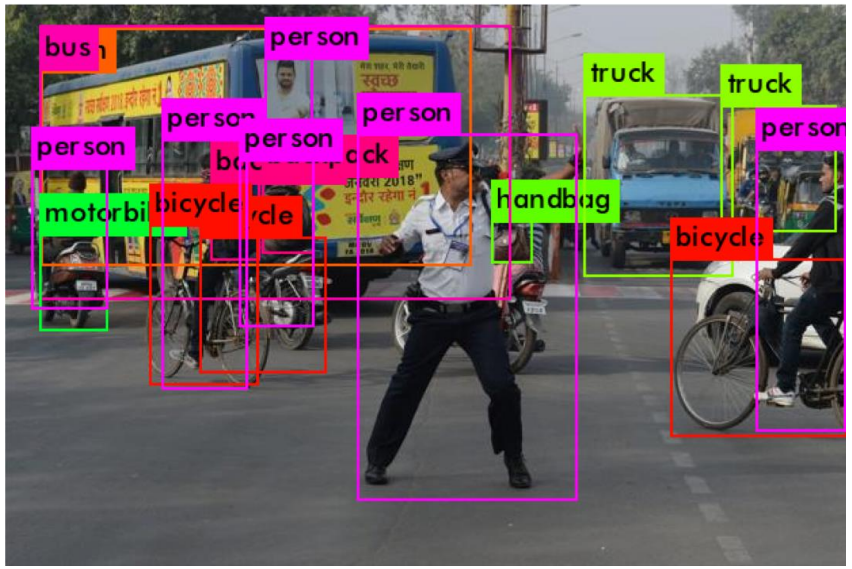
Real Time Object Recognition Backup Cameras to 60 MPH

Dr. Cheng C. Wang
Co-Founder & Senior VP Engineering/Software, Flex Logix Technologies, Inc.

cheng@flex-logix.com

Autonomous Vehicle Hardware Summit
March 27, 2019, San Jose, CA



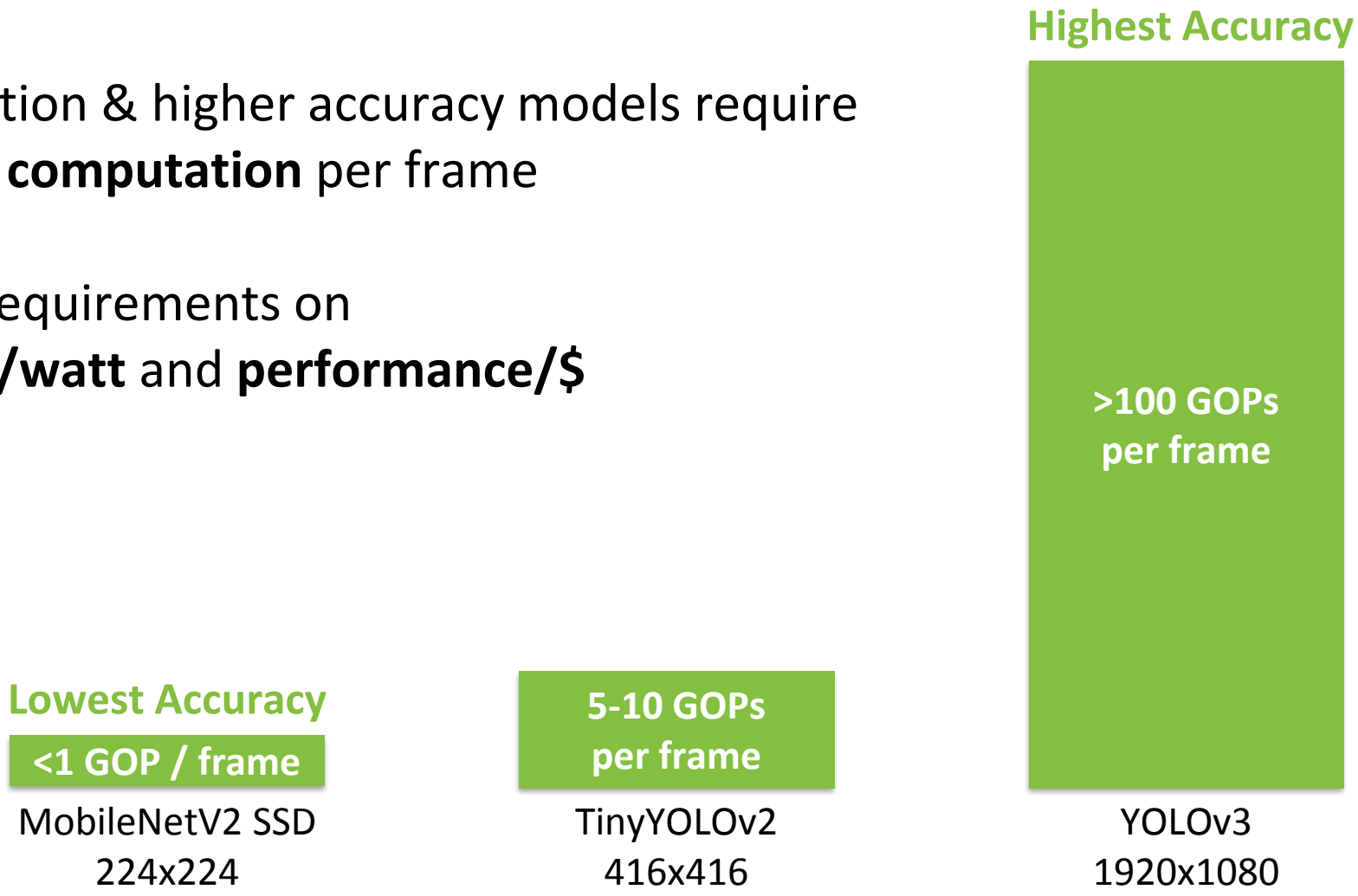
Autonomous Vehicle Inference, Applications & Requirements

- 1. Real time object recognition at 60MPH: vision and LIDAR/Radar
- 2. Real time object recognition for backup cameras
- 3. In-cabin monitoring of driver behavior, state

Require: acceptable **throughput** and **accuracy** at acceptable **power** and **cost**

