High Performance and Energy Efficient Circuit Technologies for Sub-7nm AI Accelerators and In-Memory/Near-Memory Computing

2021 NSF Workshop on Processing In Memory Technology, February 2021

Ram K. Krishnamurthy
Senior Principal Engineer, 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

Motivation: ML in IoT Platforms

NETWORK + EDGE COMPUTING
ACCELERATED BY 5G

EDGE COMPUTING | NETWORK EDGE

AUTONOMOUS DRIVING
SILICON + SERVICES
~$10B

IOT
SILICON
~$30B

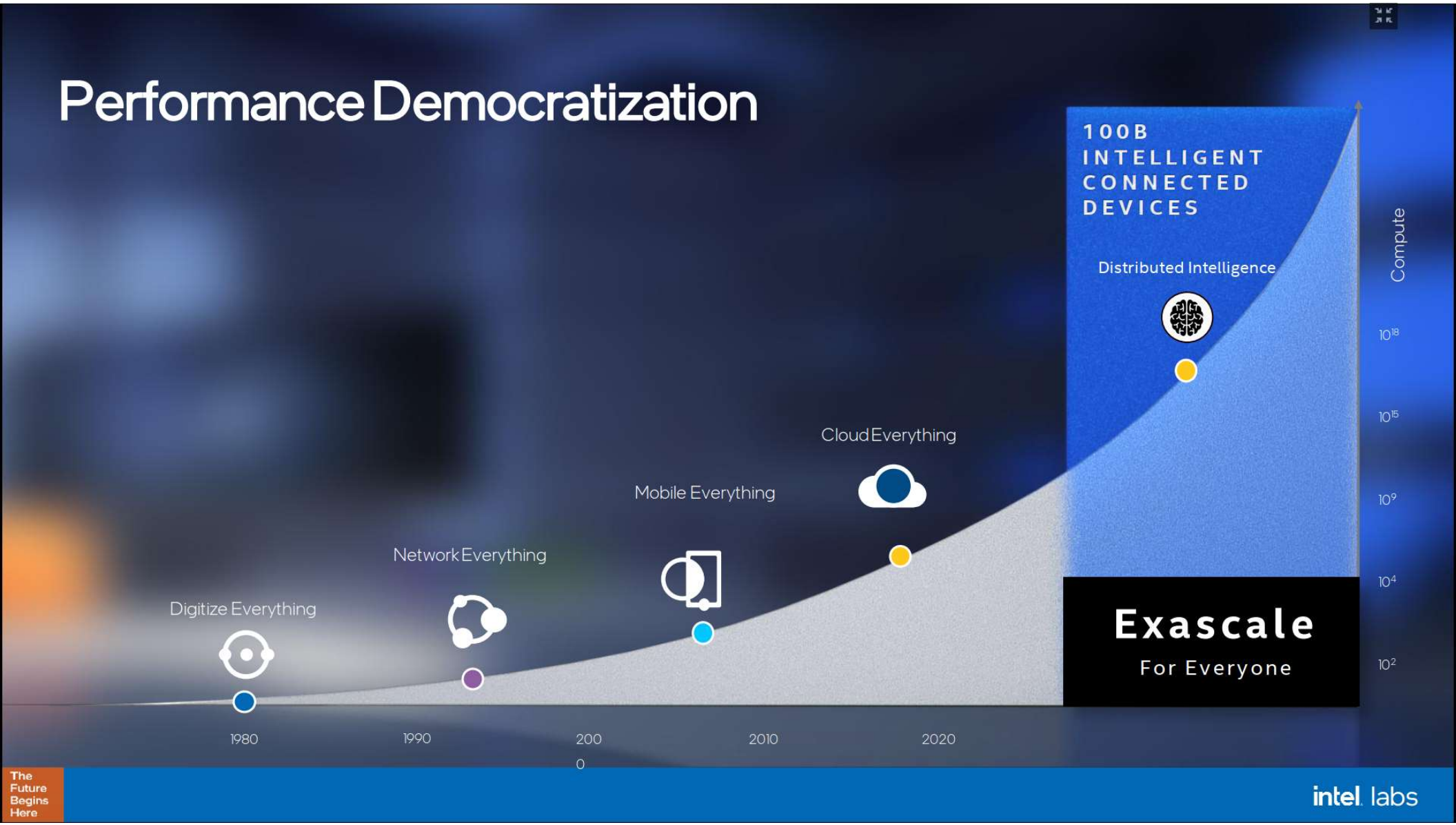
NETWORK
SILICON
~$25B

~$65B OPPORTUNITY BY 2023

2018 INTEL REVENUE
$9.5B
>20% GROWTH YOY

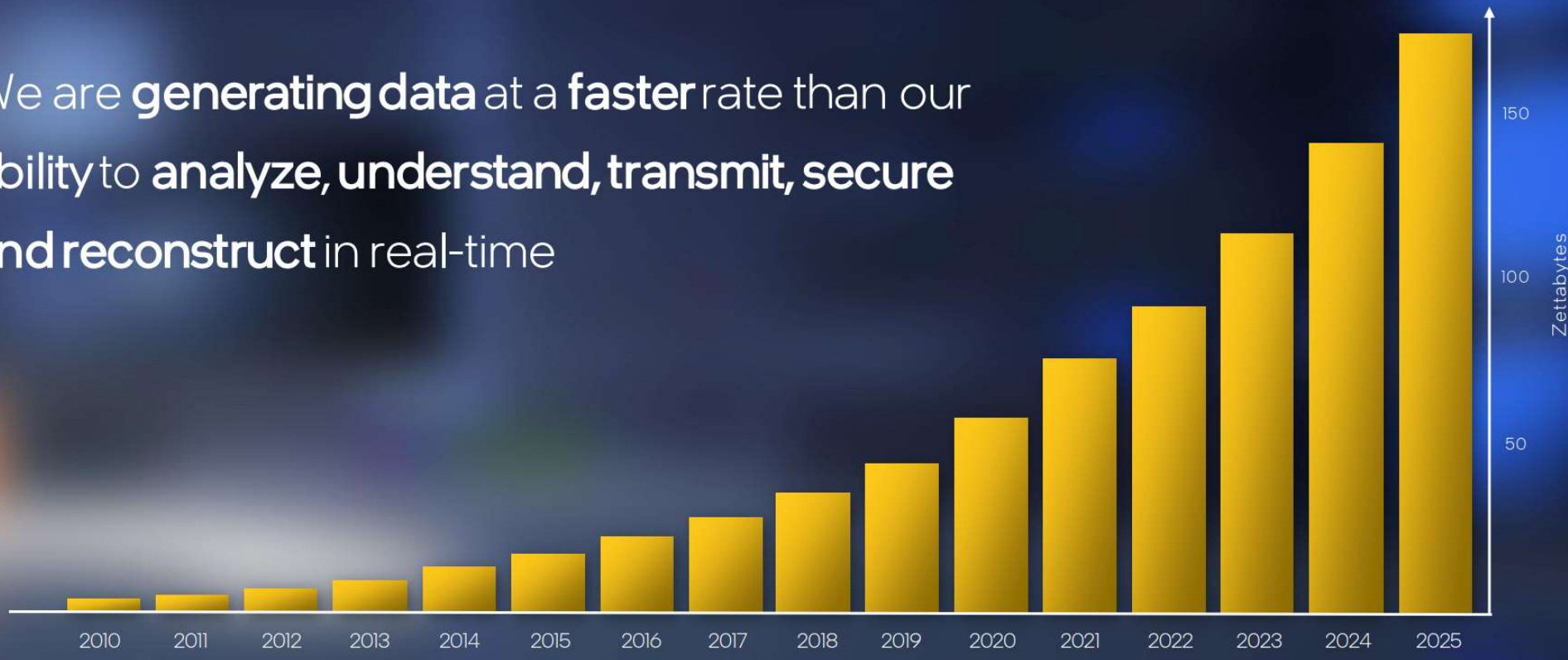
intel 2019 INVESTOR MEETING

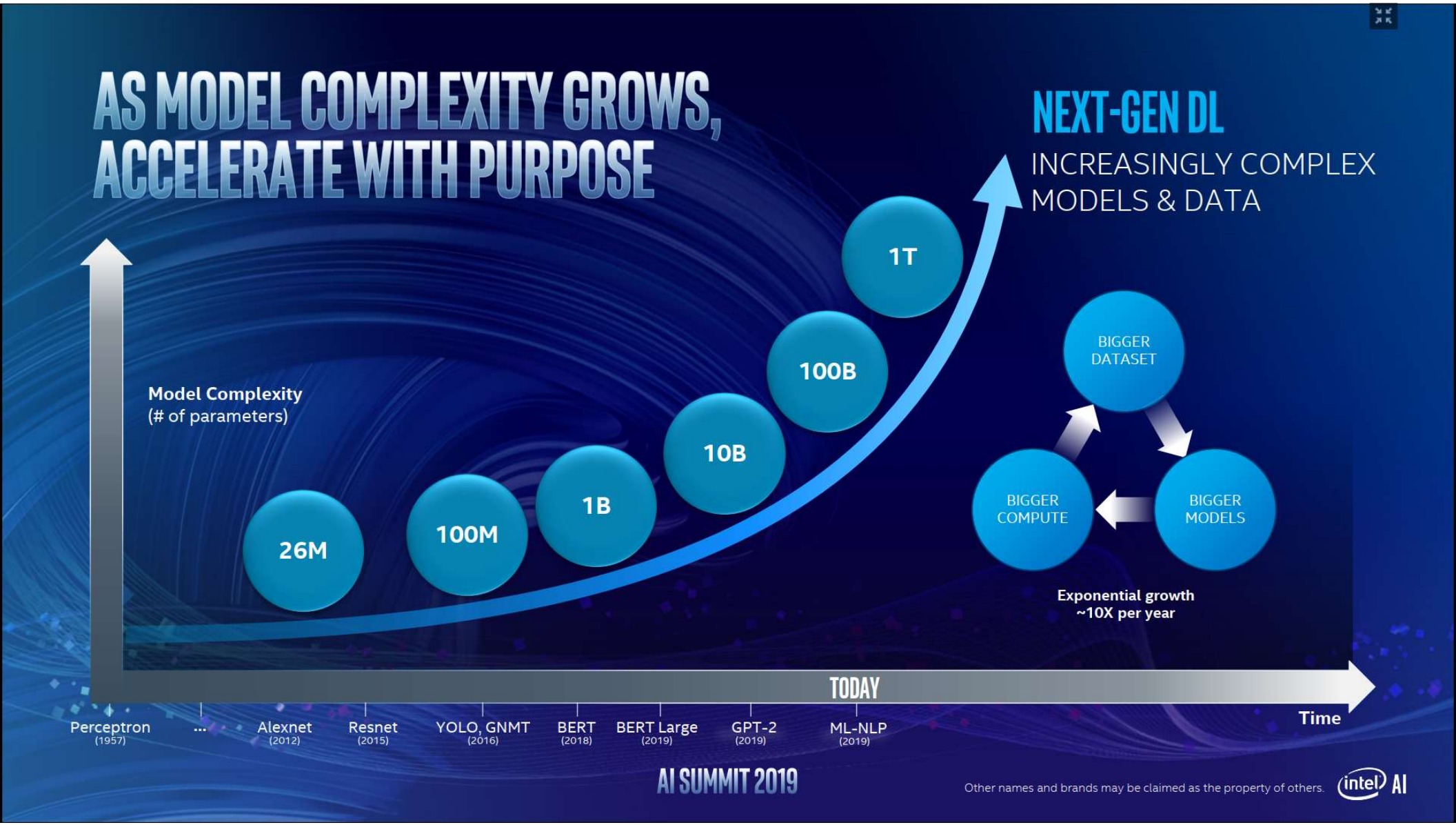
Source: Amalgamation of analyst data and Intel analysis. 2018 revenue based on a portion of DCG, PSG, IOTG and Mobileye.

