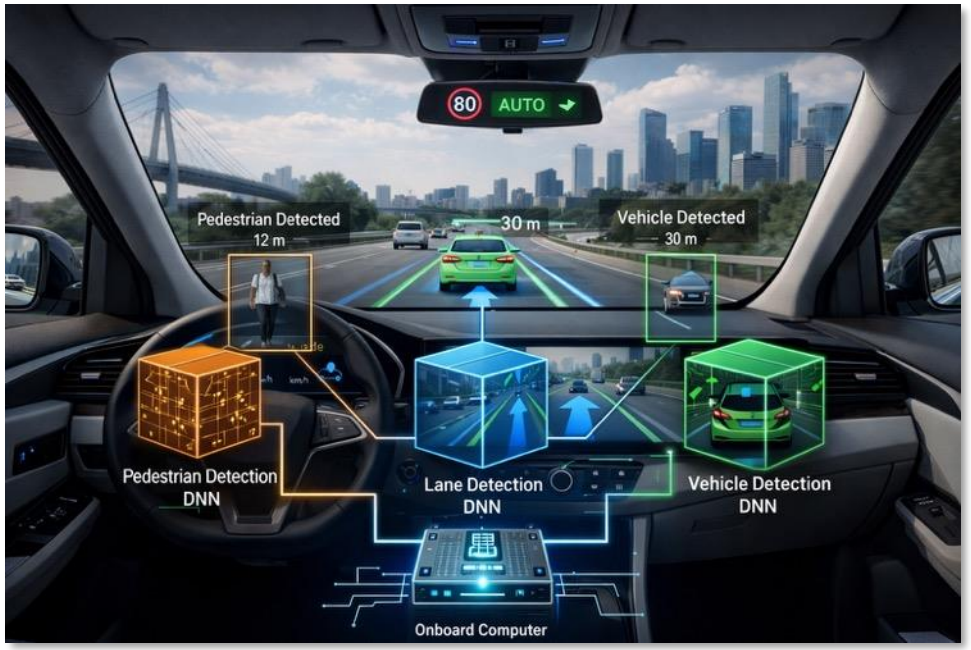
To Overlay or to Customize?
Revisiting Architectural Choices in Heterogeneous Systems

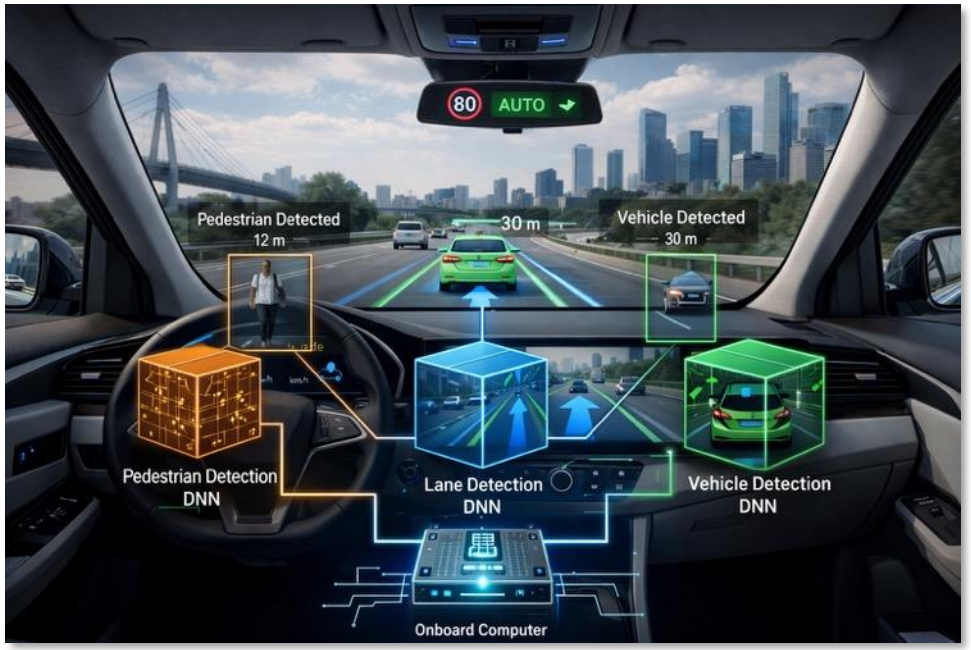
Xingzhen Chen, Shixin Ji, Zheng Dong*, and Peipei Zhou

Brown University; Wayne State University*

xingzhen_chen@brown.edu
peipei_zhou@brown.edu

https://peipeizhou-eecs.github.io/


80
AUTO
Pedestrian Detected
12 m
30 m
Vehicle Detected
30 m
Pedestrian Detection
DNN
Lane Detection
DNN
Vehicle Detection
DNN
Onboard Computer



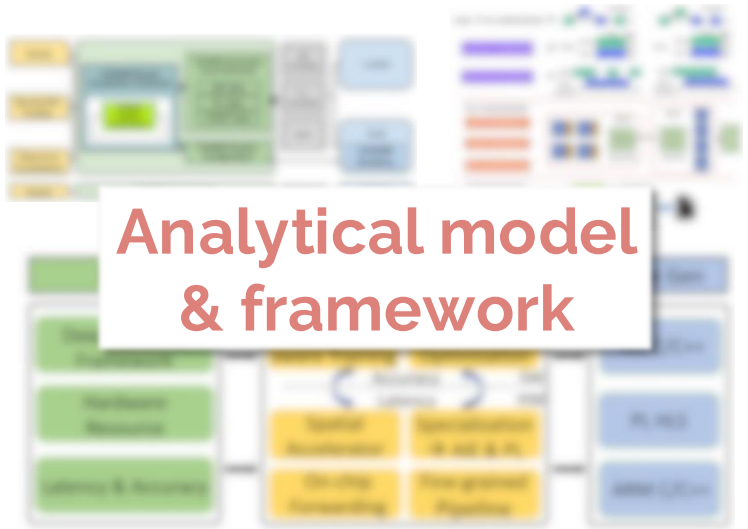
Tasks behind modern autonomous driving systems.

