Semantic Guided Fine-grained Point Cloud Quantization Framework for 3D Object Detection

<u>Xiaoyu Feng</u>, Chen Tang, Zongkai Zhang, Wenyu Sun, Yongpan Liu

Tsinghua University, Beijing, China

Self Introduction

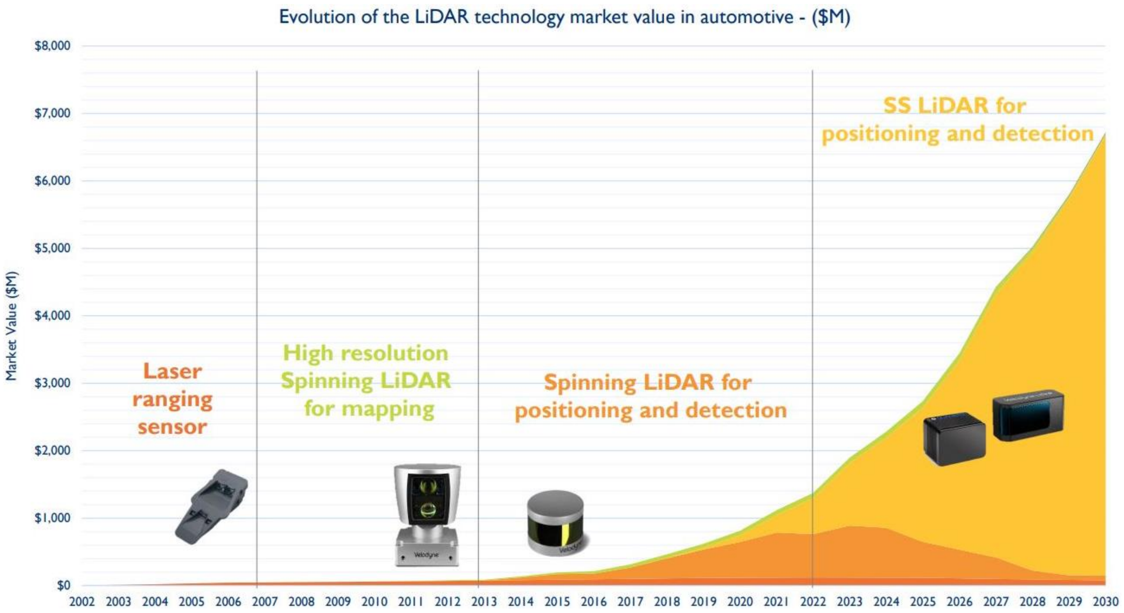
- B.Eng. Degree from Tsinghua University, Beijing, China, in 2018
- Ph.D. candidate at Tsinghua University, Beijing, China
- My research interests are in neural network algorithms & energy-efficient architecture design for 3D point cloud
  - Outdoor 3D Object Detection Algorithms
  - Energy-efficient Architecture Design for Point Cloud

Outline

- Introduction to 3D Object Detection
- Background
  - Neural Network Quantization
  - Compression for Point Cloud Networks
- Methods
- Experiment results
  - Software Evaluation
  - Hardware Evaluation
- Conclusion

Wide Utilization of 3D Object Detection

- Rapid development of Lidar/Radar pushes the wide utilization of point cloud



Yole Development, 2017

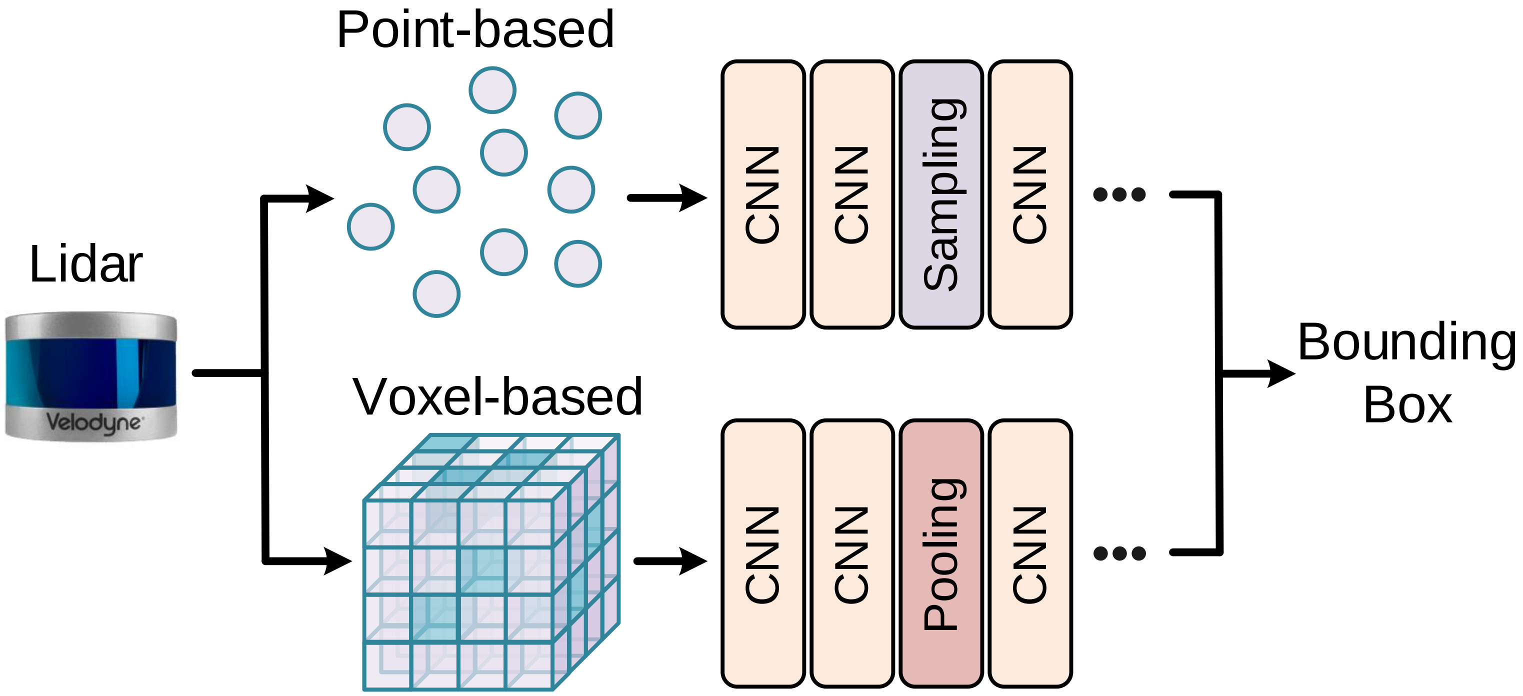
- In many point cloud based applications, 3D object detection is playing critical role

Autonomous Driving

Smart City

Network Based 3D Object Detection

- CNN is the main operator for 3D object detection
  - Two mainstreams: point and voxel based

