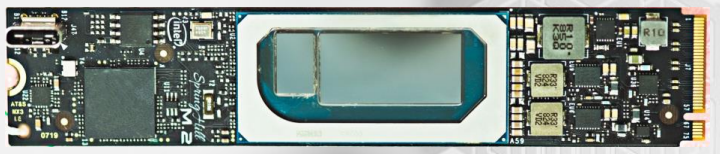
SPRING HILL (NNP-I 1000)
INTEL'S DATA CENTER INFERENCE CHIP

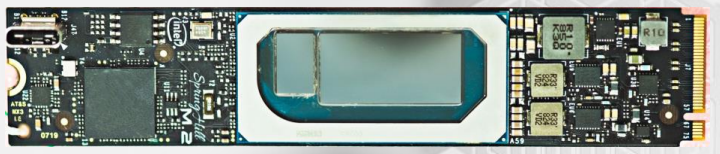
Ofri Wechsler, Michael Behar, Bharat Daga | 8/20/19

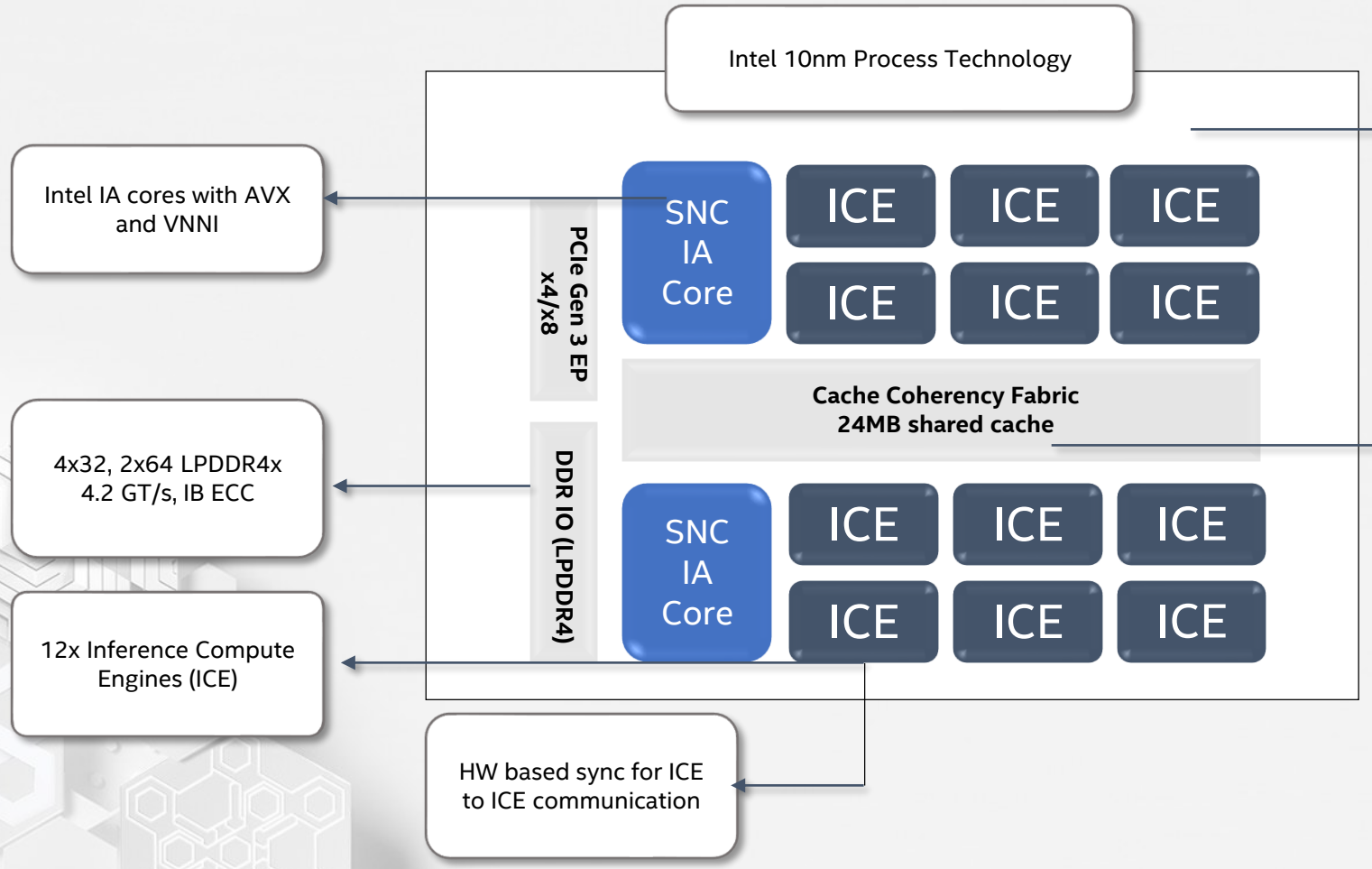
SPRING HILL (NNP-I 1000) – DESIGN GOALS



- Best in class perf/power efficiency for major data center inference workloads
  - 4.8 TOPs/W
- 5X power scaling for performance boost
  - 10-50W
- Achieve high degree of programmability w/o compromising perf/power efficiency
  - Drive AI innovation with on die Intel Architecture cores
- Data Center at Scale
  - Comprehensive set of RAS features to allow seamless deployment in existing data centers
- Highly capable SW stack supporting all major DL frameworks

INTRODUCING – SPRING HILL (NNP-I)

