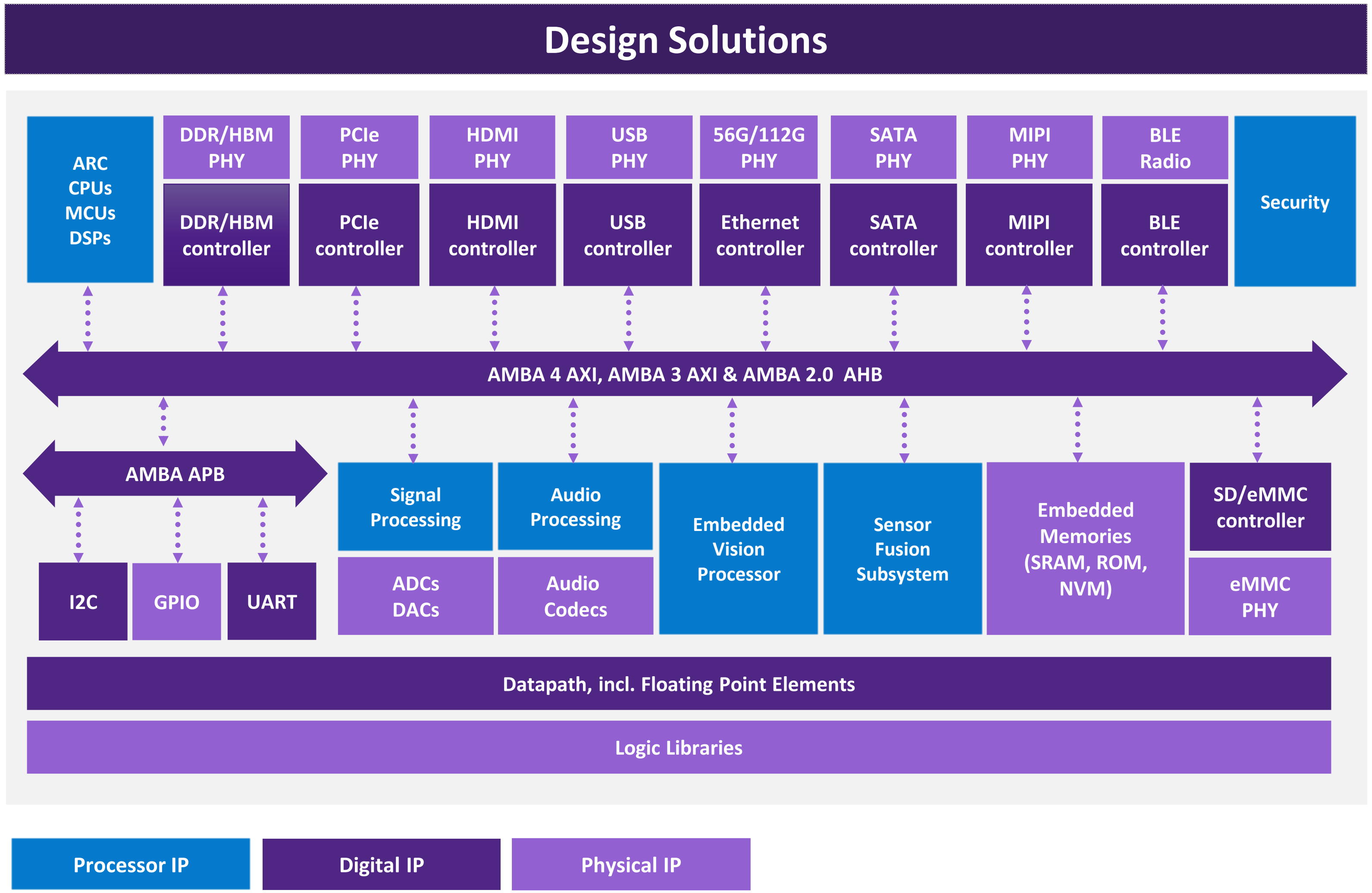
2020 embedded vision summit®

Trends in Neural Network Topologies for Vision at the Edge

Pierre Paulin
Director of R&D
September 2020

SYNOPSYS®

About Synopsys DesignWare IP



Design Solutions

ARC CPUs MCUs DSPs

DDR/HBM PHY – DDR/HBM controller

PCIe PHY – PCIe controller

HDMI PHY – HDMI controller

USB PHY – USB controller

56G/112G PHY – Ethernet controller

SATA PHY – SATA controller

MIPI PHY – MIPI controller

BLE Radio – BLE controller

Security

AMBA 4 AXI, AMBA 3 AXI & AMBA 2.0 AHB

AMBA APB – I2C, GPIO, UART

Signal Processing – ADCs DACs

Audio Processing – Audio Codecs

Embedded Vision Processor

Sensor Fusion Subsystem

Embedded Memories (SRAM, ROM, NVM)

SD/eMMC controller – eMMC PHY

Datapath, incl. Floating Point Elements

Logic Libraries

Processor IP | Digital IP | Physical IP

Growing IP Business
$700M+ in Revenue;
@ Double Digit CAGR

Broadest Portfolio
Interfaces, Analog,
Foundation IP, NVM,
Processors and More

Committed to Your Success
4200+ IP Engineers Worldwide
Dedicated to Quality

Agenda

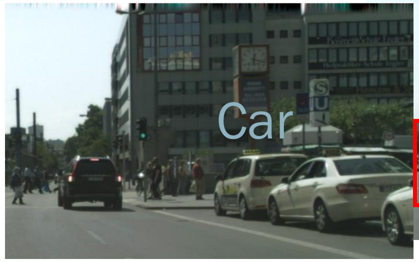
Trends in Machine Learning for Edge Applications

Key Challenges and Opportunities

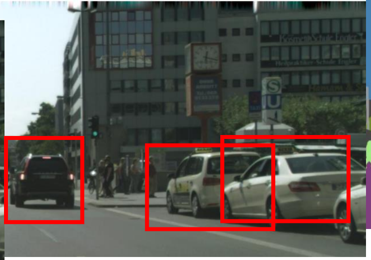
Bandwidth Optimization

The Emergence of AI-based Techniques for Embedded Systems

- Image/video
  - Classification, detection, Segmentation
    - For surveillance, AR/VR, automotive
  - Super resolution, Denoiser
    - Computational photography, MFP, DTV
  - Mostly based on Convolution Neural Networks (CNN)
- Audio, Natural Language Processing
  - Speech, Text Processing
    - Recurrent Neural Networks / LSTM
    - 1D convolutions
    - Transformers
  - Audio scene classification
    - Convolution LSTMs

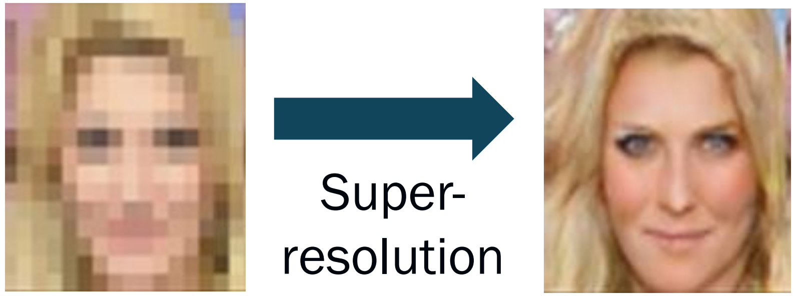


Classification

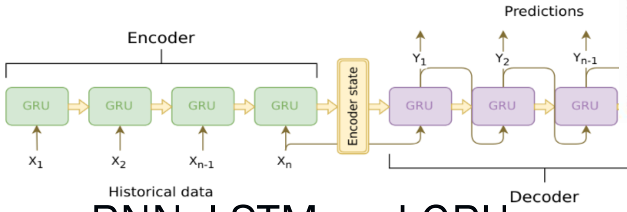


Detection

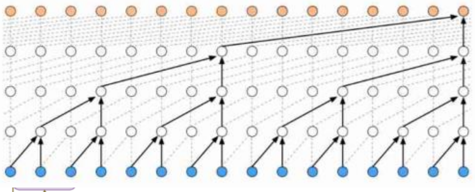
Segmentation



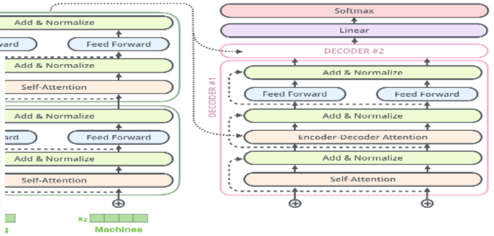
Super-resolution



RNN: LSTM and GRU



1D Convolutions

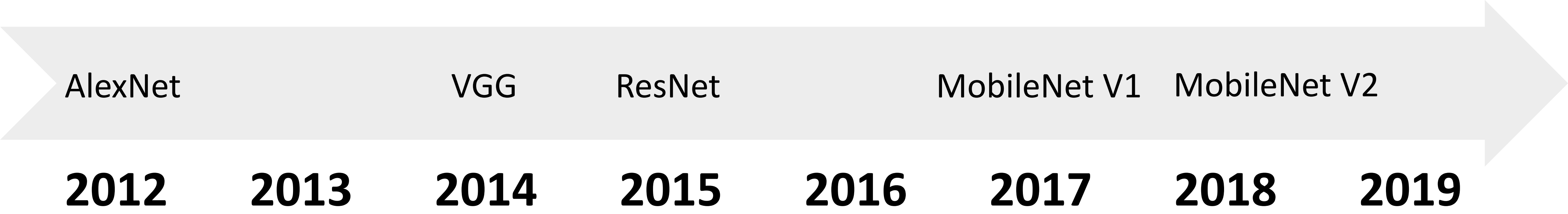


Transformers

Convolutional Neural Networks Evolving Rapidly

Classification

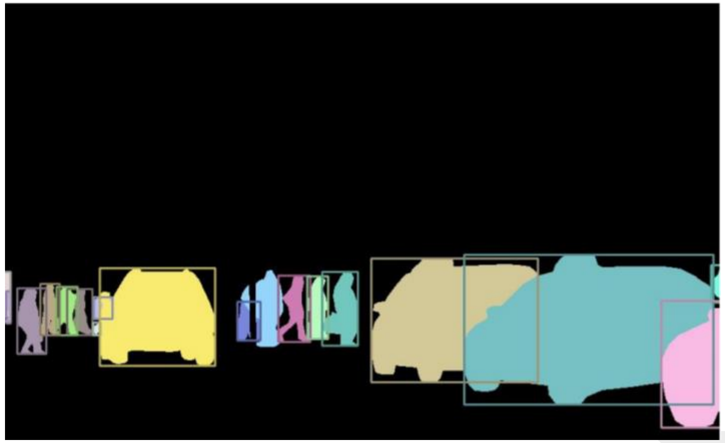
street



2012	2013	2014	2015	2016	2017	2018	2019
AlexNet		VGG	ResNet		MobileNet V1	MobileNet V2	

