ISCAS 中国科学院软件研究所学术年会'2022
暨计算机科学国家重点实验室开放周

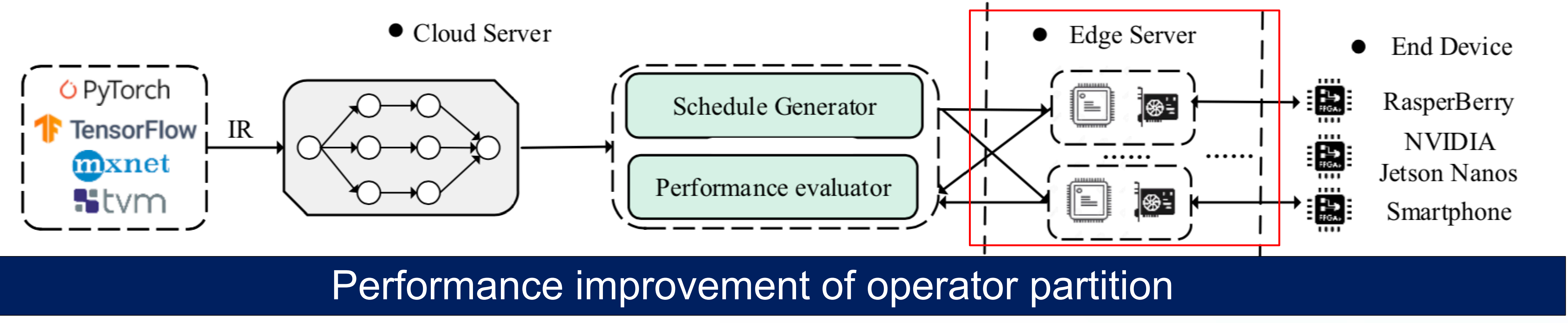
学术论文 VEE 22

EOP: Efficient Operator Partition for Deep Learning Inferece on Edge Servers

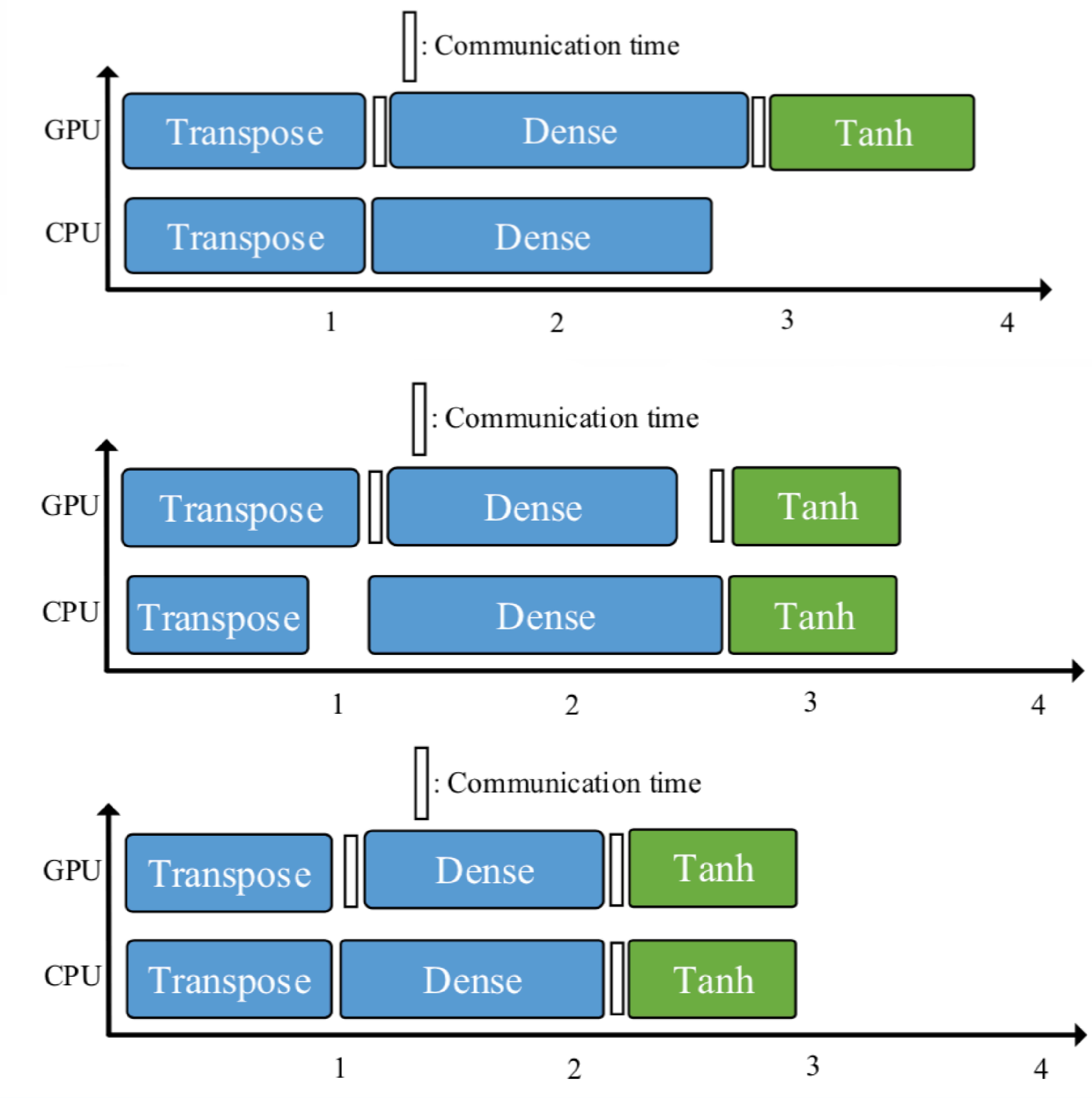
