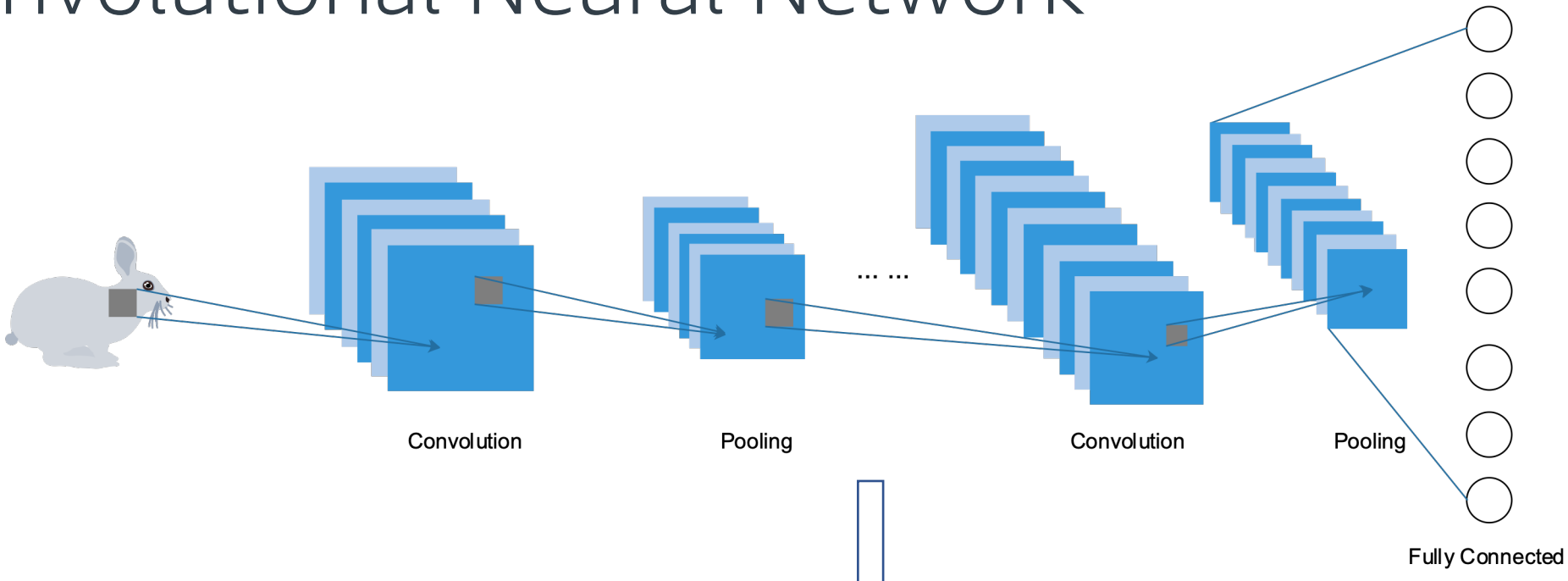
Optimizing CNN Model Inference on CPUs

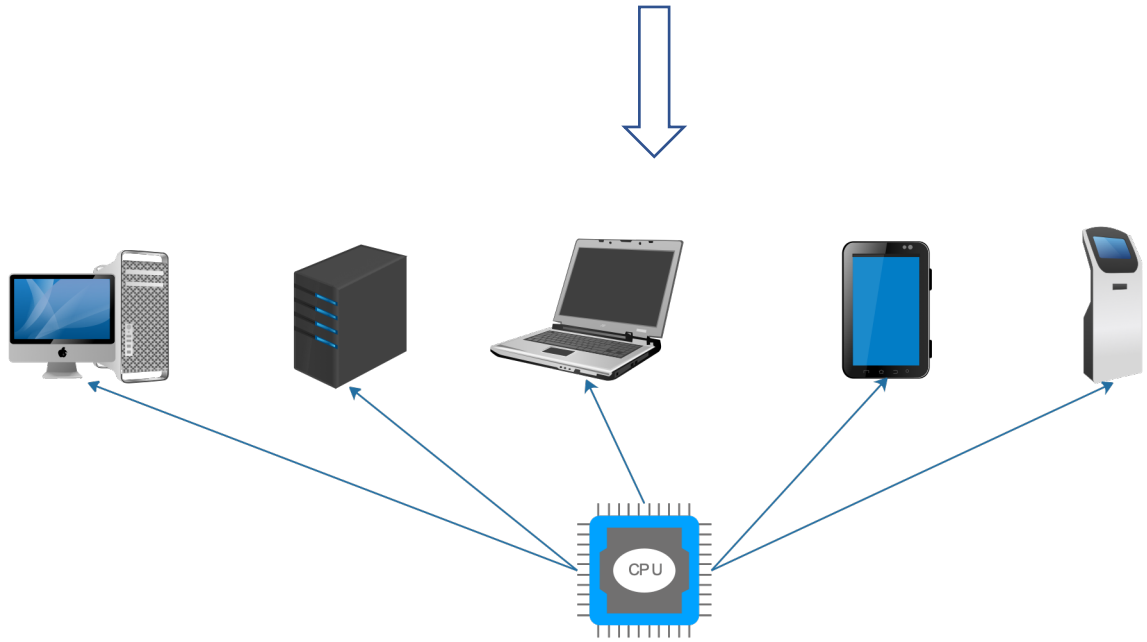
Yizhi Liu*, Yao Wang*, Ruofei Yu, Mu Li, Vin Sharma, Yida Wang