Convolutional Neural Networks Evolving Rapidly

Object Detection / instance segmentation



RCNN

FRCNN

SSD

YoloV2

Mask RCNN

YoloV3

AlexNet

VGG

ResNet

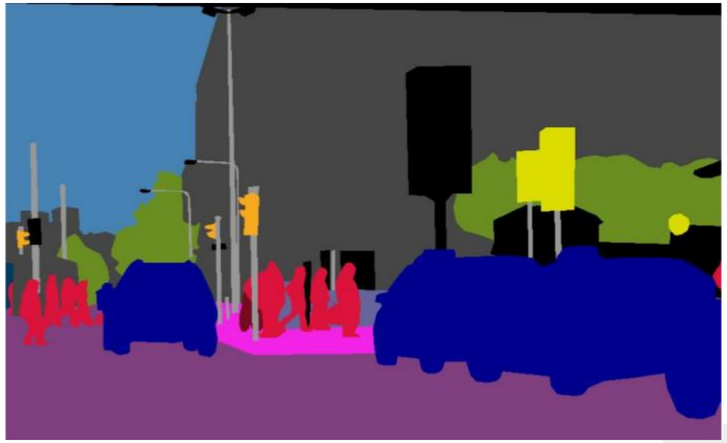
MobileNet V1

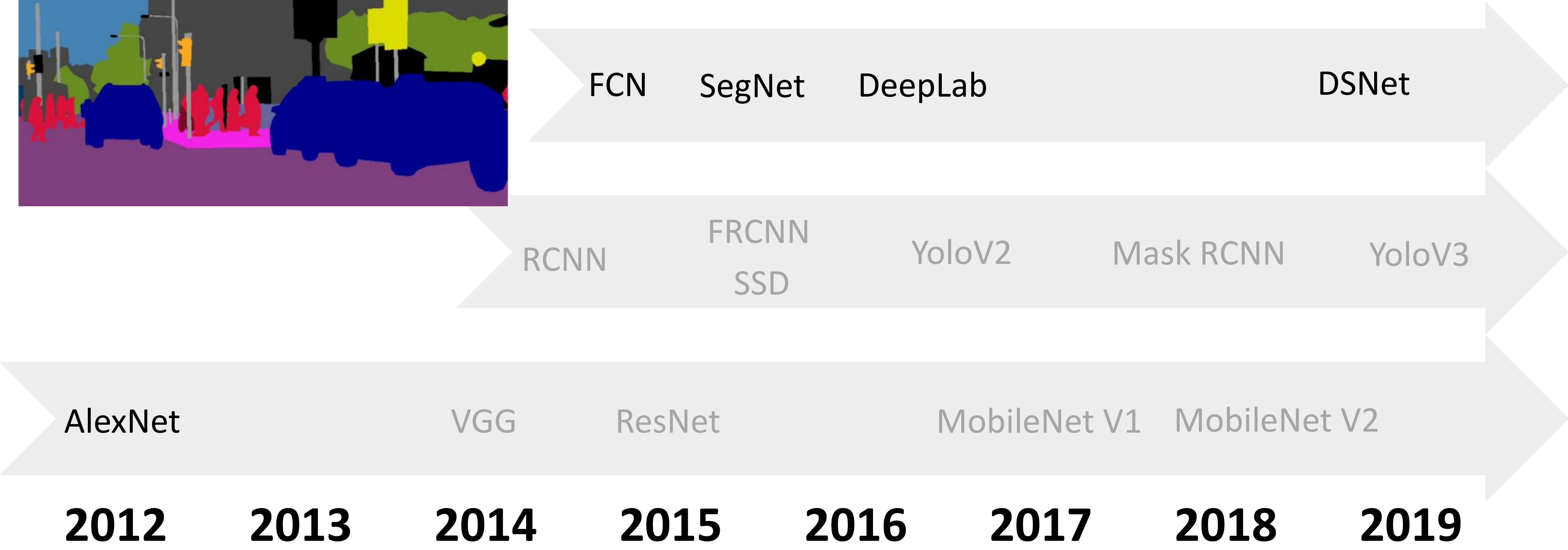
MobileNet V2

2012 **2013** **2014** **2015** **2016** **2017** **2018** **2019**

Convolutional Neural Networks Evolving Rapidly

Scene segmentation





Convolutional Neural Networks Evolving Rapidly

Panoptic Vision

PFPN
DeeperLab

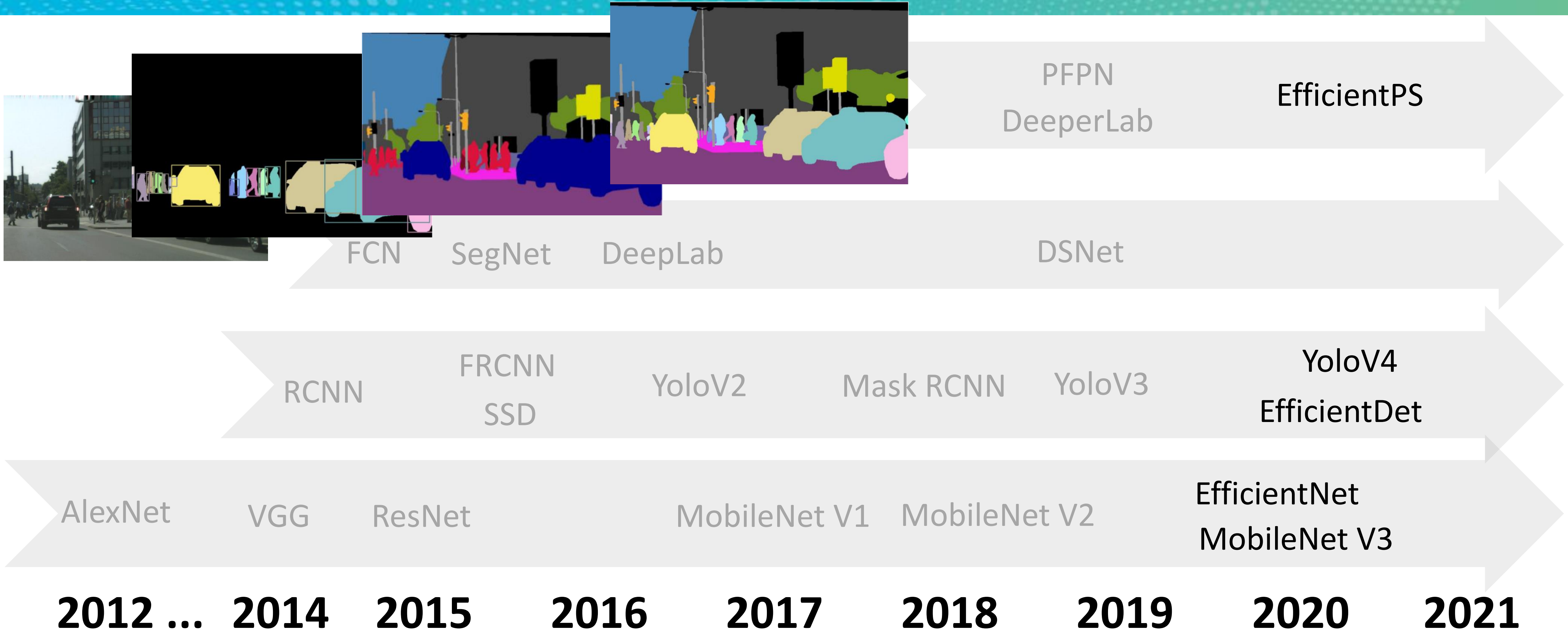
FCN SegNet DeepLab DSNet

RCNN FRCNN SSD YoloV2 Mask RCNN YoloV3

AlexNet VGG ResNet MobileNet V1 MobileNet V2

2012 2013 2014 2015 2016 2017 2018 2019

Convolutional Neural Networks Evolving Rapidly – 2020+



PFPN
DeeperLab
EfficientPS

FCN SegNet DeepLab DSNet

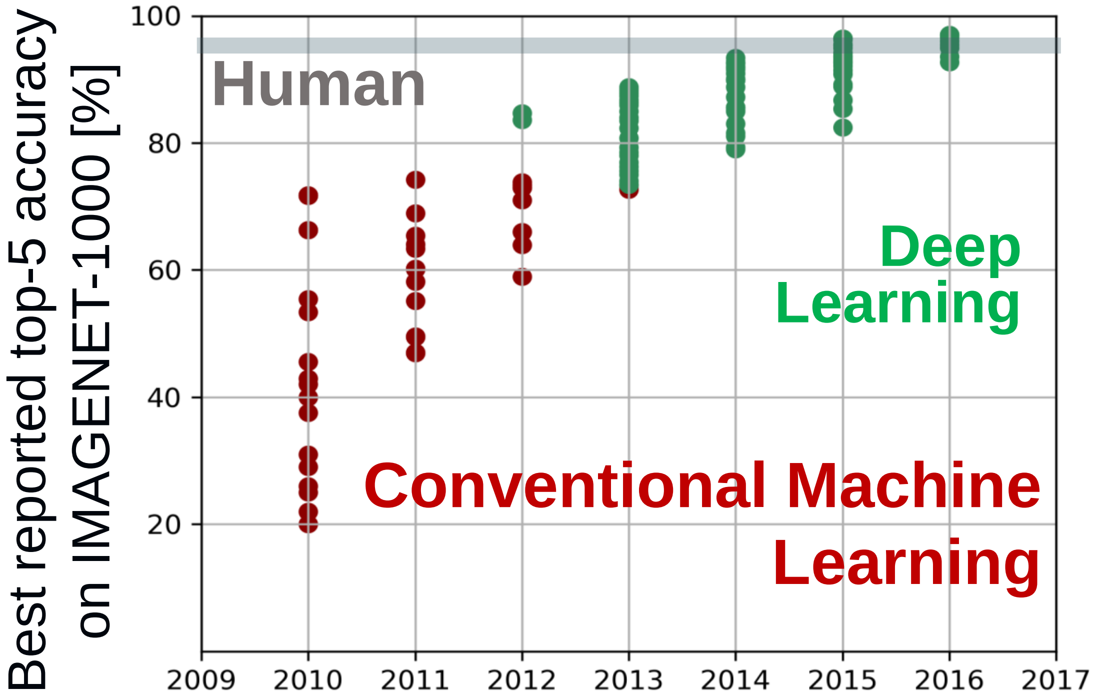
RCNN FRCNN SSD YoloV2 Mask RCNN YoloV3 YoloV4 EfficientDet

