Orchestrating a brighter world

NEC

Deep Learning with Vector Processor

NEC Labs America: Machine Learning Department

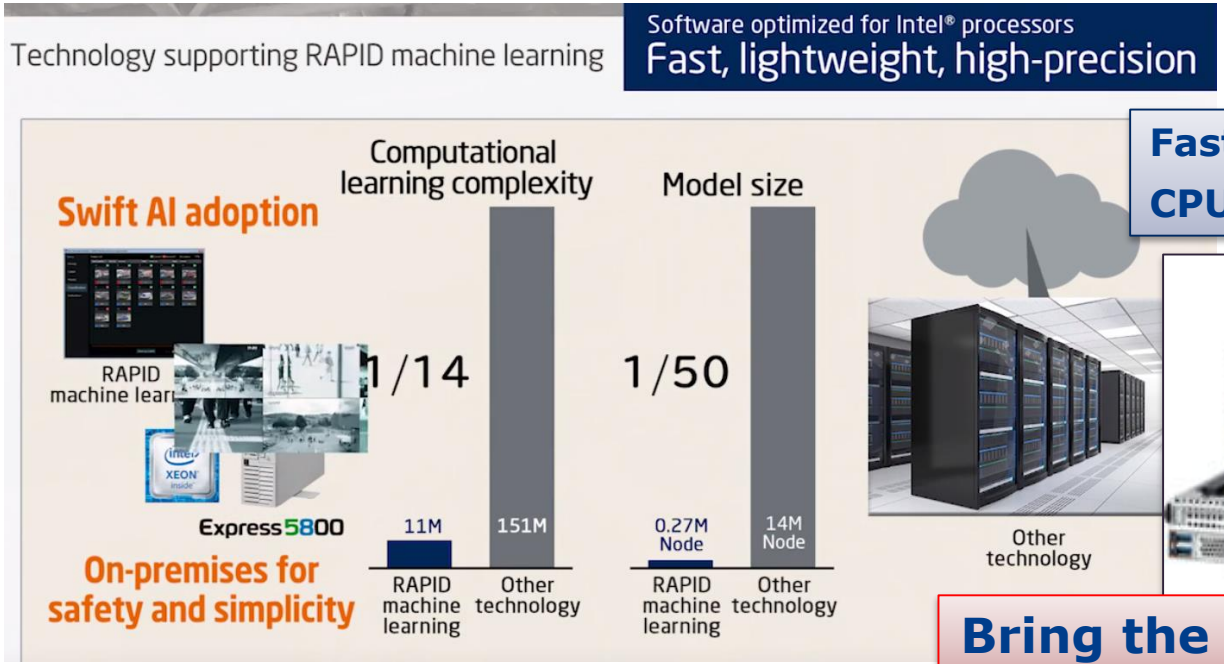
November 12, 2018

www.nec-labs.com

NEC's Machine Learning Platform: RAPID

Application Templates: Inspection, Time series analysis, Surveillance, Risk Analysis

Latest release: RAPID 2.2, July 2018



Fast, light weight, high precision

CPU, GPU, soon: VECTOR

Bring the algorithms to the data!

YouTube video: https://www.youtube.com/watch?v=sFfJngldFJs

Examples of Commercial Applications

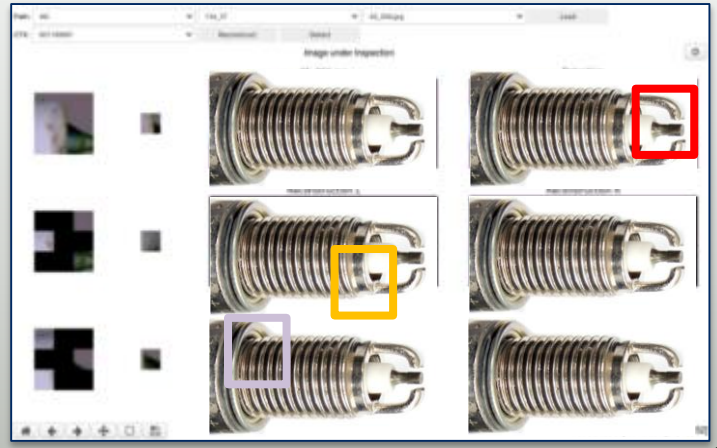
System for Unsupervised Defect Detection on Manufactured Objects

Introduced June 2017

http://jpn.nec.com/press/201706/20170621_04.html

This system can be applied for a wide range of inspection tasks for quality control of manufactured products.

Challenge: Few training data of defects



System for Spoof Detection at Tokyo Stock Exchange

Introduced March 2018

https://jpn.nec.com/press/201803/20180319_02.html

System identifies trading patterns that indicate illegal activities. This reduces drastically the amount of trades that have to be checked by experts.

売数量	値段	買数量
1,000	104	
1,000	103	
2,000	102	
3,000	101	
	100	3,000
	99	50,000
	98	1,000
	97	1,000

