A 0.11 PJ/OP, 0.32-128 TOPS, SCALABLE MULTI-CHIP-MODULE-BASED DEEP NEURAL NETWORK ACCELERATOR DESIGNED WITH A HIGH-PRODUCTIVITY VLSI METHODOLOGY

Rangharajan Venkatesan, Yakun Sophia Shao, Brian Zimmer, Jason Clemons, Matthew Fojtik, Nan Jiang, Ben Keller, Alicia Klinefelter, Nathaniel Pinckney, Priyanka Raina, Stephen G. Tell, Yanqing Zhang, William J. Dally, Joel S. Emer, C. Thomas Gray, Stephen W. Keckler & Brucek Khailany

RESEARCH TESTCHIP GOALS

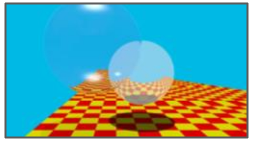
Develop and Demonstrate Underlying Technologies for Efficient DL Inference

NVIDIA RESEARCH OVERVIEW

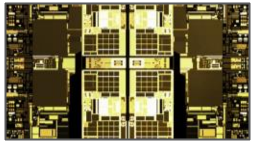
Research Teams:

Graphics, Deep Learning, Robotics, Computer Vision, Parallel Architectures, Programming Systems, Circuits, VLSI, Networks

Recent Works:



RTX



NVSwitch

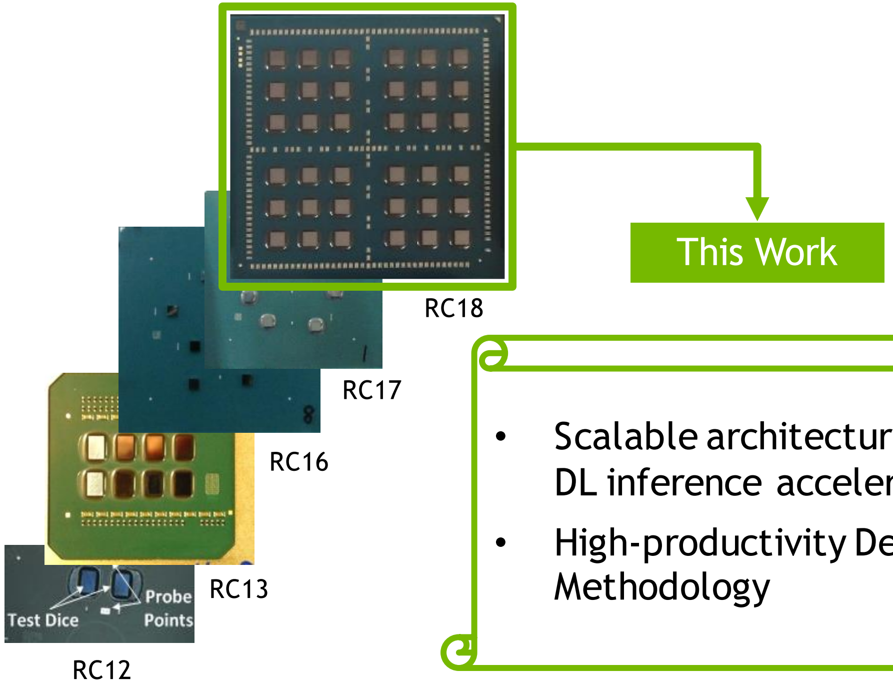
Noise-to-Noise Denoising

CuDNN

CNN Image Inpainting

Progressive GAN

TESTCHIPS

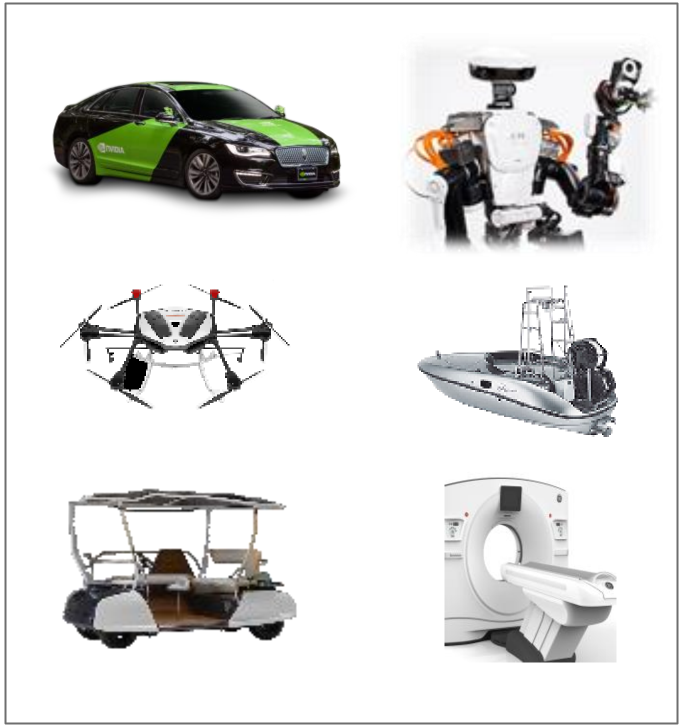


- Scalable architecture for DL inference acceleration
- High-productivity Design Methodology

