BlastNet: Exploiting Duo-Blocks for Cross-Processor Real-Time DNN Inference

Neiwen Ling[1], Xuan Huang[1], Zhihe Zhao[1], Guan Nan[2] , Zhenyu Yan[1], Guoliang Xing[1]

[1]The Chinese University of Hong Kong,

[2]City University of Hong Kong

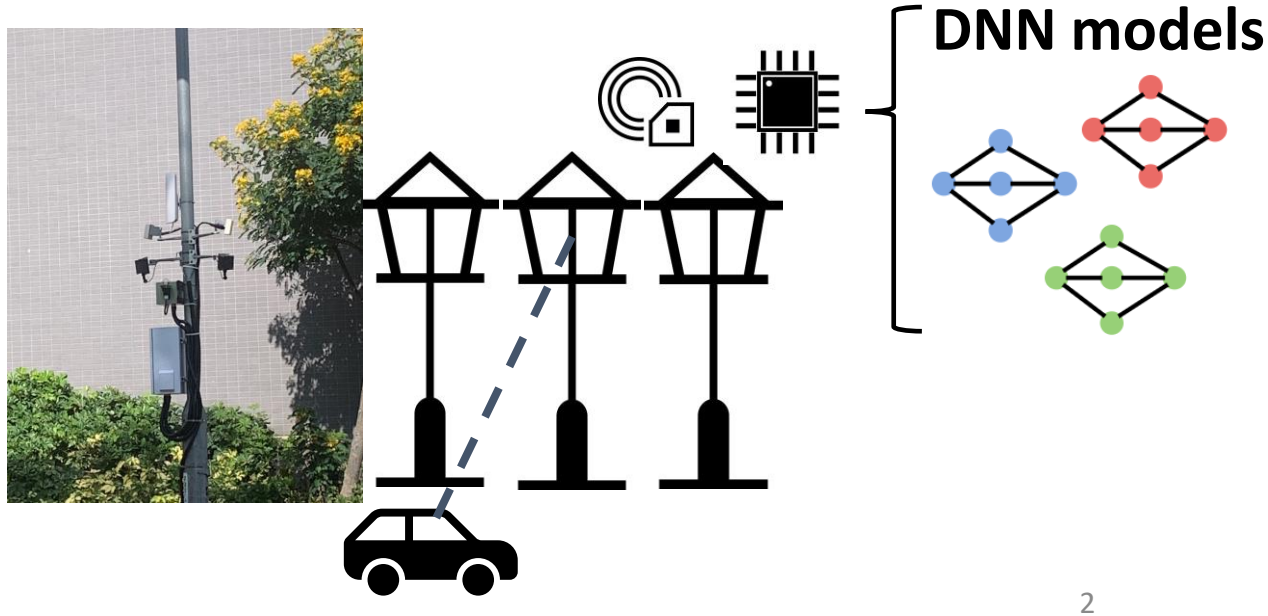
Real-time Deep Learning on the Edge

Autonomous driving
(e.g., Tesla)

Image source: tukuppt

Image source: Bernard Marr

Smart roadside infrastructure
(CUHK smart lampposts)

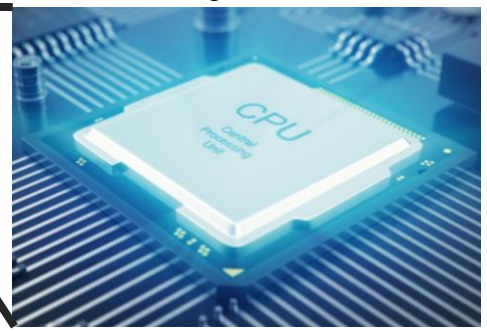


Heterogeneous Processors on Edge Platforms

Image source: EDN Taiwan

Edge Platform

Image source: How-To Geek



CPU

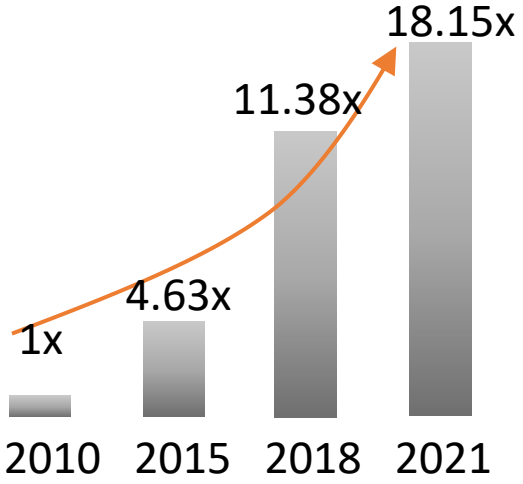
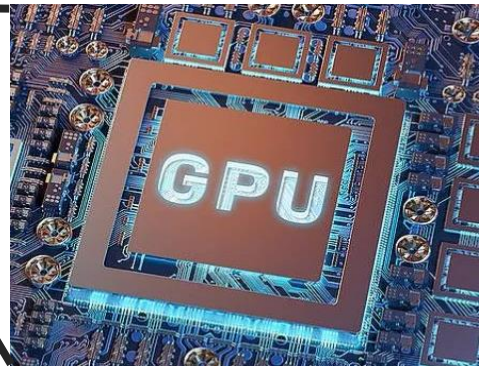
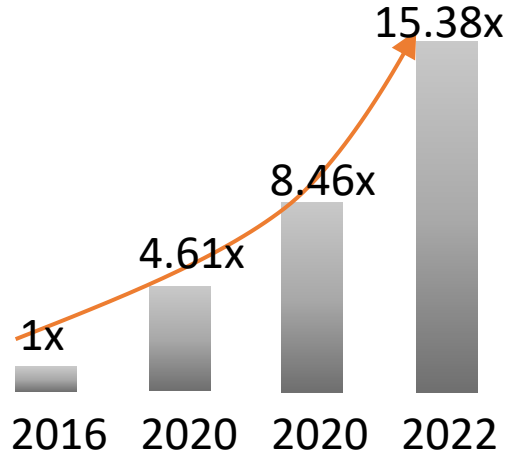


