# Online Quantization Adaptation for Fault-Tolerant Neural Network Inference

Michael Beyer[1,2], Jan Micha Borrmann[1], Andre Guntoro[1], Holger Blume[2]

[1]Bosch Corporate Research, Robert Bosch GmbH

[2]Institute of Microelectronics Systems, Leibniz University Hannover

SafeComp 2023

BOSCH

Image Source: Bosch

Image Source: Bosch

Image Source: Bosch




BOSCH

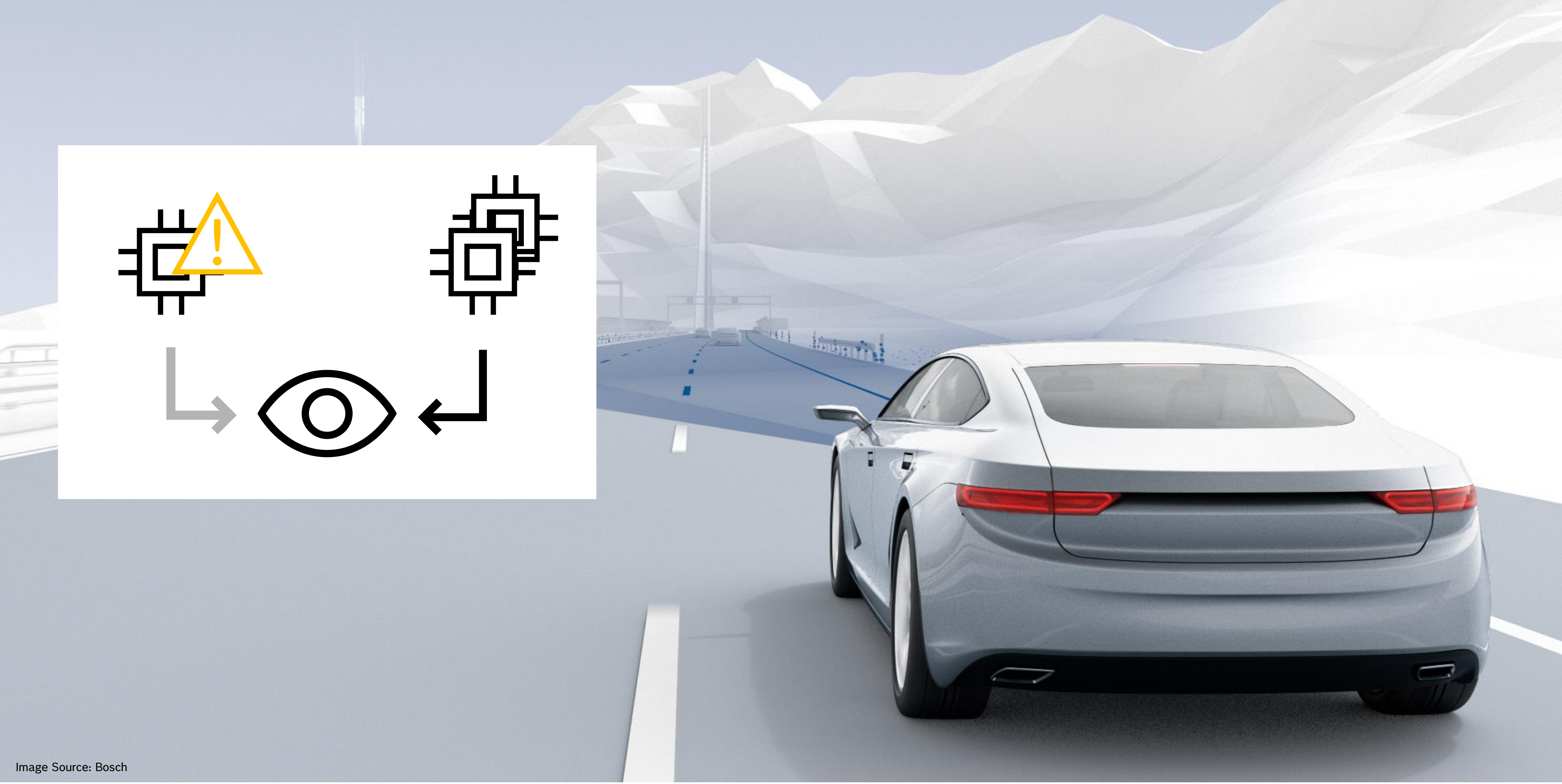


Image Source: Bosch

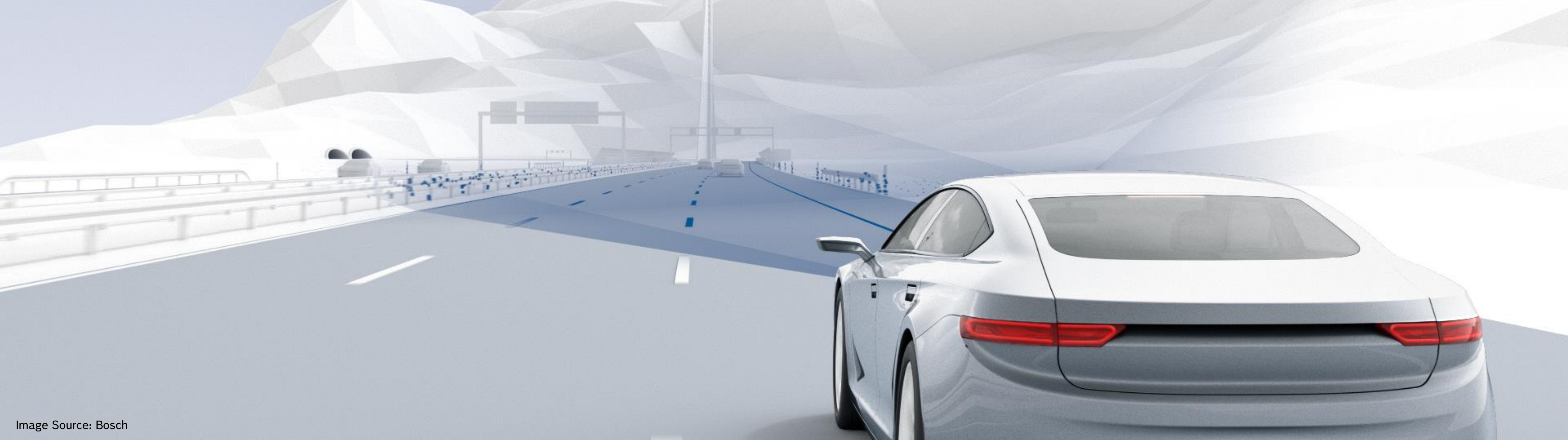


Image Source: Bosch

# Ensuring compute availability for neural network inference

Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security


BOSCH

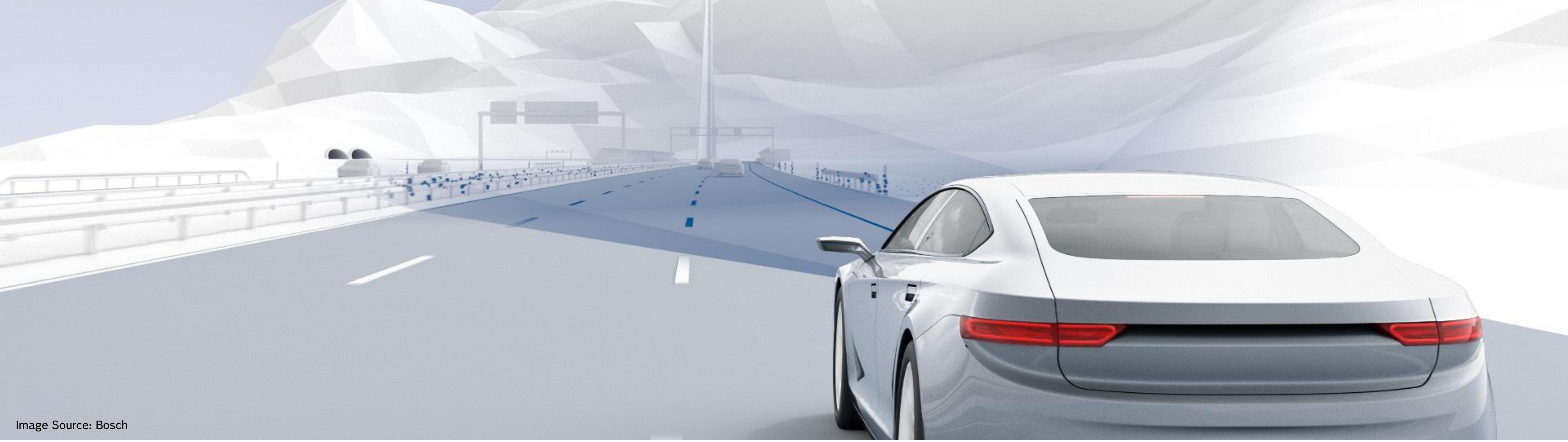


Image Source: Bosch

# Ensuring compute availability for neural network inference

- Traditional redundancy-based methods ➢ High overheads (cost, area, power)

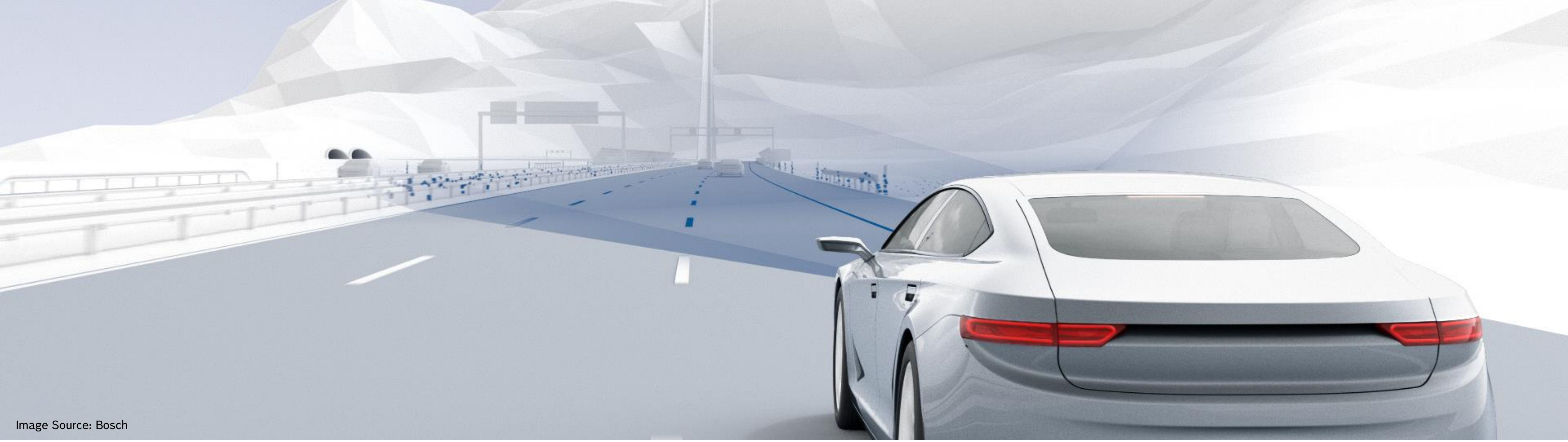


Image Source: Bosch

## Ensuring compute availability for neural network inference

- Traditional redundancy-based methods ➢ High overheads (cost, area, power)
- Adapt to HW faults by retraining NNs ➢ Not possible during runtime

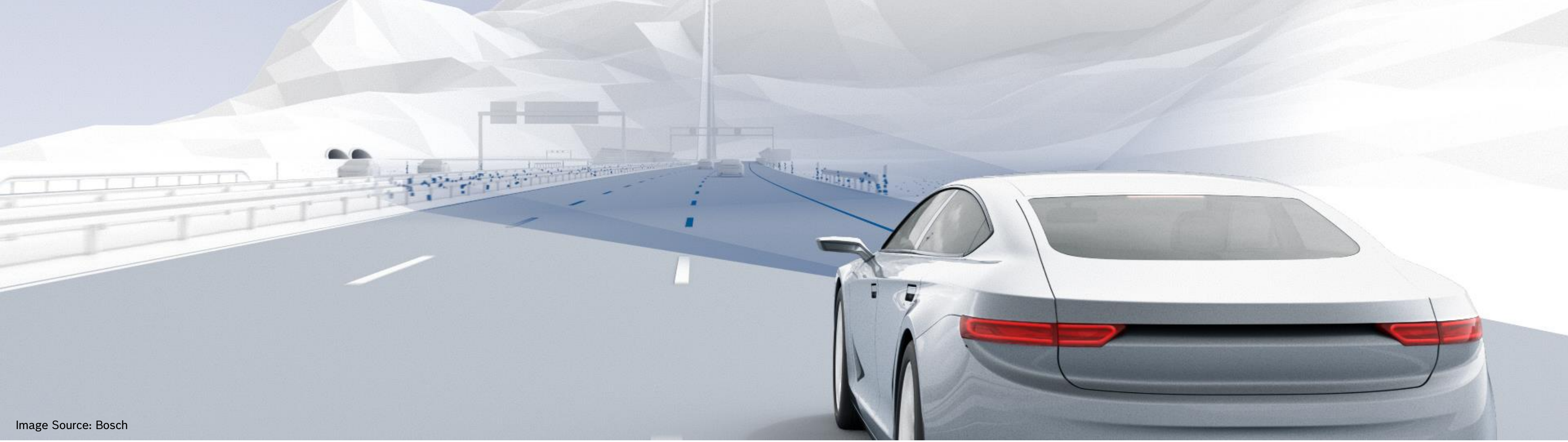


Image Source: Bosch

# Ensuring compute availability for neural network inference

| | |
|---|---|
| ▪ Traditional redundancy-based methods | ➢ High overheads (cost, area, power) |
| ▪ Adapt to HW faults by retraining NNs | ➢ Not possible during runtime |
| ▪ Masking faulty HW elements | ➢ Not guaranteed to maintain algorithmic performance |

