IEEE INFOCOM
IEEE International Conference on Computer Communications
18–21 May 2026 // Tokyo, Japan

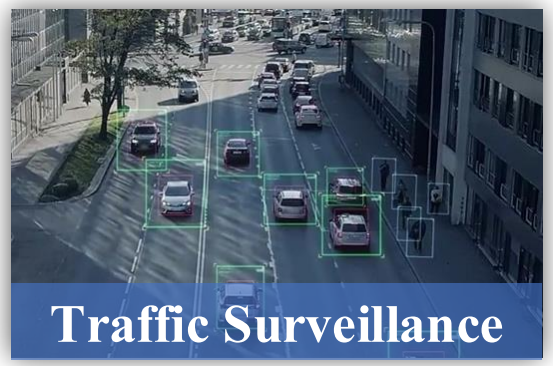
Tackling the Imbalance in Video Analytics Pipelines with Hierarchical Embodied Intelligence

Wenhui Zhou, Lei Xie*, Jingyi Ning, Shuyu Cao,
Hao Wu, Qinghua Peng, Long Fan

State Key Laboratory for Novel Software Technology, Nanjing University

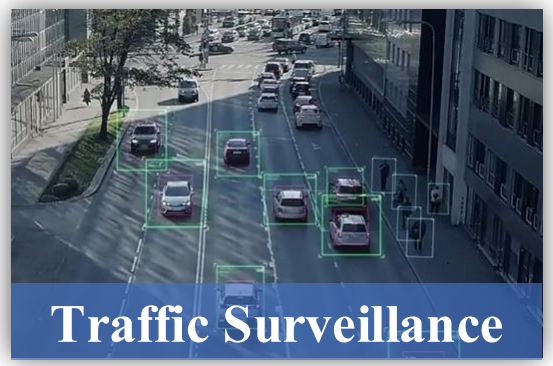
Background

- **Various Video Analytics Applications:**

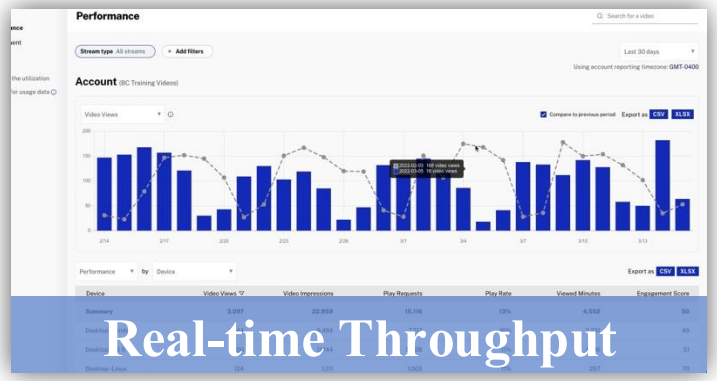


Background

➢ Various Video Analytics Applications:



➢ Quality of Experience (QoE) in Video Analytics:

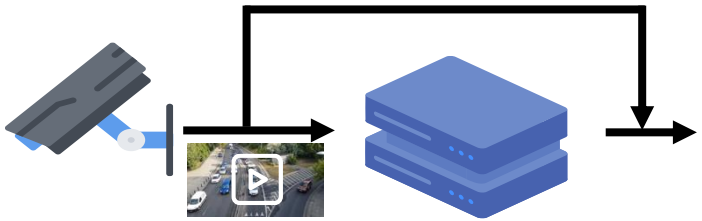


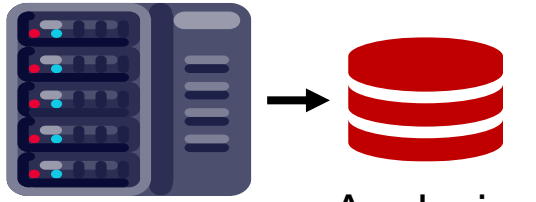
Background

- **A typical video stream analytics pipeline**

Background

- **A typical video stream analytics pipeline**



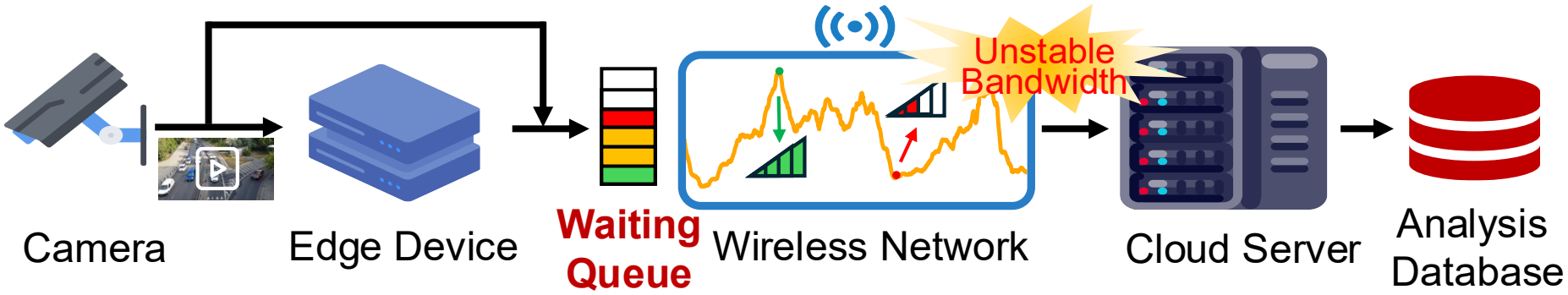


Background

- **A typical video stream analytics pipeline**

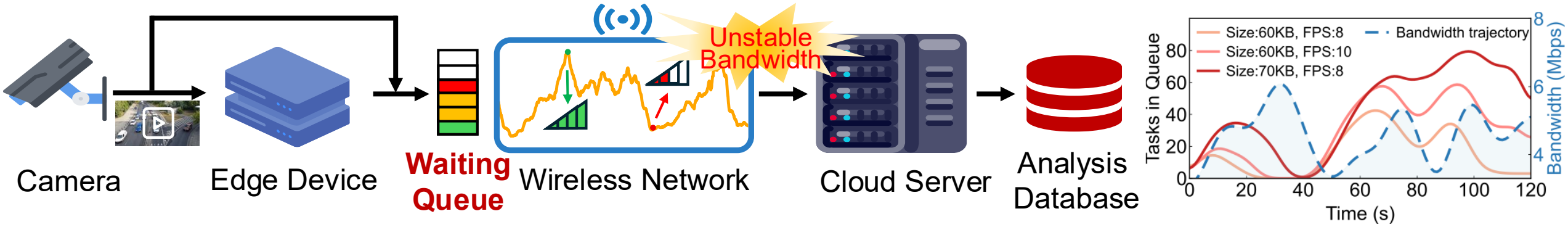
Background

- **A typical video stream analytics pipeline**



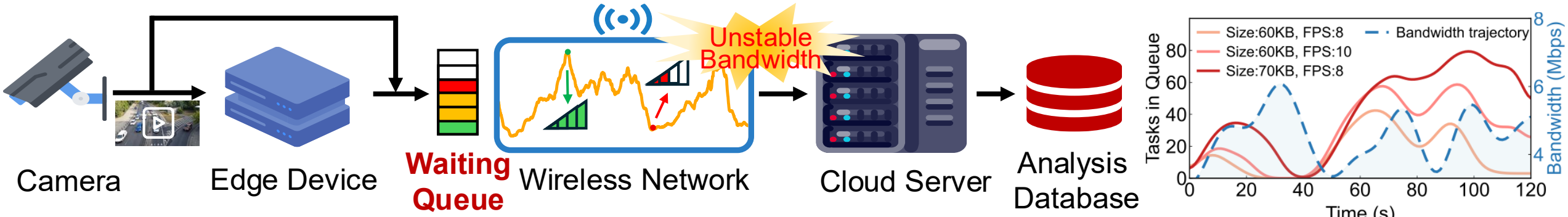
Background

➢ **A typical video stream analytics pipeline**



Background

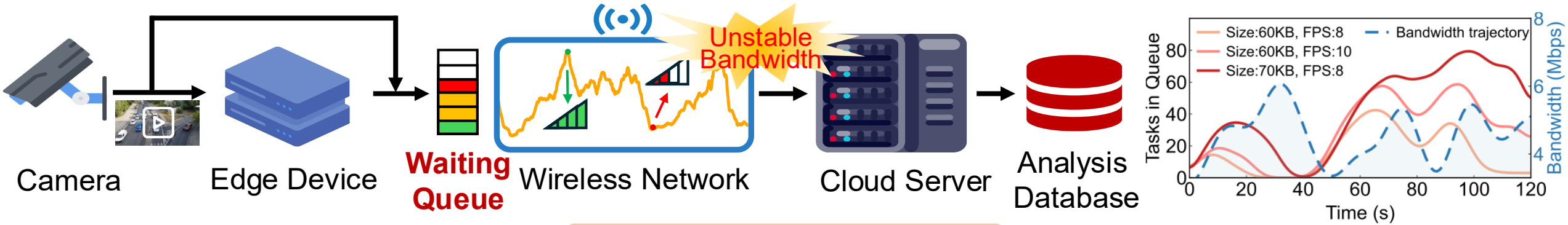
➢ **A typical video stream analytics pipeline**



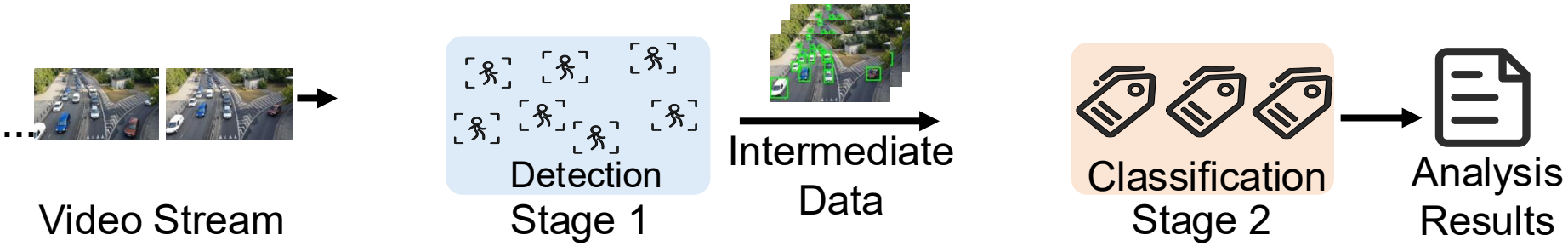
Spatial Imbalance

Background

➢ **A typical video stream analytics pipeline**

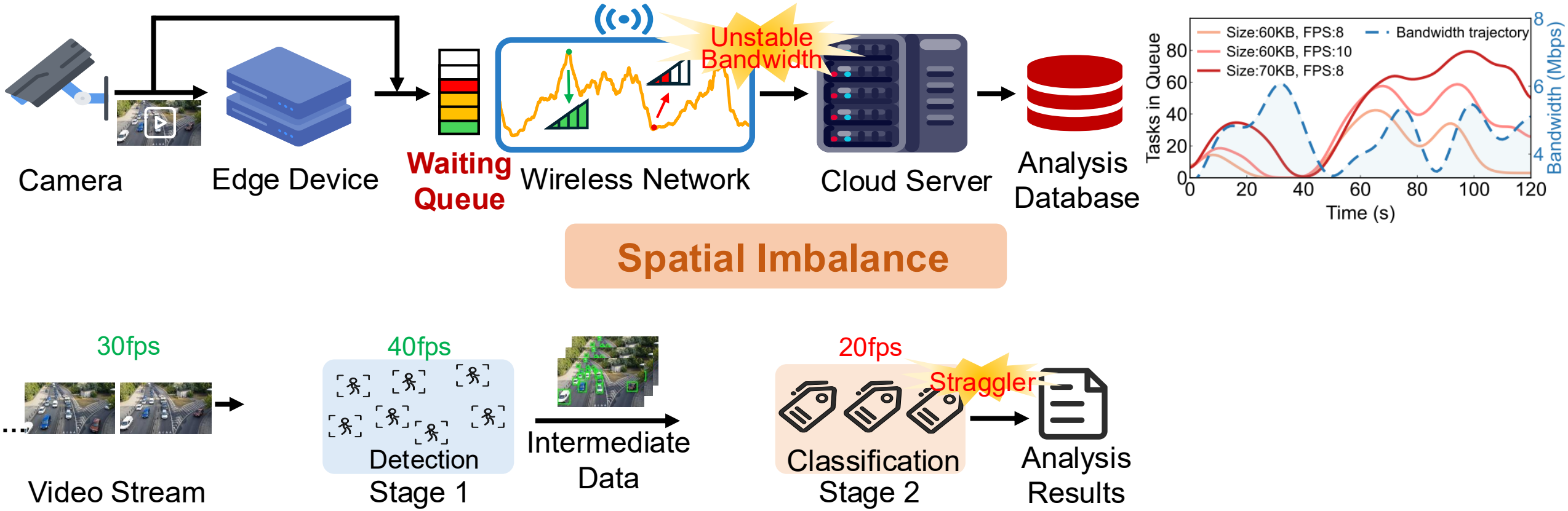


Spatial Imbalance



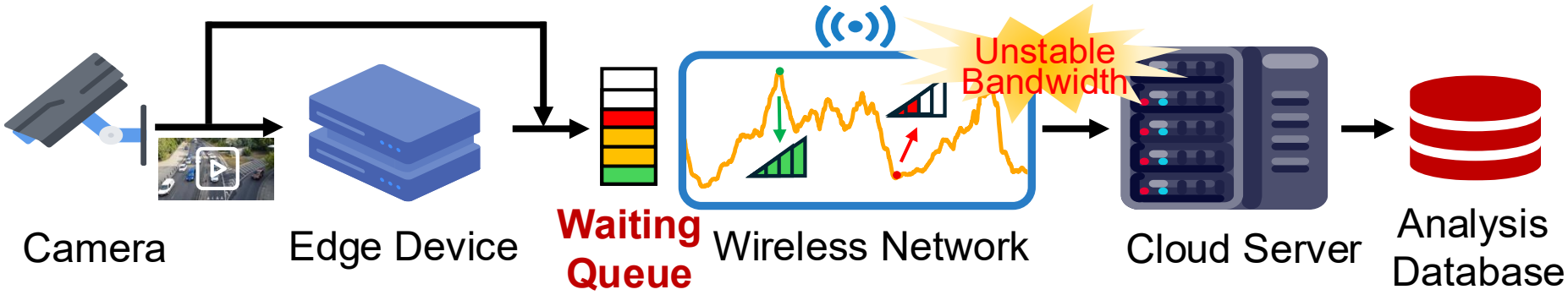
Background

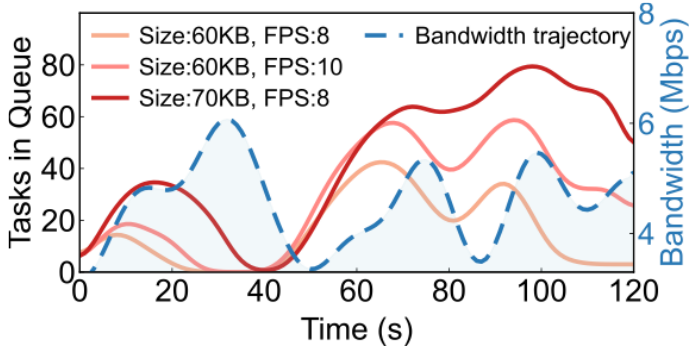
➢ **A typical video stream analytics pipeline**



