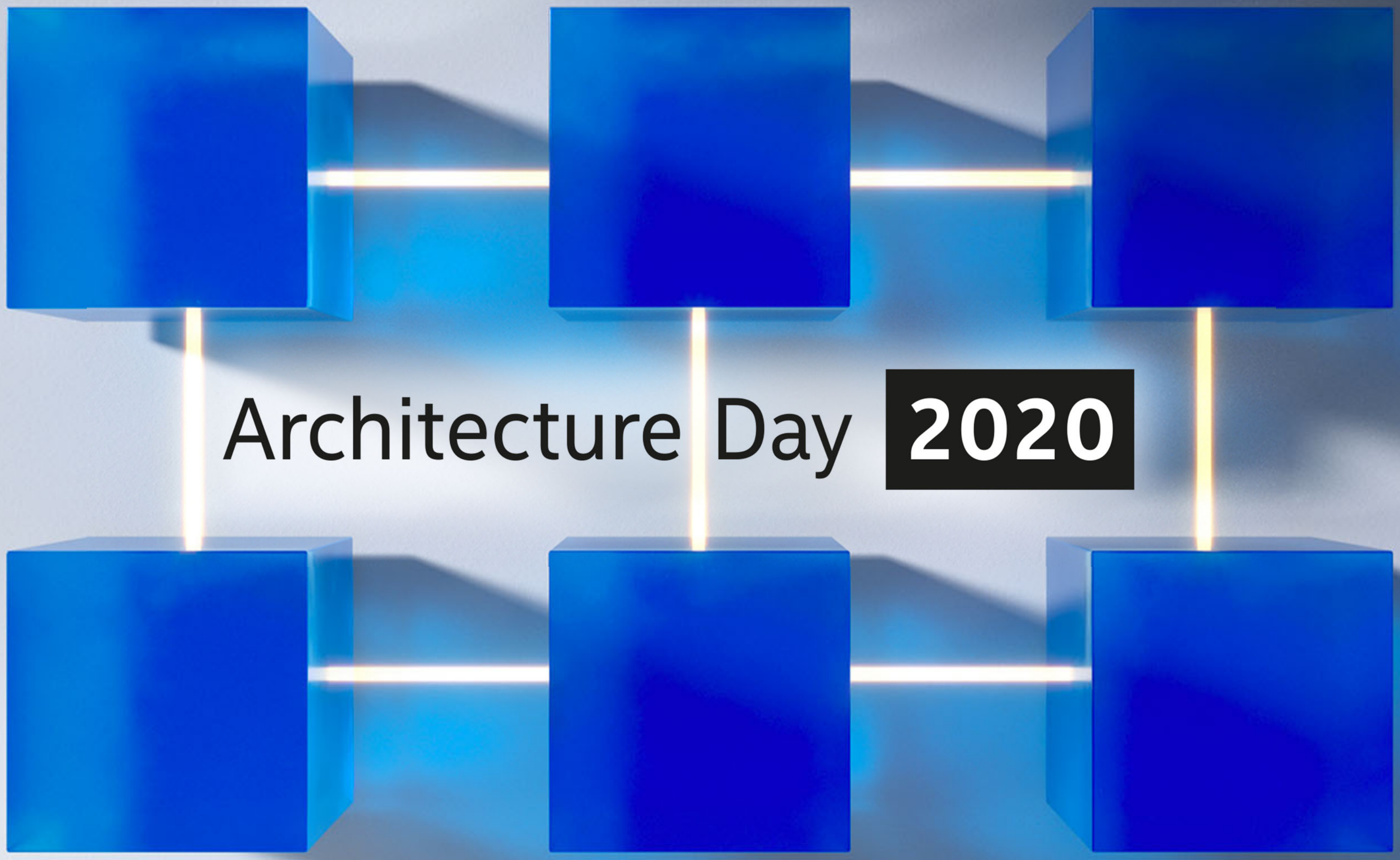


Architecture Day 2020

TECHNOLOGY
PILLARS

Disclosures

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.

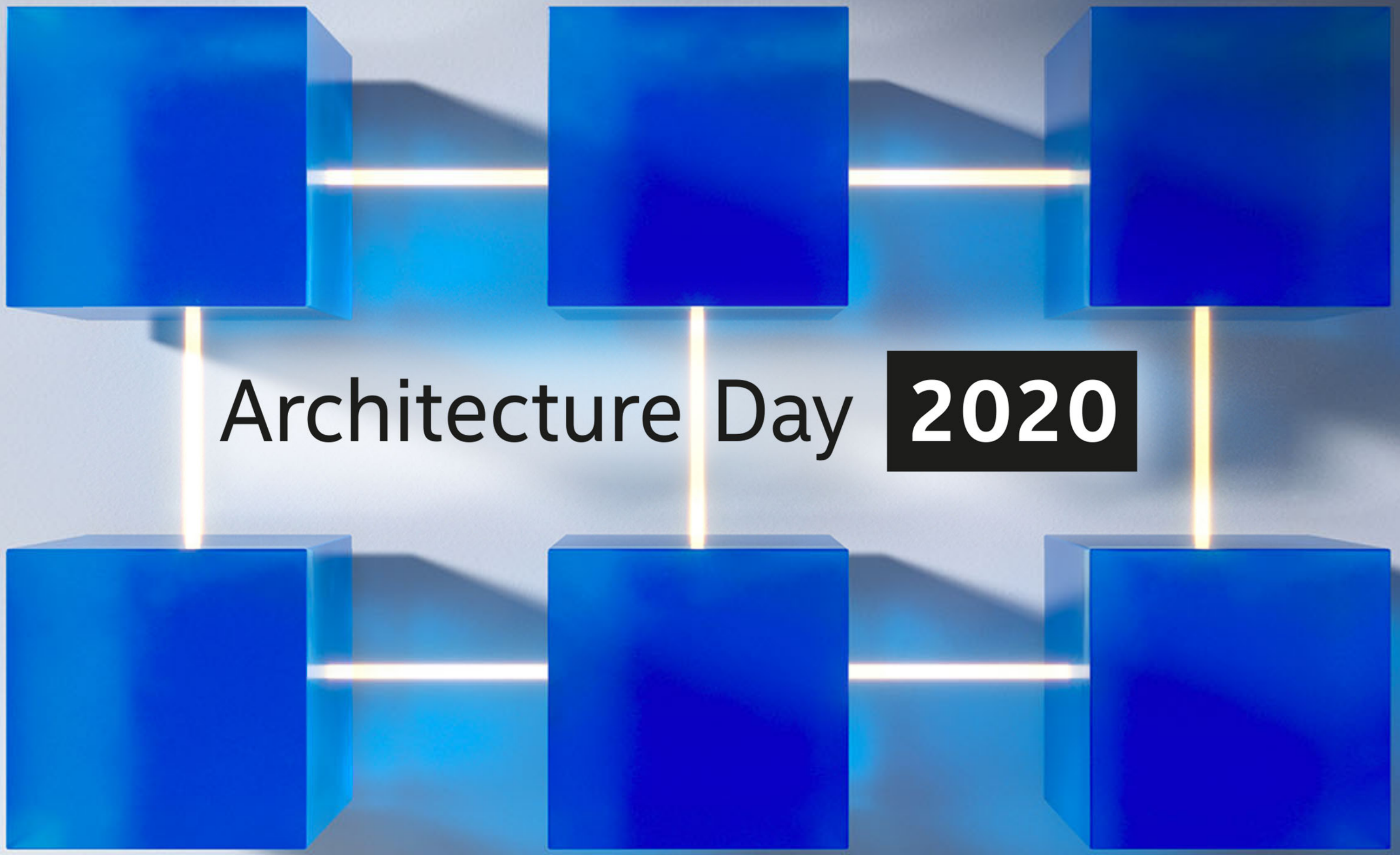
Today's Agenda

Raja Koduri

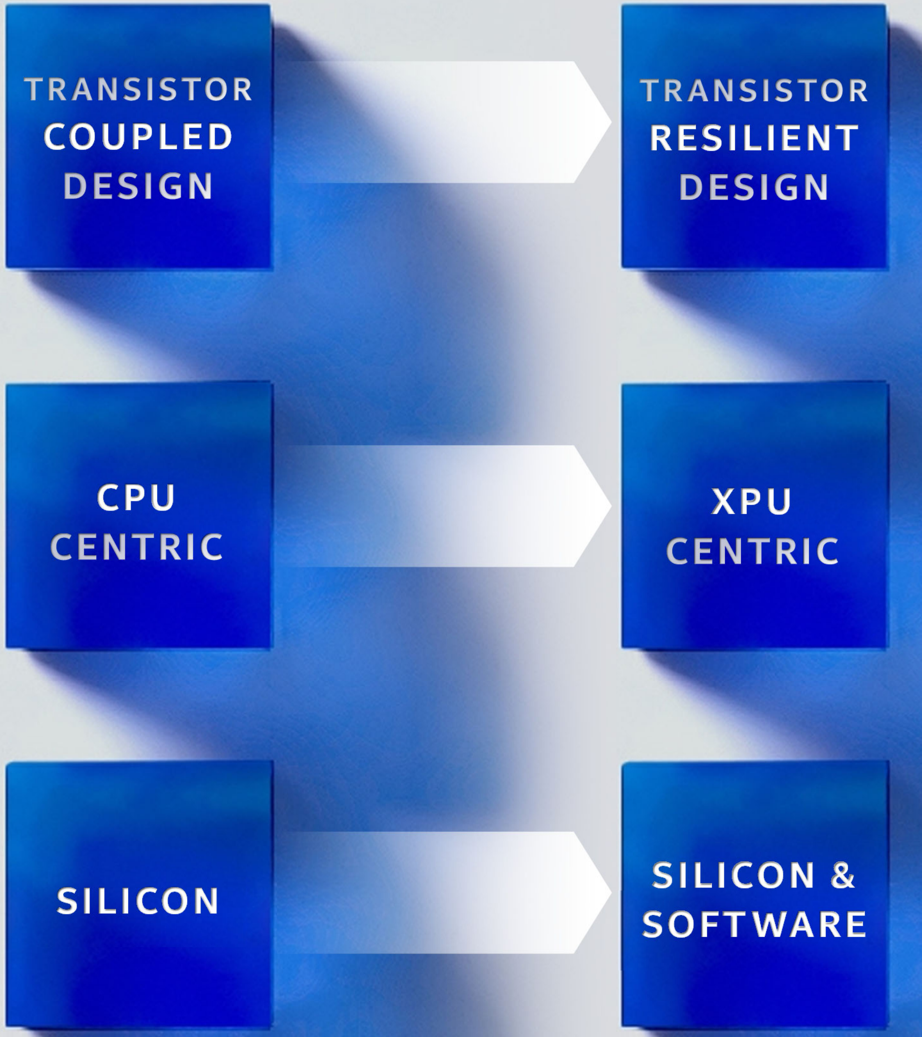
Ruth Brain – Intel Fellow, Director of Interconnect Technology and Integration
Ramune Nagisetty – Senior Principal Engineer, Director of Product and Process Integration
Boyd S. Phelps – Vice President, Client Engineering Group, General Manager, Client and Core Development Group
David Blythe – Intel Senior Fellow, Intel Architecture, Graphics and Software, Director, Graphics Architecture
Lisa Pearce – Vice President, Intel Architecture, Graphics, and Software, General Manager, Visual Technologies Team
Ravi Kuppuswamy – Vice President, General Manager, Custom Logic Engineering
Hong Hou – Vice President, General Manager, Connectivity Group
Martin Dixon – Intel Fellow, Intel Product Assurance and Security Director, Architecture
Wei Li – Vice President, Intel Architecture, Graphics and Software, General Manager, Machine Learning and Translation
Sailesh Kottapalli – Intel Senior Fellow, Intel Architecture, Graphics and Software, Chief Architect, Datacenter Processor Architecture
Brijesh Tripathi – Vice President, Intel Architecture, Graphics and Software, General Manager, Client Architecture and Immersive Computing
Rich Uhlig – Intel Senior Fellow, Director, Intel Labs

Raja Koduri

Chief Architect



Architecture Day **2020**



DESIGN

TRANSISTOR COUPLED DESIGN

TRANSISTOR RESILIENT DESIGN

ARCHITECTURE

CPU CENTRIC

XPU CENTRIC

SOLUTION

SILICON

SILICON & SOFTWARE

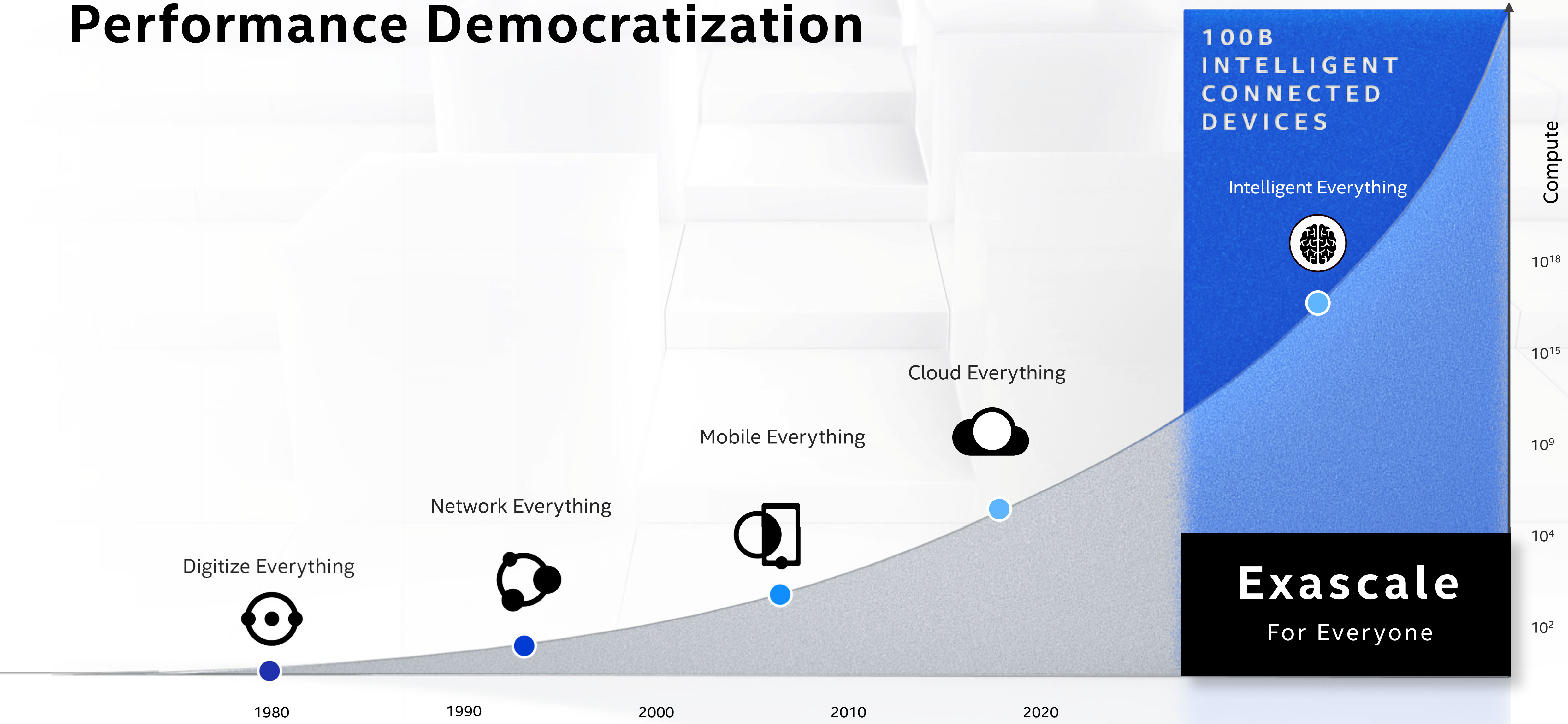
PURPOSE

To create **world-changing technology** that **enriches** the **lives** of every person on **earth**

VISION

We are on a journey
to be the trusted **performance** leader that
unleashes the potential of **data**

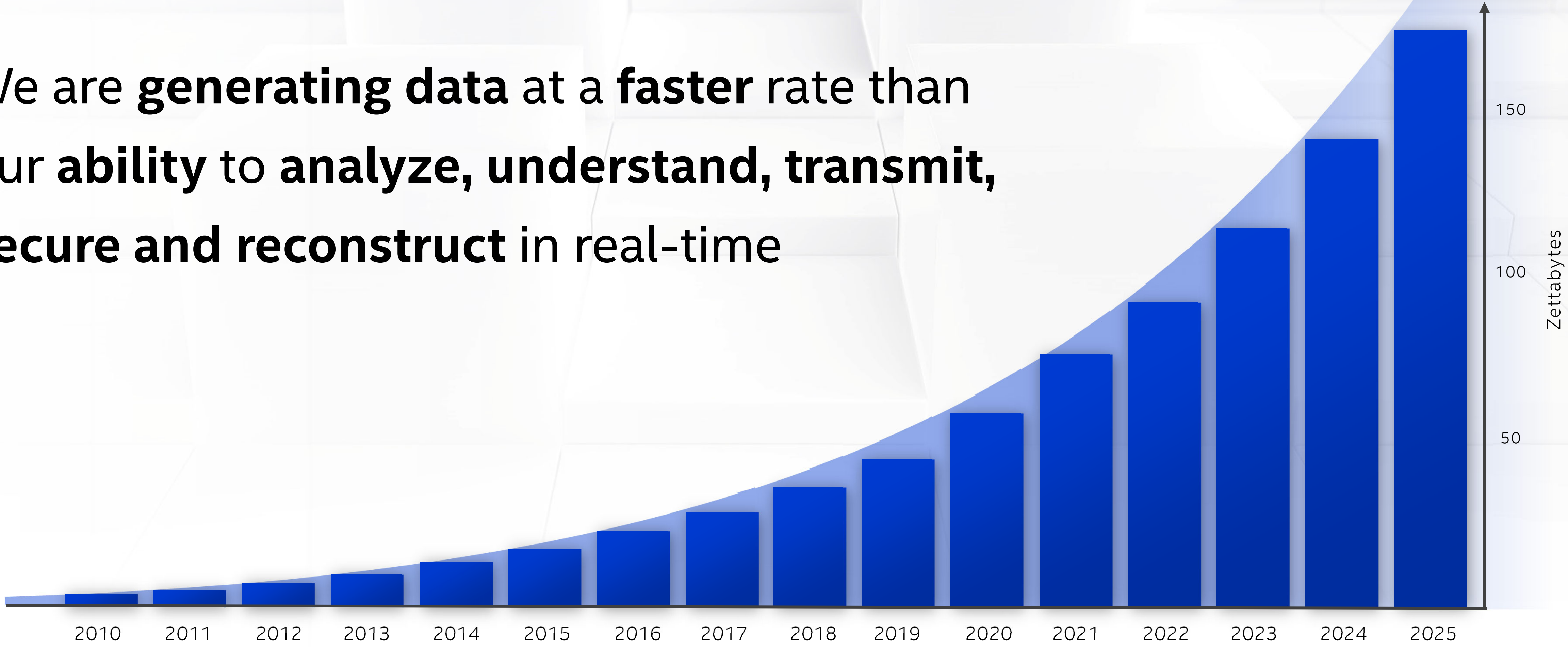
Performance Democratization



The Data Problem

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

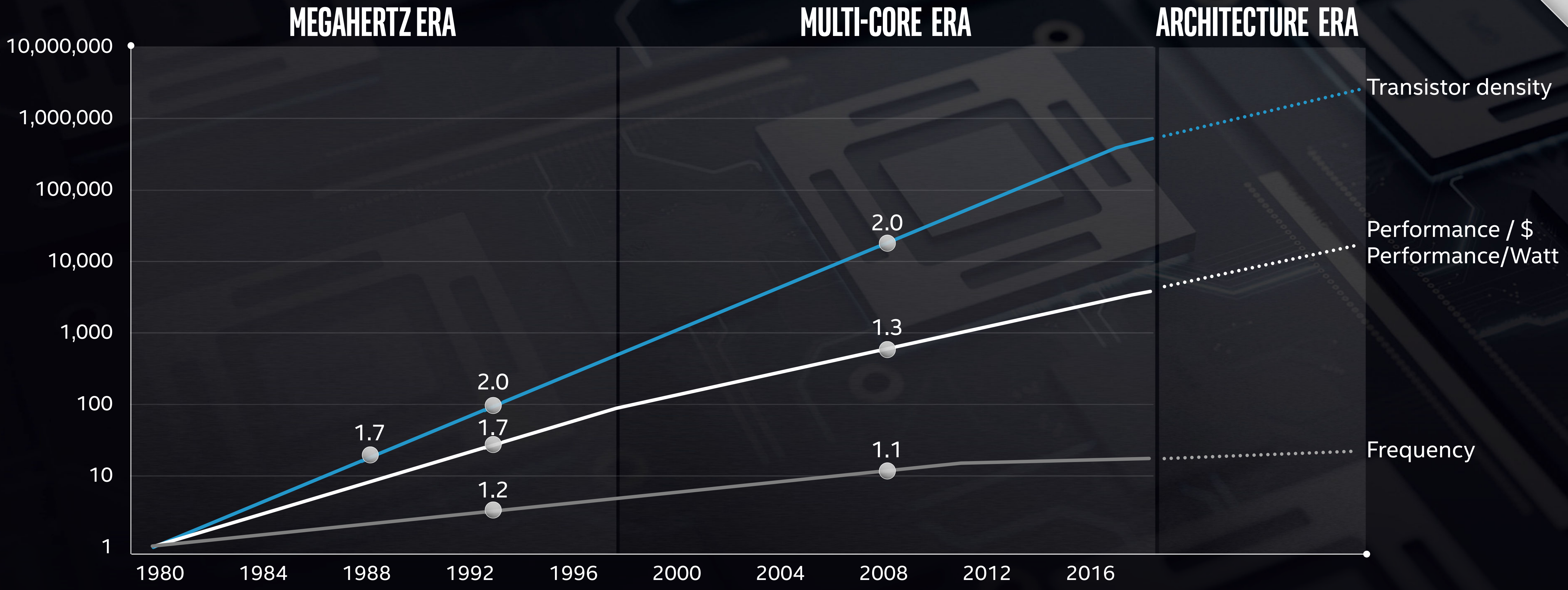
175ZB



Source: IDC Global DataSphere, May 2020

MOORE'S LAW – EXPONENTIAL ENTITLEMENT

WHAT WE SAID IN 2018



TECHNOLOGY PILLARS

- SOFTWARE
- SECURITY
- INTERCONNECT
- MEMORY
- XPU ARCHITECTURE
- PROCESS & PACKAGING

TECHNOLOGY PILLARS

SOFTWARE

SECURITY

INTERCONNECT

MEMORY

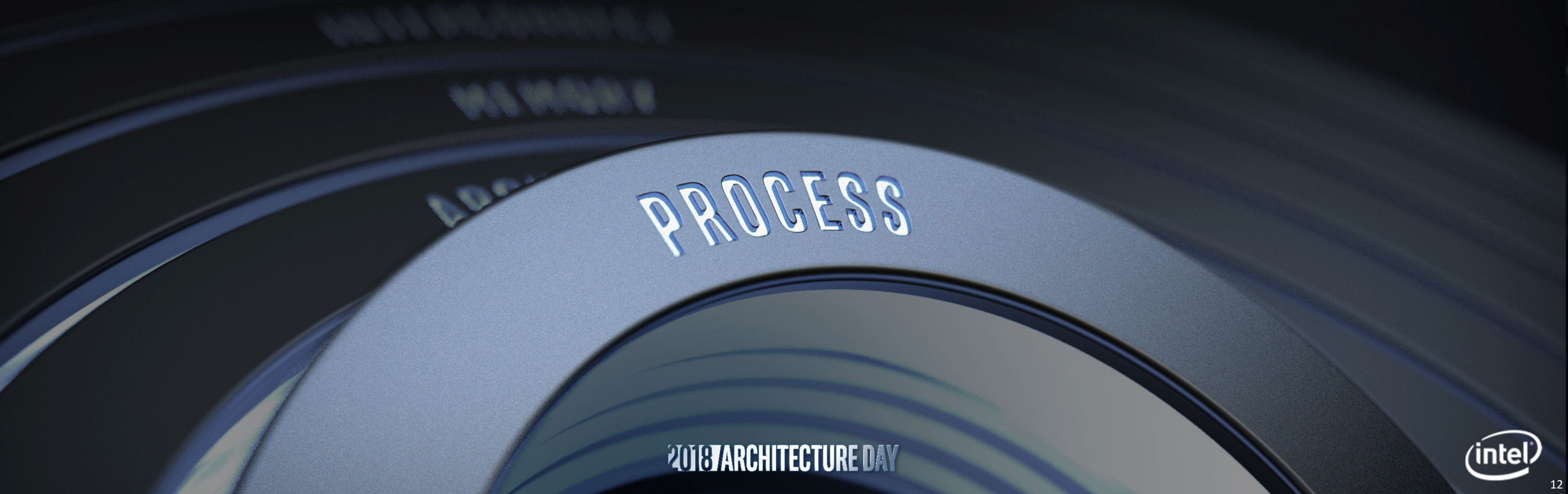
XPU ARCHITECTURE

PROCESS & PACKAGING

INTEL PROCESS AND PACKAGING

WHAT WE SAID IN 2018

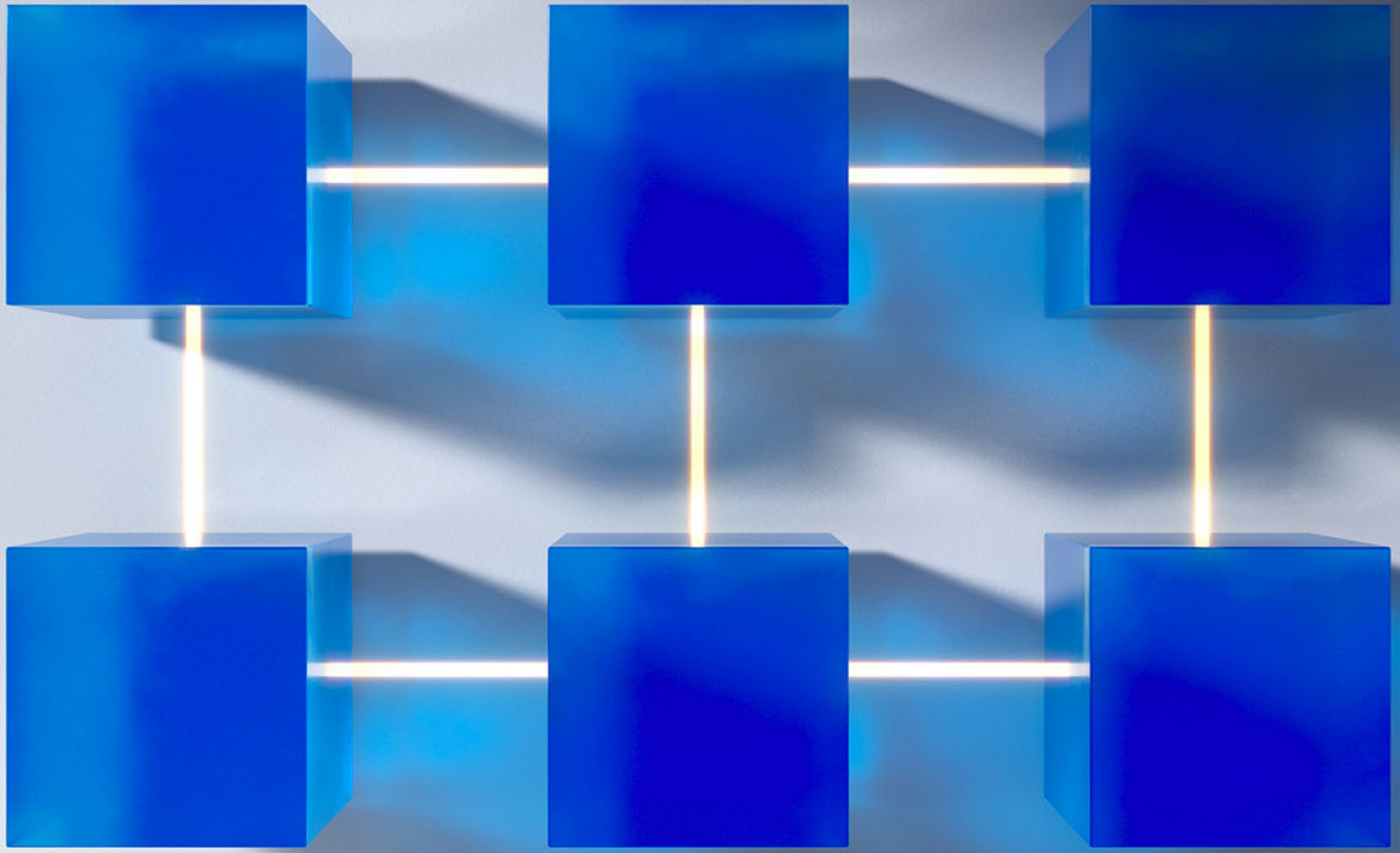
Synchronized and co-architected advances of transistors, packaging and designs are essential for the future of Moore's Law



Process Innovations

Ruth Brain

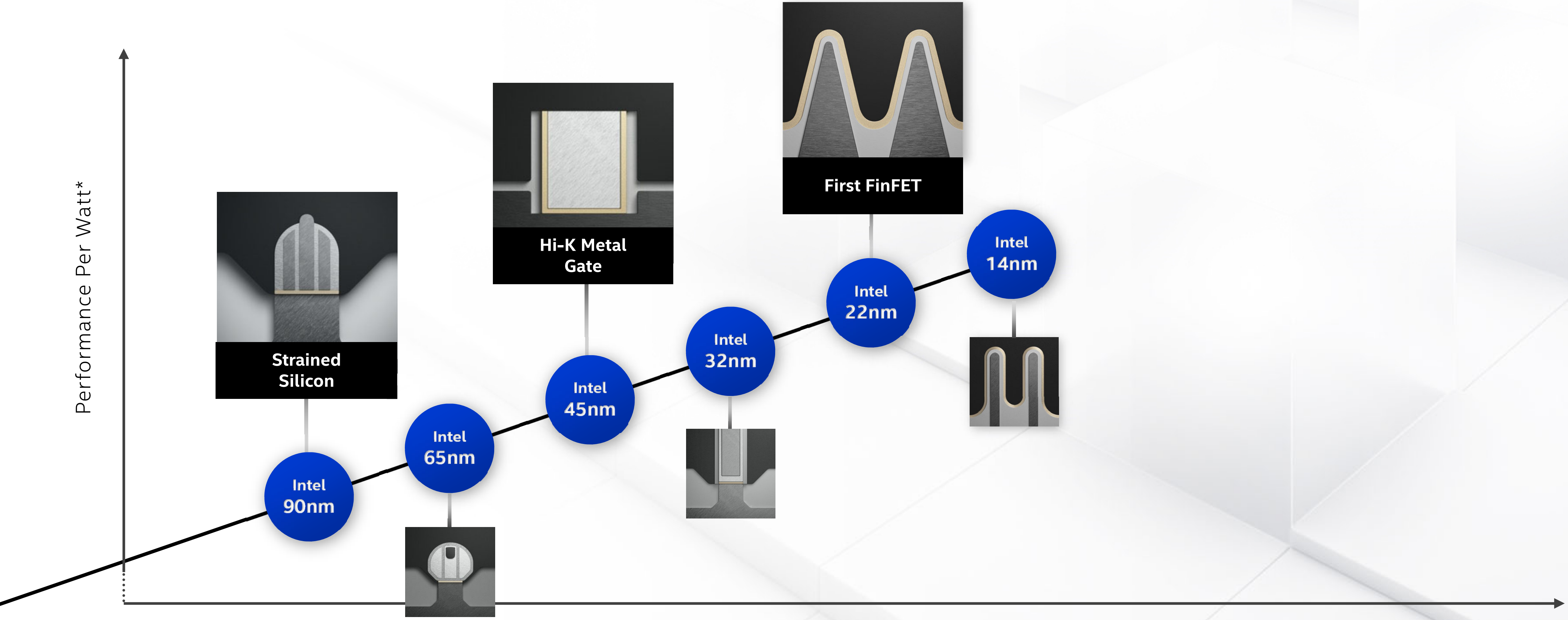
Intel Fellow,
Director of Interconnect
Technology and Integration



intel | TECHNOLOGY PILLARS

Architecture Day 2020

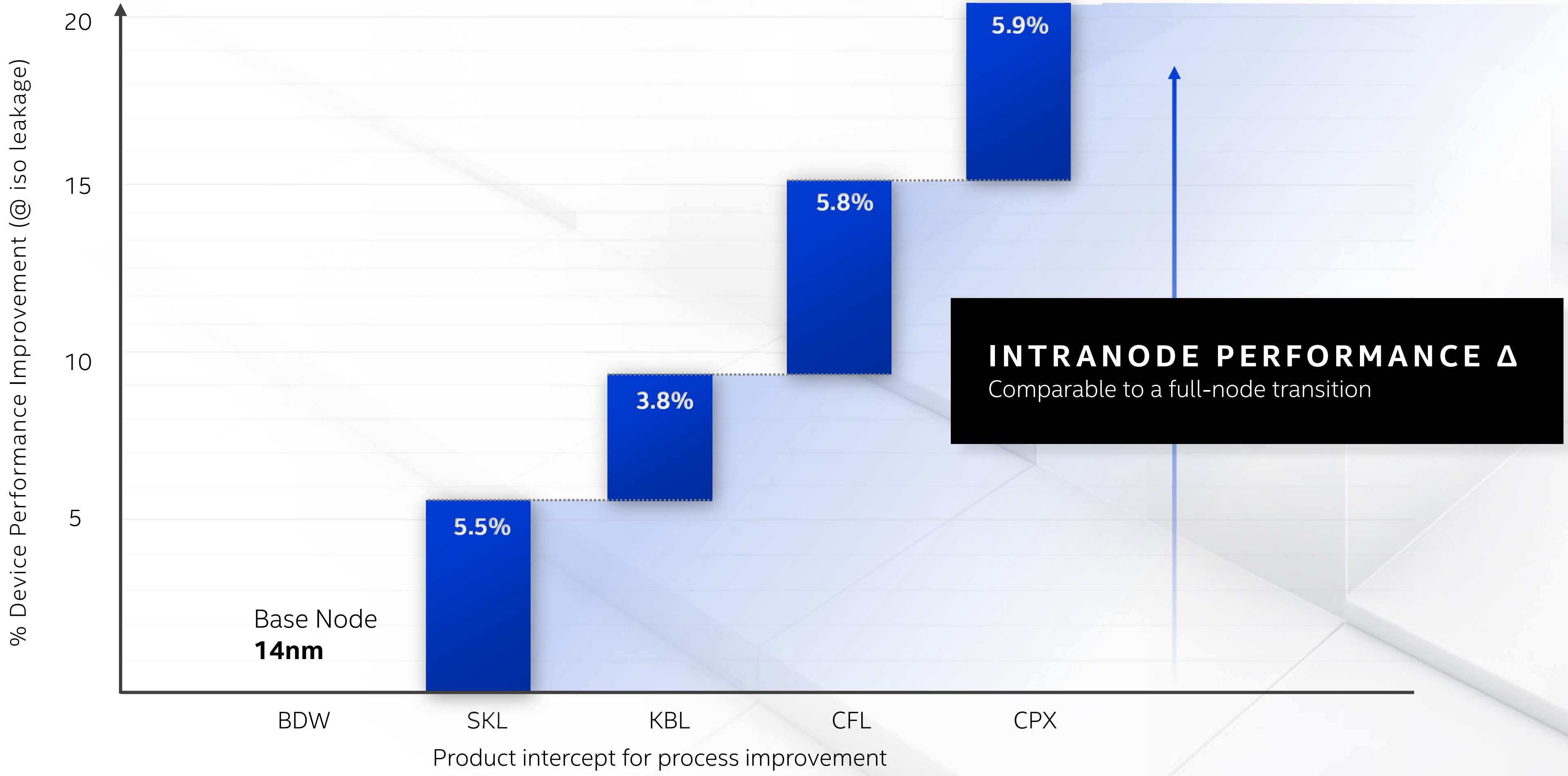
Process Technology Roadmap



*For illustrative purposes only

Refining the FinFET

Intranode enhancements deliver significant within-node performance boost

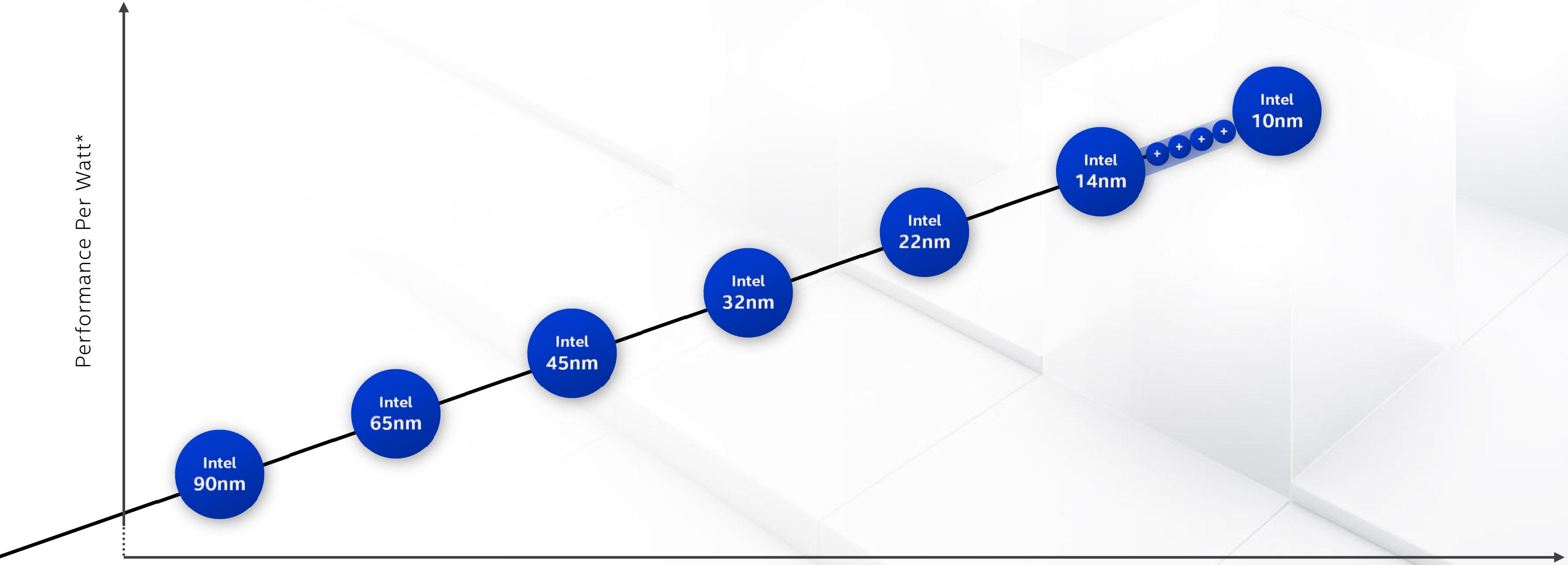


INTRANODE PERFORMANCE Δ

Comparable to a full-node transition

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Process Technology Roadmap



*For illustrative purposes only

Refining the FinFET

10nm innovations move beyond the device

Innovations

SELF ALIGNED QUAD PATTERNING (SAQP)

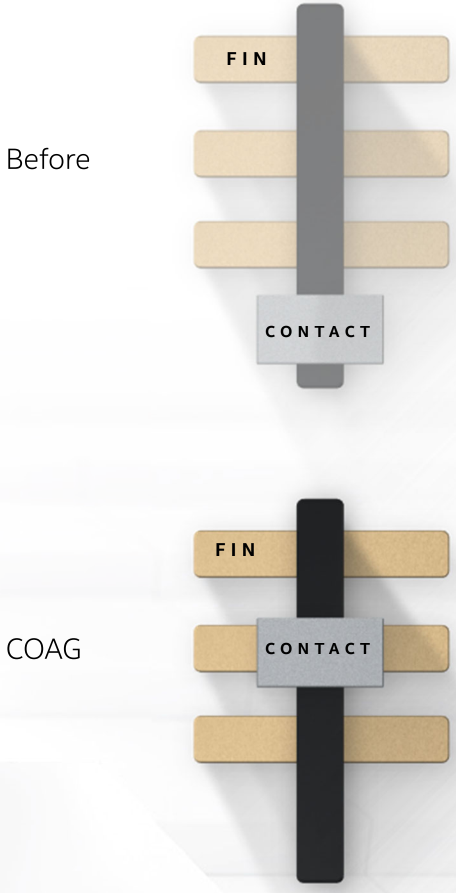
At M0 and M1 delivers 0.51x scaling of density critical layers

COBALT LOCAL INTERCONNECTS

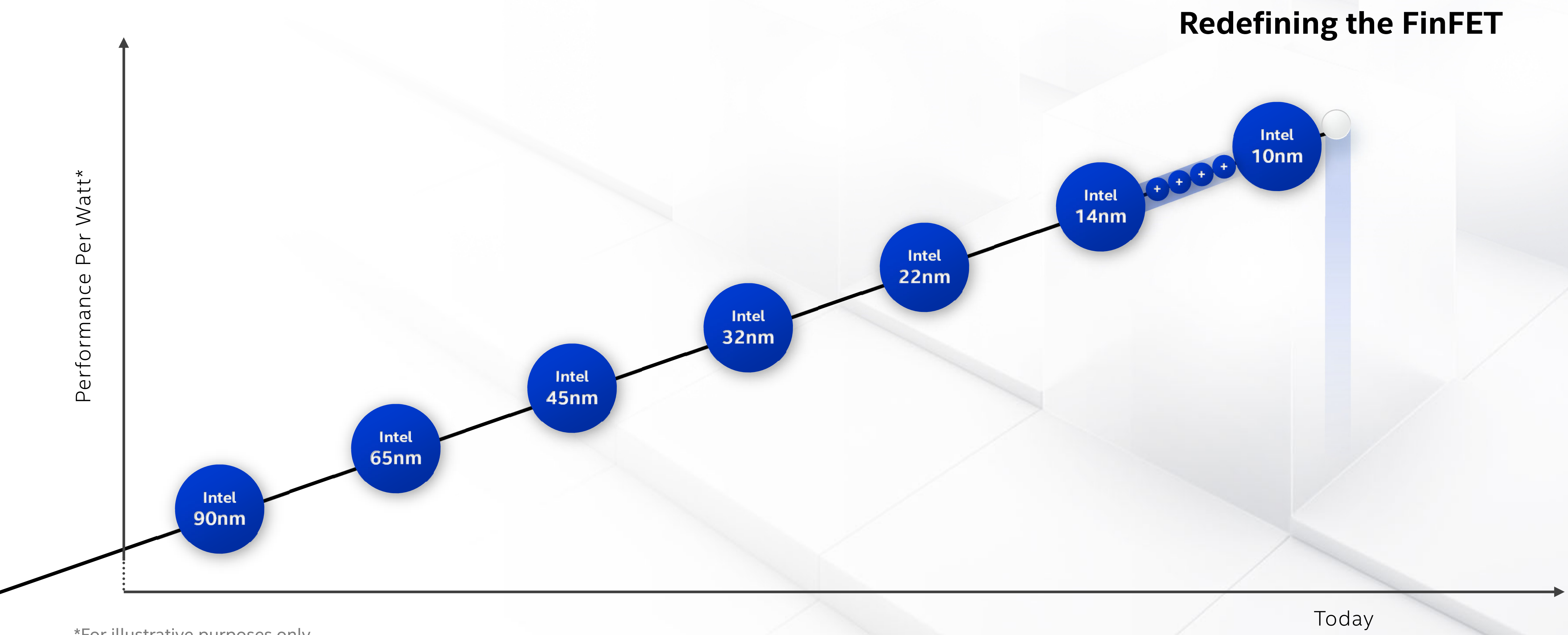
On lowest two metal layers result in 5-10x improvement in electromigration, 2x reduction in via resistance

CONTACT OVER ACTIVE GATE (COAG)

Smaller cell and improved transistor density



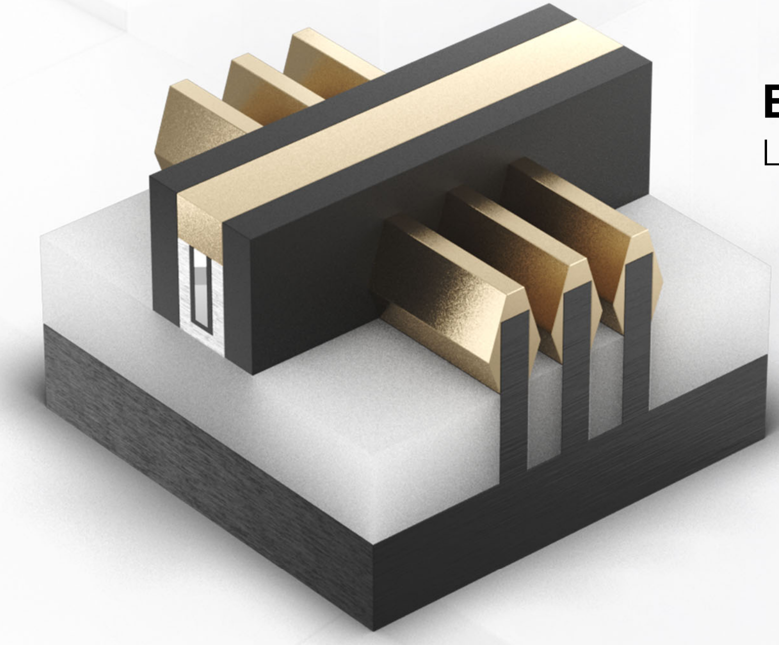
Process Technology Roadmap



*For illustrative purposes only

Redefining the FinFET

Innovation across the entire process stack, from channel to interconnects



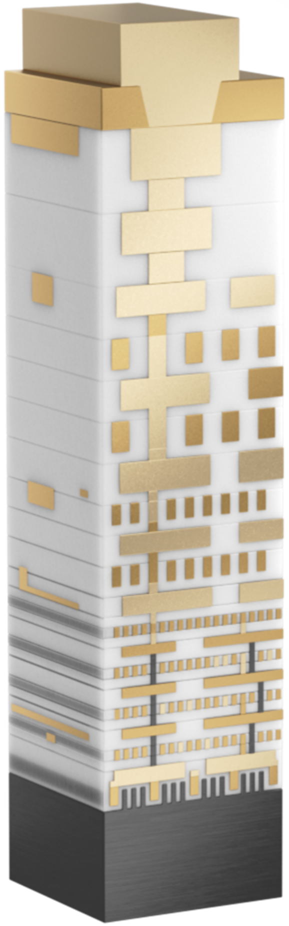
Additional Gate Pitch
Higher Drive Current

Improved Gate Process
Higher Channel Mobility

Enhanced Epitaxial Source/Drain
Lower Resistance, Increases Strain

Redefining the FinFET

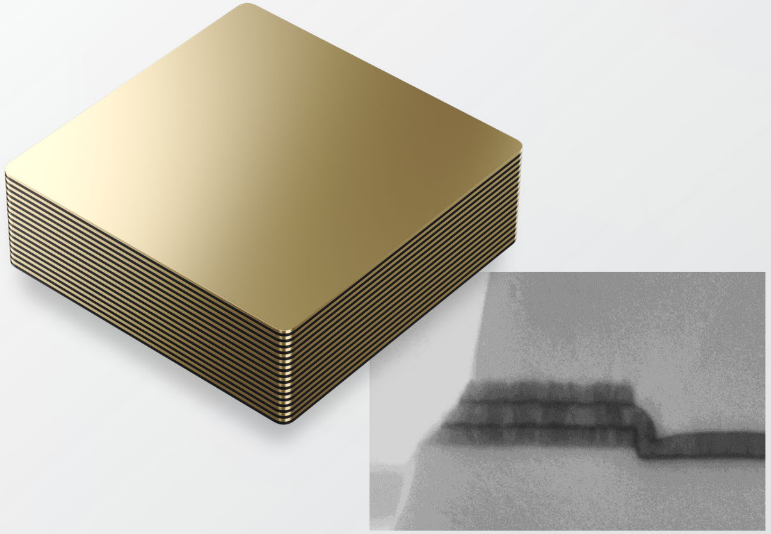
Innovation across the entire process stack, from channel to interconnects



Super MIM Capacitor
5x increase in MIM capacitance

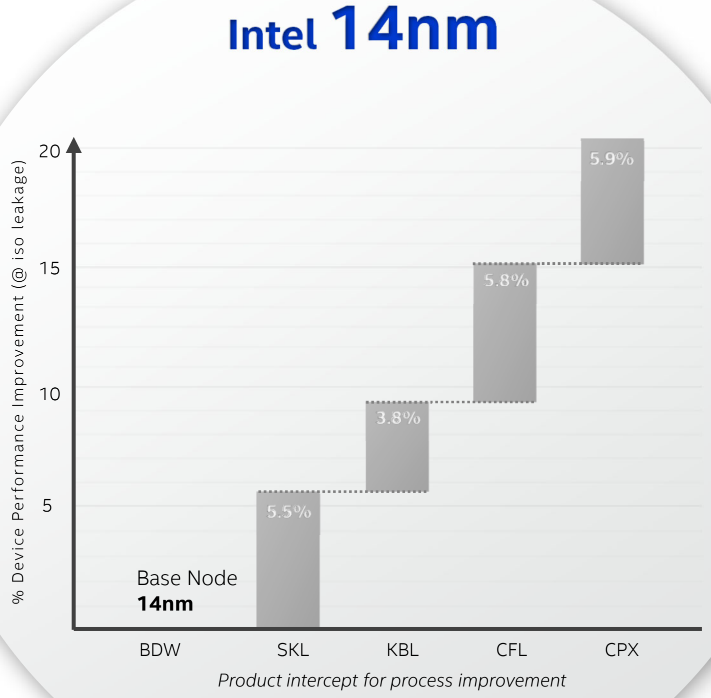
Novel Thin Barrier
reduces via resistance by 30%

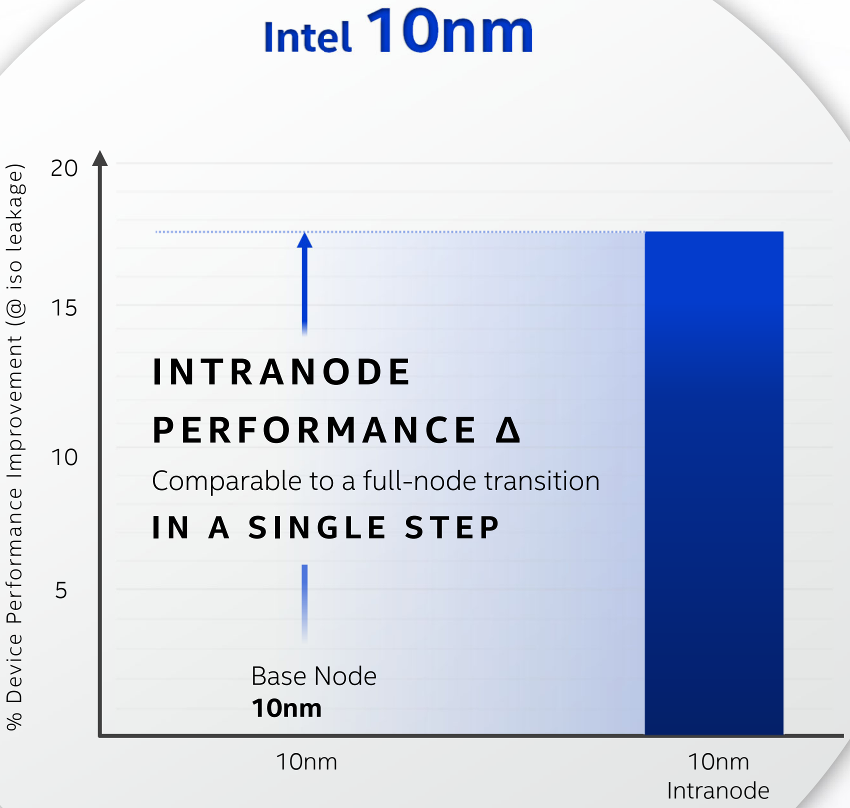
Thin layers of different Hi-K materials, each just a few Angstroms thick, stacked in a repeating "superlattice."



Largest Intranode Performance Delta In Our History

NEW DISCLOSURE



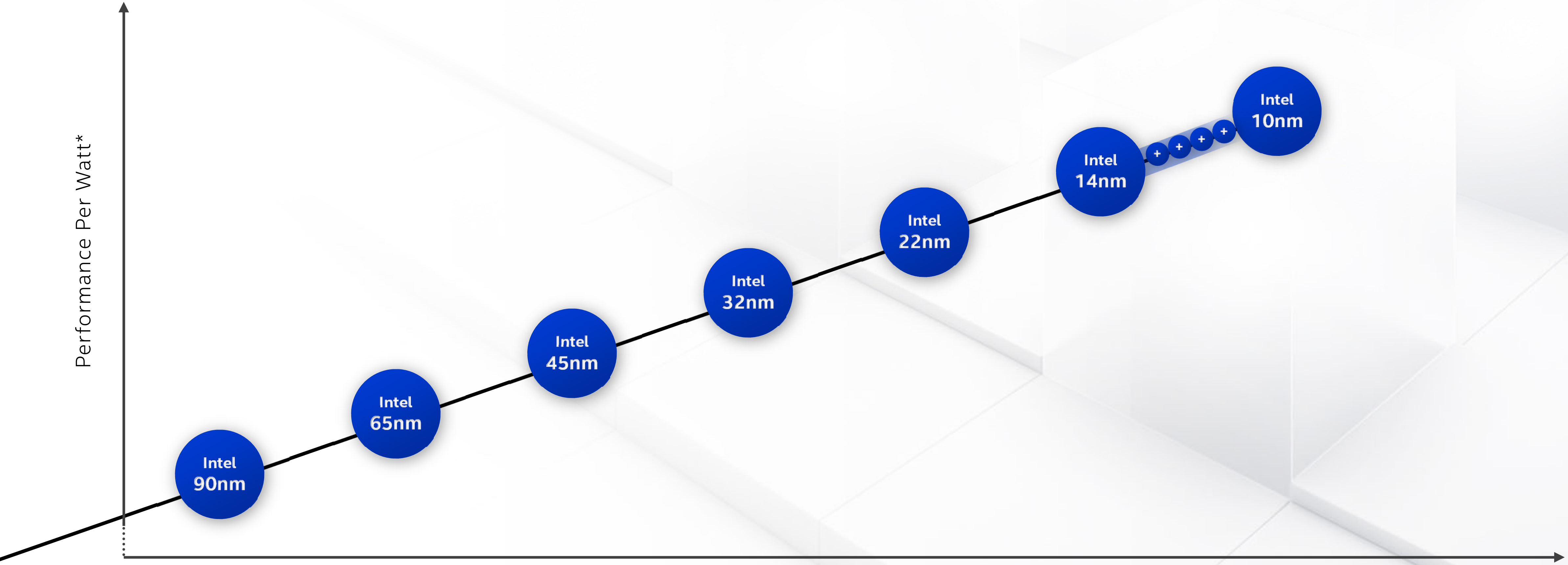


INTRANODE PERFORMANCE Δ

Comparable to a full-node transition

IN A SINGLE STEP

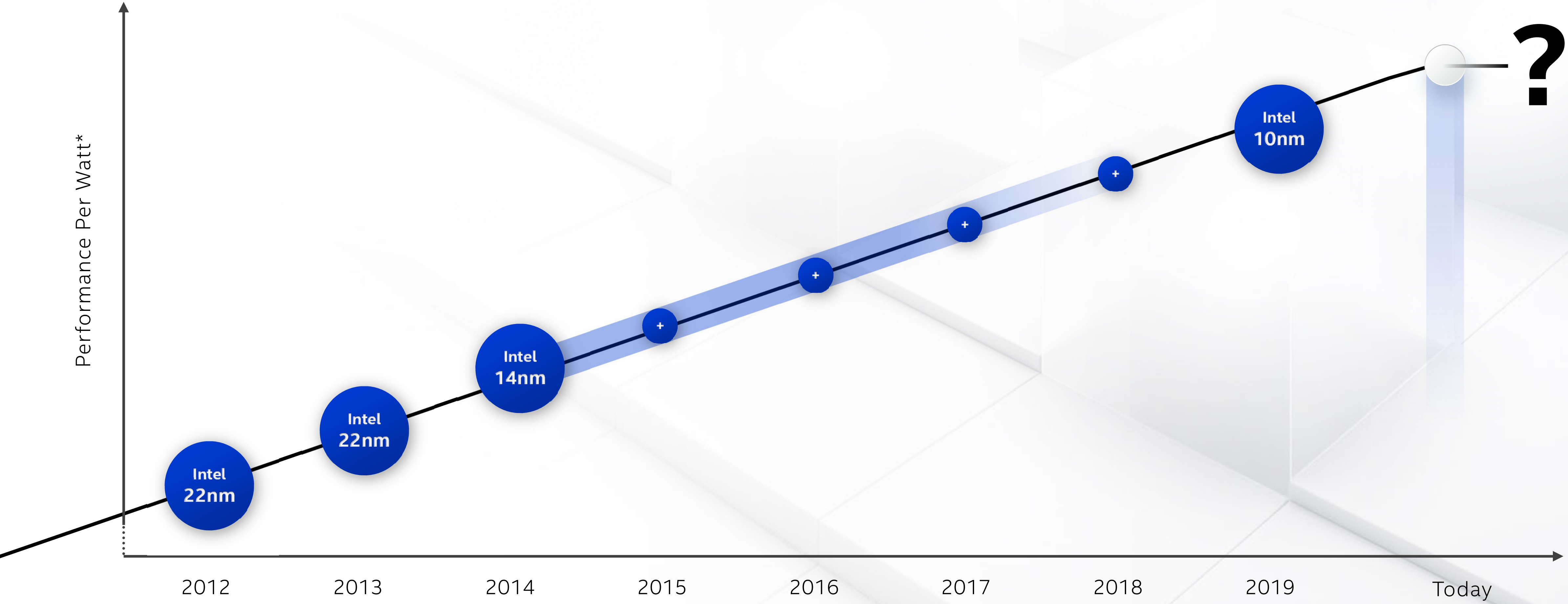
Process Nomenclature Problem



*For illustrative purposes only

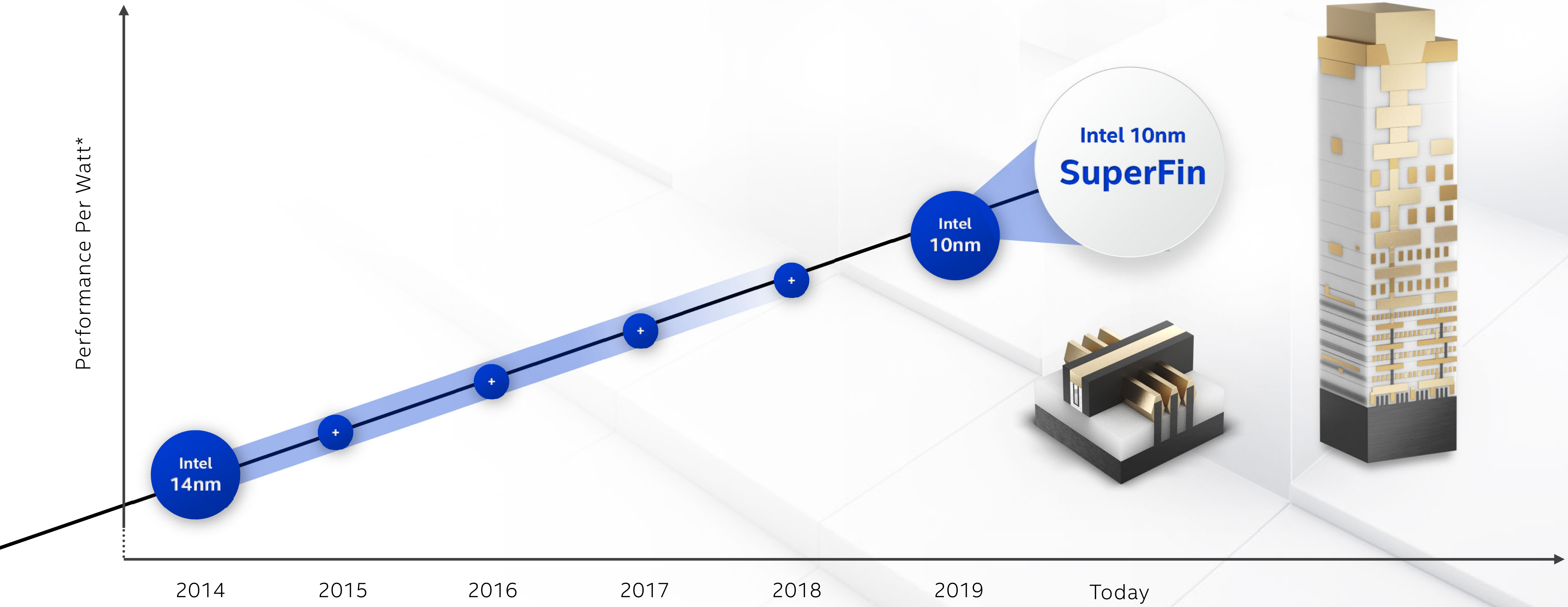
Process Nomenclature Problem

Redefining the FinFET



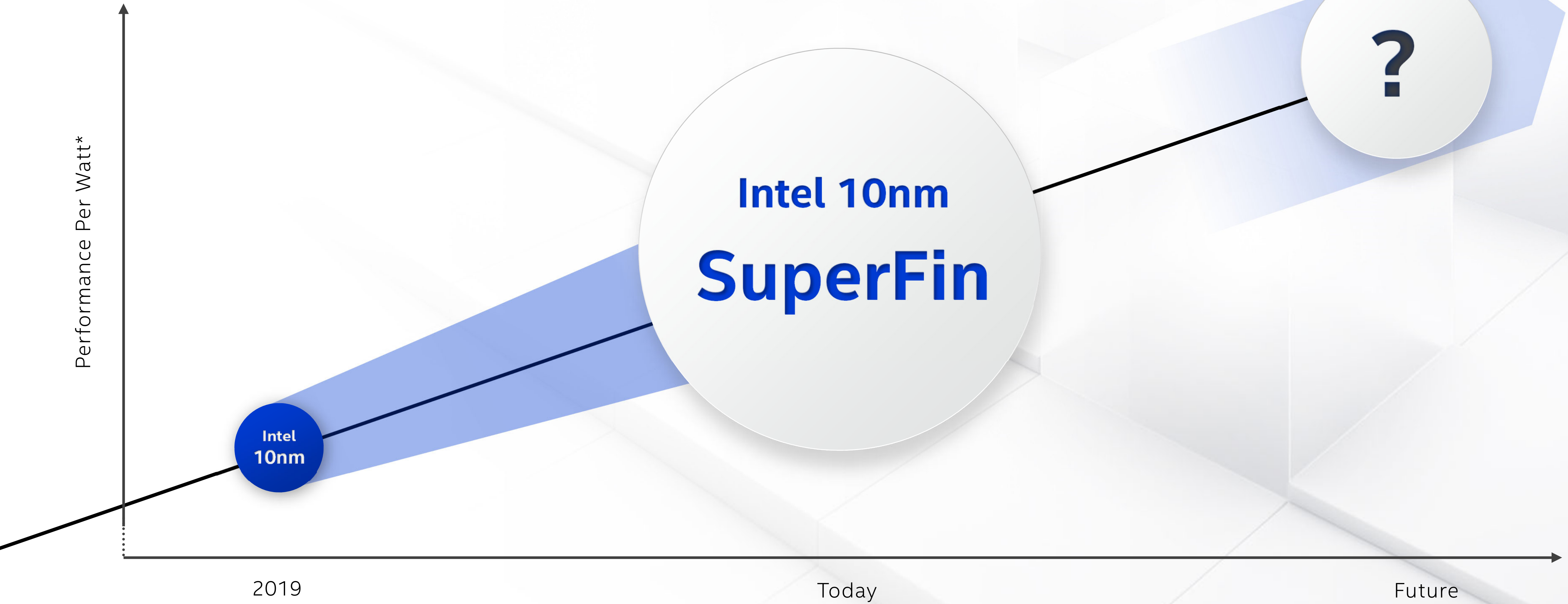
*For illustrative purposes only

SuperMIM + Redefined FinFET



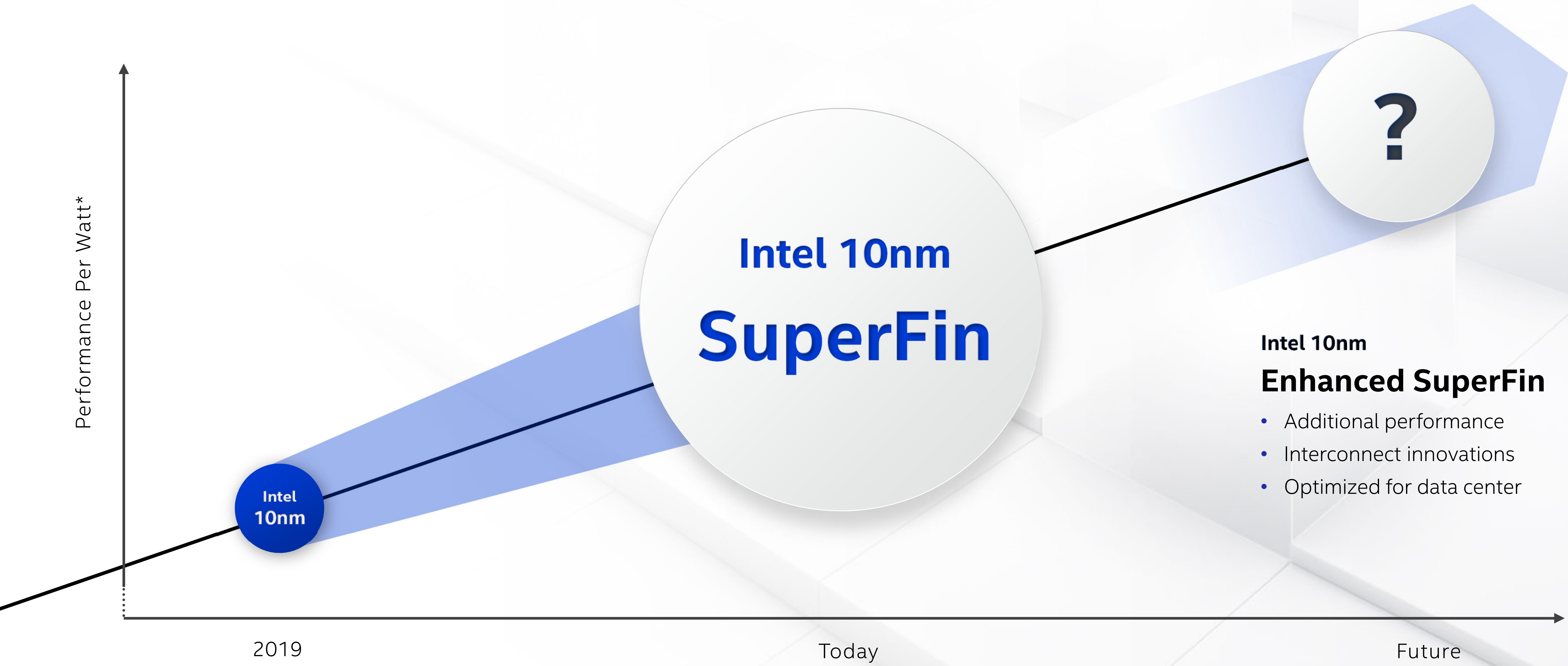
*For illustrative purposes only

Process Technology Roadmap



*For illustrative purposes only

Process Technology Roadmap



Intel 10nm

Enhanced SuperFin

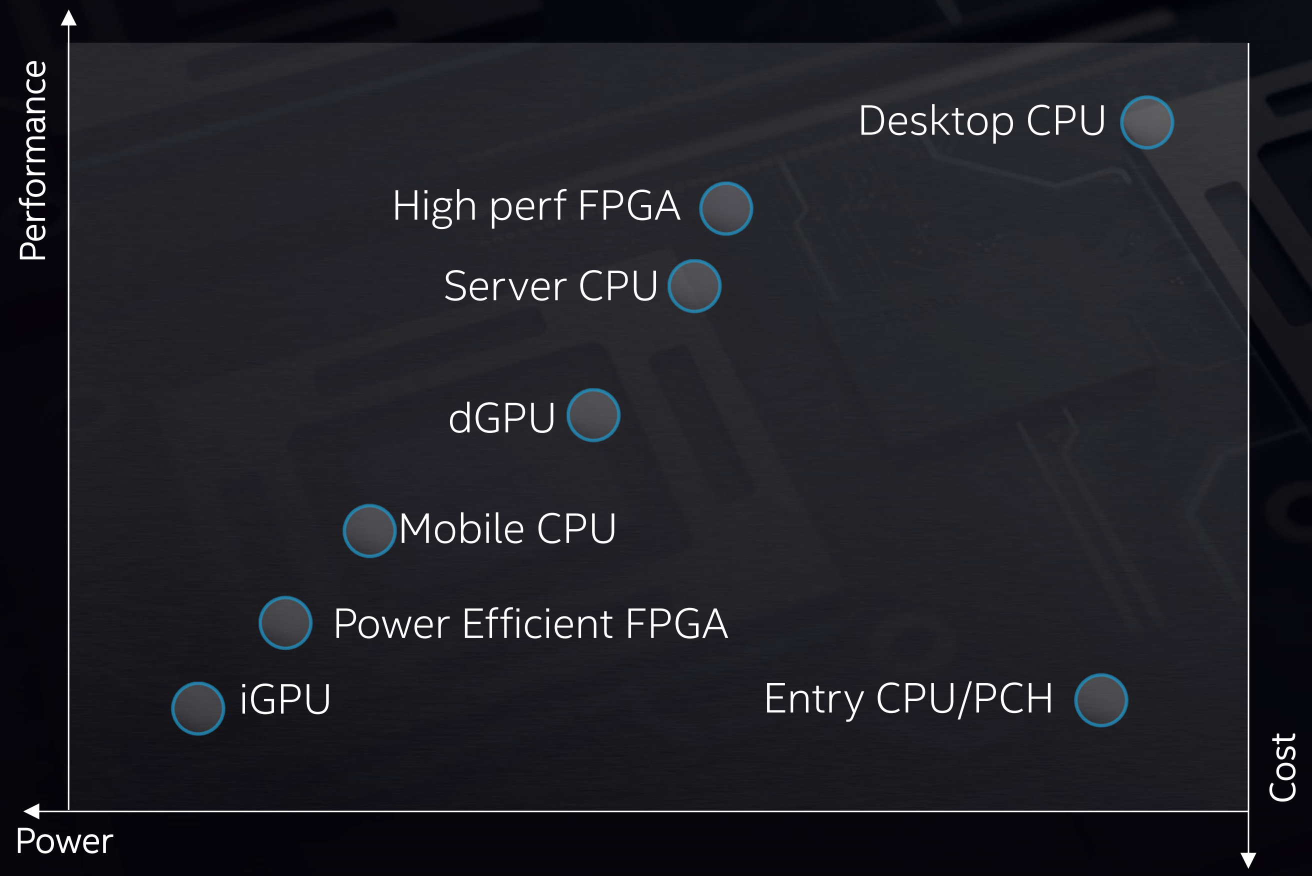
- Additional performance
- Interconnect innovations
- Optimized for data center

