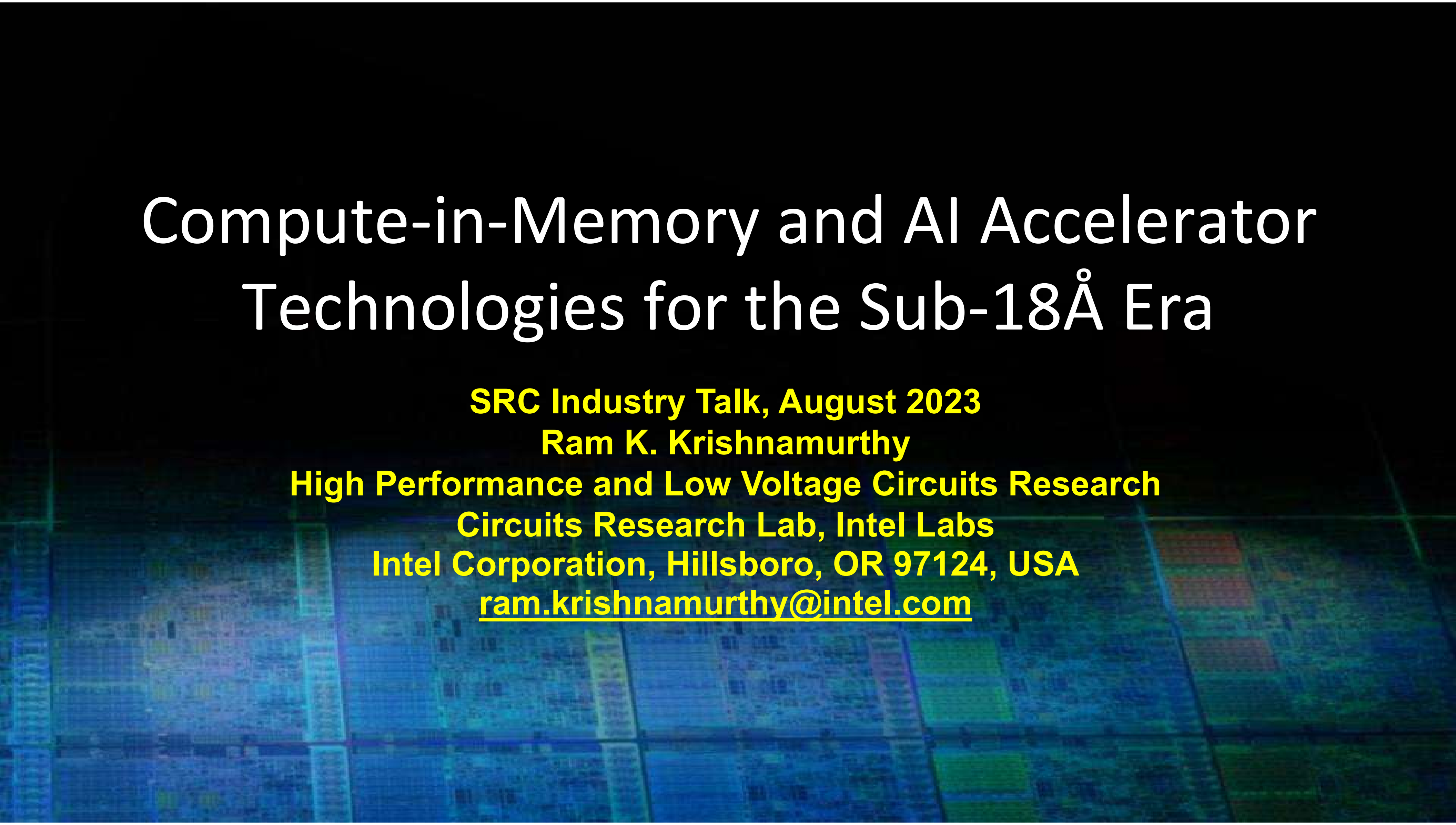
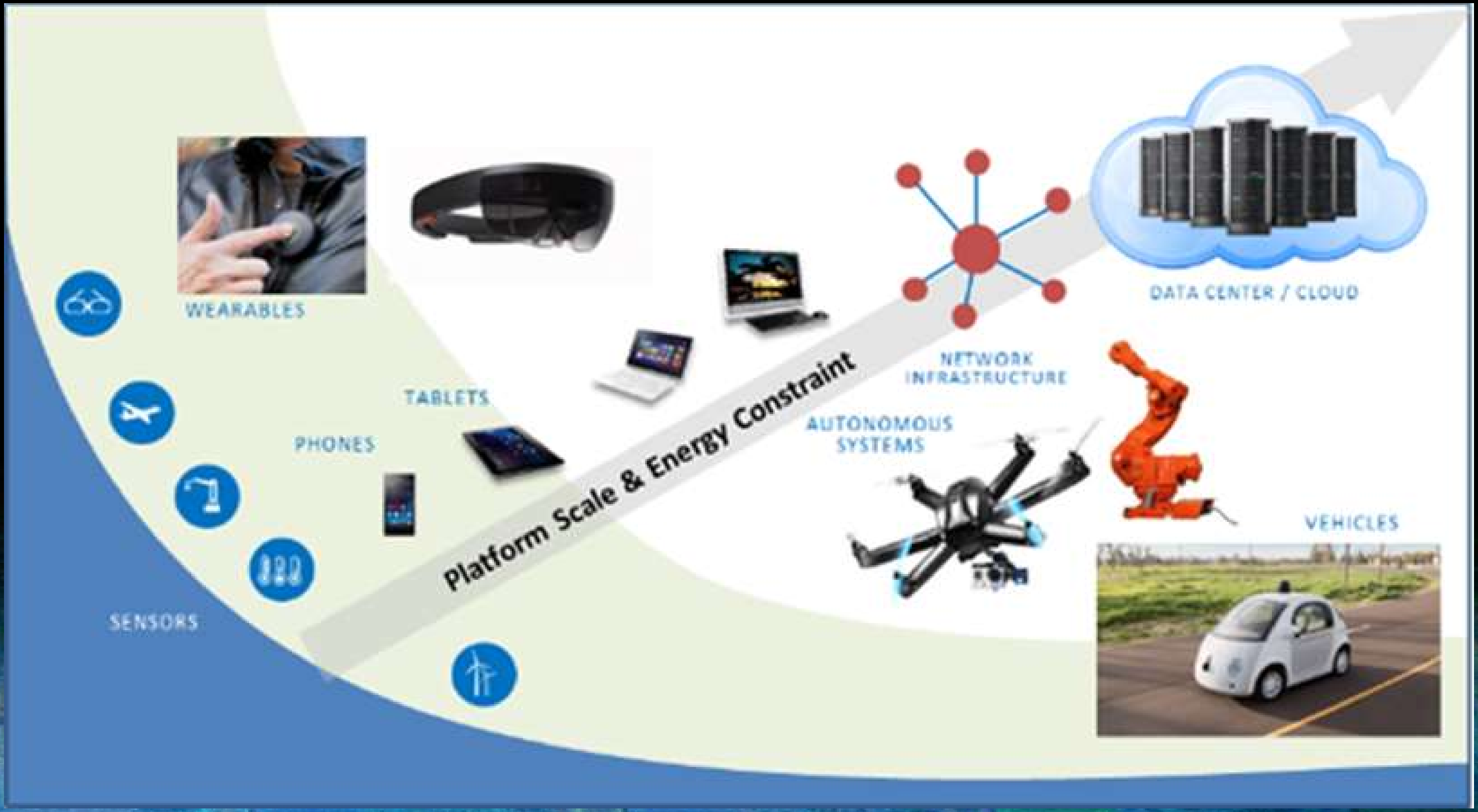


Compute-in-Memory and AI Accelerator Technologies for the Sub-18Å Era

SRC Industry Talk, August 2023

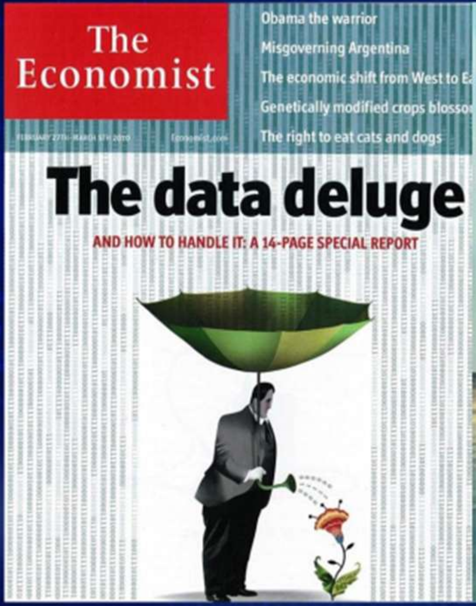
Ram K. Krishnamurthy
High Performance and Low Voltage Circuits Research
Circuits Research Lab, Intel Labs
Intel Corporation, Hillsboro, OR 97124, USA
ram.krishnamurthy@intel.com

Internet of Everything (IoE)



Need end-to-end energy efficiency, ML everywhere

DATA DEFINES THE FUTURE



The Economist

FEBRUARY 27TH–MARCH 5TH 2010 Economist.com

Obama the warrior

Misgoverning Argentina

The economic shift from West to E

Genetically modified crops blosso

The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

The Economist

Crunch time in France

Ten years on: banking after the crisis

South Korea's unfinished revolution

Biology, but without the cells

The world's most valuable resource

Data and the new rules of competition

POPULAR SCIENCE

THE FUTURE NOW

THE CONTROL CENTERS

Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate p32

OFFICER ALGORITHM

Can a Crime Be Prevented Before It Begins? p38

NEW WAYS OF SEEING

A Gallery of Extraordinary Infographics p48

PLUS

Juan Enriquez Reprograms Life

James Gleick Unsplits the Bit

AND

Lawrence Weschler Questions the Cloud

SPECIAL ISSUE

DATA IS POWER

HOW INFORMATION IS DRIVING THE FUTURE

8-PAGE SPECIAL POSTGRADUATE SURVIVAL GUIDE 8TH BIRTHDAY ISSUE!

COSMOS

THE SCIENCE OF EVERYTHING

THE END OF VIOLENCE

Steven Pinker on the new peace p36

DEFEATING POLIO

Will politics jeapordise a cure? p68

FRAUDS AND FAKES

Science's biggest scams p24

GENIUS OF DOGS

Inside the canine brain p41

IS data THE NEW GOD?

How tracking your digital trail could predetermine your future – and why you'll benefit from today's data deluge. p46

ISSUE 51 JUN/JUL 2013

A$10.95 NZ$12.90

COSMOSmagazine.com

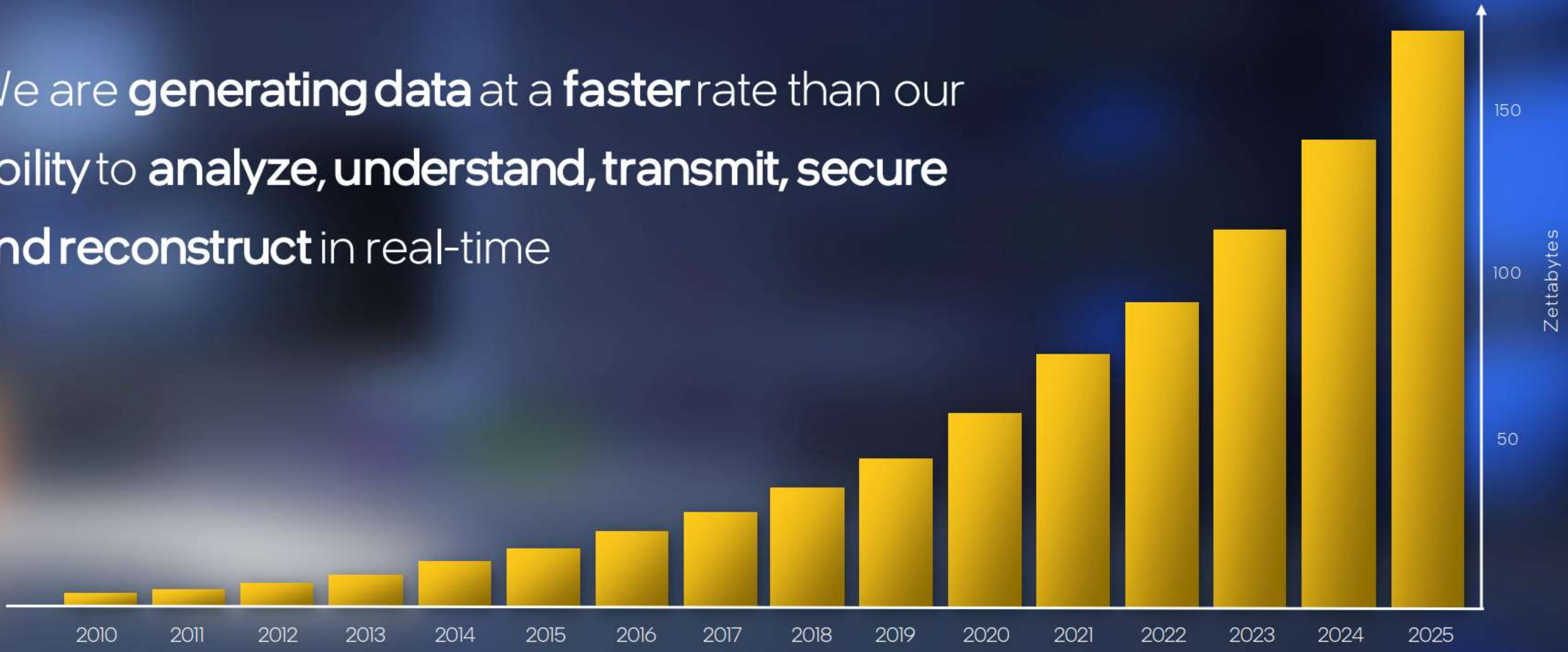
GALAXIES AND NEBULAE ▪ CANCER VACCINES ▪ WHALES ▪ FICTION ▪ REVIEWS

Other names and brands may be claimed as the property of others.