Image source: OFweek



GPU



Compute Power

Untapped Compute Power

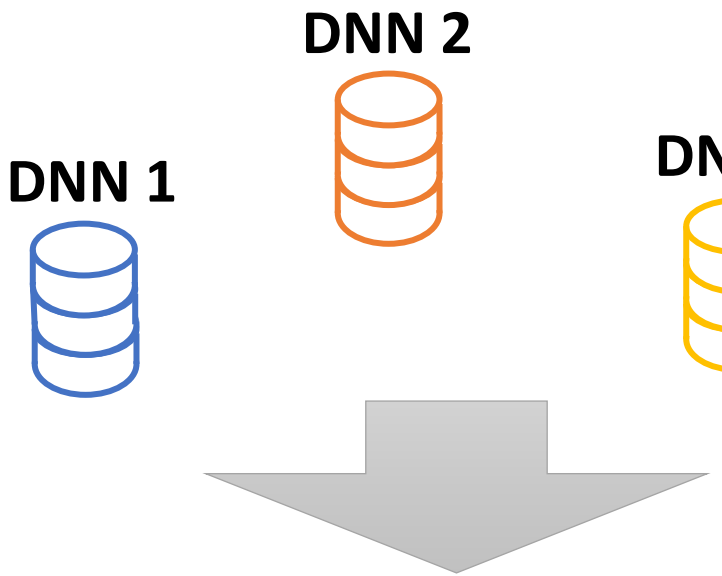
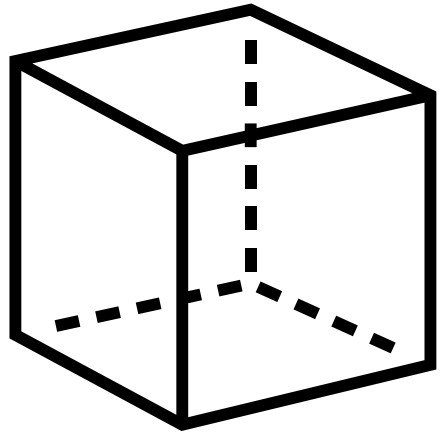
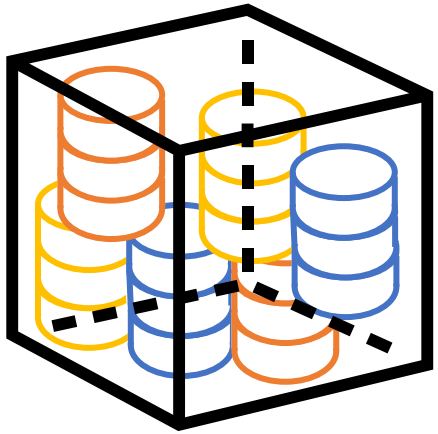


Image source: manpingou



CPU



GPU

Image source: manpingou

How to utilize both edge CPU and GPU for running multiple DNNs efficiently?

