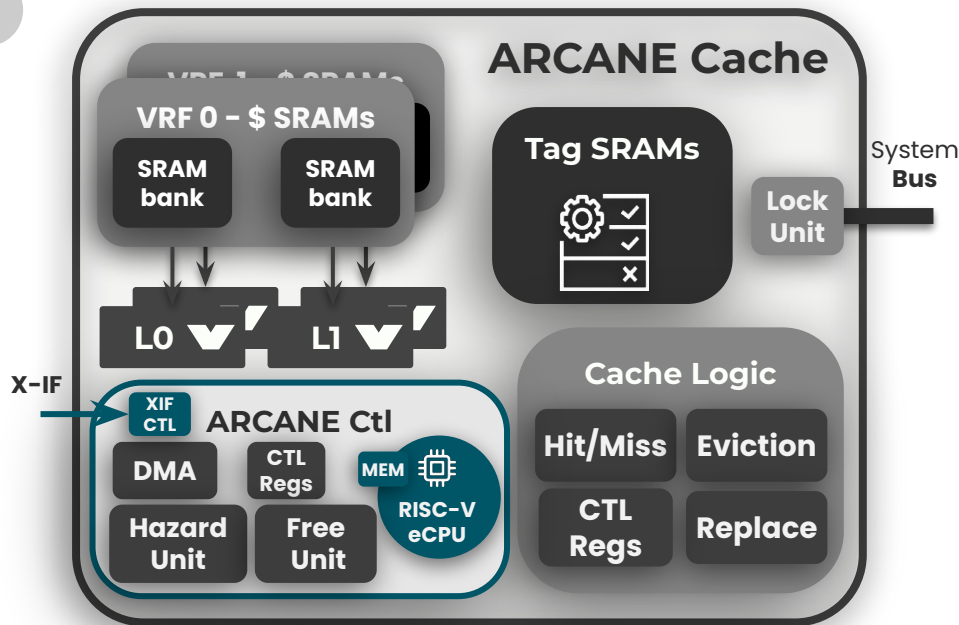
Matrix Map

Kernel Scheduler

Allocate

Kernel Execution

Writeback



Programming Model



Programmable



Data-management is now **transparent** to the host CPU

ARCANE Runtime

Kernel Decoder

Kernels Library



Kernel Queue



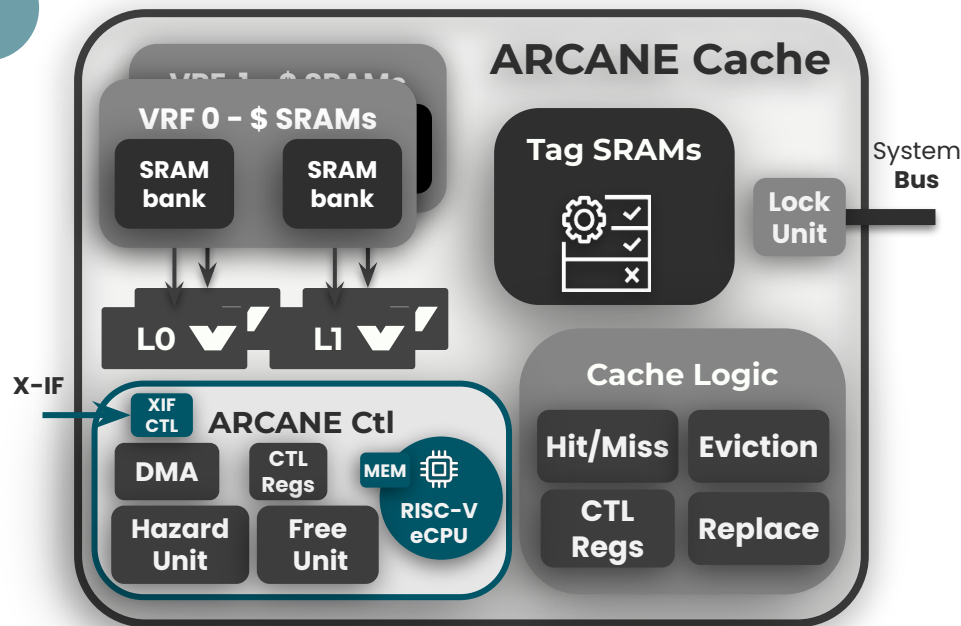
Matrix Map

Kernel Scheduler

Allocate

Kernel Execution

Writeback



Programming Model



Programmable

Data-management is now