*For illustrative purposes only

CASE FOR ADVANCED PACKAGING

WHAT WE SAID IN 2018

Transistor design target range



Transistor diversity

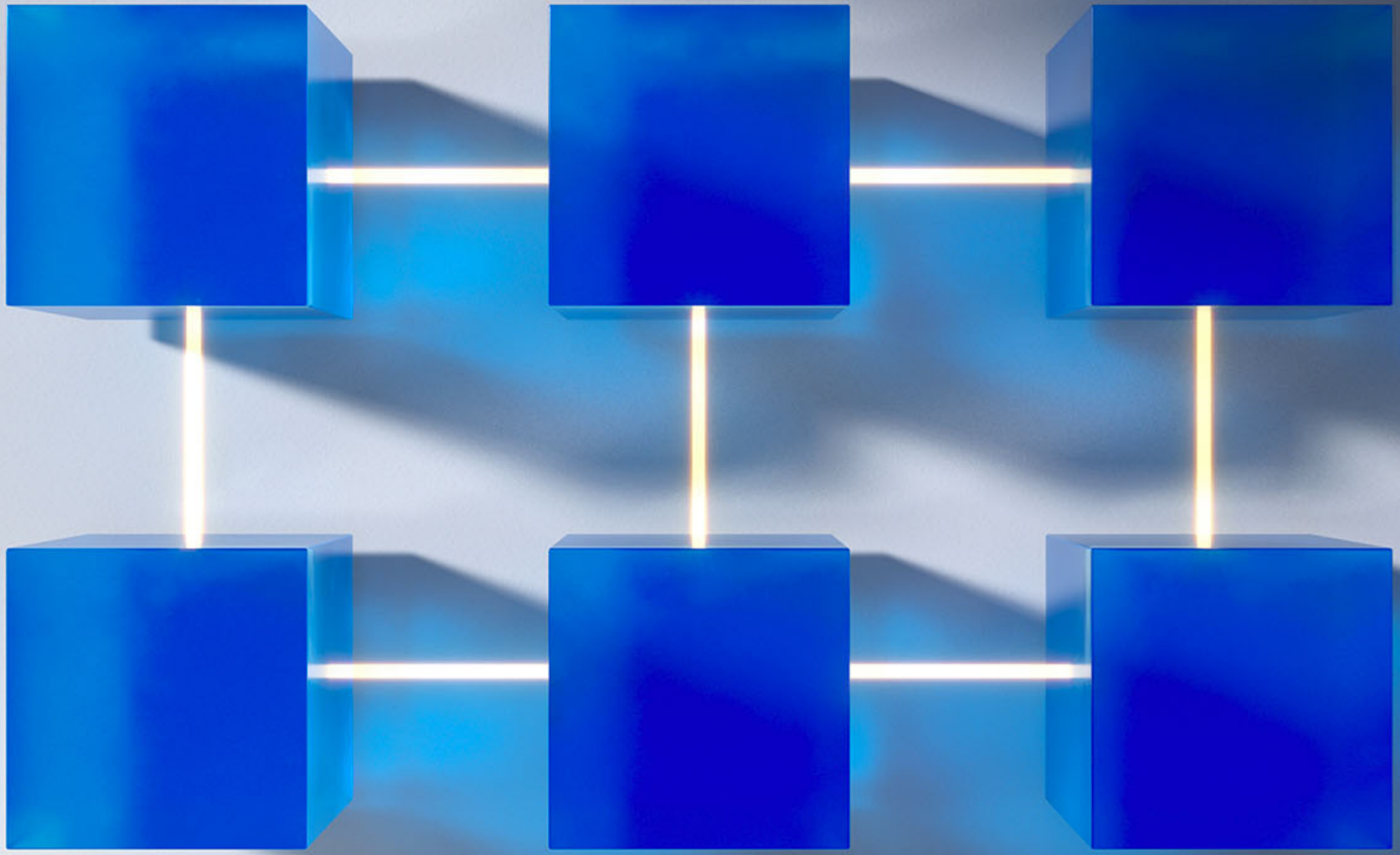
Logic Transistors
Analog/RF Transistors
High Speed Memory
Dense Memory
Non-Volatile Memory
High speed IO
Configuration Memory

No single transistor node is optimal across all design points

Advanced Packaging

Ramune Nagisetty

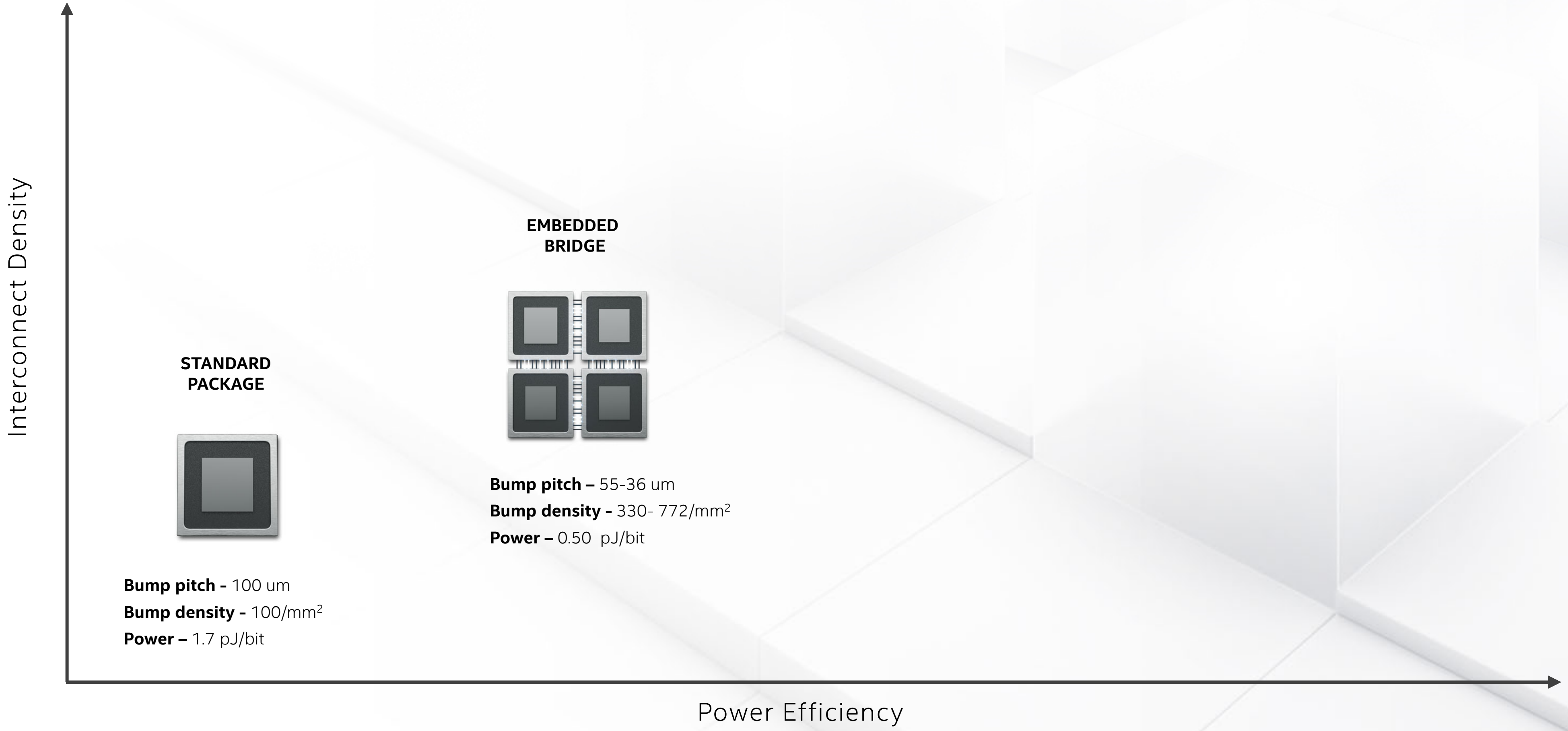
Senior Principal Engineer,
Director of Product and
Process Integration



TECHNOLOGY
PILLARS

Architecture Day **2020**

Packaging Technology Roadmap



STANDARD PACKAGE

Bump pitch - 100 um
Bump density - 100/mm^2
Power – 1.7 pJ/bit

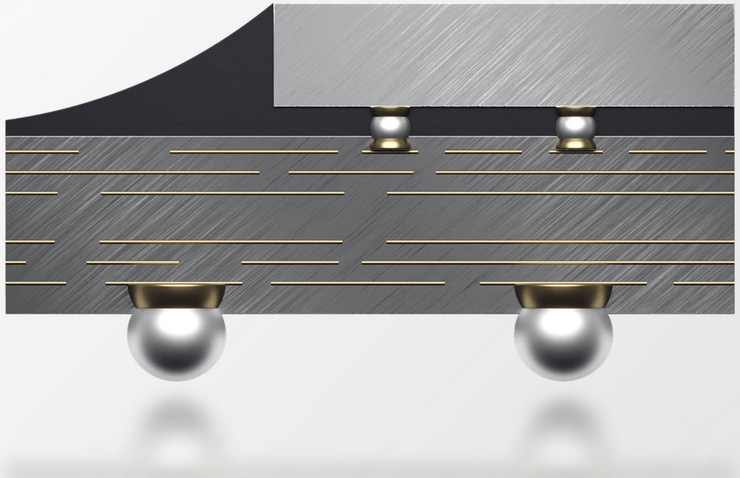
EMBEDDED BRIDGE

Bump pitch – 55-36 um
Bump density - 330- 772/mm^2
Power – 0.50 pJ/bit

2.5D Density Scaling

STANDARD PACKAGE

Typical organic package (FCBGA)



IO/mm/lyr

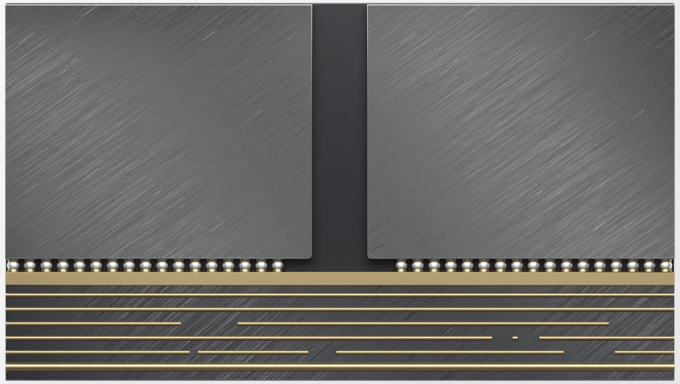
32 → 48

Flip chip bump pitch

100 µm

EMBEDDED BRIDGE

Embedding die with dense wiring in a package cavity



IO/mm/lyr

250 → 1000

HBM µ bump pitch

55 µm

Packaging Technology Roadmap

Interconnect Density

STANDARD PACKAGE

Bump pitch - 100 um
Bump density - 100/mm²
Power – 1.7 pJ/bit

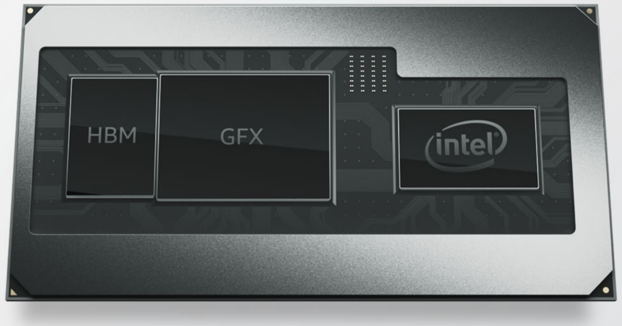
EMBEDDED BRIDGE

Bump pitch – 55-36 um
Bump density - 330- 772/mm²
Power – 0.50 pJ/bit

KABY LAKE G

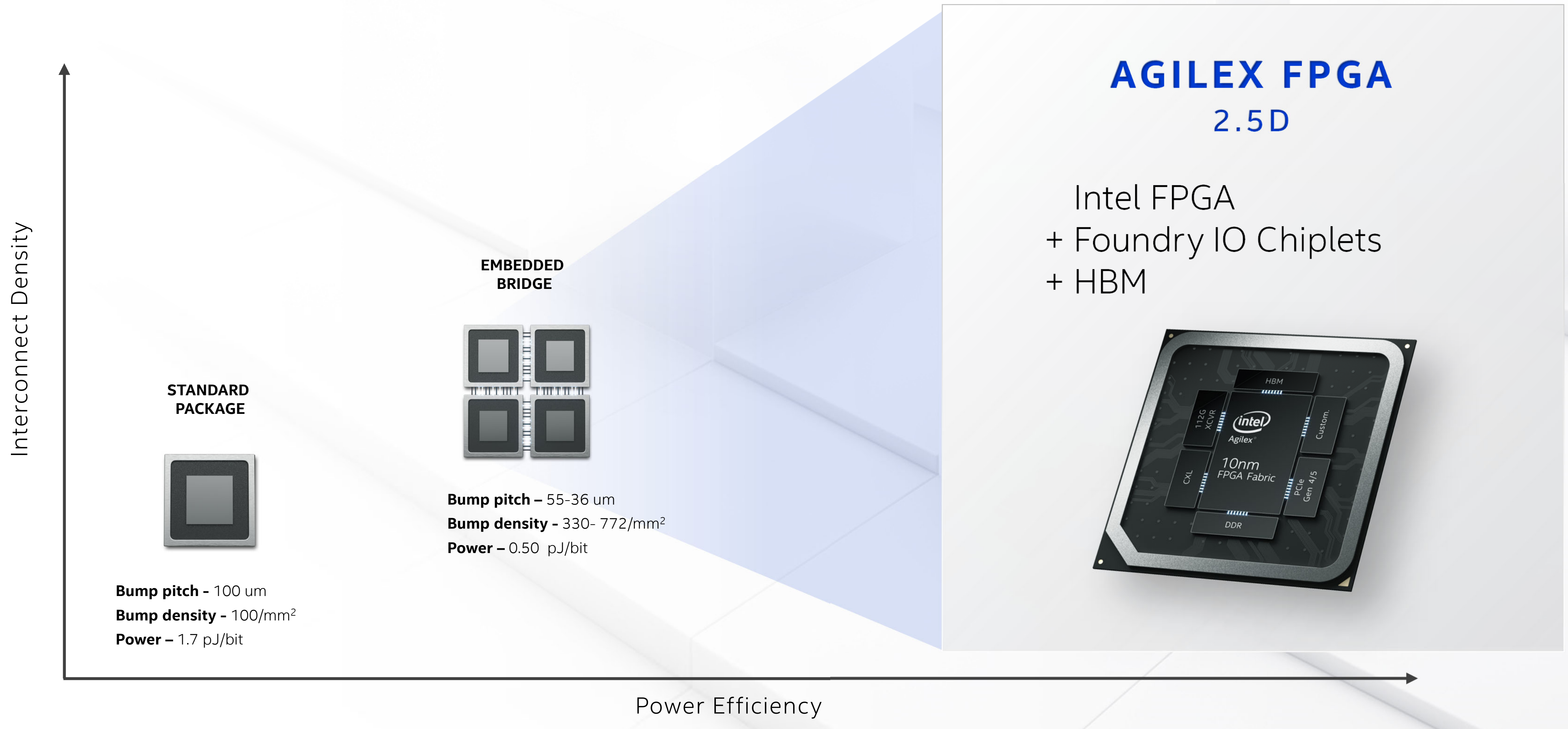
2D + 2.5D

Intel CPU
+ Foundry GFX
+ HBM



Power Efficiency

Packaging Technology Roadmap



STANDARD
PACKAGE

Bump pitch - 100 um
Bump density - 100/mm^2
Power – 1.7 pJ/bit

EMBEDDED
BRIDGE

Bump pitch – 55-36 um
Bump density - 330- 772/mm^2
Power – 0.50 pJ/bit

AGILEX FPGA

2.5D

Intel FPGA
\+ Foundry IO Chiplets
\+ HBM

Enabling an Ecosystem

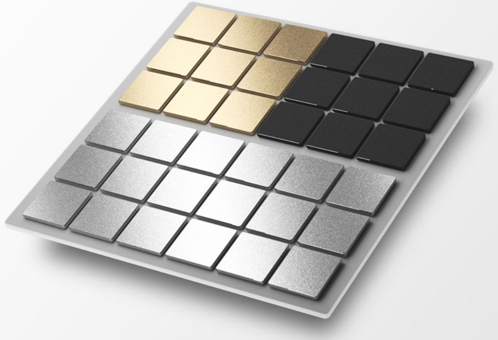
Integration enabled by standards and business models

BOARD LEVEL INTEGRATION

Standardized motherboard interfaces enable the PC ecosystem

PCIe **DDR** • • •

PACKAGE LEVEL INTEGRATION

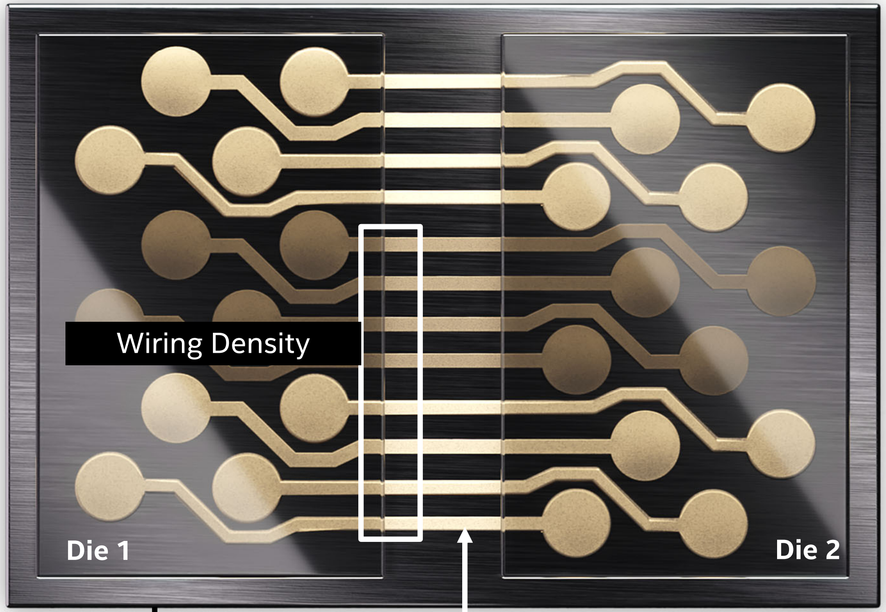


Die-to-die interfaces to enable a chiplet ecosystem

ADVANCED INTERFACE BUS (AIB)

Enabling an Ecosystem

High density die-to-die interconnects



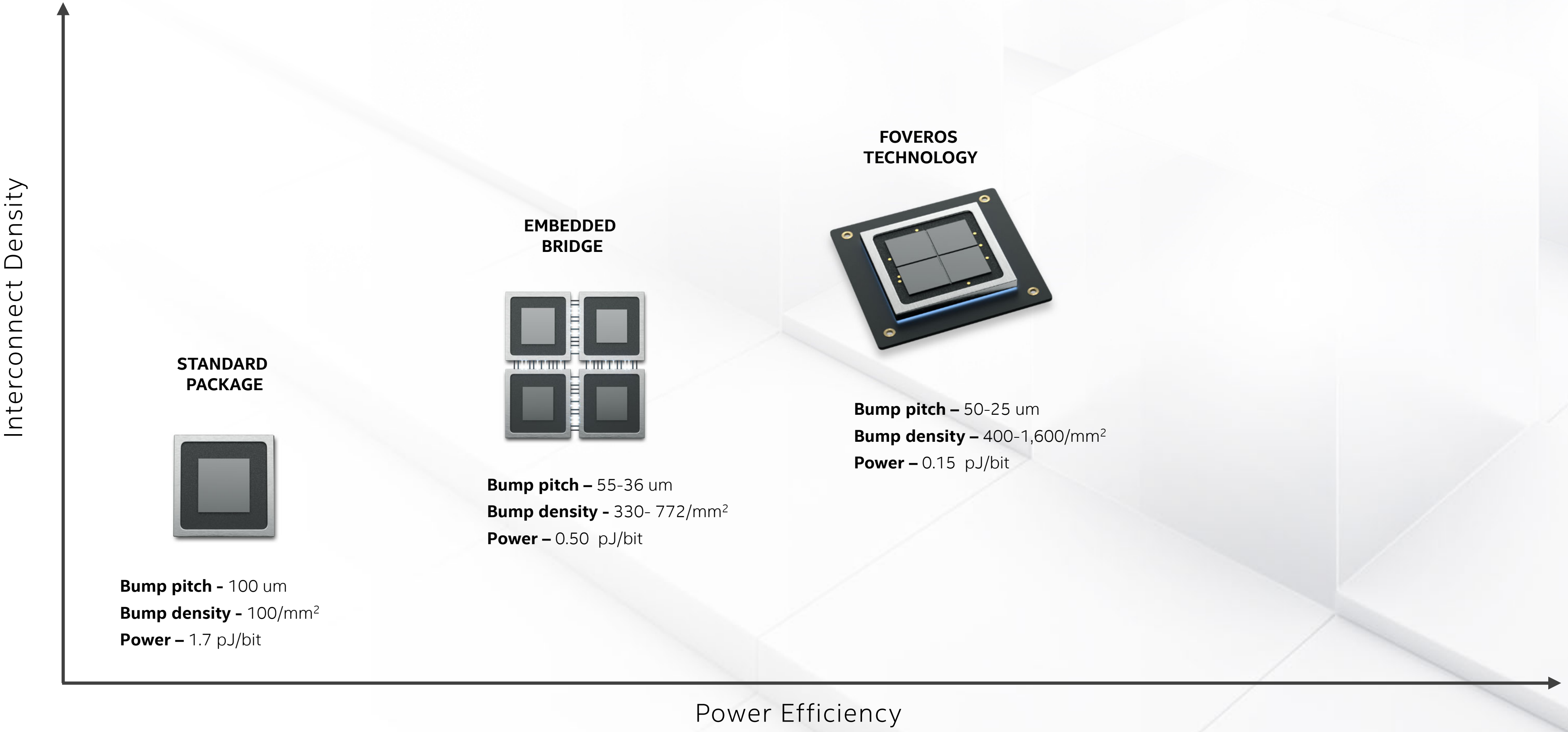
Package

Wire Width, Spacing & Dielectrics between wires

Feature	AIB 1.0	AIB 2.0
Bandwidth/wire (Gbps)	2	Up to 6.4
Bump density (um)	55	55/45/36
Bandwidth/mm shoreline (Gbps/mm)	256	1638
IO Voltage (V)	0.90	0.90/0.40
Energy/bit (pJ/bit)	0.85	0.50
Backward Compatibility	n/a	1.0

AIB Generator available at:
github.com/chipsalliance

Packaging Technology Roadmap



STANDARD PACKAGE

Bump pitch - 100 um
Bump density - 100/mm^2
Power – 1.7 pJ/bit

EMBEDDED BRIDGE

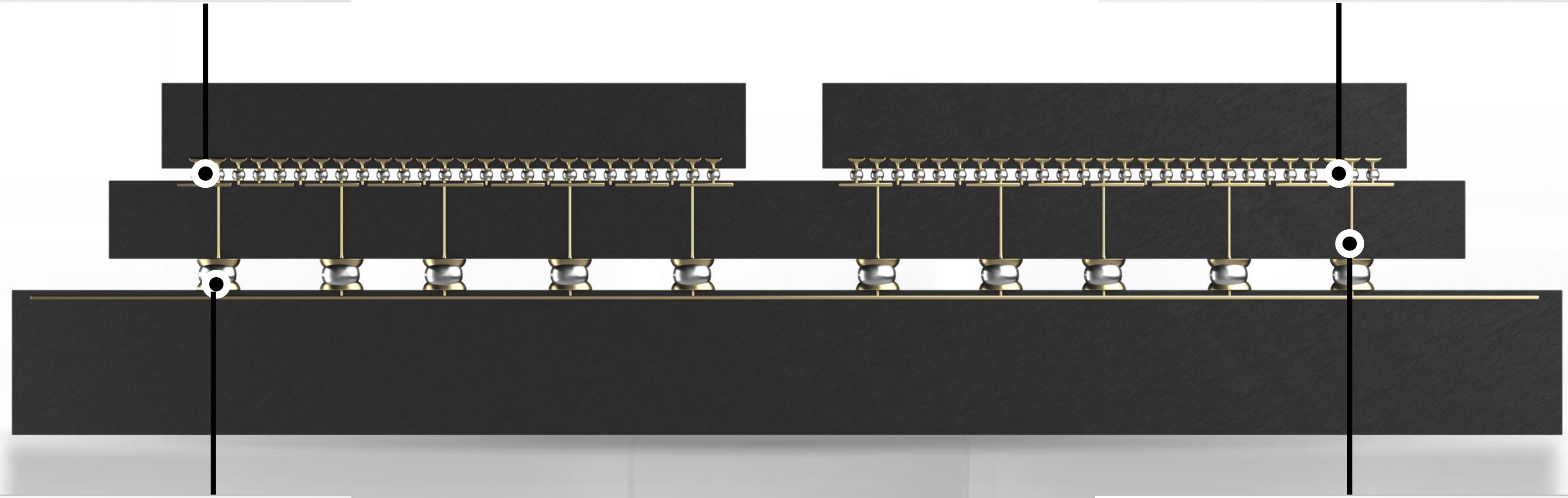
Bump pitch – 55-36 um
Bump density - 330- 772/mm^2
Power – 0.50 pJ/bit

FOVEROS TECHNOLOGY

Bump pitch – 50-25 um
Bump density – 400-1,600/mm^2
Power – 0.15 pJ/bit

3D Density Scaling

Foveros enables nearly 1,000 IO/mm2



Die to wafer stacking, face to face connection

Fine micro-bumps optimized for dense 3D and low power

Silicon to package standard C4 connection

Coarse TSV optimized for power delivery and off-package I/O

Packaging Technology Roadmap

Interconnect Density

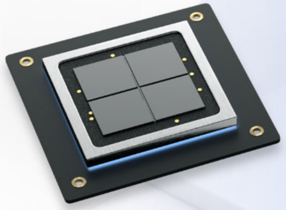
STANDARD PACKAGE

Bump pitch - 100 um
Bump density - 100/mm^2
Power – 1.7 pJ/bit

EMBEDDED BRIDGE

Bump pitch – 55-36 um
Bump density - 330- 772/mm^2
Power – 0.50 pJ/bit

FOVEROS TECHNOLOGY



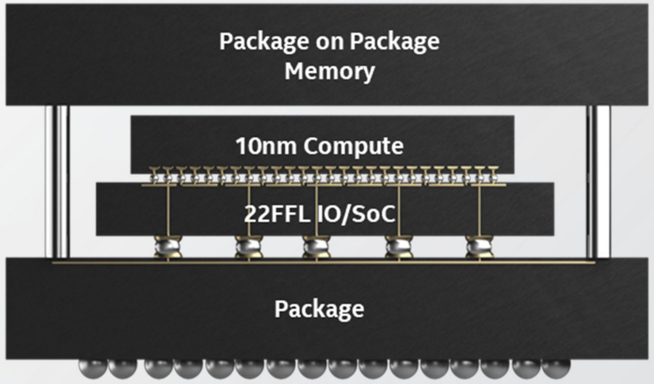
Bump pitch – 50-25 um
Bump density – 400-1,600/mm^2
Power – 0.15 pJ/bit

Power Efficiency

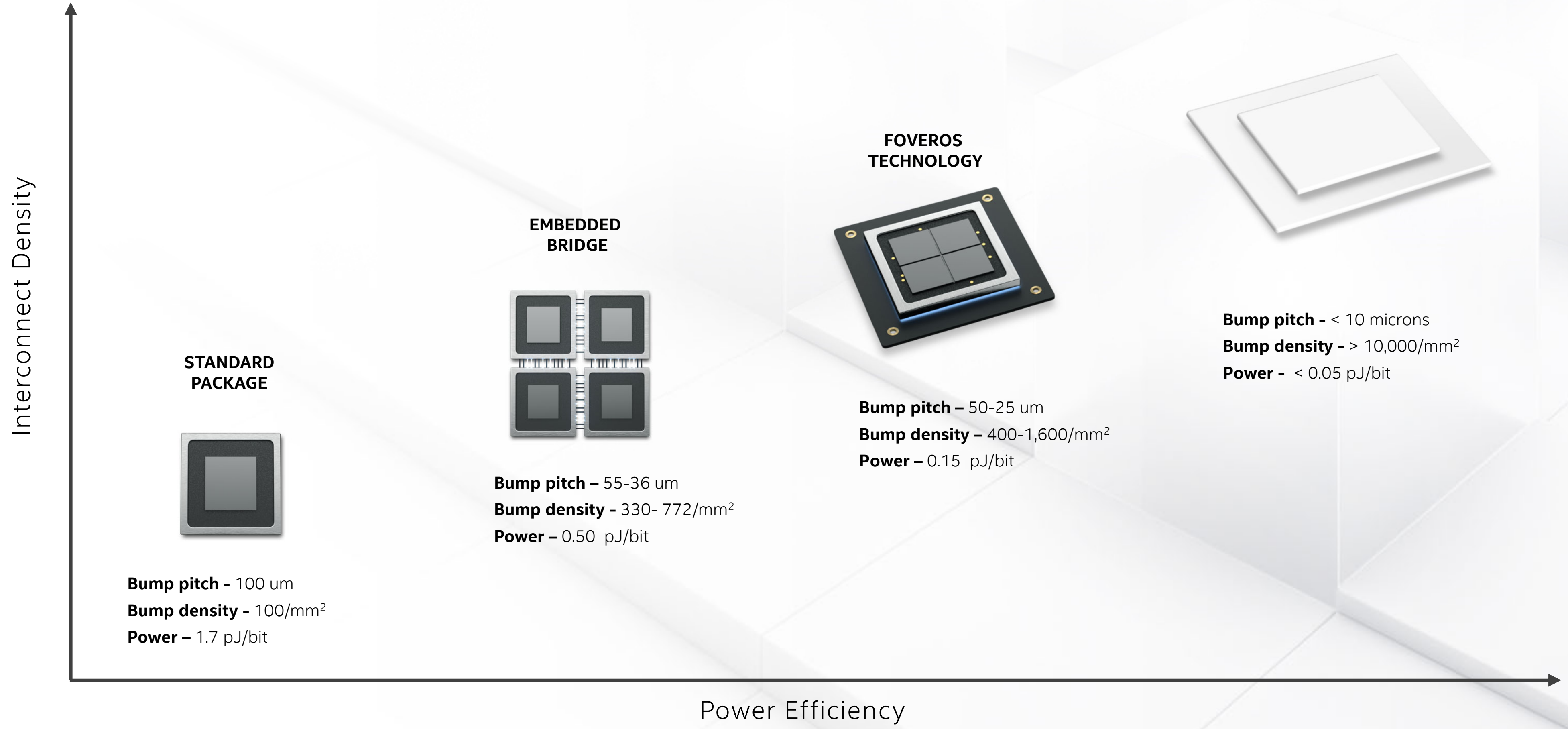
LAKEFIELD

3D

Internal Silicon on Multiple Nodes



Packaging Technology Roadmap



FUTURE

Bump pitch - < 10 micron

STANDARD PACKAGE

Bump pitch - 100 um
Bump density - 100/mm^2
Power – 1.7 pJ/bit

EMBEDDED BRIDGE

Bump pitch – 55-36 um
Bump density - 330- 772/mm^2
Power – 0.50 pJ/bit

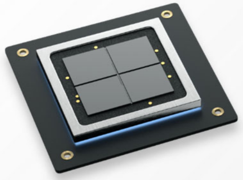
FOVEROS TECHNOLOGY

Bump pitch – 50-25 um
Bump density – 400-1,600/mm^2
Power – 0.15 pJ/bit

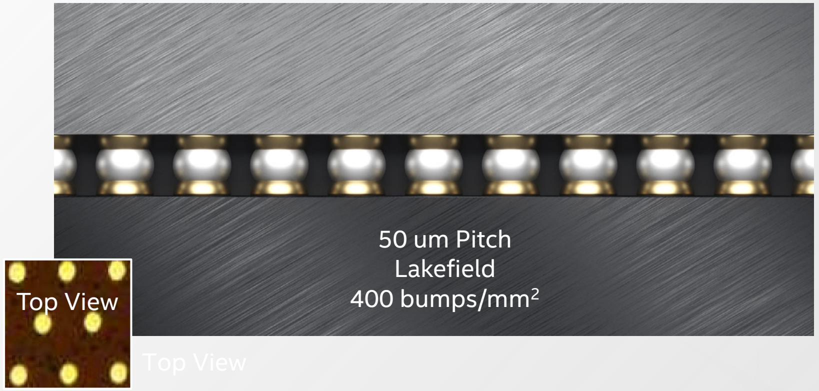
Bump pitch - < 10 microns
Bump density - > 10,000/mm^2
Power - < 0.05 pJ/bit

Hybrid Bonding

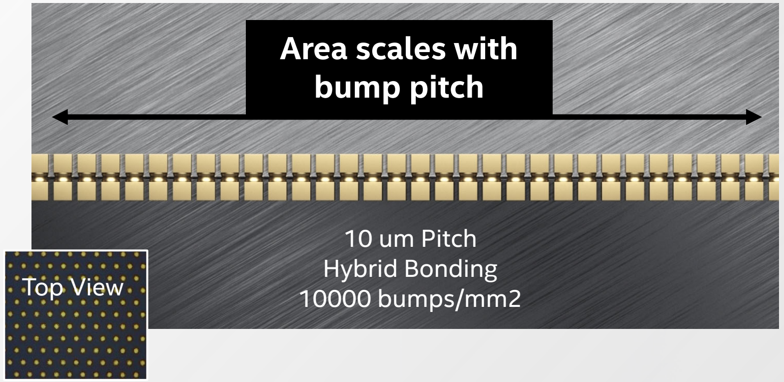
Dense vertical interconnects



FOVEROS TECHNOLOGY

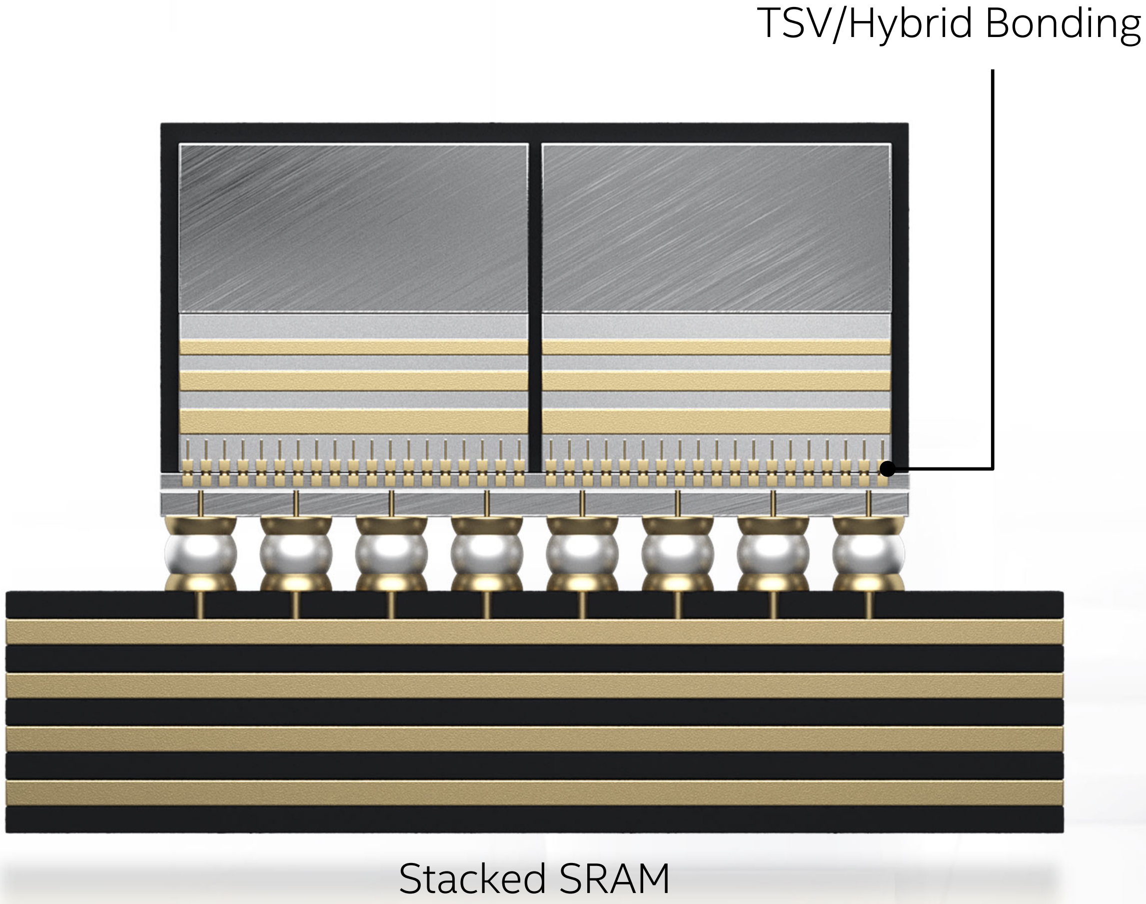


FUTURE



- Smaller, simpler circuits
- Lower capacitance
- Lower power

NEW DISCLOSURE



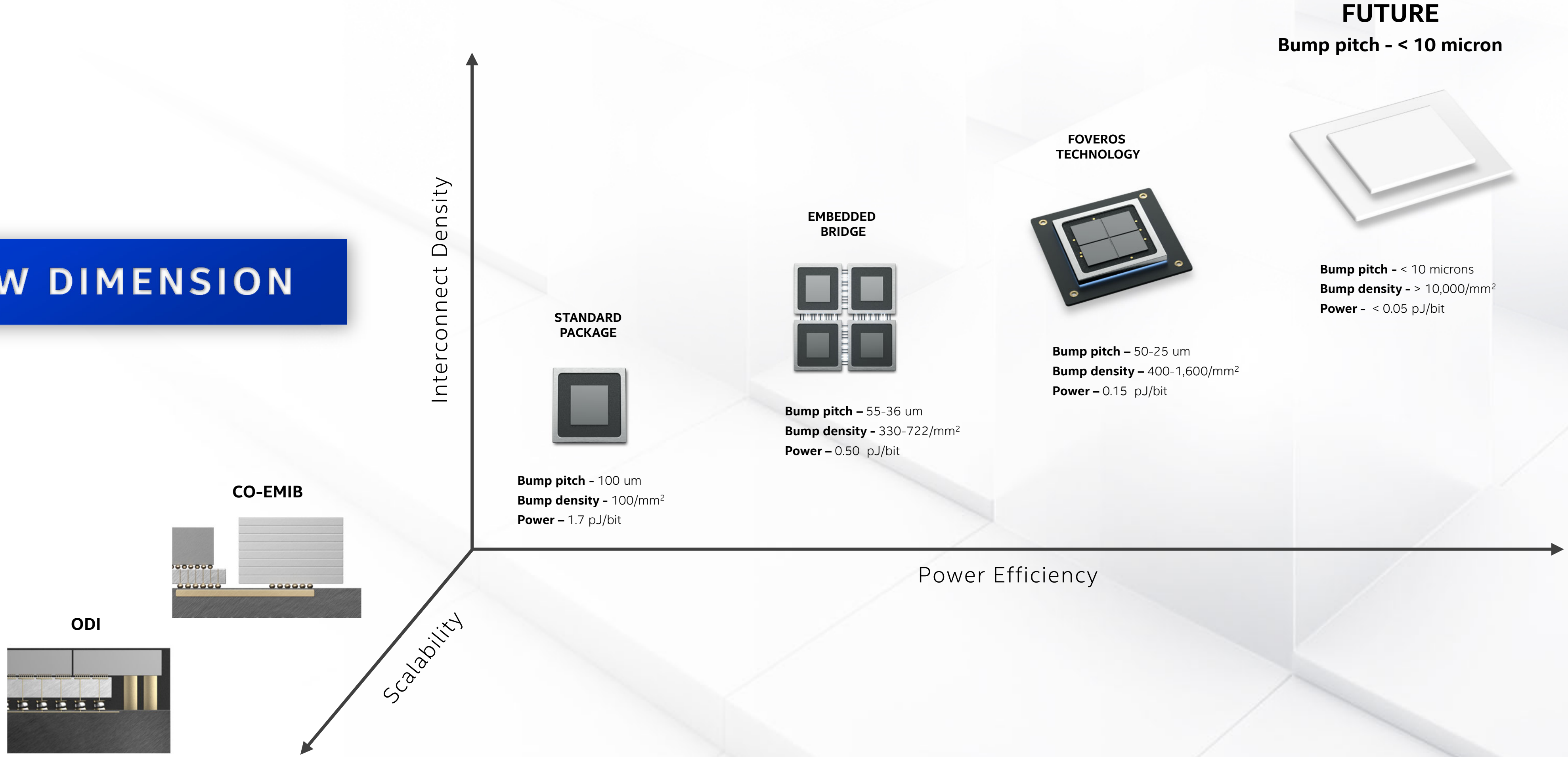
Hybrid Bonding

Enabling new architectures

TEST CHIP TAPED OUT
Q2 2020

Packaging Technology Roadmap

A NEW DIMENSION



FUTURE

Bump pitch - < 10 micron

Bump pitch - < 10 microns
Bump density - > 10,000/mm^2
Power - < 0.05 pJ/bit

FOVEROS TECHNOLOGY

Bump pitch – 50-25 um
Bump density – 400-1,600/mm^2
Power – 0.15 pJ/bit

EMBEDDED BRIDGE

Bump pitch – 55-36 um
Bump density - 330-722/mm^2
Power – 0.50 pJ/bit

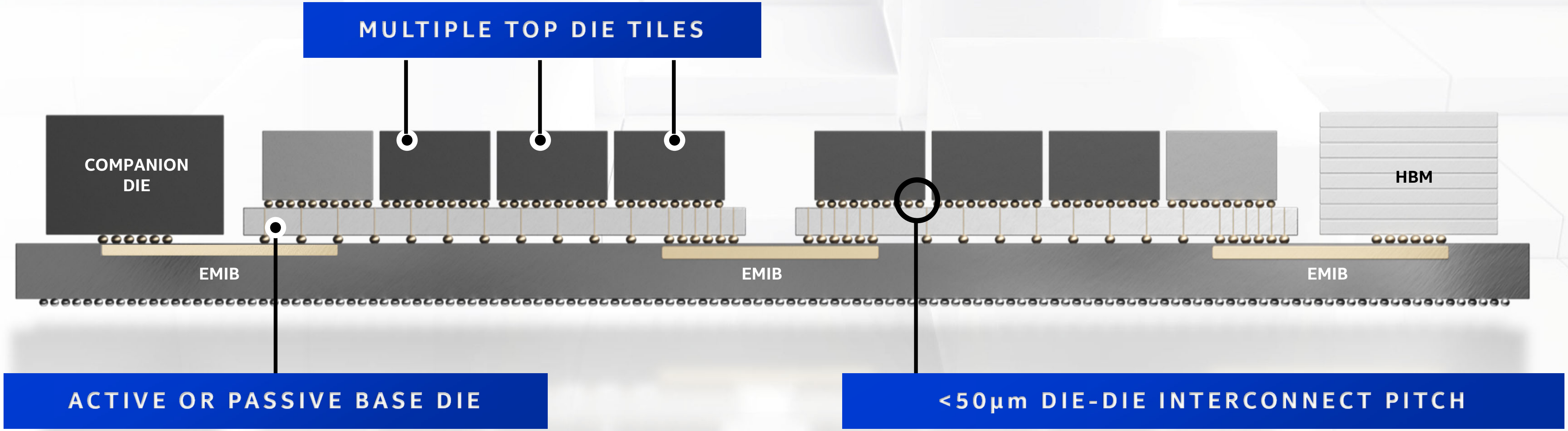
STANDARD PACKAGE

Bump pitch - 100 um
Bump density - 100/mm^2
Power – 1.7 pJ/bit

CO-EMIB

ODI

Co-EMIB - Blending 2D and 3D

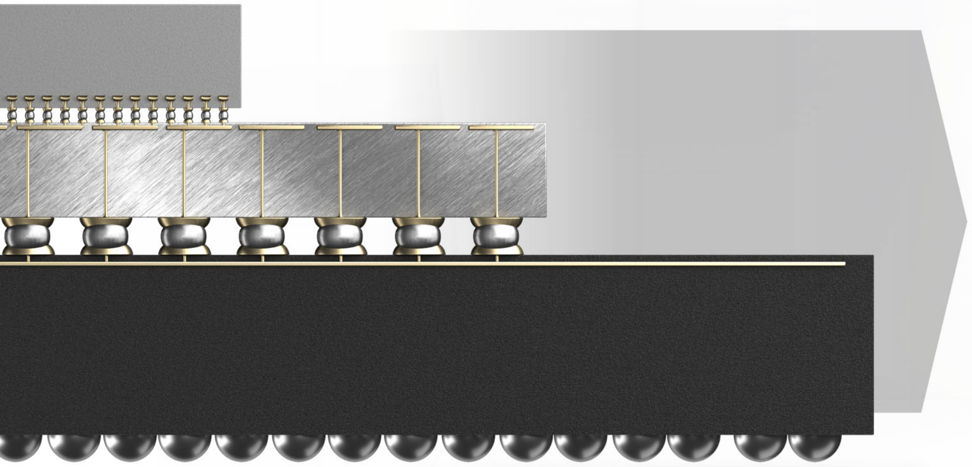


- Architecture enables large-than-reticle sized base & high density connections to companion die and stacked die complexes
- Increased partitioning opportunities

Omni-Directional Interconnect (ODI)

A new Packaging Dimension

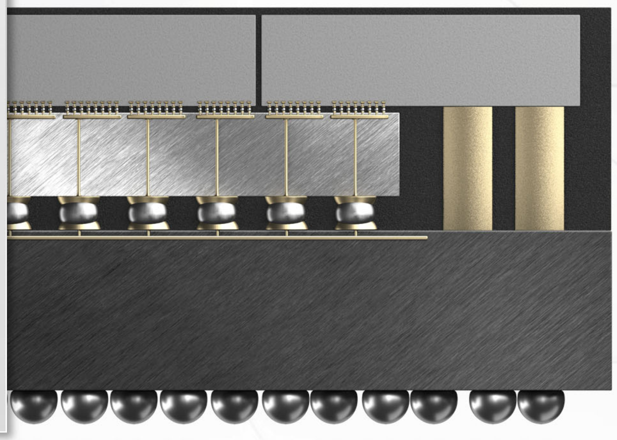
Foveros



Enables flexible design with maximum performance

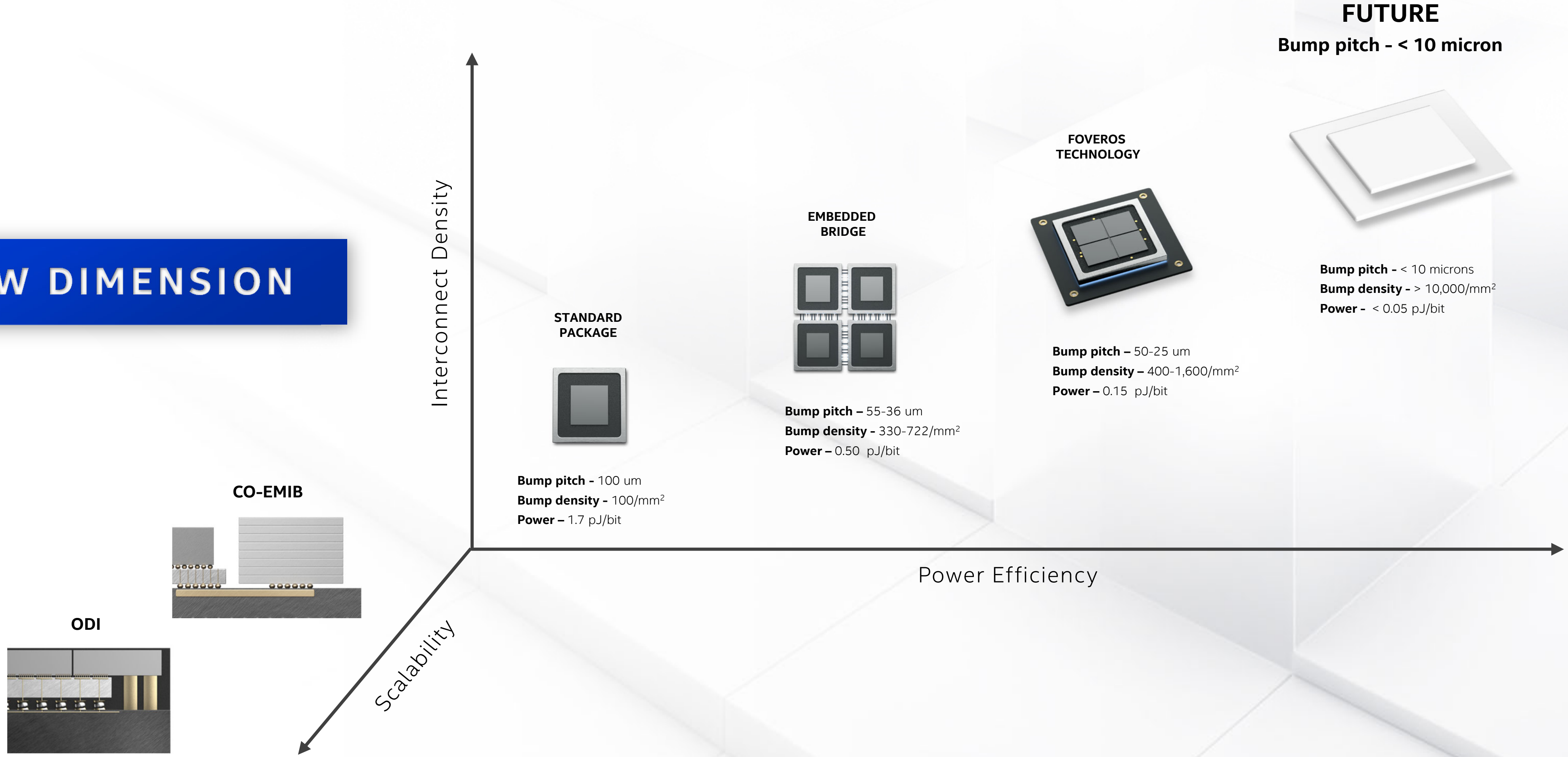
- Smaller TSV die area
- Direct power delivery
- High bandwidth interconnects

ODI



Packaging Technology Roadmap

A NEW DIMENSION



FUTURE

Bump pitch - < 10 micron

STANDARD PACKAGE

Bump pitch - 100 um
Bump density - 100/mm^2
Power – 1.7 pJ/bit

EMBEDDED BRIDGE

Bump pitch – 55-36 um
Bump density - 330-722/mm^2
Power – 0.50 pJ/bit

FOVEROS TECHNOLOGY

Bump pitch – 50-25 um
Bump density – 400-1,600/mm^2
Power – 0.15 pJ/bit

Bump pitch - < 10 microns
Bump density - > 10,000/mm^2
Power - < 0.05 pJ/bit

TECHNOLOGY PILLARS

SOFTWARE

SECURITY

INTERCONNECT

MEMORY

XPU ARCHITECTURE

PROCESS & PACKAGING

INTEL ARCHITECTURE

WHAT WE SAID IN 2018

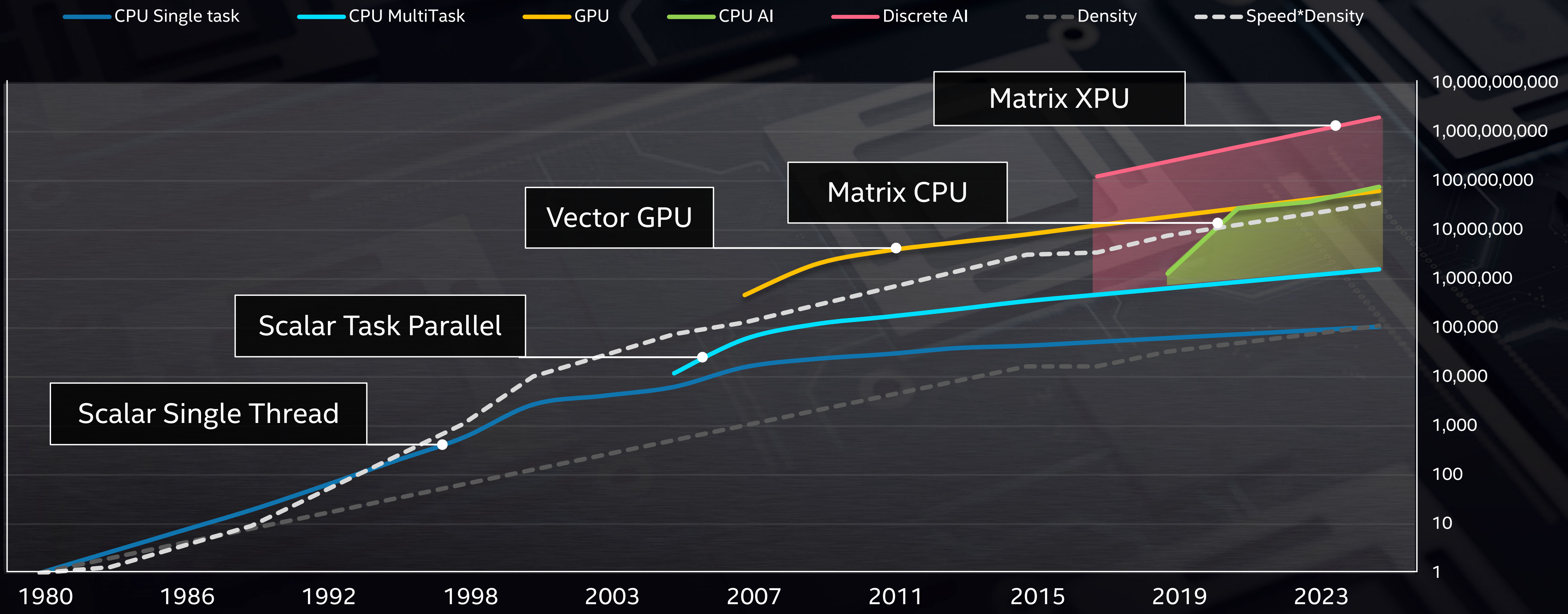
Offer a diverse mix of Scalar, Vector, Matrix and Spatial architectures deployed in CPU, GPU, FPGA and Accelerator sockets. With scalable interconnect and a single software abstraction

General Purpose Architecture Taxonomy

SCALAR	VECTOR	MATRIX	SPATIAL
CPU	GPU	AI	FPGA

MOORE'S LAW AND ARCHITECTURE

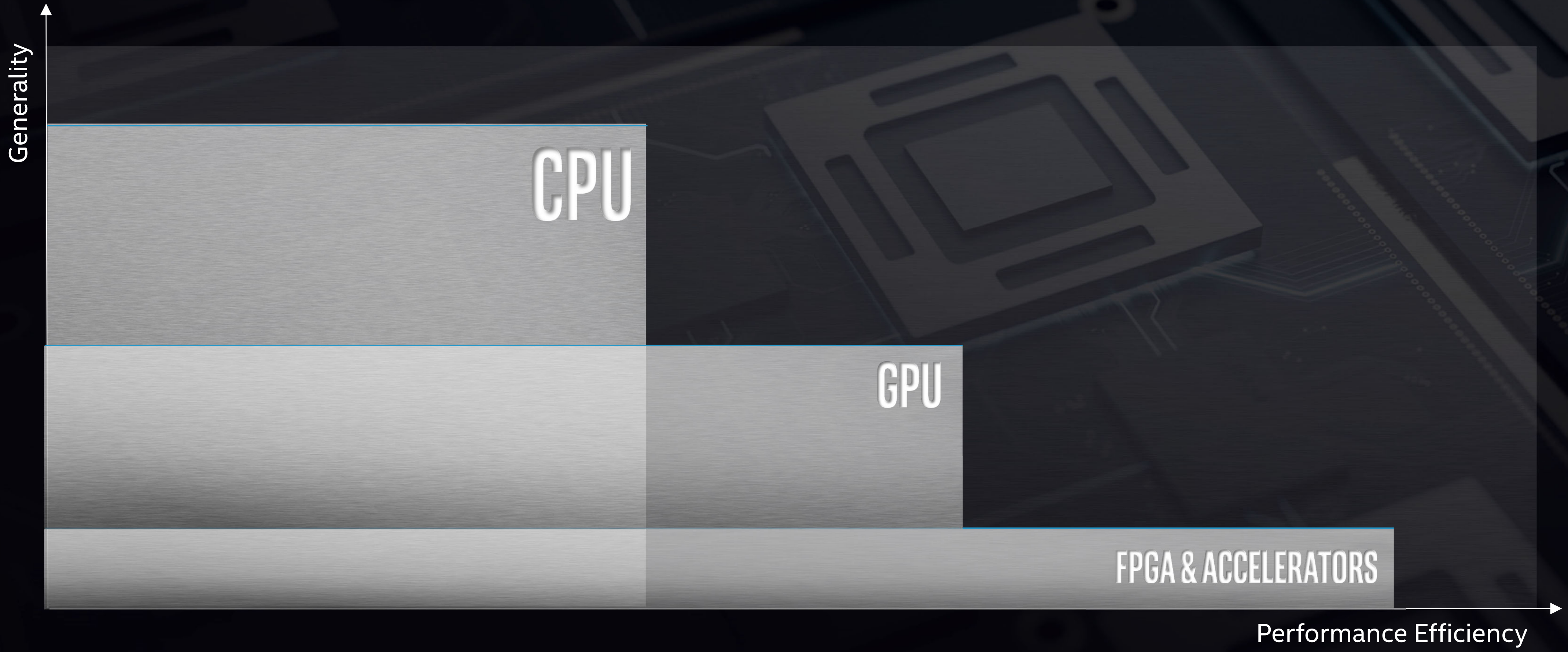
WHAT WE SAID IN 2018



ARCHITECTURE IMPACT

WHAT WE SAID IN 2018

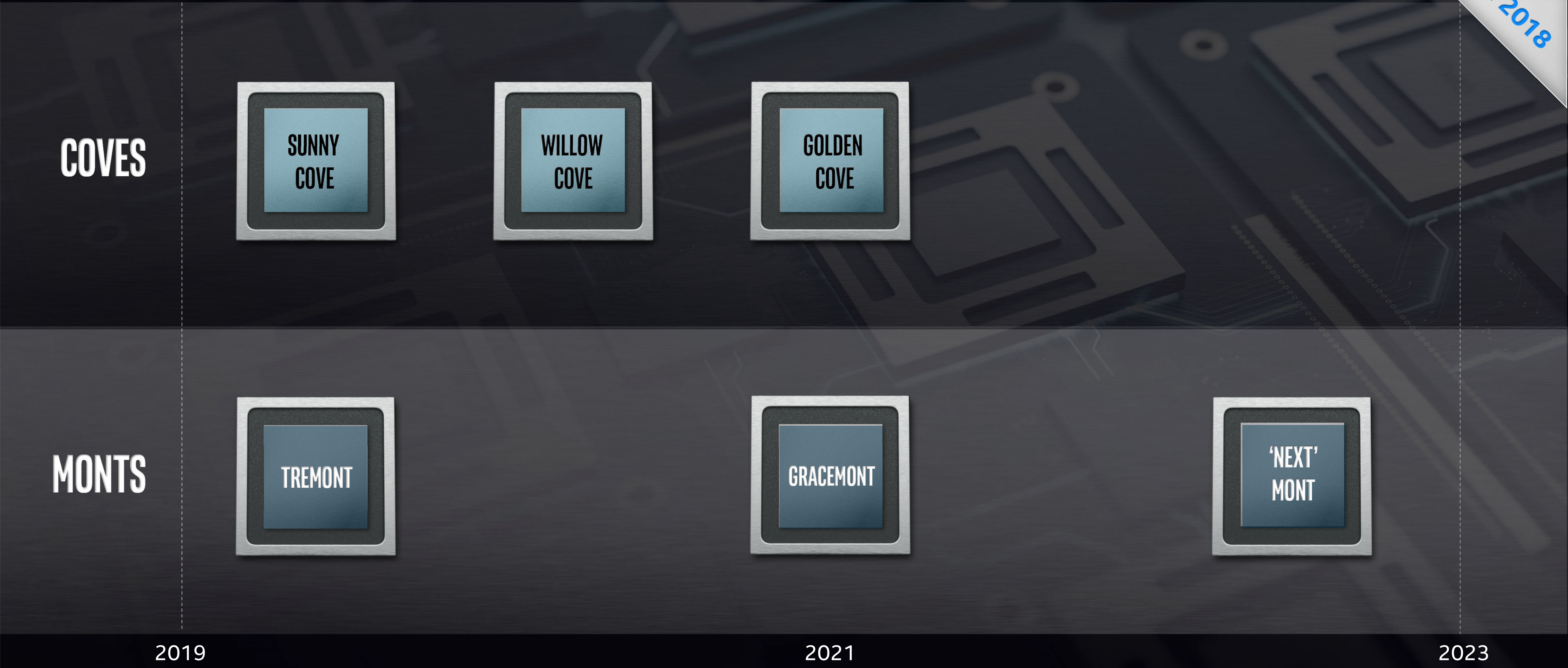
Architecture Impact = Performance x Generality



For illustrative purposes only

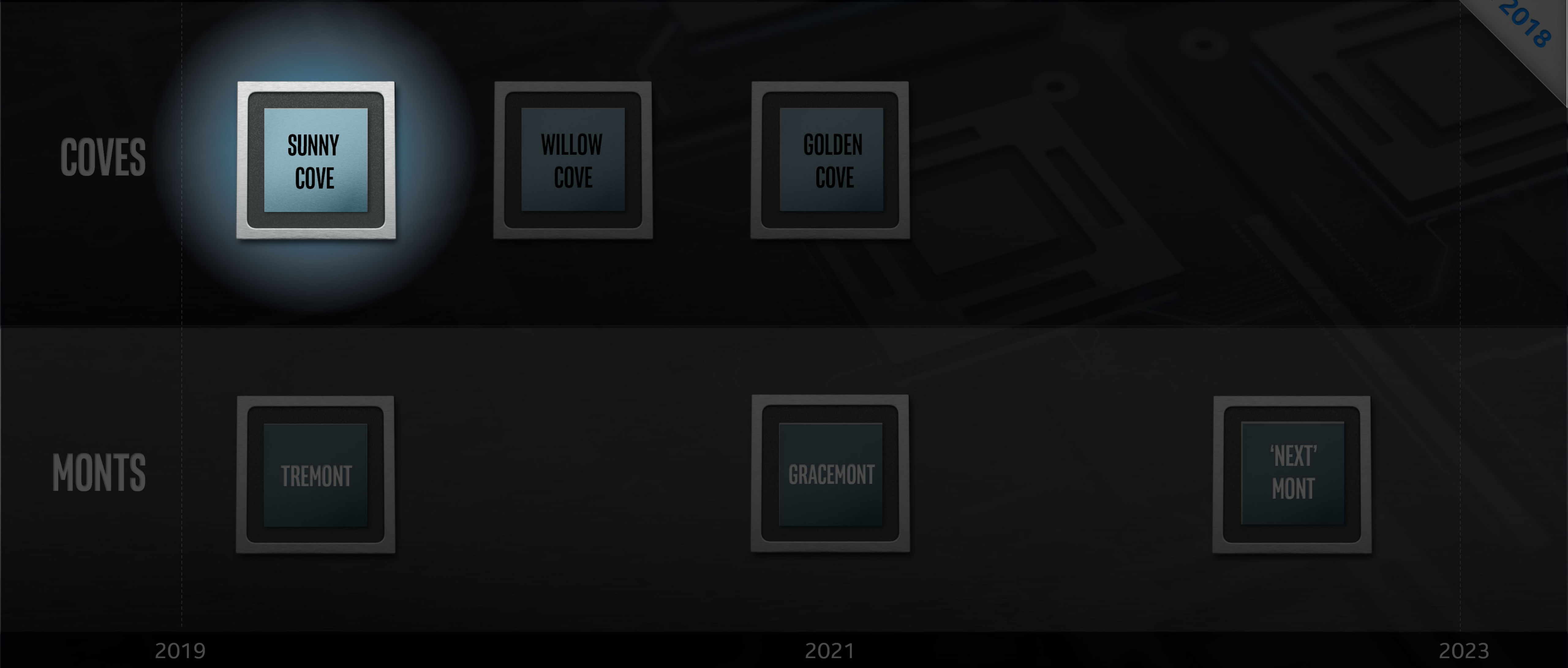
CPU CORE ROADMAP

WHAT WE SAID IN 2018



CPU CORE ROADMAP

WHAT WE SAID IN 2018



2018 ARCHITECTURE DAY

CPU CORE ROADMAP

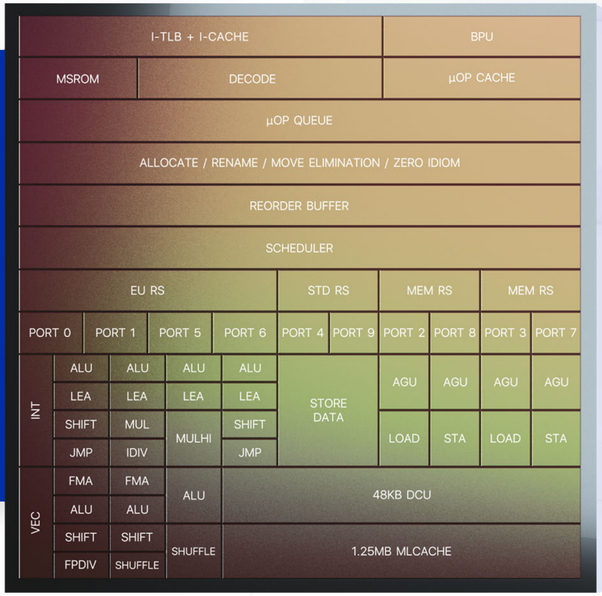
WHAT WE SAID IN 2018

Tiger Lake Architecture

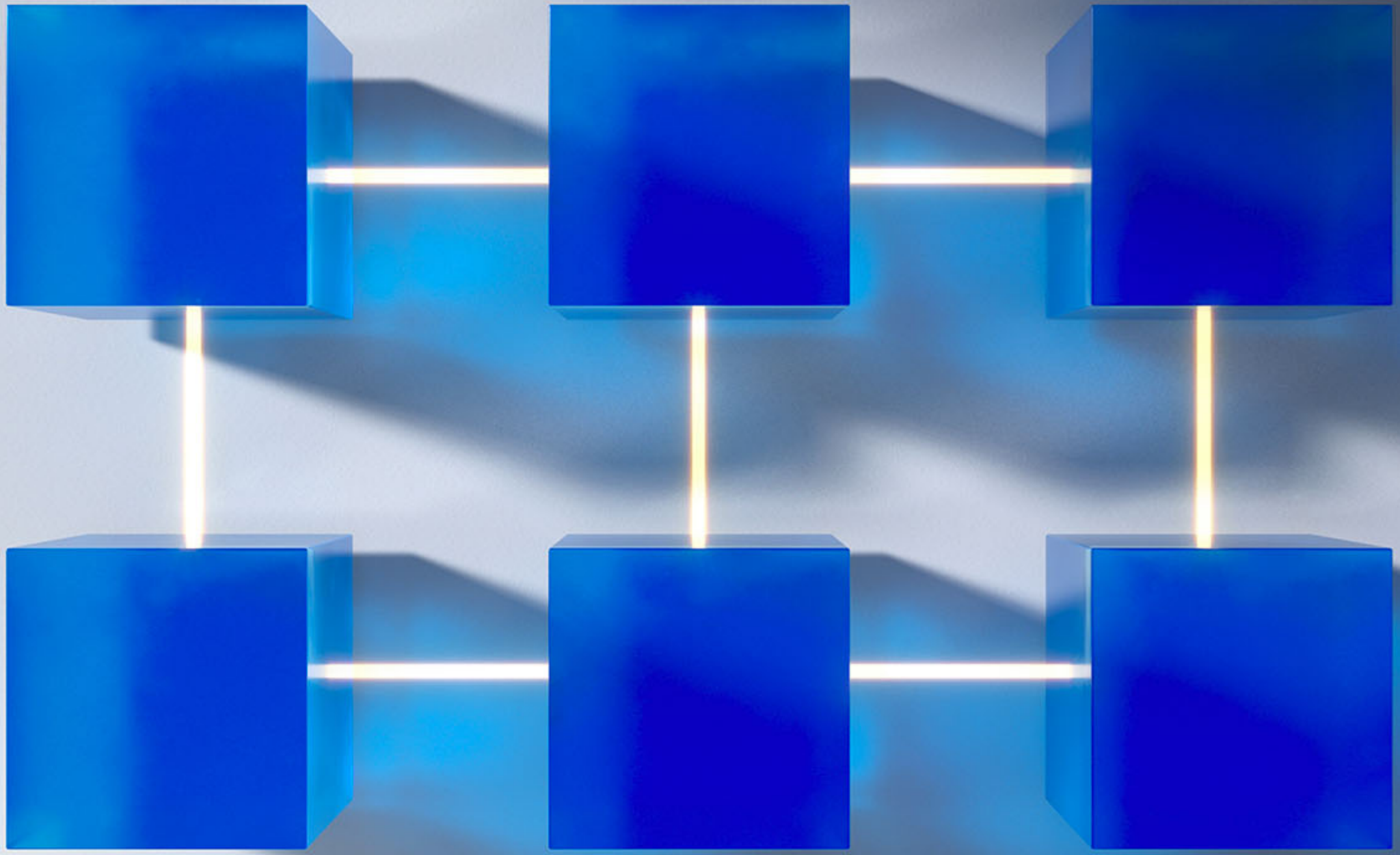
TIGER LAKE SOC

Tiger Lake Architecture

WILLOW COVE SCALAR ARCHITECTURE



TIGER LAKE SOC



Tiger Lake SOC Architecture

Boyd Phelps

CVP & GM Devices
Development Group

TECHNOLOGY
PILLARS

Architecture Day **2020**

TIGER LAKE ARCHITECTURE

Goals

- CPU Greater than generational performance in CPU
- GPU Disruptive performance in integrated graphics
- AI Scalable AI for emerging client workloads
- Increased memory & fabric efficiency for high bandwidth
- Best in class set of IPs throughout the SoC
- Continue to advance Intel's Security

