

Intel GPU Optimization Guide

Work-Group Mapping and GPU Occupancy Calculation

Learn about Intel GPU Occupancy Calculation

rakshith.Krishnappa@intel.com



intel[®]

Agenda

- Intel Data Center GPU MAX Series Architecture
- Access Intel GPUs using Intel Developer Cloud
- Intel GPU Optimization using SYCL
 - Mapping SYCL Work-Groups to Intel GPU
 - Intel GPU Occupancy Calculator
 - SYCL Kernel Launching and Profiling

Learning Objectives

- Access **Intel Data Center GPU MAX** using Intel Developer Cloud
- Explain how SYCL Work-Groups map to Intel GPU hardware
- Use **Intel GPU Occupancy Calculator** to choose work-group and sub-group sizes to maximize GPU occupancy
- **Enable profiling** for kernel execution

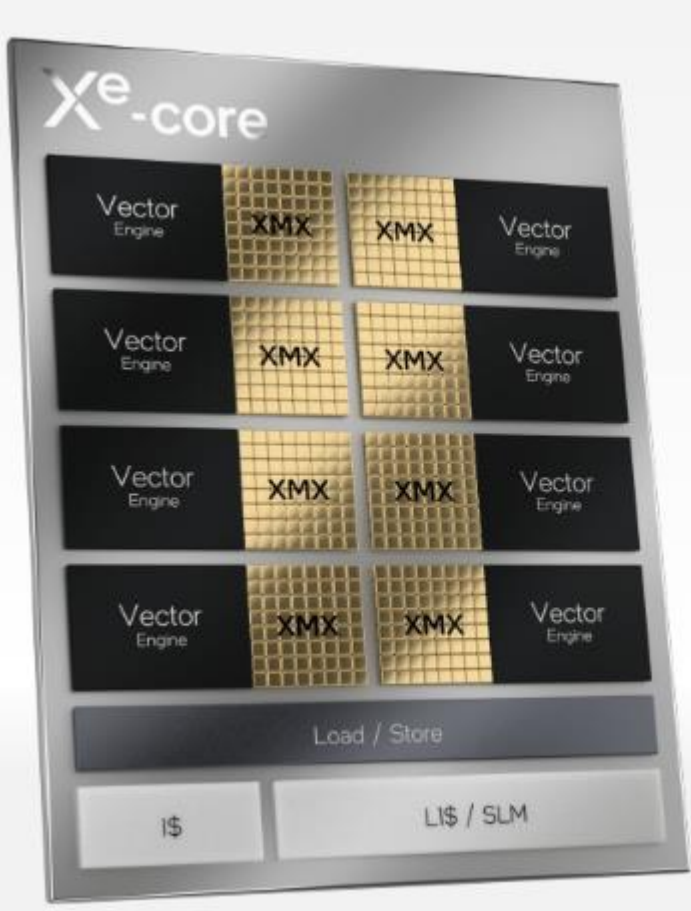
Intel Data Center GPU MAX Series

Intel's highest performing, highest density, general-purpose discrete GPU, which packs over 100 billion transistors into one package



