



IC Technology –
**What Will the Next
Node Offer Us?**

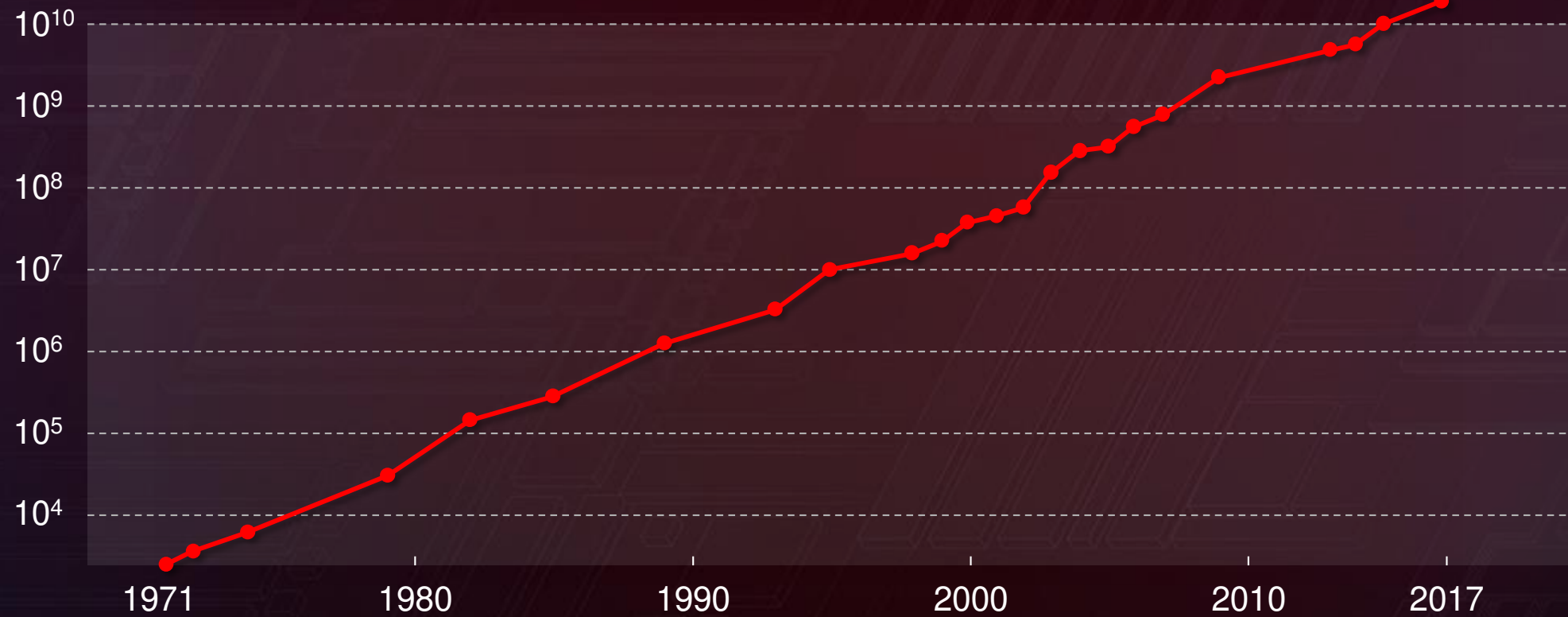
H.-S. Philip Wong

Vice President, Corporate Research, TSMC

Willard R. & Inez Kerr Bell Professor, Stanford University

MOORE'S LAW

Transistors per microprocessor

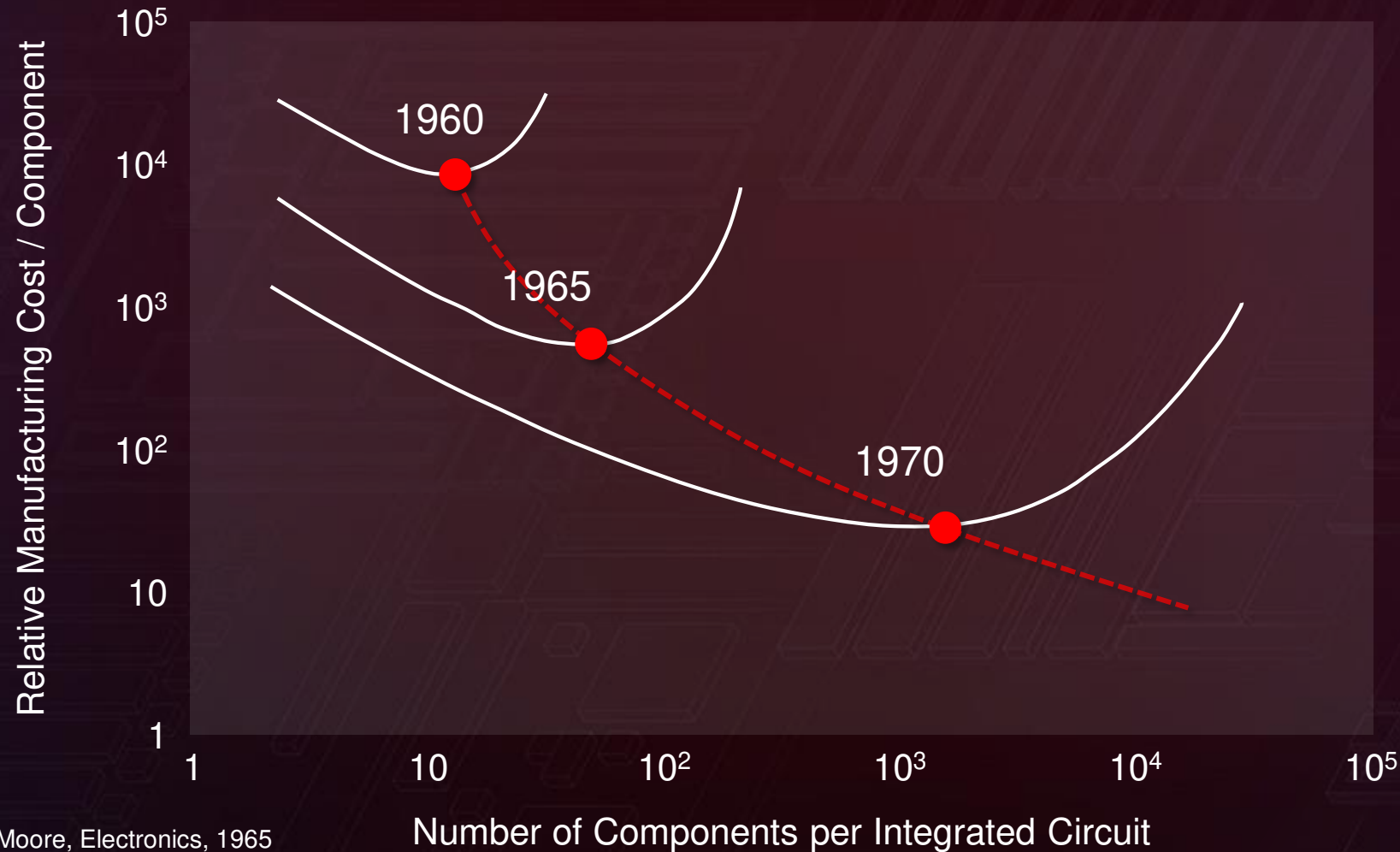


Source: Karl Rupp. 40 Years of Microprocessor Trend Data.



MOORE'S LAW

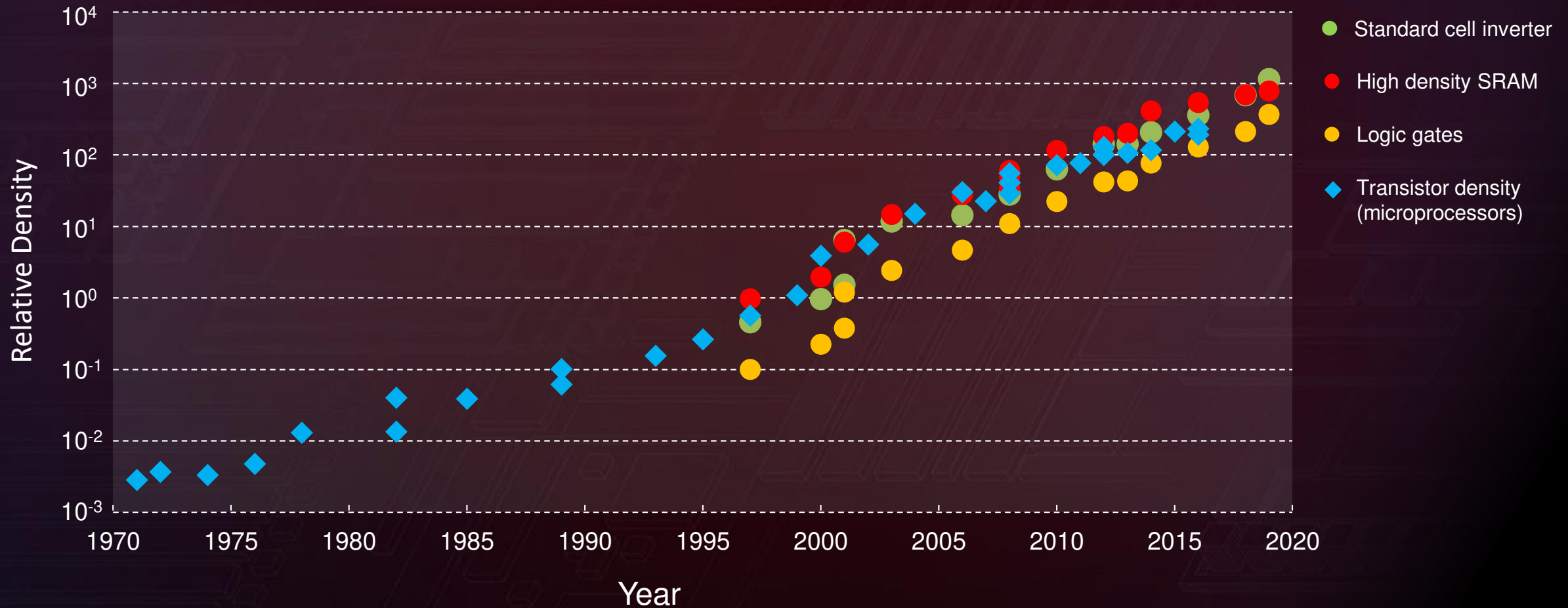
DENSITY AND COST PER FUNCTION



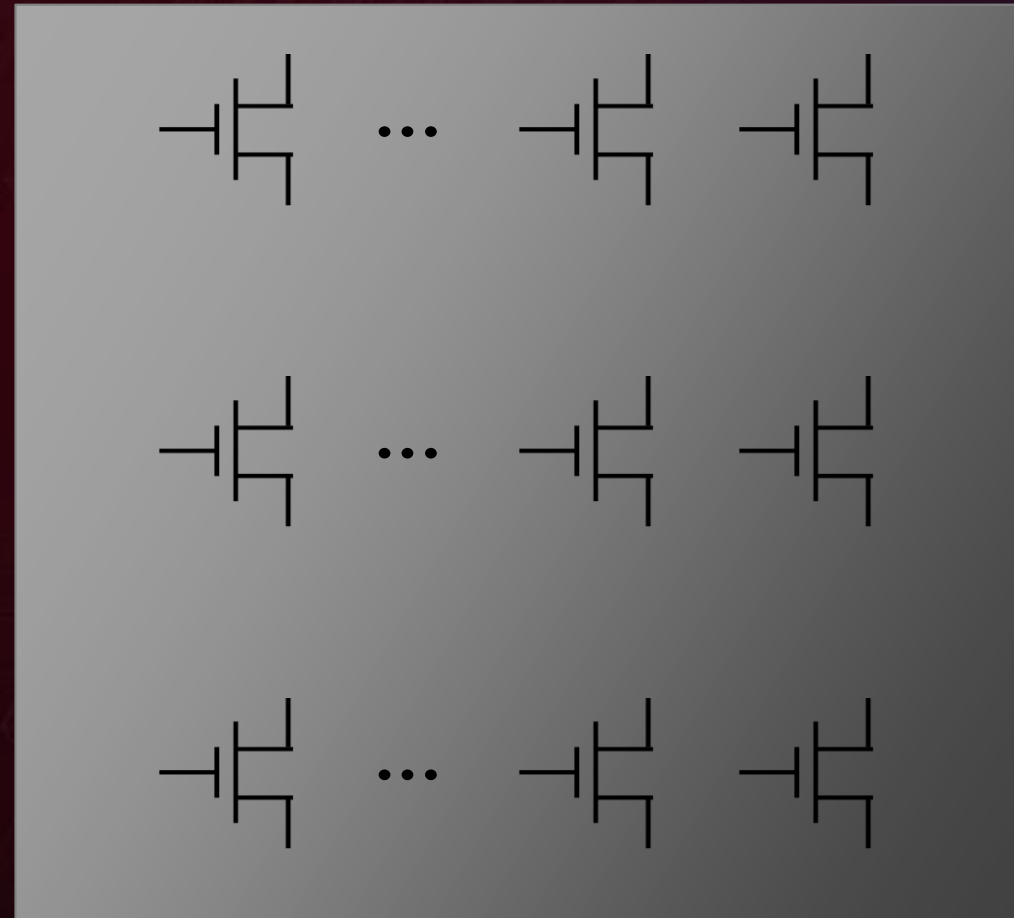
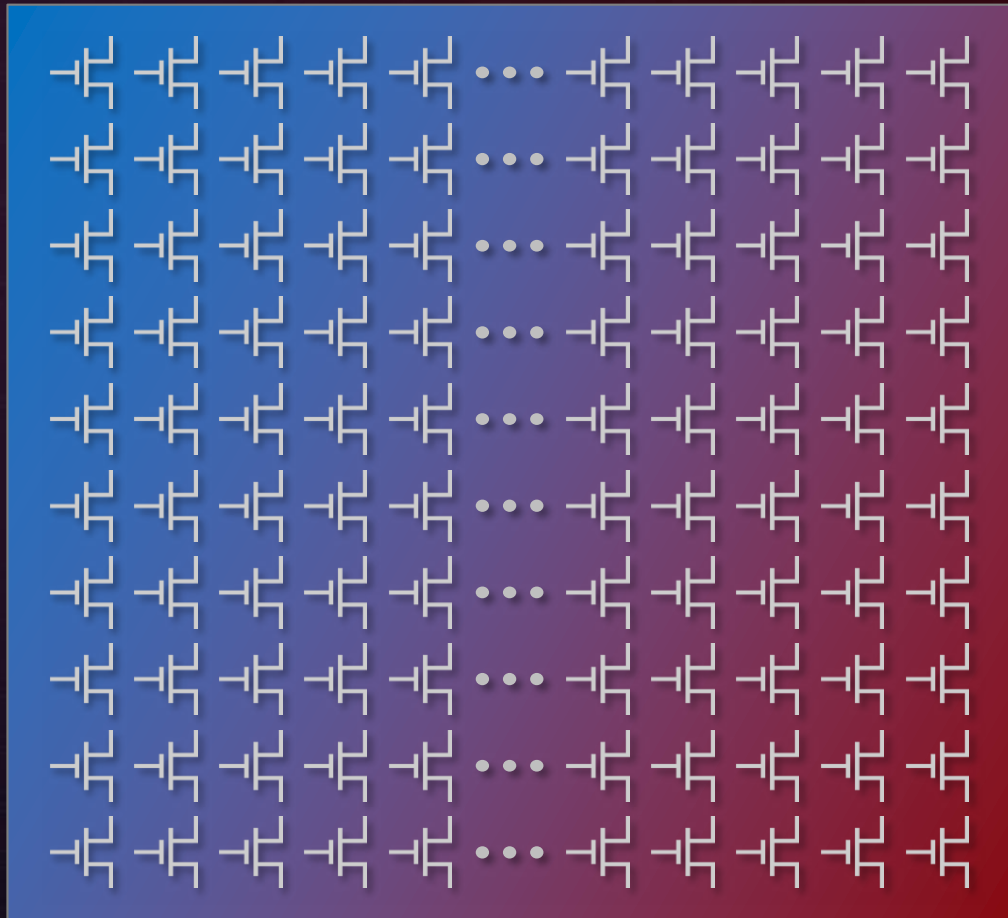
Source: G. Moore, Electronics, 1965

MOORE'S LAW IS WELL AND ALIVE

DENSITY: A NECESSARY ATTRIBUTE



IMAGINE: TRANSISTOR PERFORMANCE W/O DENSITY



IMAGINE:

TRANSISTOR PERFORMANCE W/O DENSITY

- Not enough memory
- No multi-core chips
- No accelerators
- Wire delay slows big chips.

