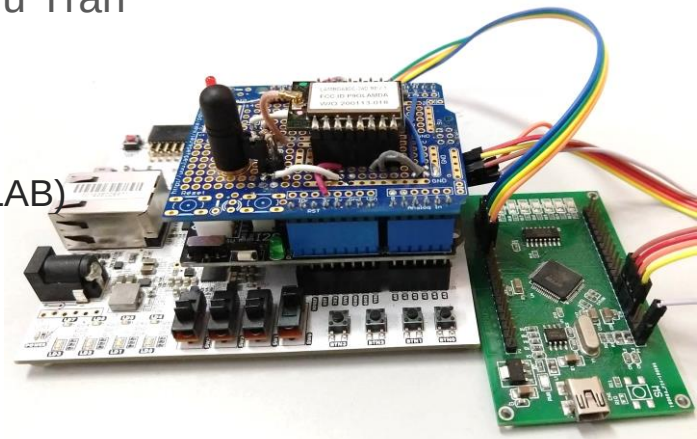
VIETNAM NATIONAL UNIVERSITY HANOI (VNU)
Information Technology Institute

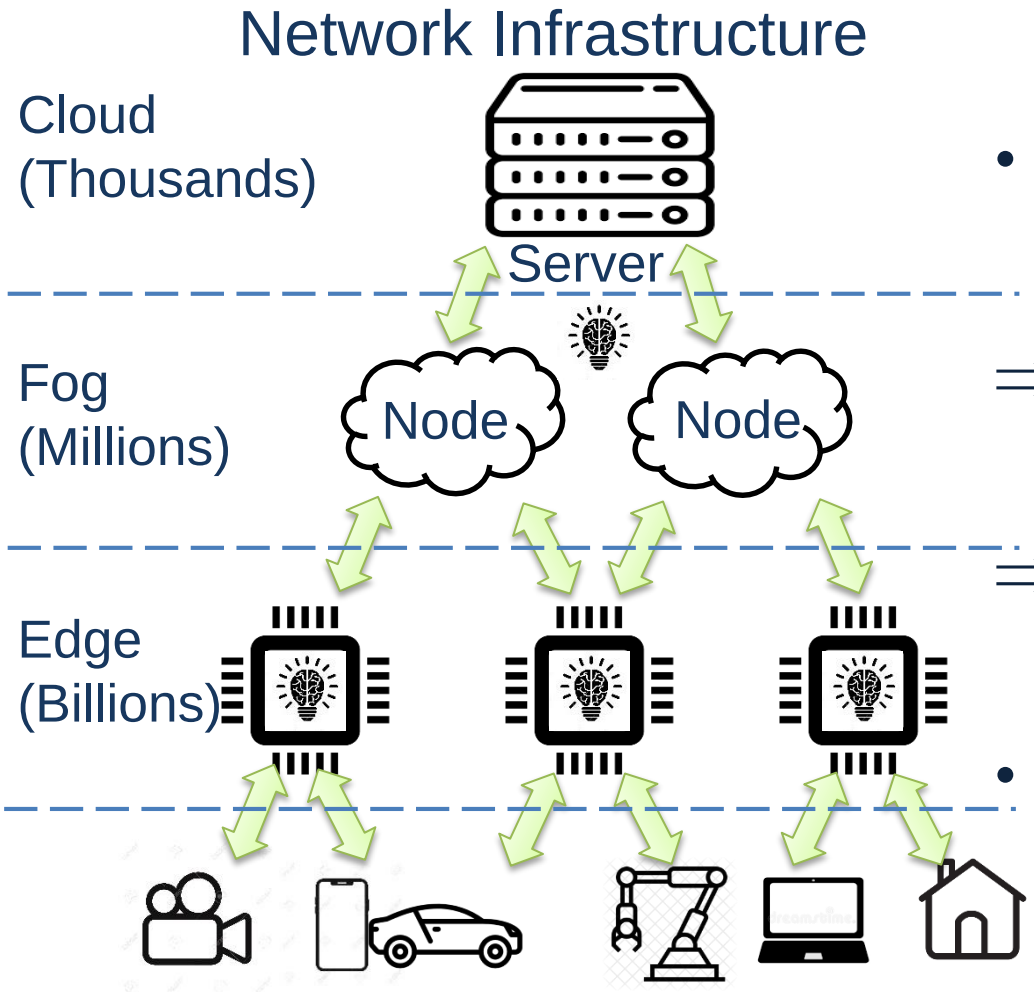
A Tiny Neural Network System base on RISC-V Processor

Ngo-Doanh Nguyen, Duy-Hieu Bui, Xuan-Tu Tran

Laboratory for Smart Integrated System (SISLAB)
Information Technology Institute (ITI)
Vietnam National University, Hanoi (VNU)
Website: http://www.iti.vnu.edu.vn



Artificial Intelligent of Things



- Billions of devices provide a huge numerous of information.

⇒ *Centralized processing methods are not suitable.*

⇒ *Decentralized processing methods are necessary.*

- AI methods provide computational power for decentralization.

⇒ ***AI for IoT is needed in decentralized processing!***

Outline

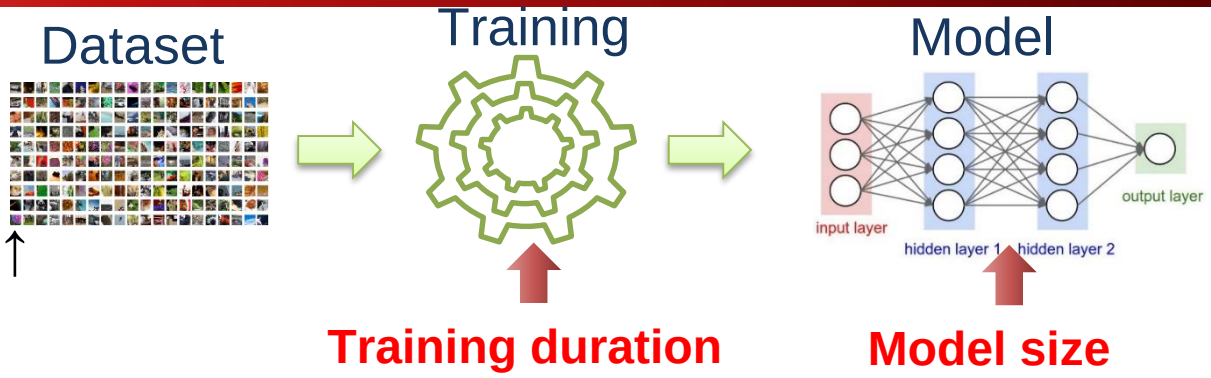
1. Motivation
2. Challenges & Solutions
3. HW Architecture for AIoT
4. ANN IP under PULPino Platform
5. CNN IP under Chipyard Platform
6. Conclusion

