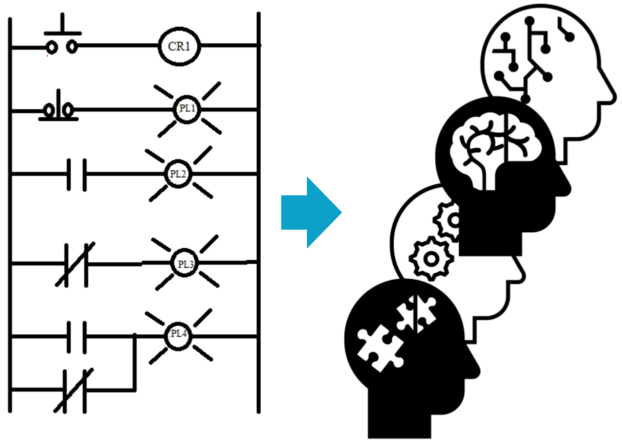
Industrial Knee-jerk: In-Network Simultaneous Planning and Control on a TSN Switch

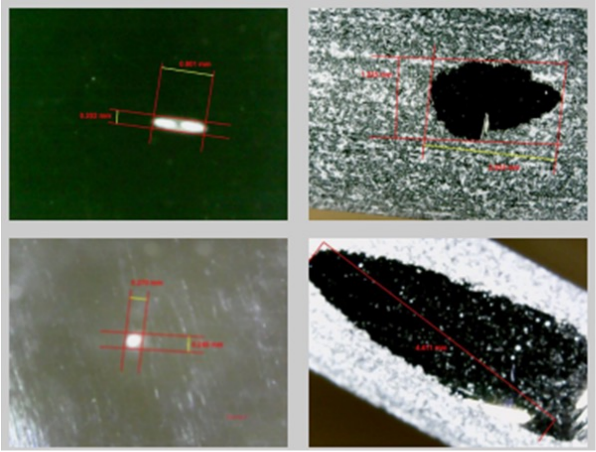
Zeyu Wang, Jingao Xu, Xu Wang,
Xiangwen Zhuge, Xiaowu He, Zheng Yang
Tsinghua University

Industry 4.0 and Autonomous Production

In the era of **Industry 4.0**, the intelligence of production lines and the autonomy of mechanical arms have gradually become highly prized goals for manufacturing factories.



Evolution of Mechanical Arms

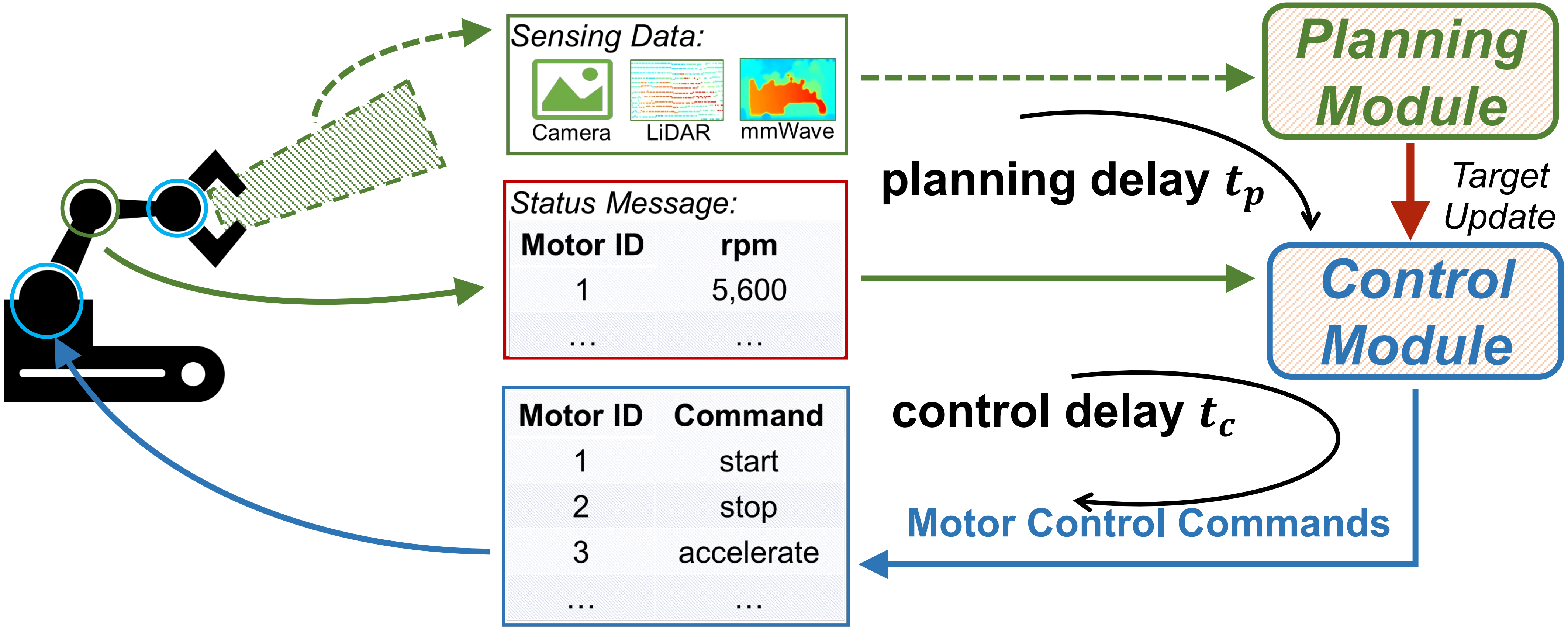


Defect Detection

Emergency braking

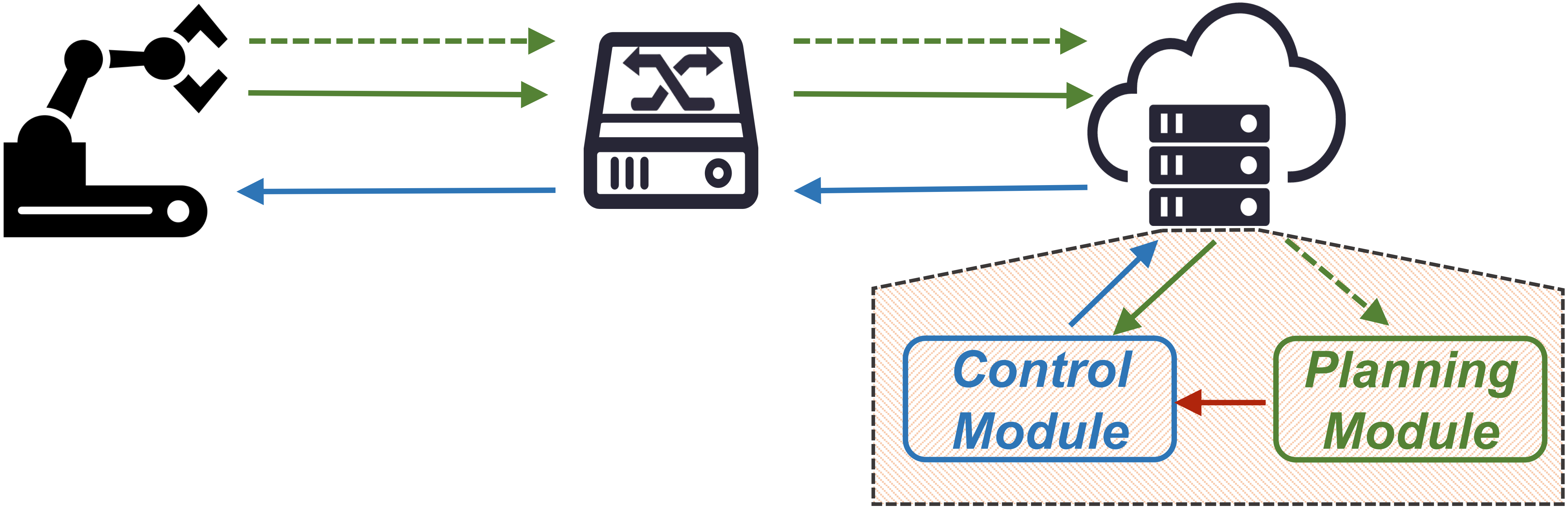
Decision Closed-Loop of a Mechanical Arm

A mechanical arm's intelligent decision closed-loop can be abstracted as **planning** and **control** two modules.



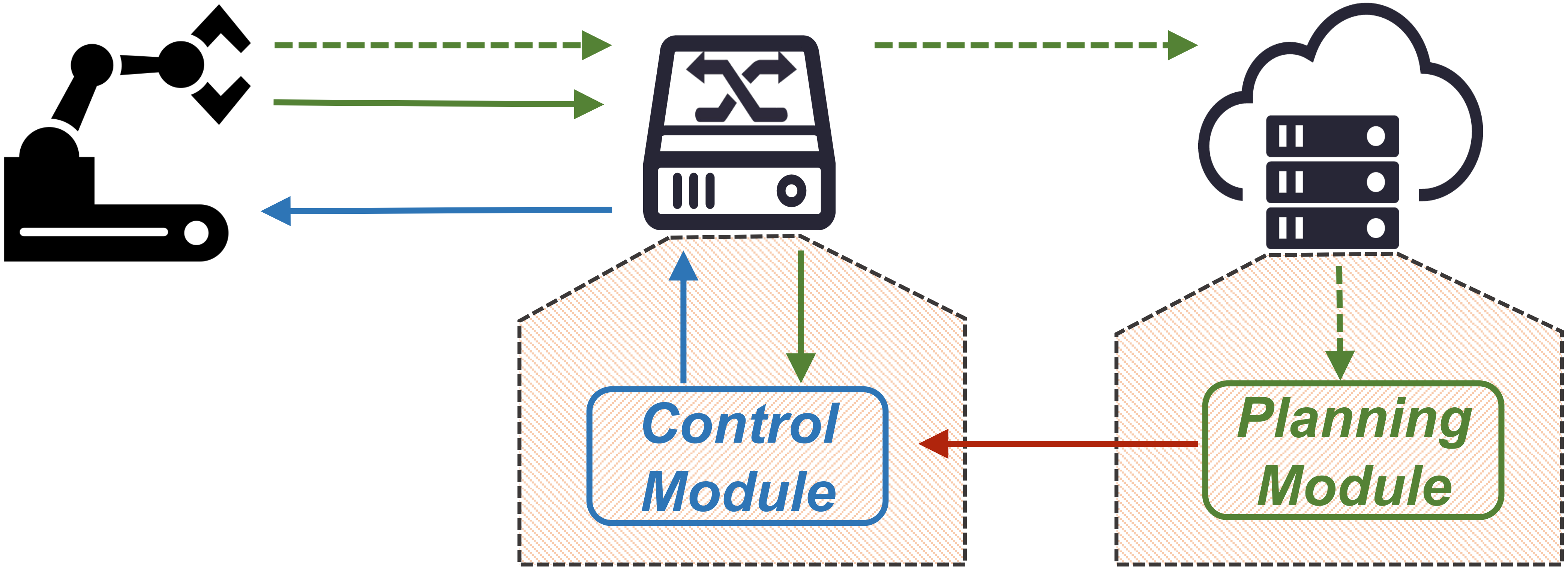
Existing Arm Control Solutions

Baseline-I: offload both the planning and control modules to a centralized **cloud/edge server**



