SHEEO: Continuous Energy Efficiency Optimization in Autonomous Embedded Systems

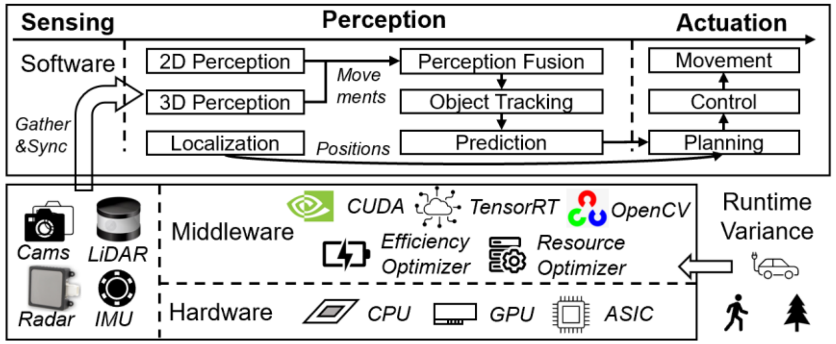
Presenter: Xinkai Wang (Shanghai Jiao Tong University)

Advisor: Chao Li, Minyi Guo

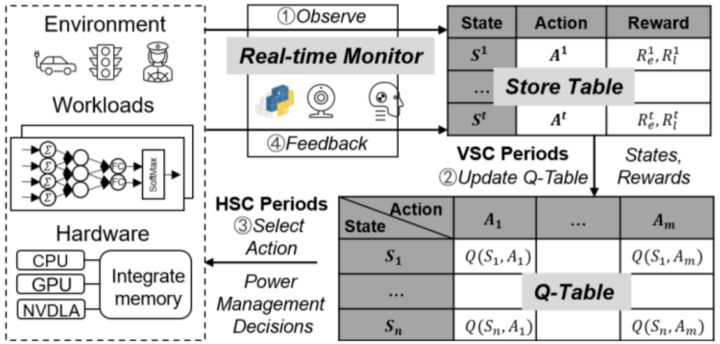
ICCD 2024, Milan, Italy

Content

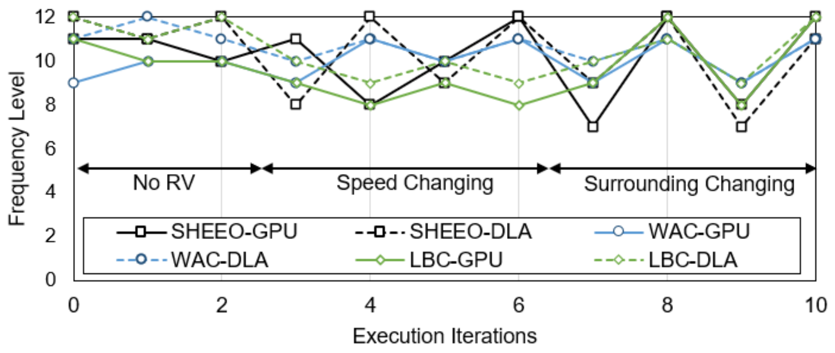
1. Background and Motivation



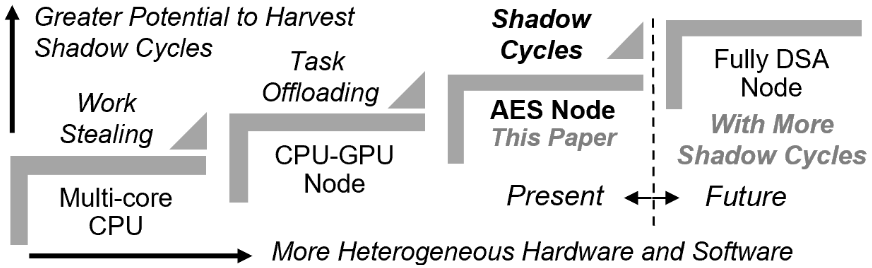
2. Design of SHEEO



3. Evaluation Results

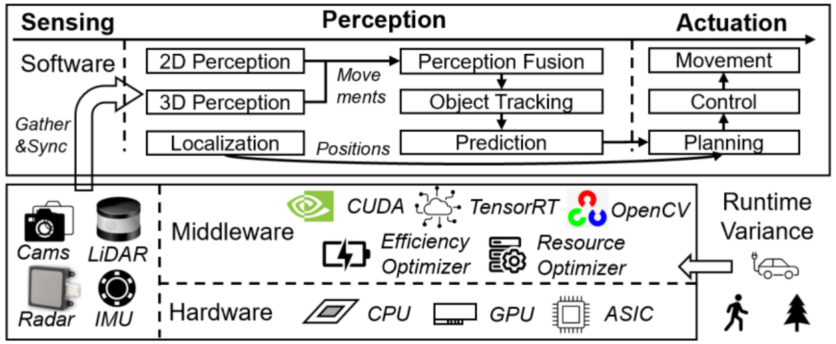


4. Conclusion and Vision

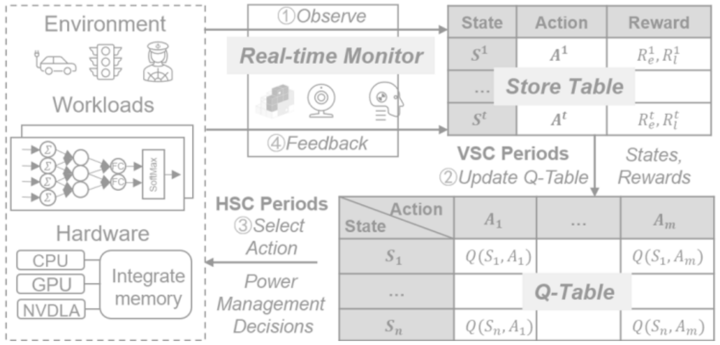


Content

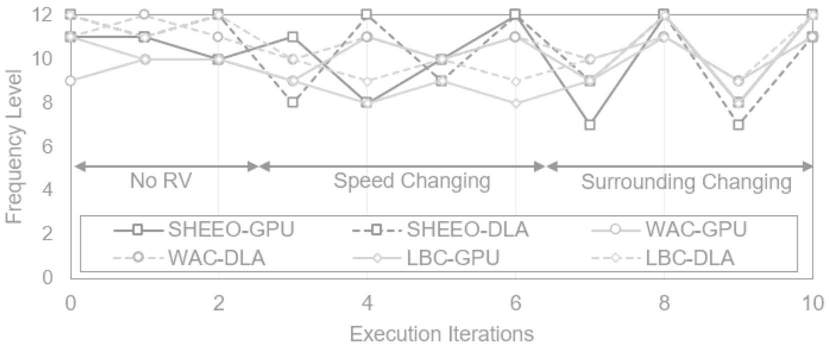
1. Background and Motivation



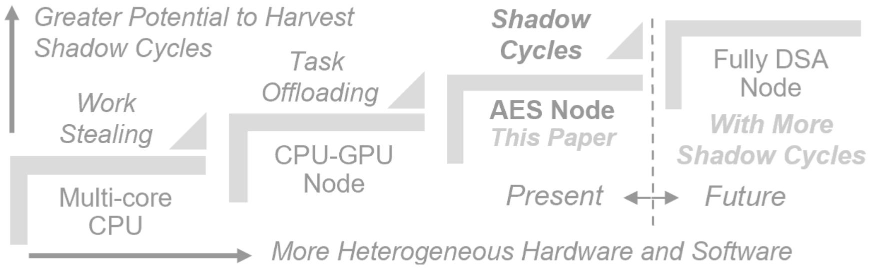
2. Design of SHEEO



3. Evaluation Results



4. Conclusion and Vision



Autonomous Embedded System (AES)

AES is promising to minimize human intervention in various tasks.

Autonomous Driving

Intelligent Manufacturing

Unmanned Drone

Space Exploration

Support

Changing Environments → *Sensing* → **AES** [**Perception** / **Energy**] → *Actuation* → *Adaptive Actions*

Autonomous Embedded System (AES)

AES is promising to minimize human intervention in various tasks.

Autonomous Driving

Intelligent Manufacturing

Unmanned Drone

Space Exploration

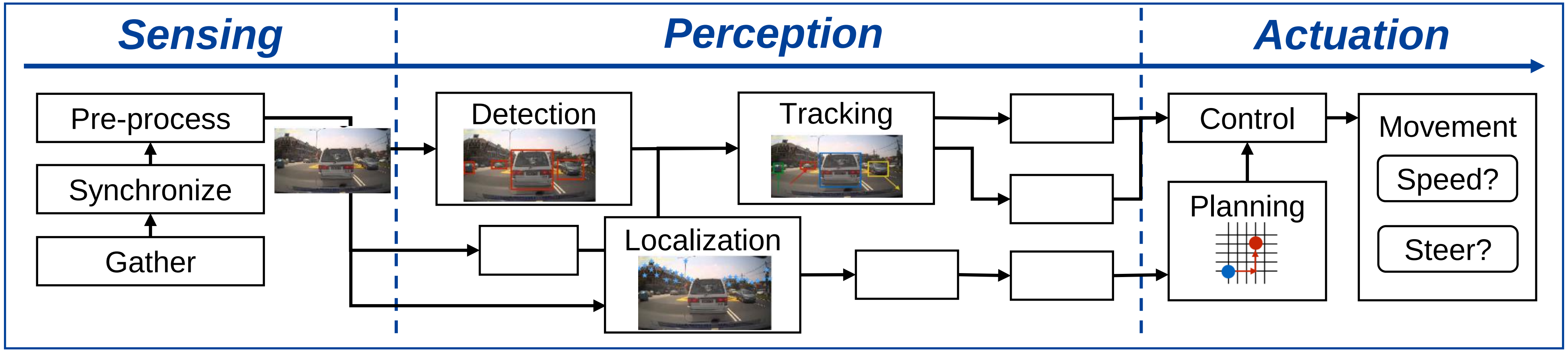
Support

Changing Environments → *Sensing* → **AES** [**Perception** / **Energy**] → *Actuation* → *Adaptive Actions*

Background: Present AES Pipeline

AES has three-stage pipeline on limited heterogeneous hardware.

