HAILO
Empowering Intelligence

A Computer Architecture Renaissance

Energy-efficient Deep Learning Processors for Machine Vision

In partnership with Embedded Vision Alliance

Uri Weiser, Emeritus Professor, Electrical Engineering Dep. at Technion IIT

Avi Baum, CTO and Co-founder of Hailo

December 2019

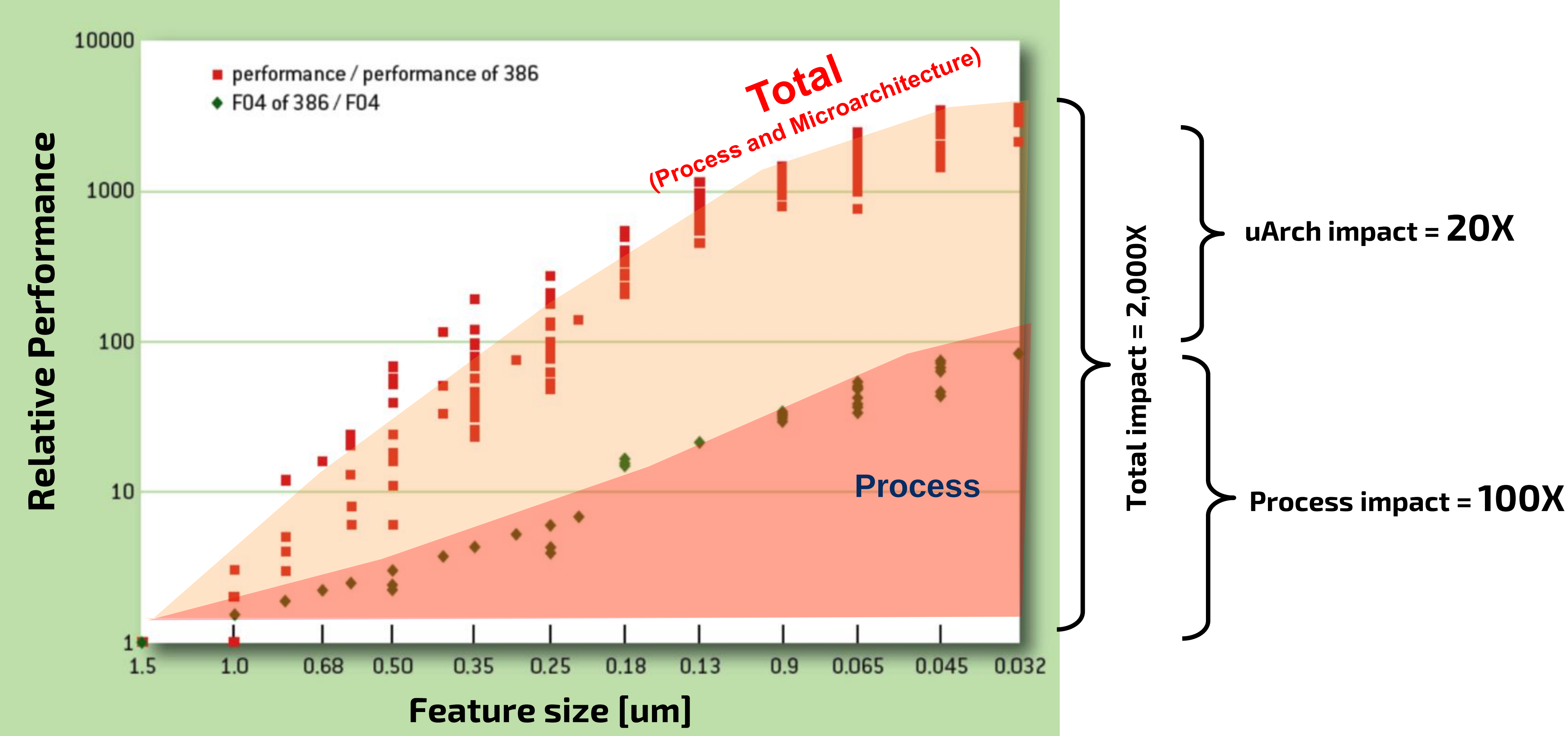
A Computer Architecture Renaissance - The Inevitable Path

- Technology environment
  - Riding the process wave
  - Implications (ISA, performance only…)
  - Hitting the wall (Process and microarchitecture)
  - Solution? ➔Accelerators
- To our rescue?
  - Killer applications = Machine Learning
  - 80/20
  - Implications – perf ⇔ power ⇔ efficiency ⇔ locality
    - Locality = Pipeline
- Implications

Performance ➔ power ➔ energy ➔ locality (Quote: Bill Dally)

Technology Environment

Performance History

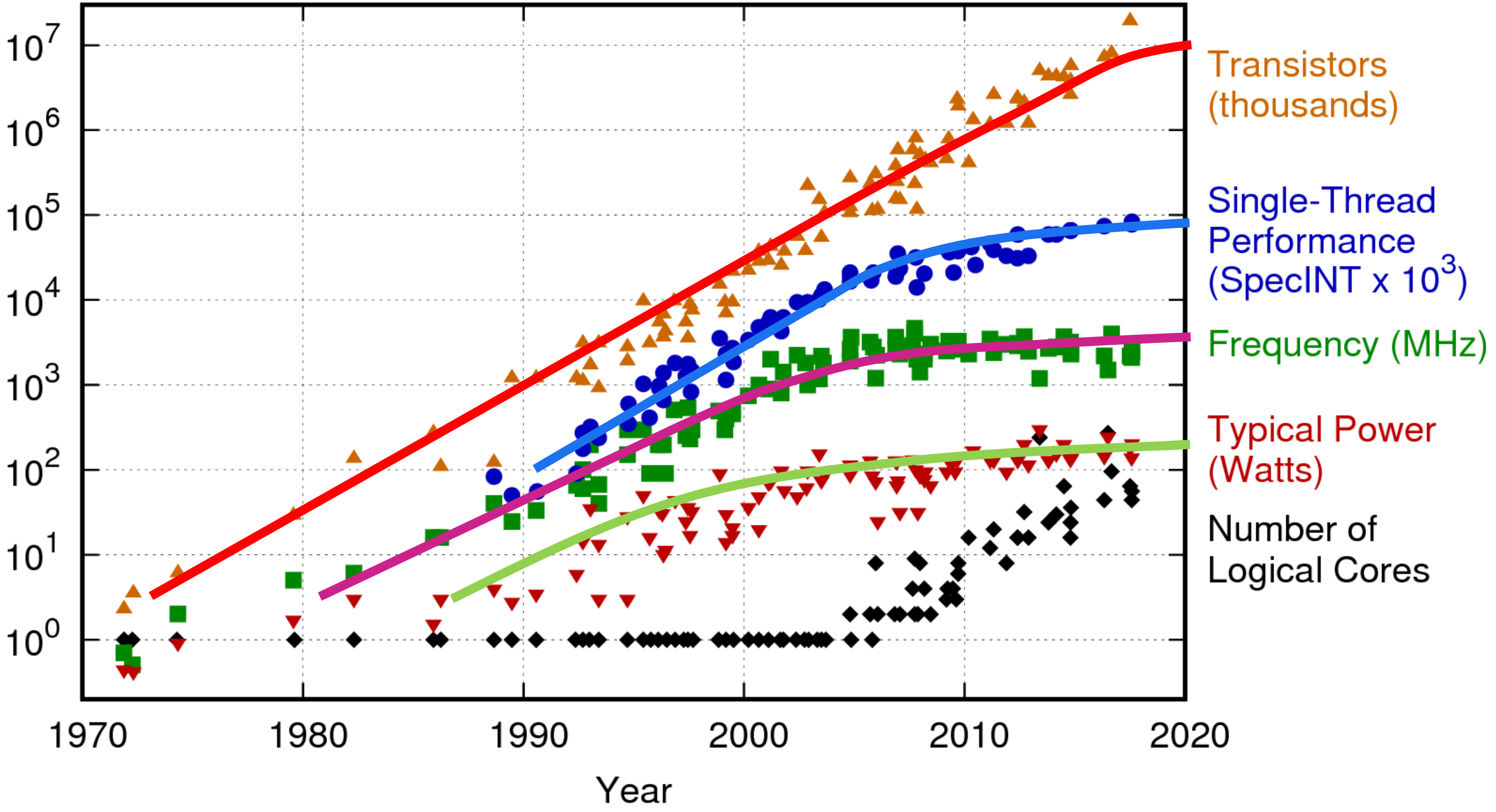


We (the architects) did an "OK-" job

*ACMqueue April 6, 2012, Processors, Volume 10, issue 4 CPU DB: Recording Microprocessor History, Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

HAILO

And.. Moore's Law is Ending



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

..and - End of Dennard Scaling

Basics: P α nCV^2f

- Moving chip design from one process generation to the next
- Keeping the die size constant

What will be the power ratio of the new device (new process)?

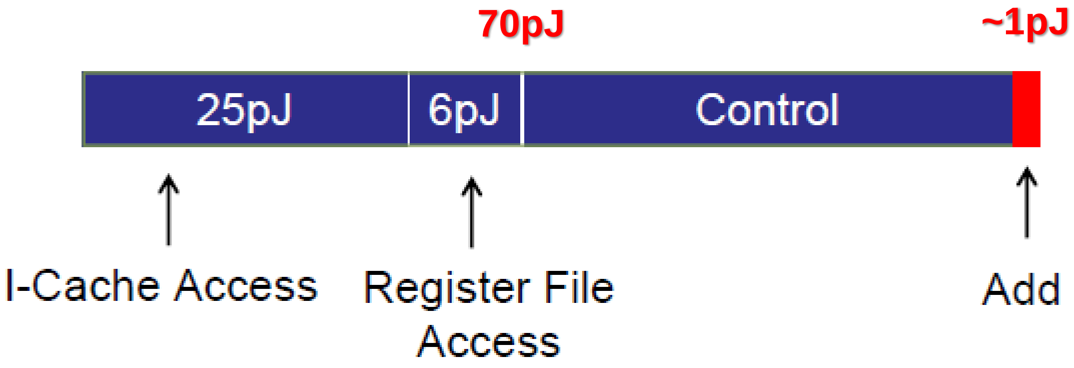
Component	Dicription	Past scaling ratio	Current scaling ratio
n	Number of transistors	2	2
C	Capacitance/Transistor	0.7	0.7
V	Voltage	0.7	0,95
f	Frequency of operation	1,4	1,1
POWER		~1	~1.4

Insight: Not all transistors can function at the same time

due to the halt in voltage scaling

The Cost of Flexibility: Instruction-based Architecture

Instruction Energy Breakdown @45nm technology

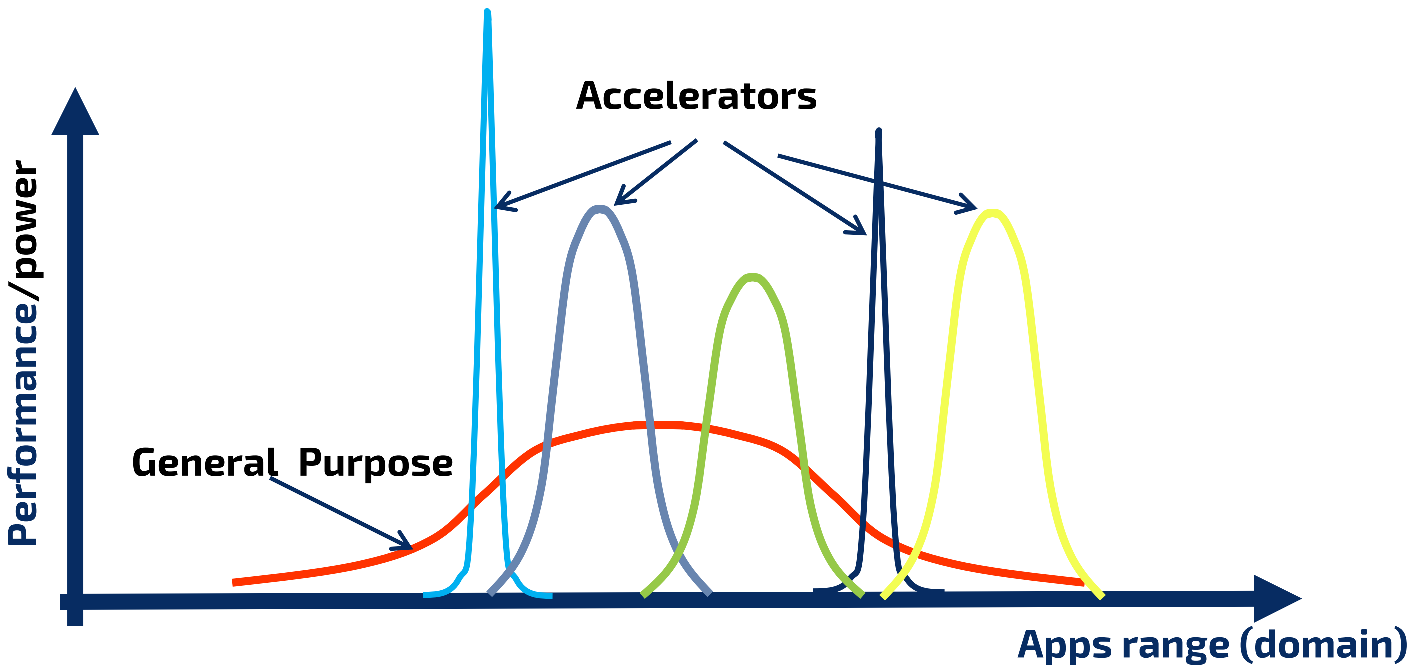


We hit the power wall due to: power limitation, Process Technology (end of Dennard scaling)

