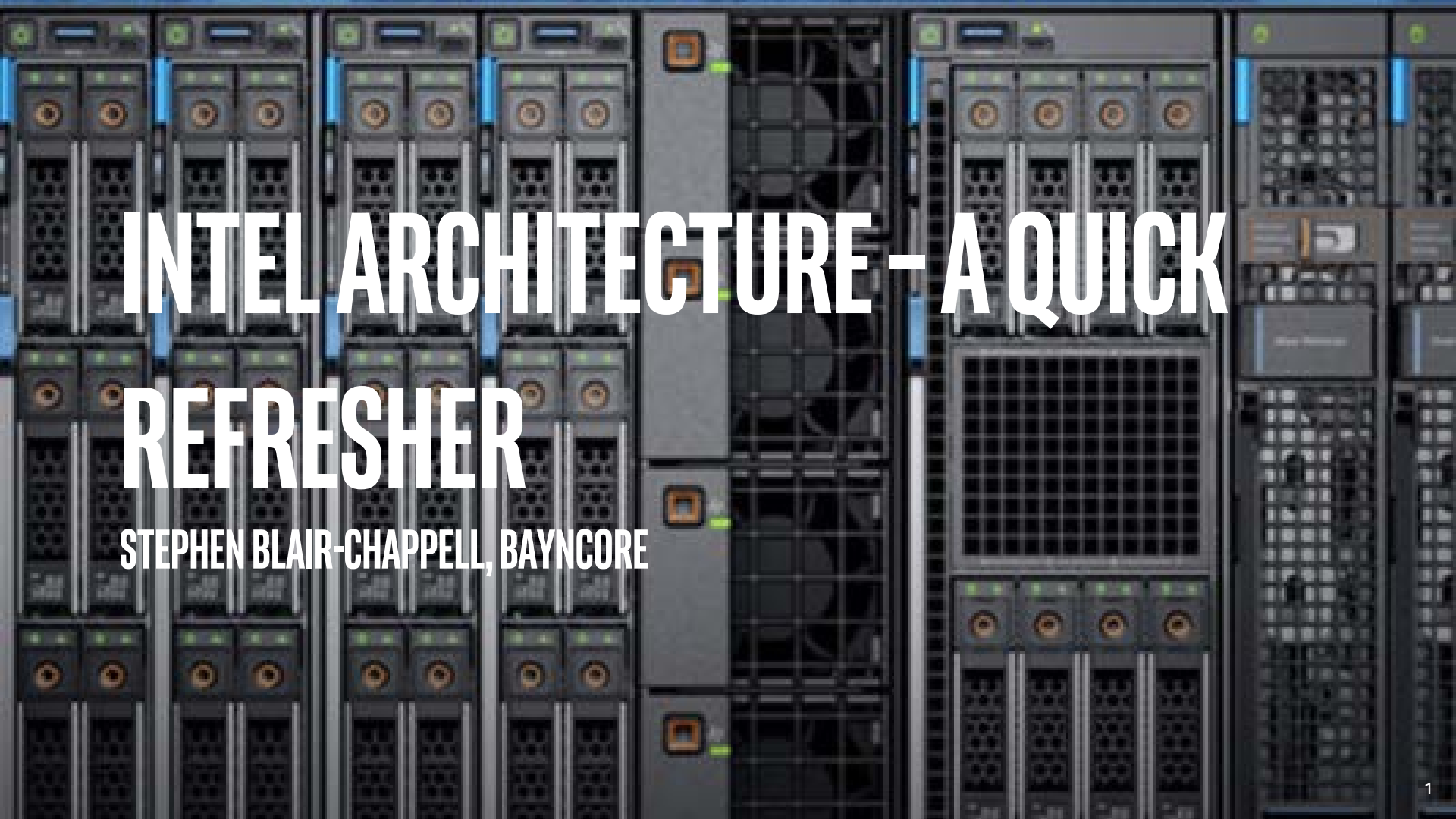


INTEL ARCHITECTURE – A QUICK REFRESHER

STEPHEN BLAIR-CHAPPELL, BAYNCORE

Three ingredients to success

Intel Silicon

1

DATACENTER

GATEWAY

EDGE

Optimised Frameworks

2

Caffe

Intel S/W & tools

3

INTEL® PARALLEL STUDIO XE

https://software.intel.com/en-us/parallel-studio-xe

intel Nervana

https://www.intelnervana.com/

Intel® VTune™ Amplifier

Visit:
www.intel.ai/technology

SPEED UP DEVELOPMENT

using open AI software

MACHINE LEARNING ← → DEEP LEARNING

TOOLKITS
App developers

ANALYTICS ZOO
Open source platform for building E2E Analytics & AI applications on Apache Spark* with distributed TensorFlow*, Keras*, BigDL

OpenVINO™
Deep learning inference deployment on CPU/GPU/FPGA/VPU for Caffe*, TensorFlow*, MXNet*, ONNX*, Kaldi*

NAUTA
Open source, scalable, and extensible distributed deep learning platform built on Kubernetes (BETA)

LIBRARIES
Data scientists

Python
- Scikit-learn
- Pandas
- NumPy

R
- Cart
- Random Forest
- e1071

Distributed
- MlLib (on Spark)
- Mahout

Intel-optimized Frameworks
TensorFlow* · Caffe2* · ONNX* · mxnet* · PyTorch* · BigDL

And more framework optimizations underway including PaddlePaddle*, Chainer*, CNTK* & others

KERNELS
Library developers

Intel® Distribution for Python*
Intel distribution optimized for machine learning

Intel® Data Analytics Acceleration Library (DAAL)
High performance machine learning & data analytics library

Intel® Math Kernel Library for Deep Neural Networks (MKL-DNN)
Open source DNN functions for CPU / integrated graphics