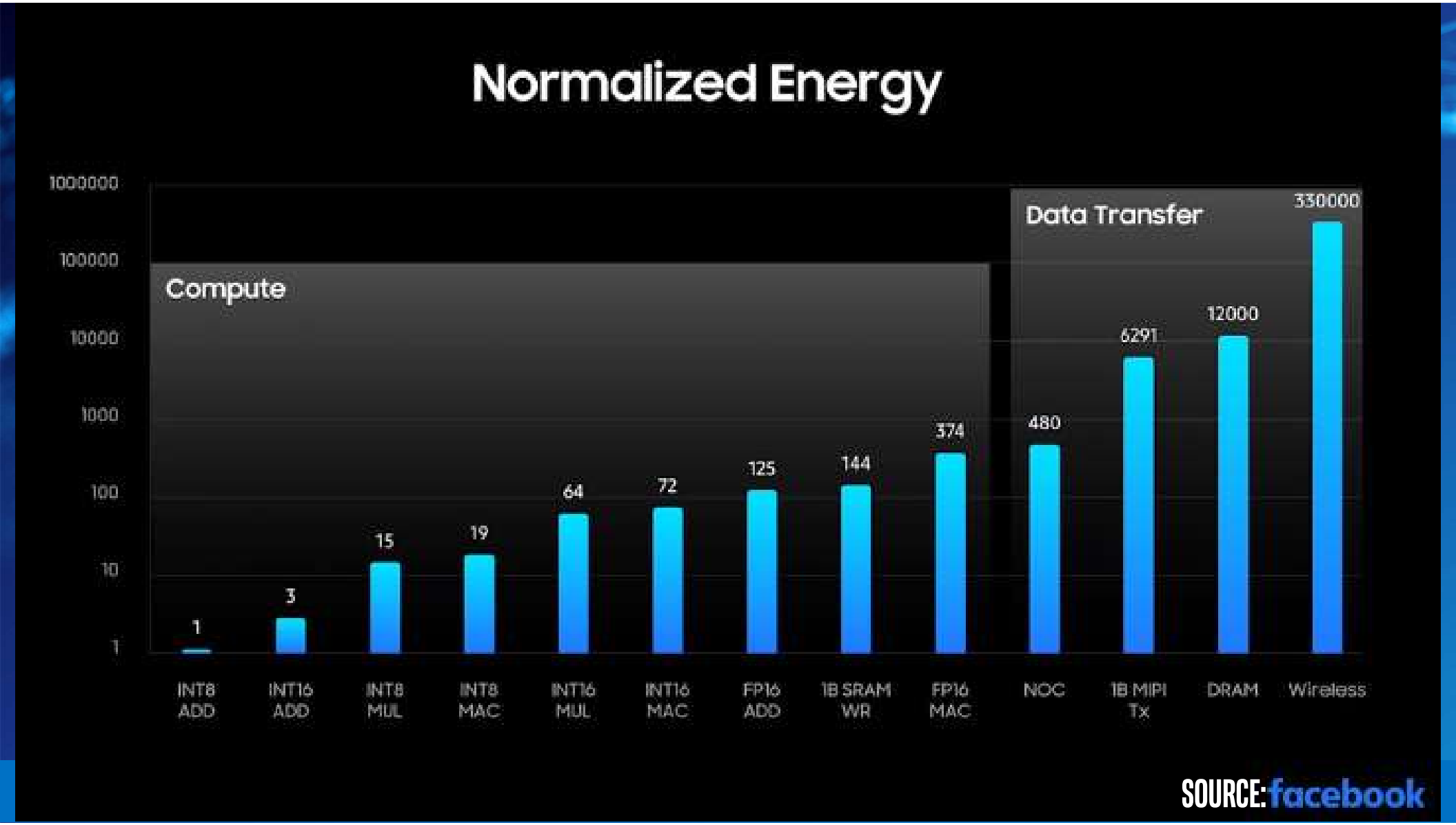

Performance Democratization
100B INTELLIGENT CONNECTED DEVICES
Distributed Intelligence
Cloud Everything
Mobile Everything
Network Everything
Digitize Everything
Exascale
For Everyone
Compute
10^18
10^15
10^9
10^4
10^2
1980
1990
2000
2010
2020
The Future Begins Here
intel labs

The Data Problem

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




AS MODEL COMPLEXITY GROWS, ACCELERATE WITH PURPOSE
NEXT-GEN DL
INCREASINGLY COMPLEX MODELS & DATA
Model Complexity (# of parameters)
26M
100M
1B
10B
100B
1T
BIGGER DATASET
BIGGER MODELS
BIGGER COMPUTE
Exponential growth ~10X per year
TODAY
Time
Perceptron (1957)
...
Alexnet (2012)
Resnet (2015)
YOLO, GNMT (2016)
BERT (2018)
BERT Large (2019)
GPT-2 (2019)
ML-NLP (2019)
AI SUMMIT 2019
Other names and brands may be claimed as the property of others.
intel AI


Normalized Energy
1000000
100000
10000
1000
100
10
1
Compute
Data Transfer
1
3
15
19
64
72
125
144
374
480
6291
12000
330000
INT8 ADD
INT16 ADD
INT8 MUL
INT8 MAC
INT16 MUL
INT16 MAC
FP16 ADD
1B SRAM WR
FP16 MAC
NOC
1B MIPI Tx
DRAM
Wireless
SOURCE: facebook

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

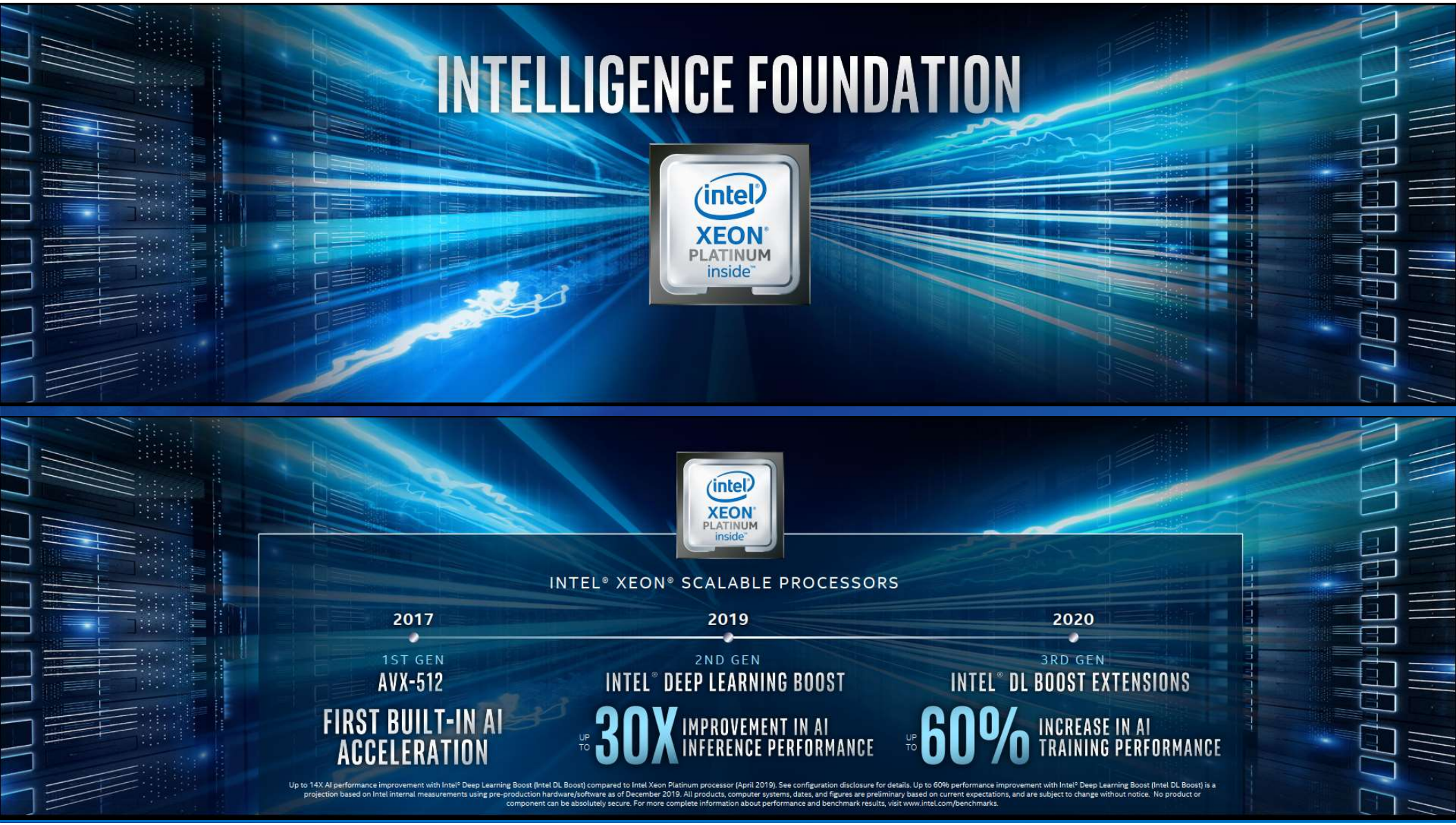
EMBEDDED ARTIFICIAL INTELLIGENCE **ACCELERATION**

HARDWARE ENHANCED SECURITY

ENHANCED AGILITY & UTILIZATION

INTEL.COM/XEONSCALABLE

UNDER EMBARGO UNTIL 2 APRIL 2019 AT 10:00 AM (PACIFIC TIME)


INTELLIGENCE FOUNDATION
intel
XEON
PLATINUM
inside™
intel
XEON
PLATINUM
inside™
INTEL® XEON® SCALABLE PROCESSORS
2017
1ST GEN
AVX-512
FIRST BUILT-IN AI ACCELERATION
2019
2ND GEN
INTEL® DEEP LEARNING BOOST
UP TO 30X IMPROVEMENT IN AI INFERENCE PERFORMANCE
2020
3RD GEN
INTEL® DL BOOST EXTENSIONS
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