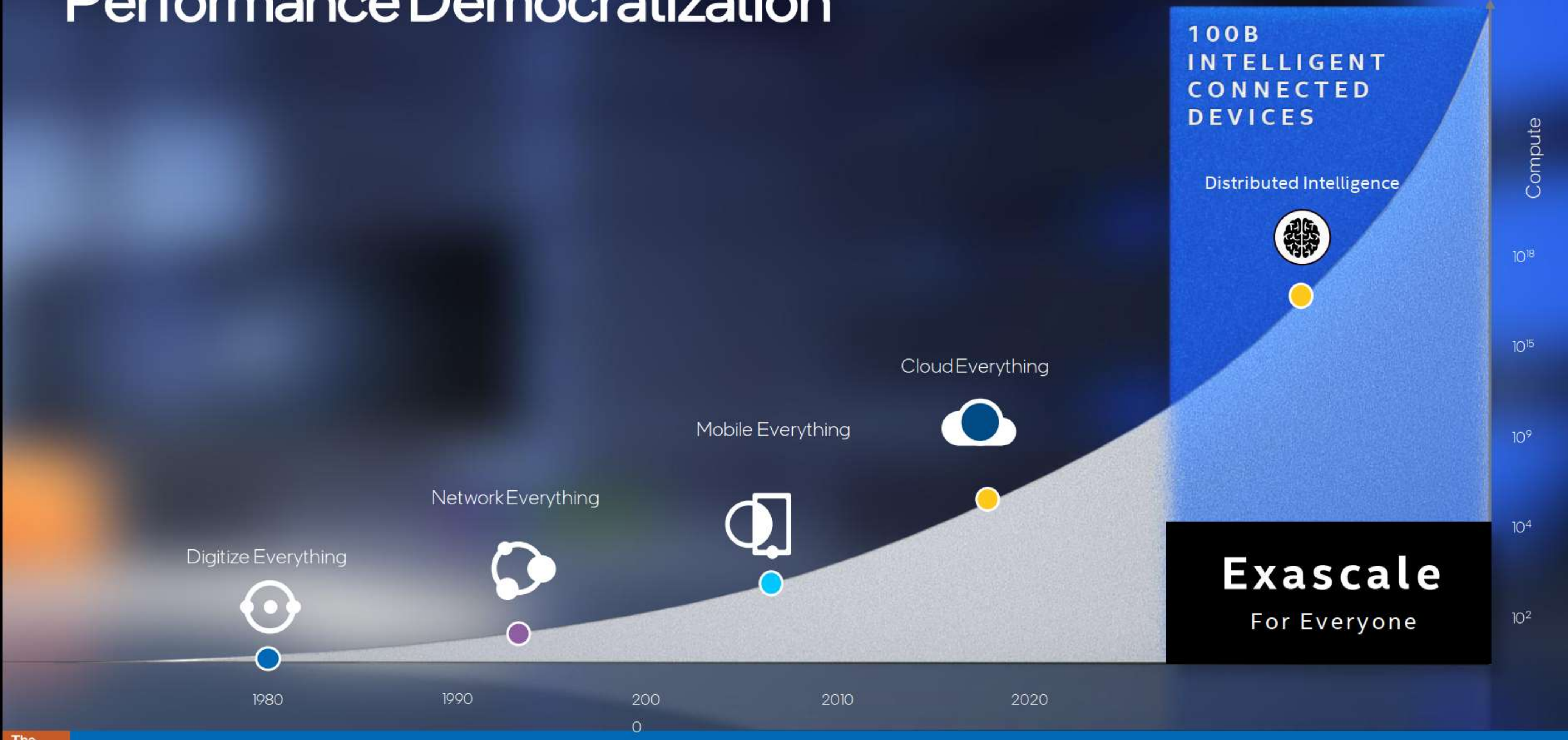
The Data Problem

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

175ZB



Performance Democratization

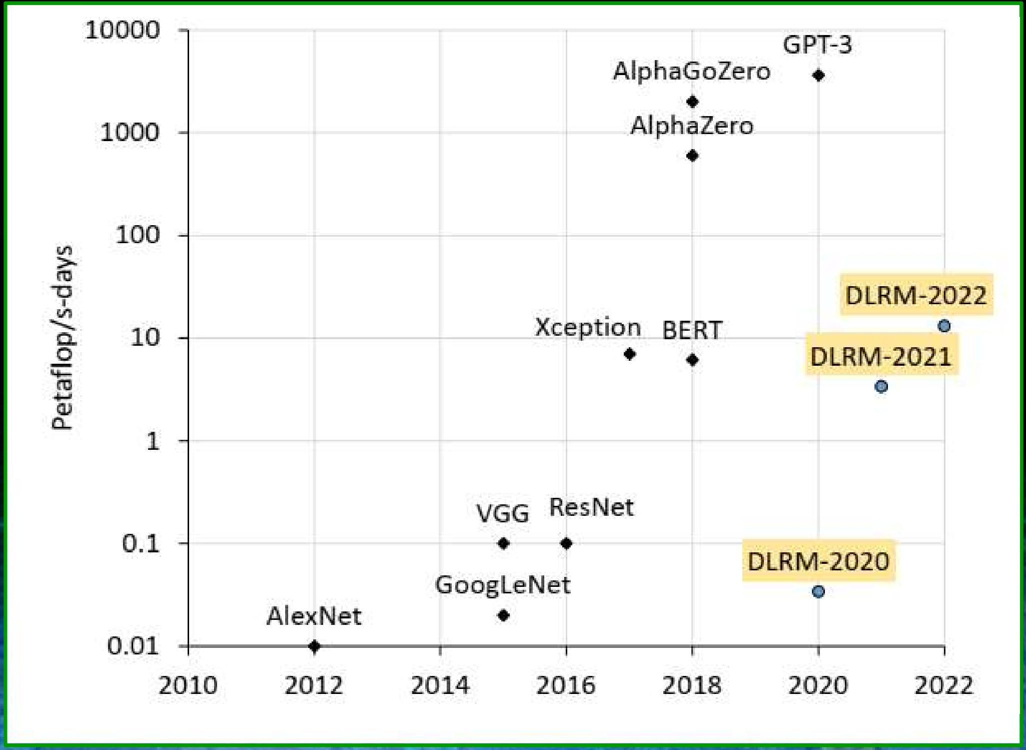


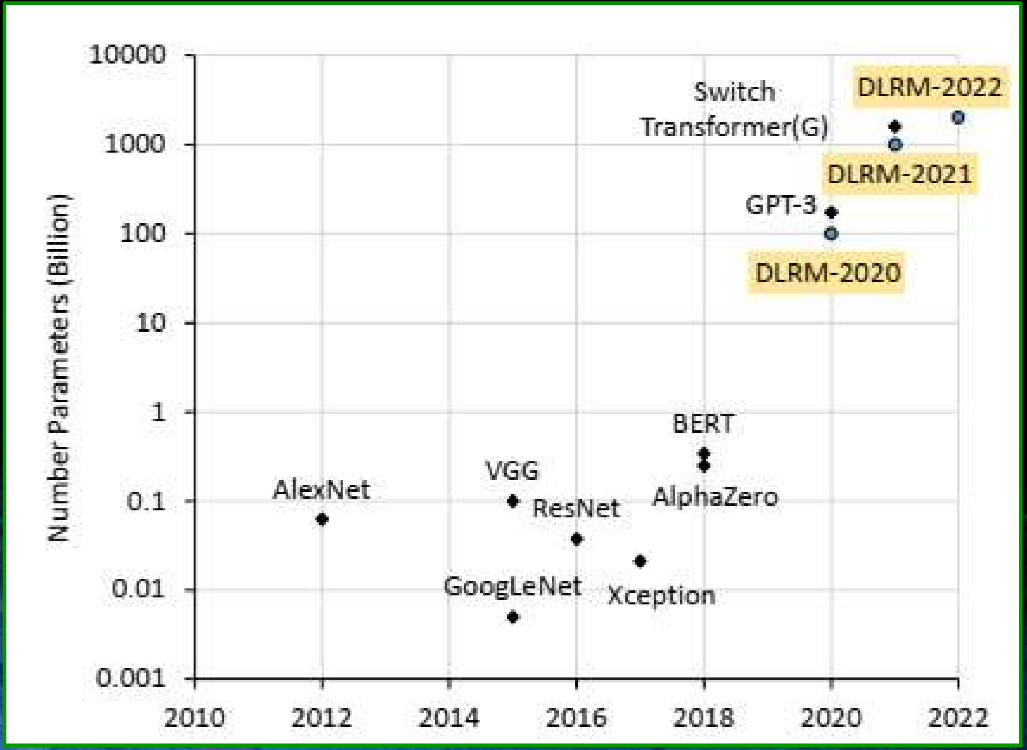
100B INTELLIGENT CONNECTED DEVICES

Exascale

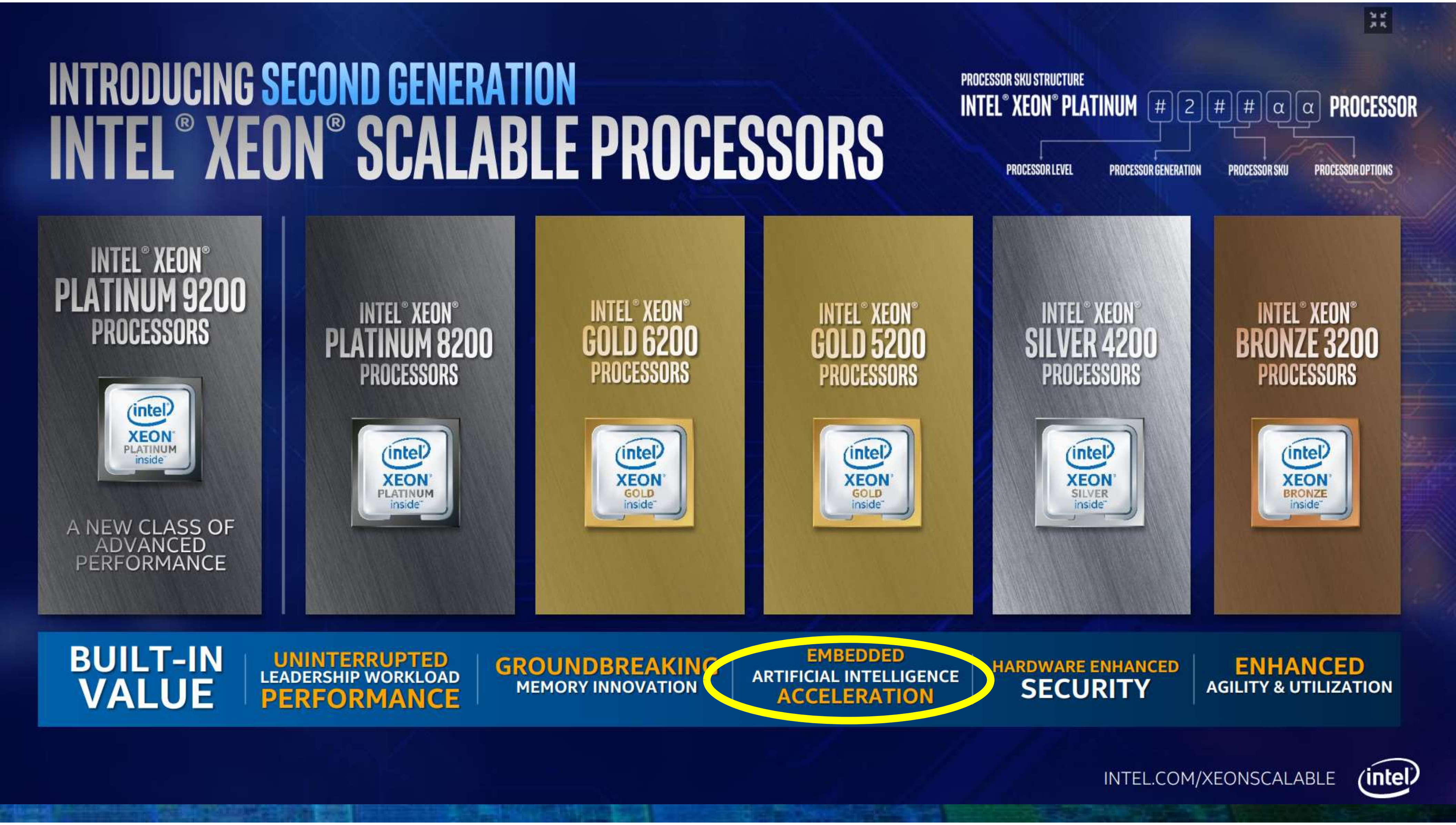
For Everyone

Compute and Memory Challenges for AI





- Compute demand growth rate: Doubling every 3-4 months
- Memory capacity growth rate: 10X per year



INTRODUCING SECOND GENERATION INTEL® XEON® SCALABLE PROCESSORS

PROCESSOR SKU STRUCTURE

INTEL® XEON® PLATINUM # 2 # # α α PROCESSOR

PROCESSOR LEVEL
PROCESSOR GENERATION
PROCESSOR SKU
PROCESSOR OPTIONS

INTEL® XEON® PLATINUM 9200 PROCESSORS

A NEW CLASS OF ADVANCED PERFORMANCE

INTEL® XEON® PLATINUM 8200 PROCESSORS

INTEL® XEON® GOLD 6200 PROCESSORS

INTEL® XEON® GOLD 5200 PROCESSORS

INTEL® XEON® SILVER 4200 PROCESSORS

INTEL® XEON® BRONZE 3200 PROCESSORS

BUILT-IN VALUE

UNINTERRUPTED LEADERSHIP WORKLOAD PERFORMANCE

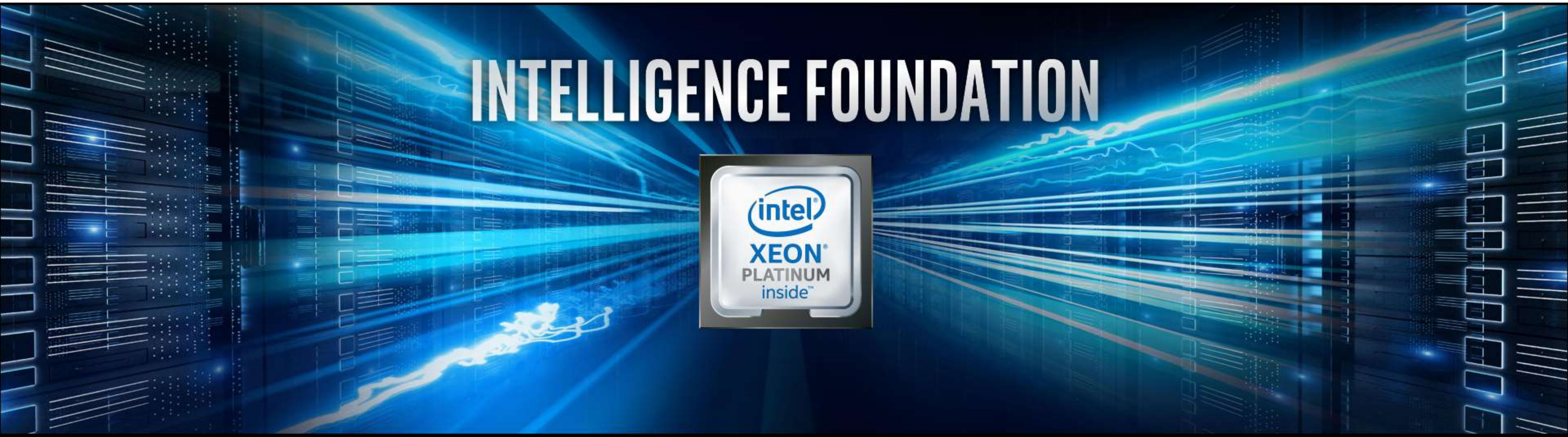
GROUNDBREAKING MEMORY INNOVATION

EMBEDDED ARTIFICIAL INTELLIGENCE ACCELERATION

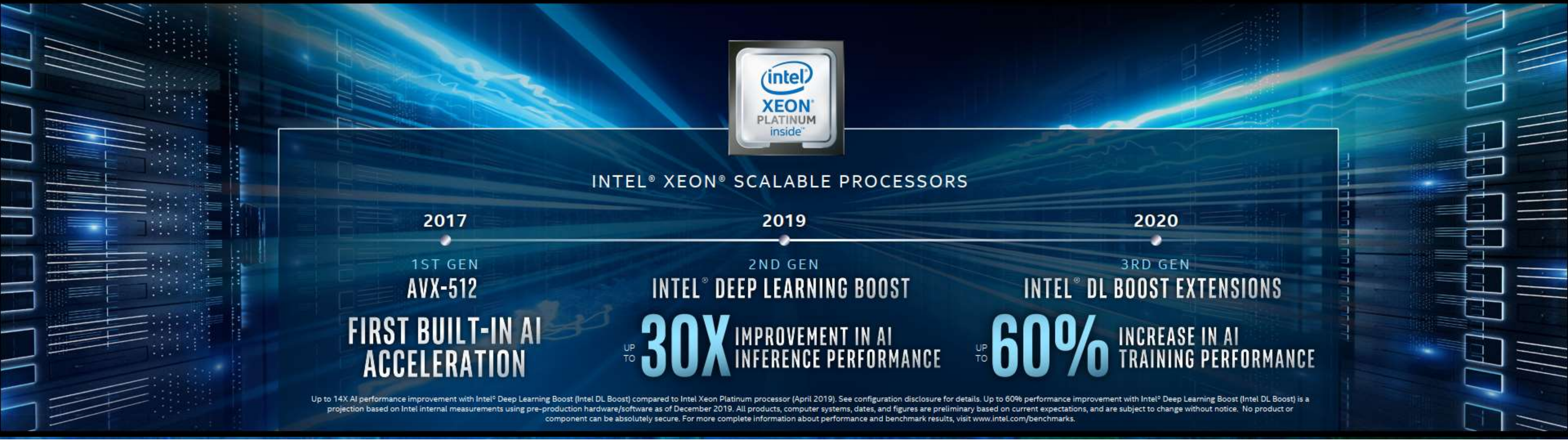
HARDWARE ENHANCED SECURITY

ENHANCED AGILITY & UTILIZATION

INTEL.COM/XEONSCALABLE



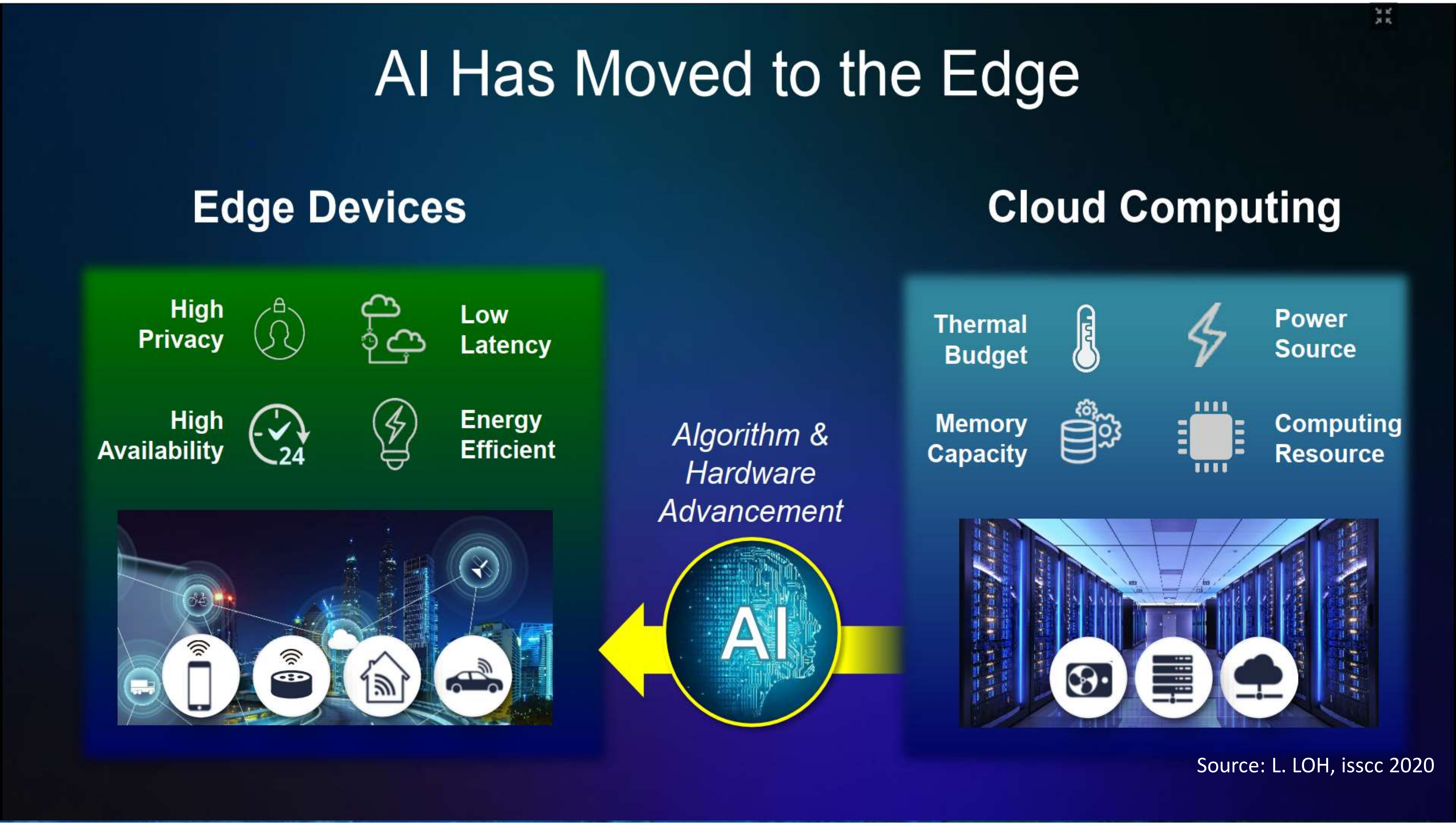
INTELLIGENCE FOUNDATION



INTEL® XEON® SCALABLE PROCESSORS

2017	2019	2020
1ST GEN	2ND GEN	3RD GEN
AVX-512	INTEL® DEEP LEARNING BOOST	INTEL® DL BOOST EXTENSIONS
FIRST BUILT-IN AI ACCELERATION	UP TO 30X IMPROVEMENT IN AI INFERENCE PERFORMANCE	UP TO 60% INCREASE IN AI TRAINING PERFORMANCE

Up to 14X AI performance improvement with Intel® Deep Learning Boost (Intel DL Boost) compared to Intel Xeon Platinum processor (April 2019). See configuration disclosure for details. Up to 60% performance improvement with Intel® Deep Learning Boost (Intel DL Boost) is a projection based on Intel internal measurements using pre-production hardware/software as of December 2019. All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. No product or component can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.



