A 2.9 TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems

Giuseppe Desoli, Nitin Chawla, Thomas Boesch, Surinder-pal Singh, Elio Guidetti, Fabio De Ambroggi, Tommaso Majo, Paolo Zambotti, Manuj Ayodhyawasi, Harvinder Singh, Nalin Aggarwal

Presenter Danilo Pau

MPSoC'17, Annecy July 7th 2017

life.augmented

MEMBER OF TECHNICAL STAFF

Outline

Background
Chip architecture
DCNN mapping
Hardware accelerators
Physical implementation
Results
Demo

Outline

- **Background**
- **Chip architecture**
- **DCNN mapping**
- **Hardware accelerators**
- **Physical implementation**
- **Results**
- **Demo**

Universal approximation theorem (Cybenko, 1989, Hornik, 1991)

For any target function

$$y = f^*(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d$$

(which is continuous and Borel measurable)

and any $\varepsilon > 0$ there exists parameters

$$h \in \mathbb{Z}^+, \boldsymbol{W} \in \mathbb{R}^{h \times d}, \ \boldsymbol{w}, \boldsymbol{c} \in \mathbb{R}^h, c \in \mathbb{R}$$

this is the dimension of the hidden layer: it is a ***parameter*** *in the theorem*

such that the ***(shallow) feed-forward neural network***

$$\tilde{y} = \boldsymbol{w} \cdot g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{c}) + c$$

approximates the target function by less than ε

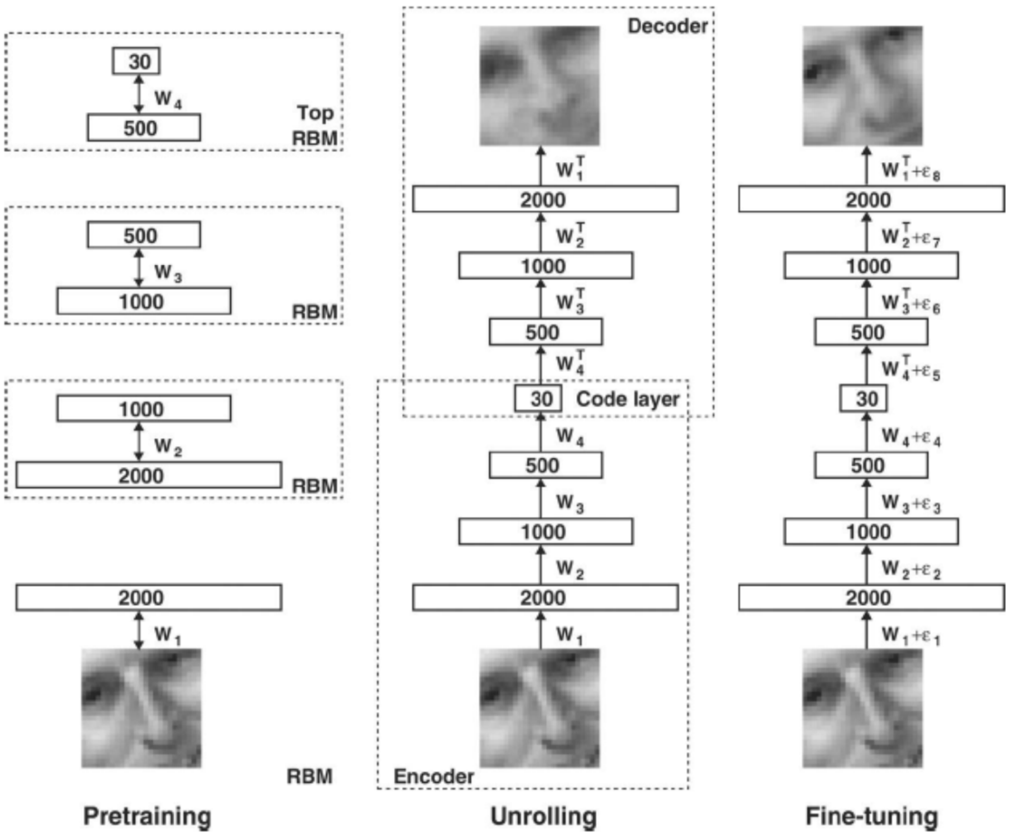
$$| f^*(\boldsymbol{x}) - \boldsymbol{w} \cdot g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{c}) + c | < \varepsilon$$

(on a compact subset of $\mathbb{R}^d$)

life.augmented

Reducing the Dimentionality of Data with Neural Networks (Hinton & Salakhutdinov, Science, 2006)

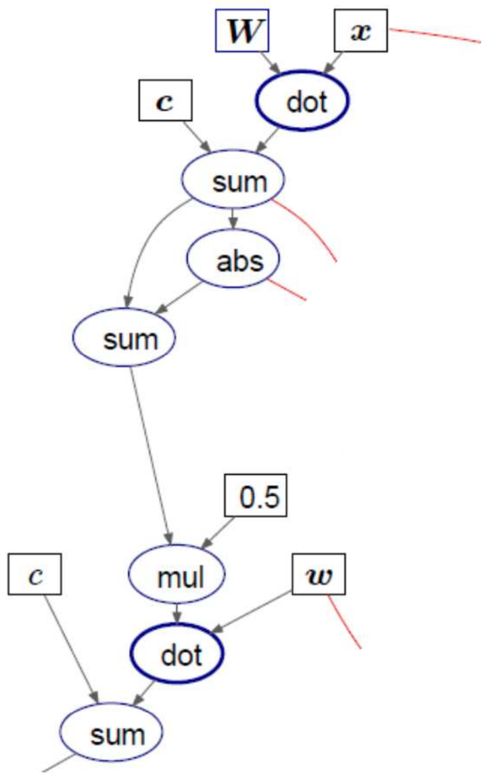
- **Generative model**: probabilistic mathematical description, it allows using the network in both directions ‘bottom-up’ and ‘top-down’.



"It has been obvious since the 1980s that backpropagation through deep autoencoders would be very effective for nonlinear dimensionality reduction, provided that computers were fast enough, data sets were big enough, and the initial weights were close enough to a good solution."
[G. Hinton et al., Science, 2006]

"All three conditions are now satisfied. Unlike nonparametric methods, autoencoders give mappings in both directions between the data and code spaces, and they can be applied to very large data sets because both the pretraining and the fine-tuning scale linearly in time and space with the number of training cases."
[G. Hinton et al., Science, 2006]

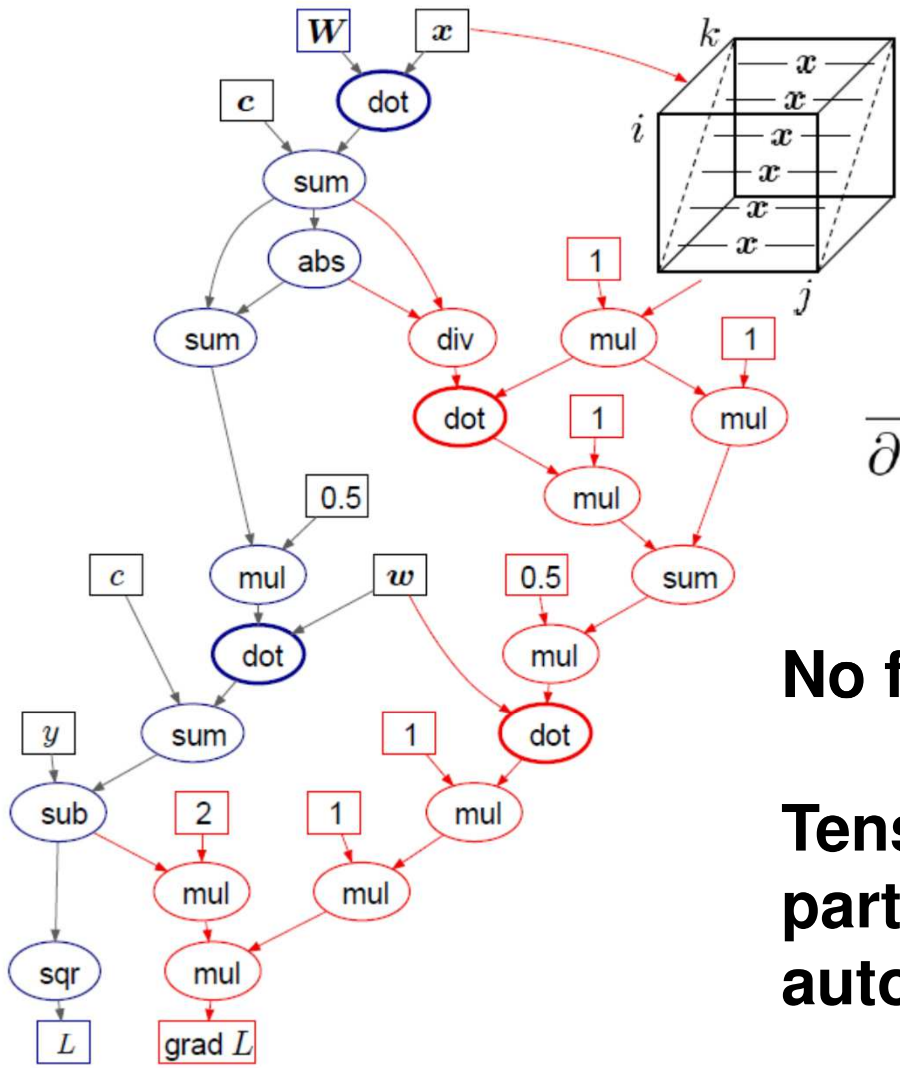
A simple network



$$(\boldsymbol{w} \cdot \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{c}) + c$$

Courtesy M. Piastra, Universita di Pavia, 2017

That can be trained by computing gradients



$$\frac{\partial}{\partial \boldsymbol{W}}(\boldsymbol{w} \cdot \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{c}) + c - y)^2$$

No fear!!!!

