What Will it Take for DNN to Go Embedded?!

Samer Hijazi

cādence®

Motivation

Cadence's mission

- Enable ***better***, ***faster***, ***cooler*** silicon systems ***sooner***

Imaging/video recognition

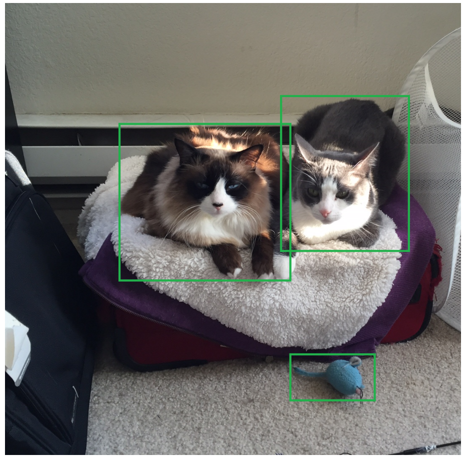
- Strong driver for creating advanced SoCs

Neural networks are a crucial innovation

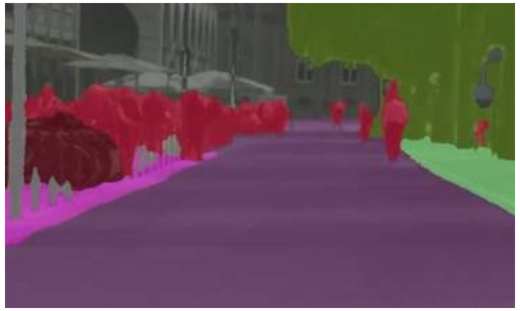
- **But, need a breakthrough in *efficiency***

Typical Computer Vision Problem

- Many classes region of interest Classifications (e.g., ImageNet and GTSB)



- Bounding Box Classification/identification.



- Pixel by Pixel segmentation.
  – If you have a 1080p image, we have 2M pixels in and 2M pixels out!
  – The most expensive from number of pixels point of view

cādence®

Why DNN Did Not Go Embedded Yet?

CNN Evolution

- Today's deep learning industry motto is "Deeper is Better"

Network	Application	Layers
LeNet-5 for MNIST (1998)	Handwritten Digit Recognition	7
AlexNet (2012)	ImageNet	8
Deepface (2014)	Face recognition	7
VGG-19 (2014)	ImageNet	19
GoogLeNet (2015)	ImageNet	22
ResNet (2015)	ImageNet	152
Inception-ResNet (2016)	ImageNet	246

The DNN Power Question

- Today's state of the art HW consume 40w/TMAC
- 4 TMAC is what be needed for a practical implementation of CNN based application.
- This means 160w!
  - Even 100 times improvement in HW efficiency is not enough.

Embedded devices power budgets, and form-factor cannot accommodate the current trend of DNN!

Courtesy of Dr. Stephen Hicks,
Nuffield Department of Clinical
Neurosciences, University of Oxford

cadence®

How to Save Power?

CNN is using excessive number of multiplies per Pixel!

- To solve this problem we can do four things

1. Optimize network architecture (Minimize the number of multiplies per pixel)
2. Optimize the problem definition (Minimize the number of needed pixels)
3. Minimize the number of bits needed to represent the network (Algorithmically reduce the cost of each multiplier)
4. Utilized and optimized HW for CNN

