2020 embedded VISION summit

Getting Efficient DNN Inference Performance: Is it Really All About the TOPS?

Gary Brown, Intel IOT Group

September 22, 2020

intel

Notice and Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles https://www.intel.com/content/www/us/en/policy/policy-human-rights.html Intel's products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

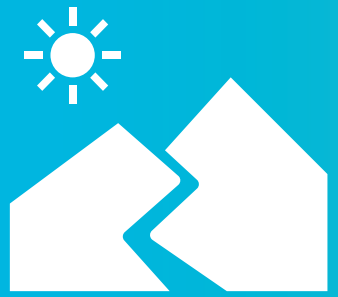
Agenda

The AI Inference Landscape and Growth of the Edge

What about Benchmarking

Comparing AI Inference Silicon: Is it Really All About the TOPS

A New Era of Efficient Architectures



The AI Inference Landscape and Growth of the Edge

The DL Inference Landscape is Vast

How to Compare

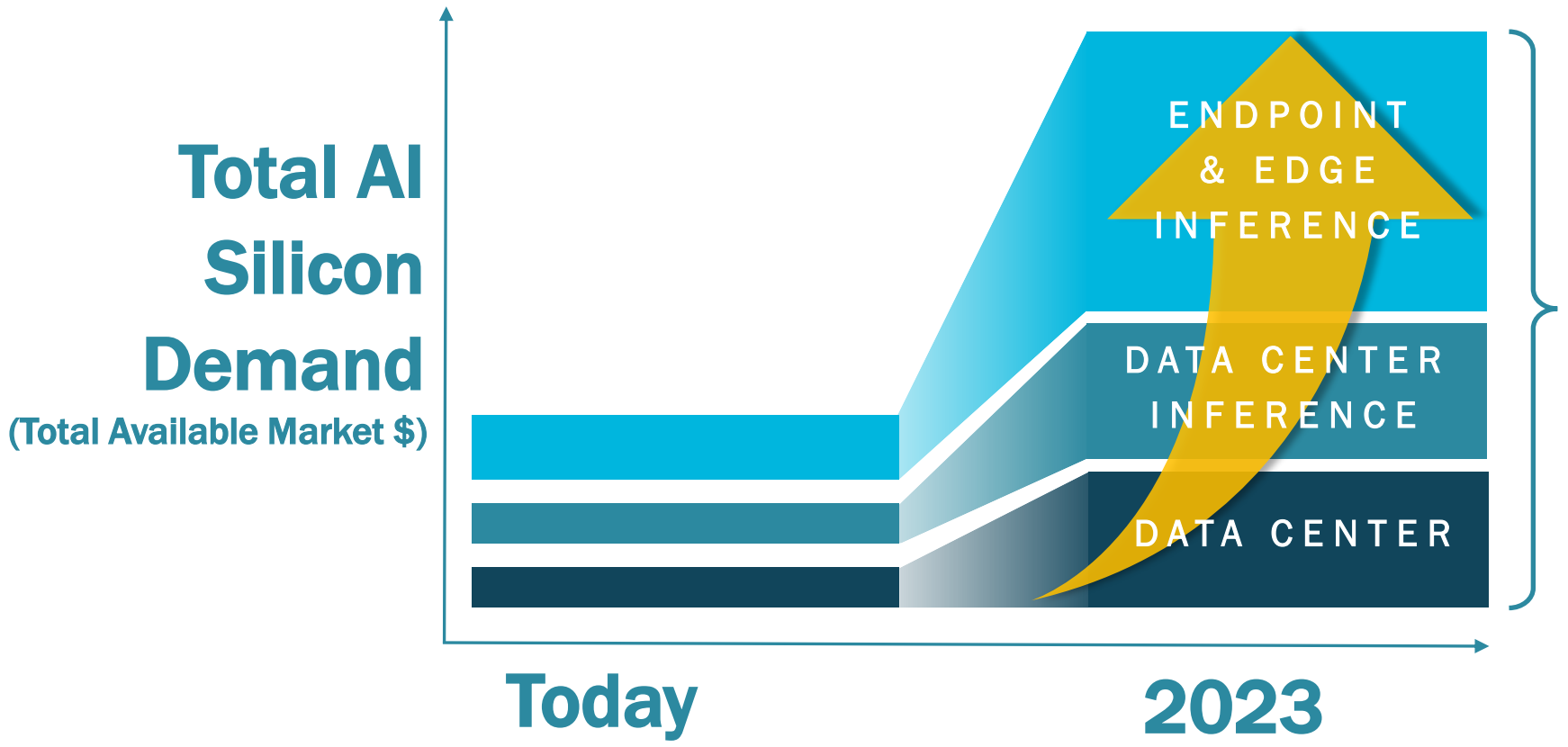
General Purpose Platforms

Edge Devices, Edge Servers

Cloud / Data Center

Purpose-built Platforms

Edge AI Silicon Opportunity Is Growing Fast

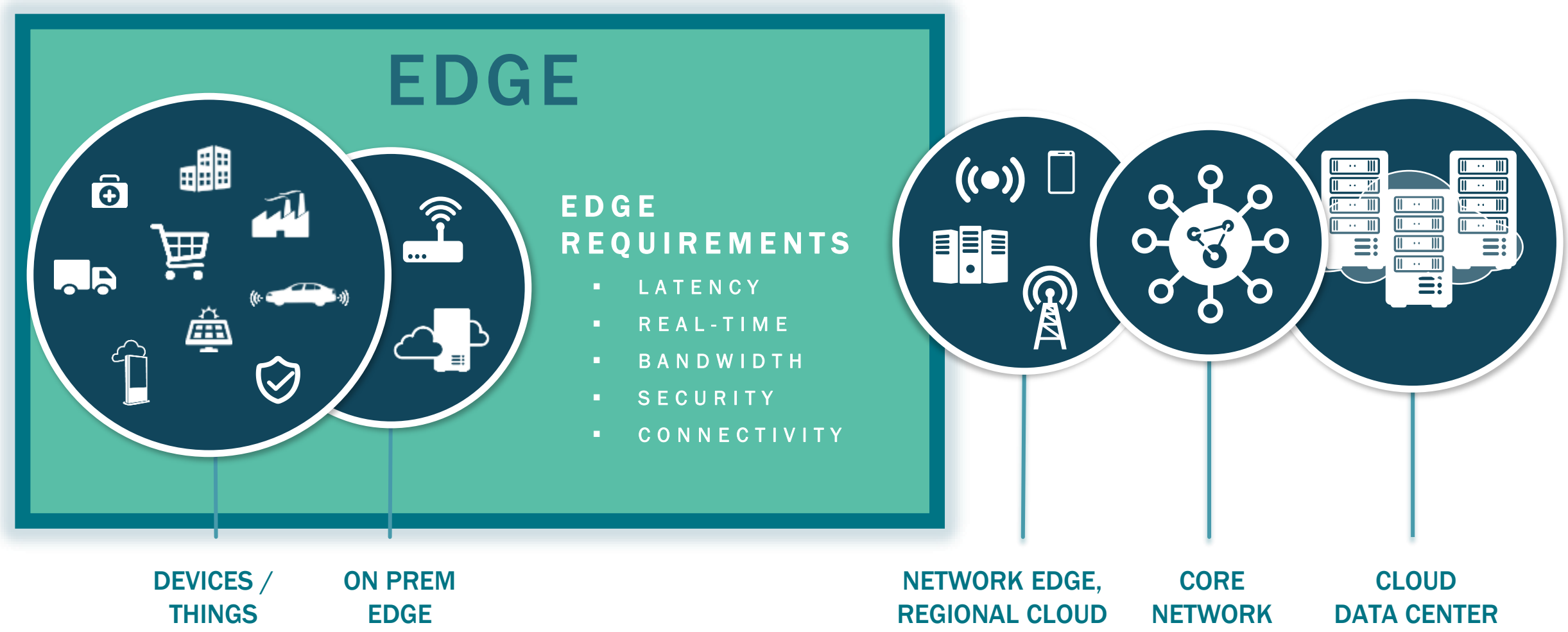


25%-35%
CAGR In AI For Silicon Industry

Growth larger at the Edge,
Compared To Data Center

Note: Endpoint and Edge excludes smartphones and client PCs Source: Intel, IDC, Gartner

AI Inference at the Edge has Many Challenges



What about Benchmarking?

