

# Micron Deep Learning Accelerator, tools and applications

March 9th 2021

©2020 Micron Technology, Inc. All rights reserved. Information, products, and/or specifications are subject to change without notice. All information is provided on an “AS IS” basis without warranties of any kind. Statements regarding products, including statements regarding product features, availability, functionality, or compatibility, are provided for informational purposes only and do not modify the warranty, if any, applicable to any product. Drawings may not be to scale. Micron, the Micron orbit logo, the M orbit logo, Intelligence Accelerated™, and other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners.

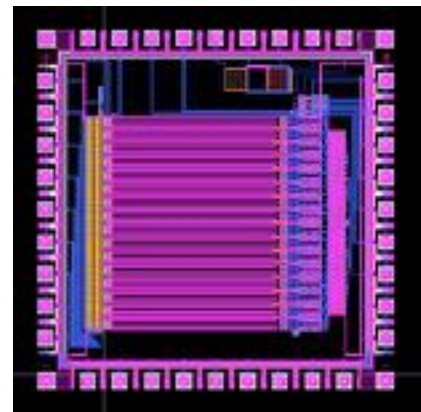


# Micron DLA

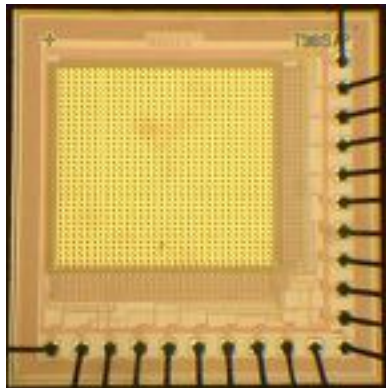
## an introduction



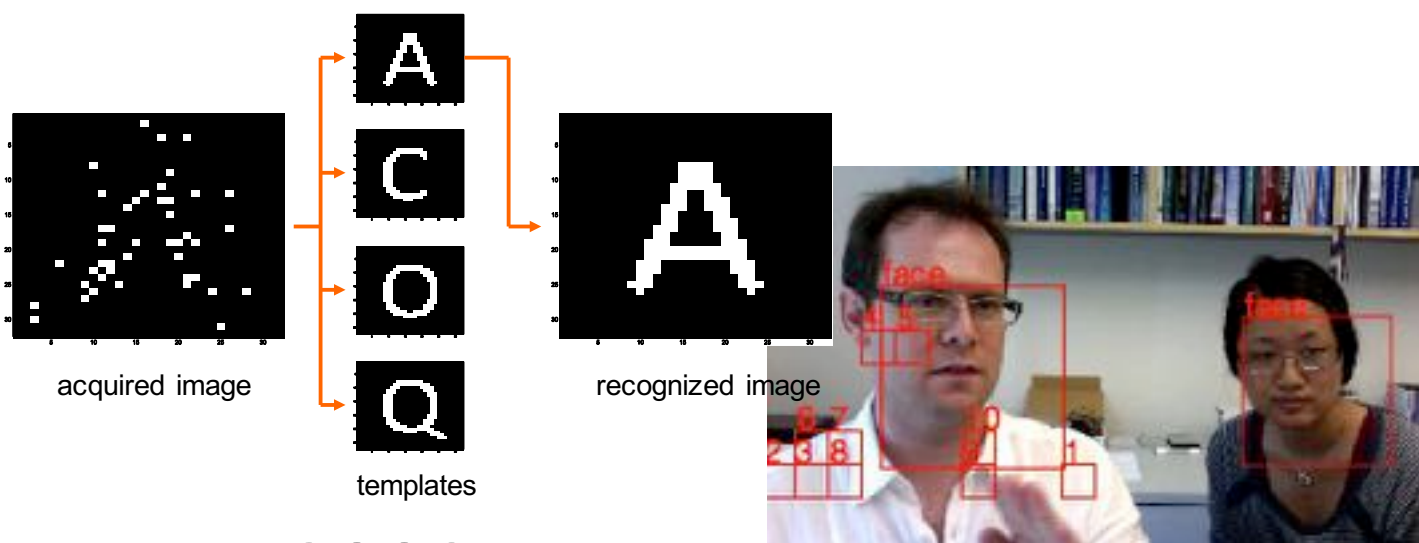
# our experience



**1998**  
neural transceiver



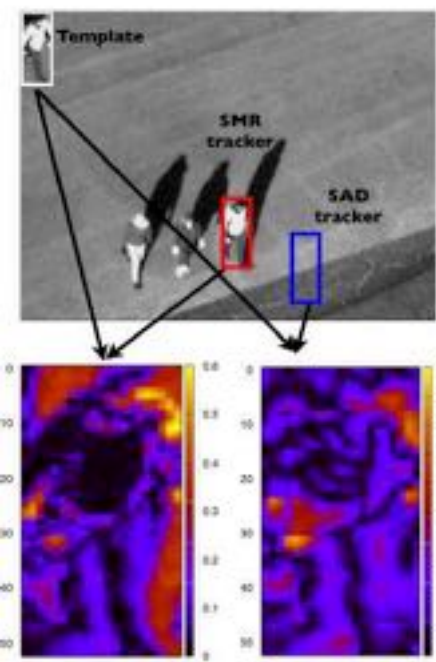
**2004**  
ALOHA  
neural imager



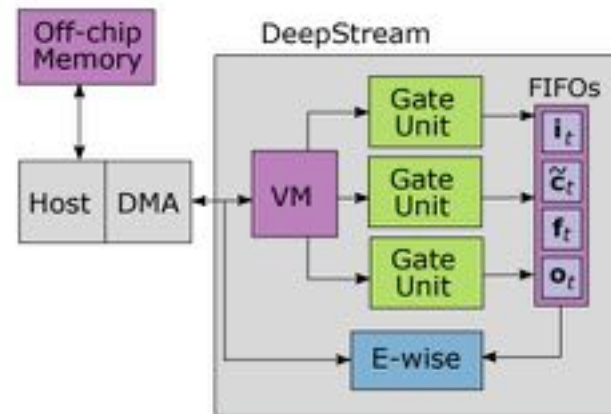
**2008**  
spike-based  
object recognition