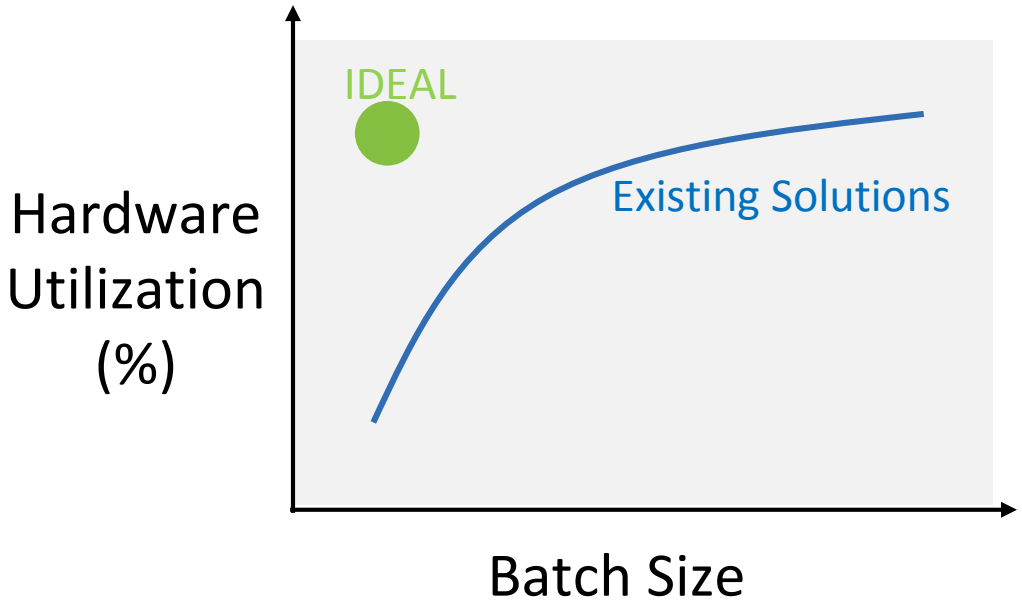
Real Time Object Detection and Recognition Models

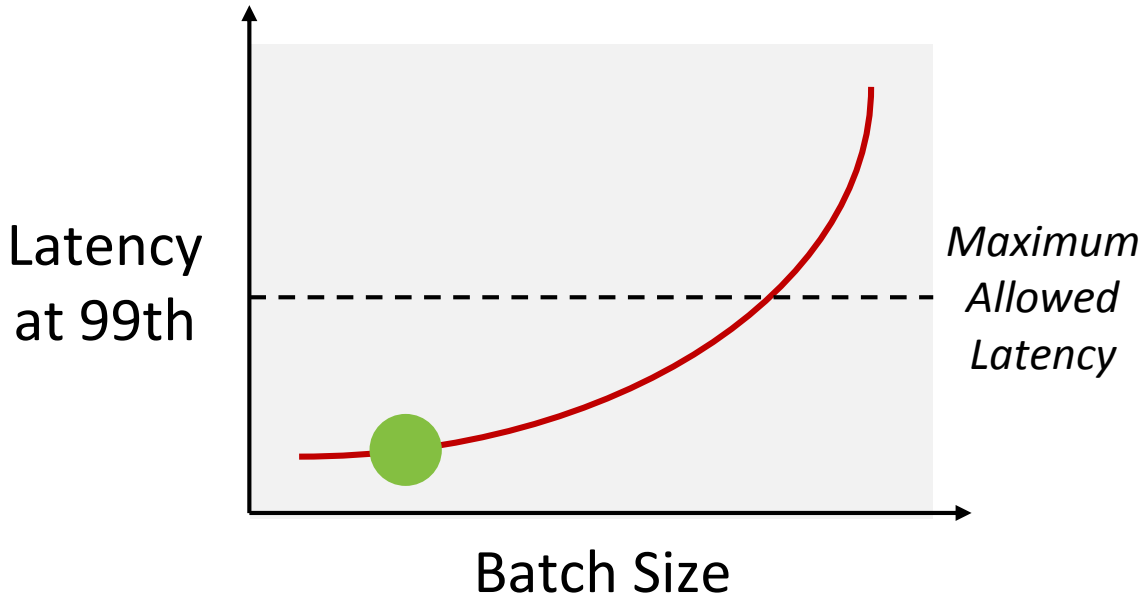
- Higher resolution & higher accuracy models require **>100x higher computation** per frame
- Places large requirements on **performance/watt** and **performance/$**



Autonomous Driving also requires low latency

Microsoft BrainWave Slide from HotChips 2018:





Ideal hardware : low latency at low batch size

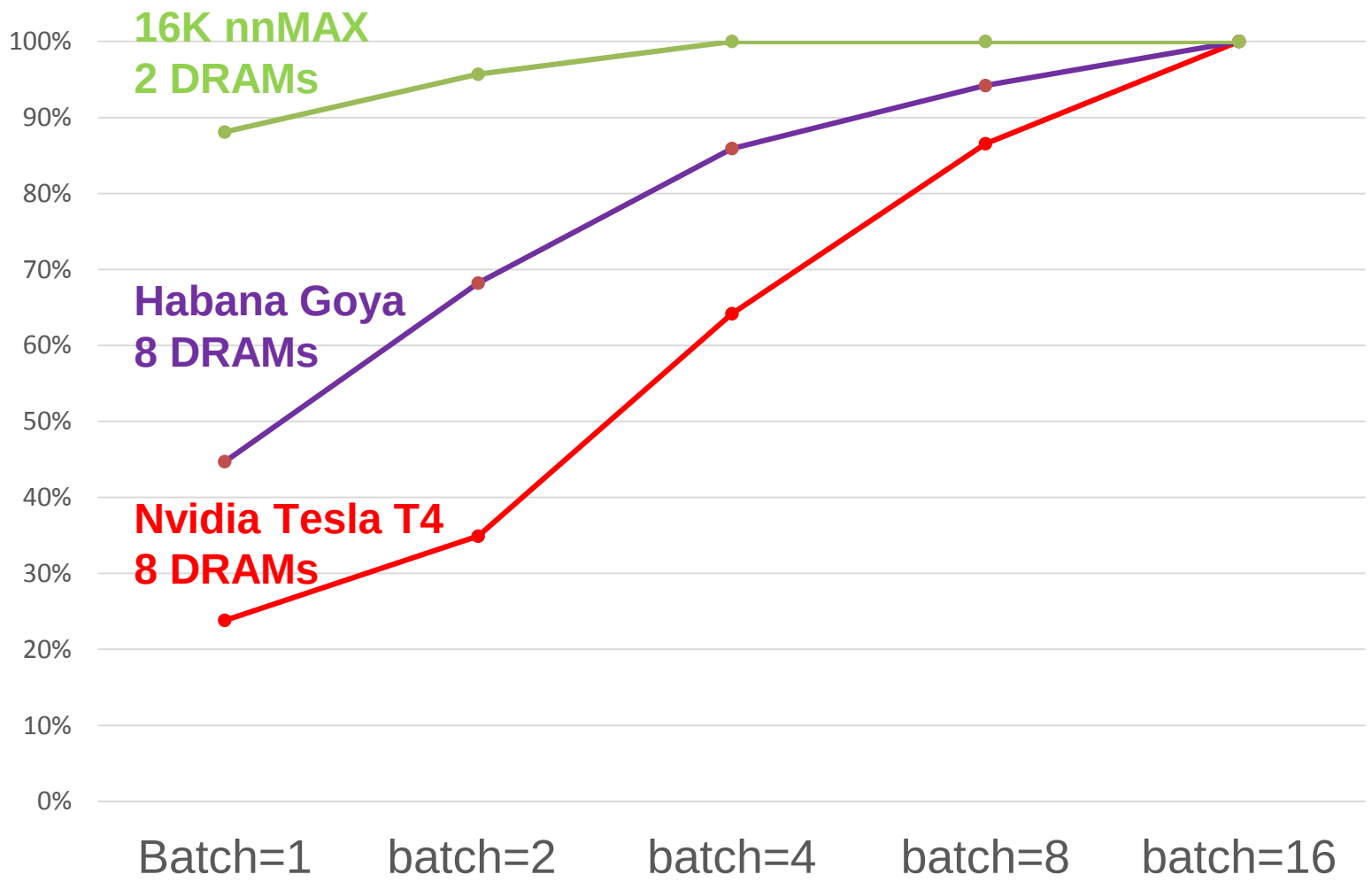
Solution for Autonomous Vehicles

- A high throughput, low power, low cost inference engine: nnMAX™
  - Maximize hardware utilization at **low batch size**
  - Minimize DRAM, maximize on-chip SRAM bandwidth utilization
- Modular: variable # of MACs, SRAM, DRAM BW optimized for application
- Scalable: double the resource, double the throughput

nnMAX: Best HW Utilization at Batch=1 with Less DRAM

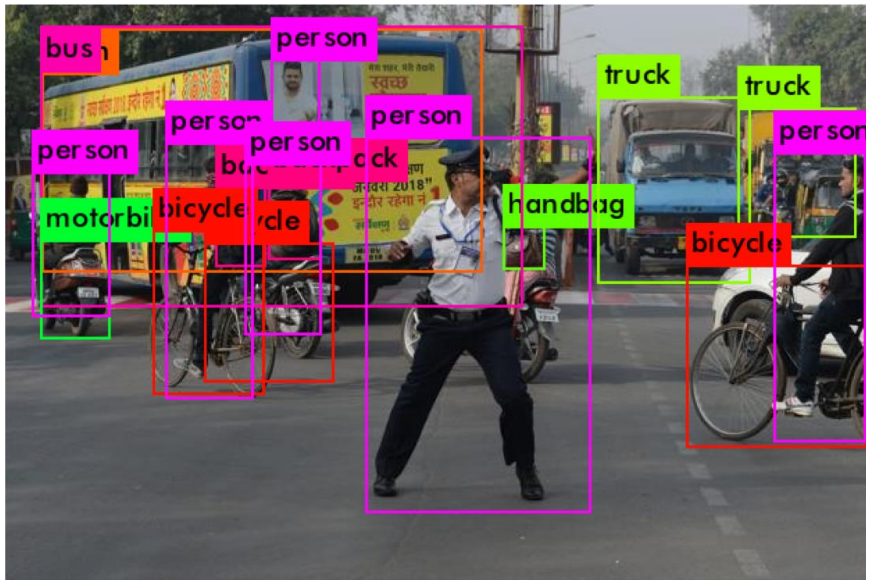
ResNet-50
Images/Sec

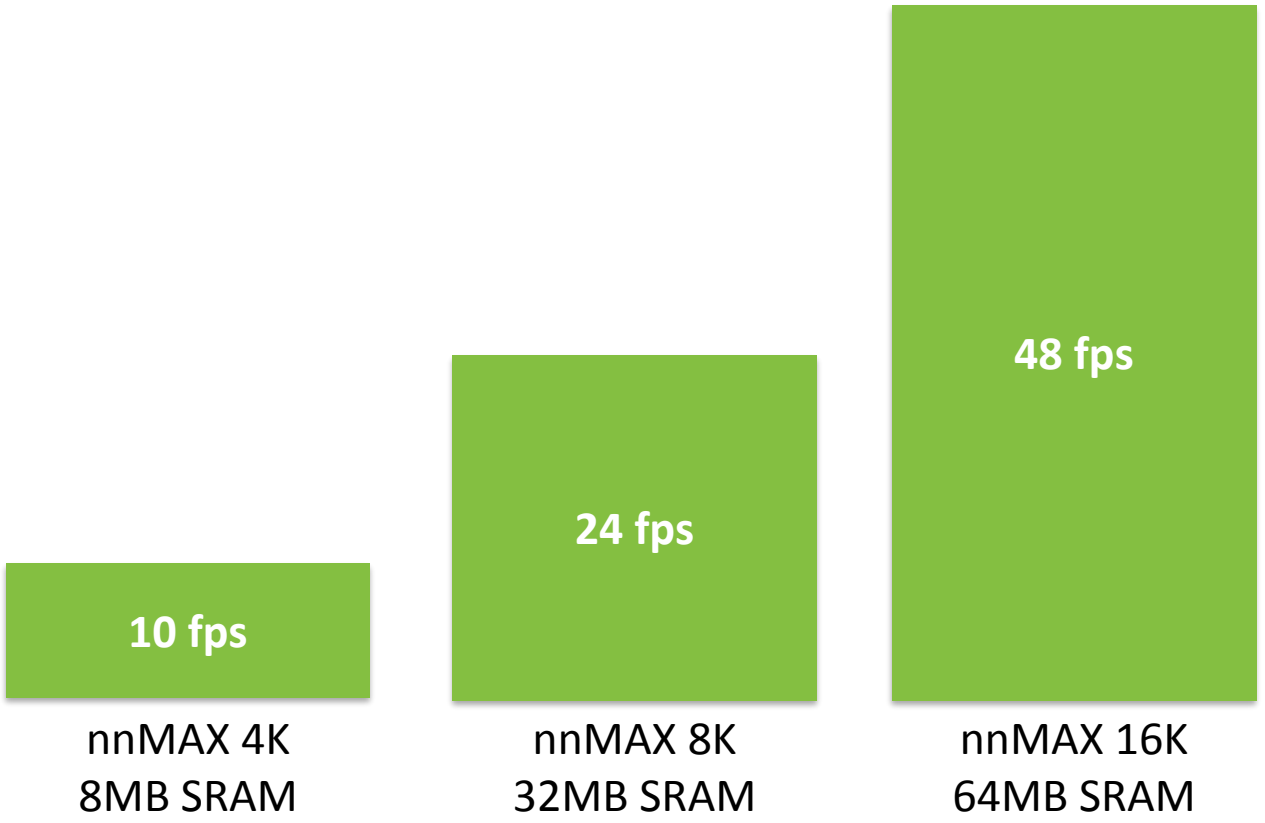
As a % of
Batch=16
Throughput



Throughput Scales with Resources

YOLOv3, 2MP per frame





nnMAX™ Inference IP

nnMAX builds on Flex Logix' leadership embedded FPGA

- EFLX® eFPGA is integrated into SoCs: like ARM does for processors
- Density and performance similar to leading FPGA chips
- Patented interconnects: XFLX™, ArrayLINX™, RAMLINX™
- Silicon proven in Sandia 180, TSMC 40/28/16/12; in fab for GF 14
- Customers announced:

BOEING®

大唐电信科技产业集团
DATANG TELECOM TECHNOLOGY & INDUSTRY GROUP

Sandia National Laboratories

DARPA

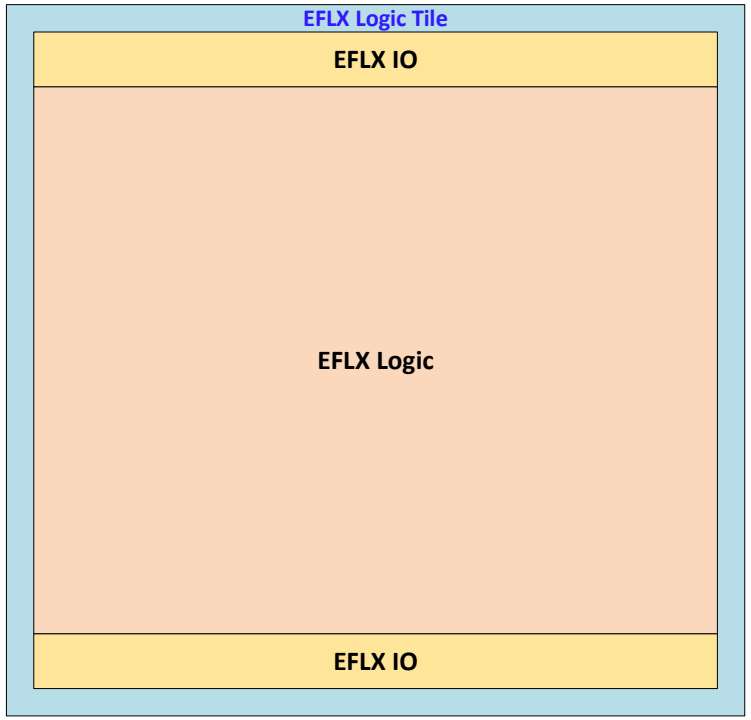
SiFive

HARVARD UNIVERSITY

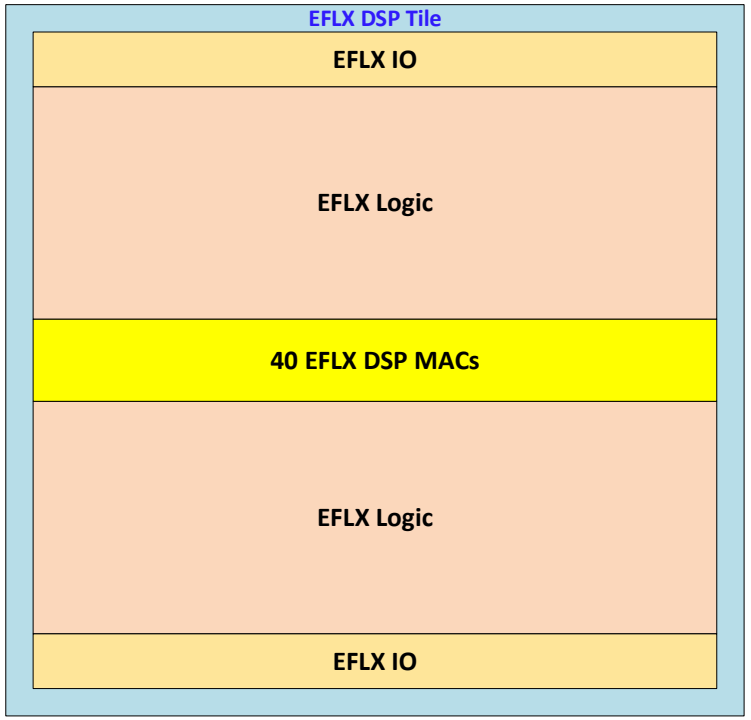
HiPer

nnMAX is an eFPGA Optimized for Inference Acceleration

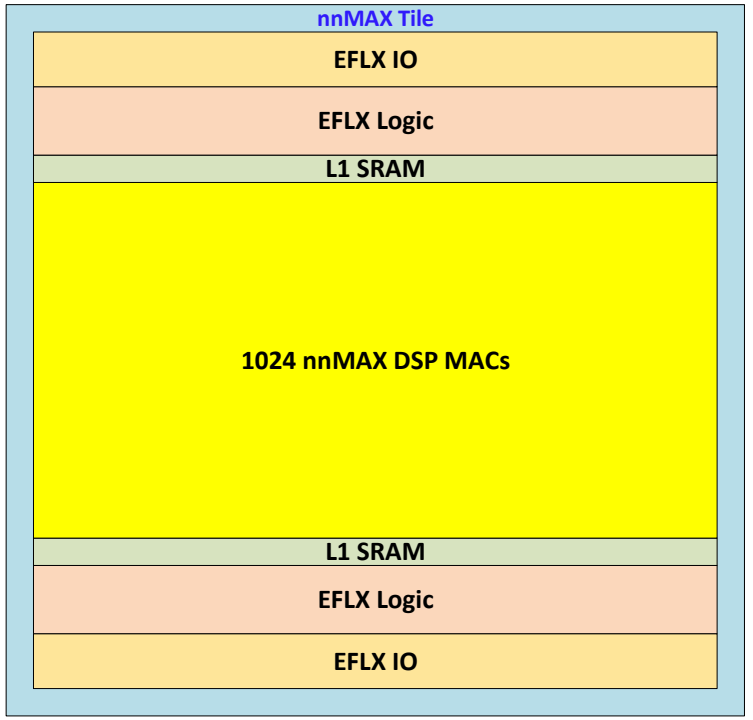
EFLX 4K eFPGA Logic



EFLX 4K eFPGA DSP



nnMAX 1K Tile*

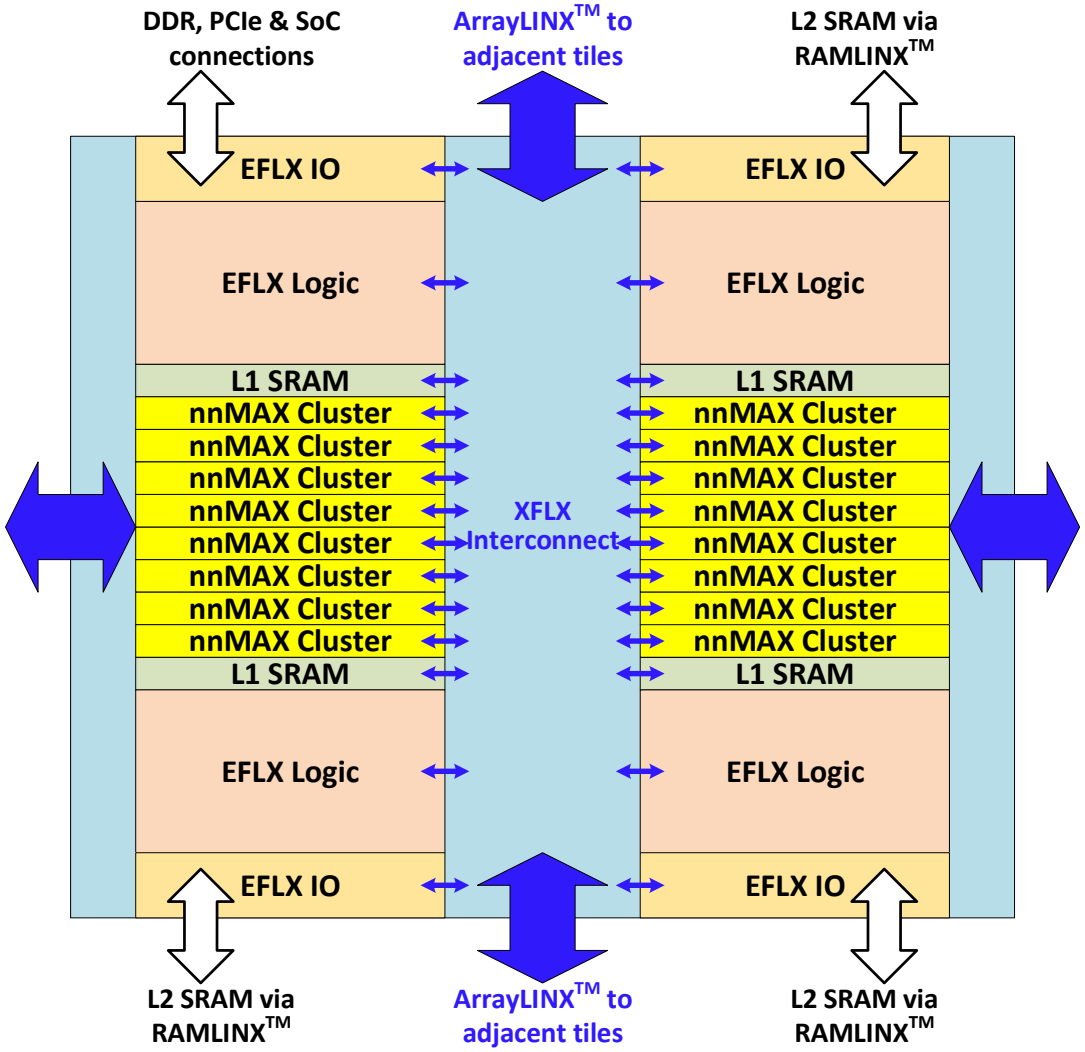


Leveraging production-proven silicon (T16FFC) & software

*architectural diagram, not to scale

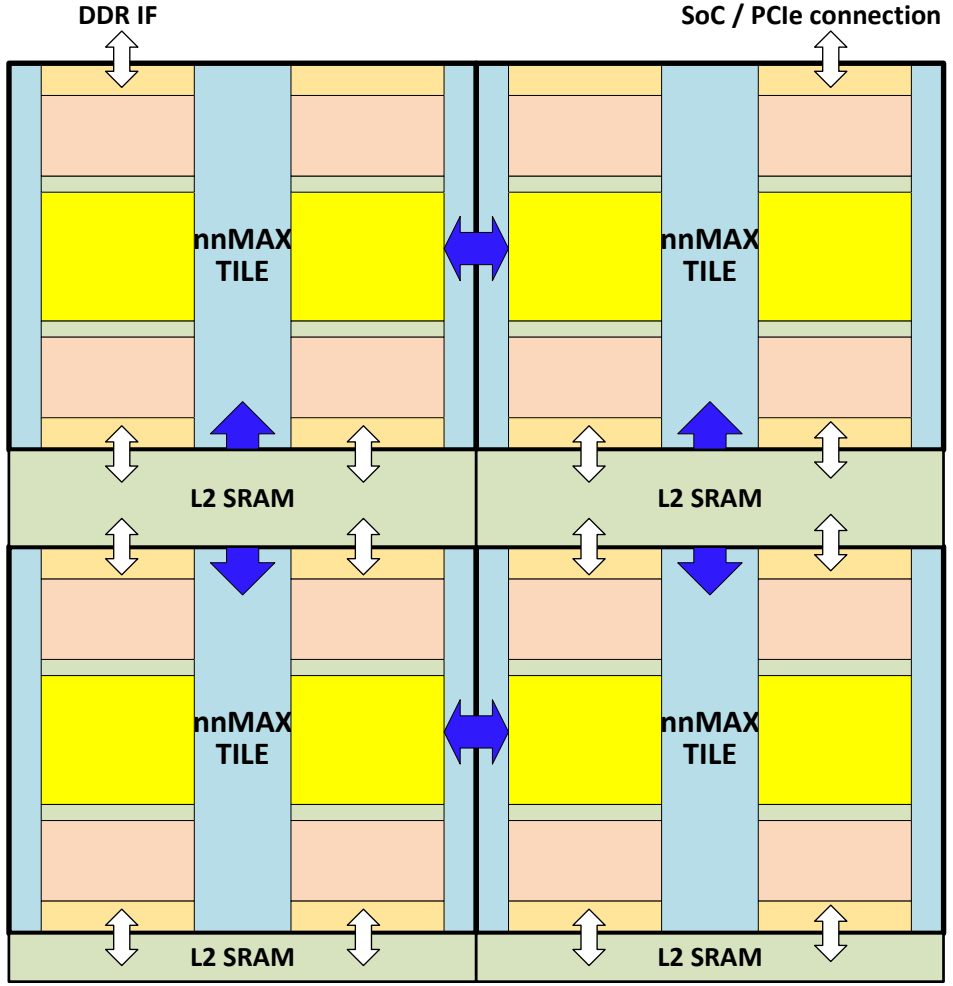
nnMAX 1K Inference Tile

- Based on silicon-proven EFLX eFPGA
  - 4mm^2 in TSMC16FFC
- 1024 configurable MACs @ 1.067GHz
  - INT8x8, INT16x8 at full rate
  - BFloat16x16, INT16x16 at half rate
  - Support mixed precision (INT8, INT16, BF16)
- Winograd acceleration for INT8
  - 2.25x performance gain for applicable layers
  - Automatically invoked by nnMAX Compiler
- Programmed by TensorFlow Lite/ONNX: multiple models can run simultaneously



nnMAX is Easy to Optimize: Modular Number of MACs, SRAM, DRAM

- nnMAX1024 tiles abut for any array size
  - Shown for a 2x2 array example
- Configurable L2 SRAM size
  - Support 1 – 4 MB per tile
- DRAM bandwidth is variable through reconfigurable I/O
  - Usually connecting to x32 or x64 LPDDR4