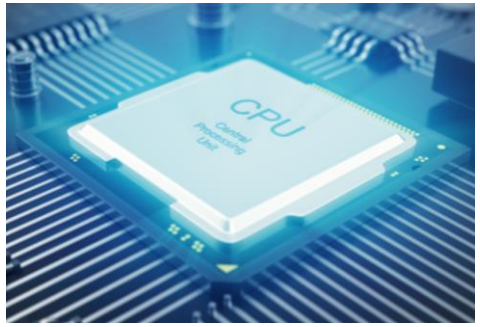


Image source: How-To Geek

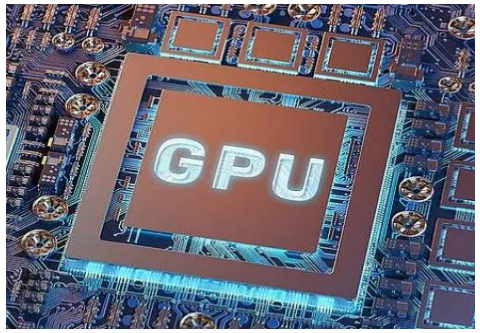
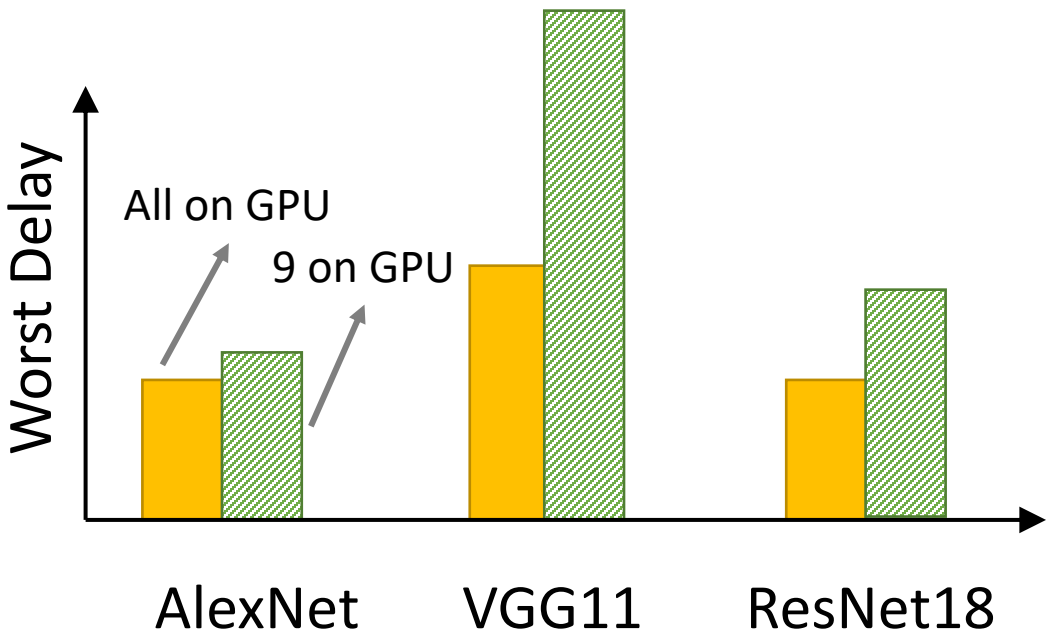
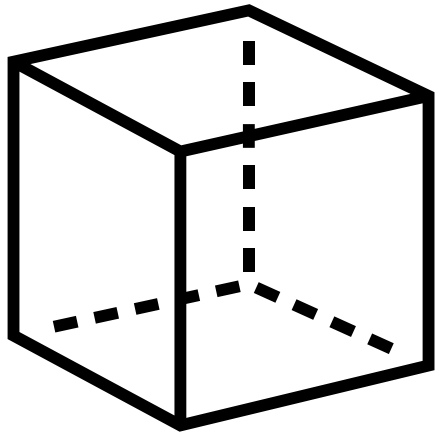


Image source: OFweek

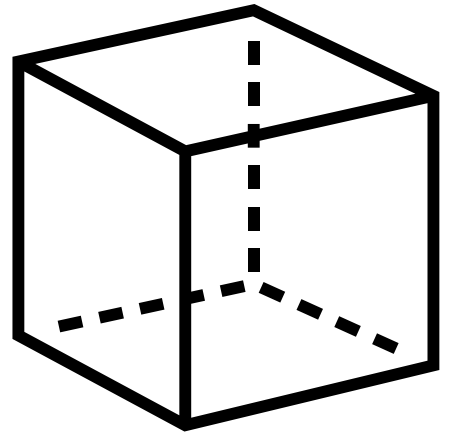
Offloading models?



