Learning Systems 2018:
Lecture **14** – Silicon Trends for Deep Learning

Crescat scientia; vita excolatur

Ian Foster and Rick Stevens
Argonne National Laboratory
The University of Chicago

Topics for Today

- Where is the semiconductor industry going in the next ten years?
- What does this mean for increasing performance for Deep Learning?
- What Architectural directions are likely?
- Survey of some new ventures aiming at the deep learning market

- How would you assess the trends?
- How would you measure performance on real applications?
- Are Deep Learning algorithms stable enough for hardware implementations?

ILP tapped out + end of frequency scaling

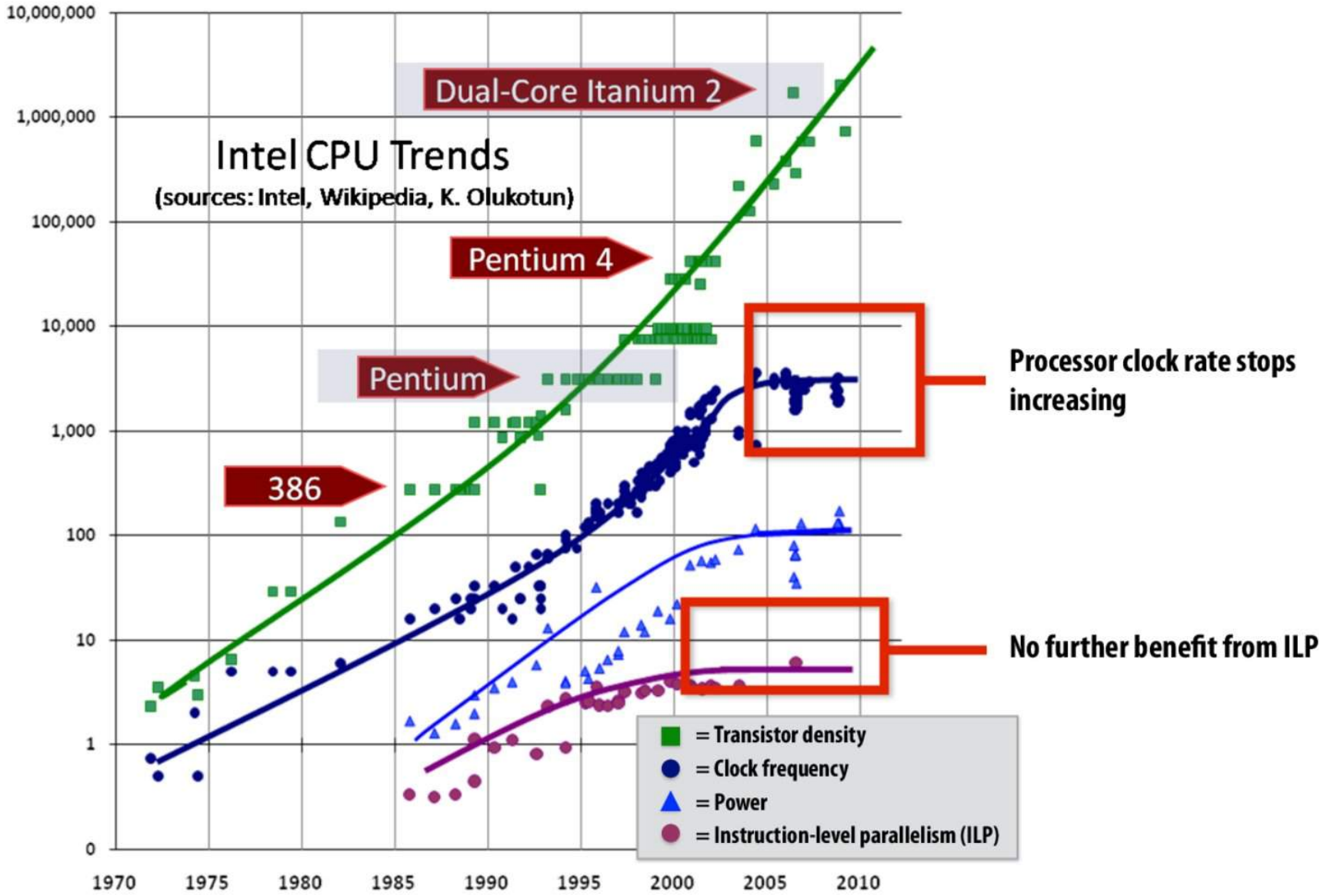
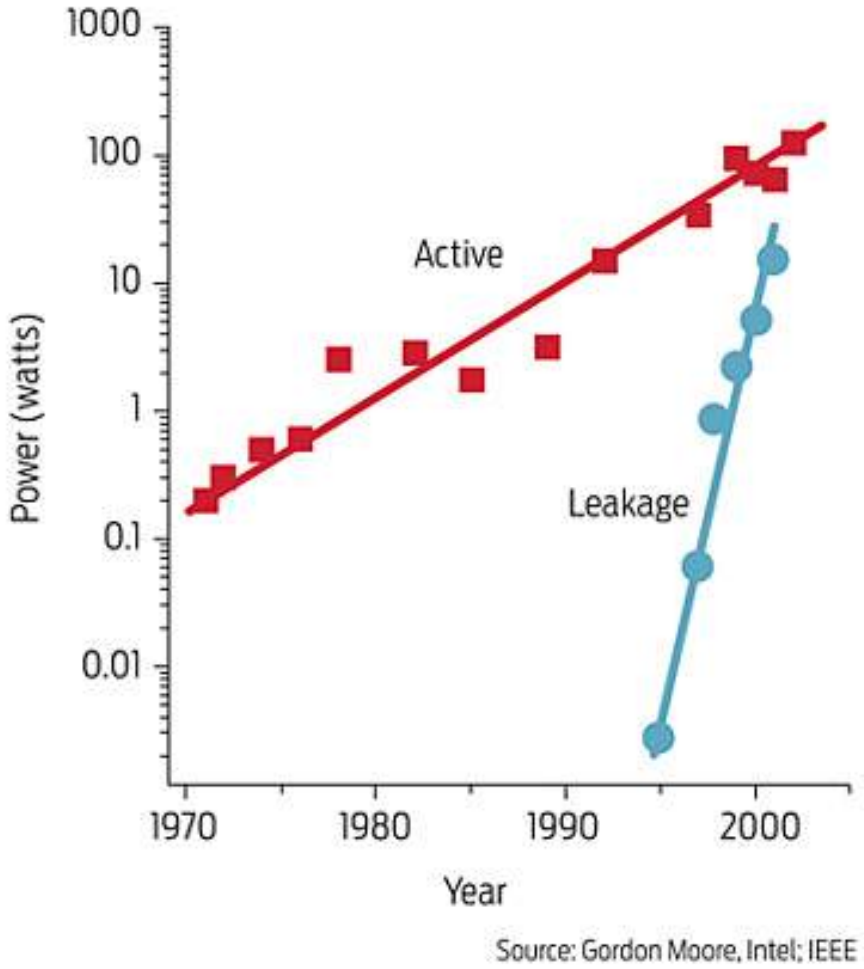
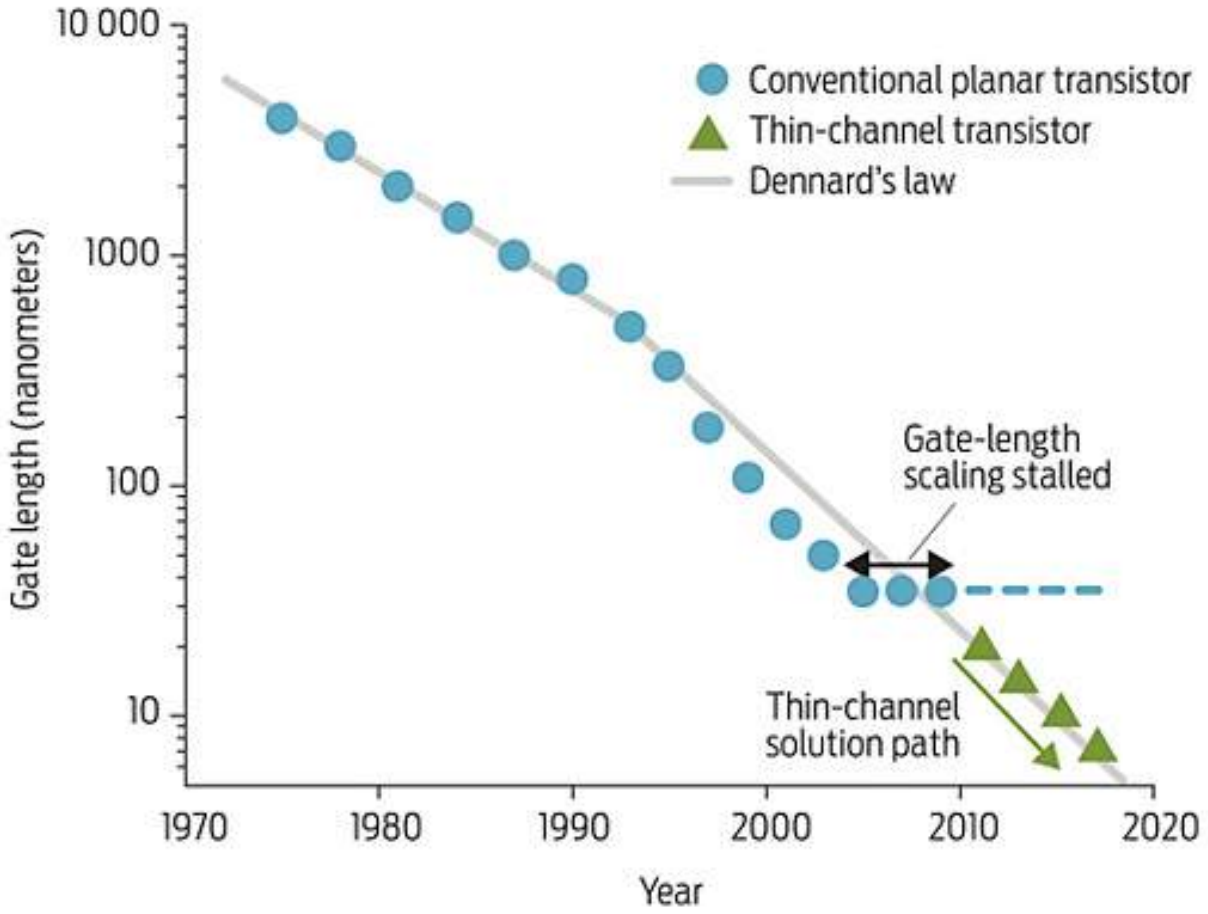


Image credit: "The free Lunch is Over" by Herb Sutter, Dr. Dobbs 2005



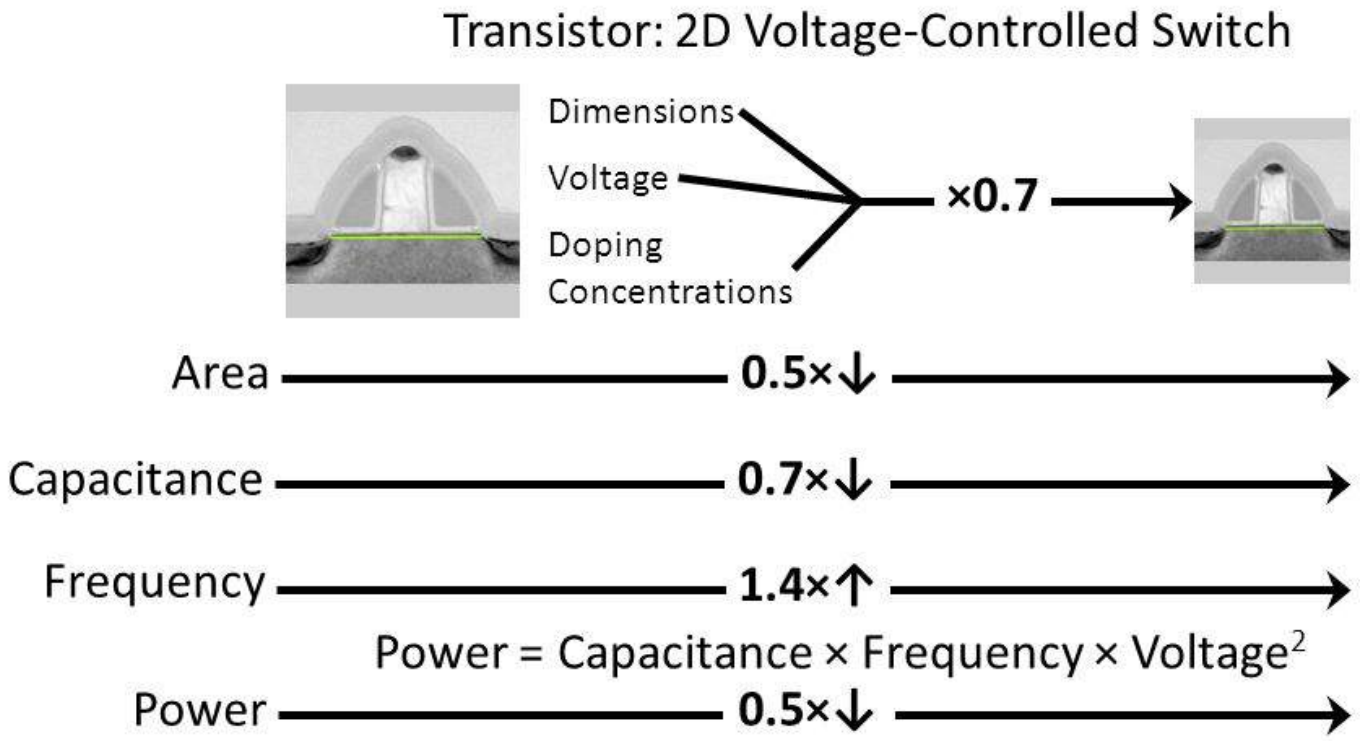
Source: Gordon Moore, Intel; IEEE



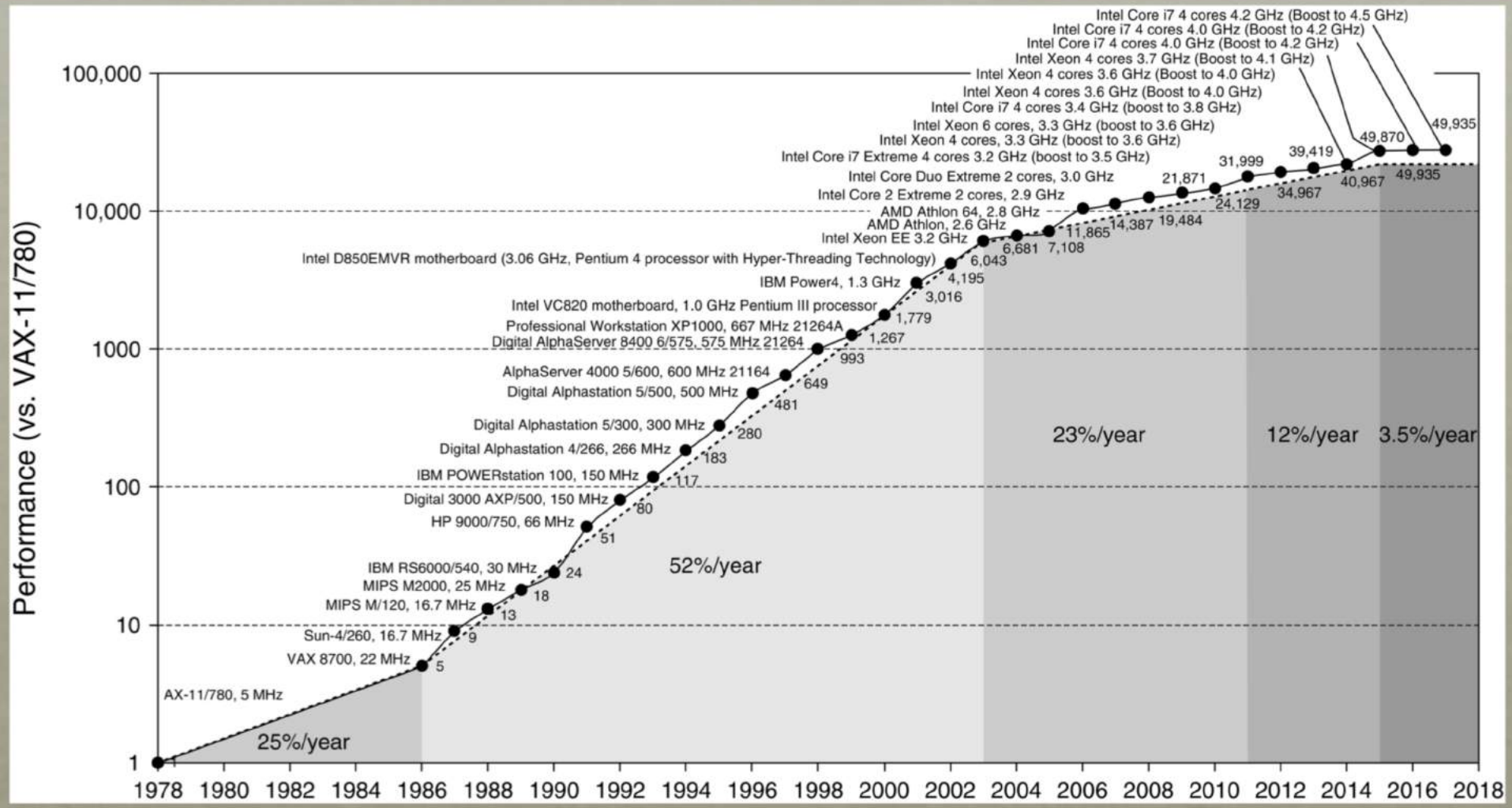
Source: Intel; Khaled Ahmed, Applied Materials

Dennard scaling:

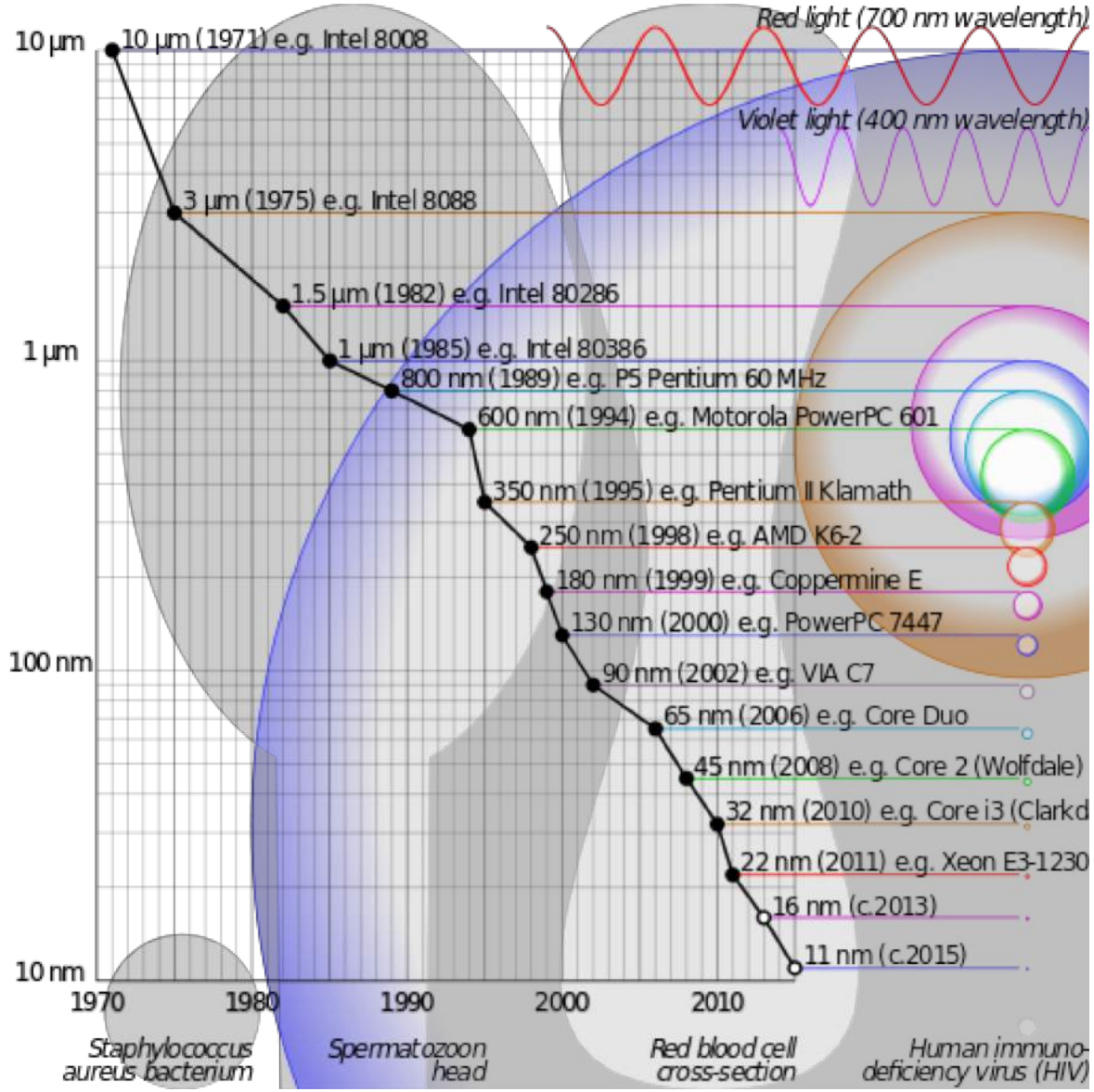
Doubling the transistors; scale their power down



Uniprocessor Performance (single core)



Performance = highest SPECInt by year; from Hennessy & Patterson [2018].


10 µm
1 µm
100 nm
10 nm
1970
1980
1990
2000
2010
10 µm (1971) e.g. Intel 8008
3 µm (1975) e.g. Intel 8088
1.5 µm (1982) e.g. Intel 80286
1 µm (1985) e.g. Intel 80386
800 nm (1989) e.g. P5 Pentium 60 MHz
600 nm (1994) e.g. Motorola PowerPC 601
350 nm (1995) e.g. Pentium II Klamath
250 nm (1998) e.g. AMD K6-2
180 nm (1999) e.g. Coppermine E
130 nm (2000) e.g. PowerPC 7447
90 nm (2002) e.g. VIA C7
65 nm (2006) e.g. Core Duo
45 nm (2008) e.g. Core 2 (Wolfdale)
32 nm (2010) e.g. Core i3 (Clarkd
22 nm (2011) e.g. Xeon E3-1230
16 nm (c.2013)
11 nm (c.2015)
Red light (700 nm wavelength)
Violet light (400 nm wavelength)
Staphylococcus aureus bacterium
Spermatozoon head
Red blood cell cross-section
Human immuno-deficiency virus (HIV)

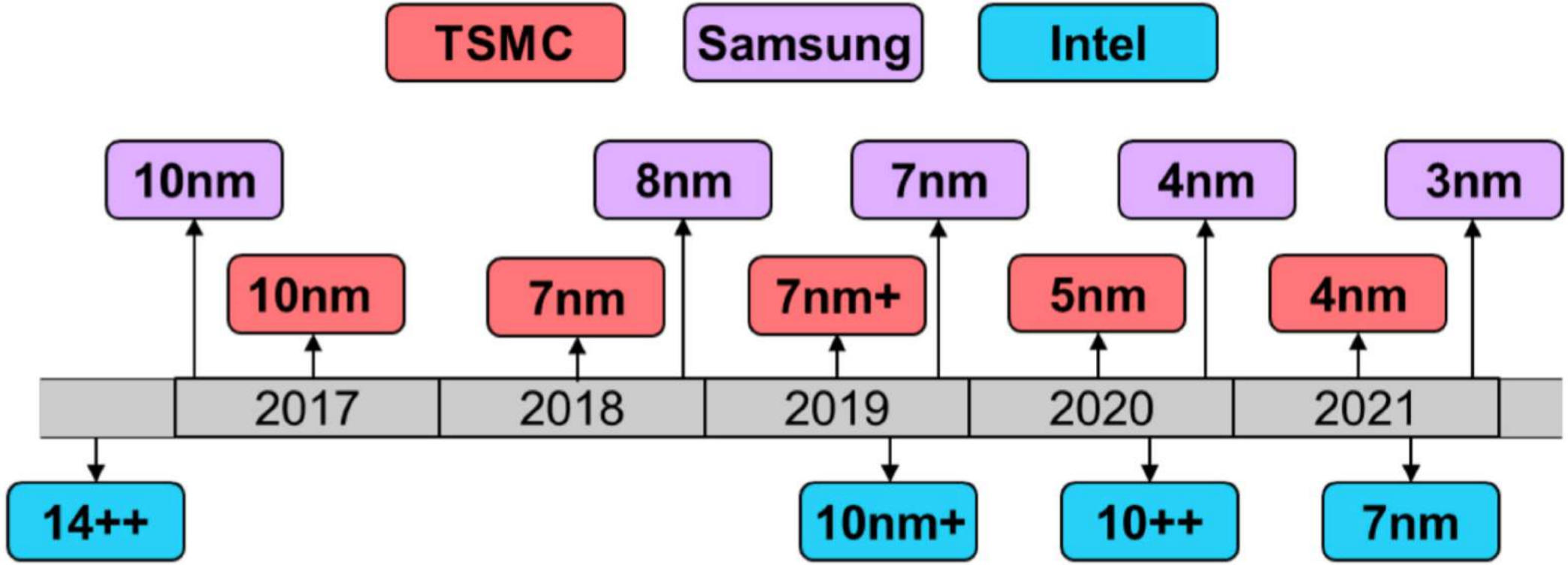


Figure 1. IC-process roadmap to 3nm. TSMC expects to be first with EUV by offering its 7nm+ process in mid-2019, but Samsung plans to lead the way to 3nm. All dates are for high-volume production. (Source: vendors, except future nodes are The Linley Group estimates)

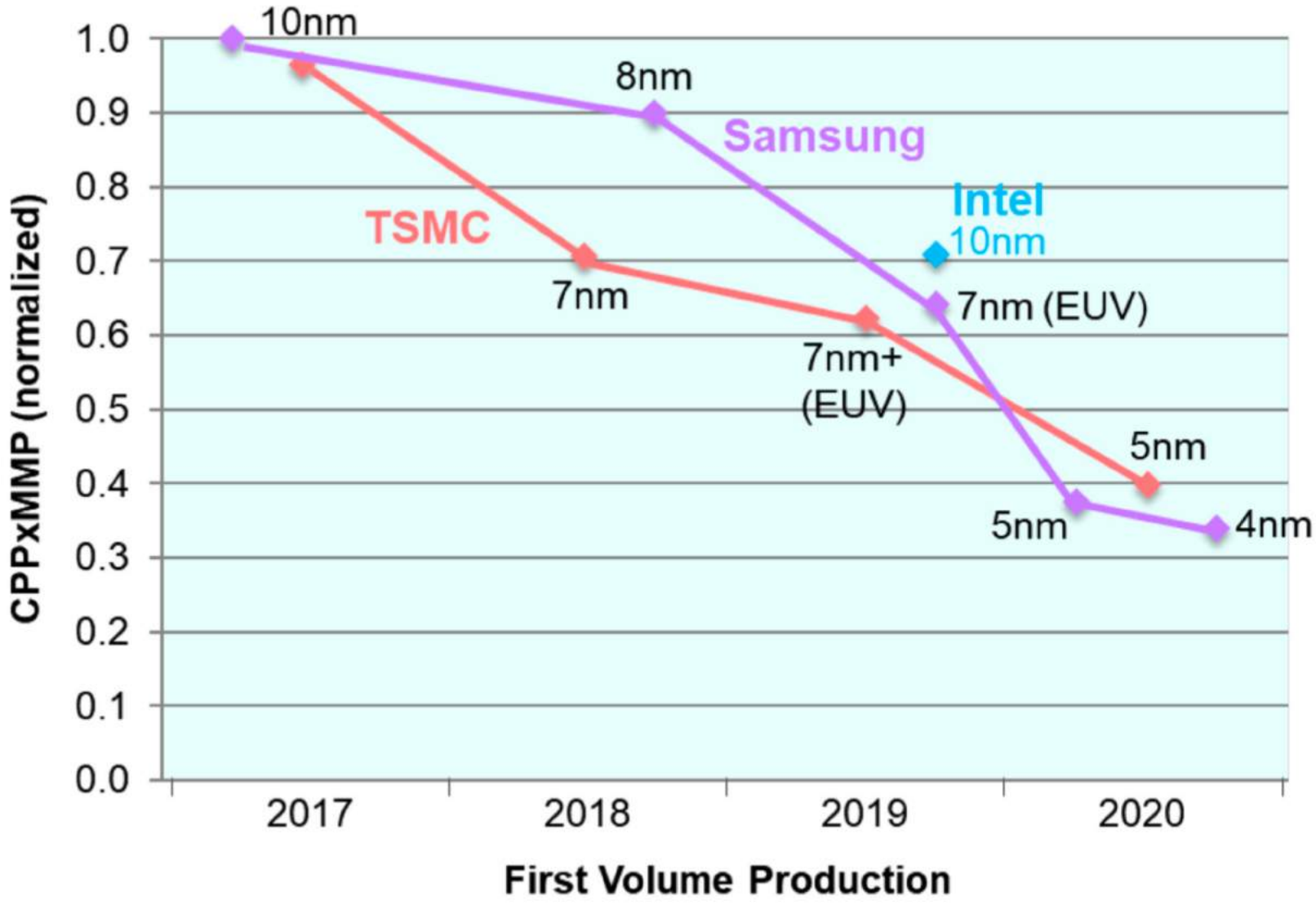


Figure 3. IC-process roadmap to 3nm. TSMC expects to be first with EUV by offering its 7nm+ process in mid-2019, but Samsung plans to lead the way to 3nm. All dates are for high-volume production. (Source: vendors, except future nodes are The Linley Group estimates)

