Smart Mobile Platform

Lyapunov Optimization for Time-Varying Queueing Systems

Prof. Joongheon Kim
Korea University, School of Electrical Engineering
https://joongheon.github.io
joongheon@korea.ac.kr

Lyapunov Control for Learning-based Systems

Problem Solving for **Utility Maximization**

- Linear Programming
- Convex Programming
- …

No Consideration on Time

What if our optimal solution introduces delays?

- Harmful for real-time mobile systems
- **Tradeoff between utility and delay**

Step 1

Time Modeling

- Time Modeling w/ Queue
- Lyapunov Drift Formulation w/ the Queue Values

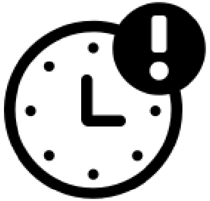
Step 2

Time-Average Optimization

- Under Queue Stability

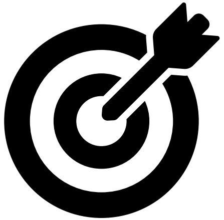
Lyapunov Control for Learning-based Systems

Delay



Tradeoff

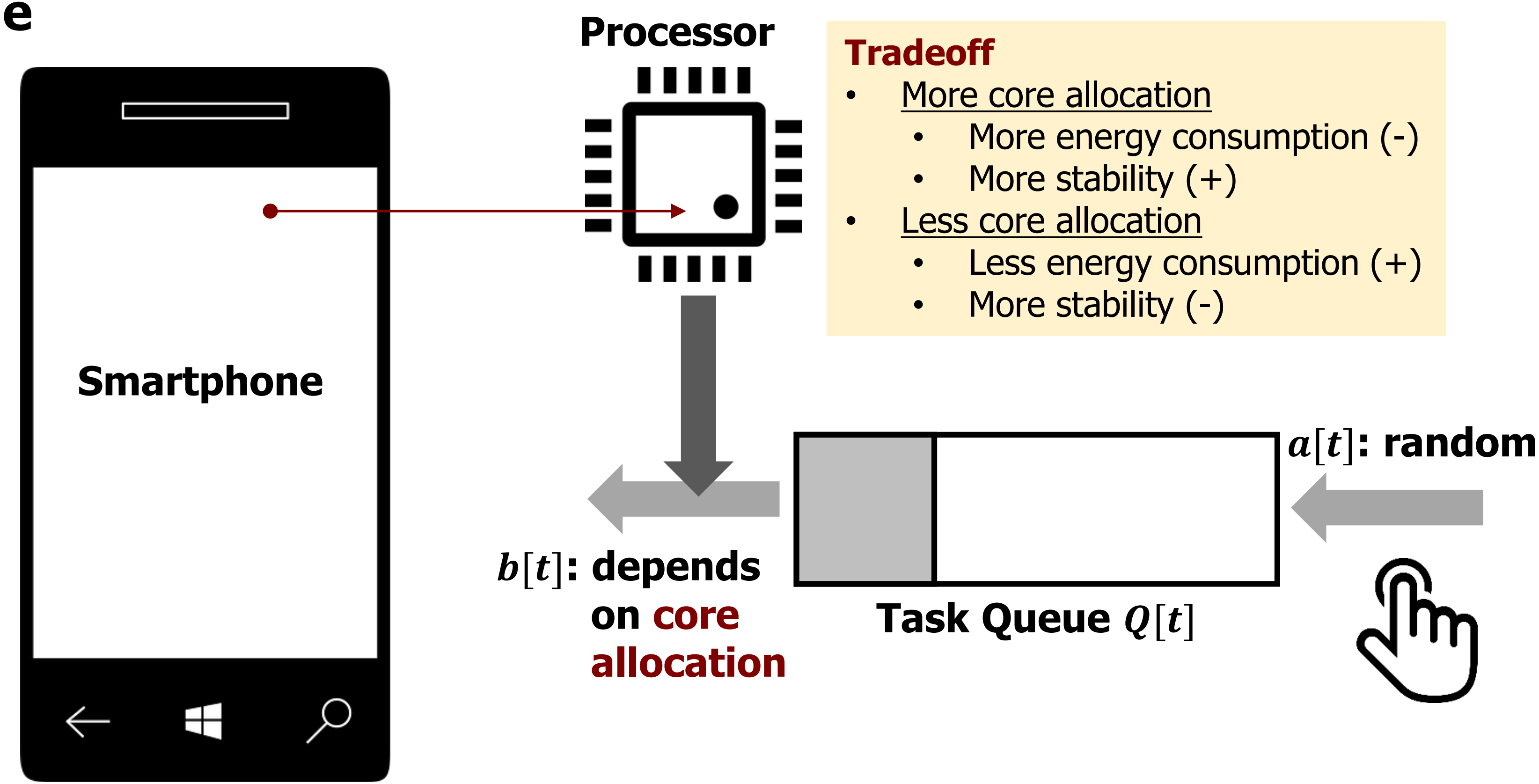
Utility



- Queue-based Model

- Energy
- Quality
- PSNR
- Accuracy
- …

- **Example**



Tradeoff
- More core allocation
  - More energy consumption (-)
  - More stability (+)
- Less core allocation
  - Less energy consumption (+)
  - More stability (-)

- **Time-Average Optimization subject to Queue Stability**

minimize: $\overline{y_0}$
subject to:
- [C1] Queue Stability

$$\overline{y_0} = \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1} y_0[\tau]$$

Lyapunov Drift

Objective Function

$$\text{Minimize: } V \cdot \boxed{y_o(\alpha[t])} + \boxed{Q[t]}\{\boxed{a(\alpha[t])} - \boxed{b(\alpha[t])}\}$$

Observation

Control

- **Time-Average Optimization subject to Queue Stability**

Observation

Control

Objective Function

$$\text{Minimize: } V \cdot \boxed{y_o(\alpha[t])} + \boxed{Q[t]}\{\boxed{a(\alpha[t])} - \boxed{b(\alpha[t])}\}$$

Objective Function (Utility)

- $\alpha[t]$: control action at time t
- $y_o(\alpha[t])$: objective w/ control action $\alpha[t]$

Stability

- $Q[t]$: queue-backlog at time t
- $a(\alpha[t])$: arrivals w/ $\alpha[t]$
- $b(\alpha[t])$: departures w/ $\alpha[t]$

- **Time-Average Optimization subject to Queue Stability**

Observation

Control

Objective Function

Minimize: $V \cdot \boxed{y_o(\alpha[t])} + \boxed{Q[t]}\{\boxed{a(\alpha[t])} - \boxed{b(\alpha[t])}\}$

Objective Function (Utility)
- $\alpha[t]$: control action at time t
- $y_o(\alpha[t])$: objective w/ control action $\alpha[t]$

Stability
- $Q[t]$: queue-backlog at time t
- $a(\alpha[t])$: arrivals w/ $\alpha[t]$
- $b(\alpha[t])$: departures w/ $\alpha[t]$

Large Delay

$Q[t] = \infty$

- **Time-Average Optimization subject to Queue Stability**

Observation

Control

Objective Function

$$\text{Minimize: } V \cdot \boxed{y_o(\alpha[t])} + \boxed{Q[t]}\{\boxed{a(\alpha[t])} - \boxed{b(\alpha[t])}\}$$

Objective Function (Utility)

- $\alpha[t]$: control action at time t
- $y_o(\alpha[t])$: objective w/ control action $\alpha[t]$

Stability

- $Q[t]$: queue-backlog at time t
- $a(\alpha[t])$: arrivals w/ $\alpha[t]$
- $b(\alpha[t])$: departures w/ $\alpha[t]$

Large Delay

$$Q[t] = \infty$$

If the arrival $a(\alpha[t])$ is controllable, it should be minimized.