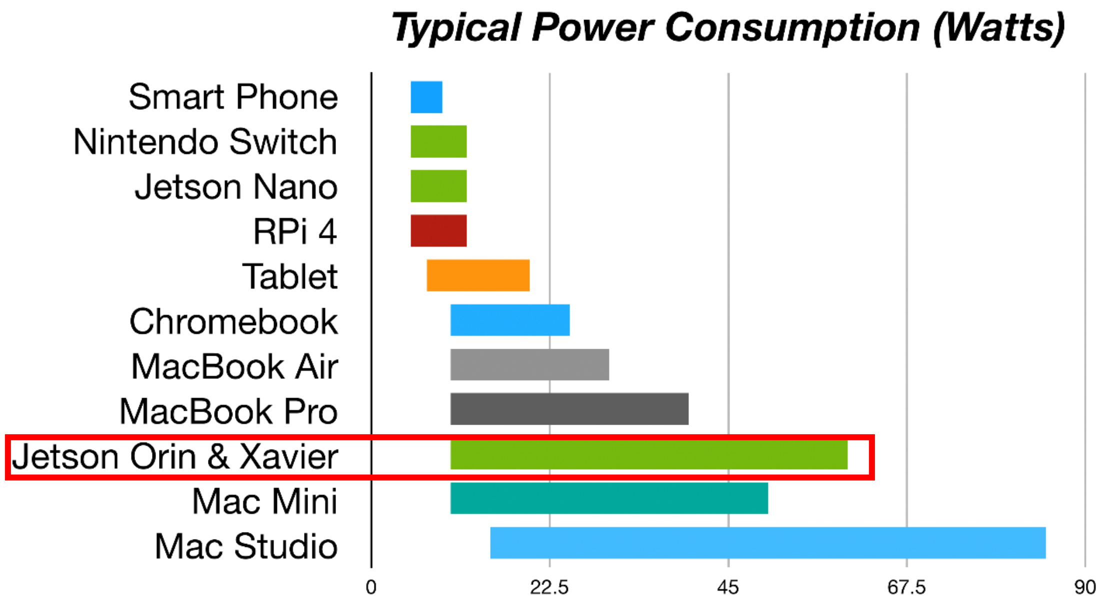


☹ Limited to small-size scenarios
☹ Information loss by sampling

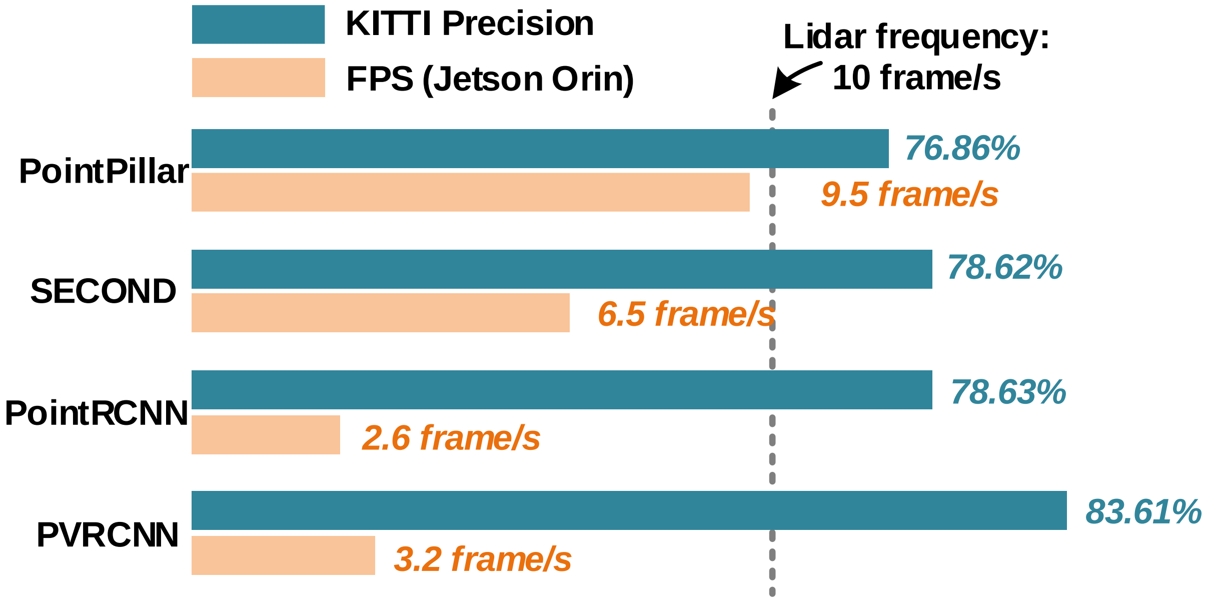
☺ Suitable for large-scale scenarios
☺ Maintain most information

Challenges of Real-time Requirement

- To meet the real-time requirement, energy-hungry hardware is required.
- However, there is still gap for real-time processing

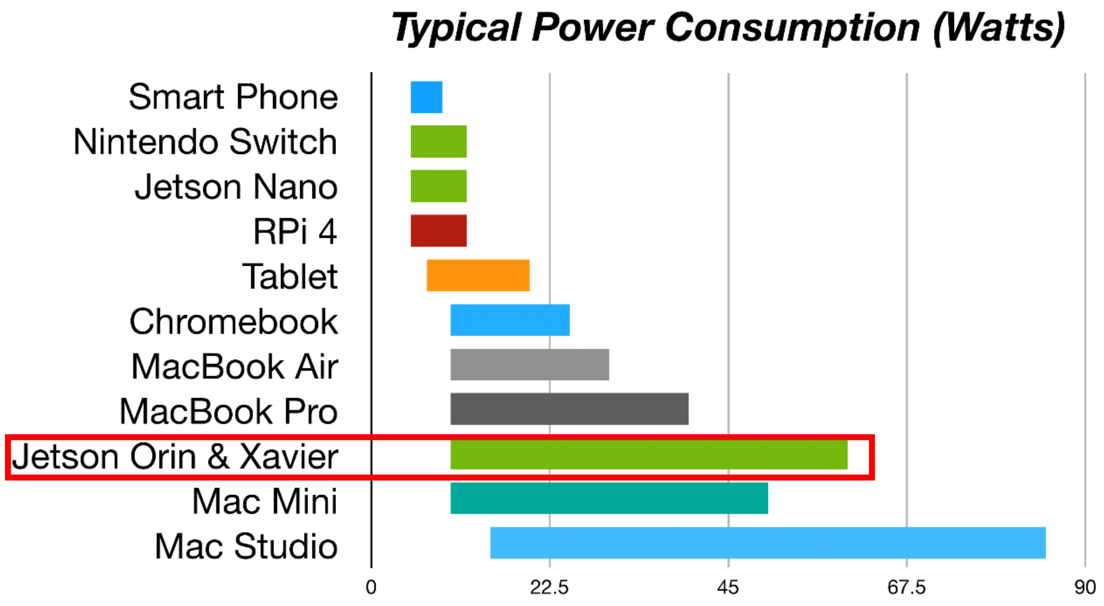


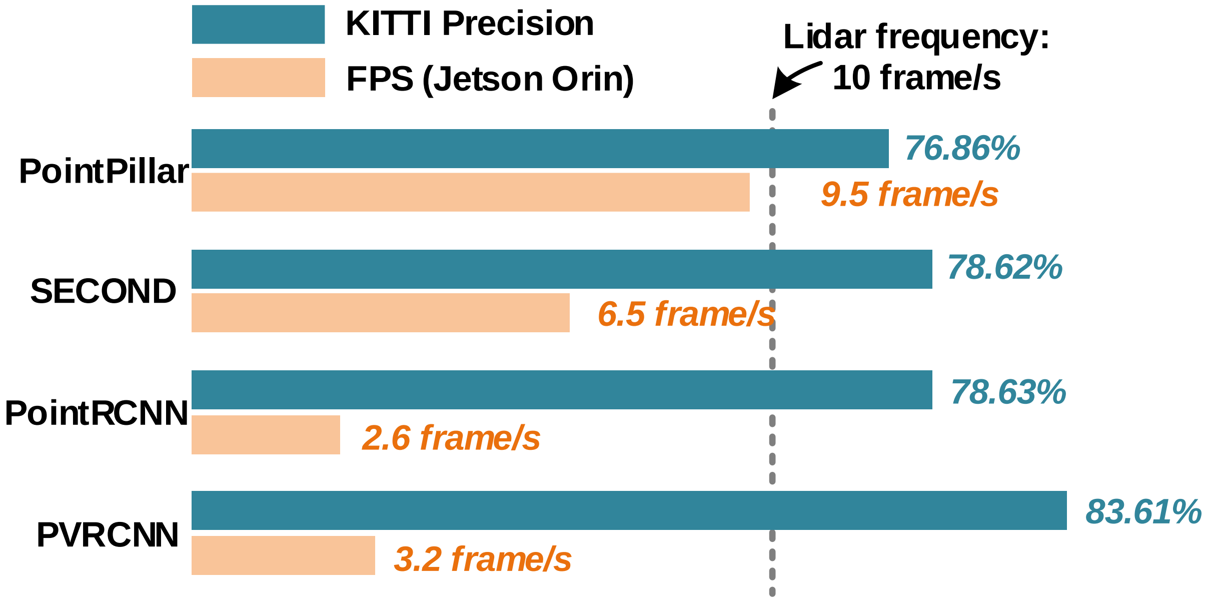
https://jetsonhacks.com/



Challenges of Real-time Requirement

- To meet the real-time requirement, energy-hungry hardware is required.
- However, there is still gap for real-time processing





Call for efficient compression for 3D object detection networks!!!

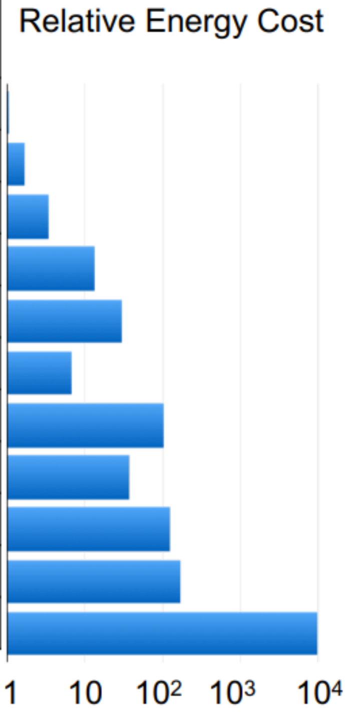
Outline

- Introduction to 3D Object Detection
- Background
  - Neural Network Quantization
  - Compression for Point Cloud Networks
- Methods
- Experiment results
  - Software Evaluation
  - Hardware Evaluation
- Conclusion

Neural Network Quantization

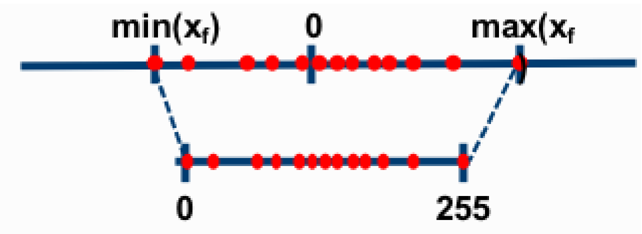
- Low-bit quantization saves the power consumption

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

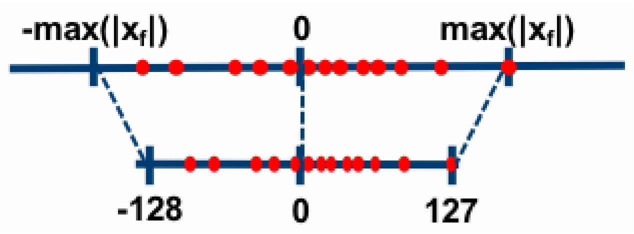


Emer et al., ISCA Tutorial (2019)

