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

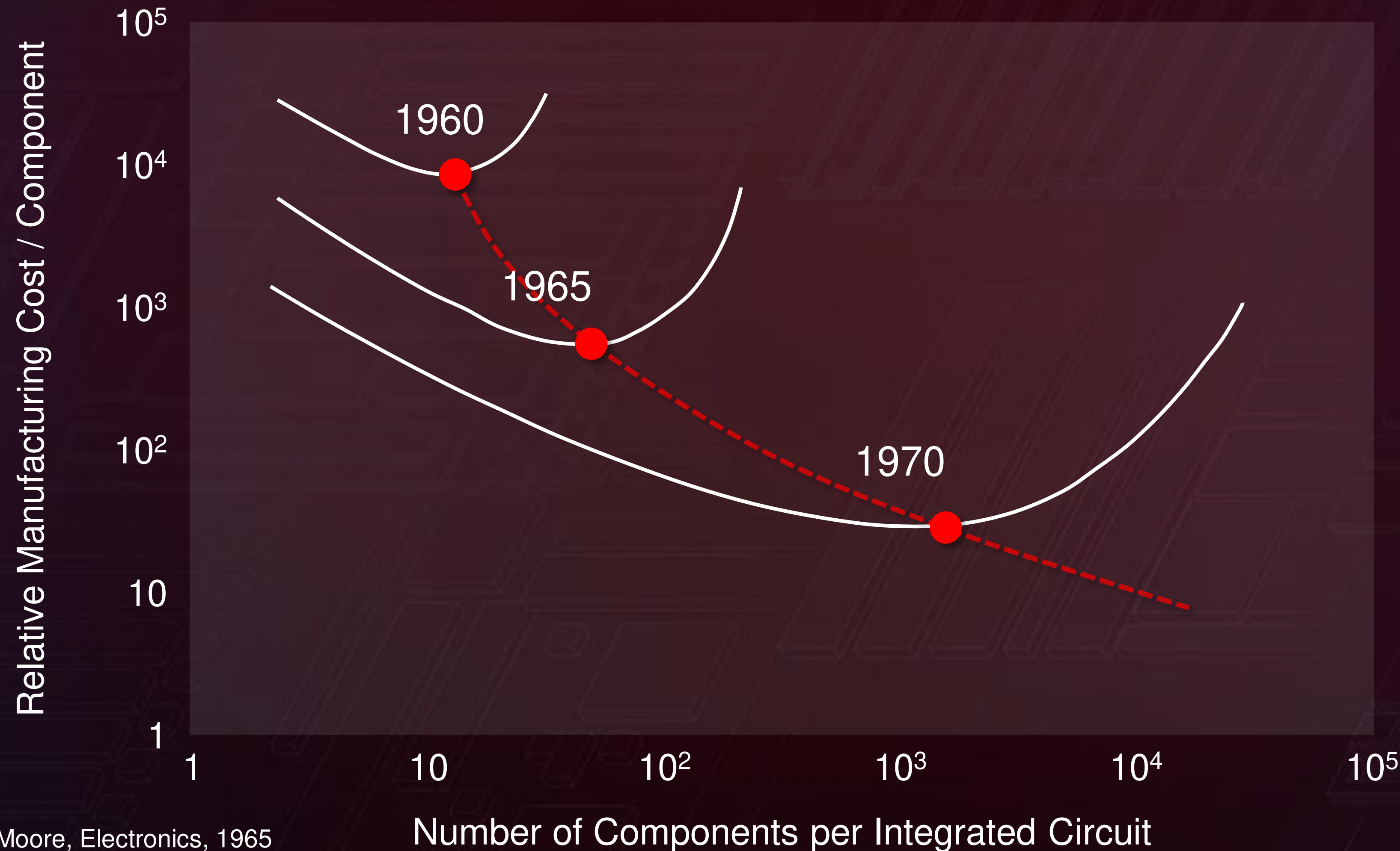
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}

1971 1980 1990 2000 2010 2017

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

© 2019 TSMC, Ltd

MOORE'S LAW
DENSITY AND COST PER FUNCTION

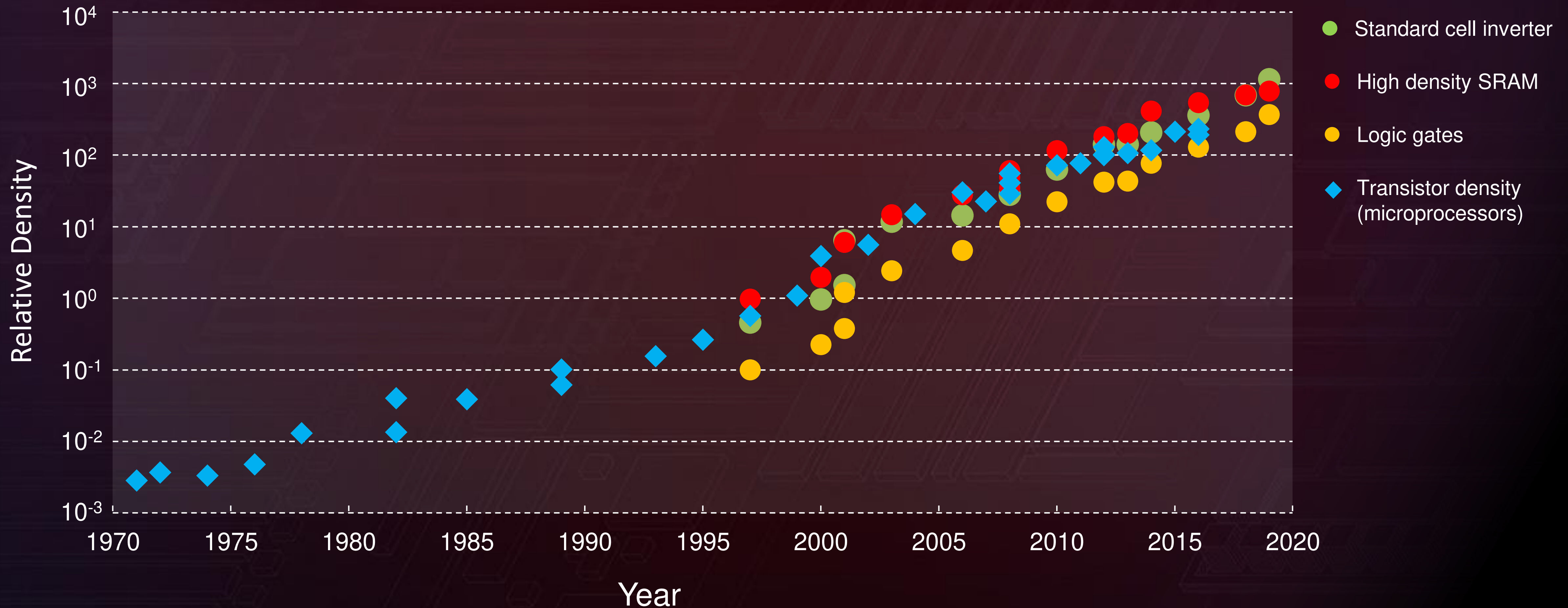


Source: G. Moore, Electronics, 1965

© 2019 TSMC, Ltd
tsmc

MOORE'S LAW IS WELL AND ALIVE

DENSITY: A NECESSARY ATTRIBUTE

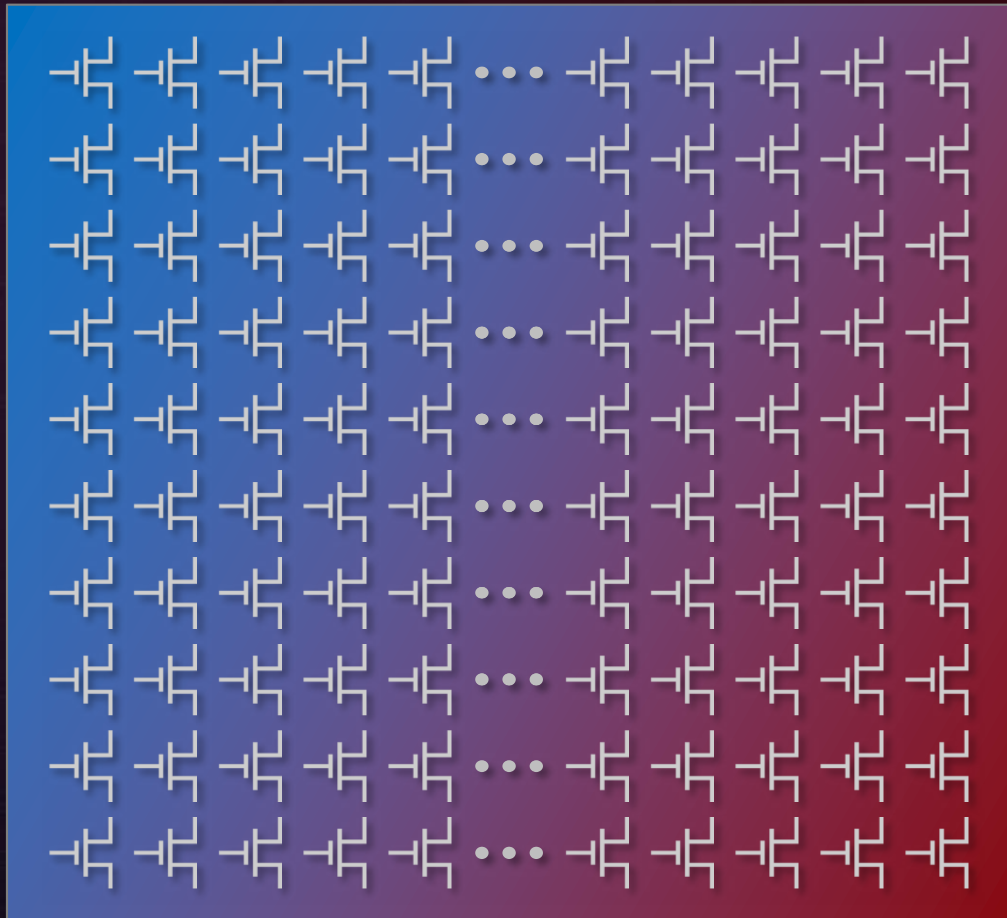


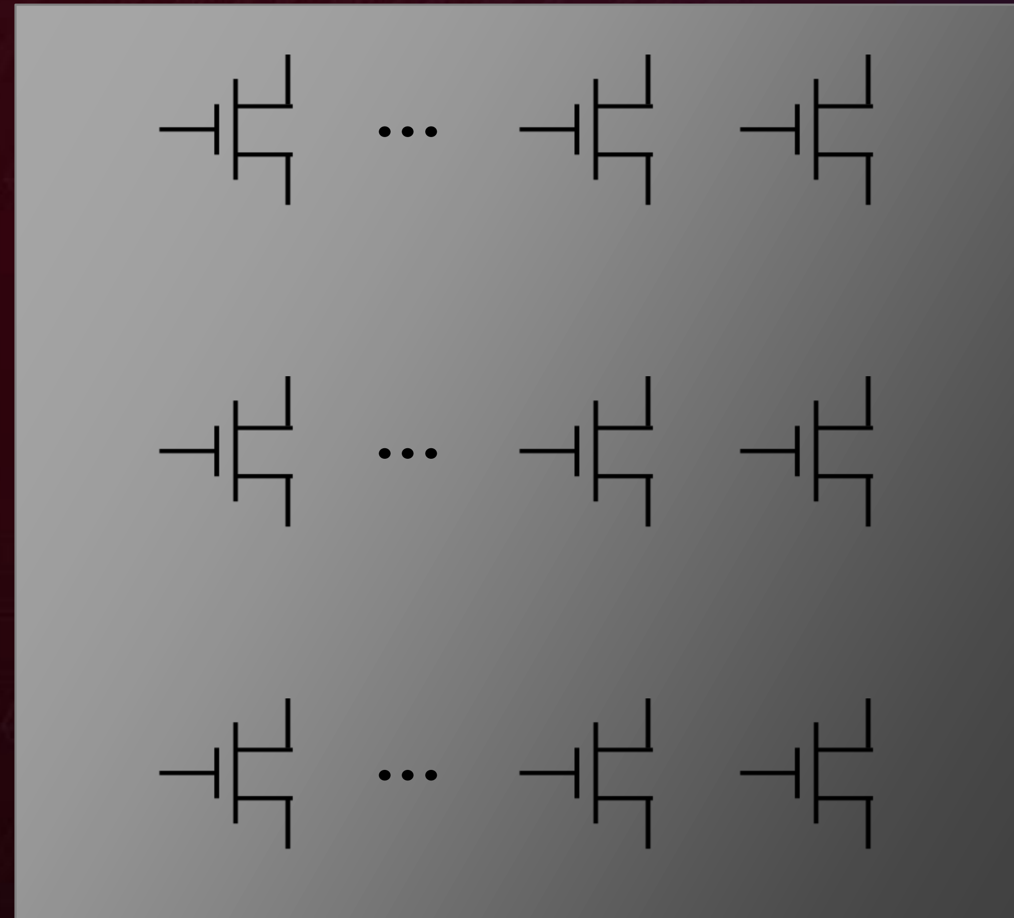
 TSMC Property

© 2019 TSMC, Ltd

tsmc

IMAGINE:
TRANSISTOR PERFORMANCE W/O DENSITY





© 2019 TSMC, Ltd

IMAGINE:
TRANSISTOR PERFORMANCE W/O DENSITY

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

© 2019 TSMC, Ltd

TECHNOLOGY **LEADERSHIP**

N7

World's first 7 nm

Participated in all the products on 7 nm

© 2019 TSMC, Ltd
tsmc

TECHNOLOGY **LEADERSHIP**

N7 N5(P)

Best performance
Highest density
Extensive EUV layers
Design ecosystem ready
In risk production