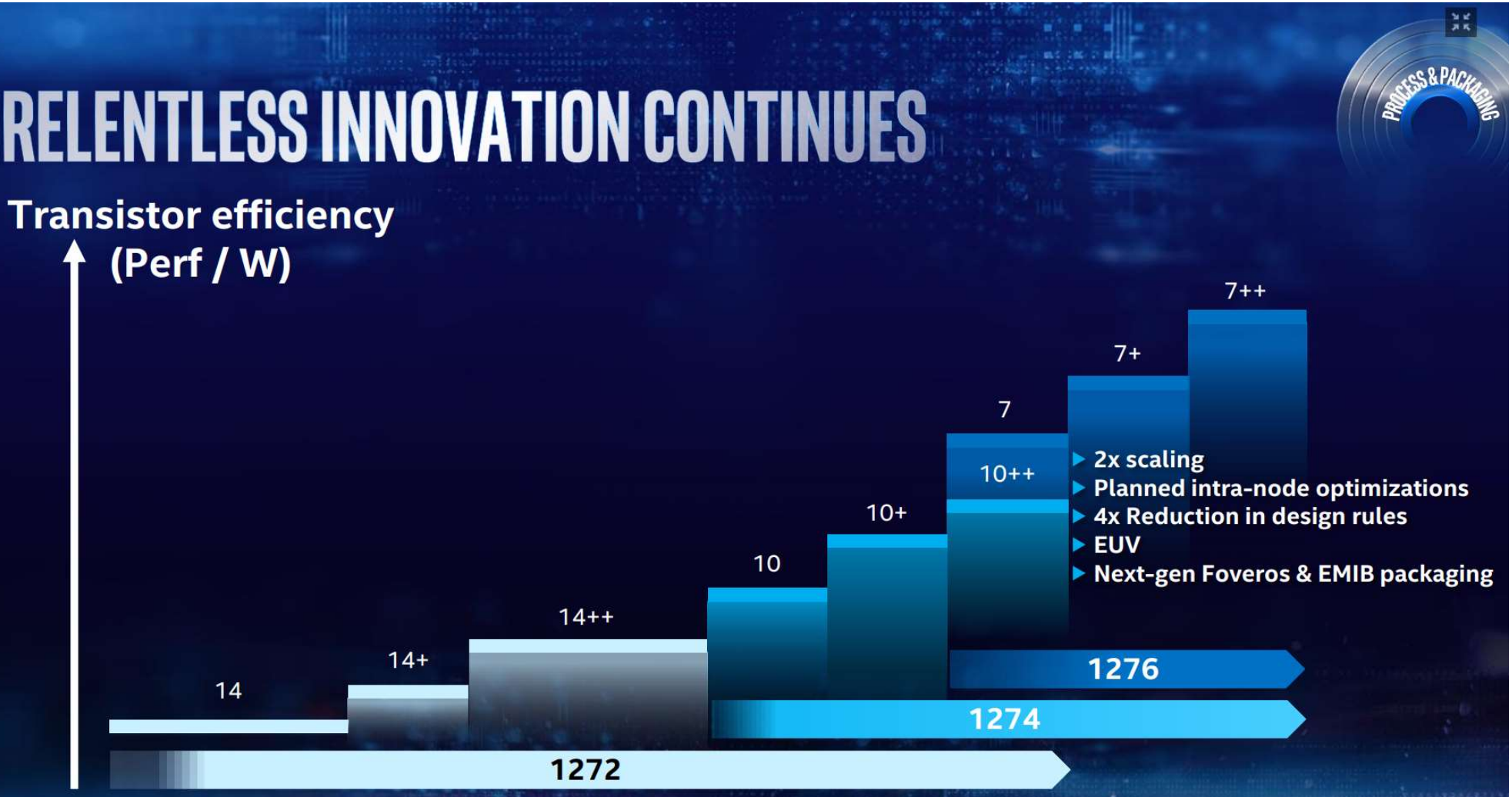
SOURCE: L. LOH, ISSCC 2020

Diversified Workload & Increasing Demands

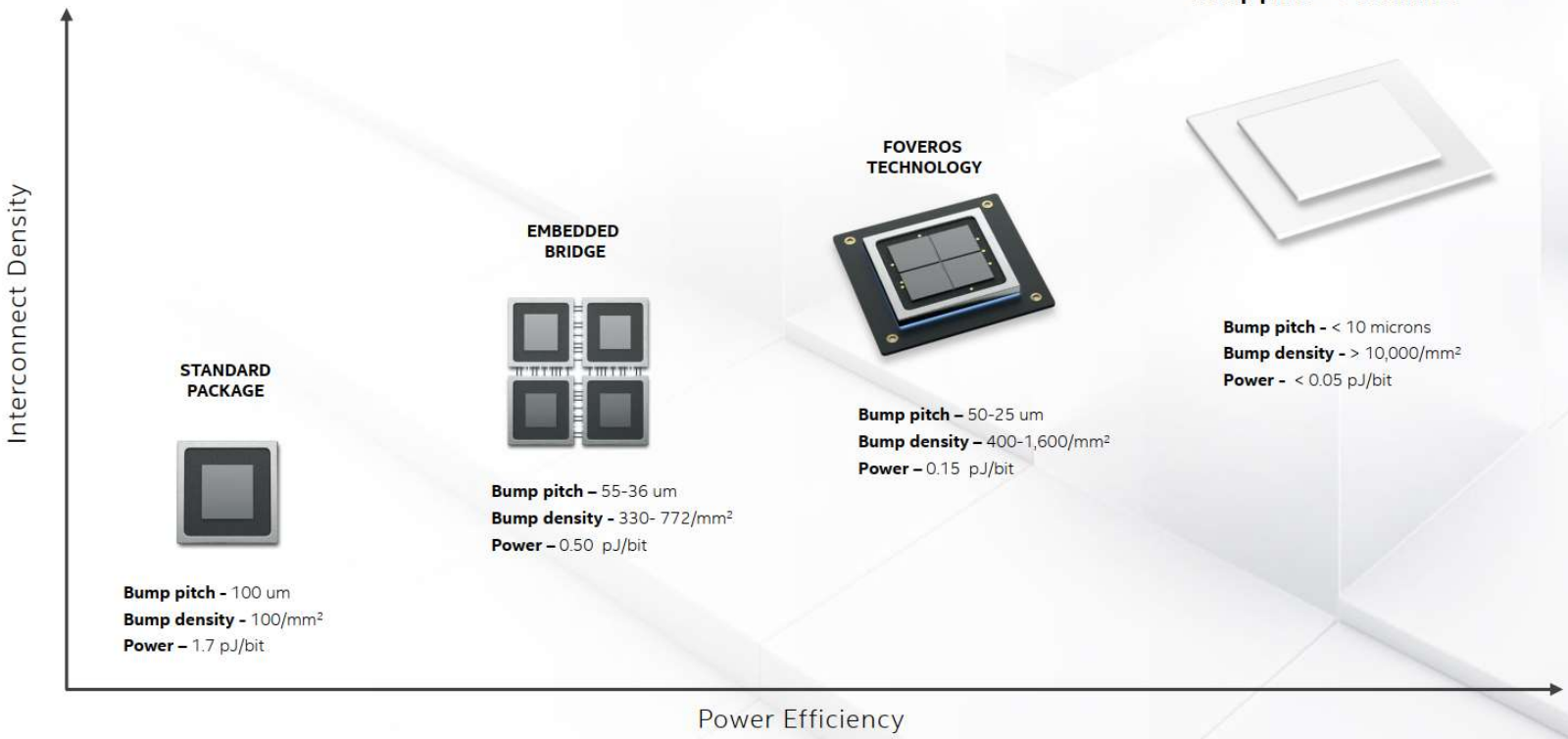
0.1 TOPS 1 TOPS/W	1 TOPS 3 TOPS/W	10 TOPS 10 TOPS/W	100 TOPS 30 TOPS/W
Vision Perception	Vision Construction	Visual Quality	Multi-Streaming

SOURCE: L. LOH, ISSCC 2020

RELENTLESS INNOVATION CONTINUES



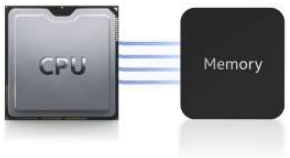
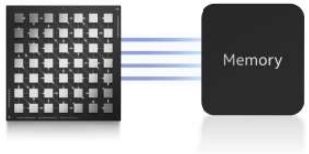
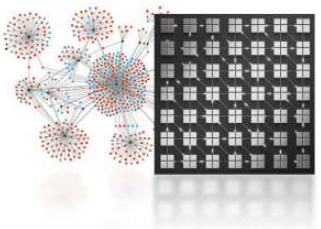
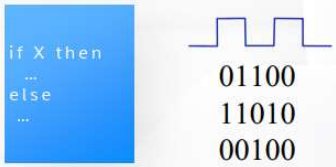
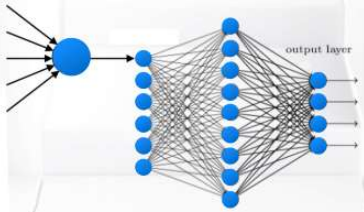
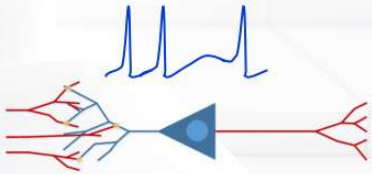
Packaging Technology Roadmap



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

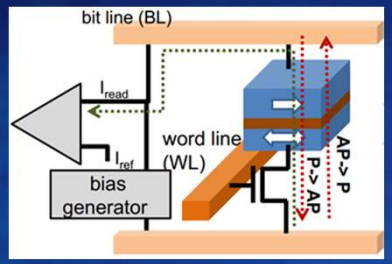
… by thinking differently

… 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
		

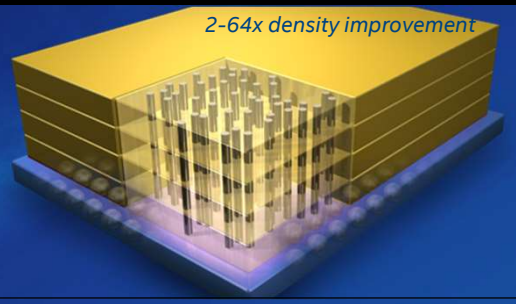
OPPORTUNITIES FOR IN-MEMORY/NEAR-MEMORY PROCESS AND CIRCUIT INNOVATION (BOTH DIGITAL AND ANALOG/MIXED-SIGNAL)

STTRAM STATE



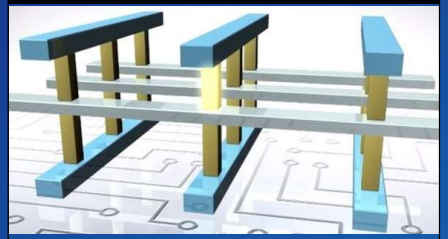
2x density improvement iso-energy

3D



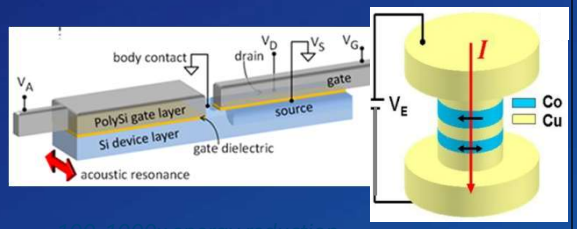
2-64x density improvement

XPOINT MEMORY



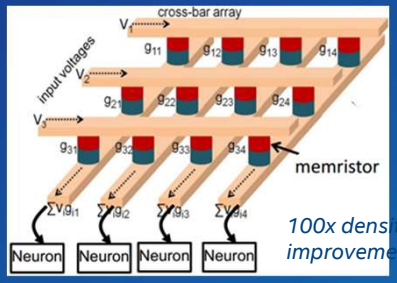
10x density improvement

SINGLE-DEVICE OSCILLATORS



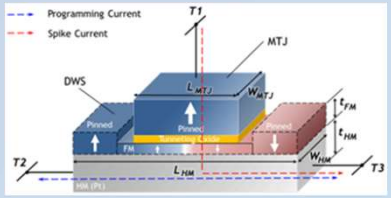
100-1000x energy reduction *[Nikonov, Young]*