• Time-Average Optimization subject to Queue Stability

Observation

Control

Objective Function

$$\text{Minimize: } V \cdot y_o(\alpha[t]) + Q[t]\{a(\alpha[t]) - b(\alpha[t])\}$$

Objective Function (Utility)

- $\alpha[t]$: control action at time t
- $y_o(\alpha[t])$: objective w/ control action $\alpha[t]$

Stability

- $Q[t]$: queue-backlog at time t
- $a(\alpha[t])$: arrivals w/ $\alpha[t]$
- $b(\alpha[t])$: departures w/ $\alpha[t]$

Large Delay

$Q[t] = \infty$

If the arrival $a(\alpha[t])$ is controllable, it should be minimized.

If the departures $b(\alpha[t])$ is controllable, it should be maximized.

- **Time-Average Optimization subject to Queue Stability**

Observation

Control

Objective Function

$$\text{Minimize: } V \cdot \boxed{y_o(\alpha[t])} + \boxed{Q[t]}\{\boxed{a(\alpha[t])} - \boxed{b(\alpha[t])}\}$$

Objective Function (Utility)

- $\alpha[t]$: control action at time t
- $y_o(\alpha[t])$: objective w/ control action $\alpha[t]$

Stability

- $Q[t]$: queue-backlog at time t
- $a(\alpha[t])$: arrivals w/ $\alpha[t]$
- $b(\alpha[t])$: departures w/ $\alpha[t]$

No Delay

$$Q[t] = 0$$

- **Time-Average Optimization subject to Queue Stability**

$$\text{Minimize: } V \cdot \overbrace{y_o(\alpha[t])}^{\text{Objective Function}} + \overbrace{Q[t]}^{\text{Observation}}\overbrace{\{a(\alpha[t]) - b(\alpha[t])\}}^{\text{Control}}$$

Objective Function (Utility)

- $\alpha[t]$: control action at time t
- $y_o(\alpha[t])$: objective w/ control action $\alpha[t]$

Stability

- $Q[t]$: queue-backlog at time t
- $a(\alpha[t])$: arrivals w/ $\alpha[t]$
- $b(\alpha[t])$: departures w/ $\alpha[t]$

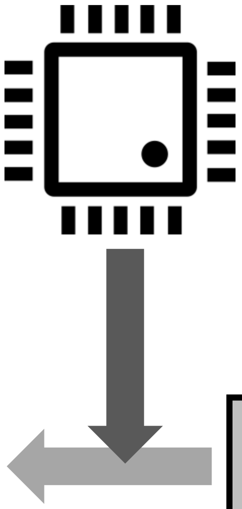
No Delay

$$Q[t] = 0$$

Utility Maximization (i.e., Minimizing the Objective Function)

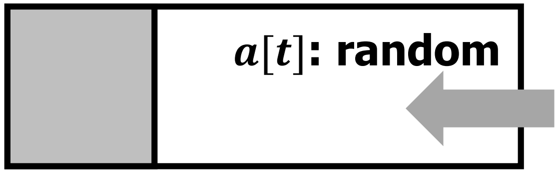
- **Research #1 (Core Allocation)**

Processor



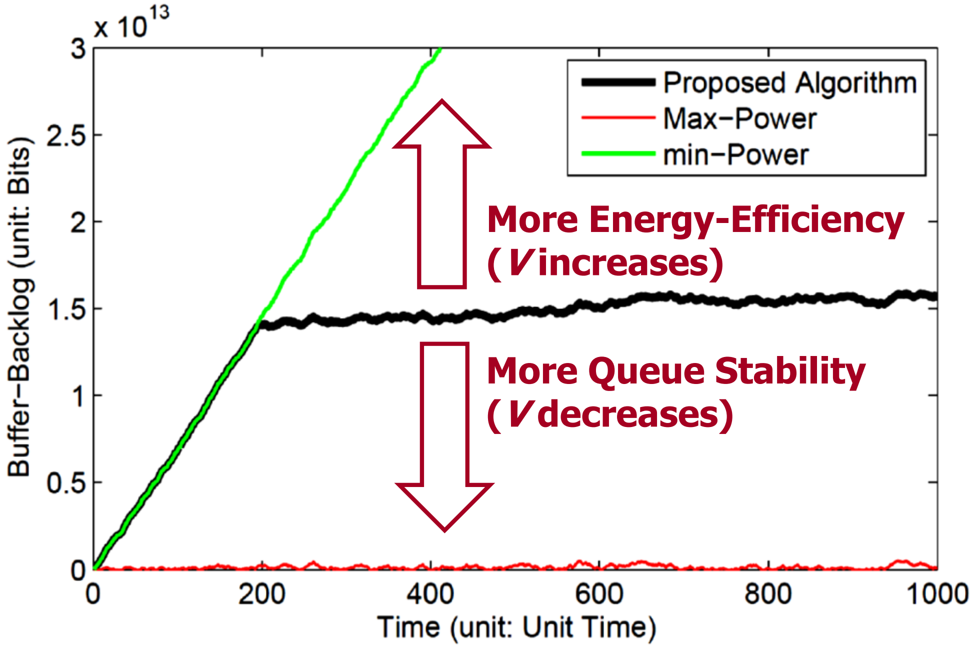
Tradeoff

- <u>More core allocation</u>
  - More energy consumption (-)
  - More stability (+)
- <u>Less core allocation</u>
  - Less energy consumption (+)
  - More stability (-)