transparent to the host CPU

No explicit synchronization

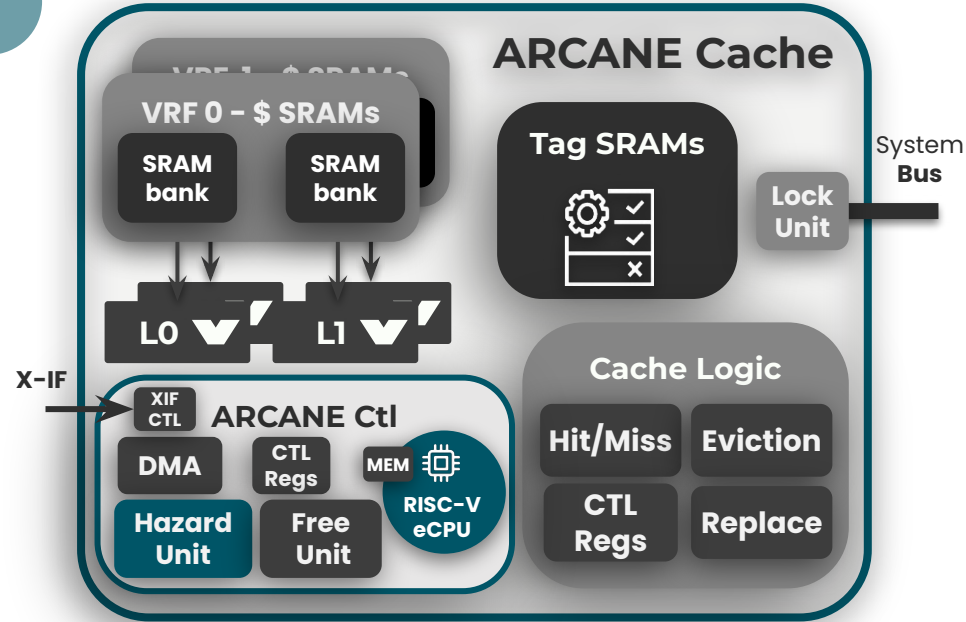
```
main.c

int main(void) {
    int A[rowsA][colsA] = {...};
    int F[rowsF][colsF] = {...};
    int R[rowsR][colsR];

    // Reservation
    _xmr_w(m0, A, rowsA, colsA);
    _xmr_w(m1, F, rowsF, colsF);
    _xmr_w(m2, R, rowsR, colsR);

    // Matrix Kernel
    _xmk0(m2, m0, m1);
    ...

    // Access to results
    print(R[0][0]);
    ...
}
```