AI: From Deep Learning to Reasoning

Deep Learning is pattern matching. Next generation: Abstract reasoning

Approach: Structured Network Learning

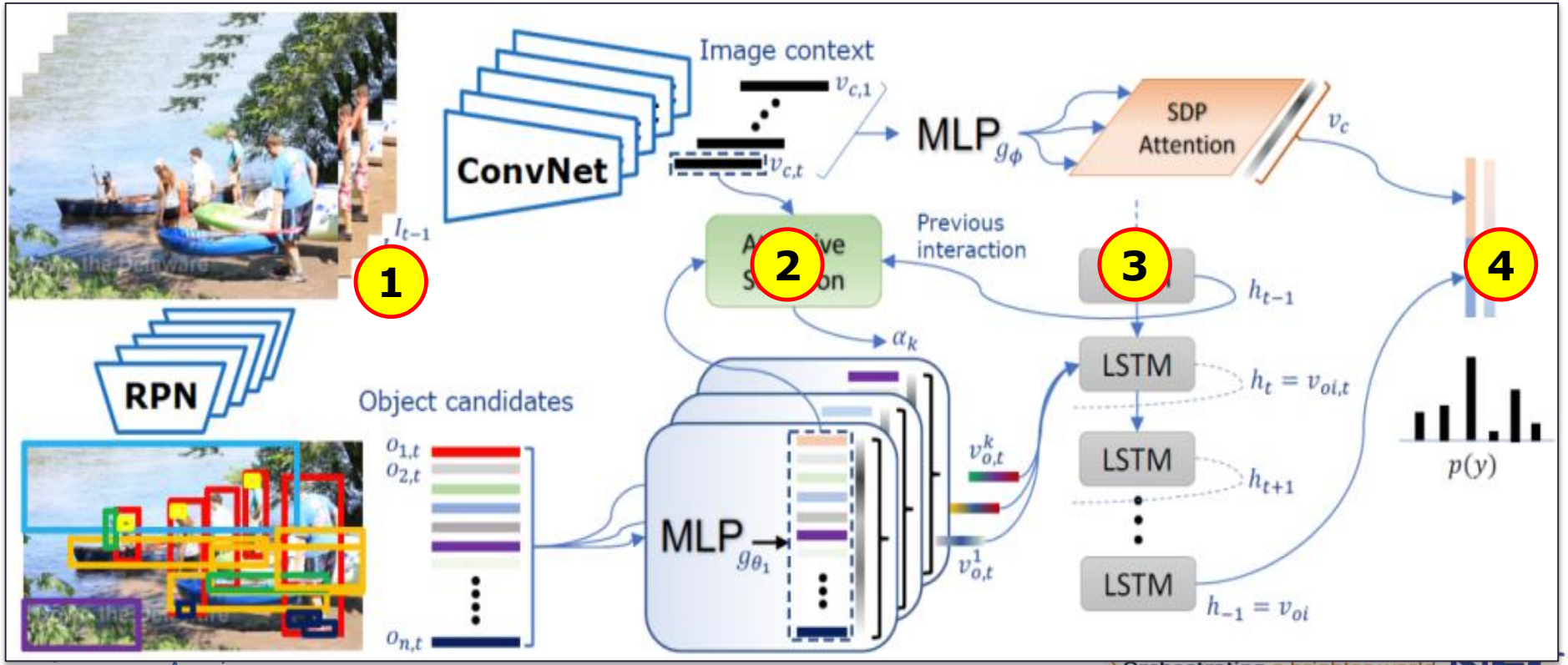
Analyze spatial and temporal relations between objects

A group of people get off of a yellow school bus with life rafts around their neck.
Relationships: [group of people, get off, yellow school bus], [group of people, with, life rafts]

"Attend and Interact: Higher-Order Object Interactions for Video Understanding"; C. Ma; A. Kadav; I. Melvin; Z. Kira; R. Hasan; H. Graf; CVPR 2018

Structured Network Learning

1. Object localization (CNN) → 2. Focus of attention (MLP, LSTM) →
3. Interactions of objects (spatial and temporal) → 4. Interpretation (MLP)



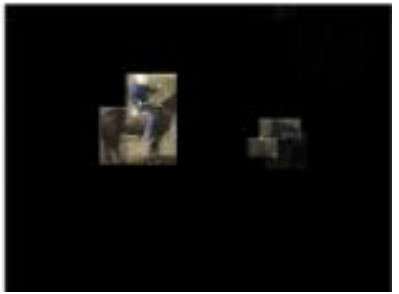
Focus of Attention + Object Relations

Example: All images show a horse and a person. But actions are different
Key: Interpretation of relations of objects







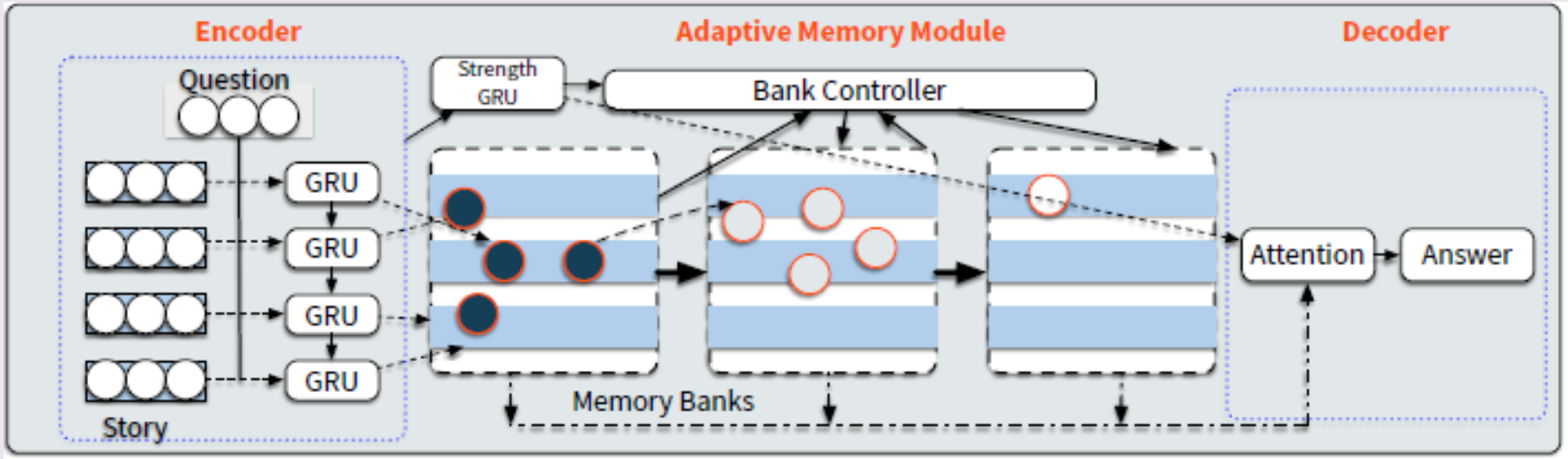


Riding Brushing Play Polo Tie up


WBC SEMIFINALS - LOS ANGELES
USA 0
JPN 0
0-1 2OUT
8 PITCHES
catching or throwi
hitting baseball 16.98
hitting baseball 18.37
catching or throwing baseball 15.73