AlexNet VGG ResNet MobileNet V1 MobileNet V2 EfficientNet MobileNet V3

2012 … 2014 2015 2016 2017 2018 2019 2020 2021

CNN Accuracy Comes at a Cost

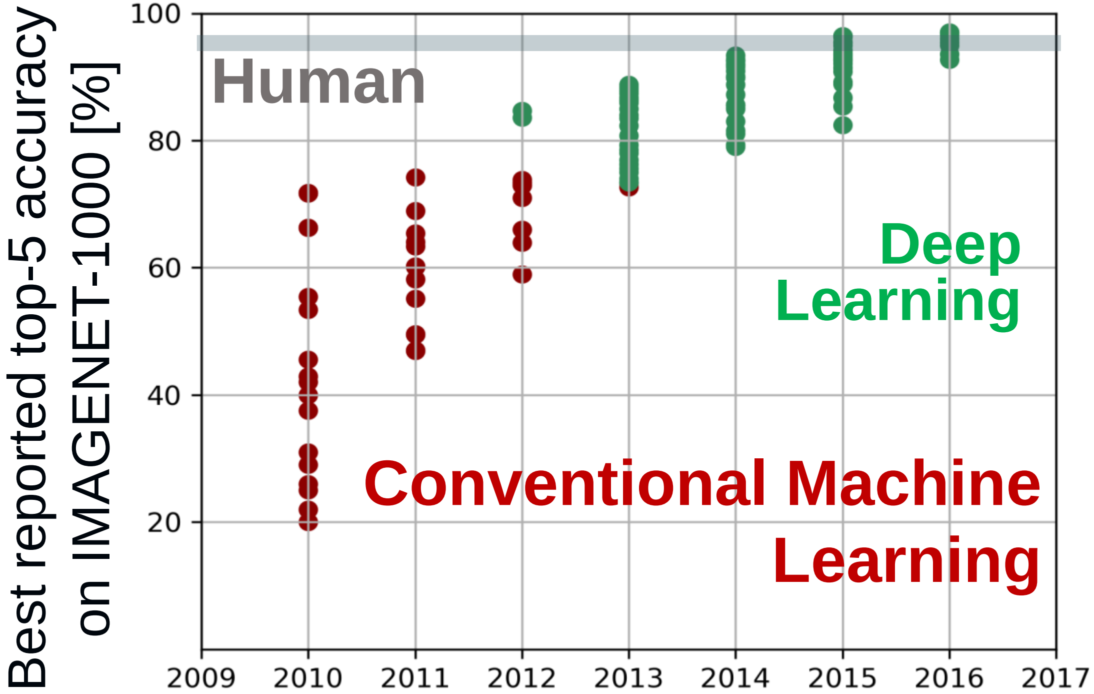
Neural network accuracy comes at a **cost** of a high workload per input pixel, large model sizes and huge bandwidth requirements

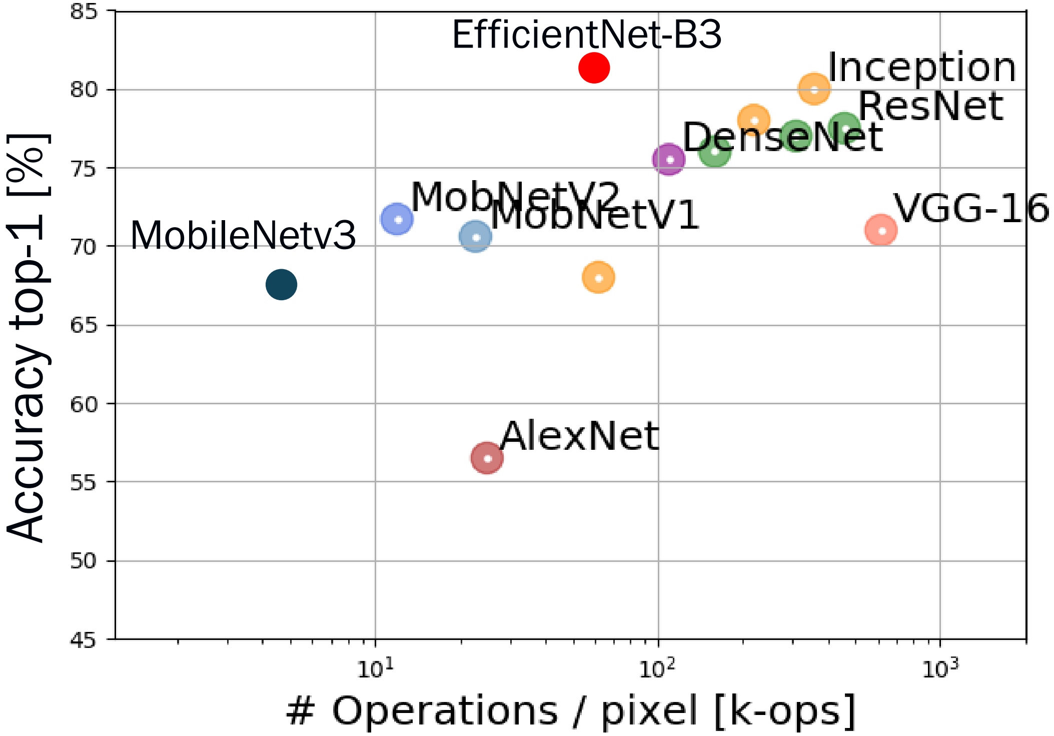


© 2020 Synopsys

CNN Accuracy Comes at a Cost

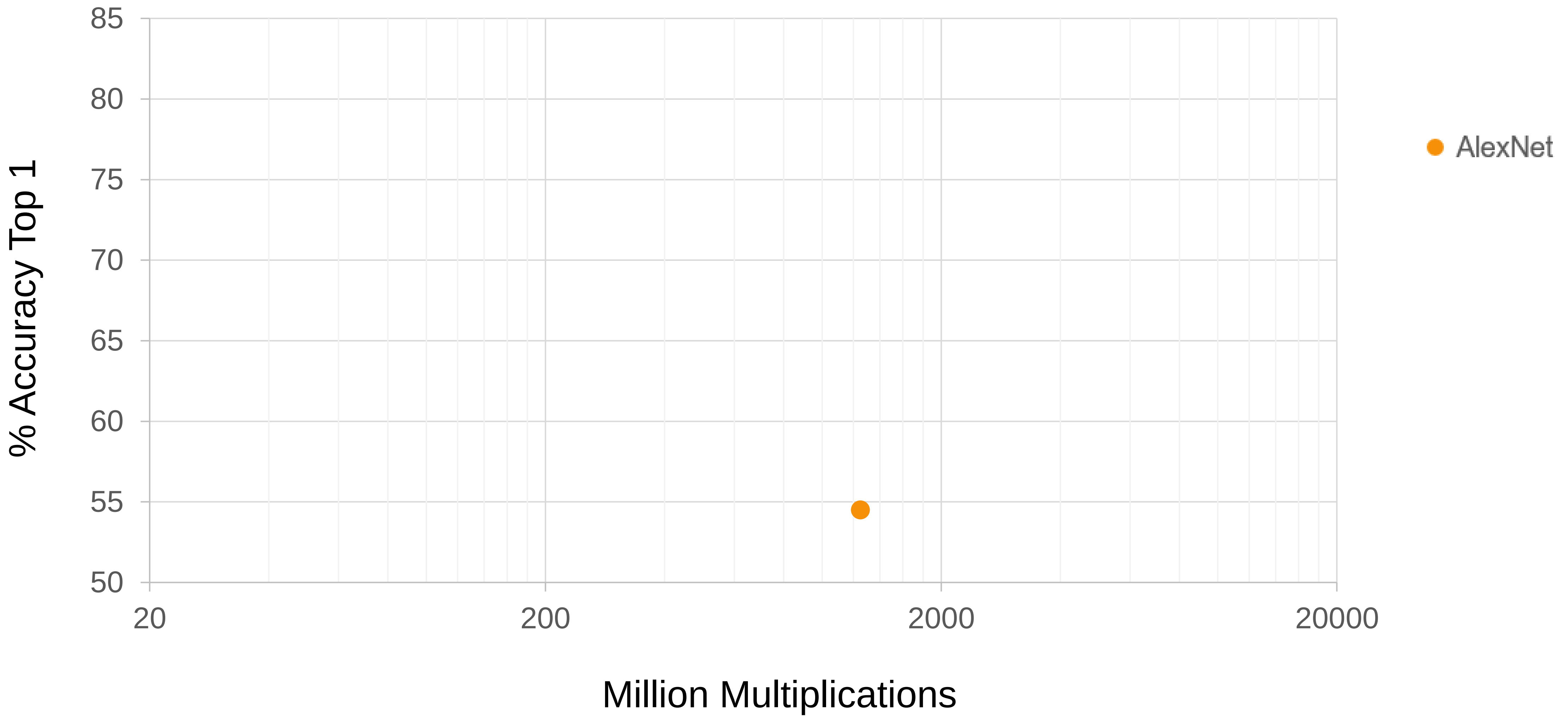
Neural network **accuracy** comes at a **cost** of a high workload per input pixel, large model sizes and huge bandwidth requirements