Outline

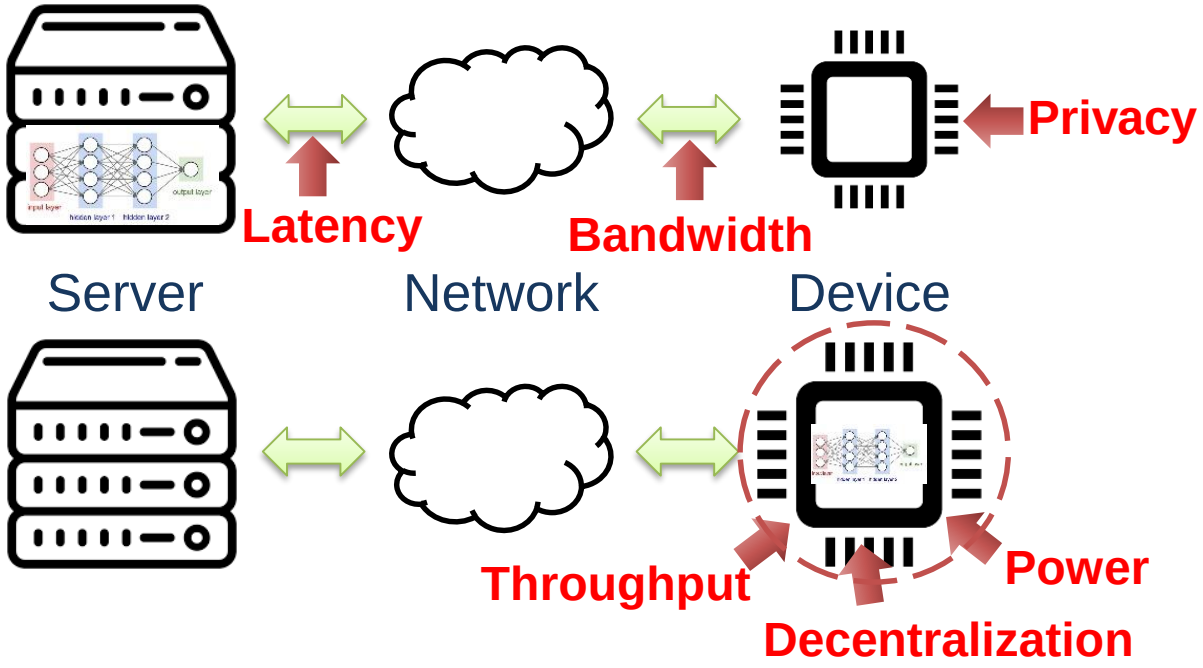
1. Motivation
2. Challenges & Solutions
3. HW Architecture for AIoT
4. ANN IP under PULPino Platform
5. CNN IP under Chipyard Platform
6. Conclusion

Challenges for AIoT

- **AI algorithm:**
  - *Complexity ↑, accuracy ↑*



- **Cloud computing:**
  - *High calculability*
  - *High resource*



- **Edge computing:**
  - *Low latency*
  - *Low cost*

⇒ ***A Tiny Neural Network System is a solution for edge devices!***

RISC-V & Opensource Ecosystem Opportunity

- Ready to use SoC platforms

Pulp-platform

- RISC-V processors: Rocket chip, Arian
- Simulator: Validator, Firesim (FPGA)
- Configurable IPs: SHA-3, testchip IPs
- Interconnect: AXI4, Tilelink

- Libraries
  - BSG BaseJump STL: clk_gen, async_fifo, synchronizers, Front-Side Bus, Network-on-chip IPs, etc.
  - Chips Alliance

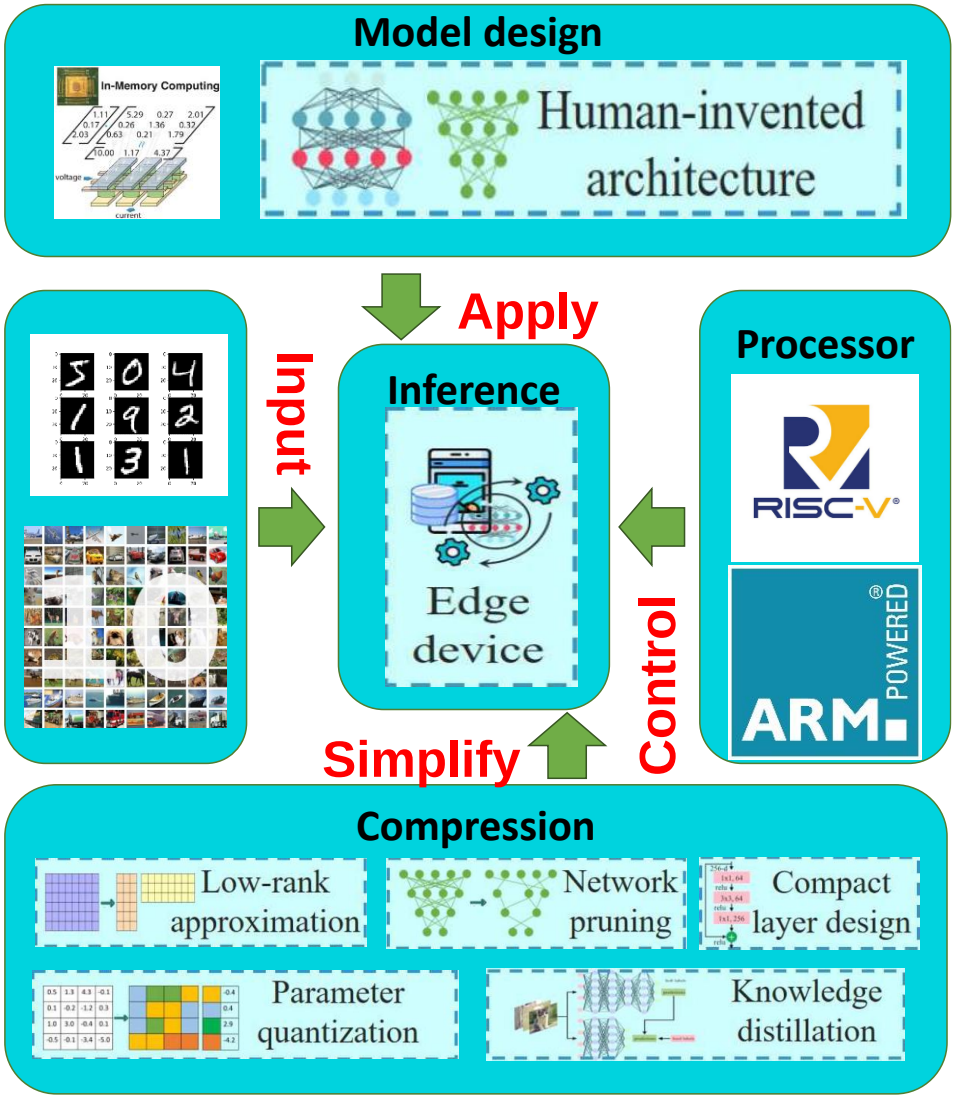
✓ **An opportunity for fast HW implementation at System level**

