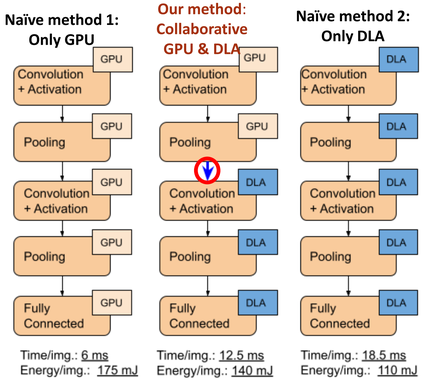
Pushing the Edge: Efficient AI Computing with NVIDIA Jetson SoCs

Ismet Dagli (Ph.D. Candidate), Justin Davis (Ph.D. Student), Mehmet E. Belviranli (Advisor), Computer Science Department, Colorado School of Mines

Energy-aware and Shared-memory-contention-aware Layer-wise Scheduling of DNN Inference

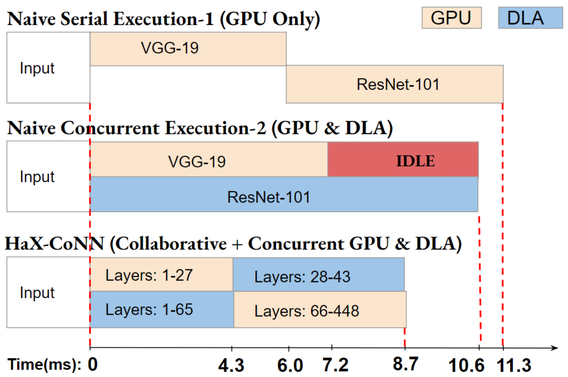
Optimizing Latency under Energy Constraint



Target Hardware: Xavier AGX & **Target Application**: GoogleNet

Take-away: Collaborative multi-accelerator (GPU & DLA) execution of a single DNN inference in a single edge device can save significant time & energy.

Significant Slowdown when DNNs are assigned to GPU & DLA

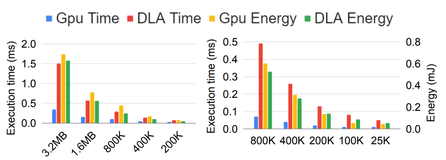


Target Hardware: Orin AGX , **Target Application**: VGG19 &ResNet101

Take-away: Ensuring optimal concurrent execution requires careful consideration of shared memory contention.

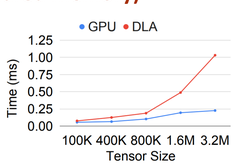
Observation-1: Getting benefits of Hardware Heterogeneity When Scheduling Tasks (layers)

GPU/DLA: Latency: 2x to 4.5x & Energy 1.1x to 2.4x



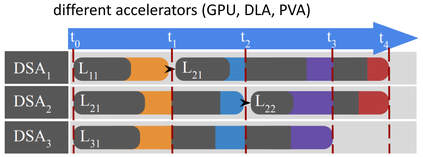
(a) Convolution (b) Pooling

Observation-2: Transition Cost between GPU and DLA (with shared memory)



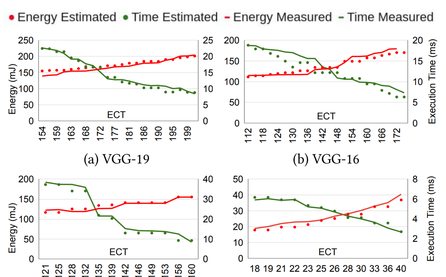
Observation-3: Slowdown Depends on Memory Demand by Layers and the Accelerator

Different layers observe varying slowdown on different accelerators (GPU, DLA, PVA)



EVALUATION (Energy)

Comparison of measured results against estimation by our model



(a) VGG-19 (b) VGG-16

FORMULATION

We integrate layer execution time, inter-accelerator transition time, and memory contention slowdown into a cost function formulate the scheduling problem.

Objective:

$$\min \quad T(L,P,S(L\to P)) \quad \text{s.t.} \quad E(L,P,S(L\to P)) < ECT$$

Total Time:

$$T(L,P,S(L\to P),TR) = \sum_{i=0}^{len(L)} (t(L_i,s(L_i)) + (TR_i \times \tau(L_i,s(L_i),OUT|IN) + (TR_i \times pipeline(L_i,S(L_i))))$$

Trans. Time:

$$TR_i = \begin{cases} 1 & \text{if } S(i) \neq S(i+1) \\ 0 & \text{if } S(i) = S(i+1) \end{cases}$$

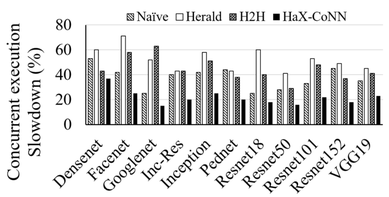
Slowdown Amount:

$$C_{L_{i,n},S(L),L} = \sum_{I_k \in Int} \frac{I(L_{i,n}\ ,\ L_{j,n}) * cont\_model(L_{i,n}\ ,\ L_n)}{t(L_{j,n},S(L_{i,n})) * len(L_n)}$$

where $1 \le j \le len(DNN_n)$, $1 \le n \le len(DNN)$, $L_{j,n} \in L_4$, $Int_k \cap [st_{i,n}, et_{i,n}] \neq \emptyset$, $Int_k \cap [st_{j,n}, et_{j,n}] \neq \emptyset$

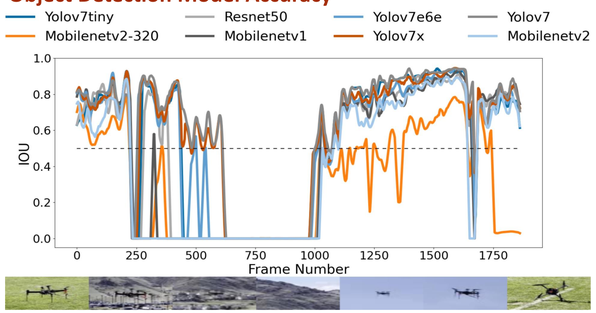
EVALUATION (Minimal Shared Memory Contention)

Slowdown amount per DNN when we run a DNN (x-axis) with other DNNs on the Orin AGX. HaX-CoNN (our work) minimizes the slowdown amount.