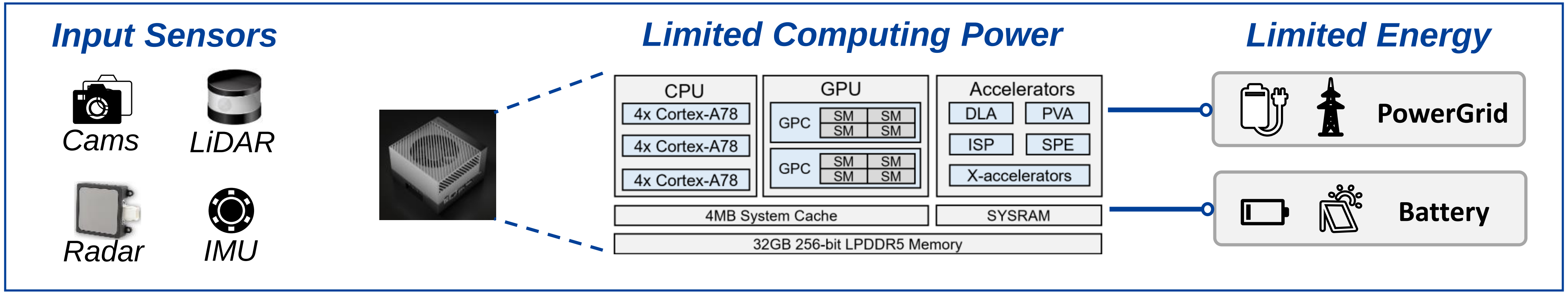
Three-Stage **Software** Pipeline



Middleware Facilities

CUDA *TensorRT* *OpenCV* *Efficiency Optimizer* *Resource Optimizer*

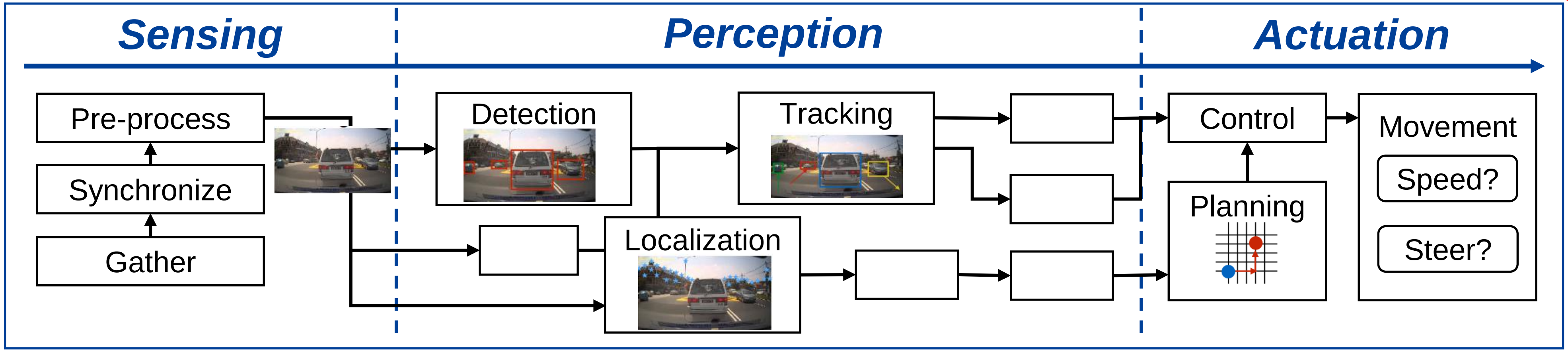
Limited **Hardware** Resources



Background: Present AES Pipeline

AES has three-stage pipeline on limited heterogeneous hardware.

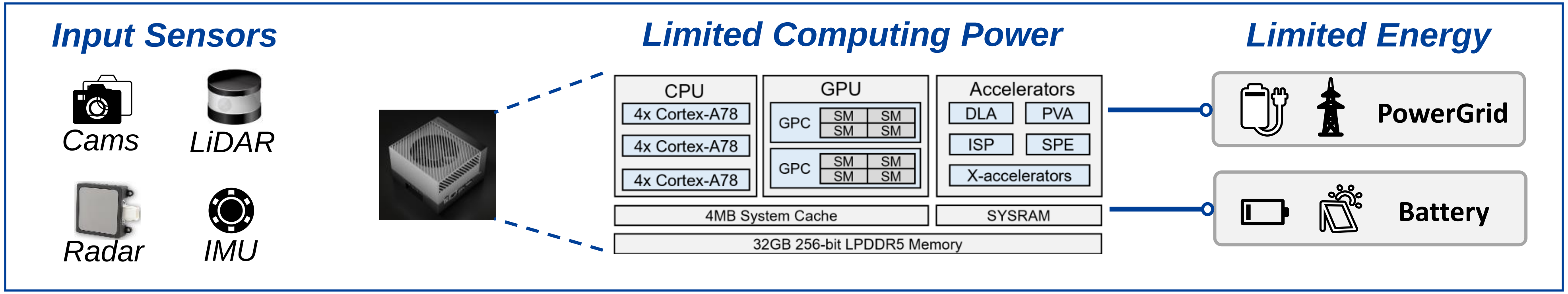
Three-Stage **Software** Pipeline



Middleware Facilities

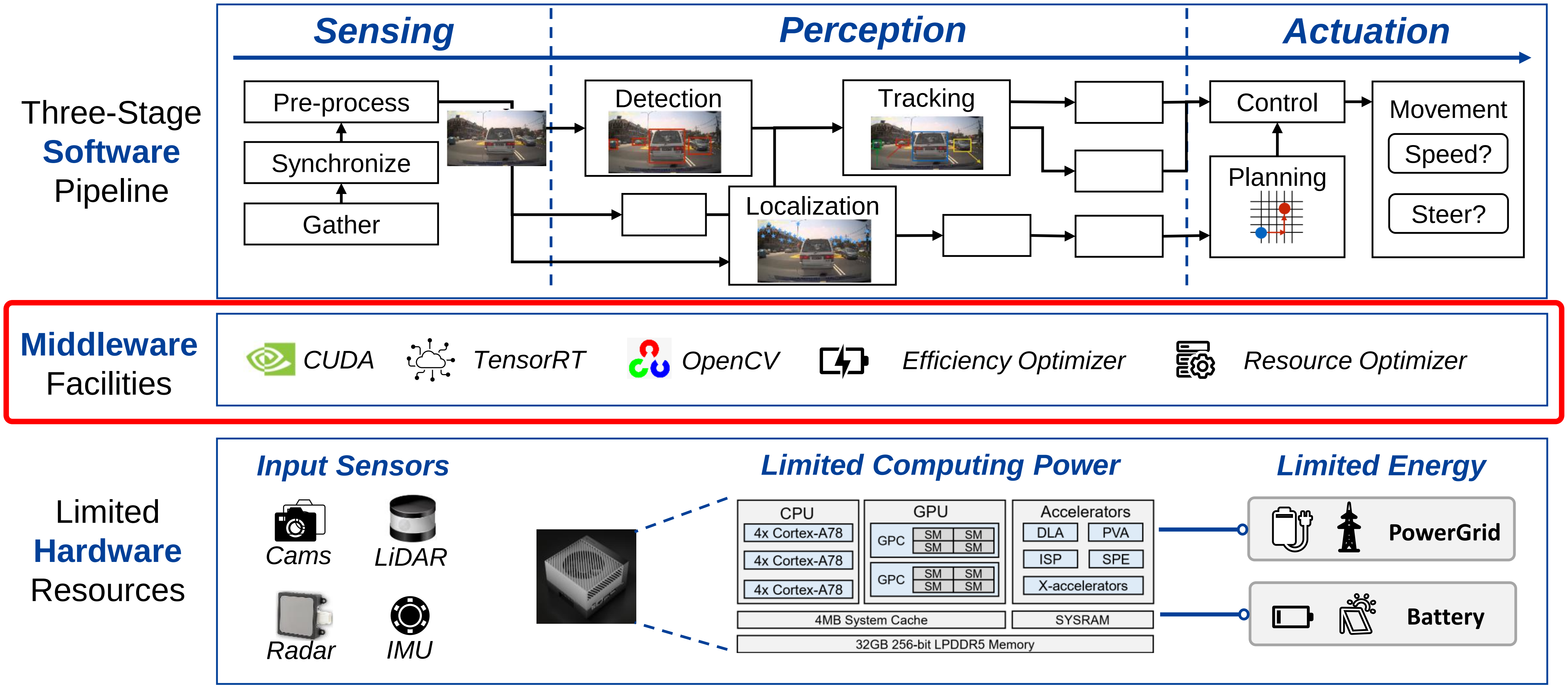
CUDA *TensorRT* *OpenCV* *Efficiency Optimizer* *Resource Optimizer*

Limited **Hardware** Resources



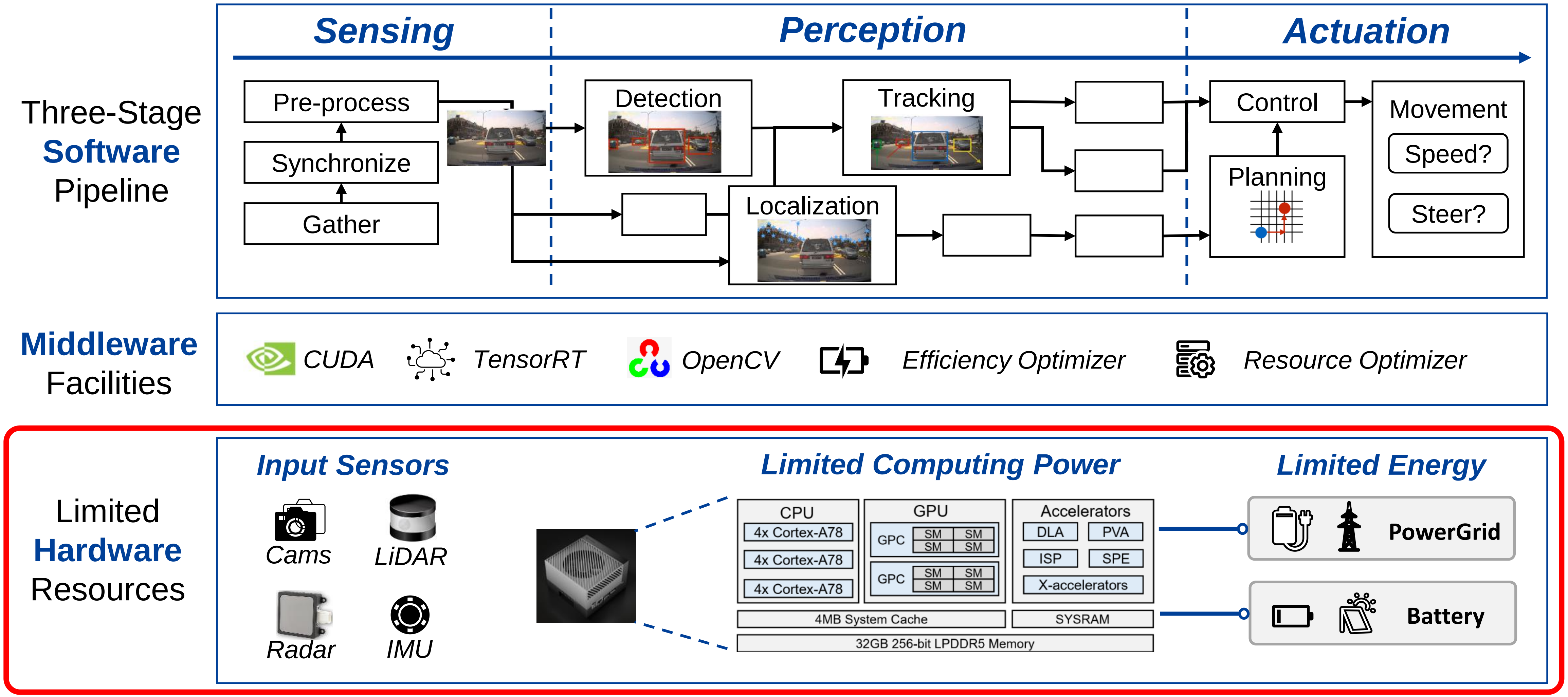
Background: Present AES Pipeline

AES has three-stage pipeline on limited heterogeneous hardware.