AI Has Moved to the Edge

Edge Devices

High Privacy

Low Latency

High Availability

Energy Efficient

Algorithm & Hardware Advancement

AI

Cloud Computing

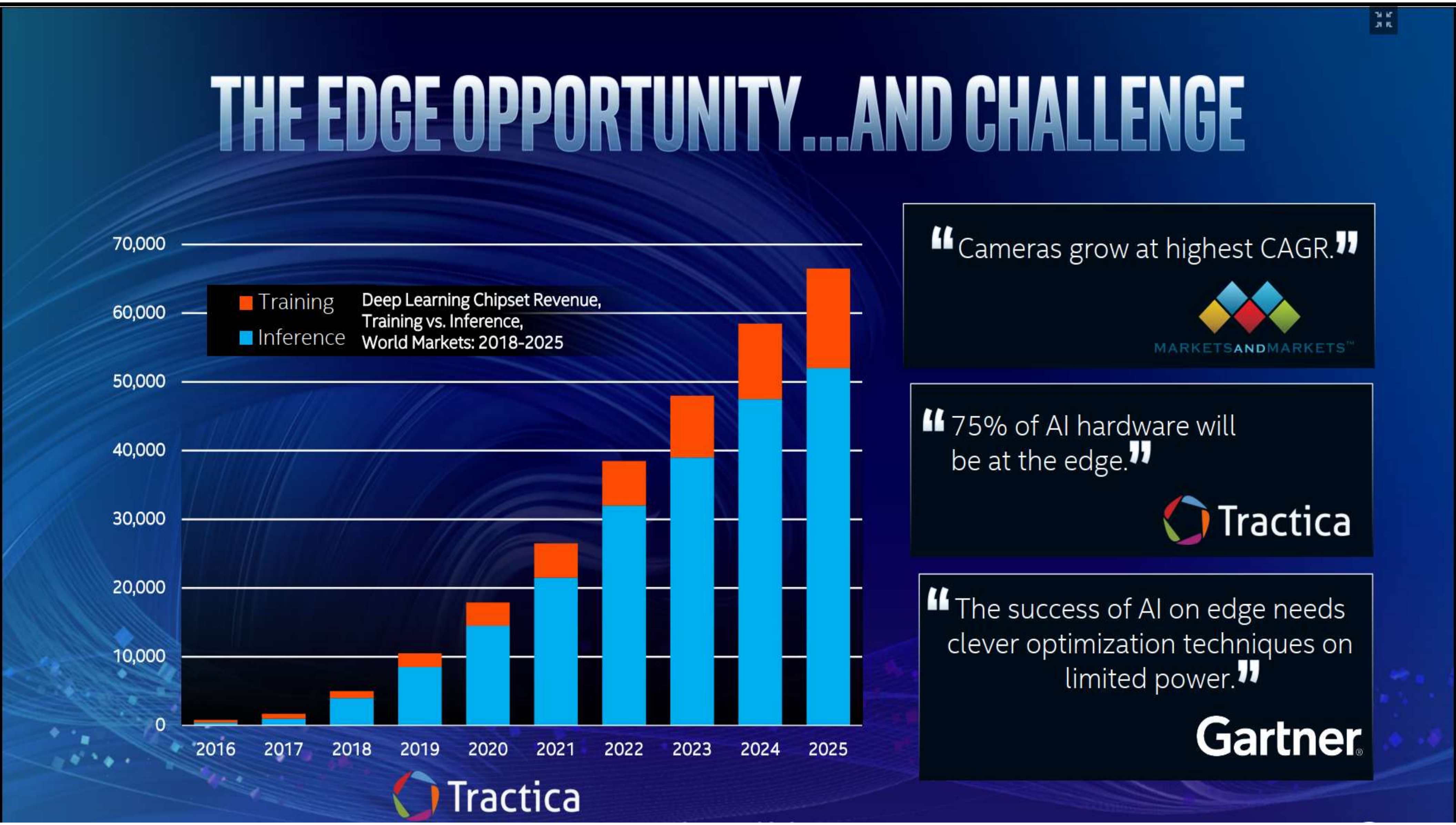
Thermal Budget

Power Source

Memory Capacity

Computing Resource

Source: L. LOH, isscc 2020



THE EDGE OPPORTUNITY...AND CHALLENGE

"Cameras grow at highest CAGR."

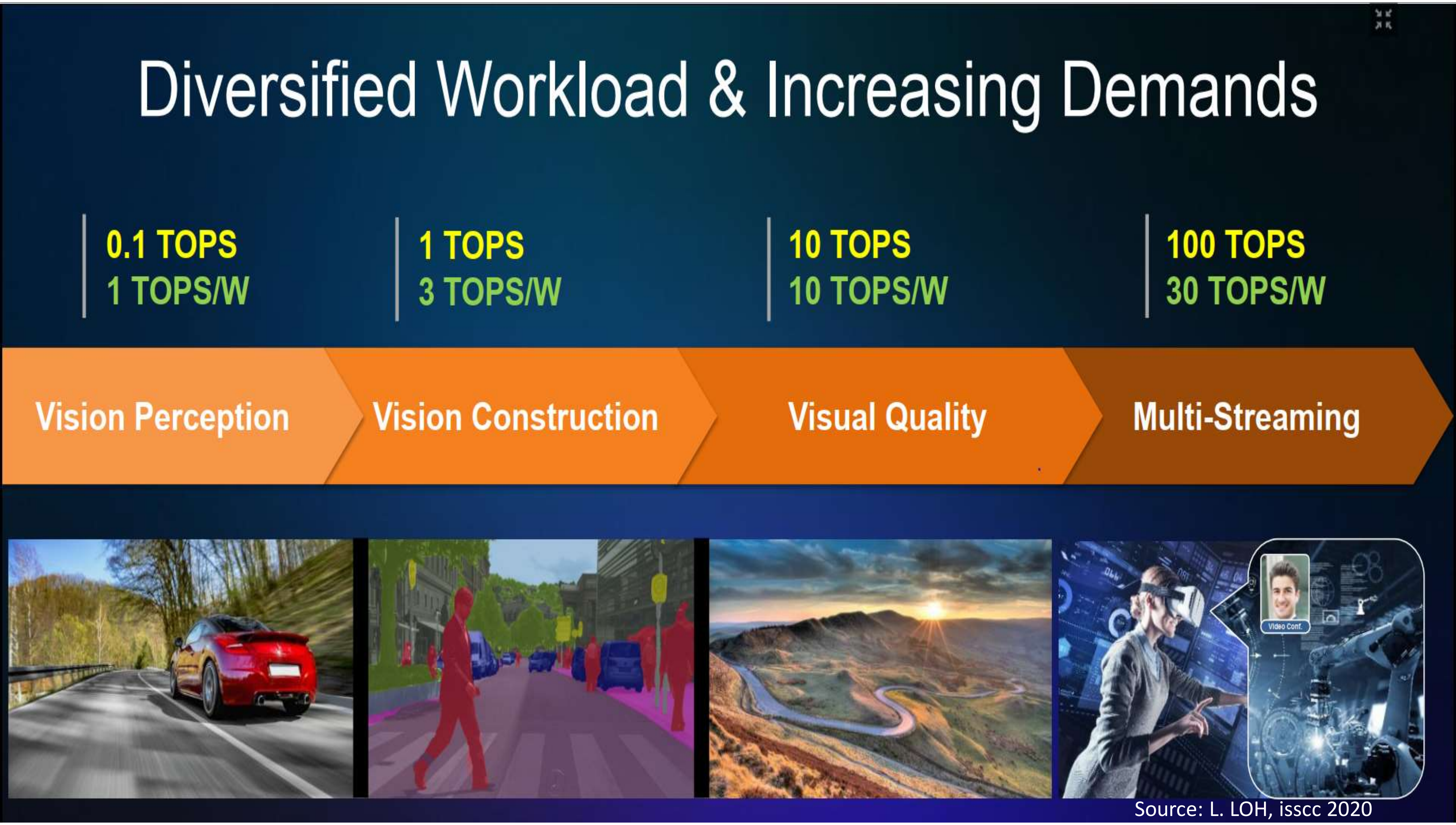
MARKETSANDMARKETS™

"75% of AI hardware will be at the edge."

Tractica

"The success of AI on edge needs clever optimization techniques on limited power."

Gartner®

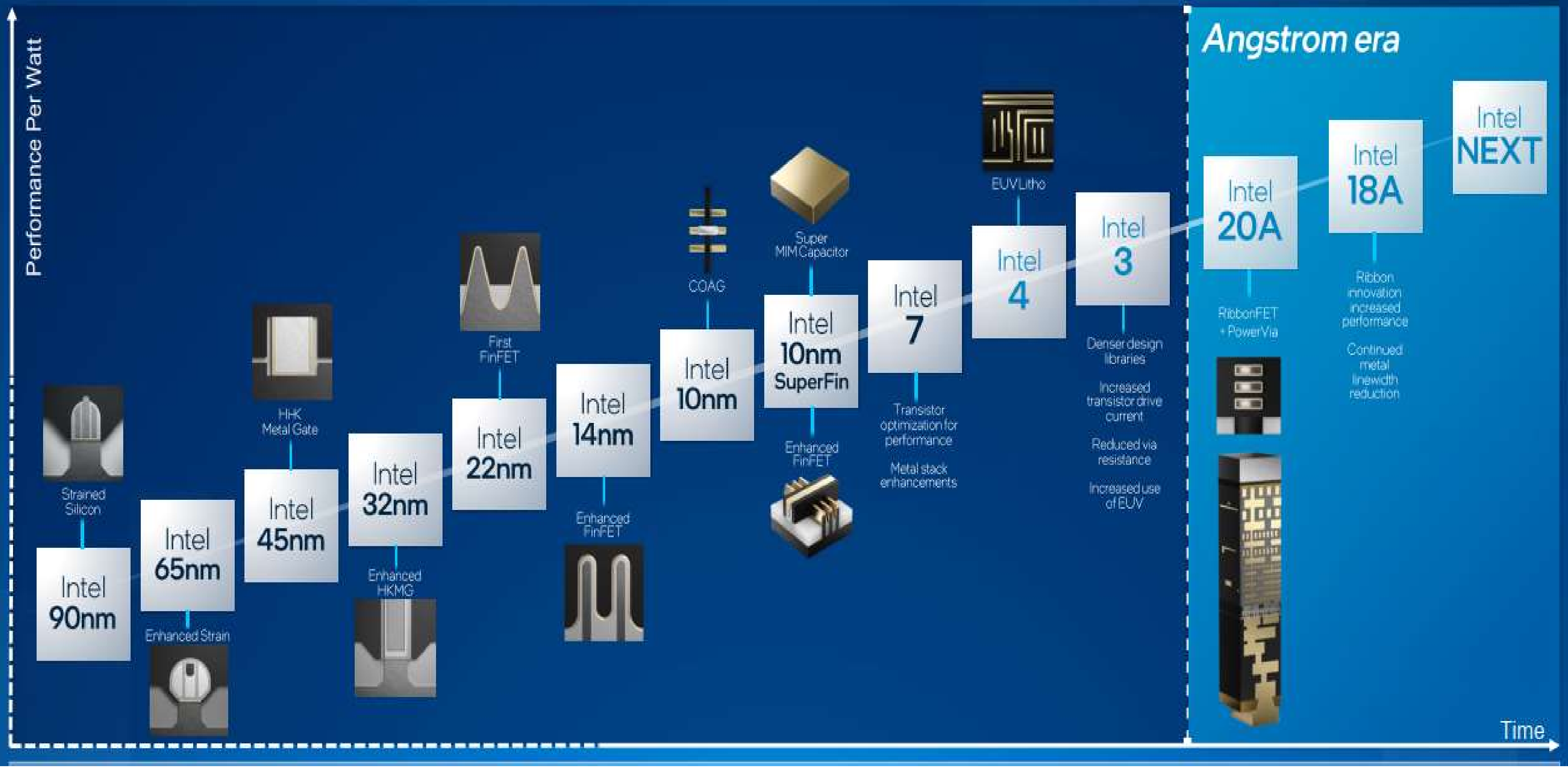


Diversified Workload & Increasing Demands

0.1 TOPS 1 TOPS/W	1 TOPS 3 TOPS/W	10 TOPS 10 TOPS/W	100 TOPS 30 TOPS/W
Vision Perception	Vision Construction	Visual Quality	Multi-Streaming

Source: L. LOH, isscc 2020

Intel Process Technology



Performance Per Watt

Time

Intel 90nm
Strained Silicon

Intel 65nm
Enhanced Strain

Intel 45nm
HI-K Metal Gate

Intel 32nm
Enhanced HKMG

Intel 22nm
First FinFET

Intel 14nm
Enhanced FinFET

Intel 10nm
COAG

Intel 10nm SuperFin
Super MIM Capacitor
Enhanced FinFET

Intel 7
Transistor optimization for performance
Metal stack enhancements

Intel 4
EUV Litho

Intel 3
Denser design libraries
Increased transistor drive current
Reduced via resistance
Increased use of EUV

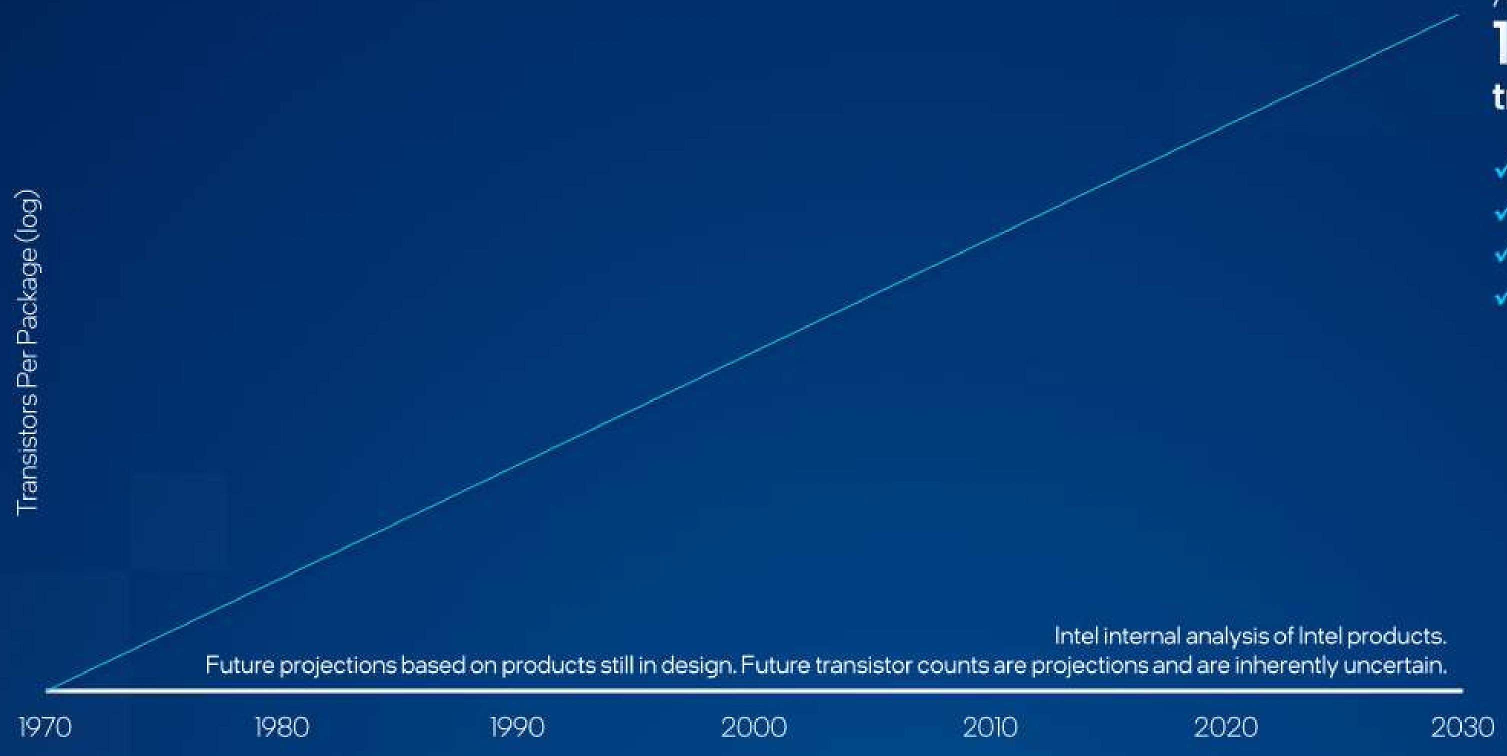
Angstrom era

Intel 20A
RibbonFET + PowerVia

Intel 18A
Ribbon innovation increased performance
Continued metal linewidth reduction

Intel NEXT

Moore’s Law



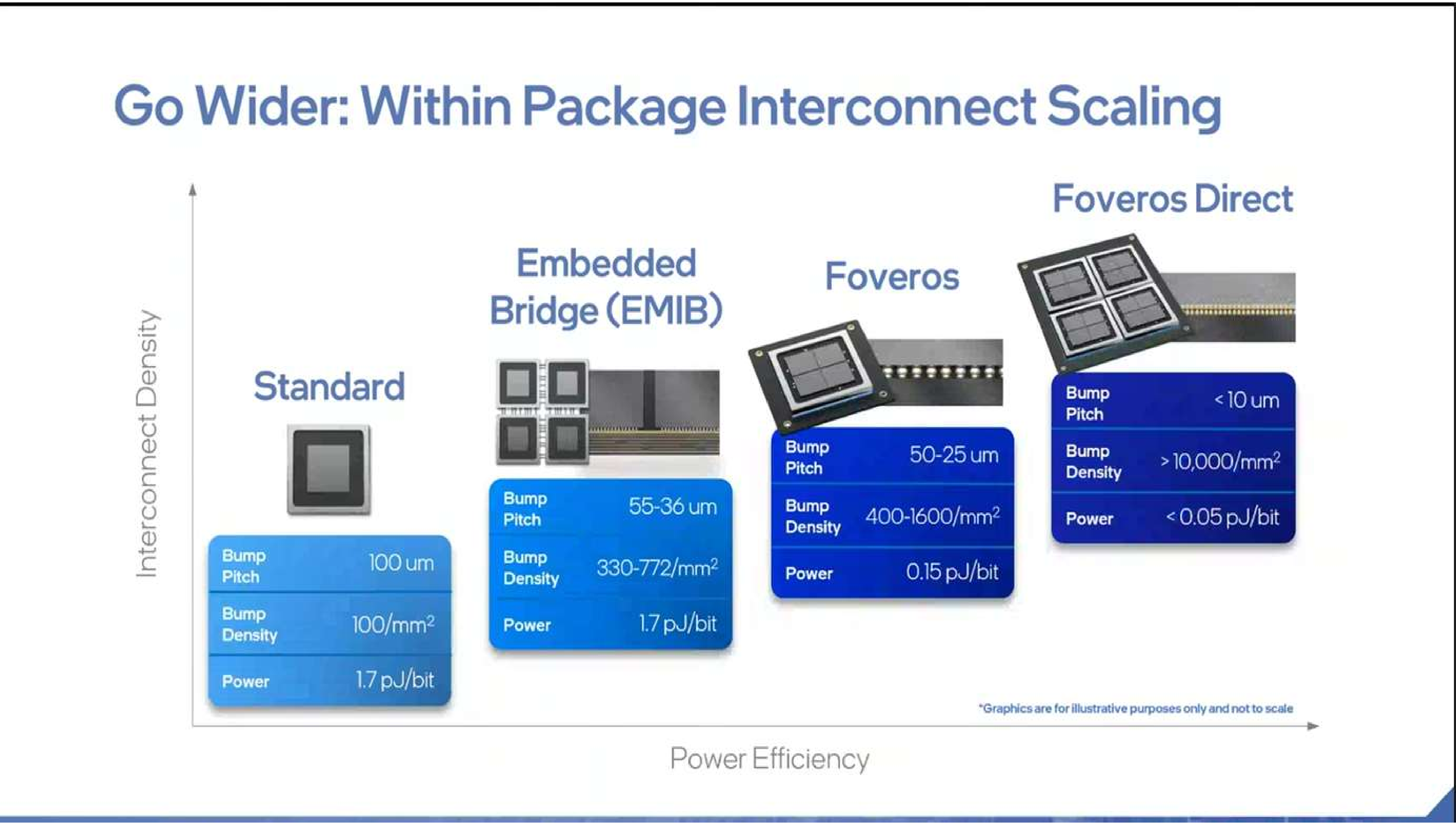
Intel internal analysis of Intel products.
Future projections based on products still in design. Future transistor counts are projections and are inherently uncertain.