- Types:



Asymmetric

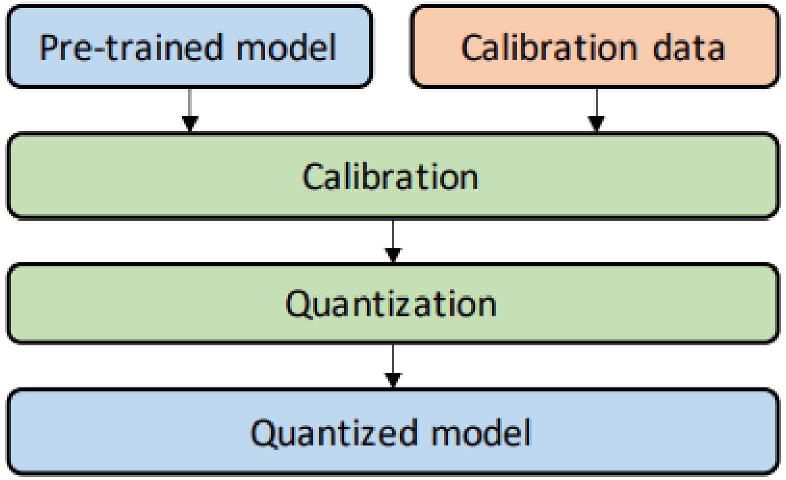


Symmetric

- Methods:
  - Post-Training Quantization
  - Quantization-aware Training

Post-Training Quantization (PTQ)

- PTQ directly quantizes the model without finetuning



Gholami et al., 2021

- Weight Correction
- Range Clipping
- Channel-wise Quantization
- Adaptive rounding

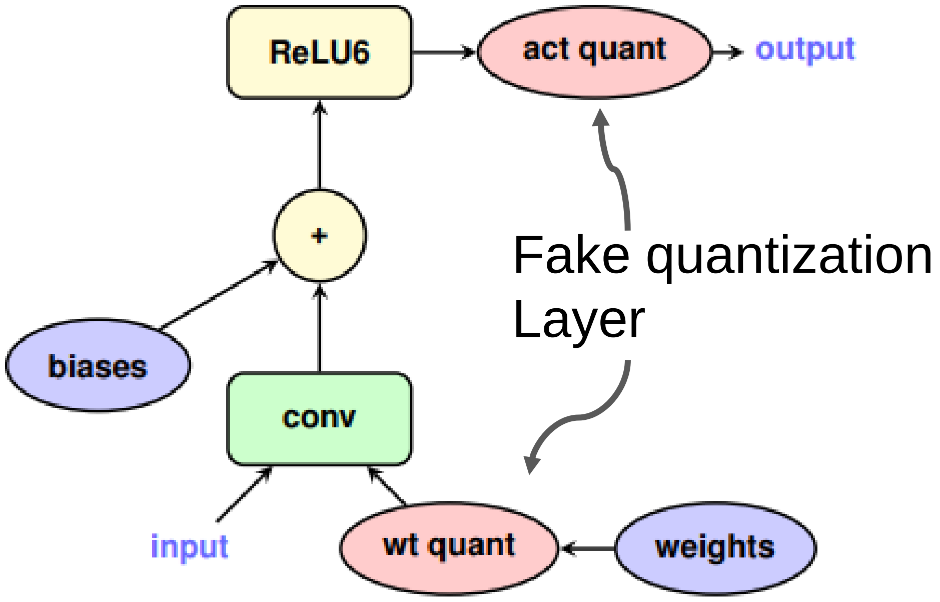
……

☺ No training

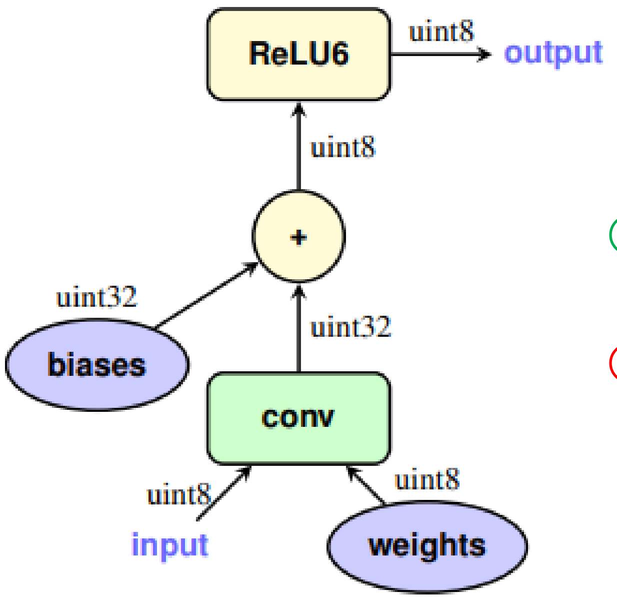
☹ Low model precision

Quantization-aware Training (QAT)

- QAT simulates the effects of quantization during training



Training



Inference

☺ High model accuracy

☹ Long training time

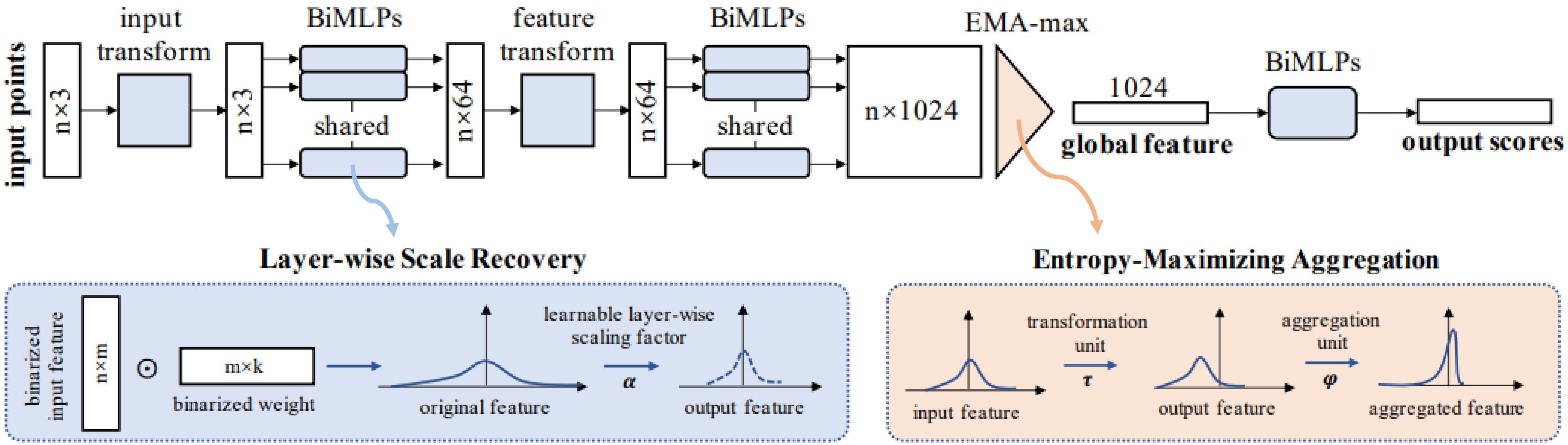
Benoit et al., CVPR, 2018

Outline

- Introduction to 3D Object Detection
- Background
  - Neural Network Quantization
  - Compression for Point Cloud Networks
- Methods
- Experiment results
  - Software Evaluation
  - Hardware Evaluation
- Conclusion

Point Cloud Quantization

- Binary Point Cloud Network

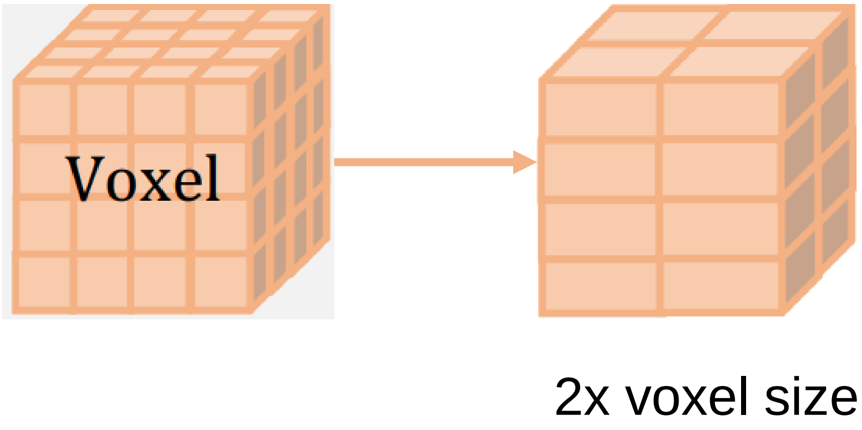


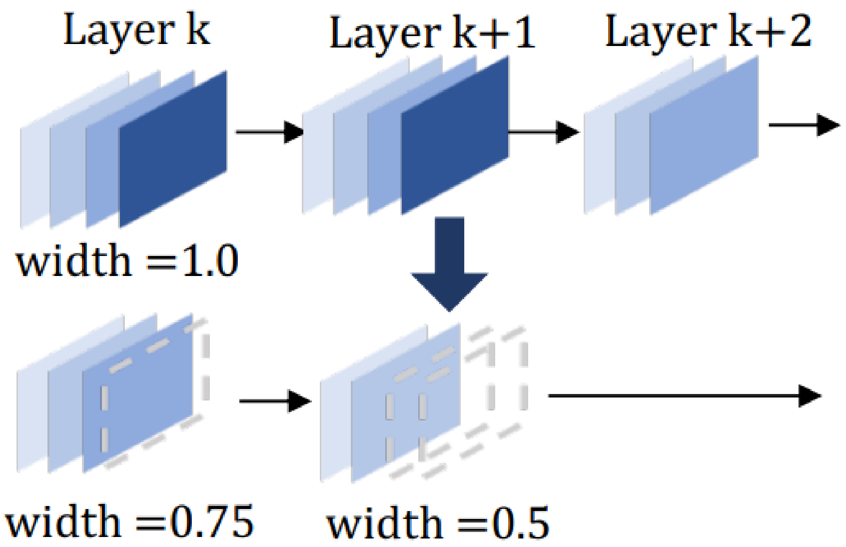
☹ Limited to small network/dataset ☹ Static quantization

