embedded VISION SUMMIT 2018

New Deep Learning Techniques for Embedded Systems

SYNOPSYS®

Tom Michiels, System Architect, Embedded Vision Processors

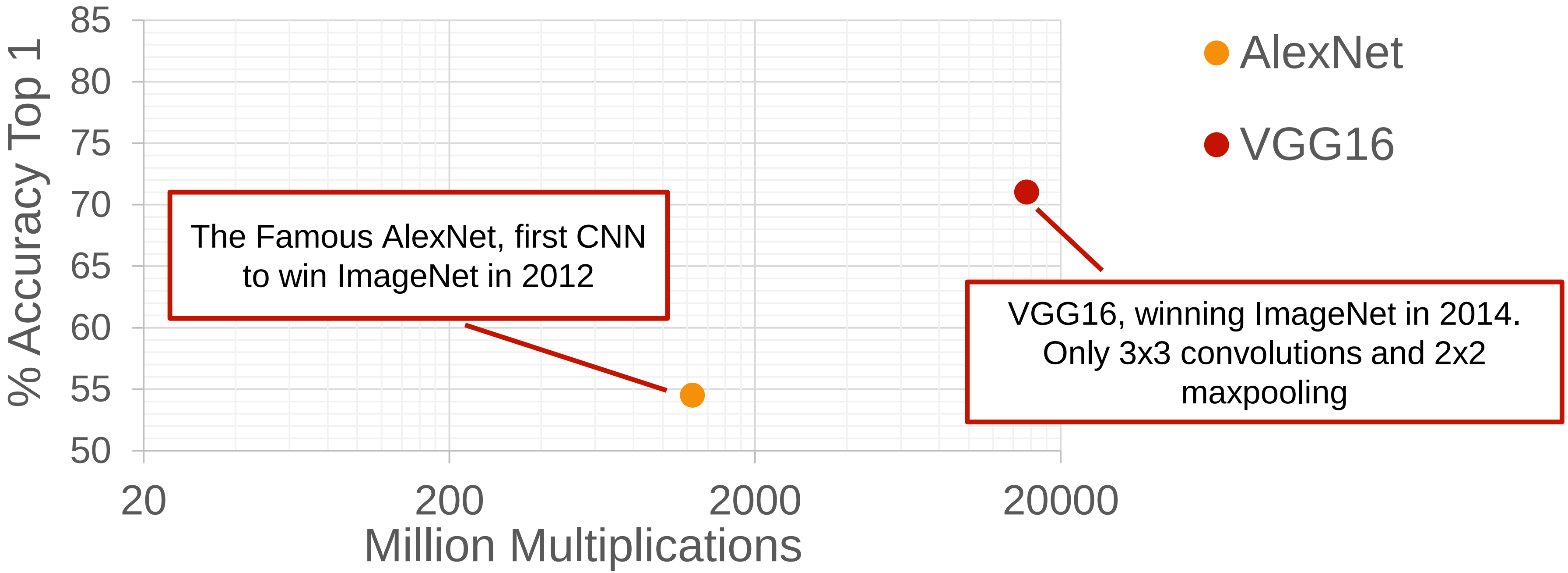
May 23, 2018

Overview

- Evolution towards more efficient and compact CNN graphs
- Memory footprint of weights and feature maps
- External memory bandwidth and impact on power
- How graph structure can impact bandwidth
    - Tiling and merging layers

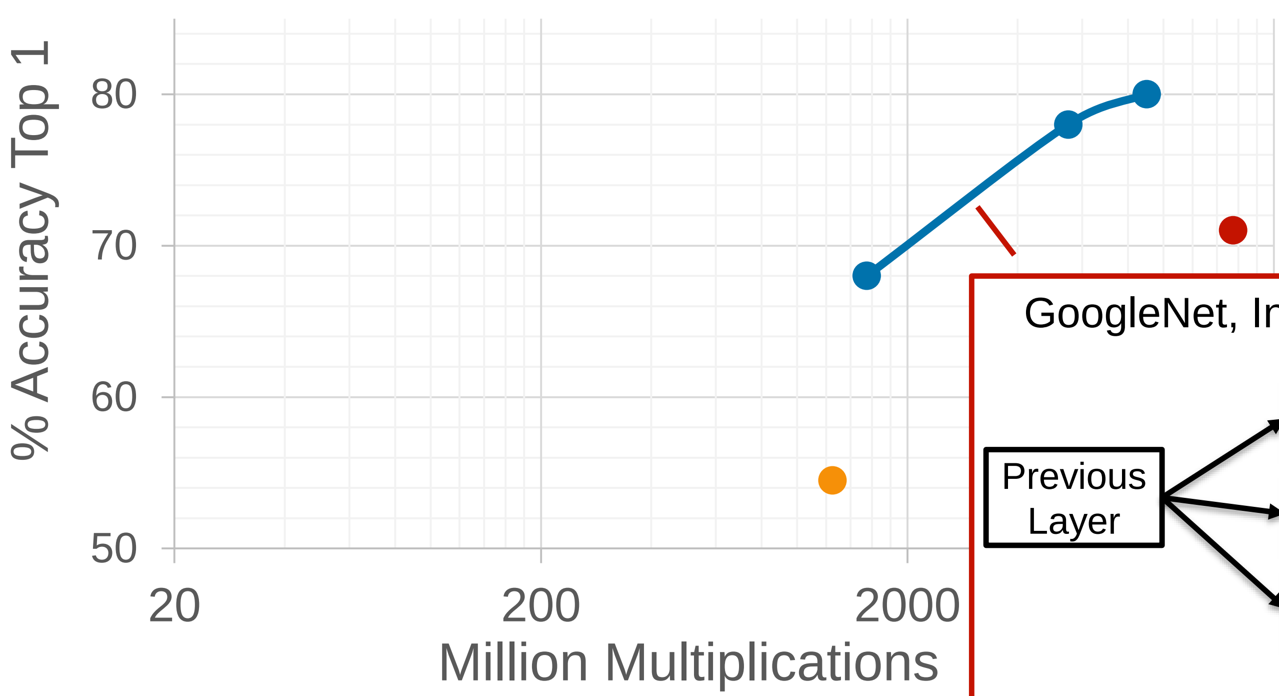
Evolution towards Compact Convolution Networks

