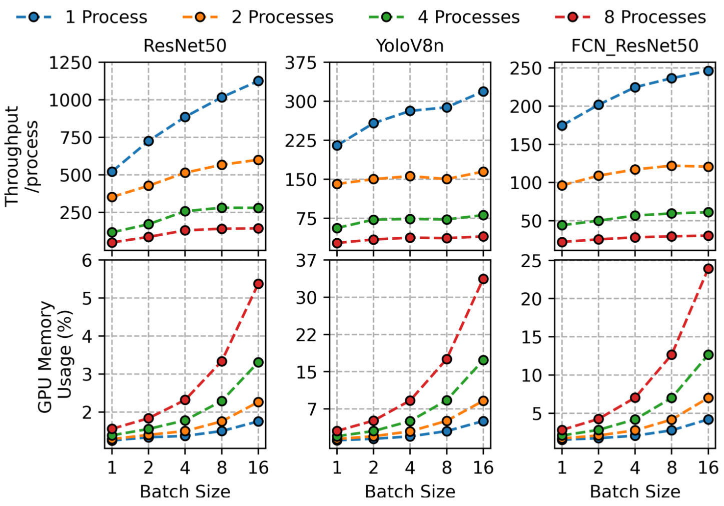
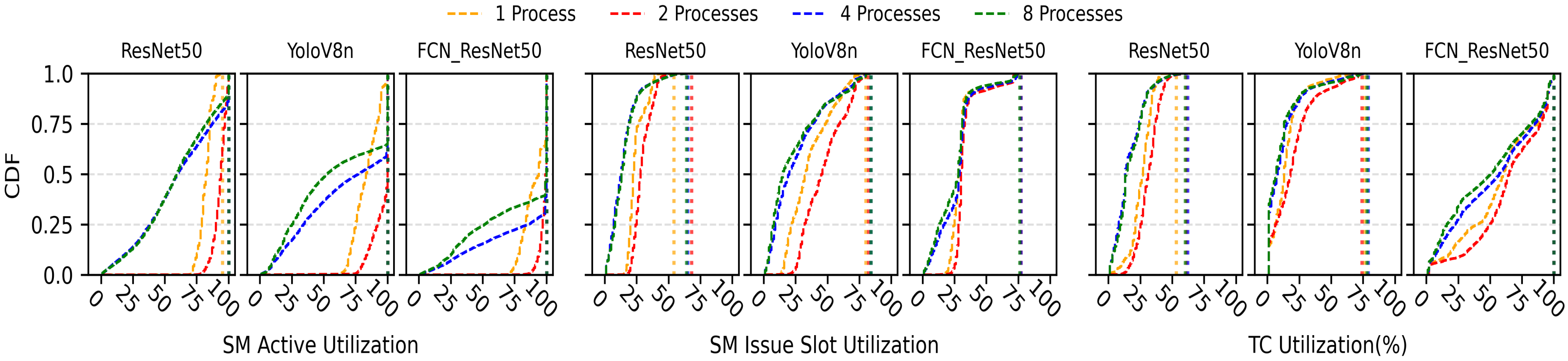
PROFILING CONCURRENT VISION INFERENCE WORKLOADS ON NVIDIA JETSON

ABHINABA CHAKRABORTY, DIDIER COLLE, MARIO PICKAVET, AKIS KOURTIS, ANDREAS OIKONOMAKIS, WOUTER TAVERNIER

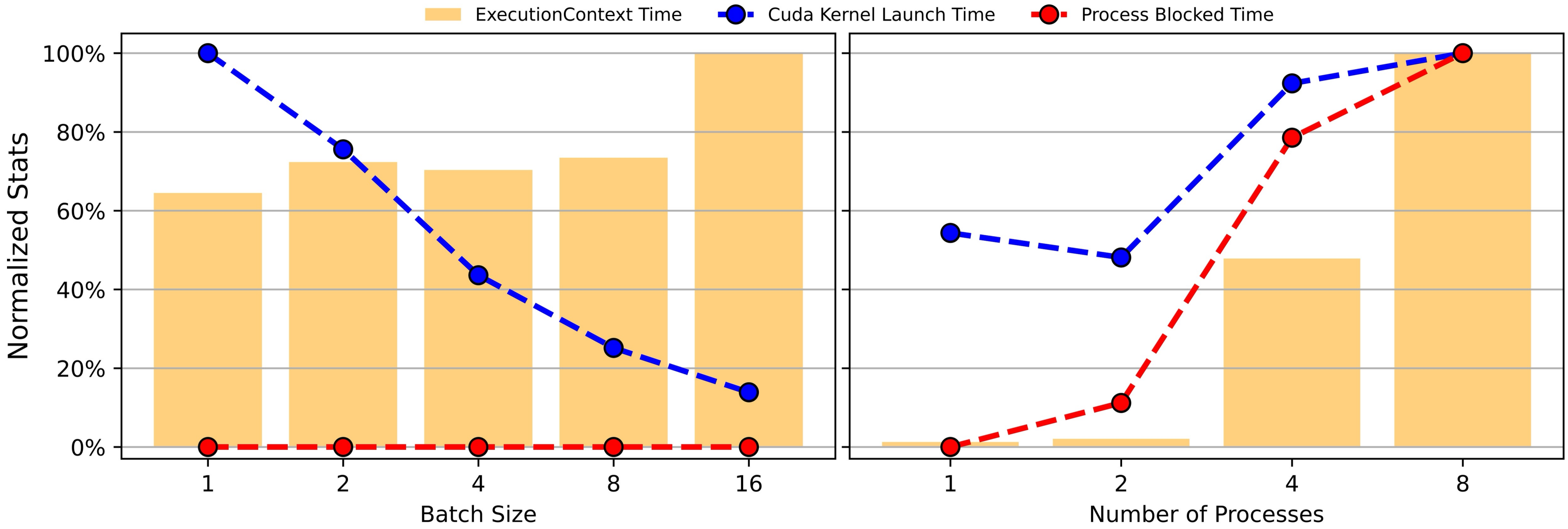
PROFILING RESULTS

1. Throughput/process increases when batch size increases
2. Throughput/process decreases when the number of processes increases
3. SM Utilisation ~ 80-100%
4. Issue Slot Utilisation ~ 25%
5. TC Utilization ~ 25%





BOTTLENECK ANALYSIS



1. EC Time increases exponentially with the growing number of processes
2. EC time increases linearly with a growing batch size.