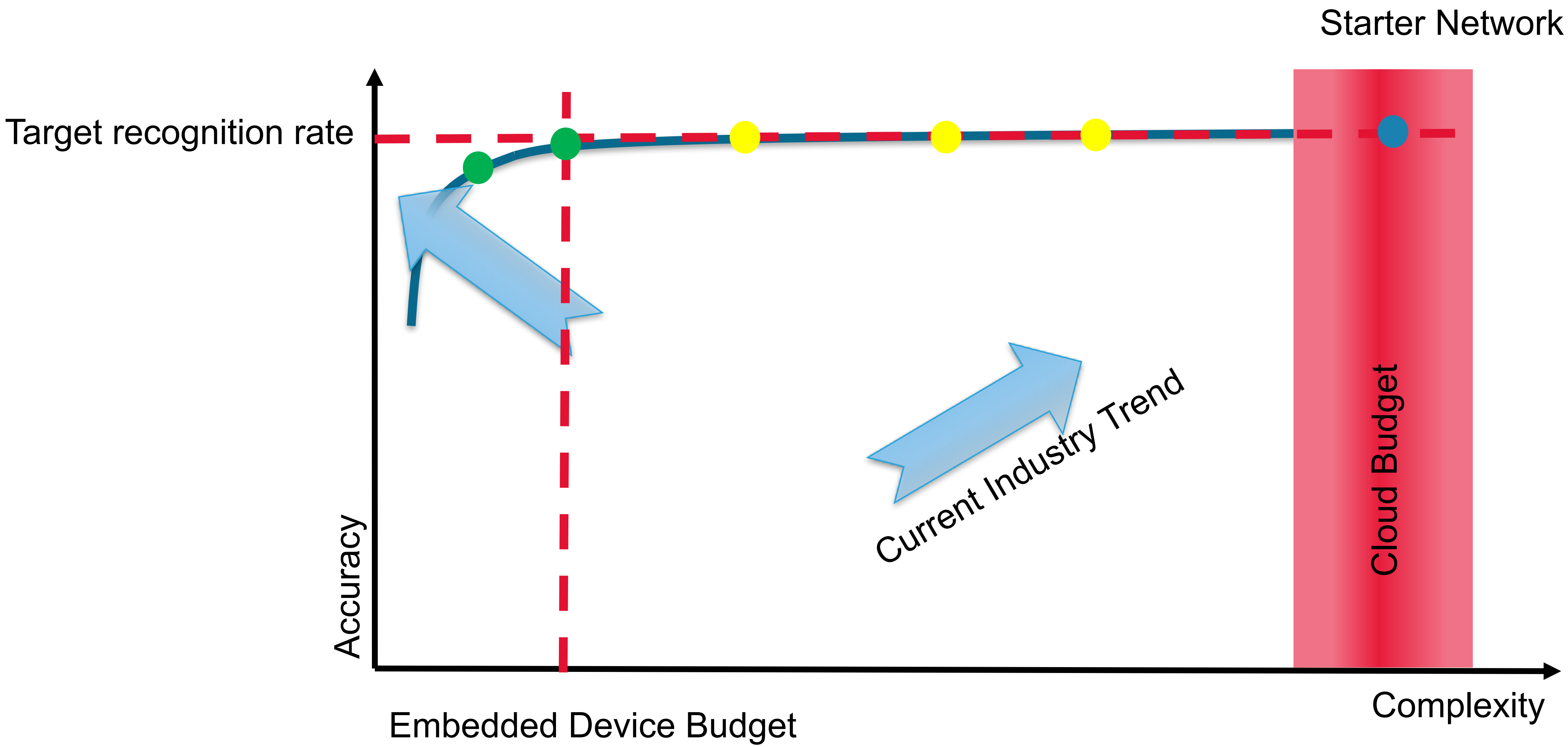
1. Optimize network architecture

2. Optimize the problem definition

3. Minimize the number of bits needed to represent the network