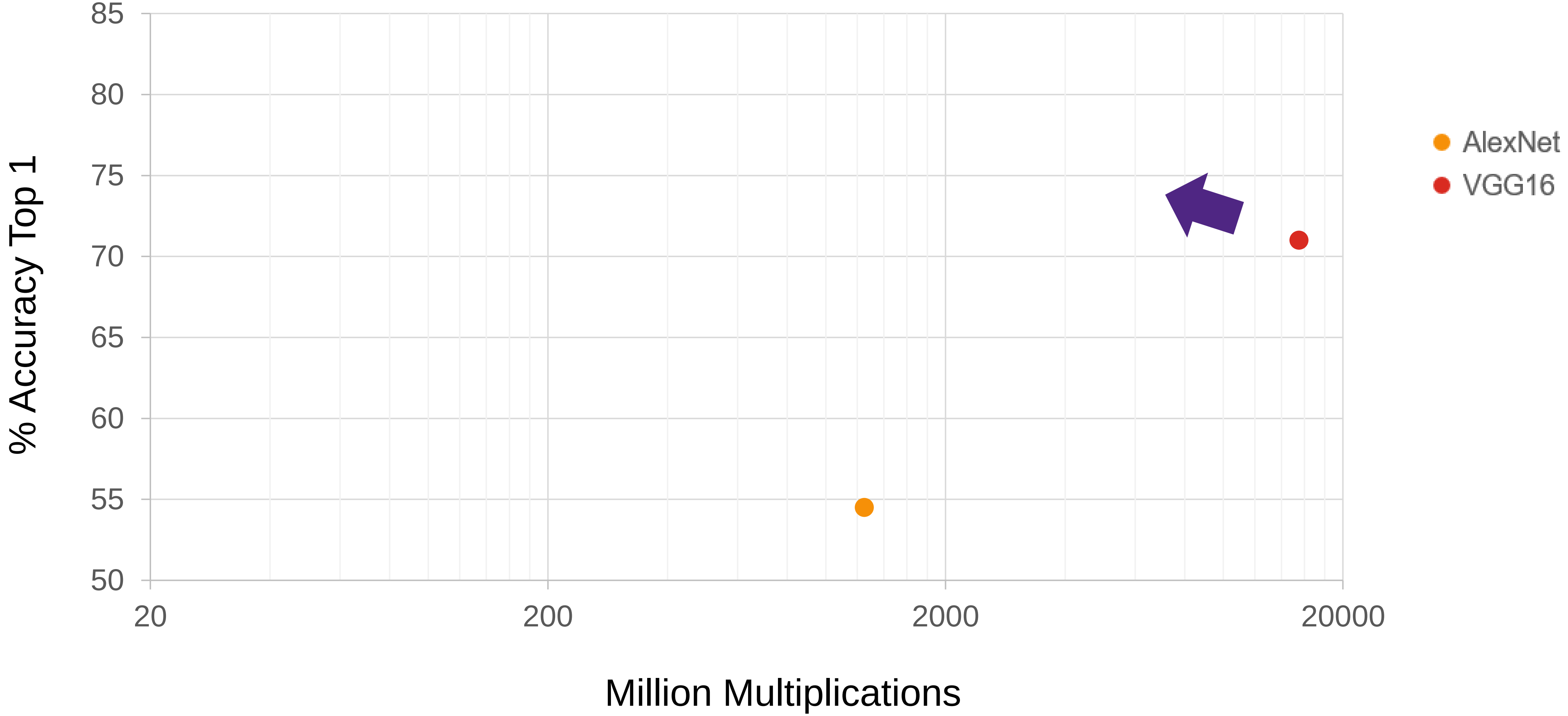
Trend 1: Reduced Compute Requirements

2012



Trend 1: Reduced Compute Requirements

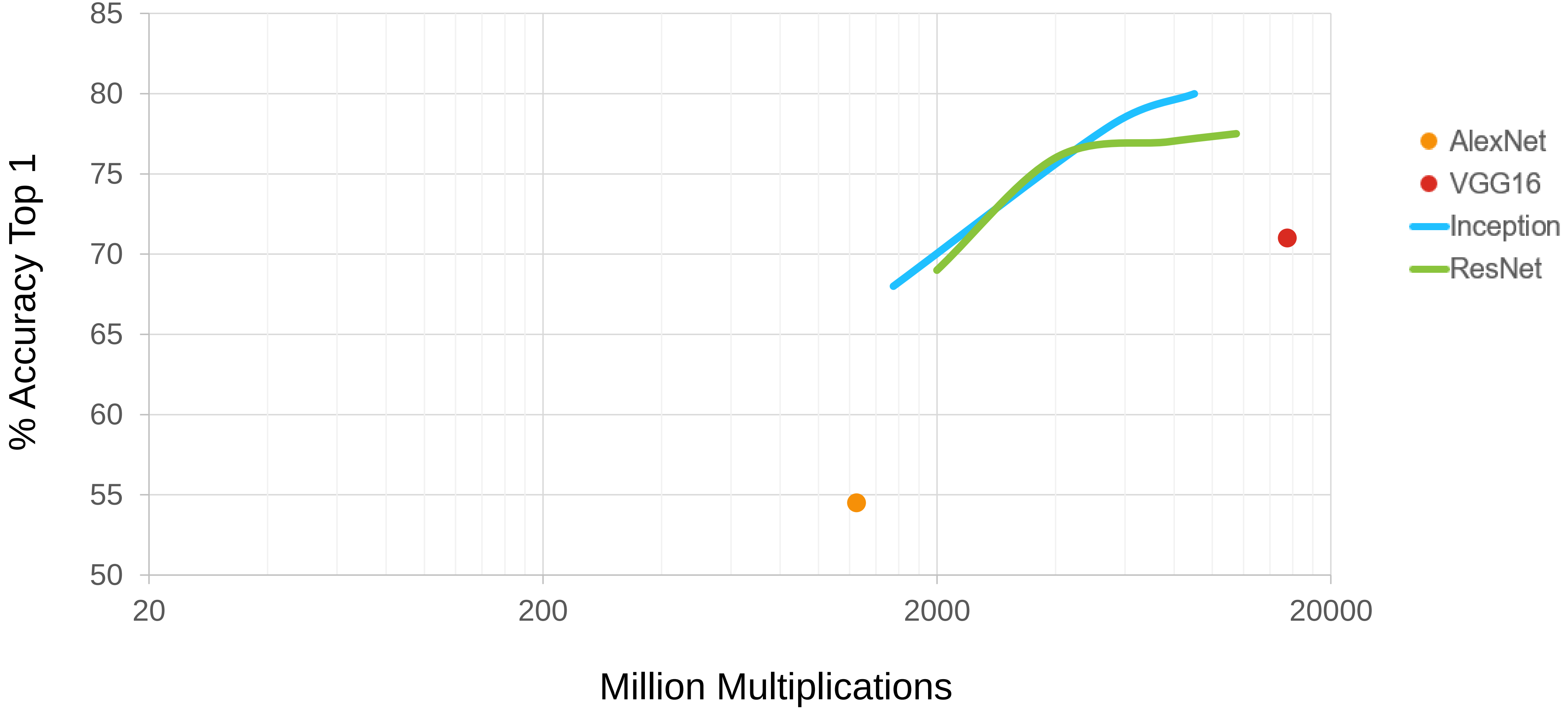
2014



Trend 1: Reduced Compute Requirements

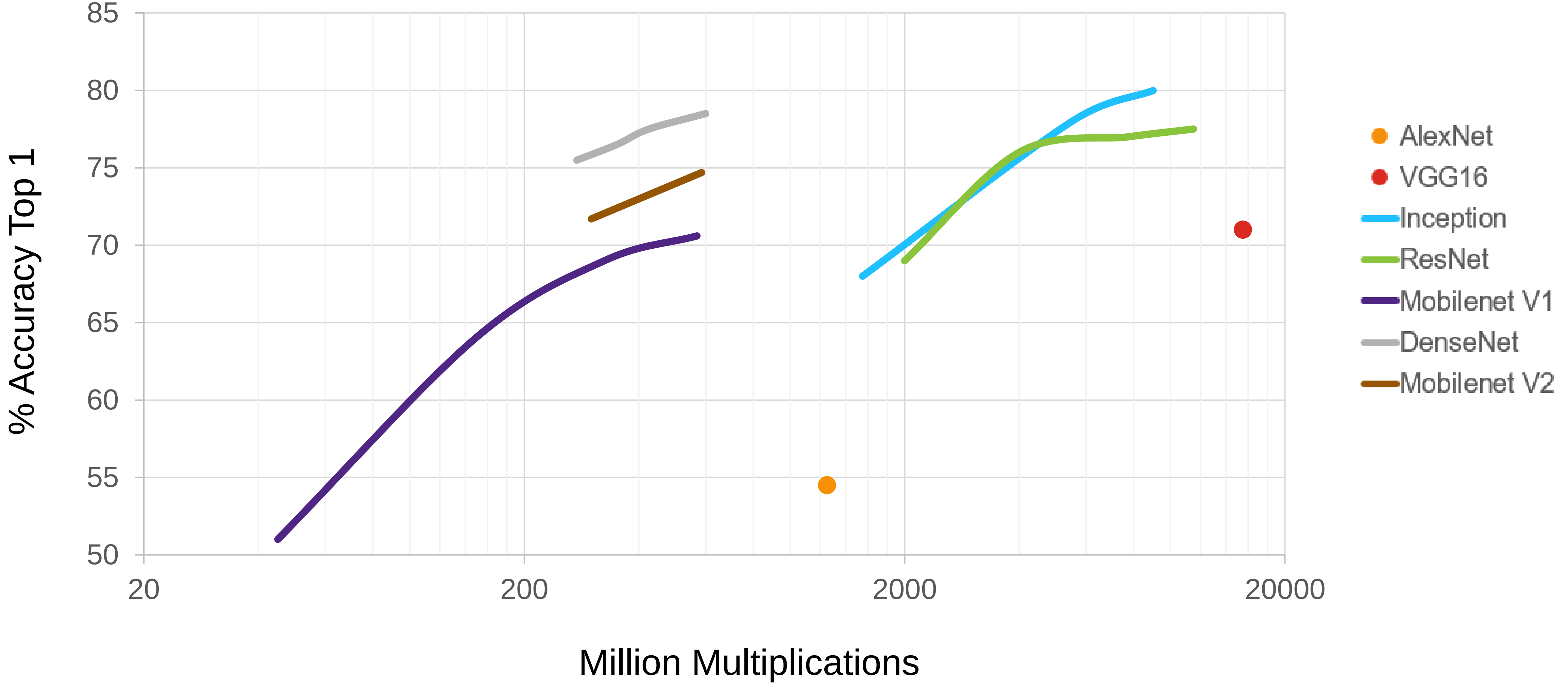
2016

2020 embedded vision summit



Trend 1: Reduced Compute Requirements

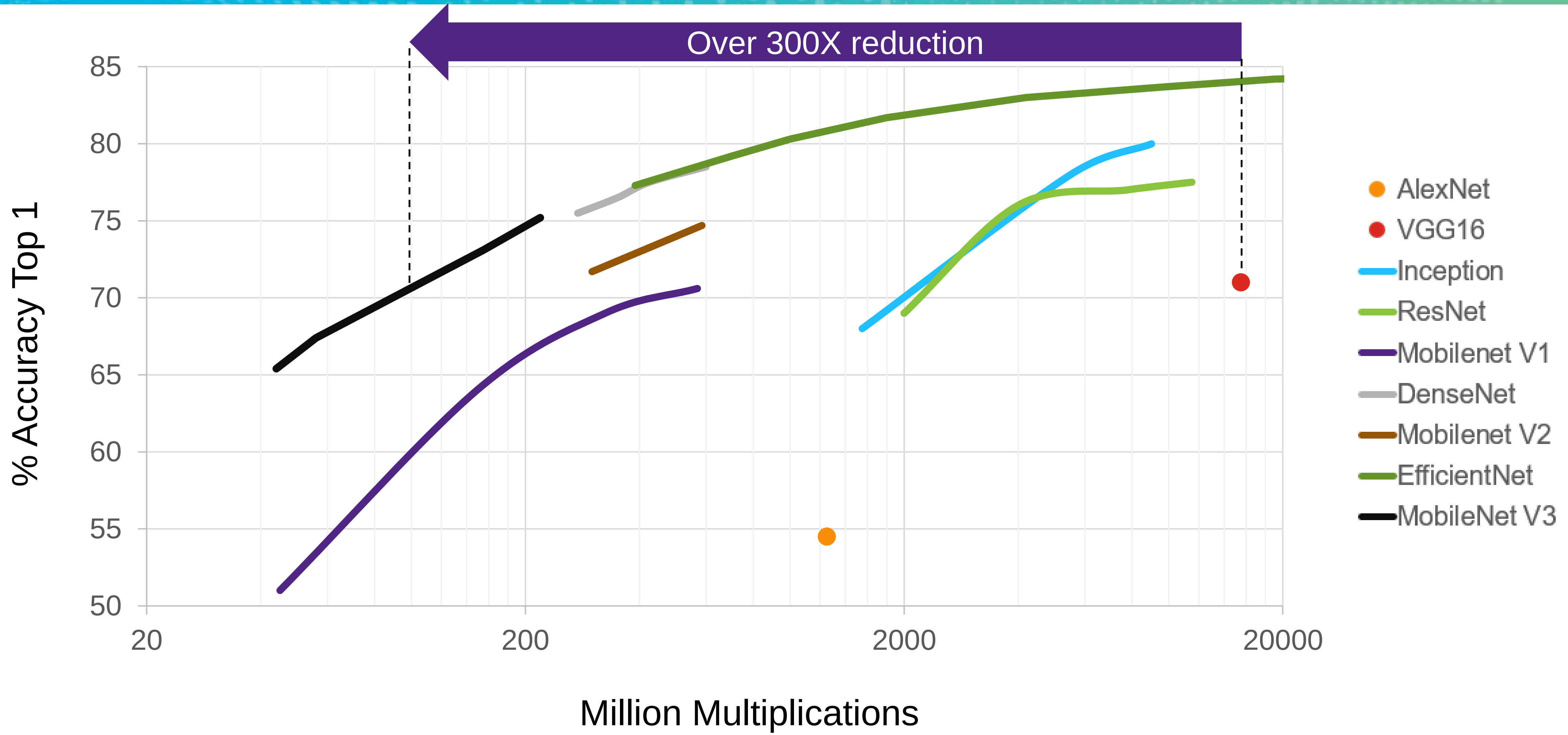
2018



Trend 1: Reduced Compute Requirements

2019

2020 embedded vision summit

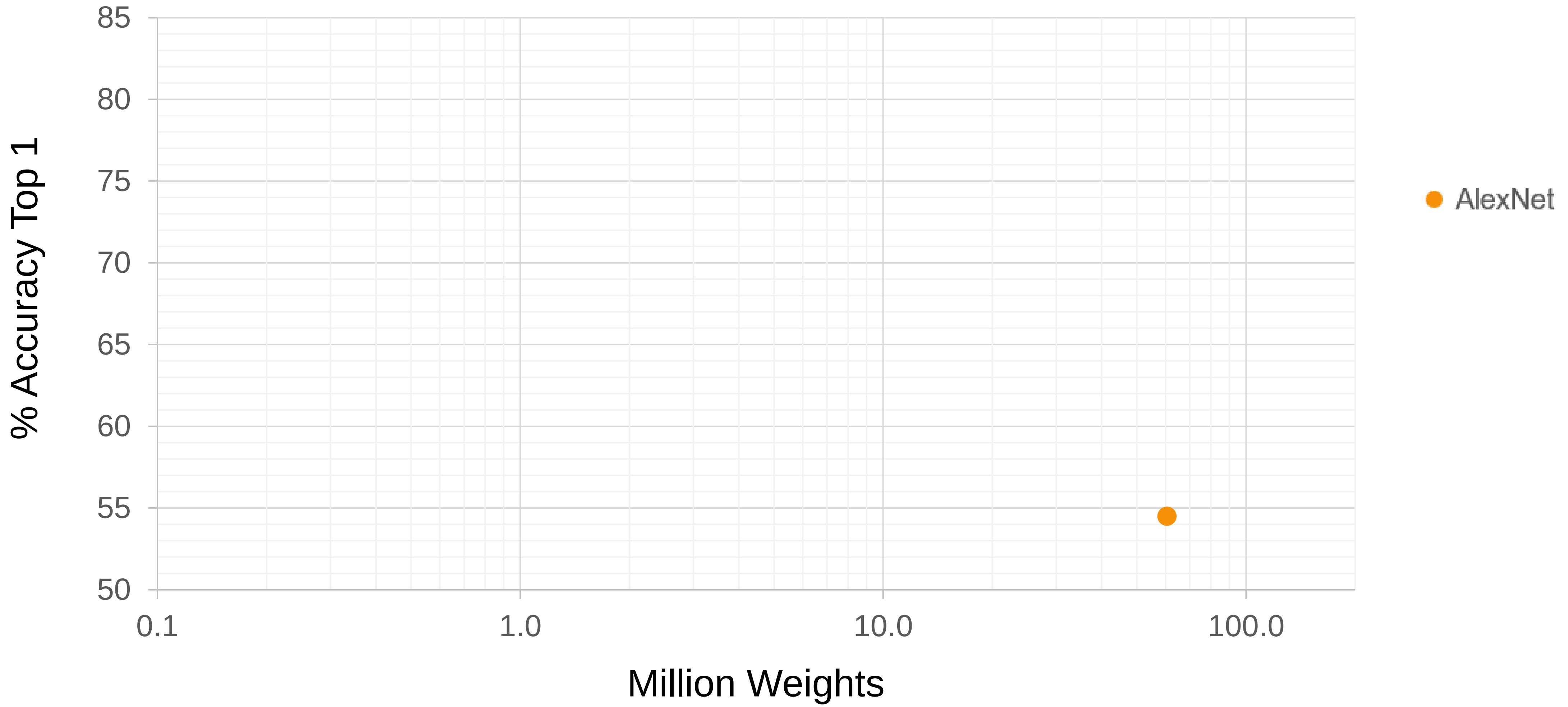


synopsys®

© 2020 Synopsys

Trend 2: Reduced Model Sizes

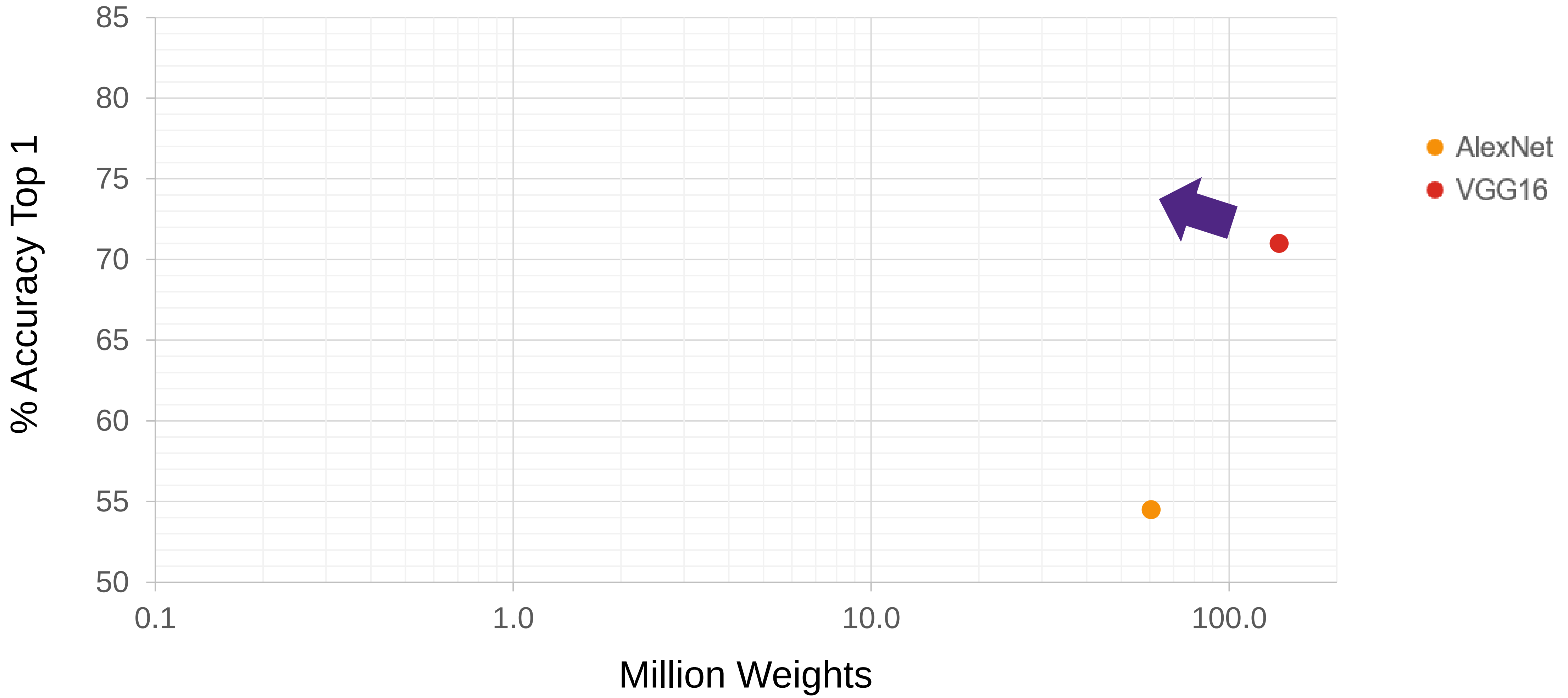
2012



Trend 2: Reduced Model Sizes

2014

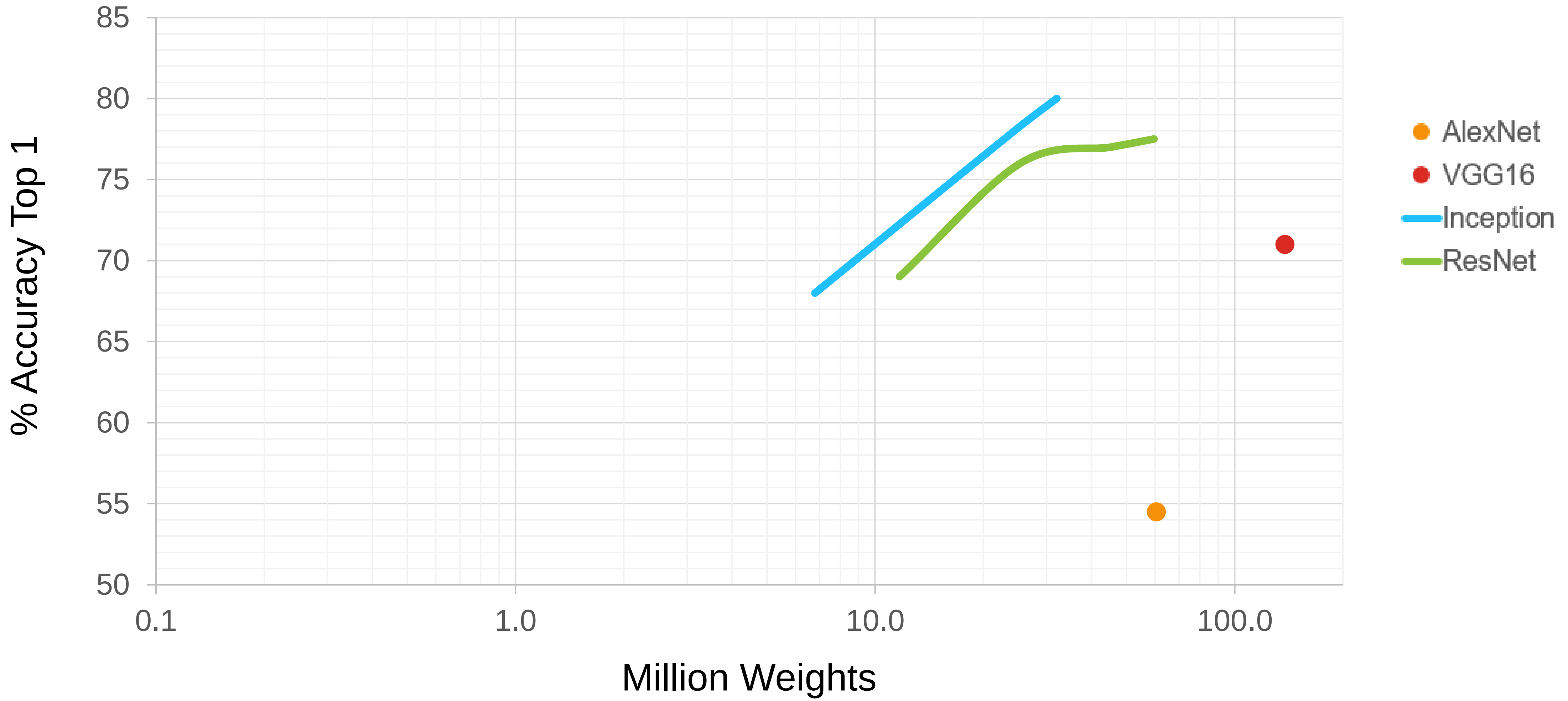
2020 embedded vision summit



SYNOPSYS®

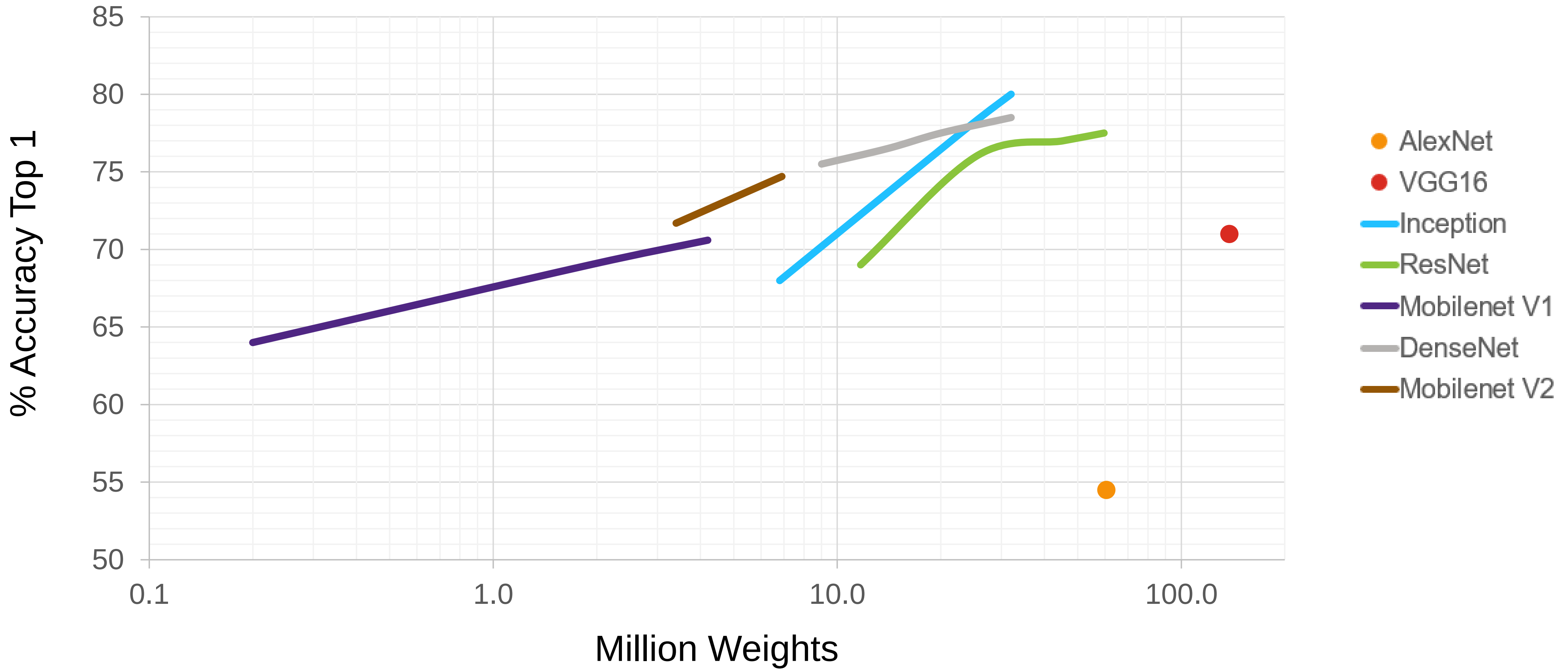
© 2020 Synopsys

Trend 2: Reduced Model Sizes

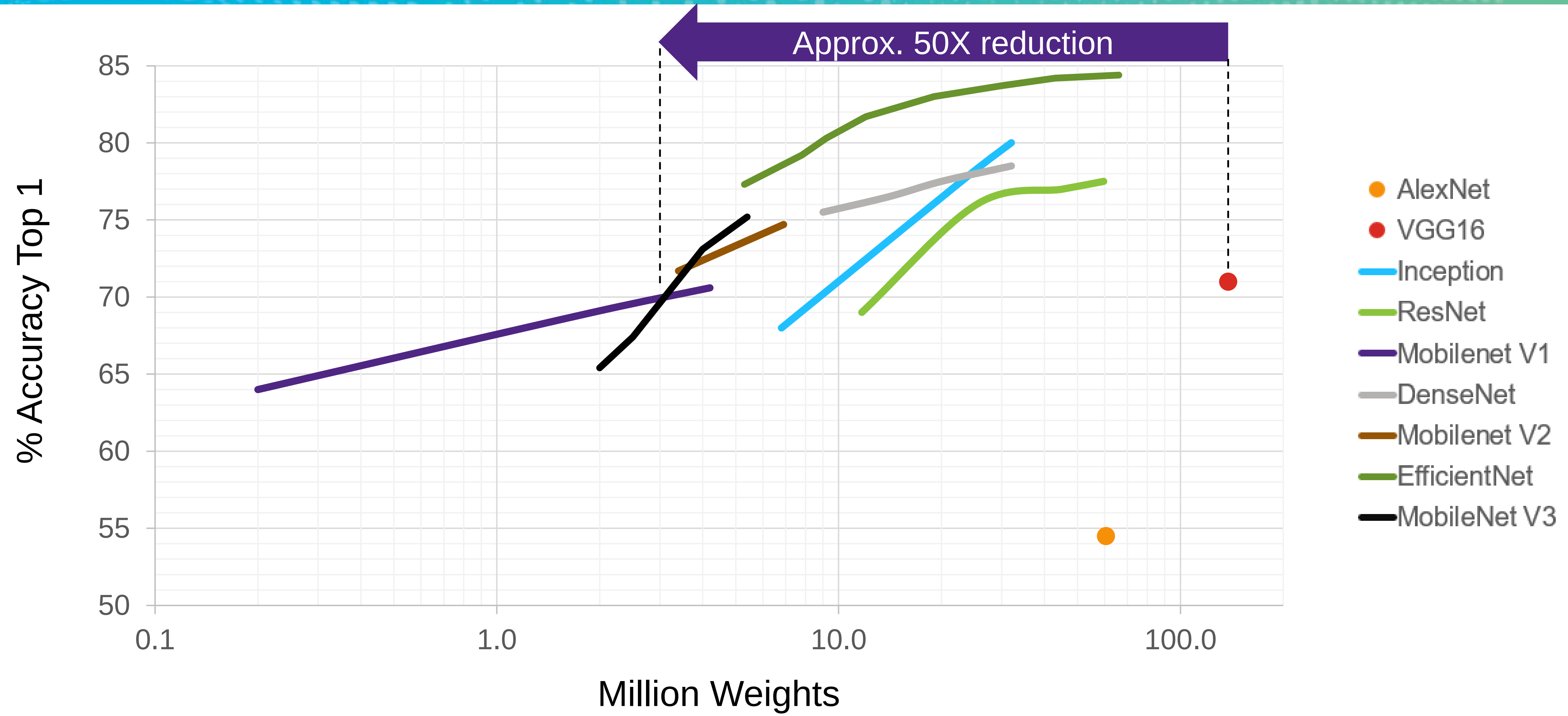