DNN models



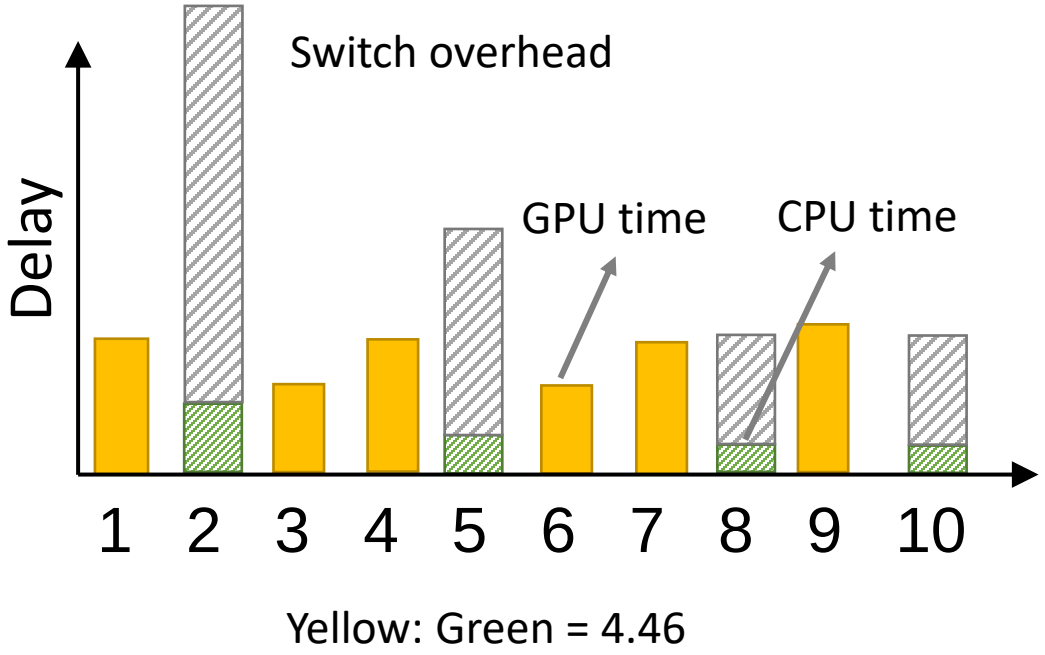
CPU



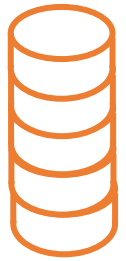
GPU

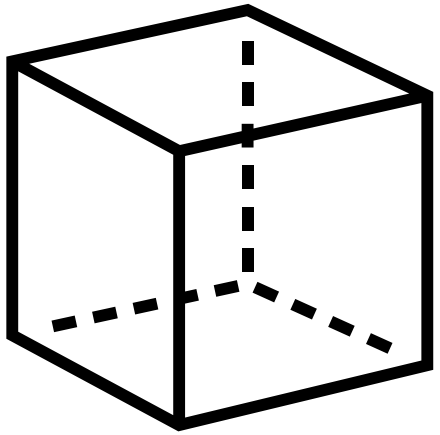
Too coarse for efficient resource utilization

Offloading layers?