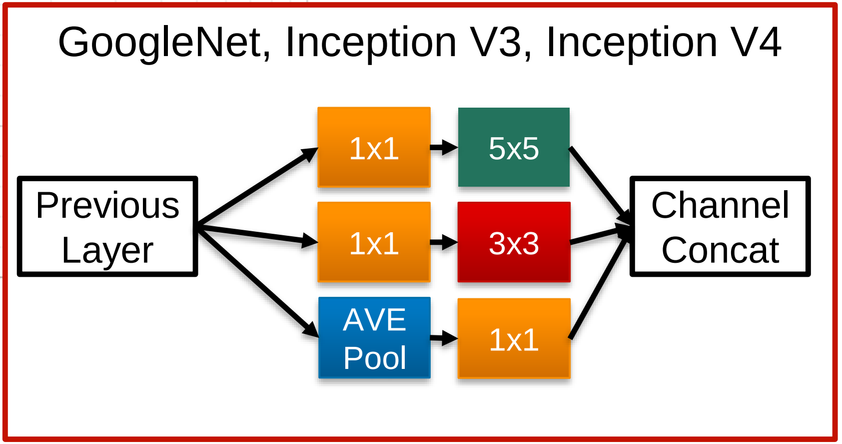
ImageNet ILSVRC-2012 Classification Accuracy vs Compute Requirements



Evolution towards Compact Convolution Networks

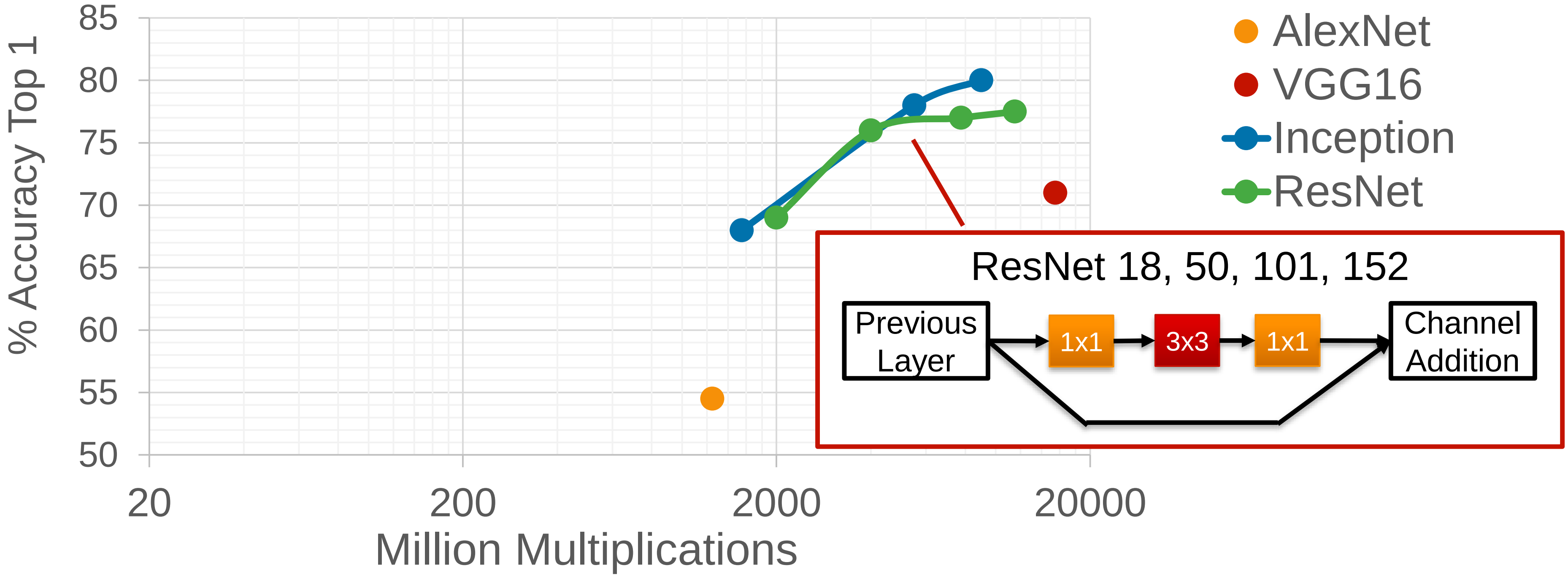
ImageNet ILSVRC-2012 Classification Accuracy vs Compute Requirements