© 2019 TSMC, Ltd
tsmc

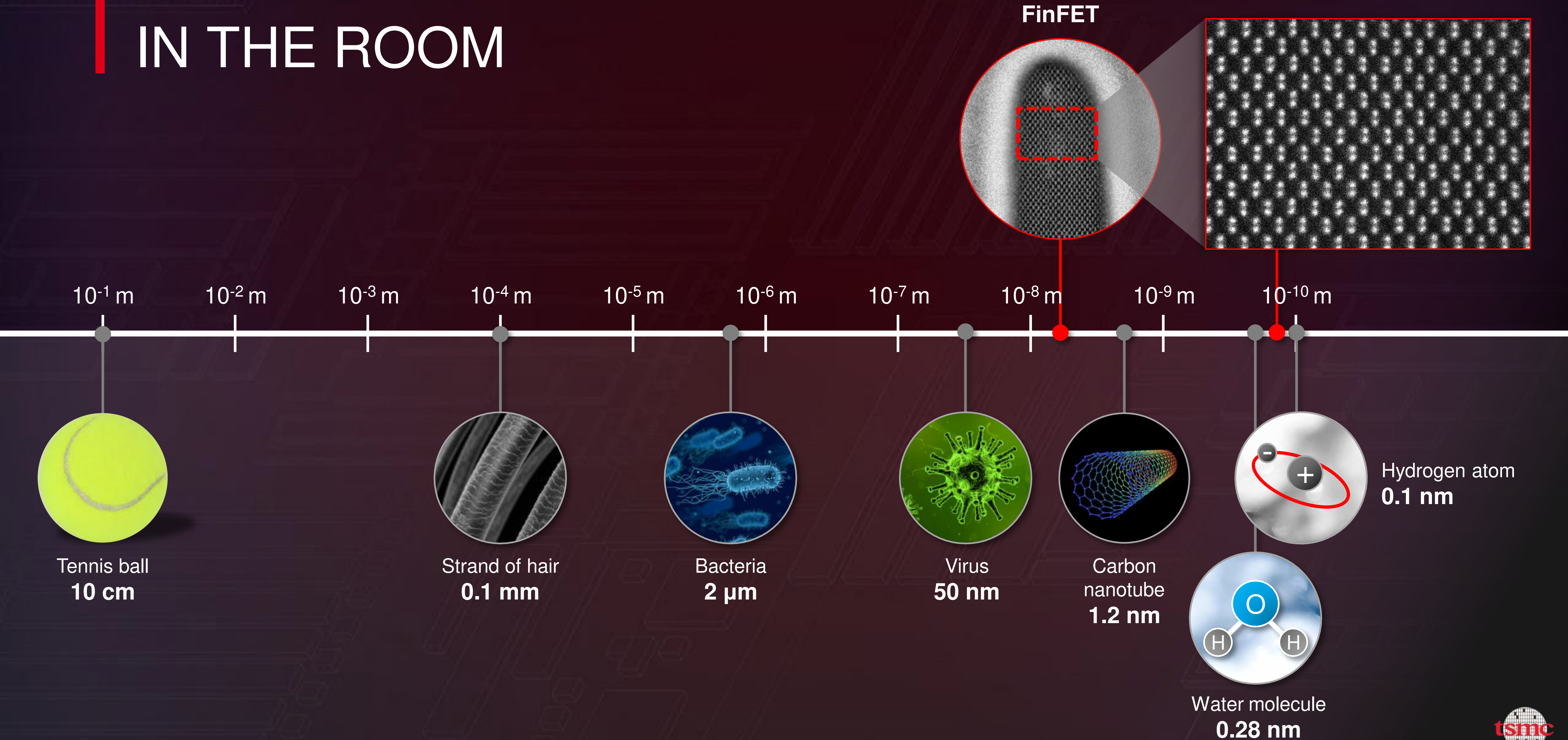
TECHNOLOGY **LEADERSHIP**

N7 N5(P) N3 …

2050 and beyond

© 2019 TSMC, Ltd

THE **ELEPHANT** IN THE ROOM



© 2019 TSMC, Ltd

MOORE'S LAW – A HISTORY OF INNOVATIONS

Dennard scaling

Strained Si, high-k / metal gate

FinFET / DTCO

CONTINUOUS BENEFITS NODE AFTER NODE

© 2019 TSMC, Ltd

MULTIPLE ROADS LEAD TO ROME

Innovations

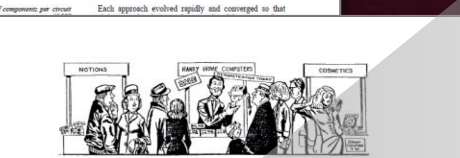
CONTINUOUS BENEFITS NODE AFTER NODE

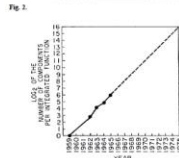
© 2019 TSMC, Ltd

INTEGRATING CHIPS INTO SYSTEMS

Cramming More Components onto Integrated Circuits

GORDON E. MOORE, LIFE FELLOW, IEEE



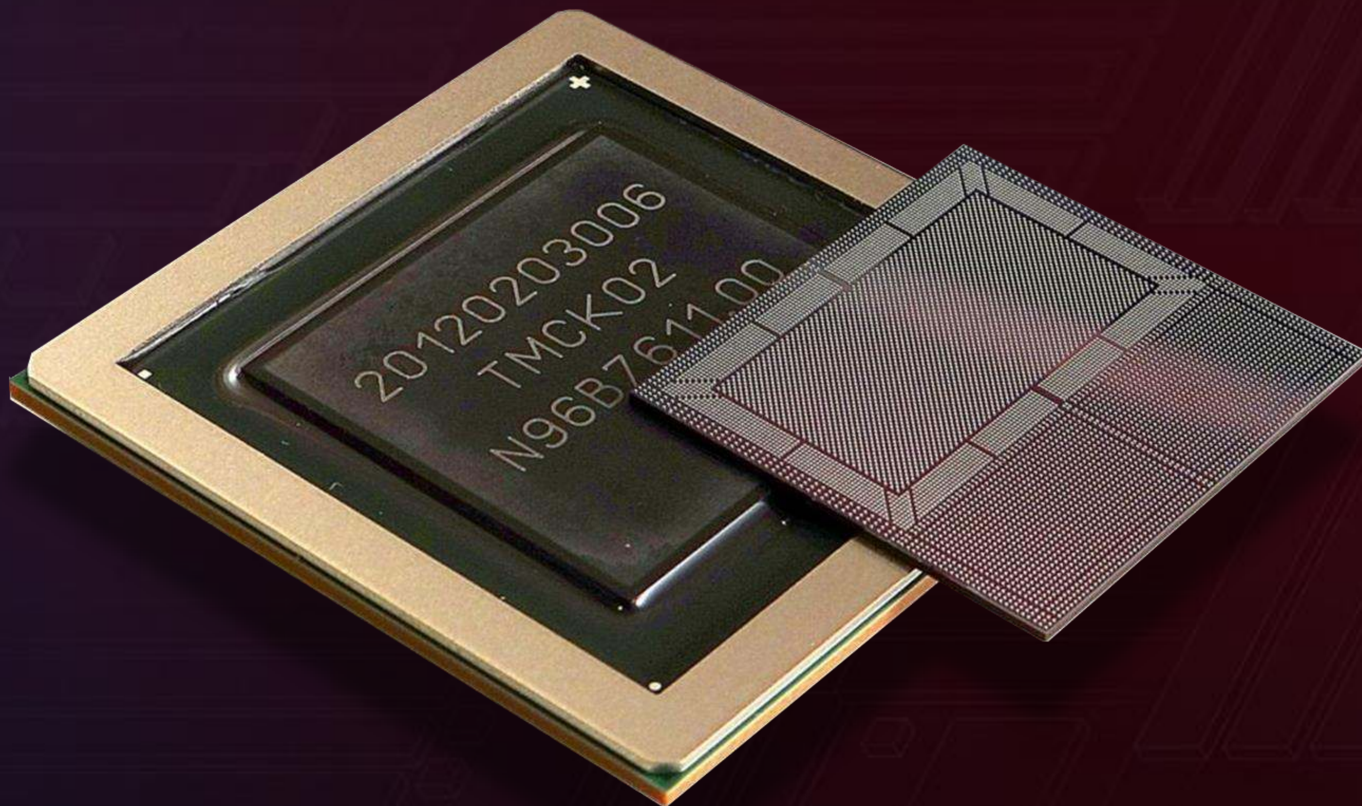


It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected. The availability of large functions, combined with functional design and construction, should allow the manufacturer of large systems to design and construct a considerable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

Source: G. Moore, Electronics, 1965

© 2019 TSMC, Ltd

CoWoS® SYSTEM INTEGRATION



TSMC CoWoS® fully assembled test chip

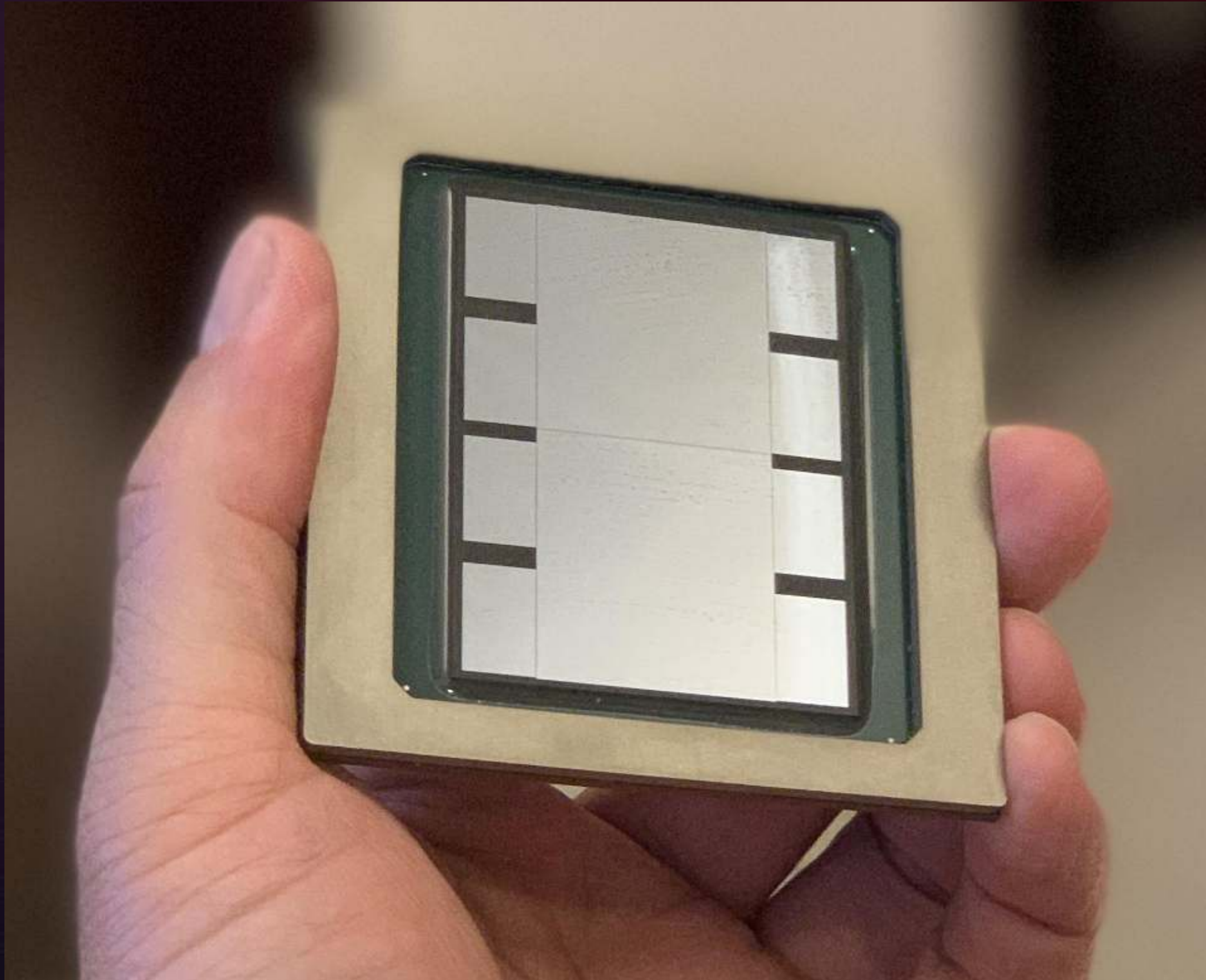
1 SoC + 2 DRAMs

Source: 2013 TSMC Technology Symposium

© 2019 TSMC, Ltd

tsmc

CoWoS® SYSTEM INTEGRATION

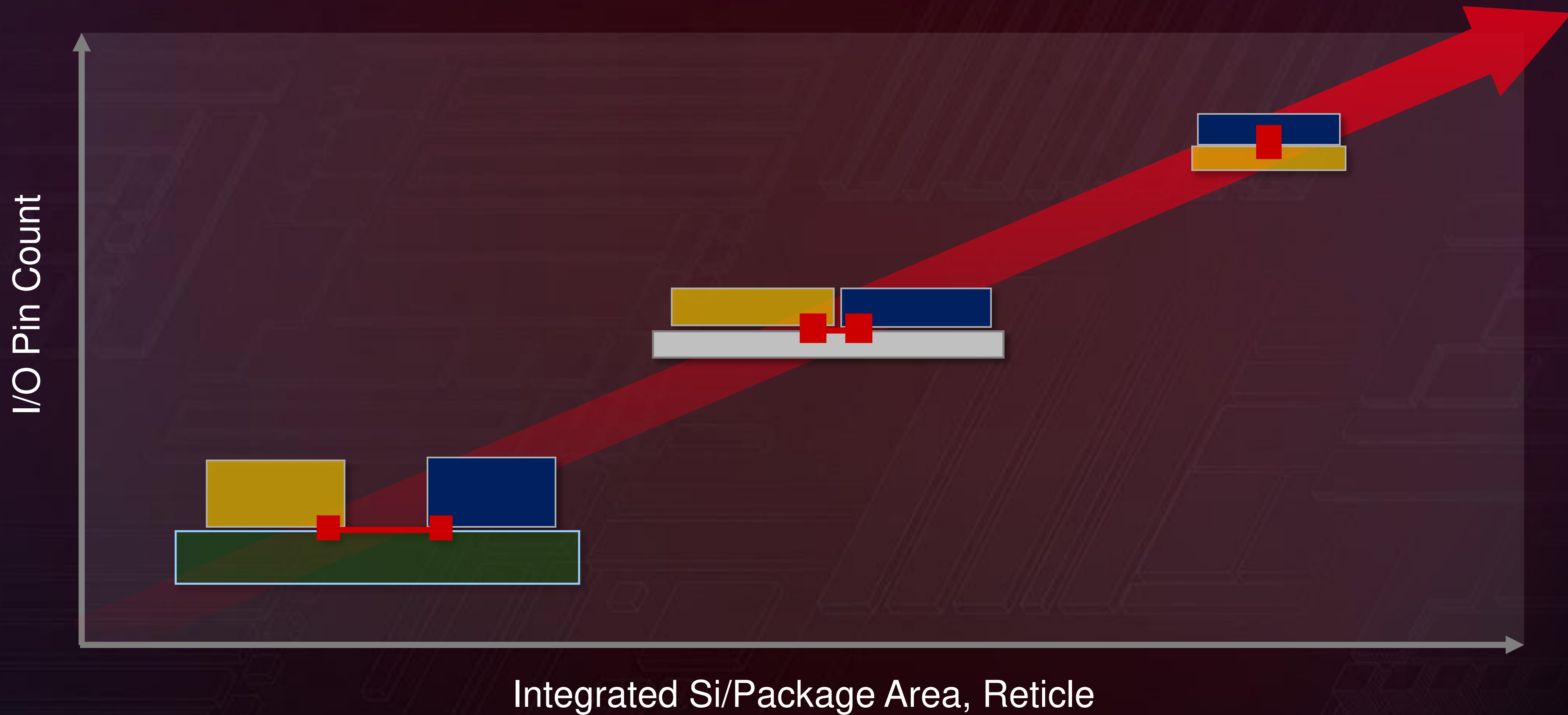


2500 mm^2 interposer:

- 2 processors (600 mm^2)
- + 8 HBM DRAM

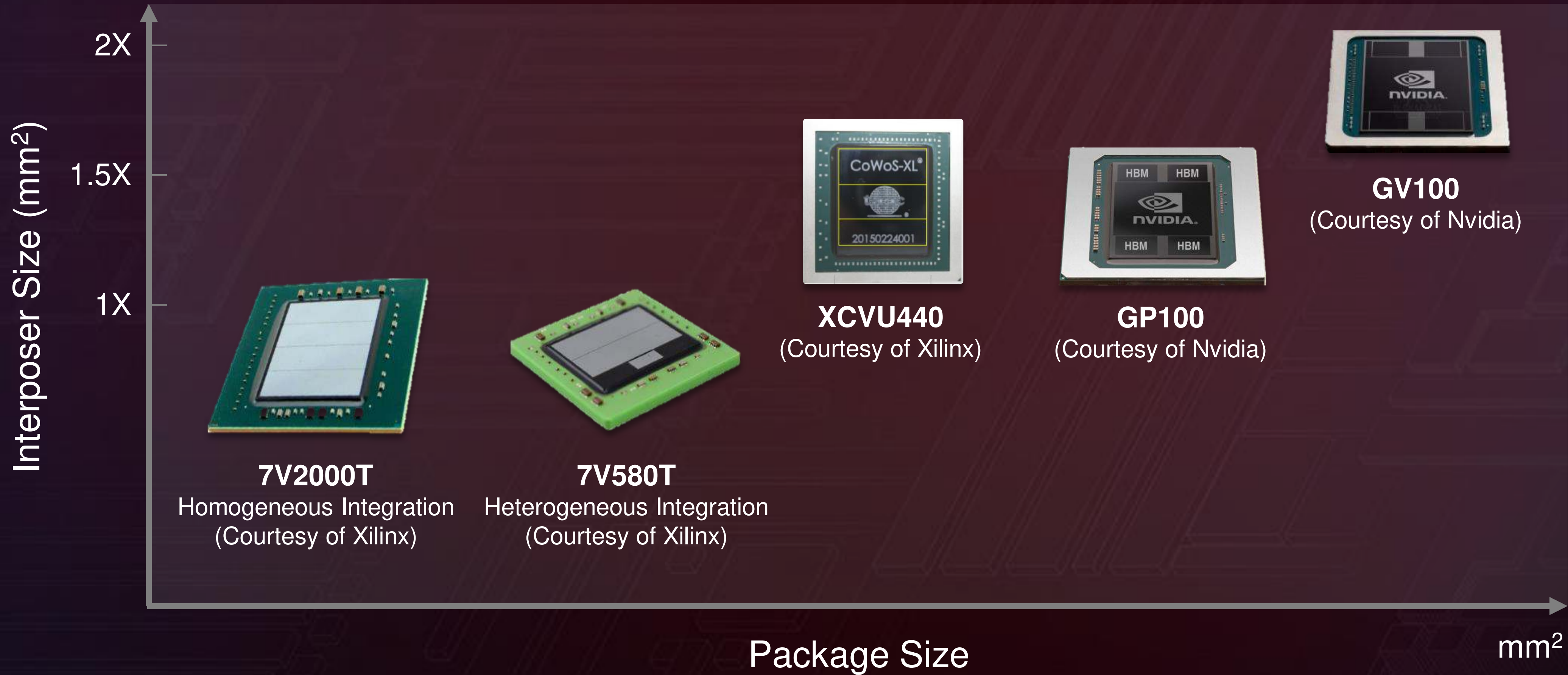
© 2019 TSMC, Ltd
tsmc

SYSTEM INTEGRATION TECHNOLOGIES



© 2019 TSMC, Ltd

CHIPLETS INTEGRATION **REDUCES** SYSTEM COST PER FUNCTION



© 2019 TSMC, Ltd

tsmc

SEMICONDUCTOR TECHNOLOGY EVOLVES

DRIVEN BY CHANGING APPLICATION LANDSCAPE

2050 and beyond

5nm CMOS
2020

7nm FinFET
2018

FinFET
1999

Transistor Scaling Principle
1974

Invention of point-contact transistor
1947

Invention of IC
1958

4G
2009

GPU (21B Transistors)
2017

Pentium CPU
1995

3G
2002

iPhone
2007

Mobile phone
1973

Flash Memory
1984

Intel 4004
1971

Transistor Radio

Mini-Computer

PC / Internet

Mobile

AI / 5G

© 2019 TSMC, Ltd

DATA MOVEMENT HITS THE MEMORY WALL

ABUNDANT-DATA APPLICATIONS: ENERGY MEASUREMENTS

AlexNet (CNN): 15% | 85%

ResNet-152 (CNN): 20% | 80%

Language Model (LSTM): 8% | 92%

• • •

Memory Compute

Deep Learning Accelerators

Source: S. Mitra (Stanford)

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

© 2019 TSMC, Ltd

DEEP NEURAL NETWORKS REQUIRE LARGE MEMORY CAPACITY

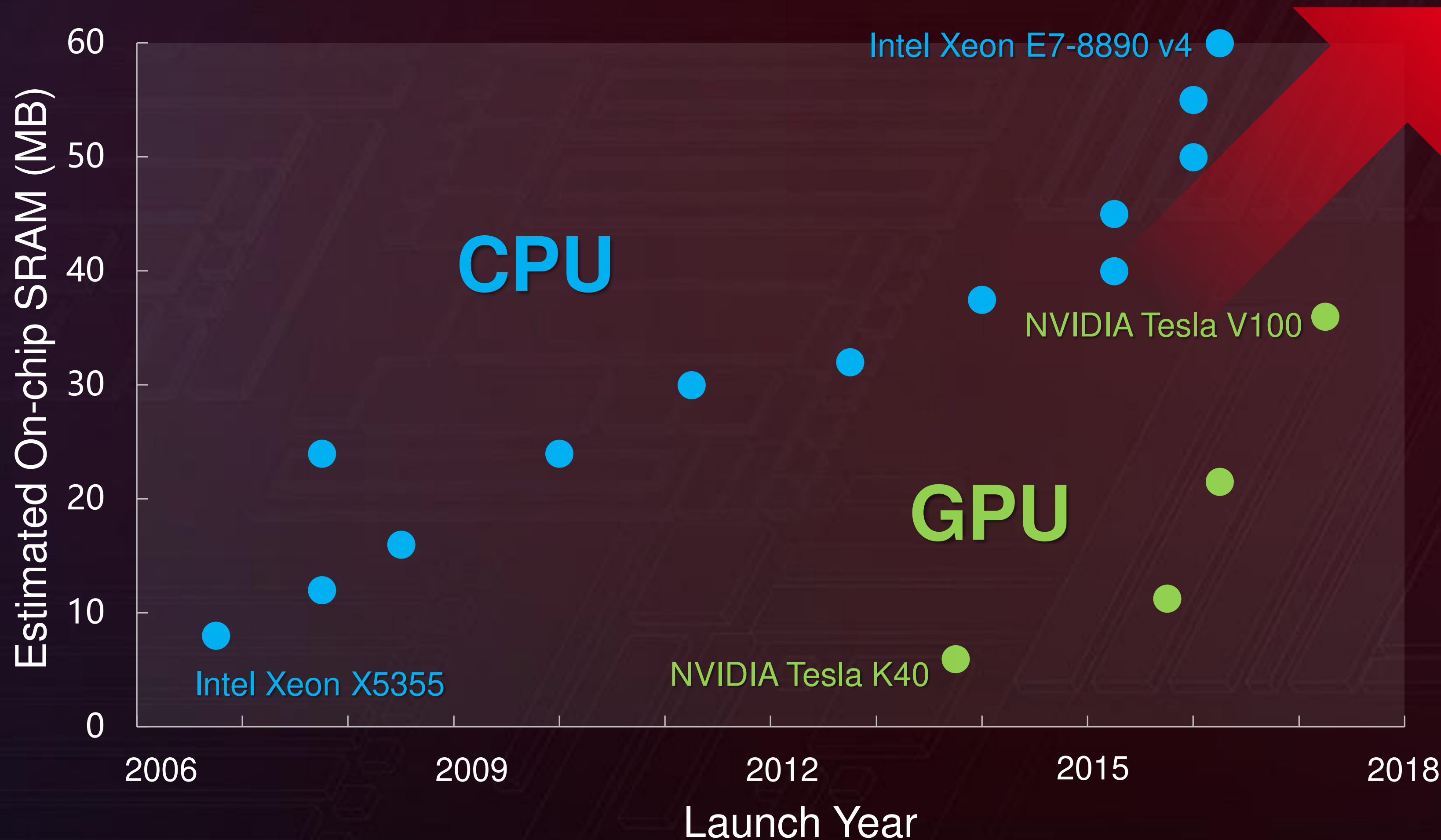
Network (application)	Type (LSTM/ CNN)	Training/ Inference	Model Size	Memory Usage (GBytes)
ResNet (vision)	CNN	Training	120 MBytes	21*
		Inference		0.12
Language Model (NLP)	LSTM	Training	2.5 GBytes	40*
		Inference		2.5

* 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)

© 2019 TSMC, Ltd
tsmc

ON-CHIP SRAM CAPACITY:
NEVER ENOUGH



3.8 Gbytes @ 1.4 nm node

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

© 2019 TSMC, Ltd

tsmc

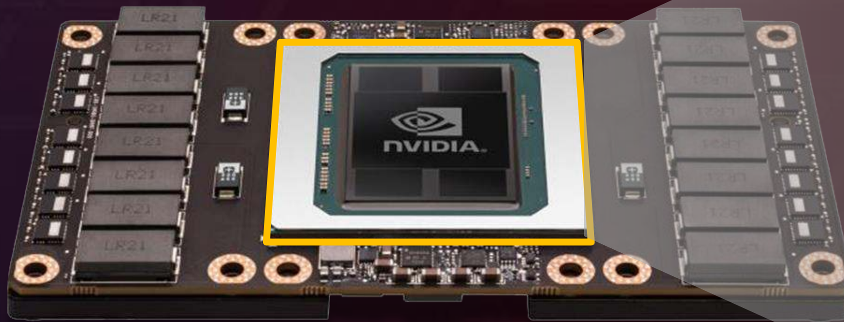
CAN WE PUT LOTS OF **MEMORY ON-CHIP**?