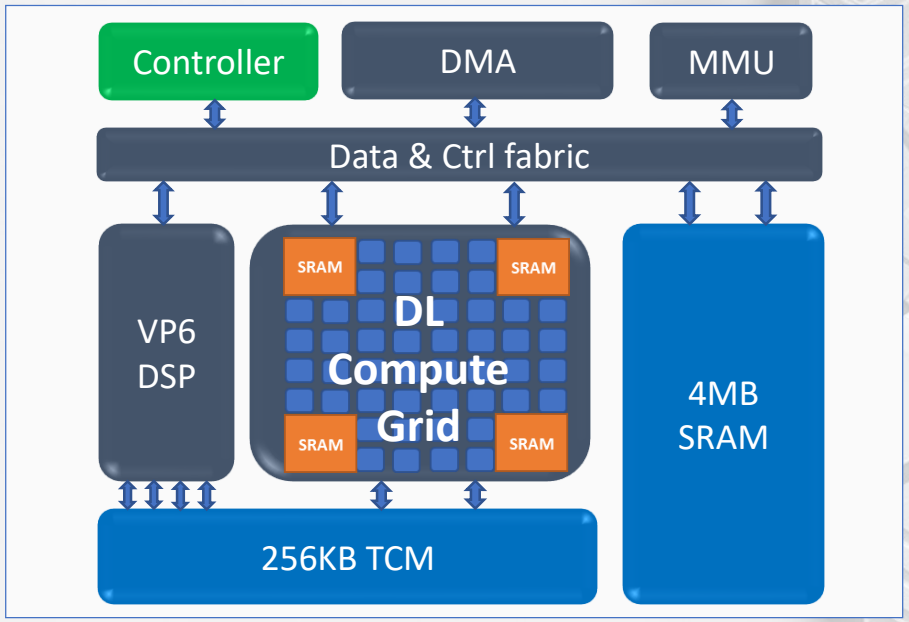




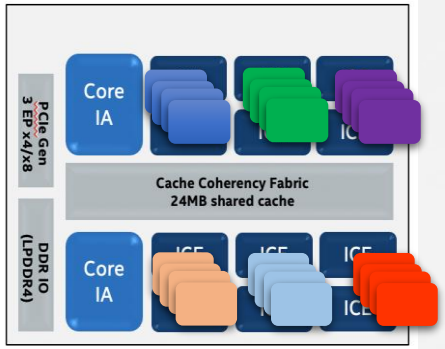
Feature	SoC
TOPs (INT8)	48 - 92
TDP	10-50w
TOPs/w	2.0-4.8 TOPs/w
Inference Engines (ICE)	10-12 ICEs
Total SRAM	75MB
DRAM BW	68 GB/s

ICE - INFERENCE COMPUTE ENGINE

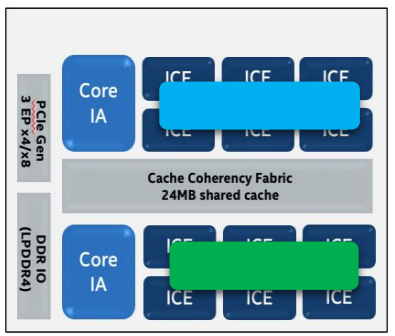
- Deep Learning compute grid
  - 4K MAC (int8) per cycle
  - Scalable support: FP16, INT8, INT 4/2/1
  - Large internal SRAMs for power efficiency
  - Non-linear ops & Pooling
- Programmable vector processor
  - High throughput: 5 VLIW 512b
  - Extended NN support (FP16/16b/8b)
- High BW data memory access
  - Dedicated DMA – optimized for DL
  - Compression/decompression unit- support for sparse weights
- Large Local SRAM
  - Persistent data and/or storage of temporal data



