Intelligent Computer Architecture & Systems Research Lab

고려대학교 컴퓨터학과
김영근

March 2023

Principal Investigator

- **Young Geun Kim (김영근)**

- **Education**
  - B.S. and Ph.D. from Korea University in 2014 and 2018, respectively (***Integrated M.S. and Ph.D. graduation in 4.5 years***)

- **Professional Experience**
  - 2022 – Present: Assistant Professor, Department of Computer Science & Engineering, Korea University
  - 2020 – 2022: Assistant Professor, School of Software, Soongsil University
  - 2019 – 2020: Postdoctoral Research Associate, School of Computing, Informatics, & Decision Systems Engineering, Arizona State University
  - 2018 – 2019: Research Professor, Department of Computer Science & Engineering, Korea University

- **Research Interests**
  - System Design and Optimization for Machine Learning
  - Energy Efficiency Optimization of Mobile/Edge Systems
  - Thermal Management of Multi-scale Heterogeneous Systems

Lab Members

M.S./Ph.D. Integrated Students

Eunjin Lee

Yonglak Son

Yerin Lee

M.S. Students

Jebum Lee

Gyudong Kim

Chanhee Park

Undergraduate Interns

Chanwoo Cho

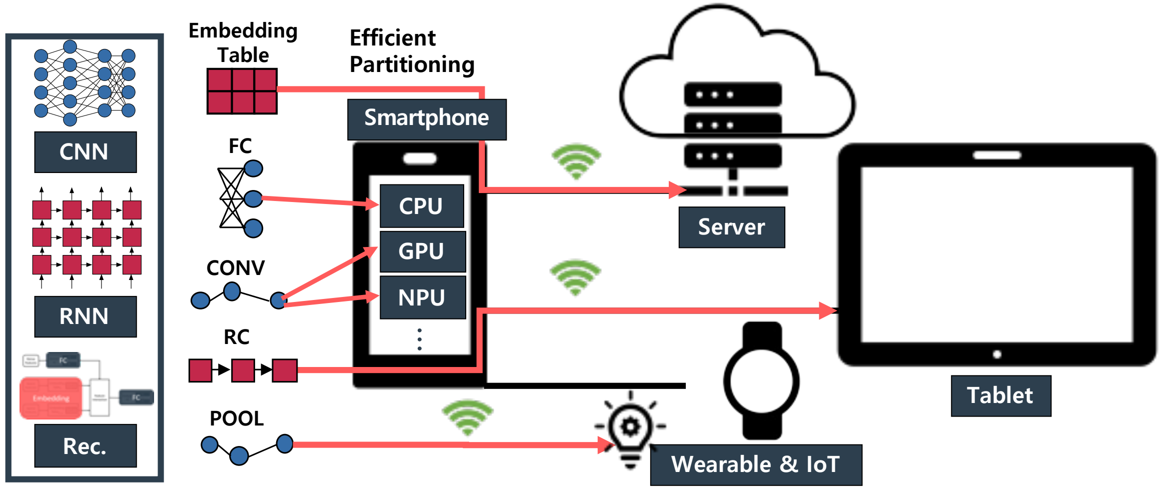
Seonghyeon Jeon

Research Interests – System Level DNN Optimization

System-level Inference Optimization

Mobile System 및 End-to-End AI Infrastructure에서 **시스템 수준의 DNN 추론 최적화** 연구

- Heterogeneous Mobile SoC에서 CNN, RNN, 등 **DNN 워크로드의 추론 성능/에너지** 최적화
- 다양한 스마트 엣지 기기에서 **System Constraint를 고려한 DNN 모델 경량화 및 정확도 평가**