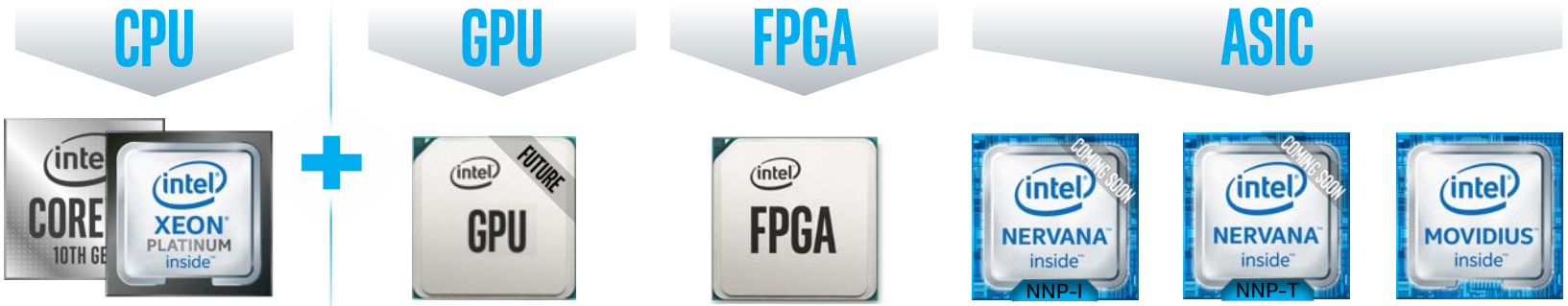
nGraph
Open source compiler for deep learning model computations optimized for multiple devices (CPU, GPU, NNP) from multiple frameworks (TF, MXNet, ONNX)

[1] An open source version is available at: 01.org/openvinotoolkit *Other names and brands may be claimed as the property of others.
Developer personas show above represent the primary user base for each row, but are not mutually-exclusive
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

DEPLOY AI ANYWHERE - INTEL® AI HARDWARE

DEVICE | INTELLIGENT EDGE | MULTI-CLOUD

OPTIMIZED FRAMEWORKS & SOFTWARE



All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
1Unified software stack development in progress DL=Deep Learning

2ND GENERATION INTEL® XEON® SCALABLE PROCESSOR

formerly known as Cascade Lake

Drop-in compatible CPU on Intel® Xeon® Scalable platform

TCO/FLEXIBILITY

Begin your AI journey efficiently, now with even more agility…

- ✓ IMT – Intel® Infrastructure Management Technologies
- ✓ ADQ – Application Device Queues
- ✓ SST – Intel® Speed Select Technology

PERFORMANCE

Built-in Acceleration with *Intel® Deep Learning Boost…*

Up to **30X** deep learning infernce throughput![1]

Throughput (img/s)

SECURITY

Hardware-Enhanced Security…

- ✓ Intel® Security Essentials
- ✓ Intel® SecL: Intel® Security Libraries for Data Center
- ✓ TDT – Intel® Threat Detection Technology

[1] Based on Intel internal testing: 1X,5.7x,14x and 30x performance improvement based on Intel® Optimization for Café ResNet-50 inference throughput performance on Intel® Xeon® Scalable Processor. See Configuration Details 3 Performance results are based on testing as of 7/11/2017(1x) ,11/8/2018 (5.7x), 2/20/2019 (14x) and 2/26/2019 (30x) and may not reflect all publically available security updates. No product can be absolutely secure. See configuration disclosure for details. ,

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. 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: http://www.intel.com/performance

