Imperial College London

Towards Efficient On-Board Deployment of DNNs on Intelligent Autonomous Systems

Alexandros Kouris, Stylianos I. Venieris, **Christos-Savvas Bouganis**

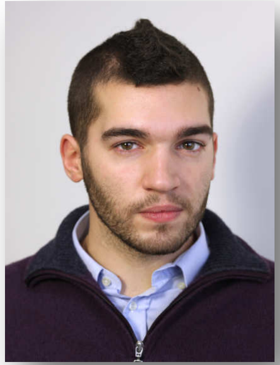
IEEE Computer Society Annual Symposium on VLSI

ISVLSI, 17 July 2019

Intelligent Digital Systems Lab
Dept. of Electrical and Electronic Engineering

www.imperial.ac.uk/idsl

Who we are



Stylianos I. Venieris
Machine Learning
(now with Samsung AIC)

Alexandros Kouris
Machine Learning,
Robotics

Konstantinos Boikos
Computer Vision, SLAM

Aditya Rajagopal
HW for Machine Learning

Manolis Vasileiadis
Computer Vision

Mudhar Bin Rabieah
Machine Learning

Nur Ahmadi
Brain-Machine Interface

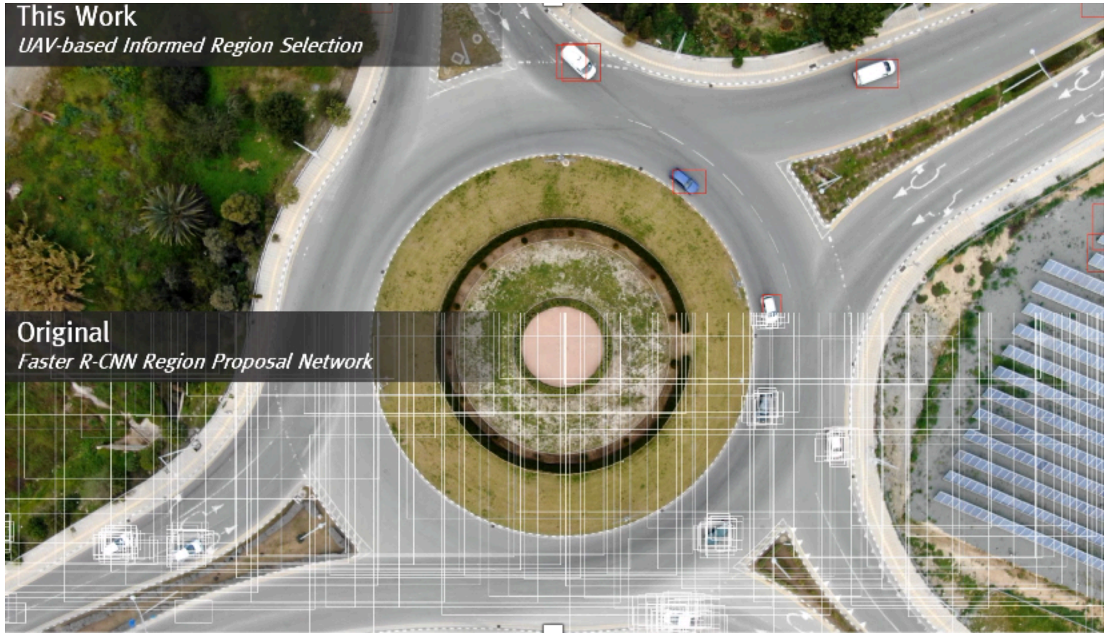
Diederik Vink
Machine Learning

Mario Lopes Ferreira
Research Assistant

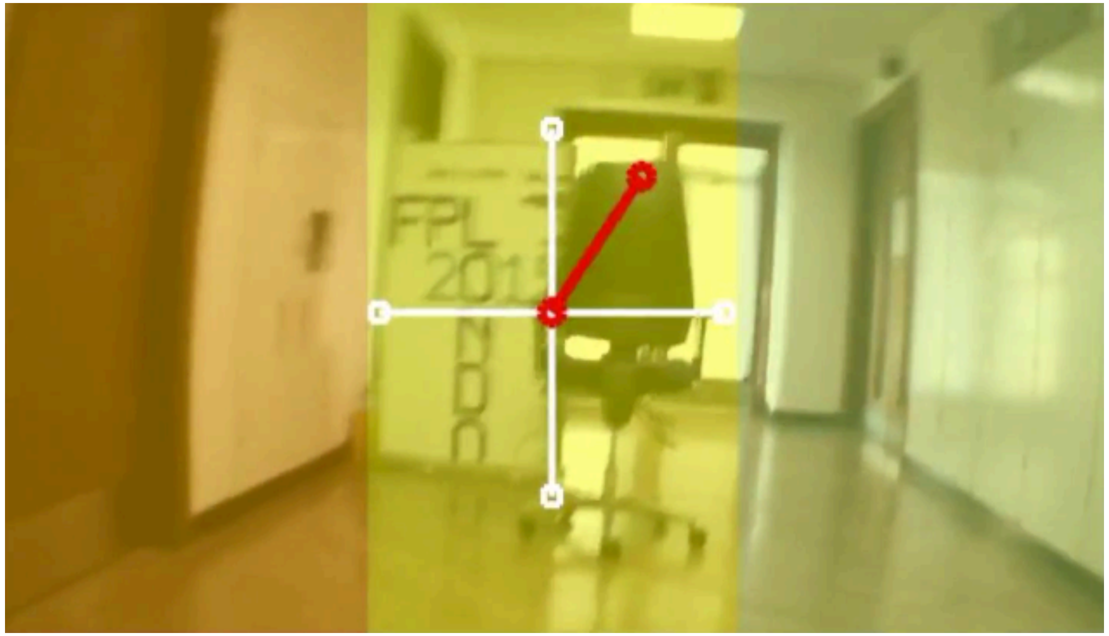
Christos-Savvas Bouganis
Lab Director
Reader at
Imperial College London

Examples of Intelligent and Autonomous Systems

Traffic Detection



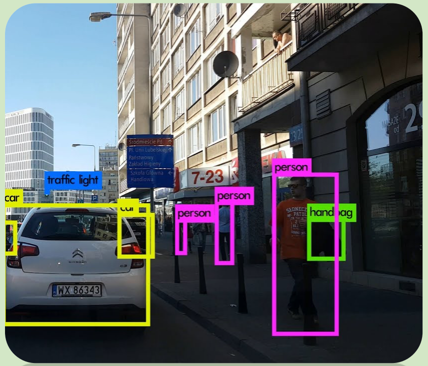
Autonomous Navigation



DNNs on Intelligent Autonomous Systems

Camera

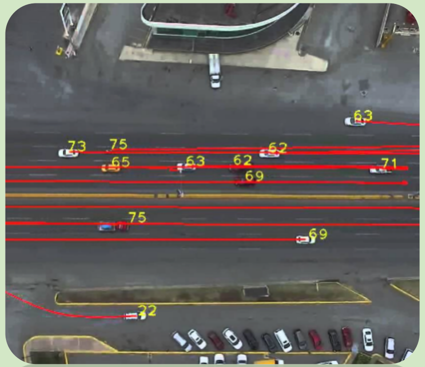
Set of CNNs



Object Detection

Semantic Segmentation

Navigation

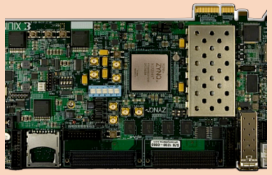


Monitoring

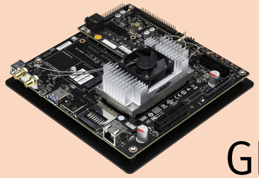
Domain Task

Mapping?