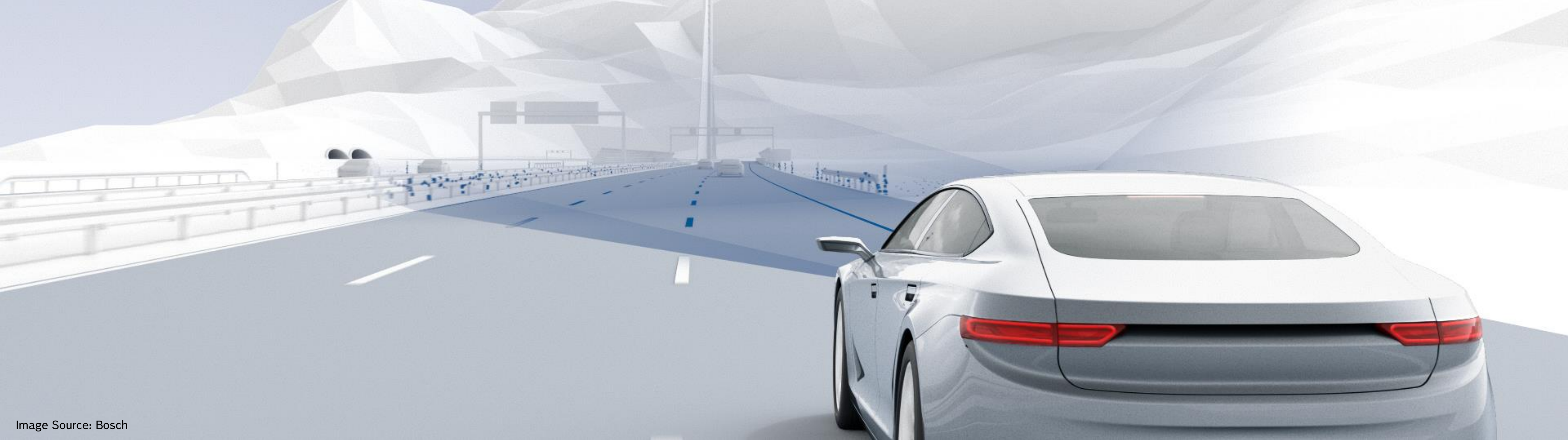


Image Source: Bosch

## Algorithmic Properties

## Dedicated Hardware Features




BOSCH

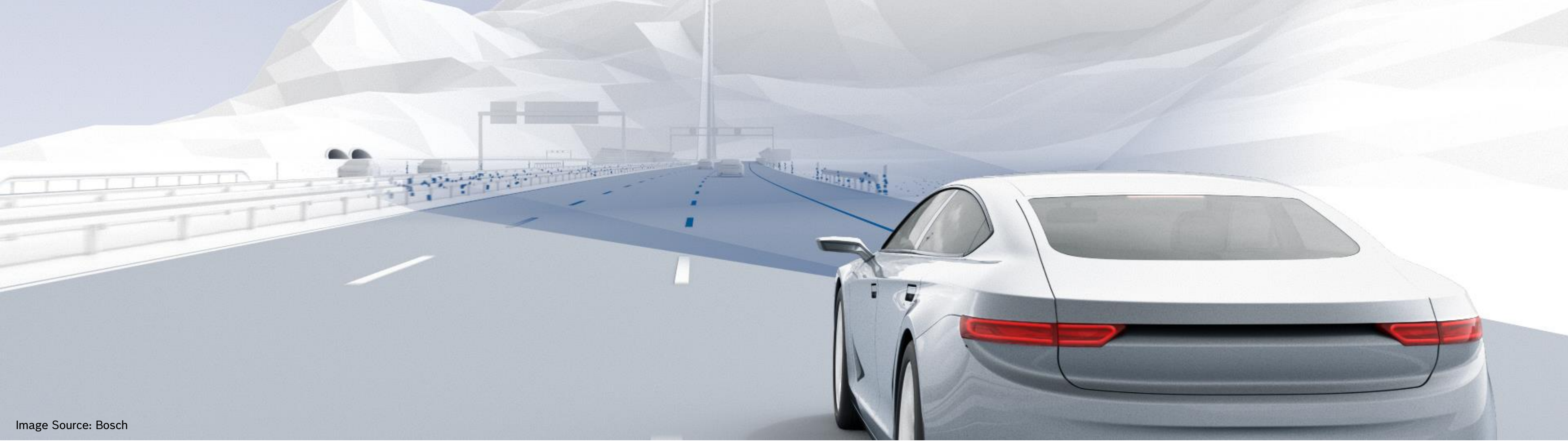


Image Source: Bosch

## Algorithmic Properties

### Quantization

- Floating-Point ⟶ Fixed-Point
- Tolerance to reduced precision computations

## Dedicated Hardware Features

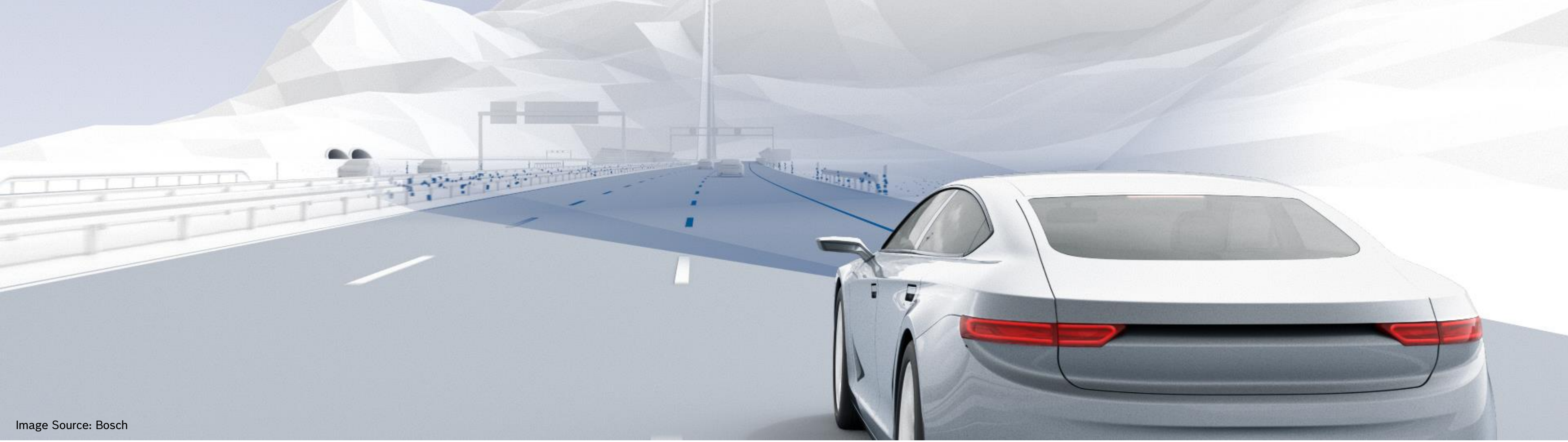


Image Source: Bosch

## Algorithmic Properties

### Quantization

- Floating-Point ⟶ Fixed-Point
- Tolerance to reduced precision computations

## Dedicated Hardware Features

### Multi-bit-width Support

- E.g., 8-bit and 2x 4-bit
- Increased compute performance and flexibility for different workloads

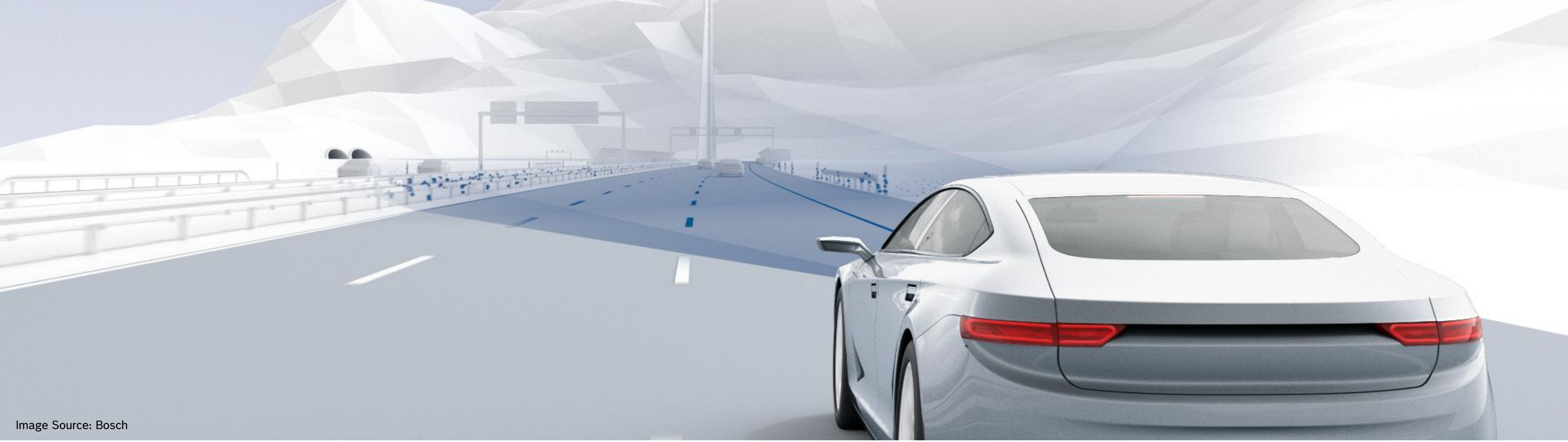


Image Source: Bosch

Algorithmic Properties

Dedicated Hardware Features

**Lightweight Fault Tolerance**

Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# AGENDA

01 **Motivation**

02 **Method**

03 **Experiments**

04 **Conclusion**


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security


BOSCH

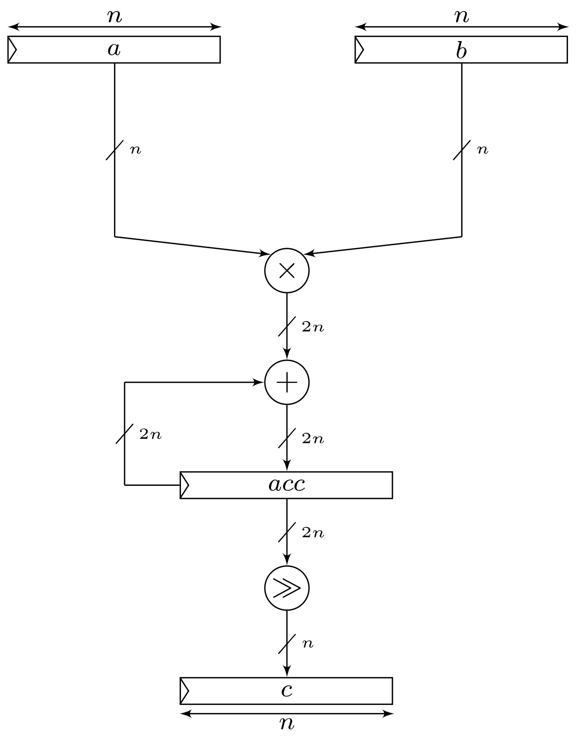
# 02

# Method

# Method
## Background – Multi-Bit-Width MAC Unit

Baseline Multiply-Accumulate (MAC) unit:

$$acc = acc + a \cdot b$$




Baseline MAC Unit


[1] Beyer, M., Gesper, S., Guntoro, A., Payá-Vayá, G., Blume, H., "Exploiting Subword Permutations to Maximize CNN Compute Performance and Efficiency", 34th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2023.

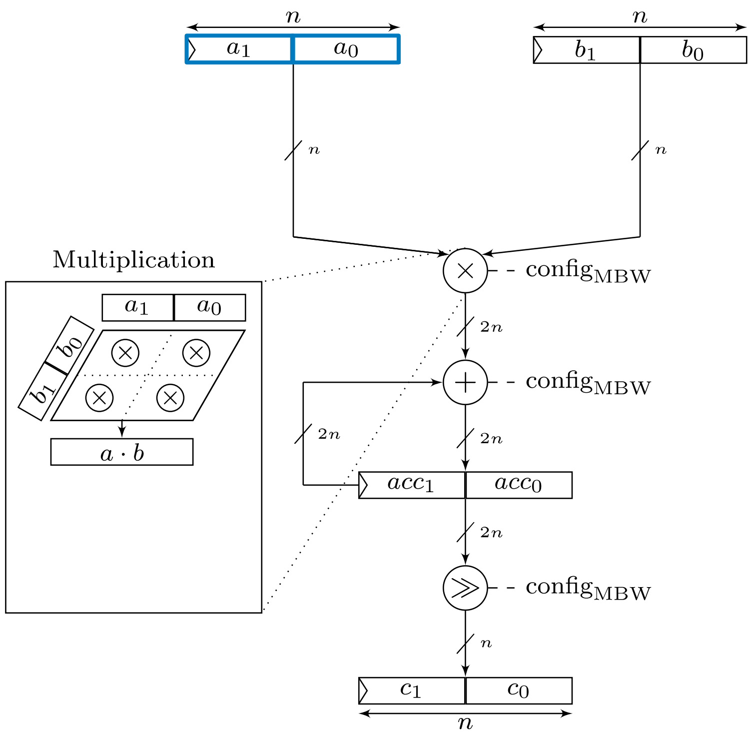
# Method
## Background – Multi-Bit-Width MAC Unit

Baseline Multiply-Accumulate (MAC) unit:

$$acc = acc + a \cdot b$$

With Multi-bit-width support:

$$acc_1 = acc_1 + \boxed{a_1} \cdot b_1$$
$$acc_0 = acc_0 + \boxed{a_0} \cdot b_0$$




MAC Unit with Multi-Bit-Width and Subword Permutation Support [1]


[1] Beyer, M., Gesper, S., Guntoro, A., Payá-Vayá, G., Blume, H., “Exploiting Subword Permutations to Maximize CNN Compute Performance and Efficiency”, 34th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2023.