TECHNOLOGY **LEADERSHIP**

N7

World's first 7 nm

Participated in all the products on 7 nm

TECHNOLOGY LEADERSHIP

N7

N5 (P)

Best performance

Highest density

Extensive EUV layers

Design ecosystem ready

In risk production

TECHNOLOGY LEADERSHIP

N7

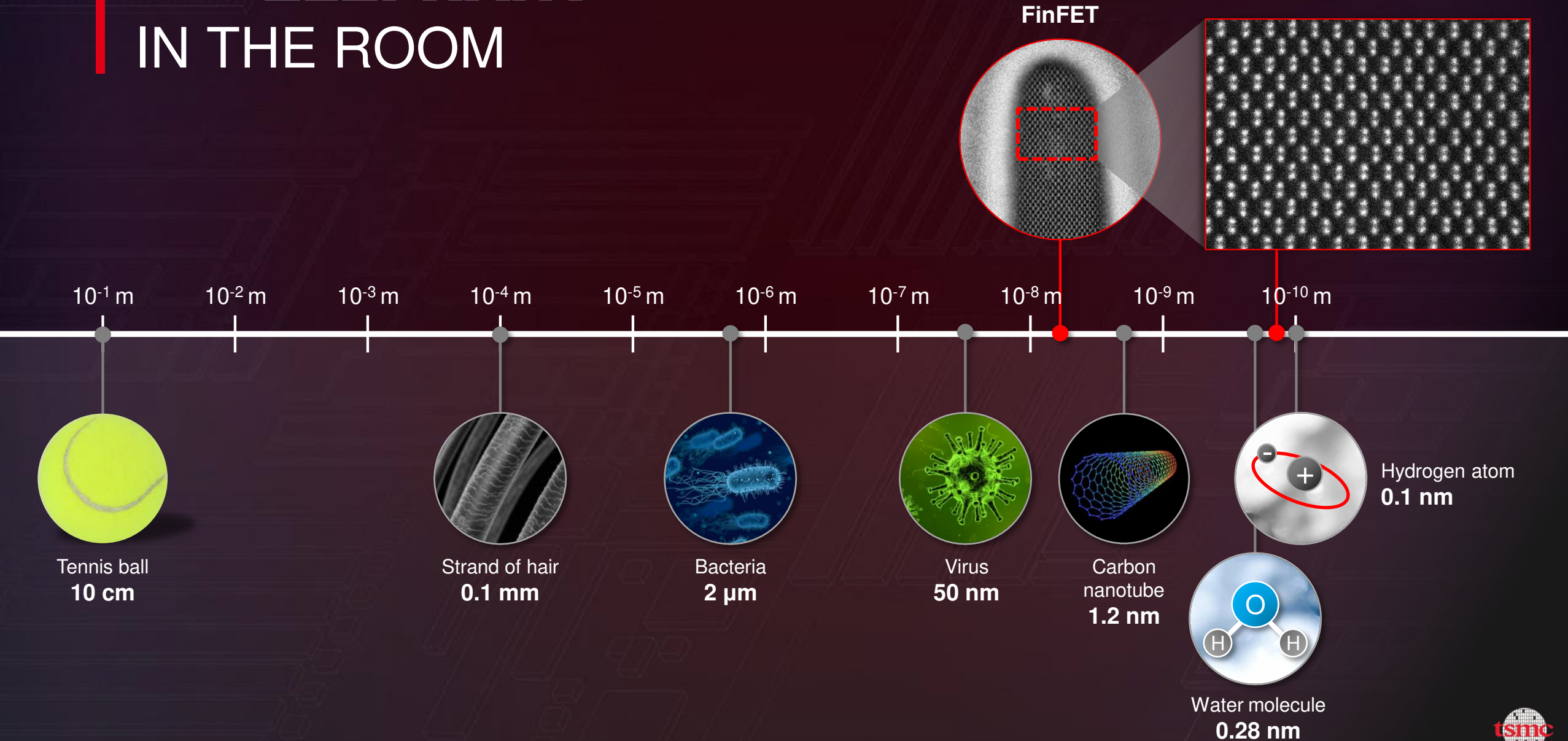
N5 (P)

N3

...

2050 and beyond

THE ELEPHANT IN THE ROOM



MOORE'S LAW – A HISTORY OF INNOVATIONS

Dennard scaling

Strained Si, high-k / metal gate

FinFET / DTCO

**CONTINUOUS
BENEFITS
NODE AFTER NODE**

MULTIPLE ROADS LEAD TO ROME

Innovations

**CONTINUOUS
BENEFITS
NODE AFTER NODE**

INTEGRATING CHIPS INTO SYSTEMS

Cramming More Components onto Integrated Circuits

GORDON E. MOORE, LIFE FELLOW, IEEE

With one cost falling as the number of components per circuit rises, by 1975 economists may determine components on a single silicon

The future of integrated electronics itself. The advance about a proliferation of electronic many new areas.

Integrated circuits will be computers—or at least their counterparts—economic control, social portable communication, research needs only a day. But the biggest potential of systems in telephone communications in digital filters will separate. Integrated circuits will and perform data processing. Computers will be gone to in completely different ways of integrated electronics as the machine instead of heat. In addition, the improvement by integrated circuits will be processing units. Machines today will be built at lower cost.

I. PRESENT AND FUTURE

By integrated electronics, analogues which are refined as well as any additional functions required by the use technologies were first wave object was to minimize the increasingly complex electronic with minimum weight. Several successively techniques thin-film structures, and semiconductors.

Reprinted from Gordon E. Moore, "Integrated Circuits," *Electronics*, pp. 10-12, December 1958.

11

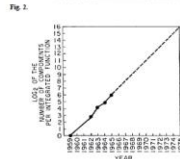
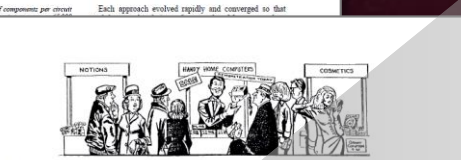


Fig. 2. Diagram to technological realization without any special

Fig. 3. It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected. The availability of large functions combined with functional design and construction, should allow the manufacturer of large systems to design and construct a considerable variety of equipment both rapidly and economically.

IX. LINEAR CIRCUITS

Integration will not change linear systems as radically as digital systems. Still, a considerable degree of integration will be achieved with linear circuits. The lack of integrative properties and inductors in the present fundamental limitations to integrated electronics in the linear area. By their very nature, such elements require the storage of energy in a volume. For high Q it is necessary that the volume be large. The incompatibility of large volume and integrated electronics is obvious from the term inductive. Certain resonance phenomena, such as those in piezoelectric crystals, can be expected to have some applications for timing functions, but inductors and capacitors will be used in for some time. The integrated RF amplifier of the future might well consist of integrated stages of gain, giving high performance at minimum cost, interspersed with relatively large tuning elements. Other linear functions will be changed considerably. The matching and tracking of similar component in integrated structures will allow the design of differential amplifiers of greatly improved performance. The use of thermal feedback effects to stabilize integrated structures to a small fraction of a degree will allow the construction of oscillators with crystal stability.

X. THE MICROAREA

Even as the microarea, structures included in the definition of integrated electronics will become increasingly important. The ability to make and assemble components small compared with the wavelengths involved will allow the use of lumped parameter design, at least at the lower frequencies. It is difficult to predict at the present time just how extensive the invasion of the microarea by integrated electronics will be. The successful realization of such items as phased-array antennas, for example, using a multiplicity of integrated microwave power sources, could completely revolutionize radar.

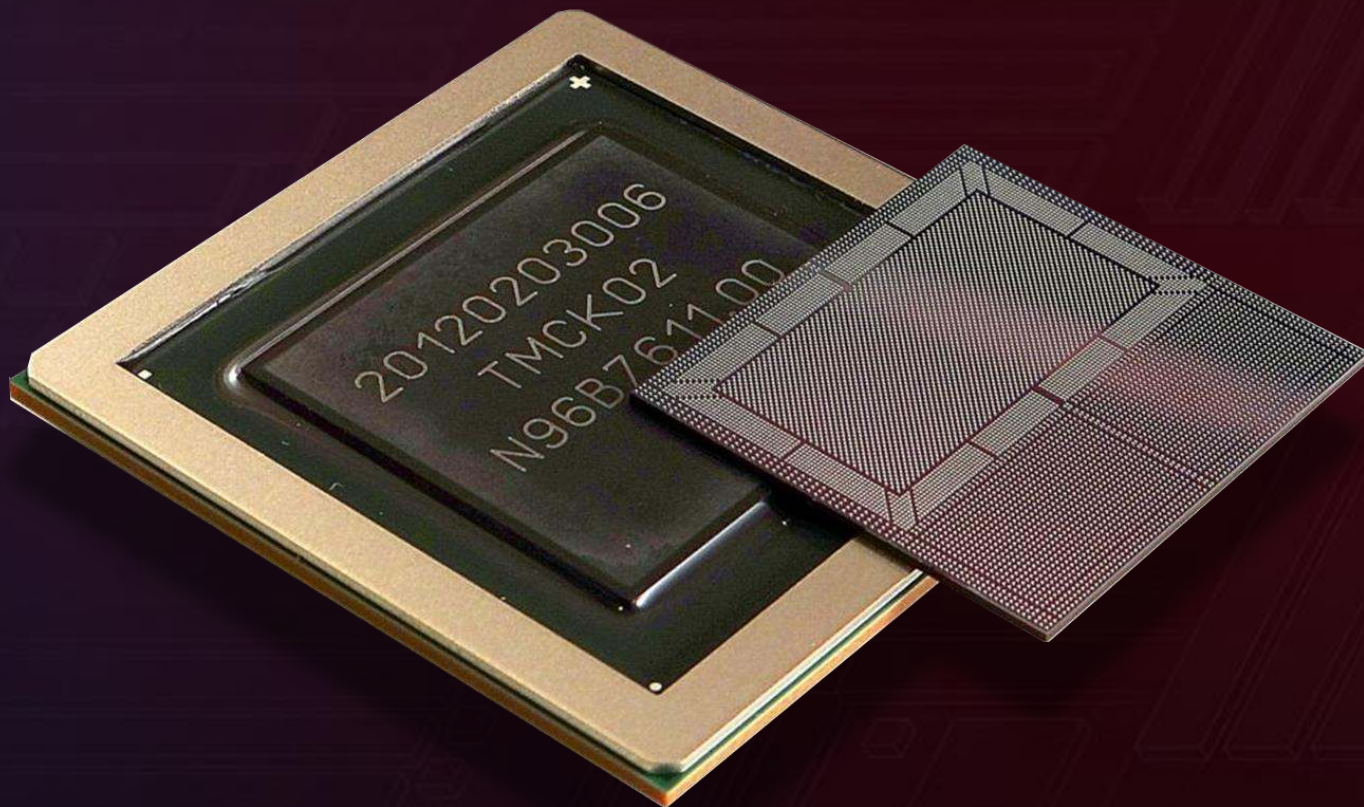
Proceedings of the IEEE, Vol. 56, No. 1, January 1968

It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected. The availability of large functions, combined with functional design and construction, should allow the manufacturer of large systems to design and construct a considerable variety of equipment both rapidly and economically.

