AI for System & System for AI

Seyeon Kim

Department of Computer Science and Engineering

Korea University, South Korea

seyeon625@korea.ac.kr

2025.04.02

2025 Spring

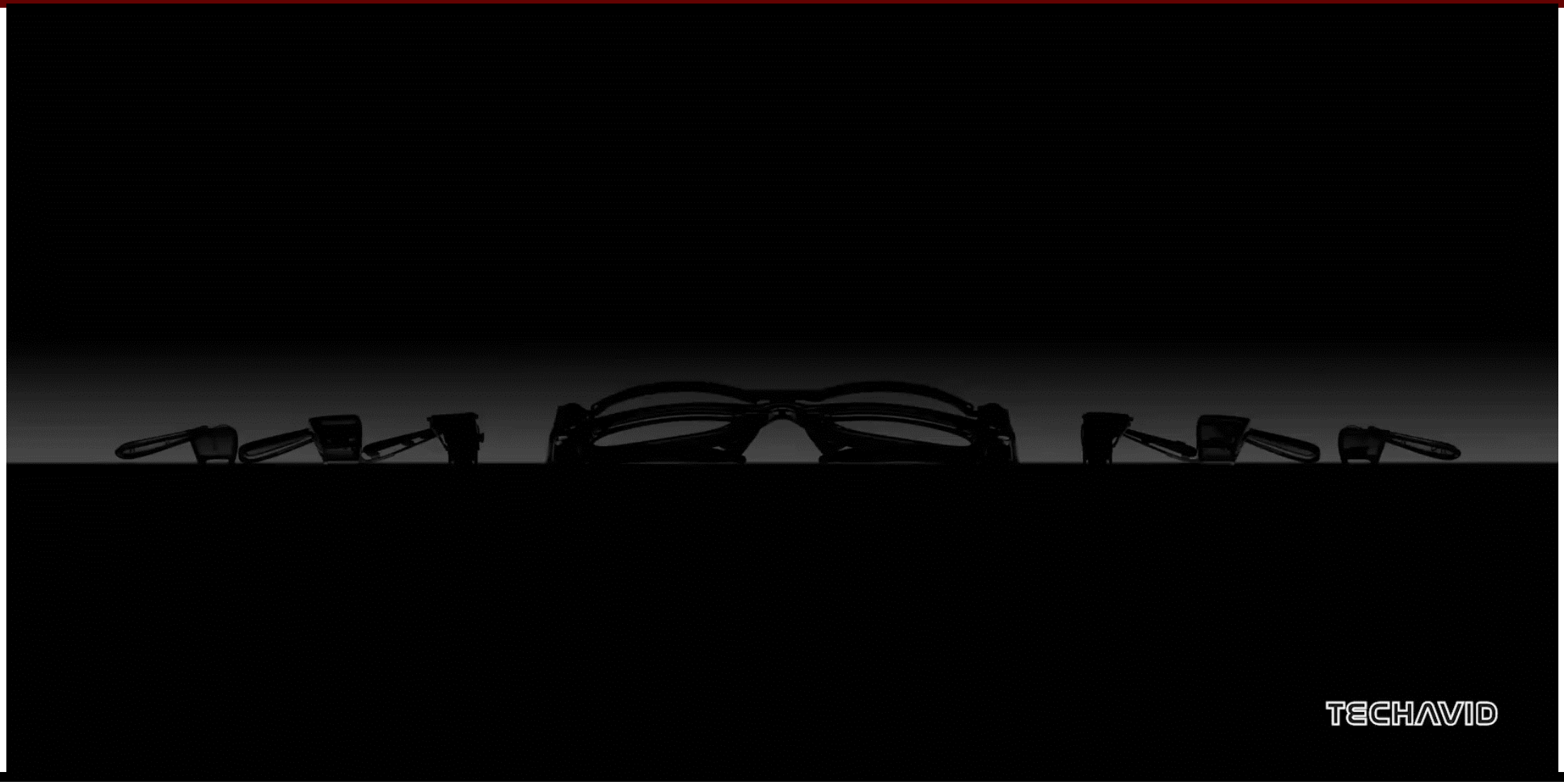
Outline

- About Instructor & Intro.
- AI for System & System for AI
  - Research 1: AI for System
  - Research 2: System for AI
- Ongoing Research Projects

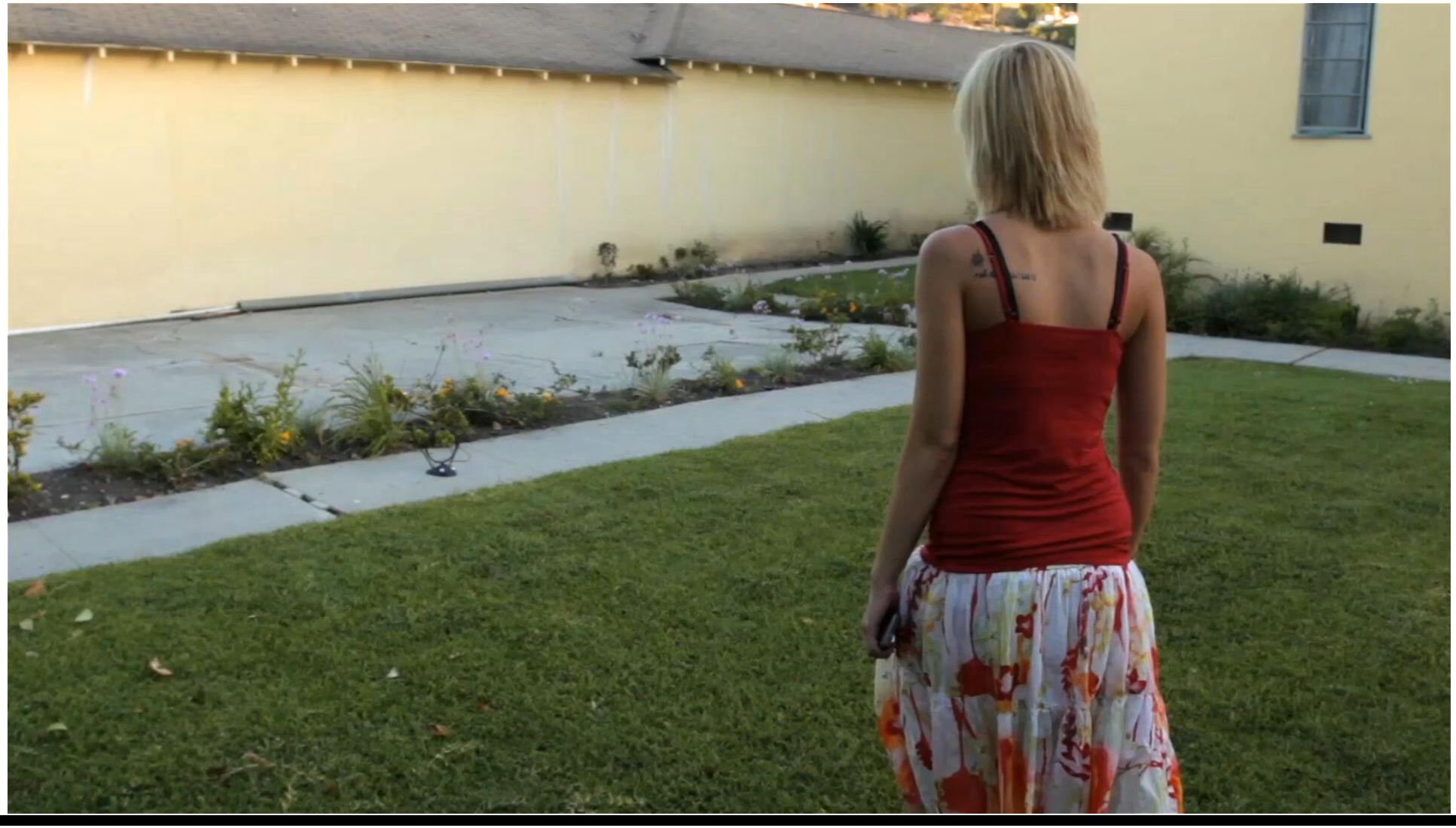
About instructor

- **Prof. Seyeon Kim**
  - System for Extended Reality and AI (SERA Lab.) (2025.03 ~)
  - Short bios
    - Assistant professor (CSE, Korea University)
    - B.S., M.S., Ph.D. (2022) from KAIST
    - 2 years of Post-doc in SNU (1Y, South Korea) and CU Boulder (1Y, USA)
    - Joined Korea University in 2025
    - Office
      - Until March: College of Science Annex (이학관별관) #204
      - From April: Hall of Informatics (우정정보관) # 506
  - Awards
    - ACM MobiSys 2021 Best Paper Award
      (Top conference in the area of mobile computing, The first best paper in Korea)

Emergence of New Devices

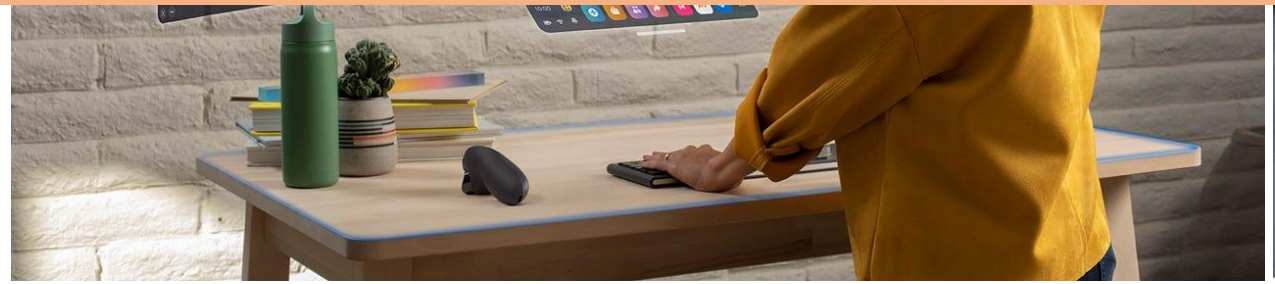


Emergence of New Applications



Limits of Hardware Advancement

Is this truly futuristic ?

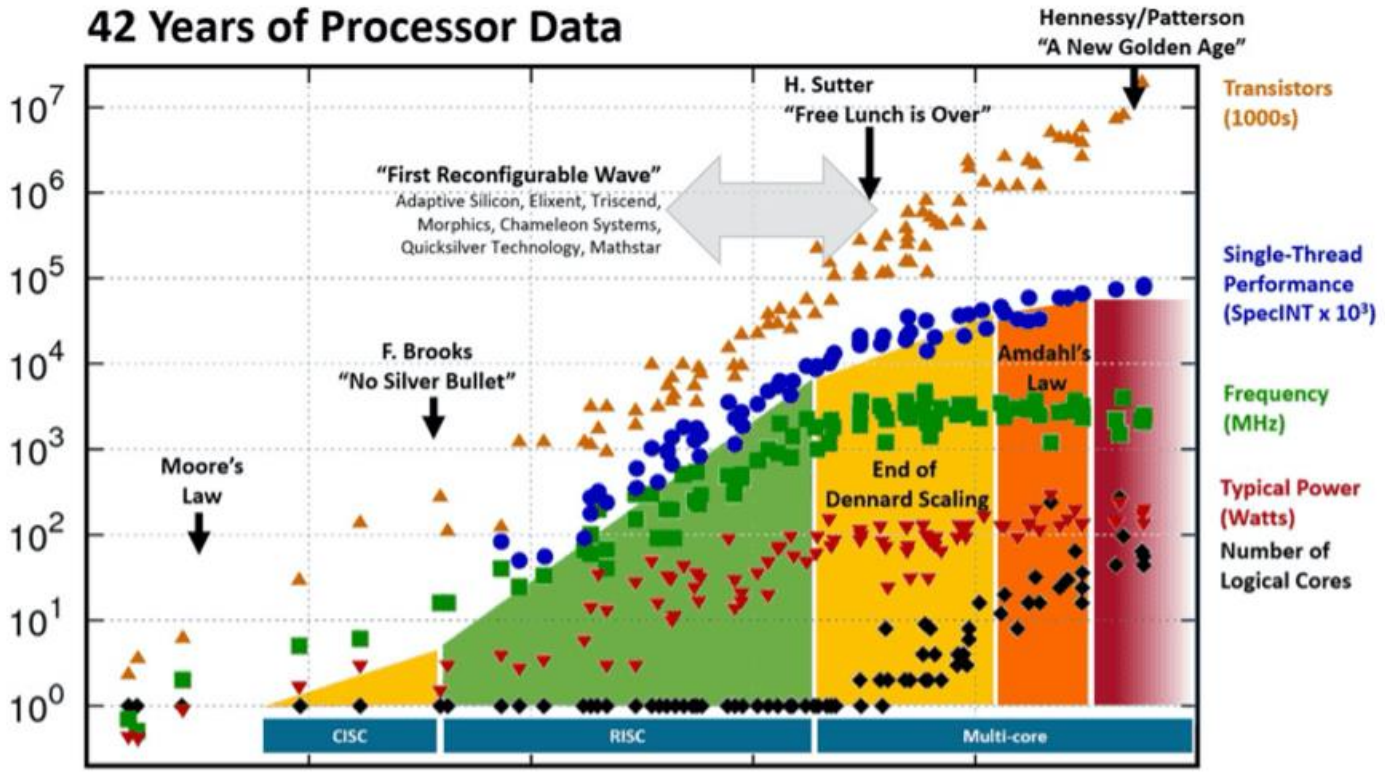


The Battery life of Meta Quest3≈1 hour

Meta Orion relies on an external computing device and a battery

Limits of Hardware Advancement

- **Dennard scaling is gone.**

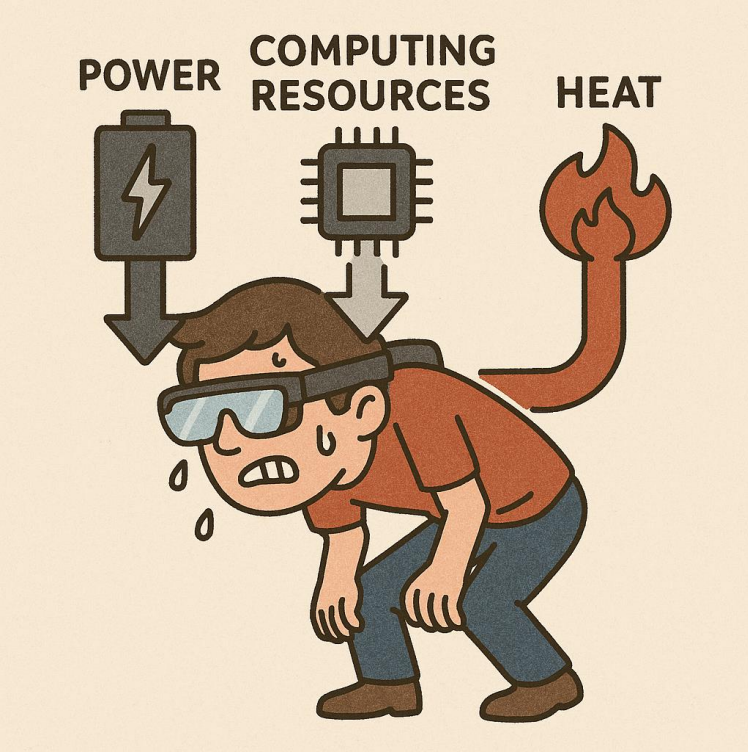


Future innovations will be driven primarily by advances in system software.