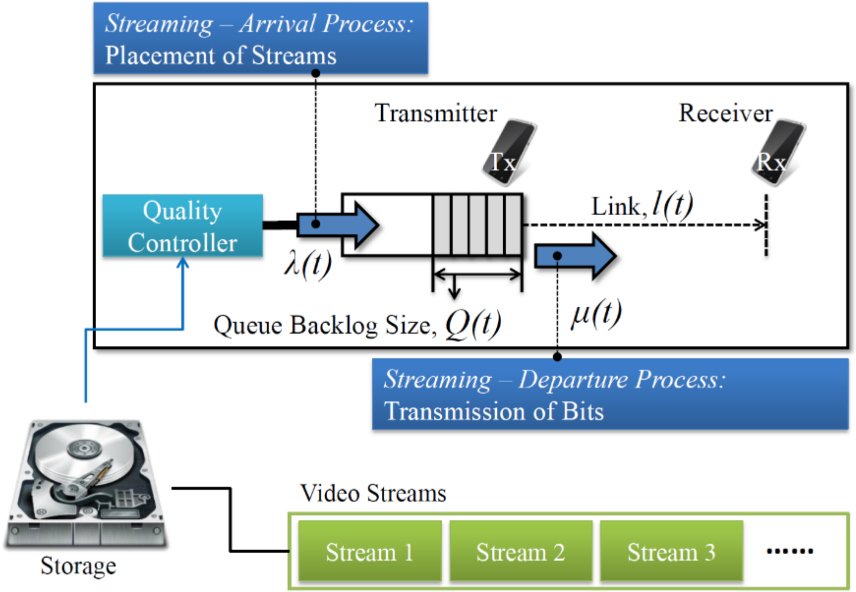


$b[t]$: depends on core allocation

Task Queue $Q[t]$



- **Research #2 (Adaptive Wireless Video Streaming)**
  - Maximization of **Time-Average Video Quality** subject to **Queue Stability**



Tradeoff

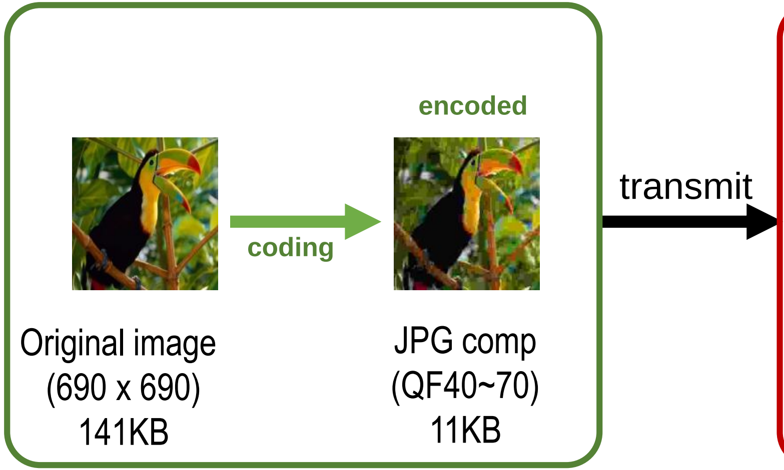
- High compression on chunks
  - Low quality on chunks (-);
  - More stabilization on queues (+)
- Less compression on chunks
  - High quality on chunks (+);
  - Less stabilization on queues (-)

J. Kim, G. Caire, and A.F. Molisch, "**Quality-Aware Streaming and Scheduling for Device-to-Device Video Delivery**," **IEEE/ACM Transactions on Networking**, 24(4):2319-2331, August 2016.

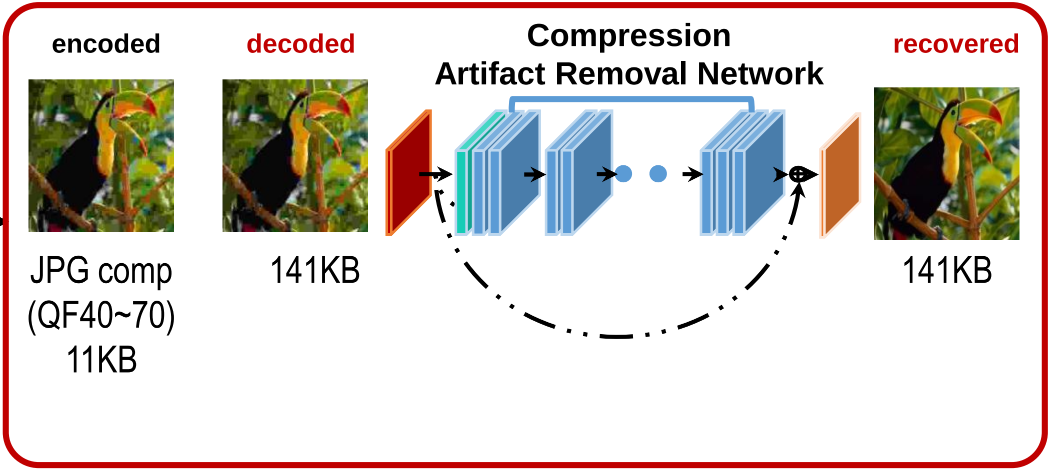
- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - JPEG image compression to reduce transmission overhead

- **Insufficient computational resource**
- **Not available to preprocess**
- **Require real-time properties**
- **Performance varies by device**

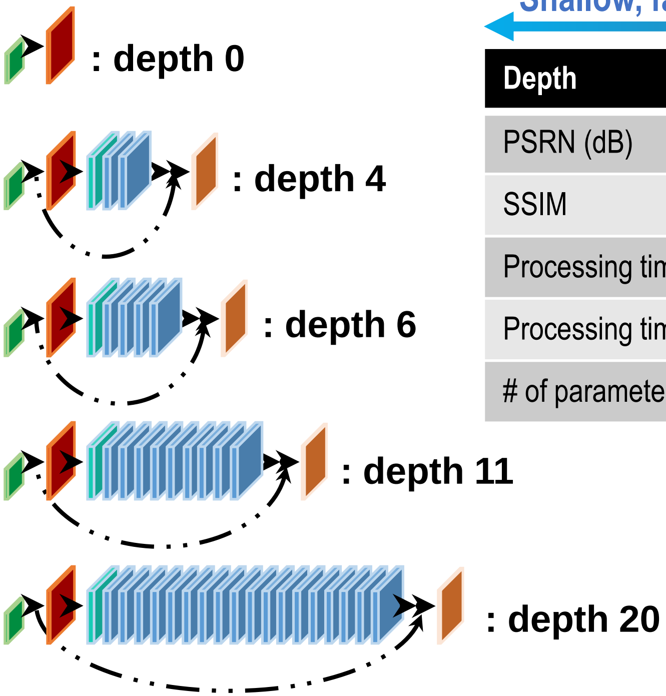
Drone



MEC



- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - Tradeoff between speed and performance over depth

