



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

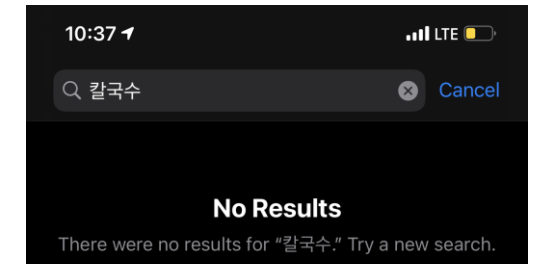
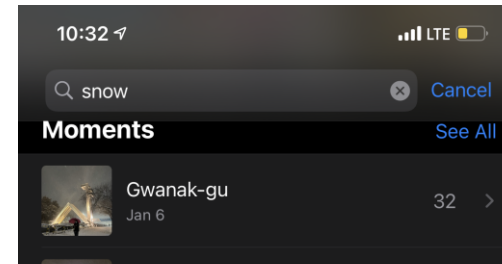
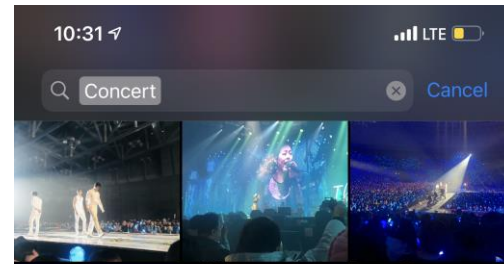
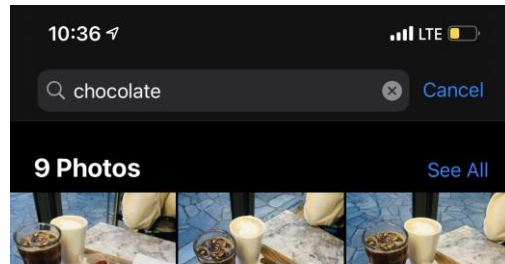
Yoona Kim †, Dusol Lee †, Geonhee Cho †  
† *Team 3 (SysMo)*



# Contents

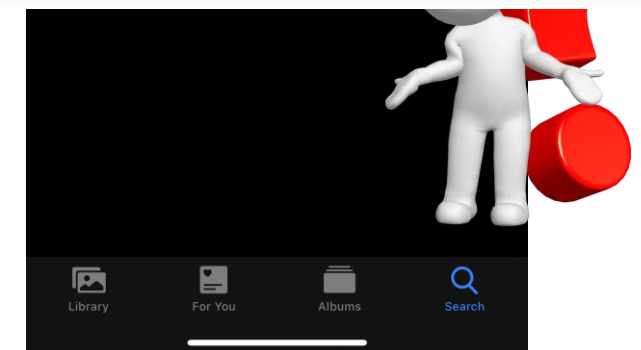
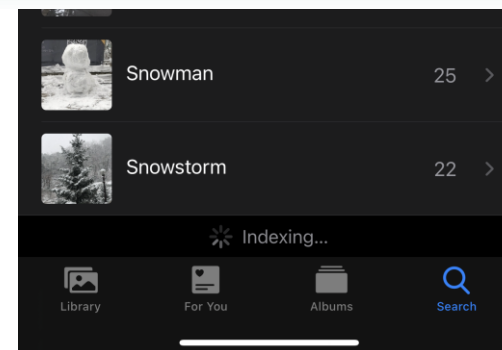
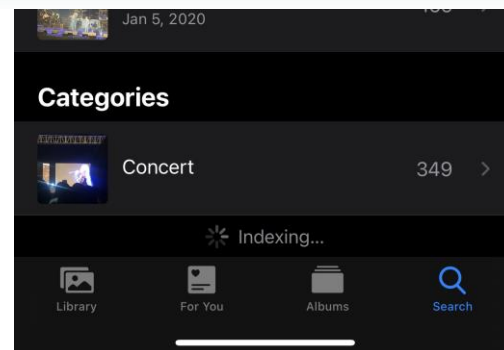
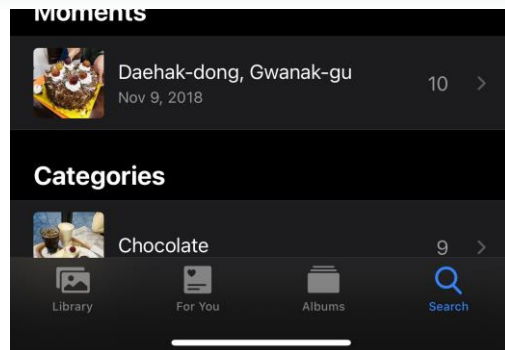
- **Motivation & Problem**
- **Related Works**
- **Key Idea & System Overview**
- **Expected Challenges**
- **Evaluation Strategy**
- **Overall Plan**
- **Deliverable**

