高性能计算(HPC)的挑战与机会
Challenges and Opportunities in HPC

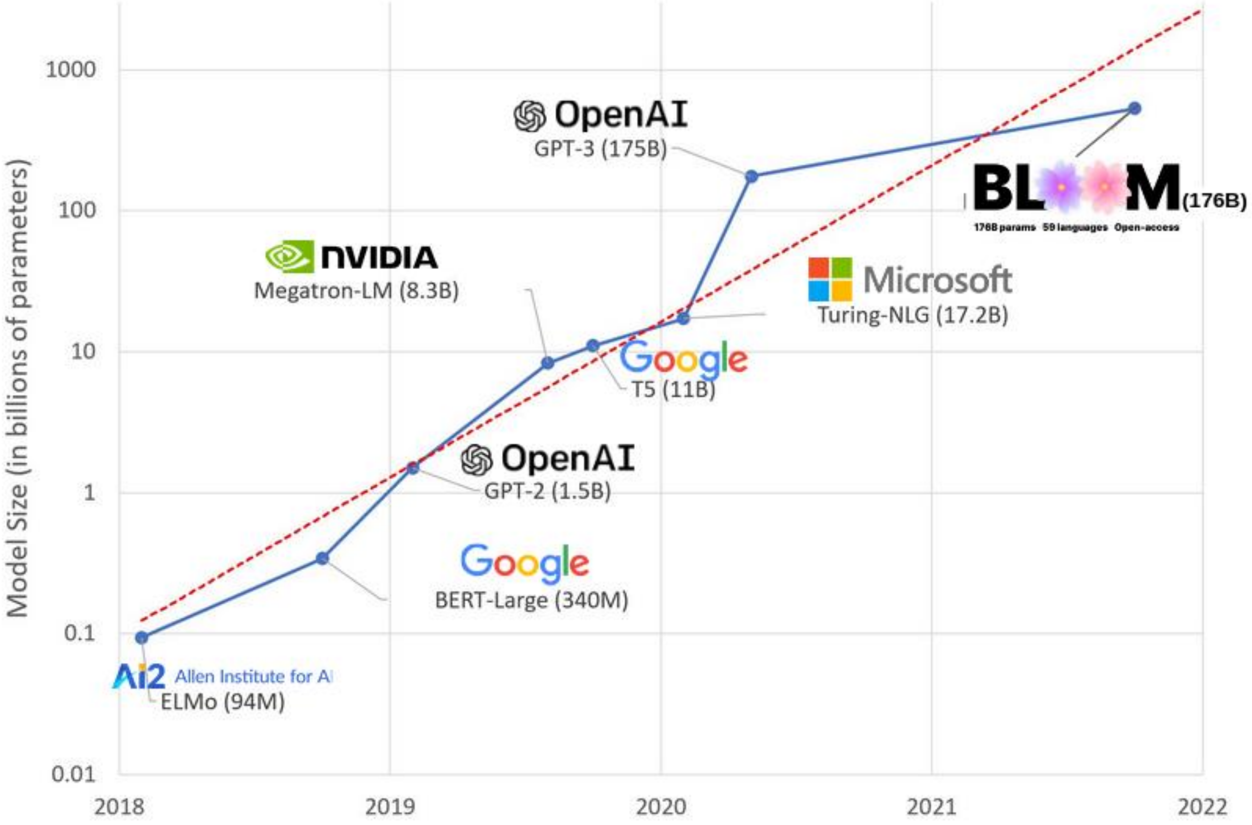
爱普存储技术(杭州)有限公司　　钟雷

Outline

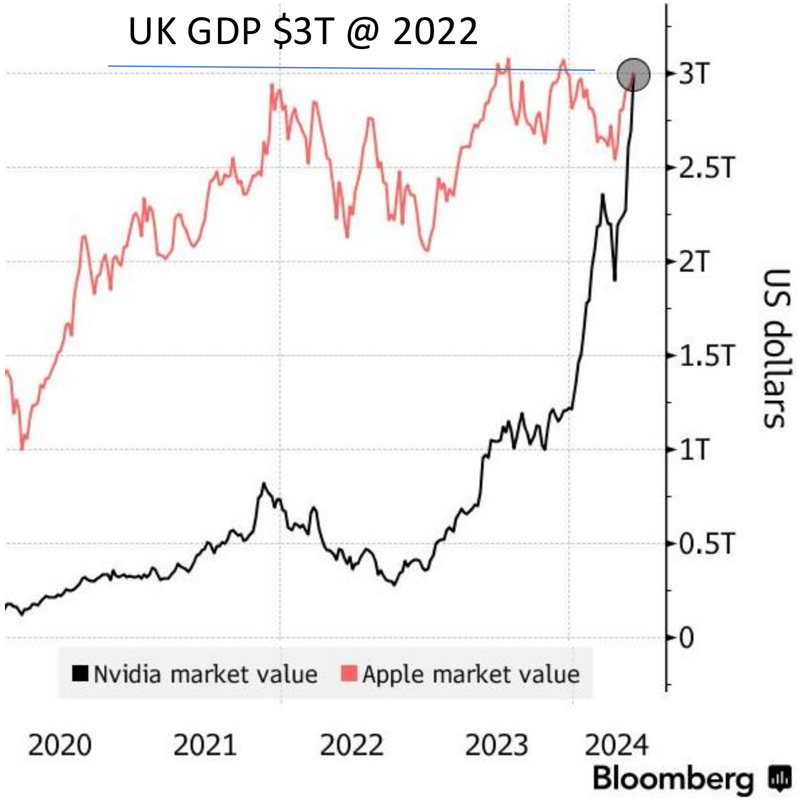
- “Two Dark Clouds” over HPC Development
  高性能计算发展的 “两朵乌云”
  - Memory Bandwidth Limit (or Information Bandwidth)
    内存带宽 / 信息带宽
  - Energy Bandwidth Limit: Power Delivery / Cooling
    能量带宽: 供电 / 冷却