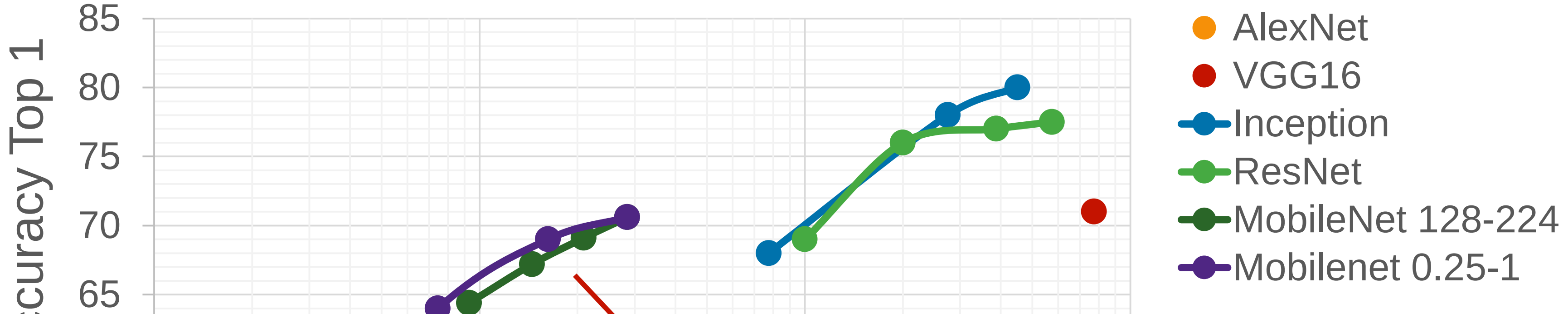


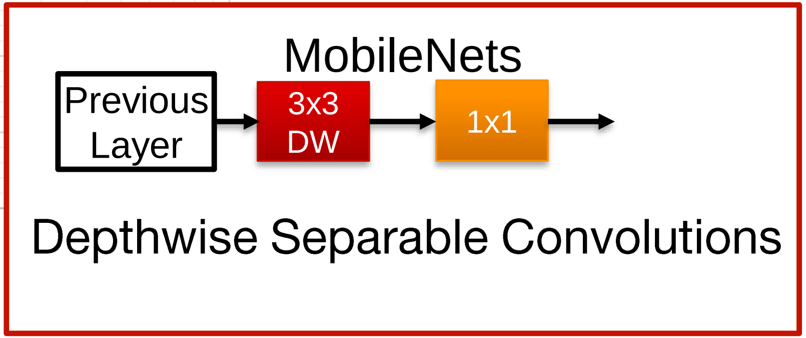
Evolution towards Compact Convolution Networks

