ONNC on NVDLA

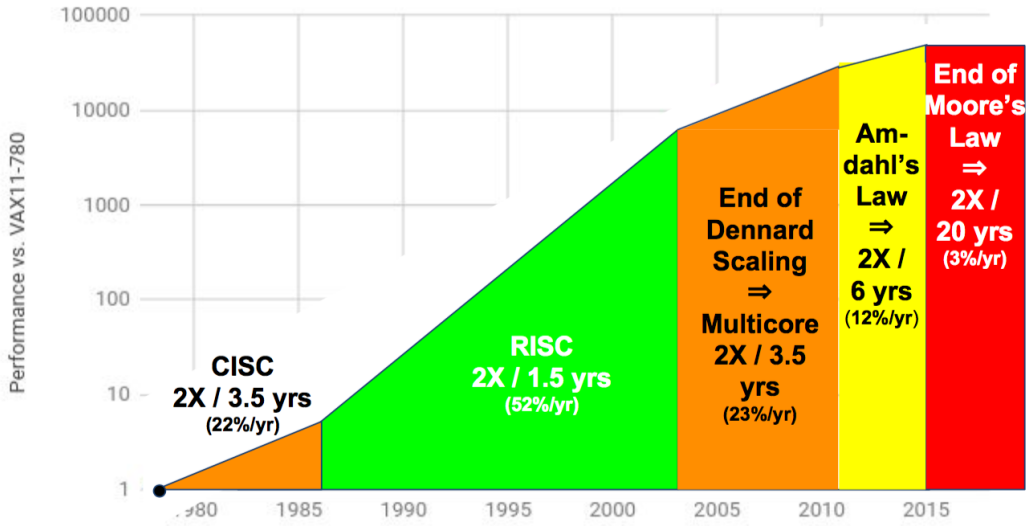
The first open-source software and hardware NVDLA system