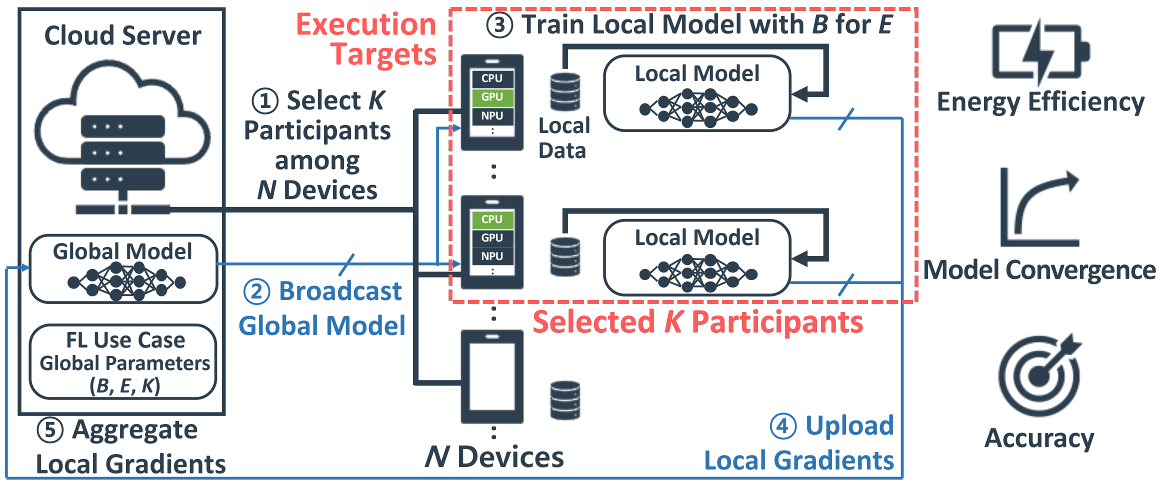


Overview of System-level DNN Optimization

Secure Deep Learning System

Mobile System의 특성을 **고려한 연합 학습 성능 및 에너지 효율 최적화** 연구

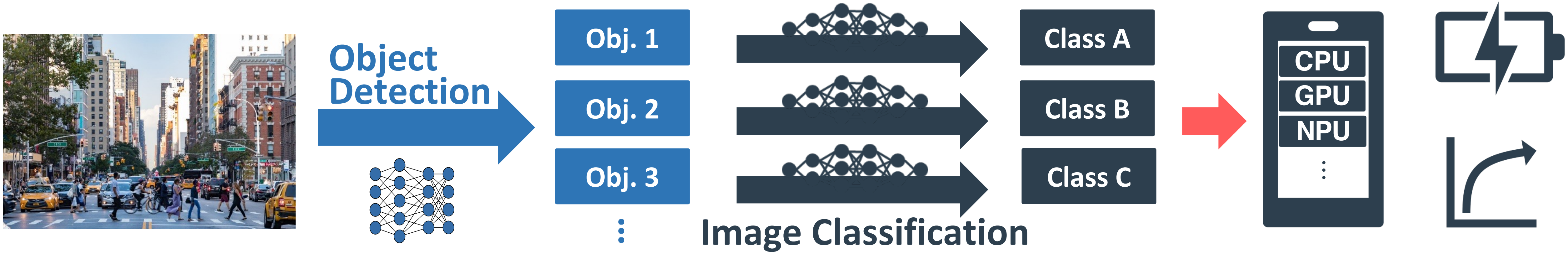
- **연합 학습에서 Mobile System의 특성을 고려**한 시스템 최적화
- **System-level Noise가 연합 학습 데이터 및 Convergence 및 에너지 효율에 미치는 영향 분석** 및 이를 고려한 FL 최적화



Overview of System-level Optimization for FL

System-Level DNN Optimization 관련 세계 최고 수준의 Publication 및 연구 경험 보유

Publications on System-Level DNN Optimization



Object Detection

Obj. 1

Obj. 2

Obj. 3

⋮

Image Classification

Class A

Class B

Class C

CPU

GPU

NPU

⋮

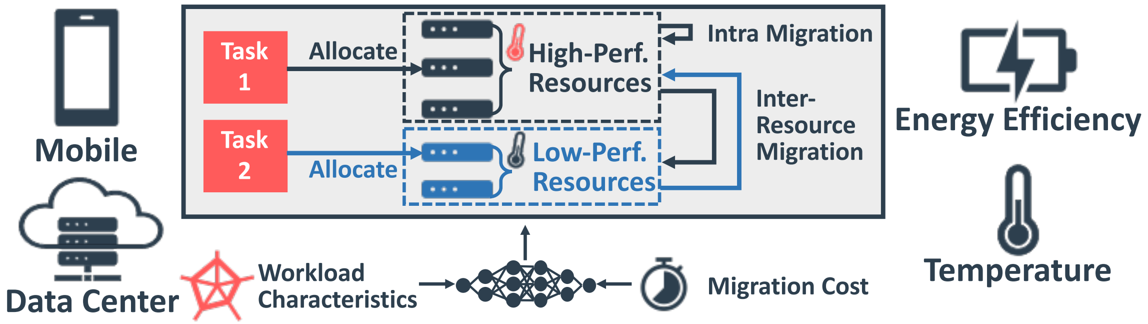
[MICRO ’20 – BK CS IF: 4] Y. G. Kim and C. –J. Wu, **“AutoScale: Energy Efficiency Optimization for Stochastic Edge Inference Using Reinforcement Learning,”** ***IEEE/ACM International Symposium on Microarchitecture*** (**Top-tier Conference**), 2020.
[MICRO ‘21 – BK CS IF: 4] Y. G. Kim and C. –J. Wu, **“AutoFL: Enabling Heterogeneity-Aware Energy Efficient Federated Learning,”** ***IEEE/ACM International Symposium on Microarchitecture*** (**Top-tier Conference**), 2021.
[IISWC ’22 – BK CS IF: 1] Y. G. Kim and C. –J. Wu, **“FedGPO: Heterogeneity-Aware global Parameter Optimization for Efficient Federated Learning,”** ***IEEE International Symposium on Workload Characterization***, 2021.
[TC ’22] S. Heidari, M. Ghasemi, **Y. G. Kim,** C. –J. Wu, and S. Vrudhula, **“CAMDNN: Content-Aware Mapping of a Network of Deep Neural Networks on Edge MPSoCs,”** ***IEEE Transactions on Computers***, 2022.

