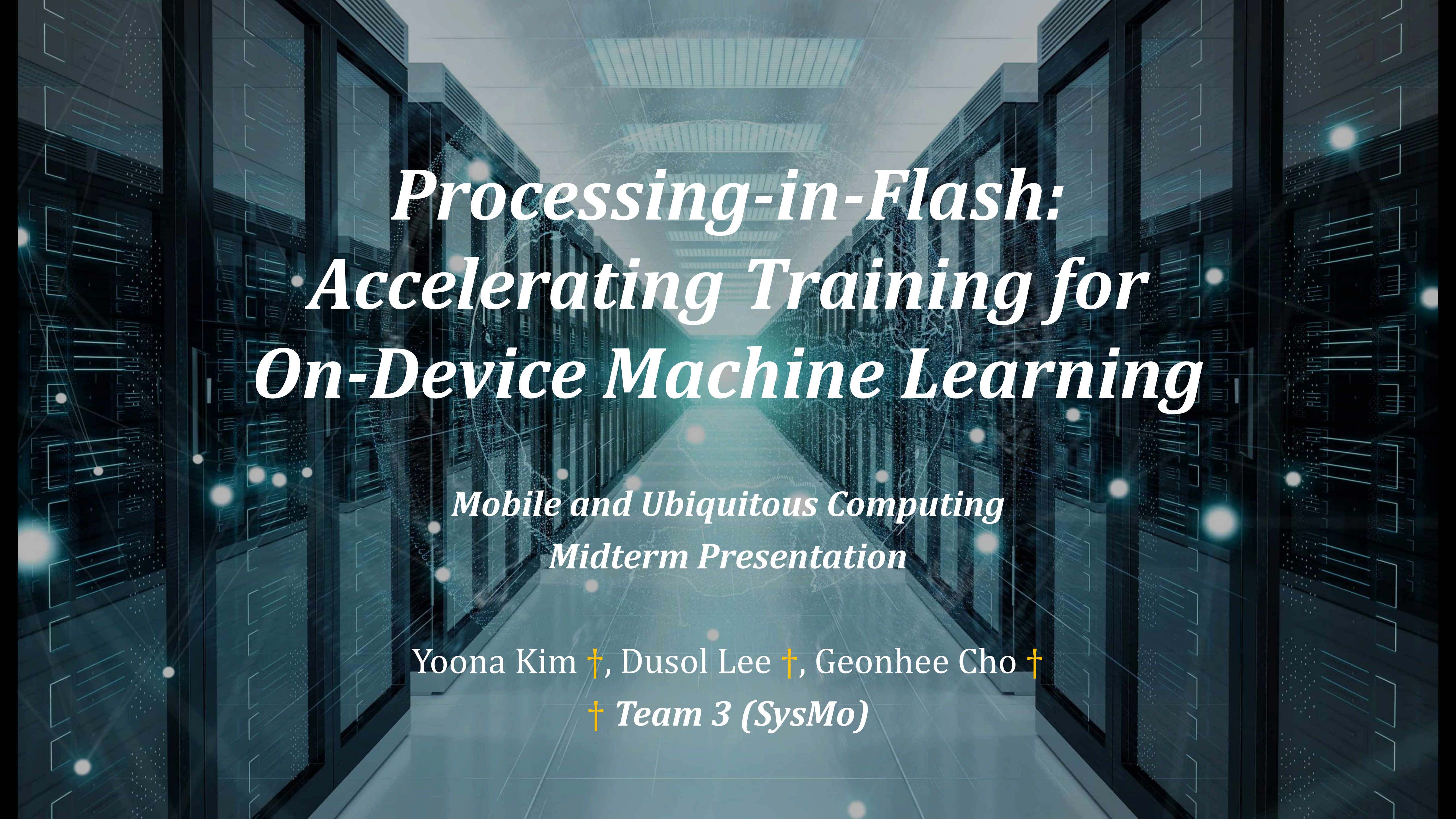


Processing-in-Flash: Accelerating Training for On-Device Machine Learning

Mobile and Ubiquitous Computing
Midterm Presentation

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

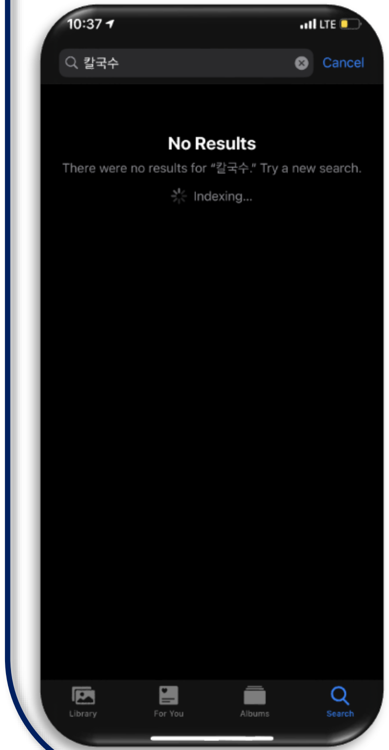
Contents

- **Project Idea**
- Preliminary Investigation
- Project Schedule
- Deliverables and Success Criteria

Remind Our Project Idea

Future mobile application ?
User-definable ML services !