RRAM/PCM ANALOG SYNAPSES



100x density improvement

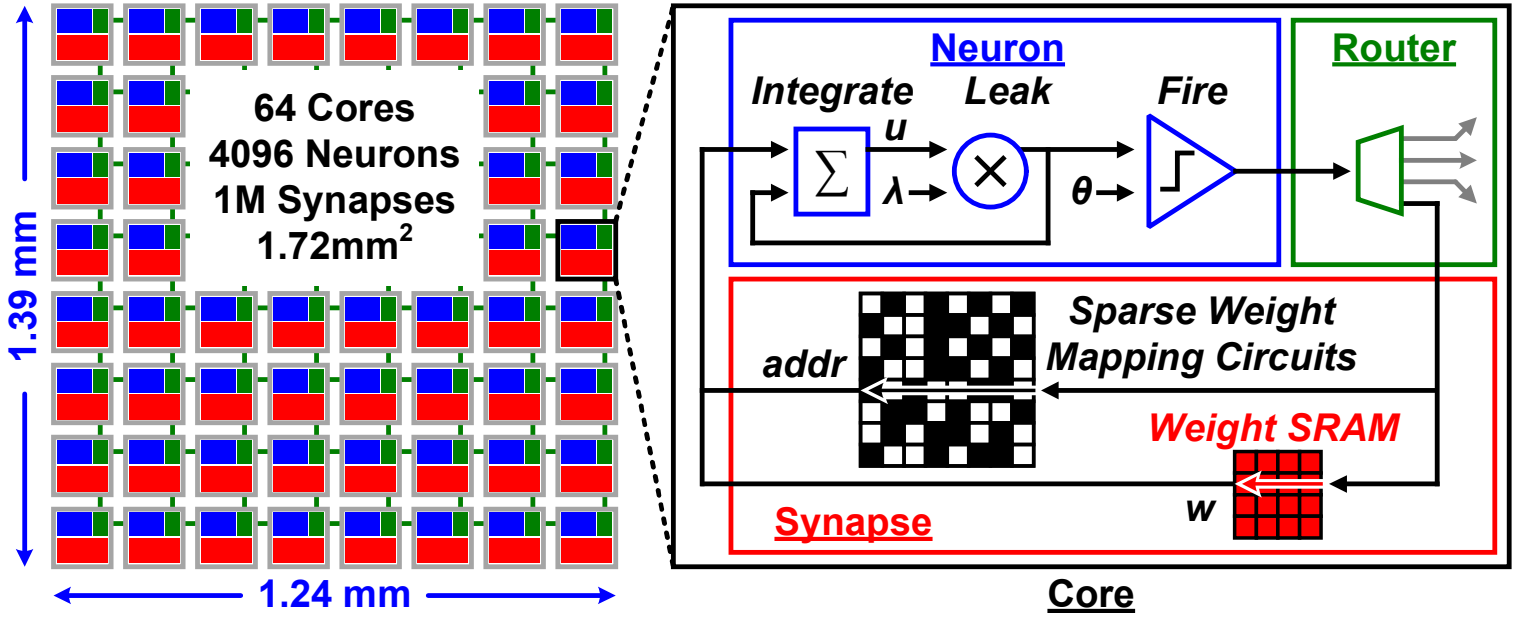
SPIN ORBIT TORQUE SYNAPSES



20x reduction in synaptic energy

[Roy]

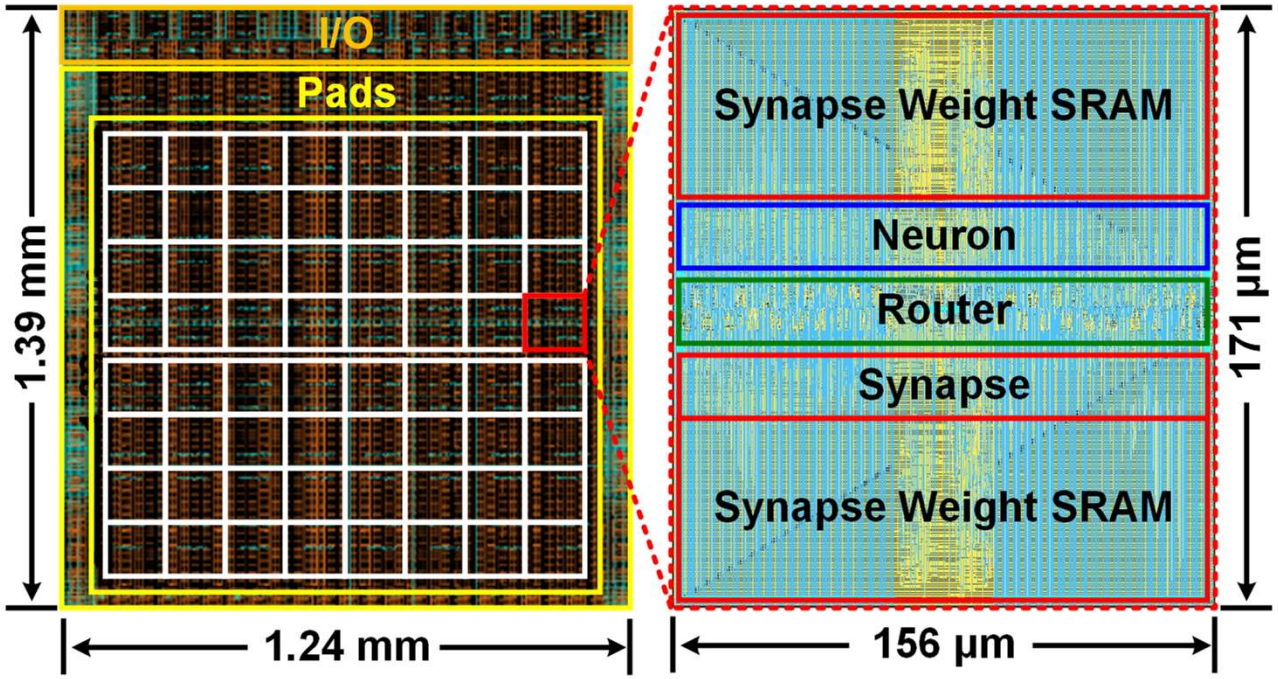
10nm 4096 Neuron Spiking Neural Network Overview



G. Chen, R. Krishnamurthy et al, IEEE Journal of Solid-State Circuits, 2019

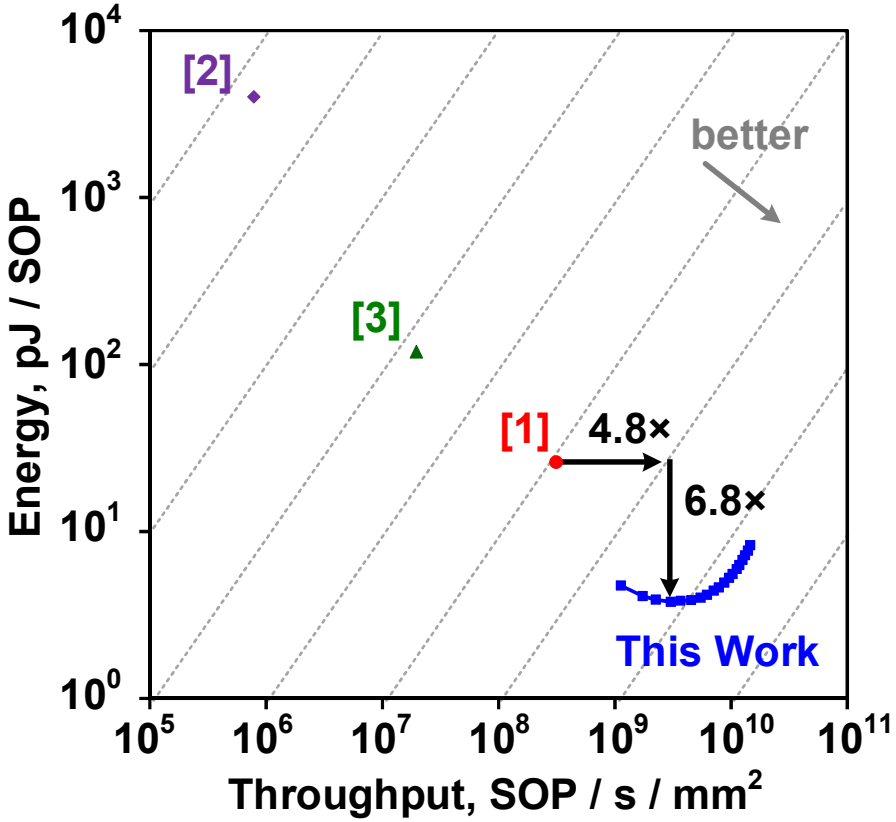
- Neuron stores membrane potentials and performs Leaky Integrate & Fire calculations
- Router multicasts neuron spikes to fan-out synapses in core across the chip
- Synapse stores weights and computes sparse structured connectivity to target neurons

10nm 4096 Neuron Spiking Neural Network Overview



Process	10nm FinFET
Area	1.72 mm^2
Neurons	4096
Synapses	1M × 7b
Throughput	25.2 GSOP/s @ 0.9V
Energy Efficiency	3.8 pJ/SOP @ 525mV
Power / Neuron	2.3 µW @ 450mV

Comparison to Previously Published SNNs

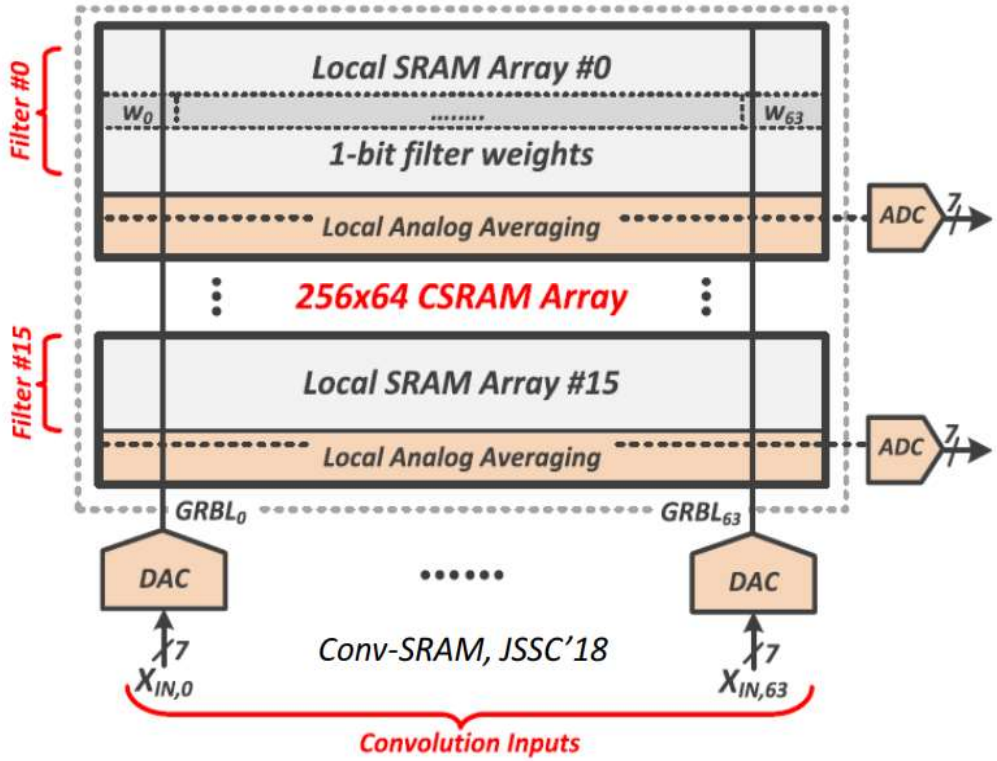


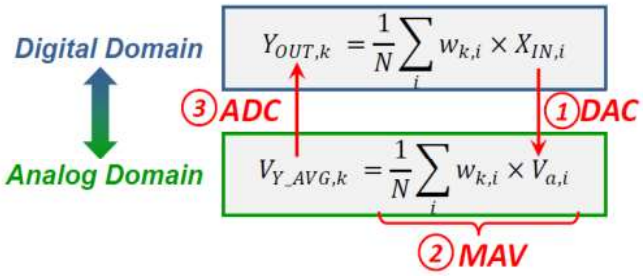
Project	This Work		[1]	[2]	[3]
Process	10nm		28nm	130nm	180nm
Area, mm^2	1.72		430	102	168
Synapse Bits	7b		1b	-	13b
Learning	Yes		No	Yes	No
Sparsity	Yes		No	Yes	No
Voltage	525mV	0.9V	0.775V	1.2V	1.8V
Freq., MHz	105	506	-	180	-
SOP/s/mm^2	3.0G	14.6G	624M	784k	19.6M
pJ/SOP	3.8	8.3	26	4000	119