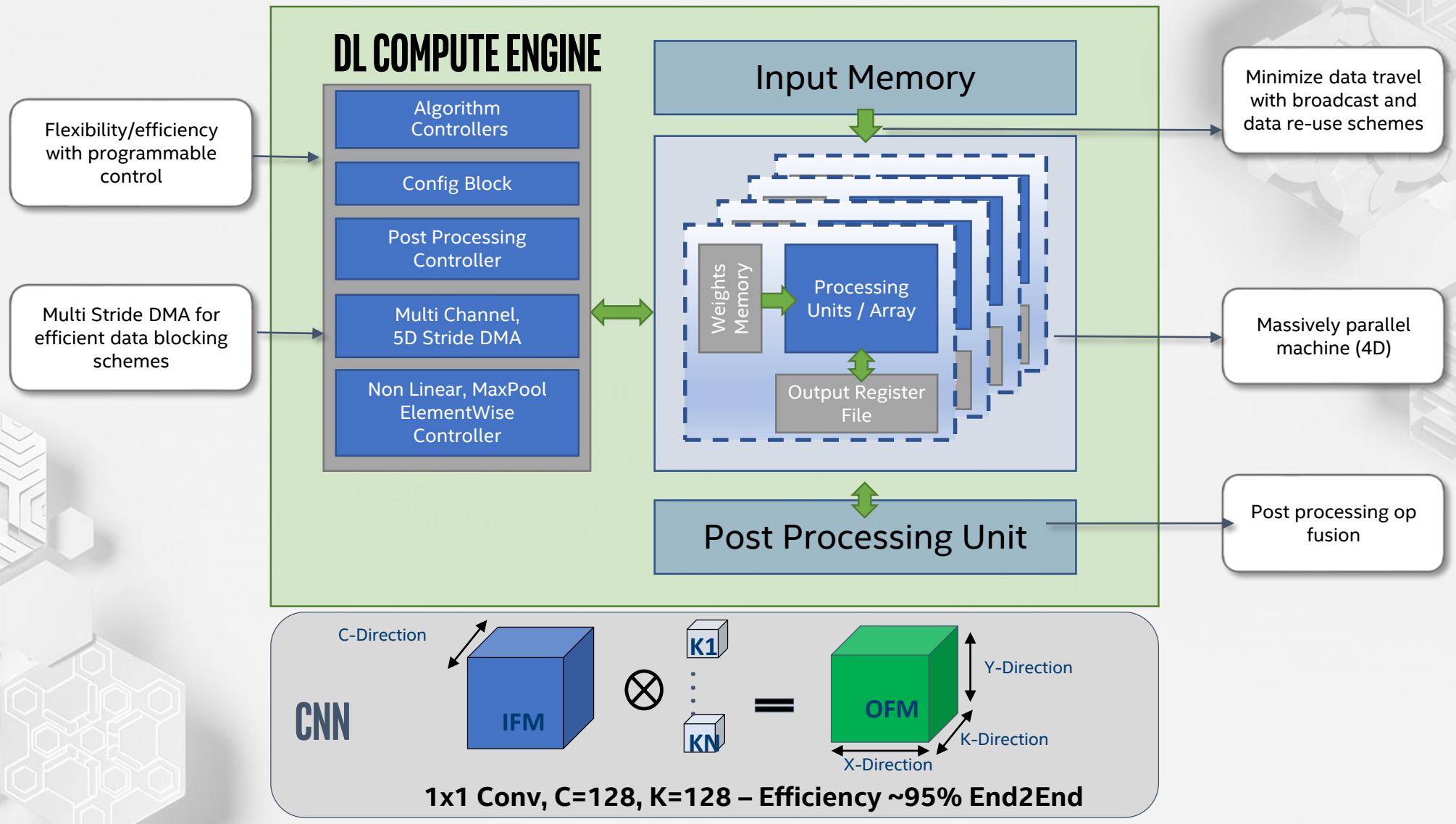
Optimized for throughput
batch: 4x6 (or 1X12)