Luba Tang
luba@skymizer.com

One-Year Improvement is 3%

Computers Stop Getting Faster

Moore’ s Law is dying

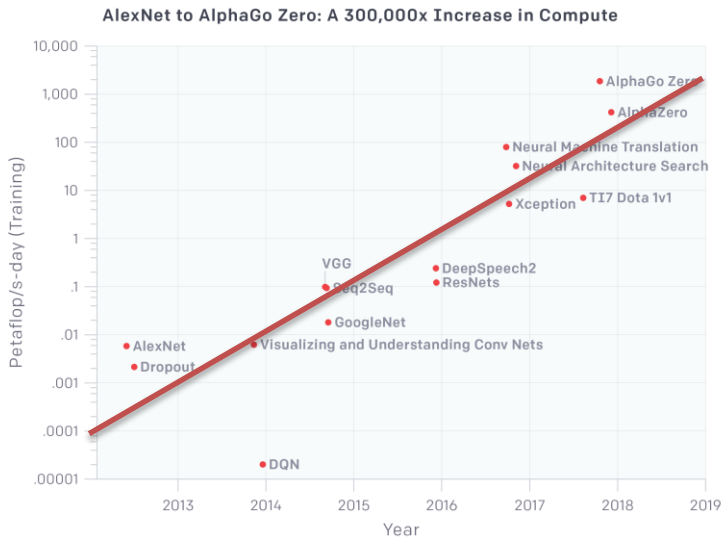


Based on SPECintCPU. Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018

3.5 monthly doubling demand

AI Needs ASICs for Computation

The computation demand has increased exponentially



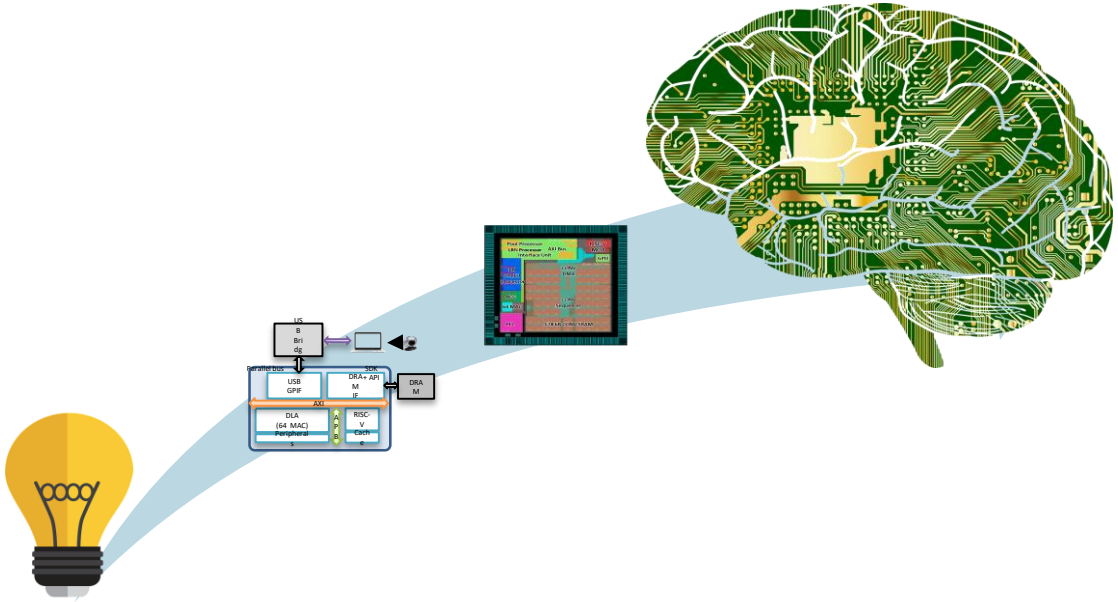
Dario Amodei and Danny Hernandez, “AI and Compute,” OpenAI Blog, May 16, 2018.

