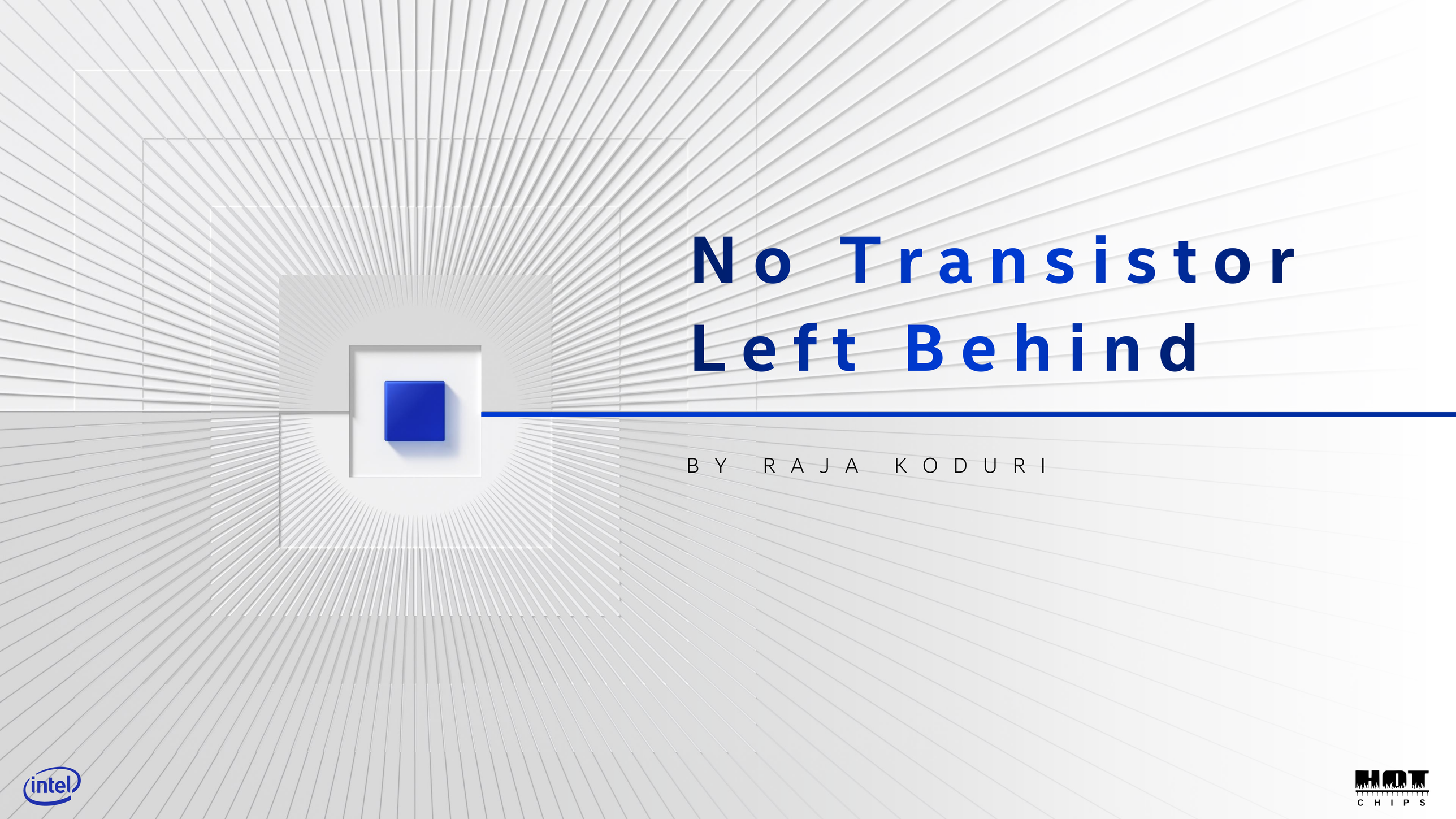


No Transistor Left Behind

BY RAJA KODURI

intel

HOT CHIPS

A tribute to

Frances Allen

“Who Helped Hardware Understand Software”

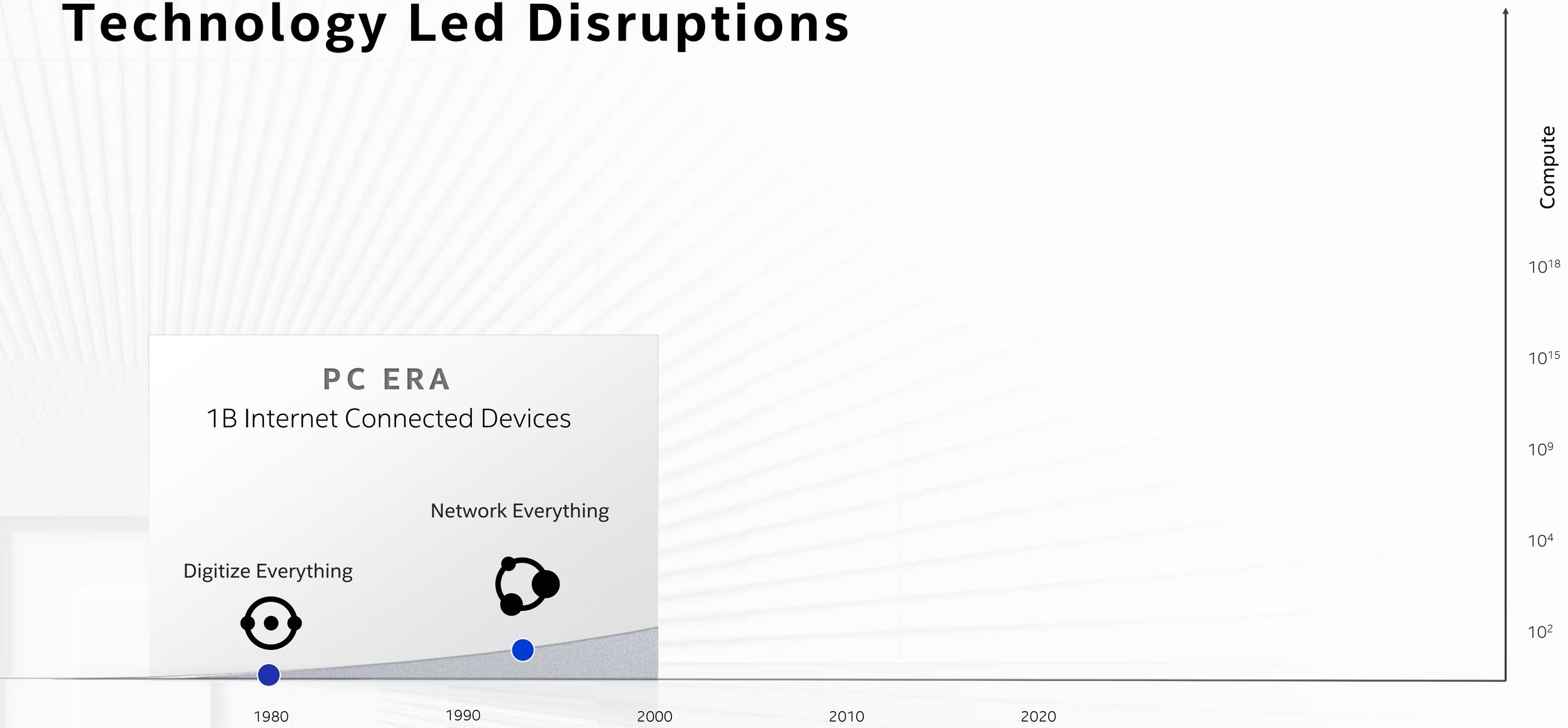
“No Transistor Left Behind”

DAVID BLYTHE, 2018

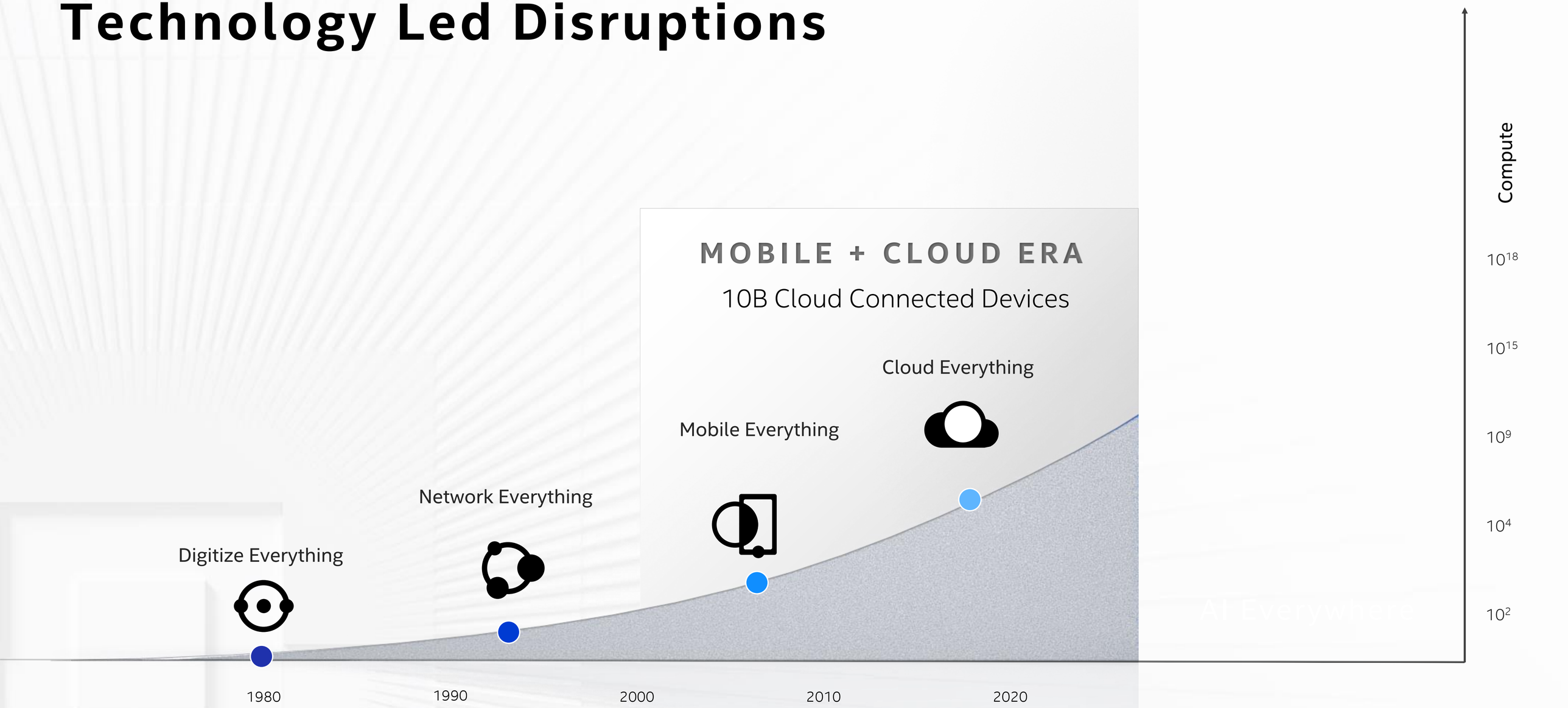
intel

HOT CHIPS

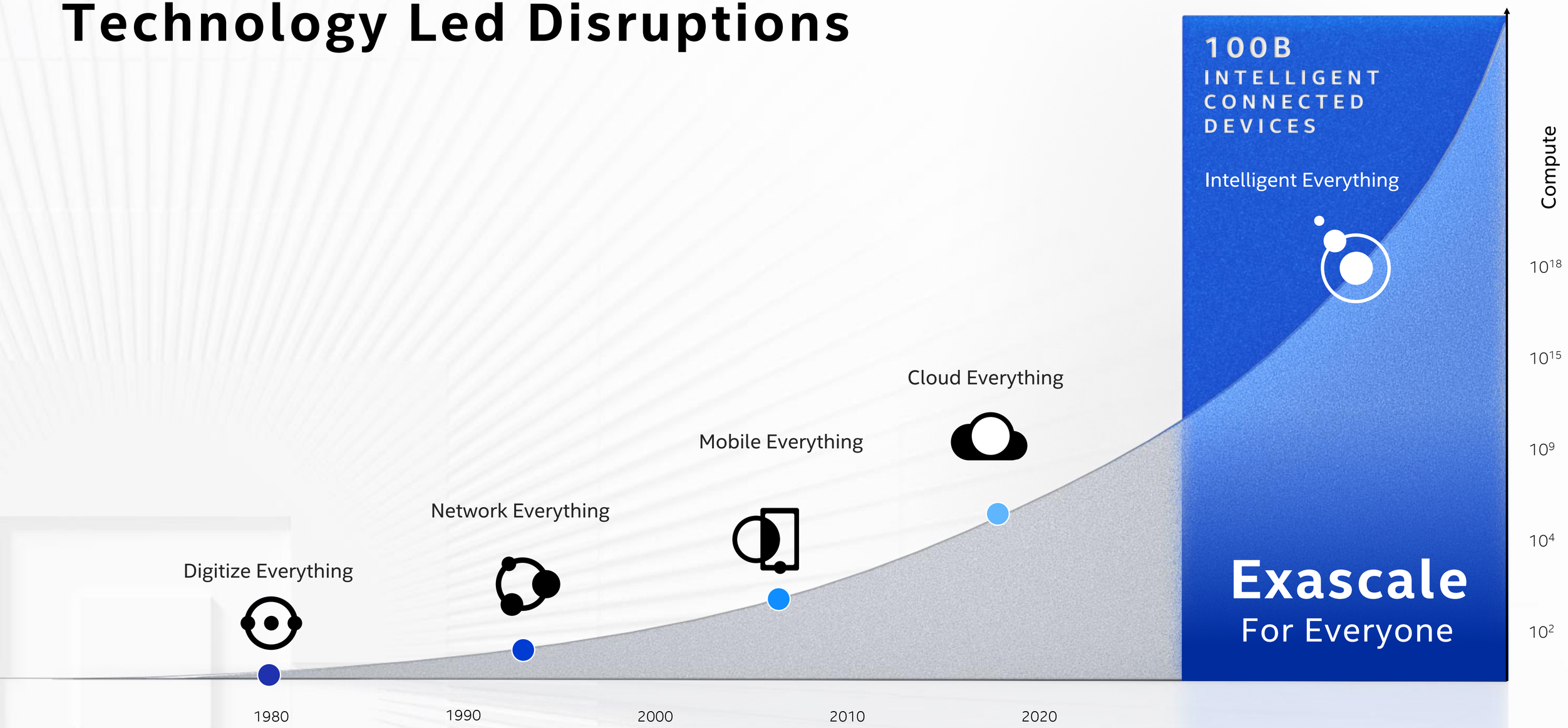
Technology Led Disruptions



Technology Led Disruptions

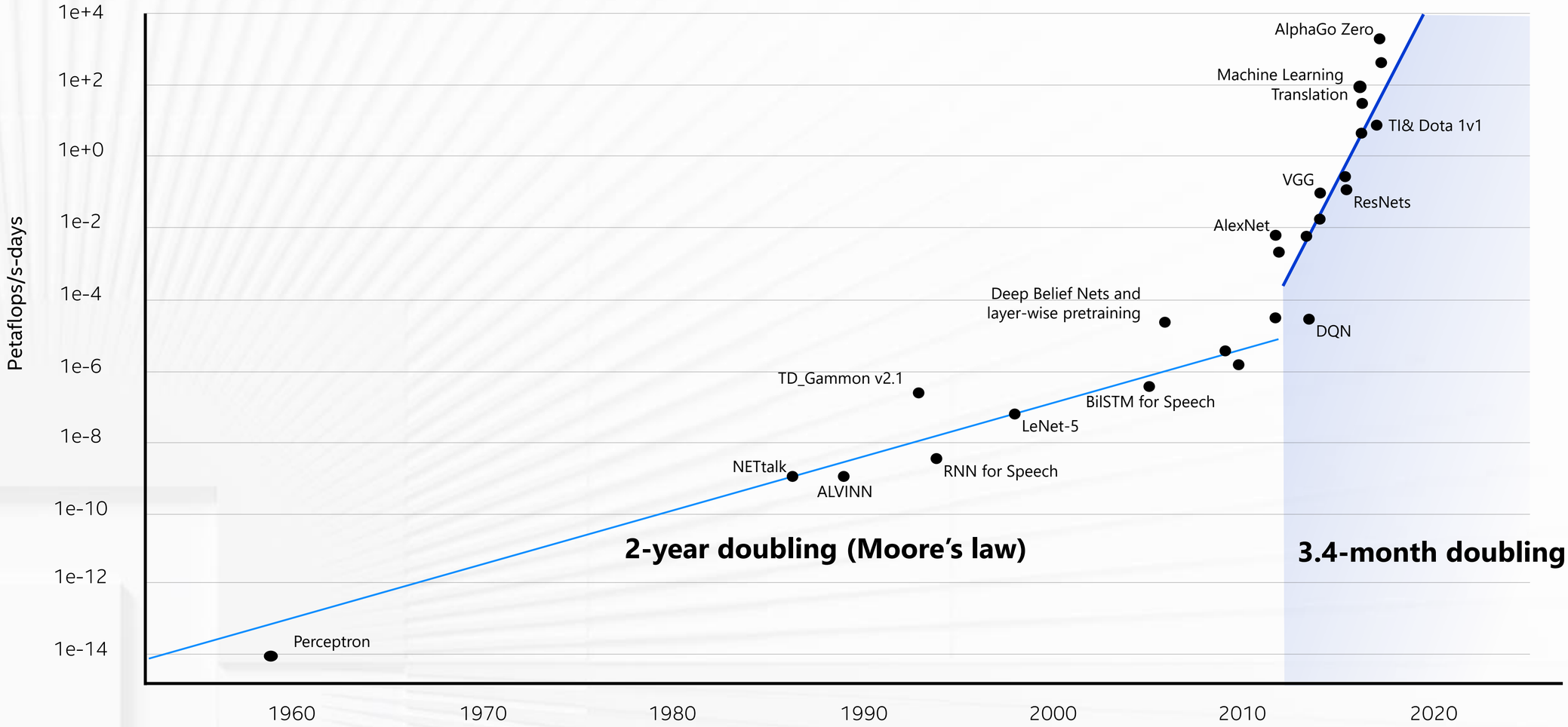


Technology Led Disruptions



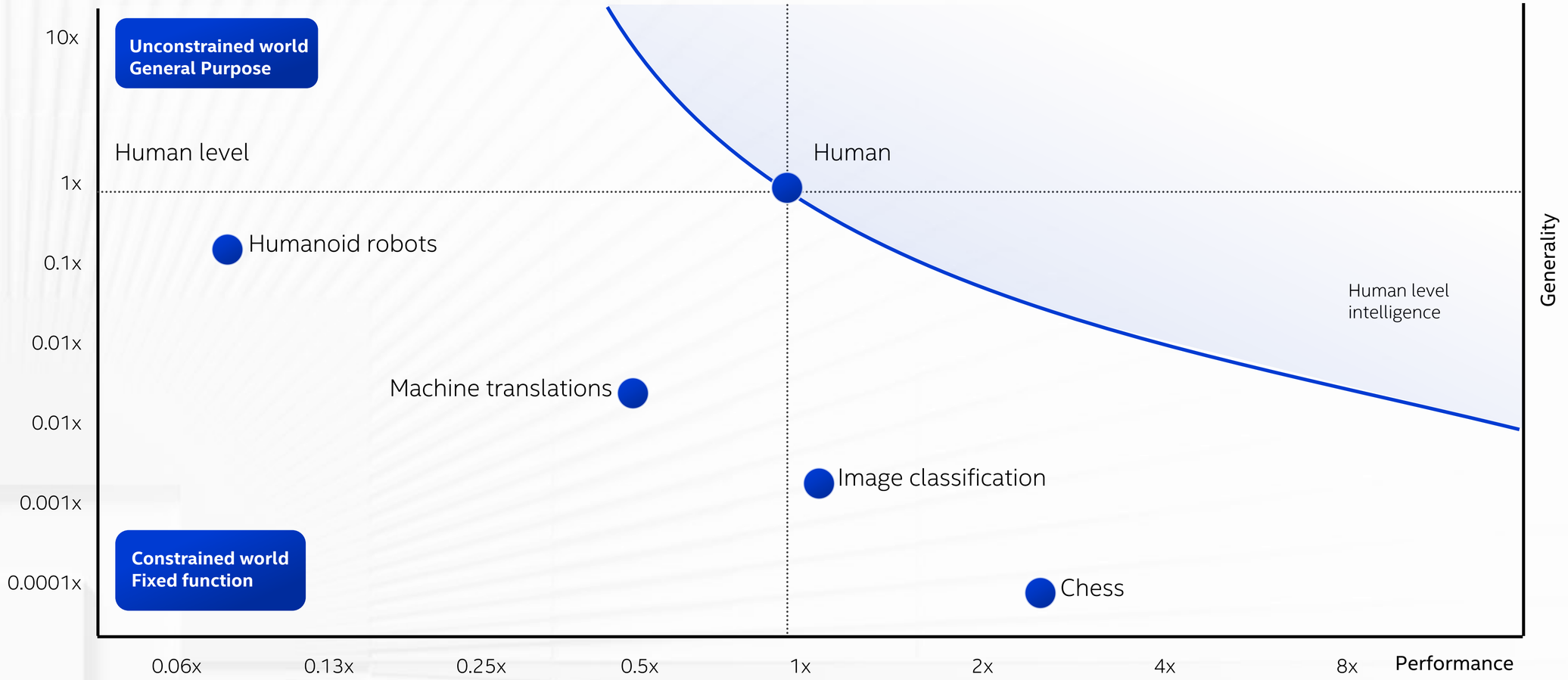
Intelligence is Expensive

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



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

Performance and Generality

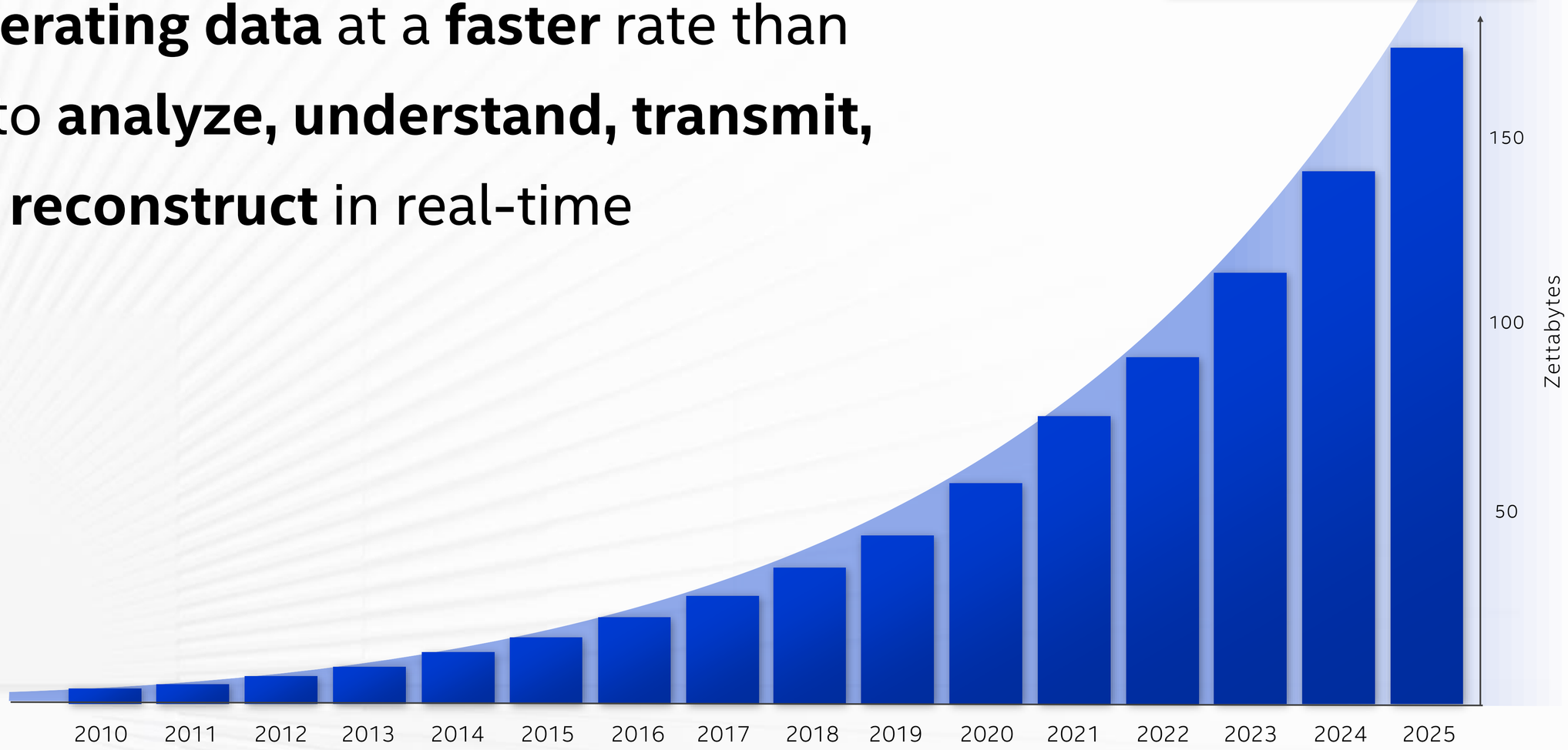


Rohrer, B., 2018. *Getting Closer To Human Intelligence Through Robotics*. Youtube. <https://www.youtube.com/watch?v=0ILdnZmp3Iw&t=2095s>