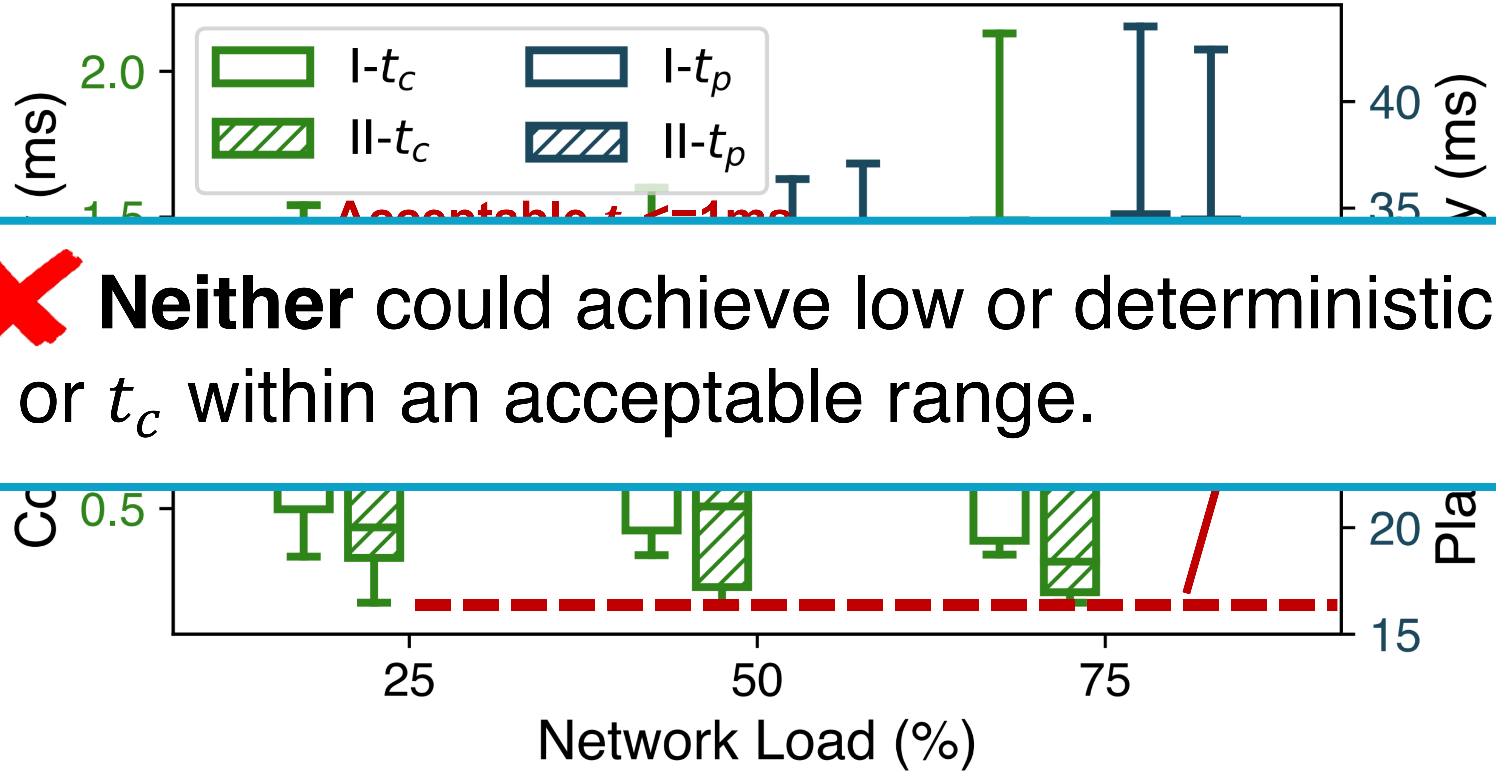
Existing Arm Control Solutions

Baseline-II: load the low-level control module onto a **network switch**, but keep the high-level yet complex planning module on **cloud/edge server**



Limitations of Current Control Solutions

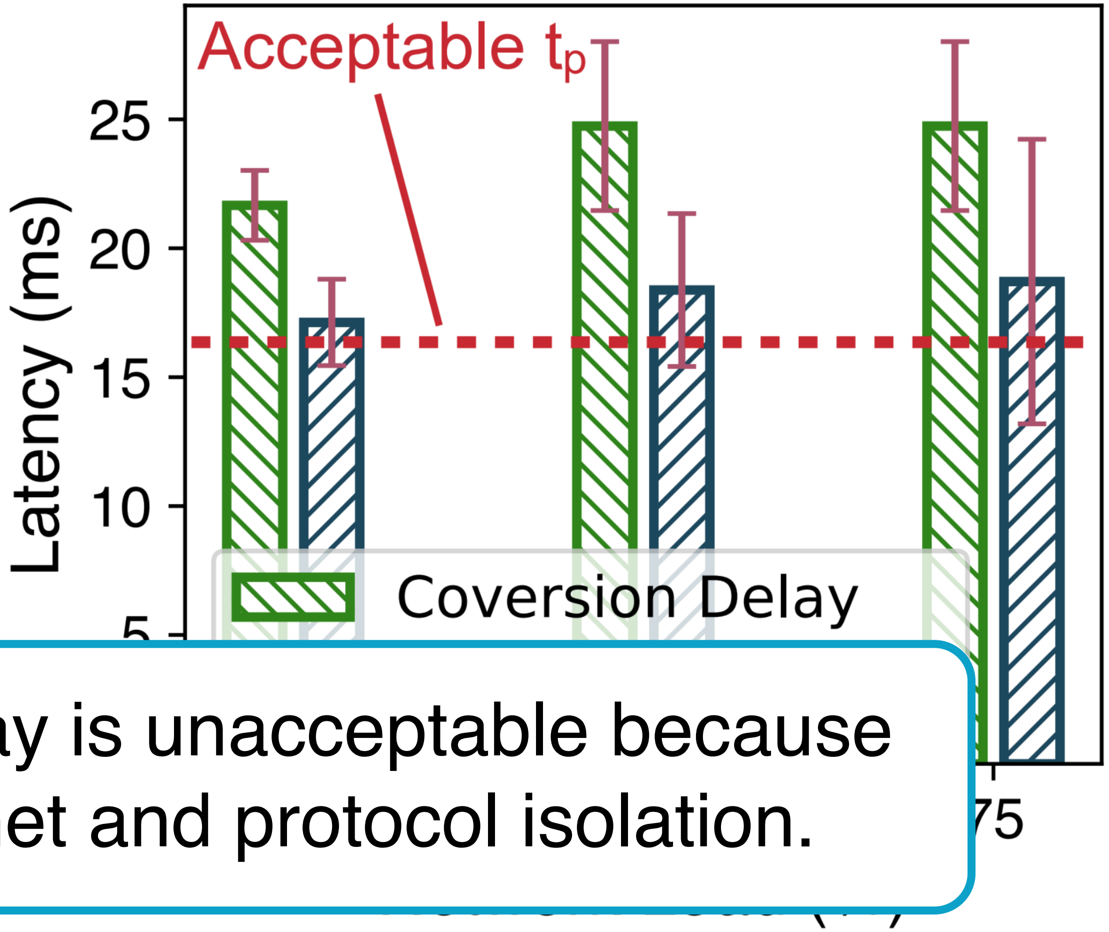
- We evaluate Baseline-I&II performance by conducting over 200 defective glass detection and grabbing tests.



❌ **Neither** could achieve low or deterministic t_p or t_c within an acceptable range.