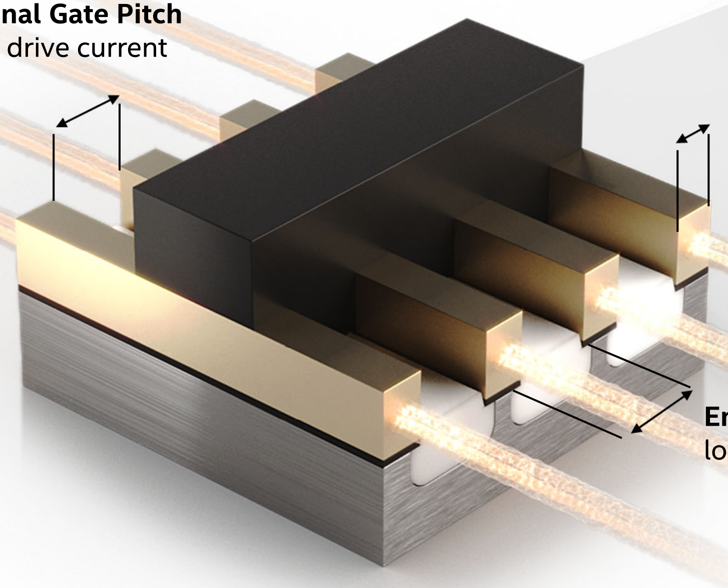
.... all at the same power envelopes with increased power efficiency

NEW DISCLOSURE

New High-Performance Transistor

Innovation across the entire process stack, from channel to interconnects

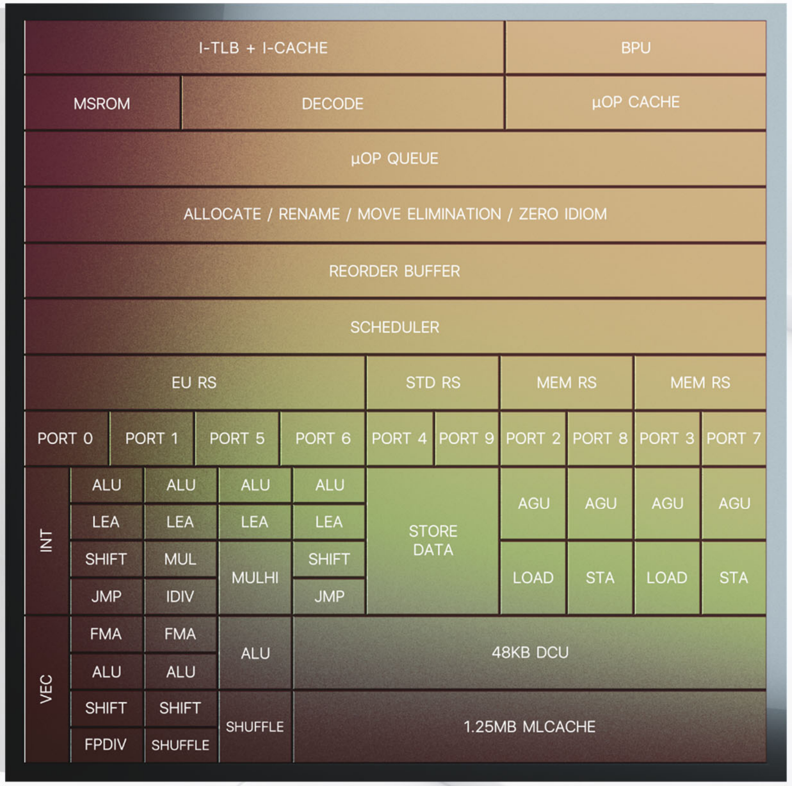
New SuperFin Tiger Lake Transistors



Additional Gate Pitch
higher drive current

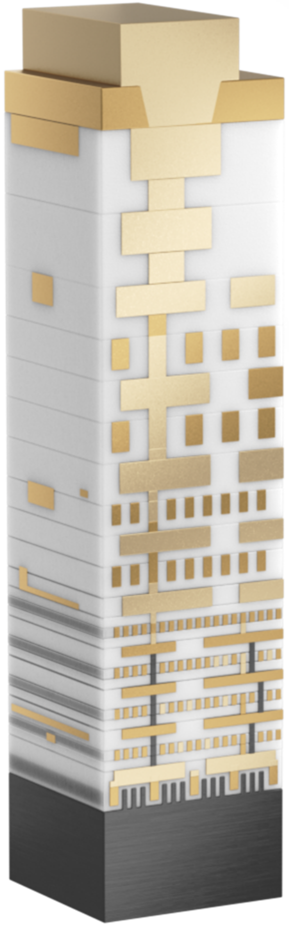
Improved Gate Process
higher channel mobility

Enhanced Epitaxial Source/Drain
lower resistance, increases strain



Improved Metal Stack

Innovation across the entire process stack, from channel to interconnects



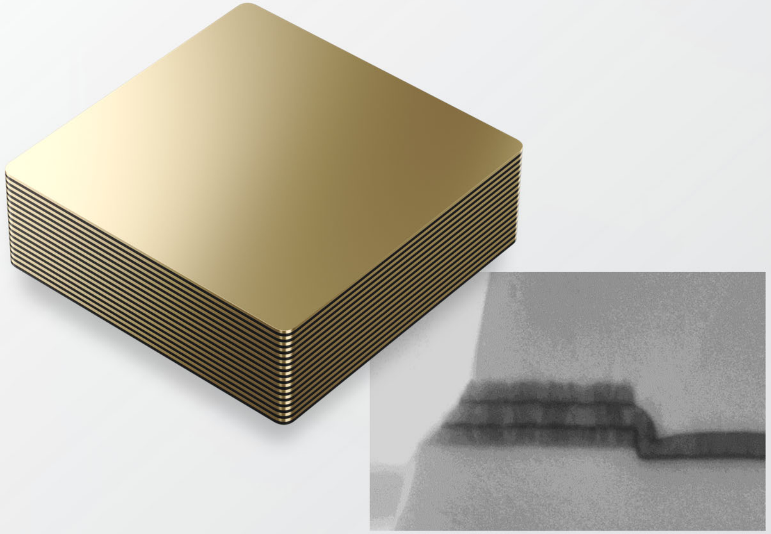
Super MIM Capacitor

5x increase in MIM capacitance

Novel Thin Barrier

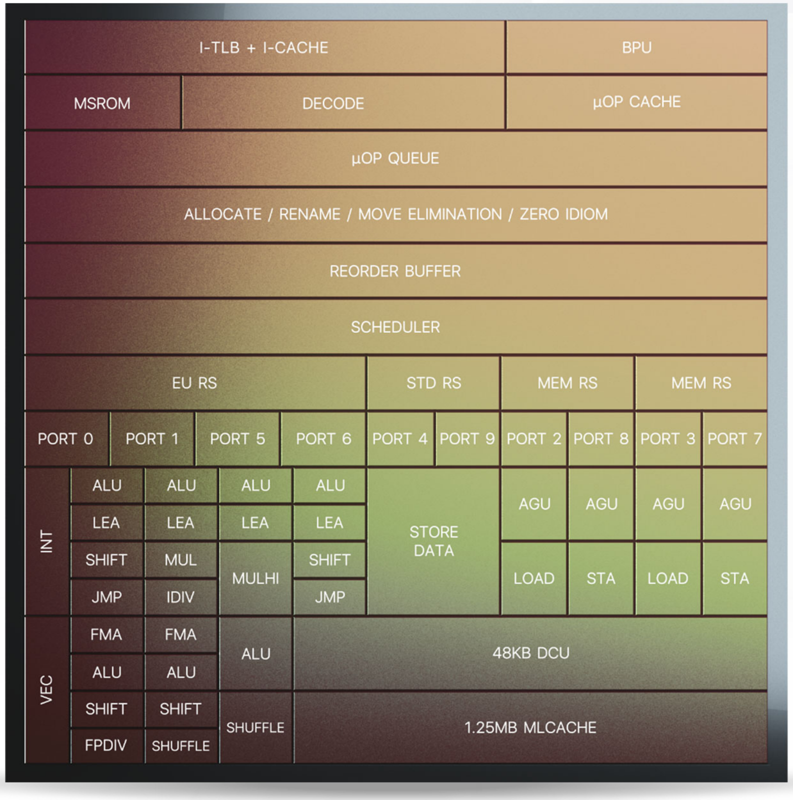
reduces via resistance by 30%

Thin layers of different Hi-K materials, each just a few Angstroms thick, stacked in a repeating "superlattice."



NEW DISCLOSURE

Willow Cove Core



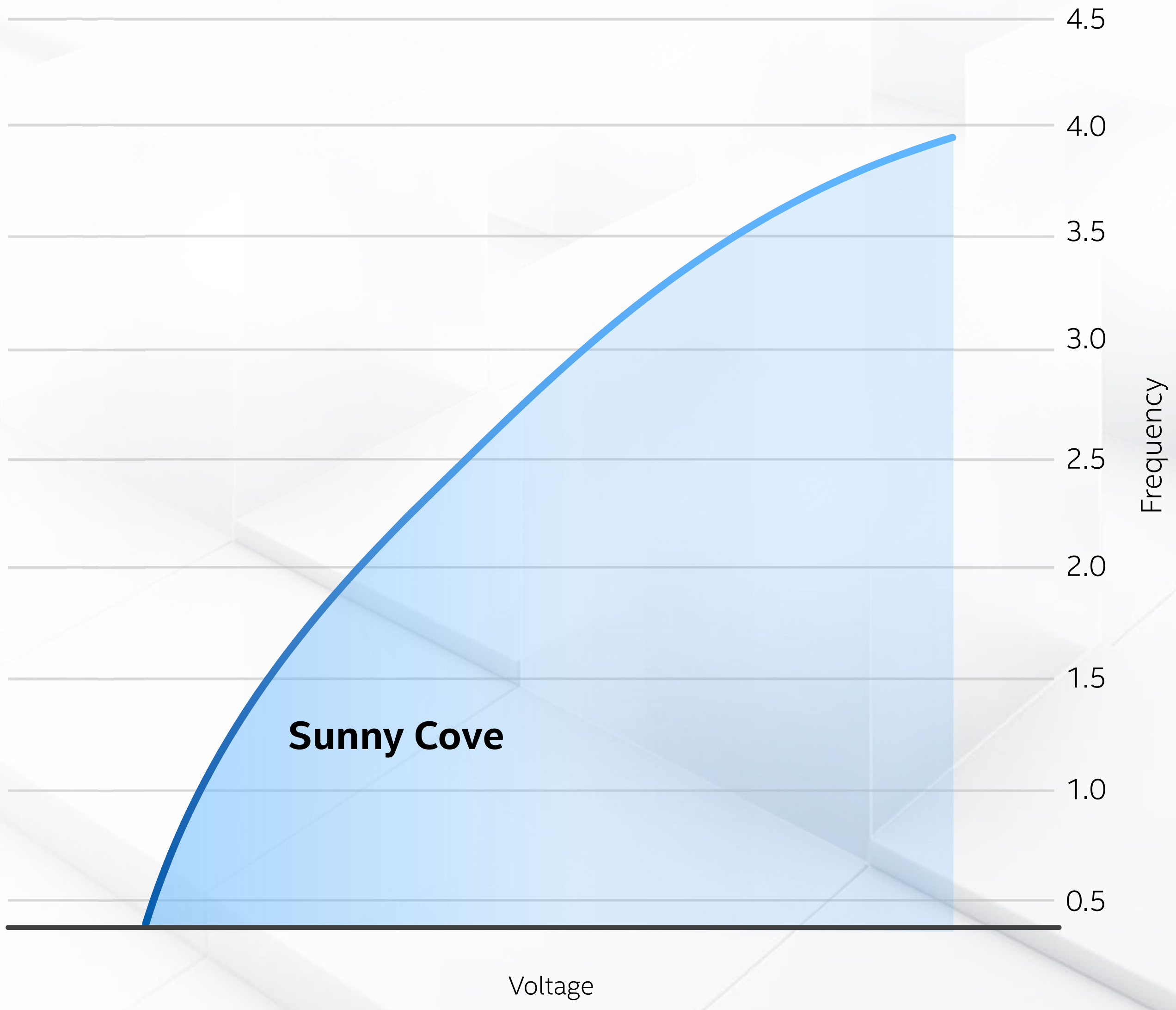
CPU Core Goals

- Build upon the Sunny Cove architectural foundation
- Redesigned caching architecture to larger non-inclusive, 1.25MB MLC
- Control Flow Enforcement technology to help protect against return / jump oriented attacks

All at a dramatic frequency increase over our prior generation!

NEW DISCLOSURE

The Result



The Result

Large Frequency Gains

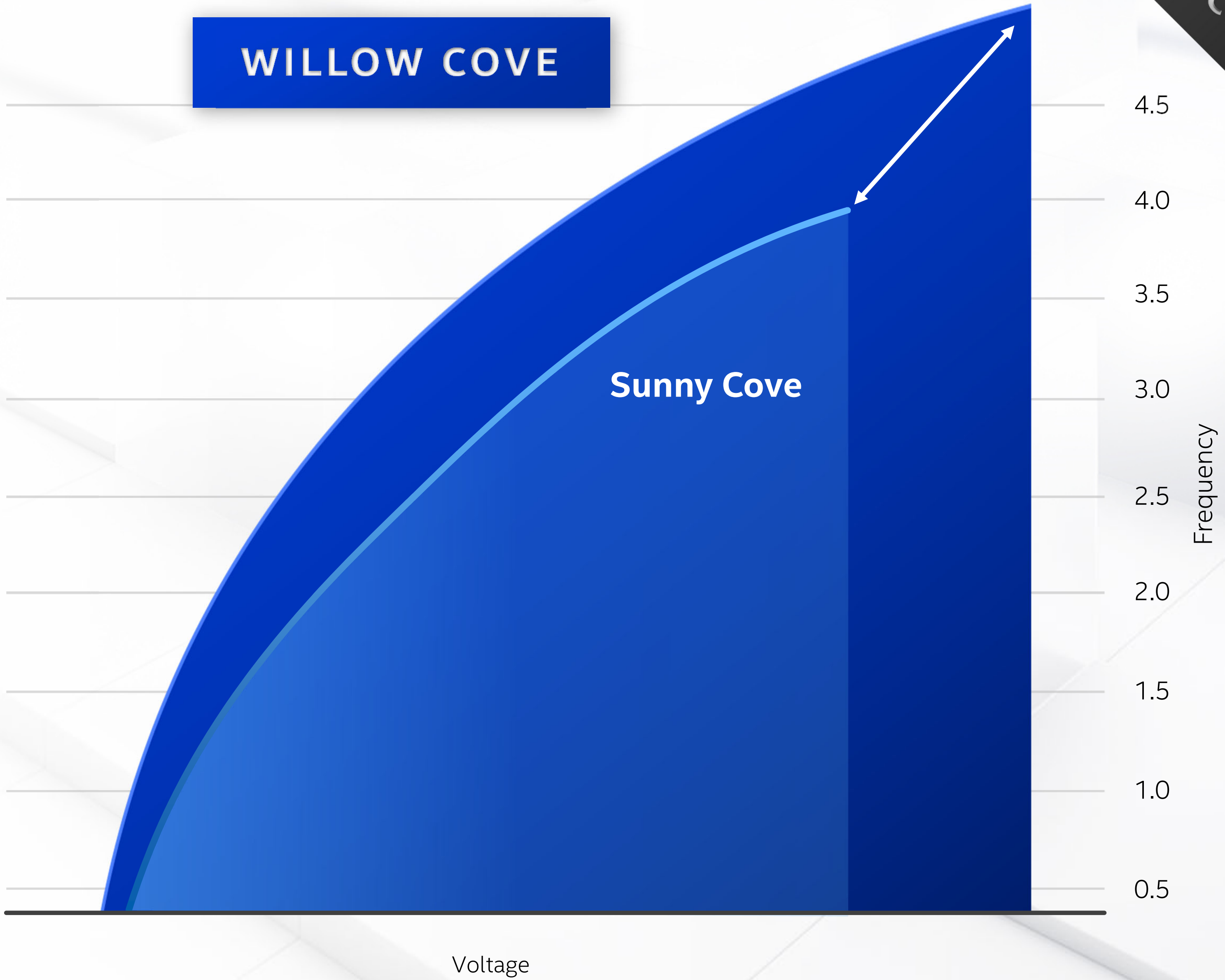
NEW DISCLOSURE

The Result

Increased Power Efficiency

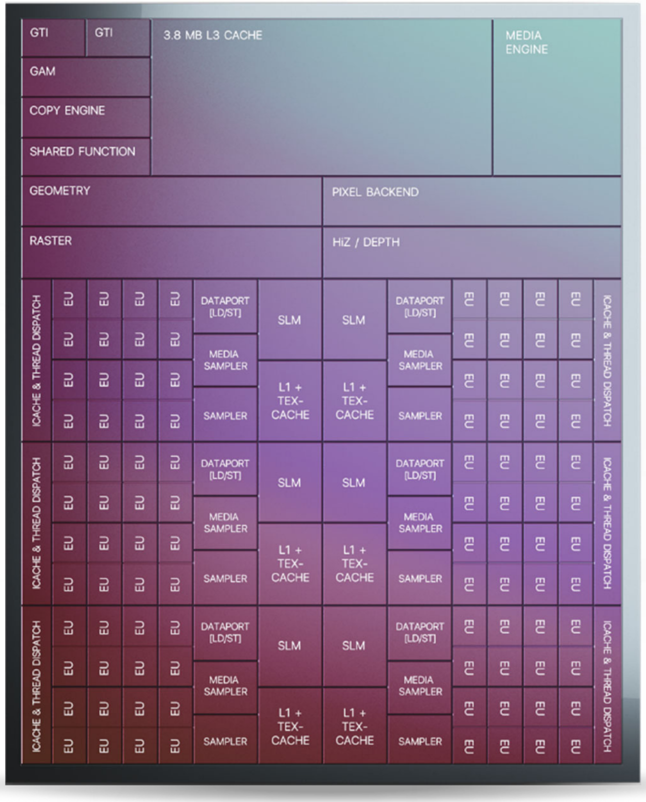
The Result

Greater Dynamic Range



X^e_{LP} Graphics

NEW DISCLOSURE



Large improvements in performance per watt efficiency

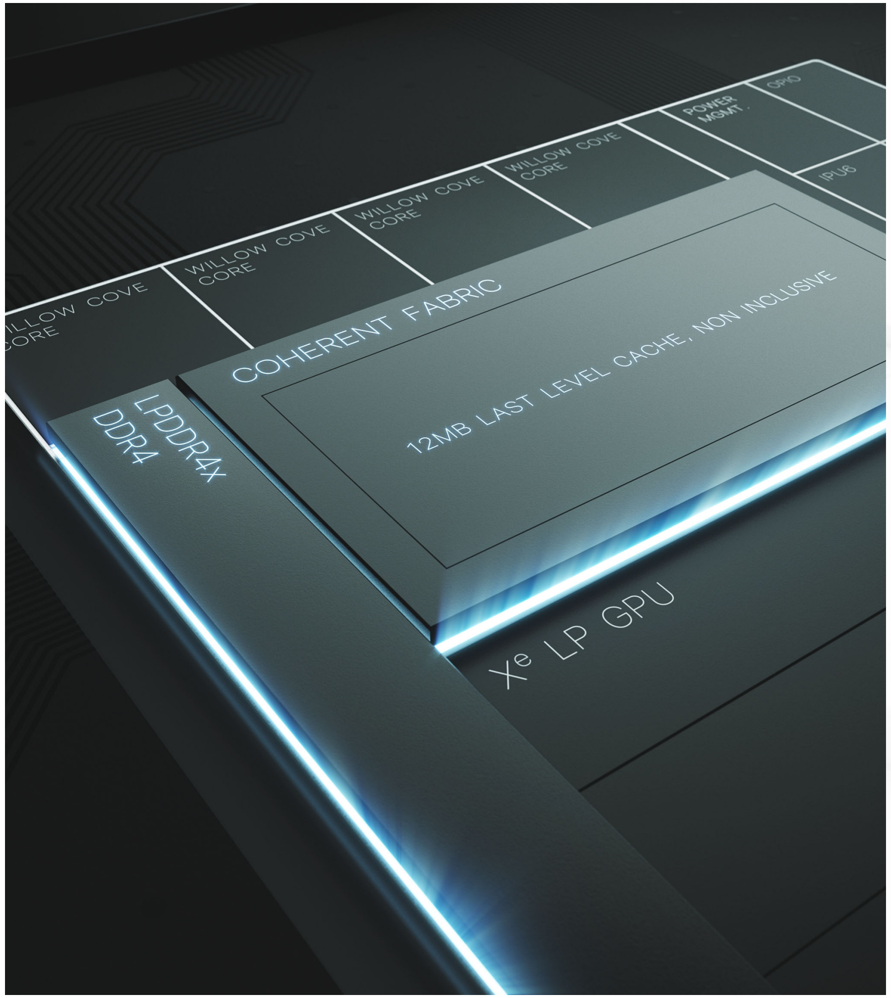
Up to 96EUs with increased capabilities

3.8MB L3 cache

Increased Memory & Fabric efficiency for high bandwidth

NEW DISCLOSURE

Fabrics and Memory



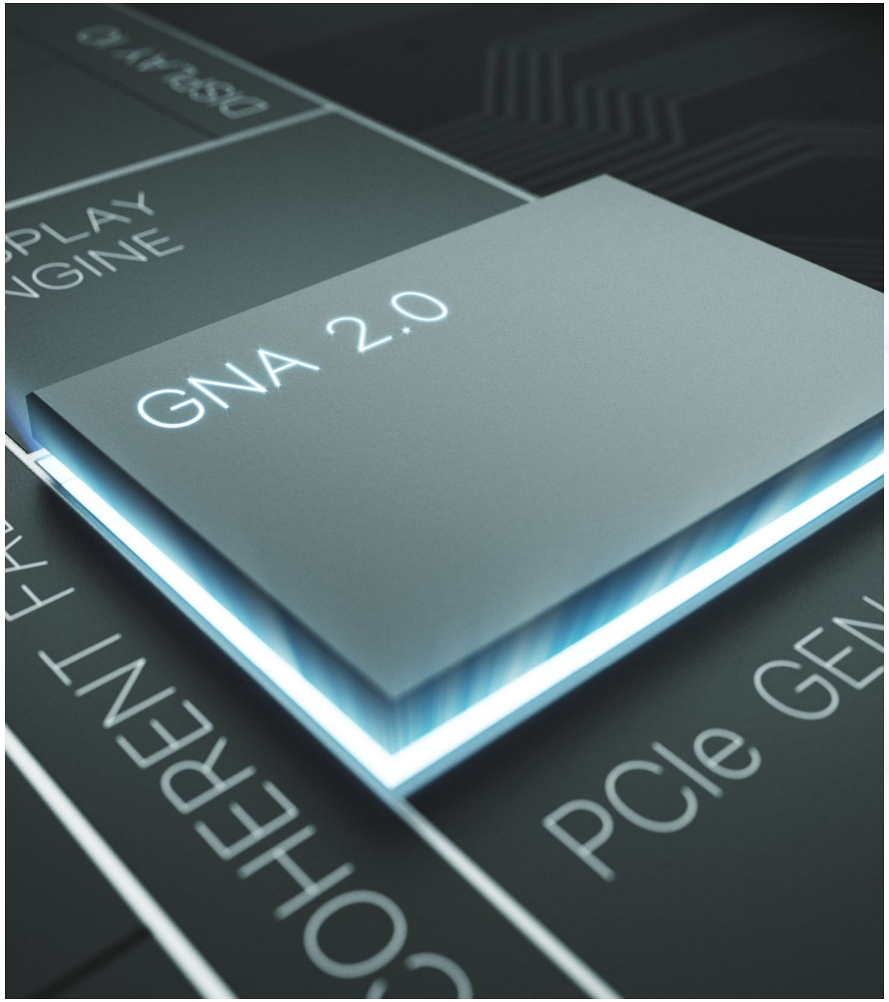
Coherent fabric and last level cache

- >2x increase in coherent fabric bandwidth
  - Dual ring microarchitecture
  - 50% LLC size increase to non-inclusive cache
  - Maintaining low hit latency

Memory

- Increases in available bandwidth throughout memory subsystem
  - Support for up to ~86GB/s of memory bandwidth
  - Dual memory controller subsystem for efficiency improvements
- Architectural support for LP4x-4267 and DDR4-3200 (initial) and up to LP5-5400
- Intel® Total Memory Encryption to protect against hardware attacks

Intel® Gaussian and Neural Accelerator (GNA 2.0)



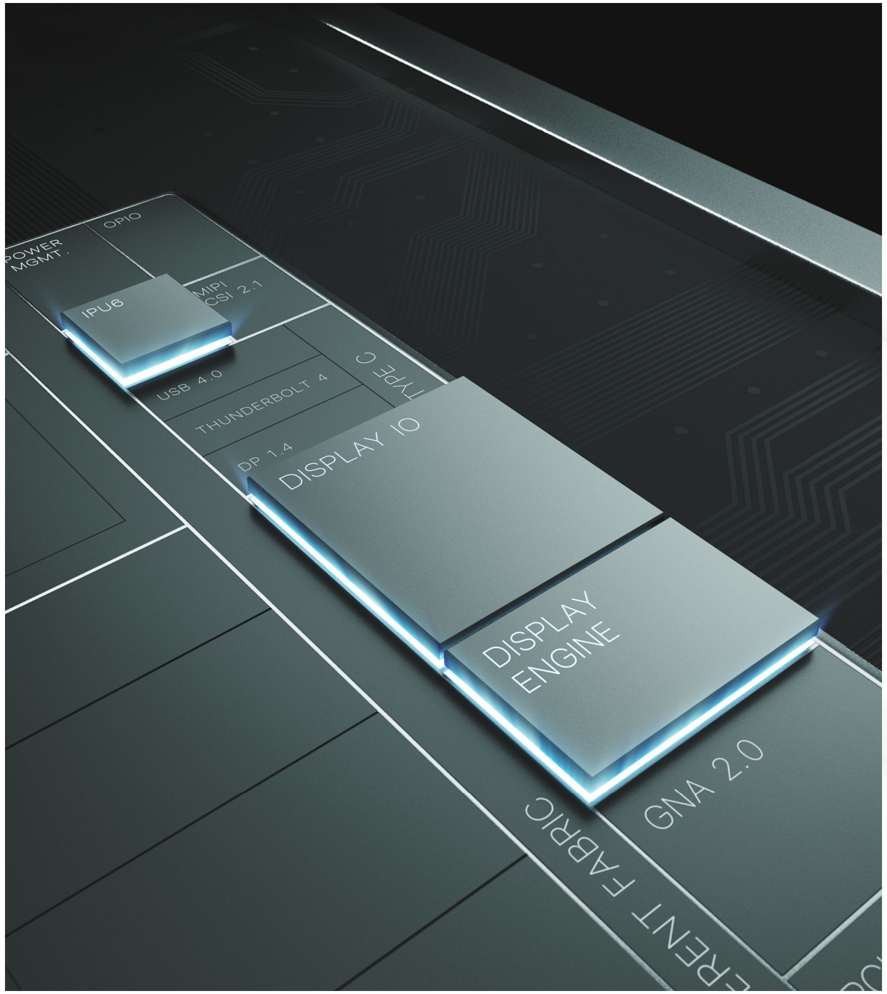
Dedicated IP for low power neural inferencing

Usages: Neural noise cancellation for high dynamic range noise

~20% lower CPU utilization on GNA

NEW DISCLOSURE

Display and IPU6



Display

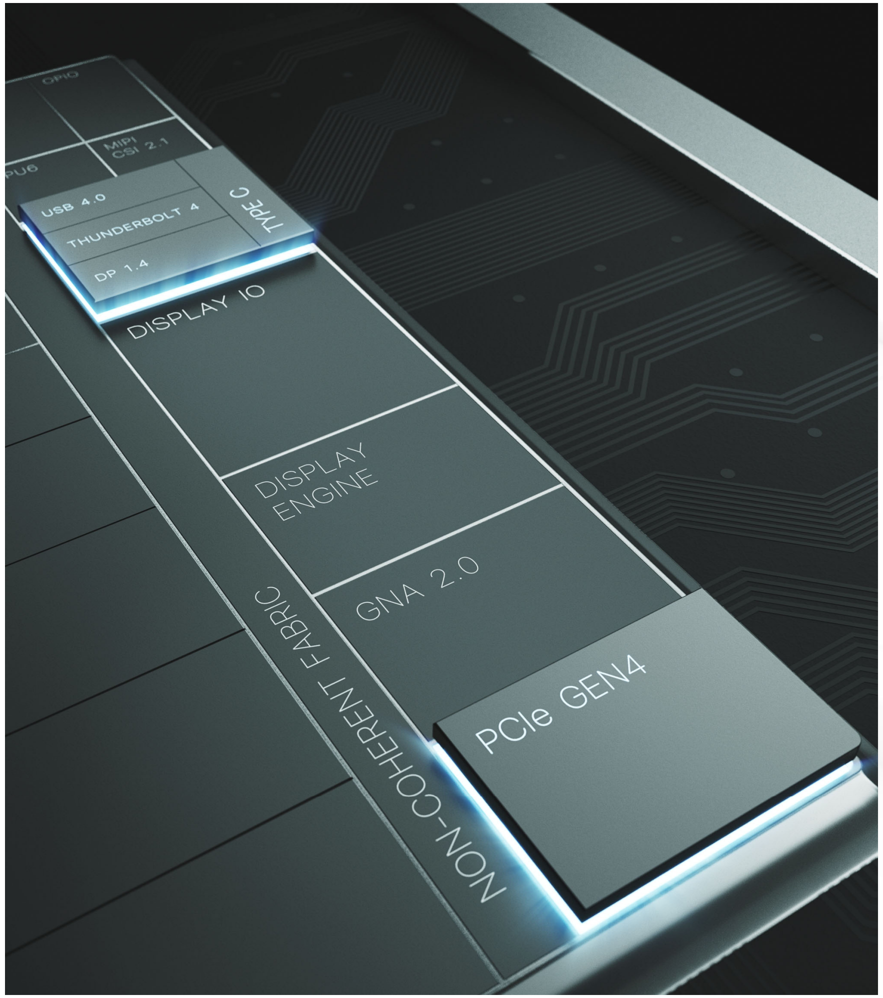
- **Goal –** More displays at higher resolution and quality
- Dedicated fabric path to memory to maintain quality of service
- Up to 64GB/s of isochronous bandwidth to memory

IPU6

- Imaging pipeline fully implemented in hardware
- Up to 6 sensors with architectural capabilities for:
  - Video up to 4K90 resolutions (initial 4K30)
  - Still image up to 42 megapixels (initial 27MP)

NEW DISCLOSURE

IO



Integrated Thunderbolt 4 and USB4 support

- Up to 40Gb/s bandwidth on each port

Integrated Display output via Type-C Subsystem

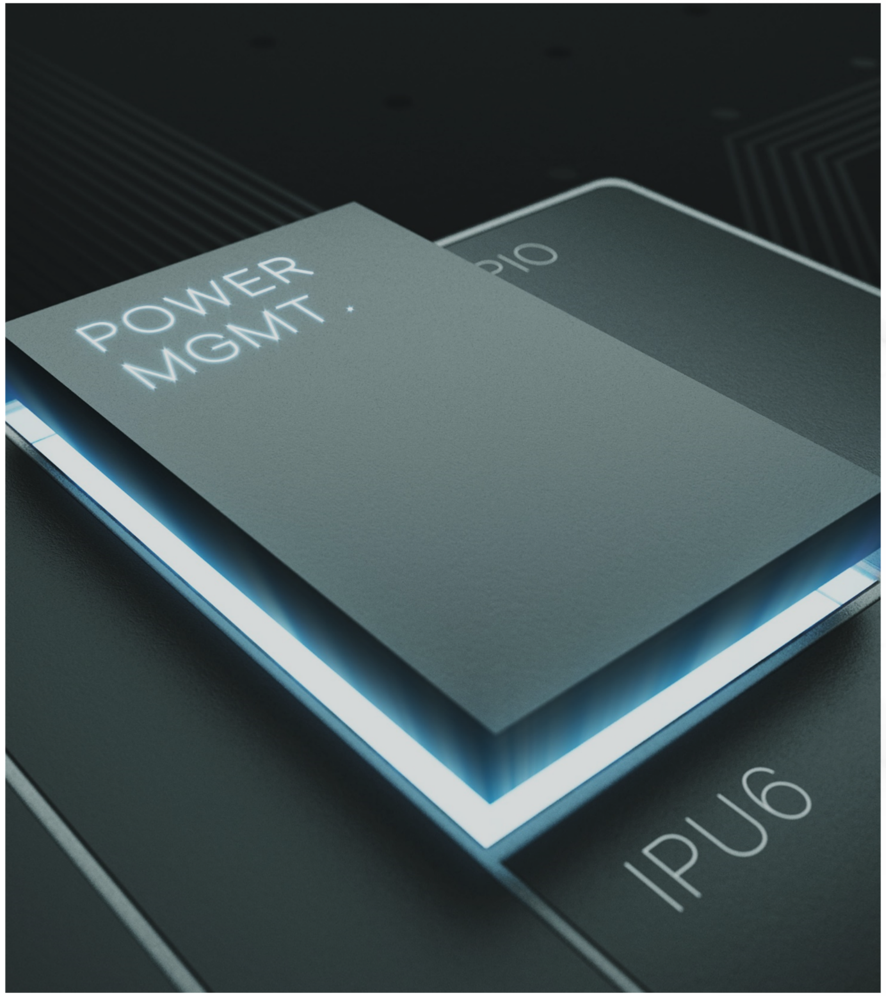
- DP alternate Mode
- DP tunneling over Thunderbolt
- DP-in ports for discrete graphics card display output to mux over type-C port

PCIe Gen4 on CPU for low latency, high bandwidth device access to memory

- Full 8GB/s bandwidth to memory
- ~100ns less latency when attached to CPU vs PCH

NEW DISCLOSURE

Power Management



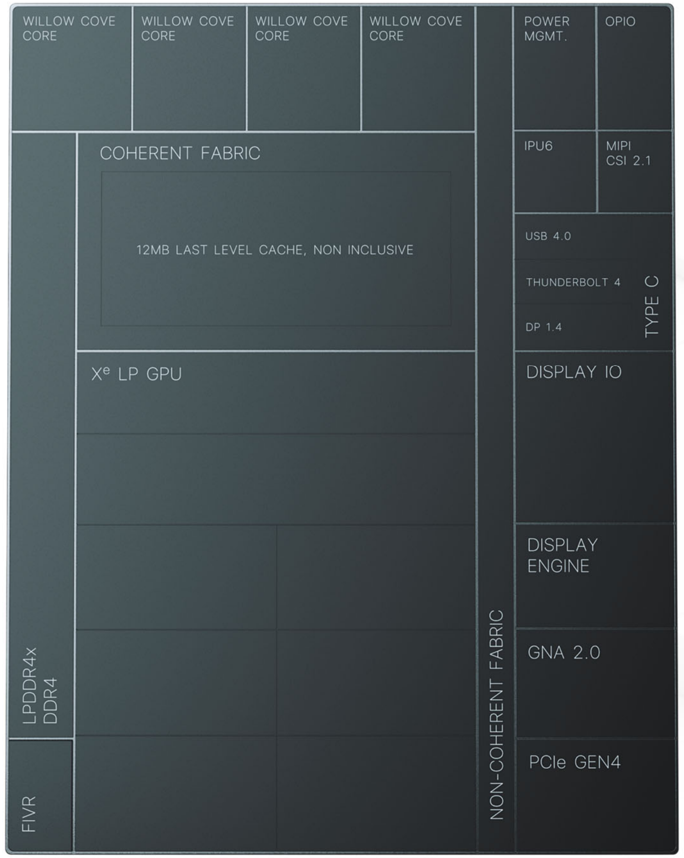
Autonomous DVFS in coherent fabric and memory subsystem to scale frequency and voltage based on bandwidth

Targeted power optimizations

- Deeper package C state turning off all clocks in CPU
- Increased FIVR efficiency at low loads
- Moved always-on logic in fabric, PCIe, Type-C and thermal sensors blocks to gated domains
- Hardware-based save and restore logic

NEW DISCLOSURE

Tiger Lake SoC Architecture

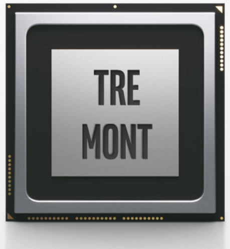
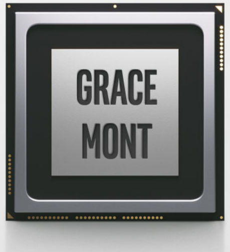


Versus Ice Lake architecture

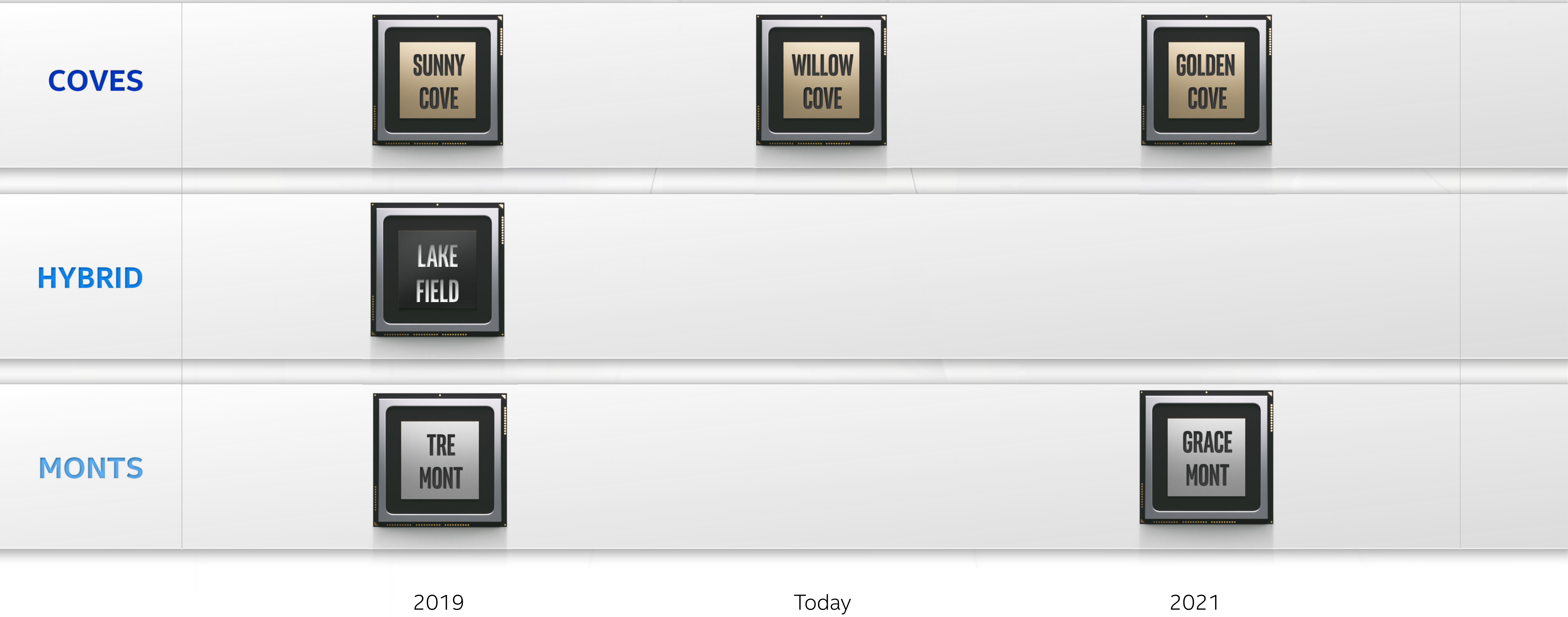
Leveraging Process Tech Improvements

- Tiger Lake SoC Architecture delivers significant advancements across a wide set of SoC IPs, with:
- More than a generational increase in CPU performance in Willow Cove CPU core
- Massive improvements in graphics power efficiency in X^e graphics
- Scalable AI for emerging client workloads
- Increased Memory & Fabric efficiency for high bandwidth
- Rich I/O … and much, much more!

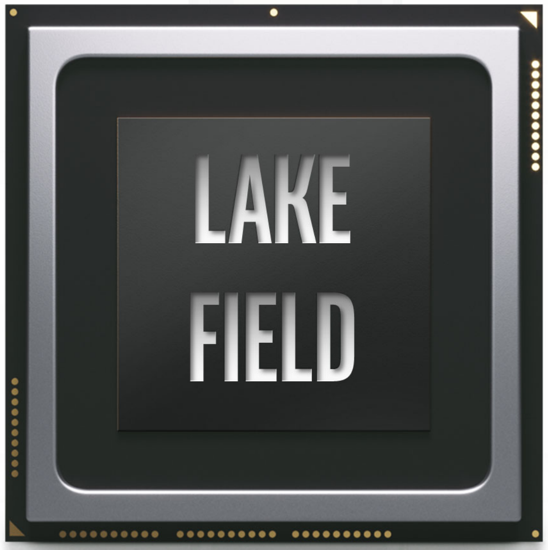
CPU Core Roadmap

	2019	Today	2021
COVES	 • ST perf • New ISA • Scalability Improved	 • Cache redesign, • New transistor optimization • Security Features	 • ST perf • AI Perf • Network/5G Perf • Security Features
MONTS	 • ST perf, • Network server perf, • Battery life perf		 • ST perf • Vector Perf

CPU Core Roadmap



	2019	Today	2021
COVES	SUNNY COVE	WILLOW COVE	GOLDEN COVE
HYBRID	LAKE FIELD		
MONTS	TRE MONT		GRACE MONT



First Hybrid x86 Architecture

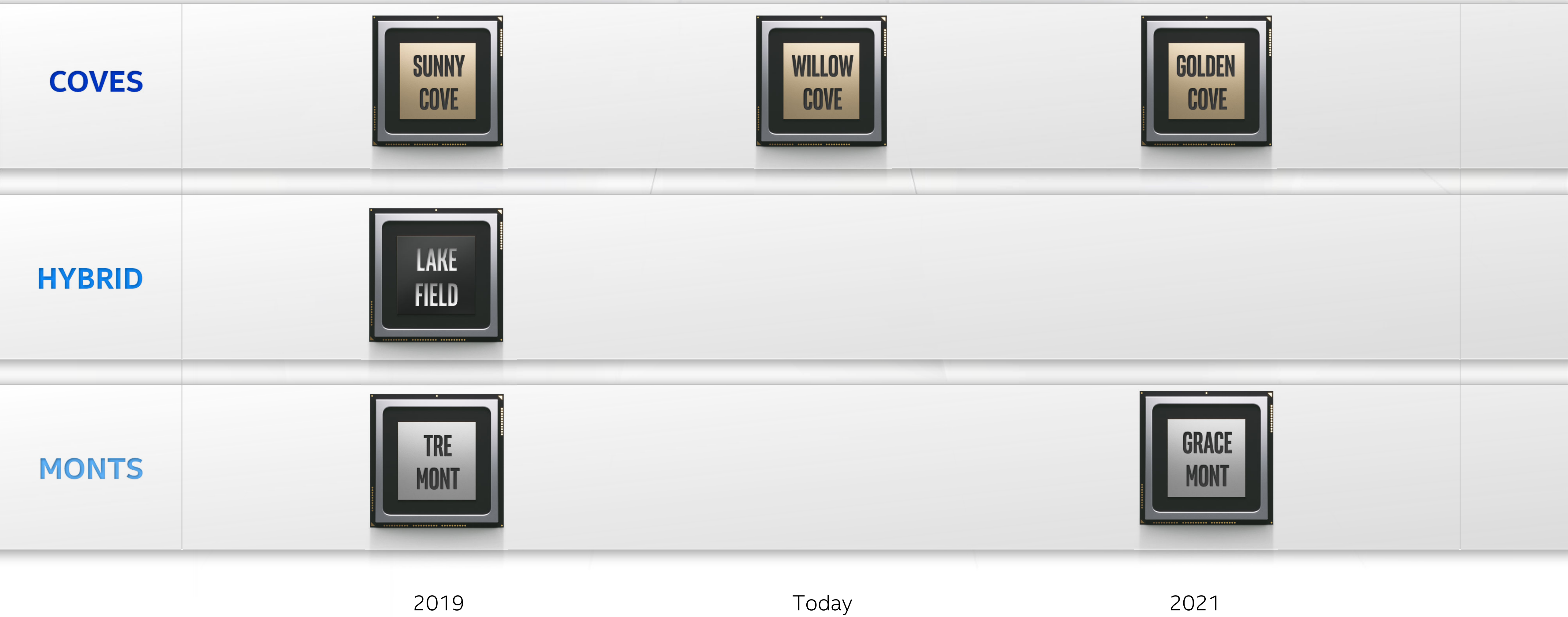
Performance improvements through Hardware and **OS optimization**

Up to **24%** Higher Power Efficiency[1]	Up to **33%** Higher Web Performance[2]
Up to **91%** Lower Standby Power[1]	Up to **17%** Higher Power Efficiency[2]

1. Compared to Previous Gen
2. Hybrid vs Non-Hybrid (4 Tremont cores)

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

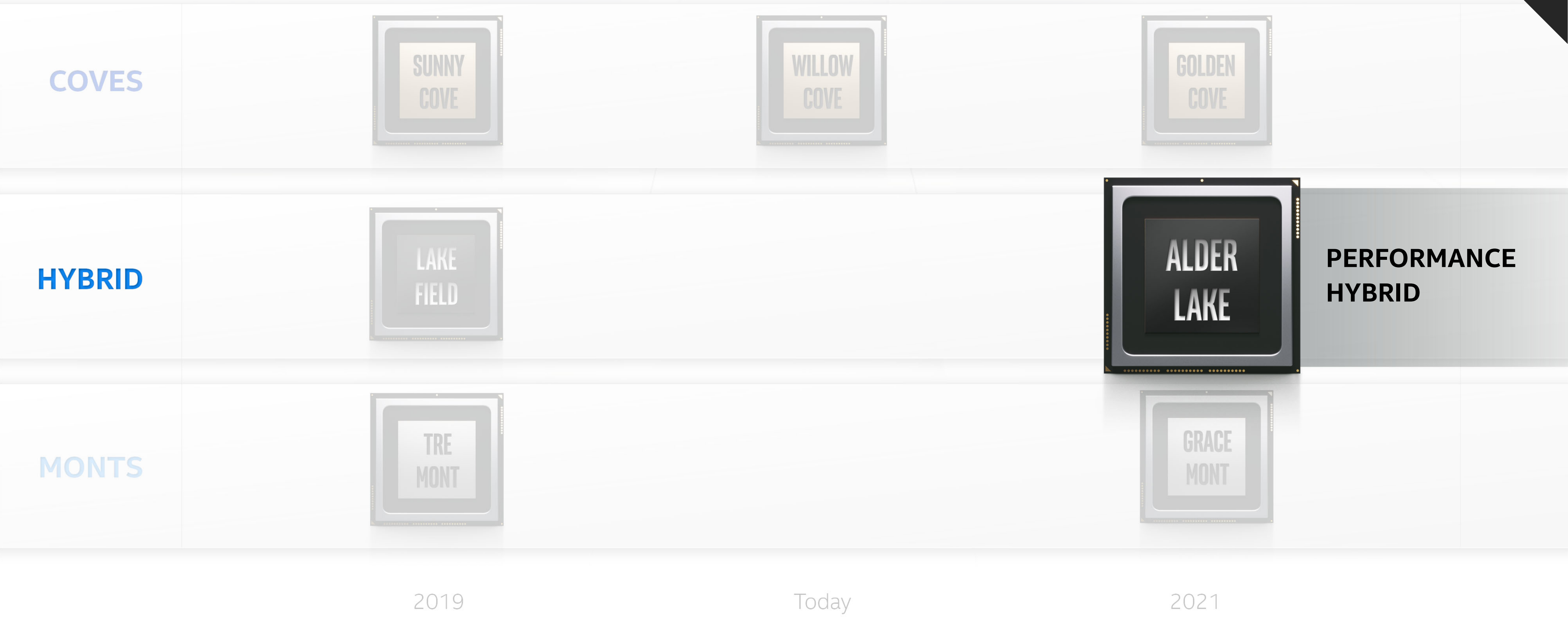
CPU Core Roadmap



	2019	Today	2021
COVES	SUNNY COVE	WILLOW COVE	GOLDEN COVE
HYBRID	LAKE FIELD		
MONTS	TRE MONT		GRACE MONT

NEW DISCLOSURE

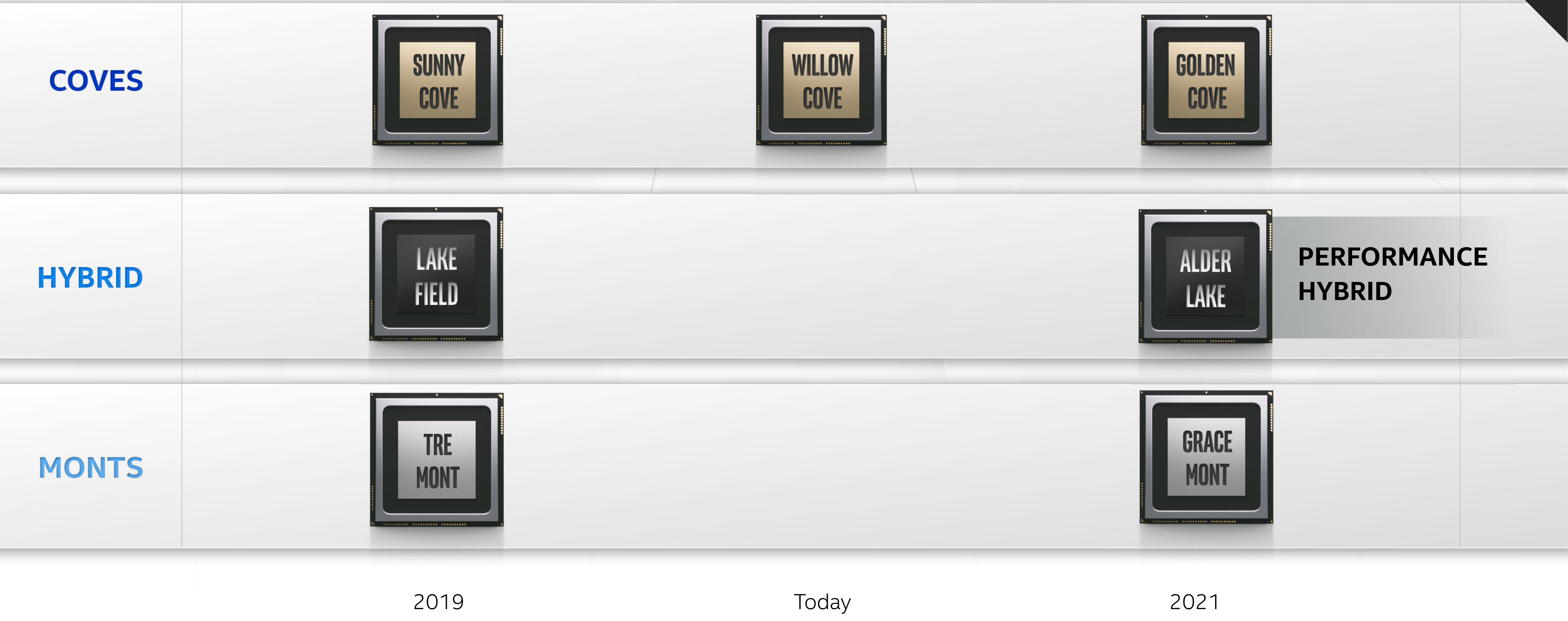
CPU Core Roadmap



PERFORMANCE HYBRID

CPU Core Roadmap

NEW DISCLOSURE



	2019	Today	2021
COVES	SUNNY COVE	WILLOW COVE	GOLDEN COVE
HYBRID	LAKE FIELD		ALDER LAKE
MONTS	TRE MONT		GRACE MONT

PERFORMANCE HYBRID

INTEL X^{e} GPU

Scalable Vector-Matrix Architecture

GPU ARCHITECTURE STRATEGY

Brand new architecture and 2 new micro architectures.

WHAT WE SAID IN 2018

2018 ARCHITECTURE DAY

X^e STRATEGY

SOFTWARE FIRST

XEON AND INTEGRATED OPEN SOURCE BASE

SCALABILITY NEXT

INSIDE CHIP, INSIDE PACKAGE, SCALE UP AND SCALE OUT

NEW WORKLOADS

AI AND VISUAL CLOUD OPTIMIZED

UPDATE IN 2019

GPU Architecture Strategy

One Architecture and 3 Micro Architectures

GPU Architecture Strategy

One Architecture and 3 Micro Architectures

X^e

X^e_{HPC}

X^e_{HP}

X^e_{LP}

HPC EXASCALE

DATA CENTER / AI

ENTHUSIAST

MID RANGE

INTEGRATED + ENTRY

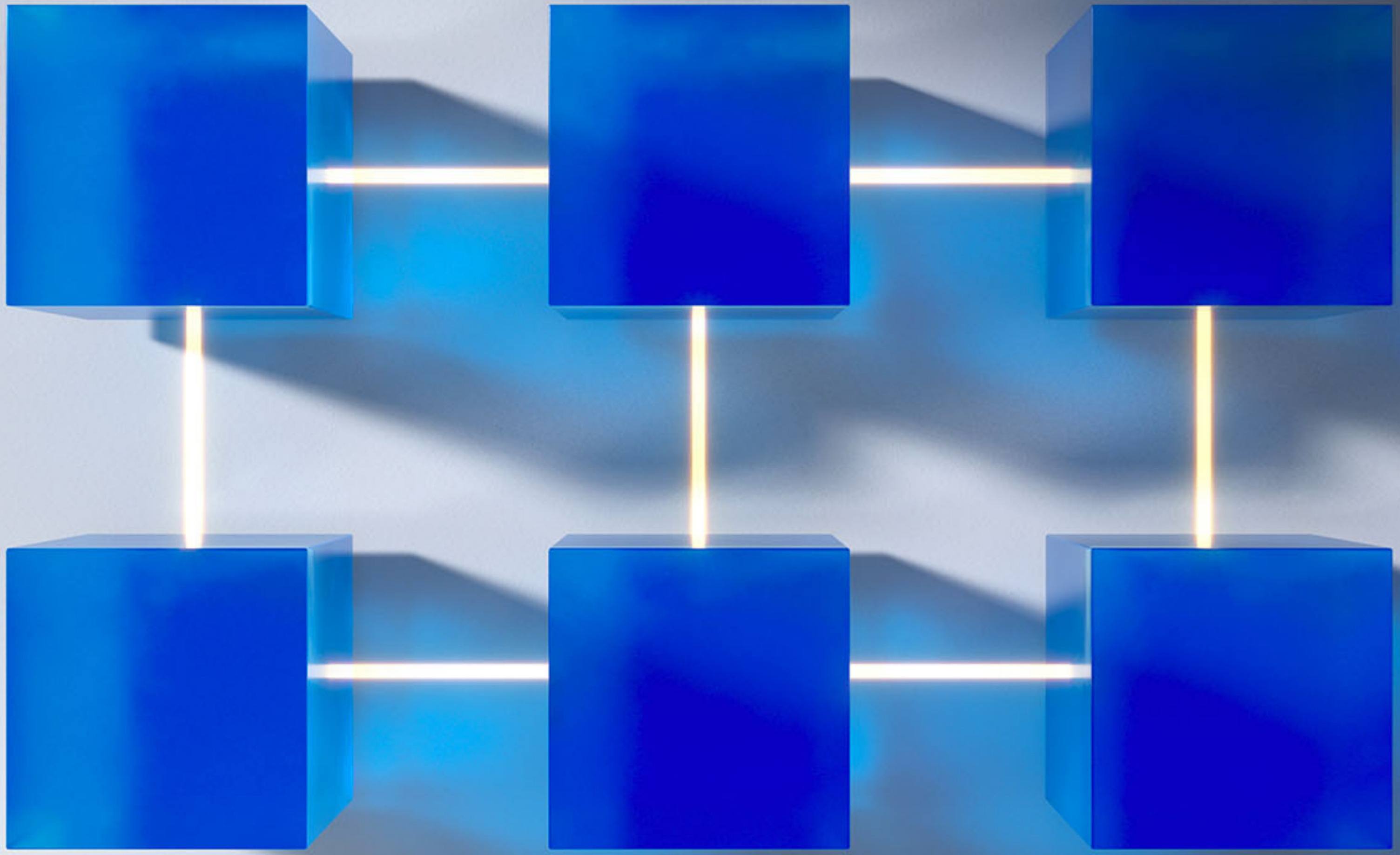
Teraflops to Peta-Ops

Xe LP

Microarchitecture

David Blythe

Senior Fellow,
Graphics Architecture



intel | TECHNOLOGY PILLARS

Architecture Day 2020

Doubling Down on Performance for Mobile

Xe LP

DESIGNED FOR PERFORMANCE

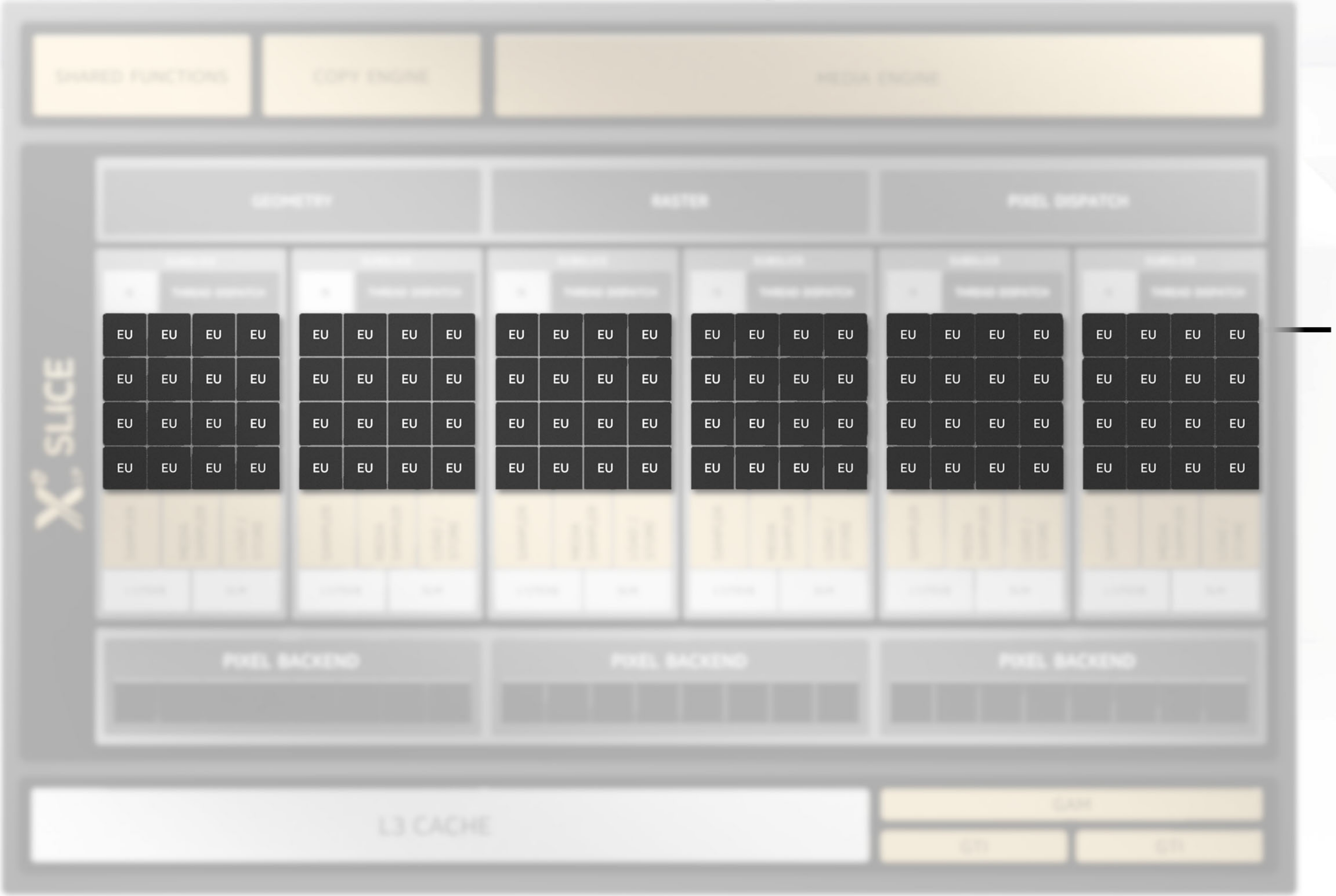
- MORE FLOPS
- MORE PERF/FLOP
- LESS POWER/FLOP

