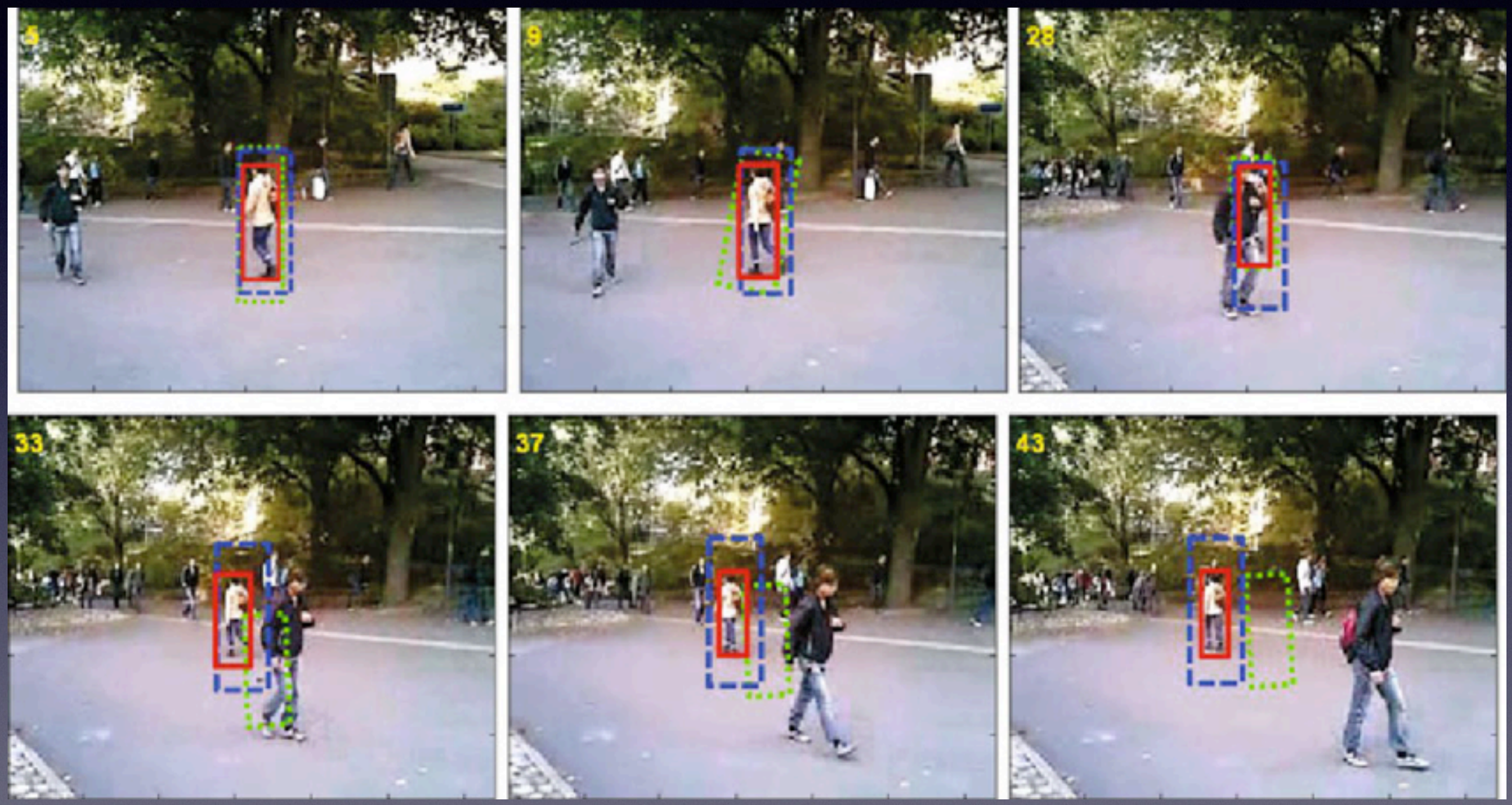
Low-Latency Proactive Continuous Vision

Yiming Gan
Department of Computer Science,
University of Rochester

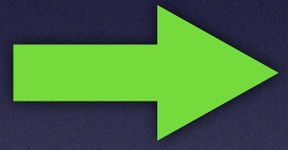
with
Yuxian Qiu, Shanghai Jiao Tong University
Lele Chen, University of Rochester
Jingwen Leng, Shanghai Jiao Tong University
Yuhao Zhu University of Rochester

