usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

FAST'25

# Archer: Adaptive Memory Compression with Page-Association-Rule Awareness for High-Speed Response of Mobile Devices

**Changlong Li**, Zongwei Zhu, Chao Wang, Fangming Liu, Fei Xu,
Edwin H.-M. Sha, and Xuehai Zhou

# Contents

1 **Background & Motivation**

2 Design

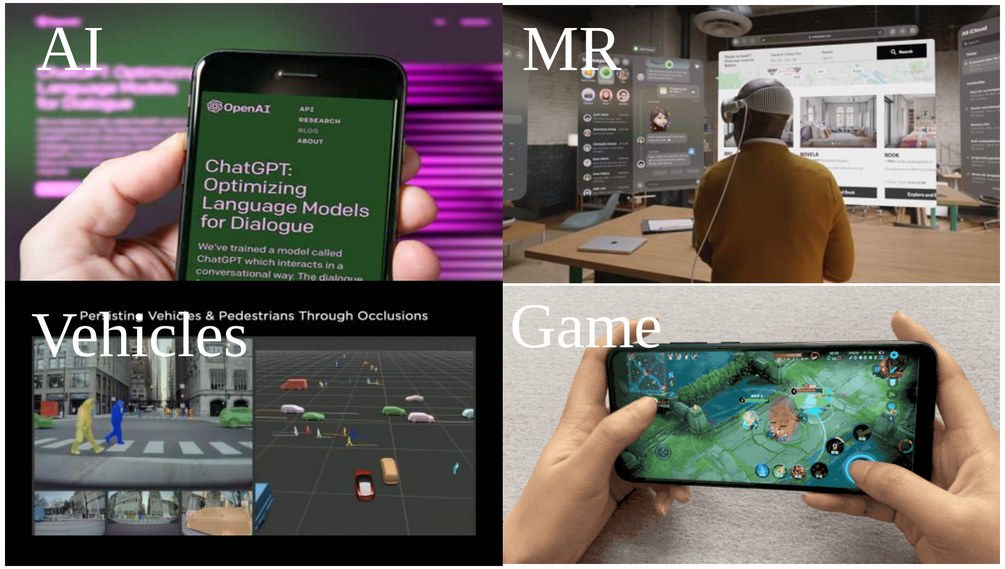
3 Evaluation

4 Conclusion

# Emerging Mobile Scenarios

- People spend more and more time on mobile devices (e.g., smartphones, autonomous vehicles, and robots)
- Emerging scenarios make the memory issue more serious
  - VR/MR, transformer-based models, mobile games, camera-based services

Mobile Devices


AI
OpenAI
ChatGPT: Optimizing Language Models for Dialogue
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue
MR
Vehicles
Persisting Vehicles & Pedestrians Through Occlusions
Game

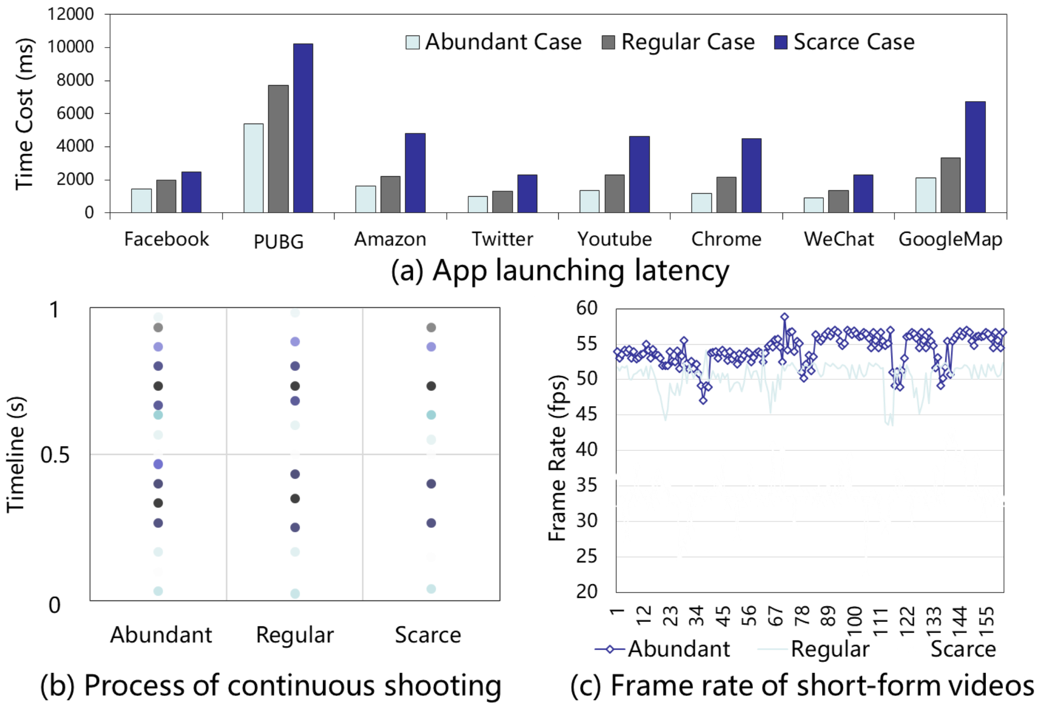
Memory-intensive apps in mobile scenarios

# Memory Issue in Mobile Scenarios

- Memory compression is widely adopted to save space
- Page compression limits system performance and degrades the user experience
  - Typical scenarios studied: **app launching**, **continuous shooting**, and **short-form video**

# Memory Issue in Mobile Scenarios

- Memory compression is widely adopted to save space
- Page compression limits system performance and degrades the user experience
  - ➢ Typical scenarios studied: **app launching**, **continuous shooting**, and **short-form video**