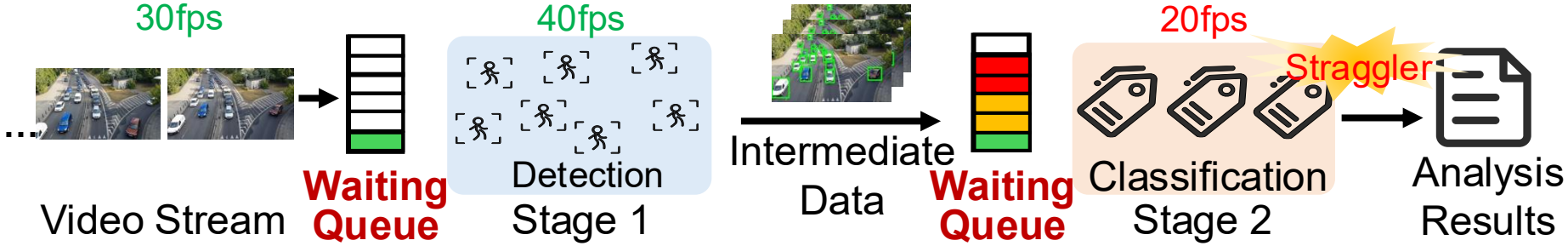
Background

➢ **A typical video stream analytics pipeline**



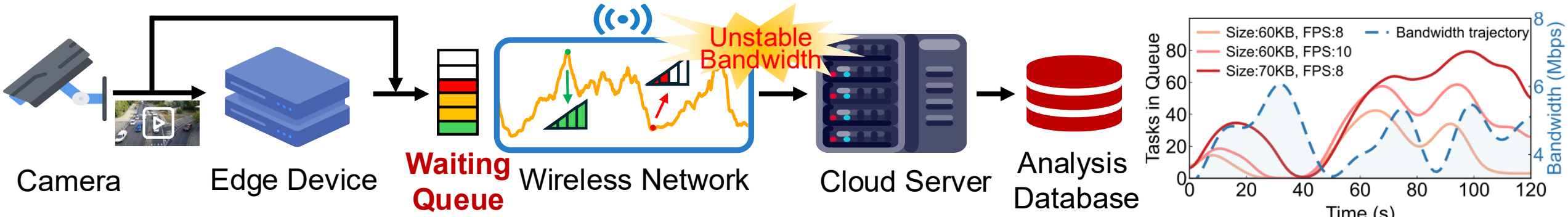


Spatial Imbalance

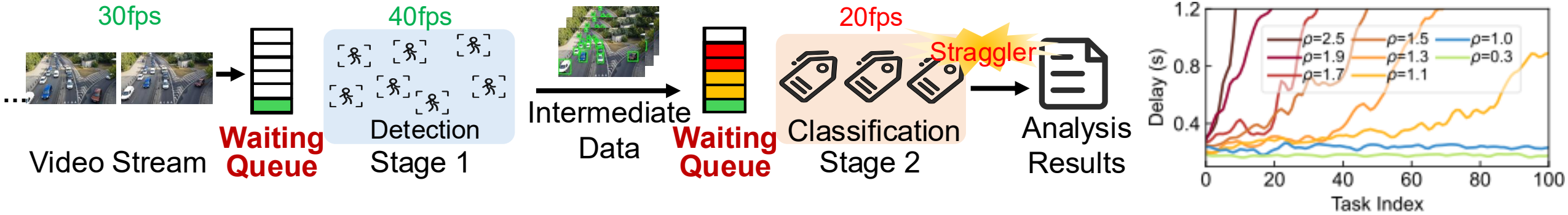


Background

➢ A typical video stream analytics pipeline

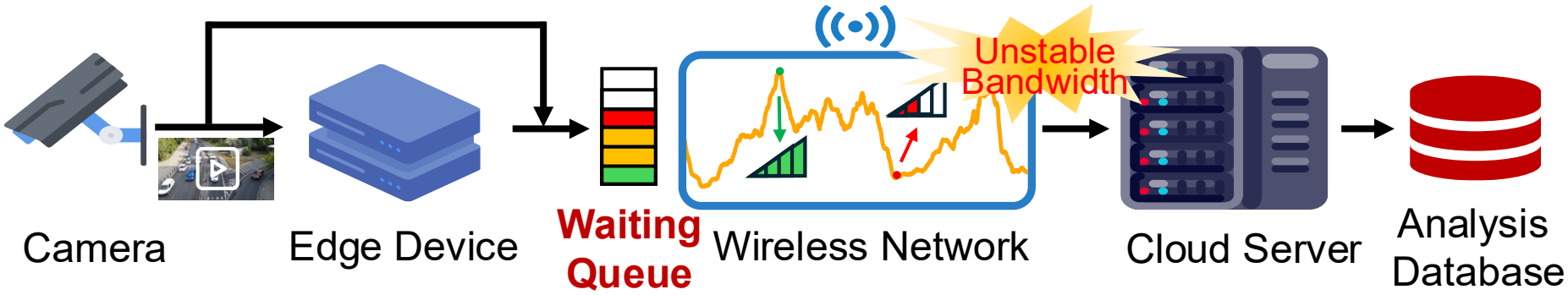


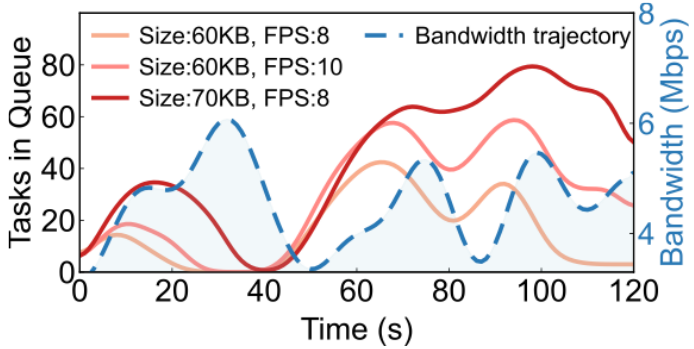
Spatial Imbalance



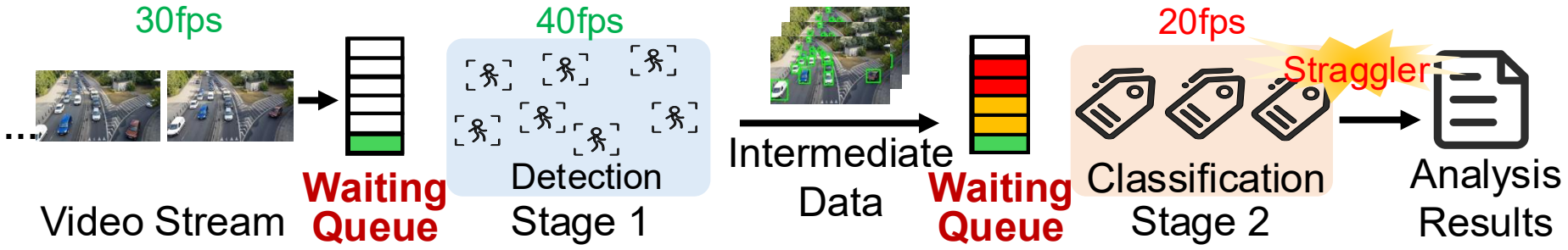
Background

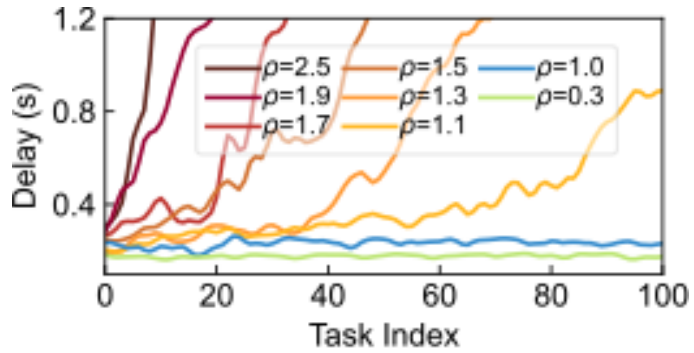
➢ A typical video stream analytics pipeline





Spatial Imbalance

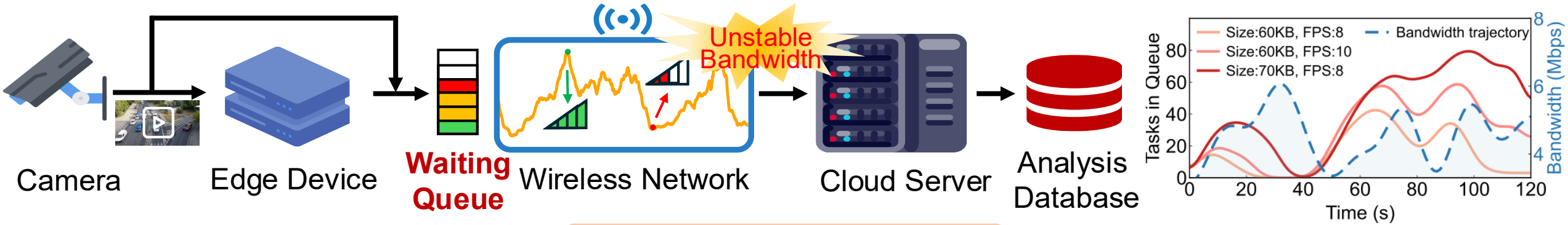




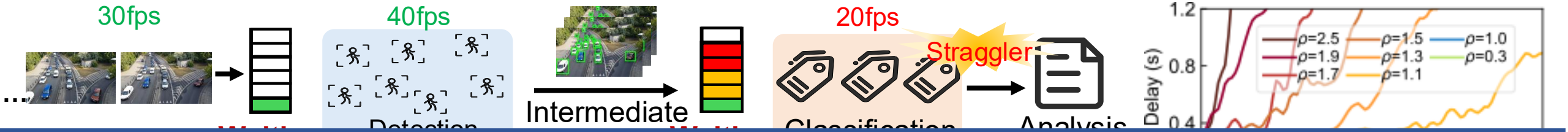
Temporal Imbalance

Background

➢ **A typical video stream analytics pipeline**



Spatial Imbalance



Adjust "knobs" can be a feasible approach

Motivation

➢ "Knob" in Video Analytics Pipelines

✓ Video Configuration (source)

frame resolution: pixel dimensions of each frame

frame rate: throughput of incoming video frames

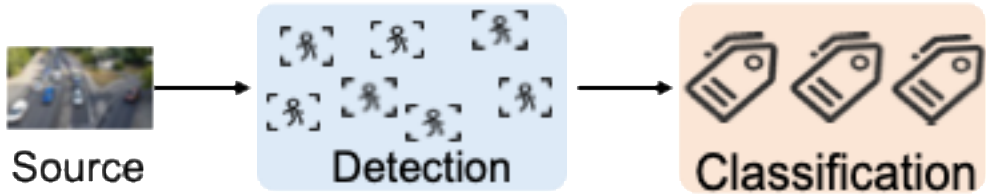
✓ Execution Acceleration (detection)

batch size: frame number in a processing segment (first for detection and others for tracking)

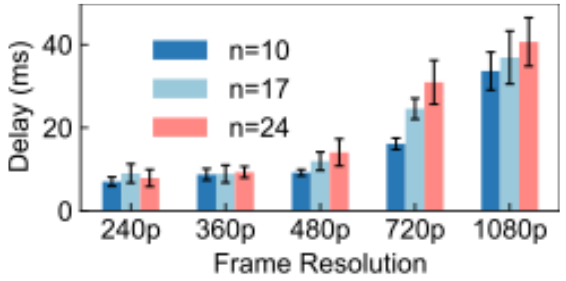
✓ Distributed Collaboration (classification)

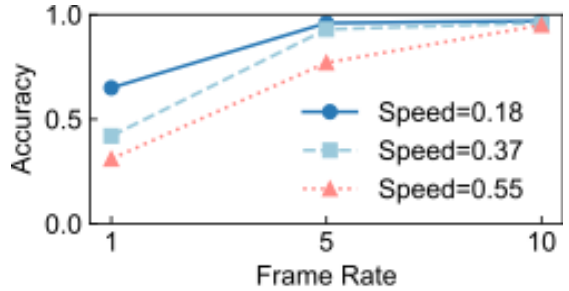
pipeline partition point: cloud–edge split between pipeline (left on edge / right on cloud)

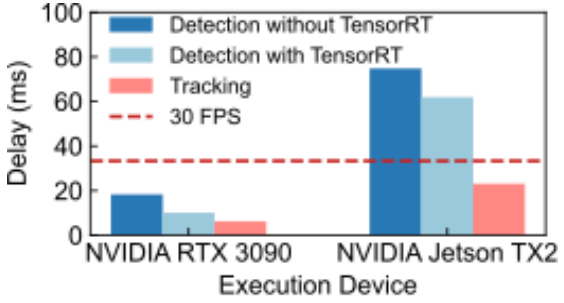
region allocation policy: assignment of detected regions to edge devices for parallel classification

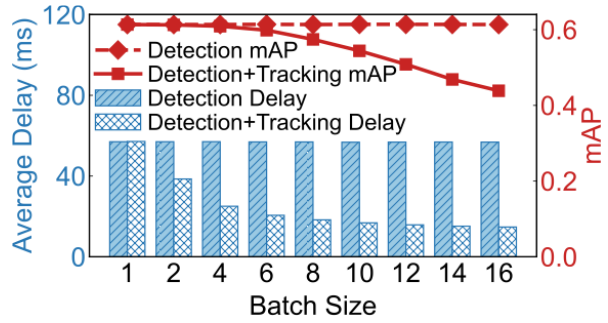


Source-Detection-Classification (SDC) Framework

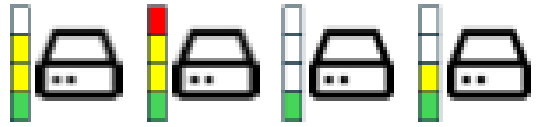








Pipeline partition point



Region allocation policy

Motivation

➢ Knob Selection

Category	Knob	Optional Values
Configuration	Frame resolution	240p,360p,480p,540p,720p,900p,1080p
	Frame rate	1, 2, 3, 4, 5, 6, 7, ..., 29, 30
	Batch size	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Offloading	Partition point	[cloud,cloud], [edge,cloud], [edge,edge]
	Region allocation	C_R^u combinations

(allocating R regions among μ devices)

Motivation

➢ Knob Selection

Category	Knob	Optional Values
Configuration	Frame resolution	240p,360p,480p,540p,720p,900p,1080p
	Frame rate	1, 2, 3, 4, 5, 6, 7, ..., 29, 30
	Batch size	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Offloading	Partition point	[cloud,cloud], [edge,cloud], [edge,edge]
	Region allocation	C_R^u combinations

(allocating R regions among μ devices)

Total knob combination number: $7 \times 30 \times 10 \times 3 \times C_{15}^{3} = 2866500$

Motivation

➢ Knob Selection

Category	Knob	Optional Values
Configuration	Frame resolution	240p,360p,480p,540p,720p,900p,1080p
	Frame rate	1, 2, 3, 4, 5, 6, 7, ..., 29, 30
	Batch size	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Offloading	Partition point	[cloud,cloud], [edge,cloud], [edge,edge]
	Region allocation	C_R^u combinations

(allocating R regions among μ devices)

Total knob combination number: $7 \times 30 \times 10 \times 3 \times C_{15}^{3} = 2866500$

Challenge I: Exponentially complex decision space for joint knob adjustment

Motivation

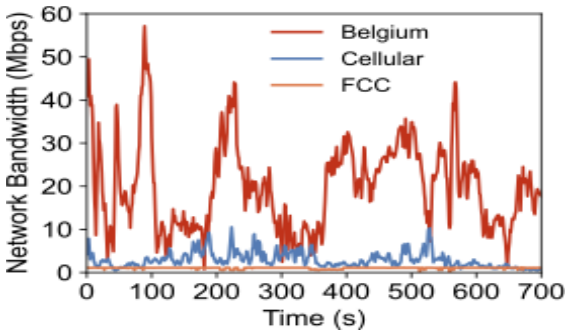
➢ Dynamic Runtime Context

✓ **Resource Context**

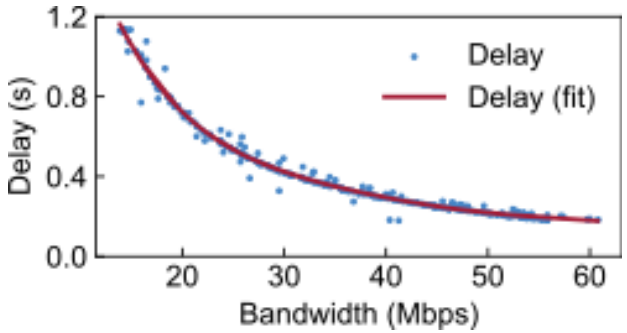
contains hardware / network of the distributed system, representing different **system supply state** during real-time video analytics.

✓ **Task Context**

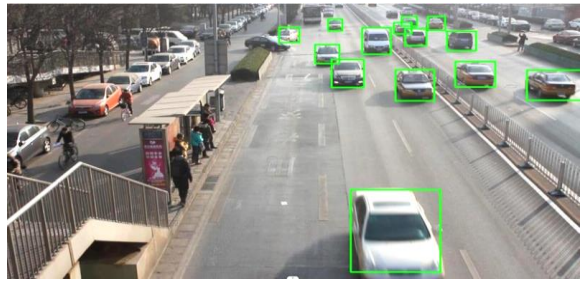
contains number / size / motion of task objects, representing different **system demand state** during real-time video analytics.



