



# 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

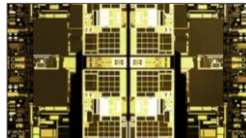
Research Teams:

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

Recent Works:



RTX



NVSwitch



Noise-to-Noise Denoising



CuDNN

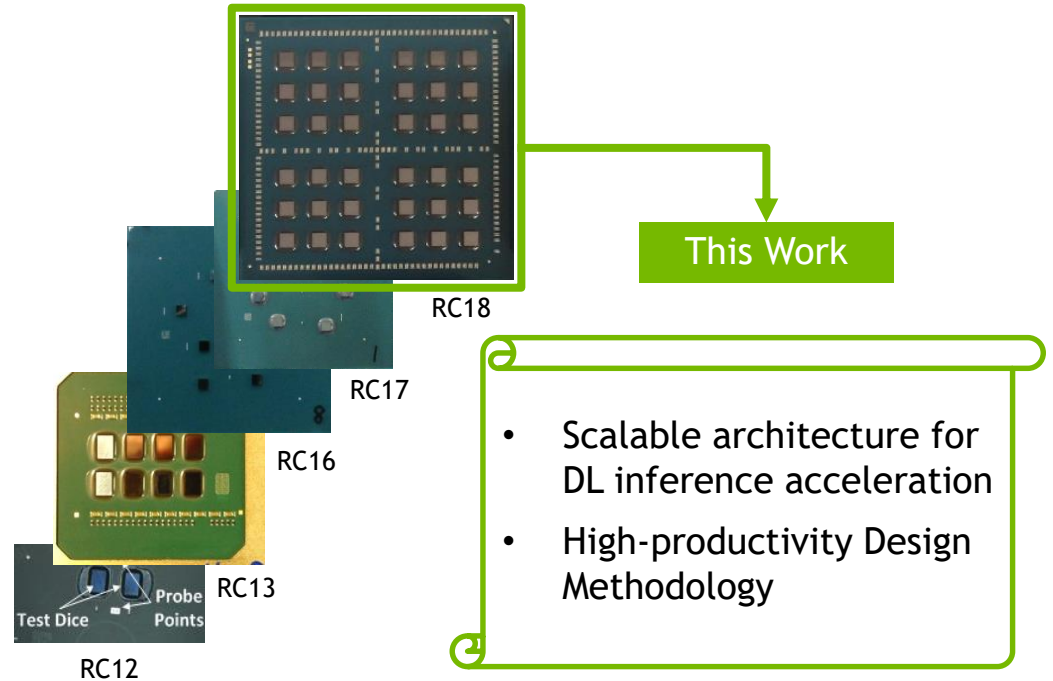


CNN Image Inpainting



Progressive GAN

## TESTCHIPS



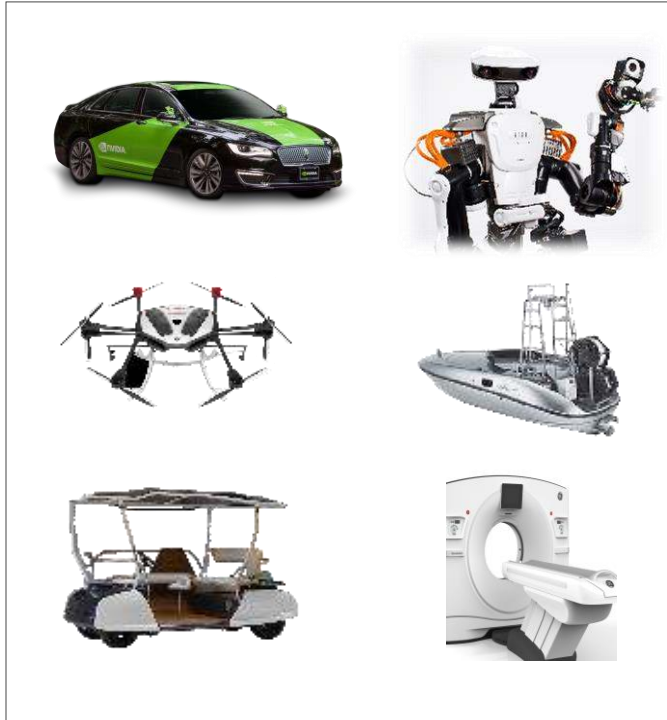


# VAST WORLD OF AI INFERENCE

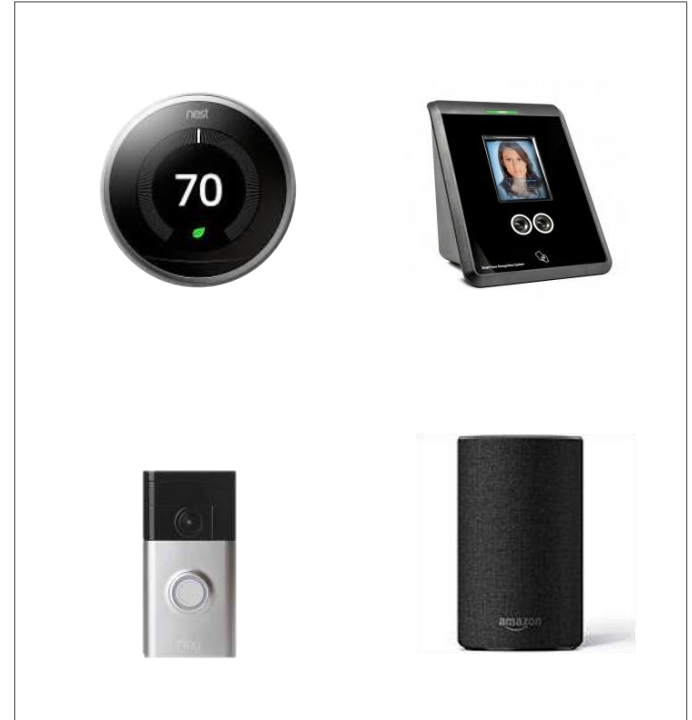
## Creating A Massive Market Opportunity