(a) App launching latency

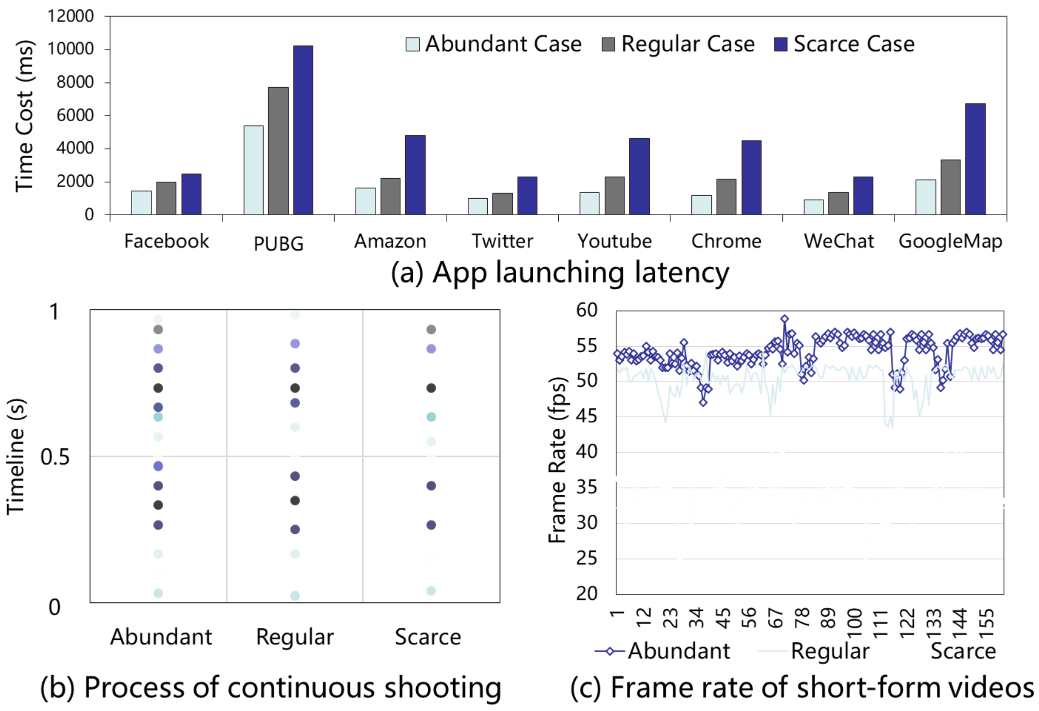
(b) Process of continuous shooting

(c) Frame rate of short-form videos

**Response time** significantly increases **with compression**.

# Memory Issue in Mobile Scenarios

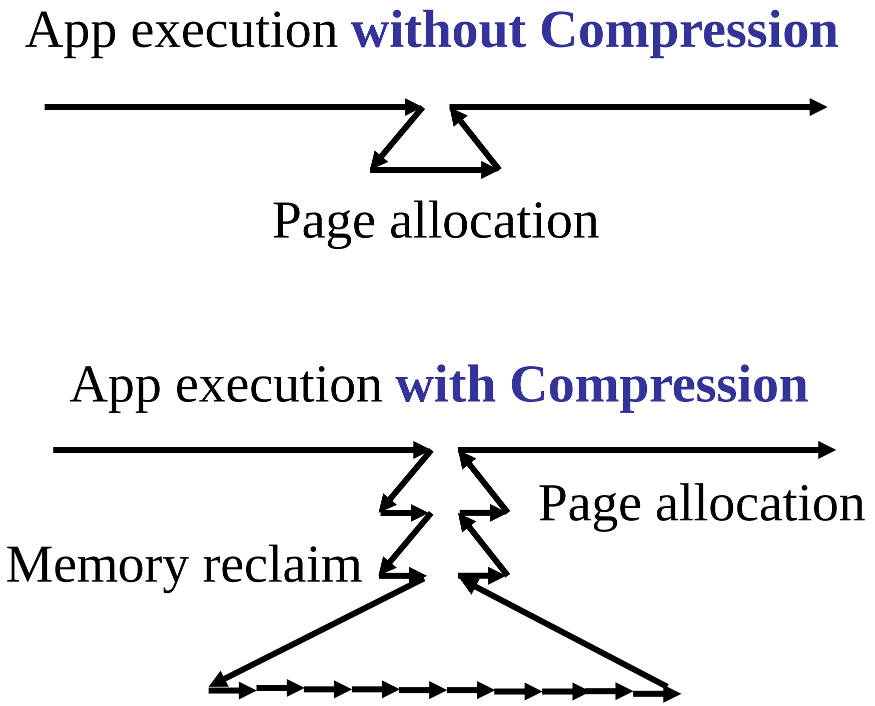
- Memory compression is widely adopted to save space
- Page compression limits system performance and degrades the user experience
  - ➢ Typical scenarios studied: **app launching**, **continuous shooting**, and **short-form video**



(a) App launching latency

(b) Process of continuous shooting

(c) Frame rate of short-form videos

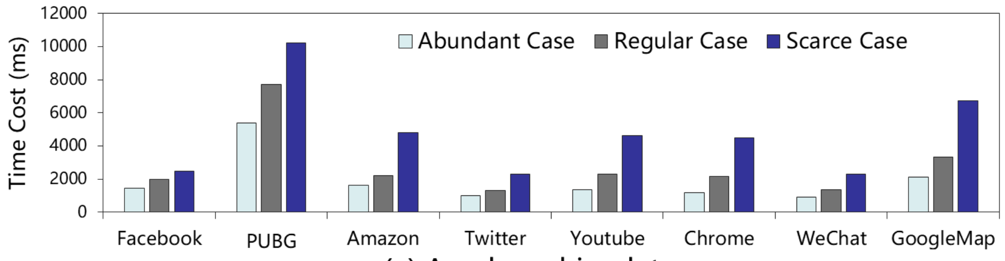

App execution **without Compression**
Page allocation
App execution **with Compression**
Page allocation
Memory reclaim

**Page-by-page compression**

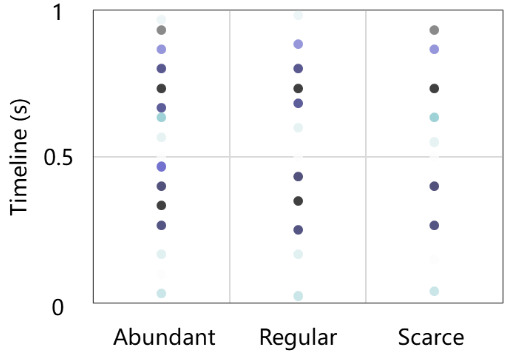
**Response time** significantly increases **with compression**.