Hey Siri !
Learn "칼국수" and Categorize It !



10:37

칼국수 Cancel

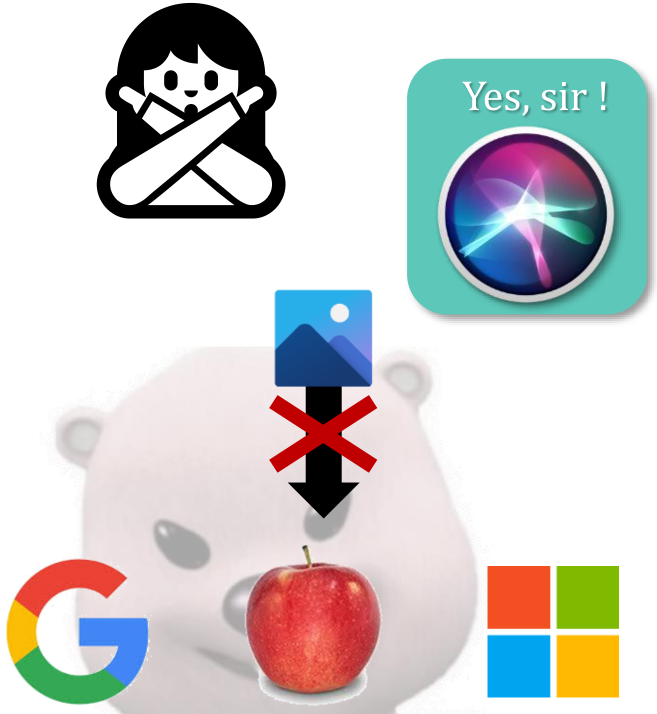
No Results

There were no results for "칼국수." Try a new search.

Indexing...

On-device training
will being essential !

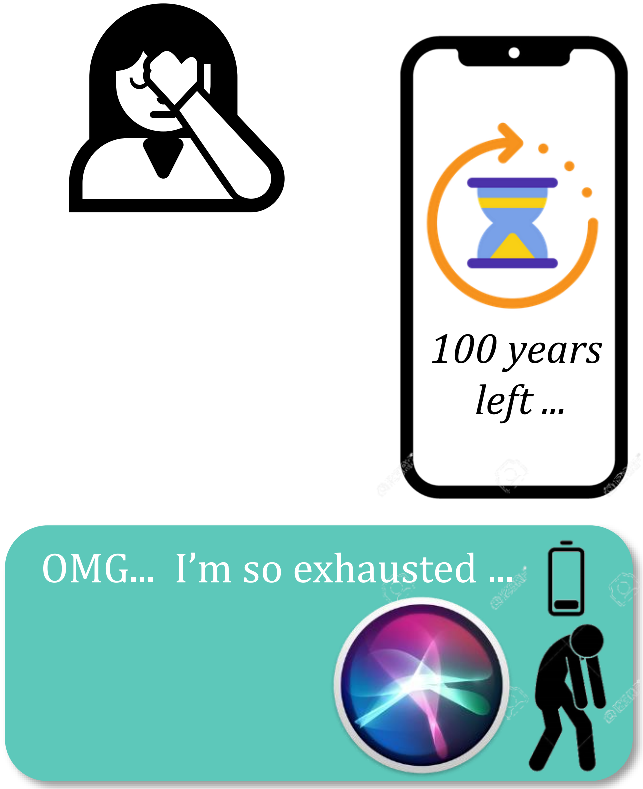
But Siri, keep my images private.



Cloud Service Providers

Can we do on-device training quickly ?
No, because of poor performance !

Why is it taking so long ...?



