Movidius

EMBEDDED DEEP NEURAL NETWORKS

“The Cost of Everything and the Value of Nothing”

David Moloney, CTO

Hot Chips 28, Cupertino, California
August 23, 2016

AGENDA

DEEP LEARNING: THE GREAT DISRUPTOR

senseFly
a Parrot company

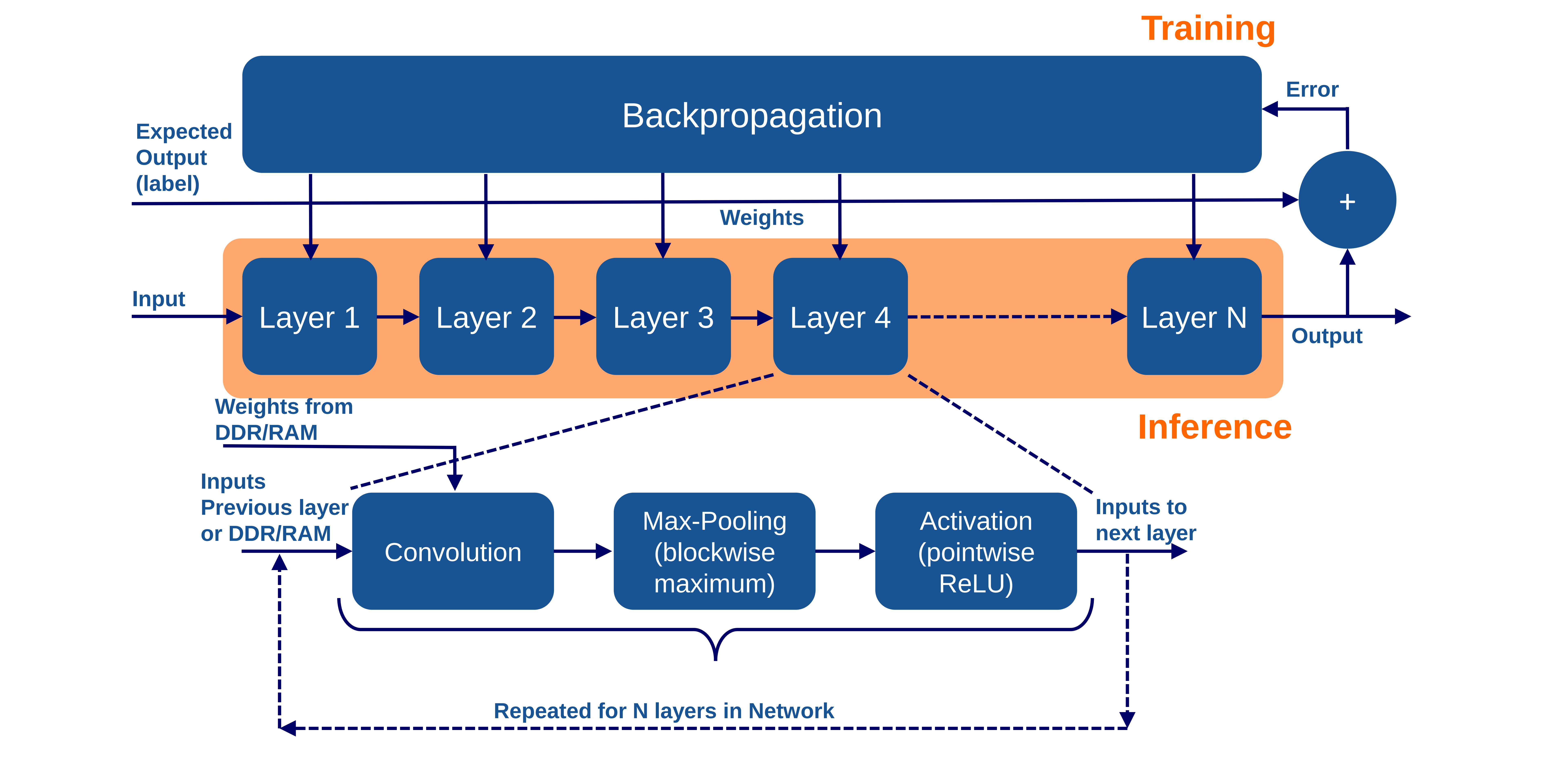
YUNEEC
ELECTRIC AVIATION

„Tesla Zooms Past BMW, Audi Limos In Europe, Closes In On Mercedes“
www.forbes.com, Oct 19th 2015

AGENDA

DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs)