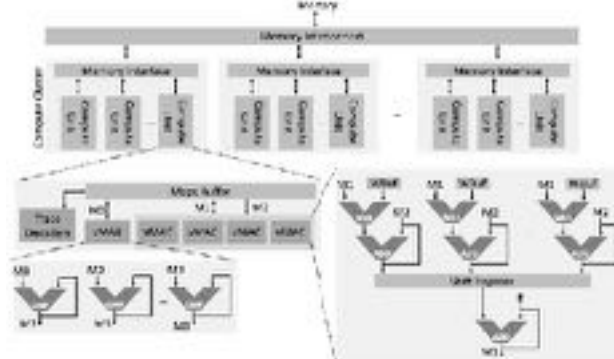
**2010**  
visual attention



**2012**  
neural tracking



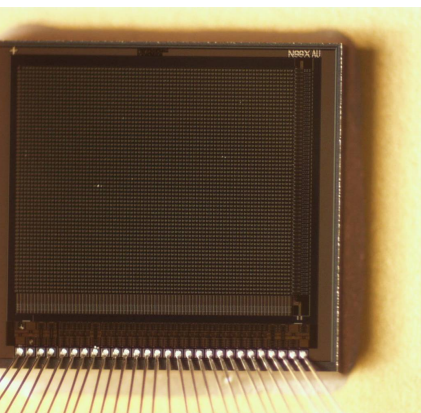
**2014**  
RNN  
accelerator



**2017**  
snowflake  
neural processor

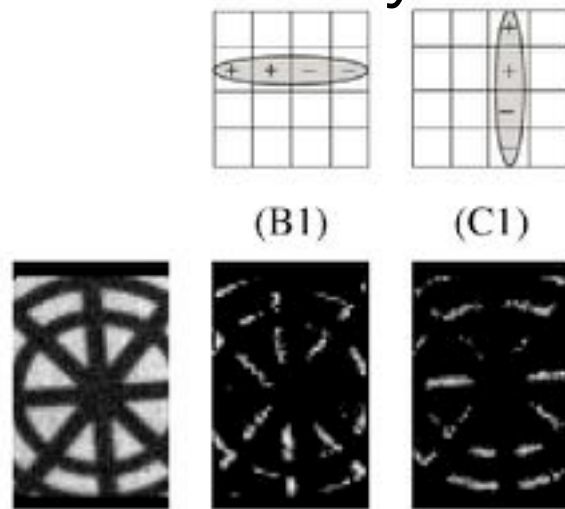


**2018**  
neural compiler

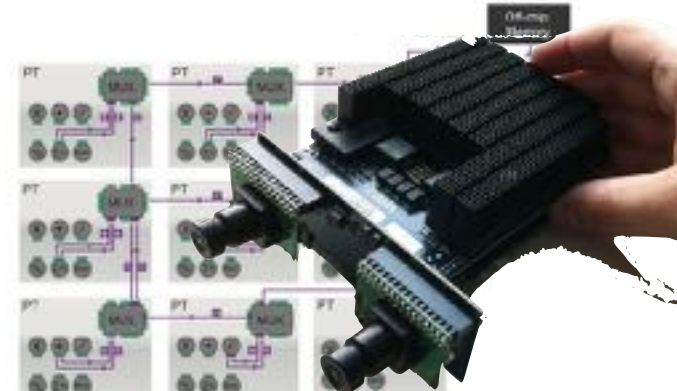


**2000**  
octopus  
retina

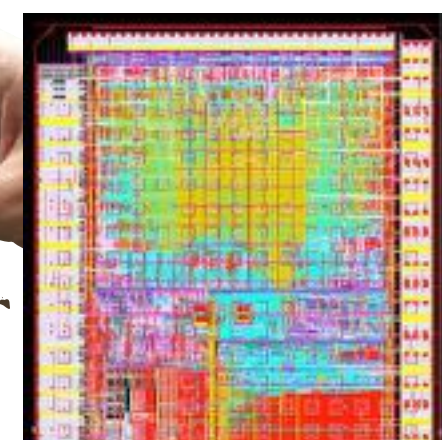
**2006**  
IFAT  
neural array



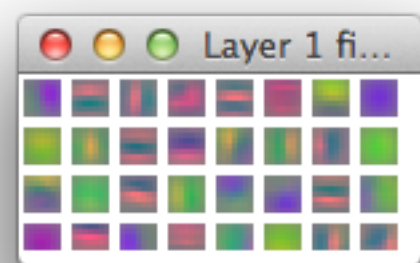
**2009**  
neuFlow  
neural  
processor



**2011**  
neuFlow SoC



**2012**  
unsupervised  
clustering  
learning



**2013**  
nn-X  
neural processor



**2016**  
e-Net

**2017**  
Linknet



**2017**  
CortexNet

**2018**  
super  
resolution

**2018**  
Predictive  
coding  
network





# Micron Deep Learning Accelerator (DLA)

*Designed for:*

- Good performance per power
- High utilization
- Efficient use of memory bandwidth
- Low latency
- Scalability: IoT to cloud

*Implemented on Micron ACS FPGAs*



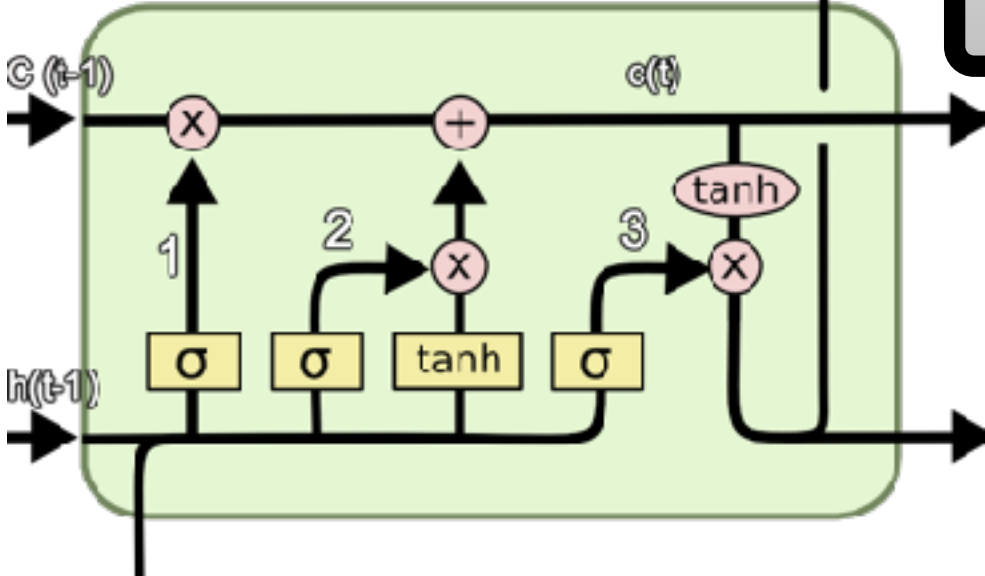
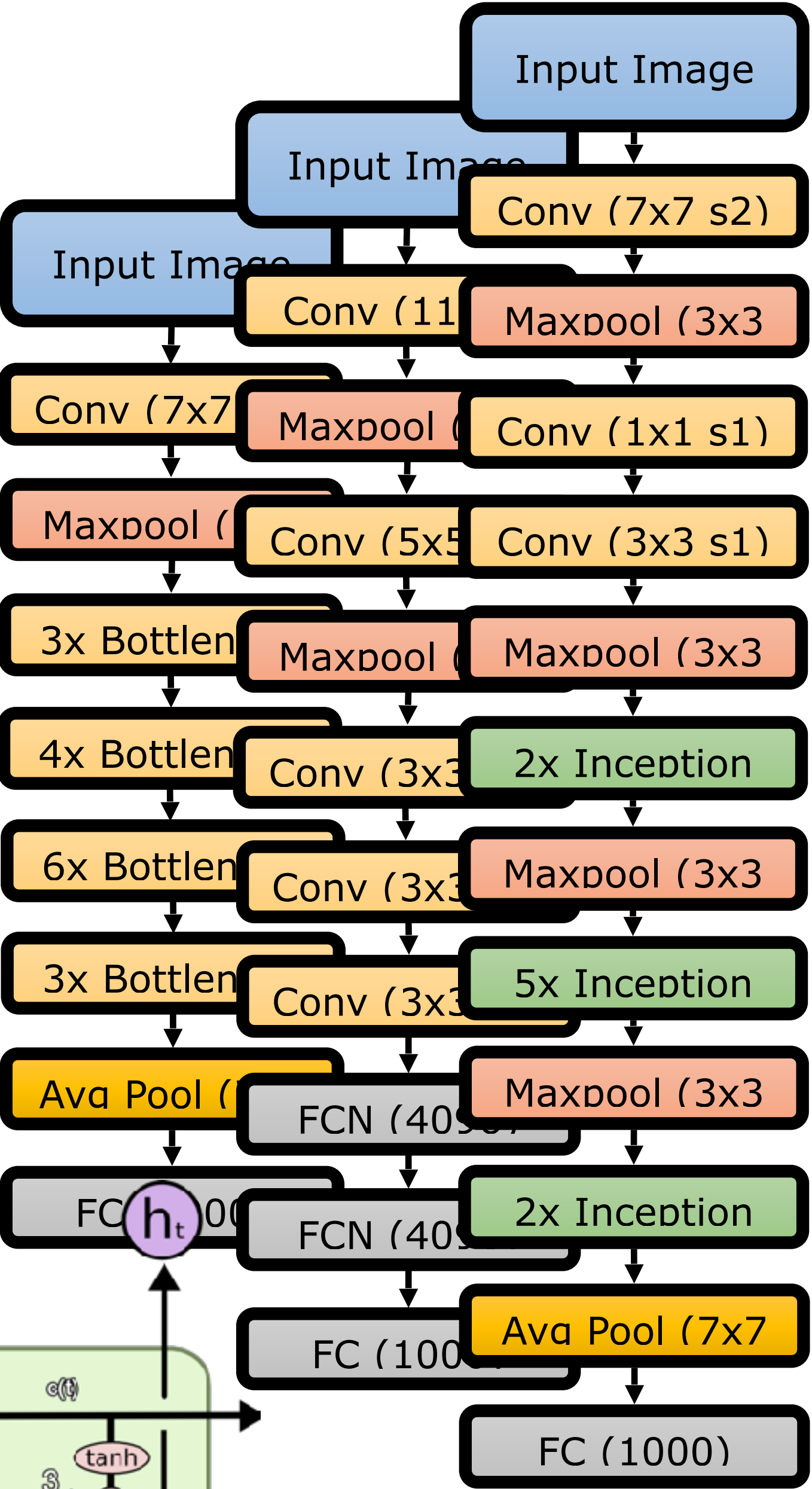