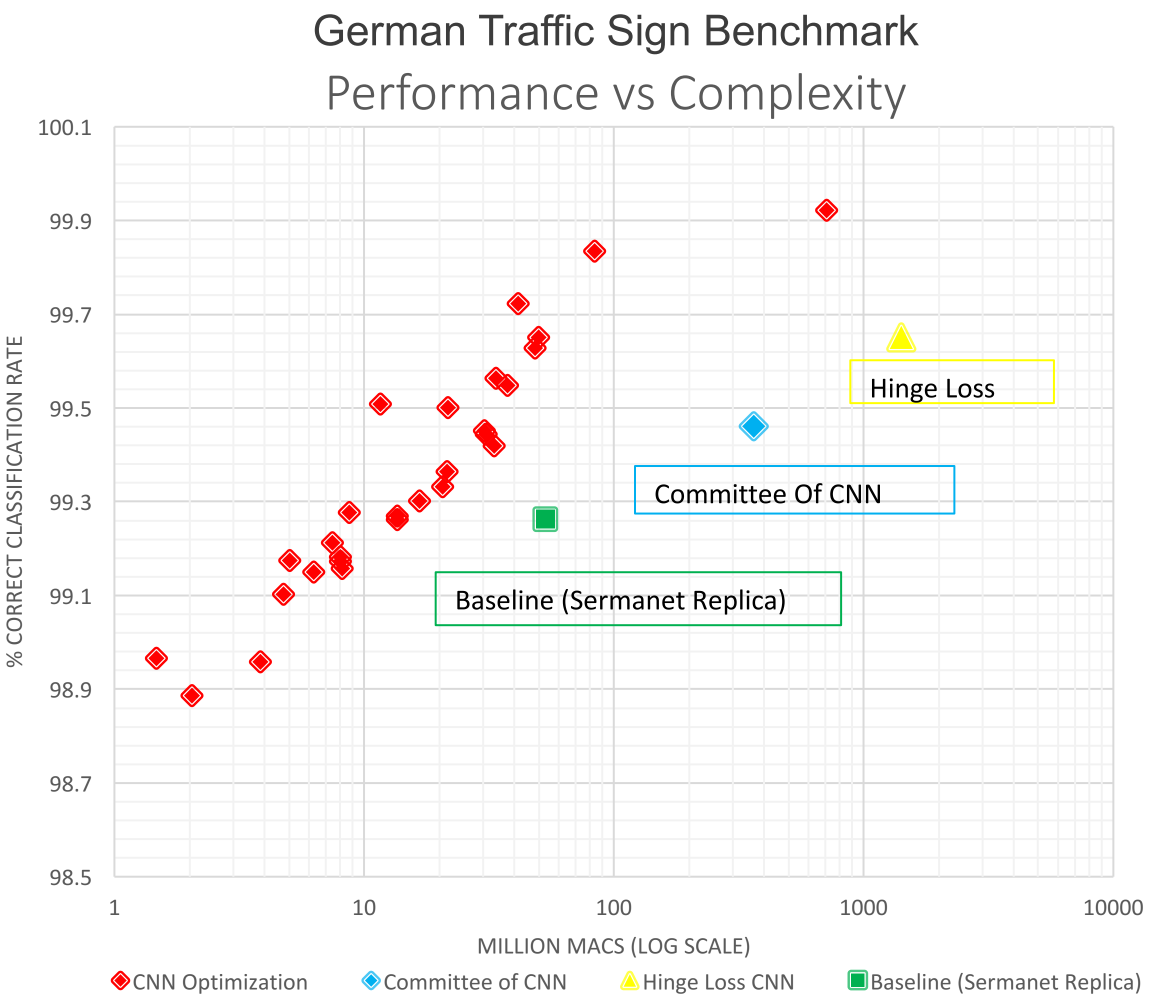
4. Utilized and optimized HW for CNN

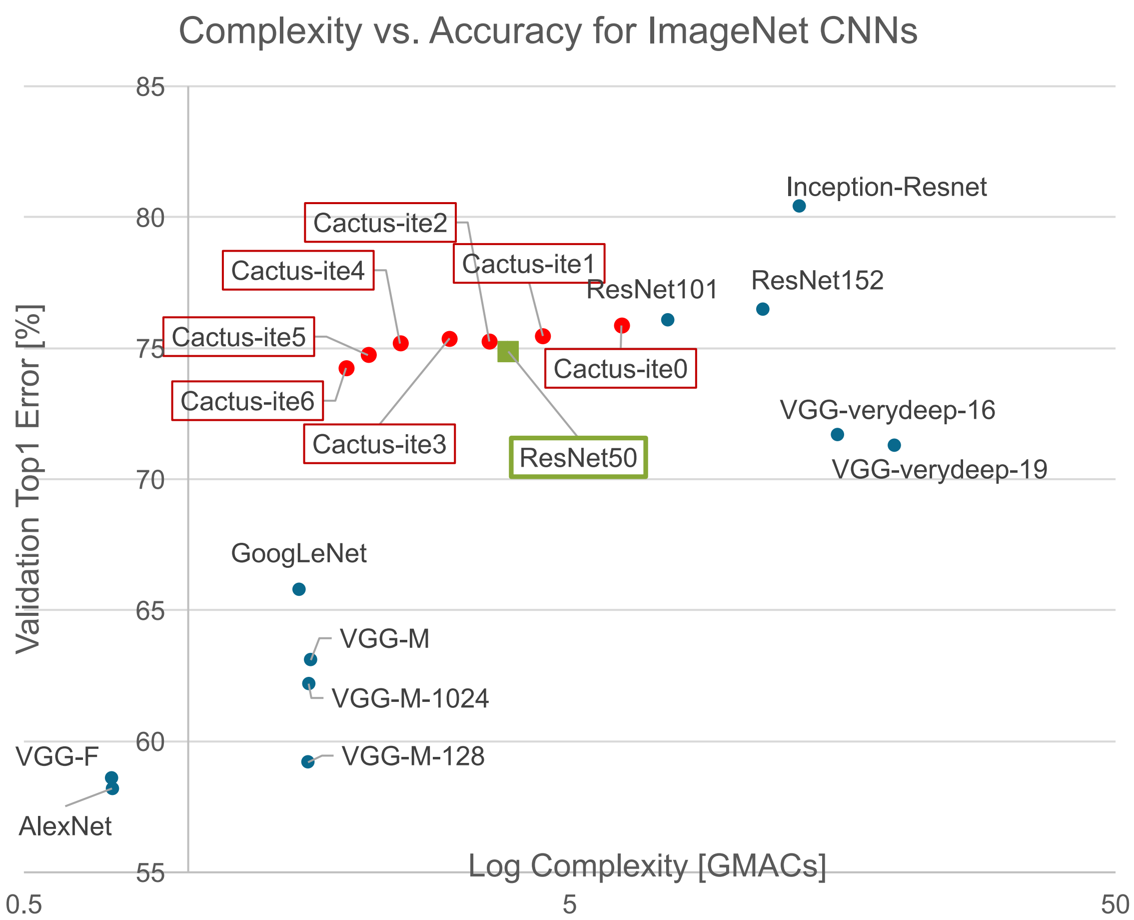
Complexity vs Performance



cadence

Can it Be Done?
You Bet!