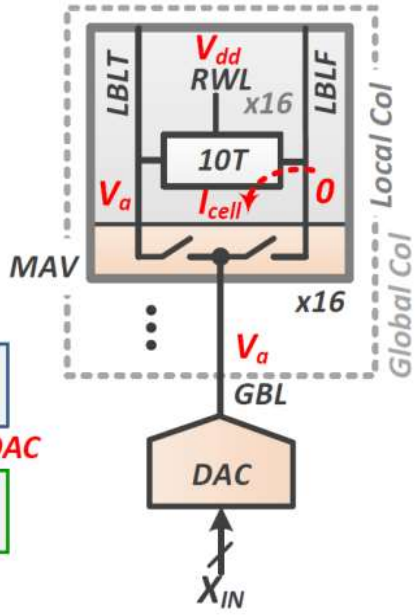
G. Chen, R. Krishnamurthy et al, IEEE Journal of Solid-State Circuits, 2019

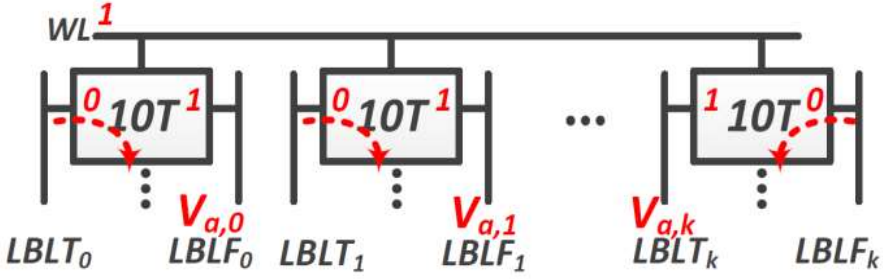
In Memory Computing

- Eliminate weight movements by computing within memory
- Exploit Kirchhoff's law for summation



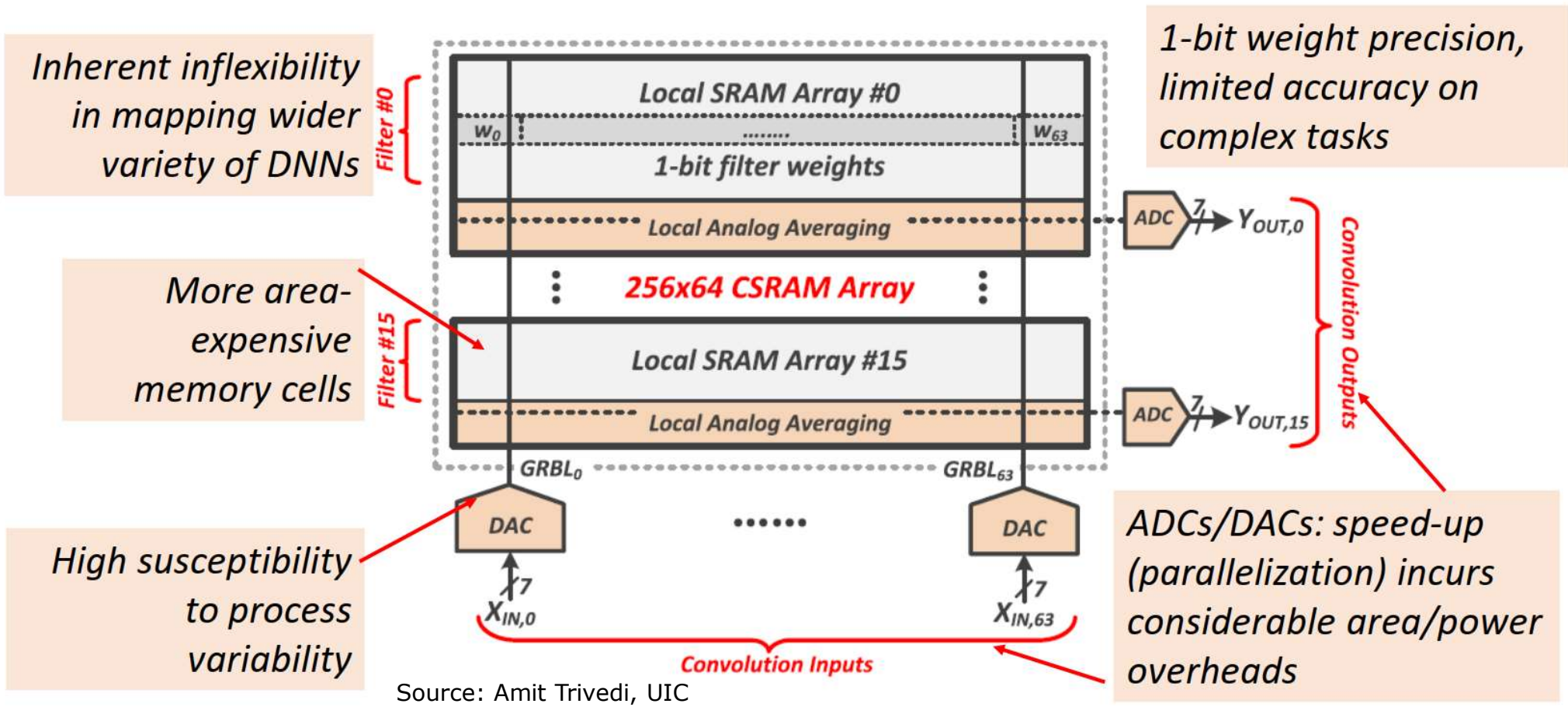




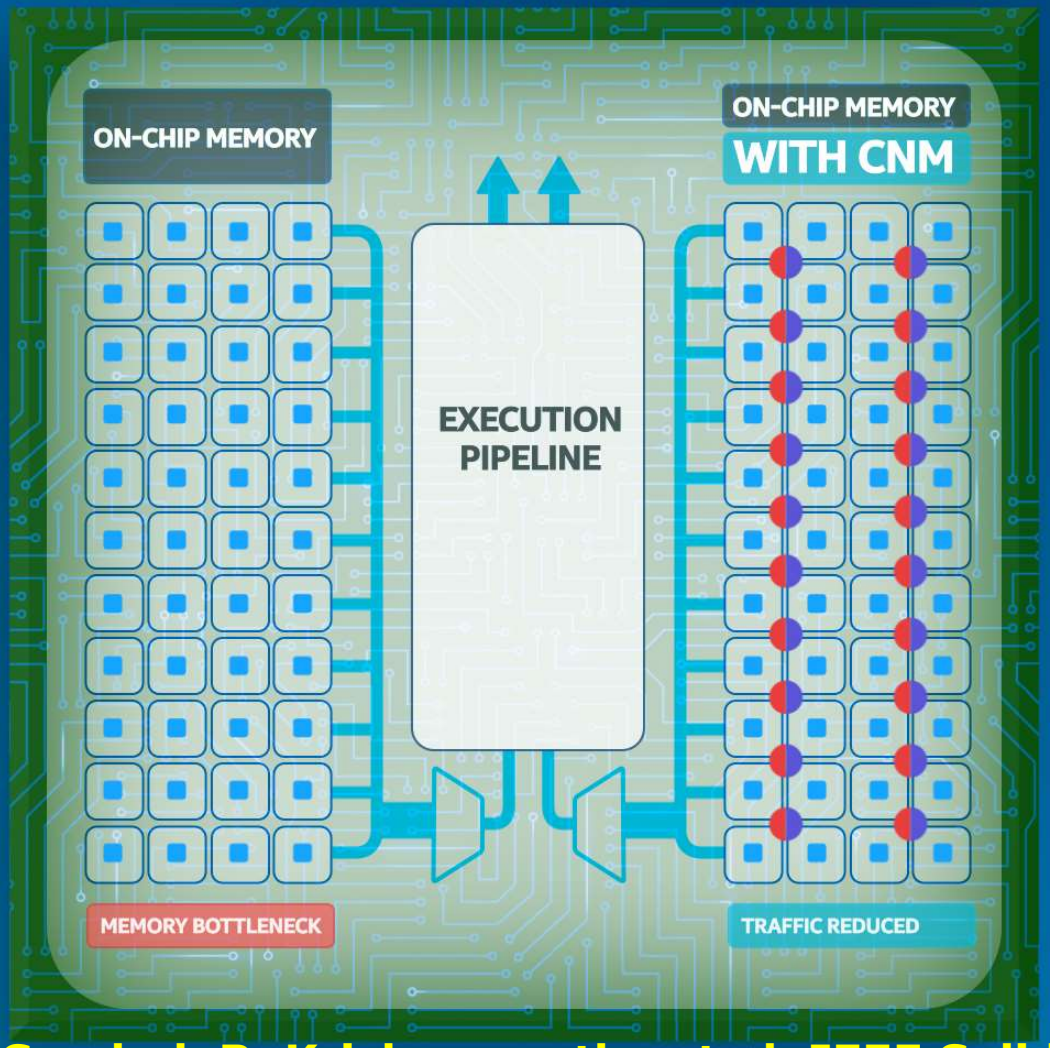


Source: Amit Trivedi, UIC

In Memory Computing Challenges



Source: Amit Trivedi, UIC



COMPUTE NEAR MEMORY (CNM)