Source: G. Moore, *Electronics*, 1965



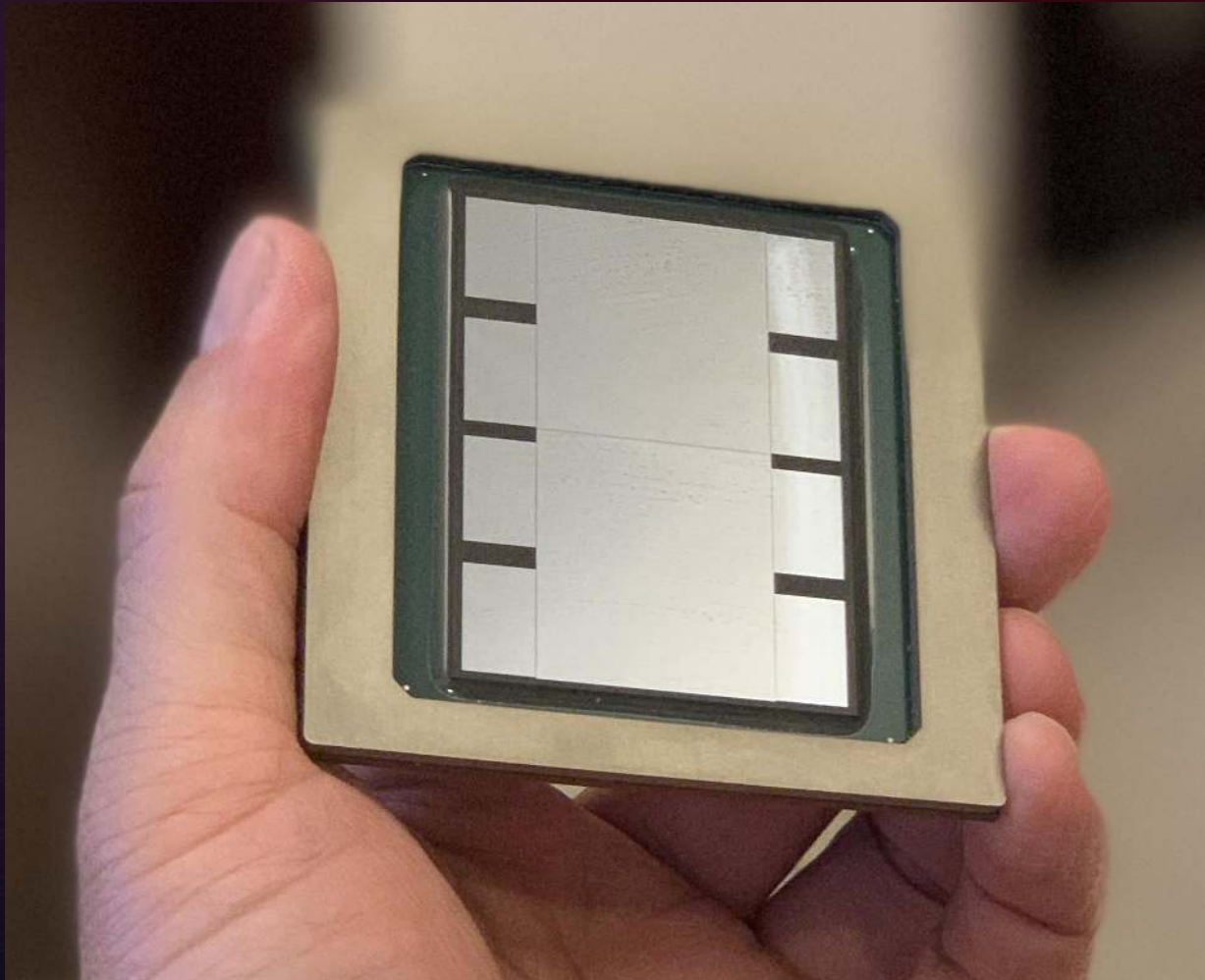
CoWoS[®] SYSTEM INTEGRATION



TSMC CoWoS[®] fully
assembled test chip
1 SoC + 2 DRAMs

Source: 2013 TSMC Technology Symposium

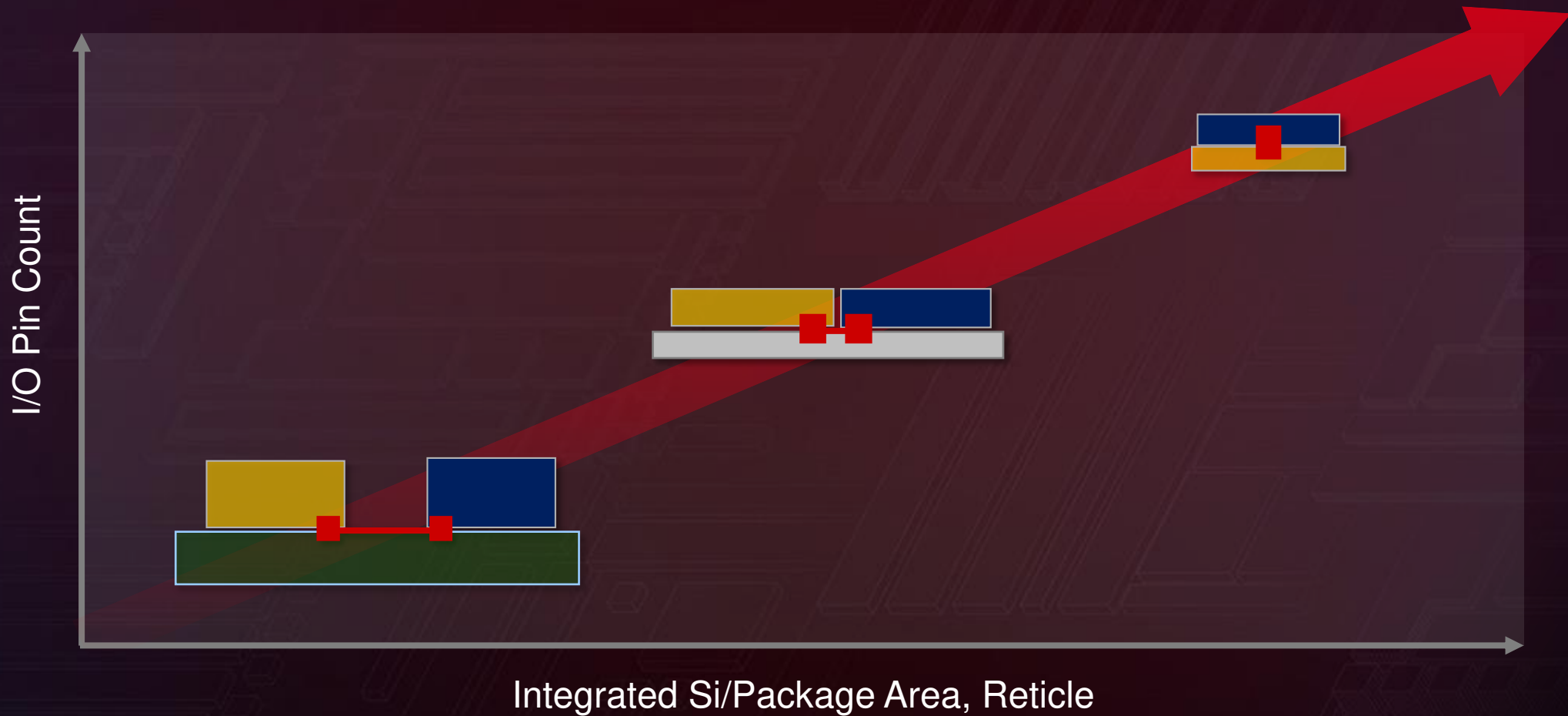
CoWoS[®] SYSTEM INTEGRATION



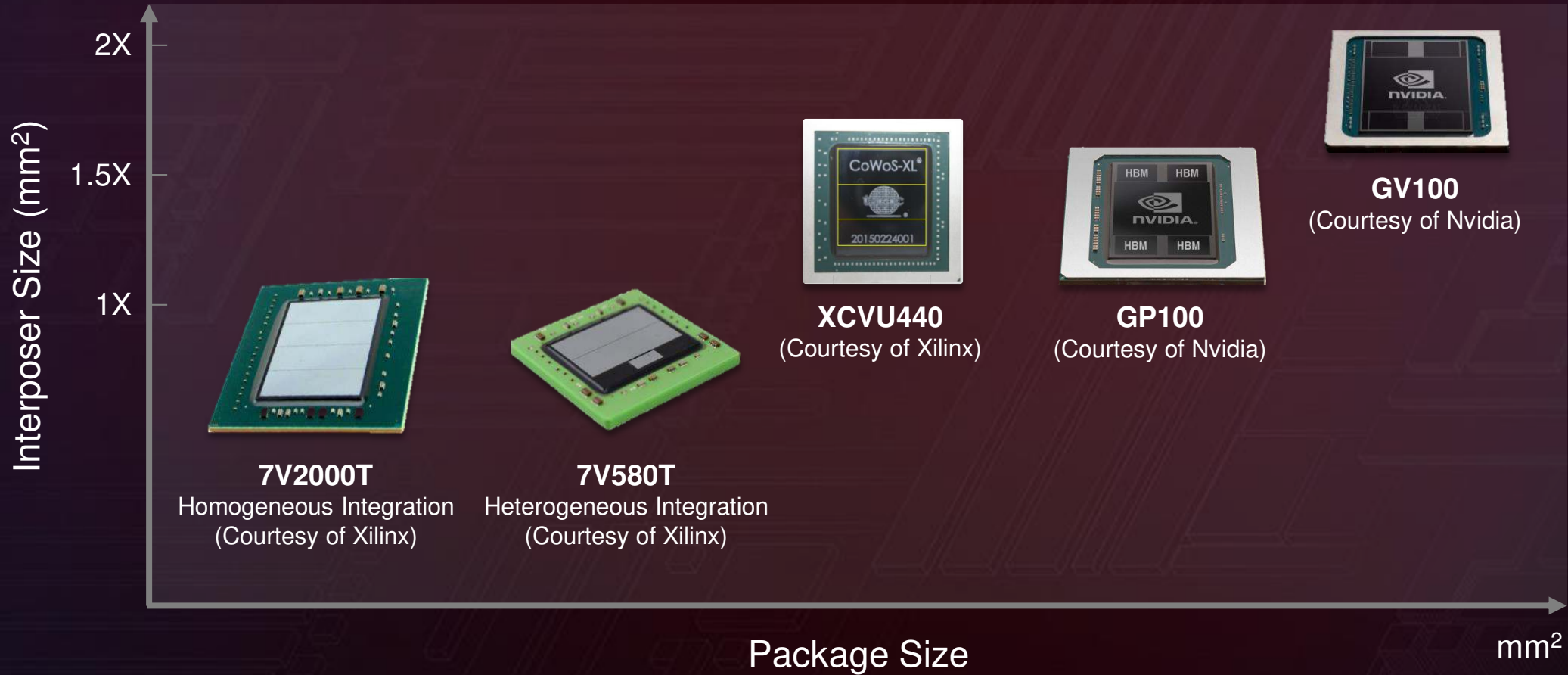
2500 mm² interposer:

- 2 processors (600 mm²)
- + 8 HBM DRAM

SYSTEM INTEGRATION TECHNOLOGIES



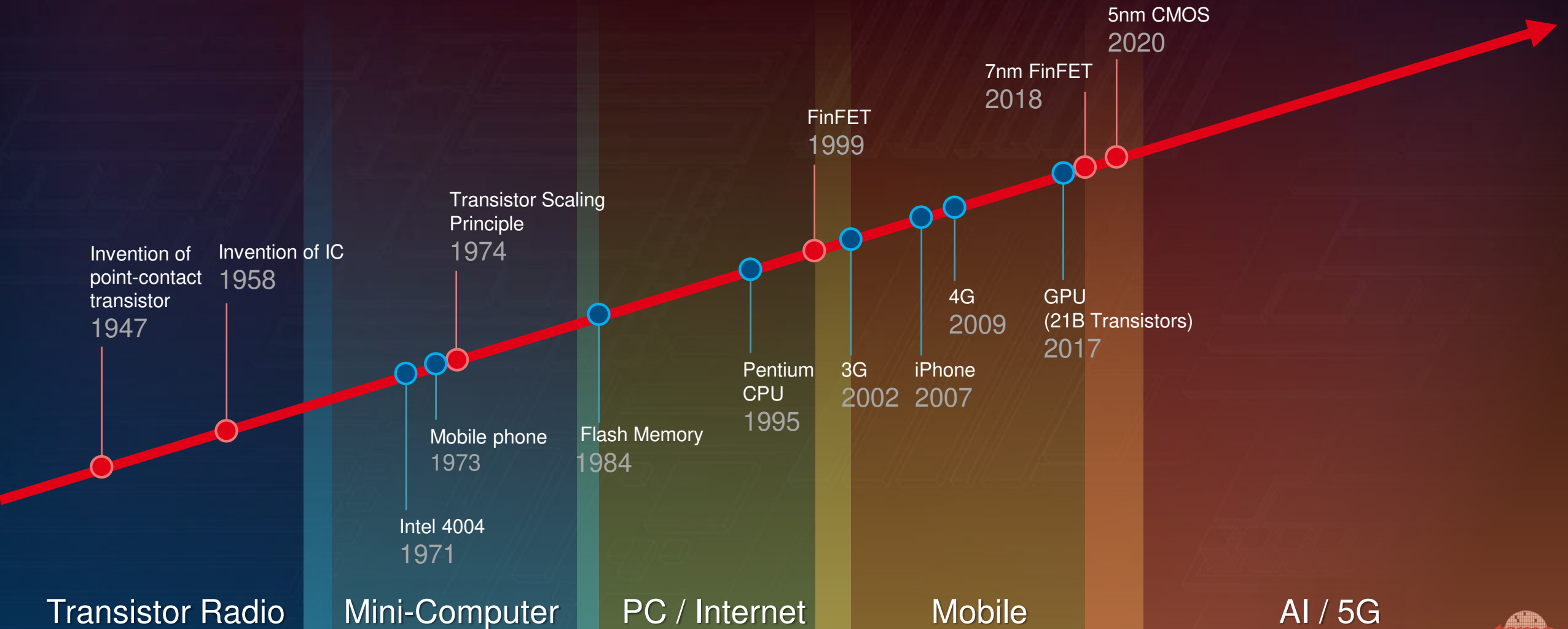
CHIPLETS INTEGRATION **REDUCES** SYSTEM COST PER FUNCTION



SEMICONDUCTOR TECHNOLOGY EVOLVES

DRIVEN BY CHANGING APPLICATION LANDSCAPE

2050 and beyond



Transistor Radio

Mini-Computer

PC / Internet

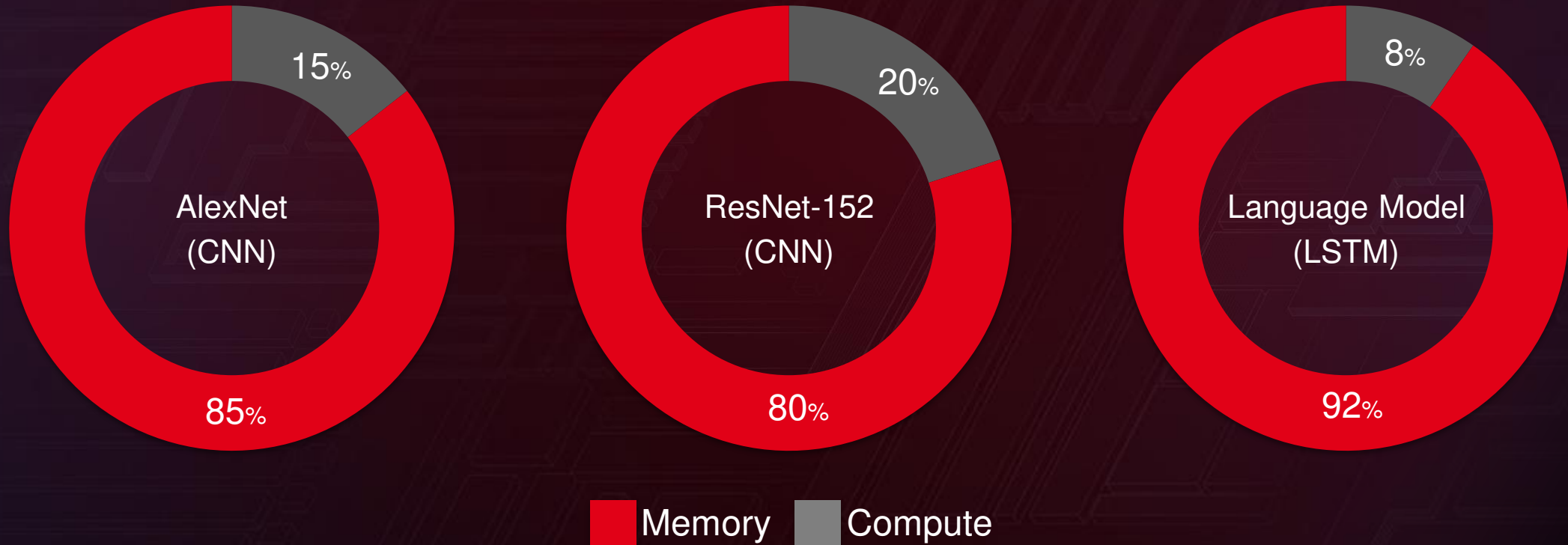
Mobile

AI / 5G



DATA MOVEMENT HITS THE MEMORY WALL

ABUNDANT-DATA APPLICATIONS: ENERGY MEASUREMENTS



Deep Learning Accelerators

Source: S. Mitra (Stanford)

Intel performance counter monitors 2 CPUs, 8-cores/ CPU + 128GB DRAM



DEEP NEURAL NETWORKS REQUIRE LARGE MEMORY CAPACITY

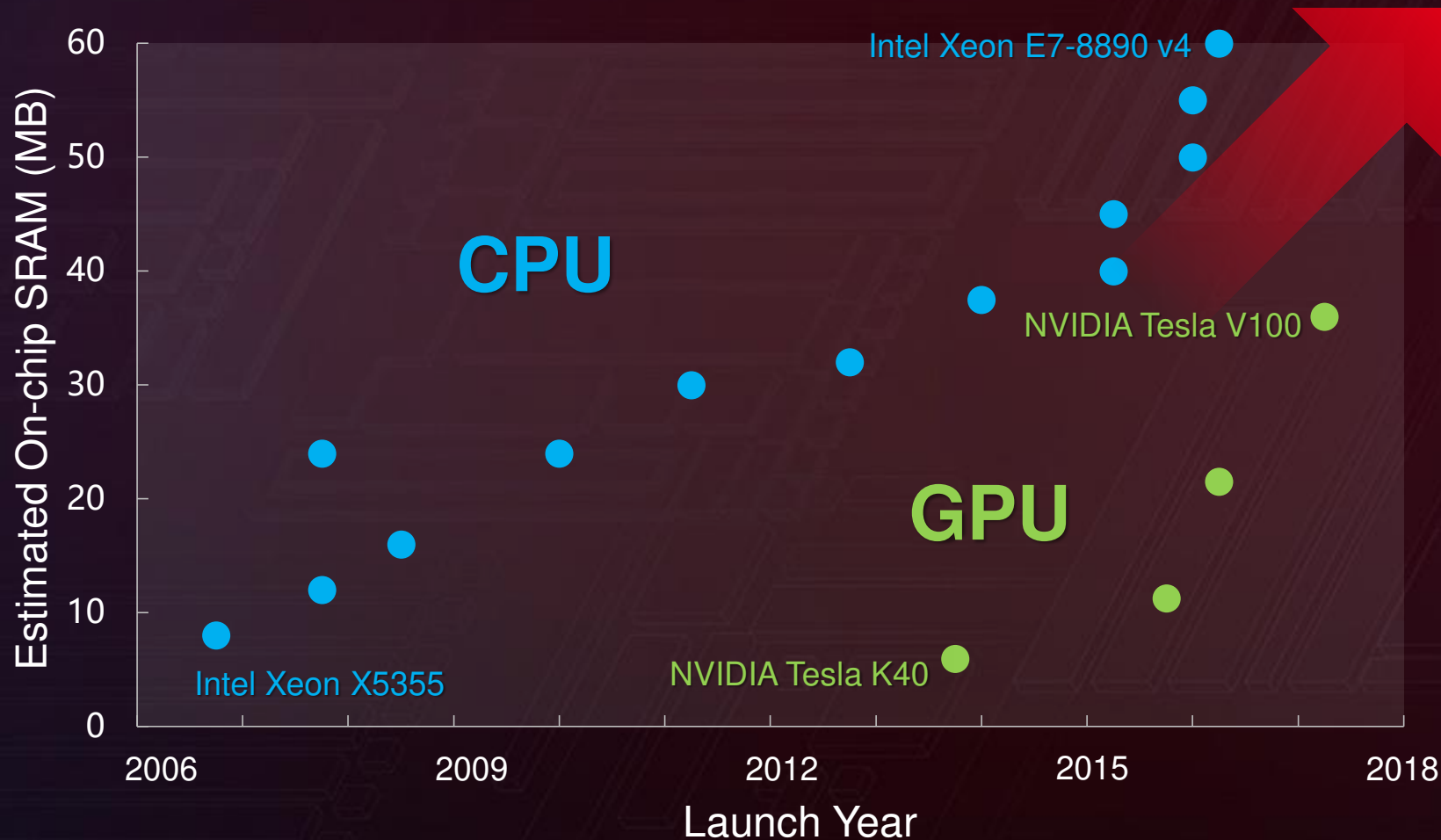
Network (application)	Type (LSTM/ CNN)	Training/ Inference	Model Size	Memory Usage (GBytes)
ResNet (vision)	CNN	Training	120 MBytes	21*
		Inference		0.12
Language Model (NLP)	LSTM	Training	2.5 GBytes	40*
		Inference		2.5

* Training memory usage: Batch size **64**, word size **64-bit**, memory can increase with greater batch sizes, footprint of activations, weights, errors and gradients.