TensorFlow, Theano, CNTK do partial derivatives computation automatically for us

Courtesy M. Piastra, Universita di Pavia, 2017

Background

Chip architecture

DCNN mapping

Hardware accelerators

Physical implementation

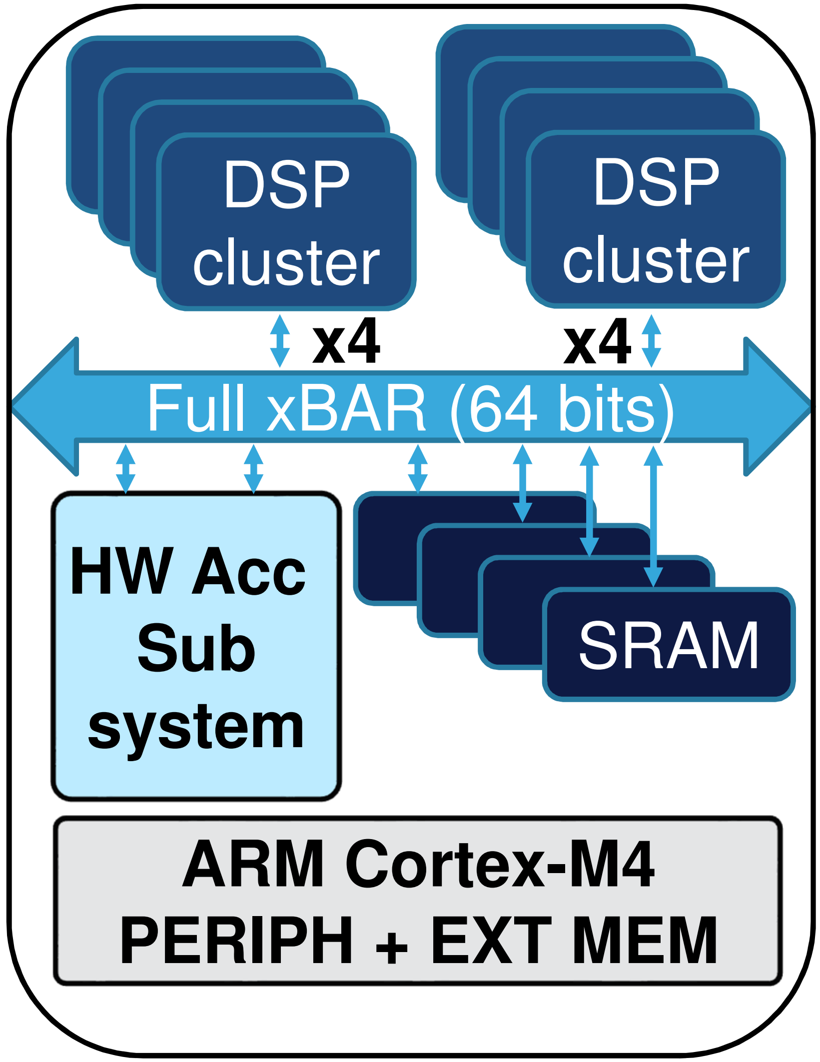
Results

Demo

A complete SoC for DCNN applications

Enabling use of DCNNs in embedded systems:

- Power efficiency, low cost
- Efficient memory hierarchy
- Flexibility to adapt to different DCNN topologies
- Input/output capability
- Integrated with state of the art Deep Learning tools



life.augmented

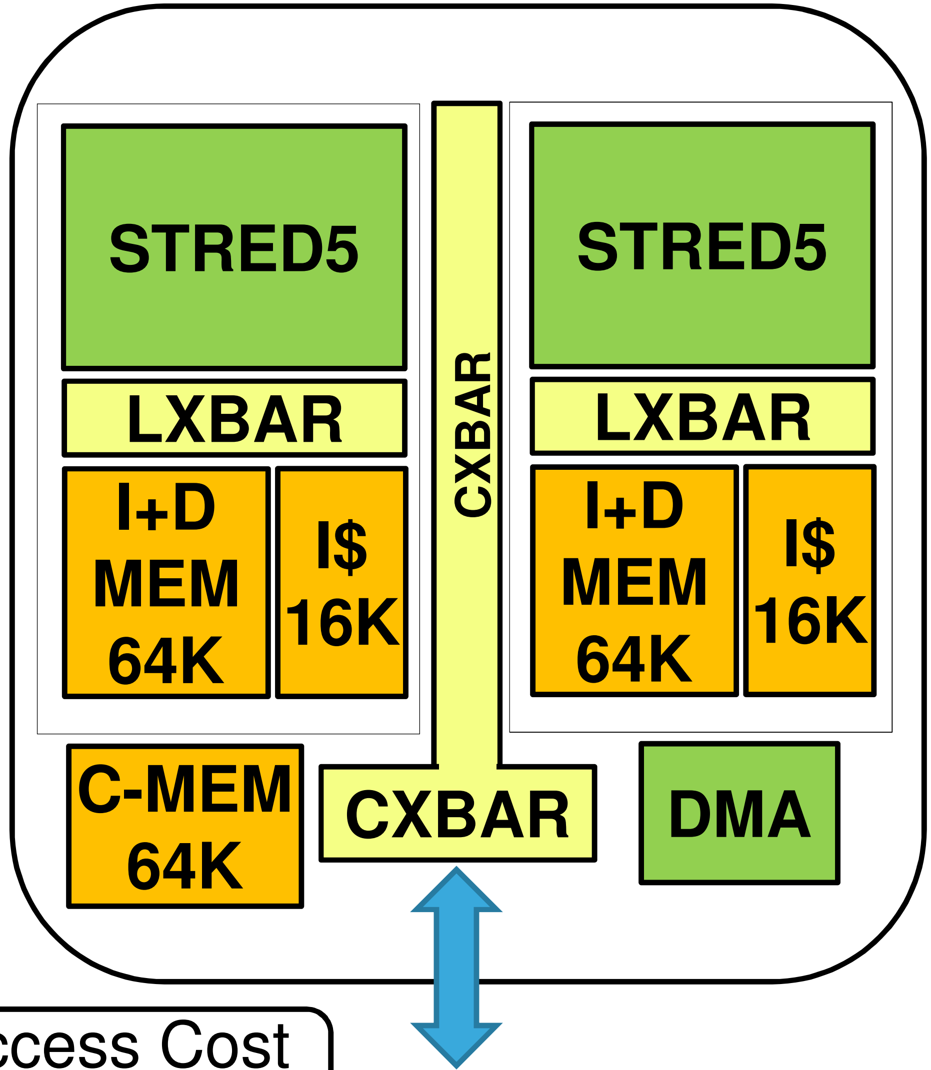
DSP Cluster

(6uW/MHz@0.6V) up to 1GHz in ST FD-SOI 28nm technology

ISA extensions for DCNN execution

8 DSP clusters, each with 2 32-bit DSPs, 4-way 16KB I-Cache, 64KB Local RAM and a shared 64KB RAM

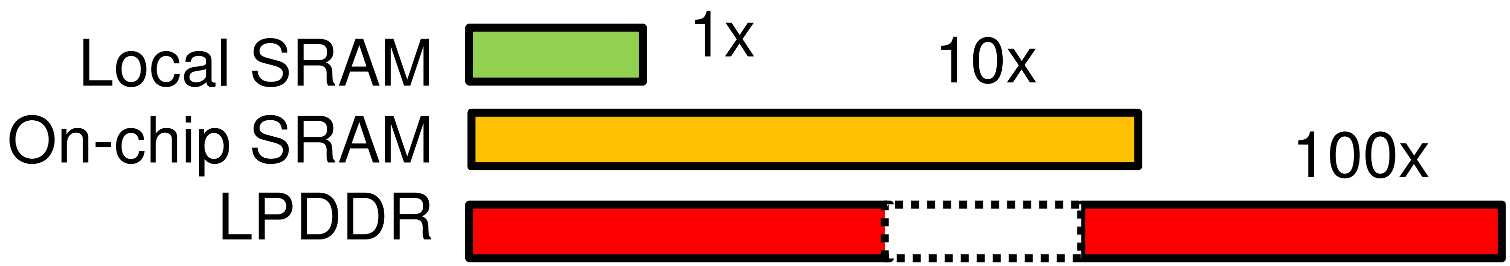
2D-DMA with independent channels, linked list, stride, padding, etc.