*Graph: Illustrative purposes only

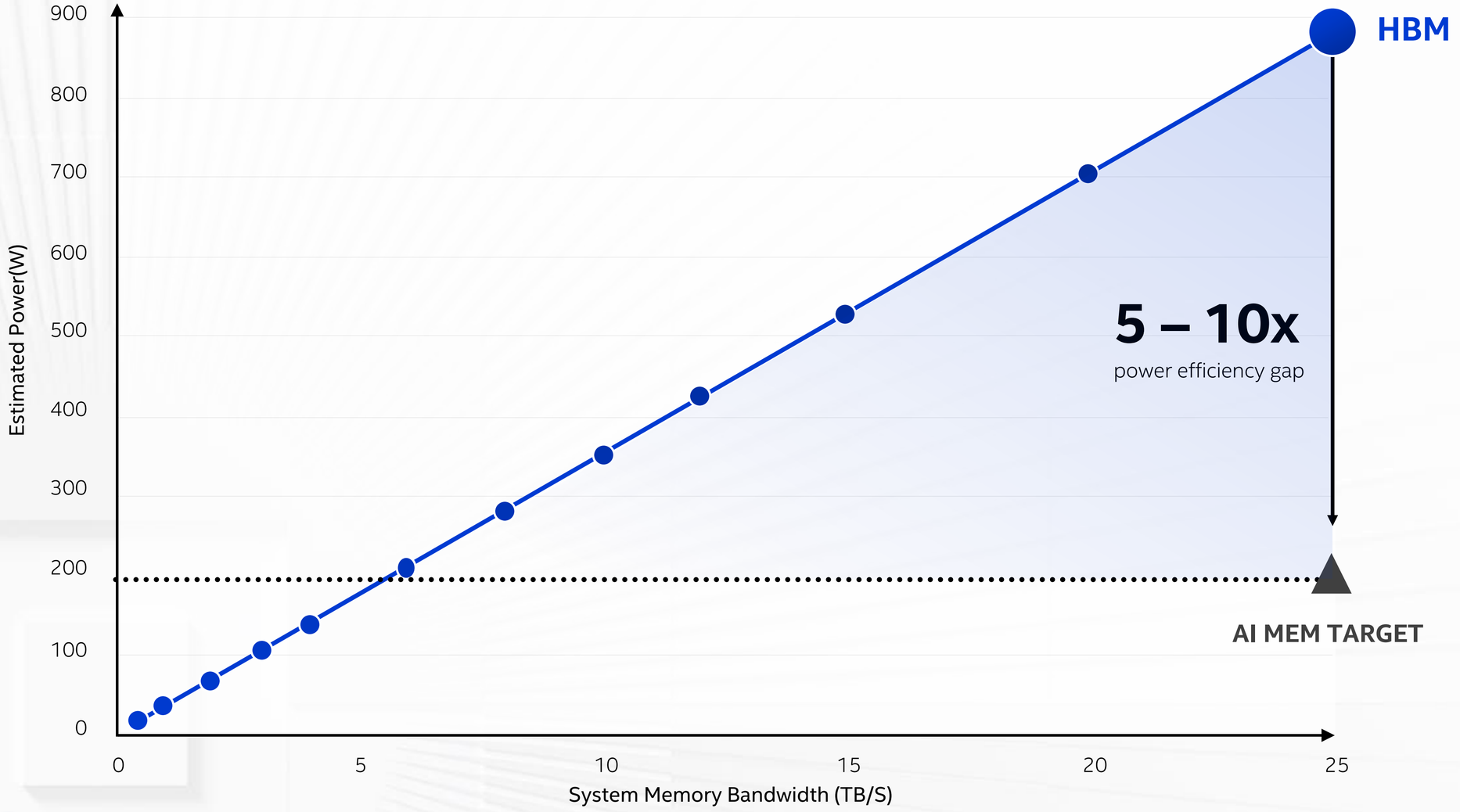
Data is Exploding

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



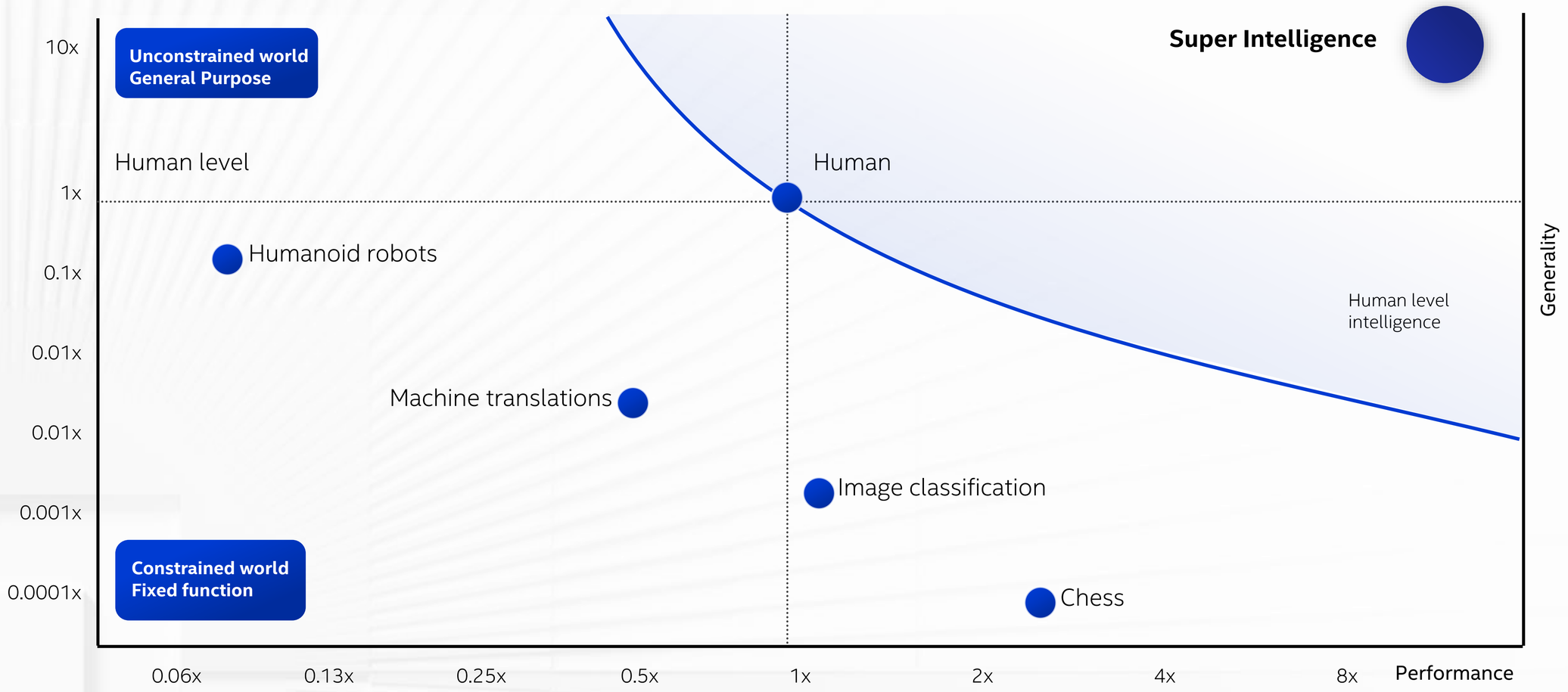
Source: IDC Global DataSphere, May 2020

Memory Wall



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

Data and Super Intelligence



Rohrer, B., 2018. *Getting Closer To Human Intelligence Through Robotics.* https://www.youtube.com/watch?v=0ILdnZmp3Iw&t=2095s

*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”

PETER LEE – VP MICROSOFT RESEARCH

“Everything that can be invented has been invented”

Charles H. Duell - US Patent Commissioner

1899

“Moore’s law won’t work at feature sizes less than a quarter of a micron”

Erich Bloch – Head of IBM Research, later Chairman of NSF

1988

Moore Sees ‘Moore’s Law’ Dead in a Decade

2007

By Mark Hachman on September 18, 2007 at 5:12 pm | 1 Comment

In a “fireside chat” with NPR “Tech Nation’s” Moira Gunn, Intel co-founder and chairman emeritus Gordon Moore said he sees his famous law expiring in 10 to 15 years.

1,901 views | Apr 29, 2010, 01:37pm

2010

Life After Moore's Law

By Bill Dally

Bill Dally is the chief scientist and senior vice president of research at NVIDIA and the Willard R. and Inez Kerr Bell Professor of Engineering at Stanford University.

Report: IBM researcher says Moore's Law at end

IBM Fellow Carl Anderson says at a conference this week that Moore's Law is hitting a ceiling, according to a report.

BY BROOKE CROTHERS | APRIL 10, 2009 7:06 AM PDT

2009

Theoretical physicist: Moore's Law has just 10 years to go

The age of silicon will come to a close but nobody knows when. Well, almost nobody.

2012

NEWS

Death of Moore's Law Will Cause Economic Crisis

"Around 2020 or soon afterward, Moore's law will gradually cease to hold true and Silicon Valley may slowly turn into a rust belt unless a replacement technology is found," says Kaku in an extract published on Salon.com website.

By John E Dunn

Techworld.com | MARCH 21, 2011 12:28 PM PT

2011

The End of Moore's Law?

The current economic boom is likely due to increases in computing speed and decreases in price. Now there are some good reasons to think that the party may be ending.

2000

by **Charles C. Mann**

“There is nothing new to be discovered in physics now”

Lord Kelvin

1900

SCI-TECH

Moore's Law limit hit by 2014?

The high cost of semiconductor manufacturing equipment is making continued chipmaking advancements too expensive, threatening Moore's Law, according to iSuppli.

BY BROOKE CROTHERS | JUNE 16, 2009 12:48 PM PDT

2009

Moore’s Law Is Dead. Now What?

Shrinking transistors have powered 50 years of advances in computing—but now other ways must be found to make computers more capable.

by **Tom Simonite**

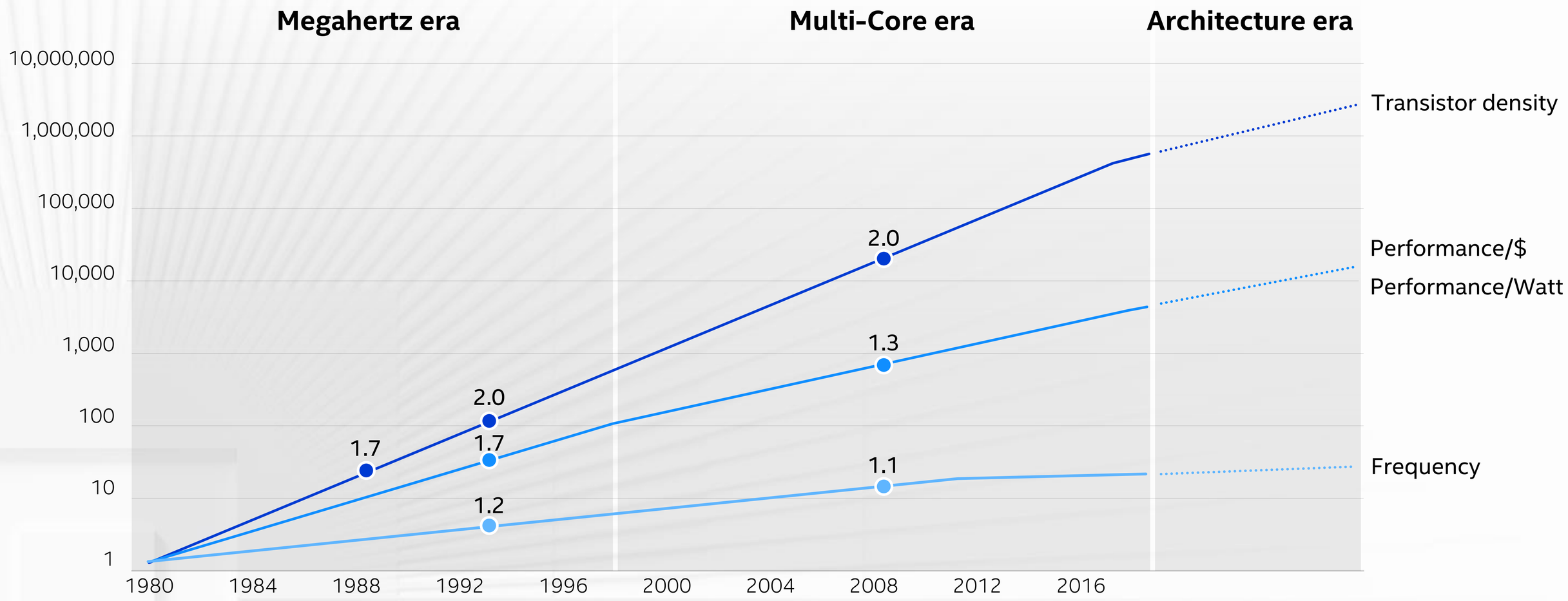
2016

May 13, 2016

intel

HOT CHIPS

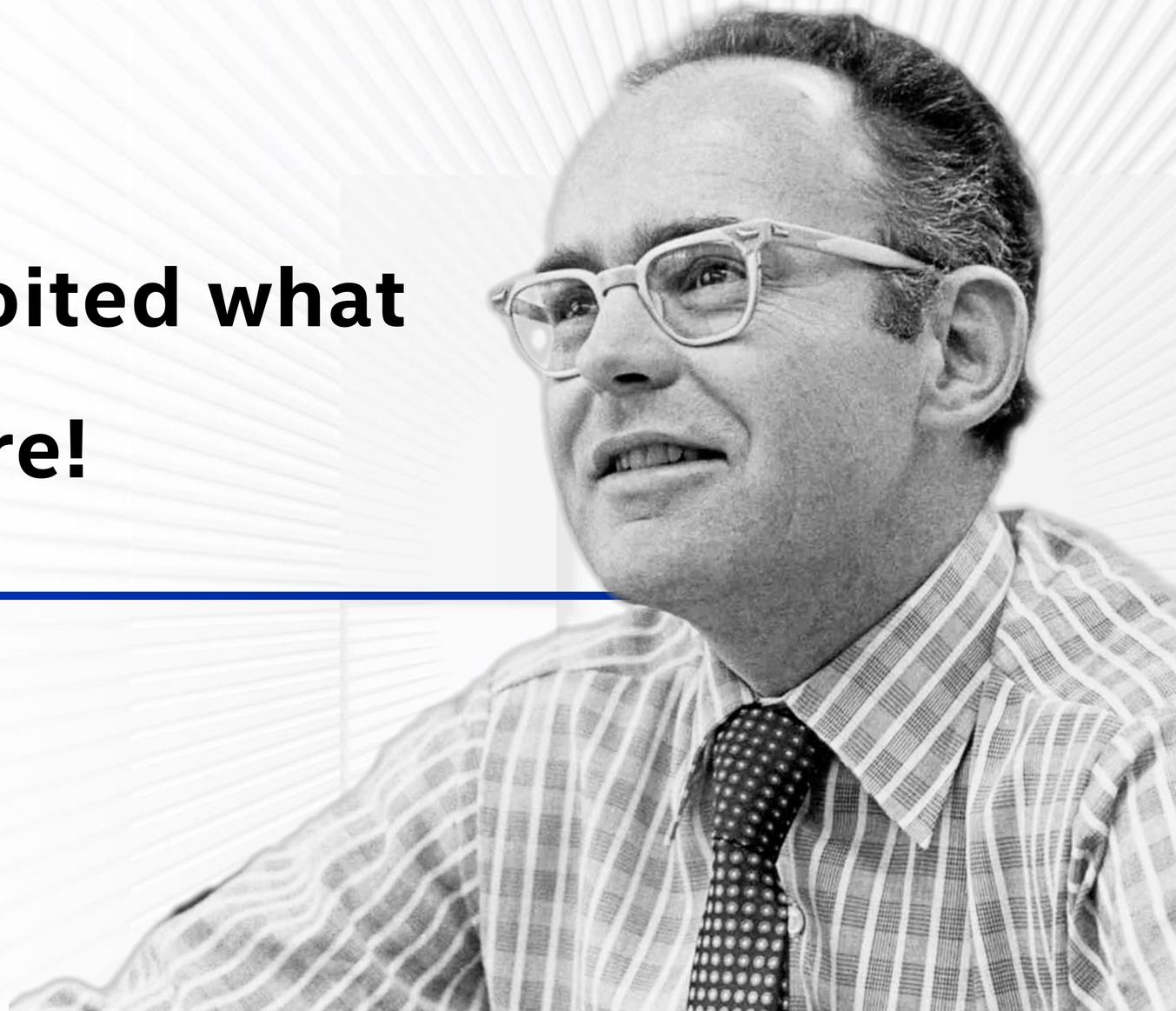
Moore's Law - Our Exponential Entitlement



Based on Intel Internal Data

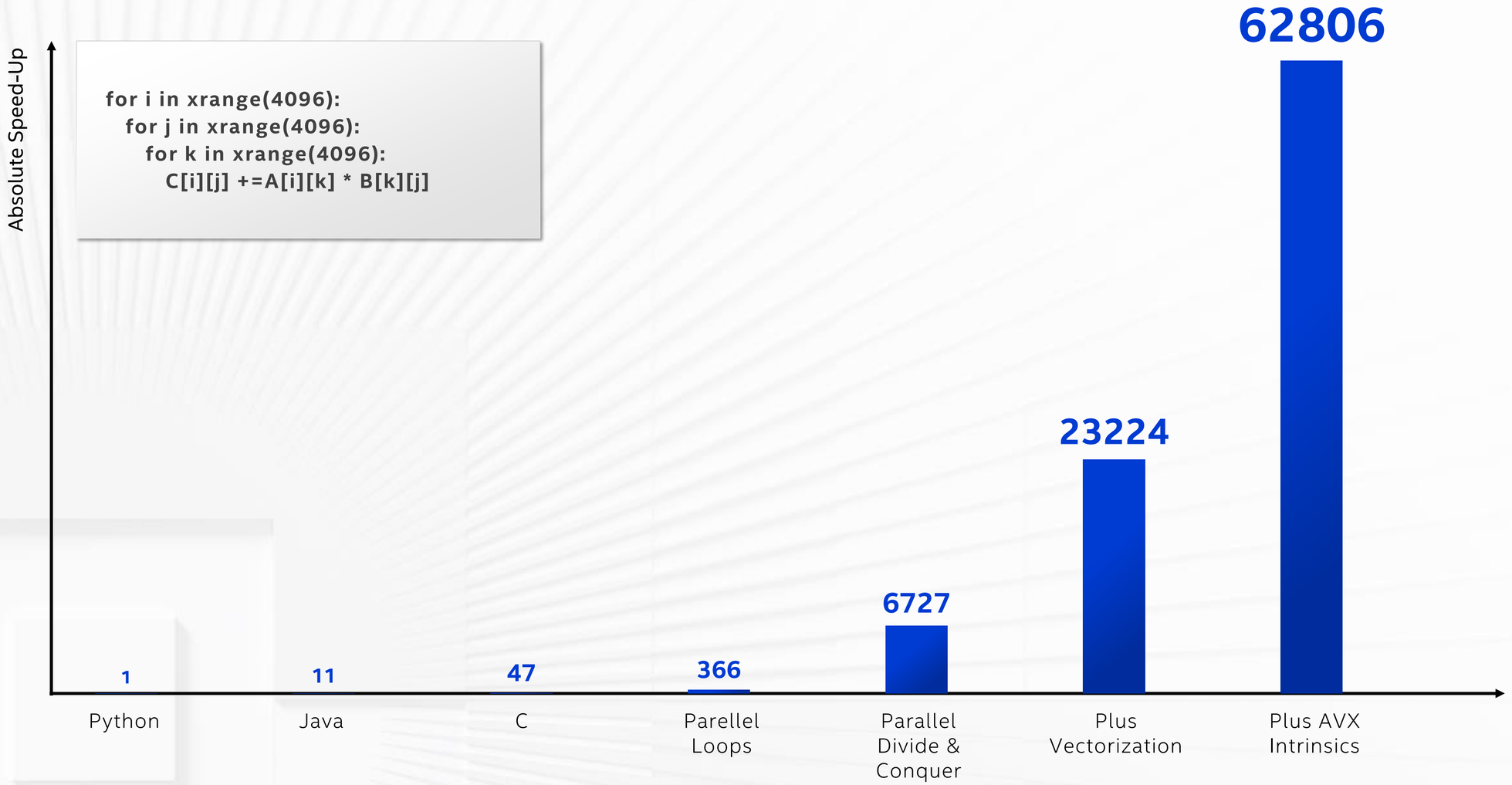
HOT CHIPS

… we haven’t fully exploited what was given to us by Moore!



