FACULTY OF ENGINEERING AND ARCHITECTURE

ISPASS

IDLab
INTERNET & DATA LAB

NETMODEL GROUP - INTERNET AND DATA SCIENCE LAB

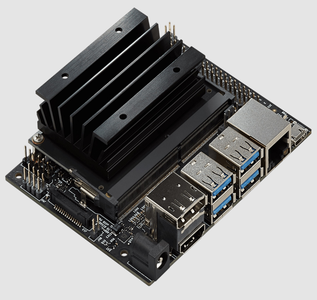
Abhinaba Chakraborty, Didier Colle, Mario Pickavet, Akis Kourtis, Andreas Oikonomakis, Wouter Tavernier

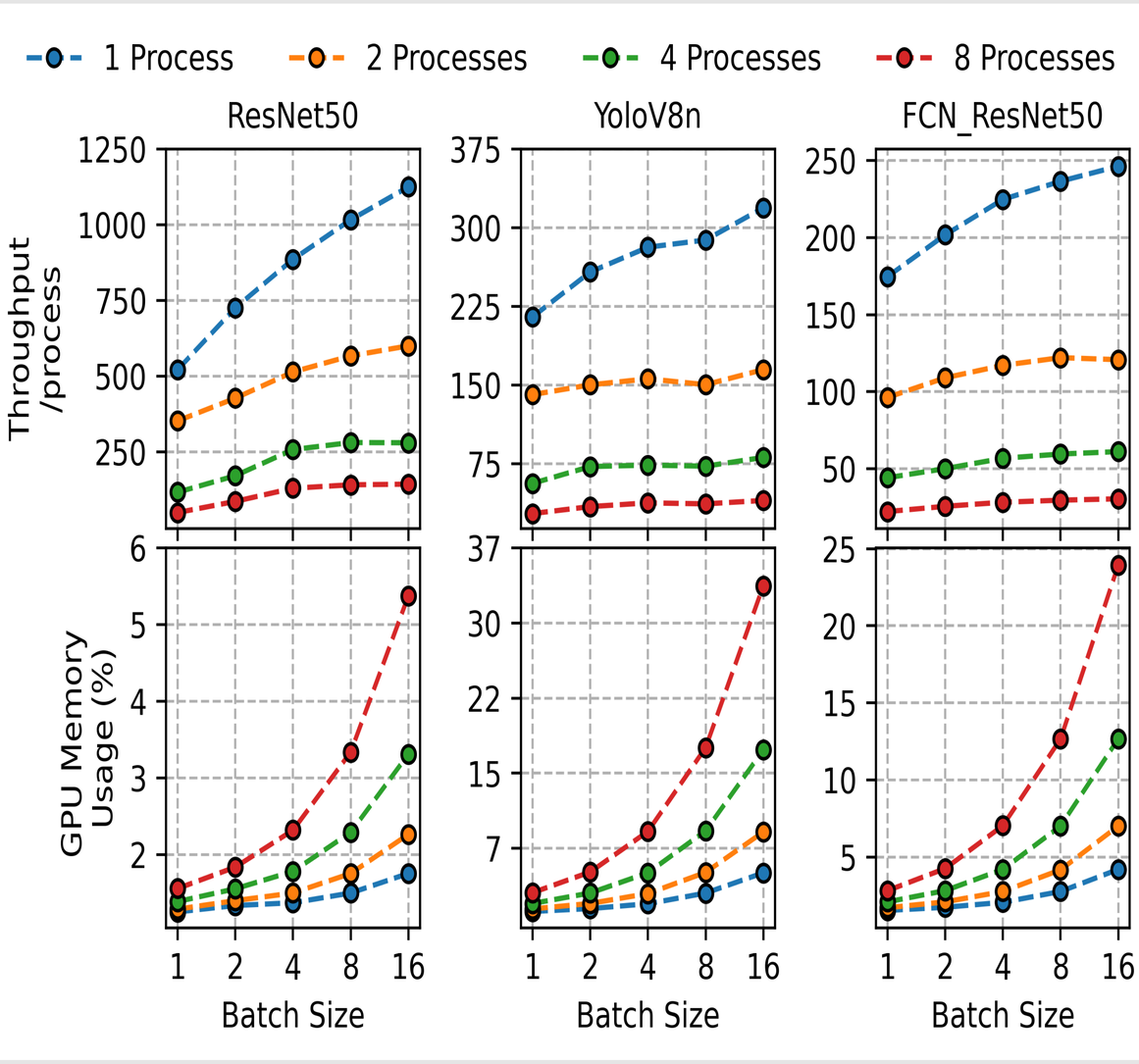
PROFILING CONCURRENT VISION INFERENCE WORKLOADS ON NVIDIA JETSON

System Throughput vs Number of Process

$process \uparrow \; throughput/process \downarrow \; GPU\ Memory\ Usage \uparrow$