GENERAL PURPOSE COMPUTERS



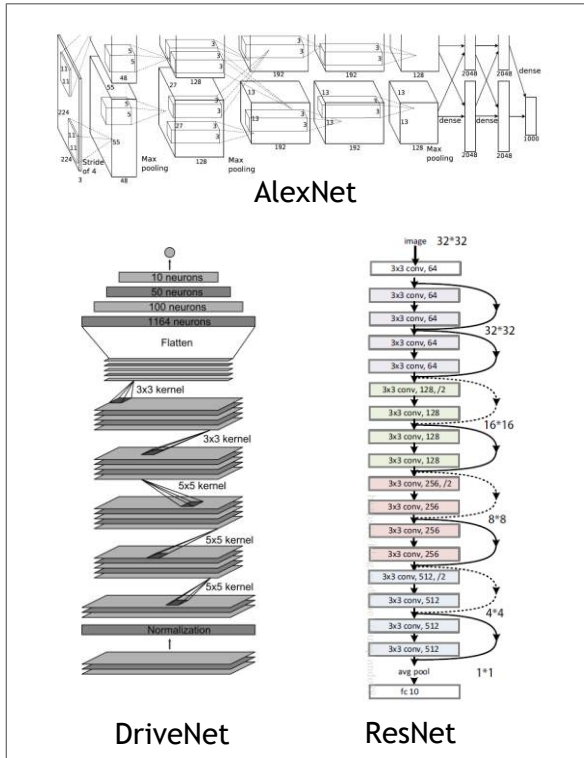
EMBEDDED COMPUTERS



EMBEDDED DEVICES

# TARGET APPLICATIONS

## Image Classification with Convolutional Neural Networks



Deep Learning Models

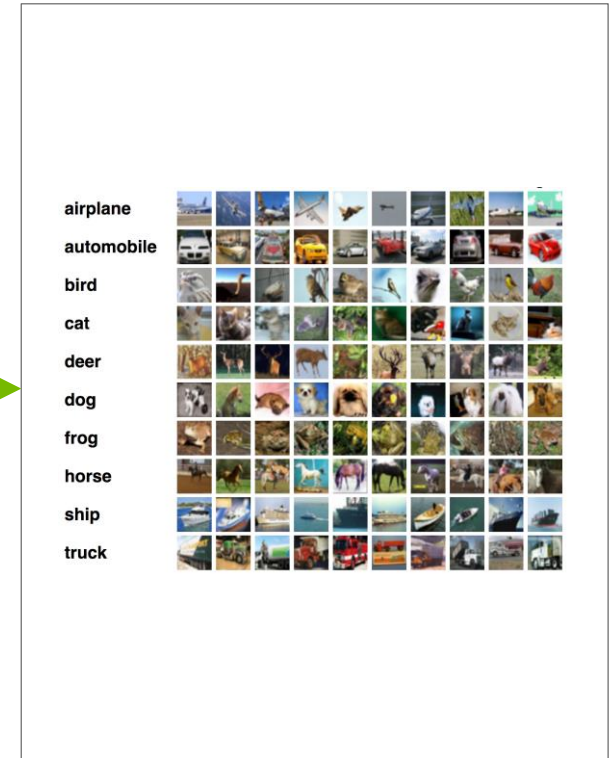
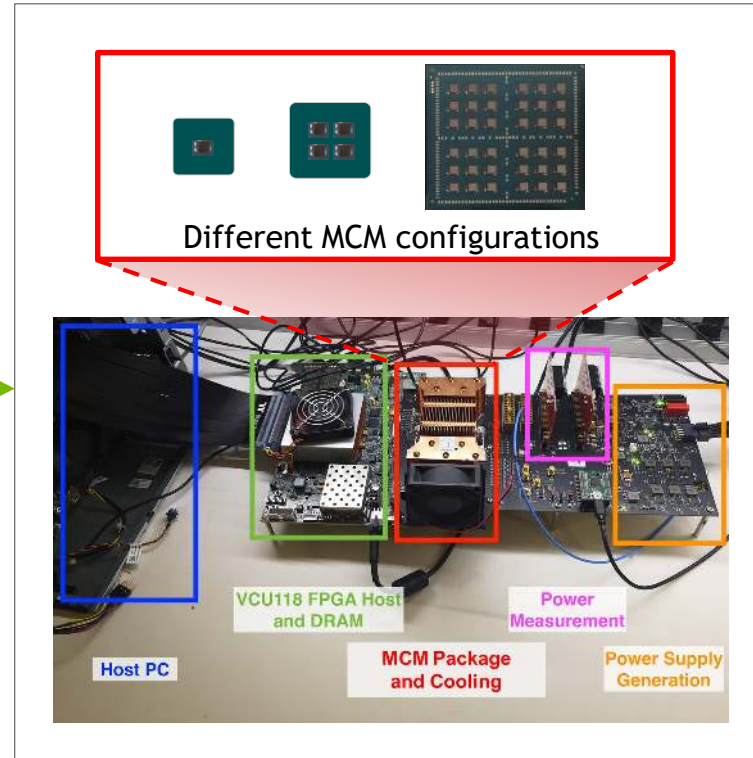
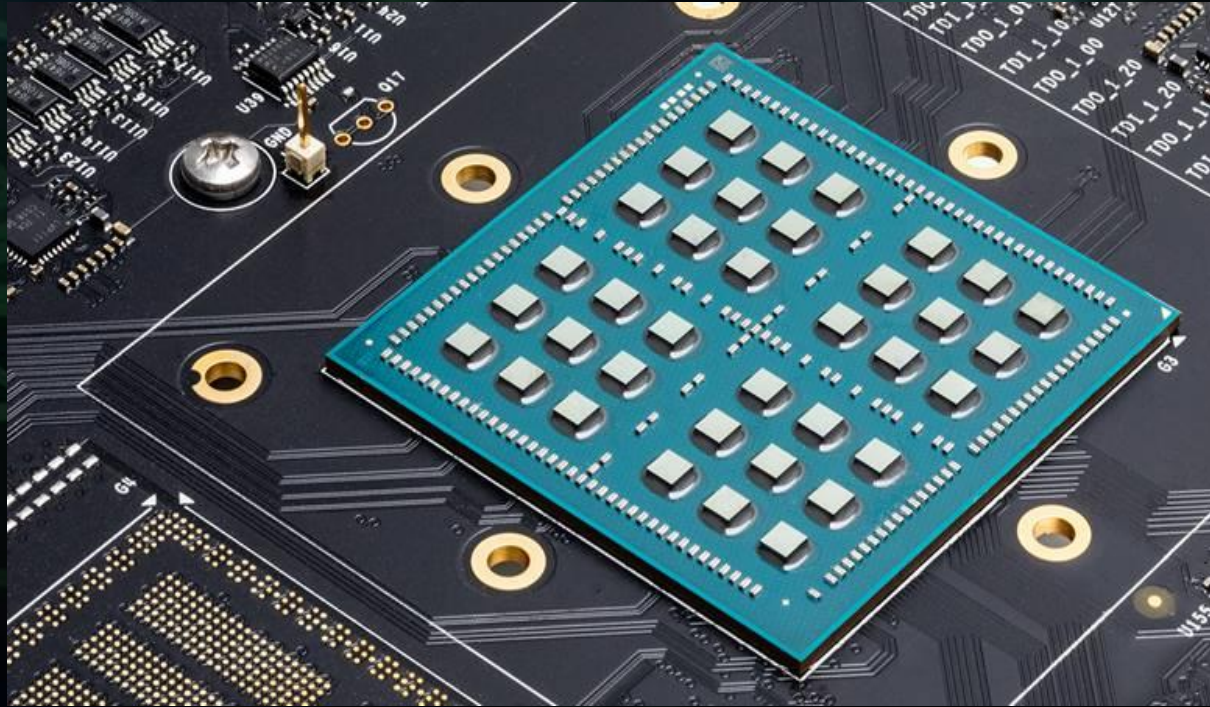