Design Choices

- **Customization accelerators:**

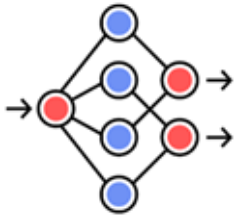
Design Choices

- **Customization accelerators:**

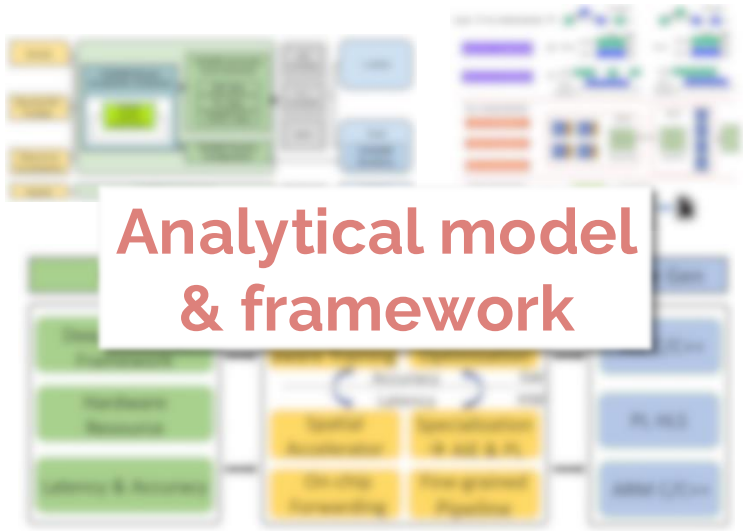


Design Choices

- **Customization accelerators:**

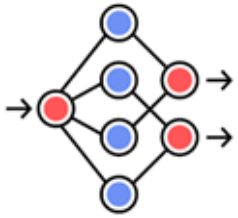


DNN model 1

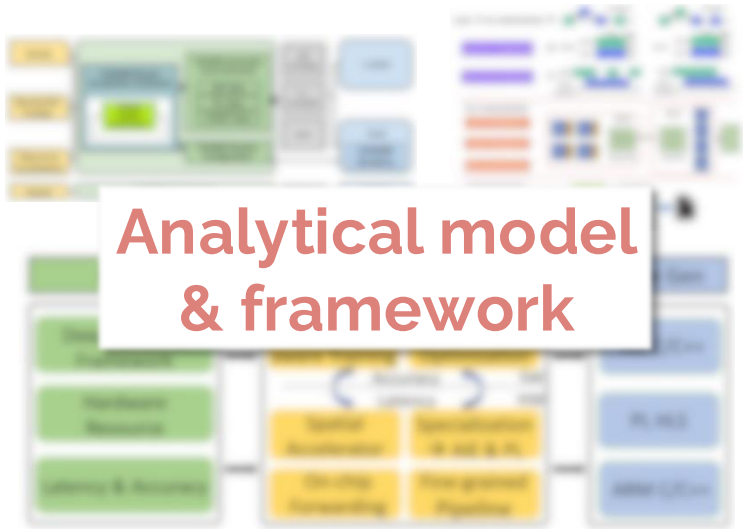


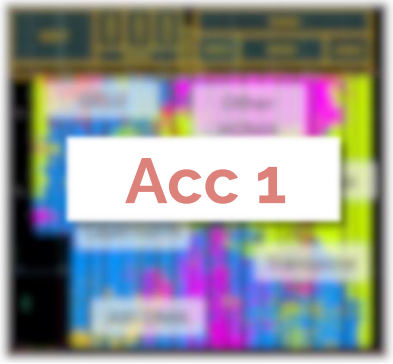
Design Choices

- **Customization accelerators:**



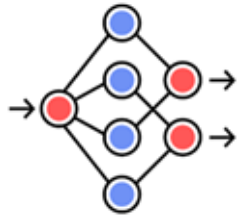
DNN model 1



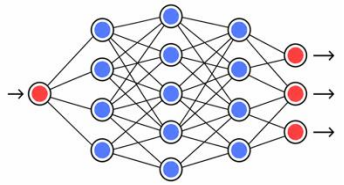


Design Choices

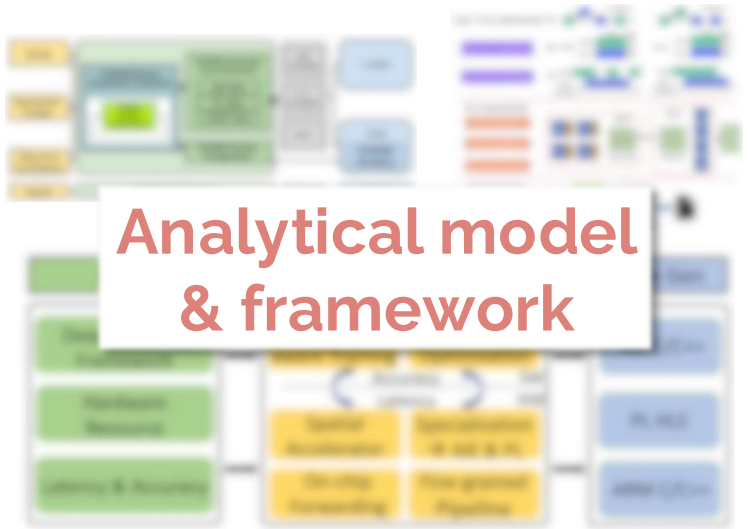
- **Customization accelerators:**



DNN model 1

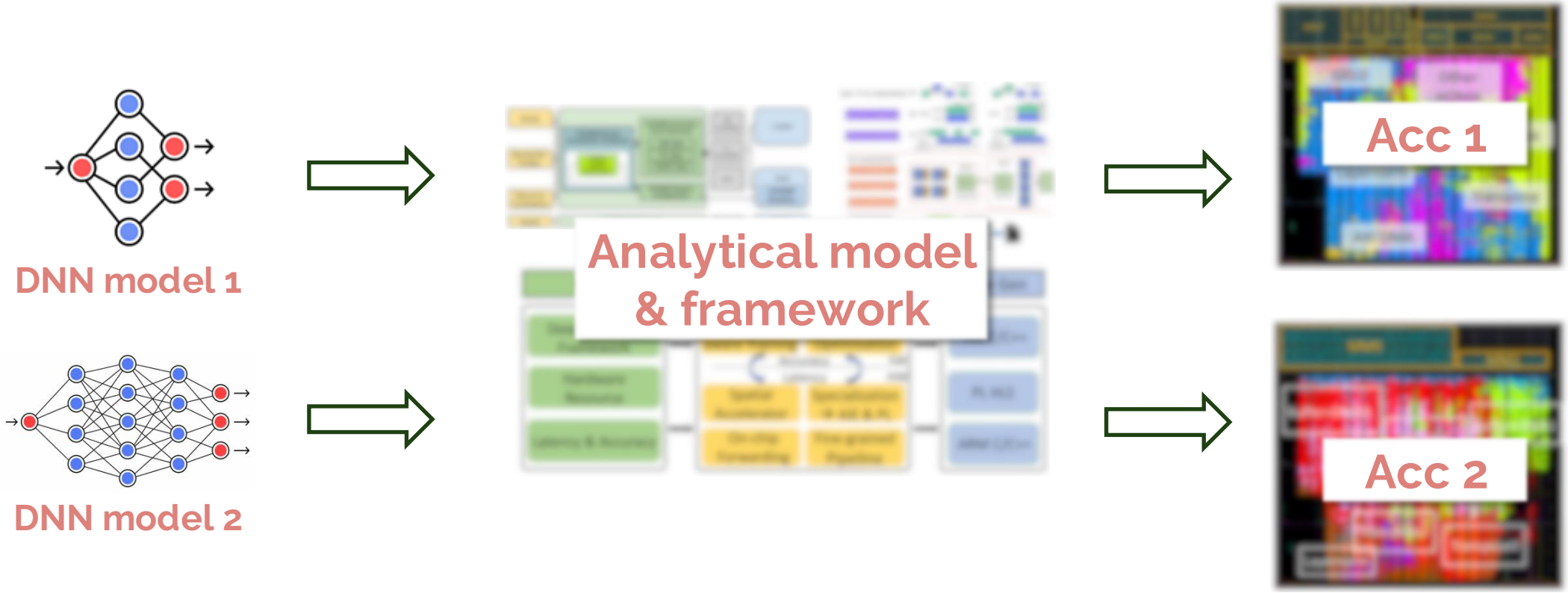


DNN model 2



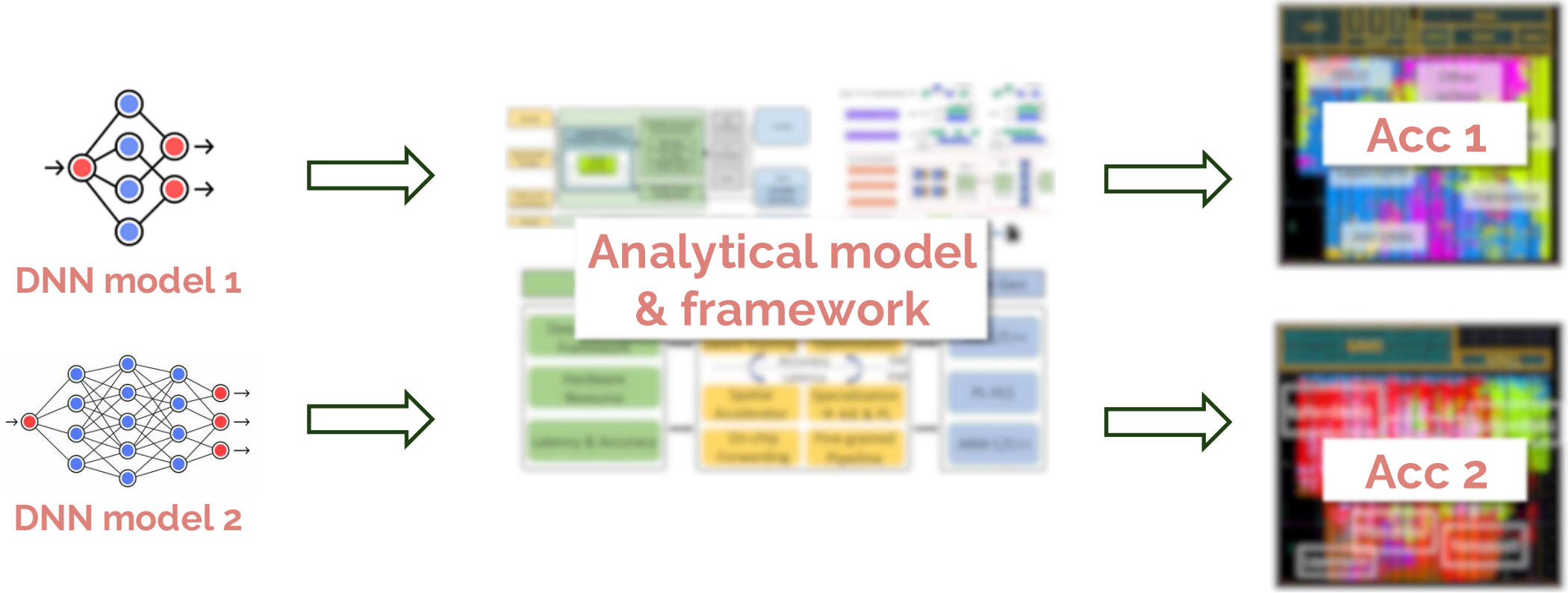
Design Choices

- **Customization accelerators:**



Design Choices

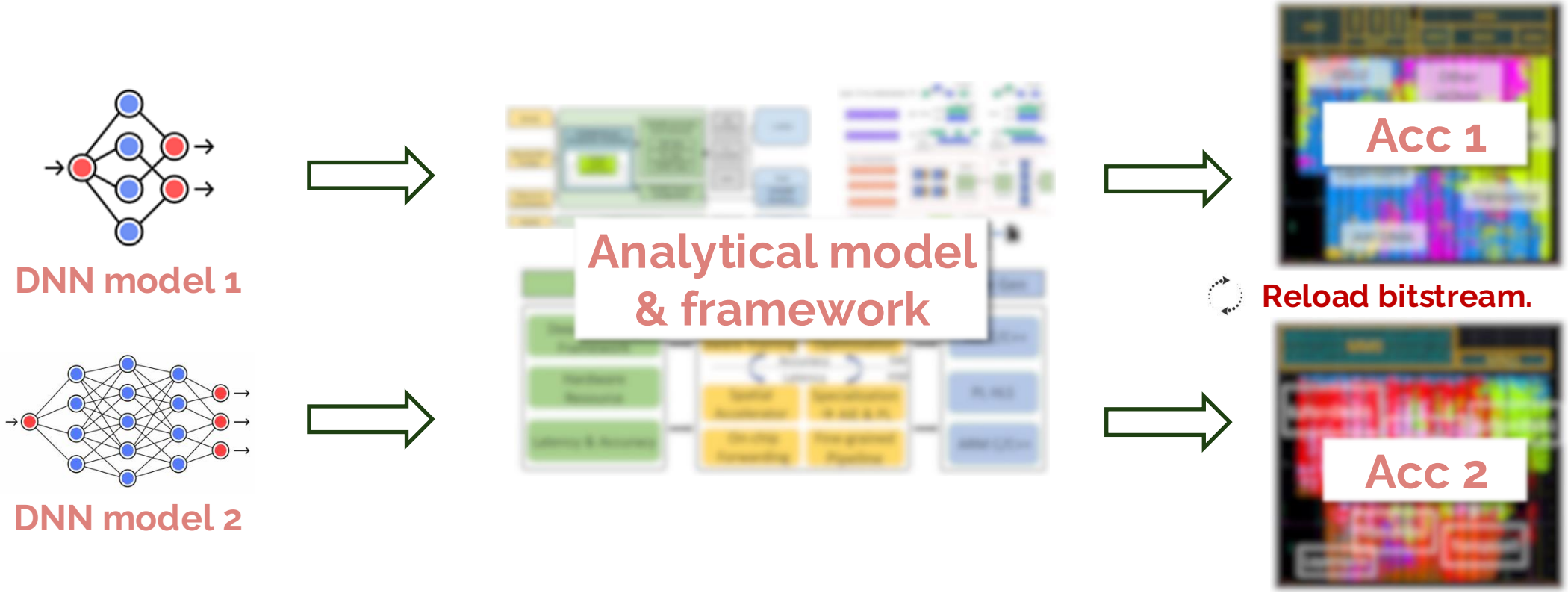
- **Customization accelerators:**



- **Dedicated architectures are instantiated for different applications.**

Design Choices

- **Customization accelerators:**



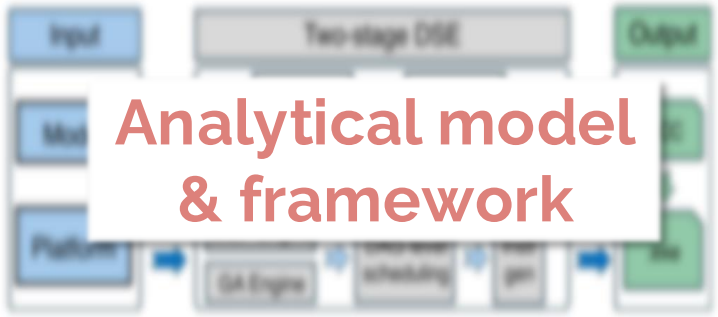
- **Dedicated architectures are instantiated for different applications.**