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

© 2019 TSMC, Ltd
tsmc

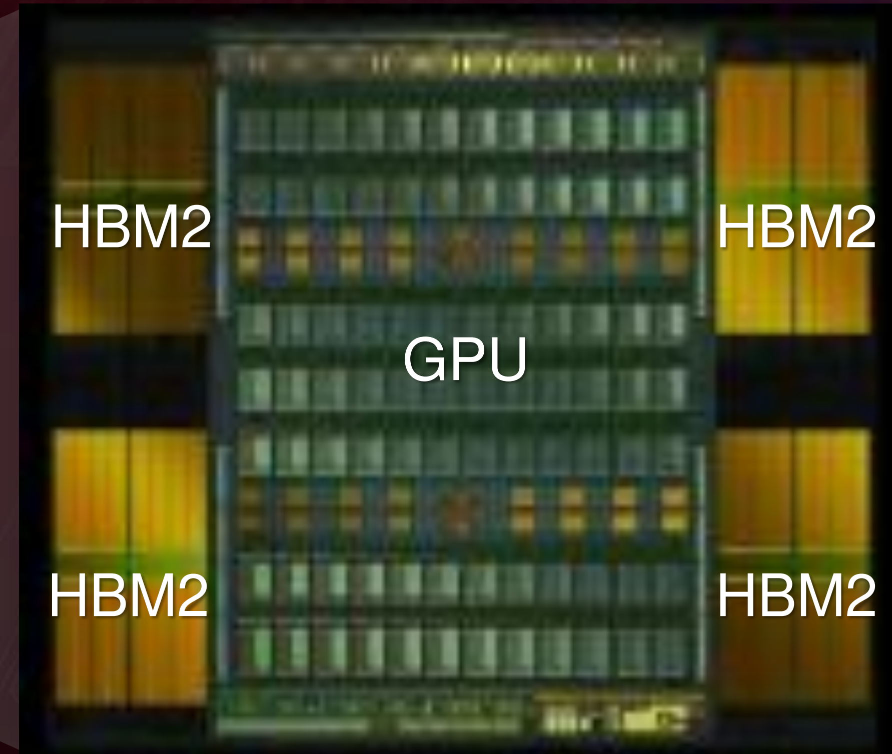
SUPER AI ACCELERATOR
ENABLED BY CoWoS®

CoWoS Module



Superior processing power that equals to 100 CPUs

Heterogeneous Integration: GPU + High Bandwidth Memory (HBM2)



>300 B transistors

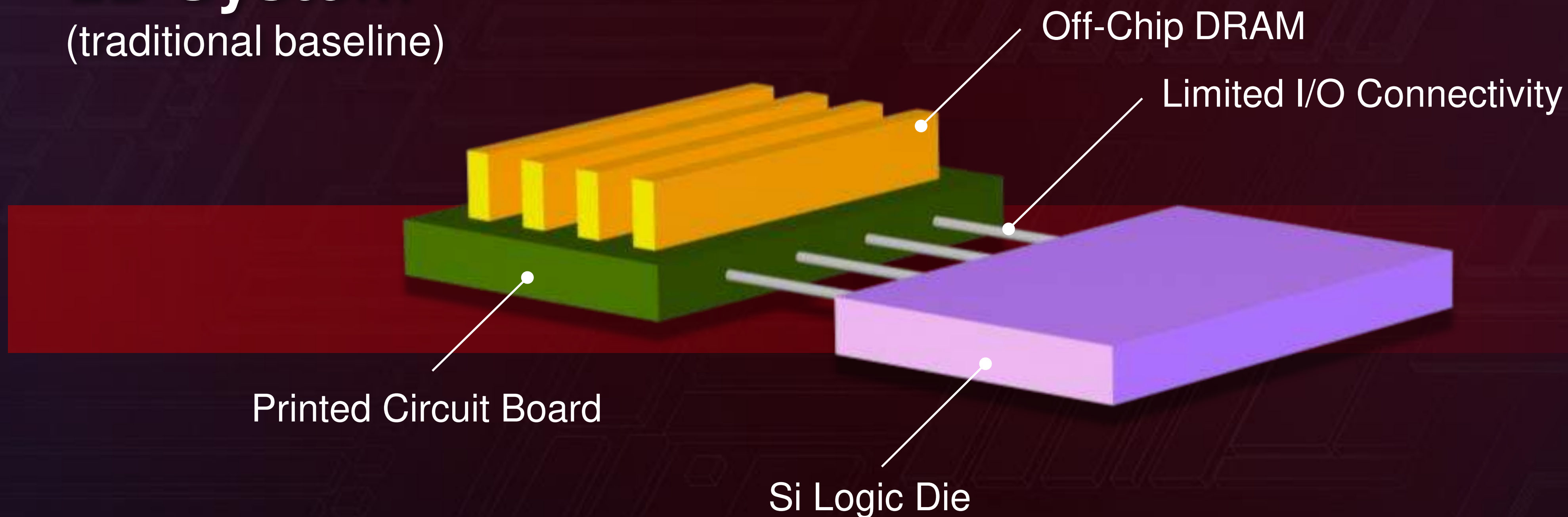
Source: “Inside Volta” , Nvidia GPU Tech. Conf. , May 10, 2017.

© 2019 TSMC, Ltd

COMPUTE-MEMORY INTEGRATION

2D System

(traditional baseline)

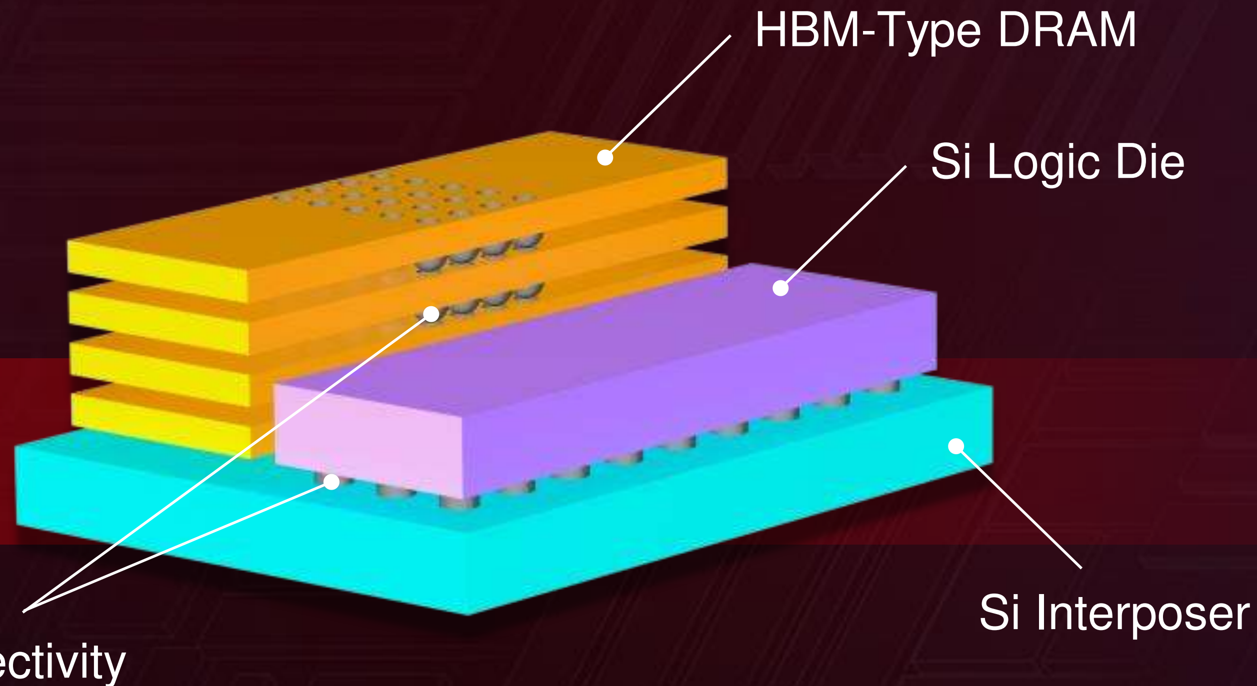


Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

© 2019 TSMC, Ltd

COMPUTE-MEMORY INTEGRATION

2.5D System

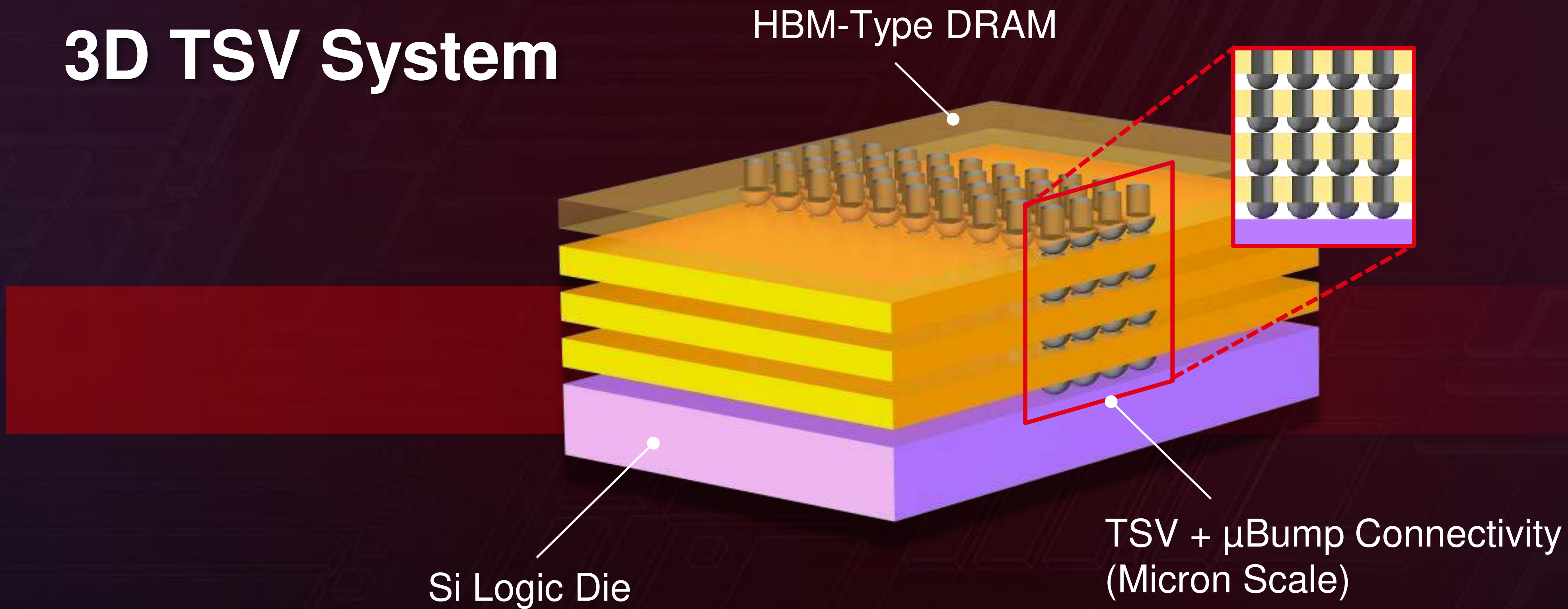


Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

© 2019 TSMC, Ltd

COMPUTE-MEMORY INTEGRATION

3D TSV System

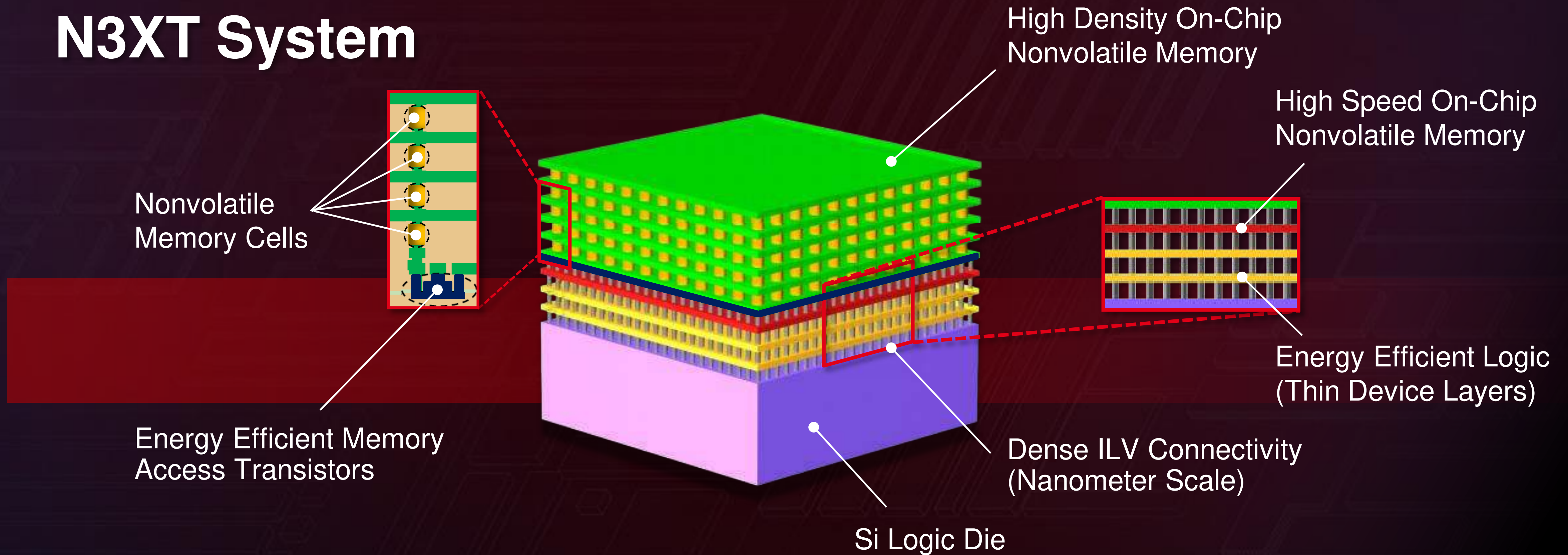


Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

© 2019 TSMC, Ltd

COMPUTE-MEMORY INTEGRATION

N3XT System

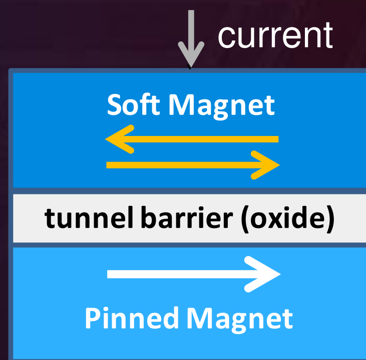


Source: W. Hwang, W. Wan, Y. Malviya, H. Li, M. Lee, M. Aly, H.-S. P. Wong, S. Mitra. Work in progress 2017 – 2019 w/ TSMC