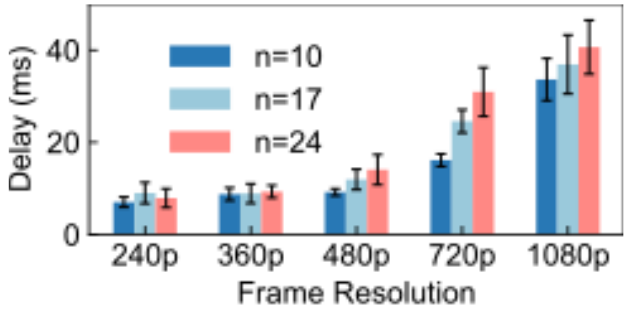
Bandwidth variation



Impact of object variation



Object variation



Impact of object variation

Motivation

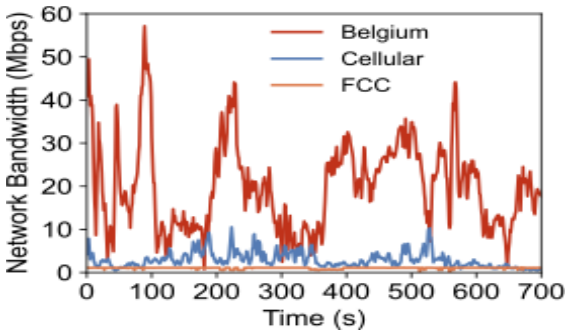
➢ Dynamic Runtime Context

✓ **Resource Context**

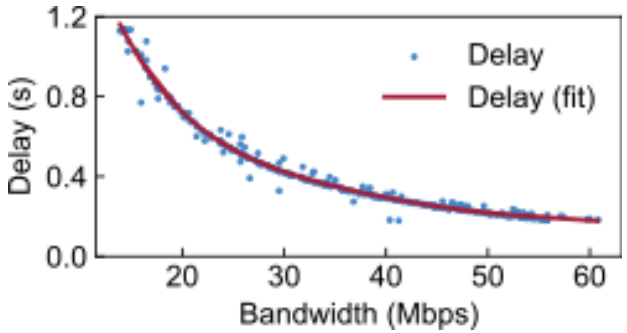
contains hardware / network of the distributed system, representing different **system supply state** during real-time video analytics.

✓ **Task Context**

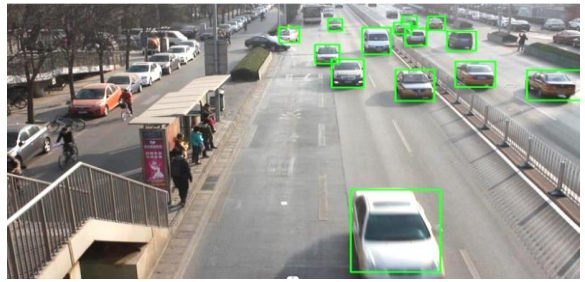
contains number / size / motion of task objects, representing different **system demand state** during real-time video analytics.



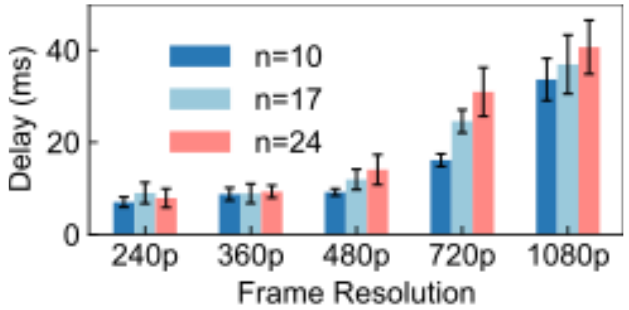
Bandwidth variation



Impact of object variation



Object variation



Impact of object variation

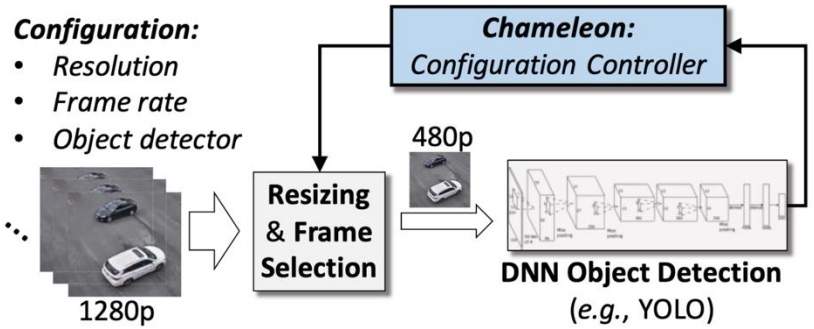
Challenge II: Dynamically fluctuating environment in adaptive knob adjustment

Motivation

➢ Existing Solutions for Knob Adjustment

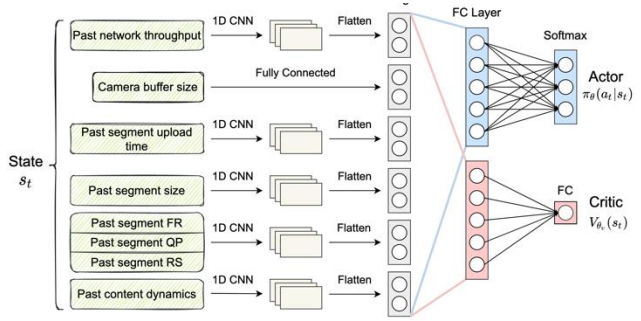
✓ Profiling-based methods:

- VideoStorm [NSDI '17], Chameleon [SIGCOMM '18], ALERT [ATC '20]
- build **off/online profiles** to guide knob tuning
- **introduce extra delays in real-time processing**



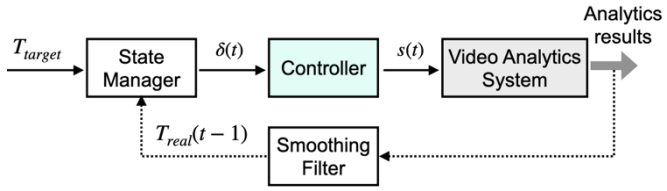
✓ End-to-end learning methods:

- CASVA [INFOCOM '22], Magic-Pipe [Middleware '21], CuttleFish [TPDS '20]
- employs a **deep reinforcement learning (DRL)** method to decide configuration
- **struggles to converge in large decision space**



✓ Negative feedback methods:

- DDS [SIGCOMM '20], FC [Sensys '21], Elf [MobiCom '21]
- Employs a **negative feedback** method to tune knob directly
- **limited to adjusting a single knob**



➢ Problem Formulation

The latency-first optimization object:

$$\begin{cases} L_t = \emptyset_L(\boldsymbol{s}_t, \boldsymbol{\tau}_t) \\ A_t = \emptyset_A(\boldsymbol{s}_t, \boldsymbol{\tau}_t) \end{cases}$$

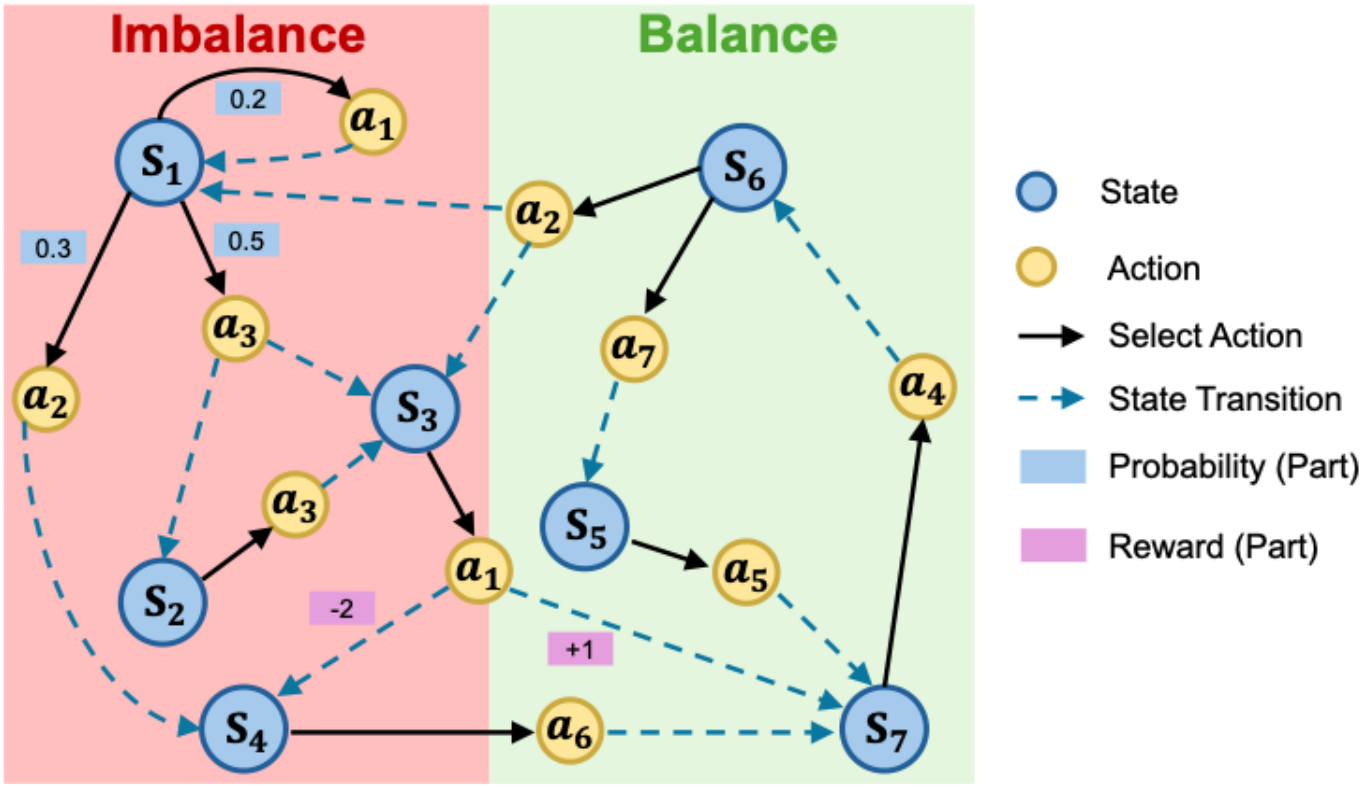
$$\max_{\boldsymbol{\tau}_t} \sum_t A_t, \qquad s.t.\ L_t \le \frac{1}{f_t}, \forall t$$

The output knob adjustment decision:

$$\boldsymbol{\tau}_t = (\tau_t^{k_1}, \tau_t^{k_2}, \dots, \tau_t^{k_n}) \in \prod_{i=1}^{n} \mathcal{X}_i$$

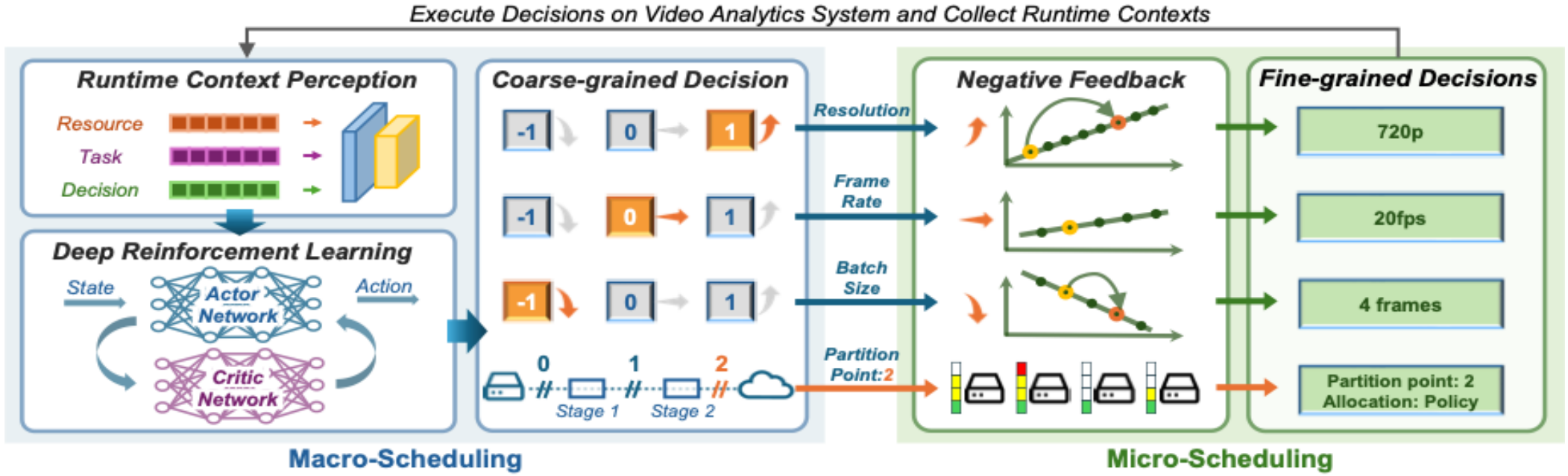
$$\mathcal{X}_i = \{x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(N_i)}\}, \quad \tau_t^{k_i} = x_i^{(j_{i,t})}$$

Decision Space Size: $\sum_{i=1}^{n} N_i$



High-dimensional Markov decision process

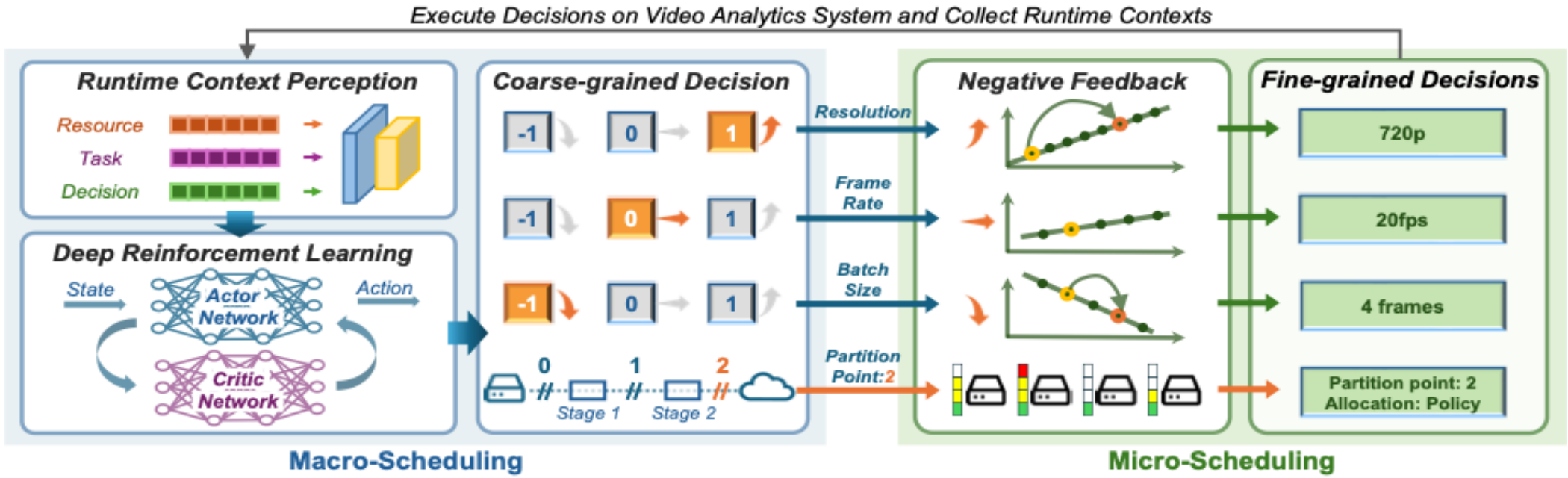
➢ Framework Overview: Hier-EI



- **Hierarchical:** macro-scheduling outputs coarse-grained decisions, micro-scheduling outputs fine-grained decisions
- **Embodied:** act as an embodied intelligence interacting with video analytics systems in collaboration and optimization

Methodology

➢ Framework Overview: Hier-EI

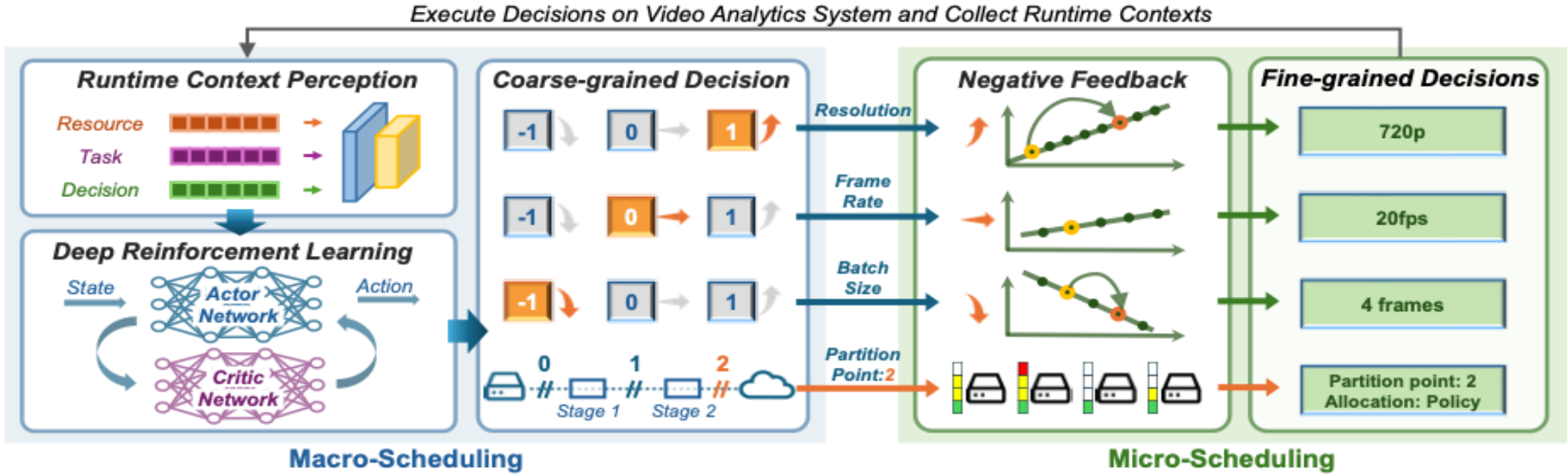


◆ **Hierarchical:** macro-scheduling outputs coarse-grained decisions, micro-scheduling outputs fine-grained decisions