Inference Matters in Embedded

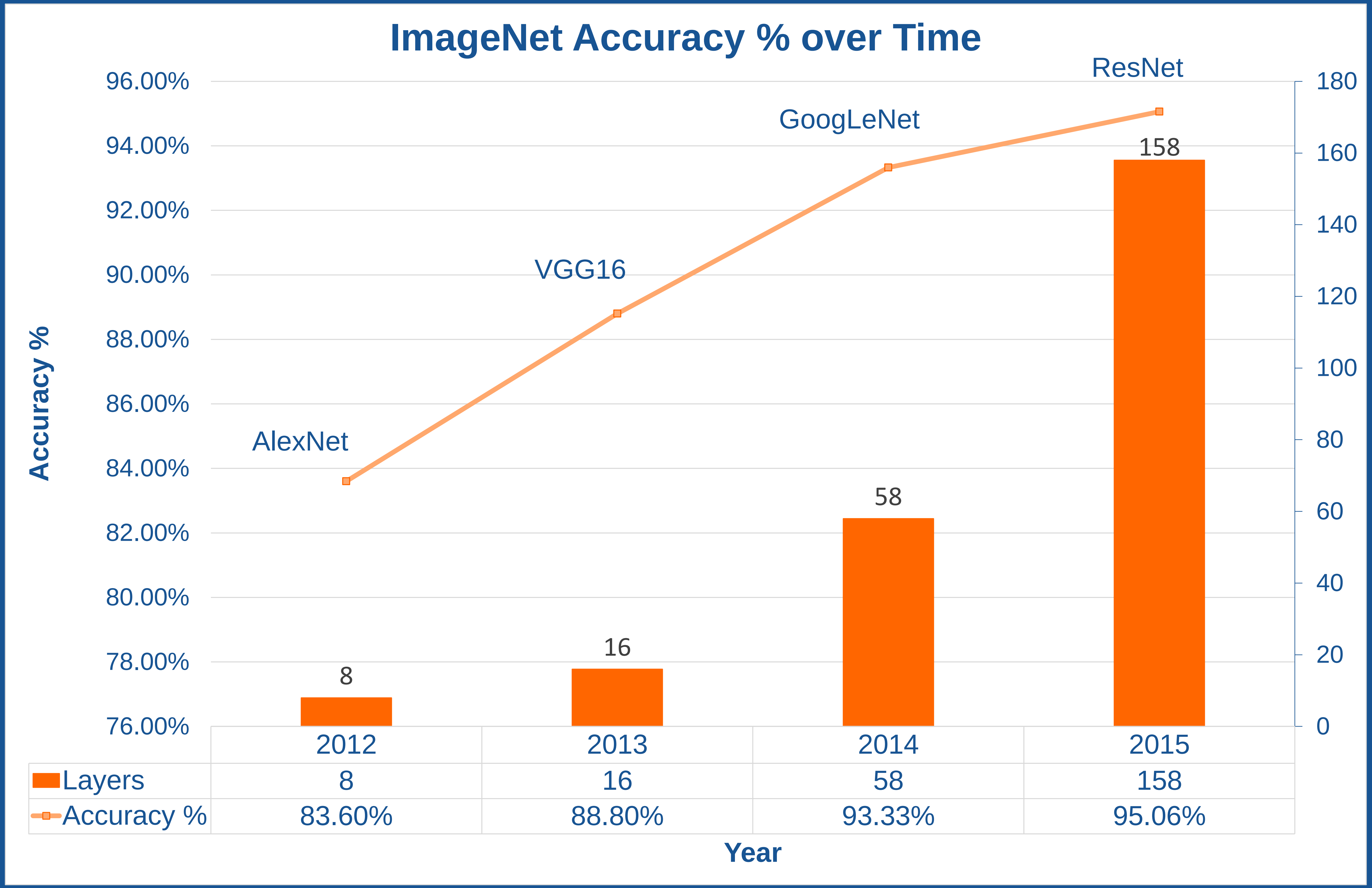


AGENDA

- DEEP LEARNING: THE GREAT DISRUPTOR
- TRAINING VS INFERENCE: INFERENCE MATTERS IN EMBEDDED
- THE DEEPER THE BETTER?
- BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE
- MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50
- OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS
- CONCLUSIONS