Qin et al., ICLR, 2021

Point Cloud Pruning

- Binary Point Cloud Network





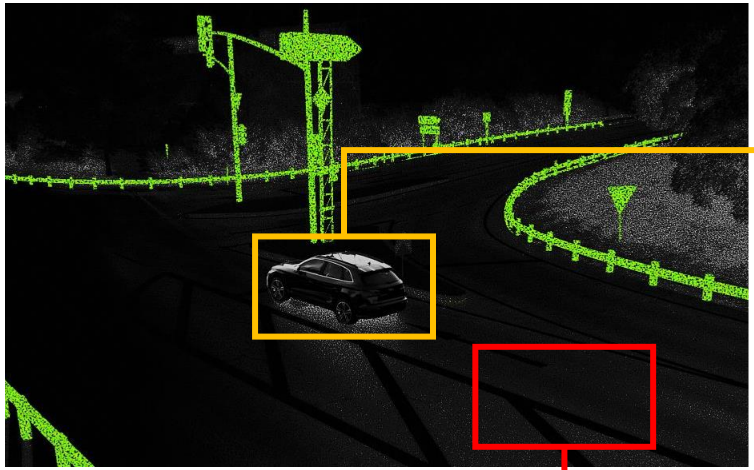
Input Compression **Layer/Channel Compression**

☹ Static pruning

Yang et al., Neurips, 2022

Dynamic Point Cloud Compression

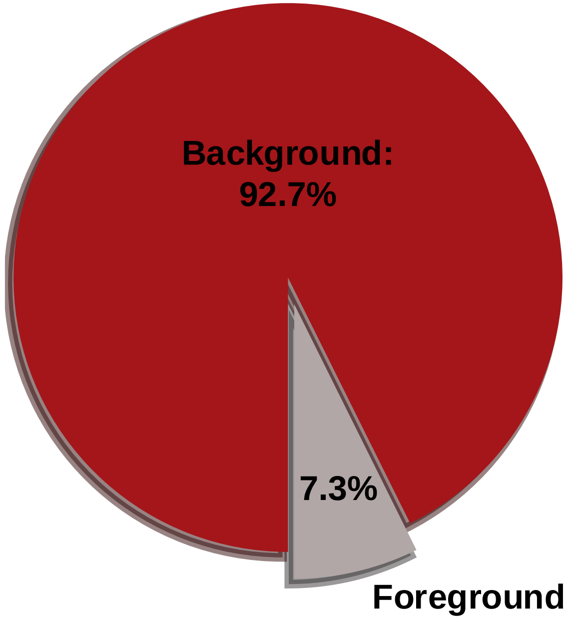
- Useful semantic information in point cloud is imbalanced!



Foreground points:
☺ Small but useful

Background points: ☹ Large but useless

KITTI Dataset



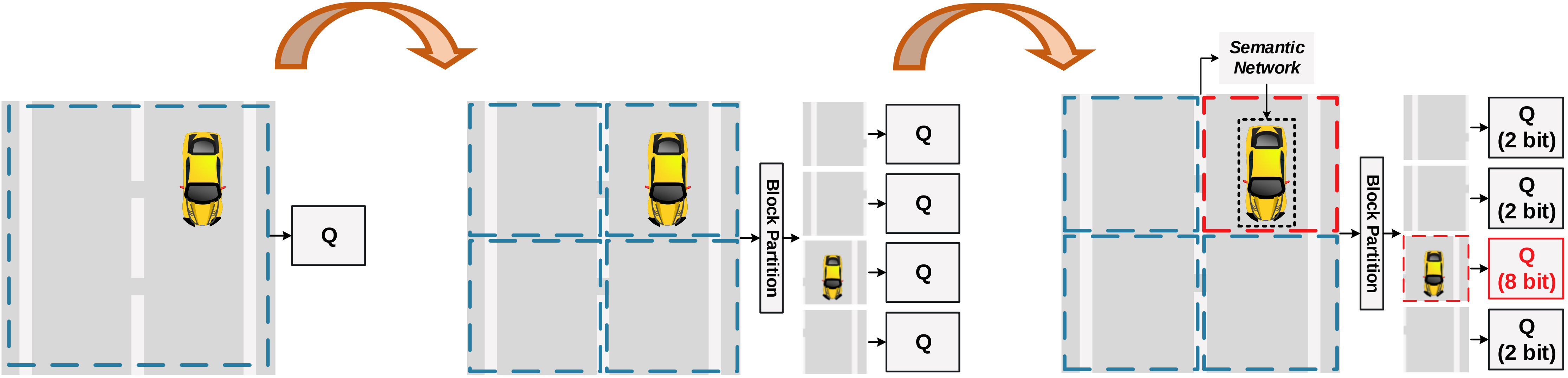
It provides opportunity for dynamic point cloud compression !!!

Outline

- Introduction to 3D Object Detection
- Background
  - Neural Network Quantization
  - Compression for Point Cloud Networks
- Methods
- Experiment results
  - Software Evaluation
  - Hardware Evaluation
- Conclusion