Design Choices

- **Overlay accelerators:**

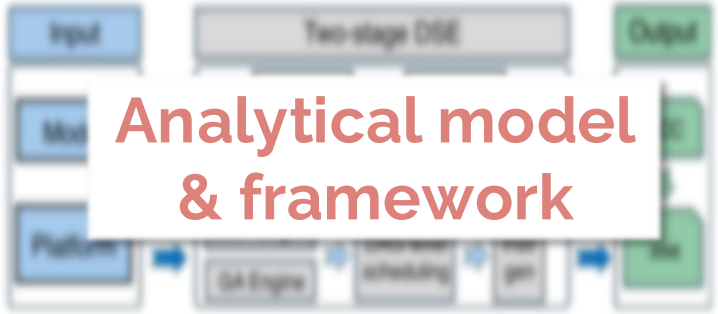
Design Choices

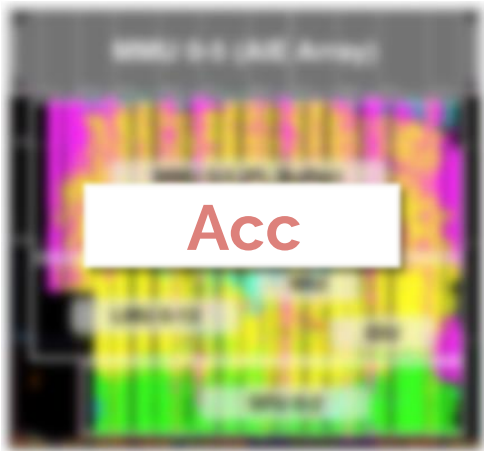
- **Overlay accelerators:**



Design Choices

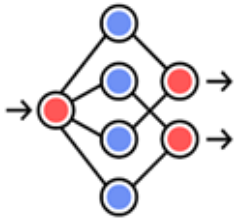
- **Overlay accelerators:**



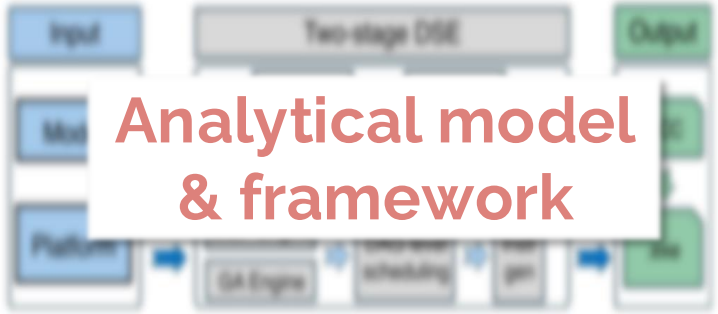


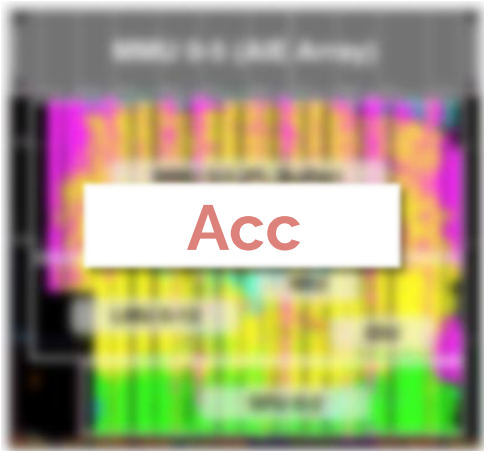
Design Choices

- **Overlay accelerators:**



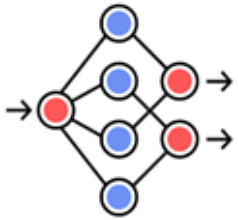
DNN model 1



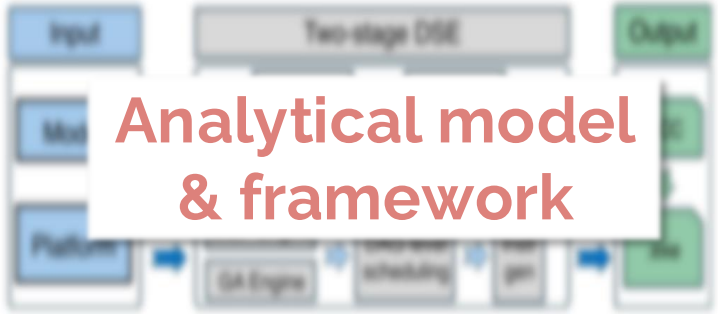


Design Choices

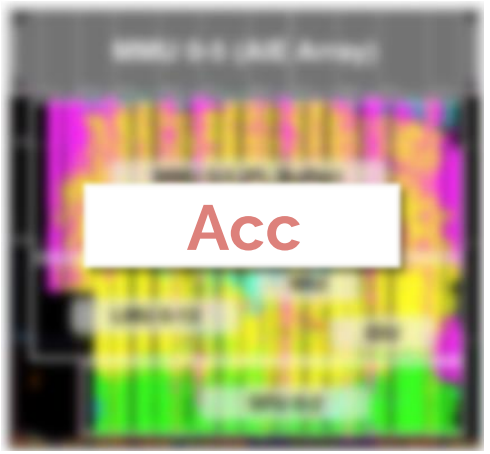
- **Overlay accelerators:**



DNN model 1

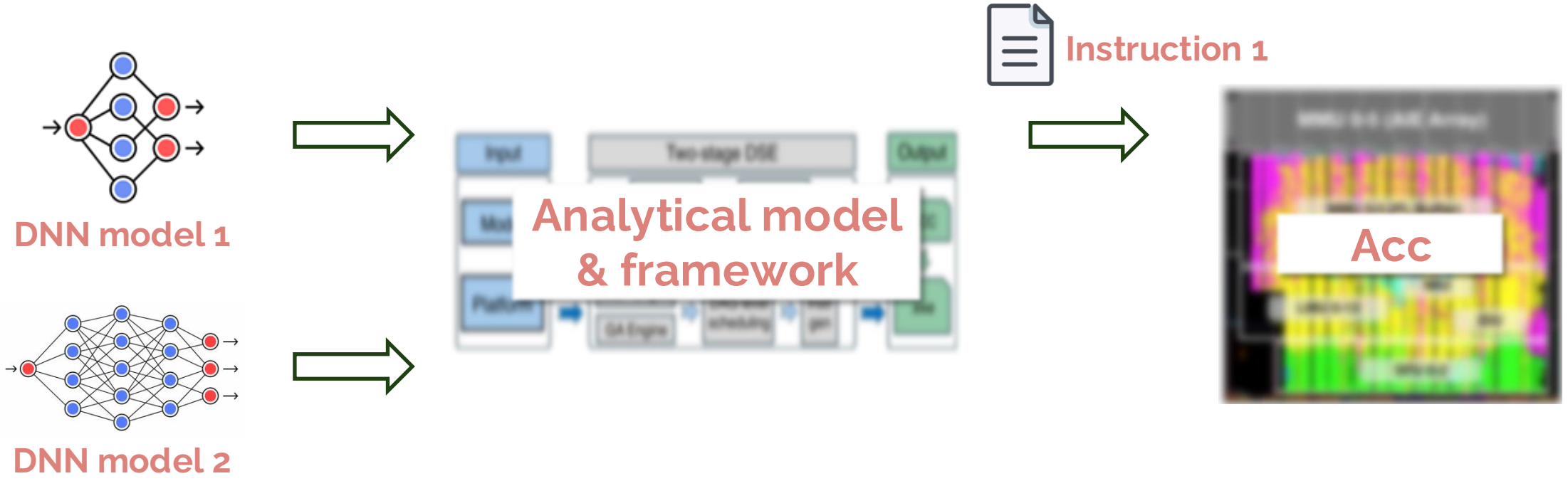


Instruction 1



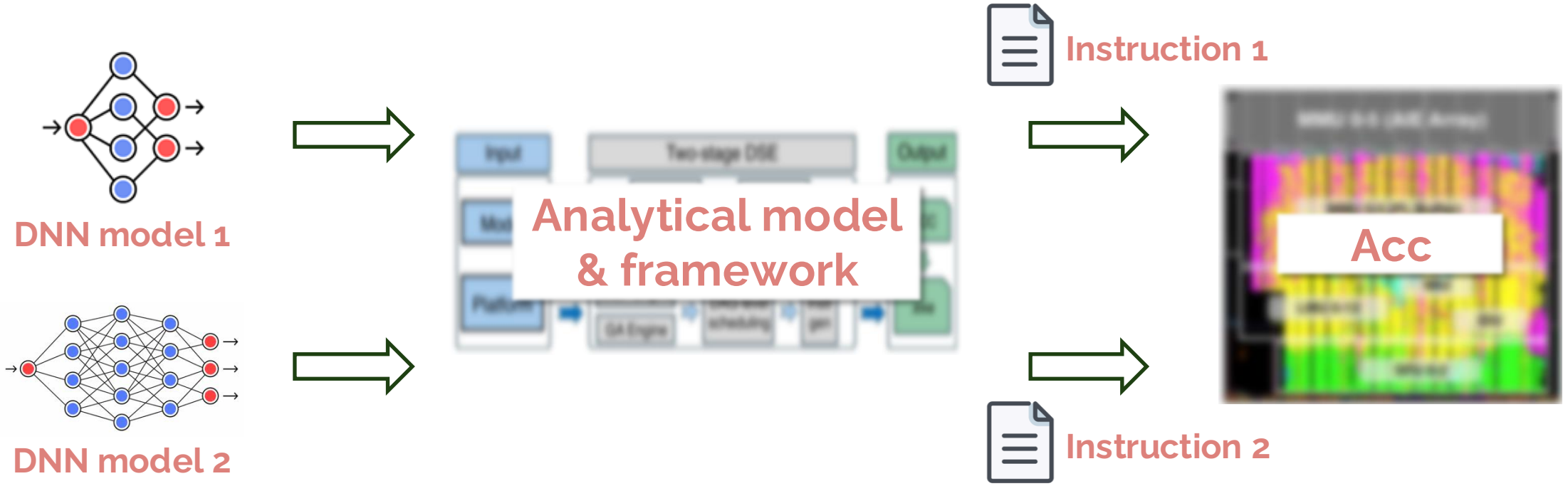
Design Choices

- **Overlay accelerators:**



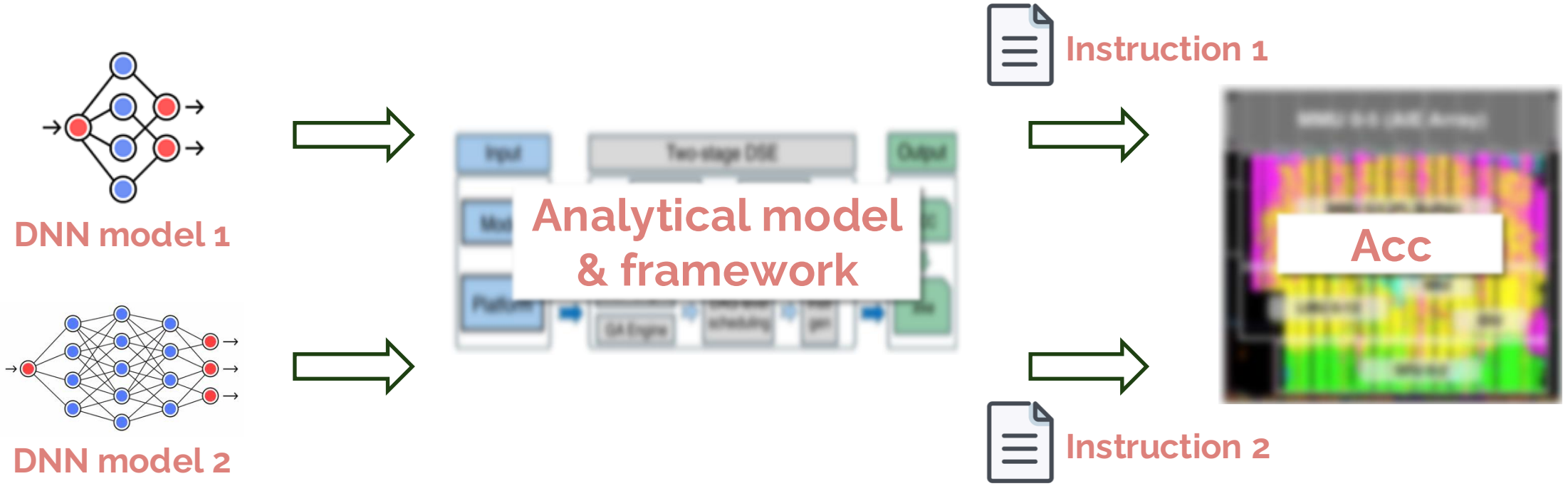
Design Choices

- **Overlay accelerators:**



Design Choices

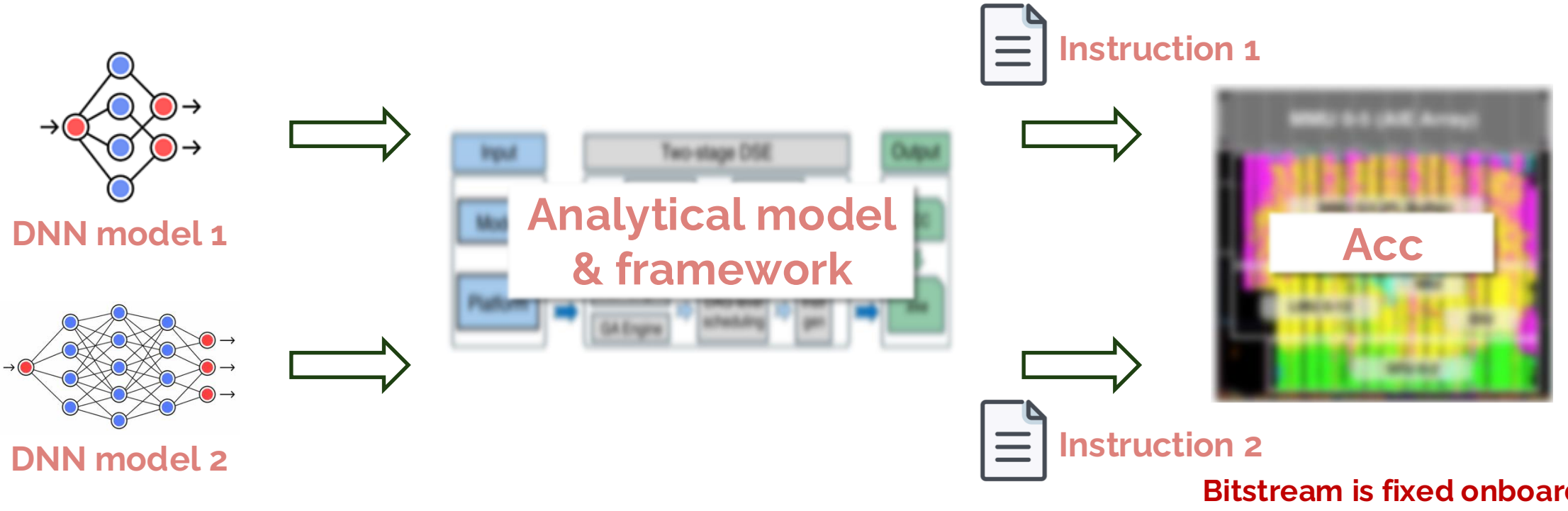
- **Overlay accelerators:**



- **Different applications share the same architecture with different instructions.**

Design Choices

- **Overlay accelerators:**



- **Different applications share the same architecture with different instructions.**

Design Choices

Design Choices

- There exists a **trade-off** between two design choices.

Customization accelerators		Overlay accelerators

Design Choices

- There exists a trade-off between two design choices.

Customization accelerators	Overlay accelerators
■ Dedicated customization for specific DNN models. ■ Better efficiency and performance. ■ Worse flexibility and extra overhead.	

Design Choices

- There exists a trade-off between two design choices.

Customization accelerators

- Dedicated customization for specific DNN models.
- Better efficiency and performance.
- Worse flexibility and extra overhead.

Overlay accelerators

- Multiple DNN models share the same architecture.
- Better flexibility and negligible overhead.
- Worse efficiency and performance.

Design Choices

- There exists a trade-off between two design choices.

Customization accelerators	Overlay accelerators
■ Dedicated customized hardware for specific DNN models. ■ Better efficiency and performance. ■ Worse flexibility and extra overhead.	■ Support multiple DNNs of the same architecture. ■ Better flexibility and negligible overhead. ■ Worse efficiency and performance.

To overlay or to Customization?

Real-time Systems

Real-time Systems

- We focus on real-time + autonomous driving systems.

Real-time Systems

- We focus on real-time + autonomous driving systems.

Real-time Systems

- We focus on real-time + autonomous driving systems.

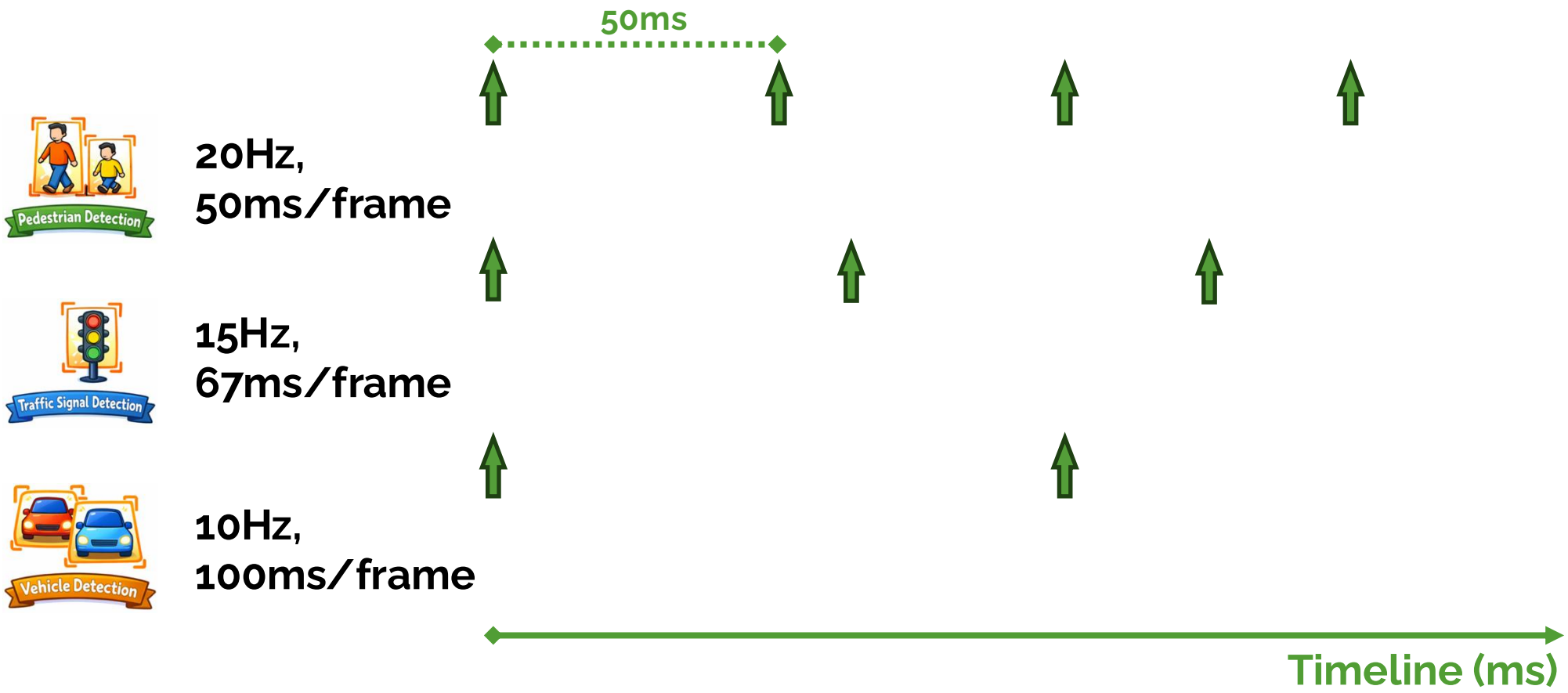
20Hz,
50ms/frame

15Hz,
67ms/frame

10Hz,
100ms/frame

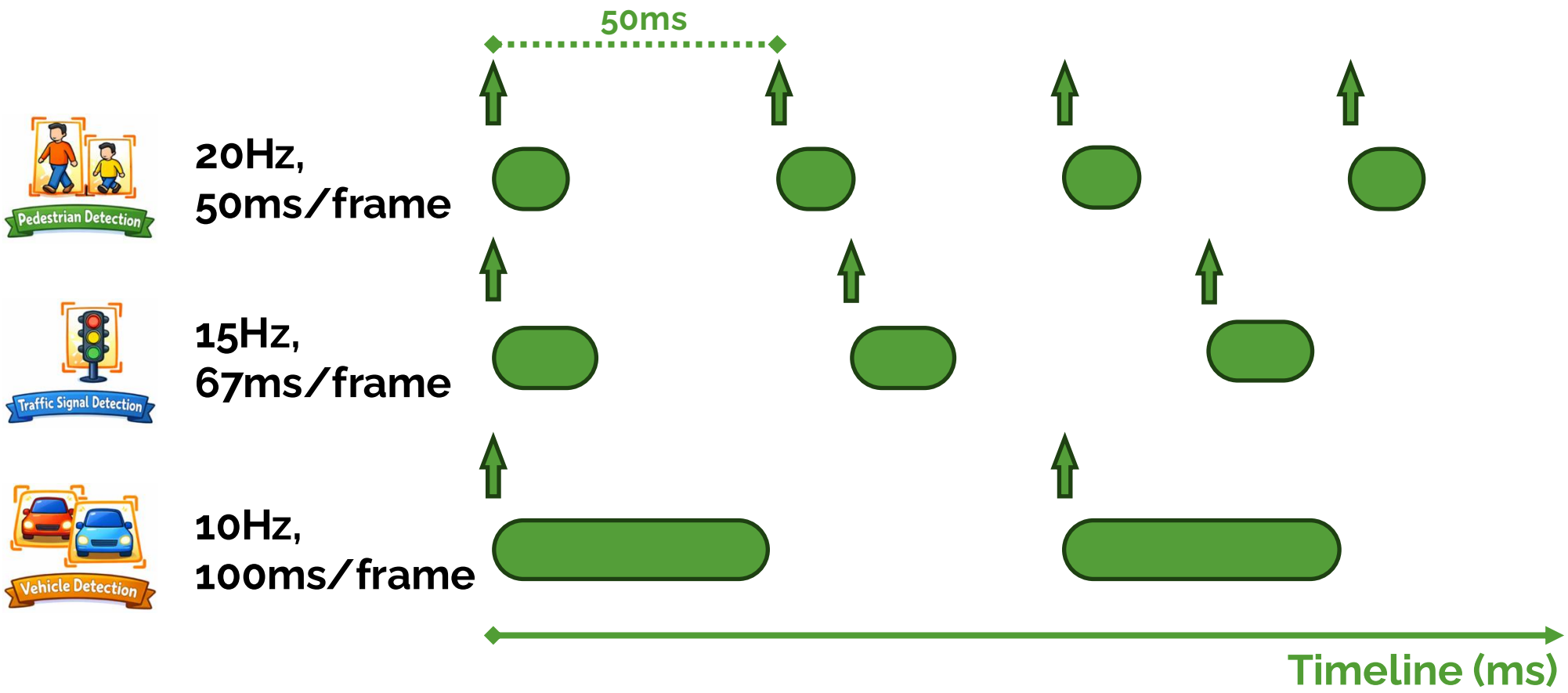
Real-time Systems

- We focus on real-time + autonomous driving systems.