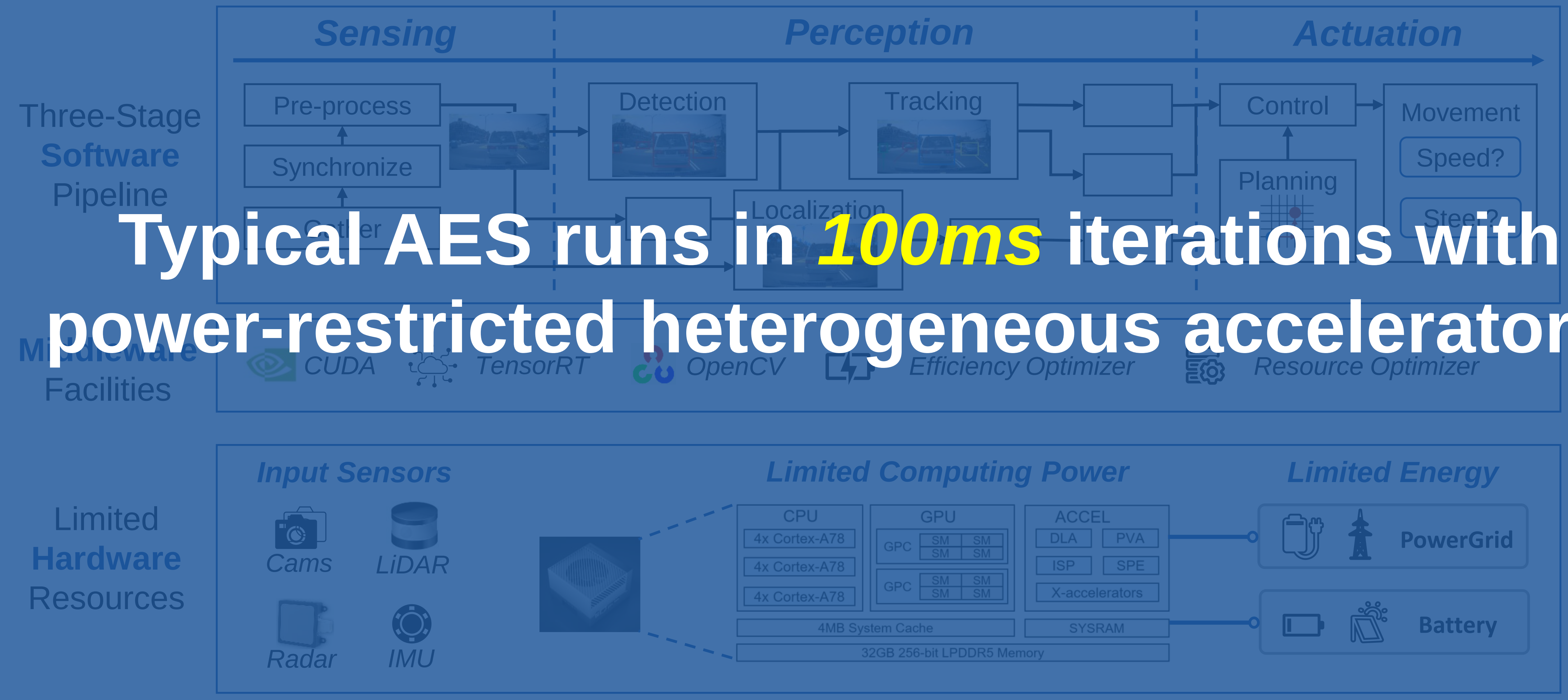
Background: Present AES Pipeline

AES has three-stage pipeline on limited heterogeneous hardware.



Background: Present AES Pipeline

AES has three-stage pipeline on limited heterogeneous hardware.



Typical AES runs in *100ms* iterations with power-restricted heterogeneous accelerators.

Background: AES Efficiency Optimization (EO)

AES faces varied environments and complex hardware to manage.

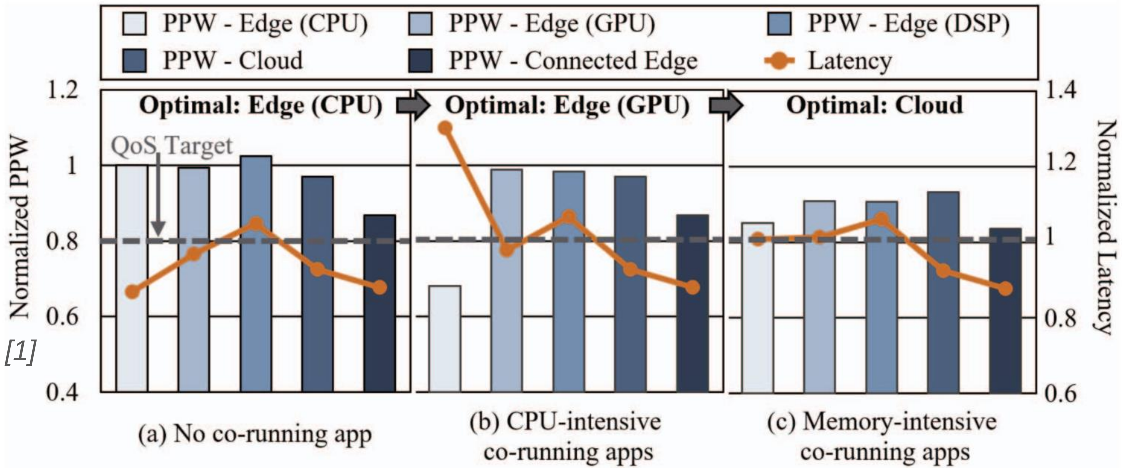
Temporal Variance: External Dynamics.

AES	Urban Road	Traffic Jam	Overtake	Highway
Dynamics	**Case 1**	**Case 2**	**Case 3**	**Case 4**
Vehicle Speed	*Low*	*Low*	*Fluctuated*	*High*
Surrounding	*Low*	*Fluctuated*	*Low*	*Low*
Task Complexity	*Low*	*High*	*High*	*Low*

Current Method	V/F ↓		V/F ↑	
Expected Method	V/F ↓	Adapting V/F with time		V/F ↓

Require ***Continuous*** *EO*

Spatial Variance: Accelerator Behavior.



[1]

Require ***Intelligent*** *EO*

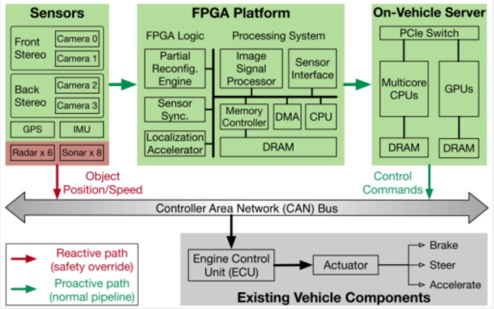
AES expects EO in *100ms* granularity within a large optimization space.

[1] Kim, Young Geun, and Carole-Jean Wu. "Autoscale: Energy efficiency optimization for stochastic edge inference using reinforcement learning." *MICRO 2020*

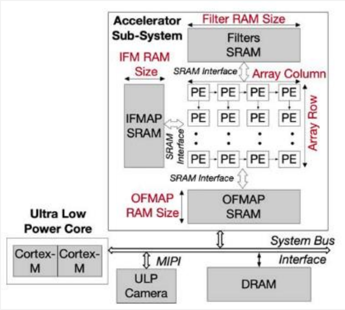
Background: Review of AES Optimization

Researchers explore resource and energy optimization separately.

Increasing Resource Heterogeneity



SOV, MICRO'20 [1]

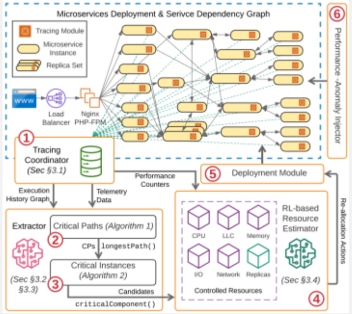


AutoPilot, MICRO'22 [2]

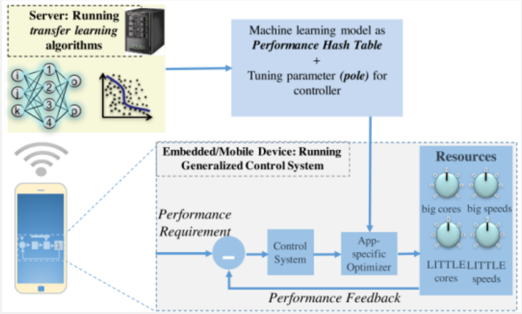
Accelerators offer unbalanced computing power

?

Reducing Energy Consumption



FIRM, OSDI'20 [3]



CALOREE, ASPLOS'18 [4]

On-device ML-based EO is costly for AES.

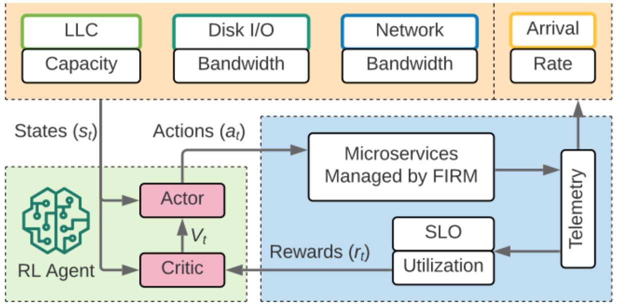
Lack of utilizing ignored hetero-computing resource for costly EO facility .

[1] Yu, Bo, et al. "Building the computing system for autonomous micromobility vehicles: Design constraints and architectural optimizations." *MICRO 2020*
[2] Krishnan, Srivatsan, et al. "Automatic domain-specific soc design for autonomous unmanned aerial vehicles." *MICRO 2022*
[3] Qiu, Haoran, et al. "FIRM: An intelligent fine-grained resource management framework for SLO-Oriented microservices." *OSDI 2020*
[4] Mishra, Nikita, et al. "Caloree: Learning control for predictable latency and low energy." *ASPLOS 2018*

Motivation: EO is Costly for AES

Desirable efficiency optimization interfere with AES pipeline.