Image Classification



# SCALABLE DEEP LEARNING INFERENCE ACCELERATOR



# MULTI-CHIP-MODULE (MCM) ARCHITECTURE

## Demonstrate:

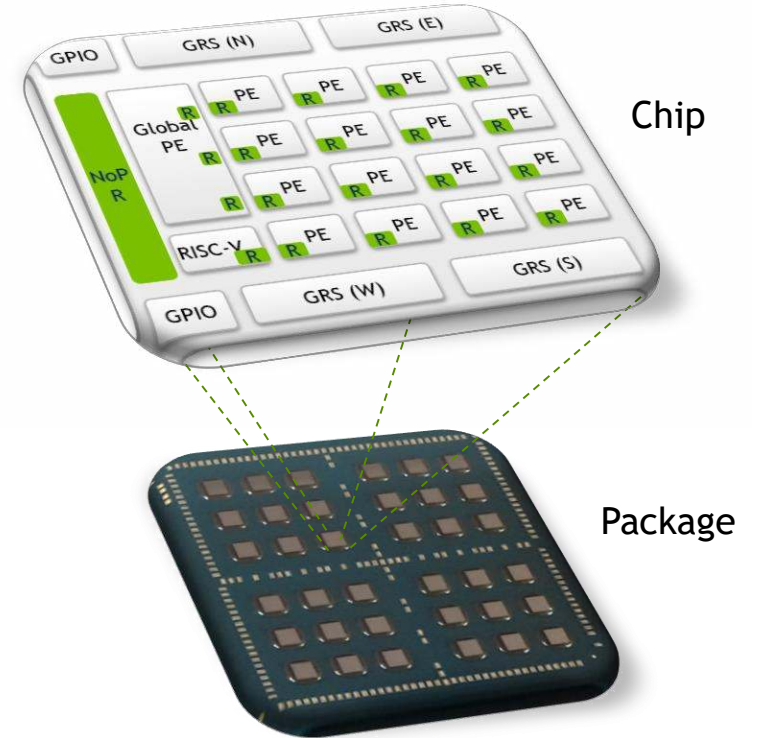
- Low-effort scaling to high inference performance
- Ground Reference Signaling (GRS) as an MCM interconnect
- Network-on-Package architecture

## Advantages:

- Overcome reticle limits
- Higher yield
- Lower design cost
- Mix process technologies
- Agility in changing product SKUs

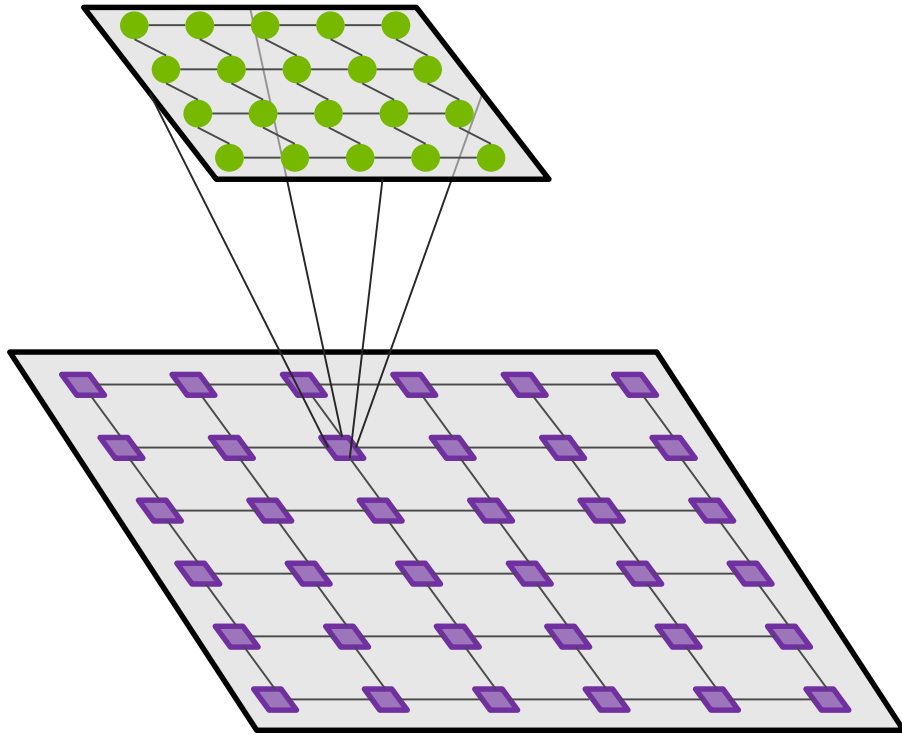
## Challenges:

- Area and power overheads for inter-chip interfaces



# 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

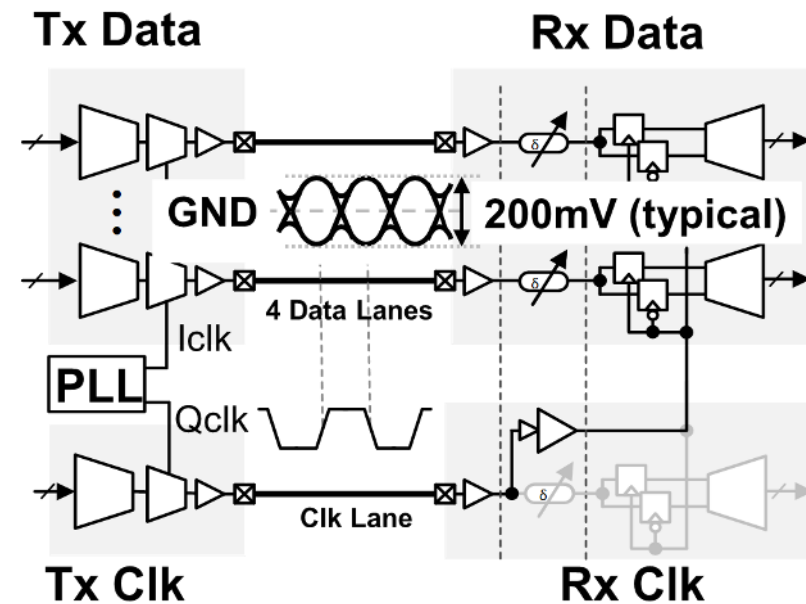
Low voltage swing (~200mV)