# any neural network

any framework:

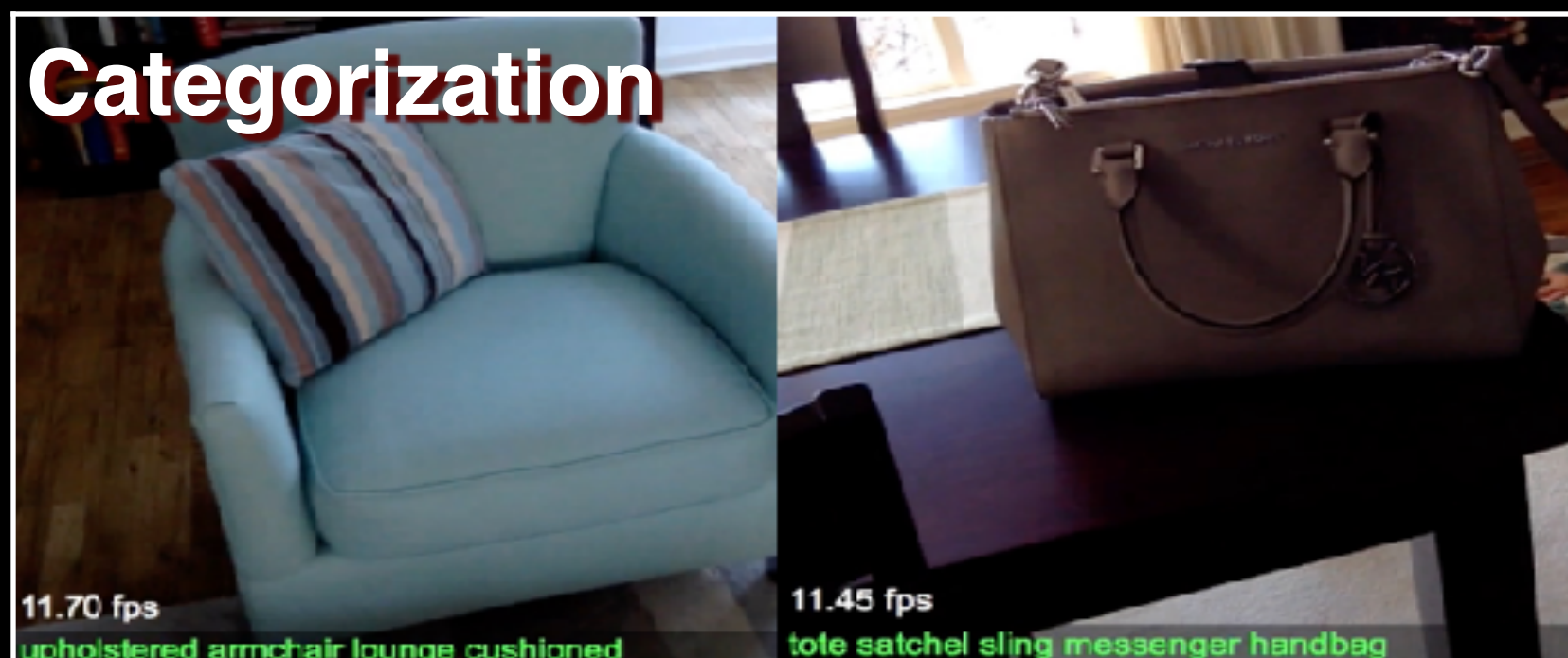


AlexNet  
ResNet  
GoogLeNet  
LinkNet  
...  
encoder-decoder  
...  
RNN  
LSTM  
training