Non Uniform Memory Access Cost
Uniform Memory Address Space

Memory Hierarchy

Energy/power x word access



4 MB (4x16x64KB) of shared RAM grouped with a 64 bits bus port x bank to sustain peak DCNN throughput

Each 64KB cut has individual sleep line control to de/activate it on demand and reduce active power consumption

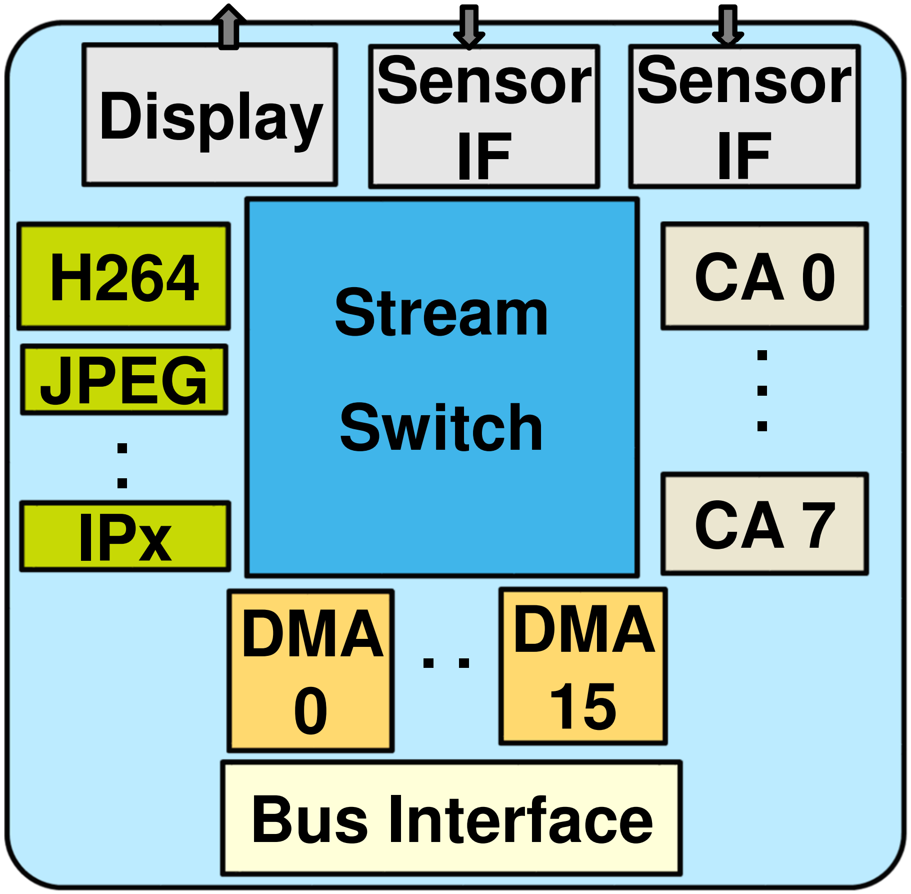
L2 SW controlled cache for feature maps and parameters

HW Accelerators subsystem

8 Conv Accelerators
16 CDNN specialized streaming DMA engines

Configurable framework supports data-flow based processing

Additional IPs
H264 SP@ML, MJPEG, 2 Census, 2 croppers, Corner detector, 4 color converter, 4 sensors input IFs, 1 DVI output IF, digital MIC array IF