1.5x Larger Engine

NEW DISCLOSURE

NEW DISCLOSURE

1.5x Larger Engine



Up to
96 EUs

1536 flops/clock

NEW DISCLOSURE

1.5x Larger Engine

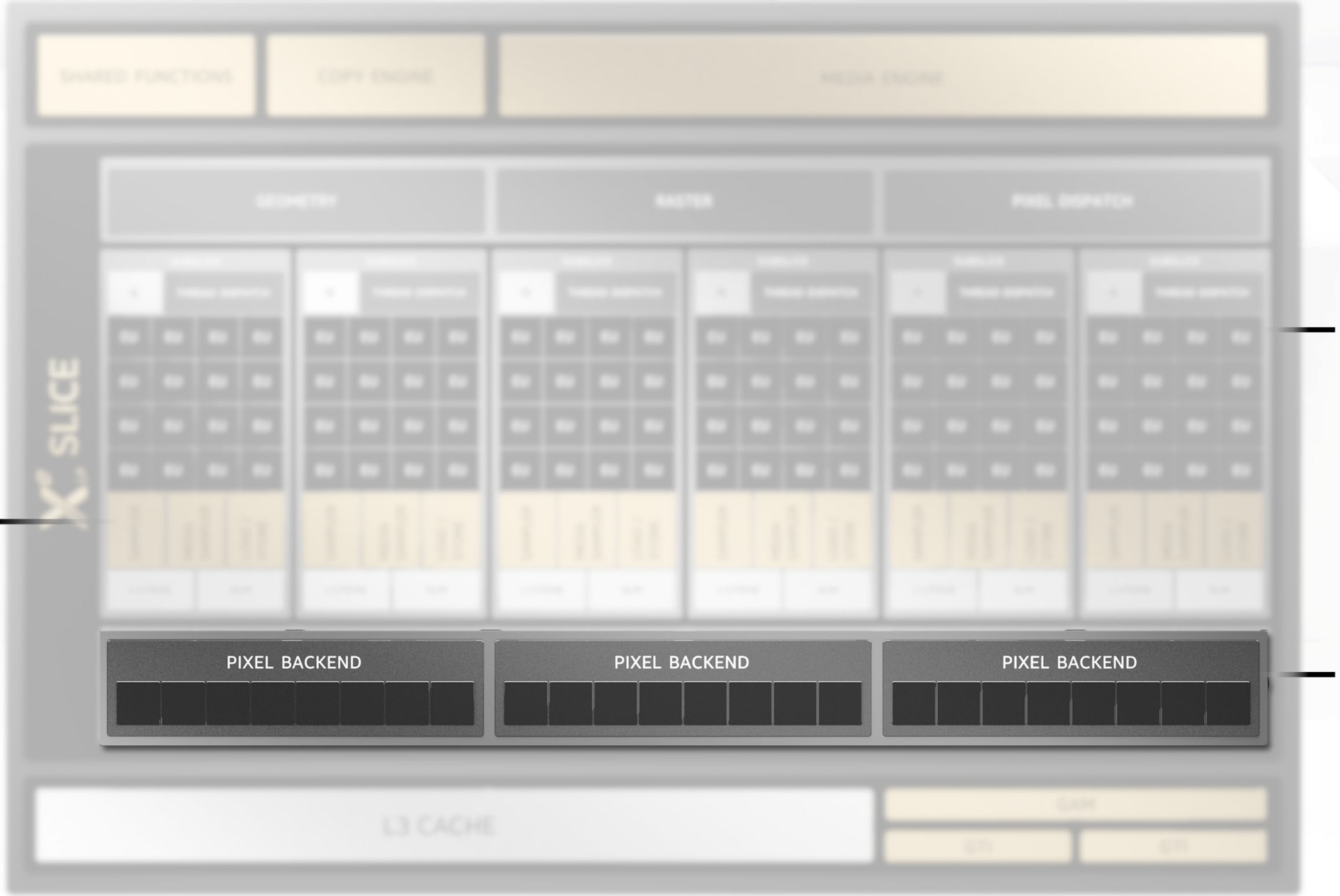
Up to
48 Texels/clock

Up to
96 EUs

1536 flops/clock

NEW DISCLOSURE

1.5x Larger Engine



Up to
96 EUs

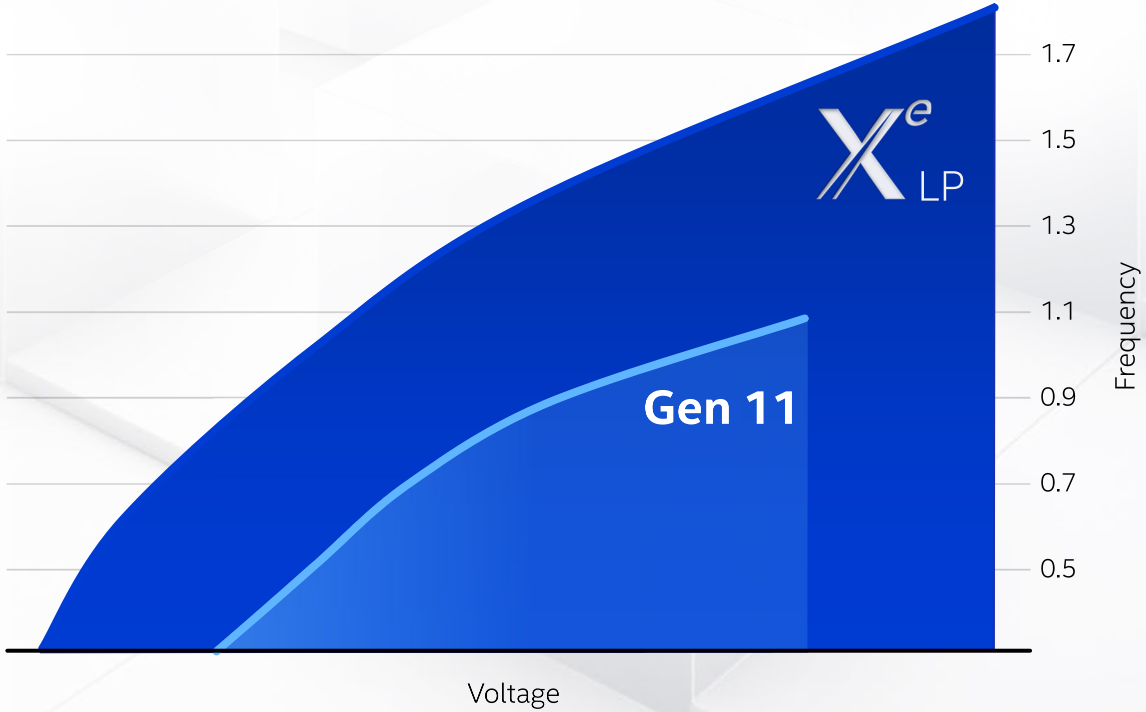
1536 flops/clock

Up to
48 Texels/clock

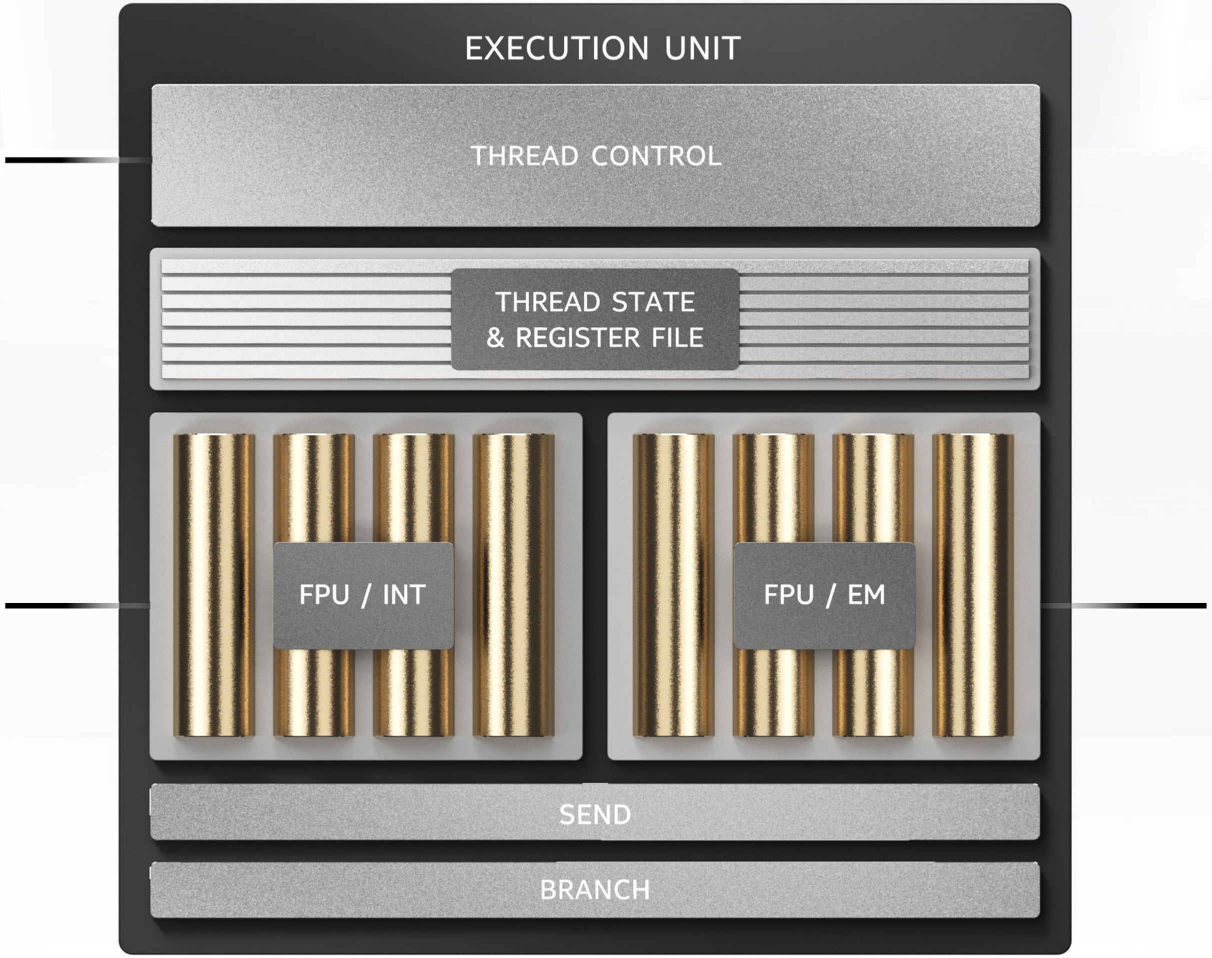
Up to
24 pixels/clock

Efficiency Improvements

- Frequency uplift at iso voltage
- Greater dynamic range
- Repipelining
- Bottlenecks analysis



GEN11 EUs

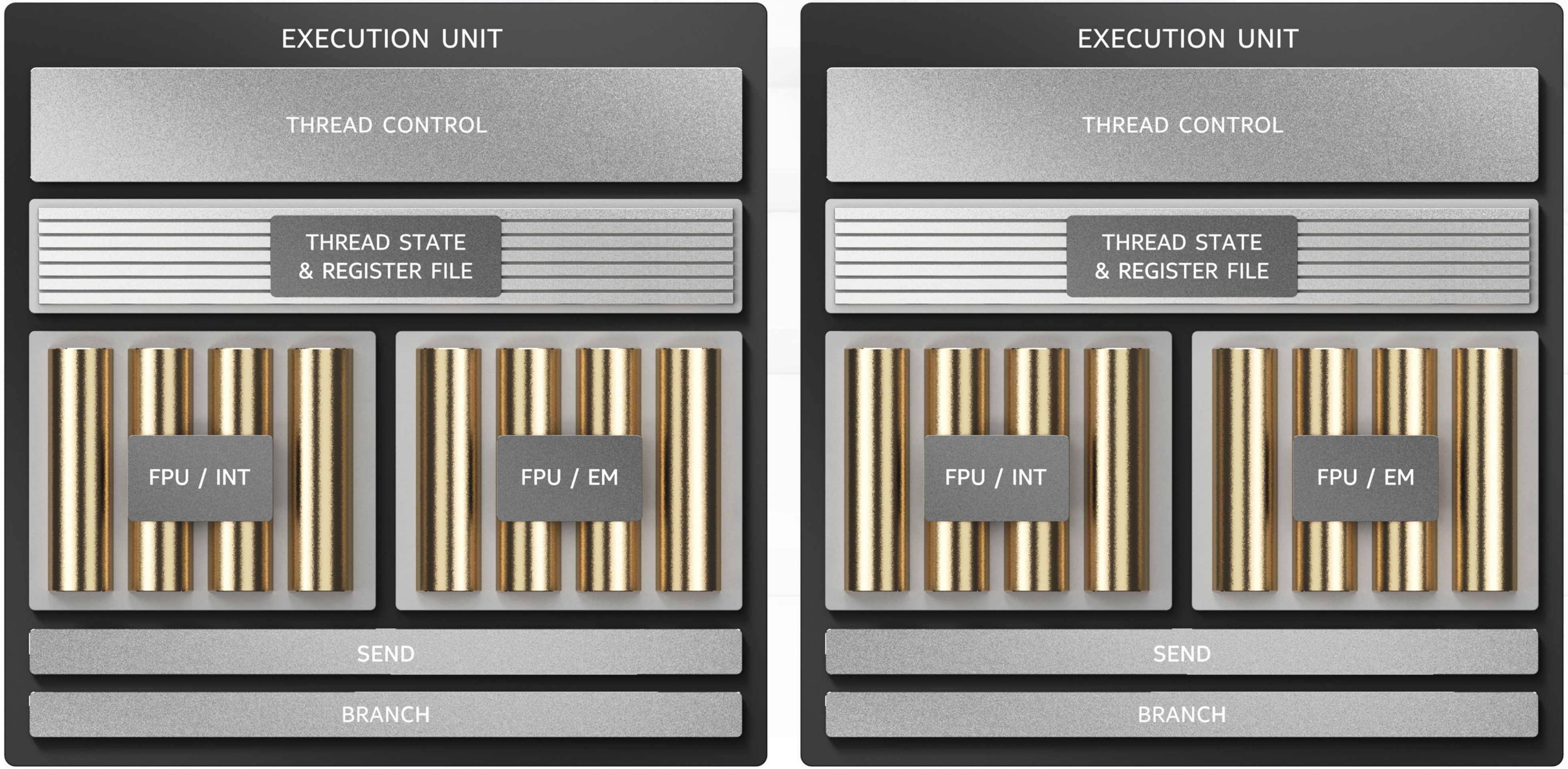


Thread control per EU

4-wide FP/INT ALU

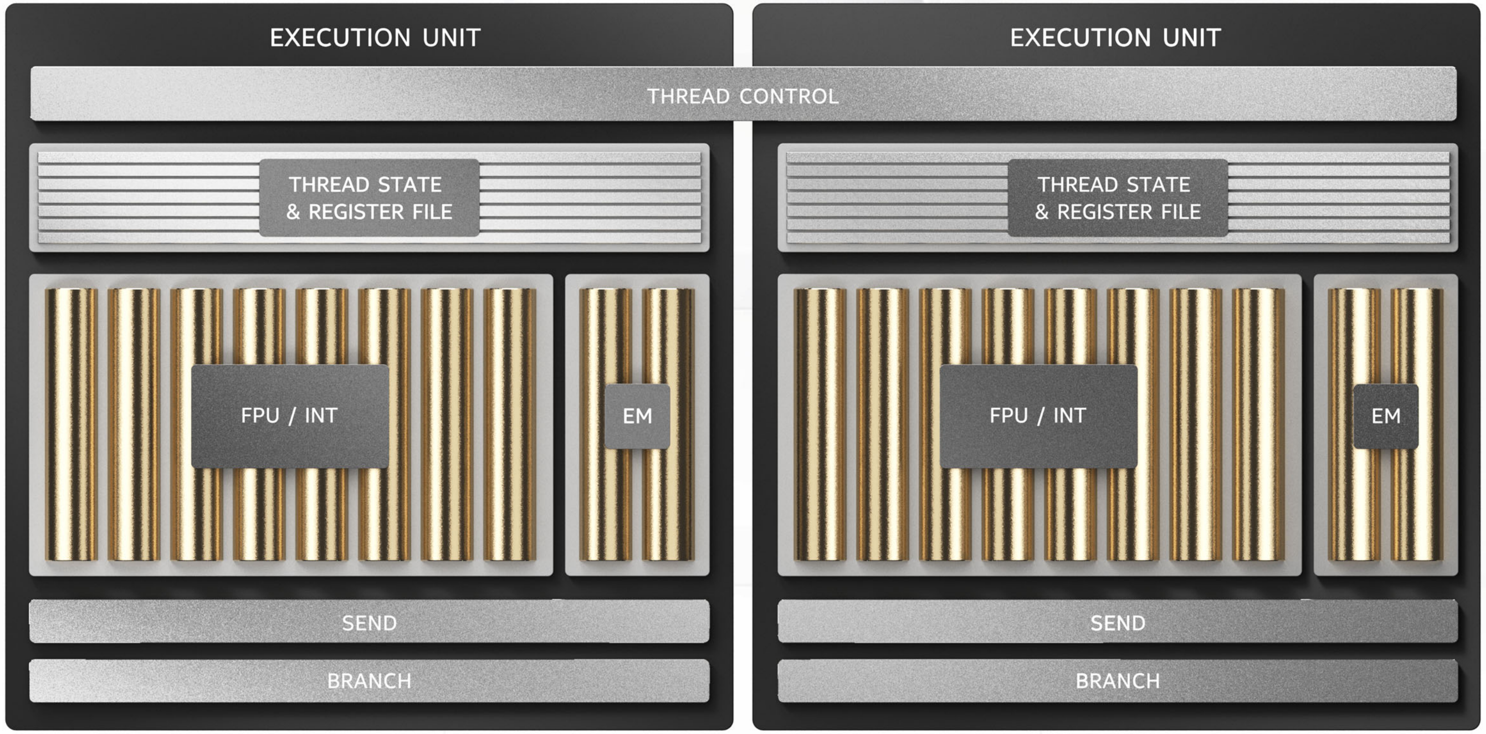
4-wide FP/Extended Math ALU

GEN11 EUs


EXECUTION UNIT
THREAD CONTROL
THREAD STATE & REGISTER FILE
FPU / INT
FPU / EM
SEND
BRANCH
EXECUTION UNIT
THREAD CONTROL
THREAD STATE & REGISTER FILE
FPU / INT
FPU / EM
SEND
BRANCH

X^e_{LP} EUs

NEW DISCLOSURE



Xe LP EUs

NEW DISCLOSURE

High-efficiency thread control

for pairs of EUs

NEW DISCLOSURE

Xe LP EUs

High-efficiency thread control

Software score boarding

8-wide FP/INT ALU

2x INT16 and INT32 rates

Fast INT8 with DP4A

X^{e}_{LP} EUs

NEW DISCLOSURE

High-efficiency thread control

Software score boarding

8-wide FP/INT ALU

2x INT16 and INT32 rates

Fast INT8 with DP4A

2-wide Extended Math ALUs

Memory System

NEW DISCLOSURE

New L1 Data Cache

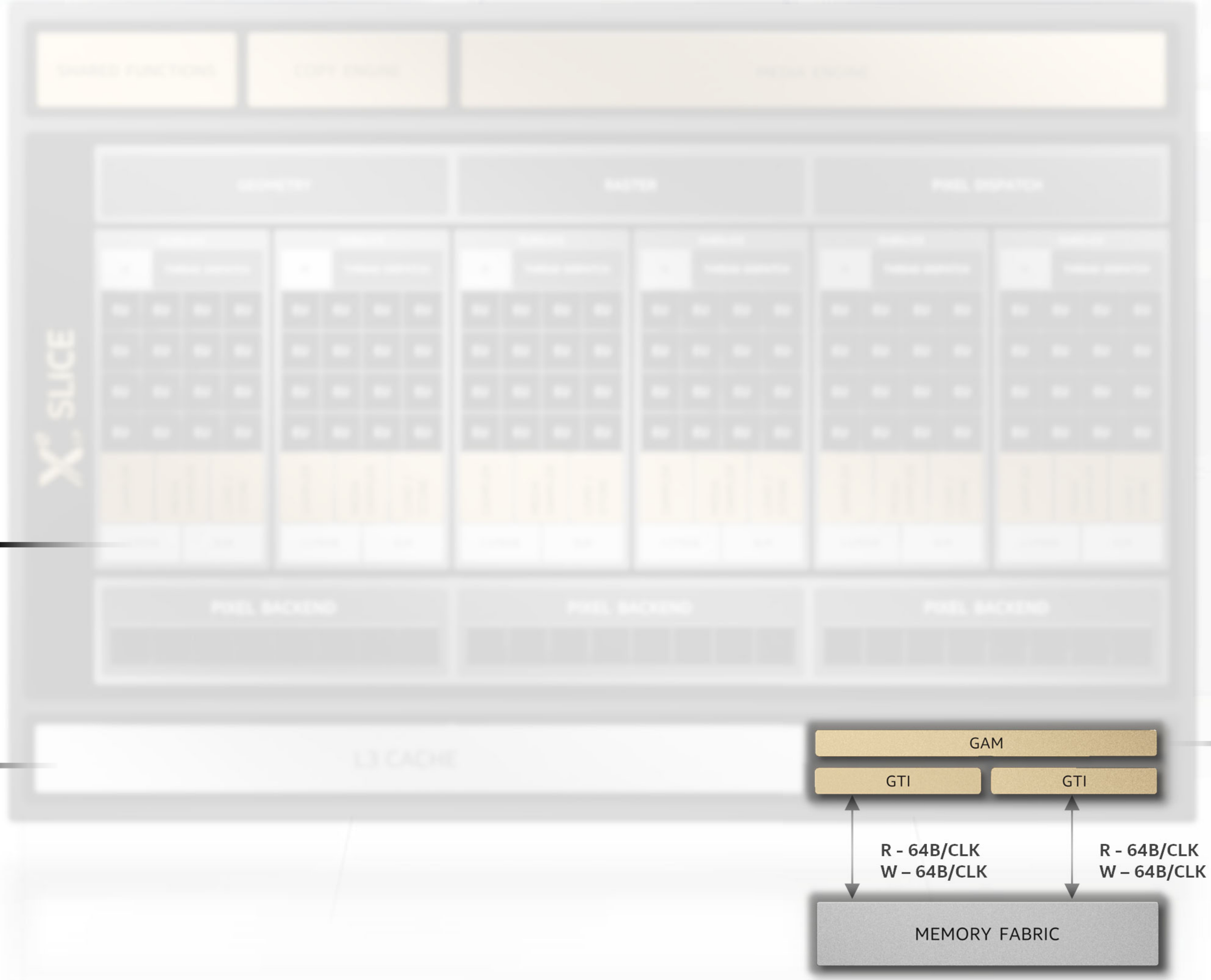
Memory System

NEW DISCLOSURE

New L1 Data Cache

Up to 16MB L3 Cache

Memory System

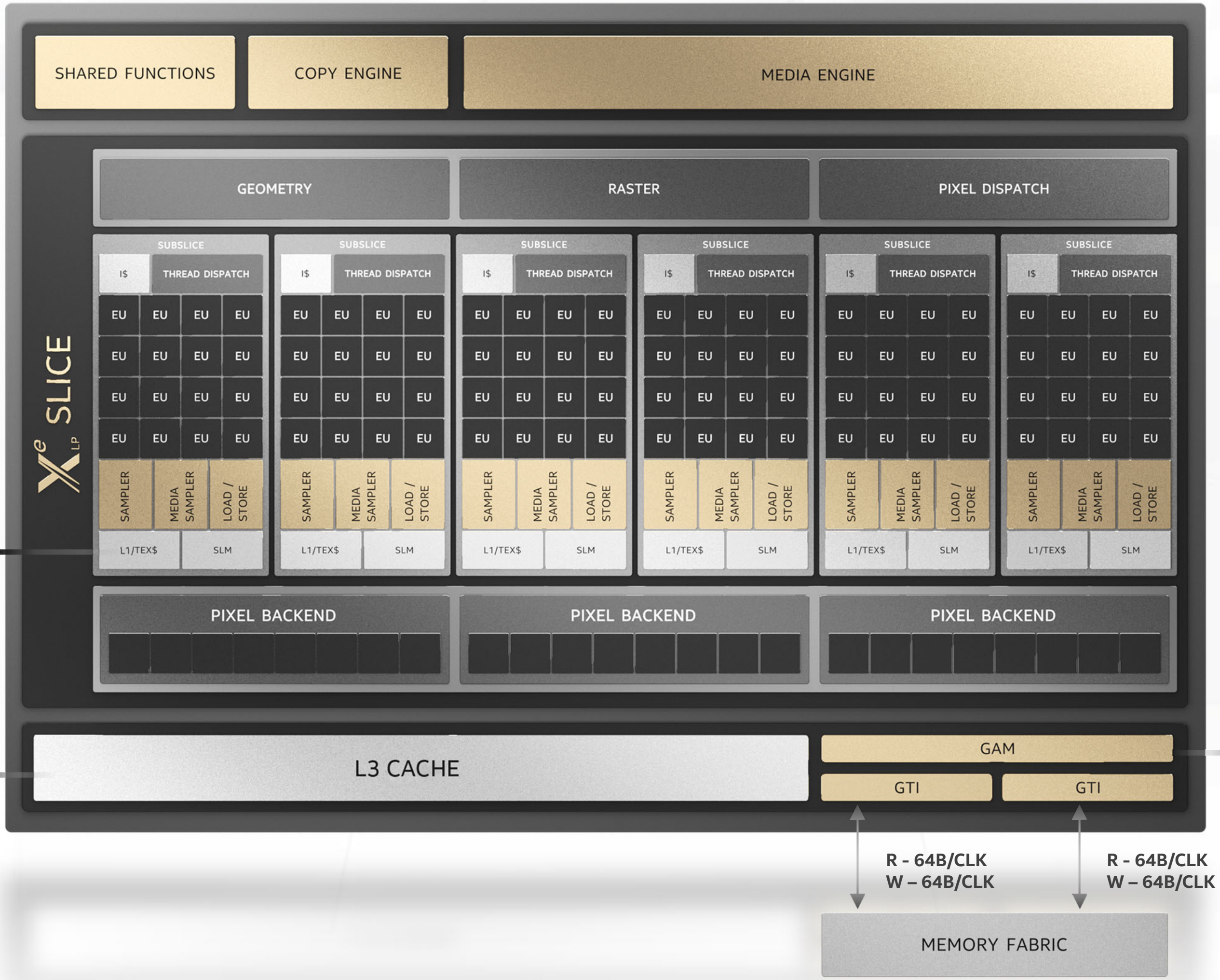


New L1 Data Cache

Up to 16MB L3 Cache

2x GTI Bandwidth

Xe LP Memory System



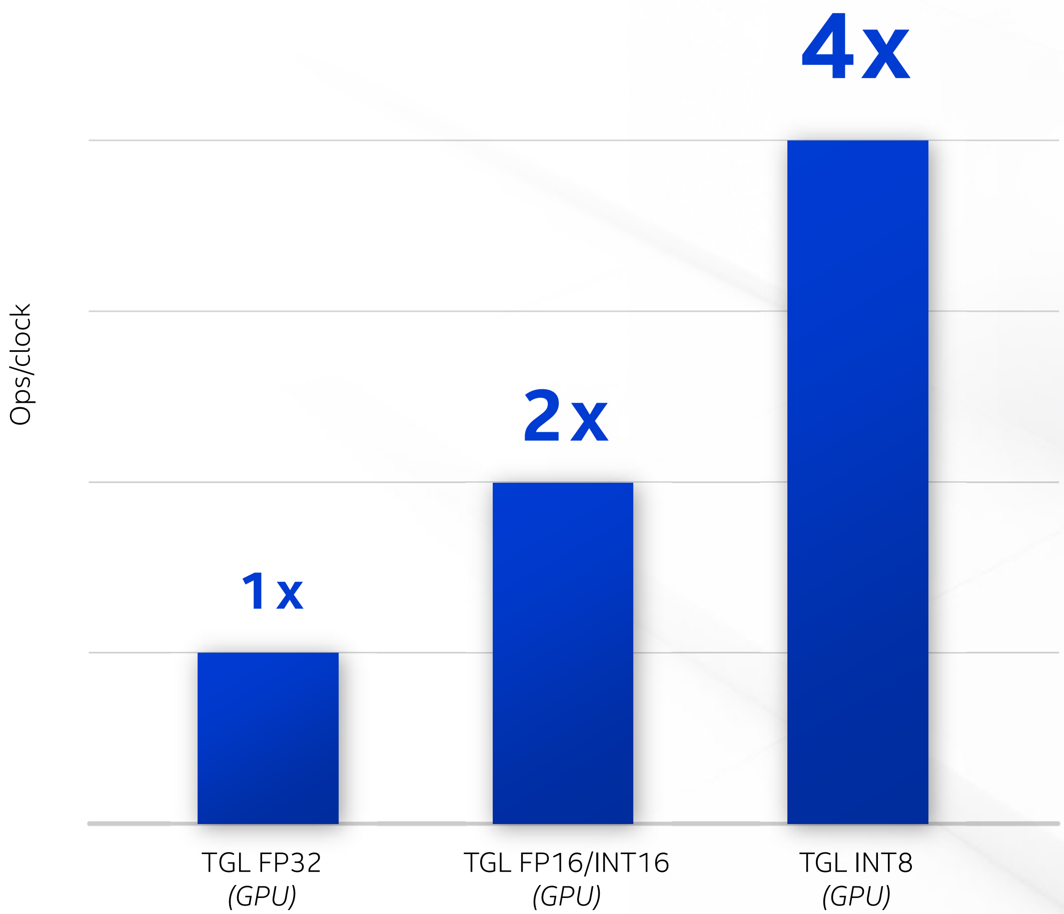
New L1 Data Cache

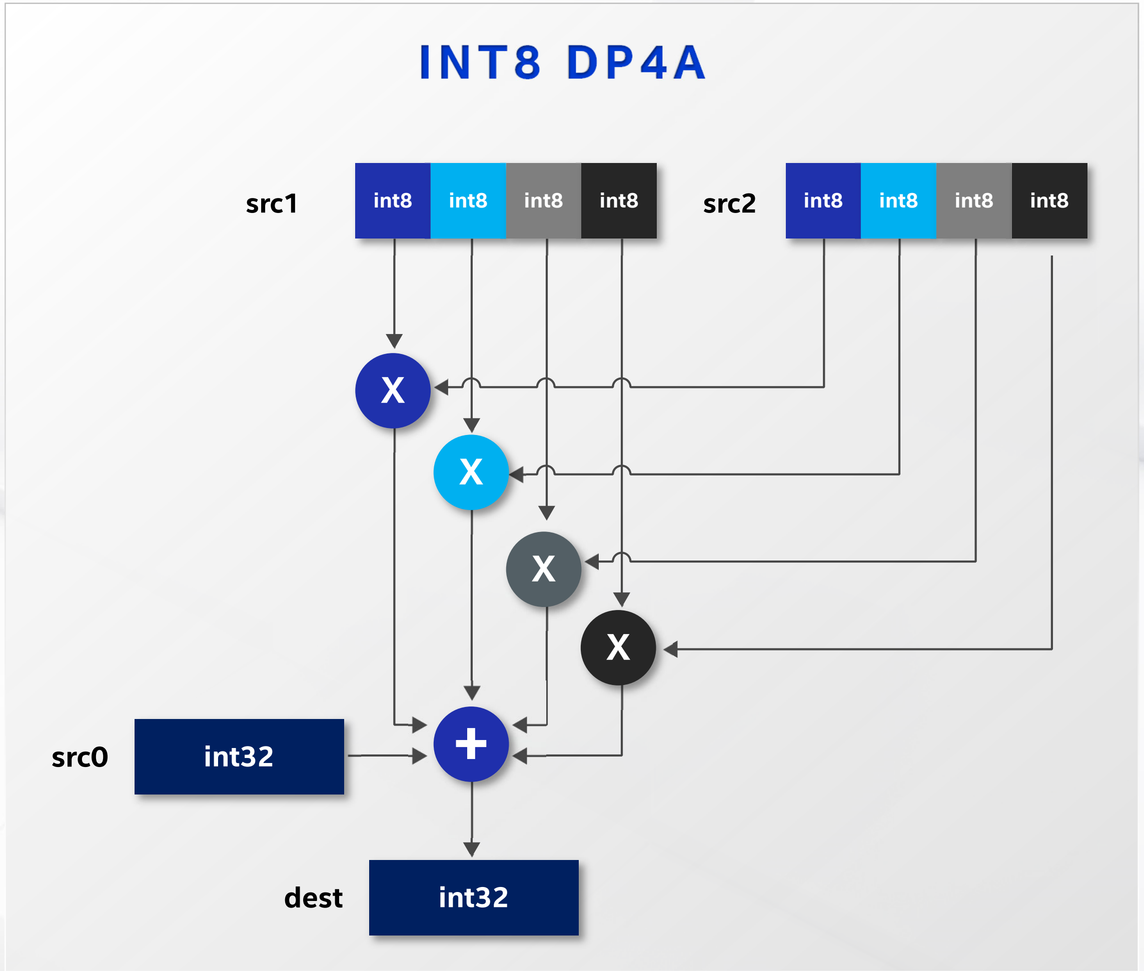
Up to 16MB L3 Cache

End-to End Compression

2x GTI Bandwidth

Diverse Datatypes for AI





Pushing the Bar on Media and Display

Xe LP

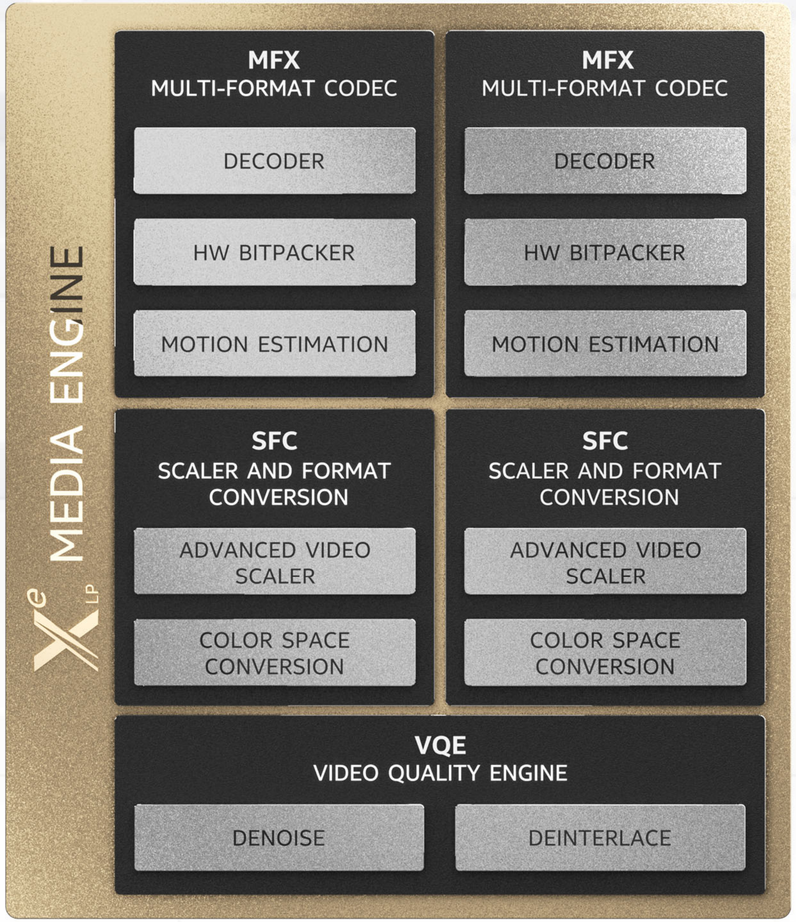
DESIGNED FOR CREATORS

- MORE PIXELS
- MORE COLORS
- LESS POWER

Xe LP Media Engine

NEW DISCLOSURE

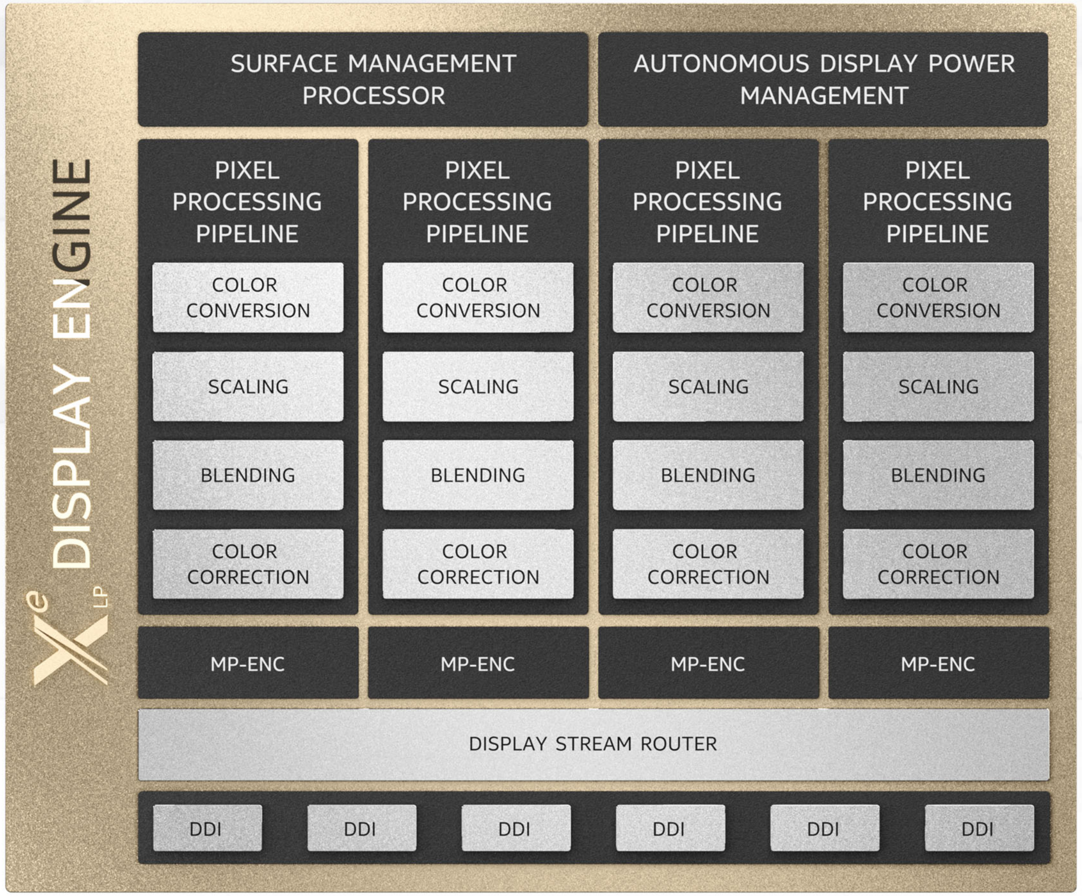
- Up to 2x encode/decode throughput
- AV1 decode acceleration
- HEVC screen content coding support
- 4K/8K60 playback
- HDR/Dolby Vision playback
- 12-bit end-to-end video pipeline



NEW DISCLOSURE

Xe LP Display Engine

- 4 display pipelines
- Dual eDP
- DisplayPort 1.4, HDMI2.0, TBT4, USB4 Type-C
- Up to 8K UHD and Ultra Wide
- HDR10 and Dolby Vision
- Up to 12-bit BT2020 Color
- Up to 360Hz and Adaptive Sync



Xe LP DISPLAY ENGINE

SURFACE MANAGEMENT PROCESSOR

AUTONOMOUS DISPLAY POWER MANAGEMENT

PIXEL PROCESSING PIPELINE	PIXEL PROCESSING PIPELINE	PIXEL PROCESSING PIPELINE	PIXEL PROCESSING PIPELINE
COLOR CONVERSION	COLOR CONVERSION	COLOR CONVERSION	COLOR CONVERSION
SCALING	SCALING	SCALING	SCALING
BLENDING	BLENDING	BLENDING	BLENDING
COLOR CORRECTION	COLOR CORRECTION	COLOR CORRECTION	COLOR CORRECTION
MP-ENC	MP-ENC	MP-ENC	MP-ENC

DISPLAY STREAM ROUTER

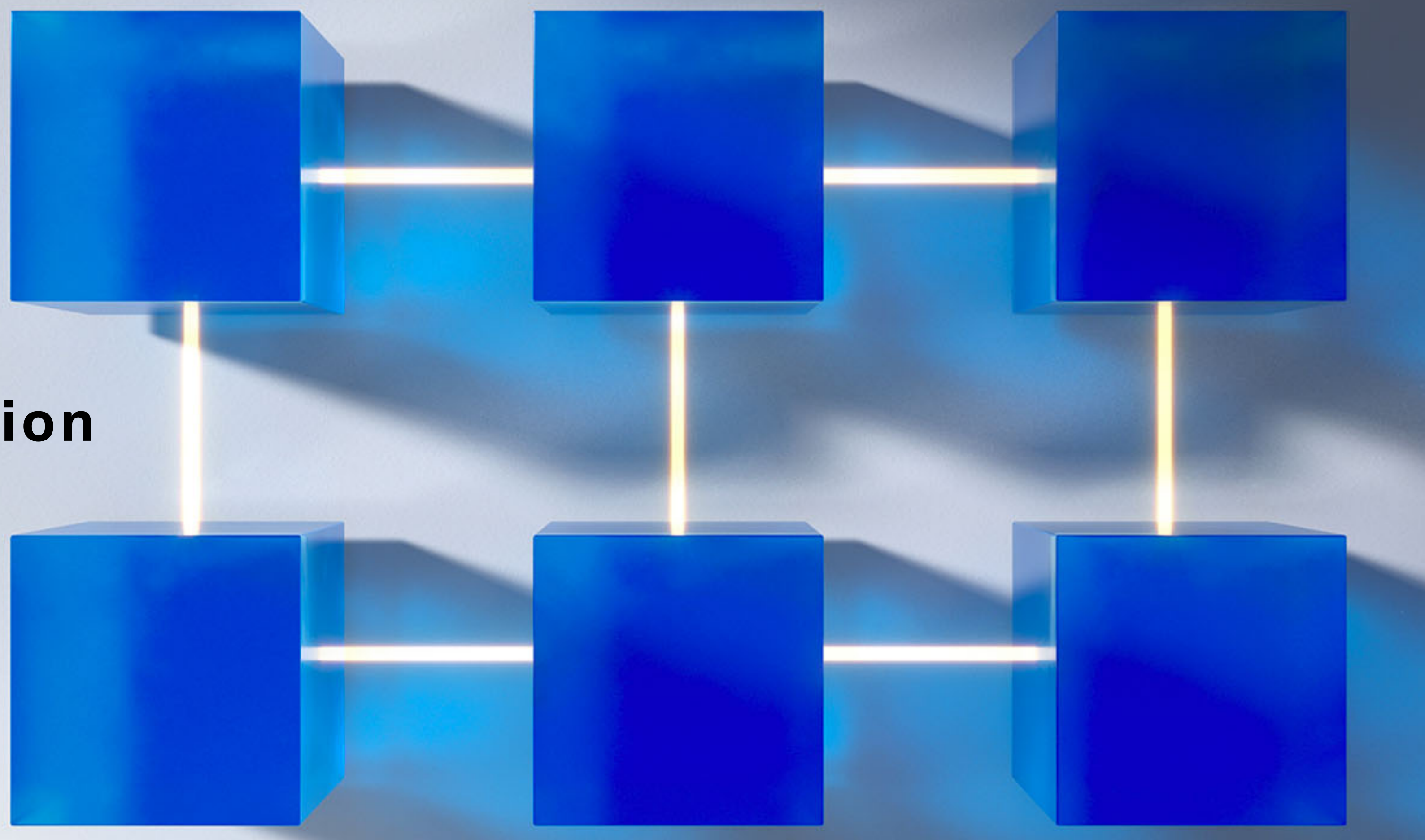
DDI DDI DDI DDI DDI DDI

Xe LP

ENGINEERED FOR EFFICIENCY

- GRAPHICS
- AI
- MEDIA
- DISPLAY

Building the Software Foundation for X^e Graphics



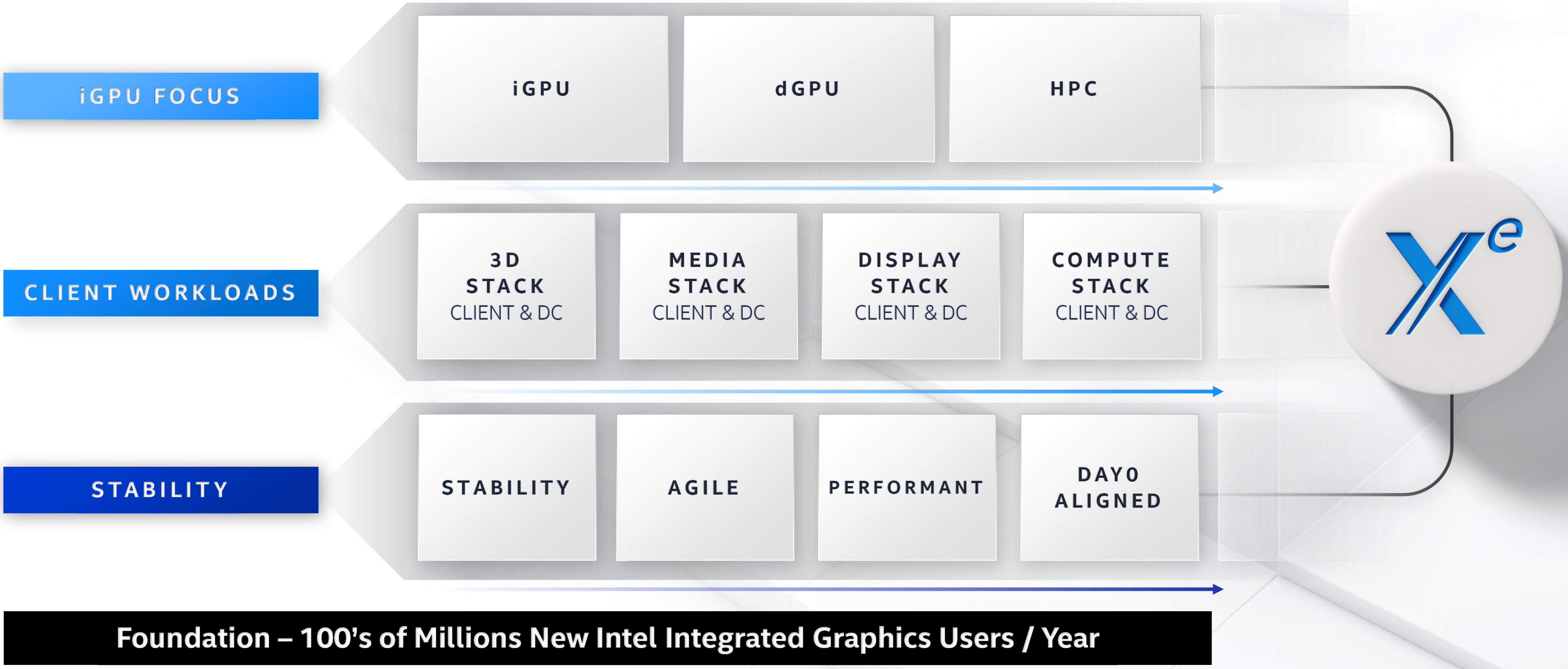
Lisa Pearce

Vice President Graphics
Software Engineering

TECHNOLOGY
PILLARS

Architecture Day 2020

Graphics Software Challenge



iGPU FOCUS

iGPU

dGPU

HPC

CLIENT WORKLOADS

3D STACK
CLIENT & DC

MEDIA STACK
CLIENT & DC

DISPLAY STACK
CLIENT & DC

COMPUTE STACK
CLIENT & DC

STABILITY

STABILITY

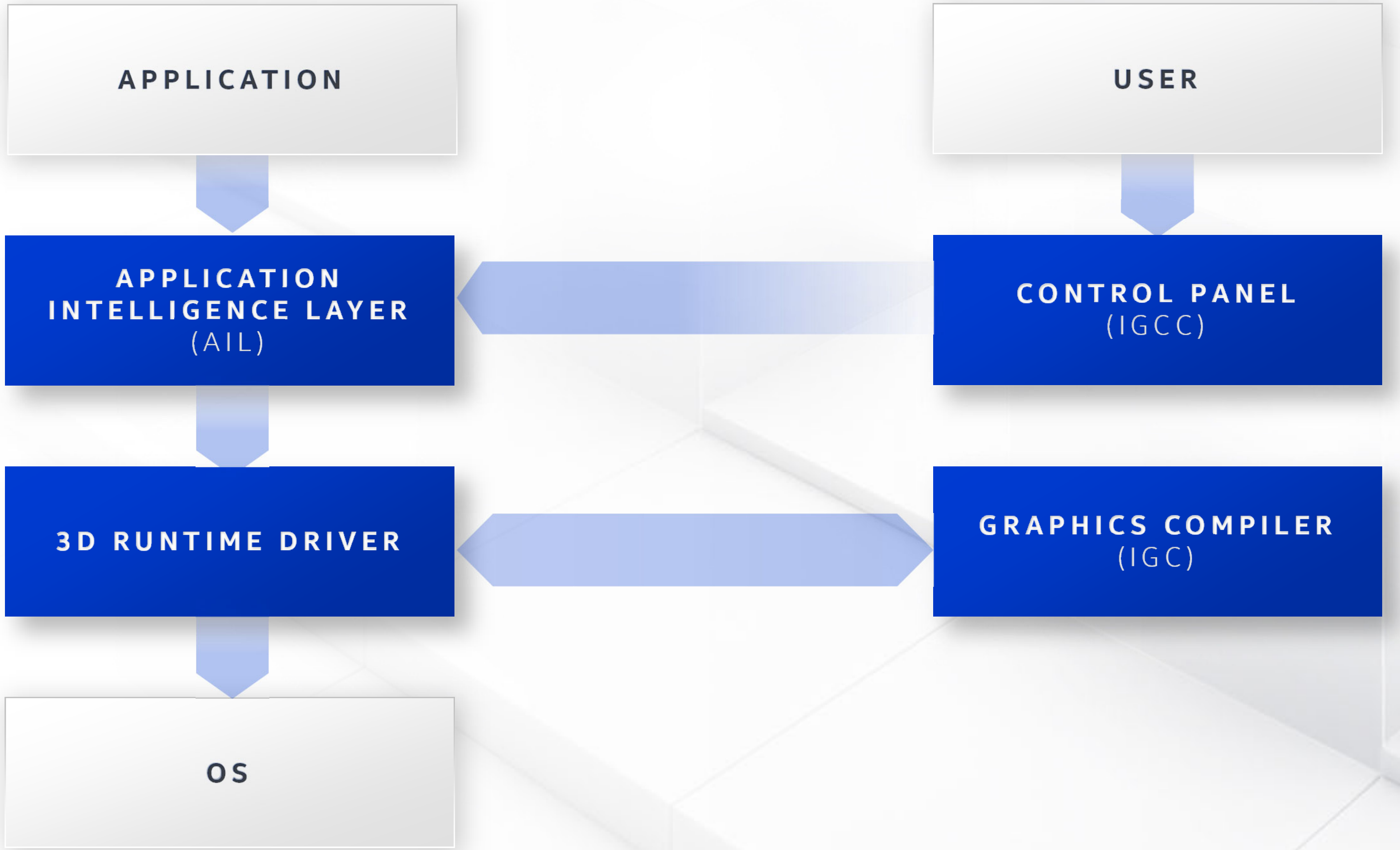
AGILE

PERFORMANT

DAY0 ALIGNED

Foundation – 100's of Millions New Intel Integrated Graphics Users / Year

Pushing Efficiency Through the Driver Stack



APPLICATION

USER

APPLICATION INTELLIGENCE LAYER
(AIL)

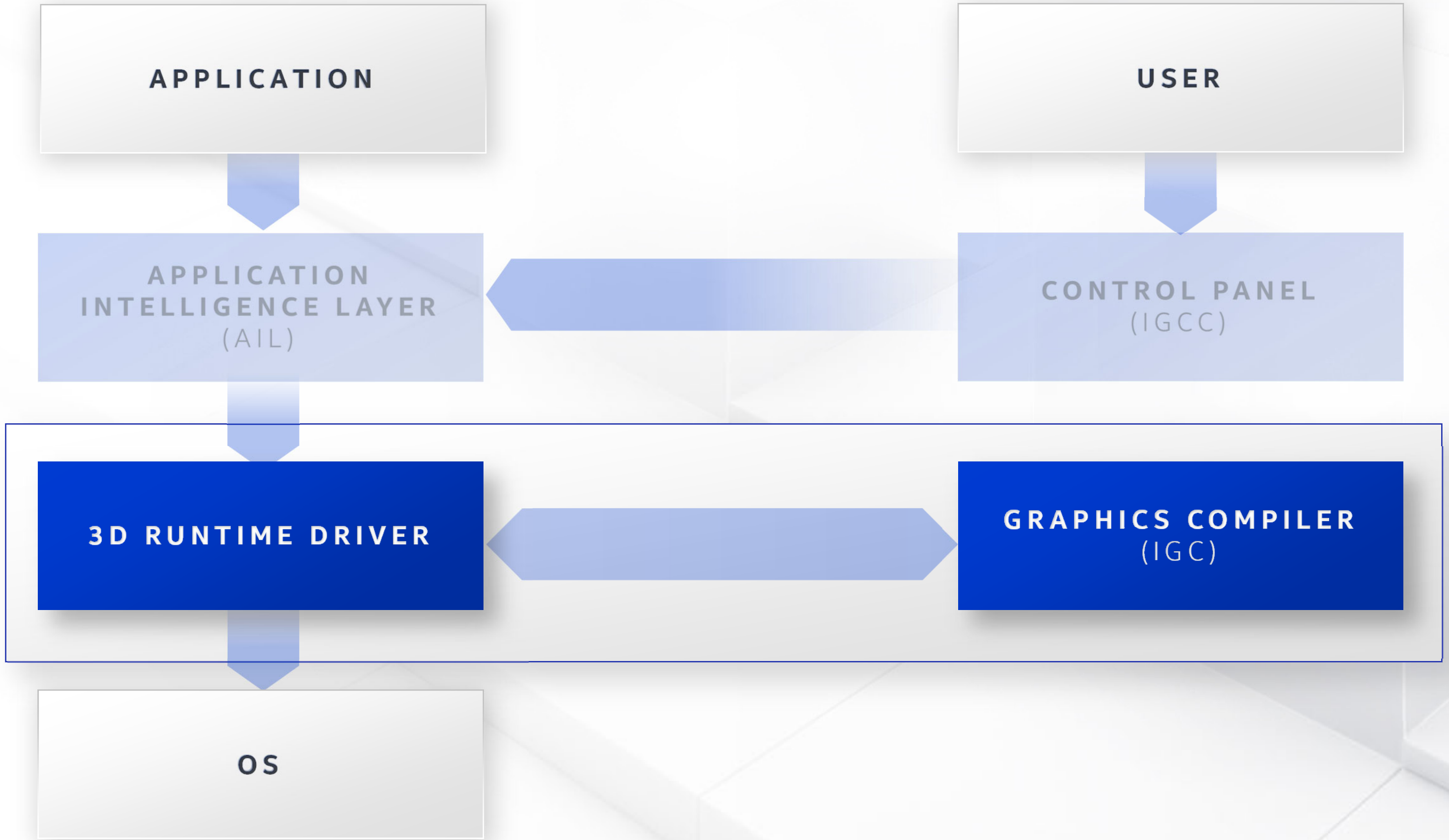
CONTROL PANEL
(IGCC)

3D RUNTIME DRIVER

GRAPHICS COMPILER
(IGC)

OS

Pushing Efficiency Through the Driver Stack



APPLICATION

USER

APPLICATION INTELLIGENCE LAYER (AIL)

CONTROL PANEL (IGCC)

3D RUNTIME DRIVER

GRAPHICS COMPILER (IGC)

OS

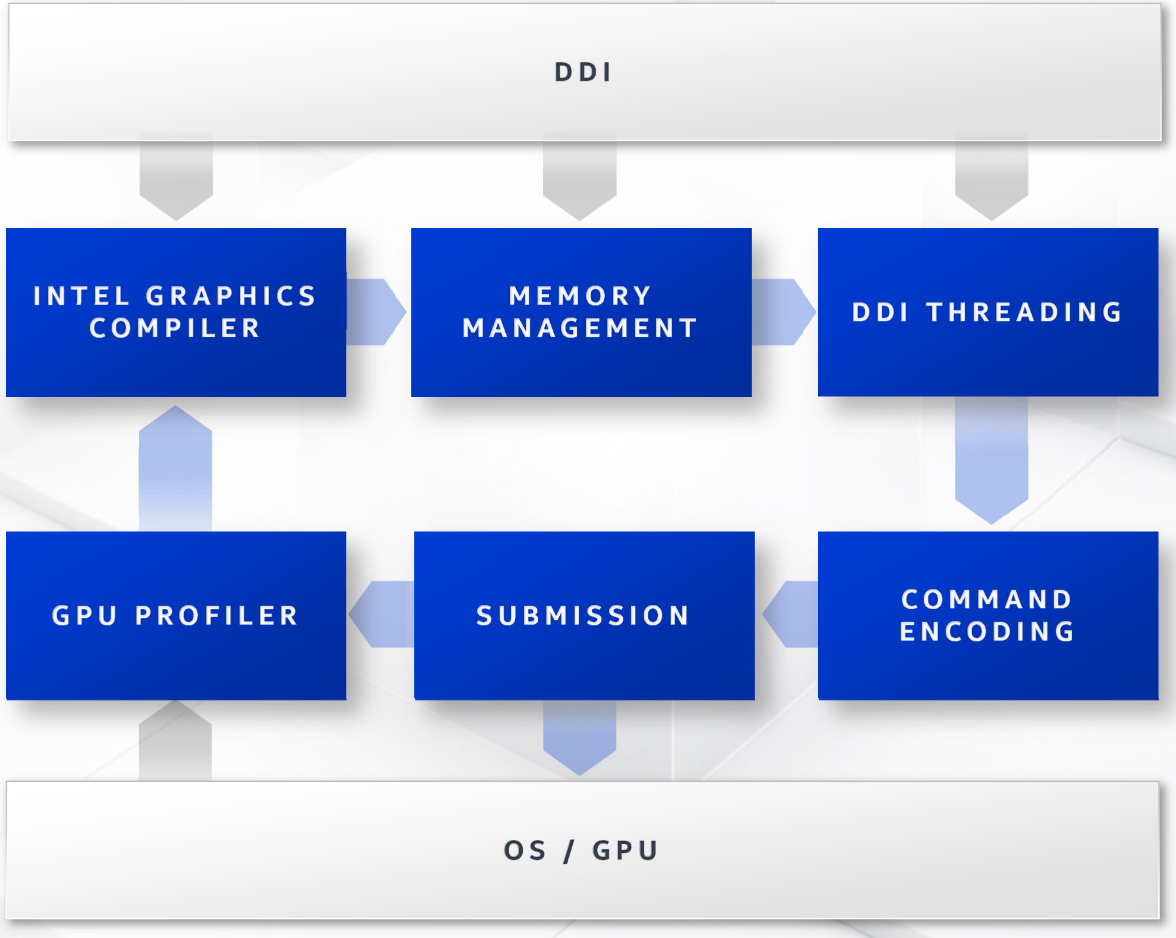
Driver and Compiler Efficiency

Compiler Improvements

- Hardware/software scheduling codesign
- Software score boarding for implicit scheduling and reduced gate count
- Support for AI-optimized instructions

New DirectX 11 driver

- Built from scratch, a strong foundation for the future of Intel GFX
- Lower overhead
  - Reduced GPU pipeline stalls
  - Reduced API latency
  - Local memory optimizations



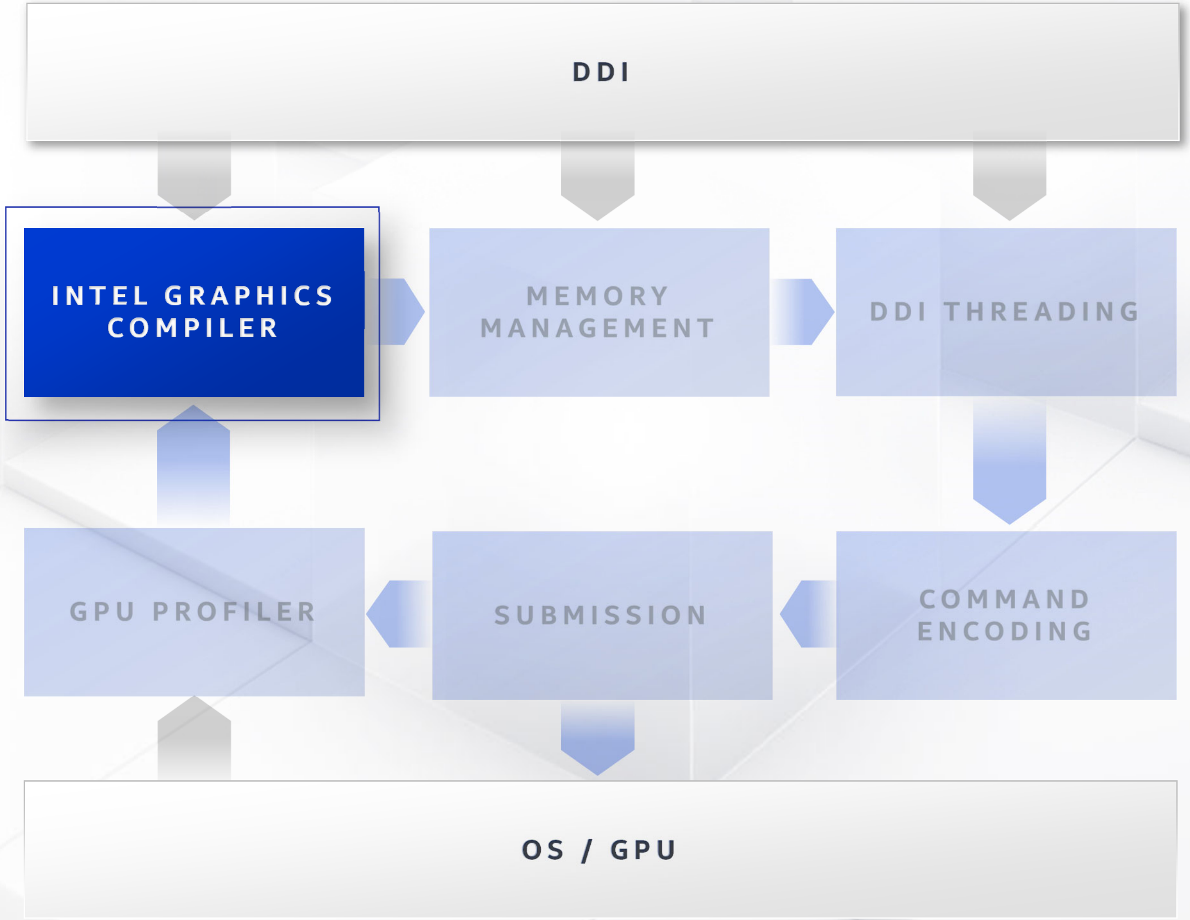
Driver and Compiler Efficiency

Compiler Improvements

- Hardware/software scheduling codesign
- Software score boarding for implicit scheduling and reduced gate count
- Support for AI-optimized instructions

New DirectX 11 driver

- Built from scratch, a strong foundation for the future of Intel GFX
- Lower overhead
    - Reduced GPU pipeline stalls
    - Reduced API latency
    - Local memory optimizations



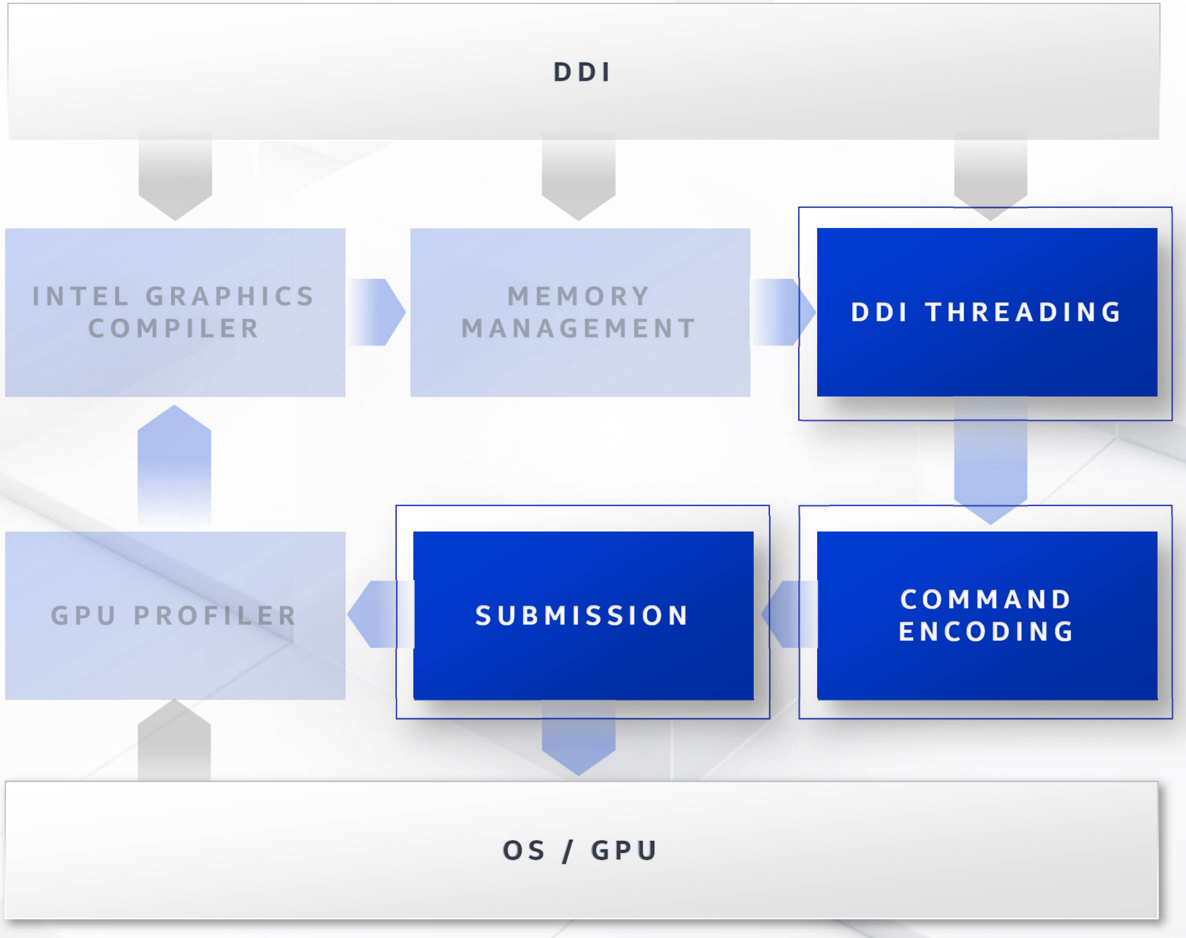
Driver and Compiler Efficiency

Compiler Improvements

- Hardware/software scheduling codesign
- Software score boarding for implicit scheduling and reduced gate count
- Support for AI-optimized instructions

New DirectX 11 driver

- Built from scratch, a strong foundation for the future of Intel GFX
- Lower overhead
  - Reduced GPU pipeline stalls
  - Reduced API latency
  - Local memory optimizations


DDI

INTEL GRAPHICS COMPILER

MEMORY MANAGEMENT

DDI THREADING

GPU PROFILER

SUBMISSION

COMMAND ENCODING

OS / GPU

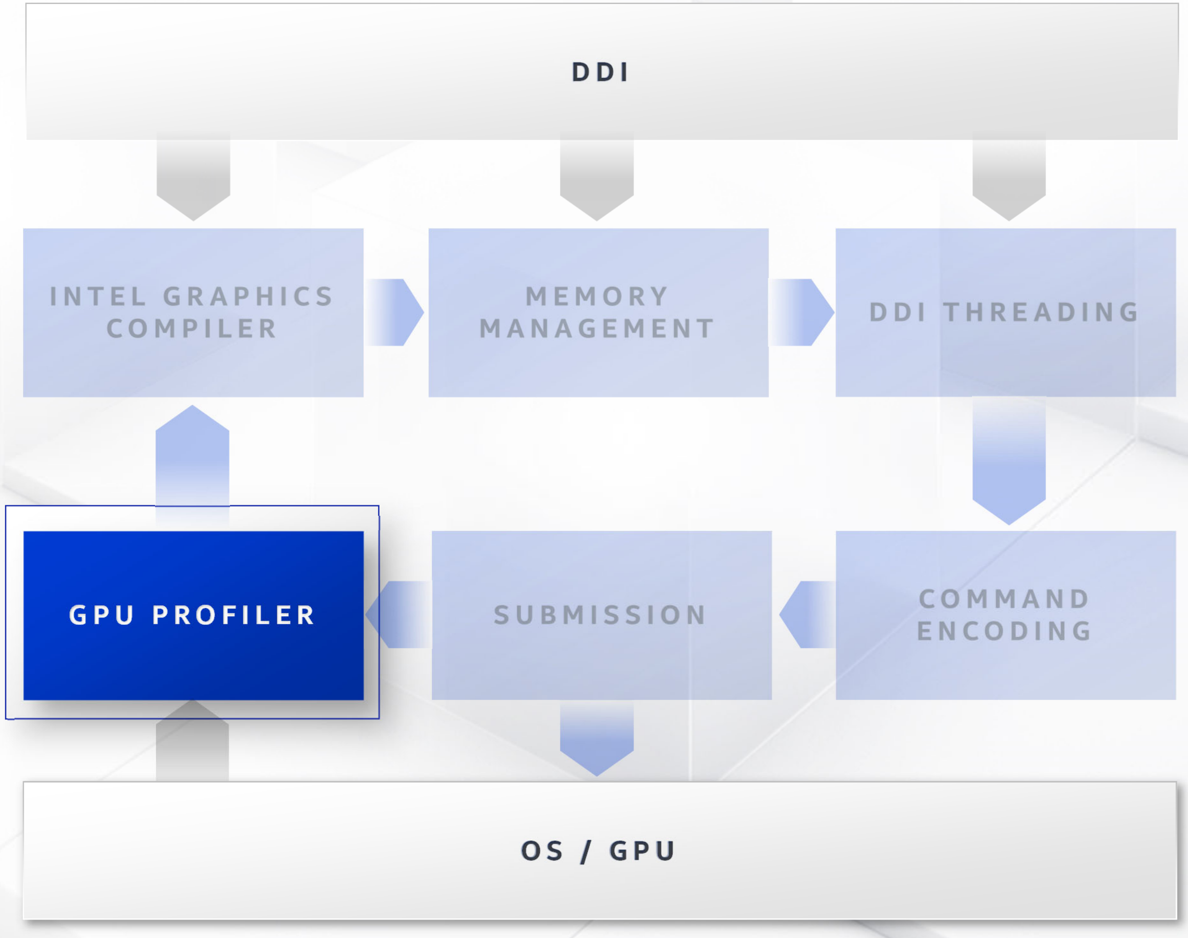
Driver and Compiler Efficiency

Compiler Improvements

- Hardware/software scheduling codesign
- Software score boarding for implicit scheduling and reduced gate count
- Support for AI-optimized instructions

New DirectX 11 driver

- Built from scratch, a strong foundation for the future of Intel GFX
- Lower overhead
  - Reduced GPU pipeline stalls
  - Reduced API latency
  - Local memory optimizations



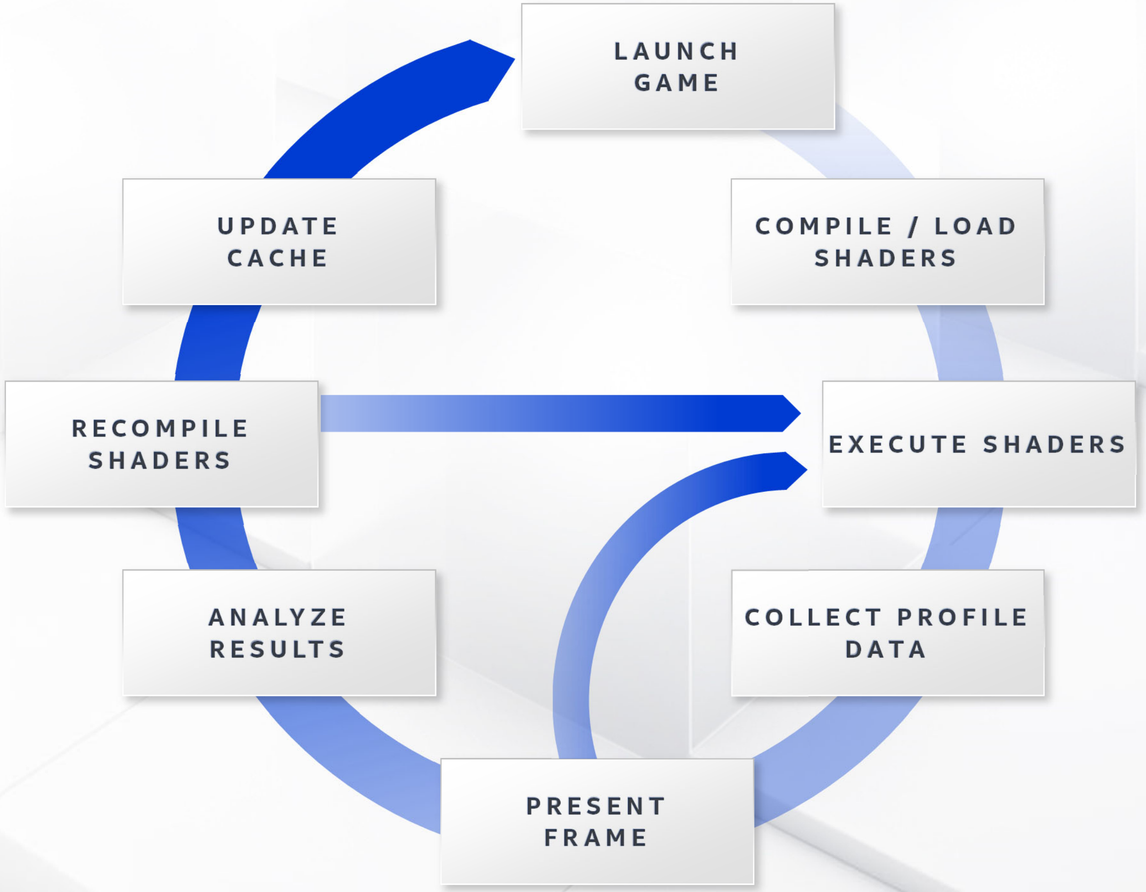
GPU Profile-Guided Optimization

Adaptive GPU optimizations

- Determined by profiling shader execution
- Utilizes advanced architecture-specific performance metrics