VAST WORLD OF AI INFERENCE

Creating A Massive Market Opportunity

GENERAL PURPOSE COMPUTERS

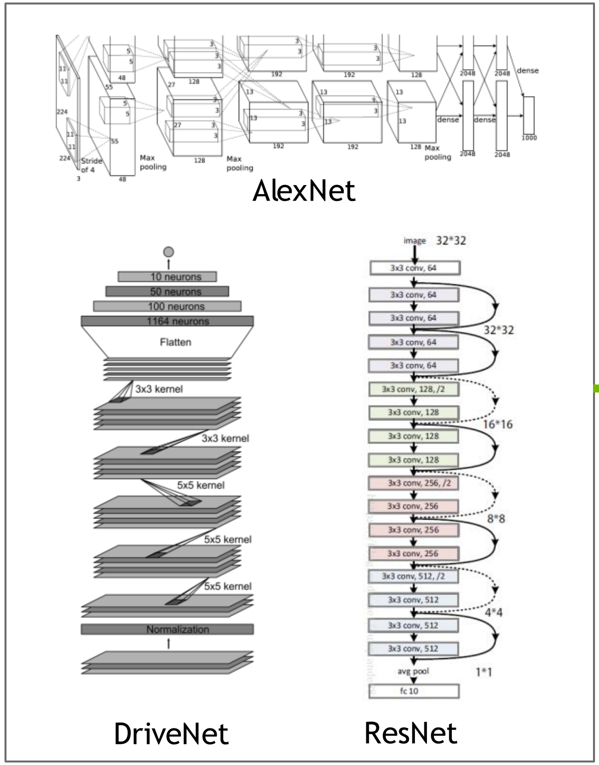


EMBEDDED COMPUTERS

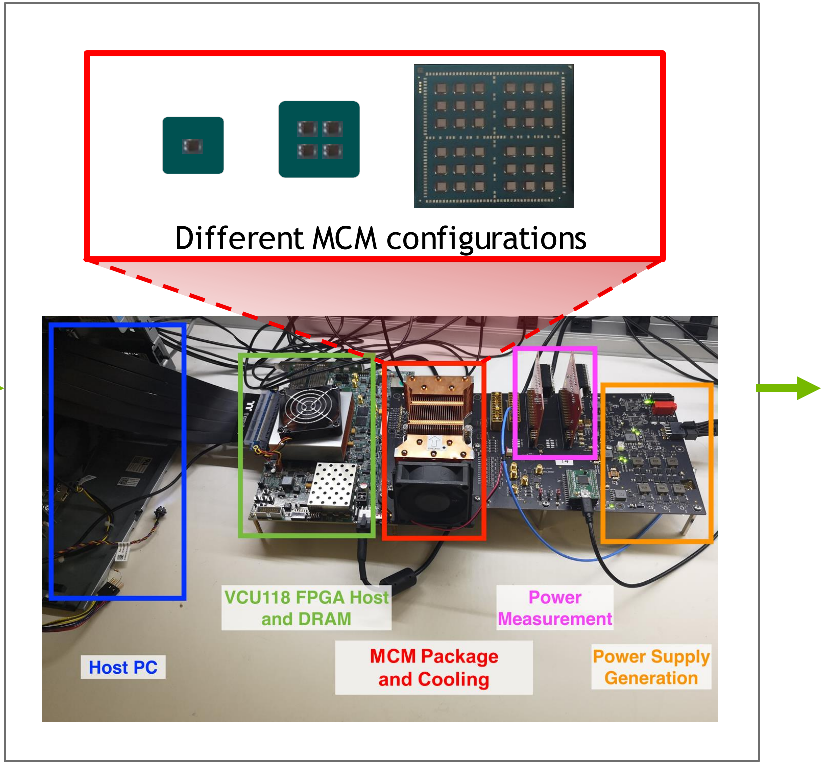
EMBEDDED DEVICES

TARGET APPLICATIONS

Image Classification with Convolutional Neural Networks



Deep Learning Models



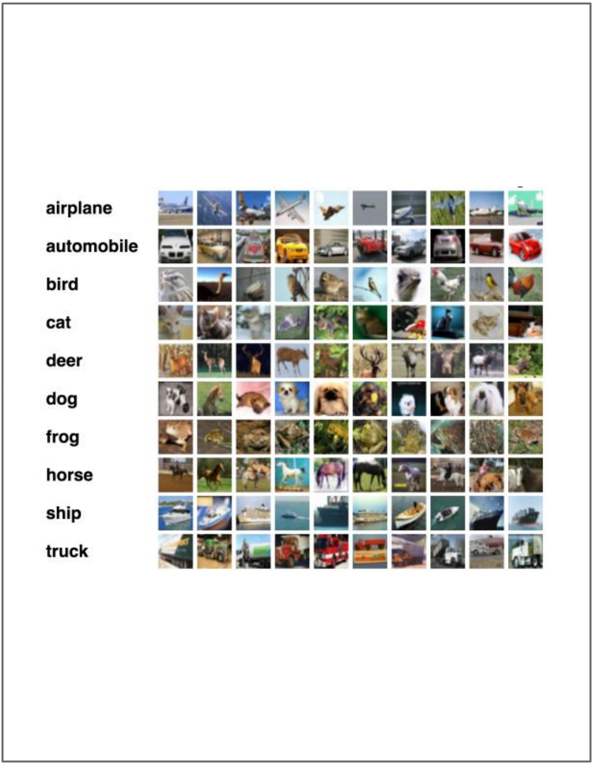
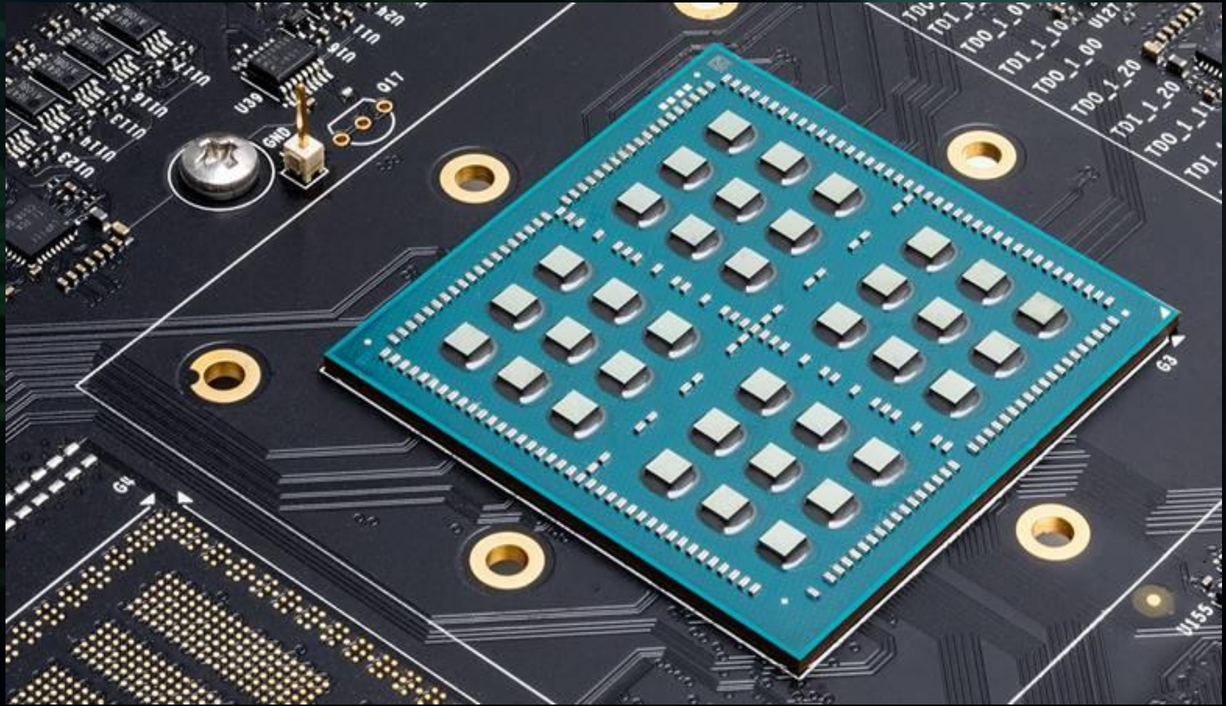


Image Classification

Ref: Krizhevsky et al., NeurIPS, 2012. Bojarski et al., CoRR 2016. He et al., CoRR 2015



SCALABLE DEEP LEARNING INFERENCE ACCELERATOR

MULTI-CHIP-MODULE (MCM) ARCHITECTURE

Demonstrate:

Low-effort scaling to high inference performance

Ground Reference Signaling (GRS) as an MCM interconnect

Network-on-Package architecture

Advantages:

Overcome reticle limits

Higher yield

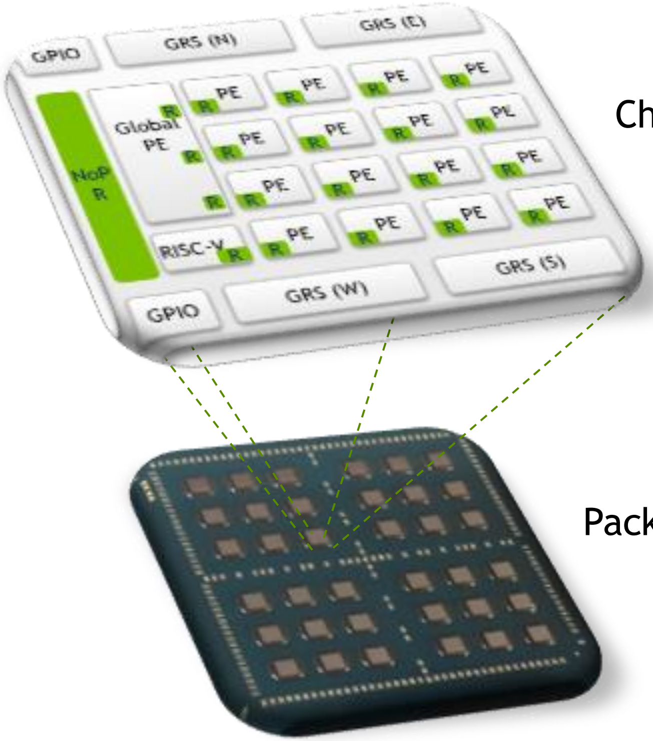
Lower design cost

Mix process technologies

Agility in changing product SKUs

Challenges:

Area and power overheads for inter-chip interfaces



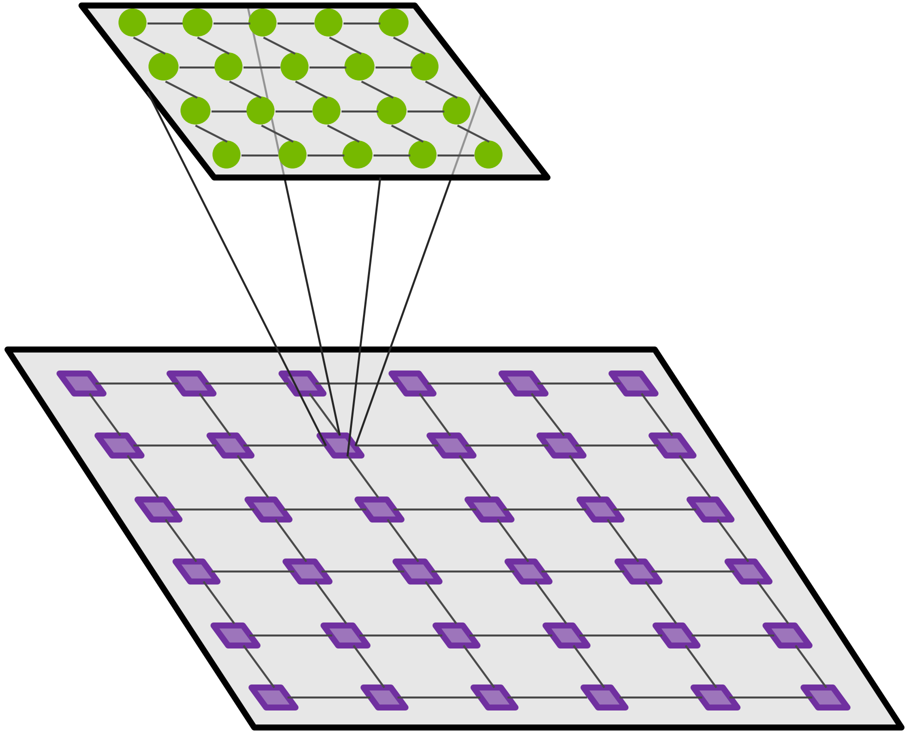
Chip

Package

Ref: Zimmer et al., VLSI 2019

HIERARCHICAL COMMUNICATION ARCHITECTURE

Network-on-Package (NoP) and Network-on-Chip (NoC)



NETWORK-ON-CHIP (NoC)

4x5 mesh topology connects 16 PEs, one Global PE, and one RISC-V

Cut-through routing with Multicast support

10ns per hop, ~70Gbps per link (at 0.72V)

NETWORK-ON-PACKAGE (NoP)

6x6 mesh topology connects 36 chips in package.

A single NoP router per chip with 4 interface ports to NoC

Configurable routing to avoid bad links/chip

~20ns per hop, 100 Gbps per link (at max)

GROUND-REFERENCED SIGNALING (GRS)