Source: M. Lee, W. Hwang, Prof. S. Mitra (Stanford), M. Aly (NTU, Singapore), Y. Wang, K. Akarvardar (TSMC)

ON-CHIP SRAM CAPACITY: NEVER ENOUGH

3.8 Gbytes
@
1.4 nm node



Source: W. Hwang, Prof. S. Mitra (Stanford)

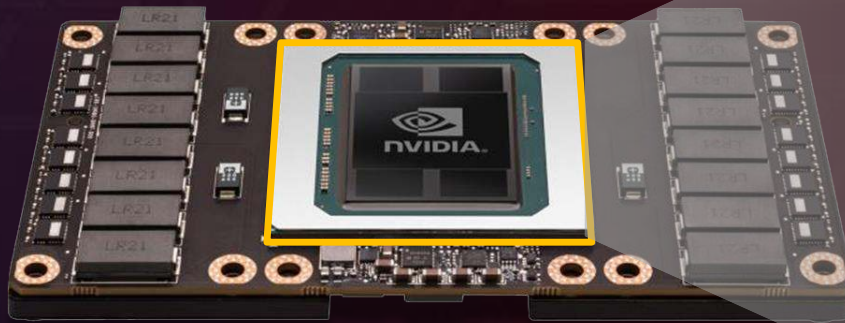
CAN WE PUT LOTS OF
MEMORY ON-CHIP?

WHAT KINDS OF MEMORY,
FOR **WHICH APPLICATION?**

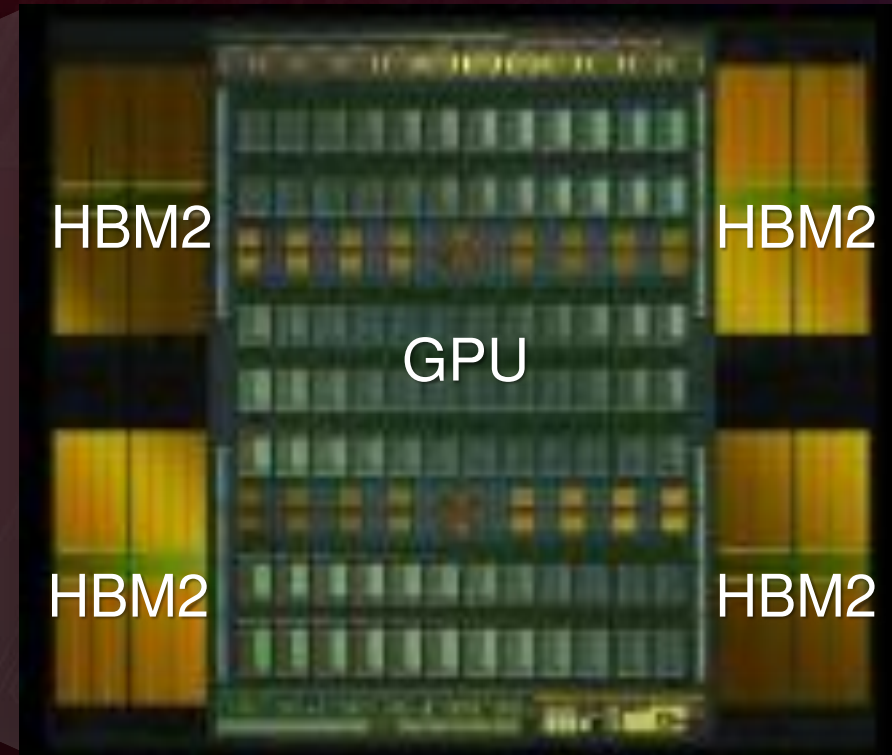
SUPER AI ACCELERATOR ENABLED BY CoWoS®

Heterogeneous Integration:
GPU + High Bandwidth Memory (HBM2)

CoWoS Module



Superior processing power that
equals to 100 CPUs

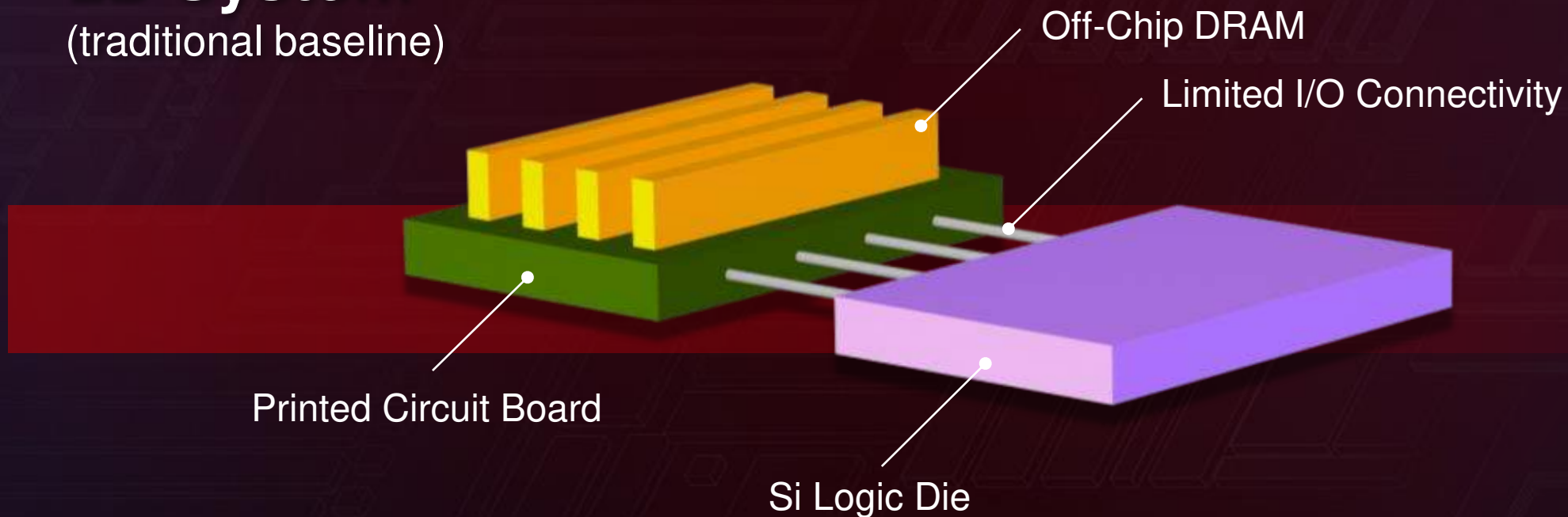


>300 B transistors

Source: "Inside Volta", Nvidia GPU Tech. Conf. , May 10, 2017.

COMPUTE-MEMORY INTEGRATION

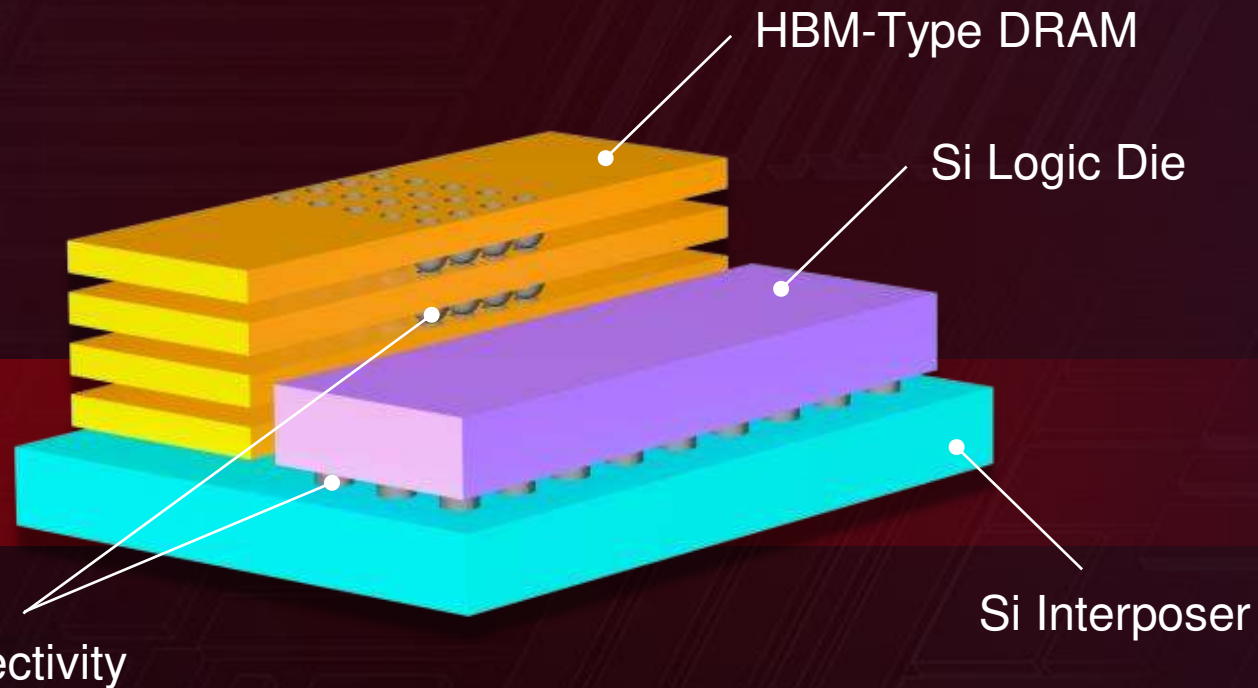
2D System (traditional baseline)



Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

COMPUTE-MEMORY INTEGRATION

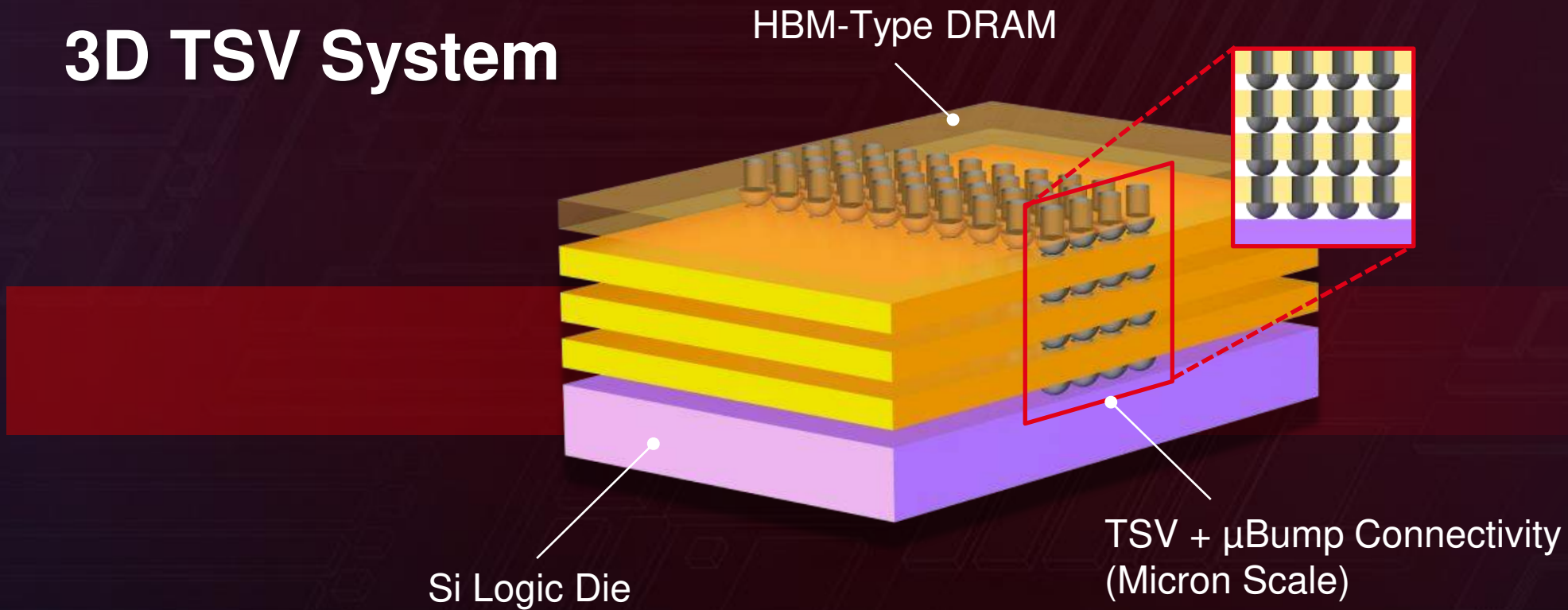
2.5D System



Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

COMPUTE-MEMORY INTEGRATION

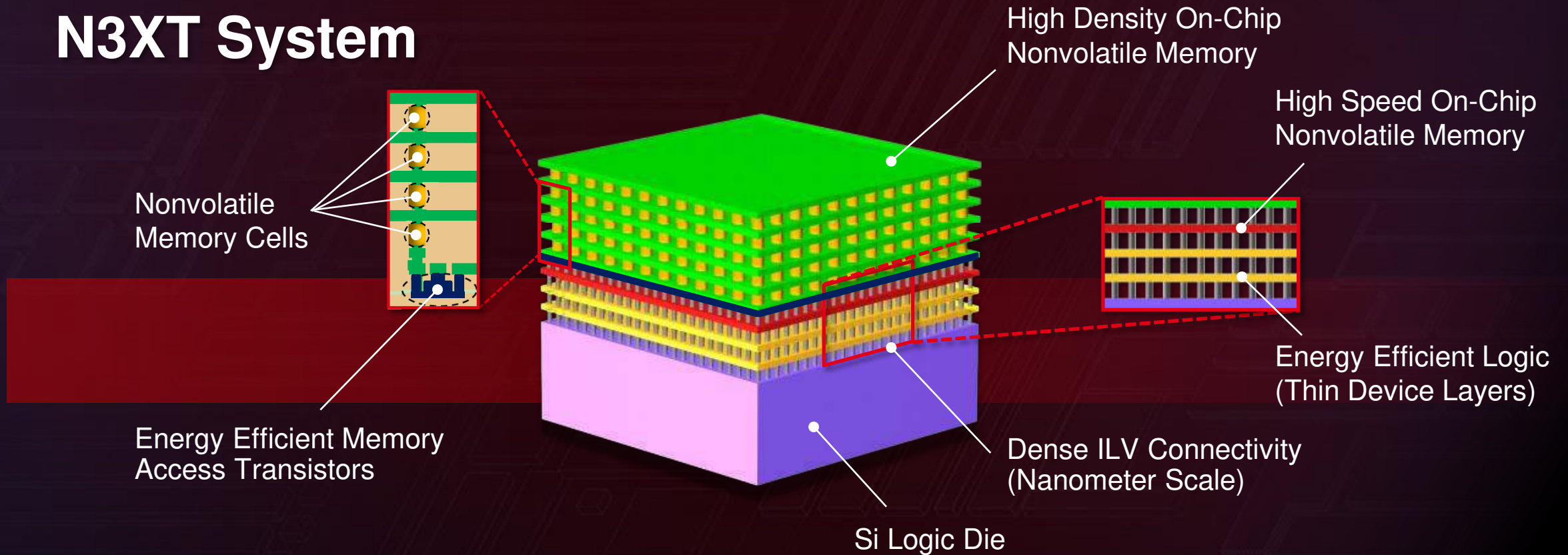
3D TSV System



Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

COMPUTE-MEMORY INTEGRATION

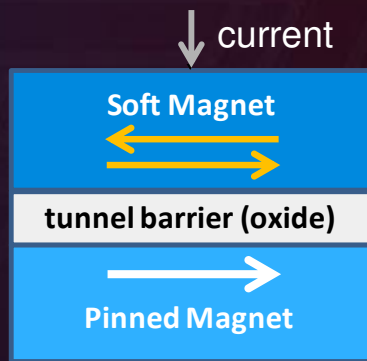
N3XT System



Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

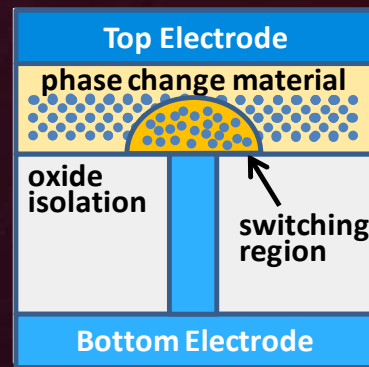
“NEW” MEMORIES FOR COMPUTE-MEMORY INTEGRATION

