IOS: **I**nter-**O**perator **S**cheduler for CNN Acceleration

Yaoyao Ding[1 2], Ligeng Zhu[3], Zhihao Jia[4],

Gennady Pekhimenko[1 2], Song Han[3]

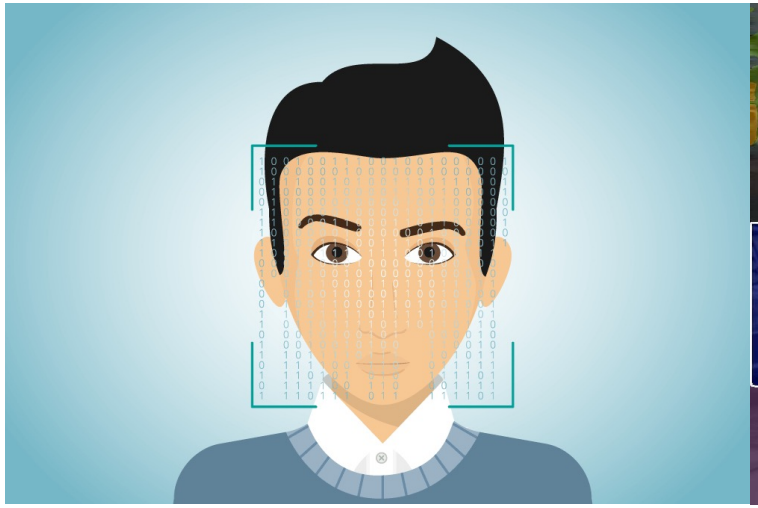
1 UNIVERSITY OF TORONTO

2 VECTOR INSTITUTE

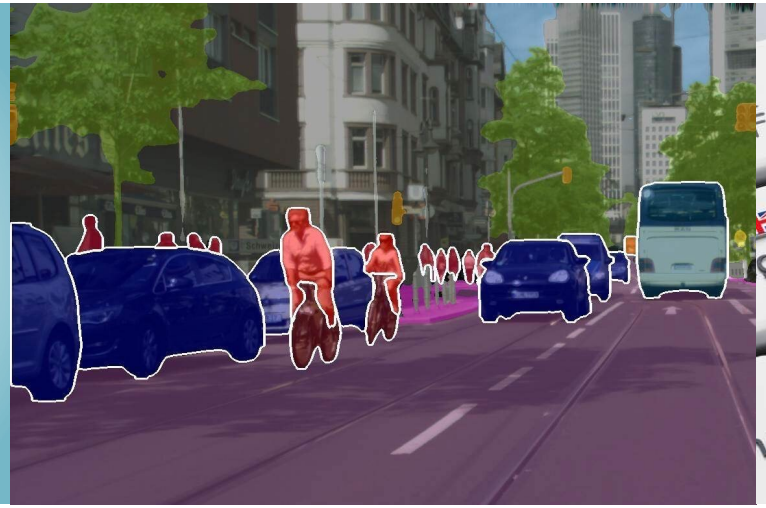
3 MIT

4 Carnegie Mellon University

Efficient Deployment of CNNs is Important



Face Recognition



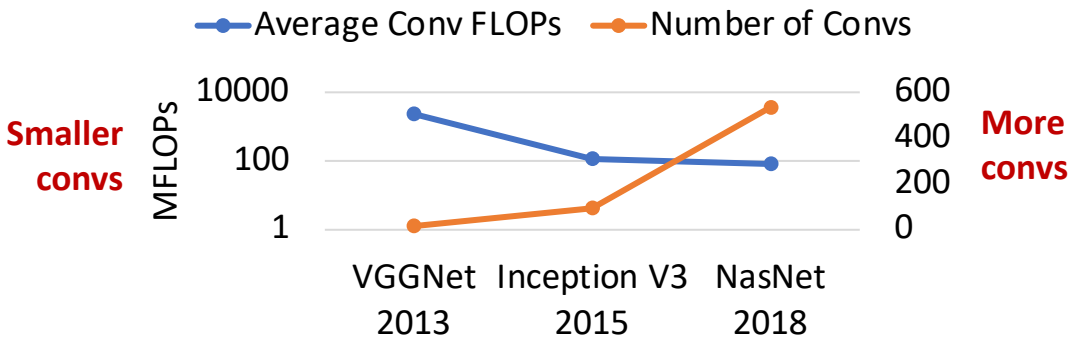
Self Driving

Language Translation

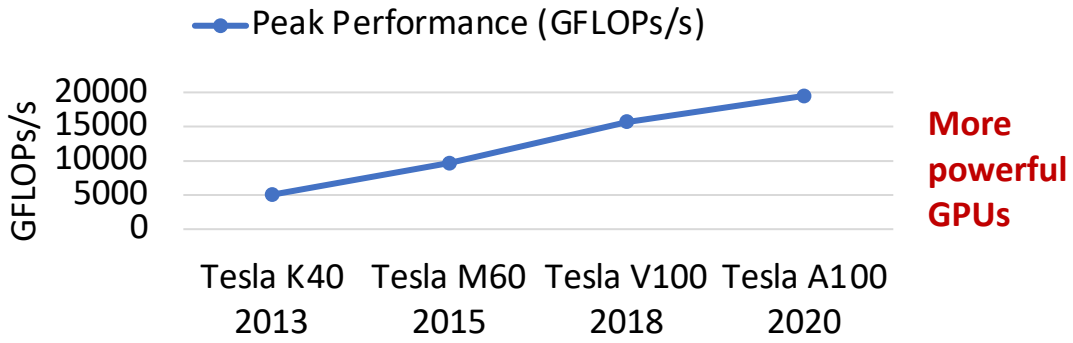
Is CNN inference in current DL libraries well utilizing underlying hardware?