High Bandwidth, Energy-efficient Inter-chip Communication

High Speed

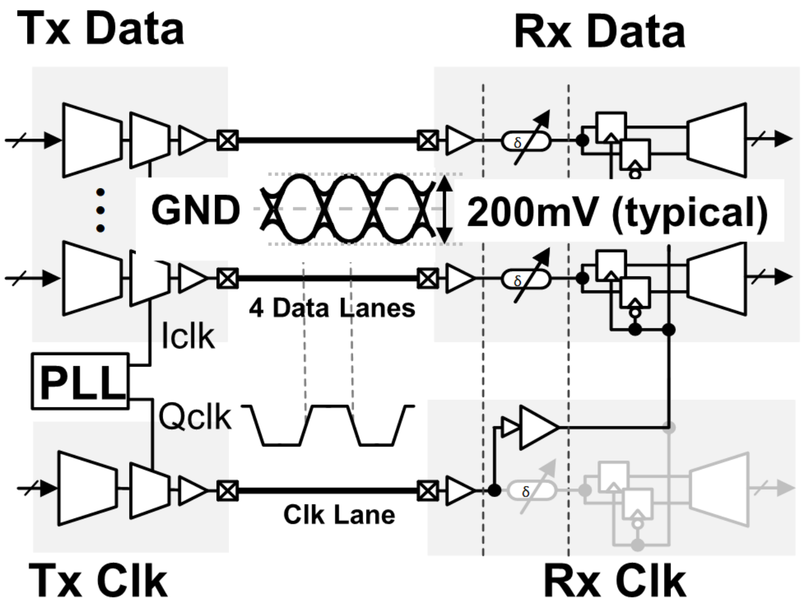
11-25 Gbps per pin

High energy efficiency

Low voltage swing (~200mV)
0.82-1.75 pJ/bit

High area efficiency

Single-ended links
4 data bumps + 1 clock bump per GRS link

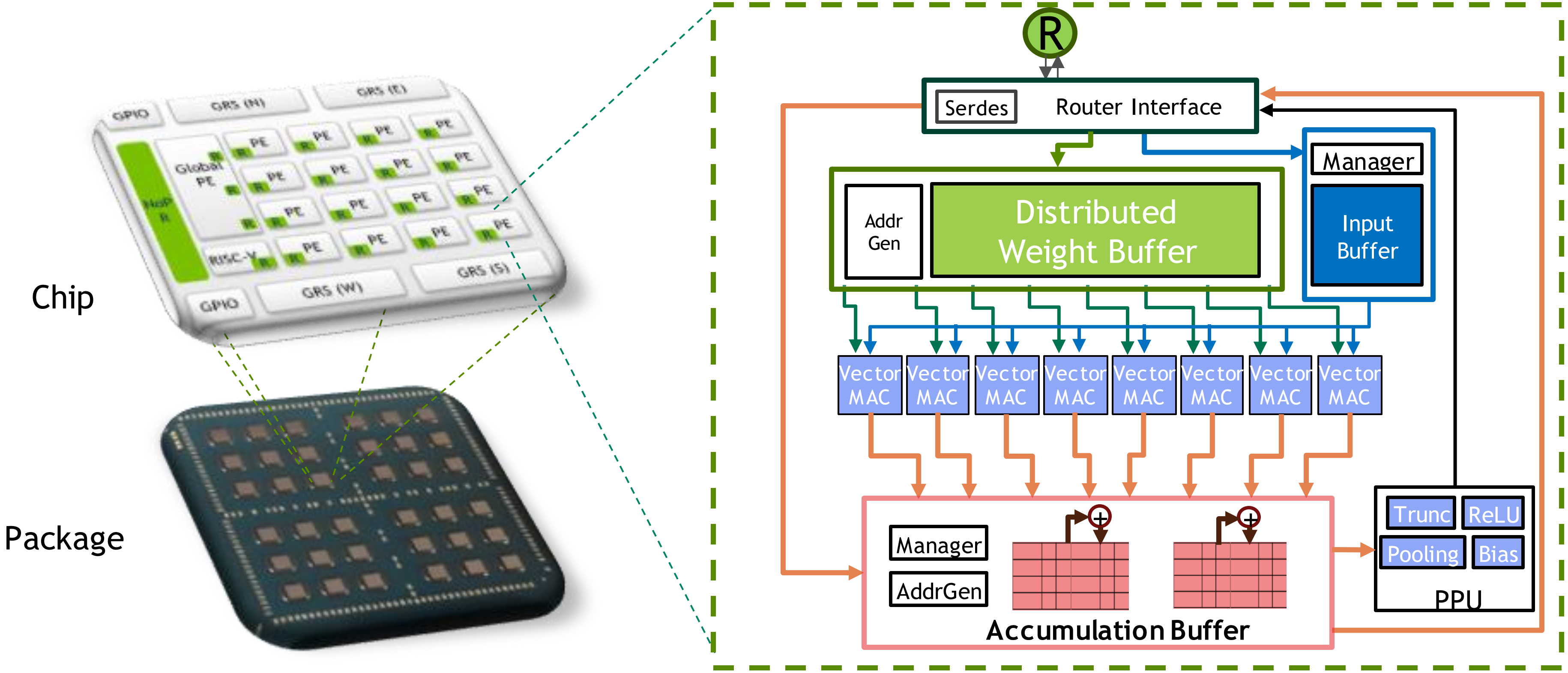


GRS Macro

Ref: Poulton et al., JSSC 2019

SCALABLE DL INFERENCE ACCELERATOR

Tiled Architecture with Distributed Memory



Ref: Zimmer et al., VLSI 2019

Ref: Sijstermans et al., HotChips 2018

SCALABLE DL INFERENCE ACCELERATOR

CNN Layer Execution

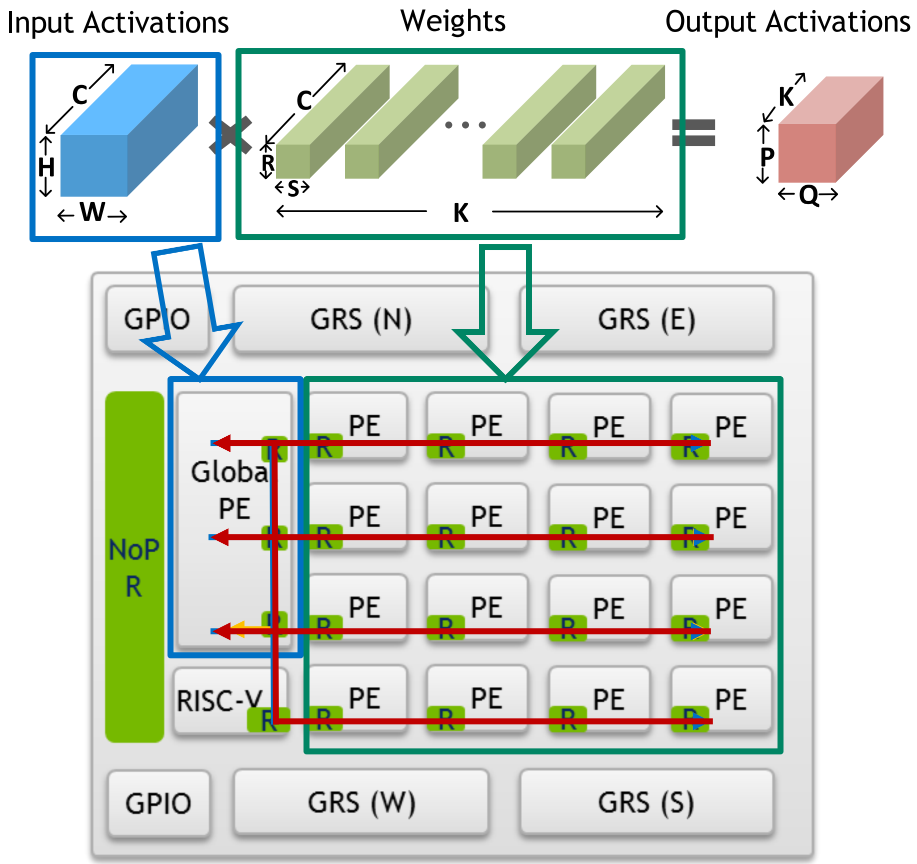
Distribute weights across PEs

Load Input Activation to Global PE

RISC-V configures control registers

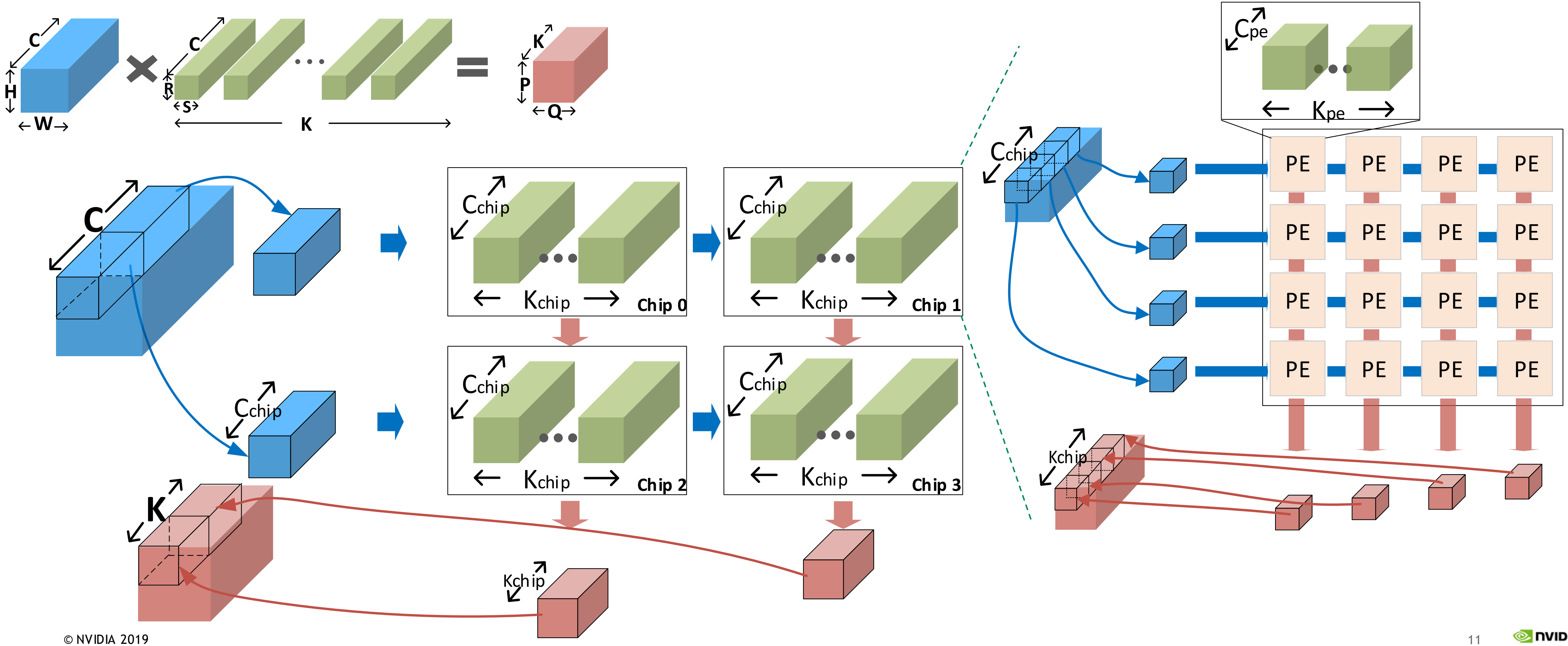
Stream input activations to PEs

Store output activations to Global PE



SCALING DL INFERENCE ACROSS NOP/NOC

Tiling Convolutional Layer Across Chips and Processing Elements



FABRICATED MCM-BASED ACCELERATOR

NVResearch Prototype: 36 Chips on Package in TSMC 16nm Technology

High speed interconnects using Ground Reference Signaling (GRS)