0.82-1.75 pJ/bit

## High area efficiency

Single-ended links

4 data bumps + 1 clock bump per GRS link

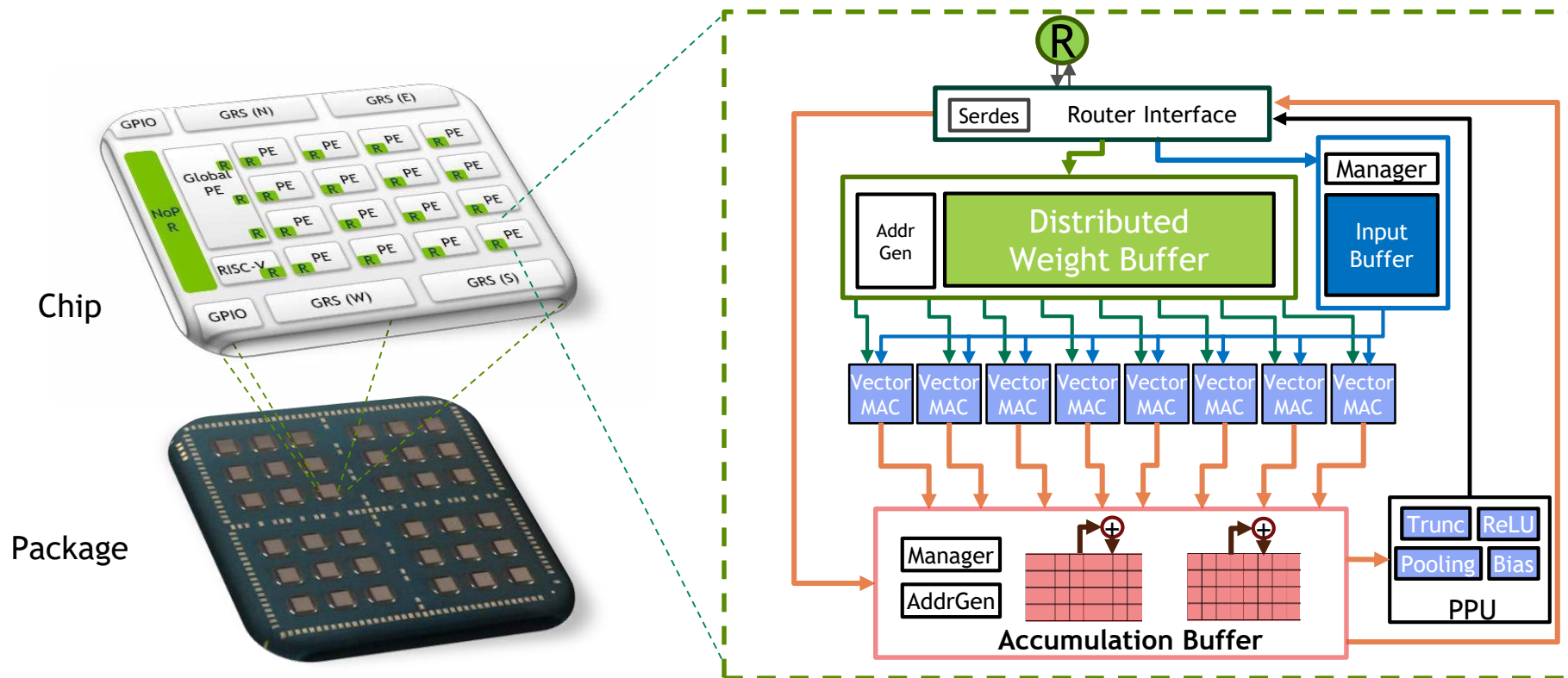


GRS Macro



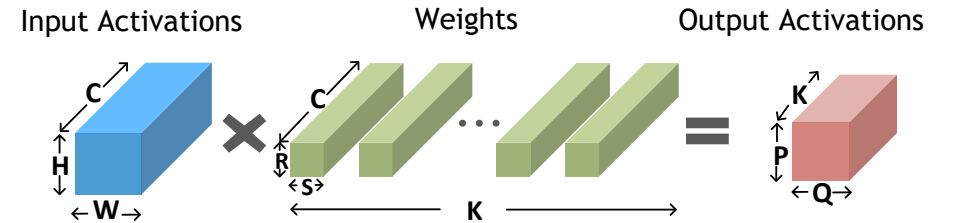
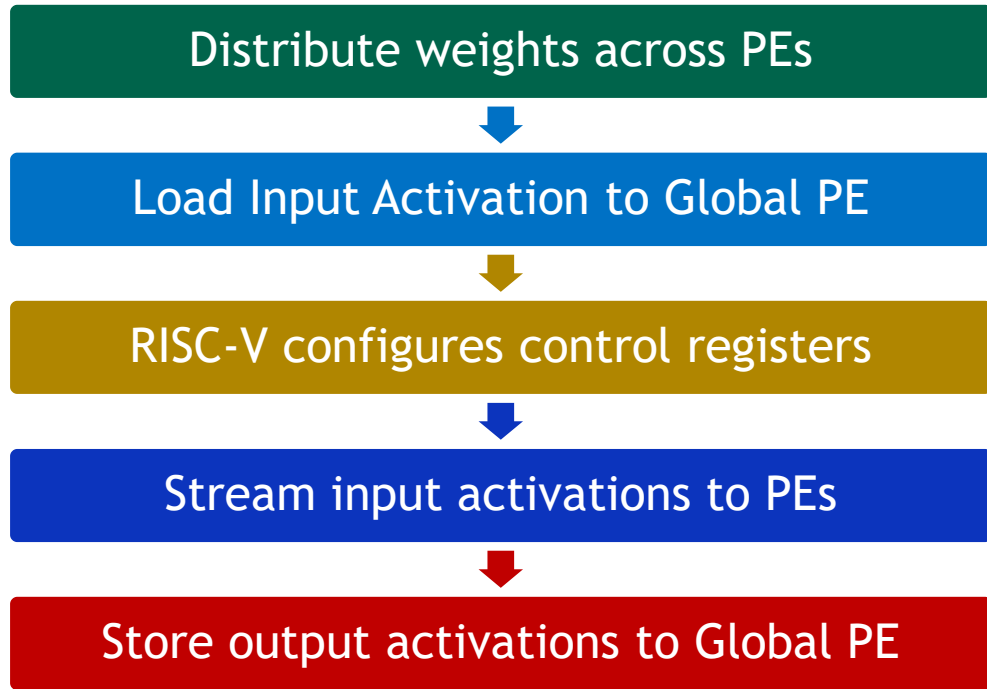
# SCALABLE DL INFERENCE ACCELERATOR