Issue: How to Deal With Large Data Sets

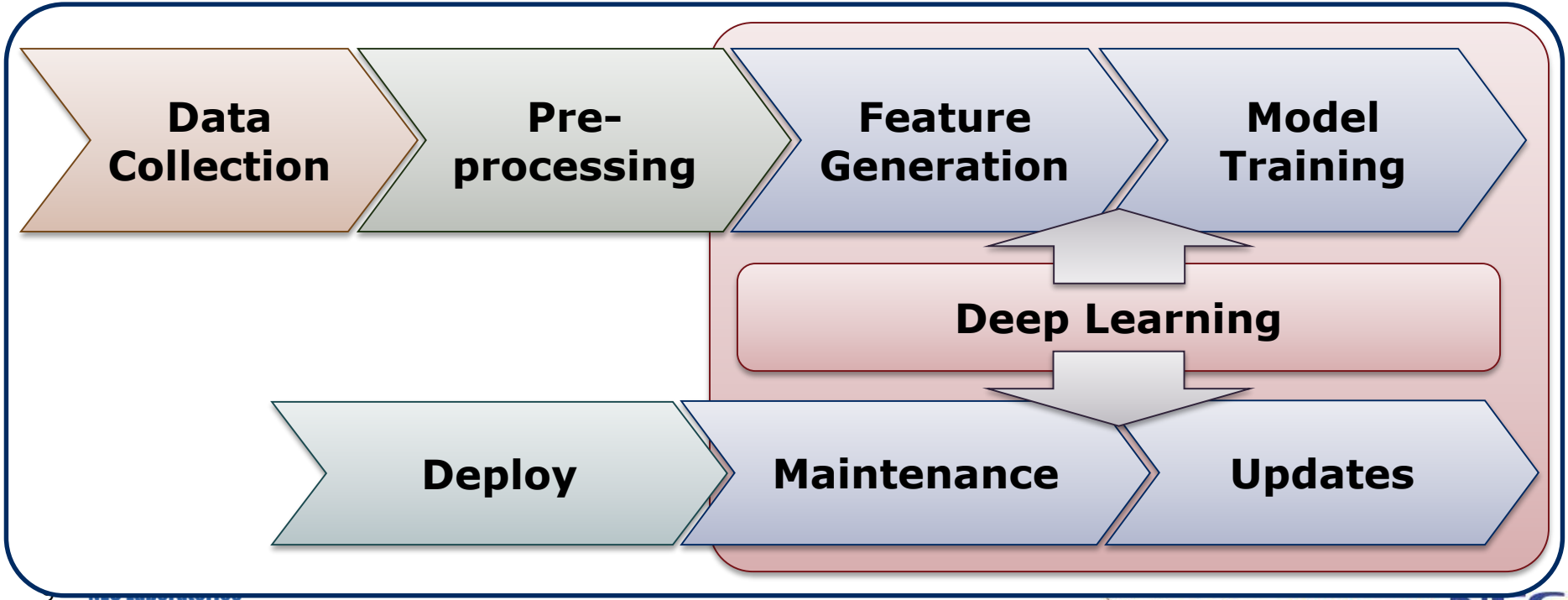
Problem: Deep Learning networks are inefficient in storing large amounts of data. E.g. cannot handle answering questions over thousands of entities.
Solution: **Adaptive Memory Networks.** Question guided network construction with entities stored in multiple memory banks based on distance from the question.
Paper: Adaptive Memory Network: ***D. Li, A. Kadav; ICLR 2018.***