ImageNet ILSVRC-2012 Classification Accuracy vs Compute Requirements

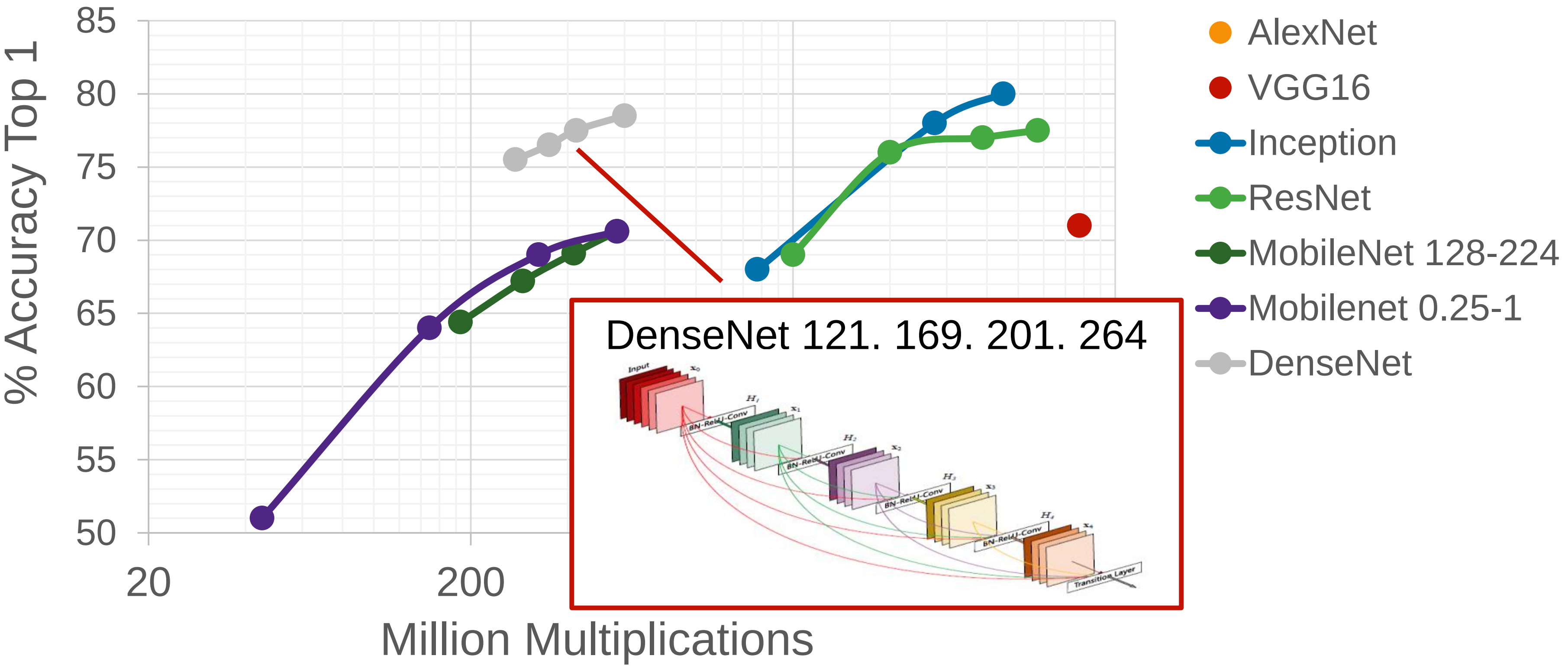


Evolution towards Compact Convolution Networks

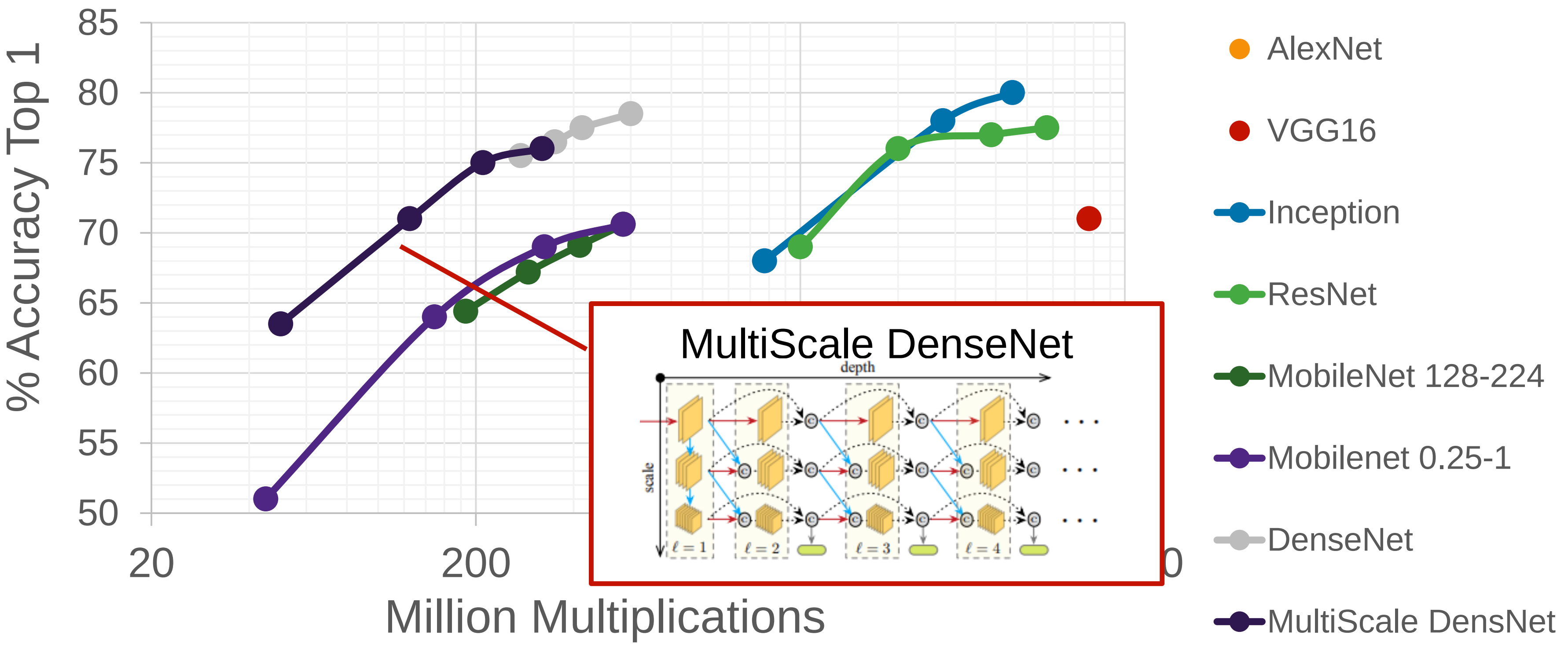




Evolution towards Compact Convolution Networks

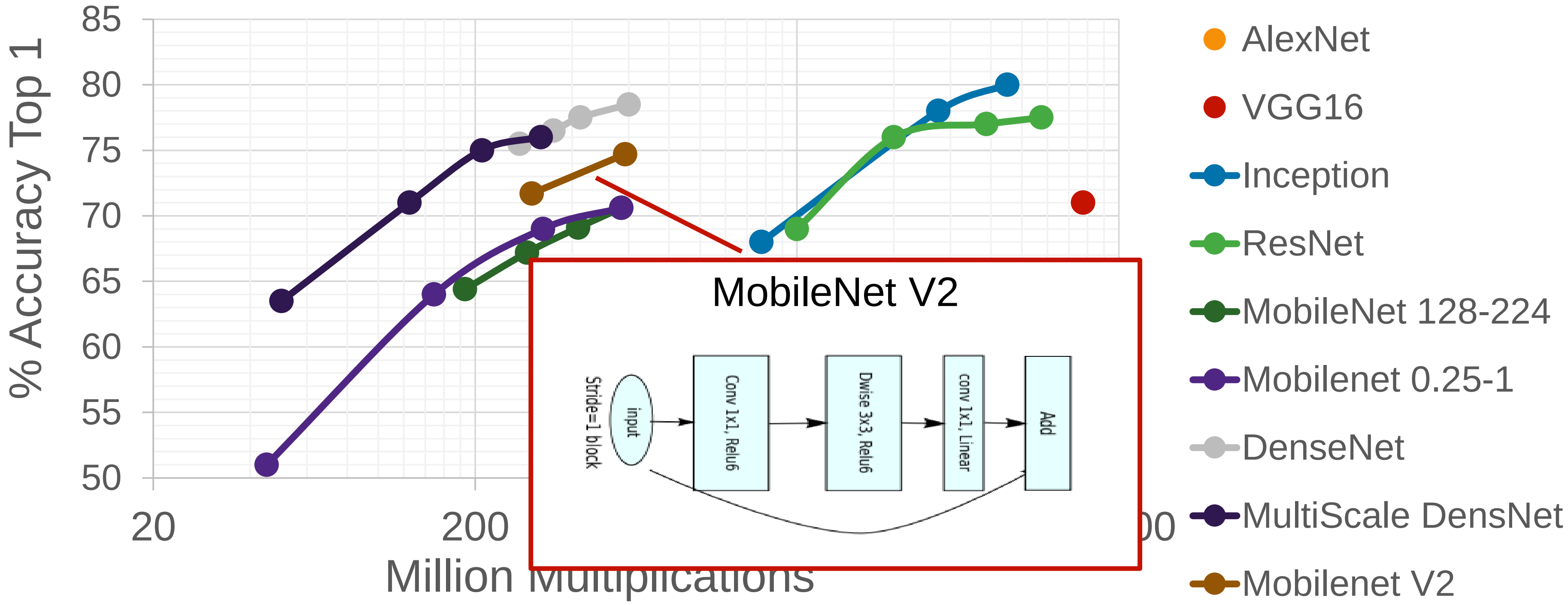


Evolution towards Compact Convolution Networks

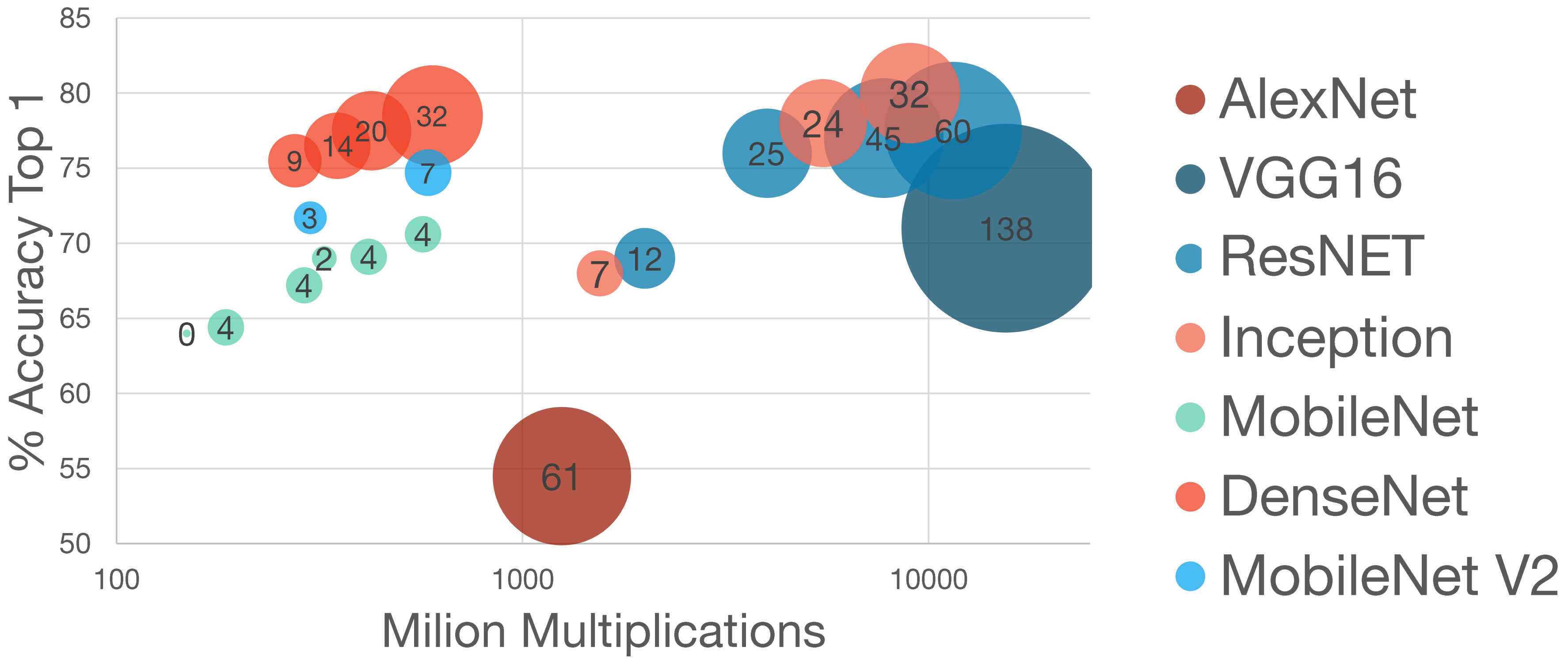


Evolution towards Compact Convolution Networks

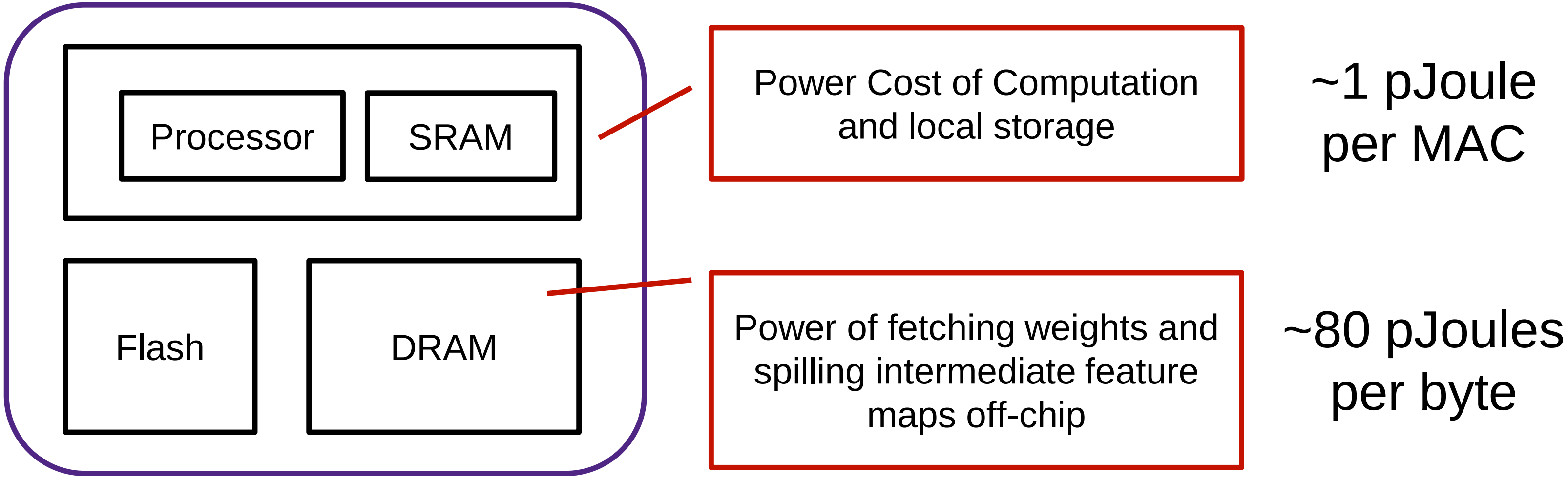
ImageNet ILSVRC-2012 Classification



Model Sizes in Million Weights



Power Cost of Mapping CNN

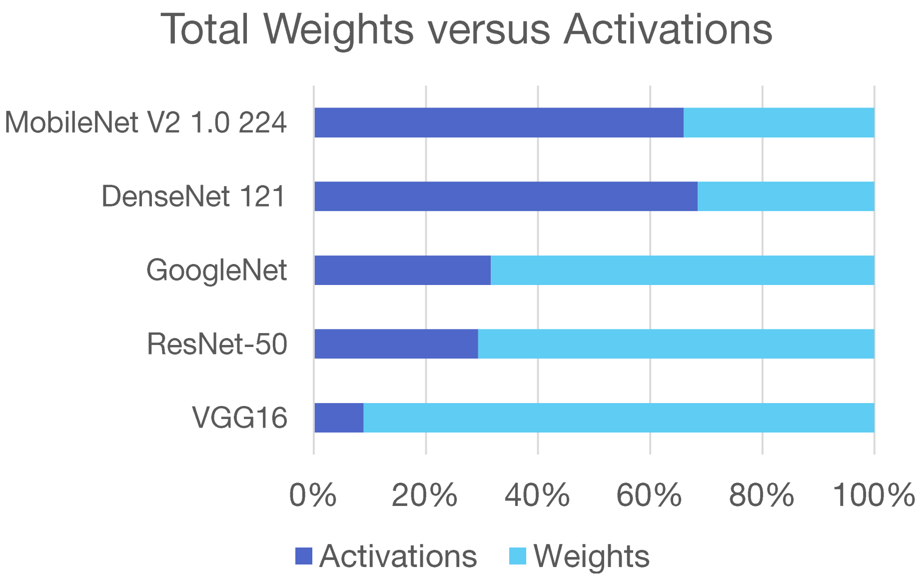