Solutions for Edge Computing

- **AI for constrained devices:**
  - *Low power*
  - *Lightweight computation*
  - *Small memory footprint*

⇒ **Proposal solutions:**

  - *Modified NN algorithms*
  - *Lightweight computational component (MAC, Pooling,…)*
  - *System Integration with low-power **RISC-V processors***
  - *Lightweight DMA design*

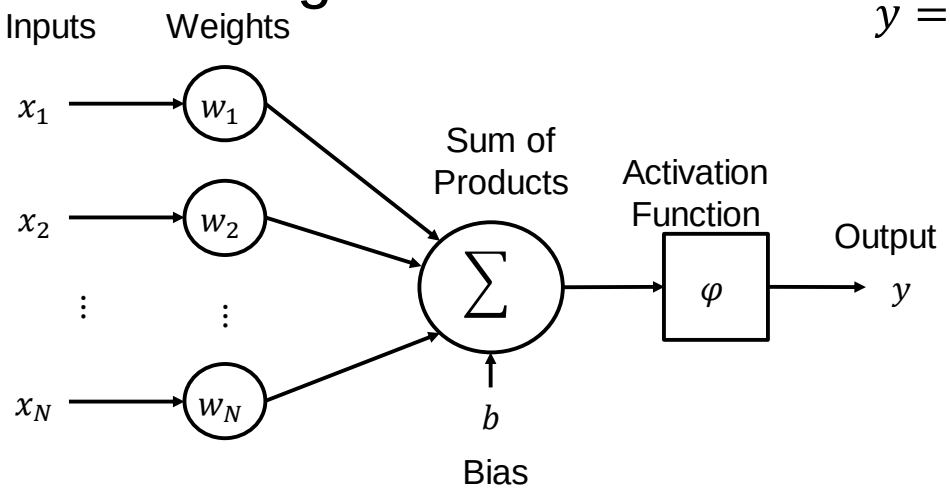


Outline

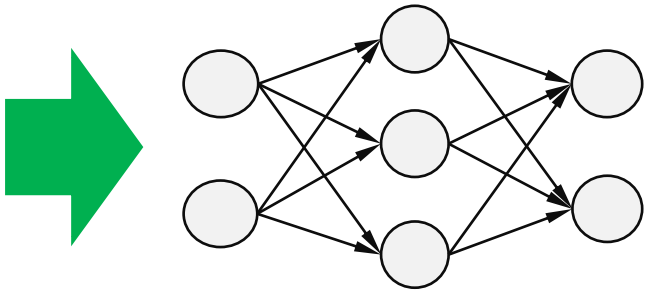
1. Motivation
2. Challenges & Solutions
3. HW Architecture for AIoT
4. ANN IP under PULPino Platform
5. CNN IP under Chipyard Platform
6. Conclusion

