Processing-in-Flash: Accelerating Training for On-Device Machine Learning

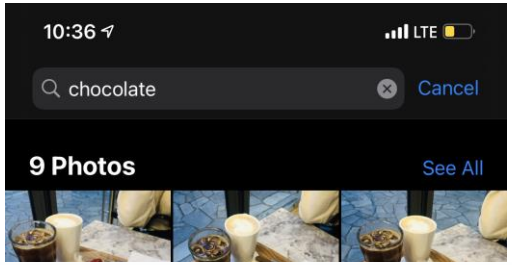
Yoona Kim †, Dusol Lee †, Geonhee Cho †

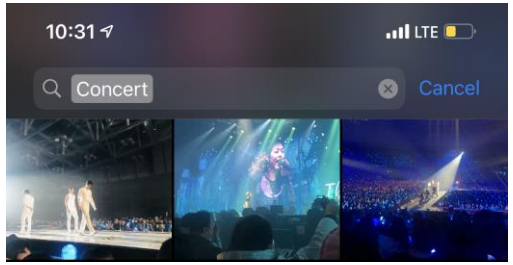
† *Team 3 (SysMo)*

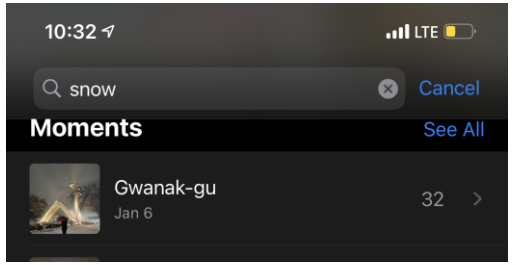
Contents

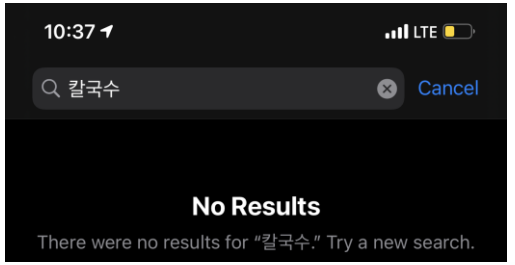
- Motivation & Problem
- Related Works
- Key Idea & System Overview
- Expected Challenges
- Evaluation Strategy
- Overall Plan
- Deliverable

We Expect User-Definable Machine Learning Services !



