The Architectural Implications of Multi-modal Detection Models for Autonomous Driving Systems

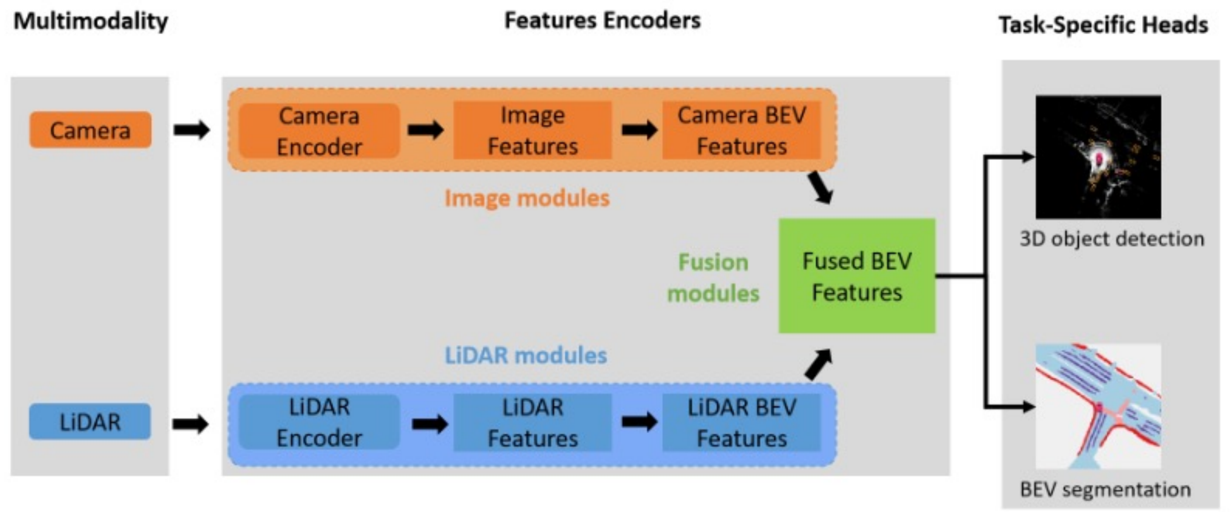
Yunge Li, Shaibal Saha, Lanyu Xu
Oakland University
Contact: {yungeli, shaibalsaha, lxu}@oakland.edu

COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF ENGINEERING AND COMPUTER SCIENCE
OAKLAND UNIVERSITY

Motivation

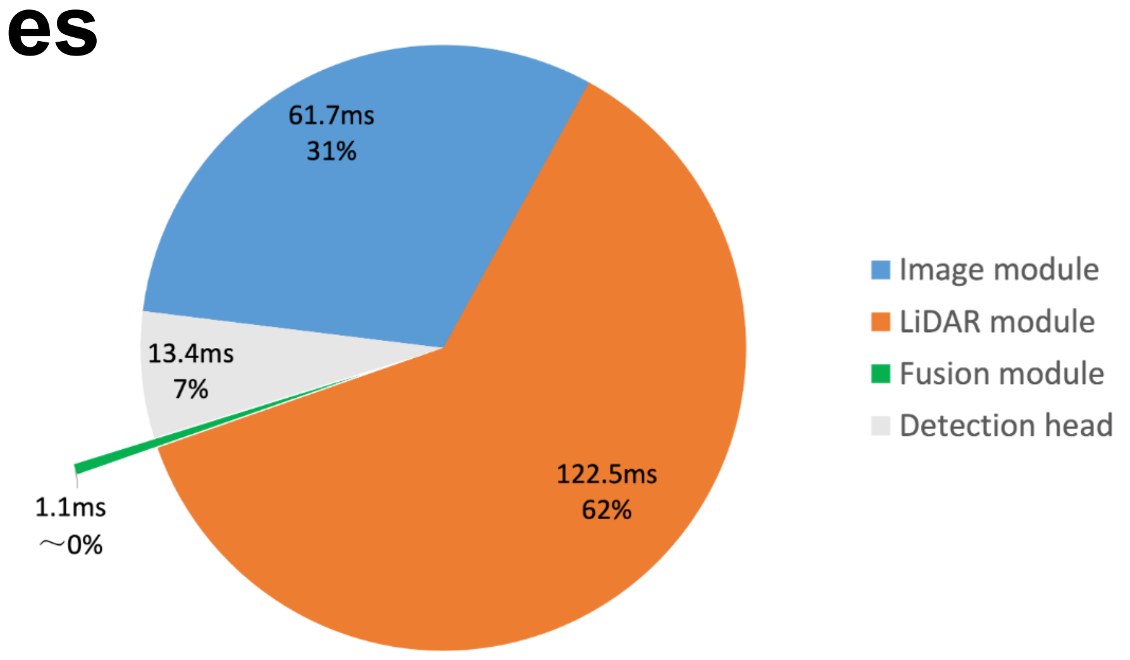
- Unique structure of FPGA and GPU.
- Deconstruct the models into modules.
- Offloading modules to FPGA to reduce power usage by GPU.