Background

Chip architecture

DCNN mapping

Hardware accelerators

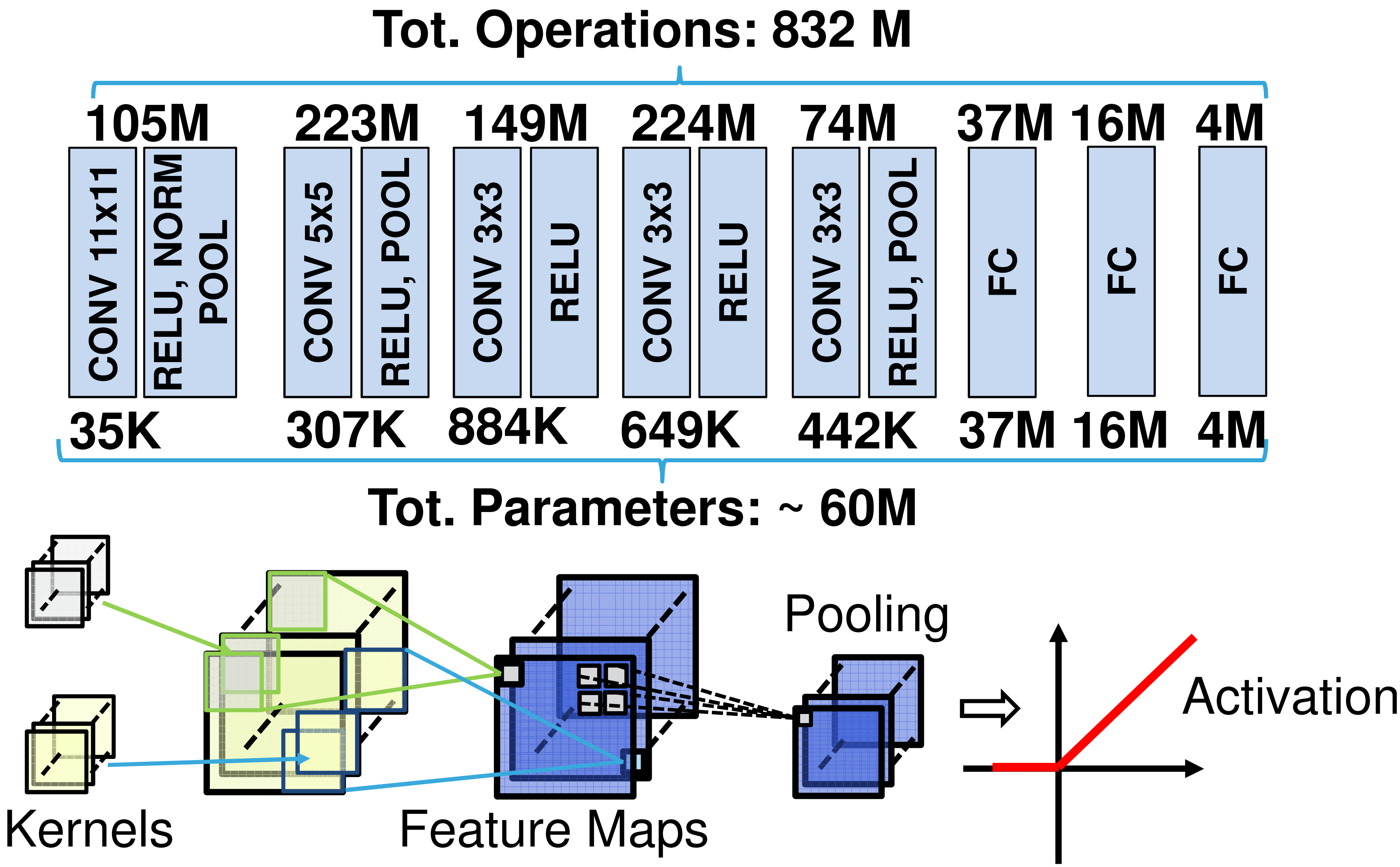
Physical implementation

Results

Demo

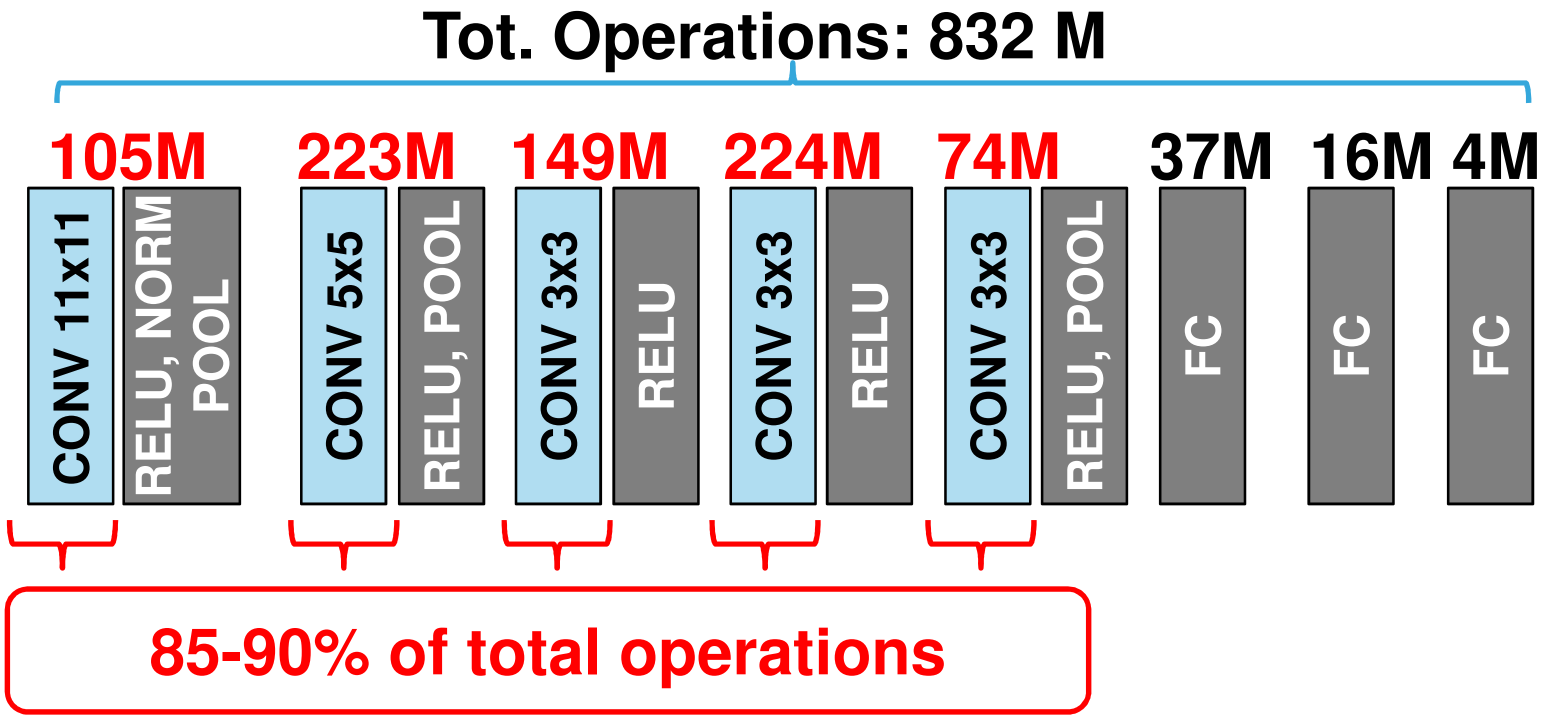
life.augmented

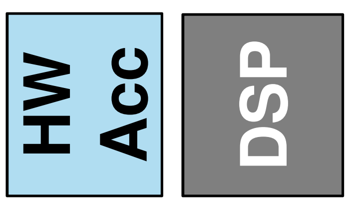
AlexNet basics



Krizhevsky et all, NIPS 2012

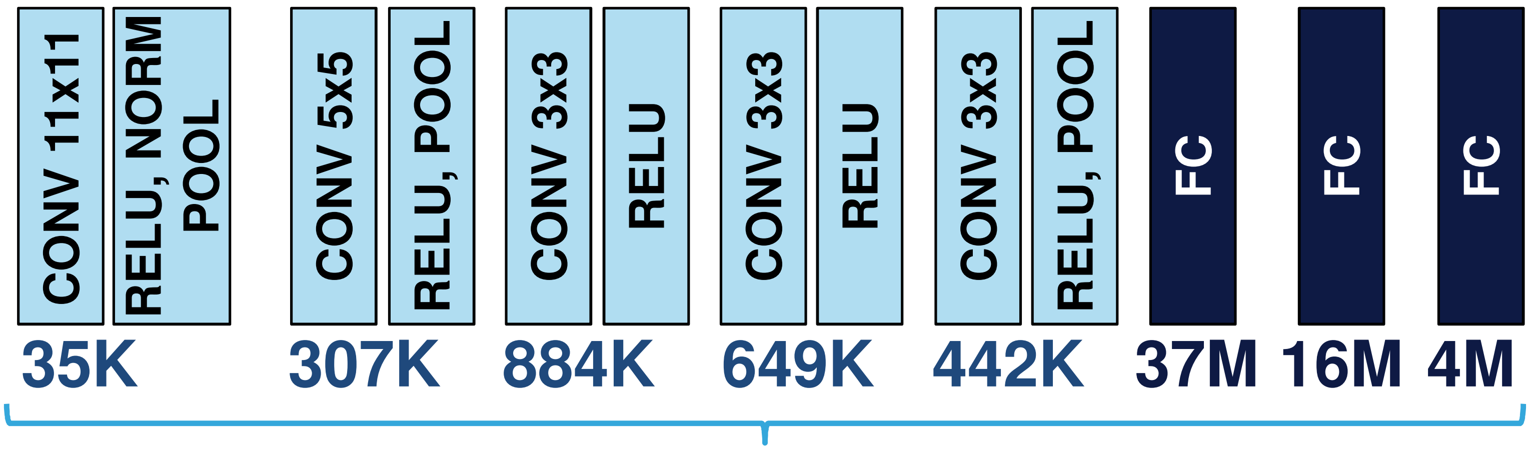
AlexNet HW/SW partitioning





- CONV layers: 1 Conv Acc → 36 MACs per cycle
- Non conv layers to DSPs to accommodate DCNN future evolution (leaky RELU, etc.)

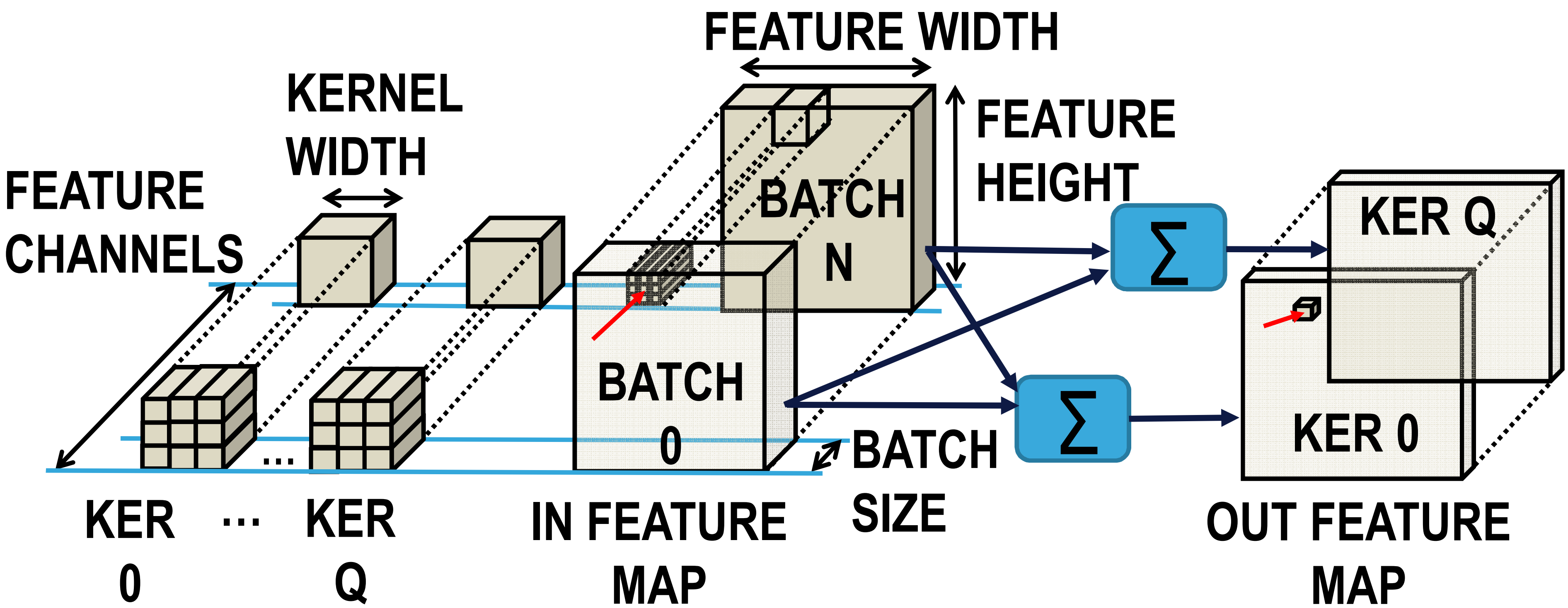
AlexNet memory footprint



Tot. Parameters: ~ 60M

- On-chip SRAM
  - **2318 KB** for parameters (8 bits non linear quant)
  - **1436 KB** for feature maps (16 bits)
- ~10 MB of external RAM for FC layers (compressed)

Logical to physical mapping



Feature maps and kernels are sliced into batches processed iteratively and results are accumulated

Batch size set x layer
Matching features and kernels parameters to HW resources and ceilings

