
NVIDIA
THE NVIDIA
DEEP LEARNING ACCELERATOR

INTRODUCTION

NVDLA — NVIDIA Deep Learning Accelerator

Developed as part of Xavier – NVIDIA's SOC for autonomous driving applications

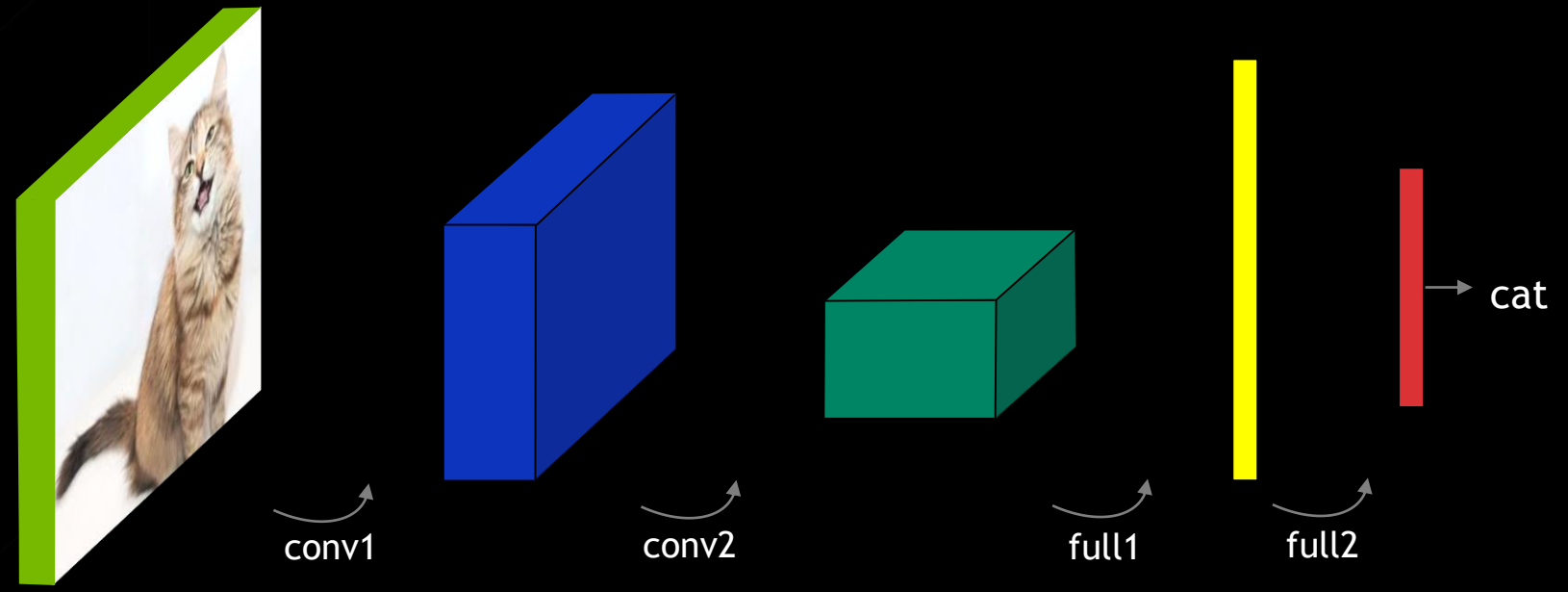
Optimized for Convolutional Neural Networks (CNNs), computer vision

Open source architecture and RTL release

- Encourage Deep Learning applications
- Invite contributions from the community
- Targeted towards edge devices, IoT
- Industry standard formats and parameterized