Yuanjia Xu, Heng Wu, Wenbo Zhang, Yi Hu

Contact: Yi Hu, huyi19@otcaix.iscas.ac.cn, 13051219797

Deployment pipeline of DL inference on edge environment



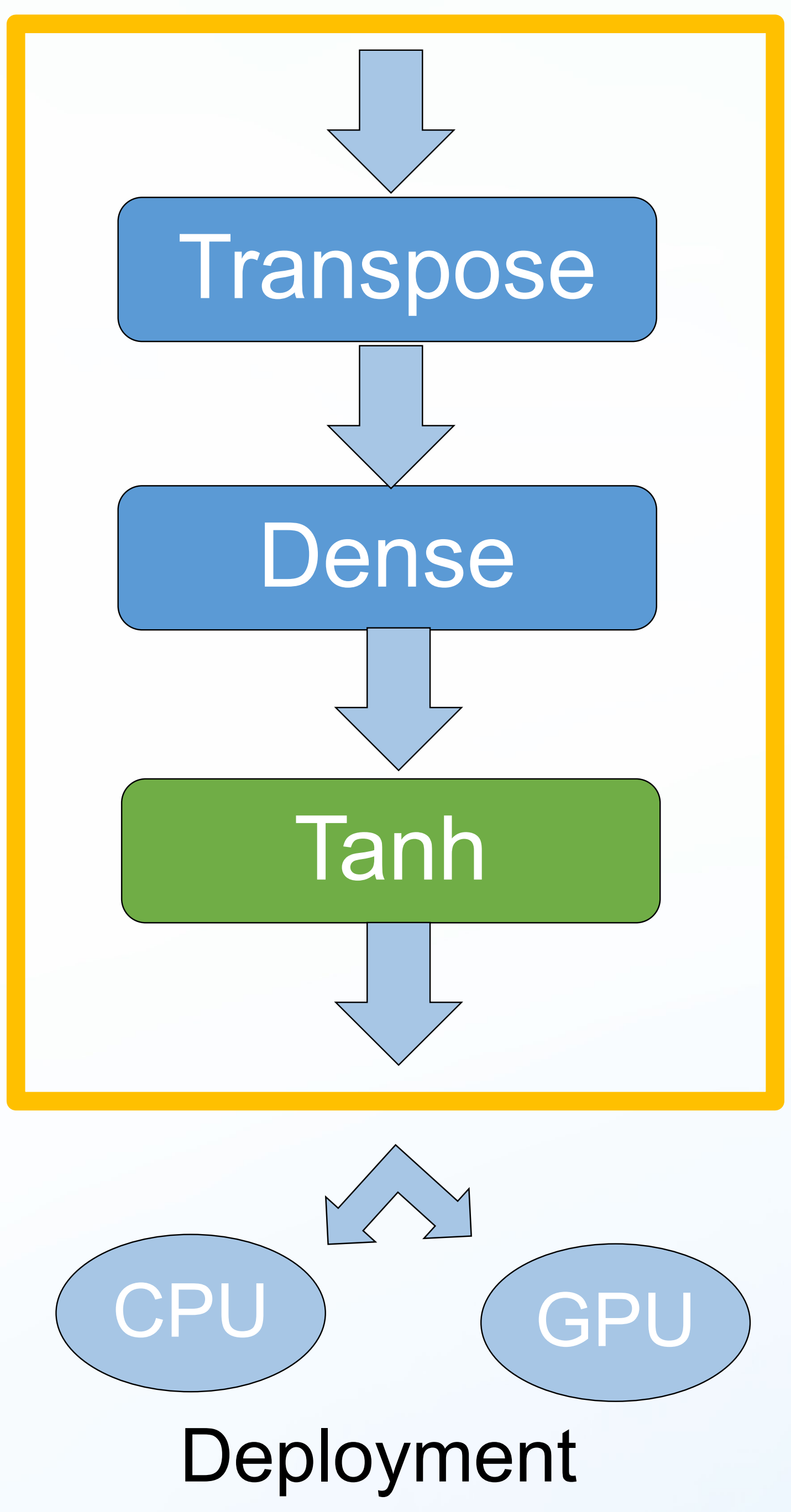
Performance improvement of operator partition