Software Productivity vs. Hardware Efficiency

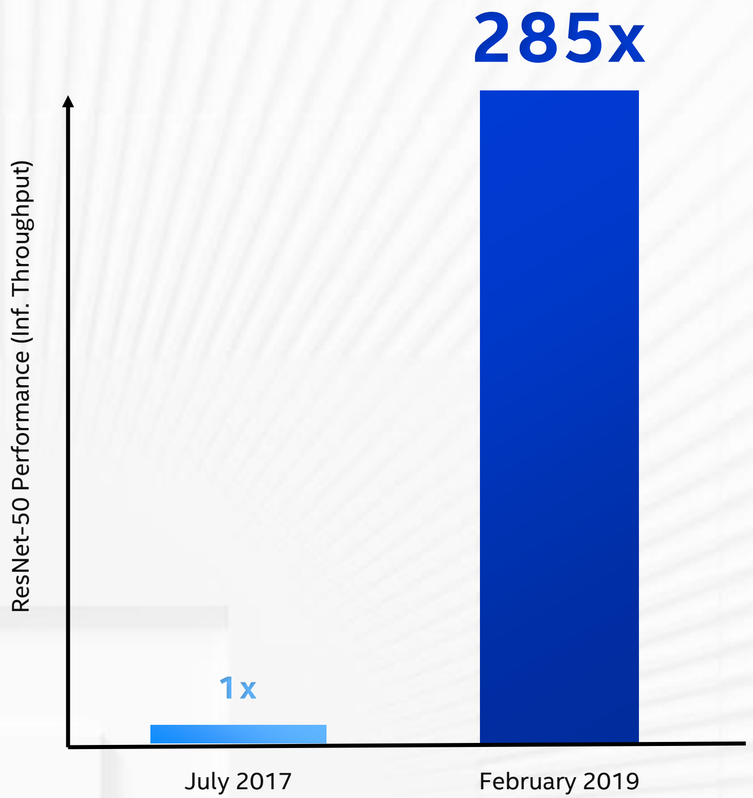
```
for i in xrange(4096):
  for j in xrange(4096):
    for k in xrange(4096):
      C[i][j] +=A[i][k] * B[k][j]
```



Leiserson, C. E., Thompson, N. C., Emer, J. S., Kuszmaul, B. C., Lampson, B. W., Sanchez, D., & Schardl, T. B. (2020). There's plenty of room at the Top: What will drive computer performance after Moore's law?. Science, 368(6495).

Training & Inference Software Speed-Up

On Intel Xeon



2S Intel® Xeon® Scalable Processor (Skylake)

INFERENCE LATENCY

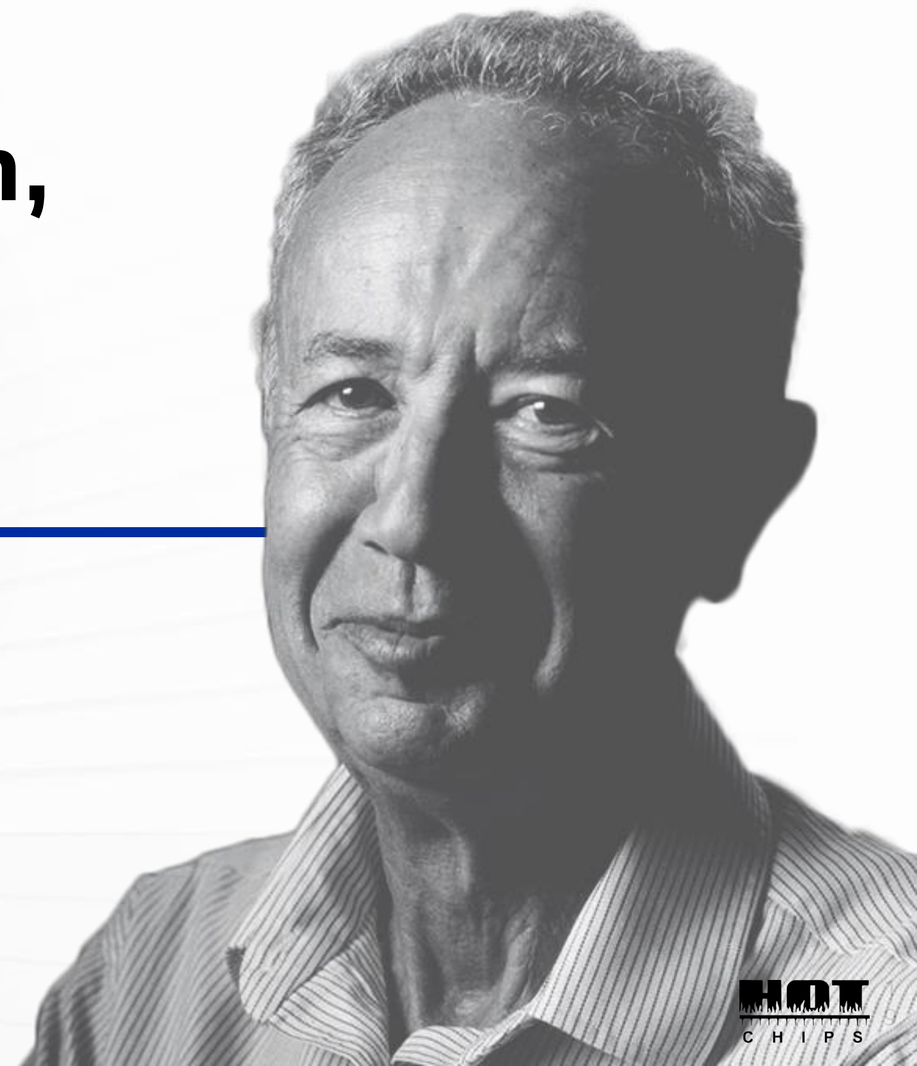
DOWN FROM **SECONDS** TO **MILLISECONDS**

TRAINING TIME

DOWN FROM **WEEKS** TO **HOURS/MINUTES**

285x inference throughput improvement with Intel® Optimizations for Caffe ResNet-50 on Intel® Xeon® Platinum 8180 Processor based on testing as of February 2019 compared to BVLC caffe performance at launch in July 2017. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to https://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

“What Andy Giveth, Bill Taketh Away”



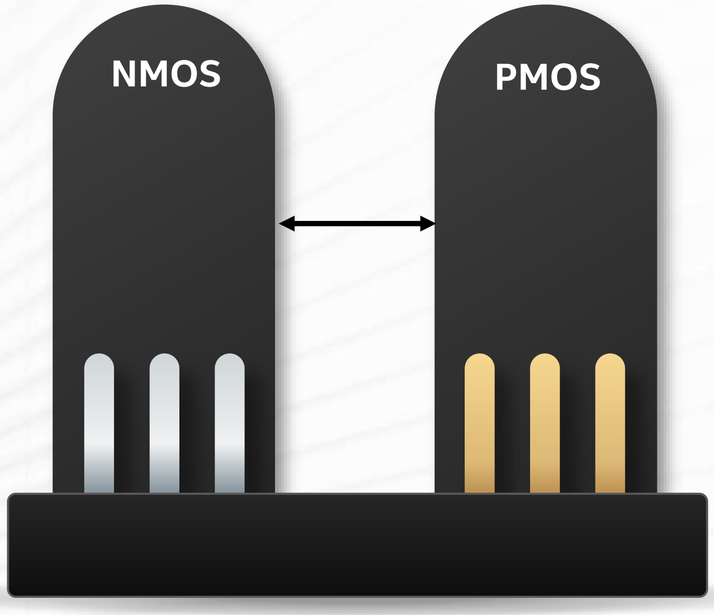
”There’s still plenty of room at the bottom”

JIM KELLER

”There’s still plenty of room at the bottom”

A path to 50x Transistor Density

Intel 10nm



FinFET

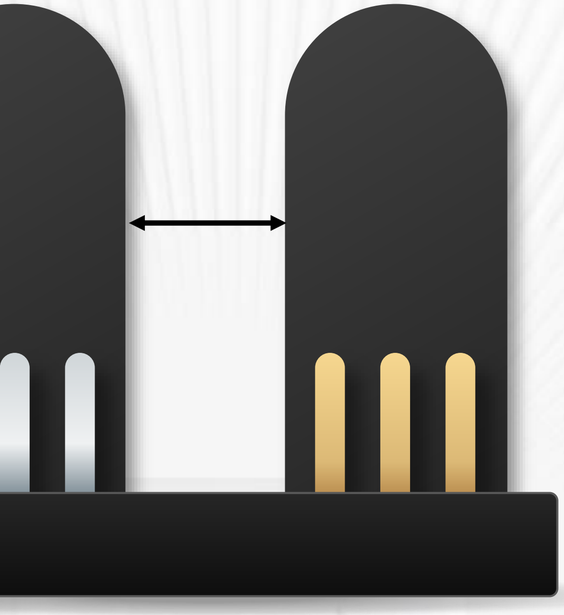
Intel 10nm

DENSITY INCREASE

x3

TOTAL

x3



FinFET

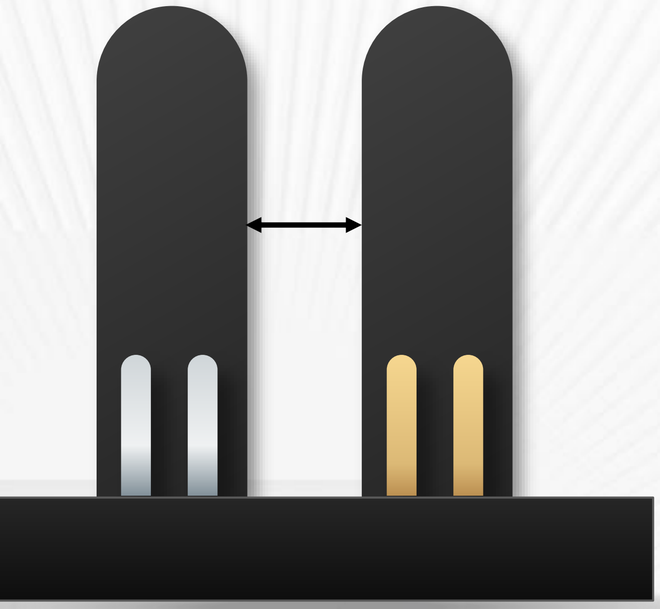
Pitch Scaling

DENSITY INCREASE

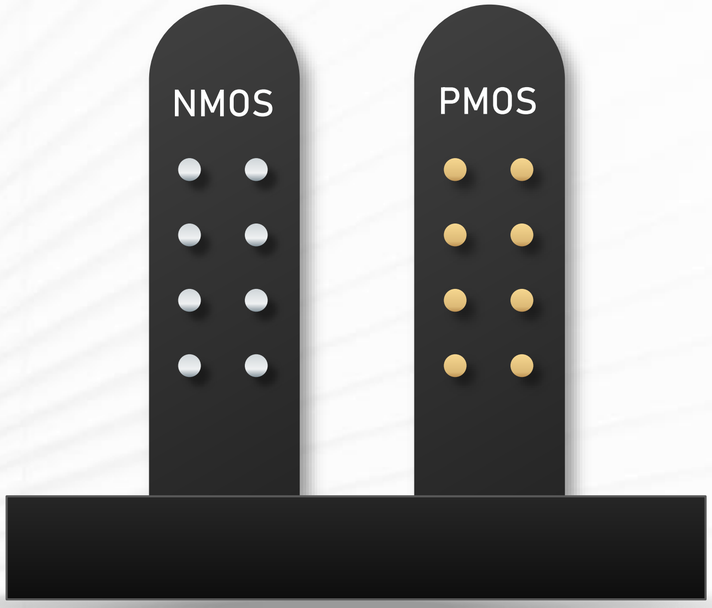
x2

TOTAL

x6



Pitch Scaling



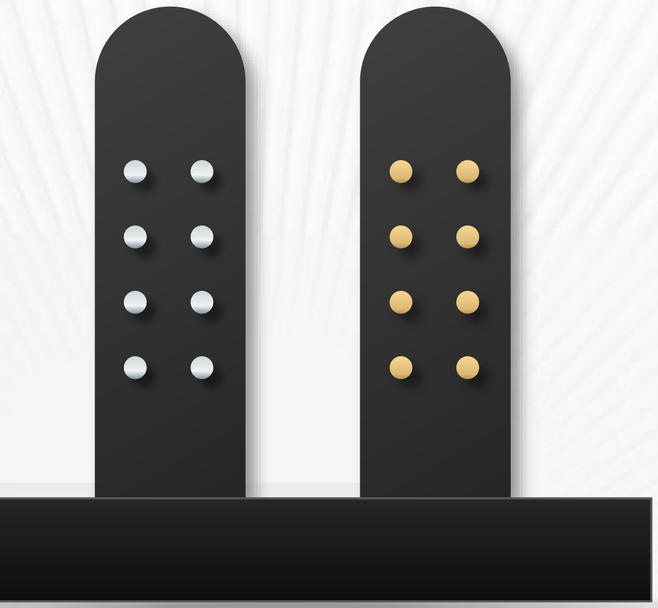
Nanowire

DENSITY INCREASE

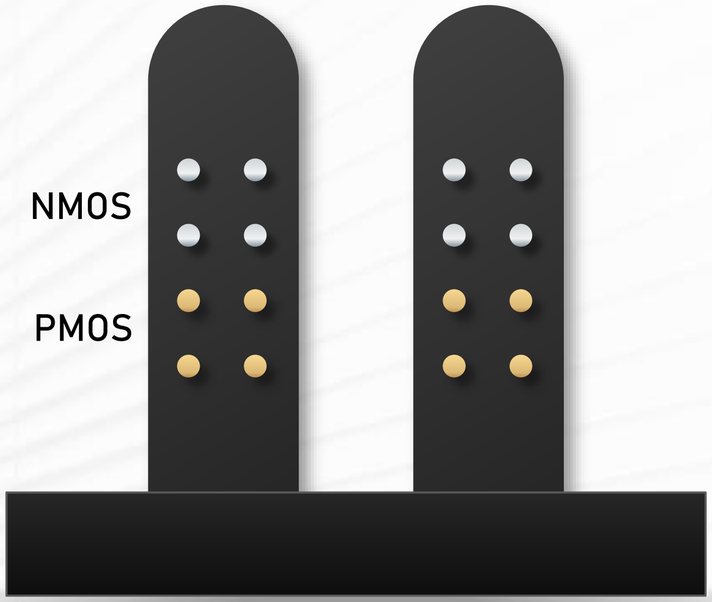
x2

TOTAL

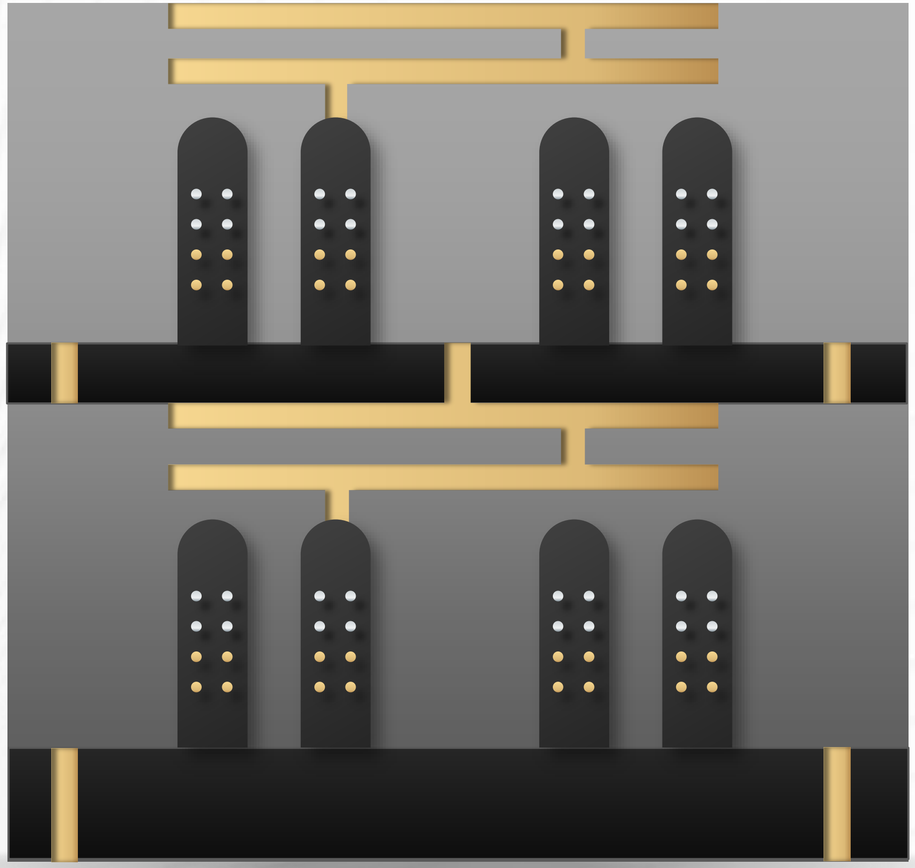
x12



Nanowire



Stacked Nanowire



DENSITY INCREASE

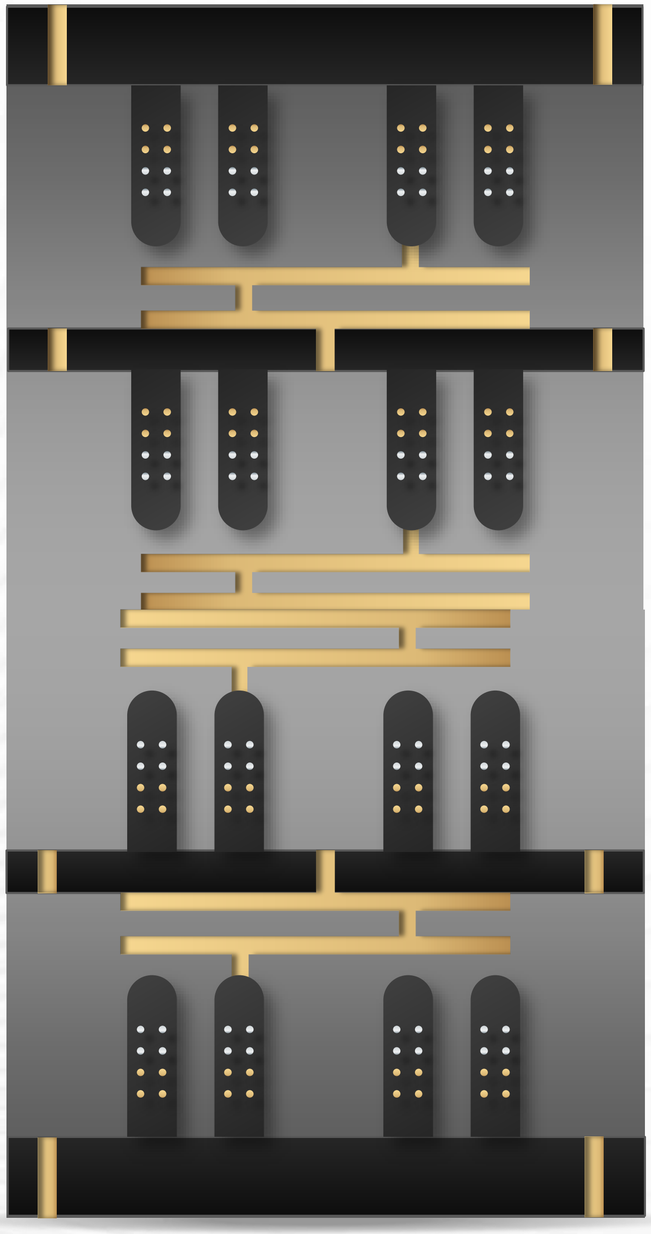
x2

TOTAL

x24

Wafer to Wafer Stacking

Die to Wafer Stacking



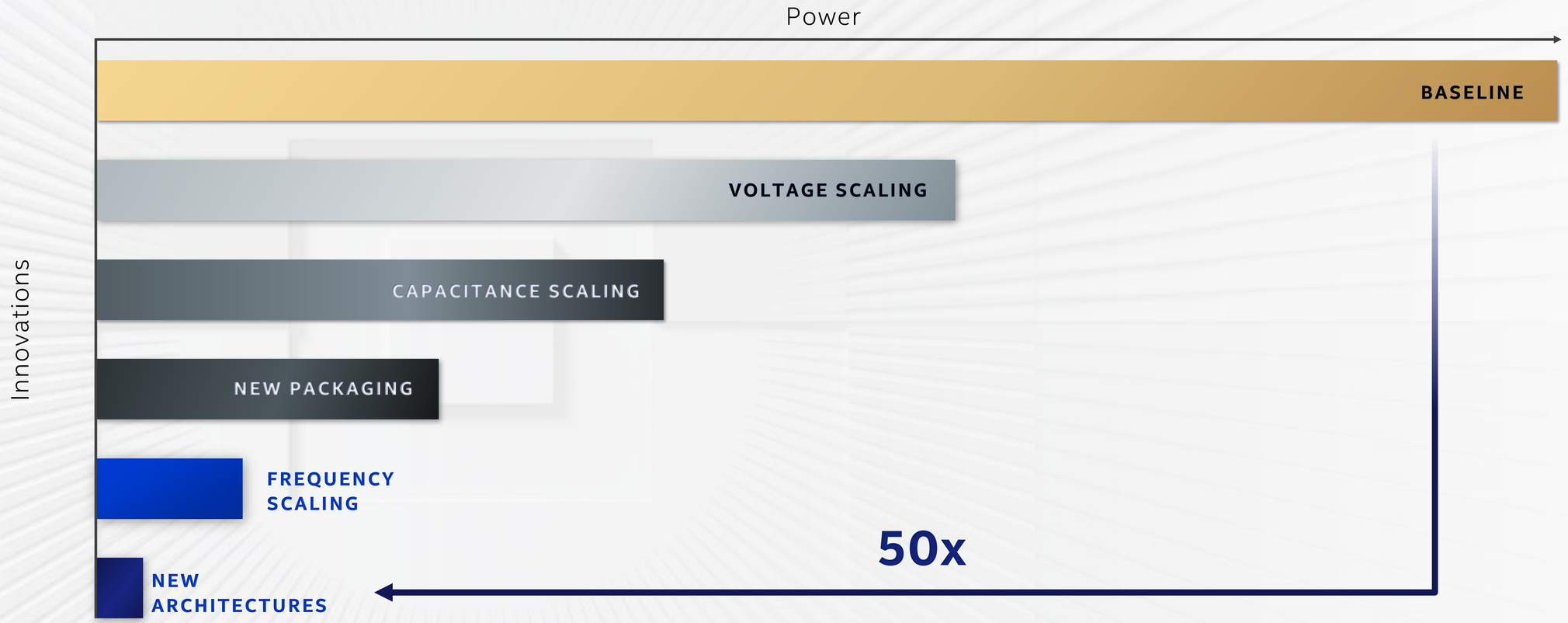
DENSITY INCREASE

x2

TOTAL

~x50

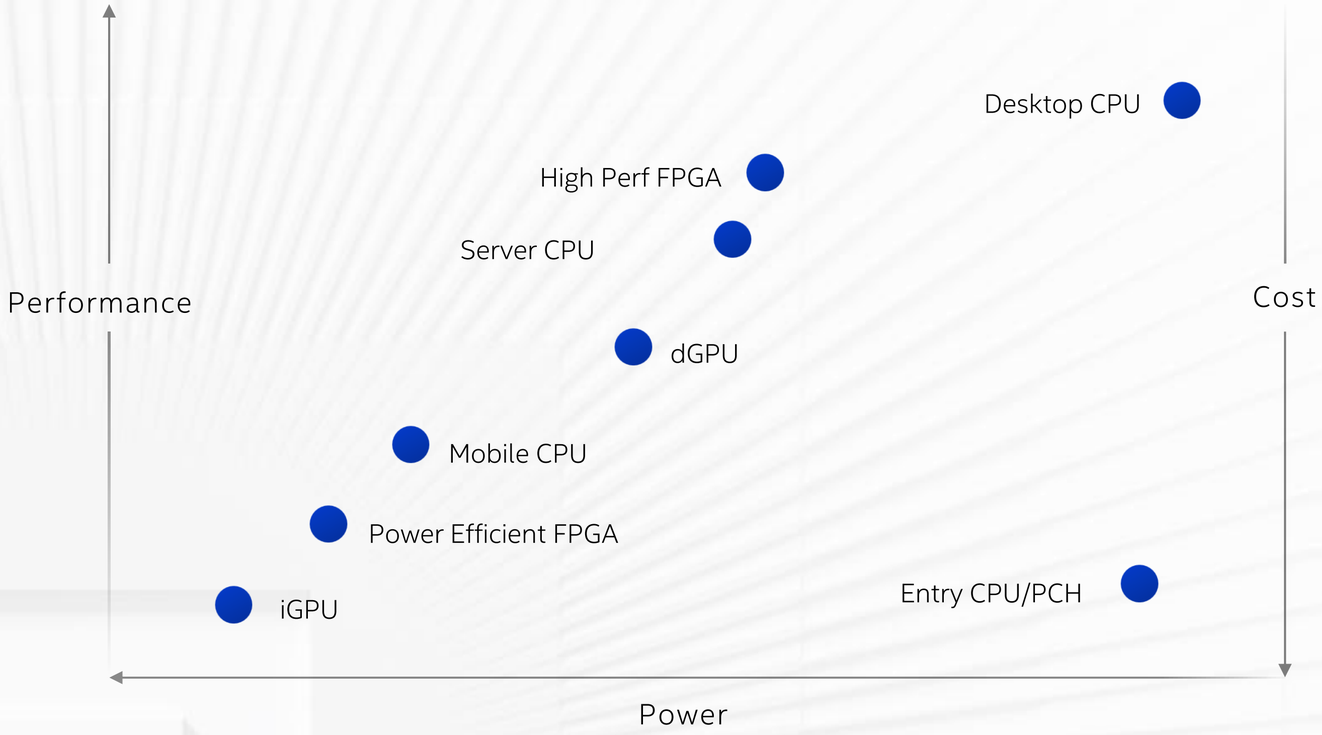
What about Power?



