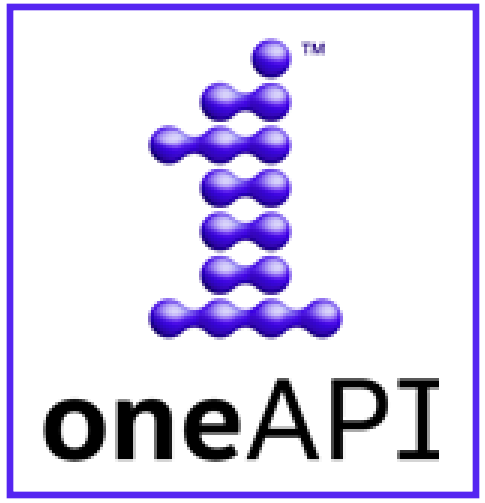


oneAPI AI Technical Advisory Board

Welcome and Thanks

- A unique opportunity to steer the parallel programming ecosystem
- A problem worth solving
  - Multi-architecture, avoiding lock-in to 1 specific hardware architecture
  - Direct and library-based programming
  - Extending existing models
  - Performant
- Your leadership, input, and feedback is critical

Rules of the Road

- DO NOT share any confidential information or trade secrets with the group
- DO keep the discussion at a High Level
  - Focus on the specific Agenda topics
  - We are asking for feedback on features for the oneAPI specification (e.g. requirements for functionality and performance)
  - We are NOT asking for feedback on any implementation details
- Please submit any feedback in writing on Github in accordance with the Contribution Guidelines at spec.oneapi.com. This will allow Intel to further upstream your feedback to other standards bodies, including The Khronos Group SYCL* specification.