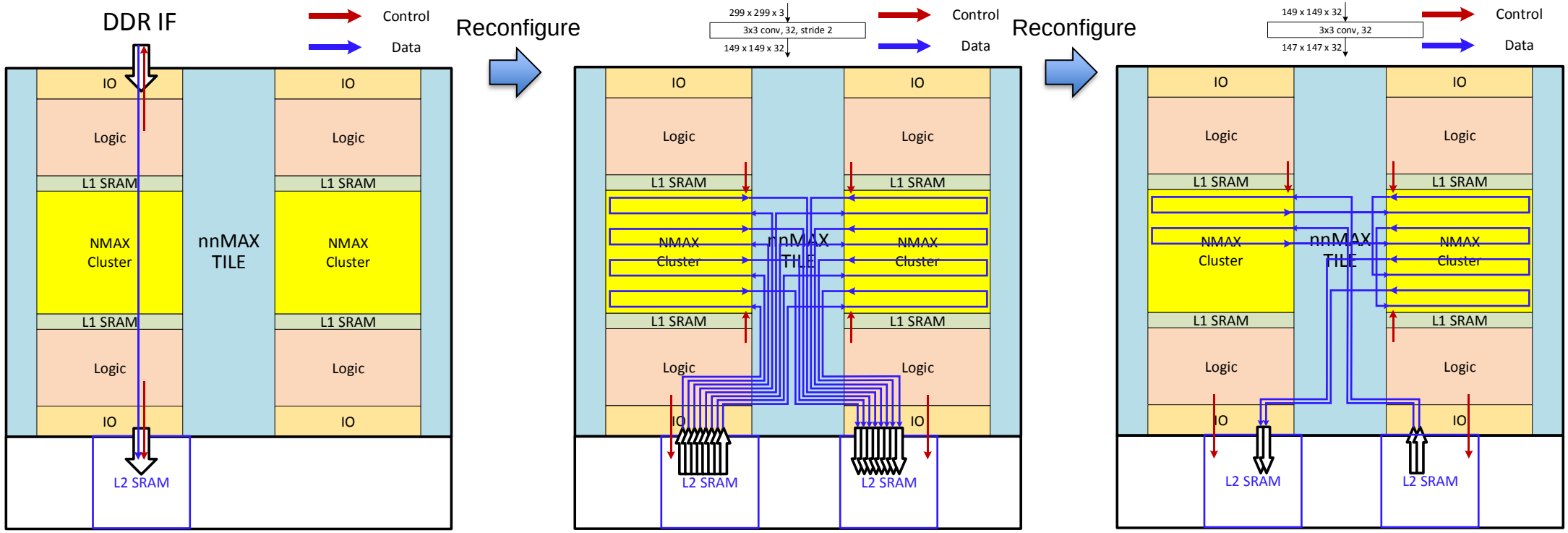
L2 SRAM is 1, 2 or 4MB per tile

Data Path Reconfigured Each Layer: High Utilization & SRAM BW

Input layer:
load DRAM into
nnMAX L2 SRAM

Layer 0:
16 nnMAX in parallel
L2 SRAM → L2 SRAM

Layer 1:
4 nnMAX in parallel, 3 in series
L2 SRAM → L2 SRAM



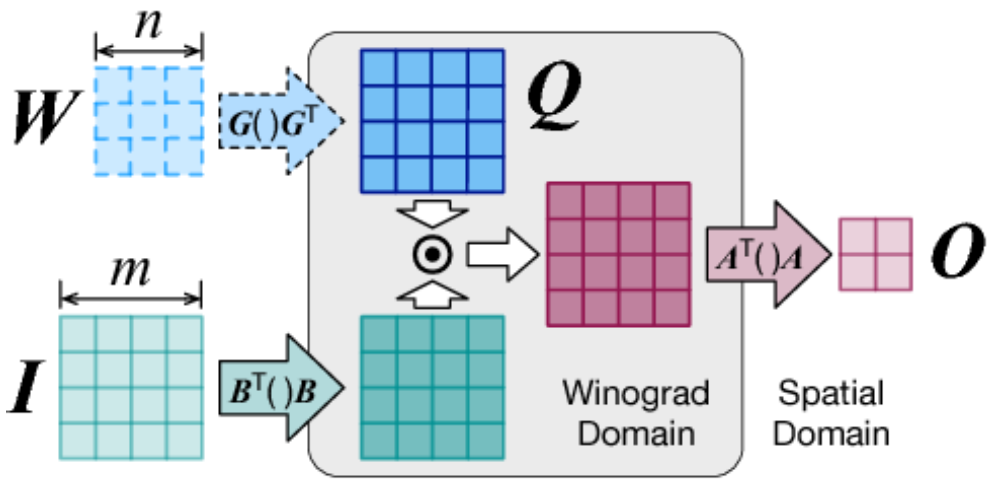
Localized data access & compute
ASIC-like performance yet fully reconfigurable

*architectural diagram, not to scale

INT8 Winograd Acceleration

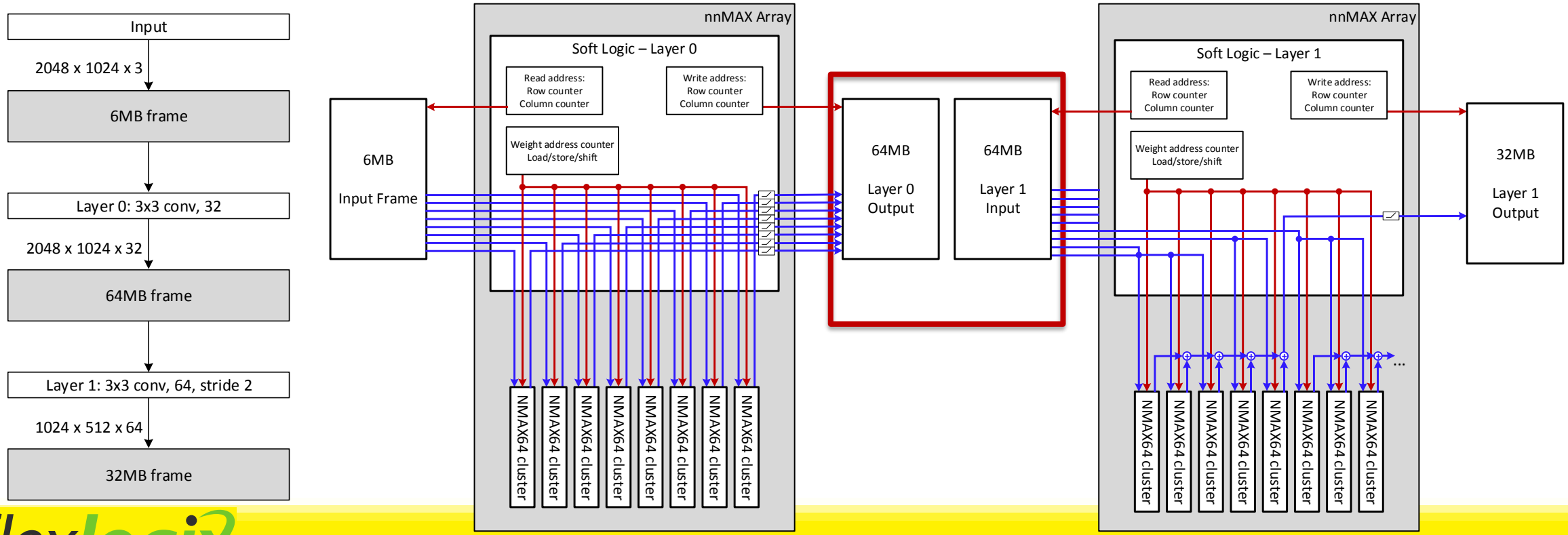
- Advantage: **2.25x** acceleration
  - Applies to 3x3 convolution layers with stride of 1
- Disadvantage: **2x** larger weights