skymizer

Shorten the distance

From AI IDEA to ASIC

System Software, Architecture Analysis and Performance Analysis Tools



skymizer

skymizer

ONNC System Software

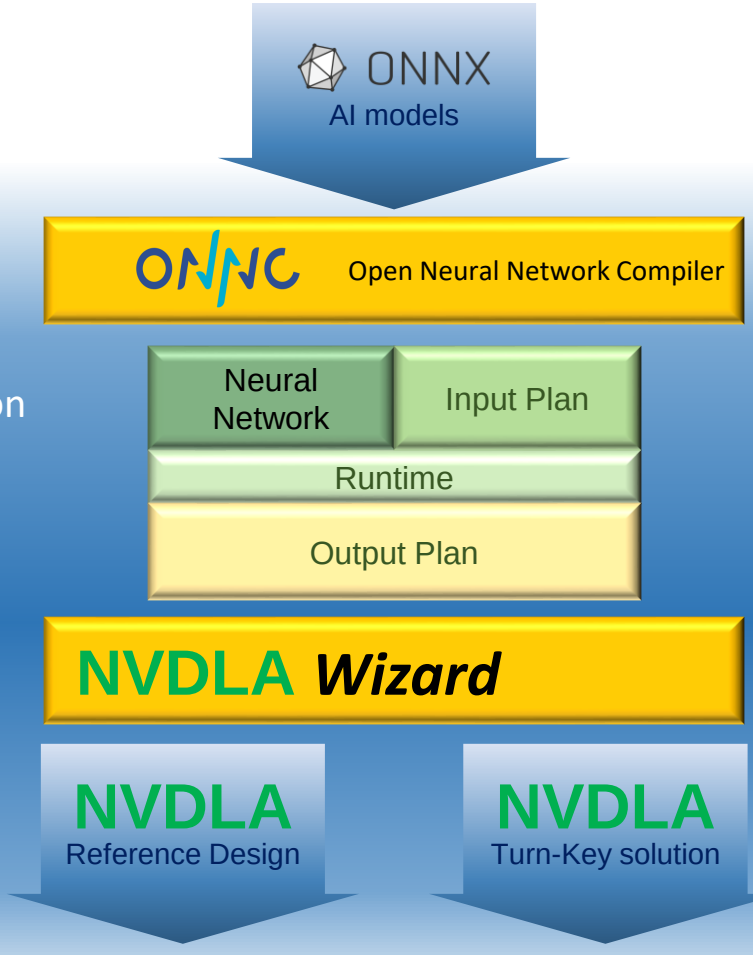
46% less memory consumption

https://onnc.ai

Architecture Design Automation on **NVDLA**

open source hardware

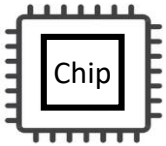
https://nvdla.org



Virtual Platform

FPGA

Emulator



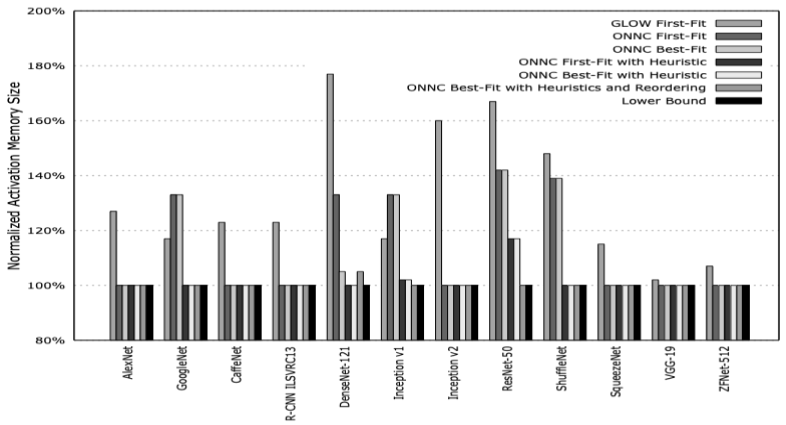
ONNC

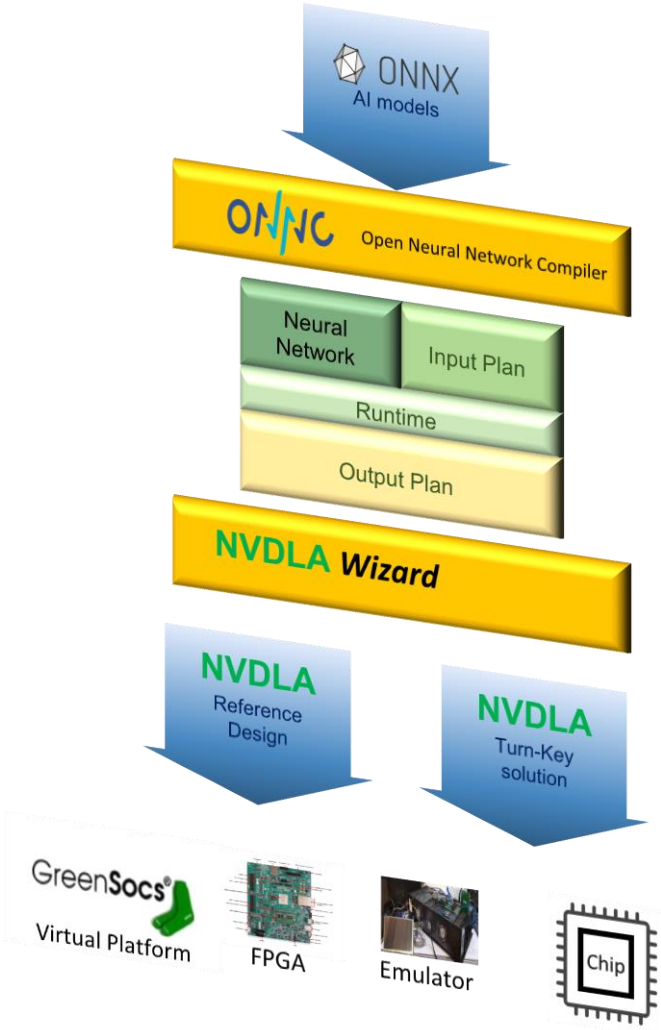
Open Source Project with Commercial Optimization

Optimizing pass	Meaning
Layer Splitting	Split one layer into pieces to open opportunities of memory allocation and tensor scheduling
Memory Allocation	To reuse local memory of DLA, to save memory usage
Dead Code Elimination	Remove unused layers
Common Subexpression Elimination	Combine tensors
Layer Fusing	Fuse layers
Tensor Scheduling	Schedule layers to adjust life-range of tensors