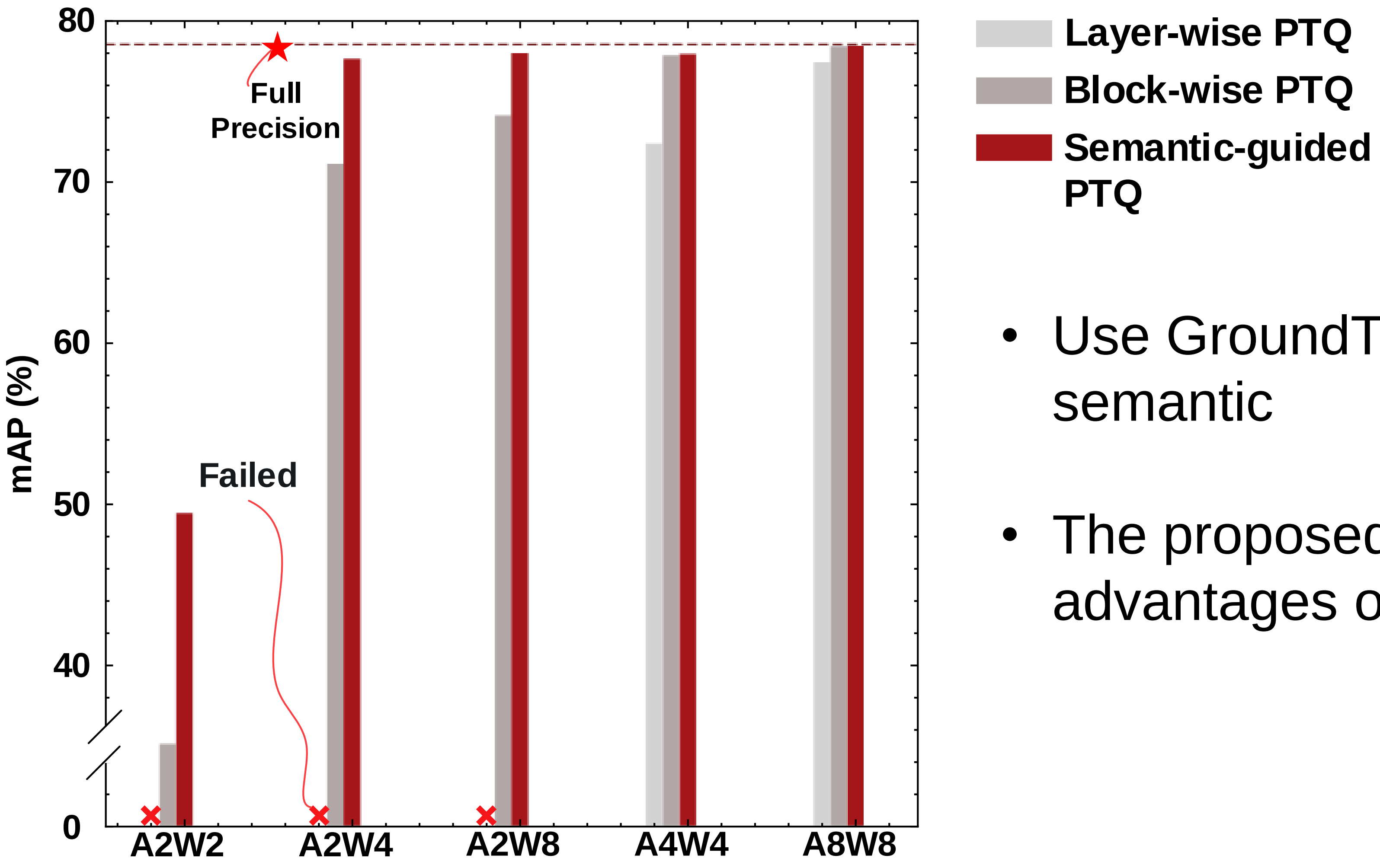
Motivation

- Semantic-guided dynamic quantization



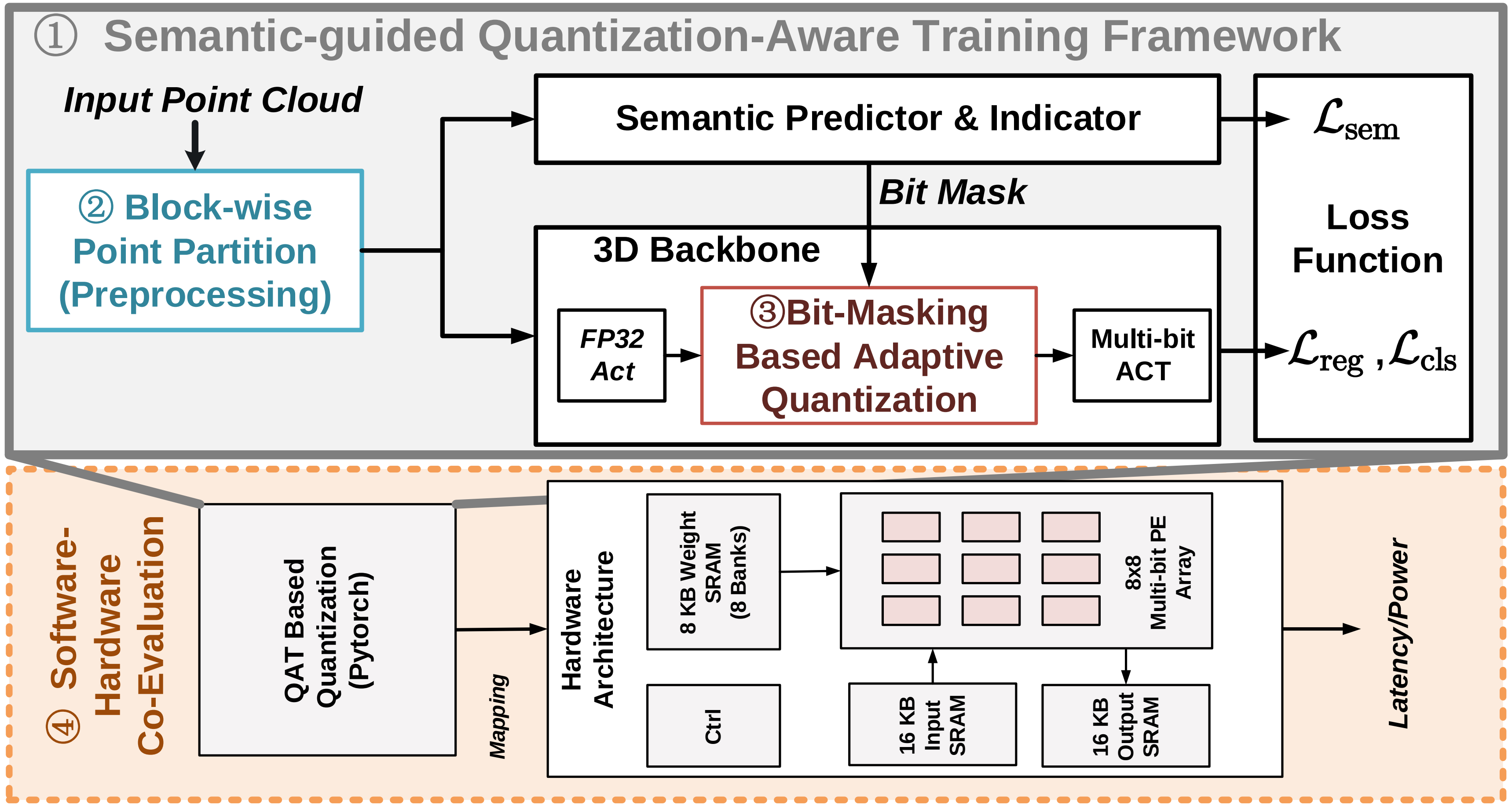
- **Baseline:** Static global quantization
- **Stage I:** Static block-wise quantization
- **Stage II:** Dynamic block-wise quantization

A KITTI Case Study



- Use GroundTruth to indicate semantic
- The proposed method shows advantages on low bitwidth

Overall Framework



①: Indicate semantic for each point with a small prediction branch

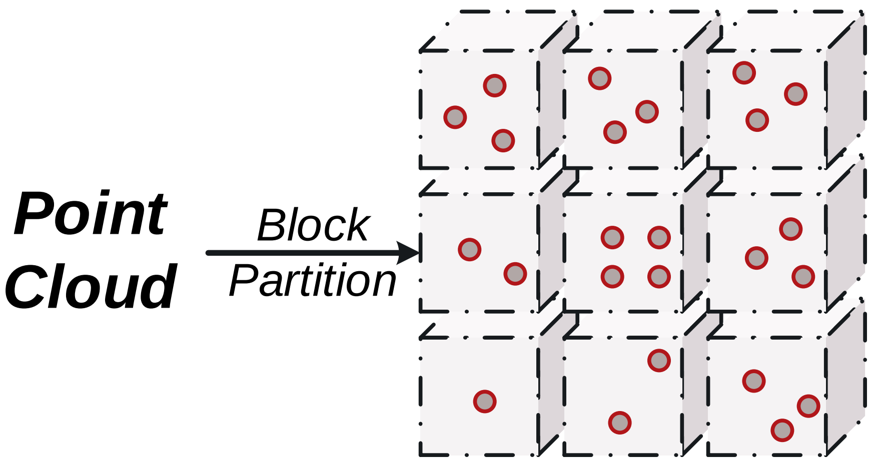
②: Partition point cloud into multiple separate blocks

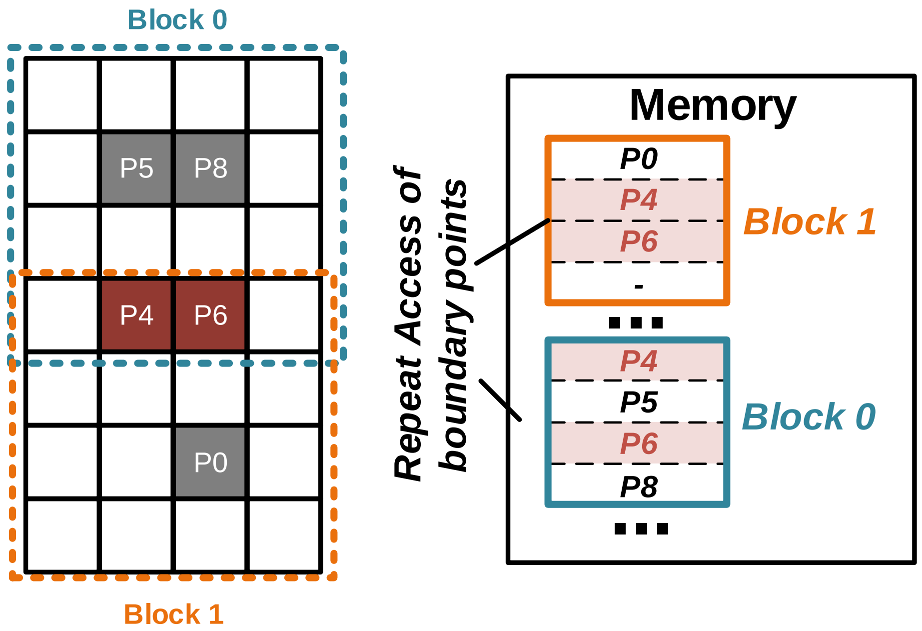
③: Bit-masking transfers the semantic indicator of each block into bitwidth