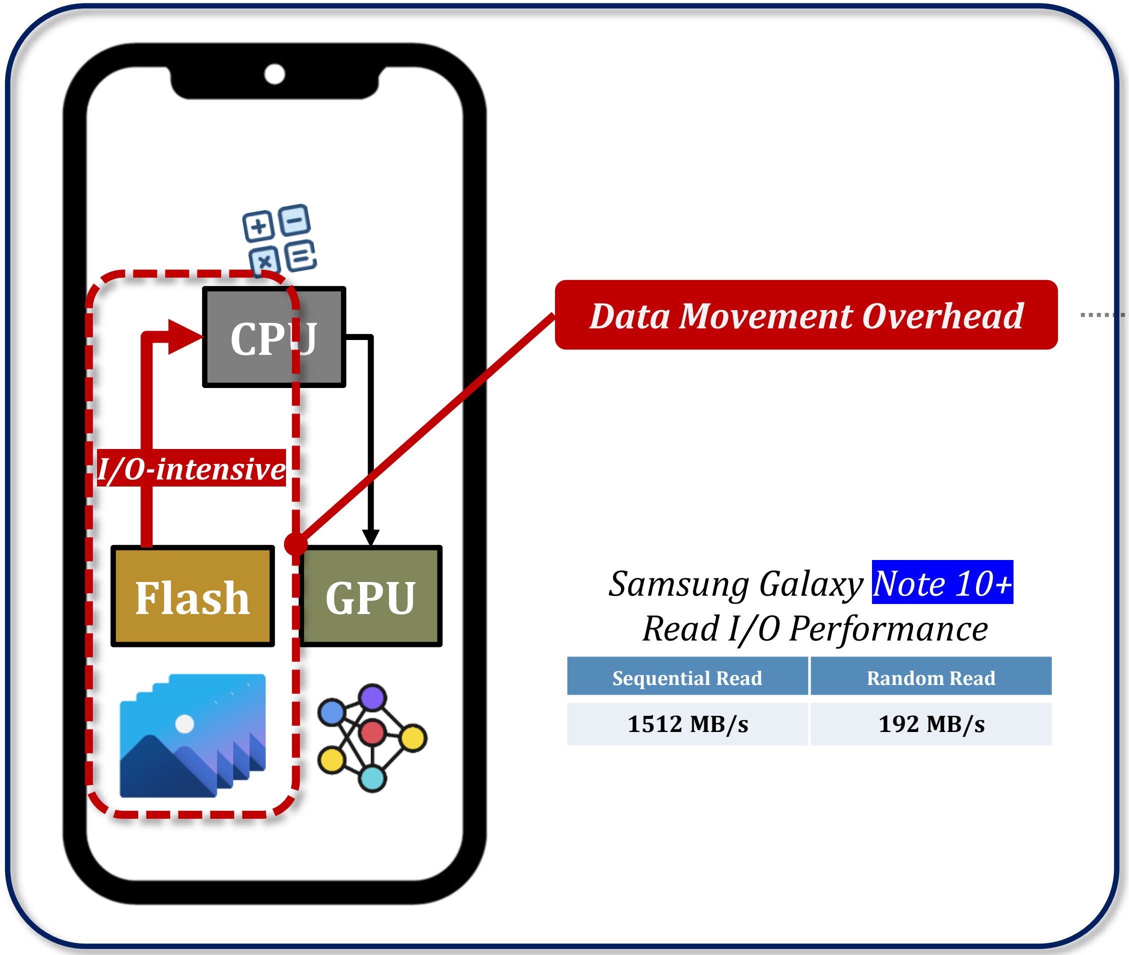
Remind Our Project Idea (Cont.)

We focus on …

Data Movement Bottleneck !