Power of computational cost and local storage can relatively easy be computed from the CNN graph structure.

How is external memory bandwidth related to the graph size and structure?

SYNOPSYS®

Size of Feature Maps & Coefficients



How is DRAM bandwidth related to the size of feature maps and the number of weights?

- For more compact graphs, the relative size of the feature maps increases.
- This is even more true when graphs run on large frame resolutions
  - Detection
  - Scene Segmentation
  - Style Transfer
- Weights can be compressed more than feature maps
  - Pruning
  - Offline Compression

How Much Bandwidth for Processing a Layer?

Example Conv4_2 of VGG16

```
DataType F_in[512][28][28];
DataType F_out[512][28][28];
DataType W[512][512][3][3];

For (m=0; m<512; m++) {
  for (x=0; x<W; x++) {
     for (y=0; y<H; y++) {
      F_out[m][x][y] = Bias[m];
      for (i=0; i<I; i++) {
        for (j=0; j<J; j++){
          for (k=0; k<512; k++) {
            F_out[m][x][y] += F_in[k][x+i][y+j] * W[m][k][i][j];
          }
        }
      }
    }
  }
}
```

400K Values

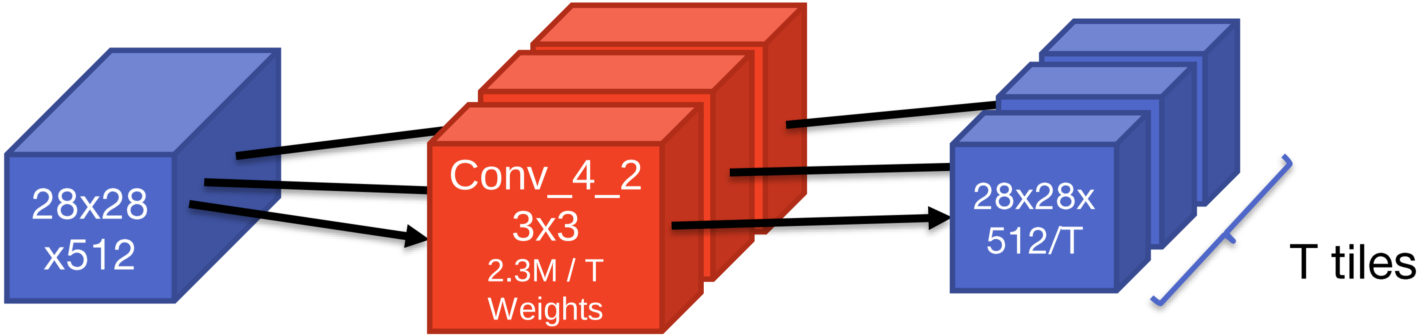
400K Values

2.4M Values

- On a memory constrained (e.g. 256KB) architecture, none of these will completely fit in the local memory
- How would we tile it to fit in local memory?