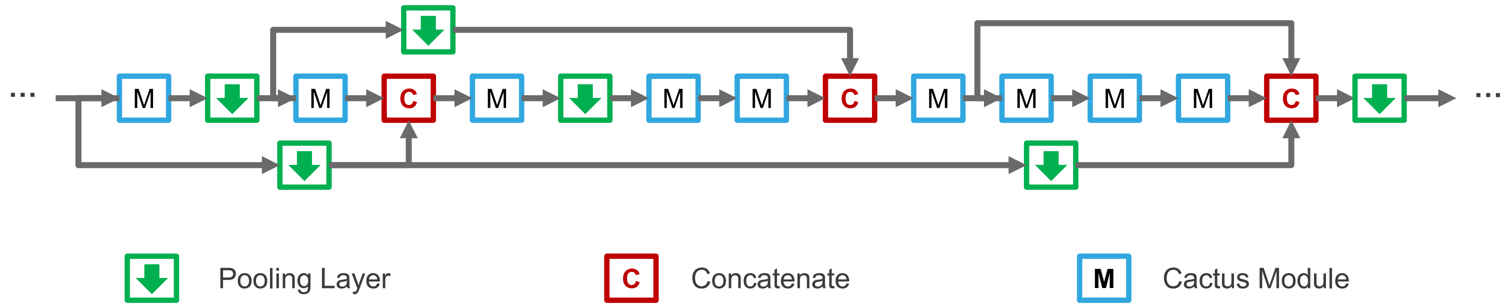


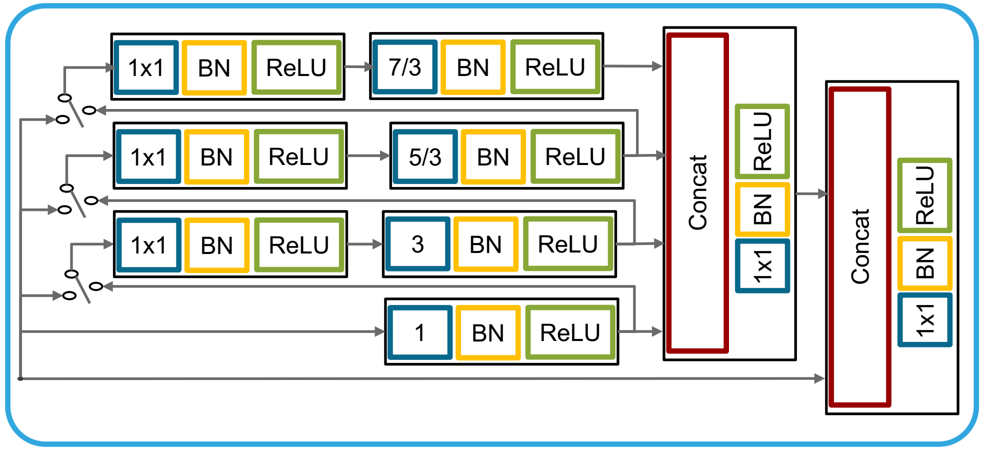
Automatic Optimizations of Network Structure.

- The ingredients:
  - A generic superset network architecture with many knobs to dial
    - CactusNet
  - Incrementally optimize the net complexity creating a family of closely related networks.
  - Sensitivity analysis:
    - Model and Measure the amount of redundancy in a network vs accuracy

CactusNet

- A general CNN reference architecture with lots of control knobs.

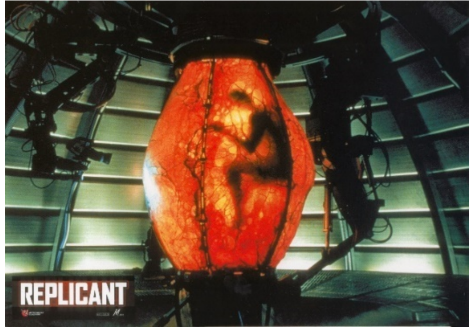




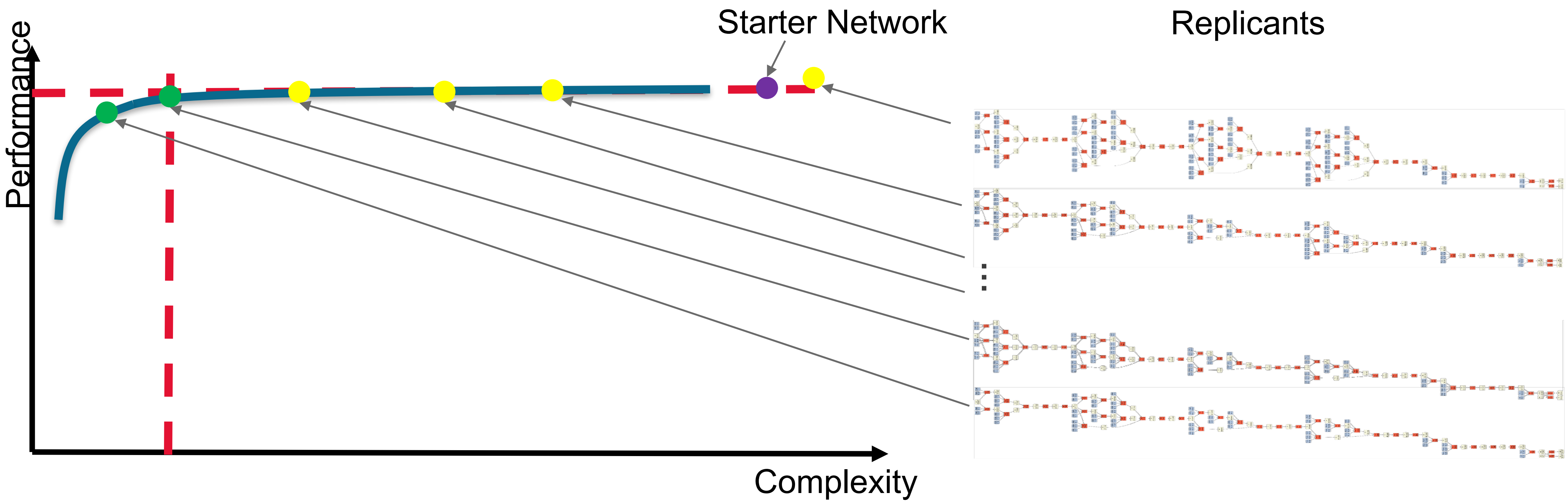
Cactus Module

cadence®

"Replicants"