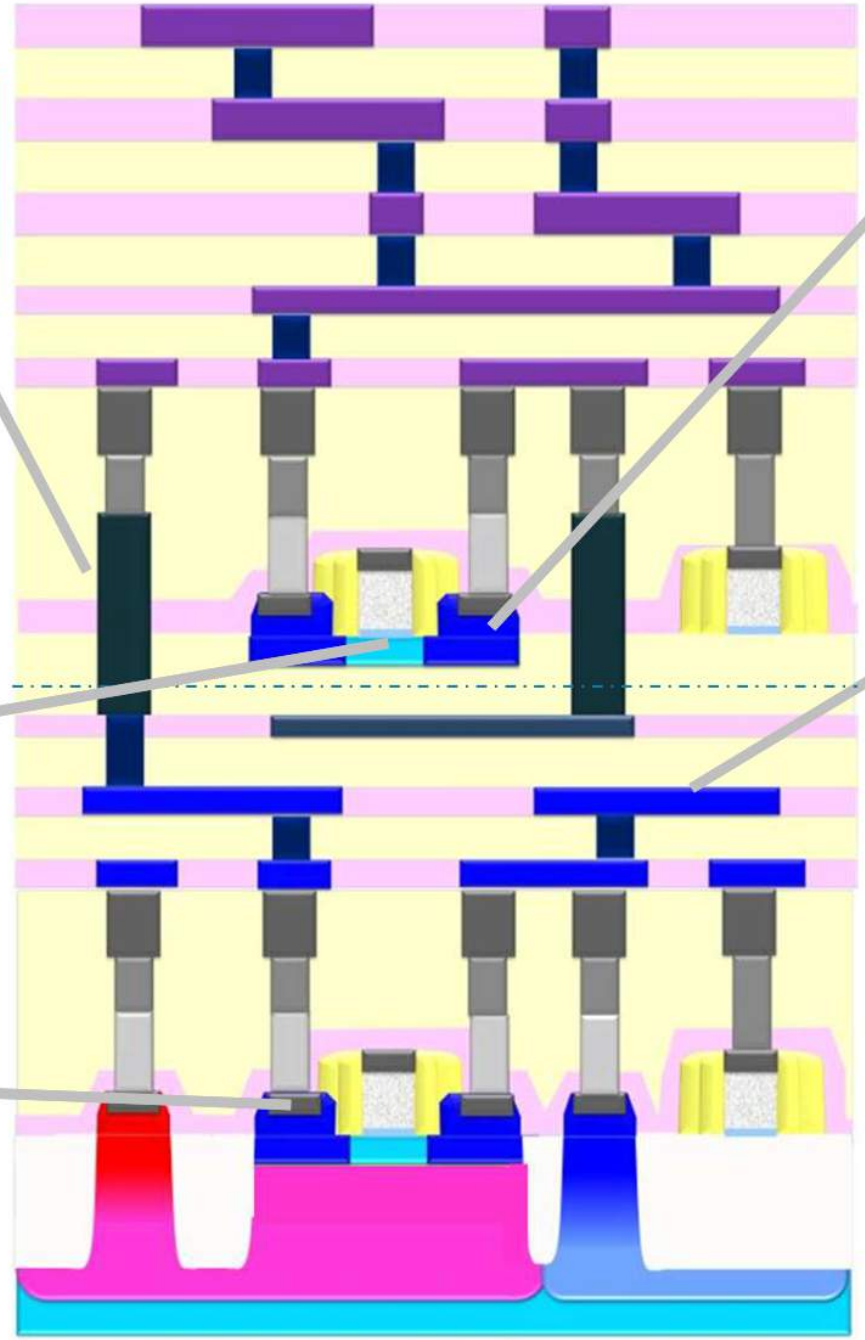
Low-resistivity 3D connections

Low thermal budget top layer

Local interconnect Level

High-quality top film

Bottom MOS FET thermal stability



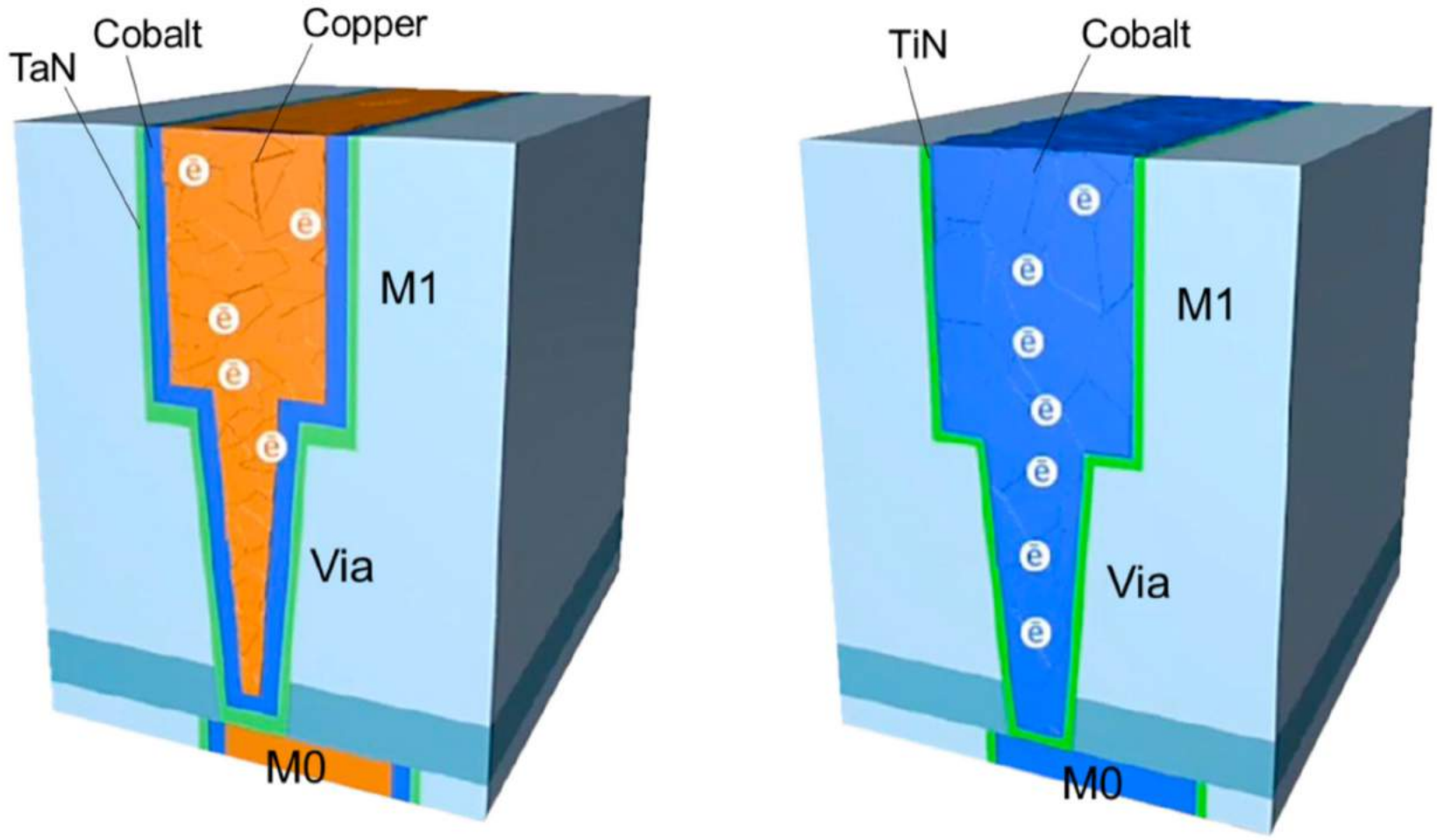


Figure 2. Copper versus cobalt interconnects. Copper requires a thicker barrier layer (TaN or TiN) than cobalt. As the trench size decreases, this barrier constrains the conductive area and reduces electron flow. (Source: Applied Materials)

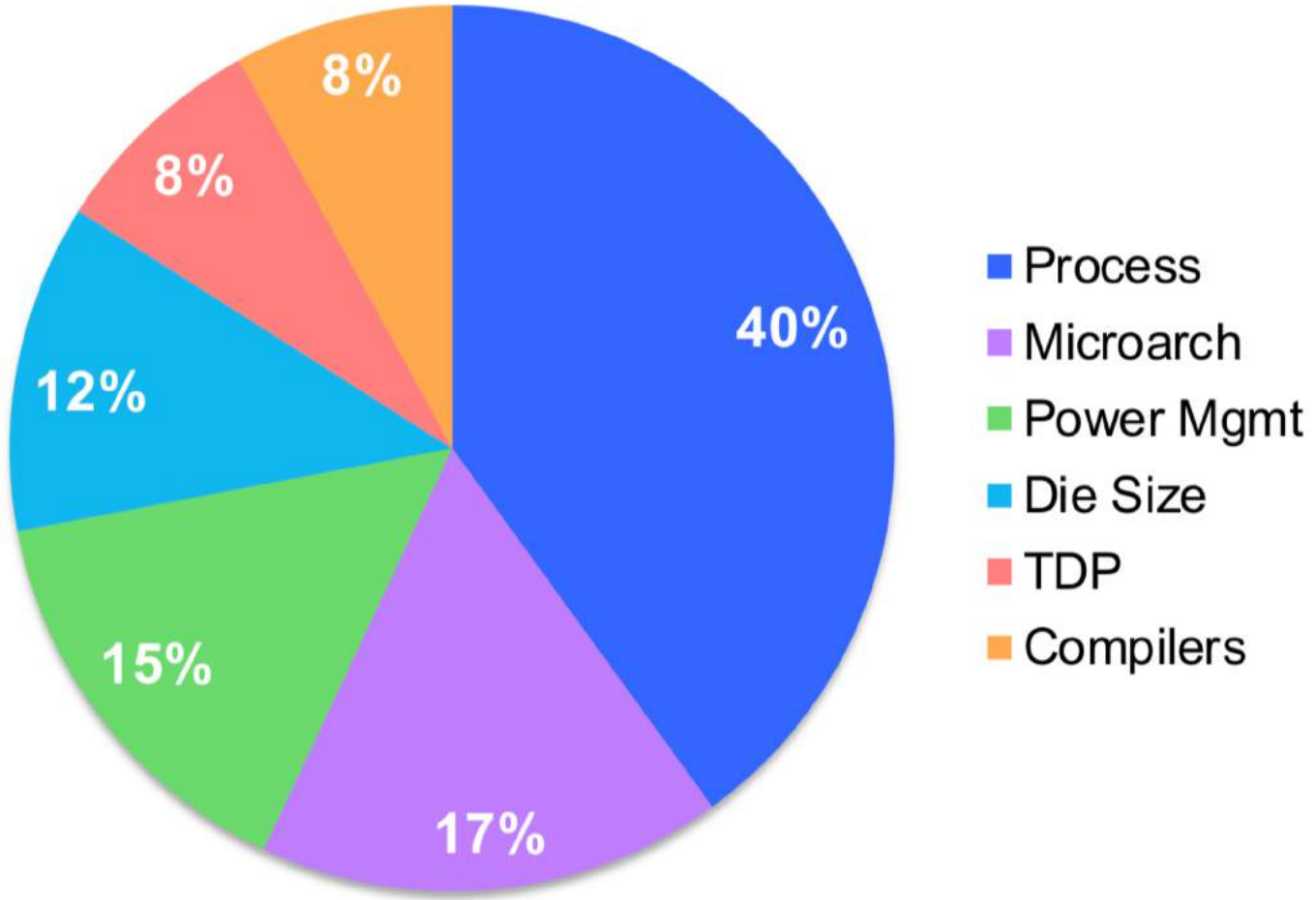


Figure 3. Processor-performance improvement factors. Although Moore's Law (IC process) has been the largest driver of increasing processor performance over the past decade, several other factors together contribute the majority of the gains.

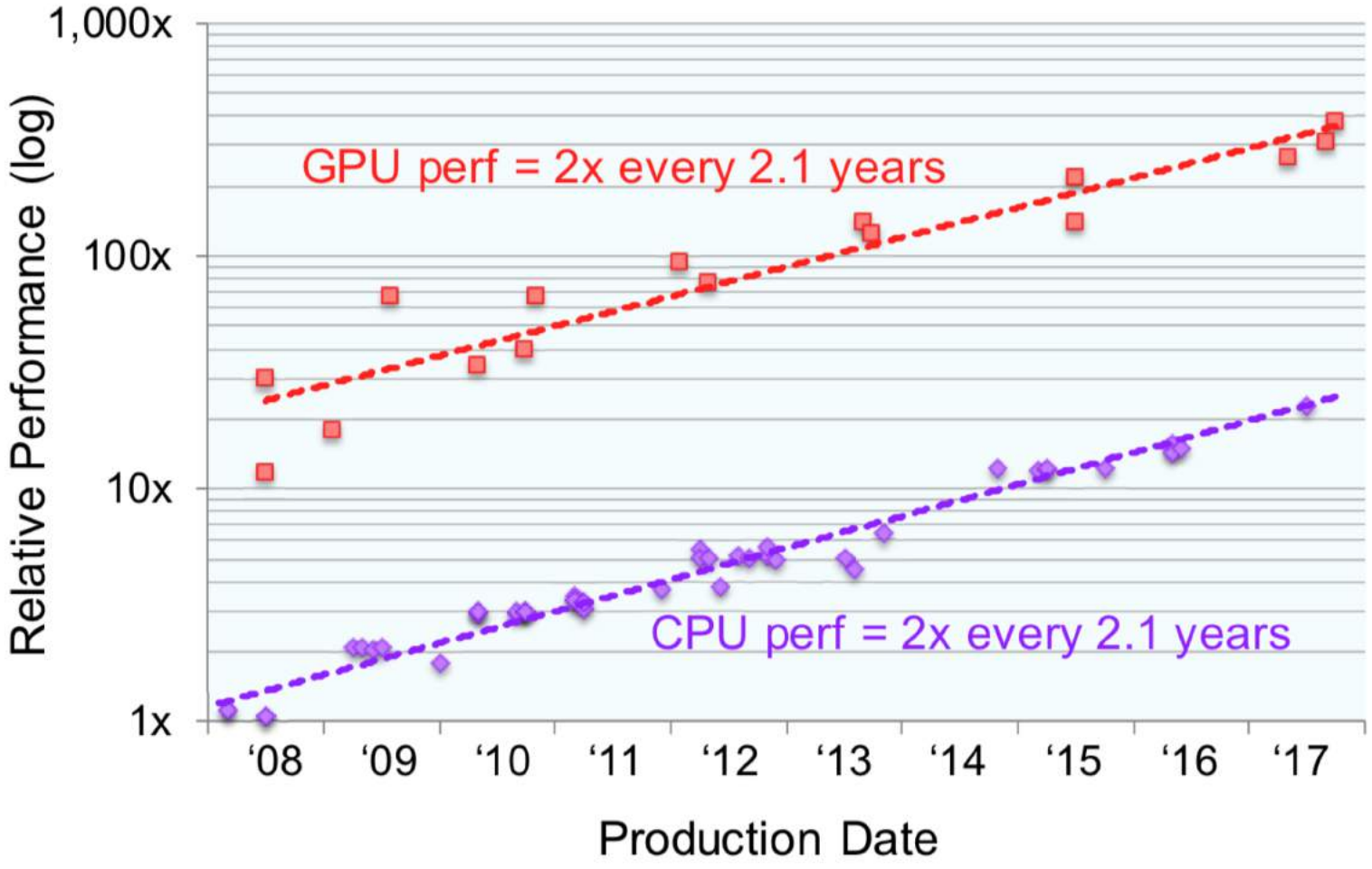
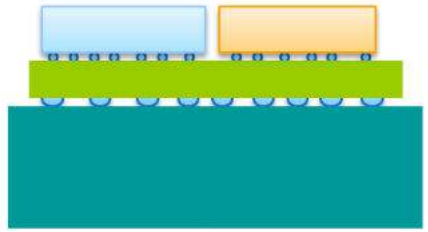
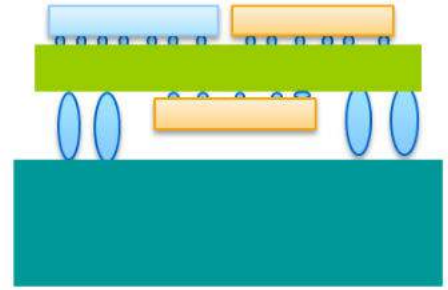


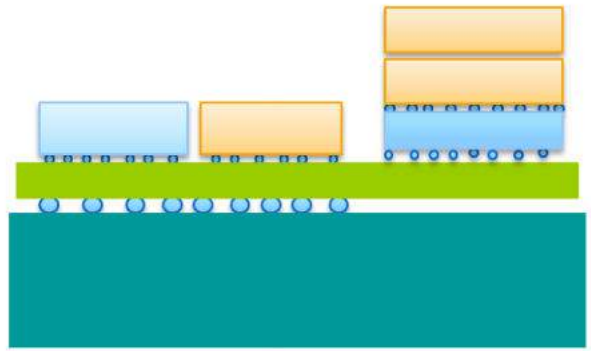
Figure 1. Processor performance trends. The performance of mainstream server CPUs (measured in SPECint_rate2006) and of gaming GPUs (measured in peak Gflop/s) have increased at similar rates over the past decade. (Data source: SPEC.org, vendors)



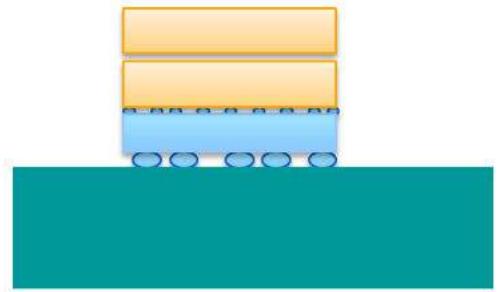
2.5D: Side-by-side die stacked on a passive interposer that includes TSVs



2.5 or 3D: Interposer with top and bottom connection



3D + Interposer: Mix of side-by-side and stacked implementations on an interposer



3D Memory on Logic: One or more DRAM die stacked directly on logic die

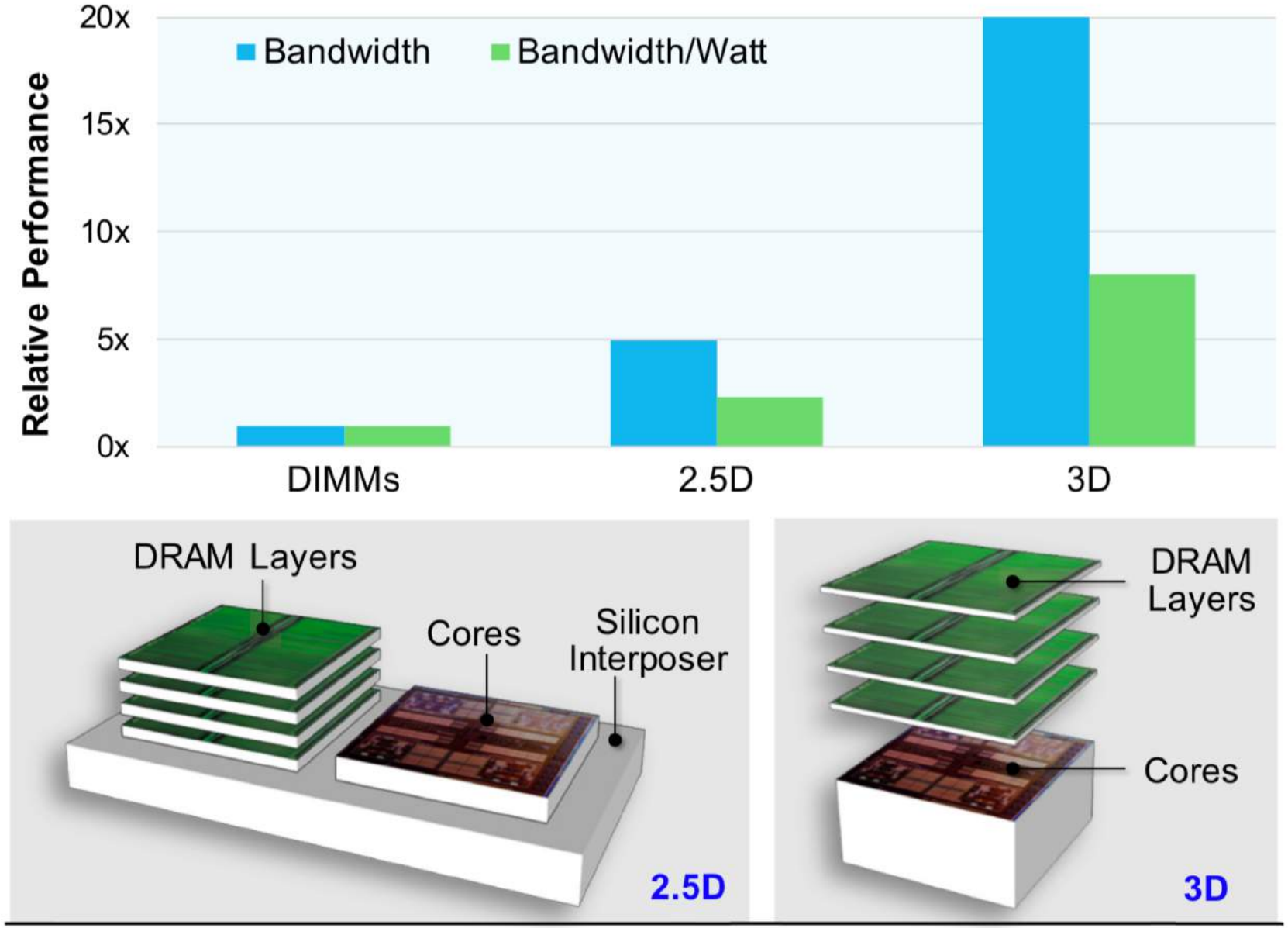
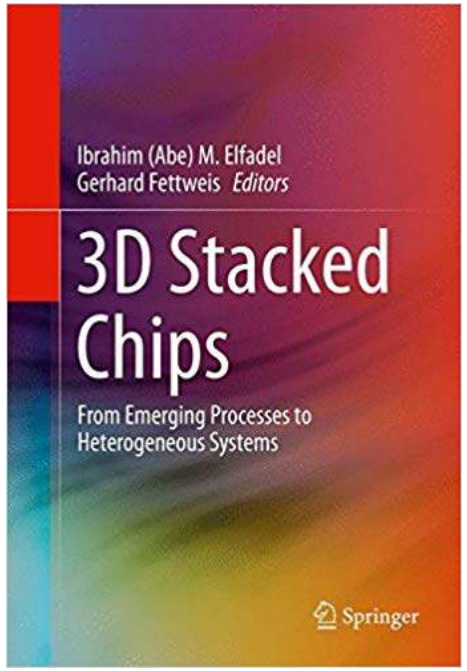
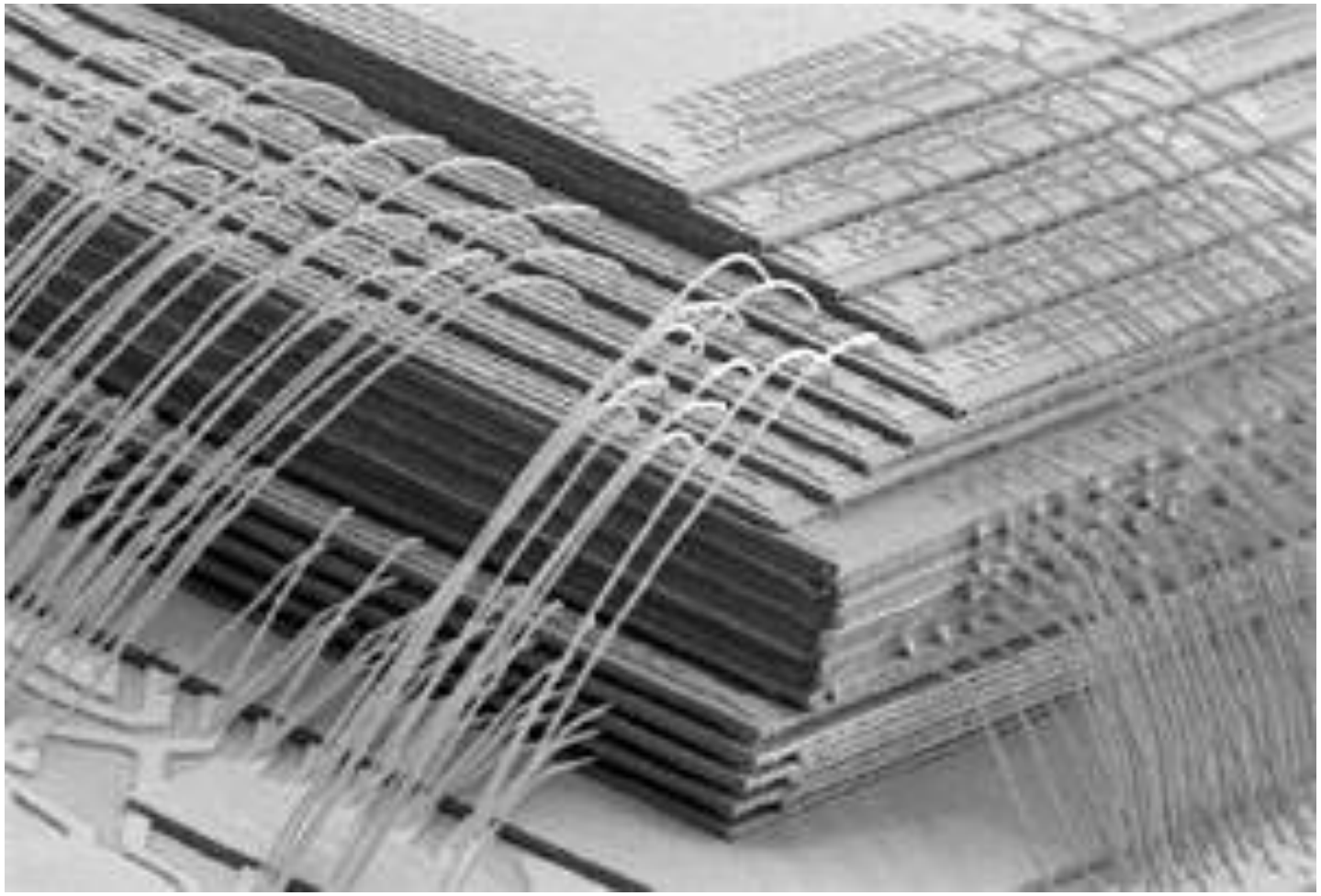
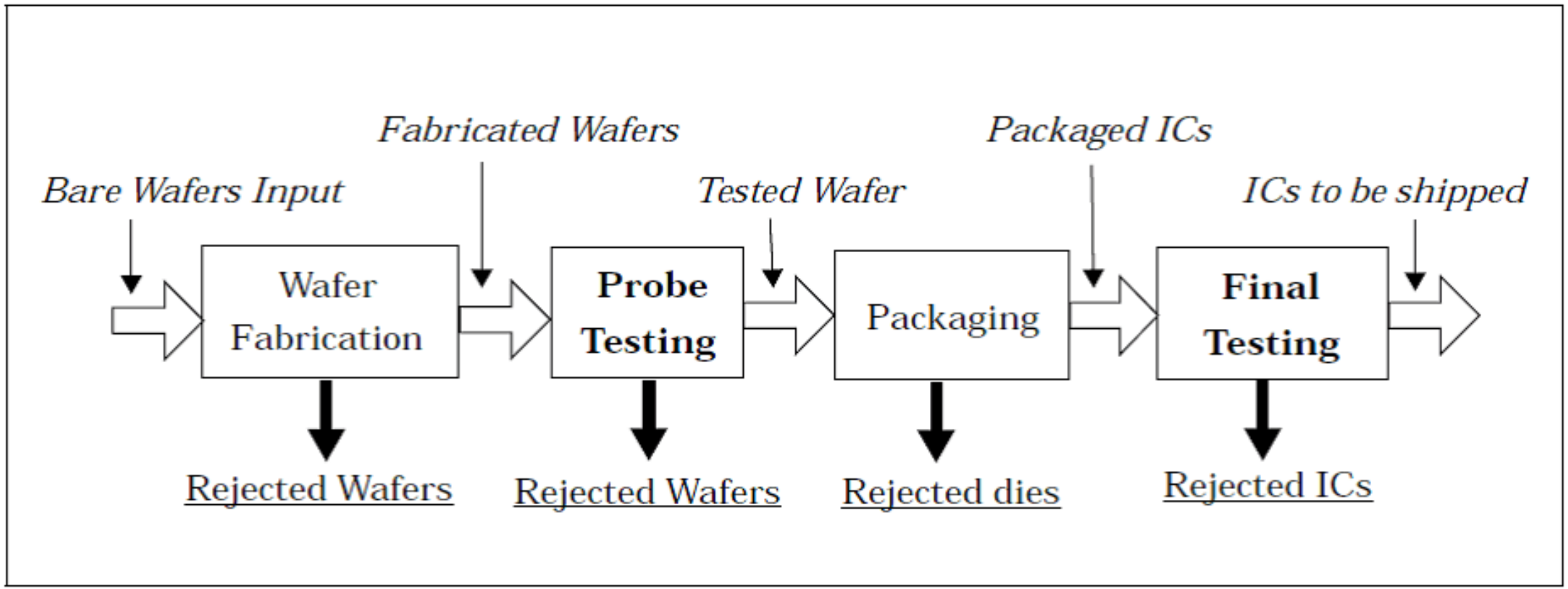
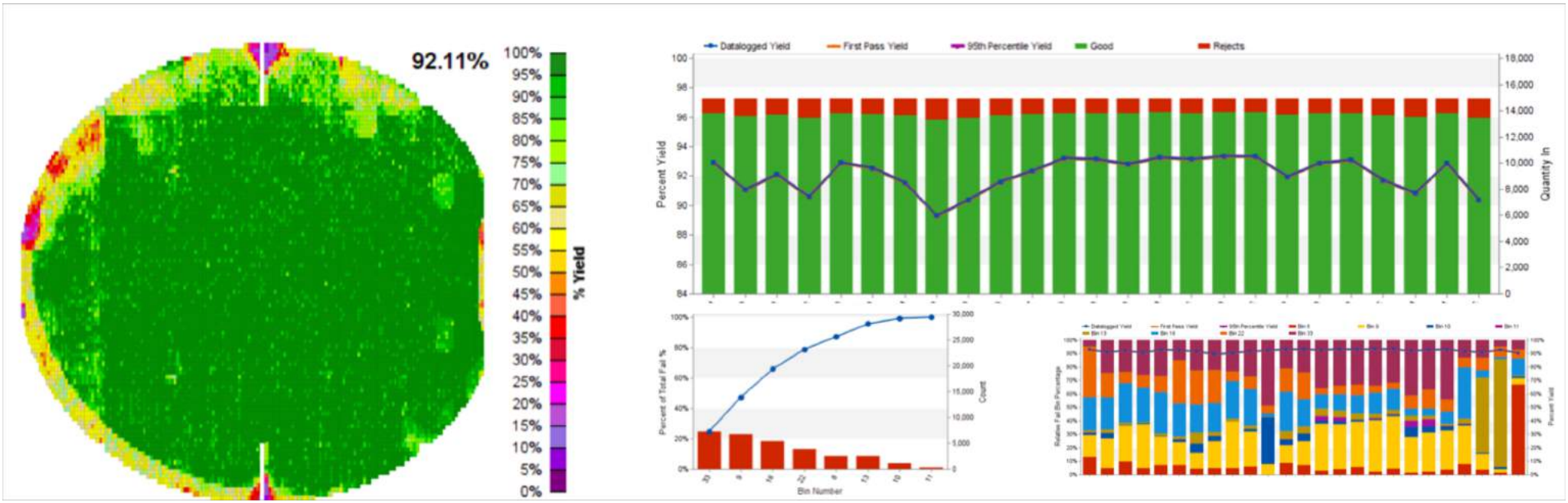


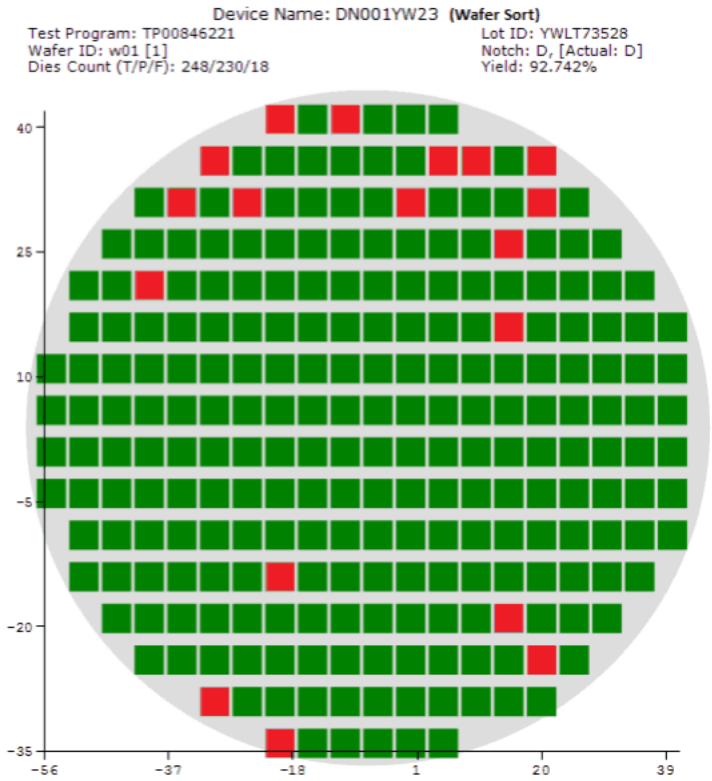
Figure 5. Improving memory packaging. Moving DRAM next to or even on top of the processor die can increase bandwidth while greatly reducing the energy required to access memory.