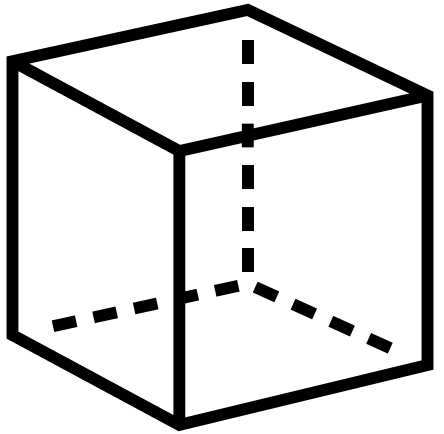


DNN layers





CPU

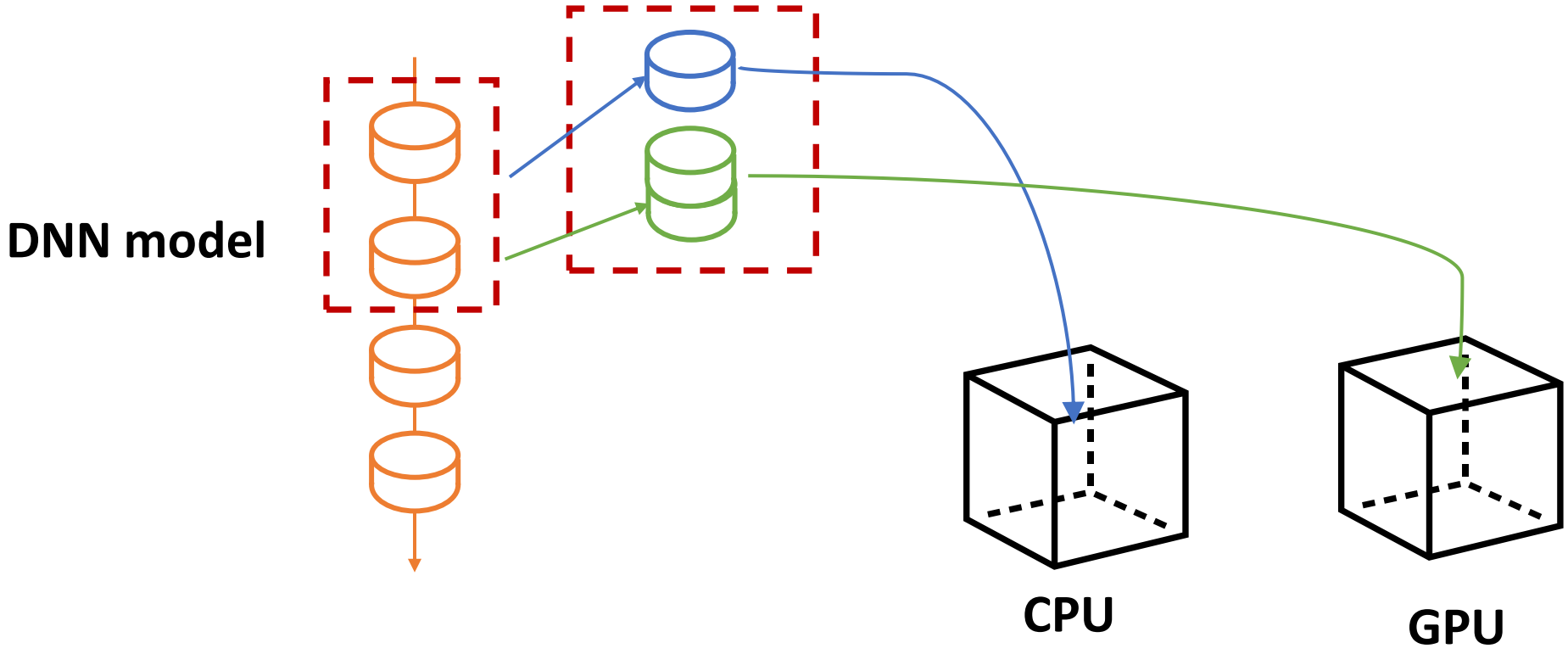


GPU

Significant idle time and switching delay

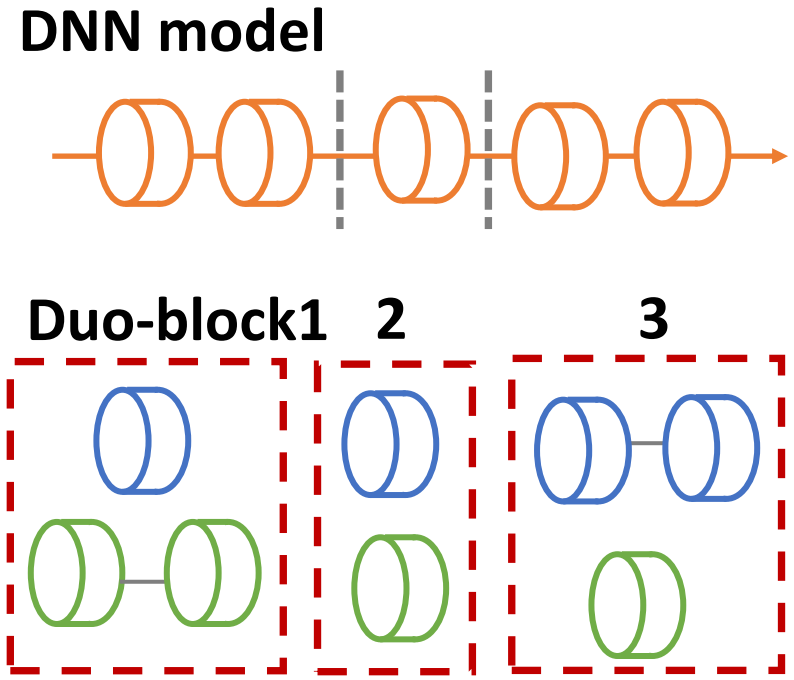
BlastNet: Block-Level model optimizAtion and Scheduling sysTem