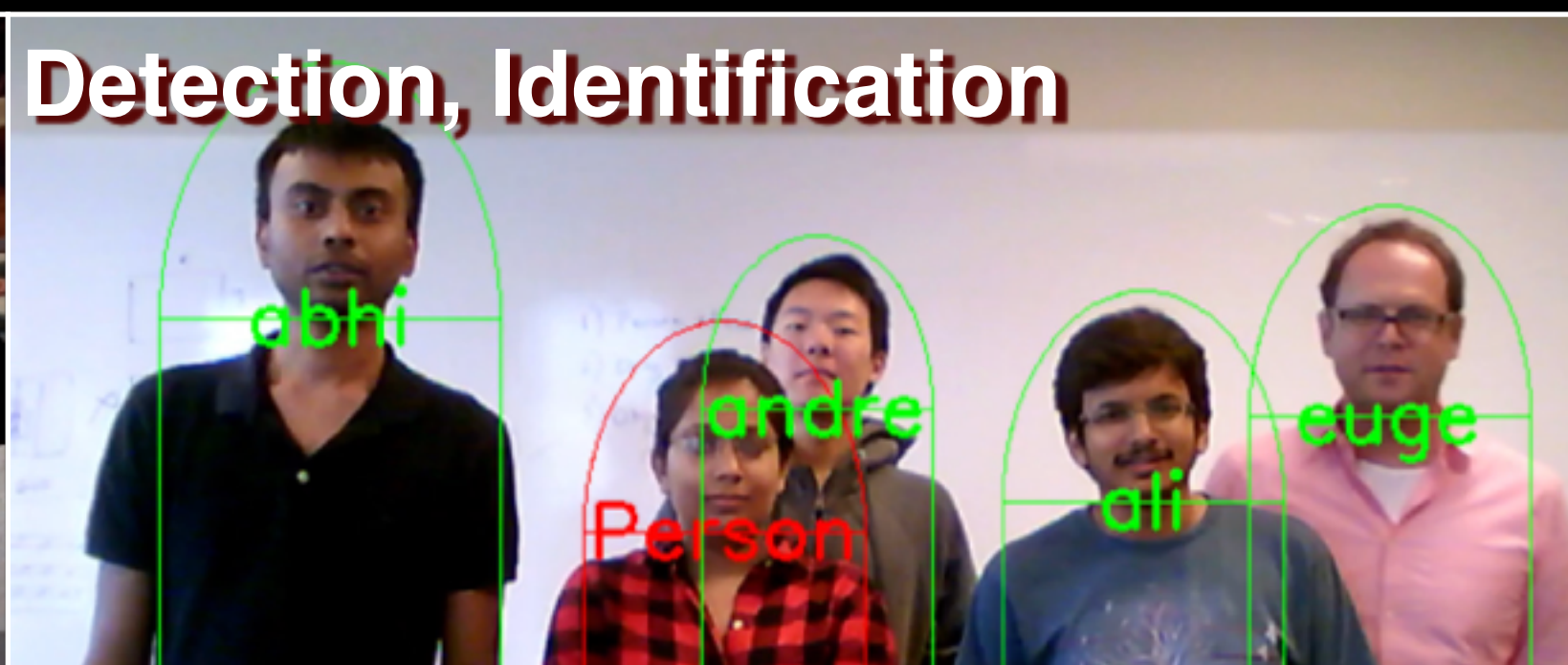




**Categorization**



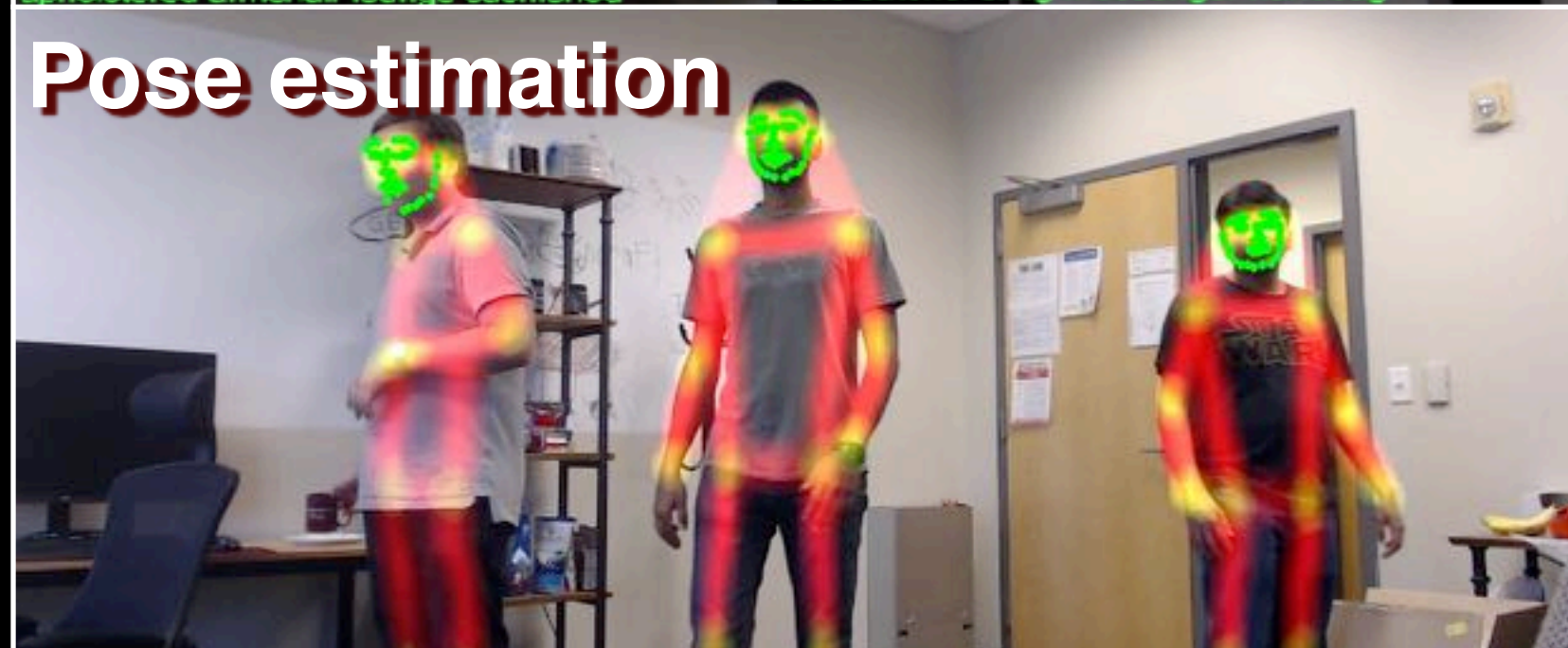
**Detection, Identification**