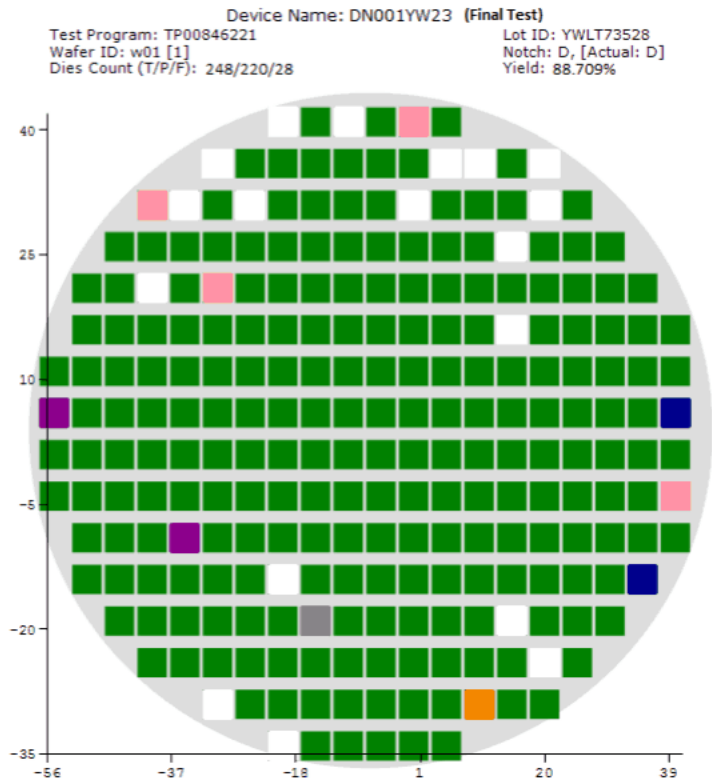

Ibrahim (Abe) M. Elfadel
Gerhard Fettweis Editors
3D Stacked Chips
From Emerging Processes to Heterogeneous Systems
Springer

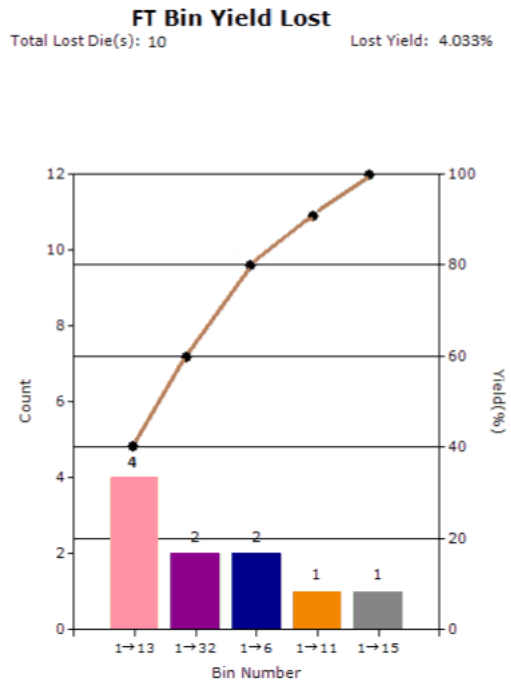


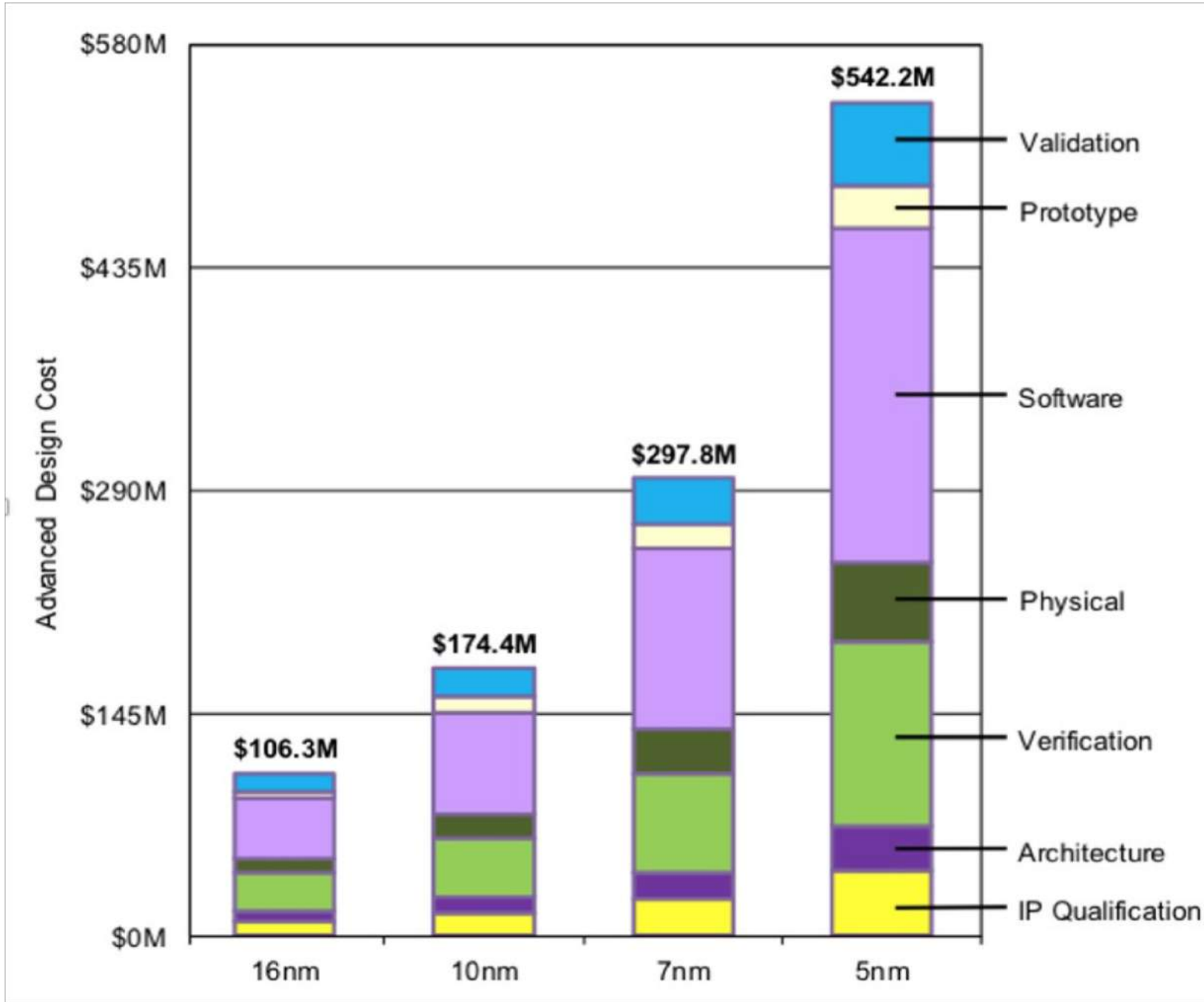

Bare Wafers Input
Fabricated Wafers
Tested Wafer
Packaged ICs
ICs to be shipped
Wafer Fabrication
Probe Testing
Packaging
Final Testing
Rejected Wafers
Rejected Wafers
Rejected dies
Rejected ICs

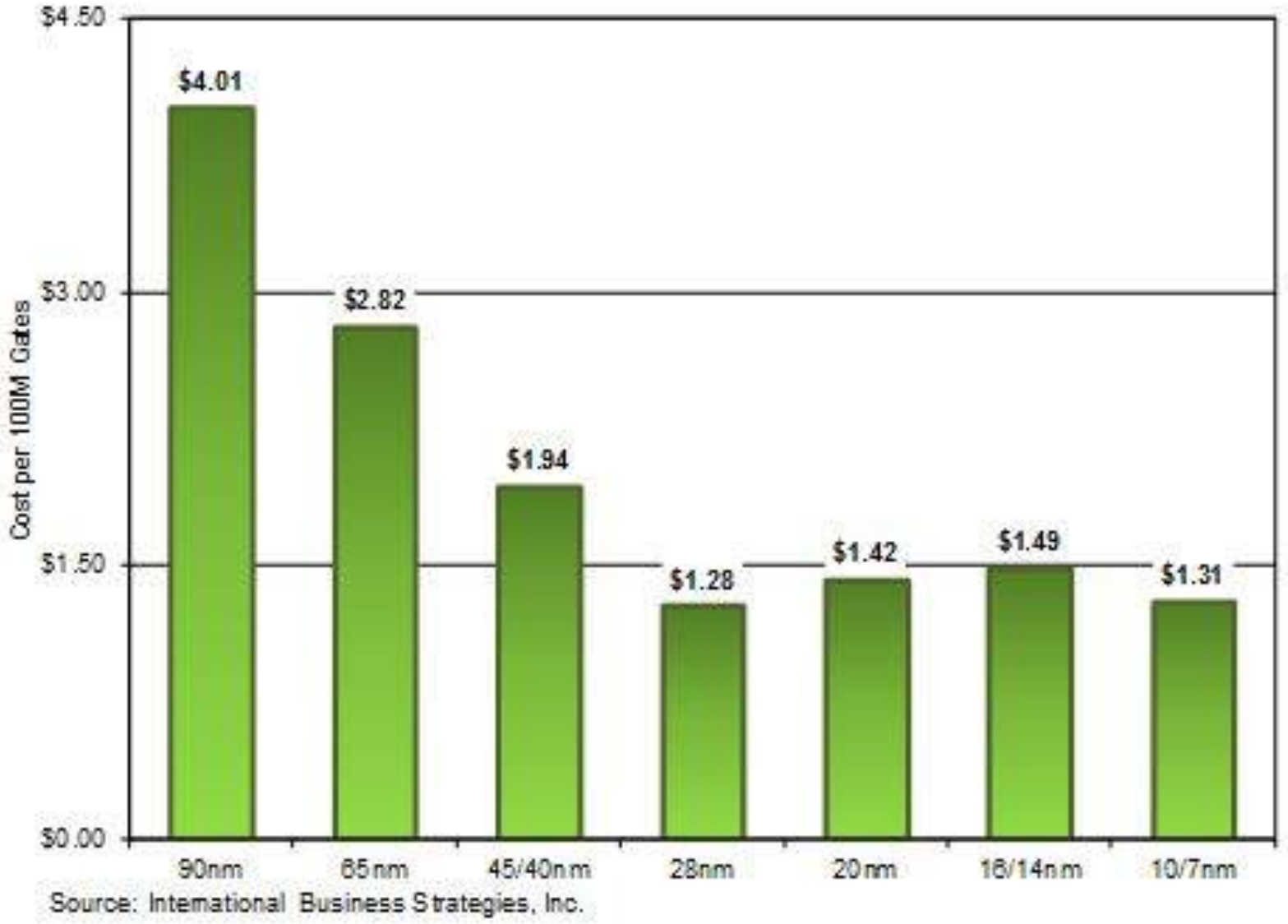

92.11%
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
% Yield
Datalogged Yield
First Pass Yield
95th Percentile Yield
Good
Rejects
Percent Yield
Quantity In
Percent of Total Fail %
Count
Bin Number

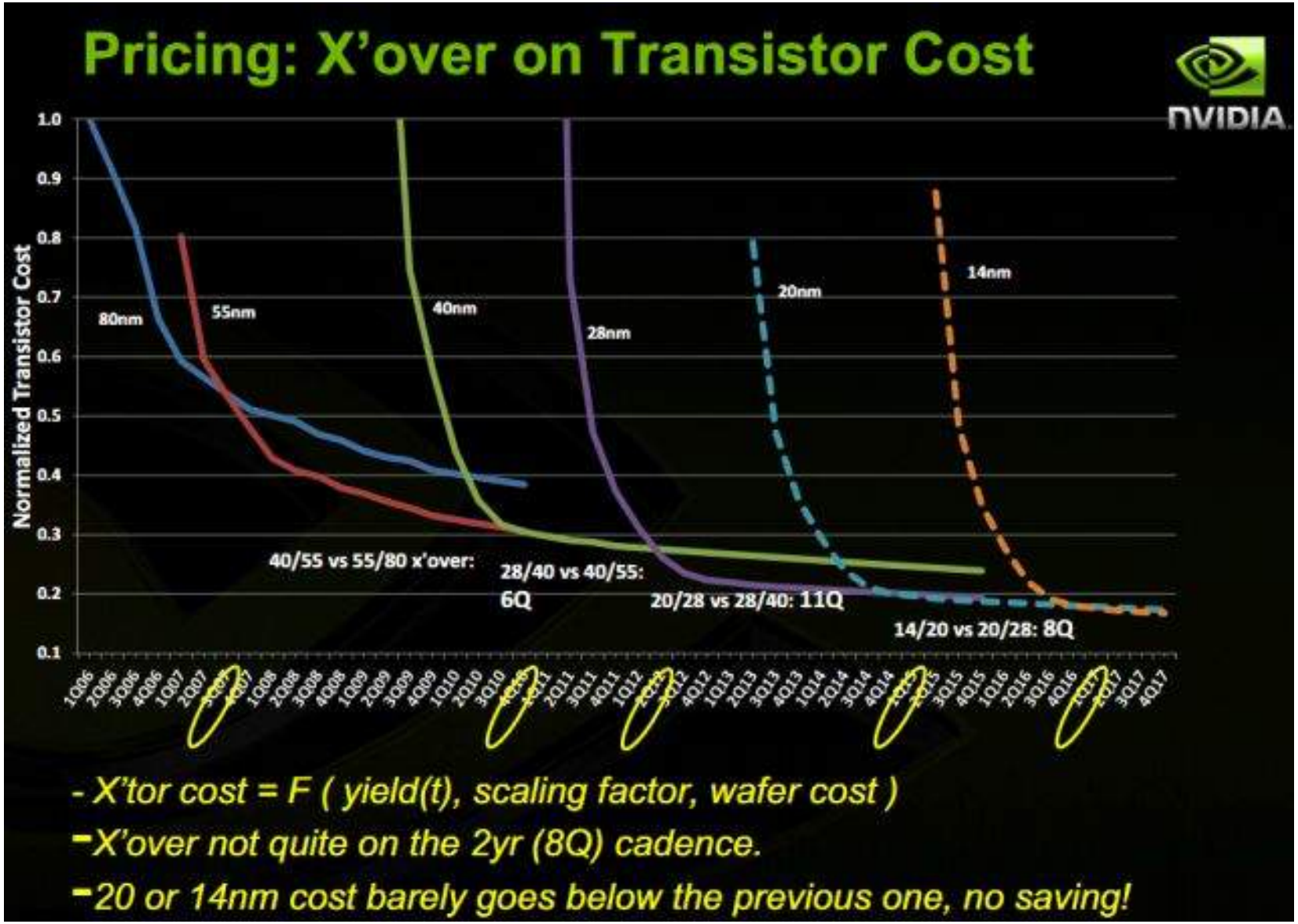

Device Name: DN001YW23 (Wafer Sort)
Test Program: TP00846221
Wafer ID: w01 [1]
Dies Count (T/P/F): 248/230/18
Lot ID: YWLT73528
Notch: D, [Actual: D]
Yield: 92.742%
40
25
10
-5
-20
-35
-56
-37
-18
1
20
39

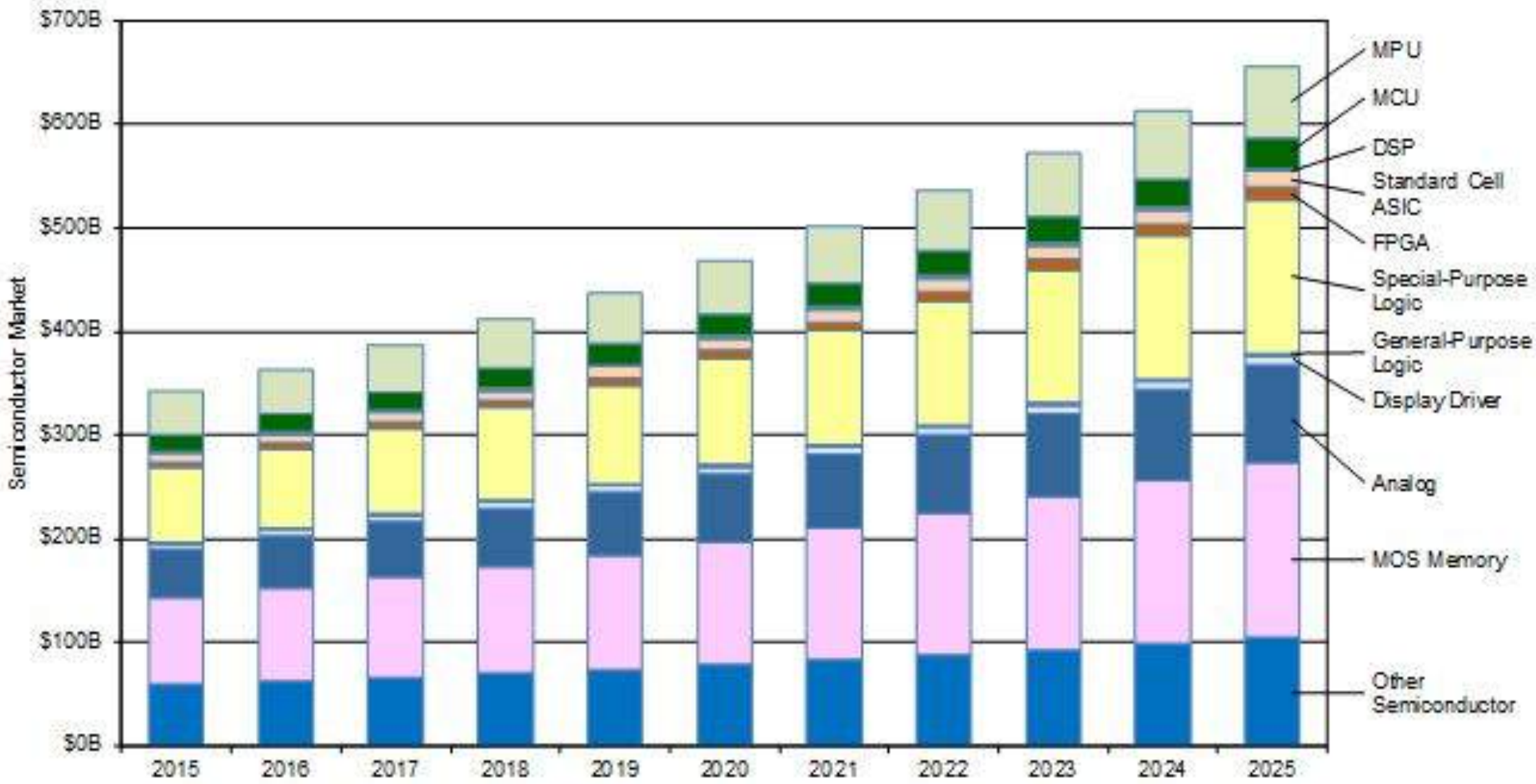

Device Name: DN001YW23 (Final Test)
Test Program: TP00846221
Wafer ID: w01 [1]
Dies Count (T/P/F): 248/220/28
Lot ID: YWLT73528
Notch: D, [Actual: D]
Yield: 88.709%
40
25
10
-5
-20
-35
-56
-37
-18
1
20
39

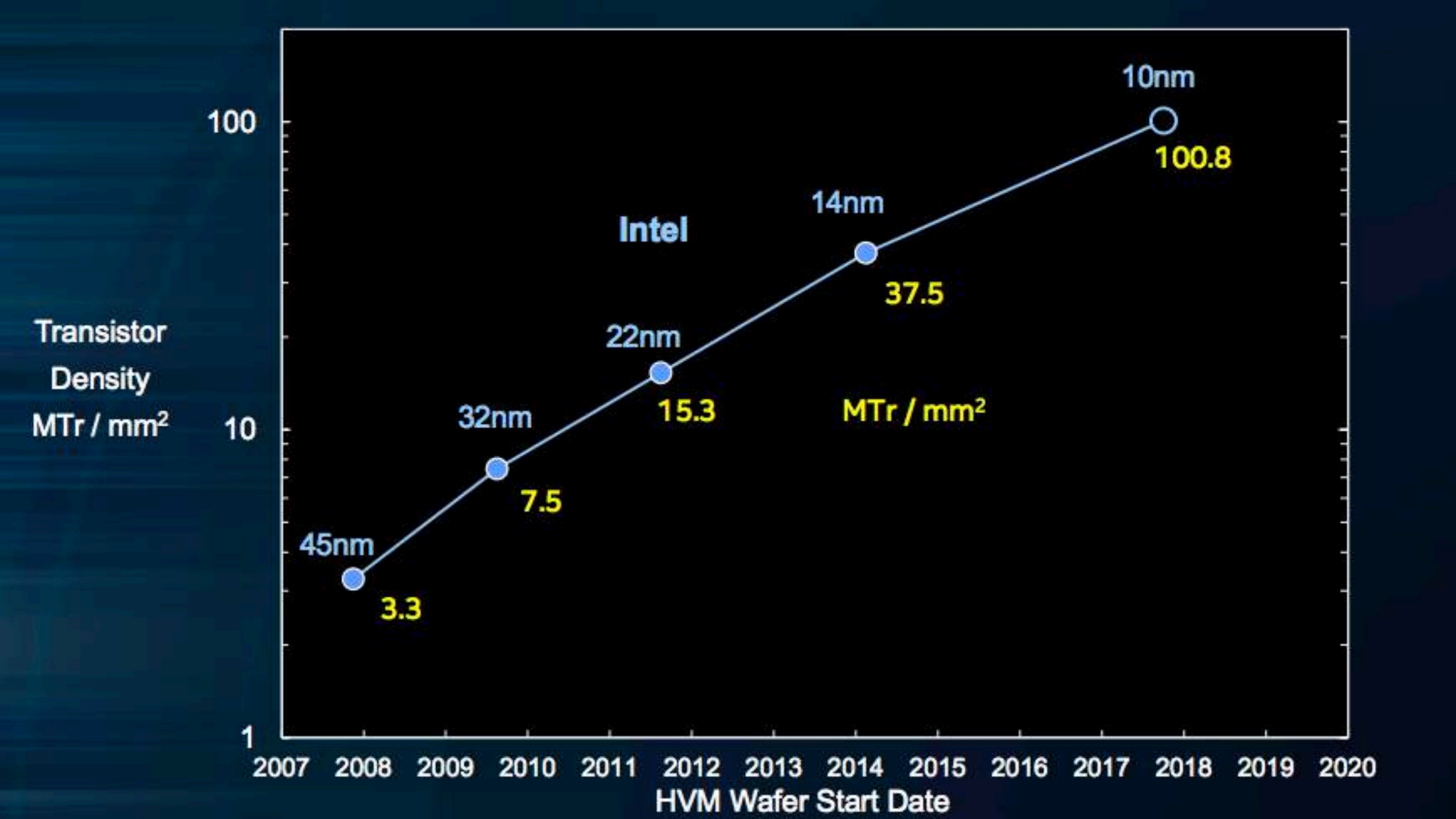

FT Bin Yield Lost
Total Lost Die(s): 10
Lost Yield: 4.033%
Count
Yield(%)
12
10
8
6
4
2
0
100
80
60
40
20
0
4
2
2
1
1
1→13
1→32
1→6
1→11
1→15
Bin Number

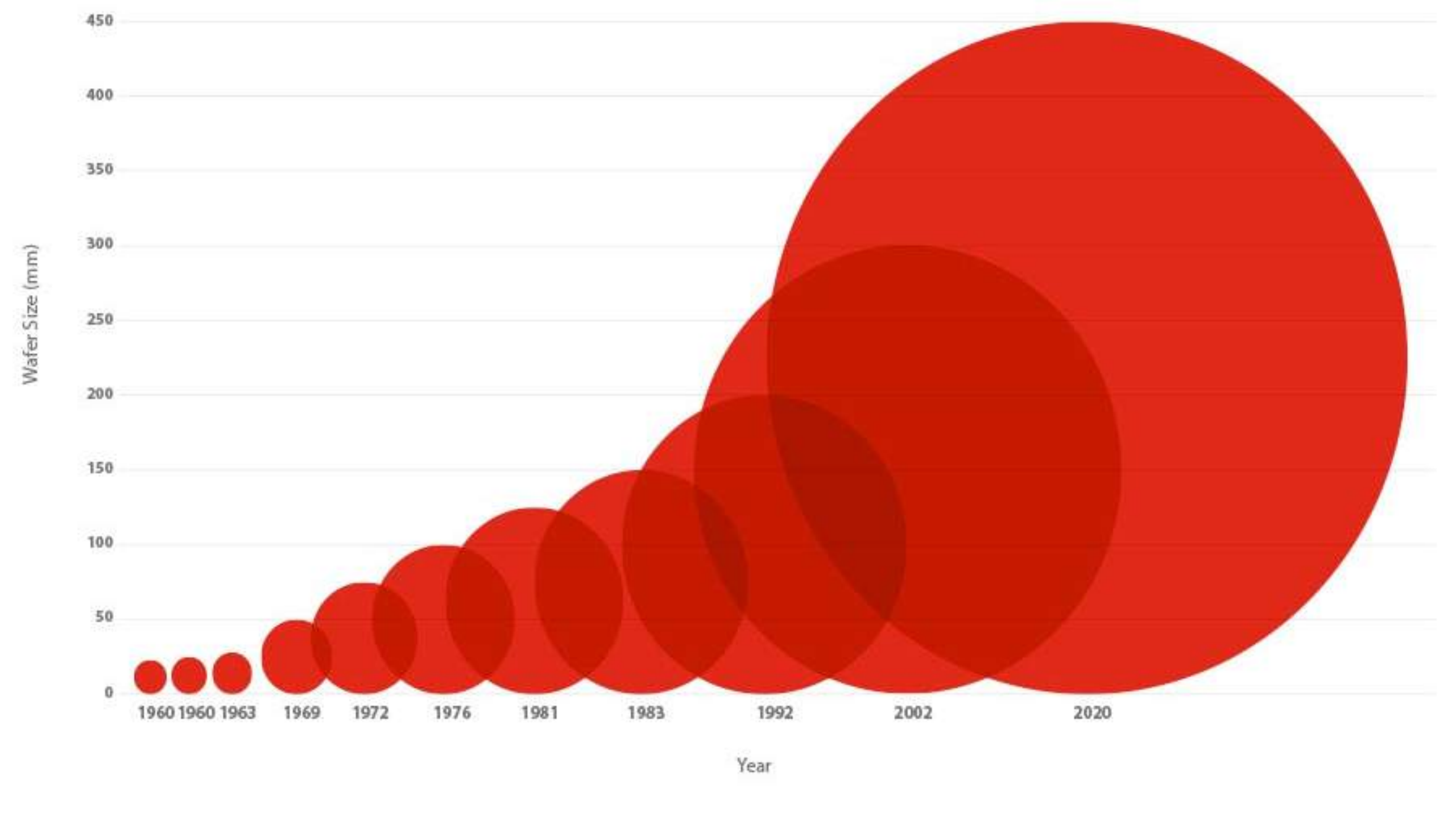

$580M
$435M
$290M
$145M
$0M
Advanced Design Cost
$106.3M
$174.4M
$297.8M
$542.2M
16nm
10nm
7nm
5nm
Validation
Prototype
Software
Physical
Verification
Architecture
IP Qualification


$4.50
$3.00
$1.50
$0.00
Cost per 100M Gates
$4.01
$2.82
$1.94
$1.28
$1.42
$1.49
$1.31
90nm
65nm
45/40nm
28nm
20nm
16/14nm
10/7nm
Source: International Business Strategies, Inc.