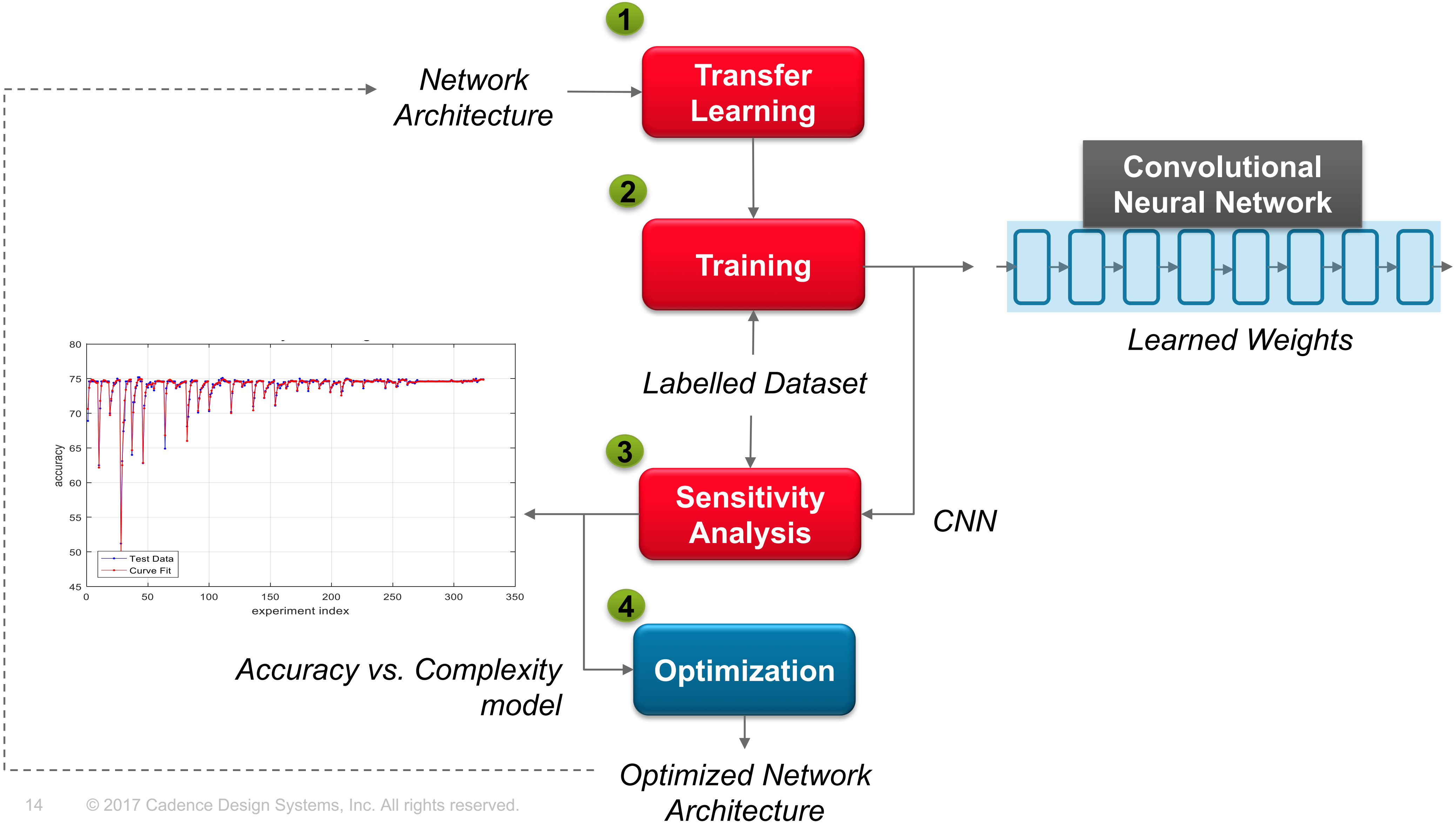


- Move the learned knowledge between networks with different architectures, why?
  - Accelerate the family of networks creation.
  - Keep the empirical cord between family members.



