NVDLA Overview

Ing-June Lu

Skymizer, TAIWAN

2020/5/26

Outline

- The “NVDLA” Story
- Features of the Accelerator
- Available Resources in the Open Source Domain
- Hardware Architecture Overview
- Hardware Configurations
- Performance Benchmark

The “NVDLA” Story

- NVDIA’s “Tegra Xavier” SOC
  - High-performance Processor for Autonomous Machines
  - 8-ARM cores+ Volta GPU+PVA+DLA
    - PVA: Programmable Vision Accelerator
    - DLA: Deep Learning Accelerator
- Announced Open-source of NVDLA in May, 2017
  - Early access in June, 2017
  - official release in September, 2017
  - “NVDLA” = NVDIA Deep Learning Accelerator
- Skymizer’s ONNC support
  - Started in September, 2018
  - Release open source in March, 2019

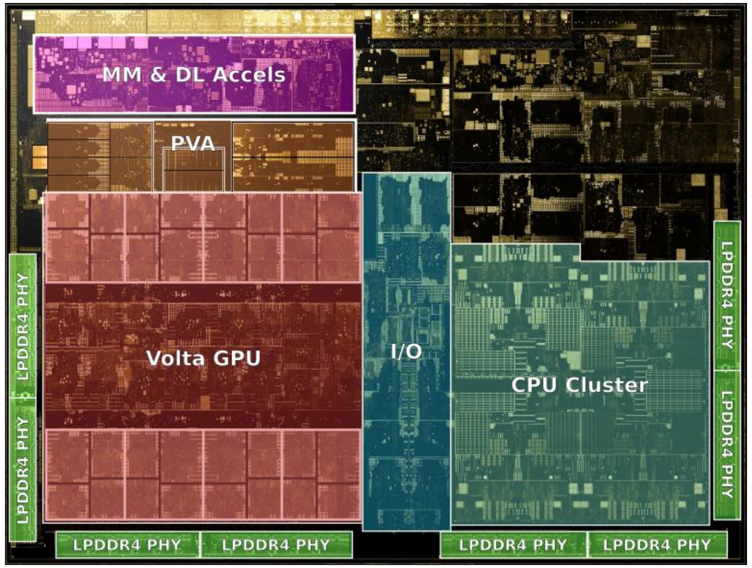


Photo adapted from
https://en.wikichip.org/wiki/File:nvidia_xavier_die_shot_(annotated).png