Limitations of Current Control Solutions

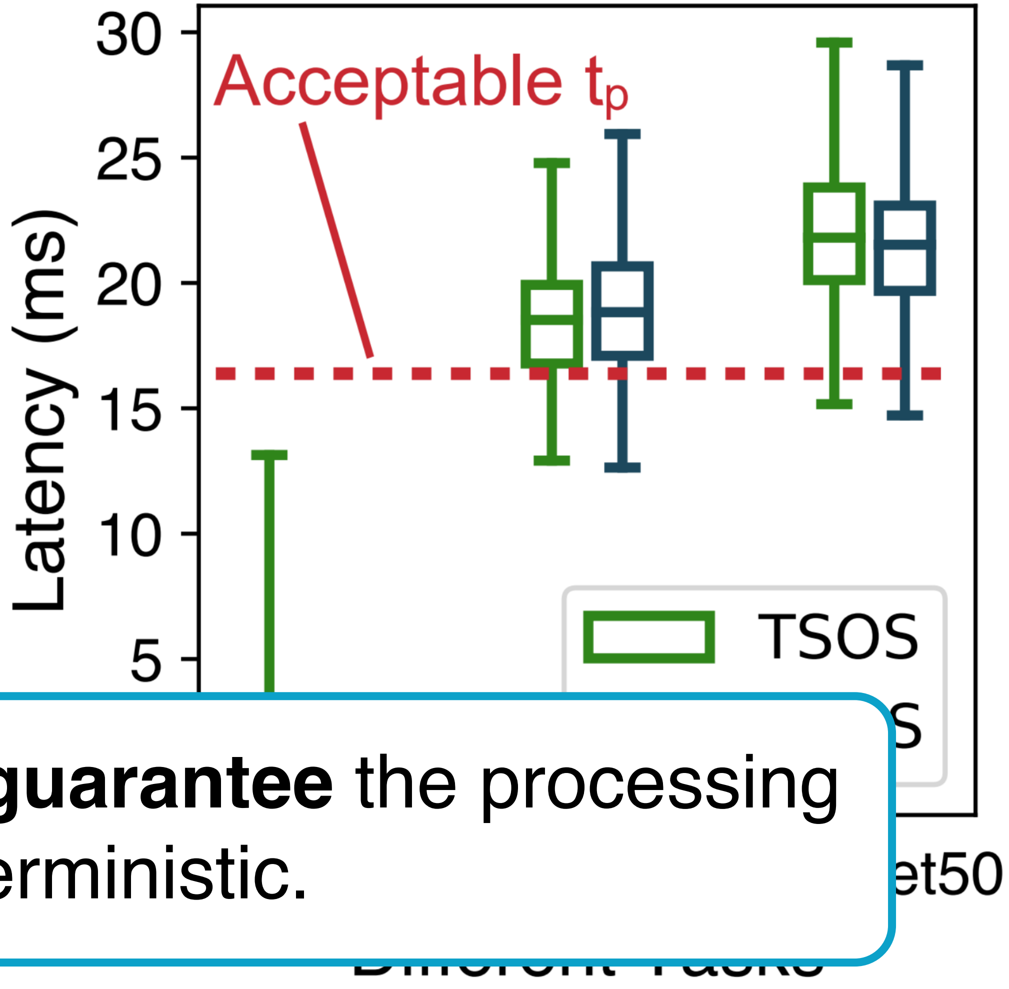
- **C1: Considerable data transmission delay**
  - Transmit frames to an edge server under different network loads
  - Transmission Delay: transmission delay on standard Ethernet
  - Conversion Delay: delay from PROFINET to standard Ethernet
  - –



❌ Data transmission delay is unacceptable because of unreliable standard Ethernet and protocol isolation.

Limitations of Current Control Solutions

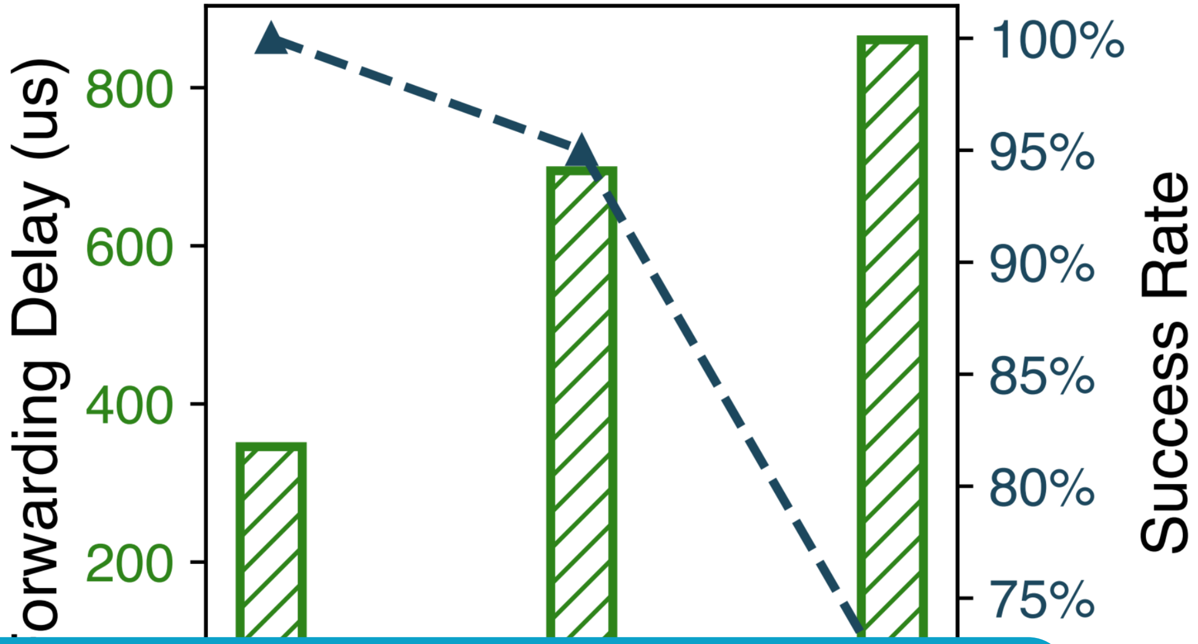
- **C2: Highly dynamic computation latency**
  - Set up an edge server running on a TSOS and RTOS, respectively
  - Run one trajectory planning(TP) algorithm on CPU
  - Run two network backbones for defect detection on GPU
  - determinism on GPU.



TSOS or RTOS **cannot guarantee** the processing latency of a specific task is deterministic.

Limitations of Current Control Solutions

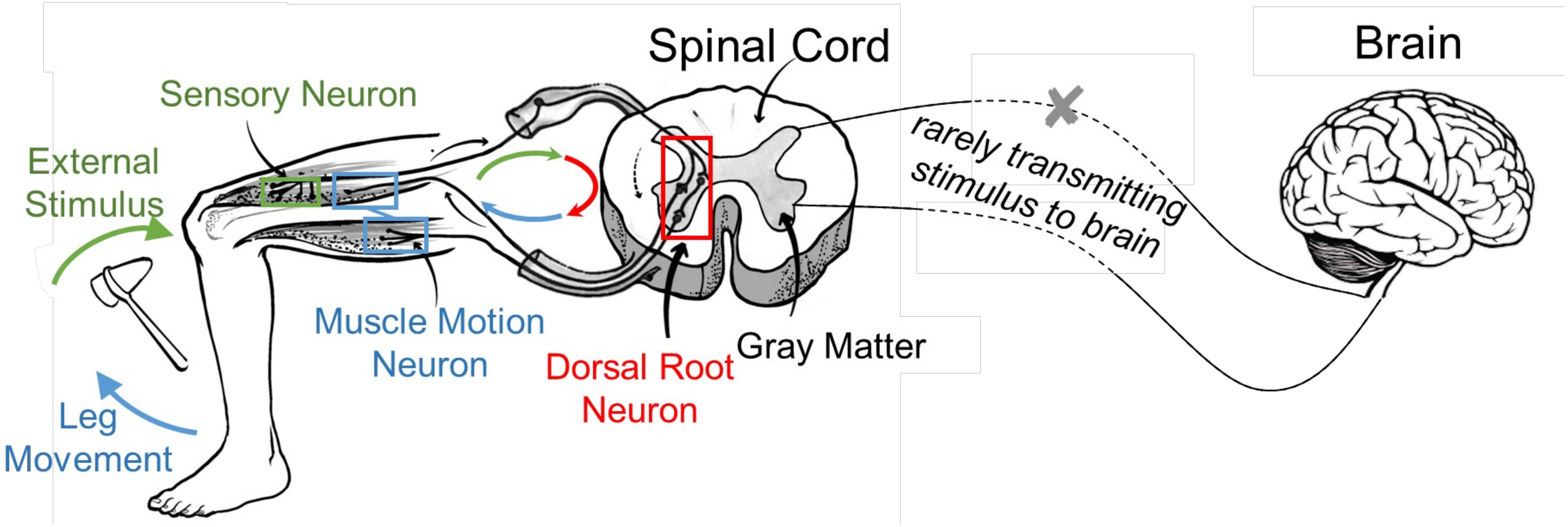
- **C3: Unreliable control packet forwarding**
  - We measure each control packet's forwarding delay and success rate under different network loads.
  - Results: when network loads > 75%, the forwarding success



Overloaded data flows result in **excessive queuing delays** and **inaccurate forwarding** of critical control packets.

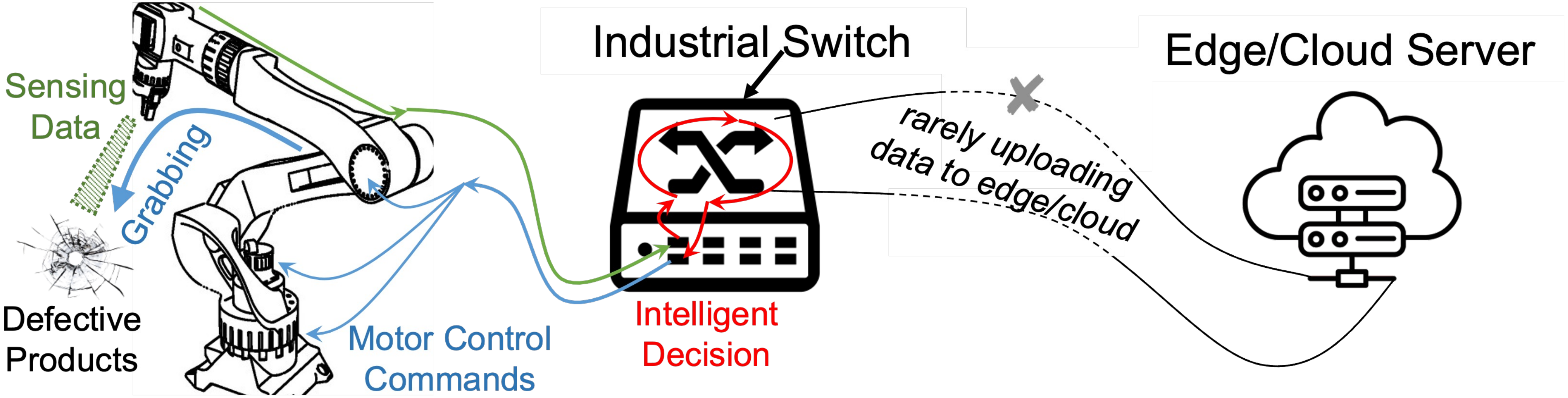
Key Insight

- **Offloading both urgent planning and control modules to the industrial switch**
  - Eliminate the data transmission delay and uncertainty.