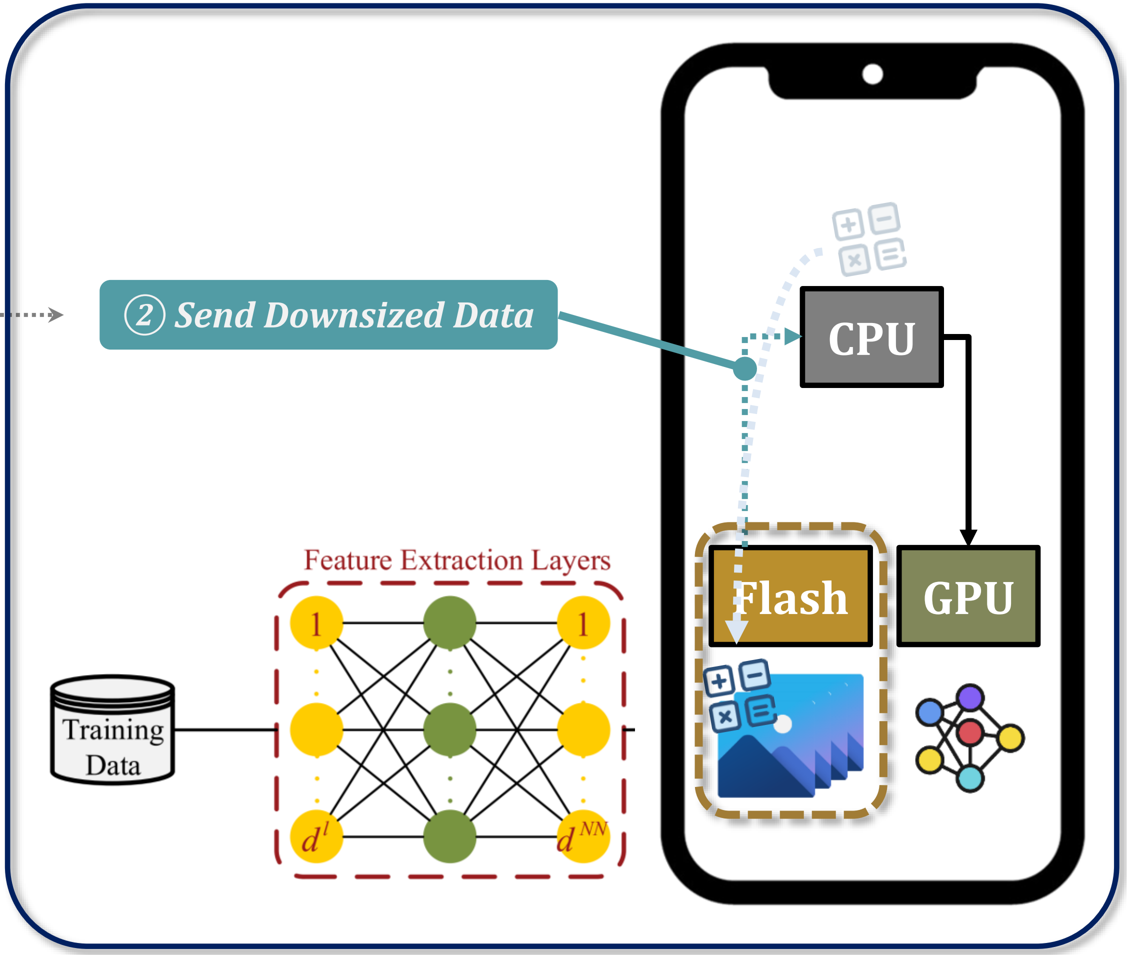
Our approach is …

Processing-in-Flash (PiF) !



Samsung Galaxy Note 10+
Read I/O Performance

Sequential Read	Random Read
1512 MB/s	192 MB/s



Questions We Need to Answer for *PiF*

① Is our hypothesis (data movement will be the bottleneck !) valid ?

② Is it possible to combine accelerator with the mobile flash chip ?

③ If possible, how to derive the specification of a suitable accelerator ?

④ Can *PiF* really do better than baseline system ?

Questions We Need to Answer for *PiF*

① Is our hypothesis (data movement will be the bottleneck !) valid ?

② **Is it possible to combine accelerator with the mobile flash chip ?**

③ **If possible, how to derive the specification of a suitable accelerator ?**

④ Can *PiF* really do better than baseline system ?

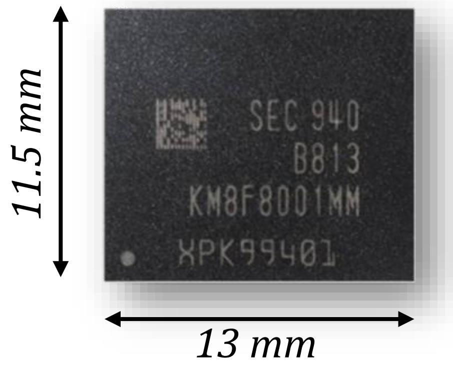
In this midterm-presentation, we will address the questions ② & ③.