Neuron Network Model

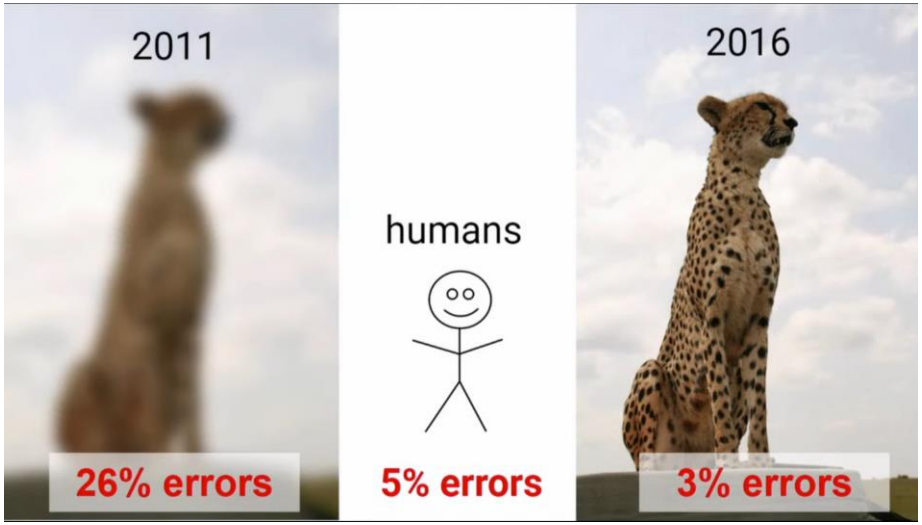
- Simulating human brains



$$y = \varphi\left(\sum_{i=1}^{N} x_i \times w_i + b\right)$$



Deep Neural Network

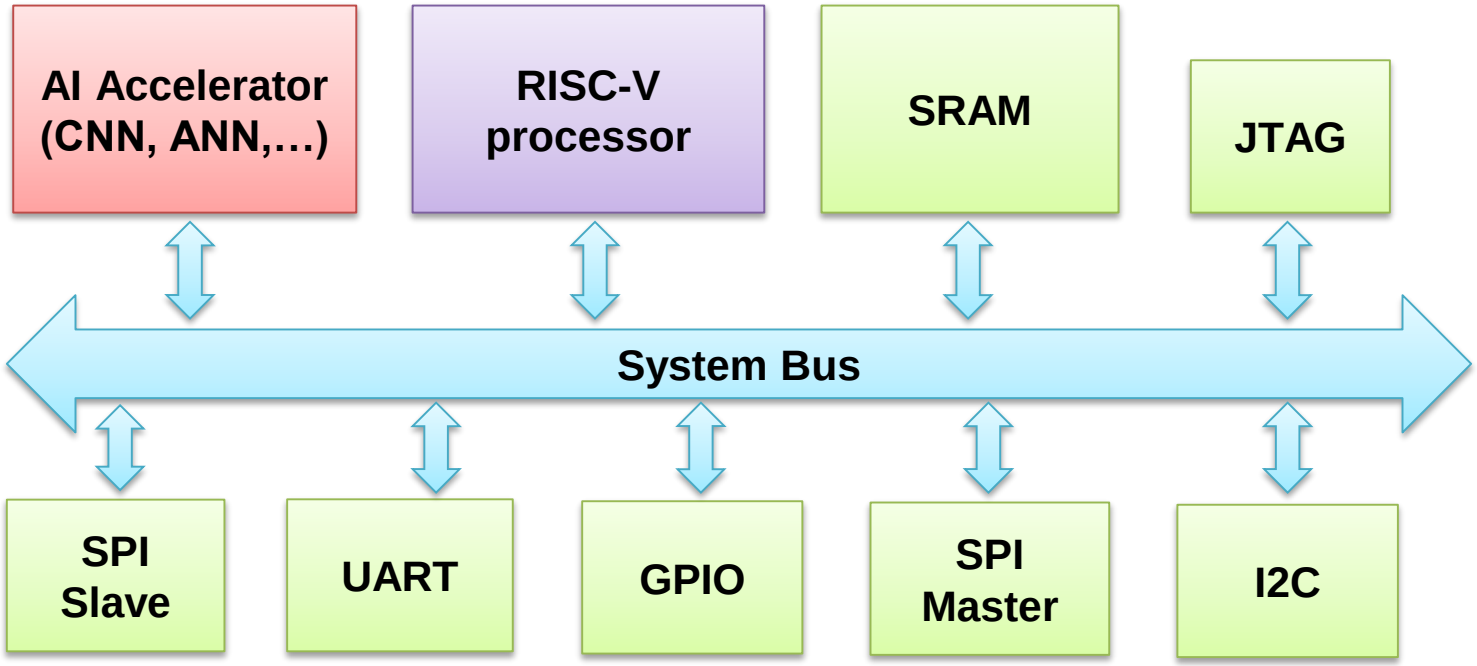


J. Dean, ISSCC2020

- ImageNet challenge
  - Classification error: 3%
  - Deep ↑, Accuracy ↑
  - CNN is one of best algorithms

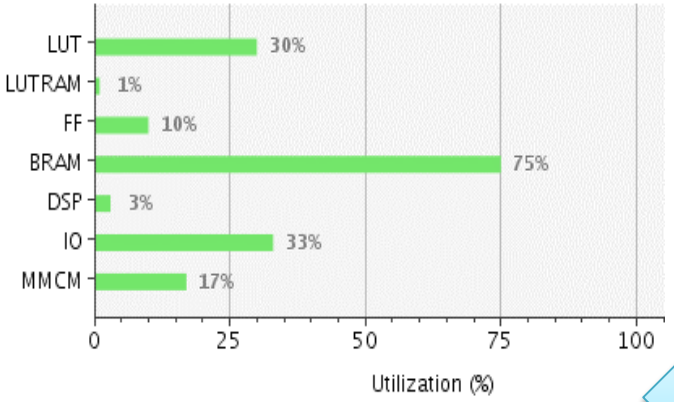
Overall Architecture for AIoT

❖ **A tiny AI accelerator connected with a RISC-V for low-cost low-power**



- AI Accelerator for high computational tasks
- Low power RISC-V core for configuration, control and data acquisition
- Communication interface (SPI, UART, I2C, …) for peripheral devices