## Tiled Architecture with Distributed Memory



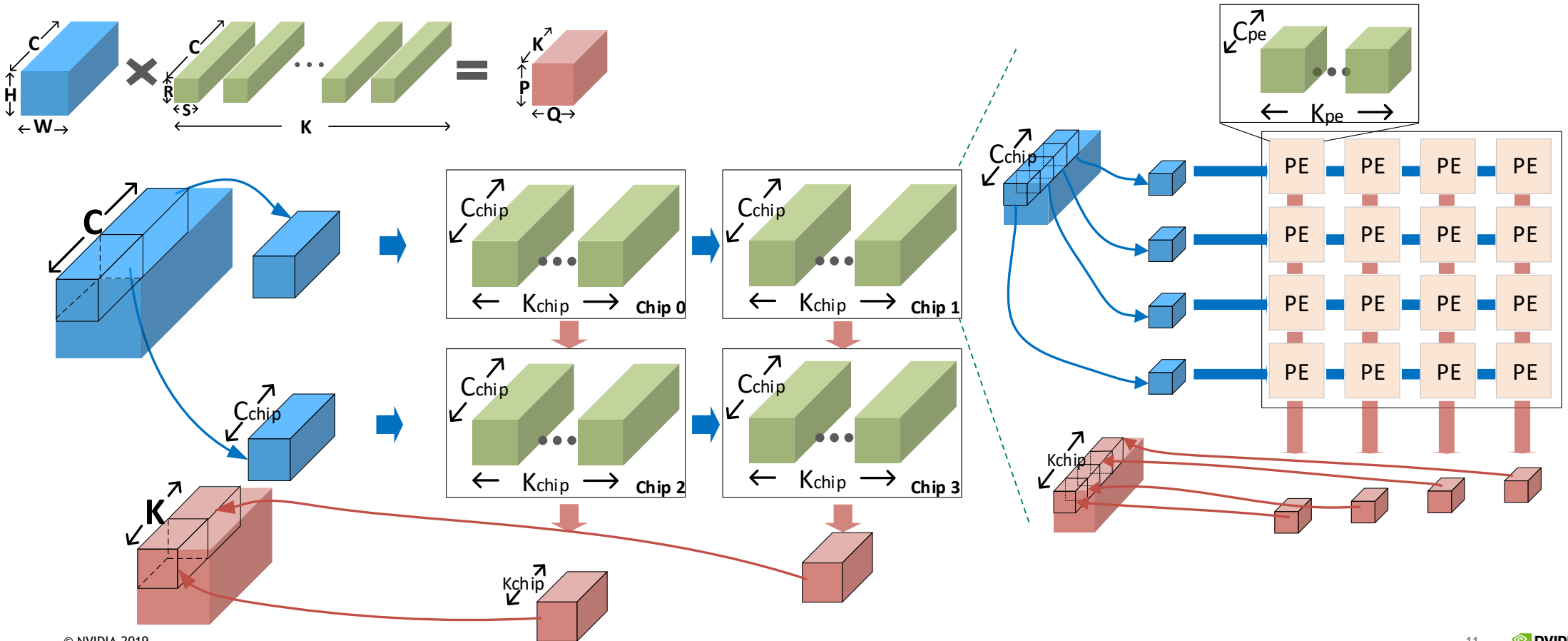
# SCALABLE DL INFERENCE ACCELERATOR

## CNN Layer Execution



# 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)

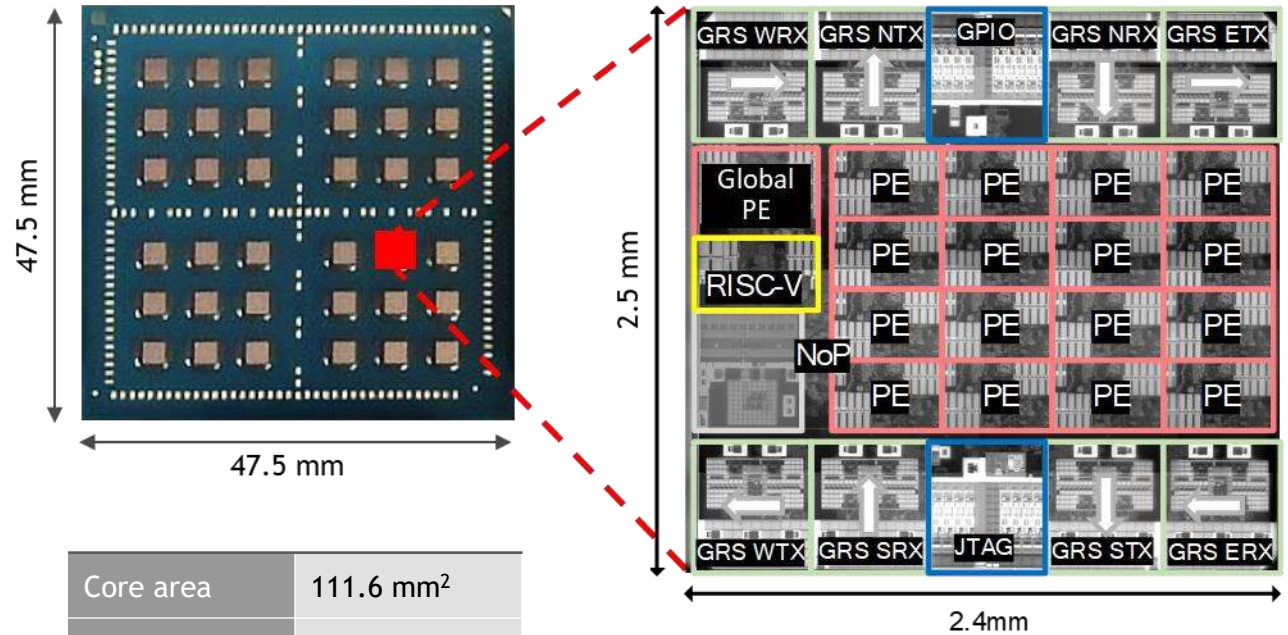
100 Gbps per link

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 <sup>2</sup>
Voltage	0.52-1.1 V
Frequency	0.48-1.8 GHz

The background features a complex network of thin, glowing green lines connecting various nodes. Some nodes are bright green, while others are a soft blue. The overall effect is a sense of interconnectedness and digital flow against a dark, almost black background.

# **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**

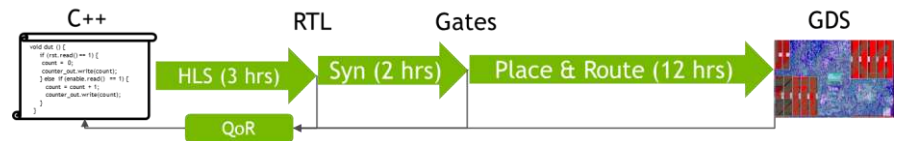
## AGILE VLSI DESIGN

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





# 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

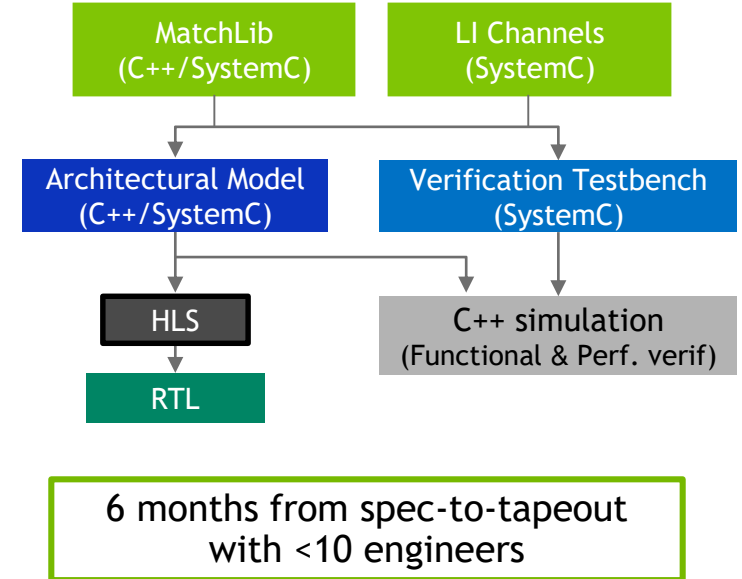
Highly-parameterized, high QoR implementation