Insight: Programmability ➔ NO efficiency

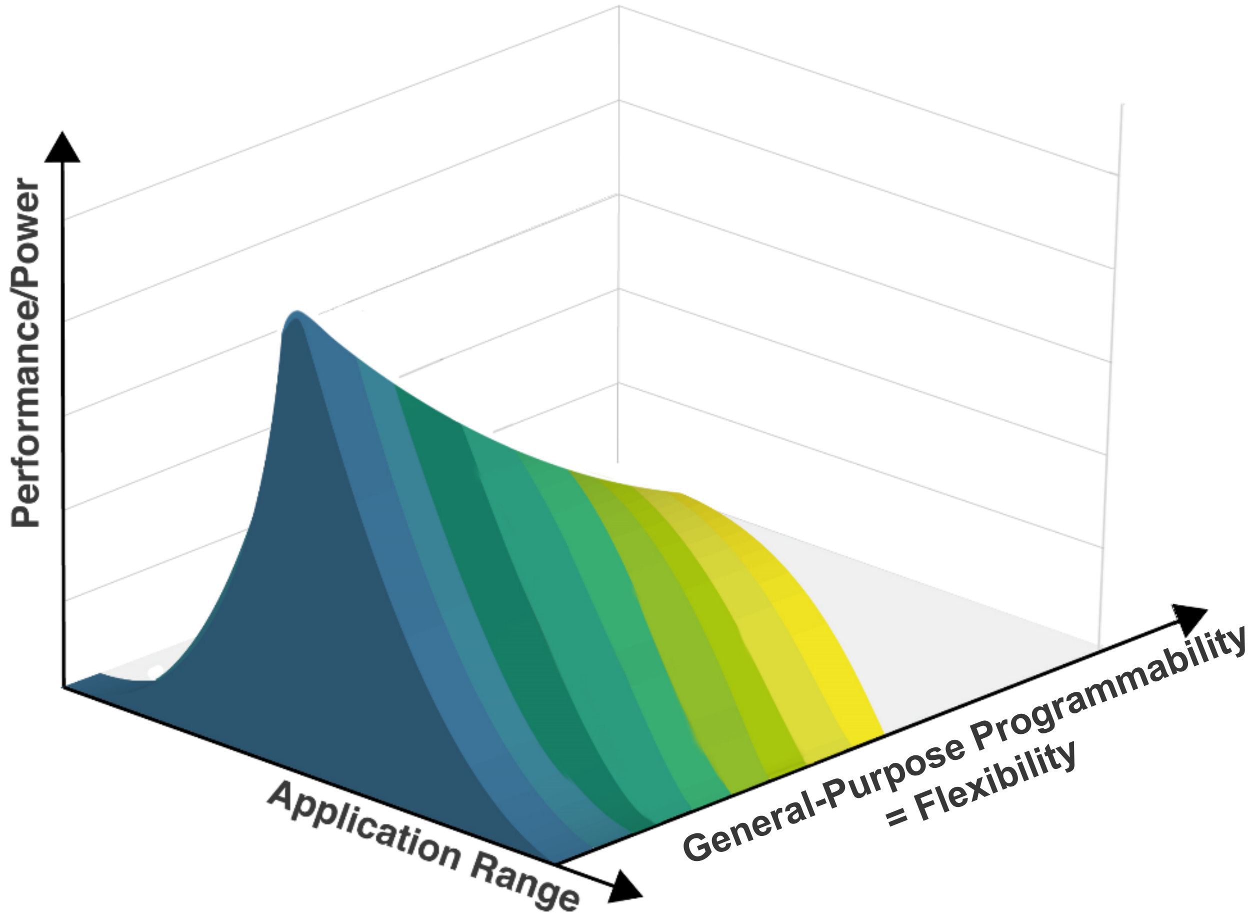
From: Mark Horowitz, Stanford "Computing's Energy Problems"

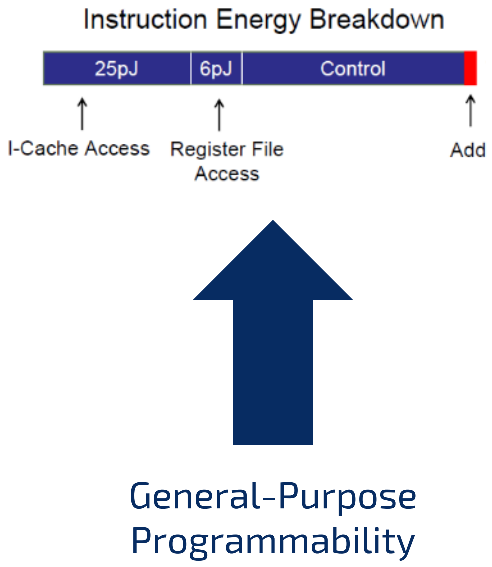
Heterogeneous/Accelerator to the rescue



Heterogeneous cores improve performance/power ratio but are designed for specific workloads ➔ trading efficiency for flexibility

Heterogeneous Computing - Flexibility Impact





Implication: Increase in flexibility = ISA based architecture → efficiency reductions

(True) Heterogeneous Usefulness

- When applications are fixed and reused there is no need for flexible/programmable instructions

- When there is a fixed MAPPING from the problem statement/application to the HW? e.g. Data-Flow ….

We have been lucky!!!

There are such applications/implementations

Killer Applications*

- There is a range of applications that map to ML

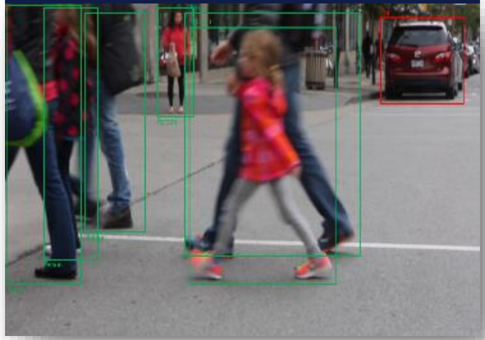
Input:	**huge amount of data**
Output:	**small amount**

- **Operations:** **many simple operations**

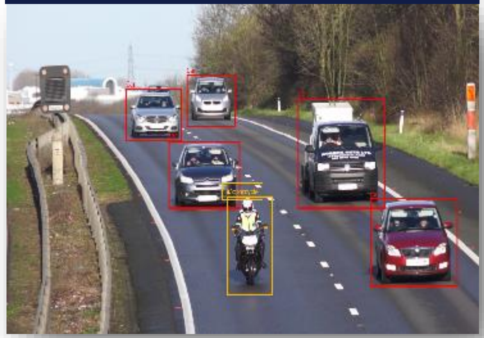
*Applications you can not effectively execute on current HW (Dr. Andy Grove)

Deep Learning is Everywhere

Pedestrian Detection



Vehicle Detection



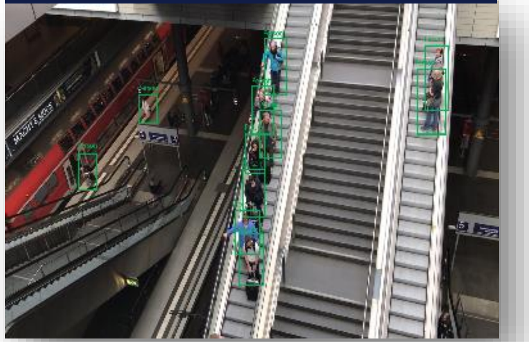
Collision Avoidance

Parking Assist

Speech Understanding

Plate / Traffic Sign Detection

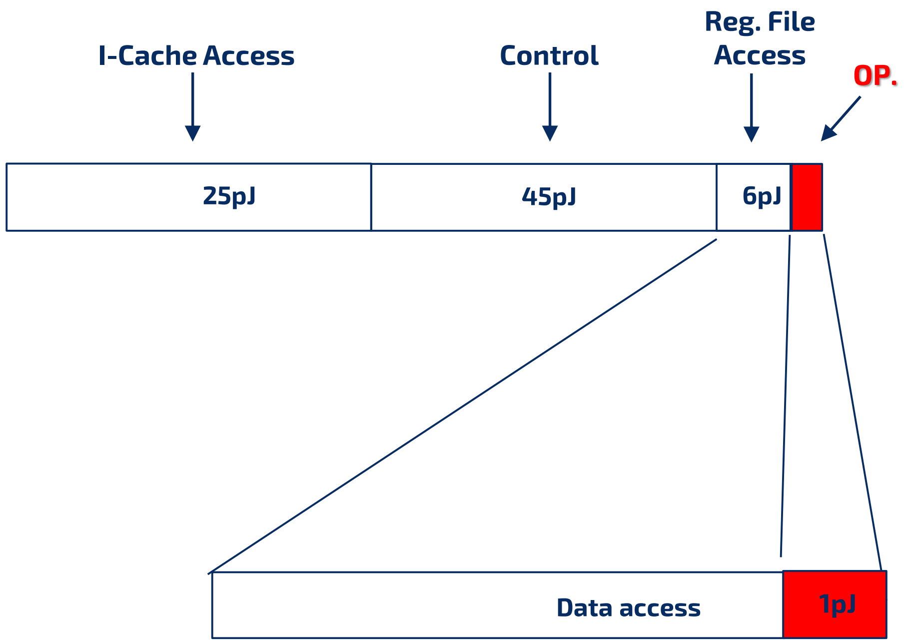
Passenger Control



Face Recognition

Efficient ML Architecture

Energy in “Data Flow” Architecture



Instruction based architecture: energy breakdown

Non-ISA based architecture: “Data flow” energy breakdown

now “Read Once“ counts!

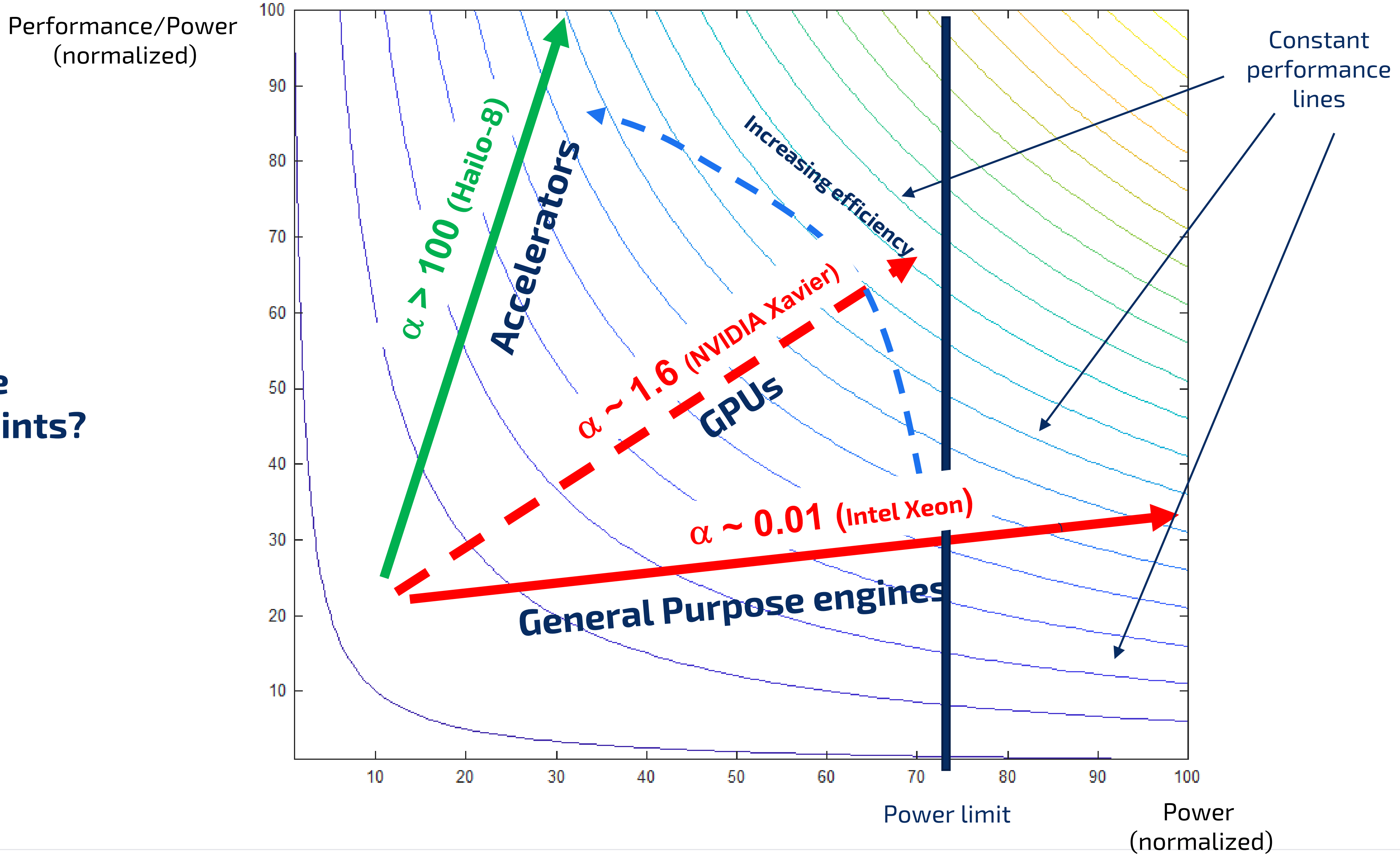
Efficient ML Balanced Design and Energy Reduction

Energy reduction

- Reduction in Computing (MACs op.)
  - Pruning
  - Prediction
  - Quantization
  - ….
- Reduction in data access and movement
  - Pipeline
  - Compression
  - …
- Efficient usage of the Hardware resources
  - Multi-Amdahl (divide effectively the resources)
  - SMT
  - ….

Performance/Power vs. Power

How to reach best performance within power constraints?