- Inevitable 3DIC Trend & CapEx-Efficient Opportunity

  3DIC的必然趋势和与之对应的高效投资机会

AI’s Incredible Growth



Model Size Growing 10X / year


UK GDP $3T @ 2022

NVDA Market Cap on par with UK GDP

AI HPC Performance: “Two Dark Clouds”

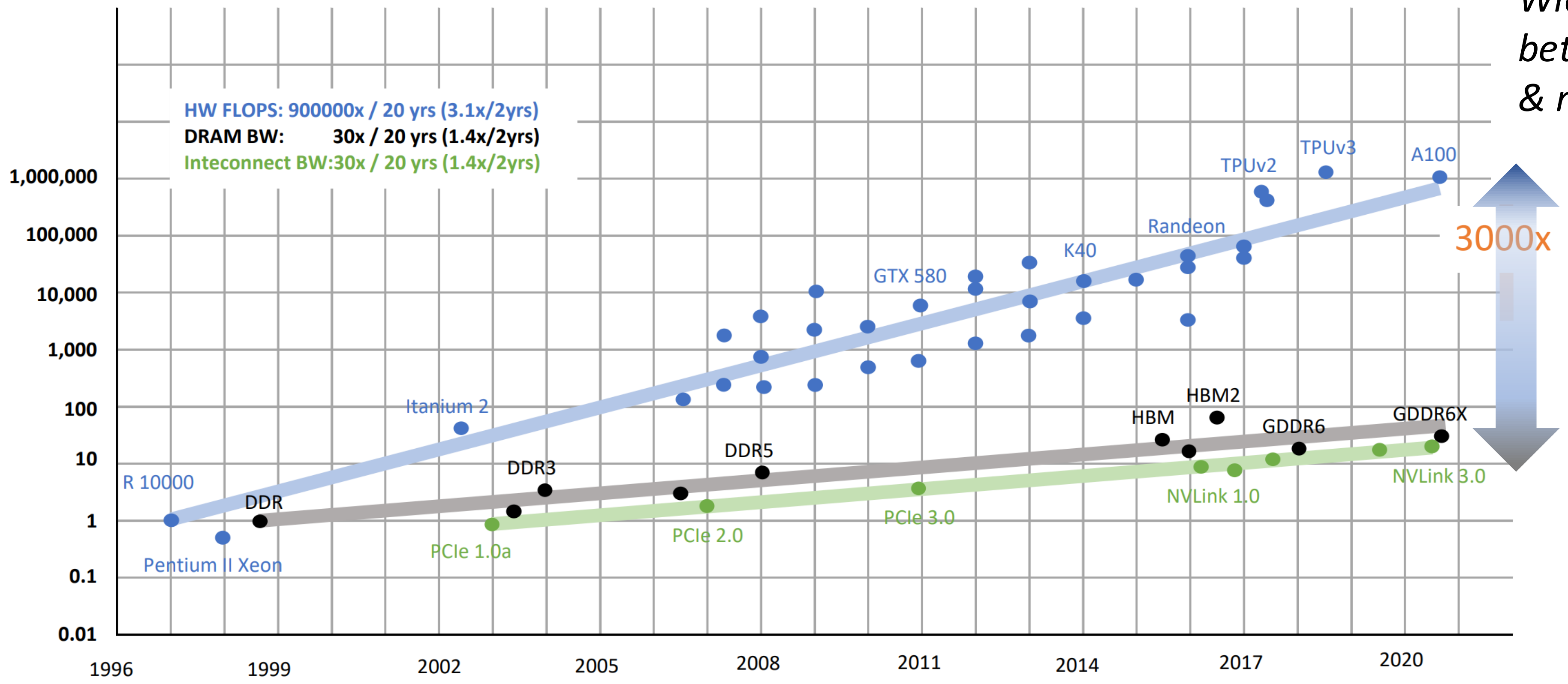
- In 1900, Lord Kelvin called out “two dark clouds” of classical physics, which lead to the discovery of quantum physics and theory of relativity