Continuous Vision: Long Frame Latency



Bottleneck: Serialization

Sensor

Light  Raw Pixels

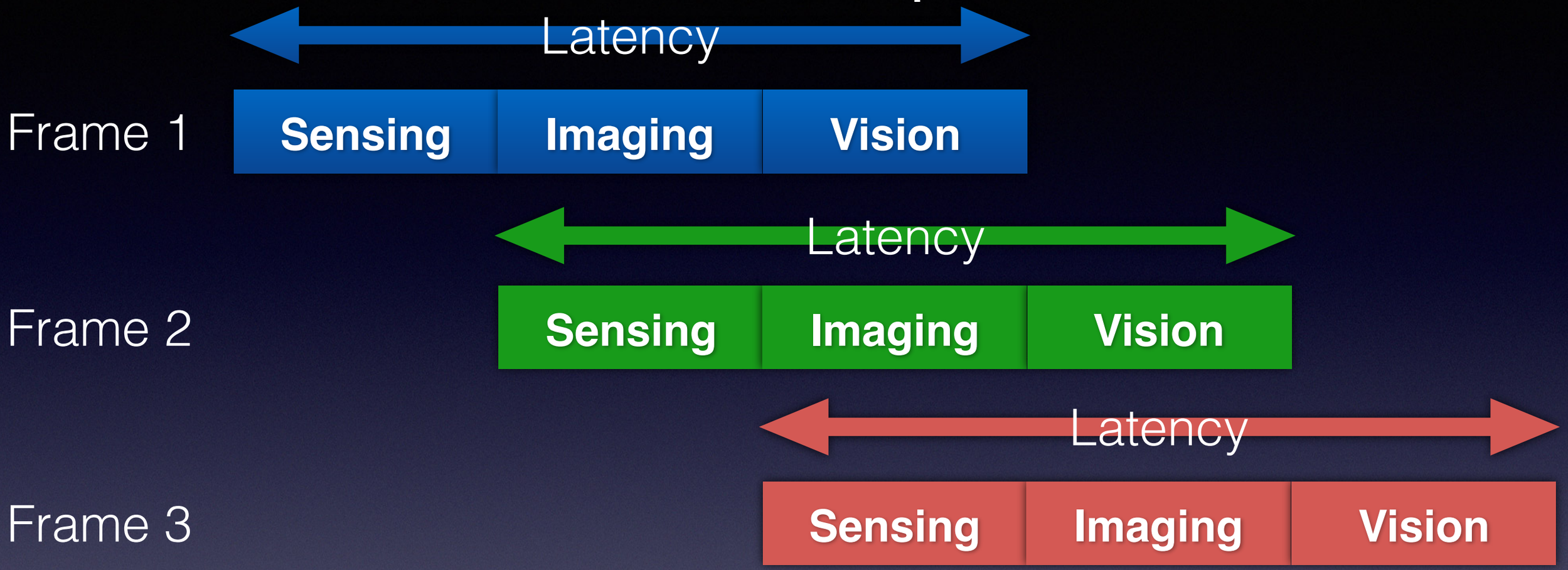
Bottleneck: Serialization

Sensor **Image Signal Processor**

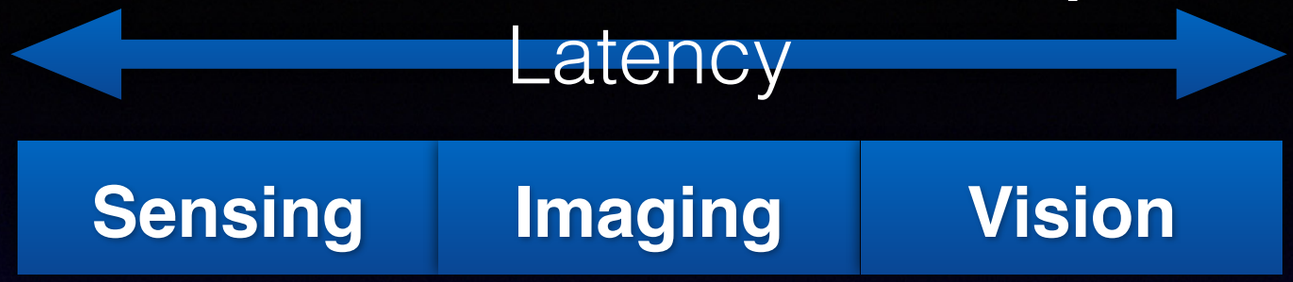
Light → Raw Pixels → RGB

Bottleneck: Serialization