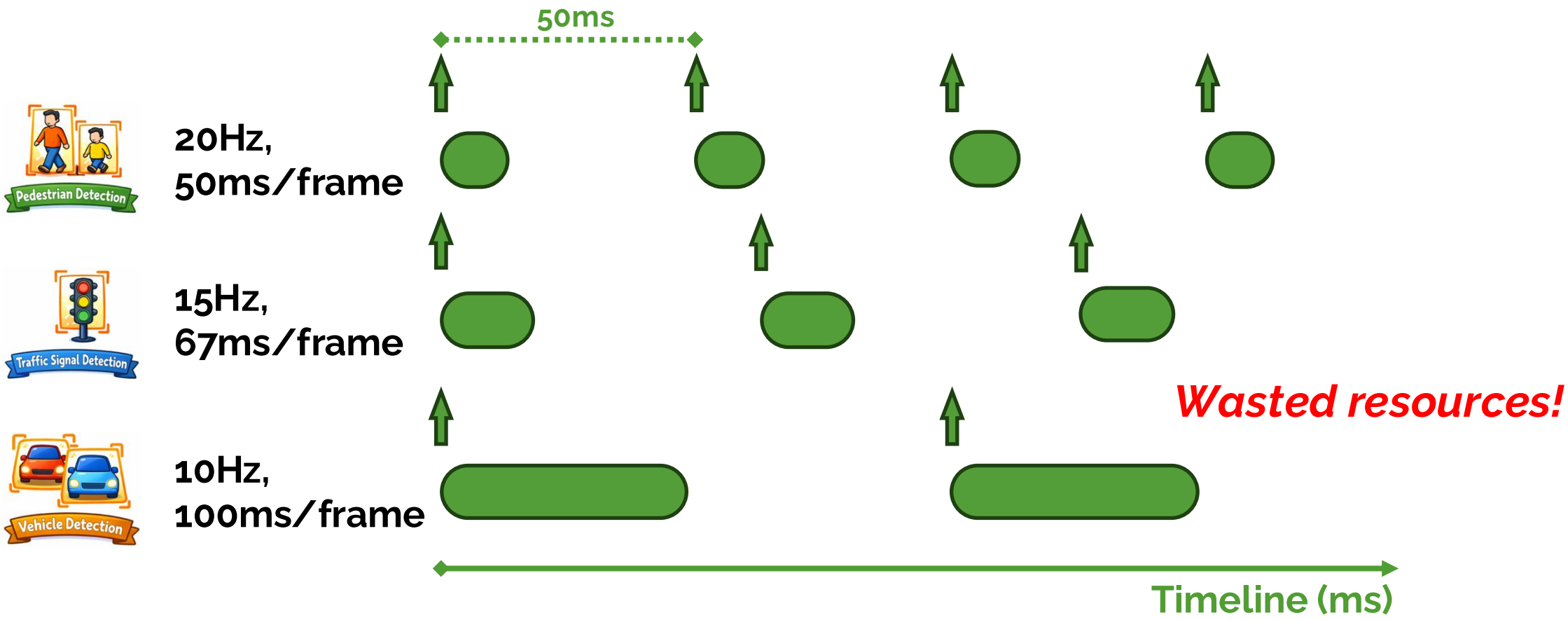
Real-time Systems

- We focus on real-time + autonomous driving systems.



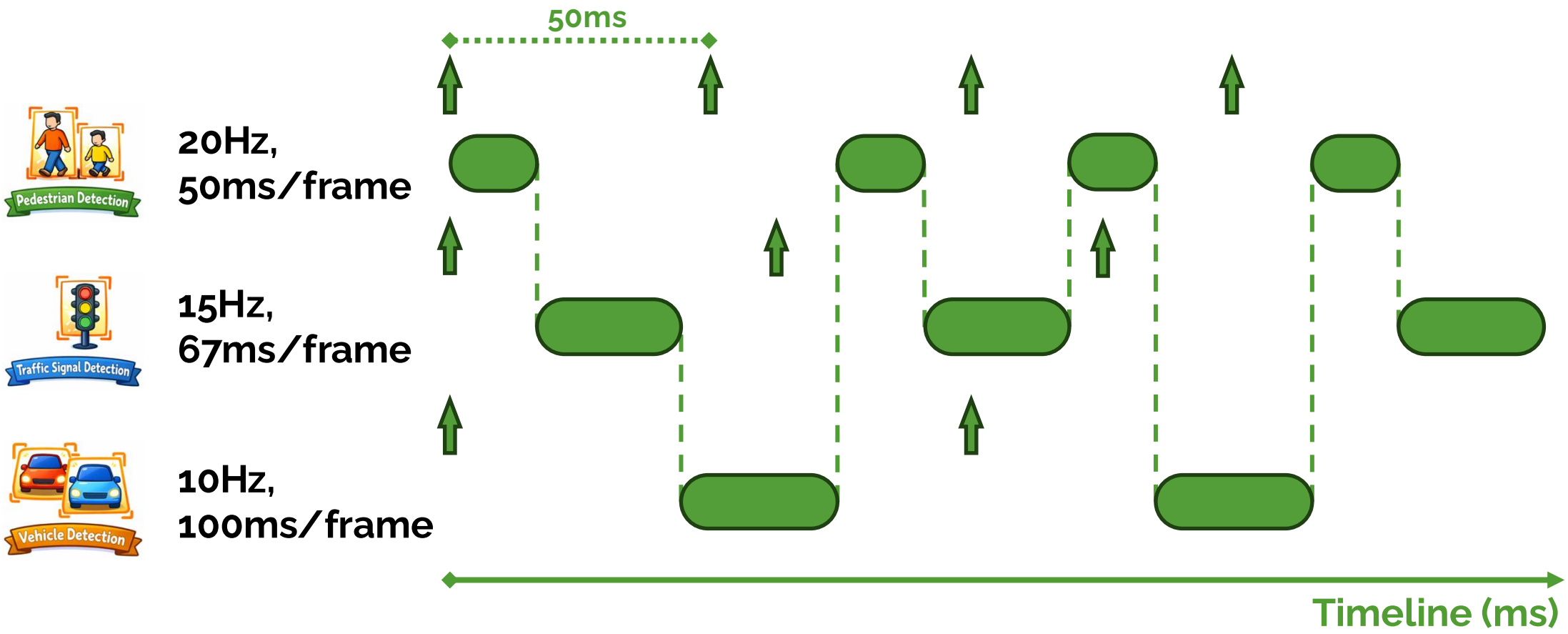
Real-time Systems

- We focus on real-time + autonomous driving systems.



Real-time Systems

- We focus on real-time + autonomous driving systems.



To Overlay or to Customize?

Revisiting Architectural Choices in Heterogeneous Systems

Challenges

Solutions

To Overlay or to Customize?

Revisiting Architectural Choices in Heterogeneous Systems

Challenges	Solutions
▪ The system requires periodic execution of multiple tasks.	▪ Integrate real-time systems into modeling.
▪ Diverse scenarios, complicated hardware, and flexible objectives.	▪ Build an analysis tool that takes safety and resources as constraints, and determines the optimal choices.
▪ Technique improvements may flip the design choices.	▪ Explore improvements in terms of key features.

Experimental Setup

Experimental Setup

- ADS Setup: Periodic tasks with hard deadline constraints.

Periodic tasks models

Experimental Setup

- **ADS Setup**: Periodic tasks with hard deadline constraints.
- **Applications Setup**: DeiT[1], MLP-Mixer[2], PointNet[3].

Periodic tasks models

	DeiT-S	DeiT-L	MLP-Mixer-S	MLP-Mixer-L	PointNet-S	PointNet-L
Setup A		20Hz		15Hz		10Hz
Setup B	20Hz			15Hz		10Hz
Setup C	20Hz			15Hz	10Hz	
Setup D	20Hz		15Hz		10Hz	

Representative applications

Experimental Setup

- **ADS Setup**: Periodic tasks with hard deadline constraints.
- **Applications Setup**: DeiT[1], MLP-Mixer[2], PointNet[3].
- **Architecture Setup**: CHARM[4], SSR[5], DORA[6].

Periodic tasks models

	DeiT-S	DeiT-L	MLP-Mixer-S	MLP-Mixer-L	PointNet-S	PointNet-L
Setup A		20Hz		15Hz		10Hz
Setup B	20Hz			15Hz		10Hz
Setup C	20Hz			15Hz	10Hz	
Setup D	20Hz		15Hz		10Hz	

Representative applications

State-of-the-art Accs.

Experimental Setup

- **ADS Setup**: Periodic tasks with hard deadline constraints.
- **Applications Setup**: DeiT[1], MLP-Mixer[2], PointNet[3].
- **Architecture Setup**: CHARM[4], SSR[5], DORA[6].
- **Evaluation Metric**: Busy ratio.

Periodic tasks models

	DeiT-S	DeiT-L	MLP-Mixer-S	MLP-Mixer-L	PointNet-S	PointNet-L
Setup A		20Hz		15Hz		10Hz
Setup B	20Hz			15Hz		10Hz
Setup C	20Hz			15Hz	10Hz	
Setup D	20Hz		15Hz		10Hz	

Representative applications

State-of-the-art Accs.

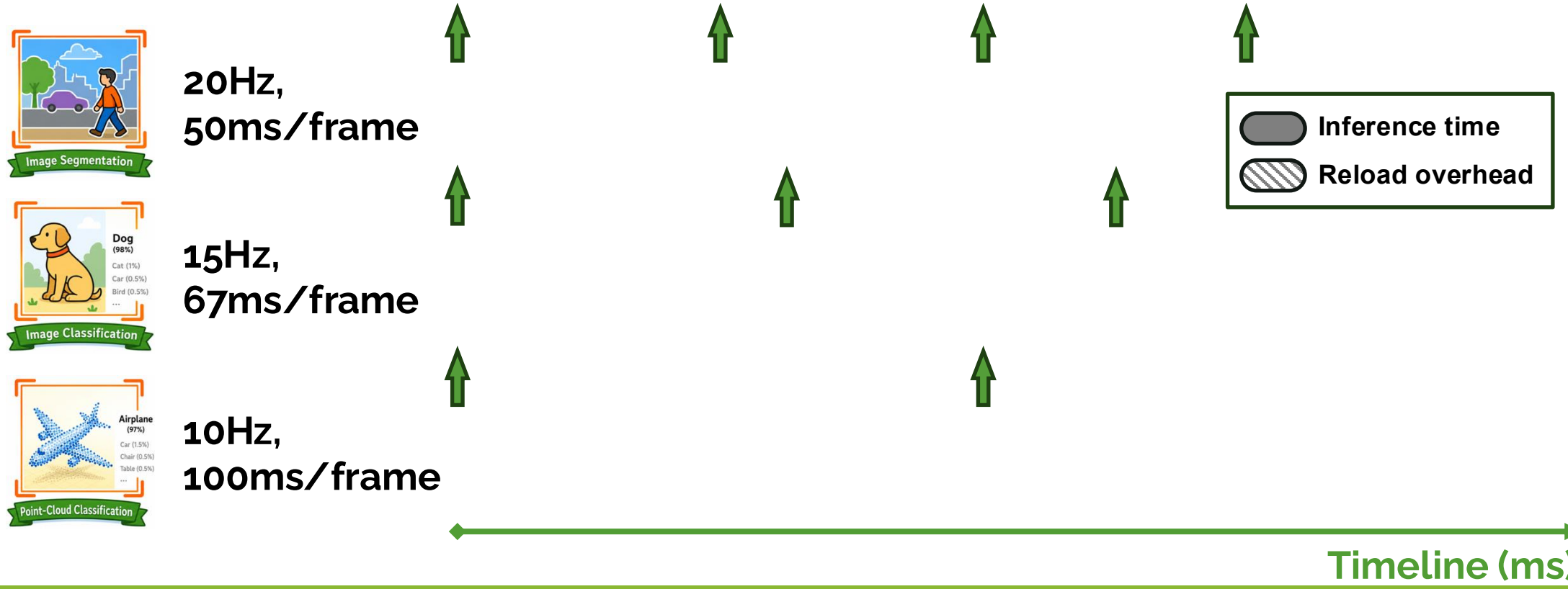
Trade-off with state-of-the-art techniques

Trade-off with state-of-the-art techniques

- Instantiate dedicated acc and reload bitstream with 20ms overhead.

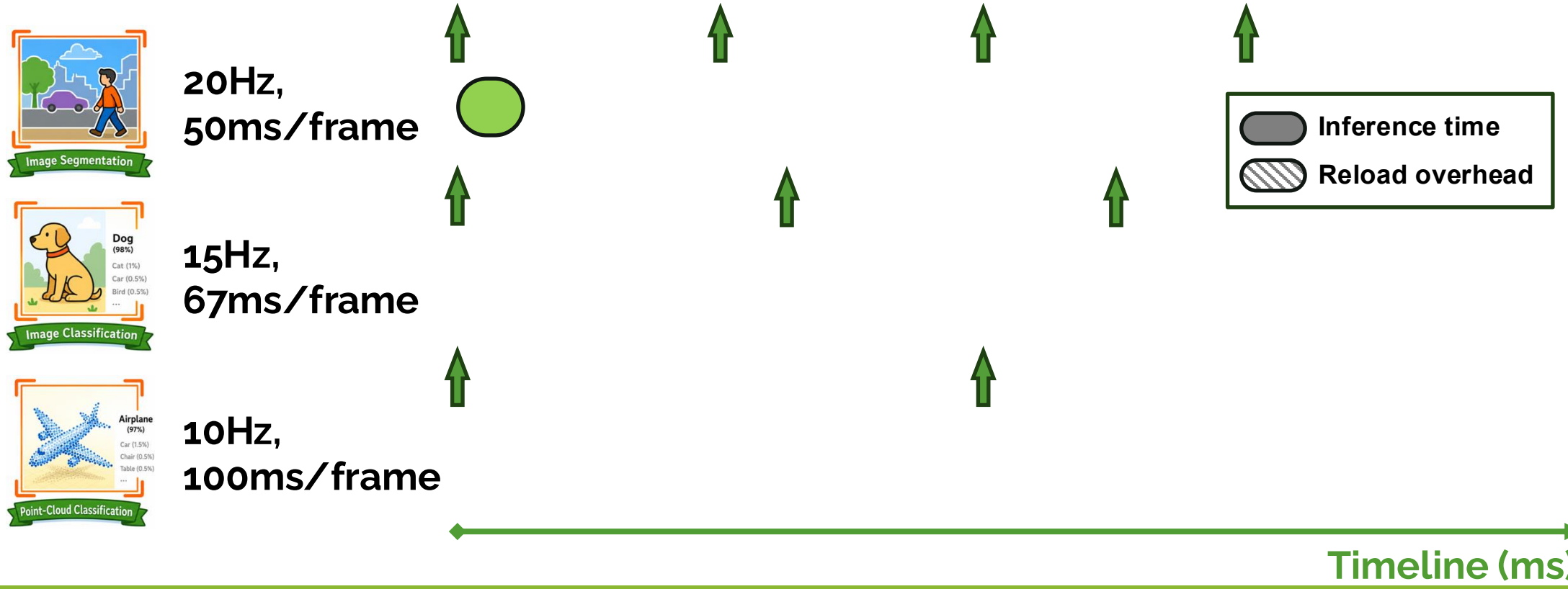
Trade-off with state-of-the-art techniques

- Instantiate dedicated acc and reload bitstream with 20ms overhead.