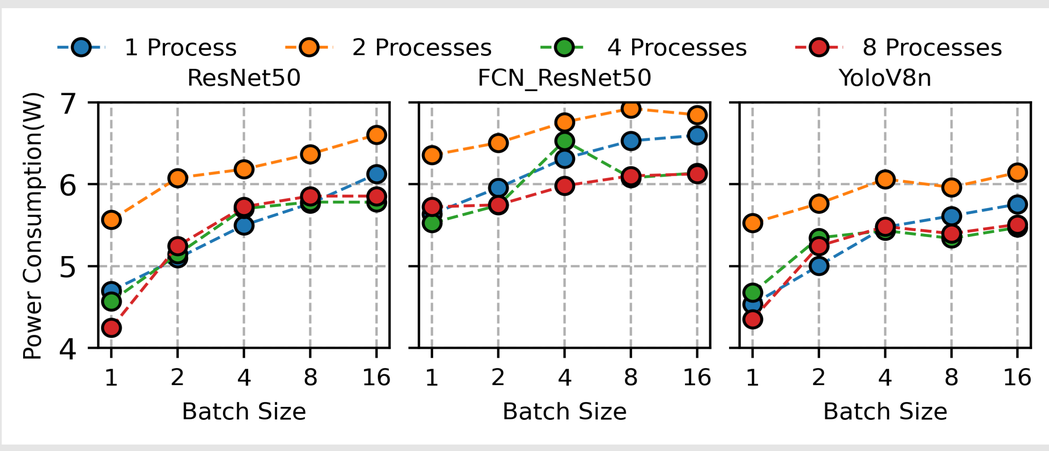
$batch \uparrow \; throughput/process \uparrow \; GPU\ Memory\ Usage \uparrow$





Power Consumption vs Number of Process

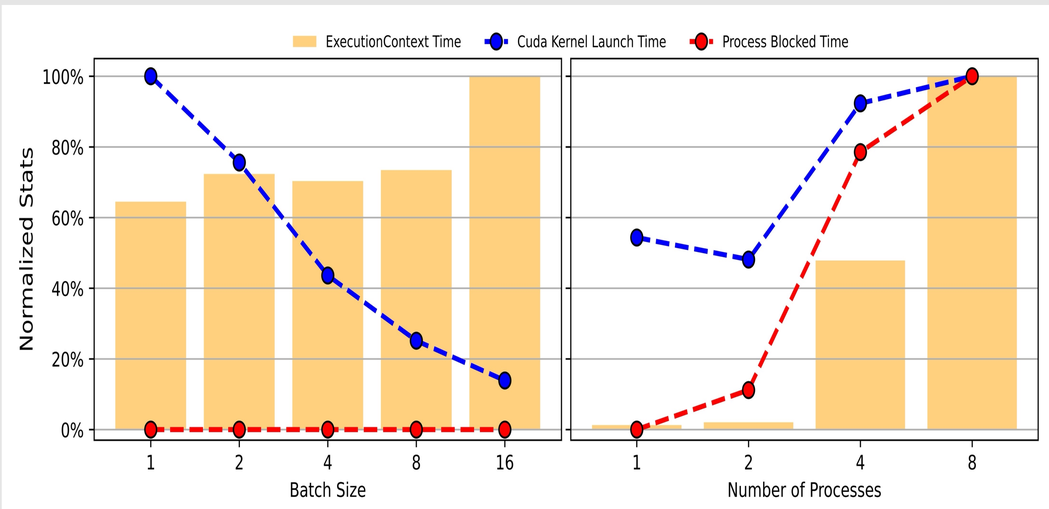
$power\ consumption \propto throughput$



Low Level Performance Metrics

$process \uparrow \; EC\ Time \uparrow \; Kernel\ Launch\ Time \uparrow$

$Batch \uparrow \; Kernel\ Launch\ Time \downarrow$

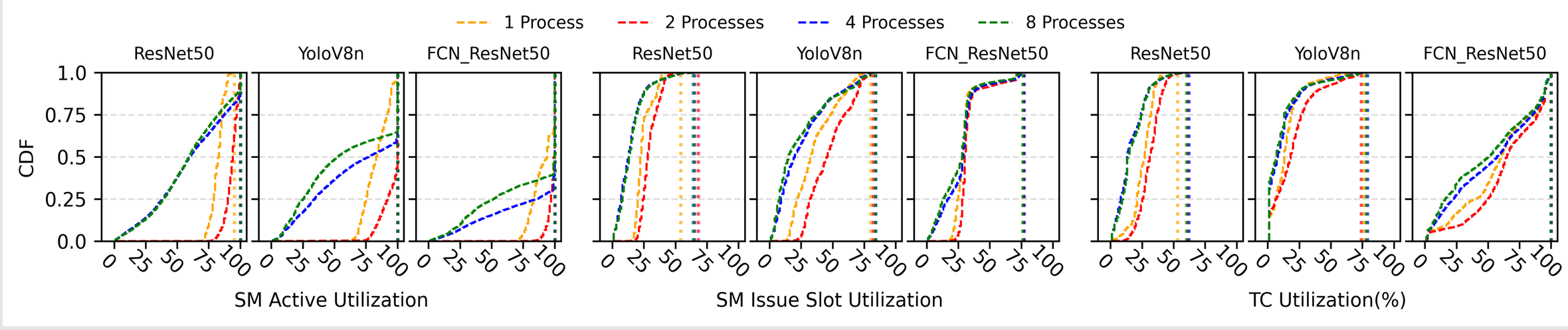


Kernel Level Performance Analysis

$SM\ Utilization \sim 80 - 100\%$

$Issue\ Slot\ Utilization \sim 25\%$

$Tensor\ Core\ Utilization \sim 25\%$



Conclusions

1. Need for Fine-Grained CPU-GPU Co-Scheduling of Threads
2. Development of orchestration techniques for concurrent DL executions across CPU-GPU

Contact

abhinaba.chakraborty@ugent.be
https://www.idlab.ugent.be/research-teams/netmodel

Universiteit Gent

@ugent

Ghent University

GHENT UNIVERSITY

imec