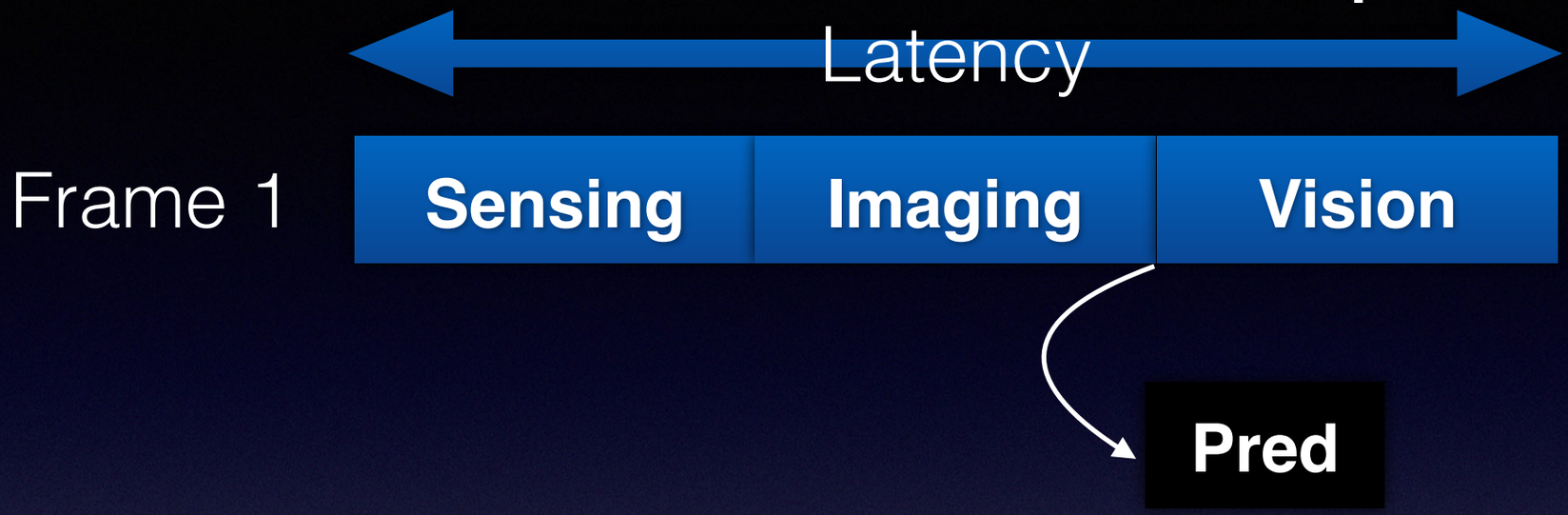
Traditional Pipeline



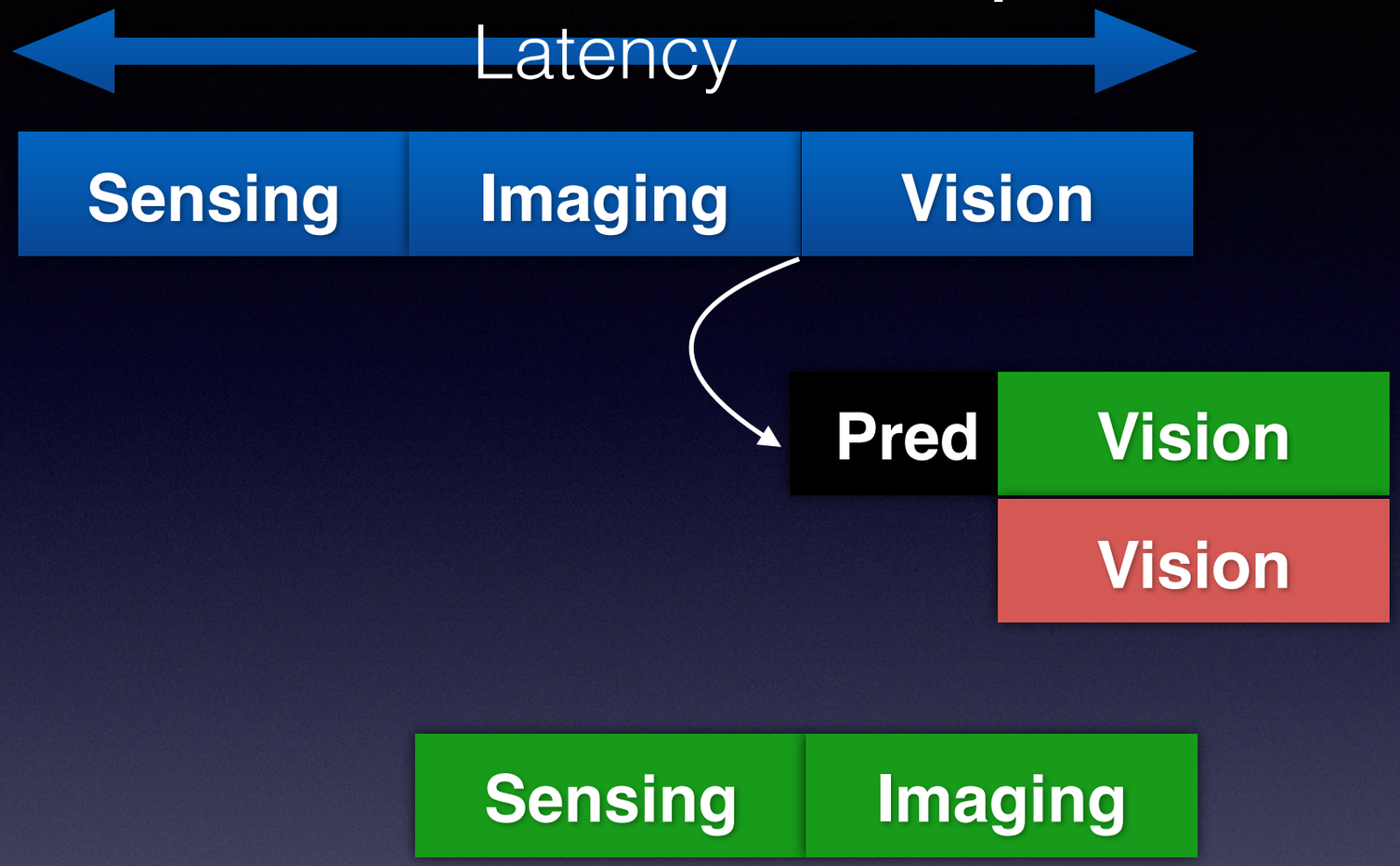
Proactive Pipeline



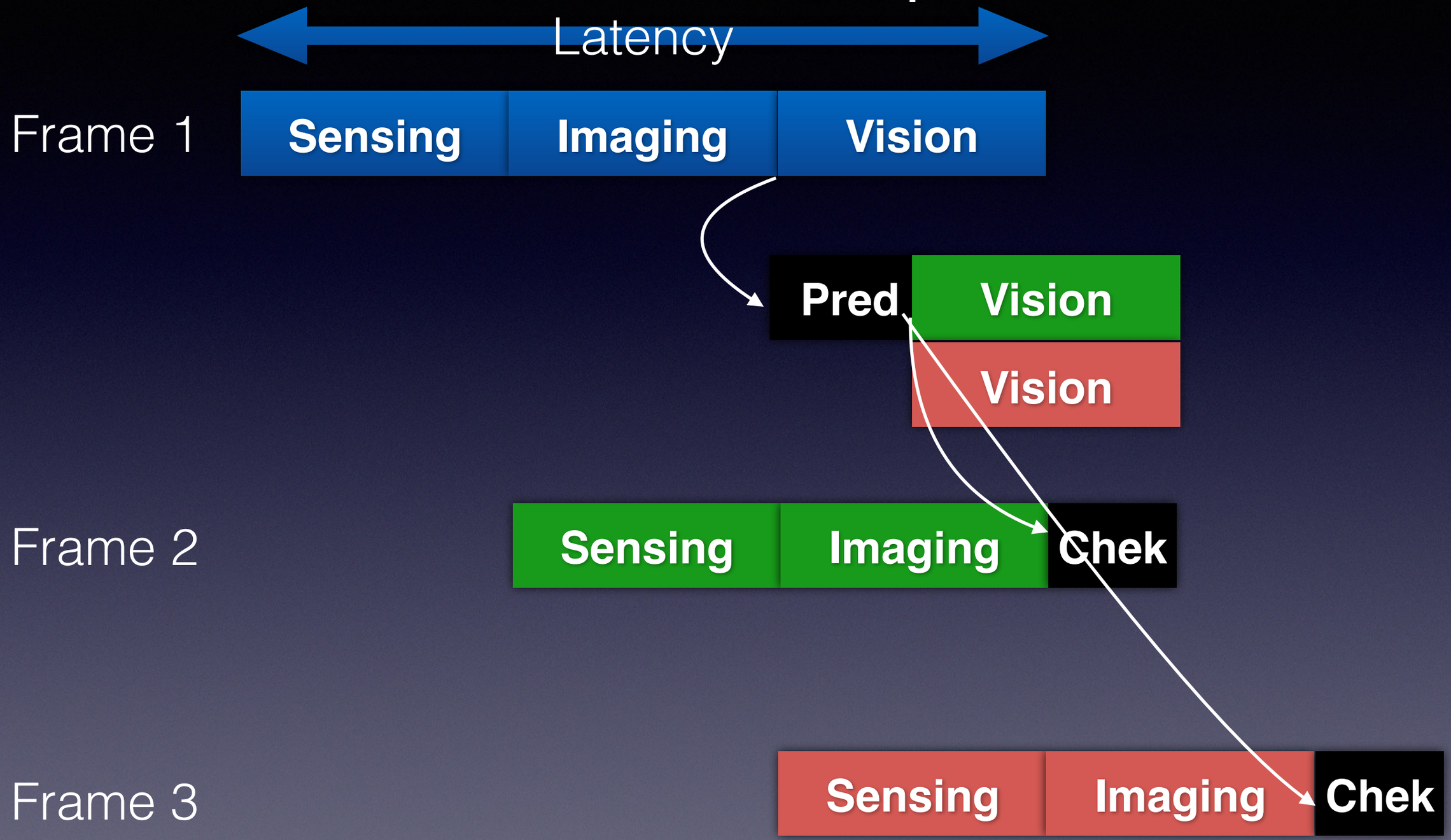
Proactive Pipeline



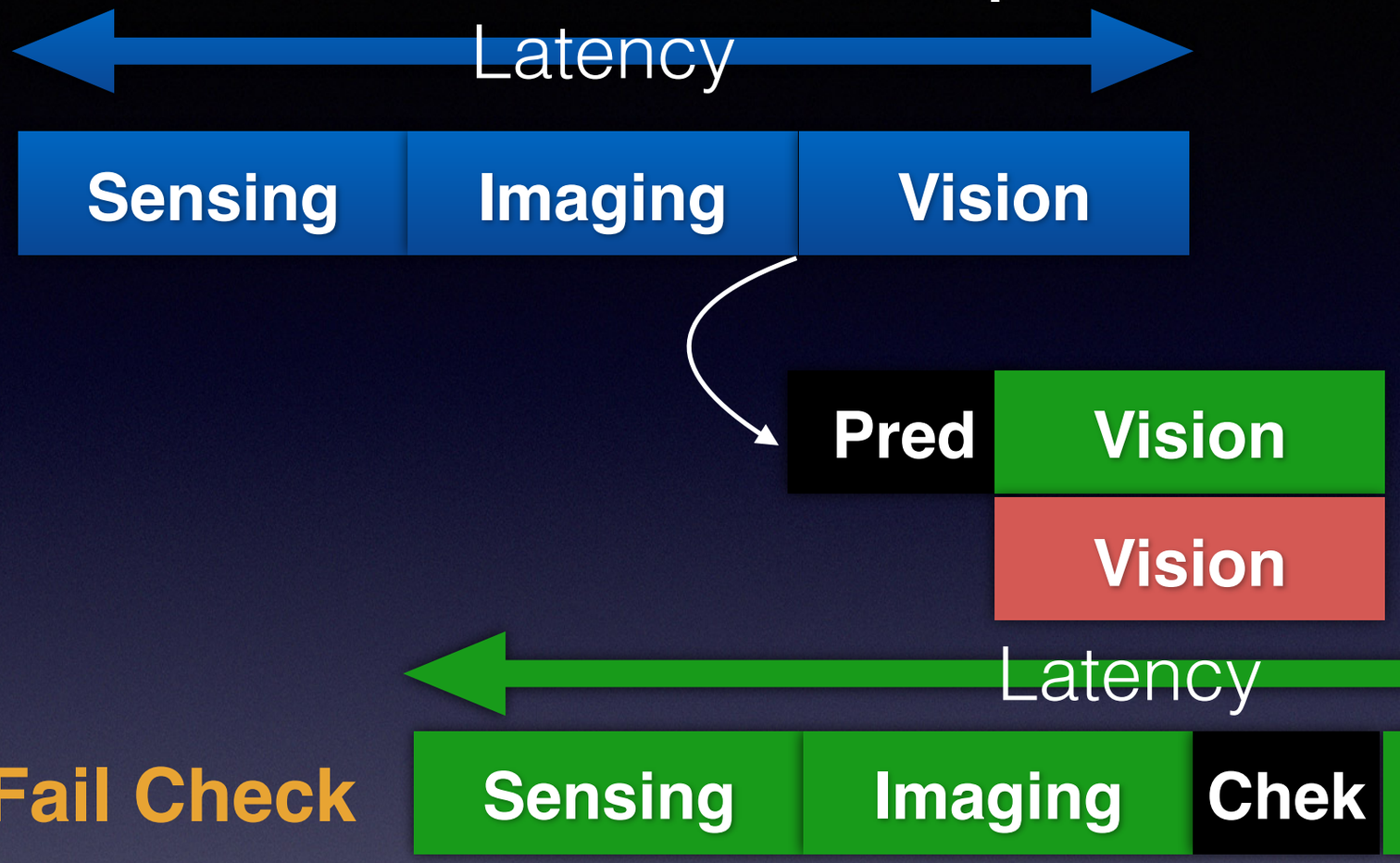
Proactive Pipeline



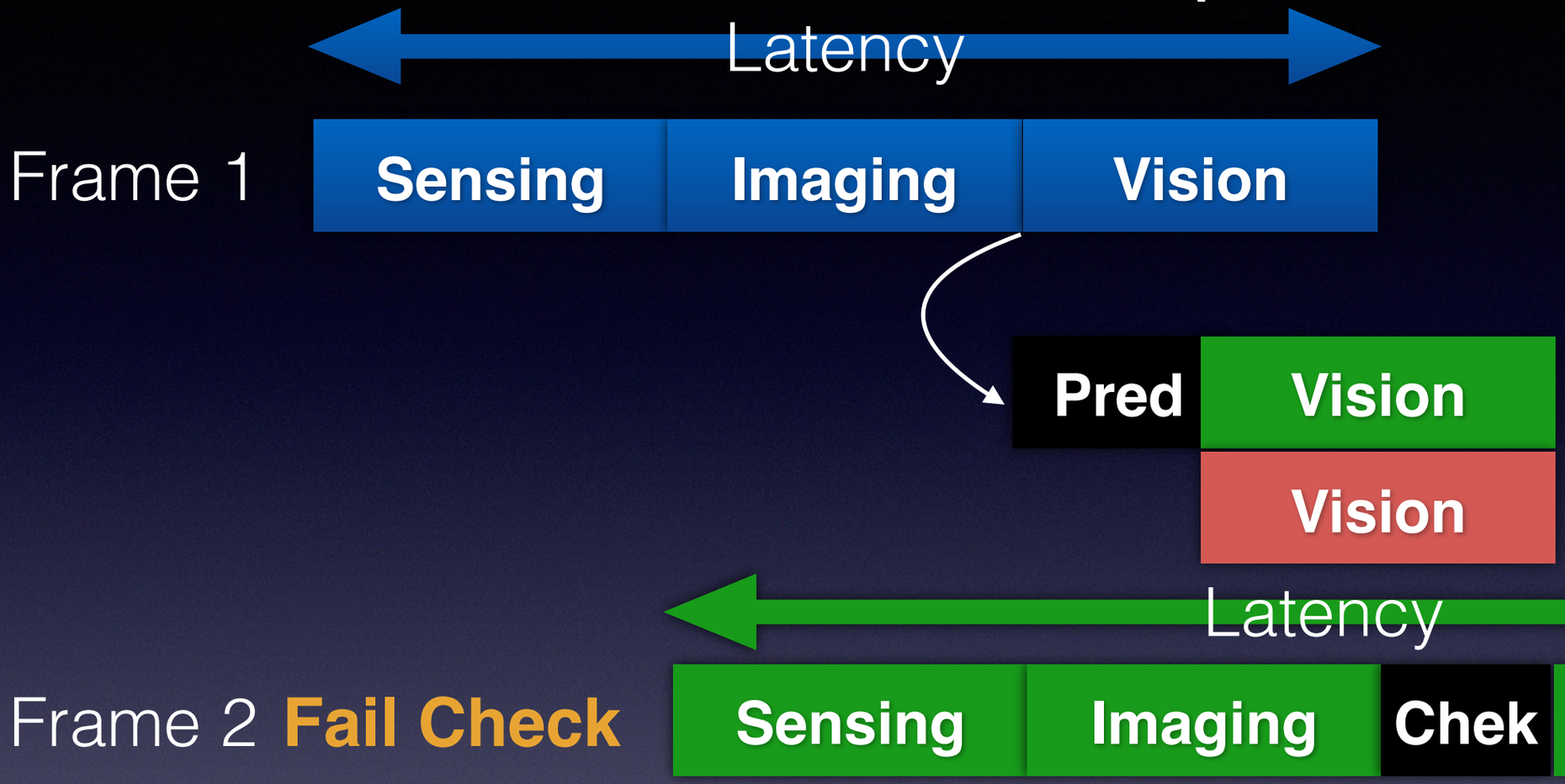
Proactive Pipeline