Equal partition

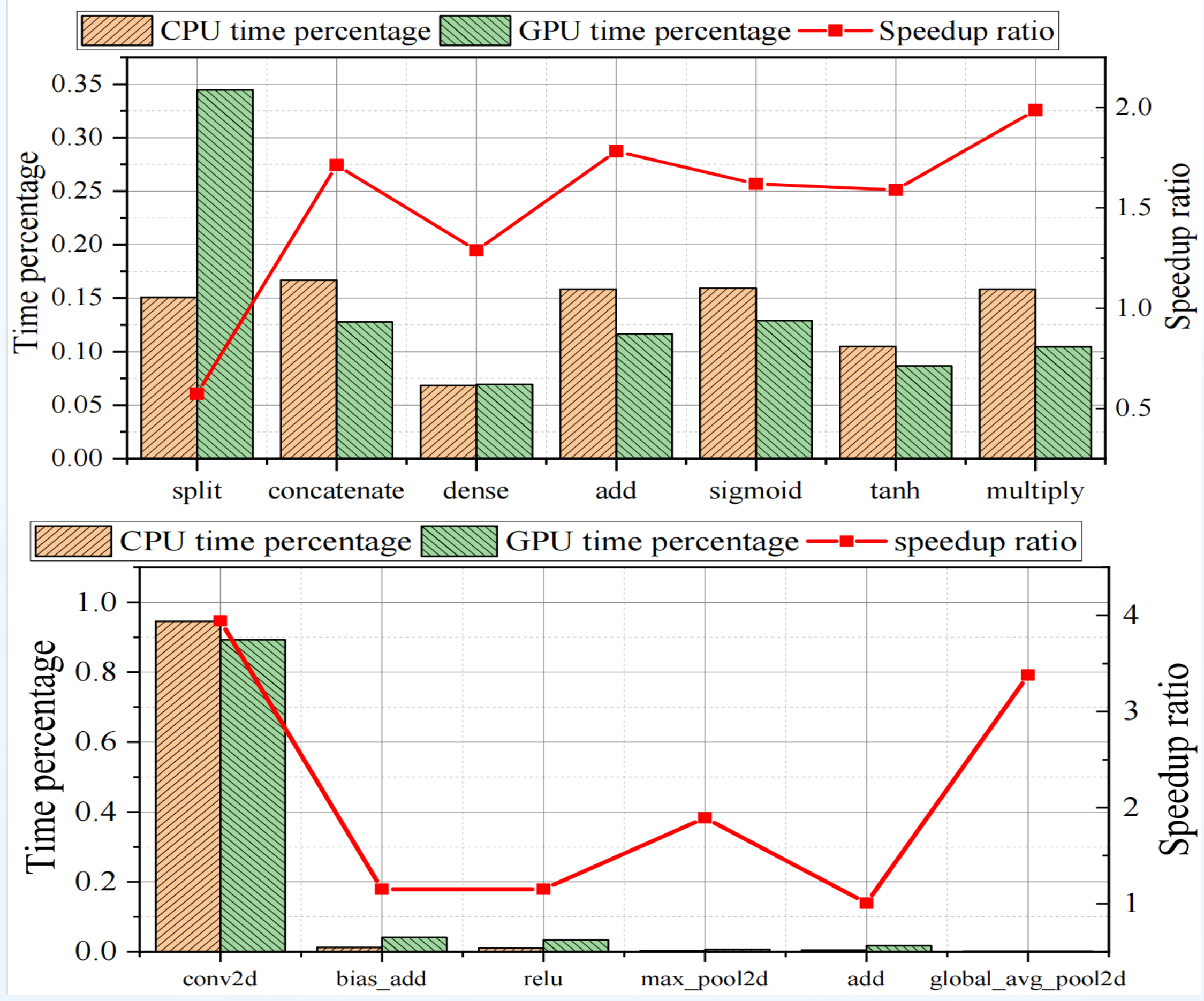
Fixed partition

Dynamic partition(ours)



