Forum on specification & Design Languages (FDL 2026)

Embodied AI Launcher: A Tool for Modeling, Simulation, and Deployment Evaluation of Transformer-Based Embodied AI Tasks at the Edge

Byeonggil Jun[1], Deeksha Prahlad[1], Mehdi Ghasemi[2], and Hokeun Kim[1]

1: Arizona State University, 2: Southern Illinois University

Guglielmo Marconi University, Rome, Italy

September 10, 2026

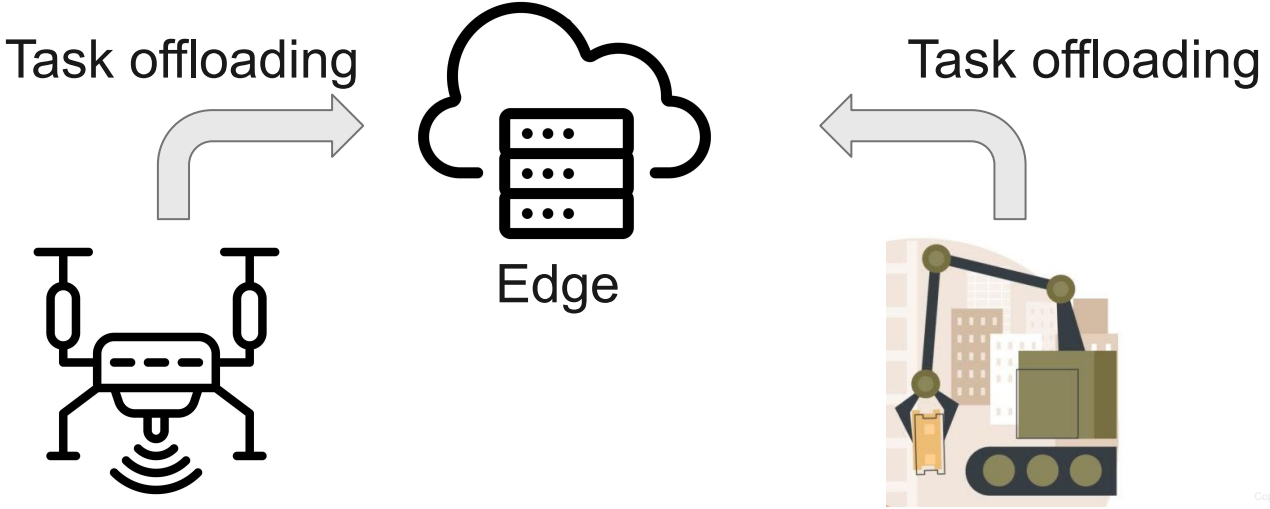
Arizona State University

Southern Illinois University CARBONDALE

ASU KIM KNOWLEDGEABLE & INTERACTIVE MACHINES

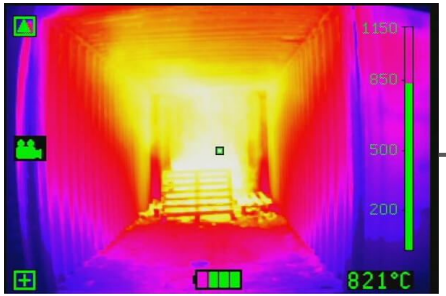
Introduction

- Embodied AI agents are emerging thanks to advances in LLM
  - Offloading ML tasks to shared servers for better schedulability
- Scheduling and orchestrating devices and tasks are challenging
- A tool for modeling, simulation and deployment evaluation of such systems are essential

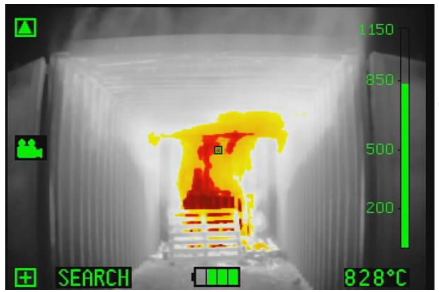


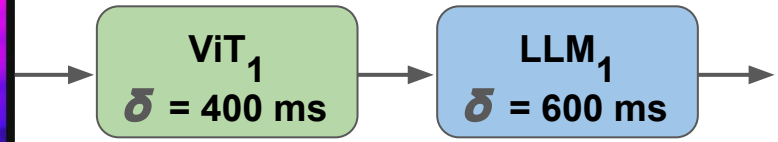
Motivating Example (Rescue Robot)