Pricing: X'over on Transistor Cost
NVIDIA
Normalized Transistor Cost
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
80nm
55nm
40nm
28nm
20nm
14nm
40/55 vs 55/80 x'over:
28/40 vs 40/55:
6Q
20/28 vs 28/40: 11Q
14/20 vs 20/28: 8Q
- X'tor cost = F (yield(t), scaling factor, wafer cost)
- X'over not quite on the 2yr (8Q) cadence.
- 20 or 14nm cost barely goes below the previous one, no saving!

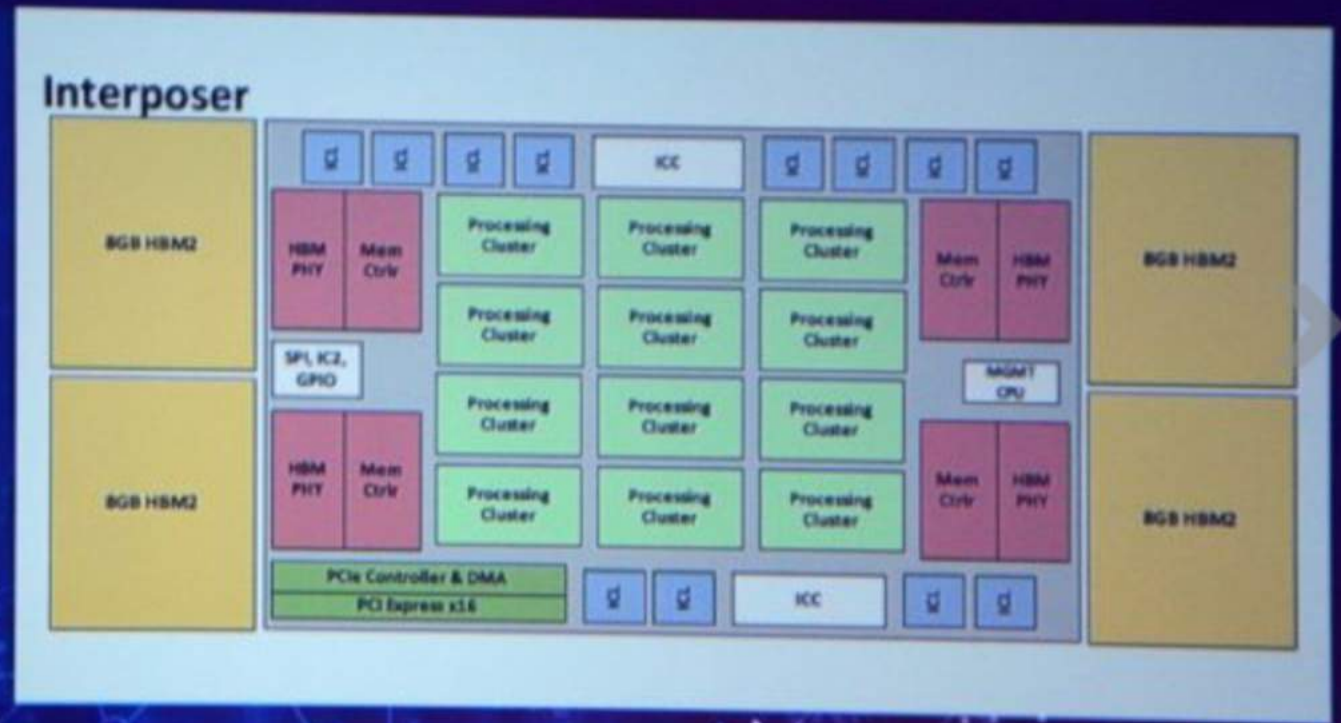

$700B
$600B
$500B
$400B
$300B
$200B
$100B
$0B
Semiconductor Market
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
MPU
MCU
DSP
Standard Cell
ASIC
FPGA
Special-Purpose
Logic
General-Purpose
Logic
Display Driver
Analog
MOS Memory
Other
Semiconductor

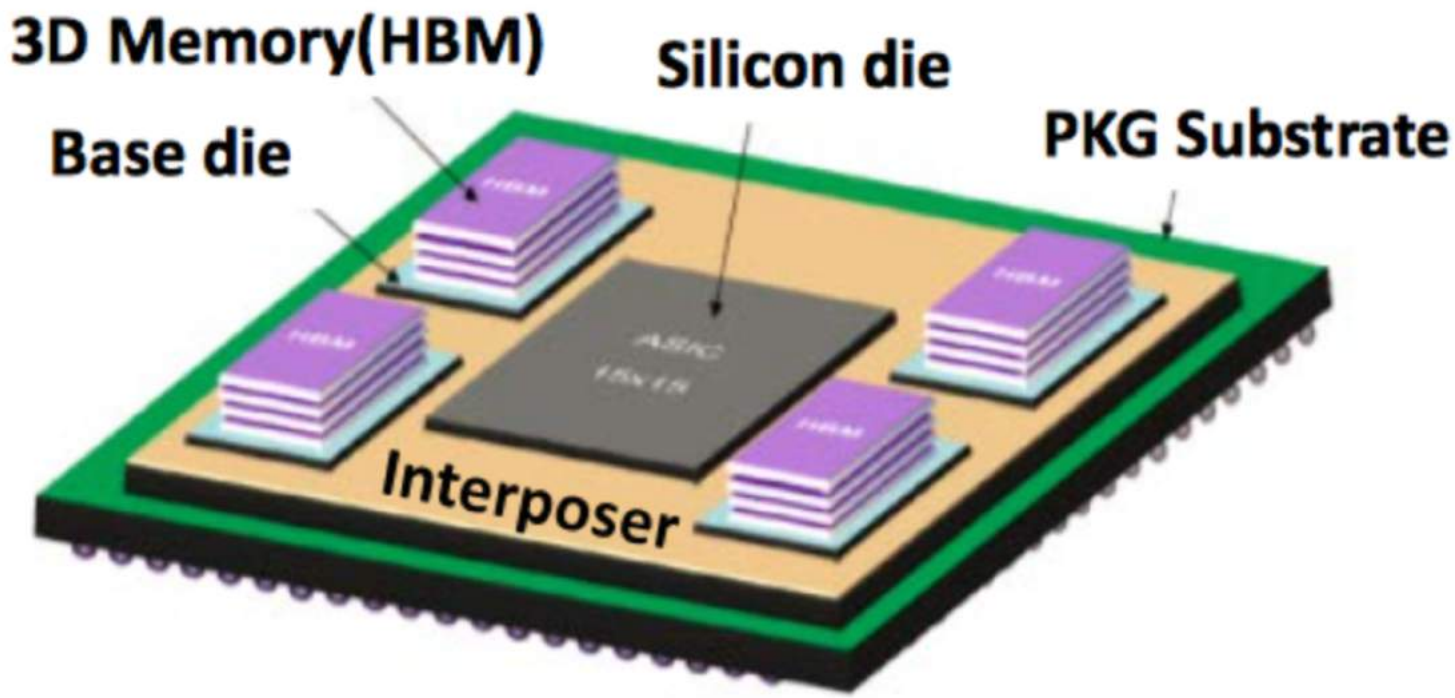

Transistor Density MTr / mm²
100
10
1
Intel
45nm
3.3
32nm
7.5
22nm
15.3
14nm
37.5
10nm
100.8
MTr / mm²
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
HVM Wafer Start Date

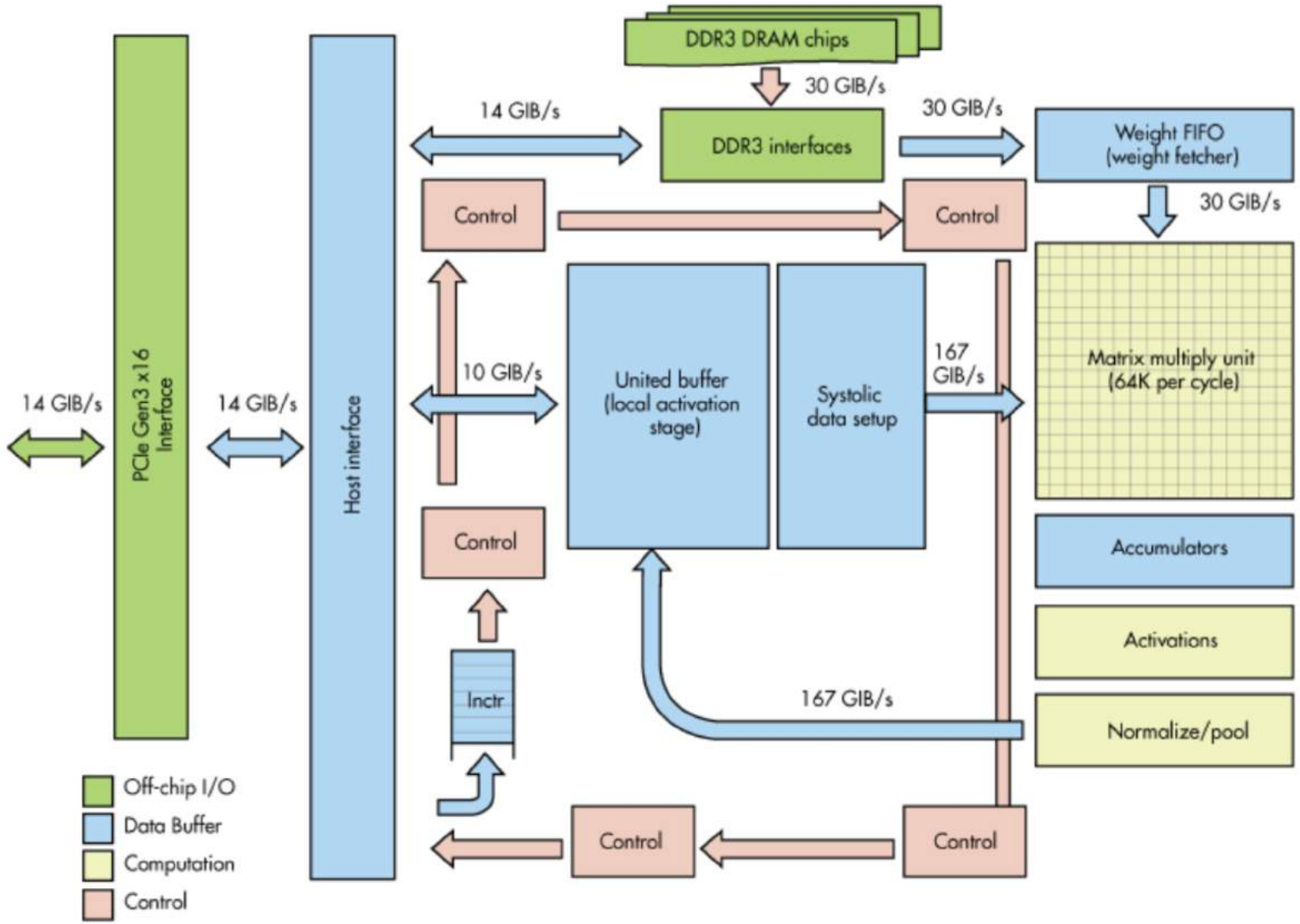

Wafer Size (mm)
450
400
350
300
250
200
150
100
50
0
1960
1960
1963
1969
1972
1976
1981
1983
1992
2002
2020
Year

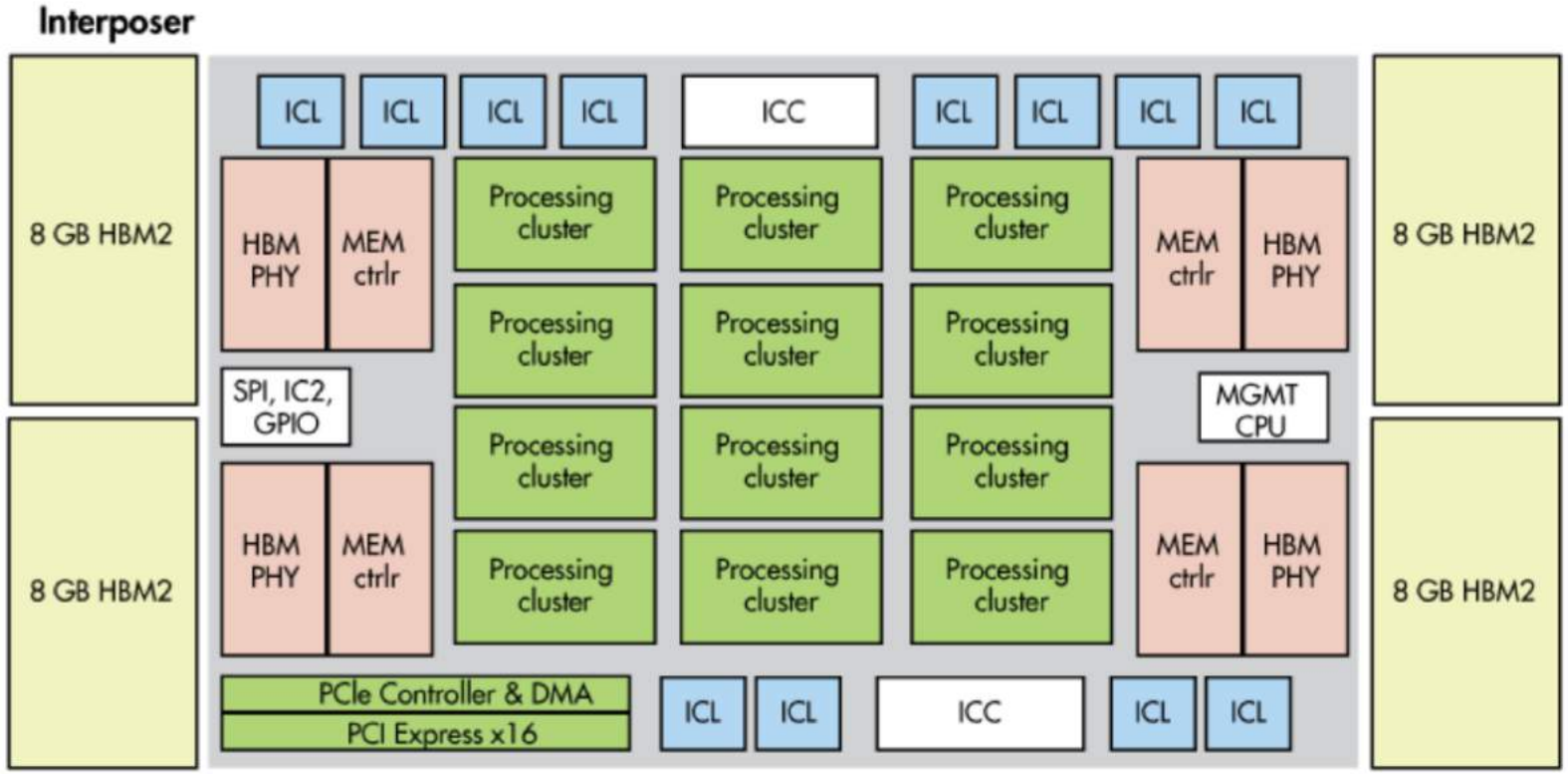
LAKE CREST DEEP LEARNING ARCHITECTURE

- Tensor based architecture
- Flexpoint®
  - Unprecedented levels of parallelism up to 10x of state-of-art
  - Low power per tensor operation
- HBM2 memory: up to12x faster than DDR4
- Proprietary inter-chip links: up to 20x faster than PCIe



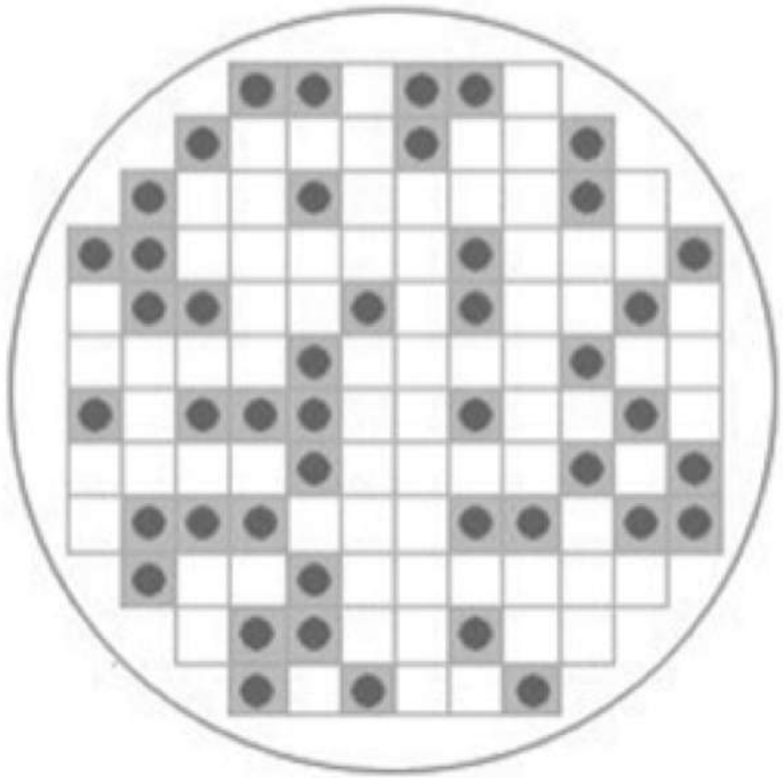

3D Memory(HBM)
Silicon die
Base die
PKG Substrate
Interposer


DDR3 DRAM chips
30 GIB/s
14 GIB/s
DDR3 interfaces
30 GIB/s
Weight FIFO
(weight fetcher)
30 GIB/s
Control
Control
14 GIB/s
PCIe Gen3 x16
Interface
14 GIB/s
Host interface
10 GIB/s
United buffer
(local activation
stage)
Systolic
data setup
167
GIB/s
Matrix multiply unit
(64K per cycle)
Control
Accumulators
Activations
Inctr
167 GIB/s
Normalize/pool
Control
Control
Off-chip I/O
Data Buffer
Computation
Control

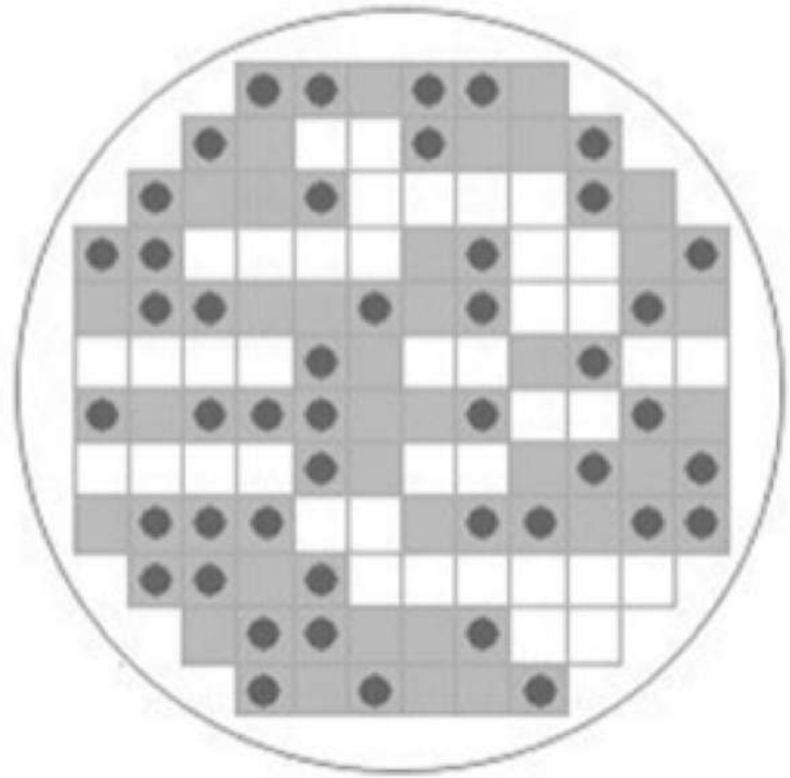


3. Intel's Lake Crest uses processing clusters optimized for AI applications.

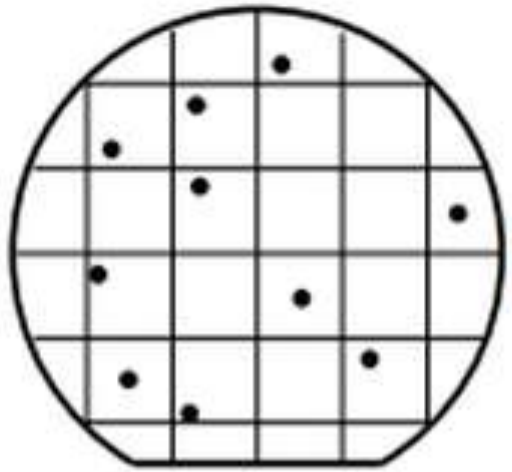
$$\frac{\text{Total number of functional chips produced}}{\text{Number of designed chips}} \times 100 = \text{Yield}$$



