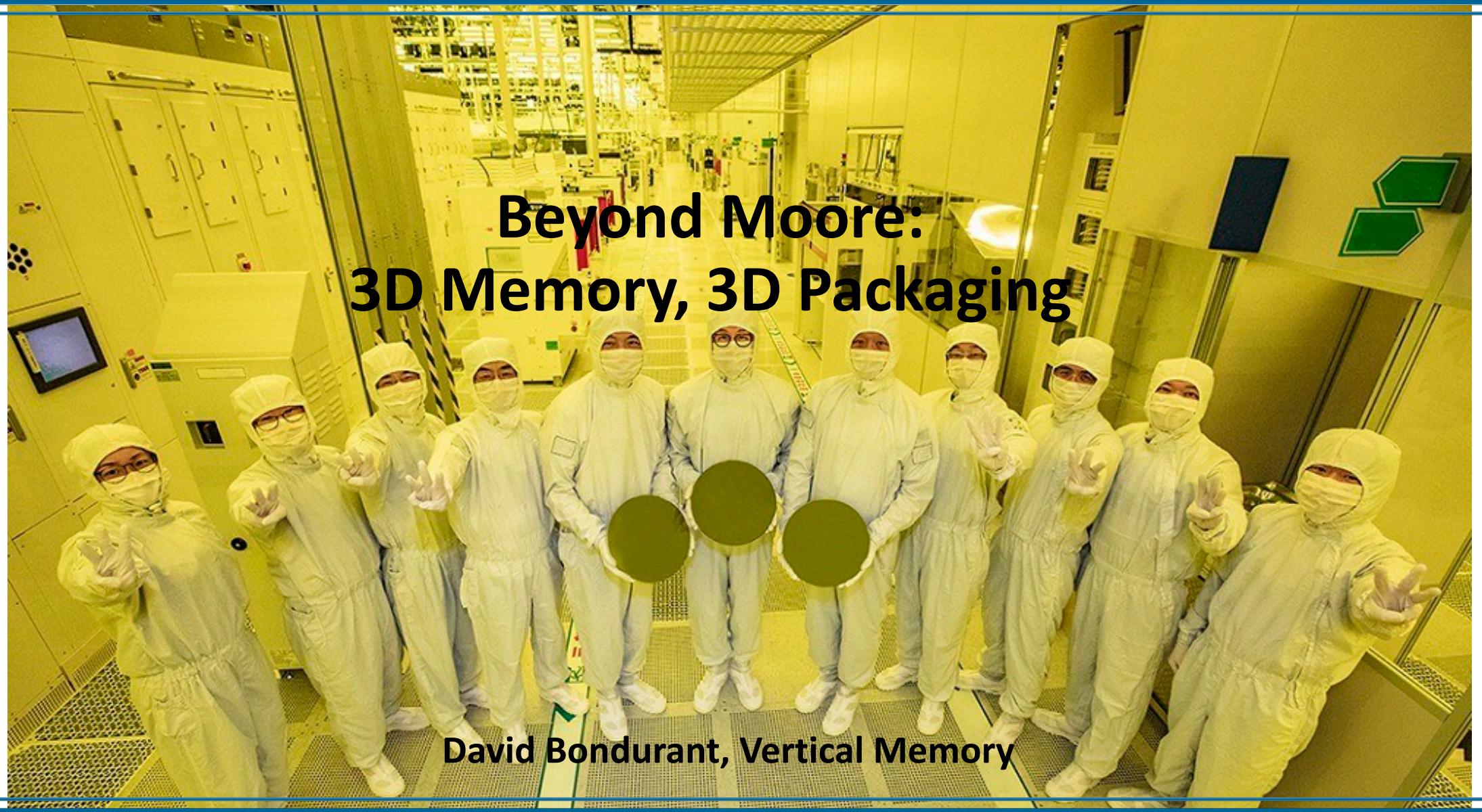


Beyond Moore:
3D Memory, 3D Packaging

David Bondurant, Vertical Memory

About Me

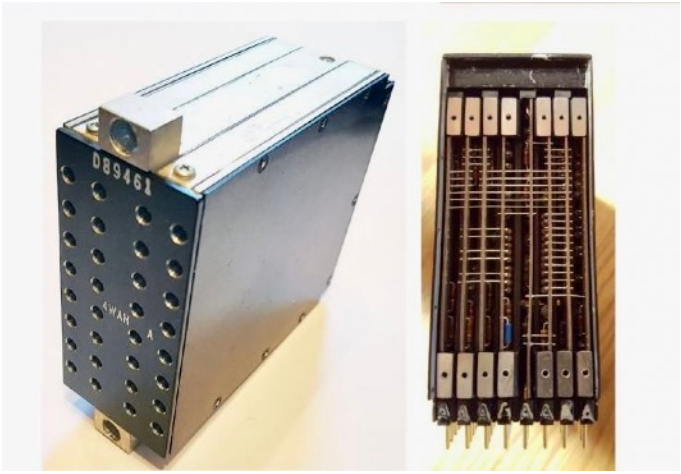
Riding the Moore's Law Wave

- BS Physics, BSEE, MBA Technology Management
- Worked on First DRAM Module at Control Data in 1971
- Designed Military Computers at Univac - 1972-80
- Designed VHSIC Microprocessors and Gate Arrays at Honeywell from 1980-88
- Marketed Emerging Memory 1988-2009
  - Ramtron (Ferroelectric RAM)
  - Enhanced Memory Systems (EDRAM, ESDRAM, HSRAM, ESRAM)
  - Simtek (nvSRAM)
  - Freescale/Everspin Technologies (MRAM)
- Consulting at Vertical Memory - 2002 - Present
- Volunteering at IEEE and Computer Society Since 2019
- Observed Semiconductor and Computer Business for 53-years

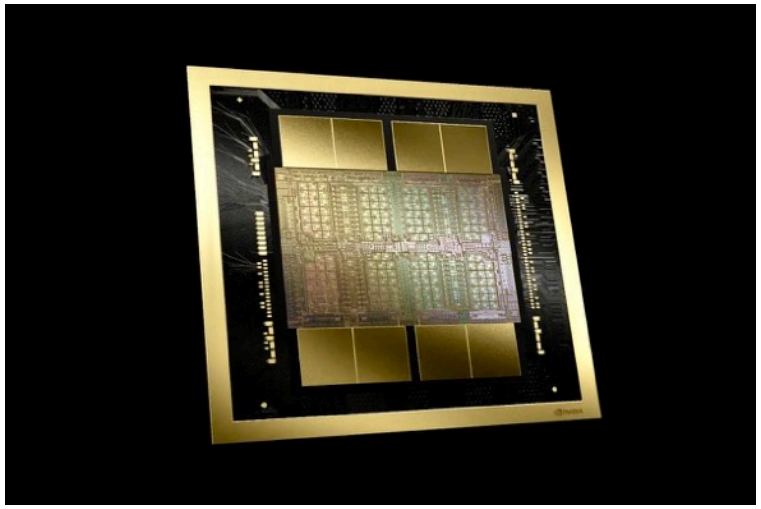
David Bondurant, PE (Emeritus), Life Senior Member

Presentation Outline

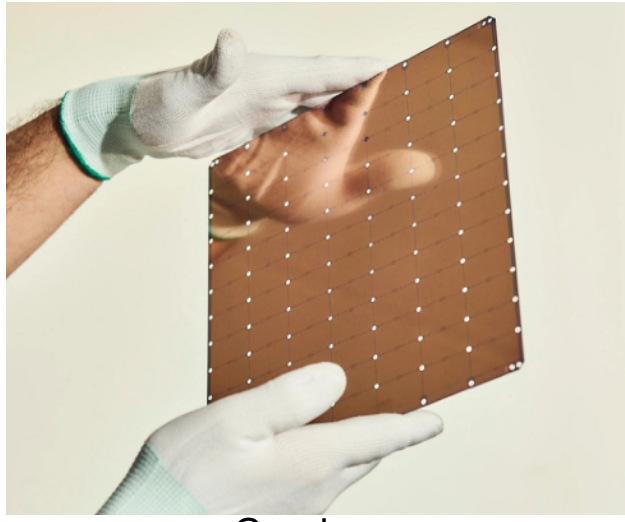
- **The First 70-Years - Moore's Law, Transistors to Many Processors on a Chip**
- **Moore's Law Hits The Wall**
- **Beyond Moore**
- **3D Packaging & Wafer-Scale Integration**
- **Vertical Memory**