CNN INFERENCE

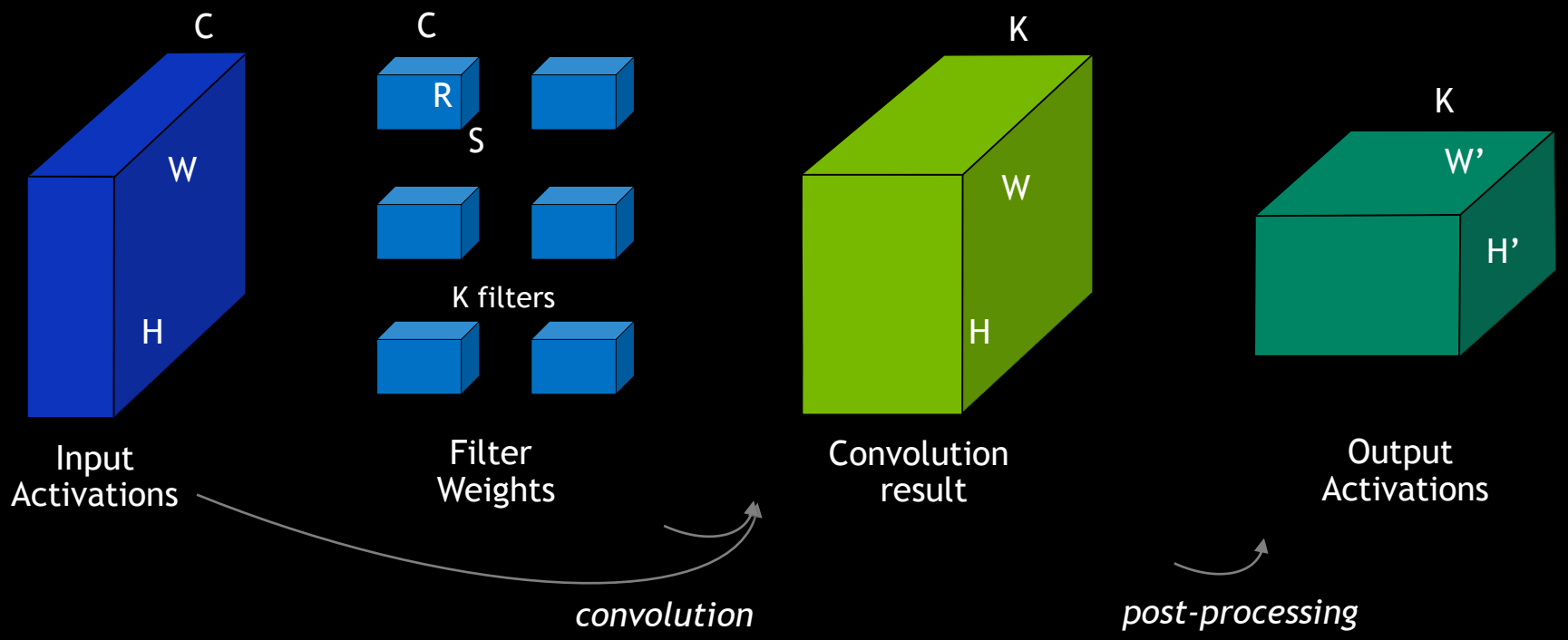
Convolutional Neural Network



Convolutional and fully connected layers

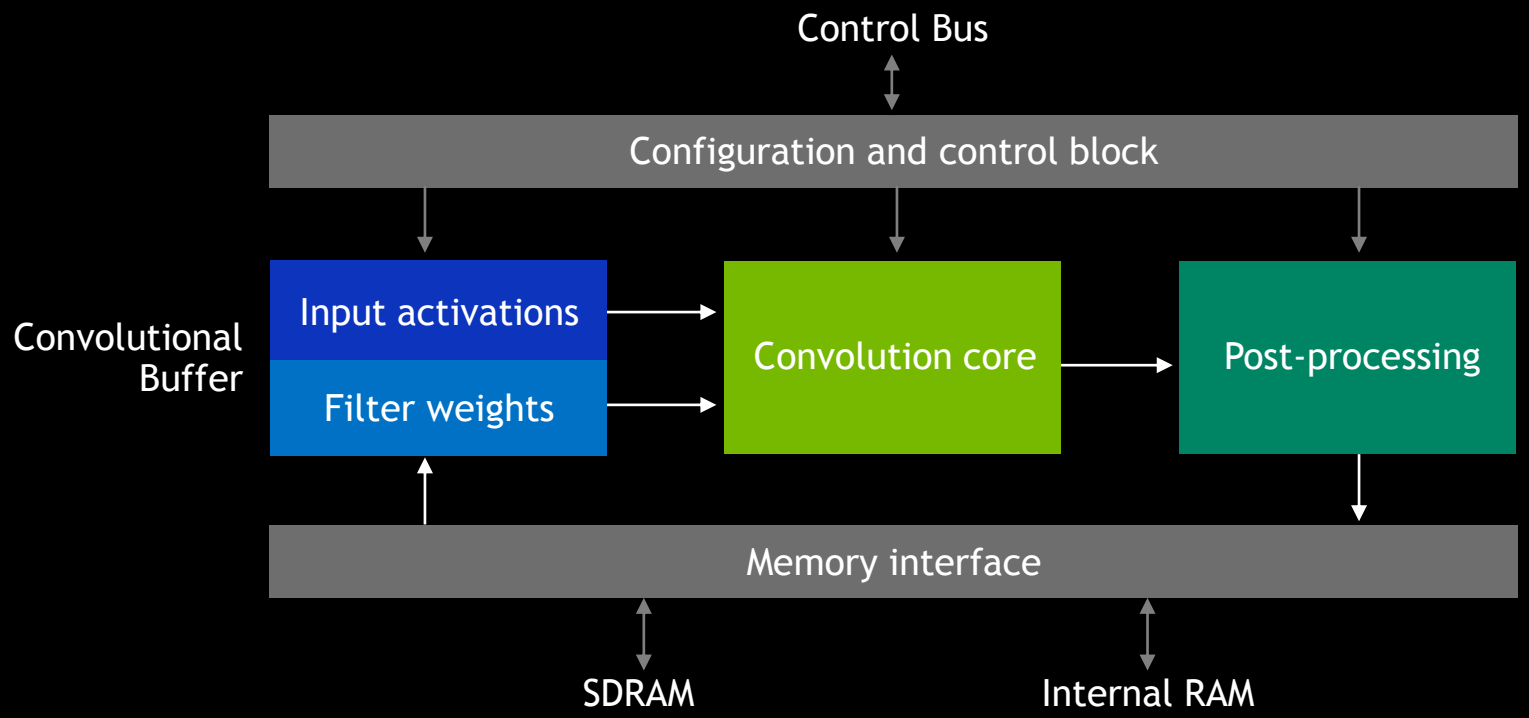
CNN INFERENCE

One Convolutional Layer



ARCHITECTURE OVERVIEW

Top Level Architecture



ARCHITECTURE OVERVIEW