Key Insight

- **Offloading both urgent planning and control modules to the industrial switch**
  - Eliminate the data transmission delay and uncertainty.

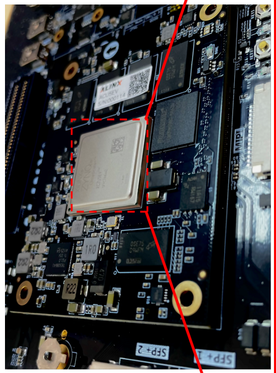


Key Insight

- **Offloading both urgent planning and control modules to the industrial switch**
  - Eliminate the data transmission delay and uncertainty.
- **Software and hardware co-design for task computing acceleration**
  - Bypass uncertain OS- or CPU-level resource allocation and task scheduling.
- **Reserving dedicated network bandwidth and time slots**
  - Avoid interference from background traffic.

Hierarchical Computing Platform

- The use of hierarchical computing devices, e.g., Xilinx Zynq, enables lightweight devices to conduct relatively complex tasks.



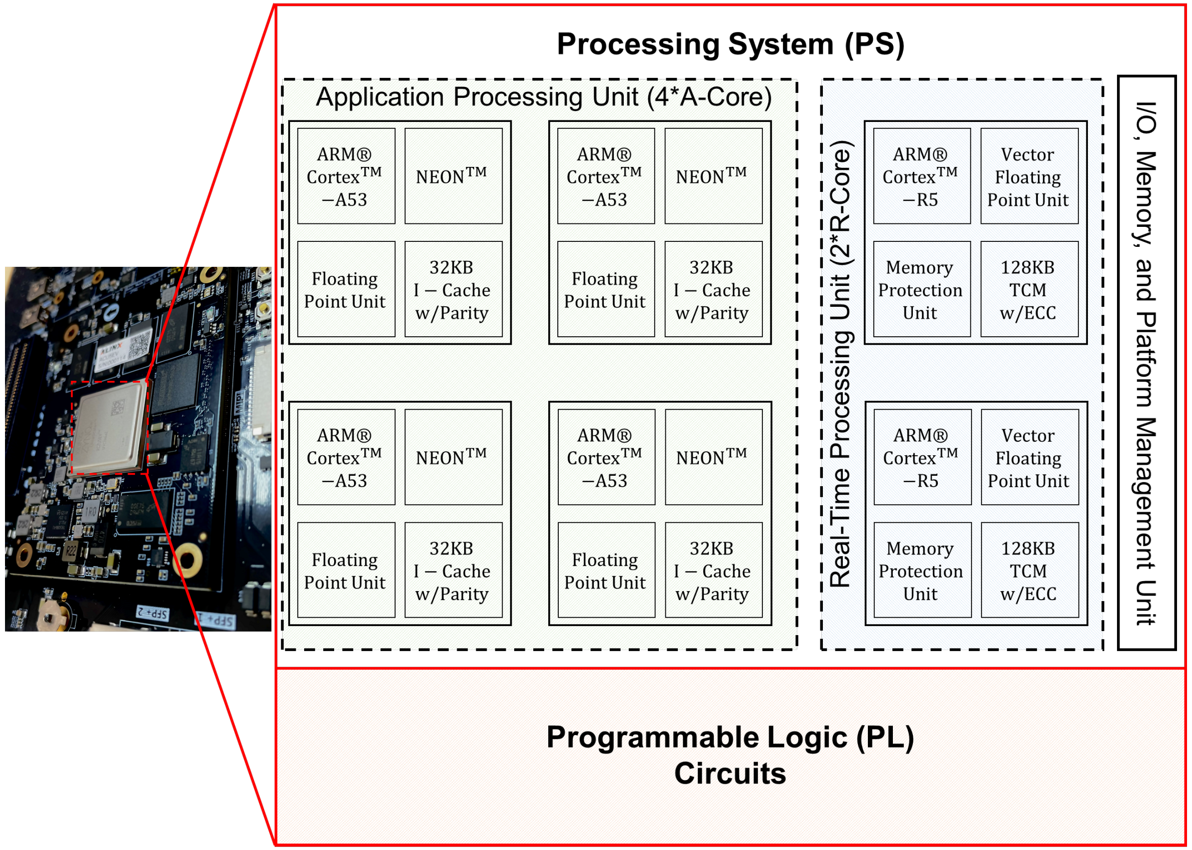
Processing System (PS)

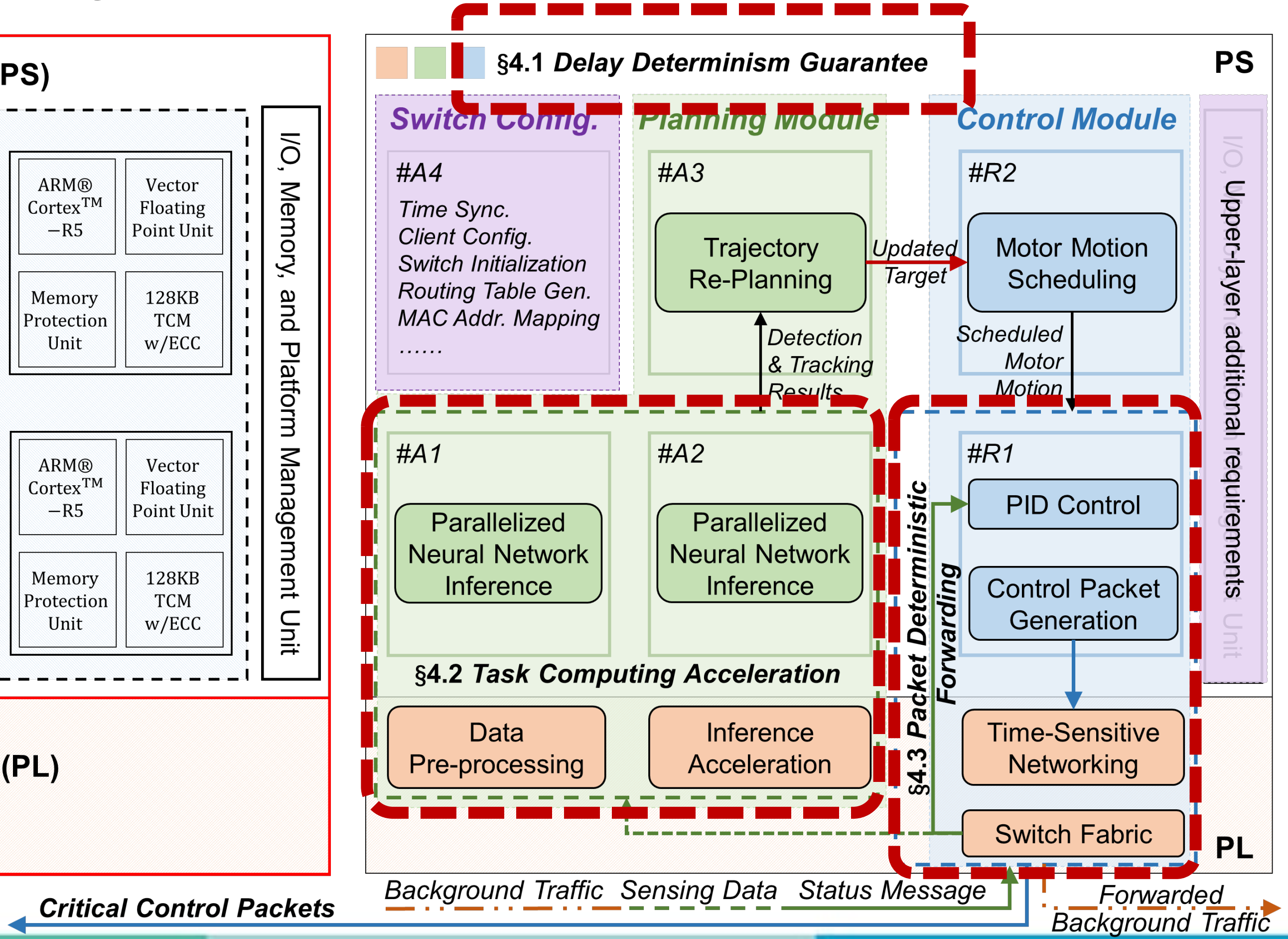
PS: quad-core Cortex-A53, for general application	PS: dual-core Cortex-R5, for real-time application

PL provides PLB, DSP, etc., for hardware design.

