DON'T BE ENCUMBERED BY HISTORY, JUST GO OUT AND DO SOMETHING WONDERFUL ...
Intel powers the world
EXASCALE FOR EVERYONE

RAJA KODURI
1st ERA IN HPC
VERTICALLY INTEGRATED SYSTEMS

PROPRIETARY HARDWARE
AND SOFTWARE

PARAGON
143 GFLOP

DELTA
8.1 GFLOP


# OF HPC SYSTEMS

10's
1000's
100,000's
1,000,000's
10,000,000's
10,000,000's
2\textsuperscript{nd} Era in HPC

Mostly based on general purpose CPU's

X86
Linux
Open Standards
Build to Order Systems

ASCi RED
1 TFLOP

TIANHE-2
34 PFLOP

# of HPC Systems


10's
1000's
10,000's
100,000's
1,000,000's
10,000,000's
NEXT ERA IN HPC
DRIVEN BY INSATIABLE AI COMPUTE
COMPUTE DEMOCRATIZATION
TECHNOLOGY LED DISRUPTIONS

1 BILLION INTERNET CONNECTED DEVICES

PC ERA
DIGITIZE EVERYTHING

NETWORK EVERYTHING


10^{18}
10^{15}
10^9
10^4
10^2
COMPUTE DEMOCRATIZATION
TECHNOLOGY LED DISRUPTIONS

1 BILLION INTERNET CONNECTED DEVICES

PC ERA
DIGITIZE EVERYTHING

1980

1990

2000

2010

2020

2030

2040

10 BILLION CLOUD CONNECTED DEVICES

MOBILE + CLOUD ERA

MOBILE EVERYTHING

CLOUD EVERYTHING

10 BILLION CLOUD CONNECTED DEVICES

NETWORk EVERYTHING

10^18

10^15

10^9

10^4

10^2

COMPUTE
COMPUTE DEMOCRATIZATION
TECHNOLOGY LED DISRUPTIONS

1 BILLION INTERNET CONNECTED DEVICES

PC ERA
DIGITIZE EVERYTHING

1980

10 BILLION CLOUD CONNECTED DEVICES

MOBILE EVERYTHING

2000

MOBILE + CLOUD ERA

2010

CLOUD EVERYTHING

2020

100 BILLION INTELLIGENT CONNECTED DEVICES

INTELLIGENCE ERA

2030

2040

100 BILLION INTELLIGENT CONNECTED DEVICES

INTELLIGENCE ERA

2030

2040

T E C H N O L O G Y  L E D  D I S R U P T I O N S

DEMOCRATIZATION
COMPUTE DEMOCRATIZATION
TECHNOLOGY LED DISRUPTIONS

1 BILLION INTERNET CONNECTED DEVICES

10 BILLION CLOUD CONNECTED DEVICES

MOBILE + CLOUD ERA

CLOUD EVERYTHING

1 BILLION CLOUD CONNECTED DEVICES

EXASCALE FOR EVERYONE

PC ERA
DIGITIZE EVERYTHING

NETWORK EVERYTHING

MOBILE EVERYTHING


1018
1015
109
104
102
EXASCALE FOR EVERYONE
Intel will deliver a diverse mix of **Scalar**, **Vector**, **Matrix** and **Spatial Architectures** designed with state of the art **Process technology**, fed by disruptive **Memory** hierarchies, integrated into systems with advanced **Packaging**, deployed at hyperscale with lightspeed **Interconnect** links, unified by a **single Software** abstraction, with benchmark defining **Security features**.
ARCHITECTURE IMPACT = PERFORMANCE X GENERALITY
ARCHITECTURE IMPACT