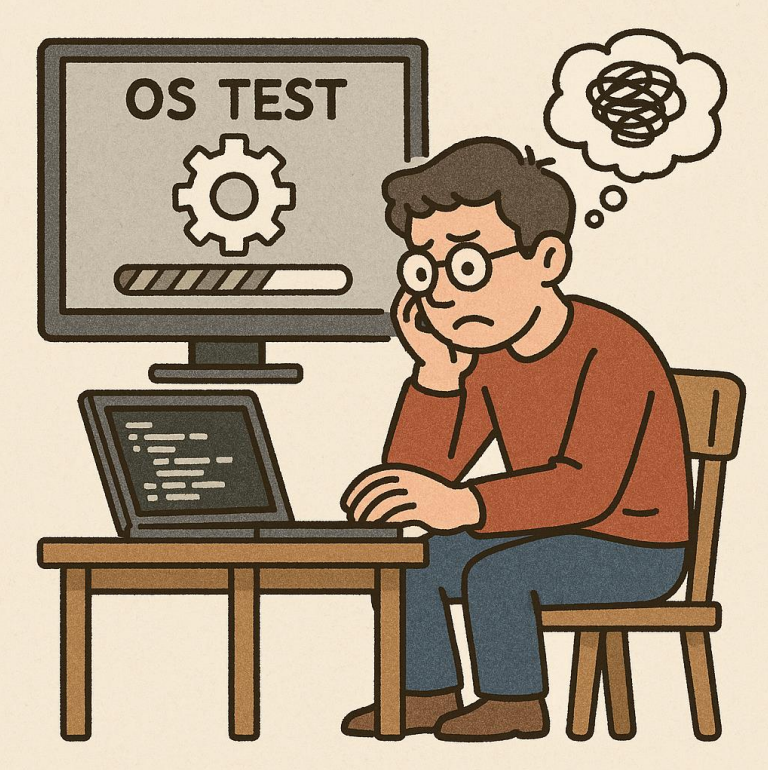
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Challenges in AI & System

- **Main challenge in AI**
  - **High resource demands**



- **Main challenge in System (Software)**
  - **Heuristic-based operating system**

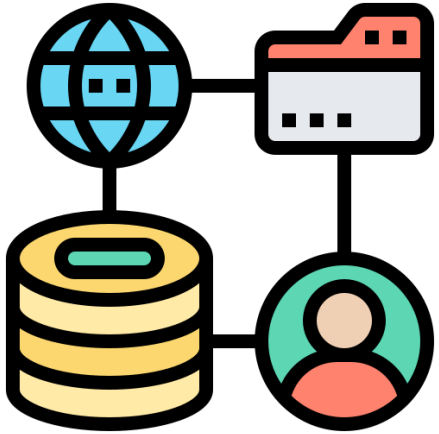


Challenges in AI & System

- **Main challenge in AI: Cost**
  - **High resource demands**
    - Computing and networking
    - Power and thermal inefficiency

AI

- **Main challenge in System**
  - **Heuristic-based operating system**
    - Empirical modeling-based rule
    - Human experience-based rule

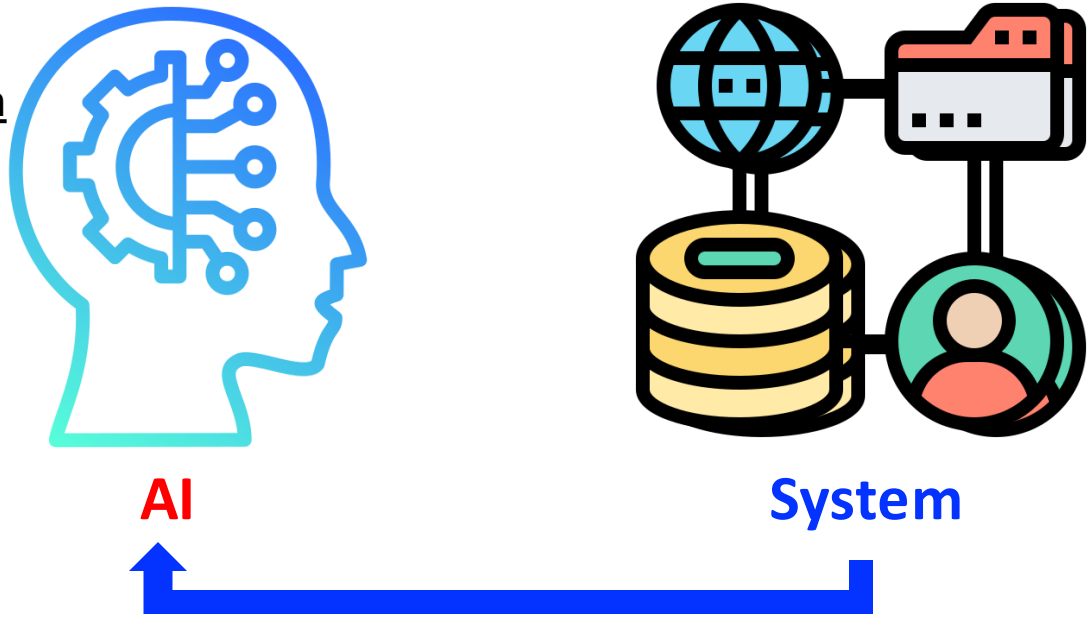


System

Q) How can we further improve the System ?

AI for System & System for AI

- **AI for System**
  - **Goal:**
    - To improve system performance
  - **Methodology**
    - Heuristic
      ⬇
      AI-based control
  - **Papers**
    - zTT (ACM Mobisys'21, Best paper)
    - R-FEC (ACM MM'22, Oral)
    - Dejavu (IEEE MASS'24)



- **System for AI**
  - **Goal:**
    - Resource-efficient AI
    - Better performance
  - **Methodology:**
    - Designing a system dedicated to AI application
  - **Papers**
    - DeltaStream (ACM MobiSys'25)
    - NeuroBalancer (IEEE TMC'25)
    - ENTRO (ACM MM'23)
    - CoActo (ACM Mobisys'24)
    - LLMem(IJCAI'24, Long talk)

AI for System

zTT: Learning-based DVFS with Zero Thermal Throttling for Mobile Devices

(ACM MobiSys'21, Best paper award)

AI for System – zTT (ACM Mobisys'21)

- **Samsung Game Optimizing Service (GOS)**
  - Samsungs' thermal solution for Galaxy devices

재난포털

KBS 뉴스

분야별 뉴스9 TV 뉴스 프리미엄K 스포츠 시사프로그램 지역뉴스

취재K

'삼성의 기술 우위는 끝나버렸다' GOS 사태의 본질

입력 2022.03.19 (08:01)

요약 GOS 사태를 통해 삼성전자의 모바일·파운드리 경쟁력 위기를 조명한다. GOS 사태는 한 편에선 삼성의 스마트폰 시장 '기술 우위가 끝나버렸음'을 상징하고, 다른 한 편에선 미래 먹을거리로 손꼽히는 파운드리 역시 기술적인 난관에 부딪혀 있음을 보여주기 때문이다.

THEELEC ≡ All News Latest Stories Semiconductor Display Panel BA

HOME > Product

Samsung CEO apologizes for Galaxy smartphone GOS controversy

Lee Sang Won | 승인 2022.03.16 17:09 | 댓글 0

Vice chairman Han also defends mobile chief's appointment as director

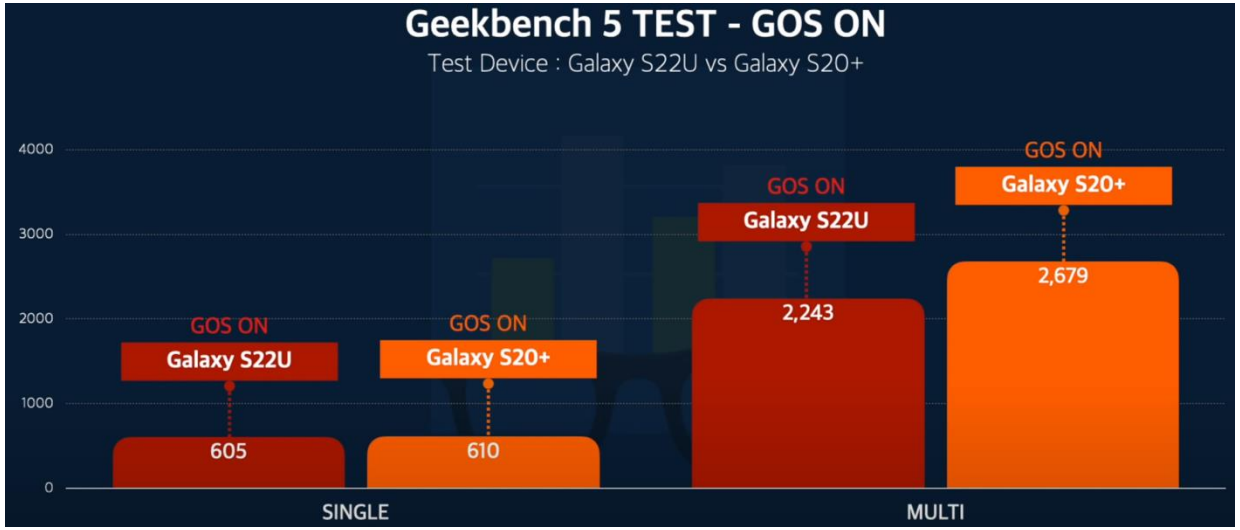
Samsung GOS caused a drastic drop in performance

AI for System – zTT (ACM Mobisys'21)

- **Samsung Game Optimizing Service (GOS)**
  - Samsungs' thermal solution for Galaxy devices
  - **GOS solution:** App-specific power (Clock frequency) control

pkgName	category	fixed
com.sec.android.app.camera	non-game	1
com.microsoft.office.word	non-game	1
com.microsoft.office.outlook	non-game	1
com.microsoft.office.excel	non-game	1
com.microsoft.office.powerpoint	non-game	1
jp.naver.line.android	non-game	1
com.instagram.android	non-game	1
com.kakao.talk	non-game	1
com.google.android.youtube	non-game	1
com.disney.disneyplus	non-game	1
com.netflix.mediaclient	non-game	1
com.netflix.Speedtest	non-game	0
tv.twitch.android.app	non-game	1

Example of app list in GOS

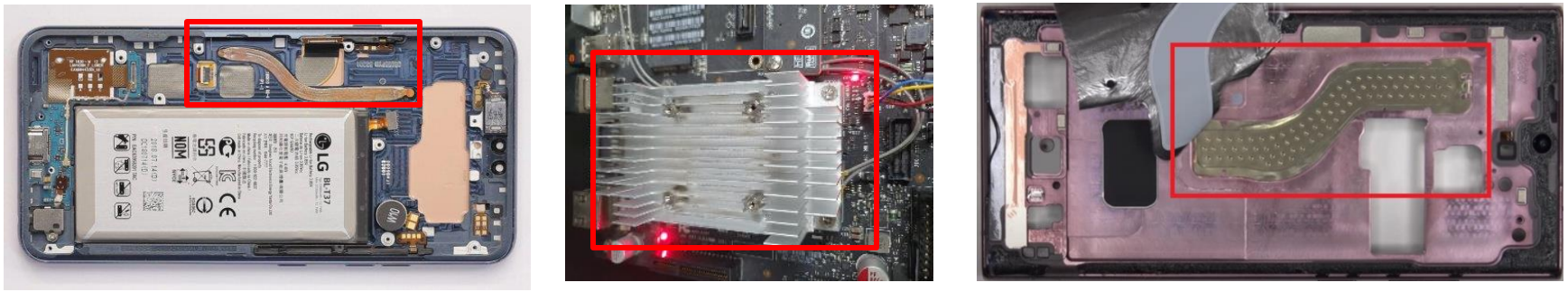


The performance of S22U is even worse than S20

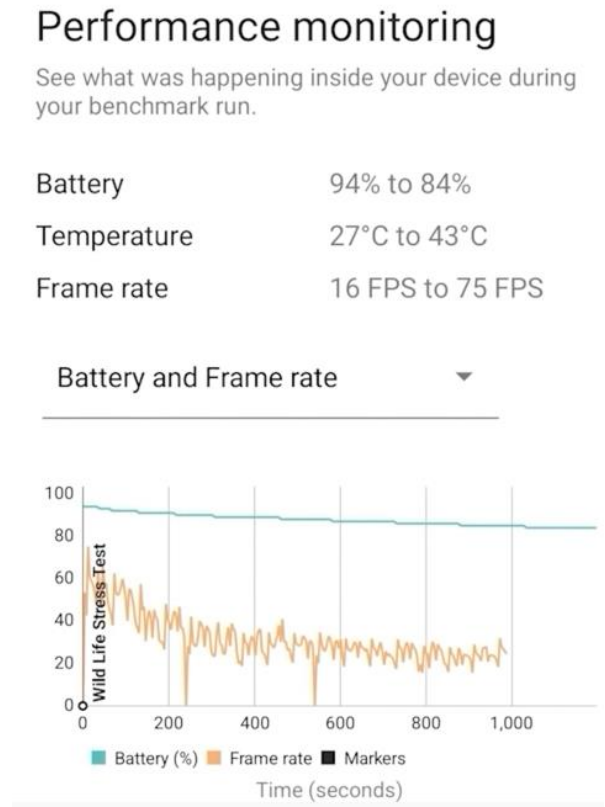
Increasing power for better performance is no longer valid !