Takeaway at this point:

- Performance (silicon technology) is king
- “Moore’s law” and Dennard scaling are scaling down (ending?)
  - 70 years of rule-based architecture; serving existing apps
- ➔ Architecture becomes the computing leading force
- New domain specific Killer apps
  - 80/20 rule
  - Non-ISA accelerators
- Properties
  - Locality – to reduce external-to-the-chip memory accesses
  - Pipelining (CNN based)
  - Gradual locality
  - Data movement resolution
  - Smart compiler

HAILO

Empowering Intelligence

A Computer Architecture Renaissance

Energy-efficient Deep Learning Processors for Machine Vision

Embedded Vision Alliance Webinar

Flow

Deep Learning as a 'Domain'

Domain Convergence

Intra-Domain Diversity

Flexibility and programmability

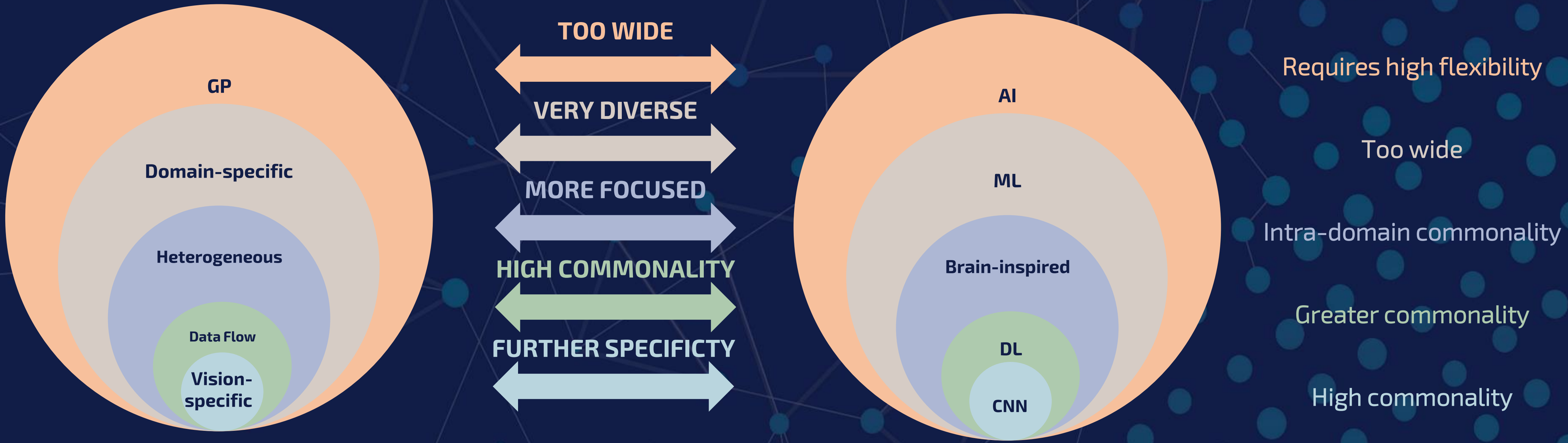
Architecture Challenges

Capacity & Efficiency

Hailo Architecture in a nutshell

Deep Learning: Domain Exploration

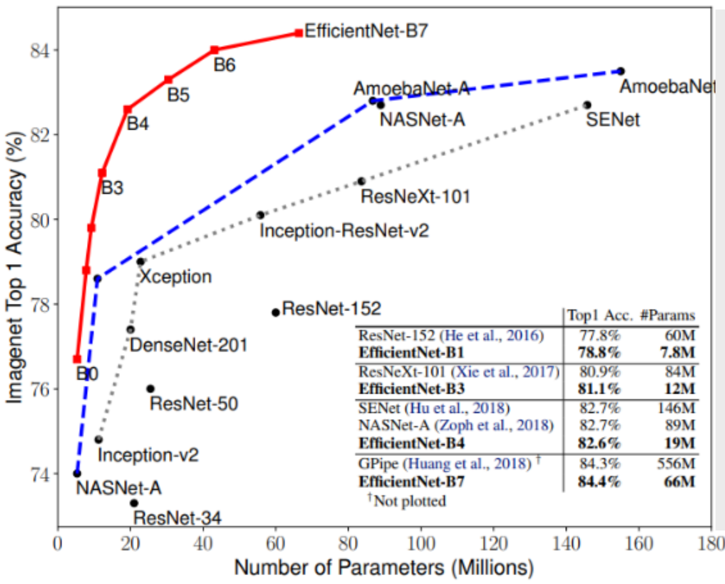
- Why DL constitutes a 'domain'?
- Since 2012 AlexNet triggered up the domain-specific thinking, by embedding **domain awareness** that was introduced to generic neural networks, accounting for spatial locality properties in images



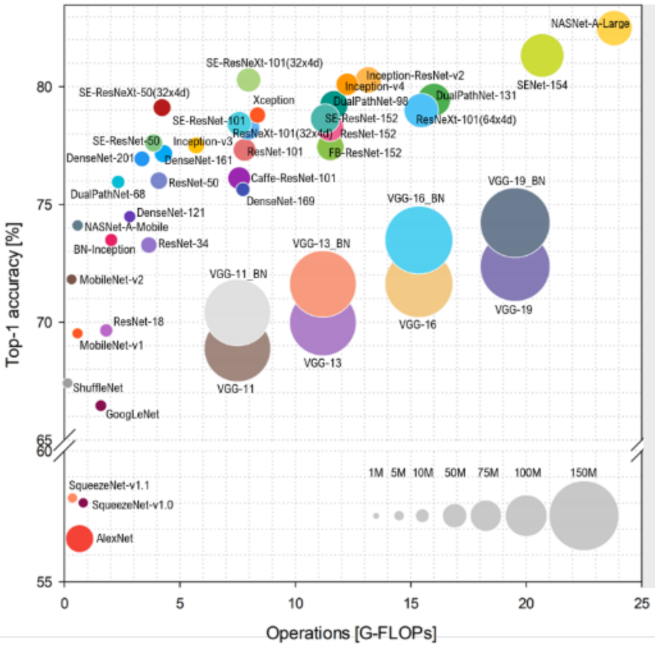
Source: Comprehensive Survey on Deep Learning Approaches (1803.01164)4.pdf

Deep Learning for Vision .. Domain Convergence

- Nearly a decade of intense research activity in this domain
- Yielding major achievements in the performance vs. accuracy trade-off
- Recent survey activity (e.g. Google's NAS) clearly indicate this model convergence trend (in terms of model size; no. of operations and accuracy)

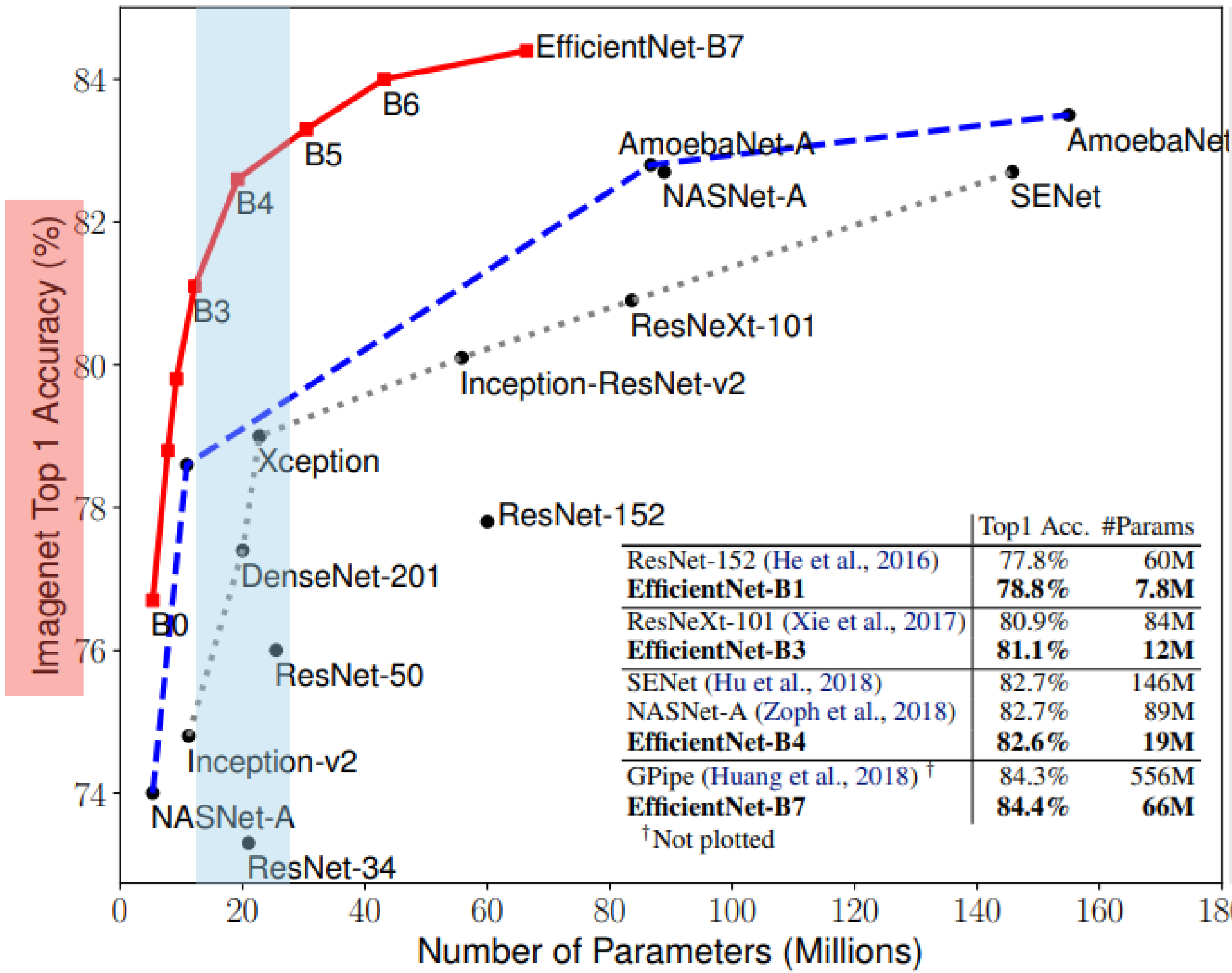


Google EfficientNet



Benchmark Analysis of Representative DNNs

Accuracy vs. Model Size



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**78.8%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.1%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.6%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

† Not plotted

Source: Google EfficientNet (1905.11946)