3-D Freon Cooled Module - 1971



Nvidia B200 Blackwell - 2024

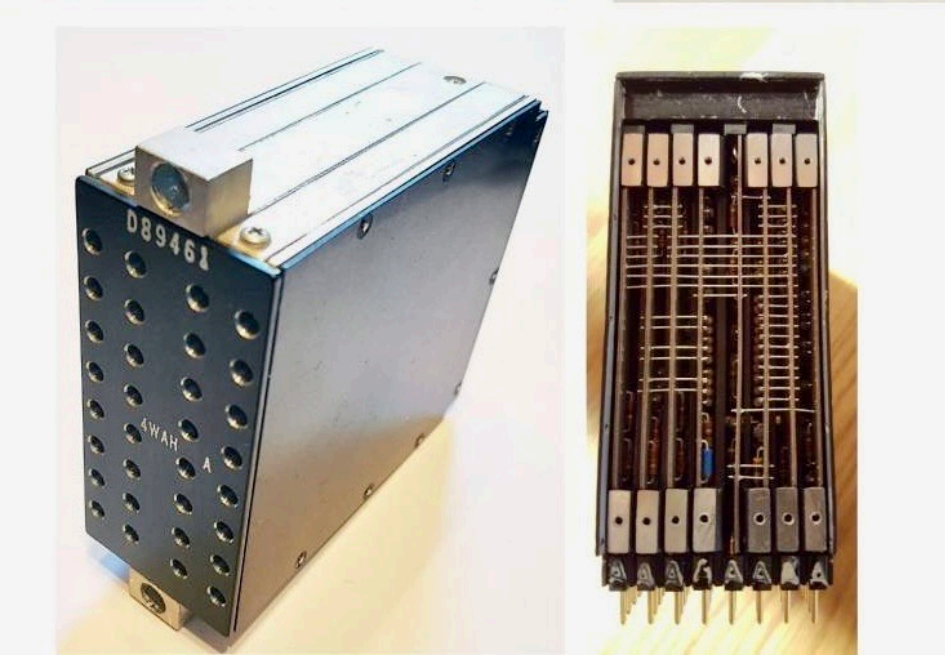


Cerebras
Wafer Scale Engine (WSE-3) - 2024

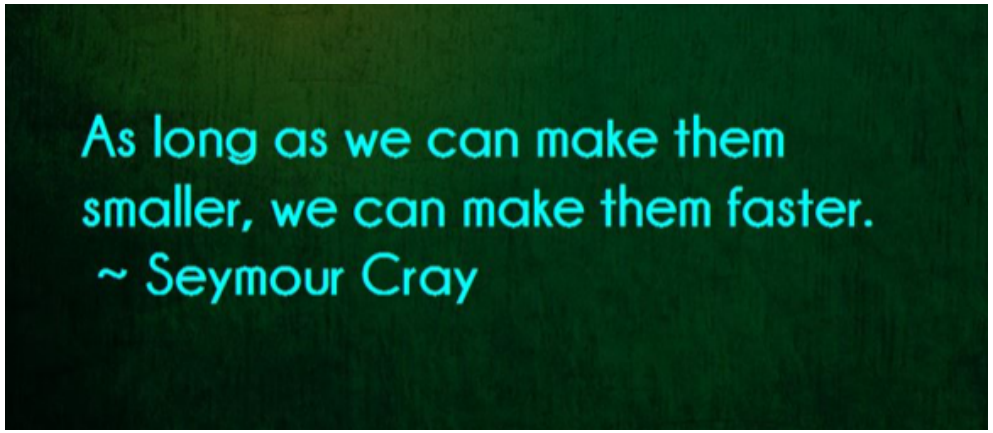
Seymour Cray

The Father of Supercomputing

- **The Transistor Was Invented in 1947**
- **Seymour Cray Joined ERA Associates in St. Paul in 1950**
- **Designed ERA's First Transistor Computer - Transtec in 1951. He designed the ERA Atlas II computer that would become the Univac 1103 in 1954. ERA would become Sperry Univac**
- **He Left Univac to Form Control Data Corporation in 1957 with William Norris, an ERA Founder**
- **Designed a Series of Supercomputers using Discrete Transistors from 1960 to 1972**
- **He formed Cray Research in 1972**
- **IEEE Computer Society created the Seymour Cray Computer Engineering Award in 1997 to honor high performance computer creativity**

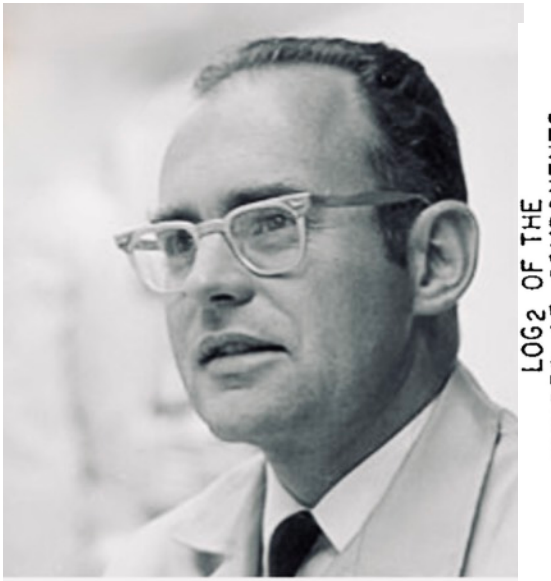


CDC 7600 Module Used 3-D Packaging - Freon Cooling

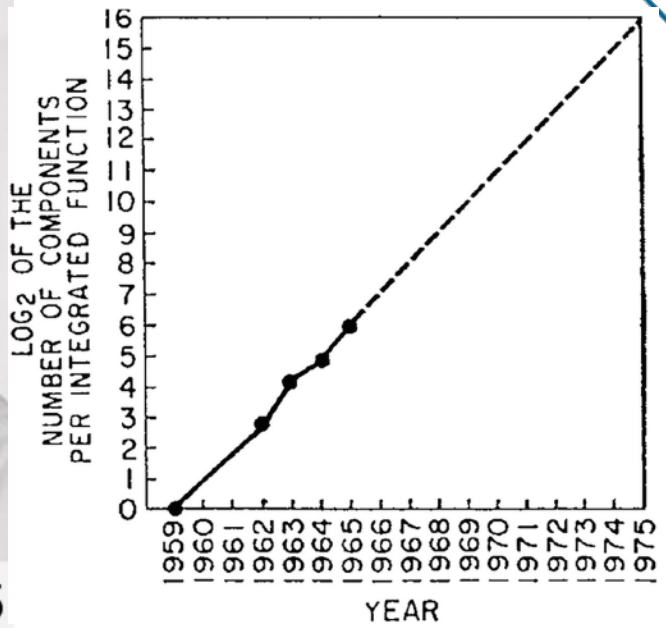


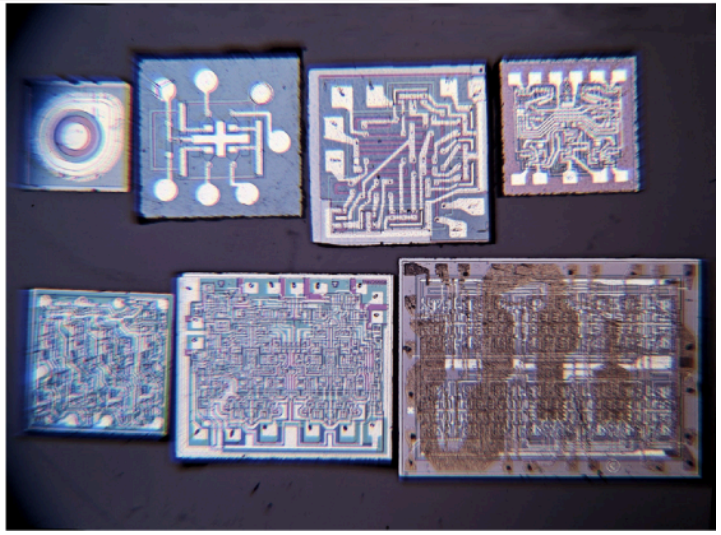
Moore's Law - IC Transistor Count

- Gordon Moore of Fairchild (later Intel) observed the number of transistors per chip doubled about every two years by 1965
- He plotted the Log scale number of transistor for 4 generations of Fairchild ICs and observed the linear increase
- Carver Mead of Caltech popularized the term "Moore's law" in 1975
- Moore's Law has defined the semiconductor industry transistor growth & economics for 50-years