# Memory Issue in Mobile Scenarios

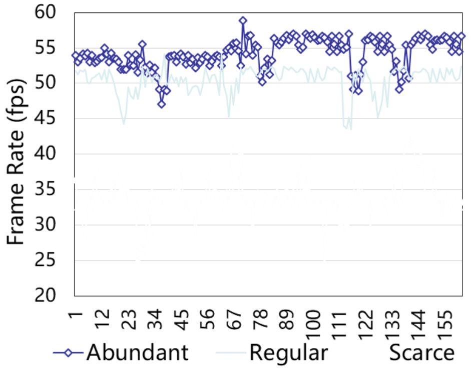
- Memory compression is widely adopted to save space
- Page compression limits system performance and degrades the user experience
  - ➢ Typical scenarios studied: **app launching**, **continuous shooting**, and **short-form video**



(a) App launching latency

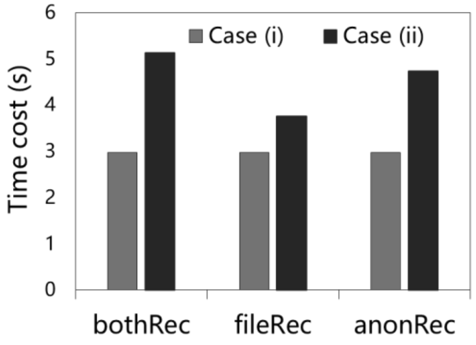


(b) Process of continuous shooting

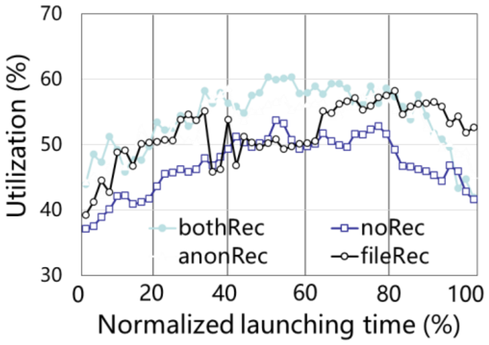


(c) Frame rate of short-form videos

**Response time** significantly increases **with compression**.

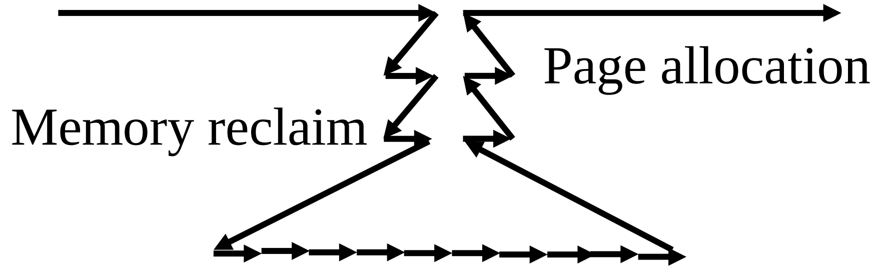


(a) App launching time



(b) CPU utilization

App execution **with Compression**



Page allocation

Memory reclaim

**Page-by-page compression**

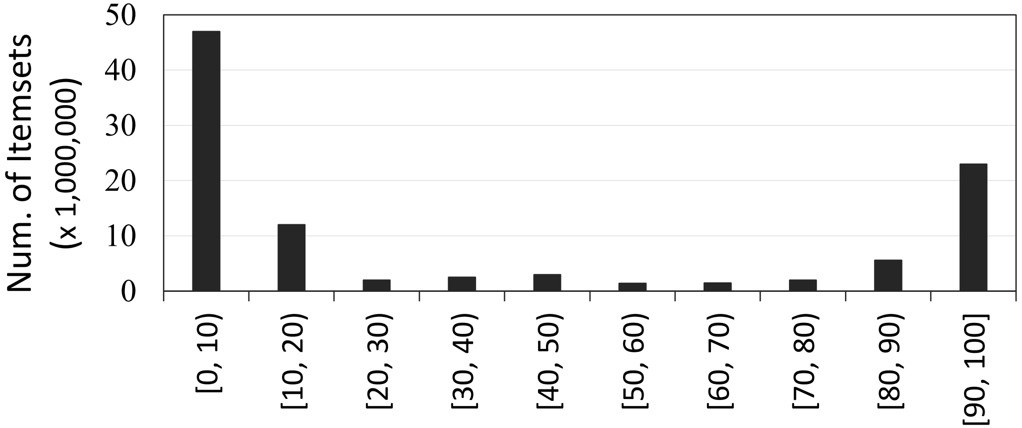
# Opportunities of Large-grain Compression

- Page compression is reasonable: minimize the read amplification
- ***Can we do large-grain compression without increasing read amplification?***

# Opportunities of Large-grain Compression

- Page compression is reasonable: minimize the read amplification
- ***Can we do large-grain compression without increasing read amplification?***
- Observation: 26.3% of anonymous pages are strongly associated with others
  - Like **Beer** and **Diaper**, many pages are always accessed together
  - The association is implicit but can be mined
  - If the system is **aware of these highly associated pages** and **does large-grain compression on them**, the performance has the potential to be significantly improved.