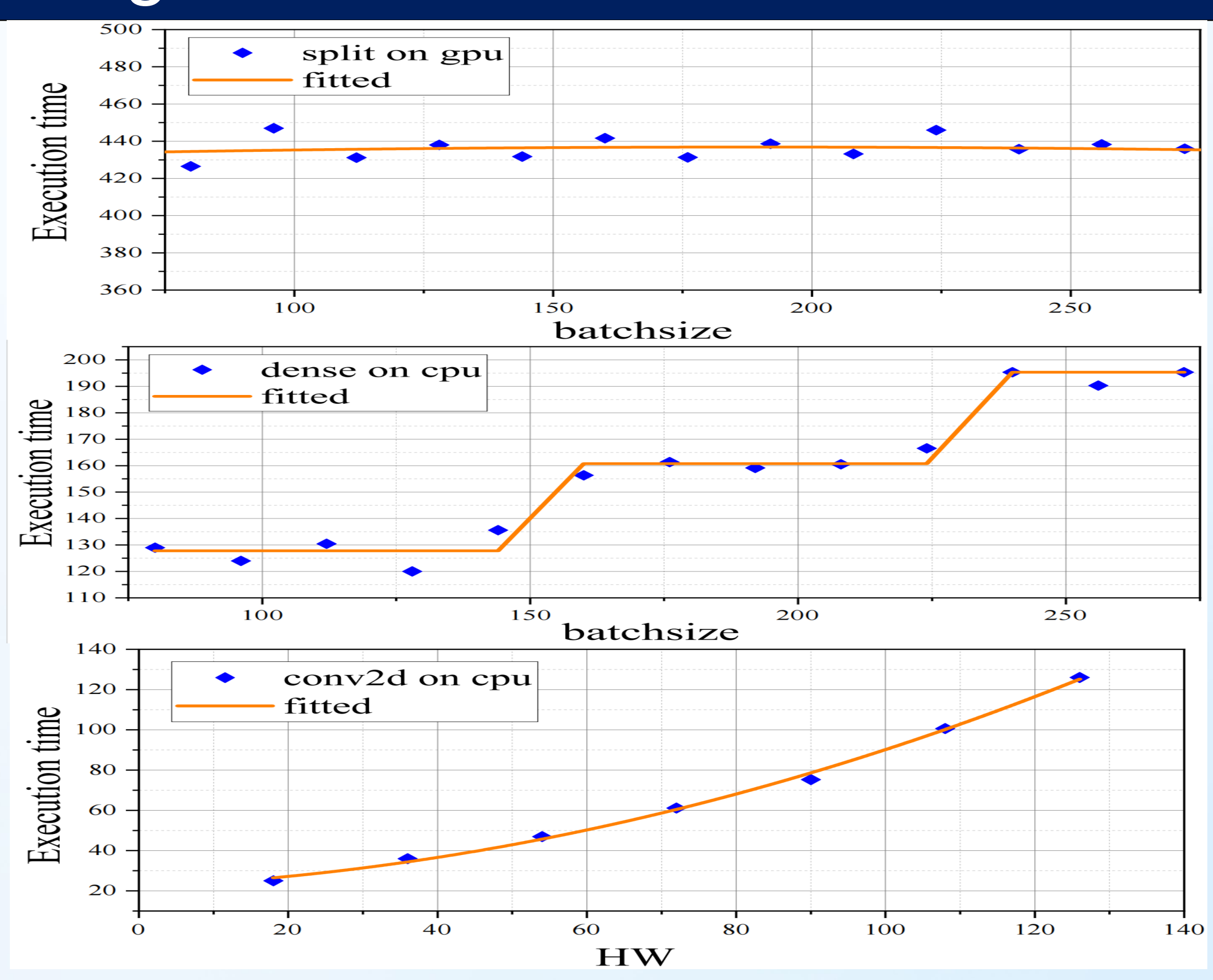
Deployment

Key technologies in EOP



RNN

CNN



Constant

Staircase