Analysis and shader recompilation occur in background

Results are cached to disk for immediate availability



LAUNCH GAME

COMPILE / LOAD SHADERS

EXECUTE SHADERS

COLLECT PROFILE DATA

PRESENT FRAME

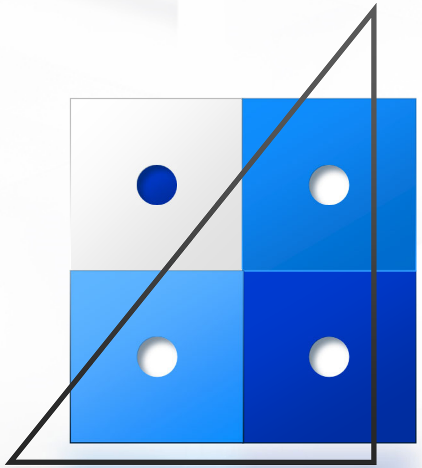
ANALYZE RESULTS

RECOMPILE SHADERS

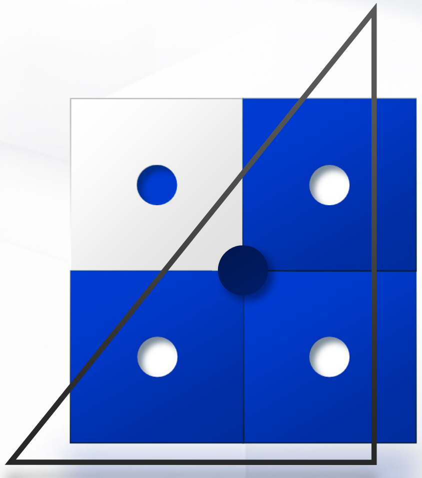
UPDATE CACHE

Variable Rate Shading

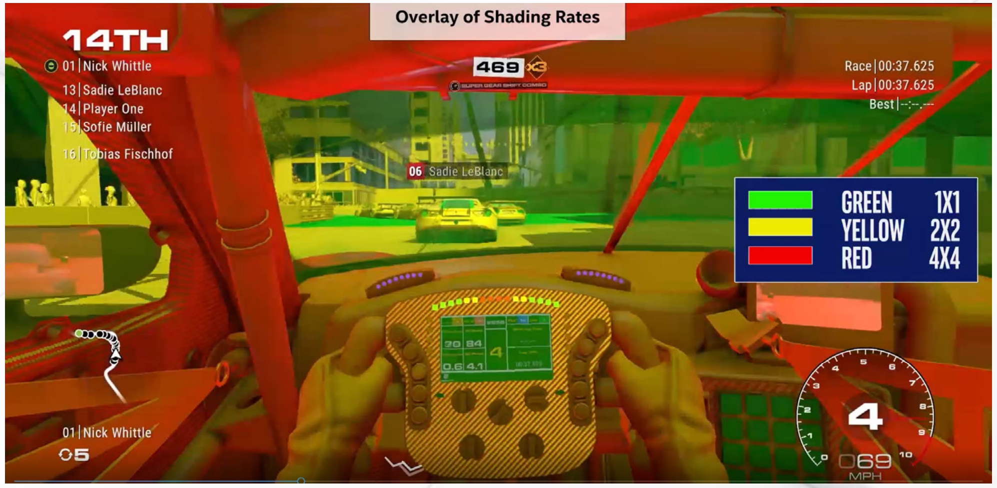
- Derived from Intel Coarse Pixel Shading
- Per draw shading rate optimization
- First Intel implementation on Ice Lake
- Similar performance boost on X^{e}-LP



Traditional 1x1



VRS 2x2



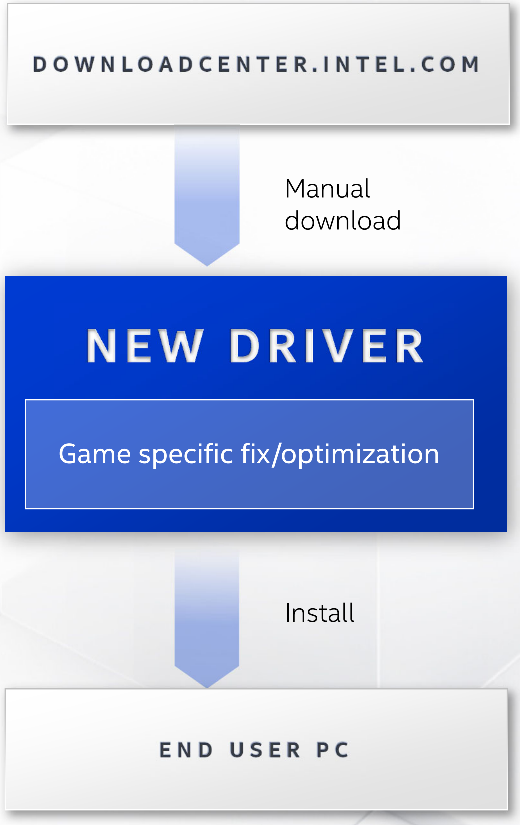
Overlay of Shading Rates

🟩	GREEN	1X1
🟨	YELLOW	2X2
🟥	RED	4X4

Instant Game Tuning

NEW DISCLOSURE

- Faster delivery of game optimizations
- Automatic tuning for specific games and SKUs
- Delivered without driver updates



DOWNLOADCENTER.INTEL.COM

Manual download

NEW DRIVER

Game specific fix/optimization

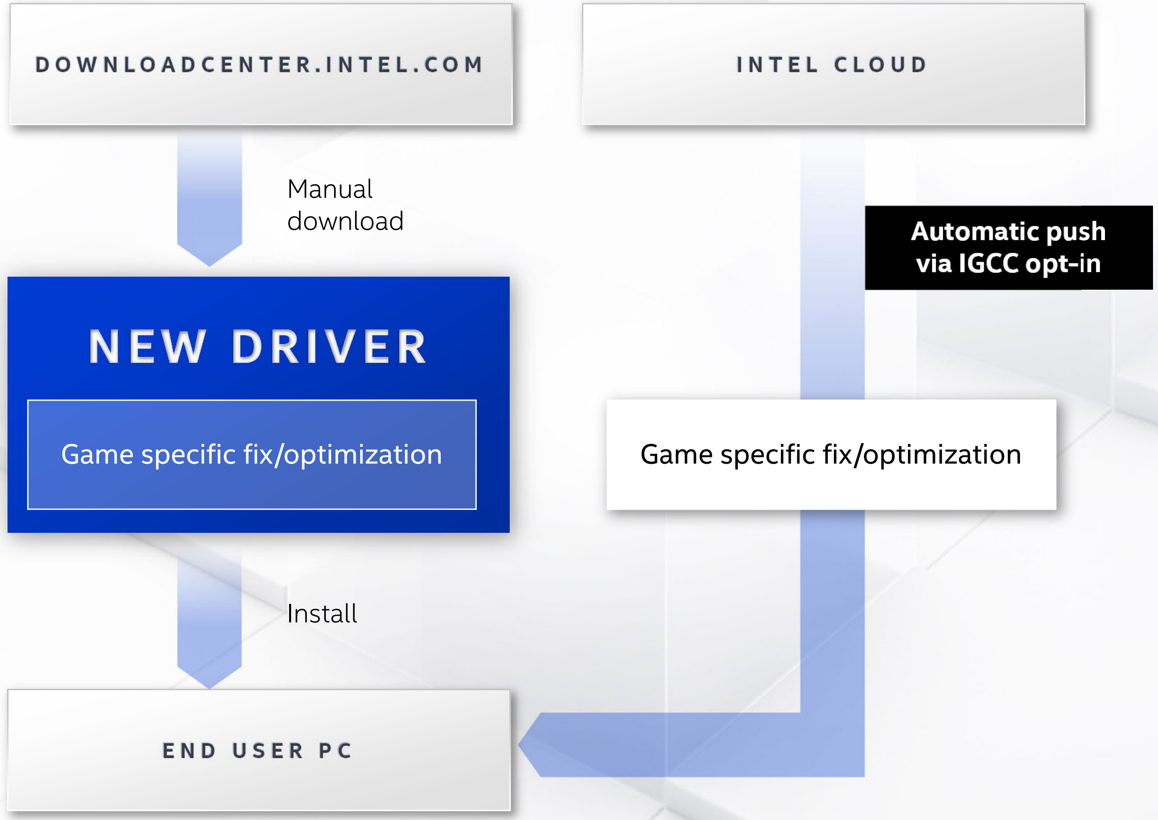
Install

END USER PC

Today

Instant Game Tuning

- Faster delivery of game optimizations
- Automatic tuning for specific games and SKUs
- Delivered without driver updates



DOWNLOADCENTER.INTEL.COM

INTEL CLOUD

Manual download

Automatic push via IGCC opt-in

NEW DRIVER

Game specific fix/optimization

Game specific fix/optimization

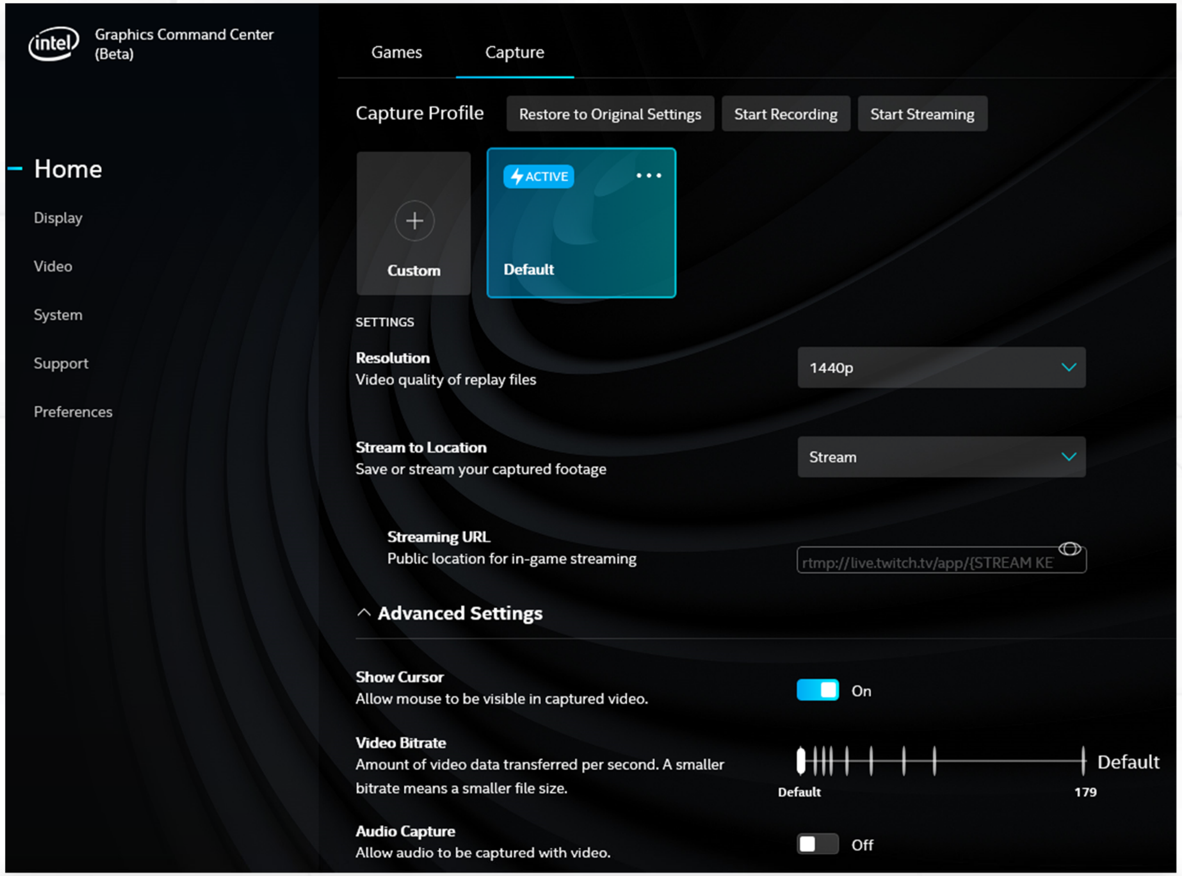
Install

END USER PC

With Instant Game Tuning

Capture and Stream

- Integrated with Intel Graphics Command Center (IGCC)
- Simple hotkeys control
- Utilizes Intel Graphics media engine
- Stream directly to Twitch and YouTube



Resolution
Video quality of replay files
1440p

Stream to Location
Save or stream your captured footage
Stream

Streaming URL
Public location for in-game streaming
rtmp://live.twitch.tv/app/{STREAM KE

Show Cursor
Allow mouse to be visible in captured video.
On

Video Bitrate
Amount of video data transferred per second. A smaller bitrate means a smaller file size.

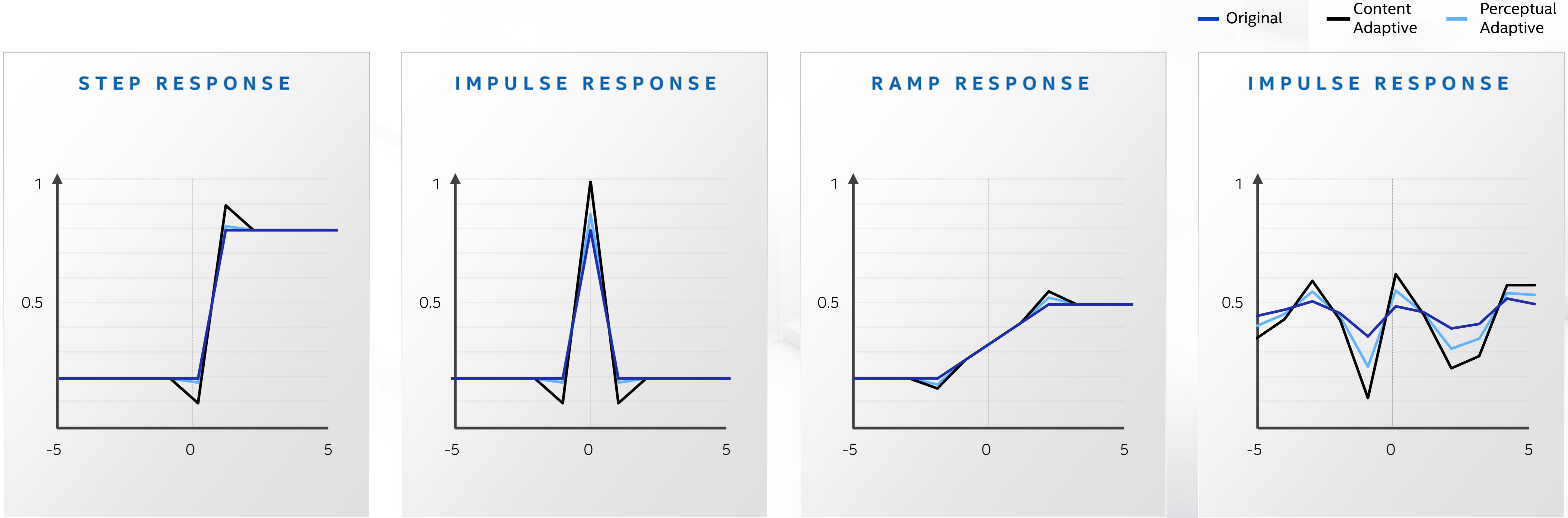
Audio Capture
Allow audio to be captured with video.
Off

Game Sharpening

- Perceptual adaptive sharpening filter
- Boosts image clarity in games
- End-users control on a per-game basis in IGCC

NEW DISCLOSURE

Game Sharpening Algorithms



Edge Enhancement without overshooting and ringing

X^e LP

Software Enablement

GRID

GEARS TACTICS

CHIVALRY II

RING OF ELYSIUM

SERIOUS SAM 4
PLANET BADASS

- Industry-wide effort to evangelize good practices
- Benefits 100s of millions of gamers worldwide

- Prepare game readiness for X^e Graphics
- Partnership with game developers to implement and optimize advanced features such as VRS

Xe LP

PRODUCTS

NEW DISCLOSURE

Xe LP PRODUCTS

TIGER LAKE

LEADERSHIP INTEGRATED GRAPHICS

DG1

GPU FOR MOBILE CREATORS

SG1

Intel® Server GPU

Xe HP

DESIGNED FOR DATA CENTER

Xe HP

DESIGNED FOR DATA CENTER

SCALE EVERYTHING

EU, FREQUENCY, BANDWIDTH, IPC, MATH

SCALE MORE

MULTI-TILE

MEDIA SUPER-COMPUTER

STREAM DENSITY, VISUAL QUALITY

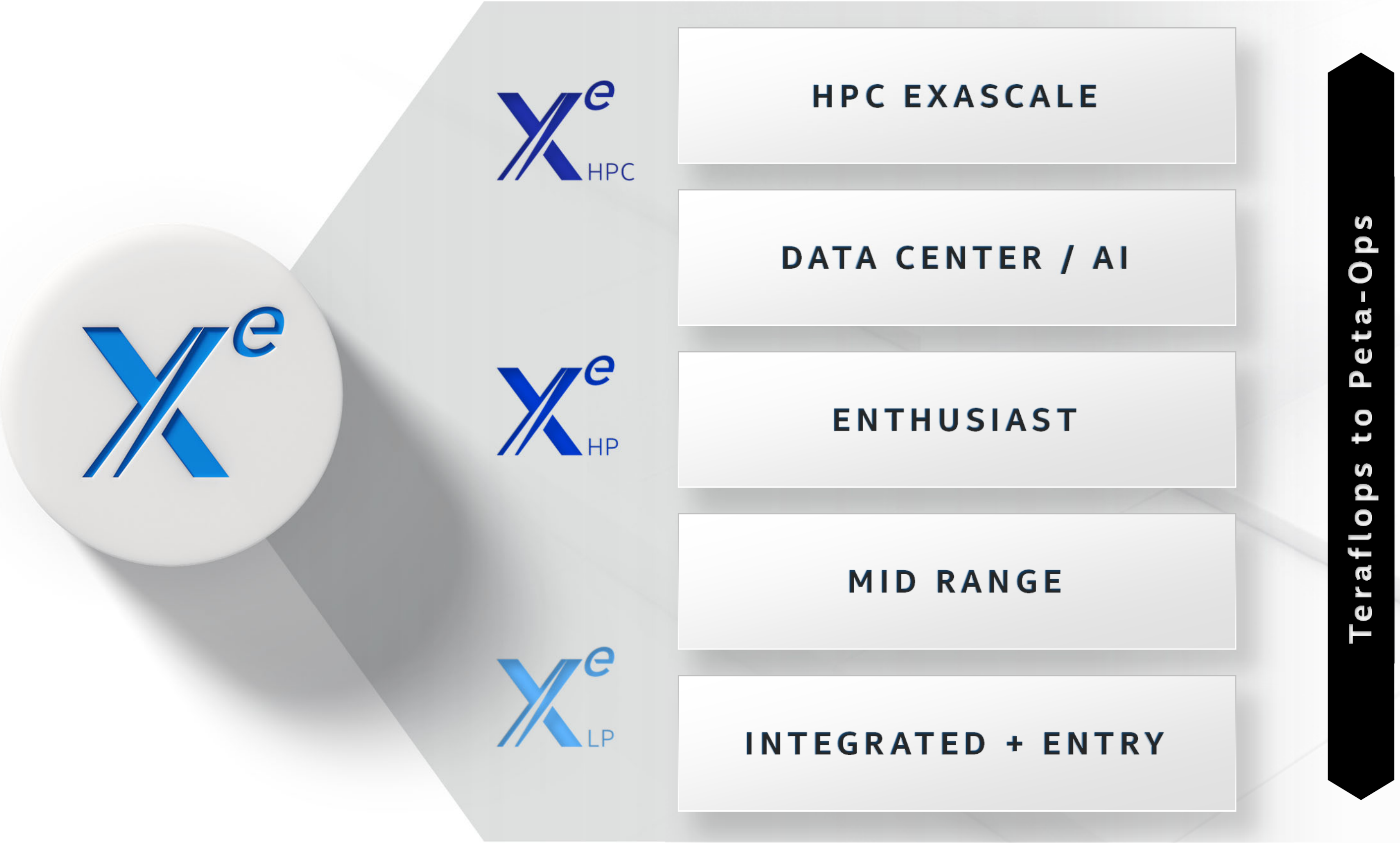
Scalability

NEW DISCLOSURE

1 Tile

2 Tile

4 Tile



X^e

X^e_{HPC}

X^e_{HP}

X^e_{LP}

HPC EXASCALE

DATA CENTER / AI

ENTHUSIAST

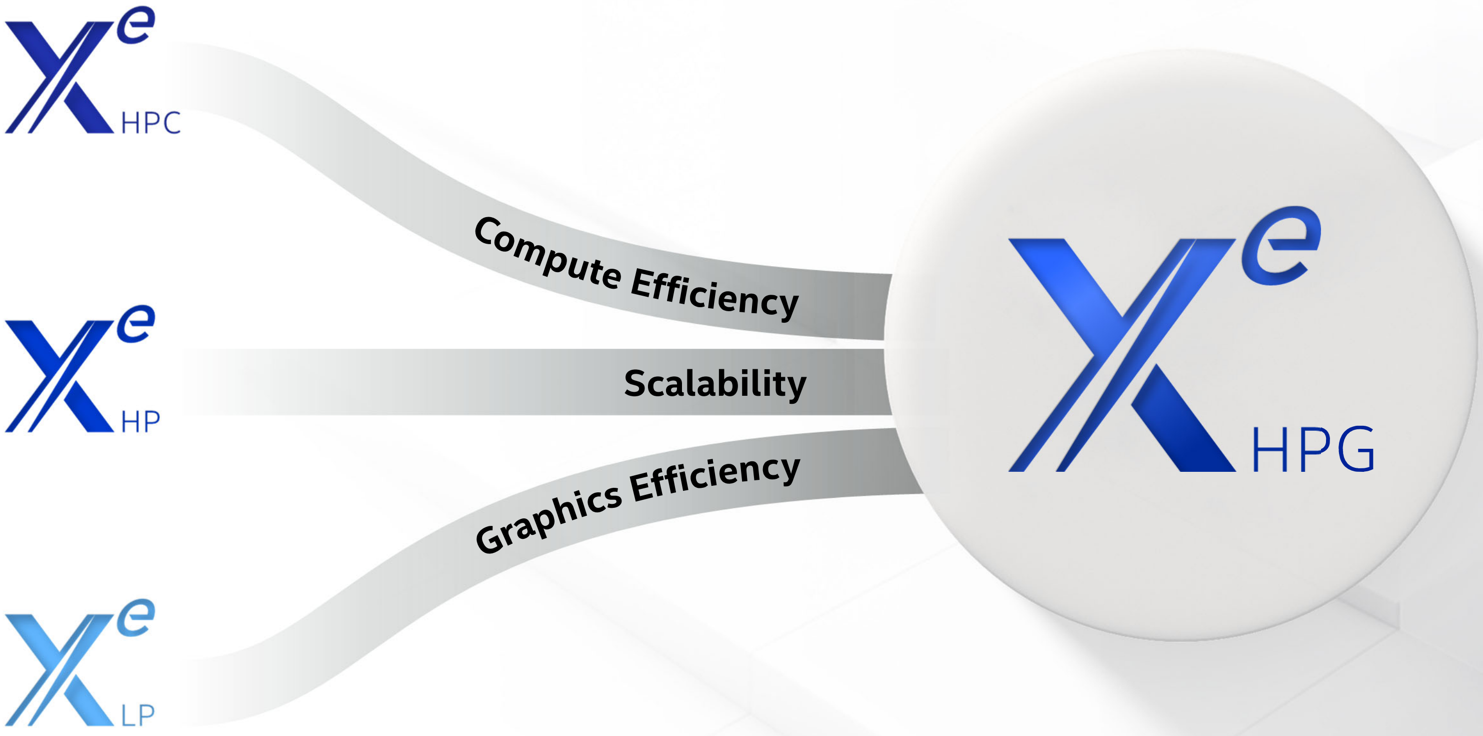
MID RANGE

INTEGRATED + ENTRY

Teraflops to Peta-Ops

What about Enthusiast Gamers?

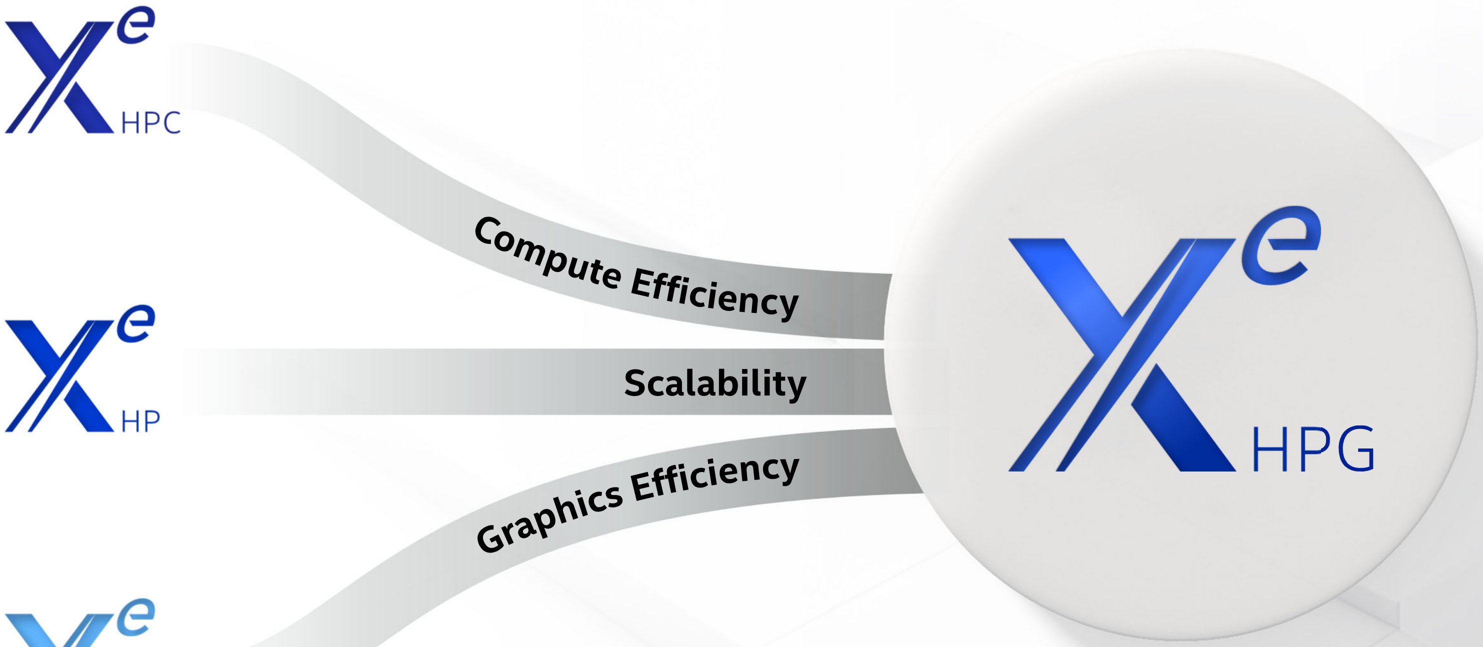
NEW DISCLOSURE



High Performance

Gaming Optimized

NEW DISCLOSURE



With Hardware Dedicated Ray-Tracing

NEW DISCLOSURE

GPU Architecture Strategy

One Architecture and **4** Micro Architectures

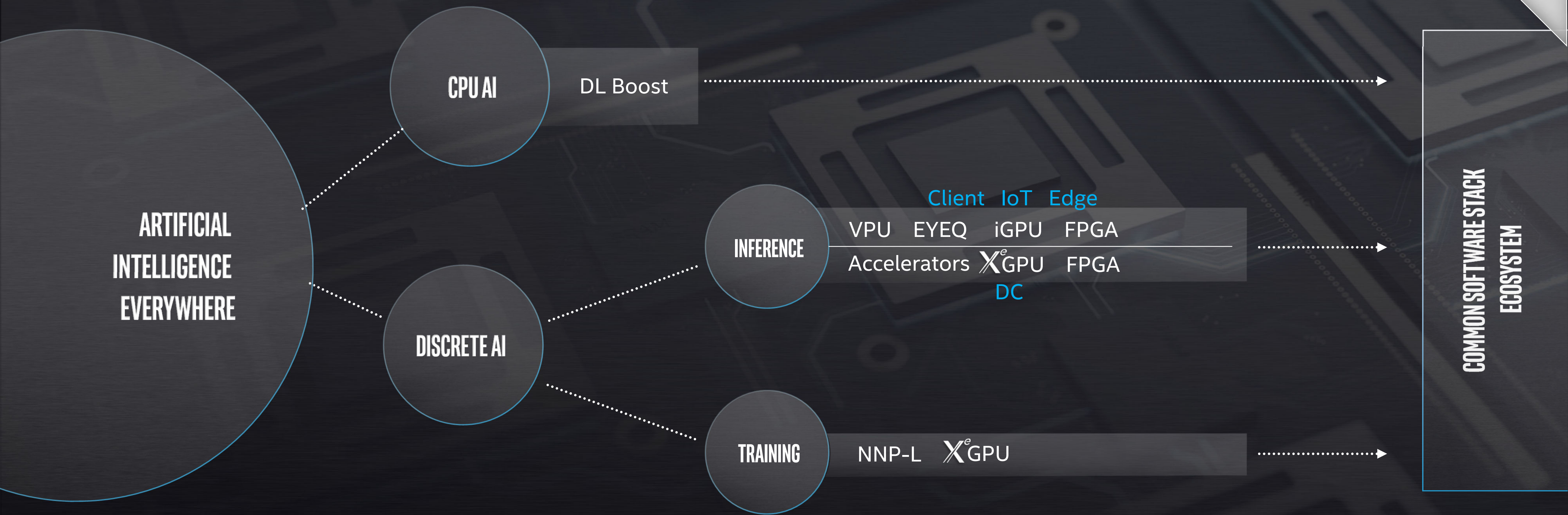
NEW DISCLOSURE

Products, Packaging and Process Overview

µArchitecture		Packaging		Process
X^e HPC	PONTE VECCHIO	FOVEROS CO-EMIB	BASE TILE	Intel 10nm SuperFin
			COMPUTE TILE	Intel Next Gen & External
			RAMBO CACHE TILE	Intel 10nm Enhanced SuperFin
			X^e LINK I/O TILE	External
X^e HP	TBA	EMIB		Intel 10nm Enhanced SuperFin
X^e HPG	TBA	STANDARD		External
X^e LP	SG1 DG1 TIGER LAKE	STANDARD		Intel 10nm SuperFin

MATRIX STRATEGY

AS WE SAID IN 2018



ARTIFICIAL INTELLIGENCE EVERYWHERE

CPU AI

DL Boost

DISCRETE AI

INFERENCE

Client IoT Edge

VPU EYEQ iGPU FPGA

Accelerators X^eGPU FPGA

DC

TRAINING

NNP-L X^eGPU

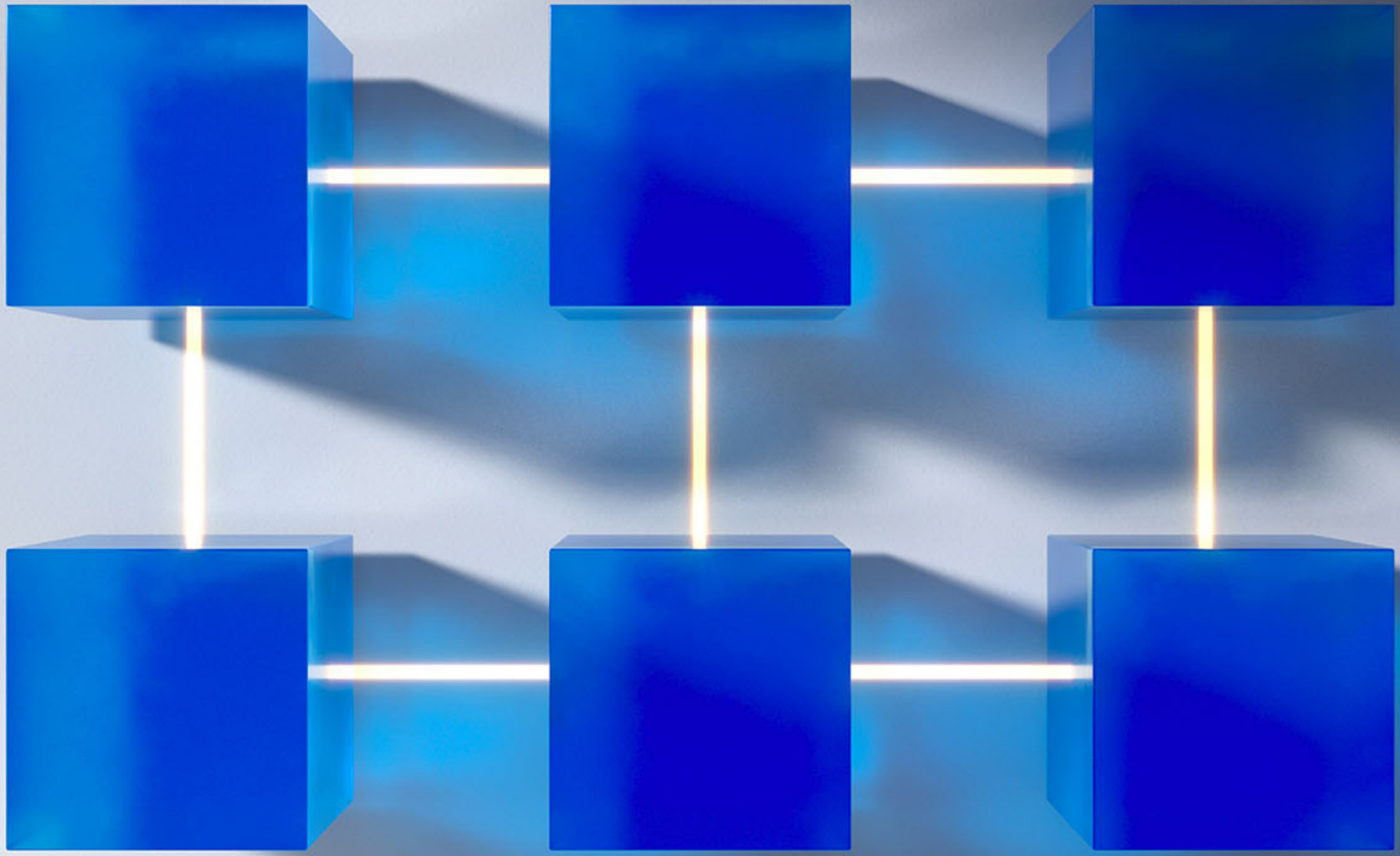
COMMON SOFTWARE STACK

ECOSYSTEM

Intel AI Portfolio

GENERAL PURPOSE ⟷ PURPOSE BUILT

BF16 - AMX	BF16 - XMX	STRATIX 10 NX	MYRIAD	GOYA
DL BOOST – VNNI – DP4A		AGILEX	MYRIAD X	GAUDI
			KEEM BAY	



Spatial Architecture

Ravi Kuppuswamy

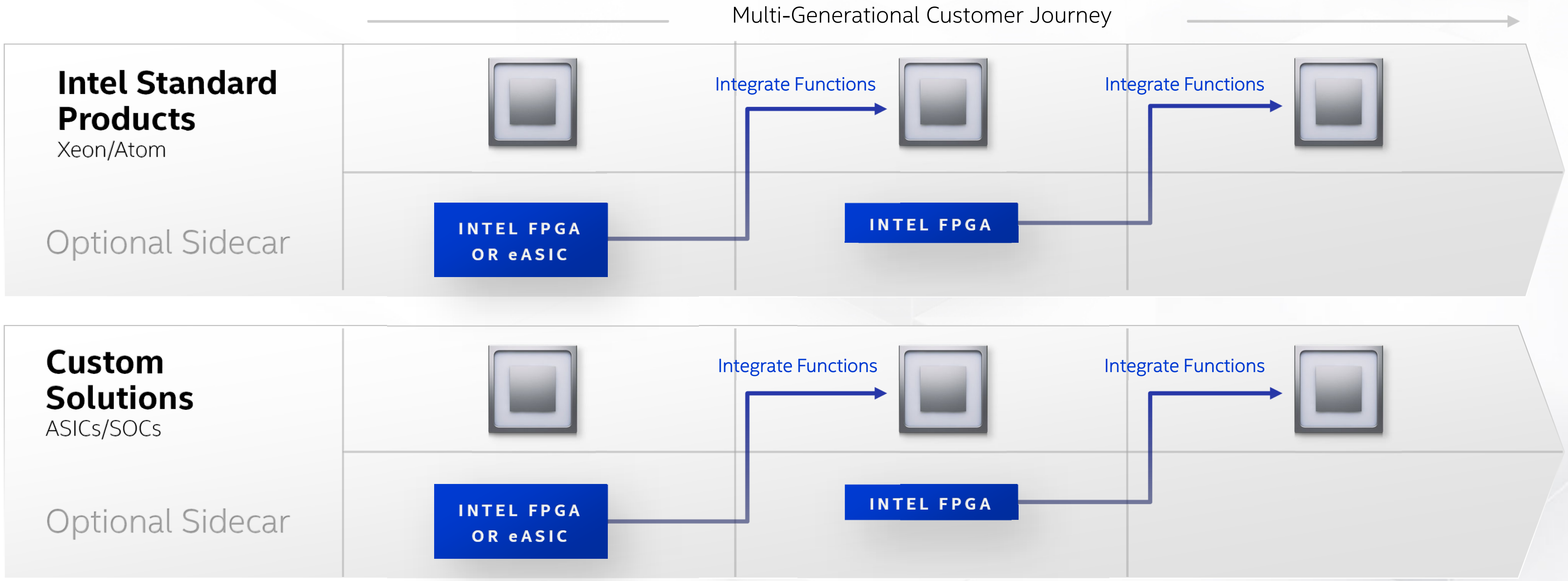
Corporate Vice President,
General Manager Custom Logic
Engineering

TECHNOLOGY
PILLARS

Architecture Day 2020

Spatial Architecture Methodology

Accelerate Key Transitions



Multi-Generational Customer Journey

Intel Standard Products
Xeon/Atom

Optional Sidecar

INTEL FPGA OR eASIC

Integrate Functions

INTEL FPGA

Integrate Functions

Example
VRAN in 5G

Custom Solutions
ASICs/SOCs

Optional Sidecar

INTEL FPGA OR eASIC

Integrate Functions

INTEL FPGA

Integrate Functions

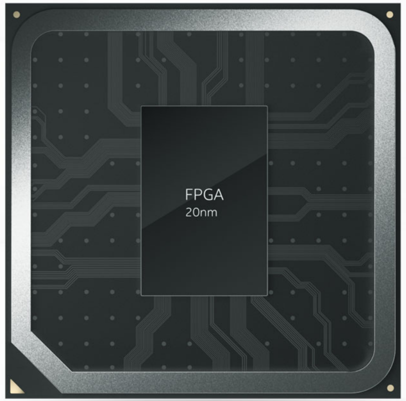
Example
SmartNIC

Repeatable model across transitions and vertical segments

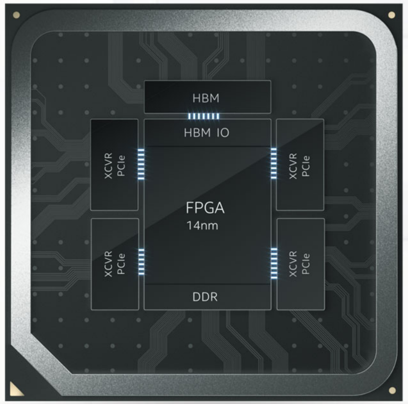
FPGA Roadmap

EMIB to Co-EMIB to Foveros

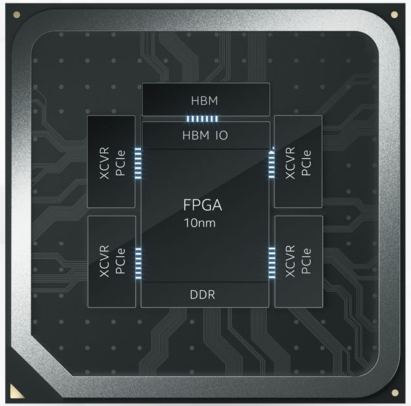
Arria®10
Production



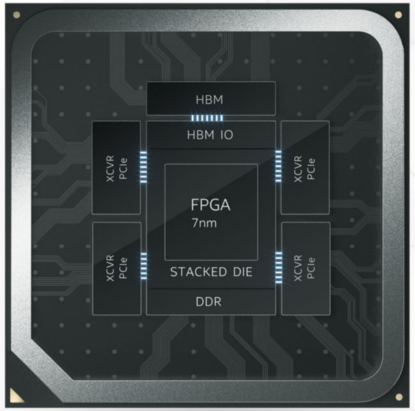
Stratix®10
Production



Agilex™
Sampling



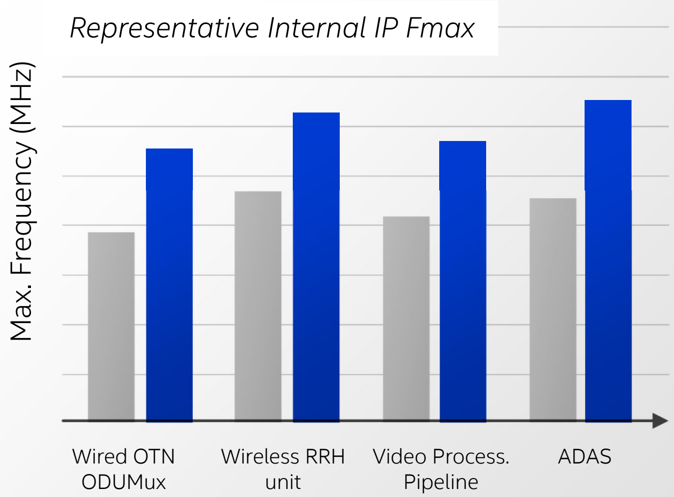
Next Gen FPGAs



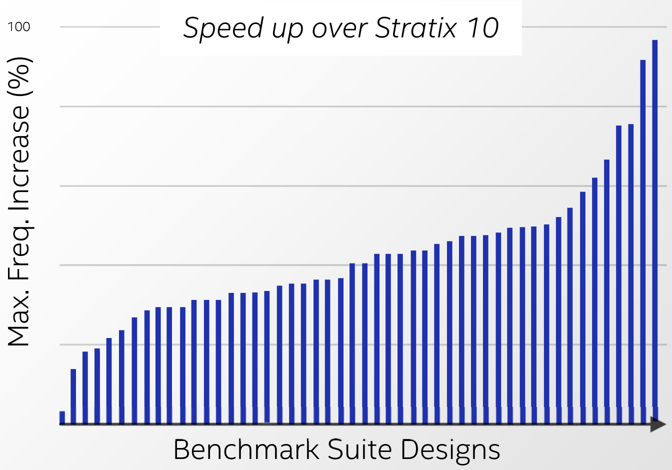
Packaging Technology	Arria®10	Stratix®10	Agilex™	Next Gen FPGAs
Monolithic	●			
EMIB (2.5D)		●	● **2nd Gen**	
Co-EMIB / Foveros (3D)				●

Agilex™ Performance/Power

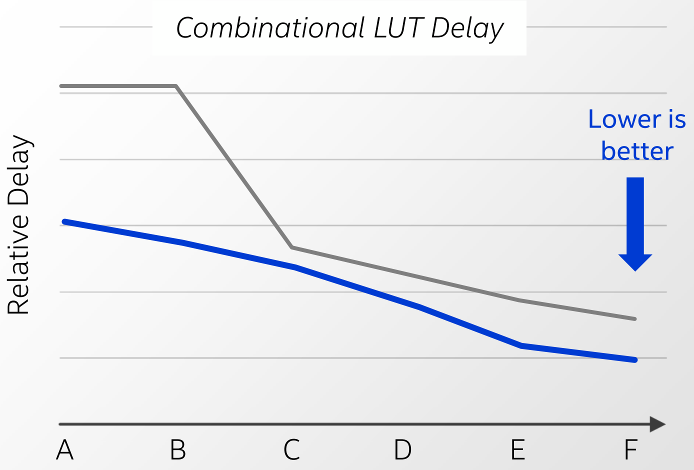
Based on Intel® 10nm Process



Agilex™ **40+% Geomean Max. Frequency**
Improvement over Stratix®10



Agilex™ **Significant Speed Up**
In Data Center, Network & Edge

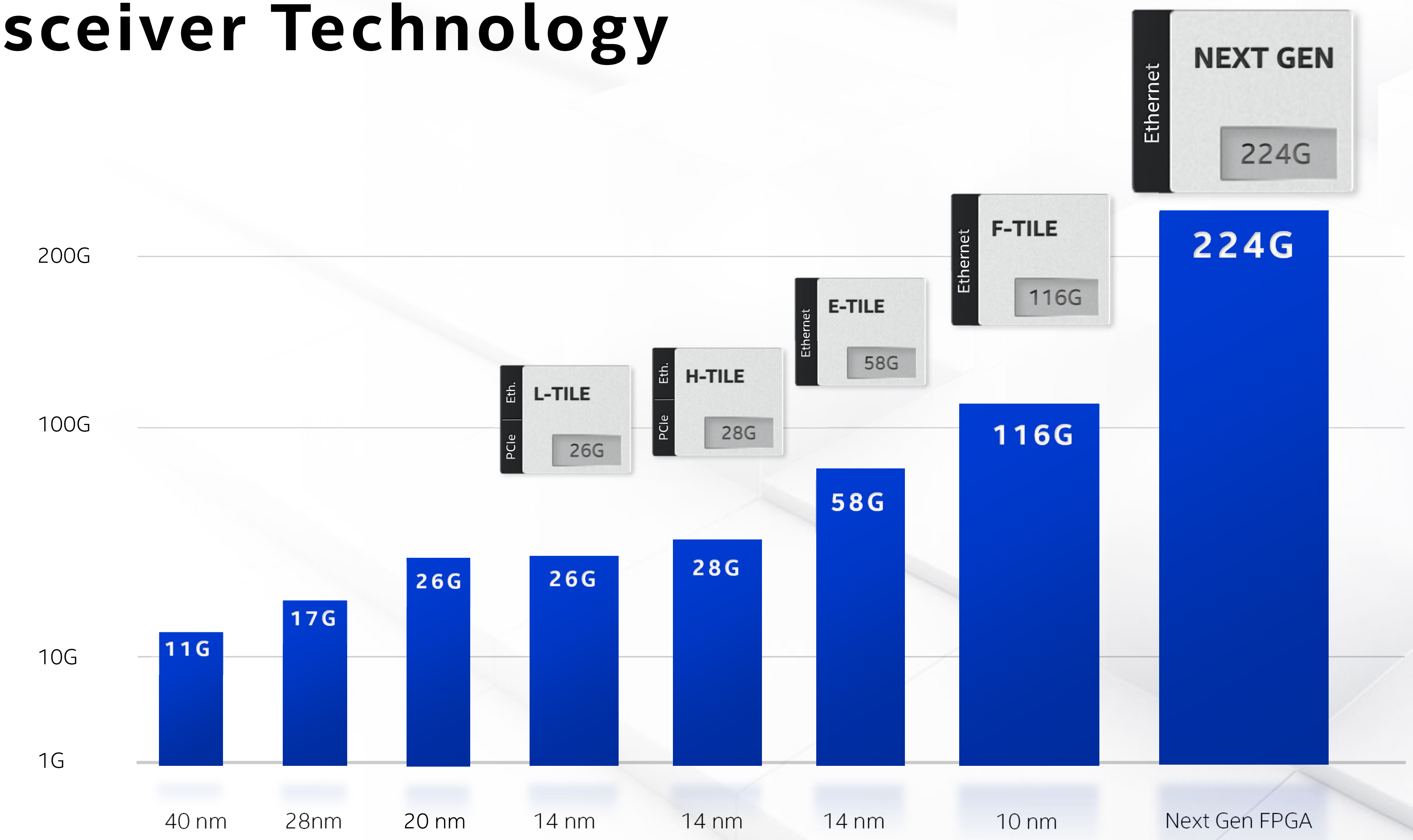


Agilex™ **Average 1.5x faster**
in Combinational LUT Delay

Intel® Agilex™ FPGAs Deliver Significantly Better Performance/Watt

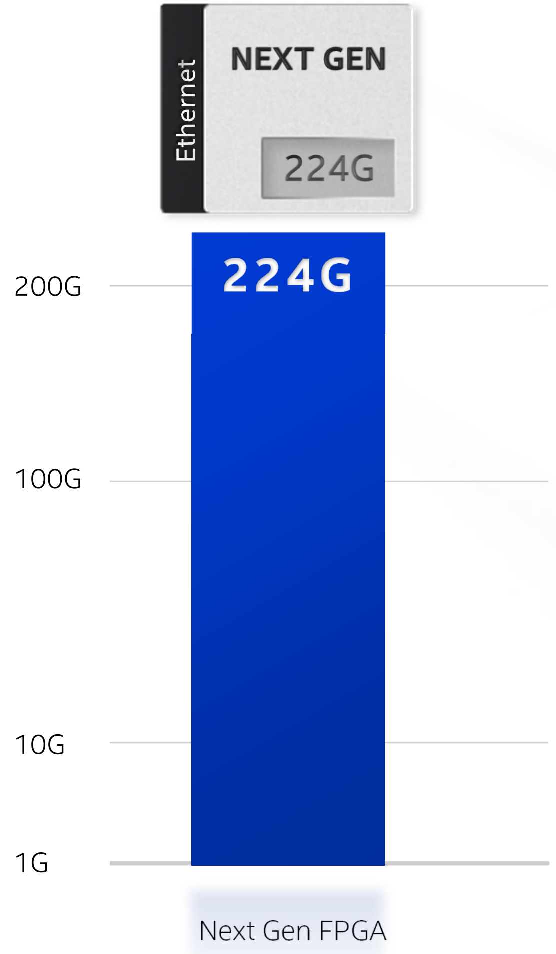
NEW DISCLOSURE

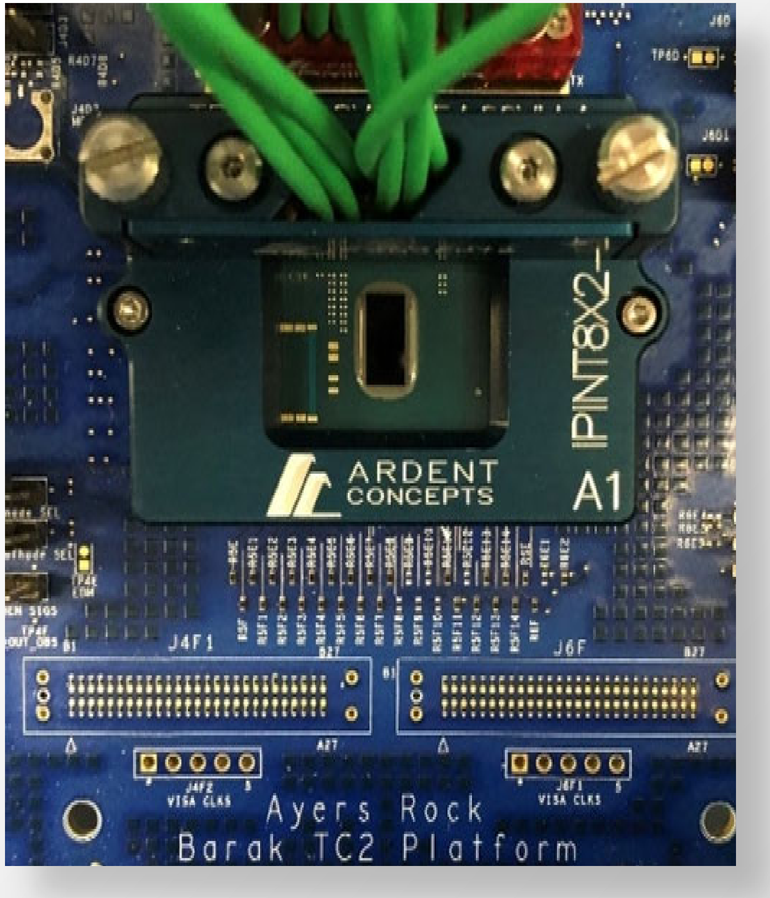
Transceiver Technology



NEW DISCLOSURE

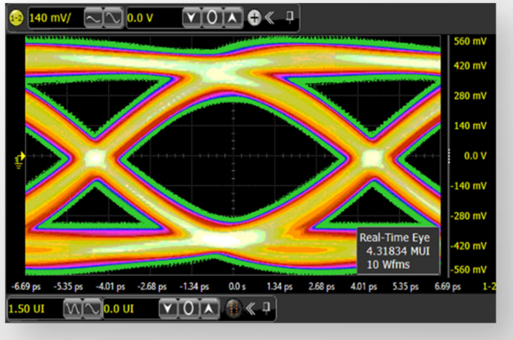
Transceiver Technology





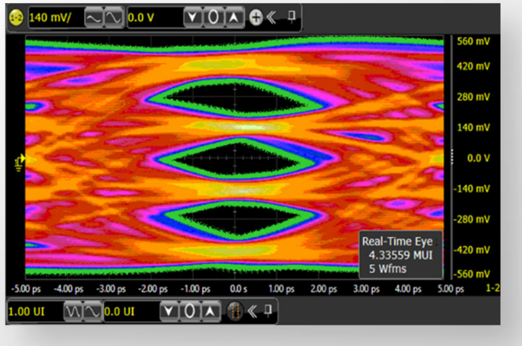
112G

NRZ



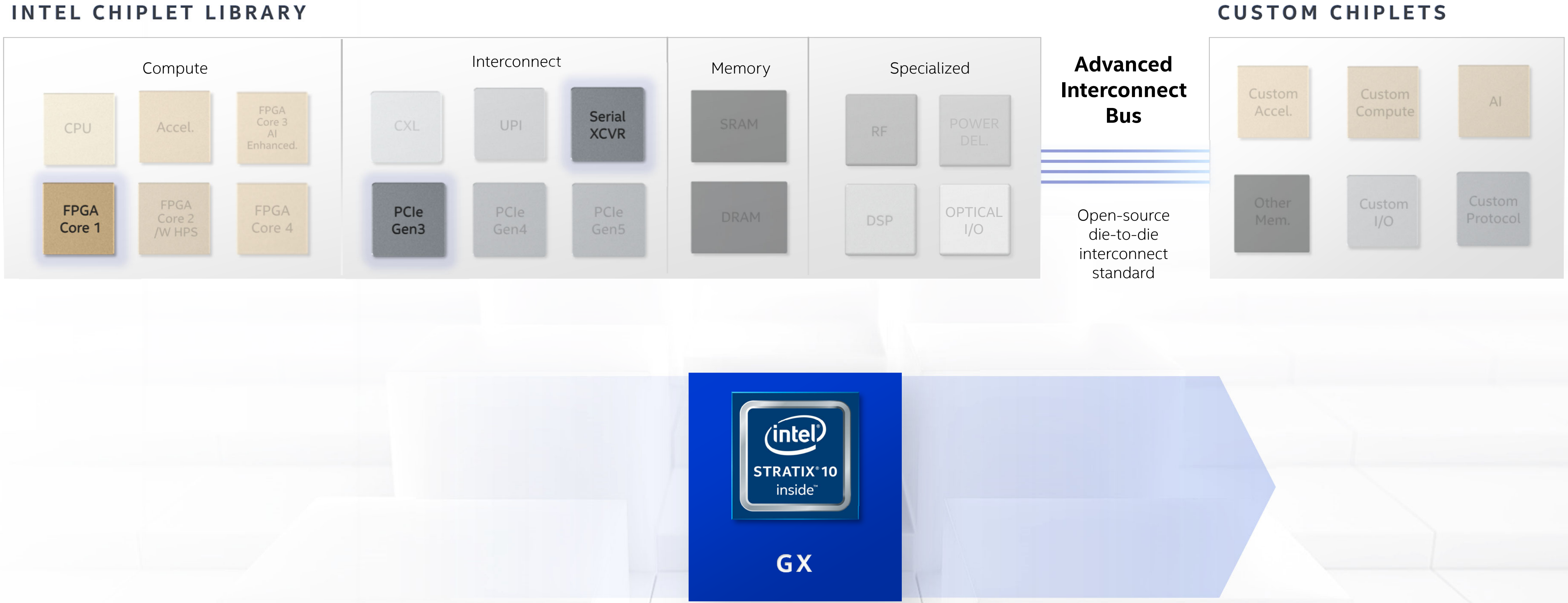
224G

PAM4



Build Advanced Spatial Architectures

Heterogenous Integration of Interoperable Chiplets



INTEL CHIPLET LIBRARY

Compute

- CPU
- Accel.
- FPGA Core 3 AI Enhanced.
- FPGA Core 1
- FPGA Core 2 /W HPS
- FPGA Core 4

Interconnect

- CXL
- UPI
- Serial XCVR
- PCIe Gen3
- PCIe Gen4
- PCIe Gen5

Memory

- SRAM
- DRAM

Specialized

- RF
- POWER DEL.
- DSP
- OPTICAL I/O

Advanced Interconnect Bus

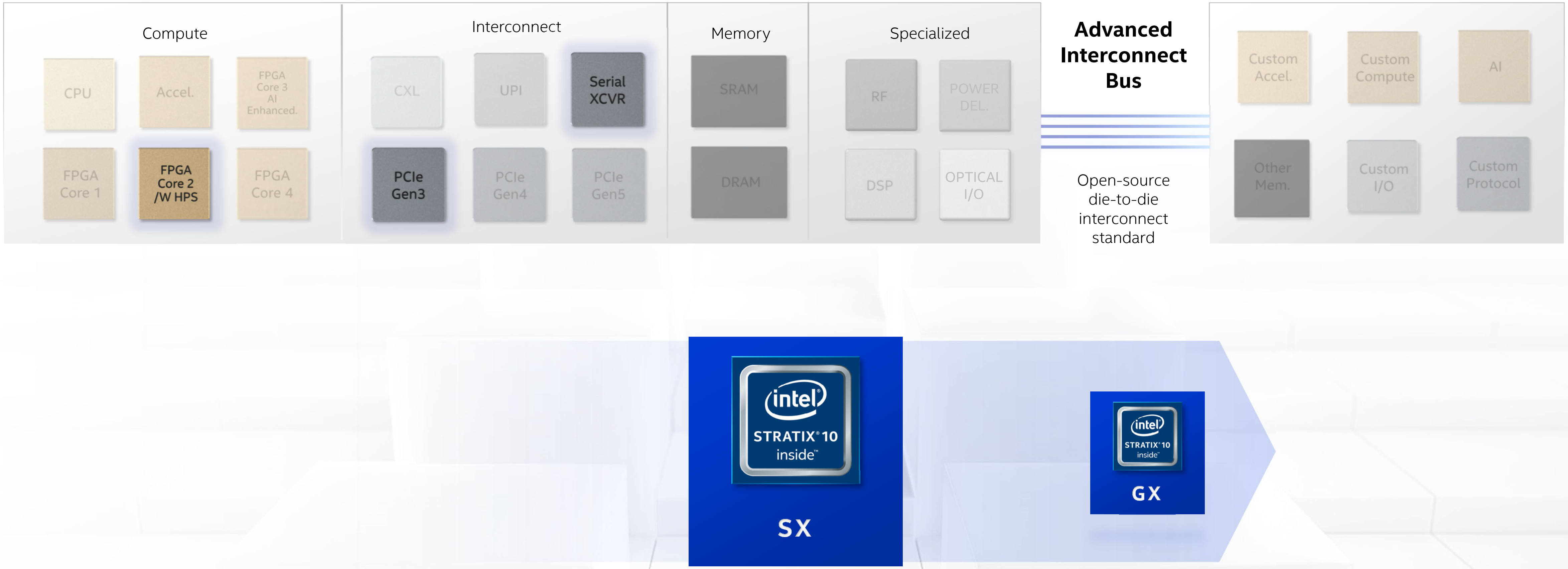
Open-source die-to-die interconnect standard

CUSTOM CHIPLETS

- Custom Accel.
- Custom Compute
- AI
- Other Mem.
- Custom I/O
- Custom Protocol

Build Advanced Spatial Architectures

Heterogenous Integration of Interoperable Chiplets



INTEL CHIPLET LIBRARY

Compute
- CPU
- Accel.
- FPGA Core 3 AI Enhanced.
- FPGA Core 1
- **FPGA Core 2 /W HPS**
- FPGA Core 4

Interconnect
- CXL
- UPI
- **Serial XCVR**
- **PCIe Gen3**
- PCIe Gen4
- PCIe Gen5

Memory
- SRAM
- DRAM

Specialized
- RF
- POWER DEL.
- DSP
- OPTICAL I/O

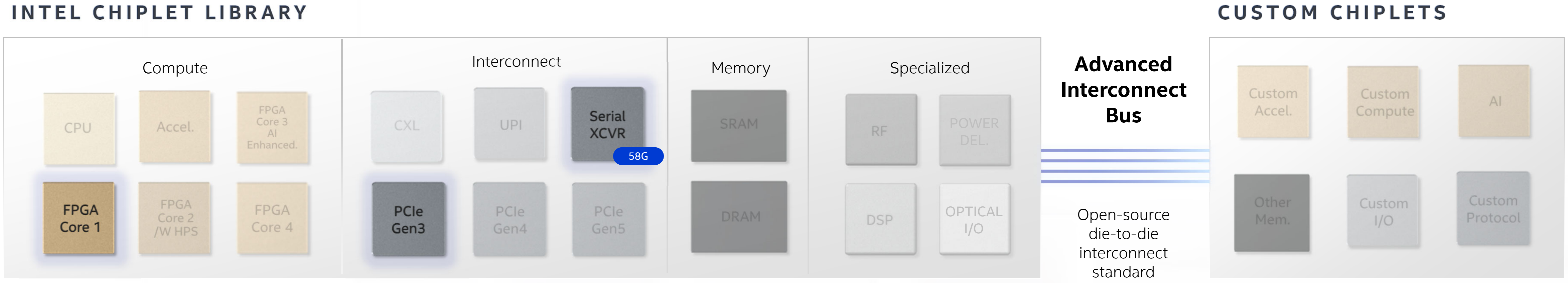
Advanced Interconnect Bus

Open-source die-to-die interconnect standard

CUSTOM CHIPLETS
- Custom Accel.
- Custom Compute
- AI
- Other Mem.
- Custom I/O
- Custom Protocol

Build Advanced Spatial Architectures

Heterogenous Integration of Interoperable Chiplets



INTEL CHIPLET LIBRARY

Compute

- CPU
- Accel.
- FPGA Core 3 AI Enhanced.
- **FPGA Core 1**
- FPGA Core 2 /W HPS
- FPGA Core 4

Interconnect

- CXL
- UPI
- **Serial XCVR** 58G
- **PCIe Gen3**
- PCIe Gen4
- PCIe Gen5

Memory

- SRAM
- DRAM

Specialized

- RF
- POWER DEL.
- DSP
- OPTICAL I/O

Advanced Interconnect Bus

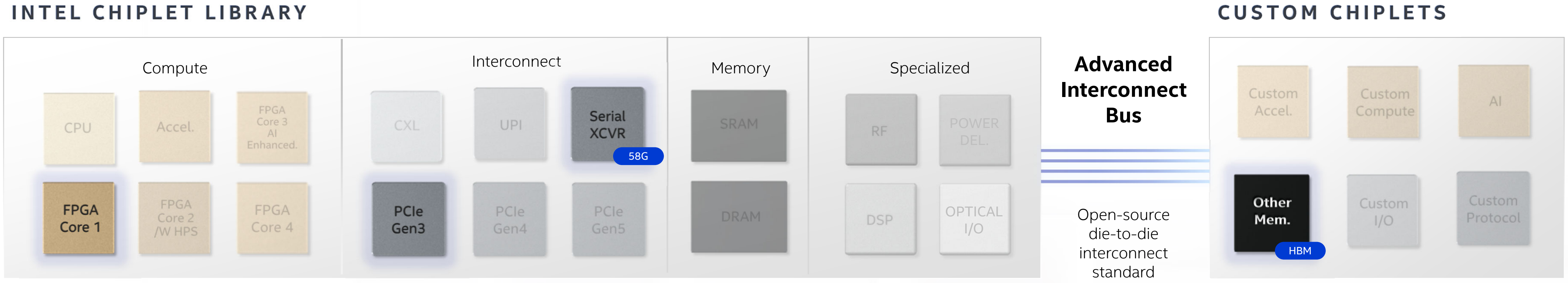
Open-source die-to-die interconnect standard

CUSTOM CHIPLETS

- Custom Accel.
- Custom Compute
- AI
- Other Mem.
- Custom I/O
- Custom Protocol

Build Advanced Spatial Architectures

Heterogenous Integration of Interoperable Chiplets



INTEL CHIPLET LIBRARY

Compute
CPU
Accel.
FPGA Core 3 AI Enhanced.
FPGA Core 1
FPGA Core 2 /W HPS
FPGA Core 4

Interconnect
CXL
UPI
Serial XCVR
58G
PCIe Gen3
PCIe Gen4
PCIe Gen5

Memory
SRAM
DRAM

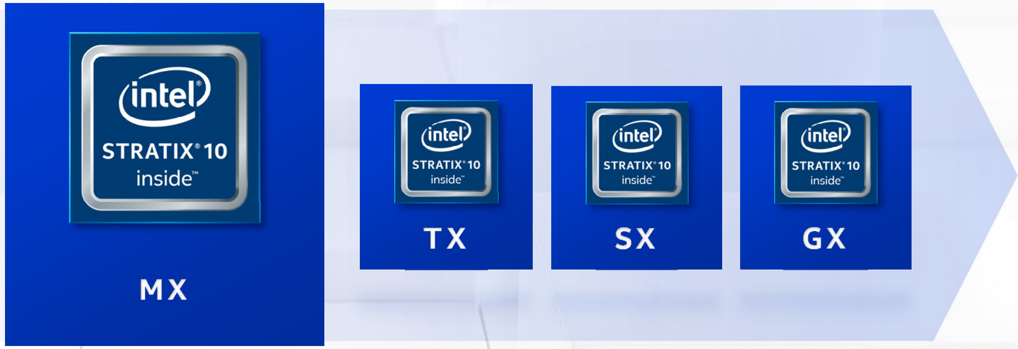
Specialized
RF
POWER DEL.
DSP
OPTICAL I/O

Advanced Interconnect Bus

Open-source die-to-die interconnect standard

CUSTOM CHIPLETS

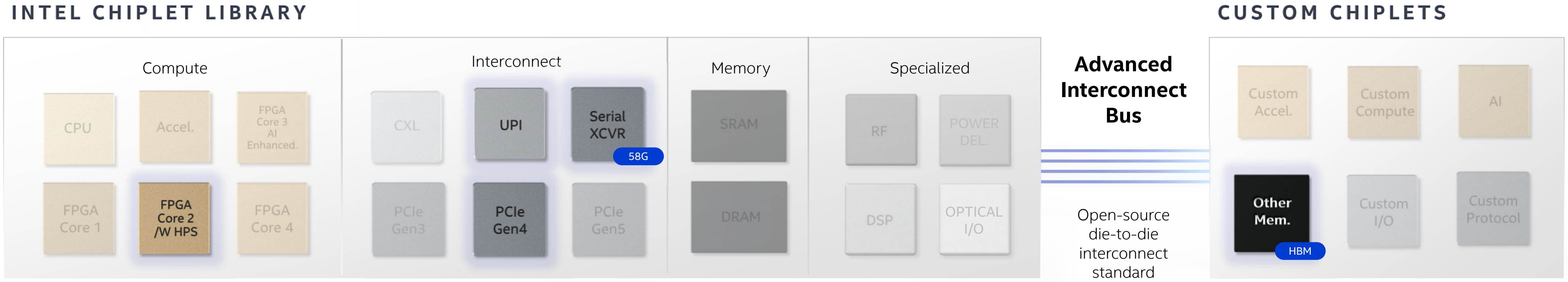
Custom Accel.
Custom Compute
AI
Other Mem.
HBM
Custom I/O
Custom Protocol



* Optional

Build Advanced Spatial Architectures

Heterogenous Integration of Interoperable Chiplets



INTEL CHIPLET LIBRARY

Compute
CPU
Accel.
FPGA Core 3 AI Enhanced.
FPGA Core 1
FPGA Core 2 /W HPS
FPGA Core 4