# Method
## Background – Multi-Bit-Width MAC Unit

Baseline Multiply-Accumulate (MAC) unit:

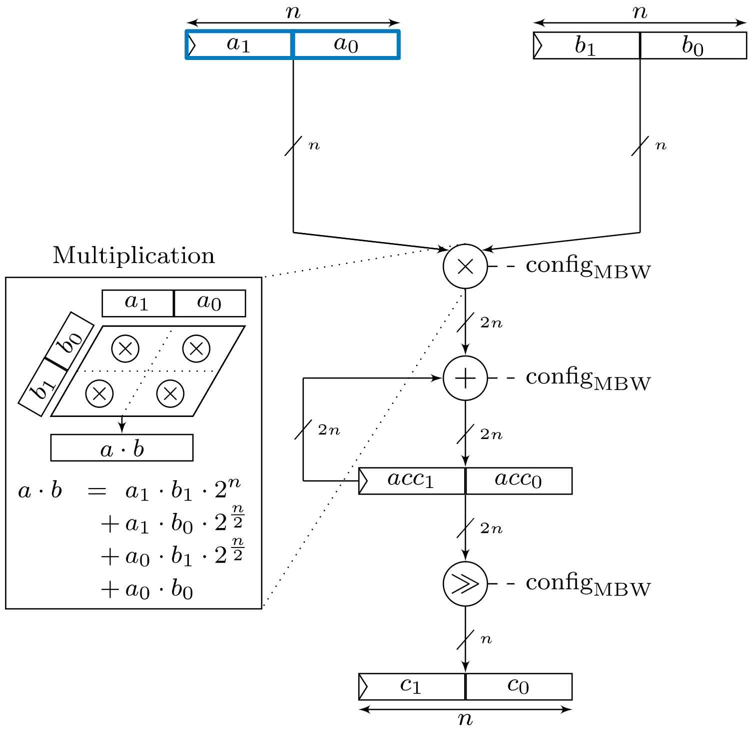
$$acc = acc + a \cdot b$$

With Multi-bit-width support:

$$acc_1 = acc_1 + \boxed{a_1} \cdot b_1$$
$$acc_0 = acc_0 + \boxed{a_0} \cdot b_0$$

Computations requiring the full precision:

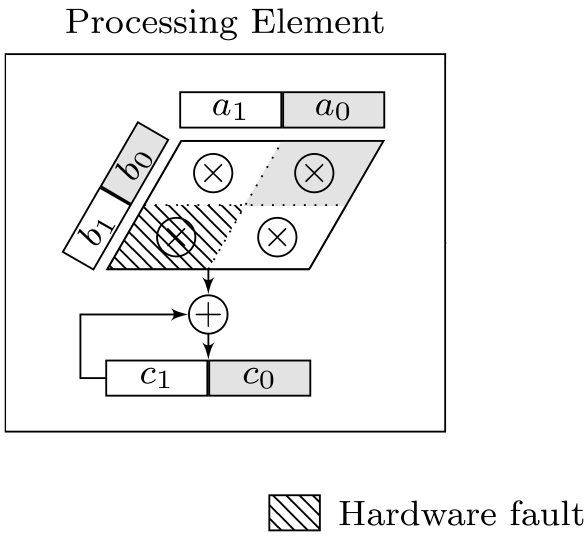
$$a \cdot b = a_1 \cdot b_1 \cdot 2^n + a_1 \cdot b_0 \cdot 2^{\frac{n}{2}} + a_0 \cdot b_1 \cdot 2^{\frac{n}{2}} + a_0 \cdot b_0$$




MAC Unit with Multi-Bit-Width and Subword Permutation Support [1]

[1] Beyer, M., Gesper, S., Guntoro, A., Payá-Vayá, G., Blume, H., “Exploiting Subword Permutations to Maximize CNN Compute Performance and Efficiency”, 34th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2023.

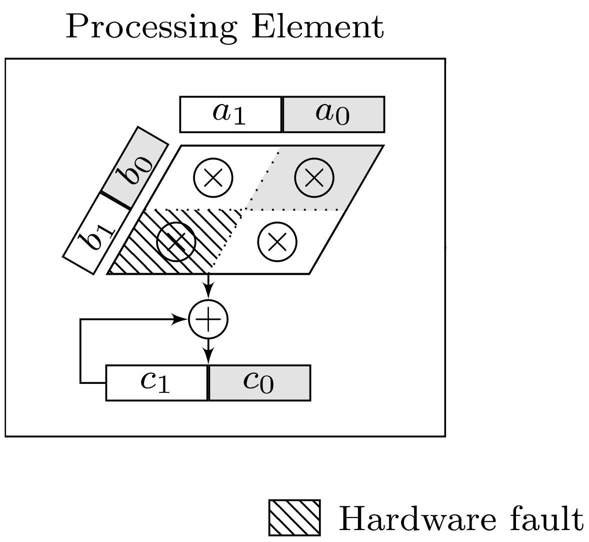
# Method
## Online Quantization Adaptation (OQA)





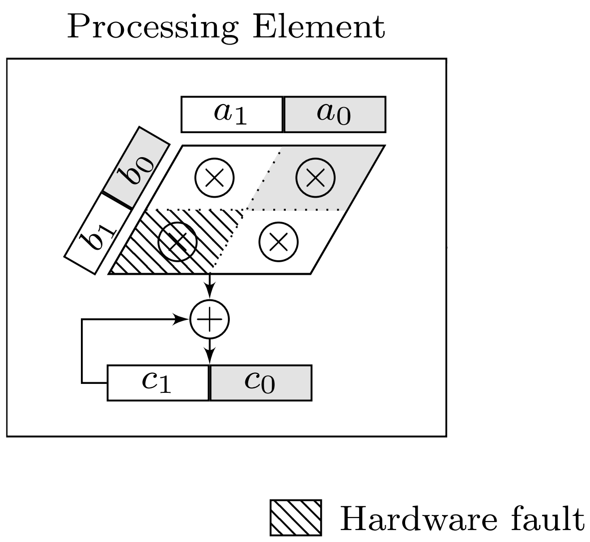
Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Method
## Online Quantization Adaptation (OQA)




- Leverage inherent redundancy for lightweight fault tolerance

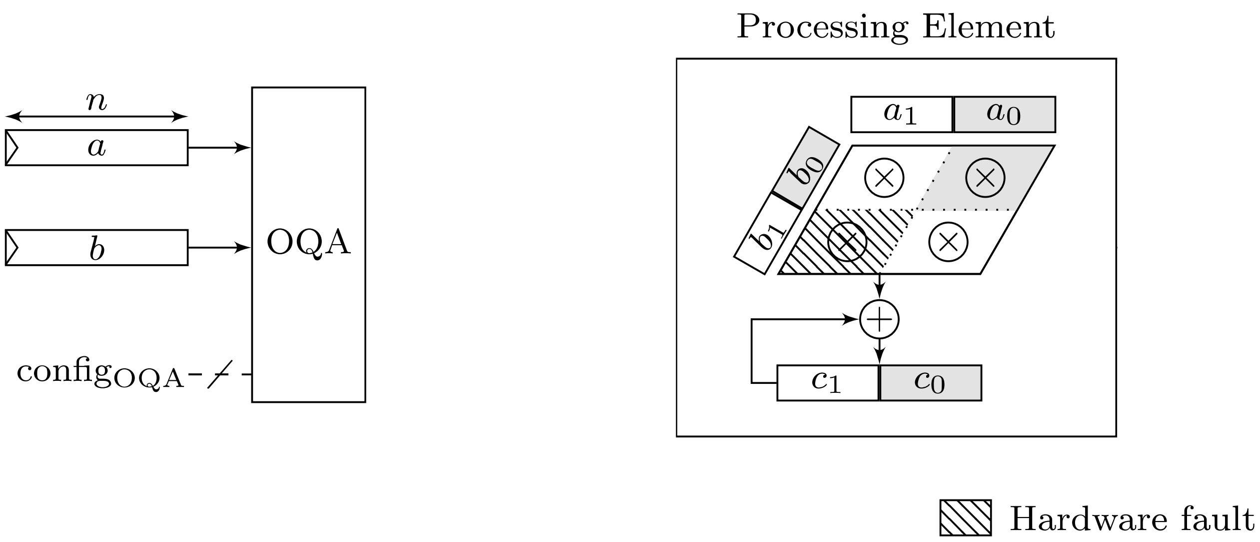
# Method
## Online Quantization Adaptation (OQA)




- Leverage inherent redundancy for lightweight fault tolerance
- Perform computations in fail-degraded operating mode
  → uphold compute capability with reduced precision


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Method
## Online Quantization Adaptation (OQA)

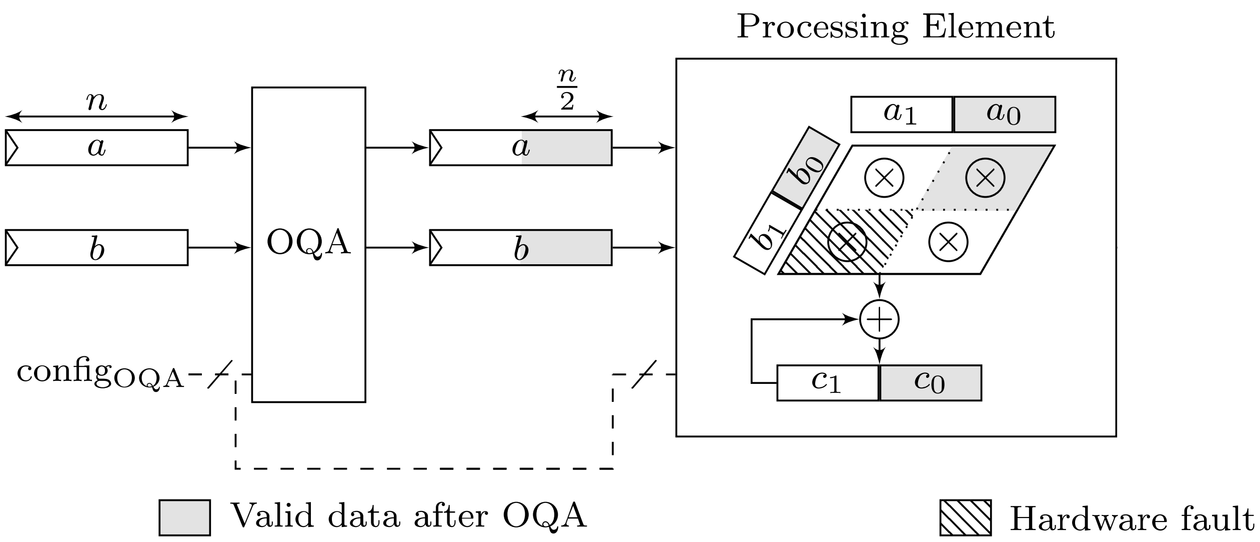



- Leverage inherent redundancy for lightweight fault tolerance
- Perform computations in fail-degraded operating mode
  → uphold compute capability with reduced precision


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

BOSCH

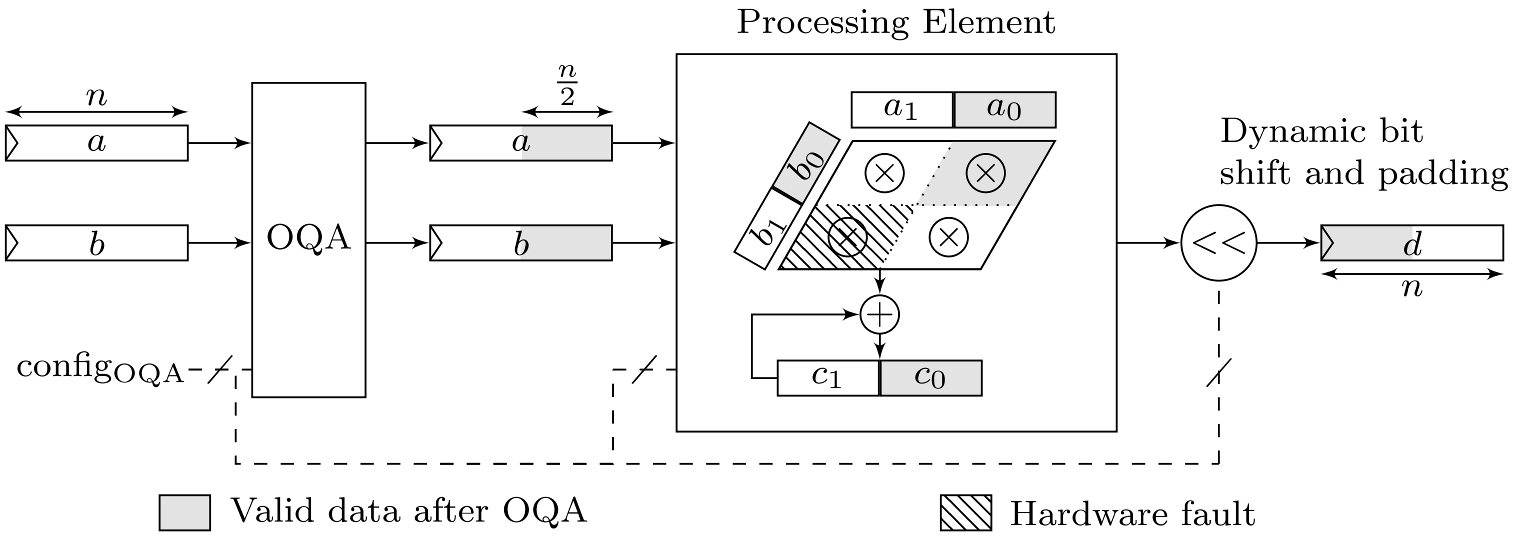
# Method
## Online Quantization Adaptation (OQA)