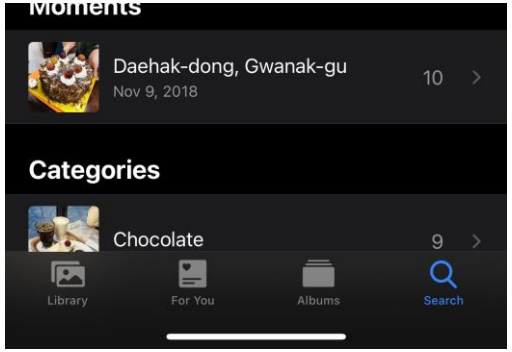


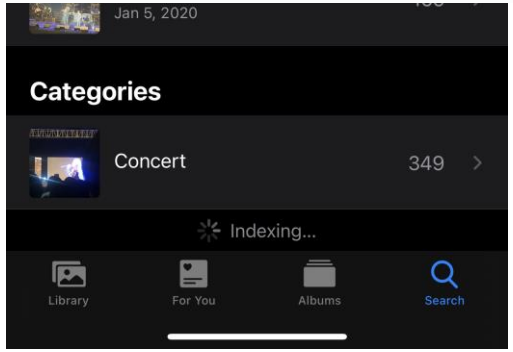


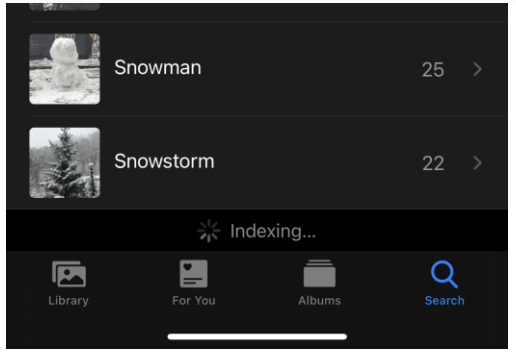


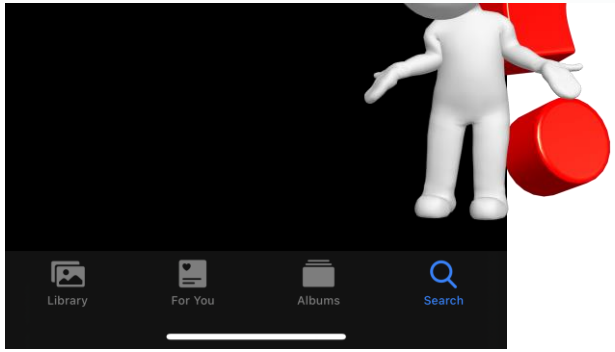
What if a user wants to add a new category… ?

The system needs to read the whole data
to re-train in order to add a new category





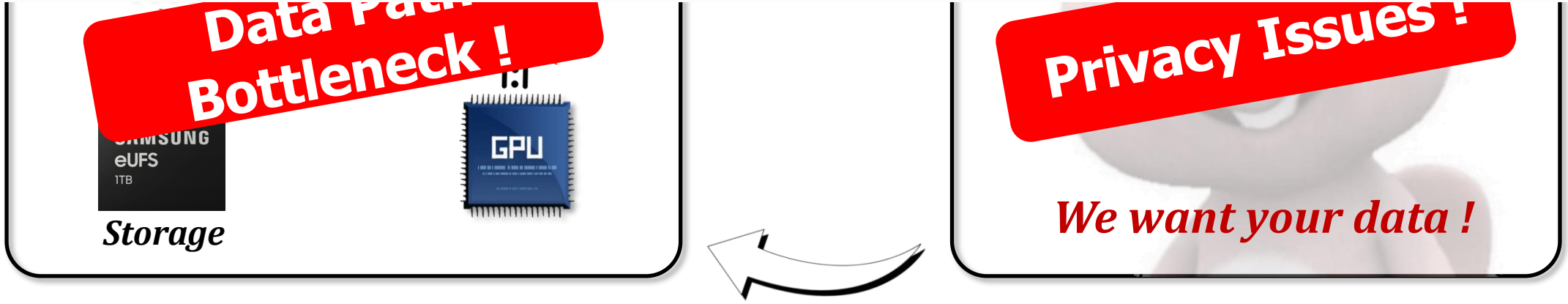




Could It be Deployed w/o Any Problems ?

- **Training on user device ?**
  - **On-device training takes too long !**
- **Offloading training to server ?**