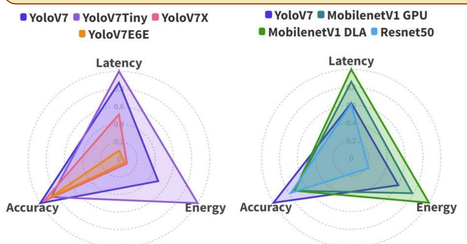
Optimizing Continuous Object Detection on Multi-Accelerator SoCs

Object Detection Model Accuracy



NO NEED FOR SOTA MODELS IN MOST SCENARIOS

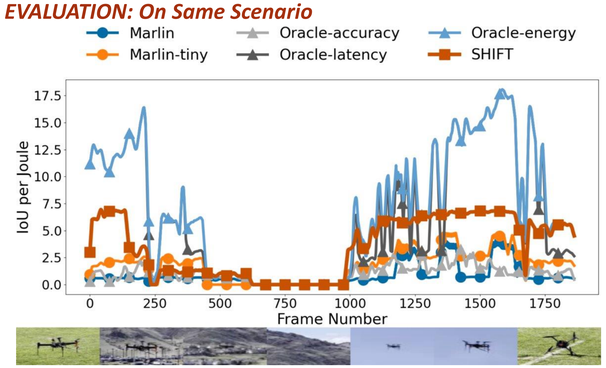
Take-away: Context-aware switching to non-SOTA object detection model architectures can improve energy/latency without significant accuracy loss



EXPLOIT NON-MONOTONIC RELATIONSHIPS BETWEEN MODEL ARCHITECTURES + HARDWARE

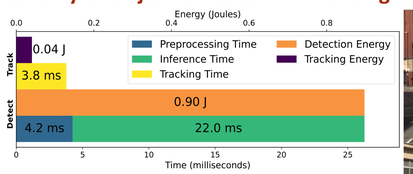
SOLUTION: Swap to models with lower computational cost if their predicted accuracy meets a threshold. Utilize a heuristic scheduler handling latency, energy, and accuracy.

EVALUATION: On Same Scenario



Target Hardware: Xavier NX + Luxonis OAK-D Lite

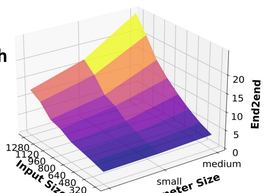
Latency of Object Detection vs. Tracking in MOT



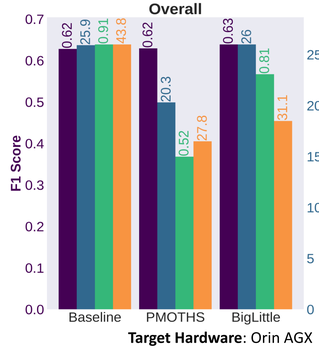
DETECTION IS THE BOTTLENECK FOR LATENCY AND ENERGY

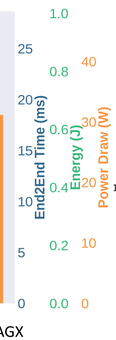
Take-away: Splitting detection can reduce computational bottlenecks on edge SoCs

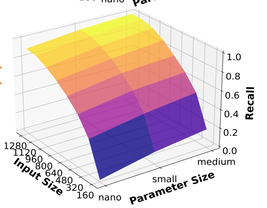
SOLUTION: Split frames at runtime and schedule models to process regions across hardware to minimize latency. Regions with low-priority objects are aggregated into a single region for efficient processing.



EVALUATION: Methods on MOT17 Dataset







Target Hardware: Orin AGX

INCREASING SINGLE MODEL ACCURACY HAS QUICKLY DIMINISHING RETURNS, MORE COMPLEX SOLUTION REQUIRED

RESULTS: SHIFT & PMOTHS

- ***SHIFT*** has up to **7.5x** reduction in energy usage and **2.5x** improvement in latency
- ***PMOTHS*** demonstrates up to **3.0x** reduction in latency, **6.5x** in energy, and **2.0x** in power draw on MOT17 dataset.

References

[1] I. Dagli, A. Cieslewicz, J. McClurg, M.E. Belviranli, "**Axonn: Energy-aware execution of DNN inference on multi-accelerator heterogeneous SoCs**", *DAC 2022*

[2] I. Dagli, M.E. Belviranli, "**Shared Memory-contention-aware Concurrent DNN Execution for Diversely Heterogeneous System-on-Chips**", *PPoPP 2024 & MICRO SRC Finalist (Top-3)*

[3] J. Davis M.E. Belviranli, "**Context-aware Multi-Model Object Detection for Diversely Heterogeneous Compute Systems**", *DATE 2024 & ASD Outstanding Paper Award*

[4] J. Davis I. Dagli, M.E. Belviranli, "**Priority-based Fast Multi-Object Tracking on Multi-Accelerator Systems**", *Under Submission*