© 2019 TSMC, Ltd

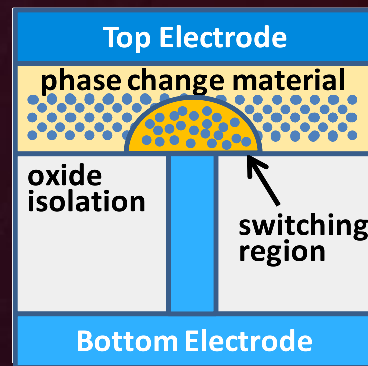
"NEW" MEMORIES FOR COMPUTE-MEMORY INTEGRATION

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



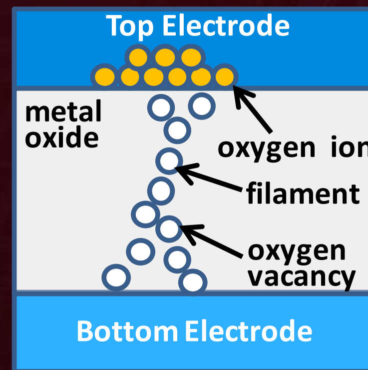
STT-MRAM

Spin torque transfer magnetic random access memory



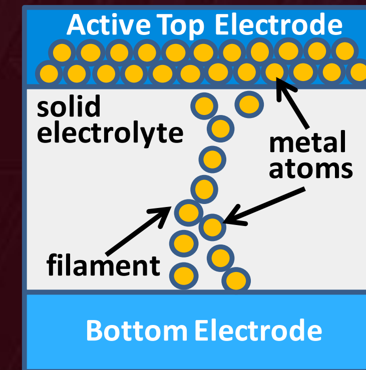
PCM

Phase change memory



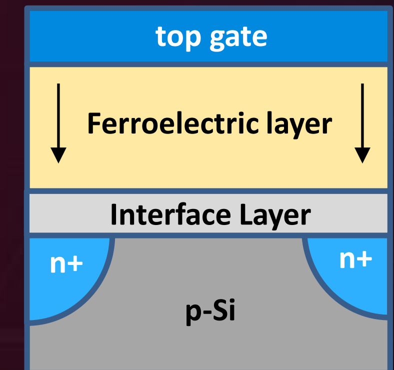
RRAM

Resistive switching random access memory



CBRAM

Conductive bridge random access memory



FERAM

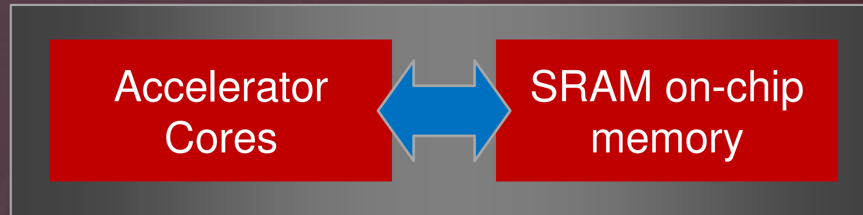
Ferro-electric random access memory

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

© 2019 TSMC, Ltd

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

2D baseline system

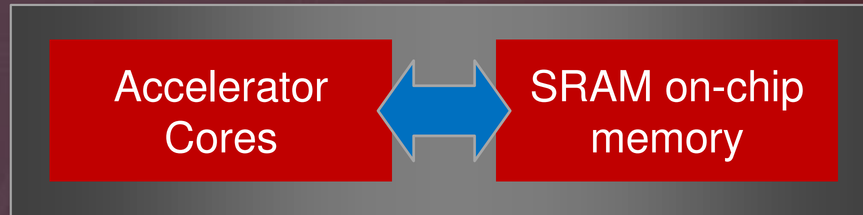


Source: Stanford/NTU: M. Aly, S. Mitra, TSMC: Yih (Eric) Wang, K. Akarvardar, 2019

© 2019 TSMC, Ltd

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

2D baseline system



Off-chip DRAM (LPDDR3)

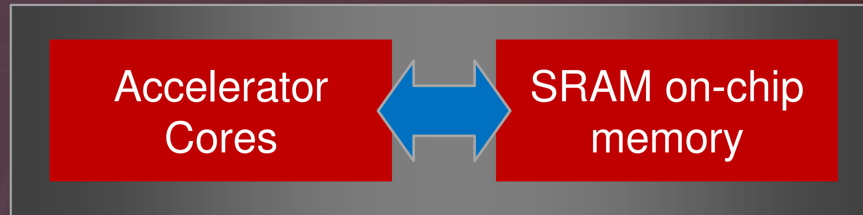
- Capacity: 4 GBytes
- Latency: 50 ns
- BW: 12 GBytes/s
- Read/write energy: 17 pJ/bit

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

© 2019 TSMC, Ltd

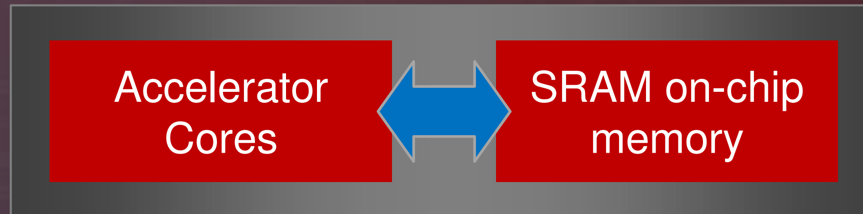
NEW MEMORY: HIGH-BANDWIDTH, HIGH-CAPACITY, ON-CHIP

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



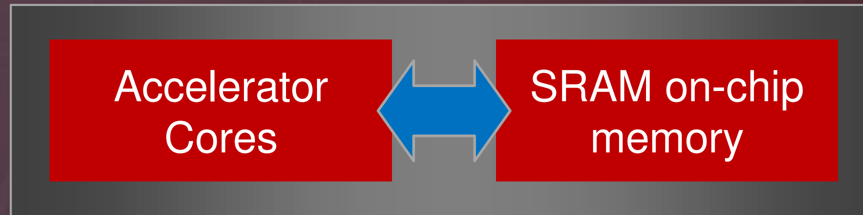
Off-chip DRAM (LPDDR3)
- Capacity: (4 GBytes minus New Mem. Cap.)
- Latency: 50 ns
- BW: 12 GBytes/s
- Read/write energy: 17 pJ/bit

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

© 2019 TSMC, Ltd

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

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

High Bandwidth, High Capacity
both critical

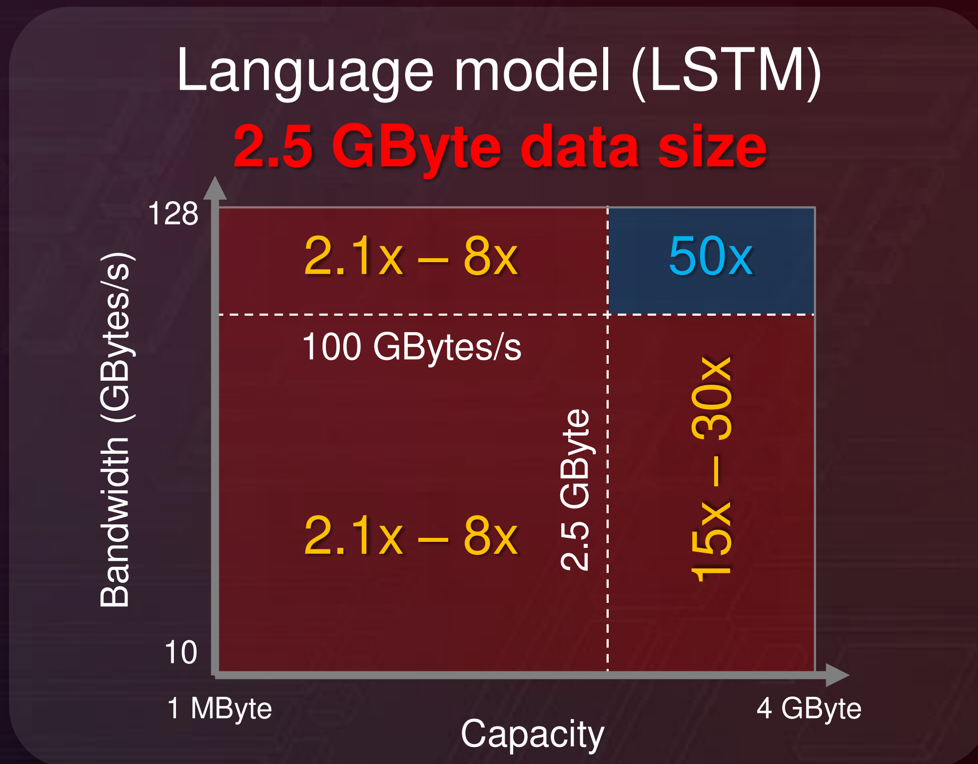
Source: Stanford/NTU: M. Aly, S. Mitra, TSMC: Yih (Eric) Wang, K. Akarvardar, 2019

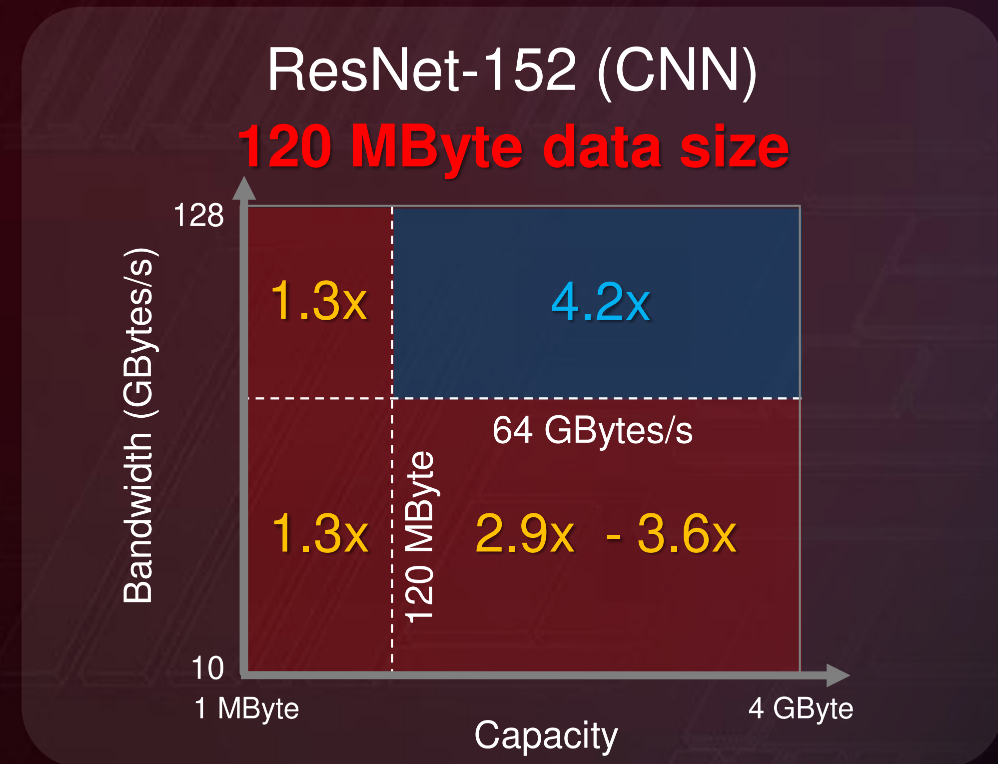
© 2019 TSMC, Ltd

NEW MEMORY ESSENTIAL REQUIREMENT

ON-CHIP CAPACITY MUST EXCEED DATA SIZE

EDP benefits





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

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

© 2019 TSMC, Ltd

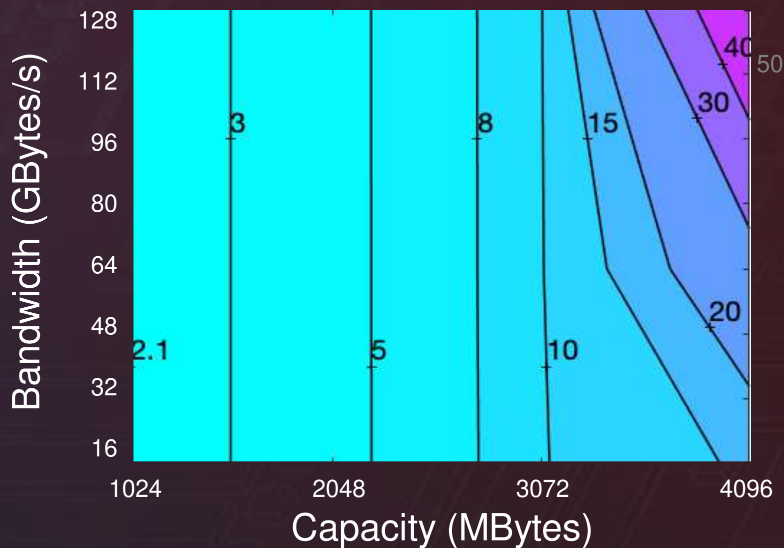
NEW MEMORY ESSENTIAL REQUIREMENT

ON-CHIP CAPACITY MUST EXCEED DATA SIZE

EDP benefits

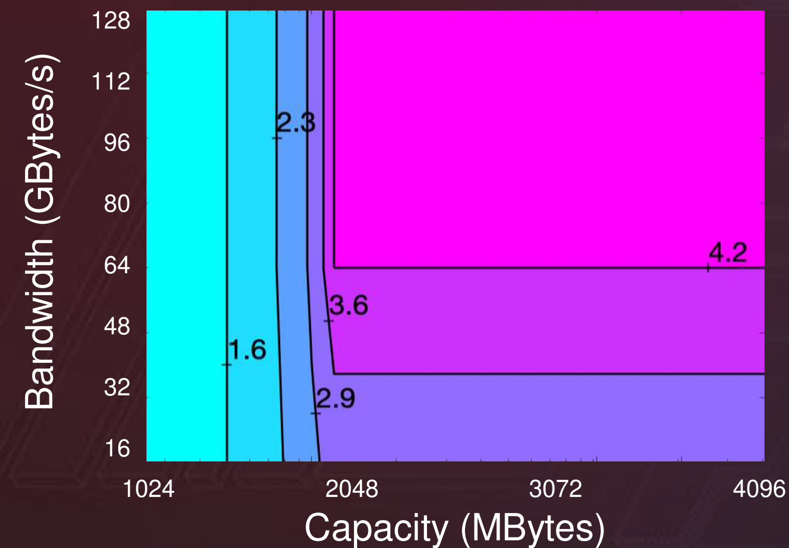
Language model (LSTM)

2.5 GByte data size



ResNet-152 (CNN)