**Tracking**



**Pose estimation**



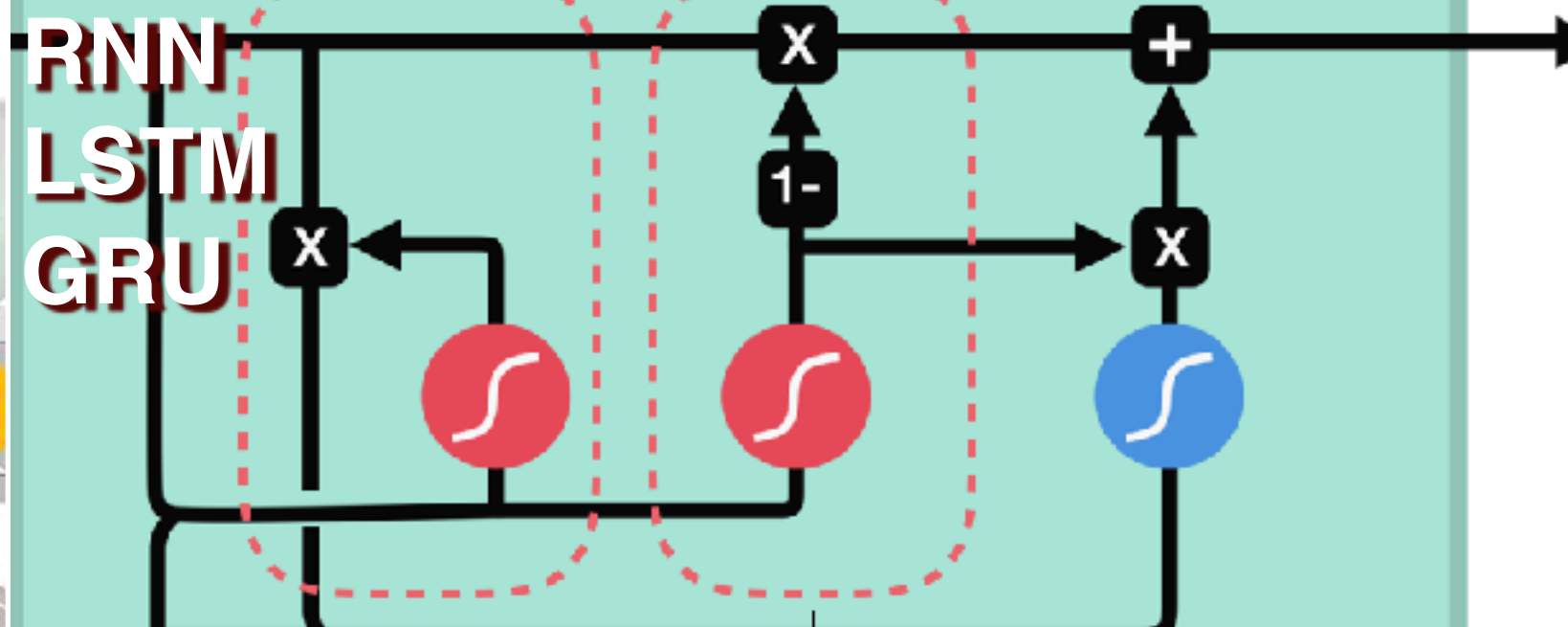
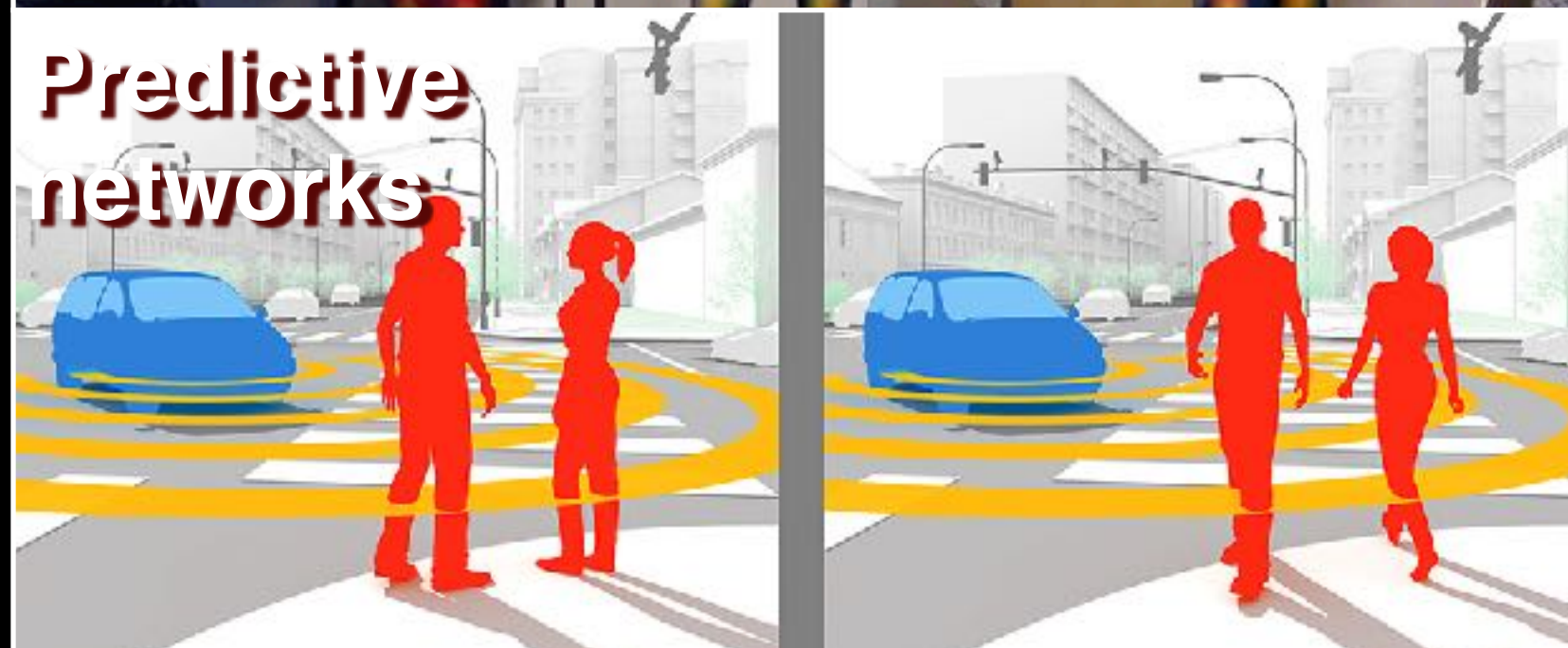
**Segmentation**