➡ **Complexity Reduction**

◆ **Embodied:** act as an embodied intelligence interacting with video analytics systems in collaboration and optimization

➢ Framework Overview: Hier-EI



- **Hierarchical:** macro-scheduling outputs coarse-grained decisions, micro-scheduling outputs fine-grained decisions → **Complexity Reduction**
- **Embodied:** act as an embodied intelligence interacting with video analytics systems in collaboration and optimization → **Dynamics Adaption**

Methodology

➢ Knob Analysis

Category	Knob	Optional Values
Configuration	Frame resolution	240p,360p,480p,540p,720p,900p,1080p
	Frame rate	1, 2, 3, 4, 5, 6, 7, ..., 29, 30
	Batch size	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Offloading	Partition point	[cloud,cloud], [edge,cloud], [edge,edge]
	Region allocation	C_R^u combinations

Monotonic Knobs

- show **predictable monotonic relationships** with performance metrics
- Adjusting monotonic knobs should **follow monotonic relationship**

Macro-scheduling: adjustment direction
Micro-scheduling: exact values

Non-monotonic Knobs

- show **unpredictable complex relationships** with performance metrics
- Adjusting non-monotonic knobs should **mine the interdependencies** between them

Macro-scheduling: partition point
Micro-scheduling: regions allocation

Methodology

➢ Knob Analysis

Category	Knob	Optional Values
Configuration	Frame resolution	240p,360p,480p,540p,720p,900p,1080p
	Frame rate	1, 2, 3, 4, 5, 6, 7, ..., 29, 30
	Batch size	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Offloading	Partition point	[cloud,cloud], [edge,cloud], [edge,edge]
	Region allocation	C_R^u combinations

Monotonic Knobs

- show **predictable monotonic relationships** with performance metrics
- Adjusting monotonic knobs should **follow monotonic relationship**

Macro-scheduling: adjustment direction
Micro-scheduling: exact values

Non-monotonic Knobs

- show **unpredictable complex relationships** with performance metrics
- Adjusting non-monotonic knobs should **mine the interdependencies** between them

Macro-scheduling: partition point
Micro-scheduling: regions allocation

Methodology

➢ Knob Analysis

Category	Knob	Optional Values
Configuration	Frame resolution	240p,360p,480p,540p,720p,900p,1080p
	Frame rate	1, 2, 3, 4, 5, 6, 7, ..., 29, 30
	Batch size	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Offloading	Partition point	[cloud,cloud], [edge,cloud], [edge,edge]
	Region allocation	C_R^u combinations

Monotonic Knobs

- show **predictable monotonic relationships** with performance metrics
- Adjusting monotonic knobs should **follow monotonic relationship**

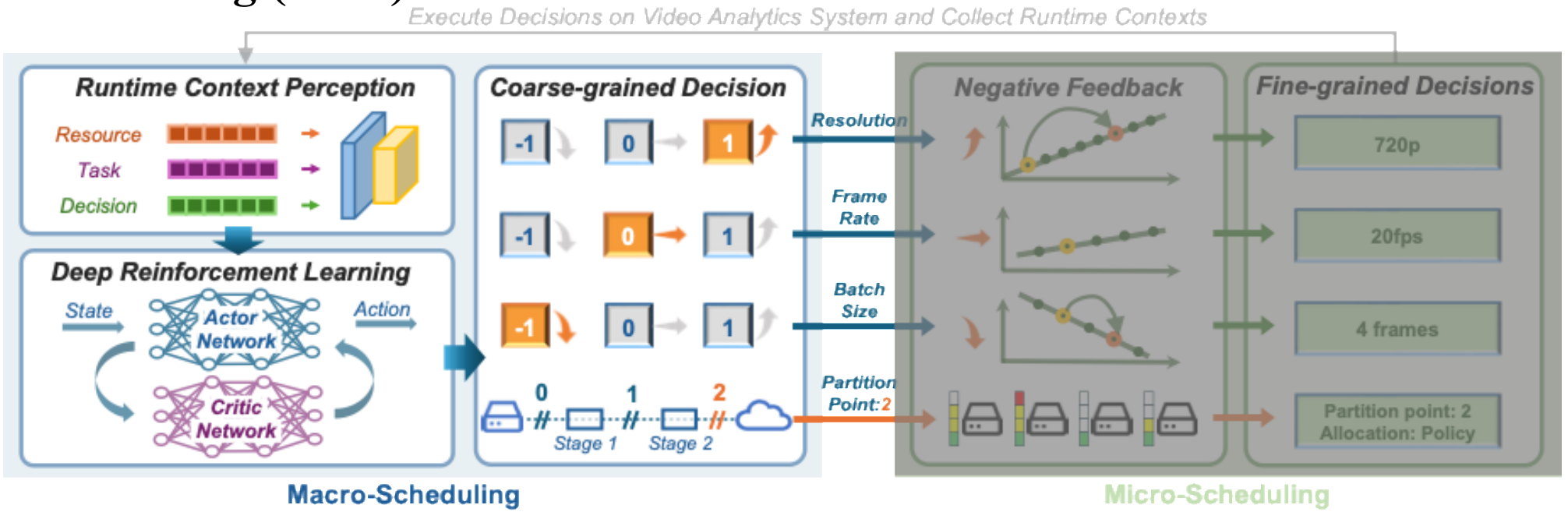
Macro-scheduling: adjustment direction
Micro-scheduling: exact values

Non-monotonic Knobs

- show **unpredictable complex relationships** with performance metrics
- Adjusting non-monotonic knobs should **mine the interdependencies** between them

Macro-scheduling: partition point
Micro-scheduling: regions allocation

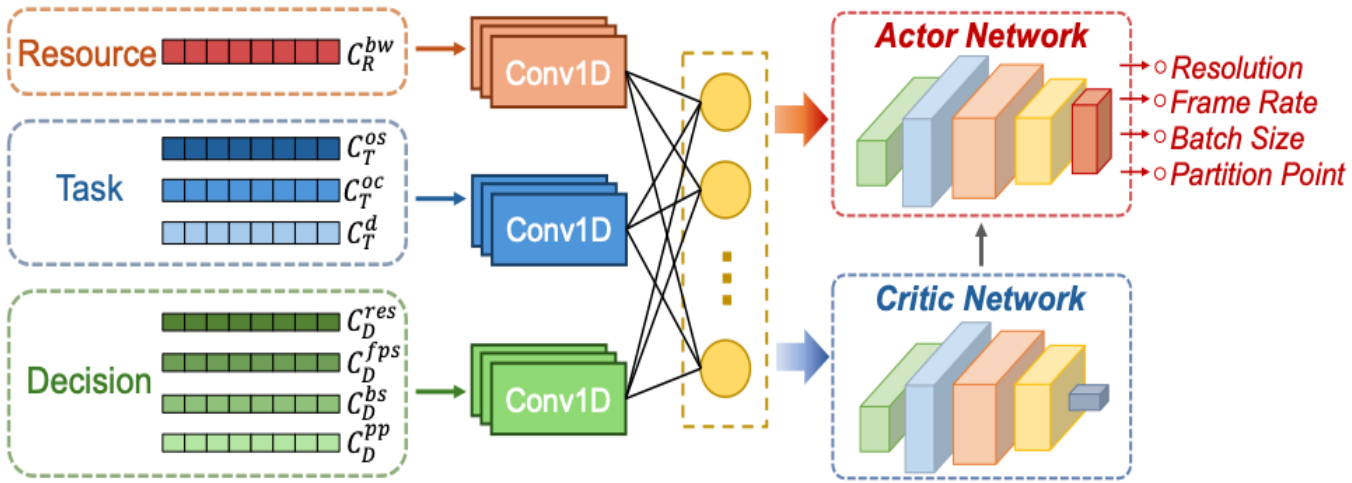
➢ Macro Scheduling (DRL)



Macro-Scheduling:

- Employ an **Actor-Critic** Deep Reinforcement Learning Model
- Output **coarse-grained decisions** to guide fine-tune in micro-scheduling
- Perform with a relatively large interval to perceive runtime context globally

➢ Macro Scheduling (DRL)



State (Runtime Context Perception)

- ✓ Perceive three runtime context: **resource** (**supply**), task (**demand**), decision (**history**)
- ✓ Use **independent Conv-1Ds** to extract underlying hidden features.

$$\boldsymbol{s}_t = (C_{R,t}^{bw}, C_{T,t}^{os}, C_{T,t}^{os}, C_{T,t}^{d}, C_{D,t}^{res}, C_{D,t}^{fps}, C_{D,t}^{bs}, C_{D,t}^{pp})$$

Action (Coarse-grained Decisions)

- ✓ For **monotonic knobs**, output adjustment directions: -1 (decrease), 0 (unchanged), 1 (increase)
- ✓ For **non-monotonic knobs**, output 0, 1, 2 representing pipeline partition point

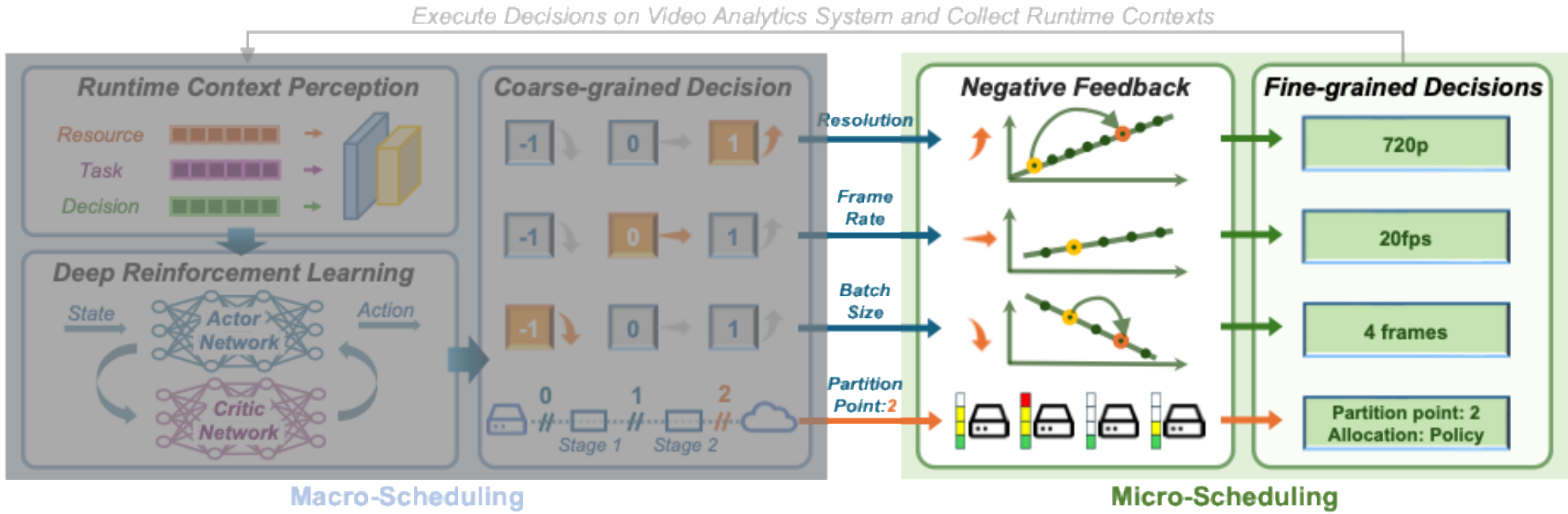
$$\boldsymbol{a}_t = \boldsymbol{\xi}_t = (\xi_t^{res}, \xi_t^{fps}, \xi_t^{bs}, \xi_t^{pp})$$

Reward (SLO-Driven QoE)

$$\boldsymbol{r}_t = \begin{cases} \beta \cdot \dfrac{1}{\max(\Delta_{L,t}, \vartheta)} + A_t, & \Delta_{L,t} \geq 0 \\ \max(\gamma \cdot \Delta_{L,t}, \theta), & \Delta_{L,t} < 0 \end{cases}$$

$$\Delta_{L,t} = \alpha \cdot \frac{1}{f_t} - L_t$$

➢ Micro Scheduling (NF)



Micro-Scheduling:

- Employs ***n* independent negative feedback methods** to fine-tune all knobs.
- Output **fine-grained decisions** to execute in video analytics systems.
- Perform with a relatively small interval to response in real time.

➢ Micro Scheduling (NF)

NF for Monotonic Knobs

AIMD-based linear feedback adjustment

additive increase and multiplicative decrease

NF for Non-monotonic Knobs

Multi-object feedback adjustment

based on historical task records

Fine-grained decision: $\tau_t = ($ $)$

➢ Micro Scheduling (NF)

NF for Monotonic Knobs

AIMD-based linear feedback adjustment

additive increase and multiplicative decrease

$$j_{i,t} = \begin{cases} \max\left(j_{i,t-1} + 1, \|\mathcal{X}_i\|\right), & \xi_t^{k_i} = 1 \\ j_{i,t-1}, & \xi_t^{k_i} = 0 \\ \left\lfloor \frac{j_{i,t-1}}{2} \right\rfloor, & \xi_t^{k_i} = -1 \end{cases}$$

$$\tau_t^{k_i} = x_i^{(j_{i,t})}, \qquad x_i^{(i_{i,t})} \in \mathcal{X}_i$$

NF for Non-monotonic Knobs

Multi-object feedback adjustment

based on historical task records

Fine-grained decision: $\boldsymbol{\tau}_t = ($ $)$

➢ Micro Scheduling (NF)

NF for Monotonic Knobs

AIMD-based linear feedback adjustment

additive increase and multiplicative decrease

$$j_{i,t} = \begin{cases} \max\left(j_{i,t-1} + 1, \|\mathcal{X}_i\|\right), & \xi_t^{k_i} = 1 \\ j_{i,t-1}, & \xi_t^{k_i} = 0 \\ \left\lfloor \frac{j_{i,t-1}}{2} \right\rfloor, & \xi_t^{k_i} = -1 \end{cases}$$

$$\tau_t^{k_i} = x_i^{(j_{i,t})}, \qquad x_i^{(i_{i,t})} \in \mathcal{X}_i$$

NF for Non-monotonic Knobs

Multi-object feedback adjustment

based on historical task records

Fine-grained decision: $\boldsymbol{\tau}_t = (\ \tau_t^{res}, \tau_t^{fps}, \tau_t^{bs},\ \ \ \ \ \ \ \ )$

➢ Micro Scheduling (NF)

NF for Monotonic Knobs

AIMD-based linear feedback adjustment

additive increase and multiplicative decrease

$$j_{i,t} = \begin{cases} \max\left(j_{i,t-1} + 1, \|\mathcal{X}_i\|\right), & \xi_t^{k_i} = 1 \\ j_{i,t-1}, & \xi_t^{k_i} = 0 \\ \left\lfloor \frac{j_{i,t-1}}{2} \right\rfloor, & \xi_t^{k_i} = -1 \end{cases}$$

$$\tau_t^{k_i} = x_i^{(j_{i,t})}, \qquad x_i^{(i_{i,t})} \in \mathcal{X}_i$$

NF for Non-monotonic Knobs

Multi-object feedback adjustment

based on historical task records

$$\tau_t^{pp} = \xi_t^{pp}$$

$$\tau_t^{ra} = \begin{cases} \{\sigma_t^0 = R_t\}, & \xi_t^{pp} \in \{0,1\} \\ \left\{ \sigma_t^\mu = \frac{\frac{1}{L_{t-1}^\mu}}{\sum_{\mu=1}^{U} \frac{1}{L_{t-1}^\mu}} \cdot R_t \right\}, & \xi_t^{pp} \in \{2\} \end{cases}$$

Fine-grained decision: $\boldsymbol{\tau}_t = (\ \tau_t^{res}, \tau_t^{fps}, \tau_t^{bs},\ \ \ \ \ \ \ \ )$

➢ Micro Scheduling (NF)