cadence

Cactus Network Compression Procedure

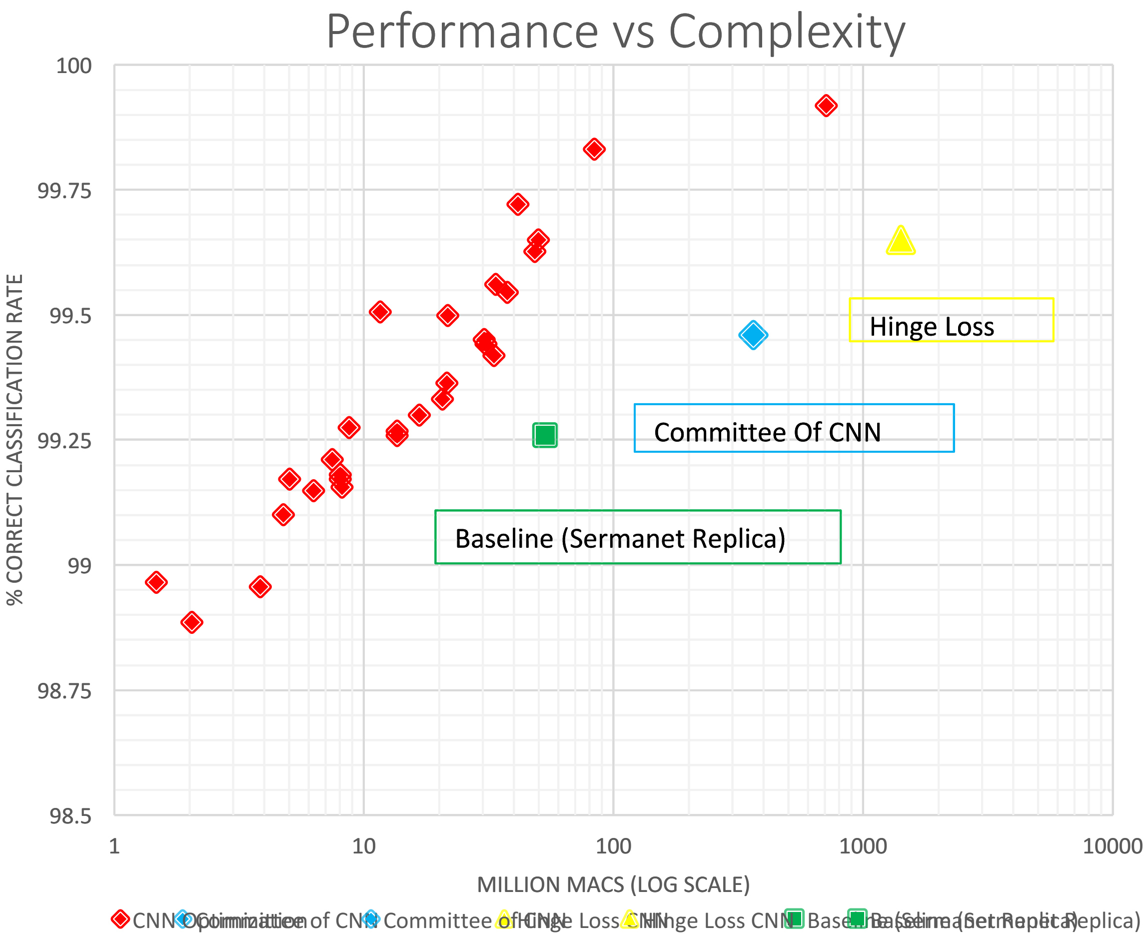


cadence

Example: German Traffic Sign Recognition Benchmark (GTSRB)

- 51840 images of German road signs in 43 classes

- Outperform every other known network on GTSRB.
- At the same performance of the next best network, CactusNet is two orders of magnitude lower complexity.