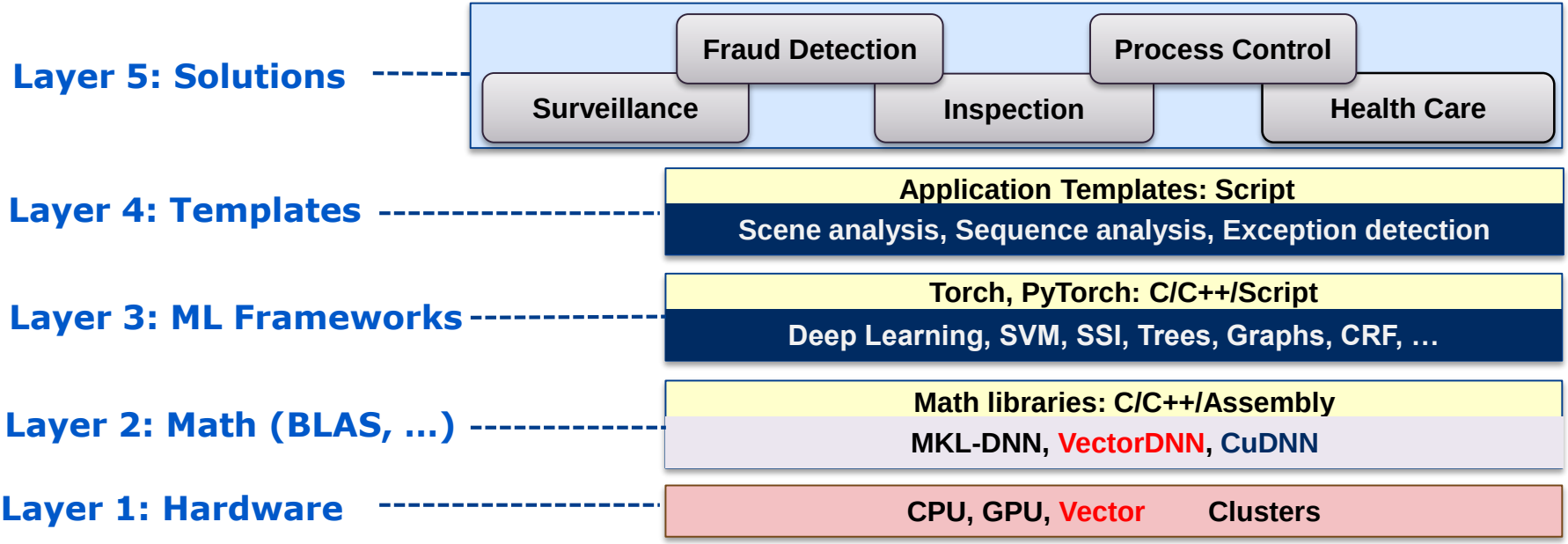
Data Analytics Work Flow

All functions must be supported to run applications efficiently

Therefore an architecture that is too specialized is usually not the best



Software Architecture



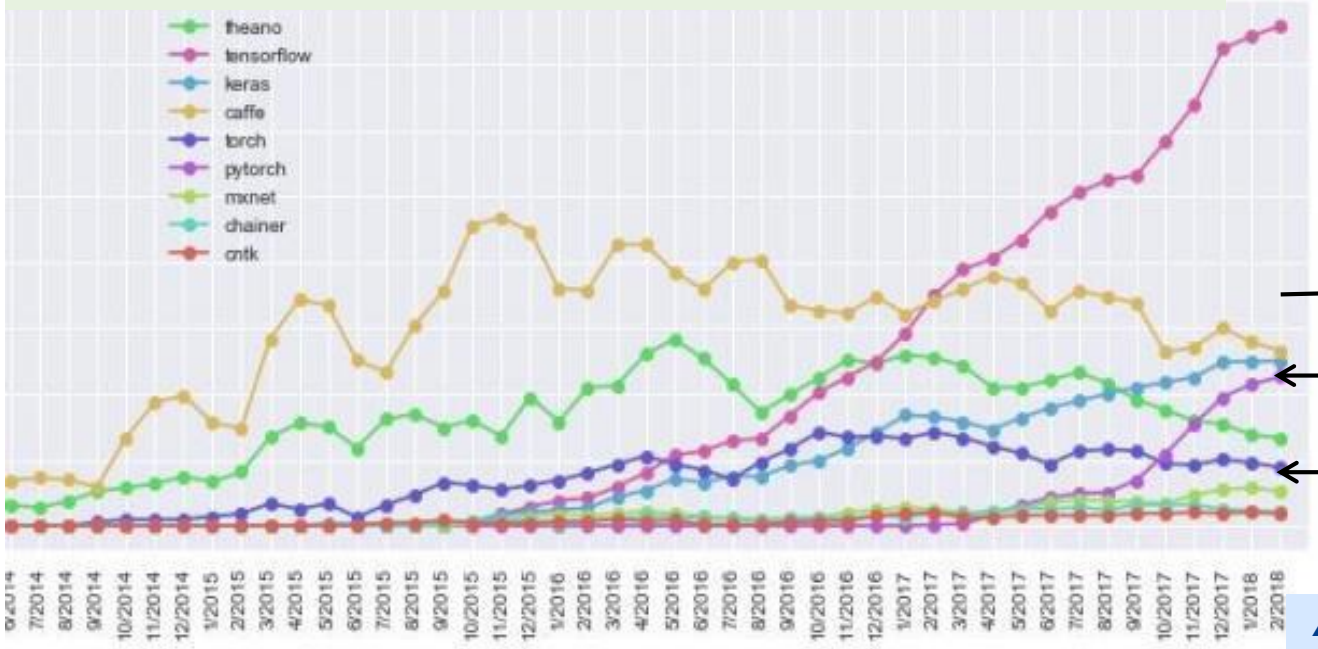
Other Critical components:

- Tools to analyze the conditions at the customer premise
- Automated evaluation of the data. Understand which Templates to use.
- Good user interface for easy deployment of solutions even by non-experts.

Machine Learning Frameworks

- Torch compiled for SX Aurora TSUBASA. Passes all tests
- Torch runs on Vector Processor. Off-loading in development

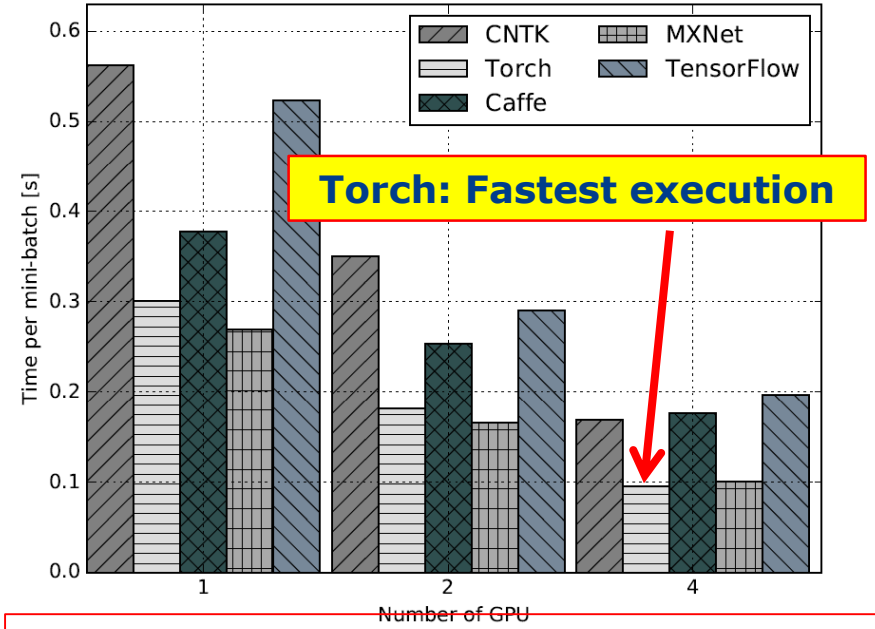
Use of ML frameworks: Number of papers that mention:

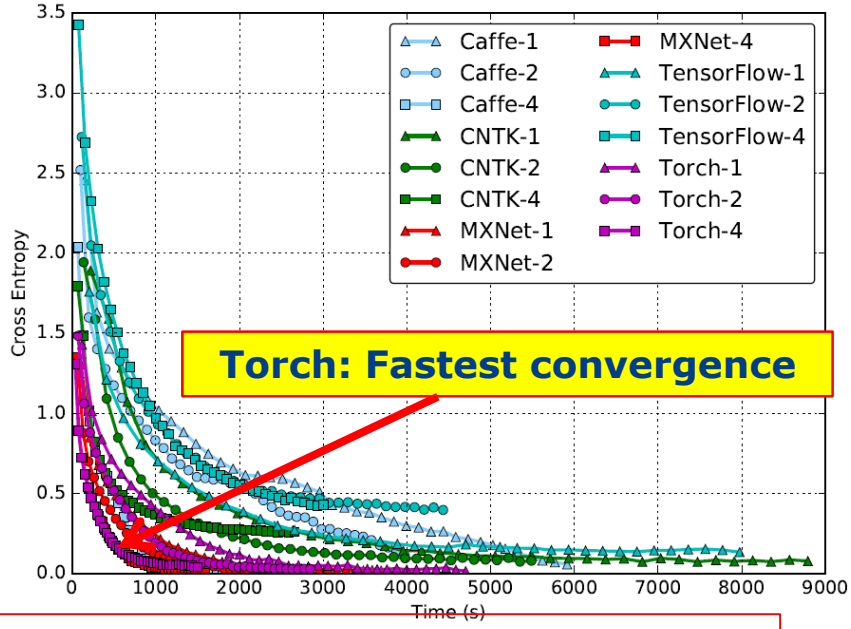


Torch: Some Benchmarks

Torch often leads worldwide competition in tests with CPU and multiple GPU's (1, 2, 4 GPU's)
Competitors here: Torch (NEC), TensorFlow (Google), CNTK (Microsoft), Caffe, MXNet

Ref: Benchmarking State-of-the-Art Deep Learning Software Tools; S. Shi, Q. Wang, P. Xu, X. Chu; Feb 2017





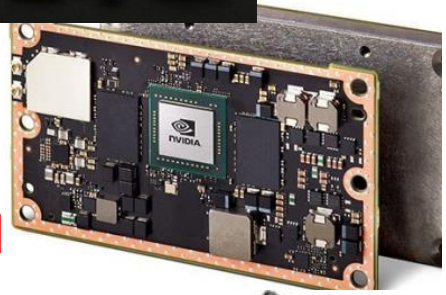
Performance comparison of ResNet-56; tested on 1, 2, 4 GPU's. http://dlbench.comp.hkbu.edu.hk/

NEC Laboratories America
Relentless passion for innovation

Orchestrating a brighter world NEC

Running on Different Hardware Platforms

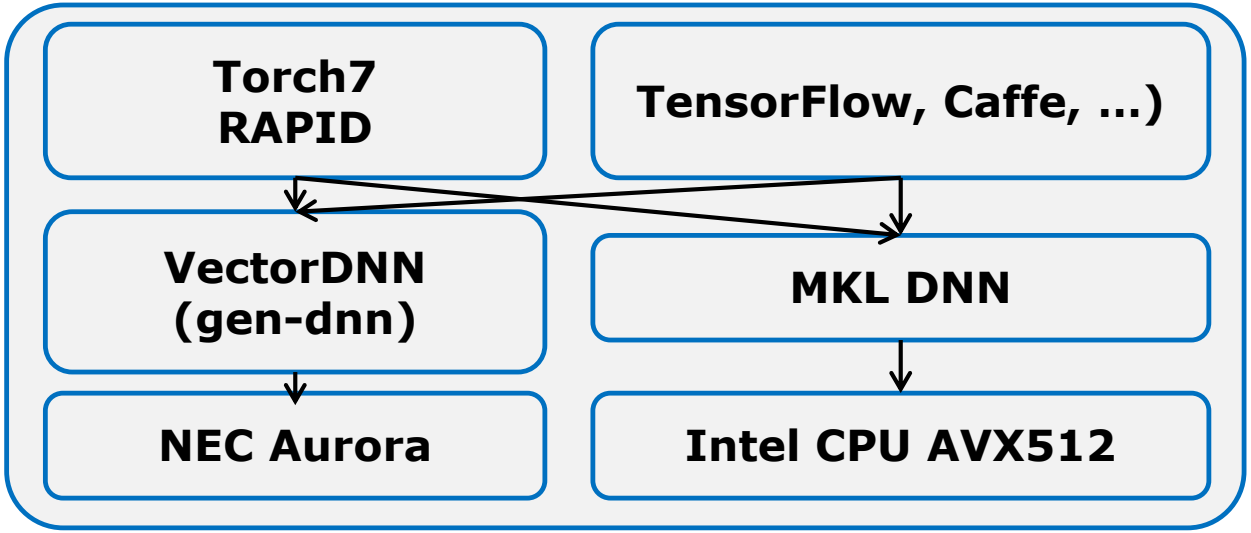
- **CPU:** Efficient parallelization
- **GPU:** Efficient multi-GPU parallelization
- **Embedded:** Cars, machining centers, …
  - **RaspberryPi**
  - **NVIDIA Jetson TX2**
- **Vector processor: Aurora with VectorDNN**
  - **Edge: Compact high-performance**
  - **Data Centers: Efficient parallelization**
- **FPGA**