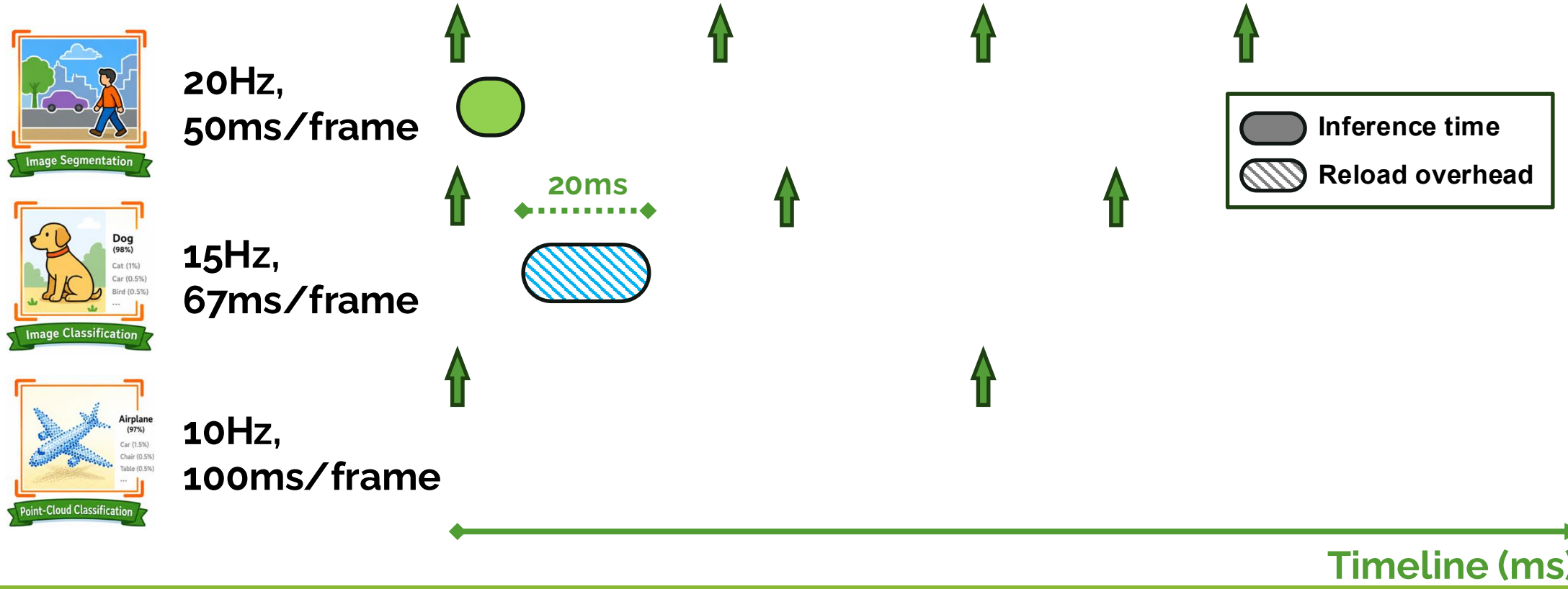
Trade-off with state-of-the-art techniques

- Instantiate dedicated acc and reload bitstream with 20ms overhead.



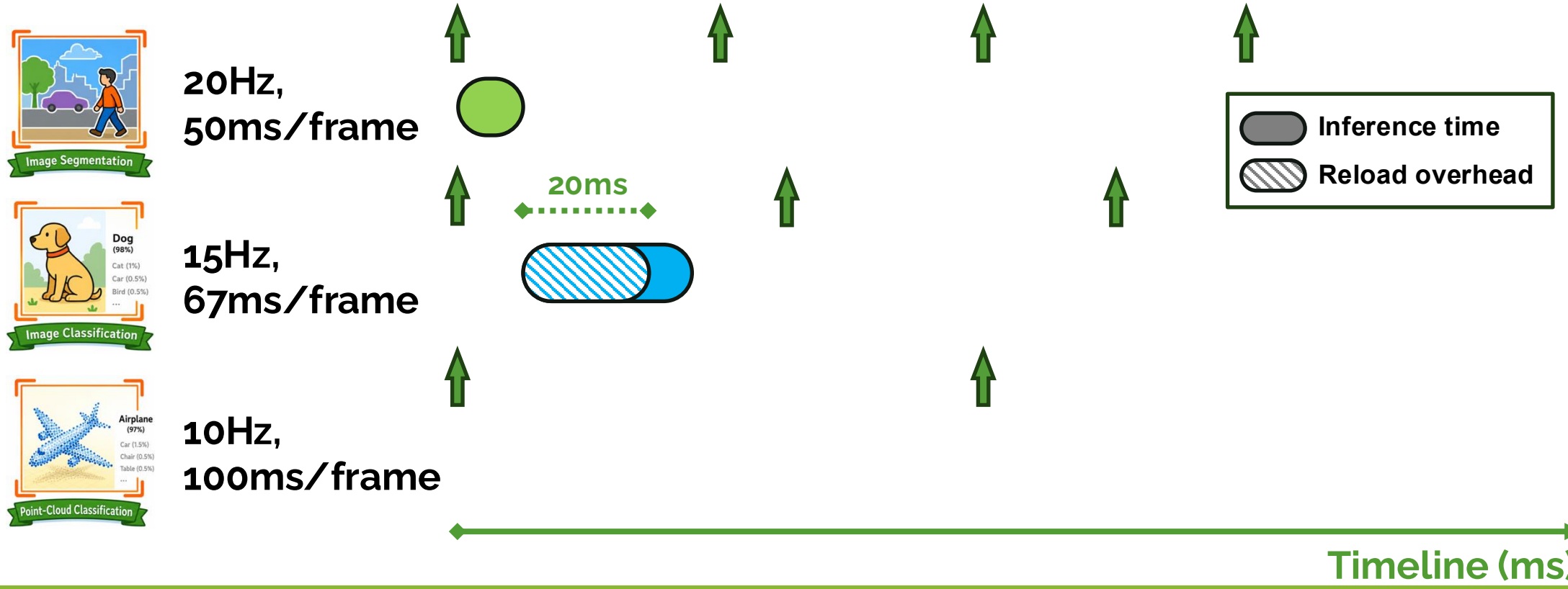
Trade-off with state-of-the-art techniques

- Instantiate dedicated acc and reload bitstream with 20ms overhead.



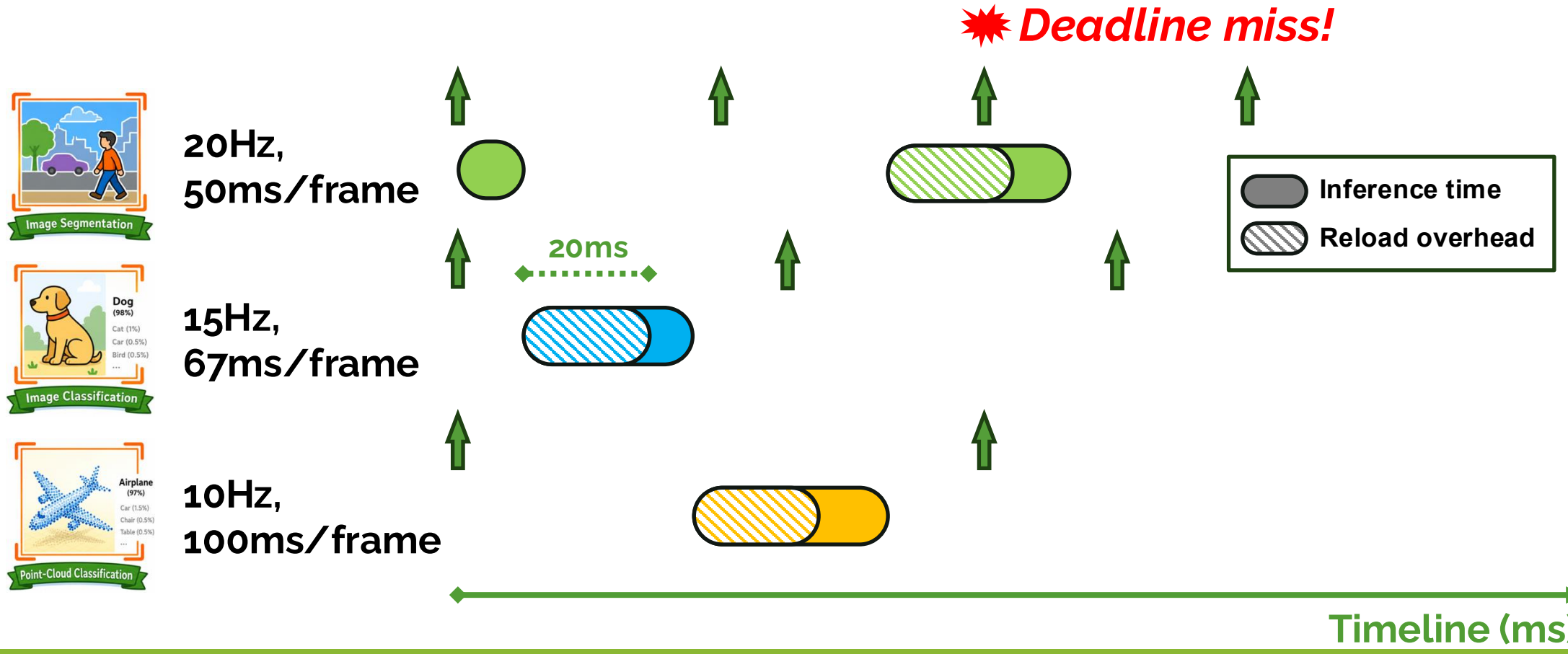
Trade-off with state-of-the-art techniques

- Instantiate dedicated acc and reload bitstream with 20ms overhead.



Trade-off with state-of-the-art techniques

- Instantiate dedicated acc and reload bitstream with 20ms overhead.



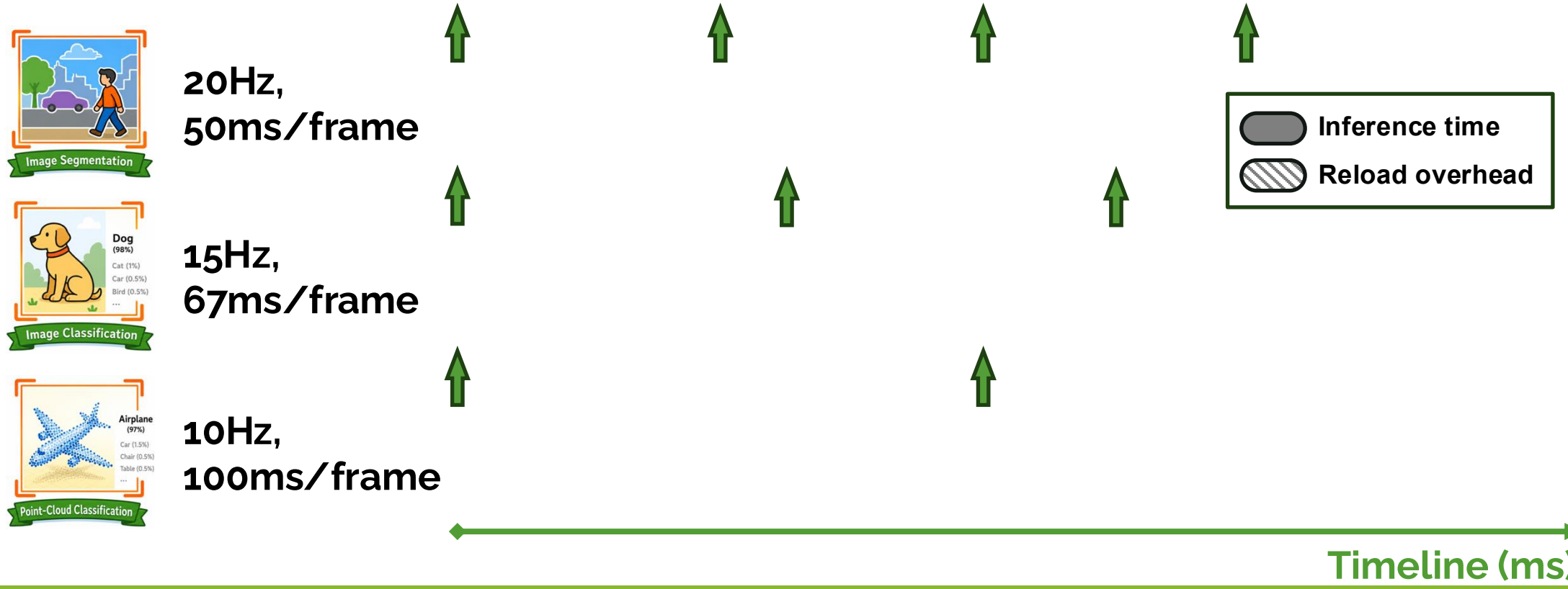
Trade-off with state-of-the-art techniques

Trade-off with state-of-the-art techniques

- Three tasks share an overlay acc with only instruction switches.