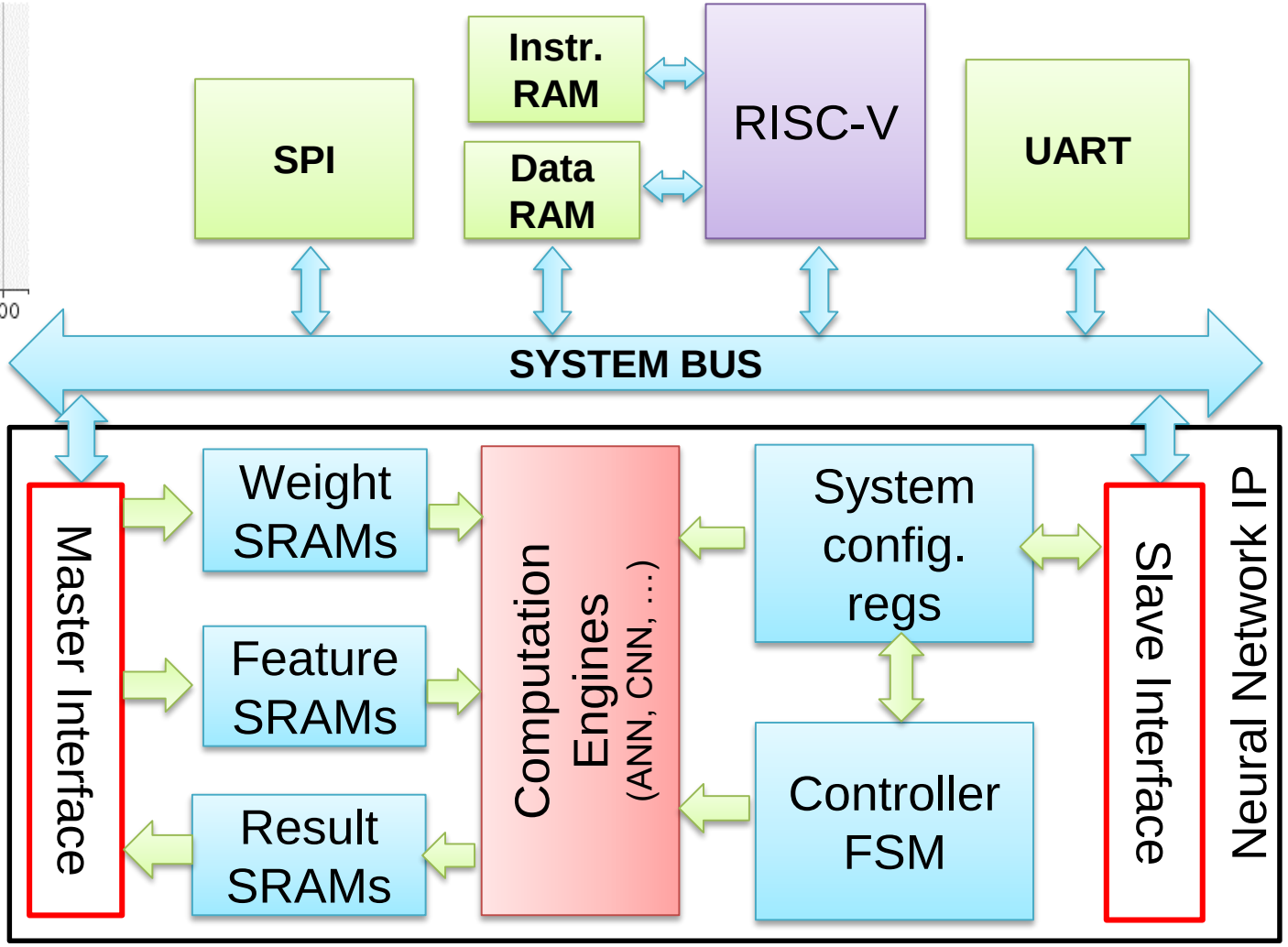
A Tiny Neural Network Accelerator



HW utilization for ANN core with VBP MAC on Arty-A7-100T

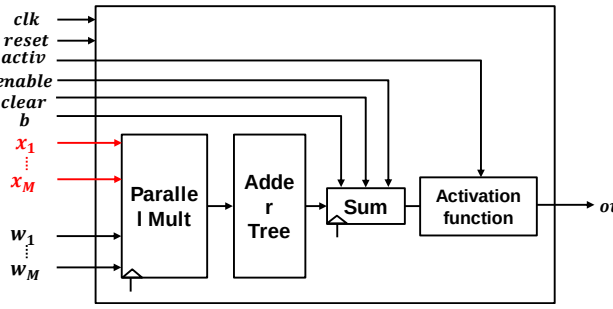
Different MACs:

- *Normal MAC*
- *Stochastic MAC*
- *Variable-bit-precision MAC*

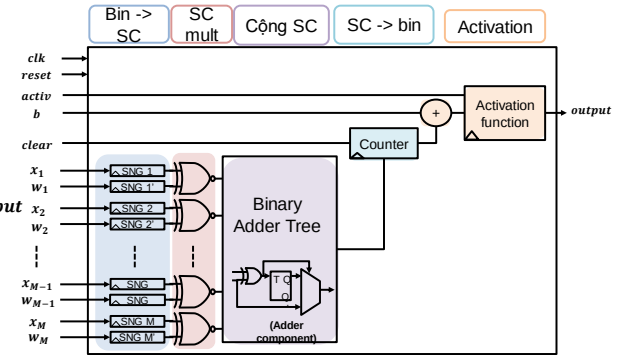


1. Nguyen *et al.*, 'An Efficient Hardware Implementation of Artificial Neural Network based on Stochastic Computing', NICS'18
2. Tran *et al.*, 'A Variable Precision Approach for Deep Neural Networks', ATC'19