AI for System – zTT (ACM Mobisys'21)

- **What's the problem ?**
  - **Thermal problem**
    - **Powerful mobile processors generate a lot of heat in a compact space**
    - **Weak cooling power**
    - **Passive cooling methods: Heat sink / Heat pipe, Vapor Chamber**



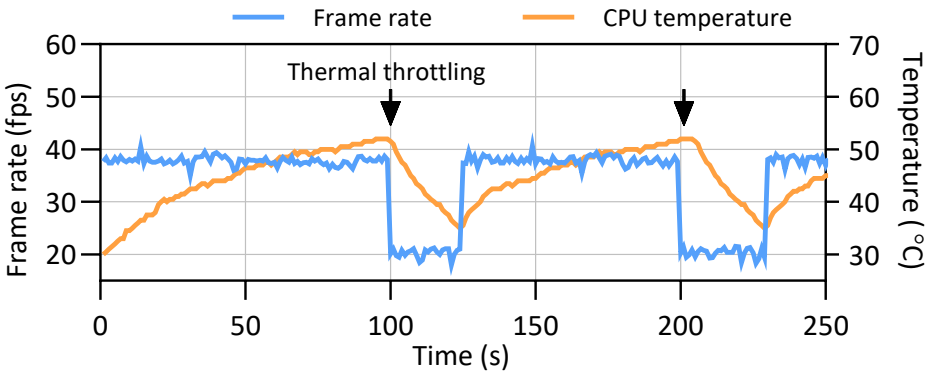
Passive cooling methods on mobile devices



Stress test with Geekbench5 on a Samsung S22+ smartphone

AI for System – zTT (ACM Mobisys'21)

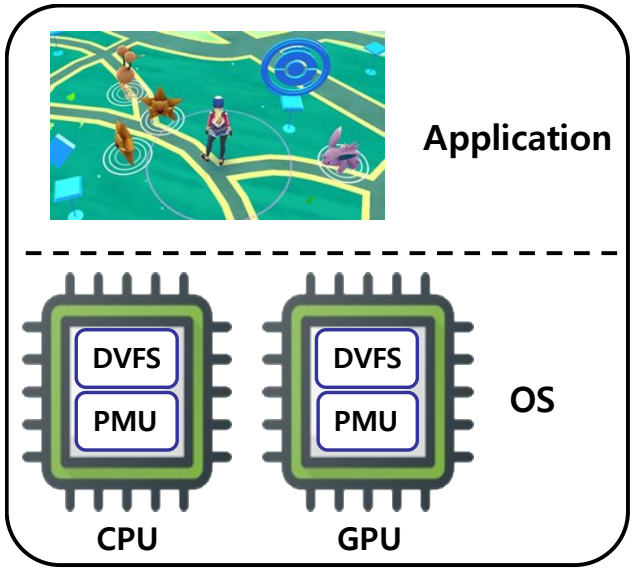
- **Existing solutions and limitations**
  - **Thermal throttling**
    - OS-level technique to prevent overheating
    - When overheated, processors lowers clock frequency



➡ Can result in significant performance degradation repeatedly.

  - **Dynamic Voltage and Frequency Scaling (DVFS)**
    - OS-level technique to improve power efficiency
    - Dynamically adjust Voltage-Frequency (VF) level.
    - Based on the predefined processor utilization levels.

➡ Application performance-agnostic control



* Governor: DVFS module
* PMU: Performance Monitoring Unit

AI for System – zTT (ACM Mobisys'21)

- **Research Goals**
  - **To overcome these limitations..**

$$\textbf{P(0):}\quad \max_{\pi} \frac{1}{T}\sum_{t=1}^{T}\left\{U(t)+\frac{\beta}{P(t)}\right\}$$

$$s.t.\quad T_C(t) \le T_{C,th}, \quad \forall t$$
$$T_G(t) \le T_{G,th}, \quad \forall t$$

$U(t)$: Utility function (User QoE) at time t
$P(t)$: Total Power consumption at time t
$T_c(t)$: CPU temperature at time t
$T_G(t)$: GPU temperature at time t
$T_{c,th}, T_{G,th}$: Threshold temperature (Constant)
π: Frequency scaling policy over t

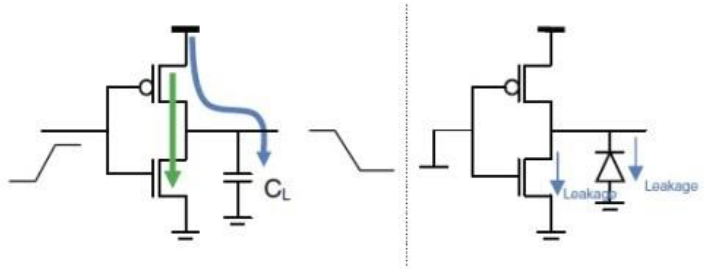
- **1. Application-aware DVFS**
  - Hybrid CPU-GPU DVFS
  - Guarantee user experience (QoE)
  - Minimize power consumption
- **2. Zero Thermal Throttling (zTT)**
  - Predict thermal budget
  - Perform DVFS within the thermal budget

AI for System – zTT (ACM Mobisys'21)

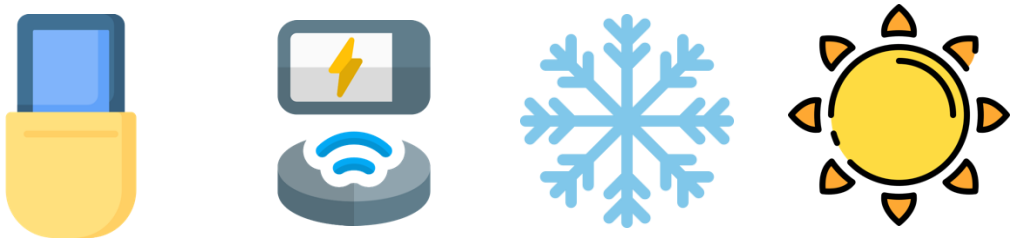
- **Difficult to predict application performance and power consumption**

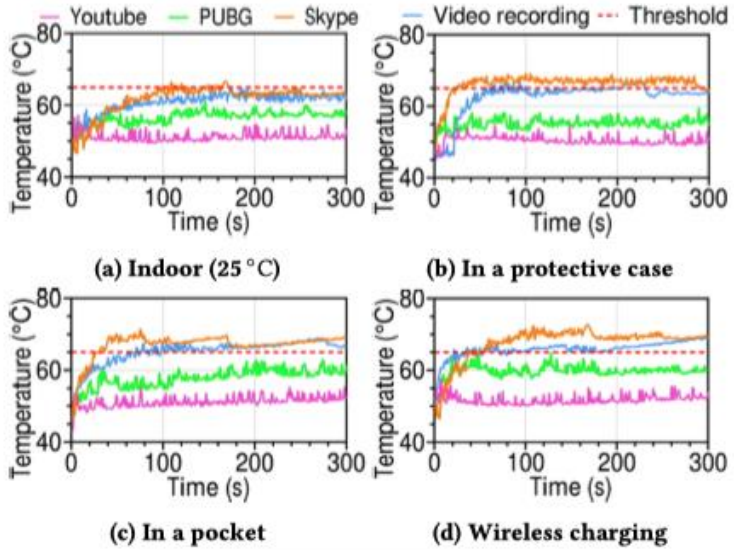
$$P_{consume} = \underbrace{P_{dynamic}}_{\propto \alpha V^2 f} + P_{short-circuit} + \underbrace{P_{leakage}}_{\propto e^{-\frac{1}{T}}}$$

***α*:** *Core utilization*
***V*:** *Voltage*
***f*:** *Frequency*
***T*:** *Temperature*



- **Difficult to predict future temperature**
  - Environmental changes
  - Thermal coupling among processors
  - Application-dependent





Mobile devices are highly affected by external environments

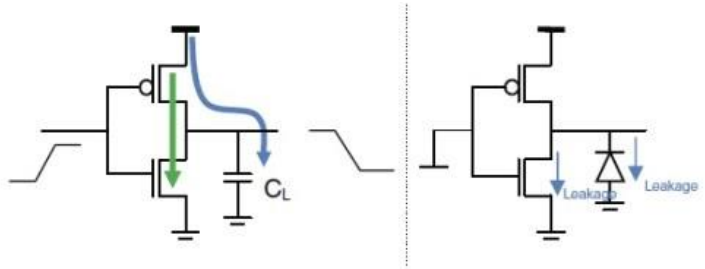
AI for System – zTT (ACM Mobisys'21)

- **Difficult to predict application performance and power consumption**

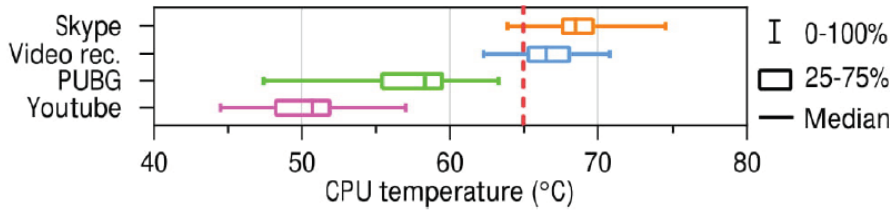
$$P_{consume} = P_{dynamic} + P_{short-circuit} + P_{leakage}$$

$$P_{dynamic} \propto \alpha V^2 f \qquad P_{leakage} \propto e^{-\frac{1}{T}}$$

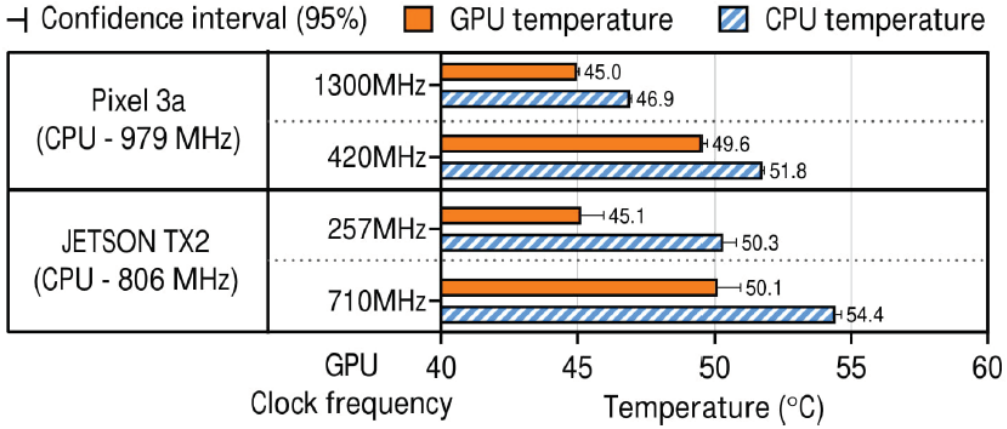
- α: Core utilization
- V: Voltage
- f: Frequency
- T: Temperature



- **Difficult to predict future temperature**
  - Environmental changes
  - Thermal coupling among processors
  - Application-dependent



CPU temperature according to application when CPU/GPU clock is fixed



CPU temperature increases when the GPU clock frequency increases while the CPU clock fixed due to thermal coupling

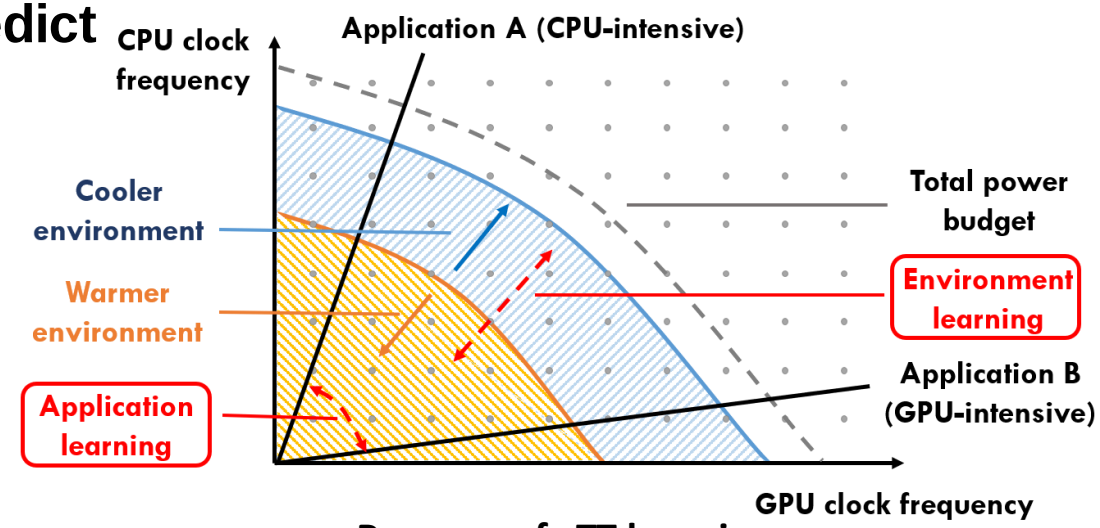
AI for System – zTT (ACM Mobisys'21)

- **Summary of research challenges**

① Difficult to predict application performance ② Difficult to predict future temperature