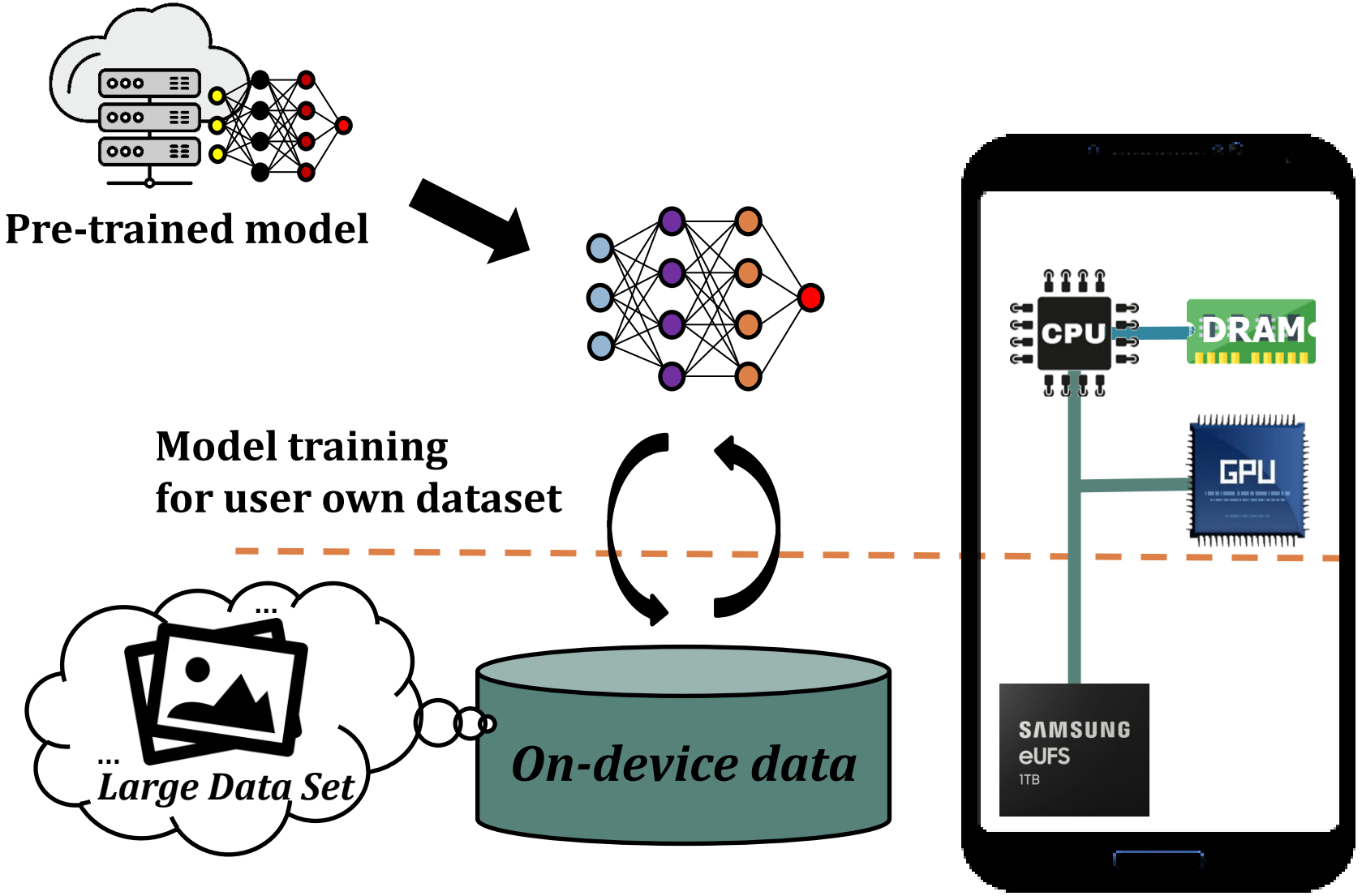
By solving the bottleneck of on-device,
We don't need to get help from