Multi-modal model on FPGA



Fuse module on FPGA?

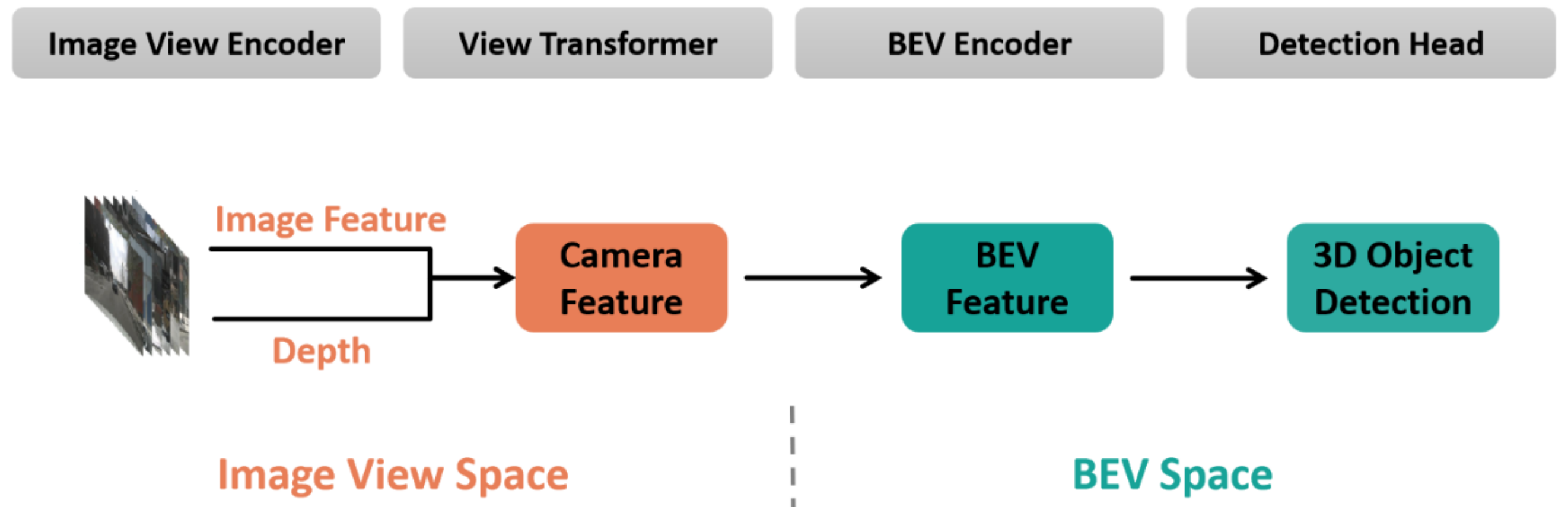
Computational bottleneck on different modules