INTEL® FPGA FOR AI

FIRST TO MARKET TO ACCELERATE EVOLVING AI WORKLOADS

- PRECISION
- LATENCY
- SPARSITY
- ADVERSARIAL NETWORKS
- REINFORCEMENT LEARNING
- NEUROMORPHIC COMPUTING
- …

DELIVERING AI+ FOR FLEXIBLE SYSTEM LEVEL FUNCTIONALITY

- AI+ I/O INGEST
- AI+ NETWORKING
- AI+ SECURITY
- AI+ PRE/POST PROCESSING
- …

REAL-TIME WORKLOADS

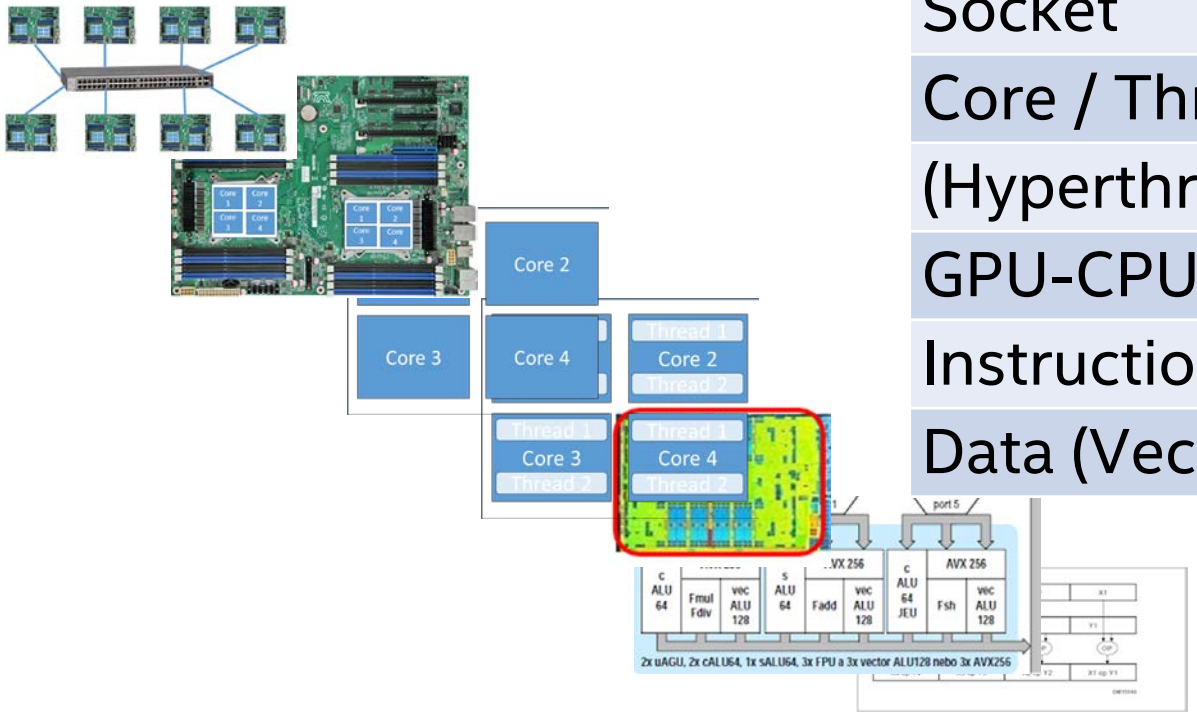
- RNN
- LSTM
- SPEECH WL

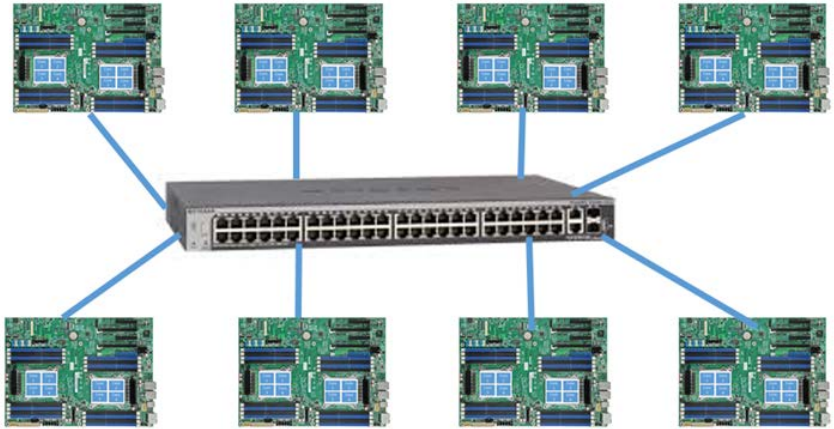
Enabling real-time AI in a wide range of embedded, edge and cloud apps

PERFORMANCE - 'IT'S ALL ABOUT PARALLELISM'

Levels of Parallelism

- Node
- Socket
- Core / Thread-Level
- (Hyperthreading)
- GPU-CPU
- Instruction (by CPU internals)
- Data (Vectorisation)





Levels of Parallelism

- Node

Socket

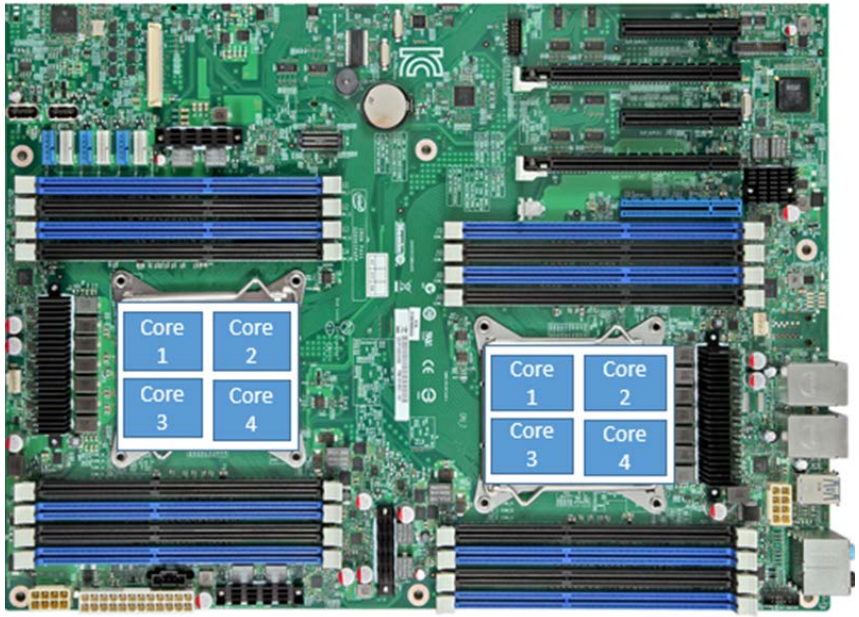
Core / Thread-Level

(Hyperthreading)

GPU-CPU

Instruction (by CPU internals)

Data (Vectorisation)



Levels of Parallelism

Node

- Socket

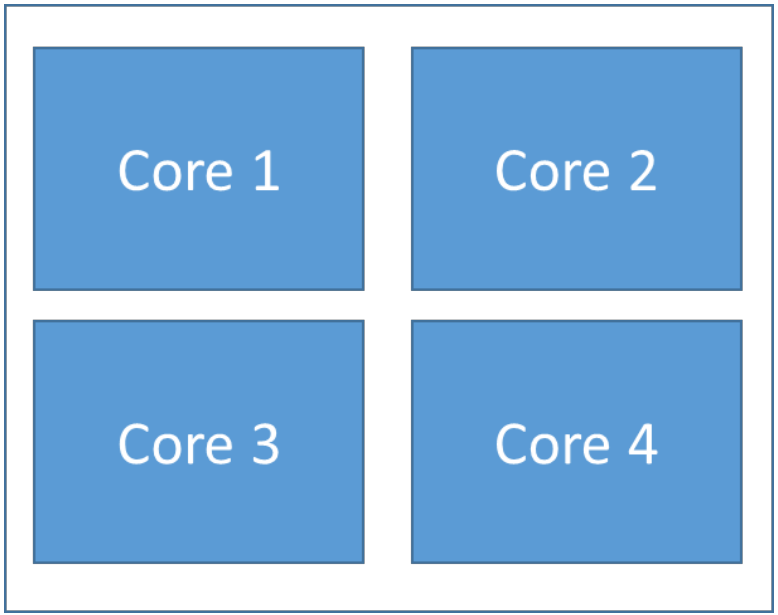
Core / Thread-Level

(Hyperthreading)