**Satellite**



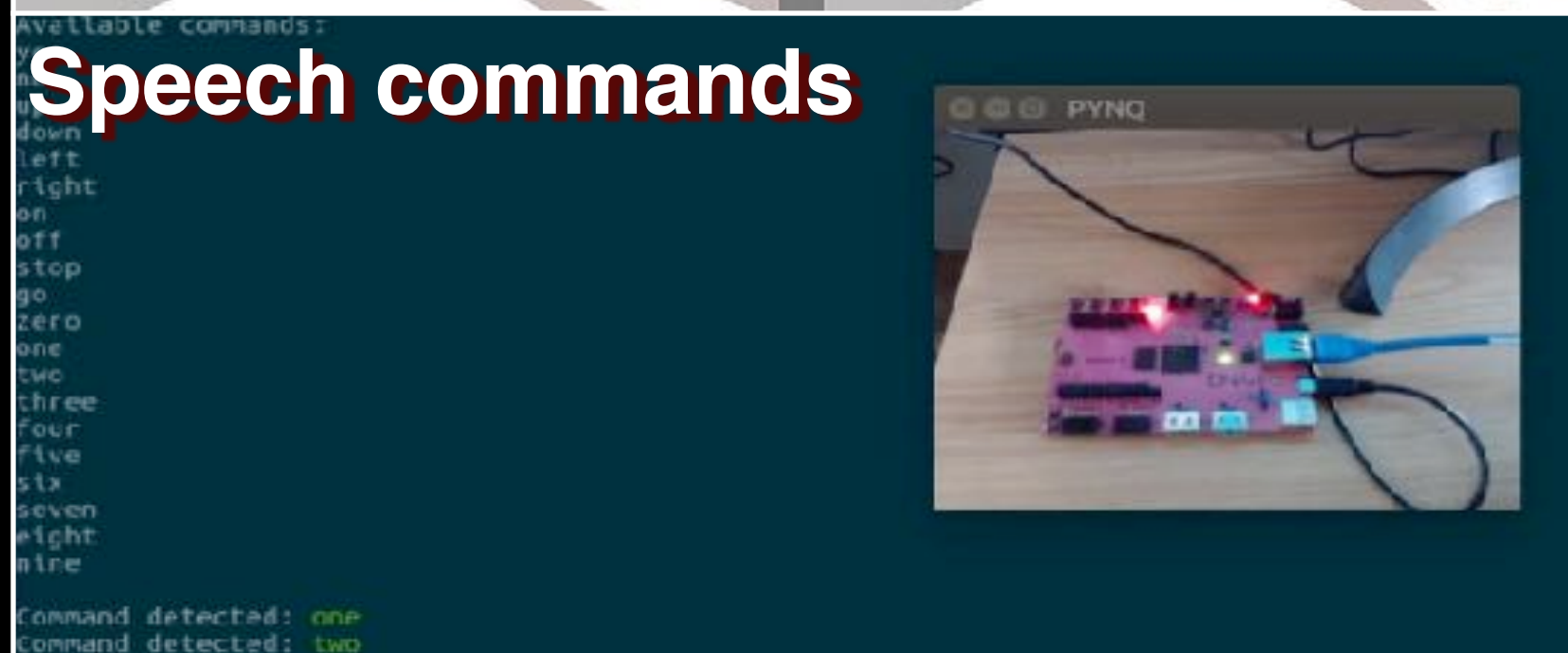
**Predictive networks**



**Speaker identification**



**Speech commands**



**GAN style-transfer**



**Reinforcement Learning**





# The Advantage

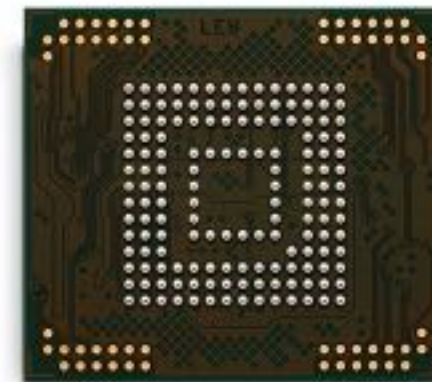


Storage

## NEURAL NETWORKS AI workloads

recommendation engines  
conversational systems

3D NAND



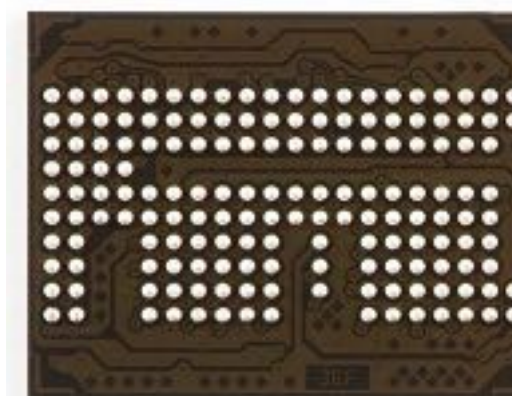
medical imaging  
satellite imaging

low-power DDR



categorization  
speech commands

machine-learning  
memory



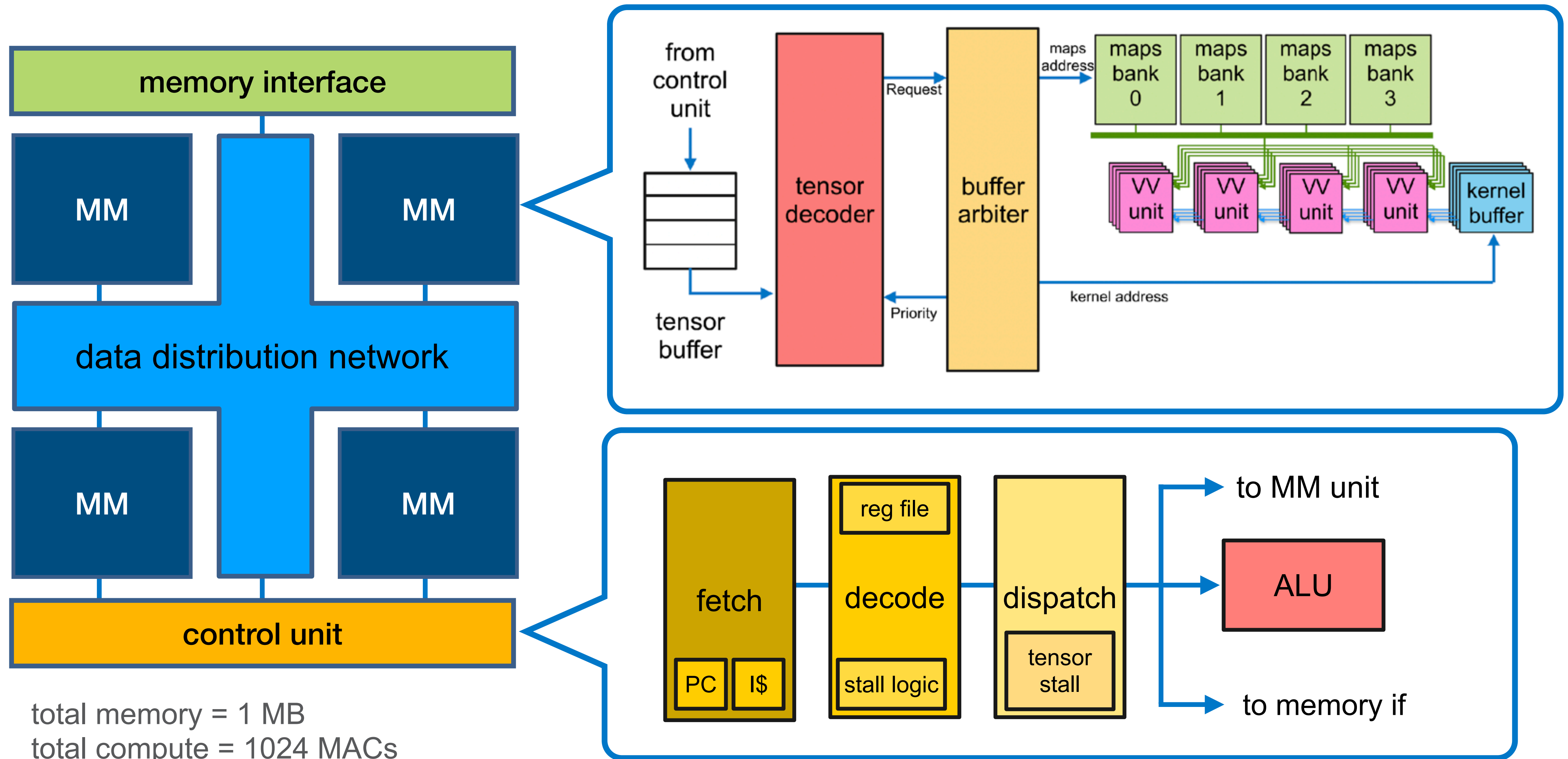
IoT ultra-low-power  
all neural networks!