120 MByte data size



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

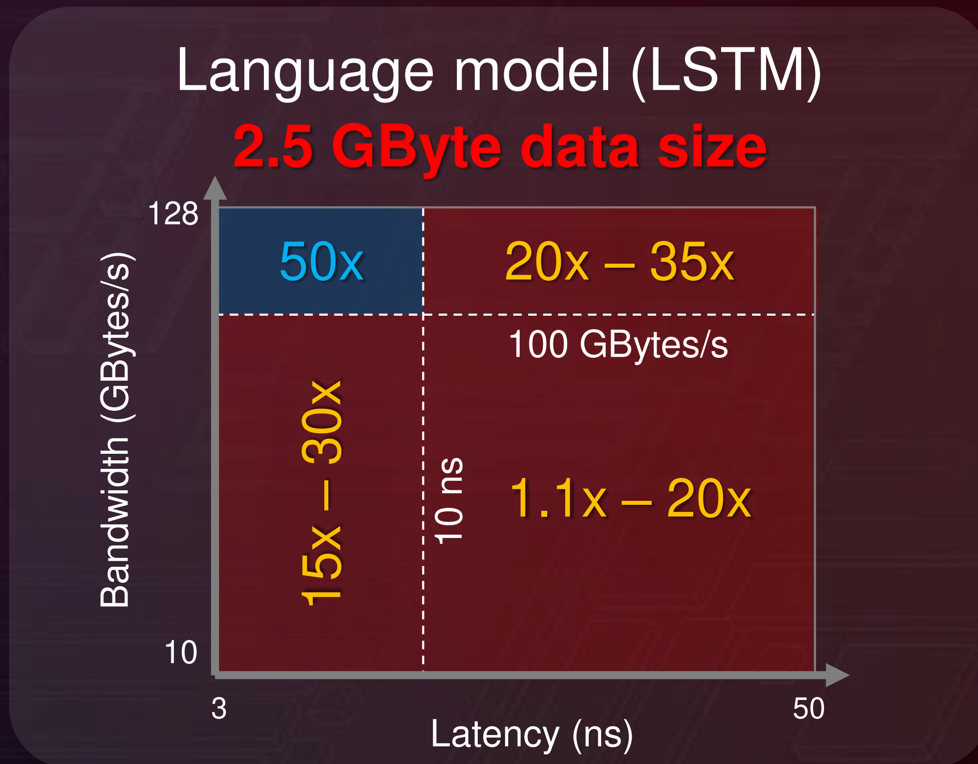
Source: Stanford/NTU: M. Aly, S. Mitra, TSMC: Yih (Eric) Wang, K. Akarvardar, 2019

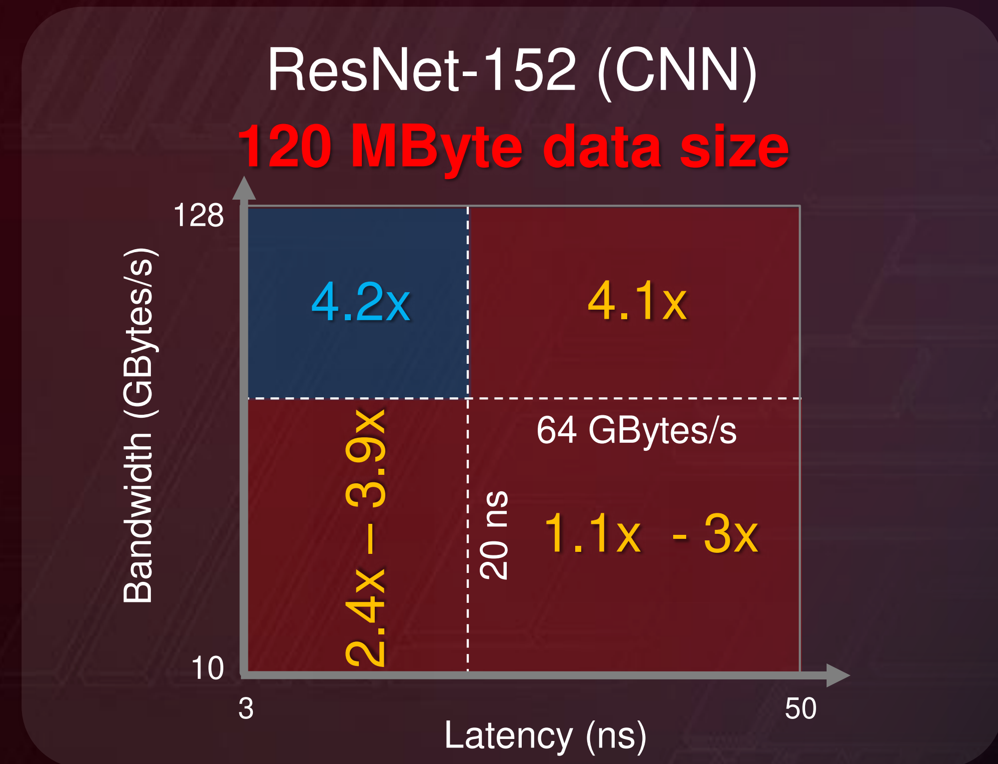
© 2019 TSMC, Ltd

NEW MEMORY ESSENTIAL REQUIREMENT

HIGH BANDWIDTH MORE CRITICAL THAN LATENCY

EDP benefits





5 pJ/bit access energy

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

© 2019 TSMC, Ltd

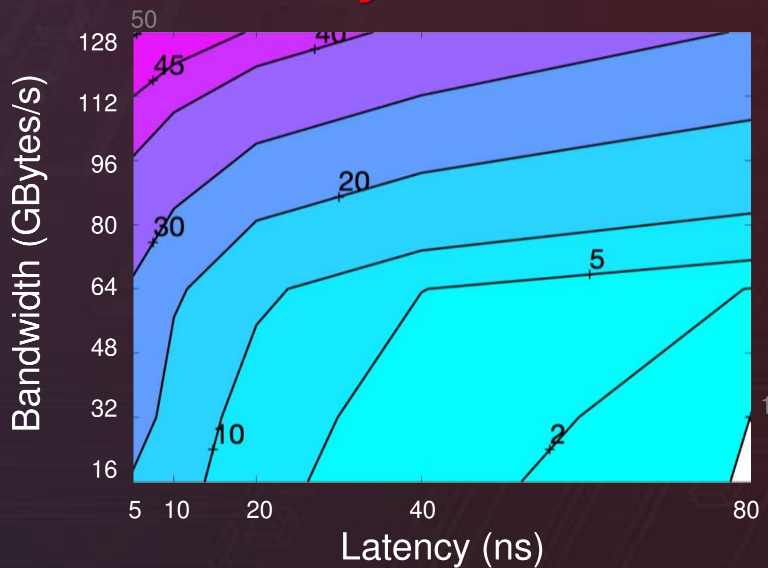
NEW MEMORY ESSENTIAL REQUIREMENT

HIGH BANDWIDTH MORE CRITICAL THAN LATENCY

EDP benefits

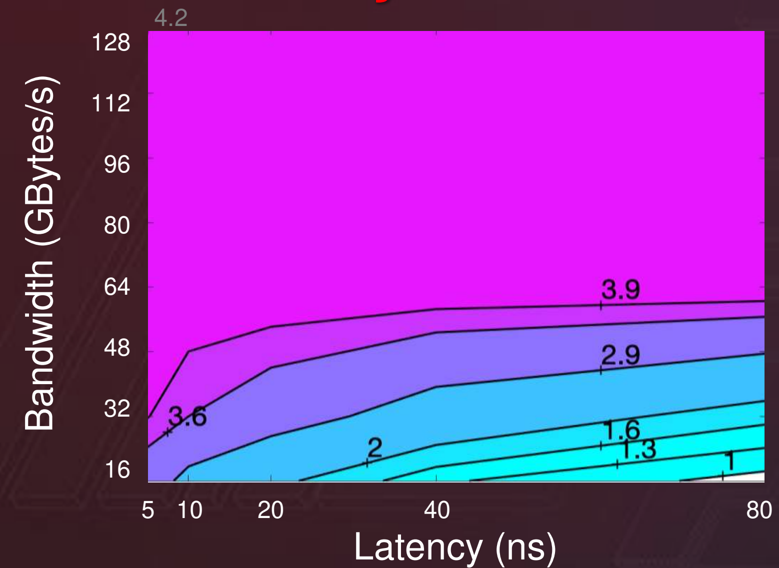
Language model (LSTM)

2.5 GByte data size



ResNet-152 (CNN)

120 MByte data size



5 pJ/bit access energy

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

© 2019 TSMC, Ltd

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

Energy × Execution Time

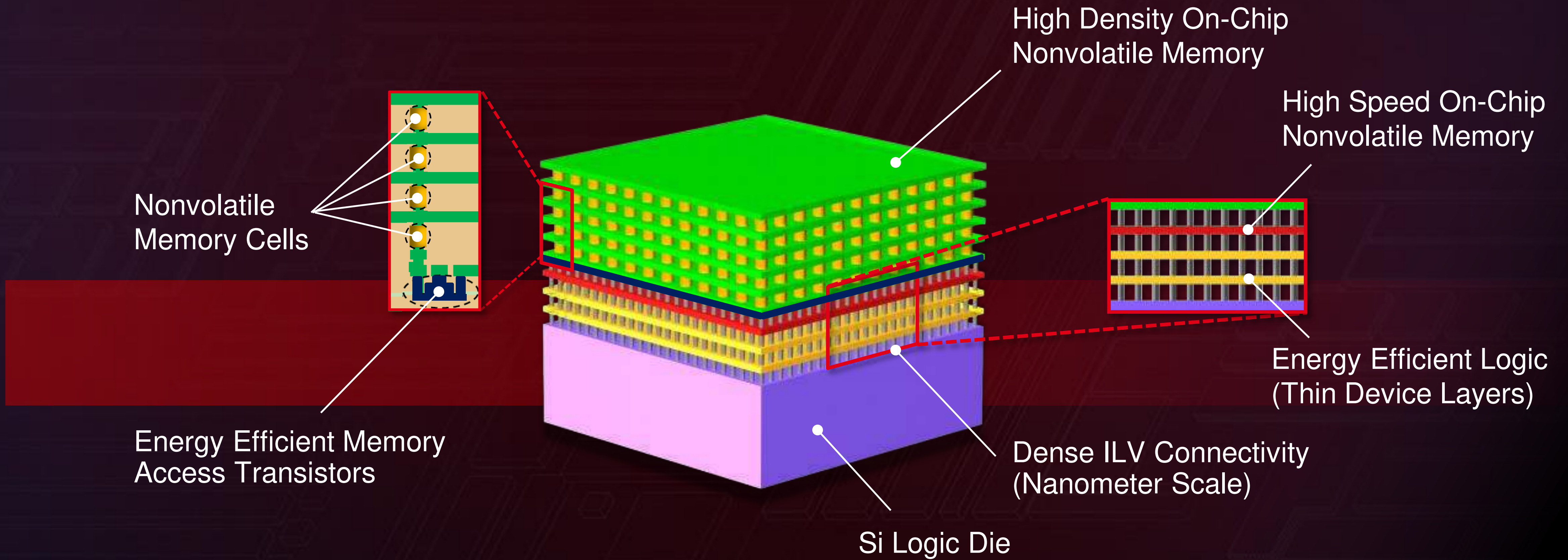
Workload: Inference on ML Accelerator	System-Level Benefits
Lang.Model (LSTM)	1971X
AlexNet (CNN)	525X
Captioning (LSTM)	320X
ResNet152 (CNN)	159X
VGG19 (CNN)	63X