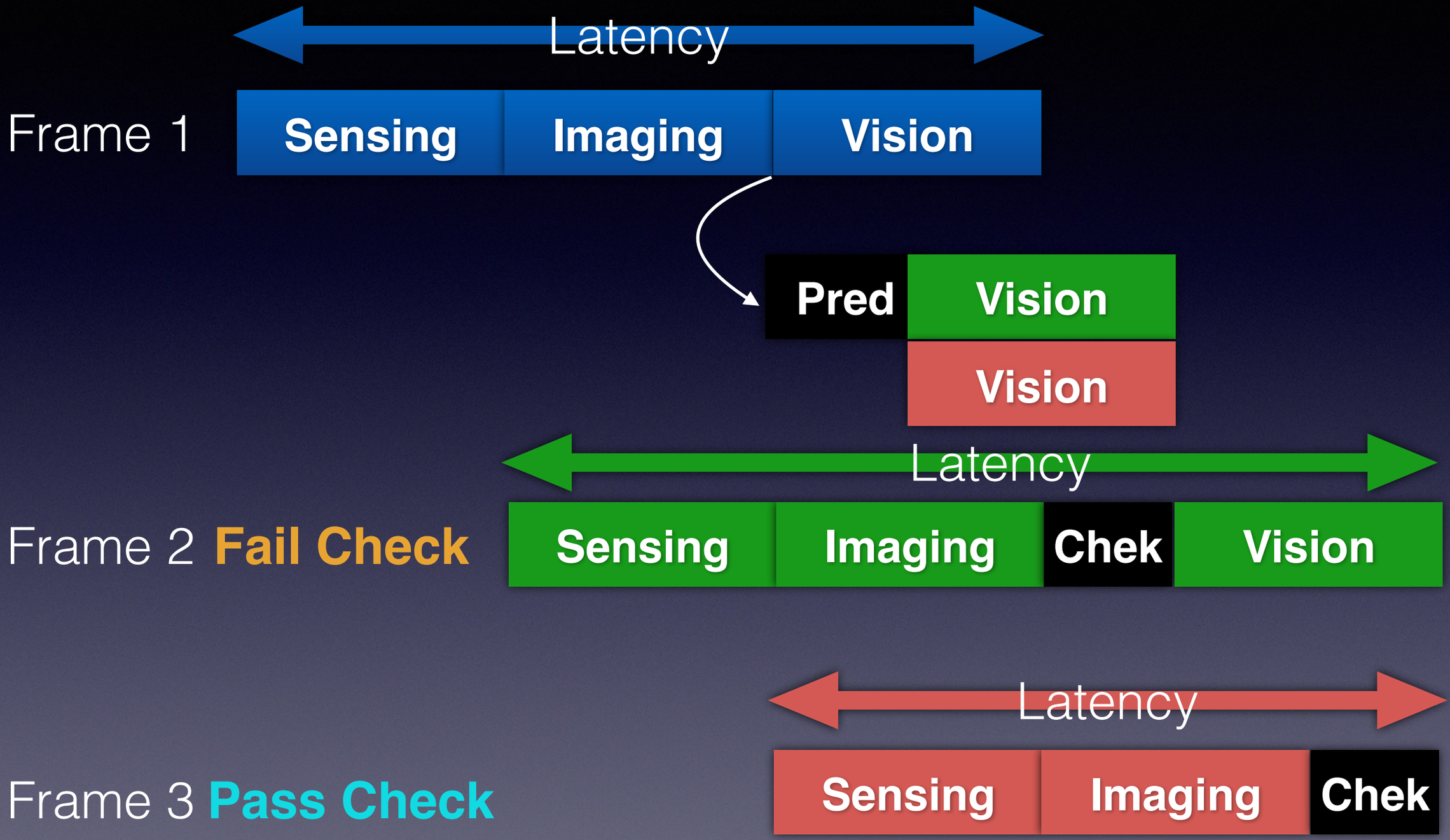
Proactive Pipeline



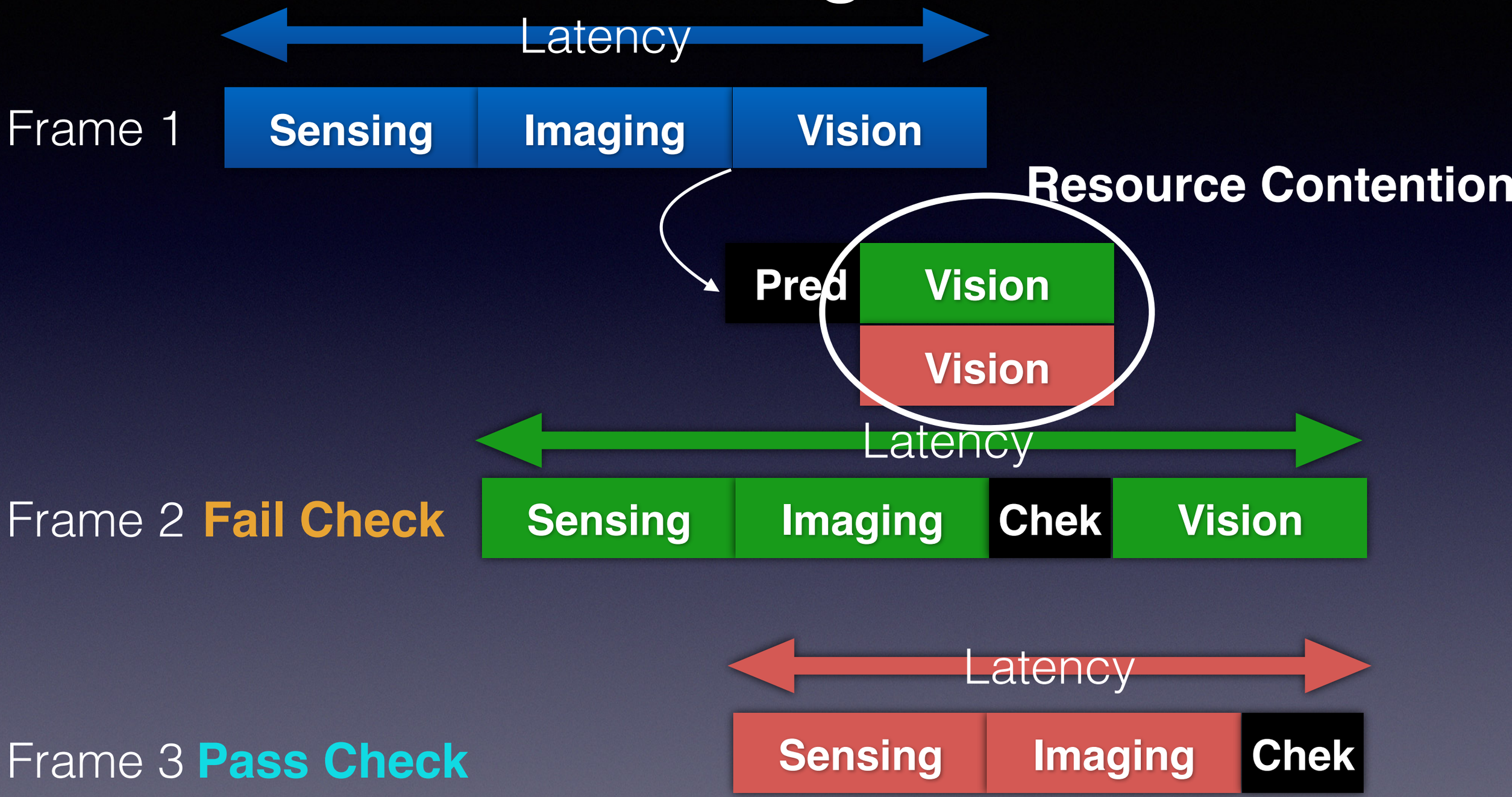
Proactive Pipeline



Gains


Latency
Frame 1
Sensing
Imaging
Vision
Pred
Vision
Vision
Latency
Frame 2 Fail Check
Sensing
Imaging
Chek
Vision
Latency
Frame 3 Pass Check
Sensing
Imaging
Chek