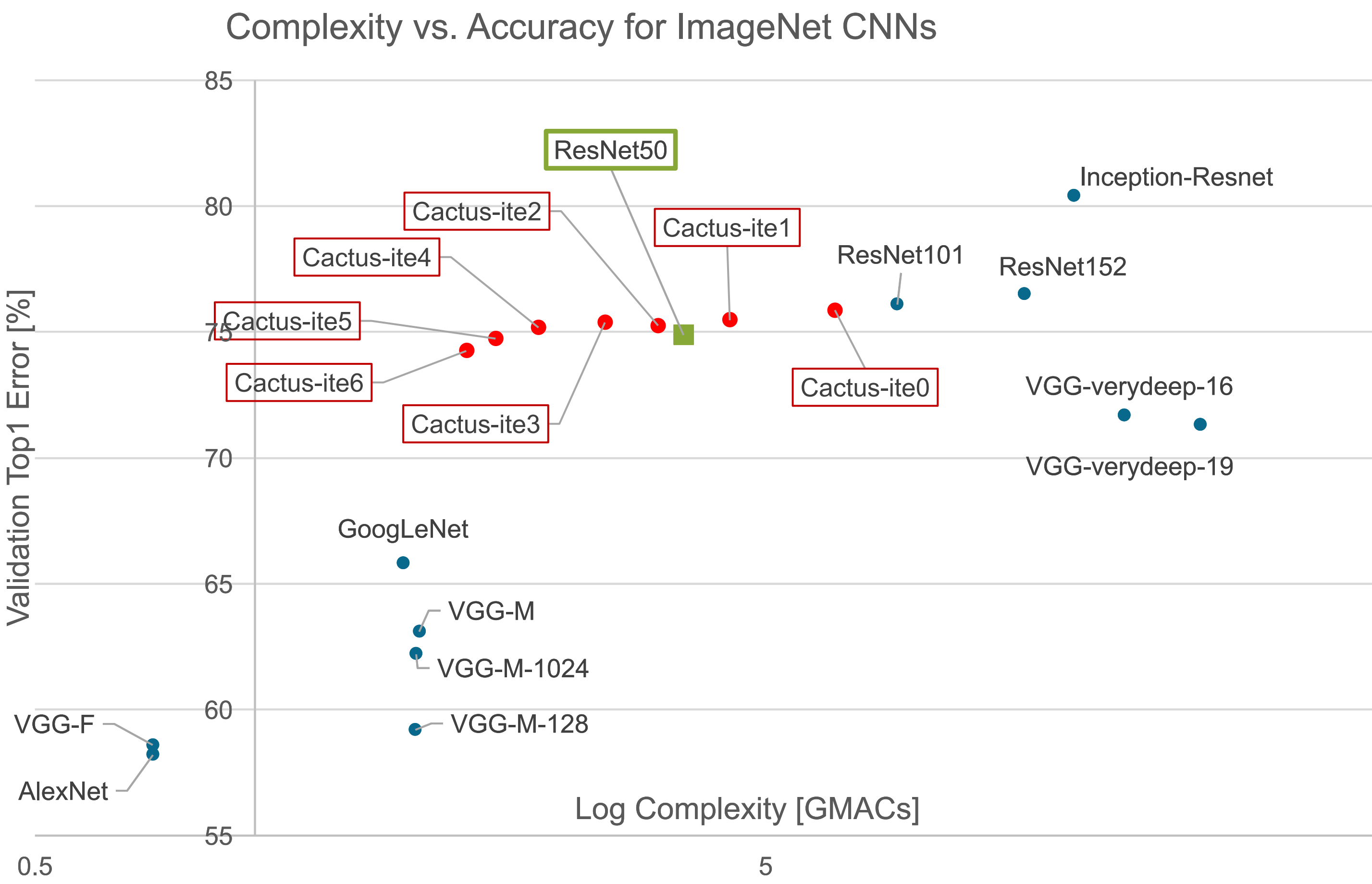


cadence

Results ImageNet (2012)

Set	Num of images	Max size	minsize
train	1281167	3456x2304	60x60
validation	50000	3657x2357	80x60
test	100000	3464x2880	63x84

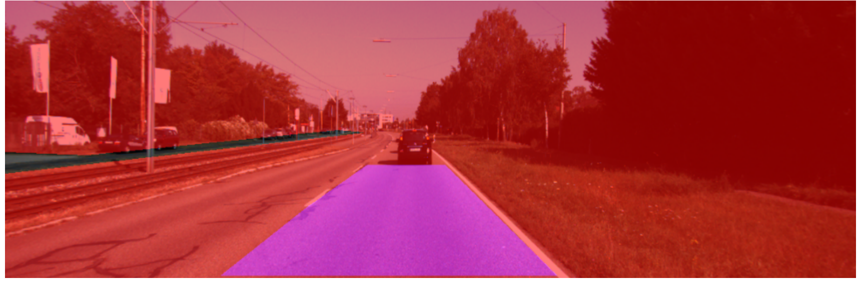
- ResNet 50 has the best Accuracy/Complexity ratio on ImageNet
- Match and outperformed ResNet50 on performance and complexity.
  - Matched ResNet50 performance at less than half of complexity.



1. Optimize network architecture
2. Optimize the problem definition
3. Minimize the number of bits needed to represent the network
4. Utilized and optimized HW for CNN

How Reduce the Problem Size for Pixel Segmentation?

KITTI Road Segmentation Dataset



- 289 training and 290 test 375x1242 images
- In segmentation 375x1242 image, one needs to solve 466k classification problems
  - These are very correlated problems; no need to solve them all.
- We conservatively redefined the ground truth description making the problem to be 22 times smaller.

NW	Precision	Recall	MaxF	Road Accuracy	Overall Acuracy	GMACS[1]
Cadence	95.52%	90.50%	92.94%	94.02%	97.74%	**10.6**
FCN[2]	94%	93.67%	93.83%	X	X	**105.3**
SegNet[3]	X	X	X	97.4%	89.7%	**112.5**

1. GMACs to process input of size 256x1280
2. http://lmb.informatik.uni-freiburg.de/Publications/2016/OB16b/oliveira16iros.pdf
3. https://arxiv.org/abs/1511.00561

cadence

1. Optimize network architecture
2. Optimize the problem definition
3. Minimize the number of bits needed to represent the network
4. Utilized and optimized HW for CNN

Minimizing Number Formats in DNN

- In Adaptive filters design, there are two general approaches to solve this problem
  1. Post training quantization.
  2. During training quantization.
     - This requires changing the training infrastructure and process.
- Redundancy in filters is typically exploited by reducing the total number of multipliers, or reducing the number format.