Linear

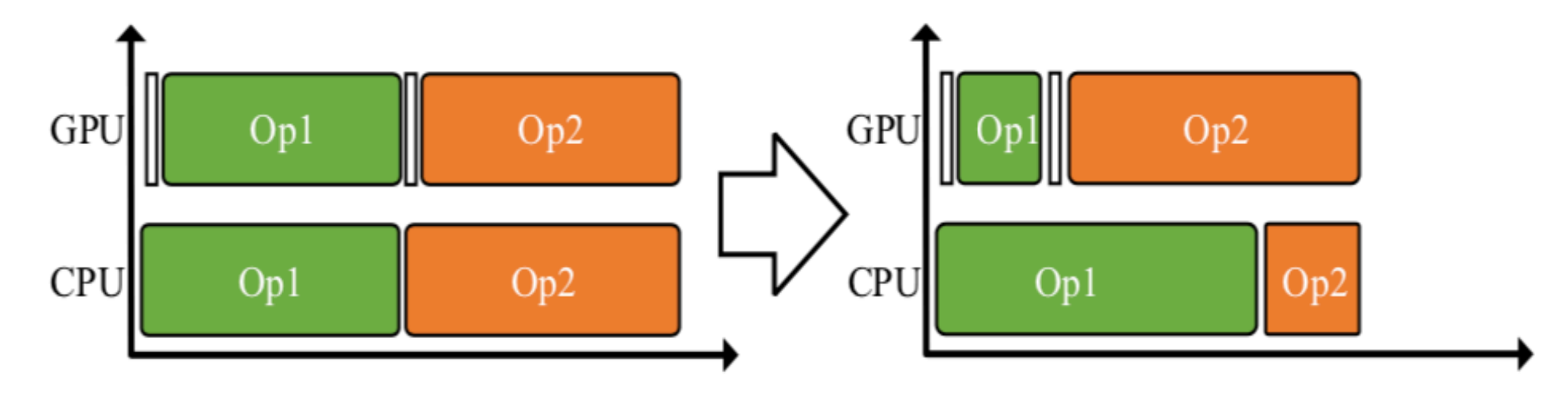
Analyzing Operators

Estimating operators performance

Min ε, *s.t.* :