Random access, non-volatile, no erase before write, on-chip integration



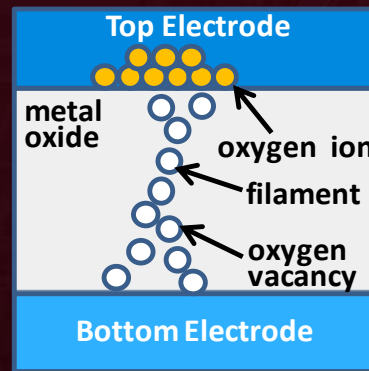
STT-MRAM

Spin torque transfer magnetic random access memory



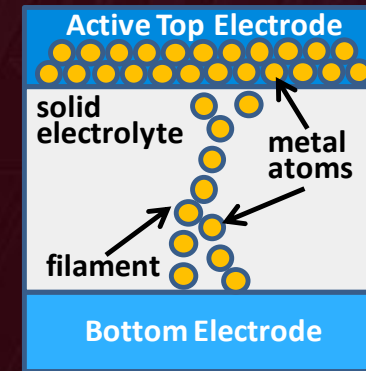
PCM

Phase change memory



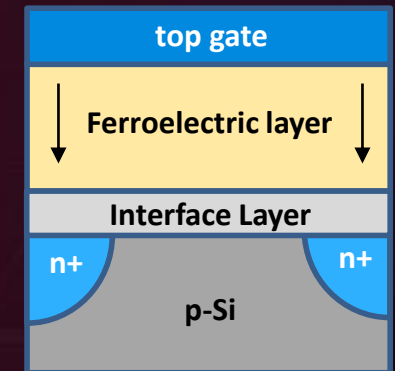
RRAM

Resistive switching random access memory



CBRAM

Conductive bridge random access memory



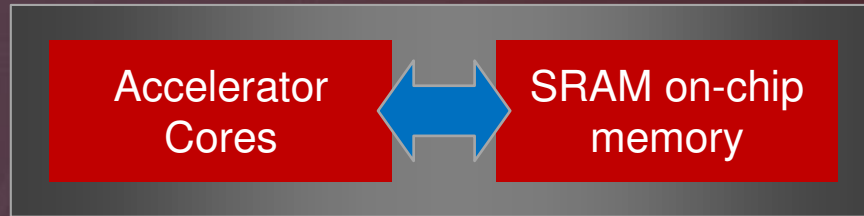
FERAM

Ferro-electric random access memory

Source: H.-S. P. Wong, S. Salahuddin, Nature Nanotech (2015)

NEW MEMORY: HIGH-BANDWIDTH, HIGH-CAPACITY, ON-CHIP

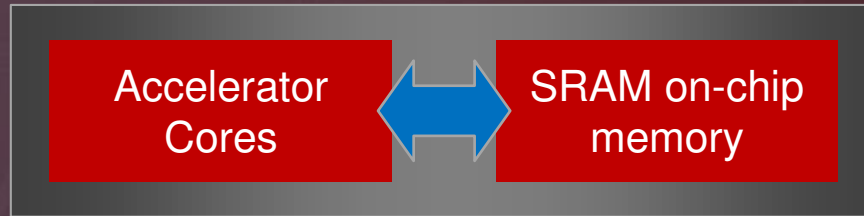
2D
baseline
system



Source: Stanford/NTU: M. Aly, S. Mitra, TSMC: Yih (Eric) Wang, K. Akarvardar, 2019

NEW MEMORY: HIGH-BANDWIDTH, HIGH-CAPACITY, ON-CHIP

2D
baseline
system

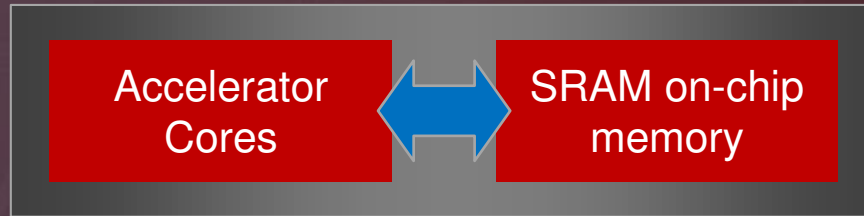


Off-chip DRAM (LPDDR3)

- Capacity: 4 GBytes
- Latency: 50 ns
- BW: 12 GBytes/s
- Read/write energy: 17 pJ/bit

NEW MEMORY: HIGH-BANDWIDTH, HIGH-CAPACITY, ON-CHIP

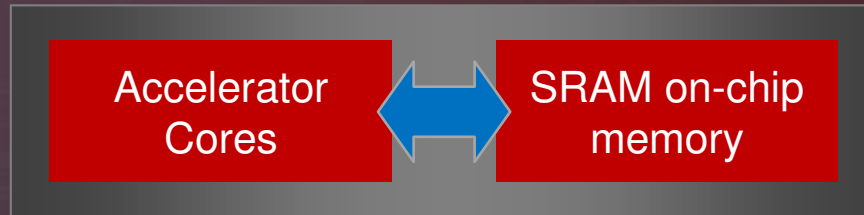
2D baseline system



Off-chip DRAM (LPDDR3)

- Capacity: 4 GBytes
- Latency: 50 ns
- BW: 12 GBytes/s
- Read/write energy: 17 pJ/bit

New system

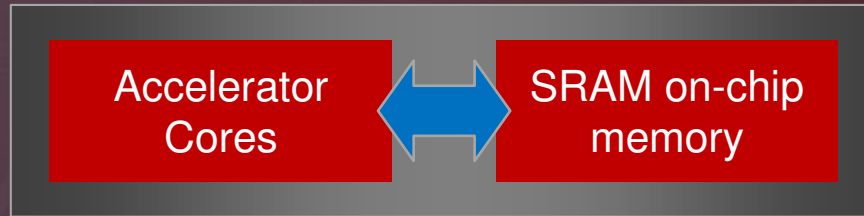


Off-chip DRAM (LPDDR3)

- Capacity: (4 GBytes minus New Mem. Cap.)
- Latency: 50 ns
- BW: 12 GBytes/s
- Read/write energy: 17 pJ/bit

NEW MEMORY: HIGH-BANDWIDTH, HIGH-CAPACITY, ON-CHIP

2D baseline system



Off-chip DRAM (LPDDR3)

- Capacity: 4 GBytes
- Latency: 50 ns
- BW: 12 GBytes/s
- Read/write energy: 17 pJ/bit

New system



On-chip New memory

- Capacity: sweep (up to 4 GBytes)
- Latency: sweep (down to 3ns)
- BW: sweep (up to 128 GBytes/s)
- Read/write energy: 5 pJ/bit

Off-chip DRAM (LPDDR3)

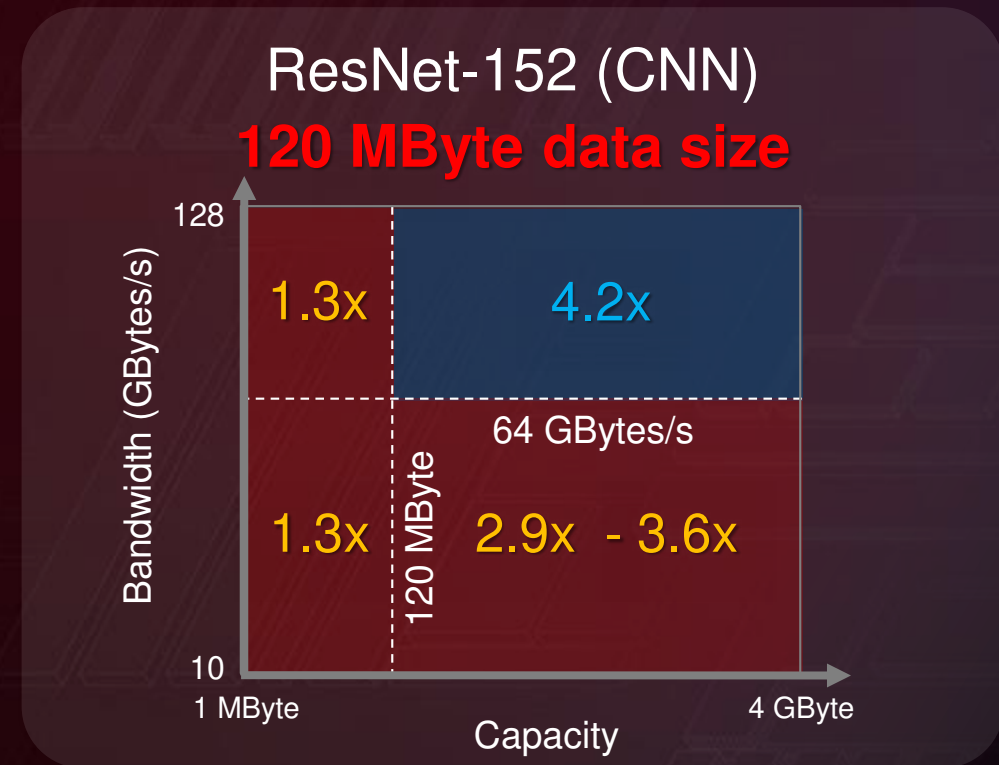
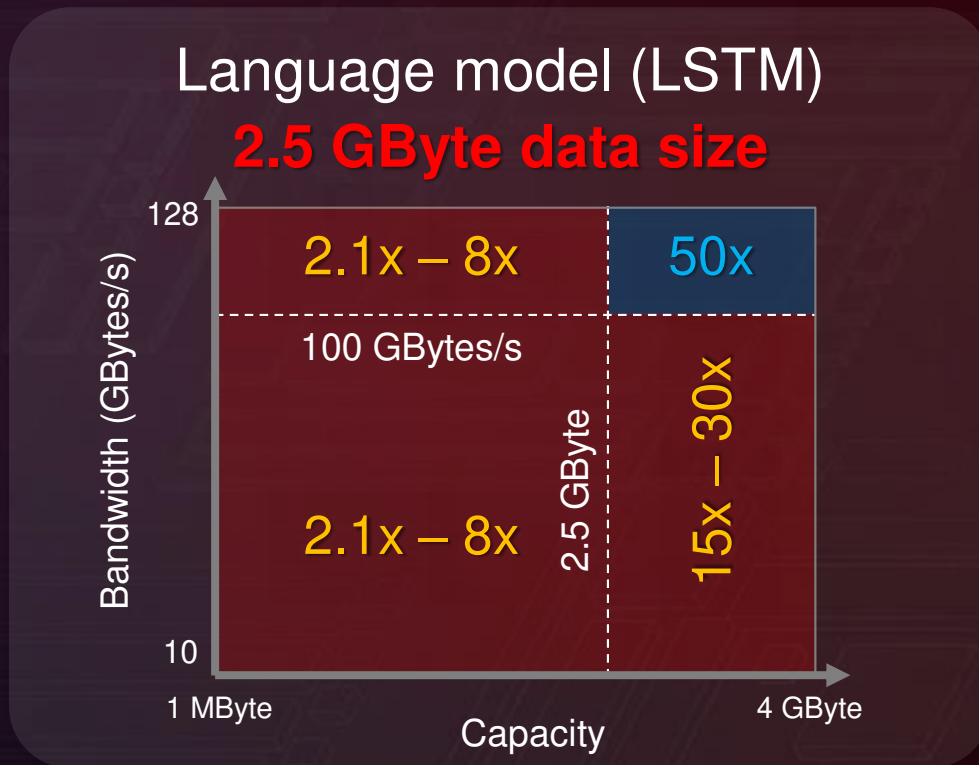
- Capacity: (4 GBytes minus New Mem. Cap.)
- Latency: 50 ns
- BW: 12 GBytes/s
- Read/write energy: 17 pJ/bit

High Bandwidth, High Capacity
both critical

NEW MEMORY ESSENTIAL REQUIREMENT

ON-CHIP CAPACITY MUST EXCEED DATA SIZE

EDP benefits



5 ns memory access latency, 5 pJ/bit access energy

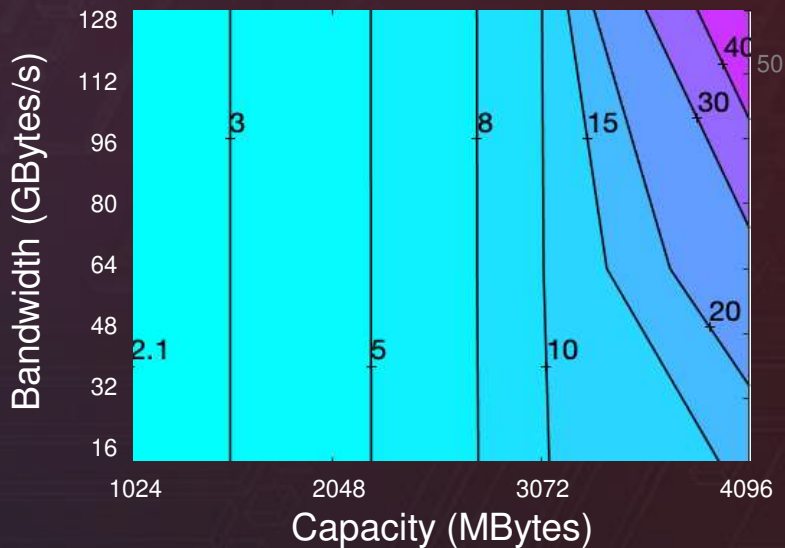
NEW MEMORY ESSENTIAL REQUIREMENT

ON-CHIP CAPACITY MUST EXCEED DATA SIZE

EDP benefits

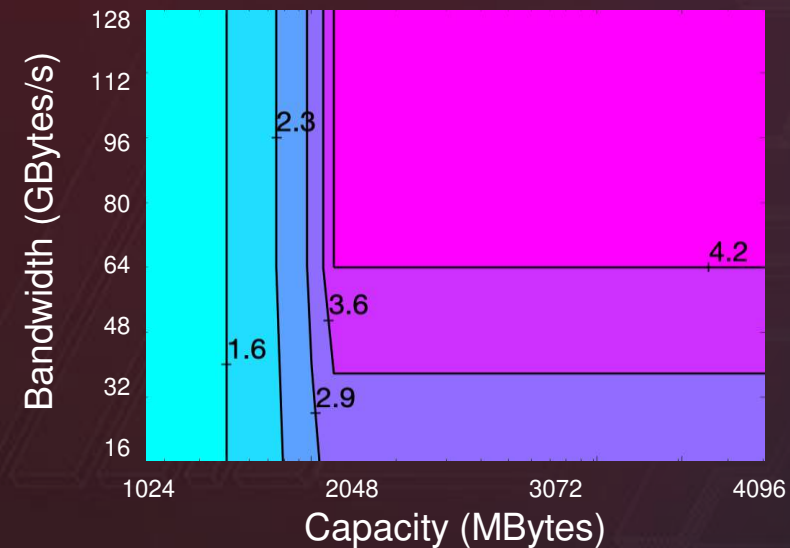
Language model (LSTM)

2.5 GByte data size



ResNet-152 (CNN)

