No Transistor Left Behind

By Raja Koduri
A tribute to

Frances Allen

“Who Helped Hardware Understand Software”
“No Transistor Left Behind”

DAVID BLYTHE, 2018
Technology Led Disruptions

**PC ERA**
1B Internet Connected Devices

Digitize Everything

Network Everything

Technology Led Disruptions

MOBILE + CLOUD ERA
10B Cloud Connected Devices

Digitize Everything
Network Everything
Mobile Everything
Cloud Everything

Technology Led Disruptions

- Digitize Everything
- Network Everything
- Mobile Everything
- Cloud Everything

100B Intelligent Connected Devices

Intelligent Everything

Exascale For Everyone

- 1980
- 1990
- 2000
- 2010
- 2020
Intelligence is Expensive

Alex-Net to AlphaGo Zero: 300,000x Increase in Compute

Source: AI and Compute • November 7, 2019 • Dario Amodei & Danny Hernandez
Performance and Generality

Performance

- Human level
- Unconstrained world
- General Purpose

Generality

- Human
- Human level intelligence
- Constrained world
- Fixed function

- Humanoid robots
- Machine translations
- Image classification
- Chess

Graph: Illustrative purposes only

Data is Exploding

We are generating data at a faster rate than our ability to analyze, understand, transmit, secure and reconstruct in real-time.
Memory Wall

Source: http://research.nvidia.com/publication/2017-02_architecting-an-energy-efficient
Data and Super Intelligence

Graph: Illustrative purposes only

"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's Law."

RICH SUTTON, THE BITTER LESSON, MARCH 2019
“The number of people predicting the death of Moore’s law doubles every two years”
Moore's Law - Our Exponential Entitlement

Based on Intel Internal Data
... we haven’t fully exploited what was given to us by Moore!
Software Productivity vs. Hardware Efficiency

for i in xrange(4096):
    for j in xrange(4096):
        for k in xrange(4096):
            C[i][j] += A[i][k] * B[k][j]
Training & Inference Software Speed-Up
On Intel Xeon

“What Andy Giveth, Bill Taketh Away”
“There’s still plenty of room at the bottom”
"There’s still plenty of room at the bottom"

A path to 50x Transistor Density
Intel 10nm

FinFET

NMOS

PMOS
Pitch Scaling

FinFET

Intel 10nm

DENSITY INCREASE
x3

TOTAL
x3

NMOS

PMOS

Pitch Scaling
Pitch Scaling  
Nanowire

Density Increase:

- NMOS
- PMOS

Total Density Increase: x2

Pitch Scaling: x6
Nanowire

Stacked Nanowire

DENSITY INCREASE

x2

TOTAL

x12
Wafer to Wafer Stacking

Density Increase

$\times 2$

Total

$\times 24$
Die to Wafer Stacking

Density Increase

x2

Total

~x50
What about Power?

- **Power**
  - BASELINE
  - VOLTAGE SCALING
  - CAPACITANCE SCALING
  - NEW PACKAGING
  - FREQUENCY SCALING
  - NEW ARCHITECTURES

50x

For illustrative purposes only
Case for Advanced Packaging

NO SINGLE TRANSISTOR NODE IS OPTIMAL ACROSS ALL DESIGN POINTS!

**TRANSISTOR DESIGN TARGET RANGE**

- Desktop CPU
- High Perf FPGA
- Server CPU
- dGPU
- Mobile CPU
- Power Efficient FPGA
- iGPU
- Entry CPU/PCH

**TRANSISTOR DIVERSITY**

- Logic Transistors
- Analog/RF Transistors
- High Speed Memory
- Dense Memory
- Non-Volatile Memory
- High speed IO
- Configuration Memory

NO SINGLE TRANSISTOR NODE IS OPTIMAL ACROSS ALL DESIGN POINTS!
Packaging Technology Improvements