Background

Chip architecture

DCNN mapping

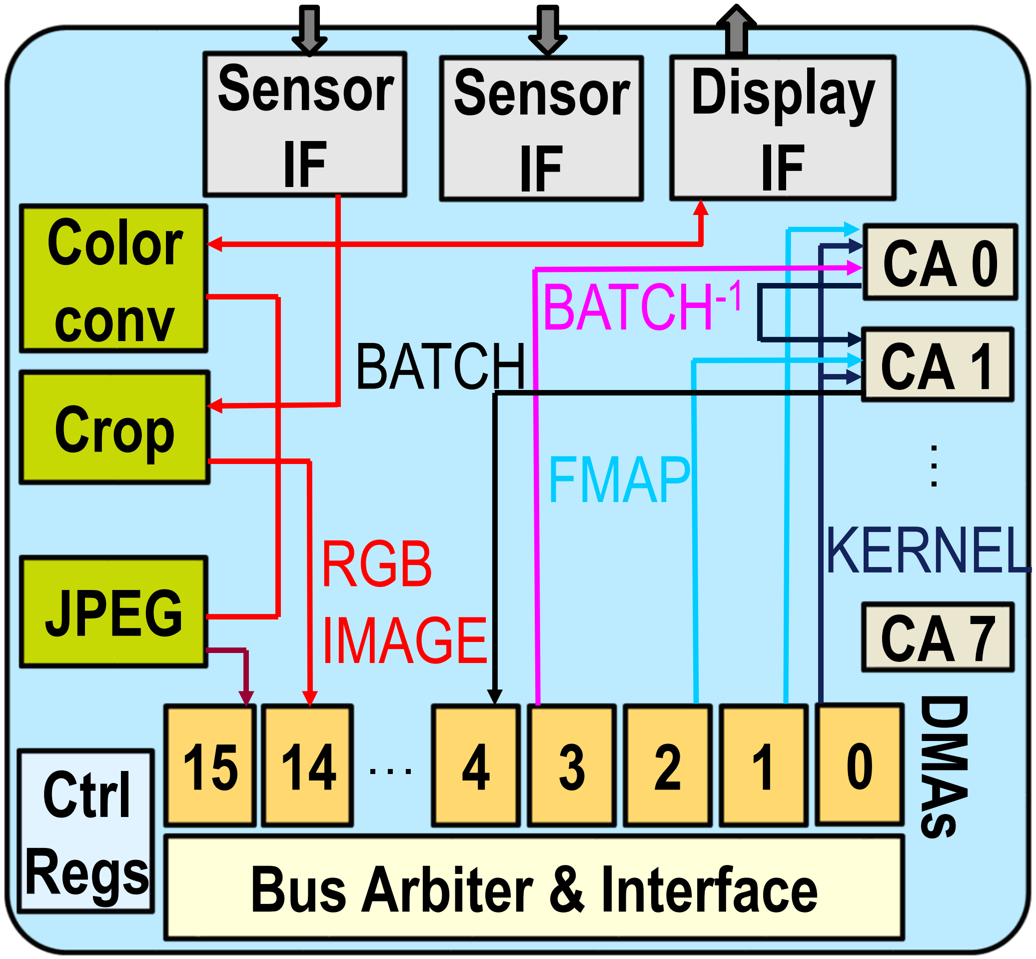
Hardware accelerators

Physical implementation

Results

Demo

Reconfigurable Accelerator Framework



DESIGN

Design time parametric

STARTUP

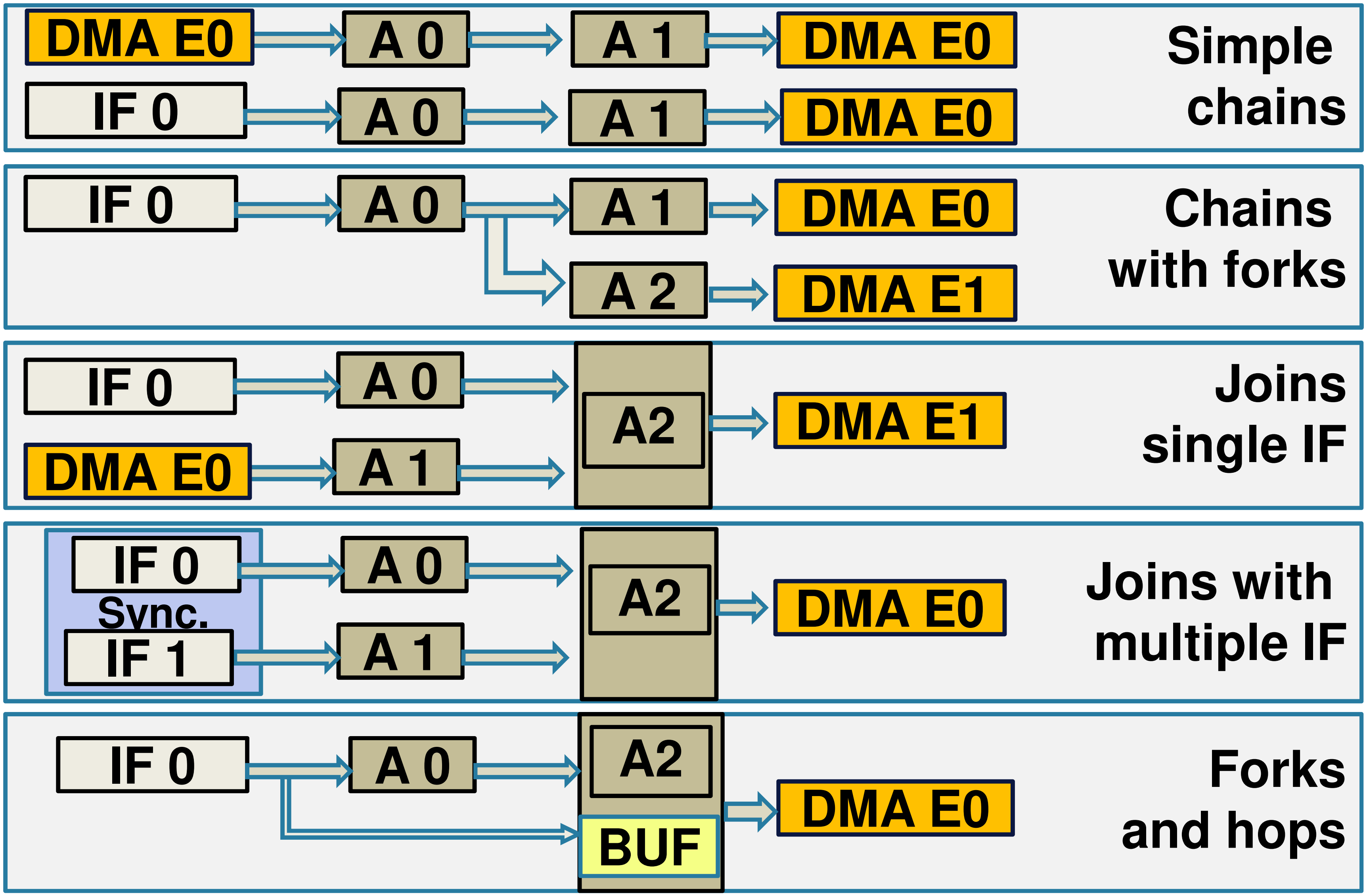
Unidirectional stream links create ad-hoc processing chains.

RUNTIME

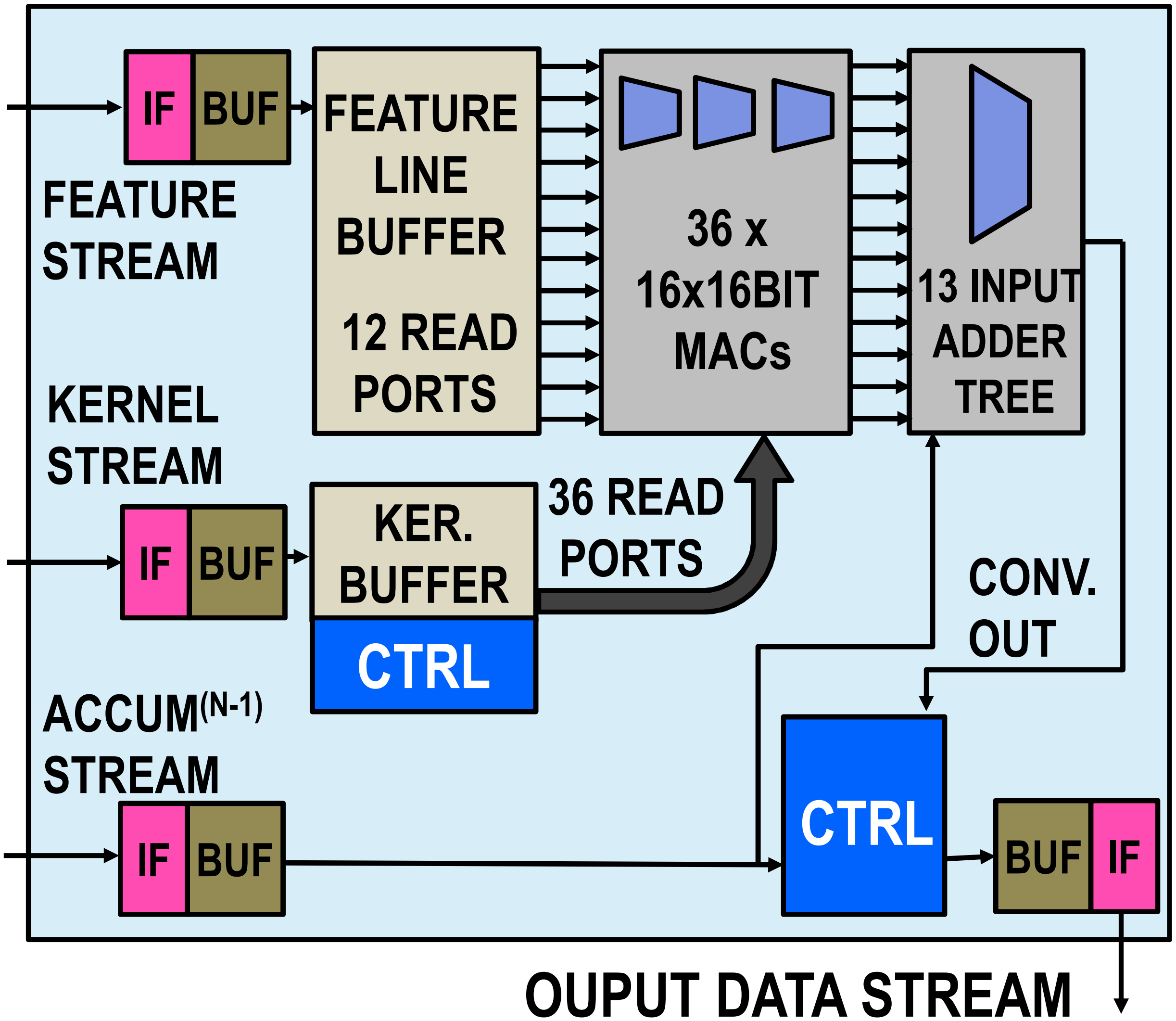
DMA engines run autonomously through linked lists and synch up with the DSP clusters