System Overview

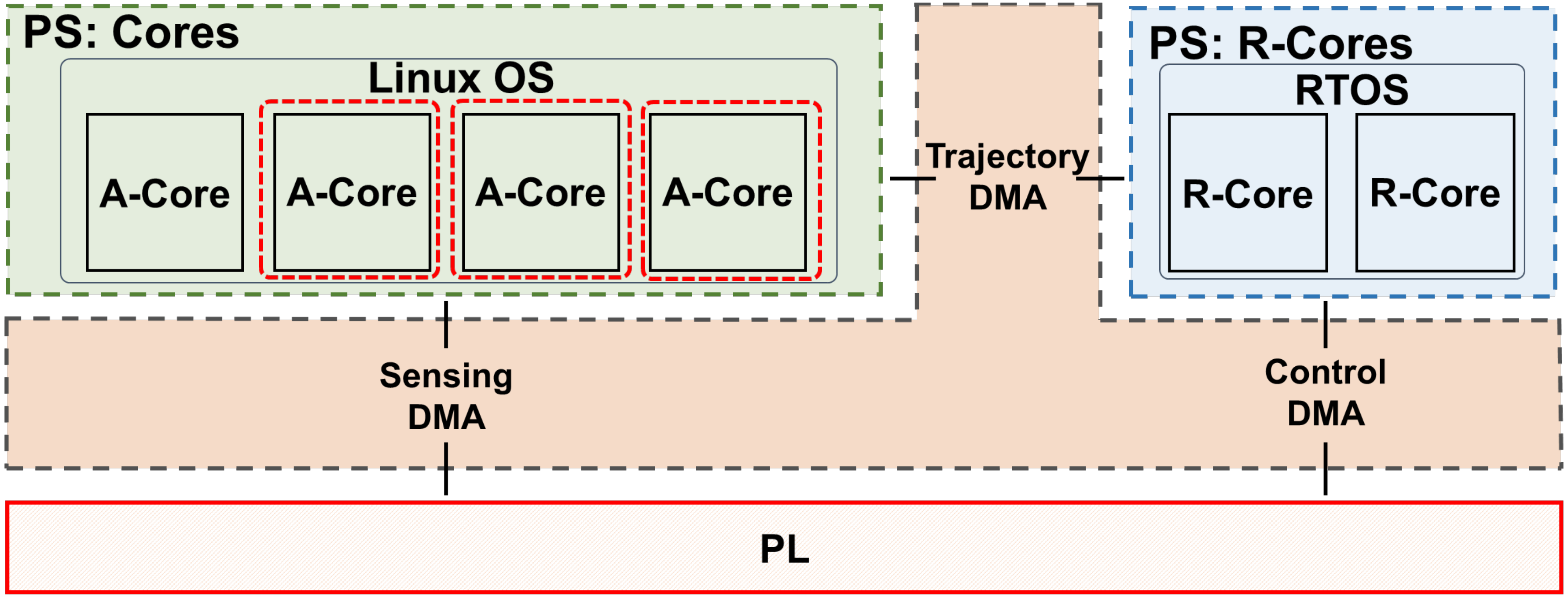
- **Netopia** is an industrial switch that simultaneously supports in-network planning and control.





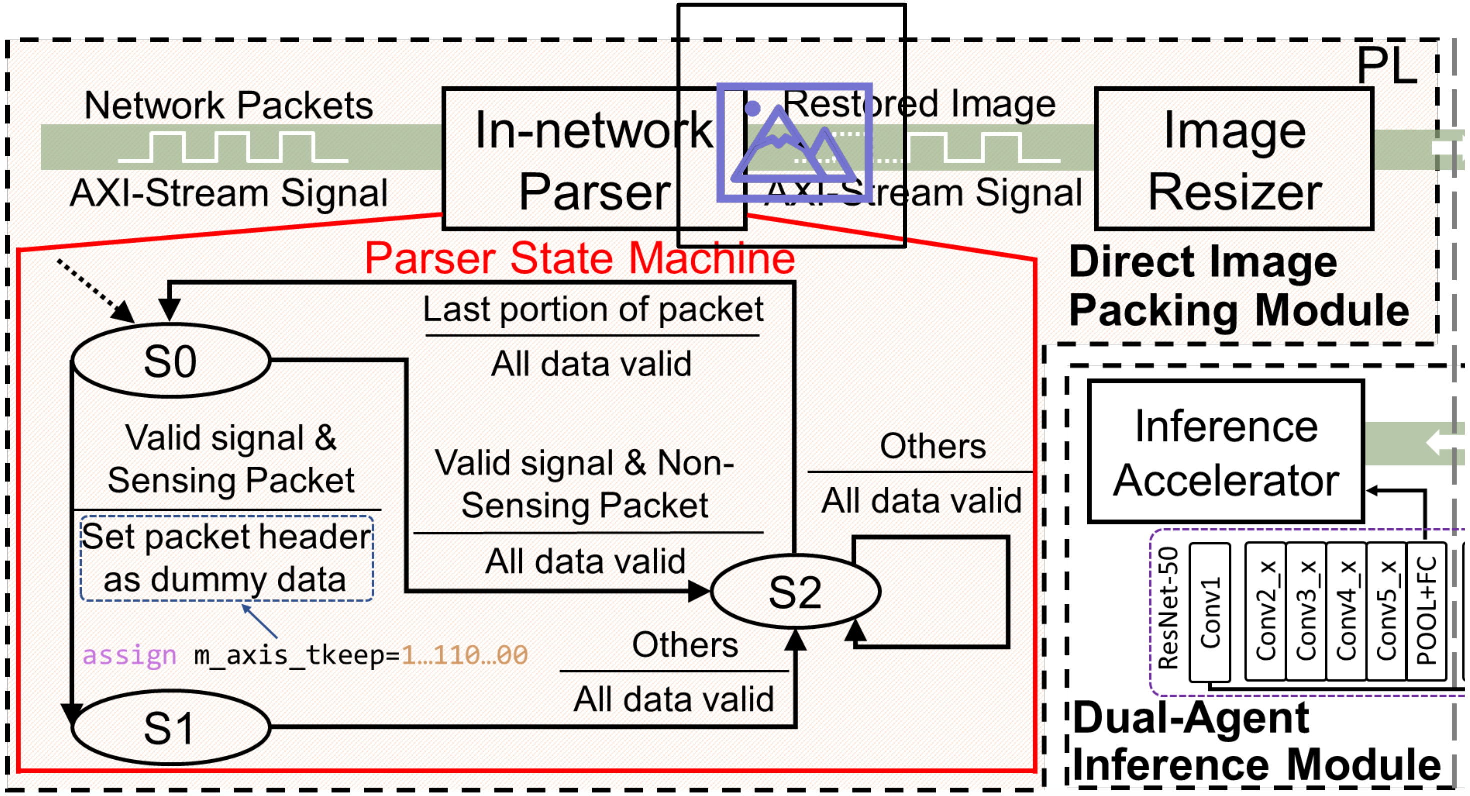
Switch Design: Delay Determinism Guarantee

- **Tri DMA for Intermediate Data Interaction**: leverage three dedicated DMAs to take over the intermediate data interaction process.
- **Sub-task Processing Determinism Guarantee**: isolated A-Core and Bare-metal R-Core keep sub-task processing away from uncertain OS scheduling



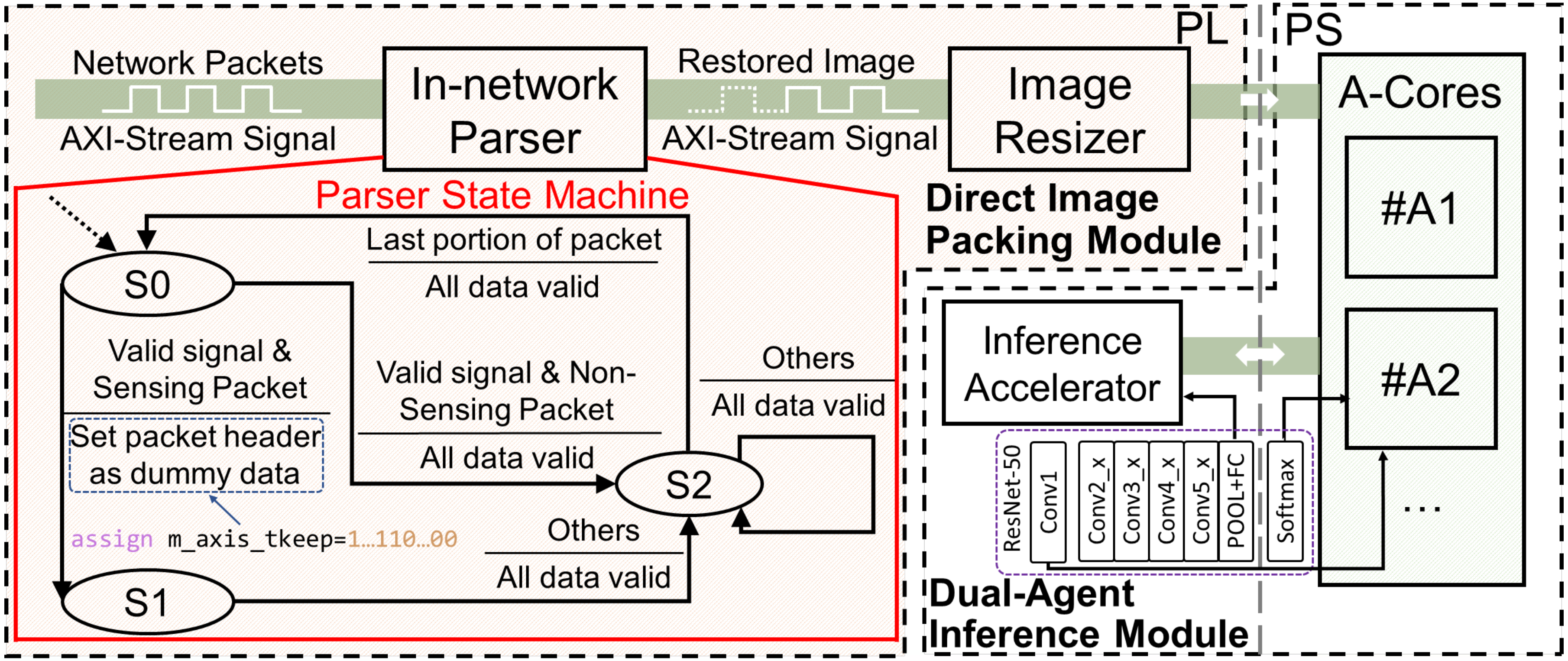
Switch Design: Task Computing Acceleration

- **Direct Image Packing**: an in-network packet parser for restoring image data from the network link layer, and then executes data pre-processing on PL.