Research Interests – Intelligent System Architecture

Intelligent System Architecture

Machine Learning 기반 시스템 성능/에너지/발열 최적화 및 안전한 시스템 구조 설계

- **Machine Learning (e.g., 강화 학습, 모방 학습, 등) 기반 시스템 성능/에너지/발열 최적화** 연구
- Mobile-Network-Cloud System Infrastructure에서 **외부 환경 및 Runtime Variance를 고려한 적응적 에너지/발열 관리 기술** 개발
- **안전한 Machine Learning 수행을** 위한 System Architecture 설계

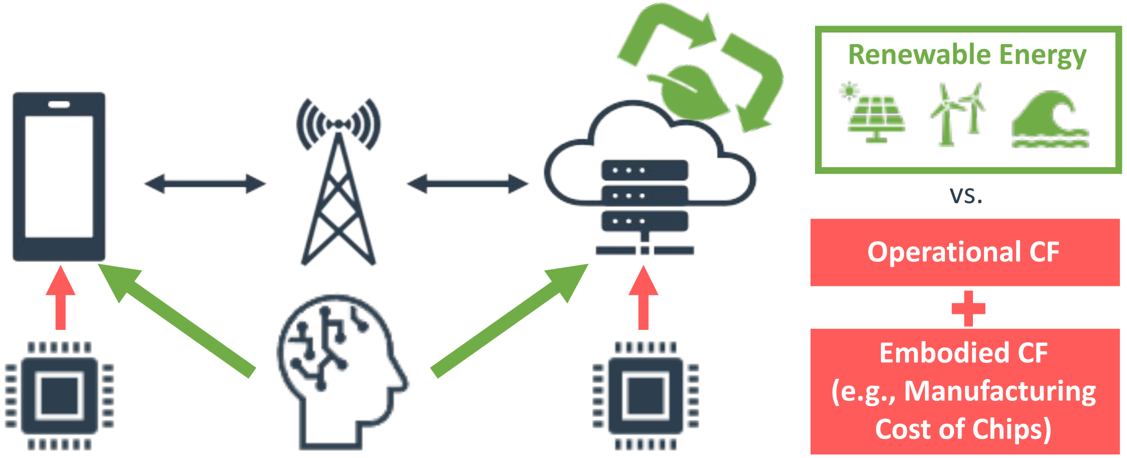


ML-based Intelligent System Architecture

Systems for Sustainable Deep Learning

End-to-End **AI Infrastructure의 Carbon Footprint 평가 및 시스템 수준 최적화**

- **Mobile, Network, Data Center의 Carbon Footprint 예측 모델 구축** 및 Global Carbon Footprint 평가
- End-to-End AI Infrastructure에서 **DNN 추론의 Global Carbon Footprint 평가 및 최적화를 위한 스케줄링** 기술 개발
- 연합 학습에서 DNN 워크로드 특성을 고려한 CF 평가 및 최적화

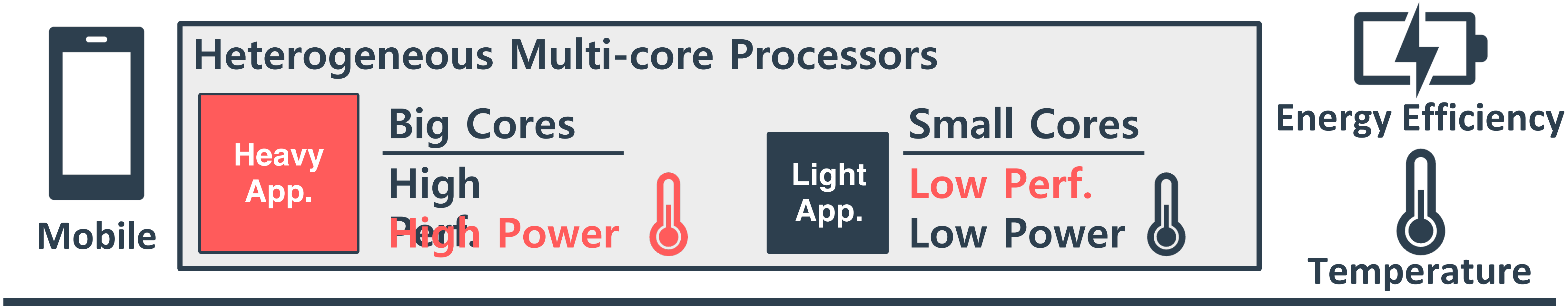


Overview of Sustainable Deep Learning System

지능형 시스템 설계 관련 세계 최고 수준 Publication 및 연구 실적 보유

Publications on Sustainable Computing

Energy/Temperature Management in Single Mobile Devices



Mobile

Heterogeneous Multi-core Processors

Heavy App.

Big Cores

High Perf.

High Power

Light App.

Small Cores

Low Perf.