Available open-source: <https://github.com/NVlabs/matchlib>

**Latency-Insensitive (LI) Channels**

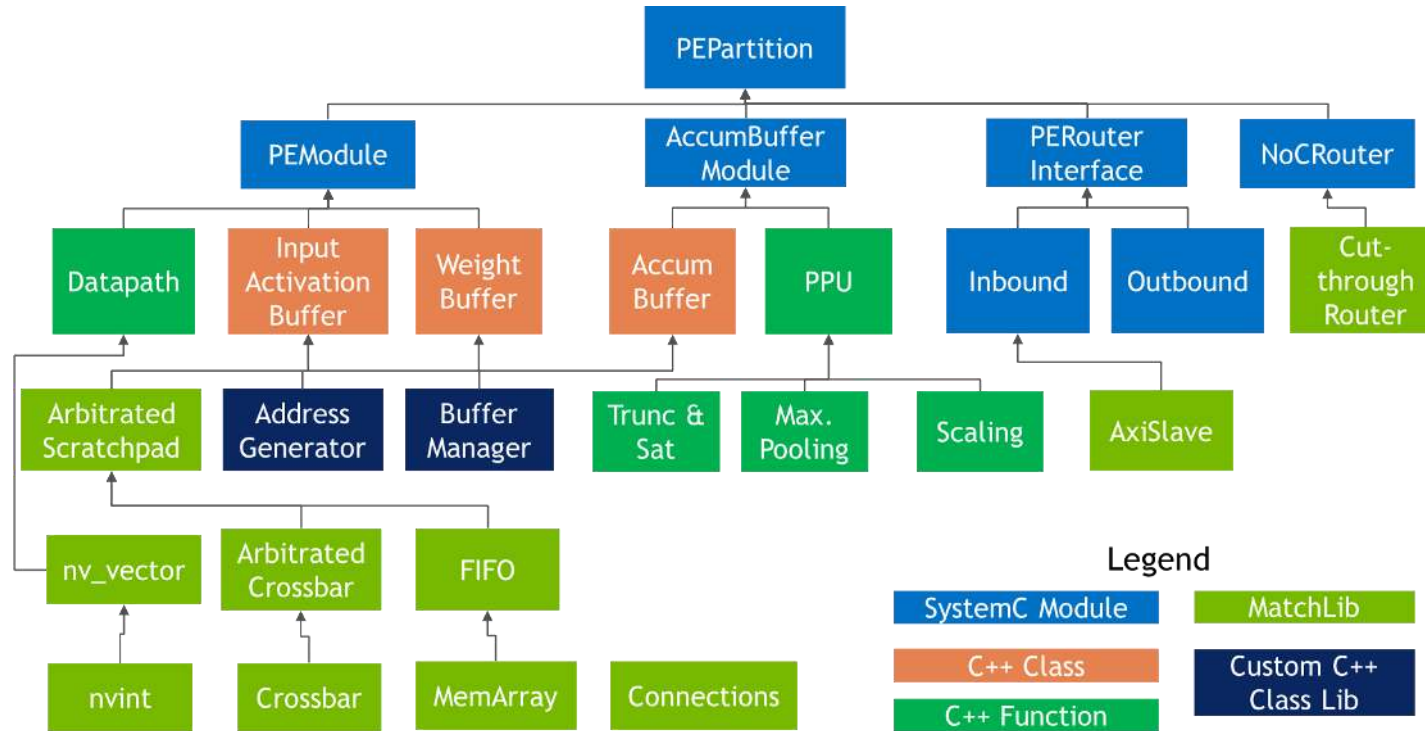
Enable modularity in design process

Decouple computation & communication architectures



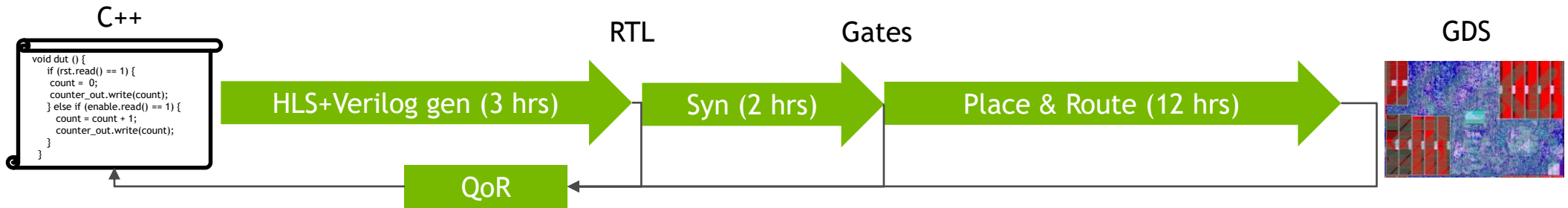
# PROCESSING ELEMENT IMPLEMENTATION

## Reuse, Modularity, Hierarchical Design



# AGILE VLSI DESIGN TECHNIQUES

## Daily “C++ to Layout” Spins



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

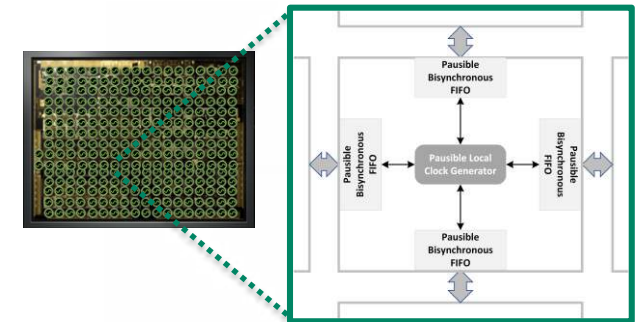
Small, abutting partitions for fast place and route iterations

Globally asynchronous locally synchronous pausable adaptive clocking scheme

Fast, error-free clock domain crossings

“Correct by construction” top-level timing closure

RTL bugs, performance, and VLSI constraints converge together





An abstract network diagram with a dark background. It features several glowing green nodes of varying sizes, connected by thin, light green lines. The nodes are scattered across the frame, with a higher density on the left side. The lines create a complex web of connections between the nodes. The overall aesthetic is futuristic and technological.