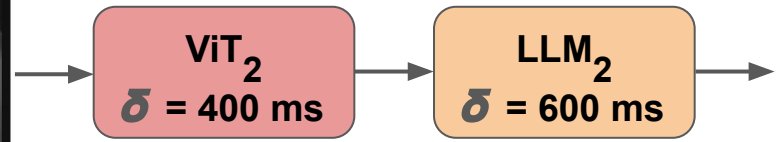
RGB-D Camera



Thermal Camera





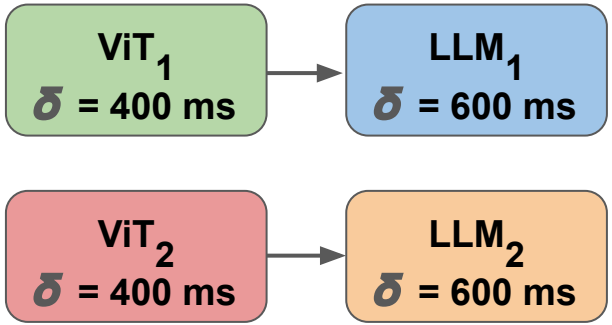


ViT: Vision Transformer Model
LLM: Large Language Model
δ: Deadline

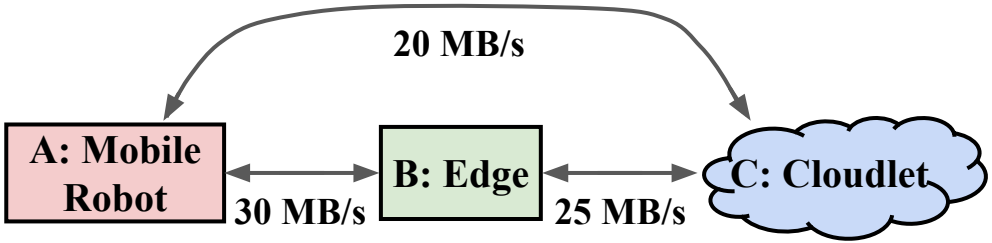
https://vatropromet.hr/en/articles/tips/how-choose-thermal-imaging-camera-firefighting-32/

What Would be Scheduling Results?

Tasks (represented with a DAG)