NF for Monotonic Knobs

AIMD-based linear feedback adjustment

additive increase and multiplicative decrease

$$j_{i,t} = \begin{cases} \max\left(j_{i,t-1} + 1, \|\mathcal{X}_i\|\right), & \xi_t^{k_i} = 1 \\ j_{i,t-1}, & \xi_t^{k_i} = 0 \\ \left\lfloor \frac{j_{i,t-1}}{2} \right\rfloor, & \xi_t^{k_i} = -1 \end{cases}$$

$$\tau_t^{k_i} = x_i^{(j_{i,t})}, \qquad x_i^{(i_{i,t})} \in \mathcal{X}_i$$

NF for Non-monotonic Knobs

Multi-object feedback adjustment

based on historical task records

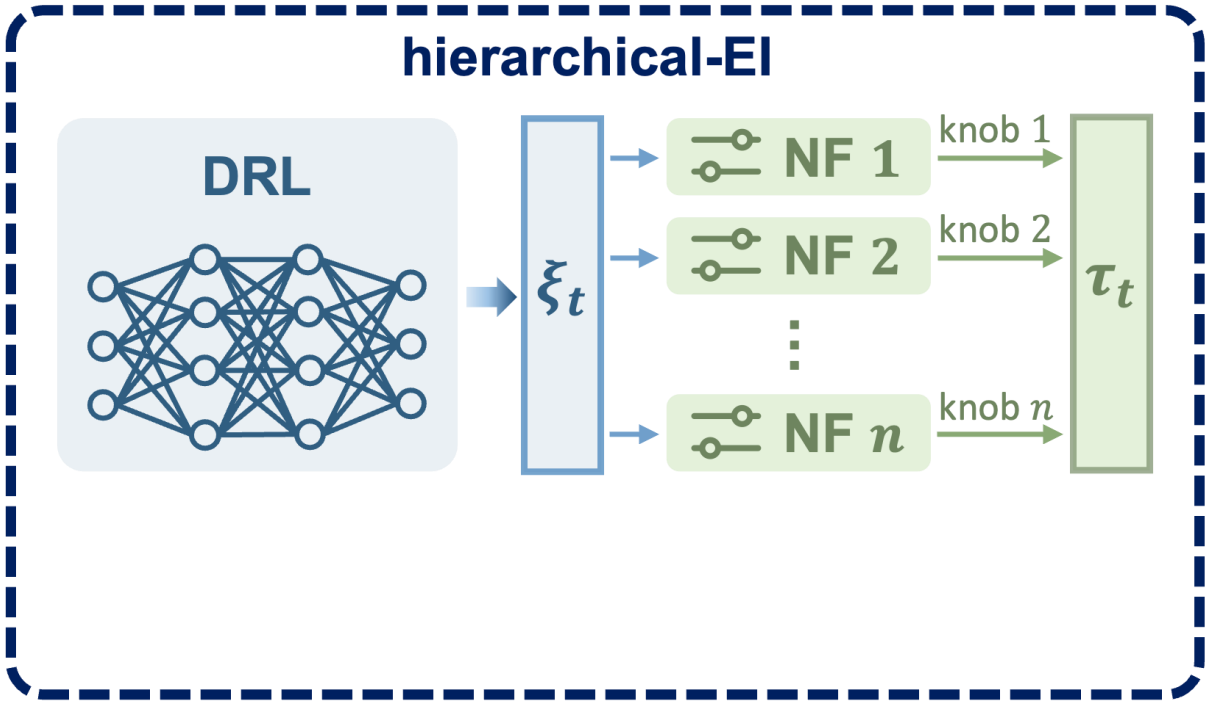
$$\tau_t^{pp} = \xi_t^{pp}$$

$$\tau_t^{ra} = \begin{cases} \{\sigma_t^0 = R_t\}, & \xi_t^{pp} \in \{0,1\} \\ \left\{ \sigma_t^\mu = \frac{\frac{1}{L_{t-1}^\mu}}{\sum_{\mu=1}^{U} \frac{1}{L_{t-1}^\mu}} \cdot R_t \right\}, & \xi_t^{pp} \in \{2\} \end{cases}$$

Fine-grained decision: $\boldsymbol{\tau}_t = (\tau_t^{res}, \tau_t^{fps}, \tau_t^{bs}, \tau_t^{pp}, \tau_t^{ra})$

➢ Hierarchical Collaboration Mechanism

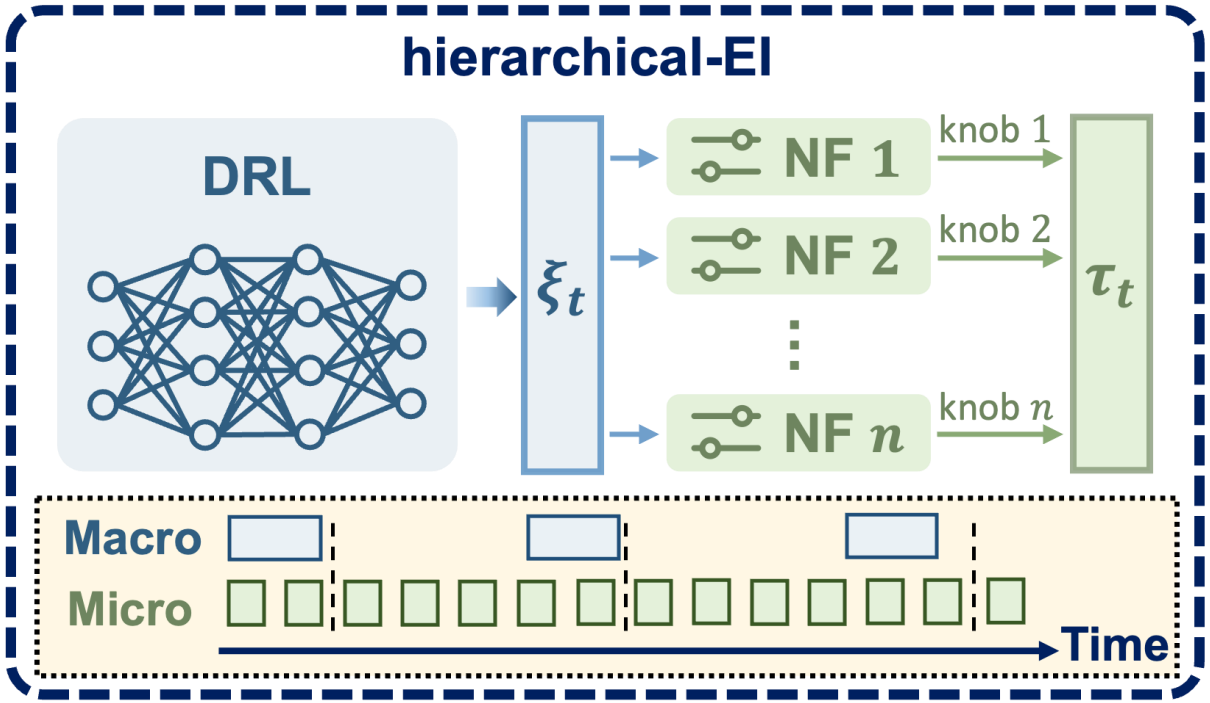
- ✓ **Macro-scheduling** uses an Actor-Critic DRL Model to output **coarse-grained decisions**
- ✓ **Micro-scheduling** uses n independent negative feedback methods to output **fine-grained decisions**.



Hierarchical Collaboration Mechanism

➢ Hierarchical Collaboration Mechanism

- ✓ **Macro-scheduling** uses an Actor-Critic DRL Model to output **coarse-grained decisions**
- ✓ **Micro-scheduling** uses n independent negative feedback methods to output **fine-grained decisions**.
- ✓ Macro and micro **collaborate asynchronously** to display respective advantages
- ✓ Micro-scheduling uses **the latest updated coarse-grained** decision to compute

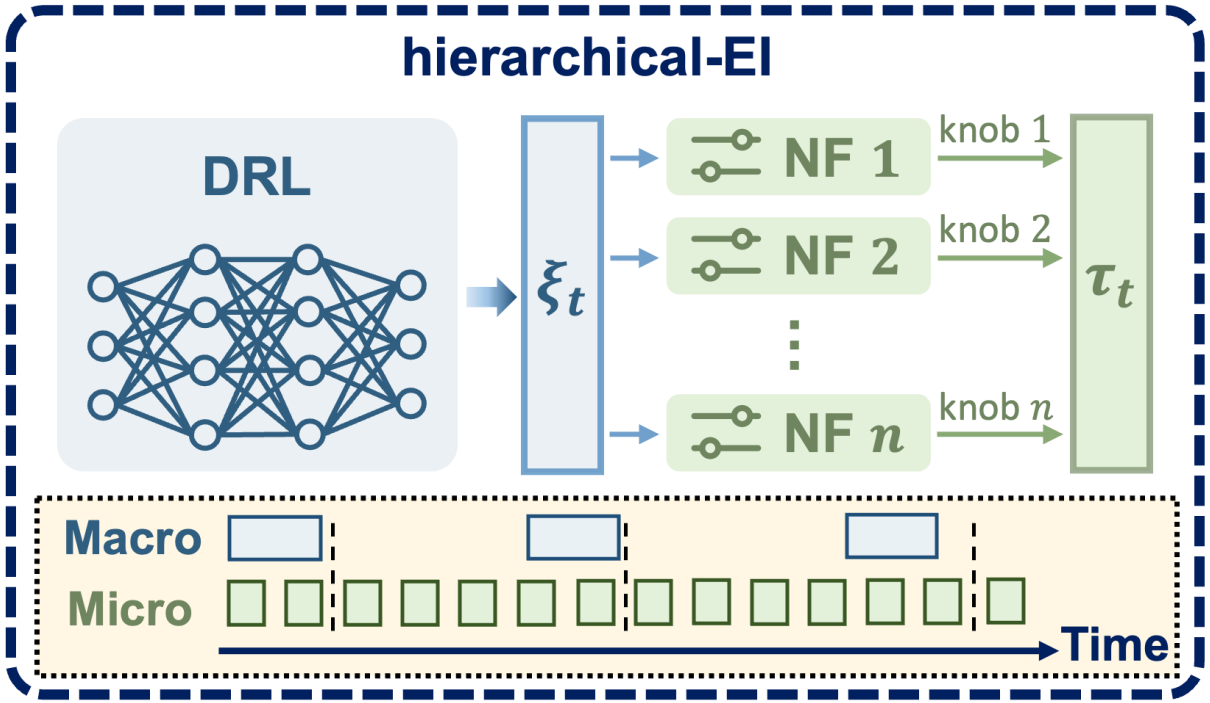


Hierarchical Collaboration Mechanism

Methodology

➢ Hierarchical Collaboration Mechanism

- ✓ **Macro-scheduling** uses an Actor-Critic DRL Model to output **coarse-grained decisions**
- ✓ **Micro-scheduling** uses n independent negative feedback methods to output **fine-grained decisions**.
- ✓ Macro and micro **collaborate asynchronously** to display respective advantages
- ✓ Micro-scheduling uses **the latest updated coarse-grained** decision to compute

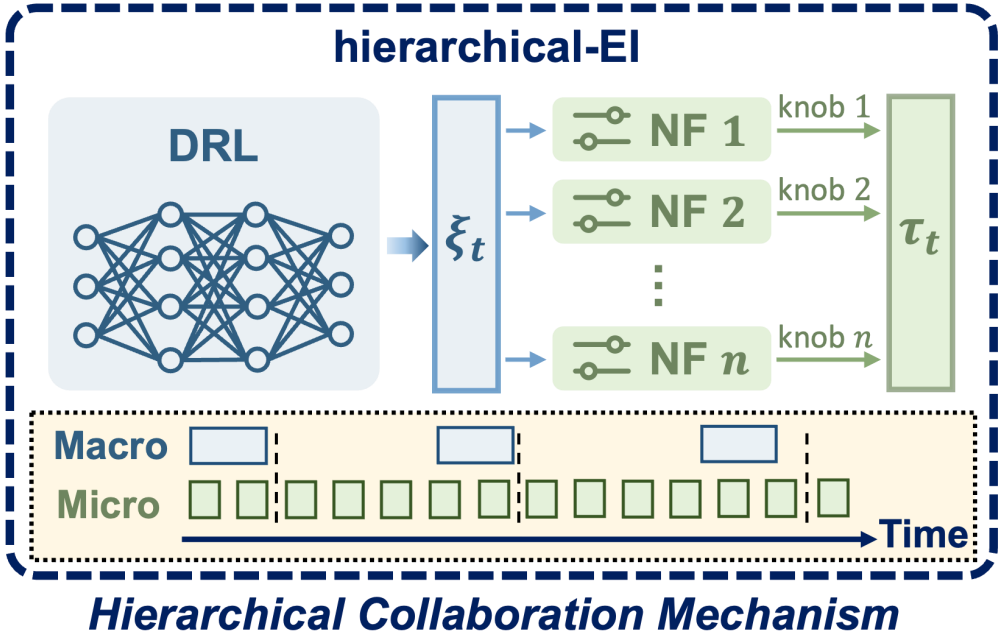


Hierarchical Collaboration Mechanism

Decompose decision space: $\sum_{i=1}^{n} N_i \rightarrow 3^n$ (1e5 → 81)

➢ Embodied Feedback Mechanism

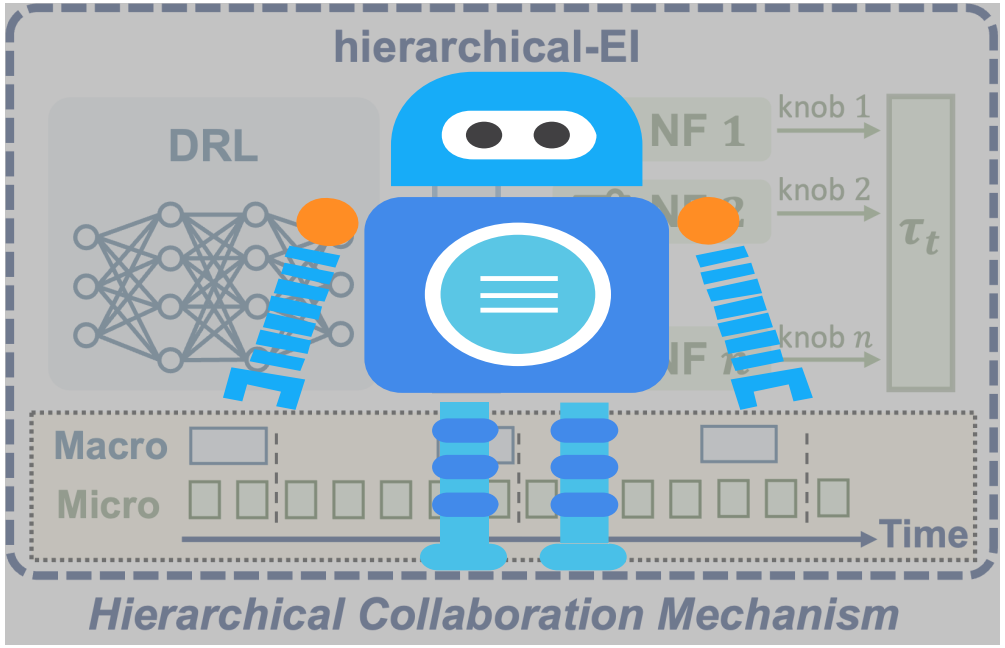
Inspired by robot control, treat two phases as a single **embodied intelligence** to optimize through **interacting with the environment**



Hierarchical Collaboration Mechanism

➢ Embodied Feedback Mechanism

Inspired by robot control, treat two phases as a single **embodied intelligence** to optimize through **interacting with the environment**



Methodology

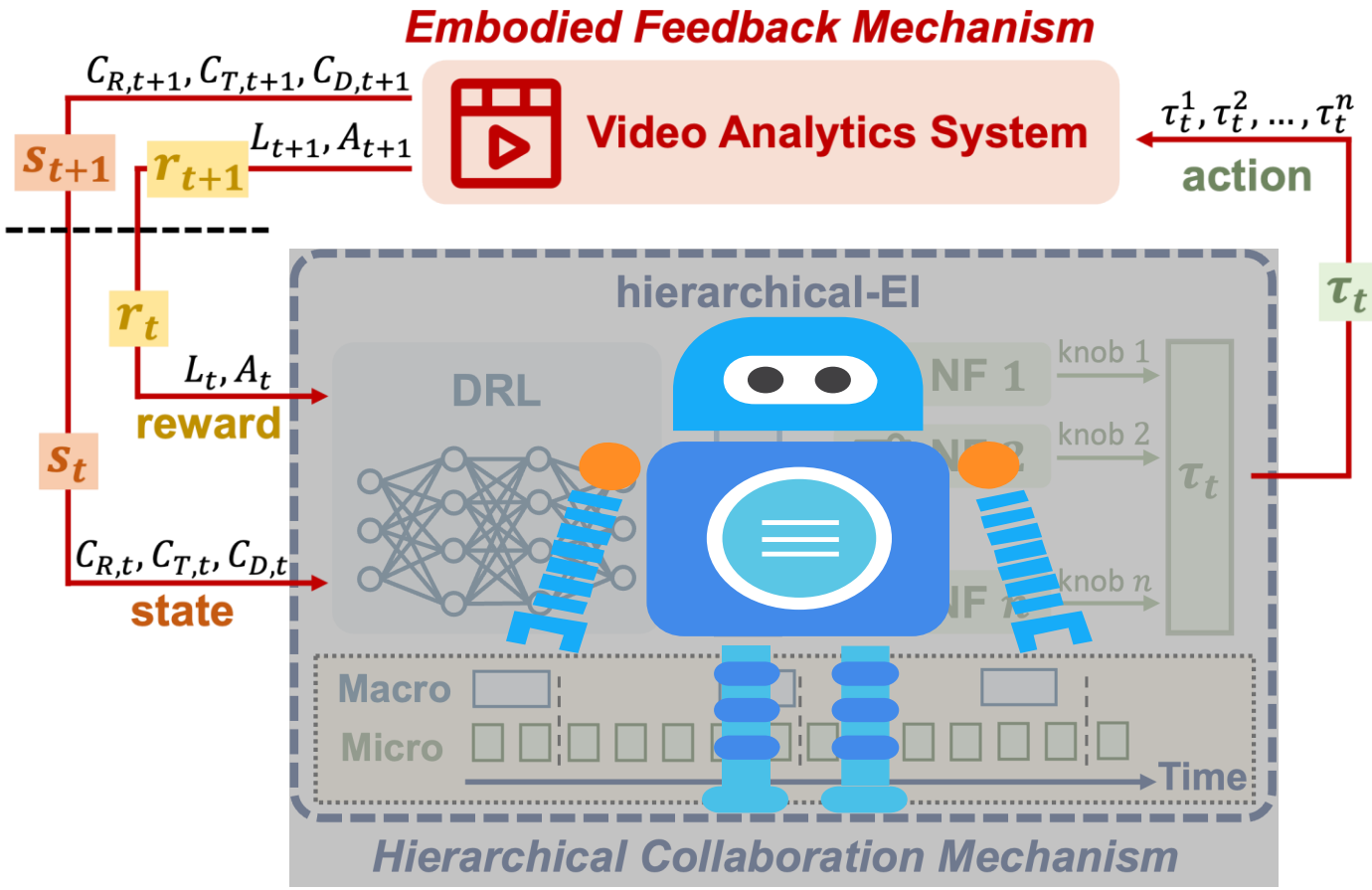
➢ Embodied Feedback Mechanism

Inspired by robot control, treat two phases as a single **embodied intelligence** to optimize through **interacting with the environment**