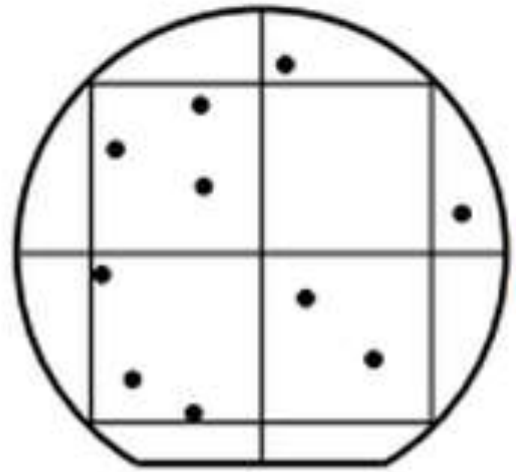
(a) About 60% yield



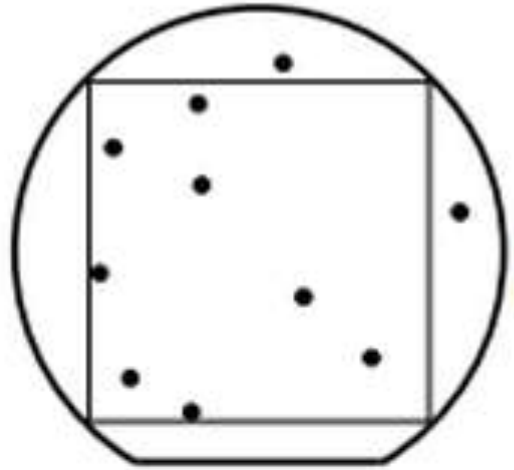
(b) About 30% yield



8 Good Dice
Out of 16 = 50%



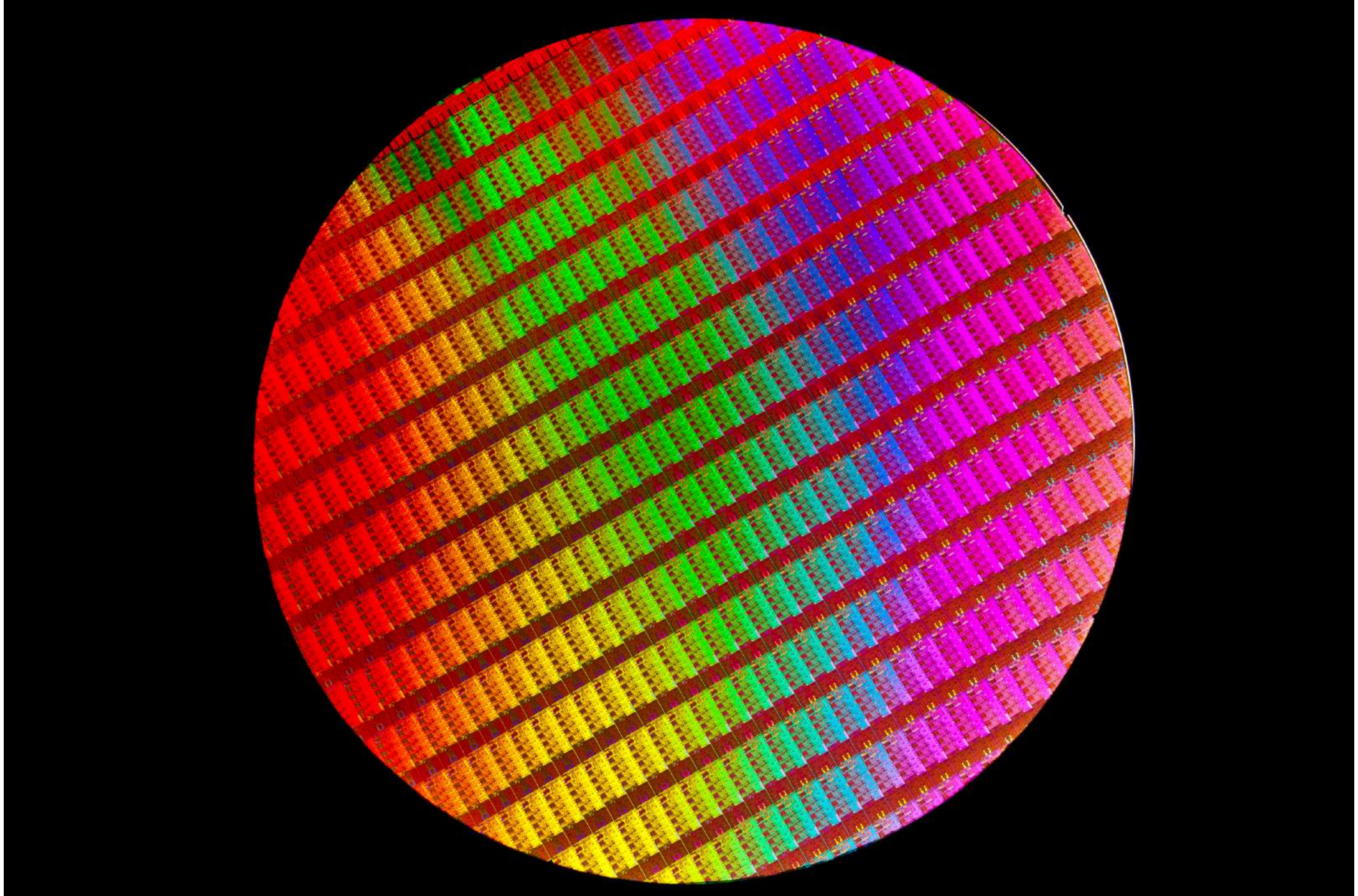
1 Good Die
Out of 4 = 25%

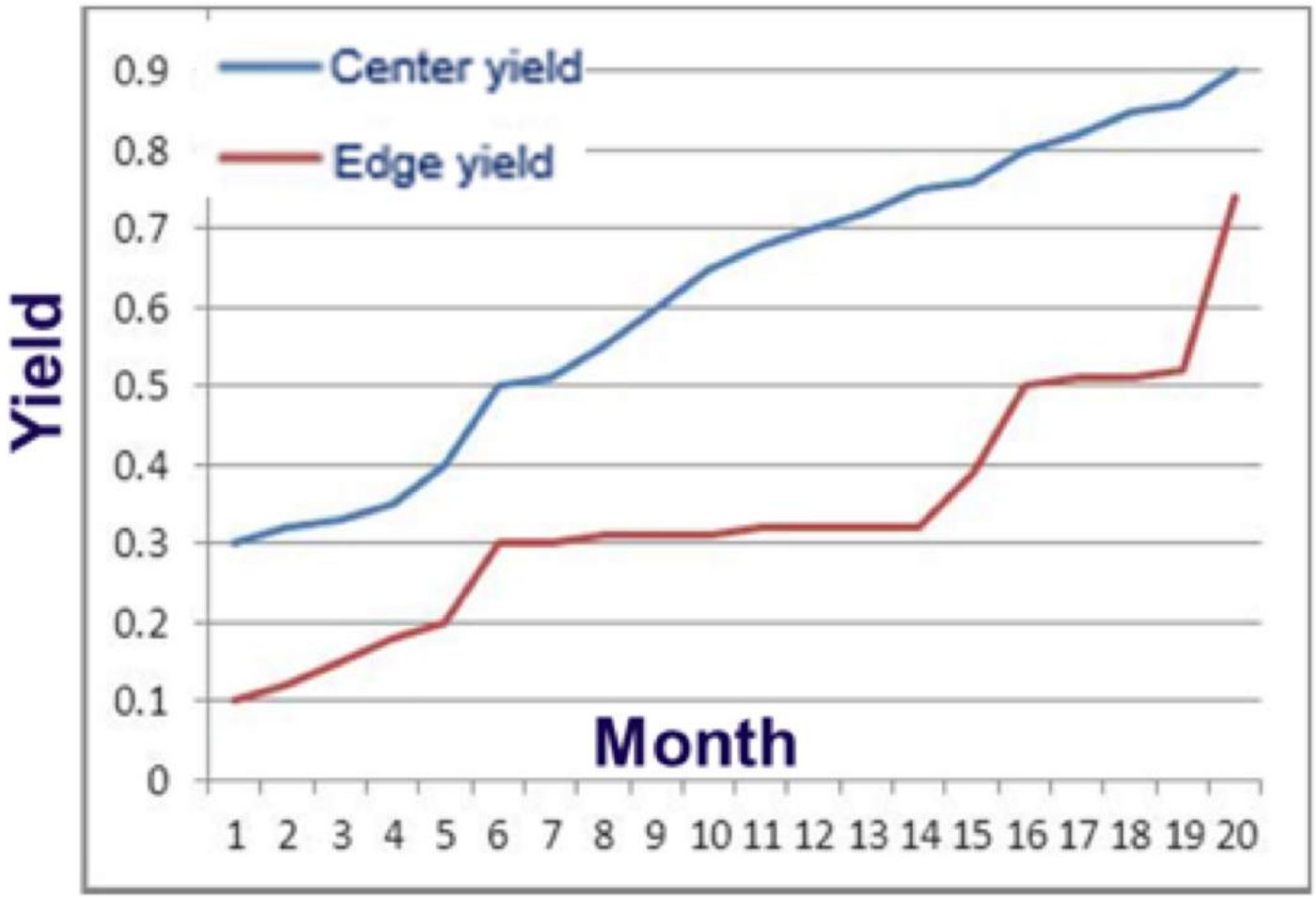


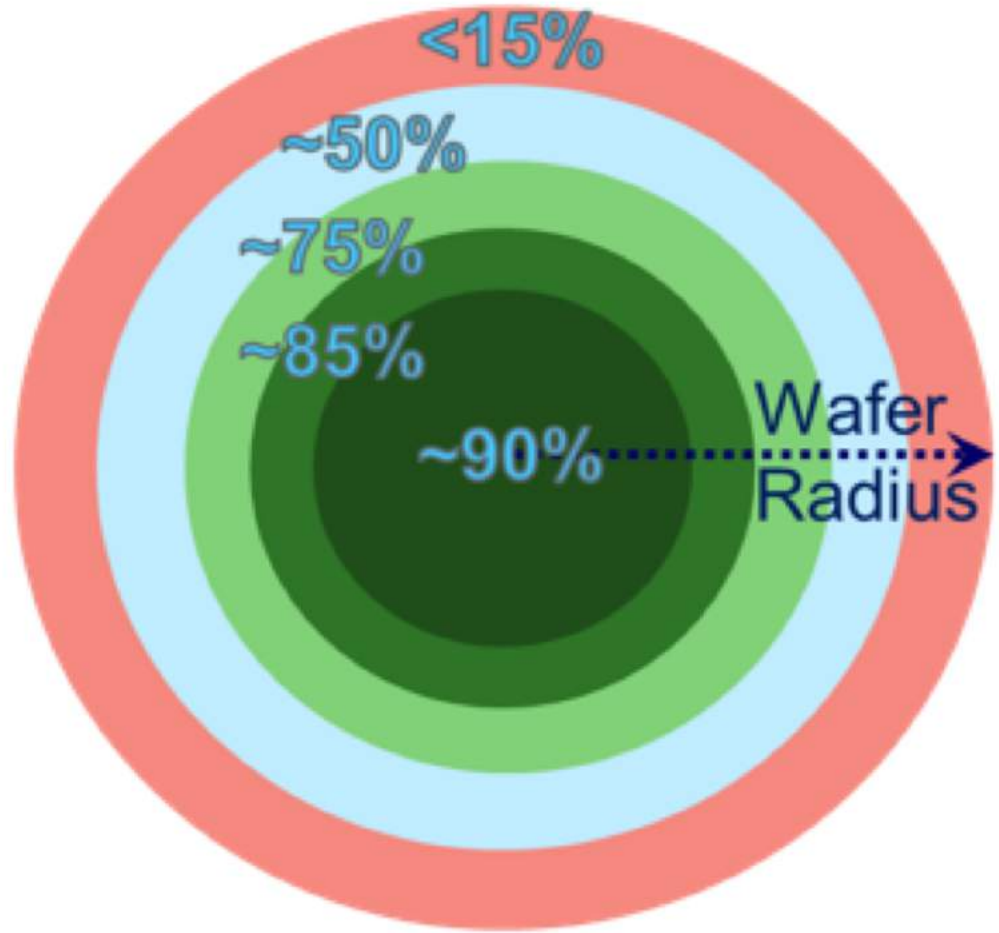
0 Good Die
Out of 1 = 0%

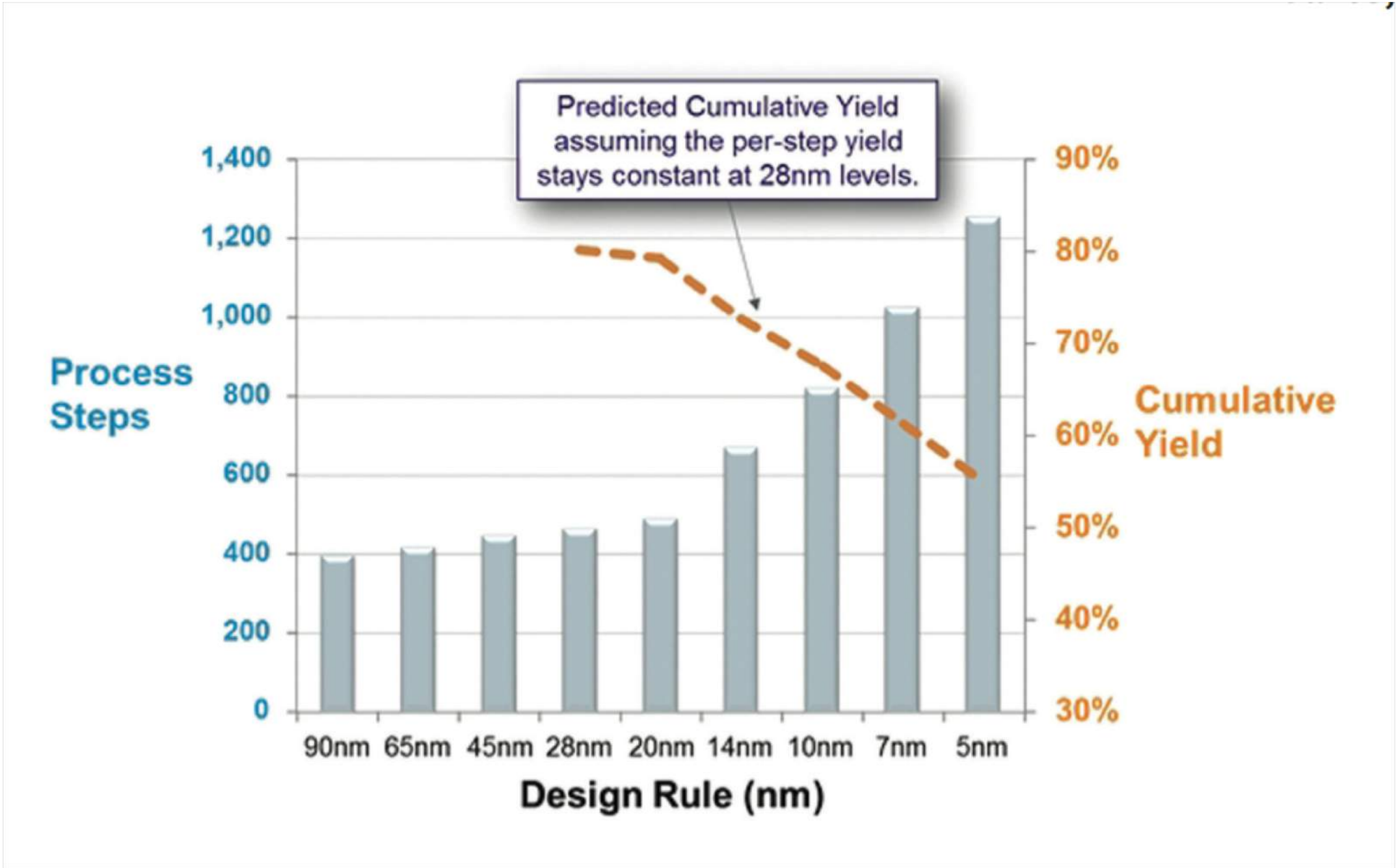
Source: ICE

7438B




Yield
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Center yield
Edge yield
Month
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20


<15%
~50%
~75%
~85%
~90%
Wafer
Radius


Predicted Cumulative Yield assuming the per-step yield stays constant at 28nm levels.
Process Steps
1,400
1,200
1,000
800
600
400
200
0
90%
80%
70%
60%
50%
40%
30%
Cumulative Yield
90nm 65nm 45nm 28nm 20nm 14nm 10nm 7nm 5nm
Design Rule (nm)

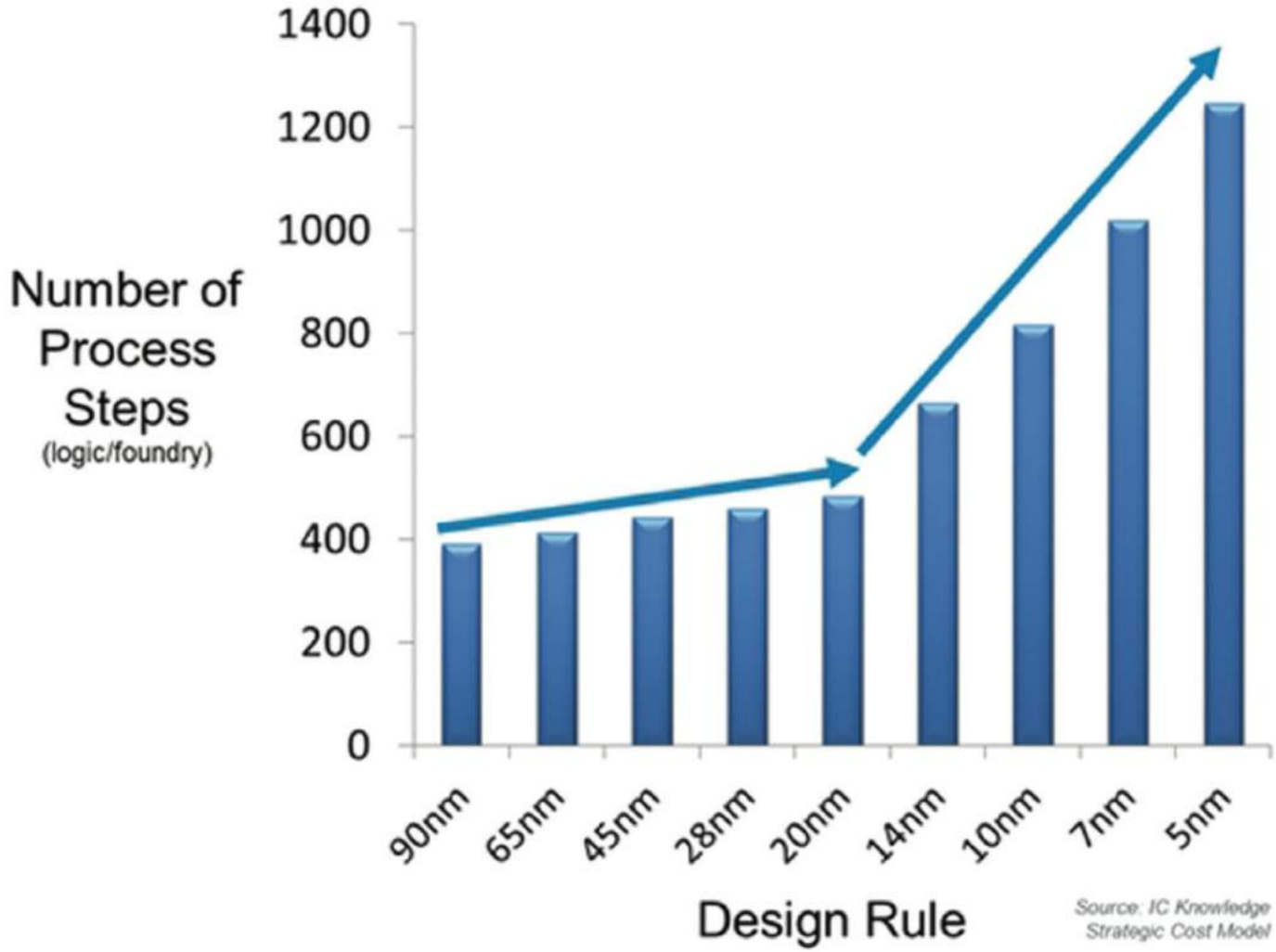
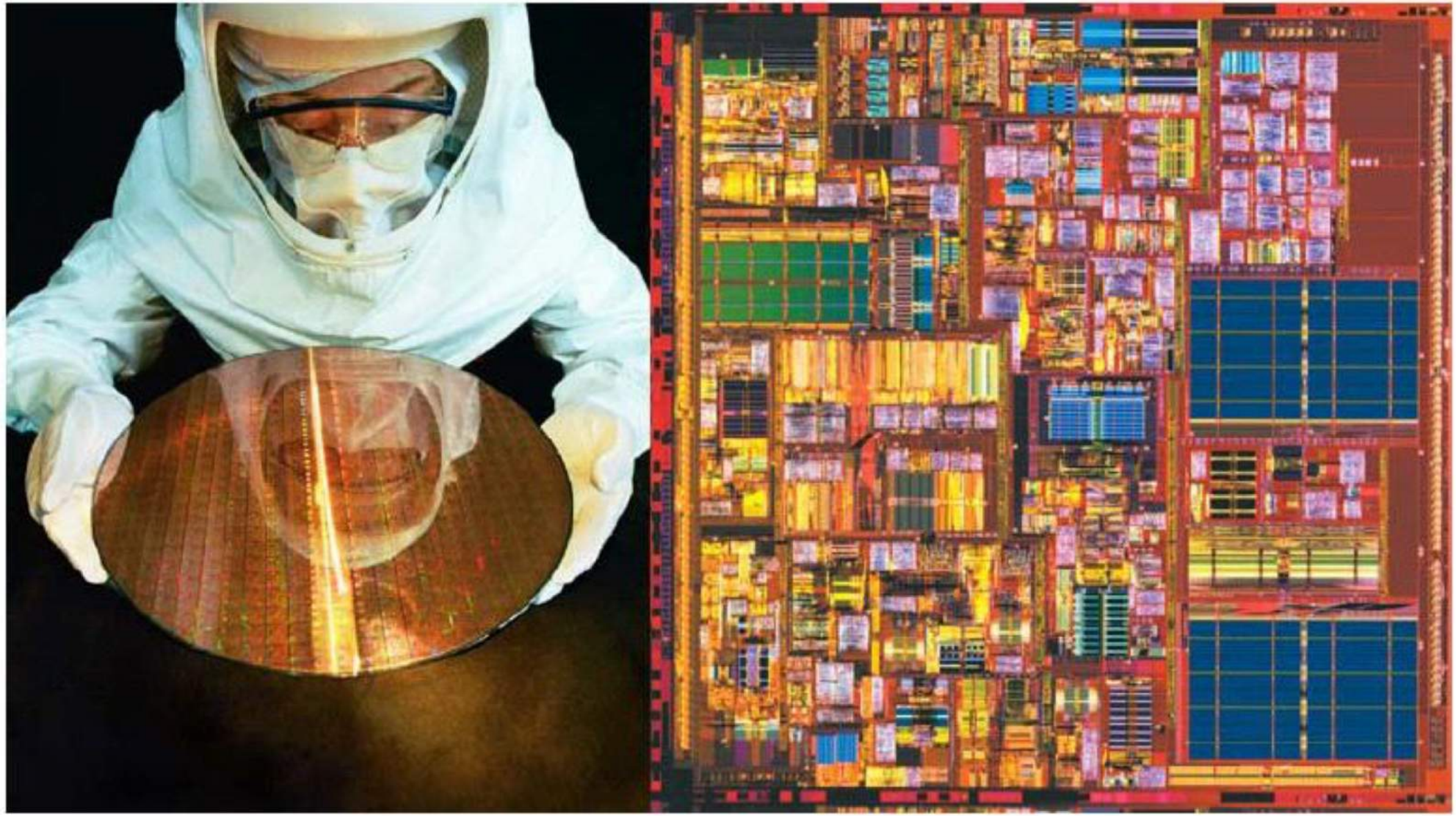


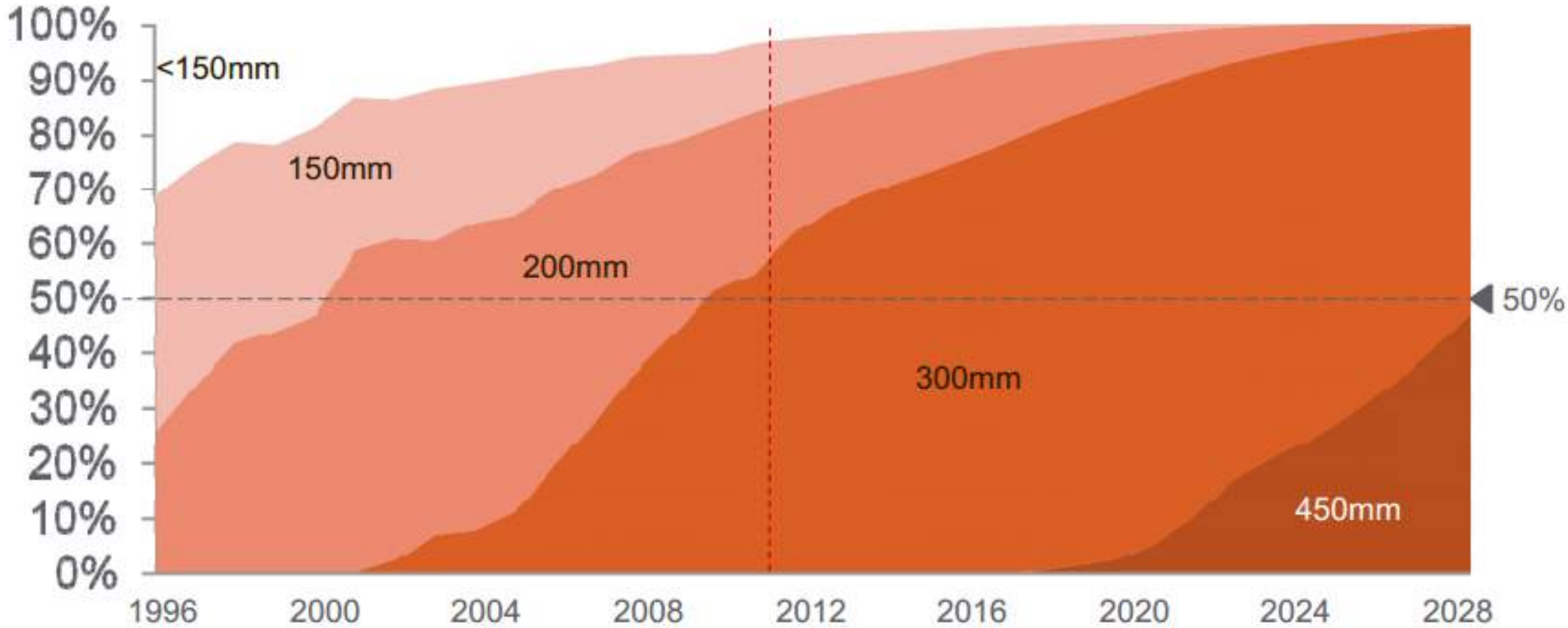
FIGURE 2. The number of process steps increases dramatically with decreasing design rule starting at the 16/14nm design node. Source: IC Knowledge Strategic Cost Model.

Silicon Wafer



300mm wafer and Pentium 4 IC. Photos courtesy of Intel.

Historic and projected semiconductor demand (% of sq. inch of Si wafer), by wafer size



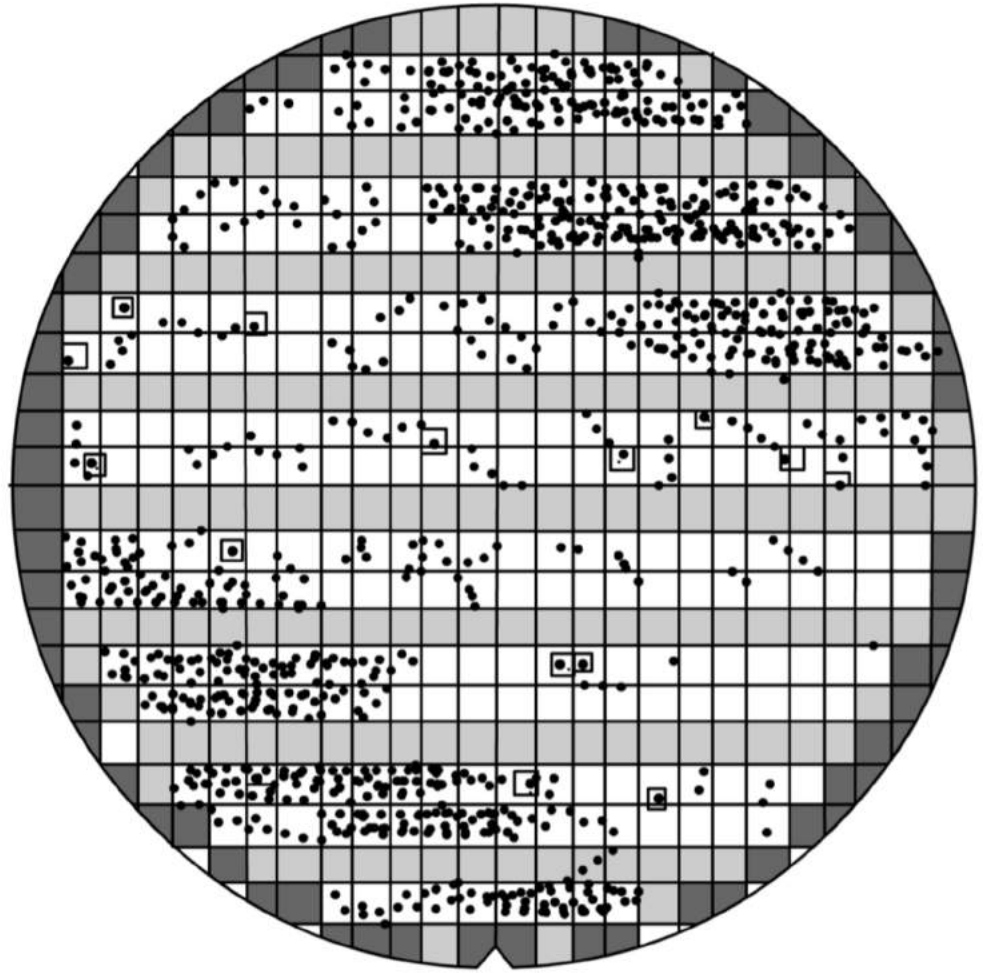
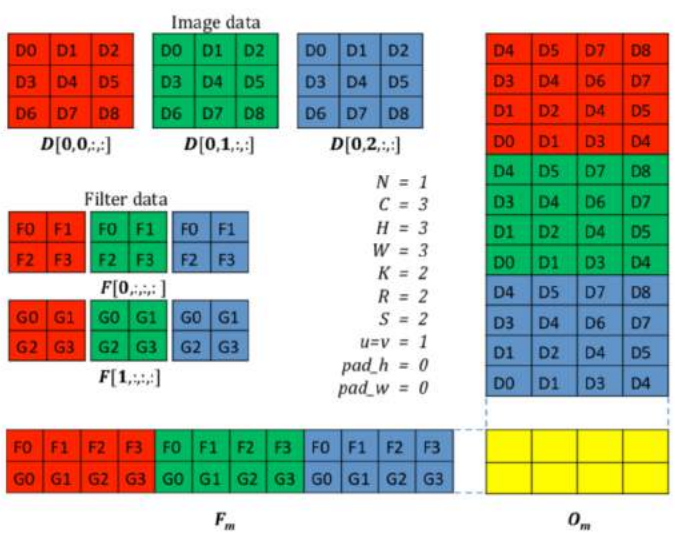
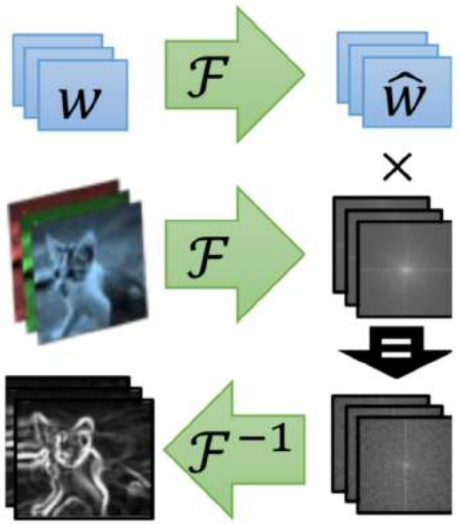


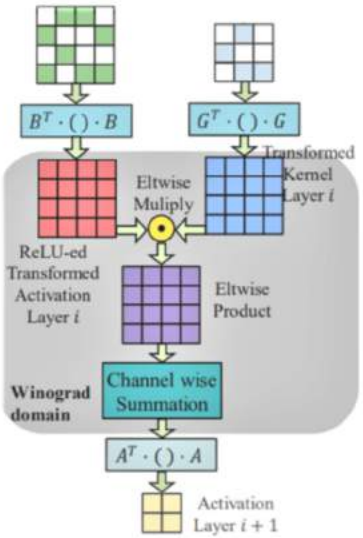
Figure 5: Clustering of defects on a wafer. While there is a random population of defects, there is clustering at the lower left corner and upper right corner of the wafer. This type of clustering happens when a certain process or a subset of processes are responsible for the defects. Source *http://www.geek.com/chips/ibm-toshiba-and-sony-form-32-nm-alliance-561548/*



(a) `im2col` (adapted from [38])



(b) FFT



(c) Winograd (adapted from [162])