Reconfigurable stream switch to achieve Data Flow Graph

Virtual Stream Links

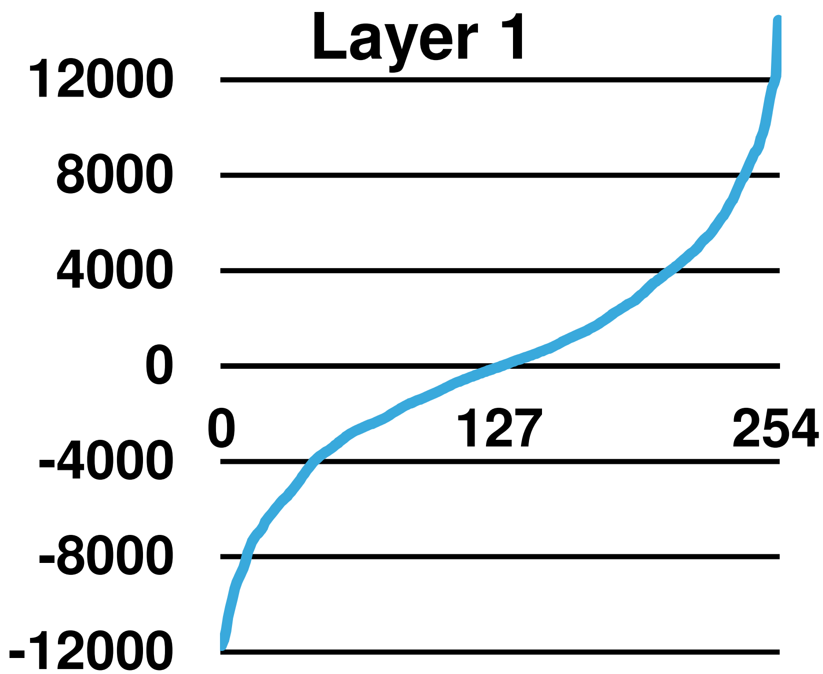


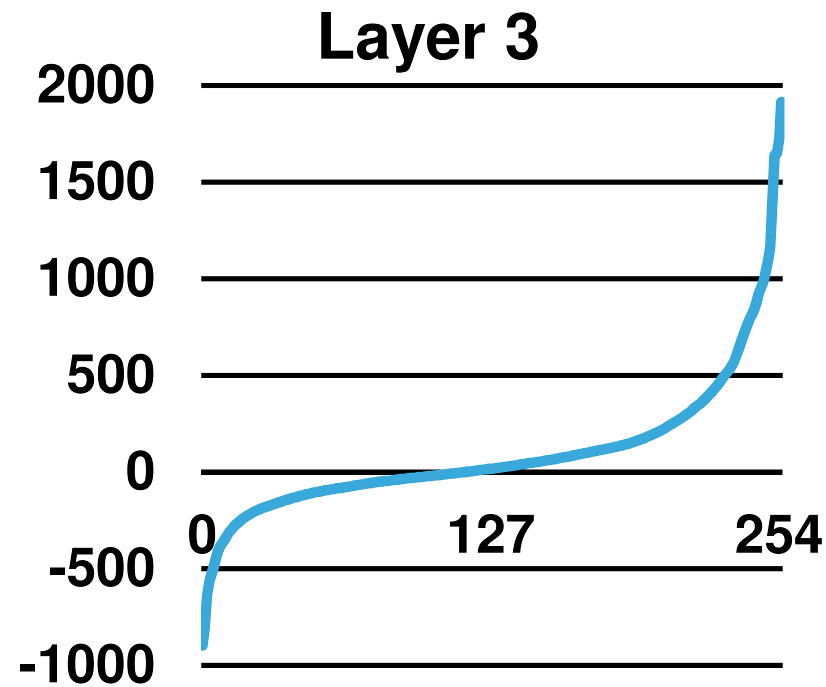
DCNN Convolution Accelerator



Kernels	1x1 to 12x12 8bit => 16 bit up to 4 in parallel
Batch size	Up to 16
Feat. size	512: ker > 6x6 1024: ker > 3x3 2048: others H/V padding H/V stride
MACs	16 bits inputs 48 bits accum. scaling, rounding, saturation

Parameter Compression





- Kernel weights can be quantized non linearly with 8 of fewer bits (e.g. with KNN),
- Convolution Accelerator supports decompression in HW
- AlexNet top-1 classification error rate increase of 0.3%

Background

Chip architecture

DCNN mapping

Hardware accelerators

Physical implementation

Results

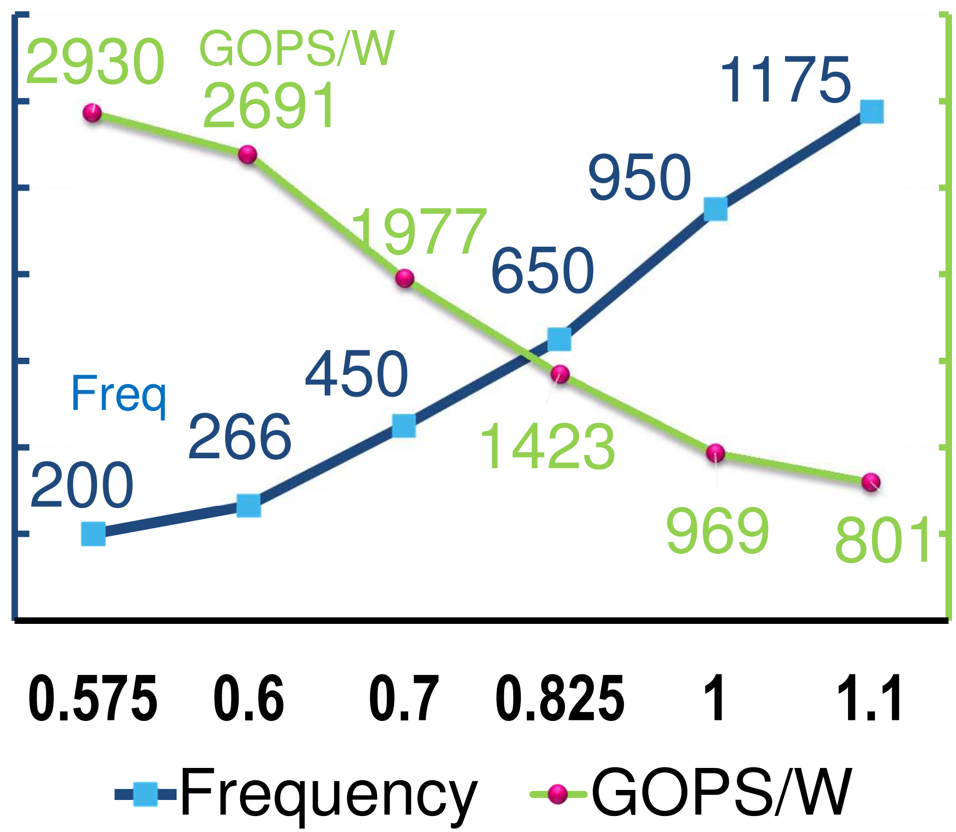
Demo

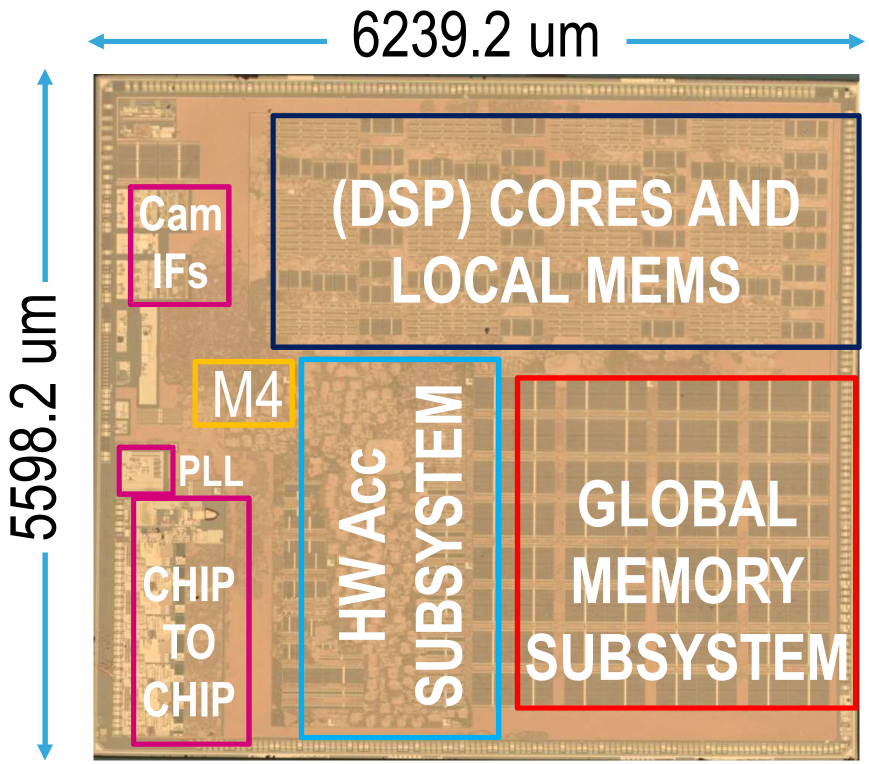
life.augmented

Ultra-Wide DVFS Range

- **LVT design** with heterogeneous Poly-Bias levels → **perf vs leakage**
- **GALS** and **low insertion delay clock networks** to minimize on chip variation margins;
- Mono Supply **memories with fine grained power switches** and sleep mode
- DVFS energy efficiency improvements via **body bias**

Wide DVFS Range Measured on AlexNet





(*) Only HW Acc avg power for AlexNet
(**) 1 MAC defined as 2 OPS (ADD + MUL)

Chip Specs

Technology	FD-SOI 28nm
Package	FBGA 15x15x1.83
Frequency	200MHz–1.175GHz
Supply voltages	0.575–1.1 V digital 1.8V I/O
Power (*)	41 mW
On-chip RAM	4x1MB 8x192KB + 128KB
Host	ARM® Cortex®-M4
No of DSPs	16
Peak DSP perf (**)	75 GOPS (dual 16b MAC loop)
No of CAs	8
CAs perf (**)	676 GOPS peak