Switch Design: Task Computing Acceleration

- **Dual-Agent Inference**: utilizes PL and two A-cores to accelerate neural network model inference.
  - Two isolated A-Cores for low-computing layers inference
  - pipelined Inference Accelerator for computation-intensive layers inference

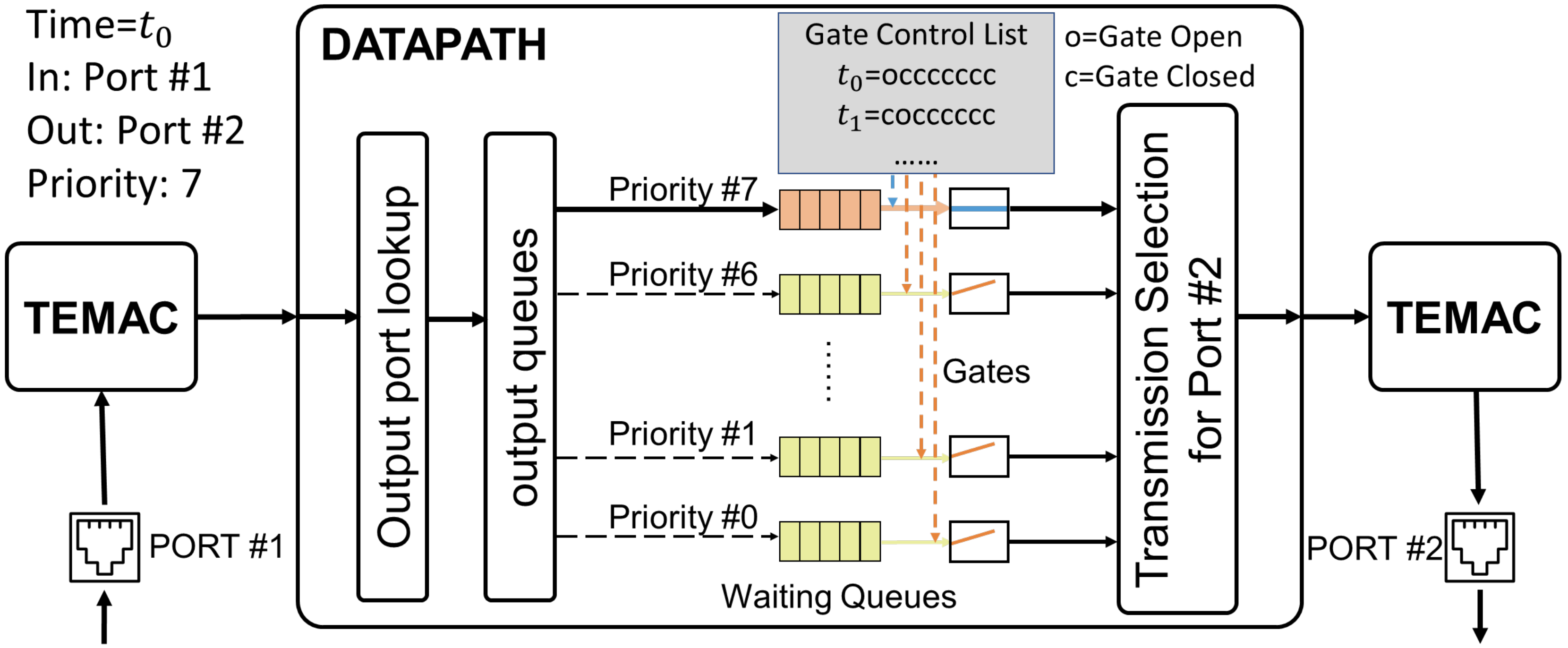


Switch Design: Packet Deterministic Forwarding

- **Time Synchronization**: all devices in the network enjoy the same global timestamp.
  - Follow IEEE 802.1AS
  - The synchronization algorithm is developed in PS using C, while the real-time clock and timestamping module are implemented in PL using Verilog.

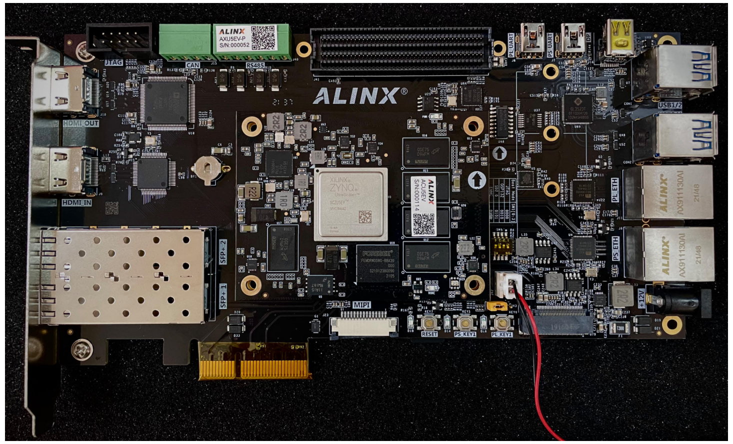
Switch Design: Packet Deterministic Forwarding

- **Time-Aware Shaper**: reserving dedicated bandwidth for critical traffic
  - Follow IEEE 802.1Qbv
  - Divide the network communication into fixed length, repeating time cycles
  - Use GCL(Gate Control List) to control the traffic transmission

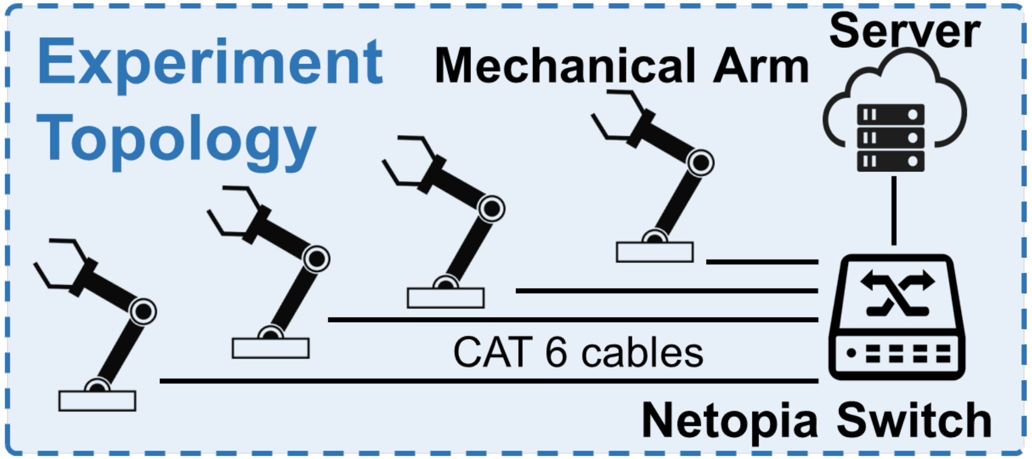


Evaluation

- Netopia switches are implemented on Zynq UntralScale+ MPSoC.
- Field studies:
  - A production line in the glass factory
  - two hours with around 1500 defect detection and grabbing tests
- Besides, conduct evaluations based on public robot control datasets



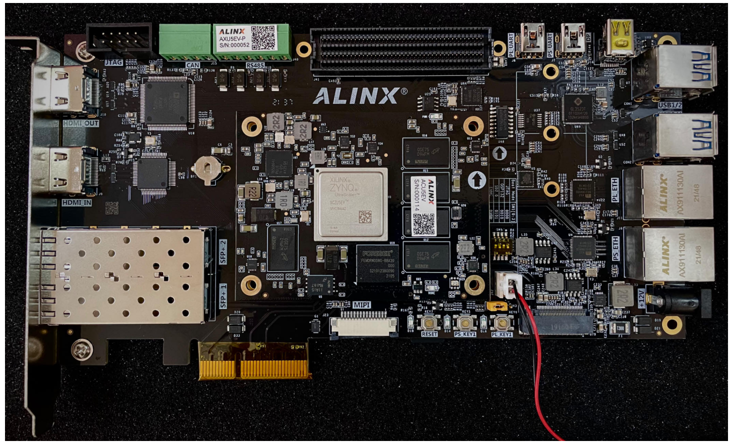
Zynq UntralScale+ MPSoC



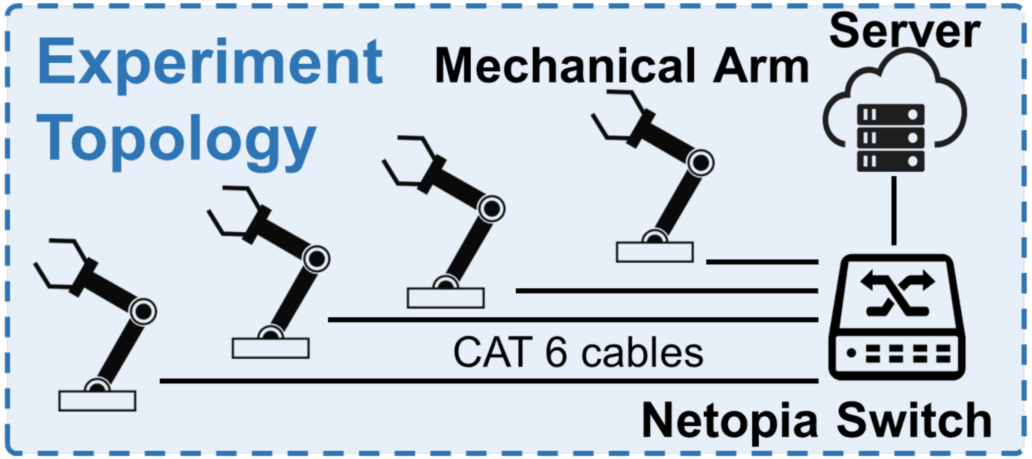
Mechanical Arm

Evaluation

- Baseline-I (IJCA'18): both the planning and control modules are offloaded to a cloud or edge server.
- Baseline-II (NSDI'22): the control module is offloaded to a network switch while the planning module is left on the server.