Association Rule Mining



Association *confidence* of page combinations (%)

# Contents

1 Background & Motivation

2 **Design**

3 Evaluation

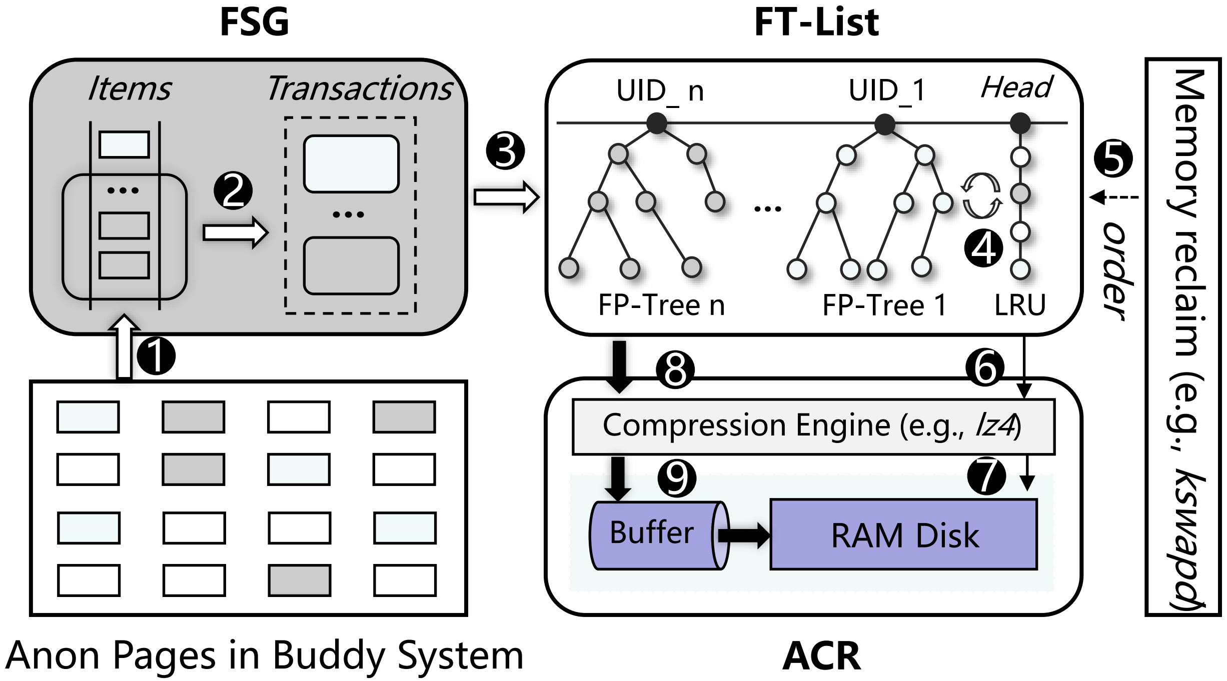
4 Conclusion

# Overview

- **Archer:** an Association-rule aware memory compression framework for the high speed response of mobile devices

# Overview

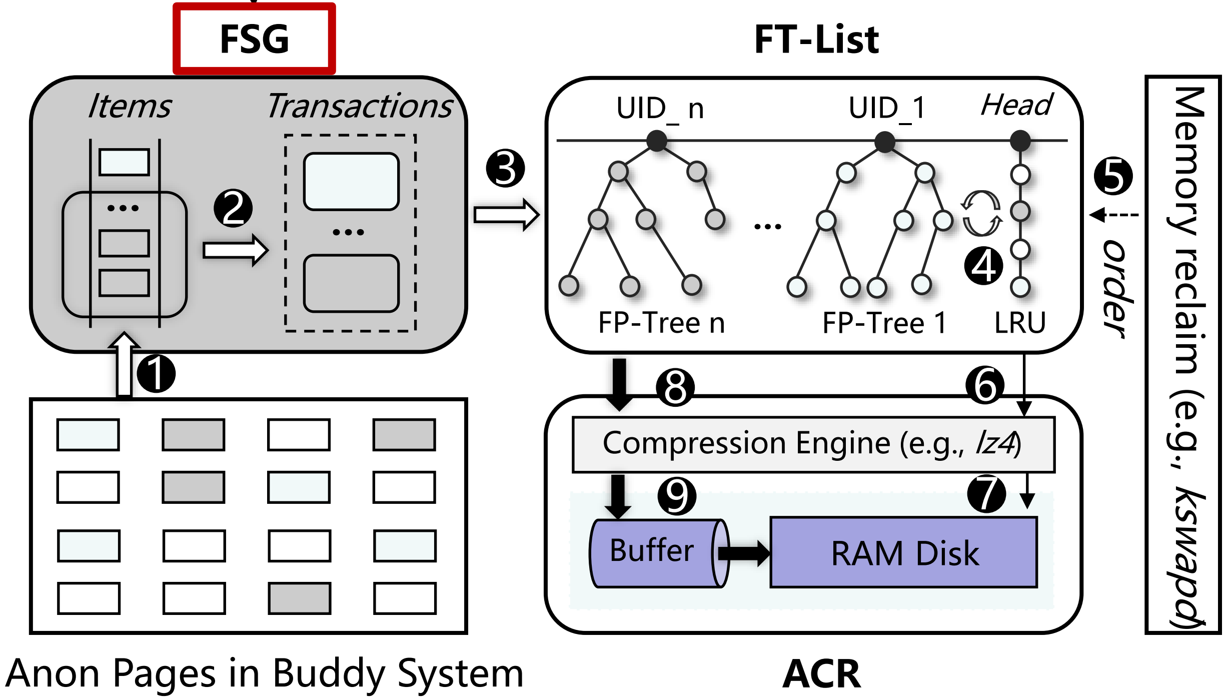
- **Archer:** an Association-rule aware memory compression framework for the high speed response of mobile devices
  - **FSG**: Footprint Stream Generator
  - **FT-List**: Frequent-pattern Tree List
  - **ACR**: Adaptive Compression Region



# Component: FSG

- **Archer:** an Association-rule aware memory compression framework for the high speed response of mobile devices
  - ➢ **FSG**: Footprint Stream Generator
  - ➢ **FT-List**: Frequent-pattern Tree List
  - ➢ **ACR**: Adaptive Compression Region