How We Benchmark Silicon Platforms – What's Missing?

MIPS or IPC

TOPS or TFLOPS

spec

EMBC EMBEDDED MICROPROCESSOR BENCHMARK CONSORTIUM

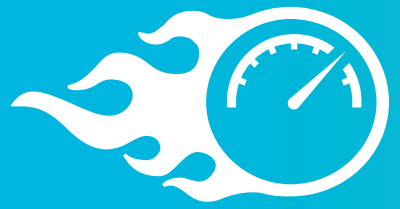
BDTi

MLPerf

Benchmarks Making Progress, and Scope is Getting Larger

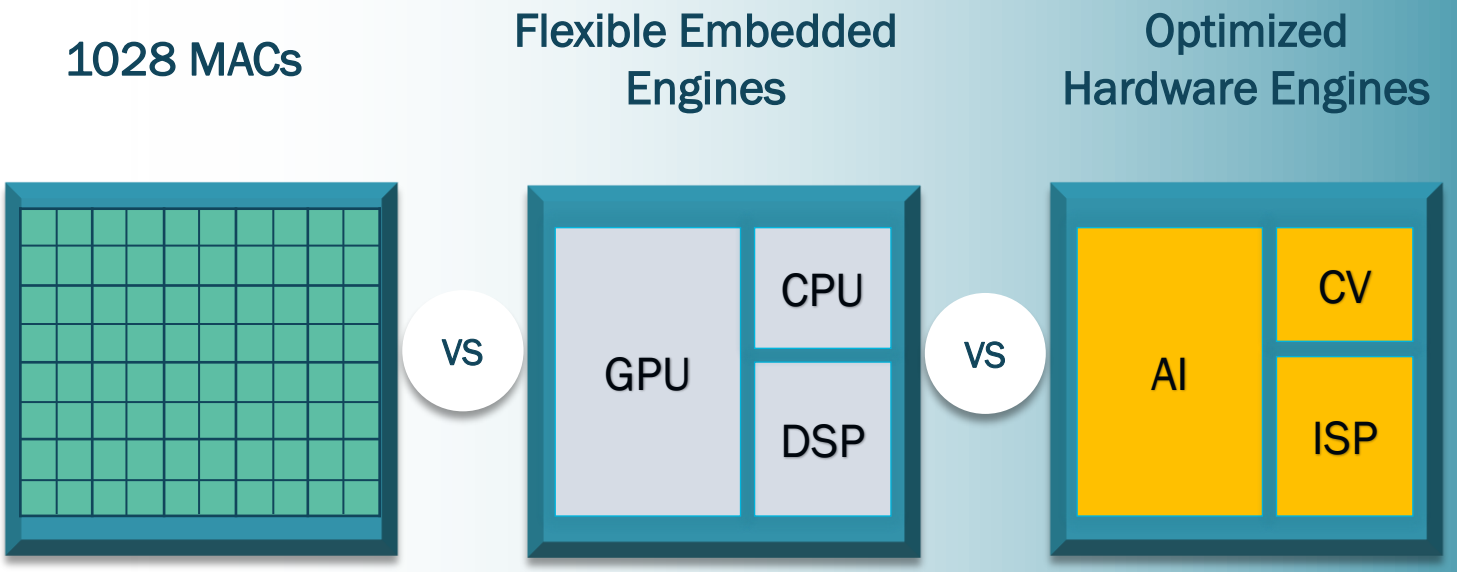
Dimensions of Comparison:

- Flexibility
- Compute Efficiency
- IOT Fortified
- Security
- Software Portability
- ...and many more



Comparing AI Inference Silicon: Is It Really All About the TOPS?

Use TOPS with Caution



- the amount of actual work accomplished
- either per clock cycle
- or per time
- or per watt
- or per $ of cost….

An Example: Which is Higher Performance?

2020 embedded VISION summit

Work accomplished per second

Child: **10** scoops/min

Commercial excavator: **1** scoop/min

Excavator: 10x slower?!