Background

Chip architecture

DCNN mapping

Hardware accelerators

Physical implementation

Results

Demo

life.augmented

AlexNet CAs Performance

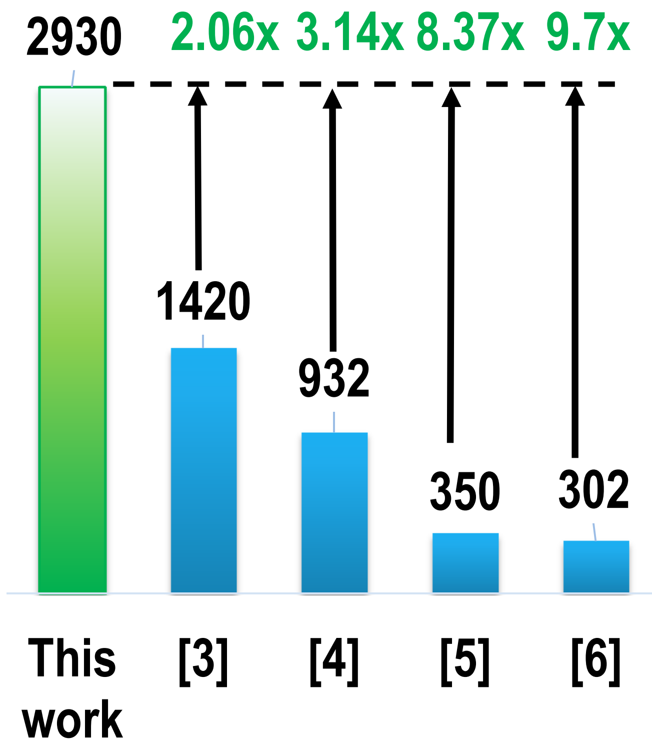
Layer	MOPS	Time [ms]	Load %	GOPs/W		Pwr [mW]	GOPs/W		Pwr [mW]
				16(F)x16(W)→16			8(F)x8(W)→16		
				max	avg		max	avg	
1	210.8	2.5	80	1228	988	86	1810	1456	58
2	447.8	6.5	86	1475	1262	54	2175	1861	37
3	299	3.6	73	1987	1445	58	**2930**	2131	39
4	224.2	2.7	73	1987	1445	58	**2930**	2130	39
5	149.6	1.8	72	1987	1434	58	**2930**	2114	39
Total	**1331.6**	**17.1**	**77**	**1677**	**1287**	**61**	**2473**	**1898**	**41**

200MHz @ 0.575V 25C, 4 chains of 2 CAs,
1 image (227x227) processing

life.augmented

Comparisons with prior work

	This work	[3]	[4]	[5]	[6]
Process (nm)	28 FD-SOI	65	65	28	65
VDD (V)	**0.575–1.1**	1.2	NA	0.9	0.82–1.17
(1)Power (mW)	**39**	45	485	15970	278
Freq (MHz)	200-1175	125	980	606	100–250
Memory (kB)	4096 8*192+128	36	44	16x2048 4096	108
Die size (mm^2)	**2.2 CAs** 34 whole chip	16	3.02	67	12.25
Peak perf (GOPS)	**676 (CAs)** 76 (DSP)	64	452	**5580**	84
Peak effic. (GOPS/W)	**2930**	**1420**	**932**	**350**	**302**



[3] J. Sim, et al. 2016
[4] T. Chen et al 2015
[5] Y. Chen, et al 2014
[6] Y. Chen et al 2016

(1) Power for best efficiency (this work @ 200 MHz, 0.575V, 25C)

Background

Chip architecture

DCNN mapping

Hardware accelerators

Physical implementation

Results

Demo

life.augmented

Development Board

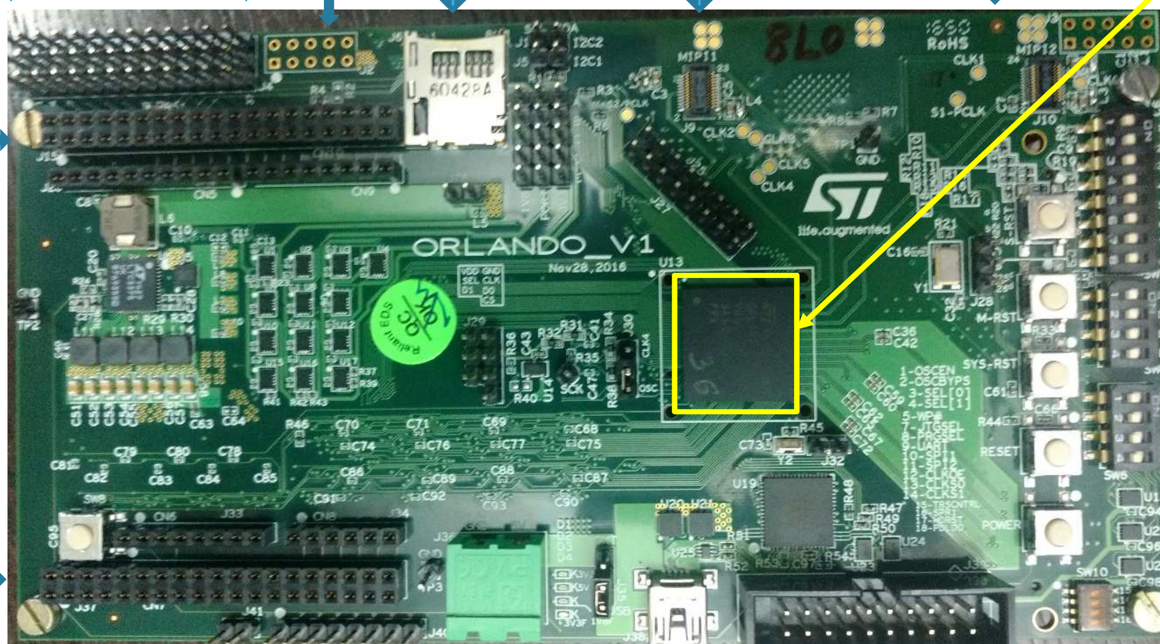
Orlando Expansion
(Non STM32ODE)

2 MIPI &
2 Parallel
Interface

ST Morpho And Arduino Expansion Headers

STM32 Open Development Environment

Motor Control
RF link
Motion Sensors
Security
Environmental Sensors
Dual Interface
Audio



JTAG

Orlando -28FDSOI @1GHz

- ARM M4 Cortex
- 8X2 REISC5 Cluster
- 8 CDNN HW Accerlator
- Image/Video Acc.
- On Chip RAM 48Mb
- 2 CSI2 Camera
- 2 Parallel Camera
- Chip to Chip Bridge 8GHz
- On Board Memory
  - Pseudo RAM128MB
  - 128MB NOR FLASH