- TCA’s 50th Anniversary Special Event – the Taiwan Semiconductor Day Forum Debuts in Tokyo on April 2
- Dr. Chen Wen stated in his keynote speech
  - ① Memory Bandwidth Limit
  - ② Energy Bandwidth Limit

Source: https://show.computex.biz/NewsReleaseDetail.aspx?index=42488&category=68

AI Computing Performance Bottleneck

- Hardware performance far outstrips memory bandwidth
- This is a problem, particularly for AI/ML computing



Widening gap between compute & memory bandwidth

Source: Morgan Stanley Research, including (e) estimates.

Performance Limit of Neuro-Network AI/ML

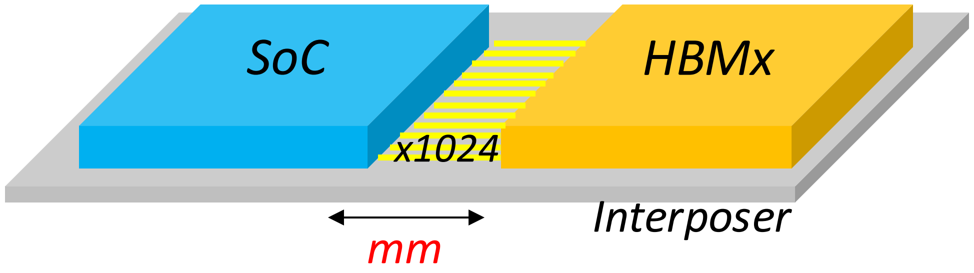
- Fundamental Computation of NN AI/ML

$$\underset{\underset{\text{Output}}{\uparrow}}{B} = \underset{\underset{\text{Large Matrix of Weights}}{\uparrow}}{W} * \underset{\underset{\text{Input}}{\uparrow}}{A}$$

- Measure of AI/ML performance

$$\text{FLOPS} = \frac{\#\text{ of FLOP}}{\text{Sec}} = \frac{\#\text{ of FLOP}}{\text{Byte}} \text{ X } \underset{\text{This is Memory bandwidth}}{\underbrace{\frac{\text{Byte}}{\text{Sec}}}}$$

Scaling Limitations of HBM in 2.5D

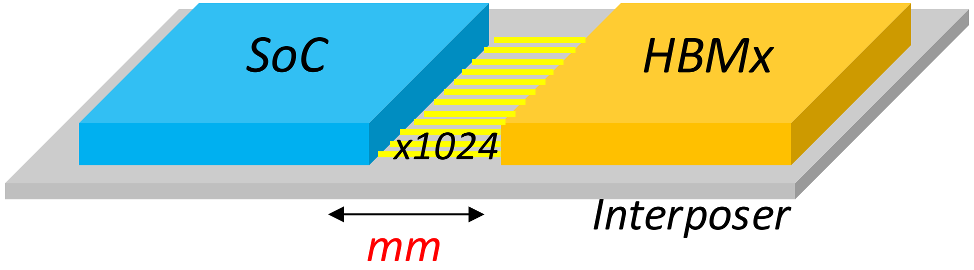


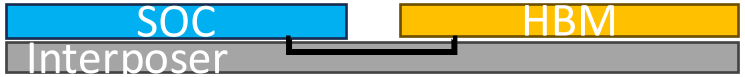
# of connections ➔ Limits Bandwidth

“beach front” of SoC ➔ Limits # of HBM placement

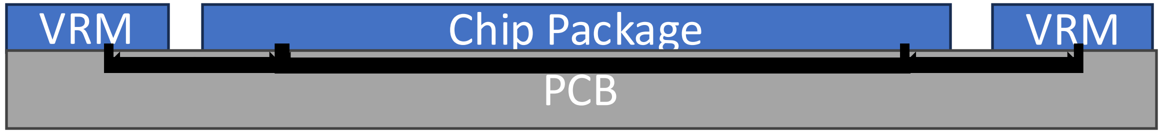
Length of connections ➔ Limits Power consumption