- REDUCE MEMORY BOTTLENECK
- INCREASE PERFORMANCE / WATT
- OPTIMIZE FOR AI OPERATIONS

AI WORKLOAD CONFIGURATION 2

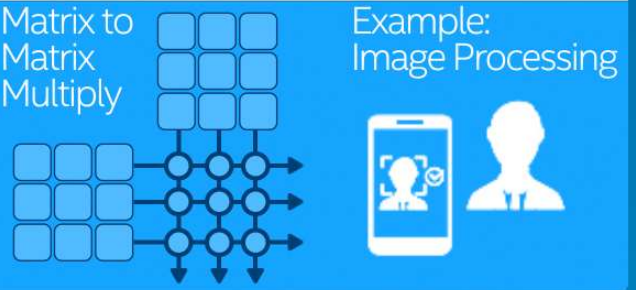
Matrix to Matrix Multiply

Example: Speech Recognition

AI WORKLOAD CONFIGURATION 2

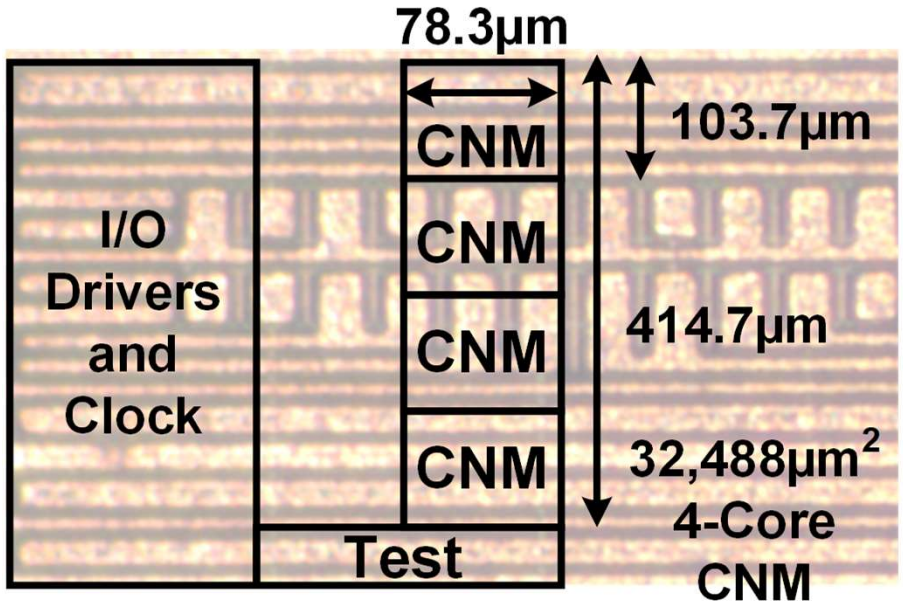
Matrix to Matrix Multiply

Example: Image Processing



E. Sumbul, R. Krishnamurthy et al, IEEE Solid-State Circuits Letters, 2020

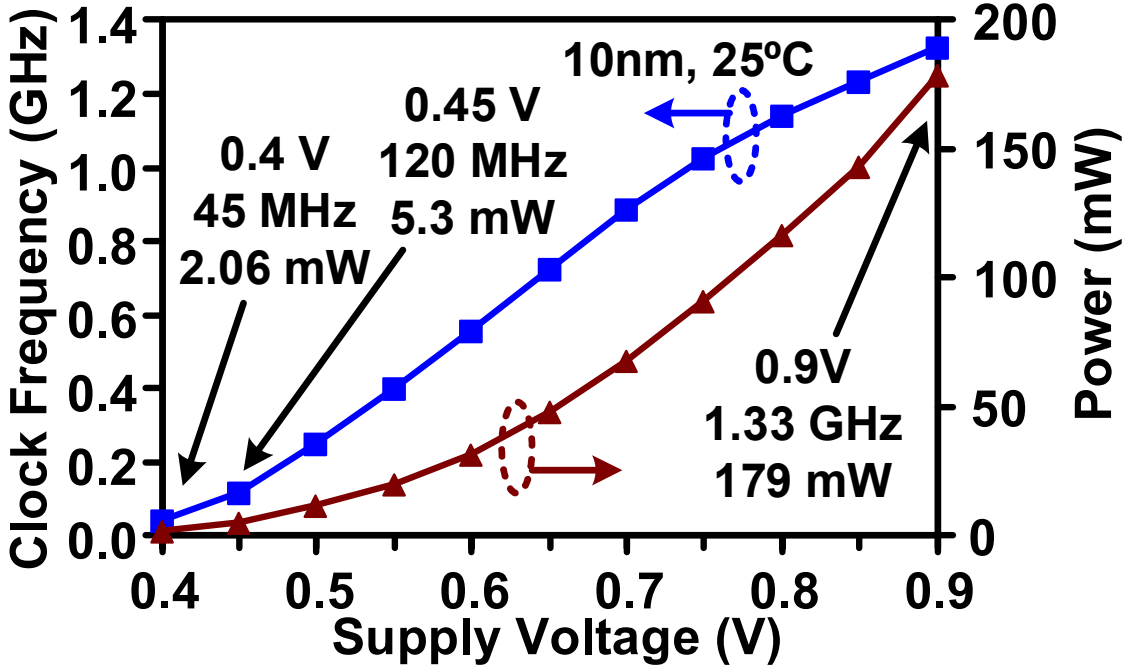
10nm Near Memory Computing AI Accelerator

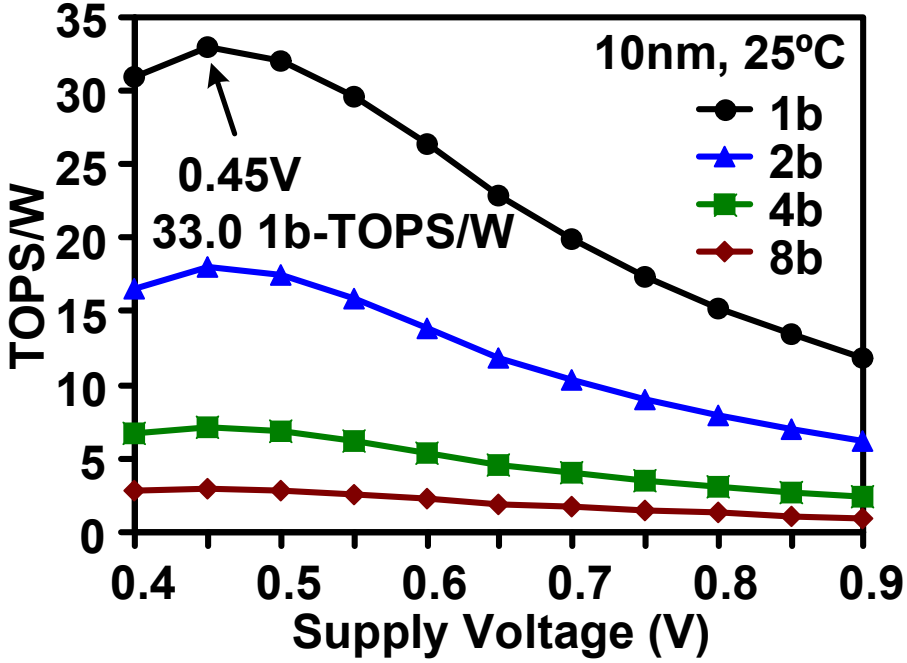


Precision	8b	4b	2b	1b
MACs	64	128	256	512
GOPS 0.9V	170	340	681	1362
TOPS/mm² 0.9V	5.2	10.5	21.0	41.9
TOPS/W 0.45V	2.9	7.1	18.0	33.0

- 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

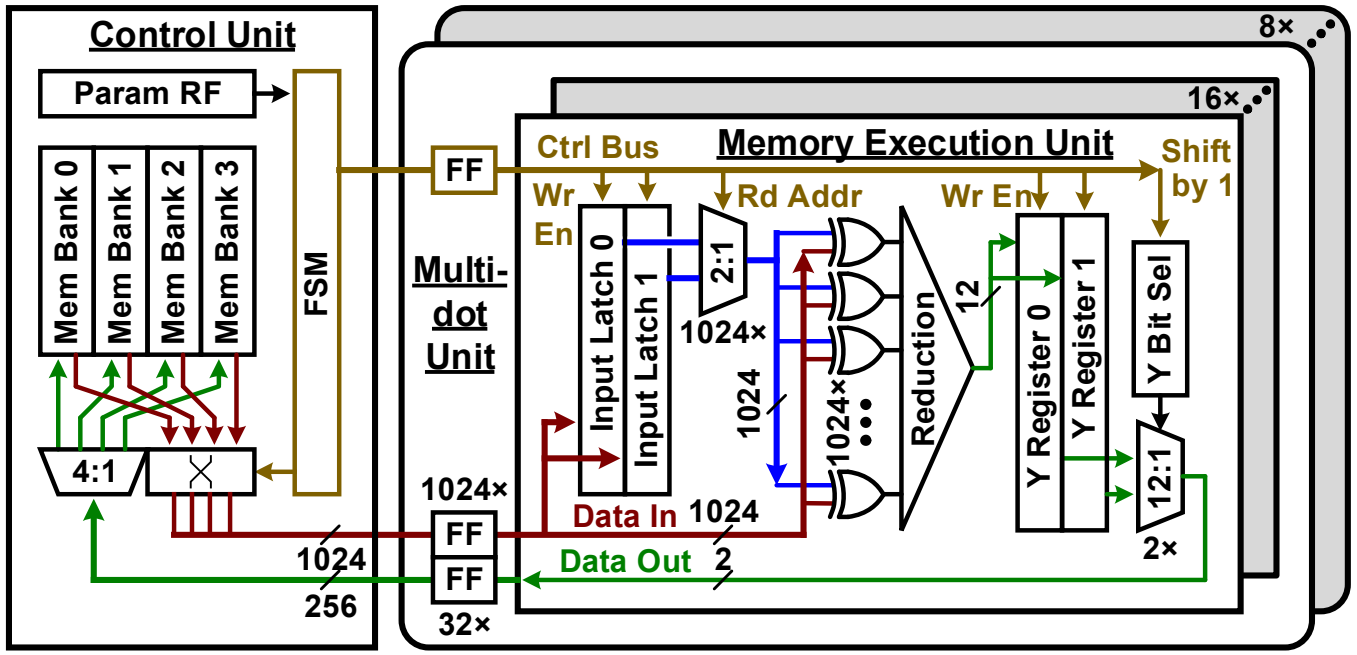




- Peak throughput of 170 8b TOPS @ 0.9V that scaled up with number of CNM cores
- NTV operation down to 450mV decreases energy by 3.1x to 2.9 8b TOPS/W
- Variable precision improves energy efficiency by 11.4x to 33.0 1b TOPS/W

E. Sumbul, R. Krishnamurthy et al, IEEE Solid-State Circuits Letters, 2020

10nm Binary Neural Network Accelerator Overview

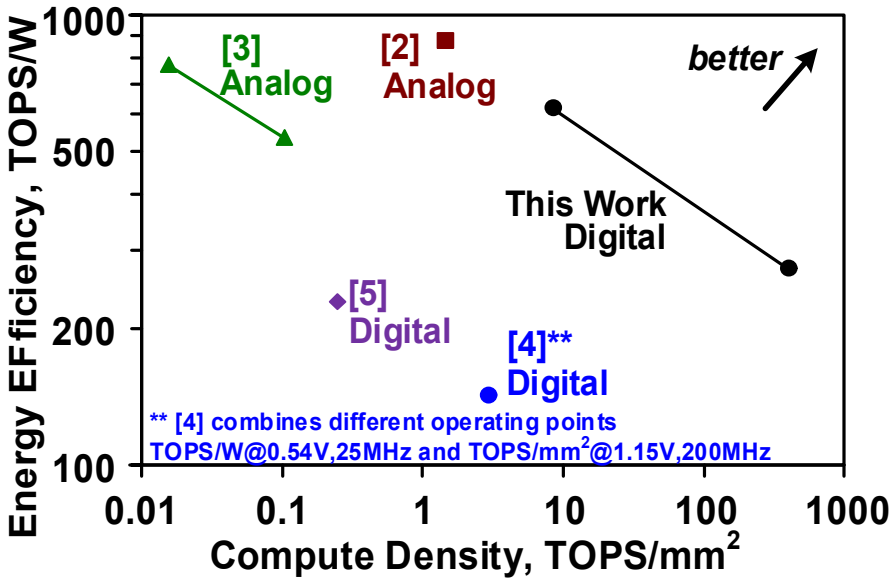


- Array of 128 Memory Execution Units (MEU) combine latch base memory and inner product compute in fine grain manner to minimize interconnect energy
- Central controller manages data flow from four 256b memory banks to MEUs
- 2 latch words per MEU enables data reuse reducing input bandwidth by 2x