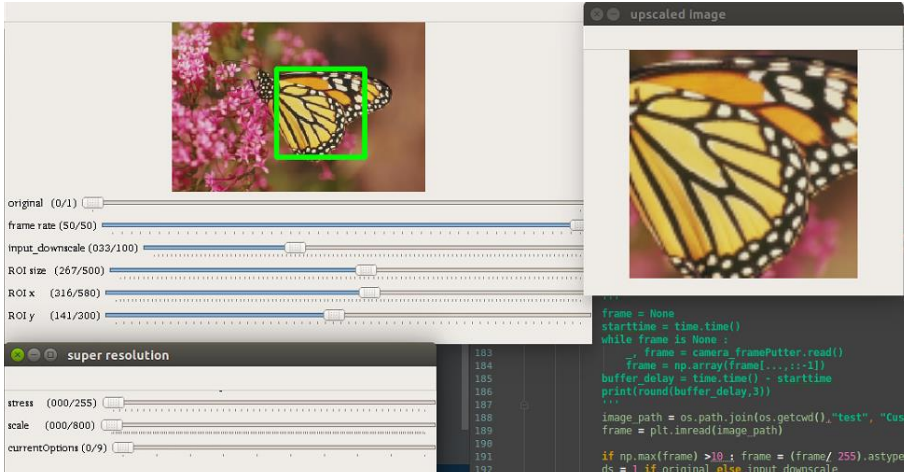


Shallow, faster, lower performance ⟷ Deeper, slower, higher performance

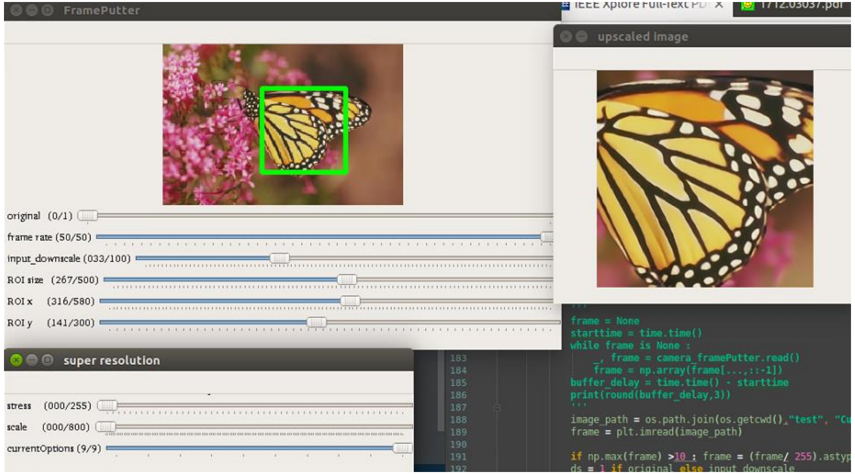
Depth	0	4	6	8	11	14	17	20
PSRN (dB)	30.4	32.56	33.01	33.229	33.379	33.435	33.495	33.523
SSIM	0.8682	0.91	0.916	0.918	0.92	0.92	0.921	0.922
Processing time (CPU)	0.002	0.321	0.5468	0.7725	0.994	1.317	1.622	1.96
Processing time (GPU)	0.001	0.01	0.012	0.0152	0.0189	0.0224	0.0262	0.0305
# of parameters	0	75K	148K	222K	333K	444K	555K	665K

(processing time have measured on butterfly, 512 x 768)

- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - Super-resolution task



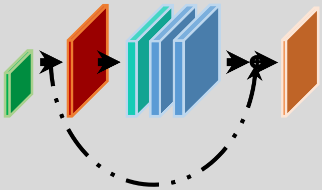
Depth 0



Depth 20

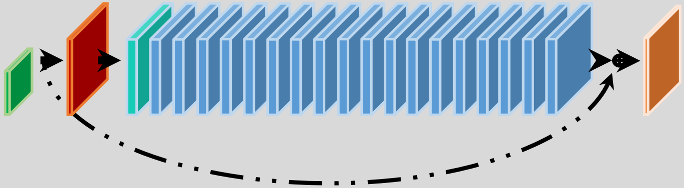
- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - Model selection

Shallow Model



- Faster, lighter
- Lower performance
- Suit for mobile, IoT devices

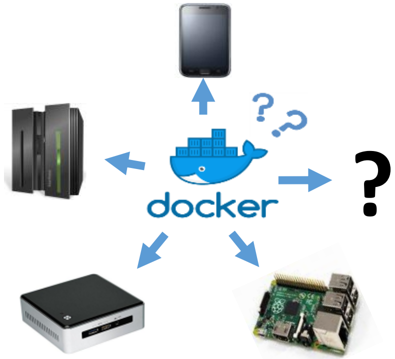
Deep Model



- Slower, heavier
- Higher performance
- Suit for server, station

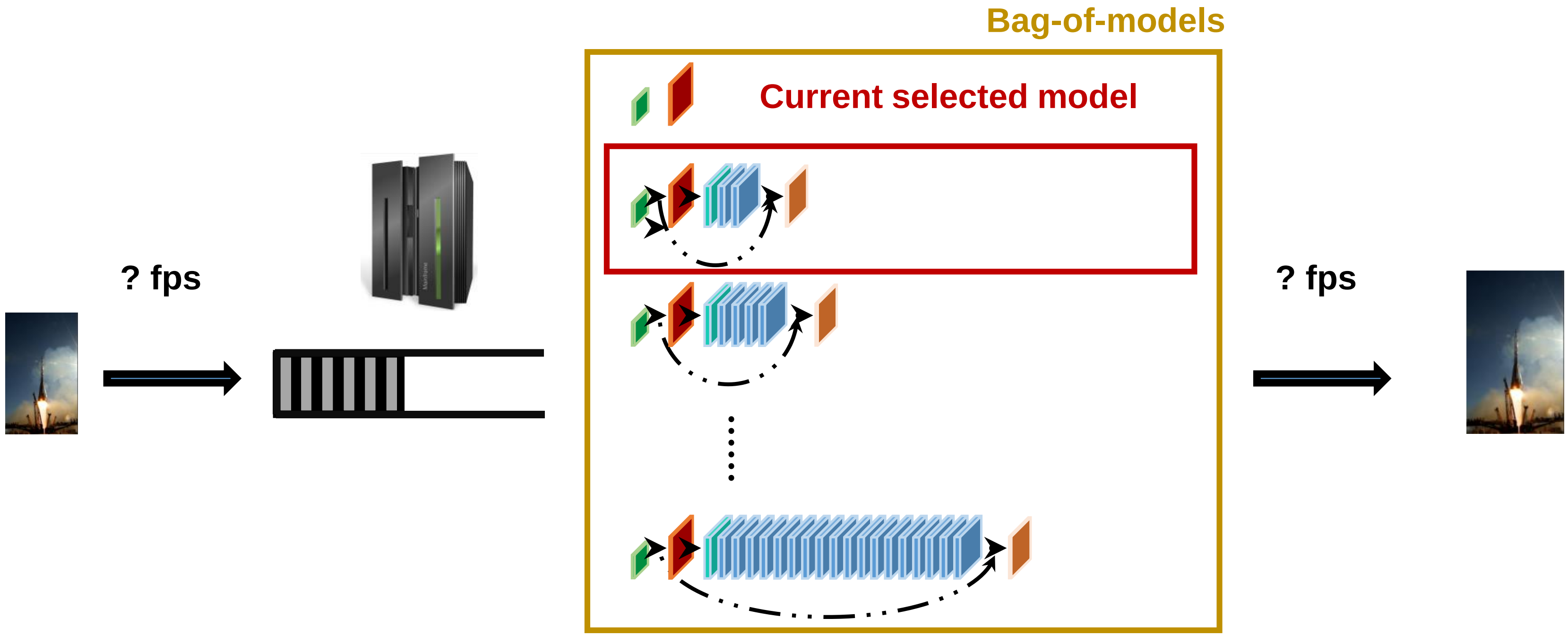
- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - Dynamic model adaptation

- Input rate (or image size) is unfixed.
- Required resolution of image is changed.
- Performance of system fluctuates frequently during running time.