Near-optimal memory allocation

4% gap to theoretical optimal result





ONNC vs NVDLA compiler

- Official NVDLA compiler is released in the binary form and it only supports limited operators and models
- NVDLA Linux drivers, UMD and KMD, are released with source code and exist as defined APIs.
- Successfully demystify the Loadable file format and the NVDLA register specification
- The NVDLA backend in ONNC compiler compiles a model into NVDLA Loadable file
- ONNC can compile 6 models and run on the NVDLA virtual platform successfully
- For the nv_small configuration, there is no official compiler available. Only one Alexnet Loadable file is available in the released testbench
- The basic ONNC support for nv_full is released in the ONNC GitHub repository V1.0 to help the research community

model	nv_full		nv_small	
	NVDLA	ONNC	NVDLA	ONNC
AlexNet	O	O	O	O
GoogleNet	x	O	x	O
CaffeNet	O	O	x	O
R-CNN ILSVRD13	O	O	x	O
DenseNet-121	x	x	x	x
Inception v1	x	O	x	O
Inception v2	x	x	x	x
ResNet-50	O	O	x	O
ShuffleNet	x	x	x	x
SqueezeNet	x	x	x	x
VGG-19	N/A	N/A	N/A	N/A
ZFNet-512	N/A	N/A	N/A	N/A

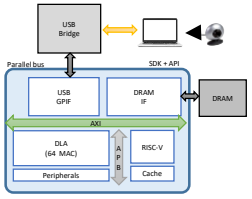
- VGG-19 and ZFNet-512 are not supported by either hardware configurations because they have layers that exceed hardware limitation.
- The other 4 unsupported models (DenseNet-121, Inception v2, ShuffleNet and SqueezeNet) need the support of more operators

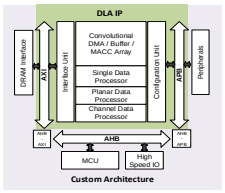
NVDLA
Reference Design

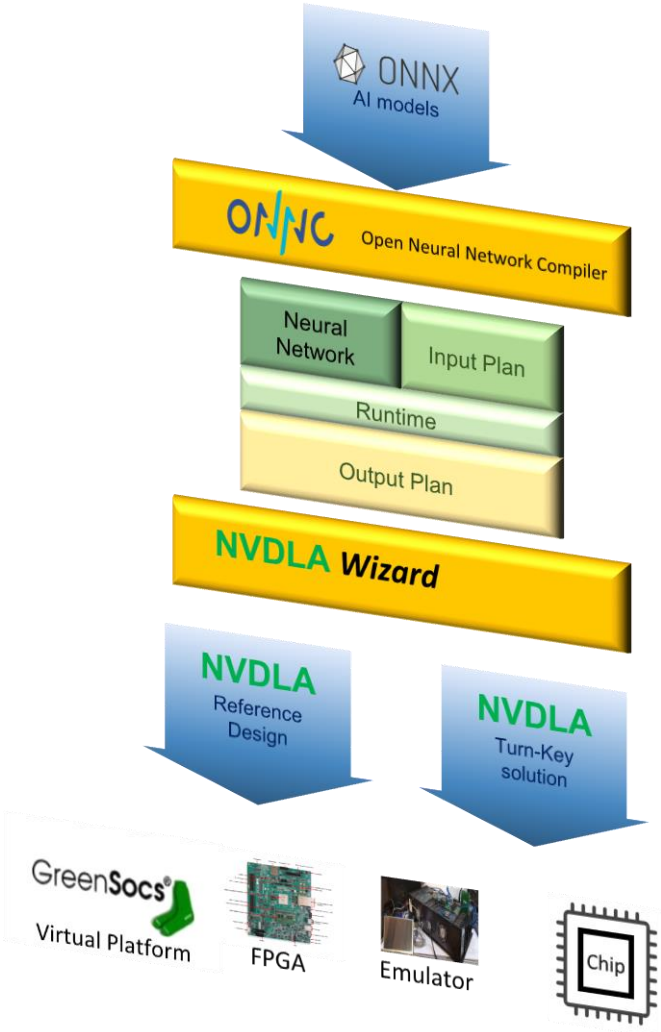
Original NVDLA opens hardware, but its software is not