Memory Considerations

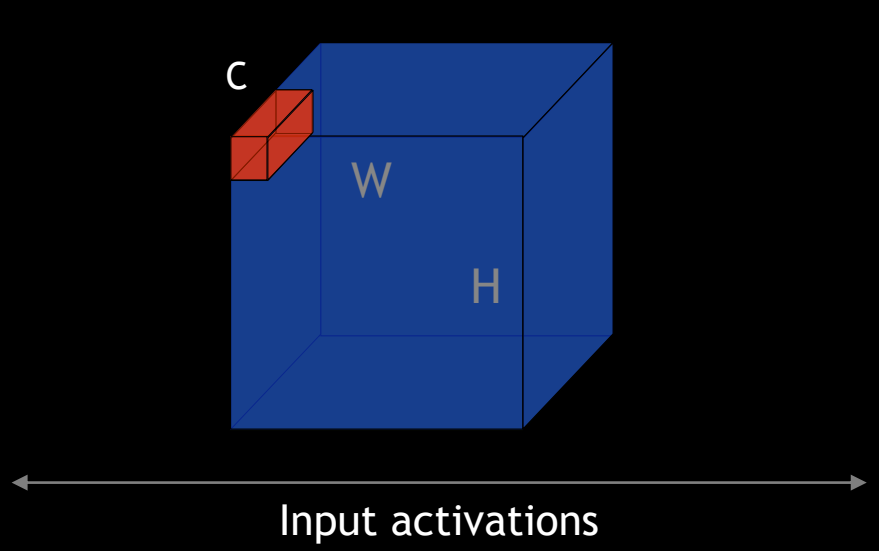
- Convolutional buffer size vs Memory Bandwidth trade off
  - If conv buffer can fit 1/N’th of total weights, activations need to be read N times
  - Example: GoogleNet layer inception 4a/3x3, 16-bit precision
    - Input activations: 1.2 MB