PERFORMANCE * GENERALITY
GENERALITY ∝ SOFTWARE STACK SCALE
GENERALITY \( \propto \) SOFTWARE STACK SCALE

<table>
<thead>
<tr>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIRMWARE</td>
</tr>
<tr>
<td>OPERATING SYSTEMS</td>
</tr>
<tr>
<td>VIRTUALIZATION/ORCHESTRATION</td>
</tr>
<tr>
<td>RUNT ImES</td>
</tr>
<tr>
<td>MIDDLEWARE</td>
</tr>
<tr>
<td>APPLICATIONS</td>
</tr>
<tr>
<td>SOLUTIONS &amp; SERVICES</td>
</tr>
<tr>
<td>DIVERSITY OF USER WORKLOADS</td>
</tr>
<tr>
<td>COMPILERS</td>
</tr>
<tr>
<td>LIBRARIES</td>
</tr>
<tr>
<td>TOOLS</td>
</tr>
</tbody>
</table>

20M DEVELOPERS
GENERALITY $\propto \frac{1}{\text{ARCHITECTURE HETEROGENEITY}}$
HETEROGENEOUS MATH IN CPU

Source: Results have been calculated based on internal Intel analysis and are provided for informational purposes only.
HETEROGENEOUS MATH IN CPU

Graph not drawn to scale
Source: Intel Projections
CPU IMPACT WITH ISA EXTENSIONS AND SOFTWARE

- CPU CORE
- DIRECT ENABLEMENT
- TOOLS
- LIBRARIES
- COMPILERS
- HETERO EXTENSIONS
CPU IMPACT WITH ISA EXTENSIONS AND SOFTWARE
DISCRETE GPUS

CPU CORE + EXTENSIONS

PERFORMANCE

GENERALITY

PROPRIETARY SOFTWARE

DISCRETE GPU IMPACT
DISCRETE GPU IMPACT

CPU CORE + EXTENSIONS

GENERALITY

PERFORMANCE
UNSCALABLE HETEROGENEOUS SOFTWARE

DIVERSITY OF USER WORKLOADS

- CPU
- GPU
- FPGA
- ACCELERATORS

COMPILERS
LIBRARIES
TOOLS

- SOLUTIONS & SERVICES
- APPLICATIONS
- MIDDLEWARE
- RUNTIMES
- VIRTUALIZATION
- OPERATING SYSTEMS
- FIRMWARE

COMPILERS
LIBRARIES
TOOLS

- SOLUTIONS & SERVICES
- APPLICATIONS
- MIDDLEWARE
- RUNTIMES
- VIRTUALIZATION
- OPERATING SYSTEMS
- FIRMWARE

COMPILERS
LIBRARIES
TOOLS

- SOLUTIONS & SERVICES
- APPLICATIONS
- MIDDLEWARE
- RUNTIMES
- VIRTUALIZATION
- OPERATING SYSTEMS
- FIRMWARE

COMPILERS
LIBRARIES
TOOLS

- SOLUTIONS & SERVICES
- APPLICATIONS
- MIDDLEWARE
- RUNTIMES
- VIRTUALIZATION
- OPERATING SYSTEMS
- FIRMWARE

INTTEL
oneAPI

No transistor left behind™
DIVERSITY OF USER WORKLOADS

OPEN

SCALAR – VECTOR – MATRIX – SPATIAL
oneAPI GOALS

DIVERSITY OF USER WORKLOADS

OPEN

SIMPLE & SCALABLE

SCALAR – VECTOR – MATRIX – SPATIAL
oneAPI GOALS

DIVERSITY OF USER WORKLOADS

OPEN

SIMPLE & SCALABLE

NO DEVELOPER LEFT BEHIND

SCALAR – VECTOR – MATRIX – SPATIAL
DIVERSITY OF USER WORKLOADS

- DIRECT PROGRAMMING LANGUAGES
- DOMAIN SPECIFIC LIBRARIES
- MIGRATION TOOLS
- ANALYSIS & DEBUG TOOLS

SYSTEM PROGRAMMING

SCALAR – VECTOR – MATRIX – SPATIAL
DIVERSITY OF USER WORKLOADS

SYSTEM PROGRAMMING

Peer to Peer Comms
Scheduler
Sync Primitives
Profile
Device & Memory management
Trace & Debug