Devices

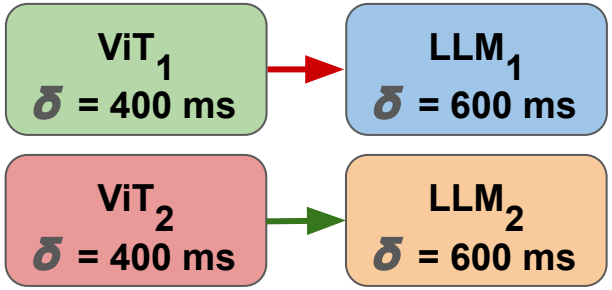


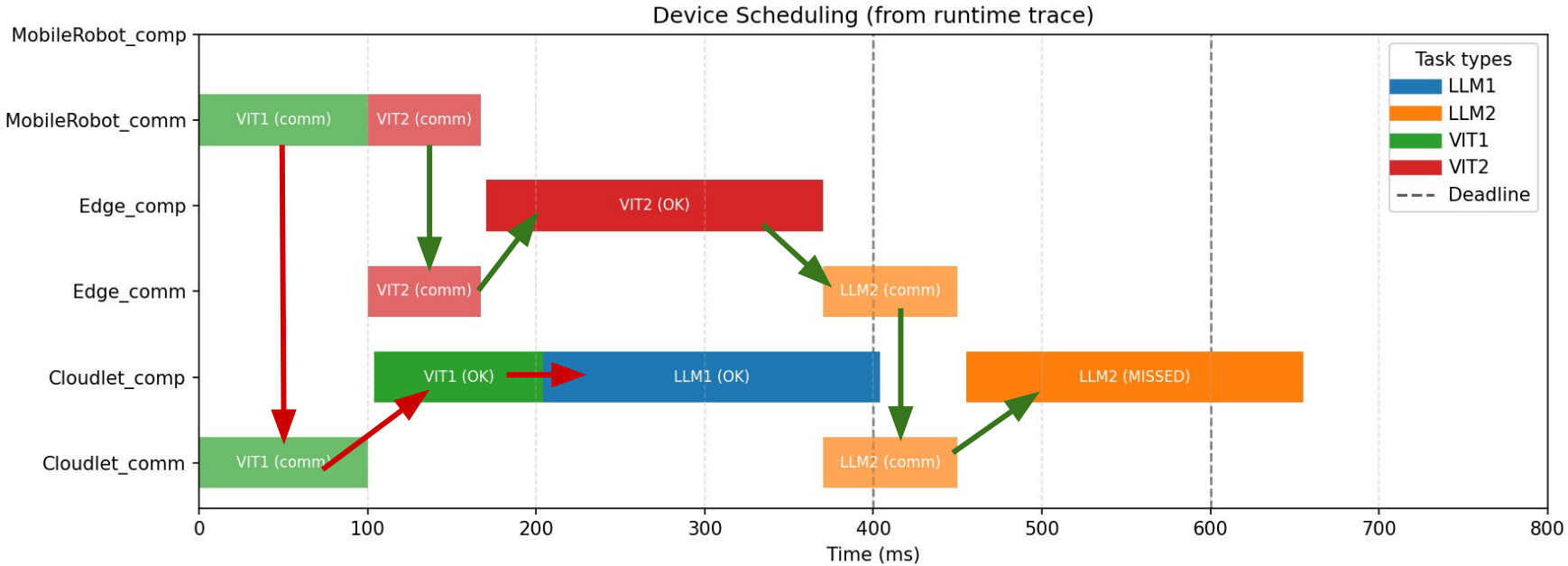
Execution Times

Execution time	A (Mobile Robot)	B (Edge)	C (Cloudlet)
ViT1 and ViT2	300 ms	200 ms	100 ms
LLM1 and LLM2	600 ms	400 ms	200 ms