THE DEEPER THE BETTER…

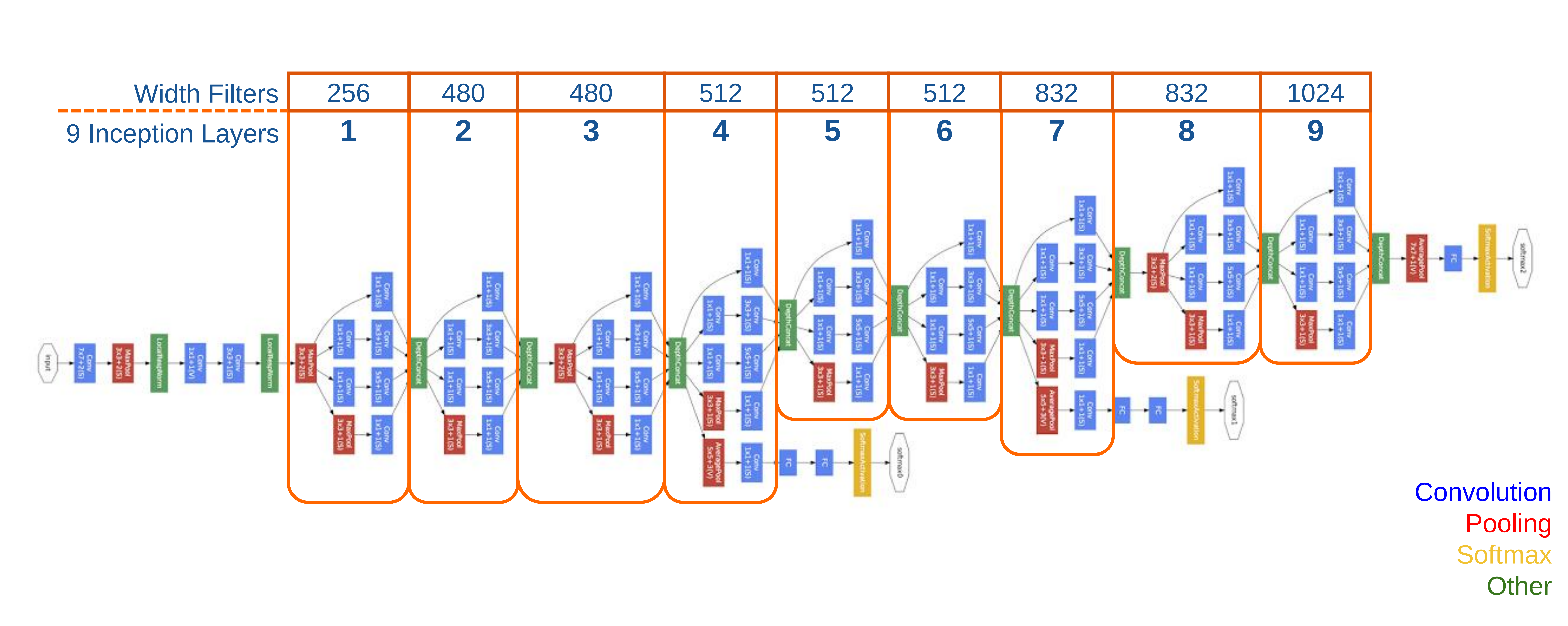
Yet With Increased Complexity



	2012	2013	2014	2015
Layers	8	16	58	158
Accuracy %	83.60%	88.80%	93.33%	95.06%