oneAPI

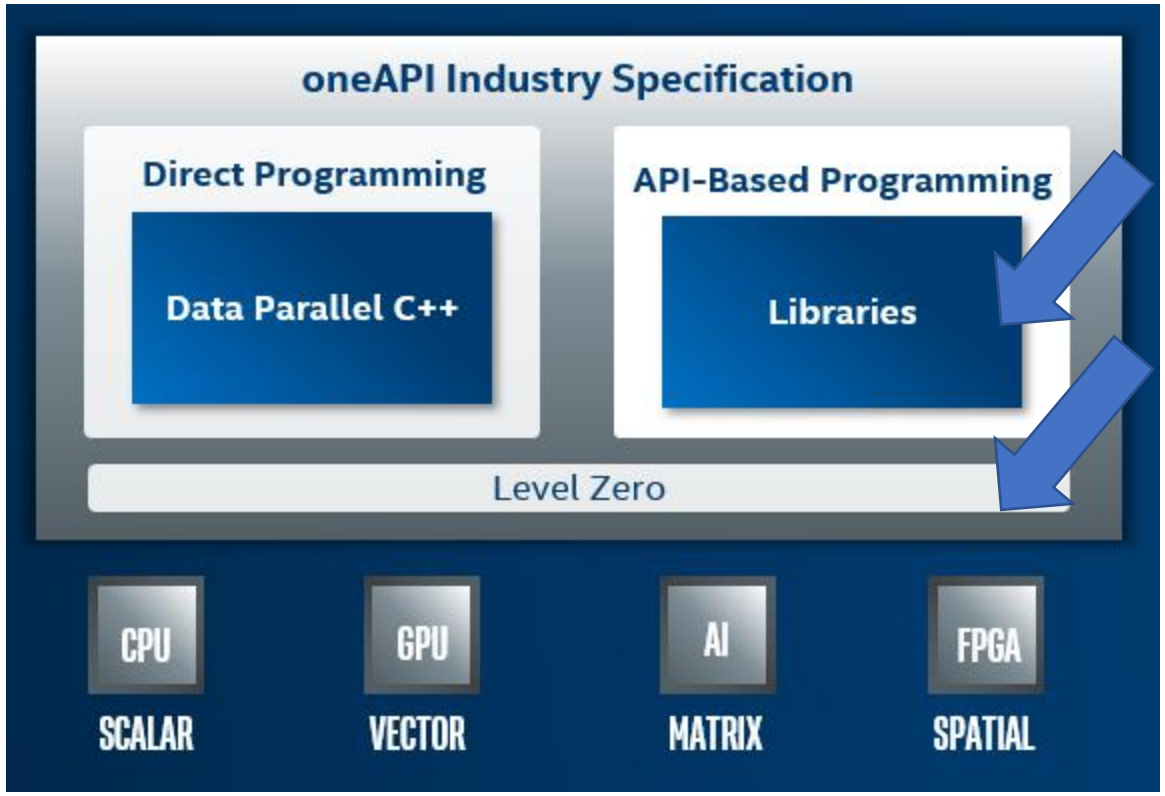
Agenda

- oneAPI Welcome & Introduction – Jeff McVeigh
- oneAPI for AI Everywhere – Wei Li
- oneDNN on ARM – Kentaro Kawakami
- oneDNN Graph API – Jian Hui Li
- Level Zero – Ben Ashbaugh
- Brainstorm Discussion: Requirements, Use Cases, Q& A - All

oneAPI

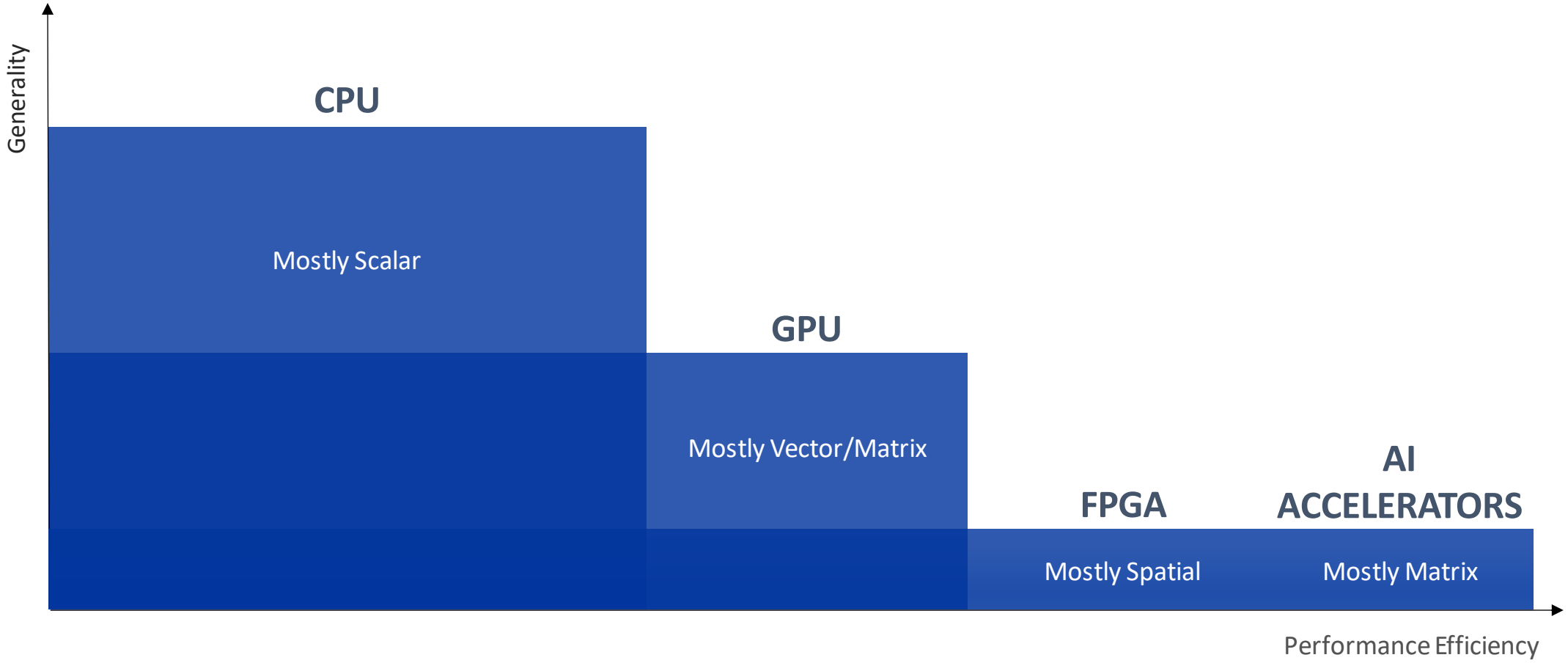
A unified programming model to simplify development across diverse architectures