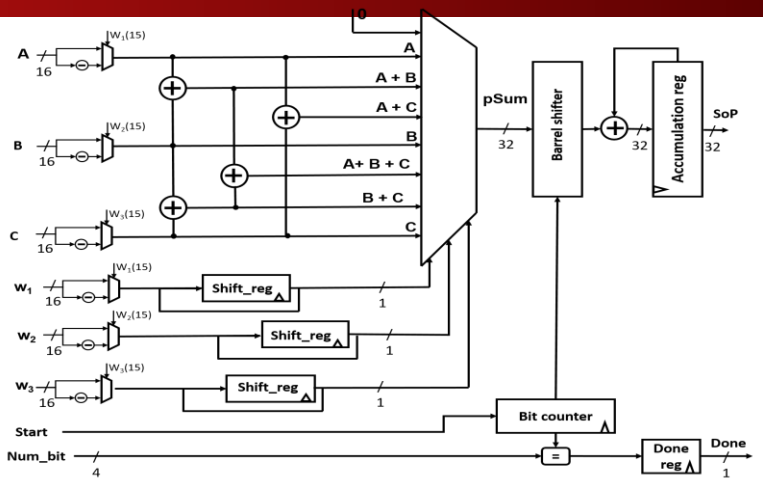
Multiple-Precision MACs



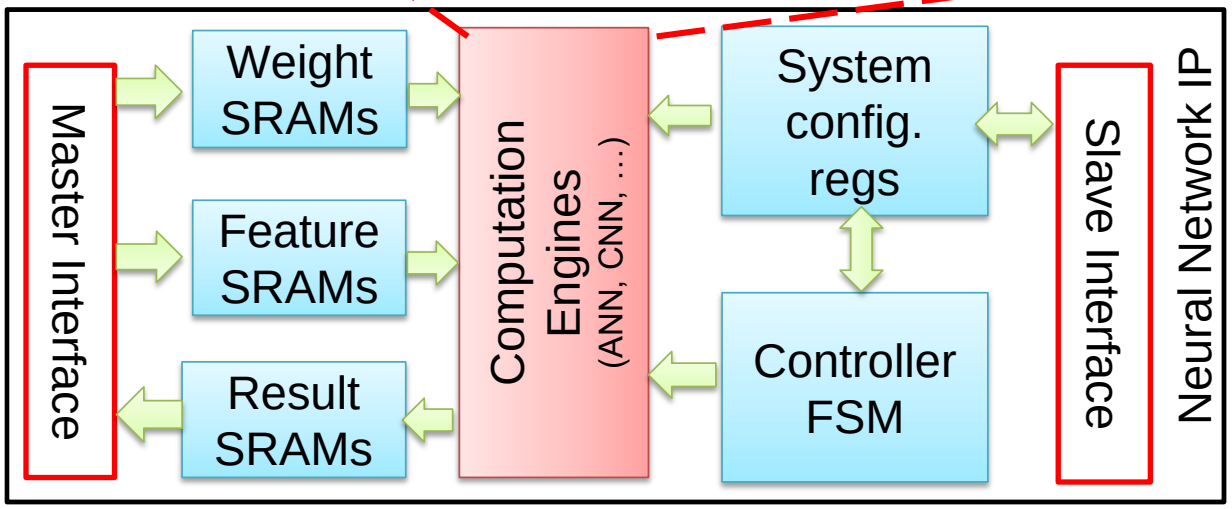
Normal MAC



Stochastic MAC



Variable bit precision MAC



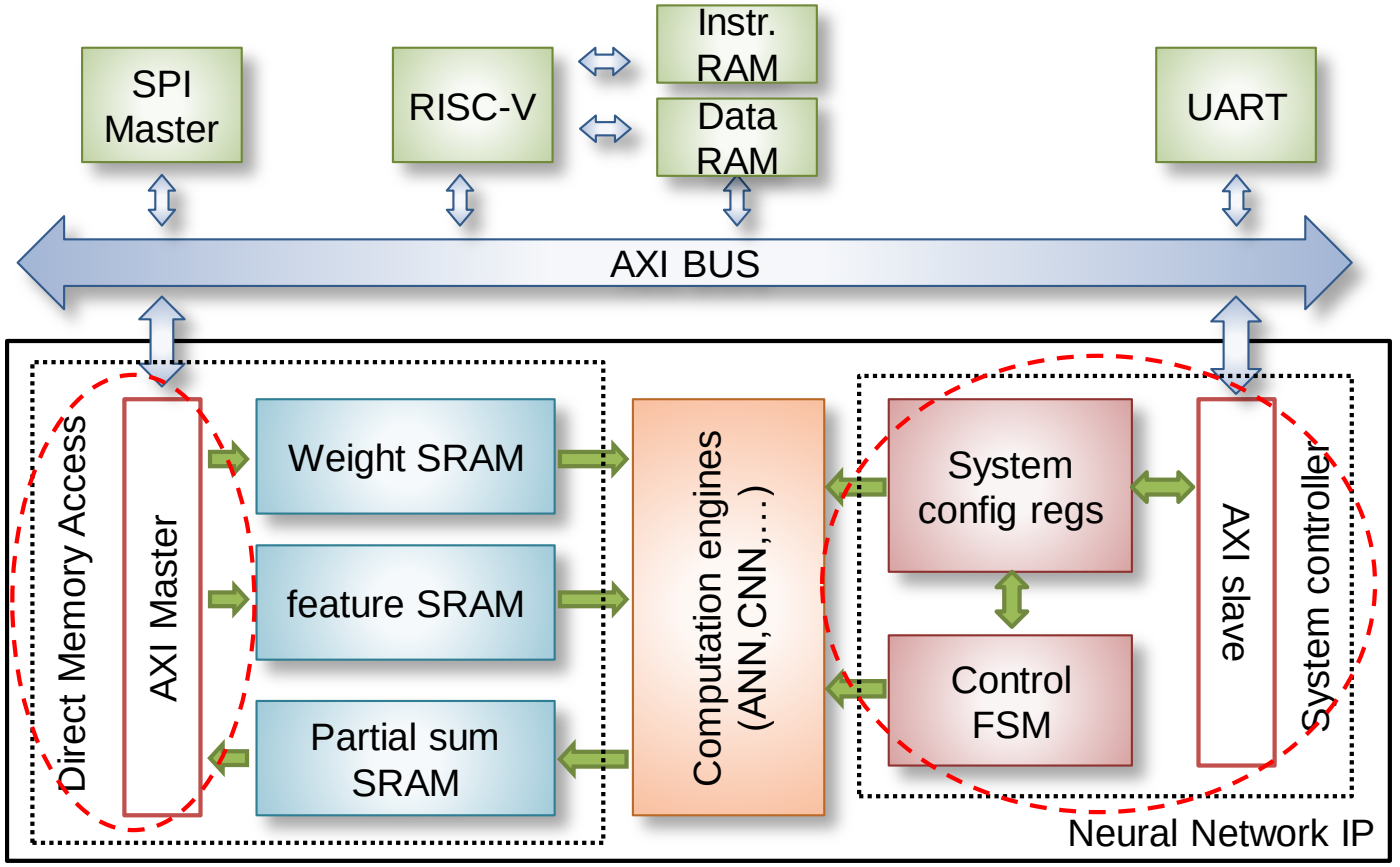
- Configure type of MAC via MMIO Registers
- Parallel computation = #MAC

Outline

1. Motivation
2. Challenges & Solutions
3. HW Architecture for AIoT
4. ANN IP under PULPino Platform
5. CNN IP under Chipyard Platform
6. Conclusion

ANN IP under PULPino platform

✓ PULPino uses RISC-Y as core processor and AXI4/AHB as bus interface

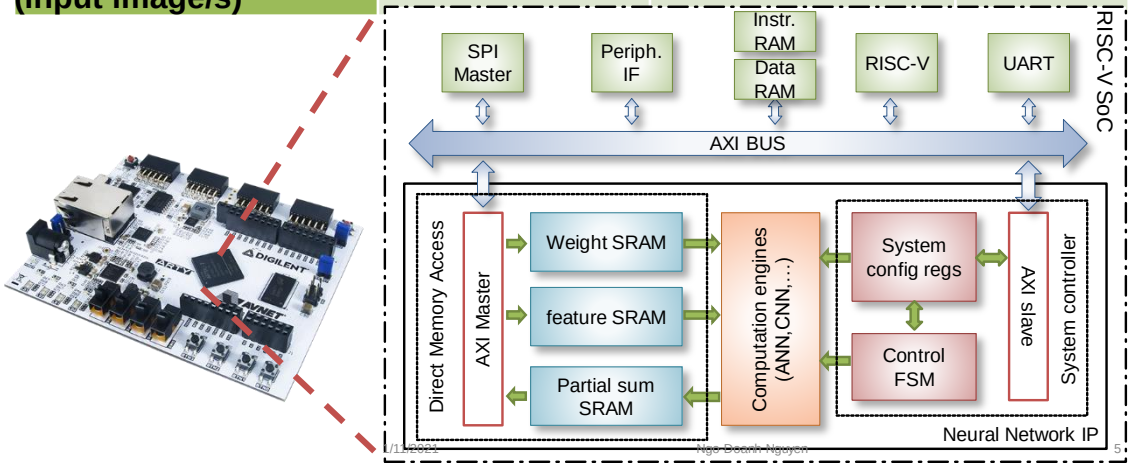


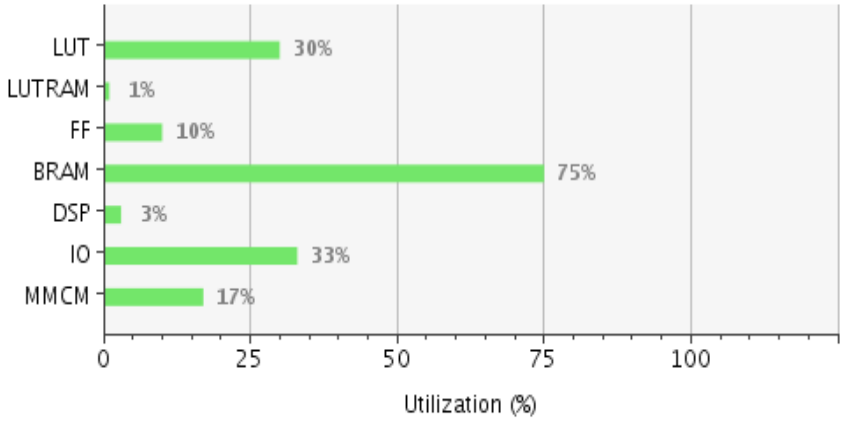
⇒ **Slave: CPU writes configurations and reads core status**