Optimized for latency
batch: 1x2

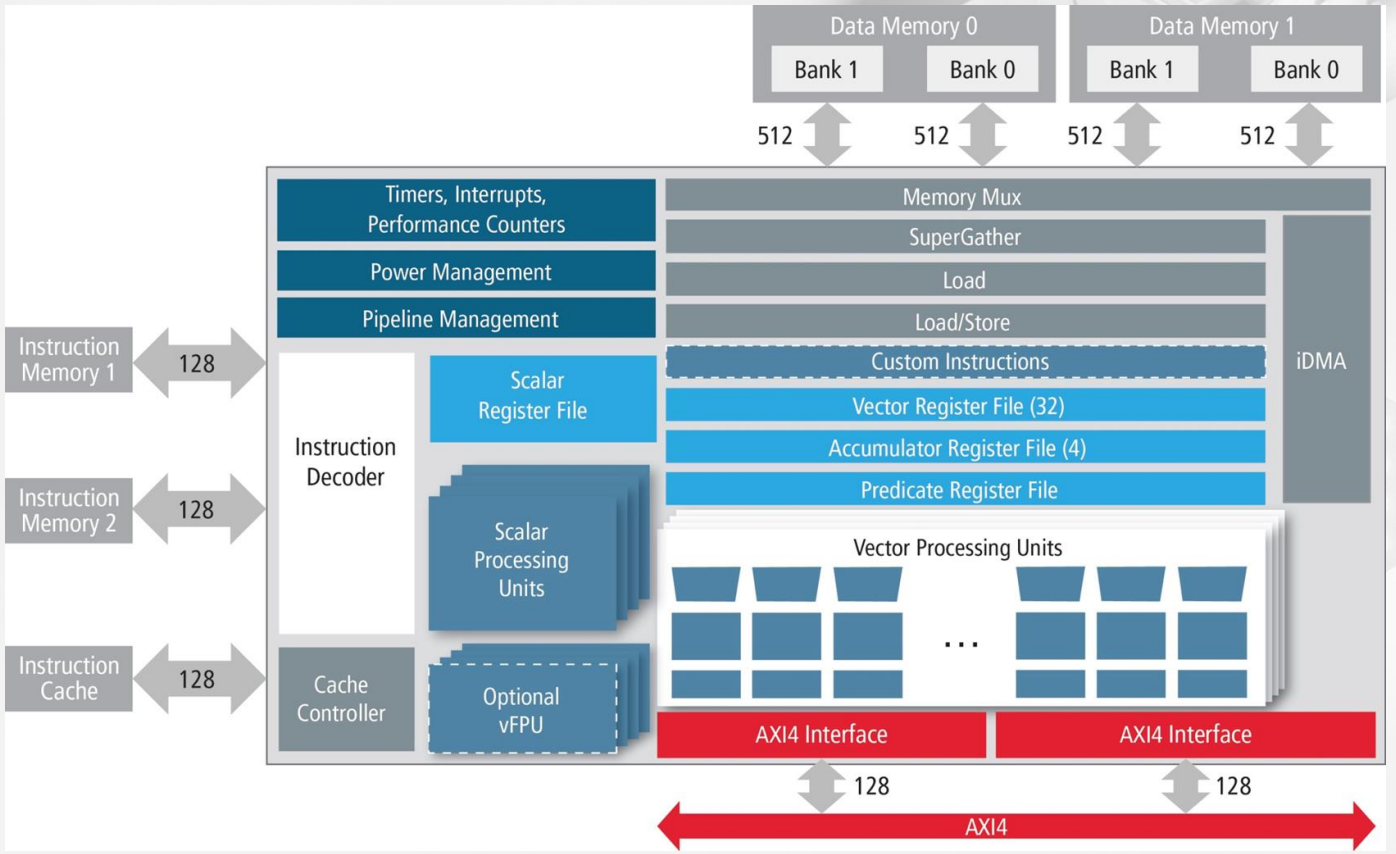


DL COMPUTE GRID

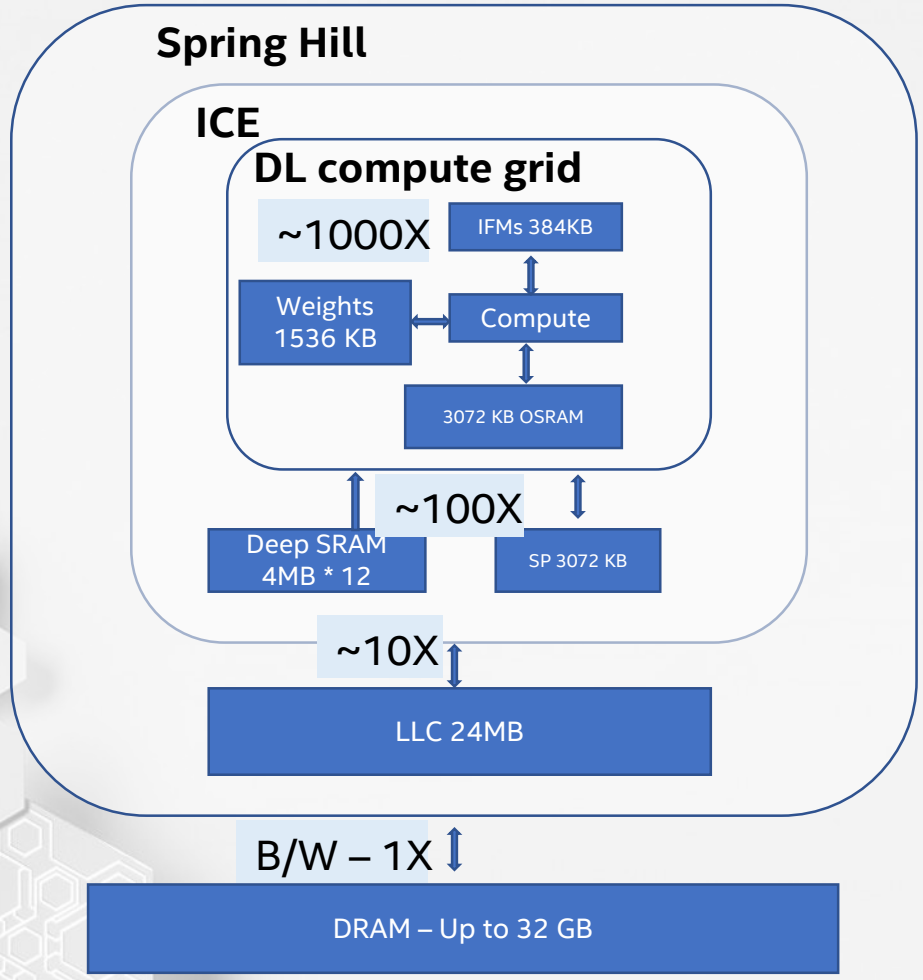


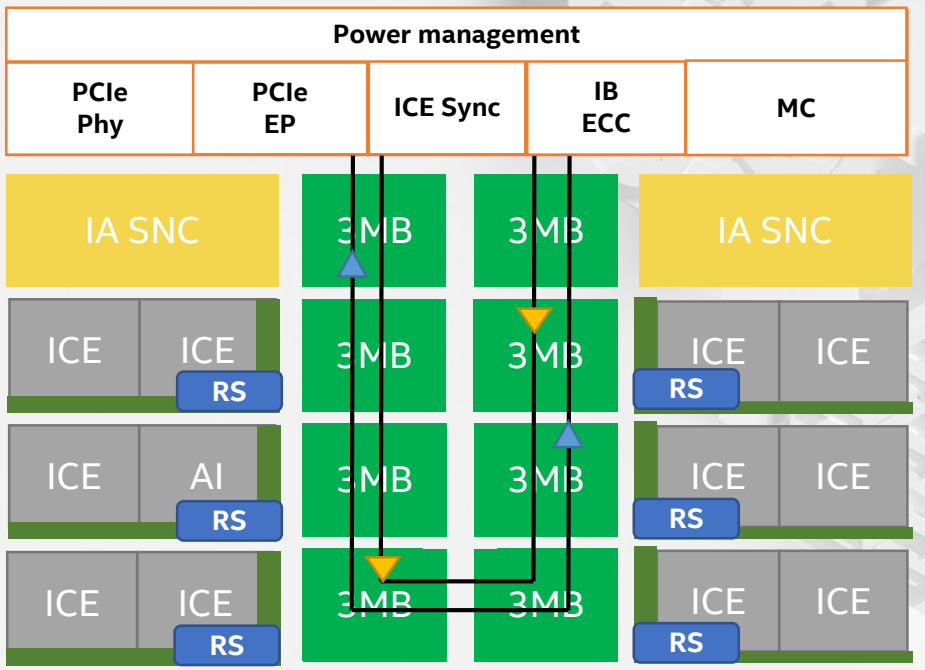
THE VECTOR PROCESSING ENGINE

- Tensilica Vision P6 DSP
  - 5 VLIW, 512b vector
  - 2 Vector Load ports
  - Scatter/gather engine
  - Data type: Int8,16,32, FP16
- Fully programmable
- Full bi-directional pipeline with the DL compute Grid
  - Shared local memory
  - HW Sync. between producer and consumer