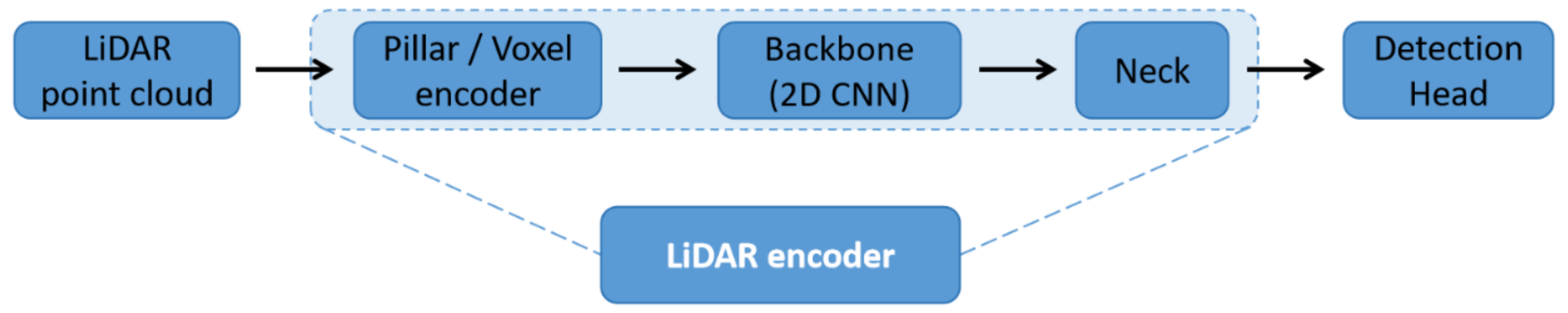
- Main bottlenecks are image, LiDAR module and Detection Head

Model Architectures

❖ BEVDet



❖ PointPillars



The design constraints are Latency, Power, Memory, Development tools

Results

Experiment Setup

Dataset:

- Kitti (PointPillars).
- NuScenes (BevDet).

Accelerating Platforms:

- Zynq UltraScale+ MPSoC ZCU104
- Nvidia GeForce RTX 3080

Metrics:

- Model Performance
  - Mean Average Precision (mAP)
  - NuScenes Detection Score (NDS)
- Efficiency Measurement Combining Latency & Power
  - Power-Delay-Product (PDP)
  - Energy-Delay-Product (EDP)

Model Performance

	BEVDet		PointPillars	
	mAP	NDS	mAP (BEV)	mAP (3D)
GPU (w/o quantization)	0.34	0.34	70.90	65.25
GPU (w/ quantization)	0.31	0.31	66.58	57.94

Resource Utilization on FPGA

	Resource utilization				Performance (GOP/s)		Memory Bandwidth (MB/s)	
	LUT	Register	DSP	URAM	DPU1	DPU2	DPU1	DPU2
Available resource on board	52161	98249	710	68	-	-	-	-
BEVDet	50951	97923	710	46	74.169	92.76	3413.031	1692.983
PointPillars	49281	97100	690	46	1.903	92.156	3890.906	1785.417

FPGA resource setup: In BEVDet, DPU1 is used for image encoder and DPU2 is utilized for BEV encoder with detection head. In PointPillars, DPU1 is utilized for LiDAR encoder and DPU2 is used for detection head.