Target Platform



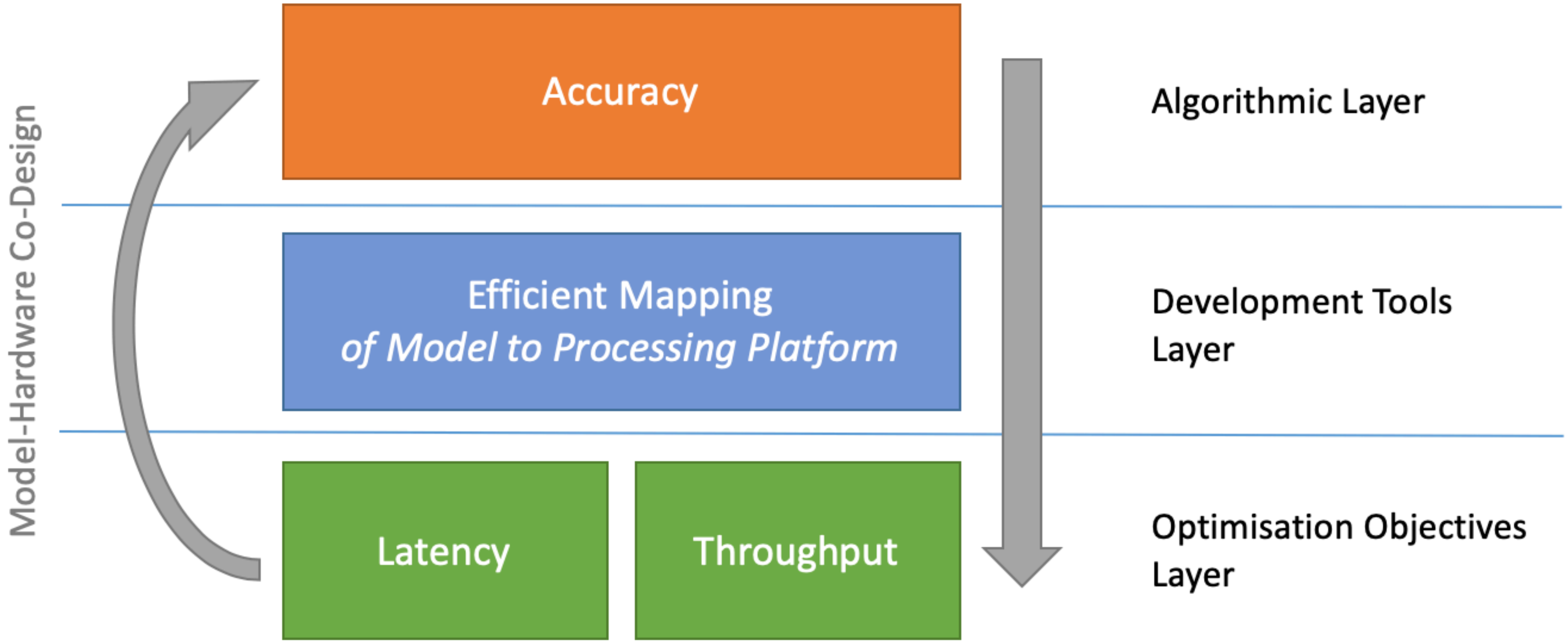
FPGA



GPU

DSP

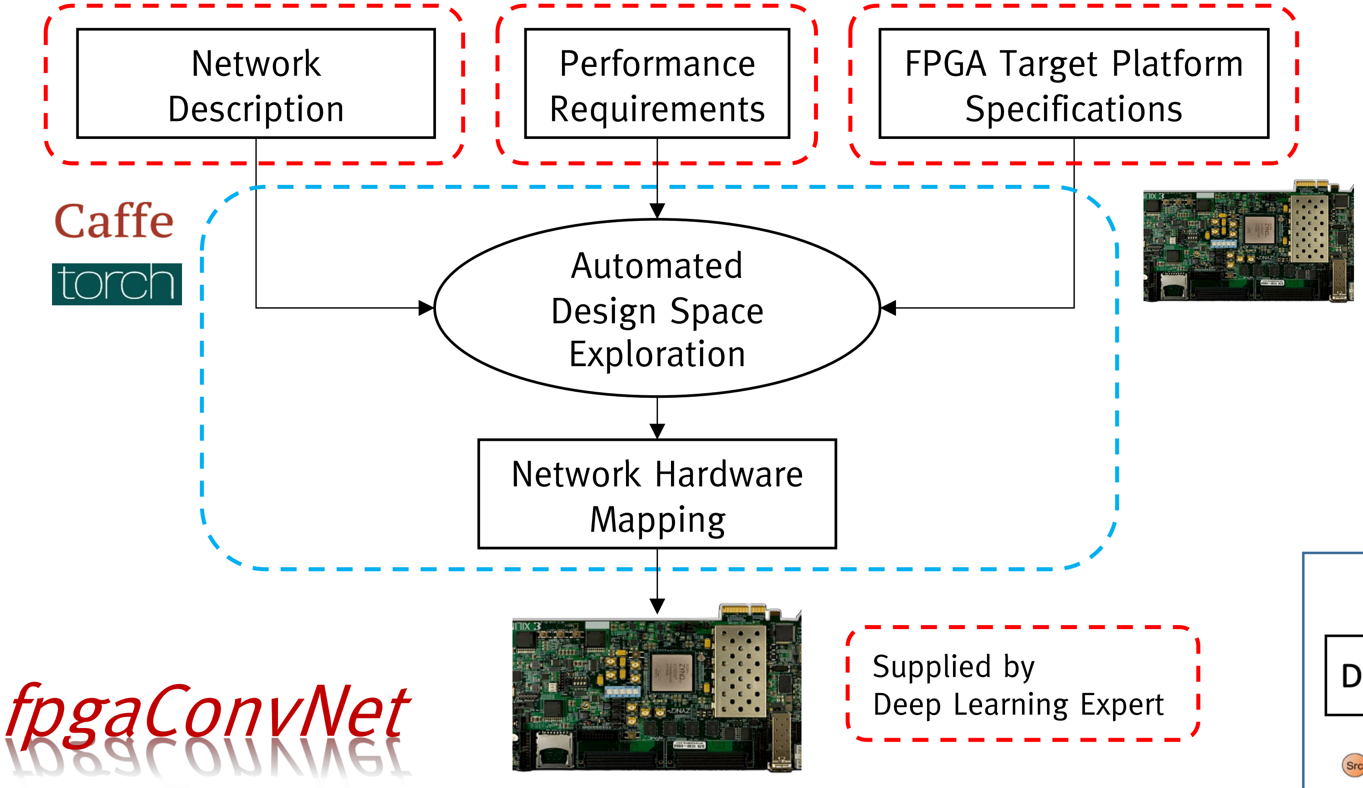
Our Approach – Intelligent Autonomous System Development Stack



Research Areas

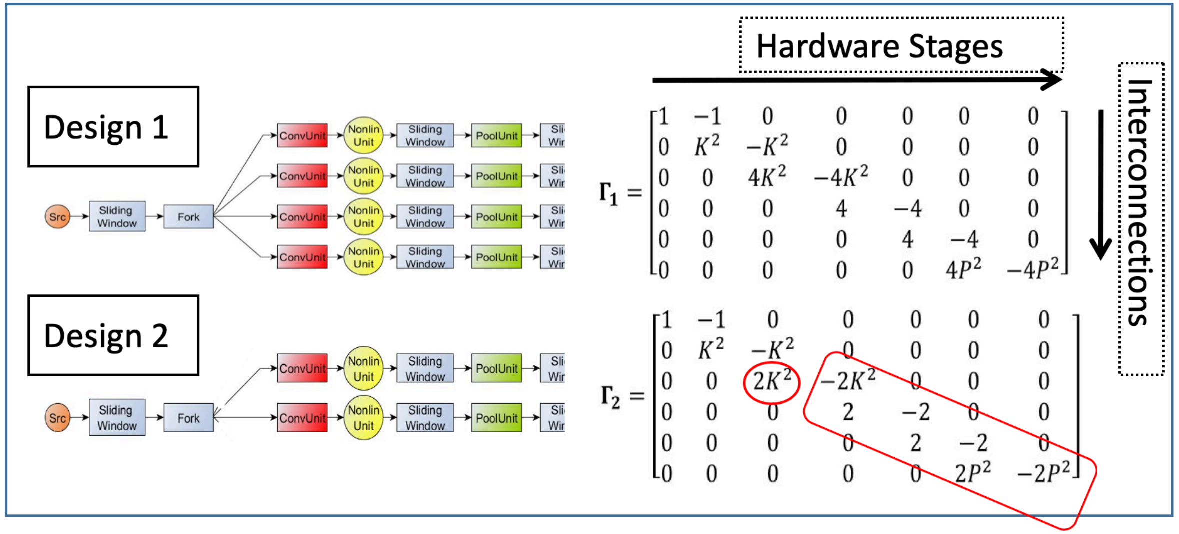
- Latency-Optimised CNN Inference
- Multi-CNN Systems
- Approximate LSTM Inference
- High-Throughput Perception