➡ **Unlike desktops and servers, there is no one-size-fits-all approach !**

- **Deep reinforcement learning (DRL)-based hybrid frequency scaling**
  - Using the history of control, we aims to predict
    - ① Application performance
    - ② Thermal budget (headroom)
  - Design a novel reward function
  - Aims to adapt quickly to changing thermal environment



Purpose of zTT learning

AI for System – zTT (ACM Mobisys'21)

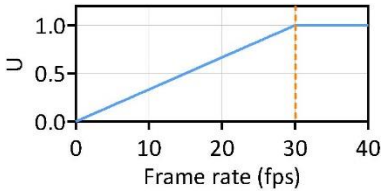
- **System Design**

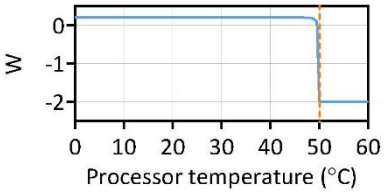
Markov Decision Process (MDP)

- **State:** $\mathbf{s(t)} \in \boldsymbol{S}$ $\quad \mathbf{s}(t) = (f_c(t), f_g(t), T_c(t), T_G(t), P_C(t), P_G(t), \mathbf{x(t)})$
  - CPU/GPU Clock frequency
  - CPU/GPU Temperature
  - CPU/GPU Power consumption
  - Frame rate
- **Action:** $\mathbf{a(t)} \in \boldsymbol{A}$ $\quad \mathbf{a}(t) = (f_c(t), f_g(t))$
- **Reward:** $\mathbf{r(t)} \in \boldsymbol{R}$ $\quad \mathbf{r}(t) = U(t) + \frac{\beta}{P(t)} + W(t)$

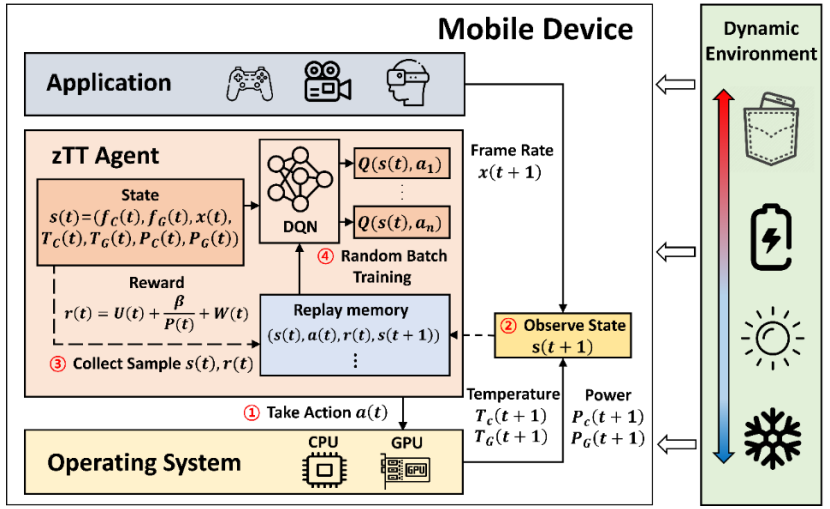
$* \ \beta: Trade-off\ term$

Utility function (User QoE) | Thermal constraints





- **DRL-based algorithm**
  - Maximizing Q-function $Q(s, a)$?
    - Maximizing $U(t)+\beta/P(t)$: QoE maximization with power efficiency
    - Maximizing $W(t)$: Zero thermal throttling at steady-state (long-term)



AI for System – zTT (ACM Mobisys'21)

- **Evaluation results**
  - **Experiment setup**

Device	JETSON TX2	Google Pixel 3a
CPU	NVIDIA Denver2 + ARM Cortex-A57	ARM Cortex-a55(LITTLE)+cortex-a75(big)
GPU	NVIDIA PASCAL GPU	Adreno 615
Memory	8GB DDR4	4GB LPDDR4X
OS	Ubuntu 16.04	Android 9.0 Pie

Experimented devices

Experimented apps

Application	Description	Device
Aquarium	WebGL-based 3D object rendering	JETSON TX2
YOLOv3	Deep learning-based object detection	JETSON TX2
Video rendering	Rendering a video with OPENCV2	JETSON TX2
Showroom VR	WebGL-based 3D object rendering	Pixel 3a
Skype	Video call	Pixel 3a
Call of duty 4	3D Mobile game	Pixel 3a

Aquarium

YOLO

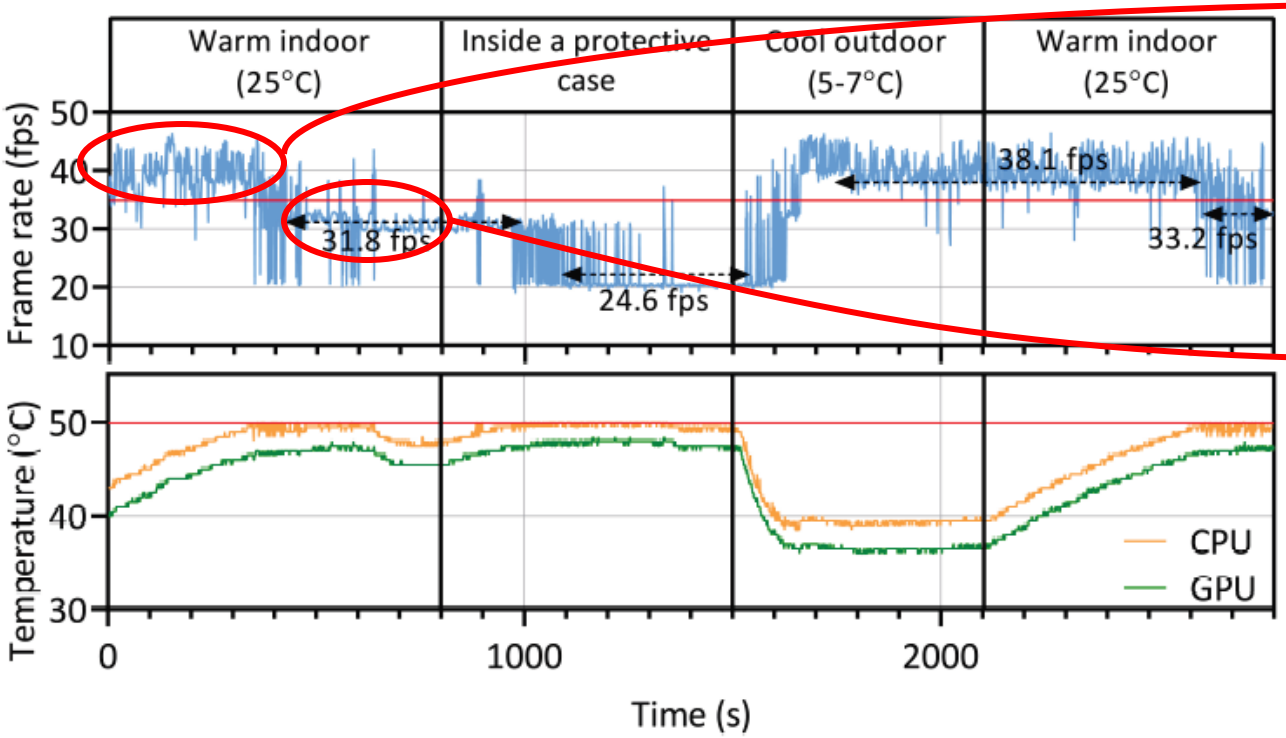
Showroom VR

Skype

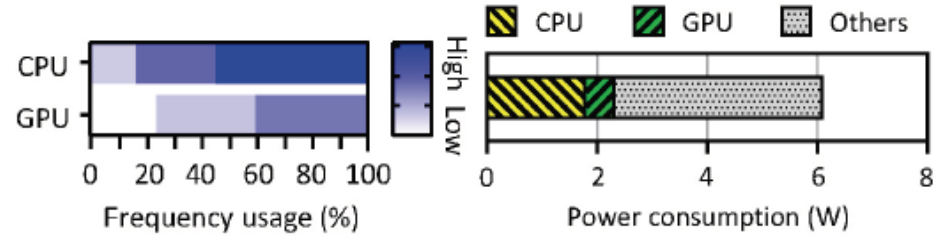
Call of duty4

AI for System – zTT (ACM Mobisys'21)

- **Evaluation results**

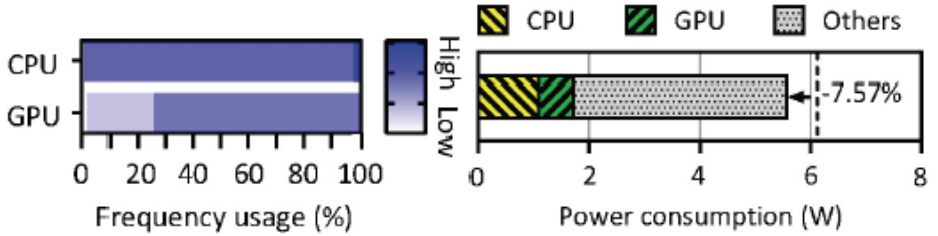


Frame rate and temperature of JETSON TX2 rendering a video while experiencing a number of environmental changes.



Before running the environment

7.5 % Total power saving



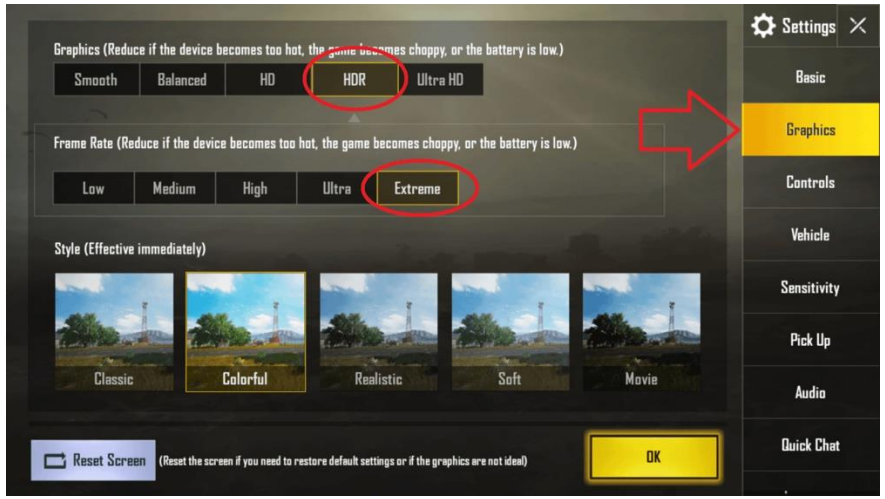
After running the environment

* Amrouch, Hussam, et al. "Npu thermal management." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 39.11(2020): 3842-3855.

AI for System – zTT (ACM Mobisys'21)

- **Key takeaway**
  - **AI can even control the OS and outperform conventional control methods**

- **Implementing zTT in practice**



PUBG Mobile

Fortnite Mobile

* https://thegameroof.com/how-to-high-fps-and-max-graphics-in-pubg-mobile-android/
* https://gamingonphone.com/guides/fortnite-mobile-best-settings-and-hud-layout-guide/

System for AI

CoActo: CoActive Neural Network Inference Offloading with Fine-grained and Concurrent Execution.

(ACM MobiSys'24)

System for AI – CoActo (ACM Mobisys'24)

- **AI-based mobile services**
  - Acceptable latency budget for human-machine interaction: <100 ms

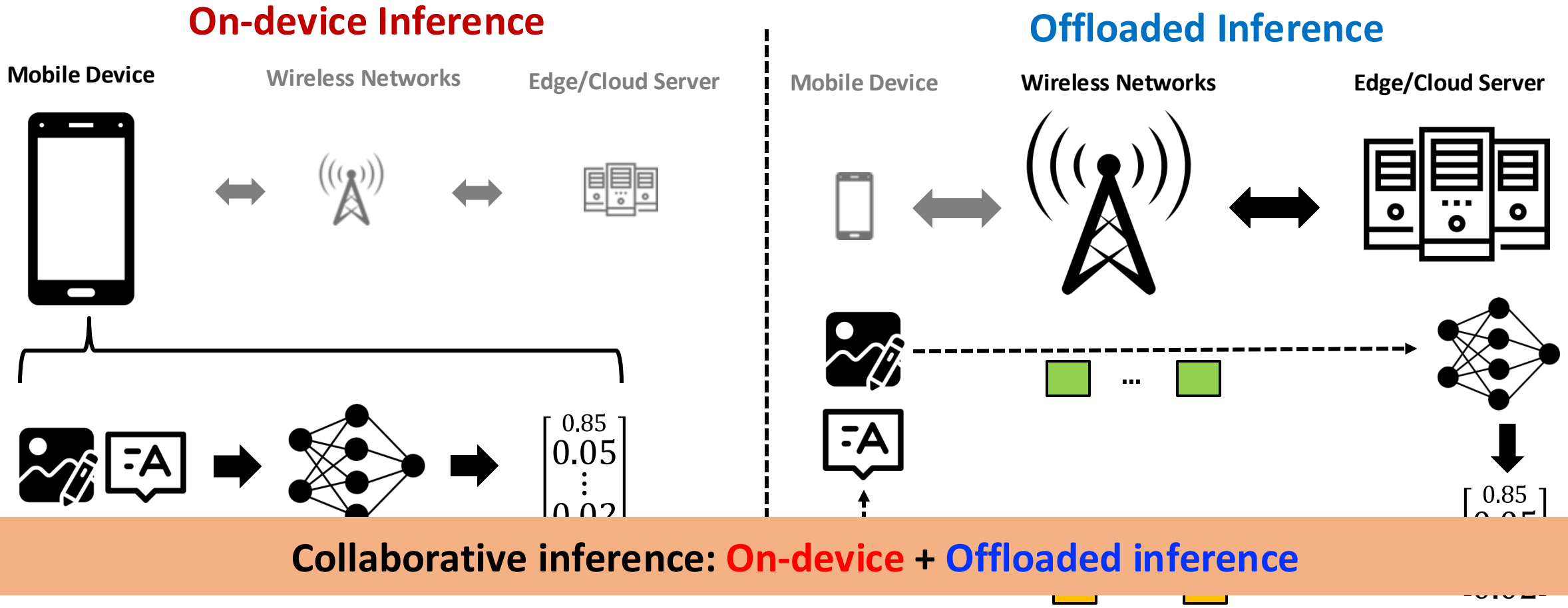
Miller, R. B. (1968). Response time in man-computer conversational transactions. Proc. AFIPS Fall Joint Computer Conference Vol. 33, 267-277.