VectorDNN: API compatible with MKL-DNN

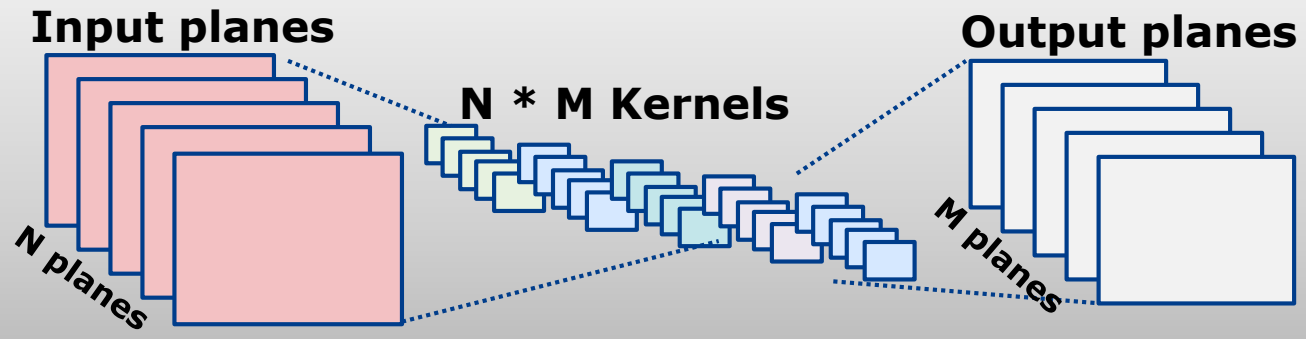
VectorDNN is designed to integrate easily with Data Analytics frameworks

Branch of MKL-DNN: https://github.com/necla-ml/gen-dnn



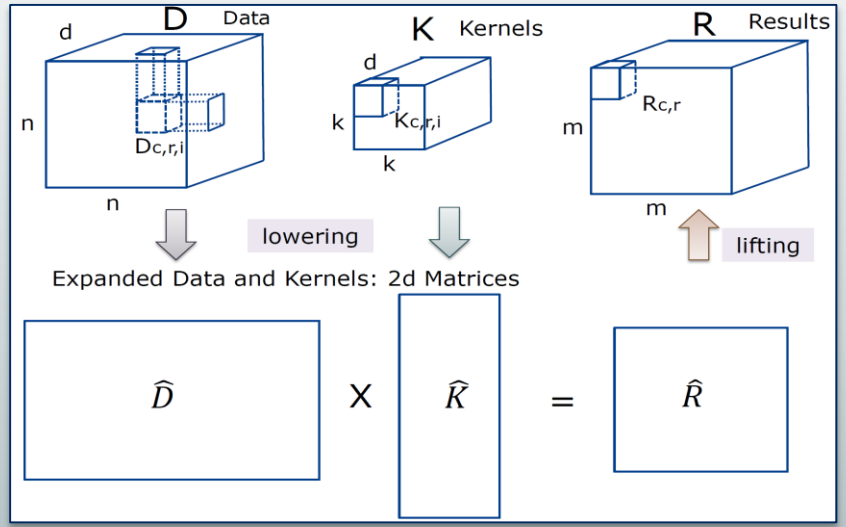
- **Passes all functional tests: benchdnn**
- **Torch compiles**
- **Performance optimizations are ongoing: JIT; optimized off-loading**

Vectorization of Convolutions



Nested loops: 8 or more

Tiling: introduces more loop levels, but allows consistent implementations across all layers



Tested 25 ways of lowering and lifting

Example: im2col

- Expensive lowering, cheap lifting
- Often used; often not optimal, but usually decent performance without specific tuning.
- Makes use of 'gemm'