Which Metric You Use Makes all the Difference

2020 embedded VISION summit

Performance / Throughput	Workload Performance per watt
Power Efficiency	Inferences/sec/watt
Cost Efficiency	Inferences per second per $
Silicon Cost Efficiency	Inferences per second per mm^2 silicon

intel

A New Era of Efficient Architectures

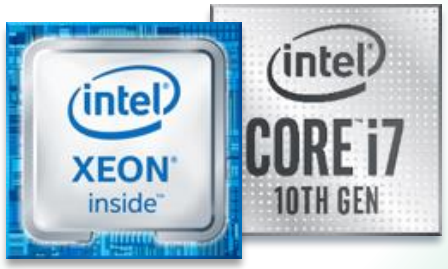
Edge AI with Intel: Examining Compute Efficiency

2020 embedded VISION summit

WORKLOAD BREADTH

AI SPECIFIC

CPU



Multi-Purpose
AI Foundation

GPU

AI, HPC, Media
& Graphics

FPGA

Real-Time &
Multi-Use DL
Inference

VPU

Edge
DL Inference

DEDICATED

habana
An Intel Company

Data Center
DL Inference
(Goya)

Data Center
DL Training
(Gaudi)

AI Acceleration with 11th Gen Intel® Core® Processors

- The latest Intel® Core™ processor for small form-factor servers, such as for digital signage with AI
- Optimized for industrial, public sector, transportation, retail, video analytics NVRs
- An efficient CPU with AI acceleration built-in using Intel® Deep Learning Boost
- Additional AI inference acceleration with Xe architecture

Intel® Xeon® Processors with Built-in AI Acceleration

2020 embedded VISION summit

- For larger servers with more video channel density
- DL inference built-into the CPU without the need for external AI accelerator
- Enough DL performance to run real-time inference across many channels of video
- No added AI accelerator card required.

NEW INTEL® XEON® GOLD 6258R PROCESSORS

UP TO 3.2X
AI INFERENCE PERFORMANCE WITH INTEL® DL BOOST
Compare to Intel® Xeon® Gold 6152 Processor

56 channels
VIDEO DECODE CAPABILITY
1080p30 Resolution, H.264/H.265

Optimized Software and Frameworks:

Caffe mxnet OpenVINO ONNX PaddlePaddle PyTorch TensorFlow

Performance results are based on testing as of 12/25/2019 and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.
Configuration Details: 1-node, 2x Intel® Xeon® Gold 6258R on Intel Reference platform with 384 GB (12 slots / 32 GB / 2933) total memory, ucode 0x500002c, HT on, Turbo on, with Ubuntu19.10, 5.3.0-24-generic, AIXPRT Image Classification AIXPRT v1.01 Intel® distribution of OpenVINO version 2019 R3 ResNet50 v1, using INT8 with Intel® DL Boost, BS=4, no.of instances =56 as of 01/29/2020, test by Intel on 12/25/2019. 1-node, 2x Intel® Xeon® Gold 6152 on Intel Reference platform with 384 GB (12 slots / 32 GB / 2933) total memory, ucode 0x2000065, HT on, Turbo on, with Ubuntu19.10, 5.3.0-24-generic, Image Classification AIXPRT v1.01 Intel® distribution of OpenVINO version 2019 R3 ResNet-50 v1 , Using INT8, BS=128, no.of instances =44 as of 01/29/2020, test by Intel on 12/25/2019.

intel

Gen 3 Intel® Movidius™ VPU for Power-Efficient Edge AI

2020 embedded VISION summit

FAST

vs. Nvidia Xavier **On Par**[1]