Simulation Result

- Visualized simulation result helps analysis

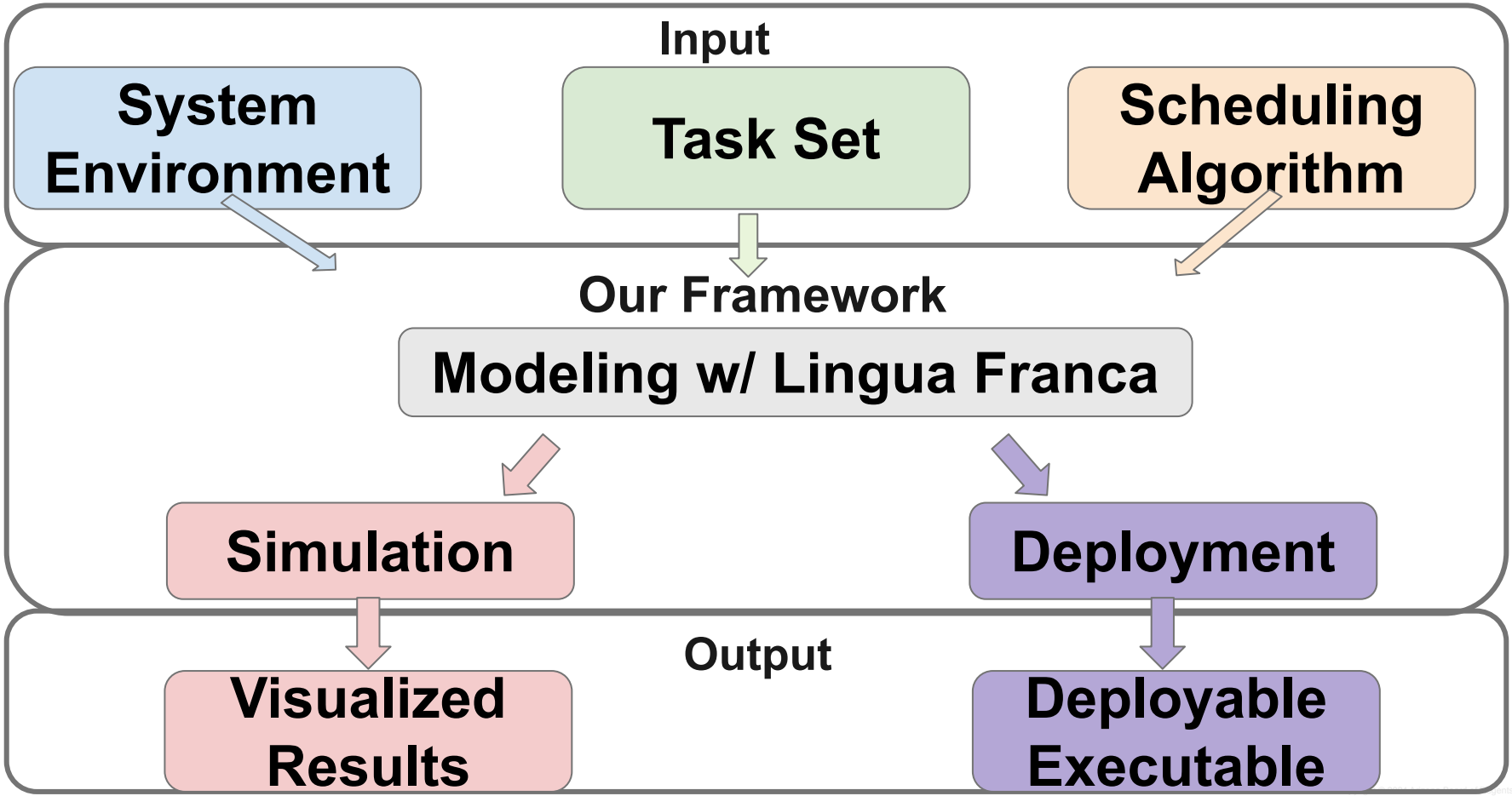




→ ViT_1 -> LLM_1 flow

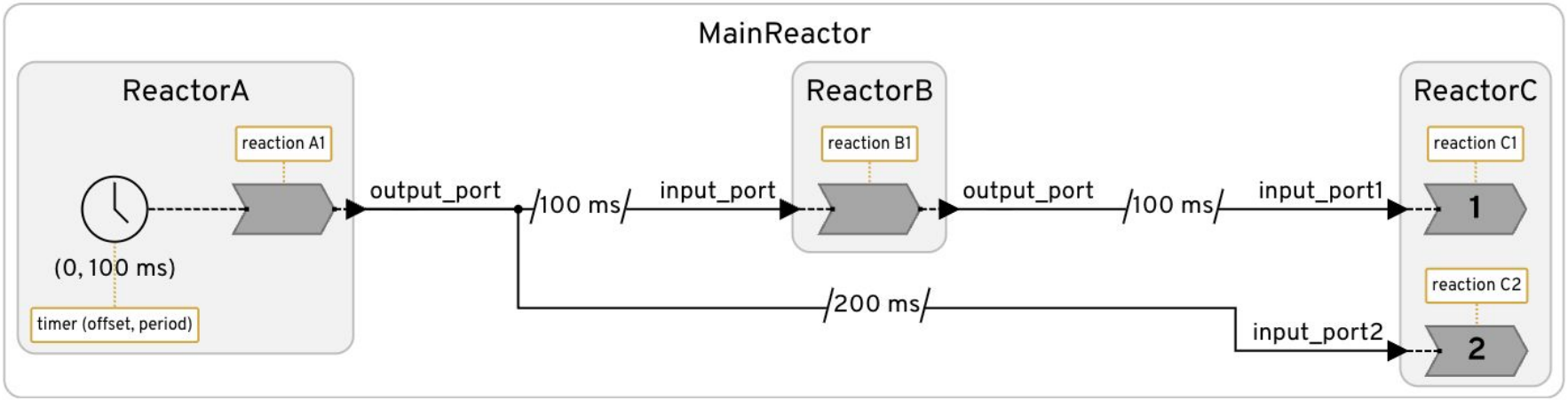
→ ViT_2 -> LLM_2 flow

Proposed Framework



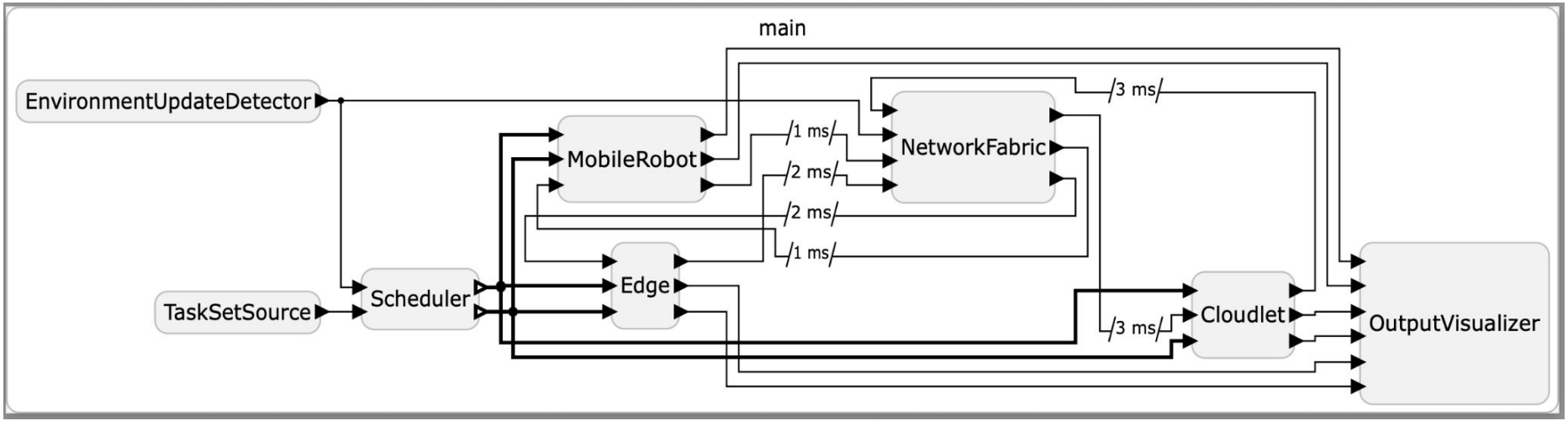
Background - Lingua Franca

- Lingua Franca (LF) is a coordination language for **reactor-oriented, time-sensitive systems**