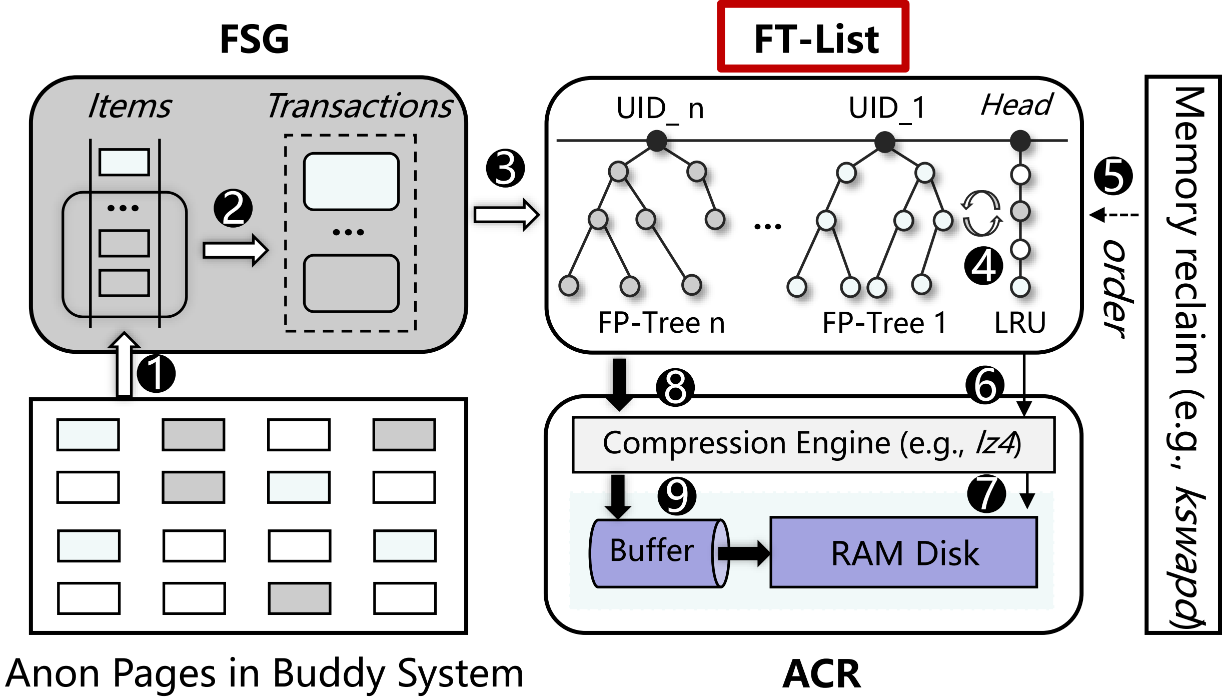
□ FSG maintains a sliding window to generate footprint stream



# Component: FT-List

- **Archer:** an Association-rule aware memory compression framework for the high speed response of mobile devices
  - ➢ **FSG**: Footprint Stream Generator
  - ➢ **FT-List**: Frequent-pattern Tree List
  - ➢ **ACR**: Adaptive Compression Region

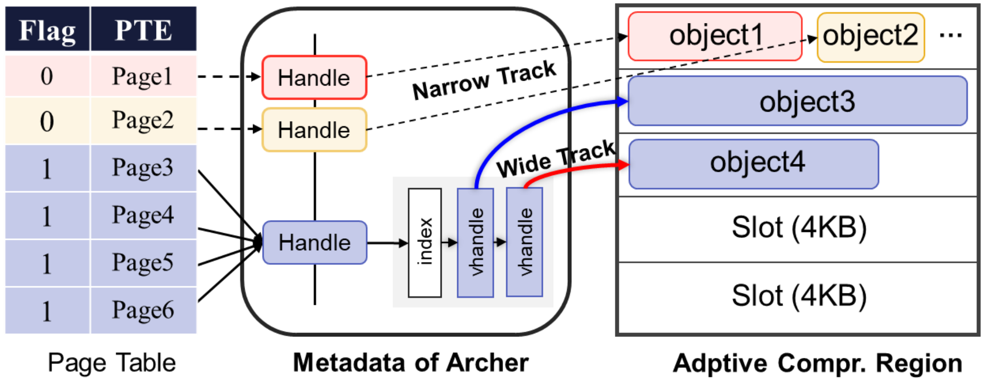
□ Codesign LRU with FP-Tree. Pages are managed and mined on the novel structure: FT-List.

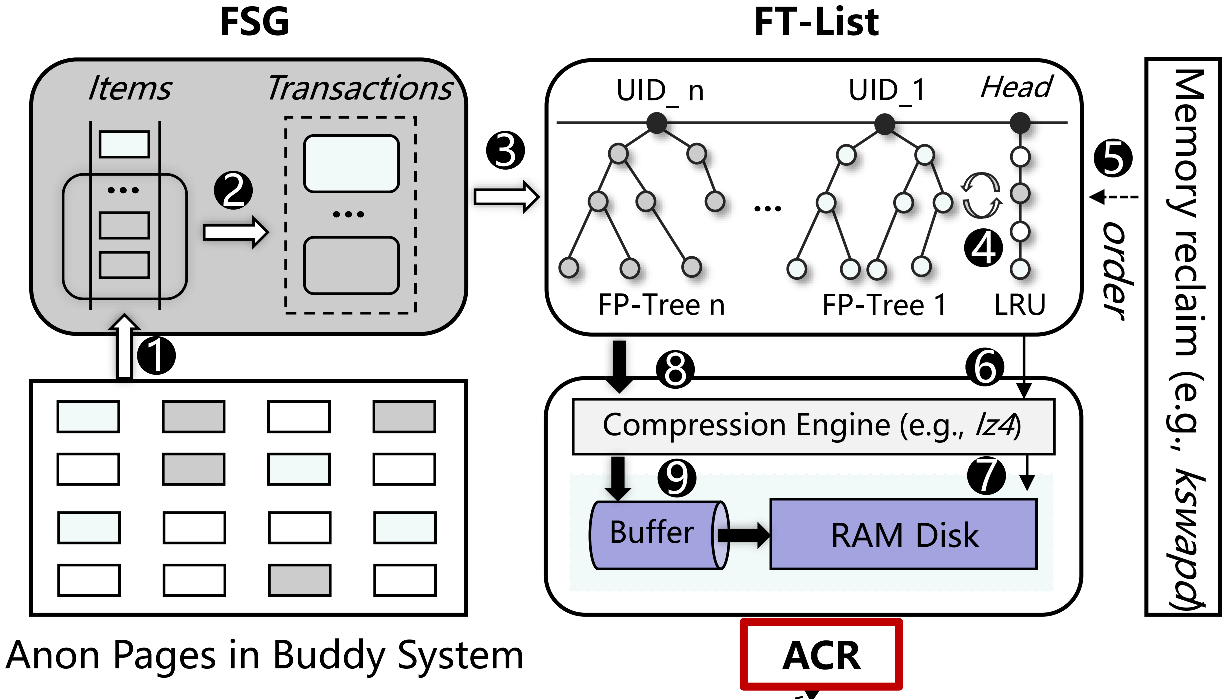


# Component: ACR

- **Archer:** an Association-rule aware memory compression framework for the high speed response of mobile devices
  - **FSG**: Footprint Stream Generator
  - **FT-List**: Frequent-pattern Tree List
  - **ACR**: Adaptive Compression Region

□ ACR is implemented based on the original ZRAM region in the Linux kernel.