ONNC is the first open source compiler supports NVDLA

- **QEMU-based virtual platform**
  - Generic and open source machine emulator and virtualizer
- **Virtual modeling for CPU and DLA**
  - Cortex-A processor
  - RISC-V processor
  - NVDLA with difference configuration
- **Performance Analysis Kit**
  - ONNC-based SDK
  - Running popular ML framework
  - Support debug and performance monitoring





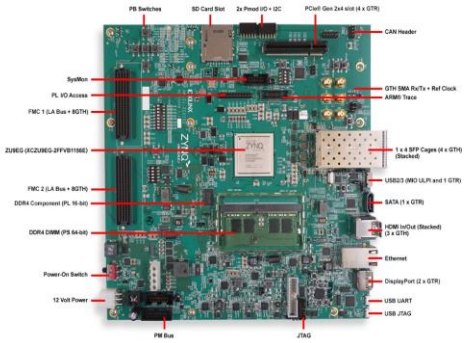


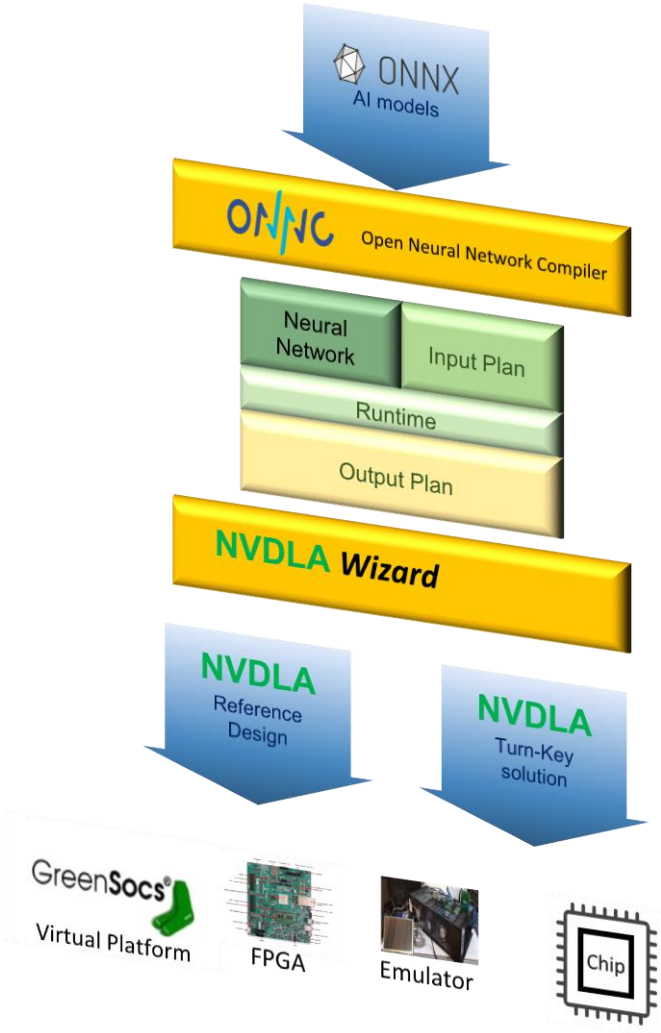
NVDLA

Turn-key solution

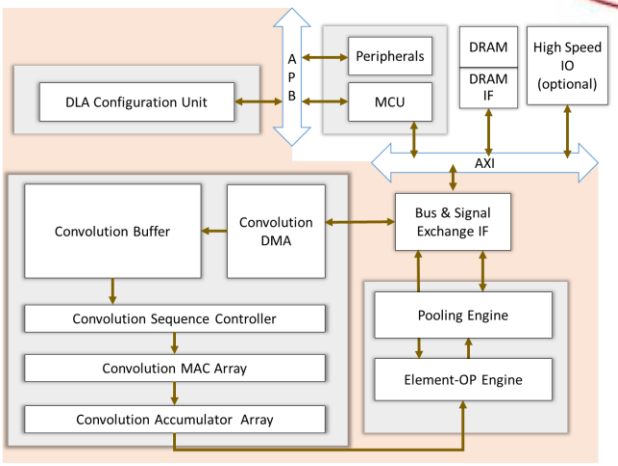
Base on NVDLA reference design, we also provide FPGA, Emulator and Design Service partner