Programming Model



Programmable



Data-management is now **transparent** to the host CPU

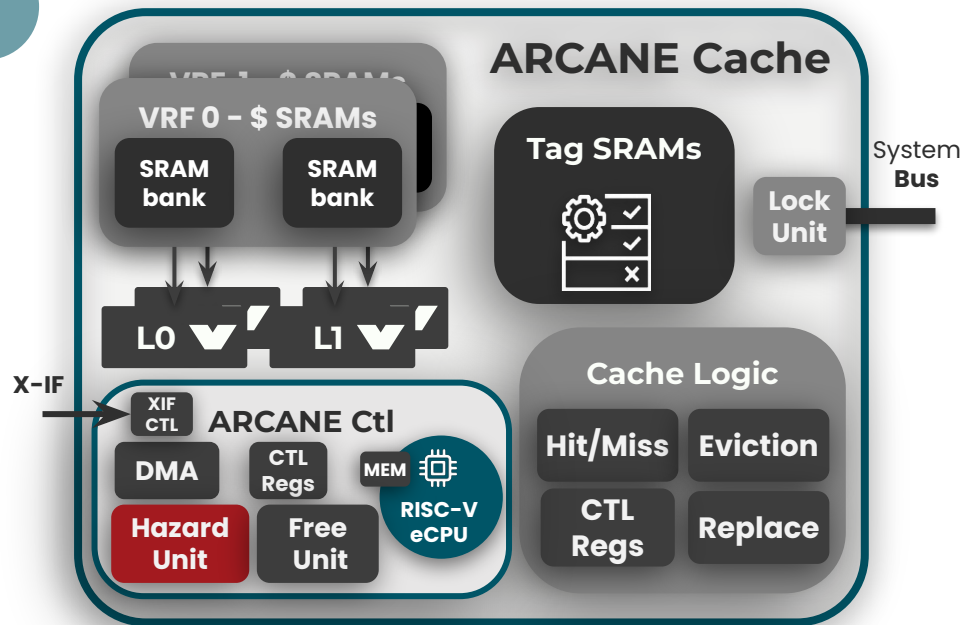
No explicit synchronization

Hazard Unit

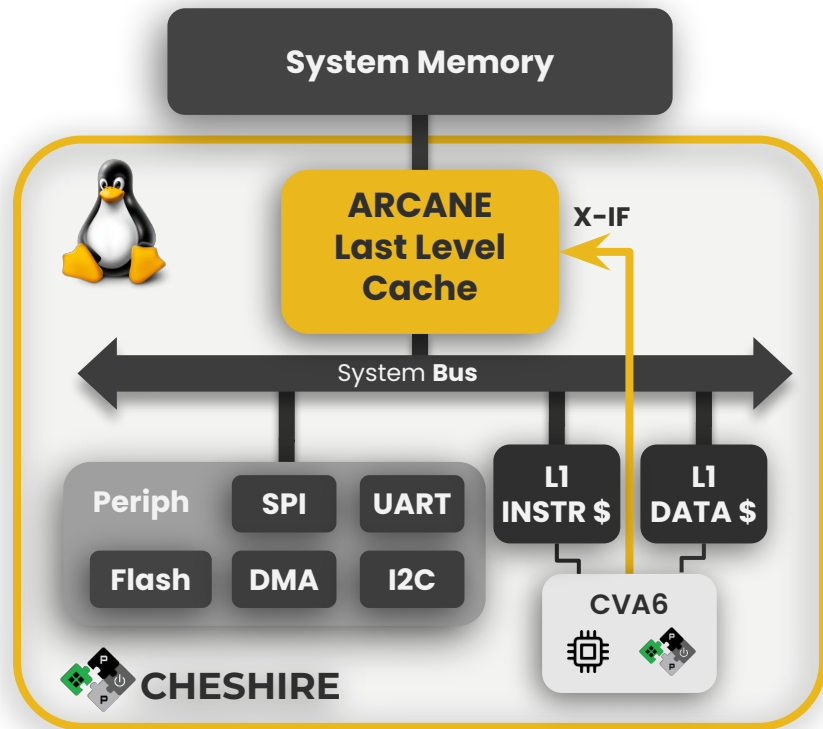
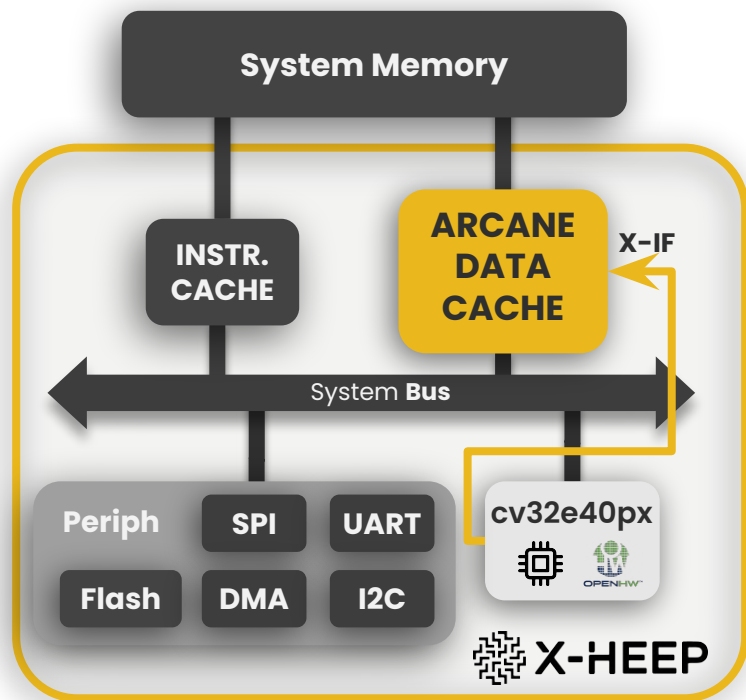