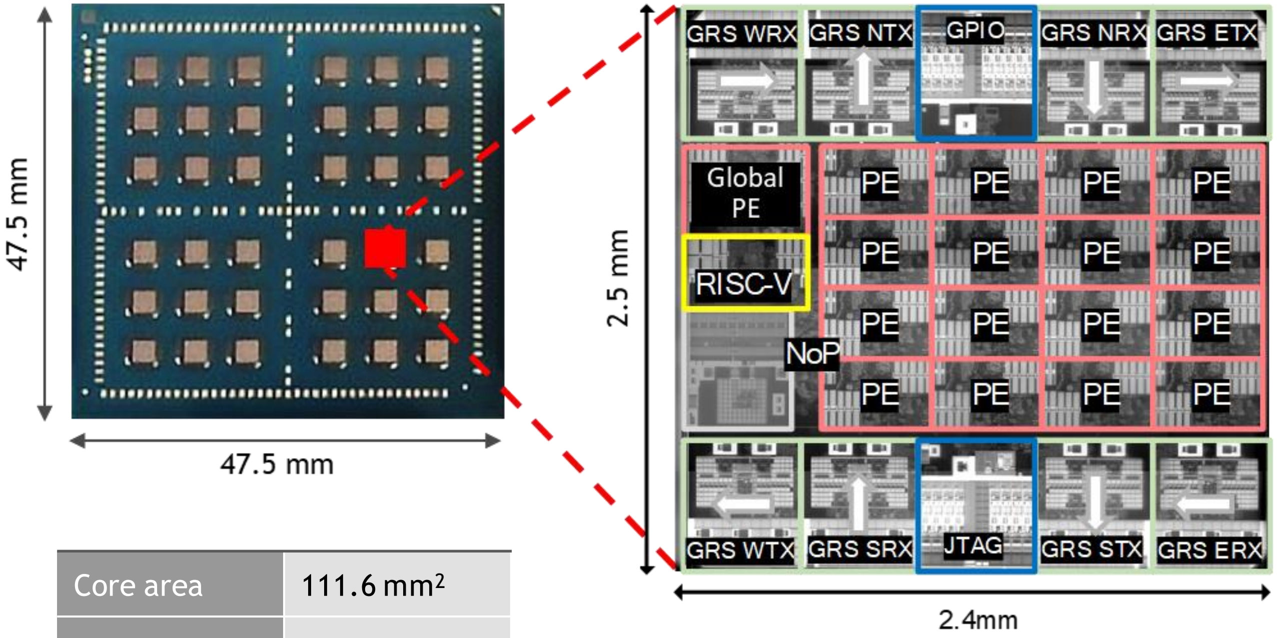
- 100 Gbps per link

Efficient Compute tiles

- 9.5 TOPS/W, 128 TOPS

Low Design Effort

- Spec-to-Tapeout in 6 months with <10 researchers



Core area	111.6 mm^2
Voltage	0.52-1.1 V
Frequency	0.48-1.8 GHz

Ref: Zimmer et al., VLSI 2019

HIGH PRODUCTIVITY DESIGN METHODOLOGY

HIGH-PRODUCTIVITY DESIGN APPROACH

Enables faster time-to-market and more features to each SoC

RAISE HARDWARE DESIGN LEVEL OF ABSTRACTION

Use High-level languages

e.g. C++ instead of Verilog

Use Automation

e.g. High-Level Synthesis (HLS)

Use libraries/generators

MatchLib

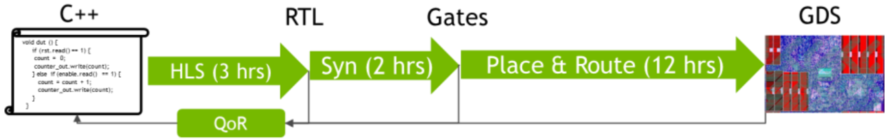
AGILE VLSI DESIGN

Small teams, jointly working on architecture, implementation, VLSI

Continuous integration with automated tool flows

Agile project management techniques

24-hour spins from C++-to-layout



```
void dut () {
    if (rst.read() == 1) {
     count = 0;
     counter_out.write(count);
    } else if (enable.read() == 1) {
      count = count + 1;
      counter_out.write(count);
    }
  }
```

OBJECT-ORIENTED HIGH-LEVEL SYNTHESIS

“Push-button” C++-to-gates flow

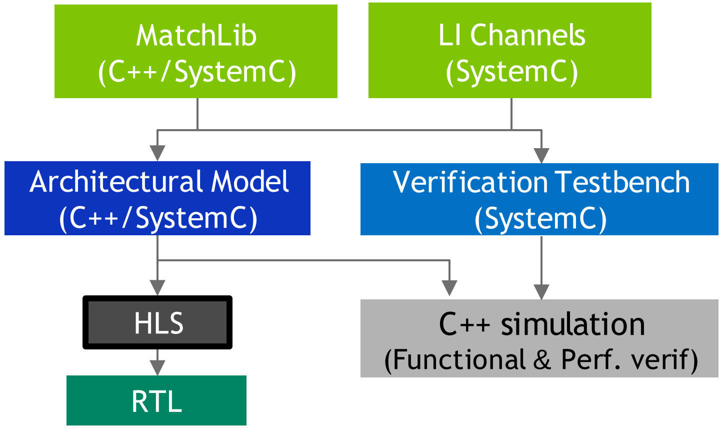
Leverage HLS tools to design with C++ and SystemC models

MatchLib: Modular Approach To Circuits and Hardware Library

“STL/Boost” for Hardware Design
Synthesizable hardware library developed by NVIDIA research
Highly-parameterized, high QoR implementation
Available open-source: *https://github.com/NVlabs/matchlib*

Latency-Insensitive (LI) Channels

Enable modularity in design process
Decouple computation & communication architectures