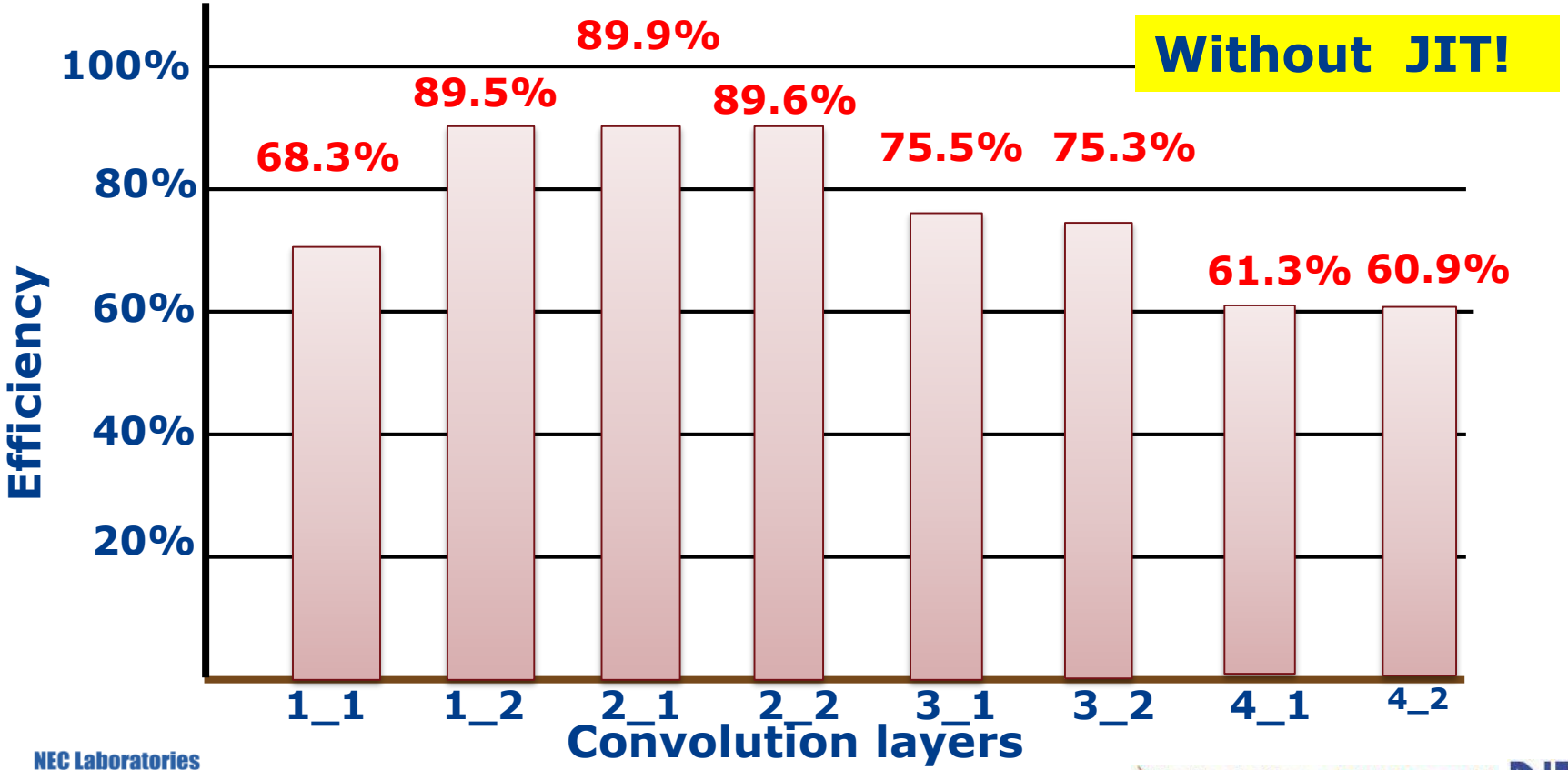
Optimizations for Deep Learning

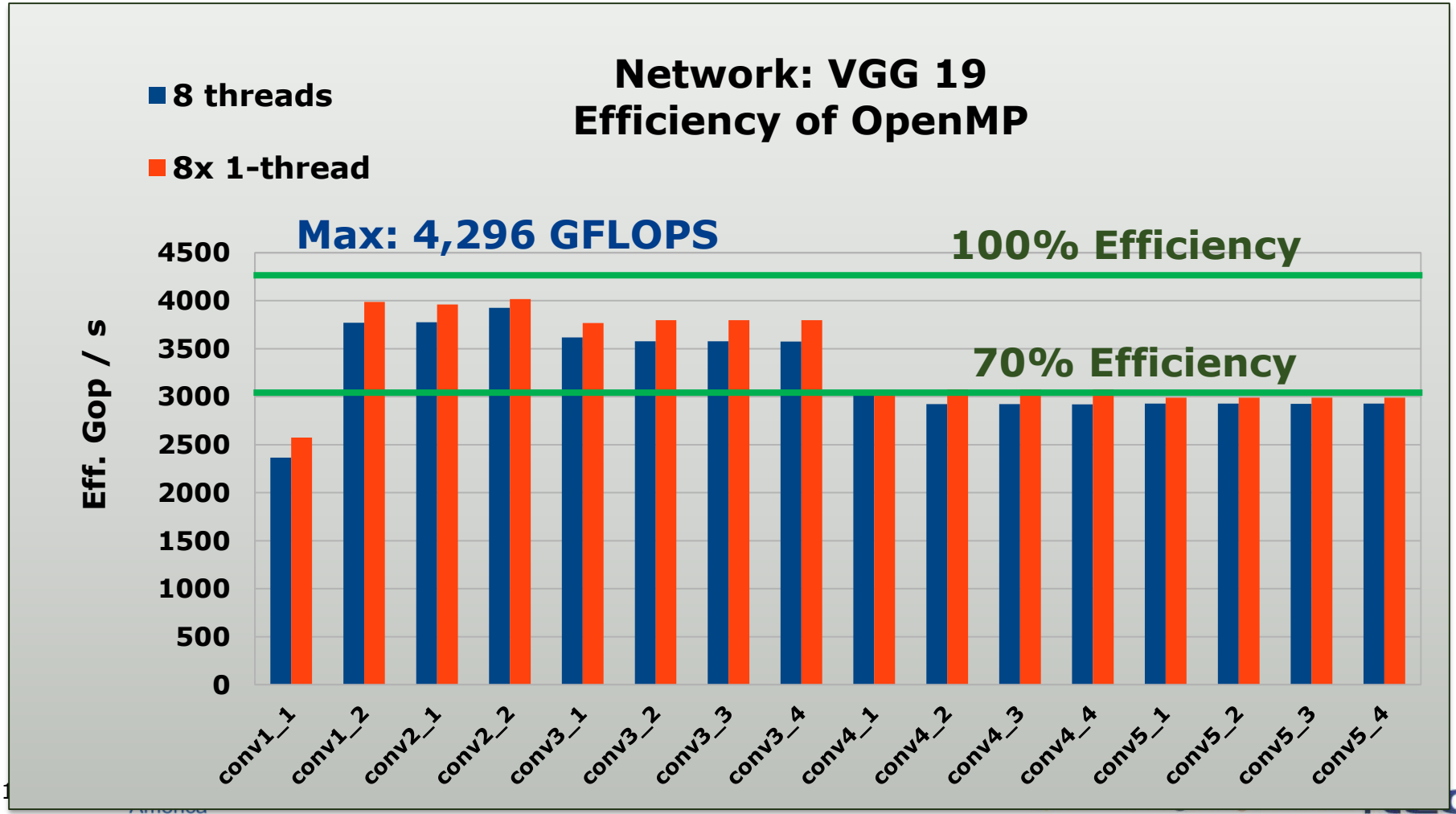
- Beside execution of the convolutions there are a number of operations to be considered.
- Most common operations are provided

Feature	Vector-DNN	Description
Interleaving	Yes	Specialized data layout
Strided	Yes	Every n'th image pixel
Padding	Yes	Pad image border with zeros
Groups	Yes	Kernel channels != image channels
Dilation	Yes	Kernel applied to dilated image region
Minibatch	Yes	Convolve multiple separate images at once
Bias	Yes	Learn bias for convolution layer output
Merging	Yes	Combine convolution with nonlinearity (Relu, …)
Forward	Yes	Forward with bias integrated
Backward	Yes	Backward wrt. data, weights, or both.