Tracks registered tensors (**xmr**)

Selectively blocks critical mem operations until data dependency is solved

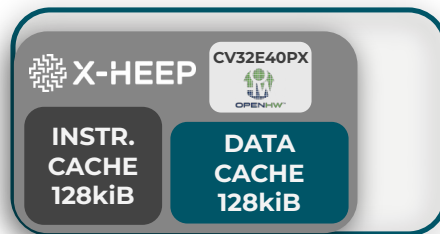
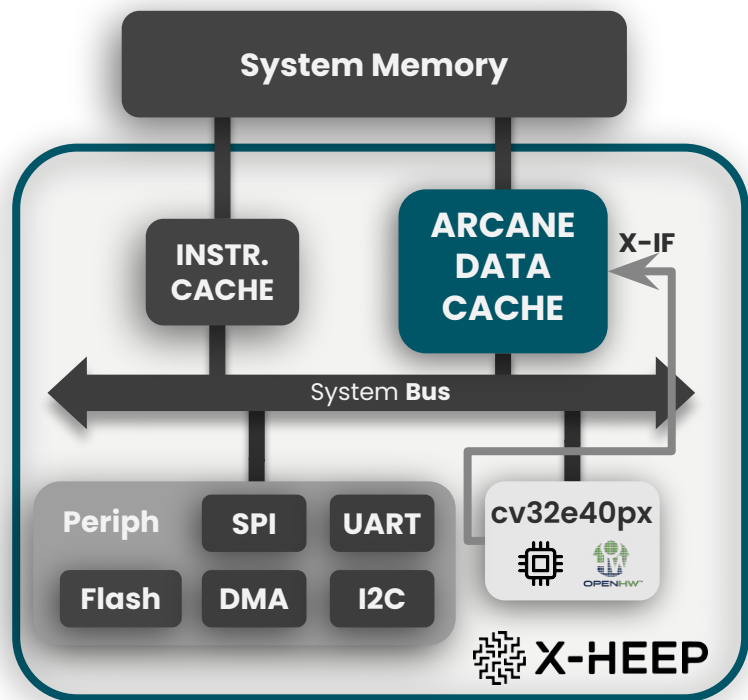


System Integration



Ottaviano et al., "Cheshire: A Lightweight, Linux-Capable RISC-V Host Platform for Domain-Specific Accelerator Plug-In"
Machetti et al., "X-HEEP: An Open-Source, Configurable and Extendible RISC-V Platform for TinyAI Applications"