COMPLEXITY OF A STANDARD NETWORK

GoogLeNet



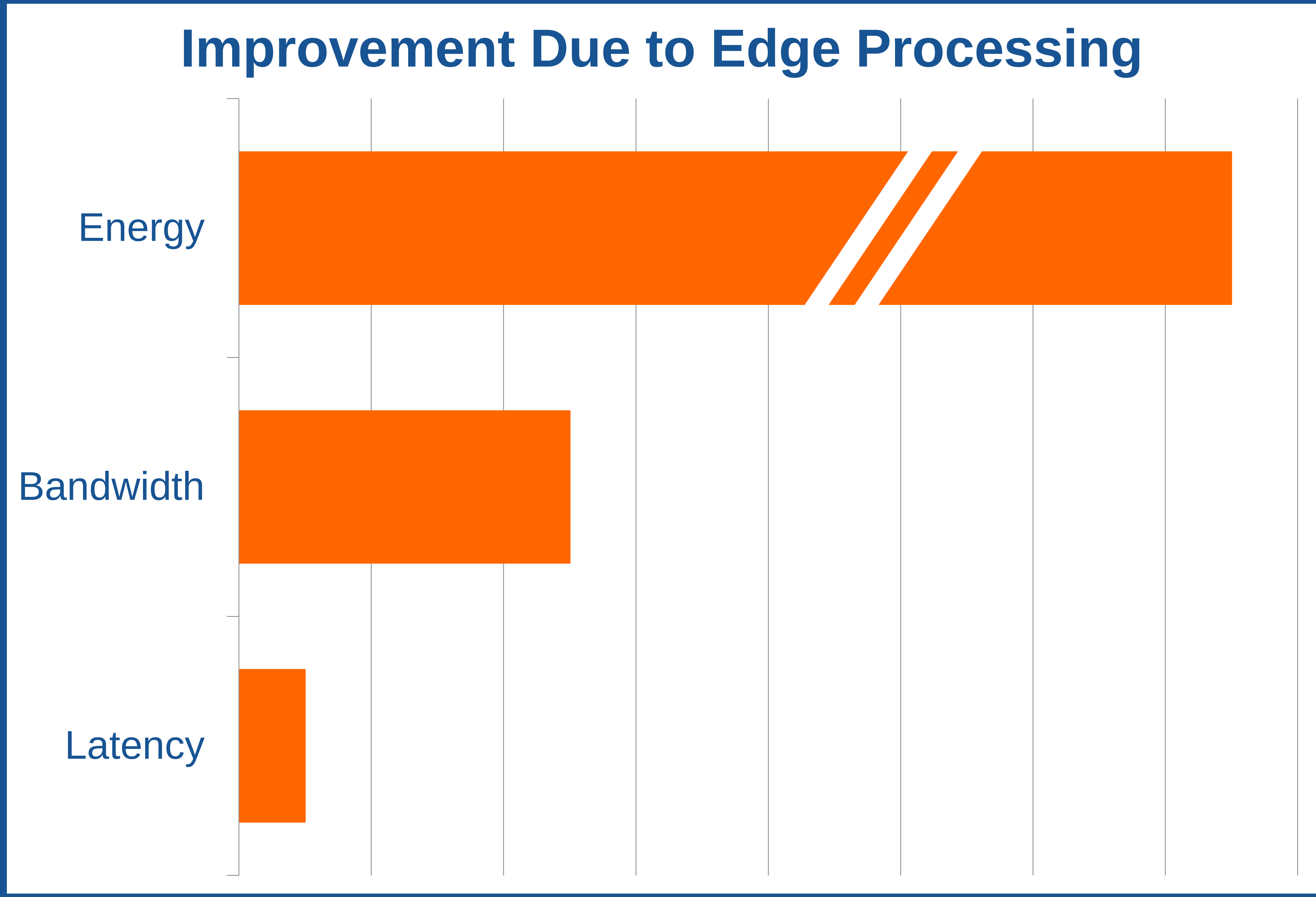
- Computational cost is increased by less than 2x compared to AlexNet (<1.5 Bn operations/ evaluation)
- 5M parameters