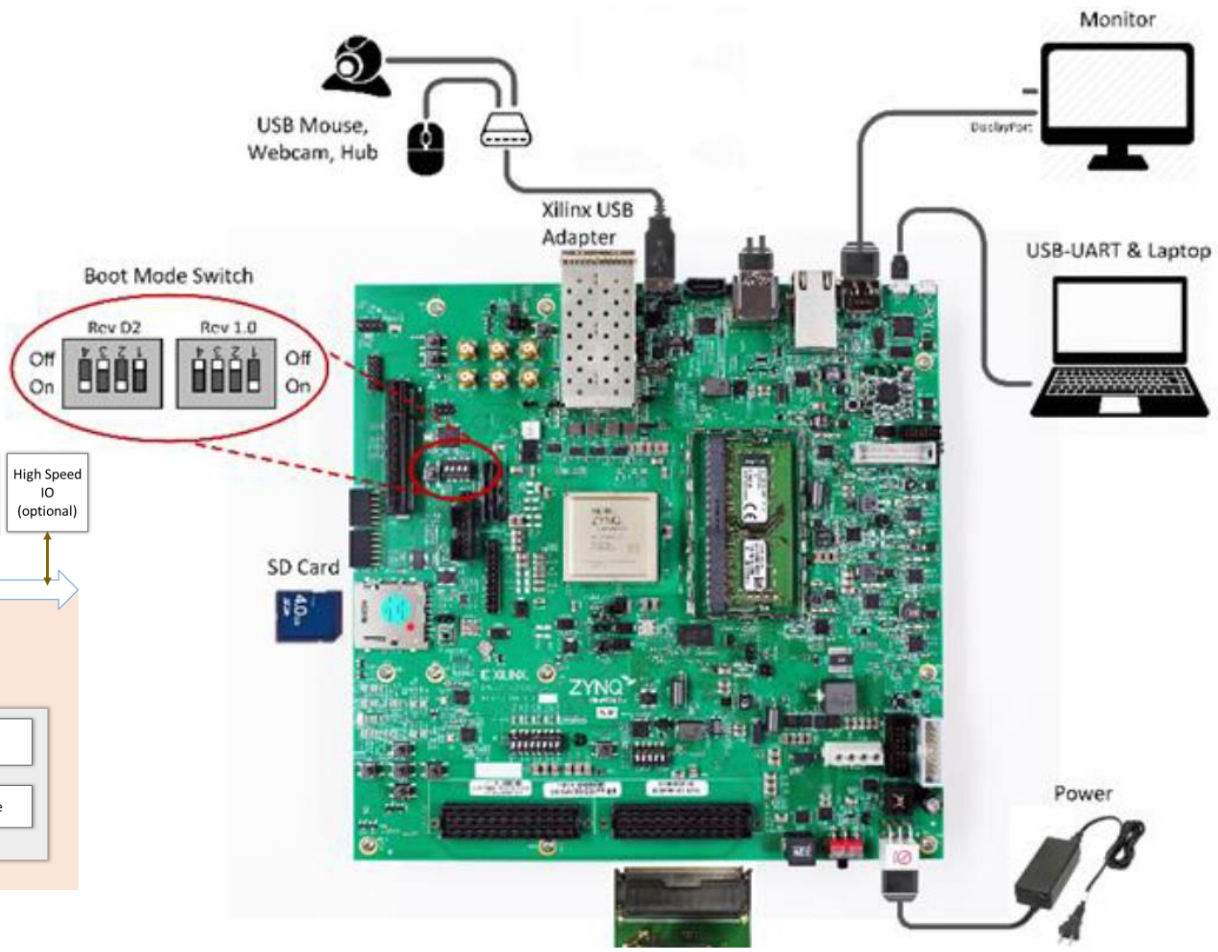
- Xilinx Zynq FPGA Board
  - Quad Arm Cortex-A53, Mali-400 GPU and H.265/264 video codec
  - FPGA programmable logic for DLA
- Synopsys HAPS FPGA system
  - Pre-integrated CA53-based Xilinx Zynq board
  - Extended FPGA programmable logic for DLA
- Integrated NVDLA (RTL/Netlist)
  - NVDLA IP with specific configuration