④: Evaluate the quantized model on both software & hardware ends

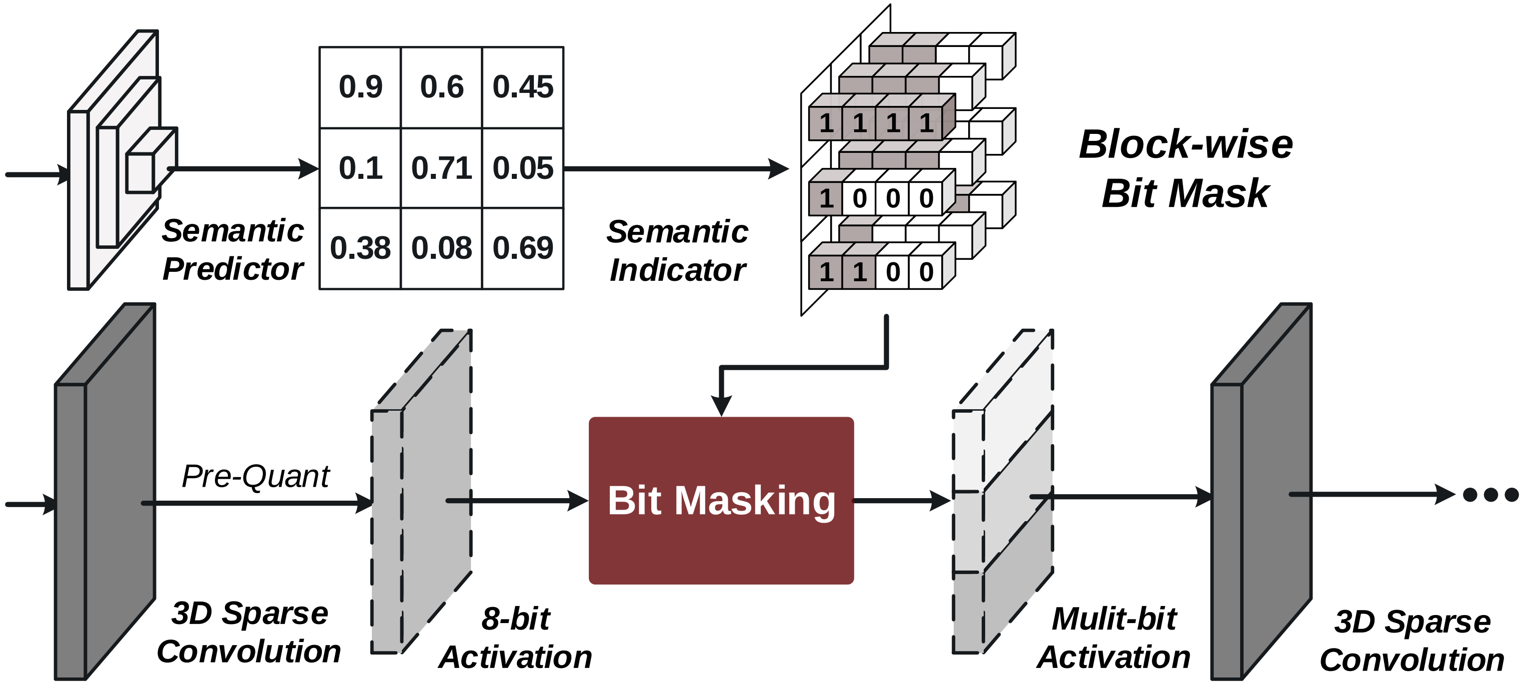
Block Partition

- Block partition narrows dynamic range and benefits quantization
  - RGB image: 8 bit
  - LiDAR point cloud: 11 bit for x, y
- Block partition results in repeat memory access for boundary points
  - Theoretical estimation: $(S+2)^3/S^3$ where S is block size



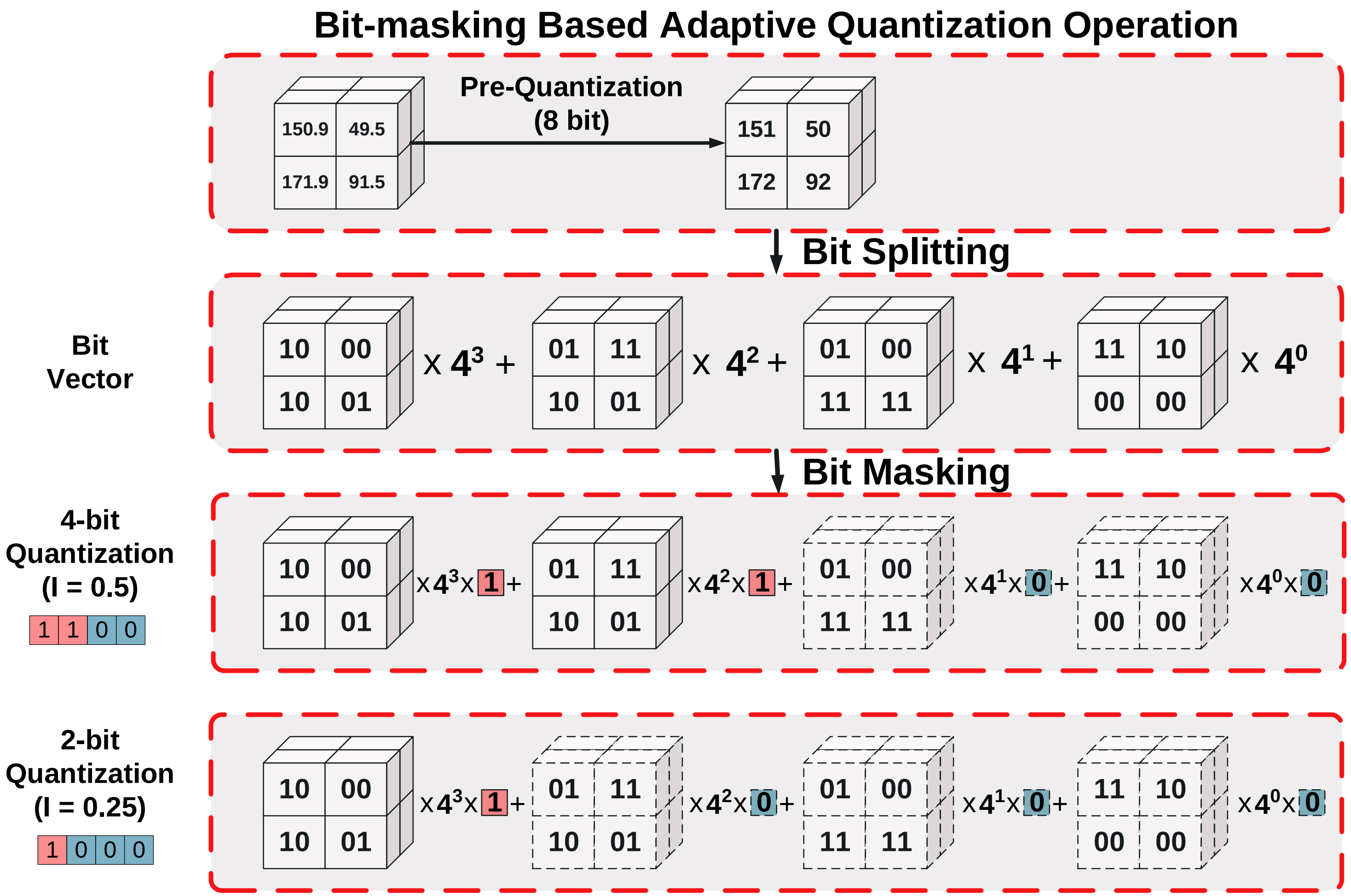


Bit-masking Based Adaptive Quantization



- For each block B_i, the semantic branch predict a 0-1 indicator I_i
- Then I_i is transferred into a 0/1 bit mask m_i
- The floating-point X_f is first 8-bit pre-quantized into X_q^{pre}
- Then X_q^{pre} is adaptively quantized into different bitwidth X_q by bit-masking

Bit-masking Based Adaptive Quantization



- A 4-bit example of bit-masking:
  - The 8-bit pre-quantized X_q^{pre} is transferred into a 0-1 bit vector v
  - v is multiplied with bit mask m and drop the LSBs
  - The higher semantic indicator is, the more LSBs are kept

Outline

- Introduction to 3D Object Detection
- Background
  - Neural Network Quantization
  - Compression for Point Cloud Networks
- Methods
- Experiment results
  - Software Evaluation
  - Hardware Evaluation
- Conclusion

Dataset Benchmark

KITTI

- Mainly ***car*** and ***pedestrian*** class
- Train on the 3712 training samples and validate on 3769 validation samples.
- 3D Average Precision for evaluation

NuScenes

- A 10-class 3D object detection dataset
- Train on 28k training frames and validate on 6k validation frames
- NuScenes Detection Score (NDS) as main evaluation metric

Quantization on KITTI

Methods	Weight	Activation	*Car* (%)	*Ped* (%)
Baseline	FP32	FP32	78.67	54.65
Layer-wise	2 bit	2 bit	75.35	38.92
Block-wise	2 bit	2 bit	77.68	47.04
This work	2 bit	2.28 bit	**78.01**	**48.63**
Layer-wise	2 bit	1 bit	62.41	16.58
Block-wise	2 bit	1 bit	66.08	21.35
This work	2 bit	1.25 bit	**71.25**	**41.41**
Layer-wise	1 bit	1 bit	62.35	16.67
Block-wise	1 bit	1 bit	64.31	19.82
This work	1 bit	1.25 bit	**70.91**	**38.71**