    - Filter weights: 360KB
  - If conv buffer is 128KB, then minimal bandwidth for activations is 3 x 1.2MB = 3.6MB
- Convolutional buffer has to be multi-ported, internal RAM can be single ported

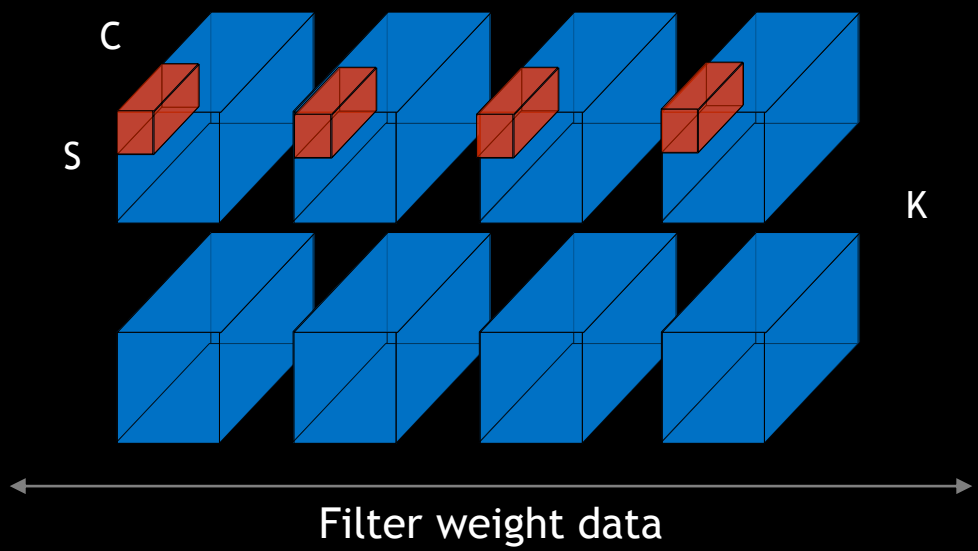
CONVOLUTION CORE

Atomic Operation (Each Clock Cycle)