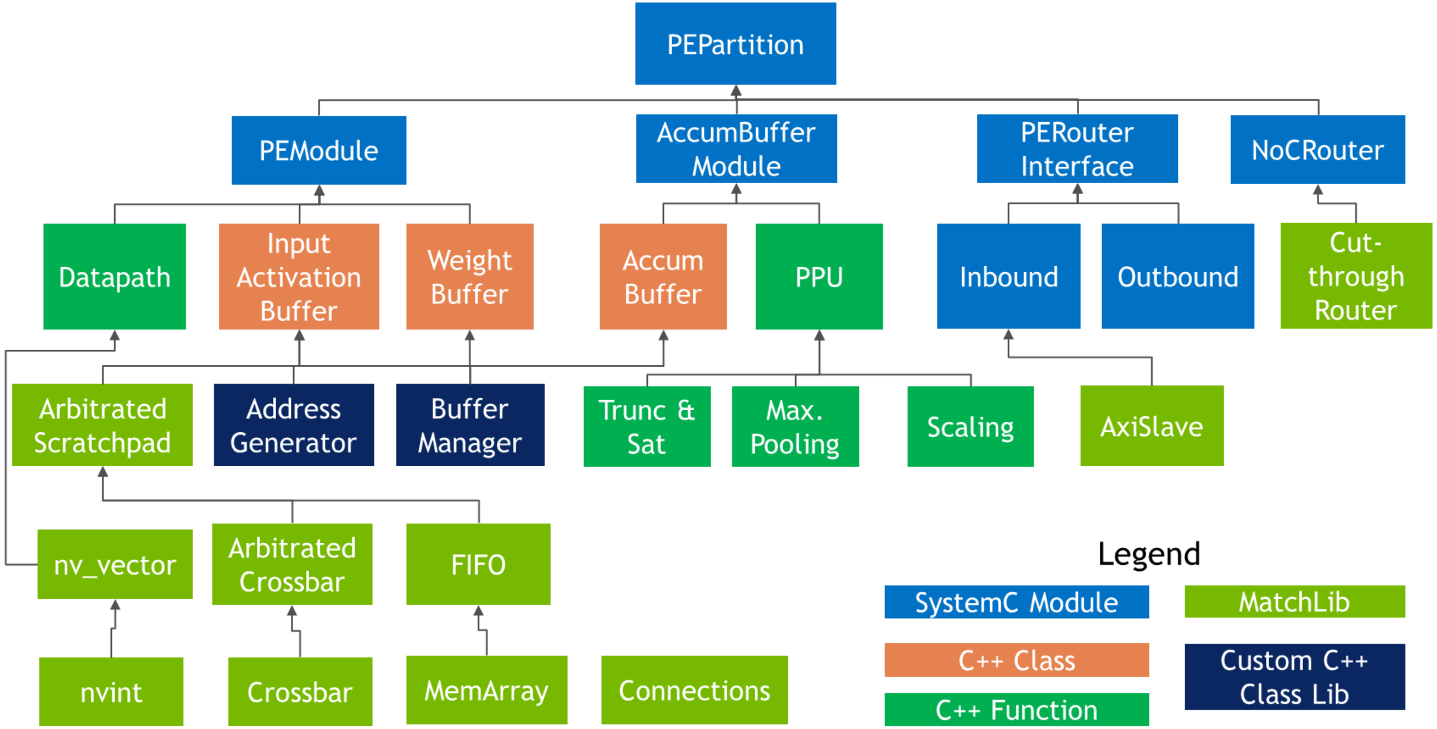


6 months from spec-to-tapeout with <10 researchers

Ref: Khailany et al., DAC 2018

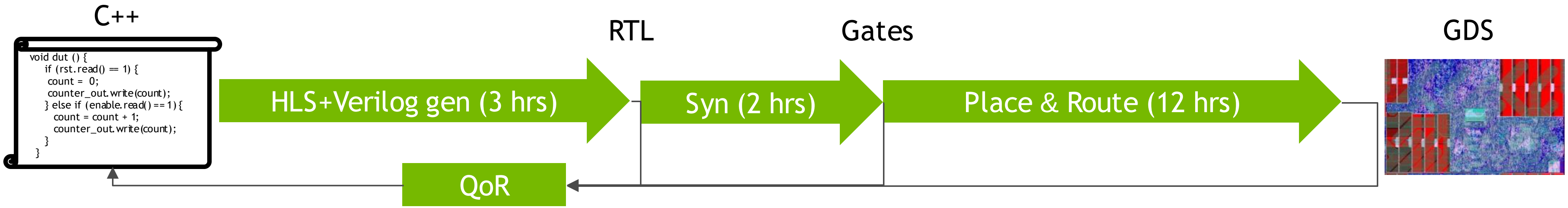
PROCESSING ELEMENT IMPLEMENTATION

Reuse, Modularity, Hierarchical Design



AGILE VLSI DESIGN TECHNIQUES

Daily “C++ to Layout” Spins



```
void dut () {
    if (rst.read() == 1) {
     count =  0;
     counter_out.write(count);
    } else if (enable.read() == 1) {
      count = count + 1;
      counter_out.write(count);
    }
 }
```

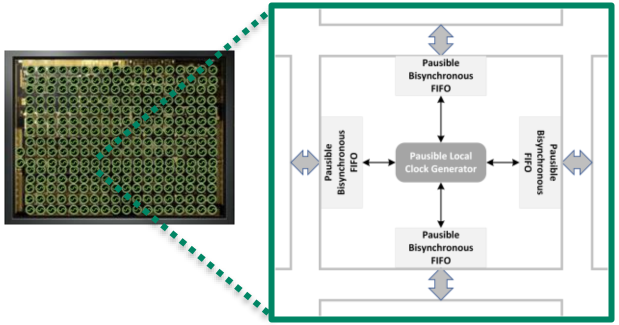
Agile, incremental approach to design closure during march-to-tapeout phase

Small, abutting partitions for fast place and route iterations

Globally asynchronous locally synchronous pausible adaptive clocking scheme

Fast, error-free clock domain crossings
“Correct by construction” top-level timing closure

RTL bugs, performance, and VLSI constraints converge together


Pausible Bisynchronous FIFO
Pausible Bisynchronous FIFO
Pausible Local Clock Generator
Pausible Bisynchronous FIFO
Pausible Bisynchronous FIFO