- Leverage inherent redundancy for lightweight fault tolerance
- Perform computations in fail-degraded operating mode
  → uphold compute capability with reduced precision


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

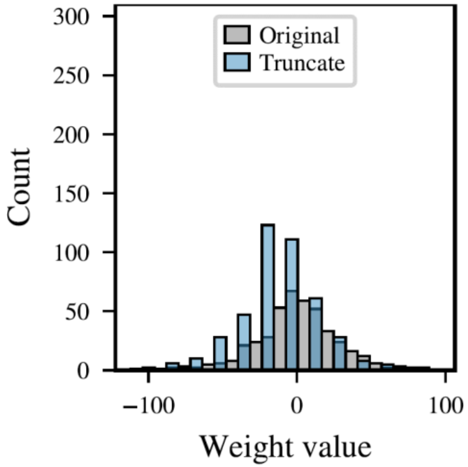
# Method
## Online Quantization Adaptation (OQA)




- Leverage inherent redundancy for lightweight fault tolerance
- Perform computations in fail-degraded operating mode
  → uphold compute capability with reduced precision


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Method
## Rounding Modes

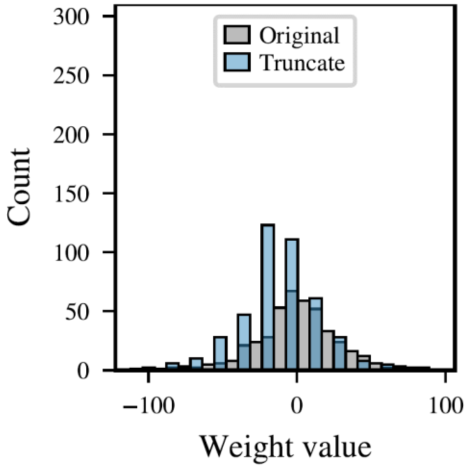


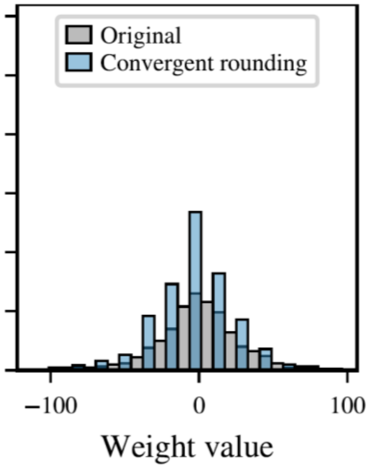

### Truncate

- Straightforward solution, simple to implement
- Adds bias and results in quantization error with non-zero mean


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Method
## Rounding Modes



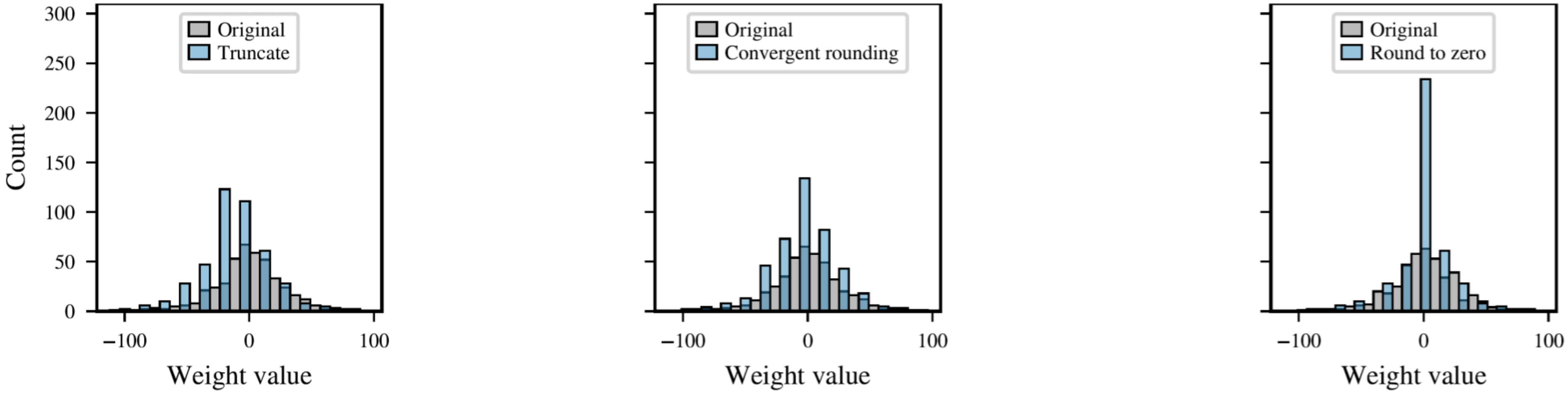

### Truncate

- Straightforward solution, simple to implement
- Adds bias and results in quantization error with non-zero mean

### Convergent Rounding

- Round ties to even
- Overestimation of values

# Method
## Rounding Modes




### Truncate

- Straightforward solution, simple to implement
- Adds bias and results in quantization error with non-zero mean

### Convergent Rounding

- Round ties to even
- Overestimation of values

### Round to Zero

- Preserve overall distribution of weights
- Attenuation rather than overestimation of values


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

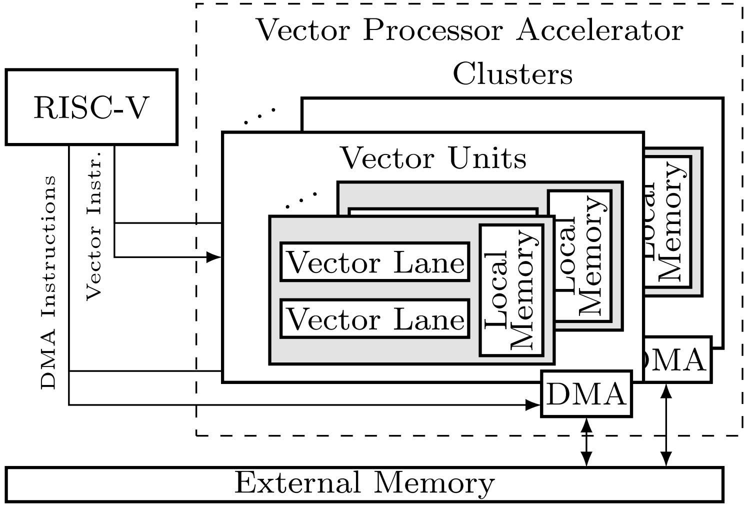
03

# Experiments

# Experiments

## Experimental Setup – Hardware Architecture

- Scalable vector processor as HW target



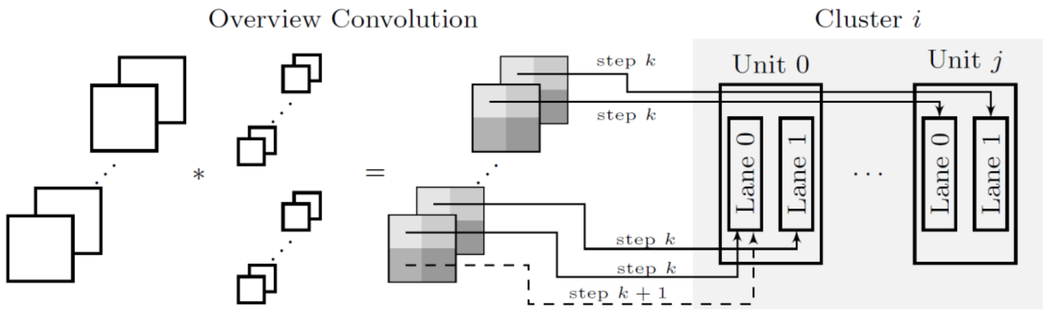

Schematic of the V²PRO Accelerator System [2]


[2] G. B. Thieu et al., "ZuSE-KI-AVF: Application-Specific AI Processor for Intelligent Sensor Signal Processing in Autonomous Driving," in 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2023.

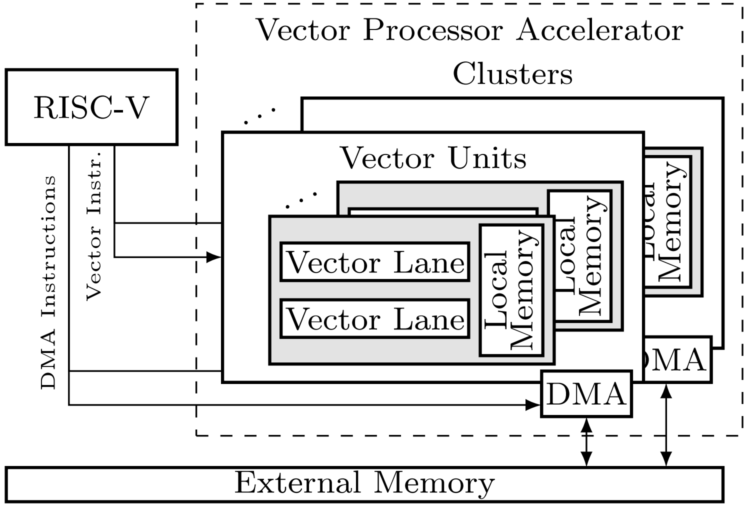
# Experiments

## Experimental Setup – Hardware Architecture

- Scalable vector processor as HW target




Mapping of convolutions on the V²PRO Accelerator System [2]

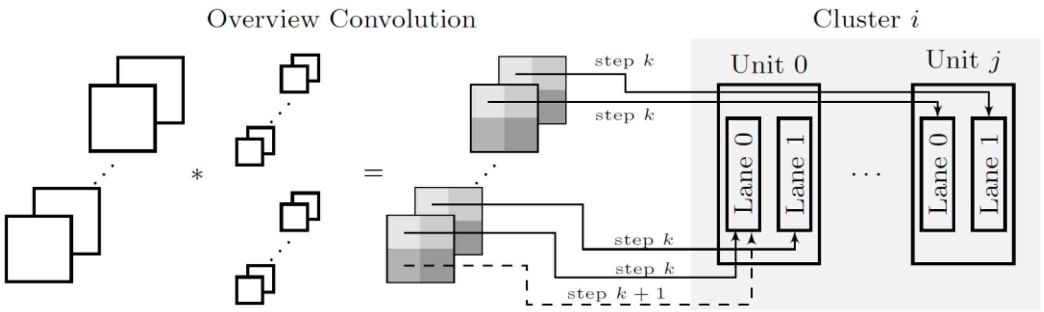



Schematic of the V²PRO Accelerator System [2]

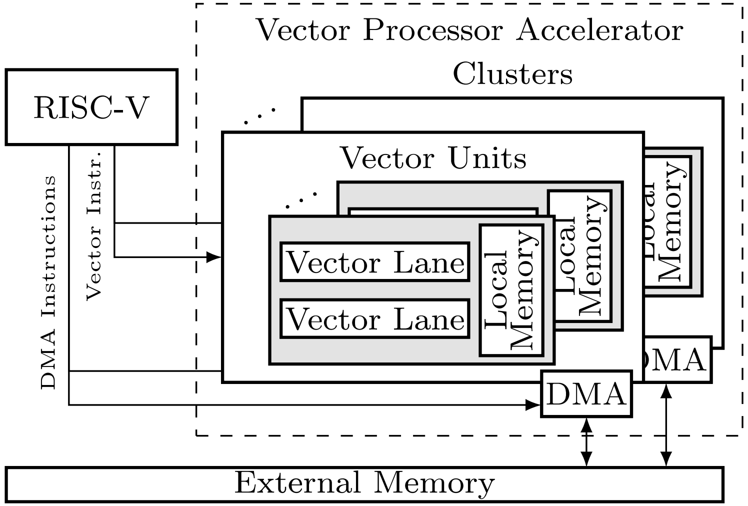
[2] G. B. Thieu et al., "ZuSE-KI-AVF: Application-Specific AI Processor for Intelligent Sensor Signal Processing in Autonomous Driving," in 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2023.

# Experiments
## Experimental Setup – Hardware Architecture

- Scalable vector processor as HW target
- Evaluate different HW configurations




Mapping of convolutions on the V²PRO Accelerator System [2]




Schematic of the V²PRO Accelerator System [2]


[2] G. B. Thieu et al., "ZuSE-KI-AVF: Application-Specific AI Processor for Intelligent Sensor Signal Processing in Autonomous Driving," in 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2023.