Latency-Optimised CNN Inference



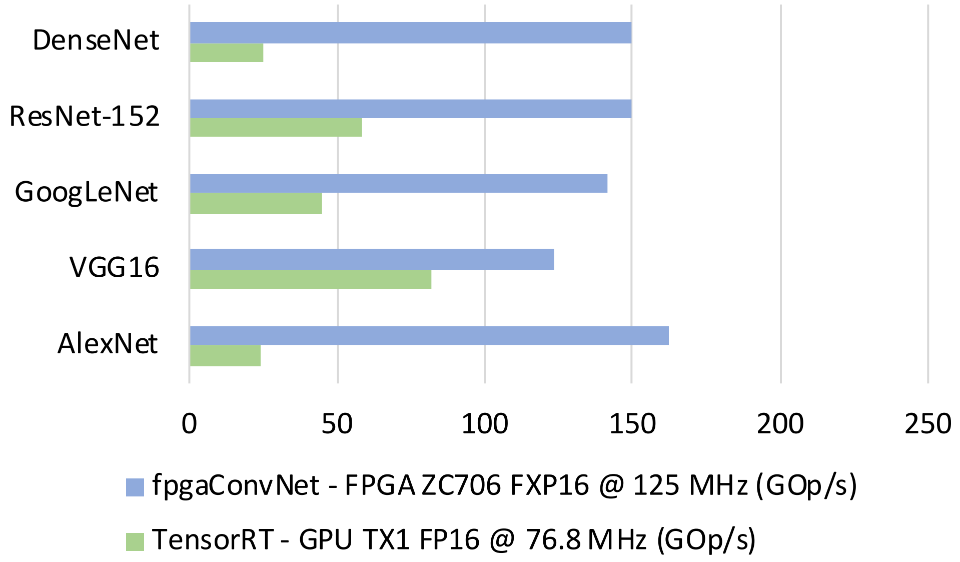
Challenges:
- High-dimensional design space
- Diverse application-level needs
- Utilise the FPGA resources
- Design automation

Synchronous Dataflow Modelling
- Capture hardware mappings as matrices
- Transformations as *algebraic operations*
- Analytical *performance model*
- Cast design space exploration as a mathematical optimisation problem

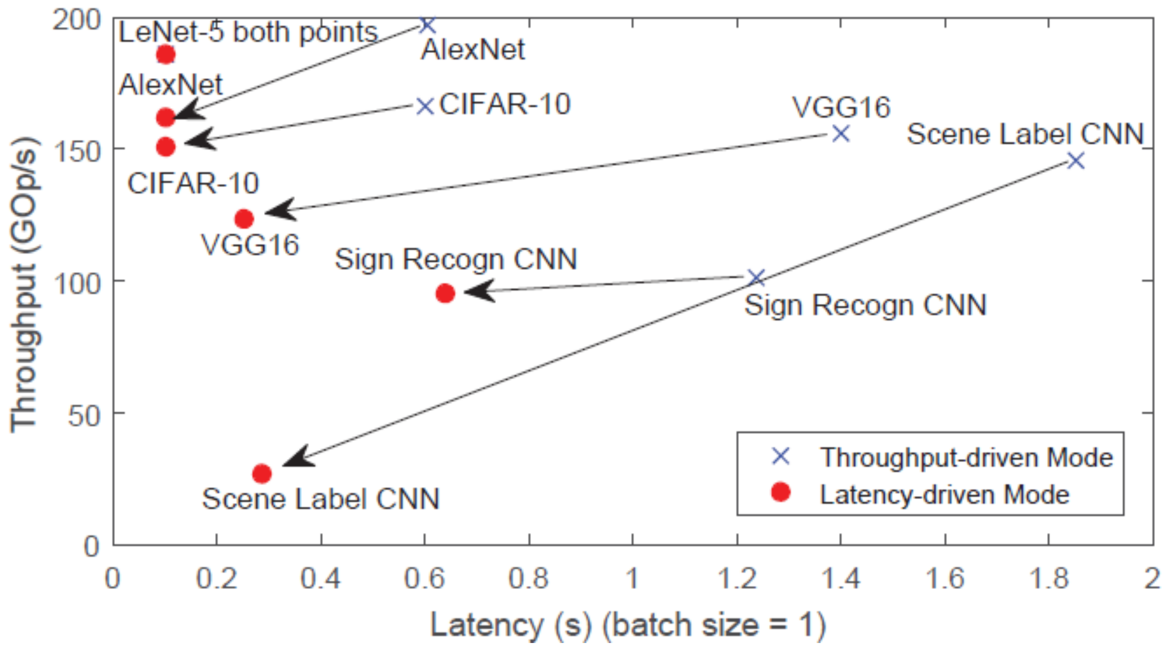


Latency-Optimised CNN Inference

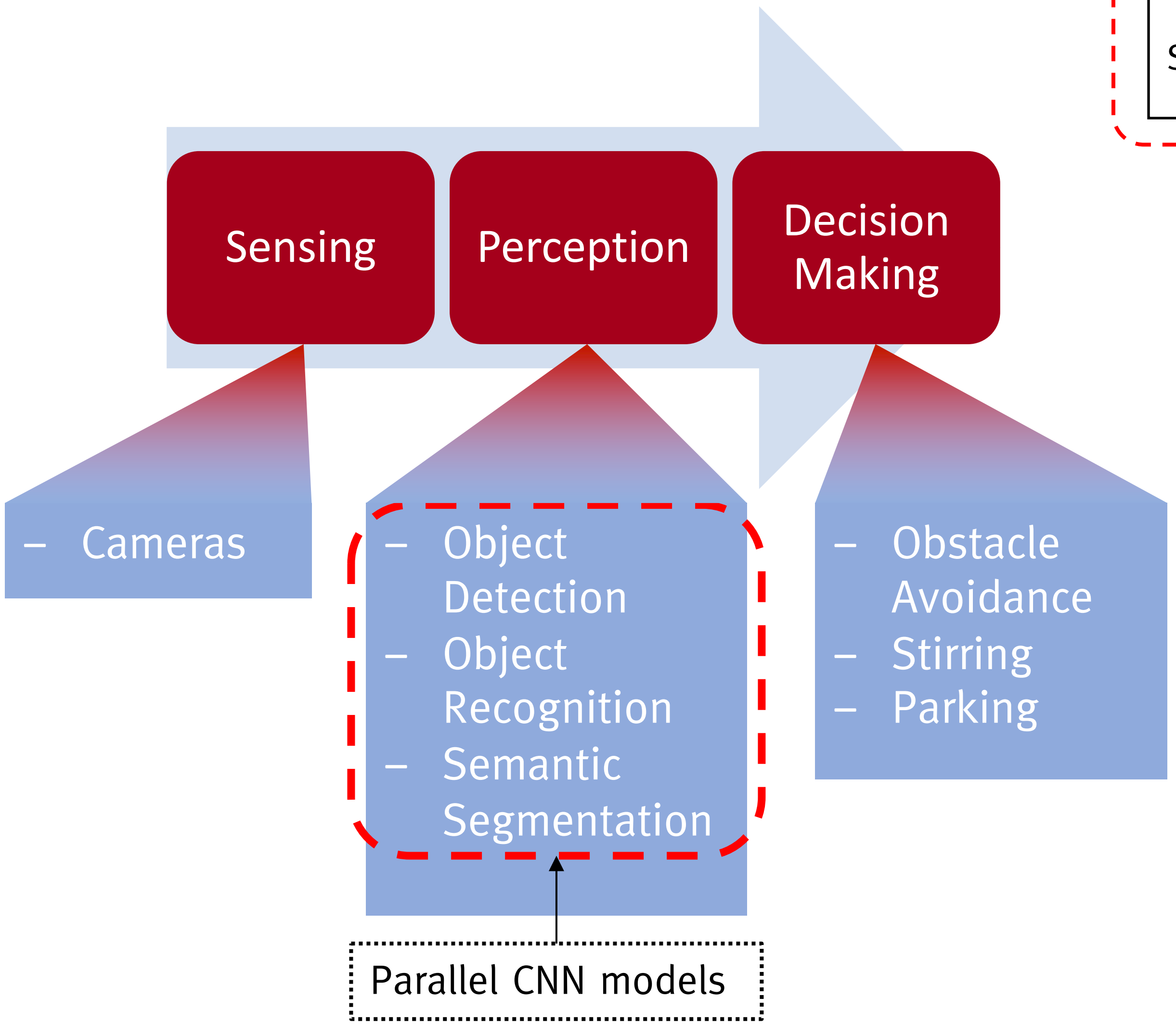
fpgaConvNet vs Embedded GPU (GOp/s)
for the same absolute power constraints (5W)