Synthesis Results

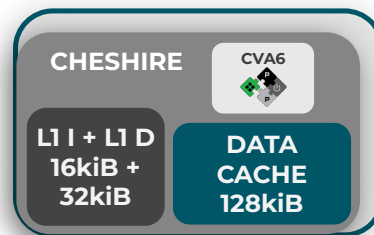
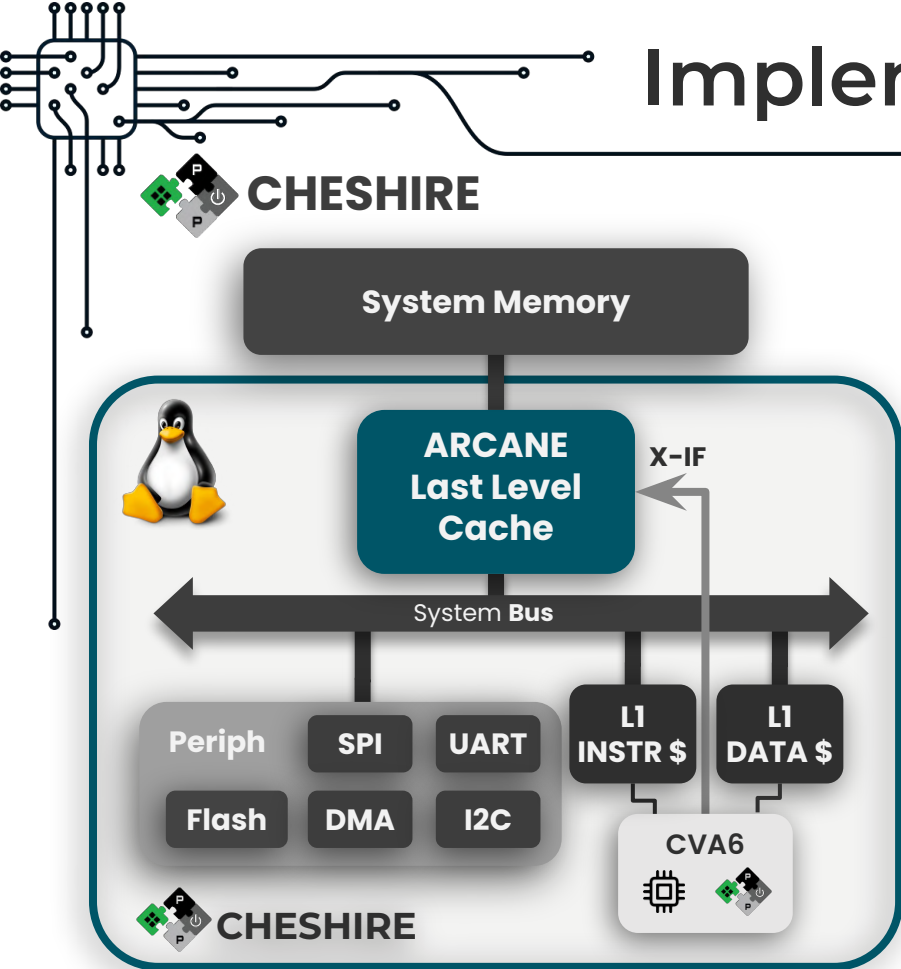