Xe Core

Building block of GPU with 8 vector engines, 8 matrix engines, SLM/L1 Cache

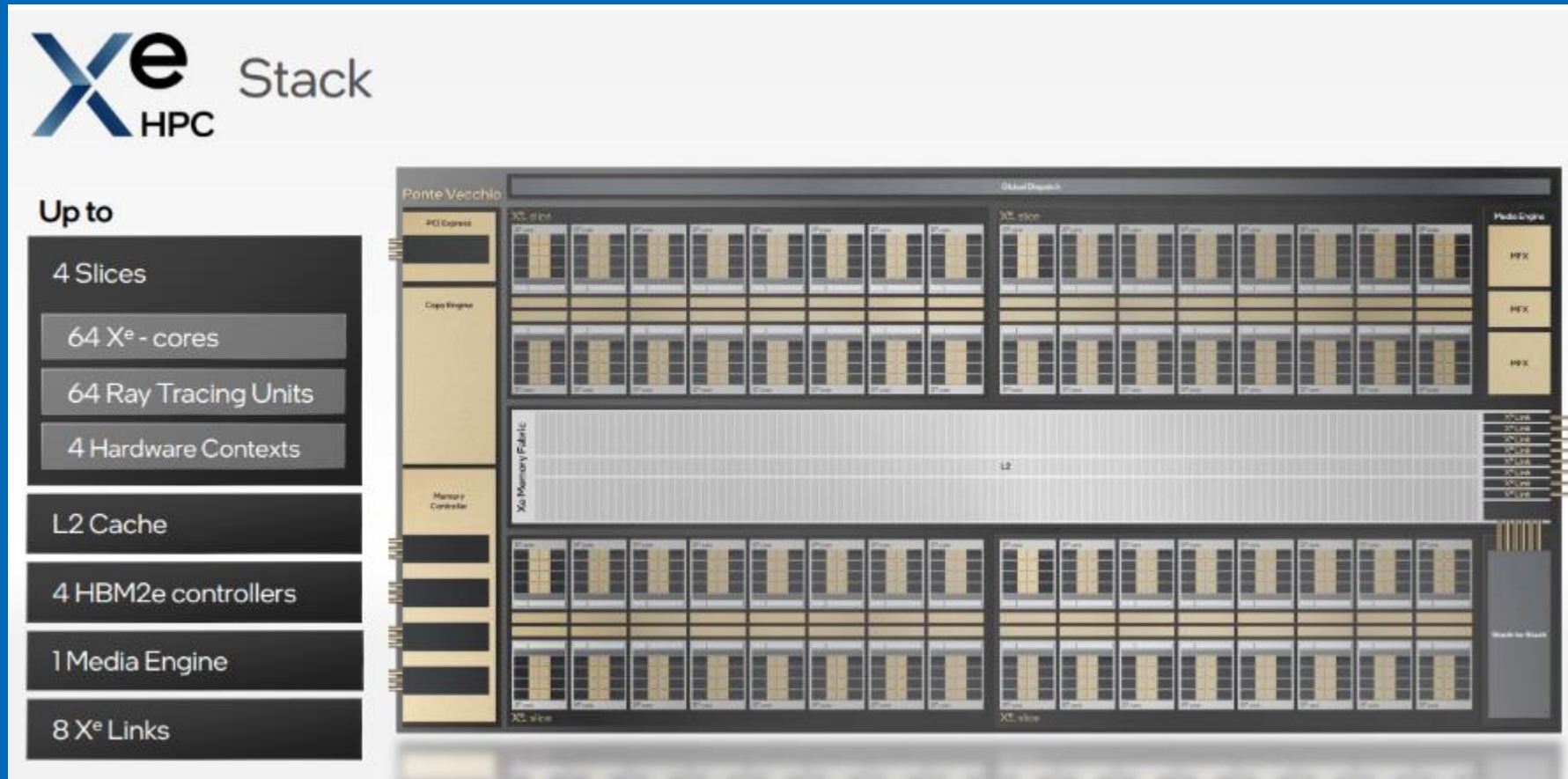


The image shows a 3D rendering of a Xe-core GPU chip on the left. The chip is labeled 'Xe-core' and features a grid of components: four 'Vector Engine' blocks, four 'XMX' (Matrix Engine) blocks, a 'Load / Store' block, and two cache blocks labeled 'I\$' and 'LI\$ / SLM'. To the right of the chip is the 'Xe-core' logo and the text 'Compute Building Block of Xe HPC-based GPUs'. Below this is a table of specifications:

8 Vector Engines	8 Matrix Engines	Load / Store 512 B/CLK
512 bit per engine	4096 bit per engine	Cache LI\$ / SLM (512KB), I\$