State-of-the-art GPGPU

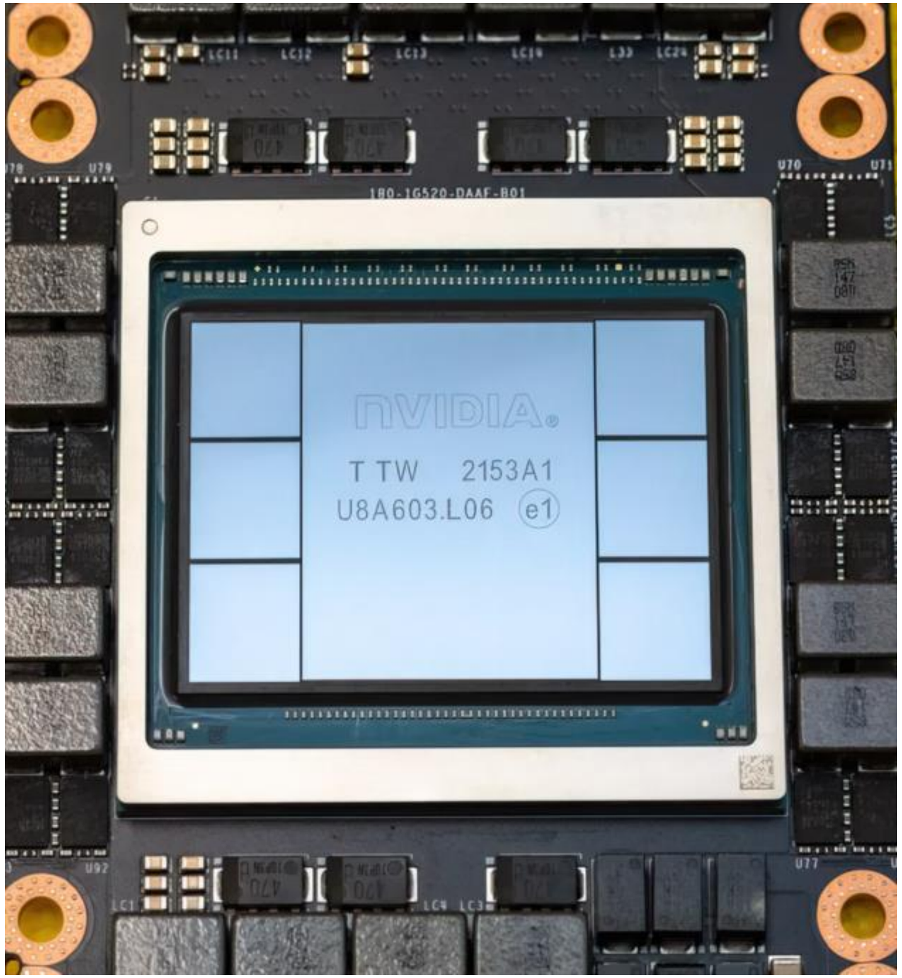




Memory bandwidth



Energy bandwidth

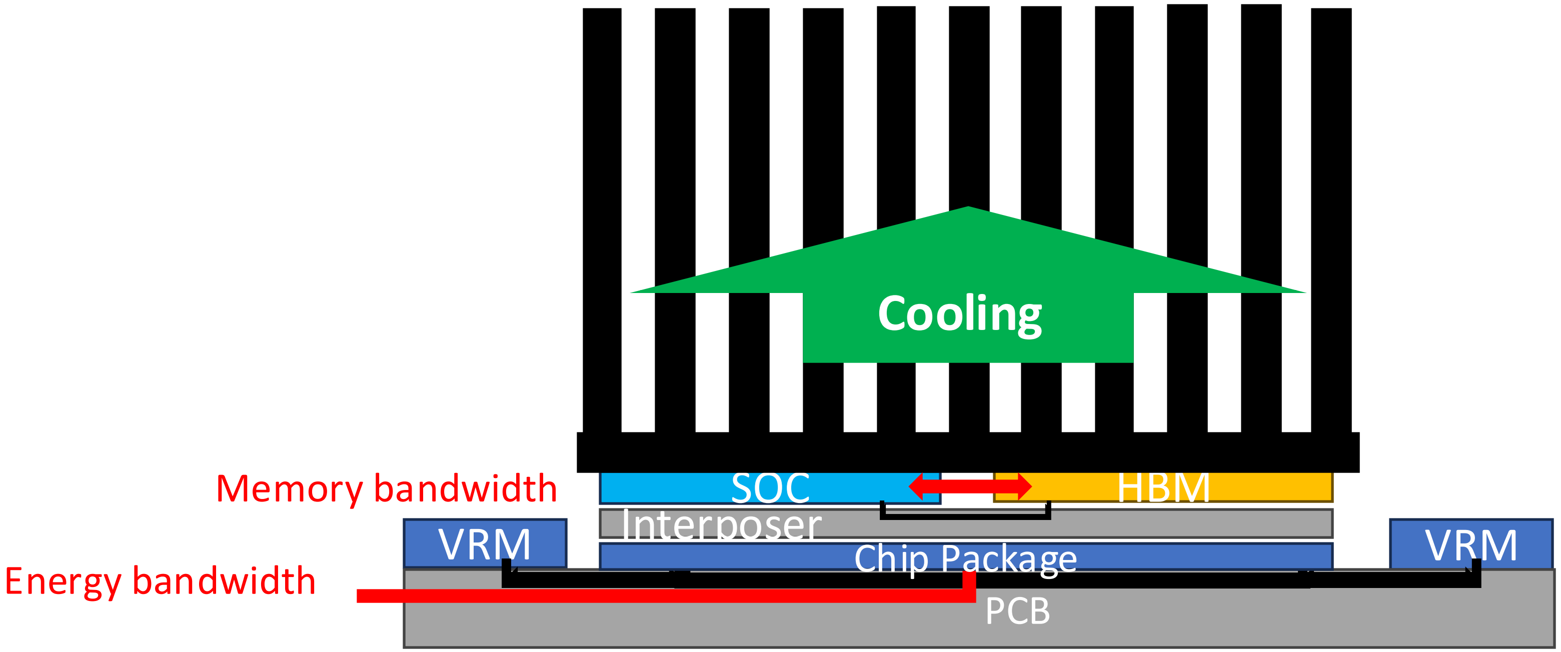


NVIDIA H100