① Perform fine-grained decisions in video analytics system to guide **task execution**

② Compute DRL reward with **performance results** and optimize model parameters

③ Perceive updated **runtime context** and perform next scheduling cycle



Implementation

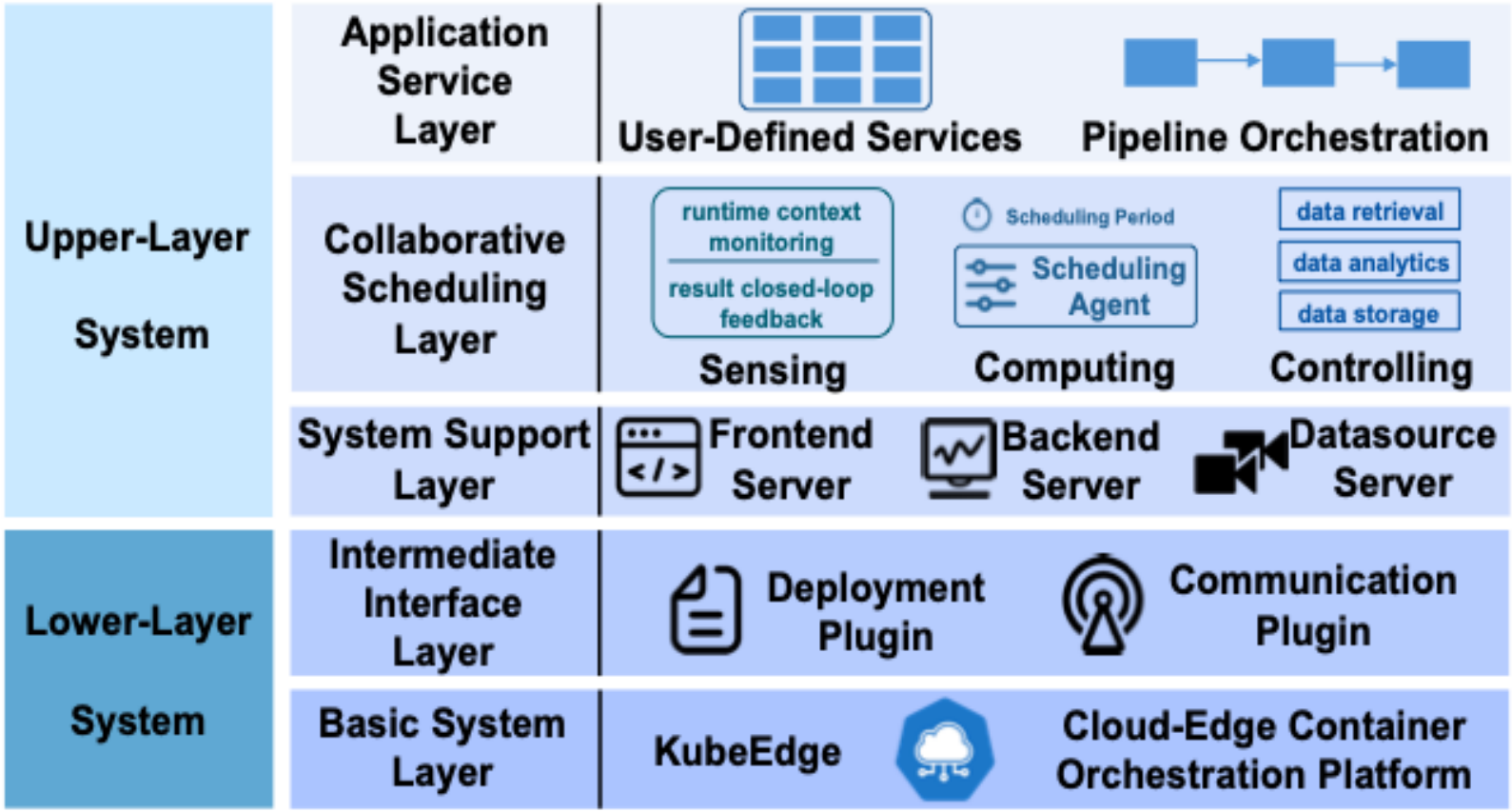
➢ Prototype System Overview

Dayu: a container-based analytics system based on KubeEdge

Dayu

Dayu is named from "大禹"

- ✓ Deployable across heterogeneous devices
- ✓ Flexible with various pipelines orchestration
- ✓ Support fine-grained user-defined scheduling



Provide infrastructure for cloud-edge collaborative stream data analytics

Dayu homepage: https://dayu-autostreamer.github.io

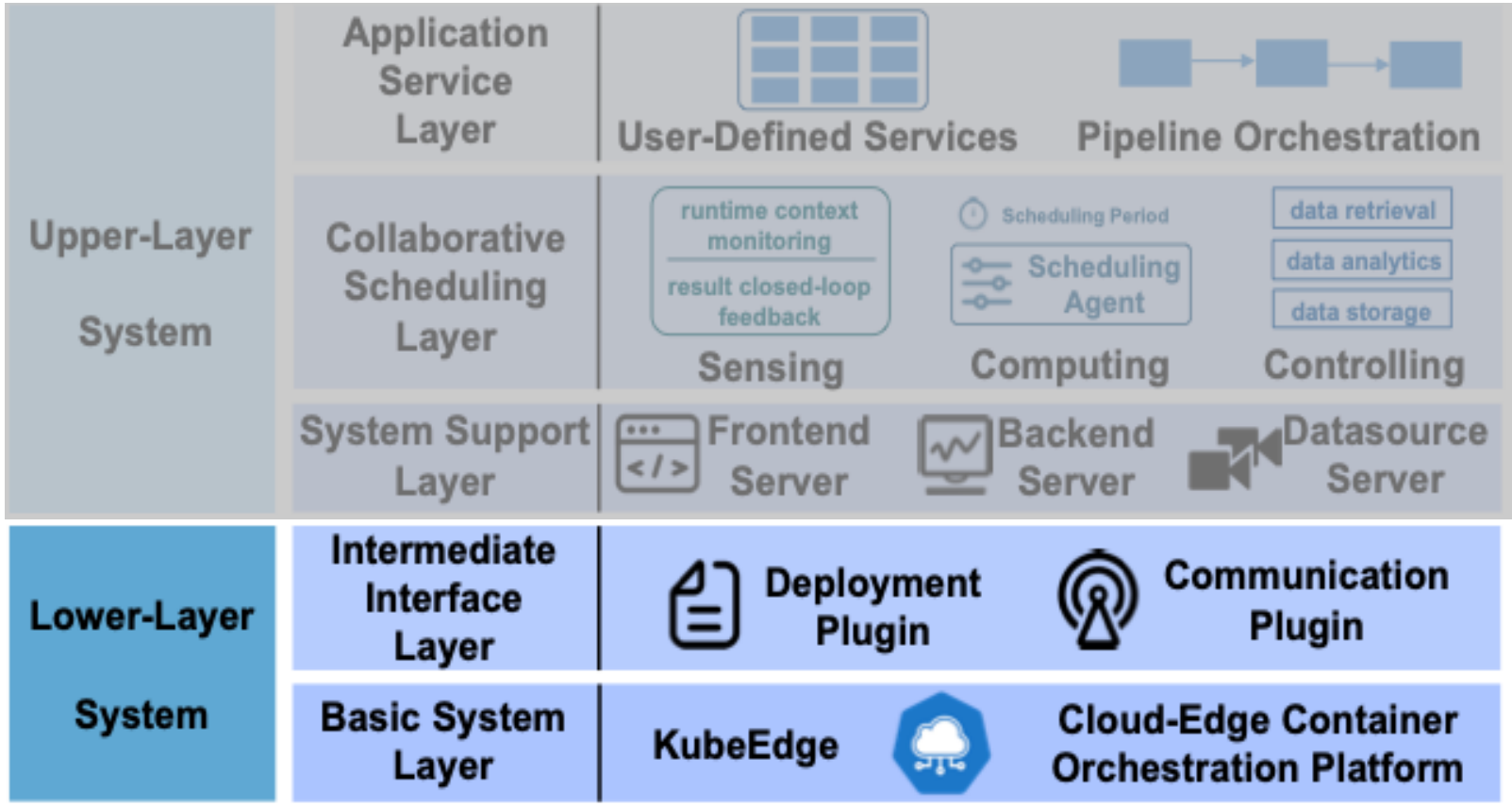
Dayu Repository: https://github.com/dayu-autostreamer/dayu

Implementation

➢ Lower-Layer System

Dayu

- ✓ Provide **container orchestration infrastructures** on distributed cloud-edge systems based on KubeEdge
- ✓ Enable **service installation** and **cross-device communication** with customized Sedna and EdgeMesh



Dayu homepage: https://dayu-autostreamer.github.io

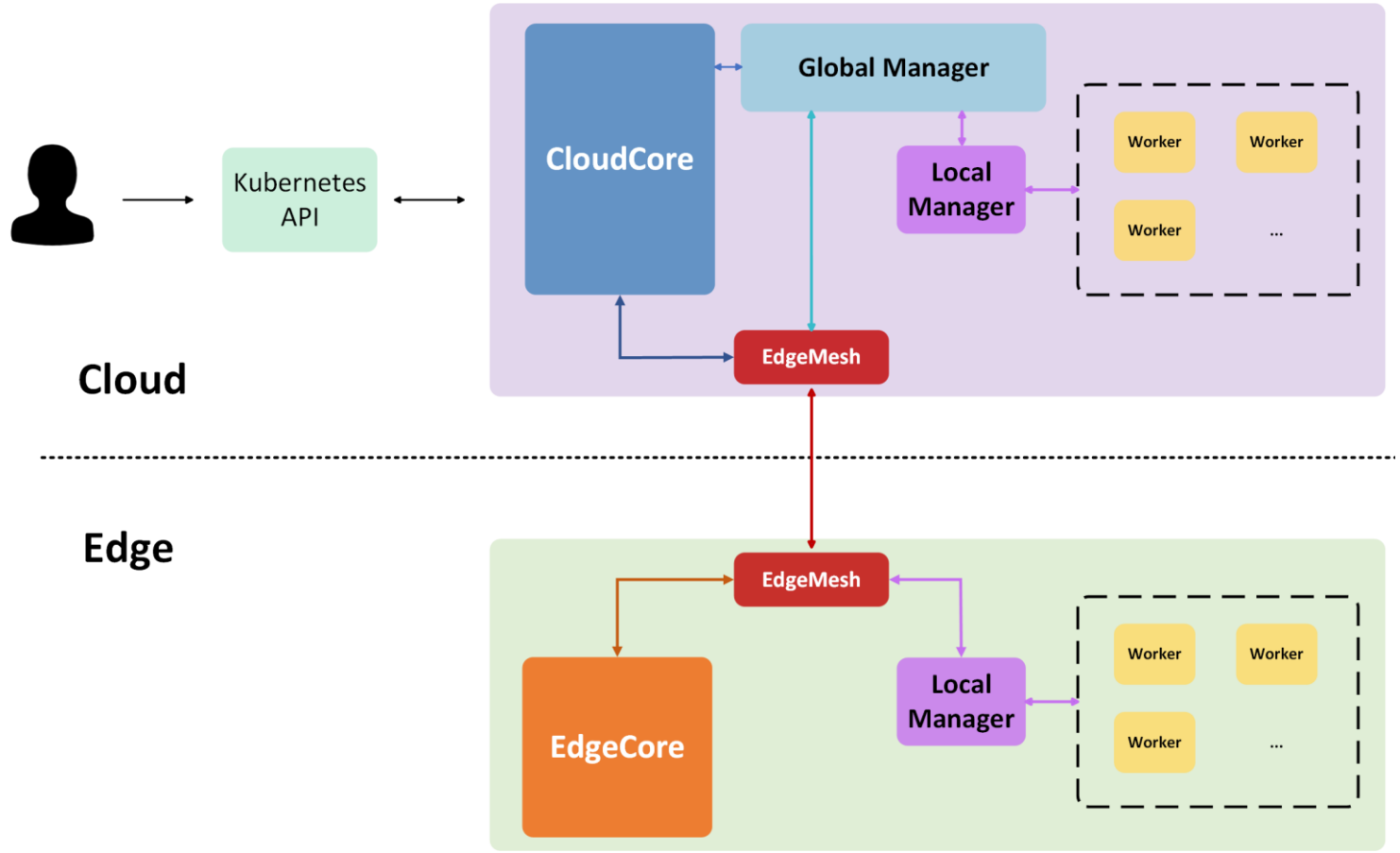
Dayu Repository: https://github.com/dayu-autostreamer/dayu

Implementation

➢ Lower-Layer System

Dayu

- ✓ Provide **container orchestration infrastructures** on distributed cloud-edge systems based on KubeEdge
- ✓ Enable **service installation** and **cross-device communication** with customized Sedna and EdgeMesh



Dayu homepage: https://dayu-autostreamer.github.io

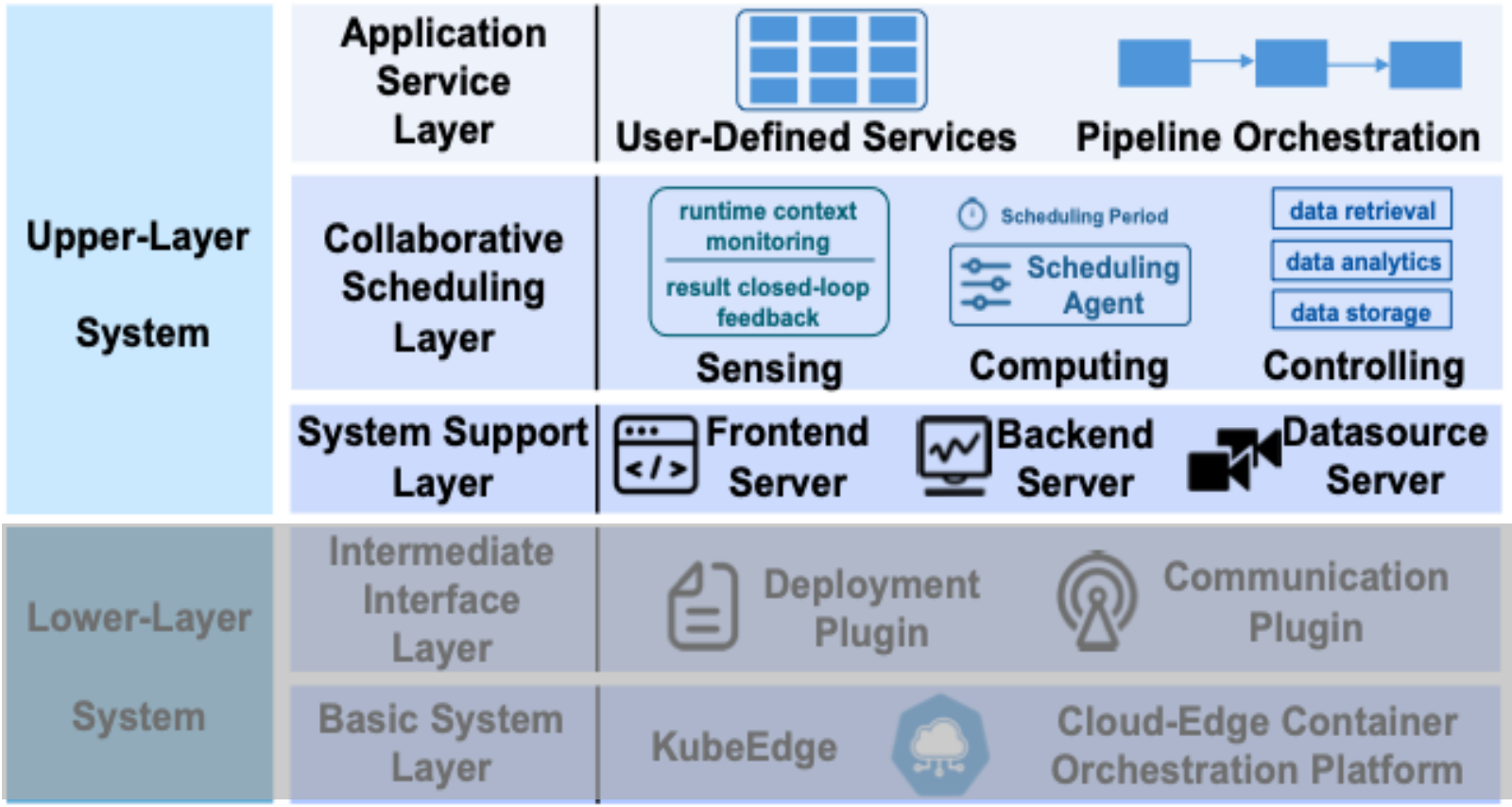
Dayu Repository: https://github.com/dayu-autostreamer/dayu