# Contents

1 Background & Motivation

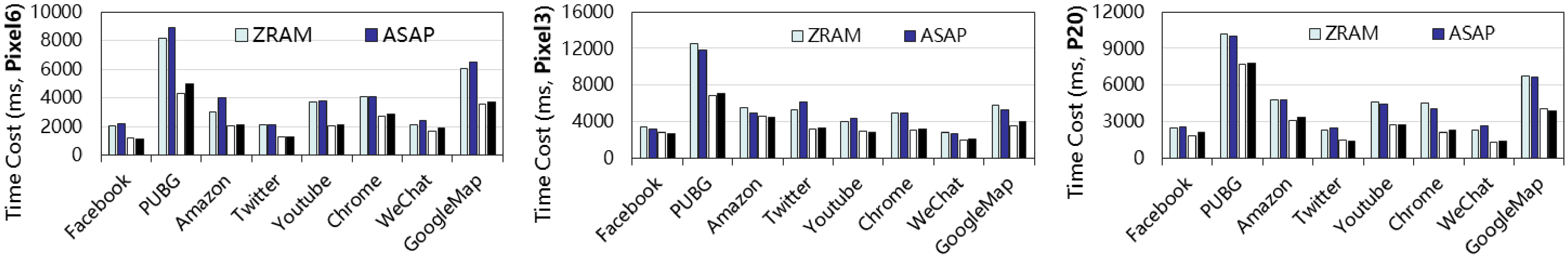
2 Design

3 **Evaluation**

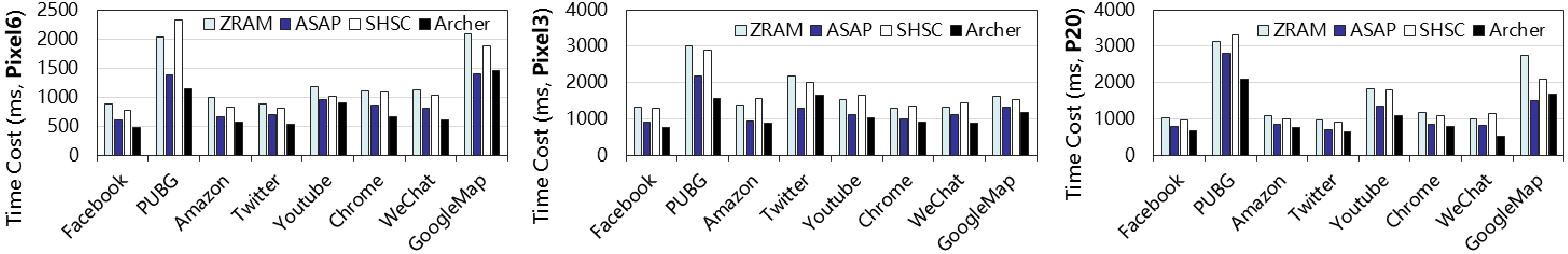
4 Conclusion

# Evaluation

- The evaluations are performed on **Google Pixel6**, **Pixel3**, and **HUAWEI P20**.
- User experience measurement: **App launching speed**, **photographic performance**, and **frame rate**.



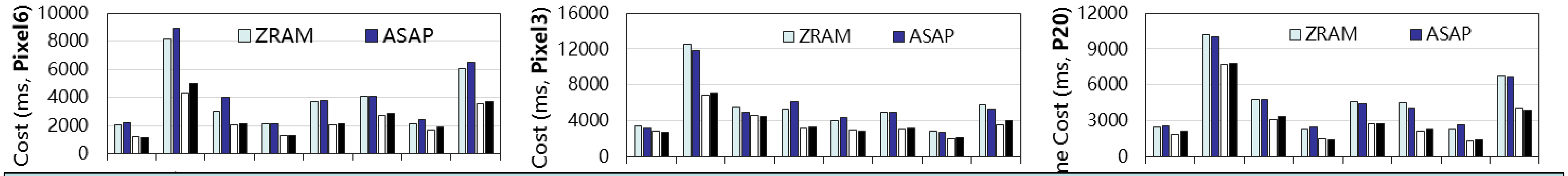
(a). Time cost of cold launching on Pixel6, Pixel3, and P20 phones



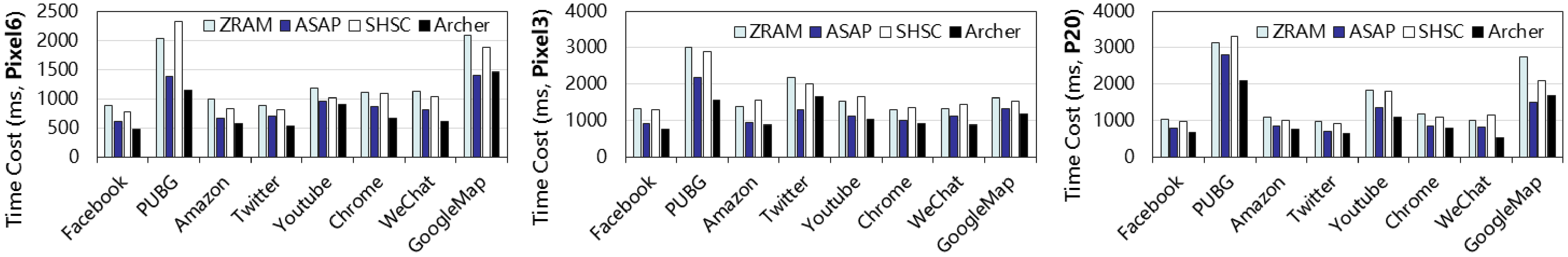
(b). Time cost of hot launching on Pixel6, Pixel3, and P20 phones

# Evaluation

- The evaluations are performed on **Google Pixel6**, **Pixel3**, and **HUAWEI P20**.
- User experience measurement: **App launching speed**, **photographic performance**, and **frame rate**.