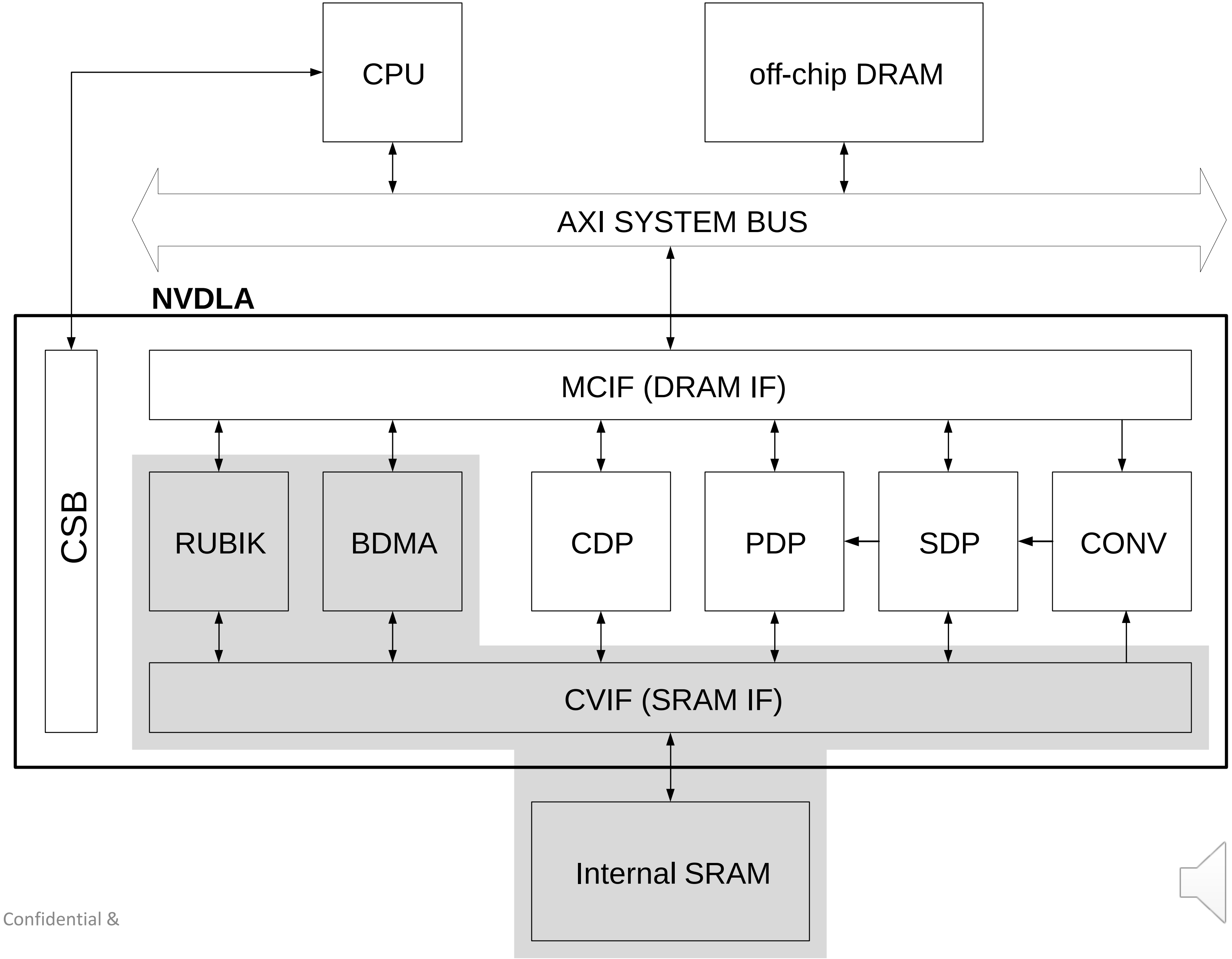
Features of the Accelerator

- Silicon Proven Production Level Design.
- Modular Design.
- True 3-D Convolution Engine.
- Scalable and Configurable.

Available Resources in the Open Source Domain

- Documentations
- Github HW Repo
  - 3 branches: nvdlav1, master, nv_small
  - RTL code, test benches and verification suites
- Github SW Repo
  - virtual platform
  - Kernel mode and user mode drivers
  - Loadable files