Gordon Moore in 1965

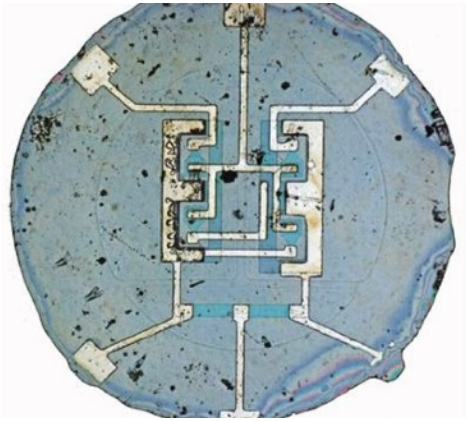




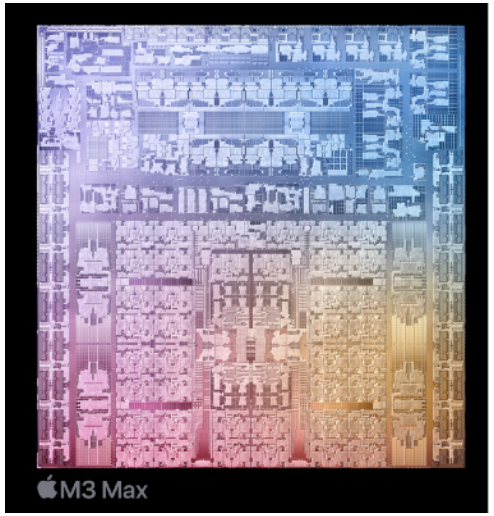
From Transistor to IC - Moore's Data

Moore's Law Hits the Wall

- The Feature Size of Integrated Circuits has scaled following Moore's law from 40 micron in 1961 to 3 nm in 2022. Shrinking the size of the transistor and allowing chips to grow from **16 transistors** to **more 92 billion on the Apple M3 Max**
- The Rate of Feature Reduction and Transistors per Chip is Slowing as we approach 1 nm (the Wall)
- NAND Flash Storage Reliability Became Limited Below 28 nm Triggering the First Beyond Moore Response - 3D NAND Required Vertical Integration



First Integrated Circuit
1961



Apple M3 Max - 2023
(3 nm, 92 Billion transistors)

	1960	1970	1980	1990	2000	2010	2020	2030	2040	2050
Semiconductor Feature Size (nm)	40,000	10,000	3,000	600	130	28	5	1	1	1
Moore's Law Scaling										
Beyond Moore										
Transistors Per Chip	16	1K	64K	16M	4B	8B	67B			

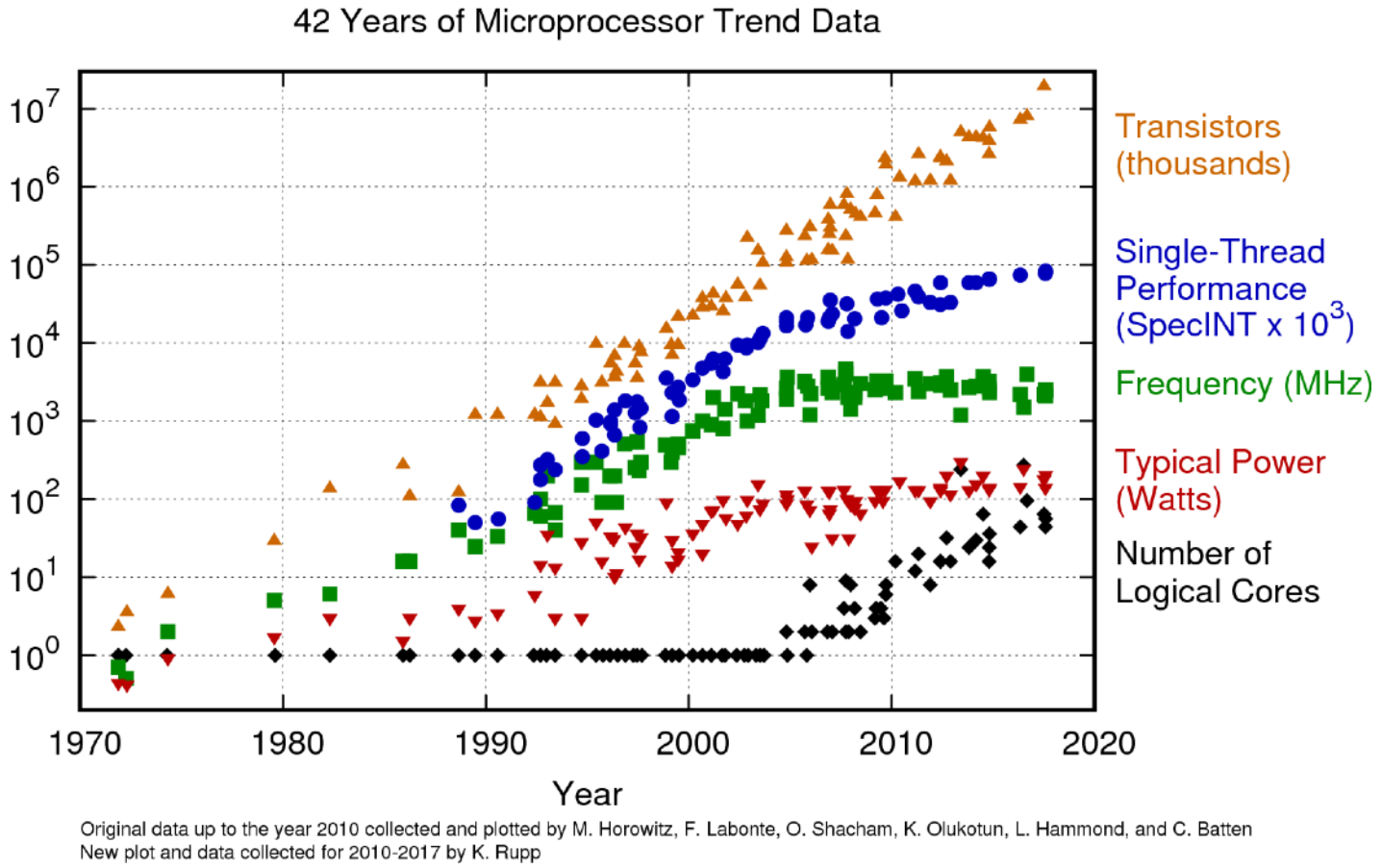
3D NAND Flash

IEEE

CMOS & Moore's Law - As long as we make them smaller, we can make them faster

CMOS Computers Hit Clock Frequency Wall Due to Power Limitations

- **Intel launched the 80386 uP in 1985 using 1.5 micron CMOS**
- **CMOS has been used in microprocessors and supercomputers every since**
- **Making transistors smaller increased the clock rate until uP clock rates peaked at a little over 5 GHz in 2005**
- **Silicon Chips became Power Limited**
- **Since then performance increases with number of processor cores**

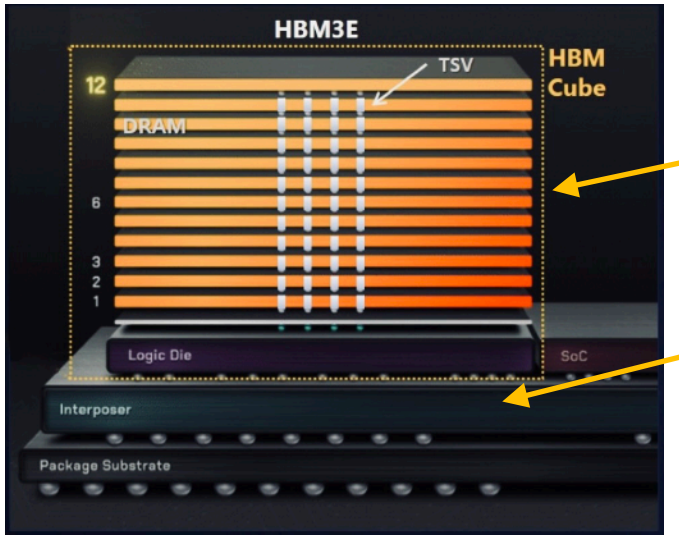


IEEE

Beyond Moore's Law - Why Process & Packaging Leadership Is Strategic

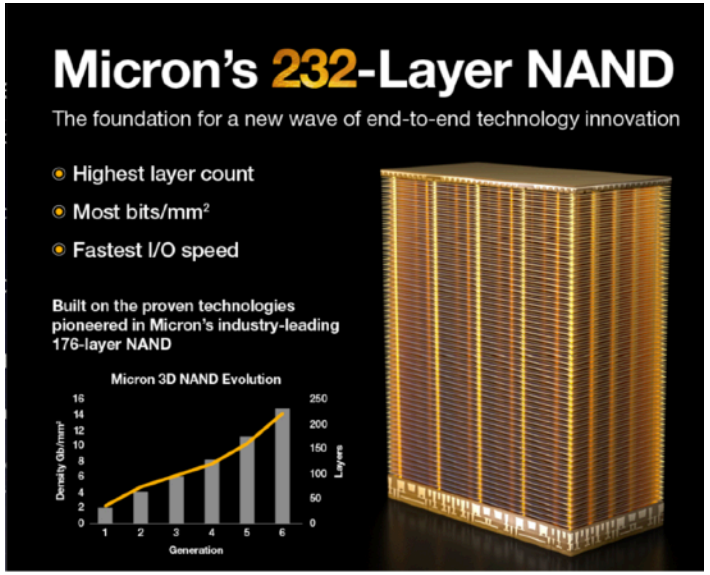
Going Vertical

- **Beyond Moore scaling requires different approaches**
  - **2.5D and 3D Heterogeneous Integration of Chiplets & Memory on Silicon Interposers (CoWoS)**
  - **3D Vertical Integration of Chips - Stacking (DRAM, Flash now)**
  - **3D Vertical Integration on Chip (NAND now, DRAM soon)**
  - **Wafer Scale Integration (Bigger Chips)**
  - **Higher dielectric constant capacitors (DRAM)**
  - **New non-volatile storage devices (Ferroelectric, Magnetoresistive, Resistive, Nanotube)**
  - **New Substrates & Devices (SiC, GaN)**