Fig. 12. Computation Methods for Convolutional Operators

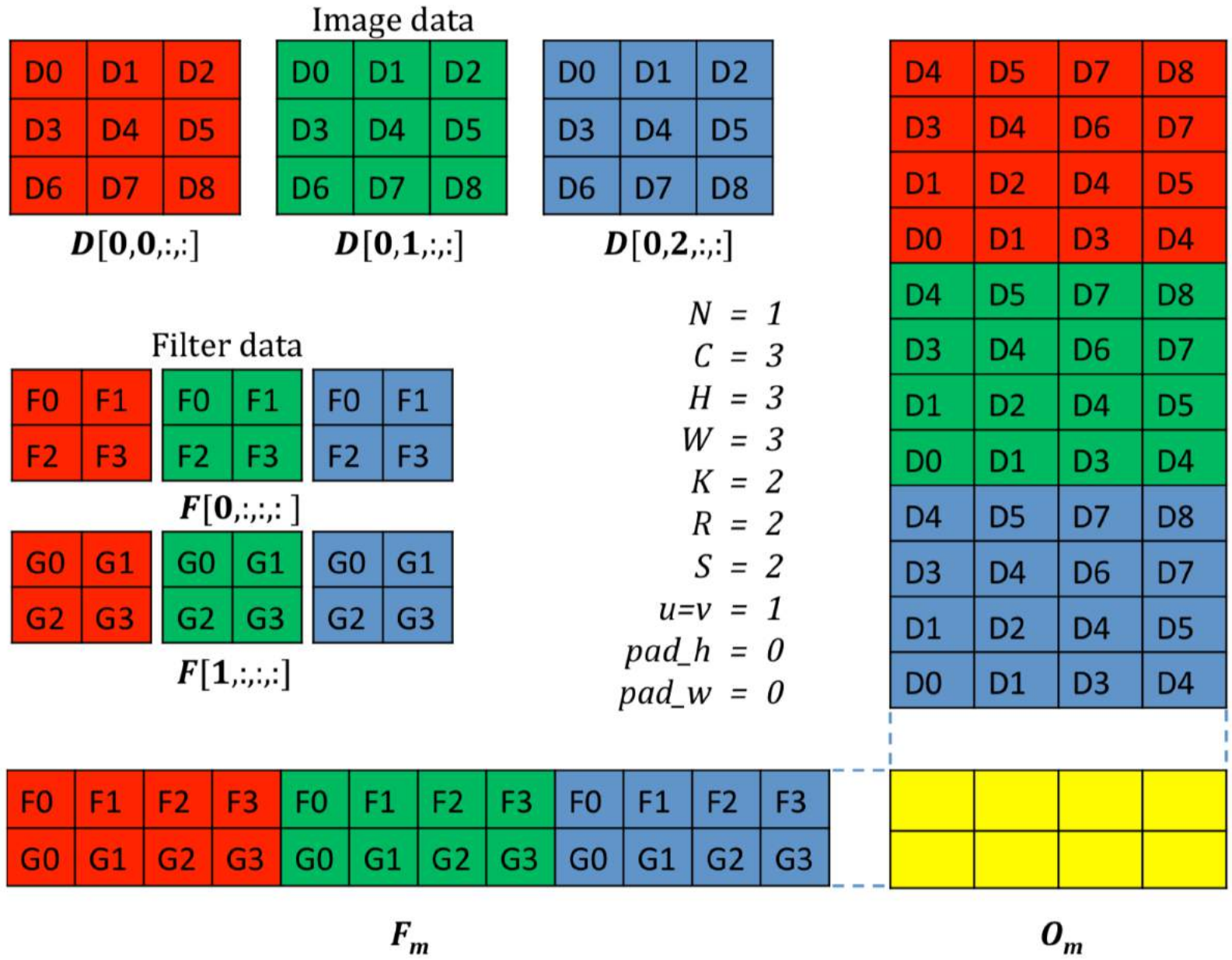
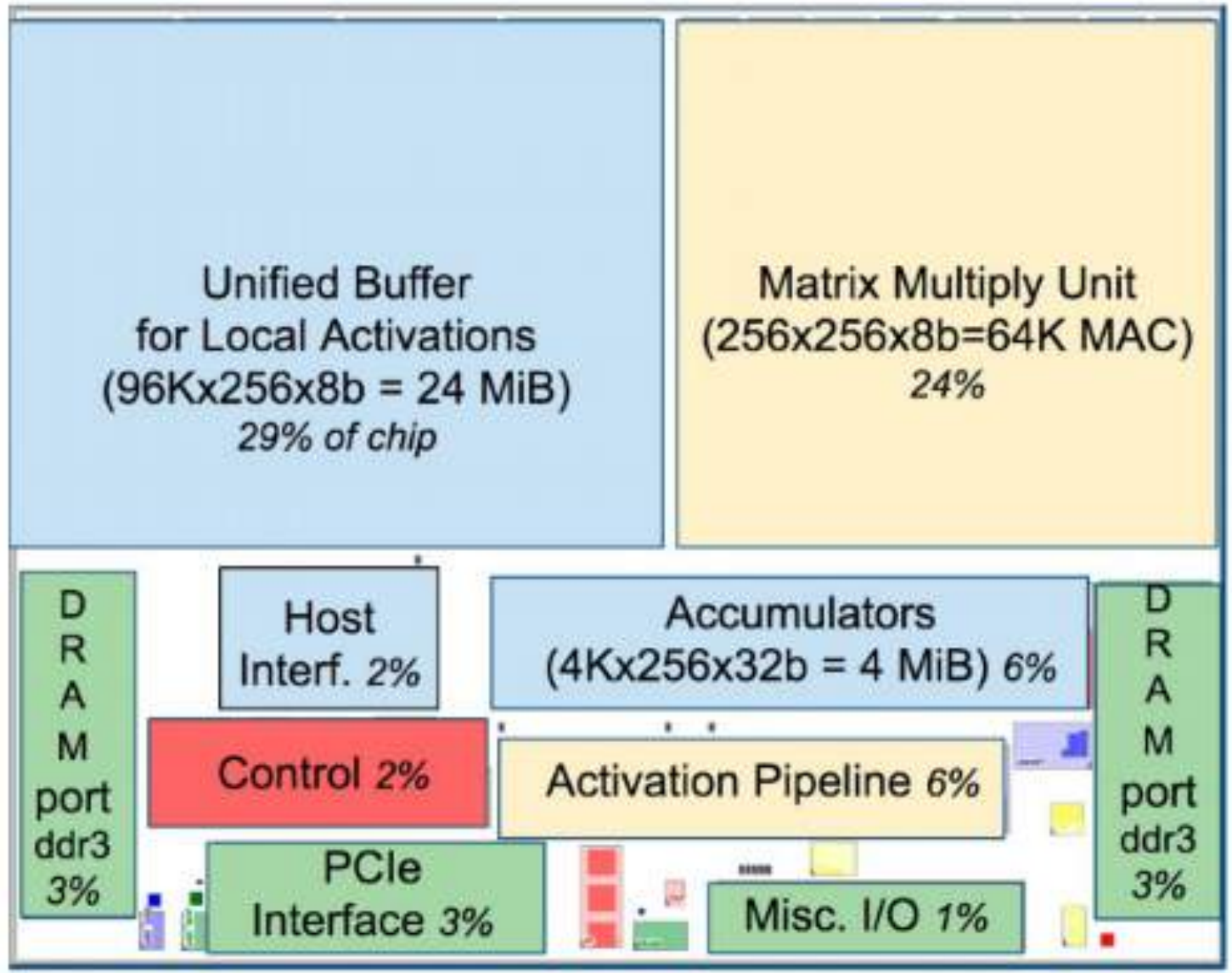
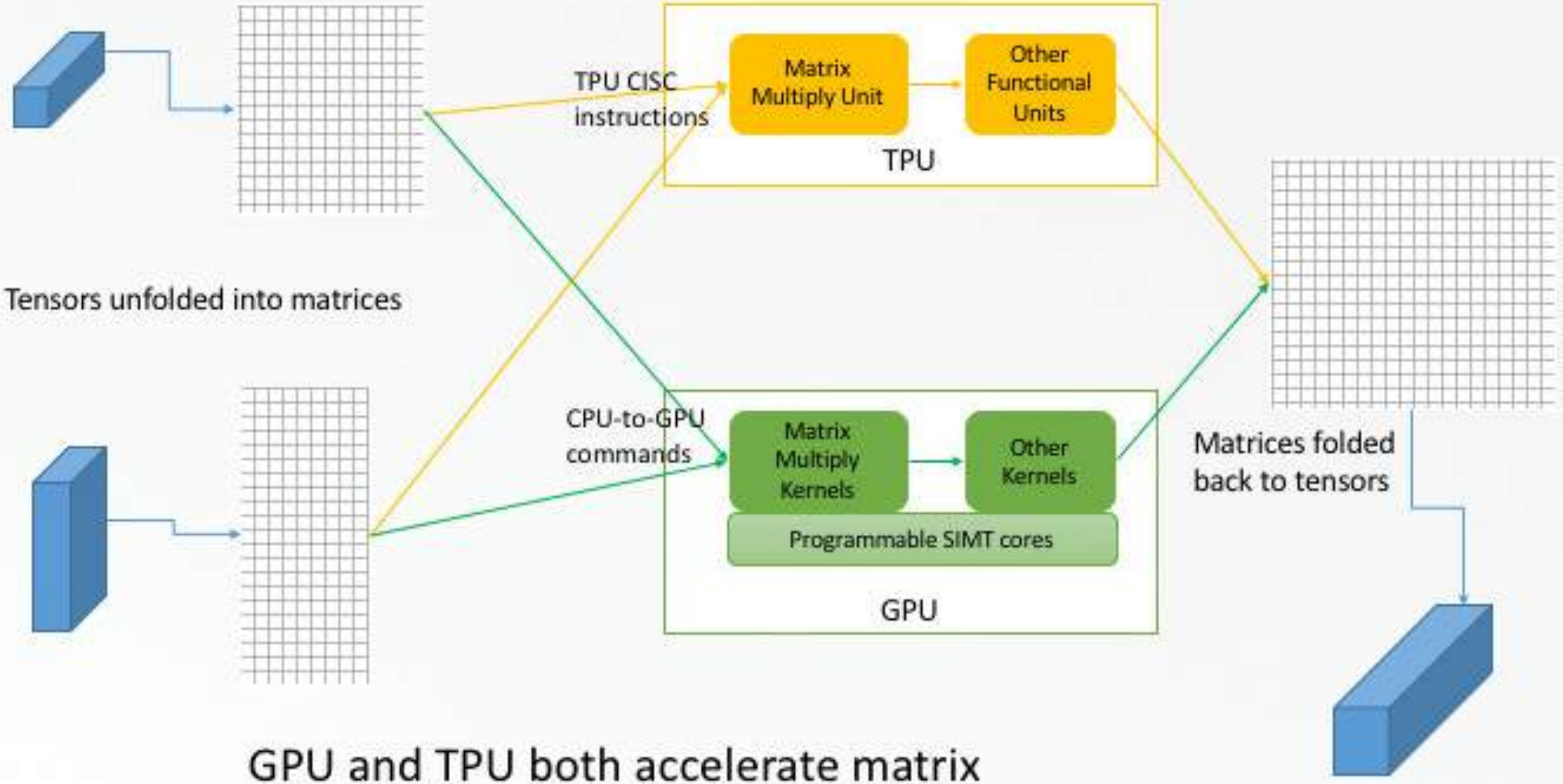


Figure 1: Convolution lowering


Unified Buffer
for Local Activations
(96Kx256x8b = 24 MiB)
29% of chip
Matrix Multiply Unit
(256x256x8b=64K MAC)
24%
D
R
A
M
port
ddr3
3%
Host
Interf. 2%
Accumulators
(4Kx256x32b = 4 MiB) 6%
Control 2%
Activation Pipeline 6%
PCIe
Interface 3%
Misc. I/O 1%
D
R
A
M
port
ddr3
3%



GPU and TPU both accelerate matrix computations from Computation Graph

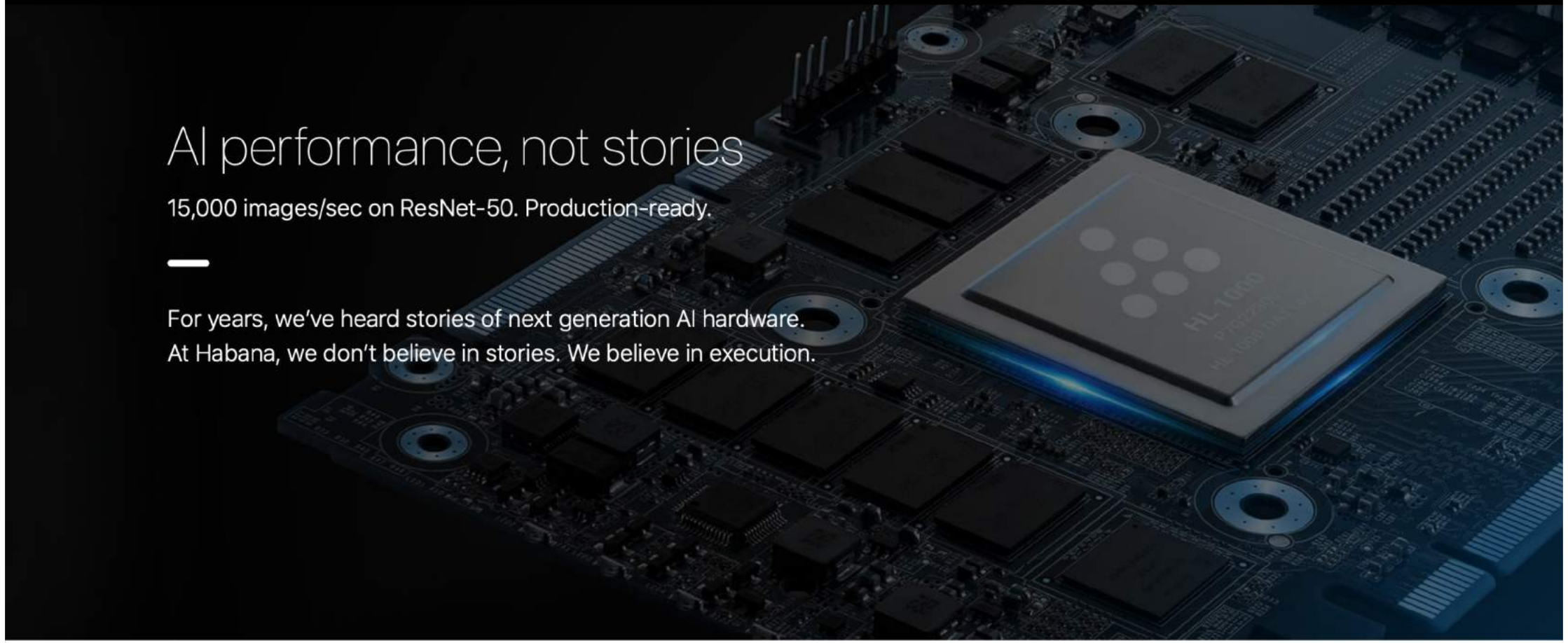
Case Studies of New DL Arch’s

Habana

Habana, an Israeli startup, is sampling the 16nm Goya chip for inference. On the popular ResNet-50 neural network, Goya achieves 15,088 images per second (IPS) with a batch size of 10, or 7,107 IPS with batch size of 1. These measured scores are 5x and 6x faster, respectively, than Nvidia's posted V100 test results. Goya also uses less power than the V100, having a 200W TDP and consuming only 100W when running at 15,000 IPS. The secretive startup has withheld all details of its custom architecture as well as its funding.

habana

Home Products News About us Contact us



AI performance, not stories

15,000 images/sec on ResNet-50. Production-ready.

For years, we've heard stories of next generation AI hardware.
At Habana, we don't believe in stories. We believe in execution.

Pure-AI processing is here

To meet the ever-increasing demand for AI workloads, the industry needs a new kind of processing.

Legacy CPU and converted-GPU architectures are not delivering the required performance.

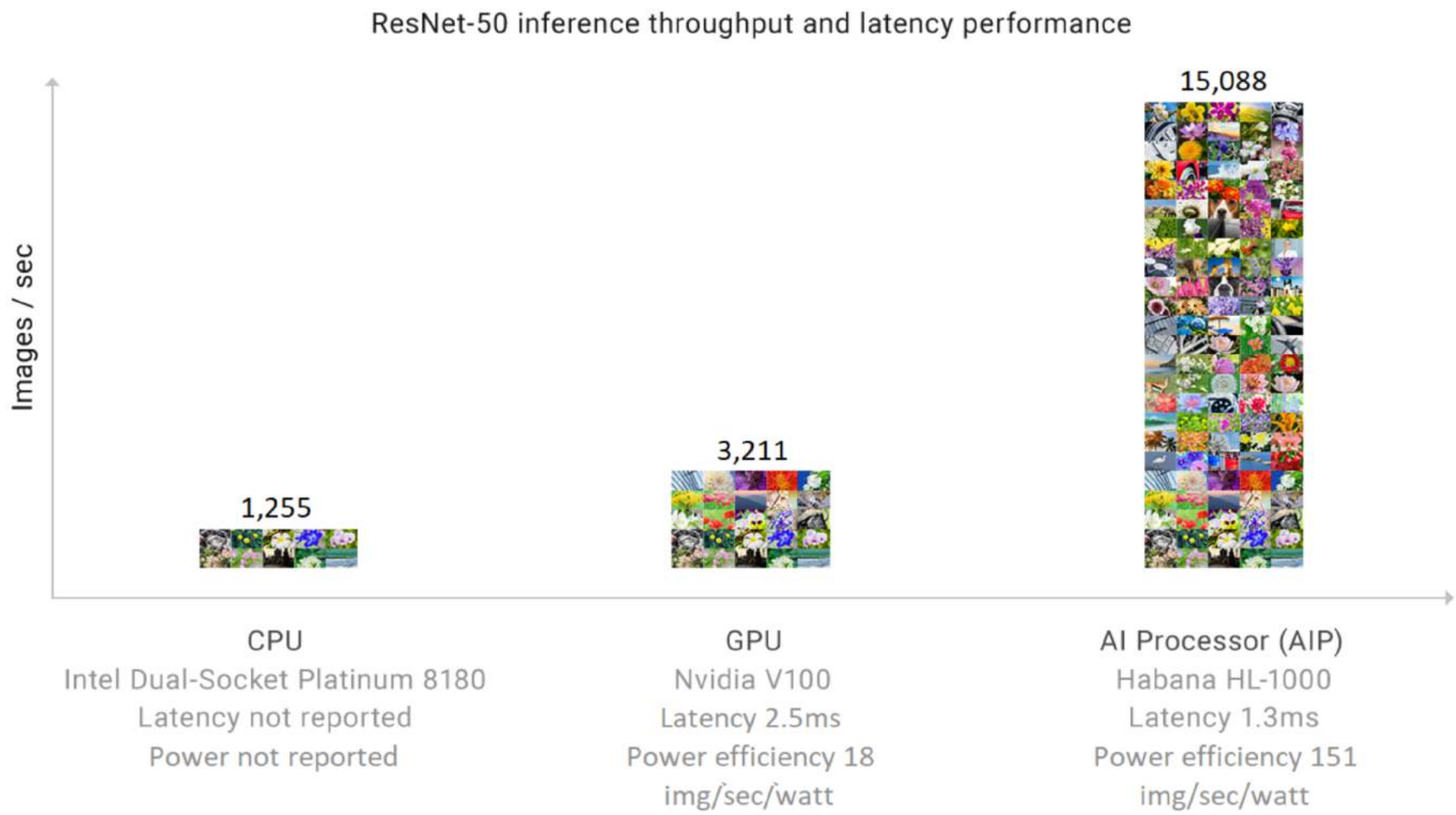
Introducing the first purpose-built AI processing, ready for production today.

Pure-AI processing is here

To meet the ever-increasing demand for AI workloads, the industry needs a new kind of processing.

Legacy CPU and converted-GPU architectures are not delivering the required performance.

Introducing the first purpose-built AI processing, ready for production today.



Inference

available now

habana Goya™

Training

HL-2000 will be sampling in Q2-2019

Training performance linearly scalable to thousands of devices

2 Terabit/second interconnect bandwidth per device

habana Gaudi™

Seamless integration with software frameworks

Habana's seamless integration with popular frameworks and ONNX exchange format allows developers to quickly use complex AI resources without having to be experts. For power users looking for further customization and optimization, Habana offers rich Software Development Tools and a Neural Network Profiler.

TensorFlow PyTorch Caffe2 mxnet ONNX

Graphcore

Having $112 million from premier venture-capital firms, Graphcore has no reason to be shy. The company is nearing production of its first chip, Colossus. According to simulations, a PCIe card using two Colossus chips will exceed 2,000 IPS for ResNet-50 training (a more complex task than inference). Such a score would beat the V100's measured performance by 50%. The chip uses an inno- vative architecture with 1,216 independent cores

One Colossal Chip

Code-named Colossus, the GC2 is among the largest processors ever fabricated. Built in 16nm, the die measures 806mm^2, just about reaching the reticle limit. The 12nm V100 die is slightly bigger at 815mm^2, but the GC2 packs 23.6 billion transistors—about 15% more than Nvidia's chip.

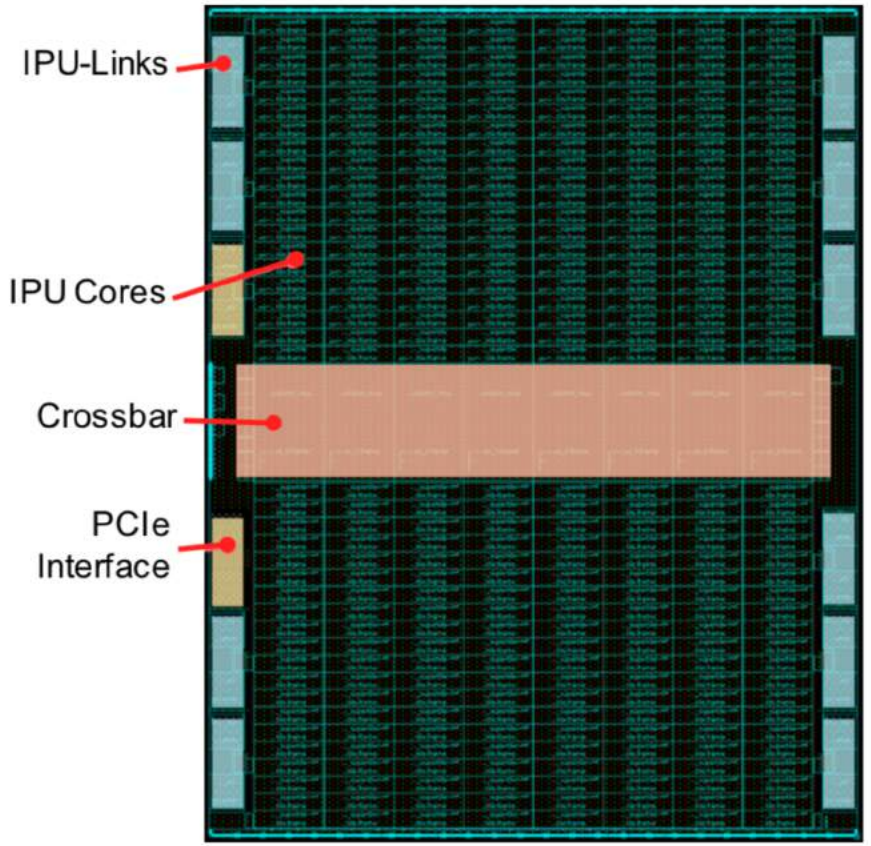


Figure 1. Graphcore GC2 die plot. At 806mm2 and 23.6 billion transistors, the 16nm chip is one of the largest processors ever fabricated. Each "supertile" visible in the figure contains four IPU cores and 1MB of SRAM. (Photo source: Graphcore)

GC2 can generate up to 250Tflop/s at a 300W TDP—twice the performance of a V100 card at a similar power. The startup offers ONNX and TensorFlow drivers, easing the implementation of DNNs on its chip.

At $112 million, Graphcore is one of the best-funded AI-chip startups. Sequoia, a top venture-capital firm, has invested $50 million; strategic investors include Bosch, Dell, and Samsung. Graphcore was founded by CEO Nigel Toon and CTO Simon Knowles. Toon was formerly CEO of base-station-processor vendor Picochip, which Mindspeed acquired in 2014, and voice-processing vendor XMOS, where the IPU project started before spinning off in 2016. Graphcore has 116 employees and two dogs, most at its Bristol, U.K., headquarters plus a handful at its rapidly ex- panding Silicon Valley office.

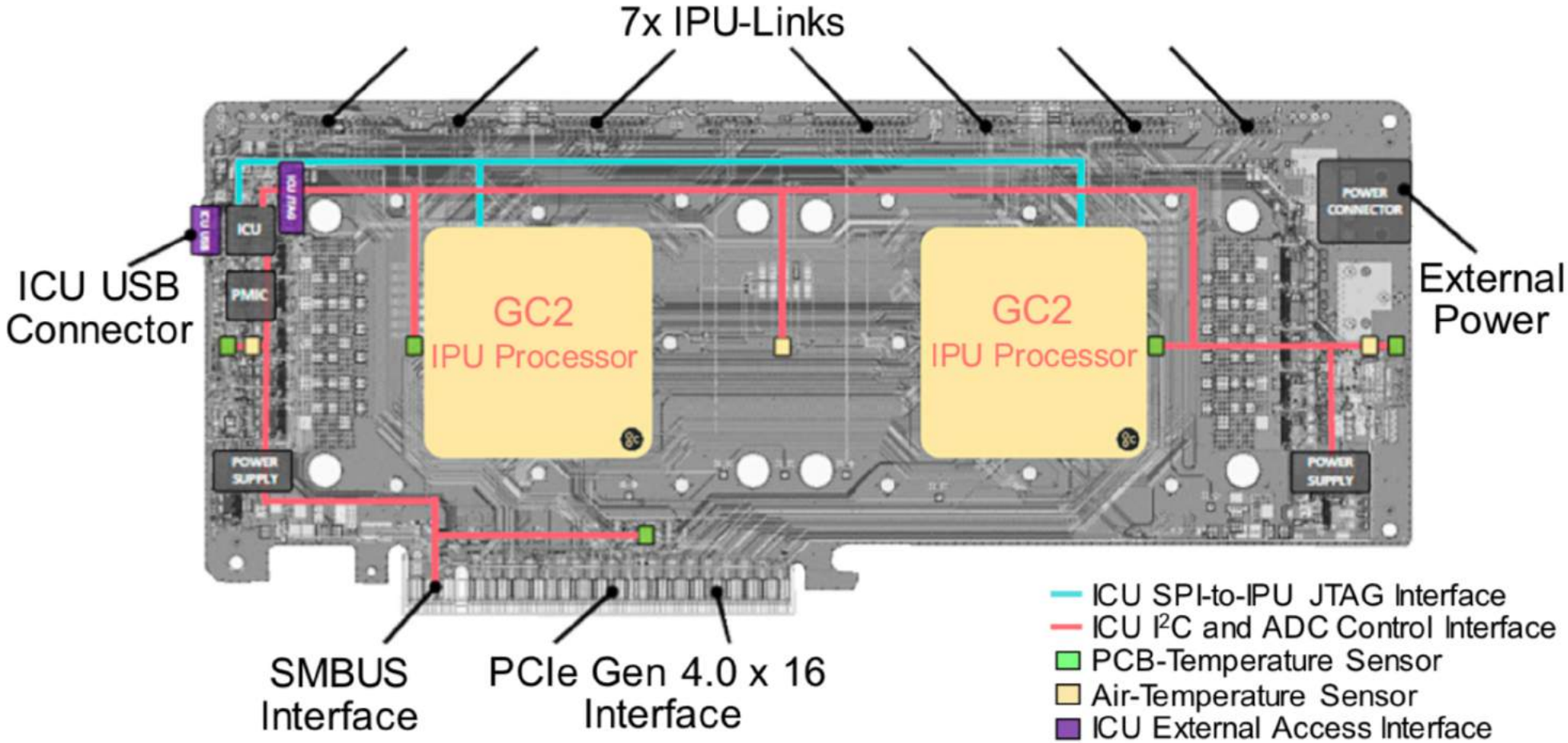


Figure 2. Graphcore C2 accelerator board. This PCI Express card combines two GC2 processor chips to deliver peak performance of 250Tflop/s in a 300W TDP. The IPU-Links can interconnect up to eight C2 boards into a single virtual accelerator. (Photo source: Graphcore)

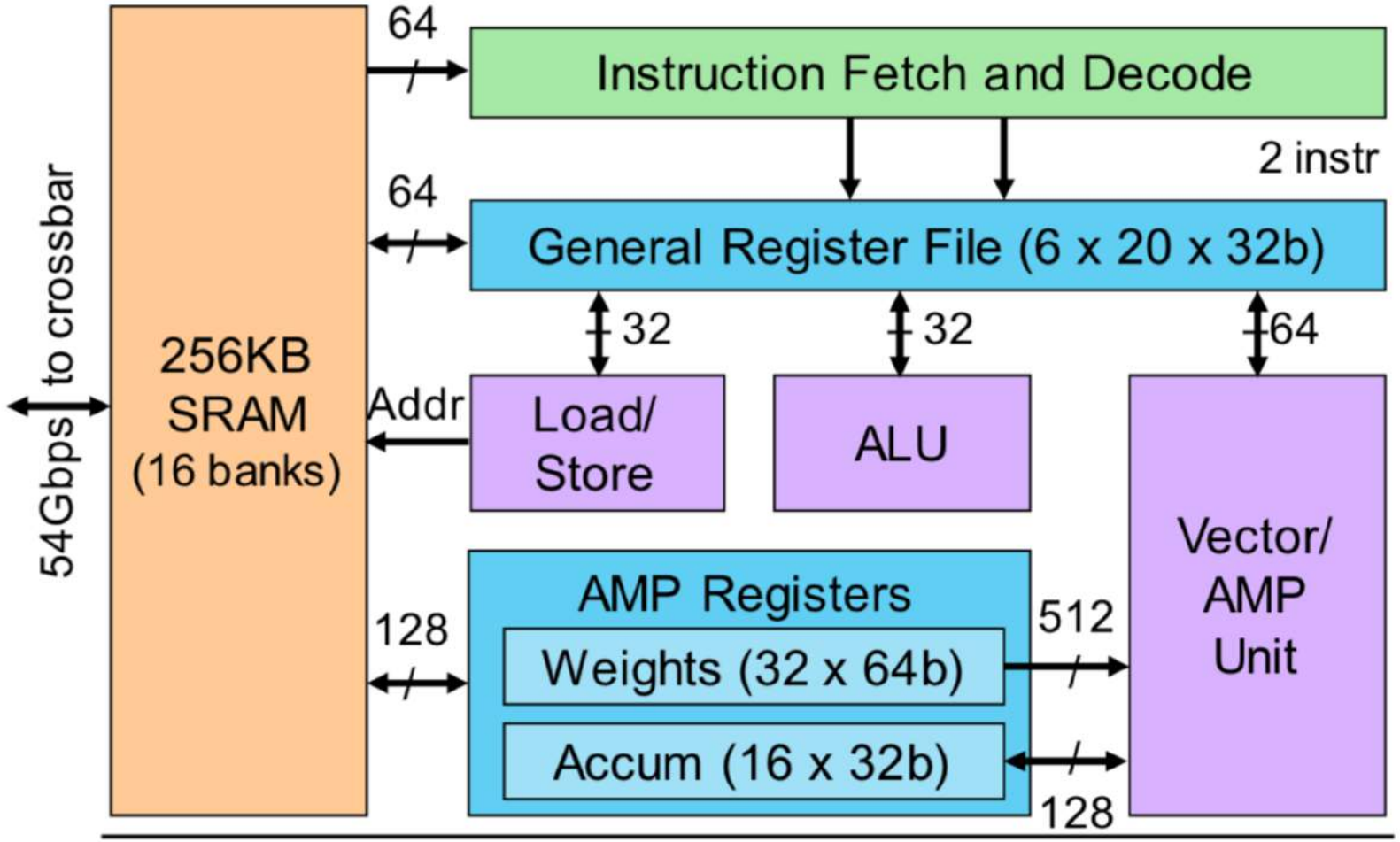


Figure 3. Simplified diagram of IPU core. This simple CPU executes two instructions per cycle, but the AMP unit performs matrix multiplication using values in the AMP registers. The general registers are replicated six times to support multithreading.