server computing.



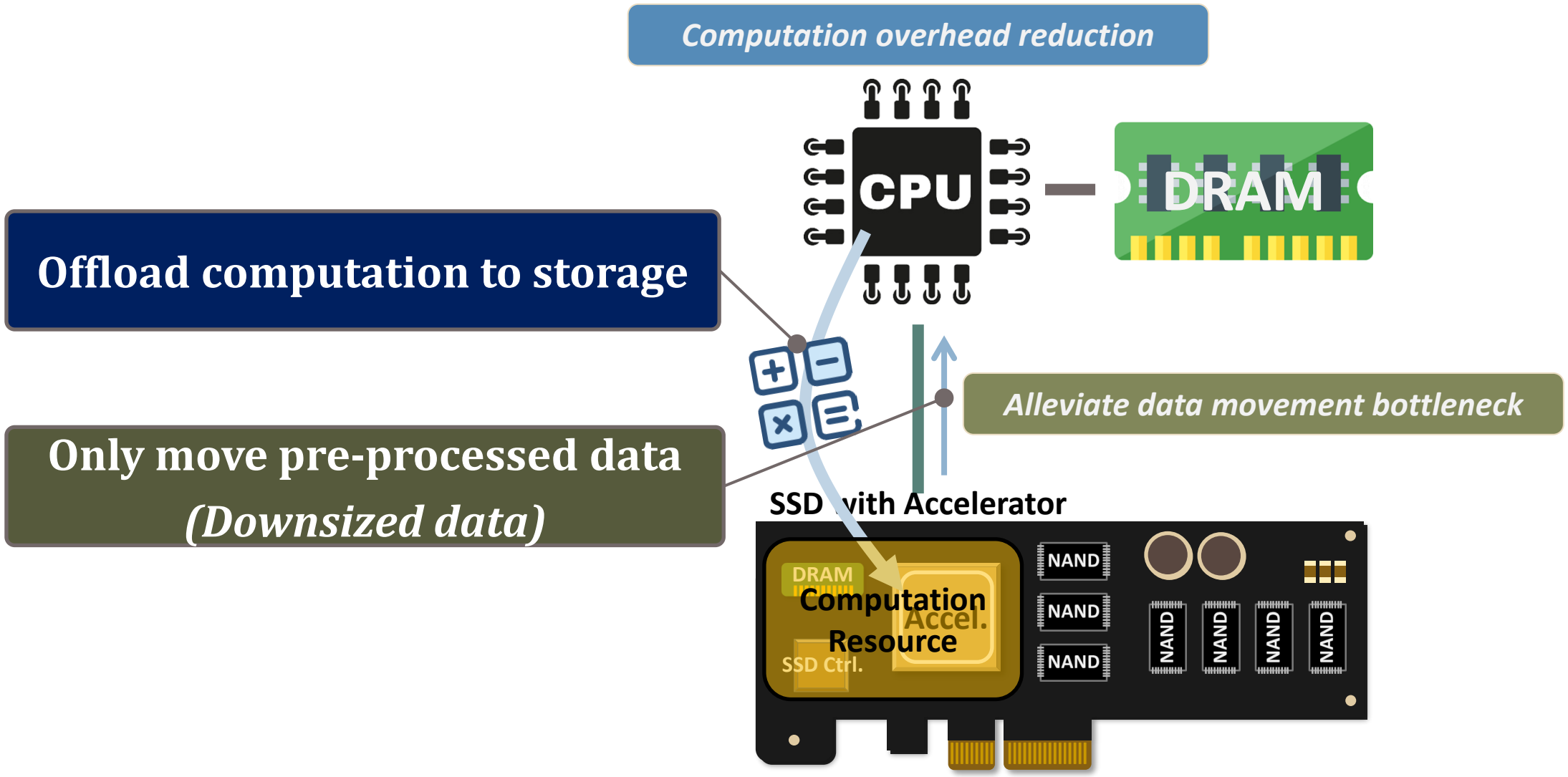
Related Works: How to Solve Data Path Bottleneck ?