Aspiring to
1 Trillion
transistors in 2030

- ✓ RibbonFET
- ✓ PowerVia
- ✓ High NA
- ✓ 2.5 / 3D packaging

Moore’s Law continues

when combining the power of processing and packaging innovation



Go Wider: Within Package Interconnect Scaling

Interconnect Density

Standard

Bump Pitch	100 um
Bump Density	100/mm^2
Power	1.7 pJ/bit

Embedded Bridge (EMIB)

Bump Pitch	55-36 um
Bump Density	330-772/mm^2
Power	1.7 pJ/bit

Foveros

Bump Pitch	50-25 um
Bump Density	400-1600/mm^2
Power	0.15 pJ/bit

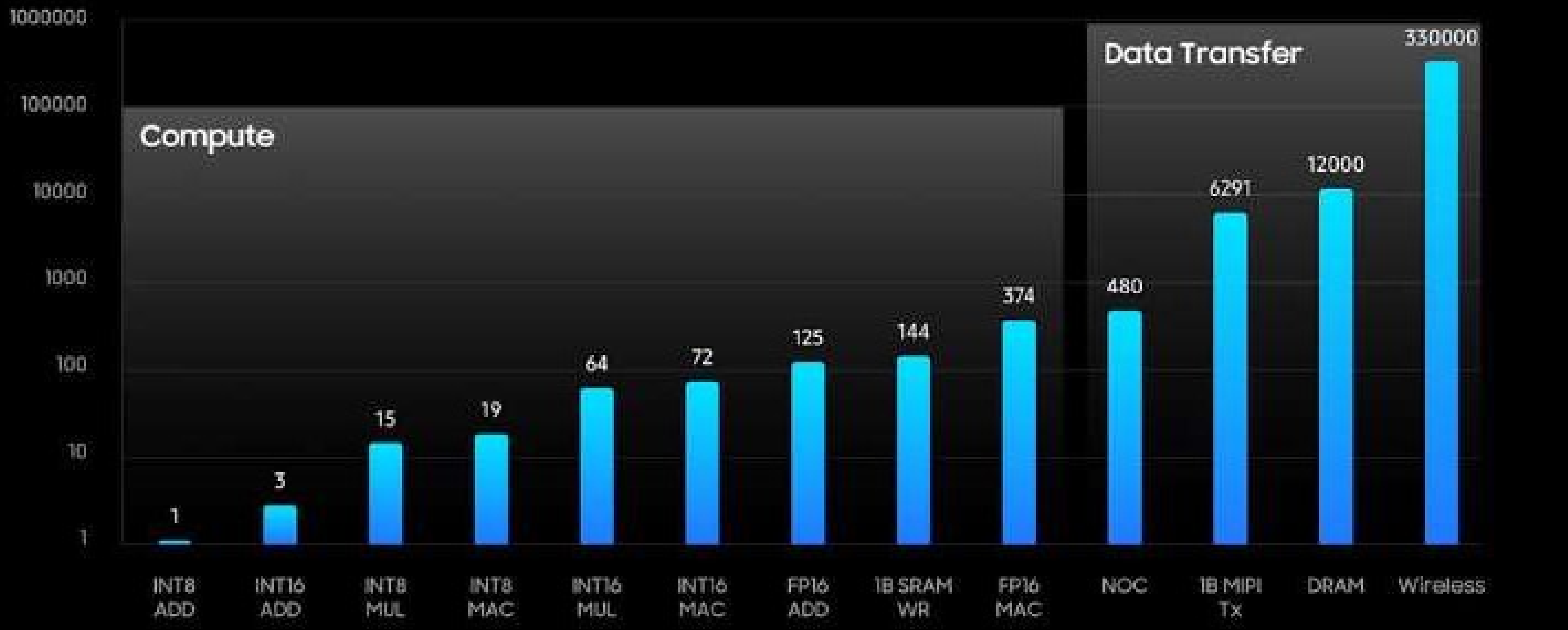
Foveros Direct

Bump Pitch	< 10 um
Bump Density	> 10,000/mm^2
Power	< 0.05 pJ/bit

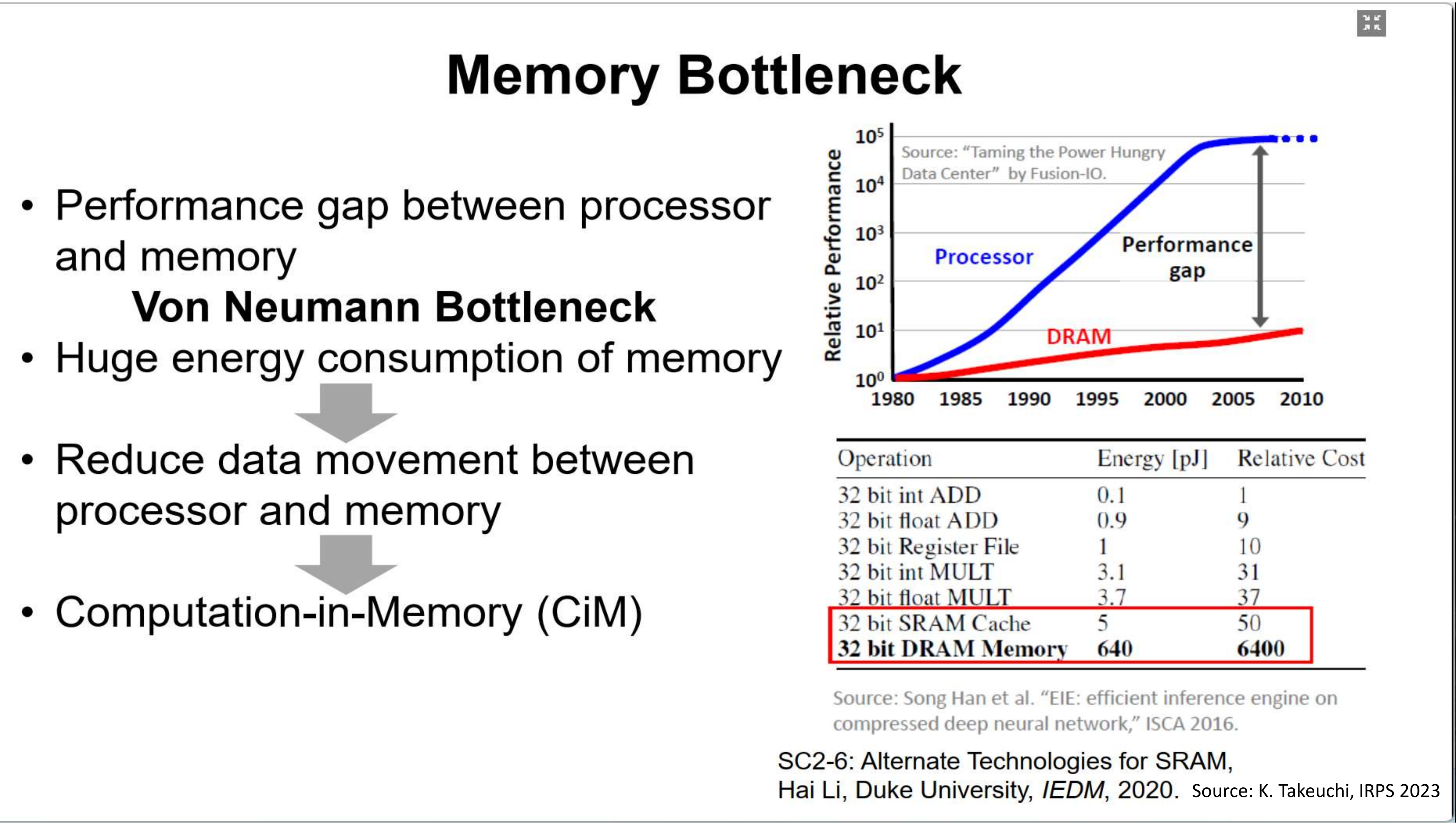
*Graphics are for illustrative purposes only and not to scale

Power Efficiency

Normalized Energy



Source: facebook



Memory Bottleneck

- Performance gap between processor and memory

  Von Neumann Bottleneck
- Huge energy consumption of memory
- Reduce data movement between processor and memory
- Computation-in-Memory (CiM)

Source: "Taming the Power Hungry Data Center" by Fusion-IO.

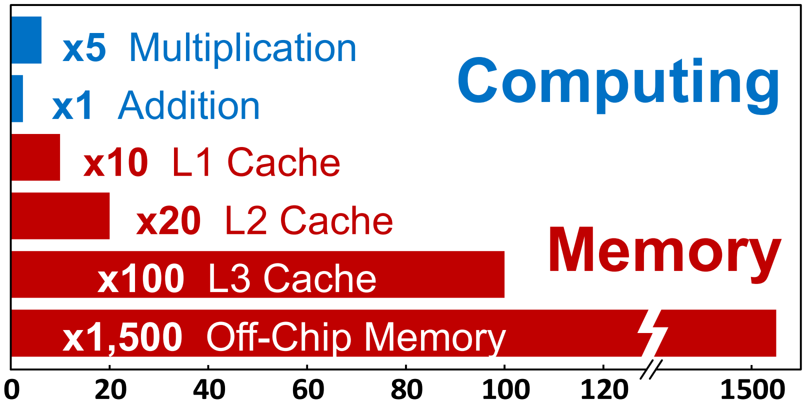
Operation	Energy [pJ]	Relative Cost
32 bit int ADD	0.1	1
32 bit float ADD	0.9	9
32 bit Register File	1	10
32 bit int MULT	3.1	31
32 bit float MULT	3.7	37
32 bit SRAM Cache	5	50
32 bit DRAM Memory	**640**	**6400**

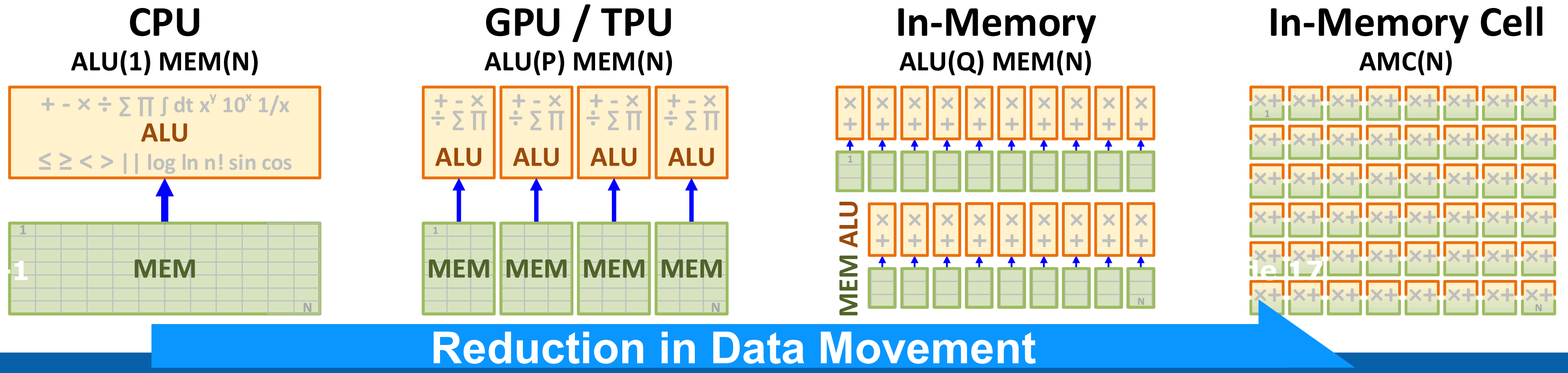
Source: Song Han et al. "EIE: efficient inference engine on compressed deep neural network," ISCA 2016.

SC2-6: Alternate Technologies for SRAM,
Hai Li, Duke University, *IEDM*, 2020. Source: K. Takeuchi, IRPS 2023

Compute-in-Memory Motivation

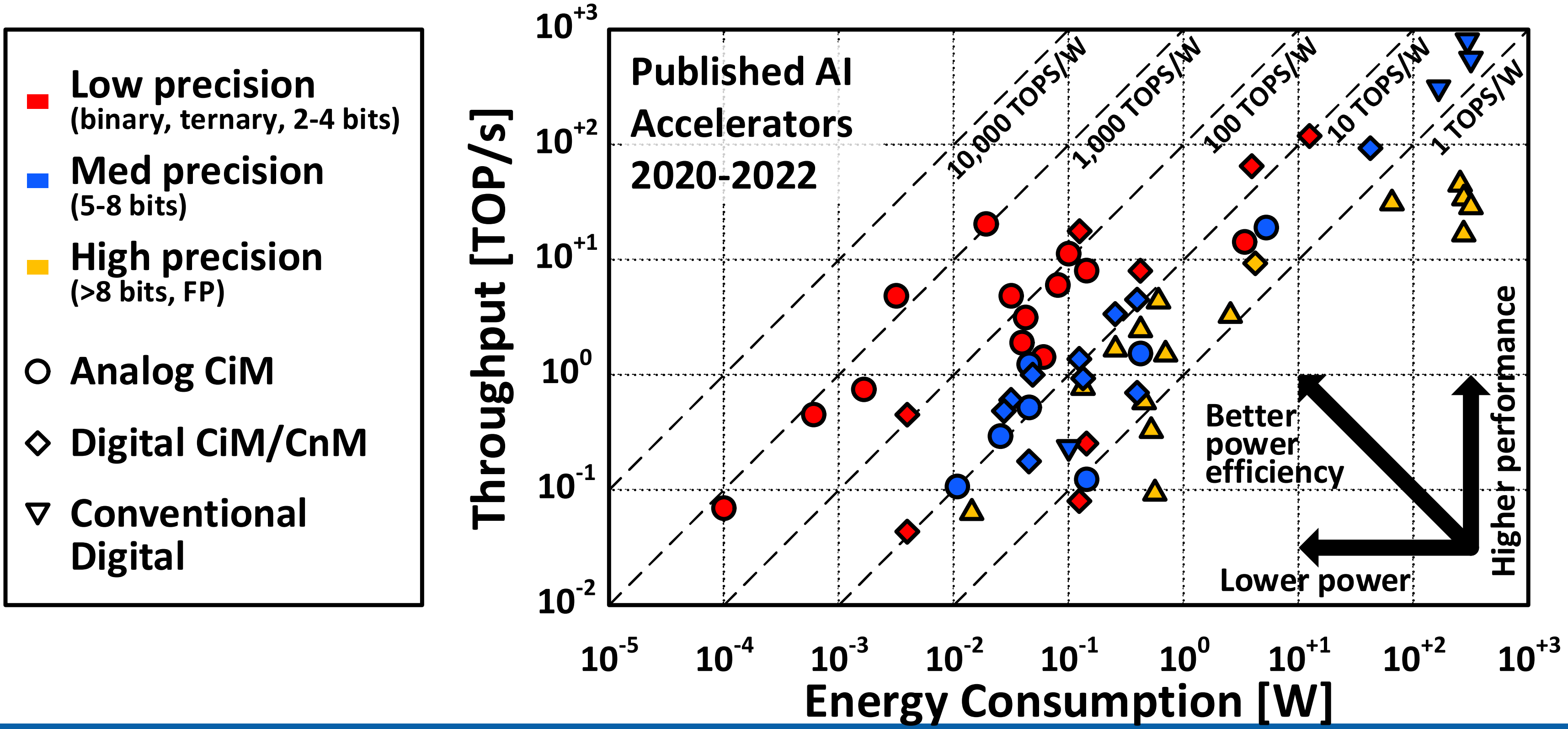
- Data movement is costly
- Multiply–accumulate (MAC) operation
- Massively parallel processing
- Beyond von Neumann architecture



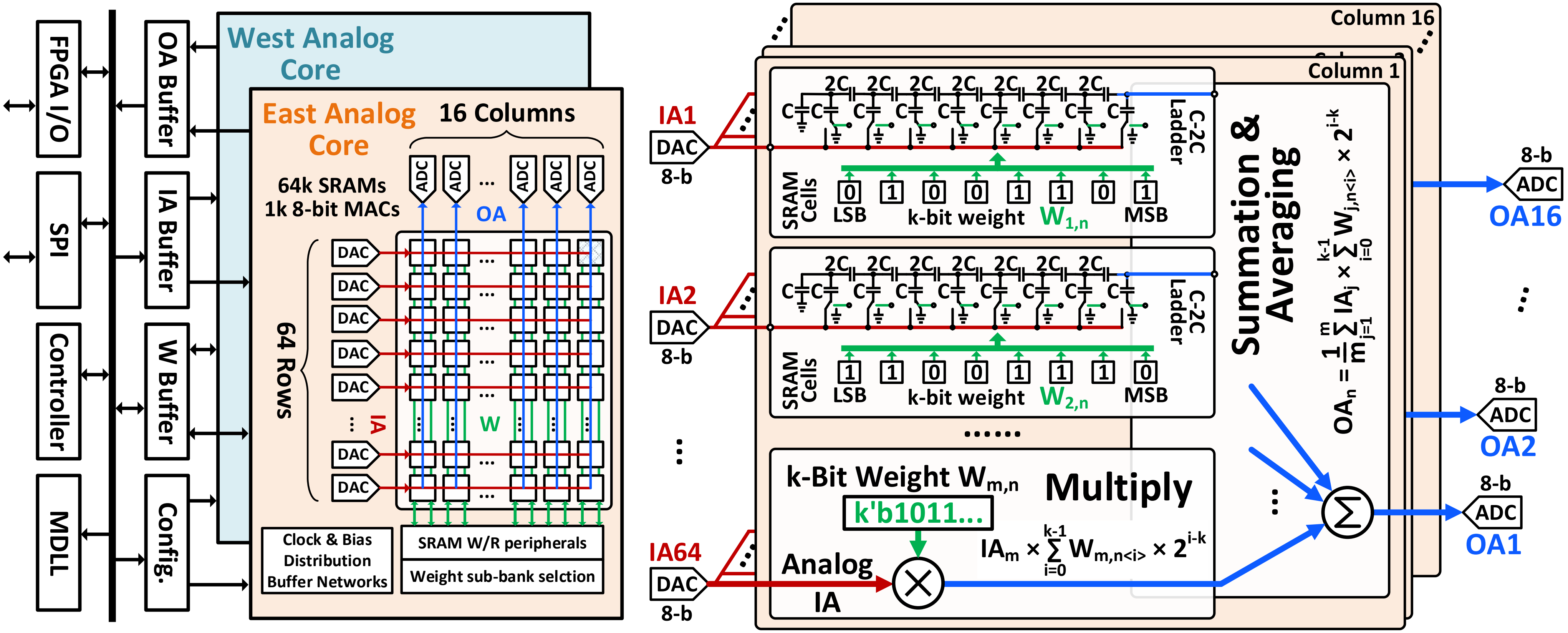


Compute-in-Memory Challenges

- TOPS/W versus Precision



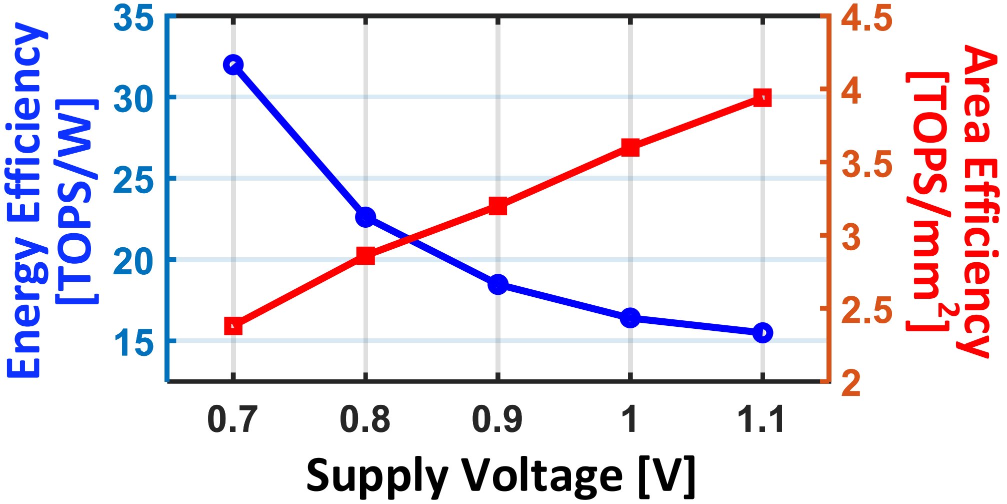
Intel Labs Analog CIM Architecture Overview