Duo-Block
a new model
inference abstraction

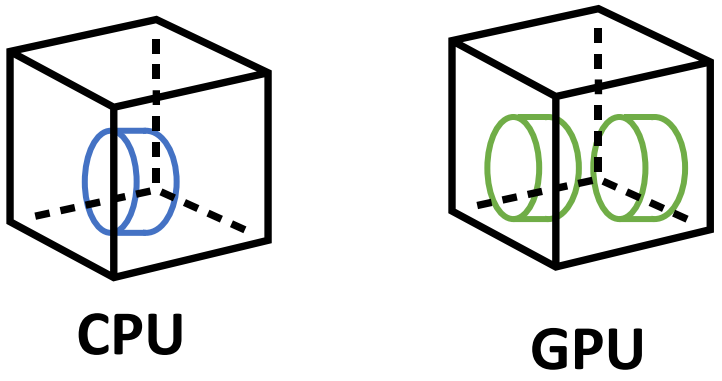


BlastNet: Block-Level model optimizAtion and Scheduling sysTem

- **Duo-block generation**

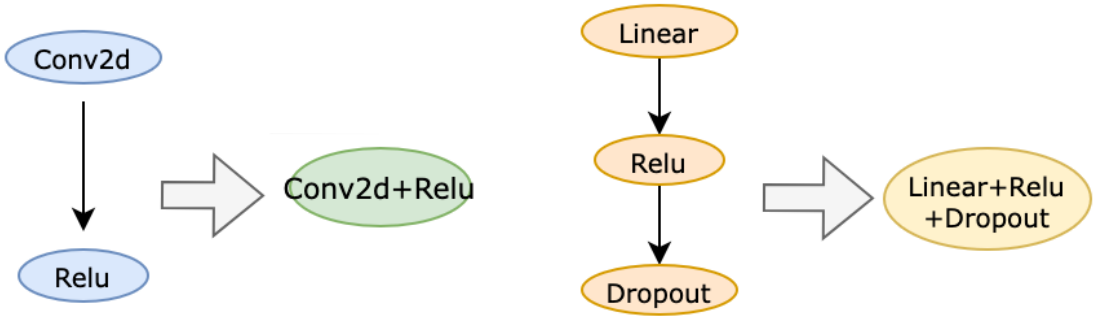


- **Dynamic** cross-processor DNN **scheduling**

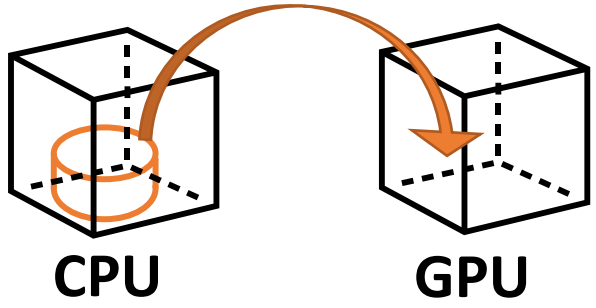


Duo-block Generation

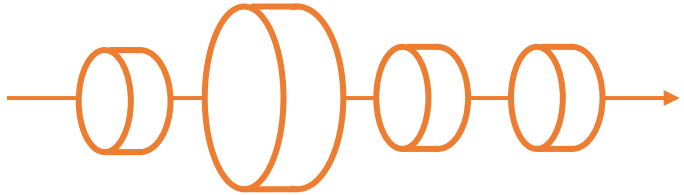
1. Preserve operator fusion



2. Switching < Execution?

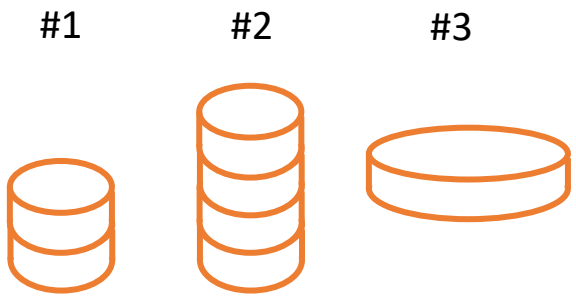


3. Optimizing bottleneck blocks

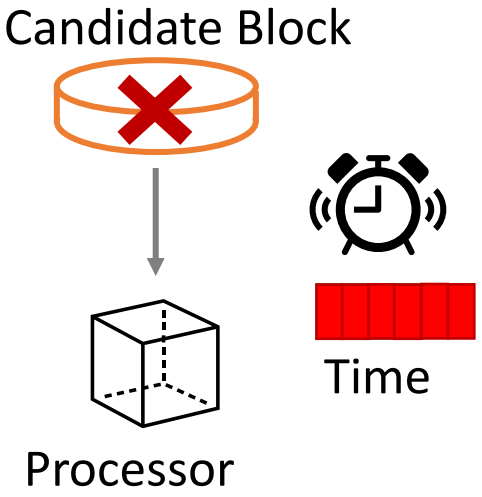


Duo-block Generation with Neural Architecture Search

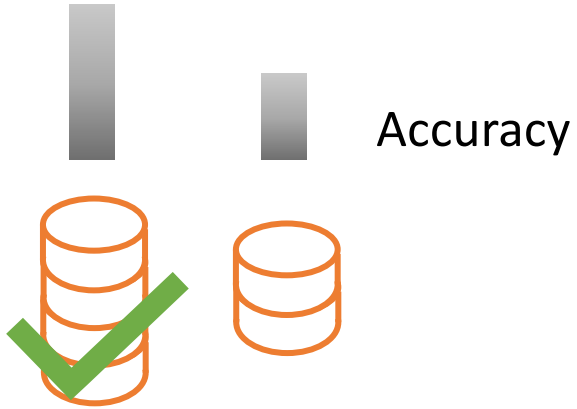
1. Candidate blocks generation with processor-friendly operators



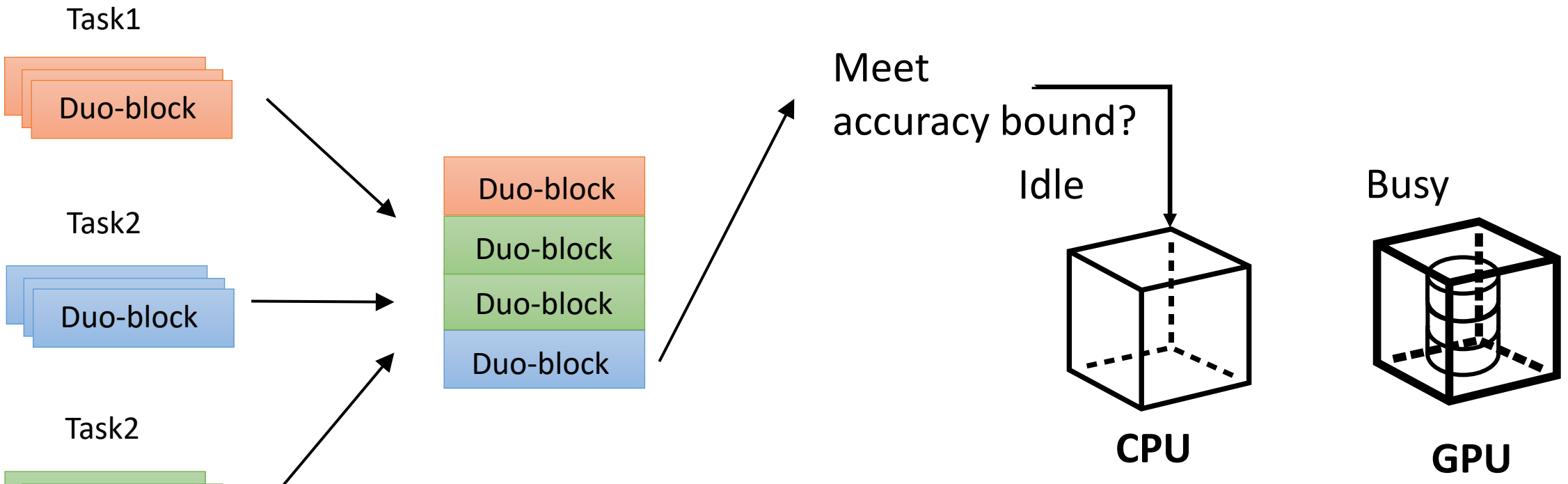
2. Optimize search space based on profiling