- Common developer experience across architectures
- Unified and simplified language and libraries for expressing parallelism
- Uncompromised native high-level language performance
- Interoperates with existing languages and libraries
- Support for CPU, GPU, AI and FPGA
- Based on industry standards and open specifications



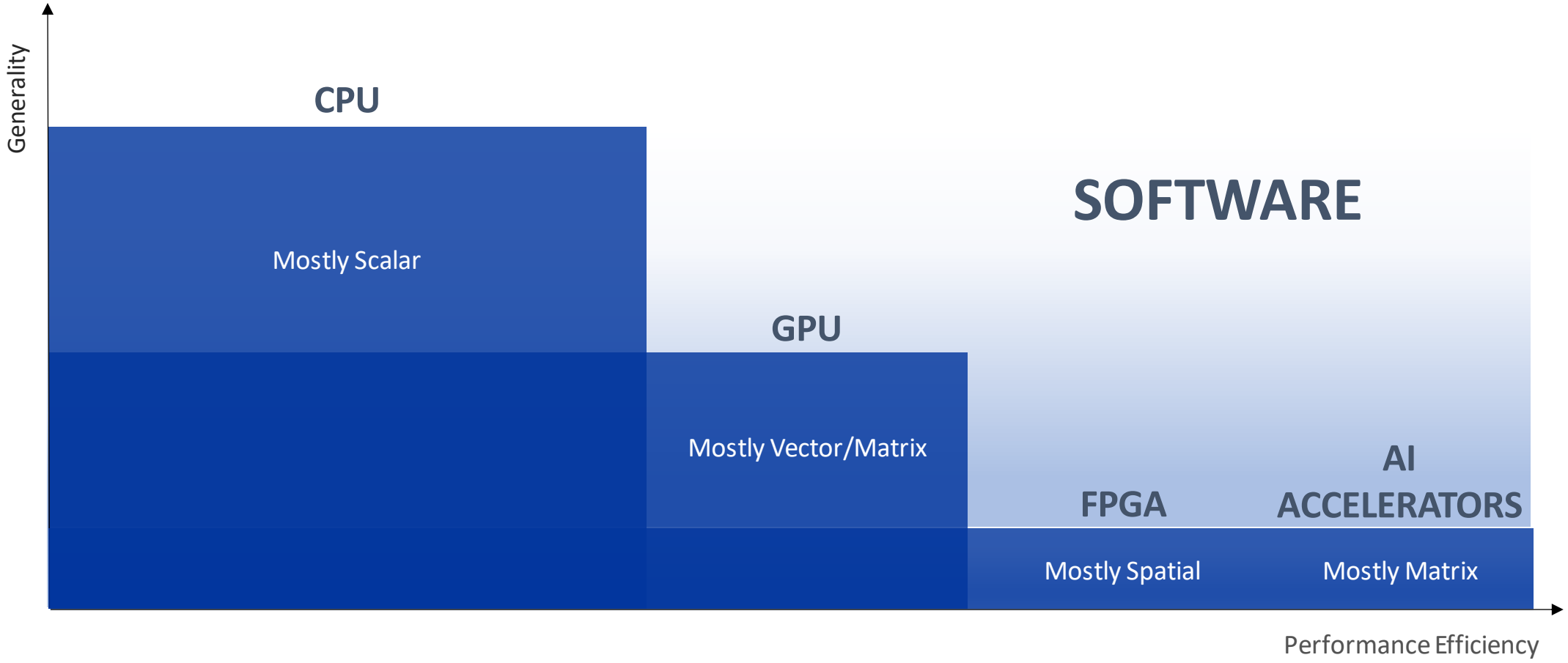
Architecture Impact

Performance x Generality



oneAPI

Increasing Generality through Software



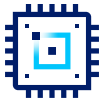
oneAPI

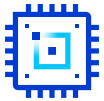
Unscalable Software for XPUs

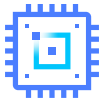
Services & Solutions

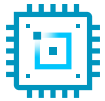
Applications

CPU	GPU	OTHER ACC.	FPGA / FF
Middleware Frameworks And Runtimes	Middleware Frameworks And Runtimes	Middleware Frameworks And Runtimes	Middleware Frameworks And Runtimes
Low Level Libraries	Low Level Libraries	Low Level Libraries	Low Level Libraries
Virtualization/ Orchestration	Virtualization/ Orchestration	Virtualization/ Orchestration	Virtualization/ Orchestration
OS	OS	OS	OS
Drivers	Drivers	Drivers	Drivers
FW IP & BIOS	FW IP & BIOS	FW IP & BIOS	FW IP & BIOS