Ref: Fojtik et al., ASYNC 2019

EXPERIMENTAL RESULTS

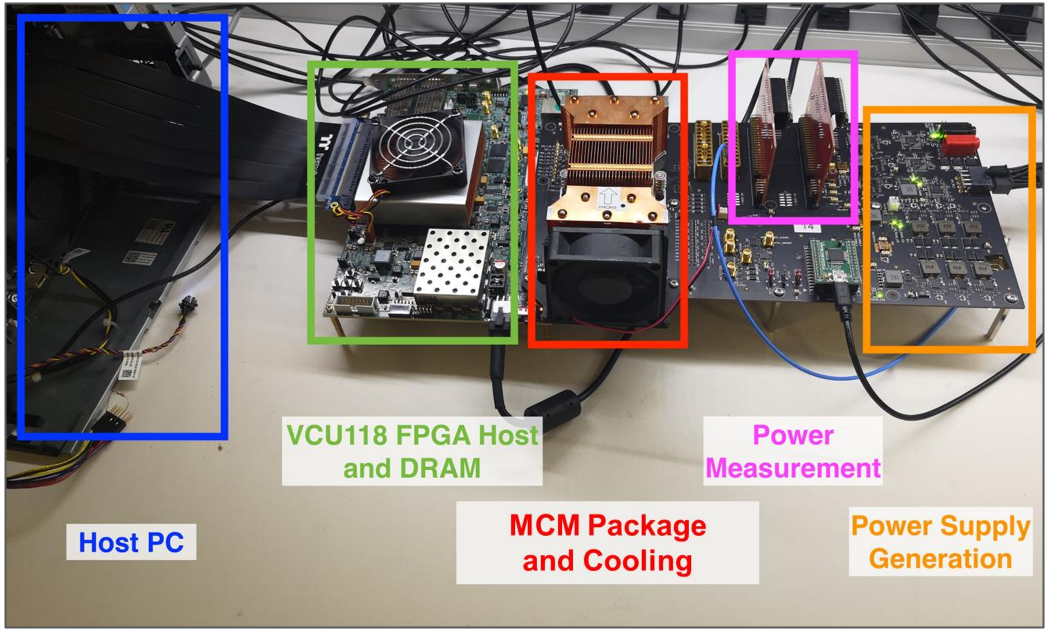
MEASUREMENT SETUP

Measurements begin after weights and activations are loaded from FPGA DRAM

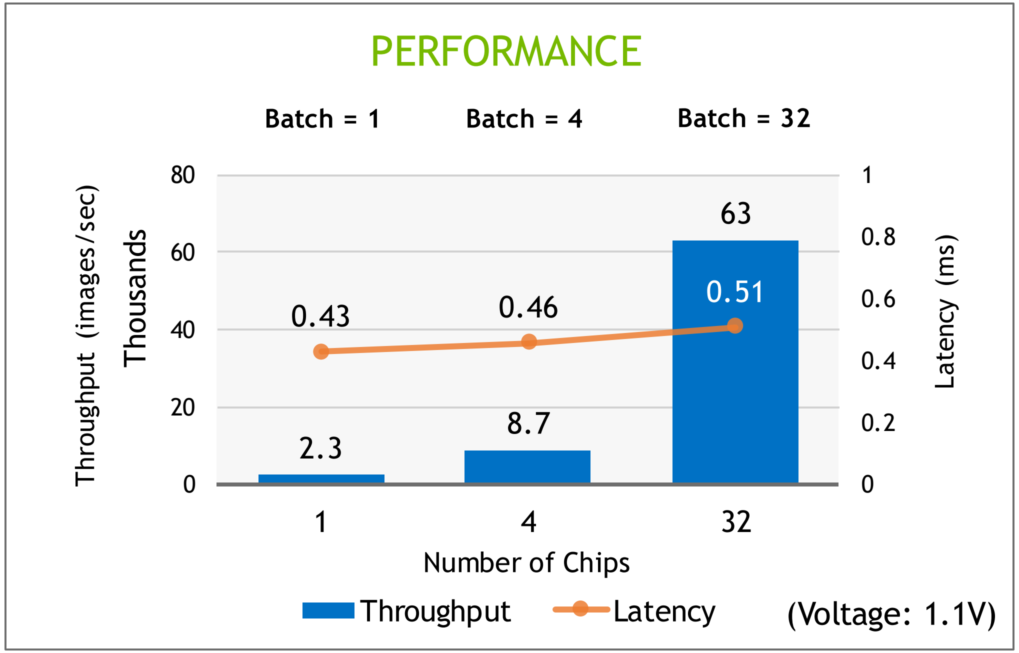
- Weights are loaded to PE memory
- Activations are loaded to Global PE

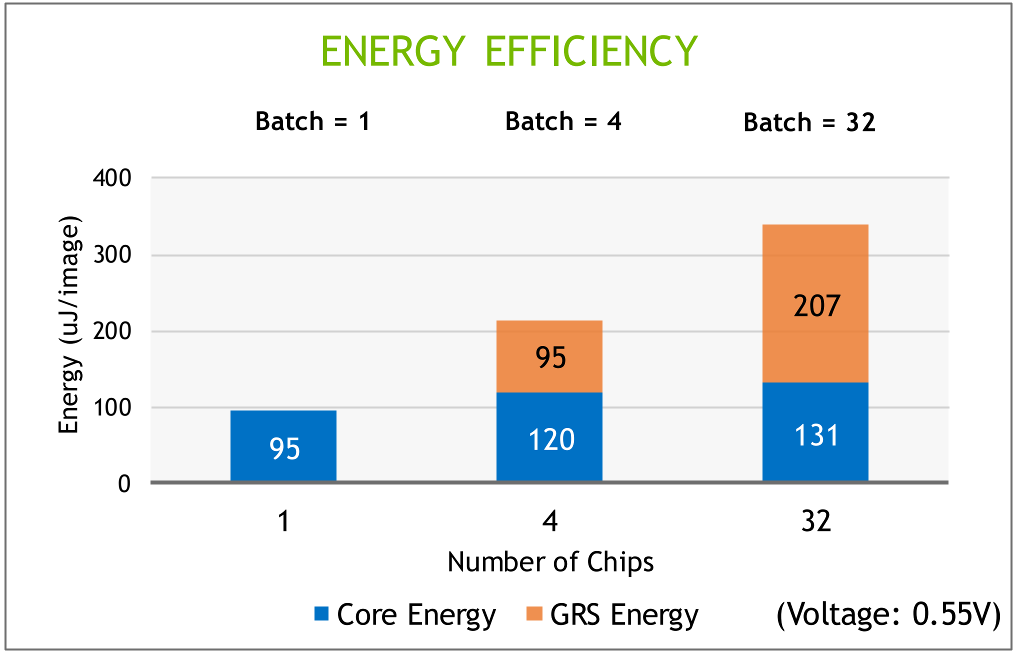
Operating points

- Max. Performance: 1.1V
- Min. Energy: 0.55V



RESULTS: WEAK SCALING WITH DRIVENET





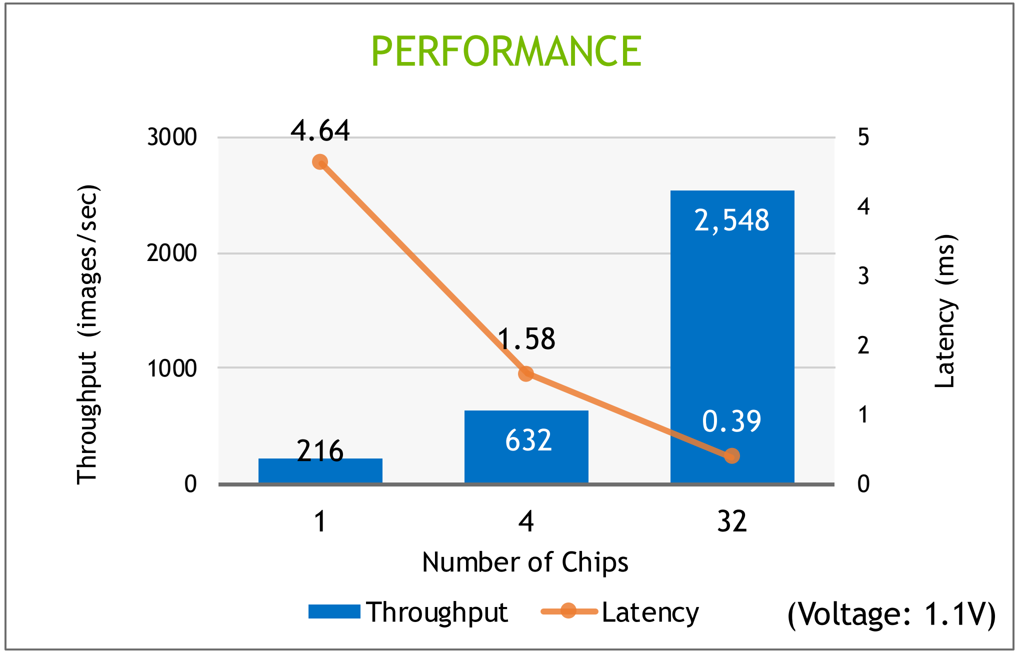
Scaling to 32 chips achieves 27X improvement in performance over 1 chip.

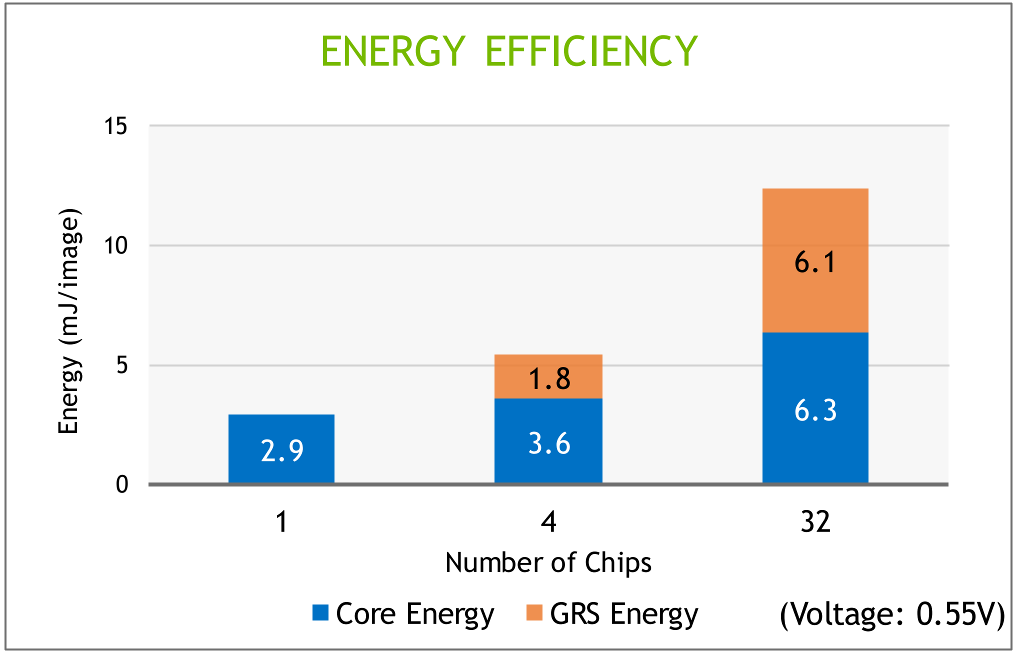
Energy proportionality in core energy consumption with weak scaling.