3D Die Stack

2.5D Packaging



IEEE

TSMC CoWoS Technology

Going Vertical

- **TSMC Pioneered Chip on Wafer on Substrate Technology**
  - **Silicon Interposer has 2X Interconnect Density of Composite Substrates**
  - **System Build Using Chiplets (SoC) and 3D Stacked HBM Memory on Silicon Interposer**
  - **Future Systems May Use Glass Substrates to provide larger, more stable interconnections**

HPE Cray Frontier Supercomputer

World's Fastest Computer (2022) Used 2.5 & 3D Integration

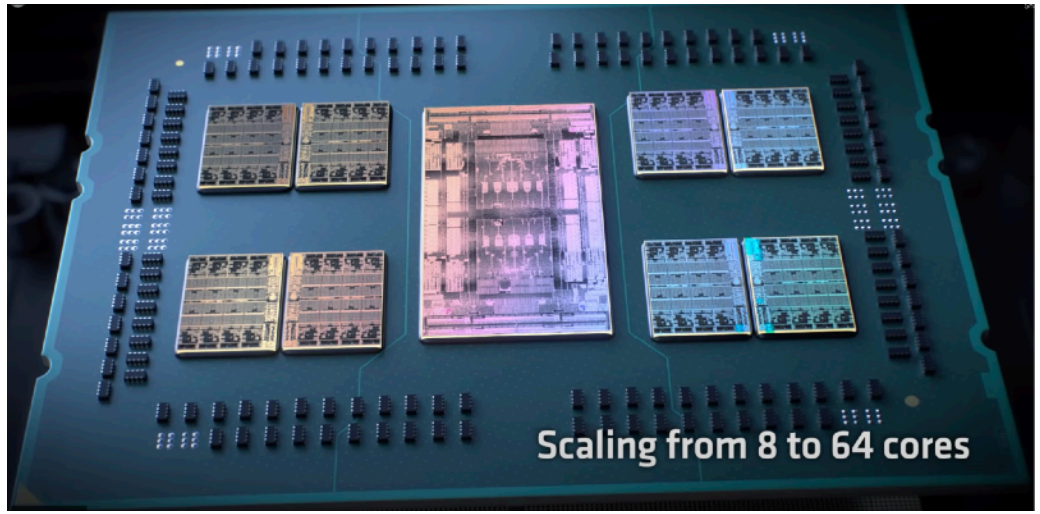
- Supercomputer Performance Leader with 1.102 exaFlops
- 9472 AMD Epyc CPU (608,208 cores), 37,888 AMD Instinct GPU (8,335,360 cores)
- 7 nm CMOS Feature Size
- 2.4 GHz Clock Speed

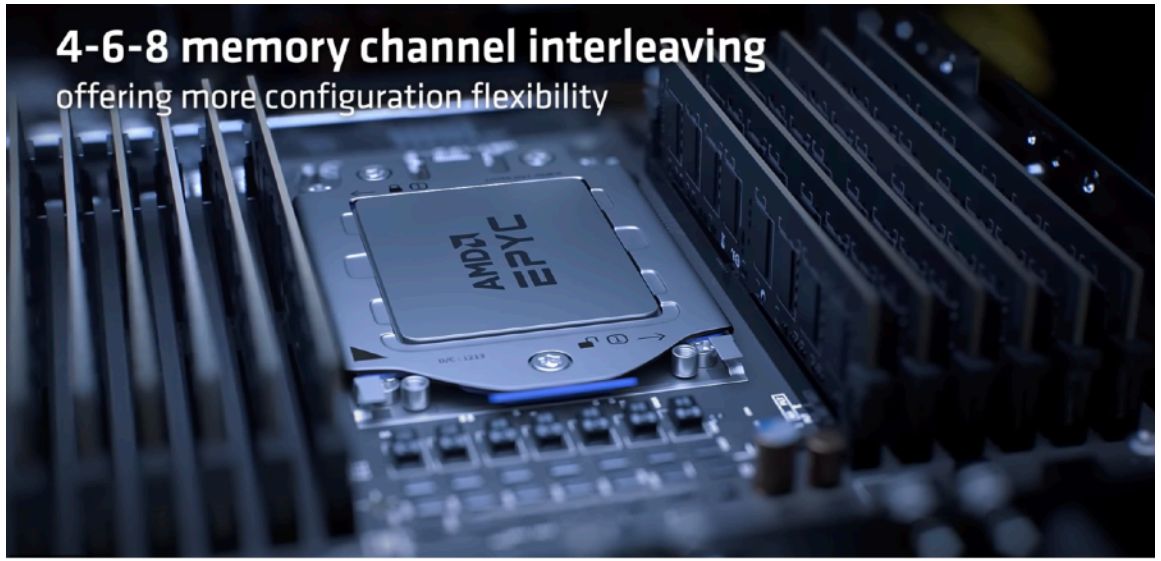
IEEE

AMD EPYC 74A53S Processor Module

Example of Heterogeneous 2.5 D Integration

- Combines 8 Processor Chiplets with 256MB L3 Cache and Interconnect Chip on Silicon Interposer (2.5D Packaging)
- 7 nm Features, 2.4 GHZ Clock
- 64 Processors
- 8 DDR Memory Channels (410GB/Sec Bandwidth)
- High Performance Ethernet Interface (200GBe)

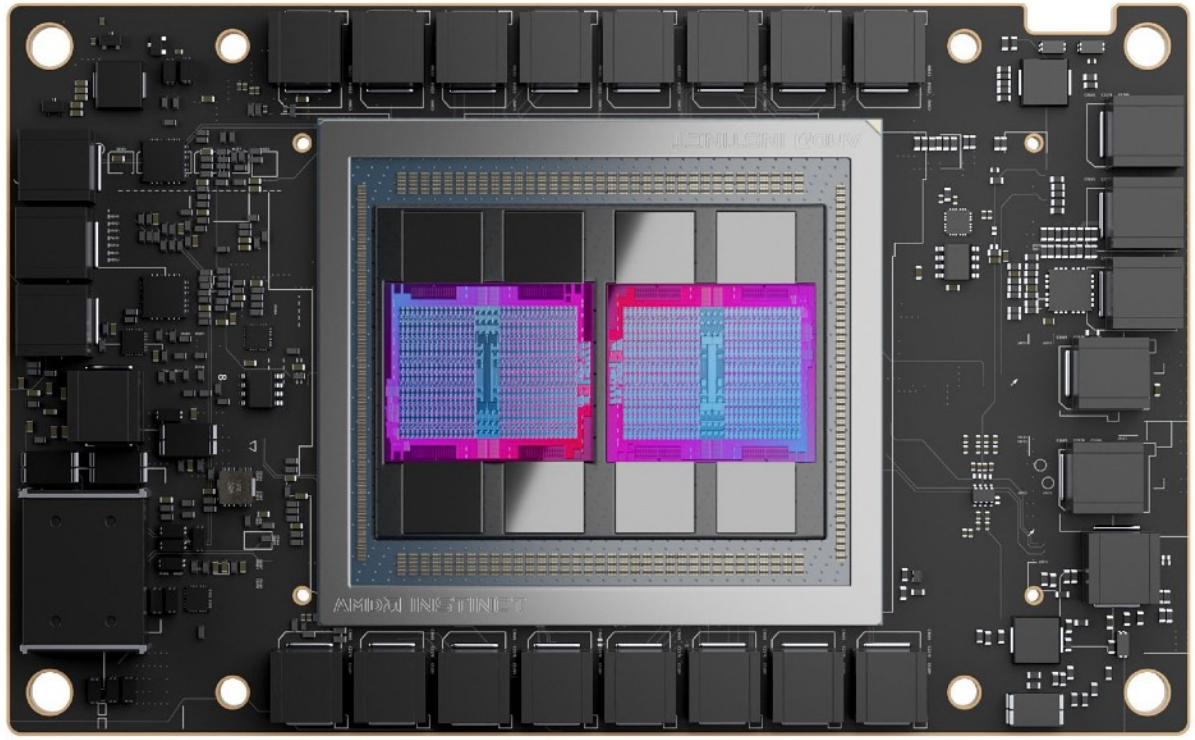




AMD 2nd Generation Instinct GPU

2.5D and 3D Integration - Complexity Increases

- 6 nm Features, 56B Transistors, 1 GHz Clock (1.7 GHZ Turbo)
- 2 GPU Chiplets - 13312 Cores, 832 TMU
- 8 HBM2E Die Stacks (8 High) - 1024 Bit Memory Bus per Stack
- 128GB Memory Capacity , 2.5GB/ Sec Bandwidth
- 500 Watts