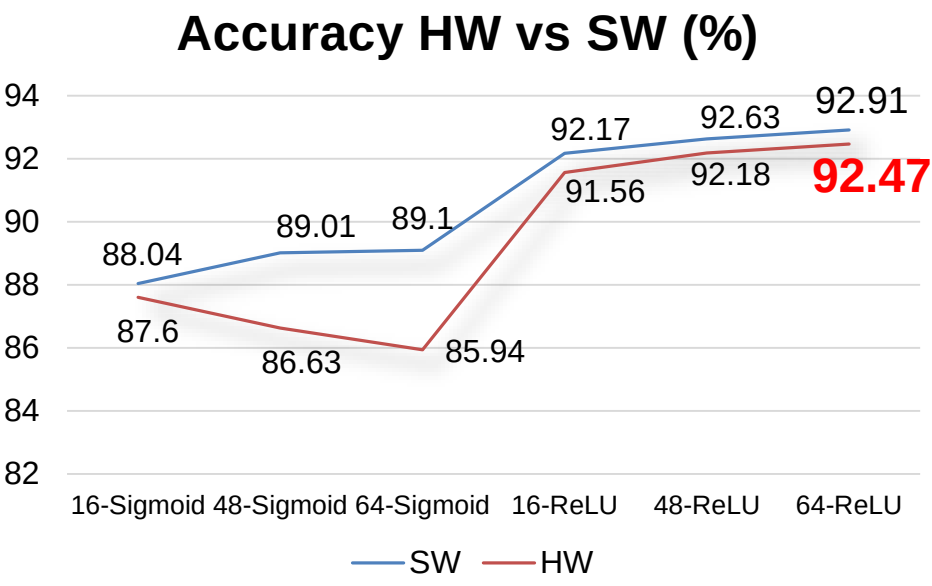
⇒ **Master + DMA: start data transfers from/to memory**

ANN IP Implementation Results on PULPino

	PULPino + ANN IP core	ANN IP core with Interface	Percentage
Frequency (MHz)	25	25	-
Slice LUTs	18873	2782	14.7%
Slice registers	13082	2745	21%
Block RAM	101.5	37.5	36.9%
DSP	8	0	0%
#cycles to load weight	-	25,811	-
#cycles to load image	-	400	-
#cycles to load bias	-	78	-
Throughput (input image/s)	-	584 (28x28)	-





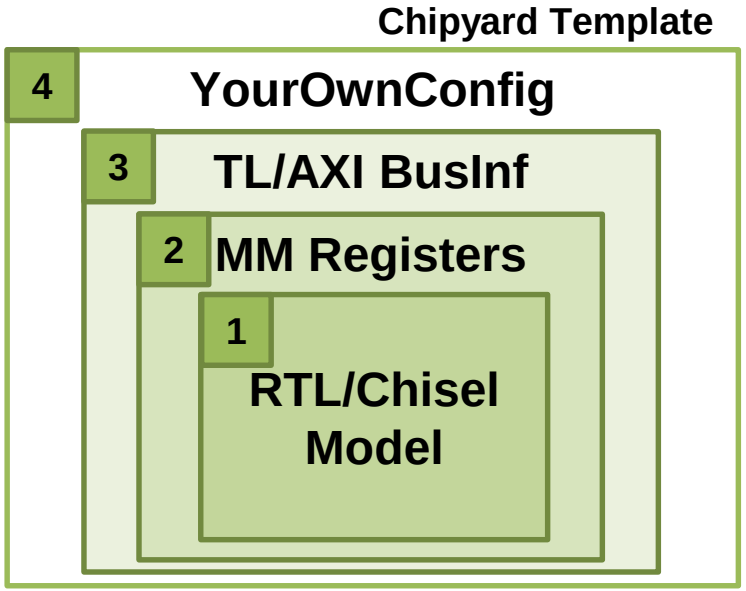


Outline

1. Motivation
2. Challenges & Solutions
3. HW Architecture for AIoT
4. ANN IP under PULPino Platform
5. CNN IP under Chipyard Platform
6. Conclusion