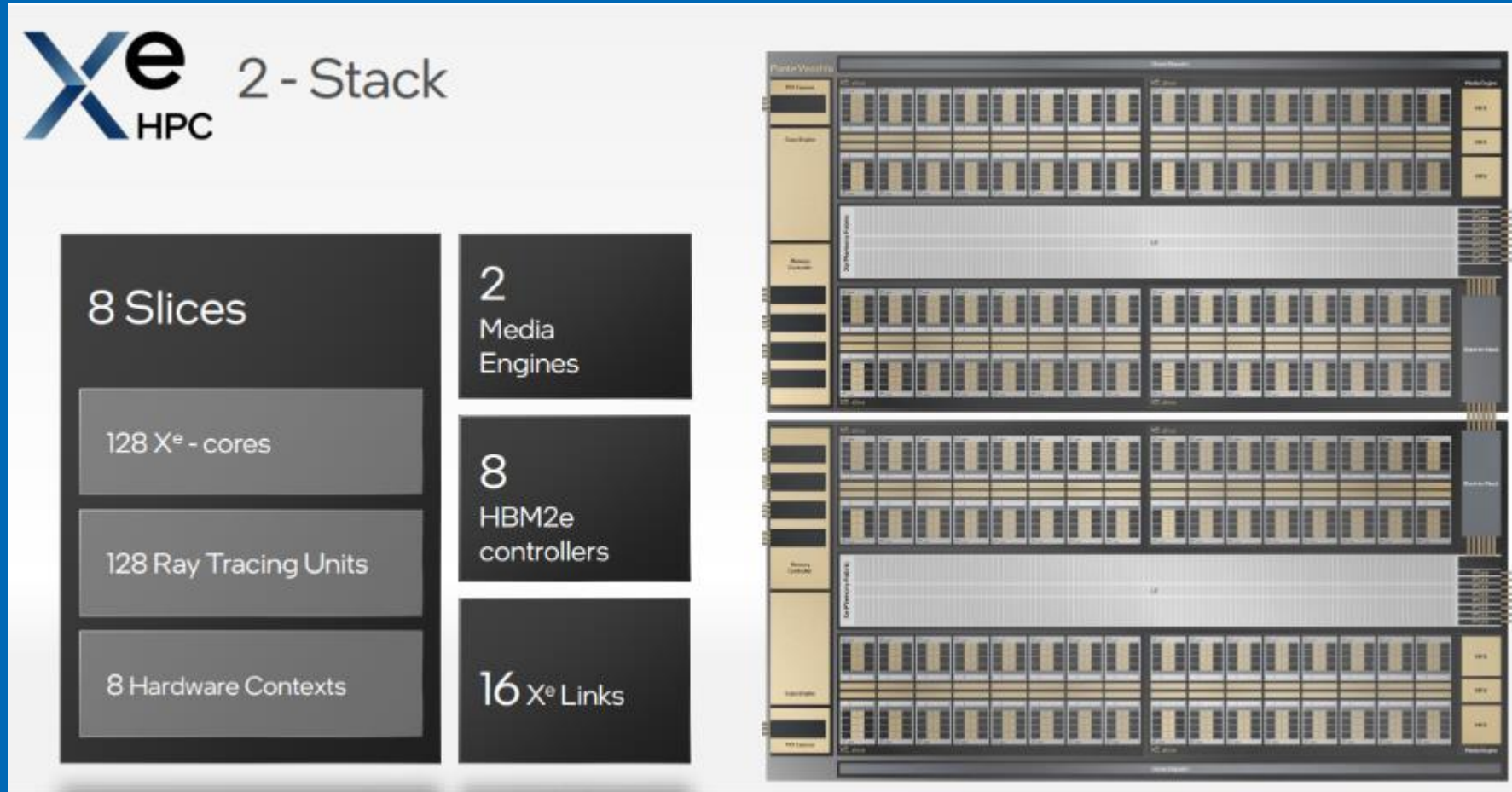
Xe Stack

Up to 4 Xe-Slices, Media Engine, L2 Cache, Memory Controllers, Xe-Links



2 Xe Stack

GPU with multiple Xe-Stack



Xe-HPC Architecture

- The Compute building block of the Xe HPC-based GPU is the Xe-Core consisting of 8 vector engines.
 - (Vector Engine formerly referred to as Execution-Units/EU, Xe-Core formerly referred to as Sub-Slice in Gen9/Gen11 Graphics HW)
- 16 Xe-Cores with a hardware context make up a Xe-Slice
- Up to 4 Xe-Slice makes Xe-Stack (with up to 64 Xe-Cores)
- 1 or more Xe-Stacks can be present in GPU

Intel Data Center GPU MAX Series

[Intel® Data Center GPU Max Series Overview](#)

Available today:

- Intel® Data Center GPU Max 1100 (56 Xe Cores)
- Intel® Data Center GPU Max 1550 (128 Xe Cores)

Intel Developer Cloud

Intel® Developer Cloud is a service platform for developing and running workloads in Intel®-optimized deployment environments with the latest Intel® processors, Intel® GPUs and performance-optimized software stacks.

- Sign-up for free
- cloud.intel.com

Hand-on Workshop

- Intel GPU Optimization using SYCL
 - Mapping SYCL Work-Groups to Intel GPU
 - Intel GPU Occupancy Calculator
 - SYCL Kernel Launching and Profiling

Resources

- SYCL Essentials training modules:
 - <https://github.com/oneapi-src/oneAPI-samples/tree/master/DirectProgramming/DPC%2B%2B/Jupyter/oneapi-essentials-training>
- Intel GPU Optimization Guide:
 - <https://www.intel.com/content/www/us/en/develop/documentation/oneapi-gpu-optimization-guide/top.html>
- SYCL Code Samples:
 - <https://github.com/oneapi-src/oneAPI-samples/tree/master/DirectProgramming/DPC%2B%2B>

Resources

- Download and Install Intel oneAPI Compiler, Libraries and Tools:
 - <https://www.intel.com/content/www/us/en/developer/tools/oneapi/base-toolkit.html>
- Build open source SYCL compiler:
 - <https://github.com/intel/llvm>
- SYCL Specification:
 - <https://registry.khronos.org/SYCL/specs/sycl-2020/pdf/sycl-2020.pdf>

Notices & Disclaimers

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

No product or component can be absolutely secure.

Your costs and results may vary.

© 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.

intel®