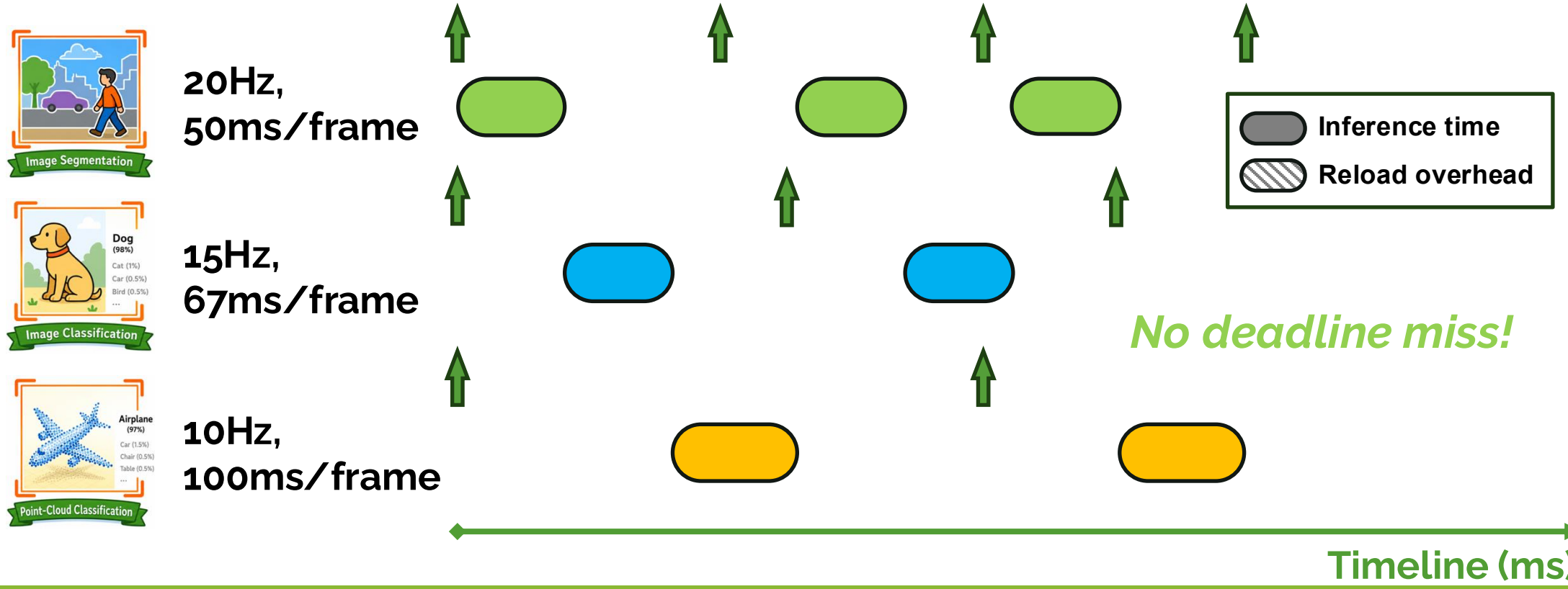
Trade-off with state-of-the-art techniques

- Three tasks share an overlay acc with only instruction switches.



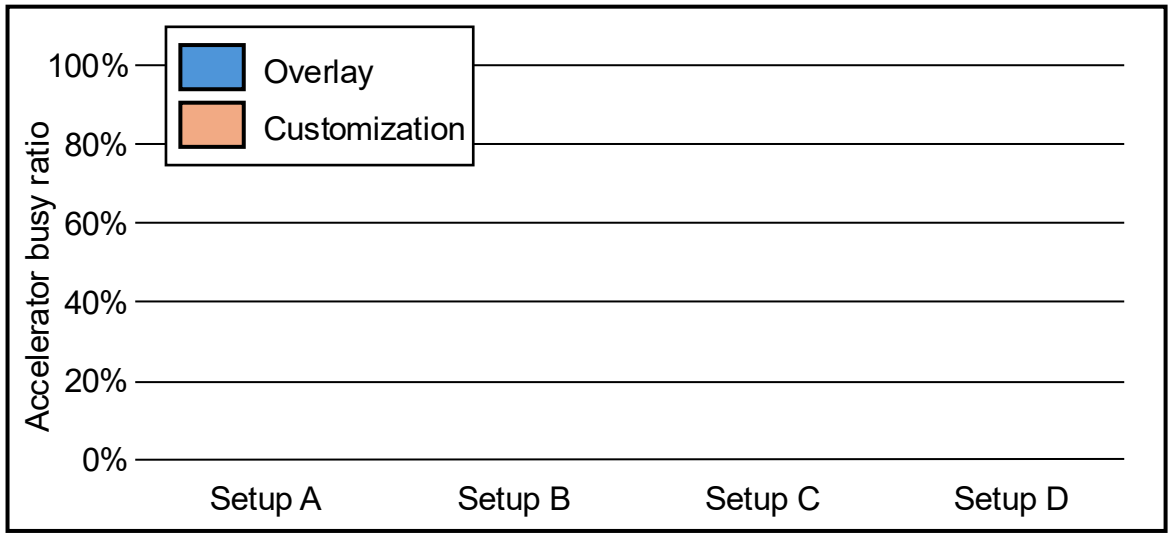
Trade-off with state-of-the-art techniques

- Three tasks share an overlay acc with only instruction switches.



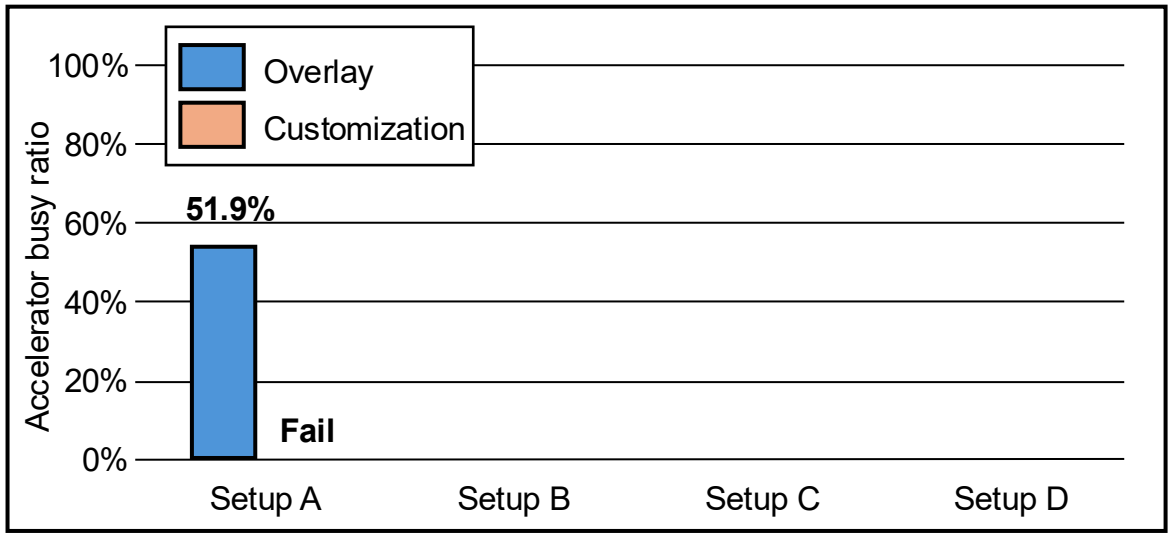
Trade-off with state-of-the-art techniques

Trade-off with state-of-the-art techniques



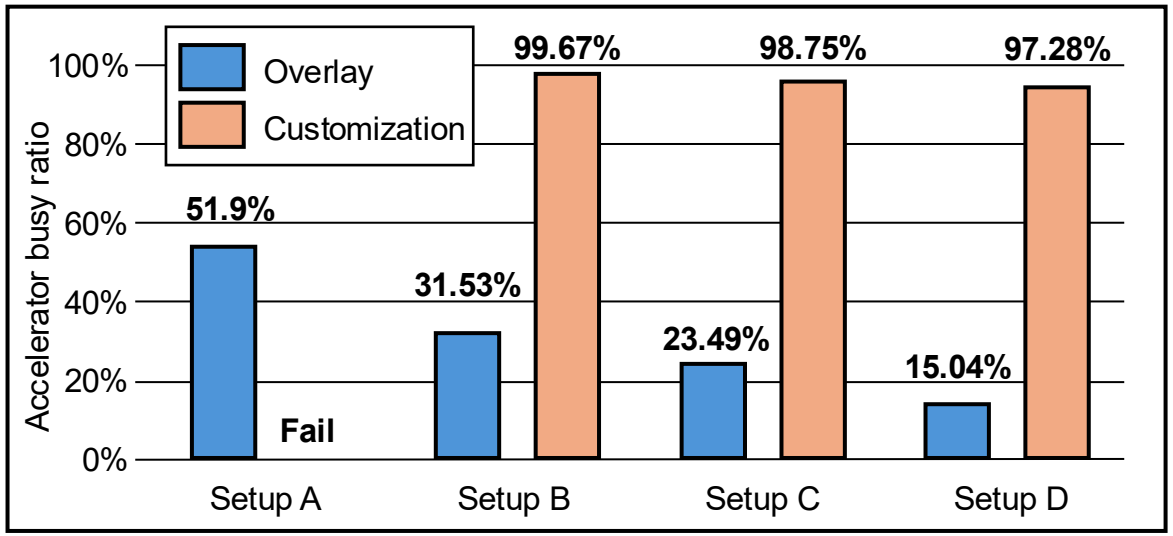
Trade-off with state-of-the-art techniques

- **Setup A: Customization choice fails with deadline miss.**



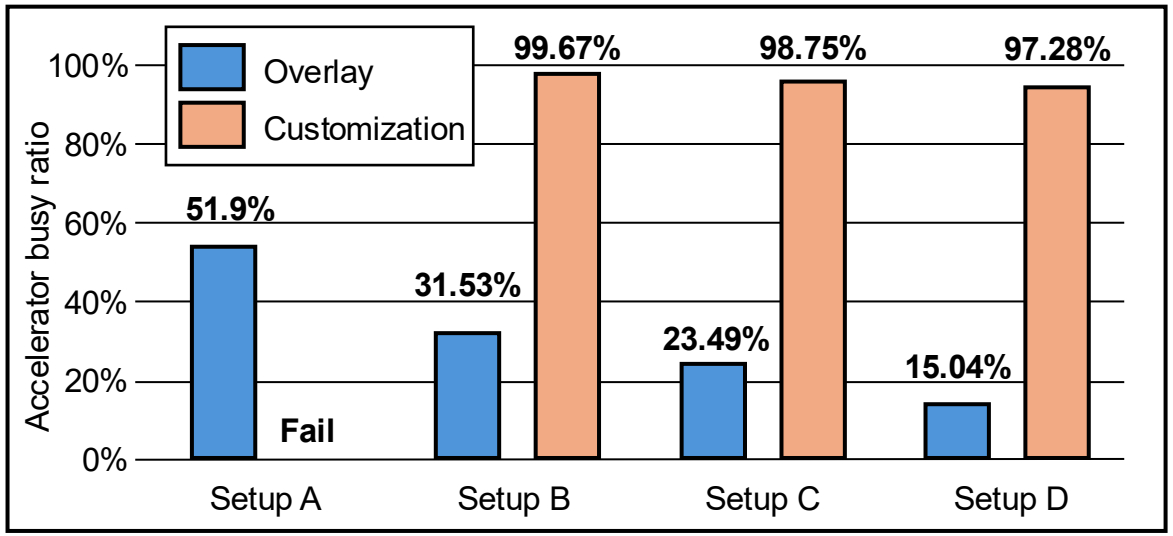
Trade-off with state-of-the-art techniques

- Setup A: Customization choice fails with deadline miss.
- Setup B~D: Bitstream reloading induces much overhead and a higher busy ratio.



Trade-off with state-of-the-art techniques

- Setup A: Customization choice fails with deadline miss.
- Setup B~D: Bitstream reloading induces much overhead and a higher busy ratio.



- Under current techniques, overlay usually outperforms customization in frequent switch scenarios.

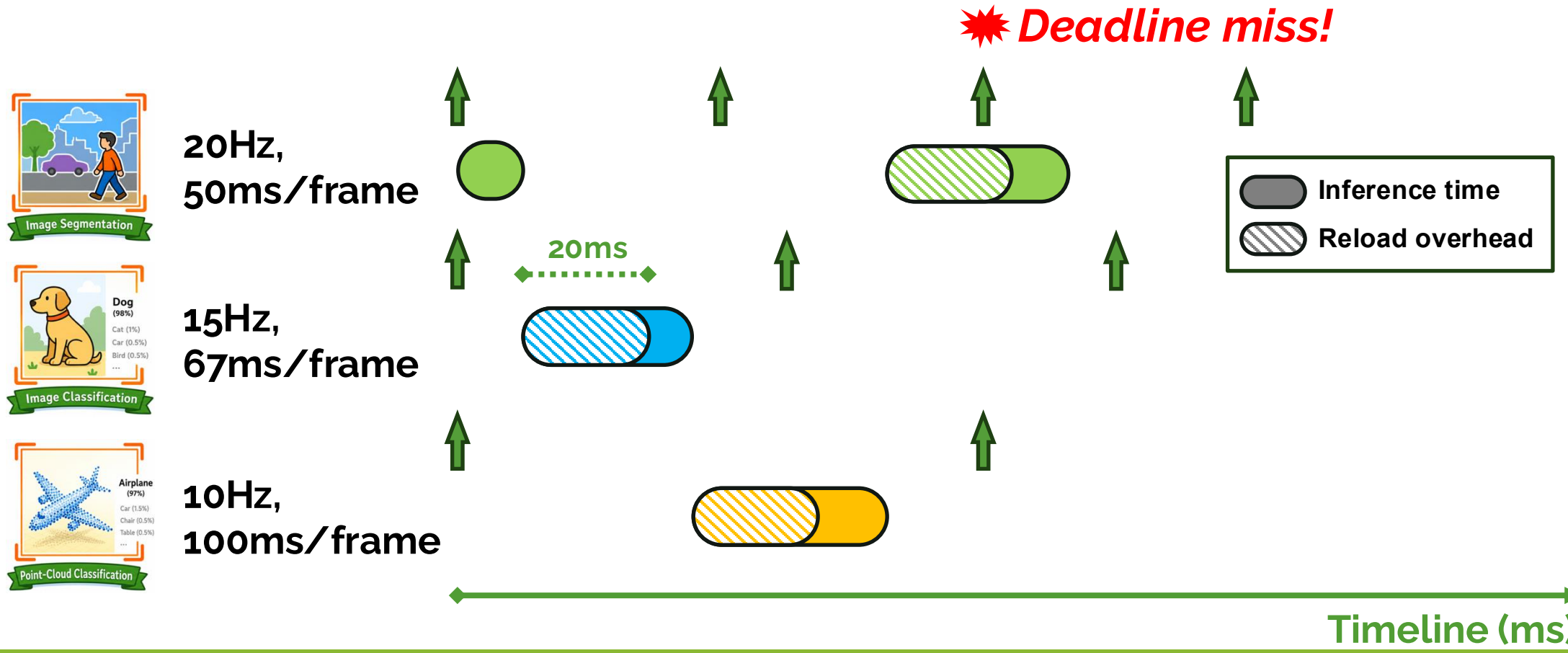
Trade-off with advanced reconfiguration

Trade-off with advanced reconfiguration

- Reconfiguration latency can be reduced in the future.