$$|t(op_i^{GPU}(\beta d), GPU) - t(op_i^{CPU}(\alpha d), CPU)| < \varepsilon$$

$$\varepsilon > 0$$

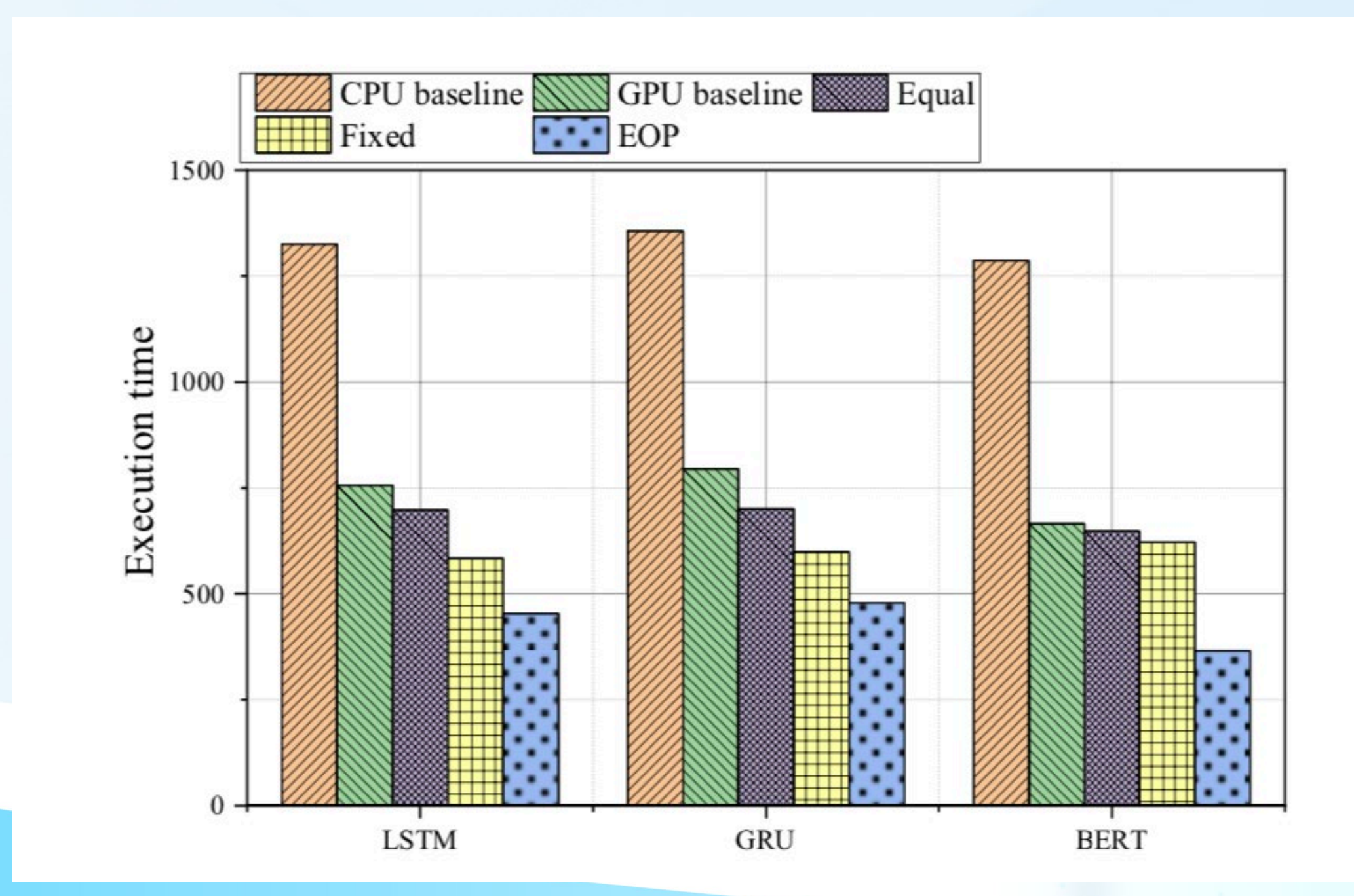


☛1.finds key operators by balancing the execution time of sub-operators on CPU and GPU
☛2.tunes two adjacent operators without partitioning according to their GPU to CPU speedups.
☛3. combines the above two mechanisms