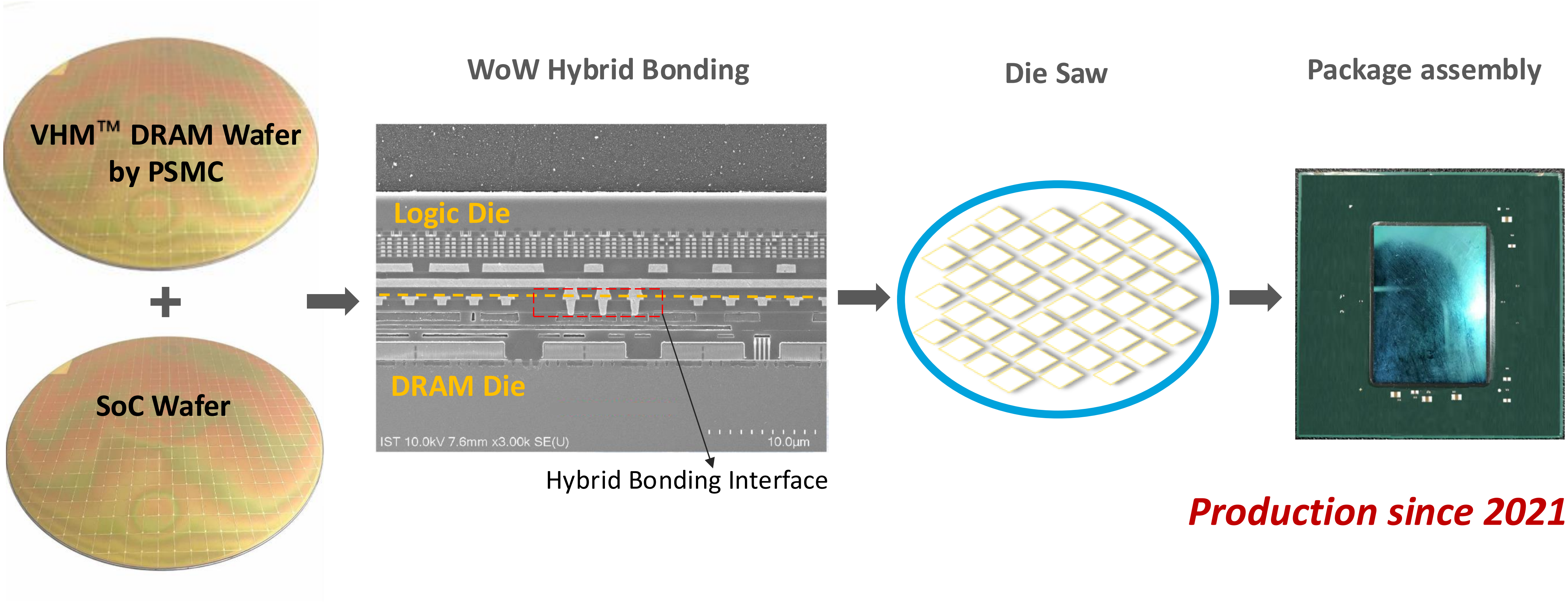
Both memory and energy are delivered horizontally over a 2D plane
“Flux density” (bandwidth/area) limited by physics