AGENDA

- DEEP LEARNING: THE GREAT DISRUPTOR
- TRAINING VS INFERENCE: INFERENCE MATTERS IN EMBEDDED
- THE DEEPER THE BETTER?
- **BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE**
- MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50
- OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS
- CONCLUSIONS

EMBEDDED PROCESSING AT THE NETWORK EDGE

Huge Benefits



- **1,000,000 x**
  more energy-efficient

- **10,000 x**
  less bandwidth consumed

- **1,000 x**
  lower latency

Other Improvements in:

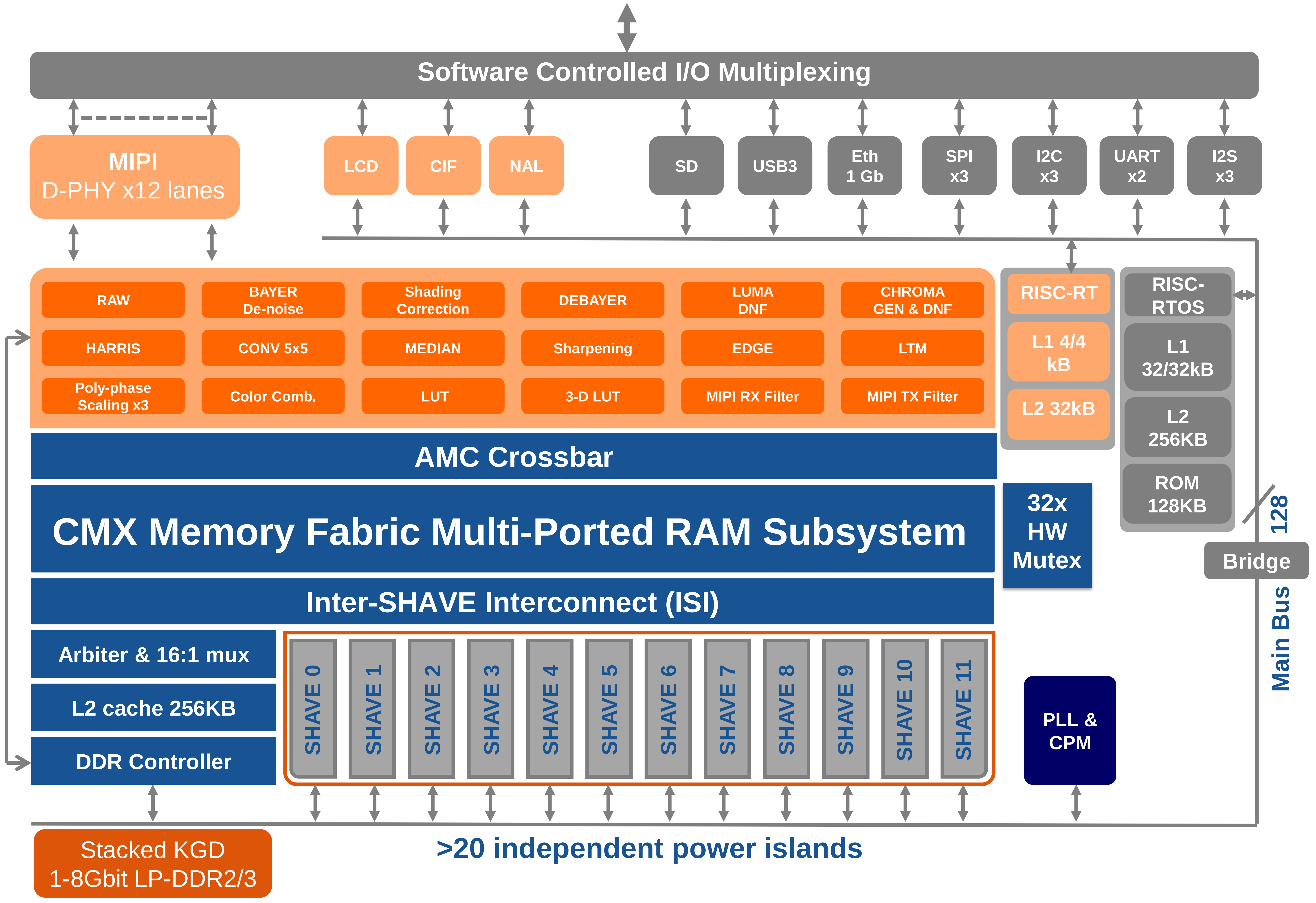
- Privacy
- Fault-tolerance/ Continuity of service