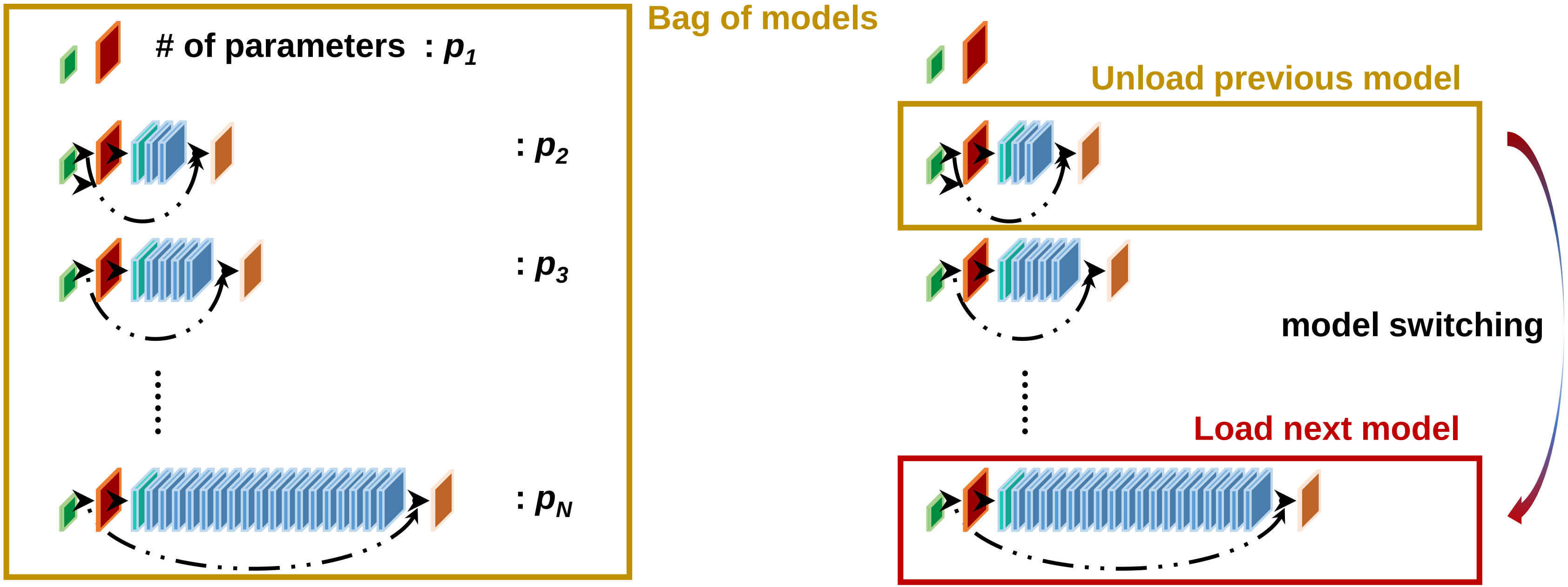


- Service provide do not have any information about clients.

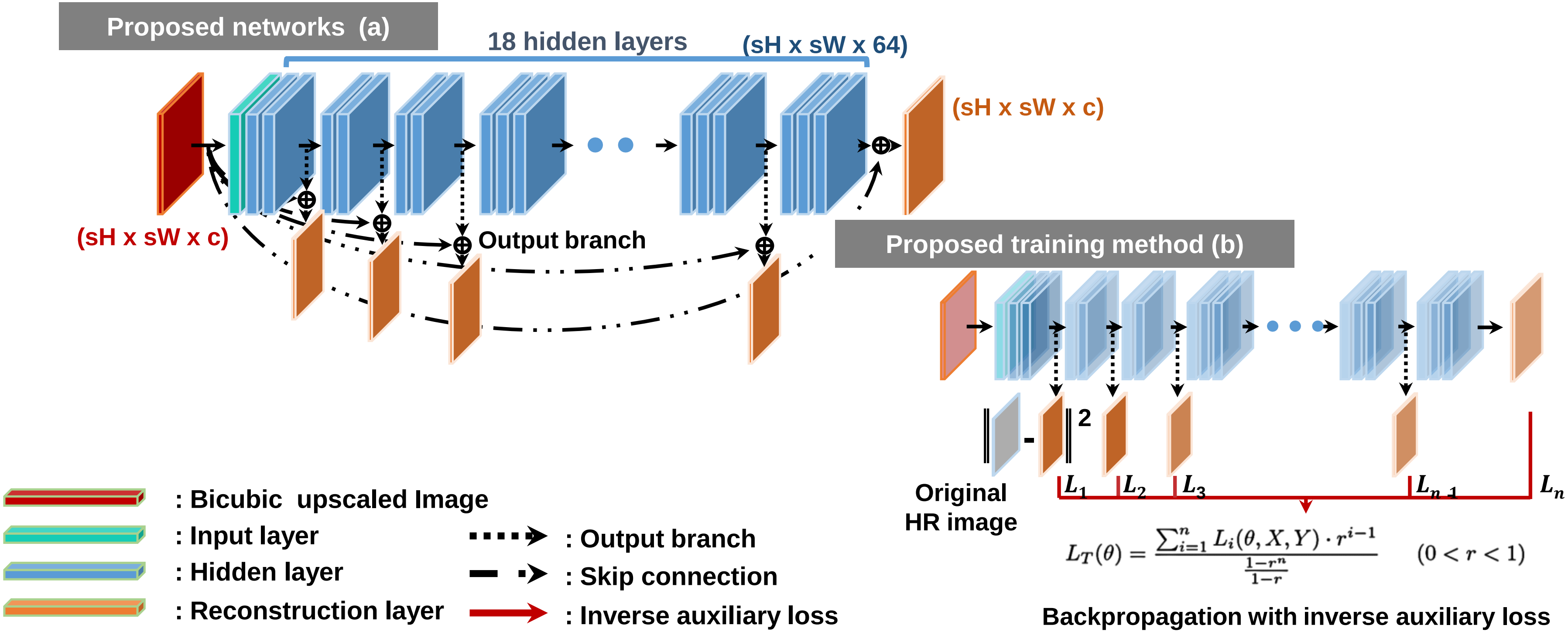
- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - Adaptation with bag-of-models



- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - Limitation to the adaptation with bag-of-models
    - **Additional memory** is required in proportion to the number of models.
    - Model switching may cause **computational overhead**.