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

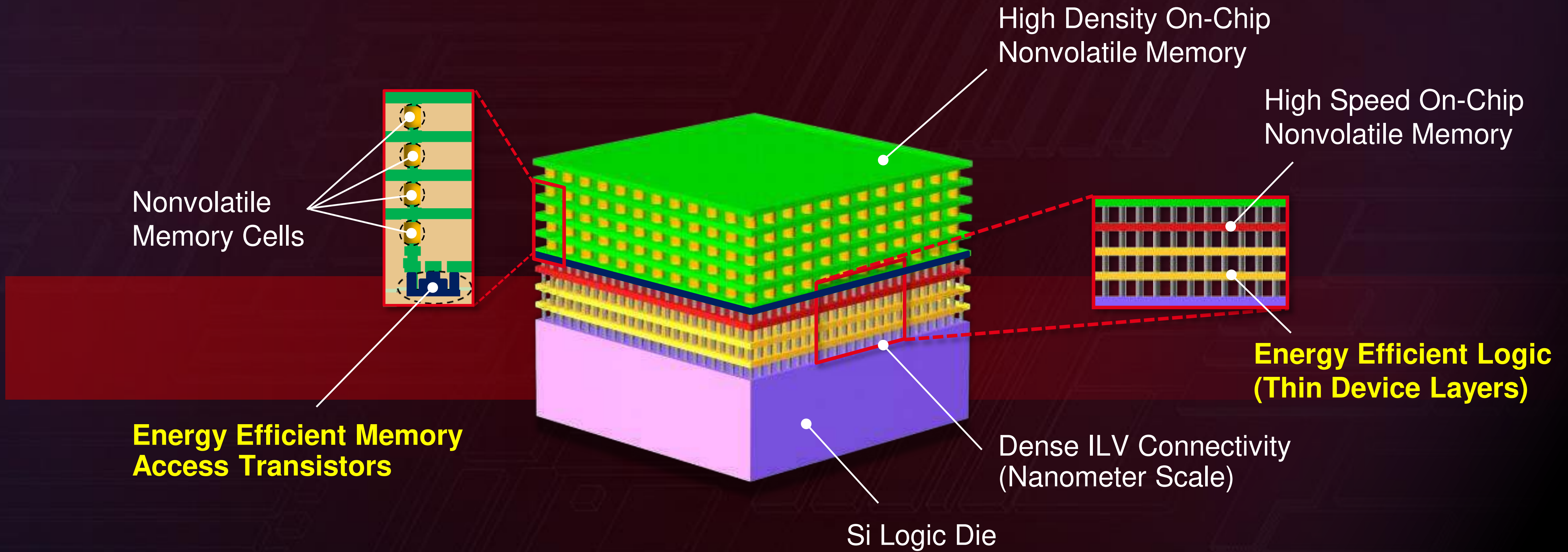
© 2019 TSMC, Ltd

N3XT SYSTEM



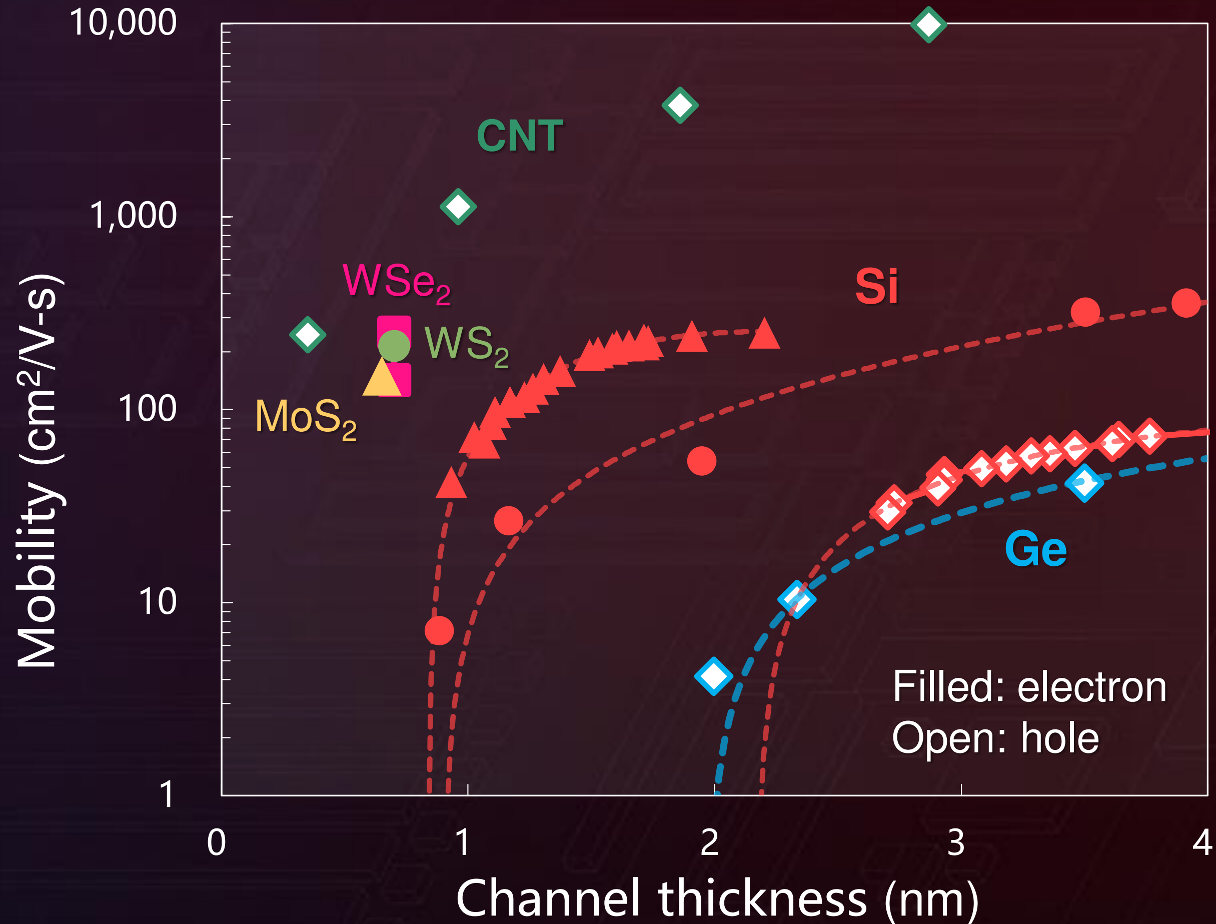
© 2019 TSMC, Ltd

N3XT SYSTEM



© 2019 TSMC, Ltd
tsmc

NANOMETER-THIN TRANSISTOR CHANNEL



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

1D carbon nanotube (CNT)

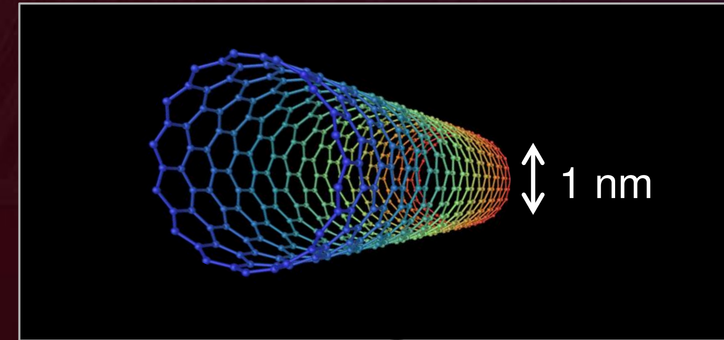


Photo credit: User Mstroeck on en.wikipedia

2D TMD (MoS_2, WSe_2, WS_2…)

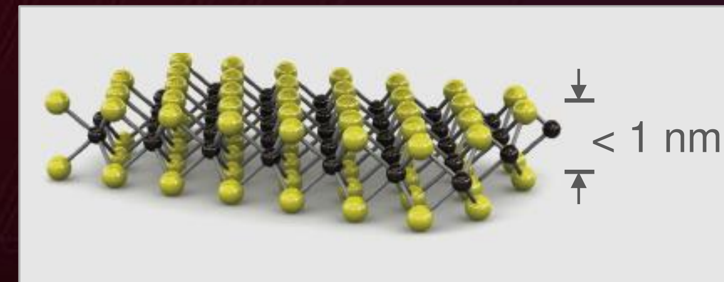
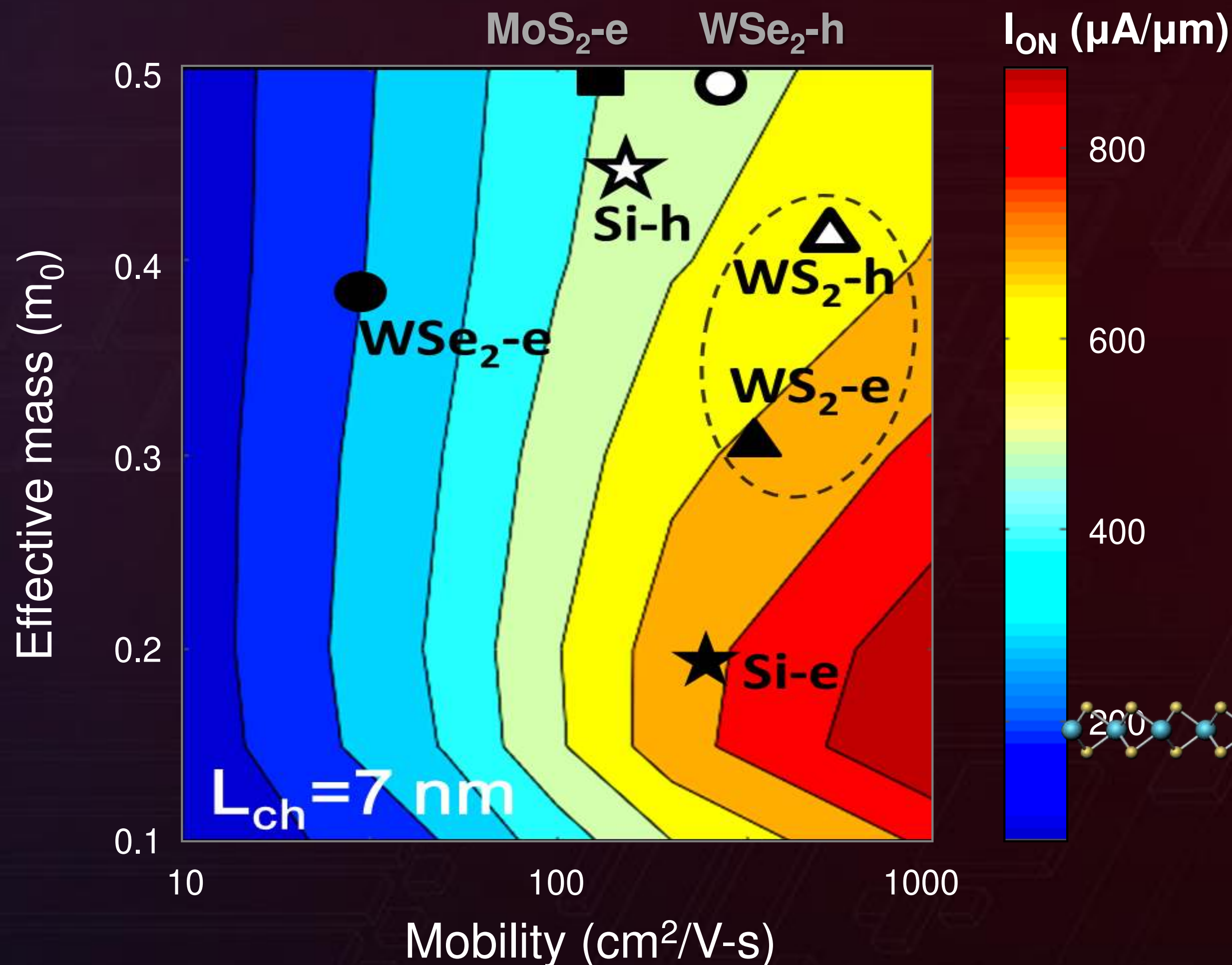
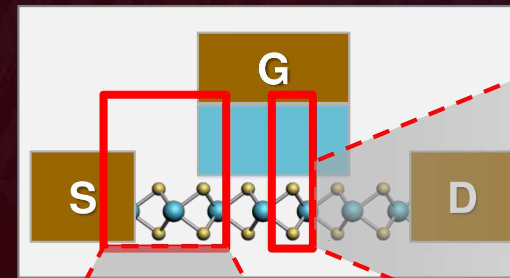


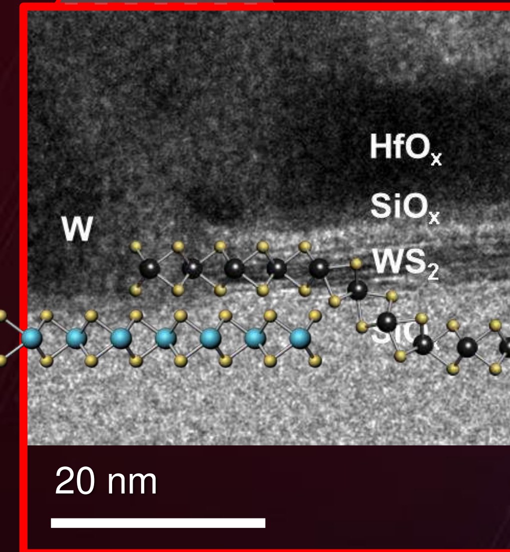
Photo credit: B. Radisavljevic et al., Nature Nanotech., p. 147, 2011

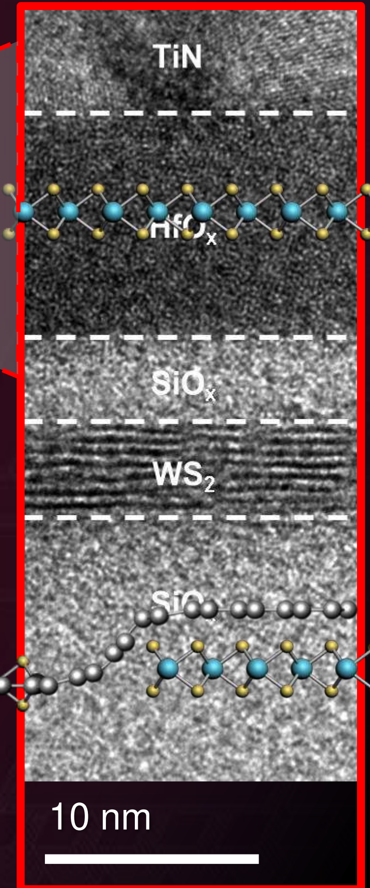
© 2019 TSMC, Ltd

2D LAYERED MATERIALS (WS_2, WSe_2)







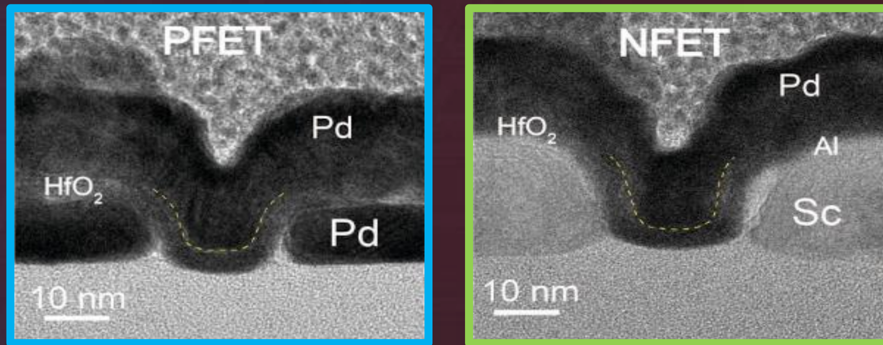


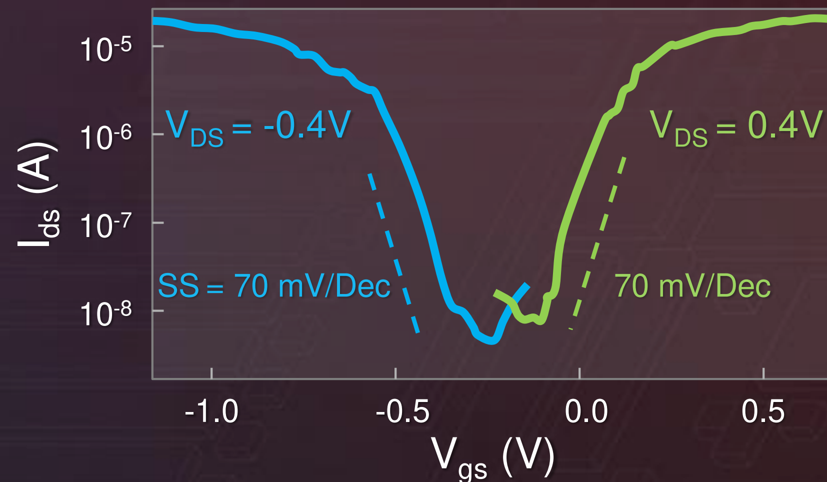
Source: C.-C. Cheng et al. (TSMC), Symp. VLSI Tech. 2019