3. Neural Architecture Search



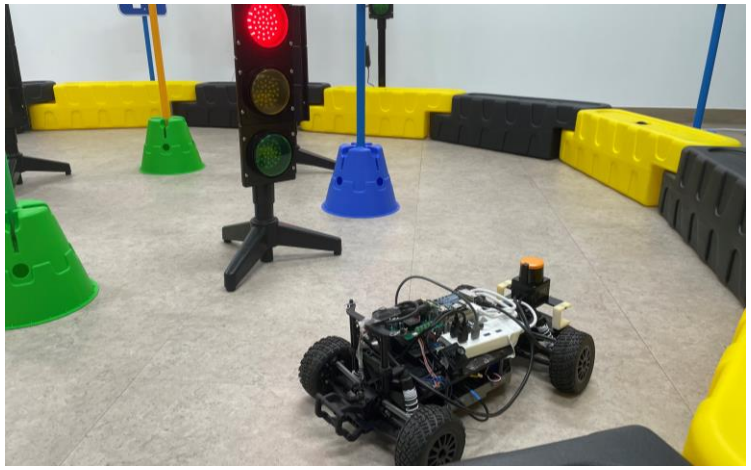
Dynamic Cross-Processor Scheduling

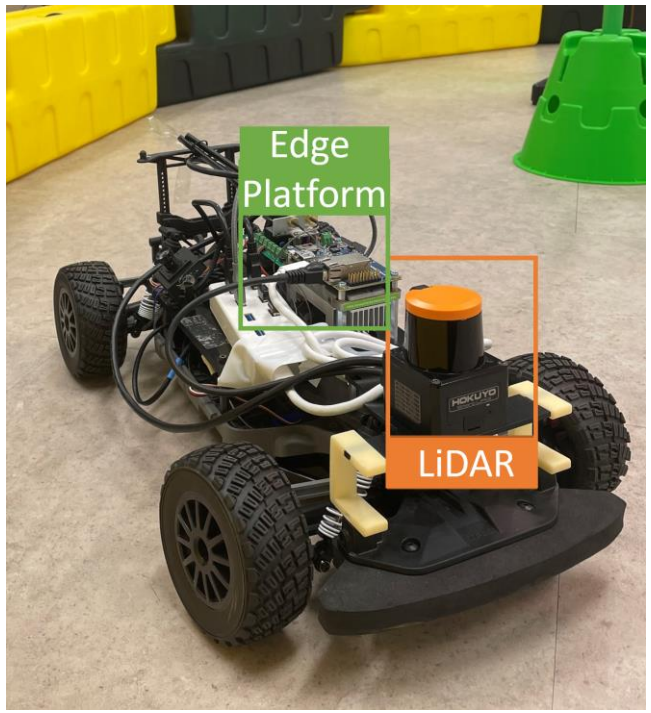


- Prioritize each duo-block based on the urgency of its model
- Choose the execution processor for each duo-block

Experiment Setup

- 3 edge platforms including an autonomous driving car





- 3 types of DL tasks
  5 DNN models
  3 datasets

Image Classification

Image source: Hertz

Object Detection

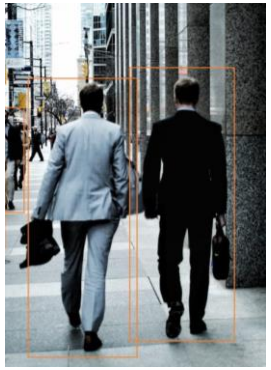
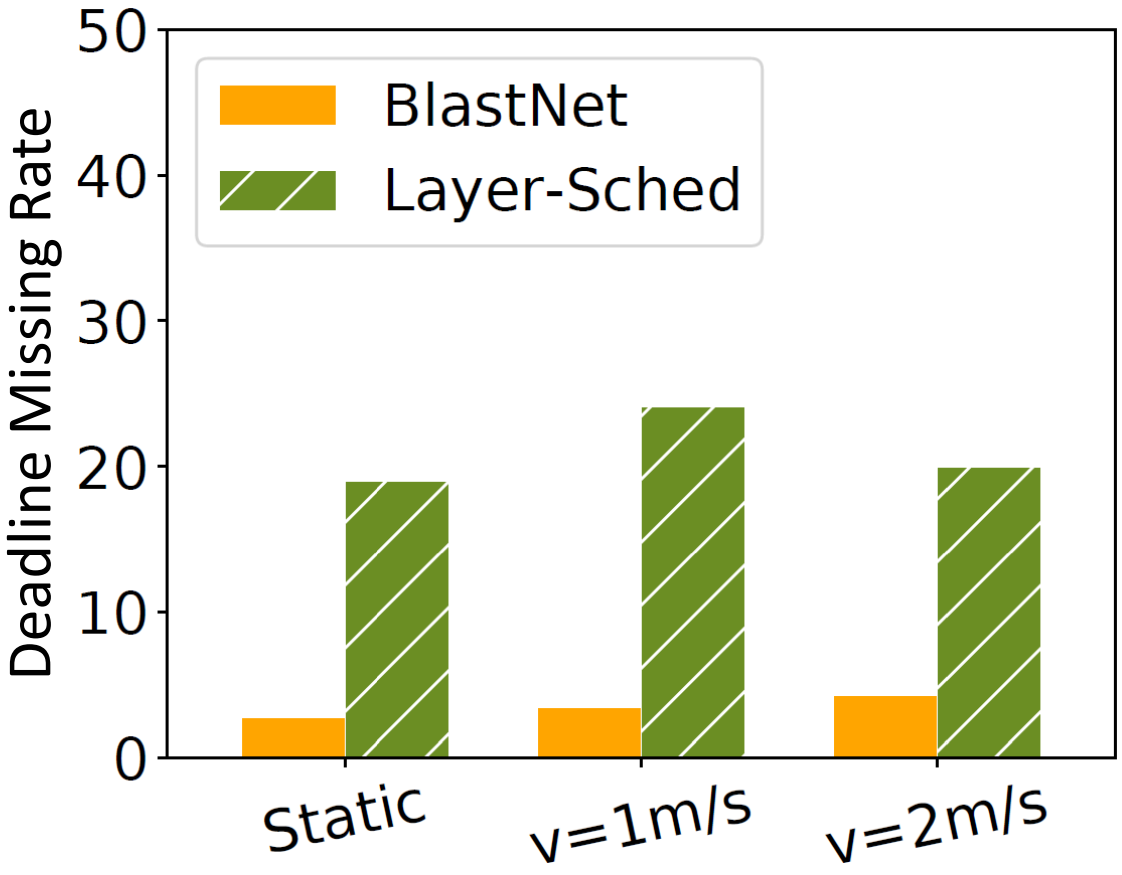