The 3rd Dimension

- Cooling is already done vertically in 3D
- Memory and Power must also be delivered vertically in 3D to breakthrough bandwidth limits



Memory in 3D | VHM™ in Production Since 2021



VHM™ DRAM Wafer by PSMC

+

SoC Wafer

WoW Hybrid Bonding

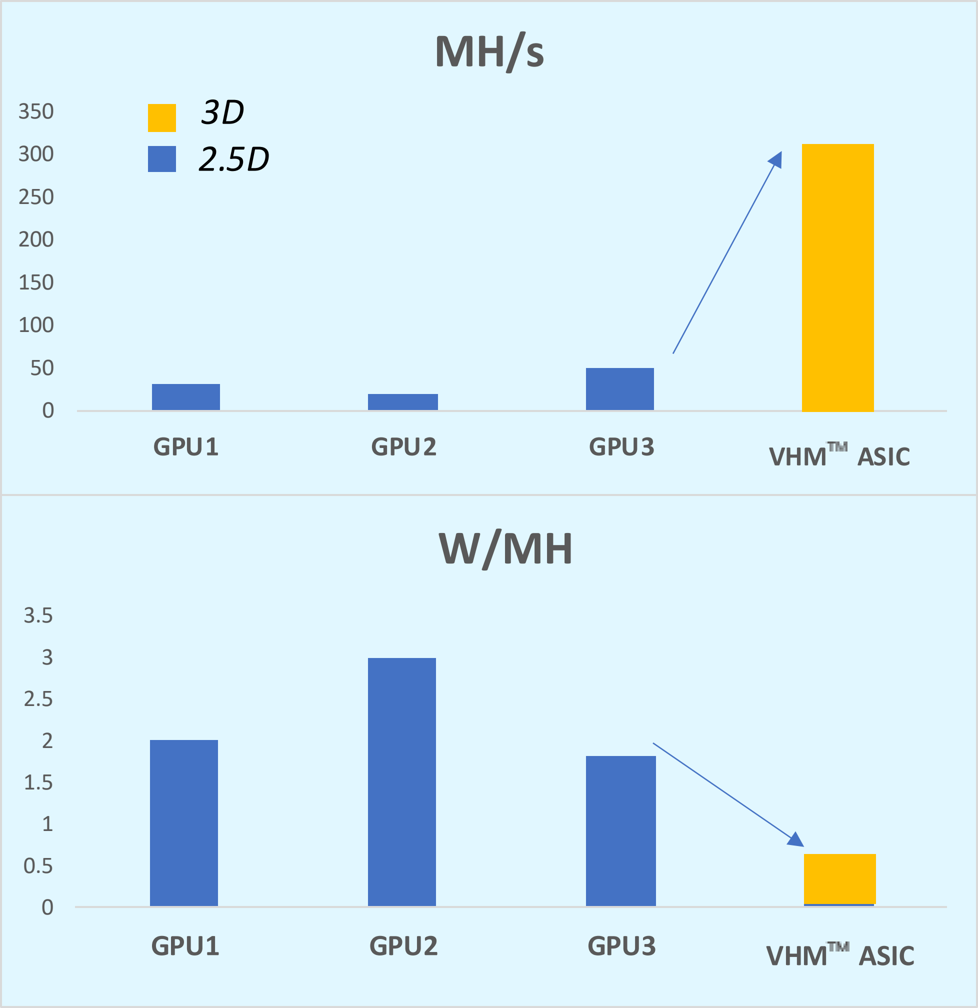
Hybrid Bonding Interface

Die Saw

Package assembly

Production since 2021

Actual Performance of 3DIC-enabled Systems