Example: 64 activations x (4 filters x 64 weights) → 4 partial outputs

256 MACs per clock cycle





CONVOLUTIONAL CORE

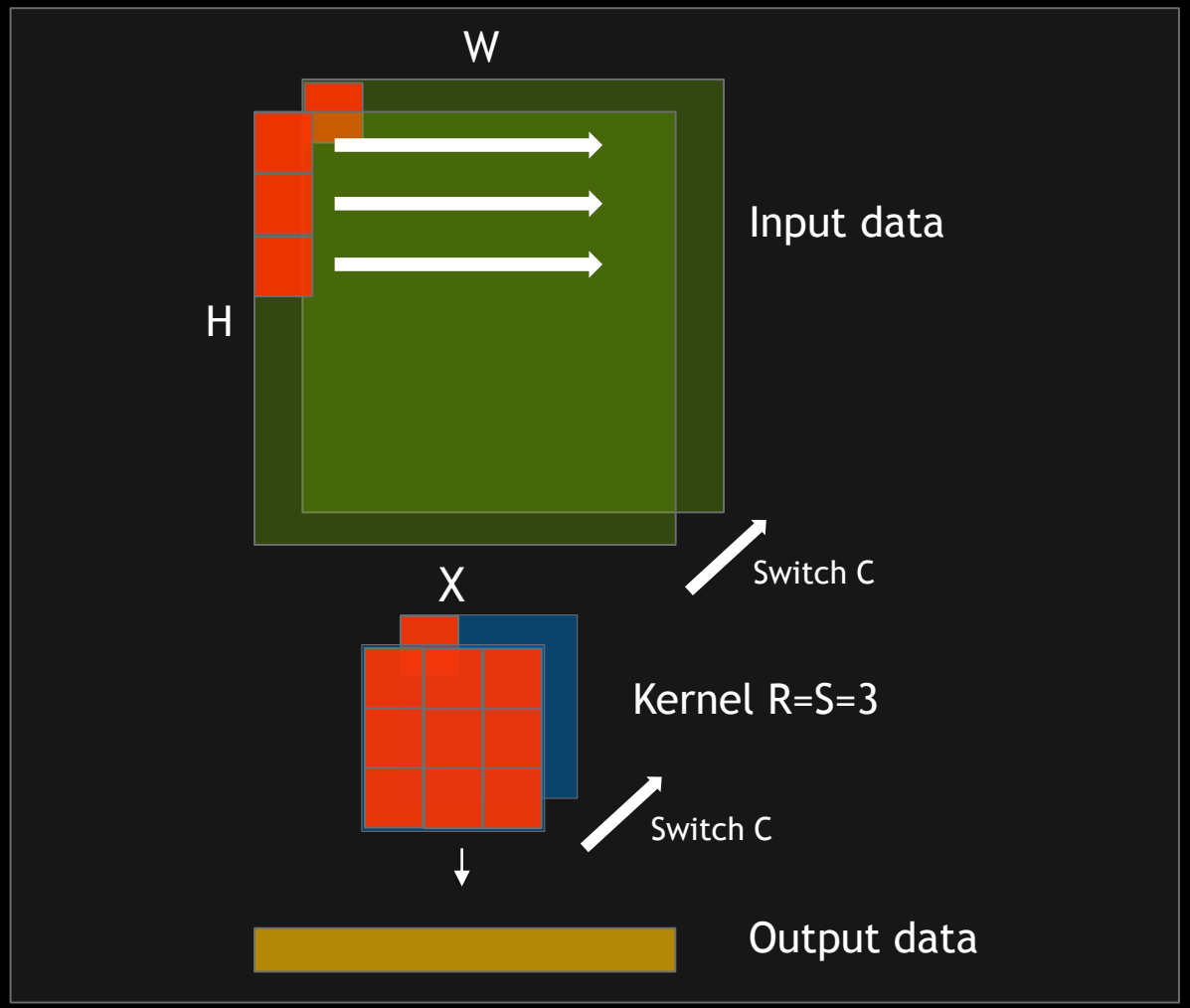
Order of Calculations

Calculate one stripe of outputs at a time

Change activations first, then weights

In the slide show:

- C-dimension not shown
- Only one kernel shown



CONVOLUTION CORE

Area and Power Considerations

More samples in C-direction per atomic operation:

- Allows for wider Wallace tree (half rather than full adders)
- Causes more wasted MACs on non-aligned boundaries

Keeping one operand (weights) of MACs constant for a number of cycles saves dynamic power and reduces data transfers

Calculate full sums rather than partial sums as latter require higher precision before rounding (better to stream/store inputs than outputs)