Amazon Web Services

*Equal contribution

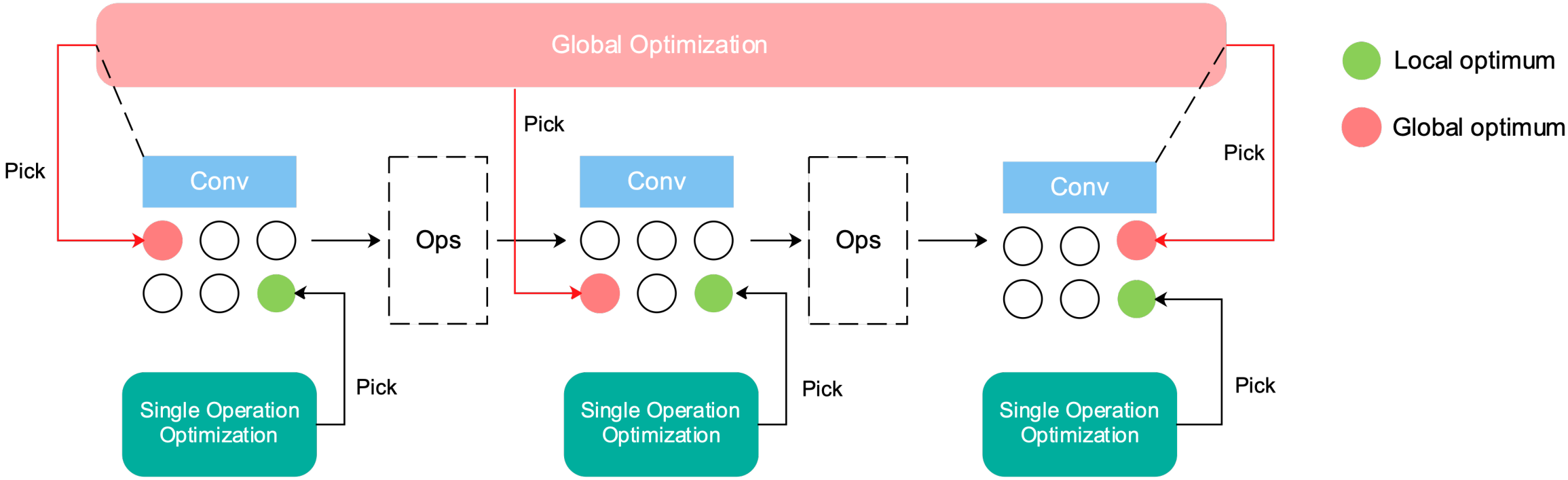
Convolutional Neural Network





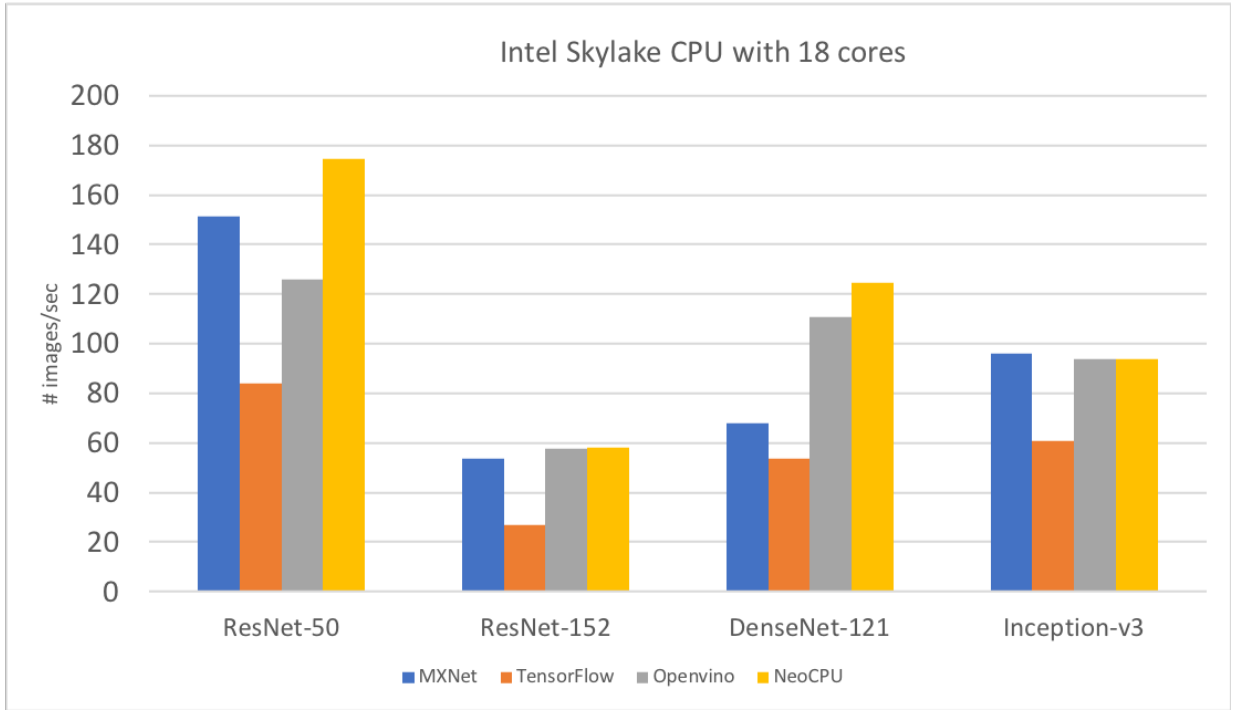
Optimization

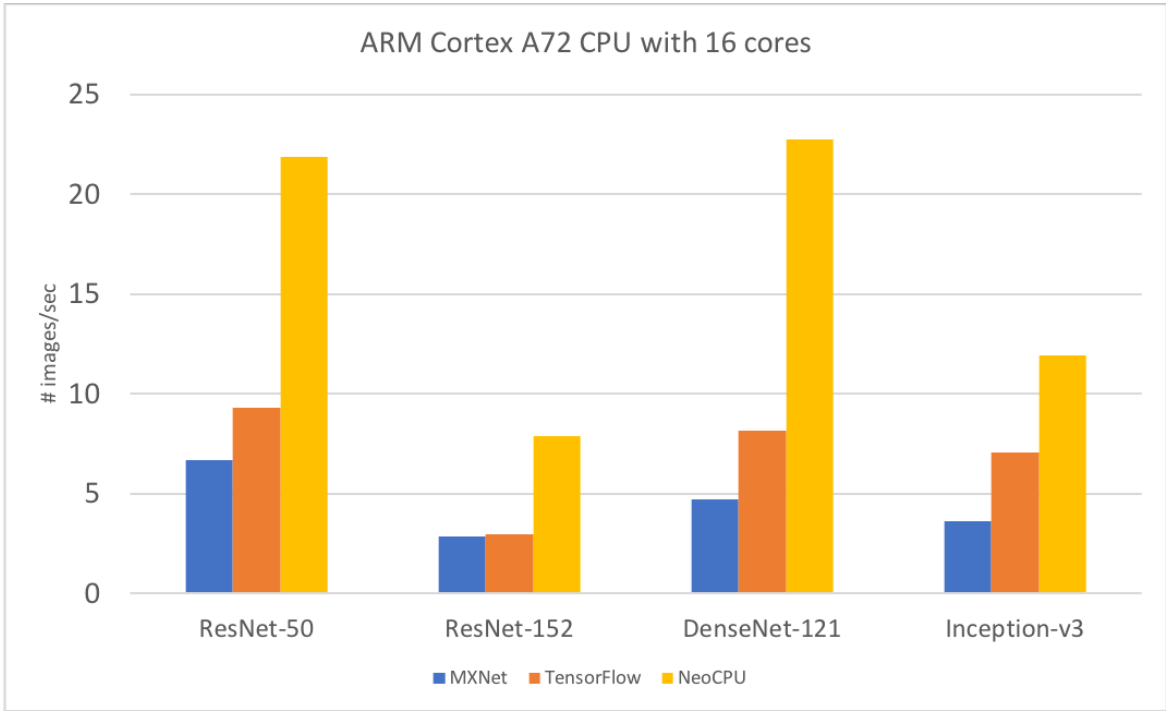
- Optimization in existing work mostly focus on **single** operator acceleration.
- We consider tensor-level and graph-level **joint** optimization.



Performance

Our solution (NeoCPU) achieved competitive performance and scalability across various of cloud and edge CPUs.





amazon

Optimizing CNN Model Inference on CPUs

USENIX ATC '19

11:50 am, Track II

Friday, July 12