How Much Bandwidth for Processing a Layer?

Example Conv4_2 of VGG16

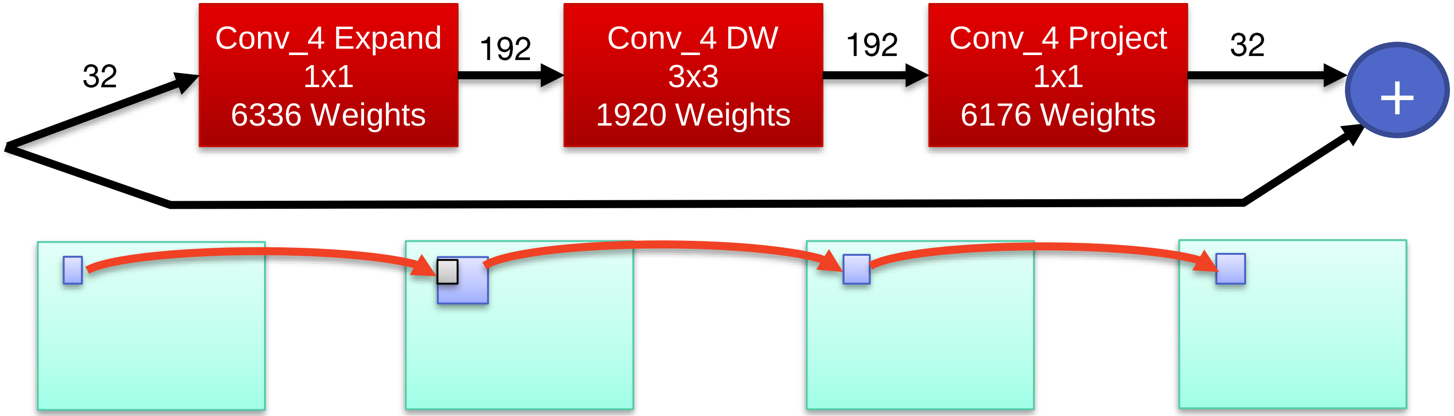


We can partition the channels of the activation of the convolution in T parts and re-read the feature maps of the previous layer T times.

For some layers/graphs the DRAM bandwidth can be much higher than the sum of the weights and the feature maps

How Much Bandwidth for Processing a Layer?

Example MobileNet V2 Layers



If the weights of multiple consecutive layers fit the local memory, we can merge the convolution layers, by in the X/Y domain and avoid spilling intermediate feature maps

For some layers/graphs the DRAM bandwidth can be much lower than the sum of weights and the feature maps

How Much Bandwidth for Processing a Layer?

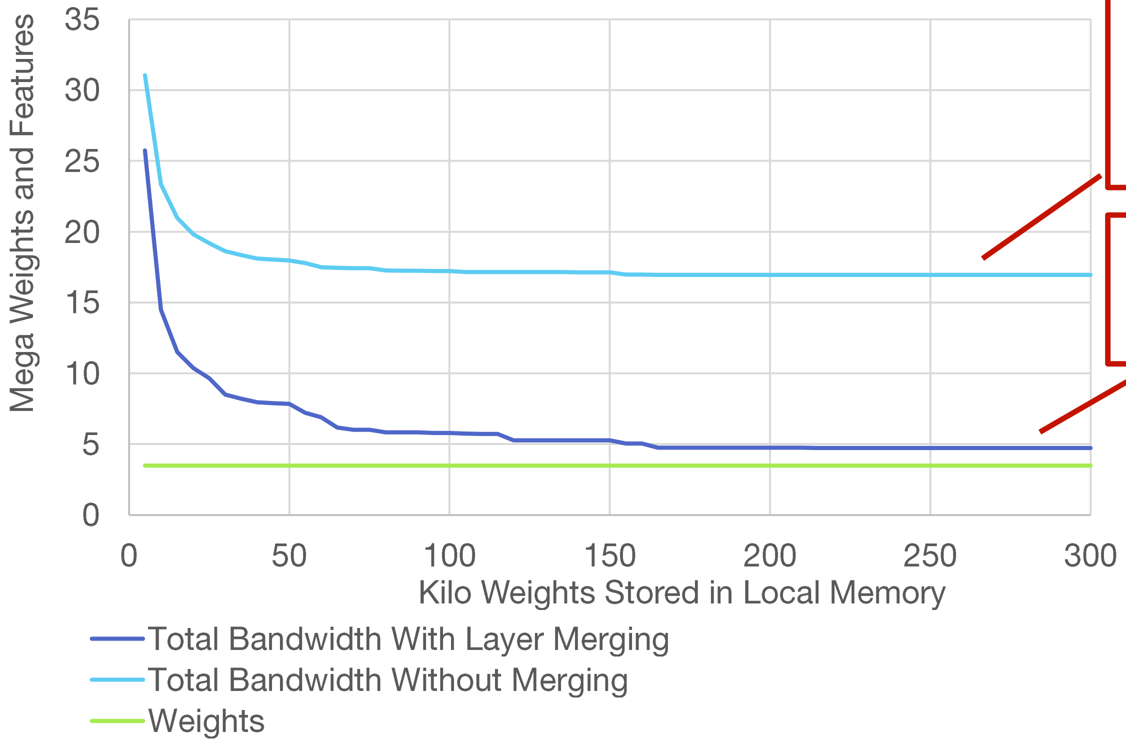
Small — Size of Weights of Layers — Large