Low Power

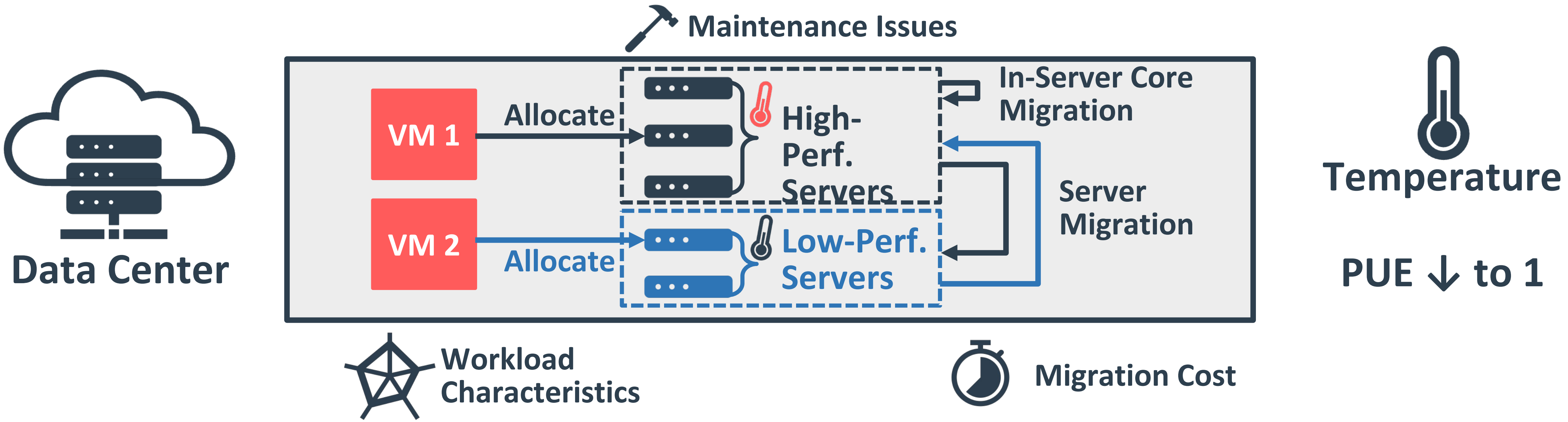
Energy Efficiency

Temperature

[TC '20] Y. G. Kim, M. Kim, J. Kong, and S. W. Chung, **"An Adaptive Thermal Management Framework for Heterogeneous Multi-Core Processors,"** ***IEEE Transactions on Computers***, 2020.
[TPDS '18] Y. G. Kim, J. Kong, and S. W. Chung, **"A Survey on Recent OS-level Energy Management Techniques for Mobile Processing Units,"** ***IEEE Transactions on Parallel and Distributed Systems***, 2018.
[TC '17] Y. G. Kim, M. Kim, and S. W. Chung, **"Enhancing Energy Efficiency of Multimedia Applications in Heterogeneous Mobile Multi-Core Processors,"** ***IEEE Transaction on Computers***, 2017. (***monthly featured paper***)
[DATE' 15 – BK CS IF: 2] Y. G. Kim, M. Kim, J. M. Kim, S. W. Chang, **"M-DTM: Migration-based Dynamic Thermal Management for Heterogeneous Mobile Multi-Core Processors,"** ***Design, Automation, and Test in Europe Conference***, 2015.

Publications on Sustainable Computing

Temperature Management in Data Centers

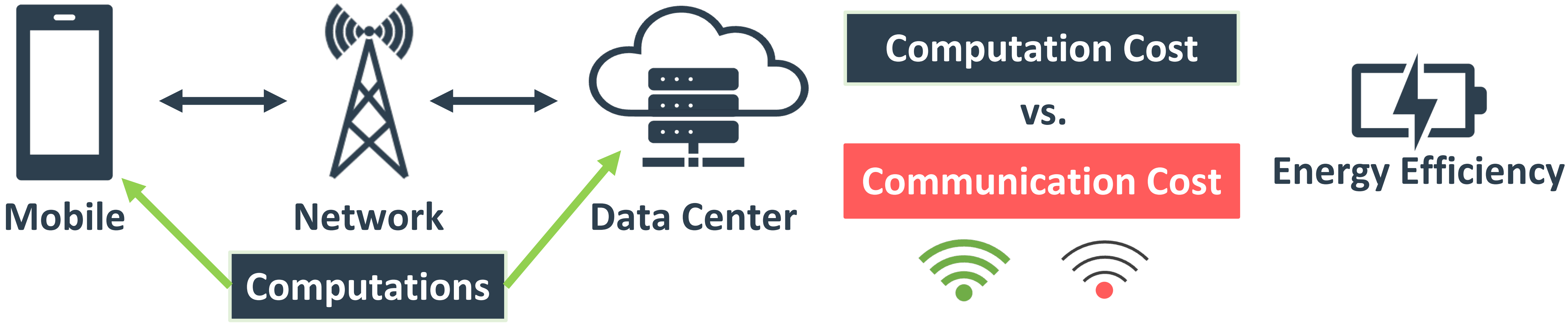


[JSA '21] **Y. G. Kim,** S. Y. Kim, S. H. Choi, and S. W. Chung, **"Thermal-aware Adaptive VM Allocation Considering Server Locations in Heterogeneous Data Centers,"** ***Journal of Systems Architecture***, 2021.