Interconnect
CXL
UPI
Serial XCVR
58G
PCIe Gen3
PCIe Gen4
PCIe Gen5

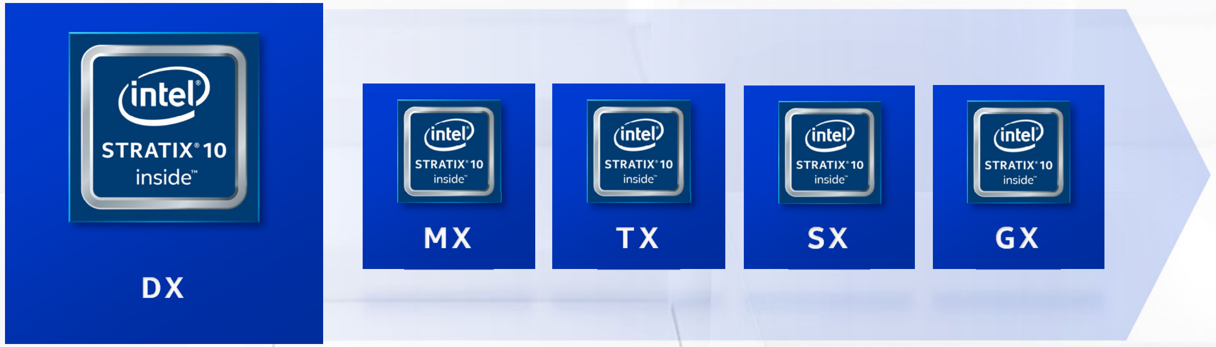
Memory
SRAM
DRAM

Specialized
RF
POWER DEL.
DSP
OPTICAL I/O

Advanced Interconnect Bus

Open-source die-to-die interconnect standard

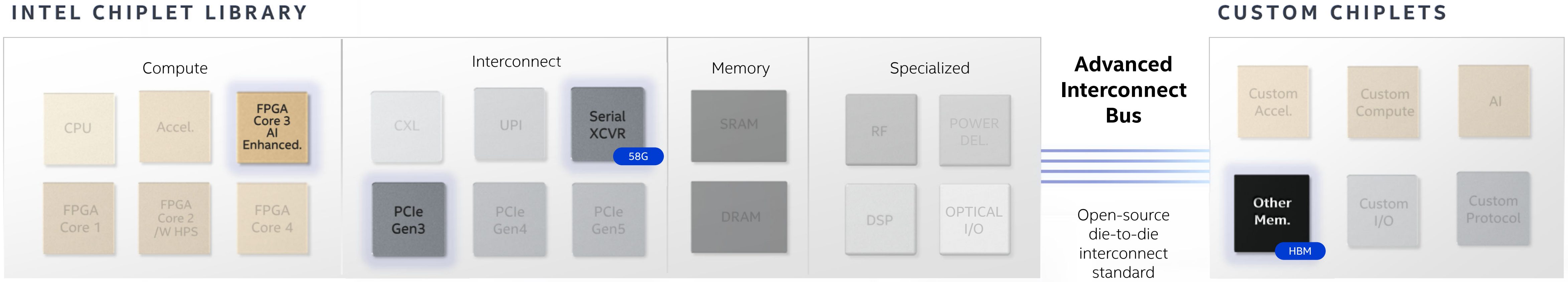
CUSTOM CHIPLETS
Custom Accel.
Custom Compute
AI
Other Mem.
HBM
Custom I/O
Custom Protocol

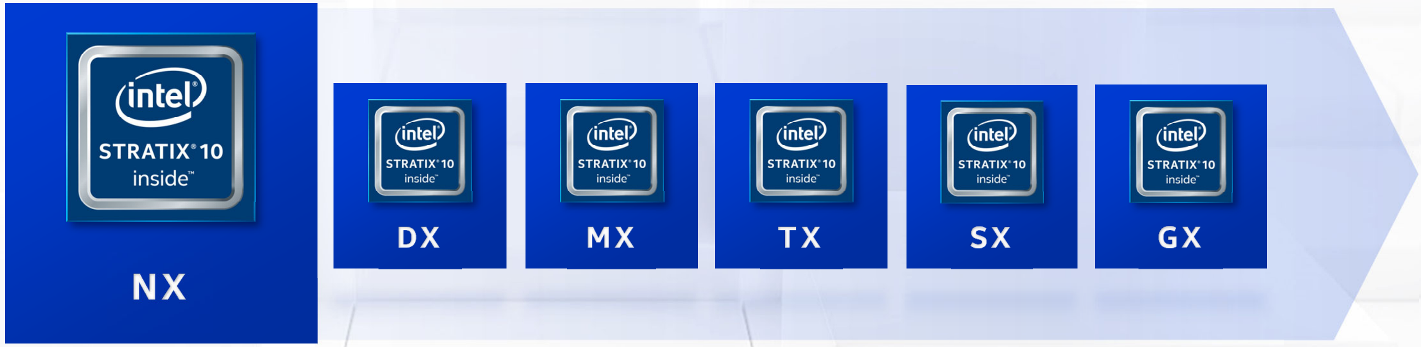


* Optional

Build Advanced Spatial Architectures

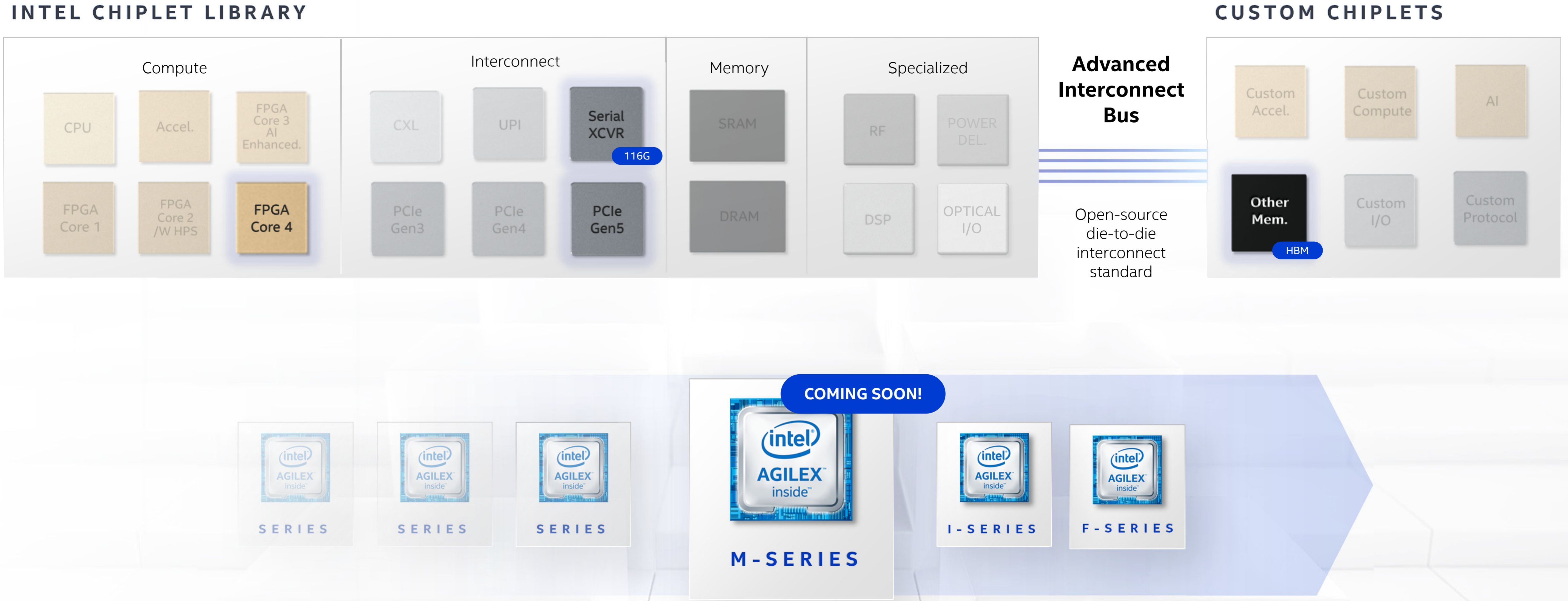
Heterogenous Integration of Interoperable Chiplets





Build Advanced Spatial Architectures

Heterogenous Integration of Interoperable Chiplets



INTEL CHIPLET LIBRARY

Compute
- CPU
- Accel.
- FPGA Core 3 AI Enhanced.
- FPGA Core 1
- FPGA Core 2 /W HPS
- **FPGA Core 4**

Interconnect
- CXL
- UPI
- **Serial XCVR** 116G
- PCIe Gen3
- PCIe Gen4
- **PCIe Gen5**

Memory
- SRAM
- DRAM

Specialized
- RF
- POWER DEL.
- DSP
- OPTICAL I/O

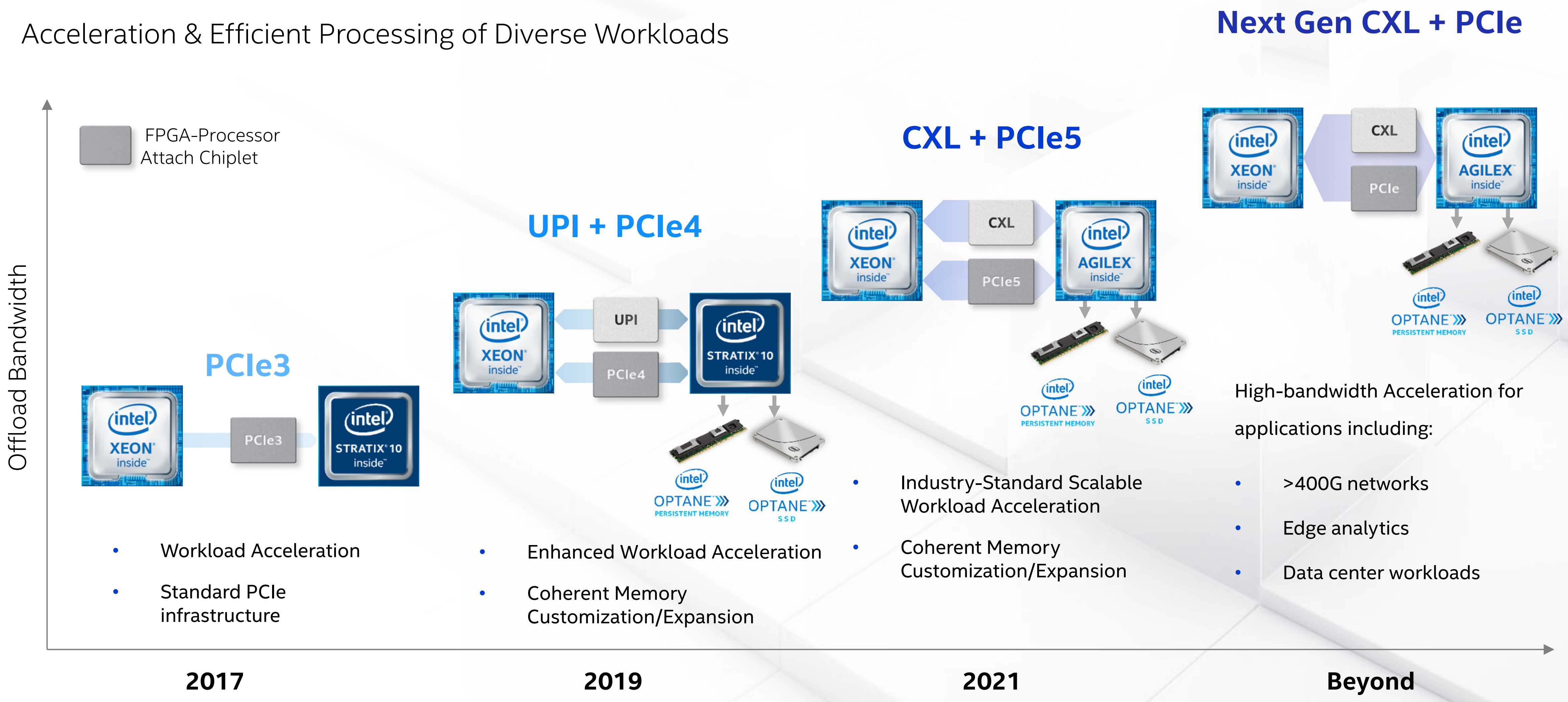
Advanced Interconnect Bus

Open-source die-to-die interconnect standard

CUSTOM CHIPLETS
- Custom Accel.
- Custom Compute
- AI
- **Other Mem.** HBM
- Custom I/O
- Custom Protocol

FPGA-Processor Attach Chiplets

Acceleration & Efficient Processing of Diverse Workloads



2017 – PCIe3

- Workload Acceleration
- Standard PCIe infrastructure

2019 – UPI + PCIe4

- Enhanced Workload Acceleration
- Coherent Memory Customization/Expansion

2021 – CXL + PCIe5

- Industry-Standard Scalable Workload Acceleration
- Coherent Memory Customization/Expansion

Beyond – Next Gen CXL + PCIe

High-bandwidth Acceleration for applications including:

- >400G networks
- Edge analytics
- Data center workloads

Intel® Stratix® 10 NX FPGA

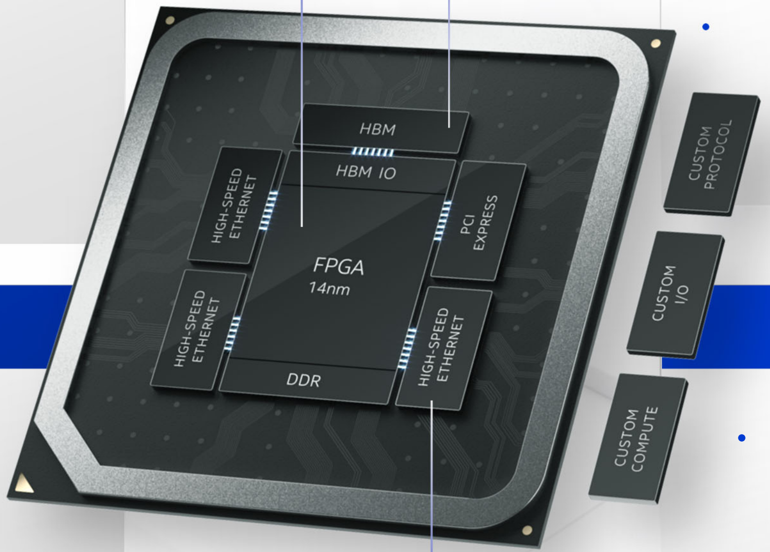
Intel's first AI-optimized FPGA

HIGH PERFORMANCE AI MATRIX BLOCKS

- Up to 15X more INT8 compute performance than today's Stratix 10 MX for AI workloads
- Hardware programmable for AI with customized workloads

HIGH BANDWIDTH NETWORKING

- Up to 57.8G PAM4 transceivers and hard Intel Ethernet blocks for high efficiency
- Flexible and customizable interconnect to scale across multiple nodes



ABUNDANT NEAR-COMPUTE MEMORY

- Embedded and customizable memory hierarchy for model persistence
- Integrated HBM for high memory bandwidth

EXTENSIBLE

- Chiplets enable easier interface customization and ASIC extensions

Matrix Compute, Memory & Networking delivers high performance HW optimized for AI

Architecture Updates Today

SCALAR

- **TIGER LAKE**
  WILLOW COVE
- **ALDER LAKE**
  PERFORMANCE HYBRID

VECTOR

- X^e LP
- X^e HPG
- X^e HP
- X^e HPC

MATRIX

- BF16 EVERYWHERE
- XMX & DP4a FOR X^e
- AMX FOR CPU'S

SPATIAL

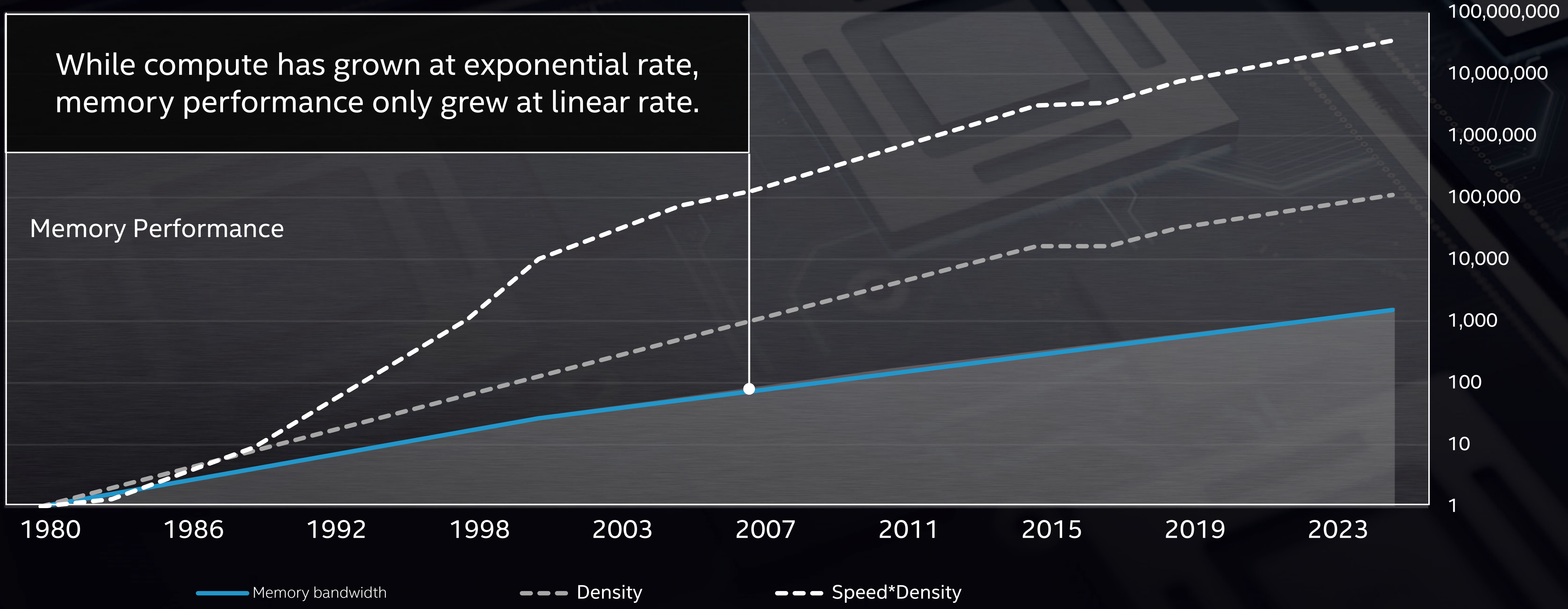
- AGILEX
- STRATIX 10 NX
- 224G TRANSCEIVER


TECHNOLOGY PILLARS
SOFTWARE
SECURITY
INTERCONNECT
MEMORY
XPU ARCHITECTURE
PROCESS & PACKAGING
intel

MEMORY PILLAR

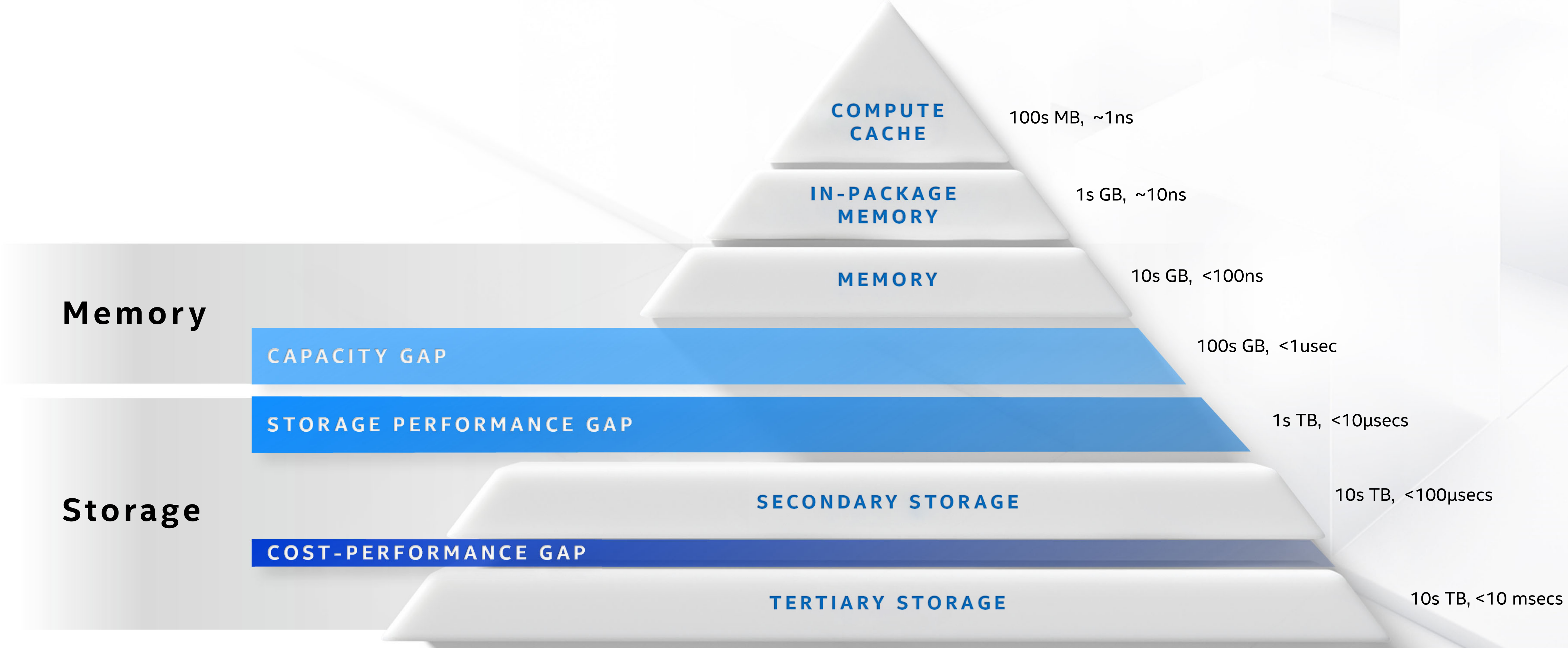
WHAT WE SAID IN 2018

Exponential advances in all levels of memory hierarchy are needed to match the ever increasing compute demand


While compute has grown at exponential rate, memory performance only grew at linear rate.

2018 ARCHITECTURE DAY

Memory and Storage Hierarchy Gaps



Memory

COMPUTE CACHE — 100s MB, ~1ns

IN-PACKAGE MEMORY — 1s GB, ~10ns

MEMORY — 10s GB, <100ns

CAPACITY GAP — 100s GB, <1usec

Storage

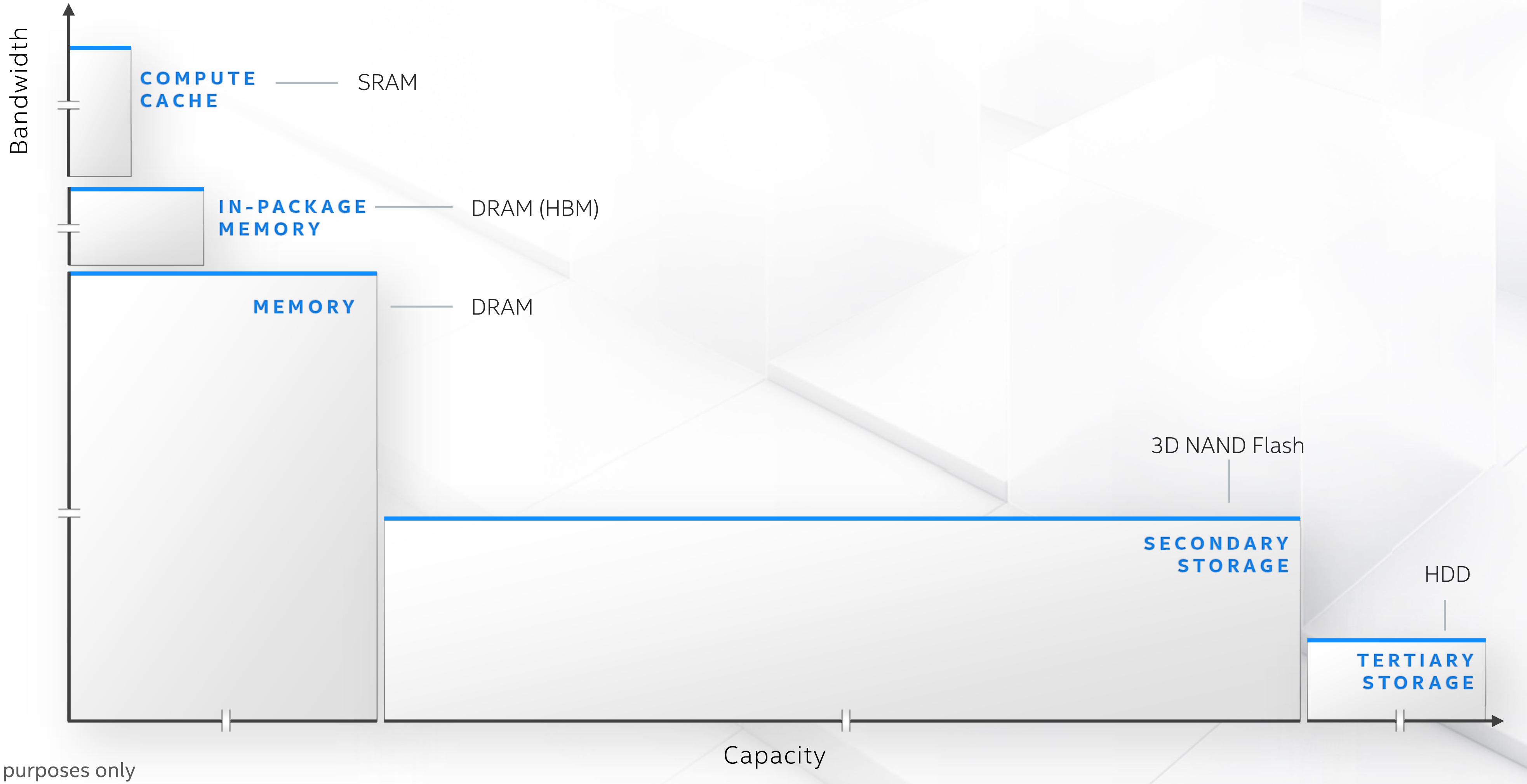
STORAGE PERFORMANCE GAP — 1s TB, <10µsecs

SECONDARY STORAGE — 10s TB, <100µsecs

COST-PERFORMANCE GAP

TERTIARY STORAGE — 10s TB, <10 msecs

Memory and Storage Hierarchy Gaps

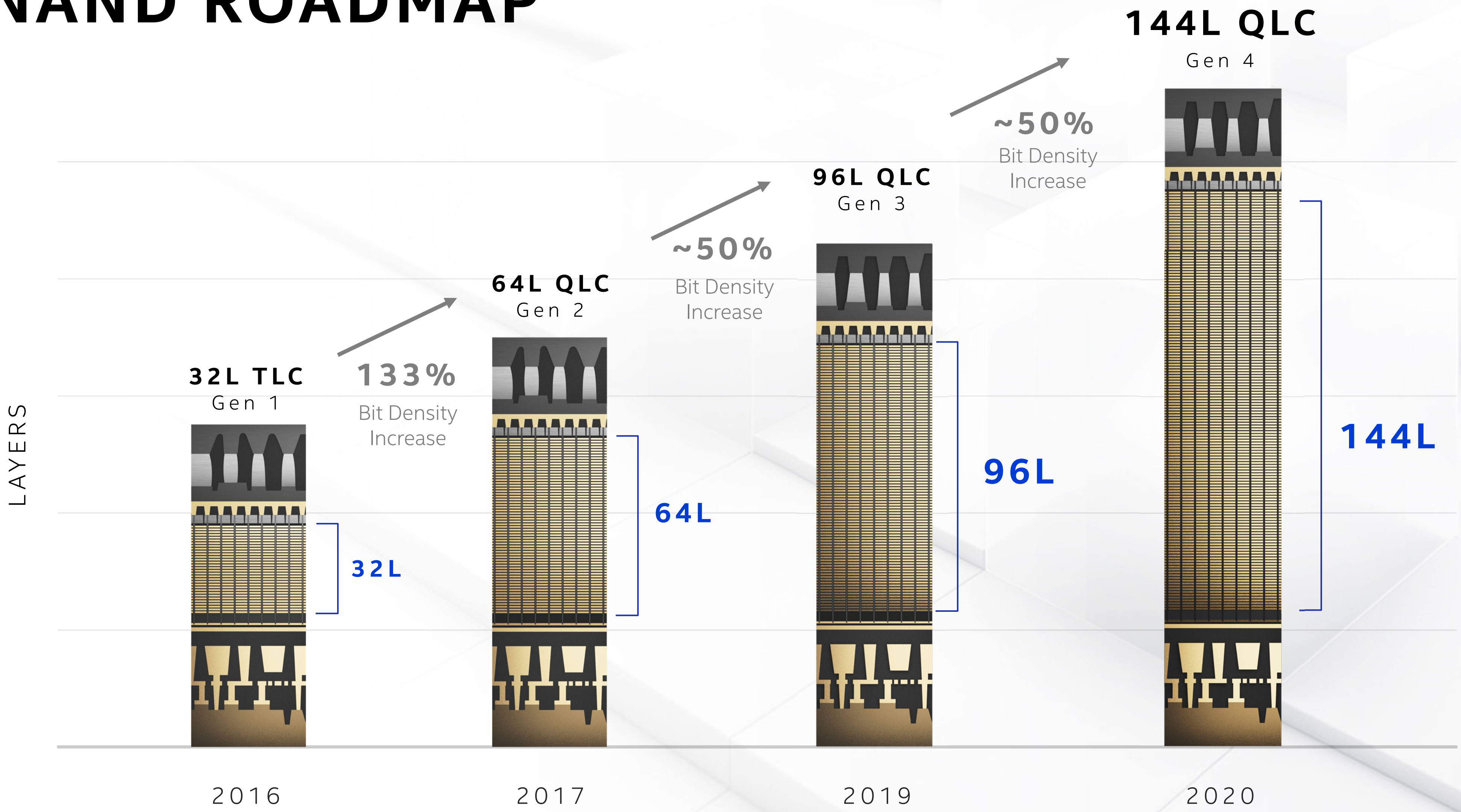


For illustrative purposes only

3D NAND ROADMAP

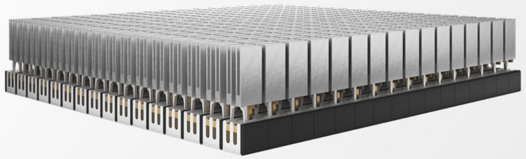
expected

3D NAND ROADMAP

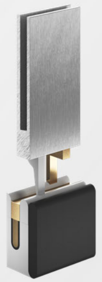


expected

Types of Memory



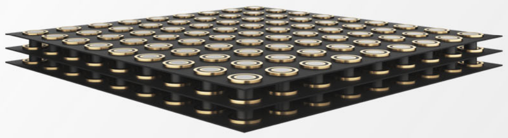
DRAM



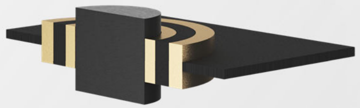
One DRAM memory cell = 1 bit

3D XPOINT

One 3D XPoint memory cell = 1 bit



INTEL 3D NAND



One 3D NAND memory cell = 1-4 bits

Types of Memory Compared

DRAM

One DRAM memory cell = 1 bit

3D XPOINT

One 3D XPoint memory cell = 1 bit

INTEL 3D NAND

One 3D NAND memory cell = 1-4 bits

3D XPoint Memory Roadmap

1st Gen

2-Deck

2nd Gen

4-Deck

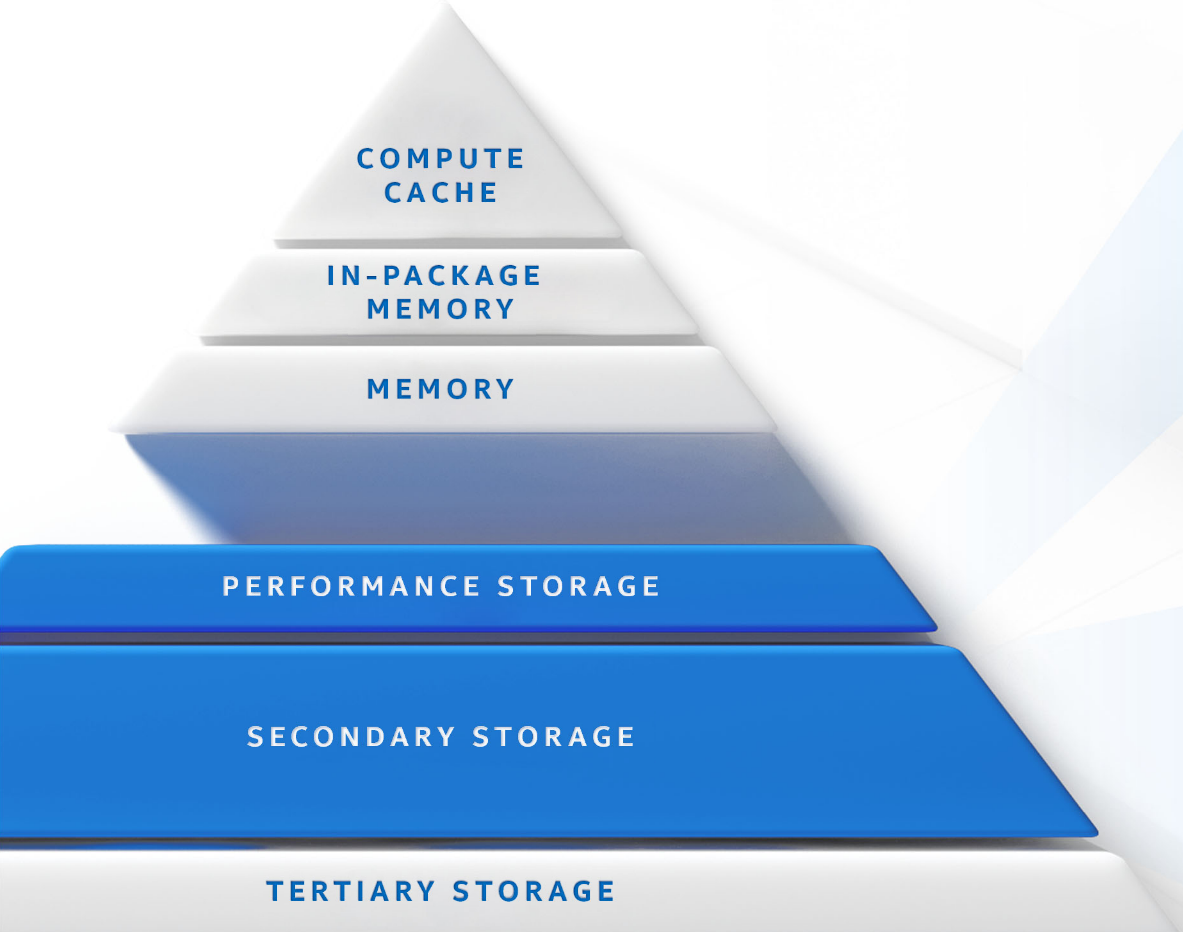
Multiple Millions of IOPS

on 2nd Generation Intel® Optane™ SSD

2017 → 2020*

*Target Production

Intel Memory Hierarchy



COMPUTE CACHE

IN-PACKAGE MEMORY

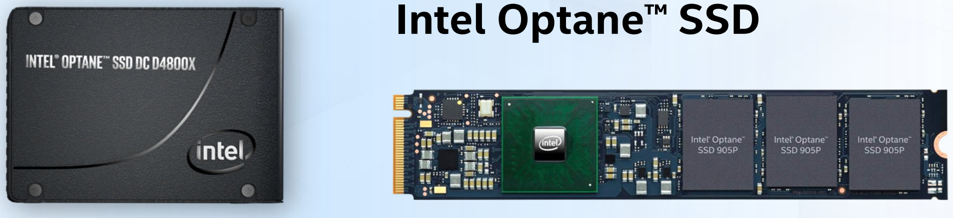
MEMORY

PERFORMANCE STORAGE

SECONDARY STORAGE

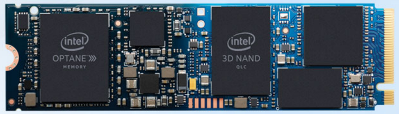
TERTIARY STORAGE

Intel Optane™ SSD



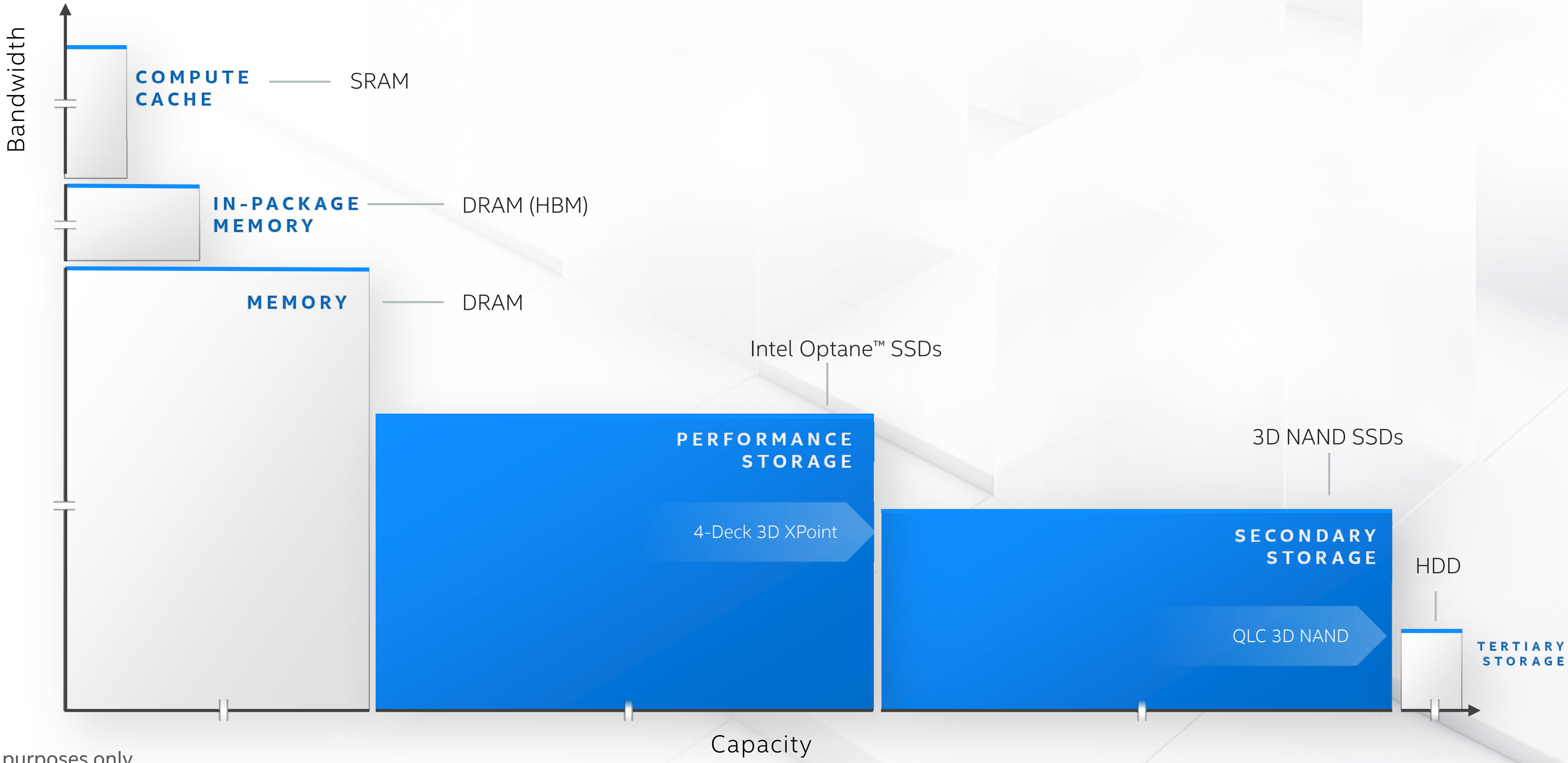
Intel Optane ™ Memory H10

Optane + 3D QLC NAND



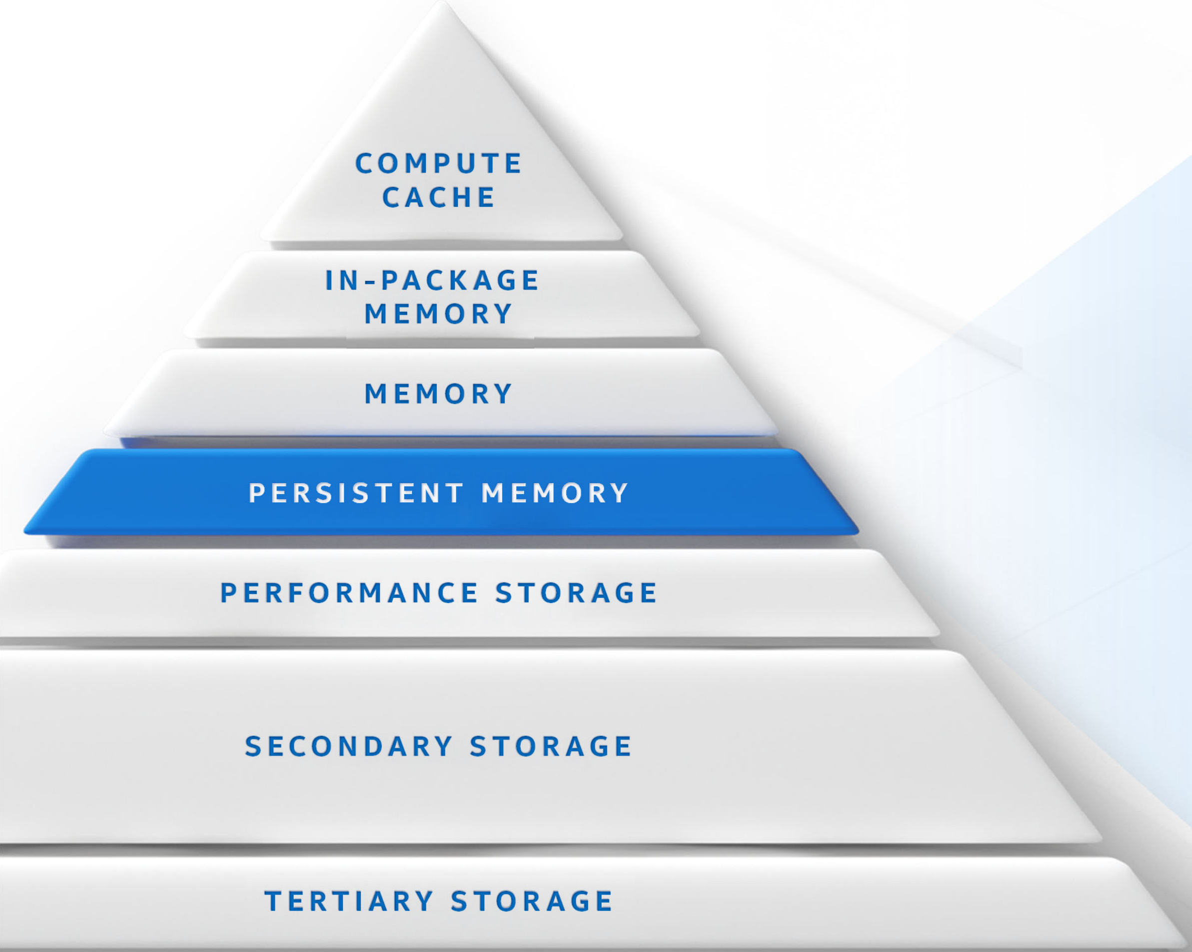
Intel 3D NAND SSD

Memory and Storage Hierarchy Gaps



For illustrative purposes only

Intel Optane™ Persistent Memory



COMPUTE CACHE

IN-PACKAGE MEMORY

MEMORY

PERSISTENT MEMORY

PERFORMANCE STORAGE

SECONDARY STORAGE

TERTIARY STORAGE

PERFORMANCE LIKE MEMORY, PERSISTENCE LIKE STORAGE

Direct load/store access to fast and large byte-addressable space

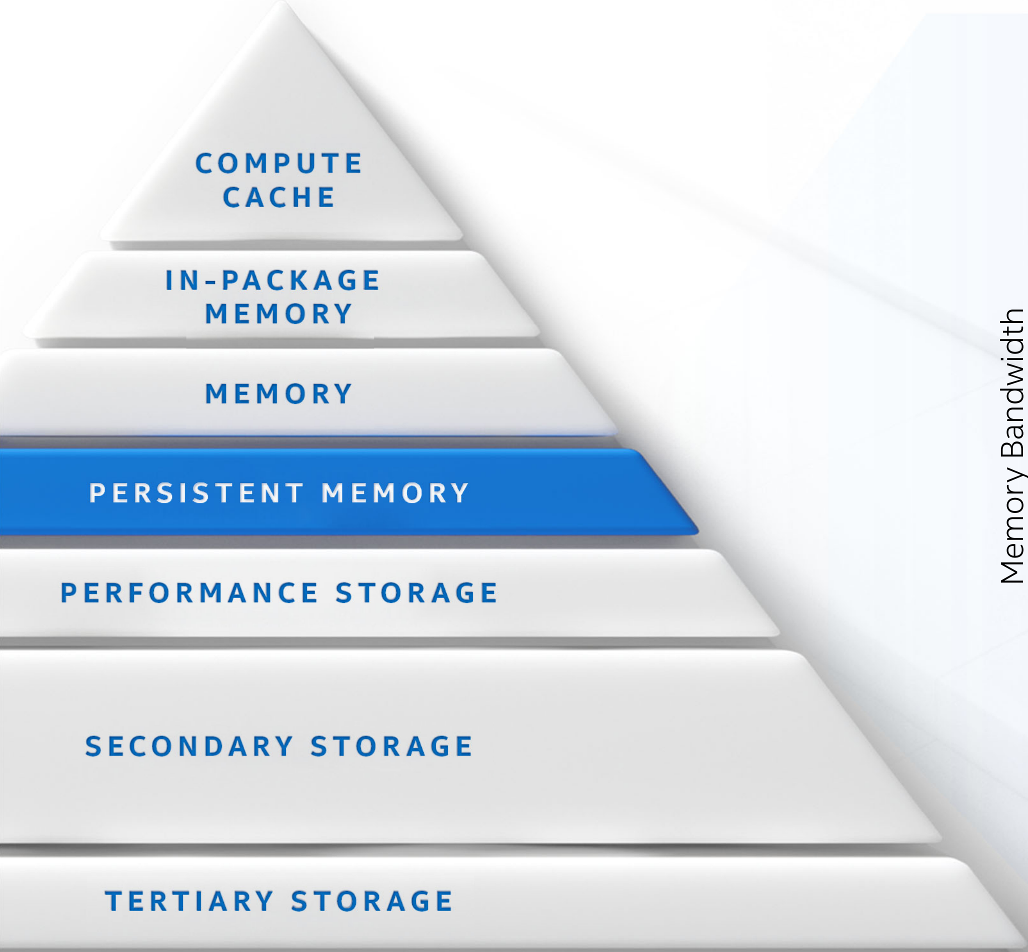
MEMORY MODE

- Higher capacity vs. DRAM at lower cost
- No application changes required

APP DIRECT MODE

- New tier of large capacity non-volatile memory (NVM)
- Innovative new usages, such as fast restart and remote access to memory

Intel Optane™ Persistent Memory



COMPUTE CACHE

IN-PACKAGE MEMORY

MEMORY

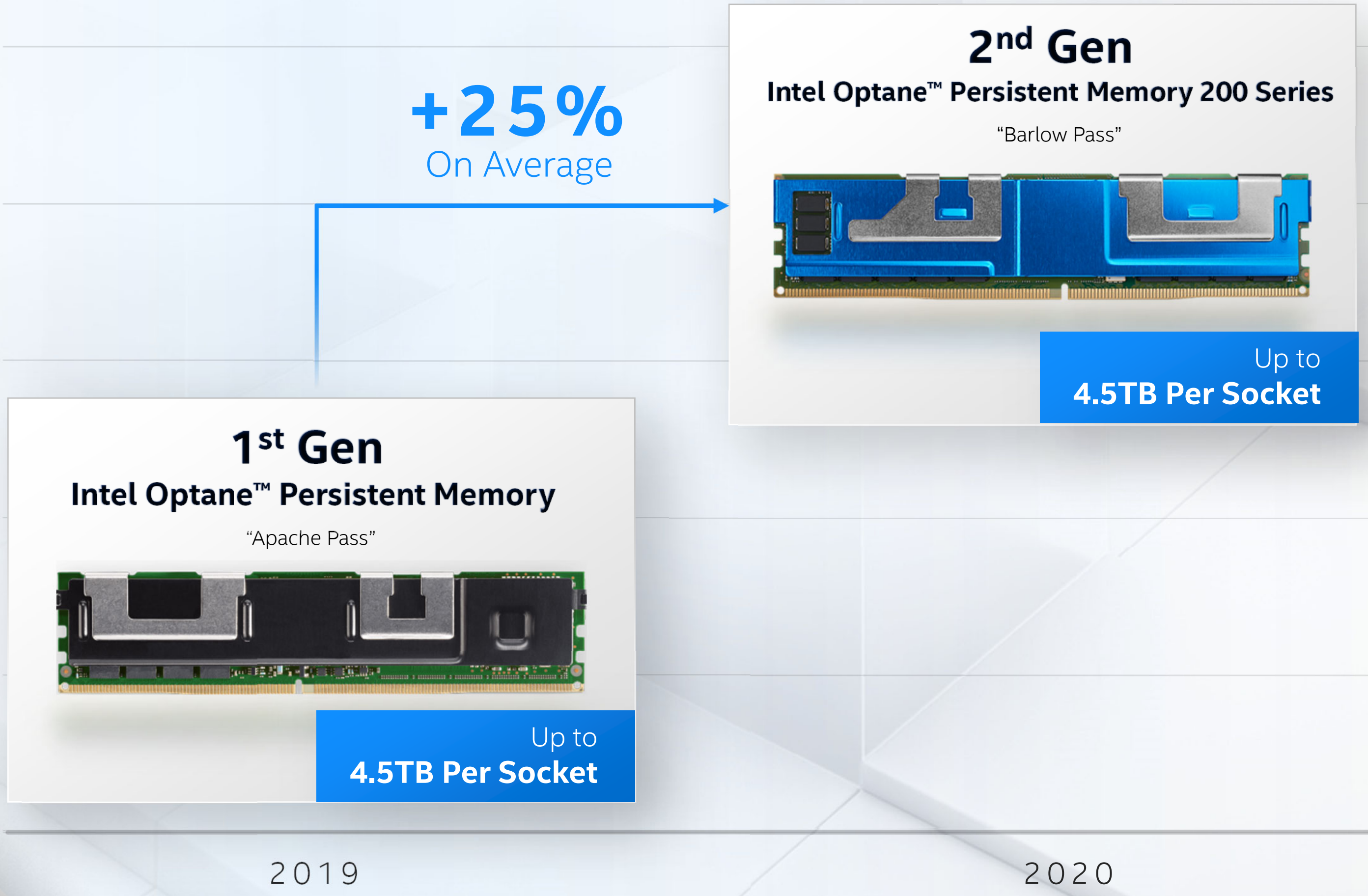
PERSISTENT MEMORY

PERFORMANCE STORAGE

SECONDARY STORAGE

TERTIARY STORAGE

Memory Bandwidth



+25%
On Average

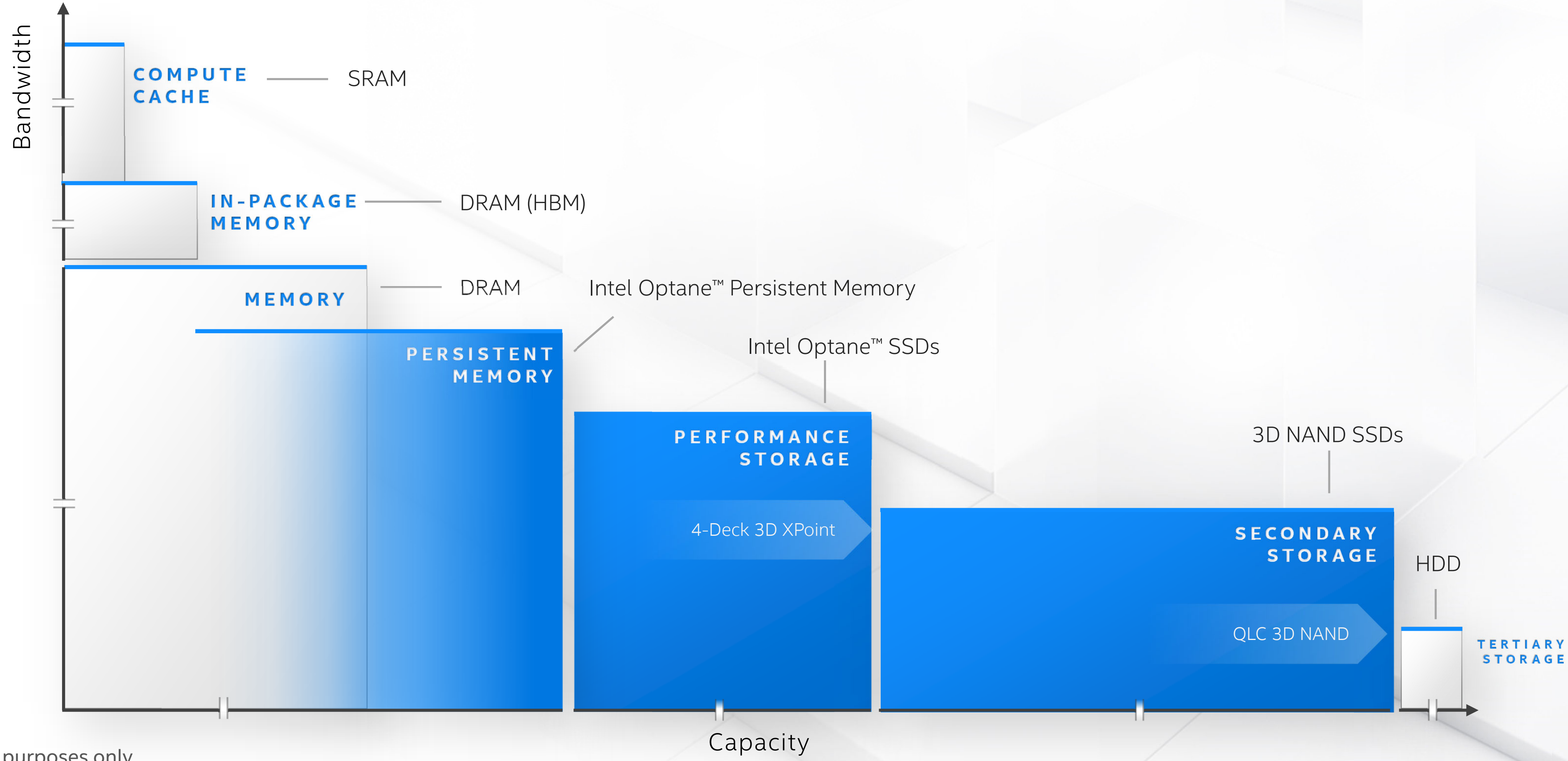
2nd Gen
Intel Optane™ Persistent Memory 200 Series
"Barlow Pass"
Up to
4.5TB Per Socket

1st Gen
Intel Optane™ Persistent Memory
"Apache Pass"
Up to
4.5TB Per Socket

2019

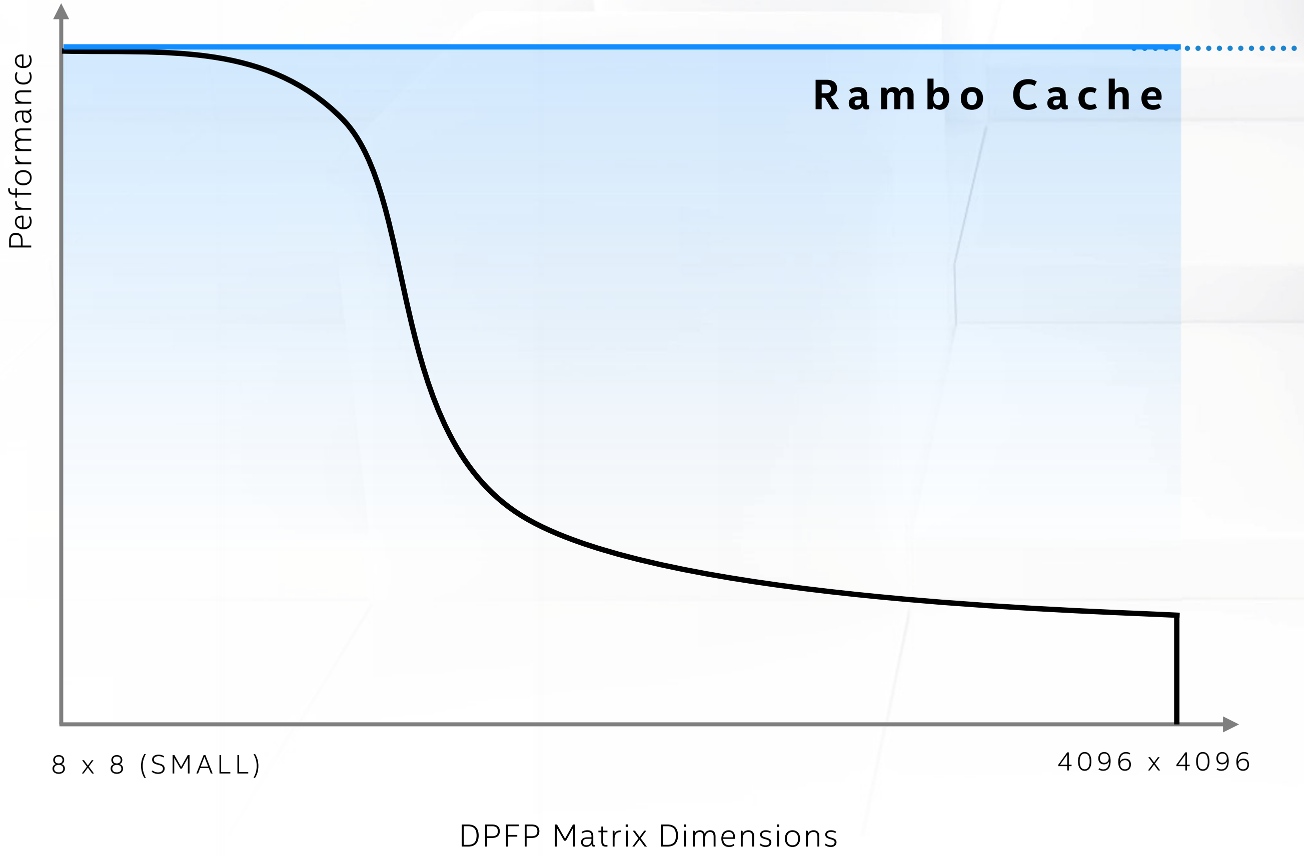
2020

Memory and Storage Hierarchy Gaps

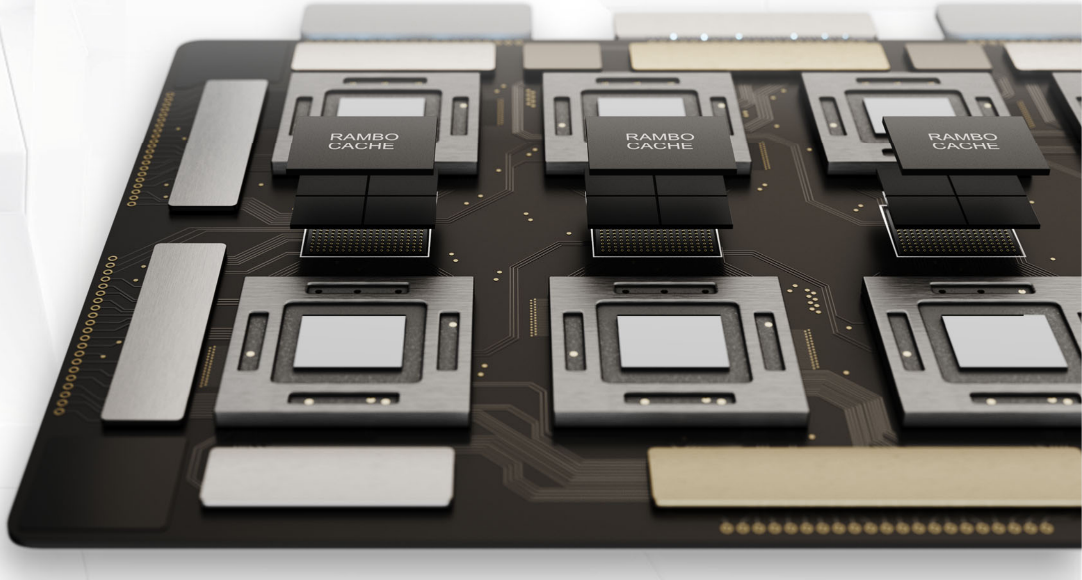


For illustrative purposes only

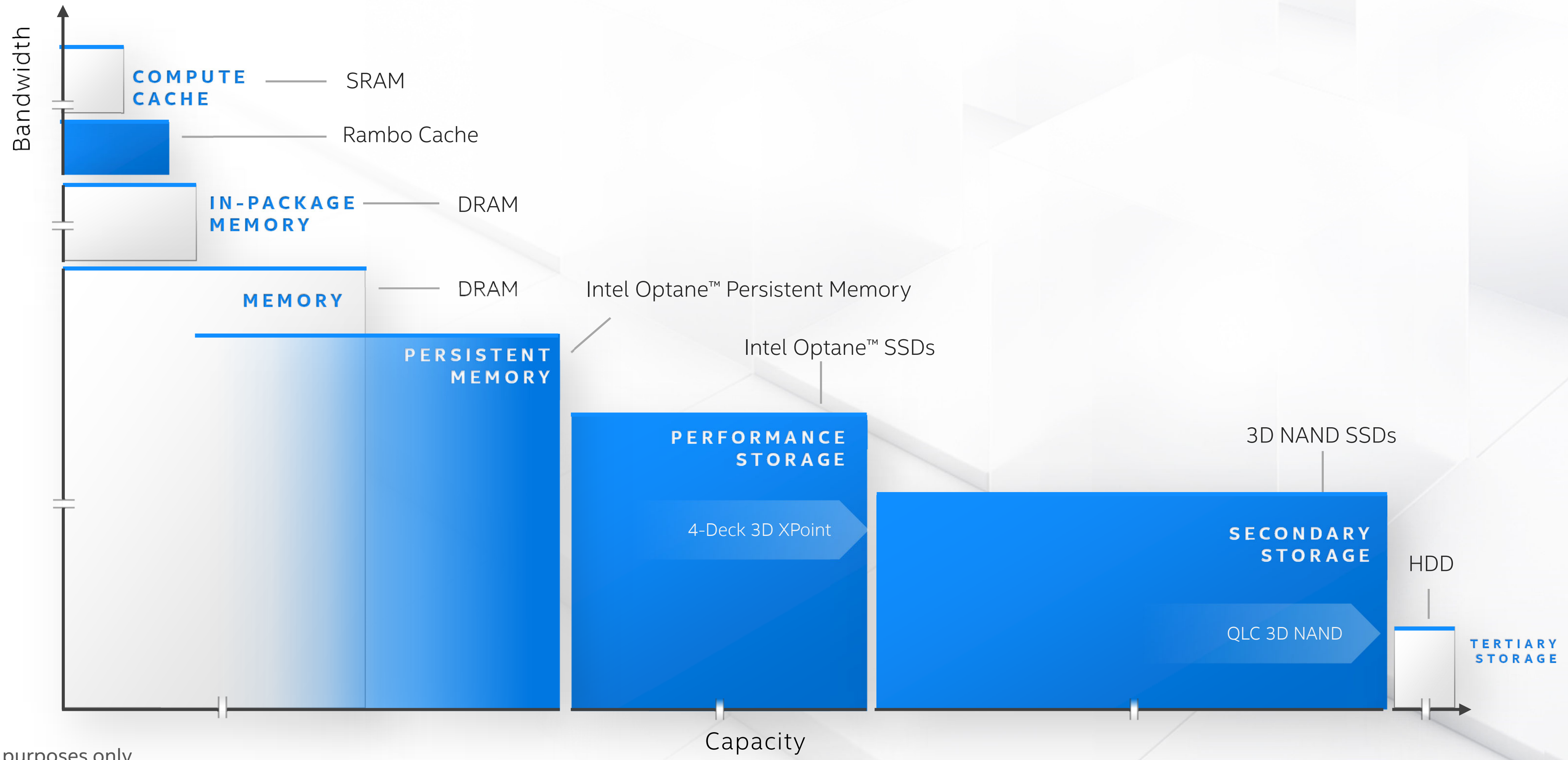
Rambo Cache



For illustrative purposes only

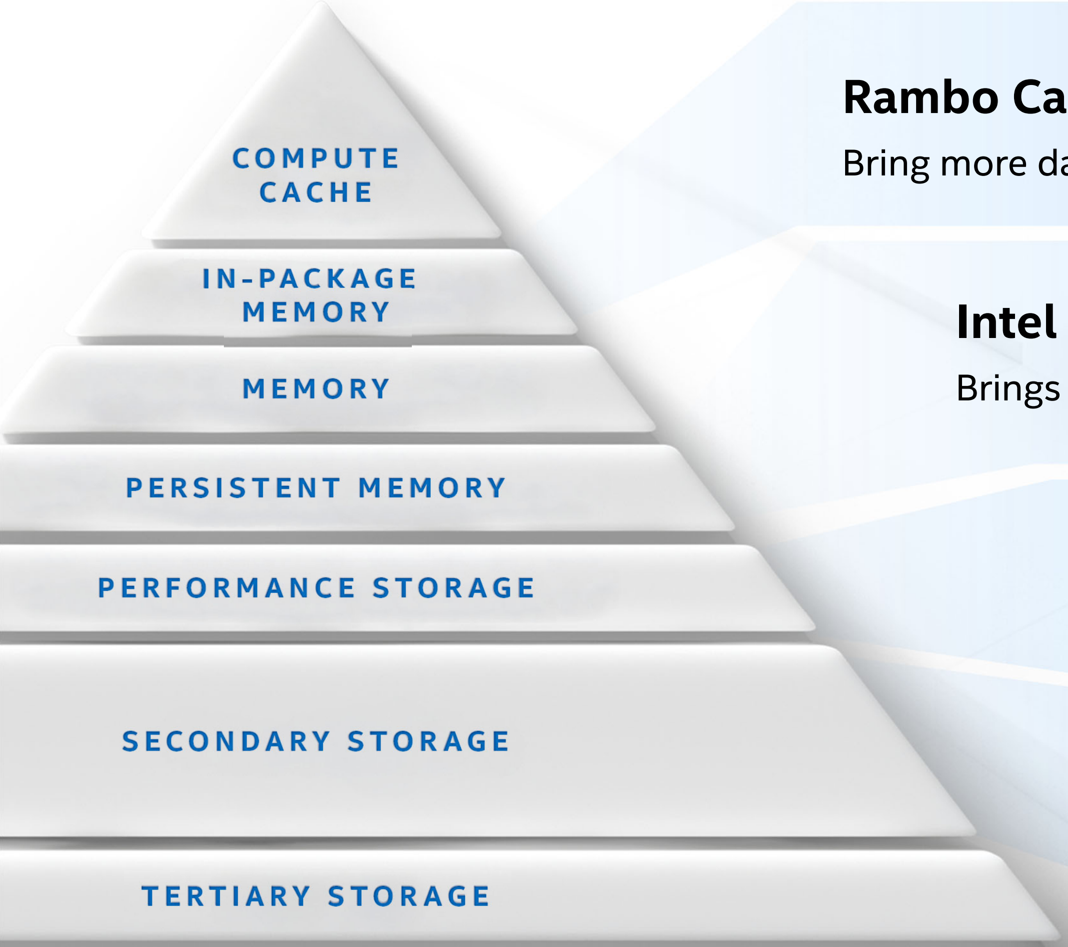


Memory and Storage Hierarchy Gaps



For illustrative purposes only

A Complete Hierarchy

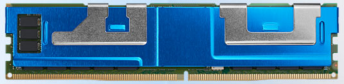


Rambo Cache

Bring more data into cores

Intel Optane™ persistent memory

Brings more data into memory



Intel Optane™ SSDs

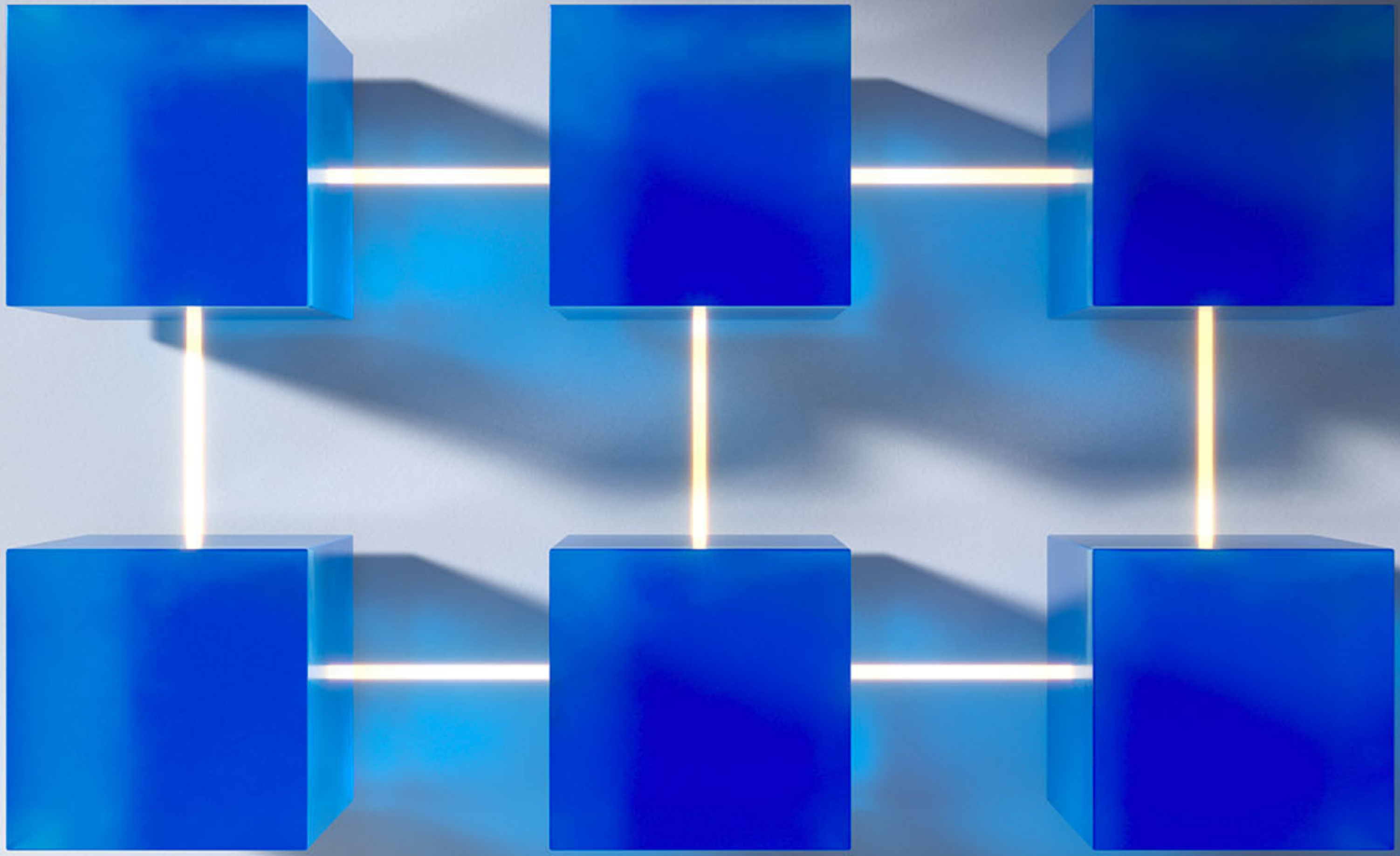
Get data from storage faster

Intel 3D NAND QLC SSDs

Efficiently scale and unlock stored data


TECHNOLOGY PILLARS
SOFTWARE
SECURITY
INTERCONNECT
MEMORY
XPU ARCHITECTURE
PROCESS & PACKAGING
intel

Interconnect from Microns to Miles



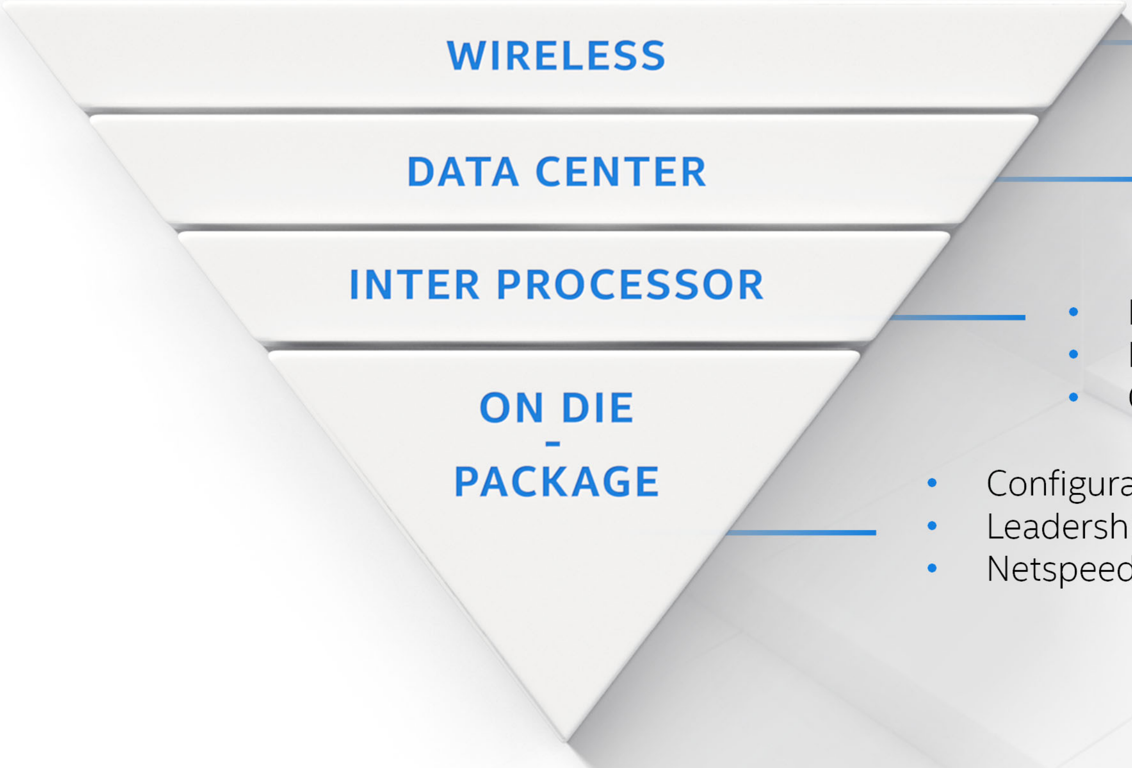
Hong Hou

Corporate Vice President,
General Manager
Connectivity Group

TECHNOLOGY
PILLARS

Architecture Day **2020**

Interconnect Hierarchy & Intel



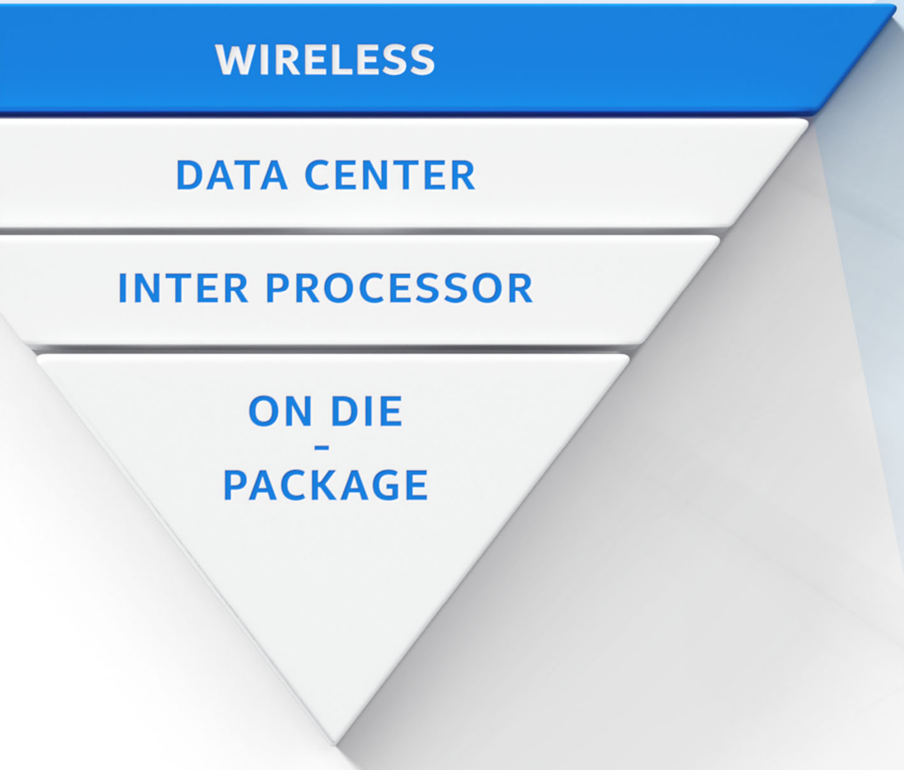
- 5G client to cloud strategy
- Compute + Comms architecture
- CPU sockets with network acceleration