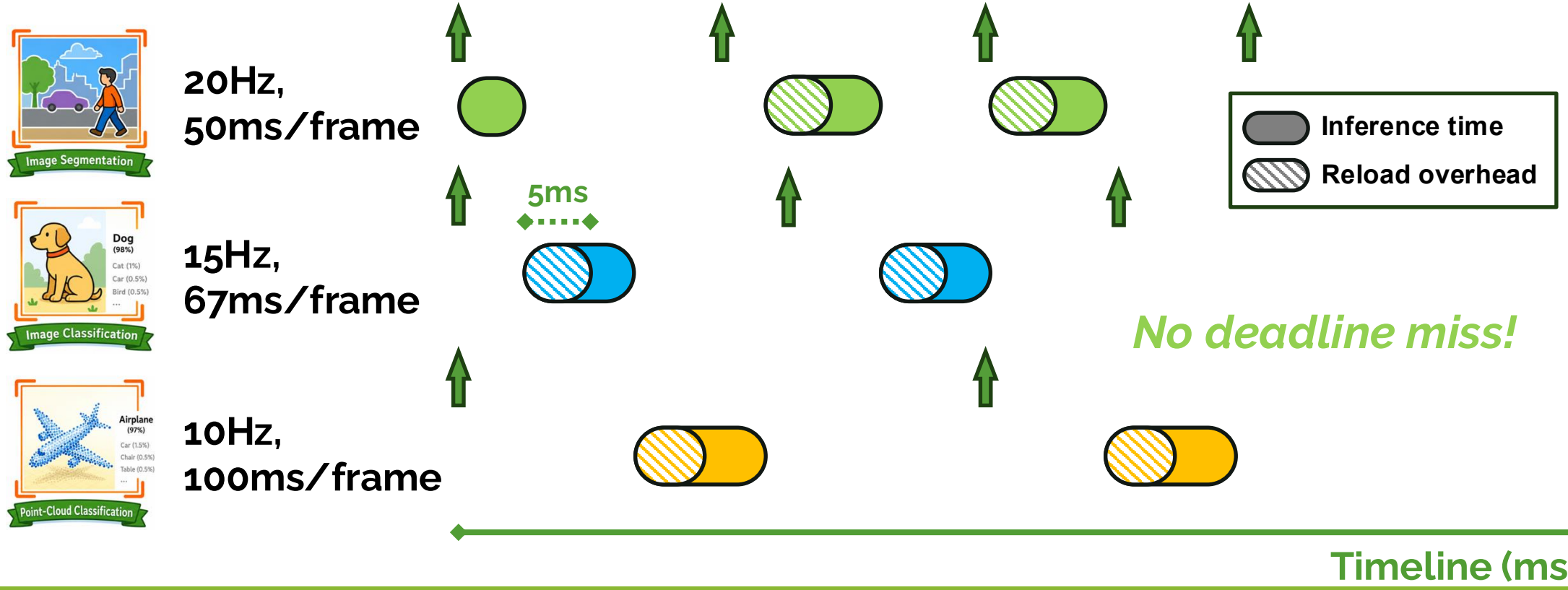
Trade-off with advanced reconfiguration

- Reconfiguration latency can be reduced in the future.



Trade-off with advanced reconfiguration

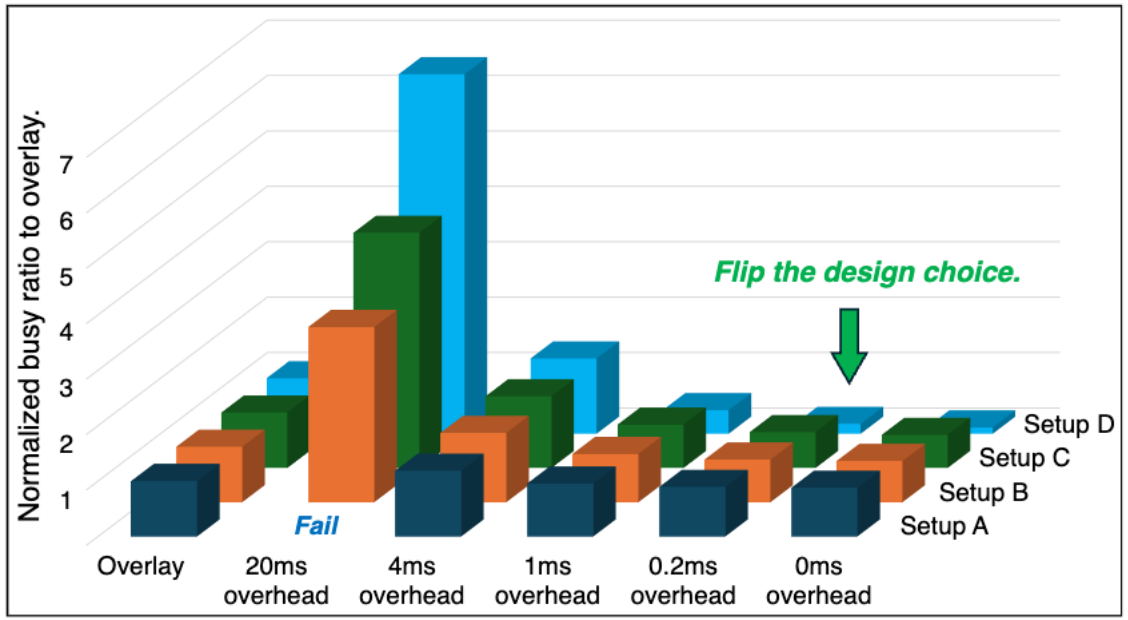
- Reconfiguration latency can be reduced in the future.



Trade-off with advanced reconfiguration

Trade-off with advanced reconfiguration

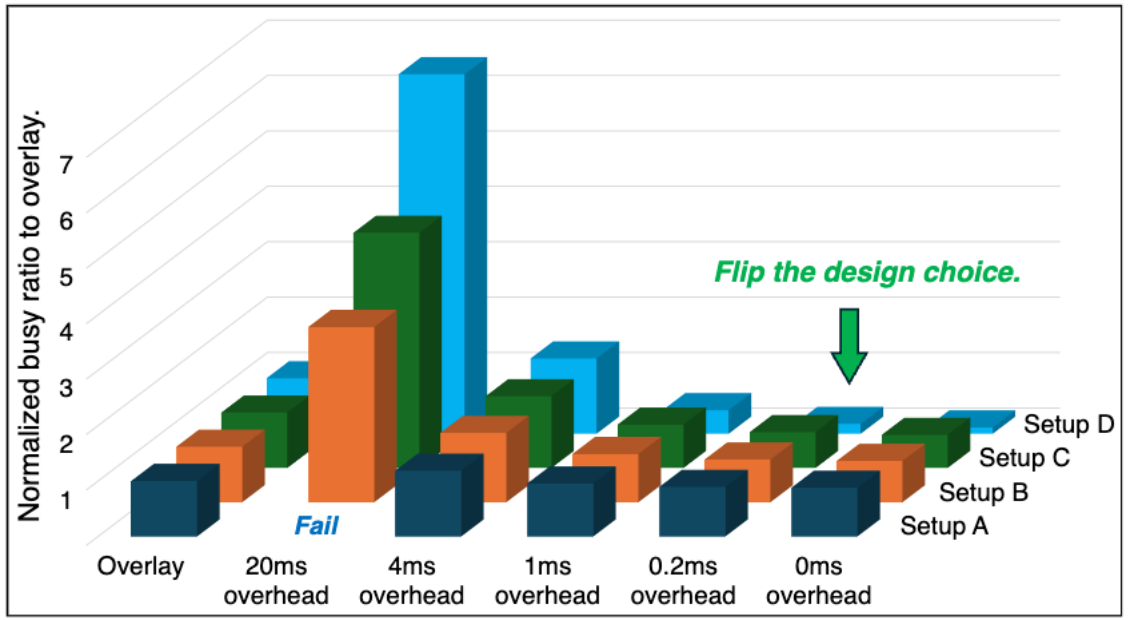
- Experiments on further advanced reconfiguration assumption.



Comparison under advanced reconfiguration techniques.

Trade-off with advanced reconfiguration

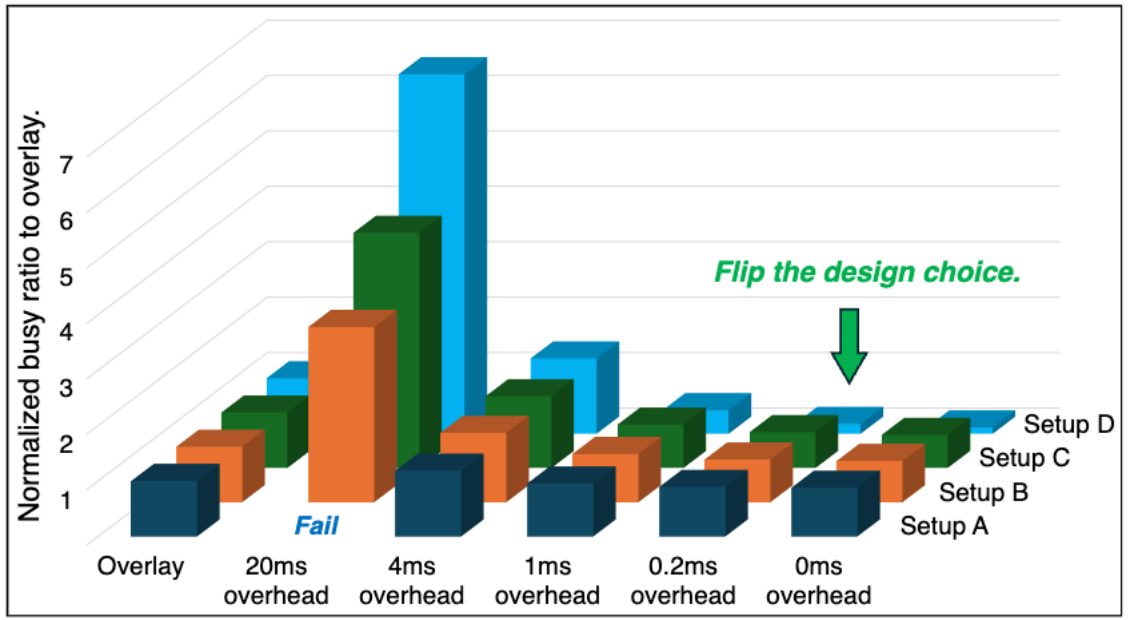
- Experiments on further advanced reconfiguration assumption.
- In current ADS scenarios, faster reconfiguration can reduce the busy ratio.



Comparison under advanced reconfiguration techniques.

Trade-off with advanced reconfiguration

- Experiments on further advanced reconfiguration assumption.
- In current ADS scenarios, faster reconfiguration can reduce the busy ratio.
- 0.2ms reconfiguration latency will flip the design choices.



Comparison under advanced reconfiguration techniques.

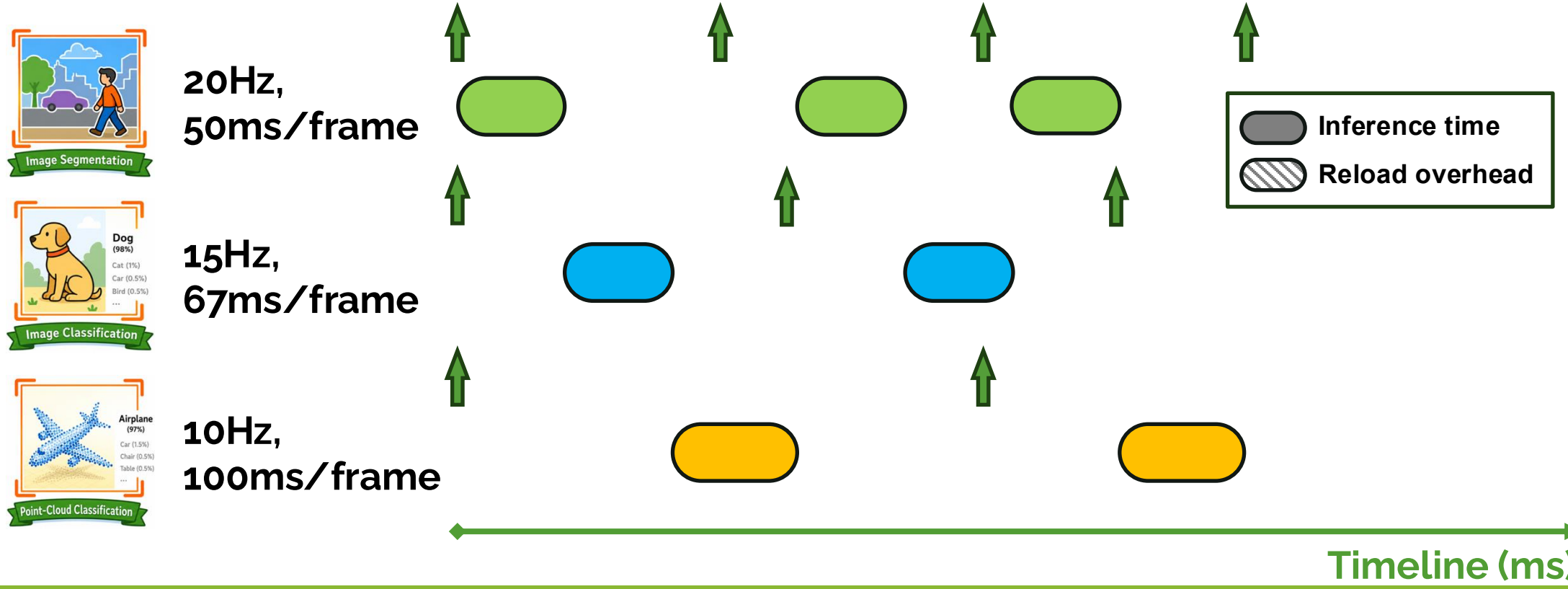
Trade-off with advanced overlay architecture

Trade-off with advanced overlay architecture

- Advanced overlay acc can be proposed in the future.