[ISLPED '19 – BK CS IF: 1] Y. G. Kim, J. I. Kim, S. H. Choi, S. Y. Kim, and S. W. Chung, **"Temperature-aware Adaptive VM Allocation in Heterogeneous Data Centers,"** ***IEEE/ACM International Symposium on Low Power Electronics and Design***, 2019.

Publications on Sustainable Computing

Computation Partitioning between Mobile and Data Centers



Mobile

Network

Data Center

Computations

Computation Cost

vs.

Communication Cost

Energy Efficiency

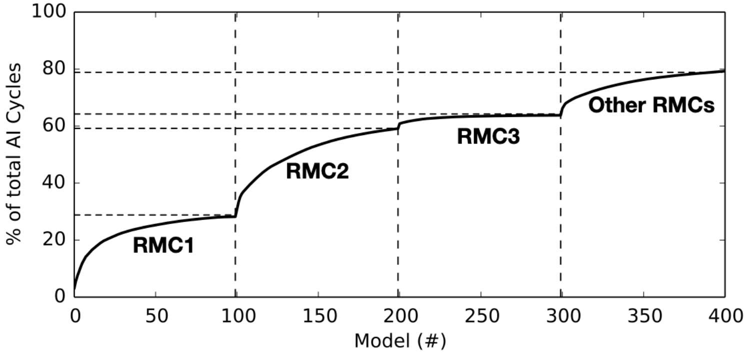
[TC ‘20] Y. G. Kim, Y. S. Lee, and S. W. Chung, **“Signal Strength-aware Adaptive Offloading with Local Image Preprocessing for Energy Efficient Mobile Devices,”** ***IEEE Transactions on Computers***, 2020.

[ISLPED ’17 – BK CS IF: 1] Y. G. Kim and S. W. Chung, **“Signal Strength-aware Adaptive Offloading for Energy Efficient Mobile Devices,”** ***IEEE/ACM International Symposium on Low Power Electronics and Design***, 2017.

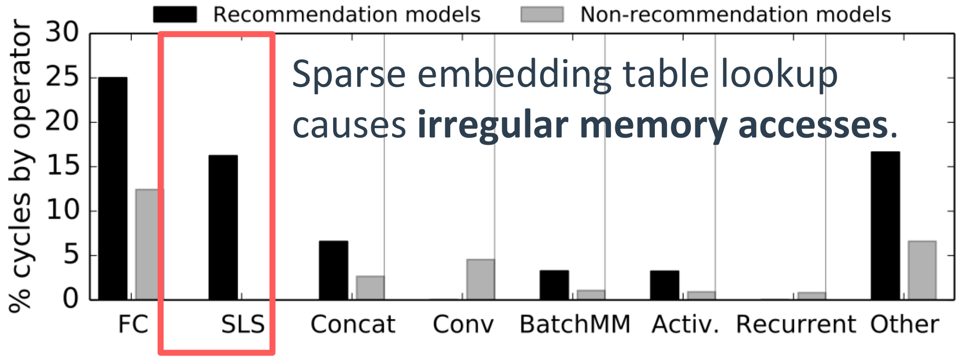
On-going Work

System Optimization for Efficient Deep Learning – Supported by NRF (147,360,000 KRW / Year)

- **Enabling Edge Recommendation (w/Federated Learning) – [HPCA 2024 - BK IF 4]**



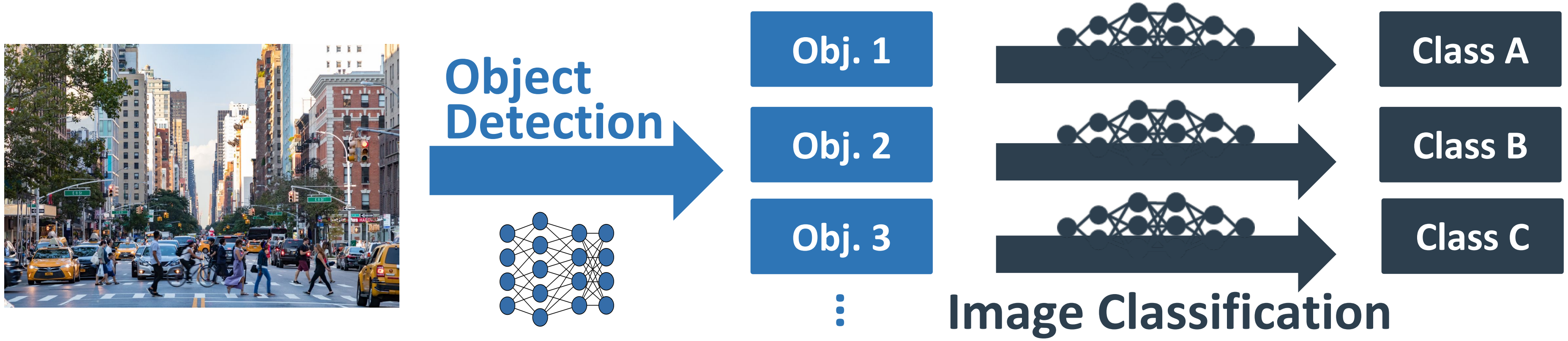
Recommendation models consume **79% of AI inference cycle** in Facebook’s data centers.