GRS energy can be reduced with sleep mode.

Ref: Bojarski et al., CoRR 2016

RESULTS: STRONG SCALING WITH RESNET-50





Scaling to 32 chips achieves 12X improvement in performance over 1 chip at Batch = 1.

Communication and synchronization overheads limit speed-up.

High energy efficiency at different scales.

Small overhead from communication as we scale number of chips.

Ref: He et al., CoRR 2015

SUMMARY

NVResearch Test Chip

Scalable Inference Accelerator

Uses MCM to address different markets with one architecture

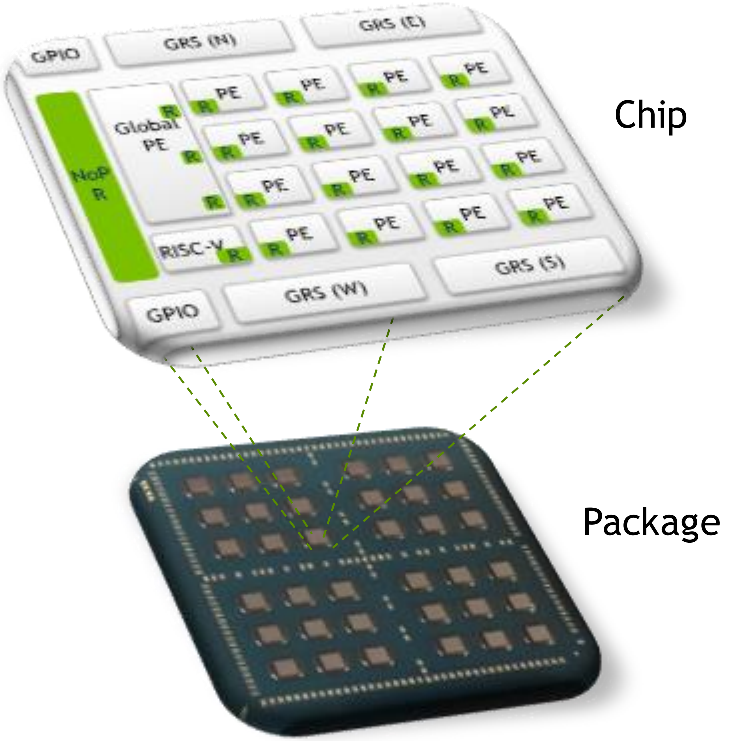
0.11 pJ/Op (8b) and 128 TOPS (111.6 core mm^2) across 36 chips connected via Ground-Referencing Signaling in a single package

Achieved 2.5K images/sec with 0.4 ms latency on ResNet-50 batch = 1

High Productivity Design Methodology

Enables faster time-to-market and more features to each SoC

10X reduction in ASIC design and verification efforts



ACKNOWLEDGMENTS

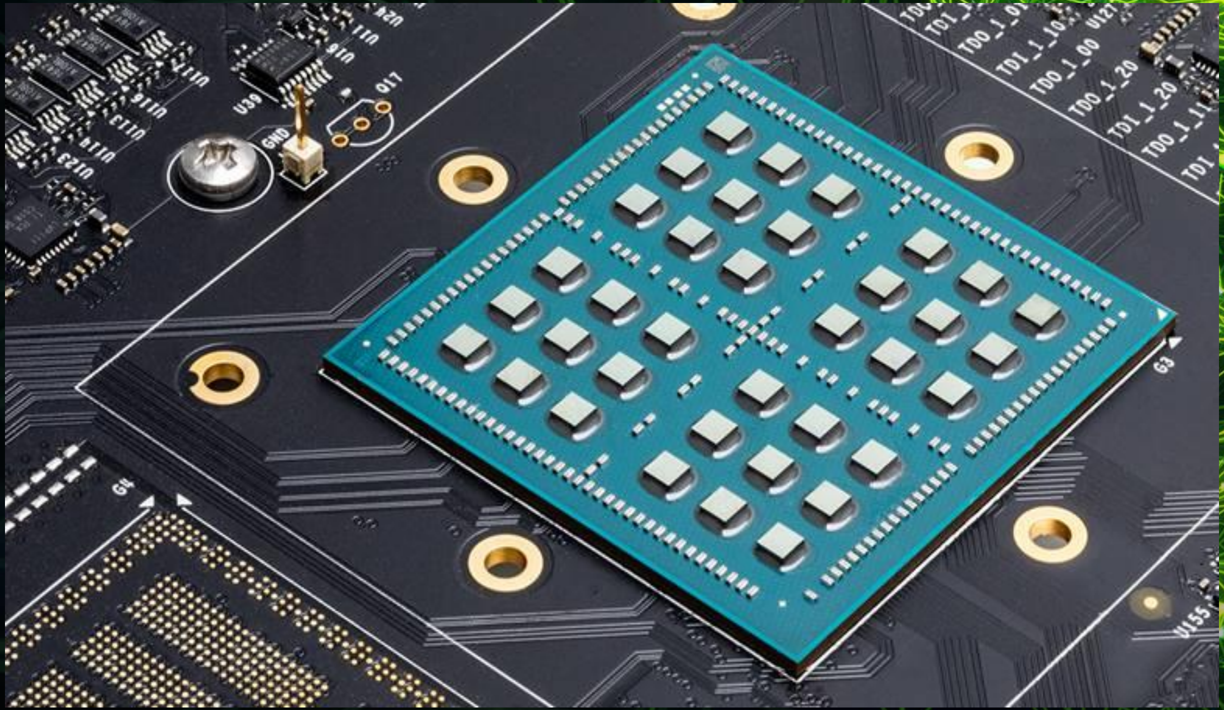
Collaborative Effort Across Architecture and Design Methodology

Research sponsored by DARPA under the CRAFT program (PM: Linton Salmon).

NVIDIA Collaborators: Frans Sijstermans, Dan Smith, Don Templeton, Guy Peled, Jim Dobbins, Ben Boudaoud, Randall Laperriere, Borhan Moghadam, Sunil Sudhakaran, Zuhair Bokharey, Sankara Rajapandian, James Chen, John Hu, Vighnesh Iyer, Angshuman Parashar for architecture, package, PCB, signal integrity, fabrication, and emulation support.

Catapult HLS team from Mentor, A Siemens Business: Bryan Bowyer, Stuart Clubb, Moises Garcia, and Khalid Islam for discussions and support.

This research was, in part, funded by the U.S. Government, under the DARPA CRAFT program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).



SCALABLE MULTI-CHIP-MODULE-BASED DEEP LEARNING INFERENCE ACCELERATOR