GPU-CPU

Instruction (by CPU internals)

Data (Vectorisation)



Levels of Parallelism

Node

Socket

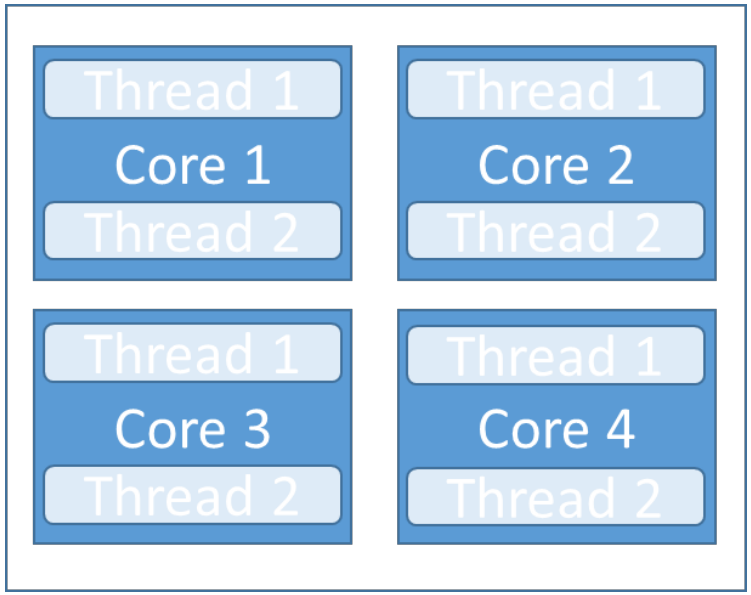
- Core / Thread-Level

(Hyperthreading)

GPU-CPU

Instruction (by CPU internals)

Data (Vectorisation)



Levels of Parallelism

Node

Socket

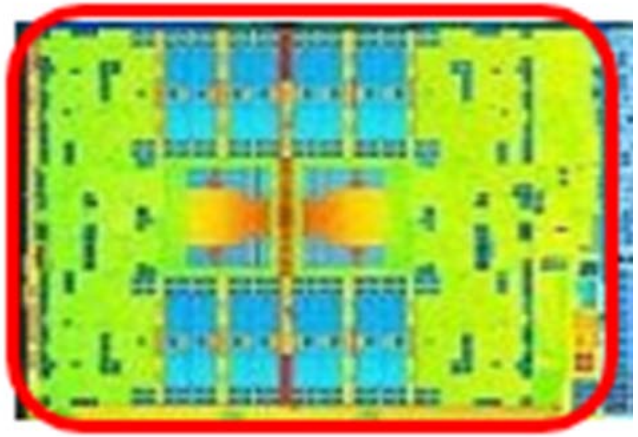
- Core / Thread-Level

(Hyperthreading)

GPU-CPU

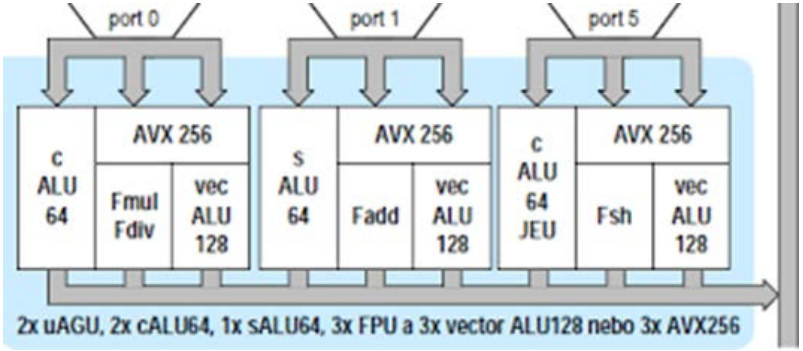
Instruction (by CPU internals)

Data (Vectorisation)



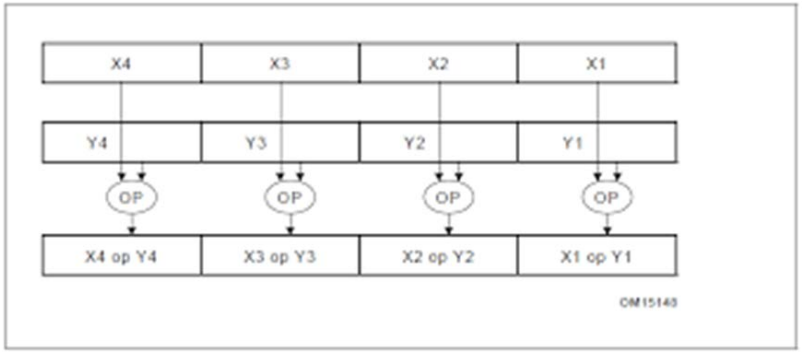
Levels of Parallelism

- Node
- Socket
- Core / Thread-Level
- (Hyperthreading)
- GPU-CPU
- Instruction (by CPU internals)
- Data (Vectorisation)



Levels of Parallelism

- Node
- Socket
- Core / Thread-Level
- (Hyperthreading)
- GPU-CPU
- Instruction (by CPU internals)
- Data (Vectorisation)



SSE2 128 bit
AVX 256 bit
AVX512 512 bit

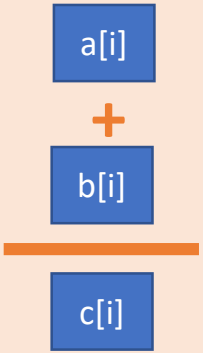
Levels of Parallelism

- Node
- Socket
- Core / Thread-Level
- (Hyperthreading)
- GPU-CPU
- Instruction (by CPU internals)
- Data (Vectorisation)

What is vectorization ?