Challenges


Latency
Frame 1
Sensing
Imaging
Vision
Resource Contention
Pred
Vision
Vision
Latency
Frame 2 Fail Check
Sensing
Imaging
Chek
Vision
Latency
Frame 3 Pass Check
Sensing
Imaging
Chek

Solutions

Snapdragon 675

CPU
4th gen Kryo CPU
Performance @ 2.0GHz
Efficiency @ 1.7GHz

Artificial Intelligence
3rd gen Qualcomm® AI Engine

Modem
Snapdragon X12
Cat 12 DL, up to 600 Mbps

Audio
Qualcomm Aqstic™ audio
Qualcomm® aptX™ audio

Display
Up to FHD+ display

GPU
6th gen Adreno GPU

DSP
6th gen Hexagon DSP
DSP Security

Camera
2nd gen Qualcomm Spectra ISP
Up to 25 Megapixels @ 30fps ZSL
48 Megapixels snapshot

Charging
Qualcomm® Quick
Charge™ 4+ technology

Commercial devices expected in Q1 2019

Solutions

Snapdragon 675

CPU
4th gen Kryo CPU
Performance @ 2.0GHz
Efficiency @ 1.7GHz

Artificial Intelligence
3rd gen Qualcomm® AI Engine

Modem
Snapdragon X12
Cat 12 DL, up to 600 Mbps

Audio
Qualcomm Aqstic™ audio
Qualcomm® aptX™ audio

Display
Up to FHD+ display

GPU
6th gen Adreno GPU

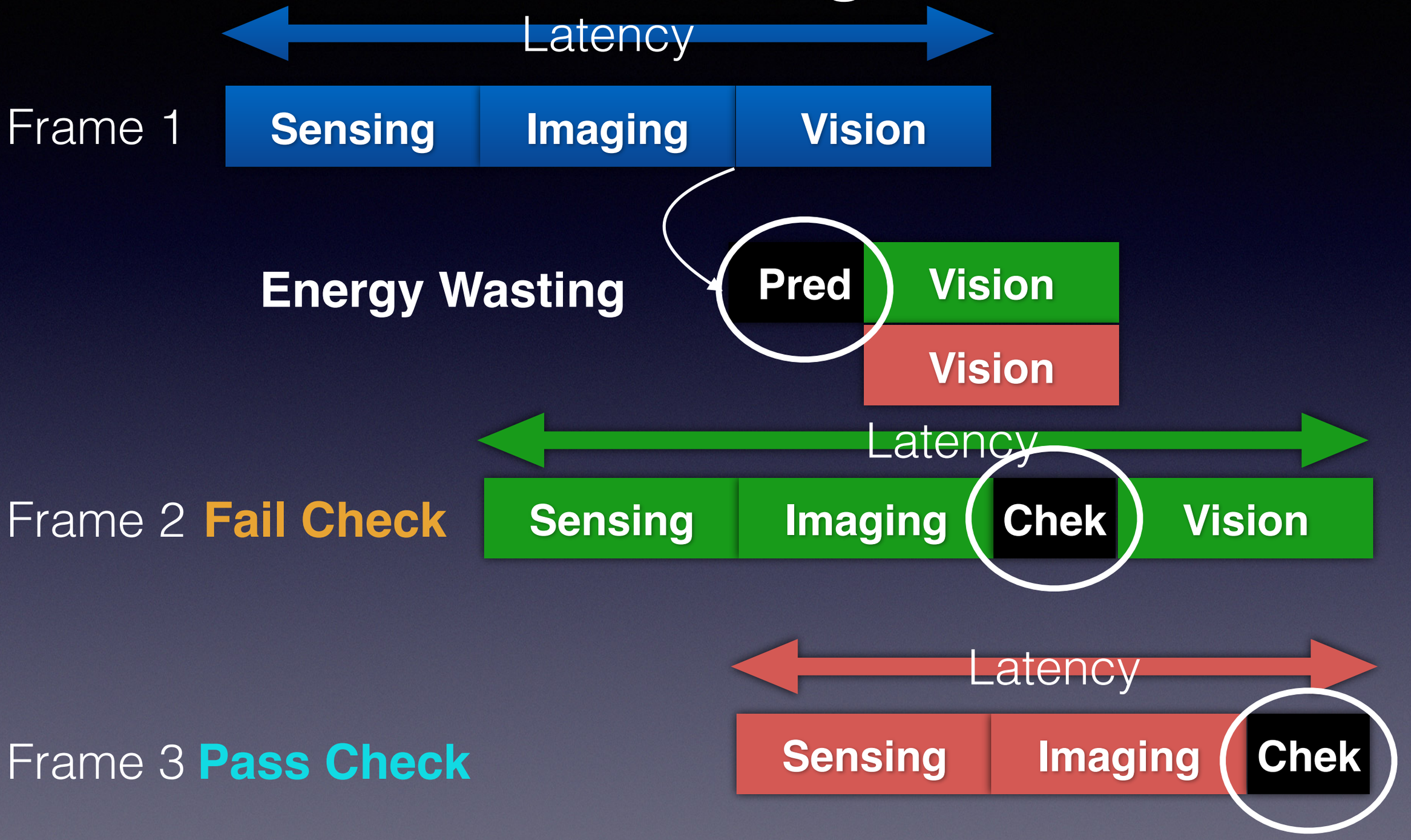
DSP
6th gen Hexagon DSP
DSP Security

Camera
2nd gen Qualcomm Spectra ISP
Up to 25 Megapixels @ 30fps ZSL
48 Megapixels snapshot

Charging
Qualcomm® Quick
Charge™ 4+ technology

Commercial devices expected in Q1 2019

Challenges


Latency
Frame 1
Sensing
Imaging
Vision
Energy Wasting
Pred
Vision
Vision
Latency
Frame 2 Fail Check
Sensing
Imaging
Chek
Vision
Latency
Frame 3 Pass Check
Sensing
Imaging
Chek

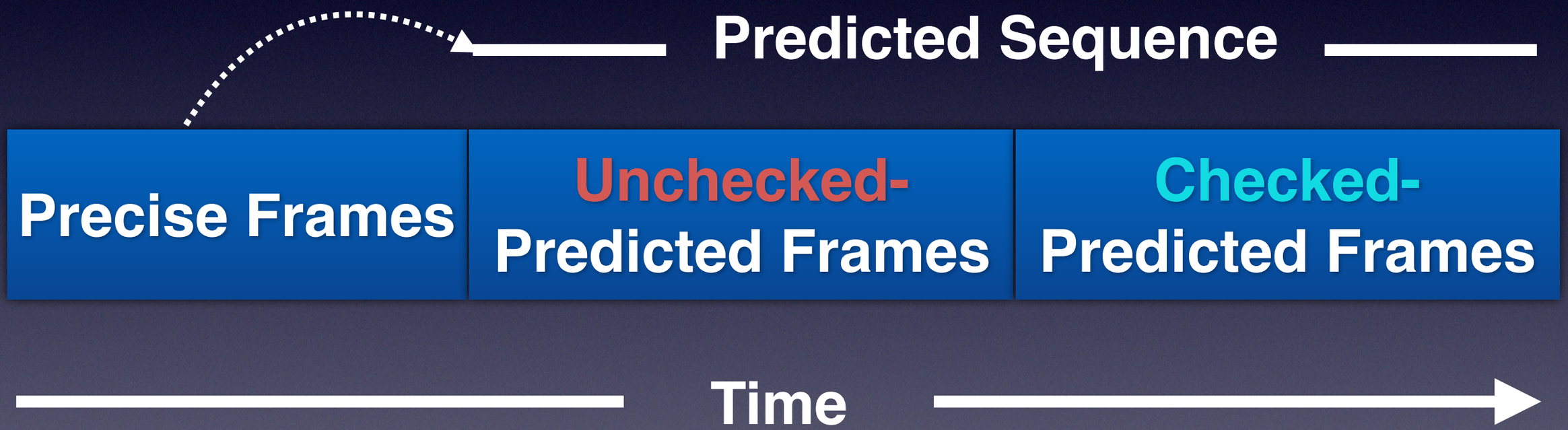
Solutions

- Relaxing Checking Criterion (Threshold T)

Solutions

- Relaxing Checking Criterion (Threshold T)
- Relaxing Checking Frequency (Degree K)

Frames Sequence



PVF Framework

Static

Dynamic

Vision Apps

SoC

CPU

NPU

DSP

GPU

ISP

BUS

Memory

Sensor

PVF Framework

Static

Dynamic

Vision Apps

Similarity T

Degree K

SoC

CPU

NPU

DSP

GPU

ISP

BUS

Memory

Sensor

“ → ”

Accuracy Target
Similarity Metric
etc.