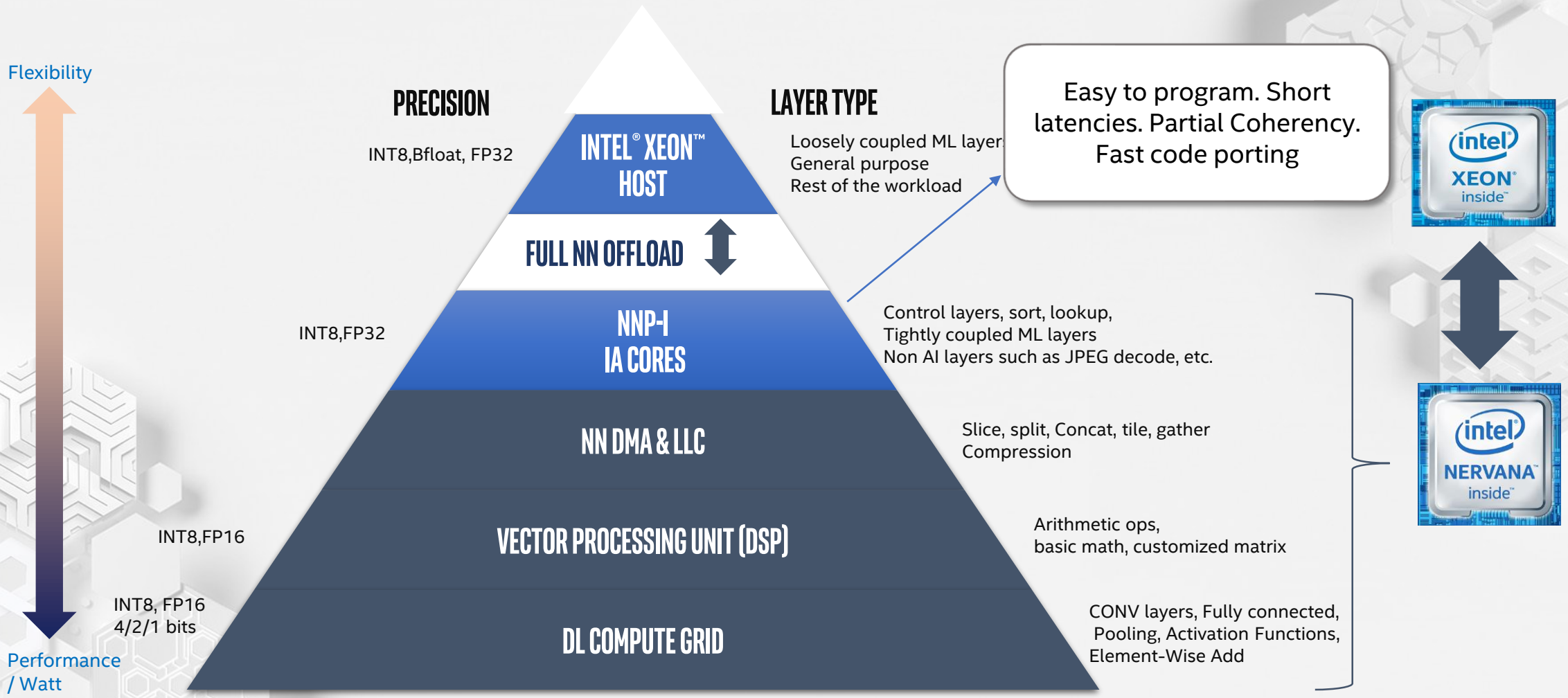
MEMORY ORGANIZATION





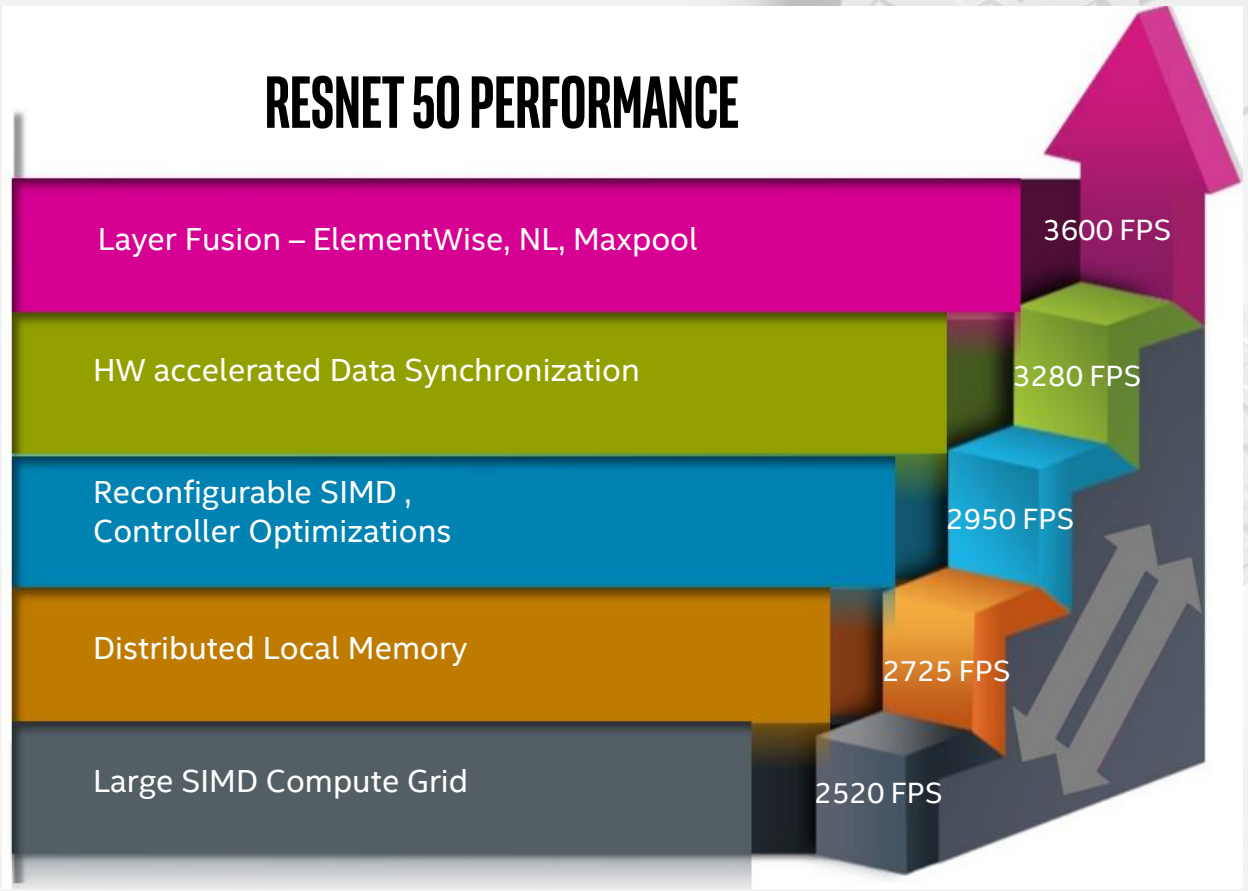
- 24MB LLC
  - Organized in 8X3MB slices
  - Shared data (ICE-to-ICE, ICE-to-IA)
  - Optimized for DL inference workloads

PROGRAMMING FLEXIBILITY AND WORKLOAD DECOMPOSITION

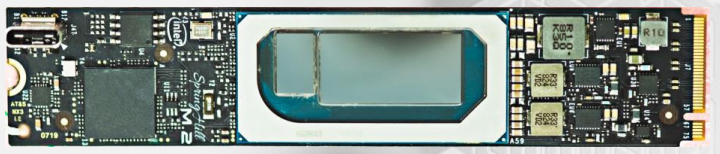


SPH (NNP-I 1000) PERFORMANCE

- **ResNet50**
  - 3600 Inferences Per Second (IPS)
  - @10W (SOC level)
  - 360 Images/Sec/W
  - 2→12 AI cores 5.85x
- **Performance/W – 4.8 TOPs/W**
- **Additional preliminary performance disclosure with MLPerf 0.5 Submission**