Trend 2: Reduced Model Sizes

2018

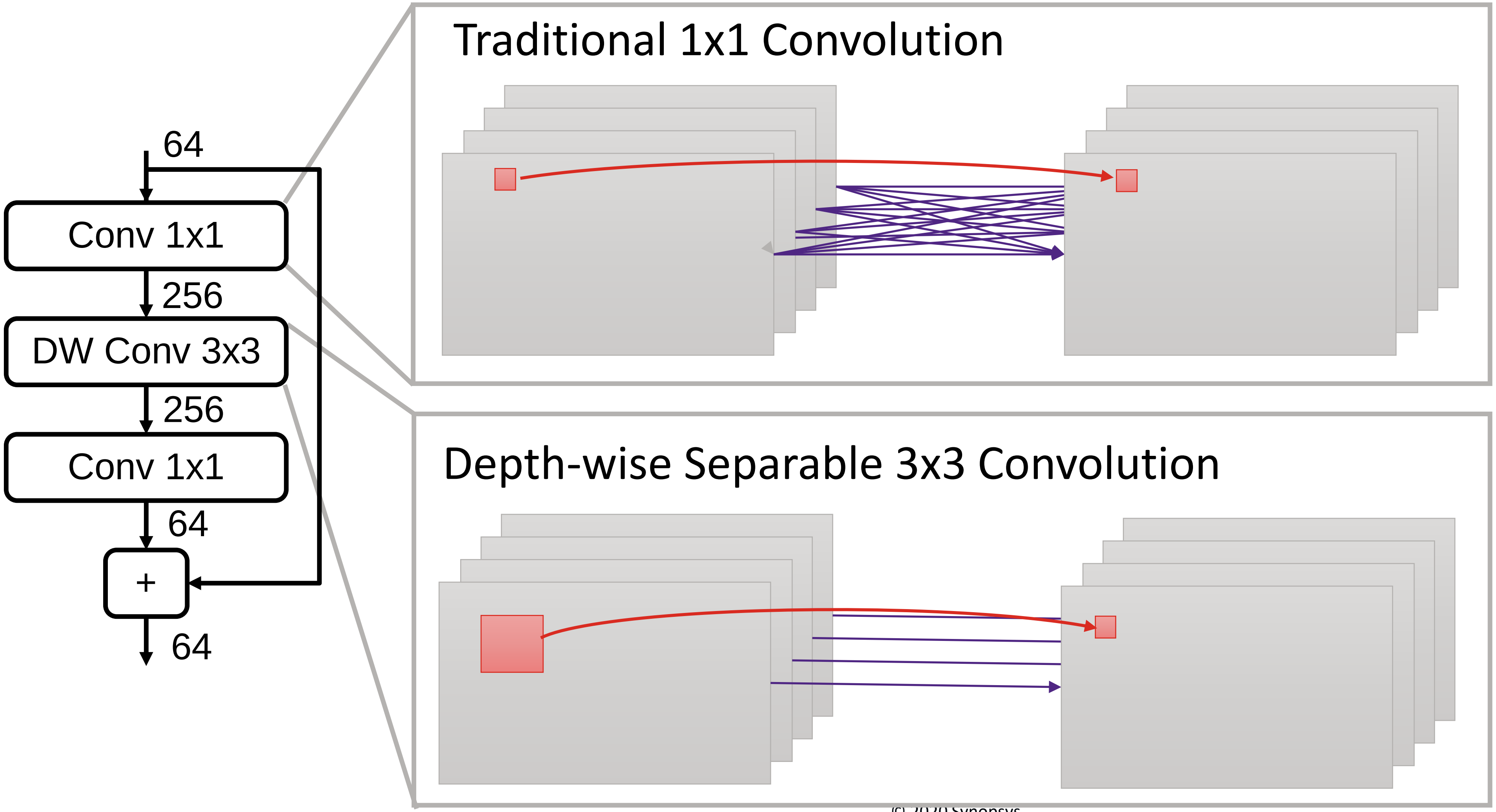


Trend 2: Reduced Model Sizes



Trend 3: Reduced Data Reuse and Parallelism

Example: Depthwise Separable Kernels used in MobileNet V2/V3, EfficientNet



64
Conv 1x1
256
DW Conv 3x3
256
Conv 1x1
64
+
64
Traditional 1x1 Convolution
Depth-wise Separable 3x3 Convolution

High Computation

High Data Reuse

High Parallelism

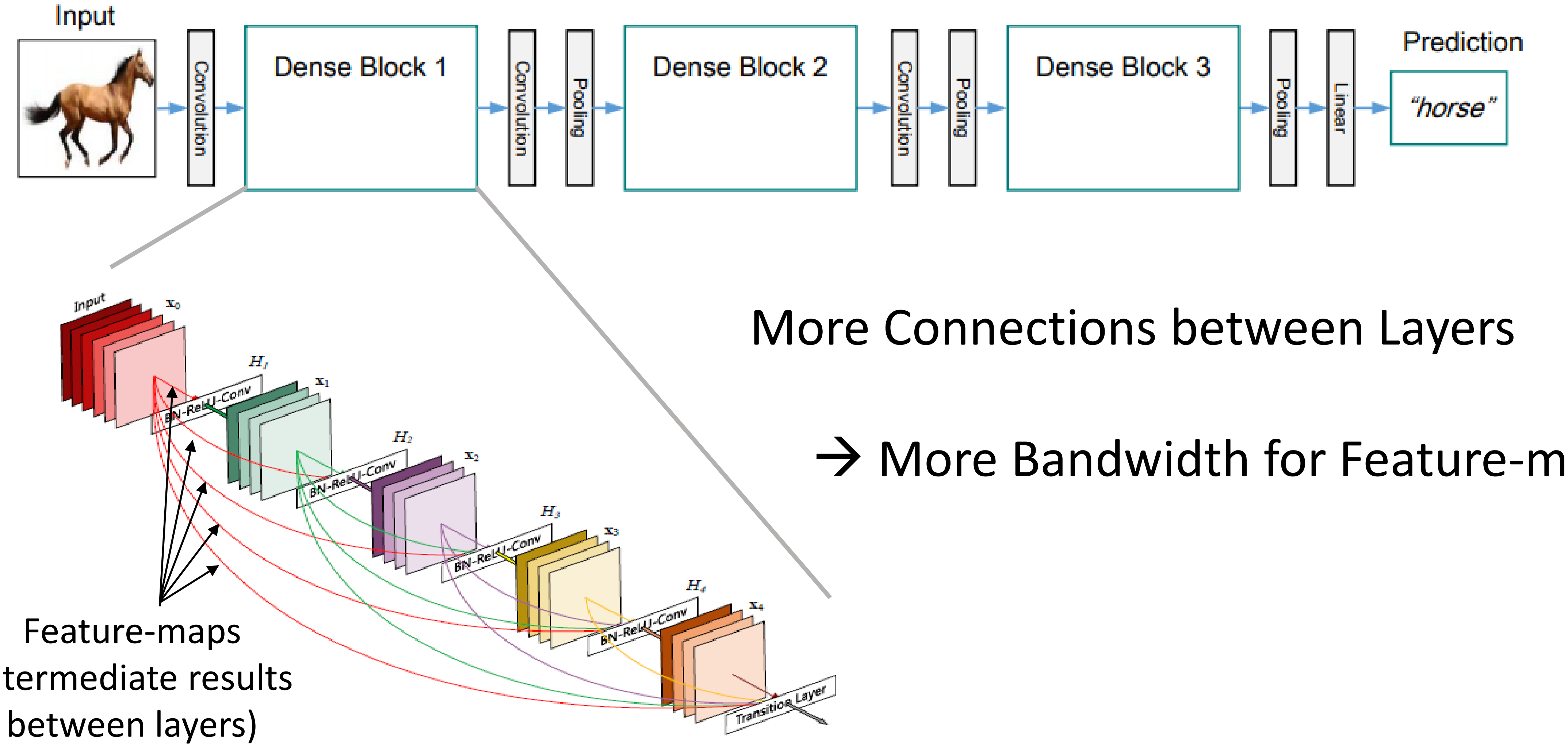
Low Computation

Low Data Reuse

Low Parallelism

Trend 4: Feature-map Bandwidth Becomes Dominant

Example: DenseNet and Multilayer DenseNet



Feature-maps (intermediate results between layers)

More Connections between Layers

→ More Bandwidth for Feature-maps

Trends in Convolutional Neural Networks Topologies

Trend 1:	Reduced Computational Requirements
Trend 2:	Reduced Model Size