System	Methodology	AlexNet Top-1 Error [%]	
		32b Float	8b Fixed
Google TensorFlow	Mapping/demapping of coefficients and data between Integer and FLP	42.1	43.4
UC Davis Ristretto	Dynamic FXP Minifloat Multiplier-free (shifts only) *Fine tuning*: FXP forward propagation FLP backward propagation	43.1	43.8
Cadence	Dynamic FXP based on forward propagation statistics	41.2	42.3

cādence

CNN Advanced Quantization

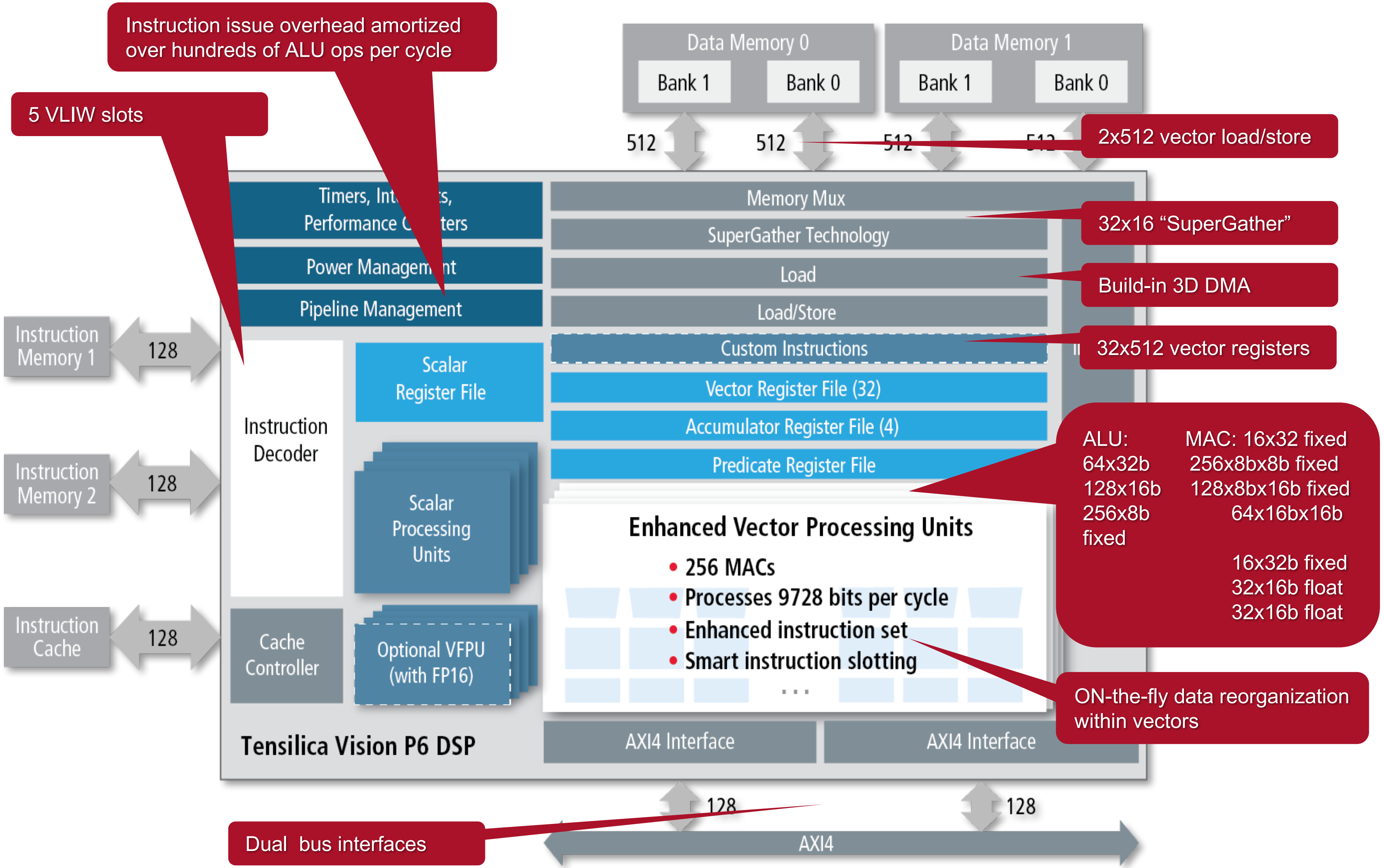
Method	Data	Coefficients	Savings vs 8b x 8b	AlexNet Top1 Error
Conventional	8 bits	8 bits	-	42.3%
Hybrid	8 bits	8 bits: 60% 4 bits: 40%	31%	43.1%

1. Optimize network architecture
2. Optimize the problem definition
3. Minimize the number of bits needed to represent the network
4. Utilized and optimized HW for CNN

Cadence Vision P6 DSP

1. A CNN optimized HW is designed to
   1. Minimized pJ/MAC
   2. Minimize data movement,
   3. Have sufficiently large MAC per sec
   4. Assure high utilization of available HW resources.

Cadence Vision P6 DSP



cādence

In summary

Take Away Points

- For CNN to achieve it s full potentials, 2 to 3 order of magnitude power efficiency improvements will be needed.
- This efficiency have to come from optimized SW/algorithms and HW IP.
- Cadence is paving the way for the semiconductor industry to offer optimized CNN capable products.
- Near future will bring us new products that would bring computer vision closer to reality.