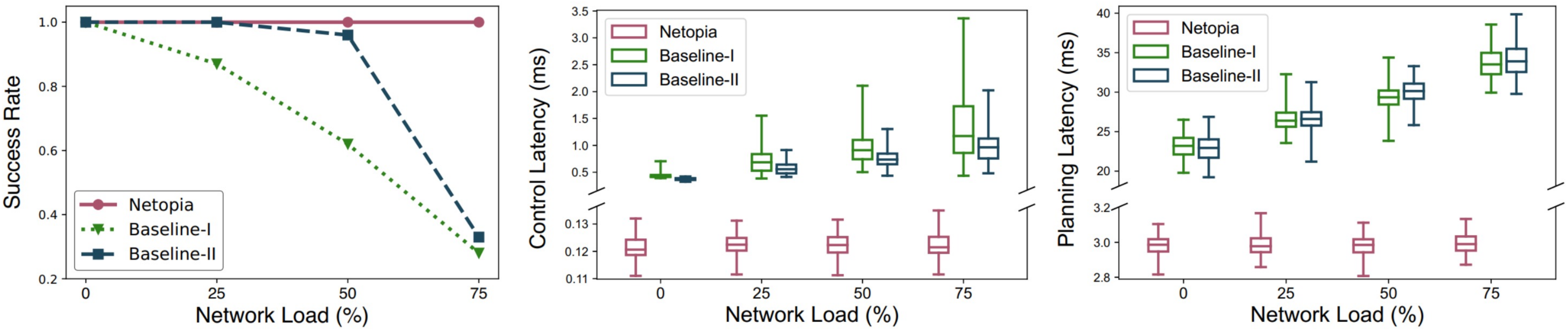
Zynq UntralScale+ MPSoC



Mechanical Arm

Overall Performance

Netopia succeeds in grabbing defective glass panes in all test cases and achieves an order of magnitude lower control and planning latency.

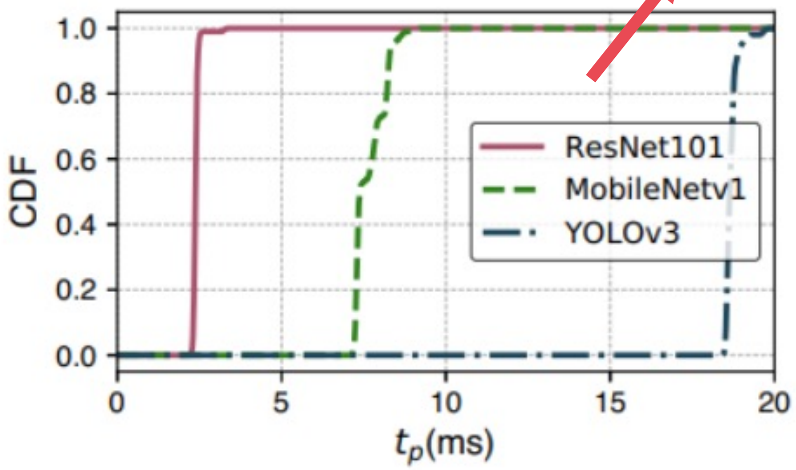


(a) Success Rate (b) Control Latency (c) Planning Latency

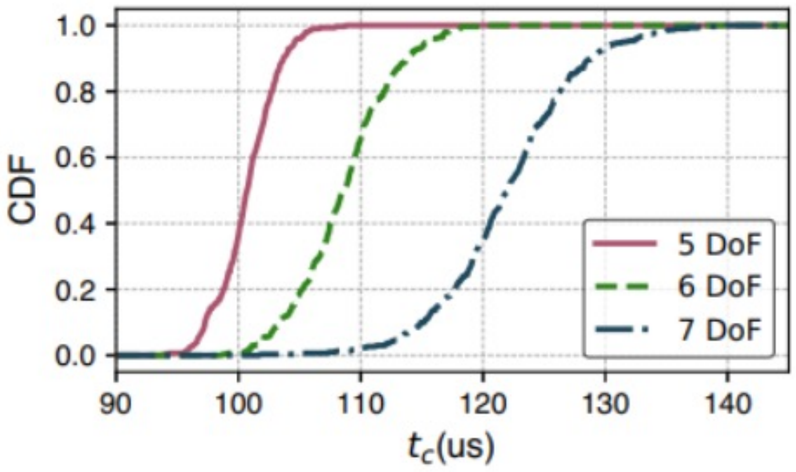
Robustness Study

Netopia is portable and scalable.

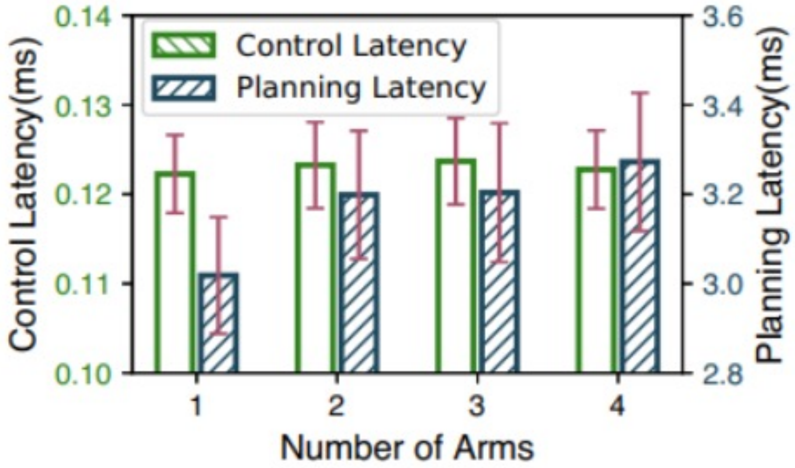
With various network models, t_p is small enough for deterministic requirements



(a) Impact of different neural network model



(b) Impact of arm's DoF

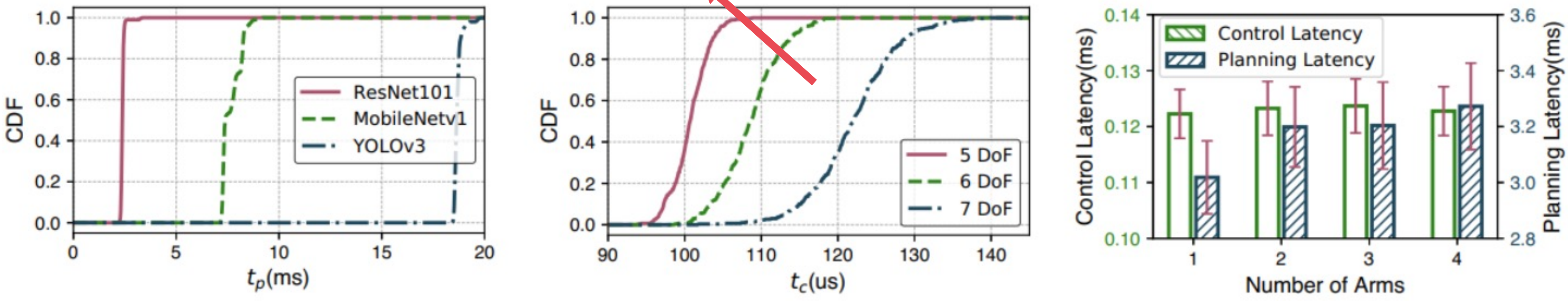


(c) Impact of number of connected arms

Robustness Study

Netopia is portable and scalable.

Despite the addition of axes in arms, Netopia will still maintain its reliability



(a) Impact of different neural network model

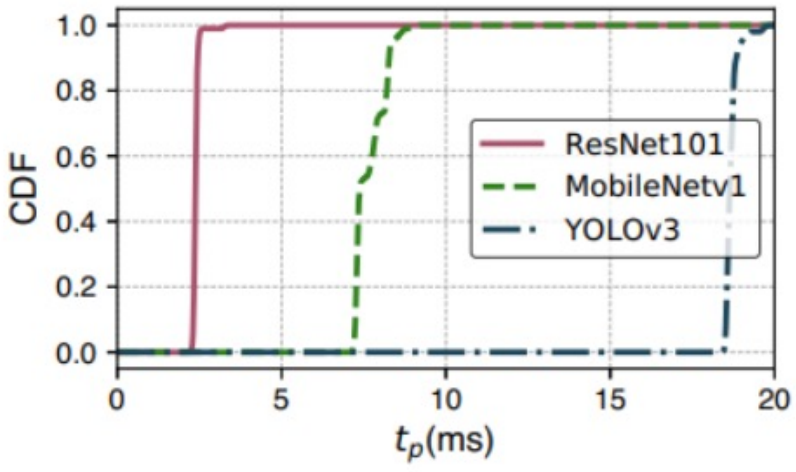
(b) Impact of arm's DoF

(c) Impact of number of connected arms

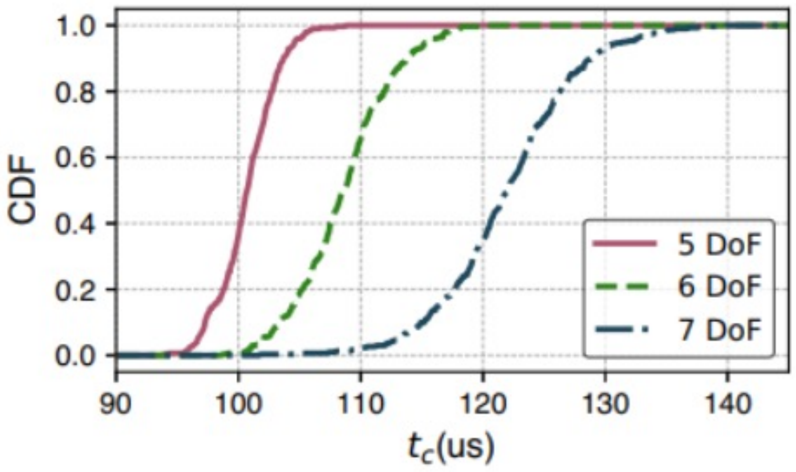
Robustness Study

Netopia is portable and scalable.