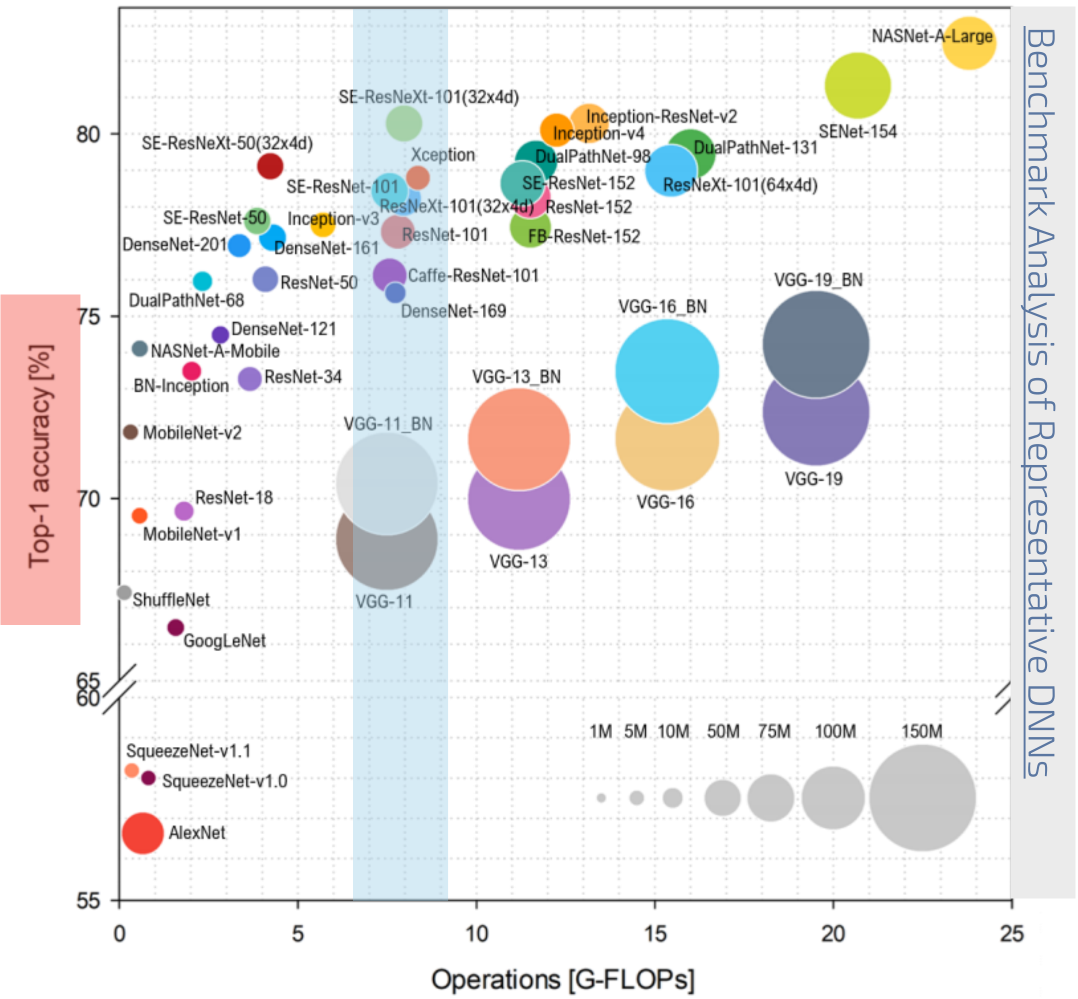
~20 M

Model-size knee-point

~86%

Asymptotic Accuracy

Accuracy vs. Compute Capacity



~150K

Operations per pixel knee-point

~86%

Asymptotic Accuracy

HAILO

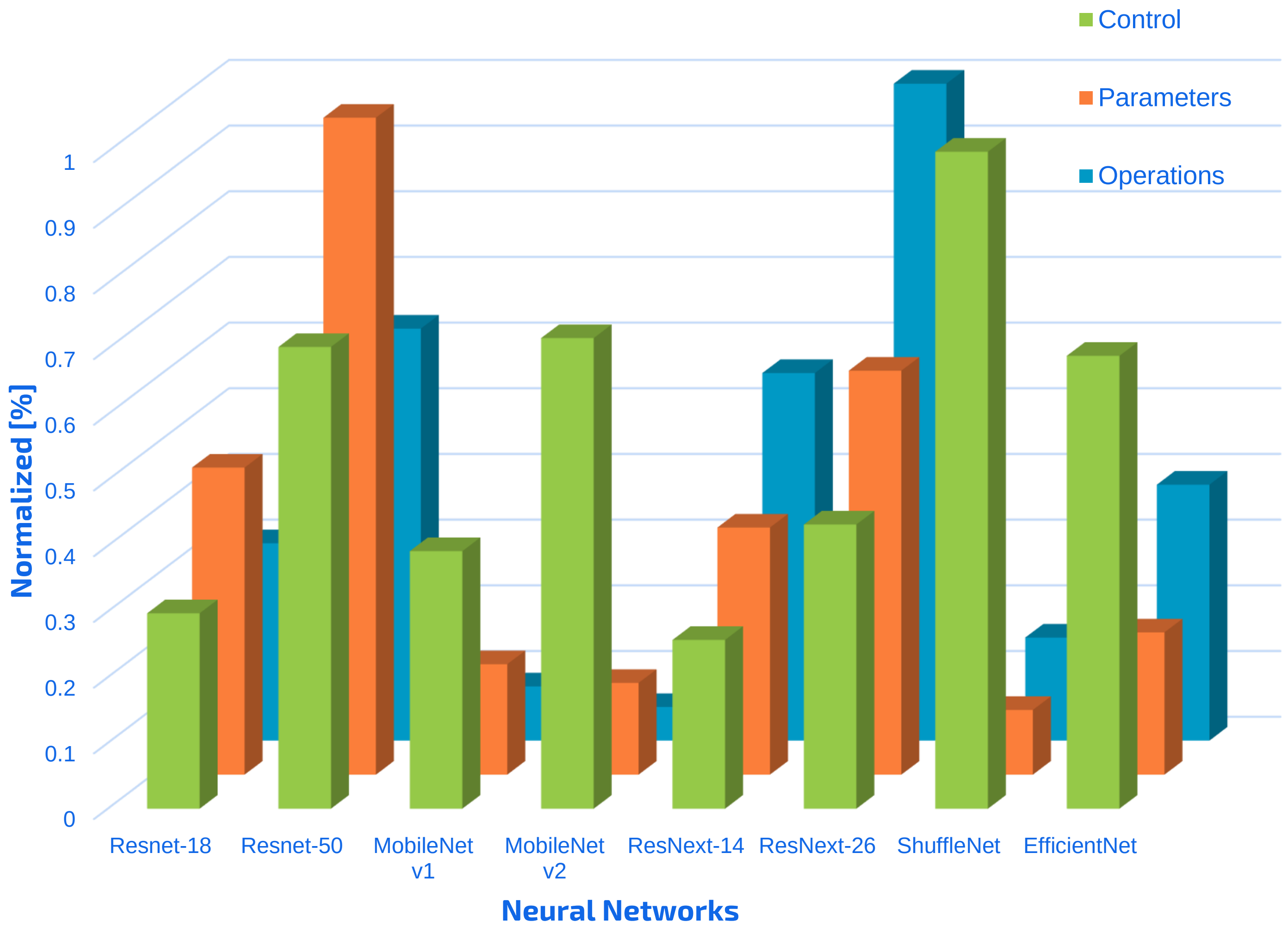
Architecture Building Block – Key Ingredients