For illustrative purposes only

Case for Advanced Packaging

TRANSISTOR DESIGN TARGET RANGE

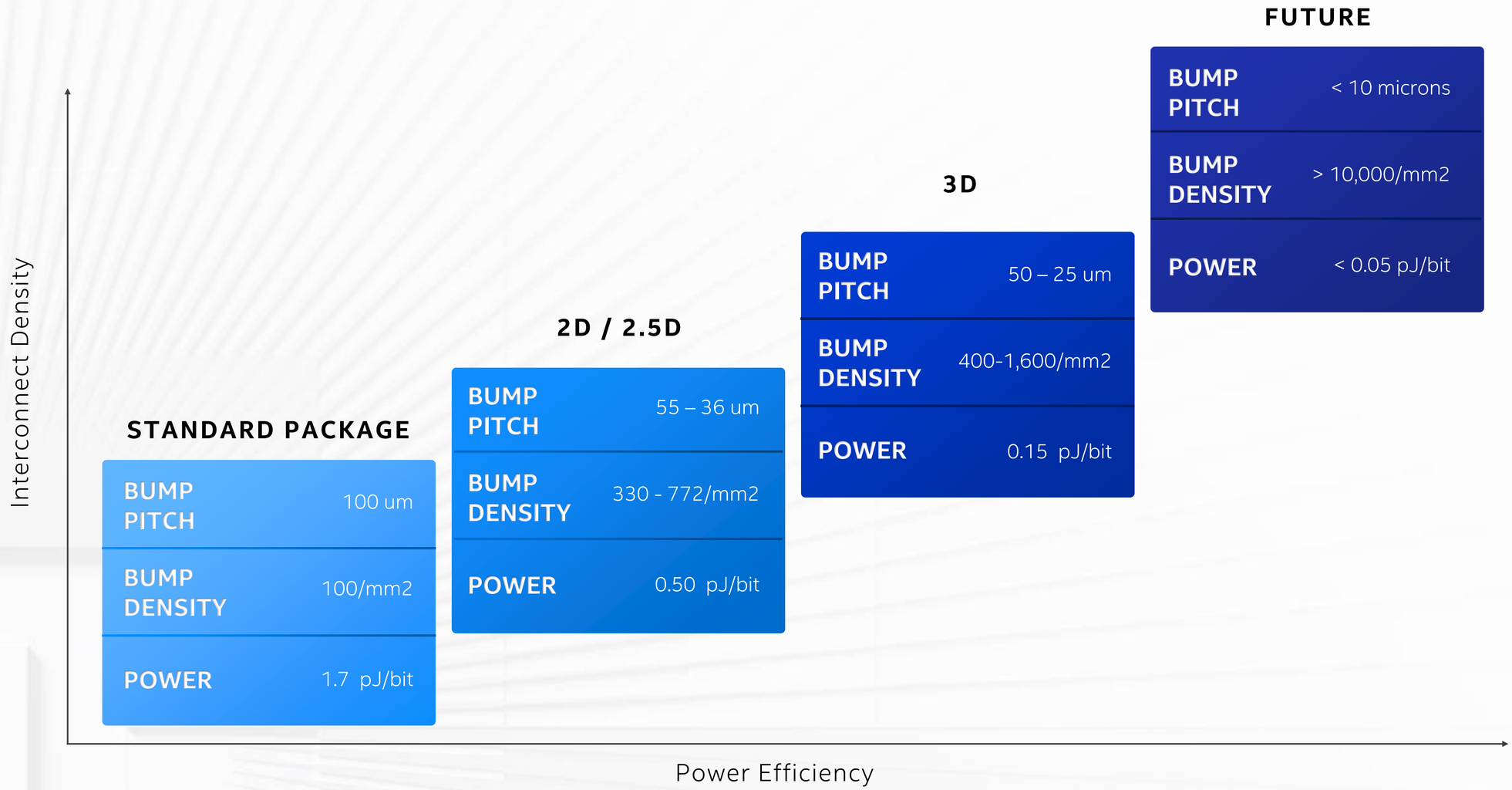


NO SINGLE TRANSISTOR NODE IS OPTIMAL ACROSS ALL DESIGN POINTS!

TRANSISTOR DIVERSITY

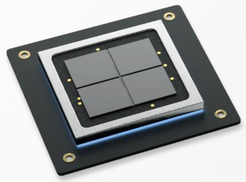
- Logic Transistors
- Analog/RF Transistors
- High Speed Memory
- Dense Memory
- Non-Volatile Memory
- High speed IO
- Configuration Memory

Packaging Technology Improvements

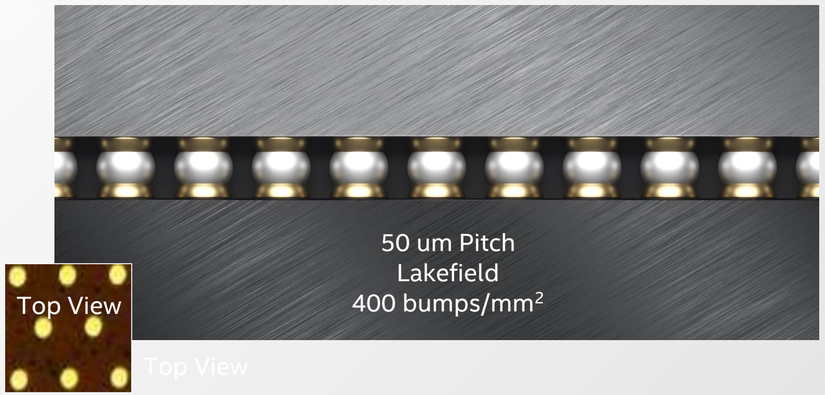


Hybrid Bonding

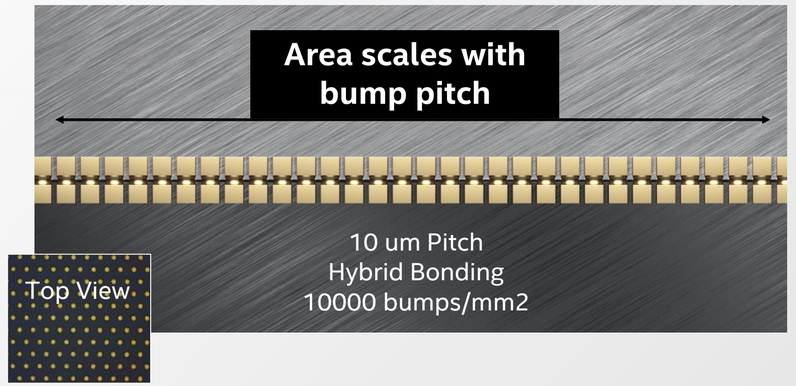
Dense vertical interconnects



Foveros Technology



Hybrid Bonding

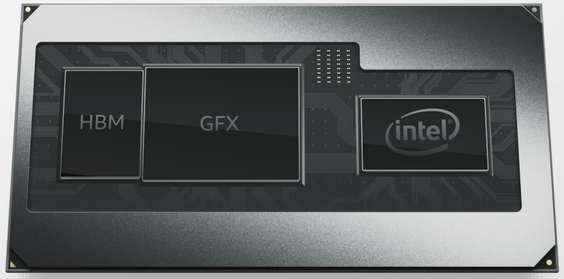


- Smaller, simpler circuits
- Lower capacitance
- Lower power

Advanced Packaging Products

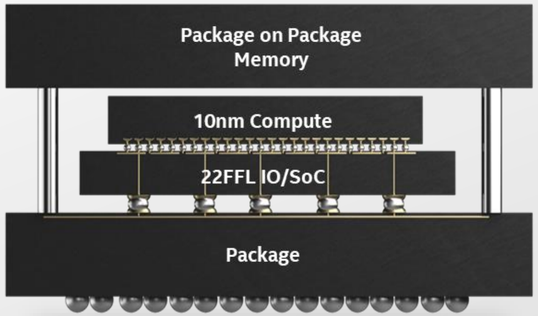
KABY LAKE G
2D

Intel CPU
+ AMD GFX
+ HBM



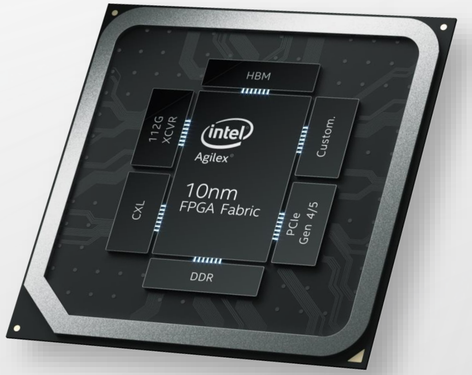
LAKEFIELD
3D

Internal Silicon on
Multiple Nodes



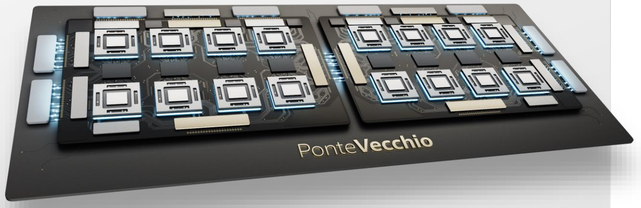
AGILEX FPGA
2.5D

Intel FPGA
+ Foundry IO Chiplets
+ HBM

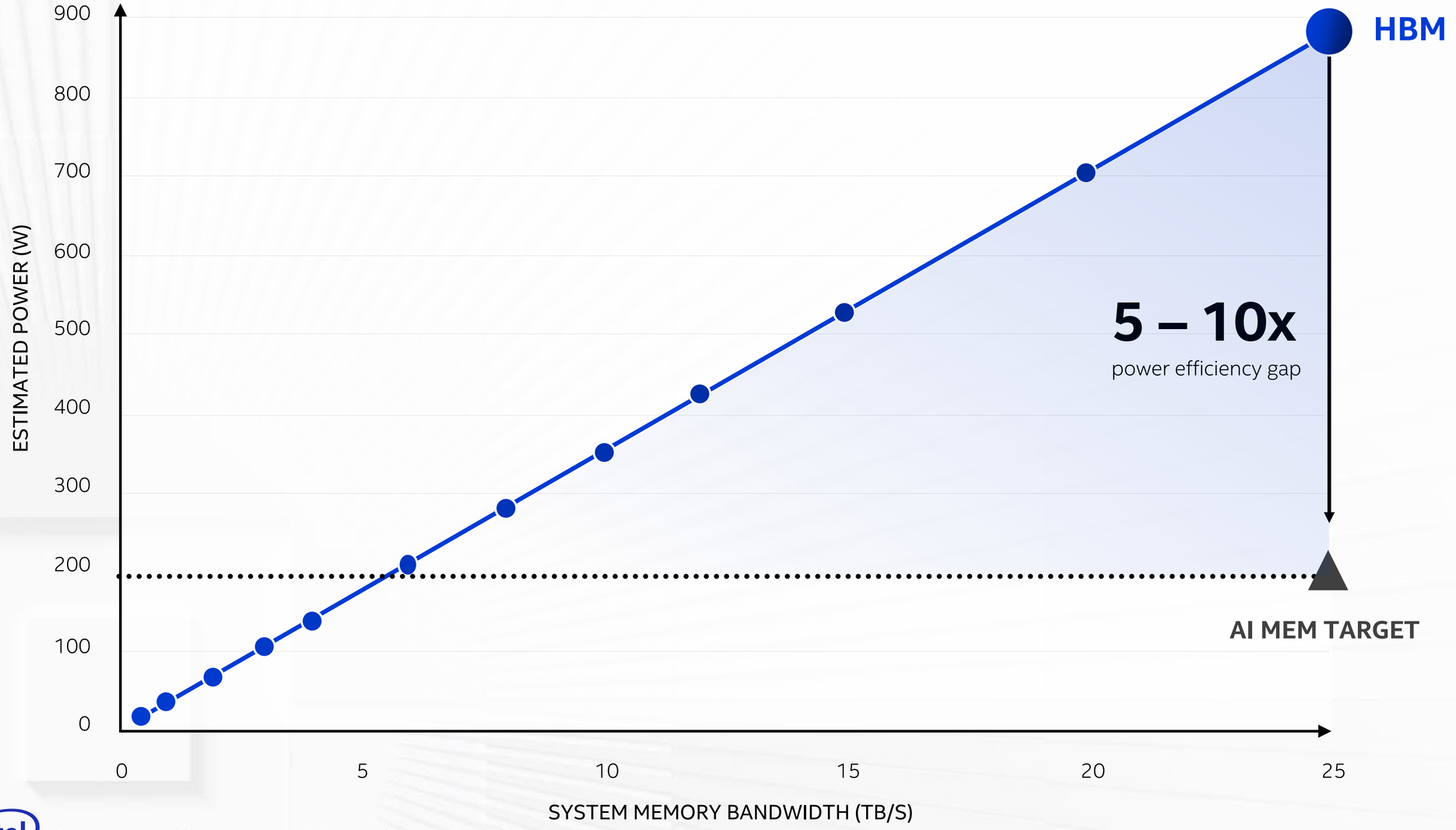


PONTE VECCHIO
3D

Internal and External
Silicon on Multiple Nodes

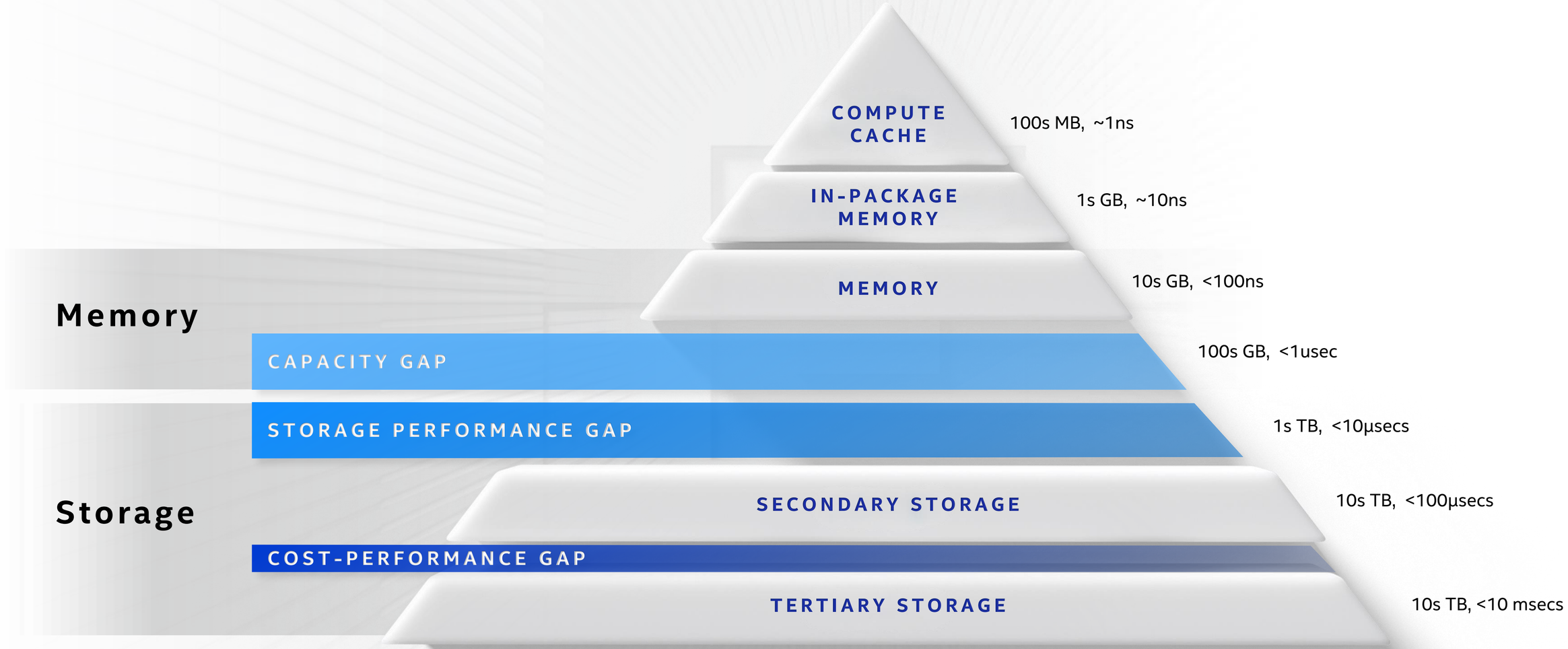


Memory Wall

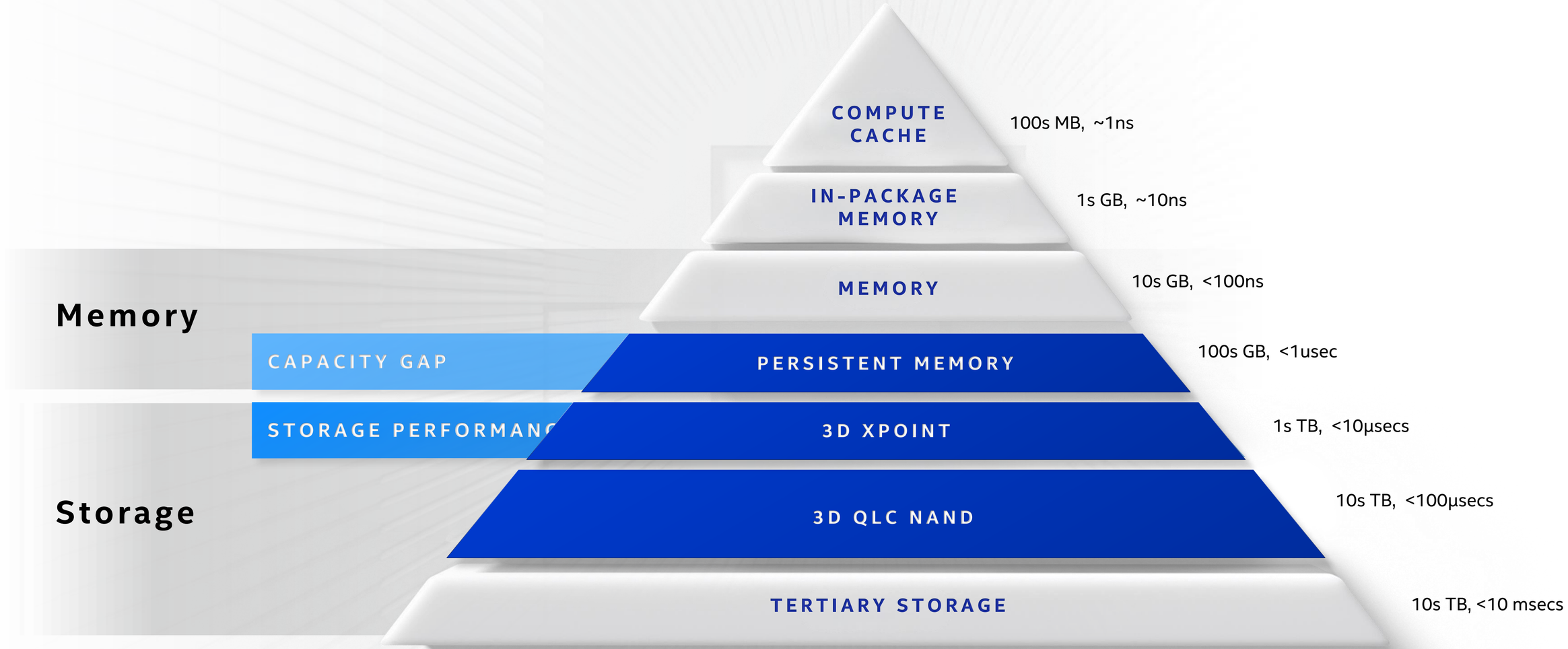


Source: http://research.nvidia.com/publication/2017-02_architecting-an-energy-efficient