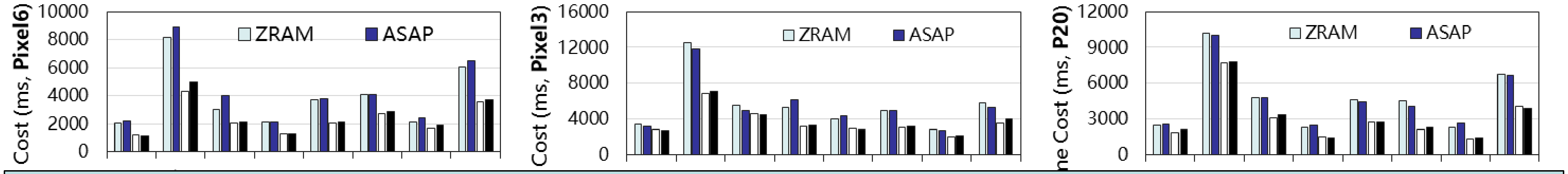
The average speed of cold launching with Archer increased by **37.2%** on Pixel6, **30.6%** on P20, and **32.9%** on Pixel3, compared to ZRAM.



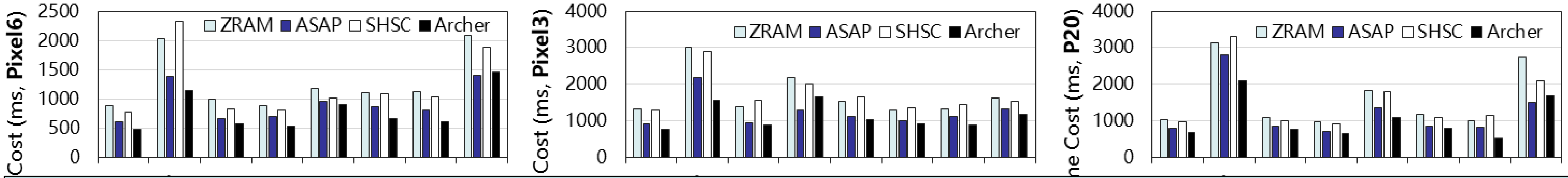
(b). Time cost of hot launching on Pixel6, Pixel3, and P20 phones

# Evaluation

- The evaluations are performed on **Google Pixel6**, **Pixel3**, and **HUAWEI P20**.
- User experience measurement: **App launching speed**, **photographic performance**, and **frame rate**.



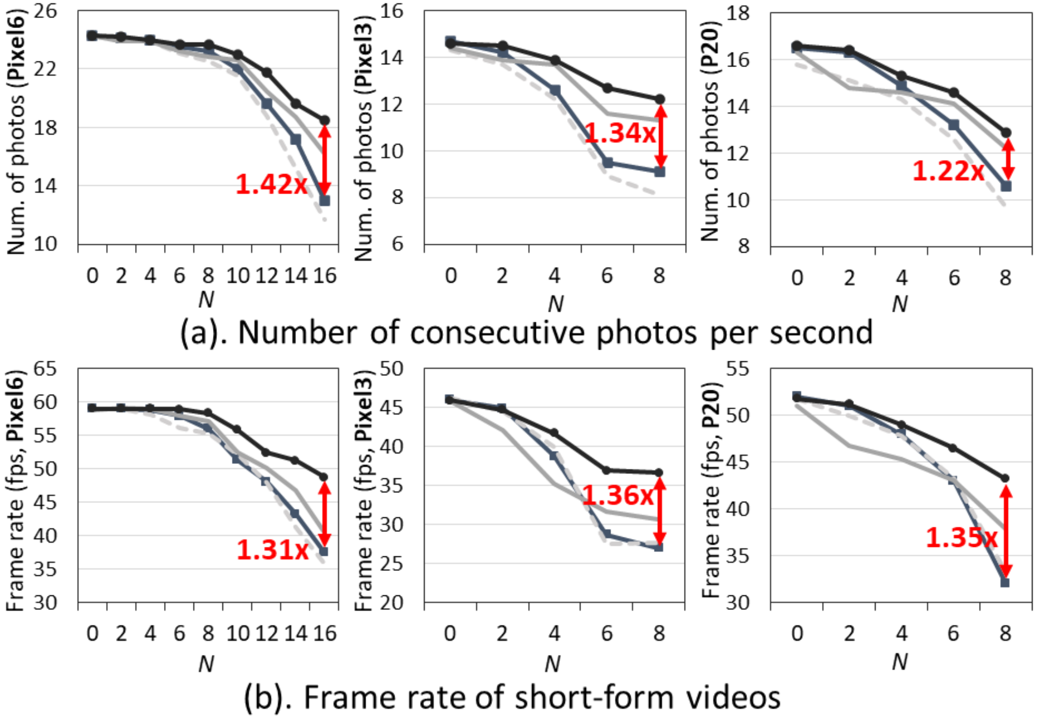
The average speed of cold launching with Archer increased by **37.2%** on Pixel6, **30.6%** on P20, and **32.9%** on Pixel3, compared to ZRAM.



The average hot launching speed on the three platforms is enhanced by **55.3%**, **47.5%**, and **29.6%**, respectively.

# Evaluation

- The evaluations are performed on **Google Pixel6**, **Pixel3**, and **HUAWEI P20**.
- User experience measurement: **App launching speed**, **photographic performance**, and **frame rate**.