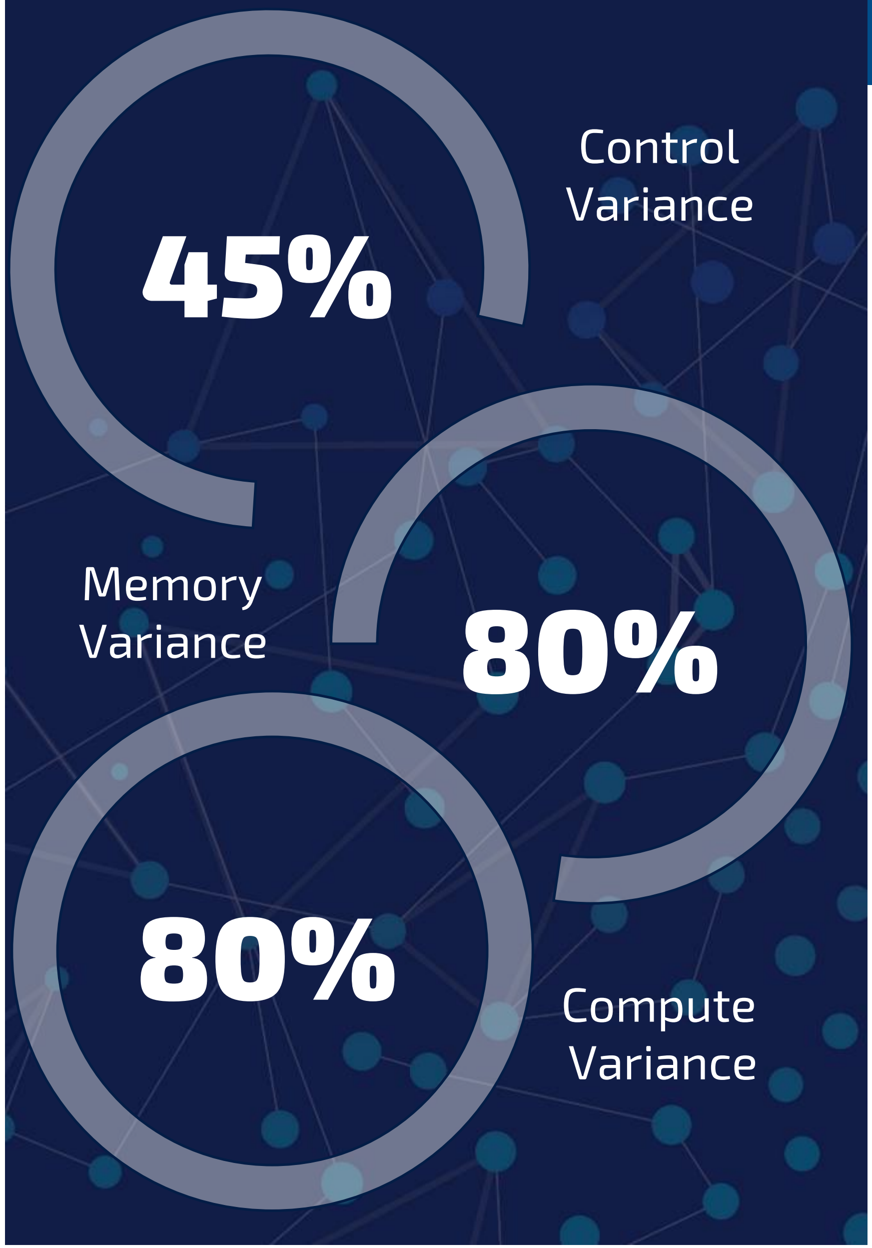
Control — Layers

Memory — Weights, Results

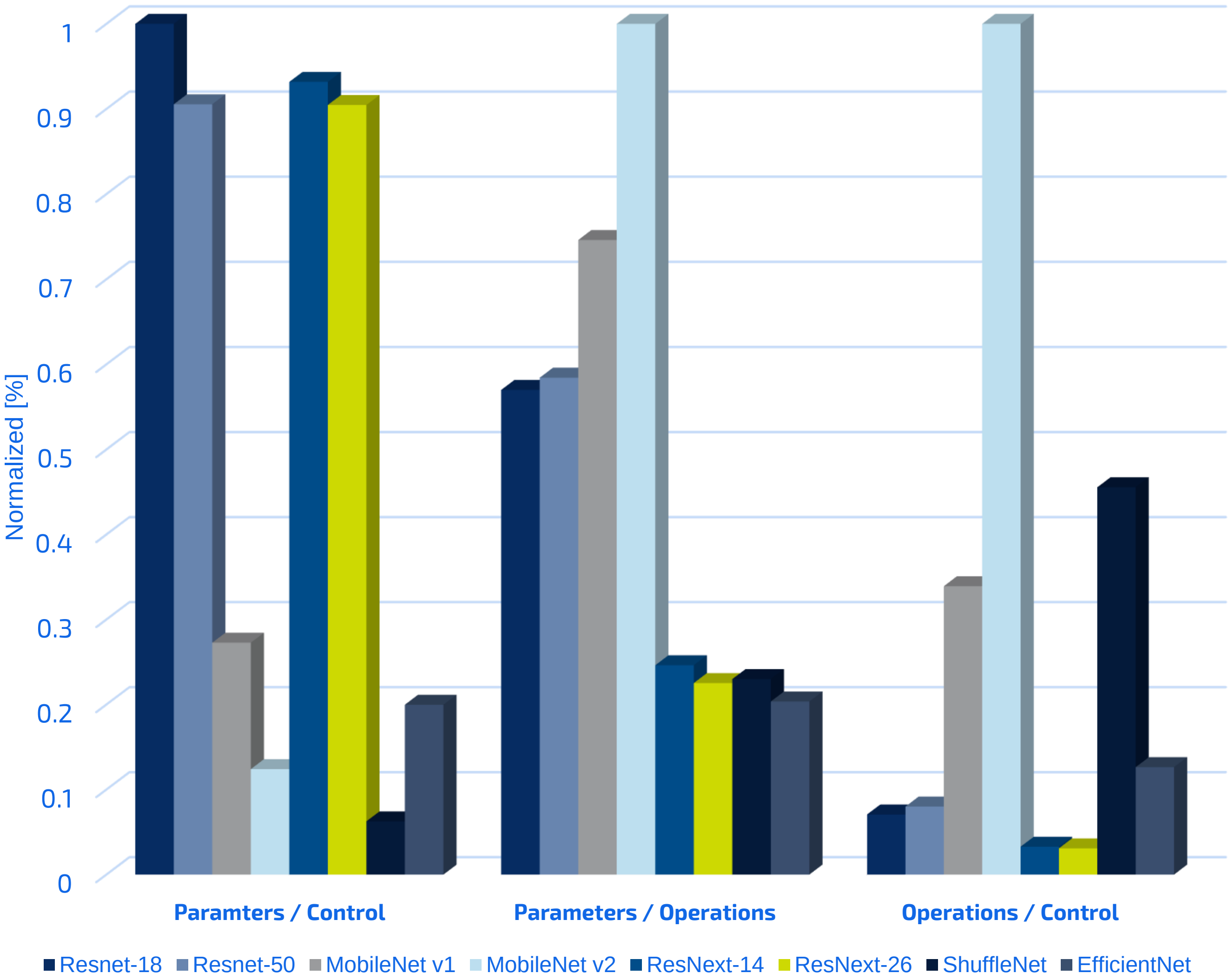
Compute — Operations

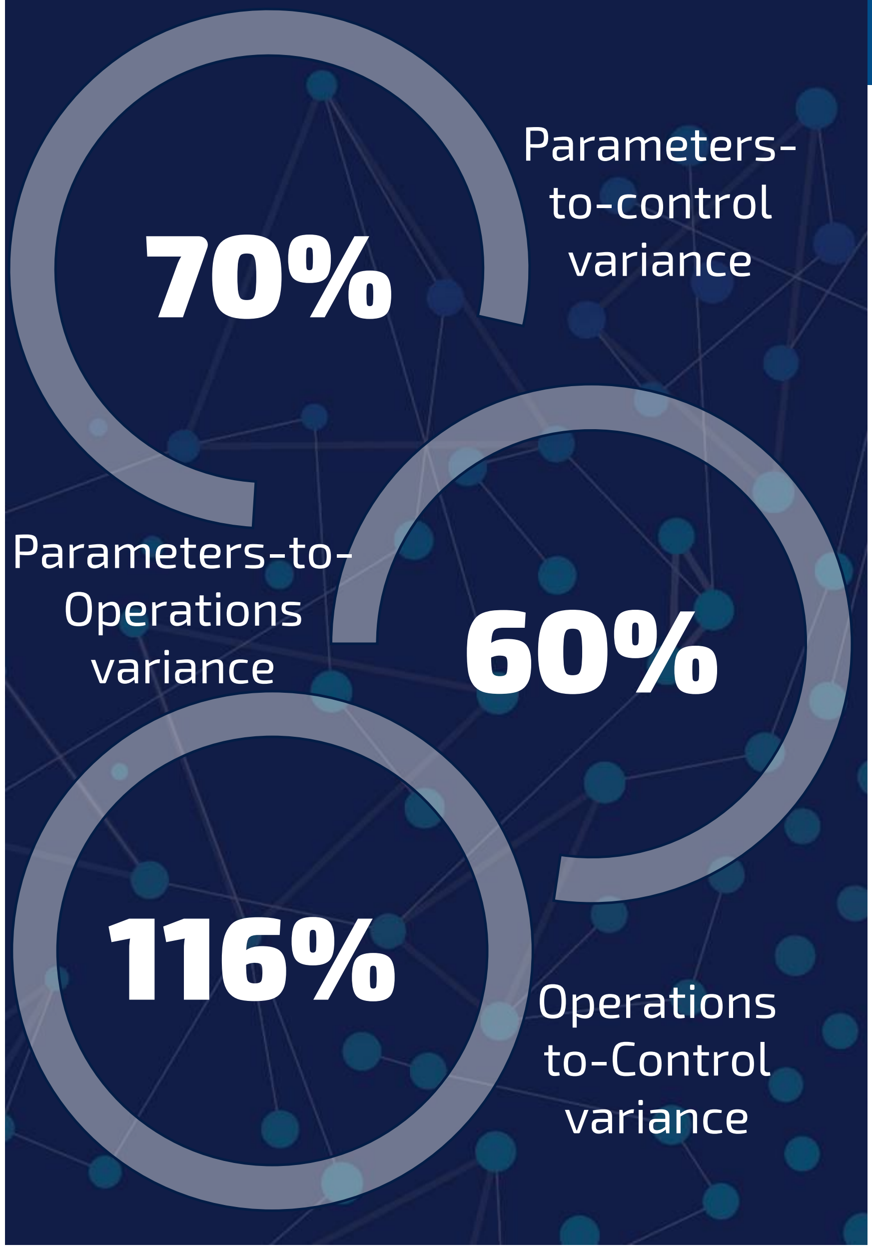
Inter-Model Diversity



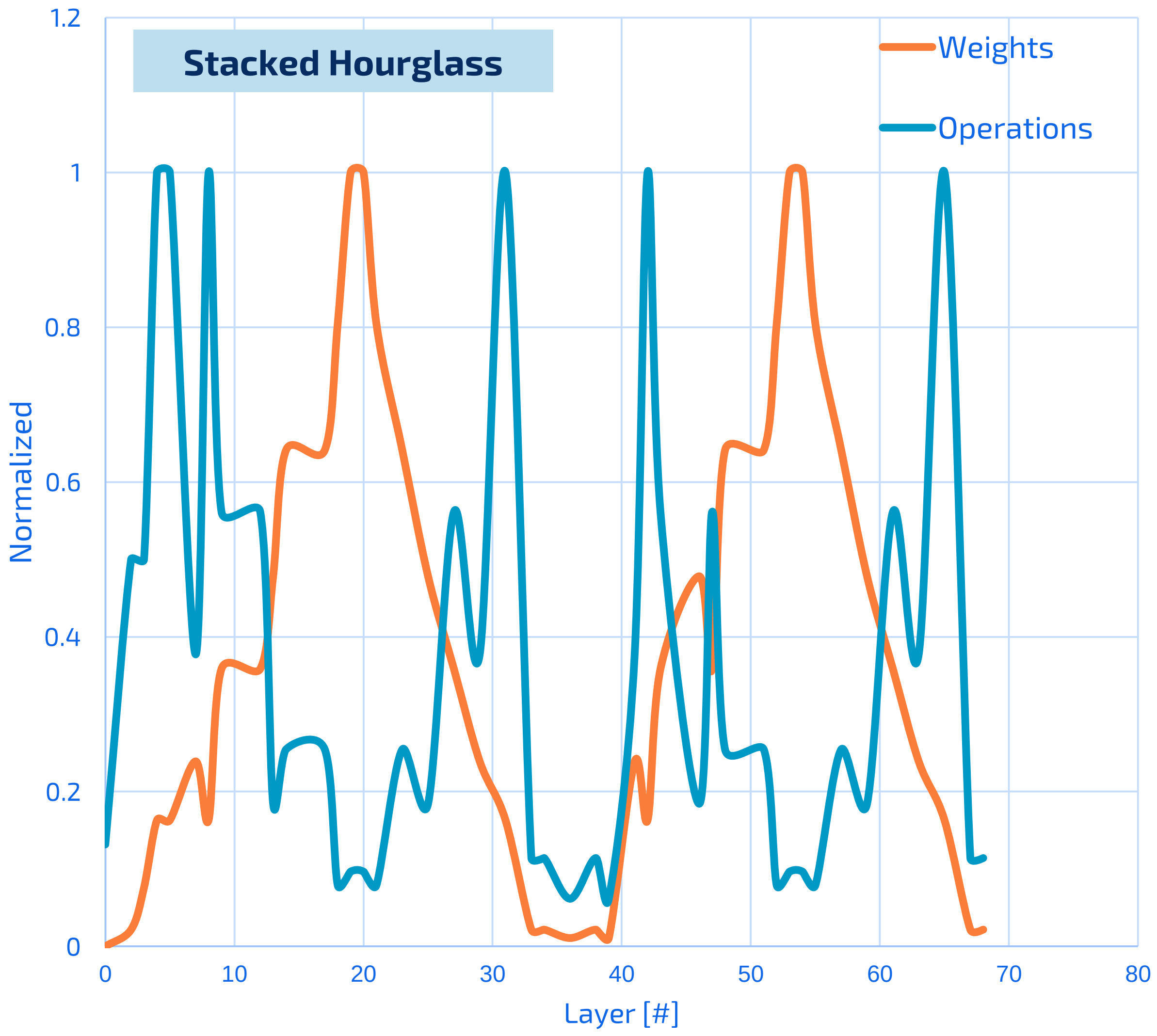


Inter- & Intra- Model Diversity





Intra-Model Diversity (an example)



80%

Per-layer weight relative diversity

86%

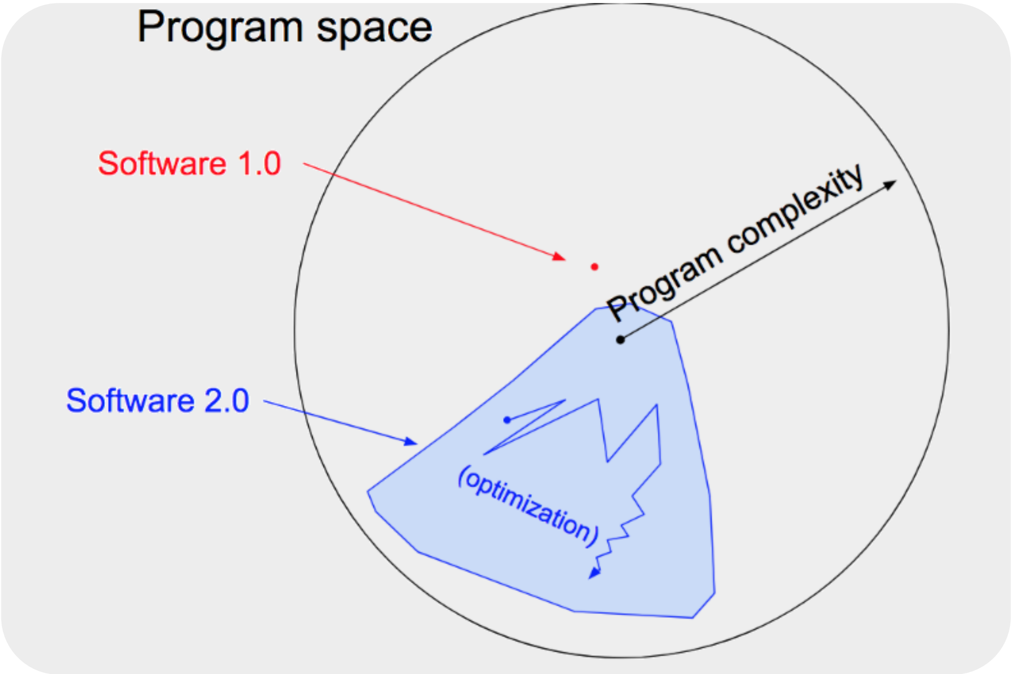
Per-layer operations relative diversity

Intra-Domain Flexibility

From application to implementation

- **Software** is a mapping from application domain to an underlying architecture

- Software is a prism for the match ("goodness-of-fit" between the domain and the machine it runs on)

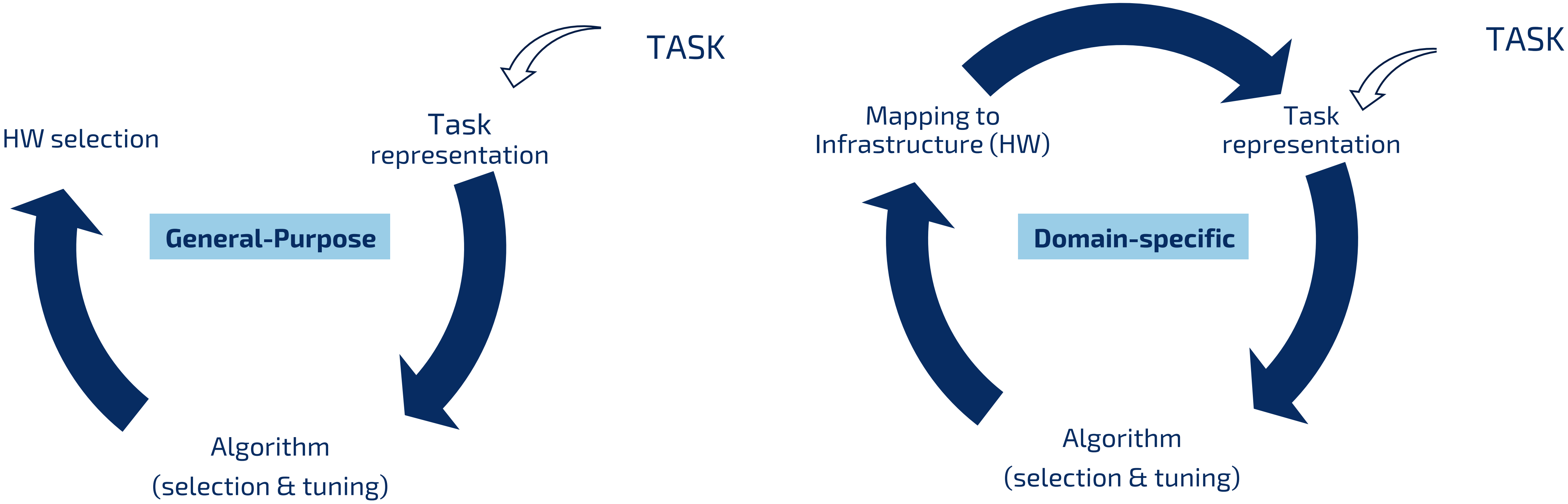


https://medium.com/@karpathy/software-2-0-a64152b37c35

Software "Program": side-by-side

	Traditional	Deep Learning
Data-driven design (example based rather than rules)	✗	✓
Data-driven control flow (runtime)	✓	✗
Defined by structure	✗	✓
'Code' is manually written	✓	✗
Deterministic behavior (for unobserved input)	✓	✗

From Application to Implementation or from Implementation to Architecture



Domain-specific architectures call for **greater coupling** between algorithms and the machine (HW) they run on

Parameter Sensitivity & Architecture Implications

- From architecture implication standpoint

  - A model ***M*** is characterized by <P, O, C>
    - P = No. of parameters (weights)
    - O = No. of operations
    - C = Control complexity (layers x layer complexity)

  - The implications are measured via
    - Performance (accuracy) – based on some quantitative quality criteria
    - Throughput
    - Latency
    - Energy

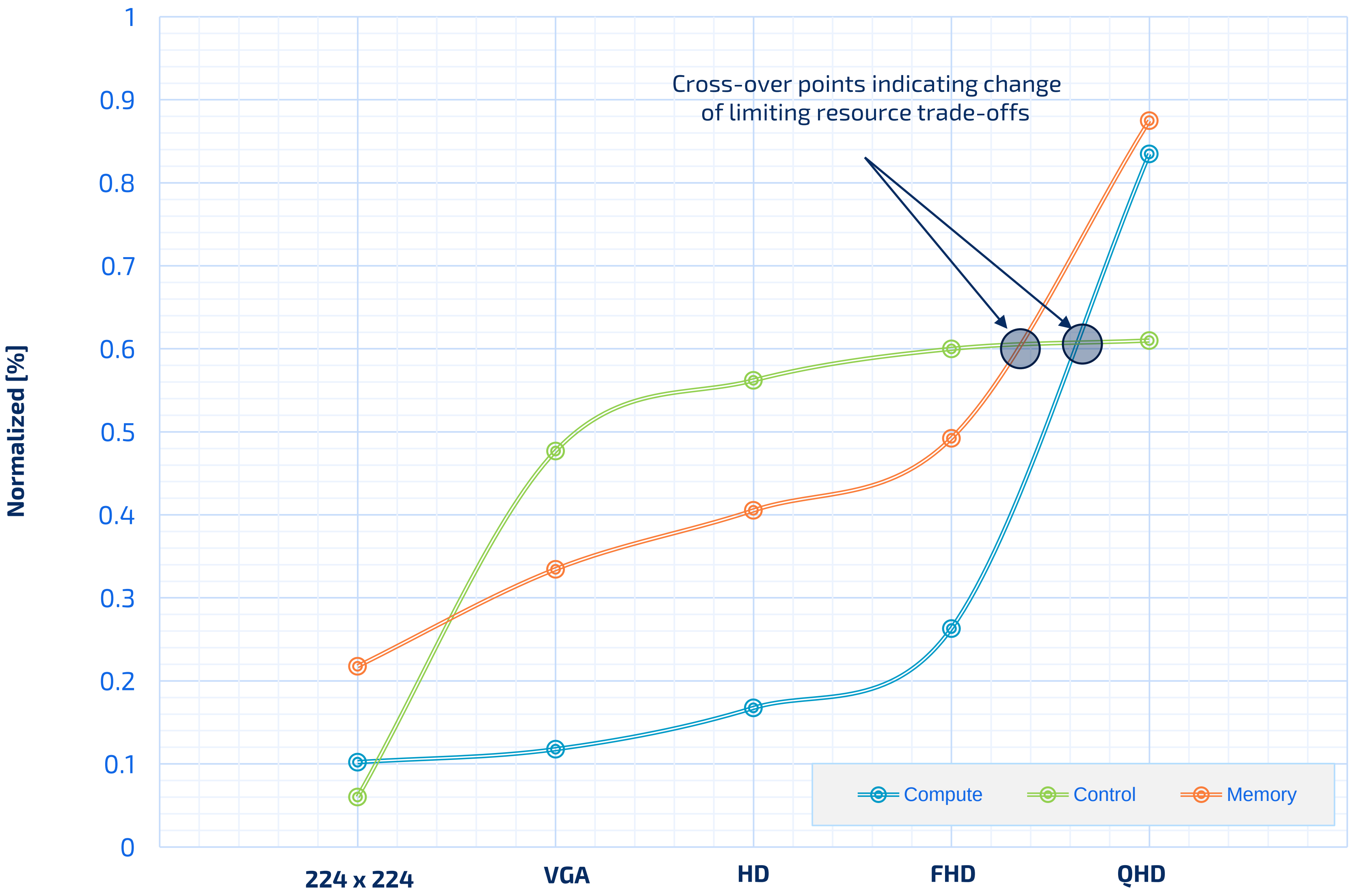
Architecture Scaling Challenges

- MLP (Fully-connected):
  - Control → L
  - Static Memory (weights) → $L * [\ W * H\ ]\ ^\wedge\ 2$
  - Dynamic Memory (results) → $L * [\ W * H\ ]$
  - Compute → $L * [\ W * H\ ]\ ^\wedge\ 2$