Image source: Innovtrics

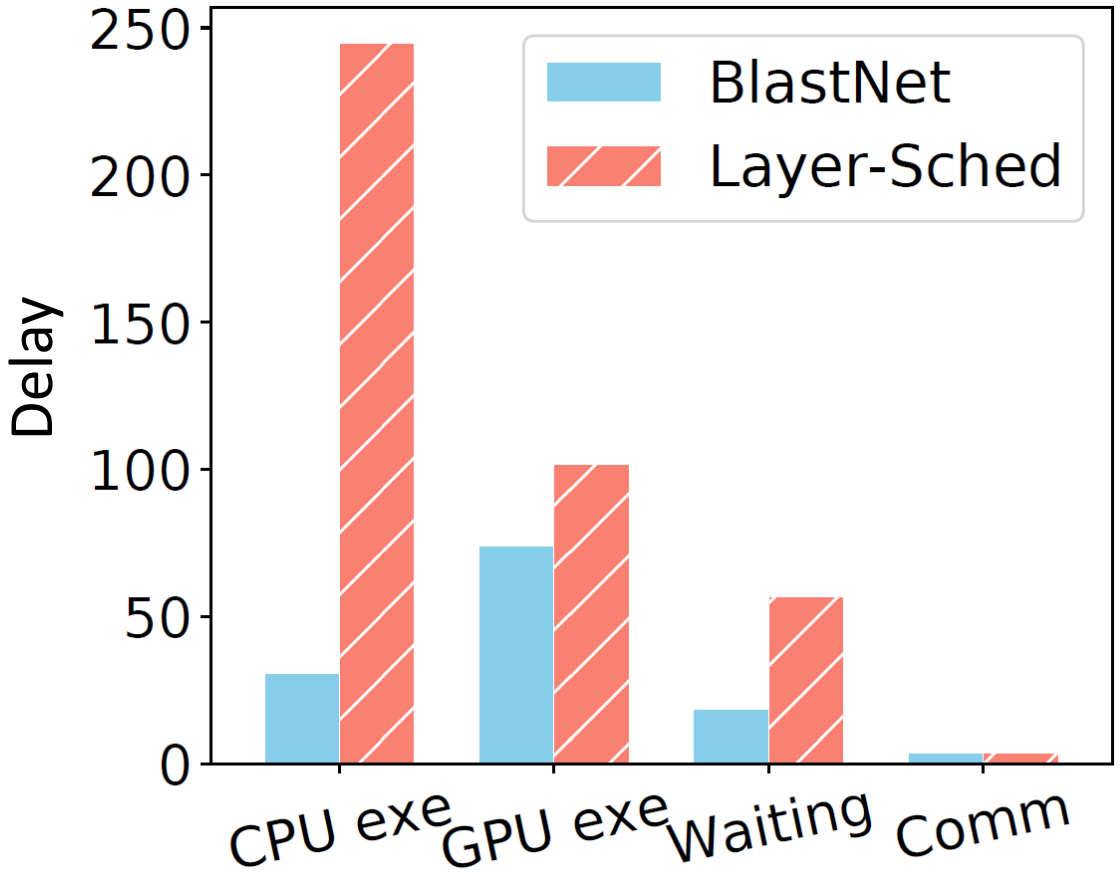
Sign Recognition

Image source: Datangxs

Experimental results: End-to-end system evaluation



< 5% deadline missing rate



Reduced CPU/GPU execution and waiting time

Conclusion & Future Work

- **BlastNet:** novel block-level model optimization and cross-processor scheduling
  - Duo-block generation
  - A dynamic block-level DNN scheduler
- **Future Work**
  - Scalability to new heterogeneous platforms
  - Uncertain workloads

Thanks for Listening!

- *Check our paper:* BlastNet: Exploiting Duo-Blocks for Cross-Processor Real-Time DNN Inference
- *Authors:* Neiwen Ling, Xuan Huang, Zhihe Zhao, Guan Nan, Zhenyu Yan, Guoliang Xing
- *Website:* http://aiot.ie.cuhk.edu.hk