Trend 3:	Reduced Data Reuse and Parallelism
Trend 4:	Feature-map Bandwidth Becomes Dominant

Examples:
MobileNet,
EfficientNet

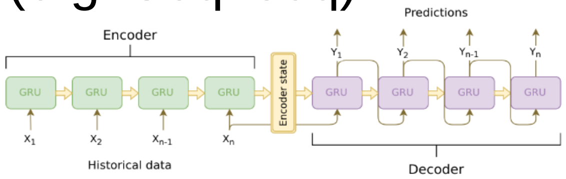
Trends in other domains like Audio & Speech

RNNs are replaced by 1D Convolutions

New: 1D Convolutions and RNN's replaced by Transformers

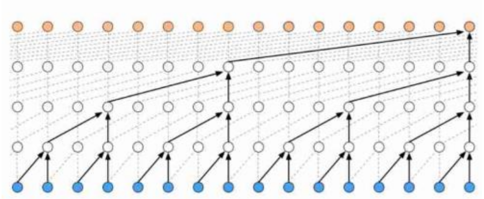
2014 - …

RNN: LSTM and GRU
(e.g. Seq2seq)



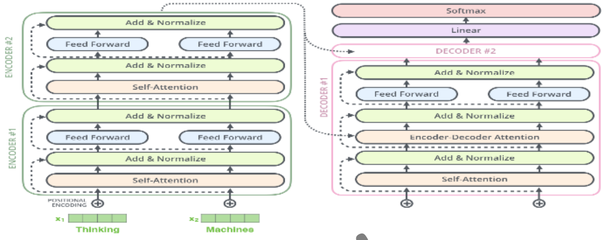
2016 - …

1D Convolutions
(e.g. WaveNet)



2019 - …

Transformers



More data-reuse, easier to parallelize and train

Key Challenges and Opportunities

- Opportunities
  - Drive towards more efficient networks focused on real world constraints
  - Well-defined, abstract high-level representation
  - Standardization of framework data representation: TensorFlow and ONNX
- Challenges