# Micron DLA

# Deep Learning Accelerator






# Micron Inference Engine Architecture



# Compiler

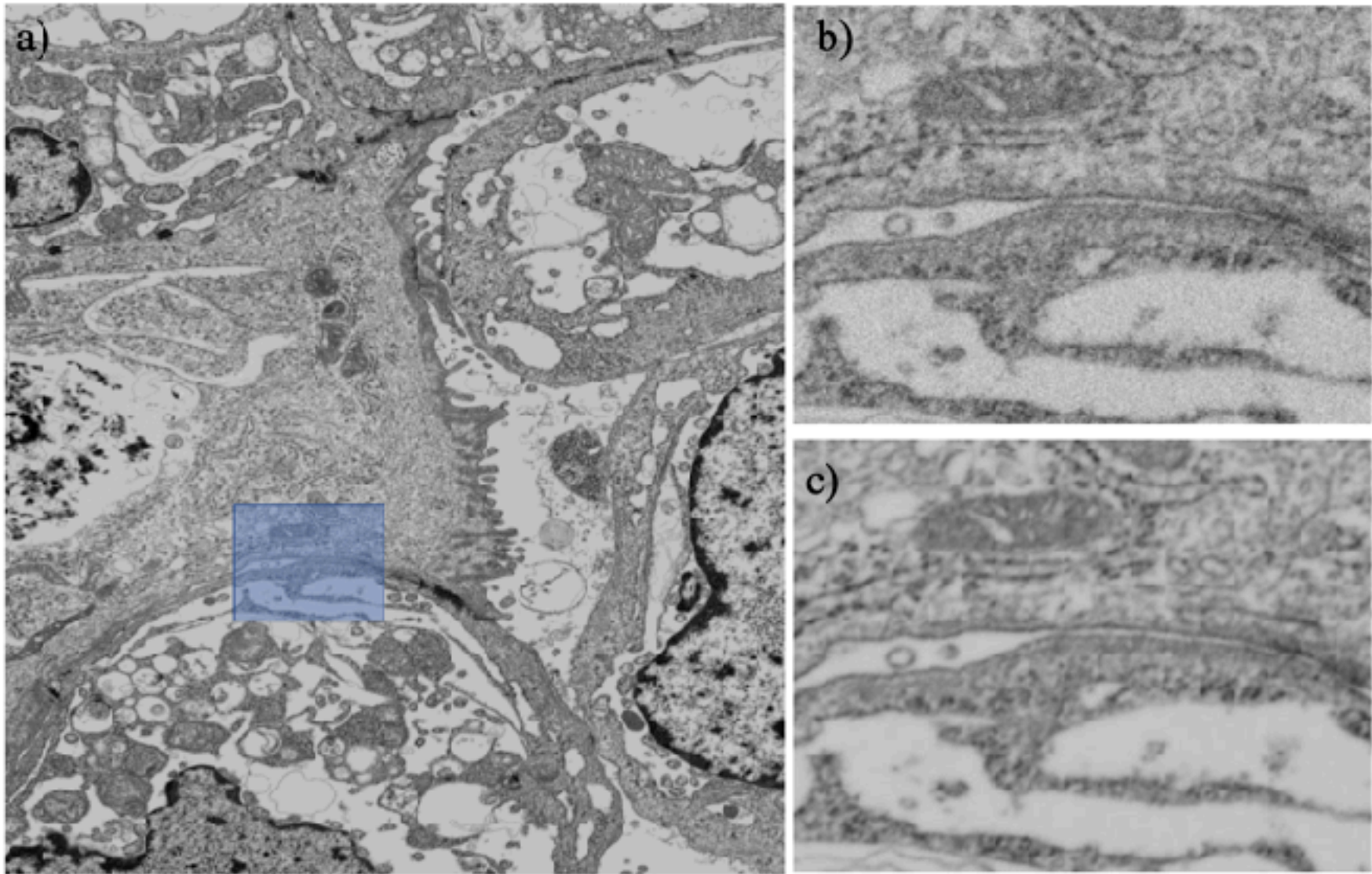
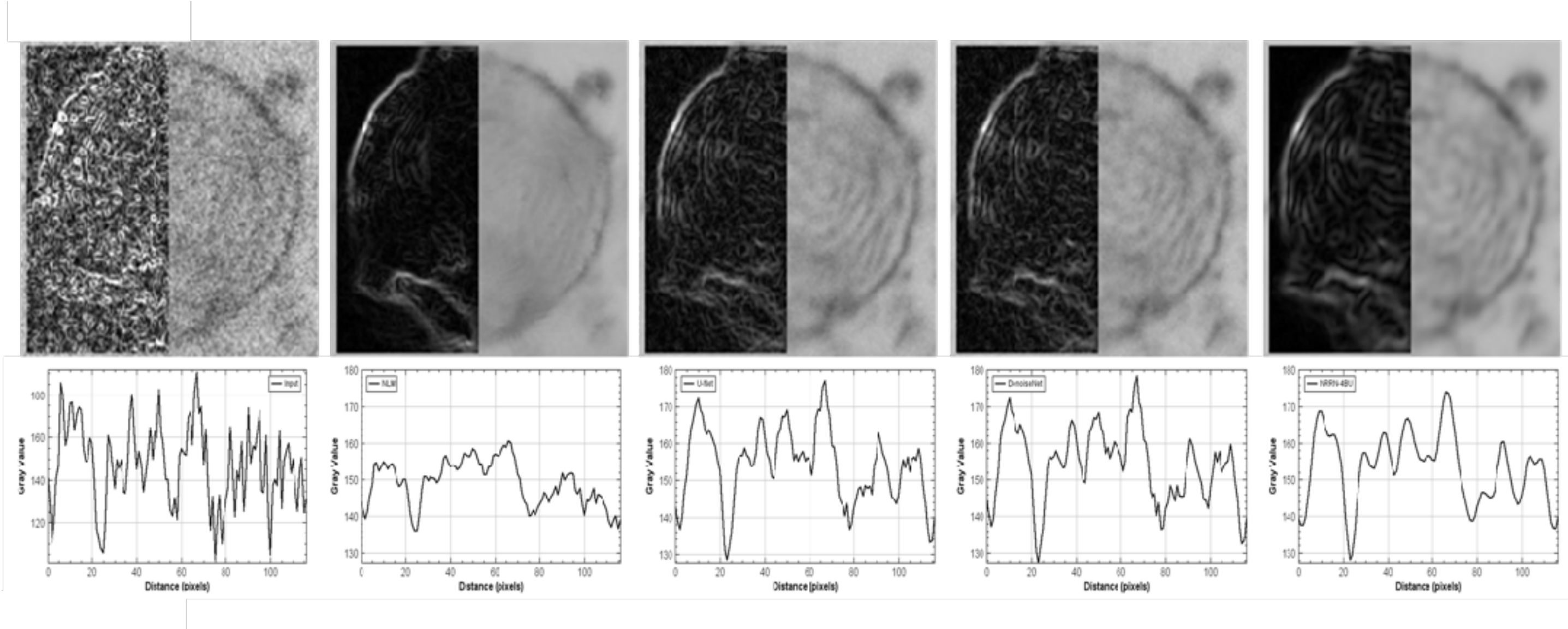


Models	CNN	RNN	Custom	
Framework	 TensorFlow	 PyTorch	 Caffe2	Others
Format	ONNX			
SDK	Python API	C API	Installer and Docs	
Compiler	AI Model Parser		AI Model Quantizer	
	DLA Optimizer		DLA Assembler	
	DLA Run Time			
Hardware	FPGA DLA		DLA	