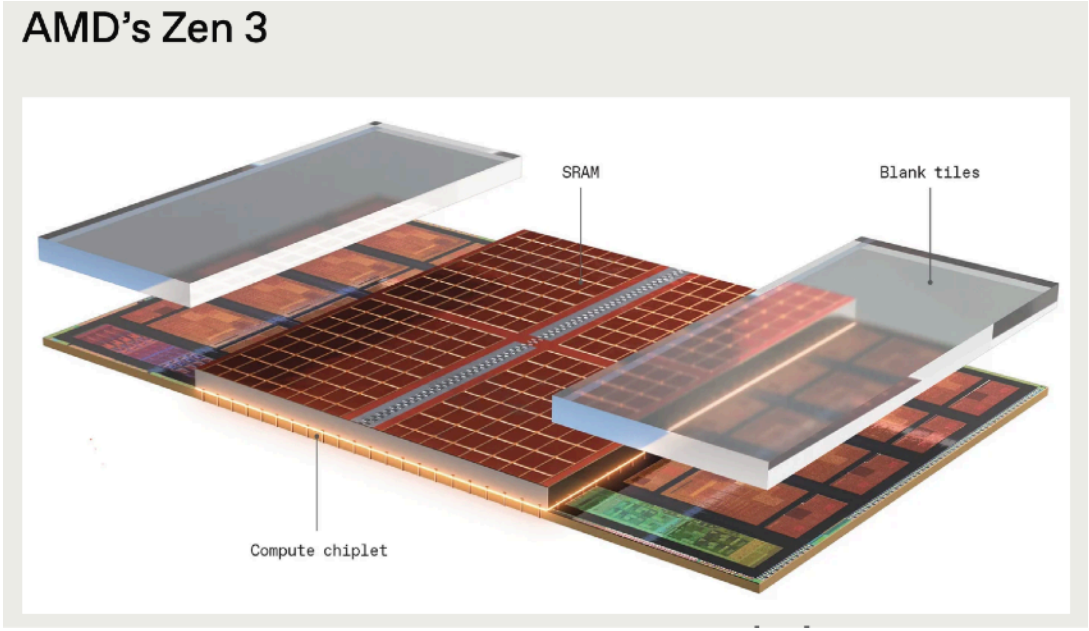


AMD Instinct MI250X GPU

3D Packaging

3D Heterogenous Integration - Complexity Increases

- AMD Zen 3 Architecture Stacked 32MByte L3 Cache on CPU Chip as early as 2020
- Heterogeneous Integration Allows CPU and SRAM Cache Chipsets to be on optimized processes
- CPU Chiplets Move to Advanced Processes
- Ryzen Versions Would Have As Much as 96MB L3 Cache Using V-Cache (Die Stacked Vertically on CPU)



AMD Zen 3 CPU Mounts Large V-Cache on CPU

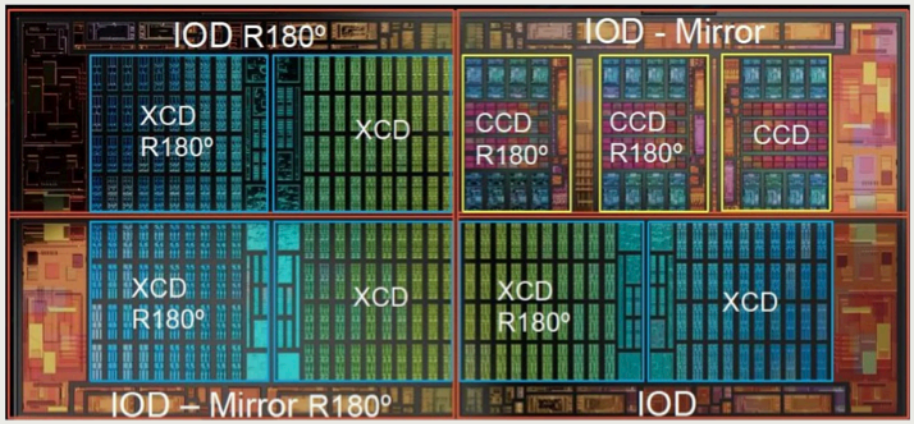
Source: 3 Ways 3D Chip Tech Is Upending Computing, IEEE Spectrum, March 2022

3D Packaging

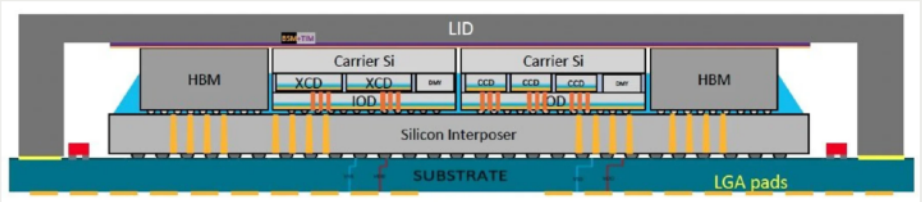
3D Heterogenous Integration - Complexity Increases

- AMD MI300a GPU Used Multiple Specialized Chiplets Stacked 3 High
- Specialized Compute, Accelerator, and I/O
- GPU Along with 3D Stacked HBM Memory Mounted on Silicon Interposer
- Interposer Has 2x Wiring Density of Composite PCBs

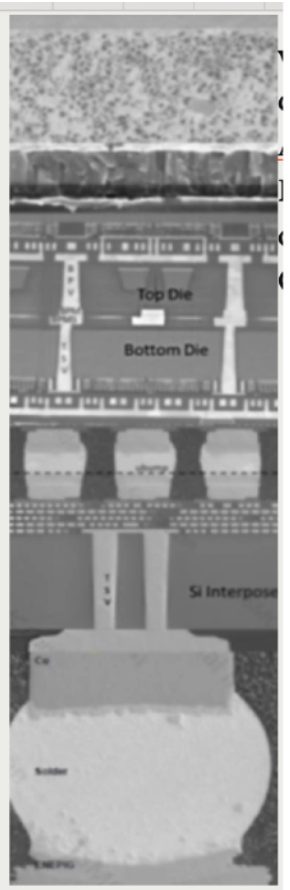
Source: AMD's Next GPU Is a 3D-Integrated Superchip, IEEE Spectrum, December 2023



To get everything to line up, the IOD chiplets had to be made as mirrors of each other, and the accelerator (XCD) and compute (CCD) chiplets had to be rotated. AMD



Compute and AI chiplets are stacked on top of I/O and cache chiplets in the MI300a. AMD

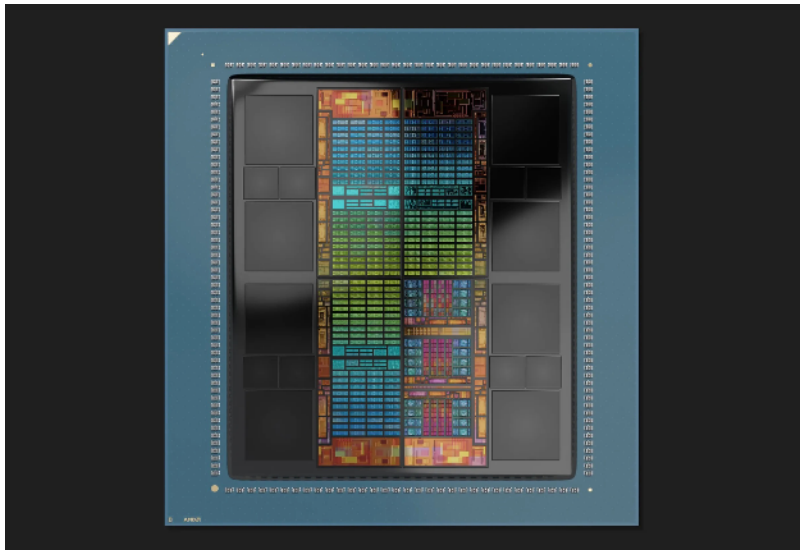


A slice of MI300 stack from the carrier silicon at the top to the solder ball at the bottom of the package. AMD

AMD 3nd Generation Instinct GPU (MI300a)

3D Integration - Complexity Increases Further

- 5 & 6 nm Features, 153B Transistors
- 3D Stacked Processor Chiplets Mixing CPU, GPU, IOD
- 8 HBM3 Die Stacks (12 24Gbit Chips High) - 1024 Bit Memory Bus per Stack
- 192GB Memory Capacity , 5.2TB/ Sec Bandwidth