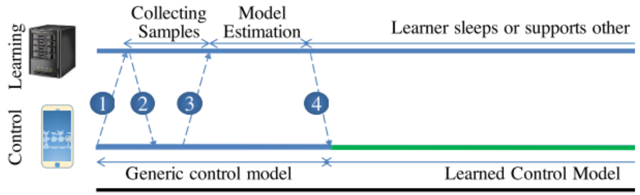
Complex ML models introduces great **training / inference overheads**.



FIRM, OSDI'20

Reinforcement Learning

Overhead=**420ms / 40ms**

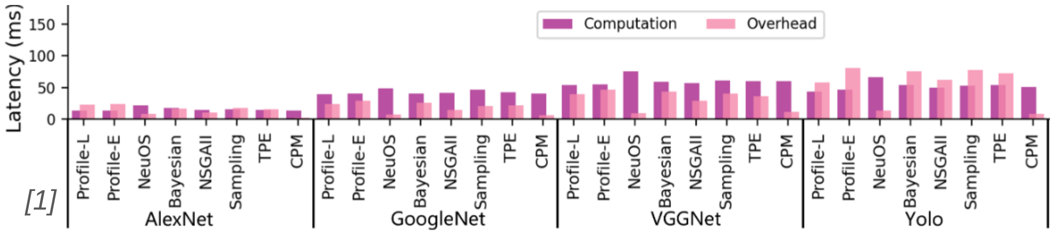


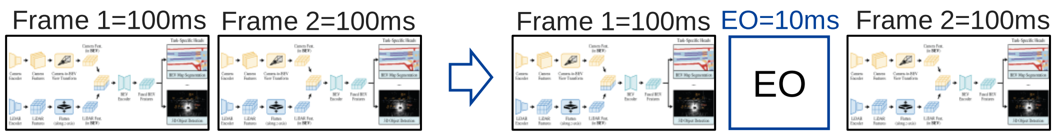
CALOREE, ASPLOS'18

Transfer Learning

Overhead=**500ms / 2ms**

Continuous EO **disturbs execution** on resource-constrained AES.





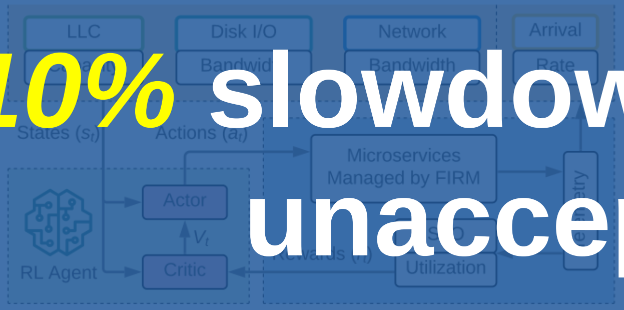
Typical EO using normal computing power causes **~10%** slowdown of each iteration.

[1] Hou, Xiaofeng, et al. "CPM: A Cross-layer Power Management Facility to Enable QoS-Aware AIoT Systems." *IWQoS 2024*

Motivation: EO is Costly for AES

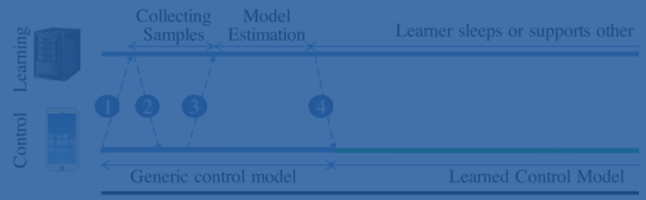
Desirable efficiency optimization interfere with AES pipeline.

Complex ML models introduces great **training / inference overheads**.



FIRM, OSDI'20

Reinforcement Learning

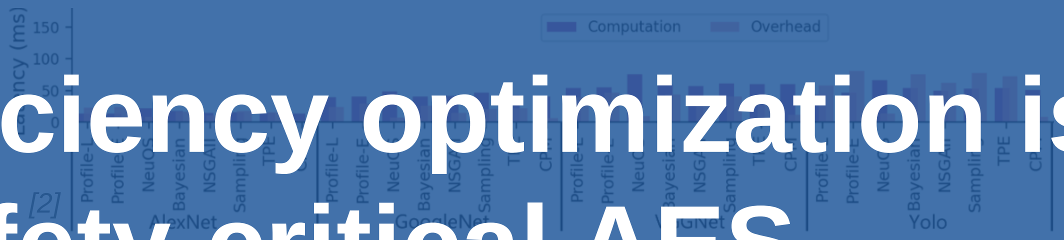


CALOREE, ASPLOS'18

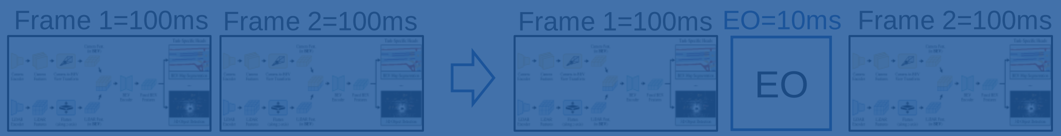
Transfer Learning

Overhead=**500ms / 2ms**

Continuous EO **disturbs execution** on resource-constrained AES.



[2]



Typical EO using normal computing power causes **~10%** slowdown of each iteration.

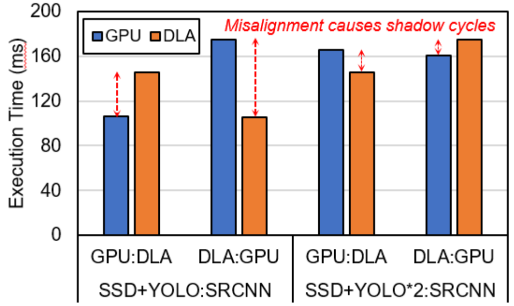
10% slowdown due to efficiency optimization is unacceptable for safety-critical AES.

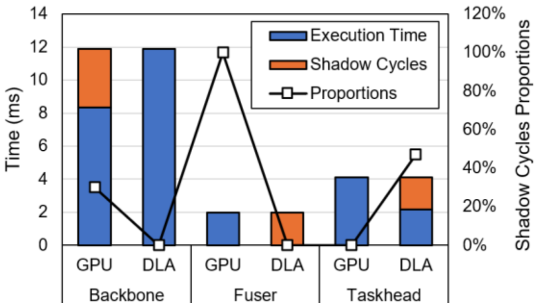
[1] Hou, Xiaofeng, et al. "CPM: A Cross-layer Power Management Facility to Enable QoS-Aware AIoT Systems." *IWQoS 2024*

Motivation: Shadow Cycles for EO

AES presents underutilized shadow cycles for deploying EO.

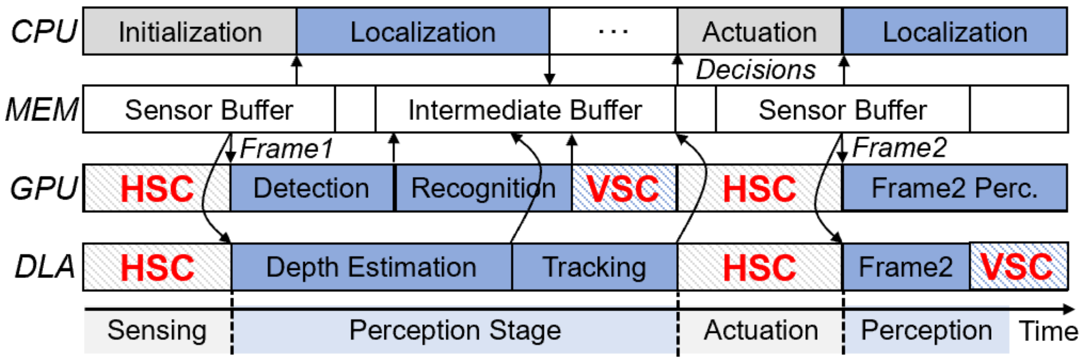
Diversity of inference and accelerator present **misaligned** execution.





*Typical inferences on GPU and DLA produces **~12%** underutilized resources.*

Shadow cycles are **underutilized and ignored** computing resources in AES.

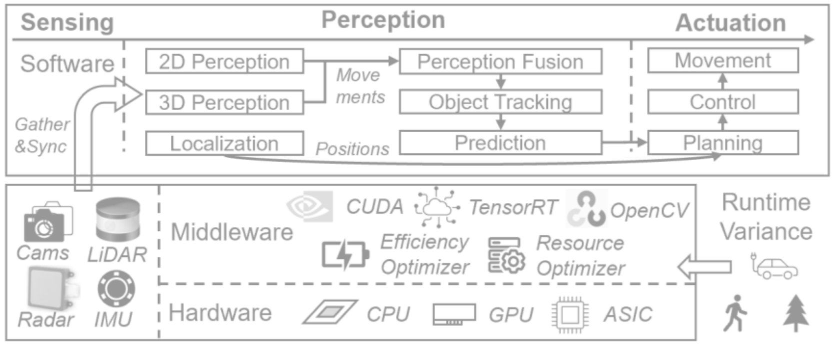


- ***Vertical Shadow Cycles (VSC)***
  Preemptive and Variable-sized
- ***Horizontal Shadow Cycles (HSC)***
  Periodic and Short-lived

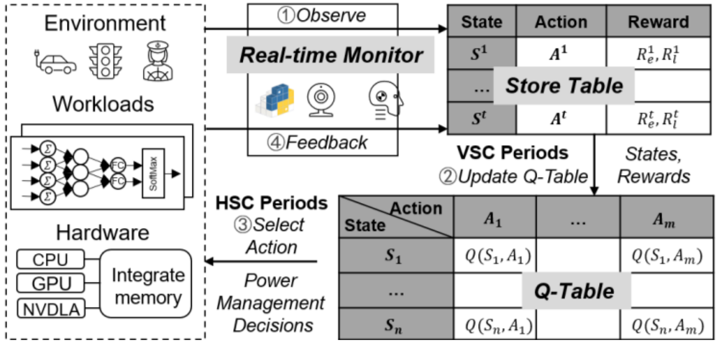
The ignored resources in AES are promising for EO without slowdown.