ChatGPT

Google Assistant

System for AI – CoActo (ACM Mobisys'24)

- **How to enable Neural Network (NN) inference on Mobile devices?**



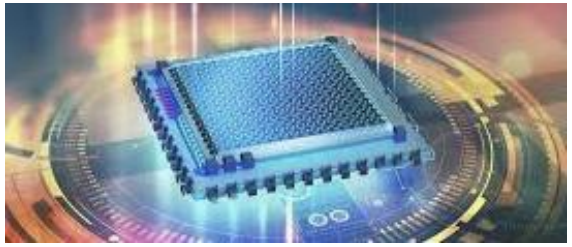
Collaborative inference: On-device + Offloaded inference

System for AI – CoActo (ACM Mobisys'24)

- **LLM with Meta Orion**

[Offloading]
Large Language Model (LLM)

Real-time translation

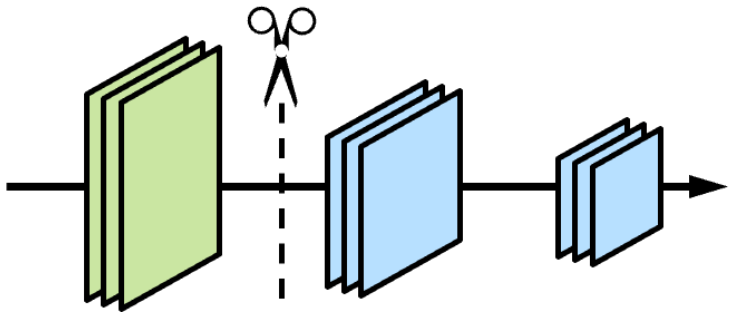


[On-device]
Small Language Model (SLM)

System for AI – CoActo (ACM Mobisys'24)

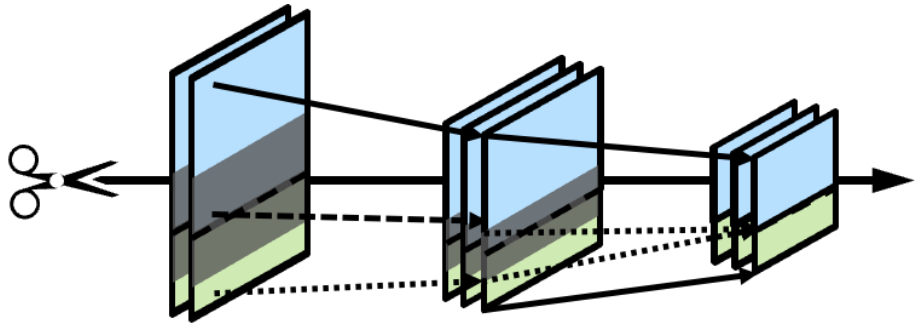
- **Mobile-Server collaborative inference**

Split computing (Vertical)



- **SPINN [Mobicom'20]**
- **Sequential execution on the mobile and the server**

Fused-Layer (FL) offloading (Horizontal)

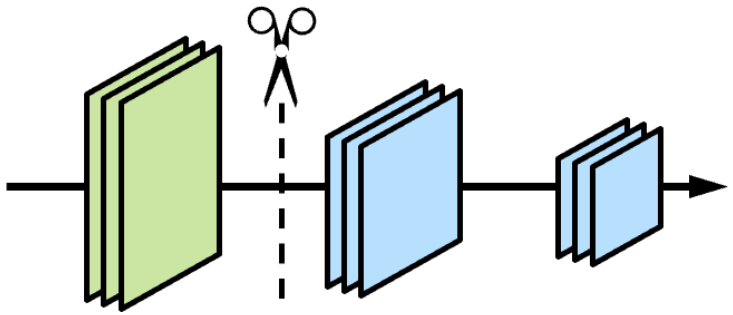


- **Partial Offloading [IEEE TPDS'23]**
- **Parallel execution on the mobile and the server**

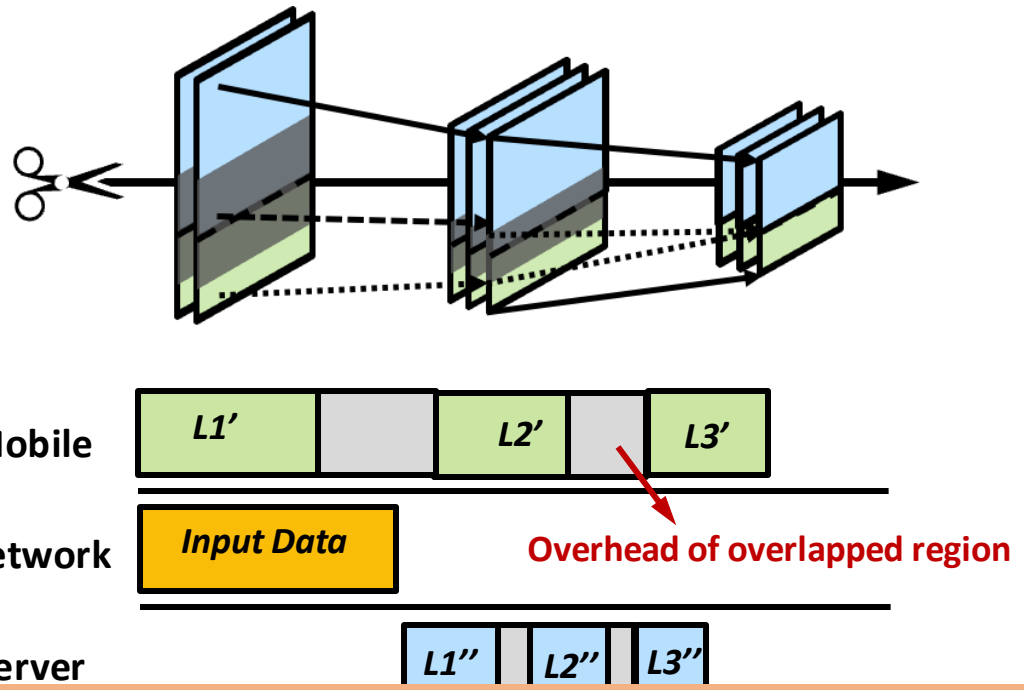
System for AI – CoActo (ACM Mobisys'24)

- **Mobile-Server collaborative inference**

Split computing (Vertical)



Fused-Layer (FL) offloading (Horizontal)

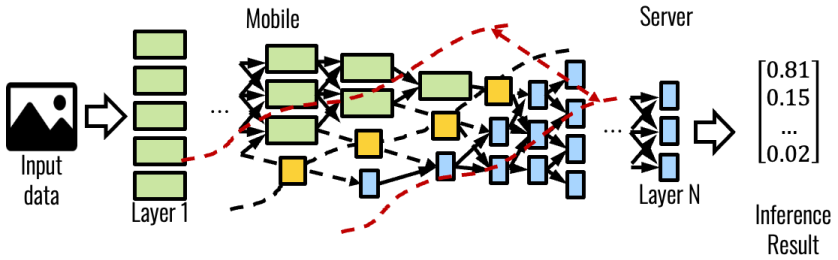


How can we further improve the performance ? Optimal pipelining

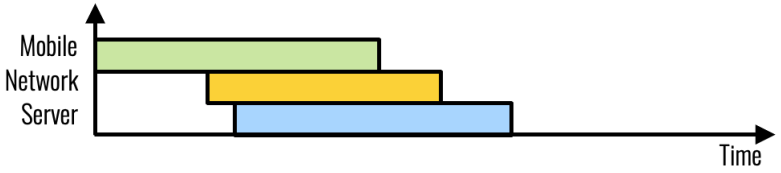
System for AI – CoActo (ACM Mobisys'24)

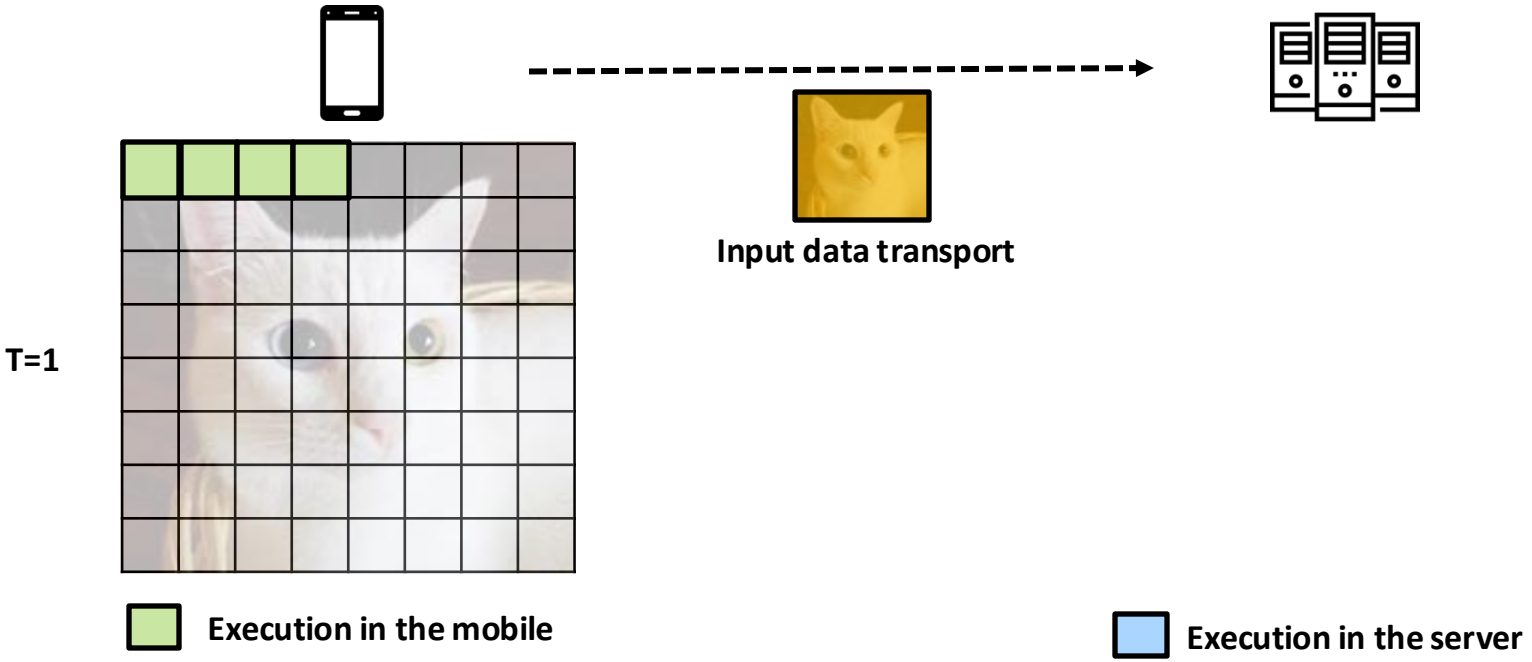
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

1) Fine-grained DNN expression



2) Concurrency of runtime resources

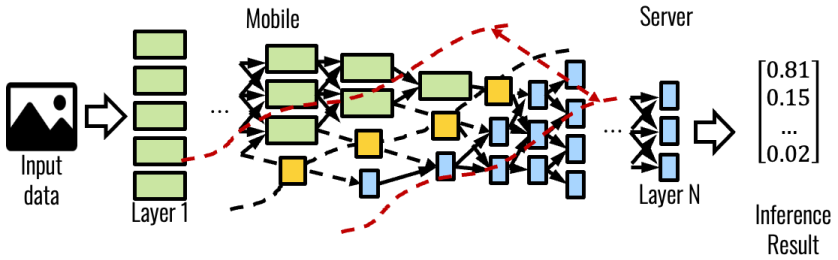




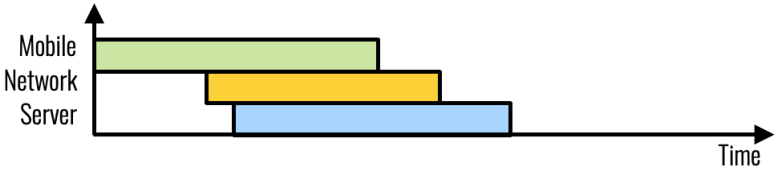
System for AI – CoActo (ACM Mobisys'24)

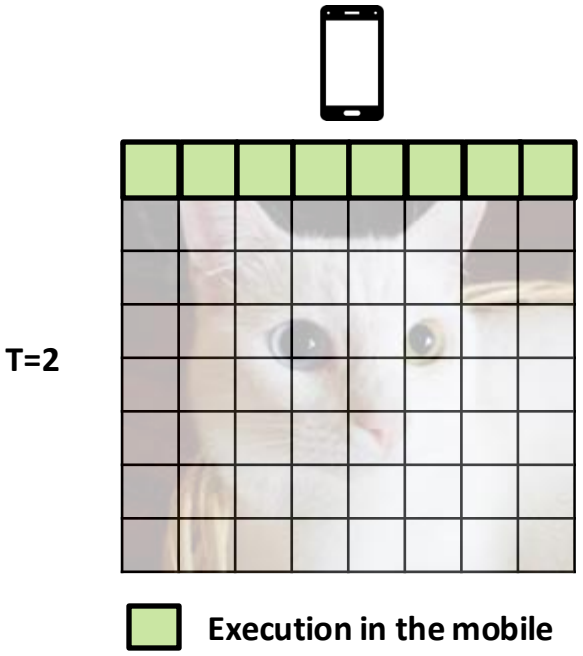
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