# EXPERIMENTAL RESULTS

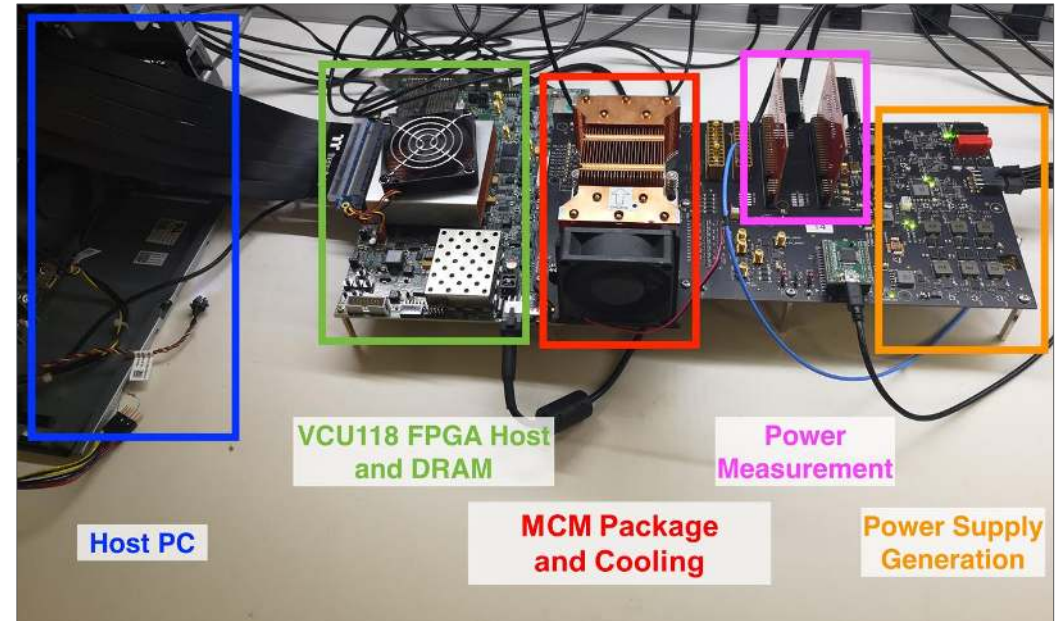
# MEASUREMENT SETUP

Measurements begin after weights and activations are loaded from FPGA DRAM

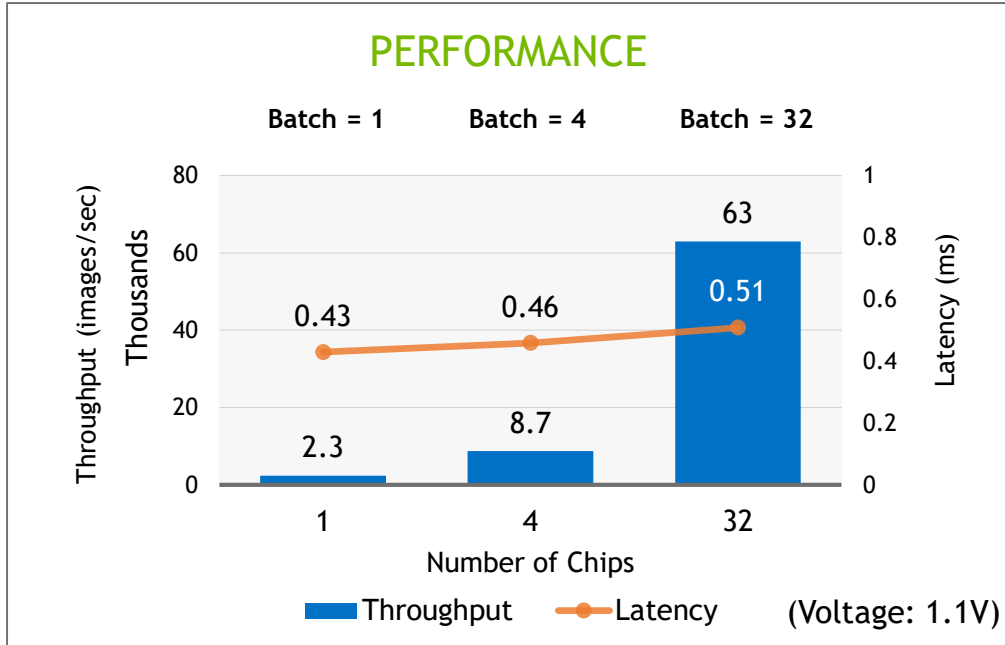
Weights are loaded to PE memory  
Activations are loaded to Global PE

## Operating points

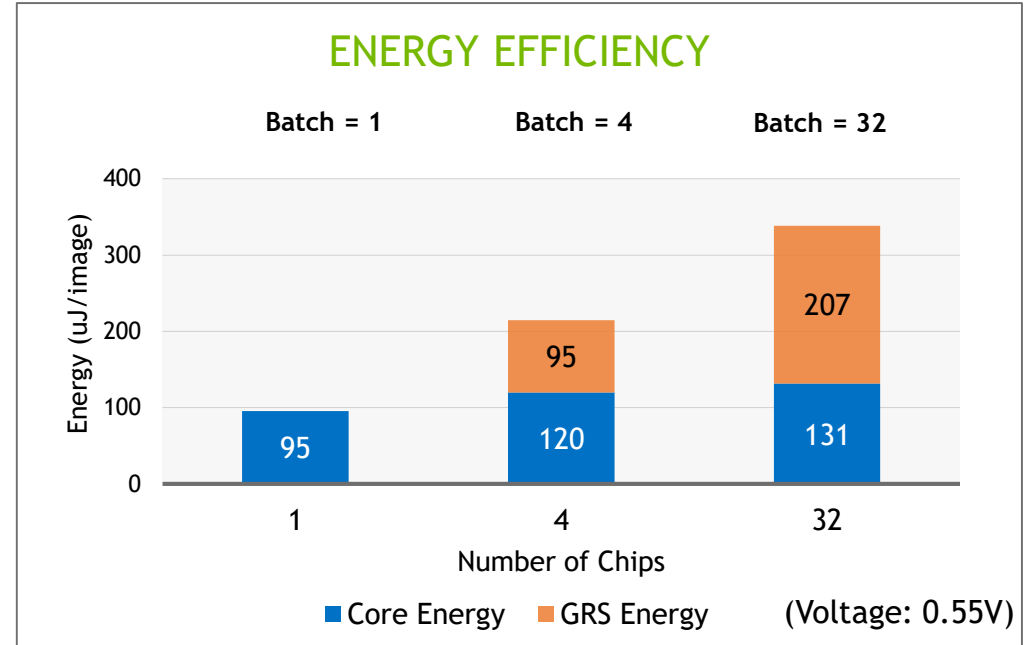
Max. Performance: 1.1V  
Min. Energy: 0.55V



# RESULTS: WEAK SCALING WITH DRIVENET

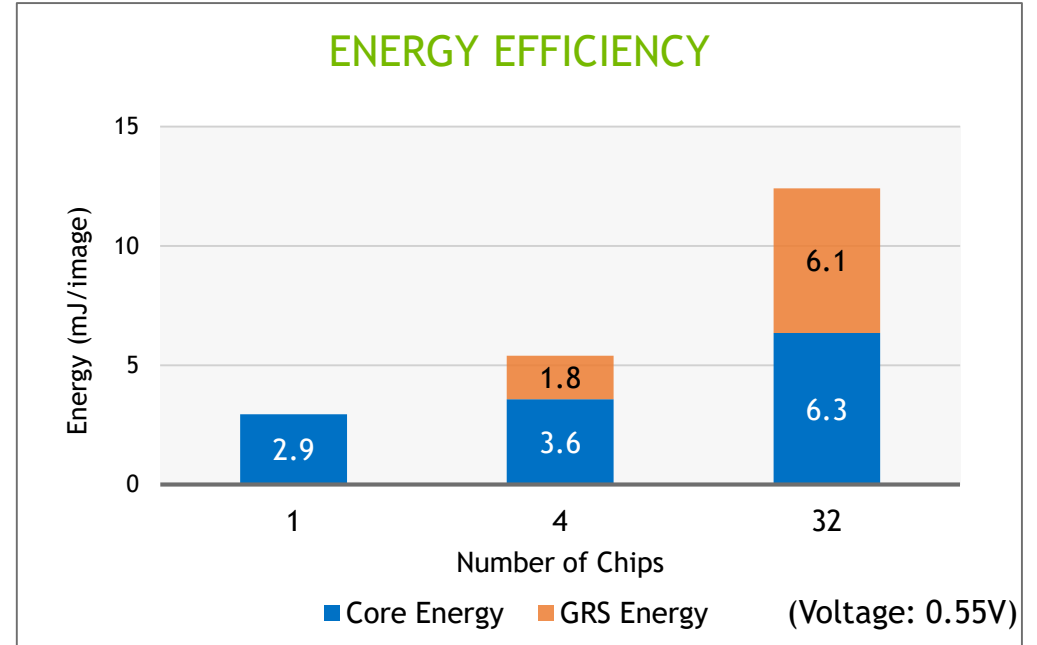
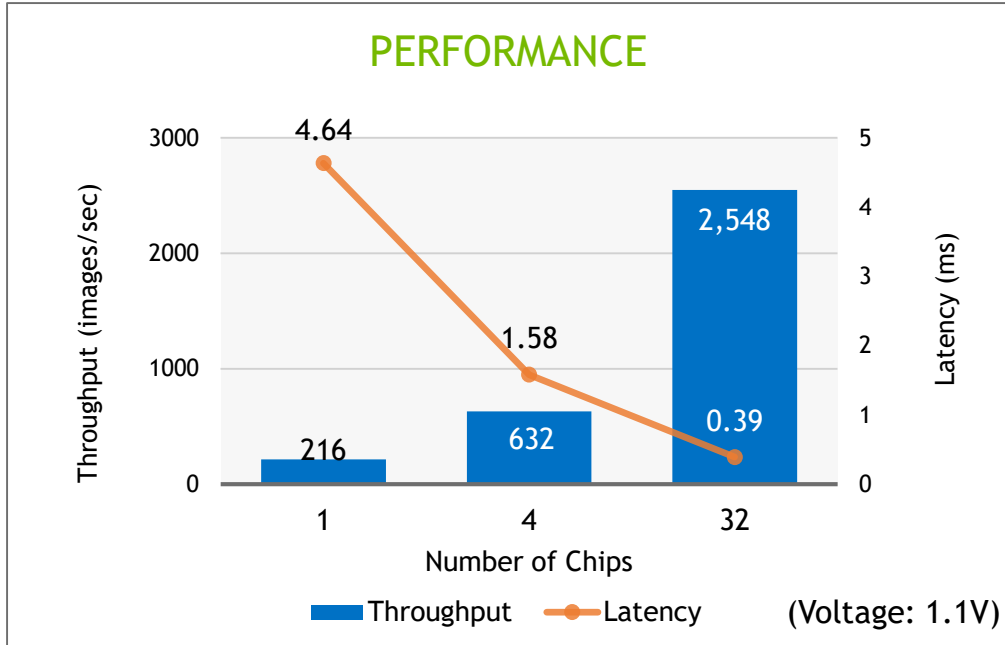