P. Knag, R. Krishnamurthy et al, IEEE VLSI Circuits Symposium, 2020

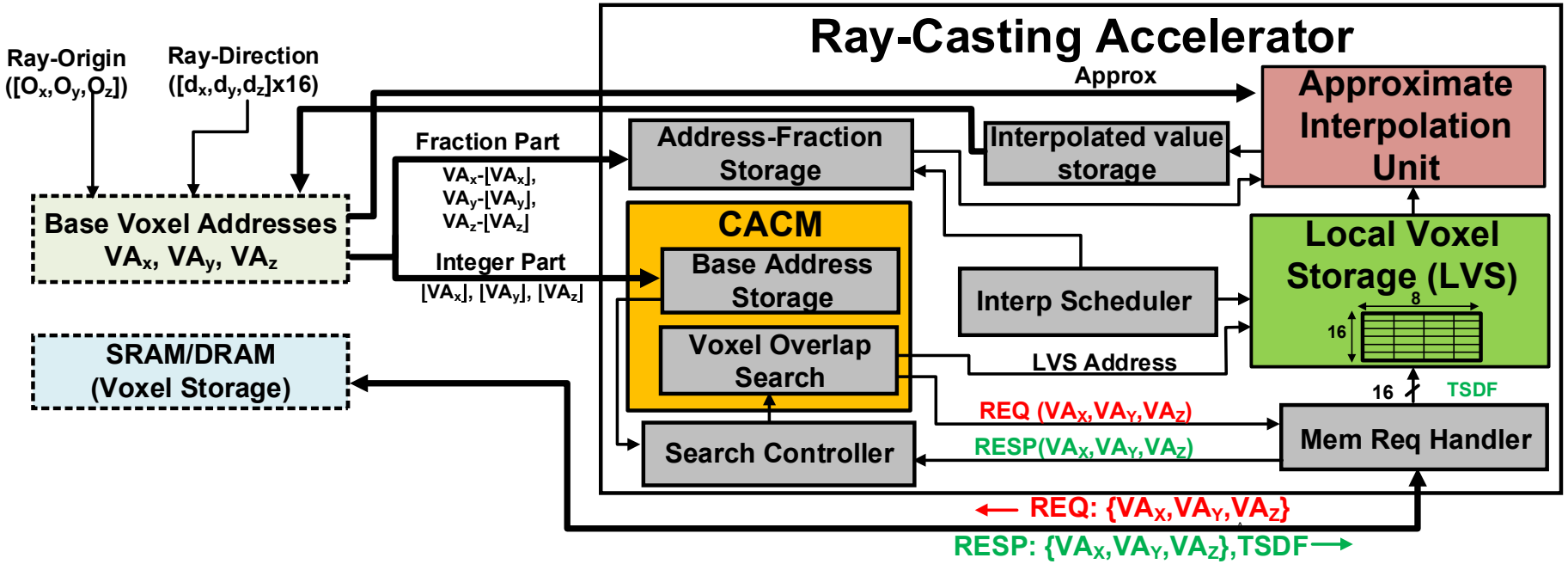
Comparison to Previously Published BNNs

	[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



- Industry-leading peak energy efficiency of 617TOPS/W
- 418 TOPS/mm^2 area efficiency

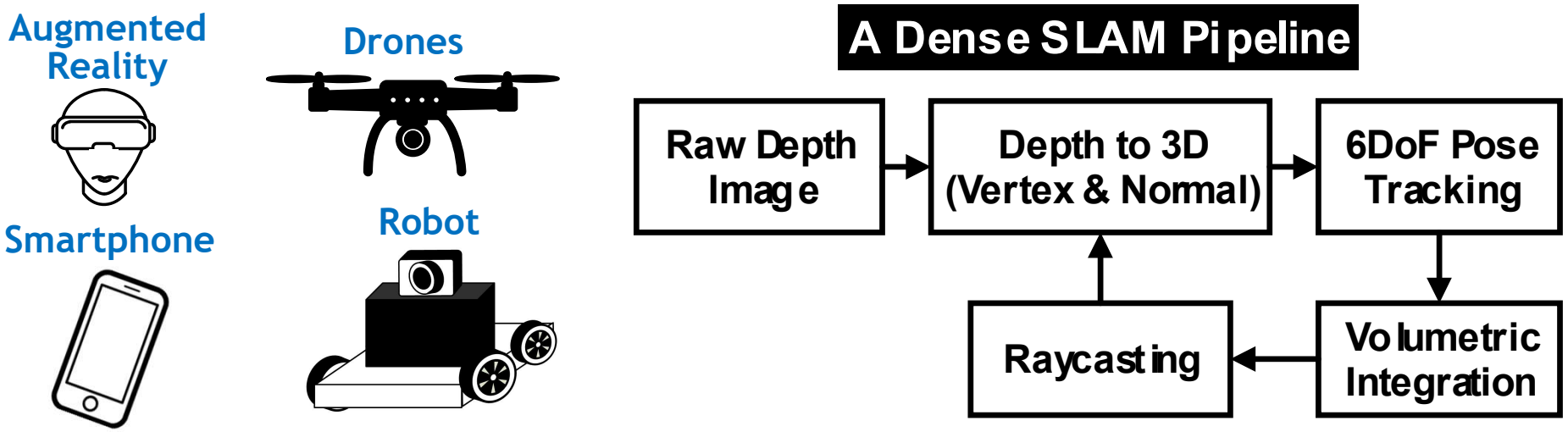
SLAM Ray-Casting Accelerator for Edge AI



- Content addressable compute memory (CACM) searches spatially overlapping voxels for rays of different pixels
- Interpolation unit with exact and approximate computation modes

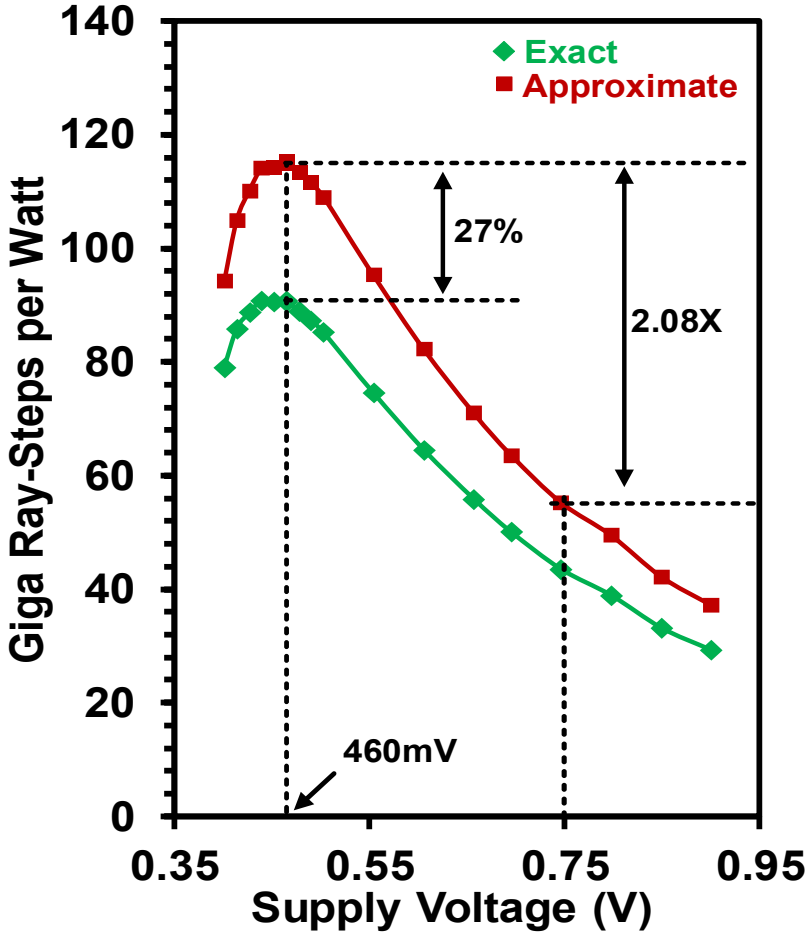
S. Hsu, R. Krishnamurthy et al, IEEE VLSI Circuits Symposium, 2020

Applications with 3D Scene Reconstruction



- Wide range of edge applications demand accurate volumetric reconstruction of surrounding scenes
- Dense Simultaneous Location and Mapping (SLAM) required for accurate surface estimation and 3D scene reconstruction

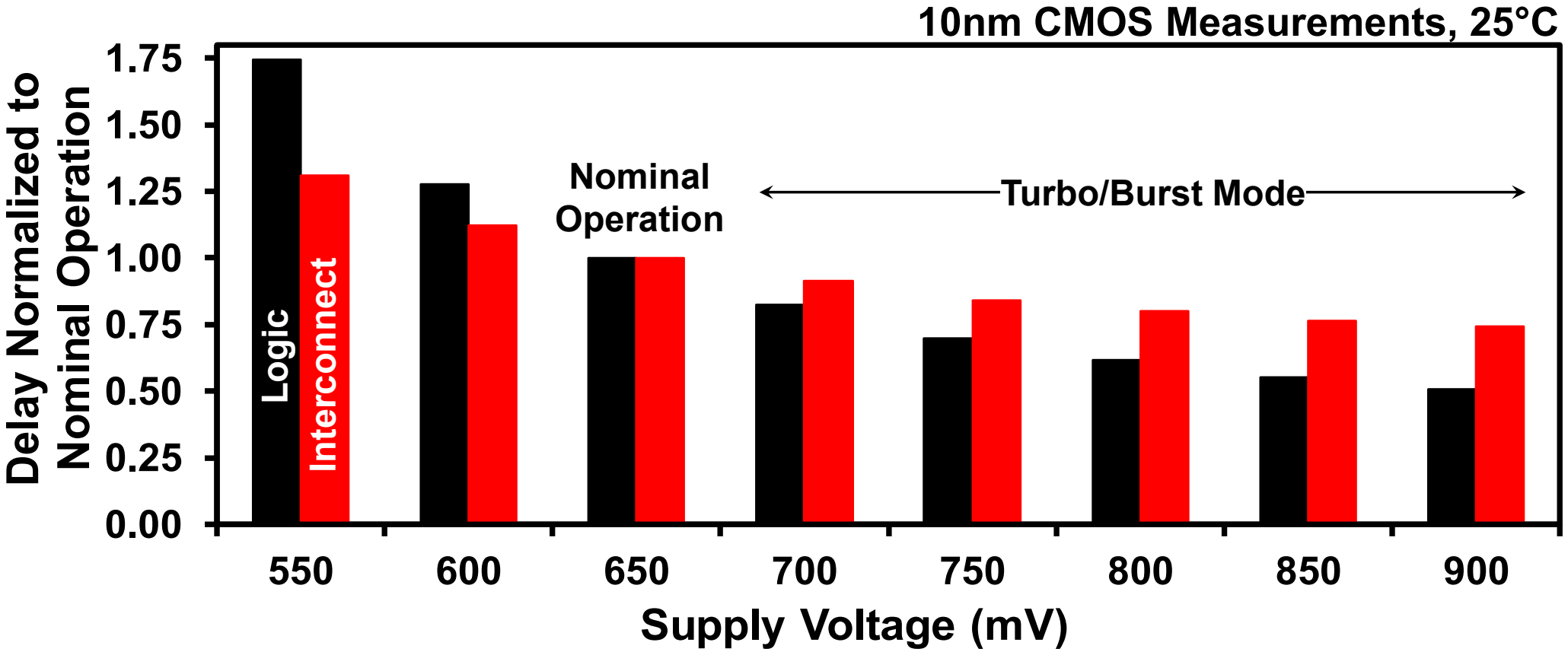
10nm Ray Casting Energy-Efficiency Measurements