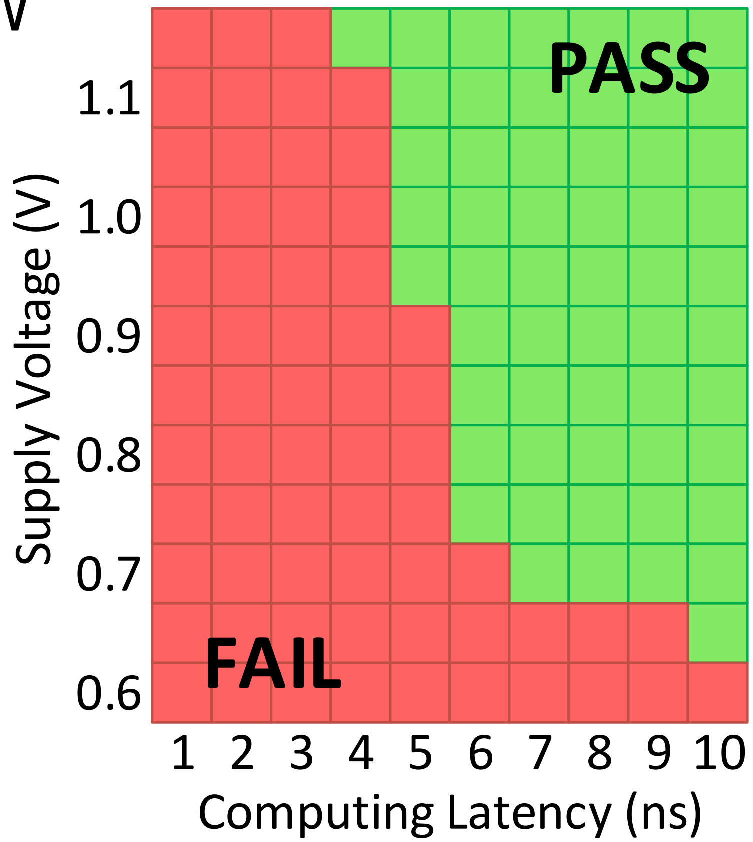


H. Wang et al, IEEE VLSI Circuits Symposium 2022

Intel Labs Analog CIM Measured Performance

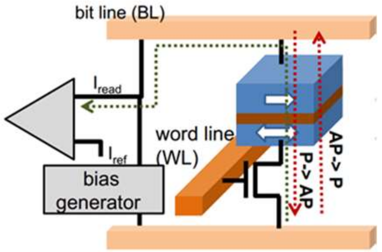
- Energy efficiency: 15.5-32.2 TOPS/W
- Area efficiency: 2.4-4.0 TOPS/mm^2
- Clock frequency: 145-240 MHz
- Supply Voltage: 0.7-1.1 V





Opportunities for in-memory/near-memory Process and Circuit Innovation
(Both Digital and Analog/Mixed-Signal)

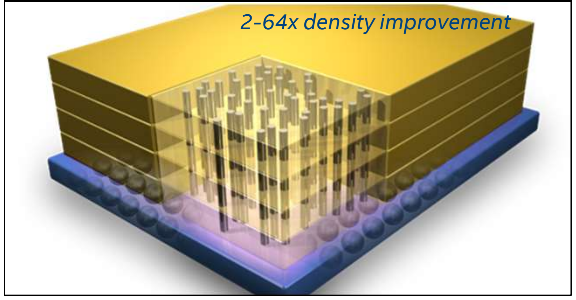
STTRAM STATE



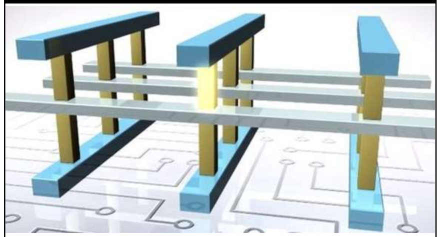
2x density improvement iso-energy

3D

2-64x density improvement

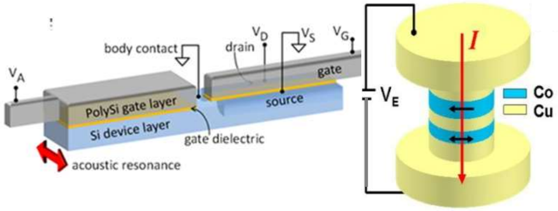


XPOINT MEMORY



10x density improvement

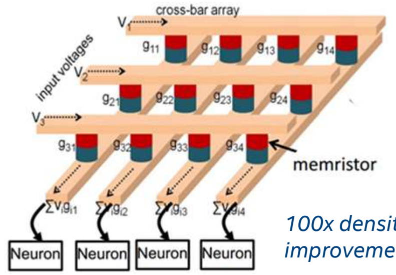
SINGLE-DEVICE OSCILLATORS



100-1000x energy reduction

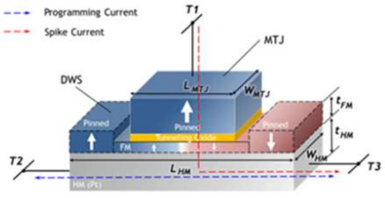
[Nikonov, Young]

RRAM/PCM ANALOG SYNAPSES



100x density improvement

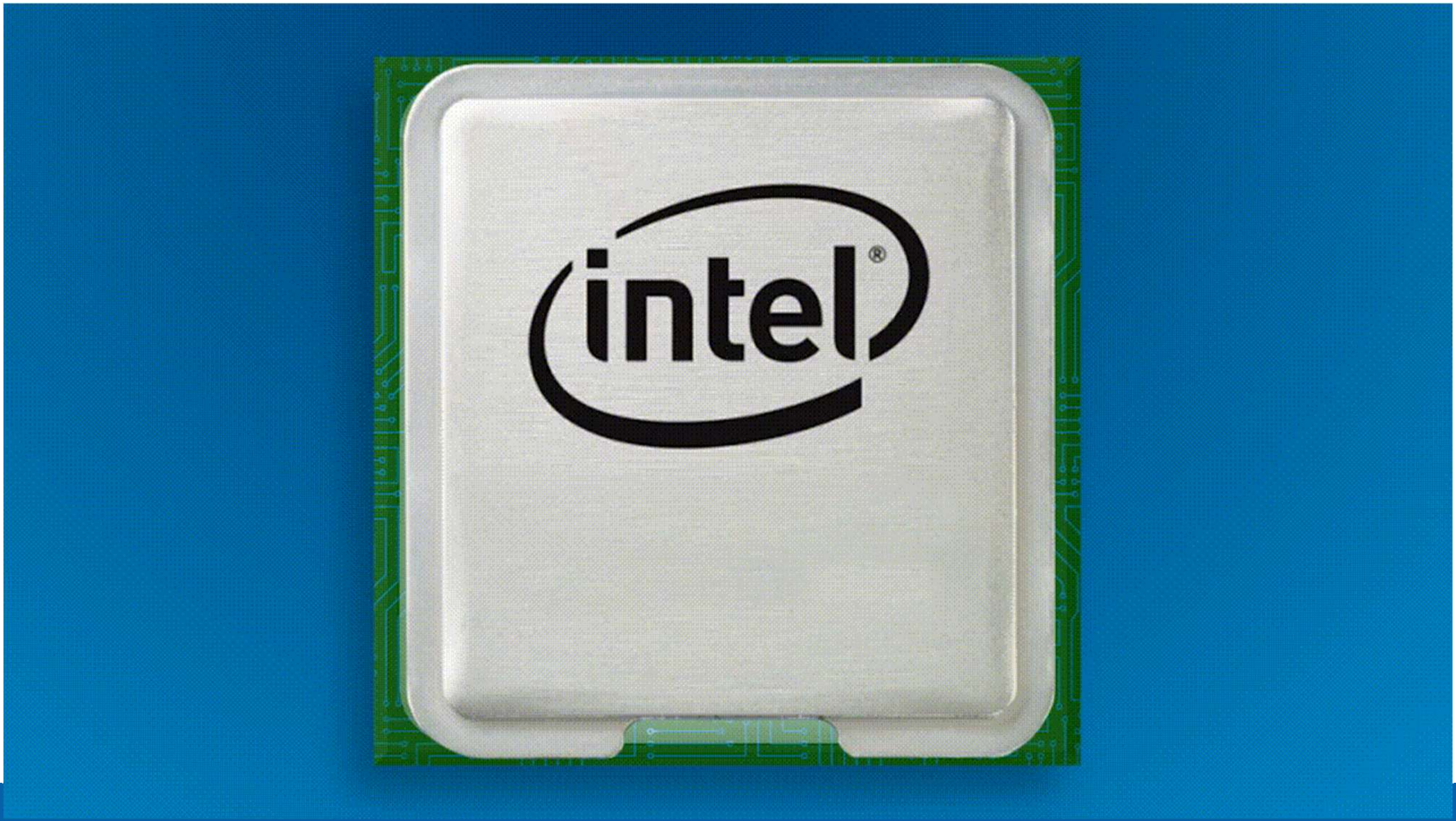
SPIN ORBIT TORQUE SYNAPSES



20x reduction in synaptic energy

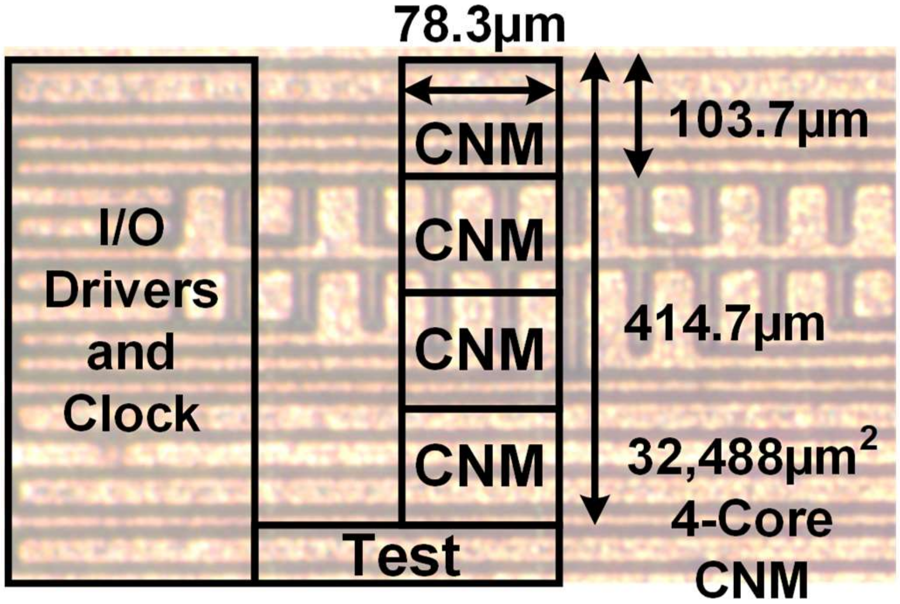
[Roy]

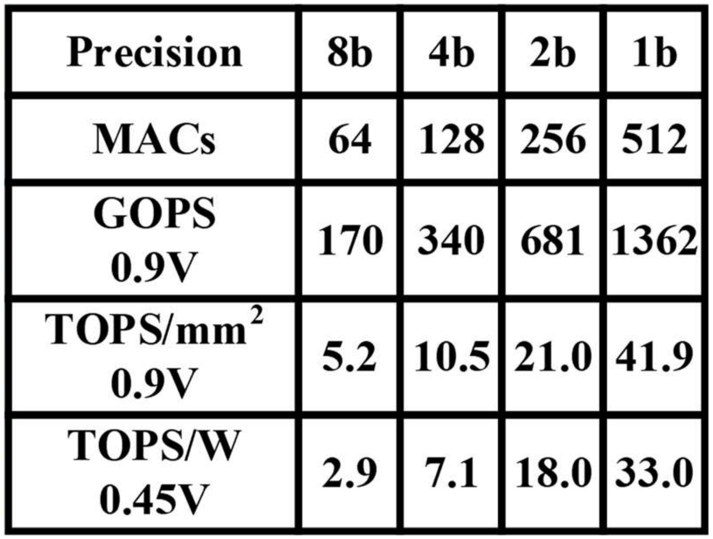
COMPUTE NEAR MEMORY CHALLENGES AND OPPORTUNITIES



E. Sumbul, R. Krishnamurthy et al, IEEE ESSCIRC 2021

10nm Near Memory Computing AI Inference Accelerator

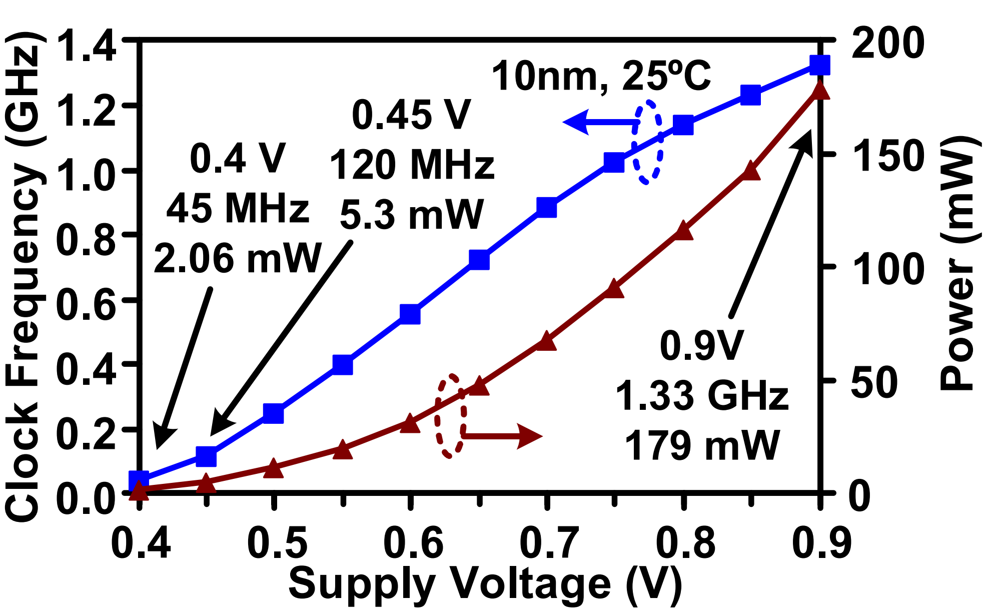


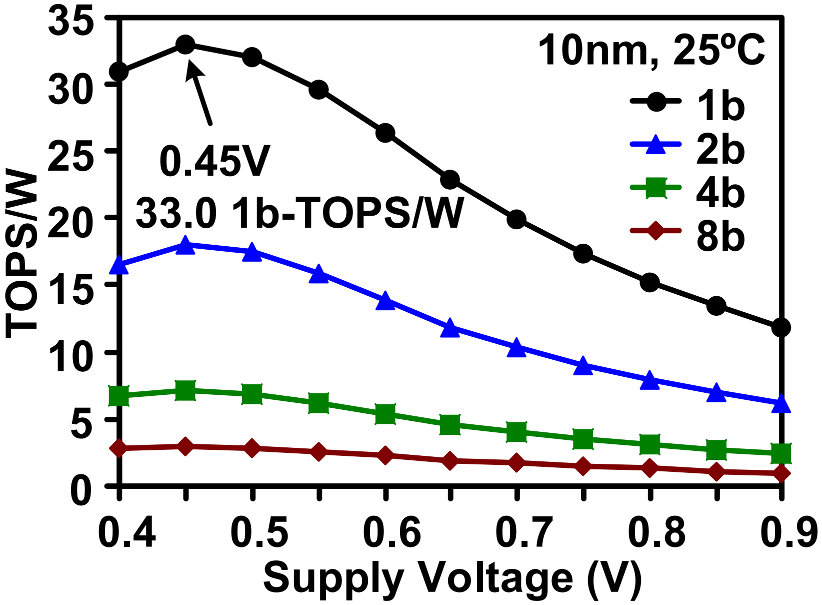


Precision	8b	4b	2b	1b
MACs	64	128	256	512
GOPS 0.9V	170	340	681	1362
TOPS/mm^2 0.9V	5.2	10.5	21.0	41.9
TOPS/W 0.45V	2.9	7.1	18.0	33.0

- 4 CNM cores with 8KB of weight memory and 64 8b multipliers
- Supports memory-intensive batch-1, large-batch, and in-place convolution

10nm Near Memory Computing Measurement Results

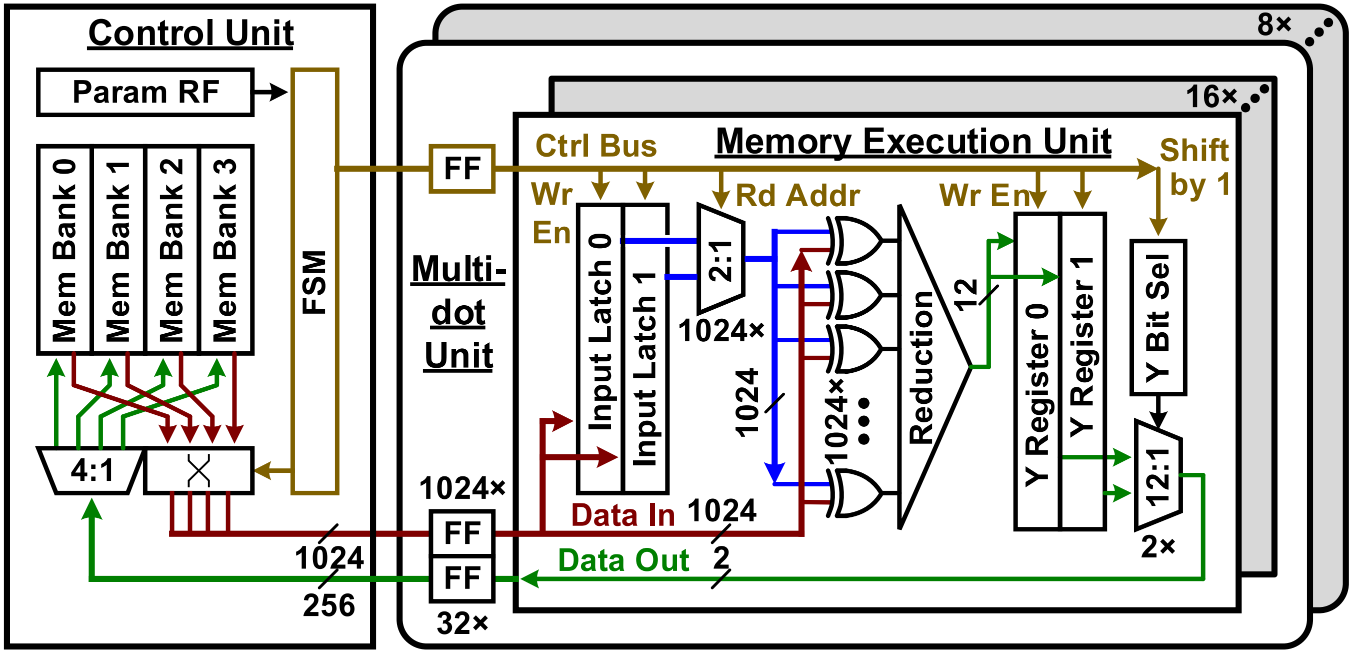




- Peak throughput 170 8b TOPS @ 0.9V that scaled up with number of CNM cores
- NTV operation down to 450mV decreases energy by 3.1x to 2.9 8b TOPS/W
- Variable precision improves energy efficiency by 11.4x to 33.0 1b TOPS/W