Next Gen MI300a Instinct GPU

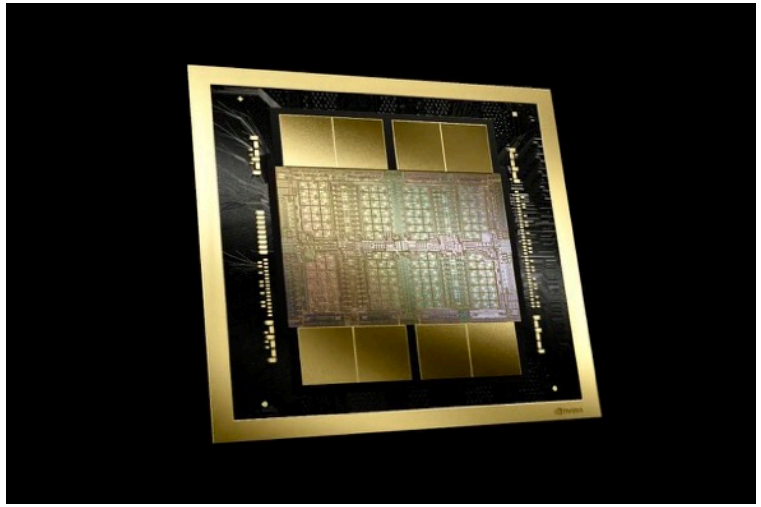
November 2024 - El Capitan Supercomputer Becomes World's Fastest Supercomputer

IEEE

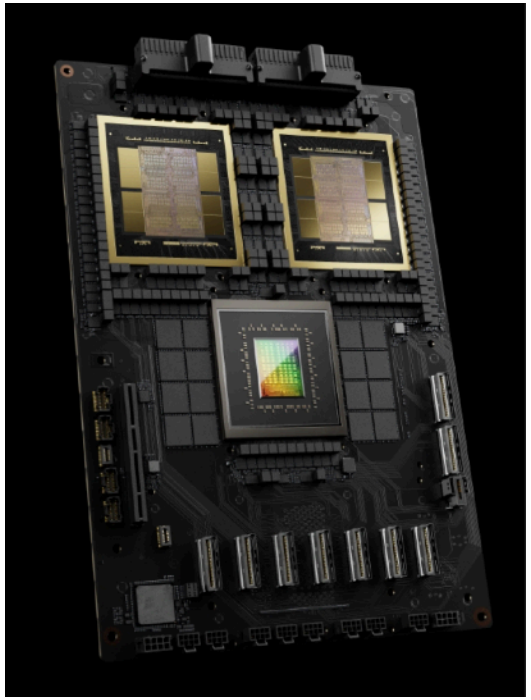
Nvidia Blackwell GPU - AI Processor

3D Integration - Complexity Increases Further

- 4 nm Features, 208B Transistors per Chiplets
- 2 GPU Chiplets, 8 HBM Die Stacks on Silicon Interposer
- 8 HBM3E Die Stacks (12 24Gbit Chips High) - 1024 Bit Memory Bus per Stack
- 192GB Memory Capacity , 8TB/s Bandwidth



Nvidia B200 Blackwell GPU



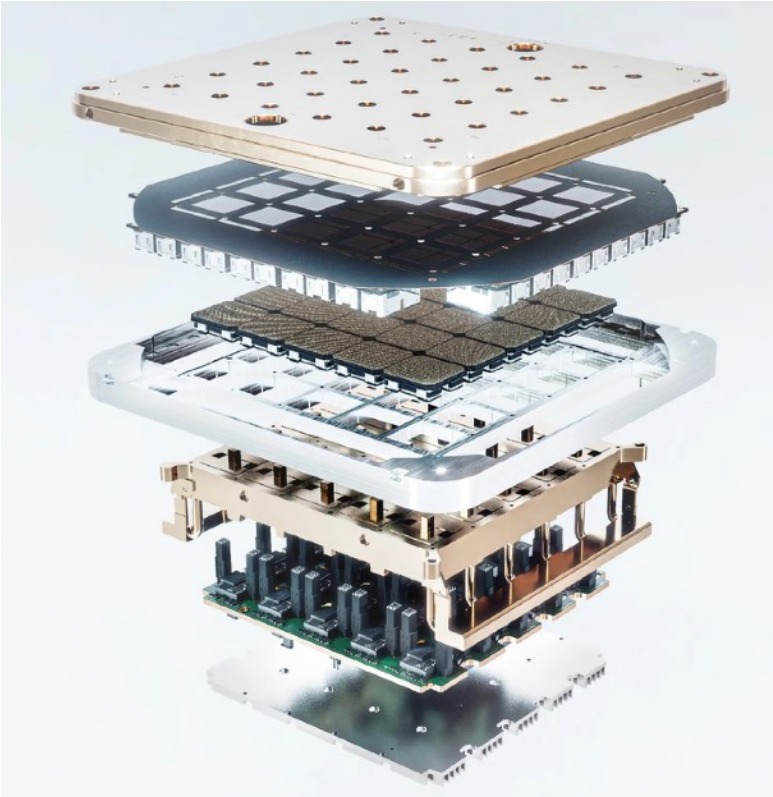
2 GPU and 1 Grace CPU Module Form Liquid Cooled Module

https://www.anandtech.com/show/21310/nvidia-blackwell-architecture-and-b200b100-accelerators-announced-going-bigger-with-smaller-data

Exceeding the 800 mm Reticle Limit

System on Wafer Increasing Silicon Interposer Size

- Gene Amdahl's Trilogy Systems attempted Wafer Scale Integration using chips mounted on Silicon Wafer as long ago as 1980s
- TSMC Claims 20 Silicon on Interconnect Fabric (Si-IF) programs since 2019
- Tesla Dojo AI System Using 5x5 array of chipsets on Si-IF now
- Nvidia Proposing Next Generation Blackwell AI System Using More Chiplets and 12 3D HBM-4 Stacks
- TSMC Predicts Full Wafer Si-IF with 40 Reticle Chiplets and 60 HBM as early as 2027



Tesla Dojo Training Tile Using TSMC System-on-Wafer

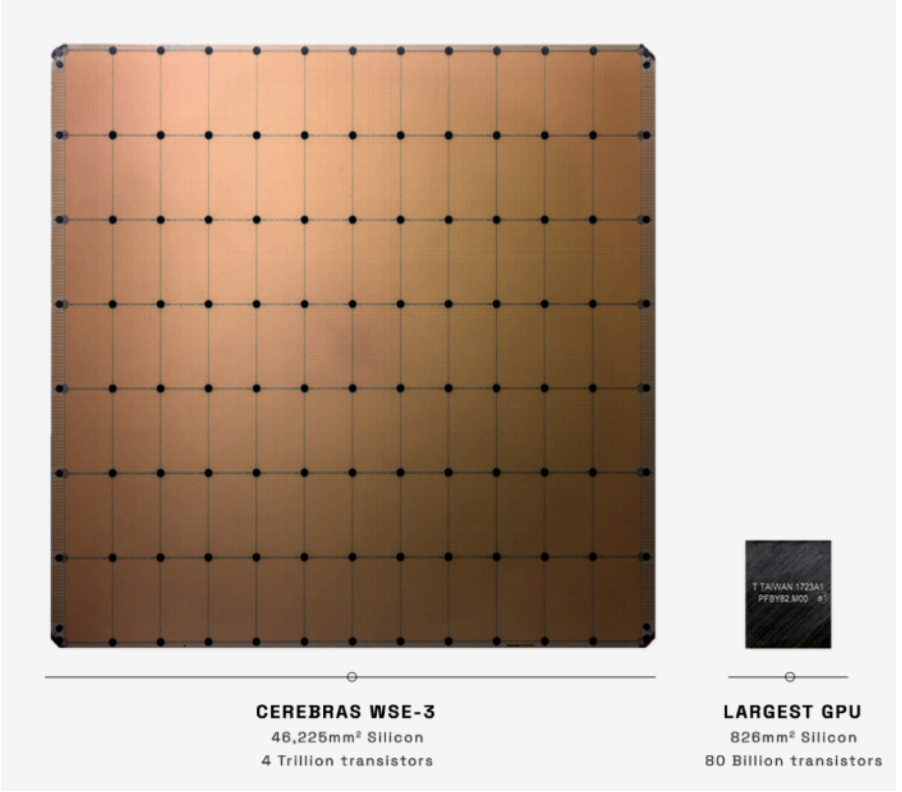
Source: https://spectrum.ieee.org/tsmc-advanced-packaging - 30 Apr 2024

IEEE

Cerebras CS-3 AI Processor

3rd Generation of Wafer Scale Integration

- 5 nm Features
- 900,000 AI Processor Cores
- 44GB eSRAM memory
- 21 PB/Sec Memory Bandwidth
- 4 Trillion Transistors
- 214 Pb/s Fabric Bandwidth
- Water-Cooled



Cerebras Wafer Scale Engine (WSE-3)