© 2019 TSMC, Ltd

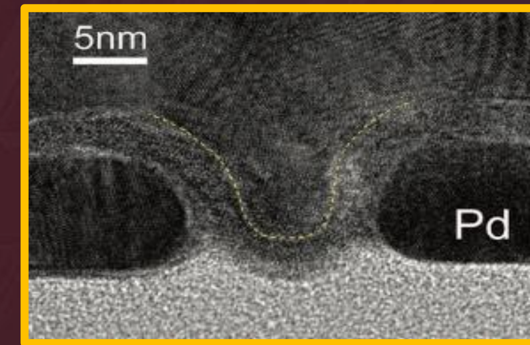
SHORT-CHANNEL
CARBON NANOTUBE TRANSISTORS

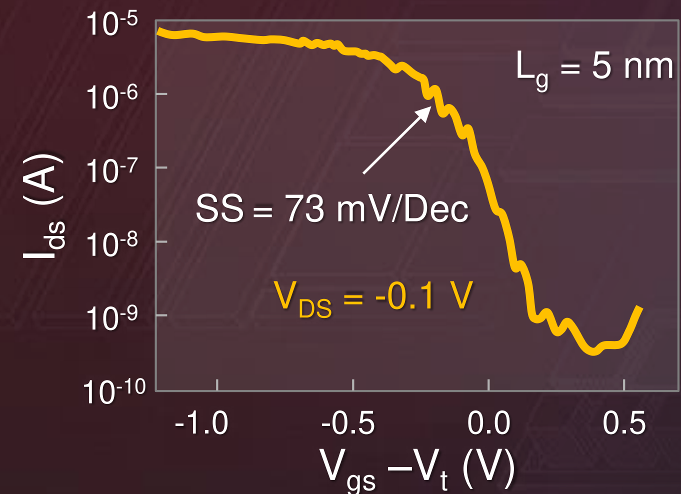
10 nm Gate Length





5 nm Gate Length

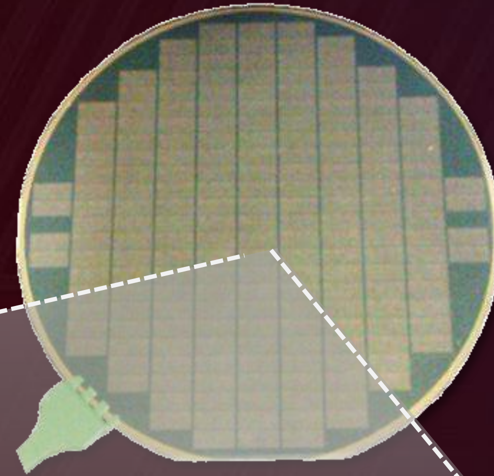


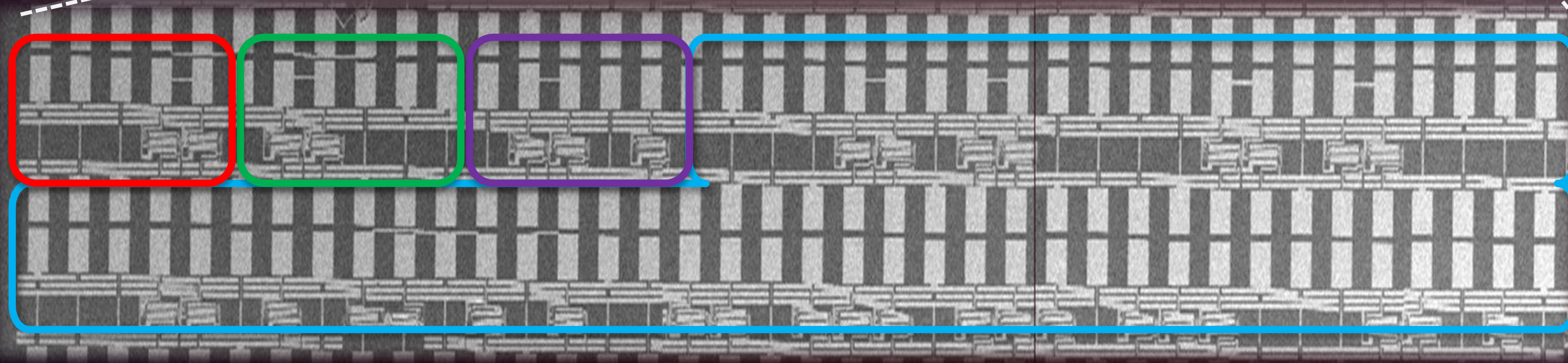


Source: C. Qiu,…L-M. Peng (PKU), Science, 2017

© 2019 TSMC, Ltd

CARBON NANOTUBE COMPUTER





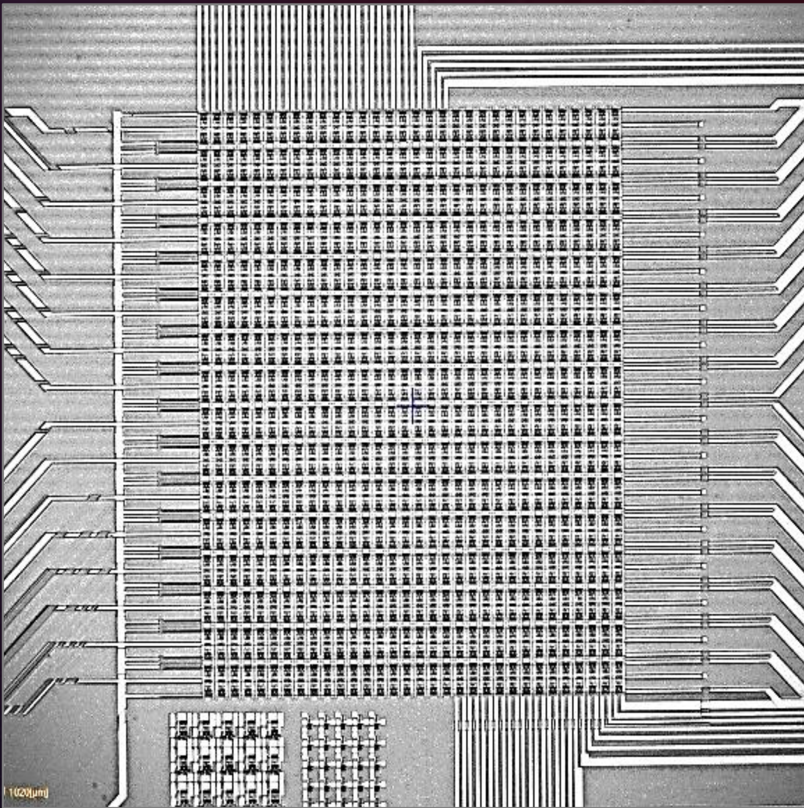
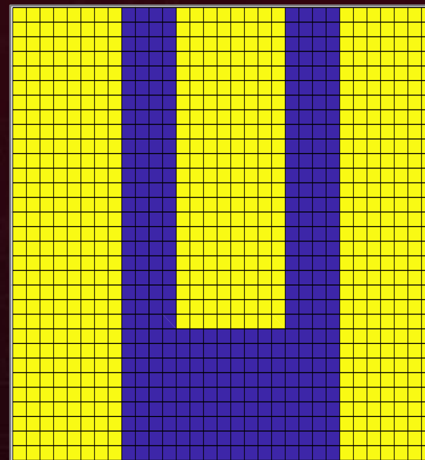
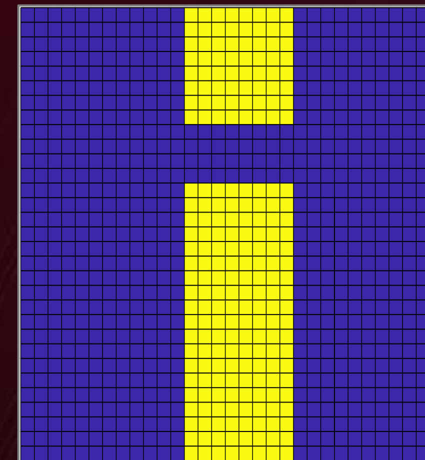
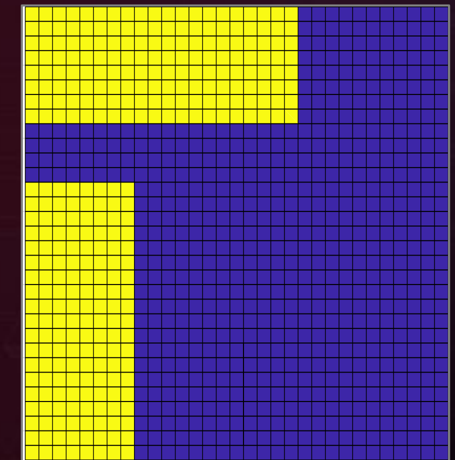
- instruction fetch
- data fetch
- arithmetic block
- write-back

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

© 2019 TSMC, Ltd

CARBON NANOTUBE FET CMOS SRAM

Kbit 6T SRAM (6144 CNFETs)

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

© 2019 TSMC, Ltd

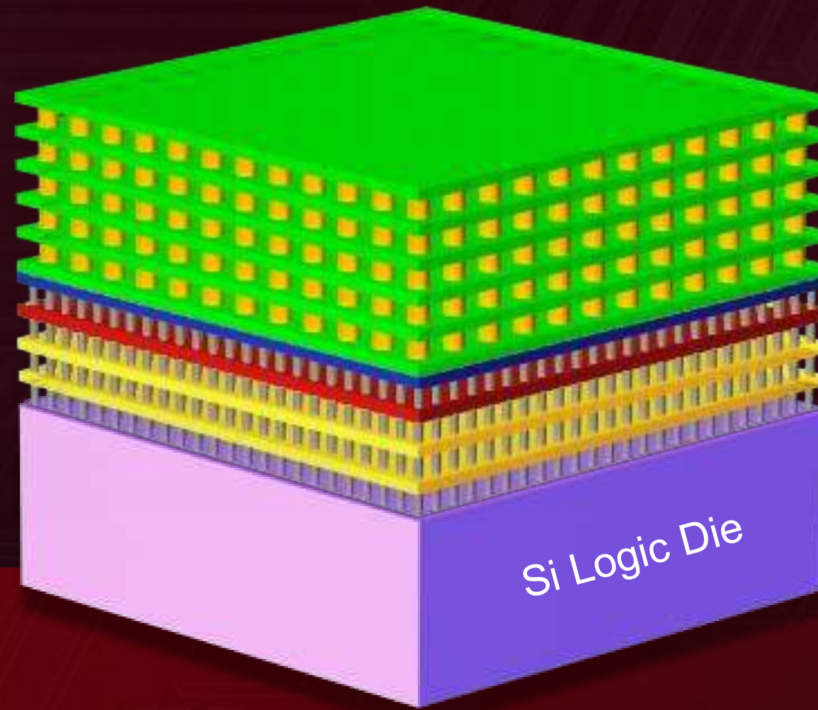
MEMORY INTEGRATION
ON LOGIC PLATFORM

Better transistor alone

© 2019 TSMC, Ltd

MEMORY INTEGRATION
ON LOGIC PLATFORM

Transistors integrated with memory in **3D**



© 2019 TSMC, Ltd
tsmc

SYSTEM INTEGRATION

A CONTINUUM FROM FAR BACK-END TO FRONT-END

Normalized Density

10^8
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0

Interposer

Chip-on-wafer
Wafer-on-wafer

Monolithic 3D

Source: IMEC

© 2019 TSMC, Ltd

SOCIETAL NEEDS FOR **ADVANCED TECHNOLOGY** IS INSATIABLE

ADVANCED TECHNOLOGY

– A KEY DIFFERENTIATOR

MULTIPLE ROADS LEAD TO ROME

Memory logic integration

Continuous transistor & memory advances

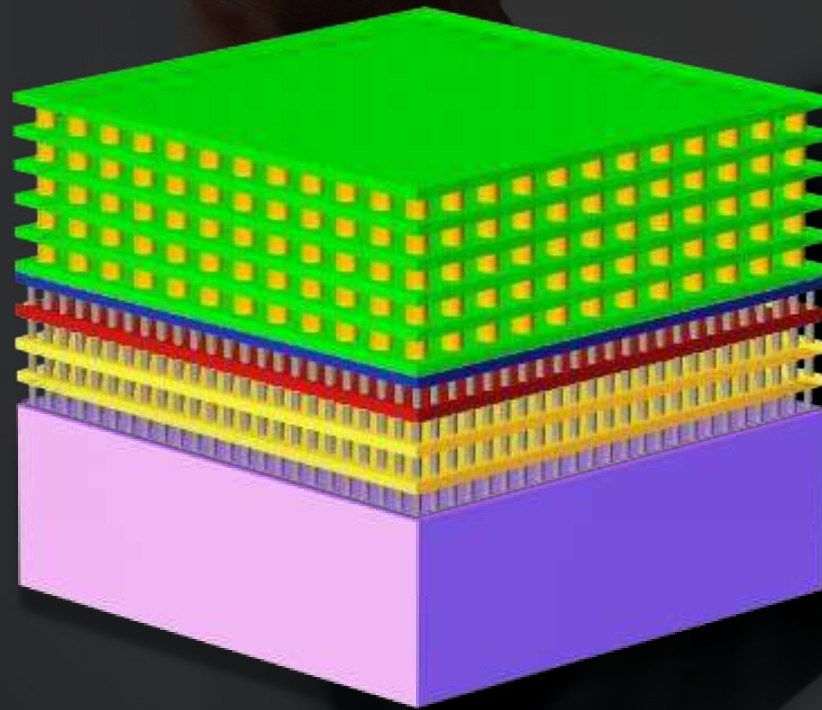
System integration with
high connectivity

CONTINUOUS BENEFITS NODE AFTER NODE

© 2019 TSMC, Ltd
tsmc ®

A CALL TO **ACTION**: EARLY ENGAGEMENT

SYSTEM ↔ TECHNOLOGY



ACADEMIA ↔ INDUSTRY RESEARCH

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**

© 2019 TSMC, Ltd