G. Chen, R. Krishnamurthy et al, IEEE European Solid-State Circuits Conference 2021

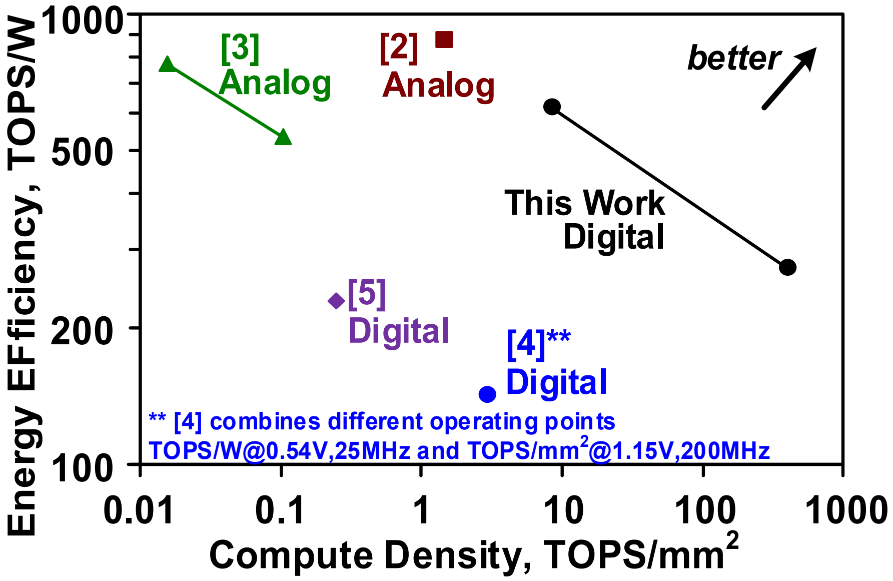
10nm Binary Neural Network Inference Accelerator



- Array of 128 Memory Execution Units (MEU) combine latch base memory and inner product compute in fine grain manner to minimize interconnect energy
- Central controller manages data flow from four 256b memory banks to MEUs
- 2 latch words per MEU enables data reuse reducing input bandwidth by 2x

Comparison to Previously Published BNNs

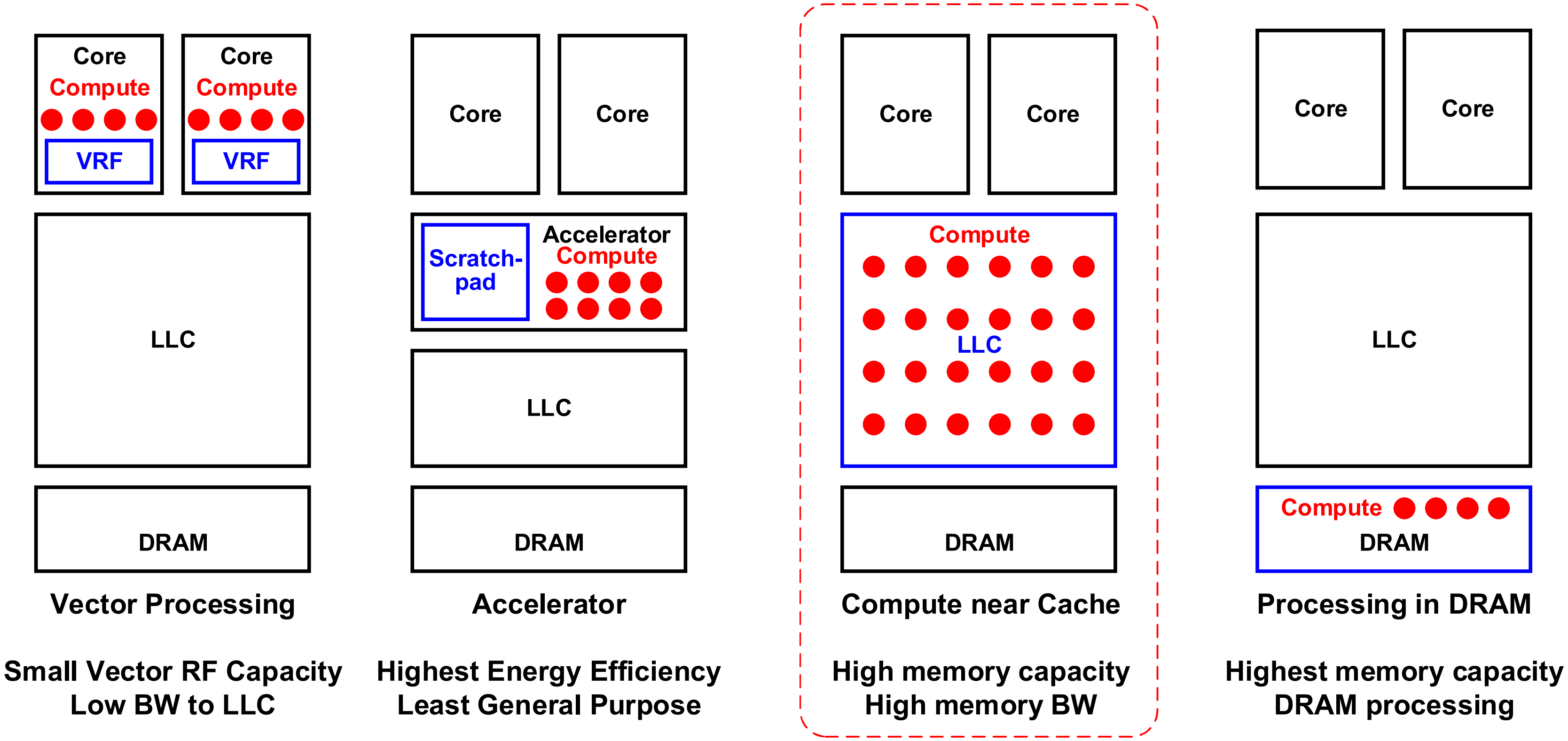
	[2]	[3]		[4]		[5]	This Work	
Digital or Analog	Analog	Analog		Digital		Digital	Digital	
Technology (nm)	65	28		65		28	10	
Area (mm^2)	12.6	4.6		4.8		1.4	0.39	
Num MACs	-	65,536		-		65,536	131,072	
Mem Capacity (KB)	295	328		104		328	161	
KB/mm^2	23	71		22		234	414	
Voltage (V)	0.68, 0.94, 1.2	0.6, 0.8, 0.6, 0.53	0.6, 0.8, 0.8, 0.8	0.54	1.15	-	0.37	0.75
Frequency (MHz)	100	1.5	10	25	200	6.0	13	622
Power (mW)	22	0.094	0.899	-	-	1.5	5.6	607
TOPS	19	0.072	0.478	-	14.9	0.35	3.4	163
TOPS/W	866	772	532	140	-	230	617	269
TOPS/mm^2	1.5	0.02	0.10	-	3.1	0.25	8.8	418



** [4] combines different operating points TOPS/W@0.54V,25MHz and TOPS/mm^2@1.15V,200MHz

P. Knag, R. Krishnamurthy et al, IEEE Journal of Solid-State Circuits Invited Paper, April 2021

Compute Near Last Level Cache (CNC)



Core Compute VRF | Core Compute VRF

LLC

DRAM

Vector Processing

Small Vector RF Capacity
Low BW to LLC

Core | Core

Scratch-pad | Accelerator Compute

LLC

DRAM

Accelerator

Highest Energy Efficiency
Least General Purpose

Core | Core

Compute LLC

DRAM

Compute near Cache

High memory capacity
High memory BW

Core | Core

LLC

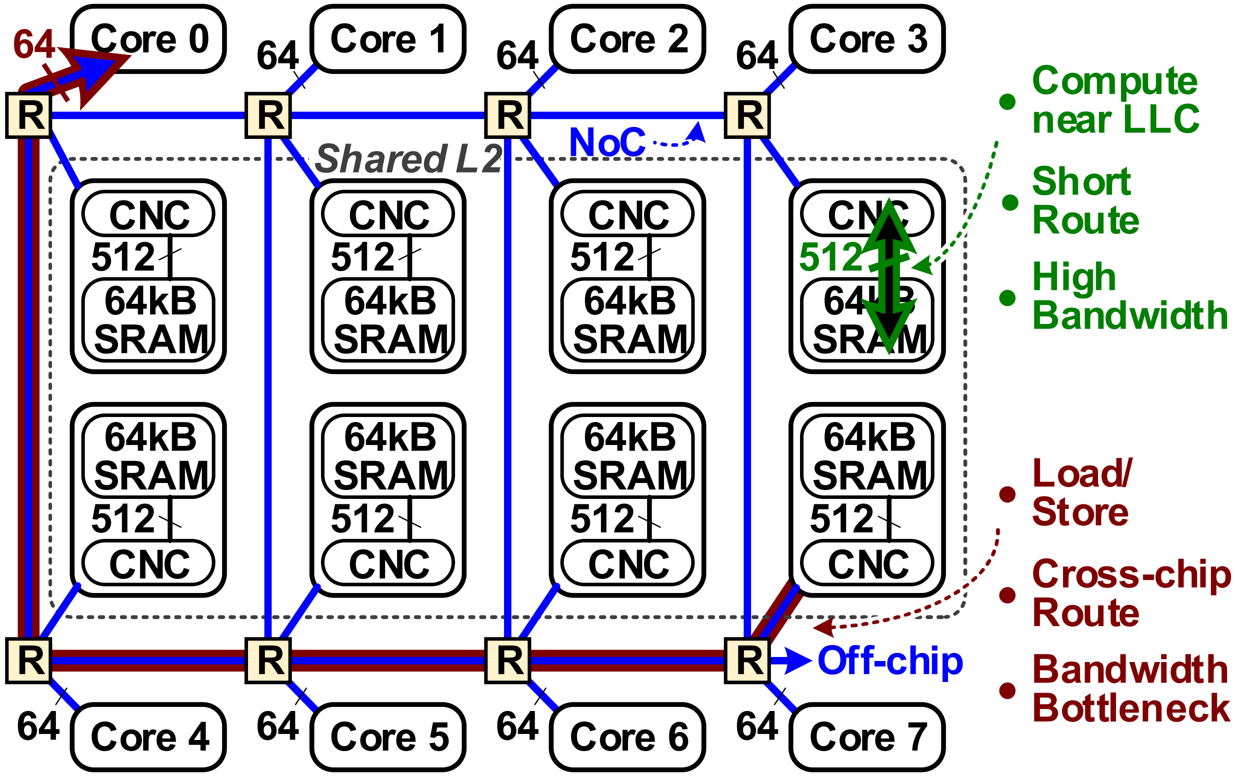
Compute DRAM

Processing in DRAM

Highest memory capacity
DRAM processing

- CNC enables fine grain mixing of near-memory vector and GP scalar computation
- High BW access to highest capacity on-chip memory instead of RF/scratchpad

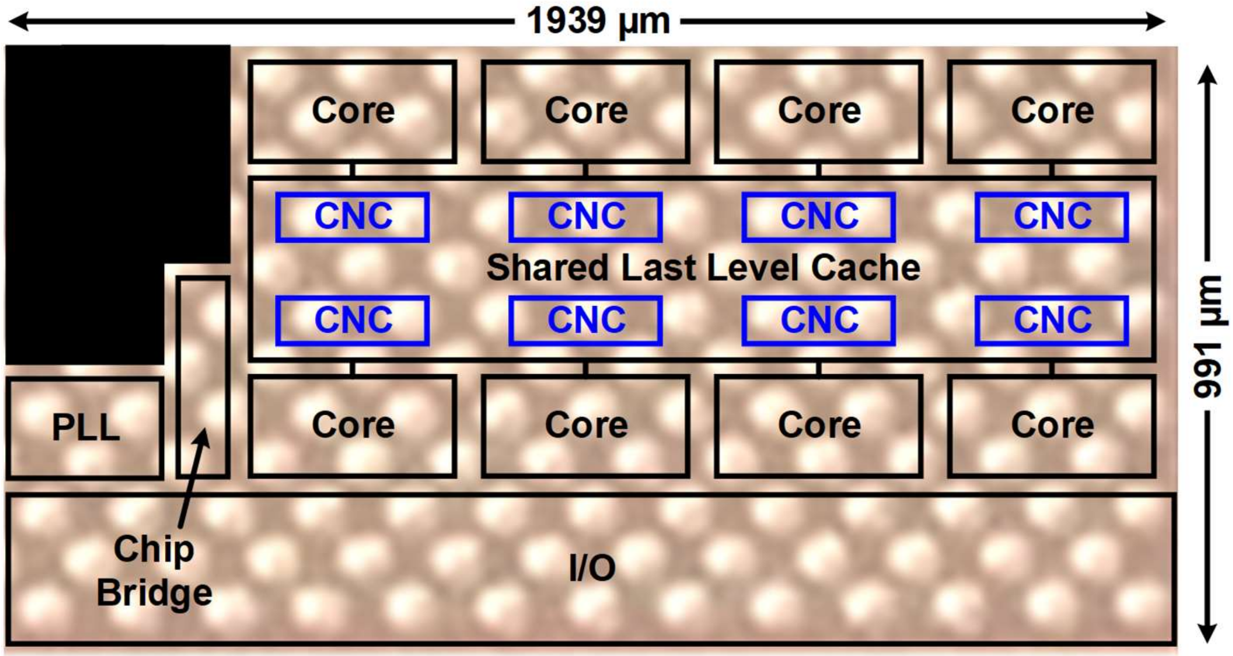
Compute Near Last Level Cache of RISC-V Multiprocessor

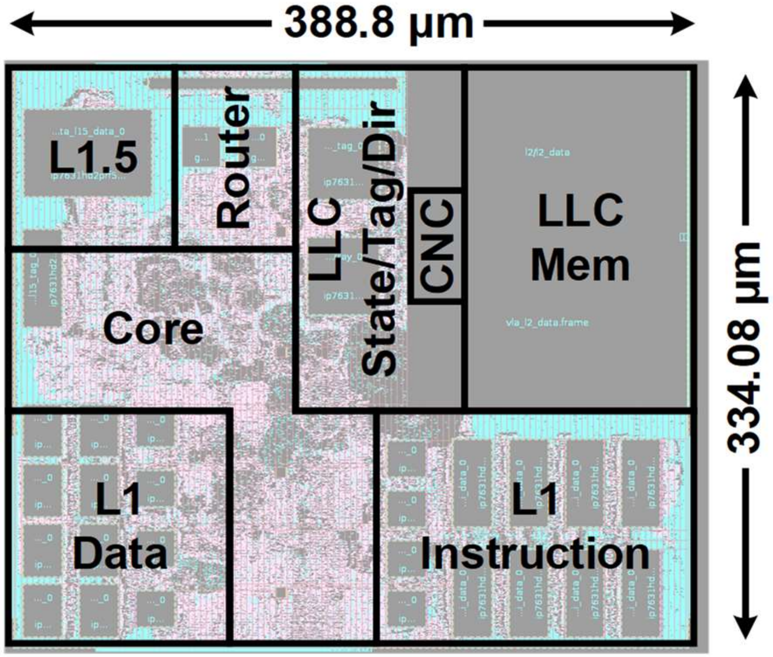


- 8-core RV64GC processor with 128 INT8 MACs near 512kB shared, distributed LLC
- CNC ISA extension with support for virtual addressing and cache coherence

G. Chen, R. Krishnamurthy et al, IEEE VLSI Circuits Symposium 2022 & JSSC Journal Invited Paper April 2023

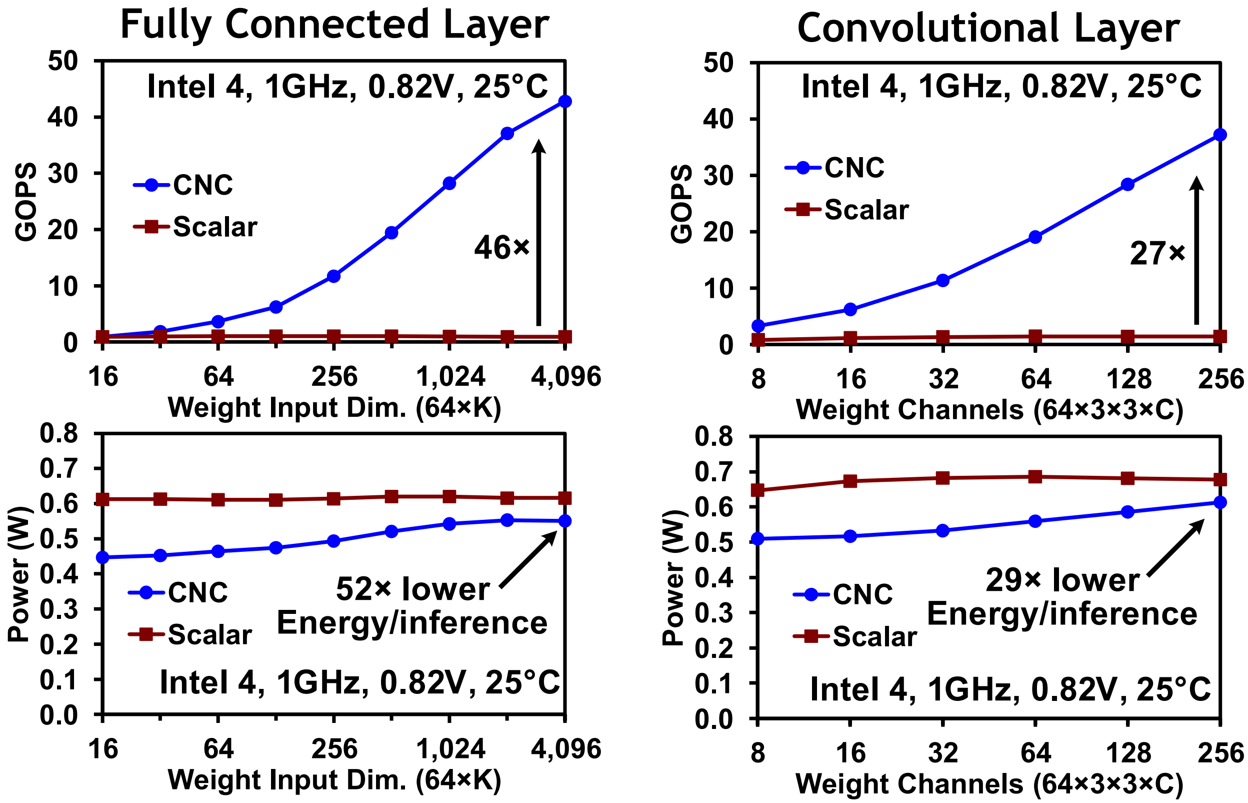
Intel 4 Silicon Implementation of 8-Core RISC-V