When multiple arms are connected, t_c remains stable while t_p increases slightly



(a) Impact of different neural network model



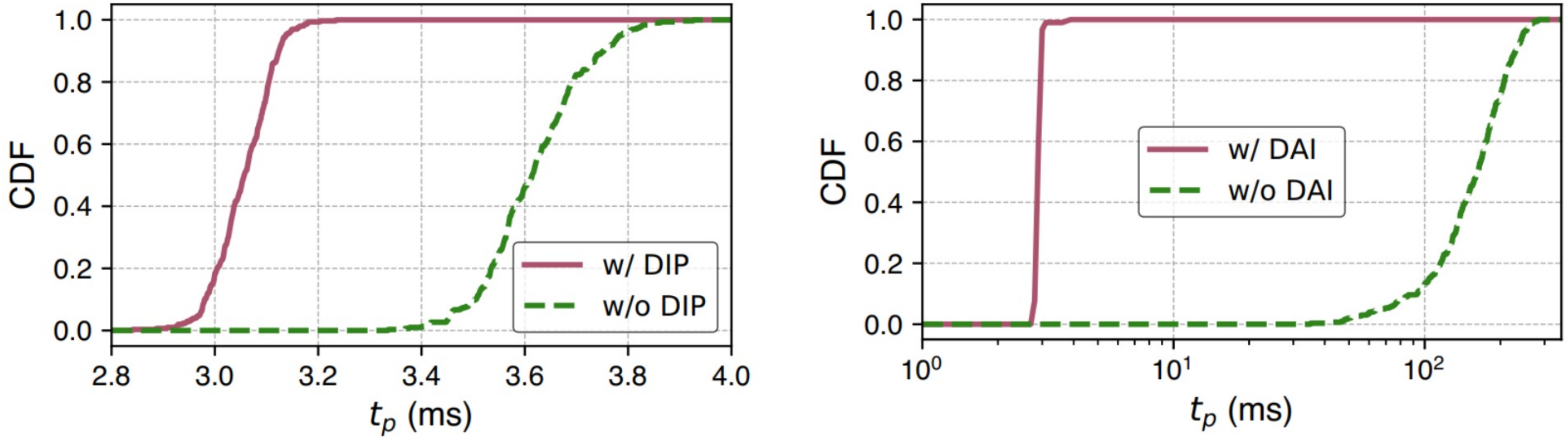
(b) Impact of arm's DoF

(c) Impact of number of connected arms

Ablation Study

Netopia has several designs to ensure determinism.

Without hardware acceleration, t_p increases significantly

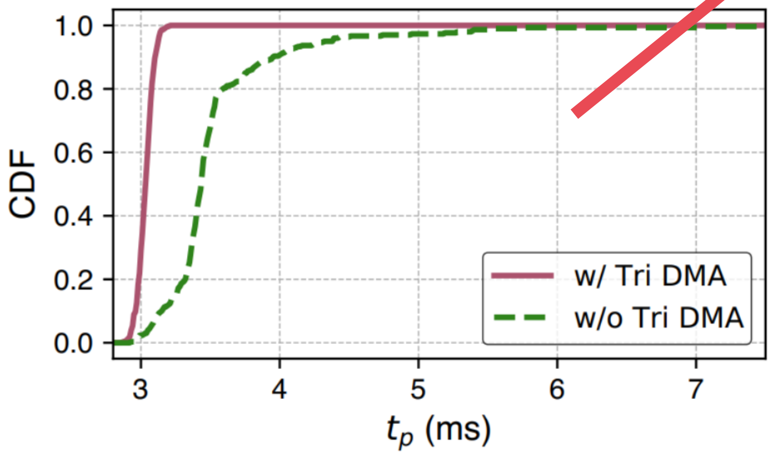


(a) Netopia w/ and w/o Direct Image Packing (b) Netopia w/ and w/o Dual-Agent Inference

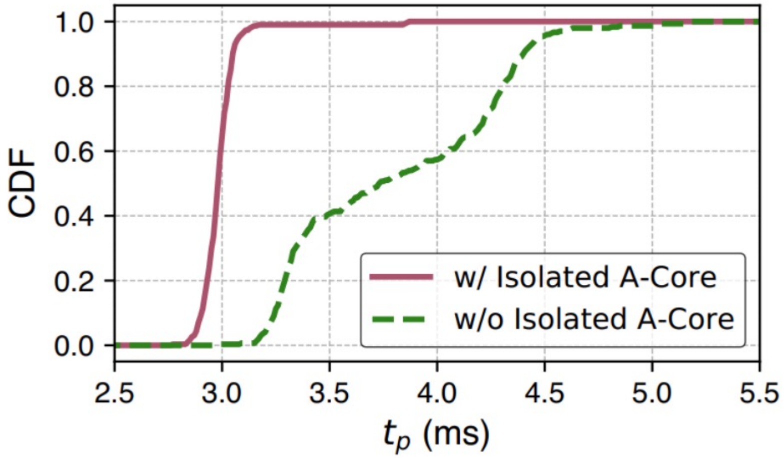
Ablation Study

Netopia has several designs to ensure determinism.

Without Tri-DMA, t_p increases significantly in the worst cases



(c) Netopia w/ and w/o Tri DMA

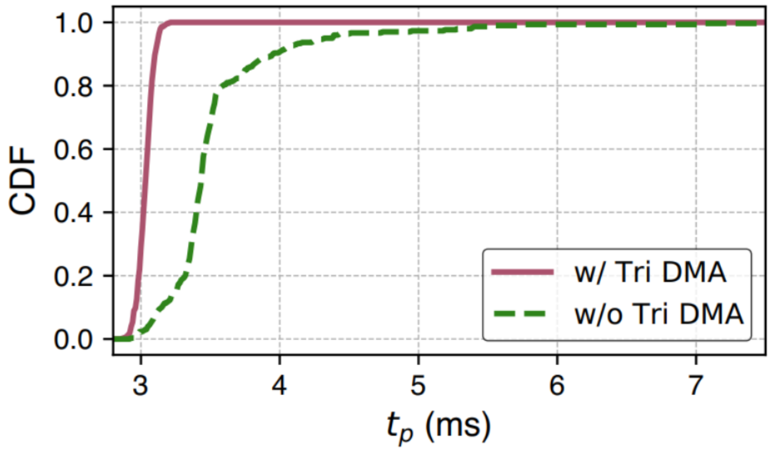


(f) Netopia w/ and w/o Isolated A-Core

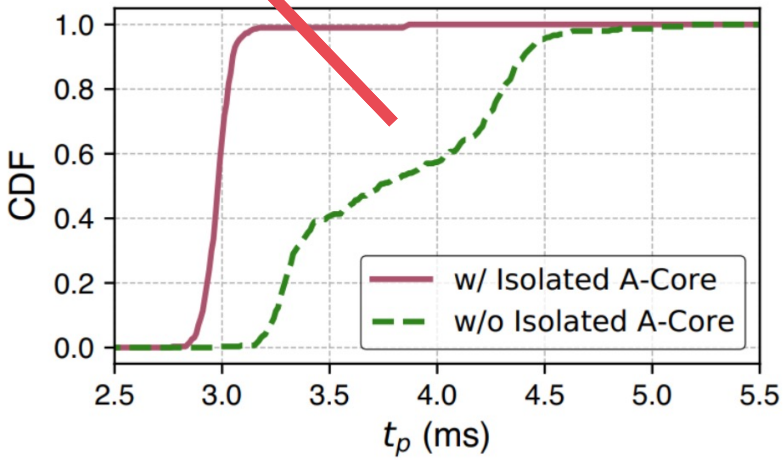
Ablation Study

Netopia has several designs to ensure determinism.

Without isolated A-Core, t_p increases due to interference from other processes



(c) Netopia w/ and w/o Tri DMA

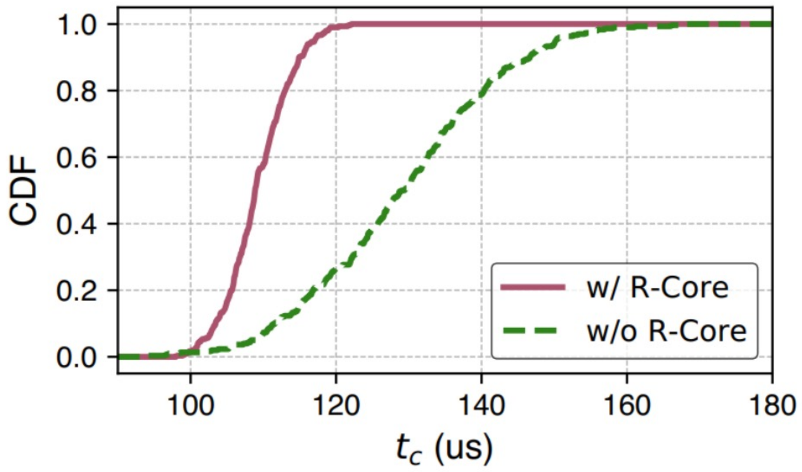


(f) Netopia w/ and w/o Isolated A-Core

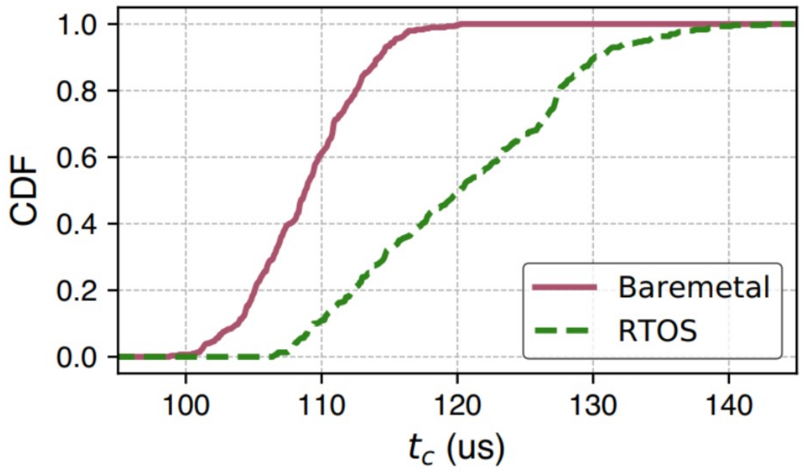
Ablation Study

Netopia has several designs to ensure determinism.

Without R-Core or RTOS, OS scheduling results in an addition in delay and jitter



(d) Netopia w/ and w/o R-Core



(e) Netopia with Bare-metal R-Core and RTOS

Conclusion

- We design and implement **Netopia**, the first industrial switch that makes both the planning and control modules compatible with in-network computing.
- We propose several technologies in Netopia to enable mechanical arms to obtain intelligent control commands with low and deterministic latency.
  - Delay Determinism Guarantee, Task Computing Acceleration, Packet Deterministic Forwarding,
- We conduct extensive evaluations to demonstrate the superior performance of Netopia.

Thanks!
Q&A