PVF Framework

Static

Dynamic

Vision Apps

Similarity T

Degree K

SoC

CPU

NPU

DSP

GPU

ISP

Predictor

BUS

Memory

Sensor

PVF Framework

Static

Dynamic

Vision Apps

Similarity T

Degree K

SoC

Runtime

CPU

NPU

DSP

GPU

ISP

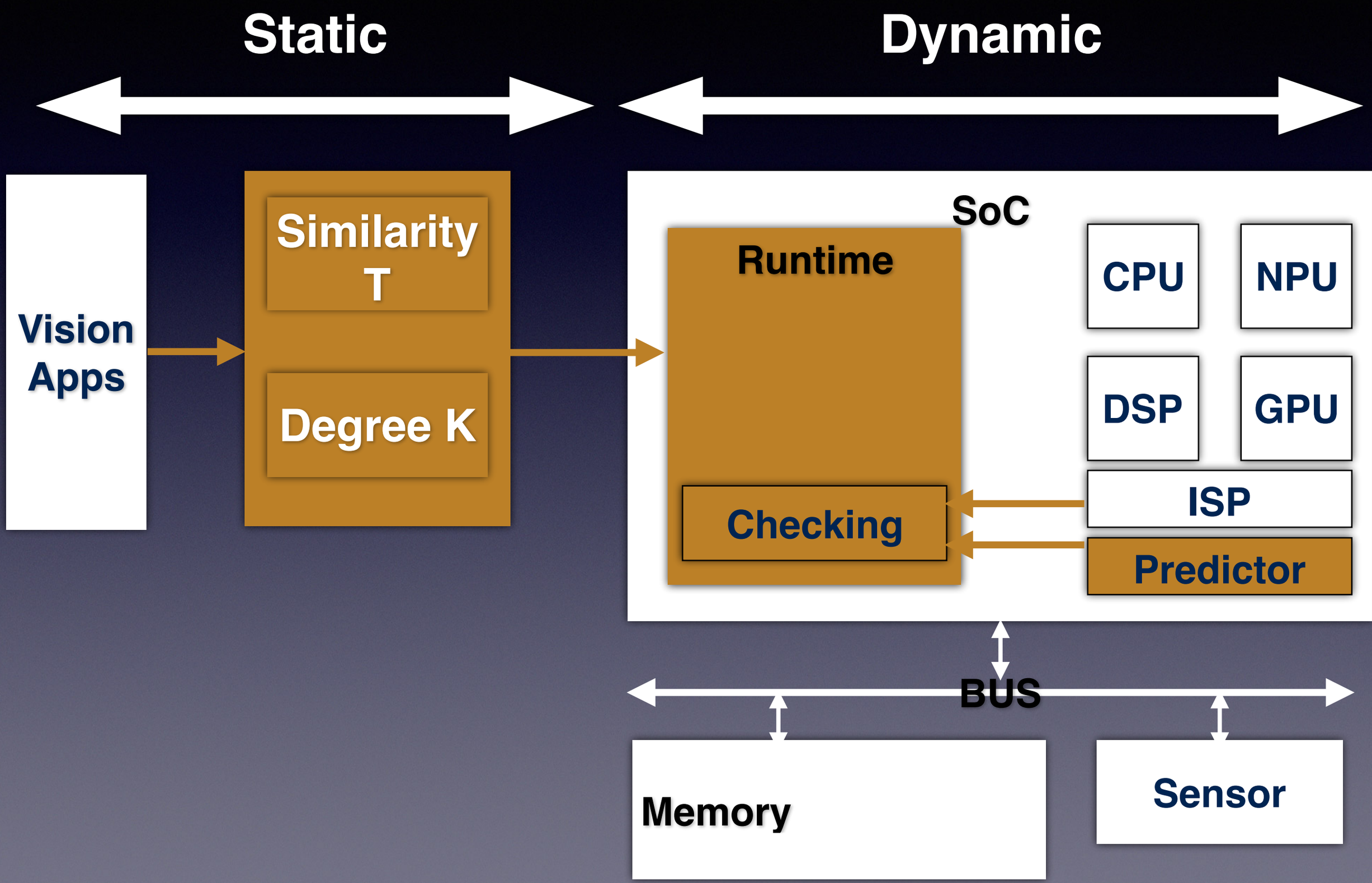
Predictor

BUS

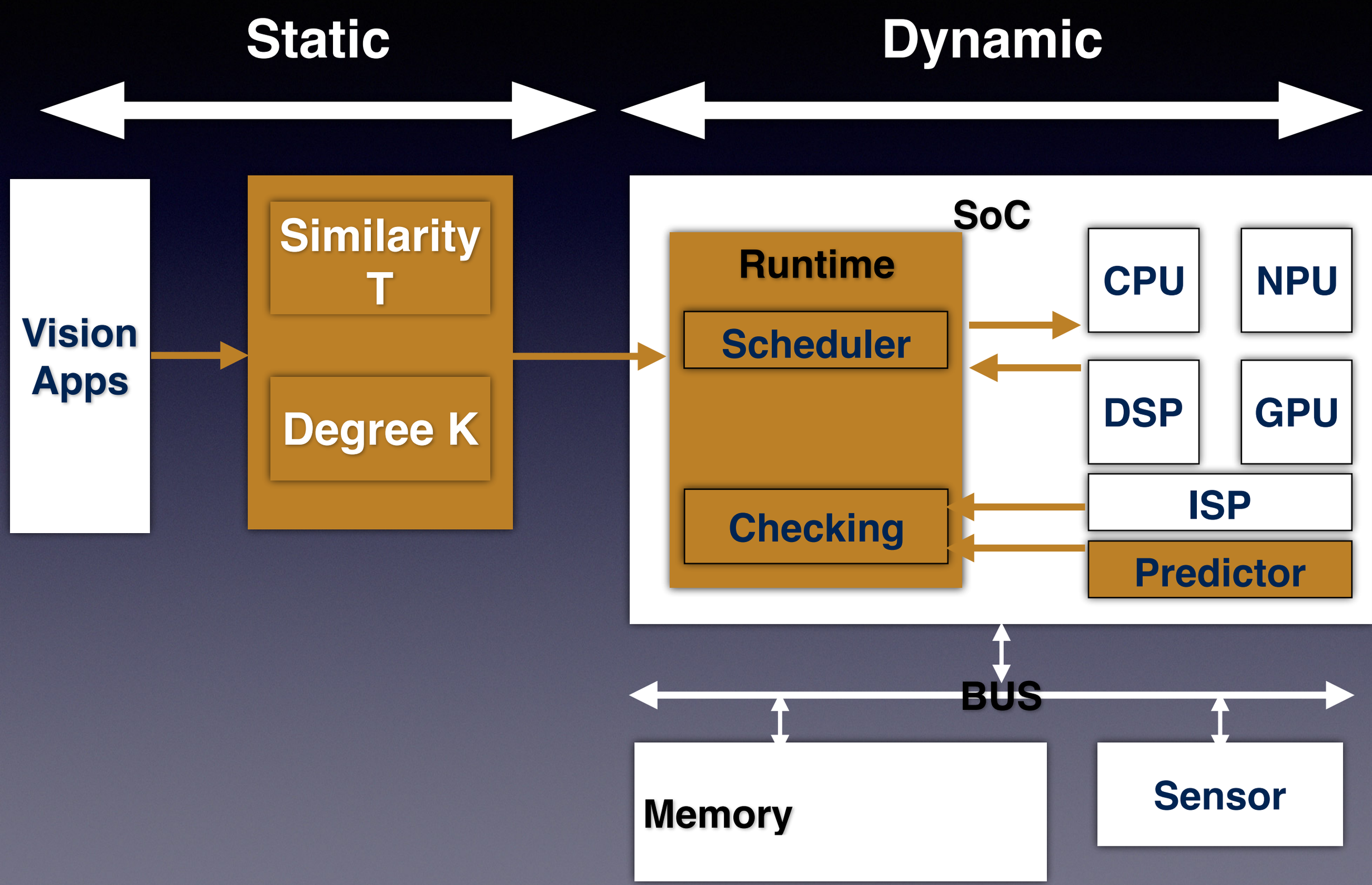
Memory

Sensor

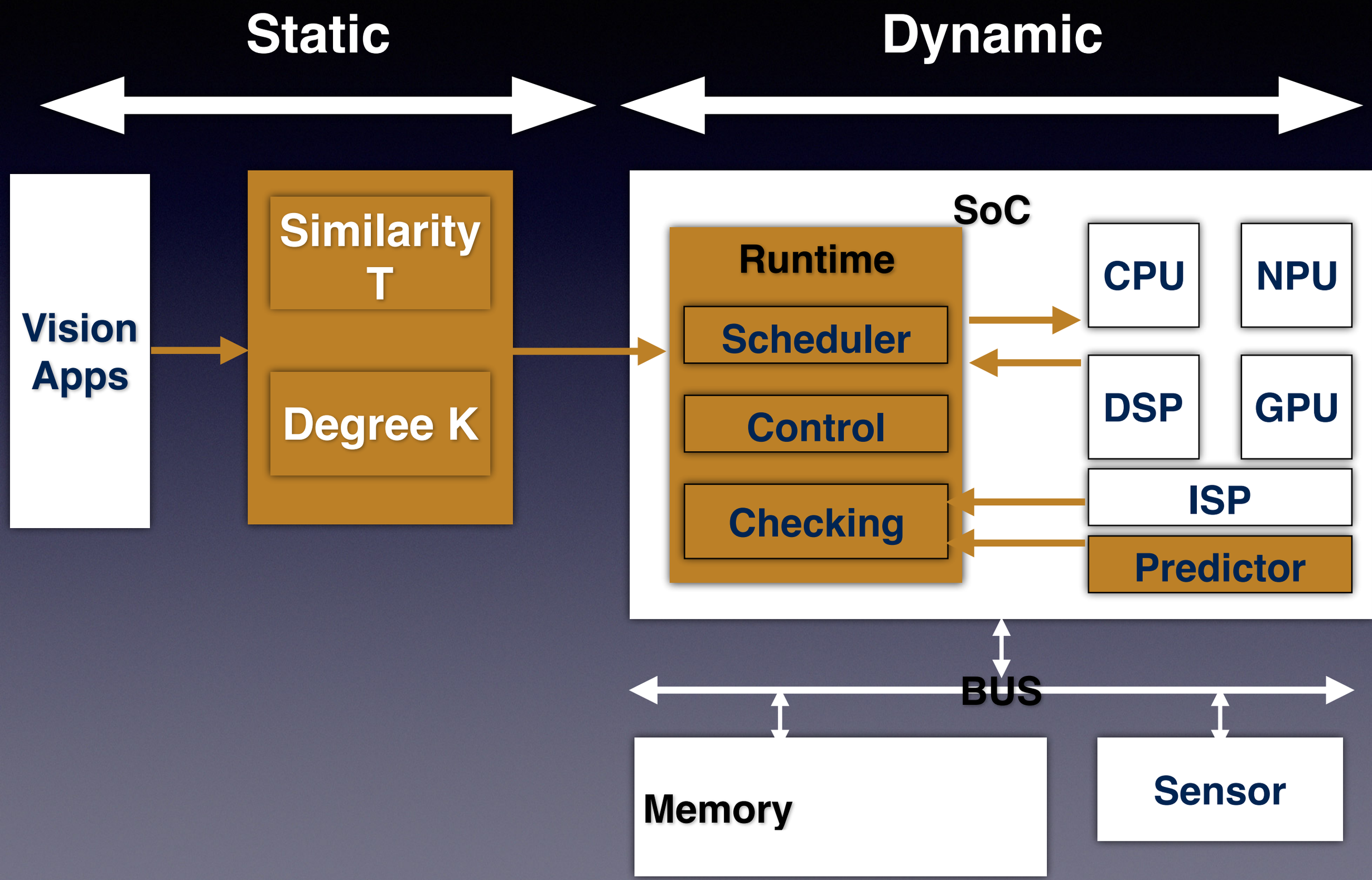
PVF Framework



PVF Framework



PVF Framework



Experimental Setup

I. In house simulator modeling state-of-the art SoCs

- Real measurement of latency and energy on different IPs.

Experimental Setup

I. In house simulator modeling state-of-the art SoCs

- Real measurement of latency and energy on different IPs.

II. RTL Implementations for NPU and Predictor

- 20x20 Systolic Array for NPU, 10x10 Systolic Array for Predictor

Experimental Setup

I. In house simulator modeling state-of-the art SoCs

- Real measurement of latency and energy on different IPs.

II. RTL Implementations for NPU and Predictor

- 20x20 Systolic Array for NPU, 10x10 Systolic Array for Predictor

III. Evaluate on Object Detection and Tracking

- KITTI dataset for object detection, VOT-challange for tracking.

Experimental Setup

I. In house simulator modeling state-of-the art SoCs

- Real measurement of latency and energy on different IPs.

II. RTL Implementations for NPU and Predictor

- 20x20 Systolic Array for NPU, 10x10 Systolic Array for Predictor

III. Evaluate on Object Detection and Tracking

- KITTI dataset for object detection, VOT-challange for tracking.