AGENDA

- DEEP LEARNING: THE GREAT DISRUPTOR
- TRAINING VS INFERENCE: INFERENCE MATTERS IN EMBEDDED
- THE DEEPER THE BETTER?
- BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE
- **MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50**
- OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS
- CONCLUSIONS

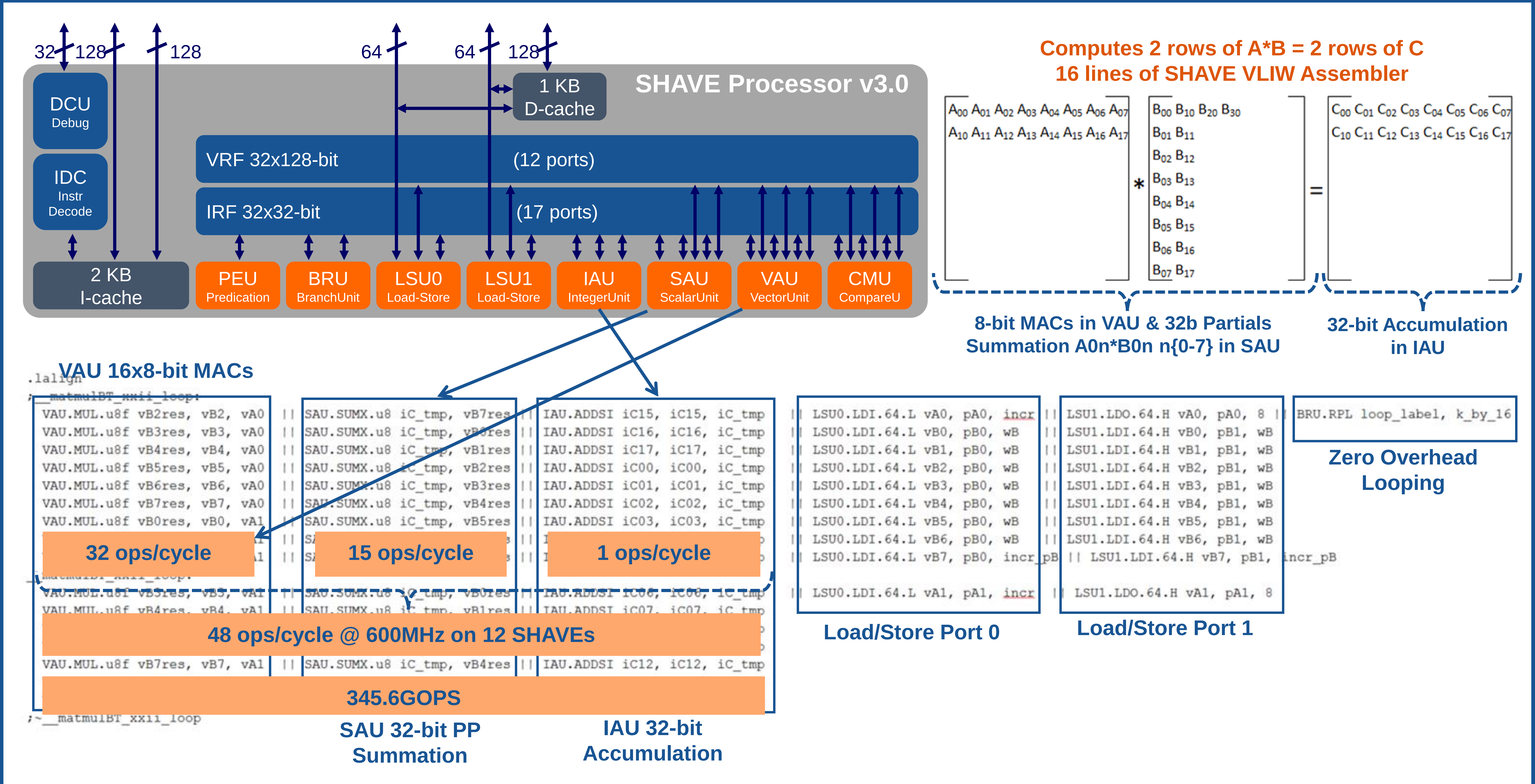
MYRIAD 2 MA2x50

Vision Processing Unit (VPU) Architecture



SHAVE PROCESSING

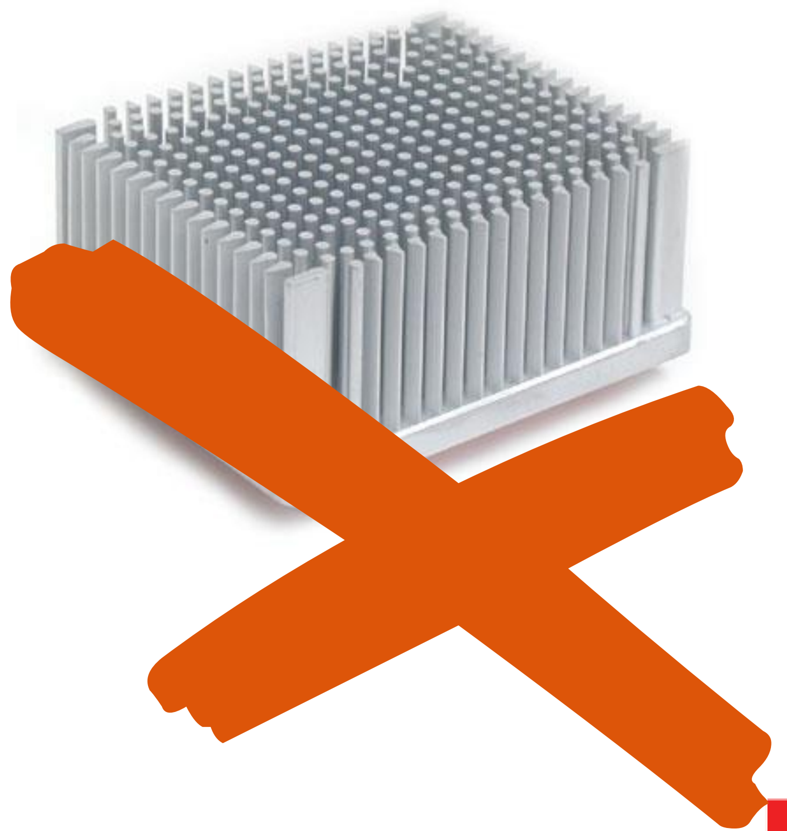
Maximising GEMM Performance



GoogLeNet RELATIVE GFLOPS/W

Performance Results

GoogLeNet Single Inference (Batch = 1) no Heatsink or Fan					
	nm Process	fps	Power (W)	GFLOPS	GFLOPS/W
GPU-A	28	15	7.5	51.37	6.85
GPU-B	20	22	7.5	75.34	10.04
Myriad 2	28	25	1.2	85.61	71.34