AREA, PERFORMANCE, AND POWER

Configurations

SMALL CONFIGURATION

INT8 data path

1 RAM interface

No advanced features

LARGE CONFIGURATION

INT8, INT16, FP16 data path

2 RAM interfaces

Integrated controller

Weight compression

…

AREA, PERFORMANCE, POWER

Small Configurations (16nm, 1GHz)

INT8 MACs (# instances)	Conv. Buffer (KB)	Area (mm2)	Memory BW (GB/s)	ResNet50		
				Perf (frames/s)	Power (mW)	Power Eff. (DL TOPS/W)
2048	512	3.3	20	269	388	5.4
1024	256	1.8	15	153	185	6.3
512	256	1.4	10	93	107	6.8
256	256	1.0	5	46	64	5.6
128	256	0.84	2	20	41	3.8
64	128	0.55	1	7.3	28	2.0

Area is synthesis area + internal RAMs, does not account for layout inefficiencies
Power is for DLA incl. internal RAMs, excluding SOC & external RAMs
Calibrated to Xavier silicon - NVIDIA flows, libraries, RAM compilers, …
*DL TOPS == #convolutional MAC operations * 2*

AREA, PERFORMANCE, POWER

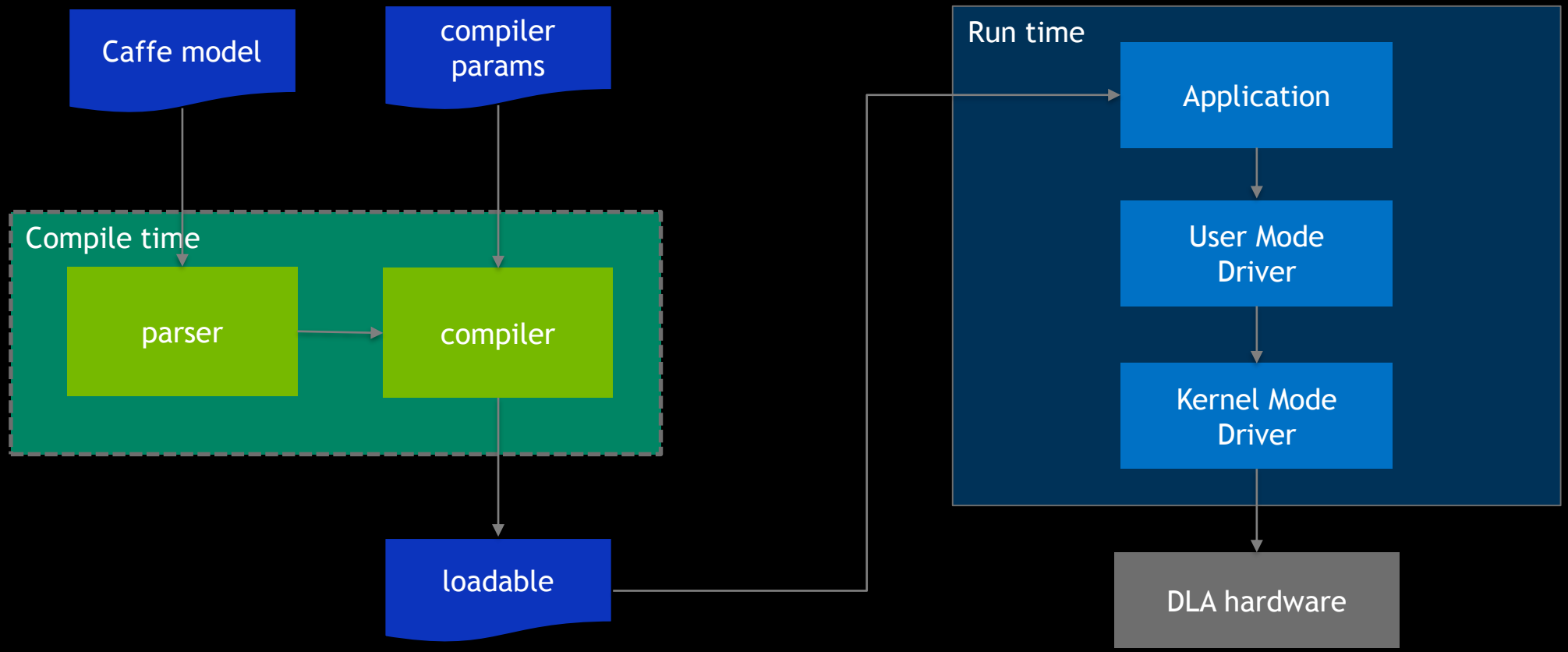
Large Configuration (16nm, 1GHz)