120 MByte data size



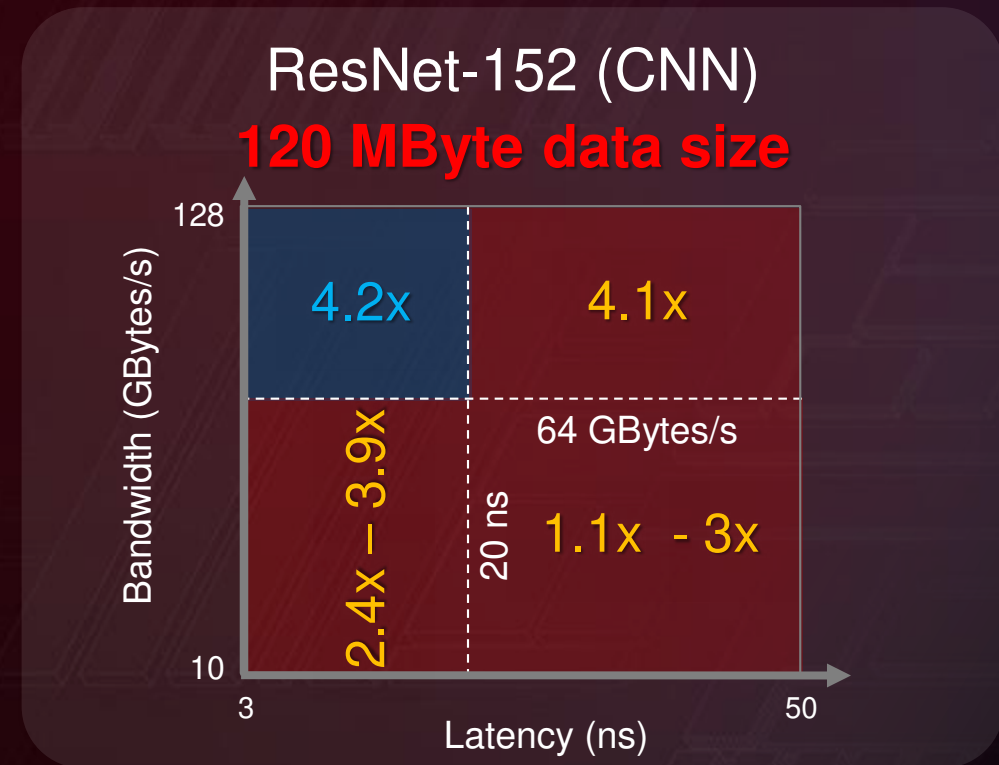
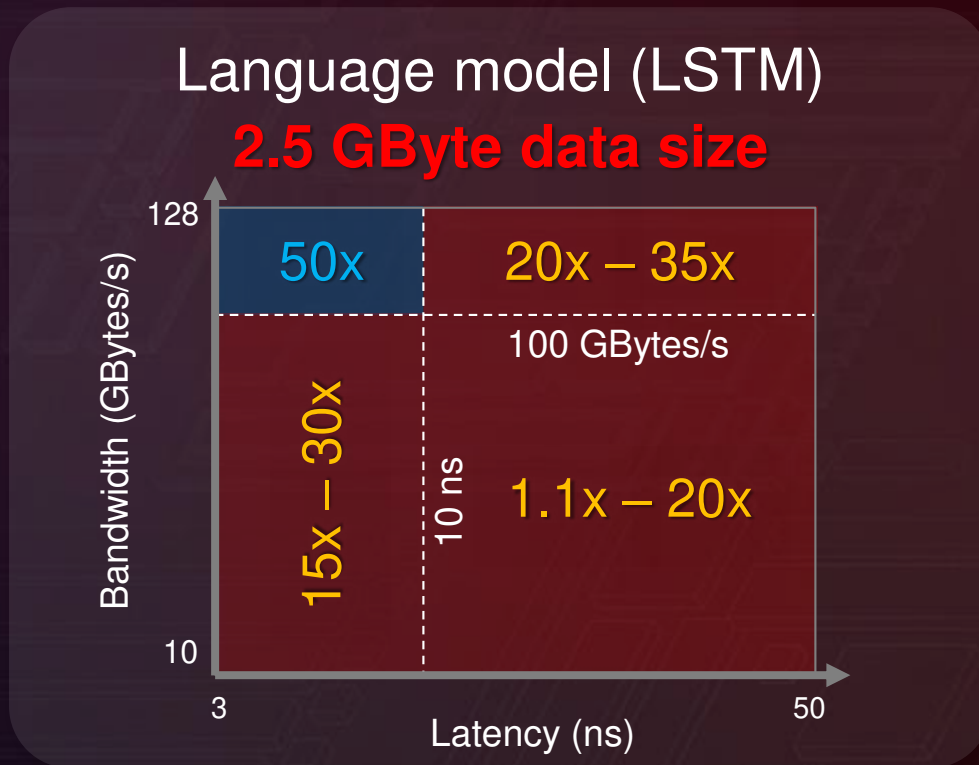
5 ns memory access latency, 5 pJ/bit access energy

Source: Stanford/NTU: M. Aly, S. Mitra, TSMC: Yih (Eric) Wang, K. Akarvardar, 2019

NEW MEMORY ESSENTIAL REQUIREMENT

HIGH BANDWIDTH MORE CRITICAL THAN LATENCY

EDP benefits



5 pJ/bit access energy

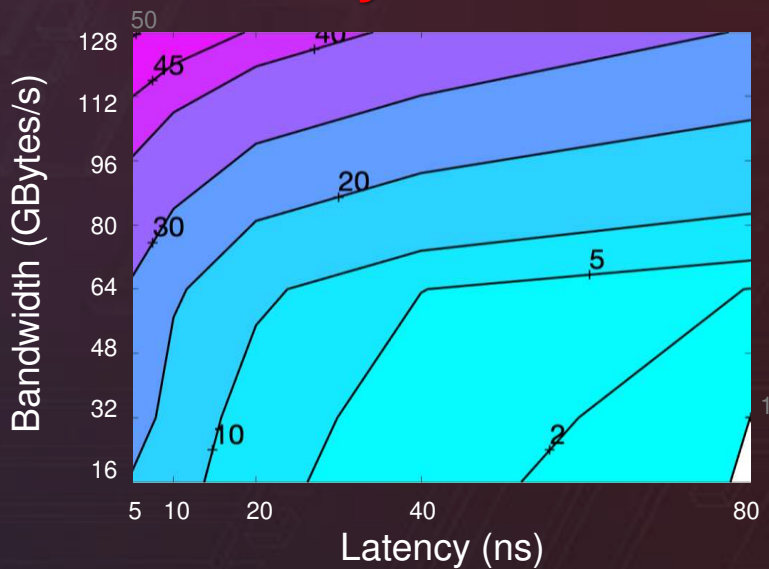
NEW MEMORY ESSENTIAL REQUIREMENT

HIGH BANDWIDTH MORE CRITICAL THAN LATENCY

EDP benefits

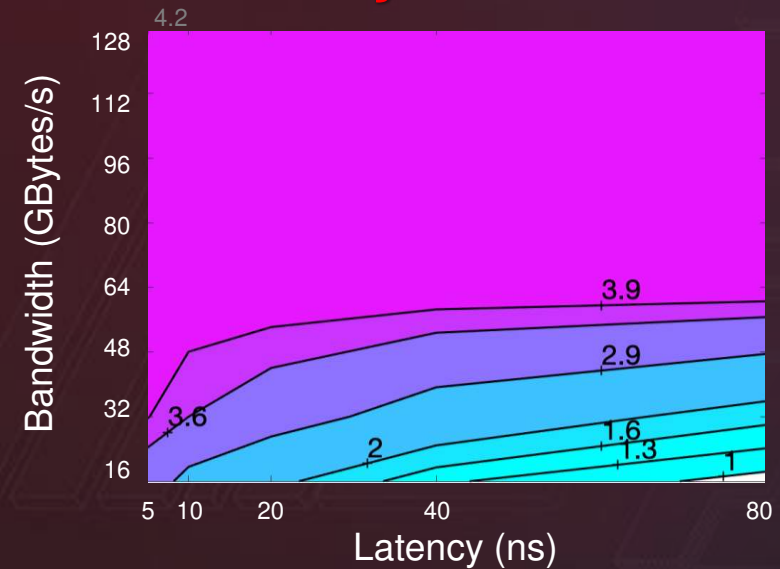
Language model (LSTM)

2.5 GByte data size



ResNet-152 (CNN)

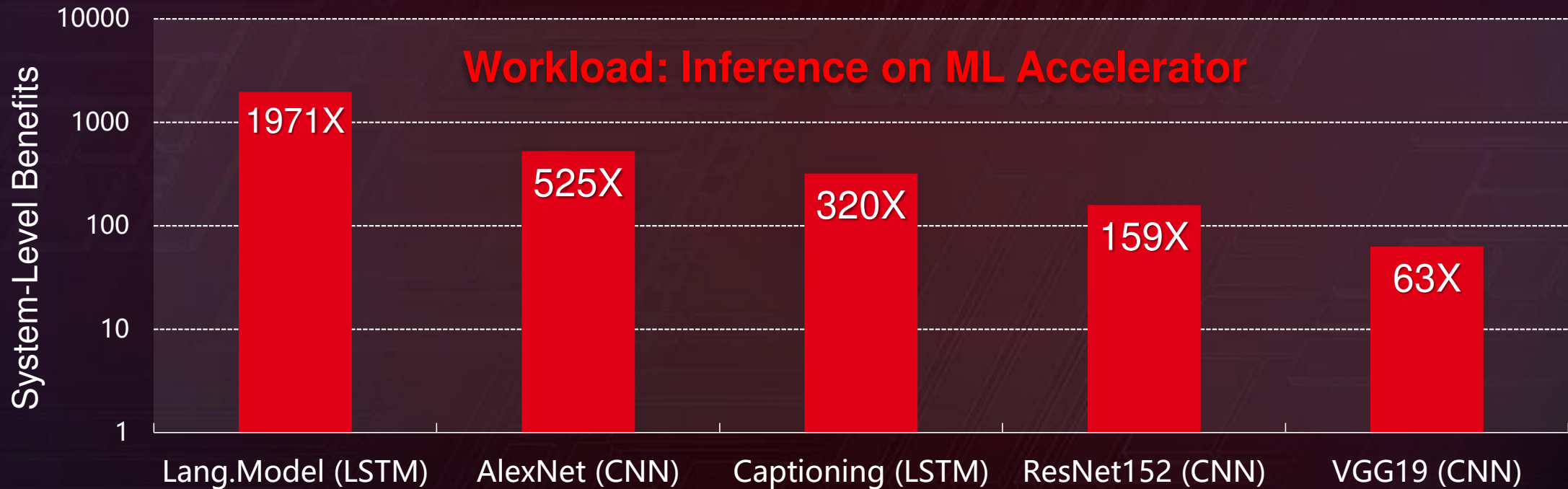
120 MByte data size



5 pJ/bit access energy

N3XT: UP TO ~2,000X ENERGY EFFICIENCY BENEFITS

Energy × Execution Time

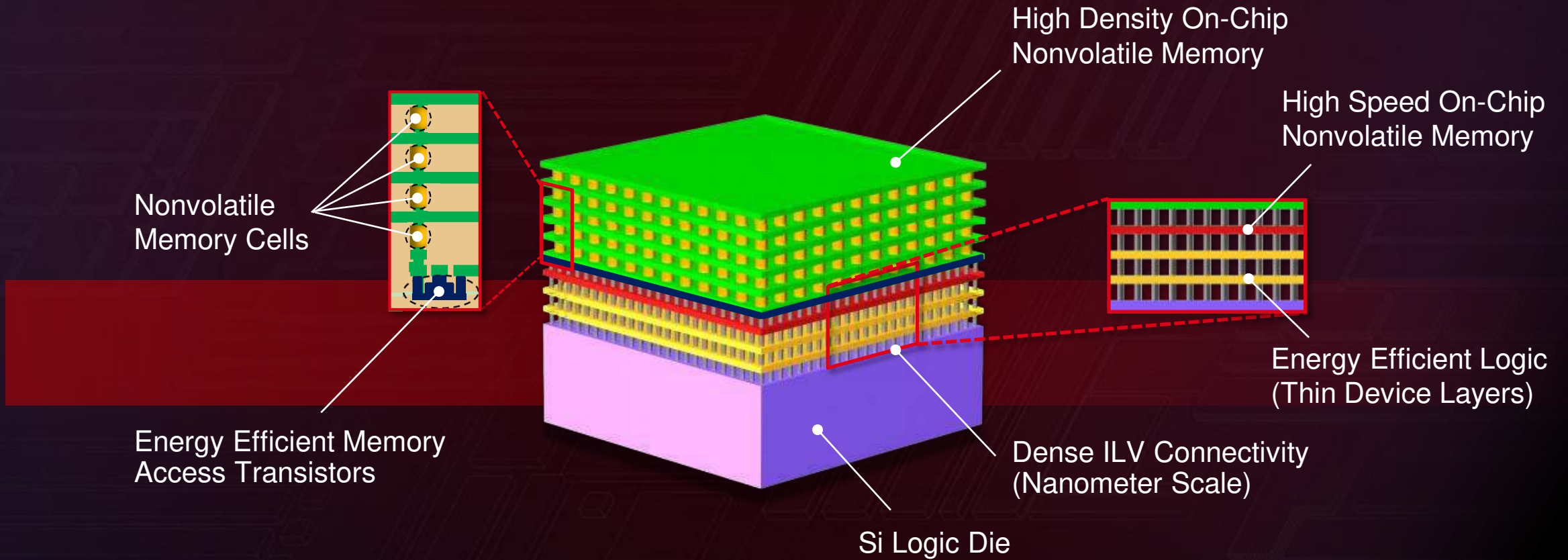


N3XT Benefits: relative to 2D Baseline System (28nm silicon CMOS, LPDDR3)
Inference: 16-bit data, batch size of 1

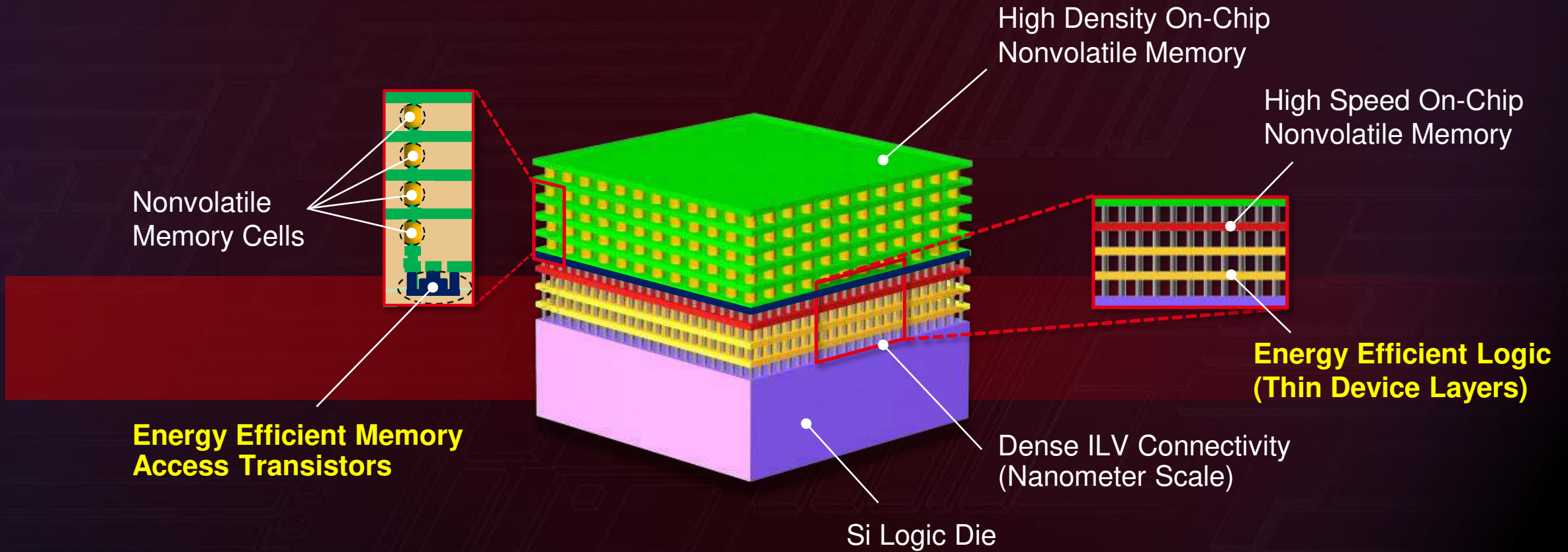
Source: Stanford/NTU: M. Aly, T. Wu, A. Bartolo, H.-S. P. Wong, S. Mitra et. al., Proc. IEEE 2019