- CNN (Machine Vision):
  - Control → L
  - Static Memory (weights) → $L * [\ K * K * Fin * Fout\ ]$
  - Dynamic Memory (results) → $L * [\ W/s * H/s\ ] * Fin * Fout$
  - Compute → $L * [\ W * H * K * K * Fin * Fout\ ]$

L – Layers ; W x H – Resolution ; F_{in}, F_{out}, *K, s – Model attributes (in/out features, kernel size, scaling);*

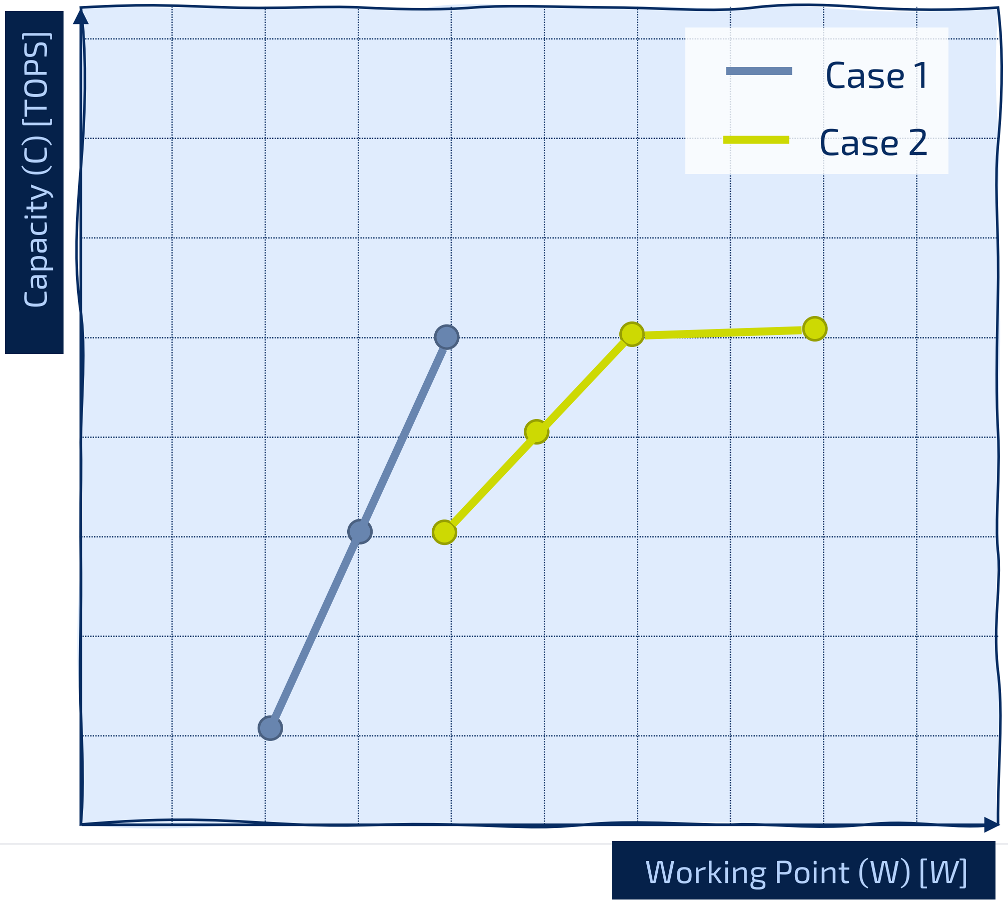
Model Size vs. Resolution, Bit-width

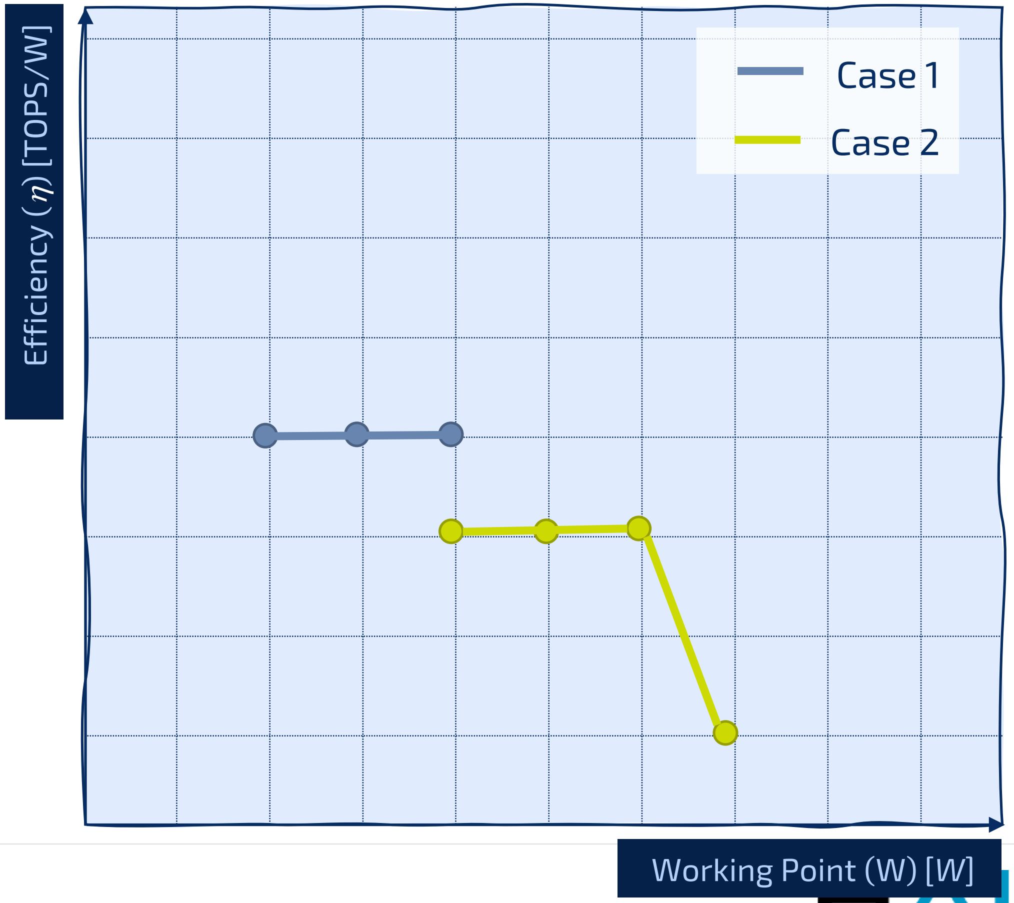


Efficiency, Capacity and Scalability

$$\eta = \frac{dC}{dW} \Leftrightarrow \mathrm{C} = \int \eta \cdot dW$$

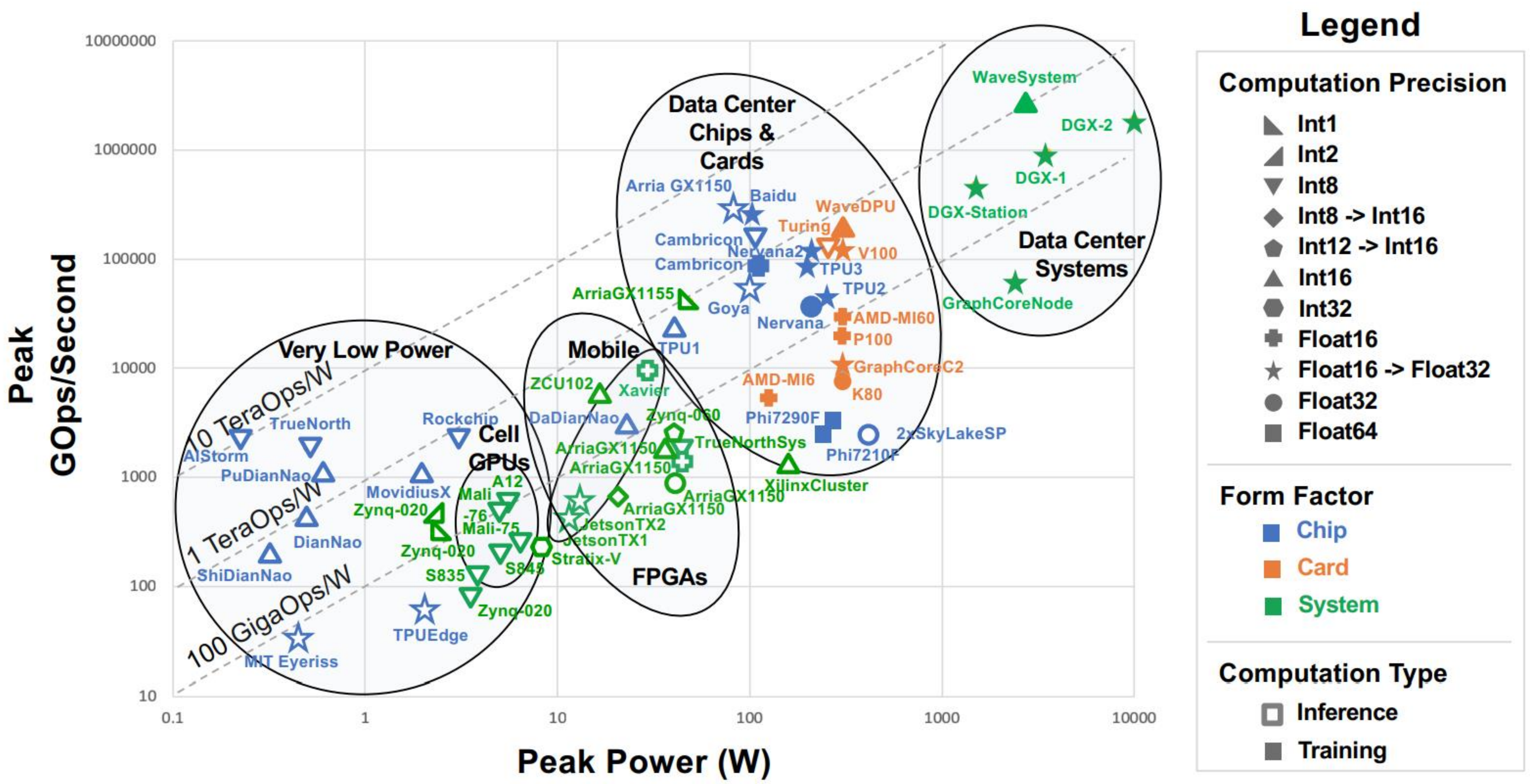
$$S = \left(\frac{d\eta}{dW}\right)^{-1}$$





HAILO

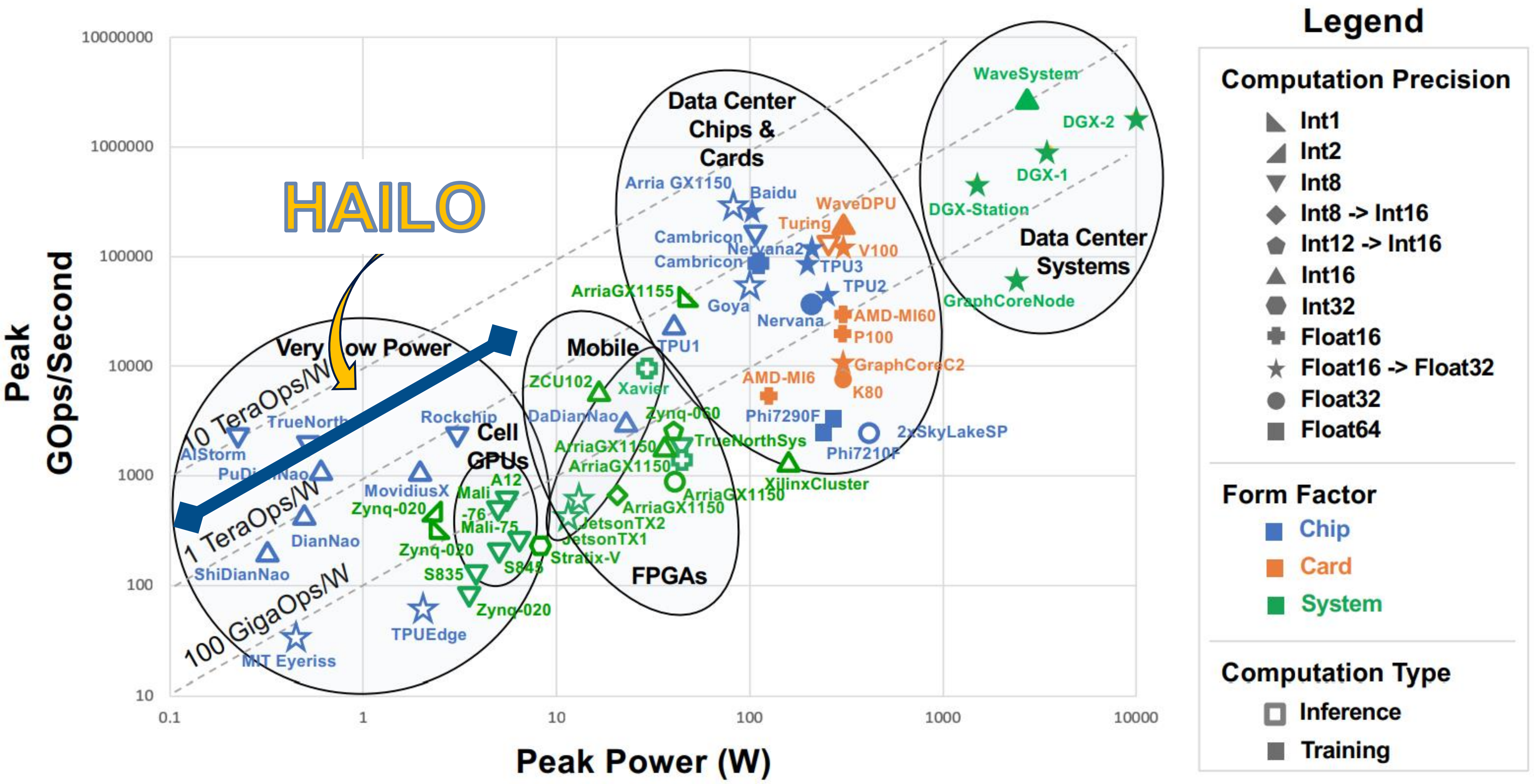
Efficiency, Capacity and Scalability



[Survey and Benchmarking of Machine Learning Accelerators (1908.11348)]