  More complex datapath
  - 3x3 weights are transformed to 4x4
  - Increase in dynamic range and fractional precision
    8b weights are transformed to 8-12b
  - Inputs and outputs also need to be transformed and de-transformed in Winograd mode
- nnMAX performs all transformation on-the-fly
  - Result: Winograd performance without weight penalties



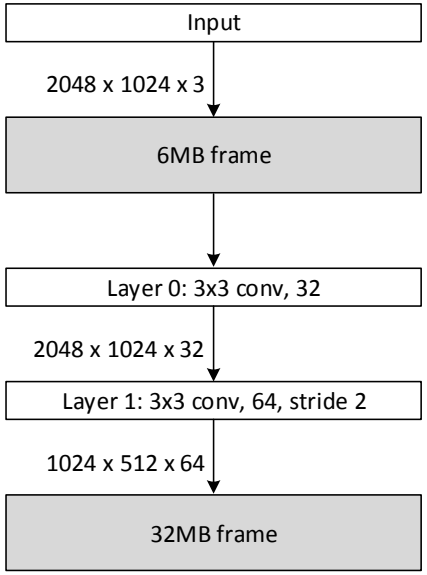
Large Activations Use Lots of DRAM Bandwidth

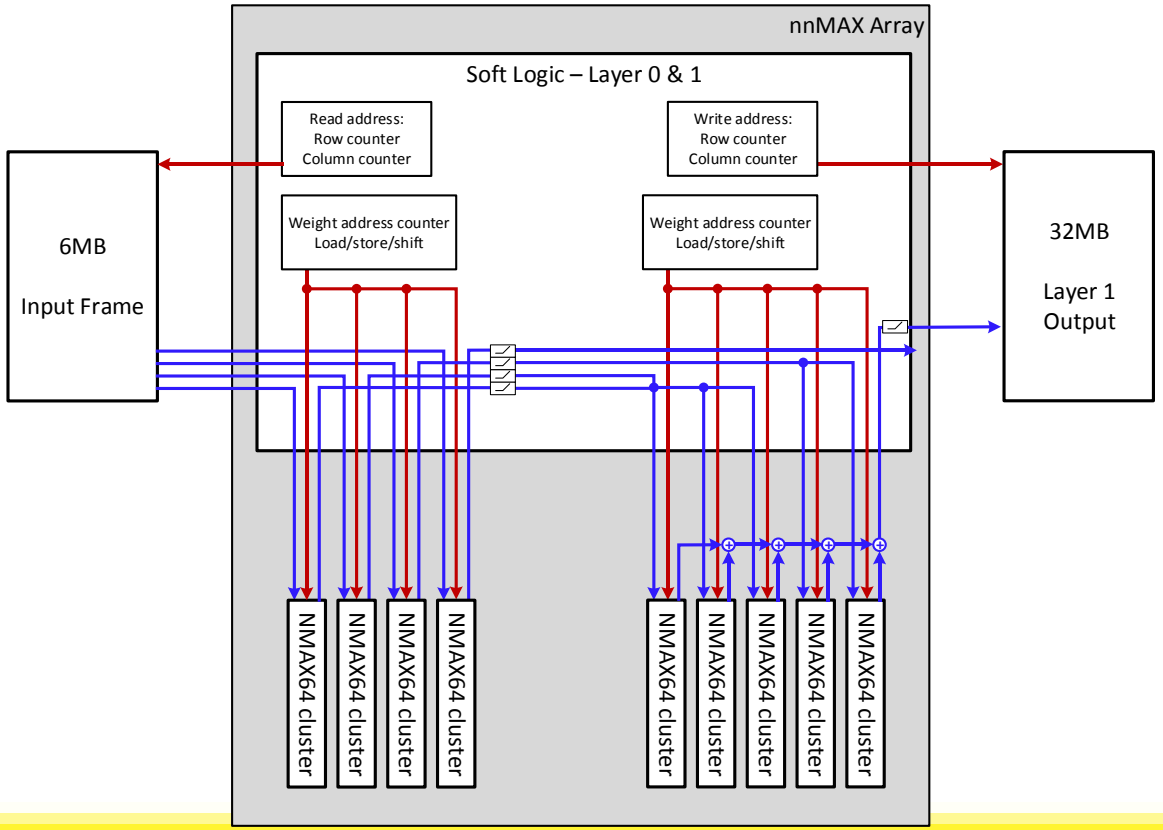
- Some layers have large intermediate frame sizes, requiring DRAM access and large DRAM BW
  - YOLOv3 example below, layer 0 outputs is **64 MB**, which may not fit in on-chip SRAM
  - Sending layer 0 output of 64 MB to DRAM (and re-reading it for layer 1) requires substantial DRAM bandwidth
  - Such large DRAM BW requirements would cause pipeline stall when processing these layers



Running Multiple Layers Cuts DRAM Bandwidth & Increases Throughput

- By running YOLOv3 layer 0 and 1 together, storing 64MB of layer 0 is avoided
  - Layer 0 activations directly streams into NMAX clusters processing layer 1
    Layer 0 & 1 (stride 2) is throughput-matched to 4:1