SCALAR – VECTOR – MATRIX – SPATIAL
oneAPI Stack

DIVERSITY OF USER WORKLOADS

DIRECT PROGRAMMING LANGUAGES
DOMAIN SPECIFIC LIBRARIES
MIGRATION TOOLS
ANALYSIS & DEBUG TOOLS

SYSTEM PROGRAMMING

SCALAR – VECTOR – MATRIX – SPATIAL
DIVERSITY OF USER WORKLOADS

DIRECT PROGRAMMING LANGUAGES

- Intel® oneAPI DPC++ Compiler
- Intel® Fortran Compiler w/ OpenMP*
- Intel® C++ Compiler w/ OpenMP*
- Intel® Distribution for Python*

SYSTEM PROGRAMMING

MAIN SPECIFIC LIBRARIES

MIGRATION TOOLS

ANALYSIS & DEBUG TOOLS

SCALAR – VECTOR – MATRIX – SPATIAL
DIVERSITY OF USER WORKLOADS

DIRECT PROGRAMMING LANGUAGES

DOMAIN SPECIFIC LIBRARIES

MIGRATION TOOLS

ANALYSIS & DEBUG TOOLS

SYSTEM PROGRAMMING

SCALAR – VECTOR – MATRIX - SPATIAL
DIVERSITY OF USER WORKLOADS

SCALAR – VECTOR – MATRIX – SPATIAL MIGRATION TOOLS

ANALYSIS & DEBUG TOOLS

SYSTEM PROGRAMMING

DIRECT PROGRAMMING LANGUAGES

DOMAIN SPECIFIC LIBRARIES

Intel® oneAPI Threading Building Blocks
Intel® oneAPI DPC++ Library
Intel® oneAPI Math Kernel Library
Intel® oneAPI Data Analytics Library
Intel® oneAPI Collective Communications Library
Intel® Video Processing Library
Intel® oneAPI Deep Neural Network Library
Intel® MPI Library
Intel® Integrated Performance Primitives
DIVERSITY OF USER WORKLOADS

- DIRECT PROGRAMMING LANGUAGES
- DOMAIN SPECIFIC LIBRARIES
- MIGRATION TOOLS
- ANALYSIS & DEBUG TOOLS

SYSTEM PROGRAMMING

SCALAR – VECTOR – MATRIX – SPATIAL
DIVERSITY OF USER WORKLOADS

DIRECT PROGRAMMING LANGUAGES

DOMAIN SPECIFIC LIBRARIES

MIGRATION TOOLS

ANALYSIS & DEBUG TOOLS

SYSTEM PROGRAMMING

SCALAR – VECTOR – MATRIX – SPATIAL
DIVERSITY OF USER WORKLOADS

DIRECT PROGRAMMING LANGUAGES

DOMAIN SPECIFIC LIBRARIES

MIGRATION TOOLS

SYSTEM PROGRAMMING

SCALAR – VECTOR – MATRIX – SPATIAL

ANALYSIS & DEBUG TOOLS

Intel® VTune™ Profiler

Intel® Advisor

GDB*

Intel® Inspector

Intel® Trace Analyzer & Collector

Intel® Cluster Checker
PUBLIC BETA AVAILABLE TODAY!

GO TO SOFTWARE.INTEL.COM/ONEAPI
TRADITIONAL
ADOPTION RATES

SERVER & PC

20M DEVELOPERS
oneAPI

DevCloud

Available Today

- No hardware acquisition
- No installation
- No downloads
- No set-up and configuration
INTEL GPU IMPACT
Billion+ users reach
ONE GPU ARCHITECTURE

TWO MICRO ARCHITECTURES

DATA CENTER / AI
ENTHUSIAST
MID-RANGE
INTEGRATED + ENTRY
TERAFLOPS
PETAFLOPS
ONE GPU ARCHITECTURE
FROM TERA FLOPS TO EXASCALE

HPC EXASCALE

DATACENTER / AI

ENThusiast

MID-RANGE

INTEGRATED + ENTRY

TERAFLOPS
HPC FEATURES

COMPUTE

MEMORY
HPC FEATURES

- Compute
- Memory
- Scalability
- AI Performance
- HPC Performance
VARIABLE VECTOR WIDTH