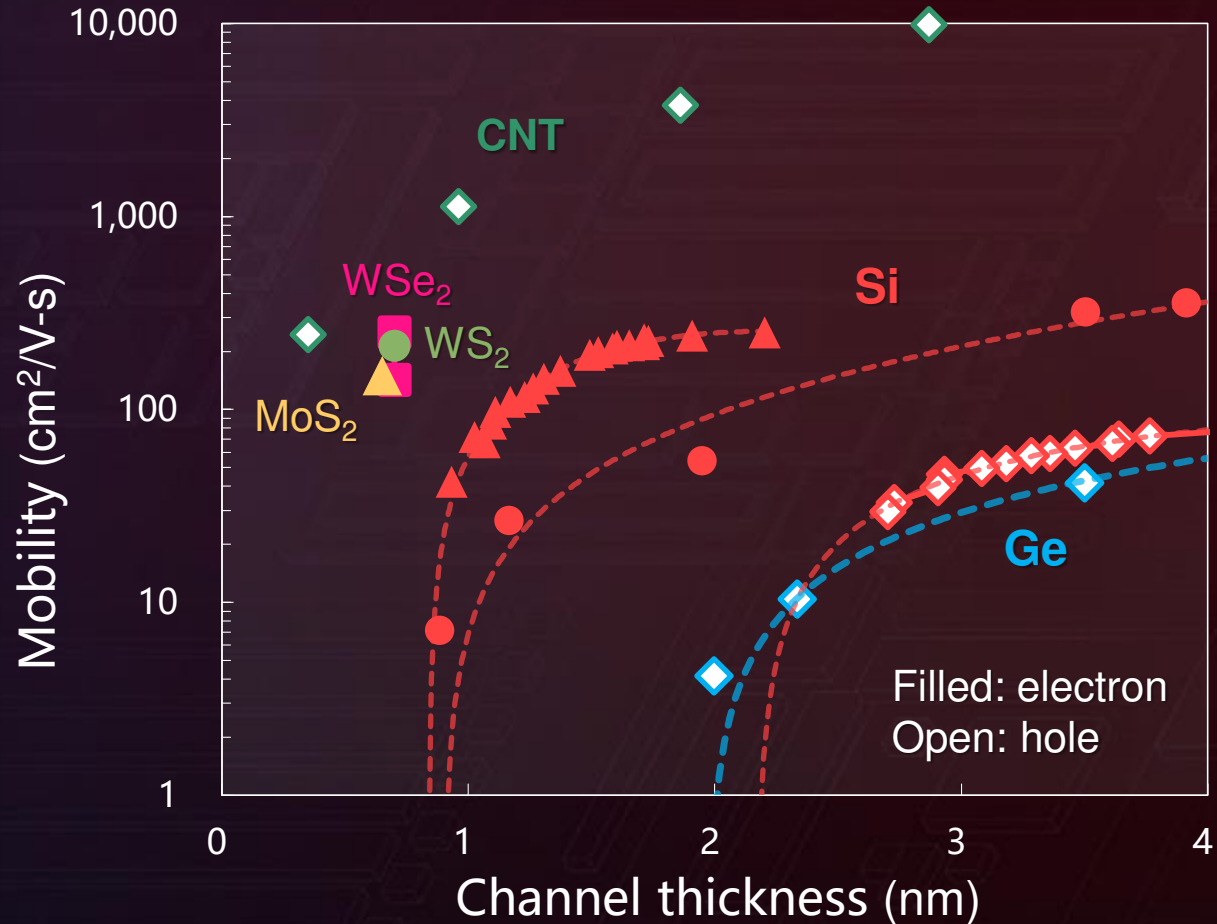
N3XT SYSTEM



N3XT SYSTEM



NANOMETER-THIN TRANSISTOR CHANNEL



Source: S.-K. Su, ... L.-J. Li (TSMC), Nature Nanotech., 2019.

1D carbon nanotube (CNT)

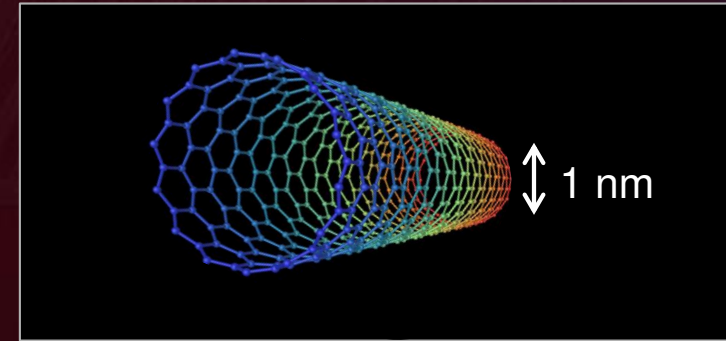


Photo credit: User Mstroeck on en.wikipedia

2D TMD (MoS₂, WSe₂, WS₂...)

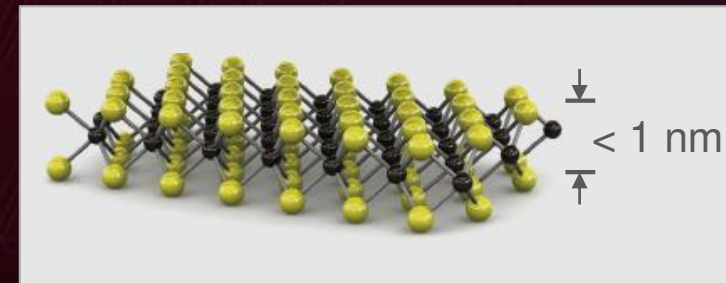
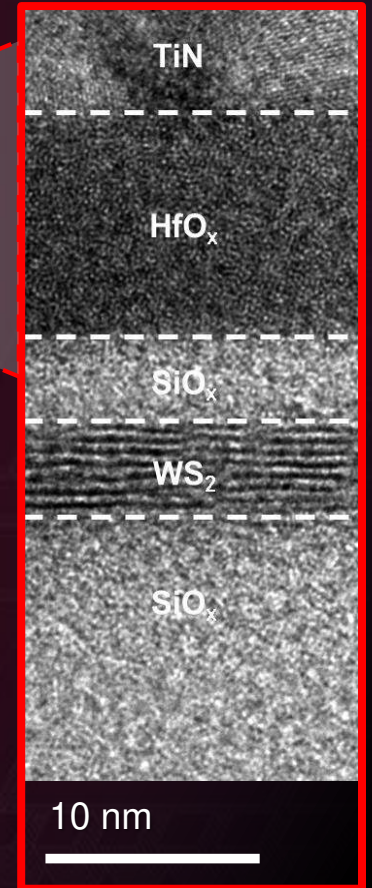
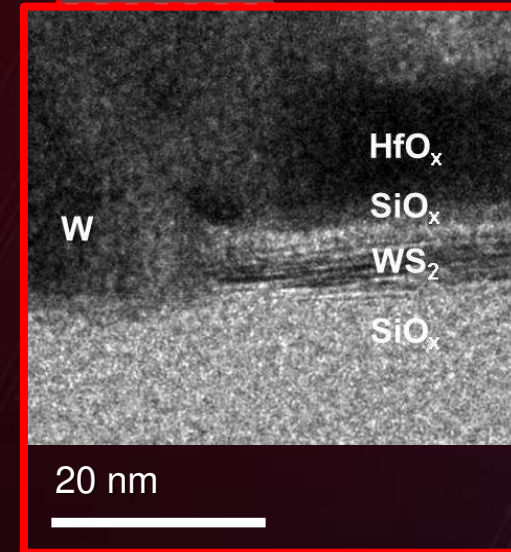
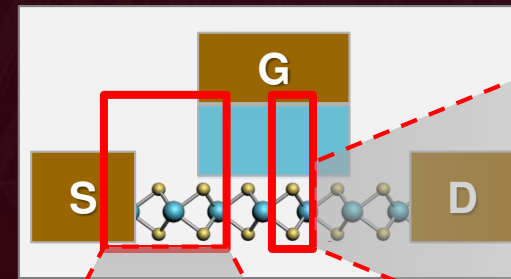
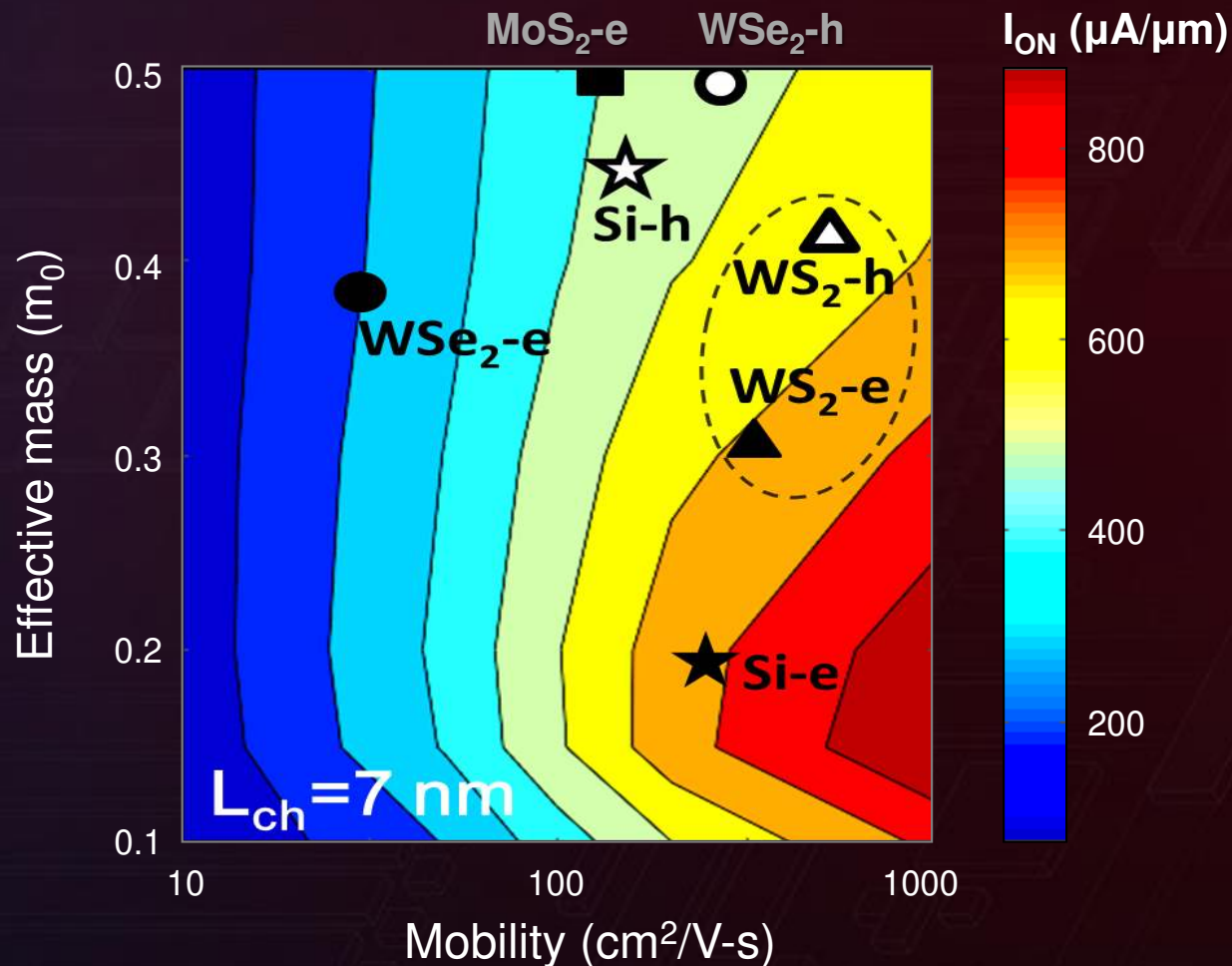


Photo credit: B. Radisavljevic et al., Nature Nanotech., p. 147, 2011

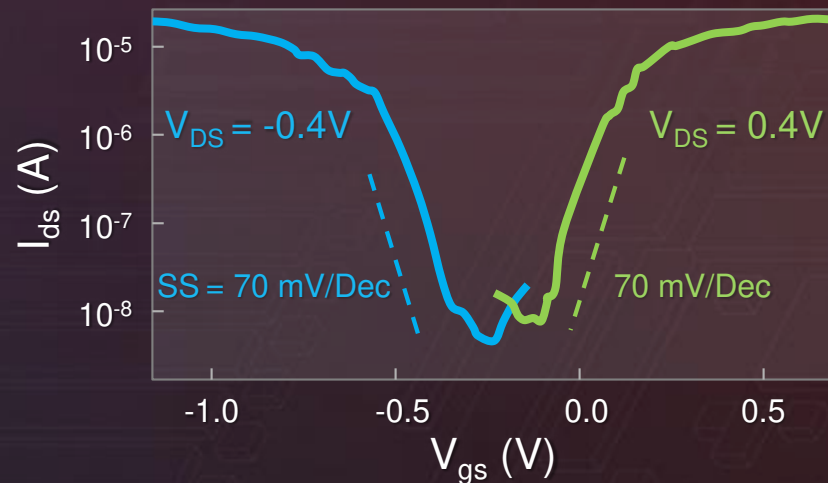
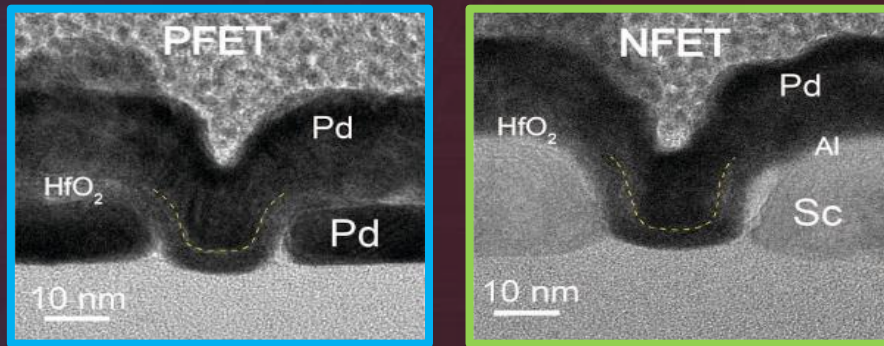
2D LAYERED MATERIALS (WS_2 , WSe_2)



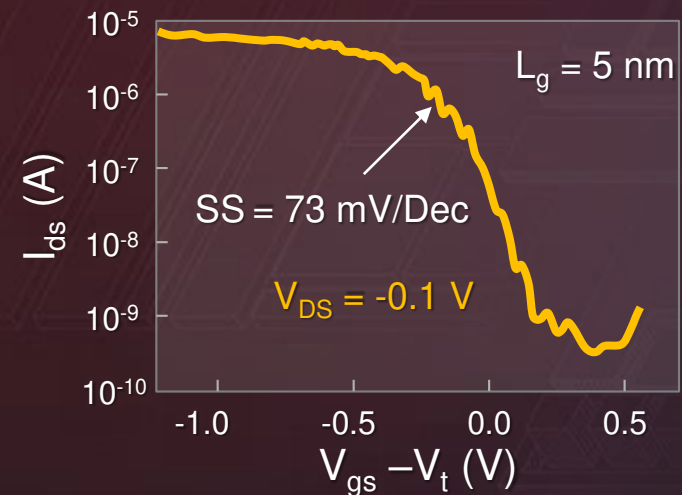
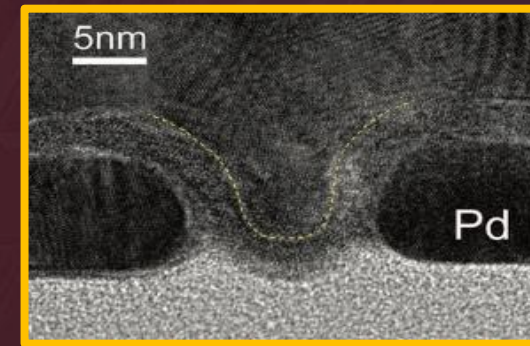
Source: C.-C. Cheng et al. (TSMC), Symp. VLSI Tech. 2019