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

Area and power do not include Tightly Coupled Memory (TCM)

SW ARCHITECTURE



NVDLA VIRTUAL PLATFORM

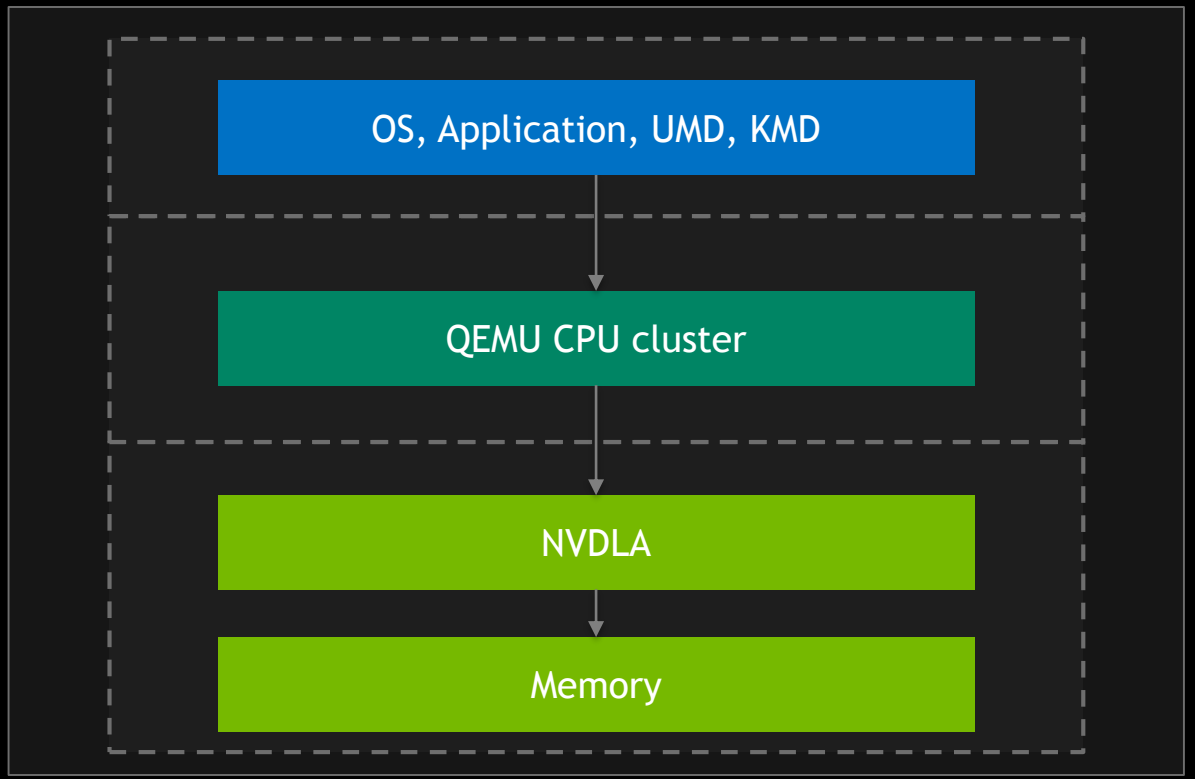
SW & RTL prototyping for all configs

Two formats for NVDLA + Memory

- C-model
- FPGA board

Easy access via cloud

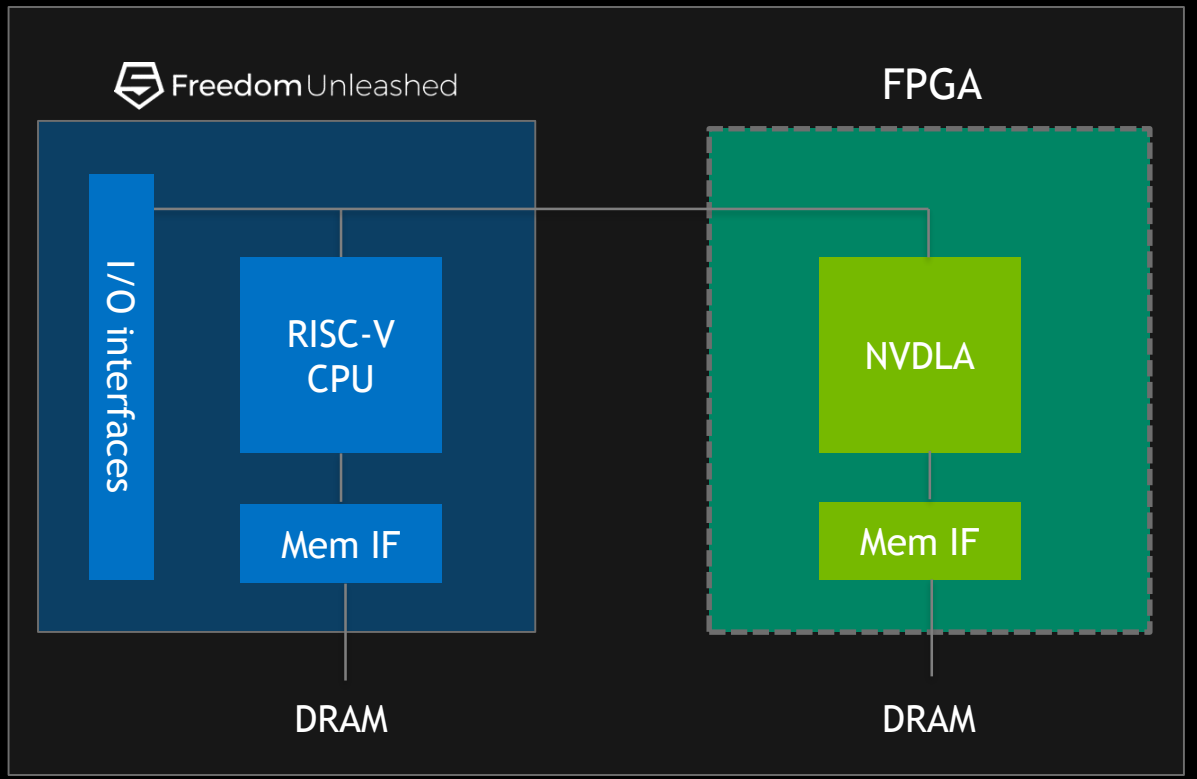
- Amazon Web Services (AWS)



OPEN SOURCE SOC PROTOTYPE

NVDLA + SiFive RISC-V

- Demo at SiFive booth
- NVDLA config
  - Small config
  - 2048 MACs
  - 512 KB
- YOLOv3 object recognition



Inserting video: Insert/Video/Video from File.
Insert video by browsing your directory and selecting OK.

File types that works best in PowerPoint are mp4 or wmv

VIDEO FILE

CHECK IT OUT FOR YOURSELF

OR CONTRIBUTE!

http://nvdla.org/index.html

Documentation and source code are available under permissive license

Community contributions under Contributor License Agreement are encouraged


NVIDIA