IV. Different Input Resolutions

Baselines

I. Base

- Baseline with traditional execution pipeline

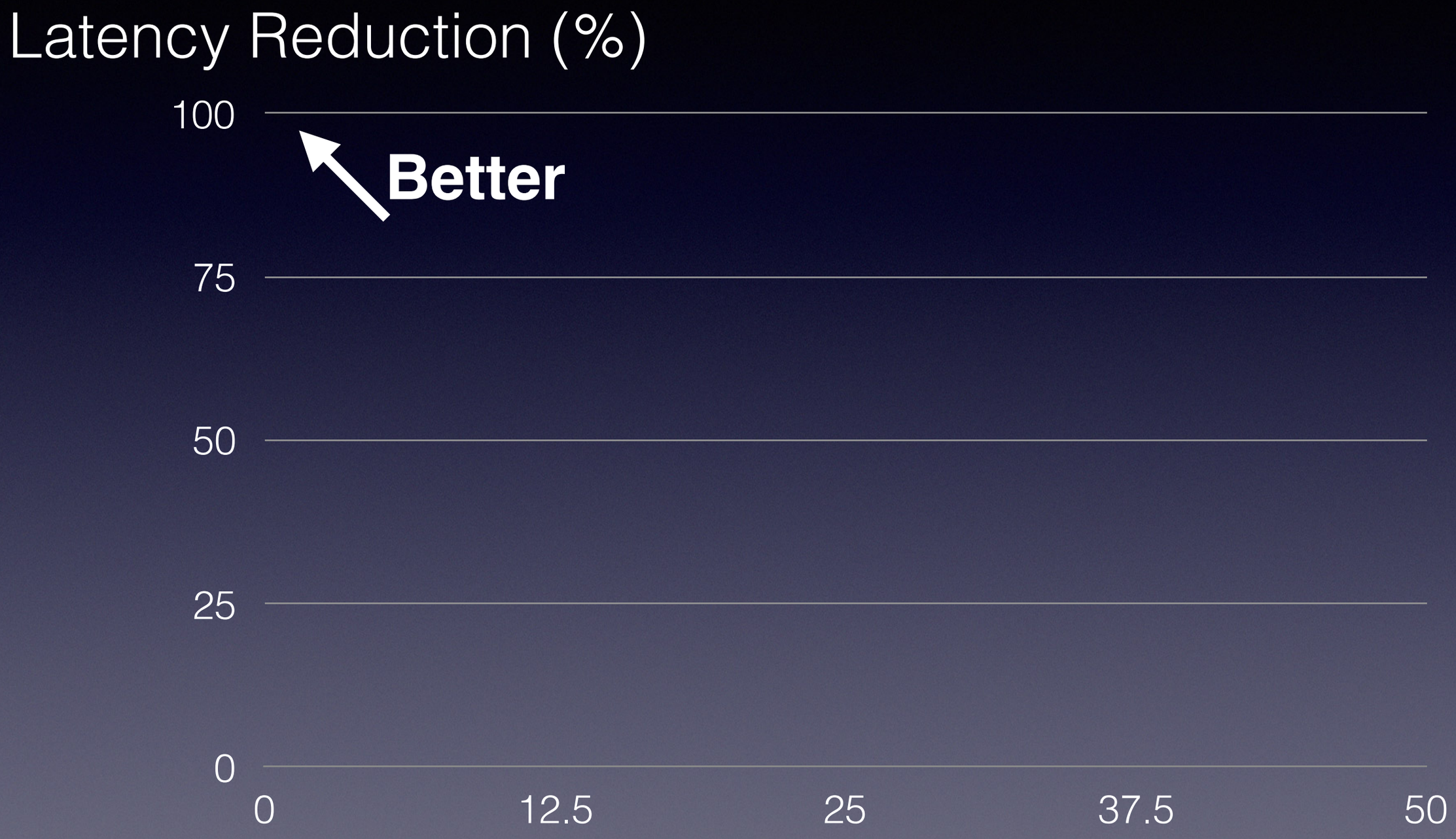
II. BO

- Baseline with optimized back-end

III. FCFS

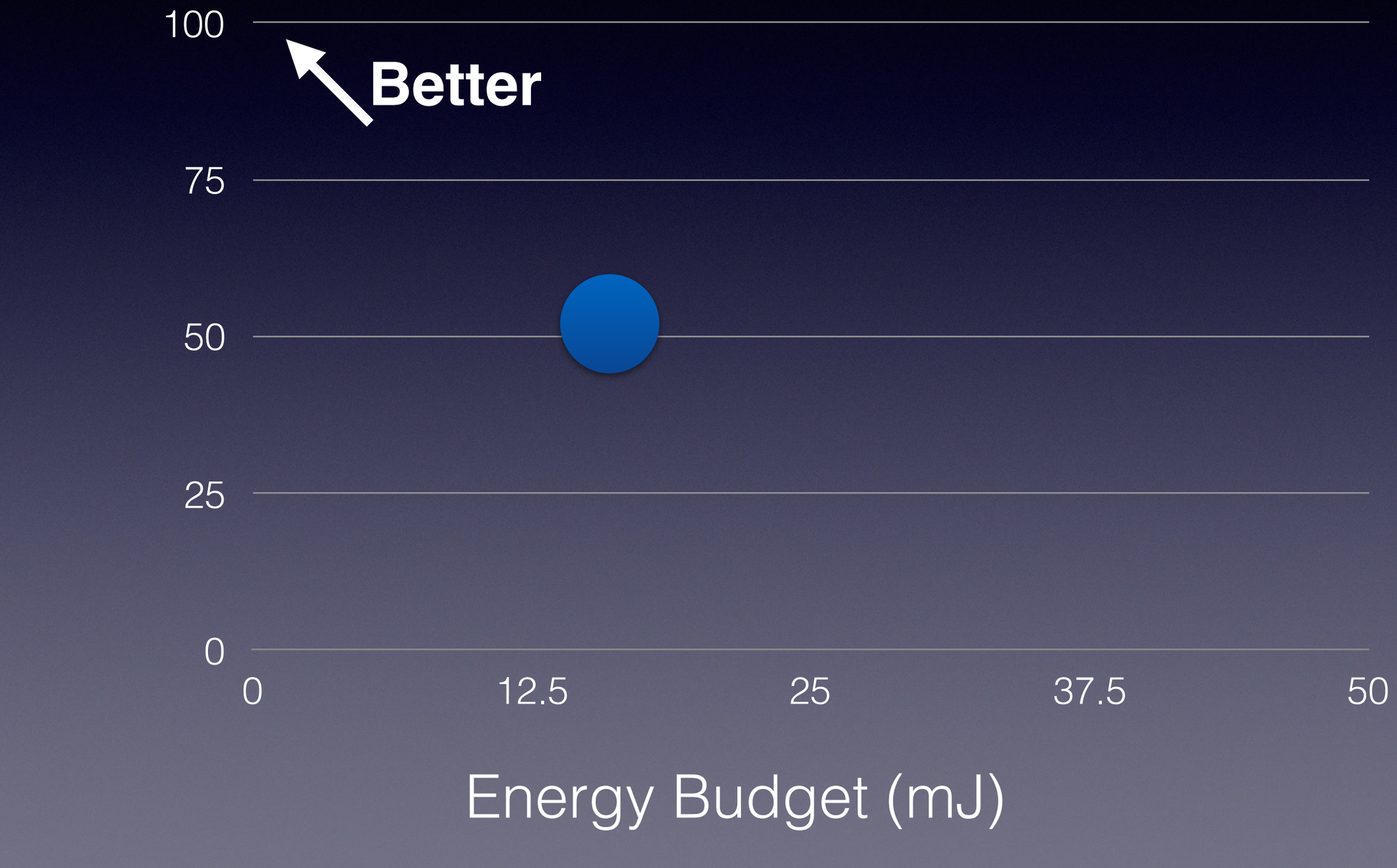
- Traditional pipeline with multiple hardware IPs

Results

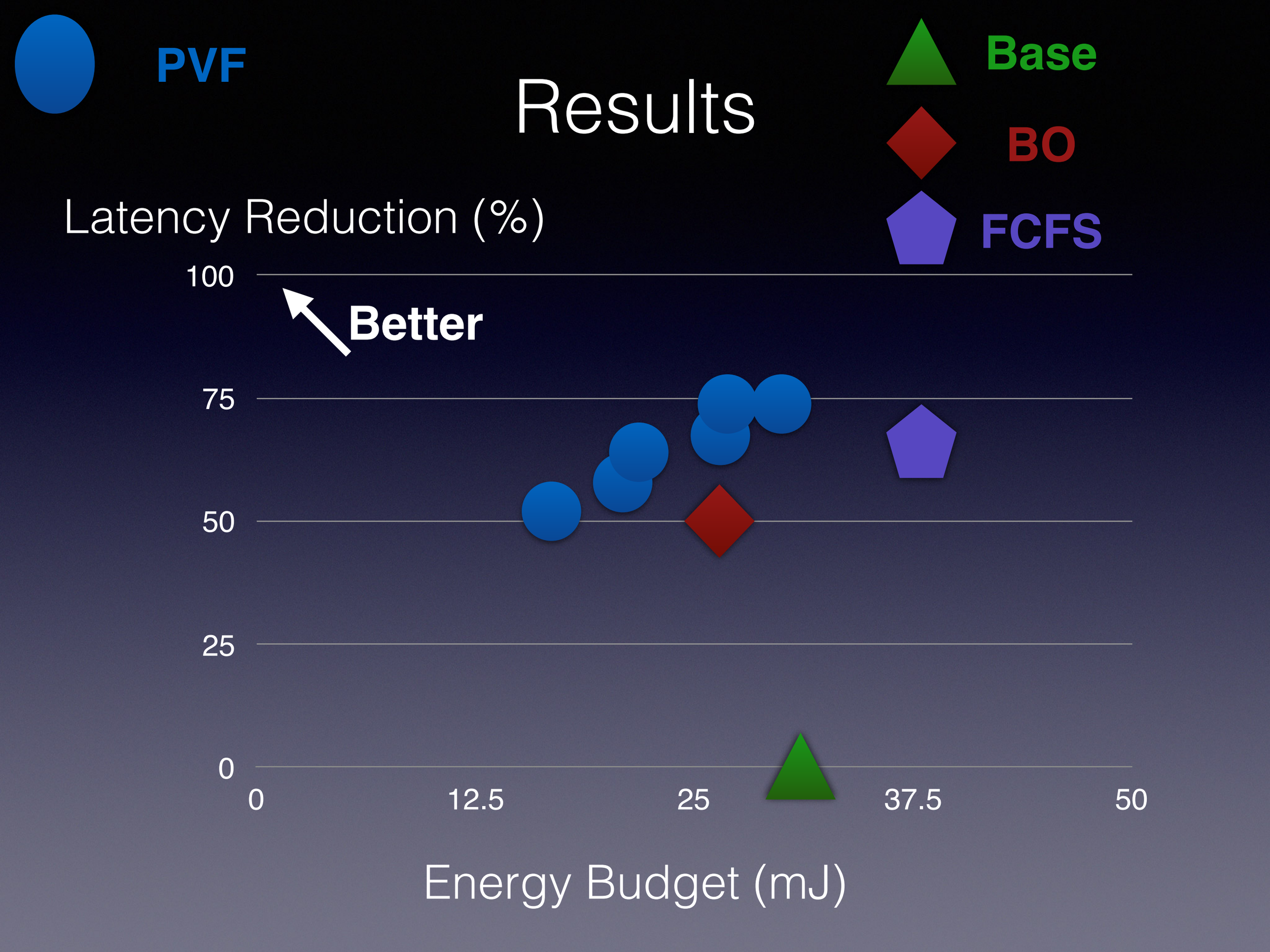

Latency Reduction (%)
100
75
Better
50
25
0
0
12.5
25
37.5
50
Energy Budget (mJ)

PVF

Results


Latency Reduction (%)
100
75