**STANDARD PACKAGE**
- **BUMP PITCH**: 100 um
- **BUMP DENSITY**: 100/mm²
- **POWER**: 1.7 pJ/bit

**2D / 2.5D**
- **BUMP PITCH**: 55 – 36 um
- **BUMP DENSITY**: 330 – 772/mm²
- **POWER**: 0.50 pJ/bit

**3D**
- **BUMP PITCH**: 50 – 25 um
- **BUMP DENSITY**: 400-1,600/mm²
- **POWER**: 0.15 pJ/bit

**FUTURE**
- **BUMP PITCH**: < 10 microns
- **BUMP DENSITY**: > 10,000/mm²
- **POWER**: < 0.05 pJ/bit

Interconnect Density vs. Power Efficiency
Hybrid Bonding
Dense vertical interconnects

- Smaller, simpler circuits
- Lower capacitance
- Lower power

Area scales with bump pitch

50 um Pitch
Lakefield
400 bumps/mm²

10 um Pitch
Hybrid Bonding
10000 bumps/mm²
Advanced Packaging Products

**KABY LAKE G**
2D
- Intel CPU
- AMD GFX
- HBM

**LAKEFIELD**
3D
- Internal Silicon on Multiple Nodes

**AGILEX FPGA**
2.5D
- Intel FPGA
- Foundry IO Chiplets
- HBM

**PONTE VECCHIO**
3D
- Internal and External Silicon on Multiple Nodes
Memory Wall

Source: http://research.nvidia.com/publication/2017-02_architecting-an-energy-efficient
Memory Hierarchy Disruptions

- **Compute Cache**: 10s MB, ~1ns
- **In-Package Memory**: 1s GB, ~10ns
- **Memory**: 10s GB, <100ns
- **Secondary Storage**: 100s GB, <1usec
- **Storage Performance Gap**: 1s TB, <10µsecs
- **Capacity Gap**: 10s TB, <100µsecs
- **Secondary Storage**: 1s TB, <100µsecs
- **Cost-Performance Gap**: 10s TB, <10 msecs
Memory Hierarchy Disruptions

- **Compute Cache**: 10s MB, ~1ns
- **In-Package Memory**: 1s GB, ~10ns
- **Memory**: 10s GB, <100ns
- **Persistent Memory**: 100s GB, <1usec
- **3D XPoint**: 1s TB, <10µsecs
- **3D QLC NAND**: 10s TB, <100µsecs
- **Tertiary Storage**: 10s TB, <10 msecs

**Memory**
- **Capacity Gap**

**Storage**
- **Storage Performance**
Compute and Memory

a Vision for Next Gen

- 10x on capacity
- 10x more B/W
- 10x lower latency
- 10x lower power
Compute and Memory

a Vision for Next Gen

NEW MEMORY?

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>10 – 100s GB</td>
</tr>
<tr>
<td>Latency</td>
<td>&lt; 10ns</td>
</tr>
<tr>
<td>Power</td>
<td>&lt; 0.5 pJ/Bit</td>
</tr>
</tbody>
</table>

TIGHTLY INTEGRATED WITH COMPUTE
“Plenty of Room at the Bottom”

“What would happen if we could arrange atoms one by one the way we want them?”

“When we get to circuits of, say 7 atoms – we will have new opportunities for design. We will manufacture in different ways”

RICHARD FEYNMAN, 1959
Beyond Exascale Compute
General Flow of Architecture

In Hardware Companies
1 Core
~25
TOTAL # OF CORES
25

1 Core
PACKAGE

x2

TOTAL # OF CORES
50

1 Core
DENSITY SCALING

x50

TOTAL # OF CORES
2500

1 Core
1 Core

PACKAGE PER BOARD

x4

TOTAL # OF CORES
10,000
“Software is Eating the World”

