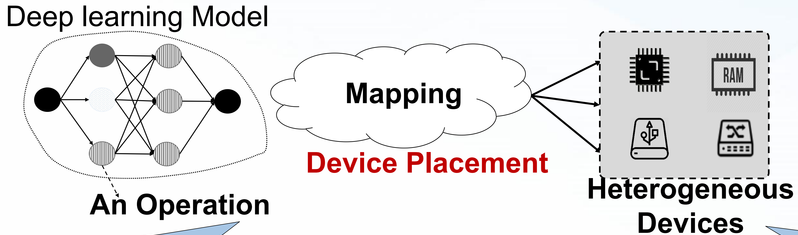
ASAP 2021 (best paper candidate)

Talos: A Weighted Speedup-Aware Device Placement of Deep Learning Models

Yuanjia XU, Heng WU, Wenbo ZHANG, Chen YANG,
Yuewen WU, Heran GAO, Tao WANG
Contact: Yuanjia XU, xuyuanjia2017@otcaix.iscas.ac.cn, 86-13474460179

Device placement for deep learning operations is challenging



An Operation (eg., Conv2D, Add) have **diverse speedups** on many different devices.

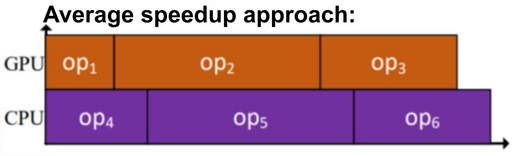
how to get the optimal placement under **diverse speedups**?

A device (eg., CPU, GPU, FPGA) have **diverse speedups** for many different operations.

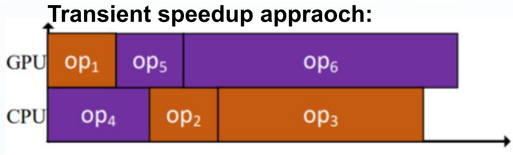
Limitations of existing device placement approaches

Existing approaches do not consdier diverse speedups, and result in **longer total operation completion time (TOCT)**:

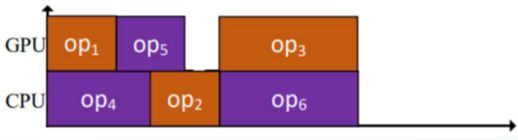
Average speedup approach:



Transient speedup appraoch:

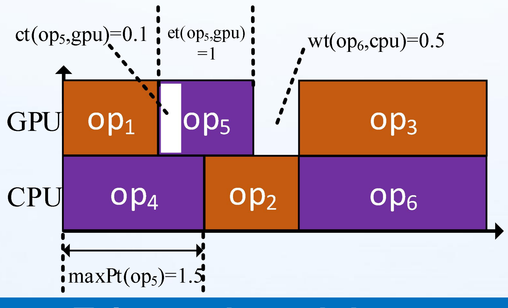


When consider the diversity, we can get the Optimal:



How to design a new diverse speedup-aware approach?

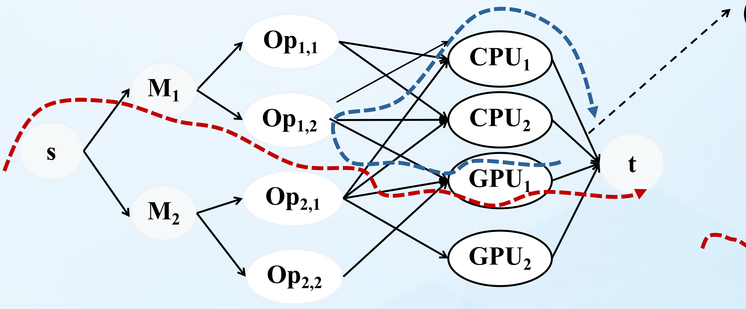
Talos: a new diverse speedup-aware approach



Talos weight-speedup to support speedup diversity:

$$ws(op_5) = \frac{et(op_5, cpu)}{wt(op_5 + ct(op_5) + et(op_5, gpu) - maxPt(op_5))}$$

Talos: using minimum cost flow to do devcie palcement



(flow, cost)

1 **Weighted-speedup**

$Op_{1,2}$: diverse operations

GPU_1: diverse devices

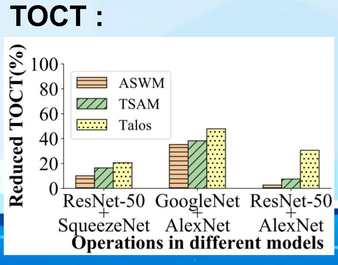
: placement decision for $Op_{1,2}$ and GPU_1

Talos reduces more total operation completion time (TOCT)

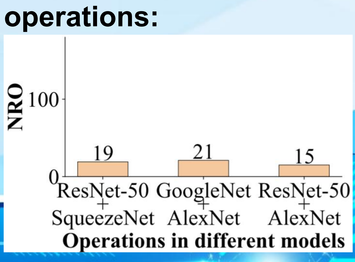
Speedup similarity among deep learning models:

Two models	Affinity value
Bert+GoogleNet	0.14
Bert+Squeeze	0.17
ResNet50+Bert	0.20
AlexNet+SqueezeNet	0.29
ResNet50+GoogleNet	0.21
GoogleNet+SqueezeNet	0.32
Bert+AlexNet	0.37
ResNet50+SqueezeNet	0.46
GoogleNet+AlexNet	0.53
RestNet50+AlexNet	0.56

Reducing 20-50% more TOCT :



Only reassigning a few operations:



See more in: https://github.com/dos-lab/talos