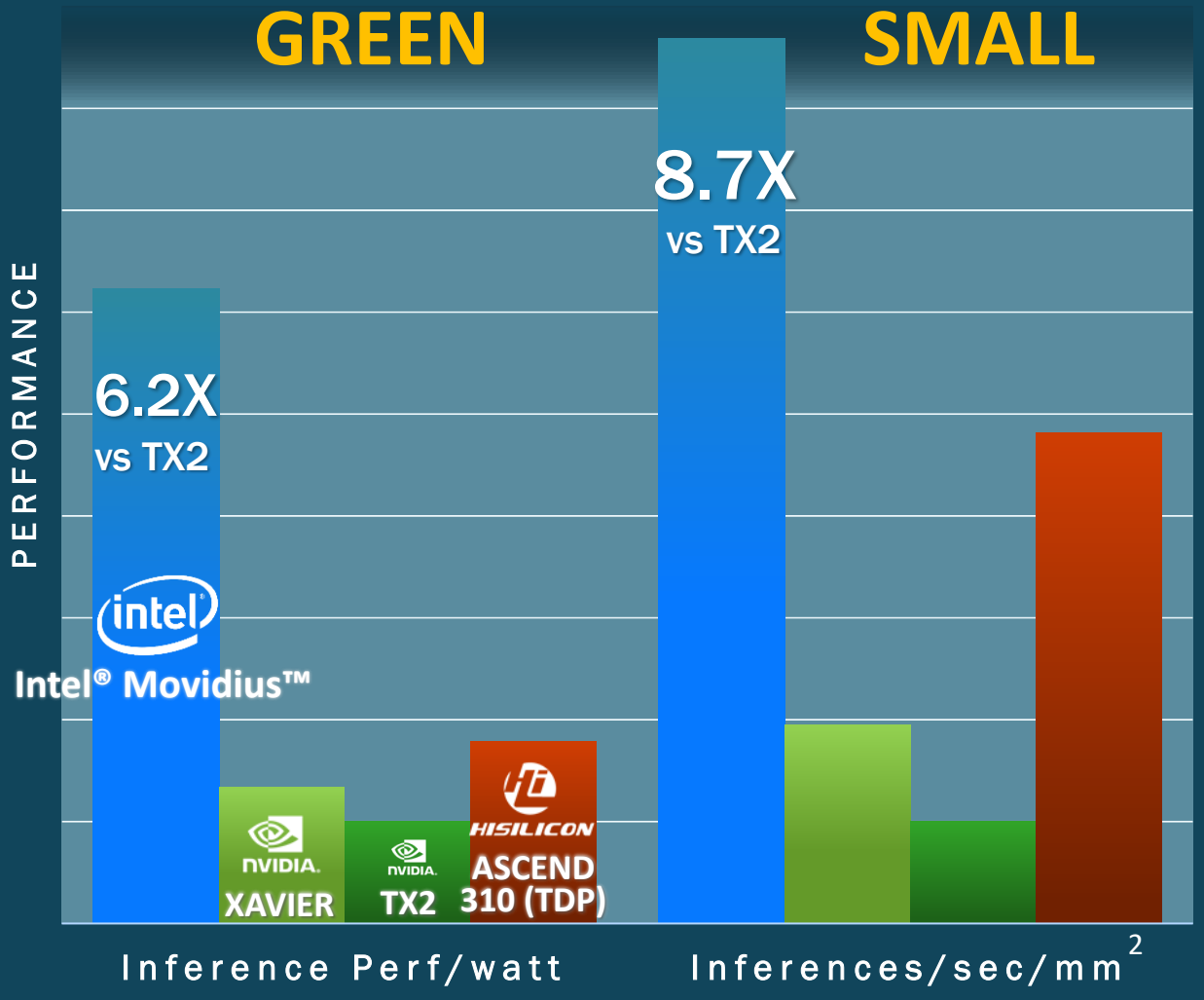
@ 1/5TH Power

1.25X Ascend 310

4X Nvidia TX2

GREEN

SMALL



EFFICIENT

4X

Inferences / Sec / TOPS
vs NVIDIA Xavier

The above is preliminary performance data based on pre-production components. Performance results are based on testing as of 11/1/2019.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. See backup for configuration details.

Comparison of Frames Per Second utilizing Resnet-50, Batch 1.