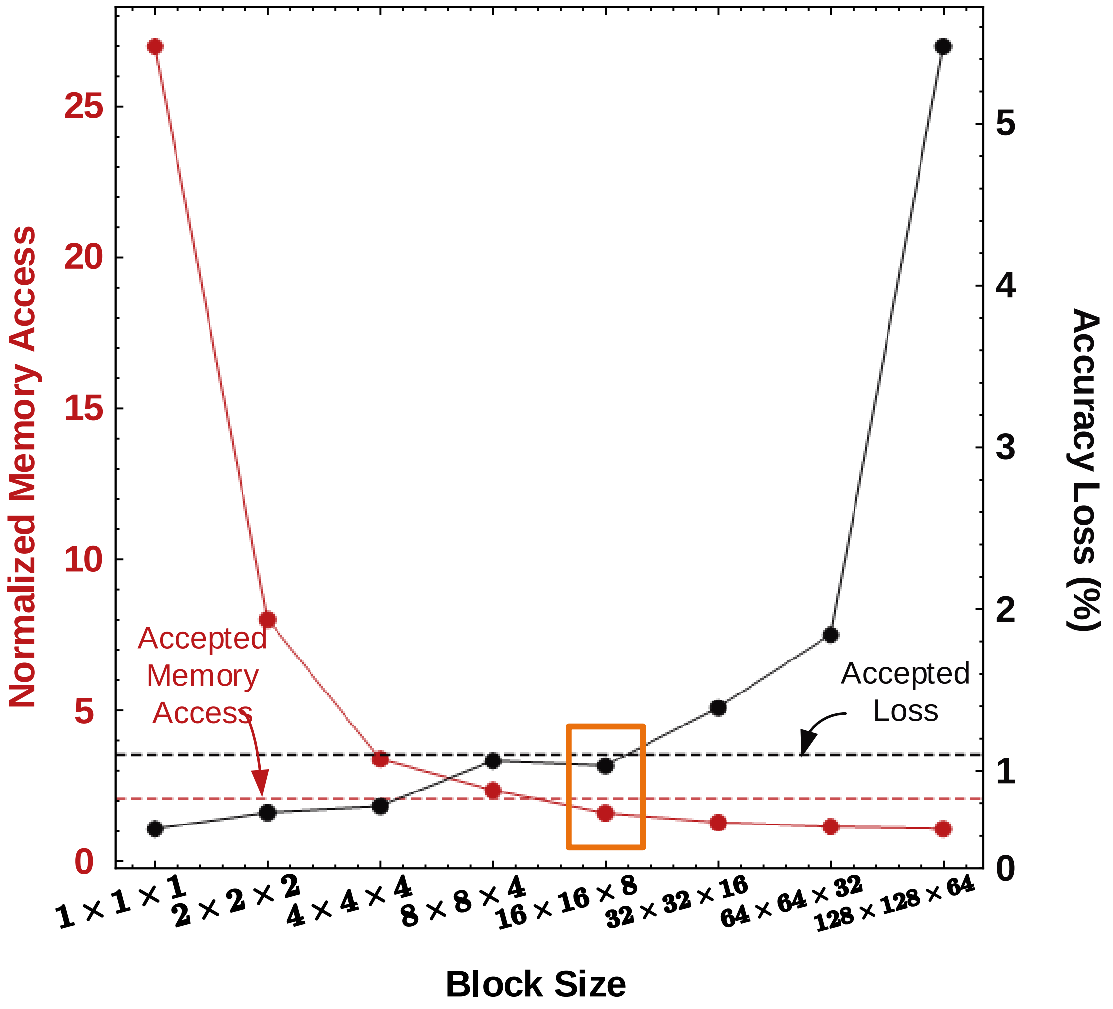
- Compared with layer-wise (global) quantization, block-wise quantization shows **2.33%/1.96%** higher ***car*** AP under **2/1 bit**.
- Adopting semantic-guided dynamic quantization improves **2.96%/7.56%** ***car*** AP under **2/1 bit**.
- The ***pedestrian*** class shows higher accuracy loss for less point resolution.

Quantization on NuScenes

Methods	Weight	Activation	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
Baseline	FP32	FP32	0.6519	0.5724	0.3045	0.2579	0.3773	0.2212	0.1826
Layer-wise	4 bit	4 bit	0.5509	0.4367	0.3426	0.2651	0.4906	0.3822	0.1941
Block-wise	4 bit	4 bit	0.6346	0.5432	0.3044	0.2598	0.3864	0.2365	0.1832
This work	4 bit	4.25 bit (Avg.)	**0.6393**	**0.5518**	**0.2994**	**0.2582**	**0.3849**	**0.2345**	**0.1889**
Layer-wise	3 bit	3 bit	0.5160	0.4014	0.3474	0.2671	0.5901	0.4382	0.2044
Block-wise	3 bit	3 bit	0.5355	0.4295	0.3542	0.2624	0.5761	0.4025	0.1967
This work	3 bit	3.51 bit (Avg.)	**0.6300**	**0.5329**	**0.3084**	**0.2582**	**0.3781**	**0.2311**	**0.1891**
Layer-wise	2 bit	2 bit	0.4786	0.3600	0.3702	0.2675	0.6657	0.5009	0.2101
Block-wise	2 bit	2 bit	0.5223	0.4141	0.3579	0.2647	0.6046	0.4231	0.1972
This work	2 bit	2.51 bit (Avg.)	**0.6038**	**0.4926**	**0.3246**	**0.2614**	**0.4057**	**0.2464**	**0.1866**

- **Achieves 8.84%/11.4%/12.52% NDS improvement on 4/3/2 bit**

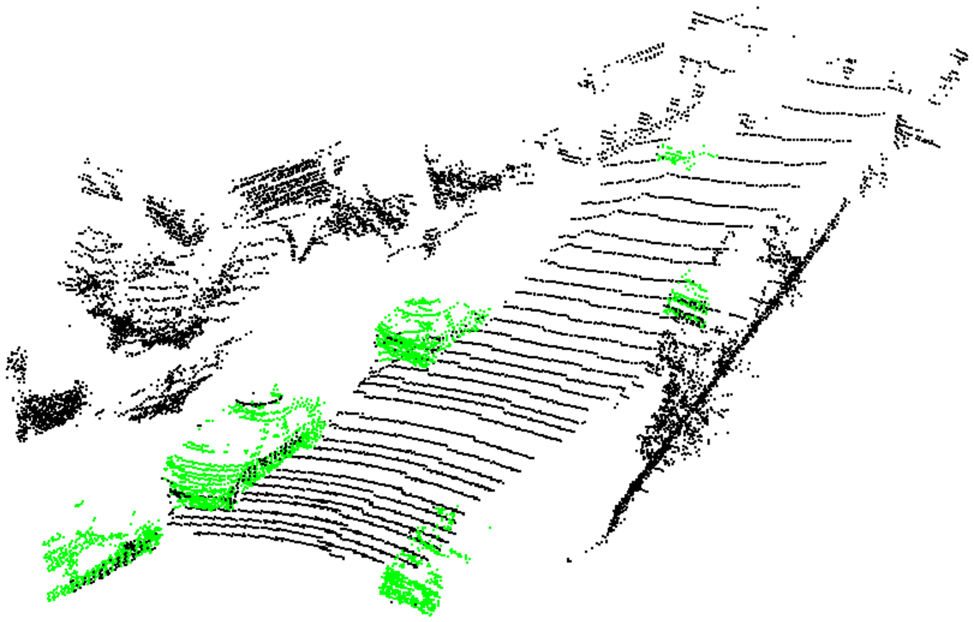
Scan of Block Size



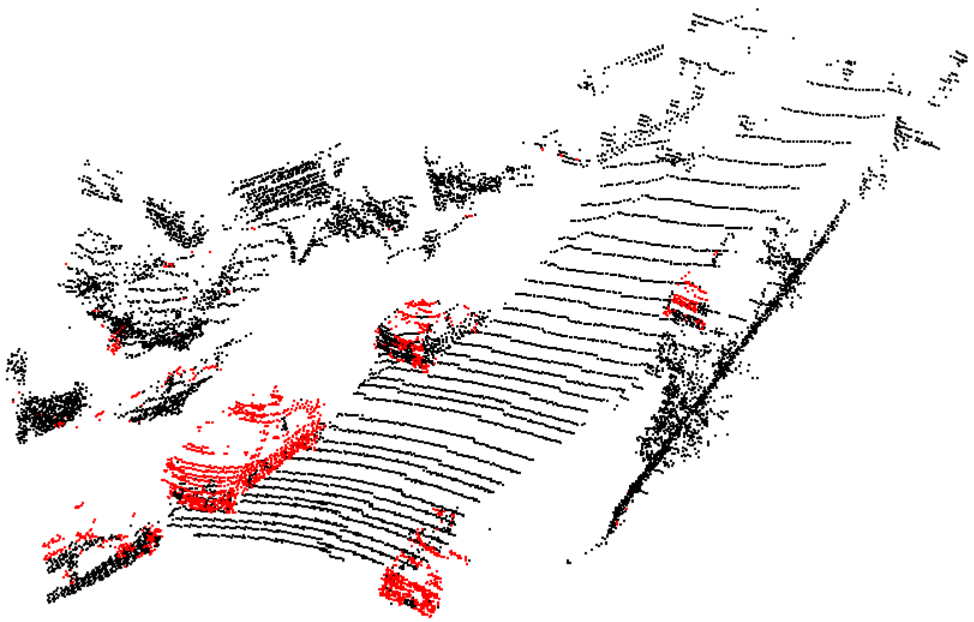
- Larger block size brings less memory access of the boundary points while higher accuracy loss
- Smaller block size brings more extra memory access but lower accuracy loss
- **In the left case, 16x16x8-size block is chosen**

Semantic Prediction Visualization

- Compared with the Ground-Truth, we can predict most foreground points with a small semantic branch