# Experiments
## Experimental Setup – Neural Networks

- ResNet18 [3] & VGG16 [4] (quantized to 8-bit)
  - CIFAR-10 [5] and GTSRB [6]


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments

## Experimental Setup – Neural Networks

- ResNet18 [3] & VGG16 [4] (quantized to 8-bit)
  - CIFAR-10 [5] and GTSRB [6]
- Simulate permanent errors & evaluate NN prediction accuracy for different error rates ($n = 200$)


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

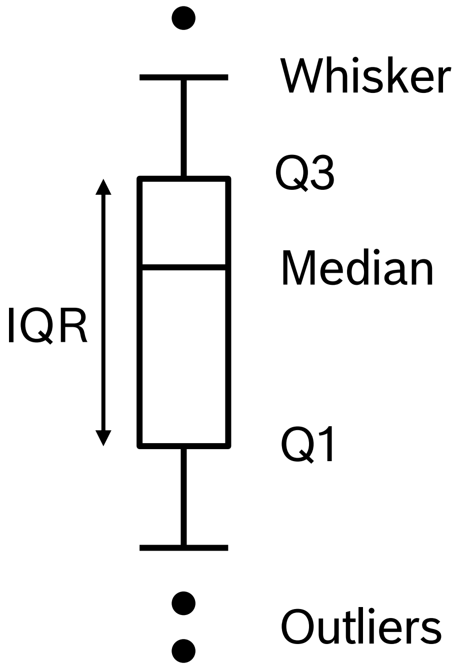
# Experiments
## Experimental Setup – Neural Networks

- ResNet18 [3] & VGG16 [4] (quantized to 8-bit)
  - CIFAR-10 [5] and GTSRB [6]
- Simulate permanent errors & evaluate NN prediction accuracy for different error rates ($n = 200$)
- Two error mitigation mechanisms for computations on faulty PEs:
  1. OQA: Values are re-quantized online, computations are performed with reduced precision
  2. Discard: Values are discarded and set to zero


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Experimental Setup – Neural Networks

- ResNet18 [3] & VGG16 [4] (quantized to 8-bit)
  - CIFAR-10 [5] and GTSRB [6]
- Simulate permanent errors & evaluate NN prediction accuracy for different error rates ($n = 200$)
- Two error mitigation mechanisms for computations on faulty PEs:
  1. OQA: Values are re-quantized online, computations are performed with reduced precision
  2. Discard: Values are discarded and set to zero

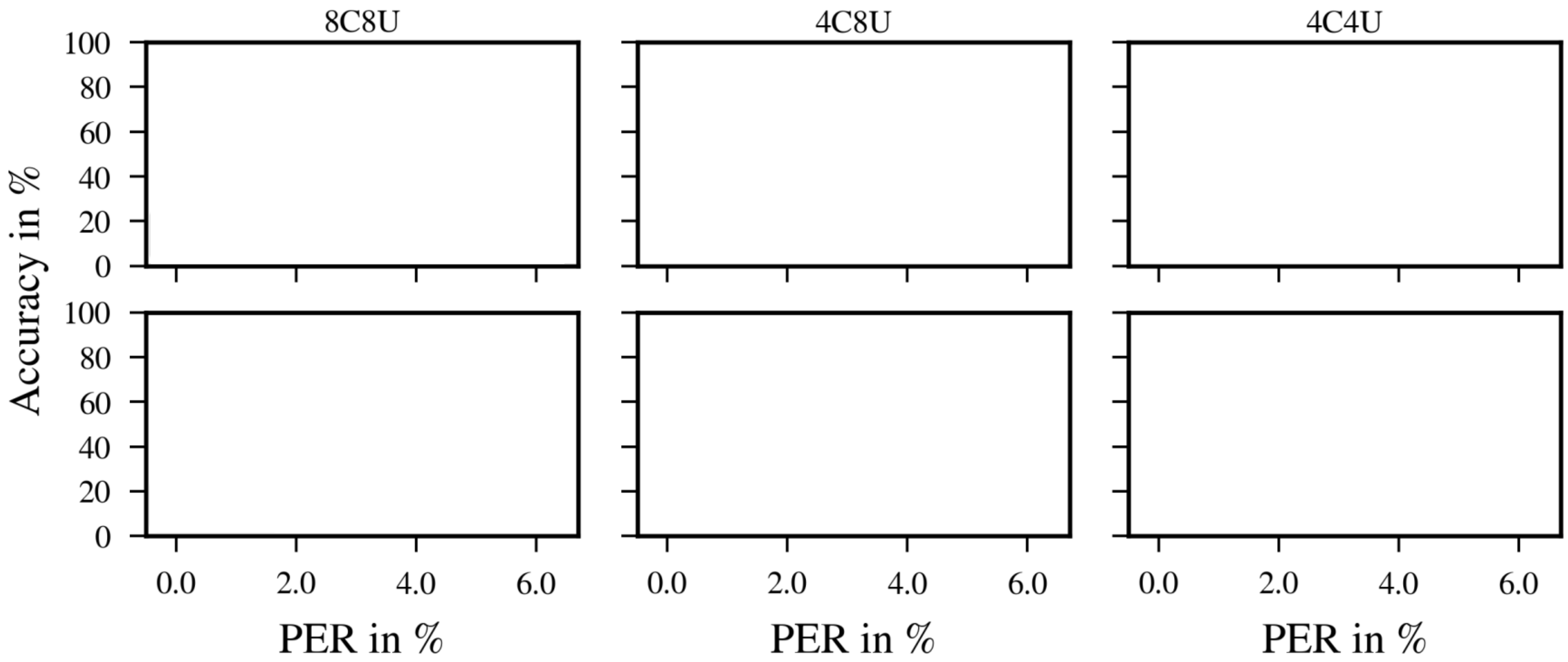



Schematic of a box plot. Higher median and lower variability is better.


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)

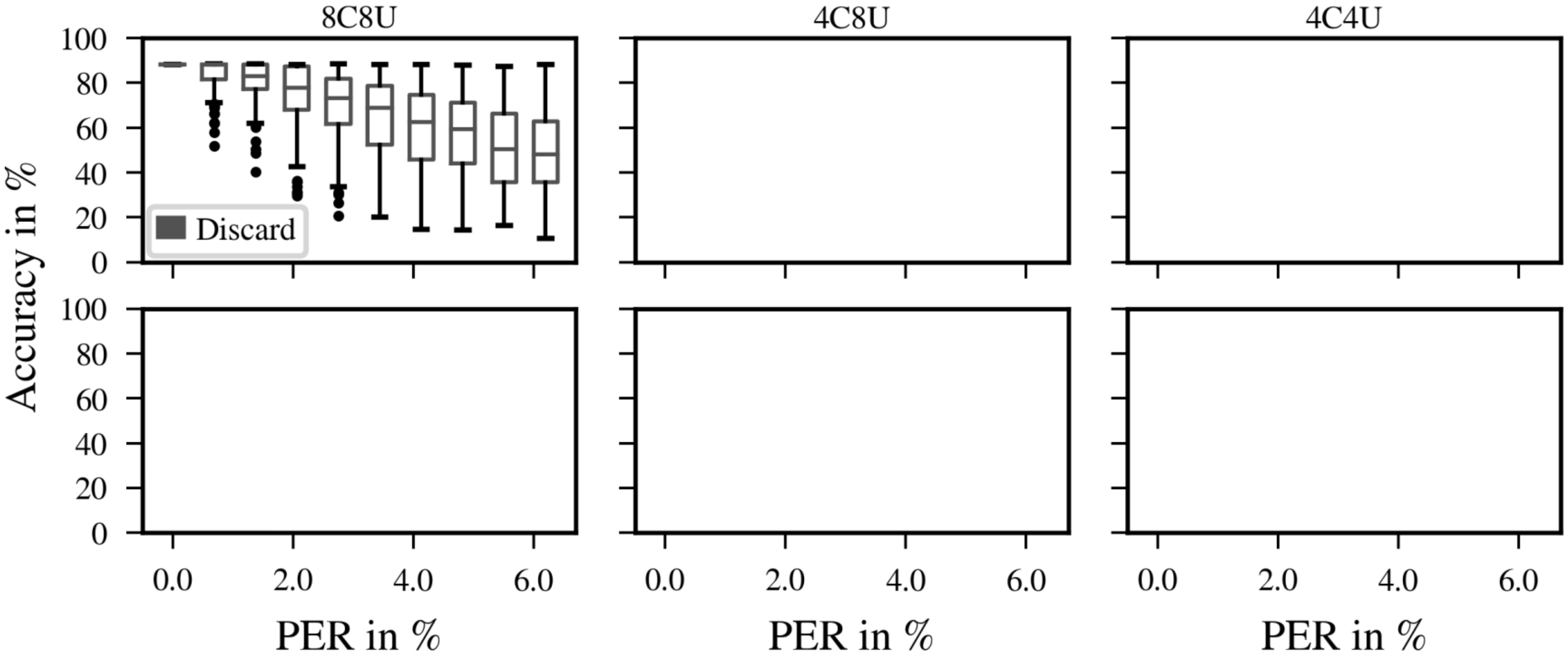



Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)

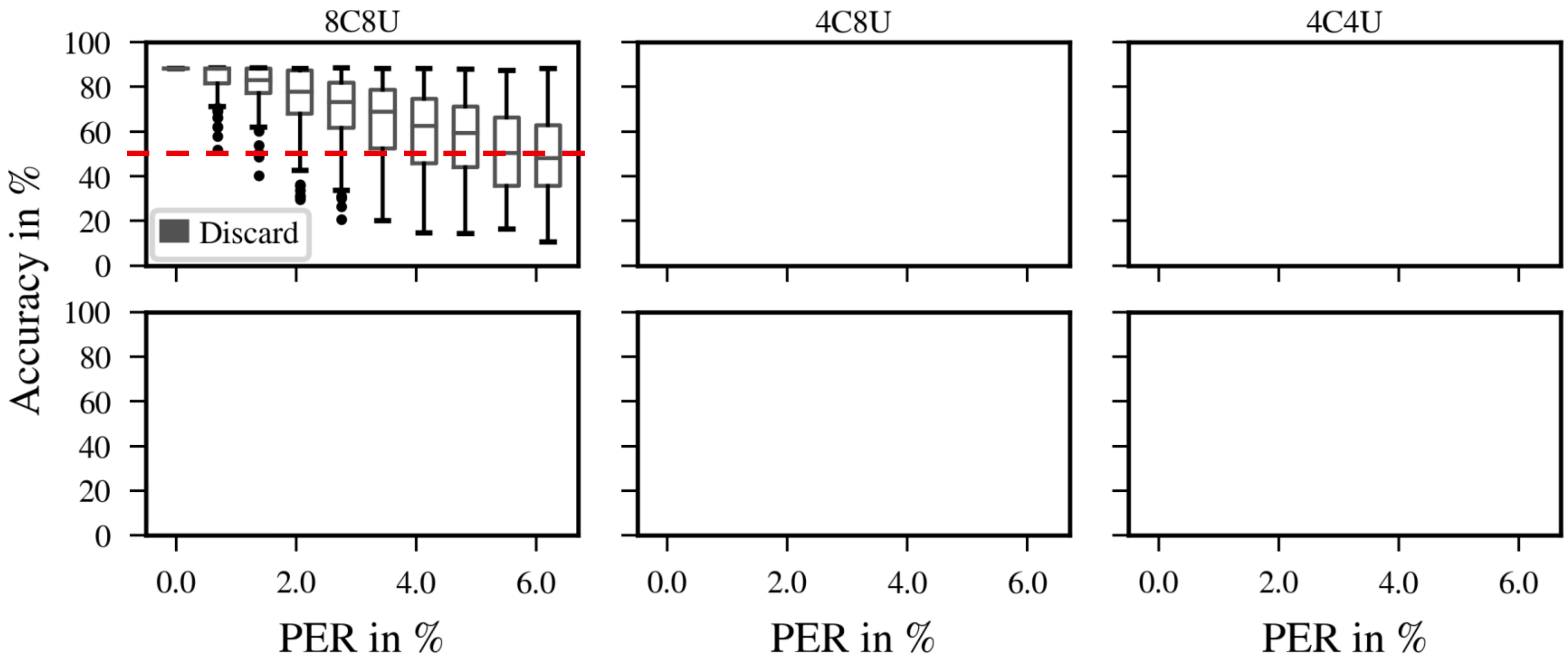



Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)

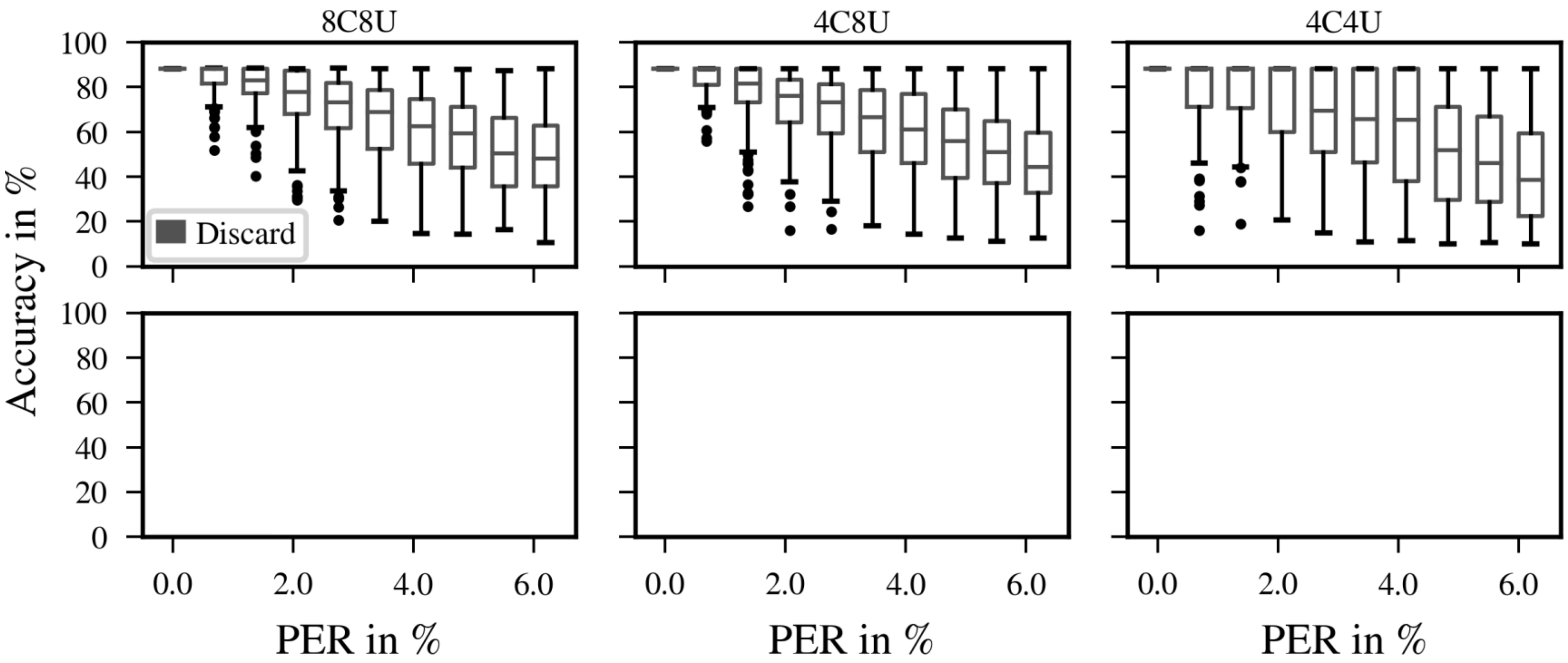



Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)