1.4x X-HEEP area
@65nm

4 VPU's
8 lanes
16kiB eMem

6x speed up on
CV32E40P

Implementation Results



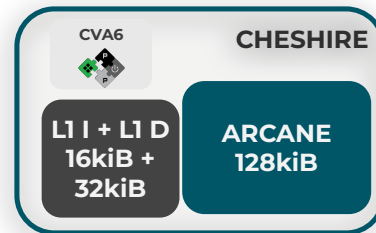
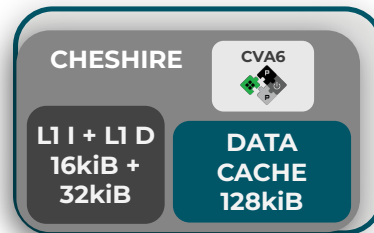
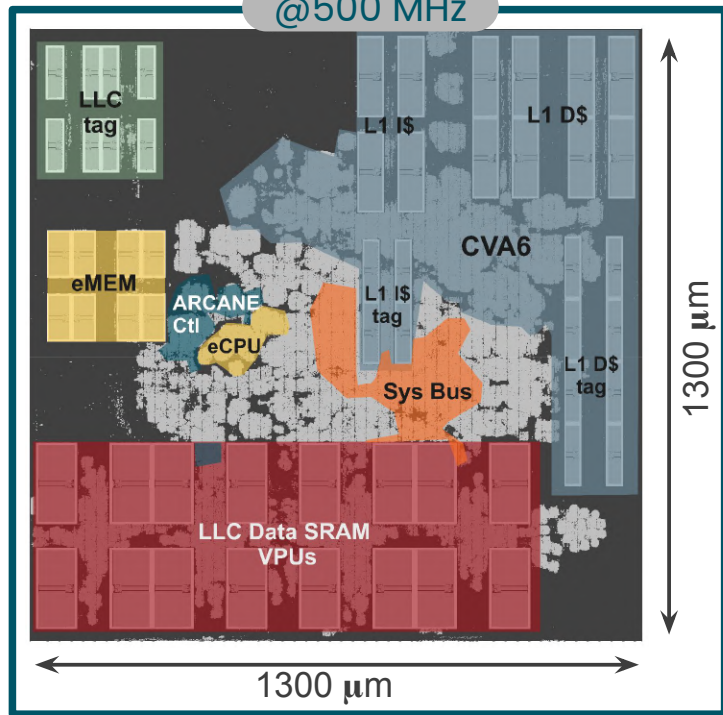
4 VPU
8 lanes
16kiB eMem