HAILO

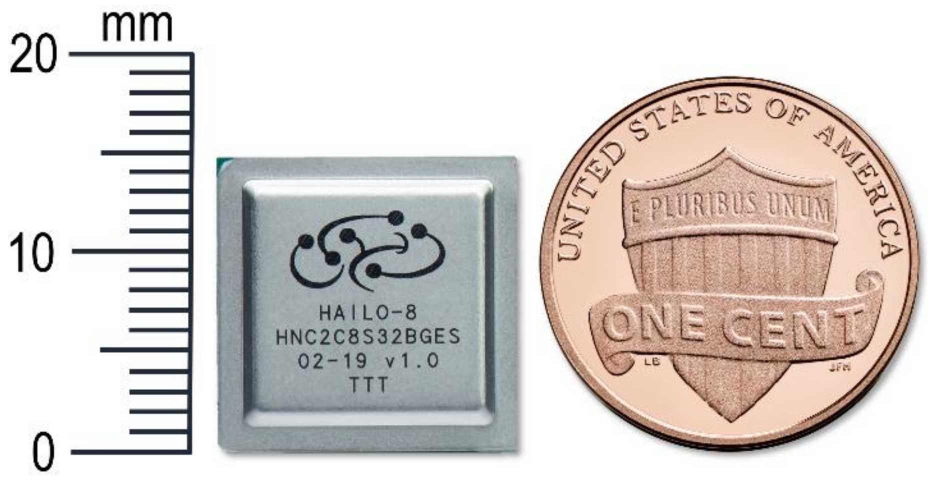
Efficiency, Capacity and Scalability



[Survey and Benchmarking of Machine Learning Accelerators (1908.11348)](https://arxiv.org/abs/1908.11348)

Hailo-8™ Highlights

The World's Most Powerful and Efficient Edge AI Processor



High Performance
26 TOPS

Automotive
ASIL-B (D)
AEC-Q100 Grade-2

Self Contained
No External DRAM

Flexible
Fully Programmable
Comprehensive SDK

High Efficiency
3 TOPS/W

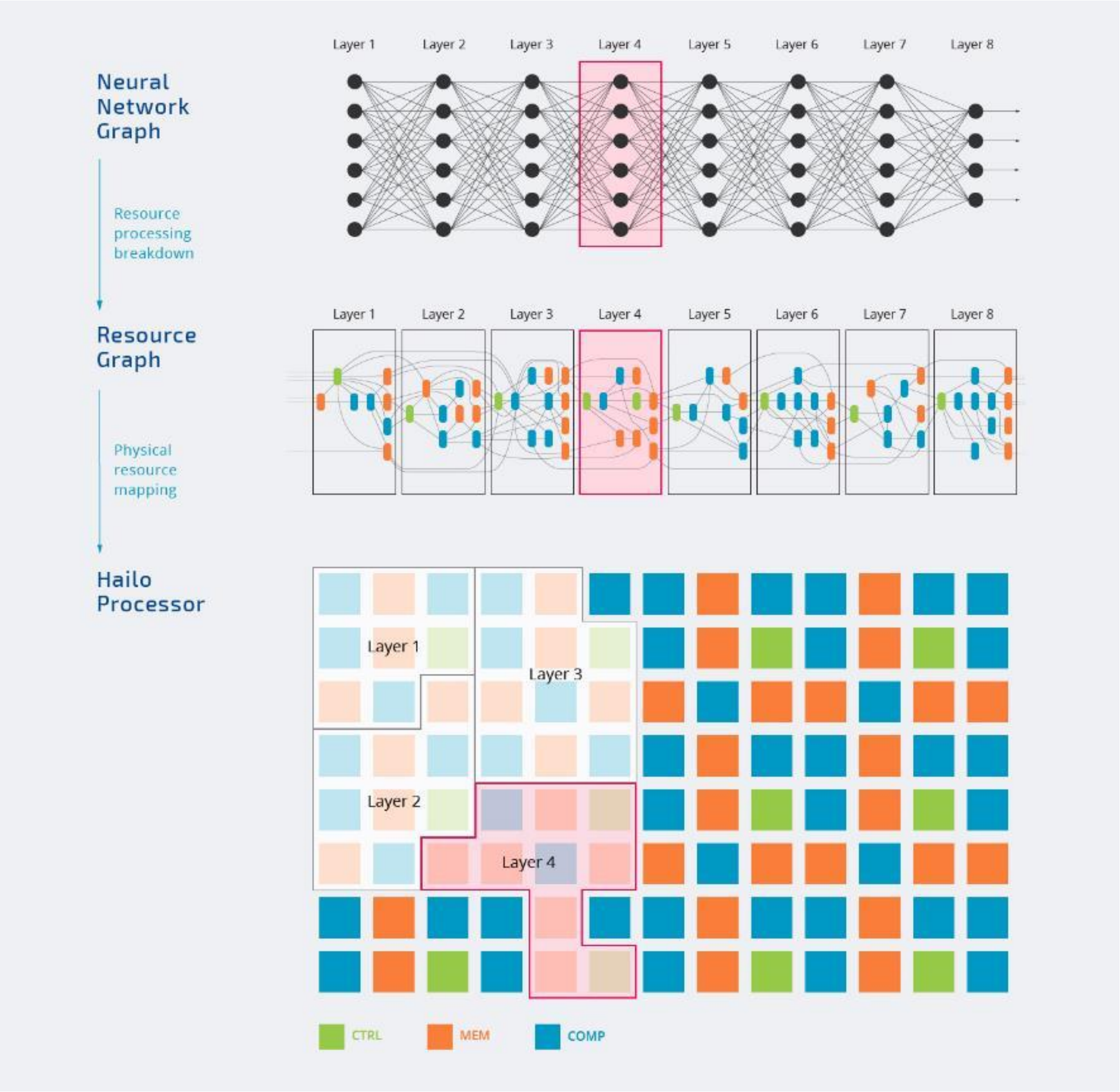
Deployment
Stand alone
Co-processor
Modes

Structure-defined Data Flow Architecture

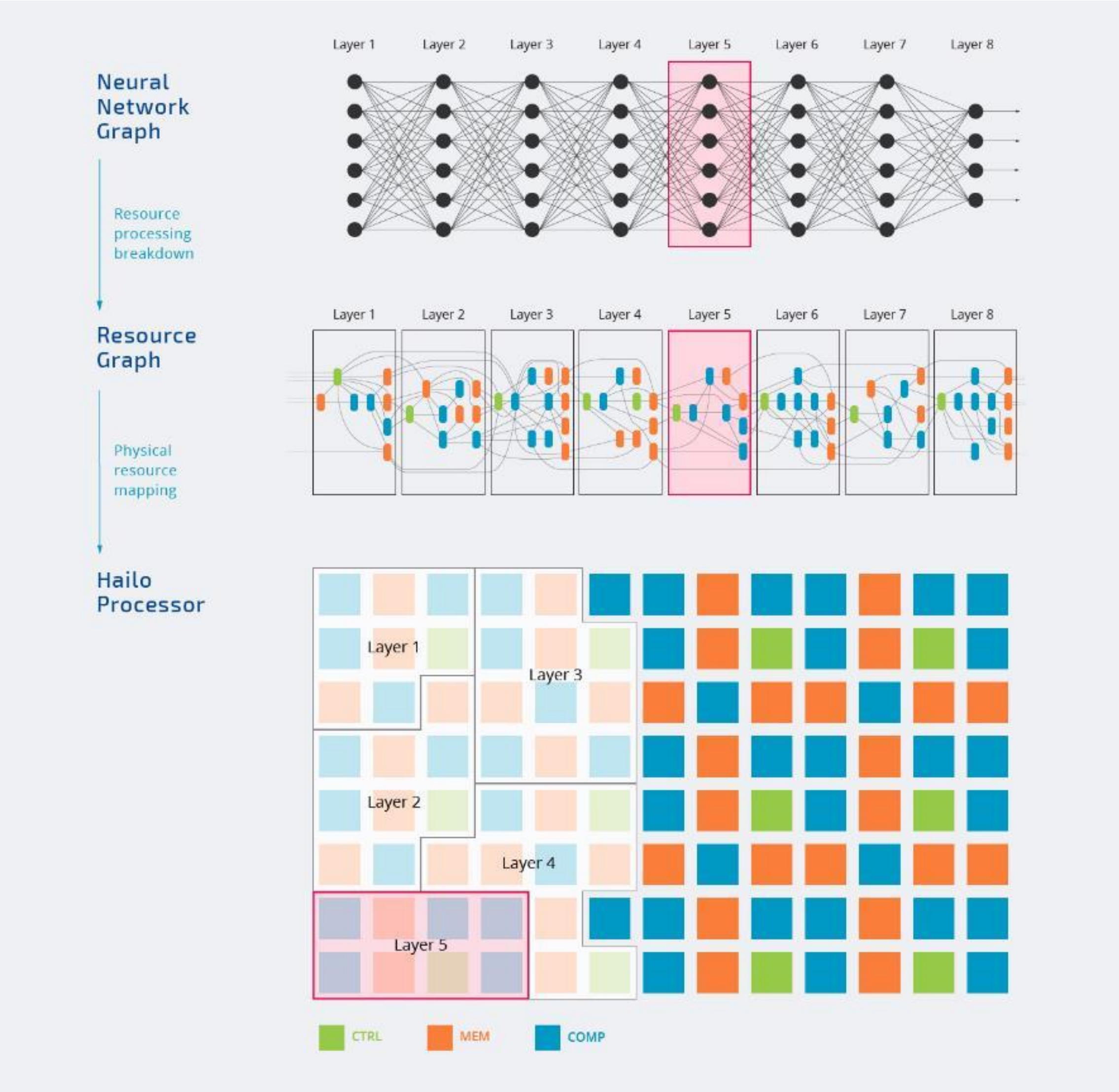
Structure-defined Data Flow Architecture

Structure-defined Data Flow Architecture

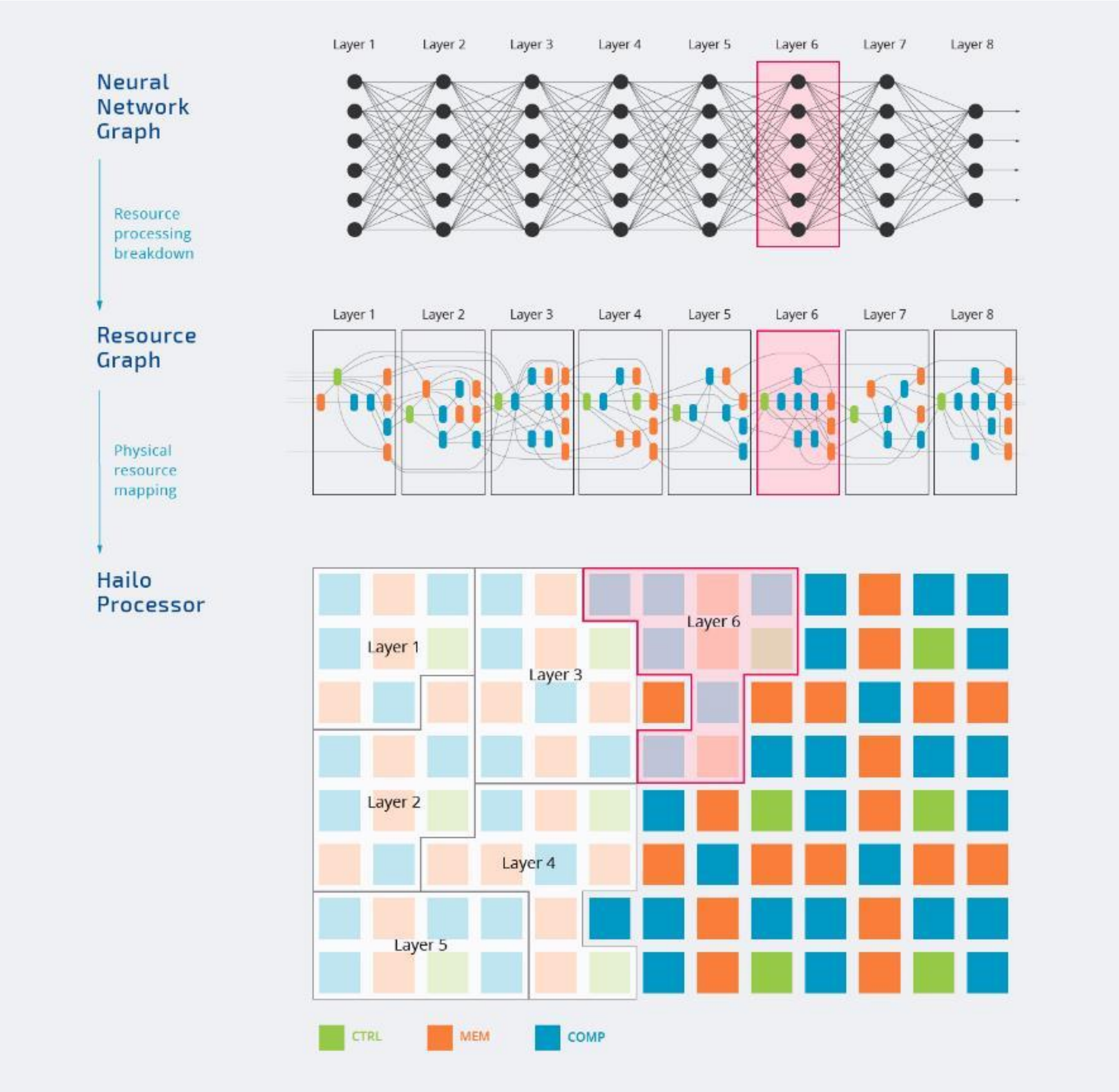
Structure-defined Data Flow Architecture