- A peak 115.3 Giga Ray-Steps/Watt energy-efficiency at 460mV, 25°C
- 27% improvement in peak energy-efficiency with Approximate Computing mode at 460mV
- Low-voltage operation provides 2.08x improvement in energy-efficiency

S. Hsu, R. Krishnamurthy et al, IEEE VLSI Circuits Symposium, 2020

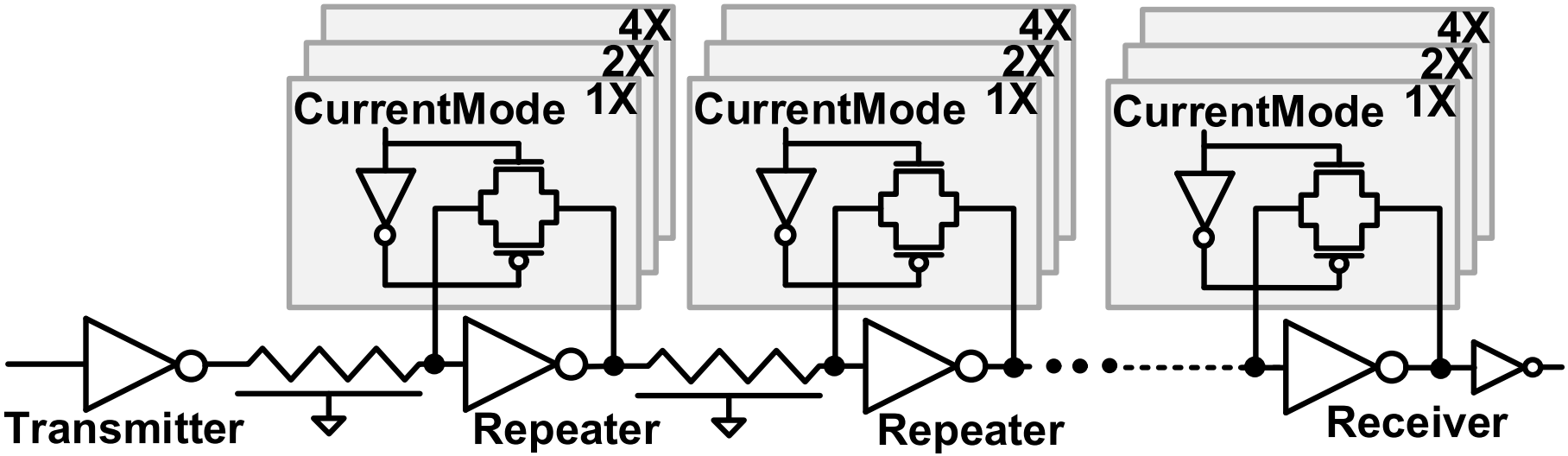
On-Die Interconnects Data Movement Challenges



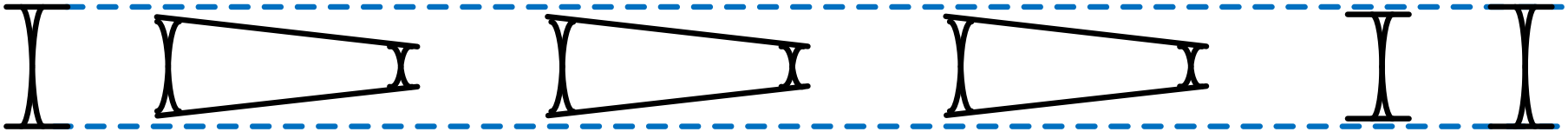
- Clock frequencies limited by interconnect RC delay at high voltages
- Reduces power/performance tunability of high-performance processors

M. Anders, R. Krishnamurthy et al, ISSCC 2020

Reconfigurable Voltage/Current-Mode Interconnects

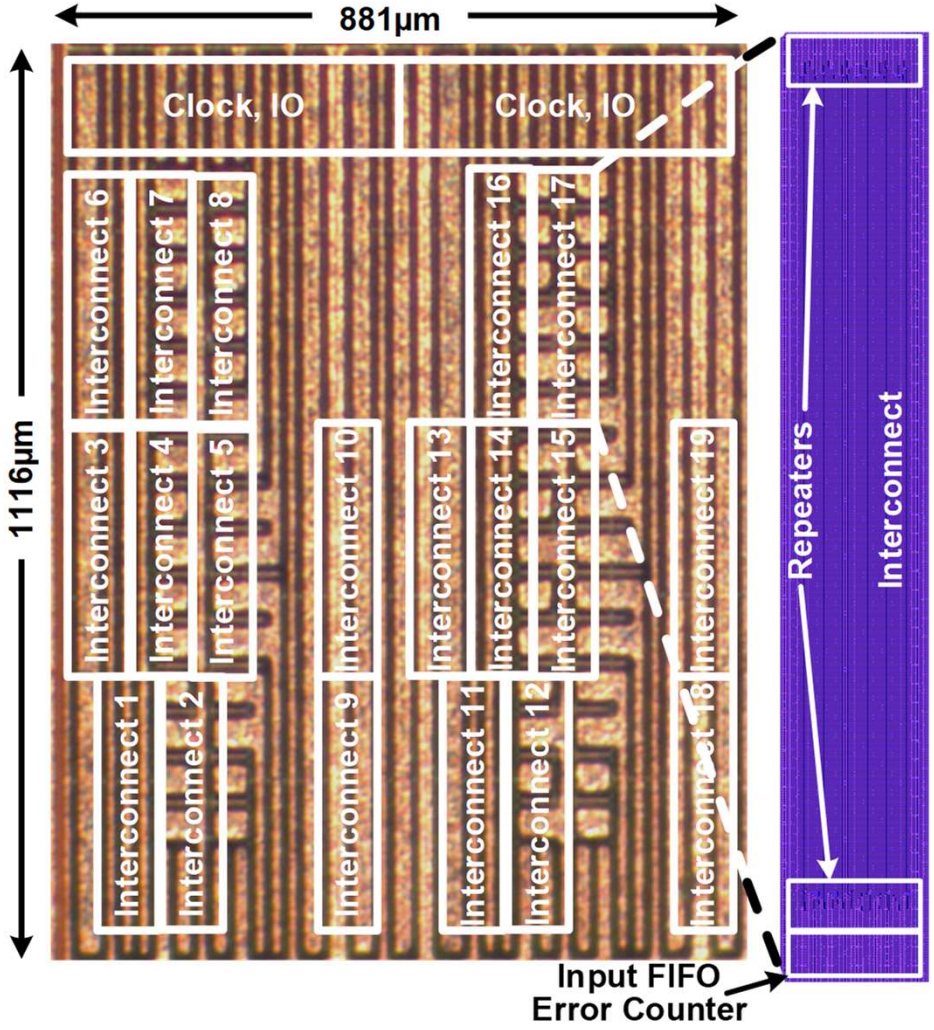


Voltage Swing in Current Mode



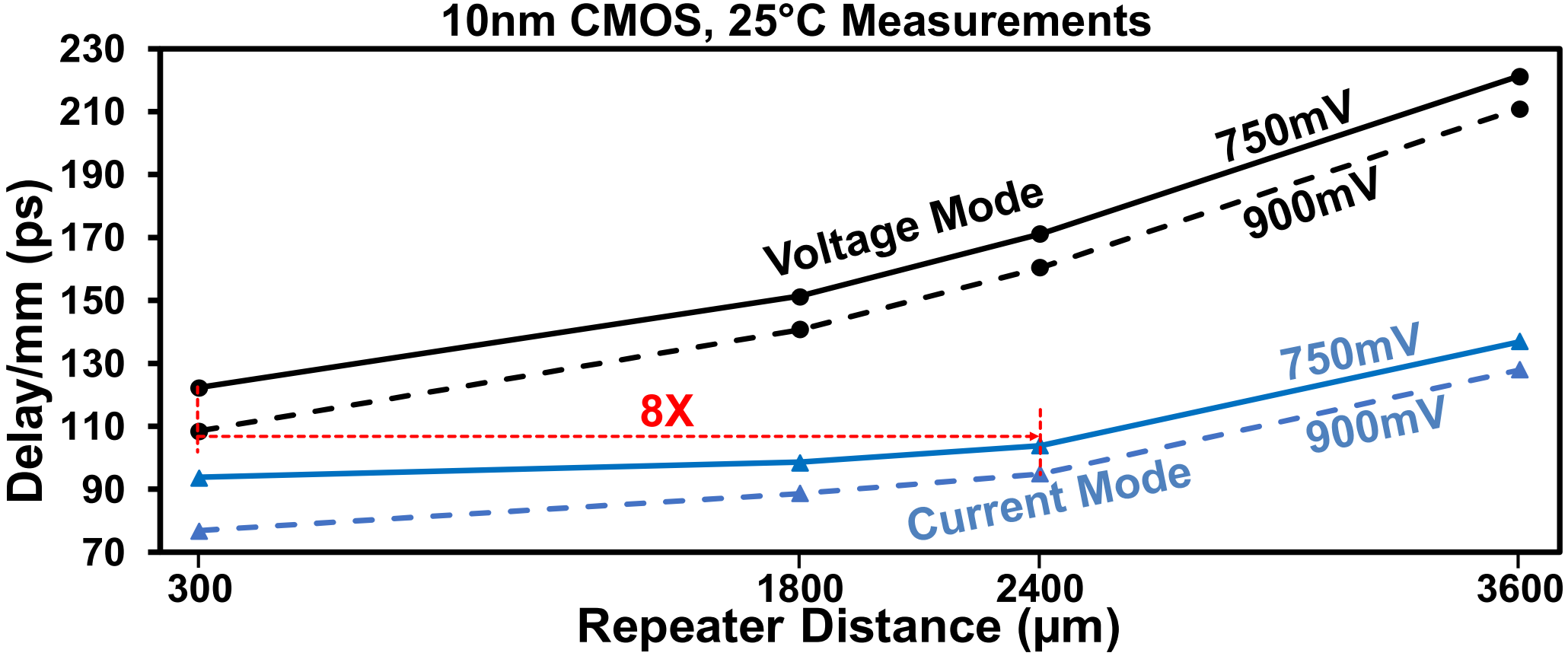
- Programmable transmission gates provide feedback resistance
  - Less than 10% additional transistor width and 5% voltage-mode delay overhead
- Voltage swing centered around inverter switching threshold
  - Swing reduces along interconnect and is amplified at each repeater

10nm Current Mode Interconnects Test Chip



Experiments	Conventional Voltage-mode Reconfigurable Current-mode Transient Current-mode
Interconnect	Metal 13, 1X - 3X minimum pitch
Repeater Distance	0.3mm – 3.6mm
Measurements	Delay, Power, Noise, Voltage Swing
Configuration Parameters	Supply Voltage Feedback Resistance Pulse Width/Delay Data Pattern

Current Mode vs. Voltage Mode Interconnects



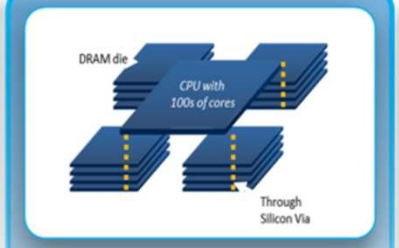
- Improved current-mode delay translated to longer inter-repeater distance
- 8X longer distance at same delay target as voltage-mode

M. Anders, R. Krishnamurthy et al, ISSCC 2020

“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