1.2x Cheshire area
@22nm

Implementation Results



@500 MHz



1.4x LLC area

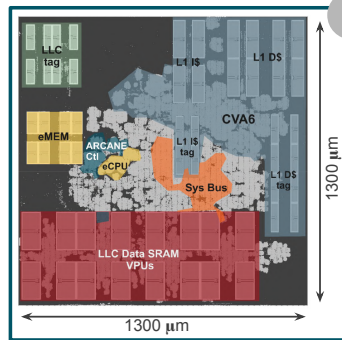
46% due to the VPU logic

36% due to eCPU and eMEM,

18% to Cache Ctl



Benchmarking



@500 MHz

int8 int16 int32



ARCANE, 2 lanes



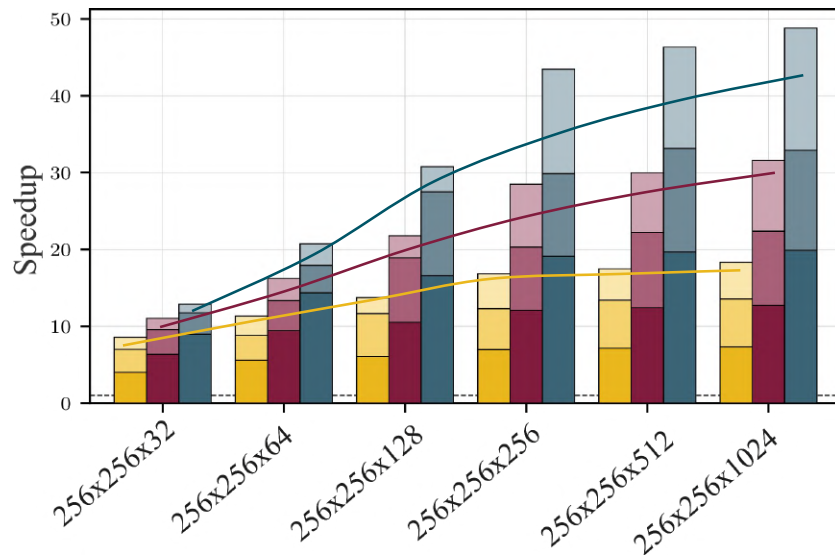
ARCANE, 4 lanes



ARCANE, 8 lanes

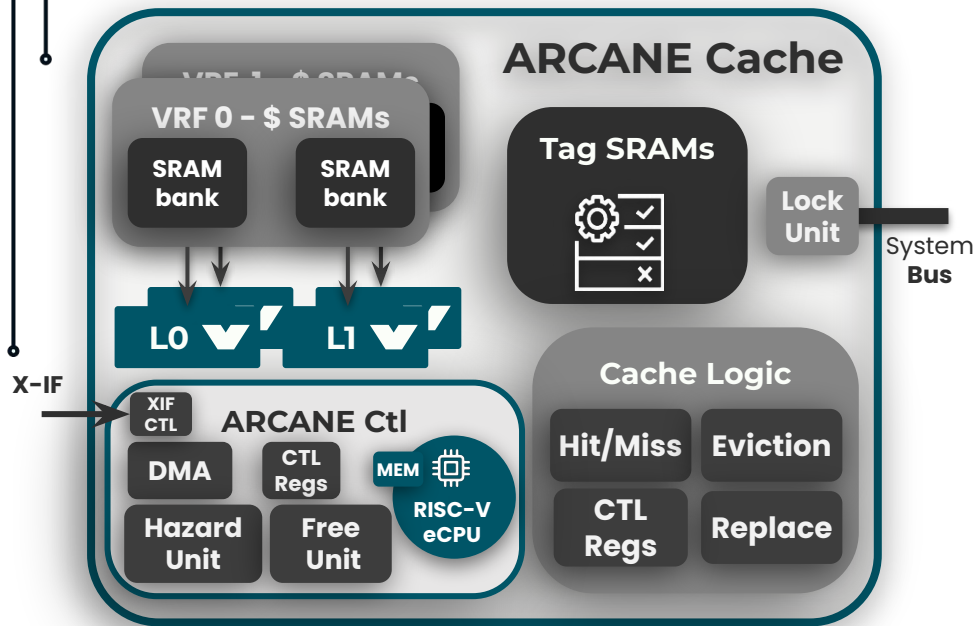
10x area efficient
14x energy efficient

MatMul



Up to 50x speed up
on CVA6

Conclusions



Drop-in replacement of a cache

Flexibility and Programmability

Scalable - from embedded to application-level systems

Demo

TinyStories - ARCANE decode

A small format language model continuing your prompt. Every GEMM runs on the ARCANE accelerator, tokens stream in as they are generated.

there was a king

there was a king who lived in a big castle. He was very proud of his crown. He liked to wear it every day. But one day, he forgot to take it off before going to bed. He felt a little uncomfortable, but he was happy that he had his crown. He put the crown on a string that was floating on the floor. He started to spin around and around. He felt dizzy and happy. He laughed and smiled. He spun faster and faster. But he did not see that his crown was on the floor. He bumped into it and lost his balance. He fell off the horse and hit his head. He felt a lot of pain. He cried and cried. He felt weak and sorry. His wife came to him and picked him up. She took him to the bathroom and cleaned his wound. She kissed it and put a bandage on his head. She said, "It's okay, the cut will heal soon. You are very brave and strong. You should be proud of yourself. And you should not forget your crown."

The king was surprised and proud. He said, "Thank you, queen. You are very kind and... truncated"

1.15 tok/s avg - 252 tok

One day, Lily met a Shoggoth

One day, Lily met a Shoggoth in the park. Lily was 3 years old and so excited to see it.

Lily asked the Shoggoth, "What do you eat?"

"I eat yummy food," he replied.

Lily asked again, "Can you bring me some yummy food?"

The Shoggoth shook his head and said, "No, I'm afraid I can't."

But then the Shoggoth started to move around and he said, "Wait! I know something that will bring you yummy food!"

Lily was so excited and asked, "What is it?"

The Shoggoth replied, "I will bring you some yummy ice cream!"