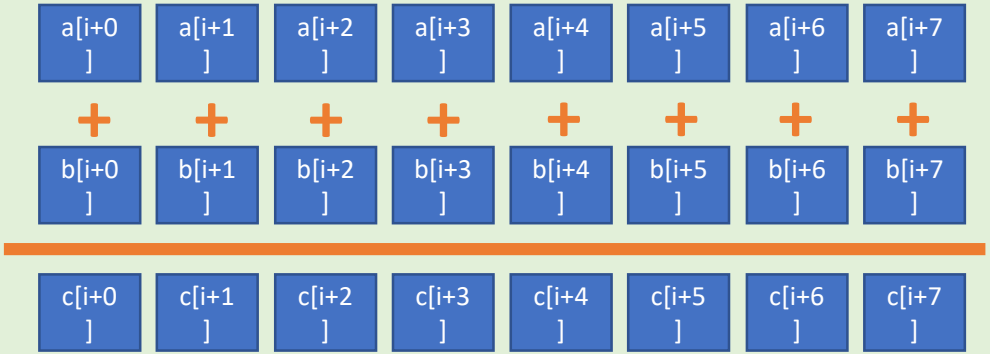
What you write

```
for(i = 0; i <= MAX; i++)
   c[i] = a[i] + b[i];
```



What the compiler might generate

```
for(i = 0; i <= MAX; i+8)
   c[i:8] = a[i:8] + b[i:8];
```



INTEL® DEEP LEARNING BOOST (DL BOOST)

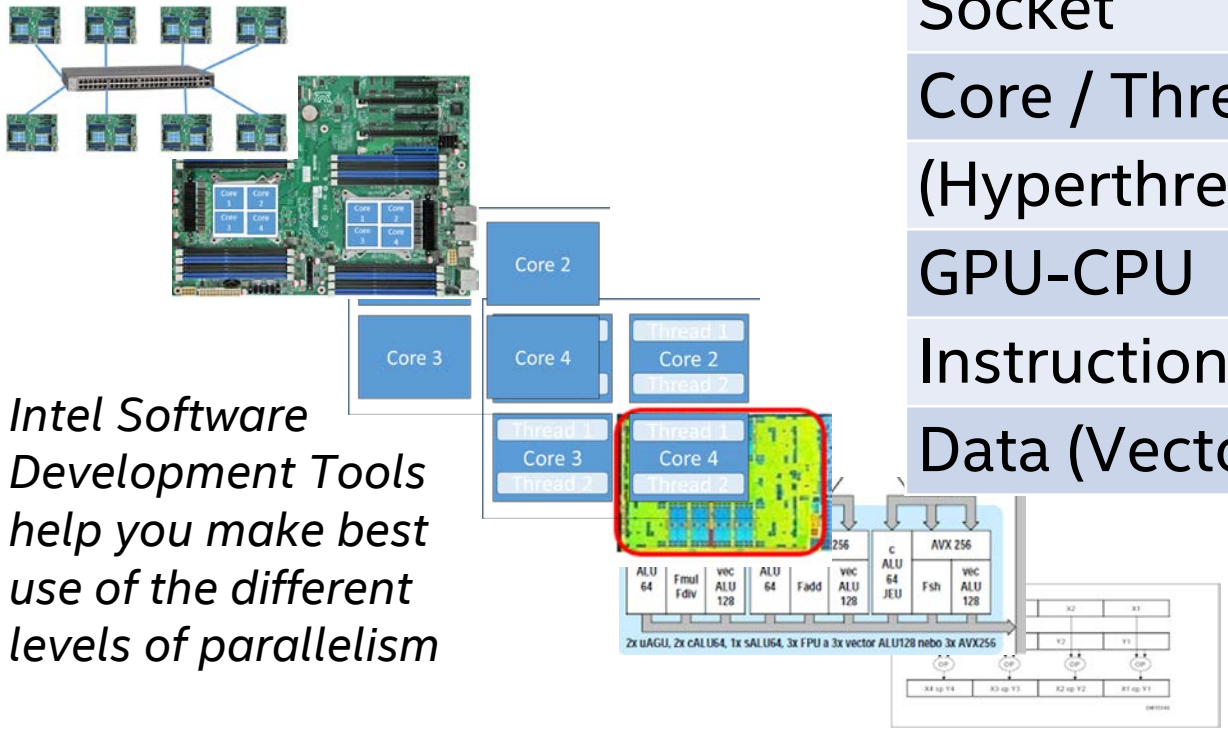
FEATURING VECTOR NEURAL NETWORK INSTRUCTIONS (VNNI)

Current AVX-512 instructions to perform INT8 convolutions: **vpmaddubsw, vpmaddwd, vpaddd**