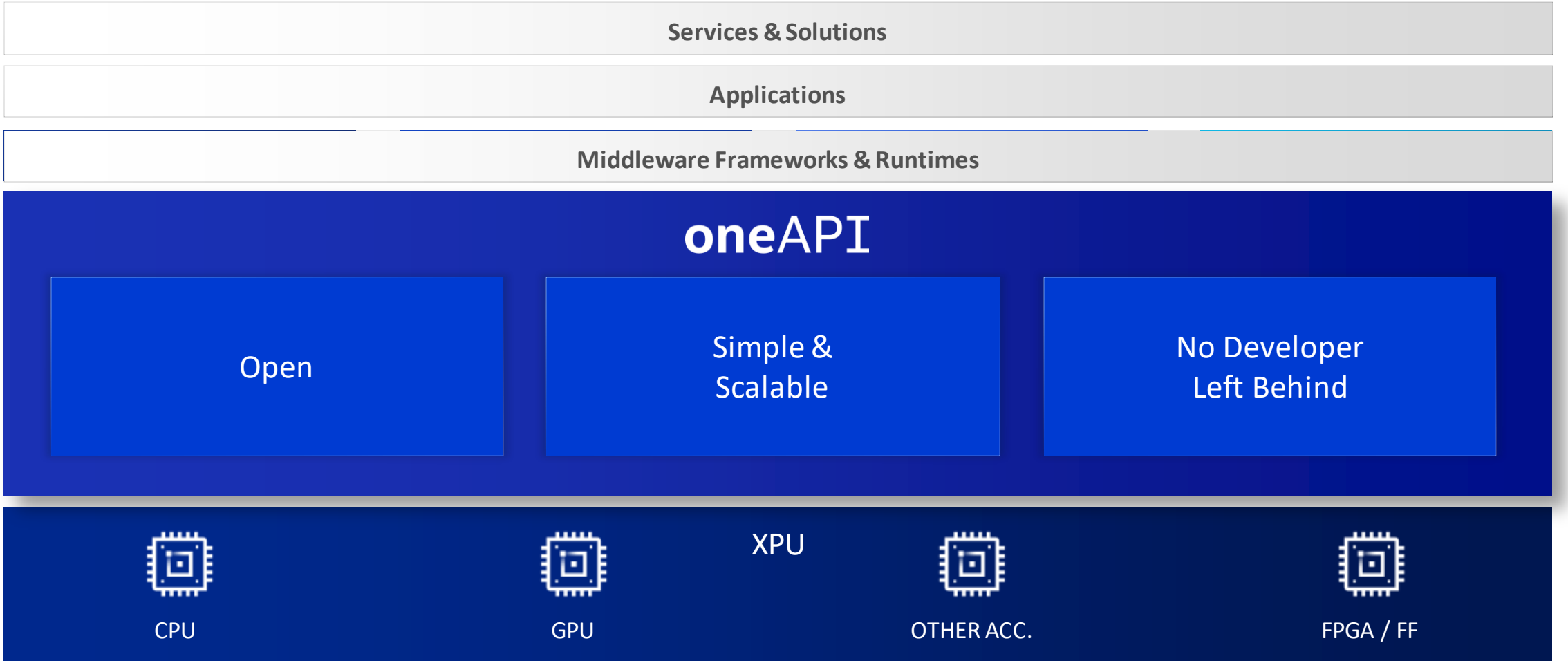








oneAPI: Scalable Initiative



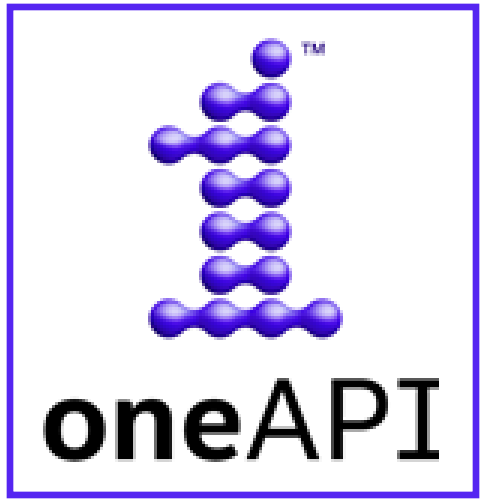
oneAPI

Request

"Our" initiative – help to frame it for your needs

Be brutally honest – good and bad

Invite your friends – and have fun

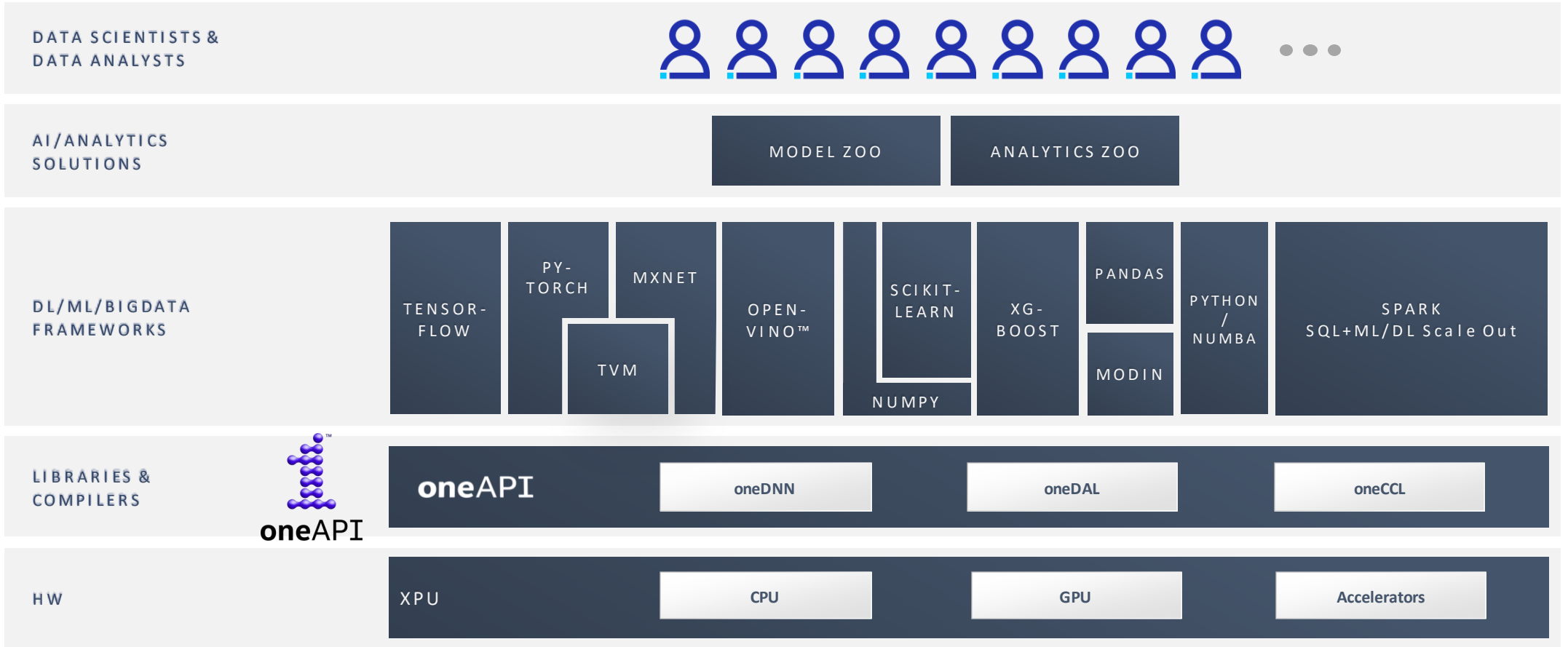


oneAPI for AI Everywhere

Wei Li
VP/GM Machine Learning Performance
Intel

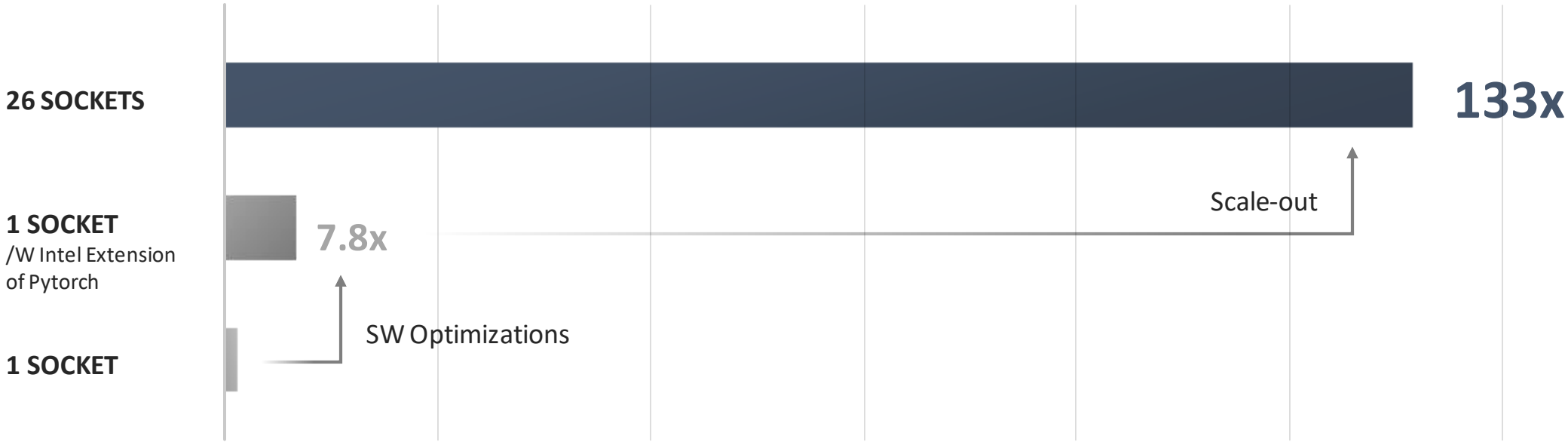

AI Everywhere
intel

Ecosystem of AI Software Stack



Scale Out DLRM Training