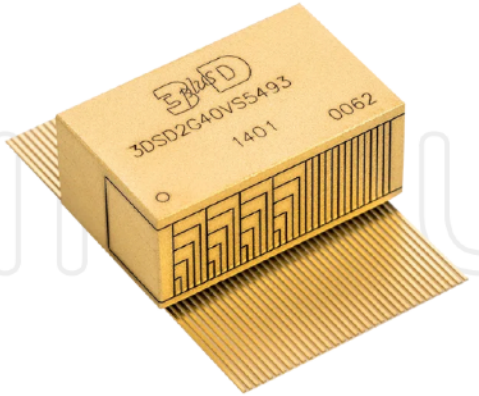
Source: https://www.cerebras.net/product-chip/

Vertical Memory

Memory Vision From 2002

- At Ramtron and Enhanced Memory, I worked with emerging memory technologies
  - Ferroelectric RAM - A non-volatile DRAM using ferroelectric materials
  - Enhanced DRAM - The fastest available DRAM combining fast DRAM and on-chip cache

- When I left Ramtron & Enhanced Memory Systems, I named my new company, **Vertical Memory**. My vision was that future memory would need to shrink to close to zero ns speed, unlimited endurance, almost no cell area, and retain data without power. At that time, I said such a memory would require new storage materials and require memory to go Vertical - vertical device structures, stacked components in a package, multiple bits per cell. Today, these visions are called **Beyond Moore**.



3D SDRAM Stack (1Gb) - 2003

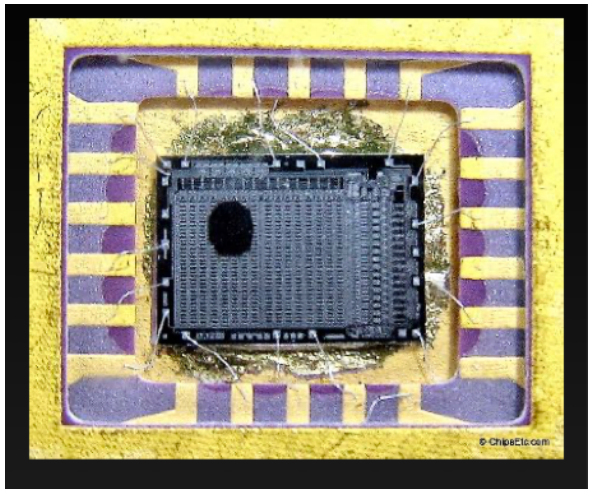
3D HBM3E (36GB) Stack - 2024

My First Job

Control Data Corporation (1971)

- CDC 7600 - The world's fastest computer in production (discrete transistors)
- Seymour Cray was developing the CDC 8600 in Chippewa Falls
- I am assigned to Memory Development and the design of first DRAM module for a CDC Supercomputer
- We used the AMS 7001, the first 1Kbit PMOS DRAM was produced in 1969. It operated at +15 Volts
- DRAM Module used 128 chips on 2 6x9 12-layer PCB with Freon Cooling. 4Kx32 Capacity. Requires conversion from PMOS to TTL to ECL

CDC Arden Hills Development Center

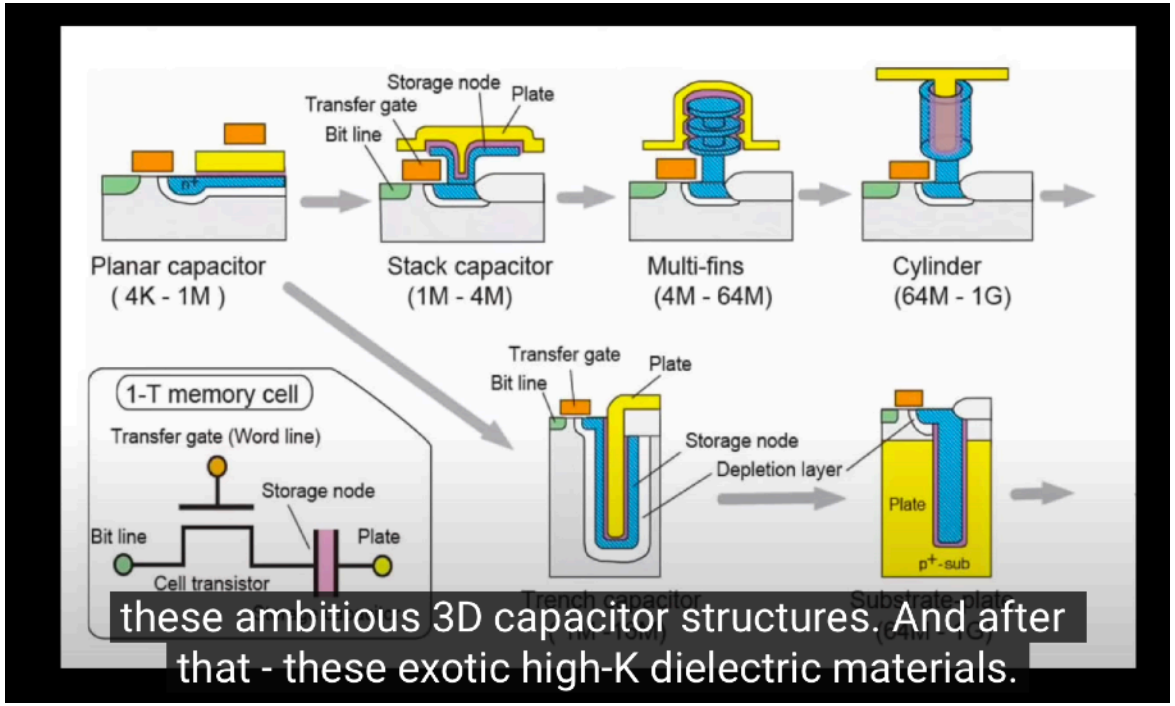


AMS 1Kbit PMOS DRAM

3D DRAM Process Integration

From Planar Process To Vertical Integration and New Materials (High-K Dielectrics)

- The First 1Kbit DRAM used a Planar capacitor
- By the 1Mbit Generation (1985) DRAM Became 3D
  - Stacked Capacitors
  - Trench Capacitors
- With Each Generation, the 3D Capacitor Structure Got More Complex
- New High-K Dielectrics Were Introduced to Increase Capacitance
  - SiN
  - HFO
  - Tantalum Pentoxide
  - BST

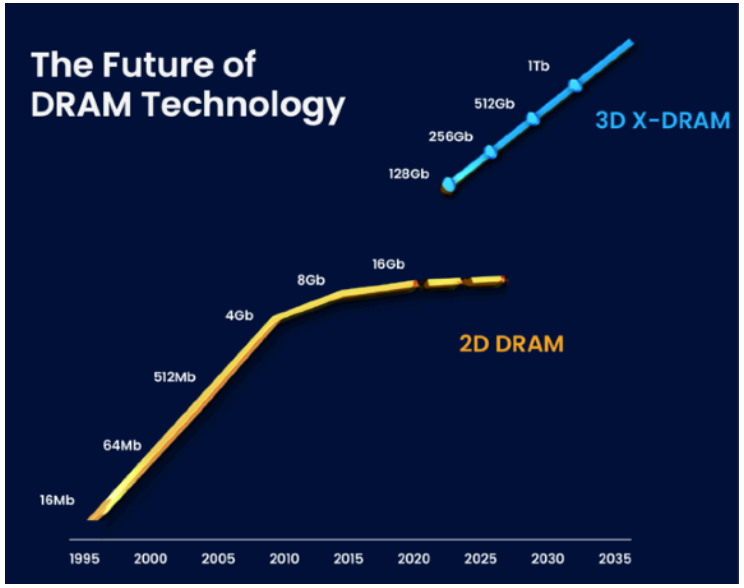


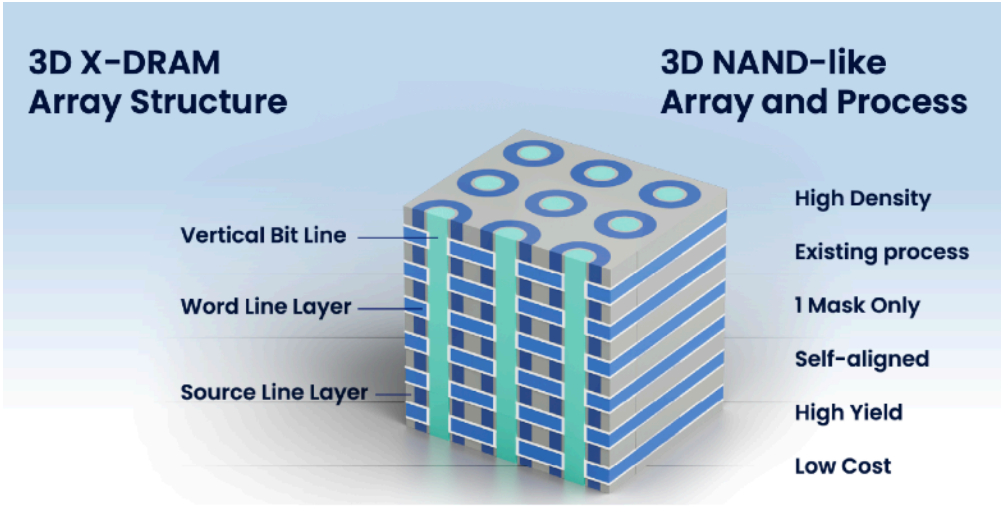
Source: How Semiconductor DRAM Went 3D - Asianometry (You Tube)