If the weights of multiple layers can be stored in local memory, spilling of intermediate feature maps can be avoided by merging layers

If the weights of a single layer do not fit, feature maps need to be read multiple times from DRAM

Bandwidth Example, MobileNet

Example MobileNet V2 1.0 224

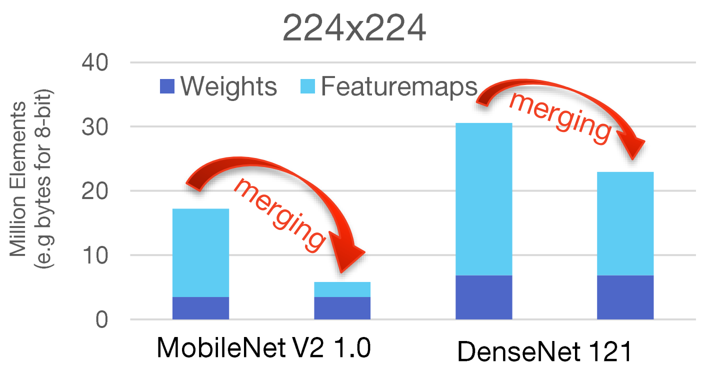


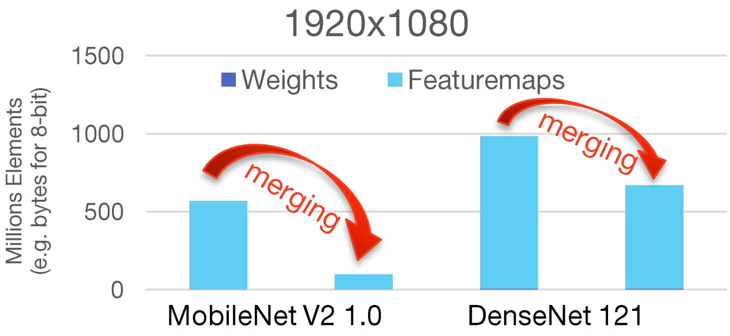
Without merging of layers, bandwidth decreases until the weight of every single layer fits in local memory

Significant bandwidth gain from merging layers already

Weight compression does not only reduce bandwidth of loading weights, but also that of spilling feature maps

Bandwidth after Merging for MobileNet vs DenseNet



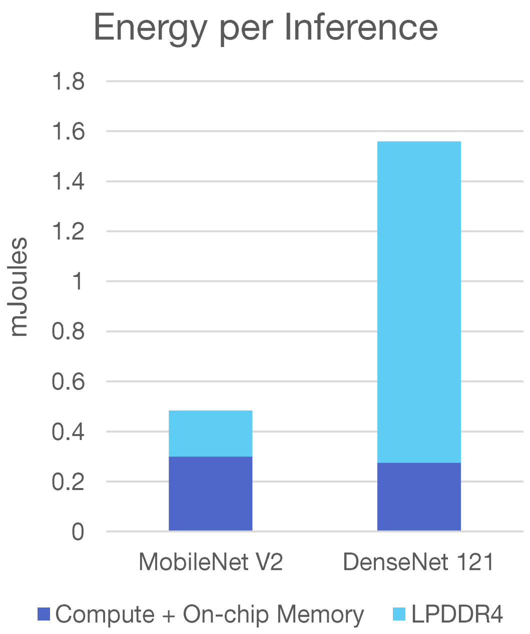


- The opportunities for layer merging depend on the network structure.
  - Densely connected layers are harder to merge because the same feature map is consumed by multiple layers

- Real applications will run on larger frames than 224x224
  - Scene Segmentation,
  - Detection (Yolo, SSD)

- Feature maps can be compressed, but not as much as coefficients.

Implications on Energy Consumption



- Differences in bandwidth can have big impact on the total energy per inference.
- The graph on the left assumes:
    - 1 pJoules per MAC for the compute and on-chip memory
    - 10 pJoules per bit to read from DRAM
    - 8-bit accuracy
    - Only counting feature maps
- Other architectural choices have impact on the bandwidth. Example: Feature map compression

Takeaways

- DRAM bandwidth can be a dominant part in the power consumption of a CNN.
- The bandwidth is determined by weights and feature maps
  - Graphs are often optimized to have small amount of weights, but the bandwidth of feature maps often dominates.
- Bandwidth of feature maps can be reduced significantly by merging convolution layers.
  - This requires an architecture that allows merged execution of convolutions.
  - Storing weights compressed in local memory not only reduces the bandwidth for fetching the weights, but also for fetching and storing feature maps
- Today's networks are optimized for weight-size and macs.
  Optimizing for bandwidth, will likely lead to new network structures.

About Synopsys

- Visit **www.synopsys.com/EV**
- Visit the Synopsys booth to check out demos:
  - *Accelerating Android Neural Network Performance with DesignWare EV6x*
  - *Real-Time Object Classification & Tracking with DesignWare EV6x*
  - *AI, 3D Imaging & SLAM on-a-Single Chip for Embedded Markets (by Inuitive)*
  - *Face Recognition for Driver Monitoring System (by PathPartner)*

Resources

- Embedded Vision Alliance Website: Software Frameworks and Toolsets for Deep Learning-based Vision Processing embedded VISION SUMMIT 2018

Thank You

Tom Michiels, System Architect, Embedded Vision Processors

SYNOPSYS®