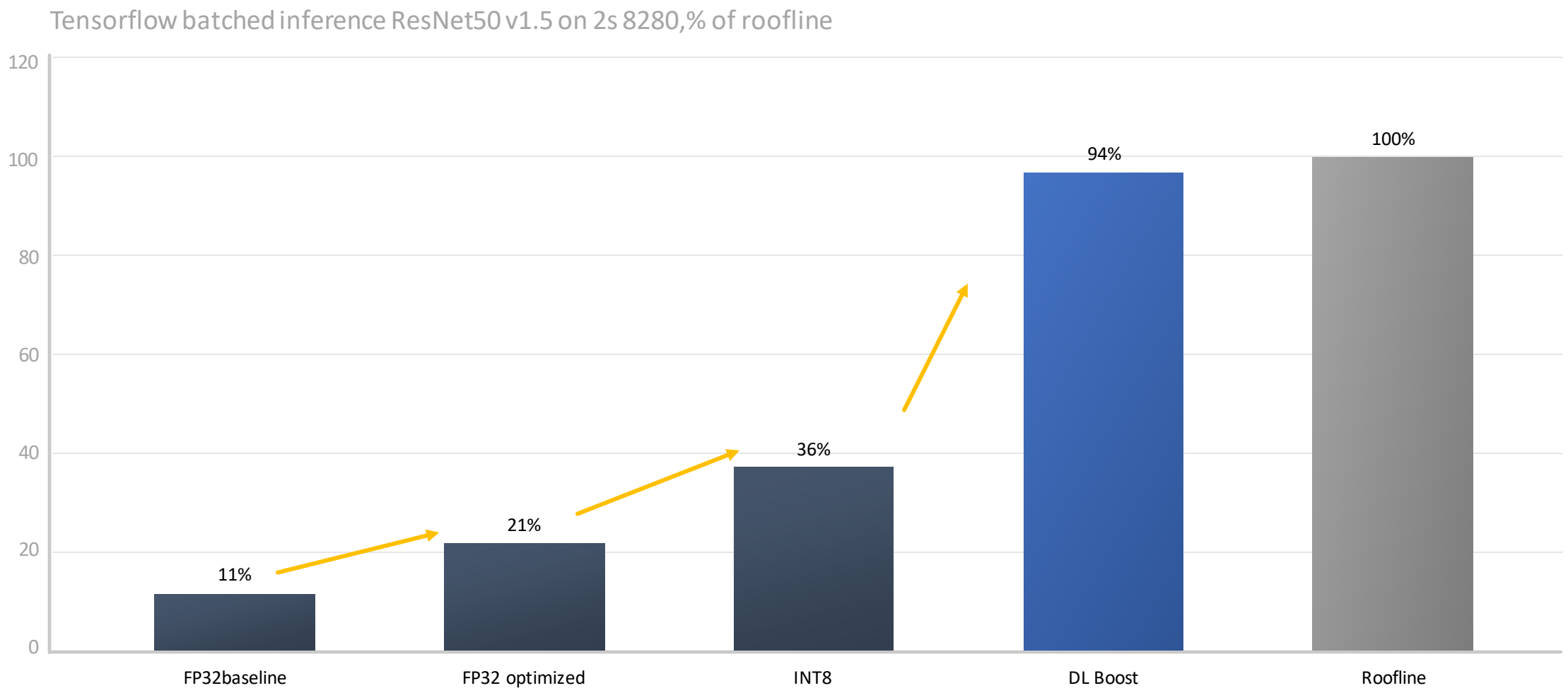
DLRM Training Performance on Xeon 8280 using PyTorch 1.4.1 - Higher is better



Date of testing: March and April 2020
CPU: Intel(R) Xeon(R) Platinum 8280 CPU | Bios: SE5C620.86B.02.01.0009.092820190230, Memory: 12x 18ASF2G72PDZ-2G6D1, Network: 2x Intel Corporation Omni-Path HFI Silicon 100 Series, OS: Cent OS 7.6, Bootline: BOOT_IMAGE=/vmlinuz-3.10.0-957.el7.x86_64 rhgb quiet intel_pstate=disable nmi_watchdog=0 intel_idle.max_cstate=1 nohz_full=2-111 LANG=en_US.UTF-8, OPA stack: 10.9.2.0-7, Pytorch: 1.4.1.