Lily was delighted and said, "Yay!" The Shoggoth was so kind and brought

8.85 tok/s avg - 373 tok

Once upon a time...

Model: TinyStories G2M Performance: Max 0.85 tok/s - 173 tok



TinyStories on llama2.c

DINOv2 ViT attention maps

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

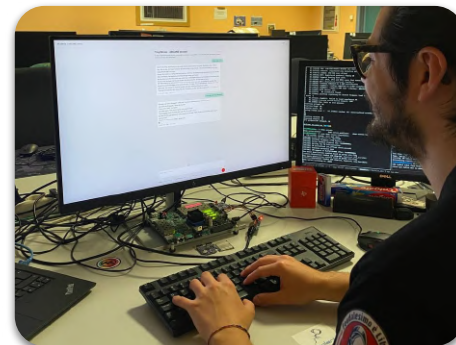
Visualizing the attention maps of the DINOv2 ViT model. The images show the model's focus on specific features in the input images.

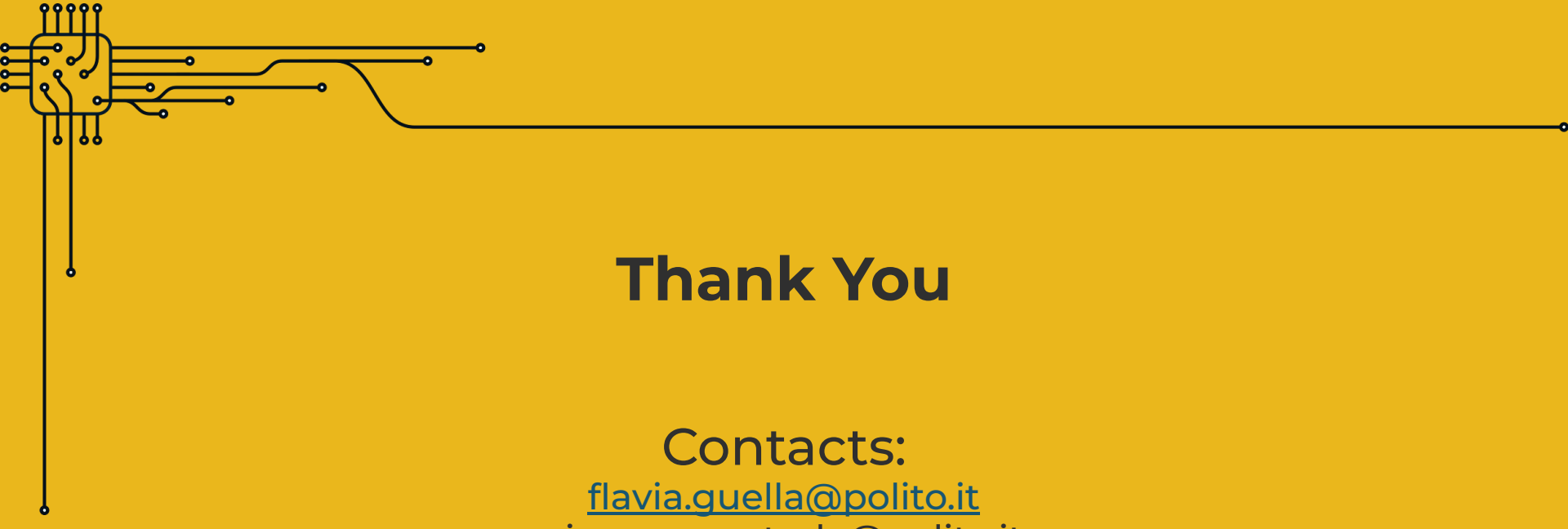
The hardware

Cheshire+ARCANE @ 80 MHz



Check out our DEMOs!





Thank You

Contacts:

flavia.guella@polito.it
vincenzo.petrolo@polito.it



Politecnico
di Torino



EPFL

RISC-V Summit Europe 2026,
Bologna, Italy