Compute Efficiency for Layers of VGG 11

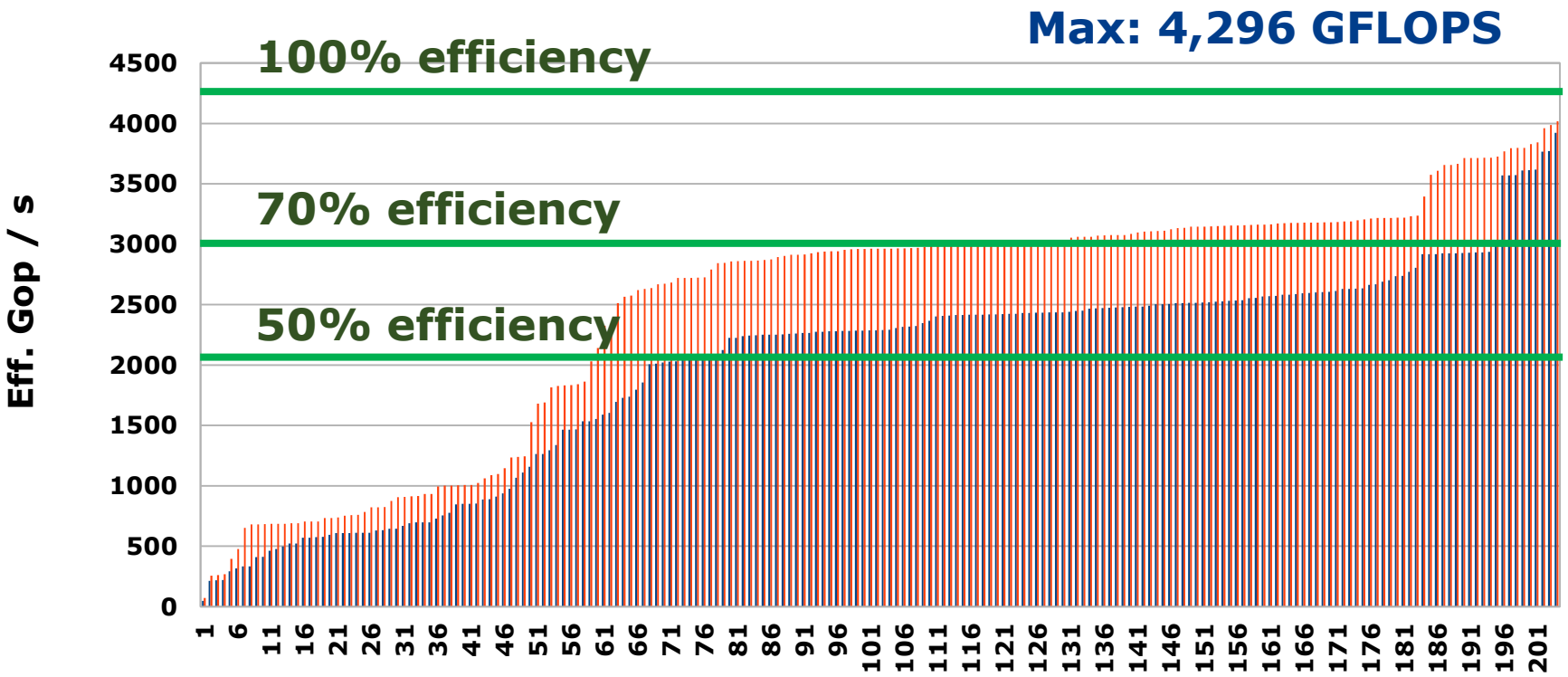
1 thread and 8 threads: Very similar efficiency → good scaling




Network: VGG 19
Efficiency of OpenMP
8 threads
8x 1-thread
Max: 4,296 GFLOPS
100% Efficiency
70% Efficiency
Eff. Gop / s
4500
4000
3500
3000
2500
2000
1500
1000
500
0
conv1_1
conv1_2
conv2_1
conv2_2
conv3_1
conv3_2
conv3_3
conv3_4
conv4_1
conv4_2
conv4_3
conv4_4
conv5_1
conv5_2
conv5_3
conv5_4

Tests with: gen-dnn, vednn

204 layers (benchdnn): alexnet, vgg, resnet, gogglenet



NEC Laboratories America
Relentless passion for innovation

Orchestrating a brighter world NEC

Conclusions

Deep Learning:

- **Vector architecture is flexible and well suited for convolution layers as well as for the other types of layers (good SIMD problem)**
- **Any of the fastest algorithms can be implemented efficiently on the vector architecture (Winograd, Cook-Toom, …).**
- **A high efficiency of ~70% is achievable with little effort.**
- **Higher efficiency with tuning (JIT).**

Data analytics is more than deep learning → key: Flexibility

- **Large memory and high data bandwidth makes vector processor attractive for large scale data analytics.**
- **Can integrate a wide range of algorithms with deep learning. Extensive library to support such functions.**
  - **e.g. collaborative filtering: Sparse matrix factorization, SVD.**