- Our tool is built upon LF because
  - LF supports **deterministic concurrency** in distributed systems
  - Simulation and deployment share the same system model, enabling scheduling evaluation before deployment

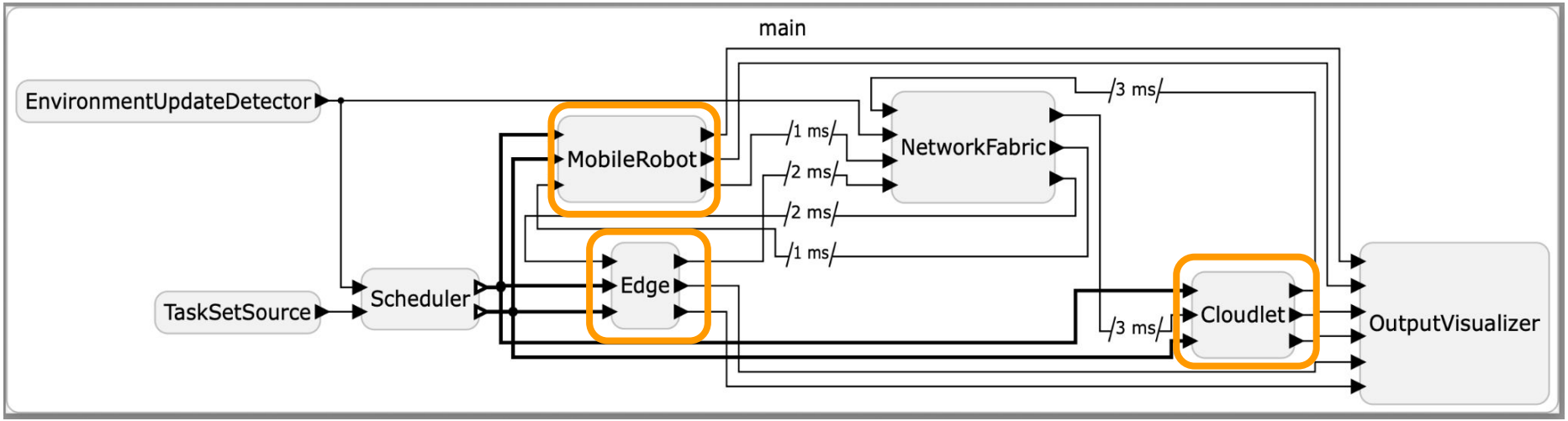


https://www.lf-lang.org/docs/

Lingua Franca Modeling (Simulaiton)