50
25
0
Better
0
12.5
25
37.5
50
Energy Budget (mJ)

Results


PVF
Base
BO
FCFS
Latency Reduction (%)
100
75
50
25
0
Better
0
12.5
25
37.5
50
Energy Budget (mJ)

Conclusion

I. Long Latency Bottleneck Continuous Vision

II. Proactive Execution Pipeline

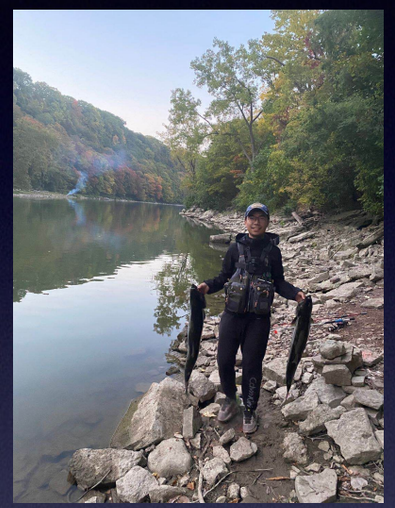
1) Leveraging Heterogeneities in Mobile SoCs
2) Relaxed Checking

III. Non-mission-critical System

Collaborators

Yuxian Qiu

Jingwen Leng



Lele Chen

Yuhao Zhu

Questions