1) Fine-grained DNN expression



2) Concurrency of runtime resources

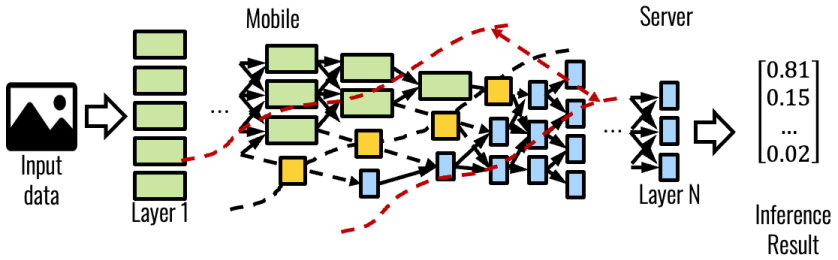




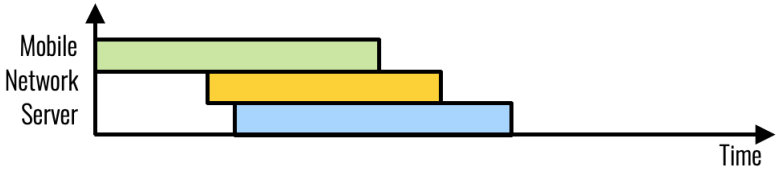
System for AI – CoActo (ACM Mobisys'24)

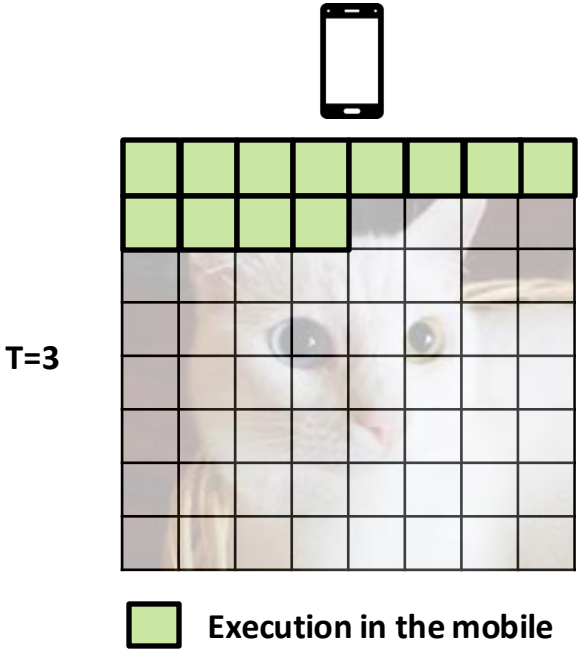
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

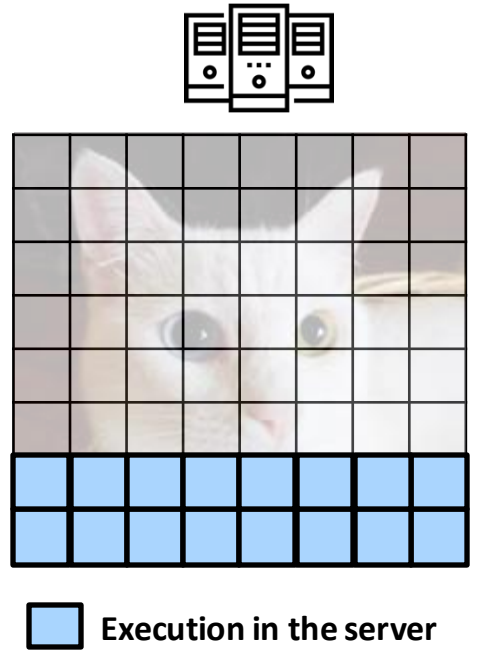
1) Fine-grained DNN expression



2) Concurrency of runtime resources



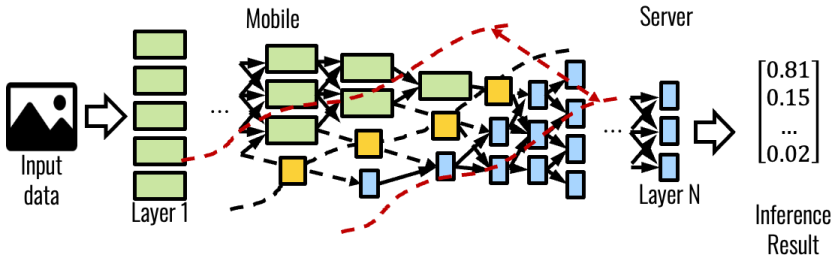




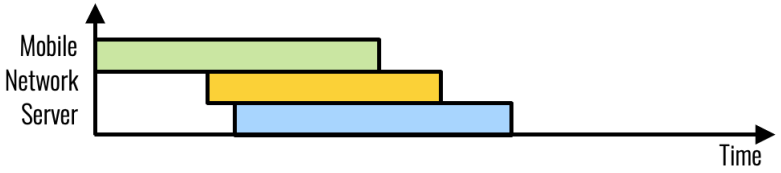
System for AI – CoActo (ACM Mobisys'24)

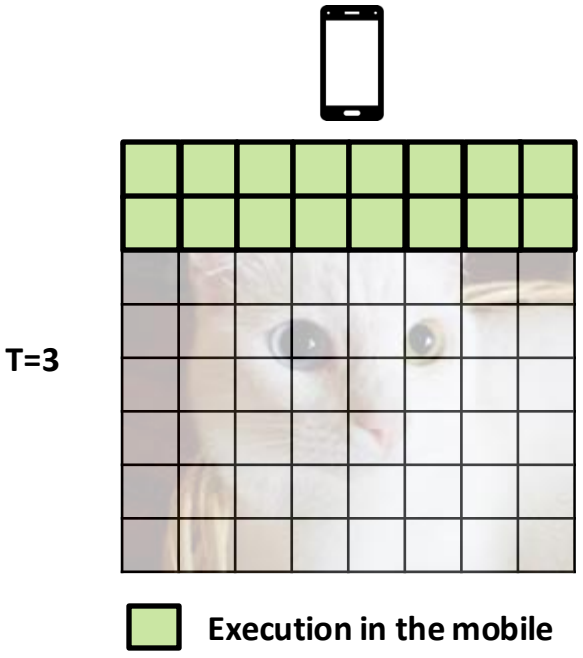
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

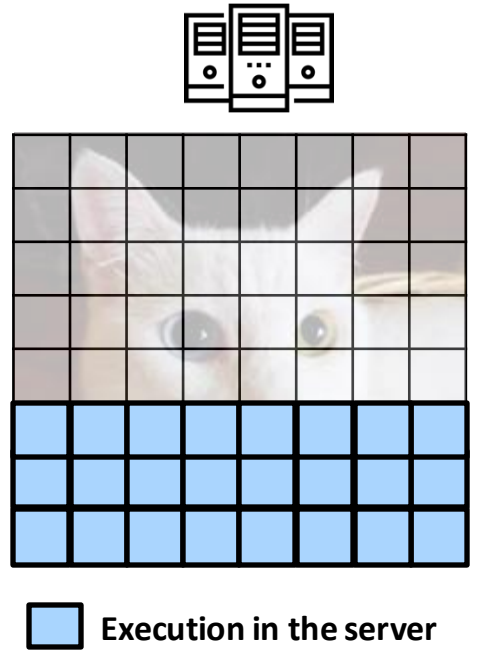
1) Fine-grained DNN expression



2) Concurrency of runtime resources



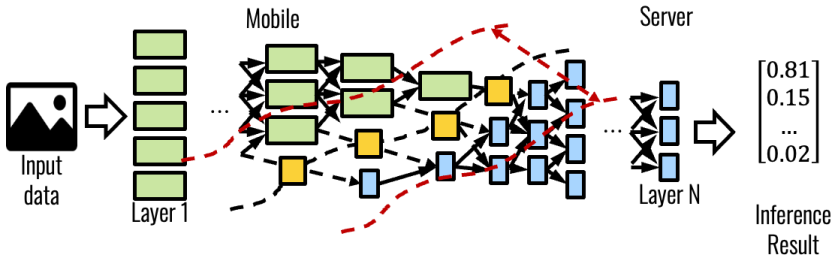




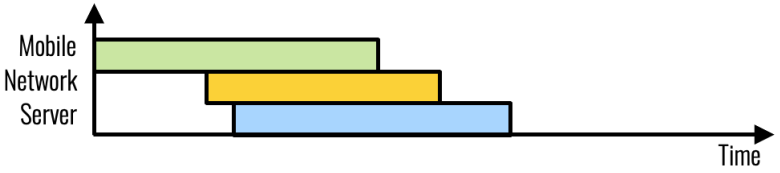
System for AI – CoActo (ACM Mobisys'24)

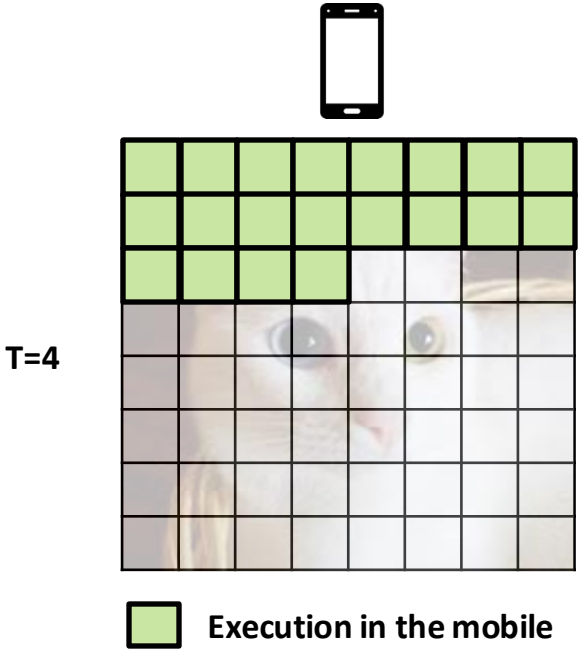
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

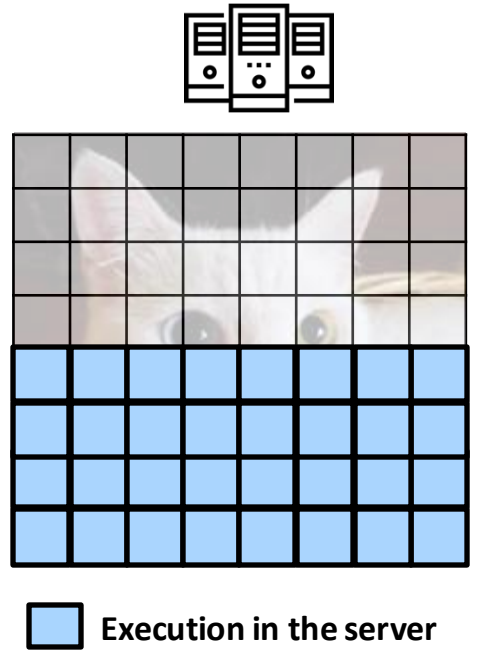
1) Fine-grained DNN expression



2) Concurrency of runtime resources



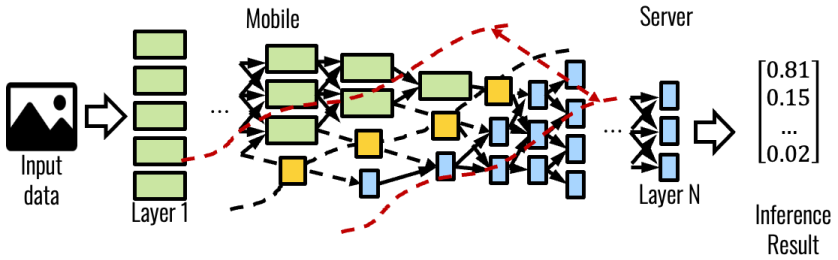




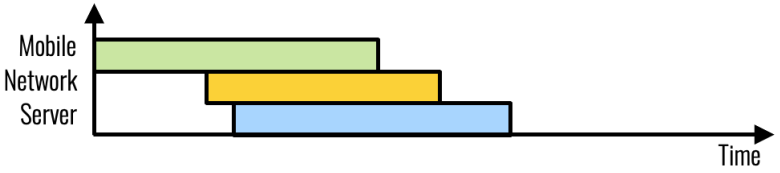
System for AI – CoActo (ACM Mobisys'24)

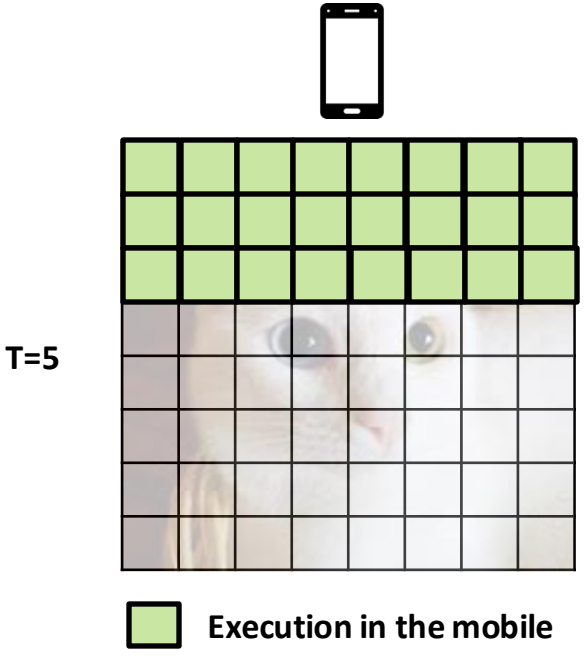
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