Contents

- Project Idea
- **Preliminary Investigation**
- Project Schedule
- Deliverables and Success Criteria

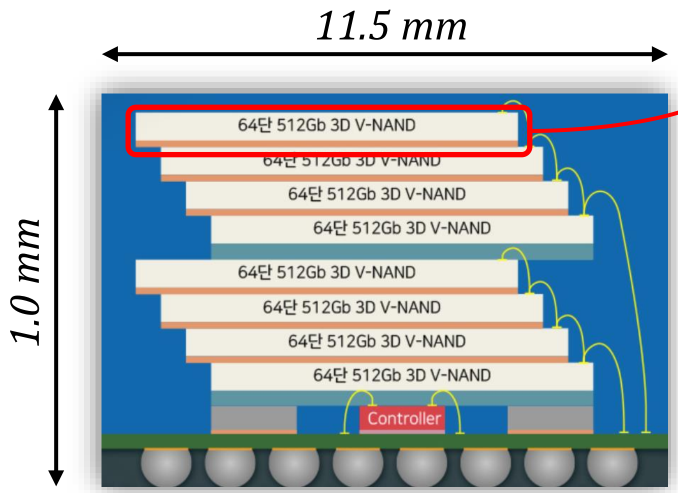
Is it possible to combine accelerator with the mobile flash chip ?

State-of-the Art Mobile Flash Package



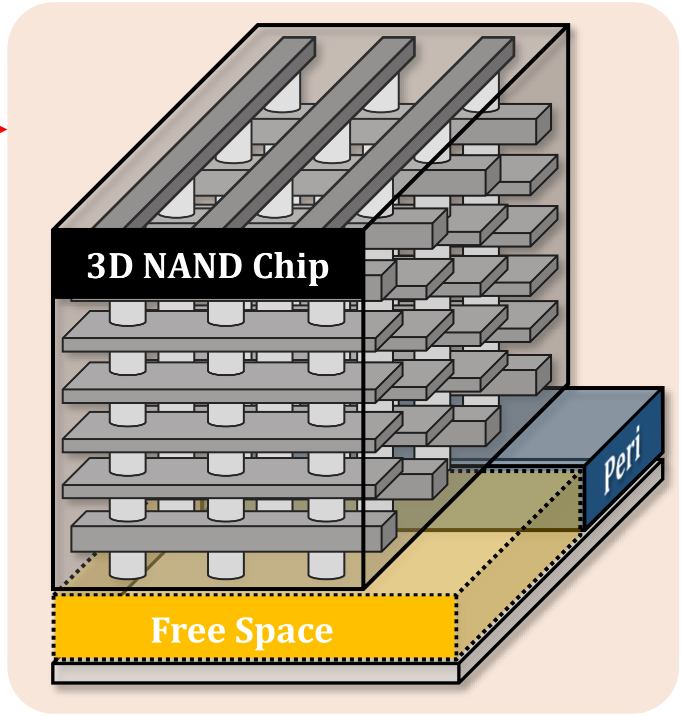
$\approx 150\ mm^2$

Inside the Flash Package: The Array of Flash Chips



512 GB eUFS Package

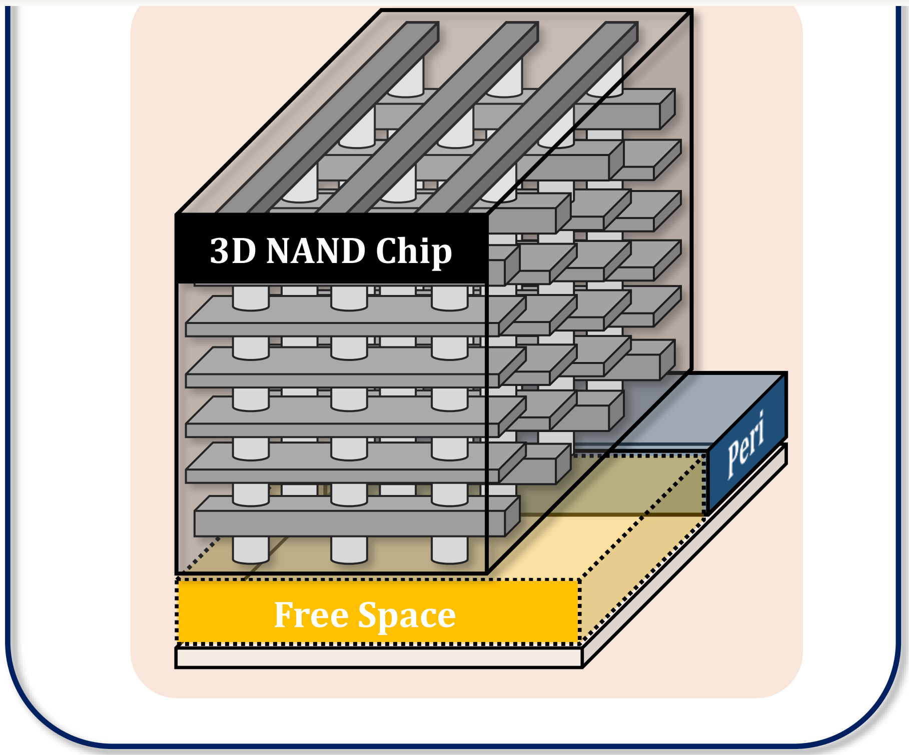
Inside the Flash Chip: Cell-over-Peri (CoP) Structure