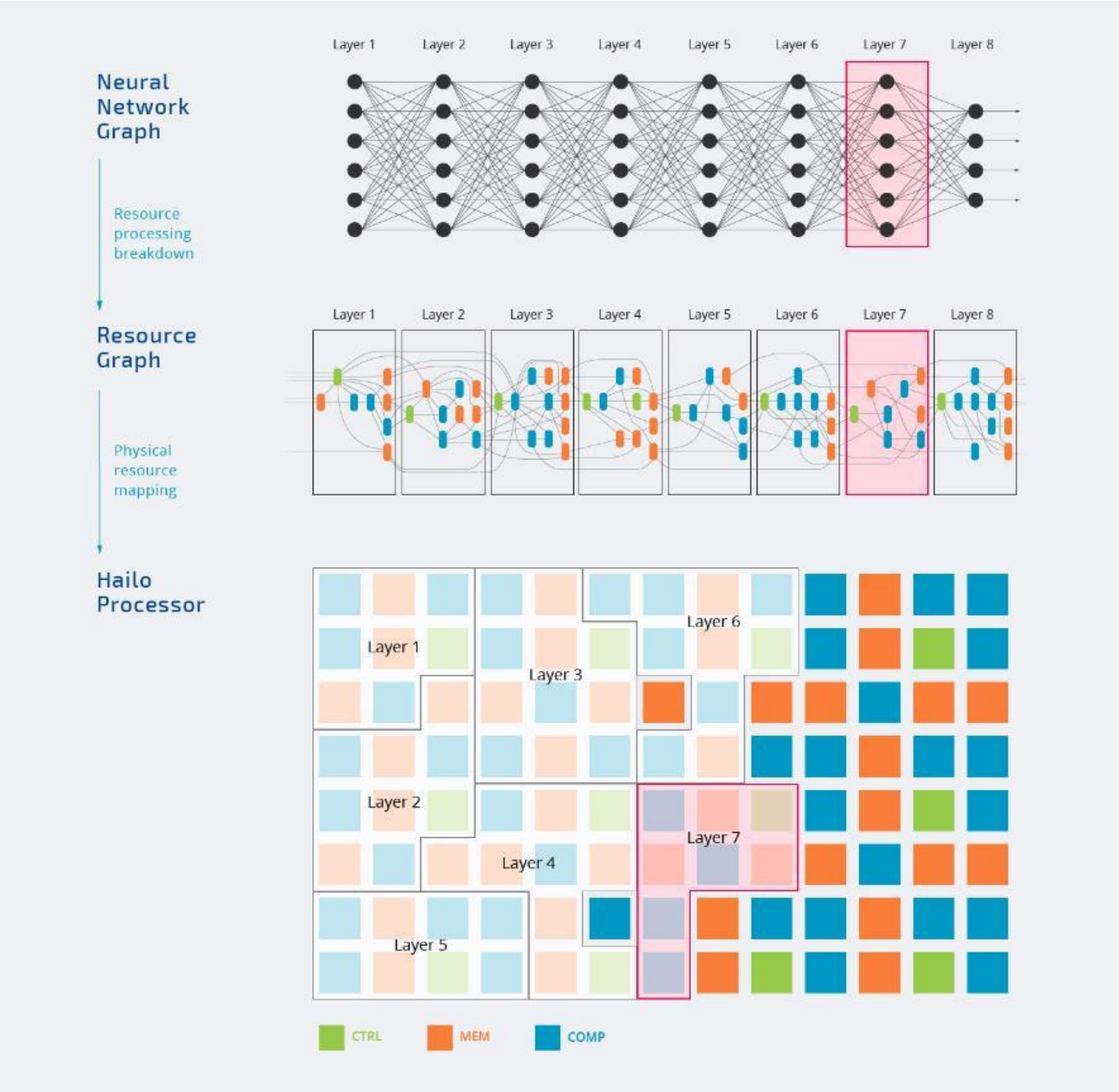
Structure-defined Data Flow Architecture



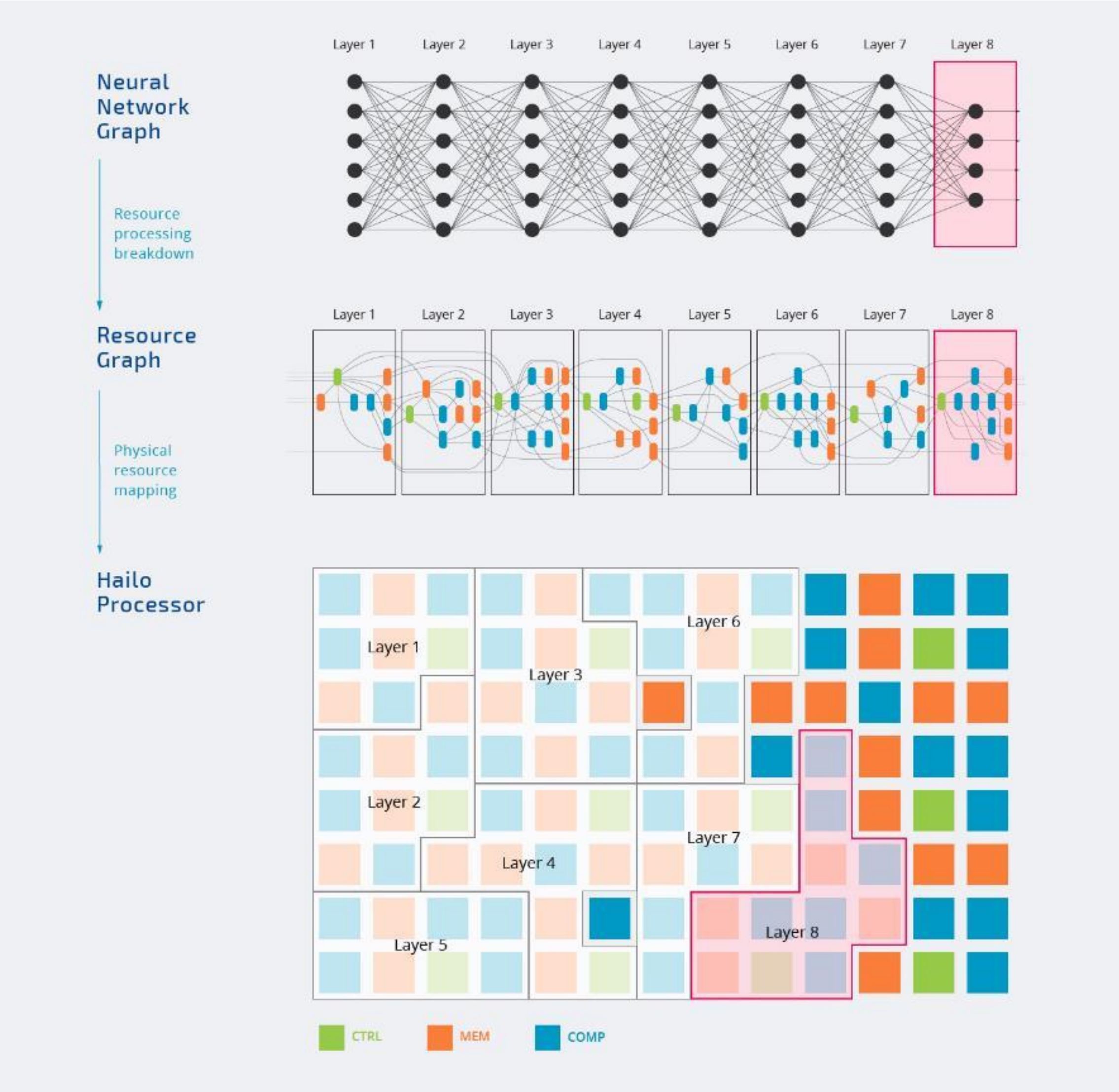
Structure-defined Data Flow Architecture



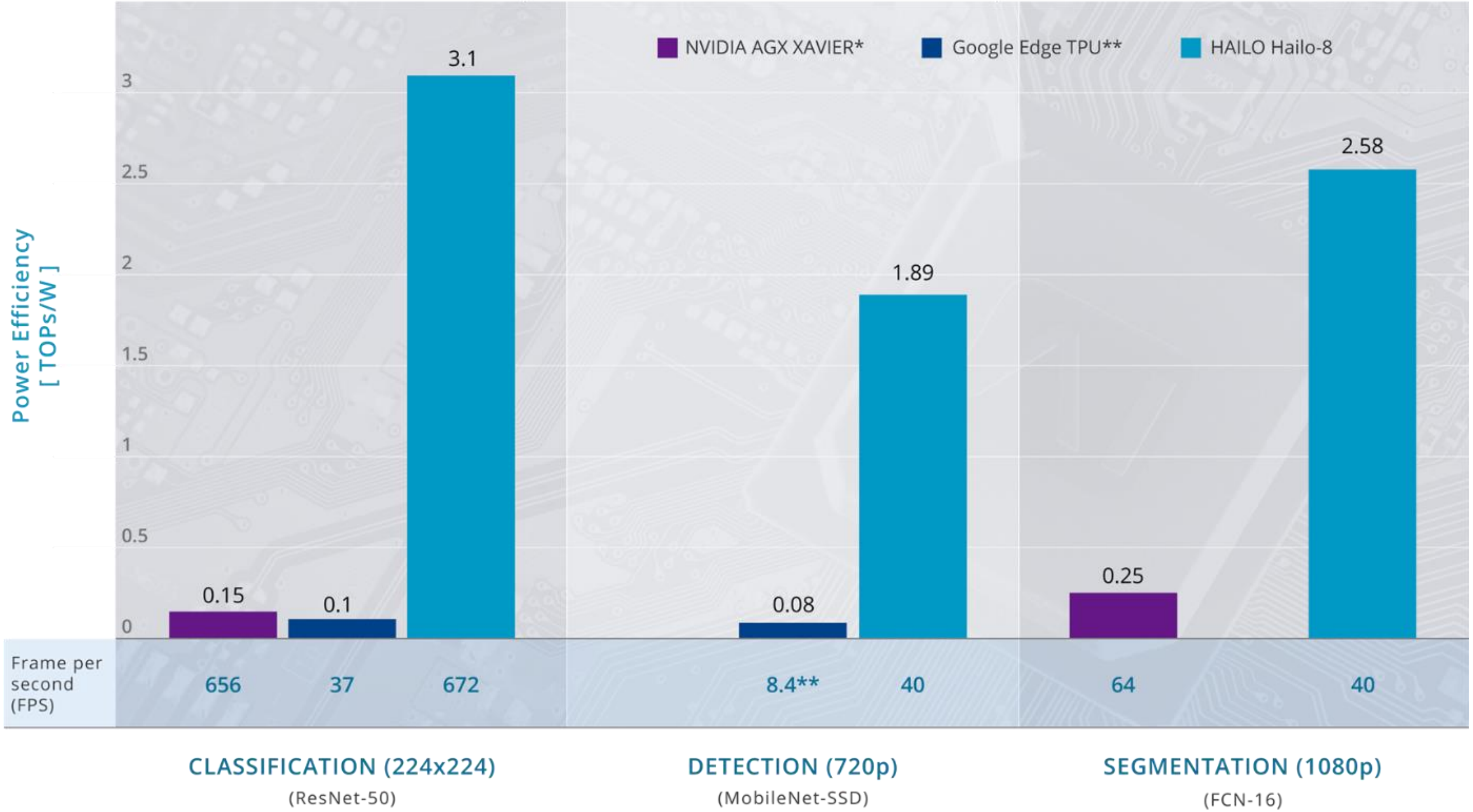
Structure-defined Data Flow Architecture



Structure-defined Data Flow Architecture



Hailo-8 Flexibility: Task Diversity



Concluding Remarks

NEXT COMPUTER ARCHITECTURE OPPORTUNITY IS DOMAIN-SPECIFIC

MACHINE VISION IS A WIDE-ENOUGH DOMAIN WITH STABLE PROPERTIES

A PROPER ARCHITECTURE SHOULD ENABLE HIGH DIVERSITY PER MODEL

THE ARCHITECTURE FLEXIBILITY SHOULD BE ACCESSIBLE TO THE PROGRAMMER

EFFICIENCY IS A KEY ATTRIBUTE FOR SCALABILITY

SCALABILITY IS A CRITICAL FACTOR FOR A VIABLE FORWARD-LOOKING ARCHITECTURE

HAILO

Empowering Intelligence

THANK YOU!

Email: contact@hailo.ai

https://www.hailo.ai

Empowering Product Creators to Harness Embedded Vision

The Embedded Vision Alliance (www.Embedded-Vision.com) is a partnership of 100+ leading embedded vision technology and services suppliers, and solutions providers

Mission: Inspire and empower product creators to incorporate visual intelligence into their products

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.Embedded-Vision.com

The Alliance enables vision technology providers to grow their businesses through leads, ecosystem partnerships, and insights

For membership, email us: membership@Embedded-Vision.com

Embedded Vision Insights
The Latest Developments on Designing Machines that See

Join us at the Embedded Vision Summit

May 18-21, 2020—Santa Clara, California

The only industry event focused on enabling product creators to create "machines that see"

- *"Awesome! I was very inspired!"*
- *"Fantastic. Learned a lot and met great people."*
- *"Wonderful speakers and informative exhibits!"*

Embedded Vision Summit 2020 highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos** of the latest apps and technologies

Visit www.EmbeddedVisionSummit.com to sign up for updates