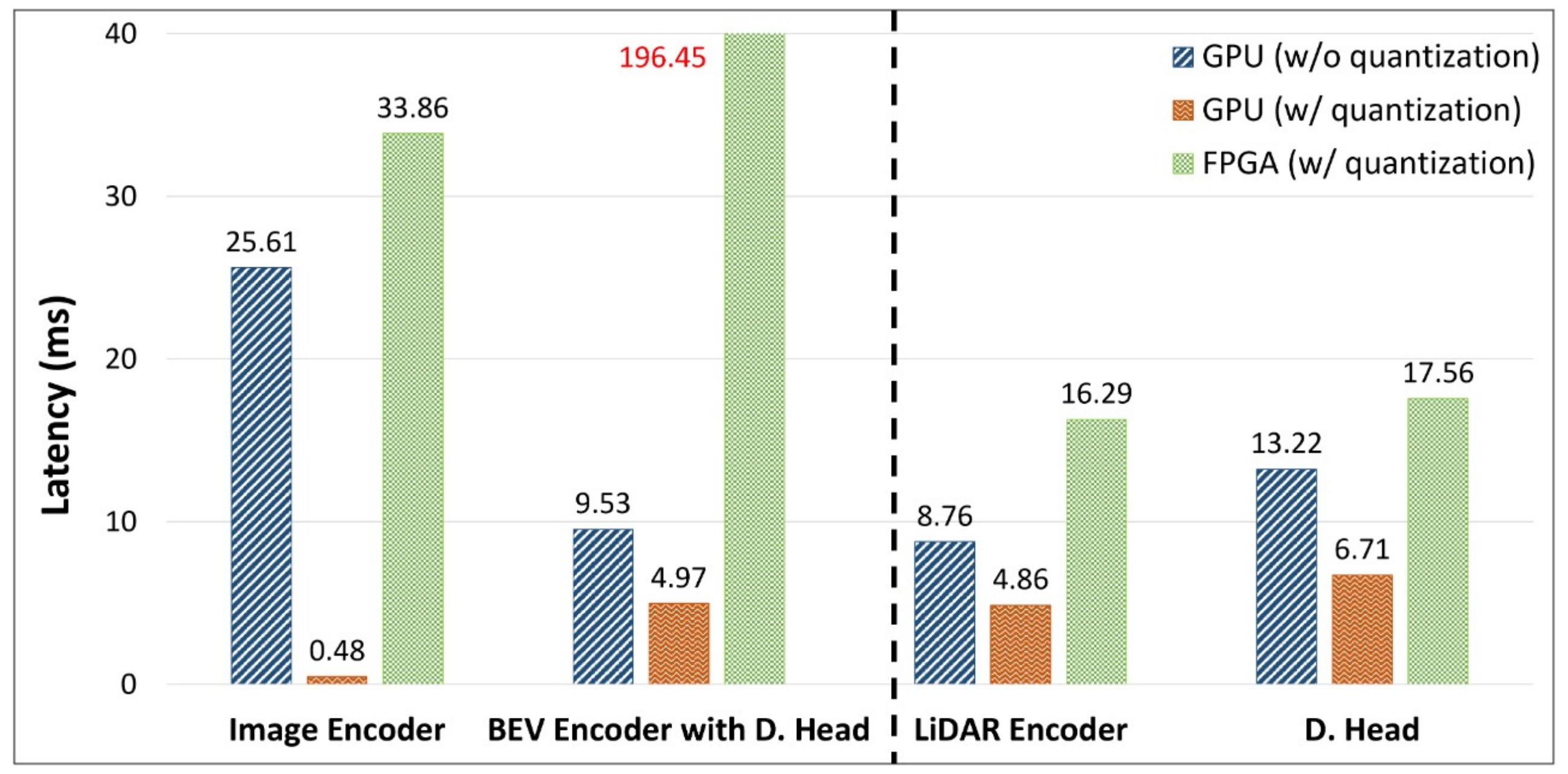
Observation 1: Both models utilize ~93% of the available resources. While detection head consumes more resources than other modules.

Power Usage on Heterogeneous Platforms

	BEVDet				PointPillars			
	Image Encoder		BEV Encoder with D. Head		Lidar Encoder		D. Head	
	Avg. Power (W)	Max Power (W)	Avg. Power (W)	Max Power (W)	Avg. Power (W)	Avg. Power (W)	Avg. Power (W)	Max Power (W)
GPU (w/o quantization)	257.2	267.3	87.5	88.2	248.1	251.6	122.3	126.5
GPU (w/ quantization)	224.5*	228.8*	107.0*	111.0*	222.9	244.2	63.1	66.4
FPGA (w/ quantization)	18.0	21.2	18.9	21.4	18.0	20.5	16.5	19.5