(a) Groundtruth



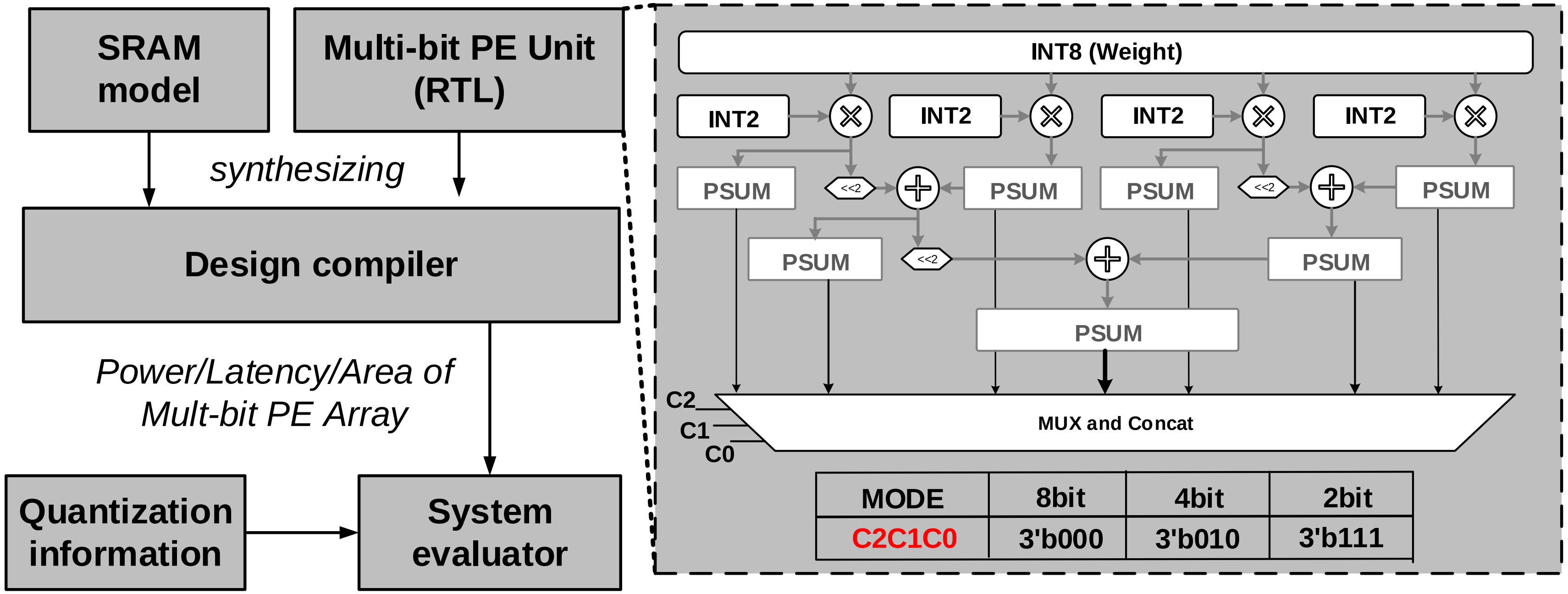
(b) Our Prediction

Outline

- Introduction to 3D Object Detection
- Background
  - Neural Network Quantization
  - Compression for Point Cloud Networks
- Methods
- Experiment results
  - Software Evaluation
  - Hardware Evaluation
- Conclusion

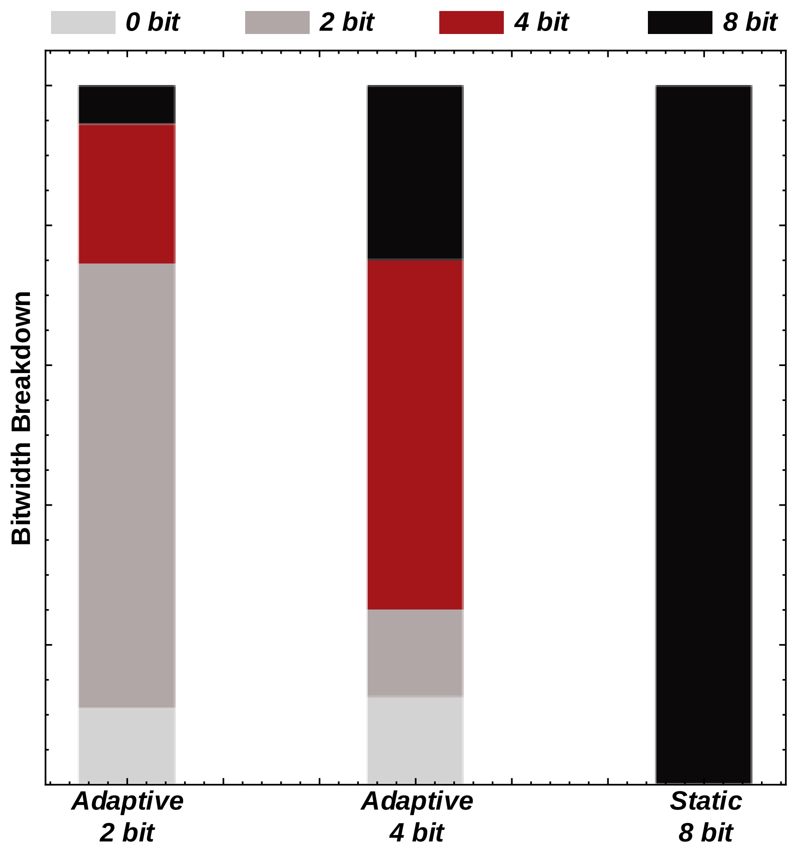
Simulation Setting

- Multi-bit reconfigurable accelerator

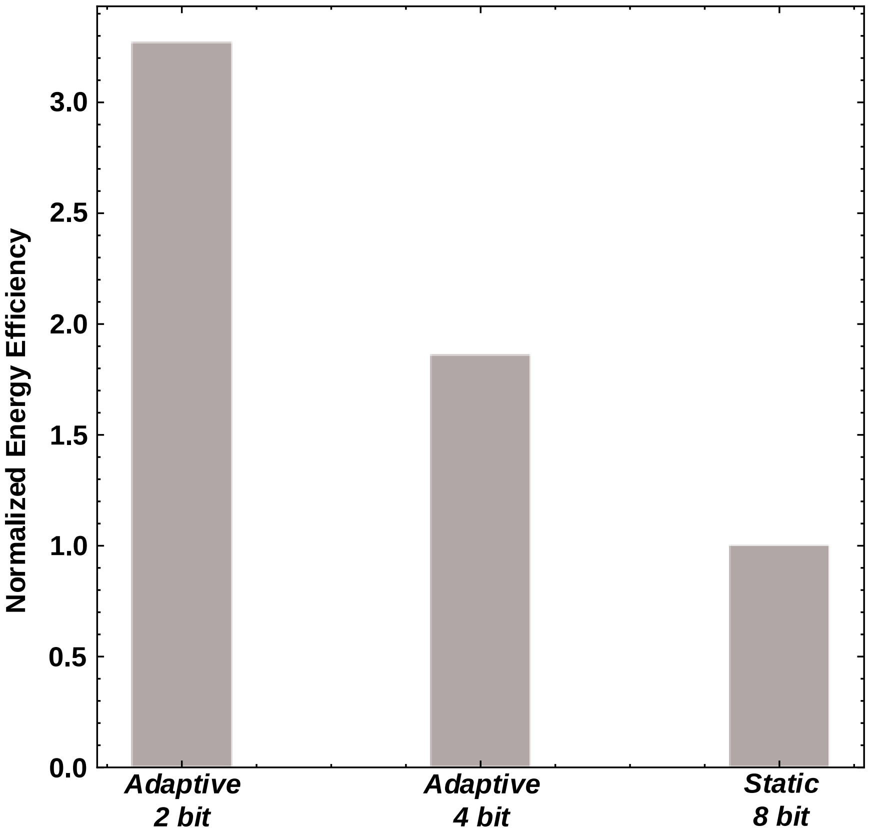


Parameter	Value
Voltage	0.81V
Frequency	200MHz
Weight SRAM	8x8KB
Input SRAM	16KB
Output SRAM	16KB
Technology	TSMC 28nm

Simulation Setting



Bitwidth Breakdown



Peak Energy Efficiency

- 3.11x (2 bit) and 1.86x (4 bit) energy efficiency compared with static 8 bit

Outline

- Introduction to 3D Object Detection
- Background
  - Neural Network Quantization
  - Compression for Point Cloud Networks
- Methods
- Experiment results
  - Software Evaluation
  - Hardware Evaluation
- Conclusion

Conclusion

- The imbalanced distribution of **semantic-rich foreground** points and **semantic-less background** points in LiDAR point cloud provides opportunity for dynamic quantization

- We design a new semantic-guided dynamic quantization for 3D object detection
  - Block-wise partition to handle the large dynamic range in 3D point cloud
  - Bit-masking based quantization to adaptively assign different bitwidth to different blocks

- On NuScenes dataset, we achieve 8.84%/11.4%/12.52% precision improvement on 4/3/2 bit compared with the global quantization baseline

Limitations and Future Work

- In this work, we mainly focus on quantizing the 3D backbone

Quantization Strategy	AP (8-bit quantization)
Only 3D backbone	78.46
Only 2D neck	78.60
Only classification head	76.83
Only regression head	**43.39**
3D backbone + 2D neck + classification head + regression head	**36.53**

Quantization sensitivity of different layers on KITTI

- 3D backbone occupies the main operation
- The FLOPs of head is extremely low while its **quantization sensitivity is very high**!
- How to effectively quantize the detection head is the future work

Outline

Thank you for listening!