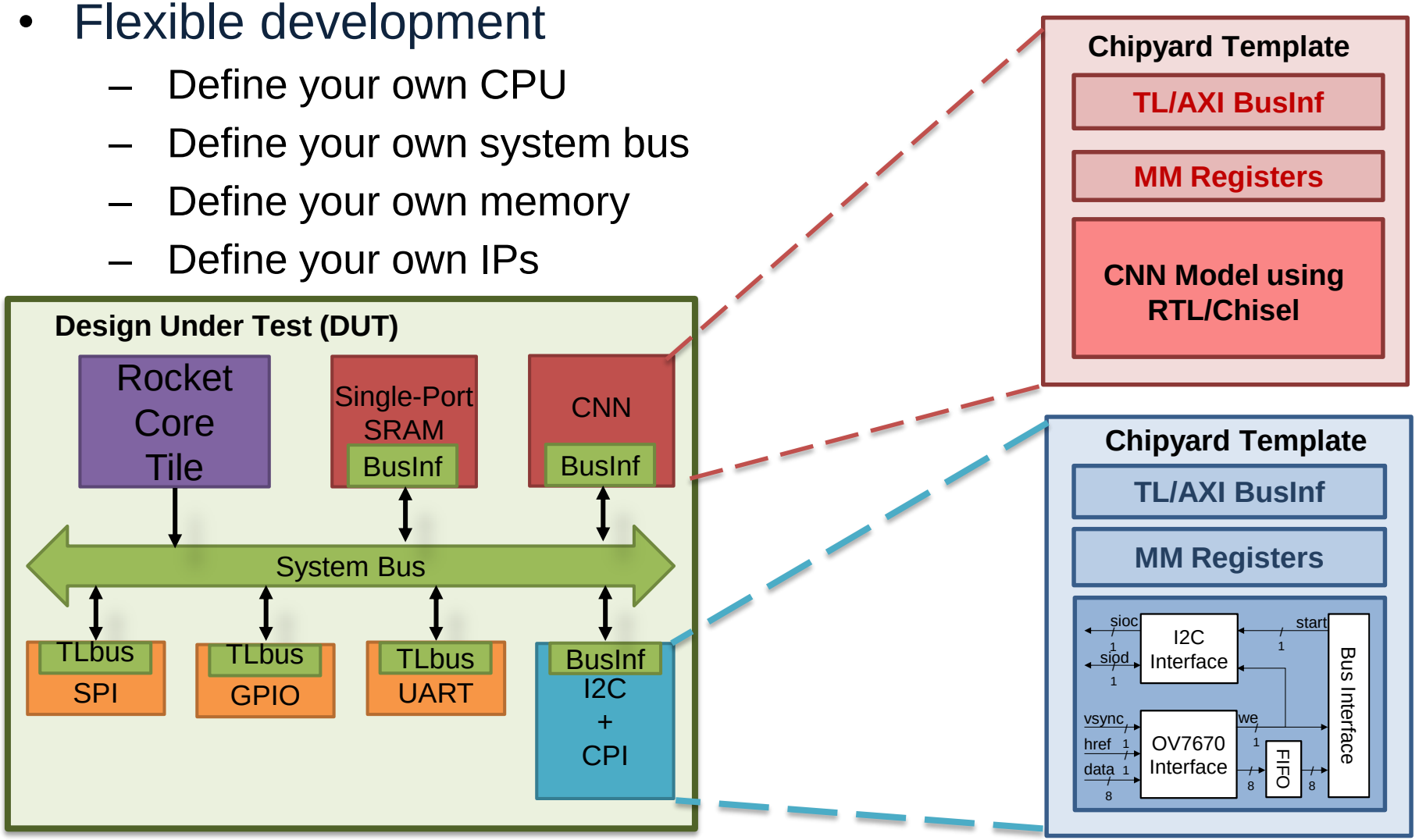
Chipyard and Integration Steps

- What is Chipyard?
  - A framework for designing and evaluating SoC
  - A collection of tools and libraries for developing SoC
- How to integrate your IPs into Chipyard?
  - IPs and Peripherals have specific addresses
  - MMIO is the easiest way to associate with RISC-V core
- Integration within 4 steps
  - Design your own module
  - Link your IPs with MM registers
  - Add your specified bus interface
  - Configure your module's params



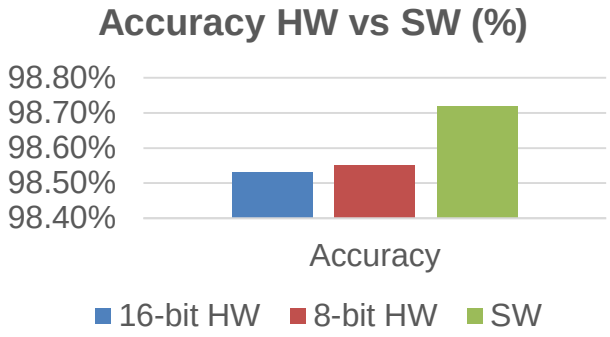
SoC with CNN IP under Chipyard

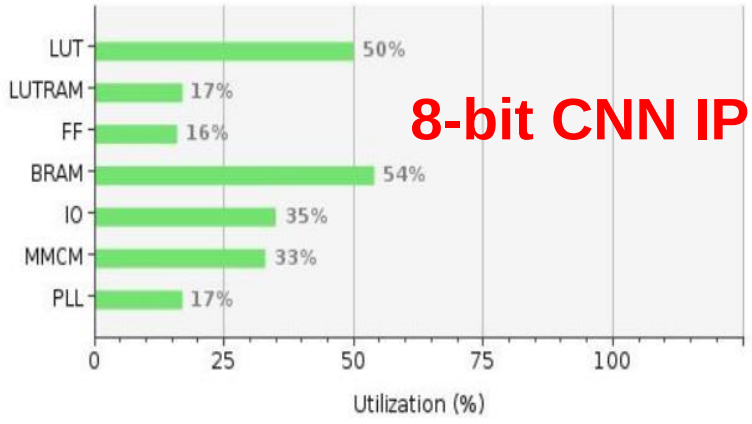
- Flexible development
  - Define your own CPU
  - Define your own system bus
  - Define your own memory
  - Define your own IPs

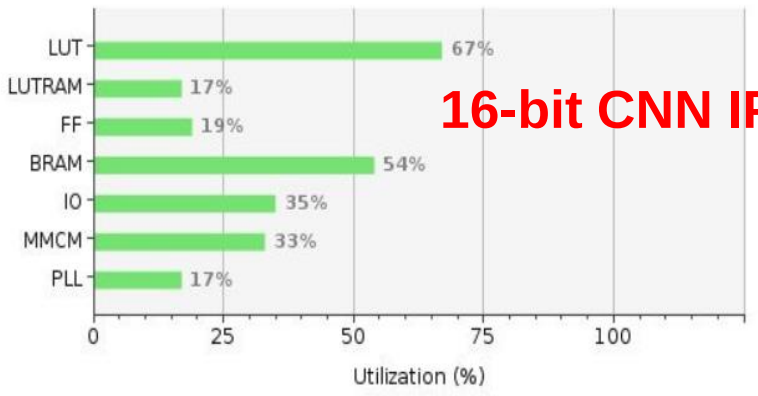


CNN IP Implementation Results on Chipyard

	8-bit SoC	8-bit CNN IP	16-bit SoC	16-bit CNN IP
Frequency (MHz)	50	50	50	50
Slice LUTs	31818 (50.19%)	7842 (12.37%)	42792 (67.50%)	18826 (29.69%)
Slice registers	19765 (15.59%)	4248 (3.35%)	24218 (19.10%)	8701 (6.86%)
Slice	9931 (62.66%)	2413 (15.22%)	12749 (80.44%)	5418 (34.18%)
Block RAM	72.5 (53.70%)	2.5 (1.85%)	72.5 (53.70%)	2.5 (1.85%)
DSP	0 (0%)	0 (0%)	0 (0%)	0 (0%)







Outline

1. Motivation
2. Challenges & Solutions
3. HW Architecture for AIoT
4. ANN IP under PULPino Platform
5. CNN IP under Chipyard Platform
6. Conclusion

Conclusion

- Open source hardware is maturing, especially for AI

Pulp-platform

- Our tiny Neural Network Systems have been demonstrated in those open source hardware.
  - ANN IP on PULPino platform
  - CNN IP on Chipyard platform
  - In future, more complex, more practical algorithms will be implemented (SNN, GCN, …)

- SISLAB is willing to support the community in Vietnam

VIETNAM NATIONAL UNIVERSITY HANOI (VNU)

Information Technology Institute

Information Technology Institute (ITI)
Vietnam National University, Hanoi (VNU)
Website: http://www.iti.vnu.edu.vn