- 1.15GHz Intel 4 test-chip runs programs in C++ with inline CNC and boots Linux
- CNC circuits add 1.4% area overhead over baseline core + LLC design
- Flip-chip packaged with PLL and 32b IO to FPGA chipset

8-Core RISC-V DNN Layer Performance in Intel 4



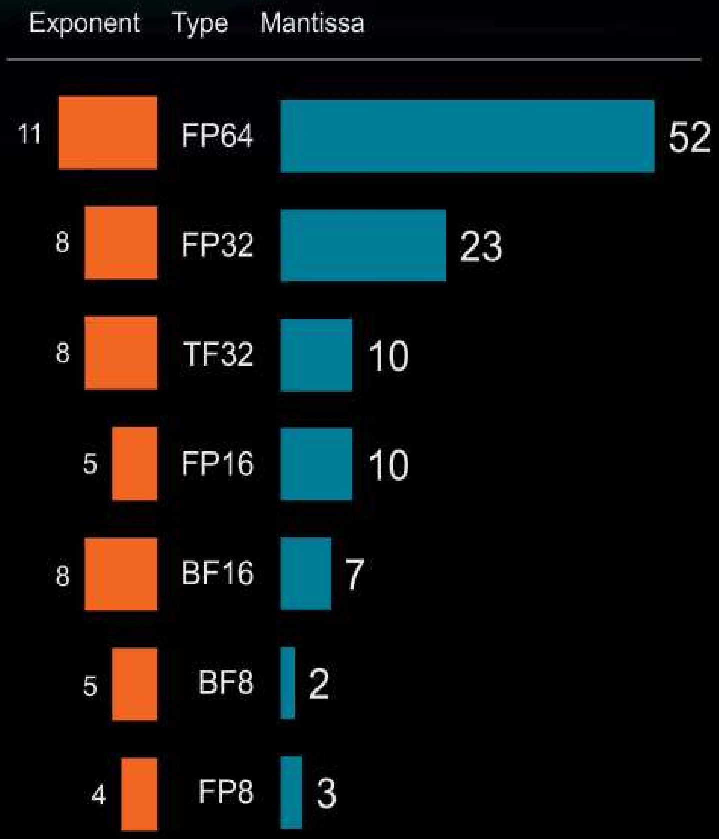
Measured for dense operation (no sparsity)

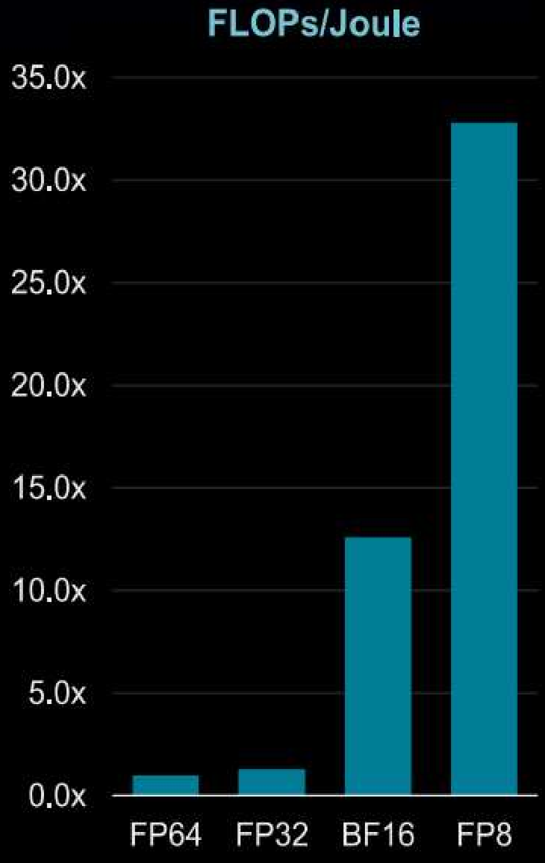
- Fully Connected Layers: up to 46× higher performance and 52× lower energy
- Convolutional Layers: up to 27× higher performance and 29× lower energy

G. Chen, R. Krishnamurthy et al, IEEE VLSI Circuits Symposium 2022 & JSSC Journal Invited Paper April 2023

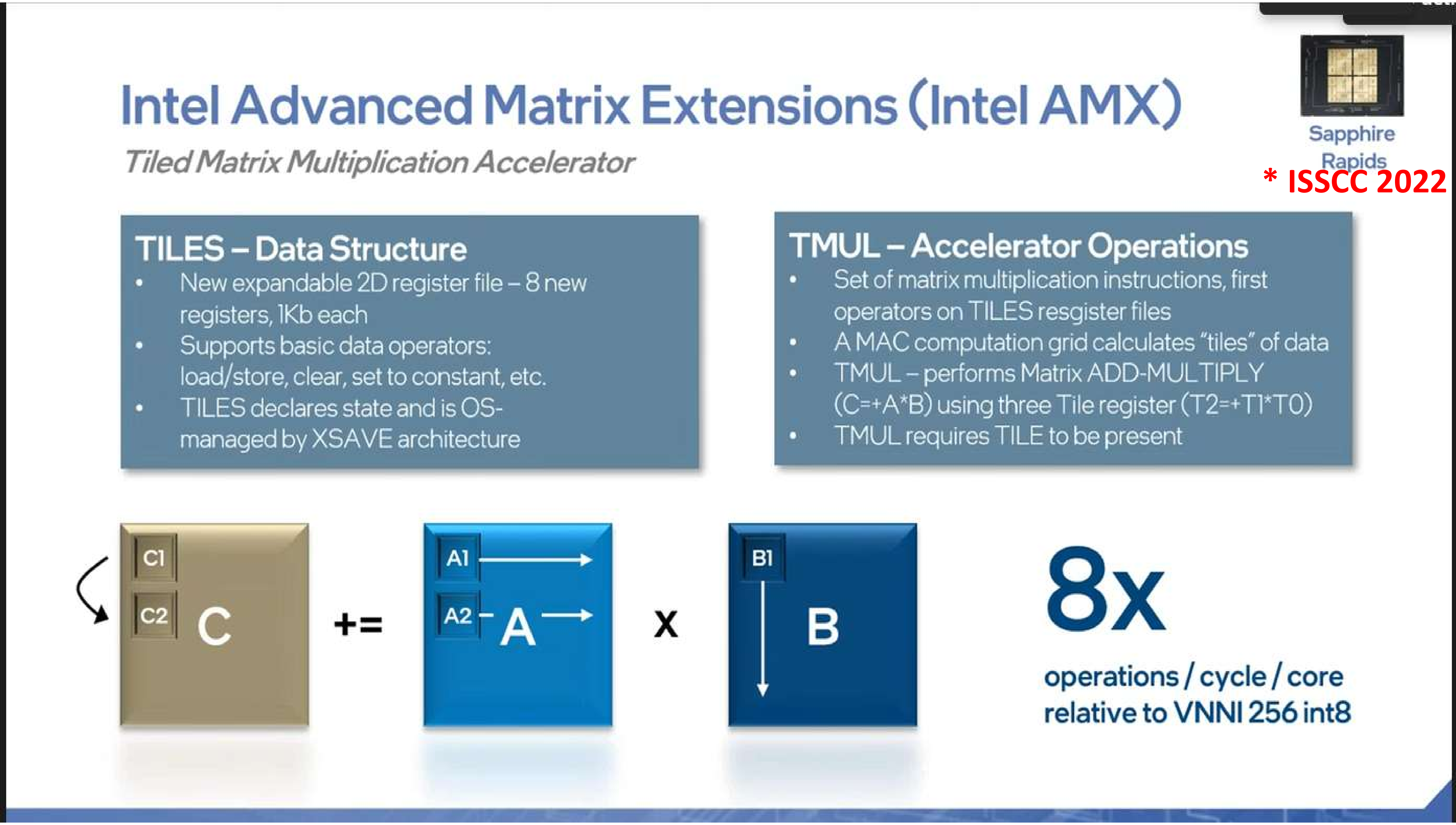
Domain-Specific Computation Enables Workload Optimization which Drives Performance and Efficiency

- Tailor architecture by application
- Adapt algorithms to use lower precision math formats for significant improvements in energy efficiency





Source: L. Su, ISSCC 2023



Intel Advanced Matrix Extensions (Intel AMX)

Tiled Matrix Multiplication Accelerator

Sapphire Rapids

* ISSCC 2022

TILES – Data Structure

- New expandable 2D register file – 8 new registers, 1Kb each
- Supports basic data operators: load/store, clear, set to constant, etc.
- TILES declares state and is OS-managed by XSAVE architecture

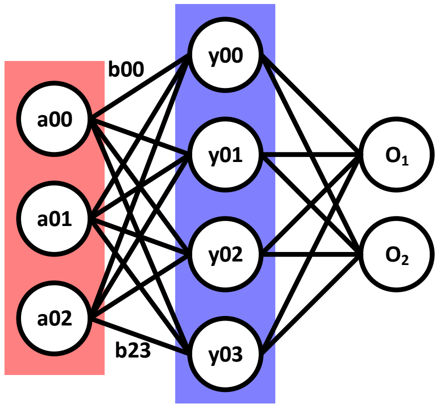
TMUL – Accelerator Operations

- Set of matrix multiplication instructions, first operators on TILES resgister files
- A MAC computation grid calculates "tiles" of data
- TMUL – performs Matrix ADD-MULTIPLY (C=+A*B) using three Tile register (T2=+T1*T0)
- TMUL requires TILE to be present

8x

operations / cycle / core relative to VNNI 256 int8

Multi-precision Neural Networks Matrix Multipliers

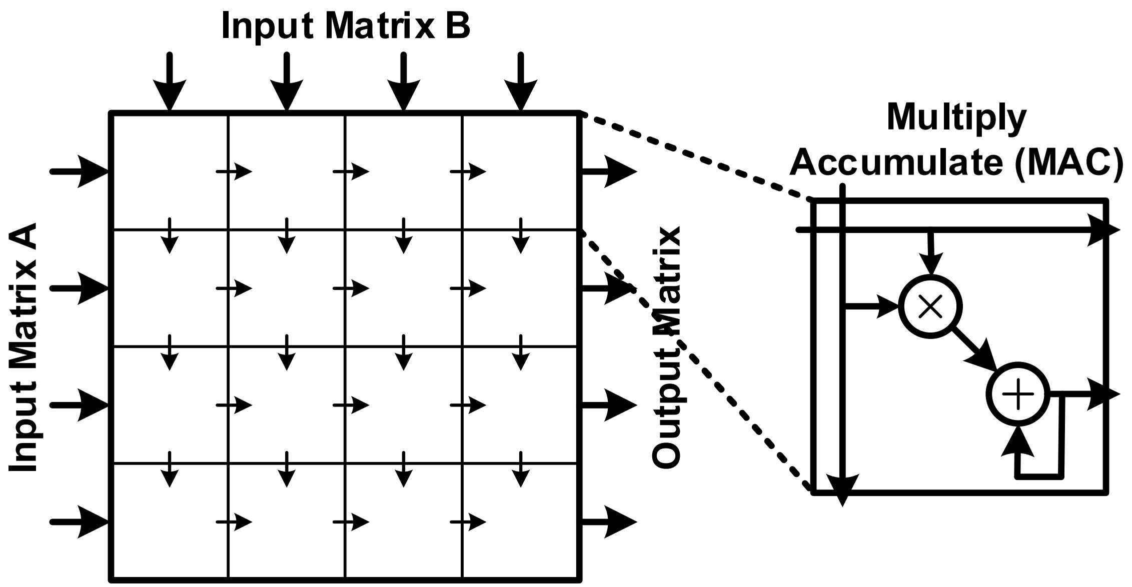


Simple Neural Network

$$\begin{array}{|c|c|c|c|}\hline a00 & a01 & a02 & a03 \\ \hline a10 & a11 & a12 & a13 \\ \hline a20 & a21 & a22 & a23 \\ \hline a30 & a31 & a32 & a33 \\ \hline \end{array} \times \begin{array}{|c|c|c|c|}\hline b00 & b01 & b02 & b03 \\ \hline b10 & b11 & b12 & b13 \\ \hline b20 & b21 & b22 & b23 \\ \hline b30 & b31 & b32 & b33 \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline y00 & y01 & y02 & y03 \\ \hline y10 & y11 & y12 & y13 \\ \hline y20 & y21 & y22 & y23 \\ \hline y30 & y31 & y32 & y33 \\ \hline \end{array}$$

- Matrix-multiply: power, performance, and area limiter
- Large matrices with many iterations
- Specialized architectures enable higher performance and energy efficiency
- Varying numeric requirements (FP16/INT16/INT8) across applications
  - Require low overhead reconfigurable circuits
- Varying matrix sparsity across applications
  - Optimized circuits can take advantage of sparsity

Variable Precision Matrix Multiply Accelerator

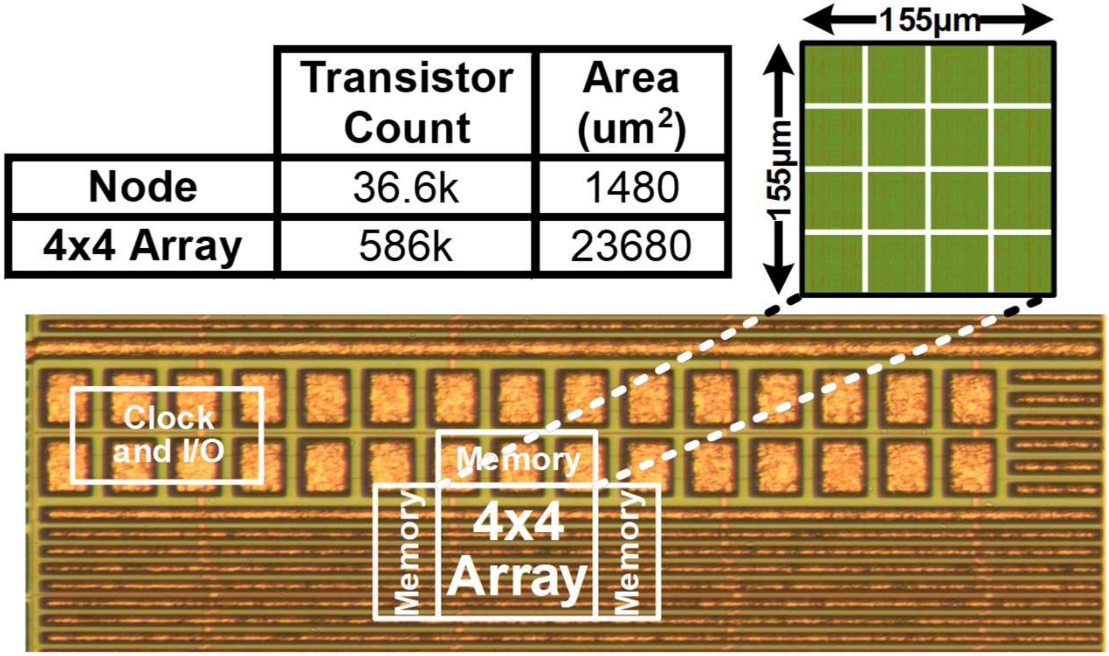


- 4x4 systolic array
- Fabric reconfigures to optimize data movement in dense/sparse mode
- Reconfigurable MAC with signed/unsigned INT16/4xINT8/FP16 support

M. Anders, R. Krishnamurthy et al, VLSI Circuits Symposium 2018

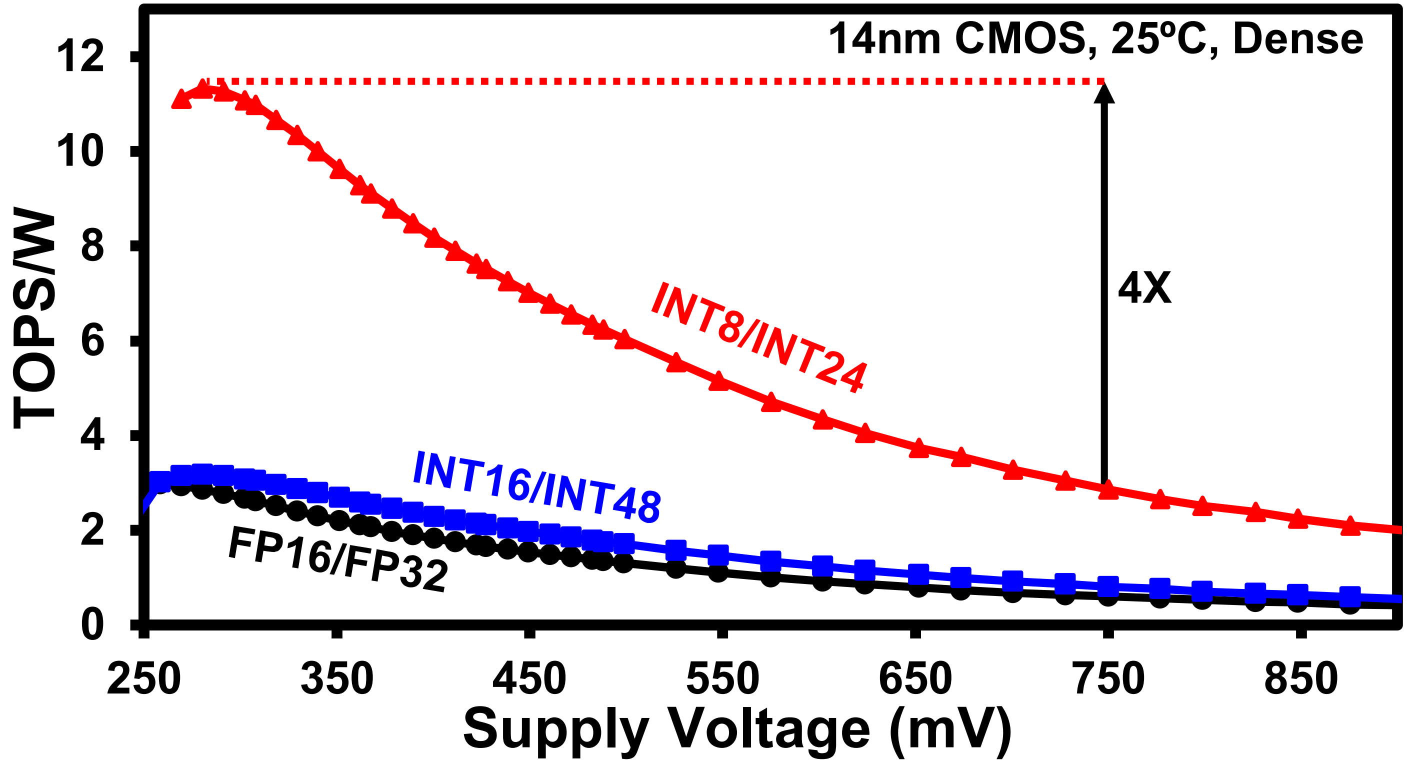
14nm Chip Micrograph and Nominal Performance