  - Optimize compute resource utilization under tight bandwidth constraints
    - Memory bandwidth not scaling with compute resources
    - Energy efficiency (mJ/frame) related to resource utilization, and memory bandwidth
  - Low-power and low-area with high flexibility
    - Adapt to constant innovation of NN-based applications
  - Complexity of NN compiler tools
    - Single biggest investment in EV project resources

Vision Applications Require Varying Levels of Performance

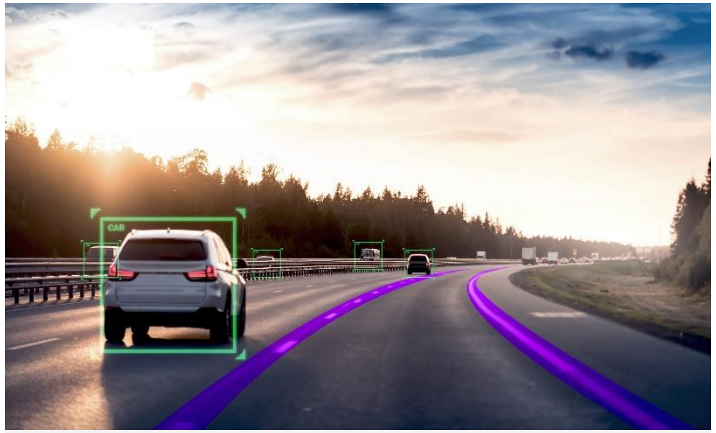
Performance requirements per application are increasing

- Facial recognition
- Always-on IoT / Smart Home
- Mid-end smartphones
- Games/toys
- Automotive in-cabin camera

<1 TOPS

- Augmented reality
- Surveillance
- Digital still cameras
- Automotive rear cameras
- High-end smartphones
- Natural language processing
- Robotics
- Drones