Implementation

➢ Upper-Layer System

Dayu

- ✓ Support **fine-grained processing, monitoring and scheduling** of video analytics pipelines.
- ✓ Support **user-defined application** with service orchestration.
- ✓ Scalable to **different core operation implementation** with hook functions



Dayu homepage: https://dayu-autostreamer.github.io

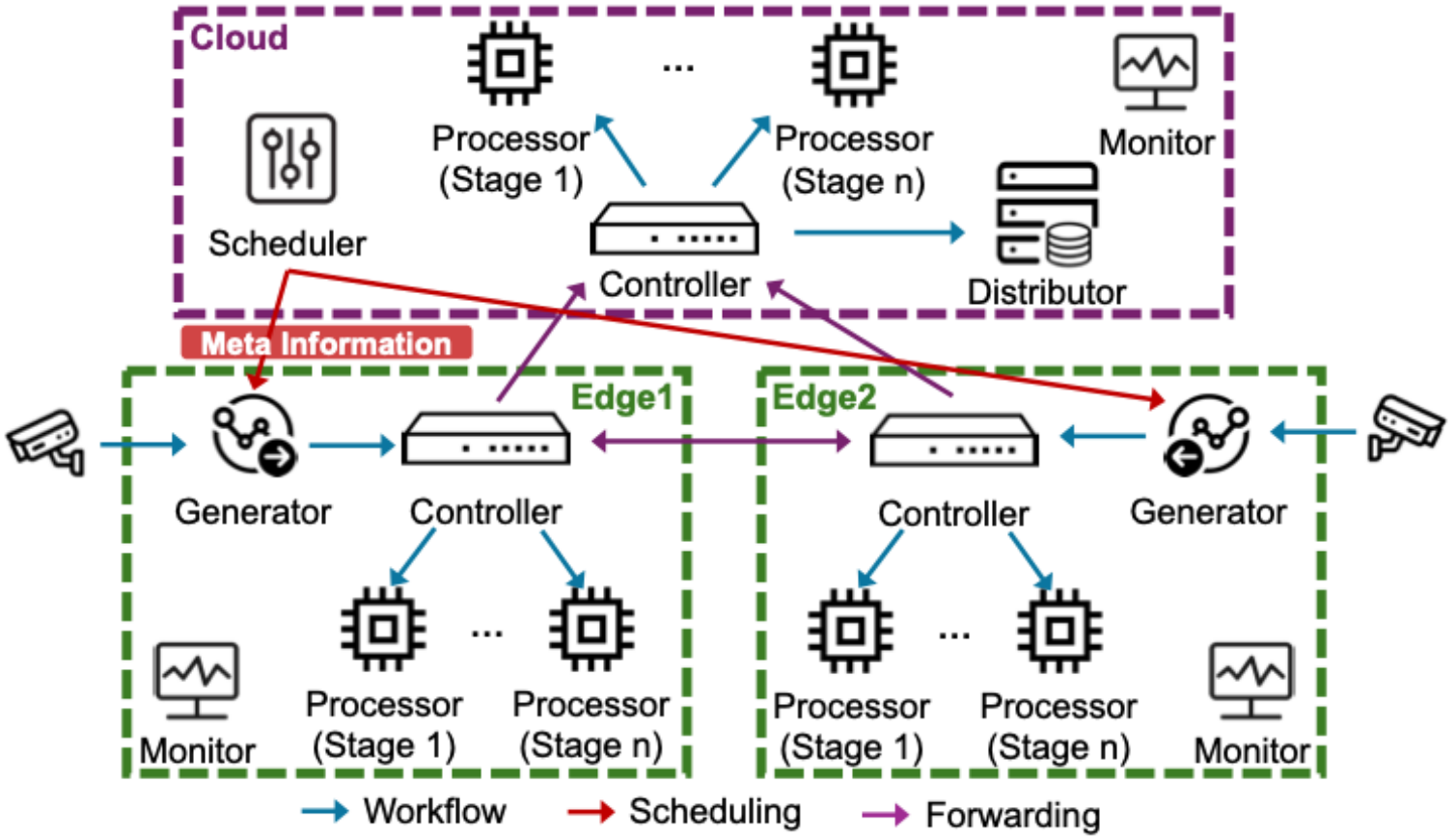
Dayu Repository: https://github.com/dayu-autostreamer/dayu

Implementation

➢ Upper-Layer System

Dayu

- ✓ Support **fine-grained processing, monitoring and scheduling** of video analytics pipelines.
- ✓ Support **user-defined application** with service orchestration.
- ✓ Scalable to **different core operation implementation** with hook functions



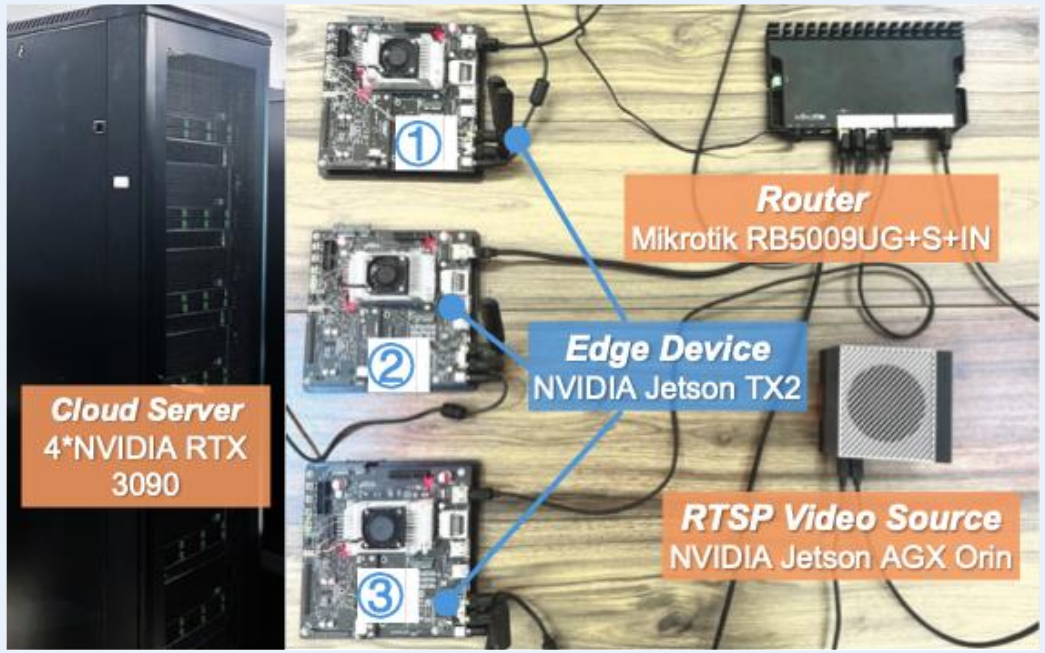
Dayu homepage: https://dayu-autostreamer.github.io

Dayu Repository: https://github.com/dayu-autostreamer/dayu

Evaluation

➢ Real-world Testbed

- ✓ **Cloud Side:** Server with **NVIDIA RTX 3090**
- ✓ **Edge Side:** 3 **NVIDIA Jetson TX2** in a LAN
- ✓ **Data Source:** RTSP with **NVIDIA Jetson Orin**
- ✓ **Network:** replaying **Belgium Dataset**



➢ Evaluation Scenarios

- ✓ **S1:** **stable** network bandwidth / **sparse** task objects
- ✓ **S2:** **stable** network bandwidth / **dense** task objects
- ✓ **S3:** **unstable** network bandwidth / **sparse** task objects
- ✓ **S4:** **unstable** network bandwidth / **dense** task objects

Imbalance Degree: S1 ☞ S4 increase

➢ Evaluation Metrics

- ✓ **Latency compliance rate:** long-term balance status
- ✓ **P95 latency:** service-level objective performance under extreme load conditions
- ✓ **Average accuracy:** overall processing quality

Evaluation

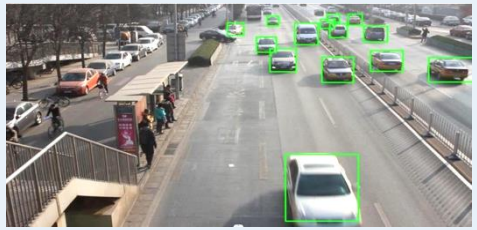
➢ Video analytics applications

1. Road Surveillance Application

- ✓ **Dataset:** Video from UA-DETRAC (total 331 minutes)
- ✓ **Ground Truth:** Annotations in dataset

2. Pedestrian Monitoring Application

- ✓ **Dataset:** Video from YouTube (total 275 minutes)
- ✓ **Ground Truth:** with "golden configuration"



Source → Car Detection → Result

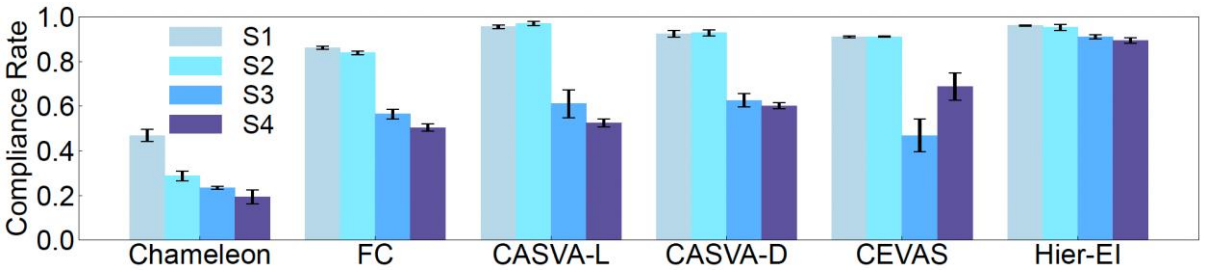
Source → Face Detection → Gender Classification → Result

➢ Baselines

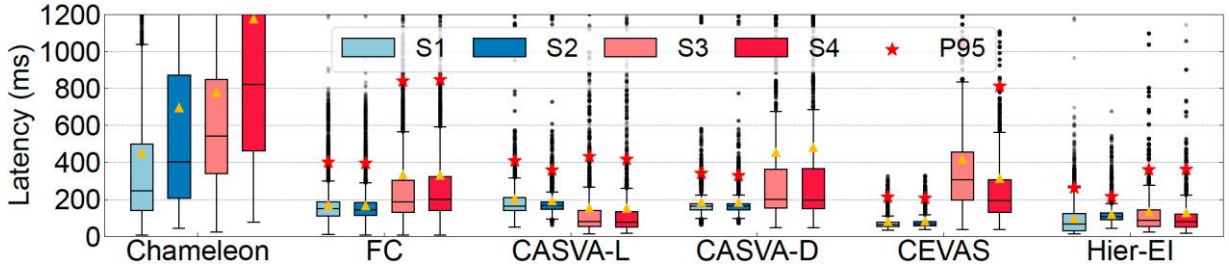
- ✓ **Chameleon [SIGCOMM '18]:** **Online profiling** to adjust video configurations
- ✓ **CEVAS [MMSys '21]:** **Offline profiling** to decide cloud-edge partition point
- ✓ **FC [SenSys '21]:** **AIMD-based negative feedback** to adjust frame resolution
- ✓ **CASVA [INFOCOM '22]:** **End-to-end DRL model (PPO)** to select video configurations
  [latency-first mode (**CASVA-L**) and delivery-first mode (**CASVA-D**)]

Evaluation

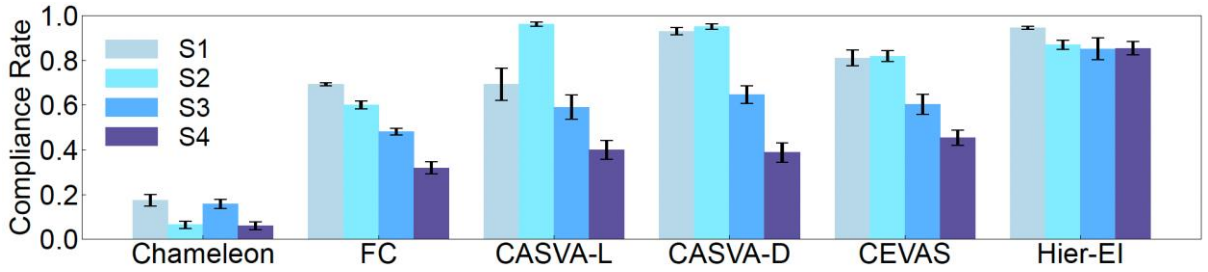
➢ Overall Performance



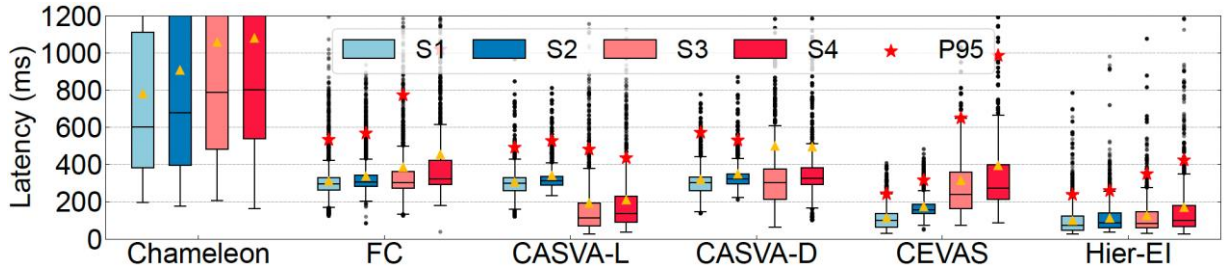
Latency compliance of road surveillance



P95 task latency distribution of road surveillance



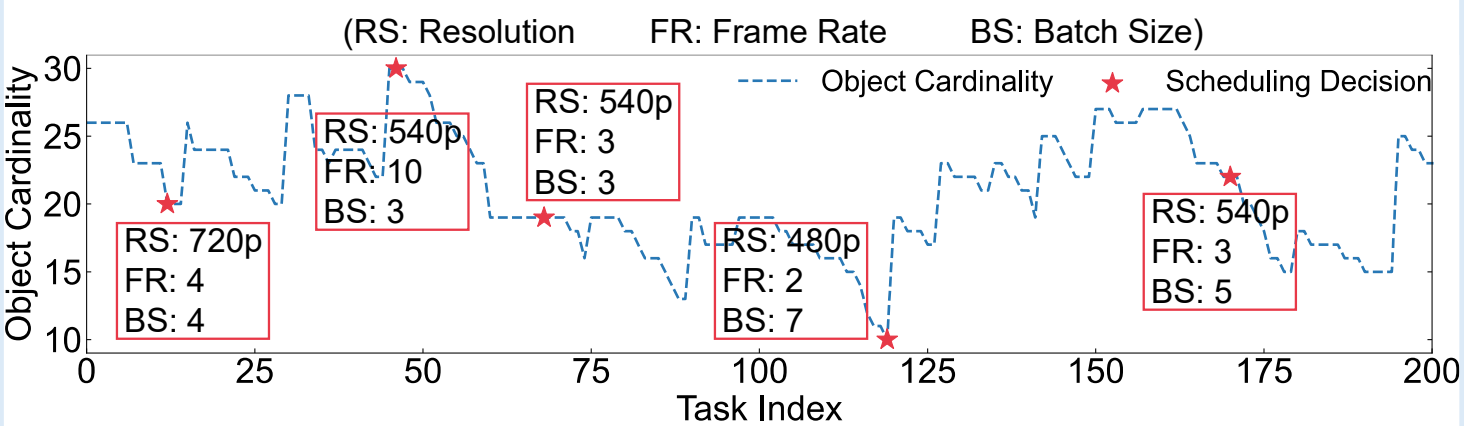
Latency compliance of pedestrian monitoring