Memory Hierarchy Disruptions

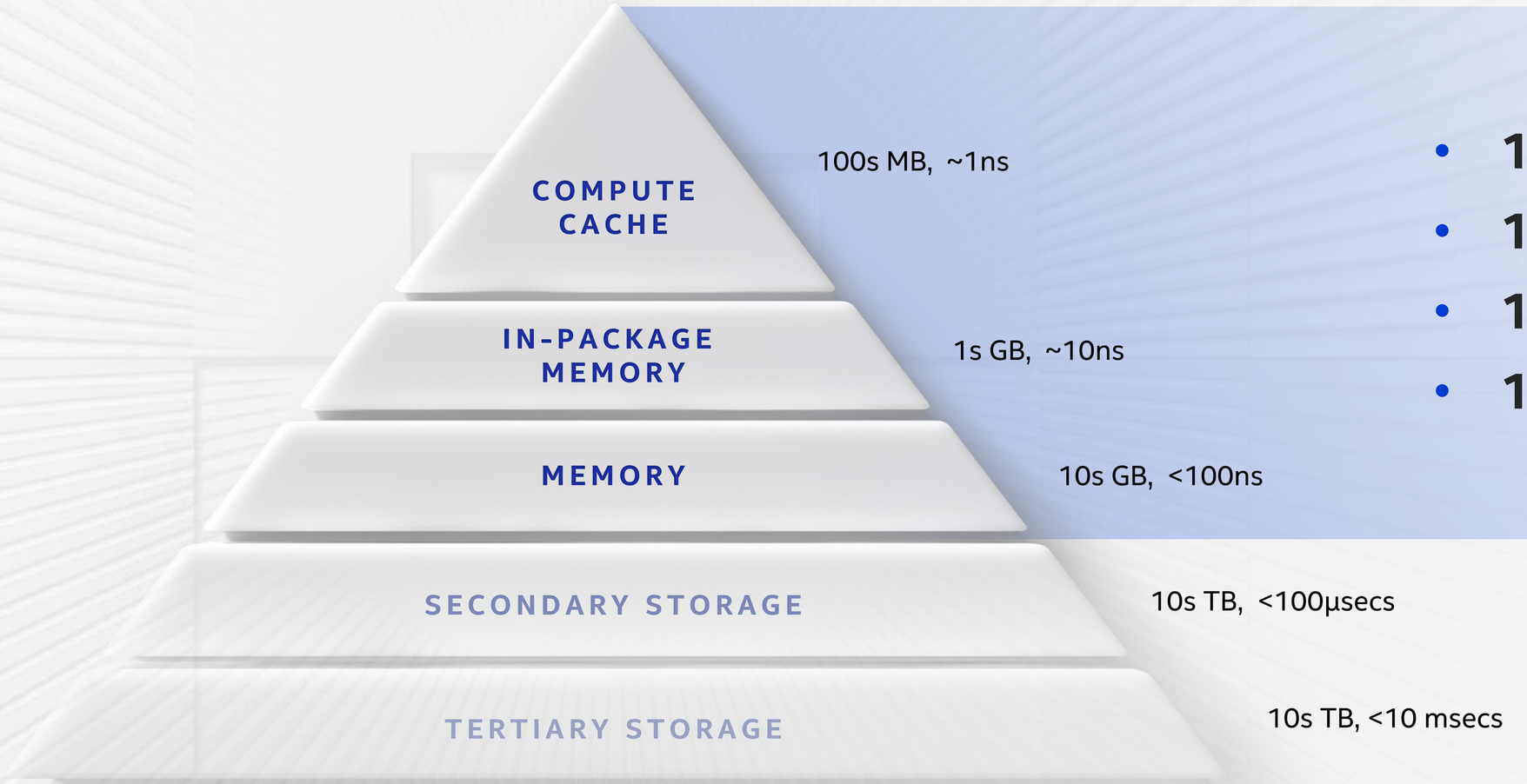


Memory Hierarchy Disruptions



Compute and Memory

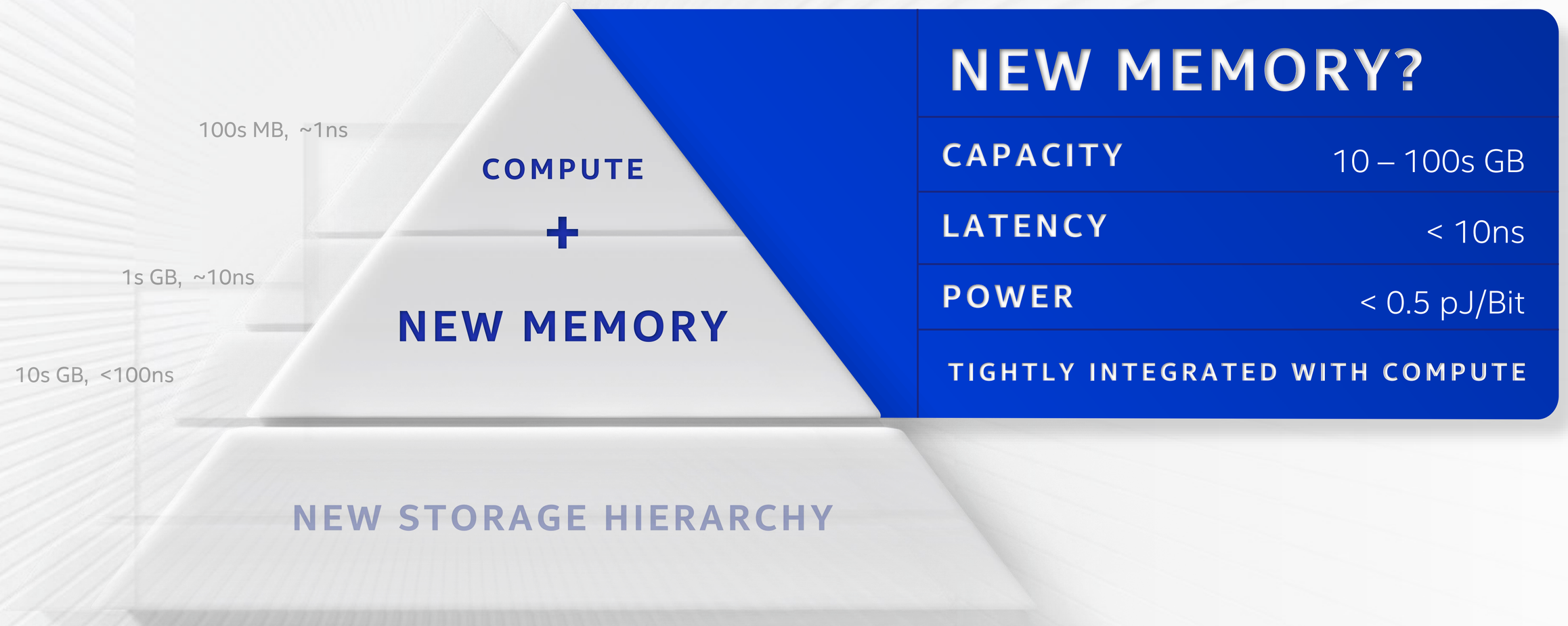
a Vision for Next Gen



- **10x on capacity**
- **10x more B/W**
- **10x lower latency**
- **10x lower power**

Compute and Memory

a Vision for Next Gen

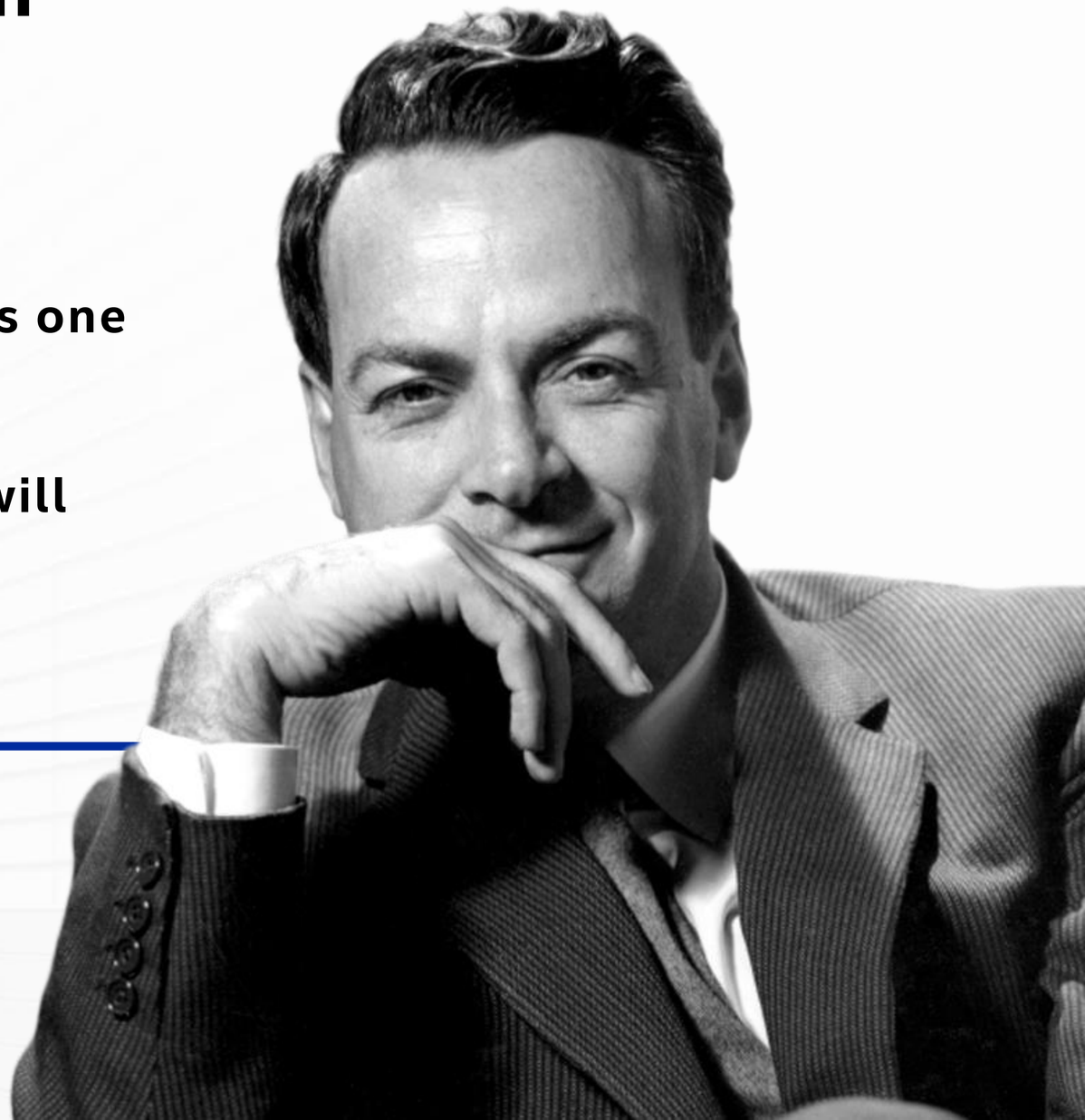


“Plenty of Room at the Bottom”

“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



TRANSISTOR SCALING

ADVANCED PACKAGING

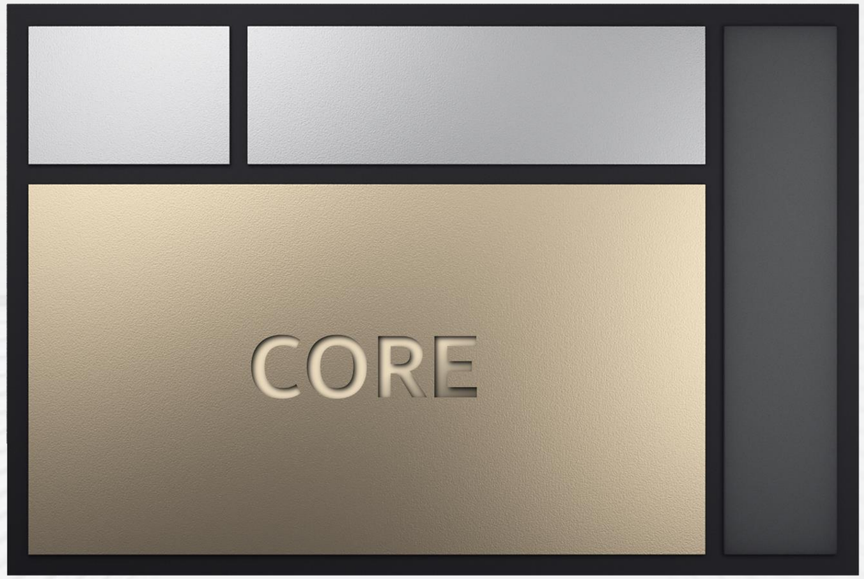
POWER SCALING

MEMORY HIERARCHY DISRUPTION

Beyond Exascale Compute

General Flow of Architecture

In Hardware Companies



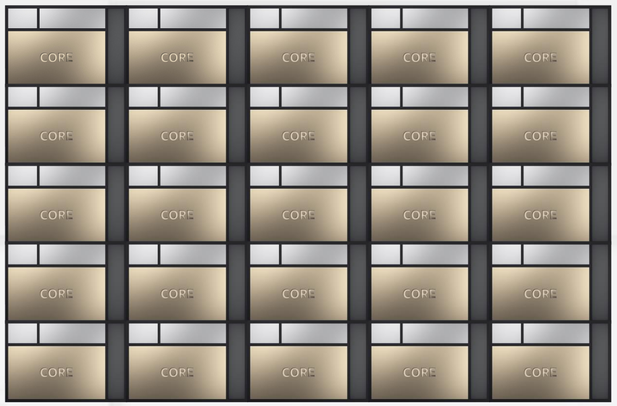
1 Core

SINGLE DIE

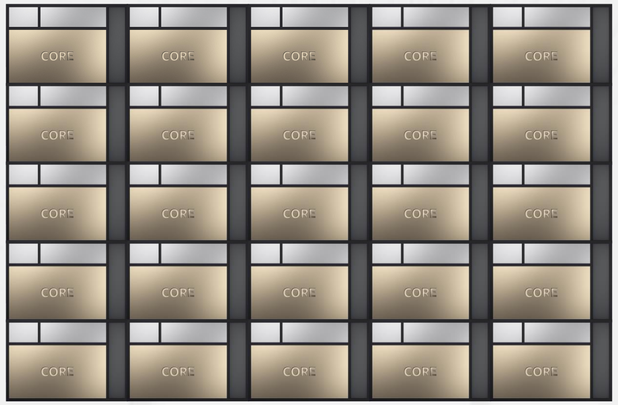
~25

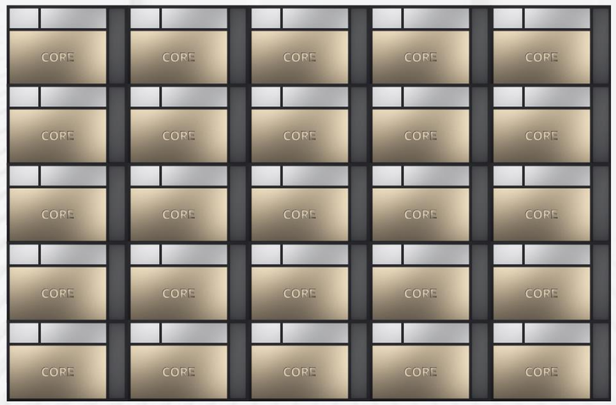
TOTAL # OF CORES

25



1 Core





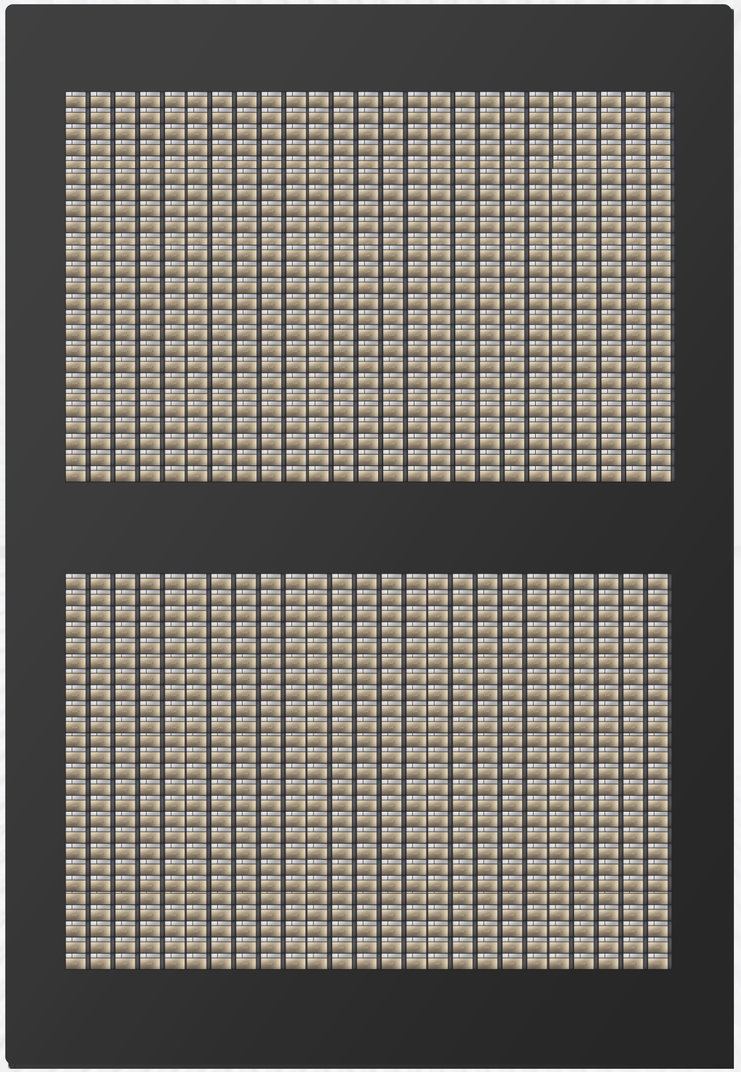
PACKAGE

x2

TOTAL # OF CORES

50

1 Core



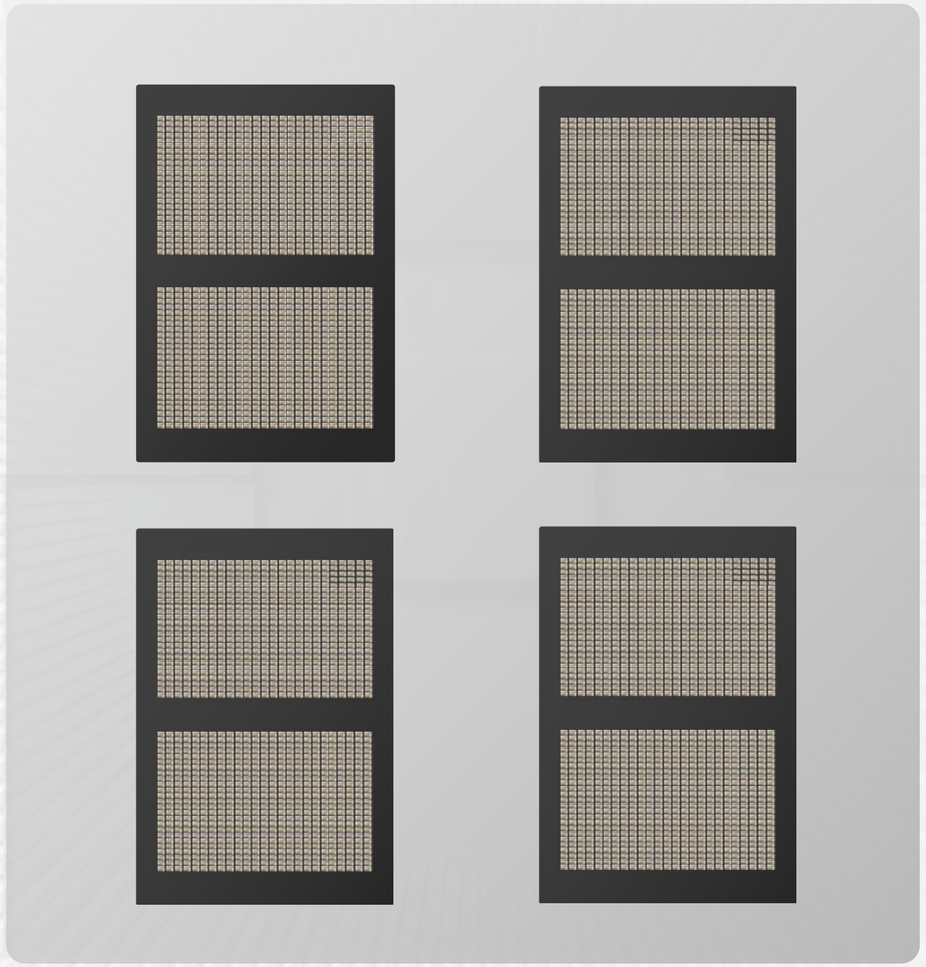
DENSITY SCALING

x50

TOTAL # OF CORES

2500

1 Core



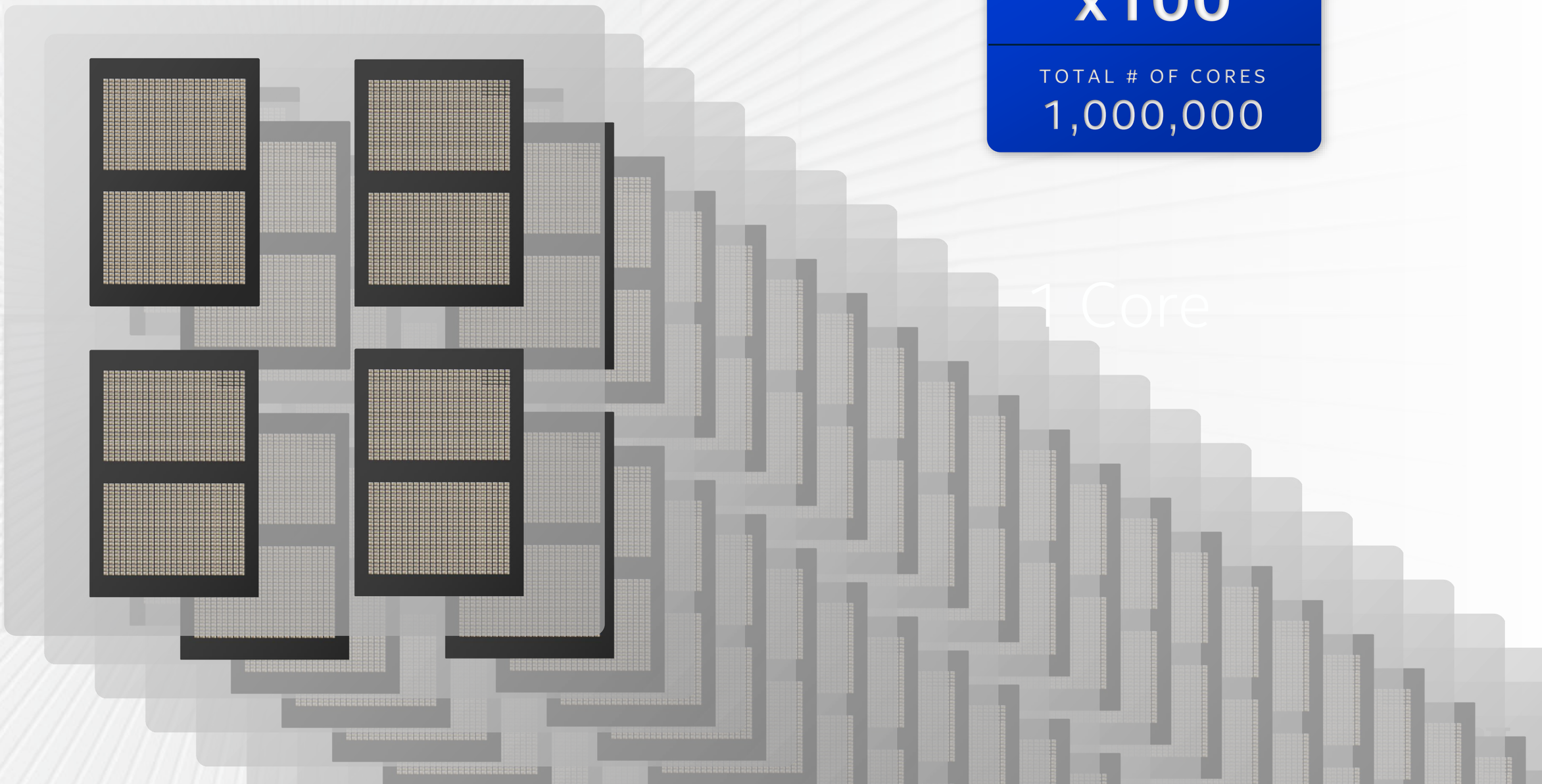
PACKAGE PER BOARD

x4

TOTAL # OF CORES

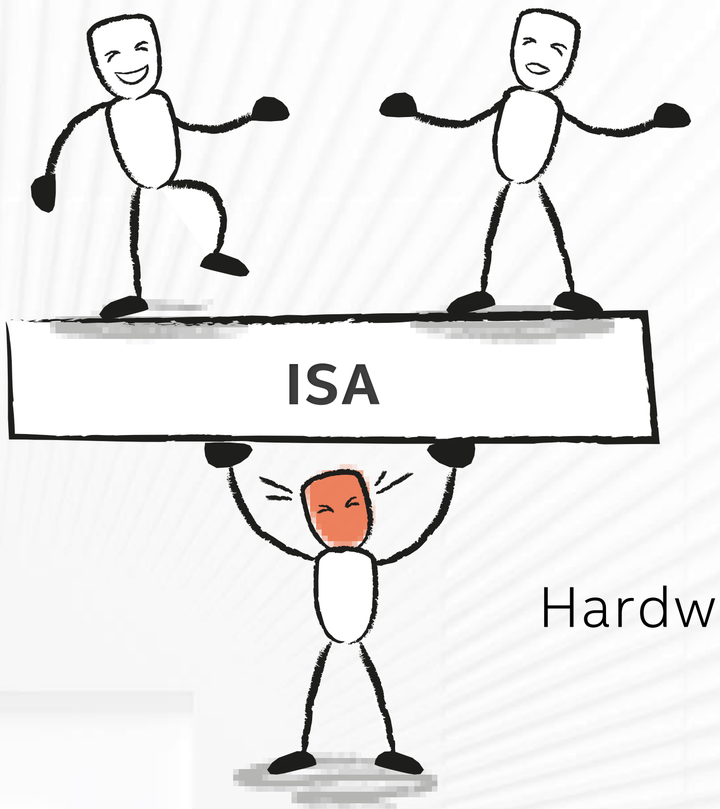
10,000

1 Core


BOARDS
x100
TOTAL # OF CORES
1,000,000
1 Core

“Software is Eating the World”