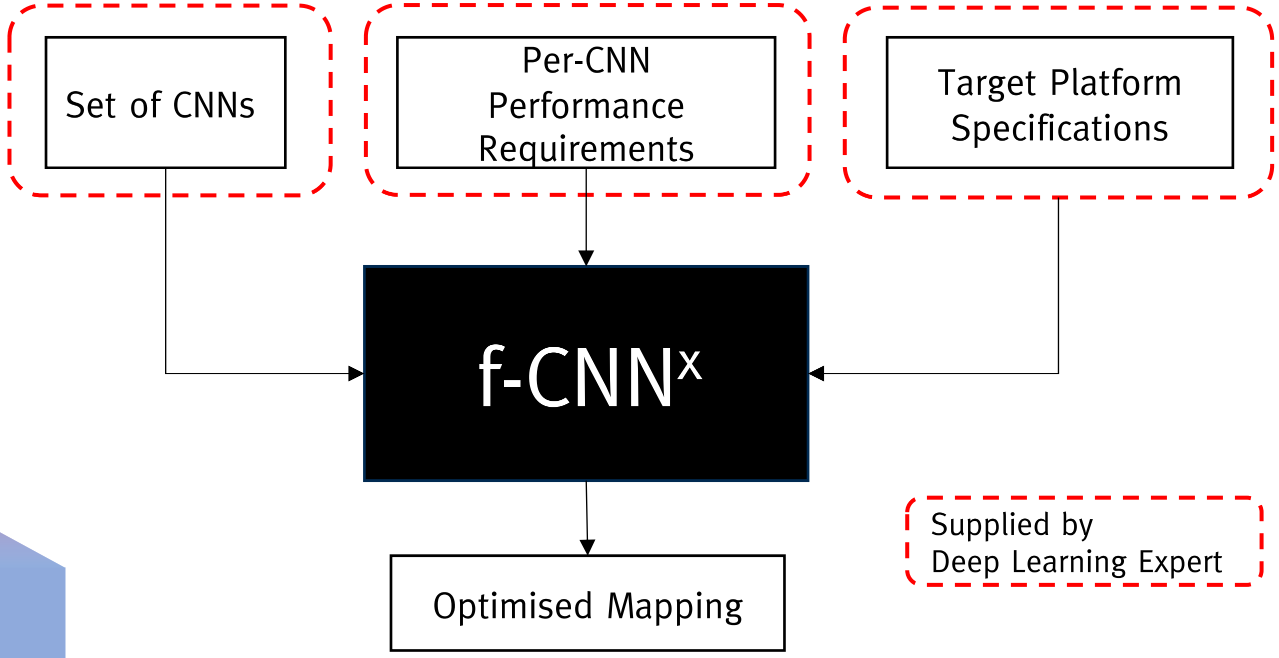


- Latency-driven scenario → batch size of 1
- Up to 6.65× speedup with an average of 3.95× (3.43× geo. mean)



Multi-CNN Autonomous Systems





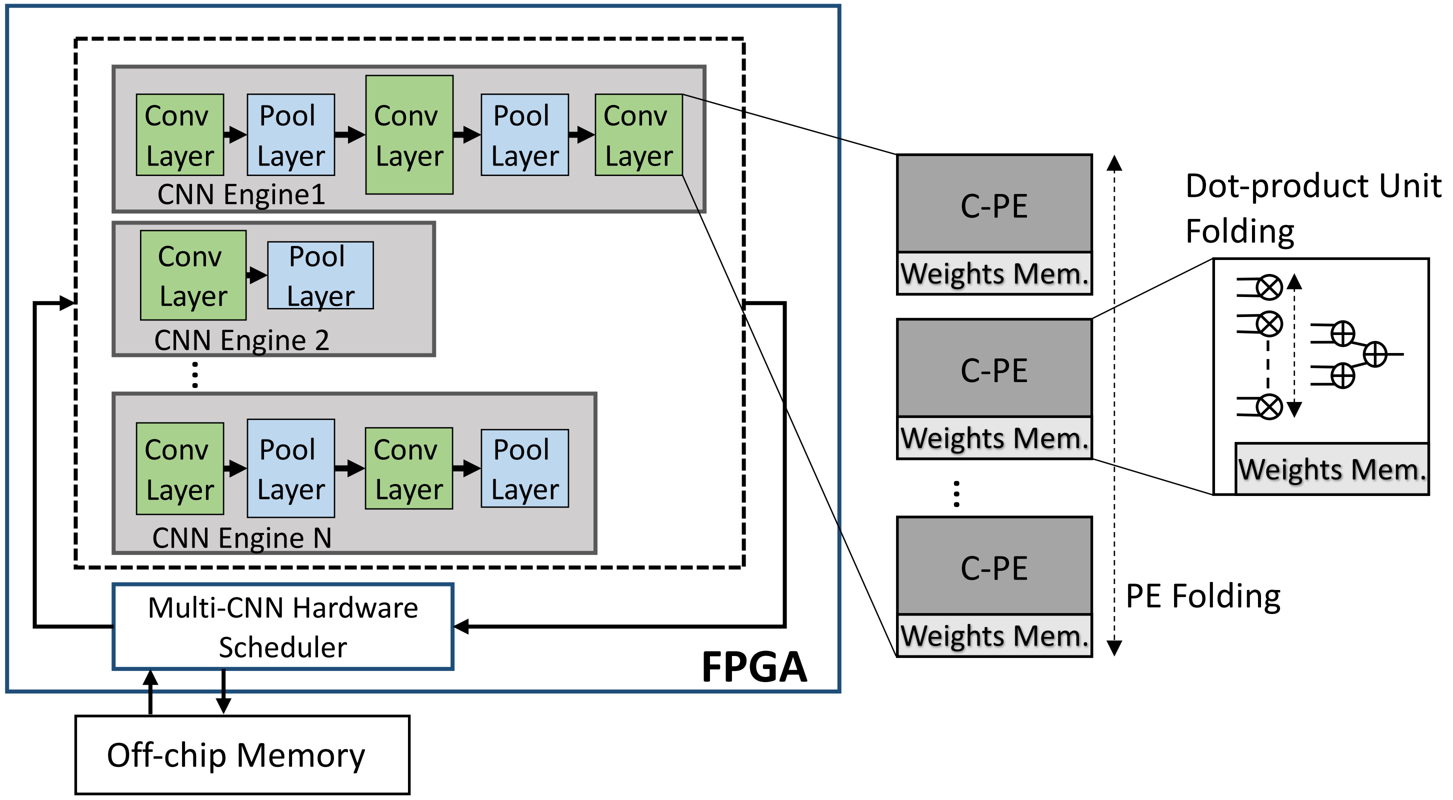
Challenges:

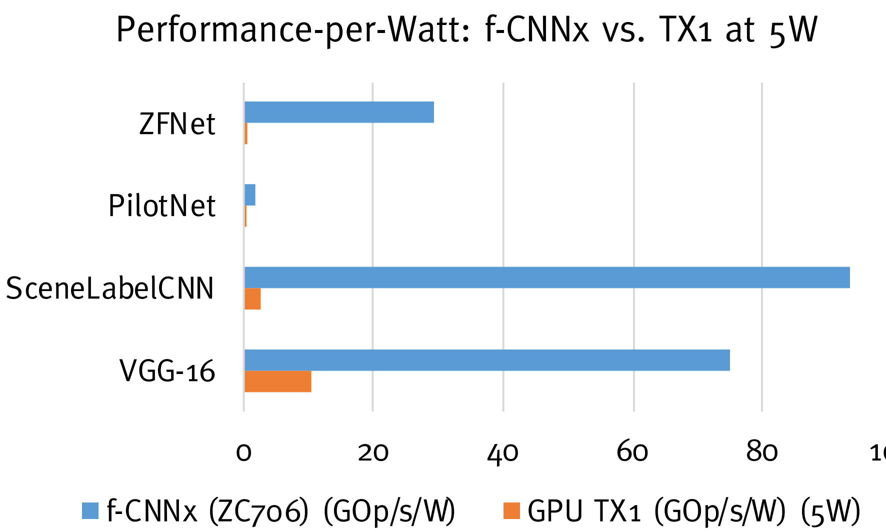
- Resource allocation among CNNs
- Design automation

Why?

- Models with different performance constraints
- Competing for the same pool of resources
- High-dimensional design space

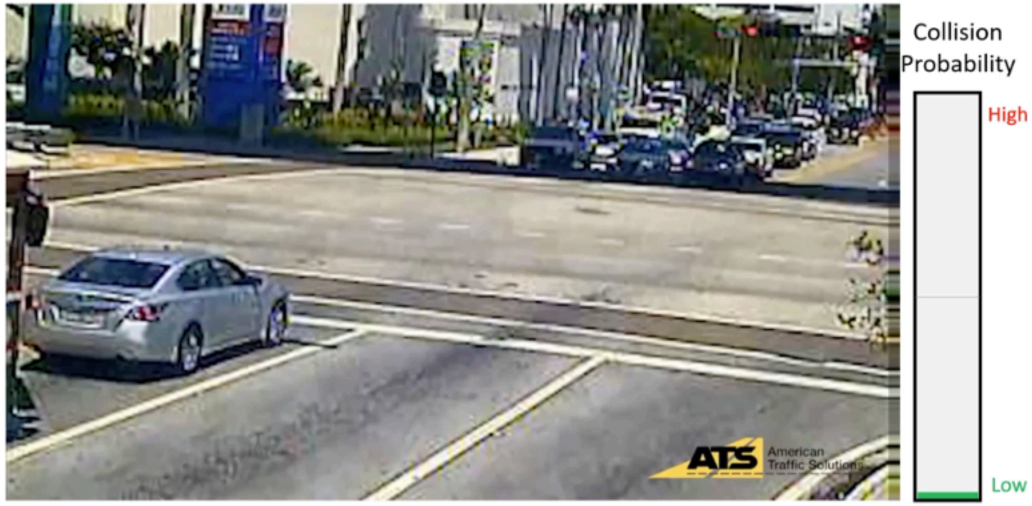
Multi-CNN Autonomous Systems

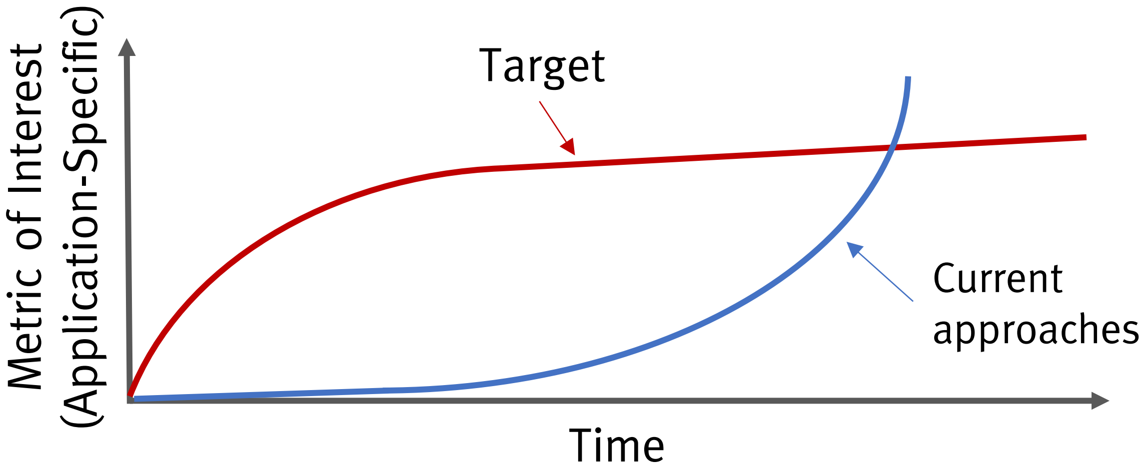


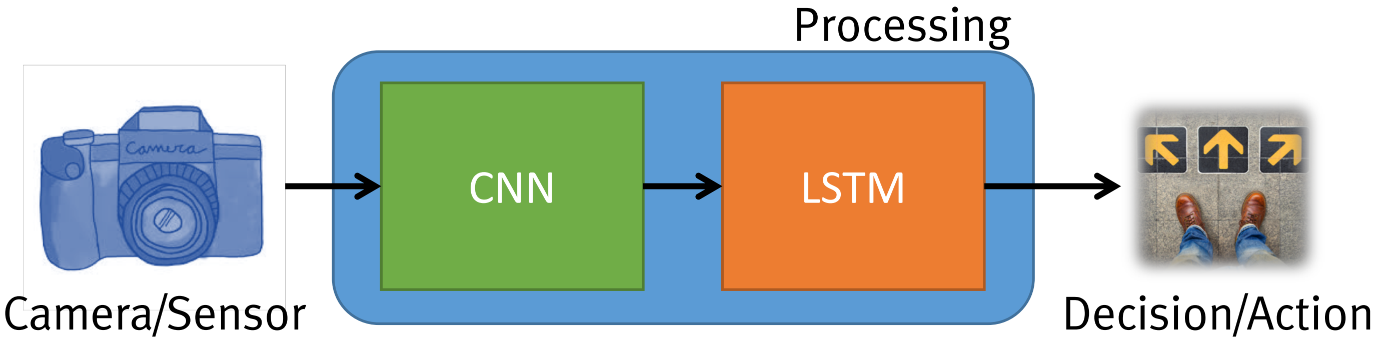


- Latency-driven scenario → batch size of 1
- Up to 19.09× speedup with an average of 6.85× (geo. mean)