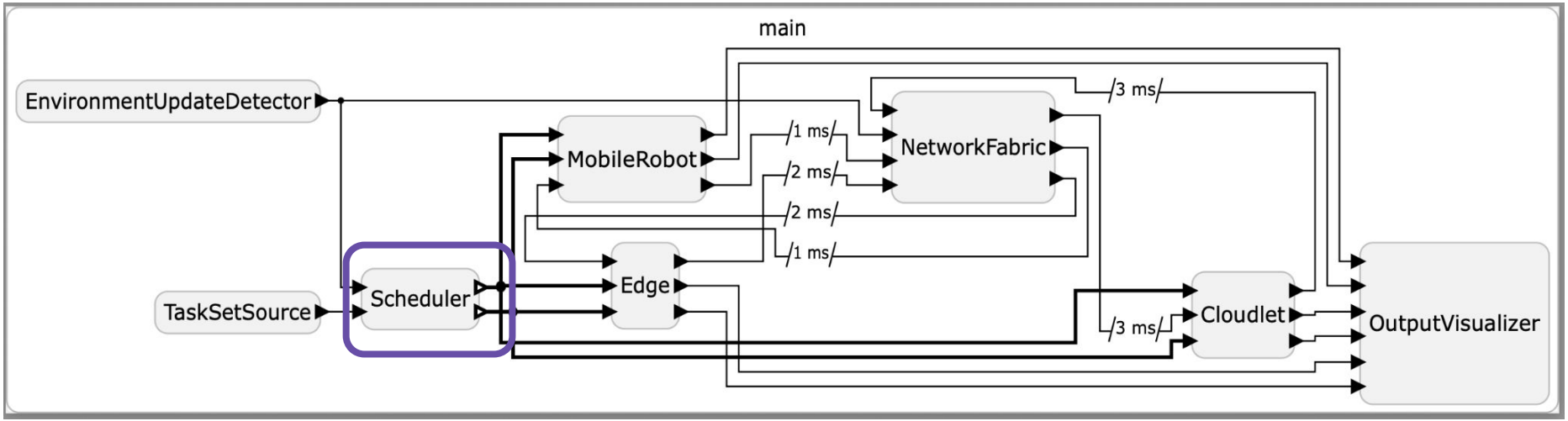
Device Reactors



- We model each device as a reactor (MobileRobot, Edge, and Cloudlet)

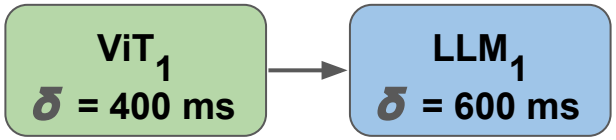
Distribute Scheduling Decision

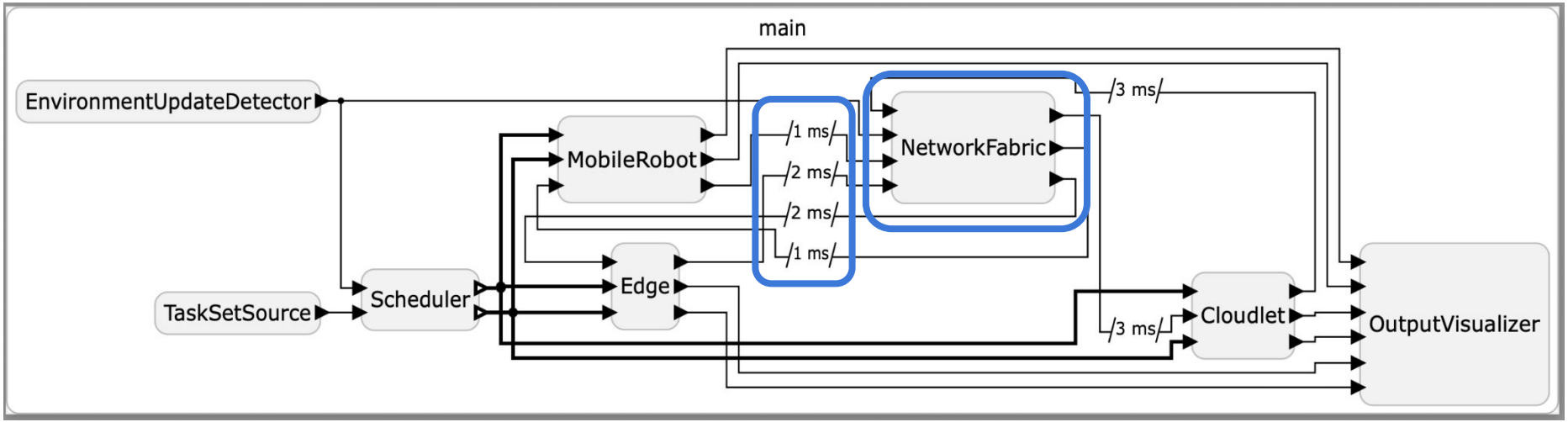
ViT_1 δ = 400 ms → LLM_1 δ = 600 ms



- At the beginning, Scheduler sends the scheduling decision
- Example scheduling and mapping
  - MobileRobot: ViT_1
  - Edge: LLM_1