- Silicon photonics architecture
- High-speed switching architecture
- Foundational and Smart NICs

- Leadership CPU Interconnect (UPI)
- New Accelerator and memory architecture
- Open eco-system and standards

- Configurable, Scalable architecture
- Leadership packaging architecture
- Netspeed

Our 5G Portfolio


WIRELESS

DATA CENTER

INTER PROCESSOR

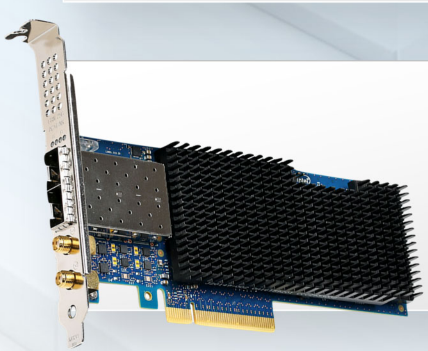
ON DIE
–
PACKAGE

Intel Atom P5900

highly integrated 10nm SOC
purpose built for 5G

Diamond Mesa

For 5G network Acceleration



Ethernet 700 Series

“Edgewater Channel,” – our first
5G-optimized network adapter

Newsroom : https://newsroom.intel.com/press-kits/5g-network-portfolio-launch/#gs.bj3636

Client Connectivity Roadmap

Capabilities

802.11ac

- Crowded Spectrum
- High Latency

THUNDERBOLT™ 3

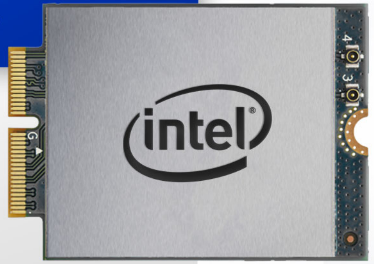
USB

MINI-DISPLAY PORT

Before 2020

INTEL WI-FI 6

- ~3x faster throughput
- ~4x network capacity
- ~75% lower latency



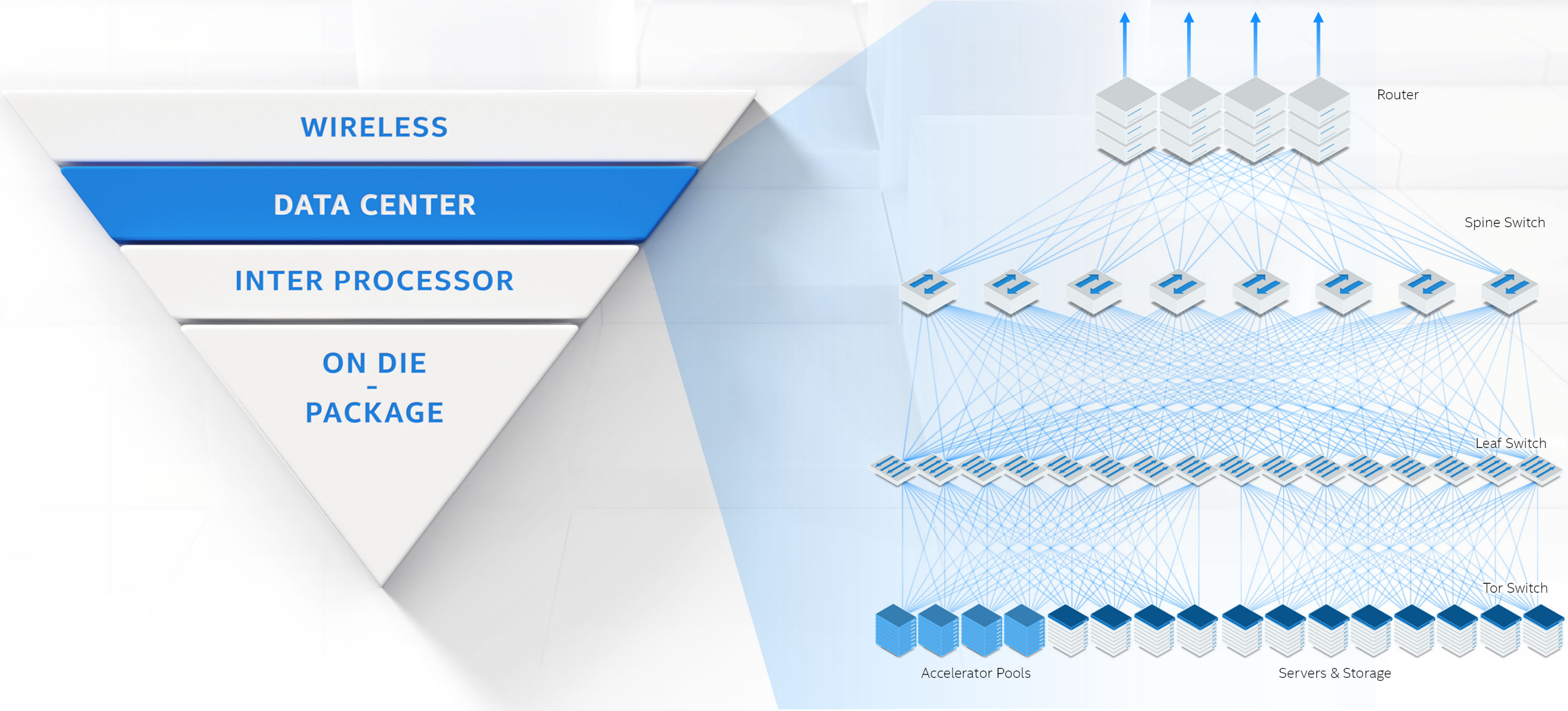
THUNDERBOLT™ 4

- Support for 2 4K monitors or one 8K display
- Accessories with four Thunderbolt ports
- USB-C Connectivity
- Always 40 Gb/s

2020

www.intel.com/wifi6disclaimers

Data Center Interconnect



WIRELESS

DATA CENTER

INTER PROCESSOR

ON DIE
–
PACKAGE

Connectivity Data Center Product Portfolio

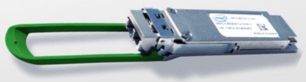
Unleashing the performance of compute at scale through innovations in end-to-end connectivity

End-to-End Co-Optimization

Photonics Integration

OPTICAL MODULES

High-bandwidth connectivity at 100G, 400G and beyond



ETHERNET SWITCH

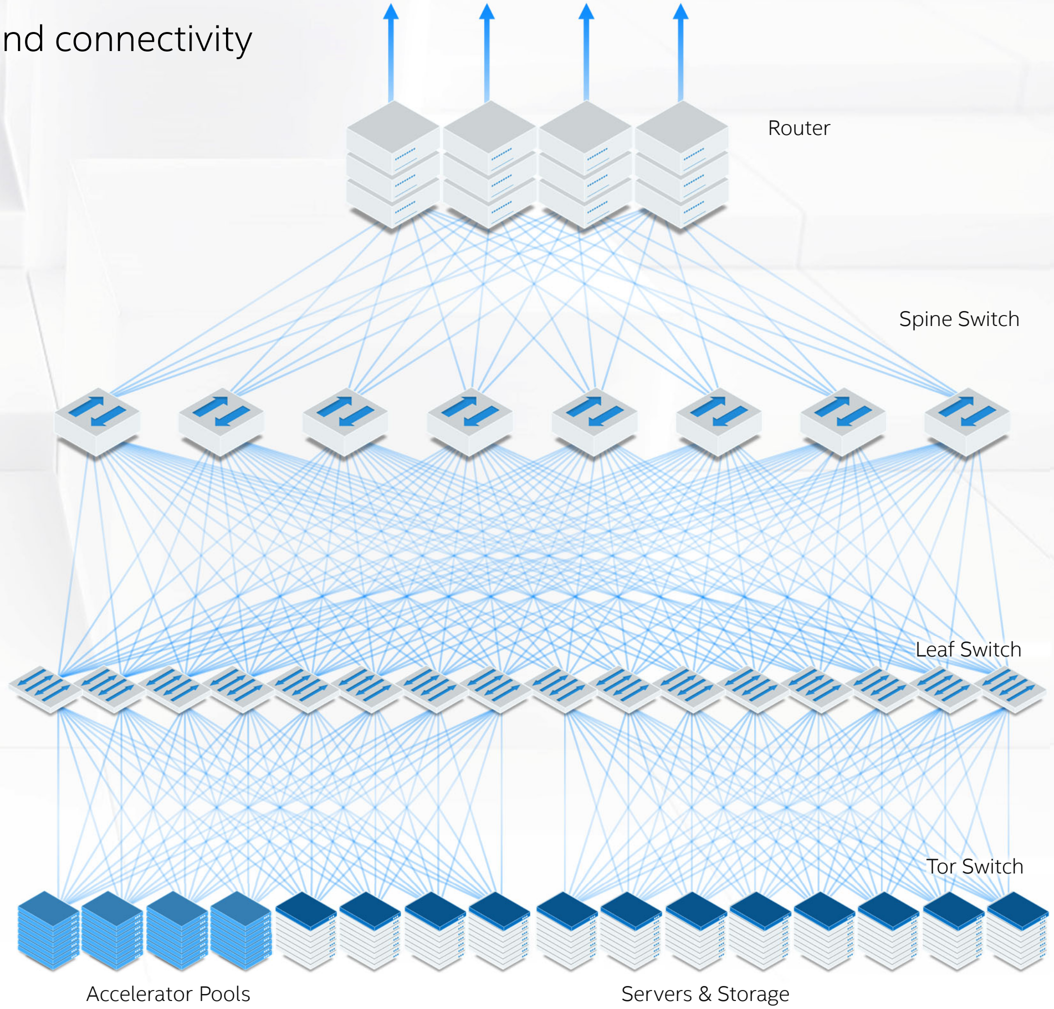
P4-programmable scale-out fabric with uncompromising performance

ETHERNET NIC

Programmable infrastructure acceleration for demanding data movement

CPUs & xPUs

Fabric-enabled endpoints aligned to accelerators & software pipelines

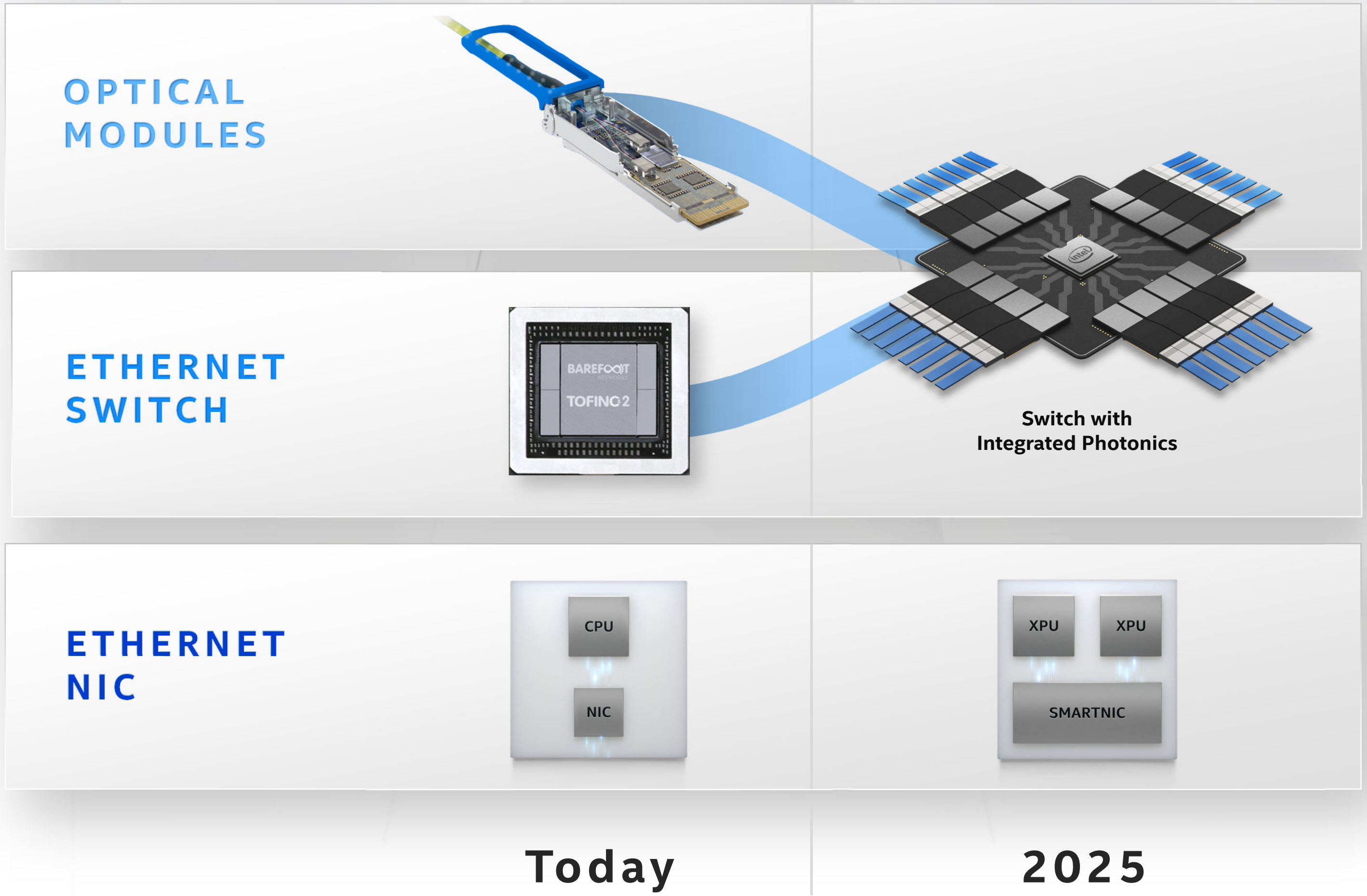


Cloud DC Infrastructure Point of Arrival

Continuous network innovation required to unbridle compute and storage at ever-increasing scale

4

Network is a Programmable Platform
Empowers customers to create new E2E applications optimized for their workloads with hardware-level performance at software-like pace of innovation



2

Photonics is integrated and ubiquitous (Optical IO)
- 20x bandwidth density
- 2x cost & power improvement

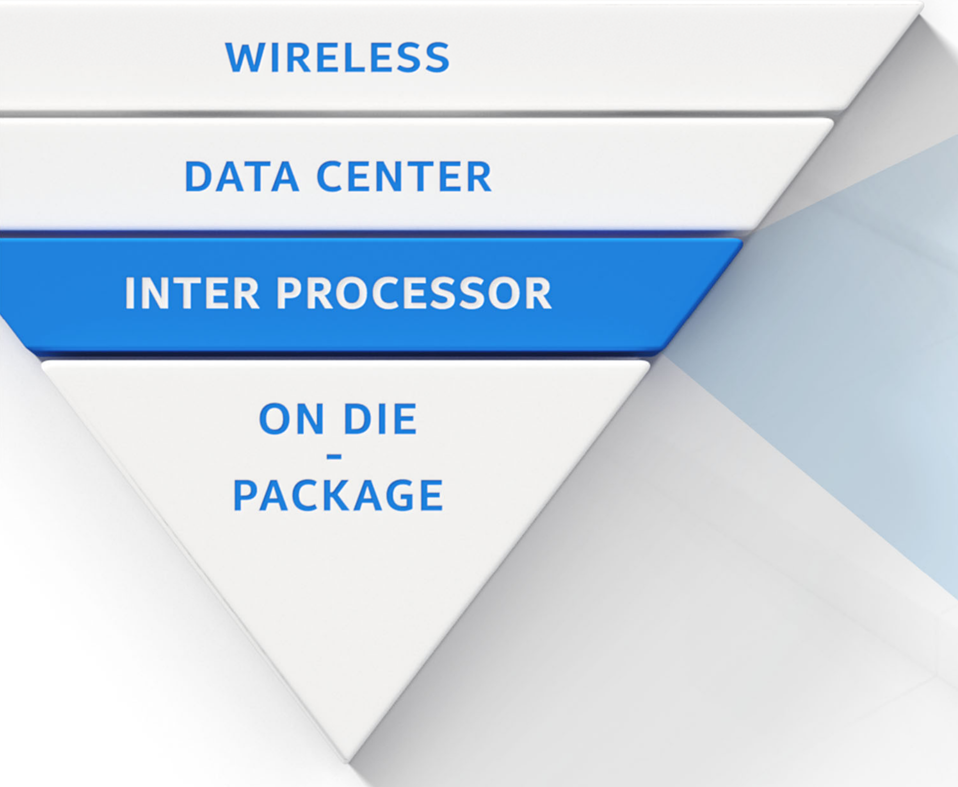
3

E2E co-optimizations unleash performance @ scale
- Telemetry, QoS, flow control
- Traffic shaping, steering, coalescing

1

SmartNIC becomes critical to offload
- Optimized for complex data flows
- Diverse workload acceleration

Inter-Processor Interconnect



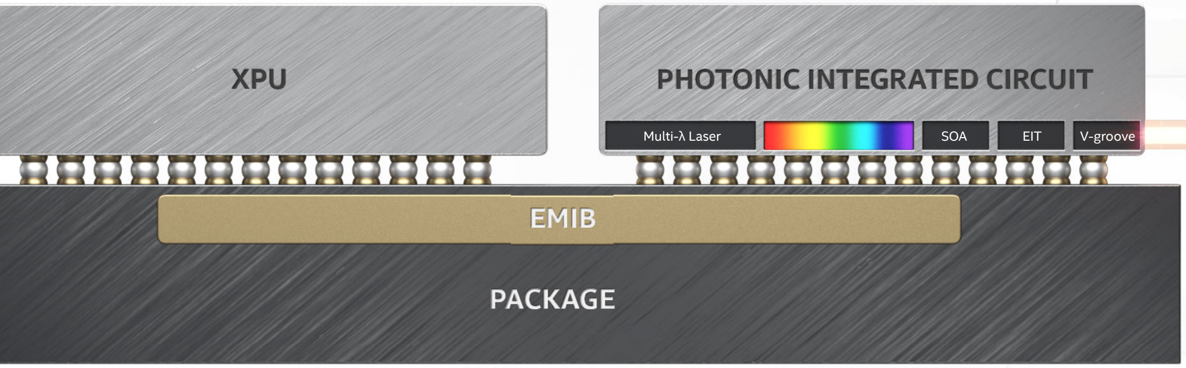
CXL

CXL provides a more **fluid and flexible** memory model
Single, common, memory address space across
processors and devices

to enable a **broad, open eco-system for**
heterogeneous computing and server disaggregation

Glimpse Into the Future

Optical IO



BREAKING THROUGH THE I/O BARRIER

Ultra-high bandwidth
~1Tbps per fiber

Reach
Order of magnitudes better than copper

Shoreline Density
>6x improvement over PCIe6

Energy efficiency
Trending to 2pJ/b (50% of PCIe6)

Latency
Comparable to electrical IO

Interconnect Leadership

Interconnect leadership is foundational to **moving the growing amount of data** and to the overall compute leadership

Intel is uniquely positioned with the broadest portfolio of industry leading technologies, products, ongoing innovation and deep investments across all interconnect layers spanning on die, on package to data center and wireless networks


TECHNOLOGY PILLARS
SOFTWARE
SECURITY
INTERCONNECT
MEMORY
XPU ARCHITECTURE
PROCESS & PACKAGING
intel

INTEL TECHNOLOGY

WHAT WE SAID IN 2018

Successful new architectures will have additional security technologies as a foundational property and priority

Security Challenge

Exponentially Increasing Surface Area

4 ARCHITECTURES

7 LEVELS OF MEMORY HIERARCHY

BILLIONS OF DEVICES

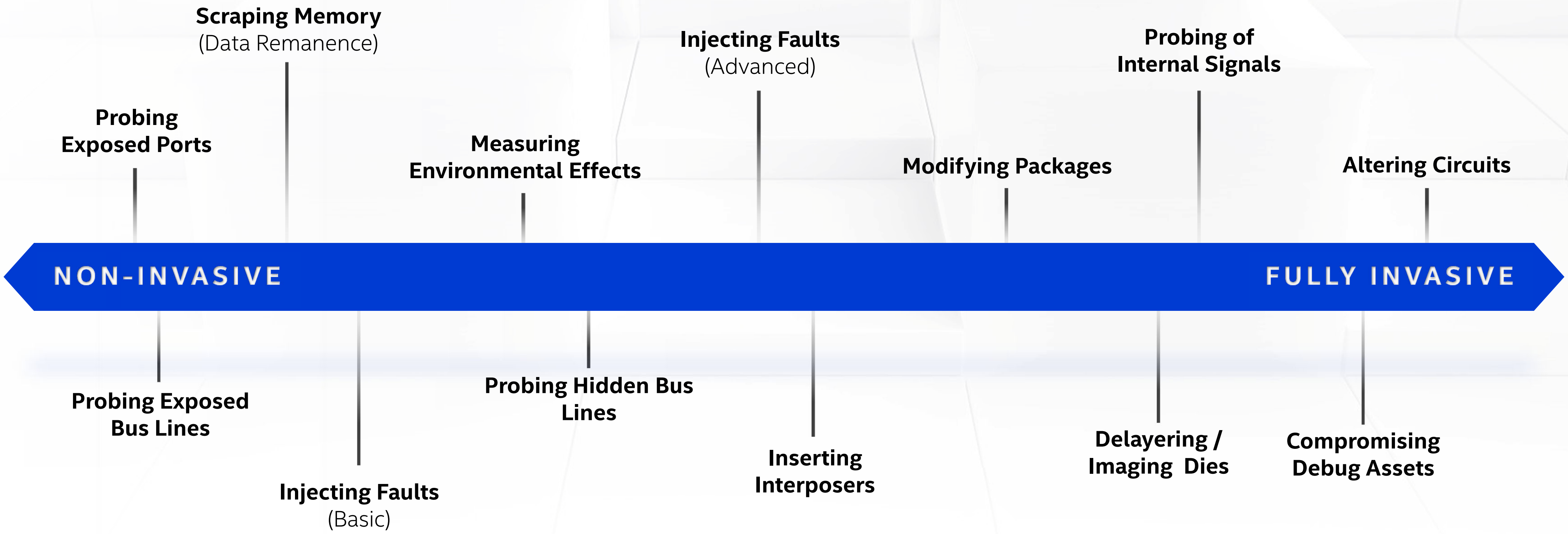
4 LEVELS OF INTERCONNECT HIERARCHY

Workloads Expand & Threat Models Evolve

SECURITY CHALLENGES

- LACK OF PHYSICAL SECURITY
- DISTRIBUTED WORKLOADS
- HETEROGENEITY

Spectrum of Physical Attacks



NON-INVASIVE

Probing Exposed Ports

Probing Exposed Bus Lines

Scraping Memory (Data Remanence)

Injecting Faults (Basic)

Measuring Environmental Effects

Probing Hidden Bus Lines

Injecting Faults (Advanced)

Inserting Interposers

Modifying Packages

Delayering / Imaging Dies

Probing of Internal Signals

Compromising Debug Assets

Altering Circuits

FULLY INVASIVE

Security Timeline

FOUNDATIONAL SECURITY

OS, BIOS, BOOT PROTECTION

PLATFORM TRUST

SECURITY CO-PROCESSOR

CRYPTO ACCELERATION

MEMORY ENCRYPTION

INTELLIGENT SECURITY ENGINES EVERYWHERE

Innovation Legacy

Today

Vision

Security Timeline

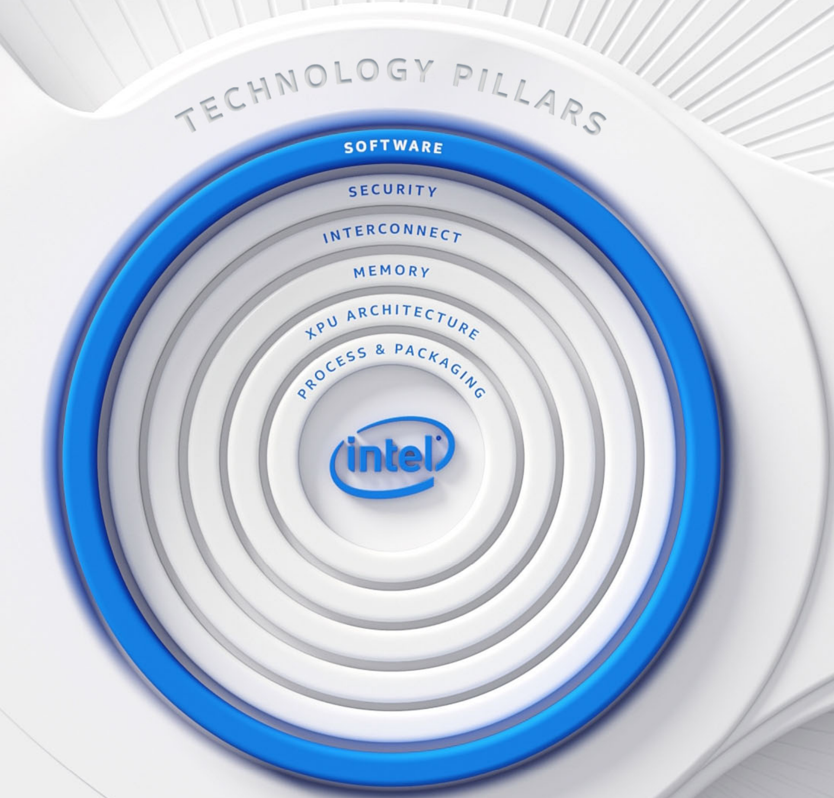
	Innovation Legacy	Today	Vision
WORKLOAD PROTECTION	VIRTUALIZATION MEASURED BOOT APPLET	APP & DATA PROTECTION	VM ISOLATION
FOUNDATIONAL SECURITY	OS, BIOS, BOOT PROTECTION PLATFORM TRUST SECURITY CO-PROCESSOR CRYPTO ACCELERATION	MEMORY ENCRYPTION	INTELLIGENT SECURITY ENGINES EVERYWHERE

Security Timeline

	Innovation Legacy	Today	Vision
SOFTWARE RELIABILITY	OBJECT ORIENTED ARCHITECTURE	CONTROL FLOW INTEGRITY (CET)	MEMORY TAGGING
WORKLOAD PROTECTION	VIRTUALIZATION MEASURED BOOT APPLET	APP & DATA PROTECTION	VM ISOLATION
FOUNDATIONAL SECURITY	OS, BIOS, BOOT PROTECTION PLATFORM TRUST SECURITY CO-PROCESSOR CRYPTO ACCELERATION	MEMORY ENCRYPTION	INTELLIGENT SECURITY ENGINES EVERYWHERE

Security Timeline

	Innovation Legacy	Today	Vision
SOFTWARE RELIABILITY	OBJECT ORIENTED ARCHITECTURE	CONTROL FLOW INTEGRITY (CET)	MEMORY TAGGING
WORKLOAD PROTECTION	VIRTUALIZATION MEASURED BOOT APPLET	APP & DATA PROTECTION	VM ISOLATION
FOUNDATIONAL SECURITY	OS, BIOS, BOOT PROTECTION PLATFORM TRUST SECURITY CO-PROCESSOR CRYPTO ACCELERATION	MEMORY ENCRYPTION	INTELLIGENT SECURITY ENGINES EVERYWHERE


TECHNOLOGY PILLARS
SOFTWARE
SECURITY
INTERCONNECT
MEMORY
XPU ARCHITECTURE
PROCESS & PACKAGING
intel

MOORE'S LAW AND SOFTWARE

WHAT WE SAID IN 2018

For every order of magnitude performance potential from new hardware architecture, there is often >2 orders of magnitude unlocked by software

HETEROGENEITY CHALLENGE

WHAT WE SAID IN 2018

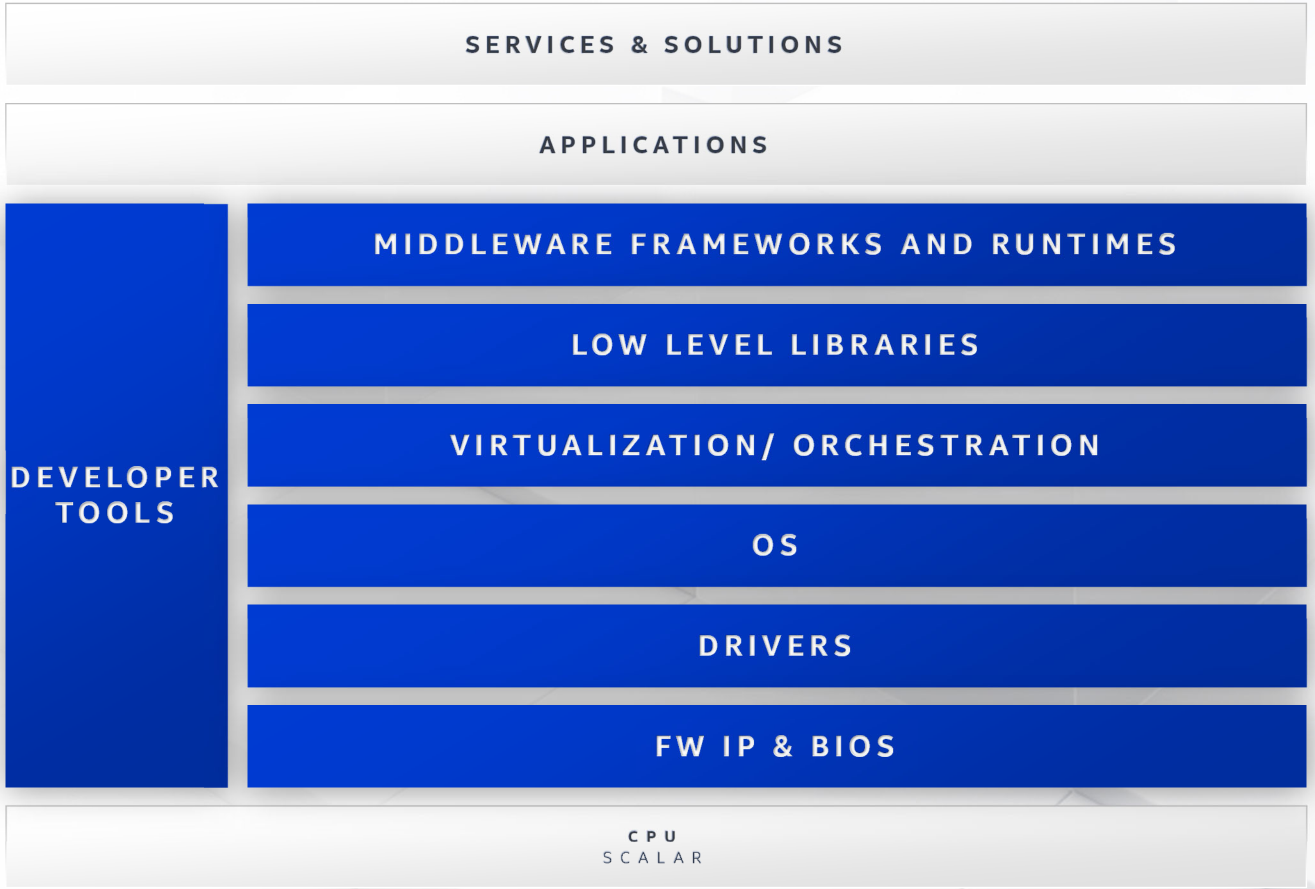
Abstractions for Scalar, Vector, Matrix and Spatial and targeted domain specific libraries will provide exa-scale computing access for everyone.
All under the 'One API' umbrella.

Architecture Impact

Generality ∝ Software Stack Scale

20M

Developers



SERVICES & SOLUTIONS

APPLICATIONS

DEVELOPER TOOLS

MIDDLEWARE FRAMEWORKS AND RUNTIMES

LOW LEVEL LIBRARIES

VIRTUALIZATION/ ORCHESTRATION

OS

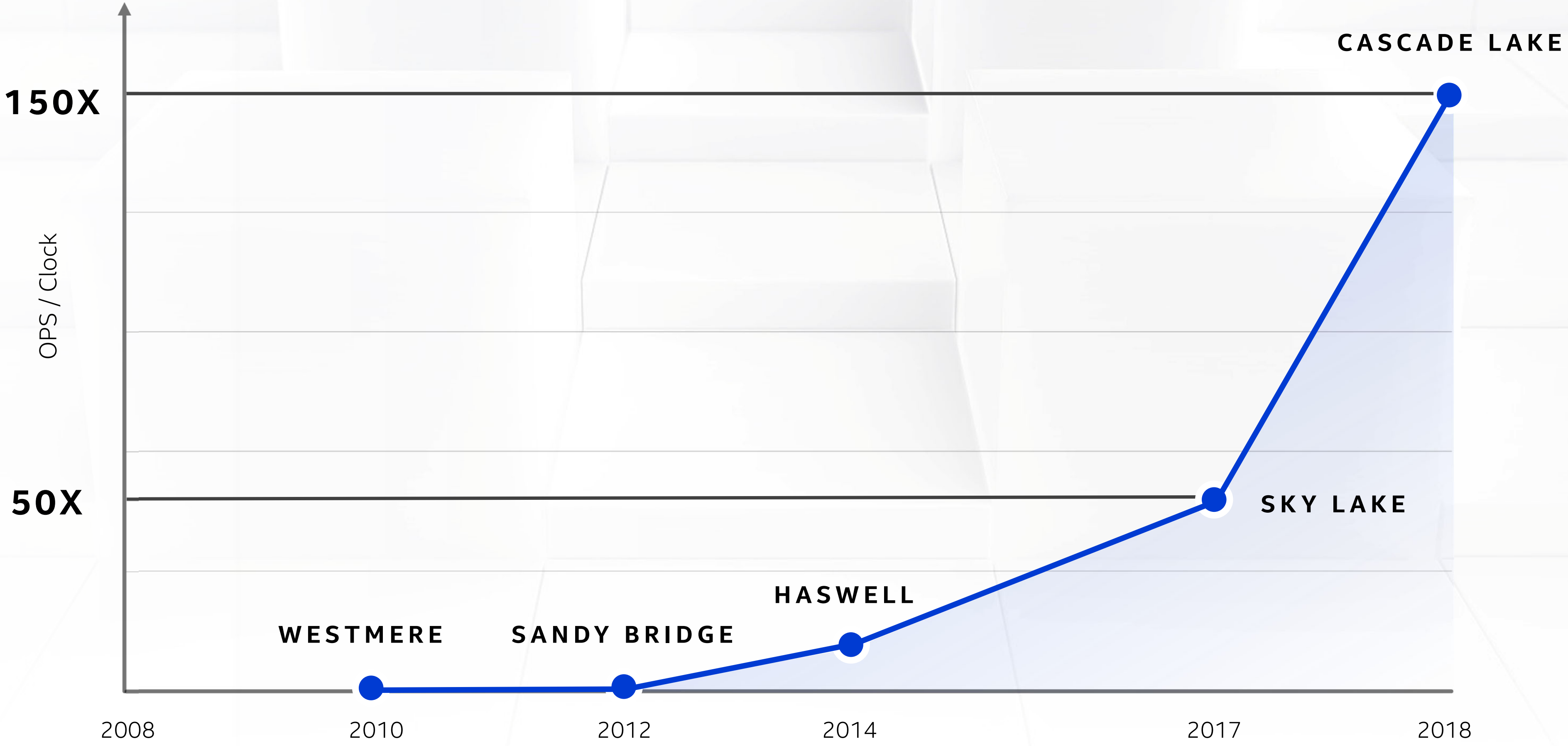
DRIVERS

FW IP & BIOS

CPU
SCALAR

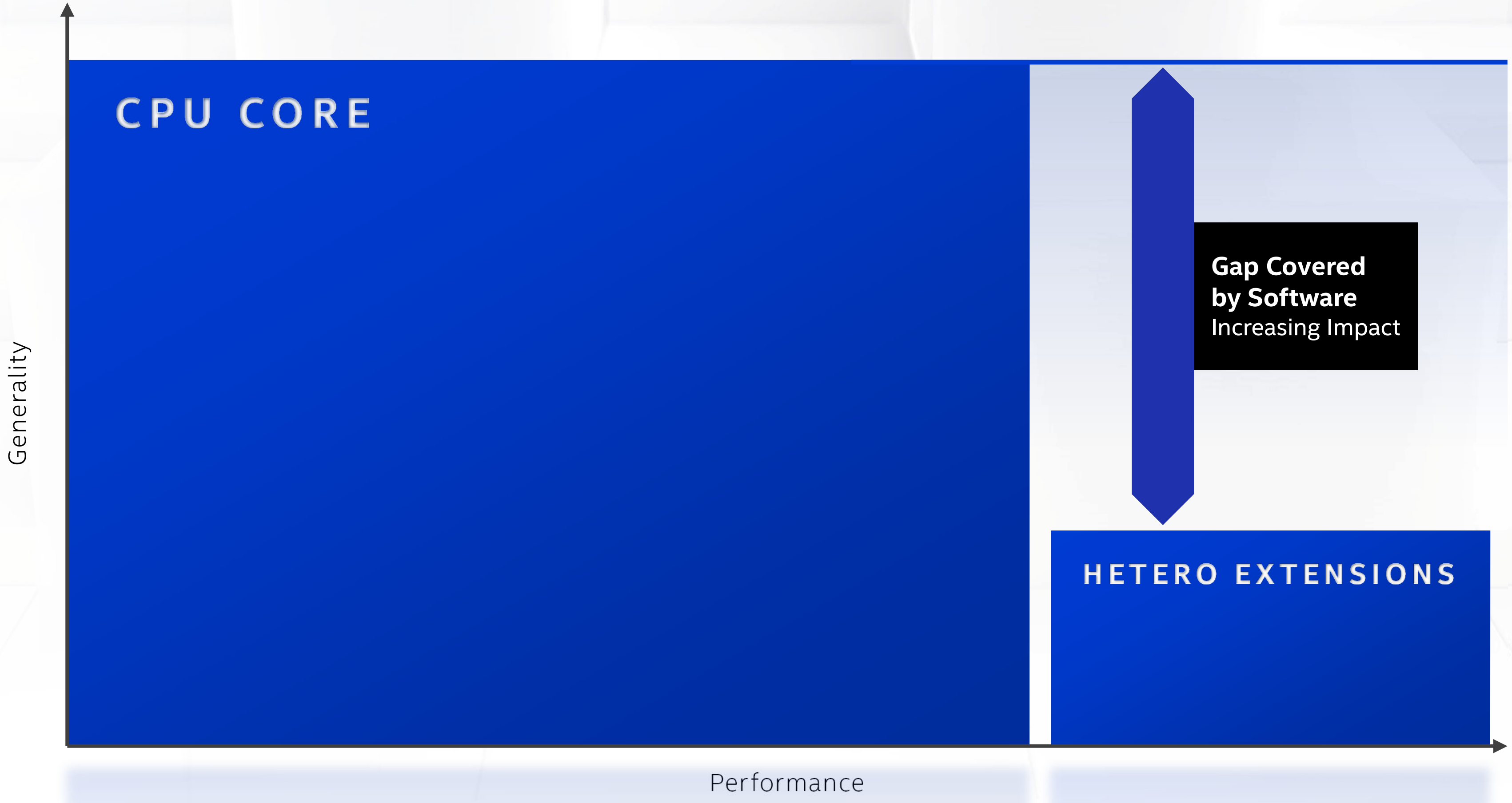
Generality $\propto$ 1/Architecture Heterogeneity

CPU AI



For illustrative purposes only

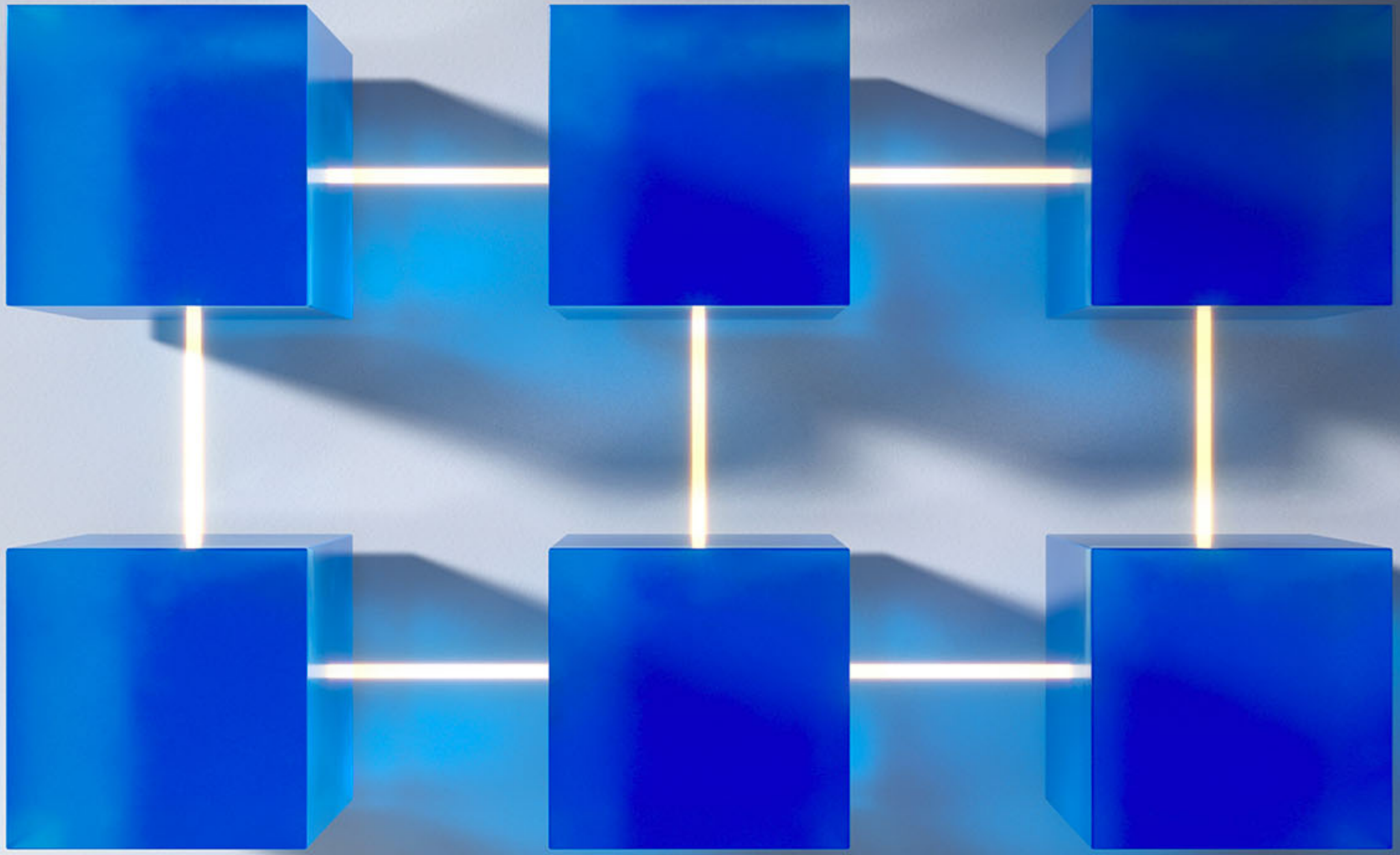
CPU Impact with ISA Extensions & Software



AI Software

Wei Li

VP & GM Machine Learning Performance



TECHNOLOGY PILLARS

Architecture Day **2020**

CHALLENGES
FOR AI SOFTWARE

DATA SIZE
Petabyte and growing

DATA TYPE
FP32, FP16, BF16, INT8, TF32

SOFTWARE SUPPORT
TensorFlow, PyTorch, MXnet and more

HARDWARE SUPPORT
CPU, GPU and XPUs

Deep Learning Progress on Intel CPU

DEEP LEARNING PROGRESS

INFERENCE LATENCY

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

TRAINING TIME

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

CUSTOMERS

- Facebook
- Alibaba ByteDance
- Google
- AWS
- Novartis
- Mastercard
- GE Healthcare
- ...

DEEP LEARNING FRAMEWORKS

- TensorFlow
- PyTorch
- ONNX Runtime
- PaddlePaddle
- MXNet
- OpenVINO
- BigDL
- Caffe
- ...

LIBRARIES & COMPILER

- MKL-DNN
- MLSL

HARDWARE

- CPU

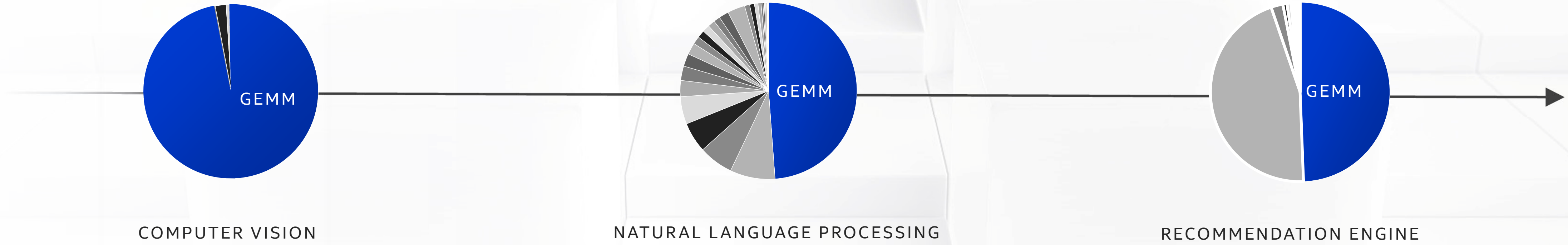
Scale Out DLRM Training

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

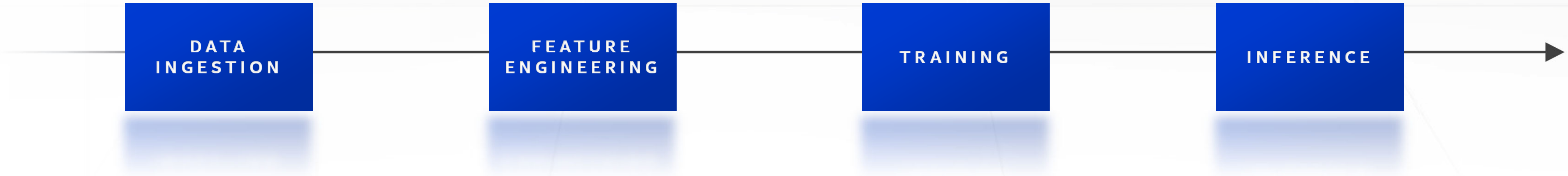
CPU: Intel(R) Xeon(R) Platinum 8280 CPU
Bios: SE5C620.86B.02.01.0009.092820190230
Memory: 12x 18ASF2G72PDZ-2G6D1
Network: 2x Intel Corporation Omni-Path HFI Silicon 100 Series
OS: Cent OS 7.6
Bootline: BOOT_IMAGE=/vmlinuz-3.10.0-957.el7.x86_64 rhgb quiet intel_pstate=disable nmi_watchdog=0 intel_idle.max_cstate=1 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.

AI Growth and End-to-End Data Pipelines

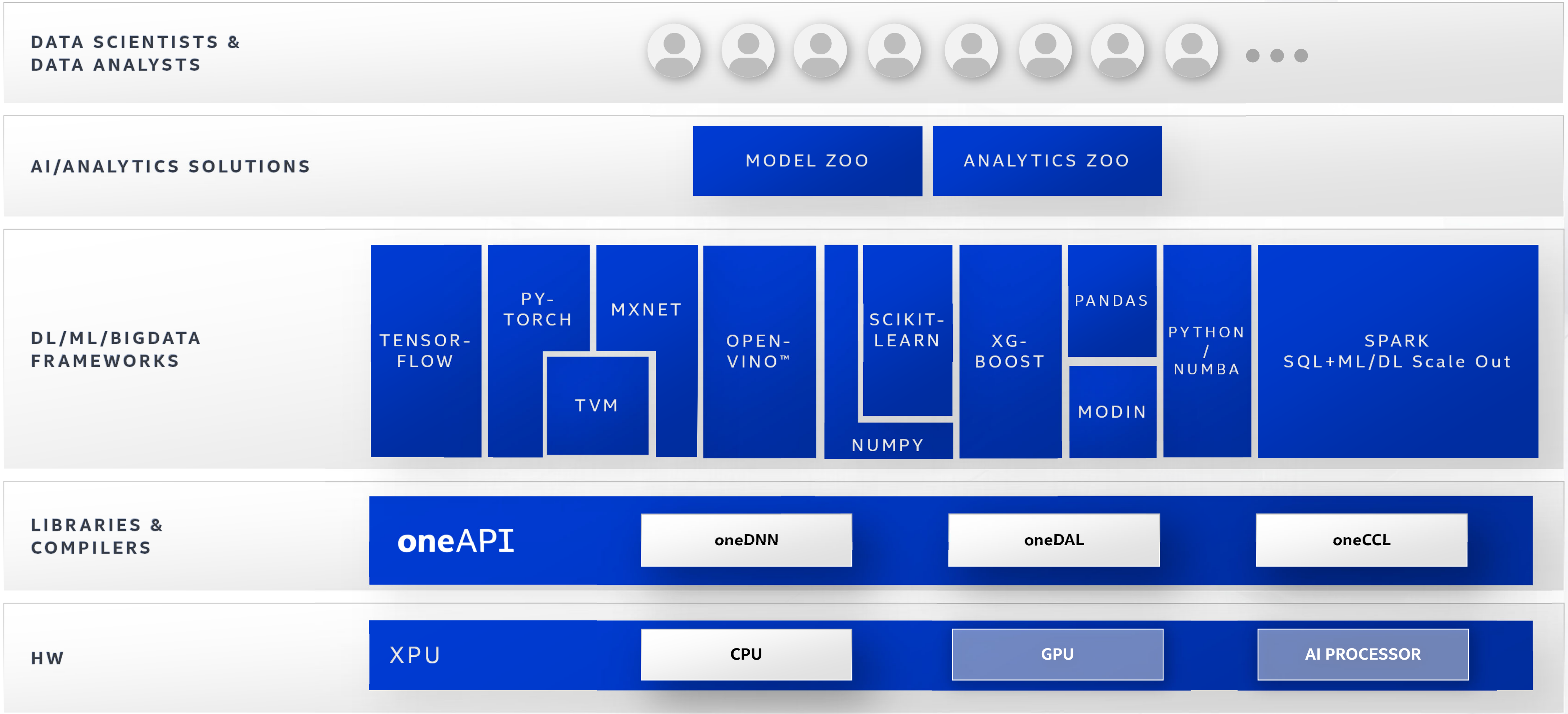
Rapidly growing deep learning with changing compute profiles



End to end DL/ML pipelines with big data



AI Software Ecosystem on Intel



DATA SCIENTISTS & DATA ANALYSTS

AI/ANALYTICS SOLUTIONS

- MODEL ZOO
- ANALYTICS ZOO

DL/ML/BIGDATA FRAMEWORKS

- TENSORFLOW
- PYTORCH
- MXNET
- TVM
- OPENVINO™
- SCIKIT-LEARN
- NUMPY
- XGBOOST
- PANDAS
- MODIN
- PYTHON / NUMBA
- SPARK SQL+ML/DL Scale Out

LIBRARIES & COMPILERS

oneAPI

- oneDNN
- oneDAL
- oneCCL

HW

XPU

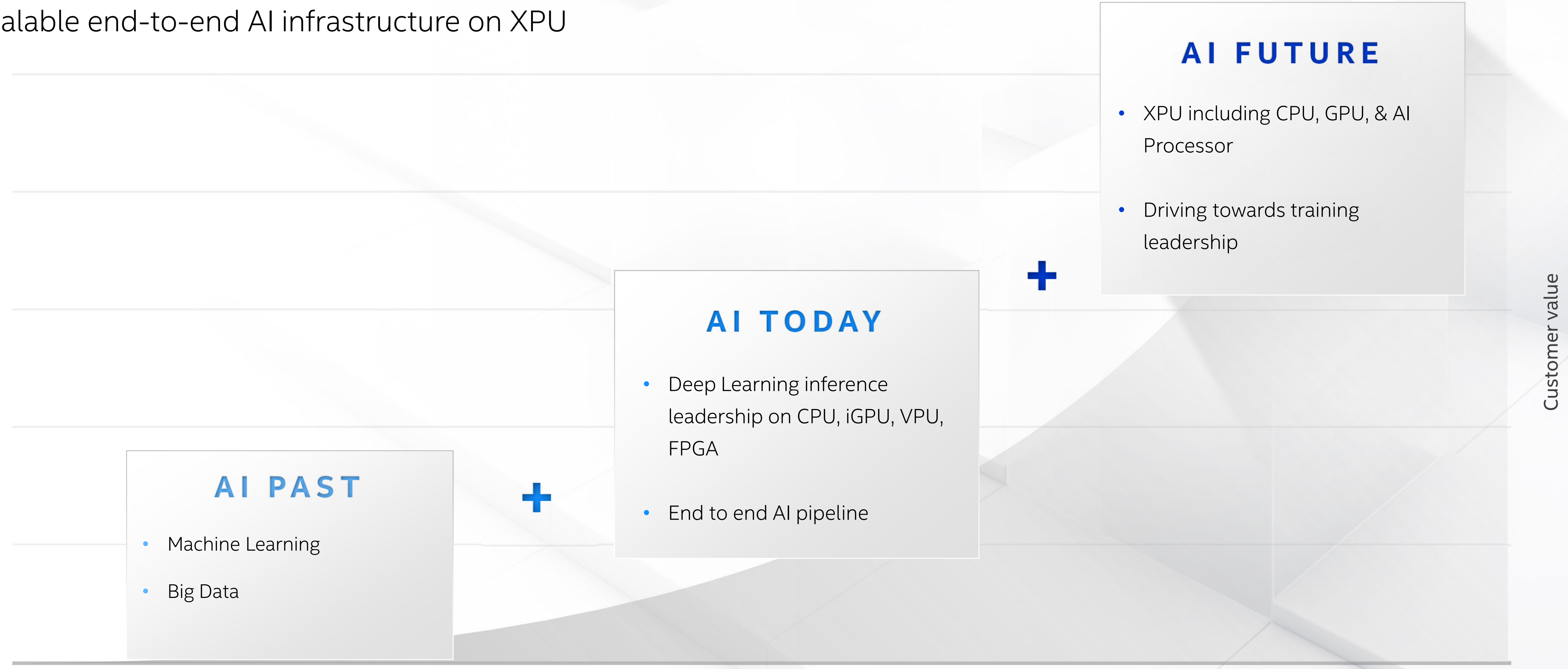
- CPU
- GPU
- AI PROCESSOR

Refer to https://software.intel.com/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

Refer to https://software.intel.com/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

Intel AI Software Roadmap

Scalable end-to-end AI infrastructure on XPU



AI PAST

- Machine Learning
- Big Data

+

AI TODAY

- Deep Learning inference leadership on CPU, iGPU, VPU, FPGA
- End to end AI pipeline

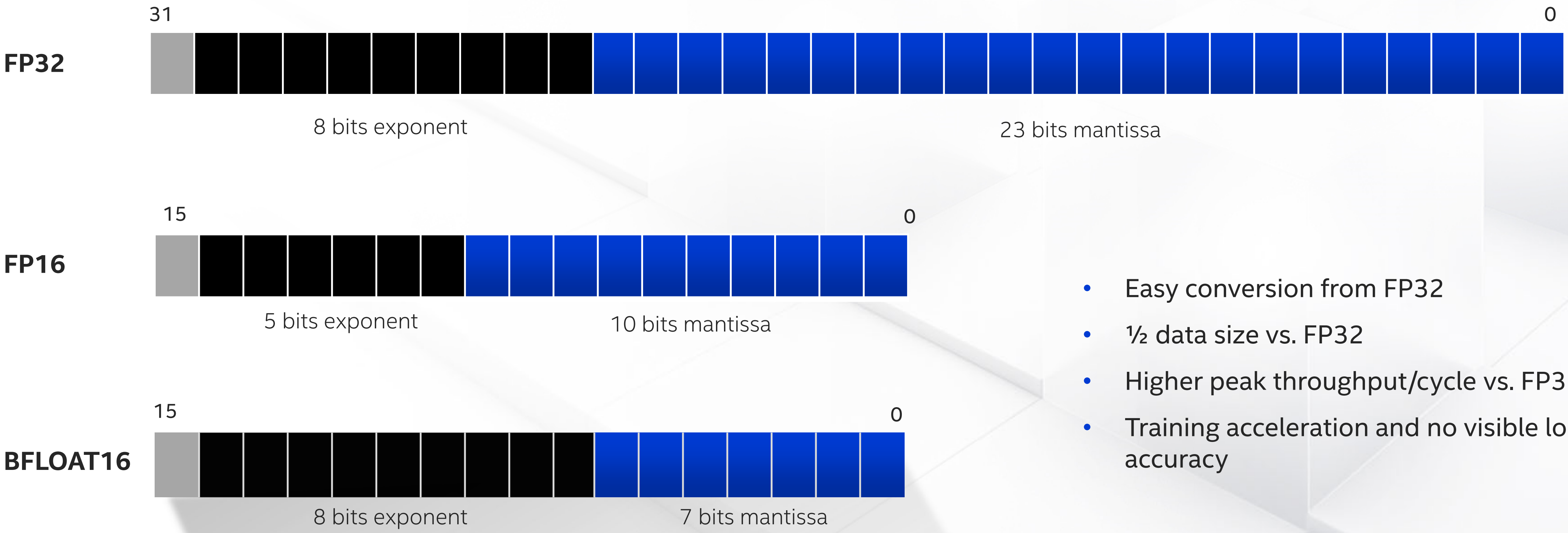
+

AI FUTURE

- XPU including CPU, GPU, & AI Processor
- Driving towards training leadership

BFLOAT16

Better performance with no visible loss in accuracy



- Easy conversion from FP32
- ½ data size vs. FP32
- Higher peak throughput/cycle vs. FP32
- Training acceleration and no visible loss in accuracy

BFLOAT16 Support on Intel XPUs

Intel 3rd Generation Xeon

(Cooper Lake)

BFLOAT16 Support in Software

USER EXPERIENCE

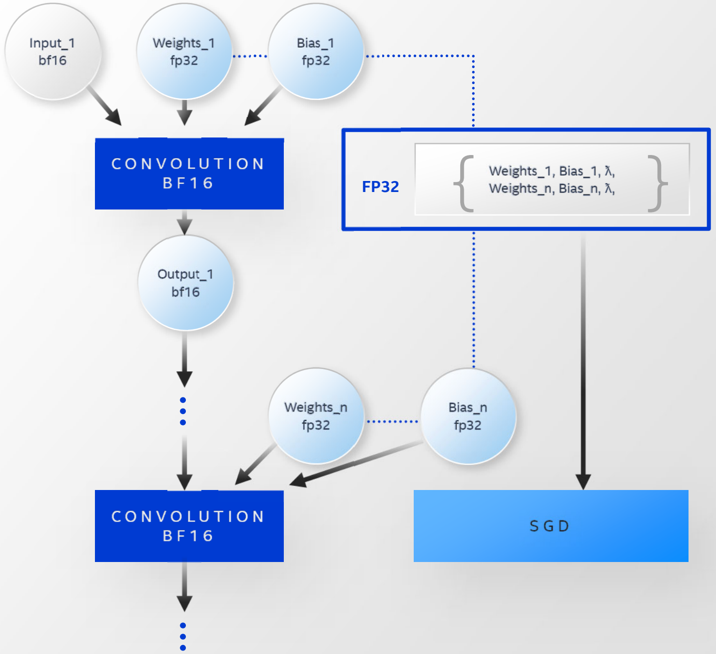
ONE LINE OF CODE CHANGE

```
import torch
import torchvision.models as models

def bf16_inference(model, x):
  x=x.to(dtype=torch.bfloat16)
   return model(x)

def bf16_train(model,
dataset, num_epochs, trained_weights):
```

