Goals

- 500X INCREASE IN COMPUTE PERFORMANCE
- SCALABLE COMPUTE & MEMORY
- PACKAGING & INTERCONNECT FOR DENSITY & SCALE
- FULL SOFTWARE STACK/PROGRAMMING MODEL
Compute Efficiency
High Performance Graphics
Scalability
Compute Density
Building Blocks

Core
Slice
Stack
Link
Xe-core

Compute Building Block of Xe HPC-based GPUs

- 8 Vector Engines
- 8 Matrix Engines
- Load/Store: 512 B/CLK
- Cache: L1$/SLM (512 KB), L1$
Up to

4 Slices
64 X^e-cores
64 Ray Tracing Units
4 Hardware Contexts
L2 Cache
4 HBM2e controllers
1 Media Engine
8 X^e Links
2 - Stack

8 Slices
- 128 Xe-cores
- 128 Ray Tracing Units
- 8 Hardware Contexts

2 Media Engines

8 HBM2e controllers

16 Xe Links
High Speed Coherent Unified Fabric (GPU to GPU)

Load/Store, Bulk Data Transfer & Sync Semantics

Up to 8 Fully Connected GPUs through Embedded Switch
Link for Scalability
8x System Compute Rates

**Vector**
- 8x Up to 32,768 FP64 Ops/CLK
- 8x Up to 32,768 FP32 Ops/CLK

**Matrix**
- 8x Up to 262,144 TF32 Ops/CLK
- 8x Up to 524,288 BF16 Ops/CLK
- 8x Up to 1,048,576 INT8 Ops/CLK
Ponte Vecchio
Ponte Vecchio

- New Verification Methodology
- New Software
- New Reliability Methodology
- New Signal Integrity Techniques
- New Interconnects
- New Power Delivery Technology
- New Packaging Technology
- New I/O Architecture
- New Memory Architecture
- New IP Architecture
- New SOC Architecture
Ponte Vecchio soc

>100 Billion Transistors
47 Active Tiles
5 Process Nodes
Ponte Vecchio

Key Challenges

- Scale of Integration
- Foveros Implementation
- Verification Tools & Methods
- Signal Integrity, Reliability & Power Delivery

Tiles:
- Compute Tile
- Rambo Tile
- Foveros
- Base Tile
- HBM Tile
- Xe Link Tile
- Multi Tile Package
- EMIB Tile
Ponte Vecchio

Compute Tiles

- Per Tile: 8 Xe-cores
- Built on: TSMC N5
- L1 Cache: 4MB Per Tile
- Bump Pitch: 36um Foveros
Ponte Vecchio

Base Tile

- Built on Intel 7 FOVEROS
- Area 640mm²
- L2 Cache 144MB
- Host Interface PCIe Gen5
- HBM2e
- MDFI
- EMIB
Ponte Vecchio

**Xe Link Tile**

- **Per Tile**: 8 Xe Links
- **Built on**: TSMC N7
- **8 ports**: Embedded Switch
- **Up to**: 90G Serdes
Accelerated Compute Systems

Ponte Vecchio x4 Subsystem
with Xe Links

+ 2S Sapphire Rapids
Software

ISA

Hardware

Raja Koduri, Intel - "No Transistor Left Behind" Hot Chips 2020 Keynote
oneAPI
Open, Standards-Based Unified Software Stack

Freedom from proprietary programming models
Full performance from the hardware
Piece of mind for developers

CPU & XPU - Optimized Stack

Applications & Services

Middleware, Frameworks & Runtimes

Low-level Libraries

Languages

Hardware Abstraction Layer

Compute Hardware

oneMKL oneDNN oneDAL oneVPL

oneTBB oneCCL oneDPL Other Libraries

DPC++ Other Languages

Level Zero

CPU GPU
Ponte Vecchio
Execution Progress

A0 Silicon Current Status

> 45 TFLOPS  FP32 Throughput

> 5 TBps  Memory Fabric Bandwidth

> 2 TBps  Connectivity Bandwidth
Ponte Vecchio Execution Progress

A0 Silicon Current Status

> 43,000 img/sec  Inference

> 3,400 img/sec  Training

Resnet 50 performance
Ponte Vecchio

The vision 2 years ago...

- Leadership Performance for HPC/AI
- Connectivity to drive scaleup and scale out
- Unified Programming Model powered with oneAPI
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.

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.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. For testing details and system configurations, please contact your intel representative. No product or component can be absolutely secure.

Results that are based on pre-production systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications, or configurations. Intel technologies may require enabled hardware, software or service activation.

Intel contributes to the development of benchmarks by participating in, sponsoring, and/or contributing technical support to various benchmarking groups, including the BenchmarkXPRT Development Community administered by Principled Technologies.

Statements in this presentation that refer to future plans and expectations are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “goals,” “plans,” “believes,” “seeks,” “estimates,” “continues,” “may,” “will,” “would,” “should,” “could,” and variations of such words and similar expressions are intended to identify such forward-looking statements. Statements that refer to or are based on estimates, forecasts, projections, uncertain events or assumptions, including statements relating to future products and technology and the expected availability and benefits of such products and technology, market opportunity, and anticipated trends in our businesses or the markets relevant to them, also identify forward-looking statements. Such statements are based on management's current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in these forward-looking statements. Important factors that could cause actual results to differ materially from the company's expectations are set forth in Intel's earnings release dated July 23, 2020, which is included as an exhibit to Intel's Form 8-K furnished to the SEC on such date, and Intel's SEC filings, including the company's most recent reports on Forms 10-K and 10-Q. Copies of Intel's Form 10-K, 10-Q and 8-K reports may be obtained by visiting our Investor Relations website at www.intc.com or the SEC's website at www.sec.gov. Intel does not undertake, and expressly disclaims any duty, to update any statement made in this presentation, whether as a result of new information, new developments or otherwise, except to the extent that disclosure may be required by law.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.
Thank you!

“Something is going to happen.”
“What is going to happen?”
“Something __________.”

bit.ly/2VEW6Dt