ONNC+NVDLA
Xilinx ZCU102
Demo Configuration





skymizer

ML Network Performance Analysis

Network	Network Computation	Network MAC Utilization	Memory Access Time %	Total Cycle	Clock Rate	Average Run Time for 1 Frame	FPS
AlexNet (224)	0.72 GMAC	27%	58%	42 M	400MHz	106.2ms	9.4
Inception v1 (224)	1.57 GMAC	82%	22%	30 M	400MHz	75.2ms	13.3
ResNet50 (224)	3.90 GMAC	94%	35%	65 M	400MHz	162.4ms	6.2
VGG16 (224)	15.47 GMAC	83%	26%	290 M	400MHz	725.9ms	1.4
MobileNet v1 (224)	0.54 GMAC	74%	50%	12 M	400MHz	28.8ms	34.7
Inception ResNet v2 (299)	13.17 GMAC	85%	24%	242 M	400MHz	604.9ms	1.7
Inception v3 (299)	5.23 GMAC	86%	23%	95 M	400MHz	238.5ms	4.2
Tiny YOLO v1 (448)	1.61 GMAC	71%	35%	35 M	400MHz	88.1ms	11.3
Tiny YOLO v2 (416)	3.28 GMAC	78%	27%	65 M	400MHz	163.2ms	6.1
Tiny YOLO v3 (416)	2.83 GMAC	91%	14%	48 M	400MHz	121.0ms	8.3
PVANet (480)	3.27 GMAC	85%	38%	60 M	400MHz	150.5ms	6.6
MnasNet (224)	1.17 GMAC	50%	52%	37 M	400MHz	91.7ms	10.9
MobileNet-SSD (480)	2.13 GMAC	78%	44%	43 M	400MHz	106.8ms	9.4

ONNC System Software

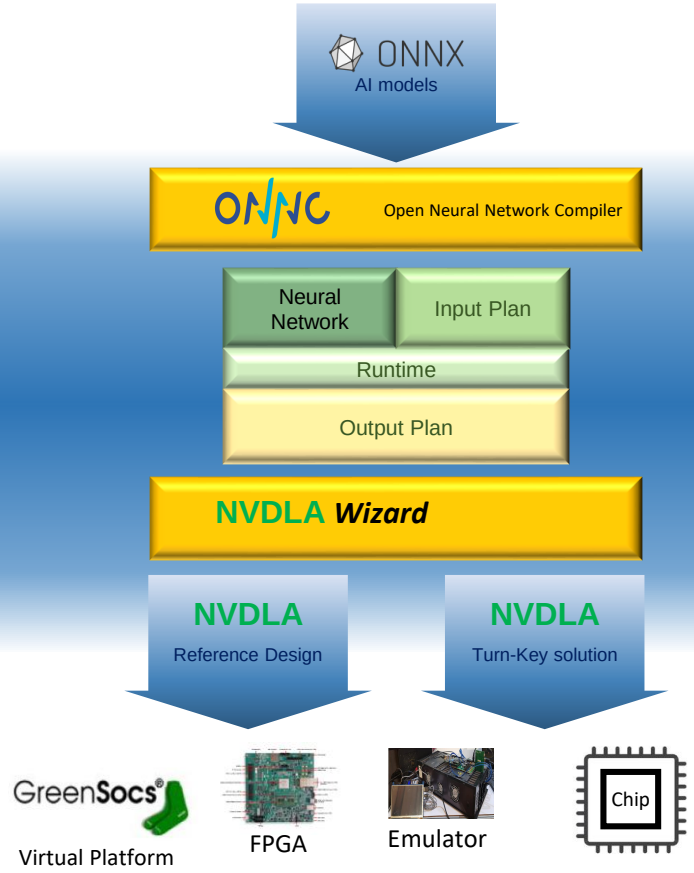
46% less memory consumption

https://onnc.ai

Architecture Design Automation on **NVDLA**

open source hardware

https://nvdla.org



Skymizer Taiwan Inc.

Contact us:

Email: sales@skymizer.com | Tel: +886 2 8797 8337

HQ: 12F-2, No.408, Ruiguang Rd., Neihu Dist., Taipei City 11492, Taiwan

skymizer