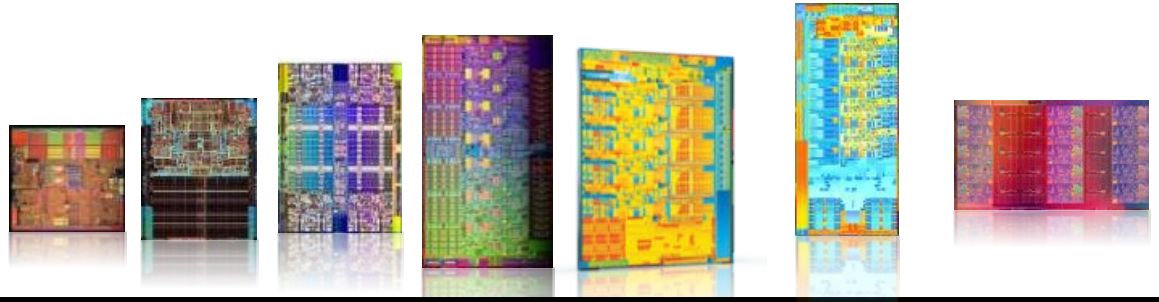
NEW AVX-512 (VNNI) instruction to accelerate INT8 convolutions: **vpdpbusd**[**]

Levels of Parallelism

- Node
- Socket
- Core / Thread-Level
- (Hyperthreading)
- GPU-CPU
- Instruction (by CPU internals)
- Data (Vectorisation)



Intel Software Development Tools help you make best use of the different levels of parallelism

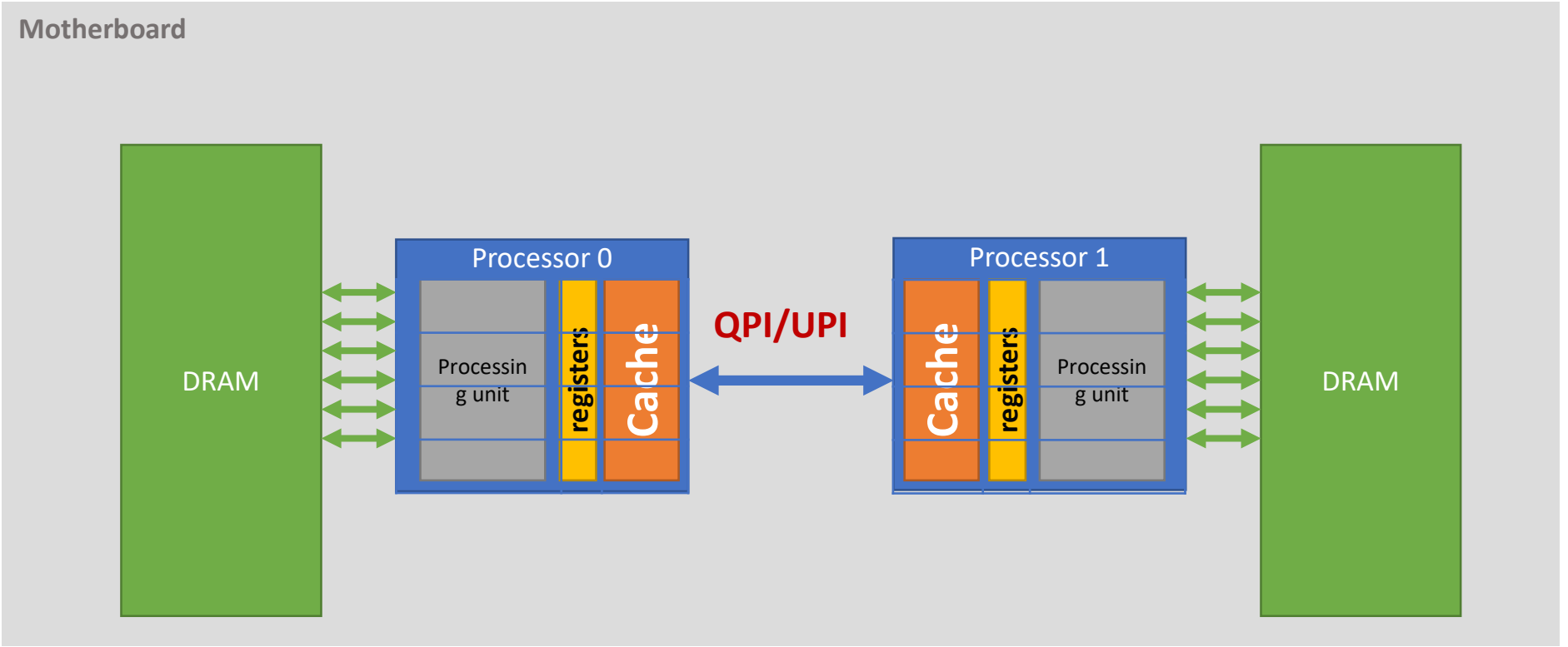


	Intel® Xeon® processor 64-bit	Intel® Xeon® processor 5100 series	Intel® Xeon® processor 5500 series	Intel® Xeon® processor 5600 series	Intel® Xeon® processor code-named Sandy Bridge EP	Intel® Xeon® processor code-named Ivy Bridge EP	Intel® Xeon® processor code-named Skylake EP	Intel® Xeon® processor code-named Cascade Lake Platinum 9200
Core(s)	1	2	4	6	8	12	28	56
Threads	2	2	8	12	16	24	56	112
SIMD Width	128	128	128	128	256	256	512	512

YET ANOTHER VIEW . . .

PERFORMANCE - ‘IT’S ALL ABOUT MEMORY’

What does a 2 sockets system looks like ?