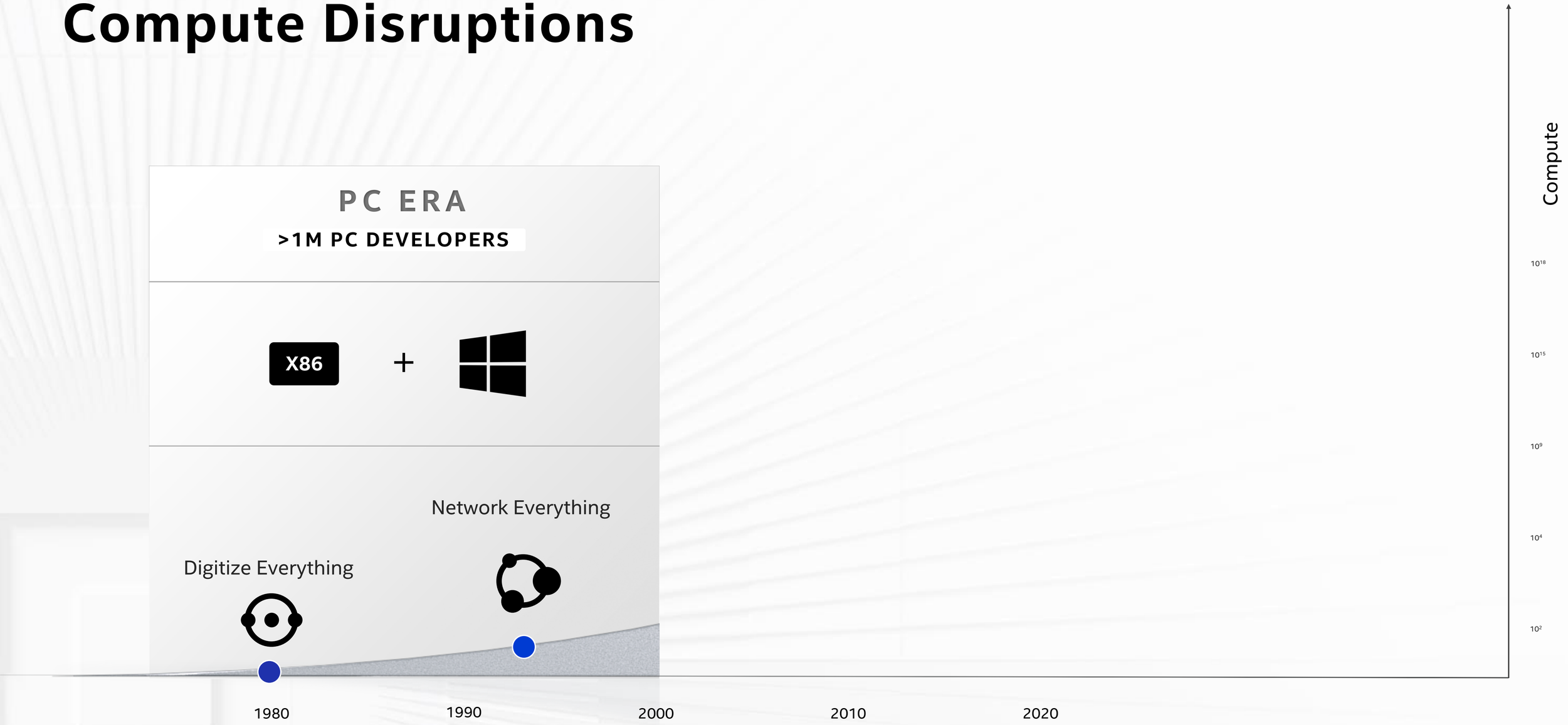
MARC ANDREESSEN, 2011



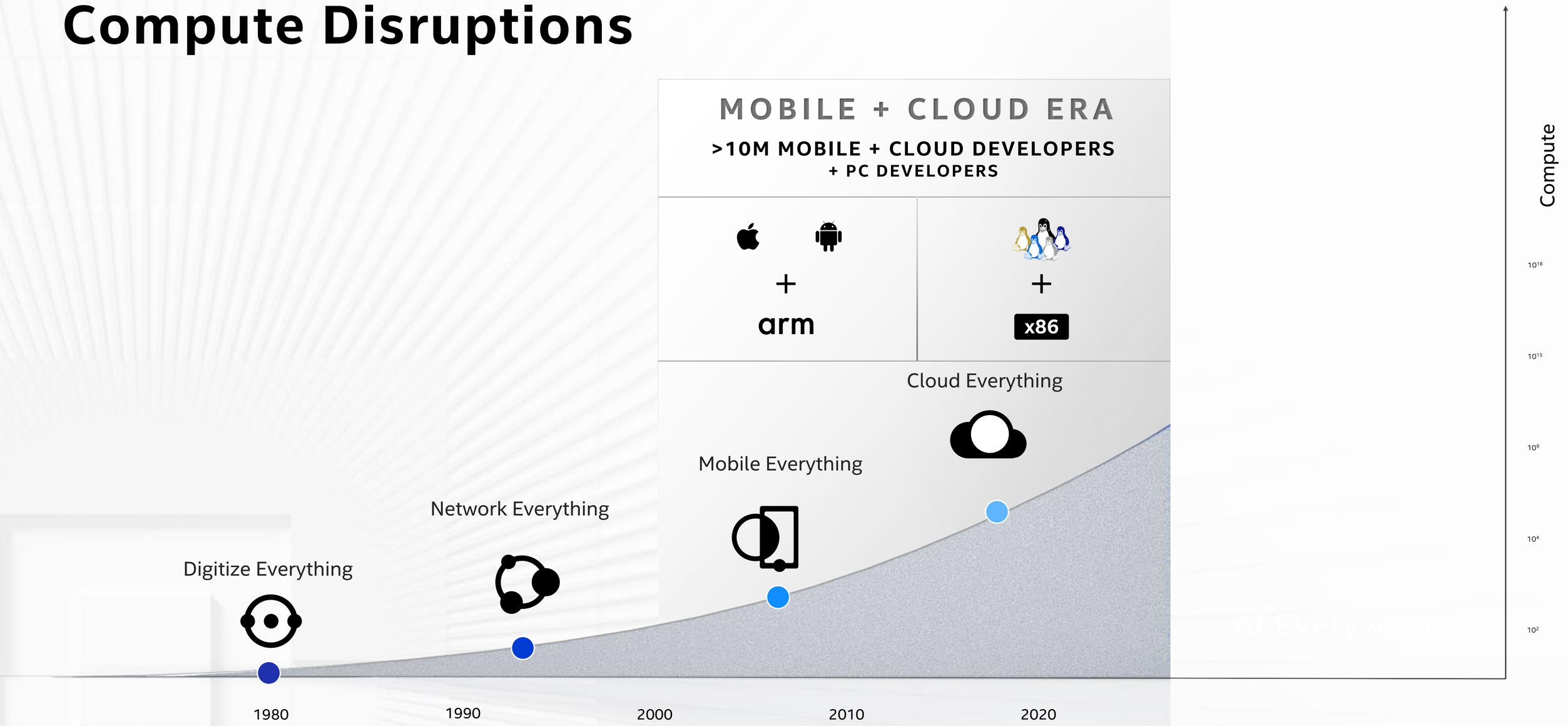
Software

Hardware

Compute Disruptions



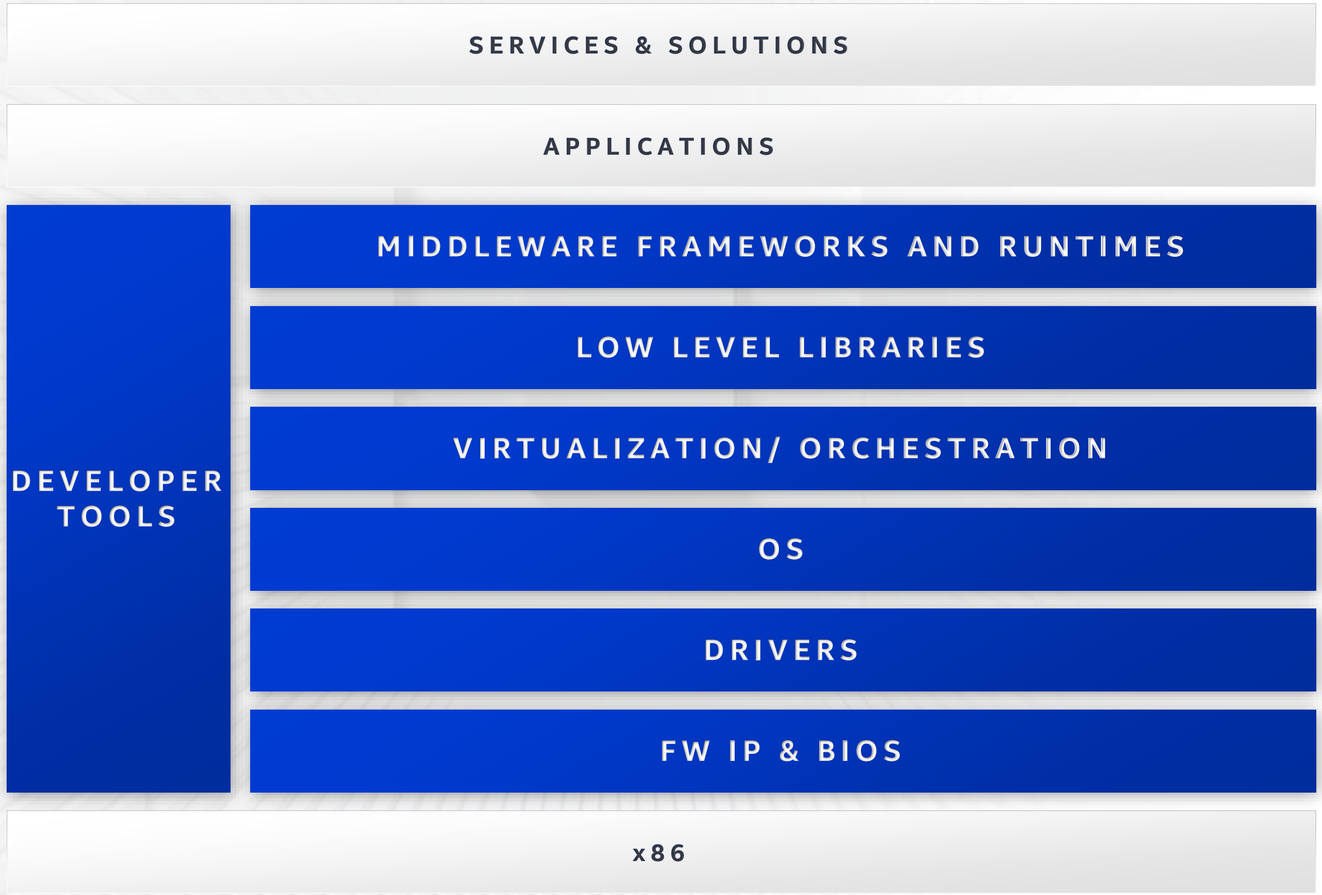
Compute Disruptions



Architecture Impact

x86 Developer Ecosystem

>20M
Developers



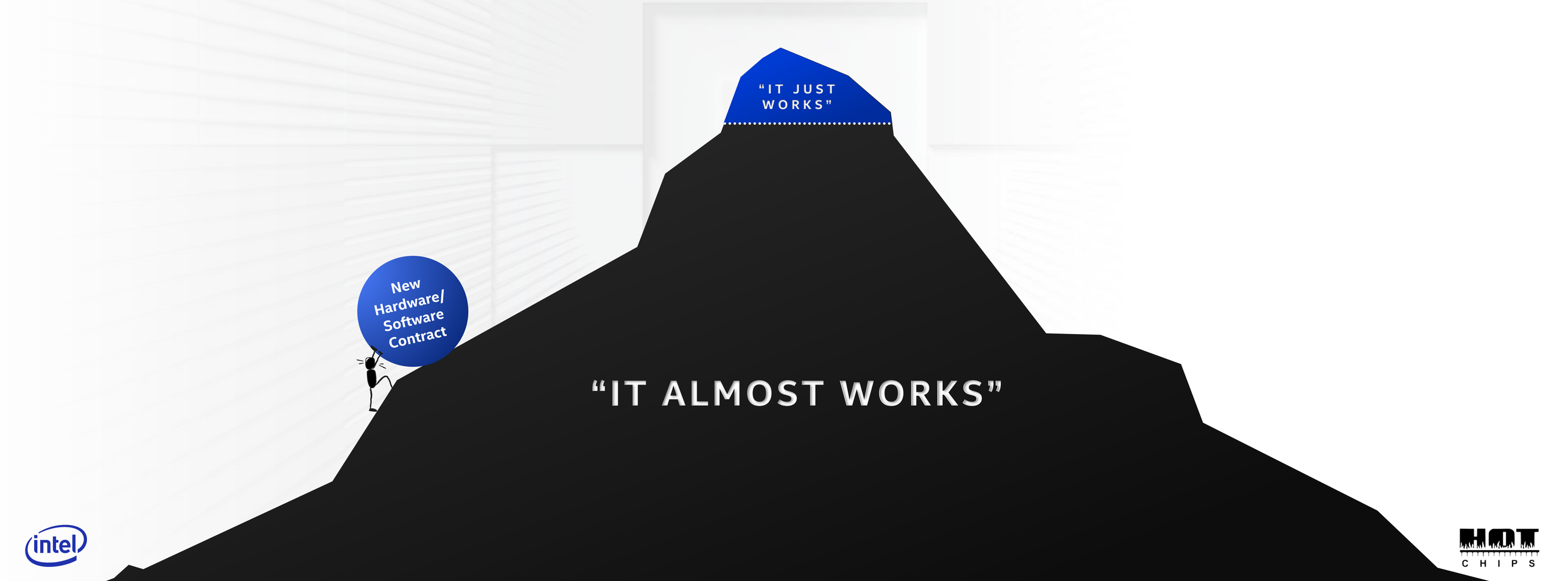
Software Stack & Developers

Stack and Swiss Cheese

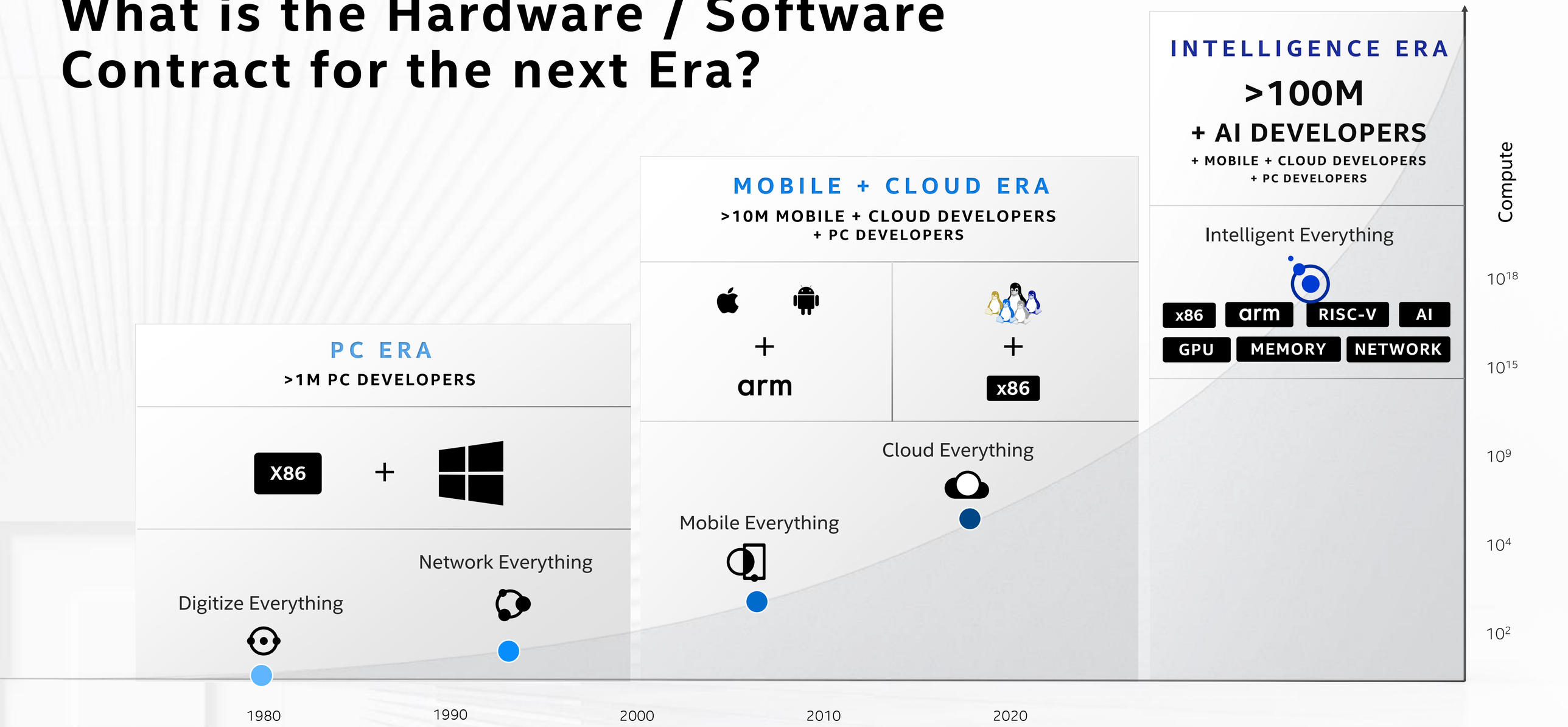
Middle is
Full of ”holes”

New Hardware / Software Contracts

The Reality…

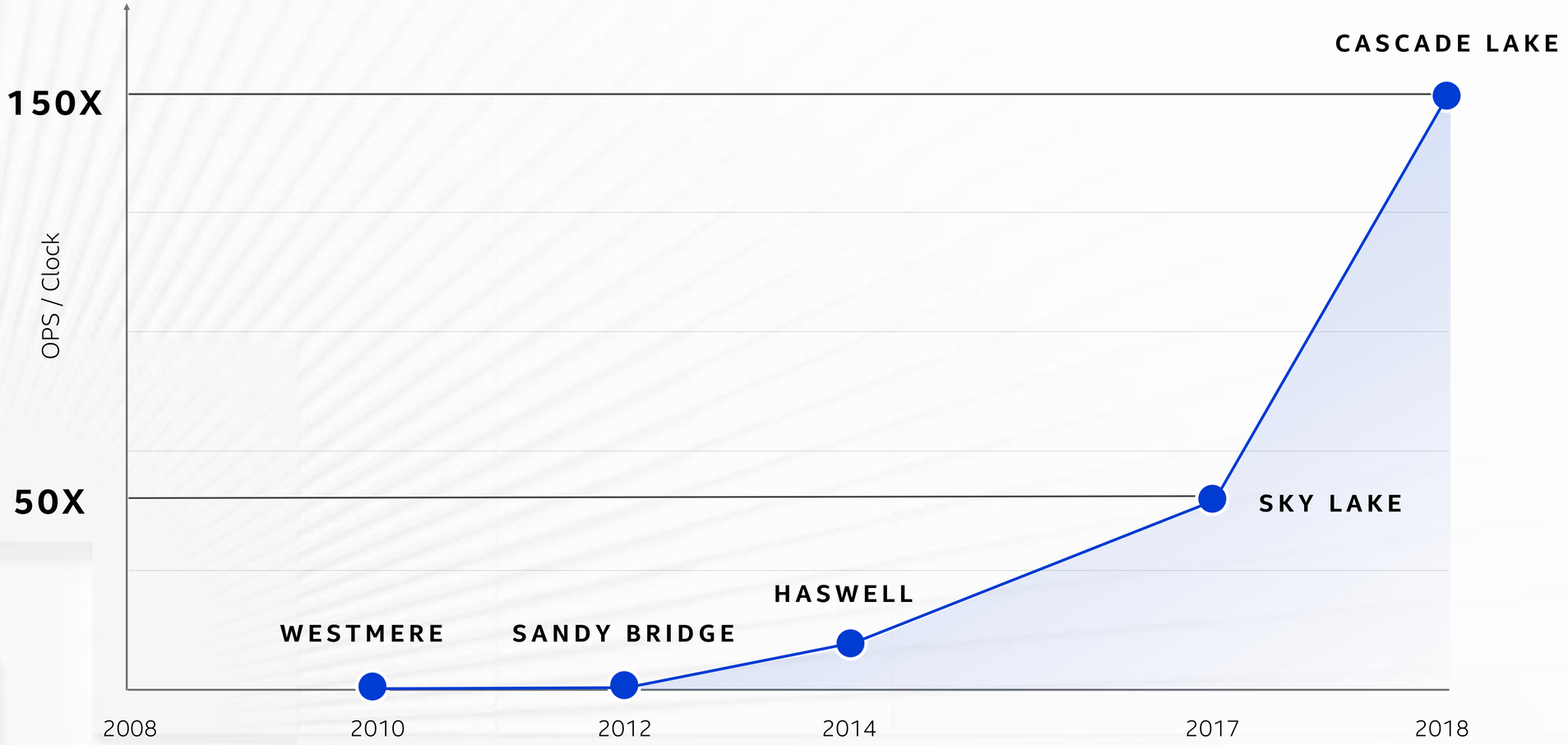


What is the Hardware / Software Contract for the next Era?