Minimizing overall execution time

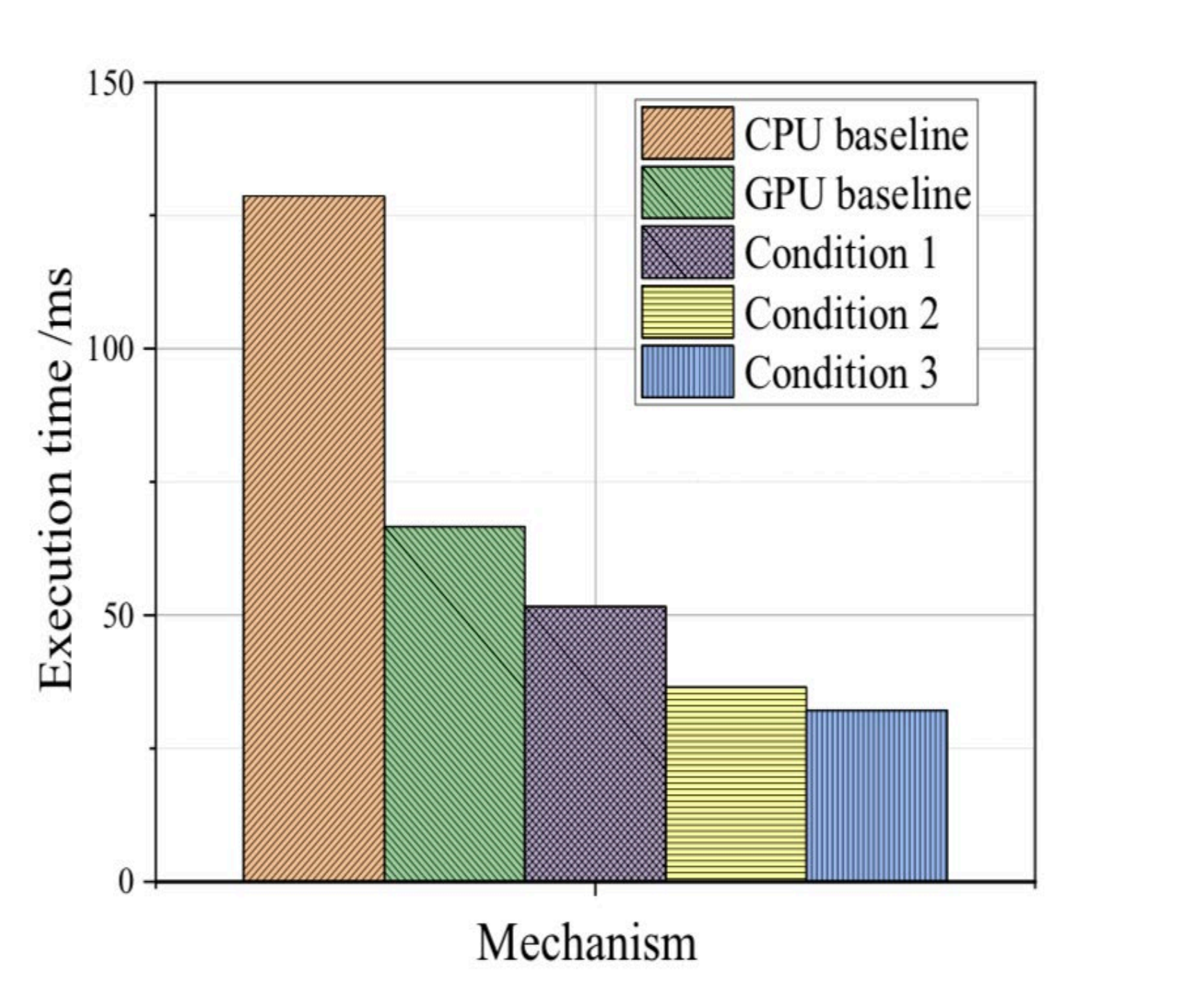
Multiple mechanisms

Experimental results



Reduce up to 1.97x overall execution time

Up to 1.45x improvement


CPU baseline
GPU baseline
Condition 1
Condition 2
Condition 3
Execution time /ms
Mechanism