Memory Hierarchy

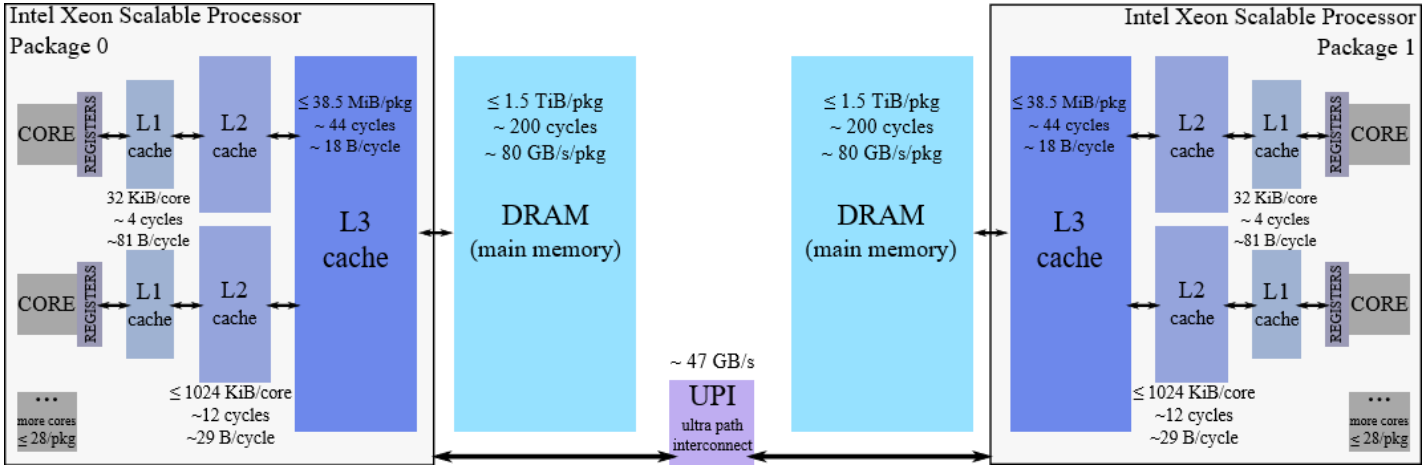
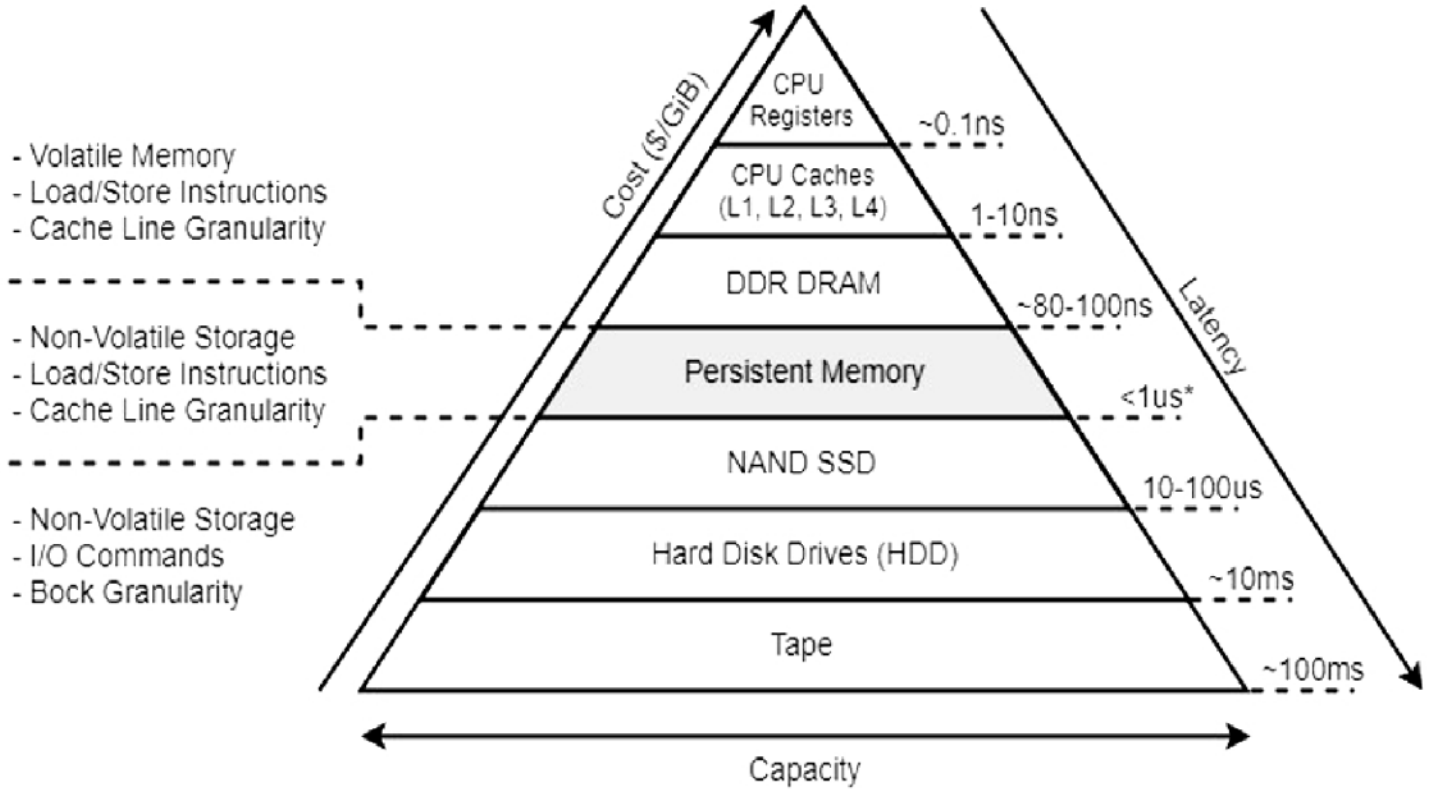
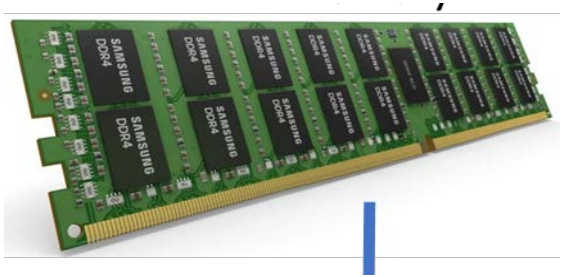