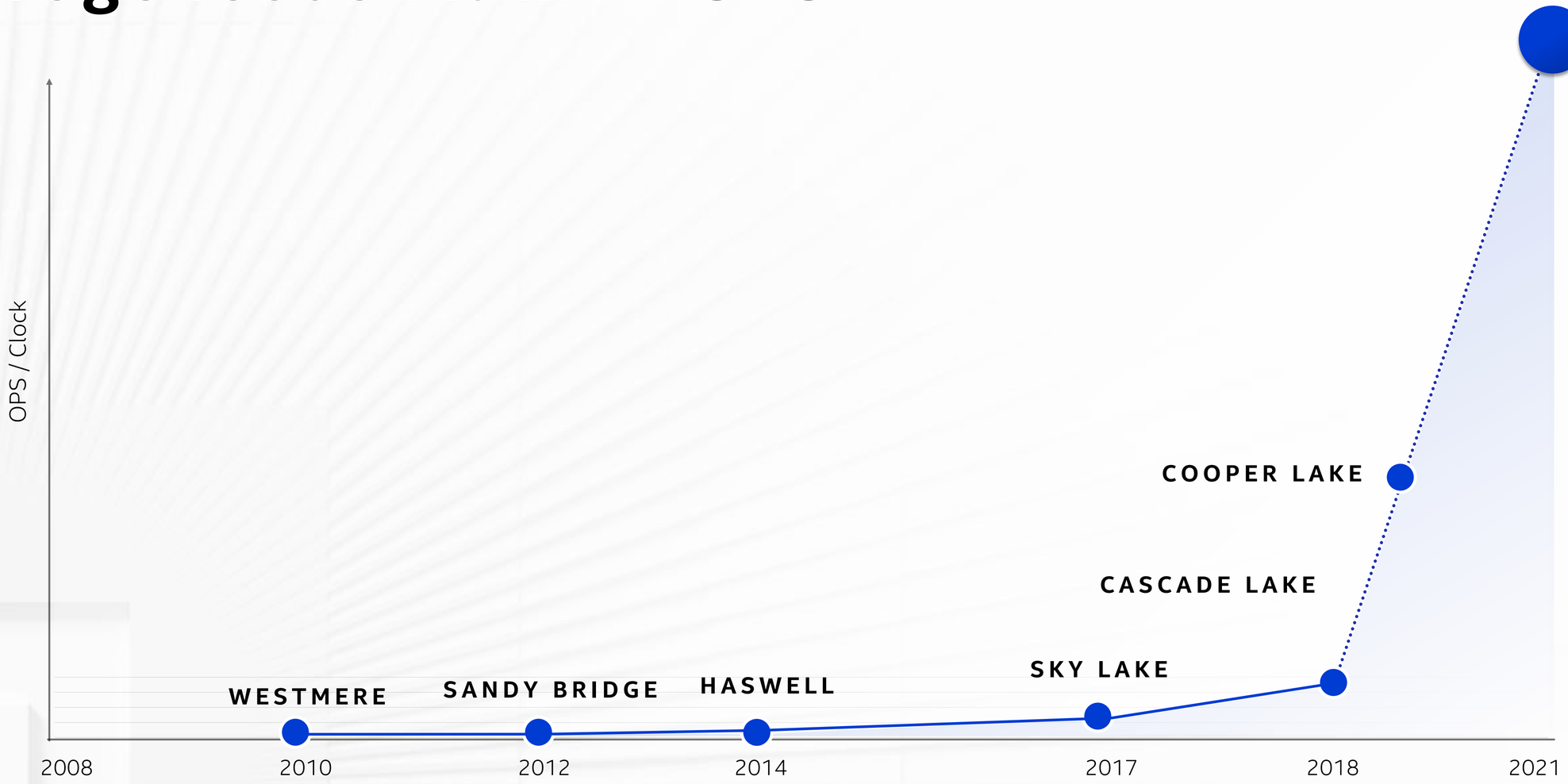
Generality $\propto$ 1/Architecture Heterogeneity

Heterogeneous Math in CPU



For illustrative purposes only

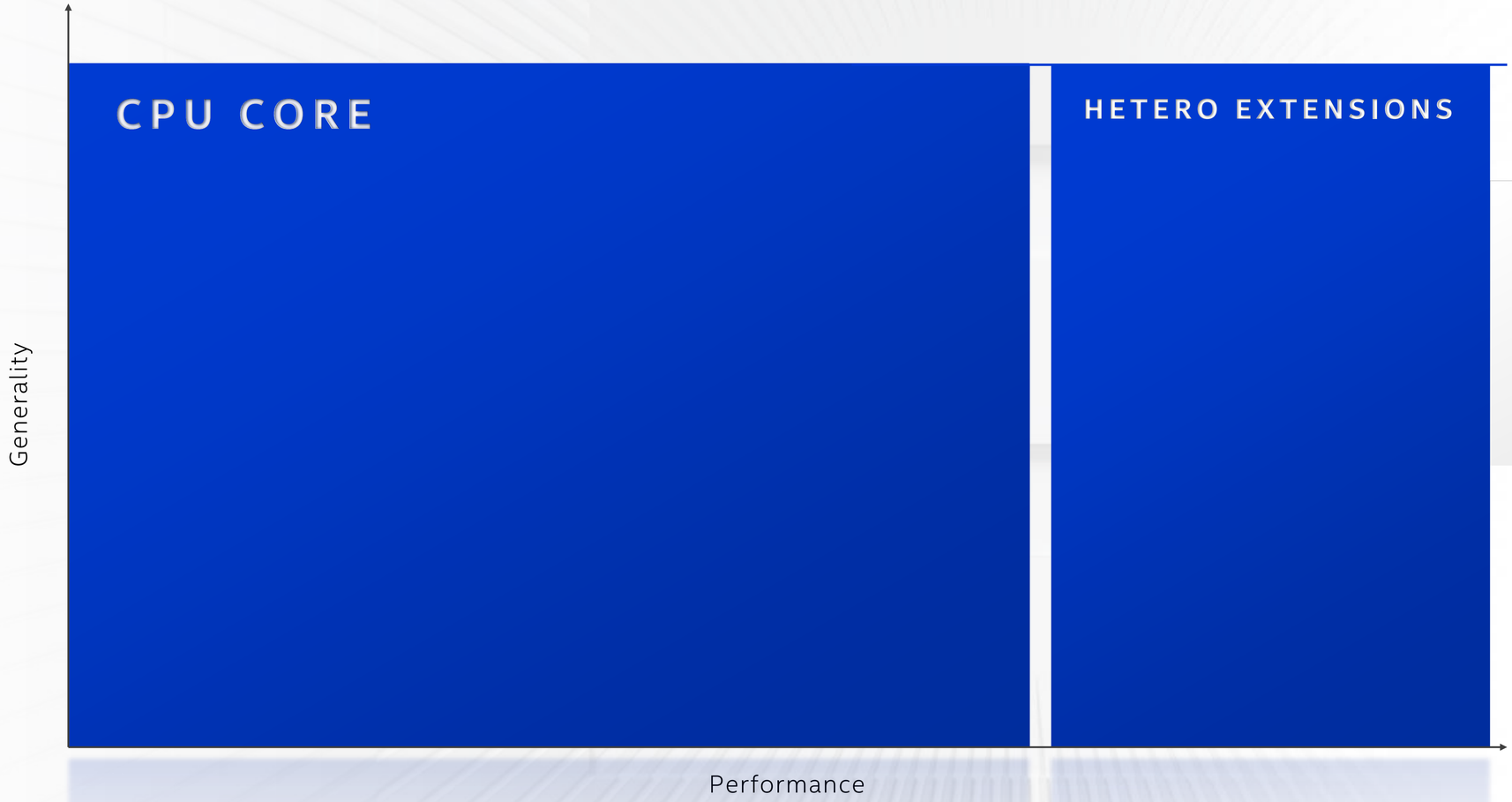
Heterogeneous Math in CPU



For illustrative purposes only

CPU Impact with ISA Extensions & Software

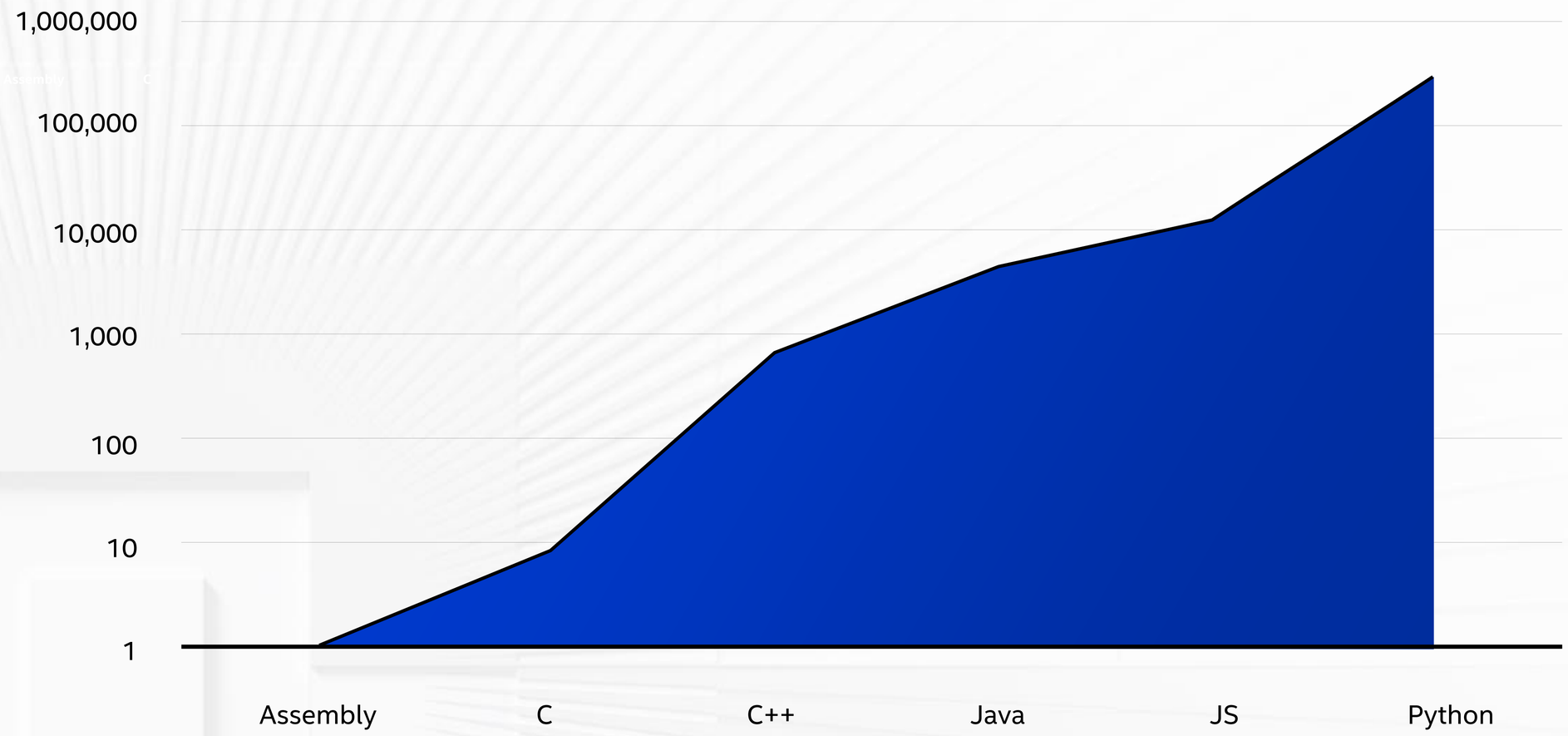
CPU Impact with ISA Extensions & Software



~3-5 YEARS TO REACH BROAD ADOPTION

Productivity and Scale

Mandelbrot Static Instructions Generated per Line of Code



For illustrative purposes only

ABSTRACTION REQUIREMENTS

SCALABLE AND OPEN

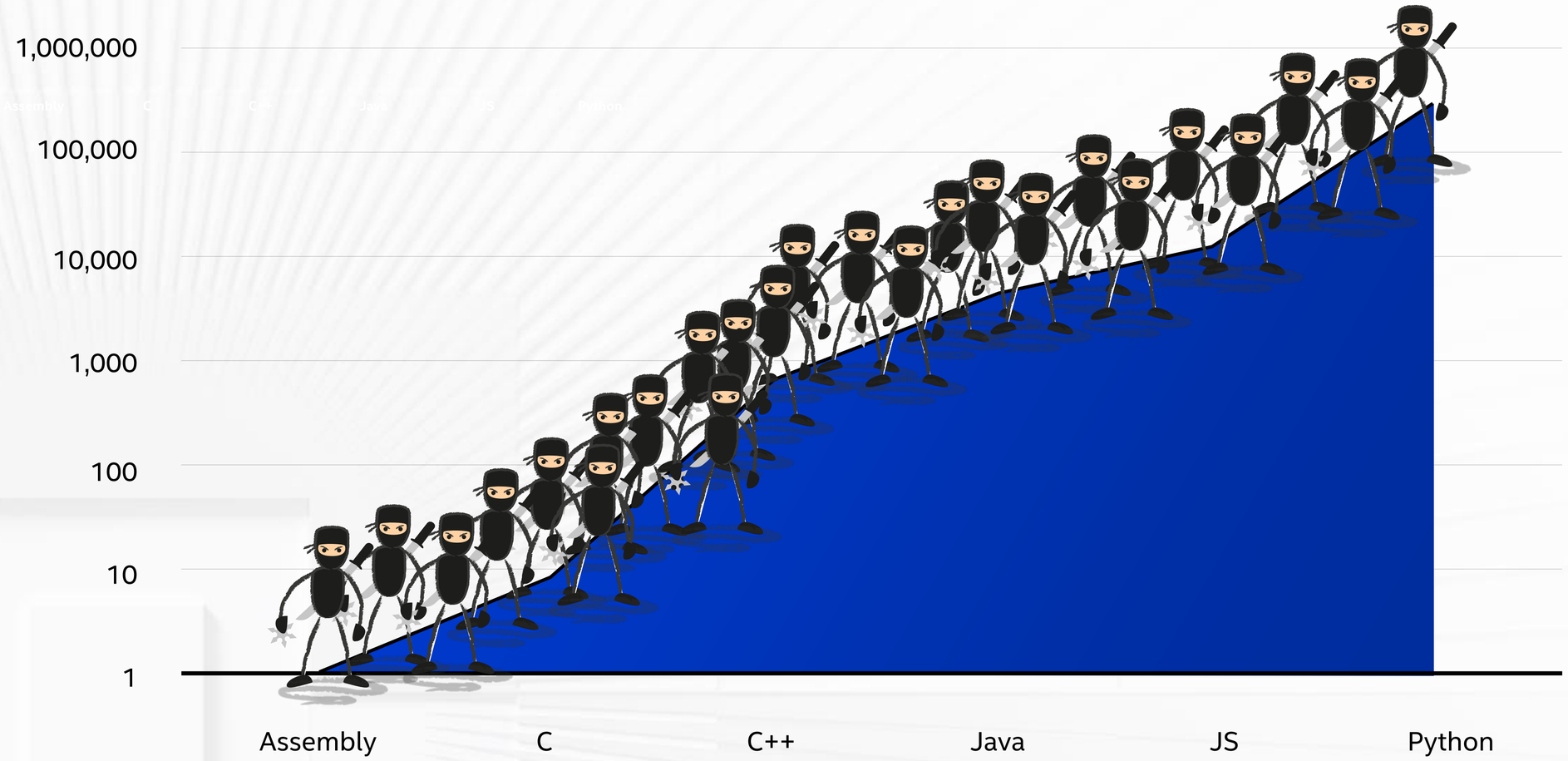
ABSTRACTION REQUIREMENTS

SCALABLE AND OPEN

ABSTRACT AT MULTIPLE LAYERS

Performance vs. Productivity

Mandelbrot Static Instructions Generated per Line of Code



For illustrative purposes only

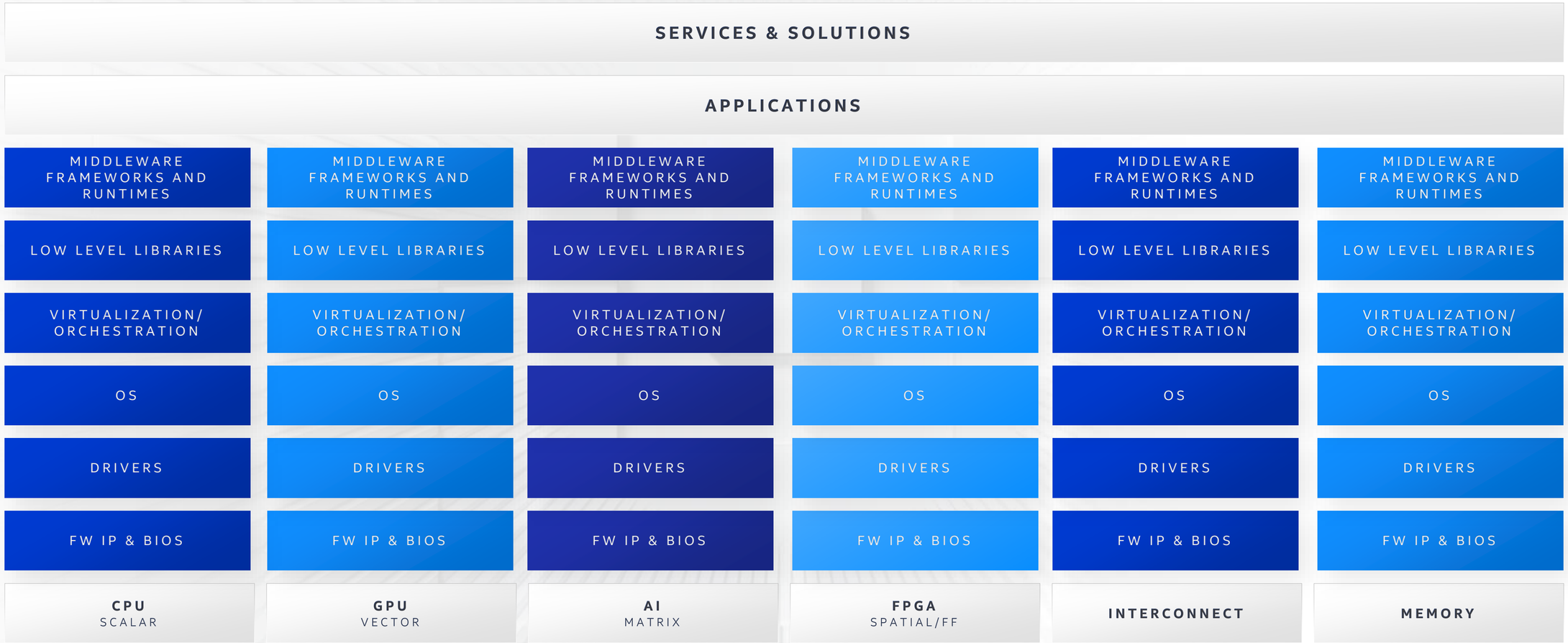
ABSTRACTION REQUIREMENTS

SCALABLE AND OPEN

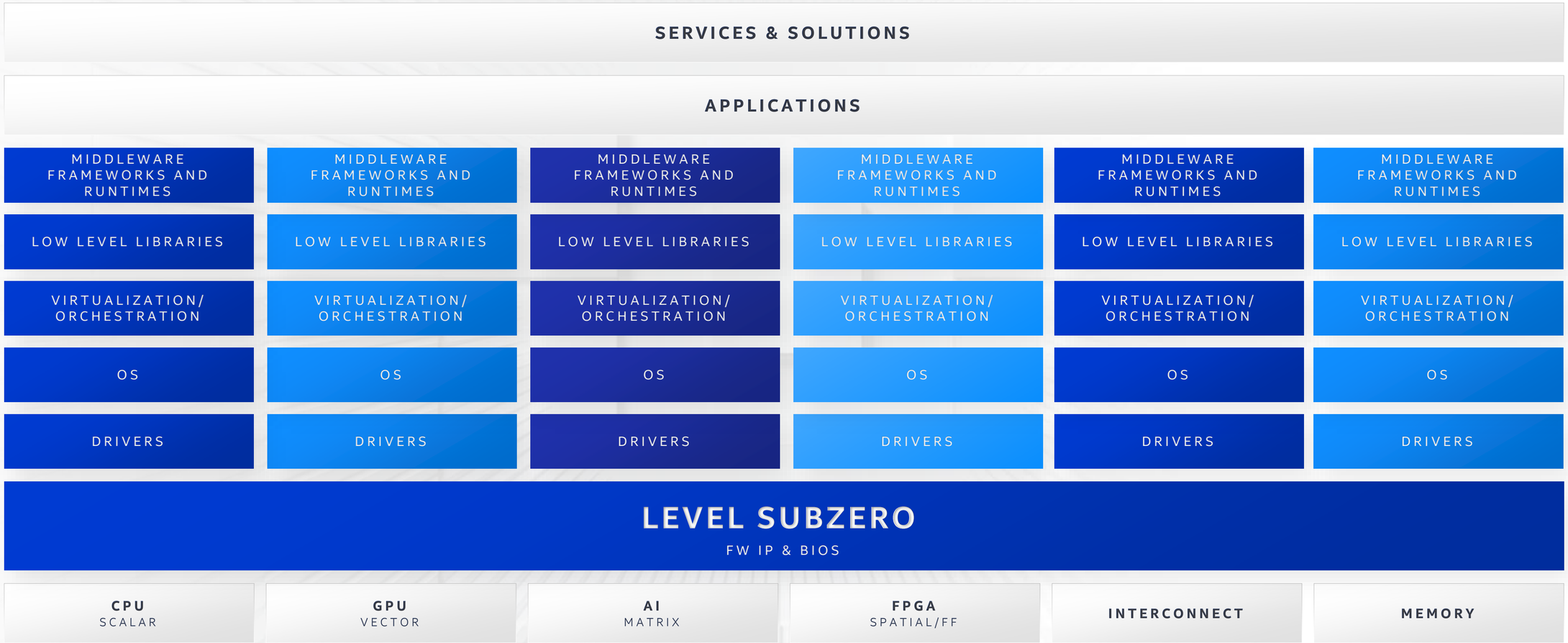
ABSTRACT AT MULTIPLE LAYERS

SUPPORT NINJA'S ACROSS THE ENTIRE STACK

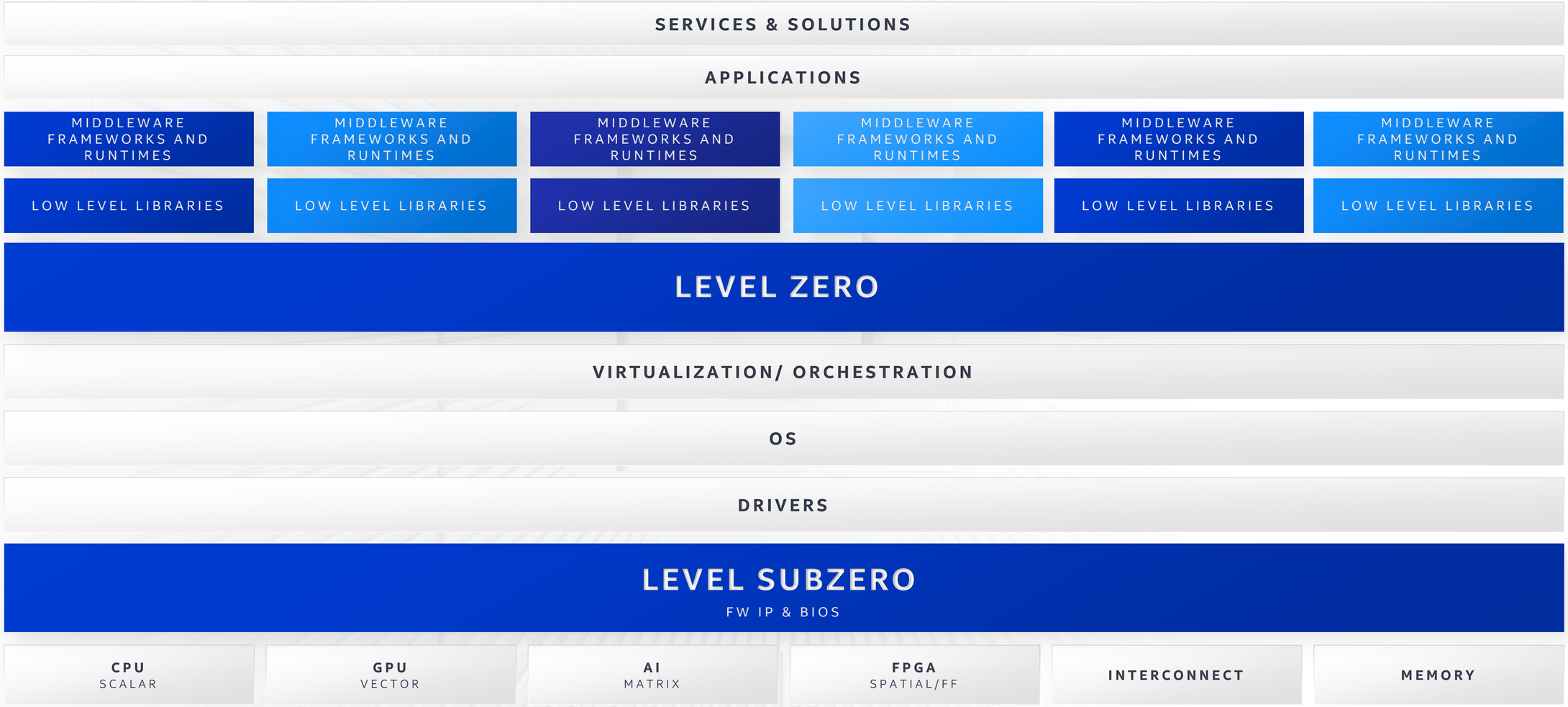
Generality ∝ 1/ Architecture Heterogeneity