FRAMEWORK INTEGRATION



Input_1
bf16

Weights_1
fp32

Bias_1
fp32

CONVOLUTION
BF16

Output_1
bf16

FP32

{ Weights_1, Bias_1, ƛ,
Weights_n, Bias_n, ƛ, }

Weights_n
fp32

Bias_n
fp32

CONVOLUTION
BF16

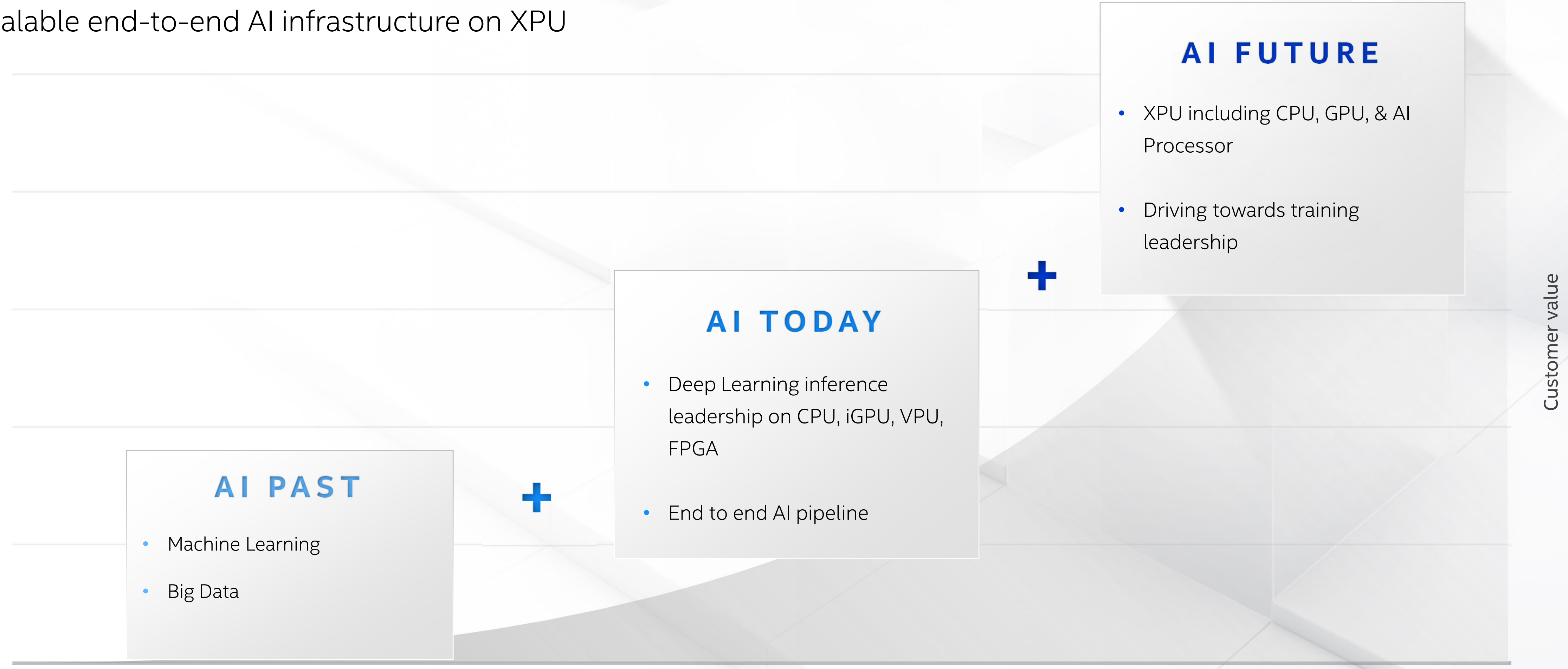
SGD

oneDNN INTEGRATION

- Same op coverage as FP32
- Provides reorders for all needed tensor formats (BF16 <=> FP32)

Intel AI Software Roadmap

Scalable end-to-end AI infrastructure on XPU



AI PAST

- Machine Learning
- Big Data

+

AI TODAY

- Deep Learning inference leadership on CPU, iGPU, VPU, FPGA
- End to end AI pipeline

+

AI FUTURE

- XPU including CPU, GPU, & AI Processor
- Driving towards training leadership

oneAPI Stack

SERVICES & SOLUTIONS

APPLICATIONS

oneAPI

DIRECT PROGRAMMING LANGUAGES

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

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 Com Library
- Intel® Video Processing Library
- Intel® oneAPI Deep Neural Network Library
- Intel® MPI Library
- Intel® Integrated Performance Primitives

MIGRATION TOOLS

- Compatibility Tool

ANALYSIS & DEBUG TOOLS

- Intel® VTune™ Profiler
- Intel*® Advisor
- GDB*
- Intel® Inspector
- Intel® Trace Analyzer & Collector
- Intel® Cluster Checker

SYSTEM PROGRAMMING

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

CPU
SCALAR

GPU
VECTOR

AI
MATRIX

FPGA
SPATIAL/FF

INTERCONNECT

MEMORY

NEW DISCLOSURE

oneAPI Roadmap

oneAPI

Beta 3

- DPC++ compiler
- Libraries
- Tools
- Developer-Domain Toolkits
- CPU, iGPU and FPGA

NOV 2019

oneAPI

Beta 8

NEW HARDWARE SUPPORT

NOW

oneAPI GOLD

PRODUCTION QUALITY AND PERFORMANCE

SPEC V1.0

2H'20

NEW DISCLOSURE

intel DevCloud

NO DOWNLOADS | **NO HARDWARE ACQUISITION** | **NO INSTALLATION** | **NO SET-UP & CONFIGURATION**

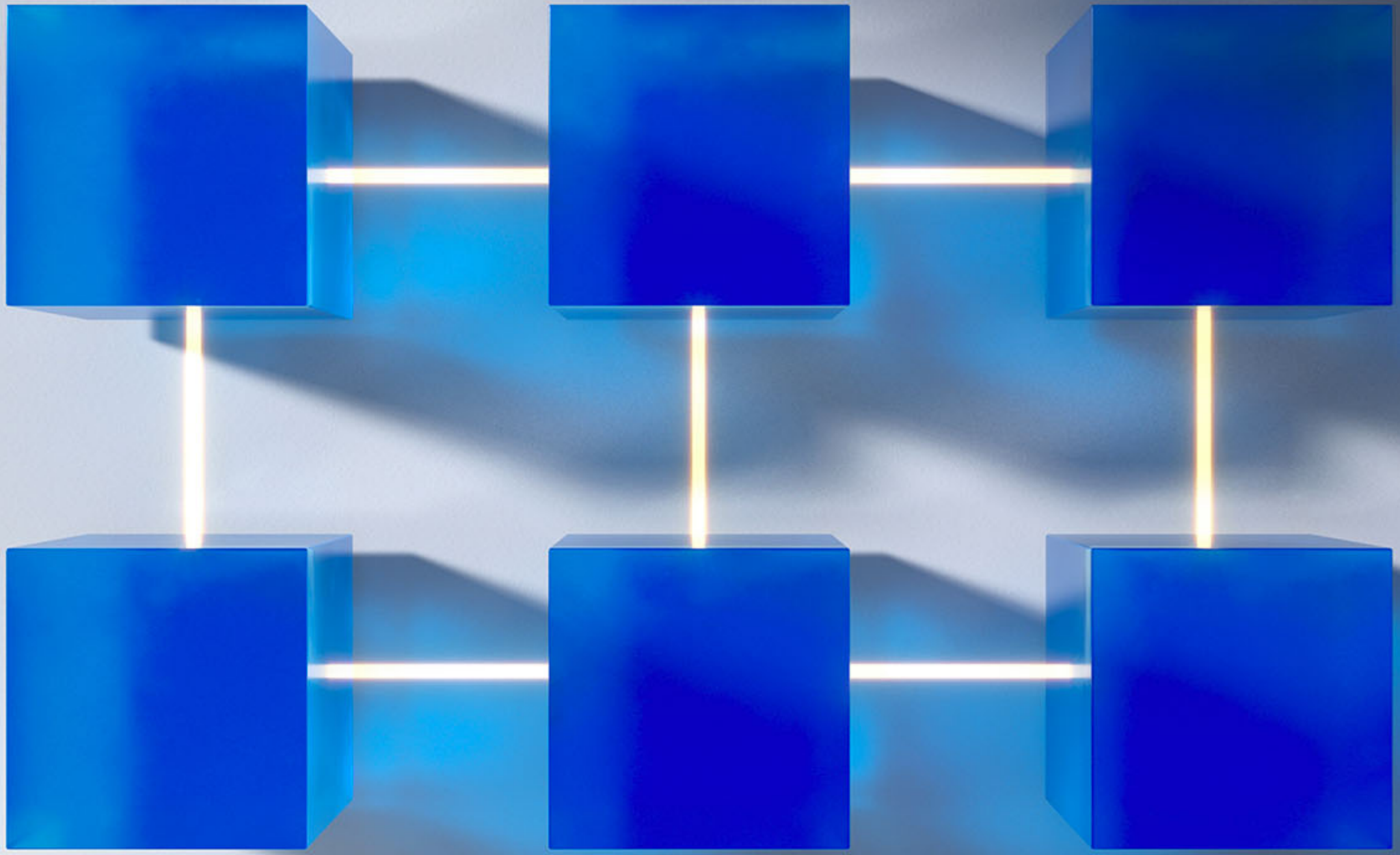
Hardware Availability

CPU

FPGA

dGPU
DG1

For Early Access Developers



Data Center Vision

Sailesh Kottapalli

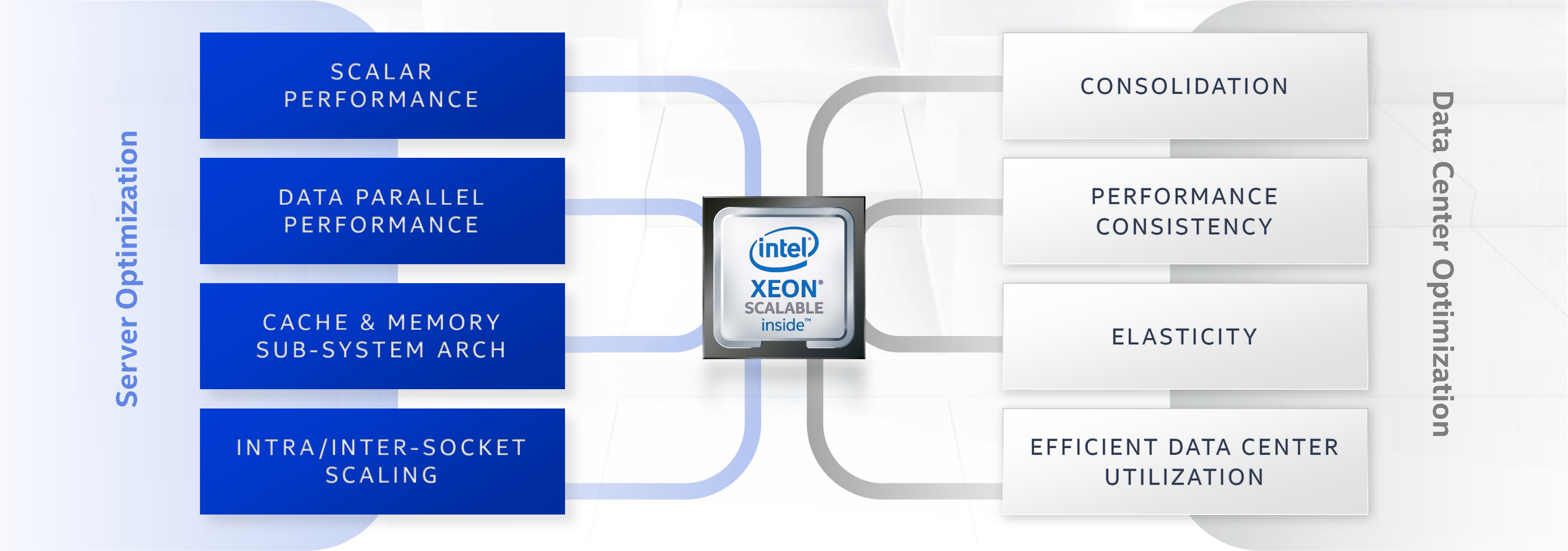
Intel Sr. Fellow, Chief Architect,
Datacenter Processor
Architecture

intel | TECHNOLOGY PILLARS

Architecture Day 2020

Compute Architecture

Best Latency & Throughput



Server Optimization

SCALAR PERFORMANCE

DATA PARALLEL PERFORMANCE

CACHE & MEMORY SUB-SYSTEM ARCH

INTRA/INTER-SOCKET SCALING

CONSOLIDATION

PERFORMANCE CONSISTENCY

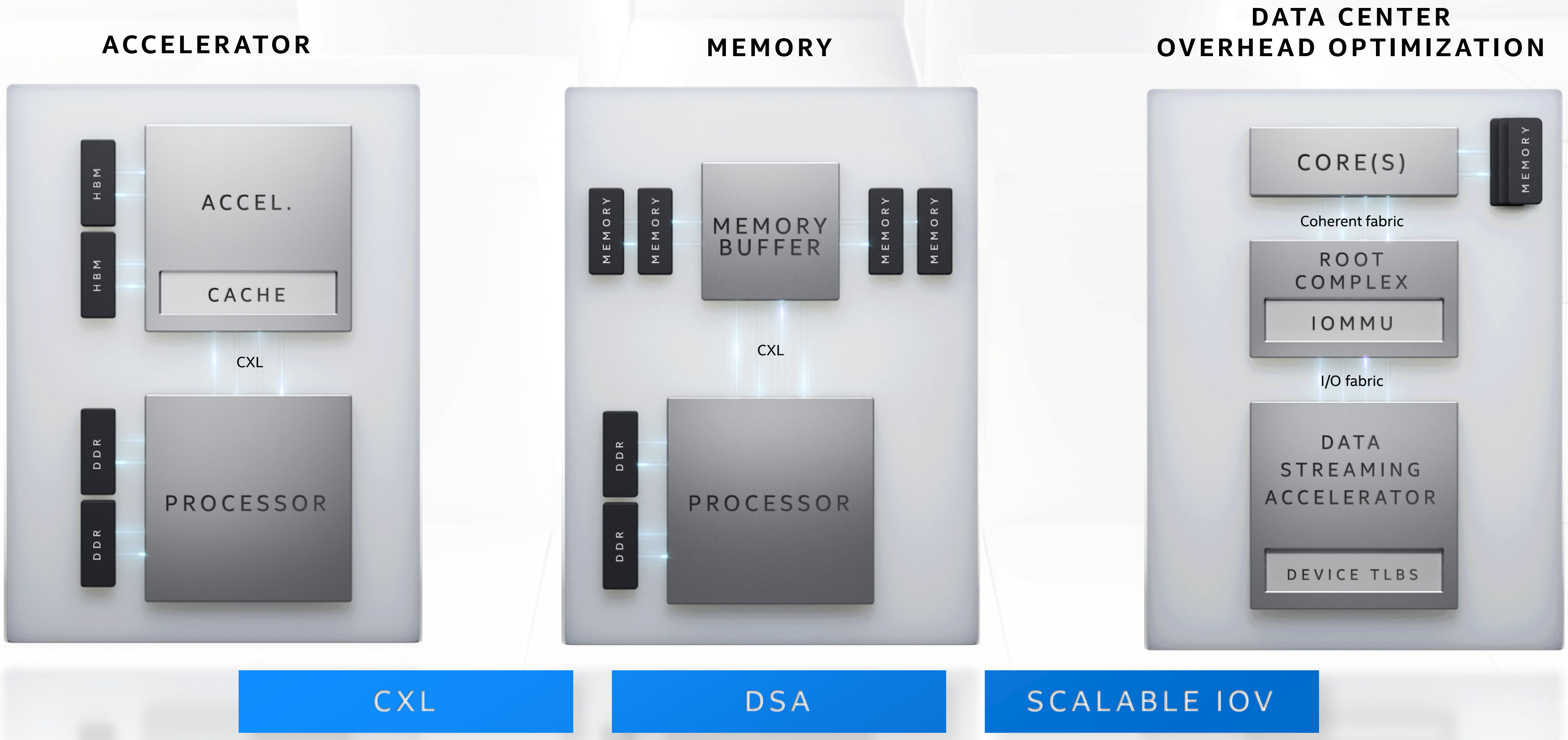
ELASTICITY

EFFICIENT DATA CENTER UTILIZATION

Data Center Optimization

Interconnect

Broad consortium, public specs and Intel HW support



Data Center Processor Roadmap

NEW DISCLOSURE

CASCADE LAKE

2019

COOPER LAKE

- Intel DL Boost
- BFLOAT16

NOW

ICE LAKE

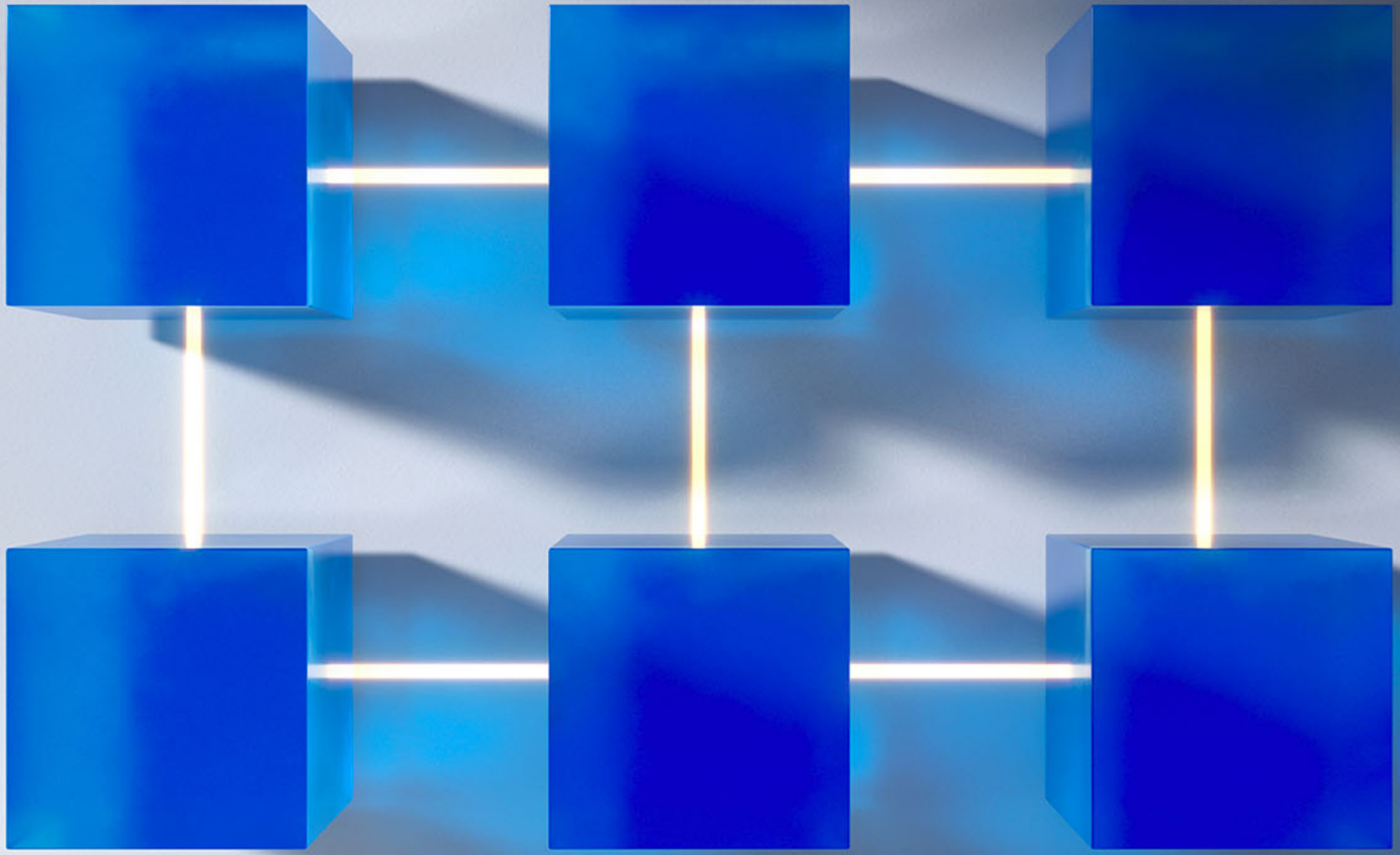
- TME
- PCIe Gen 4
- 8 memory channels
- Crypto accel.
- ICX-D

2H'20

SAPPHIRE RAPIDS

- DDR5
- PCIe Gen 5
- CXL 1.1
- AMX
- DSA

2021



Data Intelligence Client

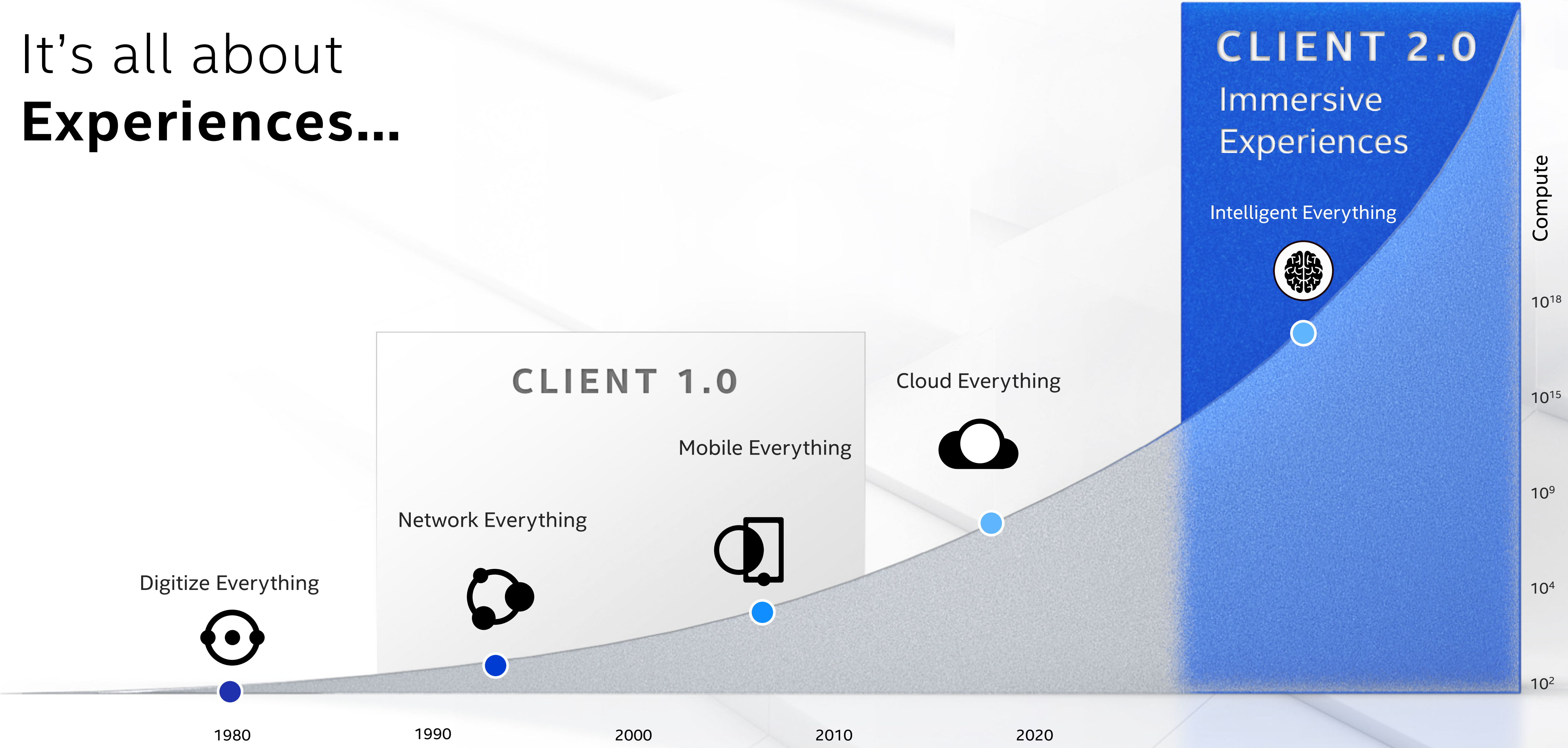
Brijesh Tripathi

Vice President,
Chief Technology Officer
Client Computing Group

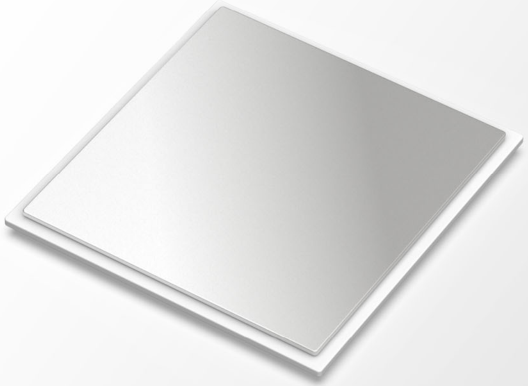
TECHNOLOGY
PILLARS

Architecture Day **2020**

It's all about
Experiences…

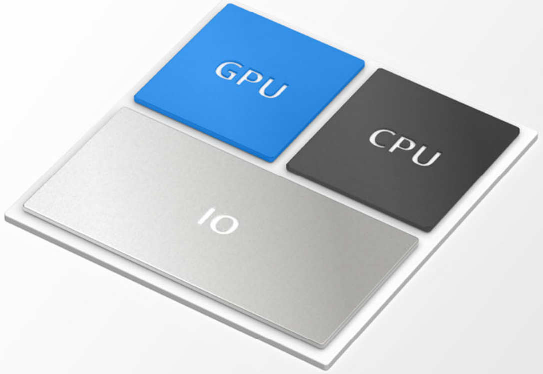


IP/SOC Methodology Change



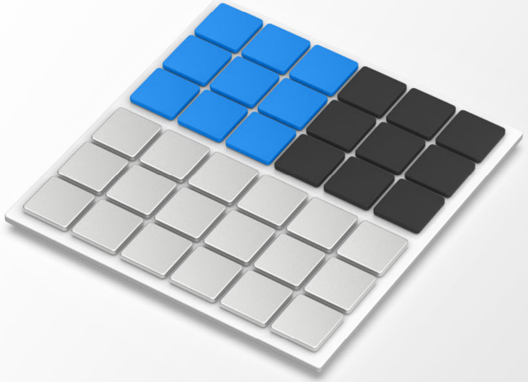
Monolithic | Integrated SOC

- Verified at SOC level
- **3-4 years** of Dev Time
- **100s of bugs** found in Silicon
- No reuse



Multiple Dies | in optimal process

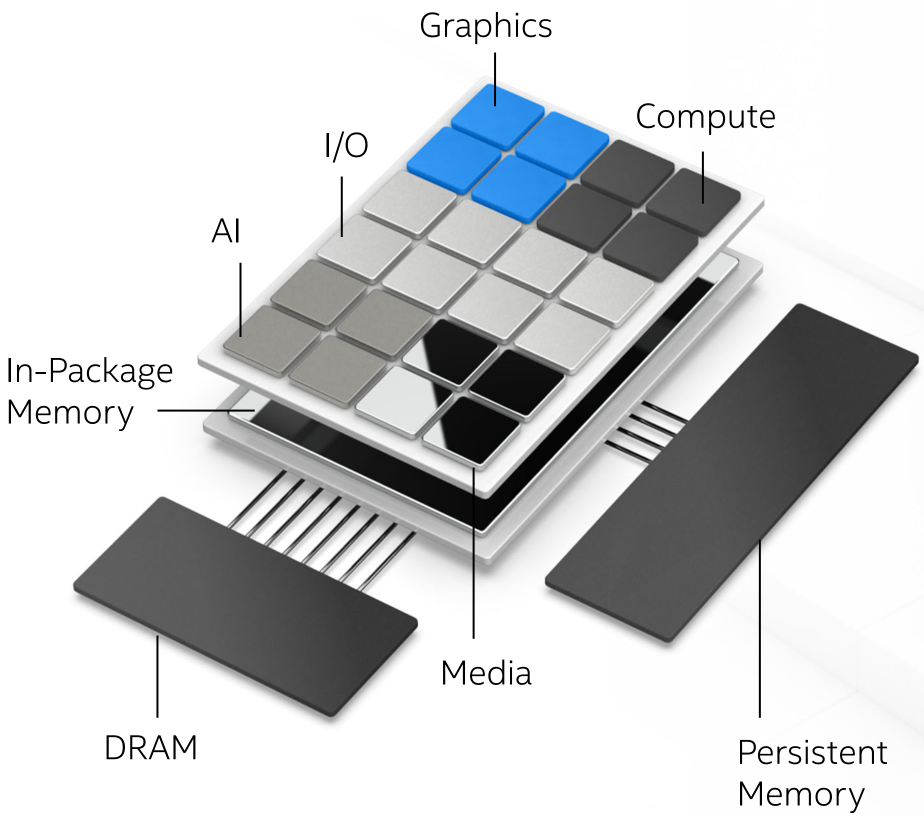
- Verified at IP level
- **2-3 years** of Dev Time
- **10s of bugs** found in Silicon
- Some reuse



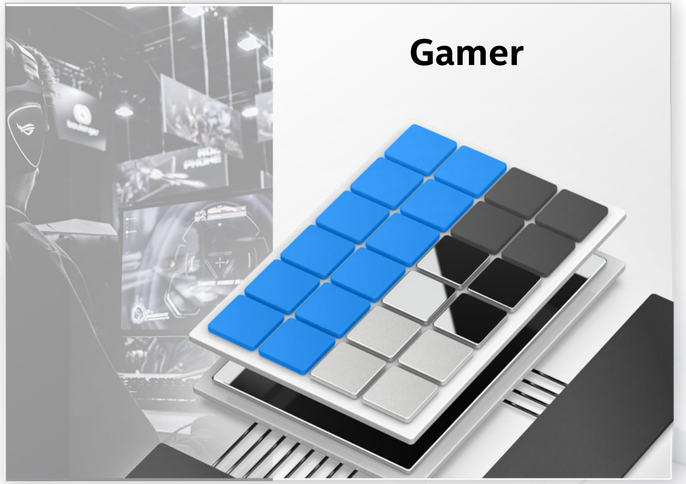
Individual IPs | in optimal process

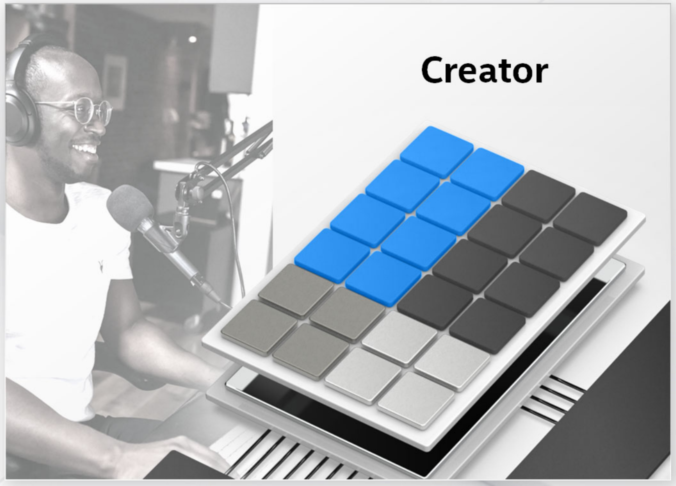
- Verified at IP/Chiplet level
- **1 year** of Dev Time
- **<10 bugs** found in silicon
- Significant reuse

Purpose Built Client



Long Term Vision





Client 2.0

EXPERIENCE FIRST

- Uniquely built to differentiate key experiences
- Uncompromised performance to deliver seamless experience

SCALABLE

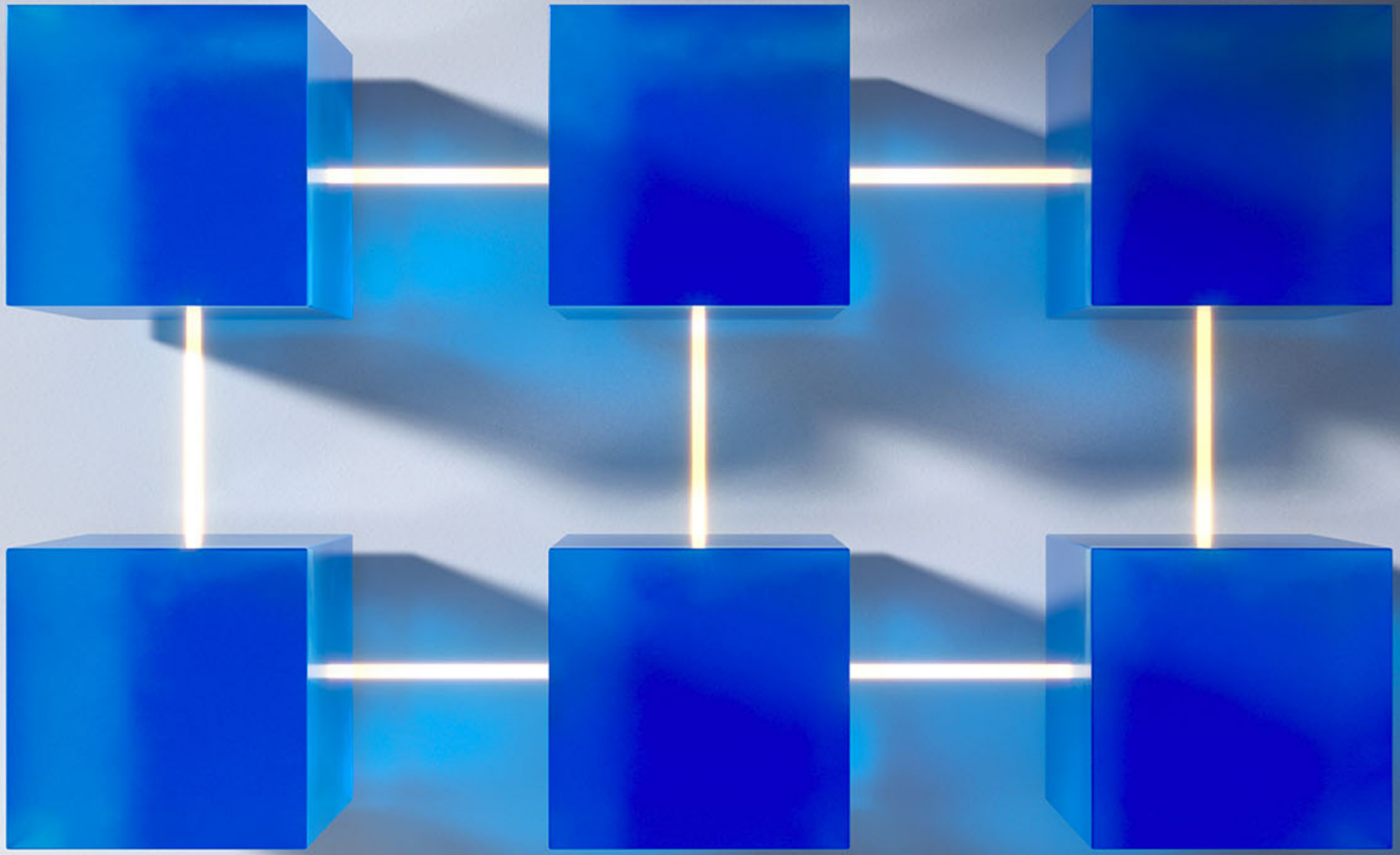
- Ability to mix and match function/IPs to processes
- React faster with reuse and enable faster TTM
- Developer Friendly Platforms

ENERGY EFFICIENT

- KEI's can benefit from using “Efficient” Cores in SOC Chiplet
- Pick low power technology for IOs

OPTIMAL USE OF MOORE’S LAW

- Focus on Performance for General Purpose Compute (CPU)
- Focus on Density for scalable compute (GPU, AI etc.)



Intel Labs

Rich Uhlig

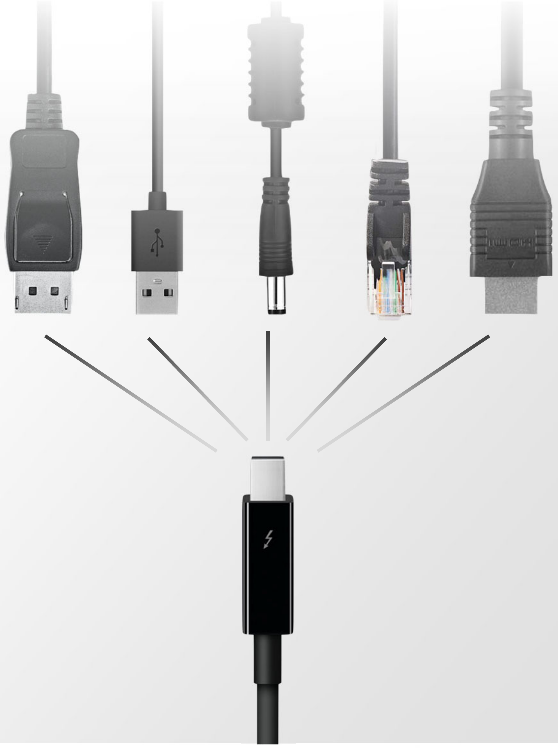
Intel Sr. Fellow, Vice President
Director, Intel Labs

TECHNOLOGY
PILLARS

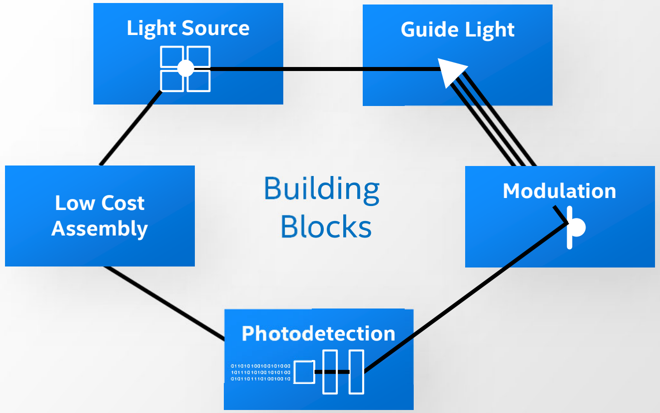
Architecture Day **2020**

Notable Intel Labs Graduates

USB & THUNDERBOLT



SILICON PHOTONICS



Light Source

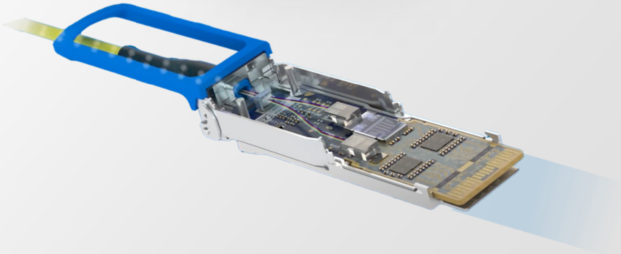
Guide Light

Low Cost Assembly

Building Blocks

Modulation

Photodetection



VIRTUALIZATION

CPU Virtualization
Virtual Machine Extensions (VMX)

Memory Virtualization
Extended Page Tables
Cache and Memory QoS

IO Virtualization
IOMMU
Scalable IOV
Data Streaming Accelerator

Intel® Virtualization Technology

Intel Labs Research today…

Key Research Focus Areas

TRUST, SECURITY, & PRIVACY

EMERGING WORKLOADS

NEXT WAVE OF AI

FUTURE OF SOFTWARE

NOVEL SENSING TECH.

NEW COMPUTE MODELS

ADVANCED MEMORY & STORAGE

NEXT-GEN CONNECTIVITY, 5G AND BEYOND

EFFICIENT DESIGN & PROGRAMMING

Intel Labs Research today…

Key Research Focus Areas

Secure Federated Learning

DARPA GARD

Ambient Computing

OMICS

Knowledgeable AI

AI Ethics

Machine Programming

4D RADAR

Neuromorphic vision

Quantum

DARPA HIVE

Cryptographic

Neuromorphic

Soft Memory

Compute Near Memory

Agile HW/SW Co-Design

DARPA IDEA

Integrated Photonics

Wireless Chip-to-Chip

Intel Labs Research today…

Key Research Focus Areas

Secure Federated Learning

DARPA GARD

Ambient Computing

OMICS

Knowledgeable AI

AI Ethics

Machine Programming

4D RADAR

Neuromorphic vision

Quantum

DARPA HIVE

Cryptographic

Neuromorphic

Soft Memory

Compute Near Memory

Integrated Photonics

Wireless Chip-to-Chip

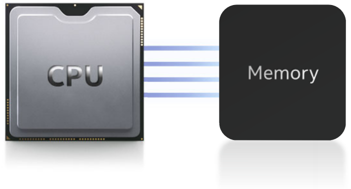
Agile HW/SW Co-Design

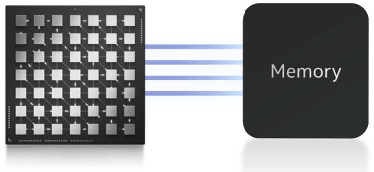
DARPA IDEA

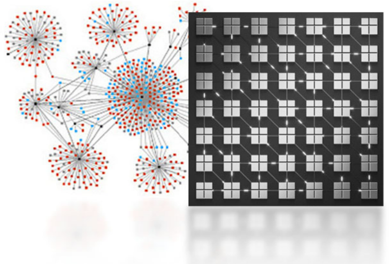
How can we improve Compute Efficiency by 100x – 1000x?

… by thinking differently

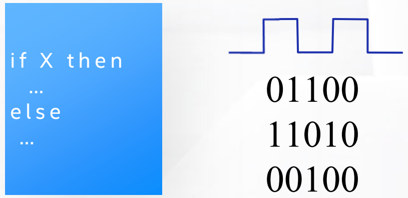
… by thinking differently

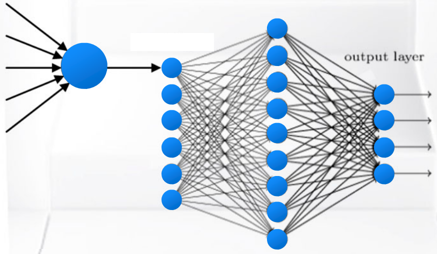


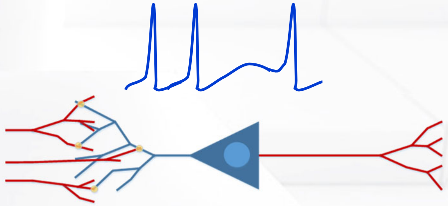




von Neumann Computing	Deep Learning (DNNs)	Neuromorphic Computing
PROGRAMMING BY ENCODING ALGORITHMS	OFFLINE TRAINING USING LABELED DATASETS	LEARN ON THE FLY THROUGH NEURON FIRING RULES
SYNCHRONOUS CLOCKING	SYNCHRONOUS CLOCKING	ASYNCHRONOUS EVENT-BASED SPIKES
SEQUENTIAL THREADS OF CONTROL	PARALLEL DENSE COMPUTE	PARALLEL SPARSE COMPUTE

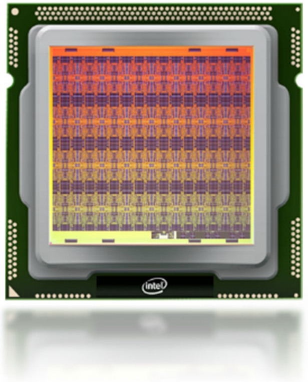




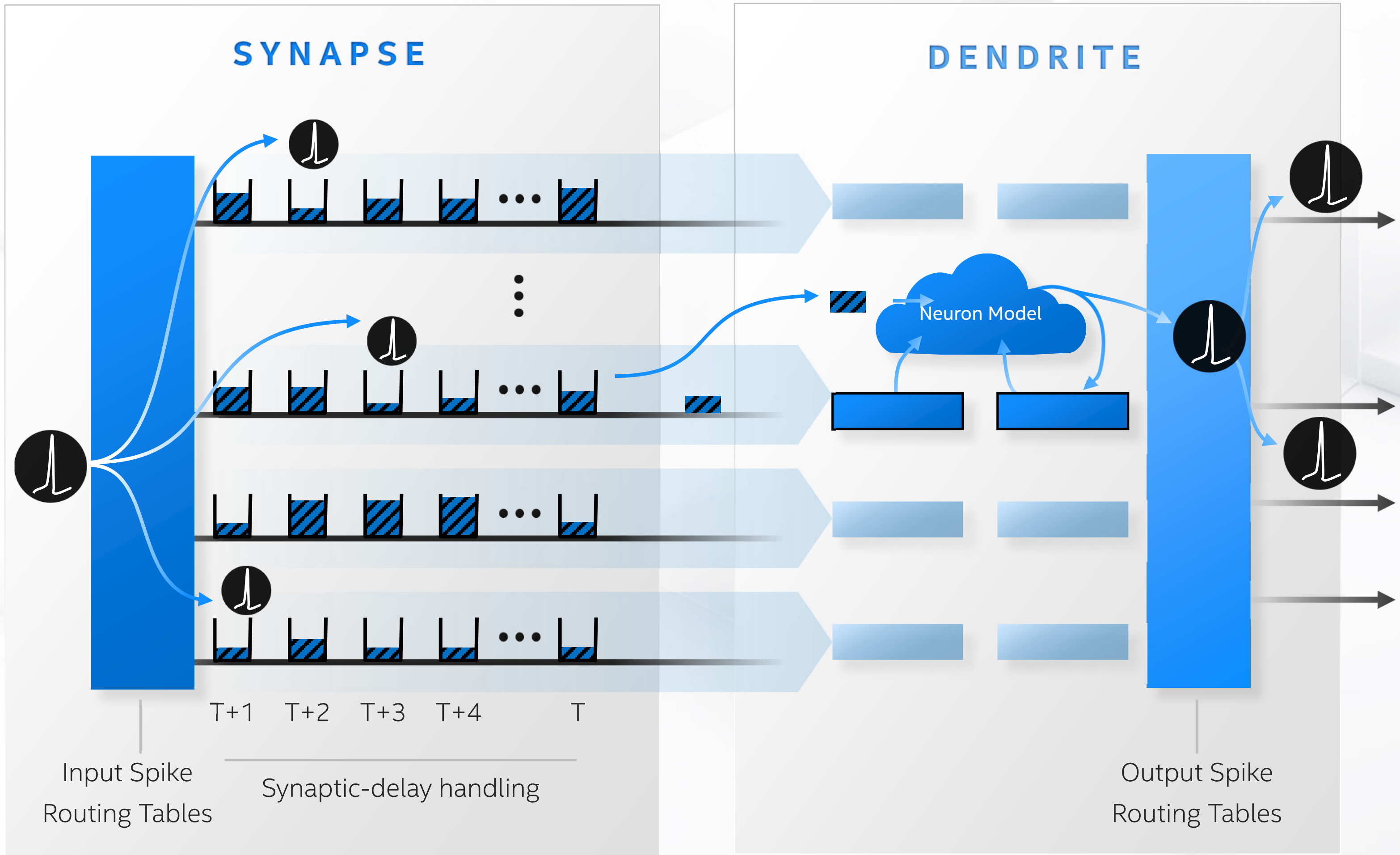


Spiking Neural Network (SNN) – Operation

INTEL LOIHI



- 128k temporal spiking neurons
- Integrated memory + compute architecture
- On-chip learning at milliwatt power levels
- Fully digital and asynchronous



An Example Usage - Sparse Coding

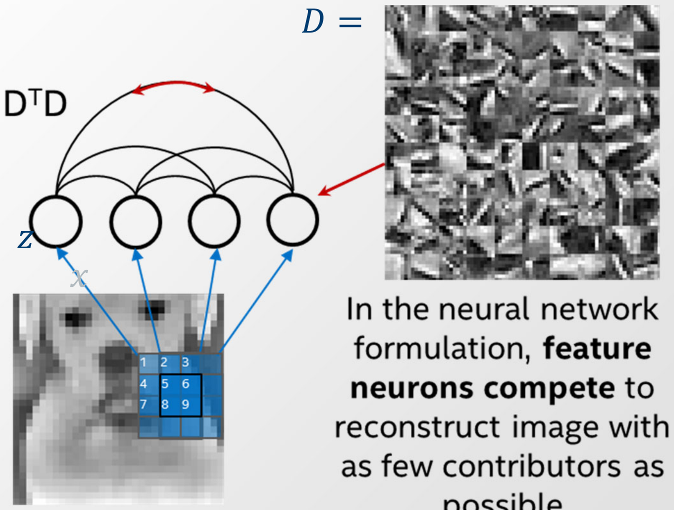
LASSO SPARSE CODING

$$\underset{z}{\operatorname{argmin}}\|x - Dz\|_2^2 + \lambda\|z\|_1$$

Input

Reconstruction

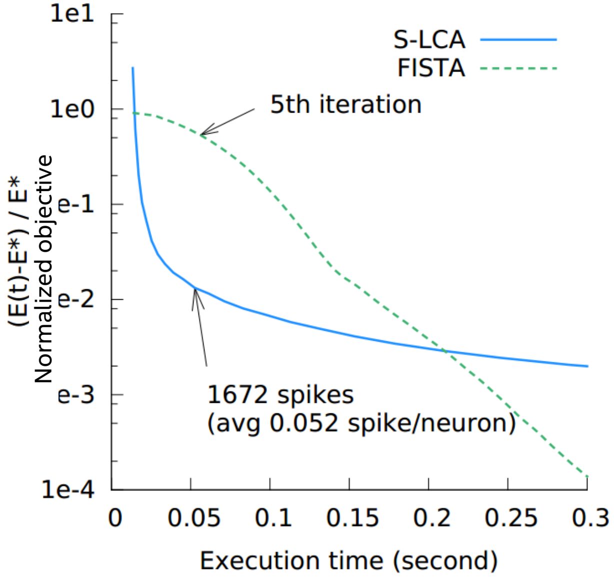
Sparse regularization

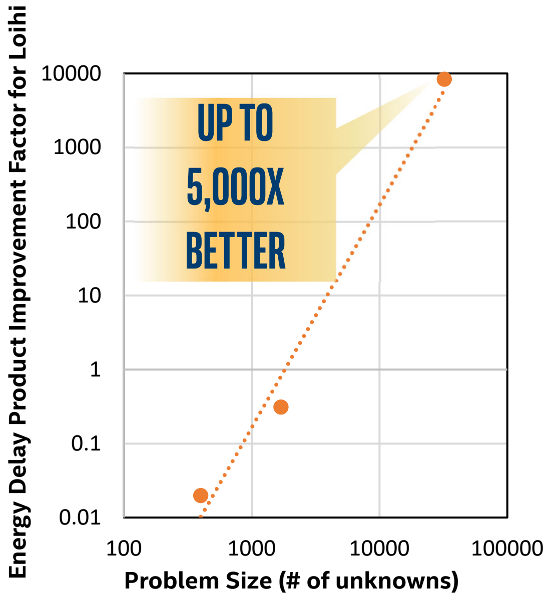


In the neural network formulation, **feature neurons compete** to reconstruct image with as few contributors as possible

Tang et al., arxiv: 1705:05475

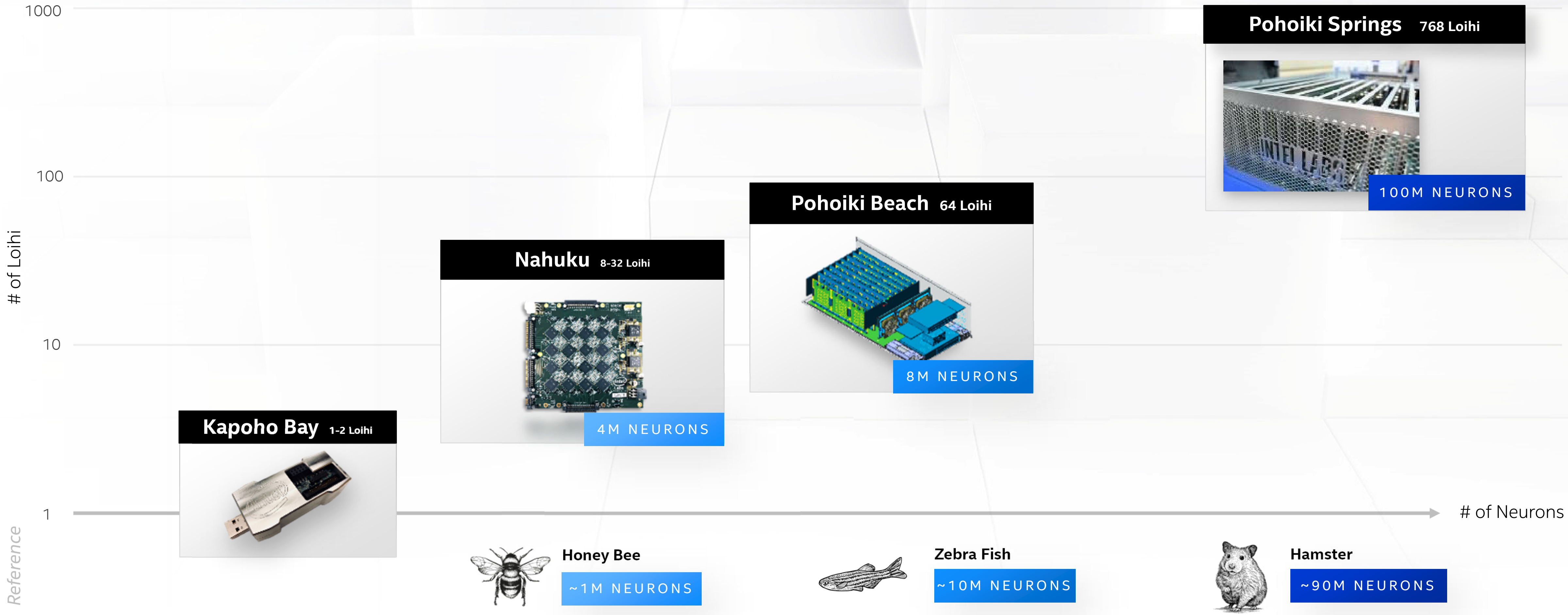
ENERGY-EFFICIENCY COMPARISON





Energy-Delay-Product Advantage

Neuromorphic Computing



Prototypes & Community

NEUROMORPHIC RESEARCH COMMUNITY

101
Engaged Groups

103
Active Projects

11
Workshops →

500+
researchers Reached

45
Publications by Community Members

Nx SDK FEATURES

Connectivity & Real Time Performance
- < 1ms command & control loopback
- Real time, event based gesture recognition

Accessibility & Extensibility
- Development Cloud
- 3rd Party Support
- Deep Network Model Support

Modularity and Composability
- Community Developed Modules
- Large Scale network connectivity

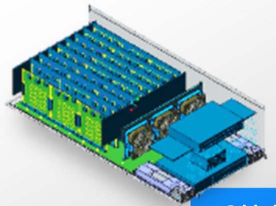
Large Scale Enhancements
- Core level Energy Management and measurement
- Large Scale Compiler and Execution Enhancements
- Simulator Support

Kapoho Bay 1-2 Loihi

Nahuku 8-32 Loihi

4M NEURONS

Pohoiki Beach 64 Loihi



8M NEURONS

Pohoiki Springs 768 Loihi

100M NEURONS

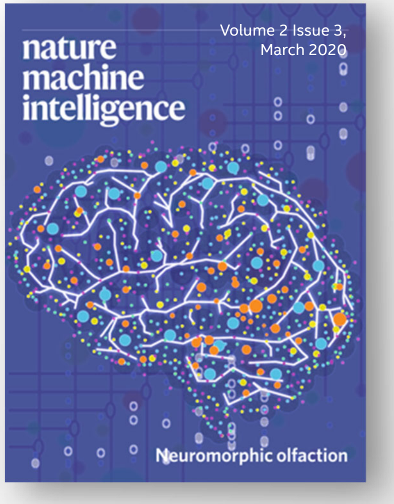
Example Applications for Loihi

SEARCHING

k-NN nearest neighbor search:
100x faster in O(1) index build

[EP Frady et al, "Neuromorphic Nearest-Neighbor Search Using Intel's Pohoiki Springs." arXiv:2004.12691]

LEARNING


nature machine intelligence

Volume 2 Issue 3,
March 2020

Neuromorphic olfaction

Rapid online learning and **robust recall** in a neuromorphic olfactory circuit

https://www.nature.com/articles/s42256-020-0159-4

SENSING


engadget

Researchers built robotic skin with a sense of touch

It's powered by Intel's brain-mimicking neuromorphic chip Loihi.

Faster real-time sensory data processing with **45x less power**

http://www.roboticsproceedings.org/rss16/p020.pdf

PURPOSE

To create **world-changing technology** that **enriches** the **lives** of every person on **earth**

VISION

We are on a journey to be the trusted **performance** leader that unleashes the potential of **data**

TECH VISION

Exascale for Everyone, leading us into **the Intelligence era,** driven by **100B intelligent connected devices**

THE DATA PROBLEM

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

PROCESS & PACKAGING

- Intel 10nm Superfin
- Intel 10nm Enhanced SuperFin
- Foveros
- Hybrid Bonding
- Co-EMIB
- ODI

ARCHITECTURE

- Willow Cove
- Tiger Lake
- Alderlake Performance Hybrid
- 224G X-ceiver
- X^e HPG
- X^e LP Unpacked
- X^e HP Preview
- DG1 / SG1
- Agilex

MEMORY

- 144L QLC 3D NAND
- 4-Deck 3D Xpoint
- Optane™ Persistent Mem.
- Rambo Cache

INTERCONNECT

- 5G Portfolio
- Client Connectivity
- Photonics
- SmartNIC

SOFTWARE

- Intel AI Software
- BFLOAT16
- oneAPI
- DG1 in Devcloud

SECURITY

- Foundational Security
- Workload Protection
- Software Reliability

DATA CENTER VISION

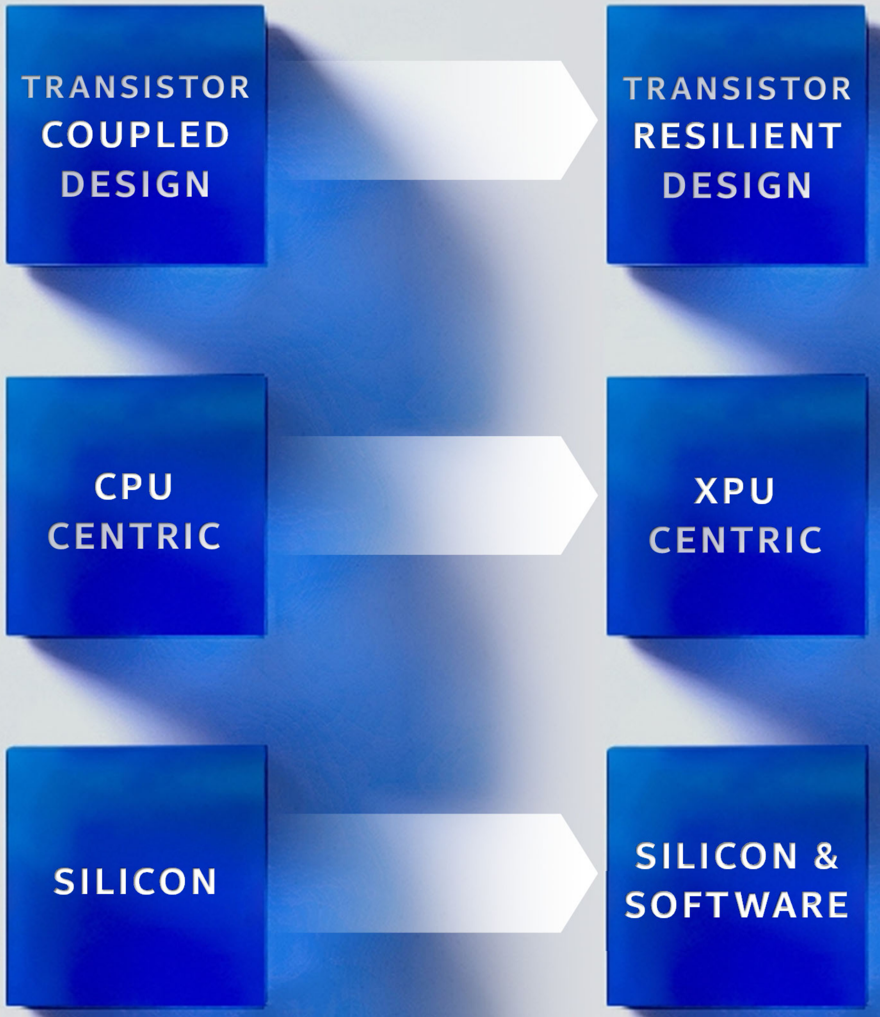
- Data Center Roadmap
- Ice Lake Xeon

INTEL LABS

- Neuromorphic Unpacking

CLIENT VISION

- IP/SOC Methodology Change
- Client 2.0

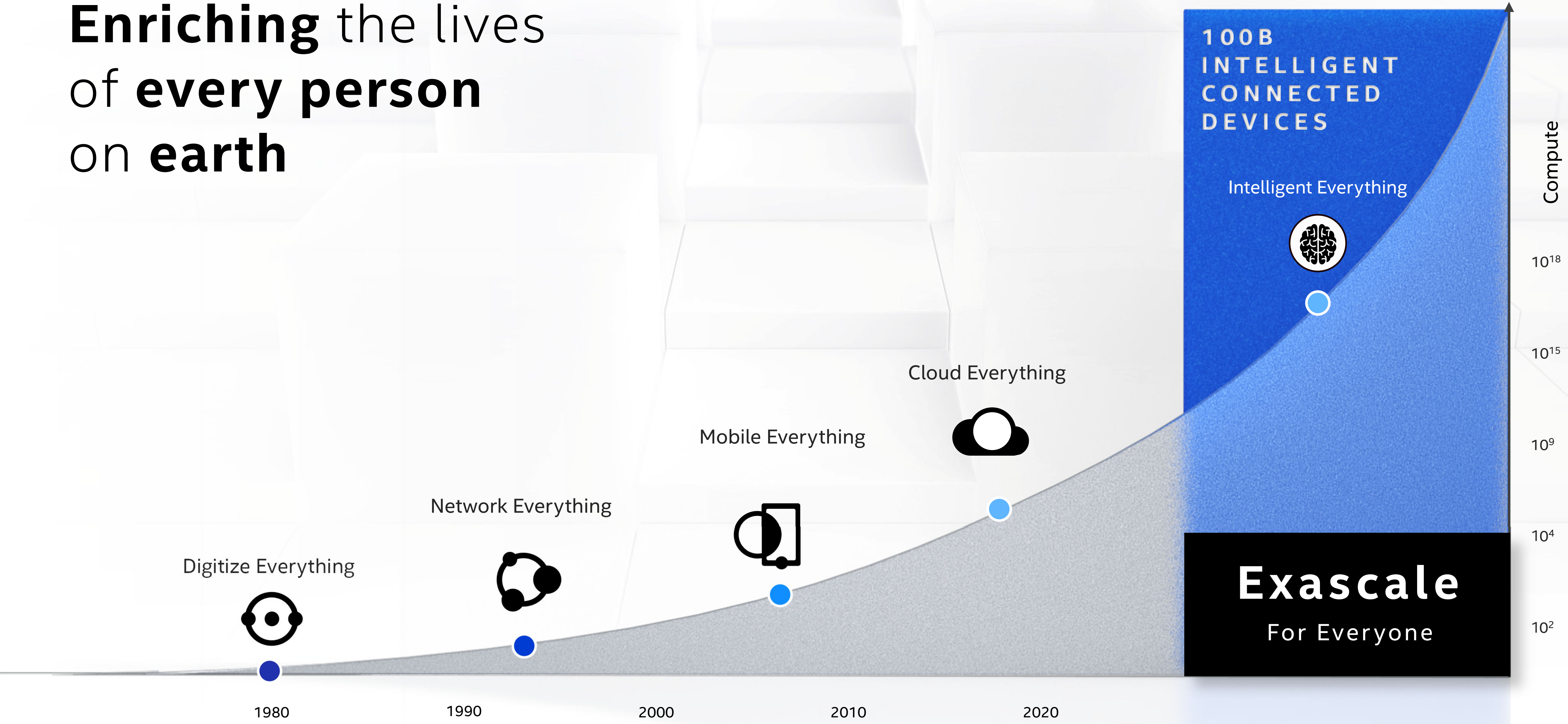


DESIGN: TRANSISTOR COUPLED DESIGN → TRANSISTOR RESILIENT DESIGN

ARCHITECTURE: CPU CENTRIC → XPU CENTRIC

SOLUTION: SILICON → SILICON & SOFTWARE

Enriching the lives of **every person** on **earth**



100B INTELLIGENT CONNECTED DEVICES

Exascale

For Everyone

PURPOSE

To create **world-changing technology** that **enriches** the **lives** of every person on **earth**

VISION

We are on a journey to be the trusted **performance** leader that unleashes the potential of **data**

TECH VISION

Exascale for Everyone, leading us into **the Intelligence era,** driven by **100B intelligent connected devices**

THE DATA PROBLEM

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

PROCESS & PACKAGING

- Intel 10nm Superfin
- Intel 10nm Enhanced SuperFin
- Foveros
- Hybrid Bonding
- Co-EMIB
- ODI

ARCHITECTURE

- Willow Cove
- Tiger Lake
- Alderlake Performance Hybrid
- 224G X-ceiver
- X^e HPG
- X^e LP Unpacked
- X^e HP Preview
- DG1 / SG1
- Agilex

MEMORY

- 144L QLC 3D NAND
- 4-Deck 3D Xpoint
- Optane™ Persistent Mem.
- Rambo Cache

INTERCONNECT

- 5G Portfolio
- Client Connectivity
- Photonics
- SmartNIC

SOFTWARE

- Intel AI Software
- BFLOAT16
- oneAPI
- DG1 in Devcloud

SECURITY

- Foundational Security
- Workload Protection
- Software Reliability

DATA CENTER VISION

- Data Center Roadmap
- Ice Lake Xeon

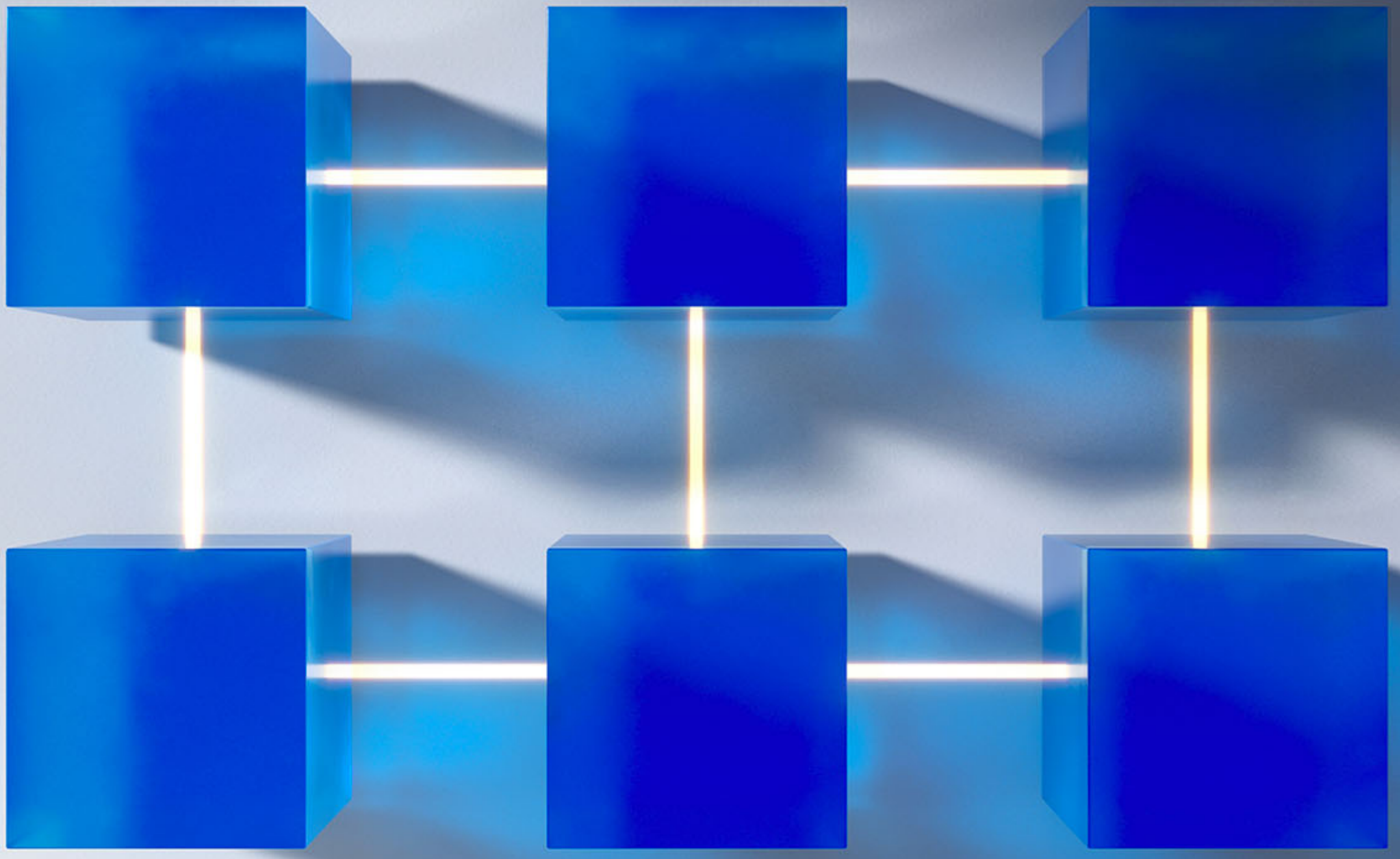
INTEL LABS

- Neuromorphic Unpacking

CLIENT VISION

- IP/SOC Methodology Change
- Client 2.0

Thank you!



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. For testing details and system configurations, please contact your intel representative. 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.

Intel contributes to the development of benchmarks by participating in, sponsoring, and/or contributing technical support to various benchmarking groups, including the BenchmarkXPRT Development Community administered by Principled Technologies.

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.