- **Mobile Environments**
  - On-Device Machine Learning
- **Server Environments**
  - Processing-in-Storage

On-Device Machine Learning



Processing-in-Storage (PiS)



Can We Use On-Device Processing-in-Storage ?

Server-Class NAND Flash *vs* ***Mobile-Class NAND Flash***

Enough Space for Large Accelerator

Large Memory on SSD

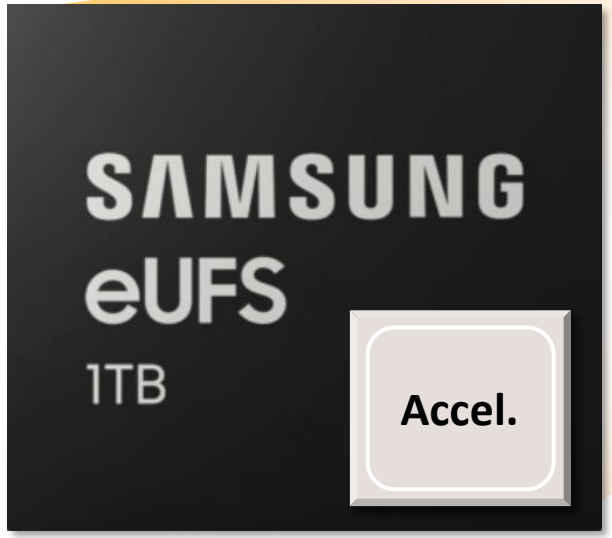
No Space for Large Accelerator

Small Memory on Embedded Flash Chip

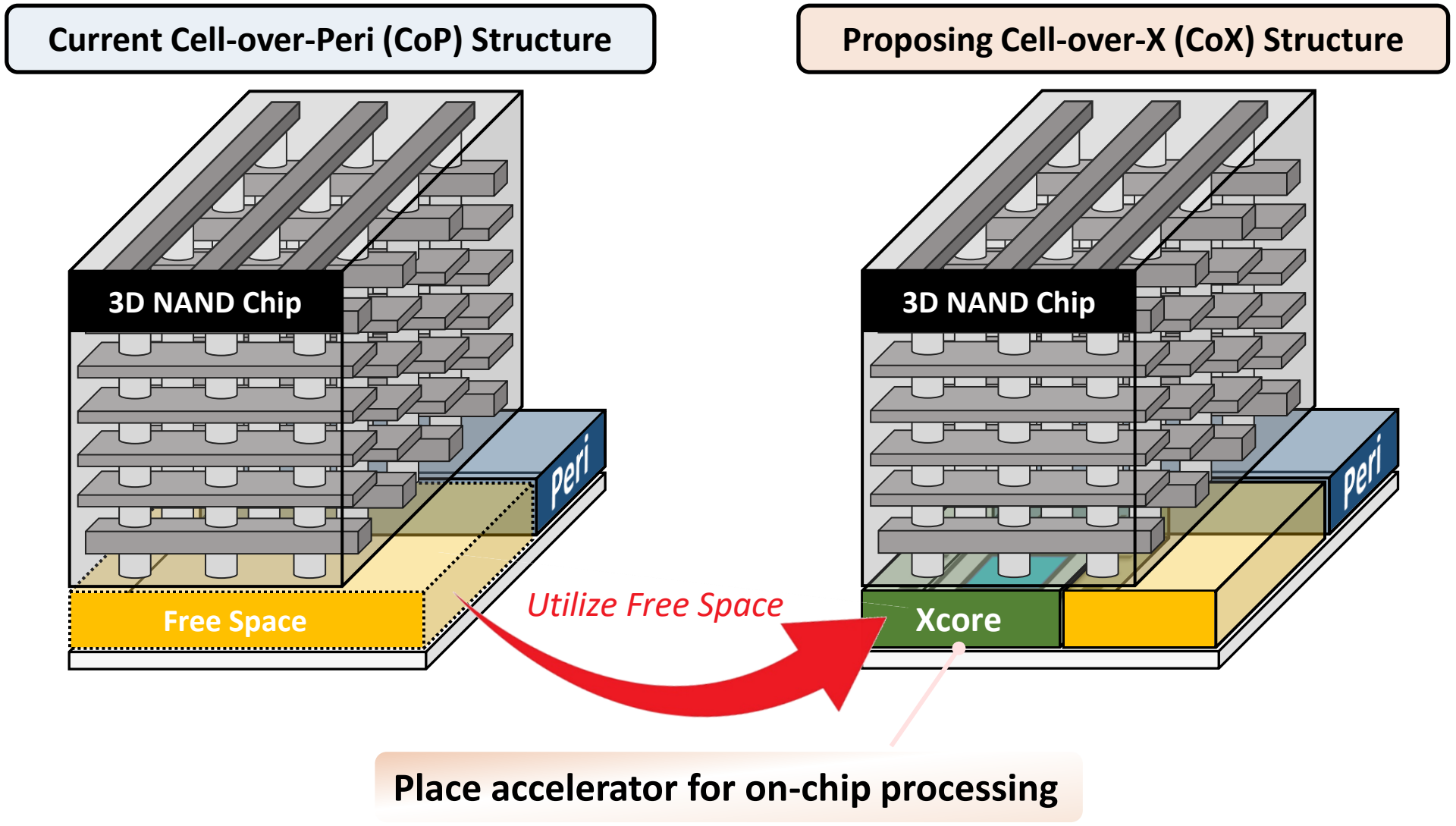
Key Idea & System Overview

➢ Processing-in-Flash (PiF)

- Same Idea w/ Processing-in-Storage !
- Can be used for mobile embedded flash chip !
- Accelerator is hide on Flash Chip !



Cell-over-X (CoX)