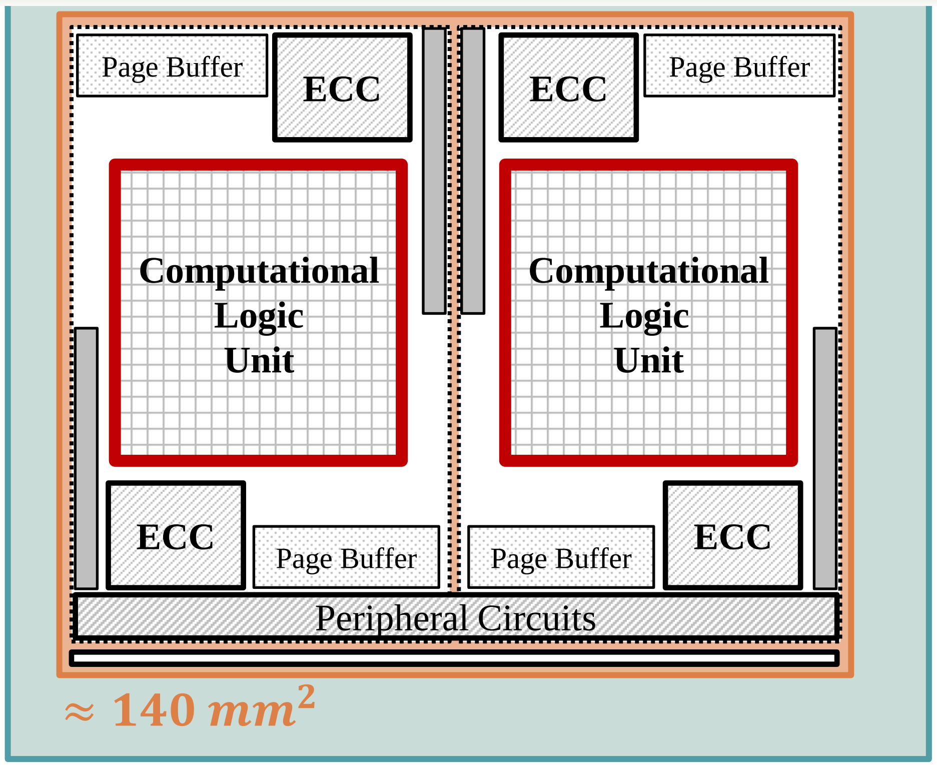


Is it possible to combine accelerator with the mobile flash chip ?

Is it possible to combine accelerator with the mobile flash chip ?

Answer: "Yes. There is enough free space"