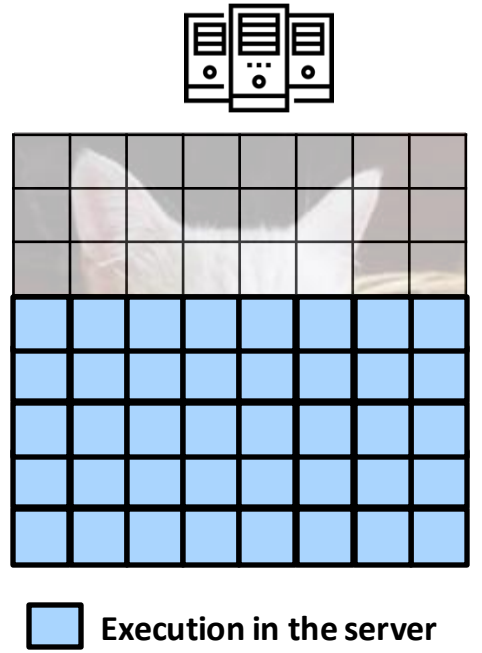
1) Fine-grained DNN expression



2) Concurrency of runtime resources



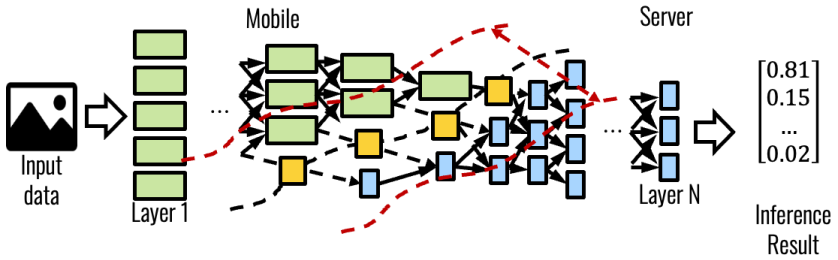




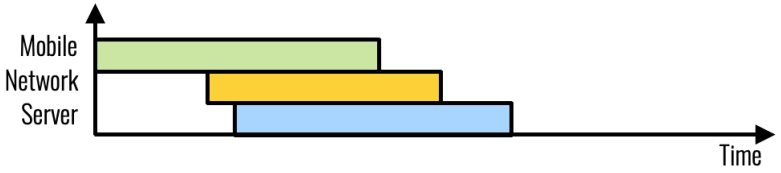
System for AI – CoActo (ACM Mobisys'24)

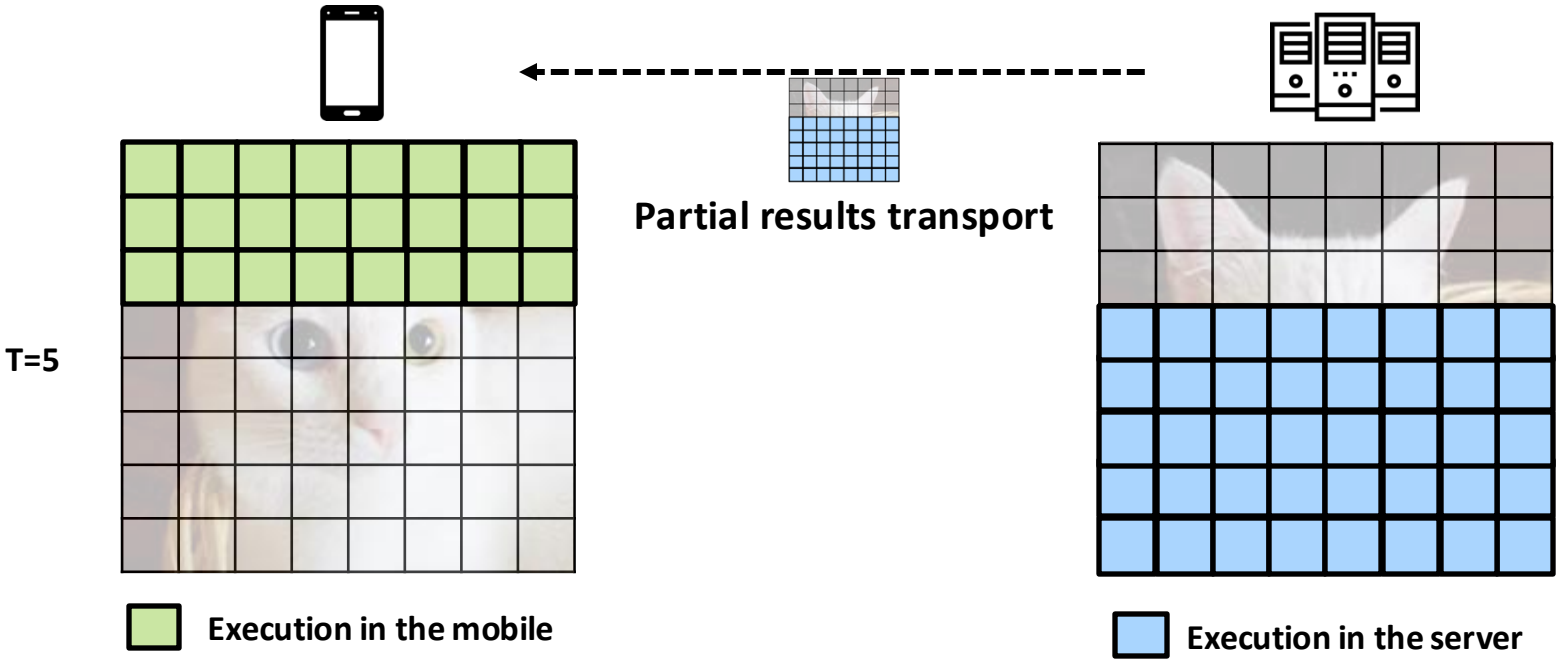
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

1) Fine-grained DNN expression



2) Concurrency of runtime resources

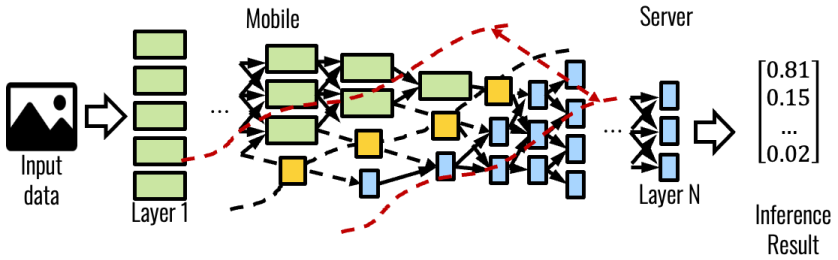




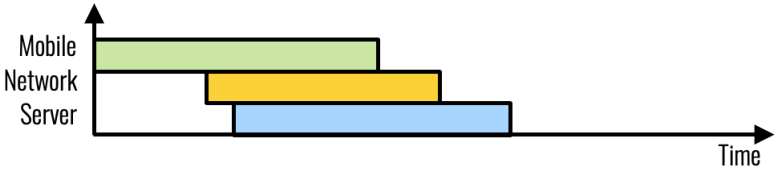
System for AI – CoActo (ACM Mobisys'24)

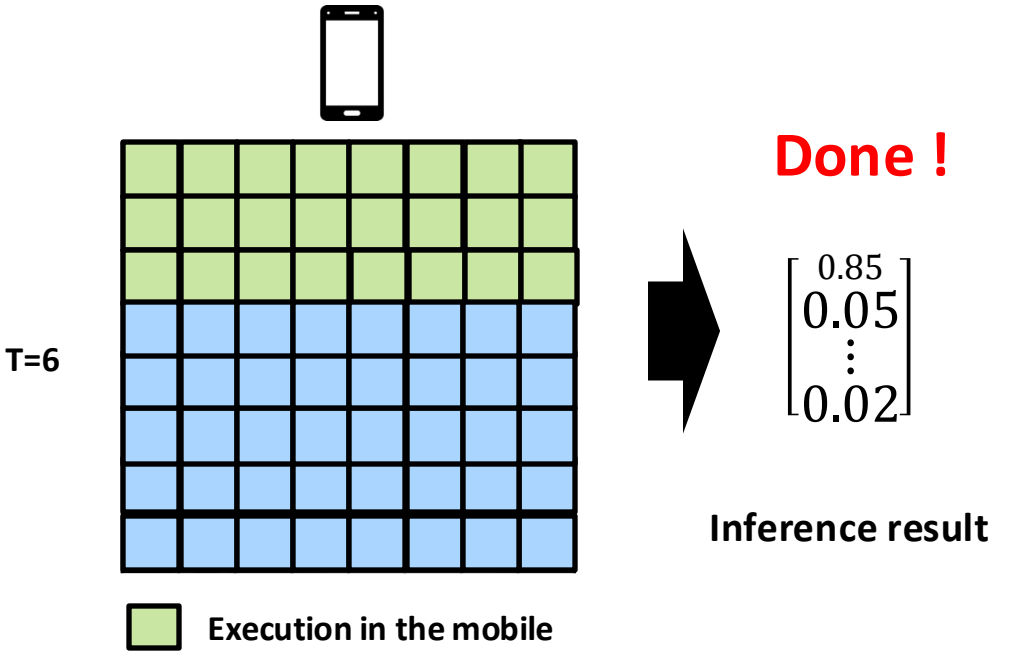
- **Our Approach: Coactive Neural Network Inference Offloading**
- **Key Design Philosophy**
  - **Minimizes idle time of mobile, network and the server.**

1) Fine-grained DNN expression



2) Concurrency of runtime resources



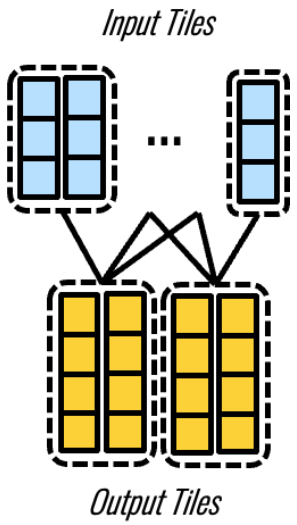


System for AI – CoActo (ACM Mobisys'24)

Overview of CoActo

1) Tile-based Partitioner (TP)

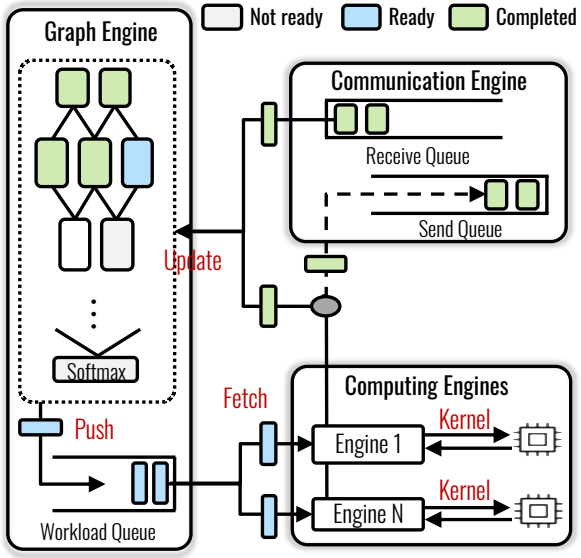
- Automatically partitions a layer-wise graph into a tile-wise graph



Tiling & Graph reconstruction

2) Asynchronous Engines (AEEs)

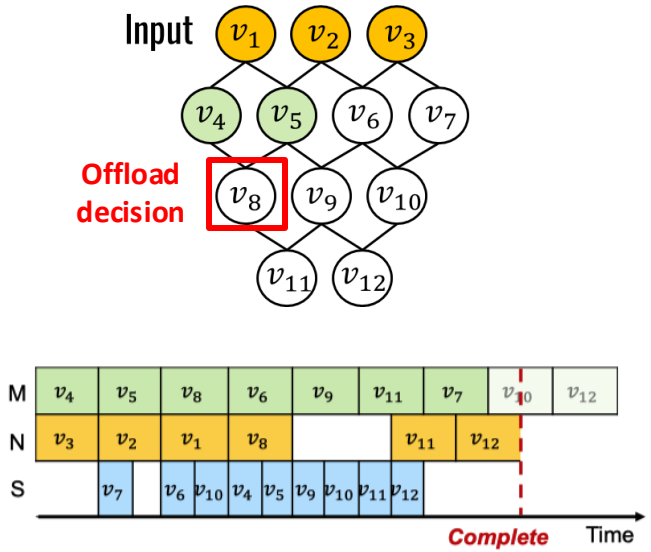
- Separate the data dependency managements and executions



Manages the workload queue for asynchronous operation

3) Dynamic Scheduler (DS)

- Efficient dynamic and adaptive tile scheduling



Schedules offloading decision

System for AI – CoActo (ACM Mobisys'24)

- **Evaluation: Experiments setup**

Jetson AGX Xavier
8 CPU Cores

Raspberry Pi 4
4 CPU Cores

Google Pixel 5
8 CPU Cores

WiFi network
(Up to 100 Mbps)

AMD Threadripper 3990X
64 CPU Cores

Tested models

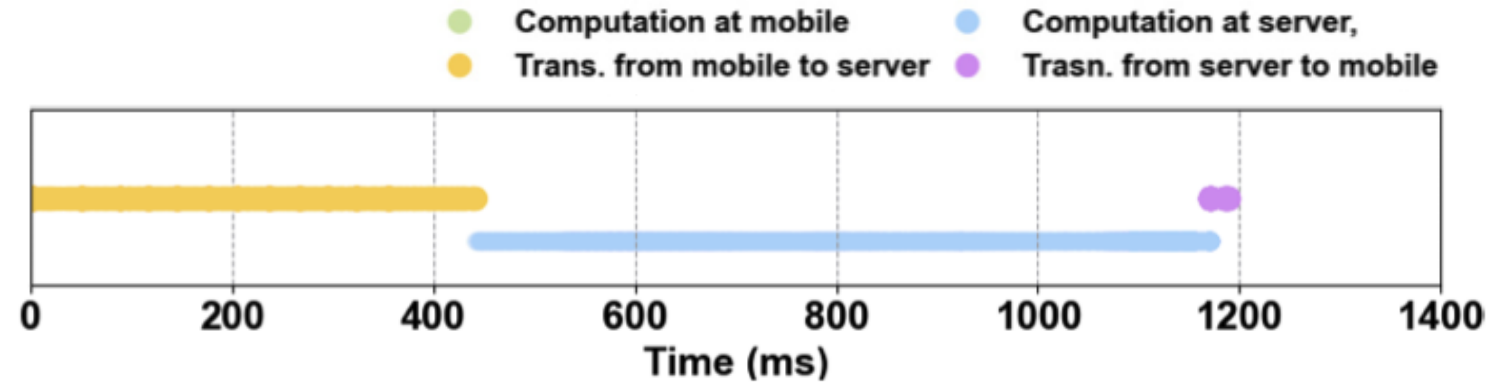
- **3 CNN models**: VGG16, ResNet50, YOLOv3
- **1 Transformer model**: BERT-base