Content

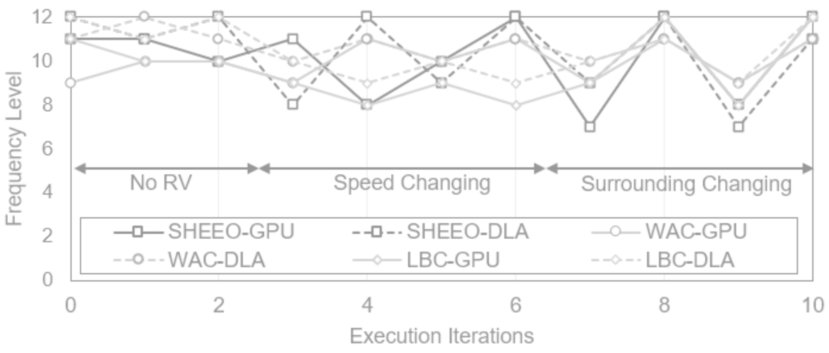
1. Background and Motivation



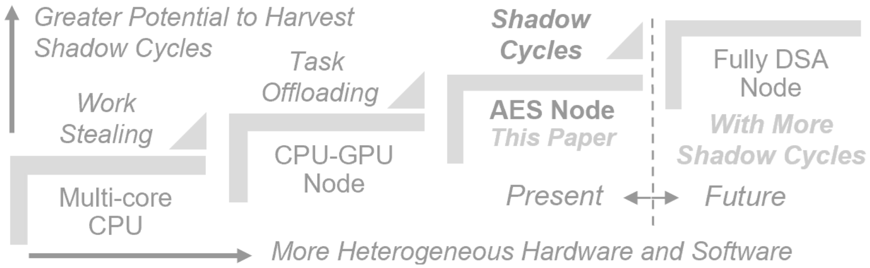
2. Design of SHEEO



3. Evaluation Results



4. Conclusion and Vision



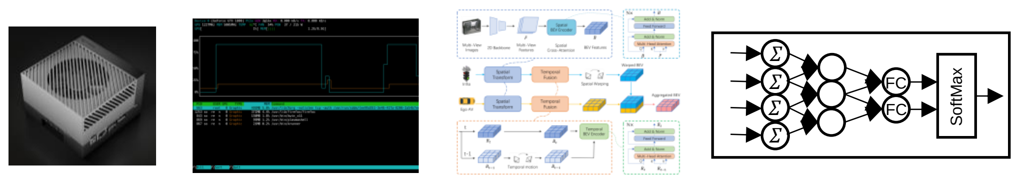
SHEEO Design Consideration

Two key issues exist in SHEEO design…

Using Heterogeneous Resource

Reducing Energy Consumption

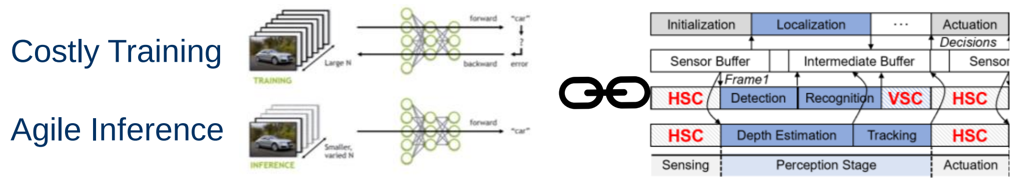
How to **Observe** AES?



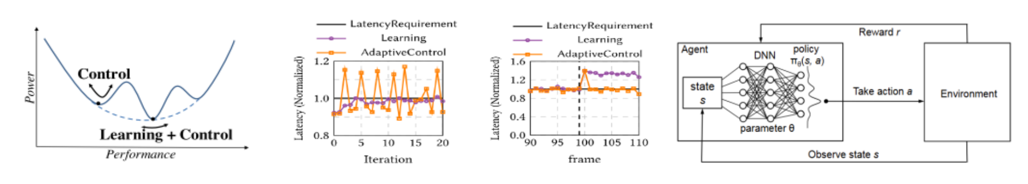
Internal Runtime Status

External Environment

How to **Optimize** AES?



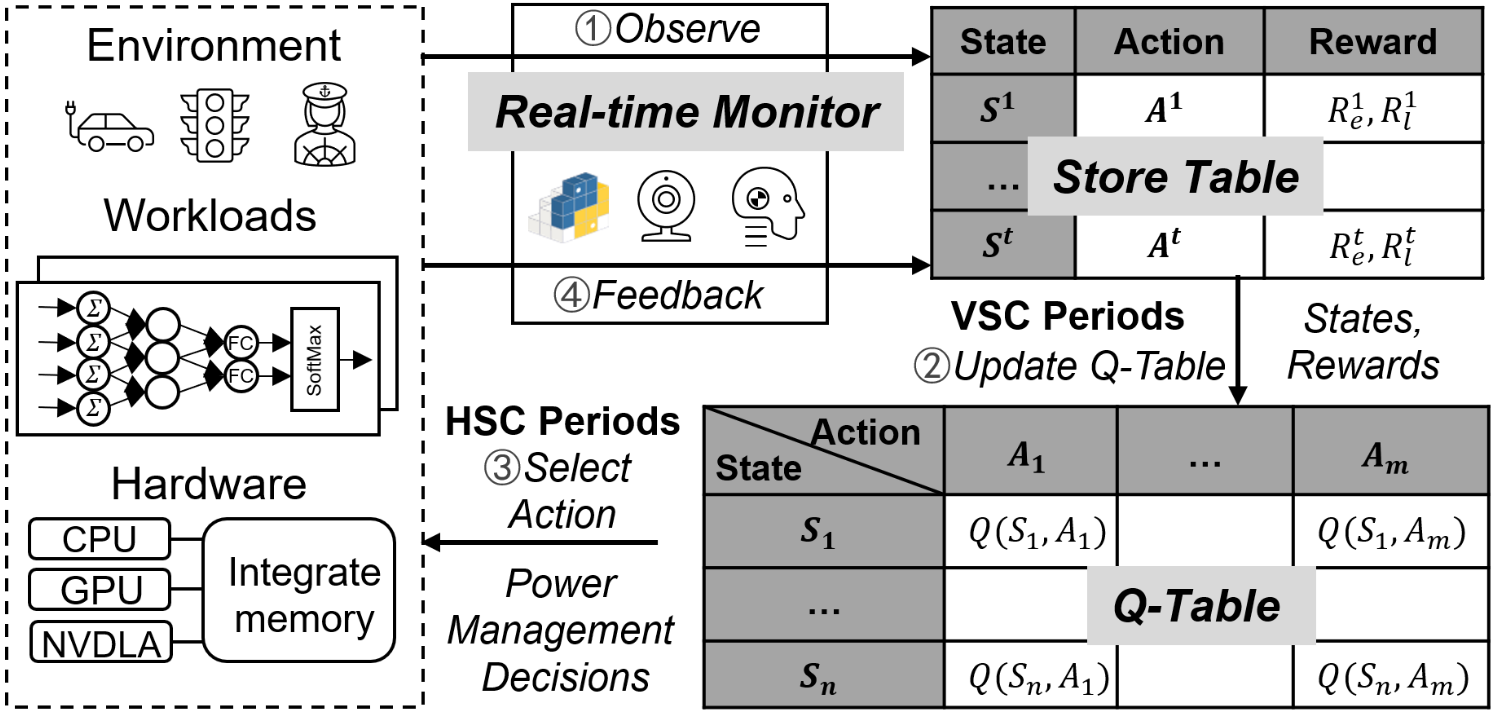
Utilization of Shadow Cycles



Learning and Control

SHEEO Overview

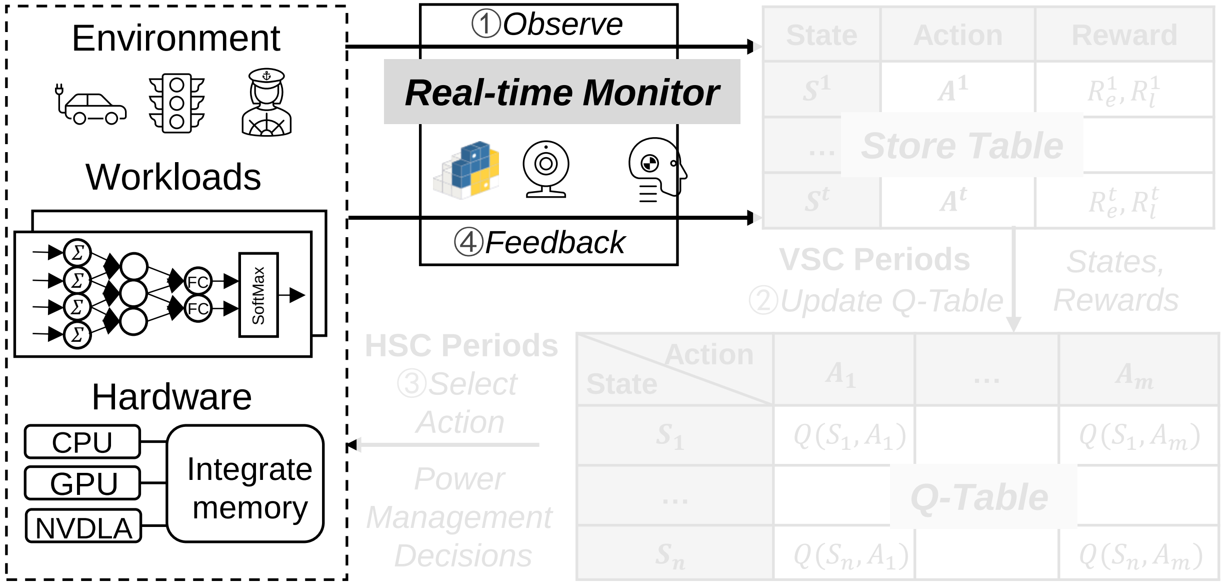
SHEEO exploits shadow cycles for continuous and intelligent EO.



- SHEEO has **two cooperative components**: real-time monitor and continuous learning manager.
- SHEEO works at OS, enhancing **power management facility** without intrusion into AES pipeline.

SHEEO Observation: Environment and Workload

SHEEO observes internal and external status via real-time monitor.



- Runtime Monitor observes the **start and end of AES tasks** for recording shadow cycles.
- It observes the environment changes and system status and records them into the **7-dimensional state table**.