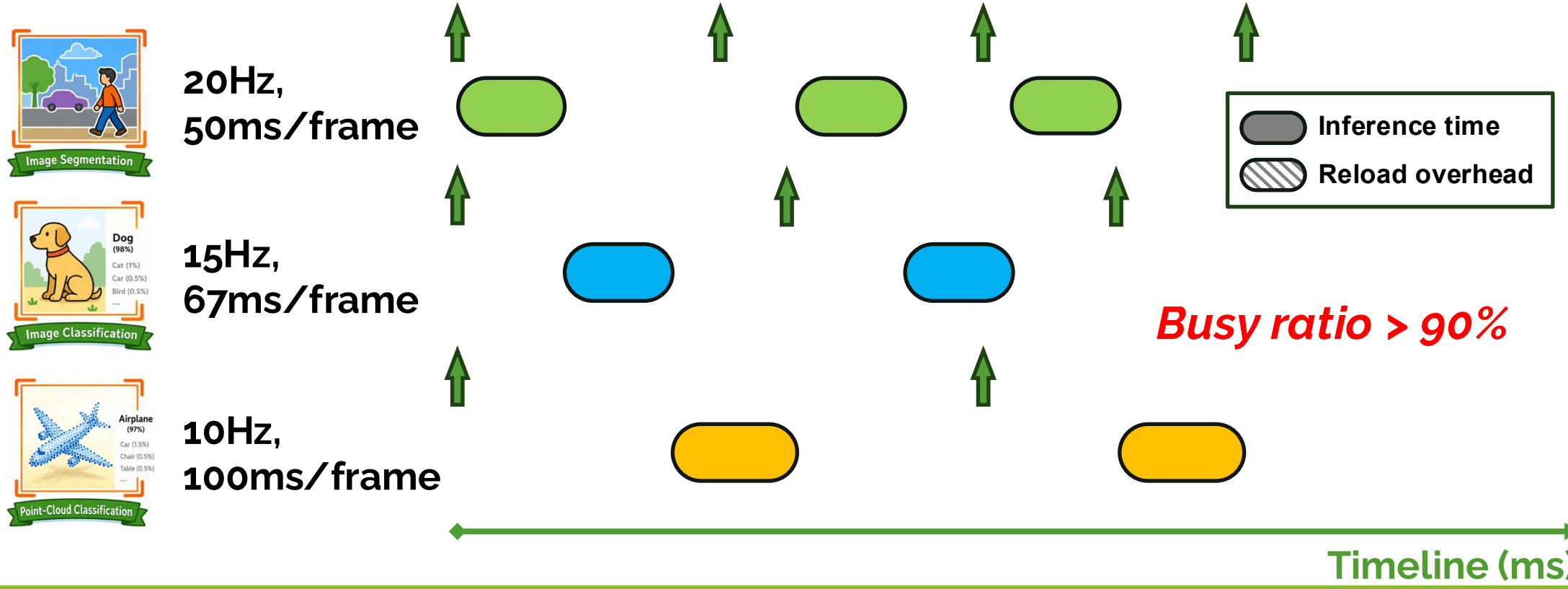
Trade-off with advanced overlay architecture

- **Advanced overlay** acc can be proposed in the future.



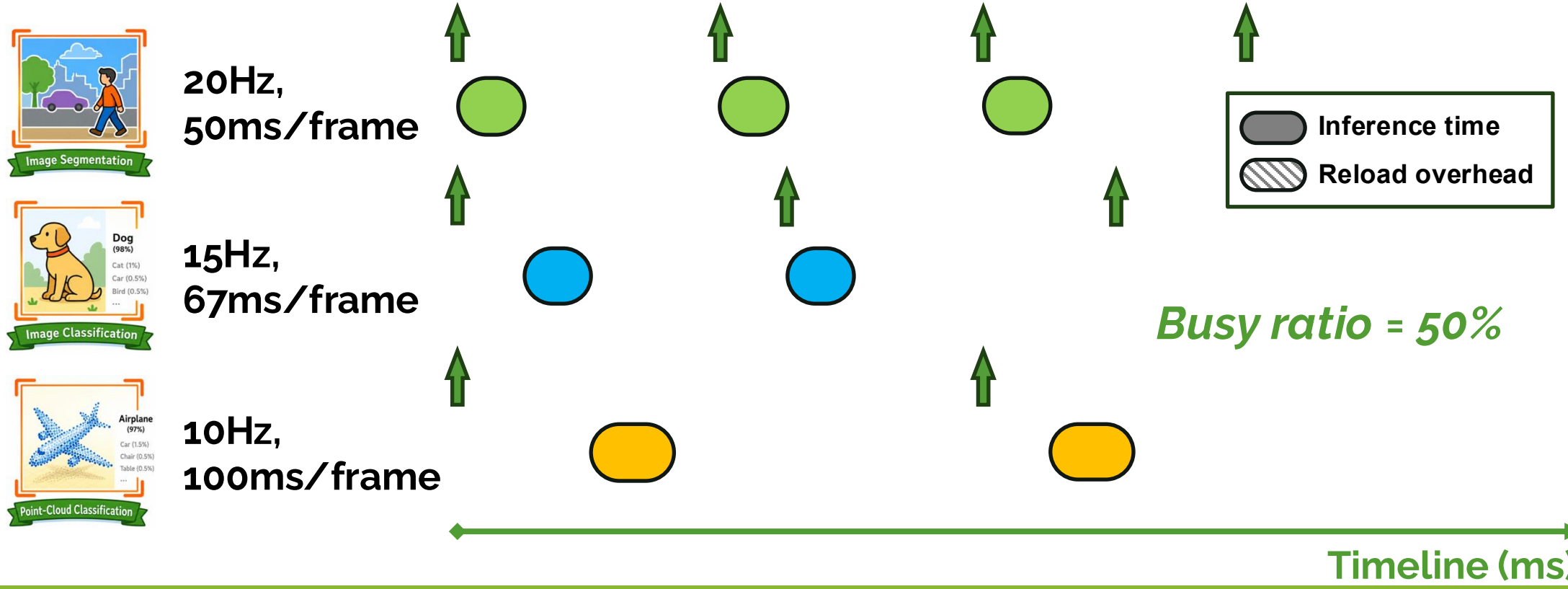
Trade-off with advanced overlay architecture

- **Advanced overlay** acc can be proposed in the future.



Trade-off with advanced overlay architecture

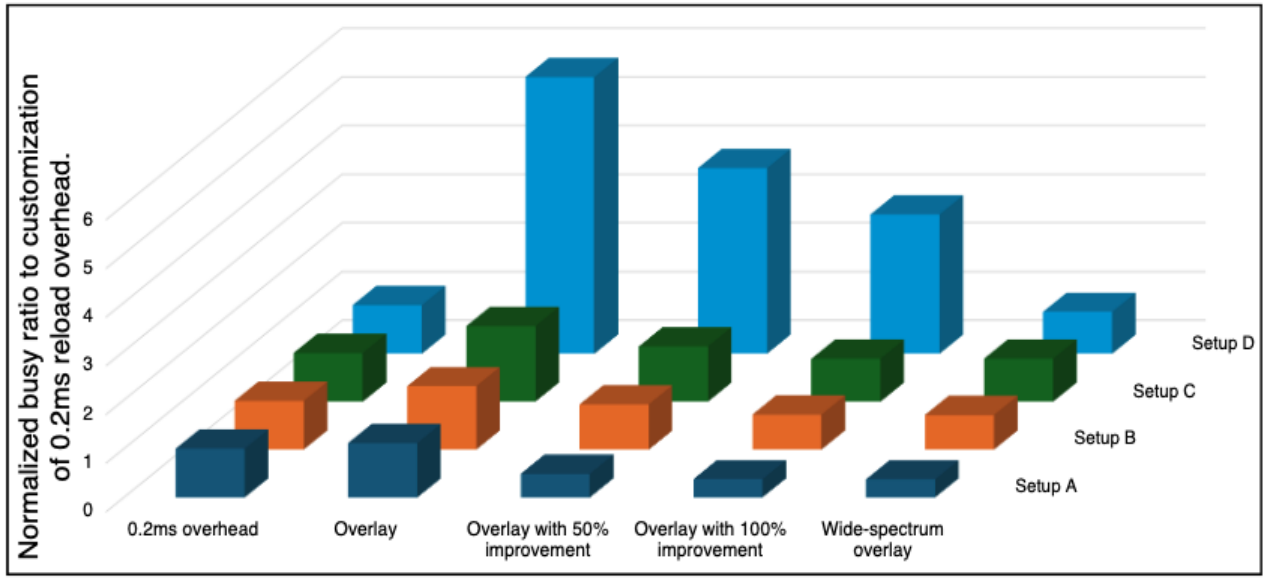
- **Advanced overlay acc can be proposed in the future.**



Trade-off with advanced overlay architecture

Trade-off with advanced overlay architecture

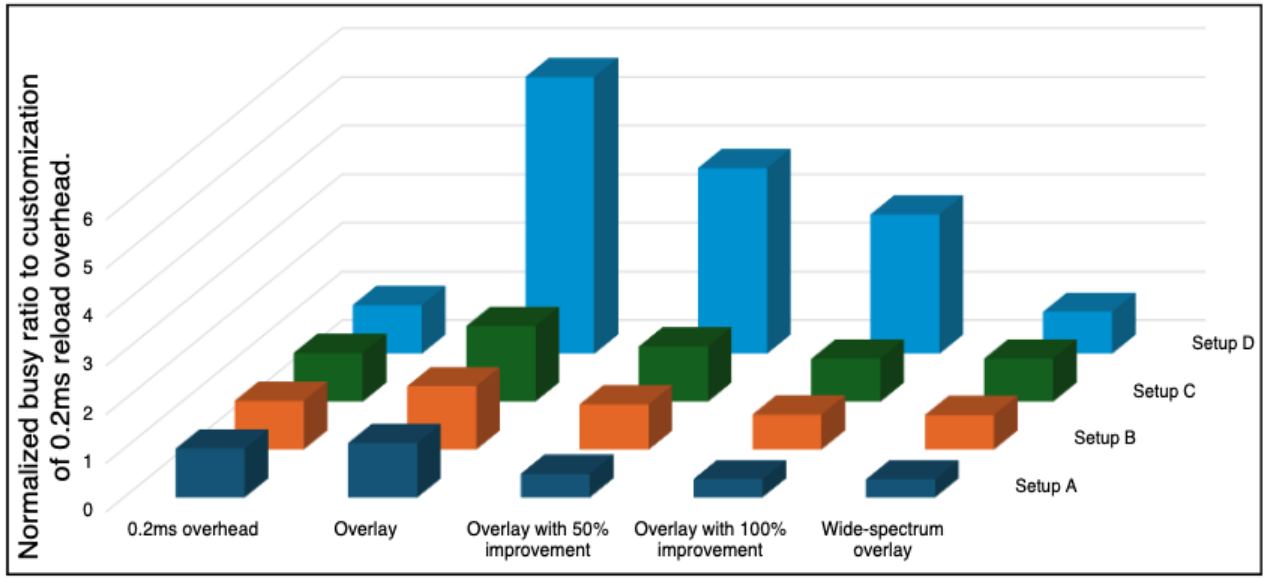
- Experiments on further advanced overlay assumption.



Comparison under advanced overlay accelerators.

Trade-off with advanced overlay architecture

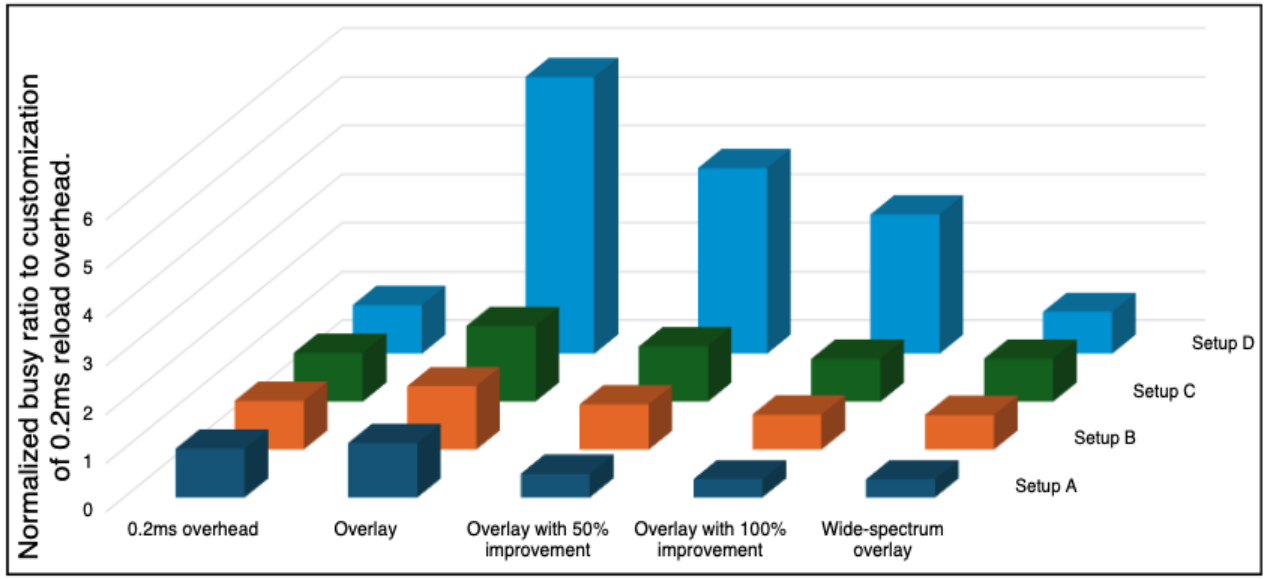
- Experiments on further advanced overlay assumption.
- Current overlay acc struggles on diverse workload scenarios.



Comparison under advanced overlay accelerators.

Trade-off with advanced overlay architecture

- Experiments on further advanced overlay assumption.
- Current overlay acc struggles on diverse workload scenarios.
- Wide-spectrum overlay that covers entire range of workload diversity is necessary.



Comparison under advanced overlay accelerators.

Key Takeaways

Key Takeaways

- Optimal architecture for one task is not necessarily the best for the whole ADS system.

Key Takeaways

- Optimal architecture for one task is not necessarily the best for the whole ADS system.
- Reconfiguration overhead is a significant factor. Even a few milliseconds' reduction affects a lot.

Key Takeaways

- Optimal architecture for one task is not necessarily the best for the whole ADS system.
- Reconfiguration overhead is a significant factor. Even a few milliseconds' reduction affects a lot.
- Customized architectures become attractive when bitstream reload overhead is sufficiently reduced.

Key Takeaways

- Optimal architecture for one task is not necessarily the best for the whole ADS system.
- Reconfiguration overhead is a significant factor. Even a few milliseconds' reduction affects a lot.
- Customized architectures become attractive when bitstream reload overhead is sufficiently reduced.
- Advanced overlays can reverse the trend again.

Thank You & Welcome to questions!

To Overlay or to Customize?
Revisiting Architectural Choices in Heterogeneous Systems

Xingzhen Chen, Shixin Ji, Zheng Dong*, and Peipei Zhou

Brown University; Wayne State University*

xingzhen_chen@brown.edu
peipei_zhou@brown.edu

https://peipeizhou-eecs.github.io/