Diagram Source: colfax.com



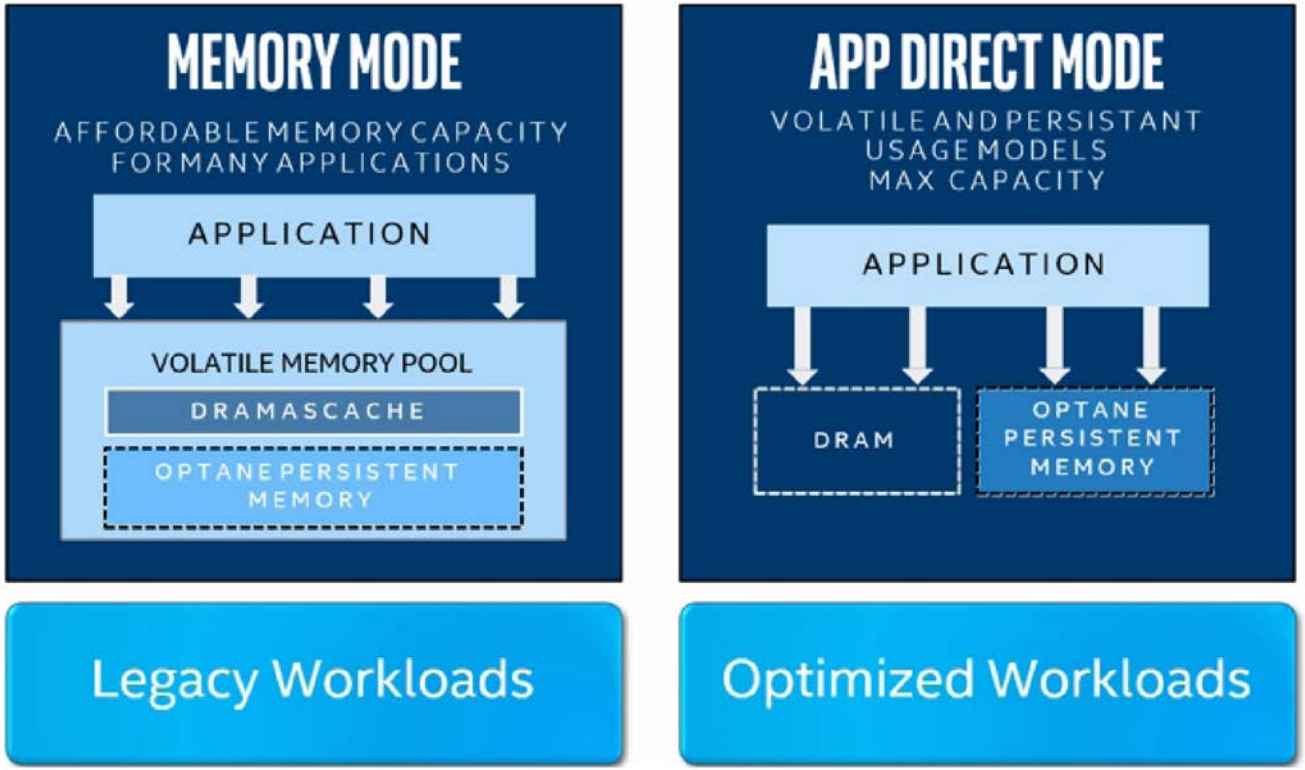
1 **Latency estimates for different storage and memory devices**

Intel® Optane™ DC Persistent Memory



- non-volatile, high-capacity memory
- near DRAM latency,
- affordable
- physically and electrically compatible with DDR4 interfaces and slots

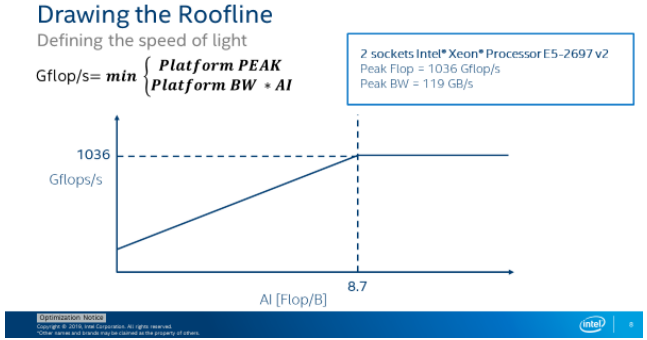
Intel® Optane™ DC Persistent Memory



Performance : A summary

- Product of **CPU** Parallelism AND **Memory**
  - See Advisor Roofline Model which combines
    - Peak Flops
    - Peak Bandwidth

  https://software.intel.com/en-us/advisor



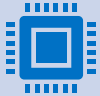
  - See Performance Optimisation and Productivity Project which combines
    - Global Efficiency,
    - Parallel Efficiency ,
    - Computational Efficiency

  https://software.intel.com/en-us/download/parallel-universe-magazine-issue-37-july-2019

Some factors in deciding ‘*What plaform/architecture should I use?*’

Factor
Cost
Performance
Accuracy
Power
Ease of Programming
Portability

Summary



Intel CPU offers multiple levels of parallelism

To get best performance you need to use these levels in your applications

Intel Libraries and Optimised Frameworks provide these ‘automatically’