Motivation for Inter-Operator Parallelization

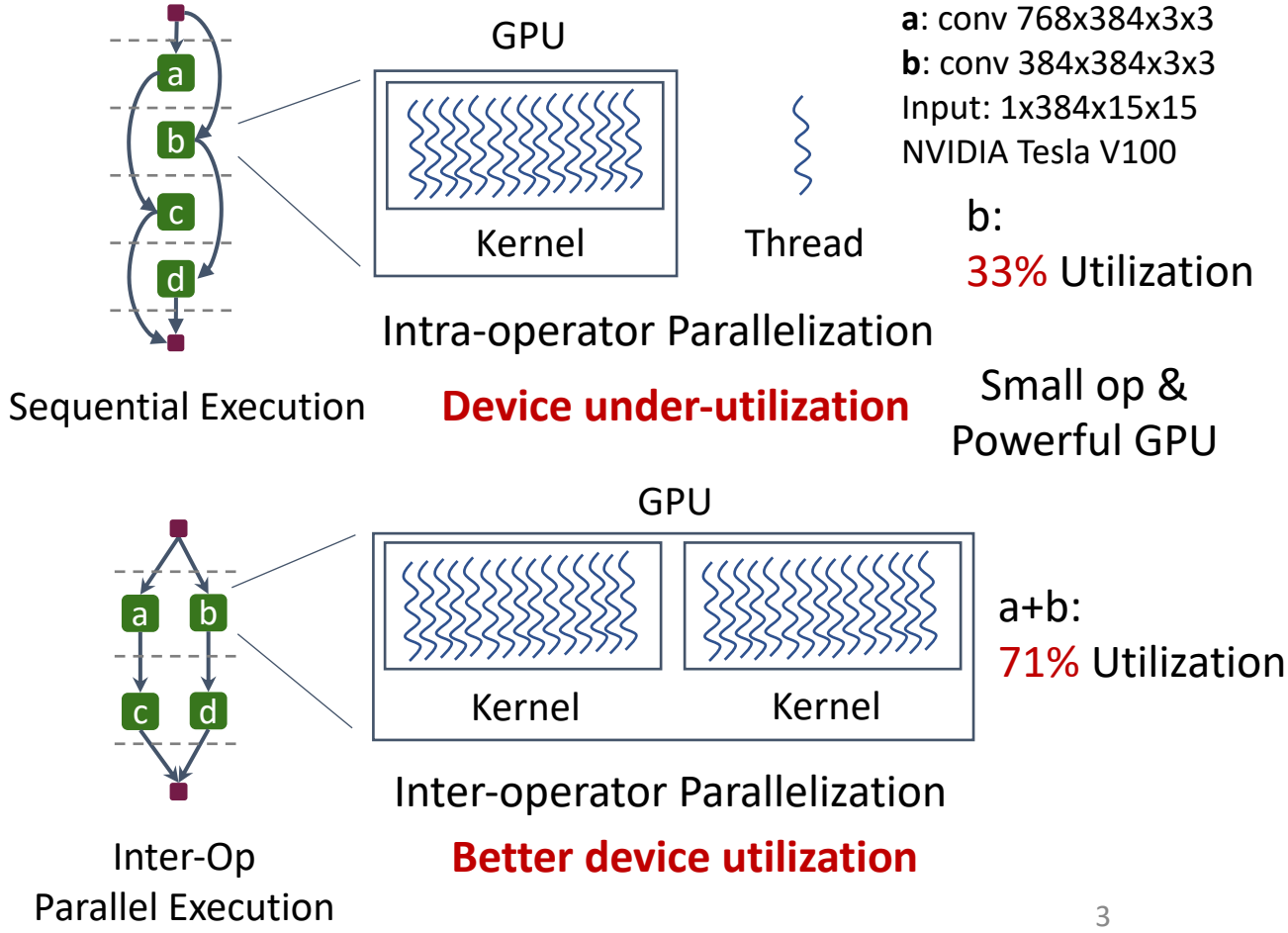
1. More small convs in CNN design.