AGENDA

HOW MOVIDIUS IS DEPLOYING STANDARD CNNs

At the Network Edge

Caffe

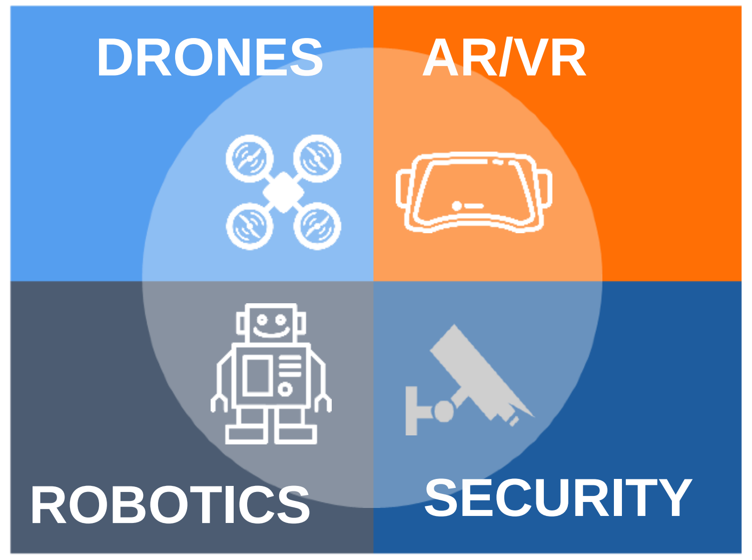
TensorFlow

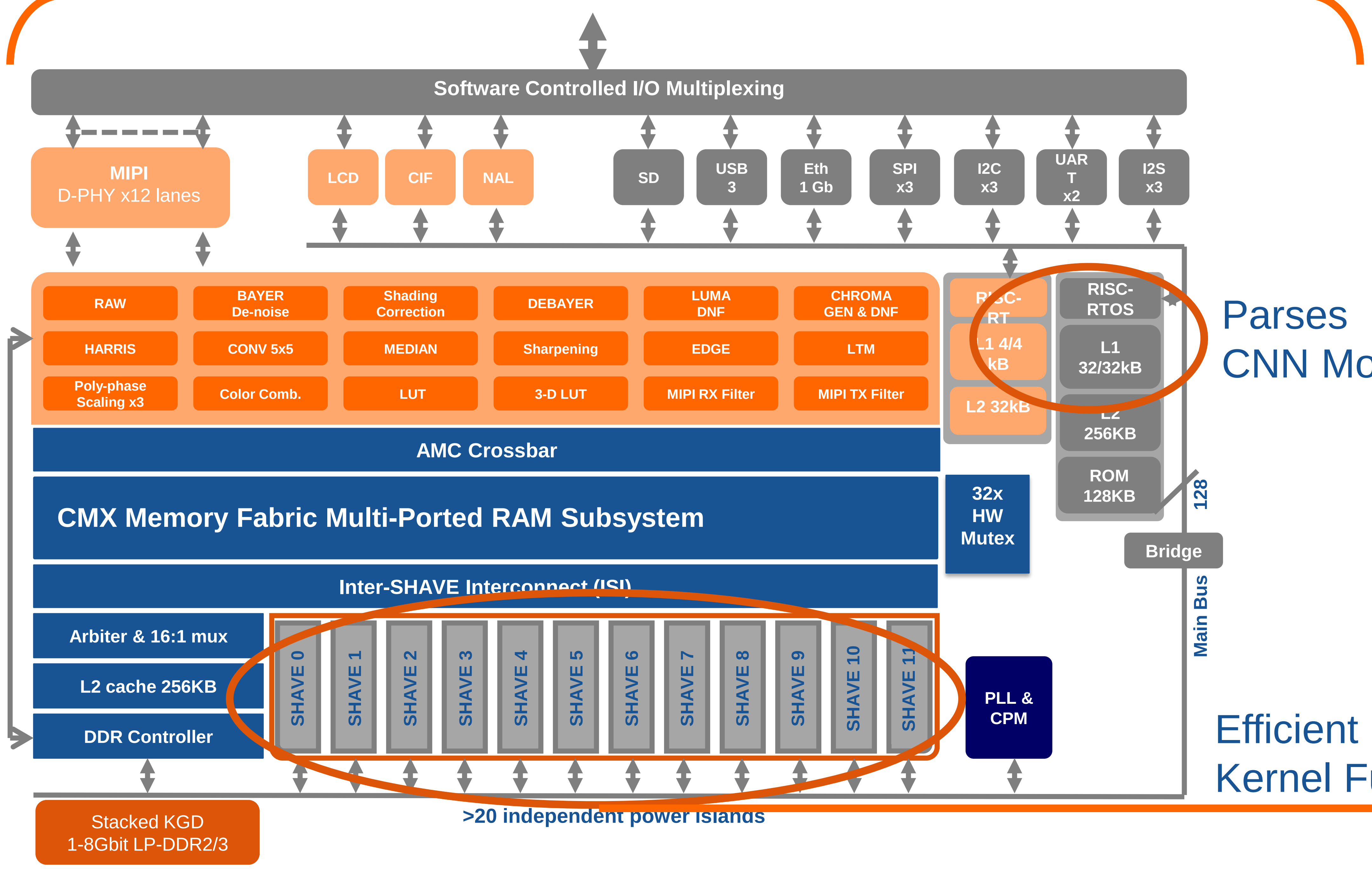
CNN Model Description

Weights

Fathom

For Integration





Parses CNN Model

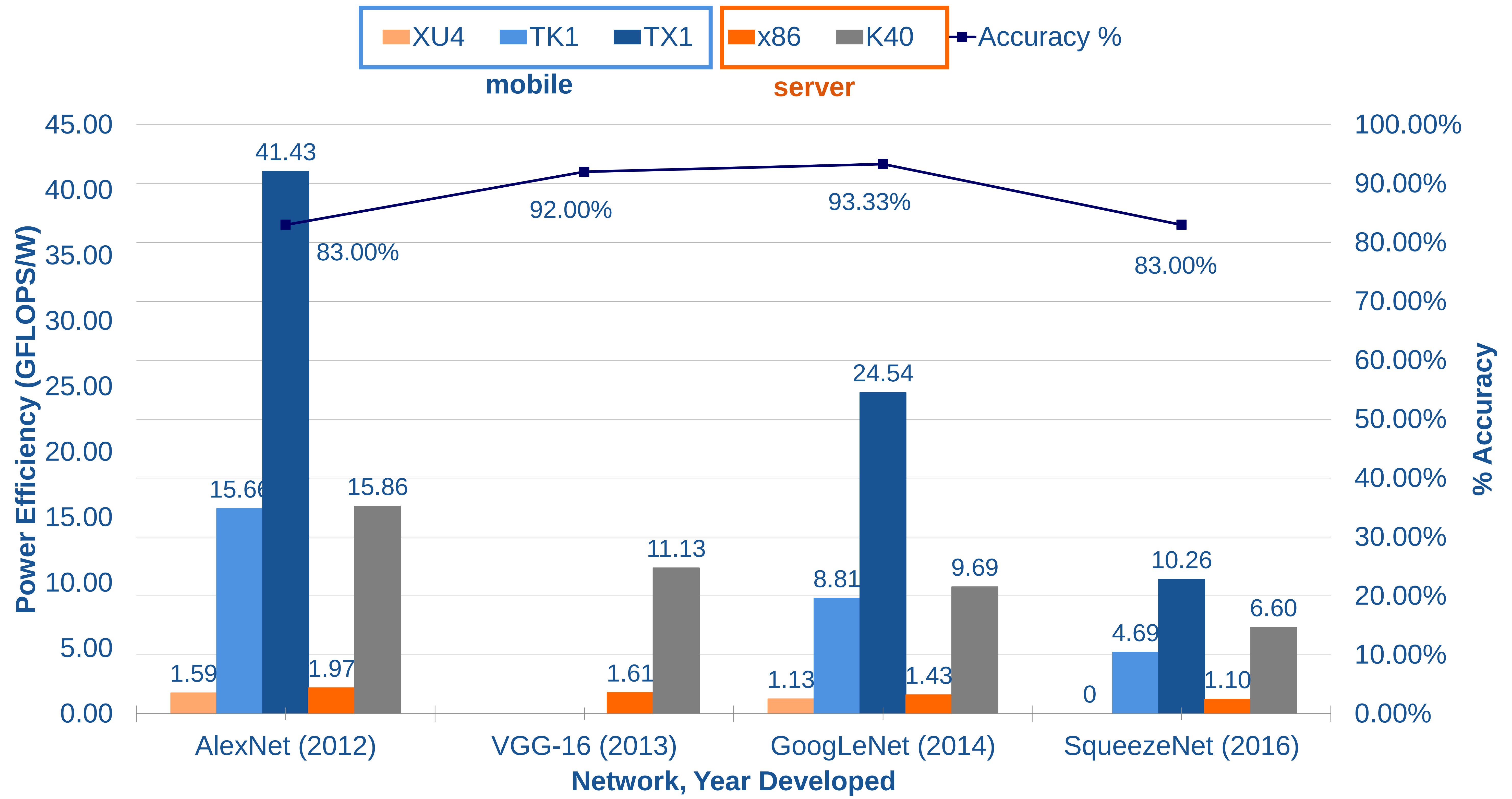
Efficient Execution of Kernel Function Libraries

AGENDA

- DEEP LEARNING: THE GREAT DISRUPTOR
- TRAINING VS INFERENCE: INFERENCE MATTERS IN EMBEDDED
- THE DEEPER THE BETTER?
- BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE
- MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50
- OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS
- **CONCLUSIONS**

IS IT WORTH IT?

Incremental Accuracy and the Power Efficiency Cost



Notes: ImageNet, Batch = 10/64, using active cooling

SUMMARY

Achievements and Future Challenges

- Deep Learning for Embedded is all about **Inference**
- Standard Networks are designed to achieve **high-accuracy**
- Embedded implementation on architectures such as Movidius VPU can achieve **significant performance results** at the network edge
- Next challenge is to further optimise networks to **maximise performance per Watt**

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°611183 (EXCESS Project, www.excess-project.eu).

EXCESS

SEVENTH FRAMEWORK PROGRAMME

Movidius

HOT CHIPS


Movidius
THANK YOU
FOR YOUR ATTENTION