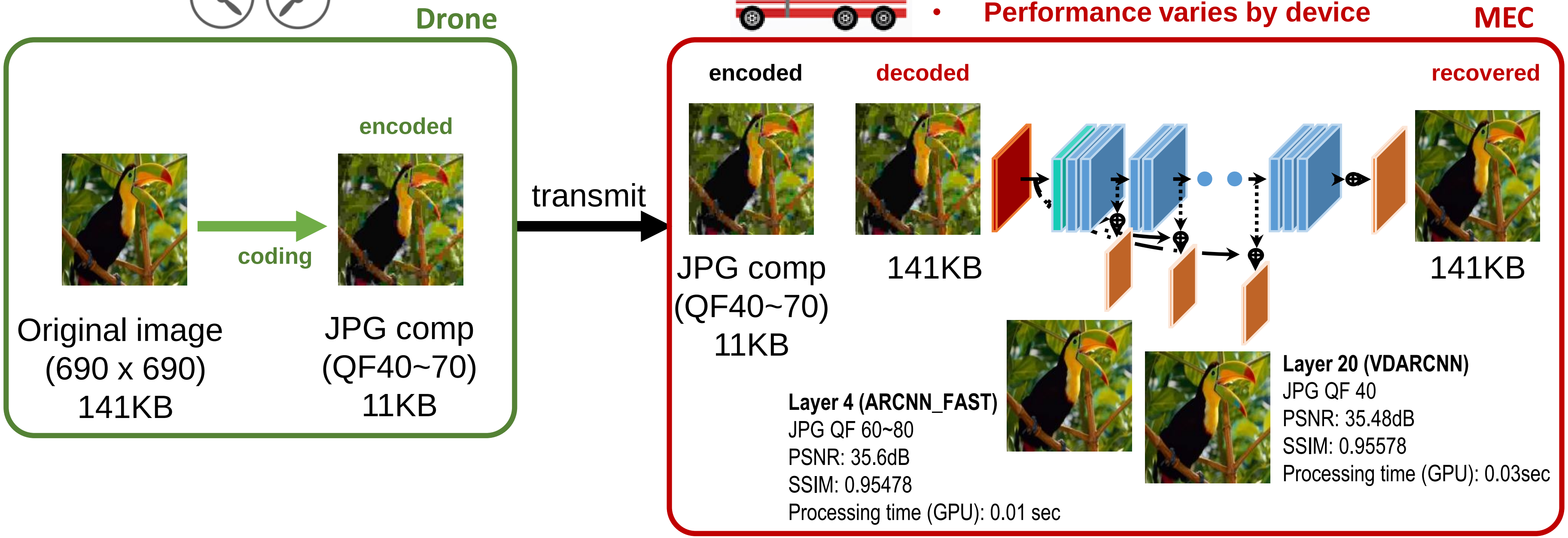


- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - Overall architecture of the proposed SR network

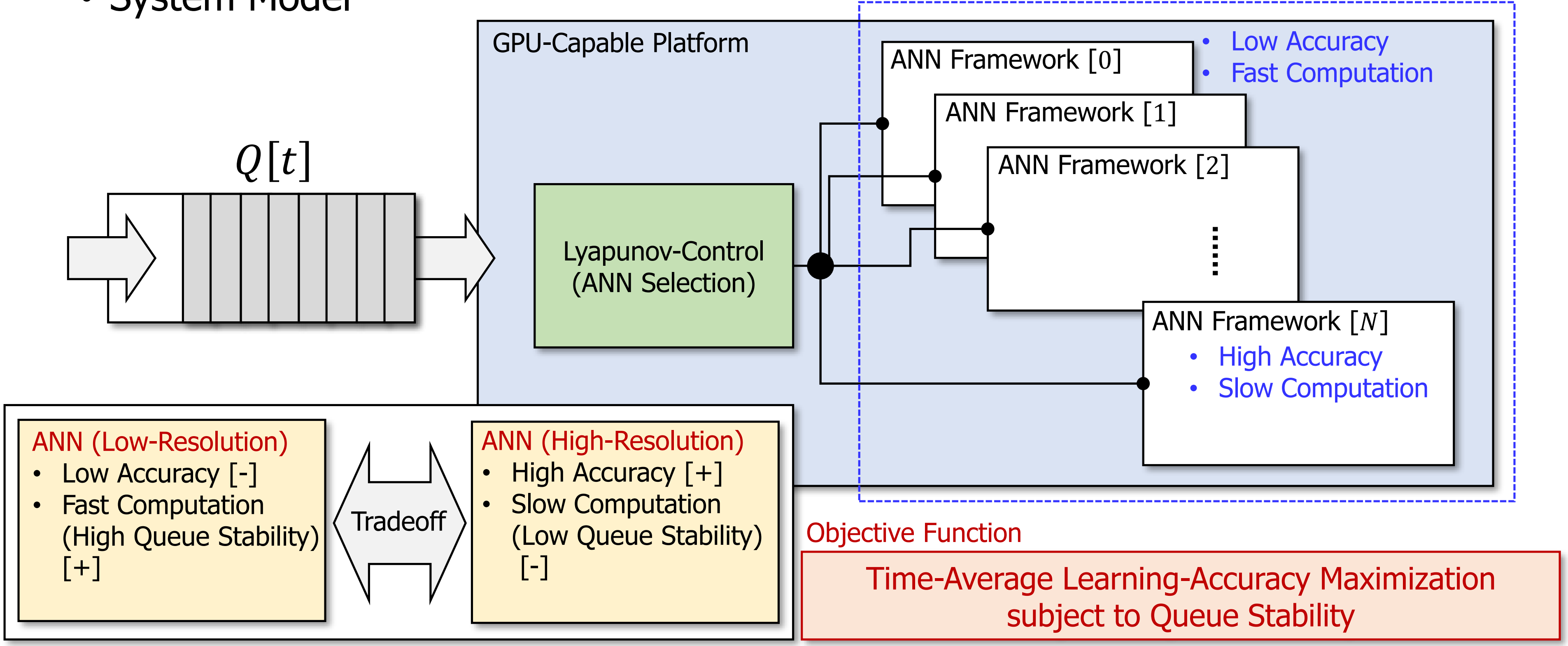


- **Research #3 (Depth-Adaptive Deep Super-Resolution Network)**
  - JPEG image compression to reduce transmission overhead

- **Insufficient computational resource**
- **Not available to preprocess**
- **Require real-time properties**
- **Performance varies by device**