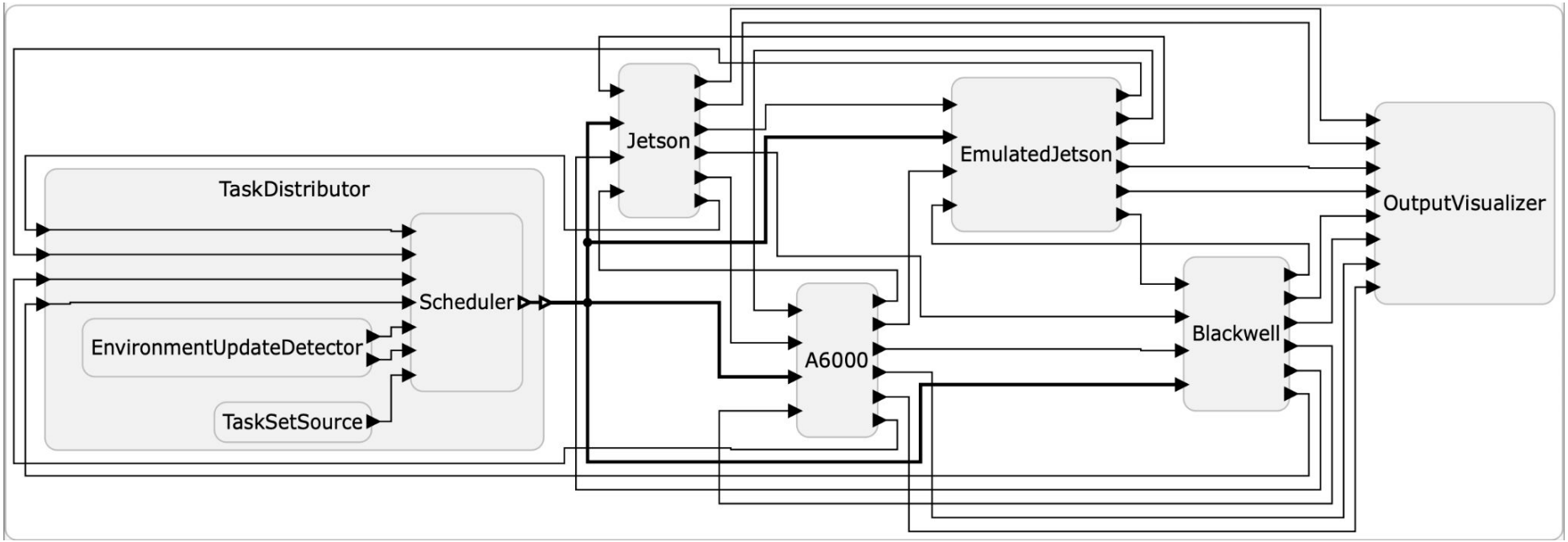
Communication Modeling





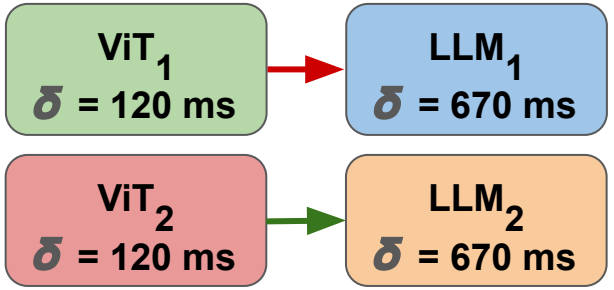
- Communication delay = latency + data size / bandwidth
- Network latency is modeled with `after` delays
- **NetworkFabric** applies additional delay (data size / bandwidth) when forwarding network messages

Lingua Franca Modeling (Deployment)