Sparse embedding table lookup causes **irregular memory accesses.**

Recommendation models have **different characteristics compared to CNNs and RNNs** due to embedding tables.

- **Optimizing Network of DNN Models at the Edge – [MICRO 2023 – BK IF 4]**



Object Detection

Obj. 1

Obj. 2

Obj. 3

⋮

Class A

Class B

Class C

Image Classification

On-going Work

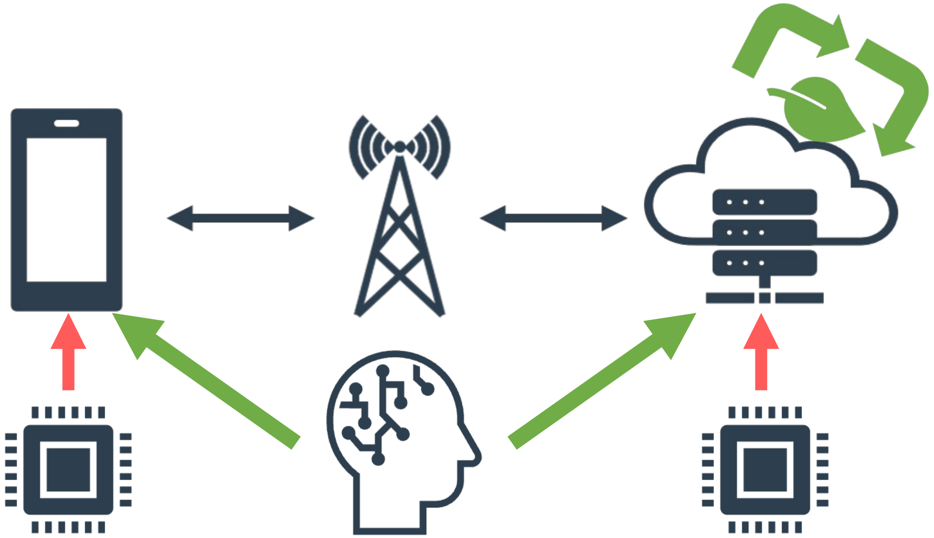
Sustainable Deep Learning – Collaboration with Meta AI Research

- Carbon Footprint-Aware Deep Learning – [ASPLOS 2024 – BK IF 4]

UNIVERSITY OF MICHIGAN

HARVARD UNIVERSITY

FACEBOOK AI



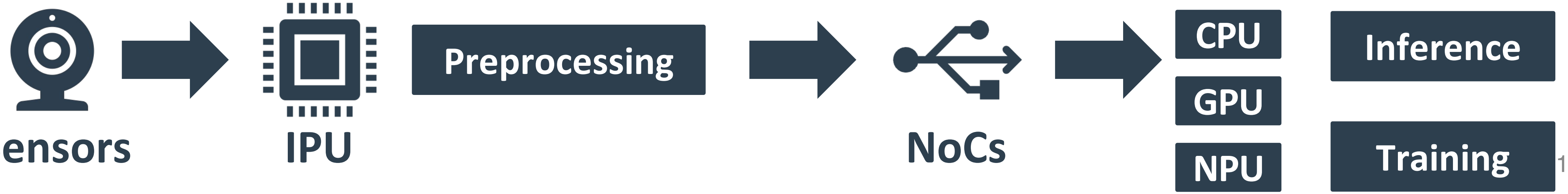
Energy/Latency/Carbon Footprint Trade-offs between Mobile and Data Centers (w/ Network)

+ Available Renewable Energy
and Amortization of Manufacturing Cost

- Analysis of End-to-End AI Pipeline for Efficient FL – [NeurIPS 2023 – BK IF 4]