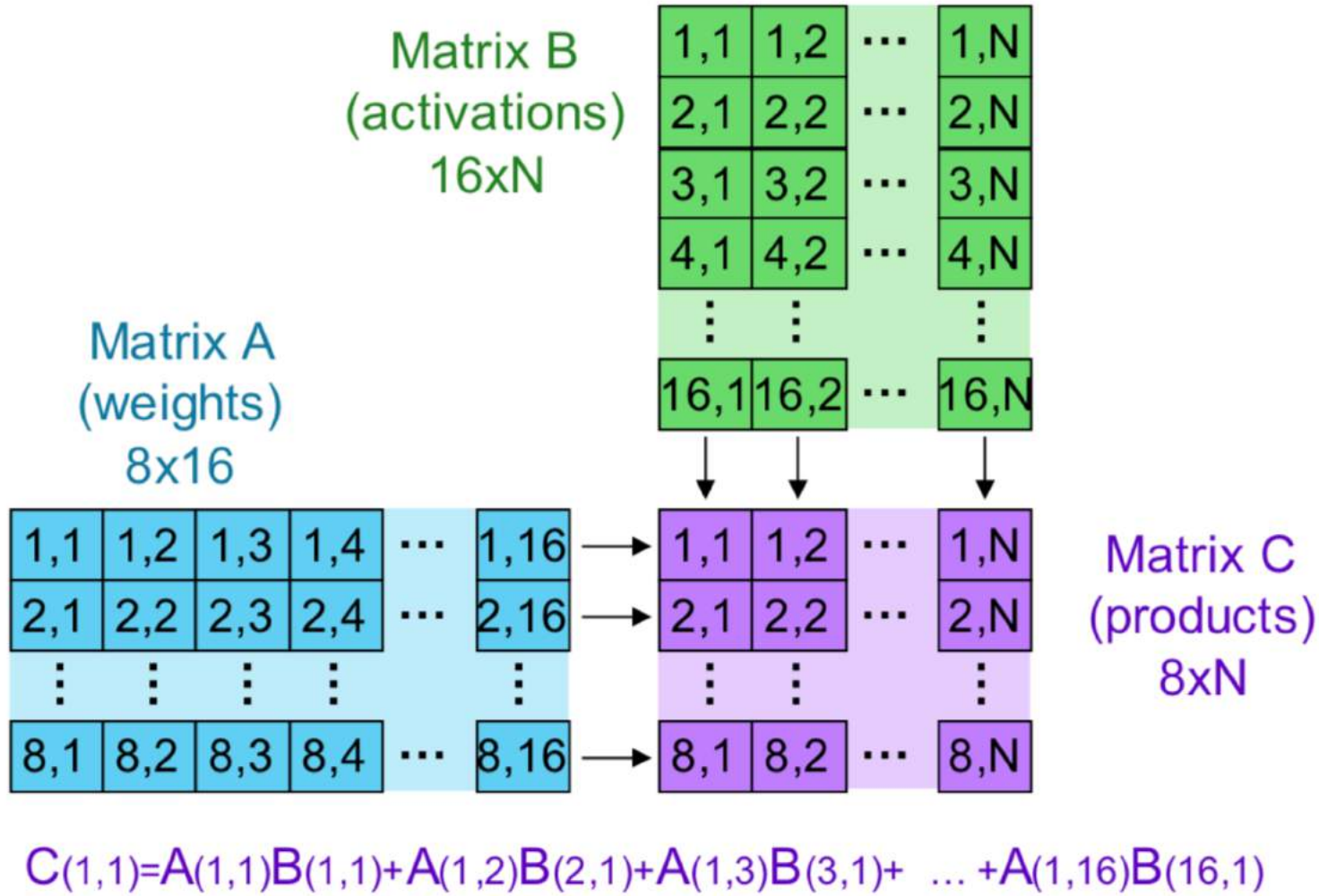


Figure 4. Matrix multiplication using the AMP unit. This unit multiplies an 8x16 matrix by a 16xN matrix to produce an 8xN matrix. Computation proceeds one column at a time, allowing arbitrary values of N.

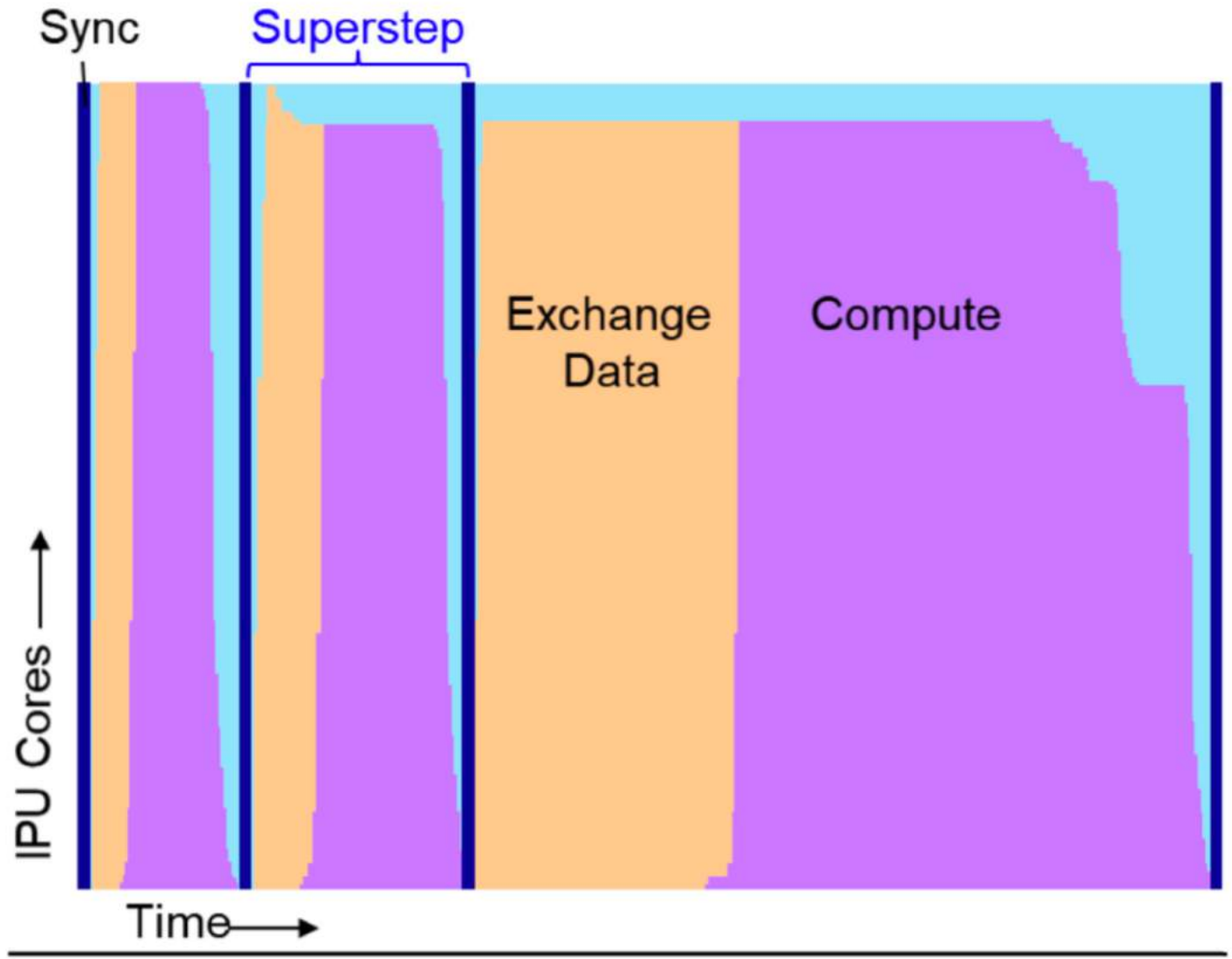


Figure 5. Timing of IPU workload. Each superstep comprises a brief synchronization period, an exchange of data and codelets among the cores, and a period when the cores execute the code until they finish. The light-blue areas show inactive cores.

Graphcore calls its software stack Poplar (after the birth-place of Tommy Flowers, creator of the original Colossus computer). It has already developed drivers for the popu-lar TensorFlow framework and the Open Neural Network Exchange (ONNX) format; it’s developing drivers for MXNet and Pytorch as well. Using this software, research- ers can easily convert existing models in these formats to run on the Graphcore processor. GC2 customers can also use the TensorFlow tools to develop new models that run on the chip.

	Graphcore C2	Nvidia Tesla V100	Google TPU1
Core Count	2x 1,216 cores	80 cores	1 core
Clock Speed (max)	1.6GHz	1.4GHz	0.7GHz
Peak Performance	125 TMAC/s	56 TMAC/s	46 TMAC/s
Data Format	FP16*	FP16*	INT8
FP32 Performance	31 TMAC/s	7 TMAC/s	Not supported
Chip RAM	2x 300MB	6MB	28MB
Board RAM	None	32GB HBM2	8GB DRAM
Host Interface	PCIe Gen4 x16	PCIe Gen3 x16	PCIe Gen3 x16
Board Power	300W	250W	50W†
Peak Perf/W	417 GM/W	224 GM/W	920 GM/W†
IC Process	16nm FF	12nm FF	28nm HKMG
Production	4Q18 (est)	4Q17	2Q15†

Table 1. Comparison of AI-accelerator cards. These PCI Express boards all target neural-network training. The Graphcore board includes two GC2 chips. *With FP32 accumulation. (Source: vendors, except †The Linley Group estimate)

Wave Computing

MICROPROCESSOR report
Insightful Analysis of Processor Technologies

Wave has spent the past two years revising its data-flow architecture. Having $117 million in funding from Dado Banatao's Tallwood Capital and others, the startup is sampling systems based on its DPU ASIC. On the basis of simulations, Wave claims a 64-DPU system could train GoogLeNet at 420,000 IPS, or about 6,500 IPS per chip, which would be 4x better than Nvidia's measured V100 performance.

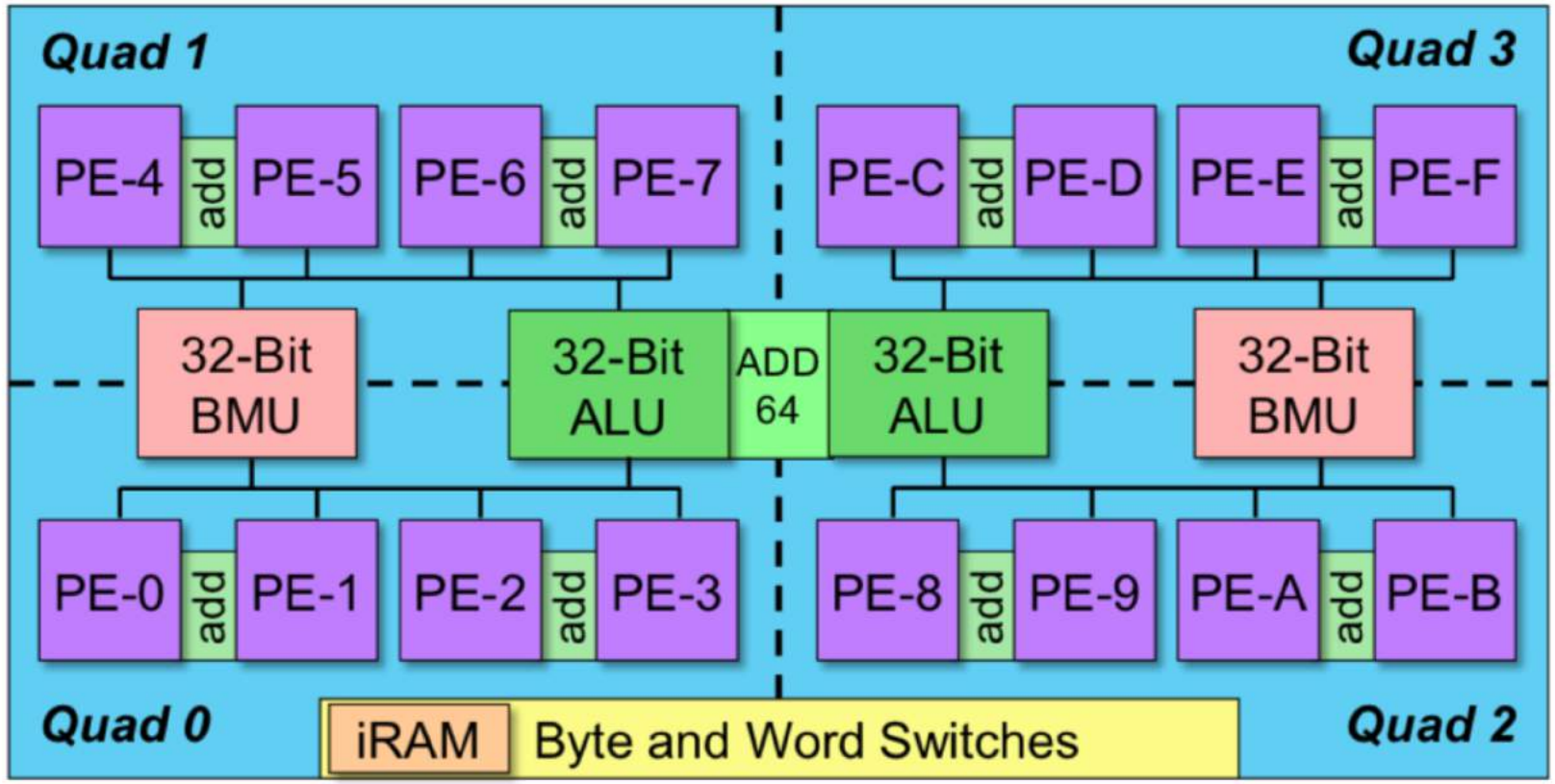


Figure 3. Wave DPU processing cluster. Each cluster contains 16 processing elements plus additional shared compute units. The switches connect to nearby clusters.

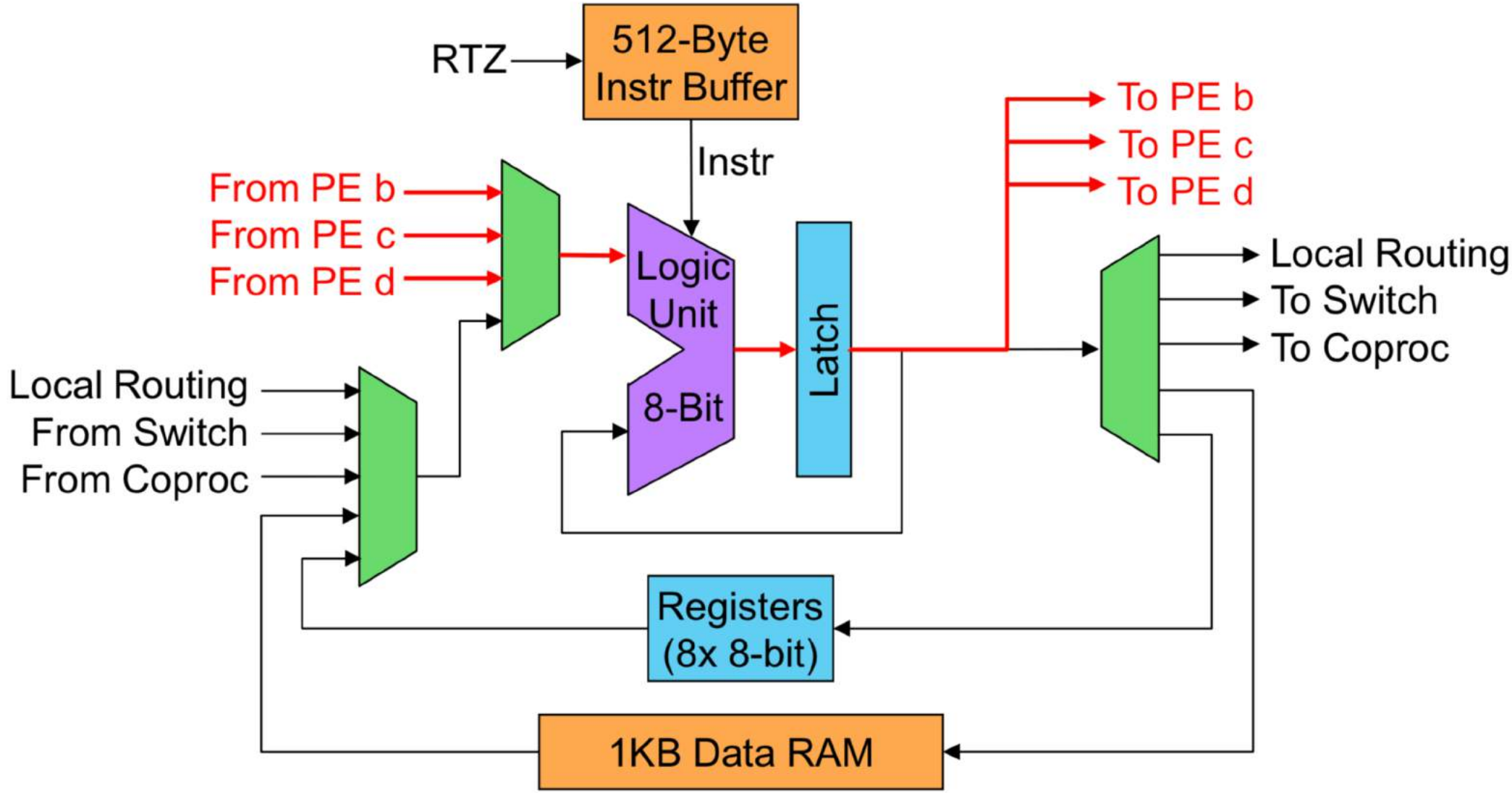


Figure 2. Wave DPU processing element. Each PE is an independent processor with its own instruction memory, data memory, registers, and compute unit. The critical timing path appears in red.

Per Cluster	Repeat Rate	8-Bit Ops Per Cluster	8-Bit Ops Per Chip	Type of Ops
16 PEs, 8 bits each	1	107 GOPS	110 TOPS	Shift, logical
8 adders, 8 bits each	1	54 GOPS	55 TOPS	Add, subtract
2 MAC units, 32 bits each	7	15 GOPS	16 TOPS	Multiply, add
2 BMUs, 32 bits each	3	18 GOPS	18 TOPS	Shift
4 addr units, 16 bits each	2	13 GOPS	13 TOPS	16b increment
Total		207 GOPS	212 TOPS	
Total – 15%		**176 GOPS**	**181 TOPS**	

Table 1. Wave DPU peak performance. Each of the 1,024 clusters can generate a peak rate of 207 billion 8-bit integer operations per second at 6.7GHz, but the company conservatively derates this number by 15% to account for internal constraints. (Source: Wave)

A single PE requires less than 0.01mm2 of die area—smaller than a Cortex-M0+ with a similar amount of memory.

The pipelined instruction buffer can deliver a new instruction every 0.1ns (10GHz), setting the machine's maximum speed.

The DPU chip contains 1,024 clusters tiled in a 32x32 ar- ray. The word switches allow any cluster to send data to any other cluster, routing the data through a series of hops until it reaches its destination. This type of mesh intercon- nect appears in many high-core-count processors, but few designers have attempted it with so many cores.

Cornami

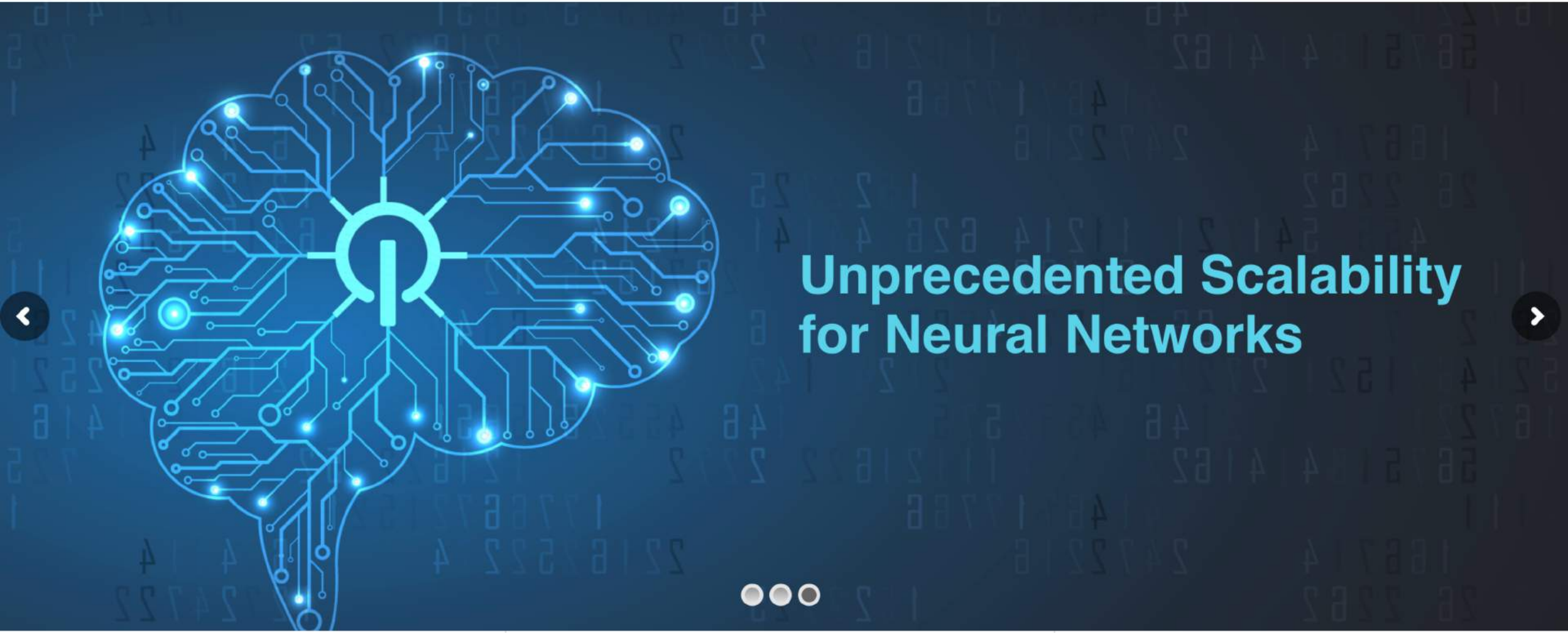
MICROPROCESSOR report
Insightful Analysis of Processor Technologies

Cornami is the latest startup from Gordon Campbell, who also founded both Seeq and Chips & Technologies. At the recent Linley Processor Conference, it disclosed ambi- tious plans to deliver a processor with 100x the performance of Nvidia's Titan V (Volta) card while using only 30W. The company has developed a unique architecture that can cre- ate dynamically reconfigurable systolic arrays using more than 10,000 independent cores, but it withheld further de- tails, including its funding and schedule. Cornami is already testing initial silicon, and we expect its first product in 1H19.

CORNAMI™

Intelligent Computing

Technology & Products | Solutions | News & Events | Blog | About Us

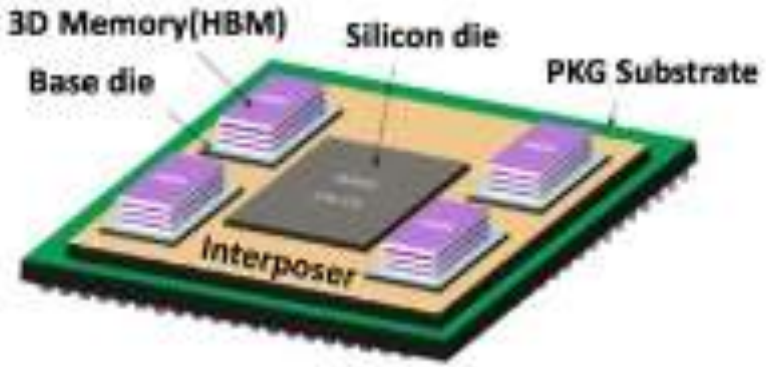


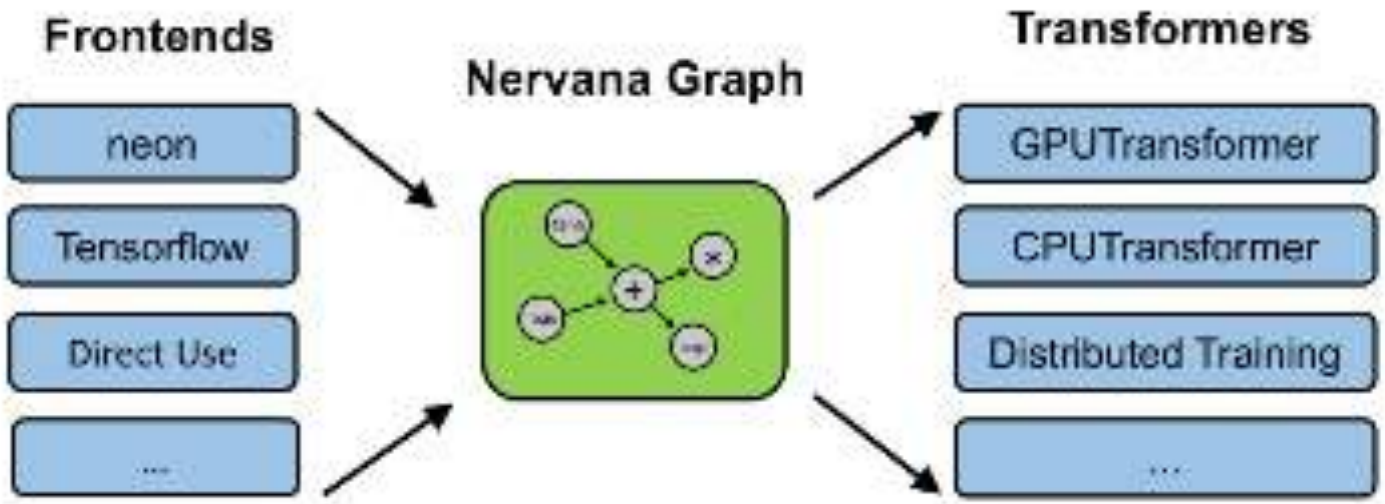
Usage Model for Dense Computational Fabric

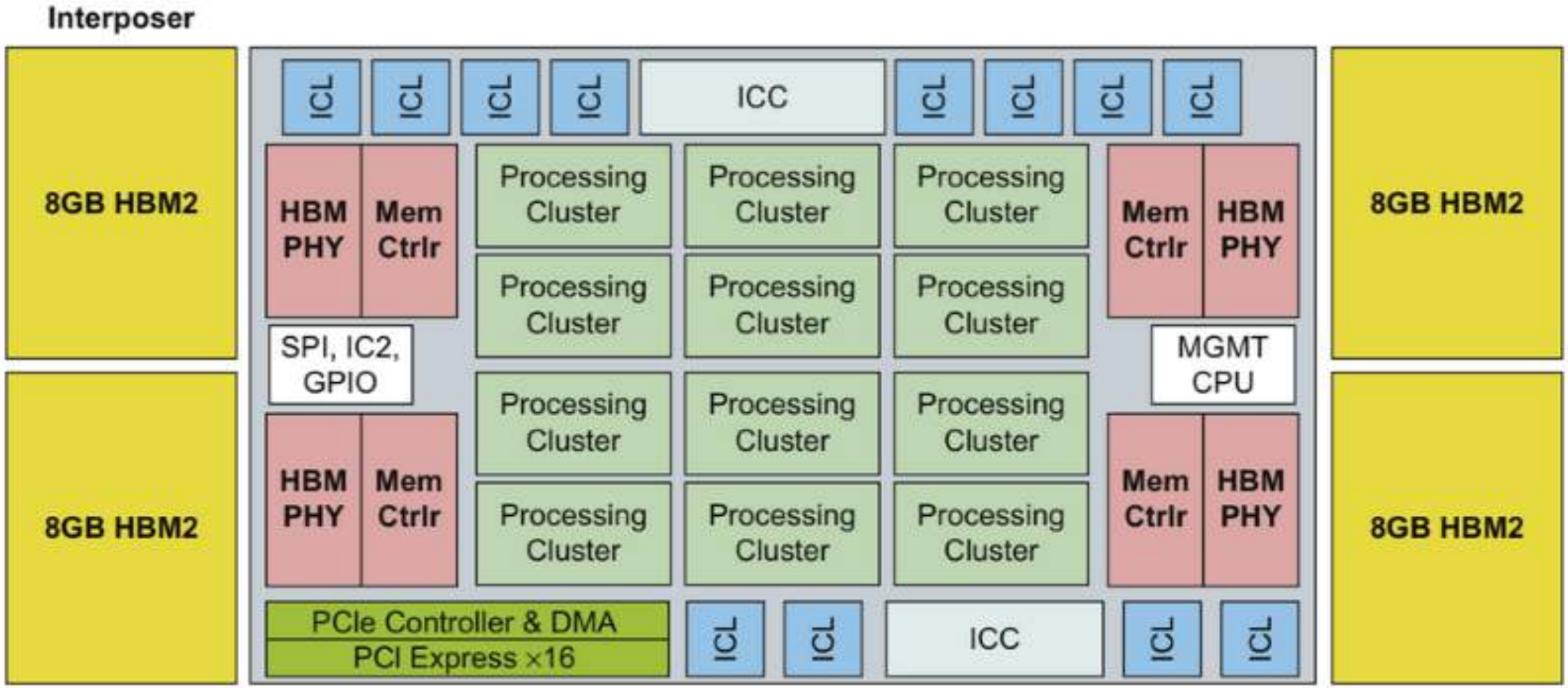
Intel Nervana

Two years ago, Intel paid about $350 million for hot AI startup Nervana Systems. But Intel decided to change the startup's original design, in part to remove its unique Flex-Point technology. The revised design, code-named Lake Crest, delivers up to 20TMAC/s at 210W, about a third of the V100's performance. After a poor customer response, the company decided not to bring Lake Crest to production. Instead, it's developing Spring Crest for production in late 2019, aiming for a 3–4x performance gain over Lake Crest.

Lake Crest


3D Memory(HBM)
Silicon die
Base die
PKG Substrate
Interposer


Frontends
neon
Tensorflow
Direct Use
...
Nervana Graph
Transformers
GPUTransformer
CPUTransformer
Distributed Training
...