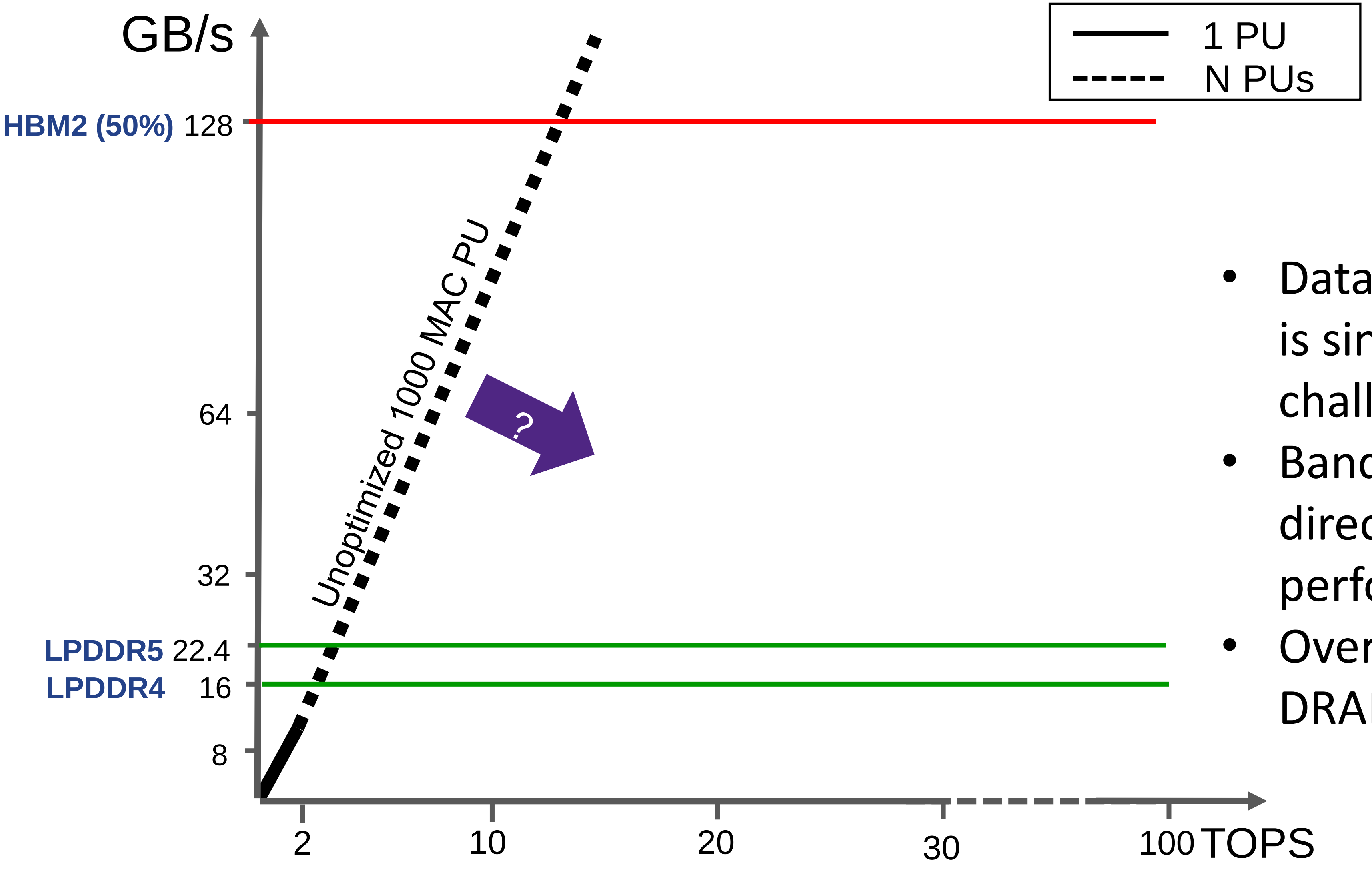
1 to 10 TOPS



- Automotive front camera
- DTV Super resolution
- Microservers (inference)
- Data center (inference)

10 to 100s of TOPS

Scaling Performance with Bandwidth Constraints



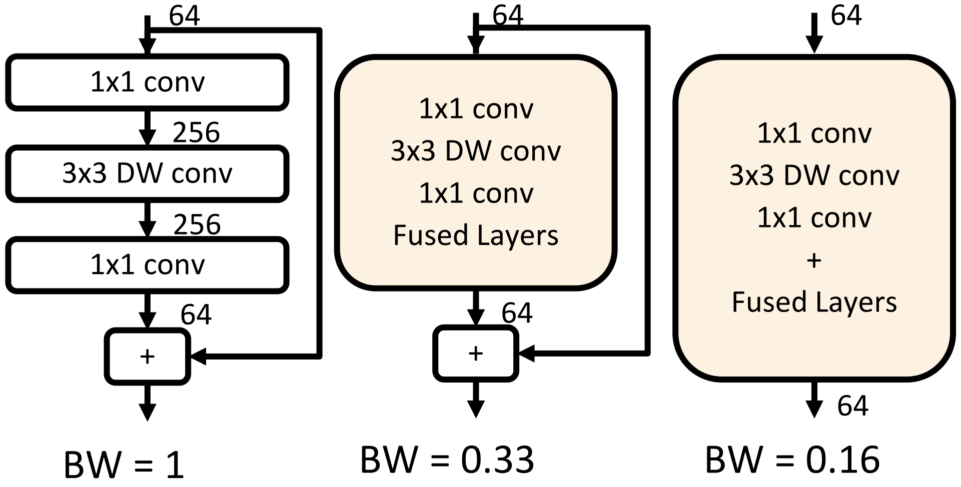
- Data bandwidth optimization is single most important challenge
- Bandwidth reduction has direct impact on performance and power
- Over 50% of SoC power is DRAM access

Bandwidth Improvement Solutions

- Coefficient Pruning
  - Coefficients with a zero value are skipped/counted
  - Modern graphs have ~60% zero coefficients
- Feature Map Compression
  - Runtime compression and decompression of feature maps to external memory
  - Approx. 40% feature-map bandwidth reduction

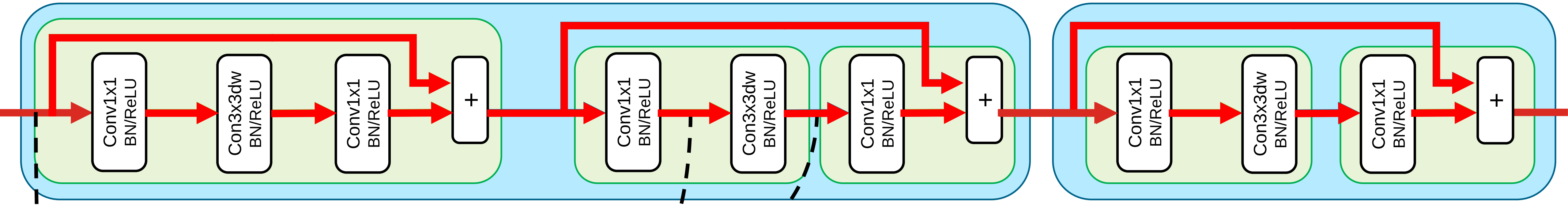
- Multi-level Layer Fusion
  - Merging multiple folded layers into single primitives reduces feature map bandwidth

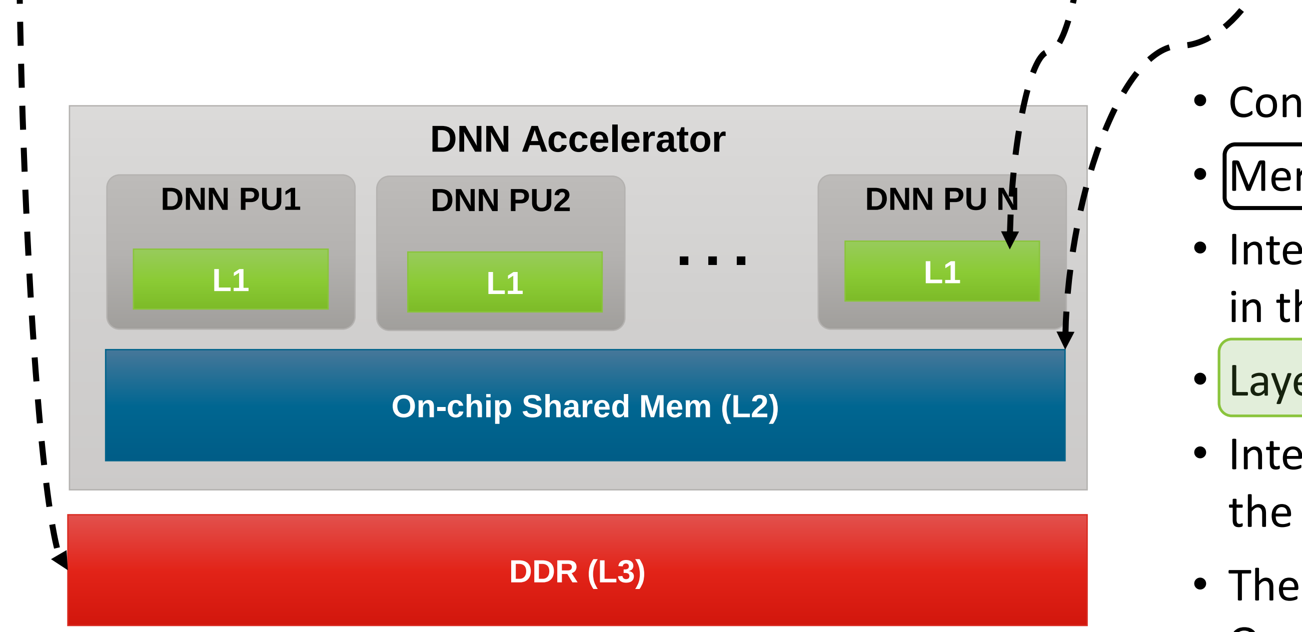
Multi-level Layer Fusion MobileNet v1/v2



Advanced Data Bandwidth Reduction Techniques

Multi-level Layer Fusion and Multi-level Tiling





- Convolutions, Pooling and Activations: merged layers
- Merged layers can be fused into layer groups
- Intermediate feature maps within a layer group are tiled to fit in the L1 Closely Coupled Memories of the DNN slices
- Layer groups can be fused into segments
- Intermediate feature maps within a segment are tiled to fit in the L2 On-chip Shared Memory
- The output of a segment may be too large (>10MB) to fit in L2 On-chip Shared Memory and is spilled to L3 DDR

Summary of Key Challenges and Opportunities

- Opportunities
  - Drive towards more efficient networks focused on real world constraints
  - Well-defined, abstract high-level representation
  - Standardization of framework data representation: TensorFlow and ONNX
- Challenges
  - Low-power and low-area with high flexibility
  - Complexity of NN compiler tools
  - Optimize compute resource utilization under tight bandwidth constraints
    - Single biggest challenge
    - Multi-level layer merging, fusion and tiling part of solution

Resources

MobileNetV2: Inverted Residuals and Linear Bottlenecks: https://arxiv.org/pdf/1801.04381.pdf

Densely Connected Convolutional Networks: https://arxiv.org/abs/1608.06993

ICNet for Real-Time Semantic Segmentation on High-Resolution Images: https://arxiv.org/abs/1704.08545

Panoptic Segmentation: https://arxiv.org/abs/1801.00868

EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks: https://arxiv.org/abs/1905.11946

YOLOv4: Optimal Speed and Accuracy of Object Detection: https://arxiv.org/pdf/2004.10934.pdf

Searching for MobileNetV3: https://arxiv.org/abs/1905.02244

Thank You

Pierre Paulin

pierre.paulin@synopsys.com

SYNOPSYS®