First Gen 3DIC-enabled Systems vs Leading GPUs

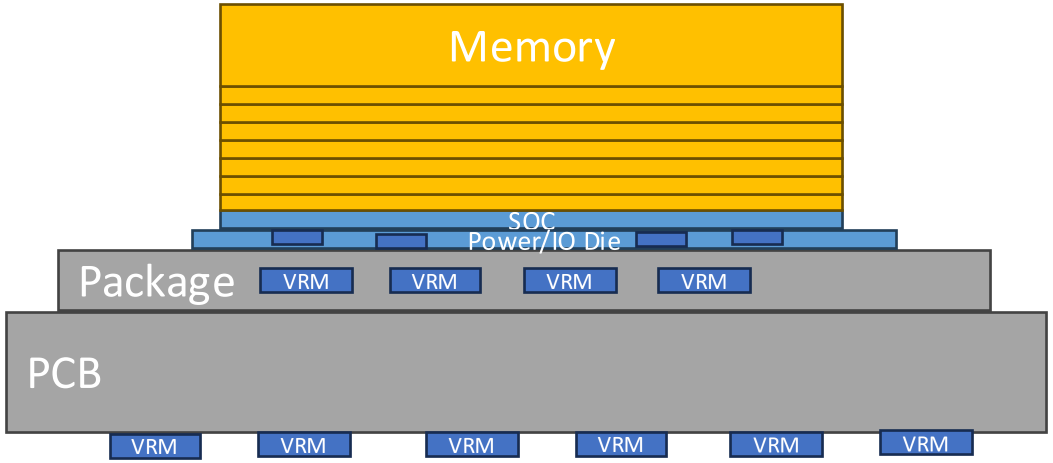
10X Performance

1/3X Power

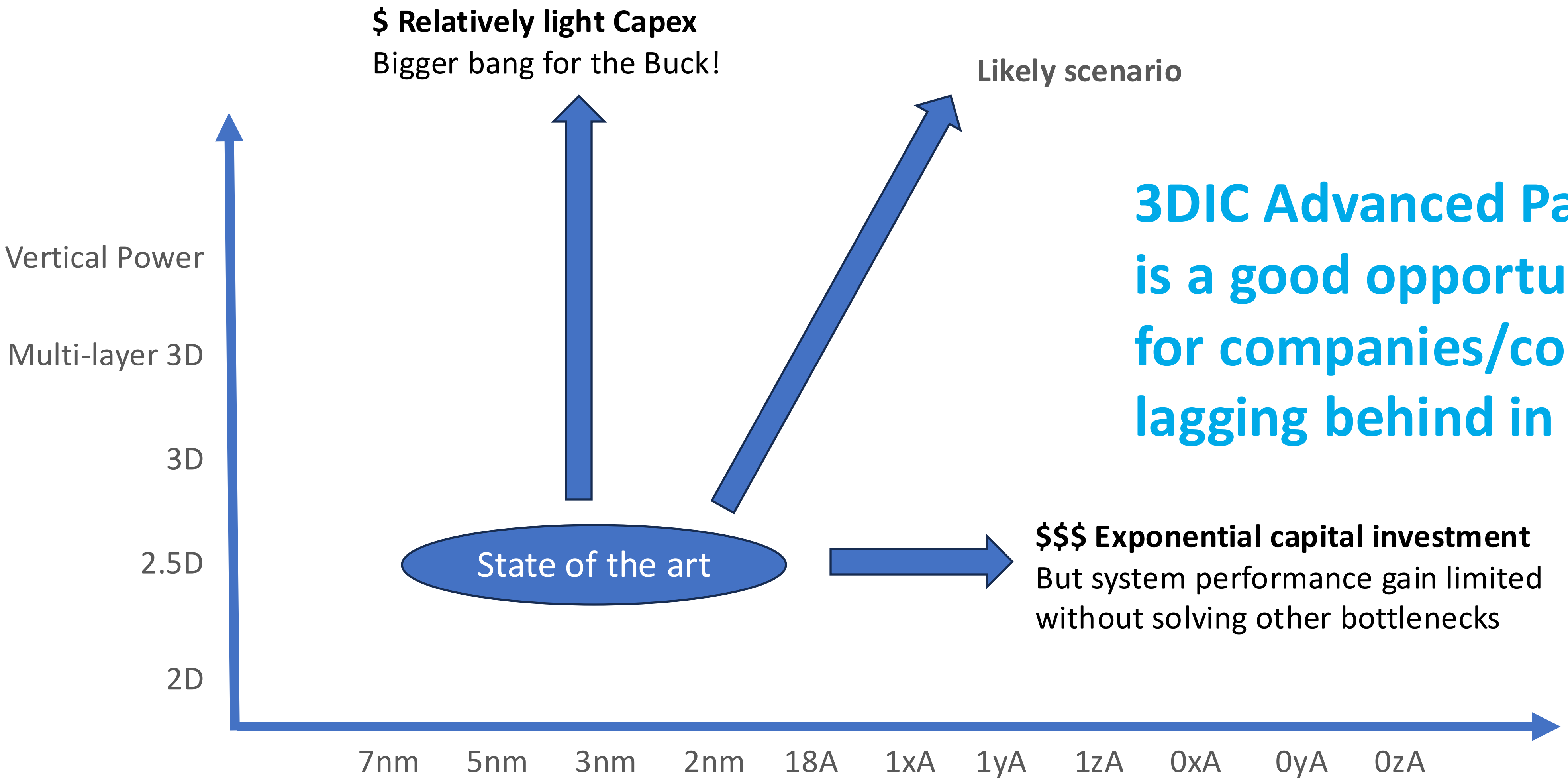
** Based on ETH mining algorithm*

Envisioned Future 3DIC HPC System

- Memory bandwidth in 3D: 10X proven, 1000X possible
- Energy bandwidth in 3D: TBD, >>2X
  - VRM placed in various places vertically



Future Roadmap: Scaling vs 3DIC Packaging



$ Relatively light Capex
Bigger bang for the Buck!