Time-constrained Approximate LSTM Inference

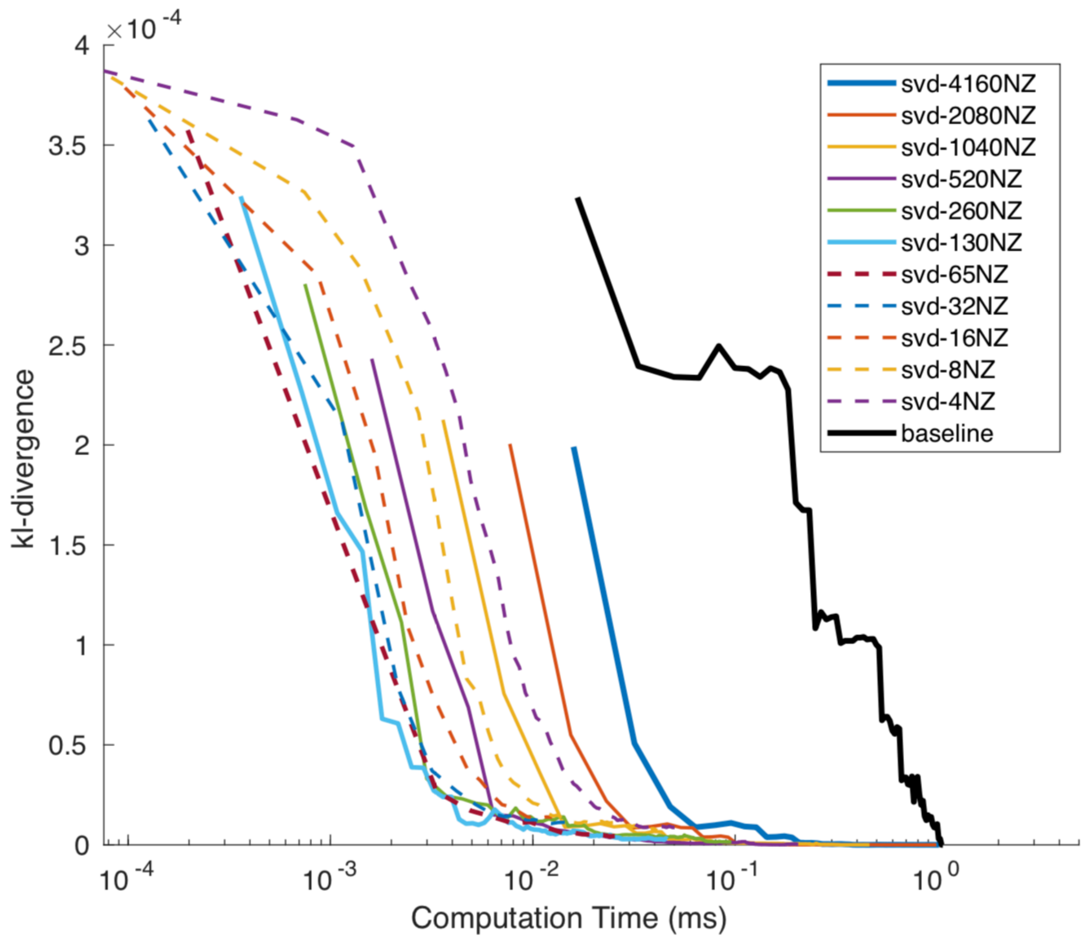




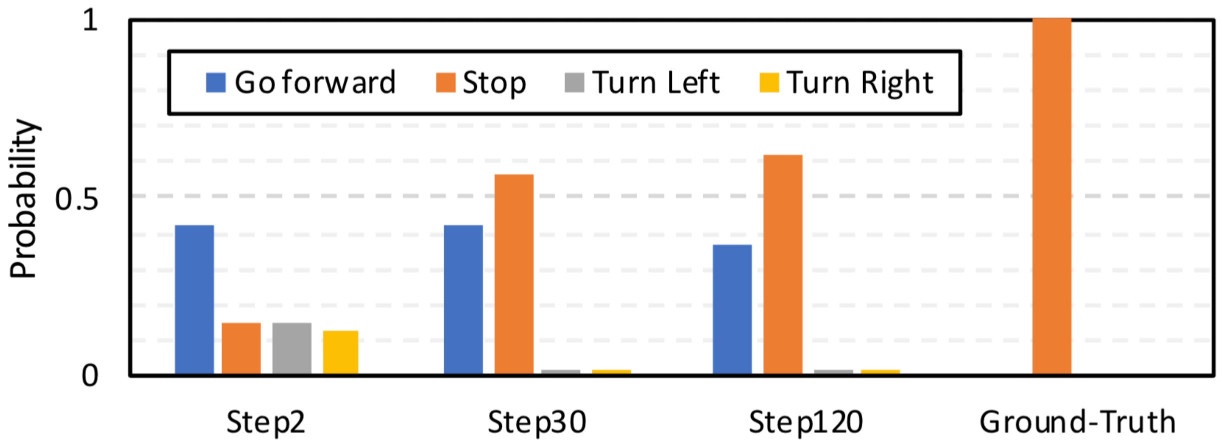


- Approximate LSTMs
  - Iterative refinement using:
    - SVD-based low-rank approximation
    - Sparsification (structured pruning)
- Co-optimise given a user-defined time budget
- Custom parametrisable architecture

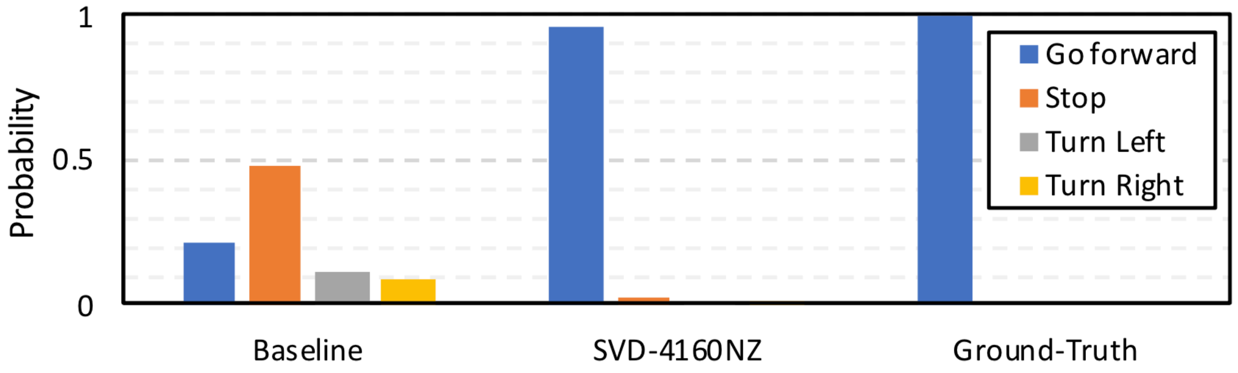
Time-constraint Approximate LSTM Inference



Progressive Inference:



Time-Constraints Inference:

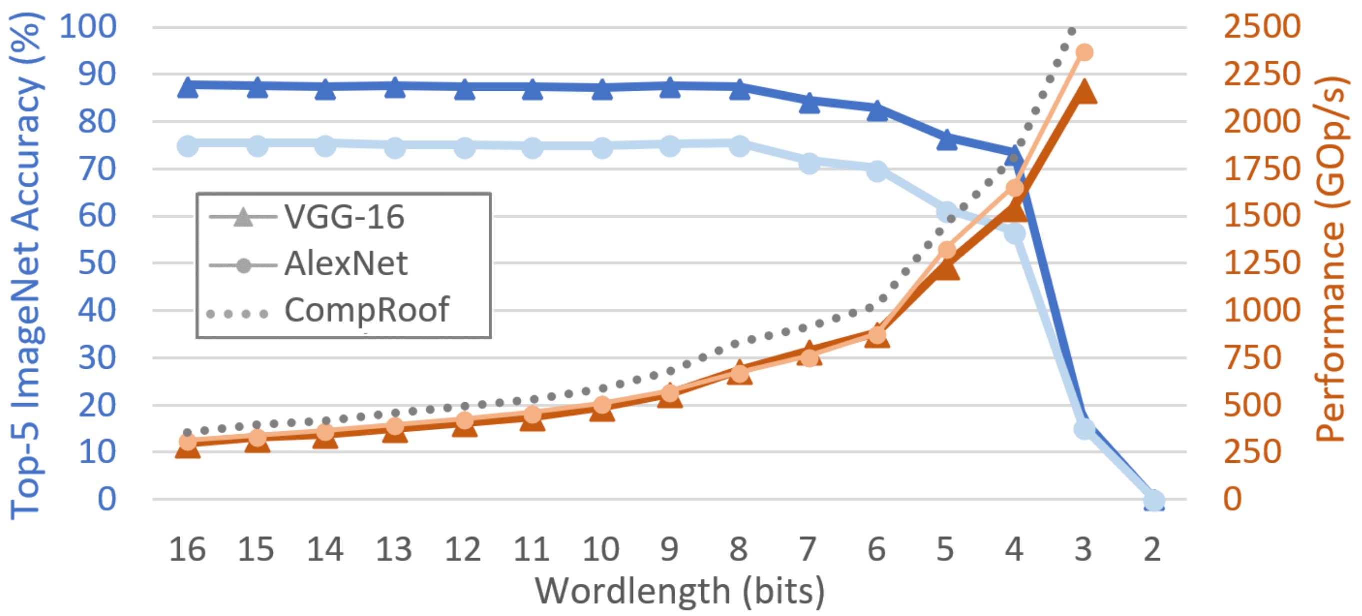


Privacy-aware High-Throughput Inference

Aim: Design an optimised HW system (performance and accuracy)

Given:

- A High-Level CNN Description (i.e. Caffe)
- A target FPGA platform
- *~~Training~~ Data* privacy, availability
- *Testing Data*
- **Target metric (top1/top-5 accuracy, …)**

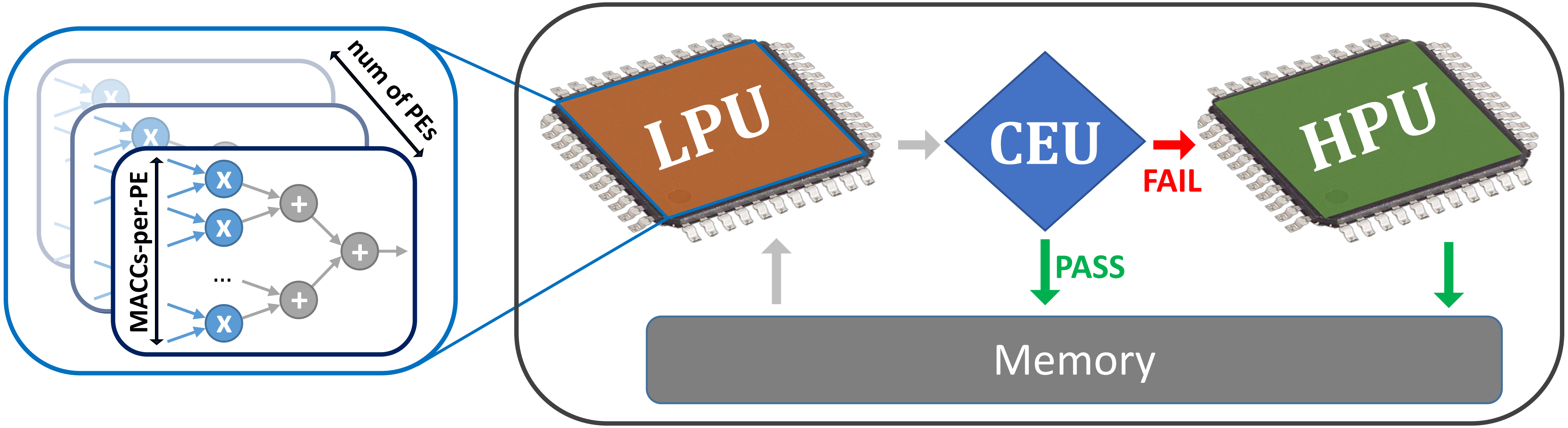


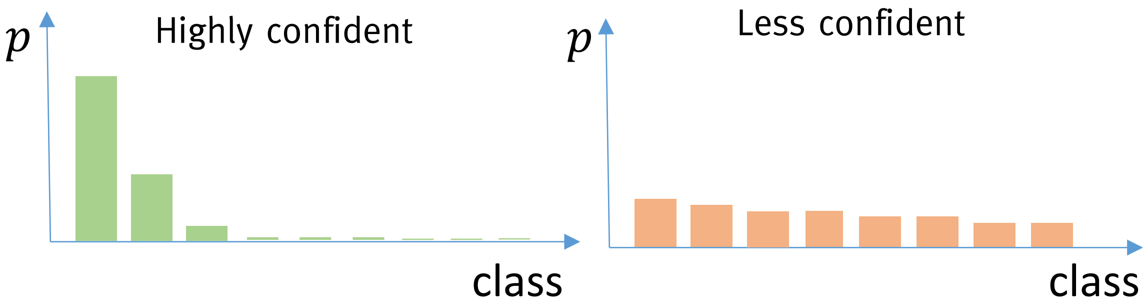
CascadeCNN:

- Exposes the application-level *error tolerance* to the Design Space Exploration
- Develops *highly parametrised search spaces* for: *quantisation* & *architectural configuration*
- Does not require access to the *training data*

Privacy-aware High-Throughput Inference

- Pushing quantization bellow limits of acceptable accuracy to gain performance (high throughput)
- Evaluation of Quality of Prediction to identify and correct error introduced by quantization