$\approx 150\ mm^2$

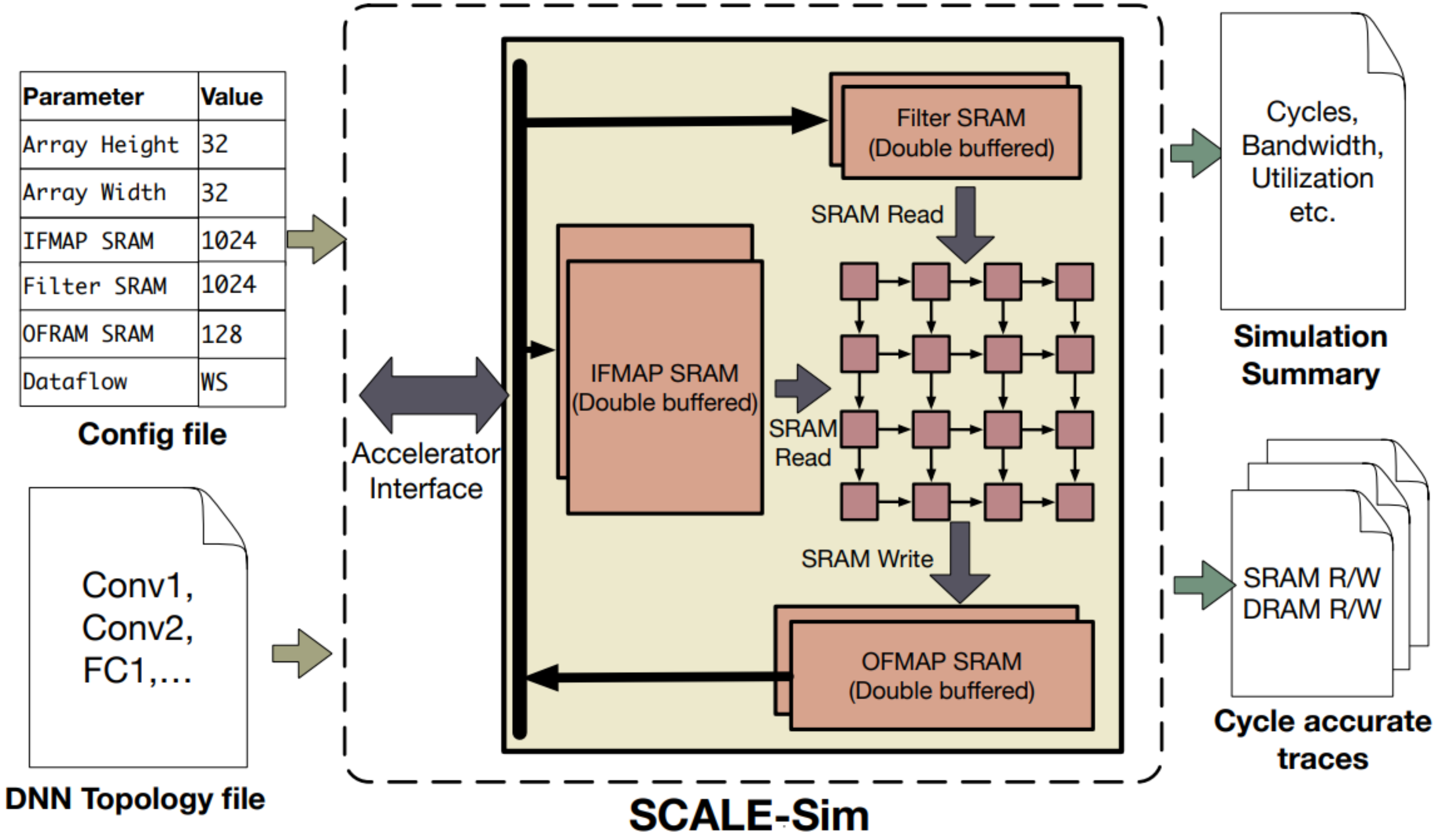
$\approx 100\ mm^2$ (70%)

	Apple A12 GPU Die Size
Size	$< 15\ mm^2$

How to derive the specification of a suitable accelerator ?

How to derive the specification of a suitable accelerator ?

Existing Accel. Design Space Exploration tool



Parameter	Value
Array Height	32
Array Width	32
IFMAP SRAM	1024
Filter SRAM	1024
OFRAM SRAM	128
Dataflow	WS

“ SCALE Sim: Systolic CNN Accelerator ”, arXiv:1811.02883 , 2018

How to derive the specification of a suitable accelerator ?

Memory Specification

Memory Spec	BW
SRAM	15 GB
NAND Interface	1.2 GB

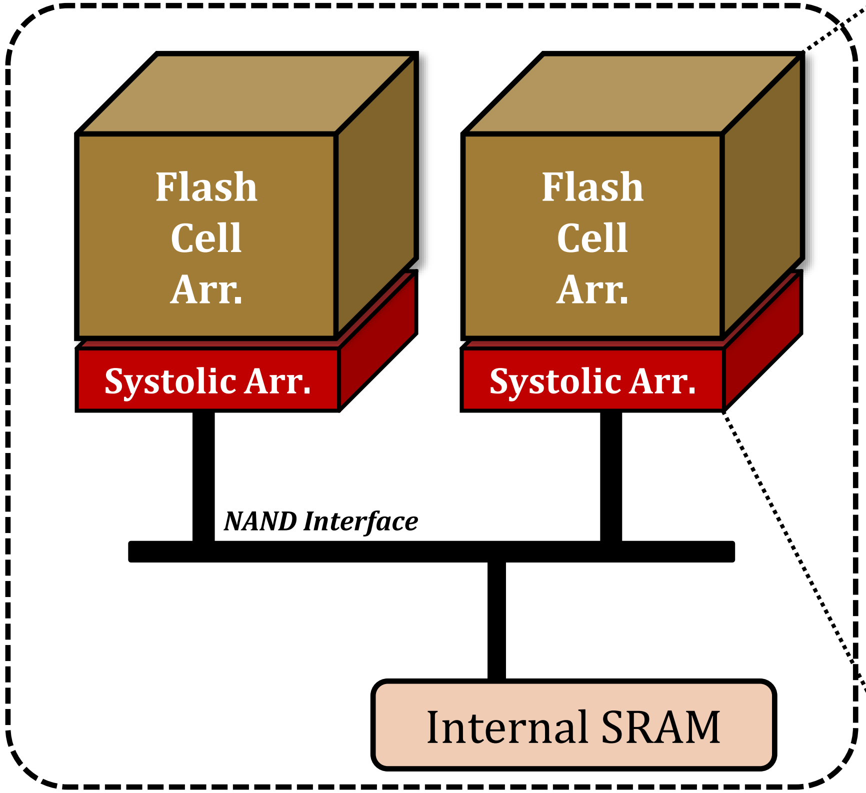
Flash Property	value
Page Size	16 KB
Cell Type	SLC
# of Plane	8
# of Chip	4