# Verticals



# Healthcare and scientific

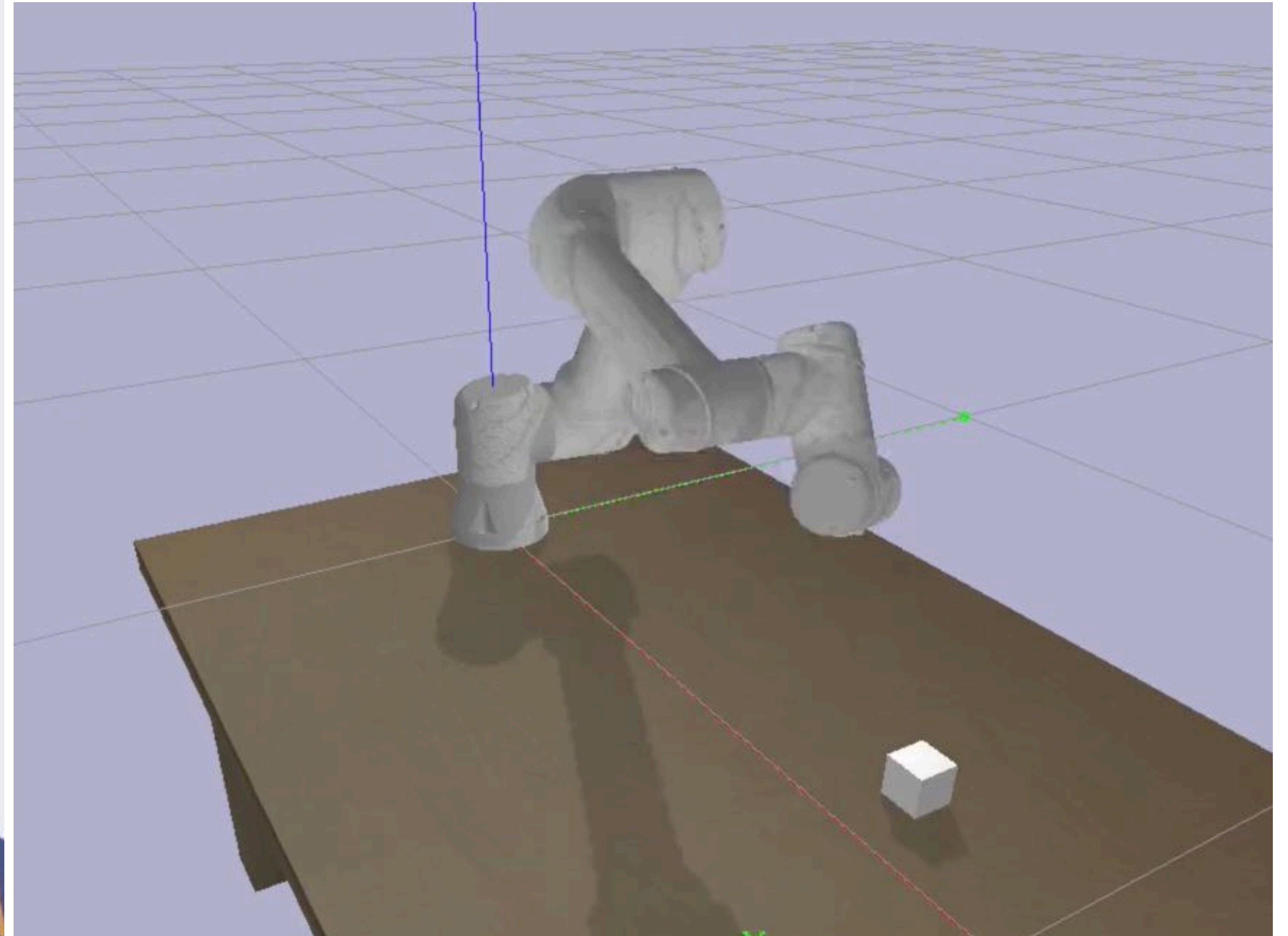
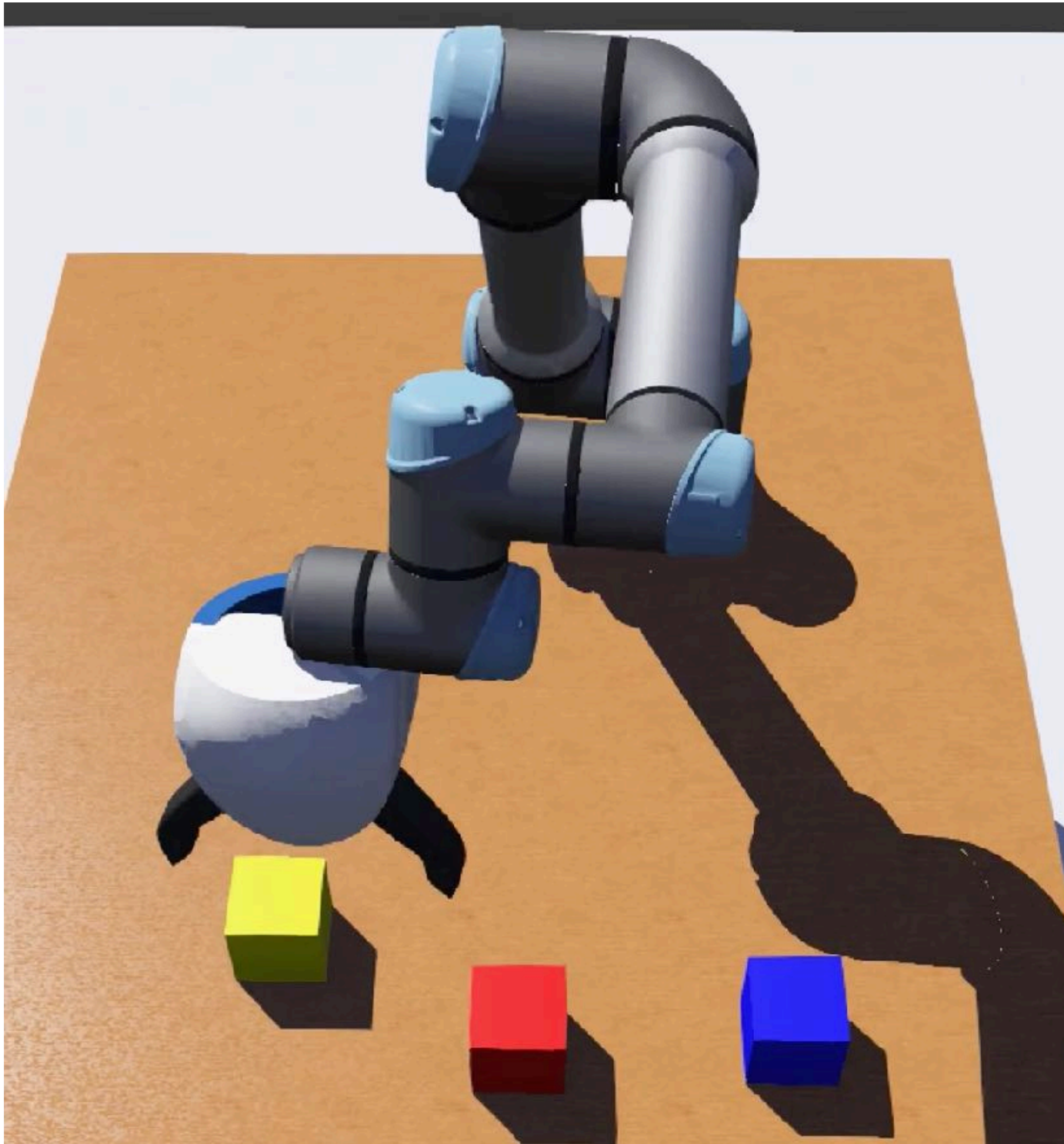


Method: NRRN-4BU, PSNR: 31.106dB, SSIM: 0.9708, IQR: 0.941



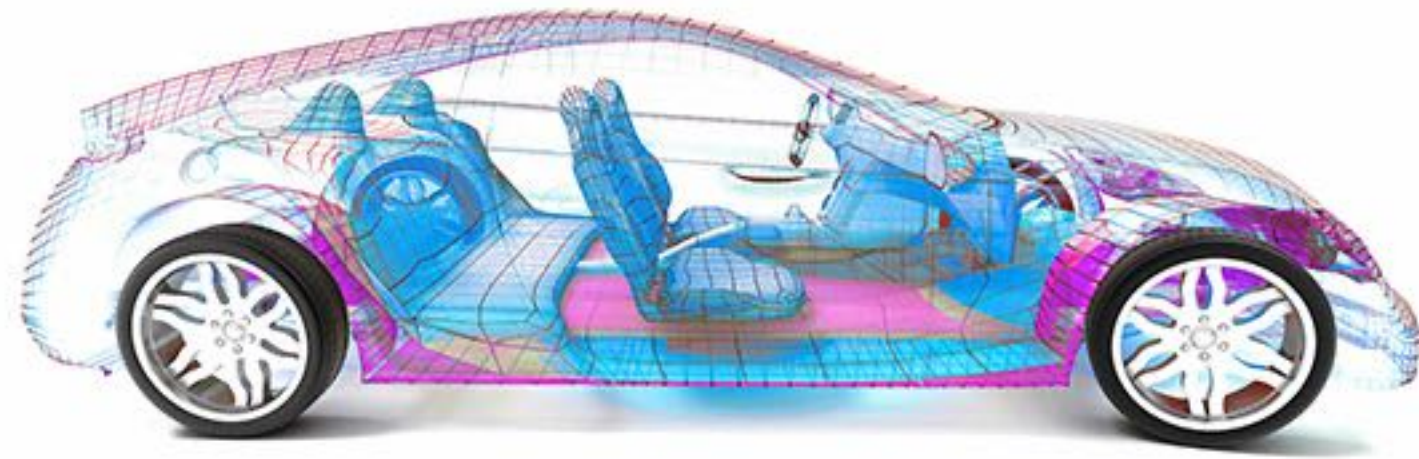


# Manufacturing and Robotics





# Autonomous systems





# Security, Smart cameras





# 3 years with OpenLab



# 3 Years with OpenLab

- Contract began in 2018
  - Boards and software stack provided to teams, collaboration begins
- Onsite visit with ProtoDUNE in June 2019
- CMS hosted in Seattle December 2019





# Current Collaboration Highlights

- Working with CMS on Scouting
  - Scouting network  $< 1 \text{ us}$  / inference
  - Support for QSFP bringup, architecture changes to minimize latency
  - Great work Thomas & team for results on the right!
- Working with ProtoDUNE to implement GNNs on hardware
  - Entirely new application for DLA, more to come