	Transistor Count	Area (um^2)
Node	36.6k	1480
4x4 Array	586k	23680



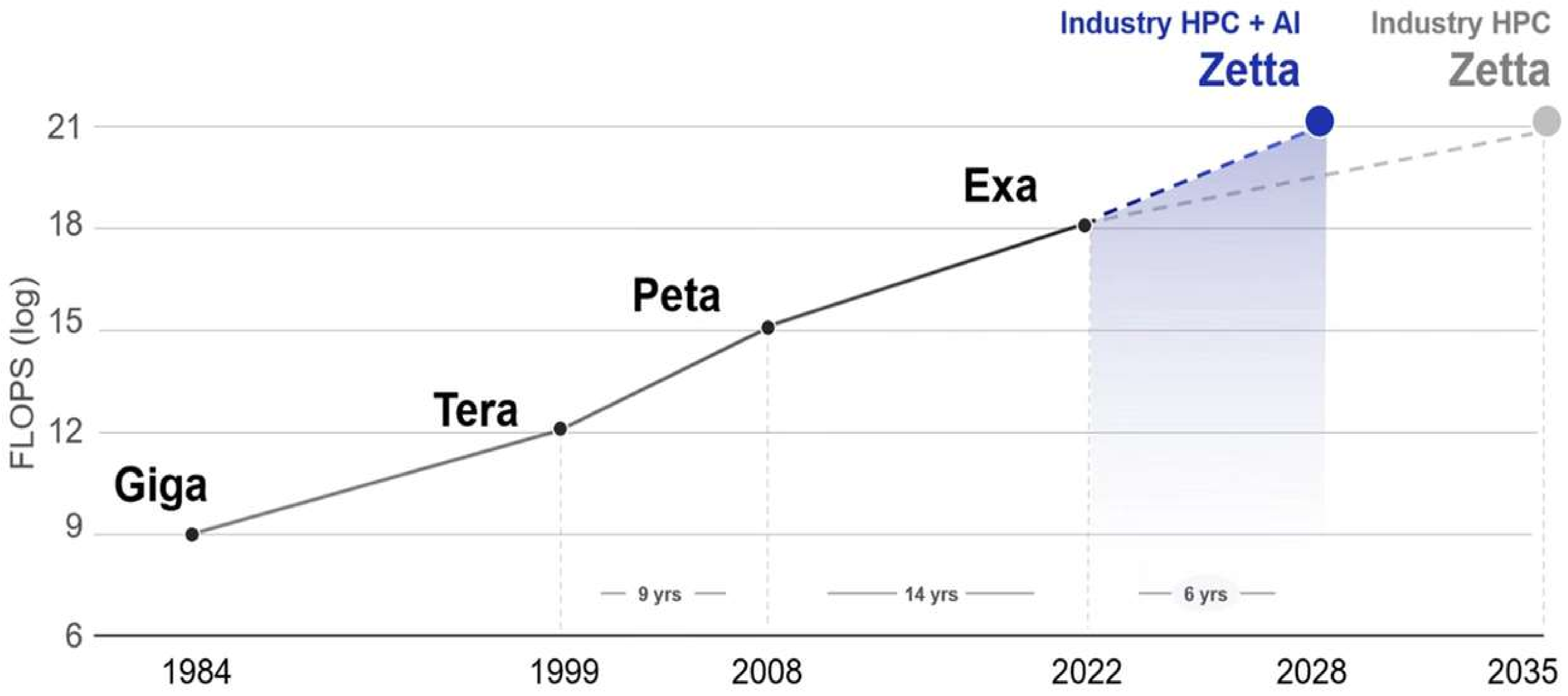
Mult/Acc Mode	Nominal (750mV, 25ºC)
FP16/FP32	800MHz, 42.7mW, 0.6TFLOPS/W
INT16/INT48	940MHz, 37.6mW, 0.8TOPS/W
INT8/INT24	1.06GHz, 47.7mW, 2.9TOPS/W

Matrix Multiplier Energy Efficiency Measurements

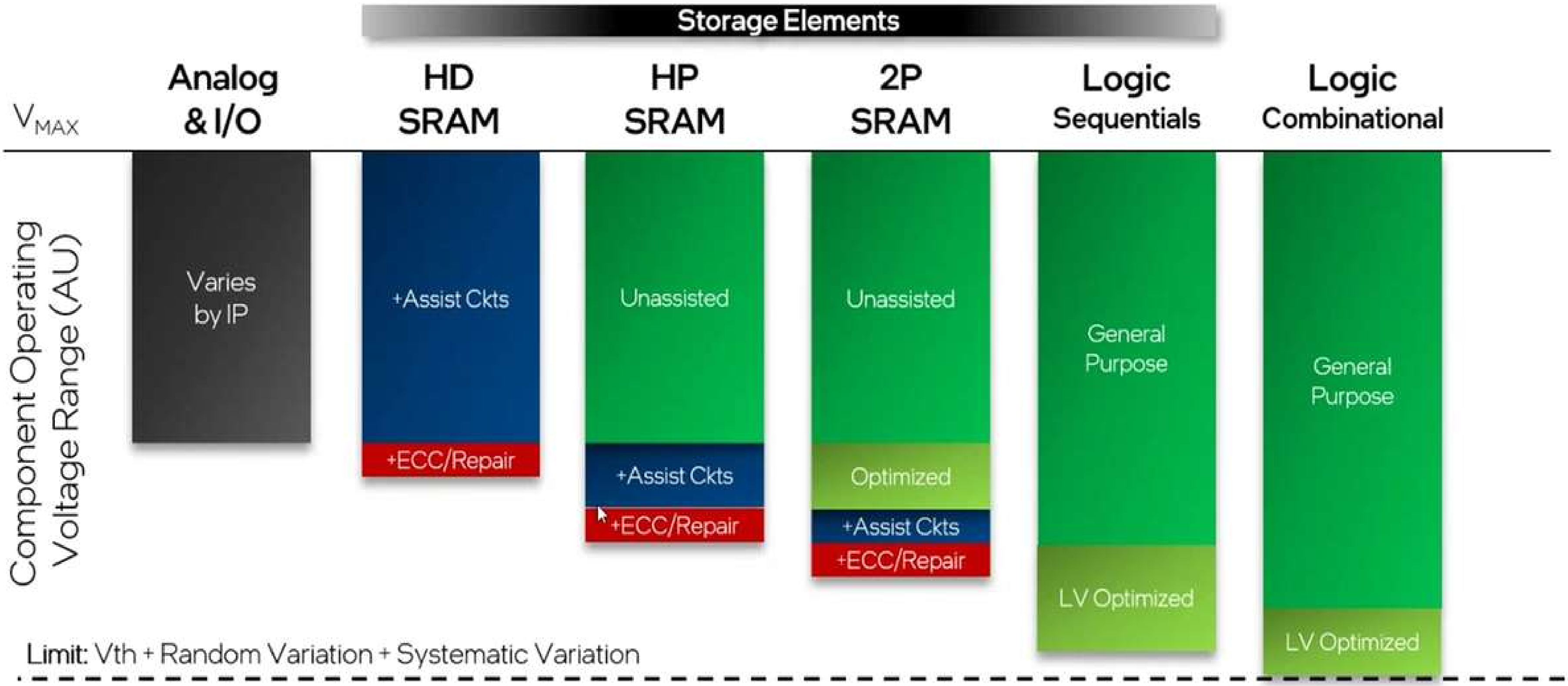


- Efficiency increases 4X from nominal 750mV to near threshold voltage
- Peak energy efficiency range from 2.97TFLOPS/W (FP16) to 11.3TOPS/W (INT8)

The Future – Zetta Flop Systems



Roadmap to Lower Voltage Operation



Storage Elements

V_{MAX}

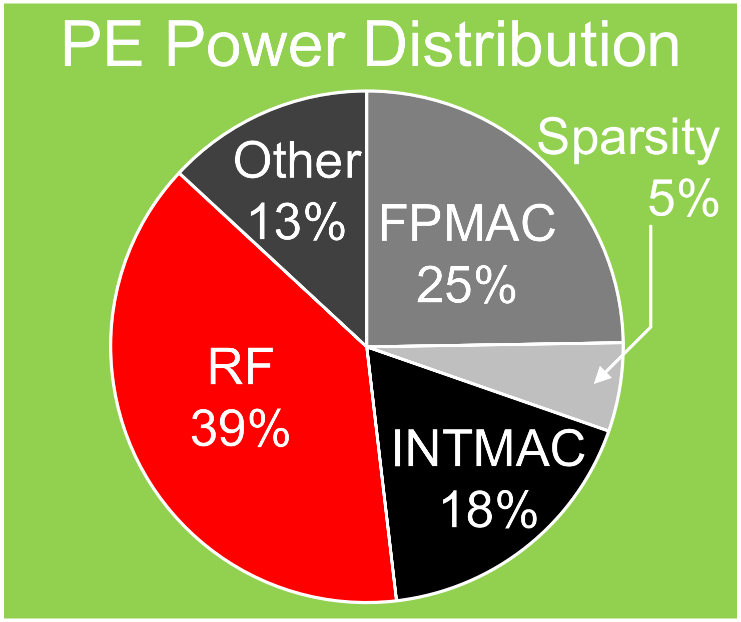
Component Operating Voltage Range (AU)

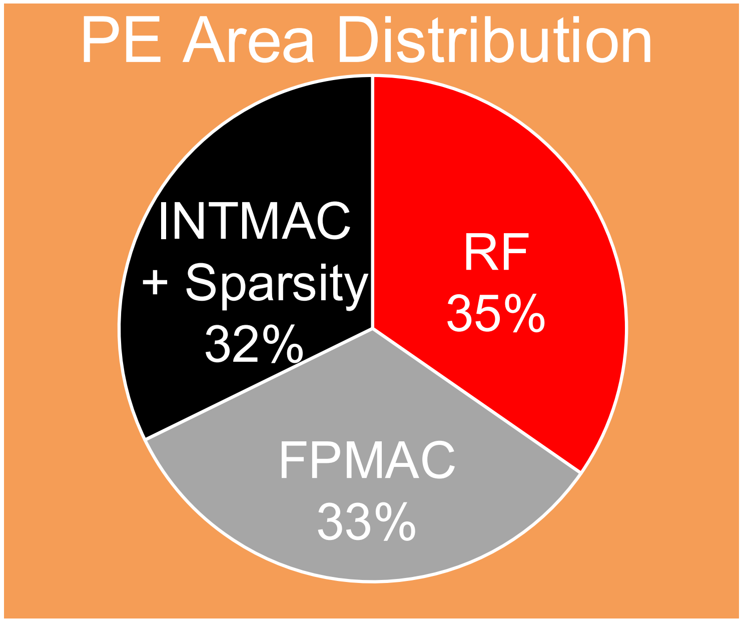
- Analog & I/O: Varies by IP
- HD SRAM: +Assist Ckts; +ECC/Repair
- HP SRAM: Unassisted; +Assist Ckts; +ECC/Repair
- 2P SRAM: Unassisted; Optimized; +Assist Ckts; +ECC/Repair
- Logic Sequentials: General Purpose; LV Optimized
- Logic Combinational: General Purpose; LV Optimized

Limit: Vth + Random Variation + Systematic Variation

Low voltage operation requires careful selection and optimization of storage elements

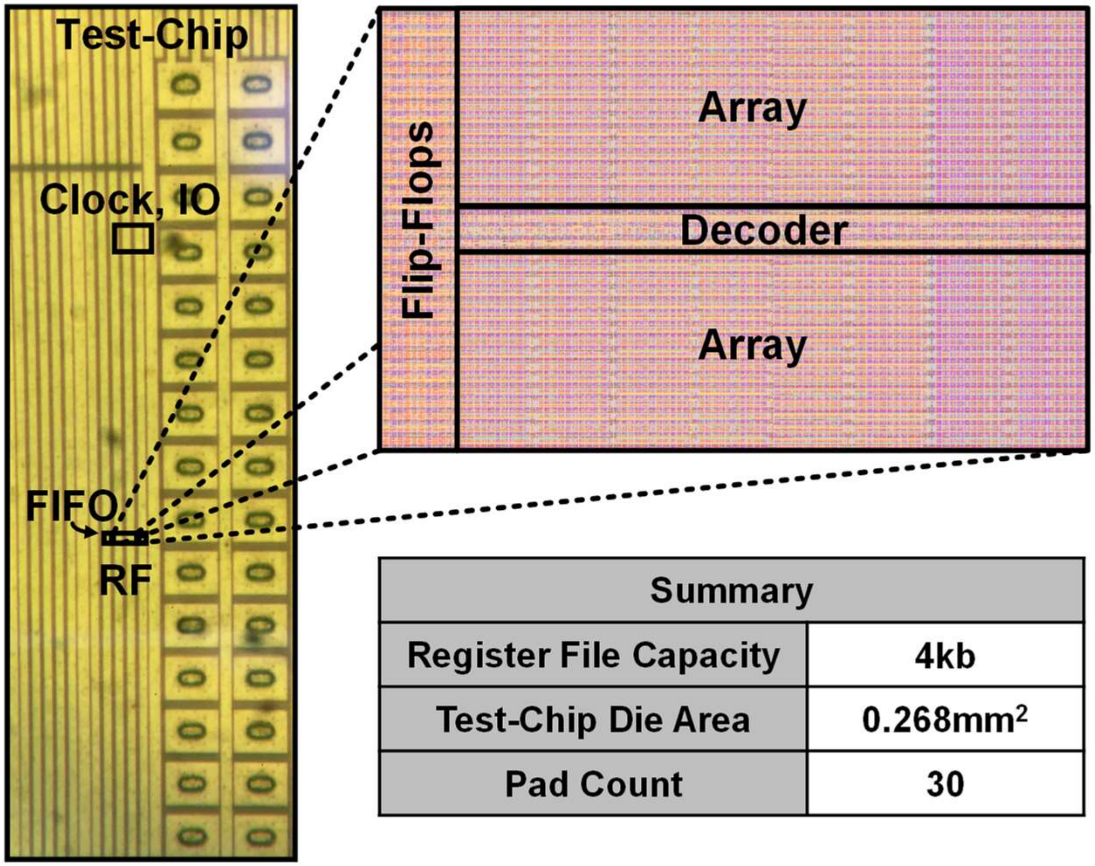
Memory Challenges in AI Accelerators





- AI accelerators are built using a large array of processing elements (PEs) containing small capacity local register files (RFs)
- Register files contribute a significant amount of power (39%) and area (35%) within the PEs

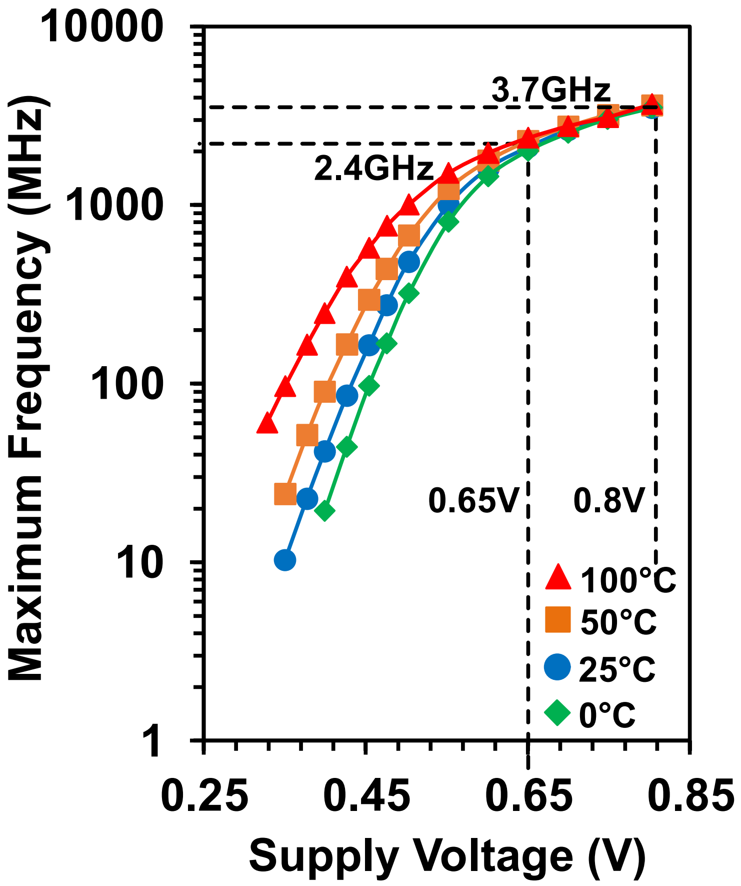
Static AI Register File Micrograph

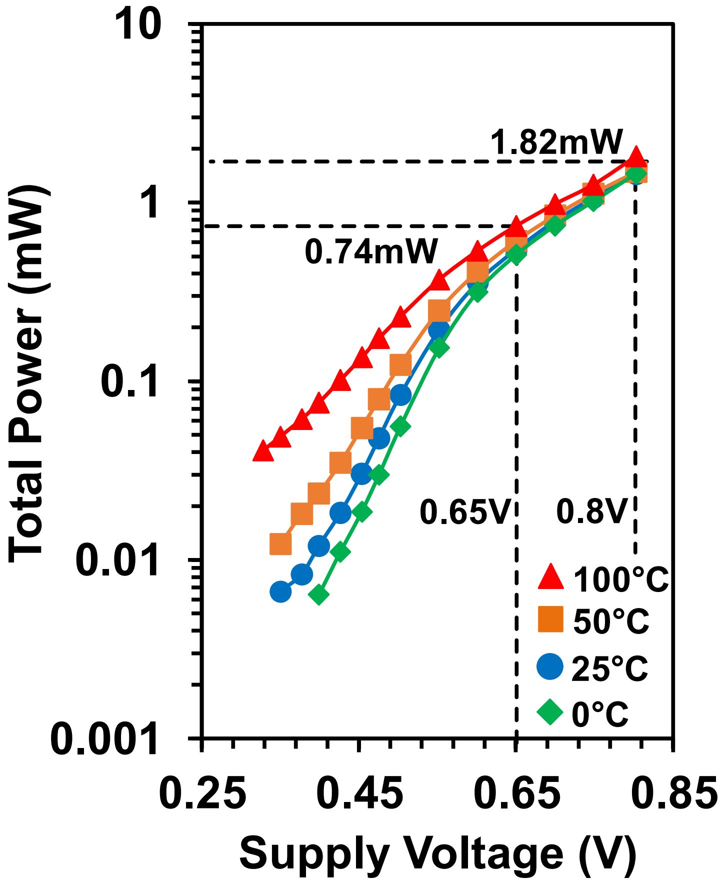


Summary	
Register File Capacity	4kb
Test-Chip Die Area	$0.268mm^2$
Pad Count	30

S. Hsu, R. Krishnamurthy et al, IEEE VLSI Circuits Symposium 2022

Register File Frequency/Power Measurements



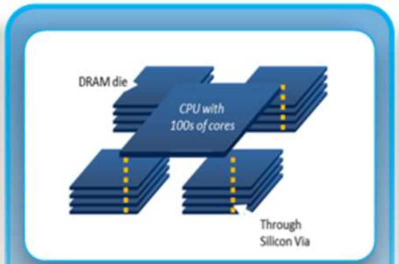


- Ultra-low voltage operation at 325mV, 100°C consuming 36.7µW, 60MHz

“Extreme” efficiency research

Extreme Energy Efficiency

Fine-Grain Power Management



Efficient Memory Subsystem

Self-Aware Computing Operation

Programming for Extreme Parallelism

System-Wide Breakthroughs Needed Across the Board

intel Look Inside™

intelligence Inside