SIMT (GPU STYLE)
SIMD (CPU STYLE)
SIMT + SIMD (MAX PERF)
VARIABLE VECTOR WIDTH

SIMT (GPU STYLE)
SIMD (CPU STYLE)
SIMT + SIMD (MAX PERF)
SCALABLE TO THOUSANDS OF EXECUTION UNITS

Xe
HPC

COMPUTE  SCALABILITY
NEW DATA PARALLEL

MATRIX ENGINE

AI DATA TYPES SUPPORTED
INT8, BF16, FP16
UPTO 32X VECTOR RATE

Source: Intel projections
40X INCREASE IN DPFP PER EU

DOUBLE PRECISION FLOATING POINT SUPPORT

Source: Intel projections
HPC FEATURES

- COMPUTE
- MEMORY

- SCALABILITY
- BANDWIDTH
- UNIFIED MEMORY
SCALABLE MEMORY FABRIC

Xe COMPUTE

XEMF

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

Xe COMPUTE
SCALABLE MEMORY FABRIC

Xe HPC

Xe COMPUTE
Xe COMPUTE
Xe COMPUTE
Xe COMPUTE
Xe COMPUTE
Xe COMPUTE
Xe COMPUTE
Xe COMPUTE

XEMF

HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
SCALABLE MEMORY FABRIC

Xe HPC

XEMF

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL

HBM CHANNEL
SCALABLE MEMORY FABRIC

Xe HPC

HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
HBM CHANNEL
RAMBO CACHE

HPC

Source: Intel projections
UNIFIED COHERENT MEMORY TO CPU AND OTHER GPUs
HPC FEATURES

- Compute
  - Scalability
  - AI Performance
  - HPC Performance

- Memory
  - Scalability
  - Bandwidth
  - Unified Memory
PATH TO EXASCALE GPU

EXASCALE CHALLENGES

- Compute Density
- Memory
- Connectivity
- Reliability
TECHNOLOGY FOR EXASCALE

NEED HUGE LEAP IN PERF/WATT AND PERF/MM^2

INTEL NEXT GEN 7nm PROCESS & FOVEROSTM PACKAGING

COMPUTE DENSITY
MEMORY FOR EXASCALE

NEED HUGE LEAP IN BANDWIDTH/WATT & FOOTPRINT/MM^2

EMIB FOR HBM & FOVEROS™ FOR RAMBO CACHE
CONNECTIVITY FOR EXASCALe

Xe LINK
SCALE OUT TO MANY GPUS/NODE
UNIFIED MEMORY
CXL BASED
RELIABILITY FOR EXASCALE

XEON™ CLASS RAS

IN-FIELD REPAIR

ECC, PARITY ACROSS ALL MEMORY AND CACHES

CONNECTIVITY
US FIRST EXASCALE COMPUTER

RICK STEVENS
ASSOCIATE DIRECTOR FOR COMPUTING, ENVIRONMENT AND LIFE SCIENCES AT ARGONNE NATIONAL LABORATORY
LEADERSHIP PERFORMANCE
For HPC, Data Analytics, AI

UNIFIED MEMORY ARCHITECTURE
Across CPU & GPU

ALL-TO-ALL CONNECTIVITY WITHIN NODE
Low latency, high bandwidth

UNPARALLELED I/O SCALABILITY ACROSS NODES
8 fabric endpoints per node, DAOS

2 INTEL XEON™ SCALABLE PROCESSORS
“Sapphire Rapids”

6 Xe ARCHITECTURE BASED GPUs
“Ponte Vecchio”

oneAPI
Unified programming model

AURORA – BRINGING IT ALL TOGETHER

DELIVERED IN 2021
PERFORMANCE INCREASE PER NODE

Source: Intel projections
THANK YOU!

RAJA KODURI
Notices & Disclaimers

Intel® technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Performance results are based on testing as of the dates shown in configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

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 information, go to 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 and SSE3 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.

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.

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 websites referenced in this document. You should visit the referenced website and confirm whether referenced data are accurate.

© Intel Corporation. Intel, Xeon, Optane, AVX, and DL Boost are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.
Business Forecast: 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.