MARC ANDREESSEN, 2011
Hardware

Software

ISA

Hardware
Compute Disruptions

PC ERA

- >1M PC DEVELOPERS

Digitize Everything

Network Everything

X86 + Windows

Compute Disruptions

MOBILE + CLOUD ERA
>10M MOBILE + CLOUD DEVELOPERS
+ PC DEVELOPERS

- Apple
- Android
- arm
- x86

Cloud Everything
Mobile Everything
Network Everything
Digitize Everything

>10M MOBILE + CLOUD DEVELOPERS
+ PC DEVELOPERS

HOT CHIPS
Architecture Impact

PERFORMANCE X GENERALITY
x86 Developer Ecosystem

>20M Developers
Software Stack & Developers

Driven by Abstraction

Affinity to Hardware

- SERVICES & SOLUTIONS
- APPLICATIONS
- MIDDLEWARE FRAMEWORKS AND RUNTIMES
- LOW LEVEL LIBRARIES
- VIRTUALIZATION/ ORCHESTRATION
- OS
- DRIVERS
- FW IP & BIOS
- x86

# OF DEVELOPERS

- 20M
- 500K
- 50K
Stack and Swiss Cheese

Middle is Full of "holes"
New Hardware / Software Contracts

The Reality...

"IT JUST WORKS"

“IT ALMOST WORKS”
What is the Hardware / Software Contract for the next Era?

**PC ERA**

- >1M PC DEVELOPERS
- 
  - **X86** + **Windows**
  - Digitize Everything

**MOBILE + CLOUD ERA**

- >10M MOBILE + CLOUD DEVELOPERS
- + **PC DEVELOPERS**
- 
  - **Apple** + **Android** + **Linux**
  - **x86** + **arm**
  - Mobile Everything
  - Cloud Everything

**INTELLIGENCE ERA**

- >100M
- + **AI DEVELOPERS**
- + MOBILE + CLOUD DEVELOPERS
- + PC DEVELOPERS
- 
  - x86, ARM, RISC-V, AI, GPU, MEMORY, NETWORK

**Timeline**

- 1980
- 1990
- 2000
- 2010
- 2020
Generality $\propto \frac{1}{\text{Architecture Heterogeneity}}$
Heterogeneous Math in CPU

OPS / Clock

150X

50X

2008 2010 2012 2014 2017 2018

WESTMERE SANDY BRIDGE HASWELL

CASCADE LAKE SKY LAKE

For illustrative purposes only
Heterogeneous Math in CPU

For illustrative purposes only
CPU Impact with ISA Extensions & Software

CPU Core

Hetero Extensions

Gap Covered by Software
Increasing Impact

Generality

Performance
CPU Impact with ISA Extensions & Software

~3-5 YEARS TO REACH BROAD ADOPTION
Productivity and Scale

Mandelbrot Static Instructions Generated per Line of Code

- Assembly
- C
- C++
- Java
- JS
- Python

For illustrative purposes only
ABSTRACTION REQUIREMENTS

SCALABLE AND OPEN
ABSTRACTION REQUIREMENTS

SCALABLE AND OPEN

ABSTRACT AT MULTIPLE LAYERS
Performance vs. Productivity

Mandelbrot Static Instructions Generated per Line of Code

- Assembly
- C
- C++
- Java
- JS
- Python

For illustrative purposes only
ABSTRACTION REQUIREMENTS

- Scalable and Open
- Abstract at Multiple Layers
- Support Ninja’s Across the Entire Stack
Generality $\propto$ 1/ Architecture Heterogeneity

<table>
<thead>
<tr>
<th>SERVICES &amp; SOLUTIONS</th>
<th>APPLICATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
<td>LOW LEVEL LIBRARIES</td>
</tr>
<tr>
<td>VIRTUALIZATION/ ORCHESTRATION</td>
<td>VIRTUALIZATION/ ORCHESTRATION</td>
</tr>
<tr>
<td>OS</td>
<td>OS</td>
</tr>
<tr>
<td>DRIVERS</td>
<td>DRIVERS</td>
</tr>
<tr>
<td>FW IP &amp; BIOS</td>
<td>FW IP &amp; BIOS</td>
</tr>
<tr>
<td>CPU SCALAR</td>
<td>GPU VECTOR</td>
</tr>
<tr>
<td>AI MATRIX</td>
<td>FPGA SPATIAL/FF</td>
</tr>
<tr>
<td>INTERCONNECT</td>
<td>MEMORY</td>
</tr>
</tbody>
</table>
Abstraction Required at Multiple Layers
## Abstraction Required at Multiple Layers