3D DRAM Process Integration - Future Direction

Vertical Stacking of Transistors and Capacitors

- DRAM Density Scaling Stopped at 10 nm and 4Gb
- Hit Moore's Law Wall in 2010
- DRAM Needs to Grow Vertically With 3D Transistors and Capacitors
- Startup Neo Semiconductor Proposing Vertical Transistors Stacking with Floating Body /Capacitors
- Proposes Roadmap starting at 128Gb DRAM in 2025





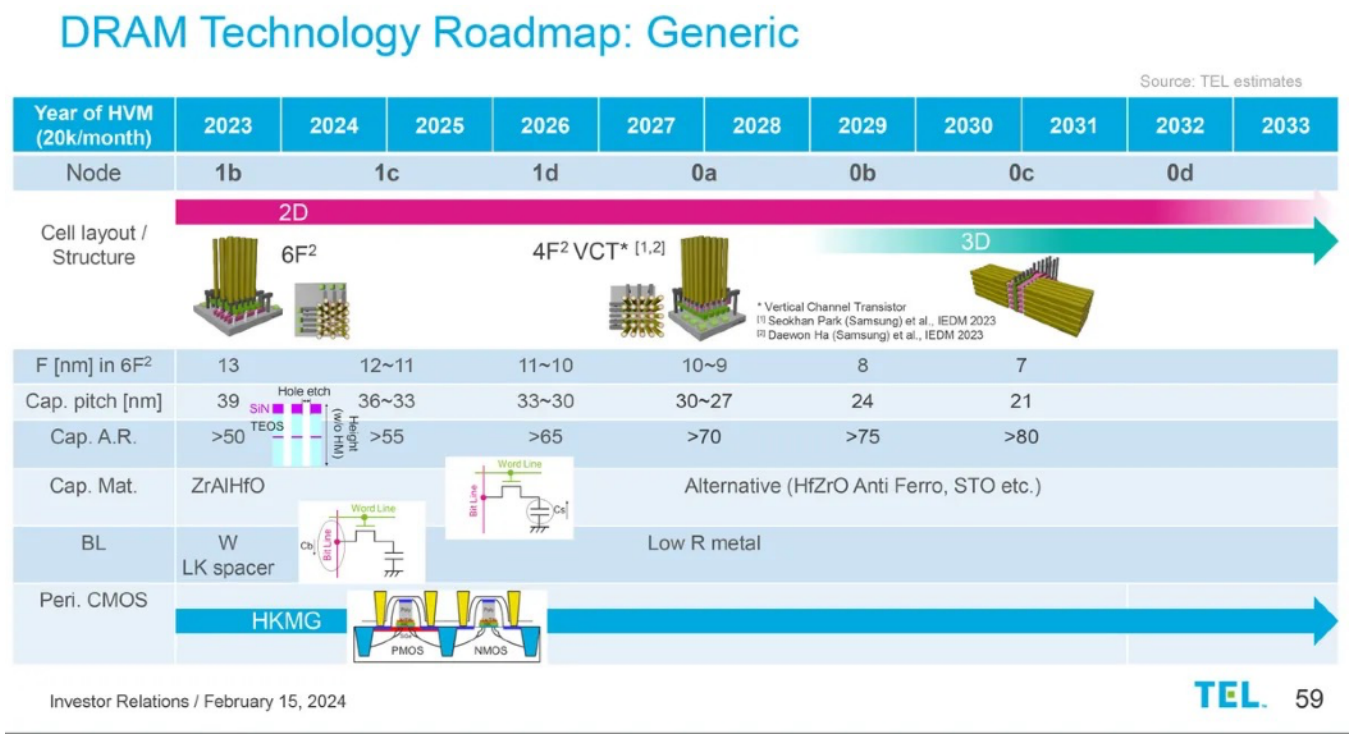
Neo Semiconductors Proposes Vertical Transistors with Floating Body Capacitors

Source: https://neosemic.com/

3D DRAM Process Integration - Future Direction

Vertical Stacking of Transistors and Capacitors

- Industry Leader Samsung Proposing Roadmap With Vertical Channel Transistors
- Vertically Stacked Capacitors
- New Materials Beyond HFO
- Process Scaling to 7 nm



Source: Samsung 3D DRAM Roadmap - Memcon 2024

Increasing DRAM Bandwidth

Multi-Processor Chips Demand Dramatic Increases in Memory Bandwidth

- Synchronous DRAM Clock Frequency Increased for 133 MHz to 4GHZ (2002 to 2024)
- Double Data Rate, Quad Data Rate I/O Introduced
- Memory Market Segments By Bandwidth
  - PC/Server Market - x1,x4,x8,x16 wide memory, DDR
  - Graphics Market - x32 wide memory, QDR
  - Low Power DRAM Market - x32 DDR 3D Memory
  - High Bandwidth Memory Market - x1024 DDR 3D Memory

3D Chip Stacking - HBM

High Bandwidth Memory Delivers Highest Bandwidth & Density

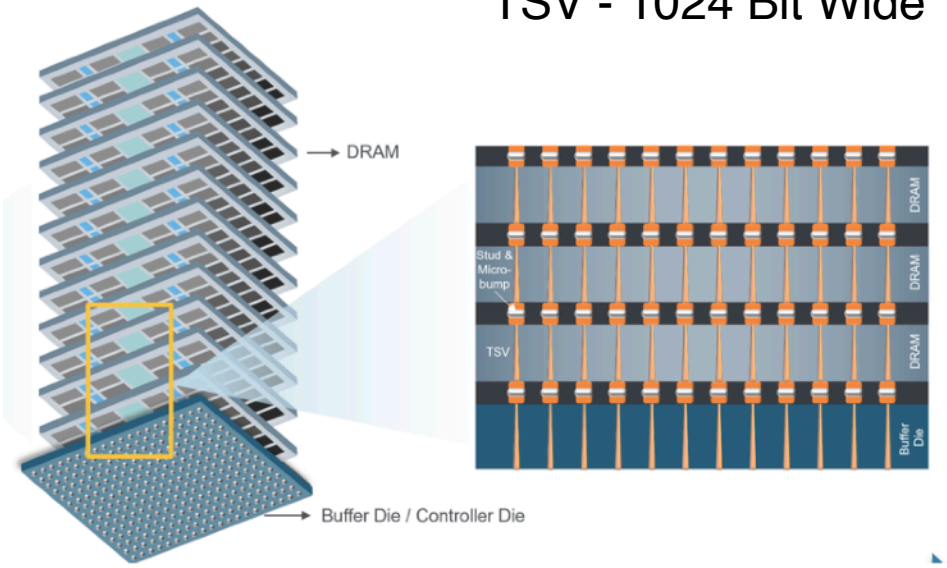
High Bandwidth Memory (HBM)

	HBM1	HBM2	HBM2E	HBM3	HBM3E	HBM 4
Year	2013	2016	2018	2022	2024	2025
Stack Height	4	8	8	12	12	16
DRAM Chip Density	8Gb	8Gb	16Gb	16Gb	24Gb	32Gb
Capacity	4GB	8GB	16GB	24GB	36GB	64GB
Clock Rate	500 MHz	1GHz	1.6GHz	3.2GHz	4.0GHz	TBD
Transfer/Second	1Gbps	2Gbps	3.2Gbps	6.4Gbps	8Gbps	TBD
Bus Width	1024	1024	1024	1024	1024	1024
Bandwidth	128GB/s	256GB/s	320GB/s	819.2GB/s	1TB/s	TBD

HBM Roadmap

12 High

TSV - 1024 Bit Wide



Source: Wikipedia, Samsung Roadmap - IEEE Life Member Conference

IEEE

Beyond Moore Is Taking Computers To New Heights

Control of Leading Edge Process & Packaging Technology Strategic to Continued Computer Development

- **Moore's Law Is Hitting the Wall**
- **Beyond Moore Technology Is Growing Chips and Systems Vertically**
- **3D Packaging Stacks Heterogeneous Chiplets 3 High and Interconnects Chiplets with Silicon Interposers**
- **3D Stacked HBM Achieving Incredible Bandwidths and Densities**
- **3D DRAM Process Technology Set To Increase DRAM Die Densities**
- **Removing Chip and System Power is the Key System Issue - Liquid Cooling Now, Cryogenic Cooling Coming**