# Motivation



*What if a user wants to add a new category... ?*

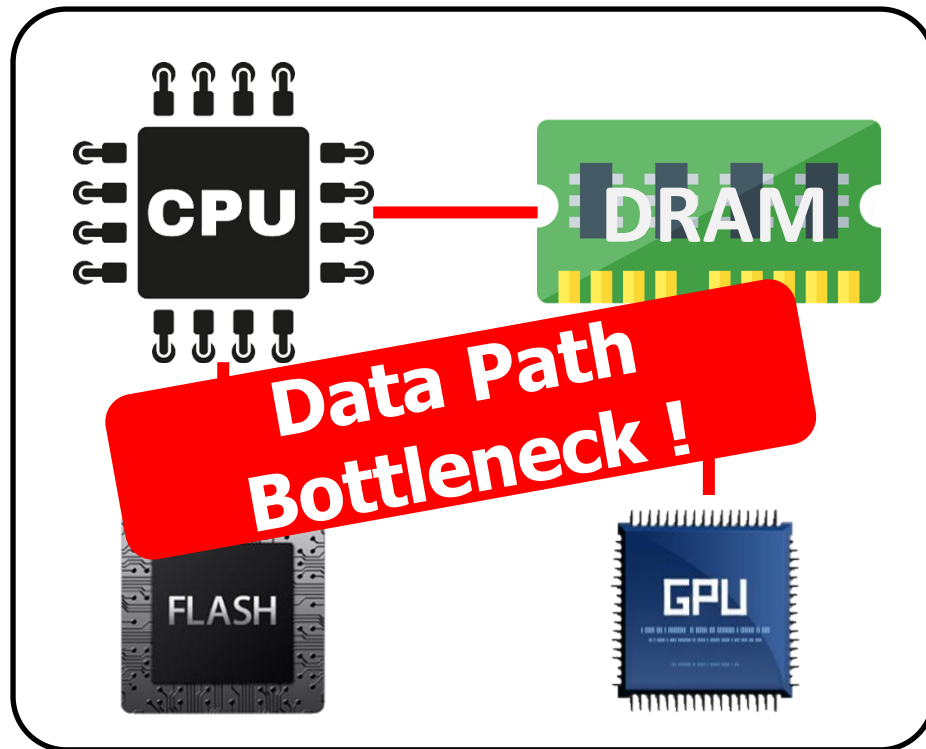
*The system needs to read the whole data  
to re-train in order to add a new category*



# Problem

- On-device training takes too long !
- User wants to keep its data private !