ASU

FACEBOOK AI



Sensors

IPU

Preprocessing

NoCs

CPU

GPU

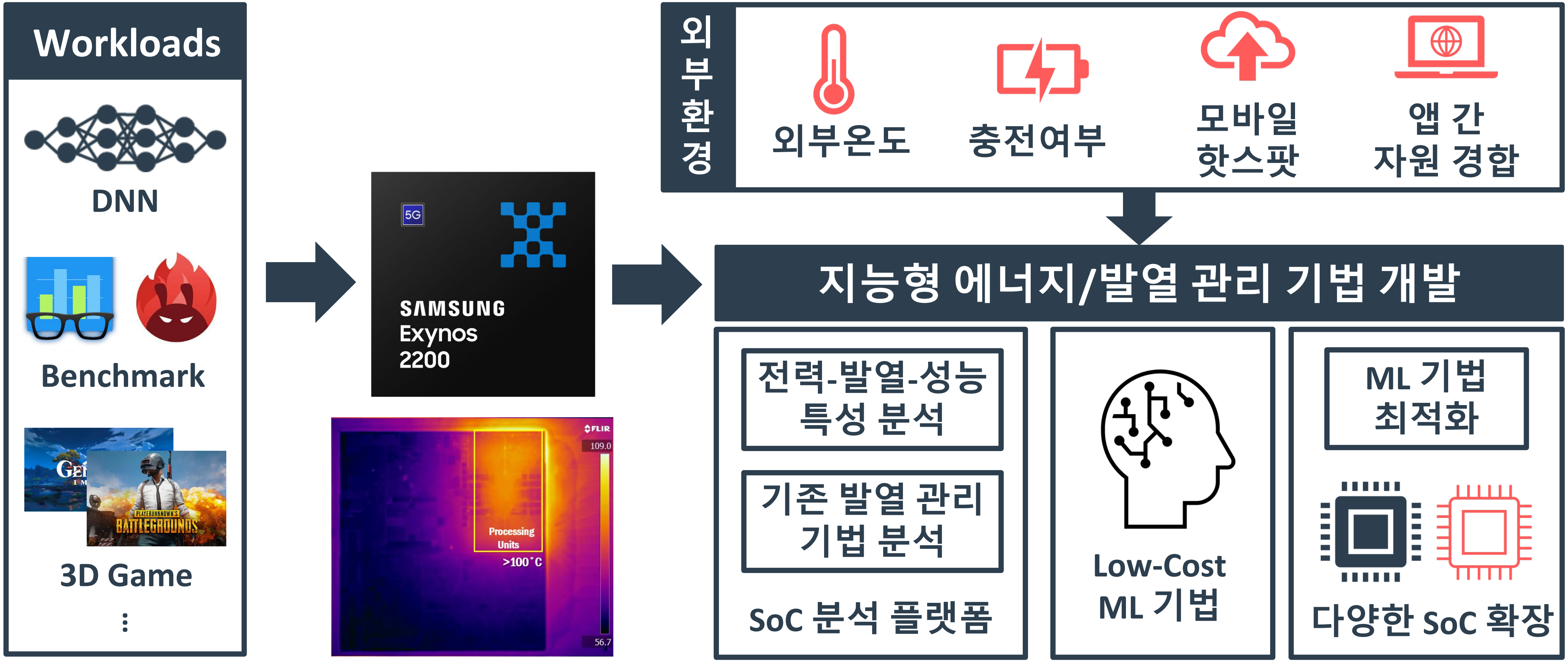
NPU

Inference

Training

On-going Work

Intelligent Energy/Temperature Management for Mobile SoCs



Workloads

DNN

Benchmark

3D Game

⋮

외부환경

외부온도

충전여부

모바일 핫스팟

앱 간 자원 경합

지능형 에너지/발열 관리 기법 개발

전력-발열-성능 특성 분석

기존 발열 관리 기법 분석

SoC 분석 플랫폼

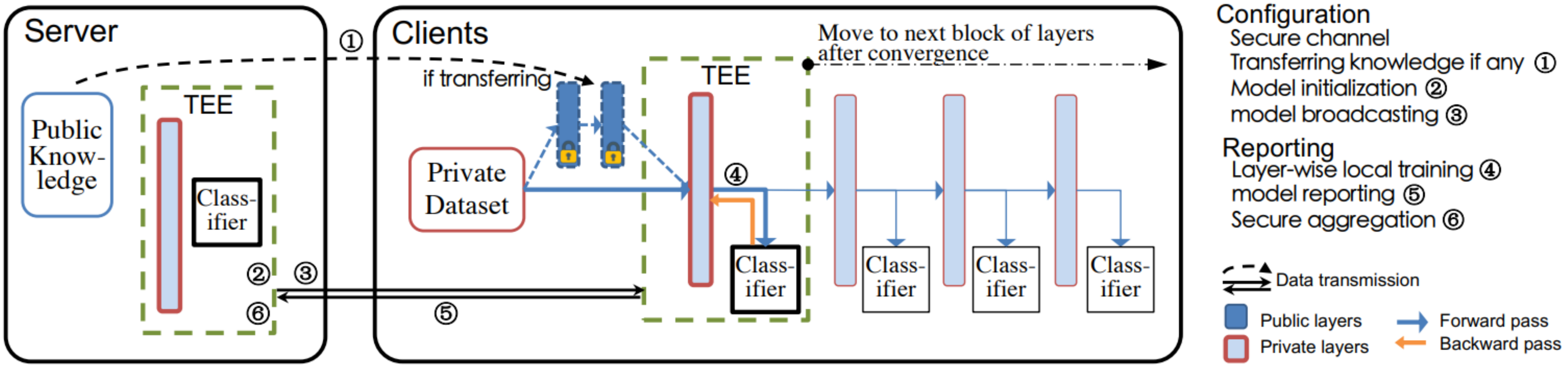
Low-Cost ML 기법

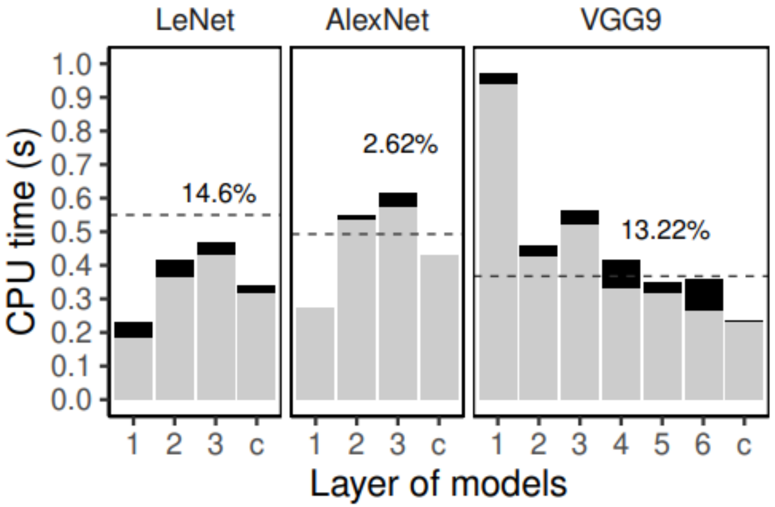
ML 기법 최적화

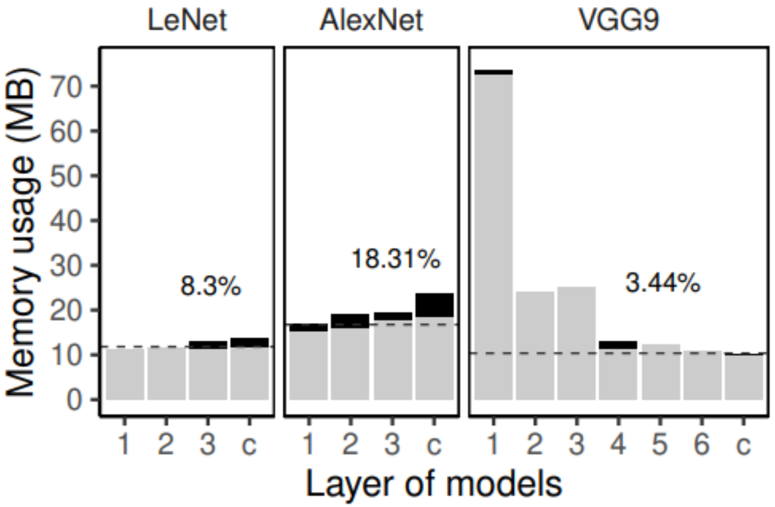
다양한 SoC 확장

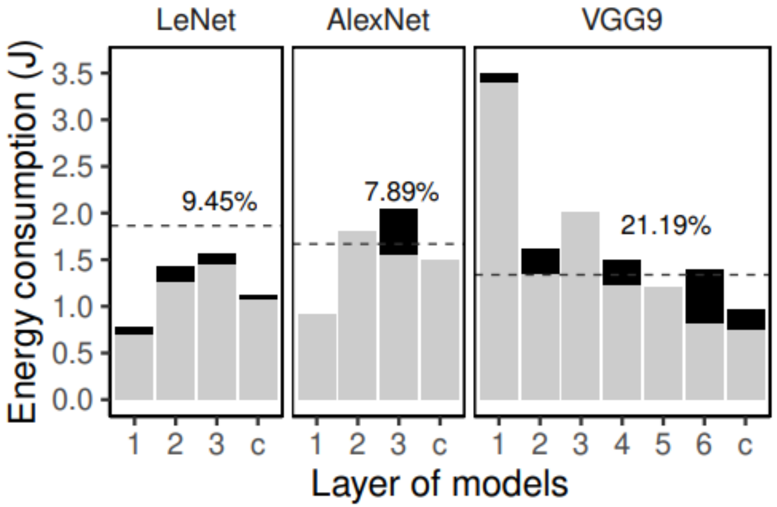
On-going Work

Privacy-Preserving Acceleration of FL









Collaborators

FACEBOOK AI

Carole-Jean Wu *Hsien-Hsin S. Lee*

YONSEI UNIVERSITY 1885 연세대학교

Dokyung Song

Haehyun Cho *Young-ho Gong*

HARVARD UNIVERSITY

David Brooks *Gu-Yeon Wei*

CORNELL TECH

Udit Gupta

UNIVERSITY OF MICHIGAN

Valeria Bertacco

Massachusetts Institute of Technology

Vivienne Sze

ASU

Sarma Vrudhula *Jeff J. Zhang* *Mehdi Ghasemi* *Soroush Heidari*

QUALCOMM SAMSUNG LG ETRI

Aaron Lamb *Taekki Kim* *Chul Keon Jin* *Jaehoon Chung* ...

Prospective Members

- **We are recruiting self-motivated students!**

- **Related Courses**
  - Computer Architecture
  - Computer System Design
  - Operating Systems
  - Artificial Intelligence
  - Deep Learning

- **Skills**
  - Programming Languages (C/C++/Python)

Thank you.

E-mail: younggeun_kim@korea.ac.kr
Lab: https://casl.korea.ac.kr