Flex Logix Technology Stack

- Hardware

InferX™ PCIe Cards

InferX Edge Inference Co-Processor ICs

nnMAX™ Inference IP

eFPGA/Interconnect Technology

- Software

TensorFlow Lite, ONNX

Software Driver

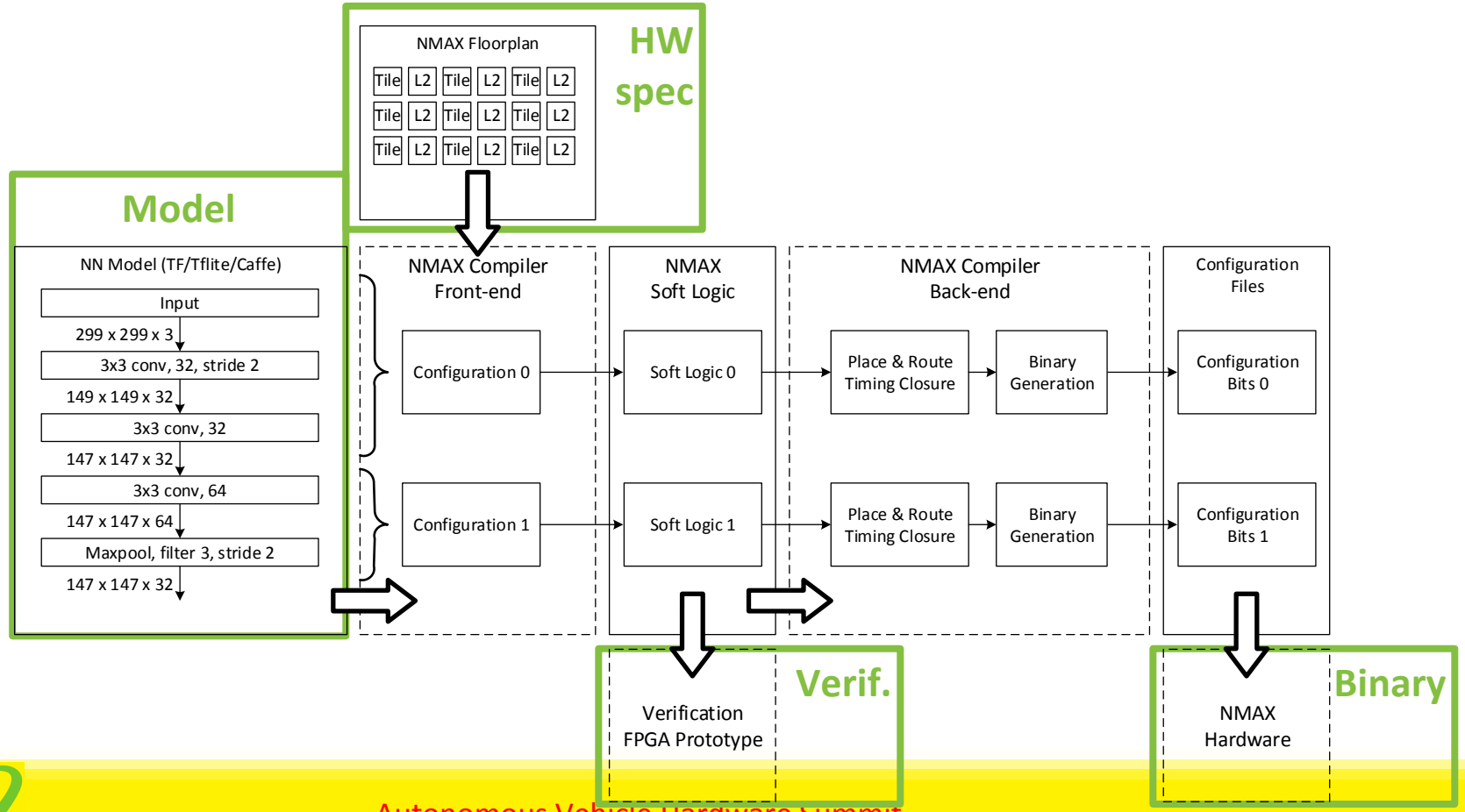
InferX/nnMAX Inference Compiler

eFPGA Place & Route Back-end

nnMAX Compiler and Driver

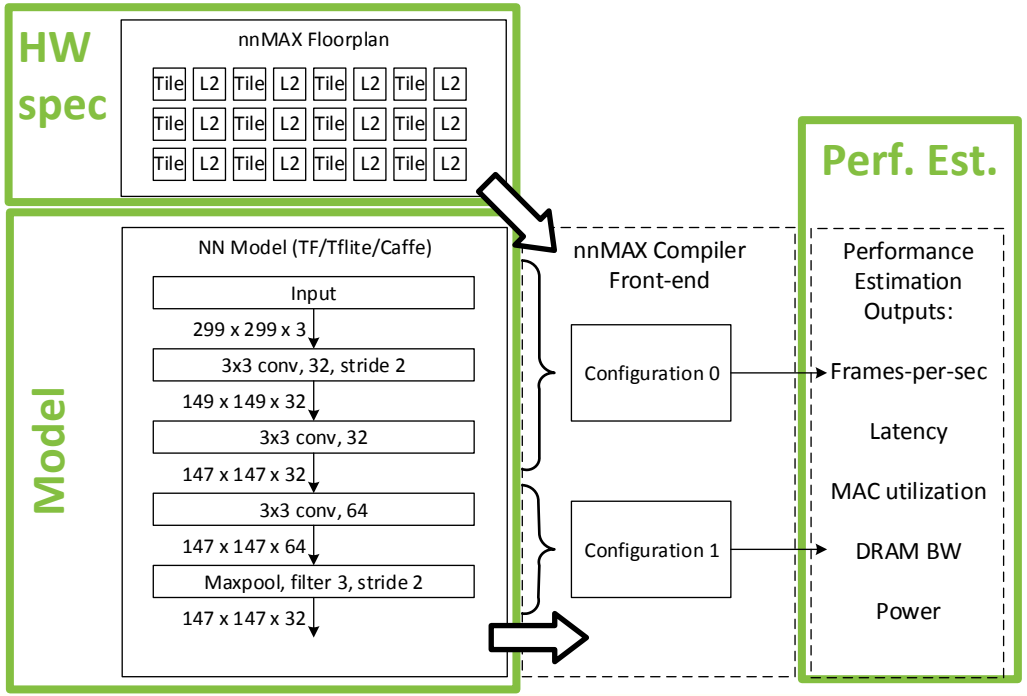
nnMAX Compiler Flow

- nnMAX Compiler front-end performs the NN model to soft-logic translation
- nnMAX Compiler back-end performs place-and-route, retiming, pipelining, and binary generation
  - Most of the nnMAX Compiler back-end is ported from EFLX Compiler, our eFPGA P&R tool



nnMAX Performance Estimation – Come see our Demo

- First part of the compiler is the performance estimation
- Accepts nnMAX floorplan and TF-lite (ONNX soon) model as input
  - Automatically partitions model across multi-layer configurations
  - Computes performance, latency, MAC utilization, DRAM BW per layer and per model




```
[INFO] *******Group Layer Native Mapping *********
[INFO] **********Performance Estimation************
[INFO] Frame/s @ 1GHz                    : 931.223
[INFO] DRAM Bandwidth (GB/s)             : 6.79634
[INFO] Latency (ms)                      : 10.7386
[INFO] TMAC/s                            : 2.46525
[INFO] MAC Utilization (%)               : 60.1869
[INFO] Area (mm^2)                       : 24.8
[INFO] SRAM Bandwidth (GB/s)             : 13.7782
[INFO] Actual Batch Size                 : 10
[INFO] Store Config in SRAM              : false
[INFO] Store Weights in SRAM             : false
[INFO] ******************************************
```

nnMAX Milestones

- Now: Performance Estimation
  - Accepts TF-Lite and (soon) ONNX models as input
- Q3
  - nnMAX IP Available for Integration
  - nnMAX silicon tapeout
    - Multiple nnMAX tiles and SRAM
    - More details at Linley Processor Conference, April 10th
- End 2019
  - nnMAX silicon ready