Interposer
8GB HBM2
8GB HBM2
8GB HBM2
8GB HBM2
ICL
ICL
ICL
ICL
ICC
ICL
ICL
ICL
ICL
HBM PHY
Mem Ctrlr
Processing Cluster
Processing Cluster
Processing Cluster
Mem Ctrlr
HBM PHY
Processing Cluster
Processing Cluster
Processing Cluster
SPI, IC2, GPIO
MGMT CPU
Processing Cluster
Processing Cluster
Processing Cluster
HBM PHY
Mem Ctrlr
Processing Cluster
Processing Cluster
Processing Cluster
Mem Ctrlr
HBM PHY
PCIe Controller & DMA
PCI Express ×16
ICL
ICL
ICC
ICL
ICL

INTEL® NERVANA™ NNP L-1000

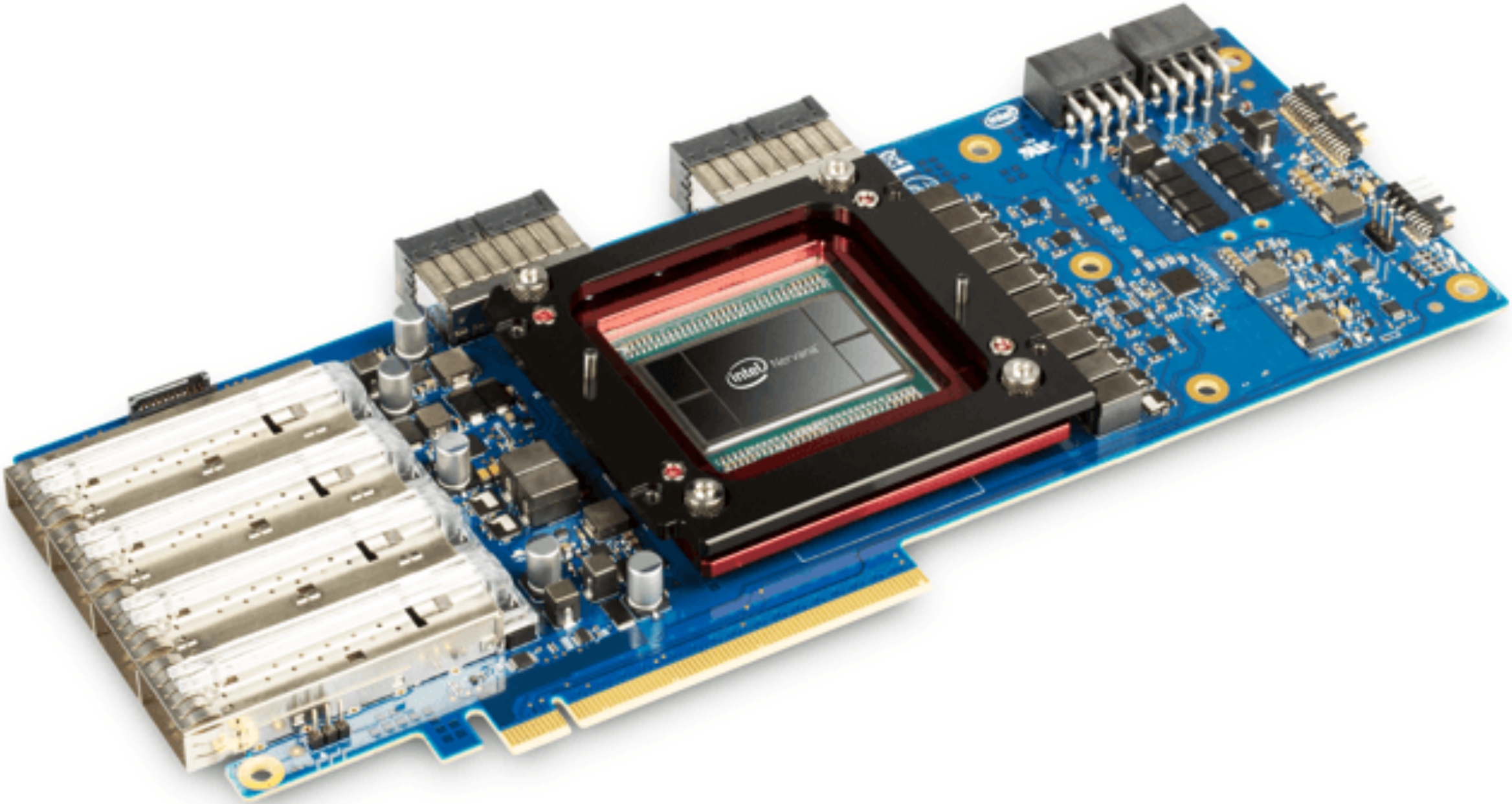
PURPOSE-BUILT FOR REAL WORLD AI PERFORMANCE

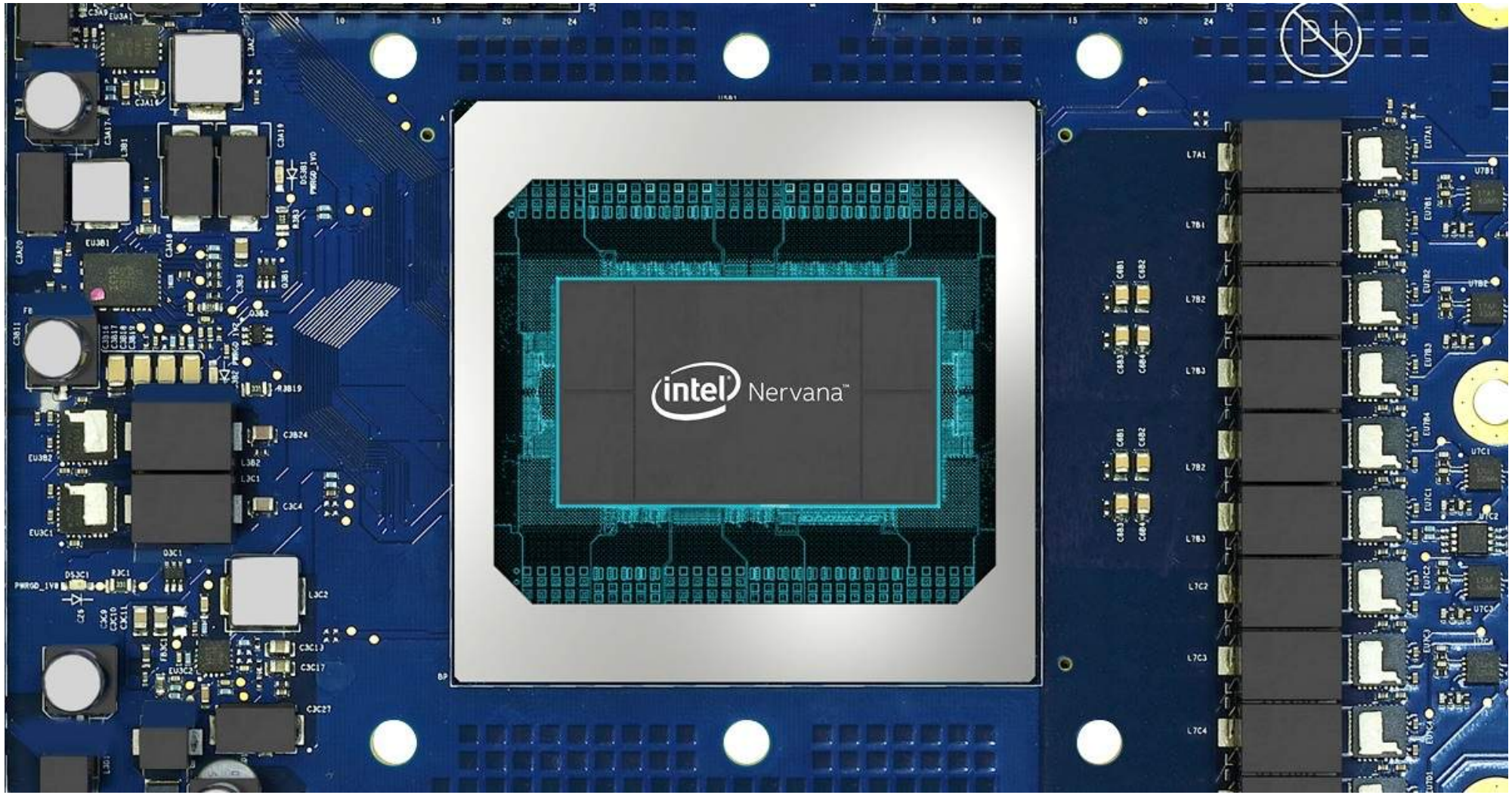
Optimized across memory, bandwidth, utilization and power

3-4x training performance of first-generation NNP product

High-bandwidth, low-latency interconnects

bfloat16 numerics


intel
Nervana


intel
Nervana™

Xilinx Alveo

Xilinx views the data center as a large new market for its FPGAs. To simplify deployment, it announced Alveo accelerator cards that hold a high-end FPGA and 16GB of DDR4 DRAM and require 225W. Xilinx preprograms the FPGA to support deep neural networks developed in Caffe, MXNet, and TensorFlow. In a server with eight Alveo cards, it ran GoogLeNet inference for batch size of 1 at 30,000 IPS, or about 4,000 IPS per card; a single V100 delivers one-third the performance but costs half as much. Alveo sets the stage for the company's next-generation products, which include dedicated AI engines.

XILINX
ALVEO™

Adaptable Accelerator Cards for
Data Center Workloads

Breathe New Life Into Your Data Center

Available Now > Media Kit >

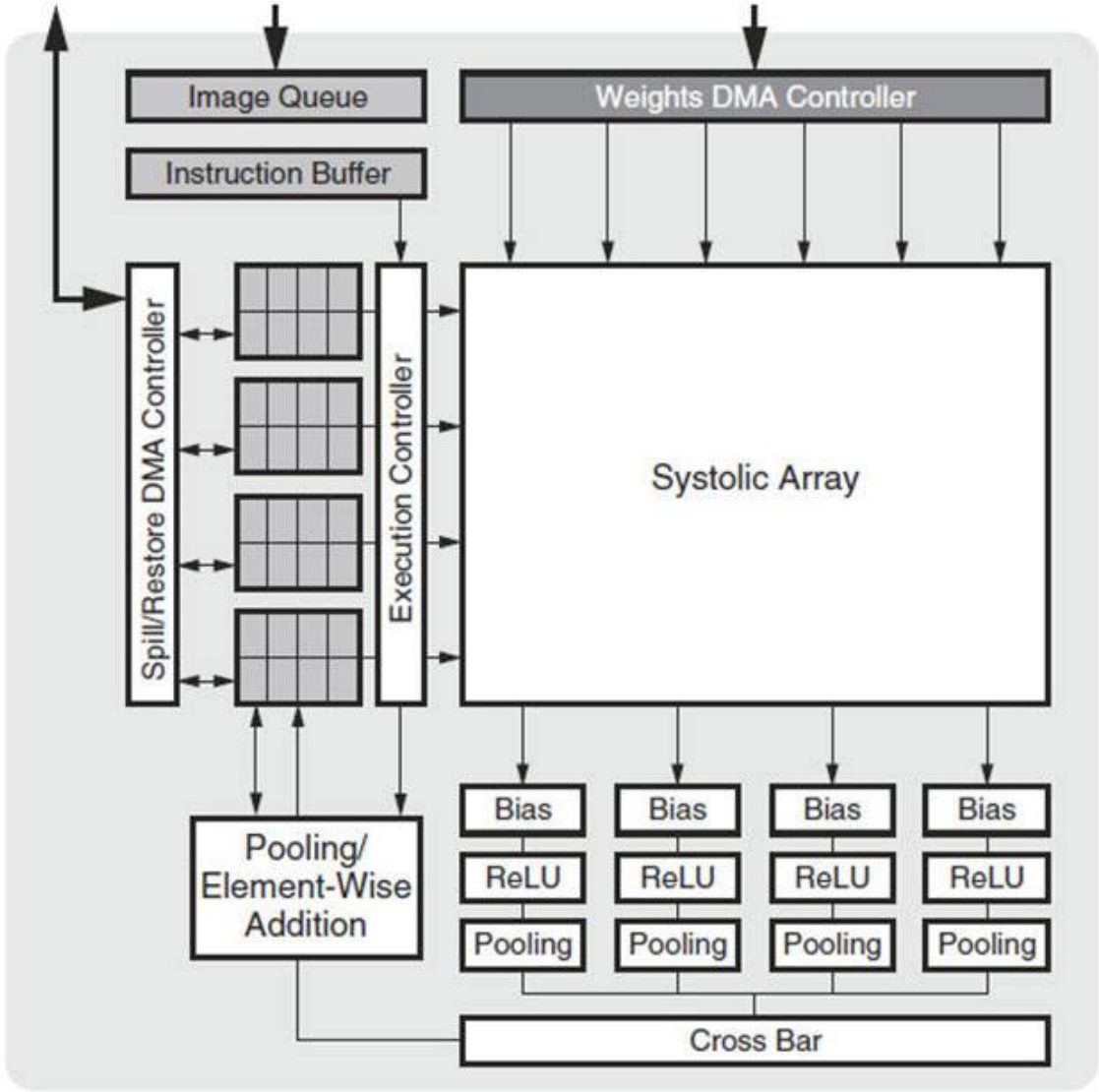
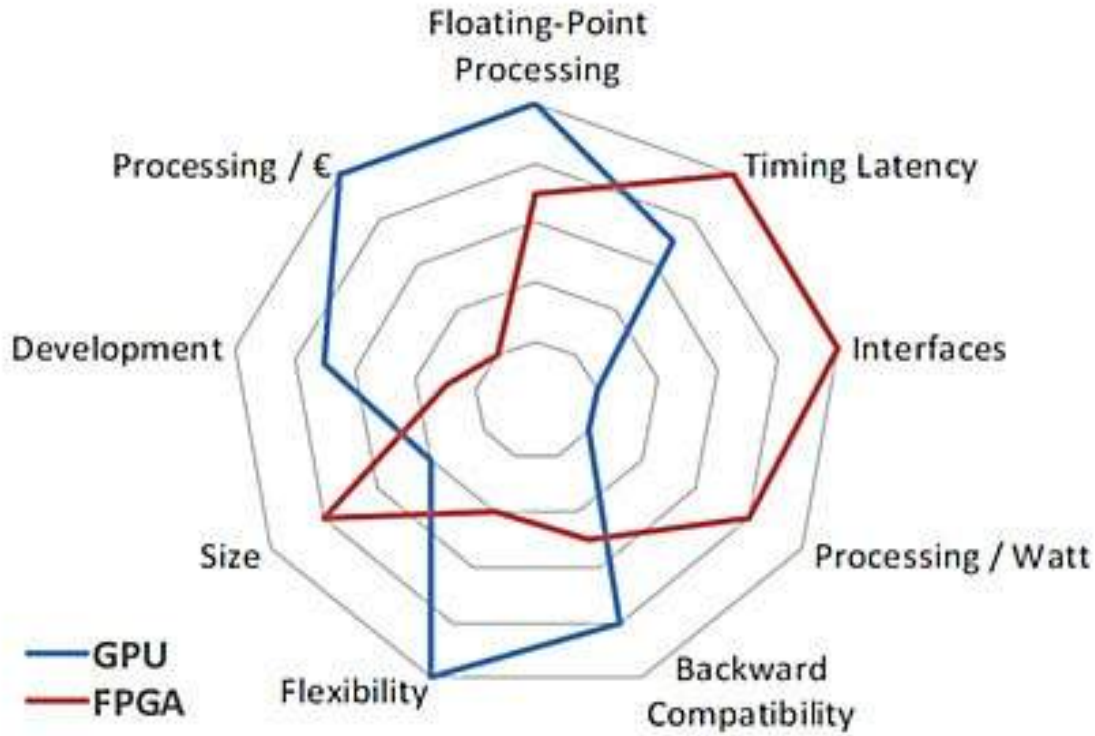
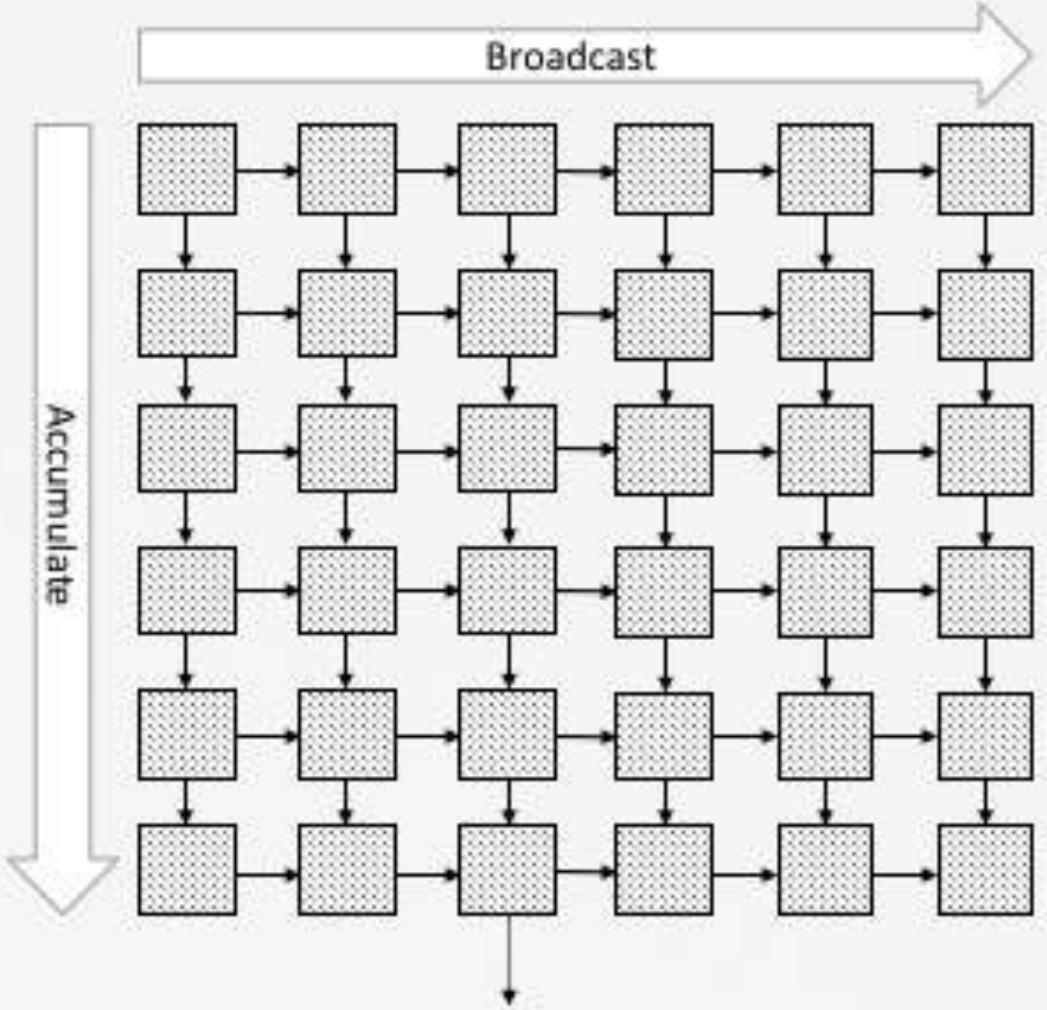


Image Queue
Weights DMA Controller
Instruction Buffer
Spill/Restore DMA Controller
Execution Controller
Systolic Array
Bias
ReLU
Pooling
Pooling/
Element-Wise
Addition
Cross Bar

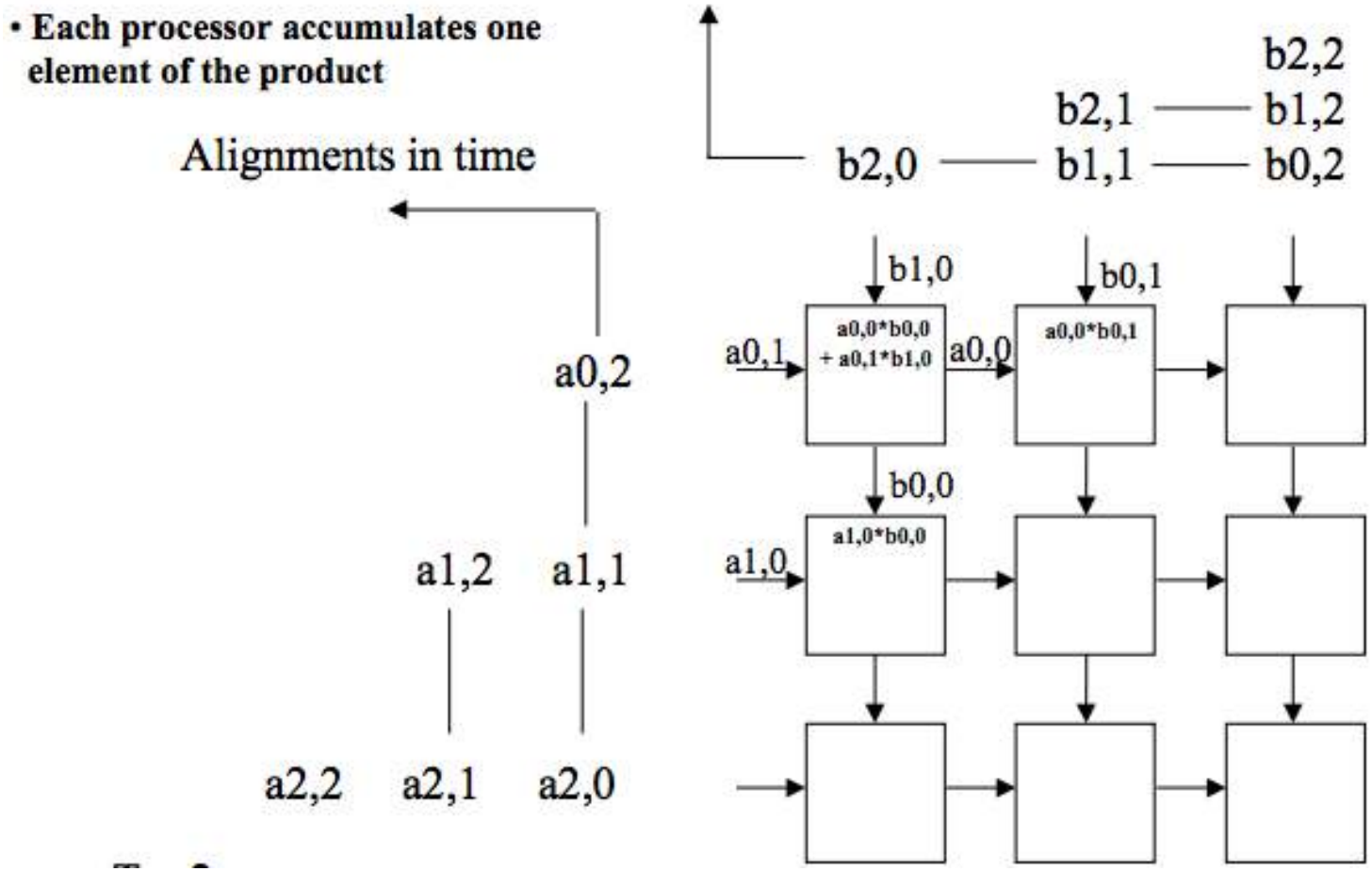

Floating-Point
Processing
Timing Latency
Interfaces
Processing / Watt
Backward
Compatibility
Flexibility
Size
Development
Processing / €
GPU
FPGA

A Systolic Array



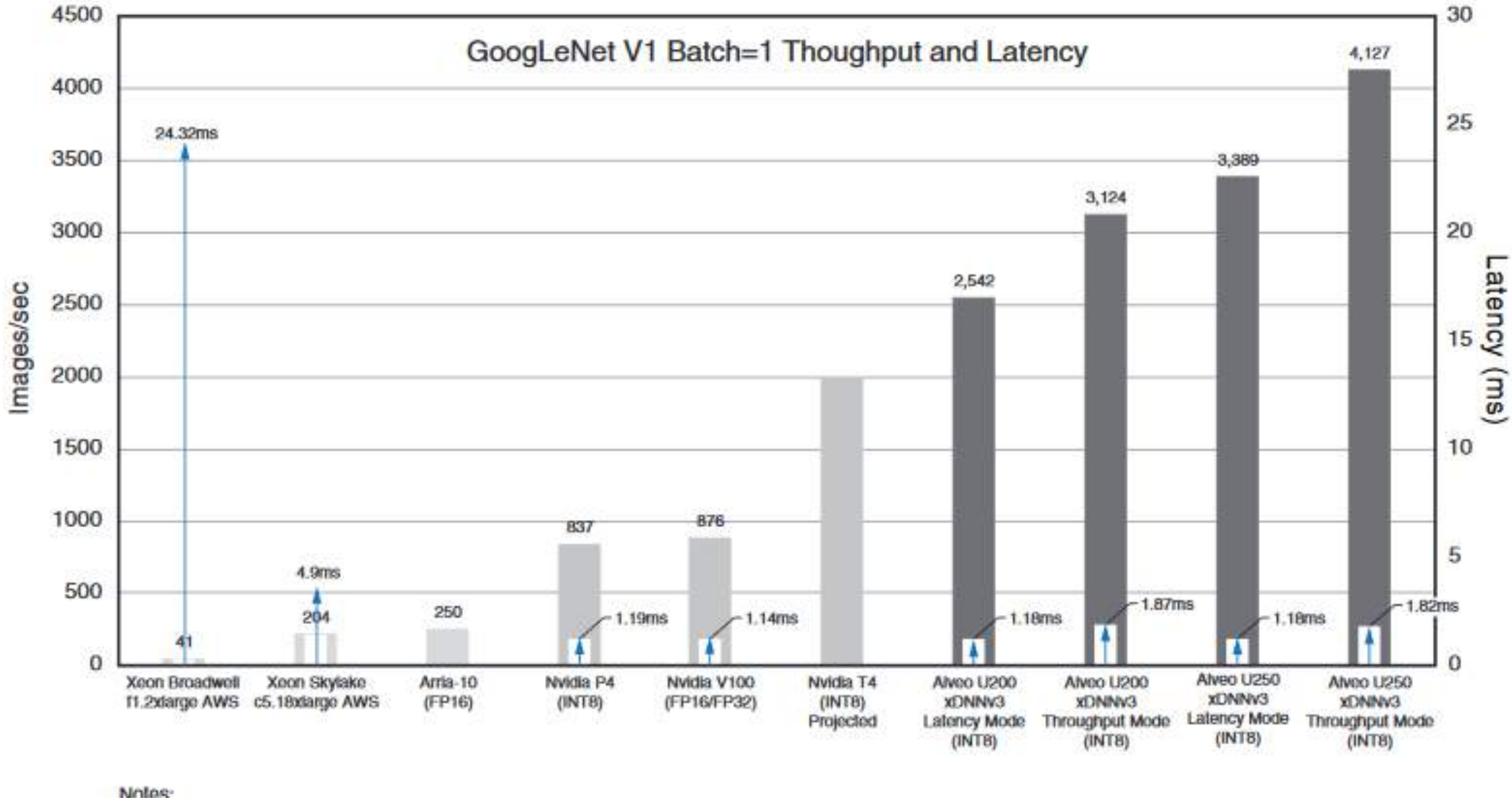
Systolic Array Example:
3x3 Systolic Array Matrix Multiplication

- **Processors arranged in a 2-D grid**
- **Each processor accumulates one element of the product**



$$y_1 = W11\,X1 + W12\,X2$$

$$y_2 = W21\,X1$$


GoogLeNet V1 Batch=1 Thoughput and Latency
Images/sec
Latency (ms)
4500
4000
3500
3000
2500
2000
1500
1000
500
0
30
25
20
15
10
5
0
24.32ms
41
4.9ms
204
250
837
1.19ms
876
1.14ms
2,542
1.18ms
3,124
1.87ms
3,389
1.18ms
4,127
1.82ms
Xeon Broadwell f1.2xdarge AWS
Xeon Skylake c5.18xdarge AWS
Arria-10 (FP16)
Nvidia P4 (INT8)
Nvidia V100 (FP16/FP32)
Nvidia T4 (INT8) Projected
Alveo U200 xDNNv3 Latency Mode (INT8)
Alveo U200 xDNNv3 Throughput Mode (INT8)
Alveo U250 xDNNv3 Latency Mode (INT8)
Alveo U250 xDNNv3 Throughput Mode (INT8)

Notes:

GoogLeNet Batch = 1 Performance, Sub-2 ms Latency							
	Nvidia	**Nvidia**	**Nvidia**	**Xilinx**	**Xilinx**	**Xilinx**	**Xilinx**
	Tesla	**Tesla**	**Tesla**	**Alveo**	**Alveo**	**Alveo**	**Alveo**
	P4	**V100**	**T4**	**U250**	**U250+DeePhi**	**AI Core**	**AI Core+DeePhi**
Performance, Images/Sec	1,215	2,716	3,500	4,127	5,365	22,500	29,250
Watts	75	250	75	225	225	275	275
List Price	$2,500	$11,500	***$3,000***	$12,995	$12,995	***$15,000***	***$15,000***
Cost Per Image, 3 Year 100% Utilization	$0.000001565	$0.000003220	$0.000000652	$0.000002395	$0.000001842	$0.000000507	$0.000000390
3 Year Images Processed, Billions	1.60	3.57	4.60	5.43	7.05	29.59	38.46
Performance/Watt	16.20	10.86	46.67	18.34	23.84	81.82	106.36
Price/Performance	$2.06	$4.23	$0.86	$3.15	$2.42	$0.67	$0.51
Price/Performance/Watt	$0.0274	$0.0169	$0.0114	$0.0140	$0.0108	$0.0024	$0.0019