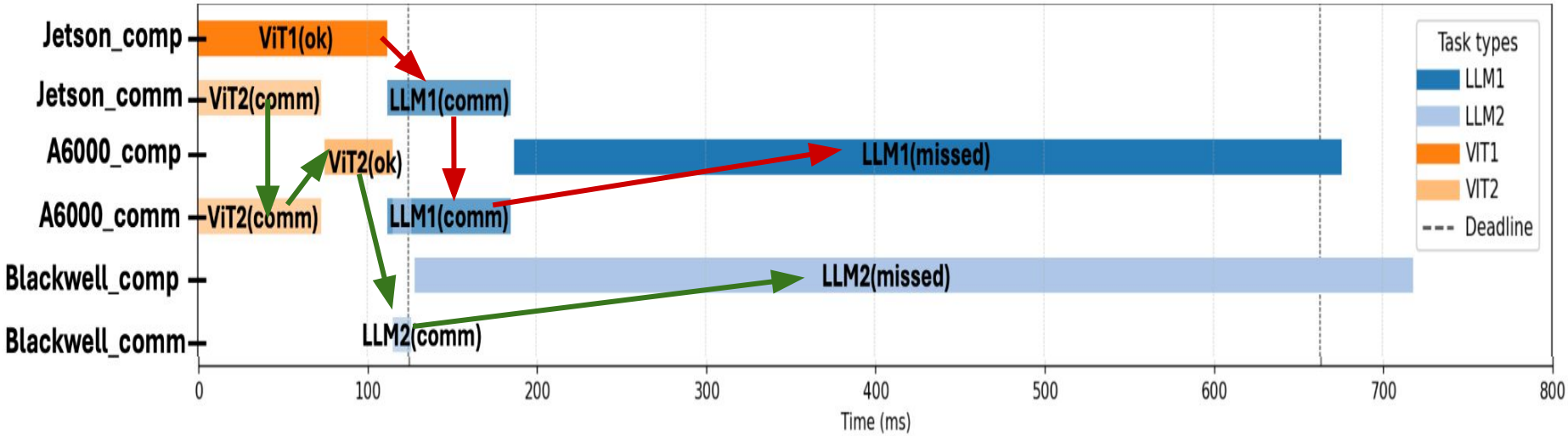


- Deployment shares the same system model with Simulation
- Each device reactor runs on physically distributed computing nodes, and communicate with each other through p2p network

Simulation Result



- Simulate a system with the profiling results of **YOLOS-S** (FP16) and **TinyLlama** (FP16) on Jetson, and two servers equipped with A6000 and Blackwell GPU

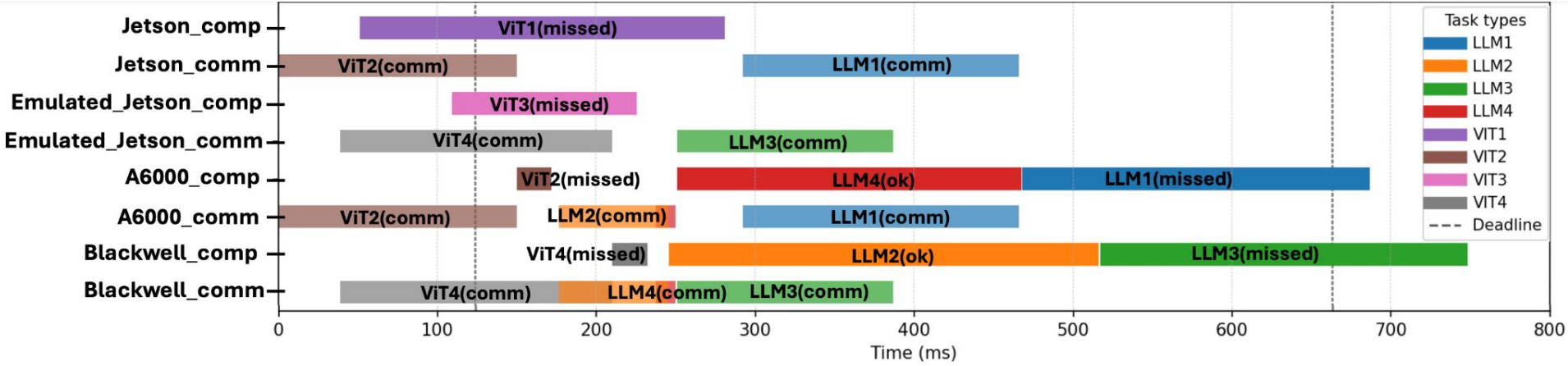


ViT1 -> LLM1 flow

ViT2 -> LLM2 flow

Deployment Result

- Deploy 4 ViT -> LLM chains on 4 devices (2 robots, 2 edge servers)
- The tool successfully coordinate distributed physical devices by respecting the task dependency
- The runtime trace enables easier analysis of the scheduling result



Conclusion

- We propose **Embodied AI Launcher** for modeling, simulation, and deployment evaluation of transformer-based AI tasks at Edge
- Our tool enables systematic analysis of computation, communication, and deadline outcomes

- **Contact**
  - byeonggil@asu.edu (Byeonggil Jun - Lead Author)
  - hokeun@asu.edu
  - https://labs.engineering.asu.edu/kim/

Thank you!

- **Our tool is publicly available in Github: https://github.com/asu-kim/embodied-ai-launcher**

ASU KIM
KNOWLEDGEABLE & INTERACTIVE MACHINES