- **Research #4 (Stabilized Computer Vision Platform)**
  - System Model



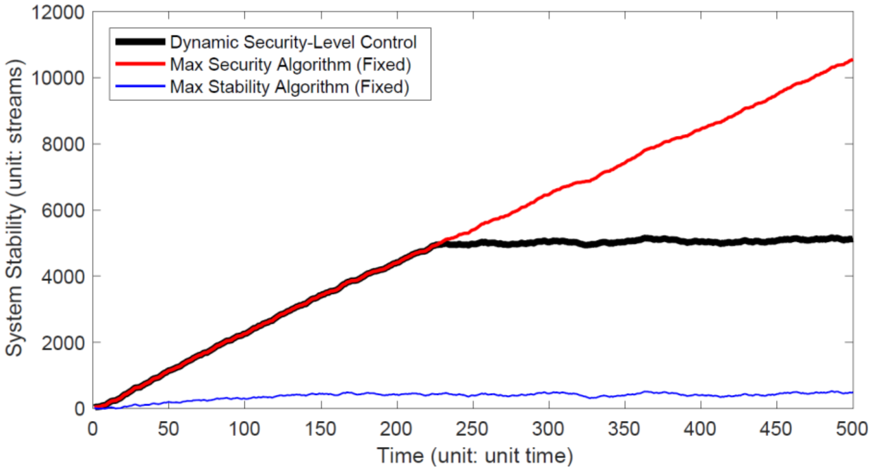
- **Research #4 (Stabilized Computer Vision Platform)**
  - Maximization of **Time-Average Learning-Accuracy** subject to Queue Stability

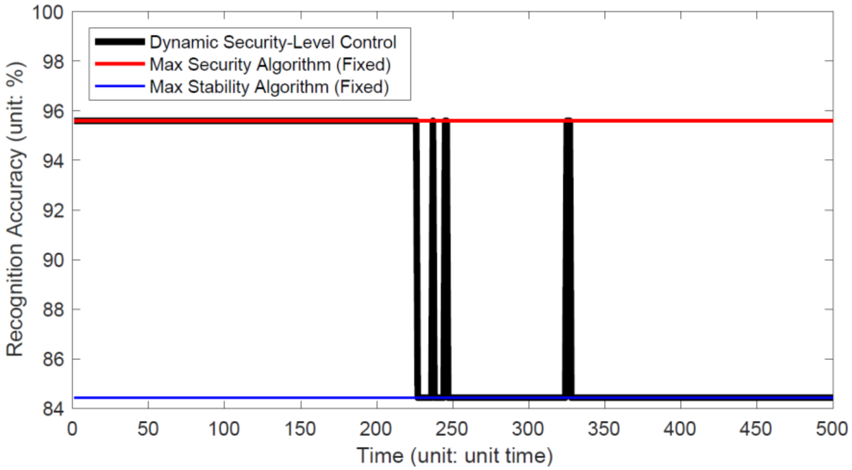
$$\alpha^*[t] \leftarrow \underset{\alpha[t]\in A}{argmax}[V \cdot Accuracy(\alpha[t]) - Q[t]\{\underbrace{a(\alpha[t])}_{\text{Not Controllable}} - b(\alpha[t])\}]$$

$$\alpha^*[t] \leftarrow \underset{\alpha[t]\in A}{argmax}\{V \cdot Accuracy(\alpha[t]) + Q[t]b(\alpha[t])\}$$

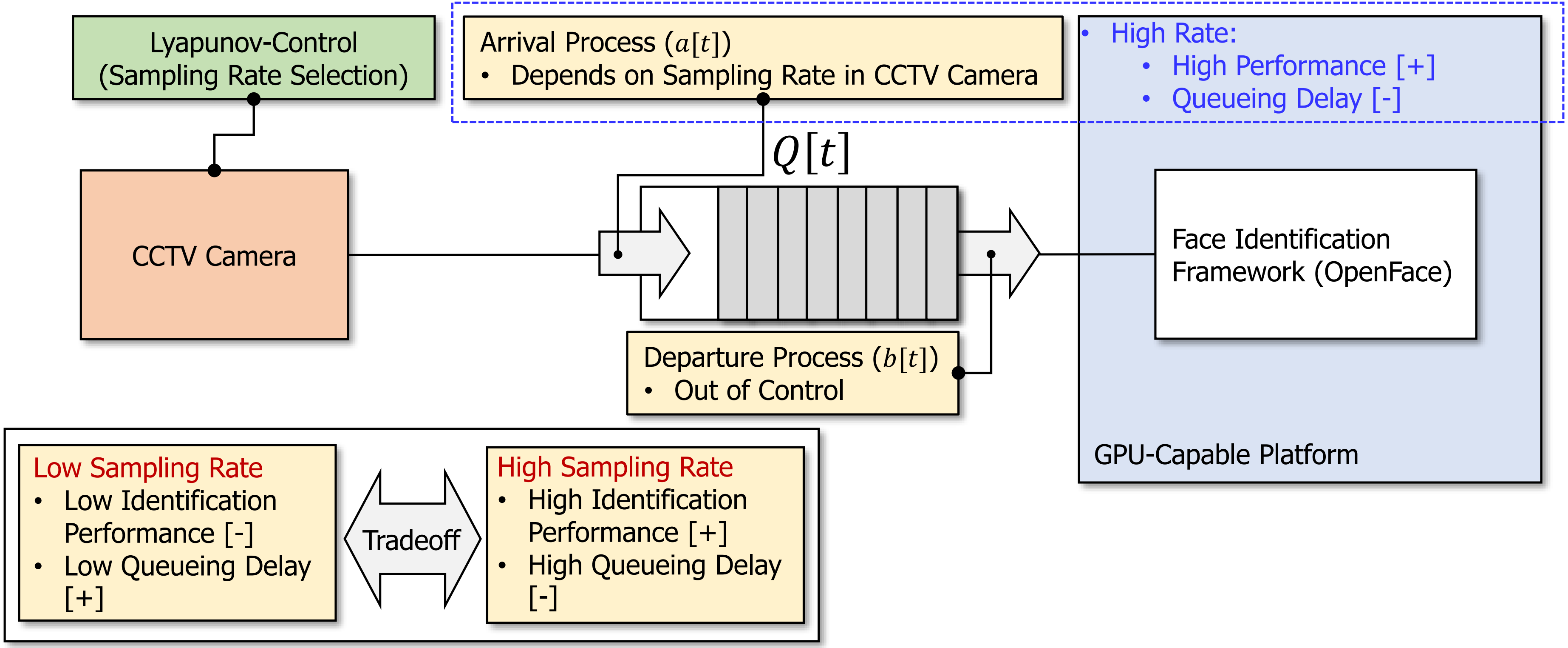
- Semantical Description
  - If the queue is near empty ($Q[t] \cong 0$),
    - Select the $\alpha[t]$ which can maximize $V \cdot Accuracy(\alpha[t])$, i.e., high learning-accuracy ANN will be selected.
  - If the queue is near overflow ($Q[t] \cong \infty$),
    - Select the $\alpha[t]$ which can maximize $b(\alpha[t])$, i.e., fast computation (i.e., low learning-accuracy) ANN will be selected.