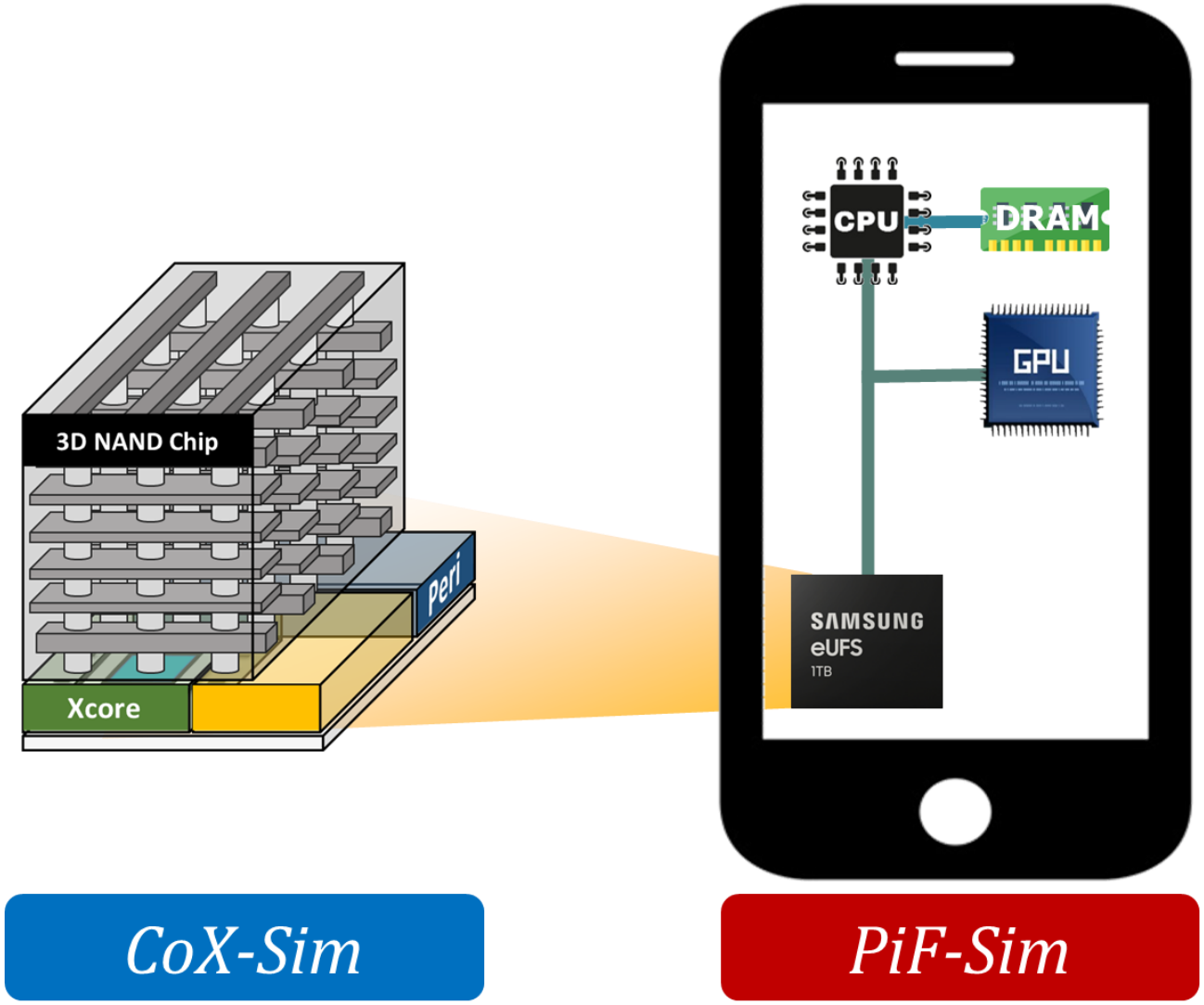


Expected Challenges

- **Physical Formulation Difficulty**
  - It is difficult to implement real NAND Flash chip
  - It is difficult to integrate the whole system on real device

Evaluation Strategy

- **Metric**
  - Scalability
  - Performance
  - Power Consumption
- **Benchmarks**
  - TBD



CoX-Sim

Design Space Exploration

PiF-Sim

System Evaluation

Overall Project Plan

#Iter.	*Objective*	*Duration*	*Misc.*
1	**Ideation & Proposal validation** Literature Review, Proposal Feedback	03/16 – 03/30	03/23 proposal (week 4)
2	**Build project environment & Design system** Build Env. for project, Design Arch. and techniques.	03/31 - 04/14	
3	**Implement Techniques & System** Implement techniques and Integrate all into system	04/15 – 05/25	05/04 demo (week 10)
4	**Evaluation** Evaluate system with training Algorithms	05/26 – 06/08	06/08 final (week 15)

Deliverable

Midterm Deliverable	*Final Deliverable*
Reasoning over project topic	*Technical Report regarding evaluations*
Design of CoX Flash chip (including tool code for design space exploration)	*Simulator (or Emulator) code for Proof-of-Concept*

Success Criteria

Performance(Processing-in-Flash) ≥ Performance (Traditional On-Device Processing)


Thank You !
Any questions or feedback are welcome.