Scaling to 32 chips achieves 27X improvement in performance over 1 chip.



Energy proportionality in core energy consumption with weak scaling.  
GRS energy can be reduced with sleep mode.

# 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.

# SUMMARY

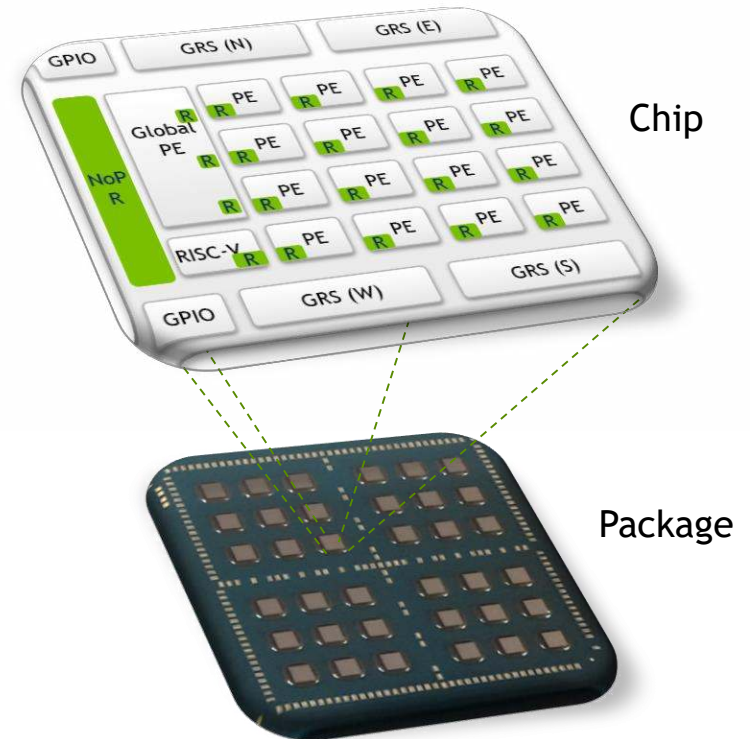
## NVResearch Test Chip

### Scalable Inference Accelerator

Uses MCM to address different markets with one architecture  
0.11 pJ/Op (8b) and 128 TOPS (111.6 core mm<sup>2</sup>) 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





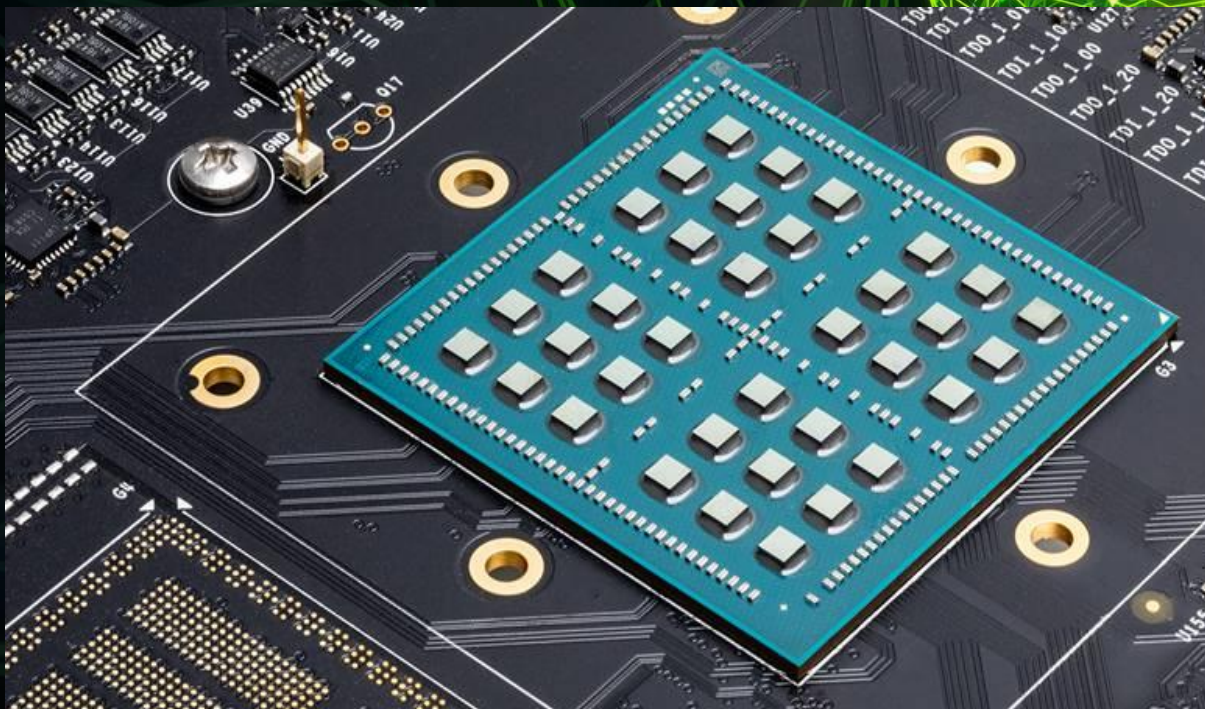
# ACKNOWLEDGMENTS

## 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.



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



**nVIDIA®**