<table>
<thead>
<tr>
<th>SERVICES &amp; SOLUTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPLICATIONS</td>
</tr>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
</tr>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
</tr>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
</tr>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
</tr>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
</tr>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
</tr>
</tbody>
</table>

## LEVEL ZERO

VIRTUALIZATION/ ORCHESTRATION

OS

DRIVERS

## LEVEL SUBZERO

FW IP & BIOS

<table>
<thead>
<tr>
<th>CPU SCALAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU VECTOR</td>
</tr>
<tr>
<td>AI MATRIX</td>
</tr>
<tr>
<td>FPGA SPATIAL/FF</td>
</tr>
<tr>
<td>INTERCONNECT</td>
</tr>
<tr>
<td>MEMORY</td>
</tr>
</tbody>
</table>
Abstraction Required at Multiple Layers

<table>
<thead>
<tr>
<th>SERVICES &amp; SOLUTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPLICATIONS</td>
</tr>
<tr>
<td>LEVEL ONE</td>
</tr>
<tr>
<td>MIDDLEWARE FRAMEWORKS AND RUNTIMES</td>
</tr>
<tr>
<td>LOW LEVEL LIBRARIES</td>
</tr>
<tr>
<td>LEVEL ZERO</td>
</tr>
<tr>
<td>VIRTUALIZATION/ ORCHESTRATION</td>
</tr>
<tr>
<td>OS</td>
</tr>
<tr>
<td>DRIVERS</td>
</tr>
<tr>
<td>LEVEL SUBZERO</td>
</tr>
<tr>
<td>FW, IP, &amp; BIOS</td>
</tr>
</tbody>
</table>

- CPU SCALAR
- GPU VECTOR
- AI MATRIX
- FPGA SPATIAL/FF
- INTERCONNECT
- MEMORY
# oneAPI Abstraction Roadmap

<table>
<thead>
<tr>
<th>SERVICES &amp; SOLUTIONS</th>
<th>APPLICATIONS</th>
<th>MIDDLEWARE FRAMEWORKS AND RUNTIMES</th>
<th>LOW LEVEL LIBRARIES</th>
<th>LEVEL ZERO oneAPI</th>
<th>VIRTUALIZATION/ ORCHESTRATION</th>
<th>OS</th>
<th>DRIVERS</th>
<th>LEVEL SUB ZERO oneAPI</th>
<th>FW IP &amp; BIOS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>oneAPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LEVEL ZERO oneAPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VIRTUALIZATION/ ORCHESTRATION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DRIVERS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LEVEL SUB ZERO oneAPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FW IP &amp; BIOS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **CPU**: SCALAR, VECTOR
- **GPU**: VECTOR
- **FPGA**: SPATIAL/FF
- **AI**: MATRIX
- **INTERCONNECT**
- **MEMORY**

**IN PLANNING**

**RELEASED**
Goal of new HW/SW Contract

No Transistor Left Behind
From Sensors to Supercomputers - 2021

- Sensors
- Edge
- Data Center

- <1M Neurons
- Tera FLOPS
- Peta FLOPS
- Exa FLOPS

- mWatts
- Watts
- kWatts
- MWatts

- Single Software Abstraction
From Sensors to Supercomputers - 2025

SENSORS

EDGE

DATA CENTER

SINGLE SOFTWARE ABSTRACTION

>PETA FLOPS

>EXA FLOPS

>ZETTA FLOPS

>1B NEURONS

EXAFLOPS ON THE EDGE ENABLE EXASCALE FOR EVERYONE
Summary

PERFORMANCE

PL ENTY OF ROOM
AT THE TOP

1000x
BY 2025

PL ENTY OF ROOM
AT THE BOTTOM

GENERALITY
EXASCALE FOR EVERYONE
“Optimism is the essential ingredient of innovation”

ROBERT NOYCE
THANK YOU

AND REMEMBER, LEAVE NO TRANSISTOR BEHIND
Legal Disclaimers

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. See backup for configuration details. 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.

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.