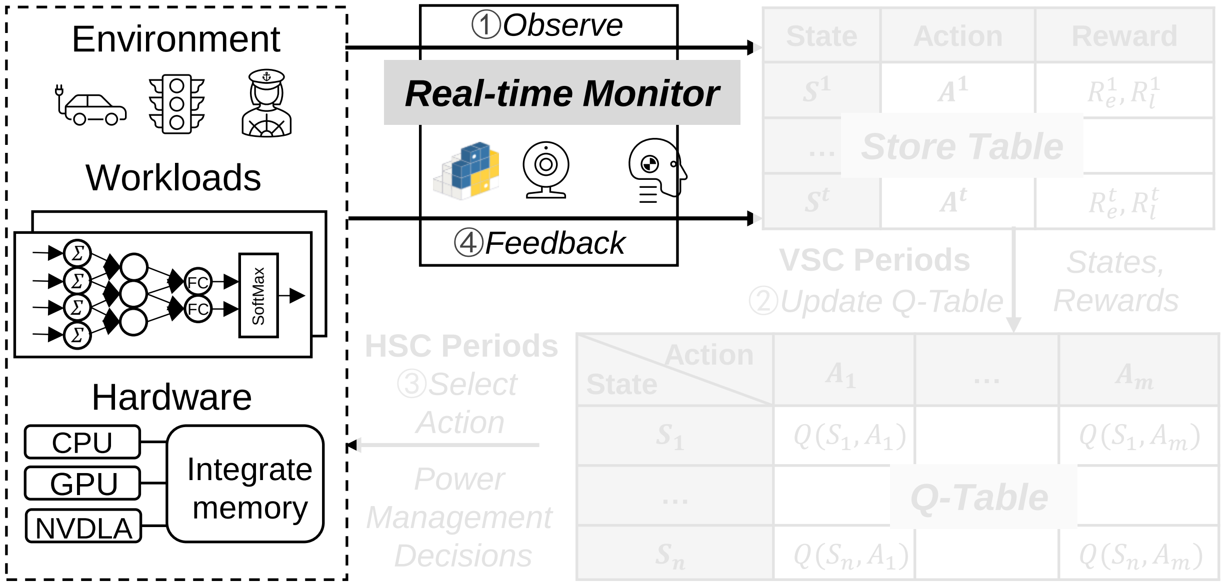
(a) Multi-dimensional state-related features

State		Descriptions	Discrete Values
Workload Features	S_{CONV}	Number of CONV layers	L (<30), M (<50), H (≥50)
	S_{FC}	Number of FC layers	L (<10), H (≥10)
	S_{RC}	Number of RC layers	L (<10), H (≥10)
Runtime Dynamics	S_{Com}	Computing units' utilization	L (<25%), M (<75%), H (≥75%)
	S_{Mem}	Memory utilization	L (<25%), M (<75%), H (≥75%)
	S_{Speed}	Current vehicle speed	L (<10mph), M (<40mph), H (≥40mph)
	S_{Var}	Variation from last execution	L (<10%), M (<30%), H (≥30%)

- **State Table for Internal Runtime Status**
  - Model software and hardware features.
  - Adjustable for actual AES workloads.
- **State Table for External Environment**
  - Model the external stochastic variance.
  - Use discrete ranges to reduce complexity.

SHEEO Observation: Environment and Workload

SHEEO observes internal and external status via real-time monitor.



(a) Multi-dimensional state-related features

State		Descriptions	Discrete Values
Workload Features	S_{CONV}	Number of CONV layers	L (<30), M (<50), H (≥50)
	S_{FC}	Number of FC layers	L (<10), H (≥10)
	S_{RC}	Number of RC layers	L (<10), H (≥10)
Runtime Dynamics	S_{Com}	Computing units' utilization	L (<25%), M (<75%), H (≥75%)
	S_{Mem}	Memory utilization	L (<25%), M (<75%), H (≥75%)
	S_{Speed}	Current vehicle speed	L (<10mph), M (<40mph), H (≥40mph)
	S_{Var}	Variation from last execution	L (<10%), M (<30%), H (≥30%)

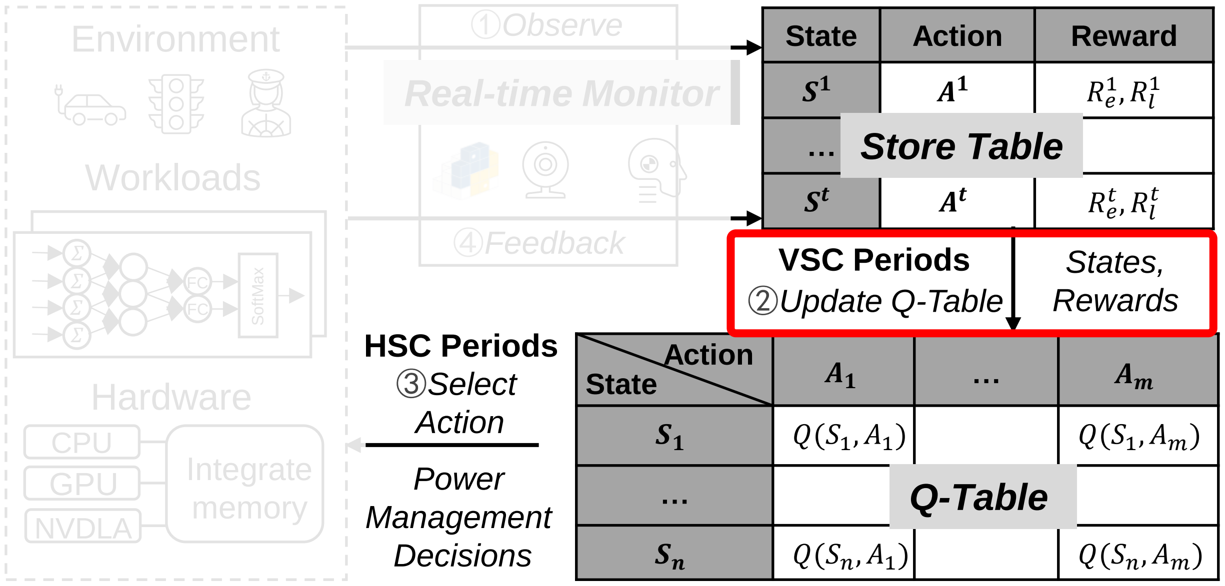
- State Table for Internal Runtime Status
  - Model software and hardware features.
  - Adjustable for actual AES workloads.

- **State Table for External Environment**
  - Model the external stochastic variance.
  - Use discrete ranges to reduce complexity.

- Runtime Monitor observes the **start and end of AES tasks** for recording shadow cycles.
- It observes the environment changes and system status and records them into the **7-dimensional state table**.

SHEEO Optimization: Learning and Control

SHEEO optimizes efficiency via continuous learning manager.



Algorithm 1: Algorithm for continuous learning

Input: Pre-trained Q-table $Q(S, A)$, learning rate α, discount factor γ, exploration probability ϵ

Output: Fine-tuned Q-table and Action A for each iteration

```
while ∃ VSC periods and ∃ Store Table do
    Calculate recorded reward R_energy and R_latency;
    For consecutive S and S' and chosen action A;
    Choose action A' with the largest Q(S', A');
    Q(S, A) ← Q(S, A) + α[R + γQ(S', A') − Q(S, A)];
    S ← S'
end
foreach HSC periods do
    Gather S and locate in Q(S, A);
    if rand() < ε then
        Choose action A randomly;
    else
        Choose action A with the largest Q(S, A);
    end
    Store Table ← (A, S') ;
end
```

Utilizing preemptive and variable-sized VSC.

Two-fold Reward

- Minimize performance slack
- Maximize energy efficiency
- Used for updating Q-table

$$R = aR_{latency} + bR_{energy} \text{ s.t. } R_l \leq \text{Constraint}$$

$$R_{latency} = \max_{t \in \text{Perception}} T_{end}^t - T_{start}$$

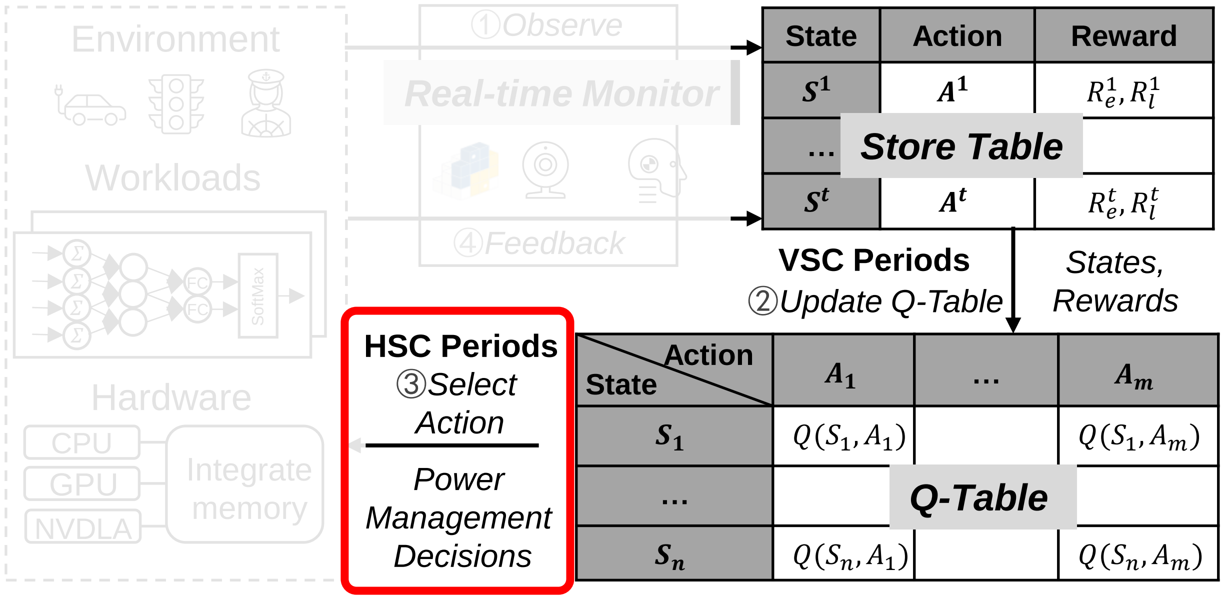
$$R_{energy} = \sum_i E_{Unit}^i = \sum_i \int_{T_b}^{T_e} P_f \delta t + P_{idle} \times t_{idle}$$

- Continuous Learning Manager adopts **reinforcement learning (Q-learning)** to optimize energy efficiency.
- It embeds continuous learning into **VSC** and action control into **HSC** to reduce power management facility costs.

> - **SHEEO learns optimal action of given states.**
>   - Consider status of consecutive iterations.
>   - Adjustable for Q-learning, DDPG, DRL, etc.

SHEEO Optimization: Learning and Control

SHEEO optimizes efficiency via continuous learning manager.



Algorithm 1: Algorithm for continuous learning

Input: Pre-trained Q-table $Q(S, A)$, learning rate α, discount factor γ, exploration probability ϵ
Output: Fine-tuned Q-table and Action A for each iteration

```
while ∃ VSC periods and ∃ Store Table do
    Calculate recorded reward R_energy and R_latency;
    For consecutive S and S' and chosen action A;
    Choose action A' with the largest Q(S', A');
    Q(S, A) ← Q(S, A)+α[R+γQ(S', A')−Q(S, A)];
    S ← S'
end
foreach HSC periods do
    Gather S and locate in Q(S, A);
    if rand() < ϵ then
        Choose action A randomly;
    else
        Choose action A with the largest Q(S, A);
    end
    Store Table ← (A, S') ;
end
```

Utilizing short-lived and periodic HSC.

7-dimensional Action

- Divided into discrete levels.
- Consider heterogeneous HW.