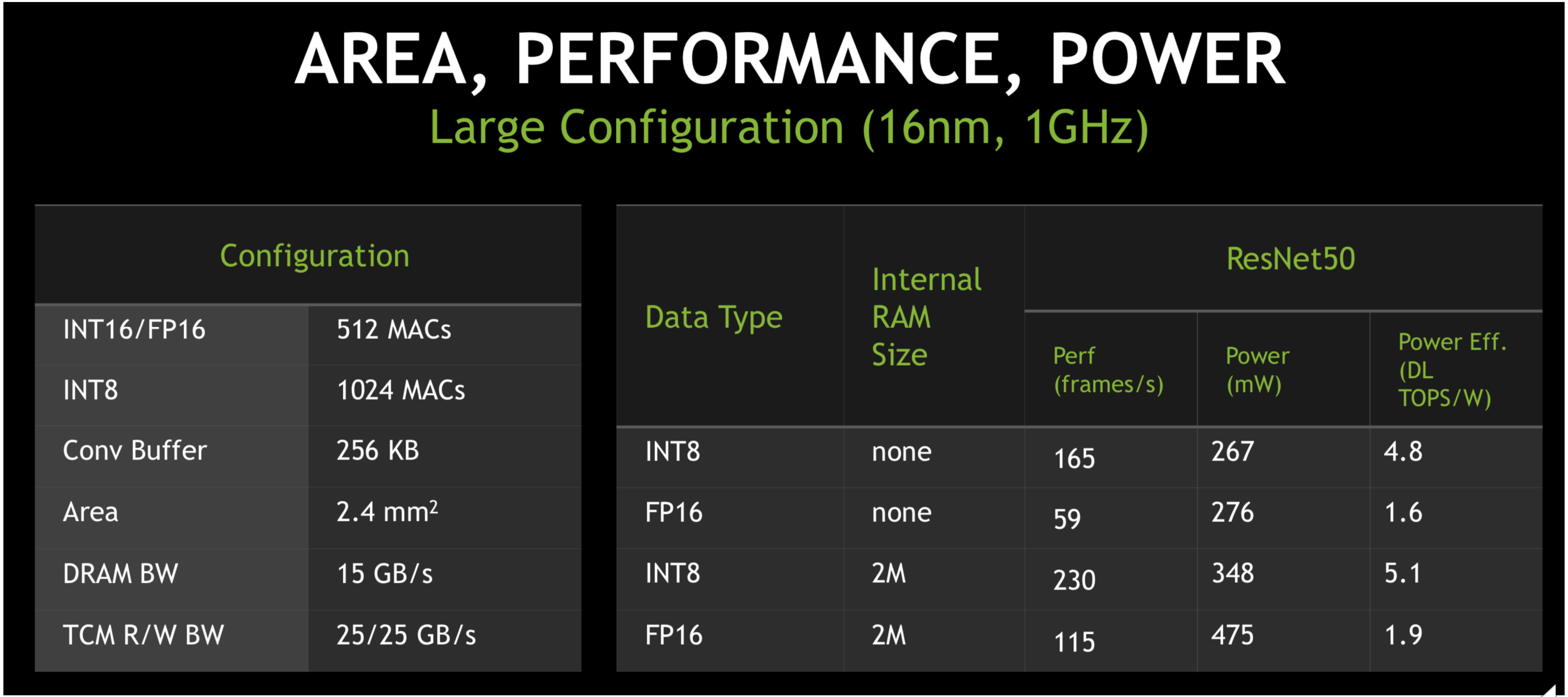
Hardware Architecture Overview



Hardware Configurations

NAME	nv_full	nv_large	nv_medium_1024_full	nv_medium_512	nv_small_256_full	nv_small_256	nv_small
DATA TYPE	FP16/INT16/INT8	INT8	INT8	INT8	INT8	INT8	INT8
# MAC (ATOMIC_K)	64	64	32	16	8	8	8
# MULT/MAC (ATOMIC_C) CUBF_BANK_WIDTH	32	32	32	32	32	32	8
CBUF_BANK_DEPTH	512	512	512	512	128	128	512
CBUF_BANK_NUM	16	16	32	32	32	32	32
SDP_BS/BN_THROUGHPUT	16	16	8	4	2	1	1
SDP_EW_THROUGHPUT	4	4	2	X	1	X	X
PDP_THROUGHPUT	8	8	4	2	1	1	1
CDP_THROUGHPUT	8	8	4	2	1	1	1
DRAM IF Data Bus Width	512	256	256	128	64	64	64
SRAM IF Data Bus Width	512	256	256	X	64	X	X
RUBIK / BDMA	YES	NO	NO	NO	NO	NO	NO

Performance Benchmark



AREA, PERFORMANCE, POWER

Large Configuration (16nm, 1GHz)

Configuration	
INT16/FP16	512 MACs
INT8	1024 MACs
Conv Buffer	256 KB
Area	2.4 mm^2
DRAM BW	15 GB/s
TCM R/W BW	25/25 GB/s

Data Type	Internal RAM Size	ResNet50		
		Perf (frames/s)	Power (mW)	Power Eff. (DL TOPS/W)
INT8	none	165	267	4.8
FP16	none	59	276	1.6
INT8	2M	230	348	5.1
FP16	2M	115	475	1.9

Table adapted from https://www.hotchips.org/hc30/2conf/2.08_NVidia_DLA_Nvidia_DLA_HotChips_10Aug18.pdf

References

- NVDLA
  - http://nvdla.org/
  - https://github.com/nvdla
- ONNC
  - https://github.com/ONNC/onnc

Skymizer Taiwan Inc.

CONTACT US

E-mail sales@skymizer.com **Tel** +886 2 8797 8337

HQ 12F-2, No.408, Ruiguang Rd., Neihu Dist., Taipei City 11492, Taiwan
BR Center of Innovative Incubator, National Tsing Hua University, Hsinchu Taiwan

skymizer

Boost deep learning accelerator
with compiler technology

https://skymizer.com