2. GPU peak performance increased



3. Intra- and Inter-operator Parallelization



Inter-Operator Scheduler (IOS)

$$Latency[S] = \min_{S' \text{ can be a last stage of } S} \left(Latency[S - S'] + StageLatency(S') \right)$$

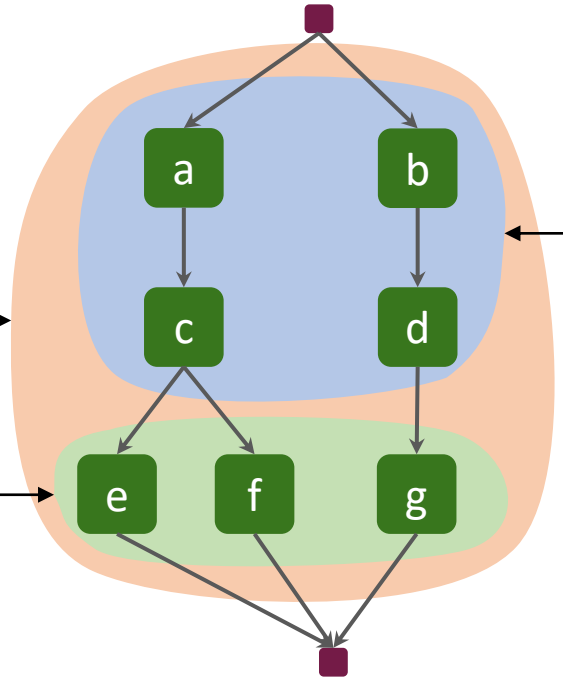
Best schedule's latency of S

S is the ops to be scheduled

S' is a candidate for last stage of S

$S - S'$ are ops remaining to be scheduled in a sub-problem

Latency of stage S'



Explore schedule space **exhaustively**

Time Complexity of IOS: $\mathcal{O}\left((n/d+1)^{2d}\right)$

$\boldsymbol{n}$: number of operators
$\boldsymbol{d}$: width of computation graph (max number of parallelizable ops)

Comparison of cuDNN-based Frameworks

TensorFlow: A widely-used machine learning framework.

TensorFlow-XLA: TensorFlow with compilation optimization.

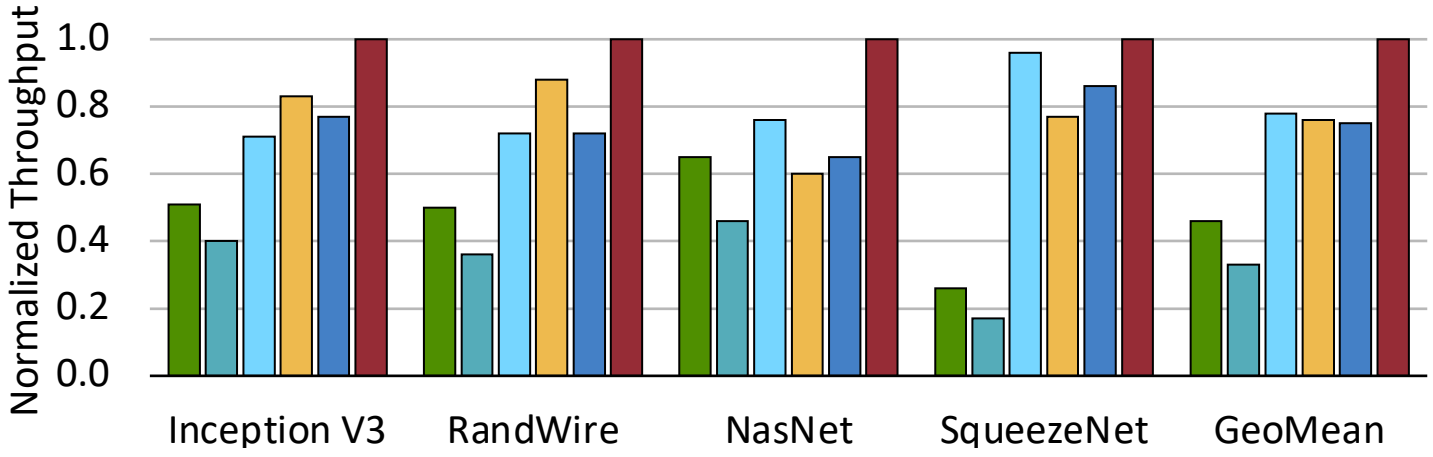
TASO: Transformation-based optimizer.

TVM-cuDNN: TVM backed with cuDNN convolution kernel.

TensorRT: NVIDIA high-performance inference engine.

IOS: Our method

Under-utilization due to sequential execution



Performance is normalized to the best framework

IOS outperforms all frameworks and achieves **1.1-1.5x** speedup.

Conclusion

- Sequential execution suffers from **under utilization** problem.
- **I**nter-**O**perator **S**cheduler (**IOS**):
  - Utilize both intra- and **inter-operator parallelism** in CNNs.
  - **Dynamic-programming** explores the schedule space **exhaustively**.
  - Time Complexity: $\mathcal{O}\left((n/d+1)^{2d}\right)$, d is usually small.
- Key Results: **1.1-1.5x** speedup on diverse CNNs.

https://github.com/mit-han-lab/inter-operator-scheduler

We provide scripts to reproduce results in every figure and table!