- **Research #4 (Stabilized Computer Vision Platform)**
  - Surveillance Platforms with Parallel Deep Learning Frameworks





- **Research #5 (Stabilized Computer Vision Platform: Revisited)**
  - System Model



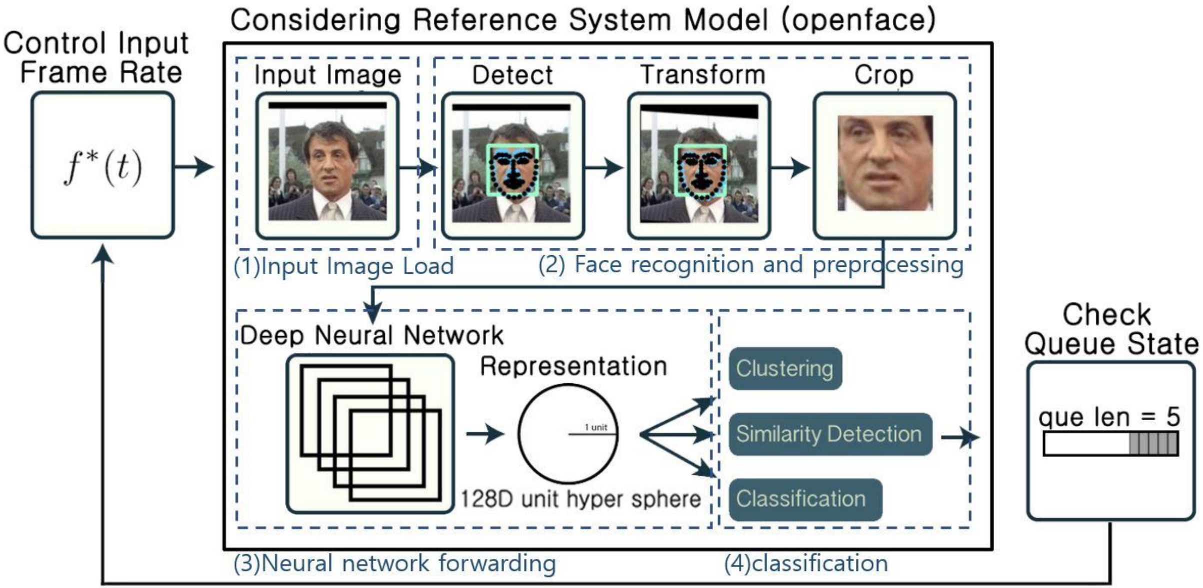
- **Research #5 (Stabilized Computer Vision Platform: Revisited)**
  - Maximization of **Time-Average Learning-Accuracy** subject to Queue Stability

$$\alpha^*[t] \leftarrow \underset{\alpha[t]\in A}{argmax}[V \cdot Accuracy(\alpha[t]) - Q[t]\{a(\alpha[t]) - \underbrace{b(\alpha[t])}_{\text{Not Controllable}}\}]$$

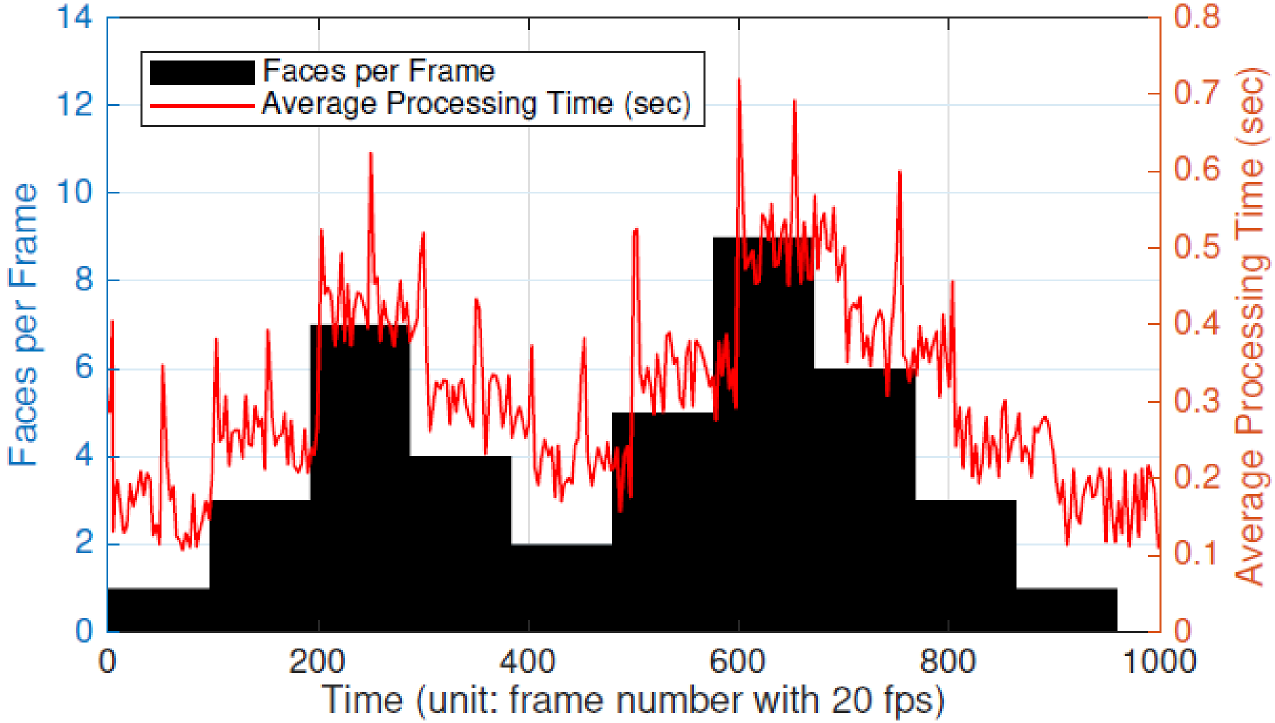
$$\alpha^*[t] \leftarrow \underset{\alpha[t]\in A}{argmax}\{V \cdot Accuracy(\alpha[t]) - Q[t]a(\alpha[t])\}$$

- Semantical Description
  - If the queue is near empty ($Q[t] \cong 0$),
    - Select the $\alpha[t]$ which can maximize $V \cdot Accuracy(\alpha[t])$, i.e., high learning-accuracy ANN will be selected.
  - If the queue is near overflow ($Q[t] \cong \infty$),
    - Select the $\alpha[t]$ which can maximize $b(\alpha[t])$, i.e., fast computation (i.e., low learning-accuracy) ANN will be selected.

- **Research #5 (Stabilized Computer Vision Platform: Revisited)**
  - Block Diagram



- **Research #5 (Stabilized Computer Vision Platform: Revisited)**
  - Implementation



- **Research #5 (Stabilized Computer Vision Platform: Revisited)**
  - Implementation