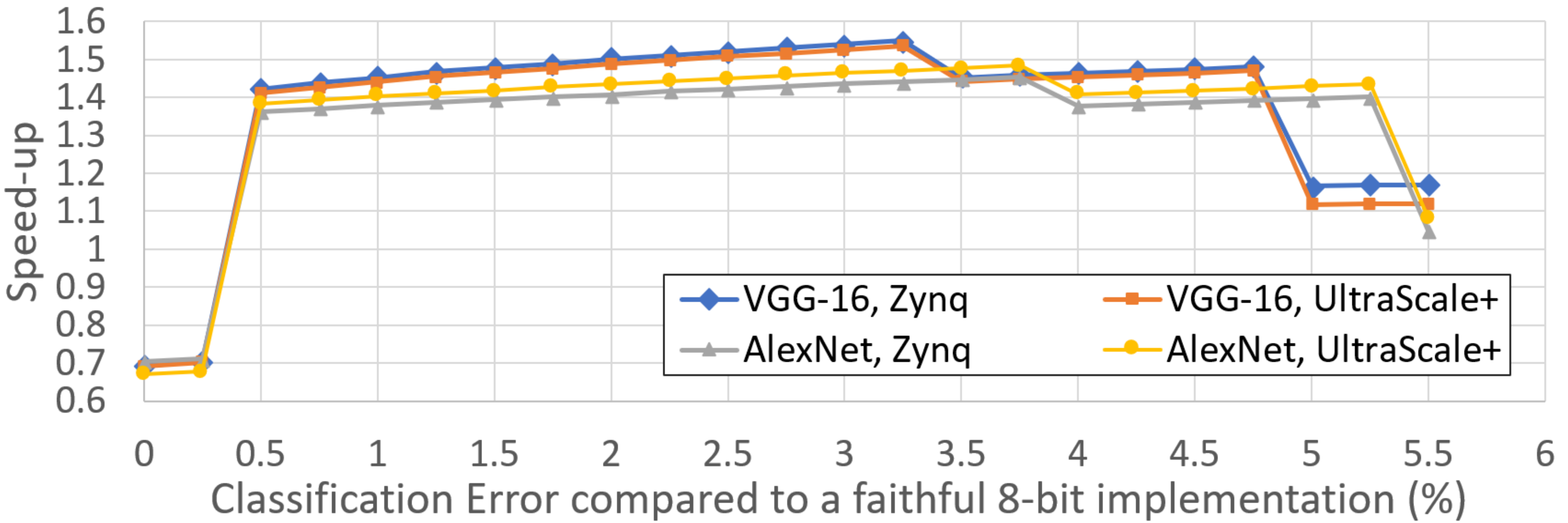


Low-Precision Unit: Degraded accuracy classification with high performance

Confidence Evaluation Unit: Identify misclassified cases

High-Precision Unit: Correct detected misclassified samples, to restore accuracy

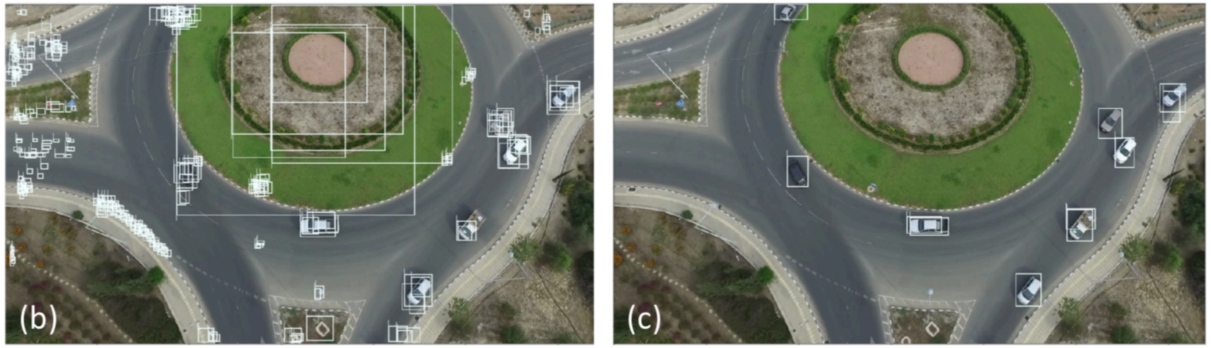
Privacy-aware High-Throughput Inference



Conclusions

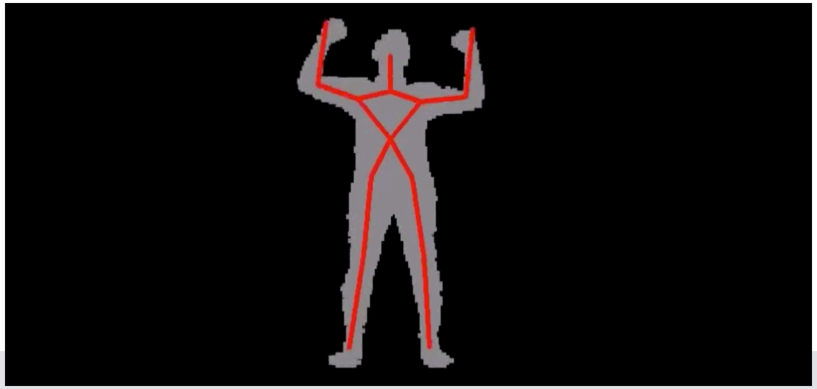
- Efficient deployment of DNNs on embedded devices requires a <u>holistic</u> approach
- Need of <u>tools</u> to help the designer to address the complexity of the design process

Traffic Detection

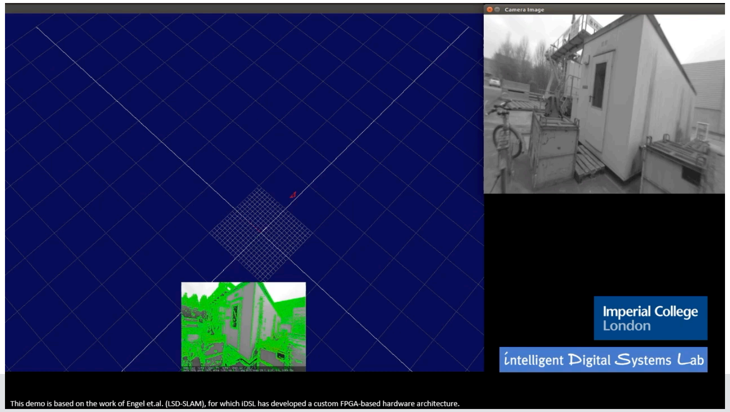


Autonomous Navigation

Pose estimation using ML



Embedded SLAM



Selected Publications

- Stylianos I. Venieris and Christos-Savvas Bouganis. 2016. ***fpgaConvNet: A Framework for Mapping Convolutional Neural Networks on FPGAs.*** *In 2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). 40–47.*
- Stylianos I. Venieris and Christos-Savvas Bouganis. 2017. ***fpgaConvNet: A Toolflow for Mapping Diverse Convolutional Neural Networks on Embedded FPGAs.*** *In NIPS 2017 Workshop on Machine Learning on the Phone and other Consumer Devices.*
- Stylianos I. Venieris and Christos-Savvas Bouganis. 2017. ***fpgaConvNet: Automated Mapping of Convolutional Neural Networks on FPGAs*** (Abstract Only). *In Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM, 291–292.*
- S. I. Venieris and C. S. Bouganis. 2017. ***Latency-Driven Design for FPGA-based Convolutional Neural Networks.*** *In 2017 27th International Conference on Field Programmable Logic and Applications (FPL).*
- Alexandros Kouris, Stylianos I. Venieris, and Christos-Savvas Bouganis. 2018. ***CascadeCNN: Pushing the performance limits of quantisation.*** *In SysML.*
- Stylianos I. Venieris, Alexandros Kouris, and Christos-Savvas Bouganis. 2018. ***Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions.*** *In ACM Computing Surveys 51, 3, Article 56 (June 2018), 39 pages.*
- Alexandros Kouris, Stylianos I. Venieris, and Christos-Savvas Bouganis. 2018. ***CascadeCNN: Pushing the Performance Limits of Quantisation in Convolutional Neural Networks***. *In 2018 28th International Conference on Field Programmable Logic and Applications (FPL).*
- S. I. Venieris and C. S. Bouganis. 2018. ***f-CNNx: A Toolflow for Mapping Multiple Convolutional Neural Networks on FPGAs.*** *In 2018 28th International Conference on Field Programmable Logic and Applications (FPL).*
- C. Kyrkou, G. Plastiras, T. Theocharides, S. I. Venieris, and C. S. Bouganis. 2018. ***DroNet: Efficient Convolutional Neural Network Detector for Real-Time UAV Applications.*** *In 2018 Design, Automation Test in Europe Conference Exhibition (DATE). 967–972.*
- Michalis Rizakis, Stylianos I. Venieris, Alexandros Kouris, and Christos-Savvas Bouganis. 2018. ***Approximate FPGA-based LSTMs under Computation Time Constraints.*** *In Applied Reconfigurable Computing - 14th International Symposium, ARC 2018, Santorini, Greece, May 2 - 4, 2018, 3–15.*
- Alexandros Kouris and Christos-Savvas Bouganis. 2018. ***Learning to Fly by MySelf: A Self-Supervised CNN-based Approach for Autonomous Navigation.*** In IEEE/RSJ International Conf. on Intelligent Robots and Systems (IROS), 2018
- Stylianos I. Venieris, Alexandros Kouris and Christos-Savvas Bouganis. 2019. ***Deploying Deep Neural Networks in the Embedded Space,*** *in* MobiSys18: 2nd International Workshop on Embedded and Mobile Deep Learning (EMDL)
- Alexandros Kouris, Stylianos I. Venieris, Michalis Rizakis, and Christos-Savvas Bouganis. 2019. ***Approximate LSTMs for Time-Constrained Inference: Enabling Fast Reaction in Self-Driving Cars*** [Under Review – available on arXiv: https://arxiv.org/pdf/1905.00689.pdf]
- Alexandros Kouris, Christos Kyrkou and Christos-Savvas Bouganis. 2019. ***Informed Region Selection for Efficient UAV-based Object Detectors: Altitude-aware Vehicle Detection with CyCAR Dataset***, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019 [to appear]