*Mobile Device*



*Cloud Service Providers*



# Related Works

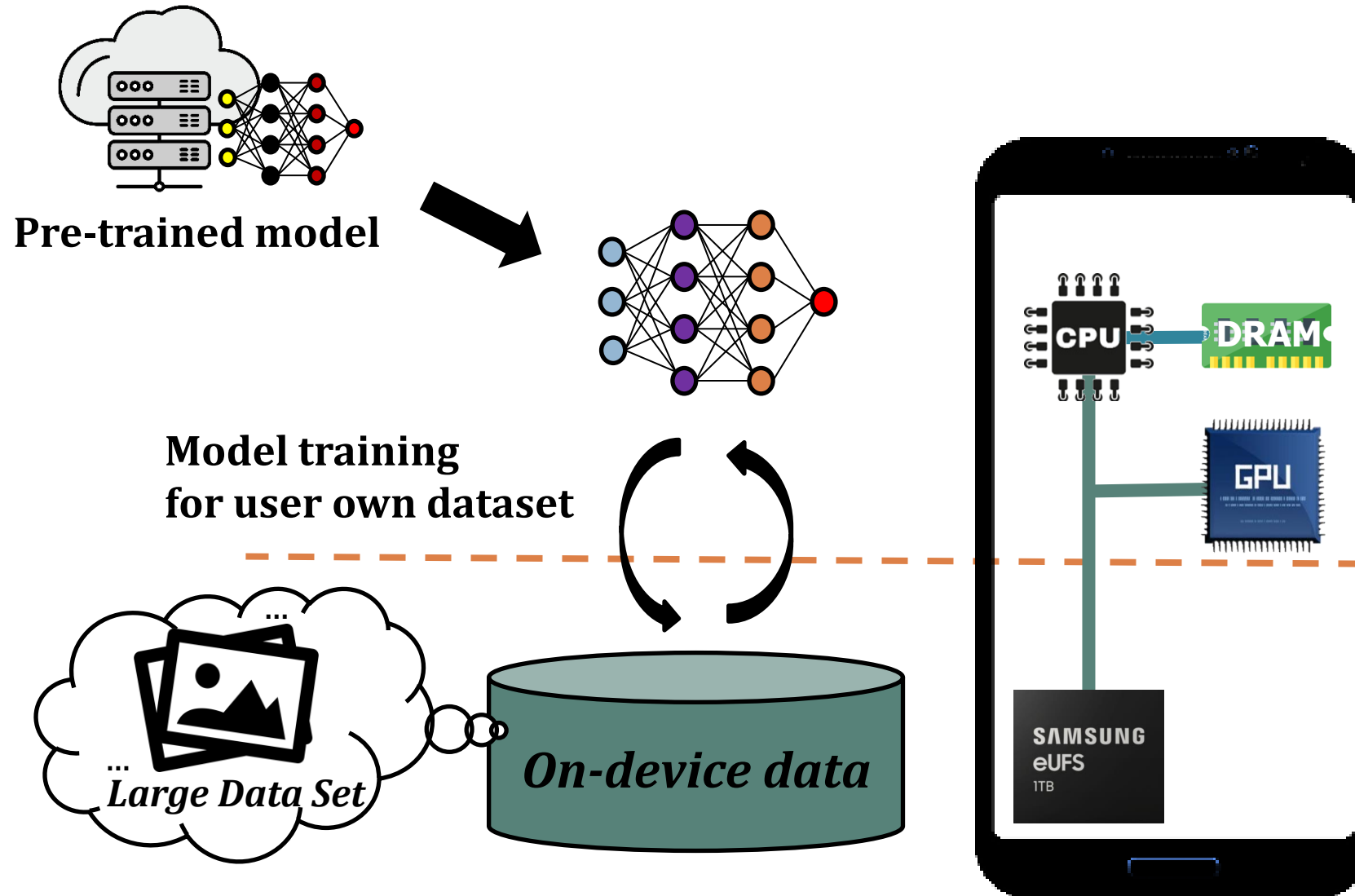
## ➤ Mobile

- ▶ On-Device Machine Learning
- ▶ Federated Learning

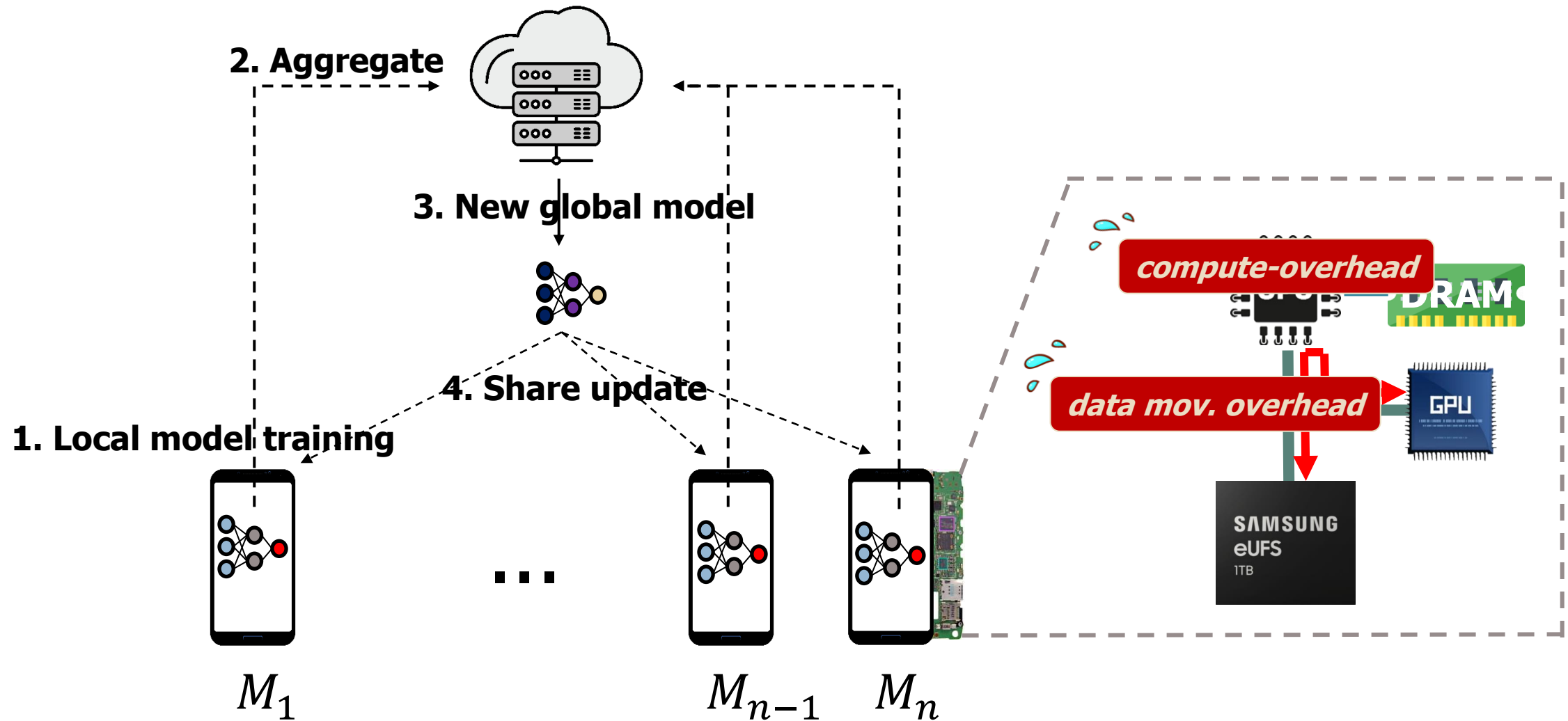
## ➤ Server Computers

- ▶ Processing-in-Memory
- ▶ Processing-in-Storage

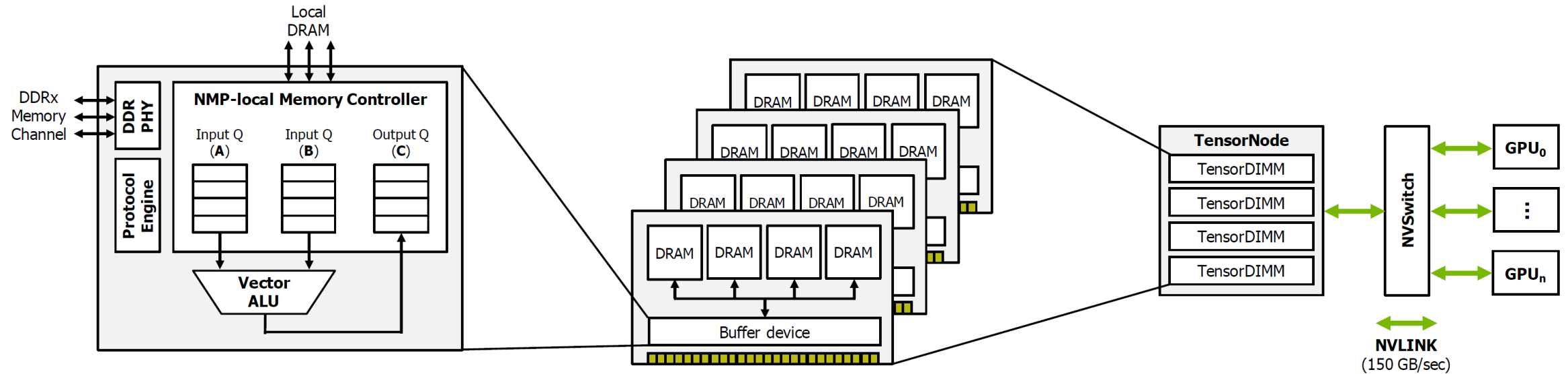
# On-Device Machine Learning



# Federated Learning

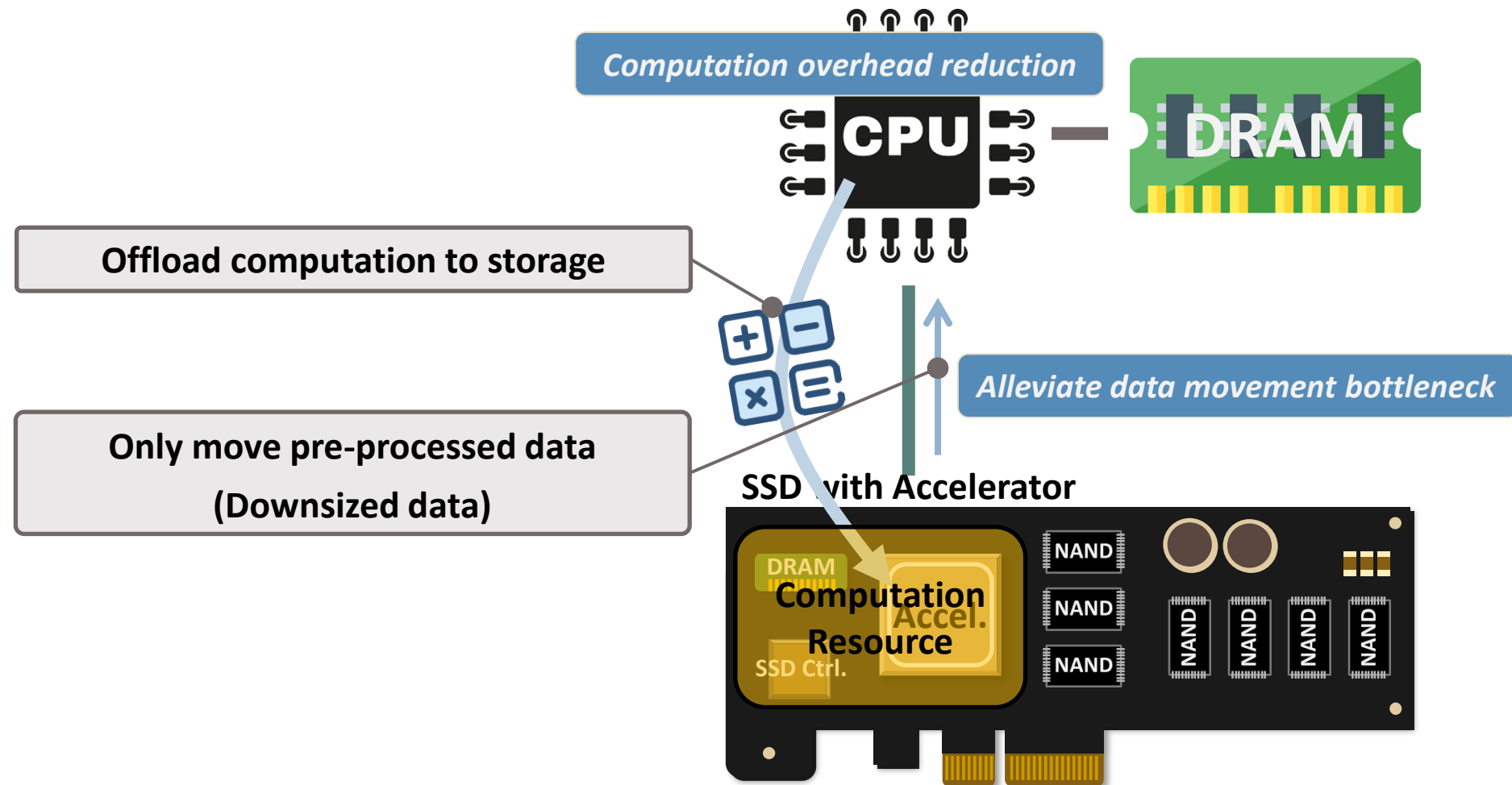