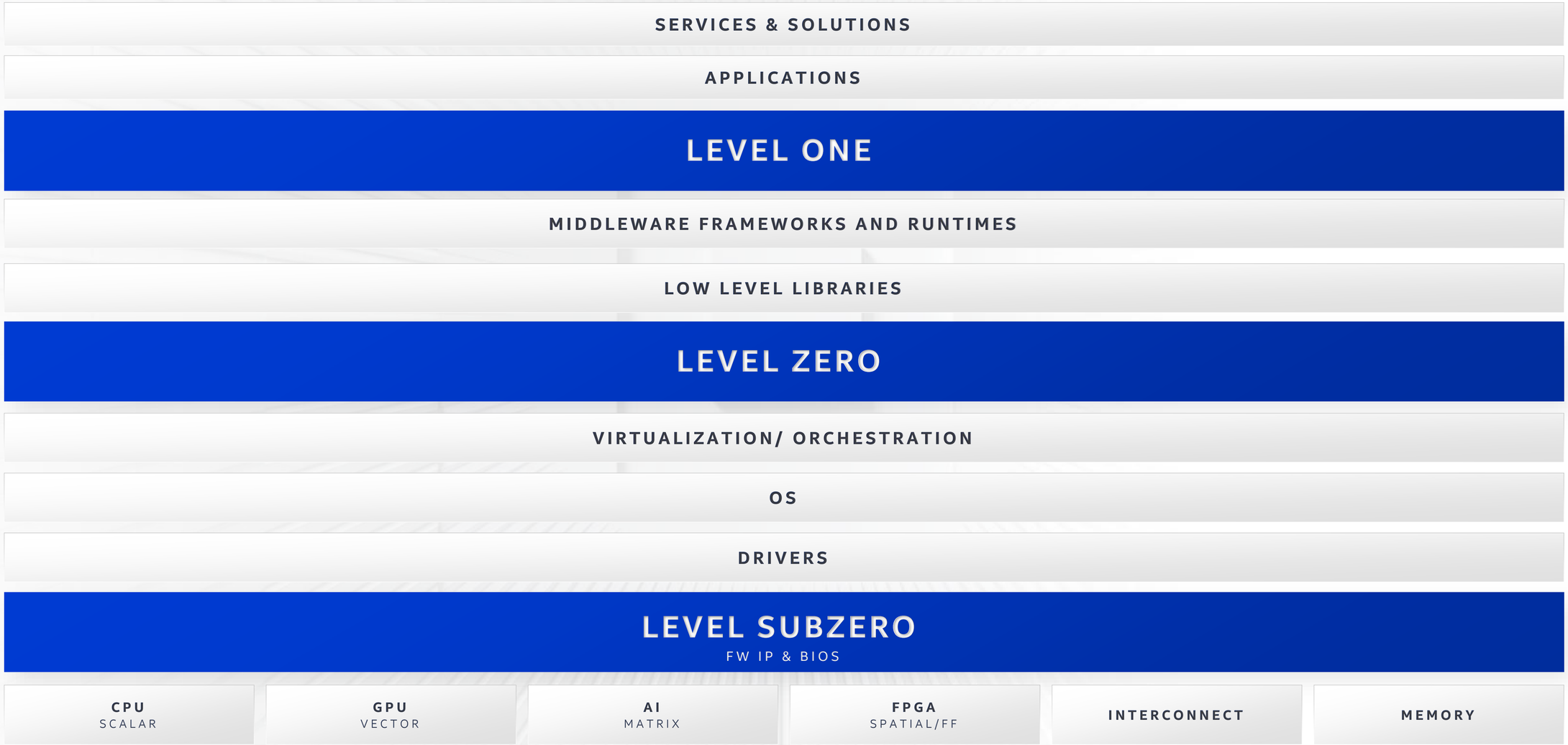
Abstraction Required at Multiple Layers



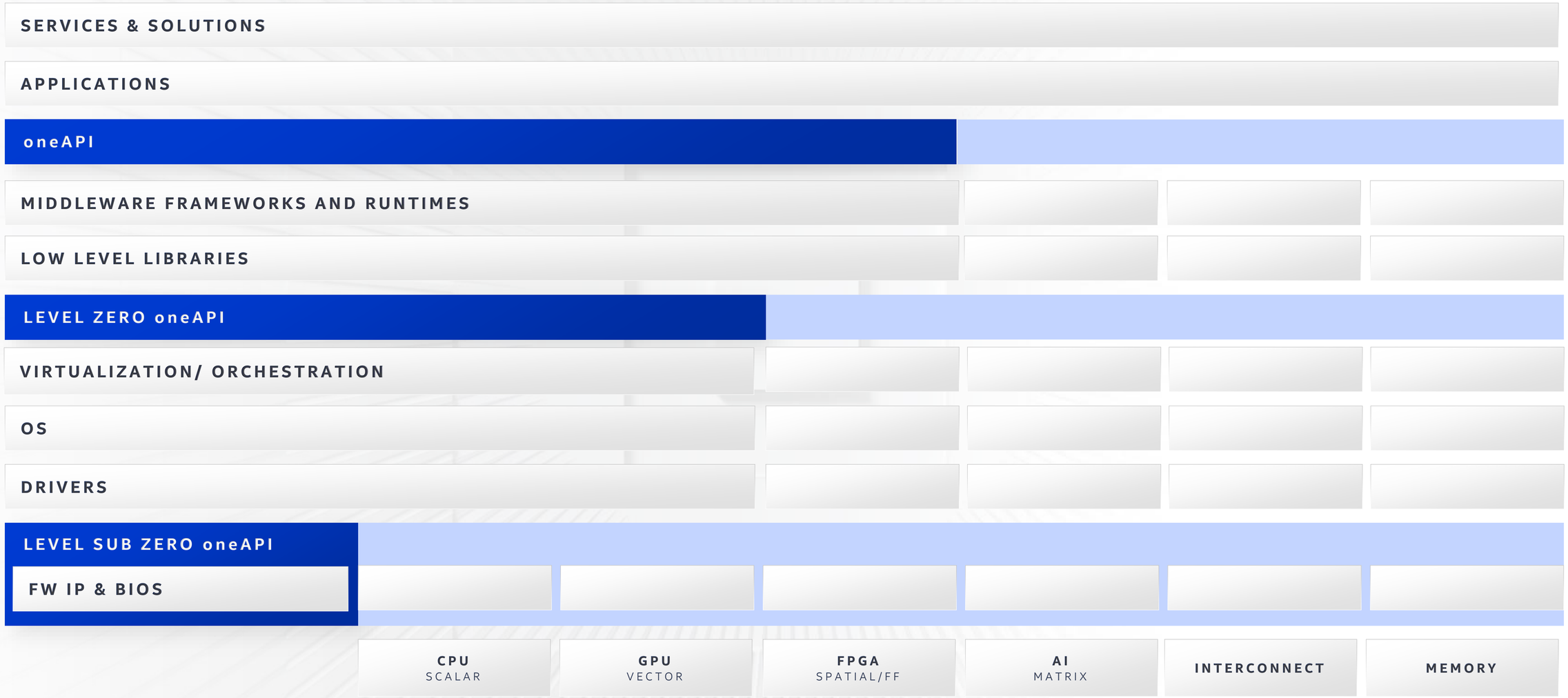
Abstraction Required at Multiple Layers



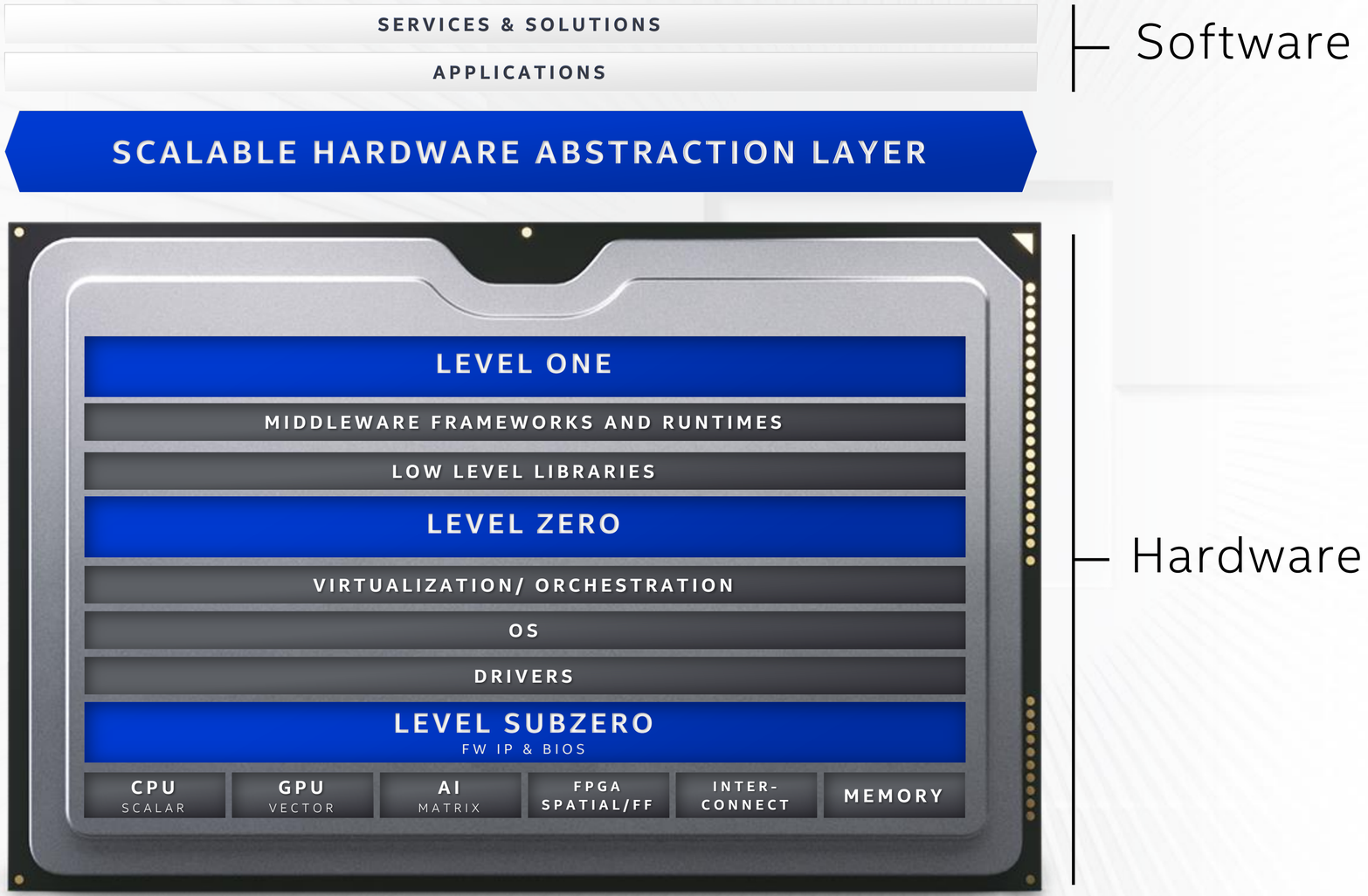
Abstraction Required at Multiple Layers



oneAPI Abstraction Roadmap



Hardware / Software Contract Layer



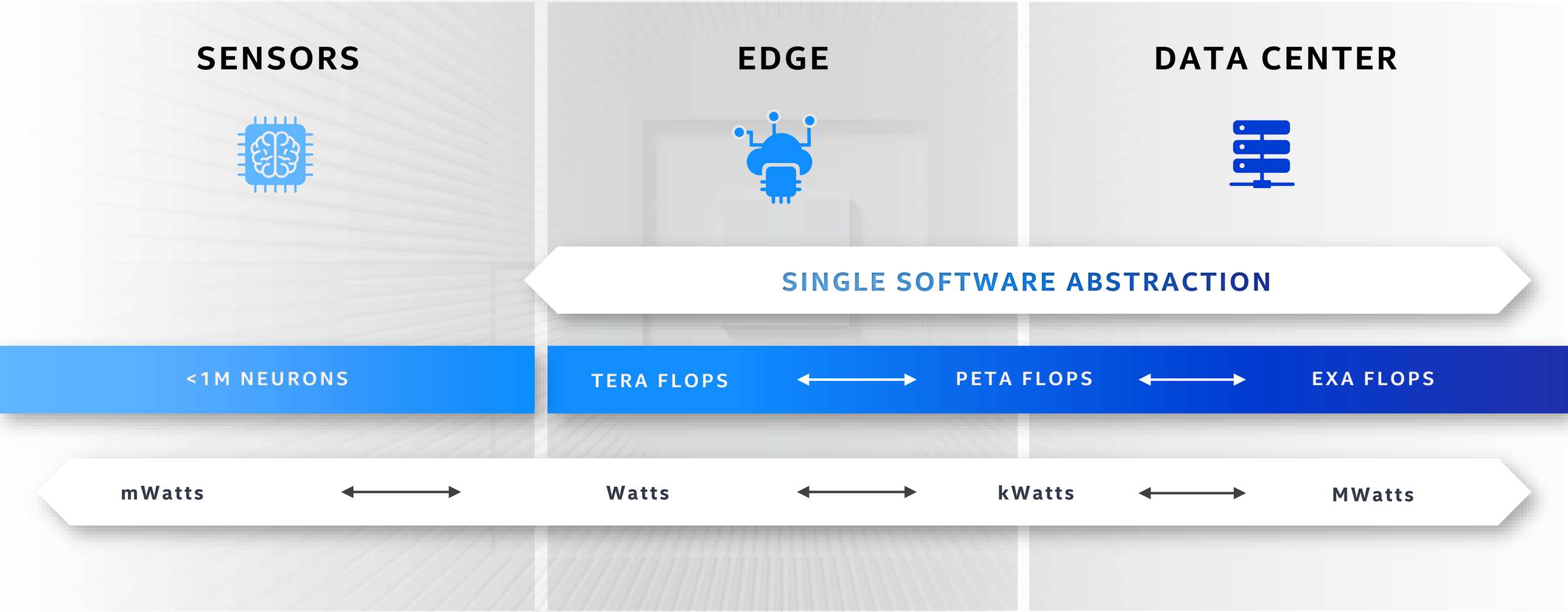
Software

Hardware

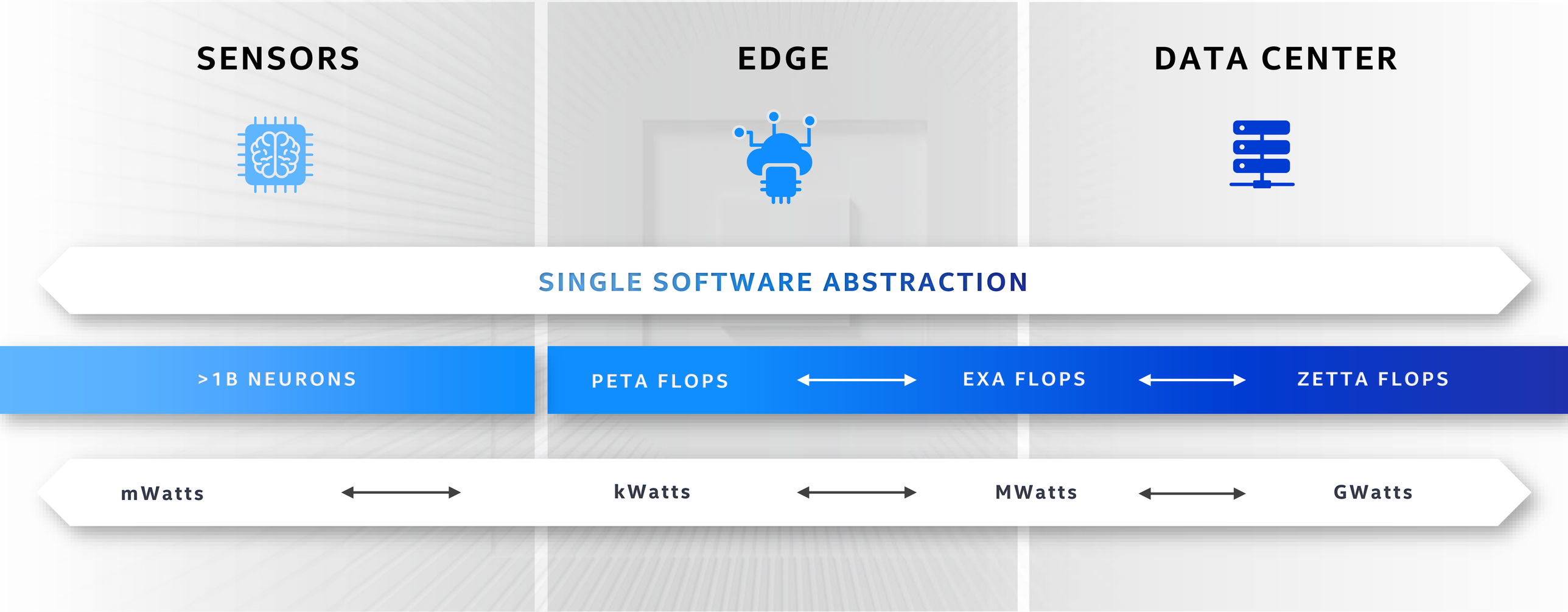
Goal of new HW/SW **Contract**

No Transistor Left Behind

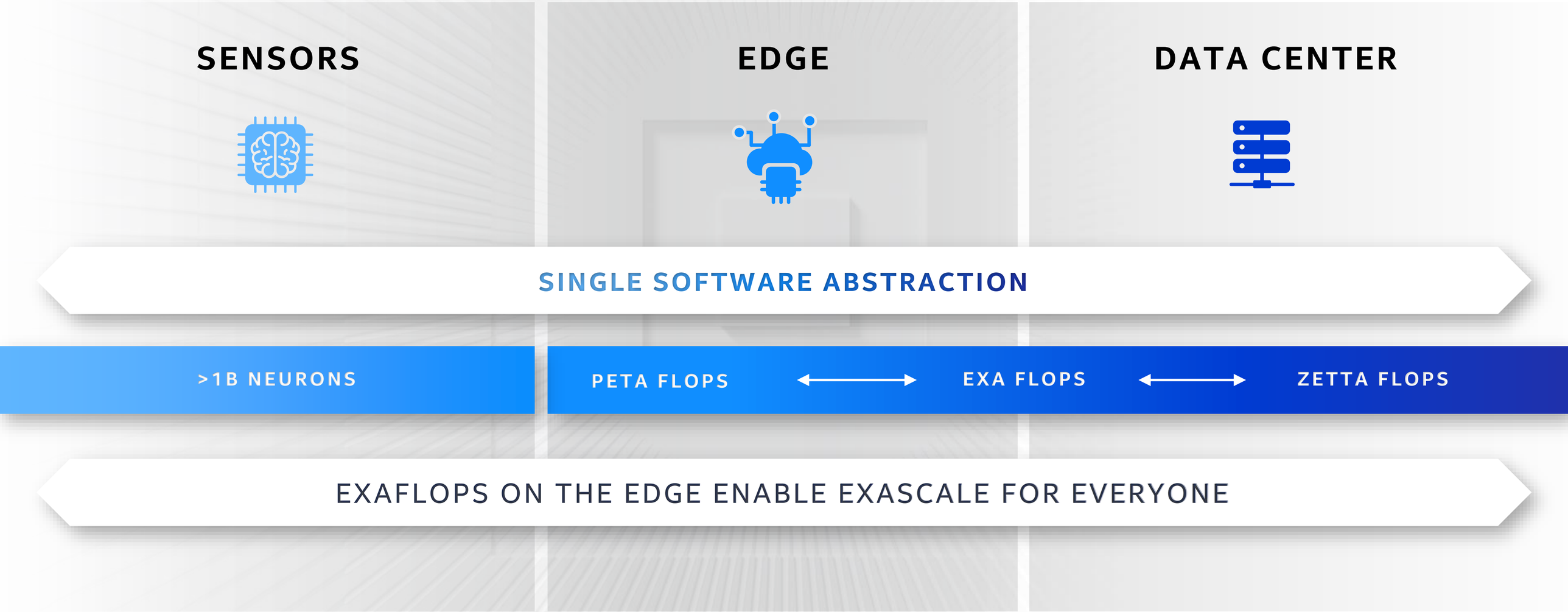
From Sensors to Supercomputers – 2021



From Sensors to Supercomputers – 2025



From Sensors to Supercomputers – 2025



Summary


PERFORMANCE
PLENTY OF ROOM AT THE TOP
1000x
BY 2025
GENERALITY
PLENTY OF ROOM AT THE BOTTOM


SOFTWARE
SECURITY
INTERCONNECT
MEMORY
XPU ARCHITECTURE
PROCESS & PACKAGING
EXASCALE
FOR
EVERYONE

“Optimism is the essential ingredient of innovation”

ROBERT NOYCE


THANK YOU
AND REMEMBER, LEAVE NO TRANSISTOR BEHIND
intel
HOT CHIPS

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.