Baselines

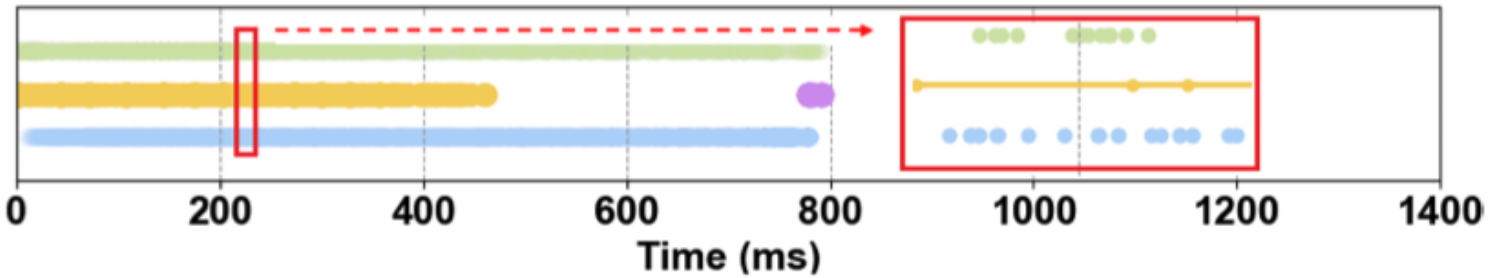
- Cloud-only
- On-device
- SPINN [Mobicom'20] (SOTA split computing)
- Fused-Layer (FL) offloading

System for AI – CoActo (ACM Mobisys'24)

- **Evaluation results**
  - **Timelines of each tile with VGG16 - Batch size 8.**



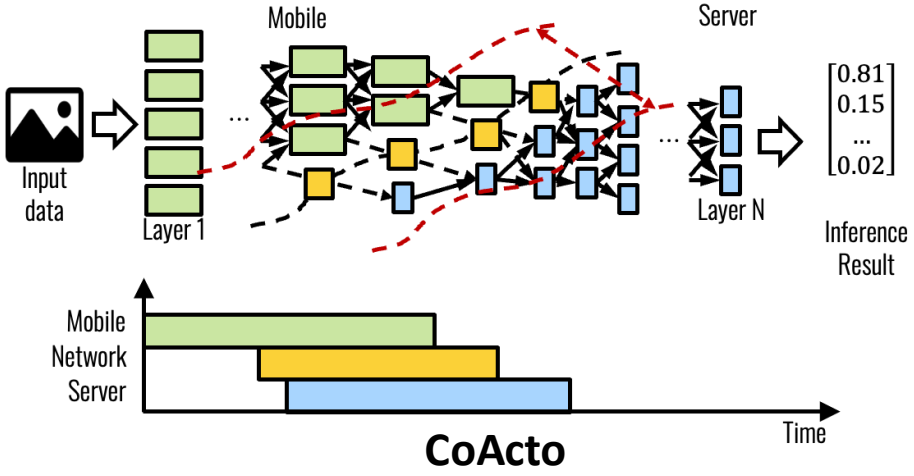
(1) SPINN [ACM Mobicom'20]

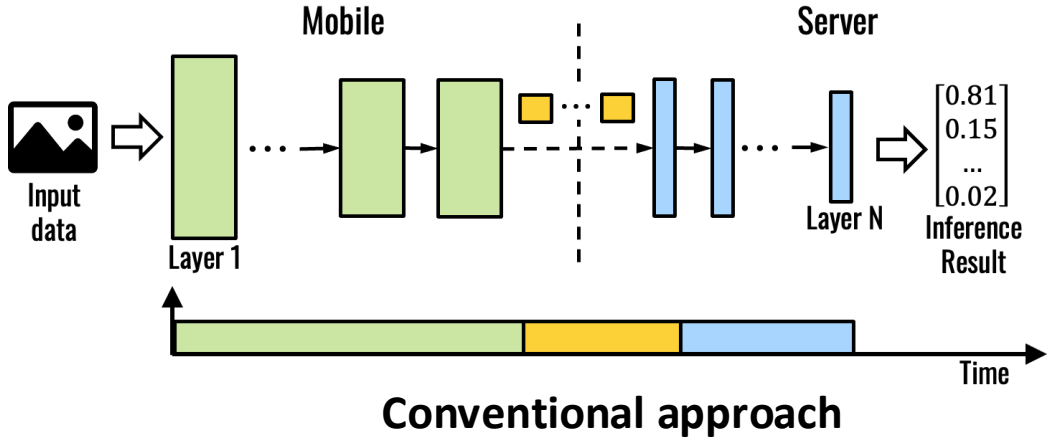


(b) CoActo

System for AI – CoActo (ACM Mobisys'24)

- **Conclusion**
  - **CoActo achieves latency reduction by overlapping computing and communication times.**





- **Key takeaway**
  - **Co-managing the networking and computing can be a new solution in emerging AI apps and new devices.**

Ongoing Research Projects

- **I'm leading 5 projects**
  - **6 Ph.D, 3 Master, 1 Undergraduate student (Co-advising)**
  - **+1 Ph.D. student, 2 Master students (Co-advising, Korea University)**

Google

LLM on CXL (Compute Express Link) Memory Controller


Scaling SFU (Selective Forwarding Unit) with SmartNICs


Real-time Volumetric Video Streaming


FM (Foundation model)-based Biosensing


5G+OpenRan and Low-Latency Networking

- **Other projects in collaboration with:**

SAMSUNG

RL-based Time-aware Cloud Scheduler (Samsung in-Compony Cloud)

HKUST

TEXAS STATE UNIVERSITY 1899

WiFi Localization with Fingerprints Augmentation

CableLabs®

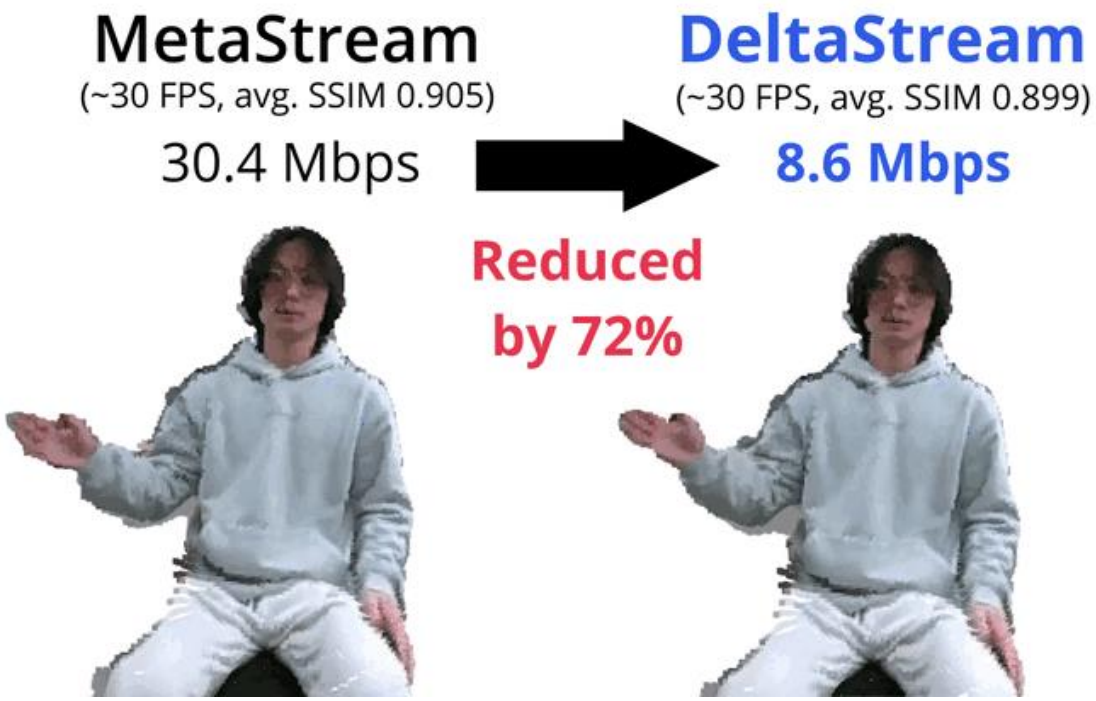
Research on ECN (Explicit Congestion Notification, TCP)

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Crowdsourced RSRP for Accurate Measurements with Mobile Phones and SDRs

Ongoing Research Projects

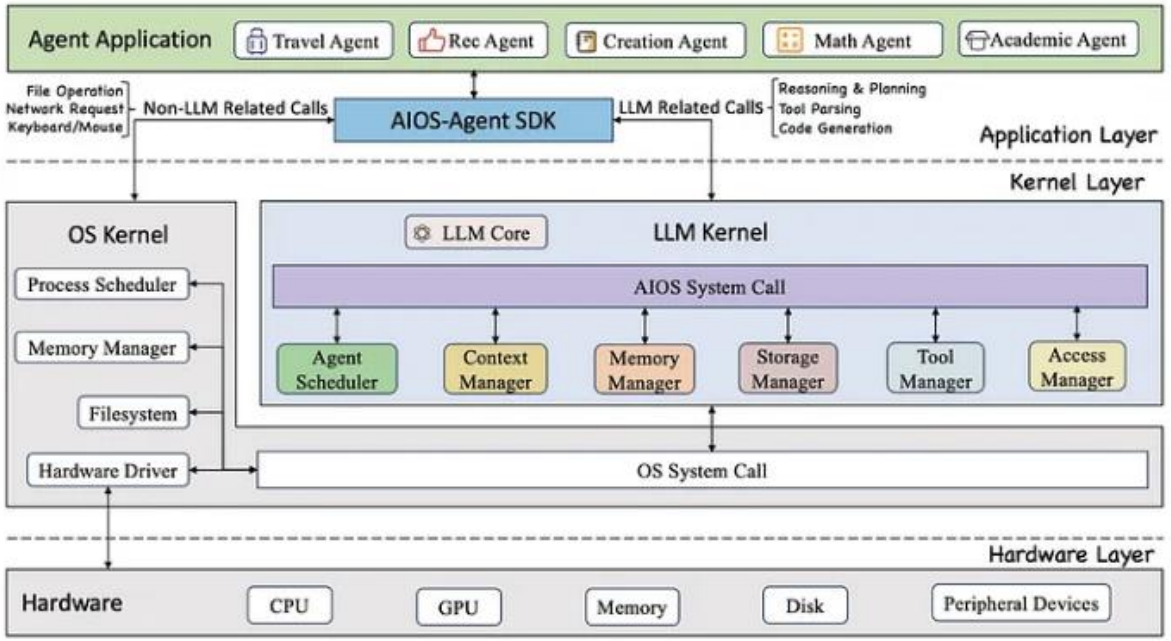
- **Real-time Volumetric Video Streaming (Accepted, ACM MobiSys'25)**



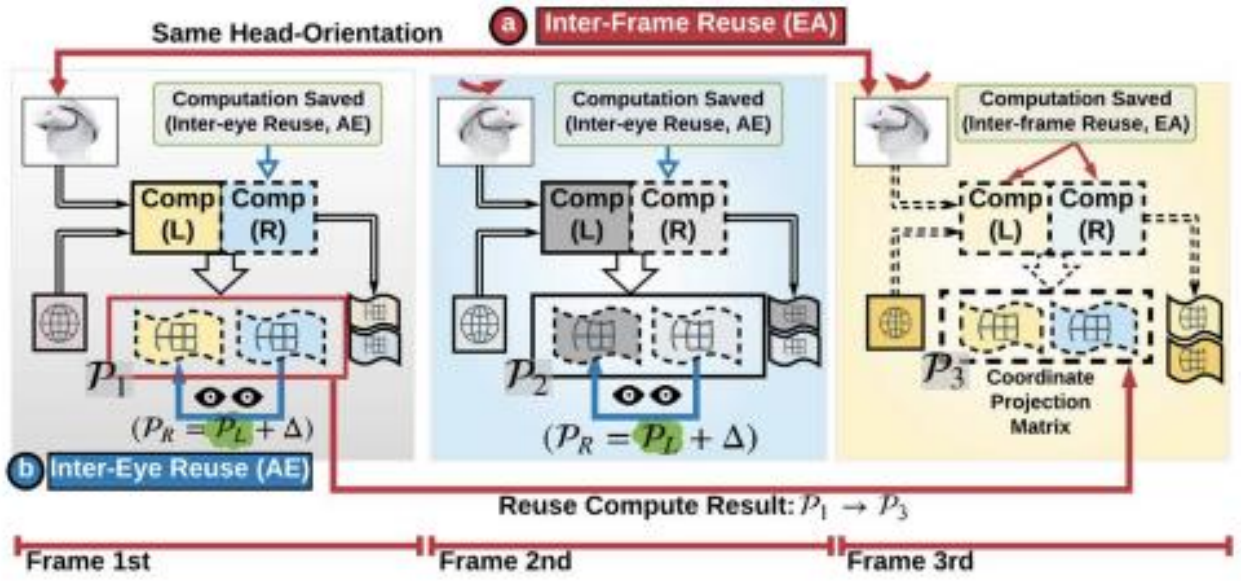
- **3D Avatar Streaming for 3D Video Call**

Ongoing Research Projects

- **LLM-based Operating System**



- **System Optimization for Rendering pipeline on AR Glass**



Q&A

Thank you !