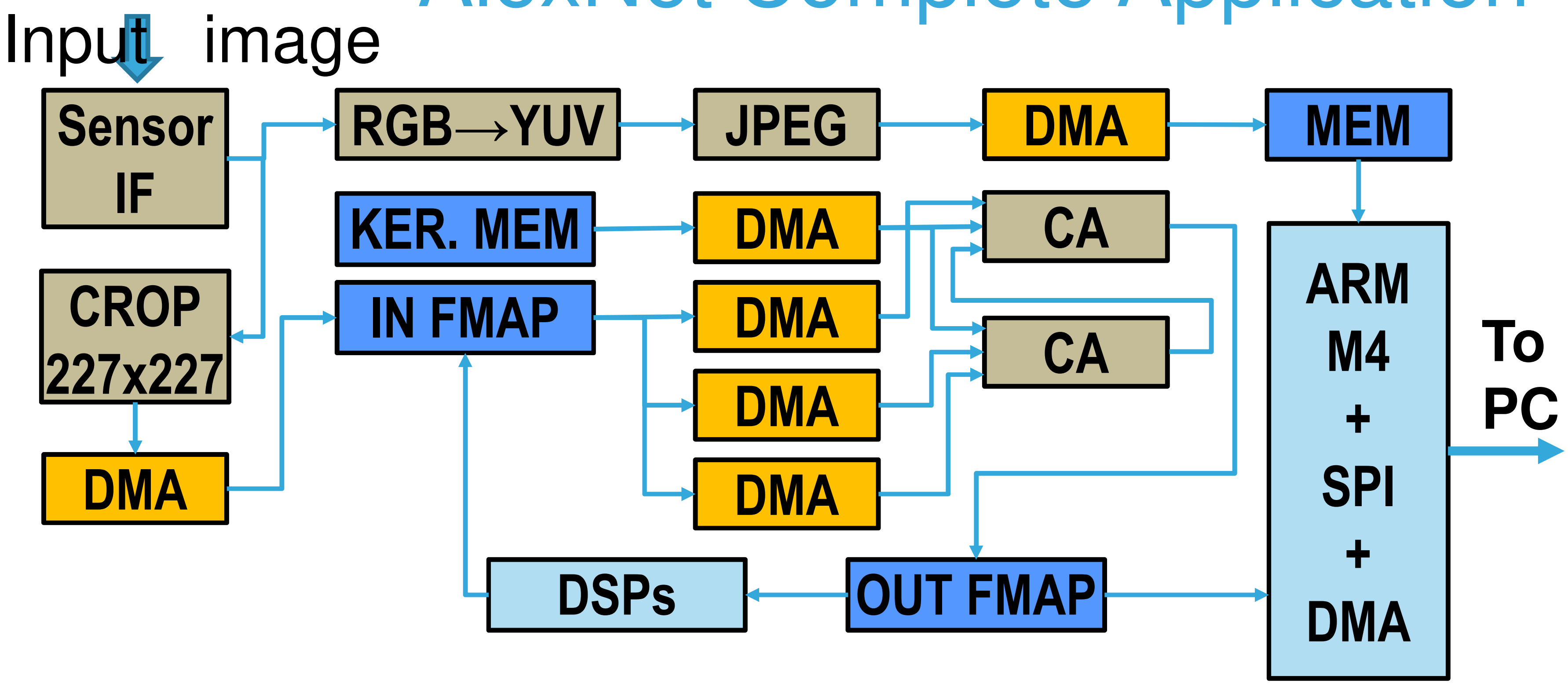
Tools

- GCC Tool Chain
- ARM tool Chain
- **model2plateform**-CDNN engine programmer

User Switch Setting

life.augmented

AlexNet Complete Application



37.5 mW @ 200MHz, 0.6V

10 FPS (38 ms DSPs, 62 ms 2 chained CAs)

Dynamic: 10 mW CAs + 17 mW system

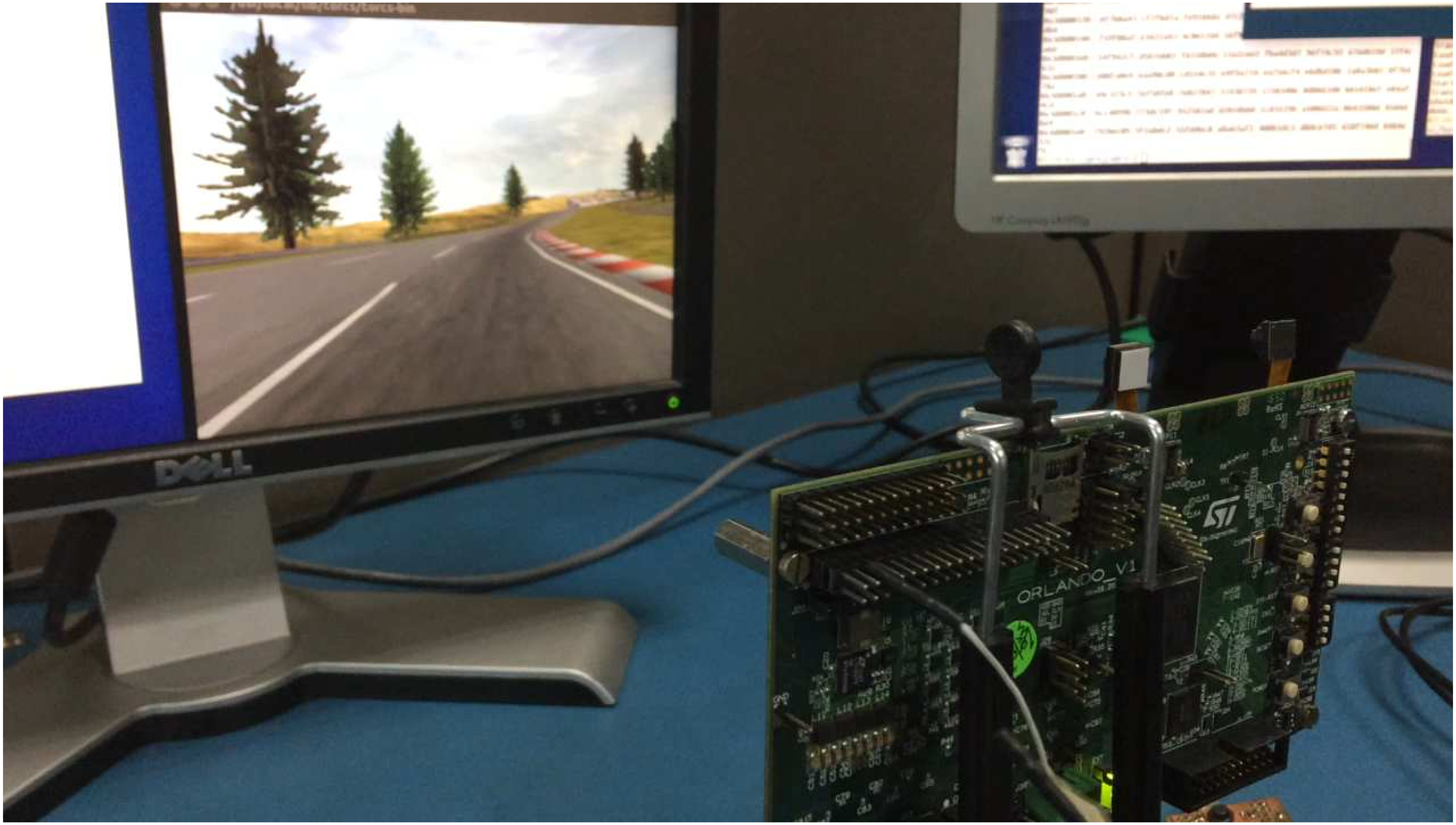
Static: 0.6 mW CAs + 9.9 mW system

life.augmented

AlexNet

Face Expression

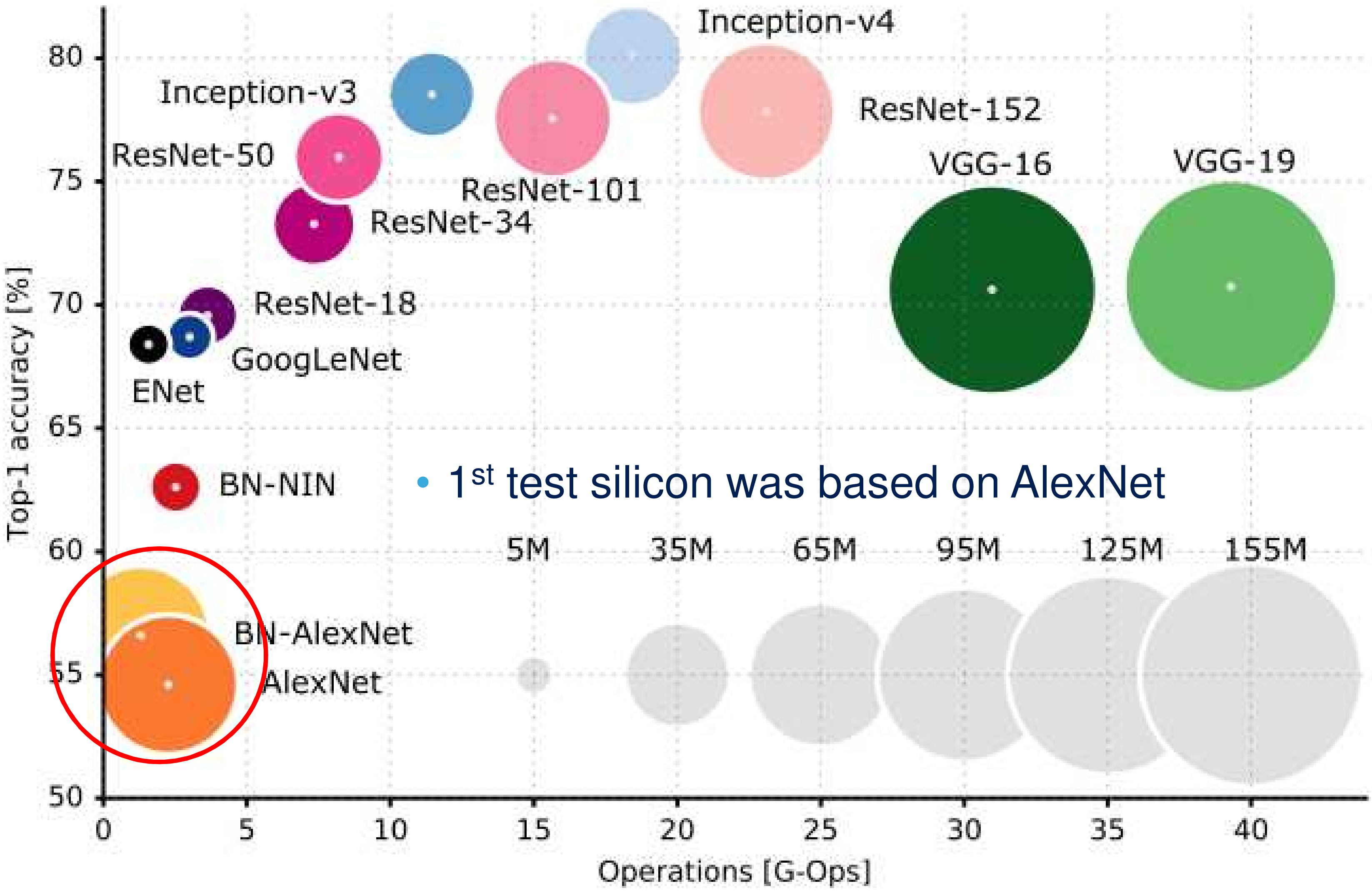
Lane Keeping



Car Race


/usr/local/lib/torcs/torcs-bin
Selected Network
Conv Layers: 5
FC Layers: 3
life.augmented
DELL
HP Compaq LA1951g

SoC Evolution



- 1st test silicon was based on AlexNet

danilo.pau@st.com