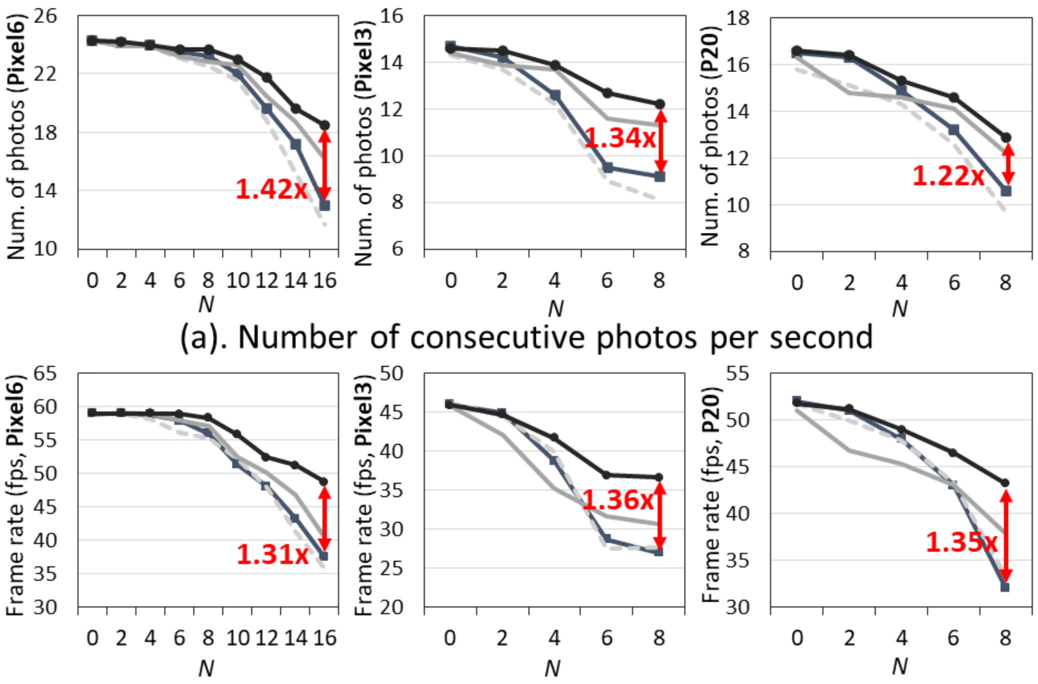
(a). Number of consecutive photos per second

(b). Frame rate of short-form videos

Benefit on photographic and frame rate

# Evaluation

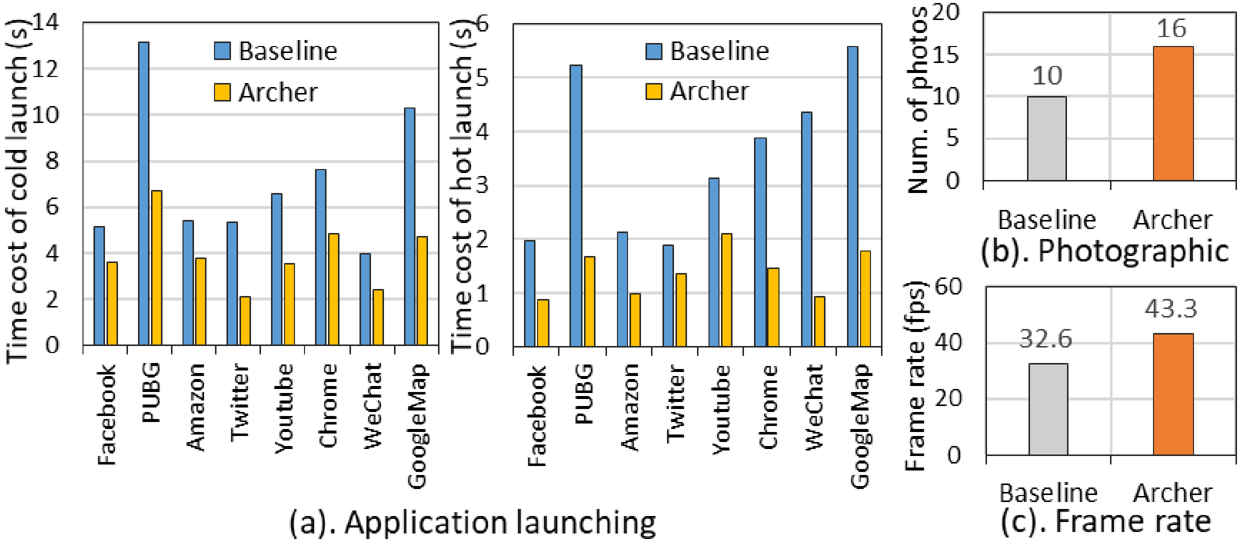
- The evaluations are performed on **Google Pixel6**, **Pixel3**, and **HUAWEI P20**.
- User experience measurement: **App launching speed**, **photographic performance**, and **frame rate**.



(a). Number of consecutive photos per second

(b). Frame rate of short-form videos

Benefit on photographic and frame rate



(a). Application launching

(b). Photographic

(c). Frame rate

- □ The tail launching latency is reduced by **44.9%** for cold launches and **60.3%** for hot launches.
- □ The worst-case performance in continuous shooting and frame rendering improves by**1.6x** and **1.3x**.

# Contents

1 Background & Motivation

2 Design

3 Evaluation

4 **Conclusion**

# Conclusion

**Problem: page compression is not suit for memory intensive applications in mobile scenarios**

The user experience degraded when suffering high memory pressure

**Key insight**: by **mining the page association rule**, large-grain compression is feasible in mobile systems

**Challenge**: collecting page footprint dynamically and in lightweight, do compression adaptively

**Archer**: **Adaptive Memory Compression with Page-Association-Rule Awareness for High-Speed Response of Mobile Devices**

Archer can improve the app launching speed by 1.55x on average over the state-of-the-art mechanisms.

usenix
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

FAST'25

# THANK YOU!

**Changlong Li**, Zongwei Zhu, Chao Wang, Fangming Liu, Fei Xu,
Edwin H.-M. Sha, and Xuehai Zhou

***Author info:*** *https://faculty.ecnu.edu.cn/_s16/lcl1/main.psp*