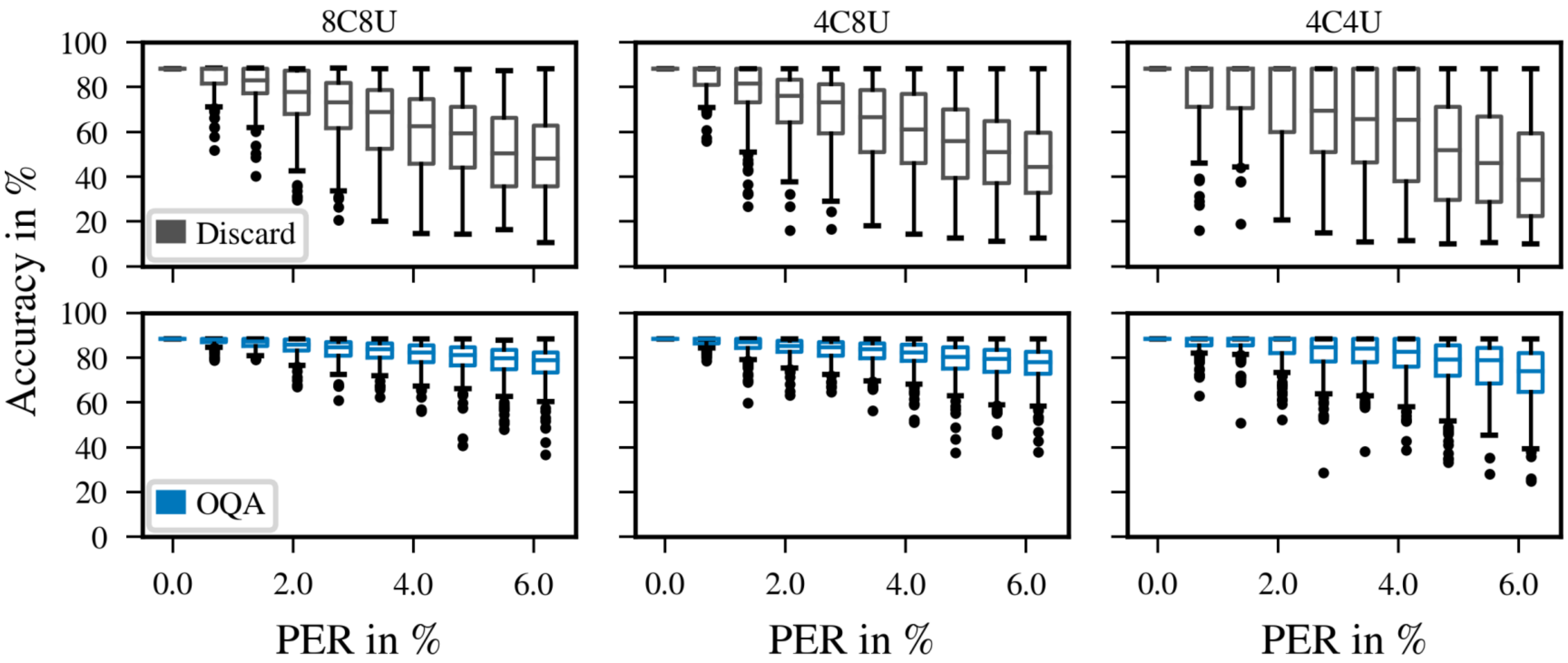



Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)

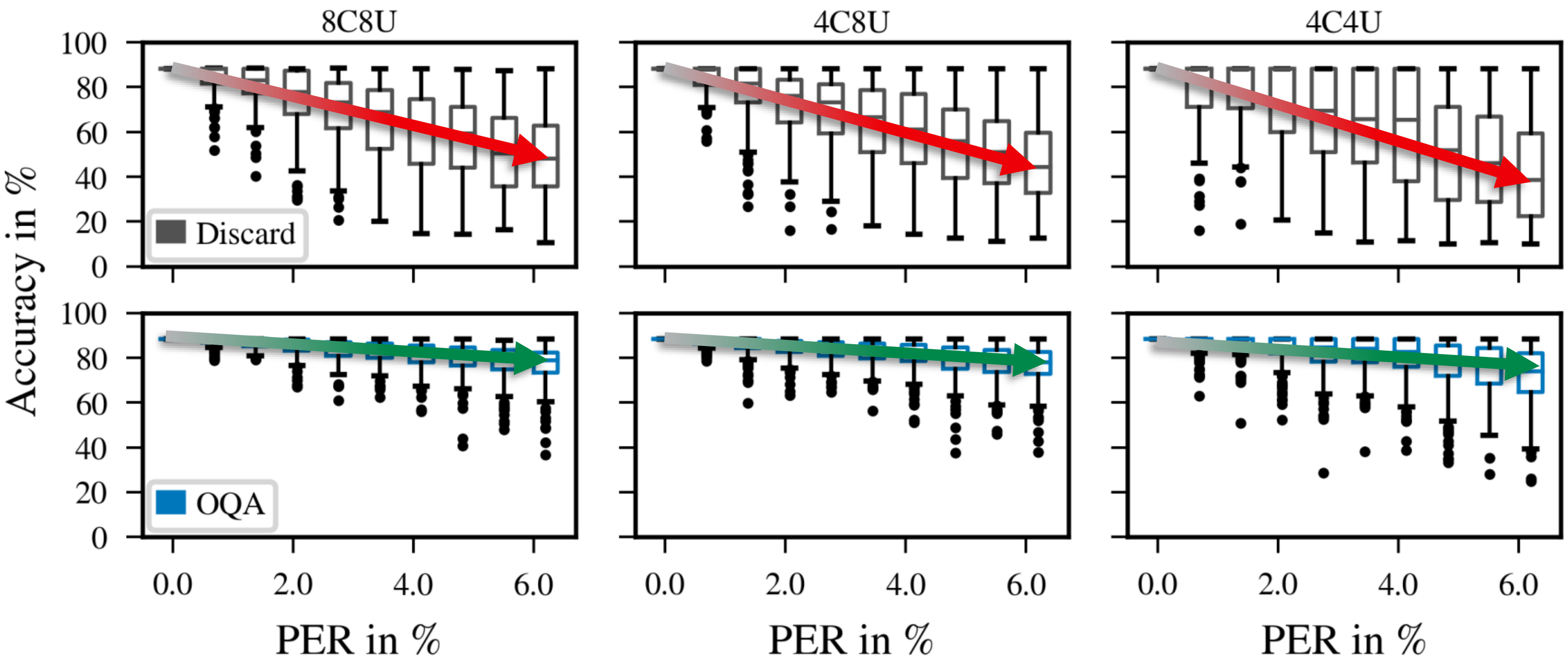



Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)

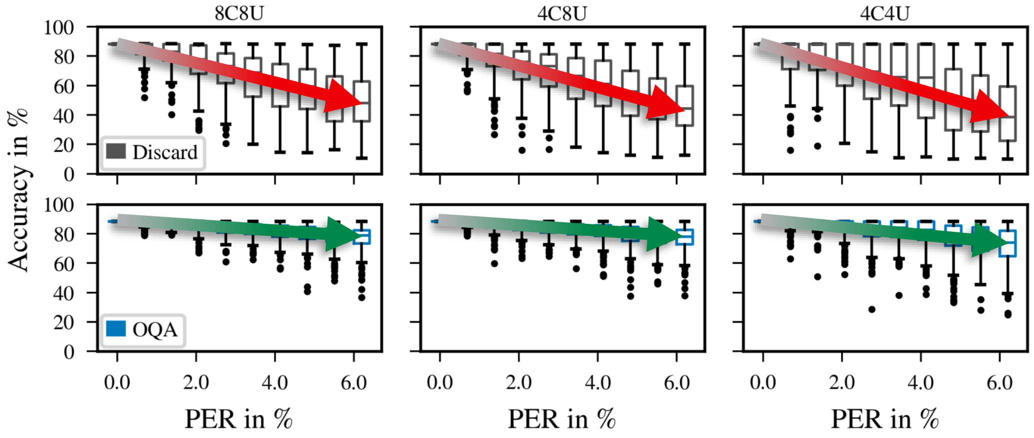



Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

**ResNet18 (CIFAR-10)**




**VGG16 (CIFAR-10)**

Higher median and lower variability is better

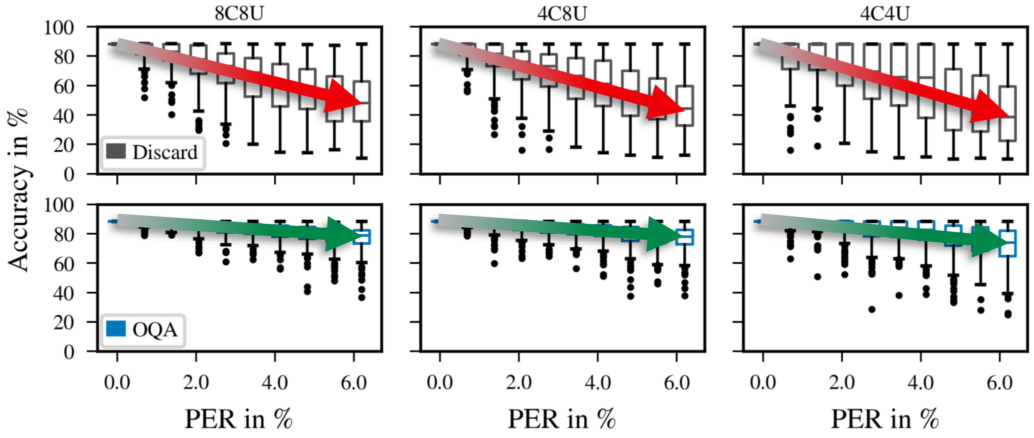




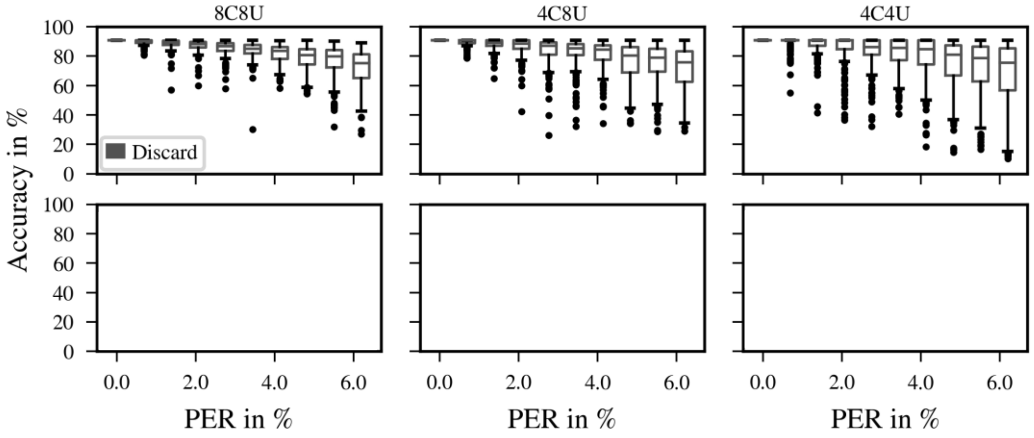
BOSCH

# Experiments
## Results

### ResNet18 (CIFAR-10)




### VGG16 (CIFAR-10)

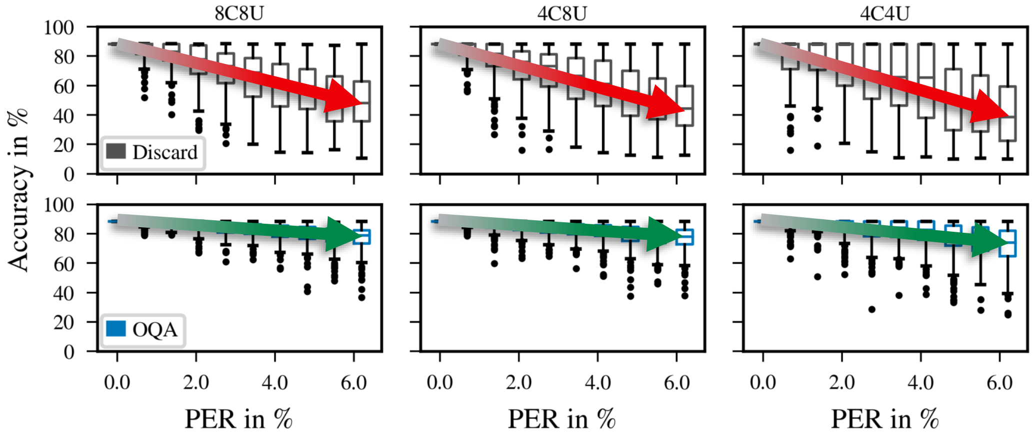



Higher median and lower variability is better

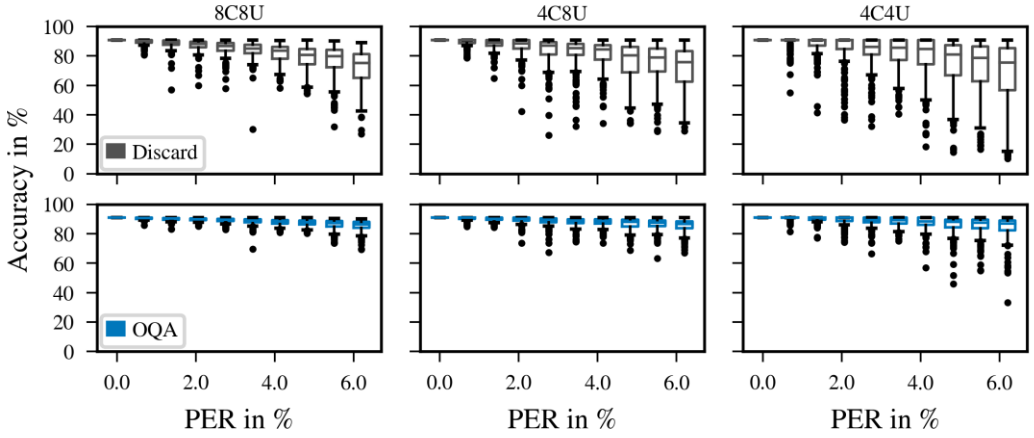

Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)