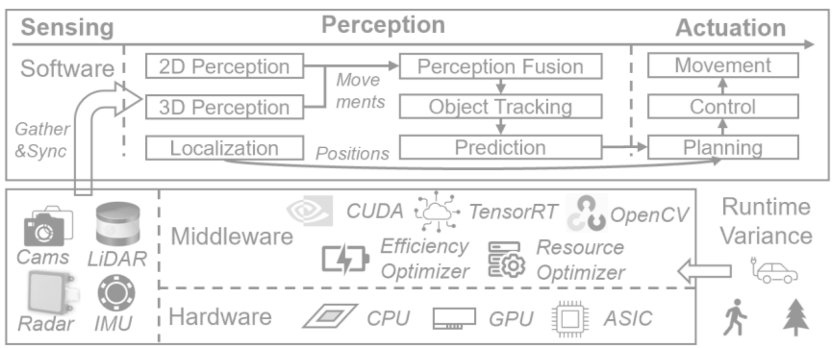
Actions	Descriptions
S_X	Turn On/Off Component X
N_{CPU}	Number of Active Core
F_{CPU}	V/F Level of CPU Core
F_{GPU}	V/F Level of GPU
F_{DLA1}	V/F Level of DLA1
F_{DLA2}	V/F Level of DLA2
F_{MEM}	V/F Level of Memory

- Continuous Learning Manager adopts **reinforcement learning (Q-learning)** to optimize energy efficiency.
- It embeds continuous learning into **VSC** and action control into **HSC** to reduce power management facility costs.

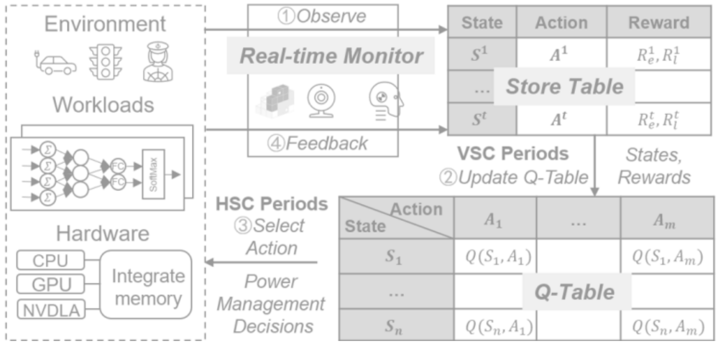
- **SHEEO controls the action in next iteration.**
  - Use ϵ to control exploitation & exploration.
  - Agilely make decisions with largest $Q(S, A)$.

Content

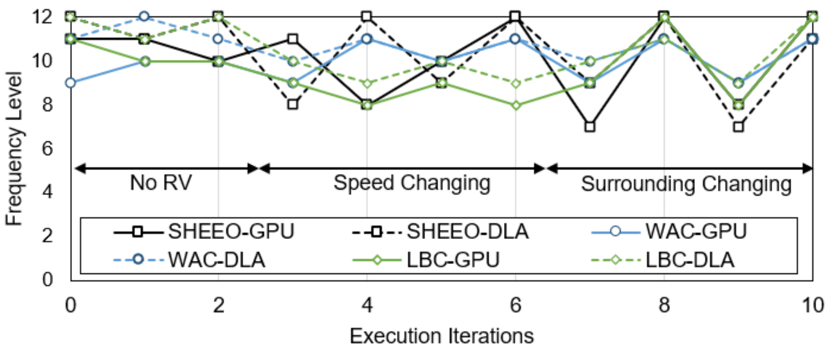
1. Background and Motivation



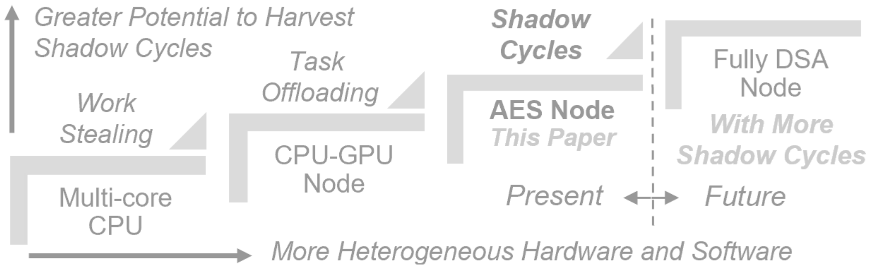
2. Design of SHEEO



3. Evaluation Results



4. Conclusion and Vision



Evaluation Methodology and Settings

We evaluate SHEEO extensively on commercial edge platforms.

Evaluated Platform Specifications	
Device	*Nvidia Jetson AGX Orin Module*
CPU	*8-core ARM Cortex-A78 v8.2*
GPU	*1792-core Ampere GPU with 56 tensor cores*
Memory	*32GB 256-bit LPDDR5*
Accelerator	*2x NVDLA v2, 1x PVA v2*
System	*Linux 5.19.104-tegra with Jetpack 5.1.1*
Software	*CUDA 11.4 and TensorRT 8.5.2*

Evaluated Workloads	
SSD	*ssd-mobilenet-v1 for object detection.*
YOLO	*yolov3-tiny-416 for image detection.*
SRCNN	*super-resolution-bsd500 for image reconstruction.*
Evaluated Baselines	
Workload-Aware Control [1]	*Use static model to profile workloads.*
Learning-Based Control [2]	*Use ML model and historical statistics.*
Oracle	*Offline optimal energy efficiency scheme.*

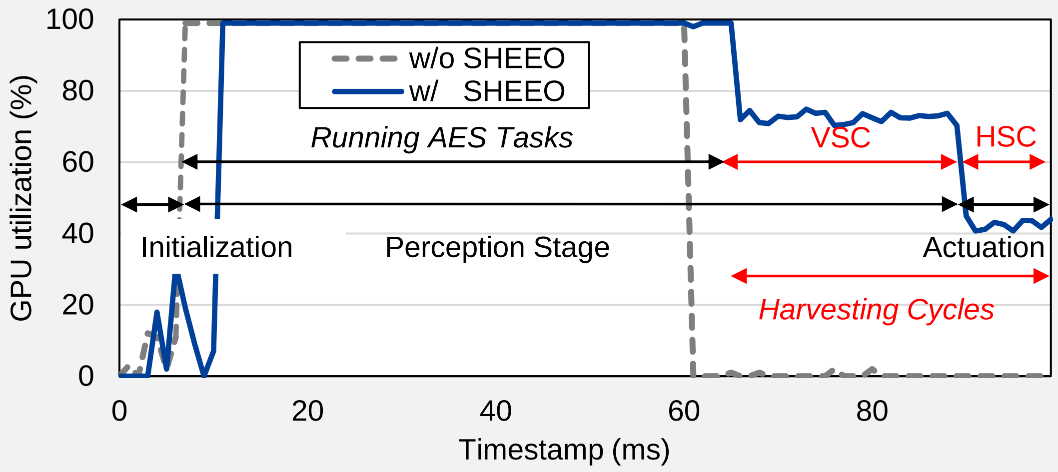
- We evaluate SHEEO on **Nvidia Jetson AGX Orin** and consider shadow cycles on **GPU and NVDLA.**
- We set learning rate as 0.9, discount factor as 0.1, and exploration probability as 0.1.
- We mimic the perception stage of AES pipeline with three networks with different combinations.

[1] Bateni, Soroush, and Cong Liu. "NeuOS: A Latency-Predictable Multi-Dimensional Optimization Framework for DNN-driven Autonomous Systems." *USENIX ATC 20*
[2] Mishra, Nikita, et al. "Caloree: Learning control for predictable latency and low energy." *ASPLOS 2018*

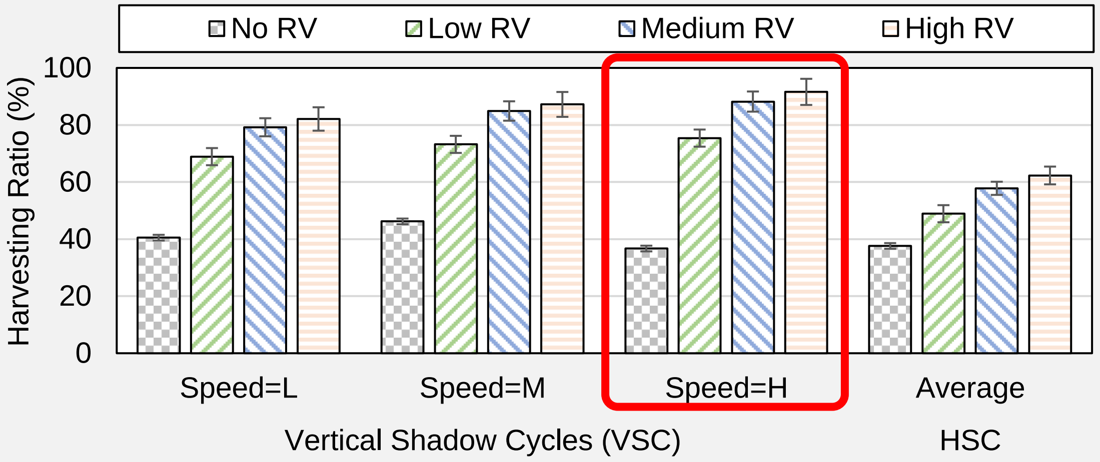
Harvesting Heterogeneous Resources

SHEEO harvests shadow cycles in AES under various scenarios.

An example iteration in AES pipeline with SHEEO harvesting shadow cycles.



SHEEO harvests up to 88% vertical shadow cycles and 51% horizontal shadow cycles.

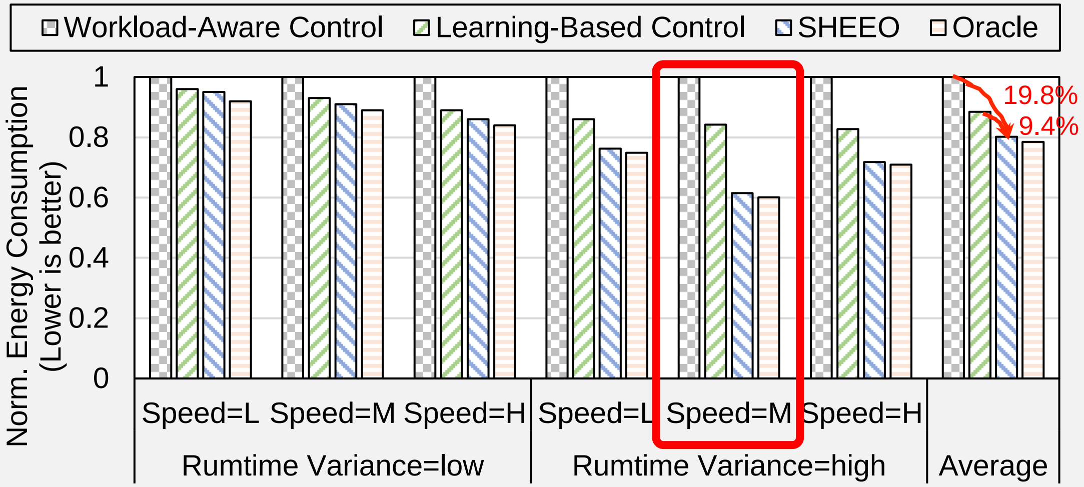


- *Q-learning is too lightweight to occupy all the shadow cycles, but more advanced facilities could utilize better.*
- *The harvesting ratio is higher under higher vehicle speed and higher runtime variance scenarios.*