Likely scenario

$$$ Exponential capital investment
But system performance gain limited
without solving other bottlenecks

3DIC Advanced Packaging is a good opportunity for companies/countries lagging behind in scaling

Positioning of apmemory solutions in AI/HPC

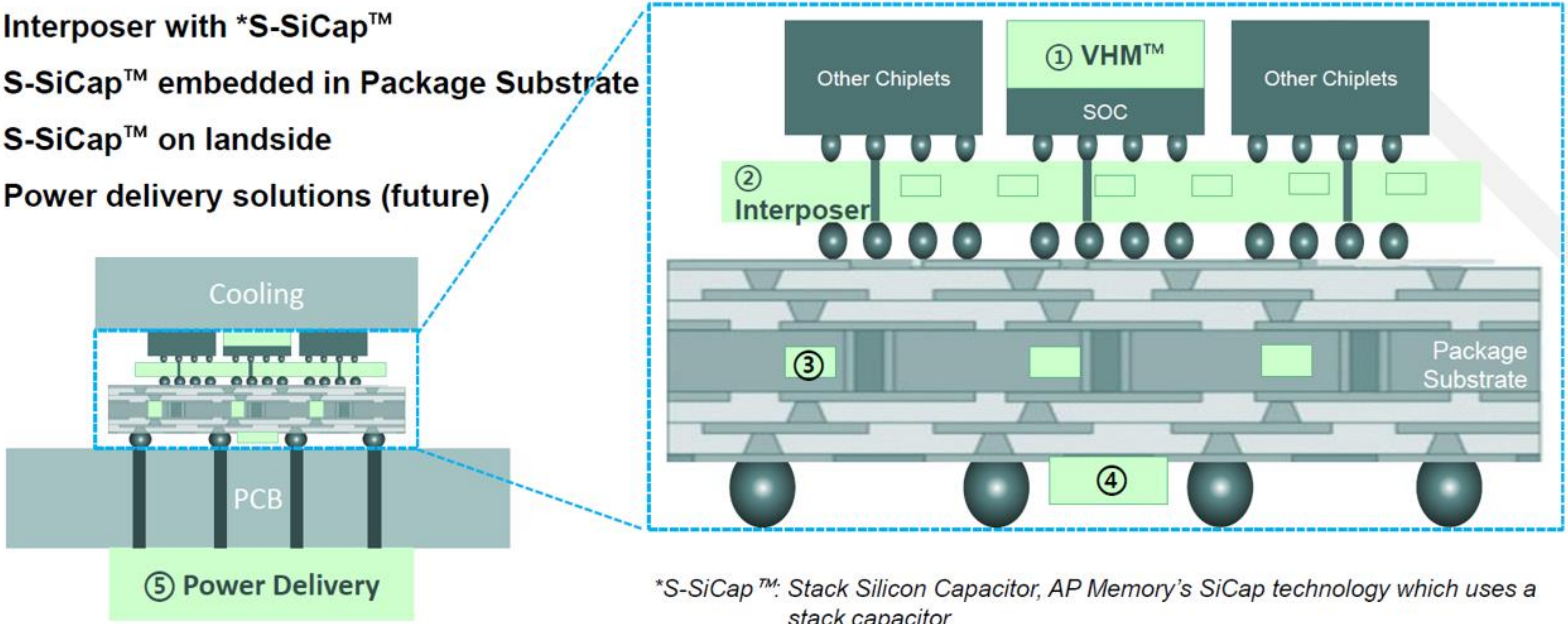
① VHM™ stacked with Compute

② Interposer with *S-SiCap™

③ S-SiCap™ embedded in Package Substrate

④ S-SiCap™ on landside

⑤ Power delivery solutions (future)



*S-SiCap™: Stack Silicon Capacitor, AP Memory's SiCap technology which uses a stack capacitor

About apmemory

apmemory

联系我们

IoTRAM^{TM} – The Ideal Memory for IoT

Leading in product innovation and market share

VHM^{TM} – The AI/HPC Memory Solution

World's first 3D integration of DRAM and SoC chips

S-SiCap^{TM} – The Performance Enabler

Highest capacitance density for high performance applications