### VGG16 (CIFAR-10)

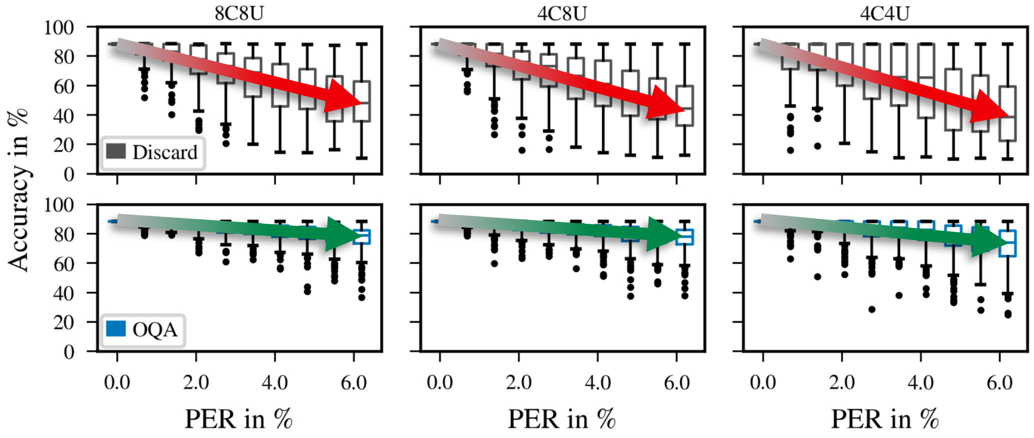



Higher median and lower variability is better

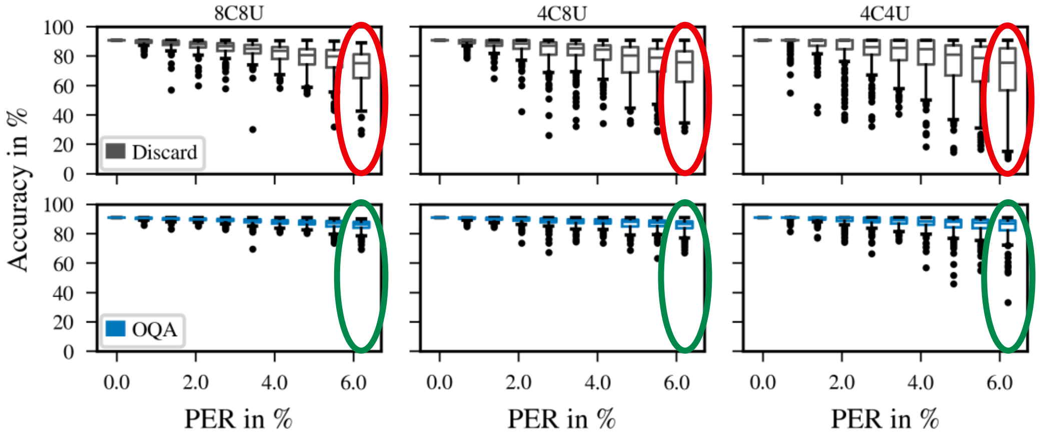

Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18 (CIFAR-10)