26 sockets refers to 13 dual-socket nodes and 1 socket refers to 1 node of Intel Xeon 8280. 1 socket optimization was achieved with Intel extension of PyTorch.
See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
Refer to https://software.intel.com/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

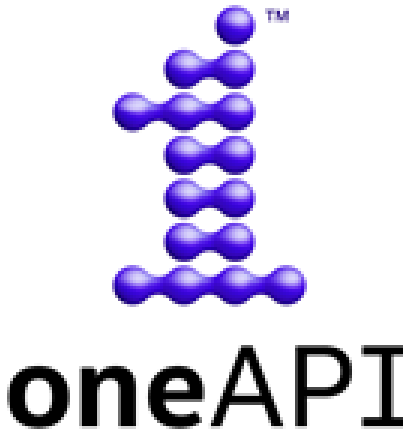
oneAPI

Path to CPU Performance Roofline



DL Boost Test configs: TensorFlow v2.3; Test by Intel; Test date: 30/06/2020; Platform: Intel(R) Xeon(R) Platinum 8280 CPU; #Nodes: 1; #Sockets: 2; Cores/socket: 28; Threads/socket: 56; HT: On; Turbo: On; BIOS version: SE5C620.86B.02.01.0008.031920191559; System DDR Mem Config: 12 slots / 16GB / 2933; OS: CentOS Linux 7; Kernel: 3.10.0-957.el7.x86_64.
See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
Refer to https://software.intel.com/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

oneAPI



Thank You!

http://oneapi.com