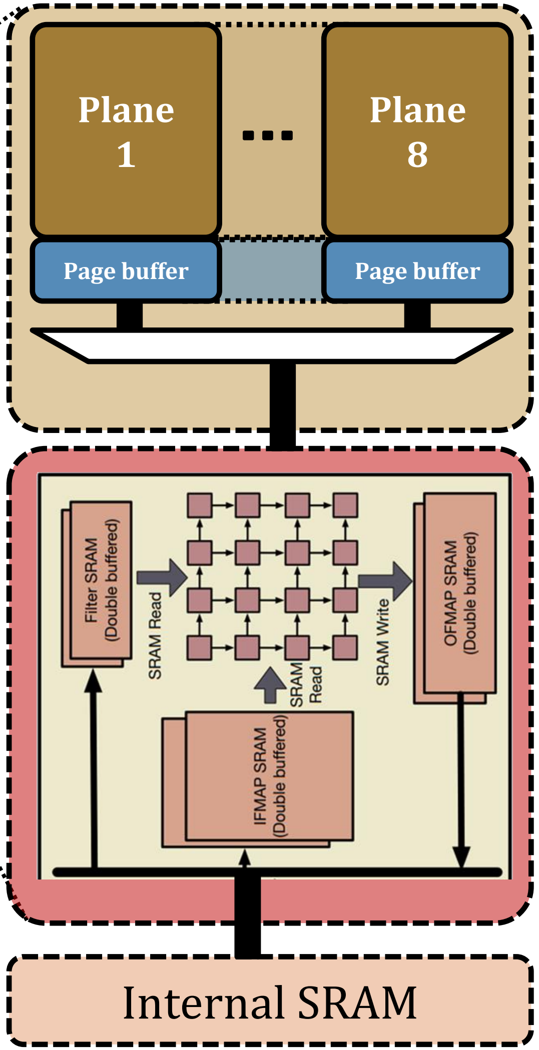
⋮

Accel. Data Path

Data	Data path
Input	Page Buf. Interface
Weight	Page Buf. Interface
Medium Result	SRAM Interface
Final Result	SRAM Interface

Flash Package




Plane 1
...
Plane 8
Page buffer
Page buffer
Filter SRAM (Double buffered)
SRAM Read
SRAM Read
SRAM Write
OFMAP SRAM (Double buffered)
IFMAP SRAM (Double buffered)
Internal SRAM

CoX-Sim

(For Only Internal-Flash Package Exploration)

Contents

- Project Idea
- Preliminary Investigation
- **Project Plan**
- Deliverable and Success Criteria

Overall Project Plan

① **Is our hypothesis (data movement will be the bottleneck !) valid ?**

- ▸ Android Reference Board (HiKey 960, etc.) or Pixel Phone

② Is it possible to combine accelerator with the mobile flash chip ?

③ How to derive the specification of a suitable accelerator ?

④ **Can *PiF* really do better than baseline system ?**

- ▸ Micro Benchmark w/ Flash Chip Simulator (CoX-Sim)
- ▸ Macro Benchmark w/ Whole System Simulator (T.B.D.)

In Final-presentation, we will answer all the questions.
(Especially focused on ④)

Contents

- Project Idea
- Preliminary Investigation
- Project Schedule
- **Deliverable and Success Criteria**

Deliverables and Success Criteria

Deliverables	*Success Criteria*
On-Device Training Benchmark Results	*Verify that data movement is the bottleneck*
~~*Quantitative investigation of mobile flash packages and flash chips*~~	~~*Verification of whether it is possible to mount an accelerator on the flash chip*~~
Flash Chip Simulator (a.k.a. CoX-Sim) ≈ 80%	*Proving that the PiF performs better*
Whole System Simulator (or Emulator)	
Macro & Micro Benchmark Results	

Thank You !

Any questions or feedback are welcome.