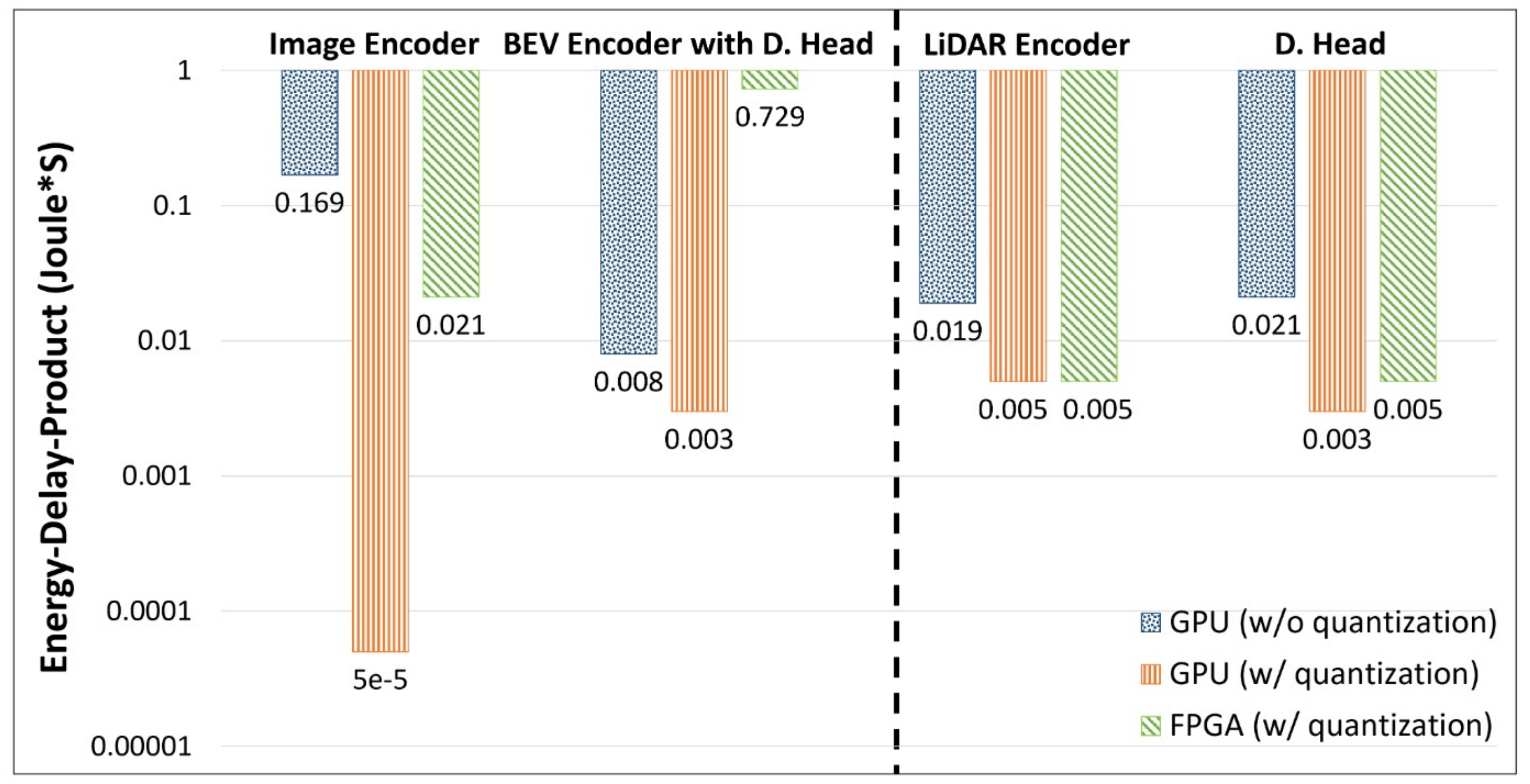
Observation 2: For GPU, the Image encoder in BEVDet and LiDAR encoder in PointPillars consume more power than the Detection head. However, the scenario is the opposite for the FPGA.

Power-Delay-Product (PDP)



Observation 3: The quantized BEVDet on the GPU is superior for PDP performance. Lower PDP achieved on FPGA for LiDAR encoder and detection head.

Energy-Delay-Product (EDP)



Observation 4: The quantized image encoder and BEV encoder with a detection head perform better on the GPU. The quantized LiDAR encoder and its detection head are equally efficient on both the GPU and FPGA.

Future Work

- ❖ Evaluate the Fusion module on FPGA.
- ❖ Focus on the different hardware communication before designing software-hardware multi-modal model.
- ❖ Hardware-friendly different ONNX models for each modules.

Acknowledgment

This work was partly supported by the U.S. National Science Foundation under Grants CNS-2245729 and Michigan Space Grant Consortium 80NSSC20M0124.