P95 task latency distribution of pedestrian monitoring

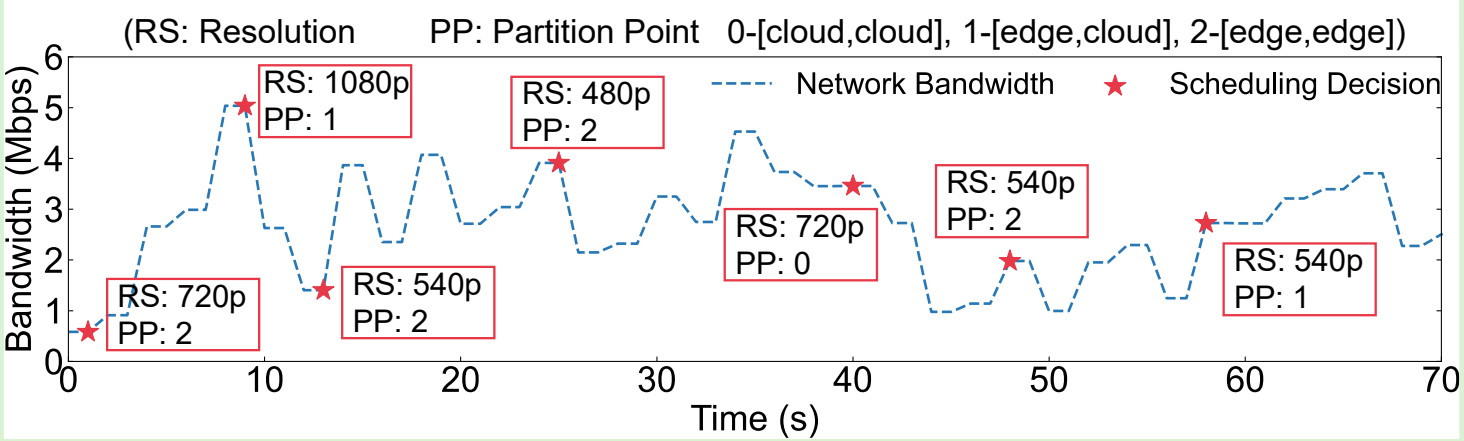
- ✓ Hier-EI has a **2×** improvement on latency compliance and a **56.26%** reduction on P95 latency, mitigating the imbalance in real time while maintaining high QoE.
- ✓ Hier-EI improves latency compliance by **3.6×** and reduce P95 latency by **67.4%** (**306ms**) in extreme scenario **S4**.
- ✓ Hier-EI has **a low CV of 4.59%** in latency compliance across all scenarios and applications.

➢ Case Study for Adaptive Scheduling



Task Context Adaption

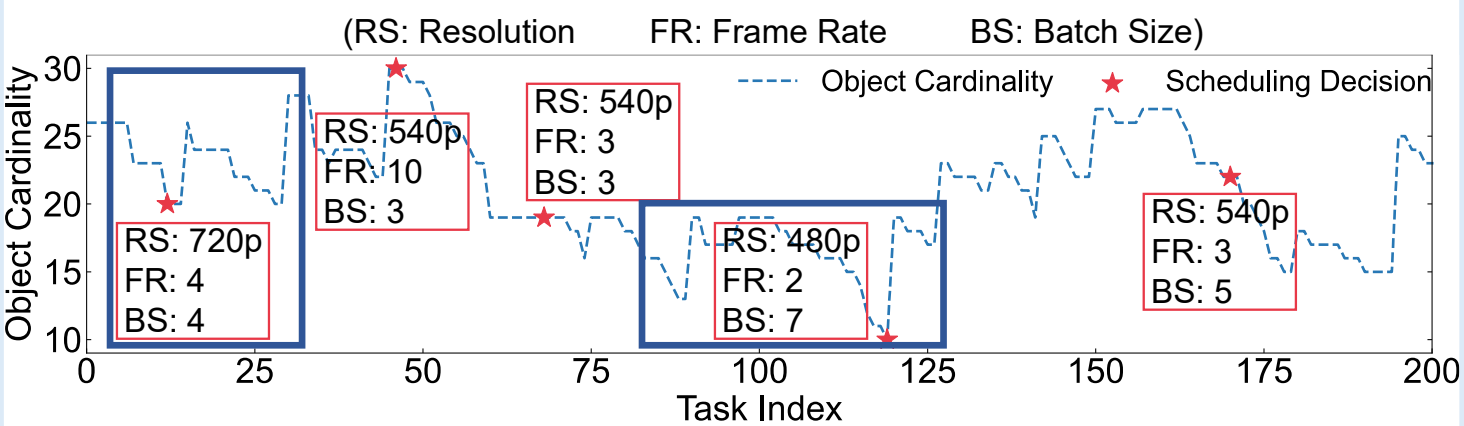
- ✓ Hier-EI reduces frame resolution during **high object density** to lower computational load
- ✓ Hier-EI adjusts frame rate/buffer size as **object motion slows** to accelerating processing



Resource Context Adaption

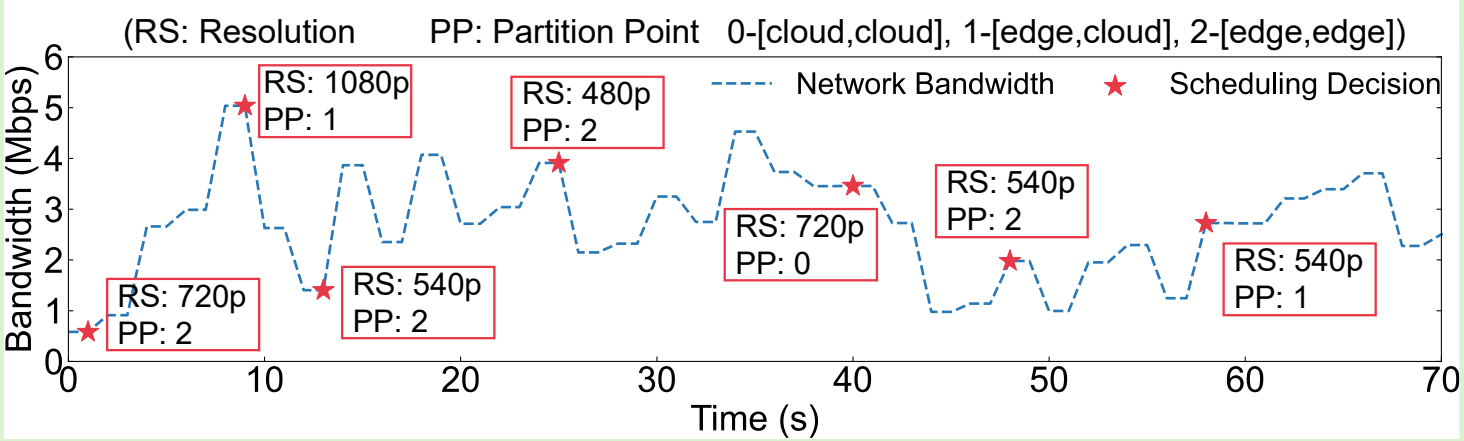
- ✓ Hier-EI also adjusts knobs to adapt to resource variation like **network bandwidth**
- ✓ Hier-EI exhibits **predictive behaviors** to make conservative decisions despite high bandwidth

➢ Case Study for Adaptive Scheduling



Task Context Adaption

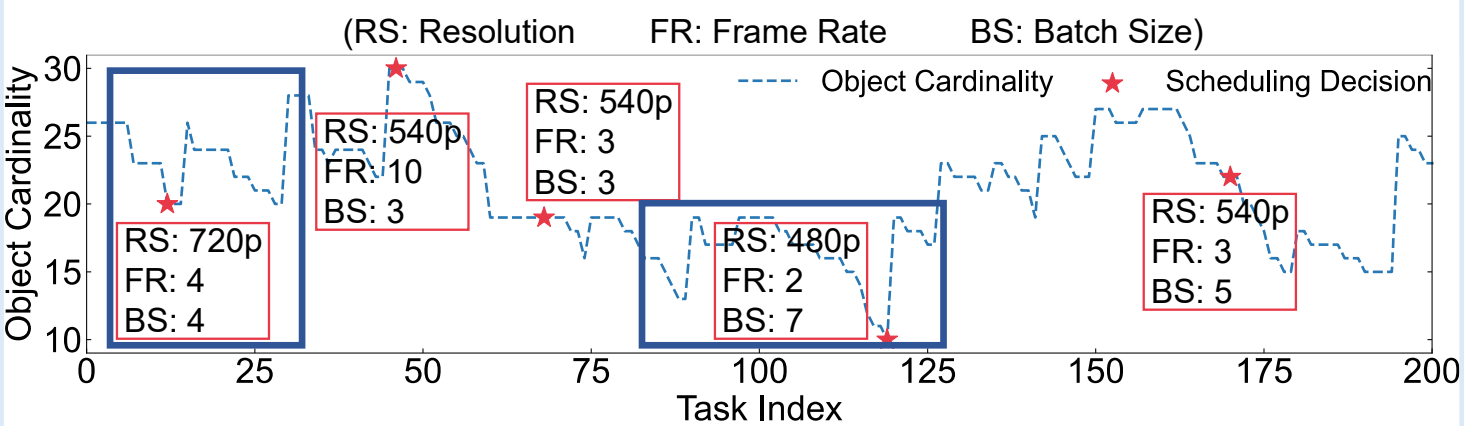
- ✓ Hier-EI reduces frame resolution during **high object density** to lower computational load
- ✓ Hier-EI adjusts frame rate/buffer size as **object motion slows** to accelerating processing



Resource Context Adaption

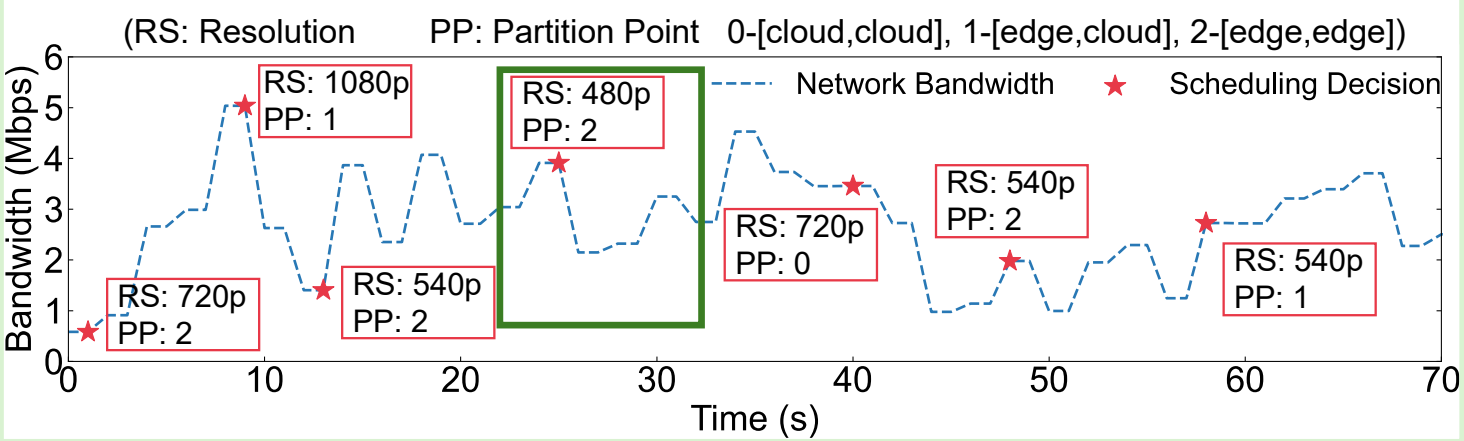
- ✓ Hier-EI also adjusts knobs to adapt to resource variation like **network bandwidth**
- ✓ Hier-EI exhibits **predictive behaviors** to make conservative decisions despite high bandwidth

➢ Case Study for Adaptive Scheduling



Task Context Adaption

- ✓ Hier-EI reduces frame resolution during **high object density** to lower computational load
- ✓ Hier-EI adjusts frame rate/buffer size as **object motion slows** to accelerating processing



Resource Context Adaption

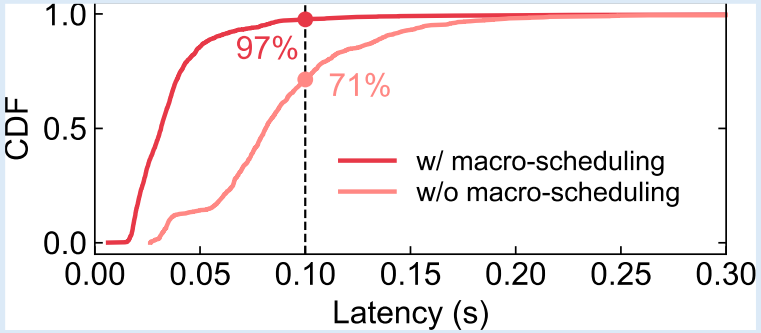
- ✓ Hier-EI also adjusts knobs to adapt to resource variation like **network bandwidth**
- ✓ Hier-EI exhibits **predictive behaviors** to make conservative decisions despite high bandwidth

Evaluation

➢ Ablation Study

No Macro

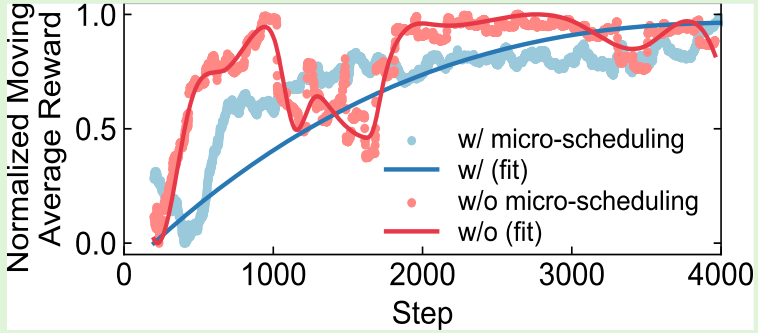
Hier-EI *v.s.* n independent NFs



- ✓ w/o macro-scheduling has **a 26% latency drop** at 0.1-second threshold
- ✓ Macro-scheduling **give global guidance** to coordination

No Micro

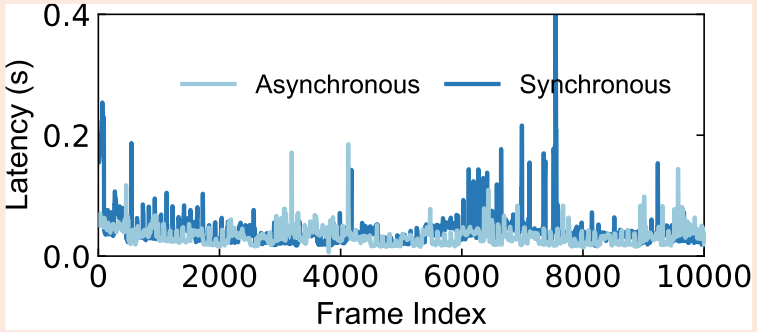
Hier-EI *v.s.* end-to-end EI



- ✓ w/o micro-scheduling **struggles to converge** in an exponentially complex decision space
- ✓ Micro-scheduling **splits high-dimensional complexity**

No Async

Asynchronous *v.s.* Synchronous



- ✓ synchronous collaboration has **poorer adaptability** to sudden changes
- ✓ Asynchronous collaboration **reserves respective advantages**

Conclusion

Tackling the Imbalance in Video Analytics Pipelines with Hierarchical Embodied Intelligence

- ✓ A **two-phase hierarchical** scheduling framework with **embodied intelligence** to eliminate the imbalance in real-time video analytics pipelines.
- ✓ A **hierarchical asynchronous collaboration mechanism** to reduce complexity and an **embodied closed-loop feedback mechanism** to adapt to dynamics.
- ✓ Implement **a prototype system based on KubeEdge** to validate the performance of our Hier-EI in real-world environment.
- ✓ Improve latency compliance ratio by 3.6× and reduces P95 latency by 67.4% compared to state-of-the-art scheduling methods.

Opensource Code: https://github.com/dayu-autostreamer/dayu

Hier-EI Homepage

Dayu Homepage

Thanks for listening! Q & A

Wenhui Zhou, Lei Xie*, Jingyi Ning, Shuyu Cao, Hao Wu, Qinghua Peng, Long Fan

State Key Laboratory for Novel Software Technology, Nanjing University

For detailed information and opensource code:

Hier-EI Homepage

Dayu Homepage

Dayu Repository

For contact:

Dislab @ Nanjing University

Lei Xie

lxie@nju.edu.cn

Wenhui Zhou

whzhou@smail.nju.edu.cn