SHORT-CHANNEL CARBON NANOTUBE TRANSISTORS

10 nm Gate Length

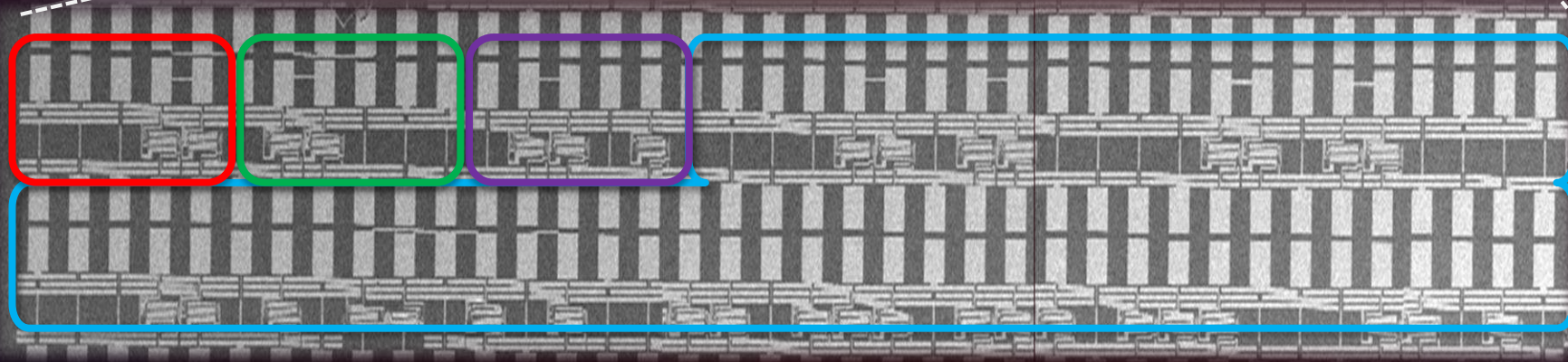
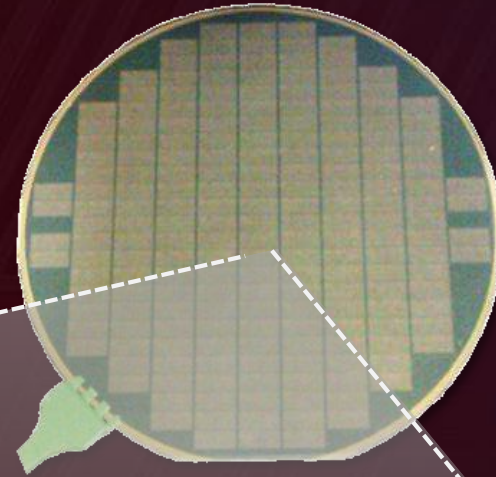


5 nm Gate Length



Source: C. Qiu, ...L-M. Peng (PKU), Science, 2017

CARBON NANOTUBE COMPUTER



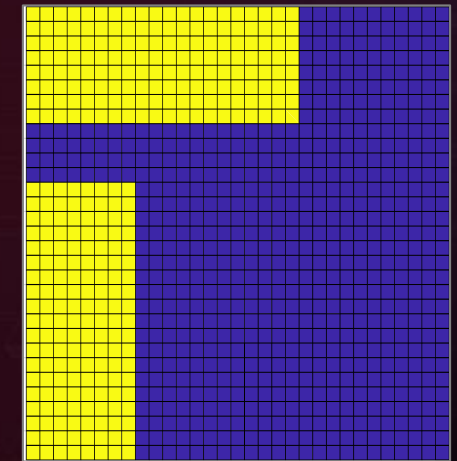
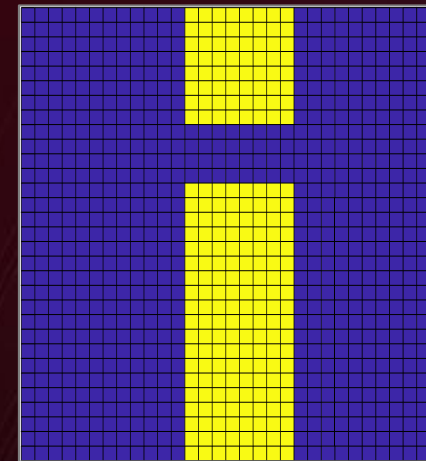
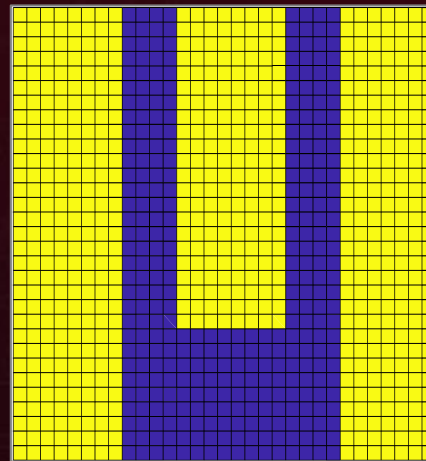
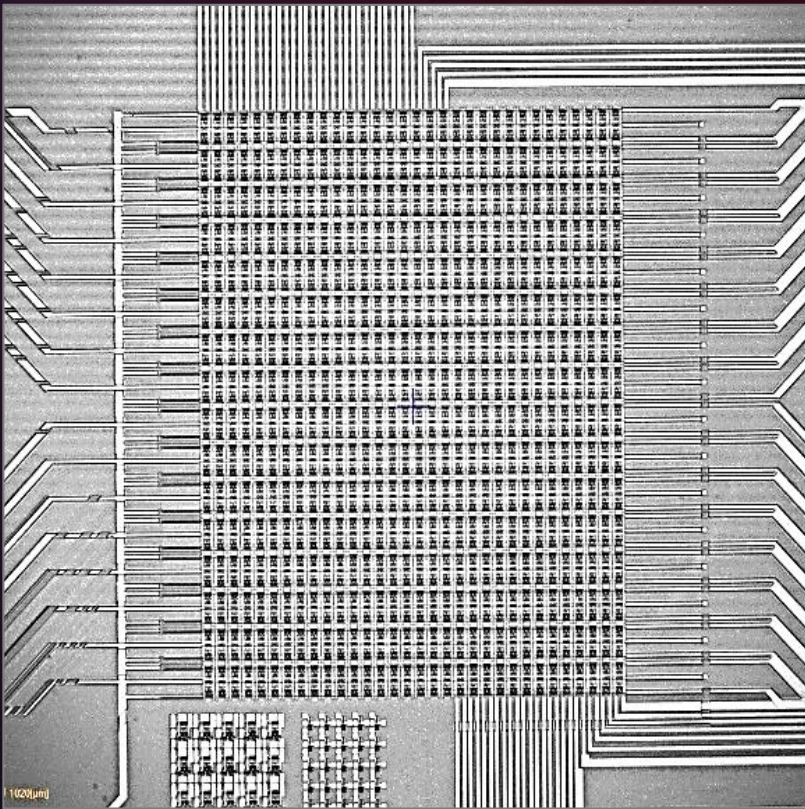
- instruction fetch
- data fetch
- arithmetic block
- write-back

Source: M. Shulaker, ... H.-S. P. Wong, S. Mitra (Stanford), Nature, 2013



CARBON NANOTUBE FET CMOS SRAM

Kbit 6T SRAM (6144 CNFETs)



Source: P. Kanhaiya, ... M. Shulaker (MIT), Symp. VLSI Tech., 2019

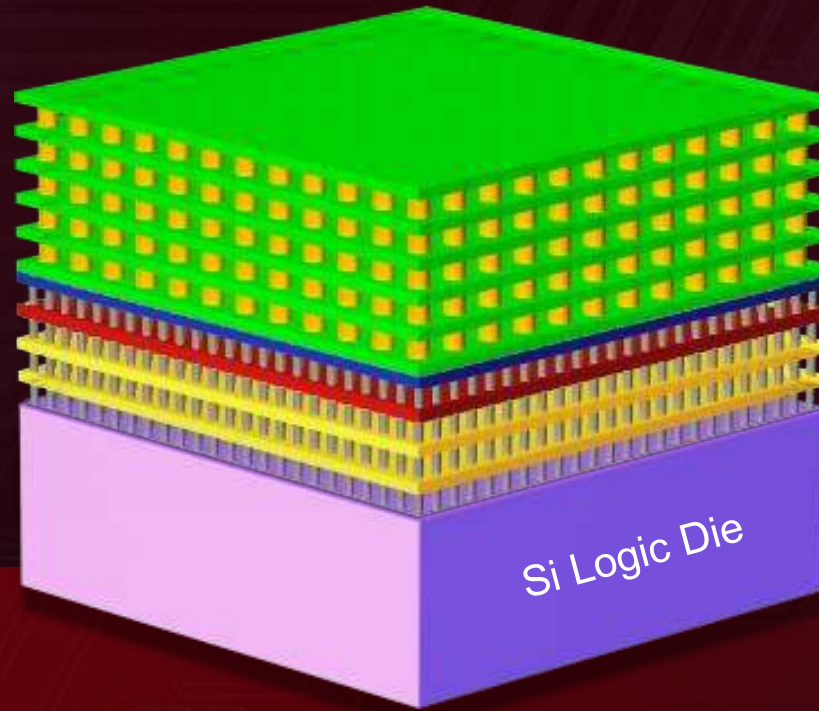
MEMORY INTEGRATION ON LOGIC PLATFORM

✗ Better transistor alone



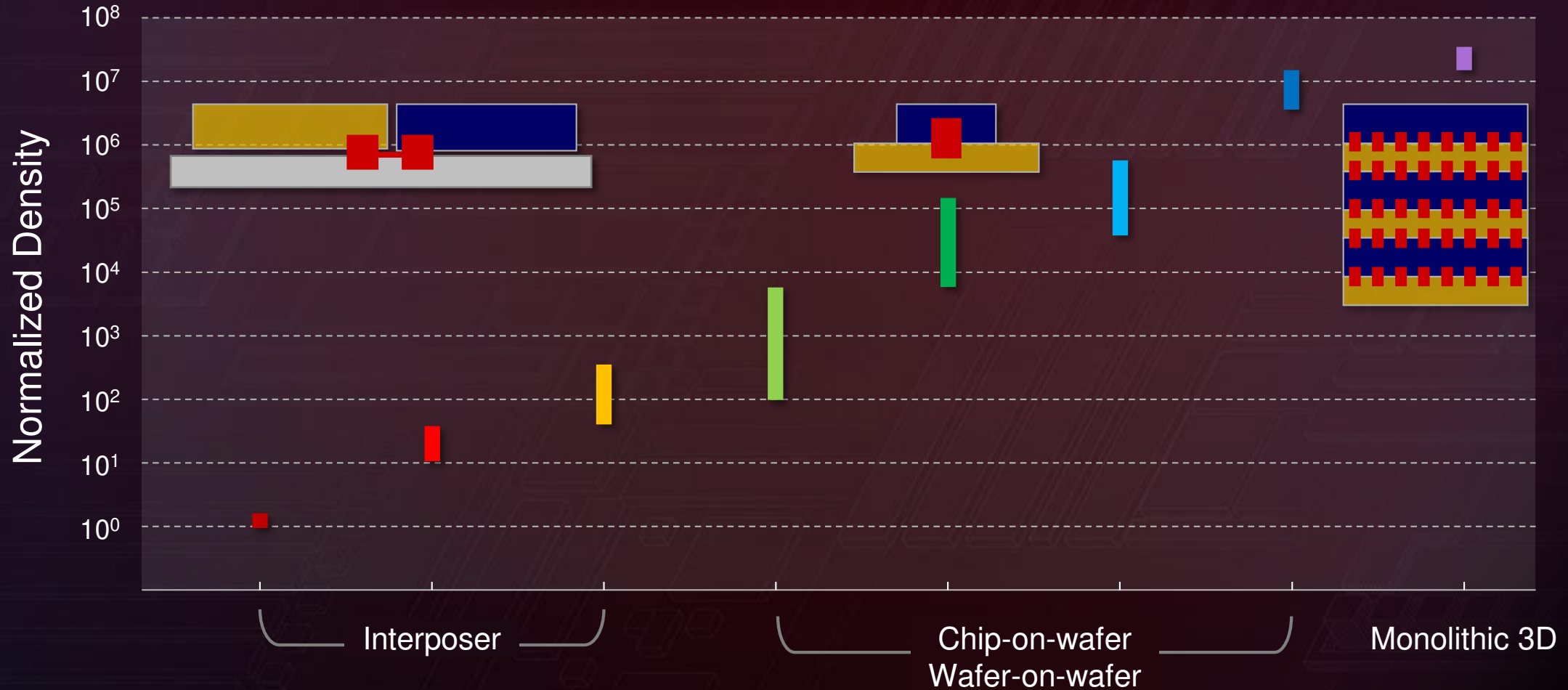
MEMORY INTEGRATION ON LOGIC PLATFORM

✓ Transistors integrated with memory in **3D**



SYSTEM INTEGRATION

A CONTINUUM FROM FAR BACK-END TO FRONT-END



Source: IMEC



SOCIETAL NEEDS FOR **ADVANCED TECHNOLOGY** IS INSATIABLE

ADVANCED TECHNOLOGY

– A KEY DIFFERENTIATOR

MULTIPLE ROADS LEAD TO ROME

Memory logic integration

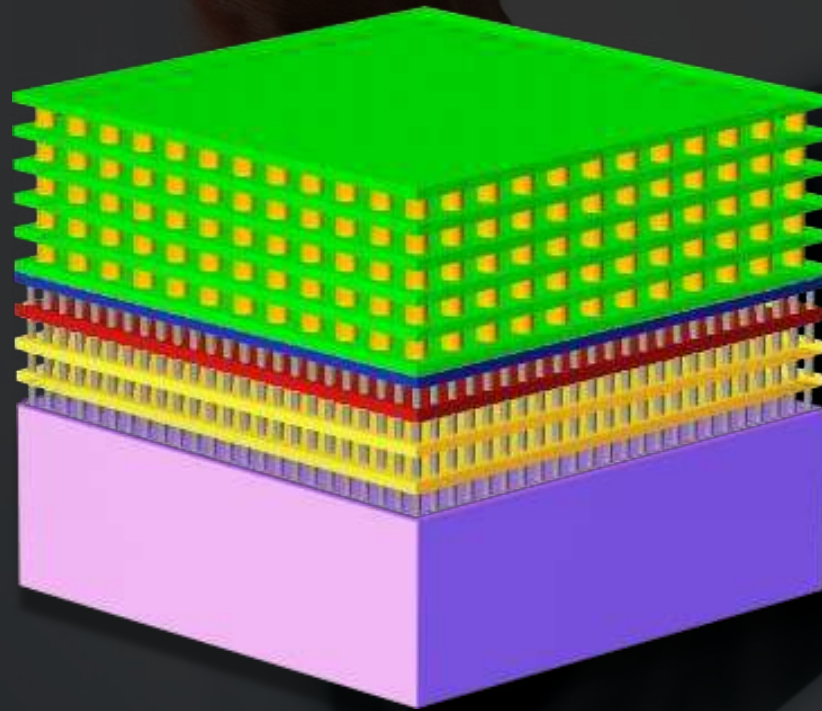
Continuous transistor & memory advances

System integration with
high connectivity

**CONTINUOUS
BENEFITS
NODE AFTER NODE**

A CALL TO **ACTION**: EARLY ENGAGEMENT

SYSTEM ↔ **TECHNOLOGY**



ACADEMIA ↔ **INDUSTRY RESEARCH**



| End of Talk

Questions?

MULTIPLE ROADS LEAD TO ROME

Memory logic integration

Continuous transistor & memory advances

System integration with high connectivity

**CONTINUOUS
BENEFITS
NODE AFTER NODE**

tsmc

COMMITTED TO PROVIDING THE MOST
ADVANCED TECHNOLOGIES