# Processing-in-Memory (PiM)



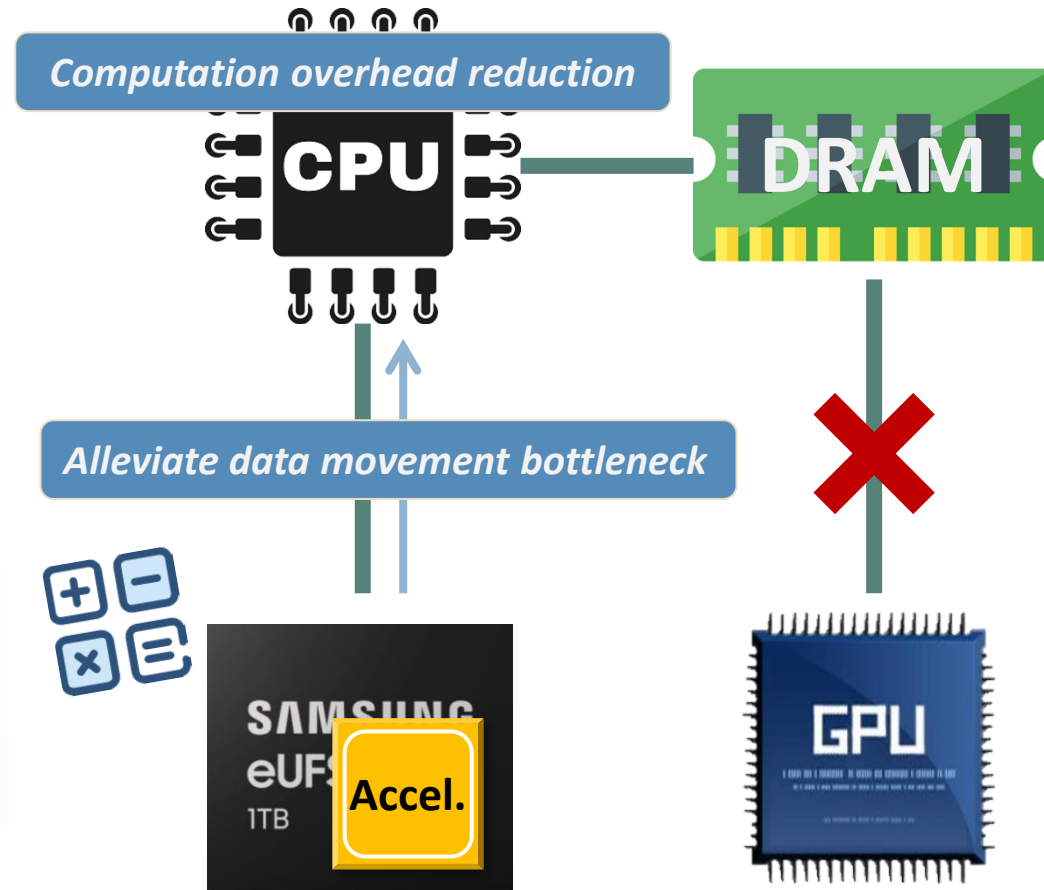


# Processing-in-Storage (PiS)



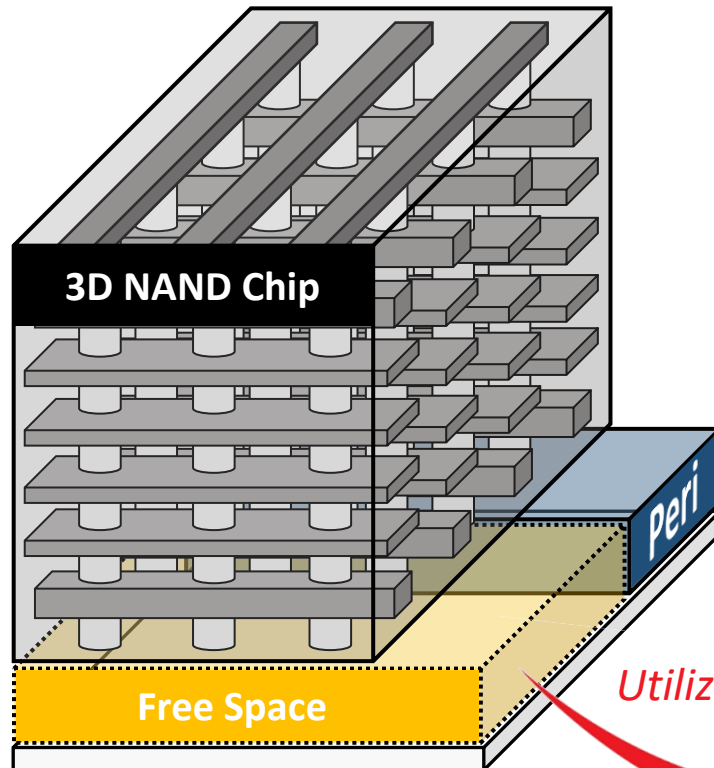
# Key Idea & System Overview

## ➤ Processing-in-Flash (PiF)

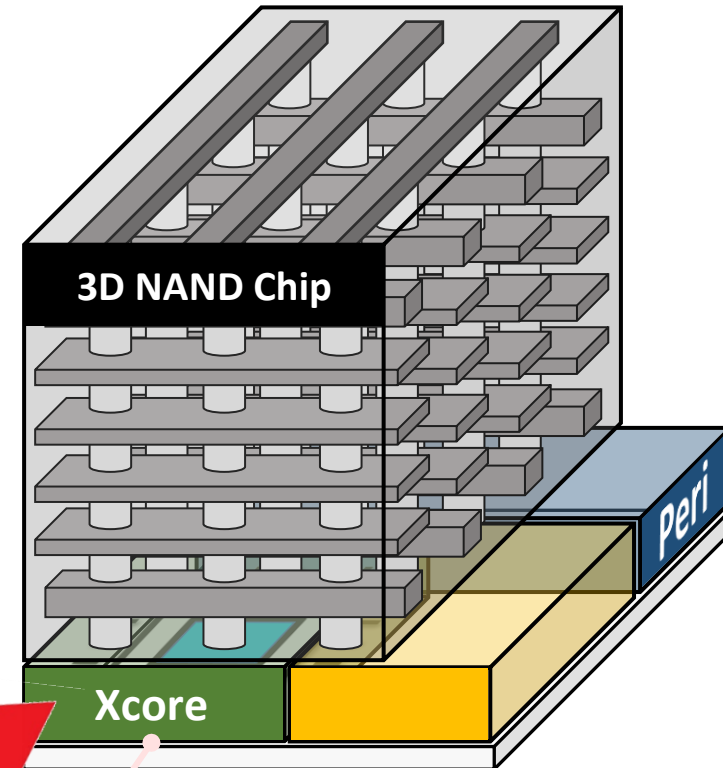


# Cell-over-X (CoX)

Current Cell-over-Peri (CoP) Structure



Proposing Cell-over-X (CoX) Structure



Place accelerator for on-chip processing

# 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

# Overall Project Plan

<i>#Iter.</i>	<i>Objective</i>	<i>Duration</i>	<i>Misc.</i>
1	<b>Ideation &amp; Proposal validation</b> Literature Review, Proposal Feedback	03/16 – 03/30	03/22 proposal (week 4)
2	<b>Build project environment &amp; Design system</b> Build Env. for project, Design Arch. and techniques.	03/31 - 04/14	
3	<b>Implement Techniques &amp; System</b> Implement techniques and Integrate all into system	04/15 – 05/25	05/04 demo (week 10)
4	<b>Evaluation</b> Evaluate system with training Algorithms	05/26 – 06/08	06/08 final (week 15)

# Deliverable

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

## *Success Criteria*

*Performance(Processing-in-Flash)  $\geq$  Performance (Traditional On-Device Processing)*



The background is a digital illustration of a server room. It features long, symmetrical aisles of server racks on both sides. The racks are dark with glowing blue light patterns and small circular lights. The floor is a light blue-grey. The ceiling has a series of rectangular light fixtures. The overall atmosphere is high-tech and digital, with a color palette dominated by blues and greys. The text 'Thank You!' is centered in a large, white, serif font.

# Thank You !

*Any questions or feedback are welcome.*