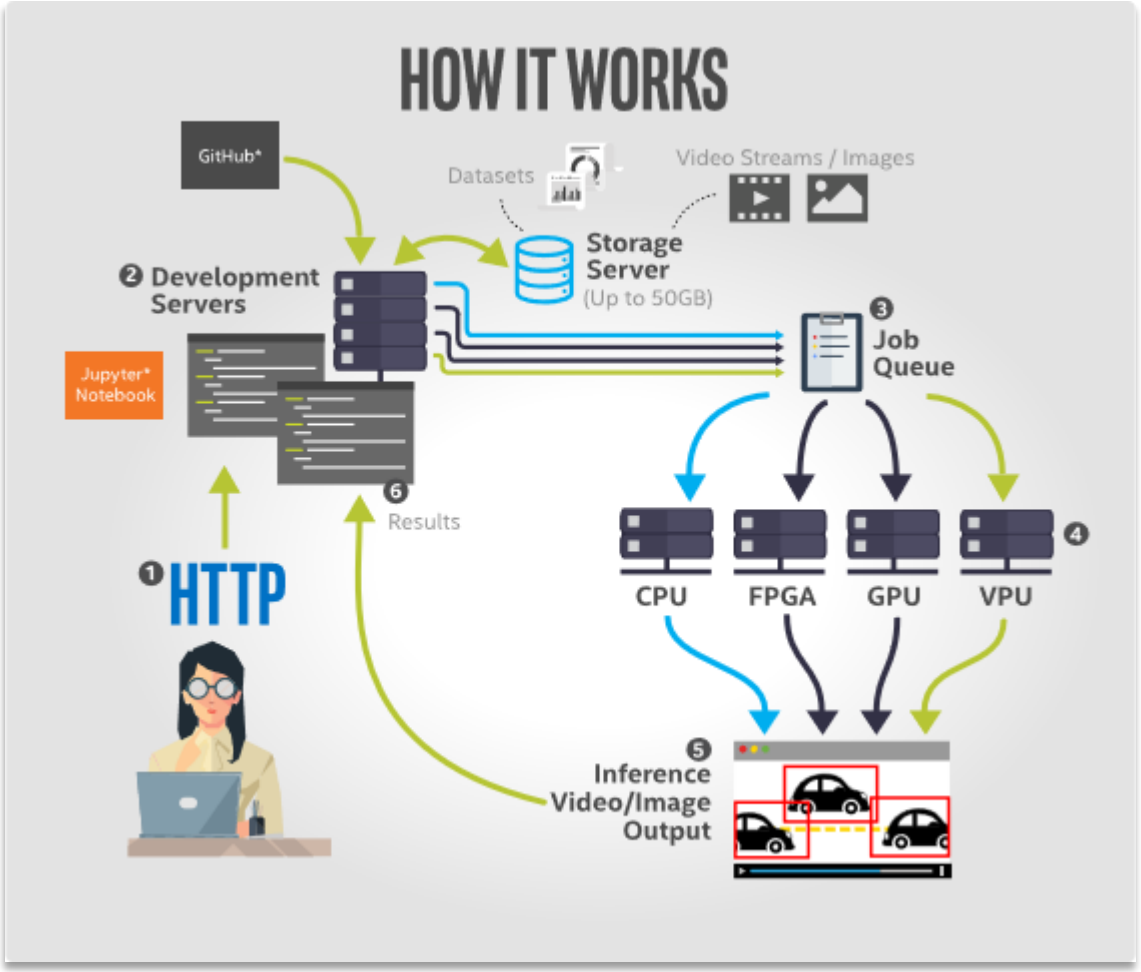
1. Gen 3 Intel® Movidius™ VPU throughput within 10% vs Xavier throughput.

intel

How Developers Can Get Productive Quickly

INTEL® DEVCLOUD FOR THE EDGE

https://devcloud.intel.com/edge



Conclusion: There's More to Comparison Than TOPS

- Simple benchmarks such as TOPS aren't effective at comparing architectures
- Better comparison is workload or application performance
- And there are many more dimensions of value in addition to performance, such as compute efficiency
- Intel's CPUs enable built-in AI inference, and VPUs offer efficiency
- Getting started is easy with Intel® DevCloud for the Edge

Thank you!

Some Useful Resources

Intel® vision and edge AI products	https://www.intel.com/visionproducts
Getting started with Intel® DevCloud	https://devcloud.intel.com/edge
OpenVINO™ toolkit documentation	https://docs.openvinotoolkit.org