SUMMARY



- Best in class perf/power efficiency for major data center inference workloads
- Scalable performance at wide power range
- Achieve high degree of programmability w/o compromising perf/power efficiency
- Data Center at Scale
- Spring hill solution -- Silicon and SW stack – sampling with definitional partners/customers on multiple real-life topologies
- Next 2 generations in planning/design

LEGAL NOTICES & DISCLAIMERS

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

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 The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

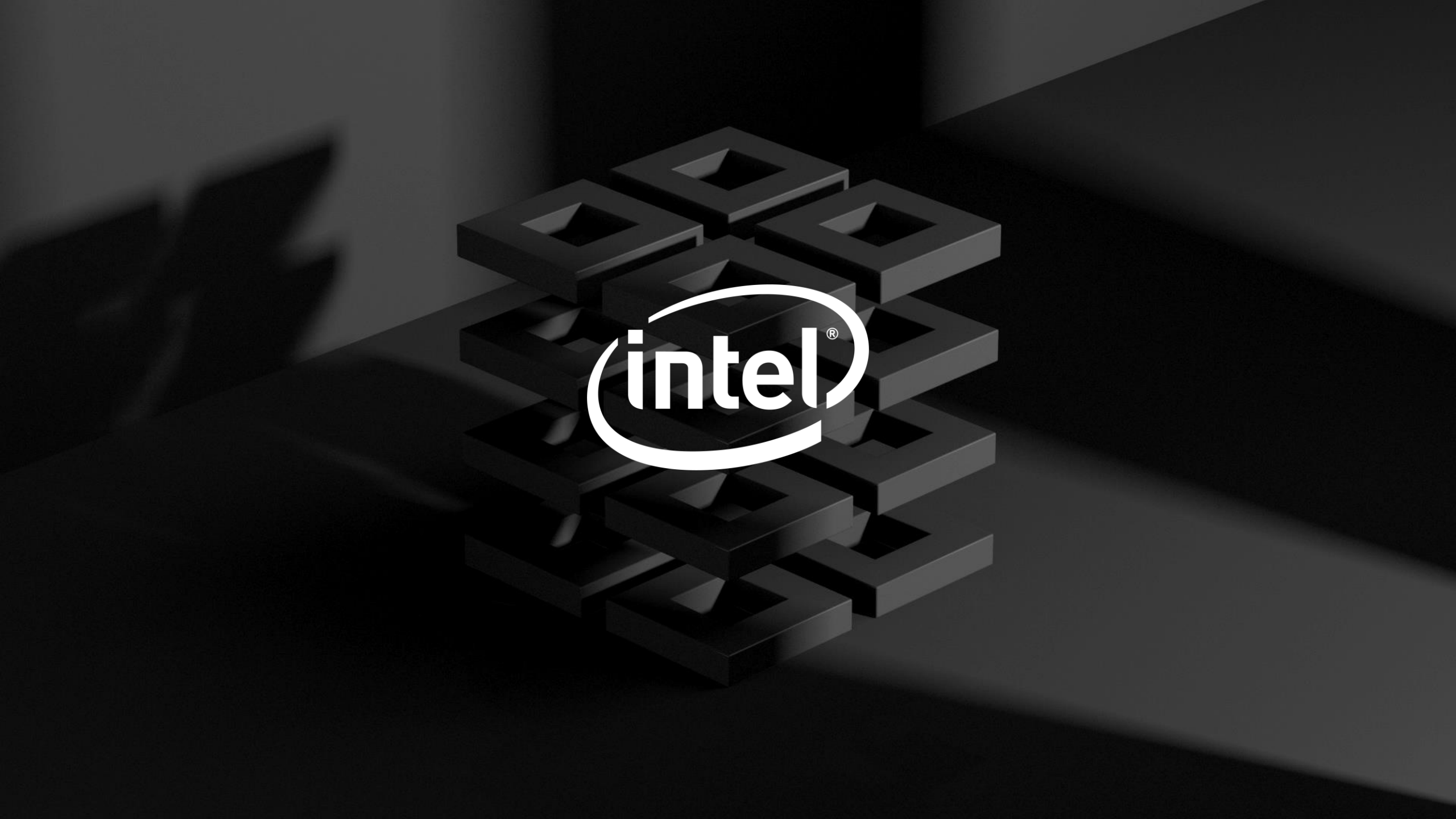
No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

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. Notice Revision #20110804.

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel, the Intel logo, Intel Inside, Nervana, and others are trademarks of Intel Corporation in the U.S. and/or other countries. Other names and brands may be claimed as the property of others.


intel®