### VGG16 (CIFAR-10)




Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

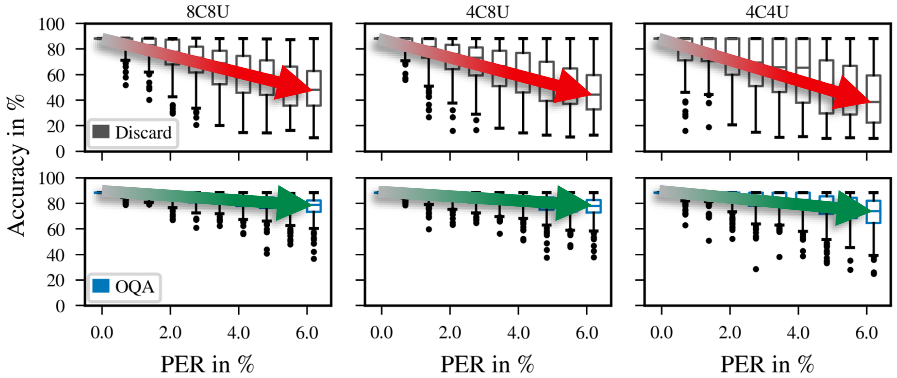

BOSCH

# Experiments
## Results

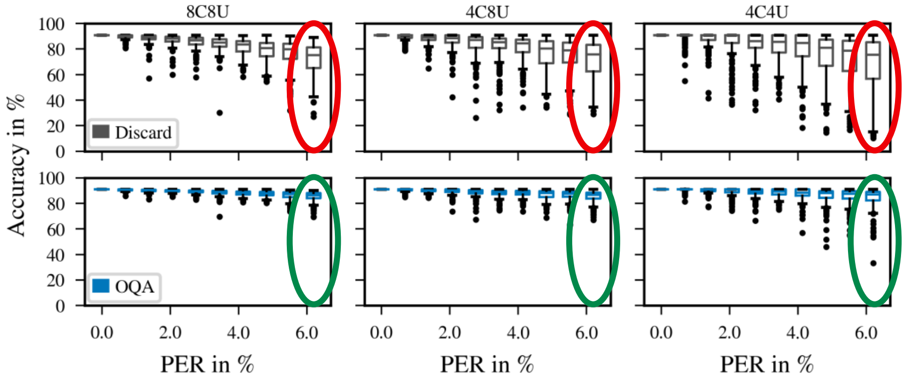
CFAR-10

GTSRB

### ResNet18




### VGG16

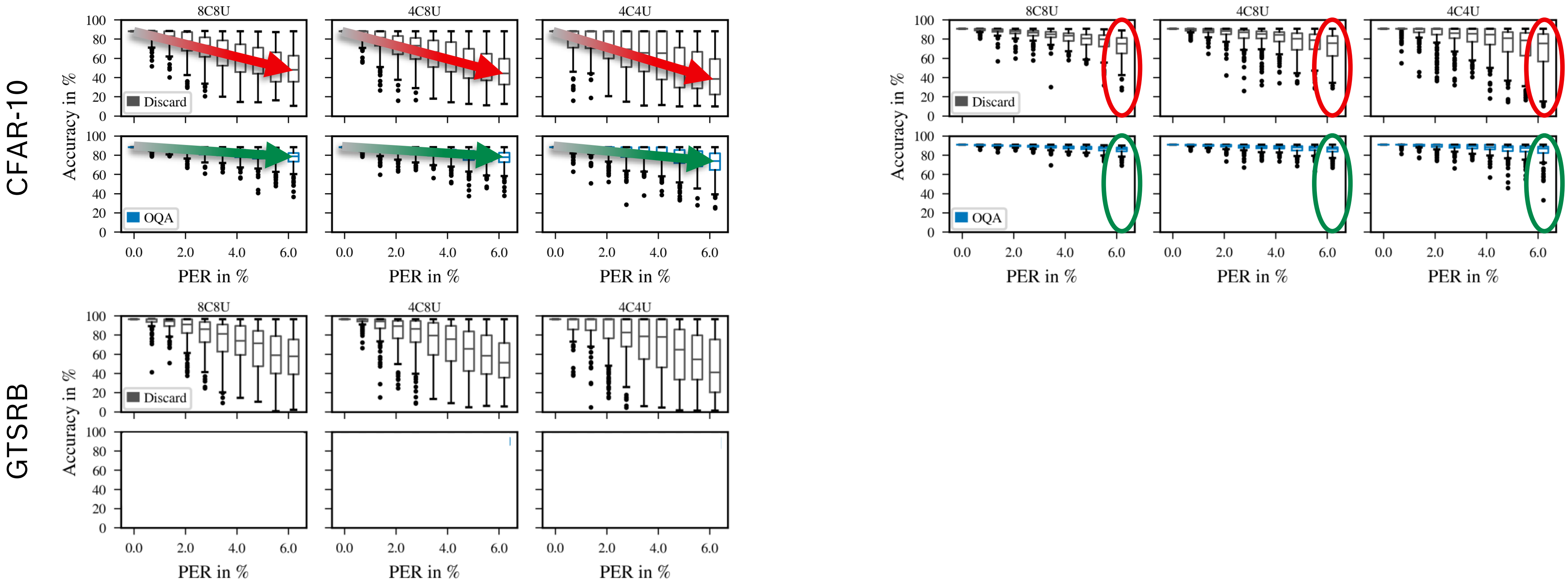



Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

**ResNet18** **VGG16**

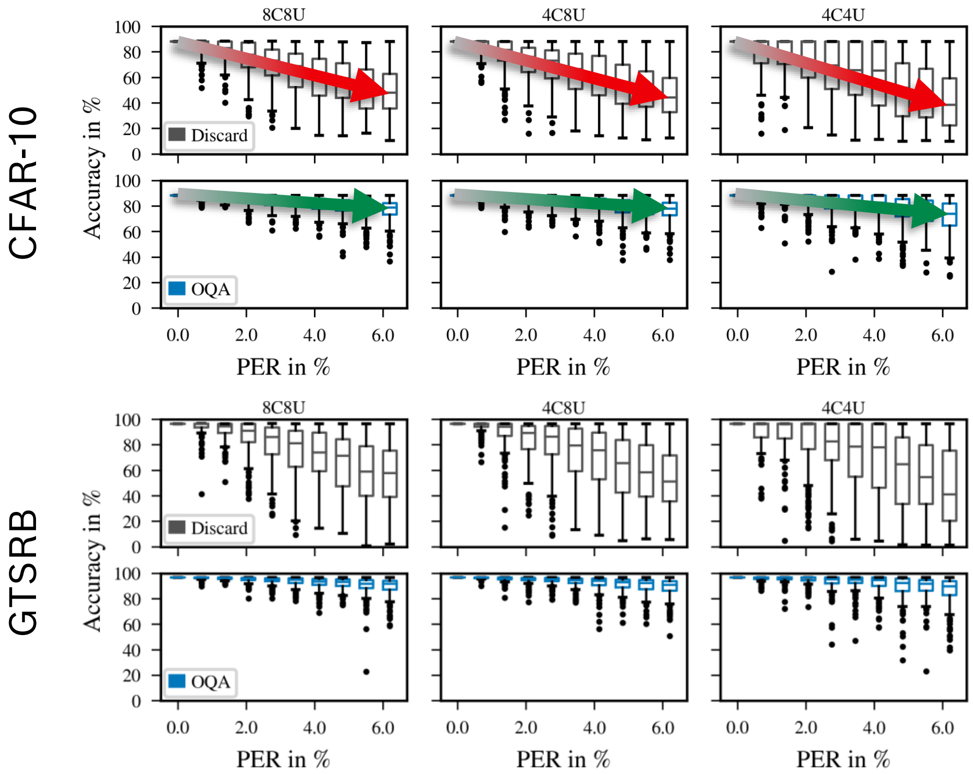



Higher median and lower variability is better

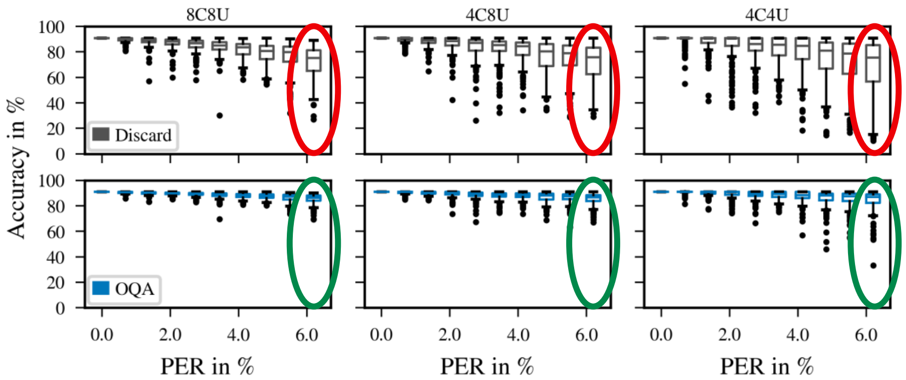

Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results

### ResNet18



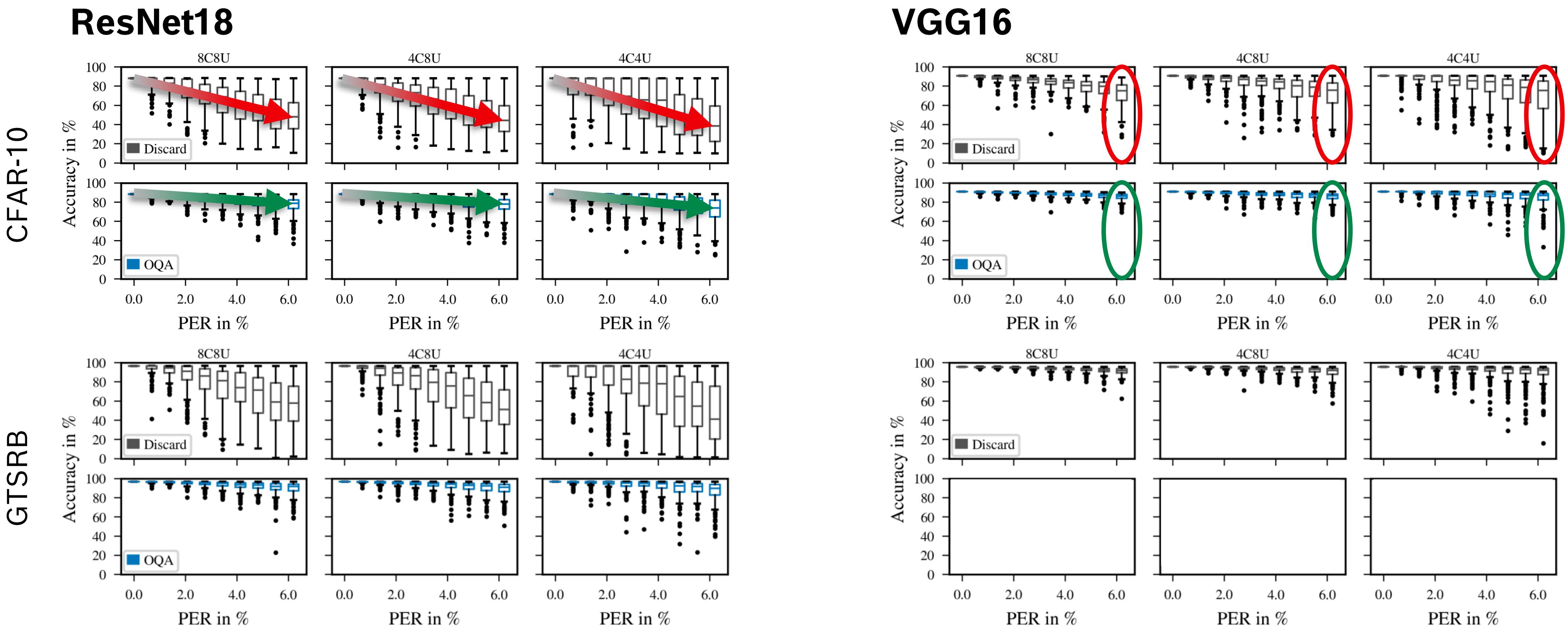

### VGG16




Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

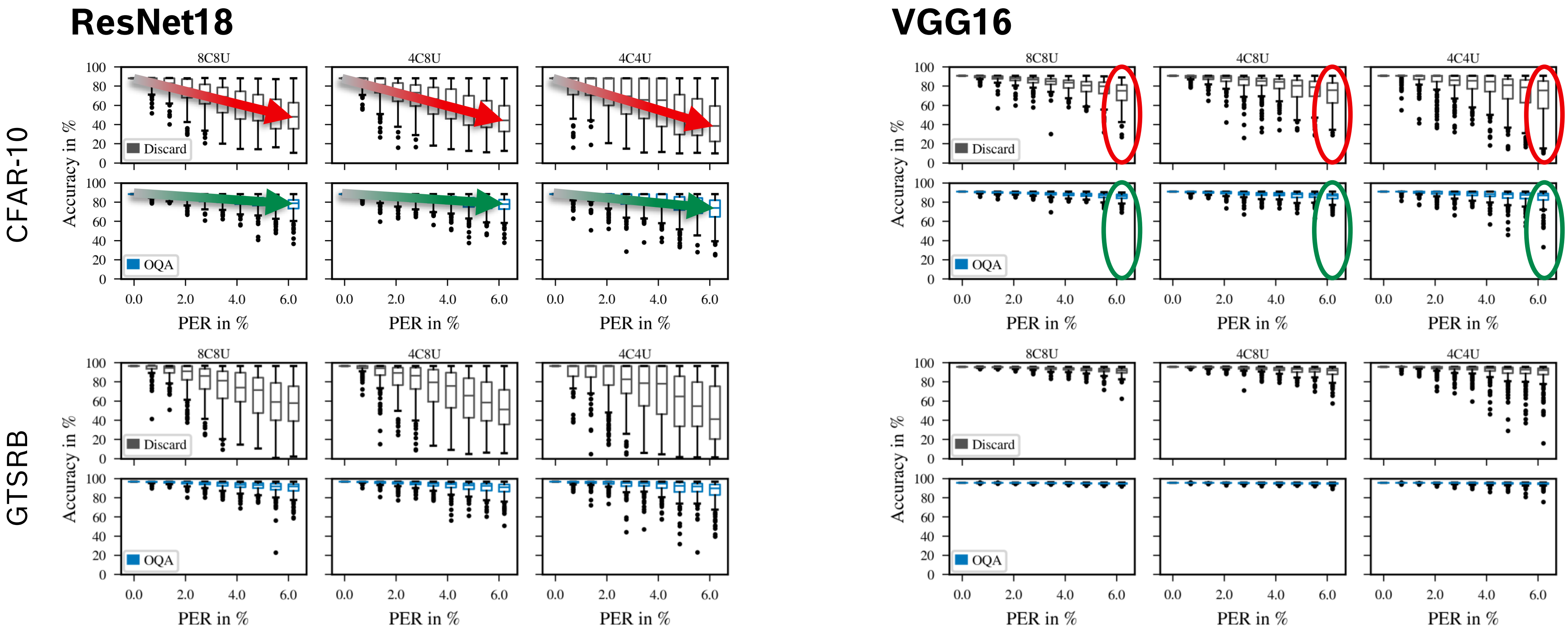
# Experiments
## Results




Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Experiments
## Results




Higher median and lower variability is better


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

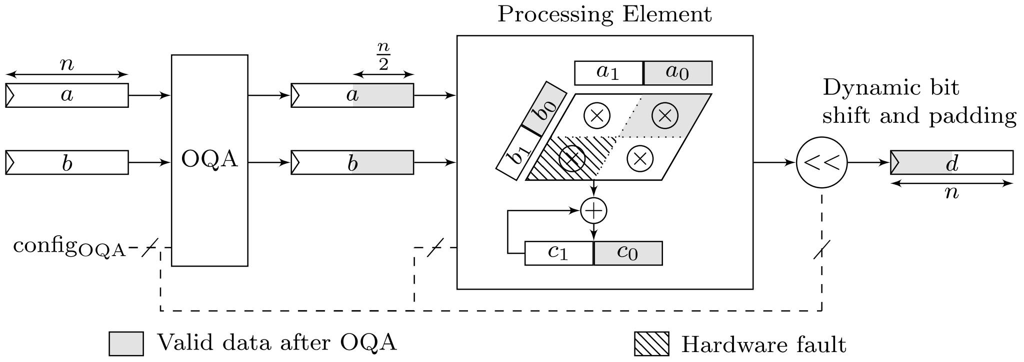
# 04

# Conclusion

BOSCH

# Conclusion
## Online Quantization Adaptation (OQA)

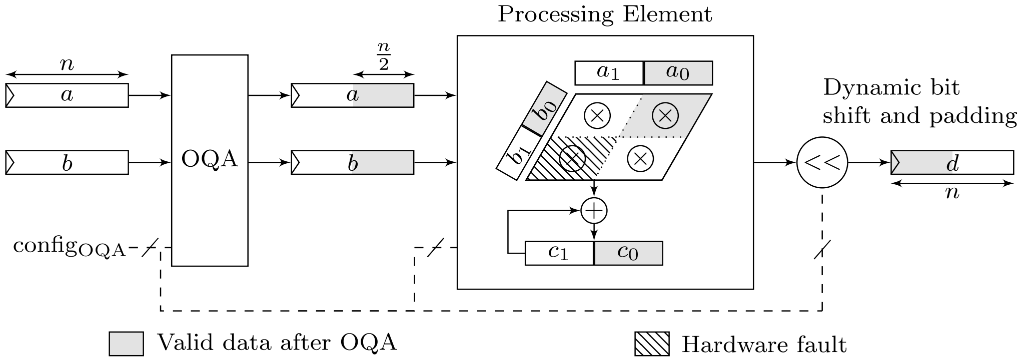
- OQA can preserve a NN's classification performance consistently





Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

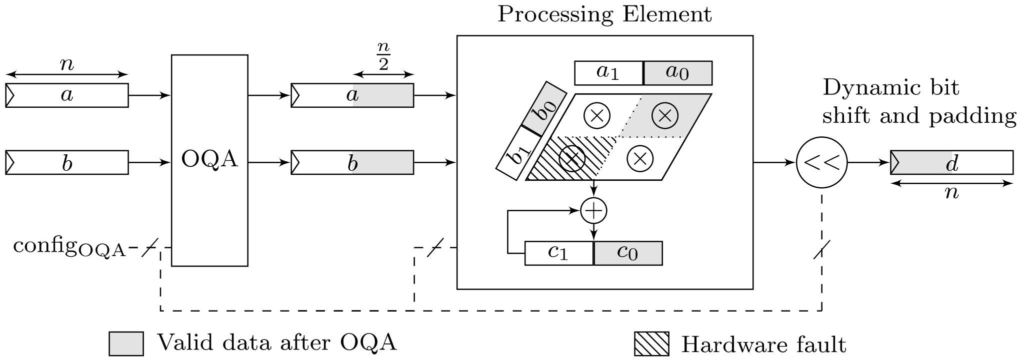
# Conclusion
## Online Quantization Adaptation (OQA)




- OQA can preserve a NN's classification performance consistently
- Low variability of NN prediction performance
  - Higher confidence in predictions made by NN executed on faulty HW


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

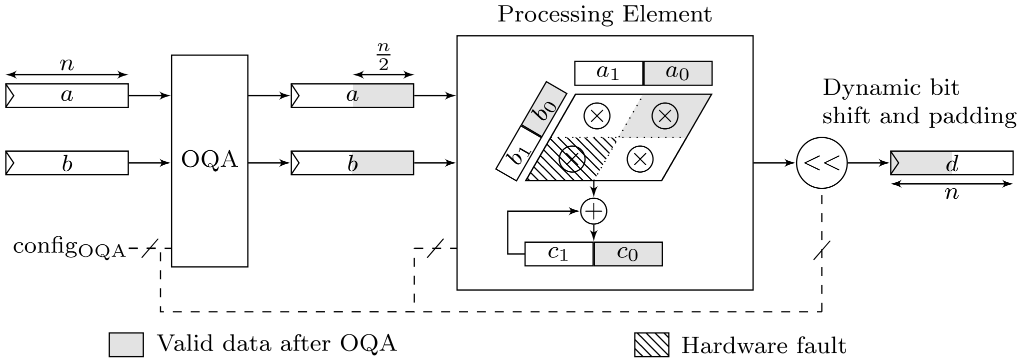
# Conclusion
## Online Quantization Adaptation (OQA)




- OQA can preserve a NN's classification performance consistently
- Low variability of NN prediction performance
  - Higher confidence in predictions made by NN executed on faulty HW
- NNs retain at least original error-free accuracy when considering top-2 predictions

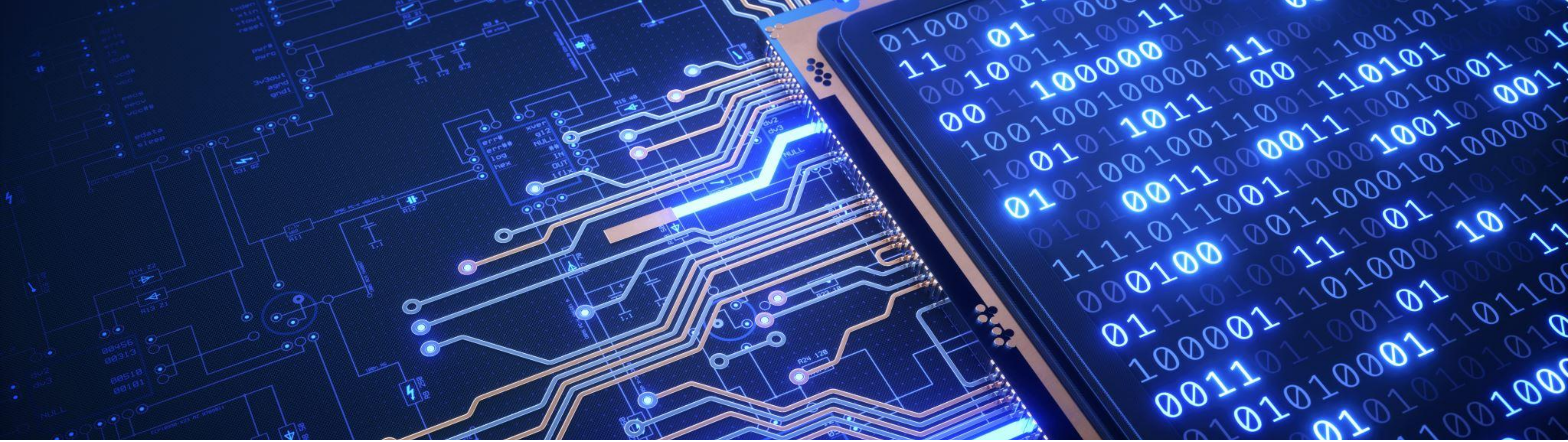

Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security

# Conclusion
## Online Quantization Adaptation (OQA)




- OQA can preserve a NN's classification performance consistently
- Low variability of NN prediction performance
  - Higher confidence in predictions made by NN executed on faulty HW
- NNs retain at least original error-free accuracy when considering top-2 predictions
- Lightweight solution through dual-use of existing HW


Michael Beyer | michael.beyer2@de.bosch.com | 2023-09-22 | 42nd International Conference on Computer Safety, Reliability and Security



# Thank you

michael.beyer2@de.bosch.com

# References


[1] M. Beyer, S. Gesper, A. Guntoro, G. Payá-Vayá, H. Blume, "Exploiting Subword Permutations to Maximize CNN Compute Performance and Efficiency", 34th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2023.

[2] G. B. Thieu et al., "ZuSE-KI-AVF: Application-Specific AI Processor for Intelligent Sensor Signal Processing in Autonomous Driving," in 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2023.

[3] K. He, et al., "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[5] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Technical report (2009).

[6] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, "Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition," Neural Netw. 32, 323–332 (2012).