Reducing Energy Consumptions

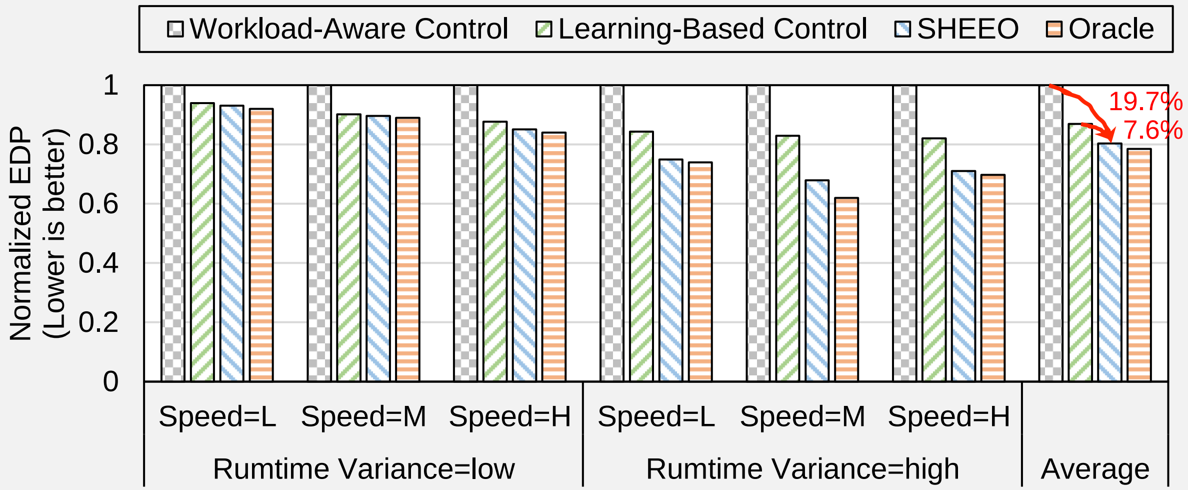
SHEEO improves energy efficiency of AES under various scenarios.

SHEEO provides a more agile and accurate power management solution.



- *SHEEO achieves better energy efficiency under higher runtime variance and medium vehicle speed.*

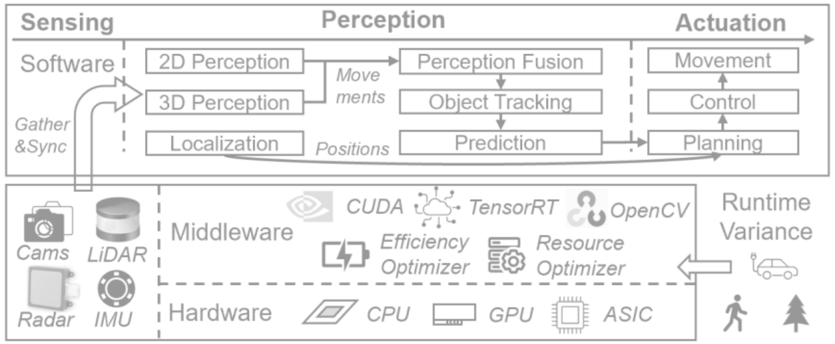
SHEEO achieves better energy-delay product (EDP) compared with baselines.



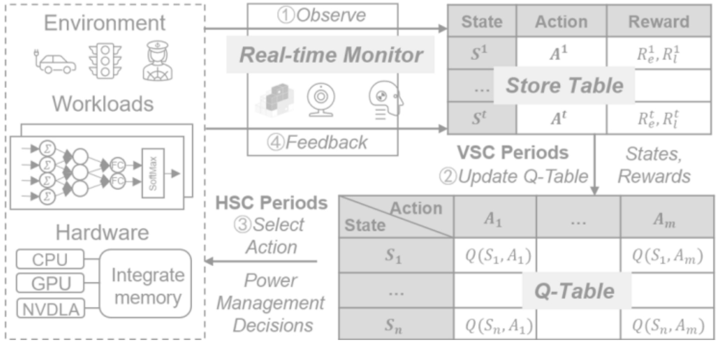
- *SHEEO improves energy efficiency with a little sacrifice of execution performance.*

Content

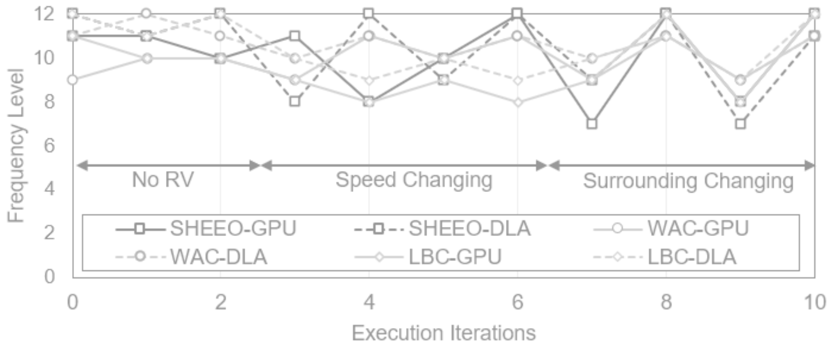
1. Background and Motivation



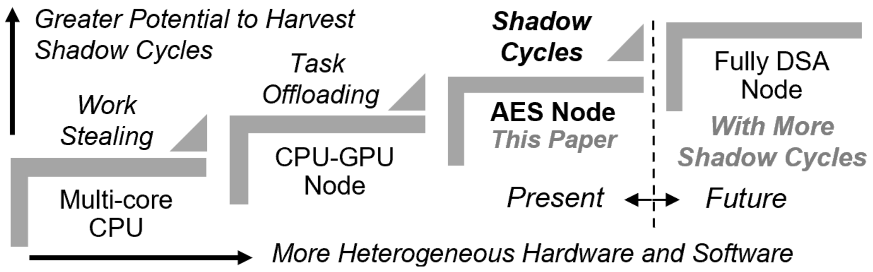
2. Design of SHEEO



3. Evaluation Results



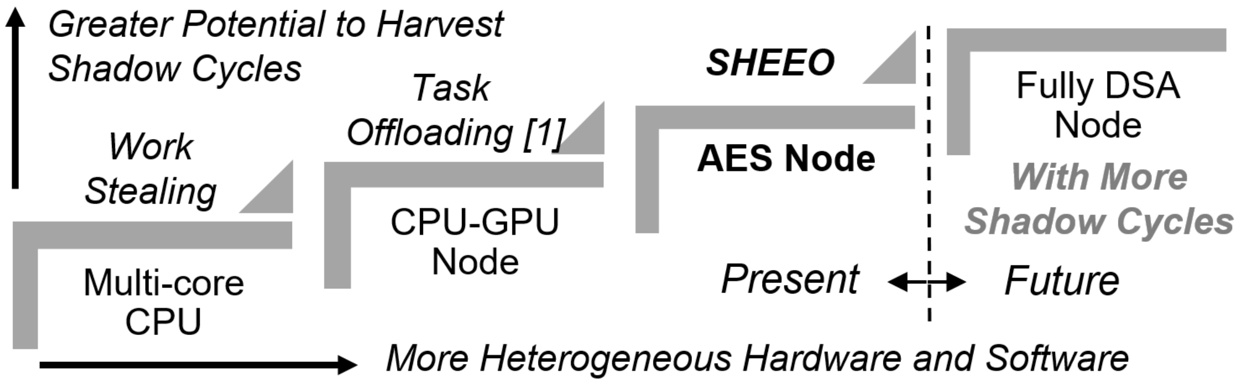
4. Conclusion and Vision



Vision: Future of Autonomous Systems

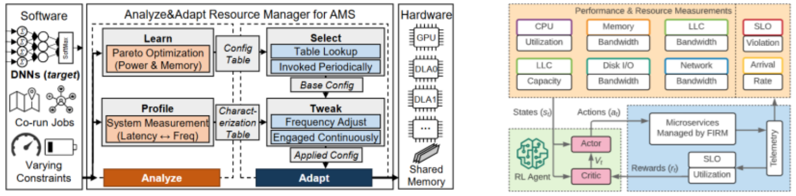
Towards better accelerator usage for efficient system operations.

Future **accelerator-rich architecture** produce **more shadow cycles to exploit!**

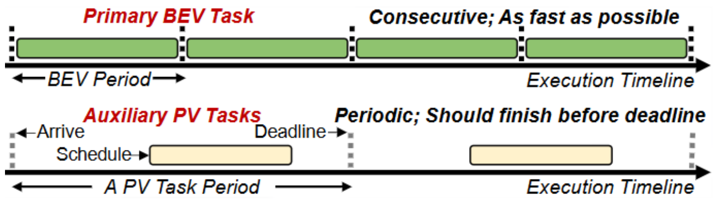


Future **artificial brains** require **more smart facilities and workloads to deploy!**

ML-based Resource Mgmt. [2]



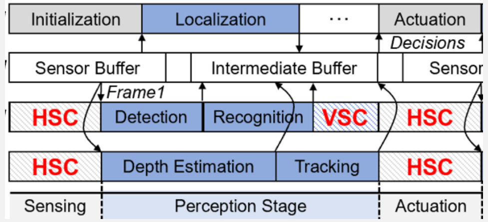
Perspective View Taks in BEV AES [3]



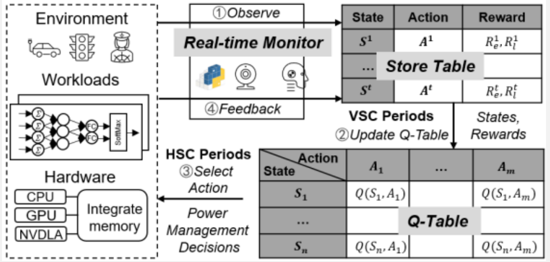
Moving Towards Efficient Autonomous Systems!

[1] Xinkai, Wang, et al. "Not All Resources are Visible: Exploiting Fragmented Shadow Resources in Shared-State Scheduler Architecture." *SoCC 2023*
[2] Lingyu, Sun, et al. "A2: Towards Accelerator Level Parallelism for Autonomous Micromobility Systems." *TACO 2024*
[3] Lingyu, Sun, et al. "Jigsaw: Taming BEV-centric Perception on Dual-SoC for Autonomous Driving." *RTSS 2024*

Conclusion



- We analyze **shadow cycles** in **autonomous embedded systems** and the need to **continuously optimizing** energy efficiency.



- SHEEO is an intelligent continuous energy efficiency optimizer that first **embeds the costly facility into shadow cycles**.
- SHEEO **harvests** the underutilized hetero-computing resources and **improves** AES energy efficiency under various scenarios.




- By exploiting **“free”** shadow cycles within more heterogeneous accelerators, AES can enable a wide range of intelligent **management facilities and autonomous workloads**.

Thank You! Q & A

SHEEO: Continuous Energy Efficiency Optimization in Autonomous Embedded Systems

Discussion: Xinkai Wang (Presenter), unbreakablewxk@sjtu.edu.cn