A Vision for Modular, Ubiquitous & Scalable Compute Systems

Prepared by Johanna Swan, Intel Fellow & Director of Package Research & System Solutions

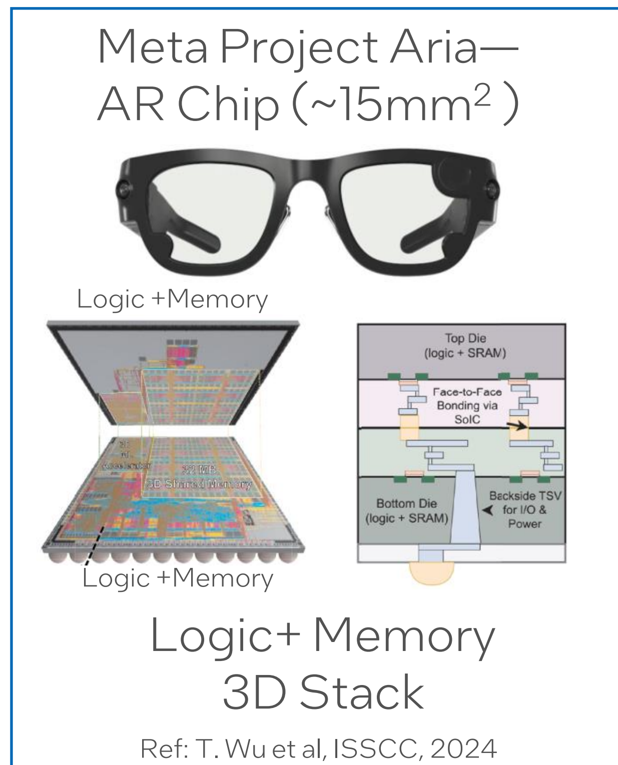
Presented by Bernd Waidhas

Outline

1. Today's Domain Specific Systems & Trends
   - Key thrusts to drive scale for high volume manufacturing
2. Interconnect Scaling
   - Solder vs. Hybrid Bonding
   - Implications & Interactions
   - Challenges & Opportunities
3. System Expansion & Connectivity
   - Methods to Expand the Package
   - Implications & Interactions
   - Challenges & Opportunities
4. Advanced Memory Integration
   - Stacked vs. Side-by-Side on Interposer or Package
   - Implications & Interactions
   - Challenges & Opportunities
5. Summary

Domain Specific Compute System Trends

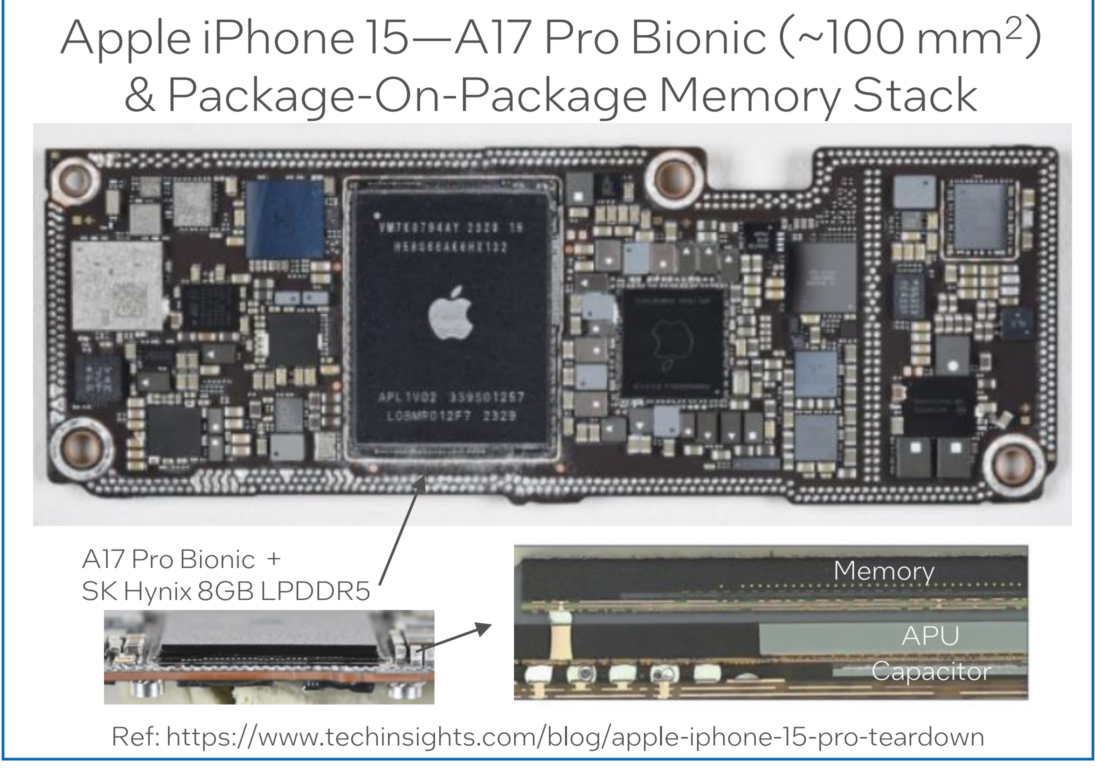
Meta Project Aria—AR Chip (~15mm^2)



Logic+ Memory
3D Stack

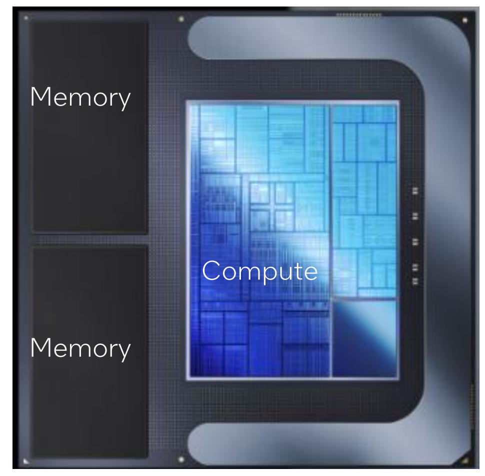
Ref: T. Wu et al, ISSCC, 2024

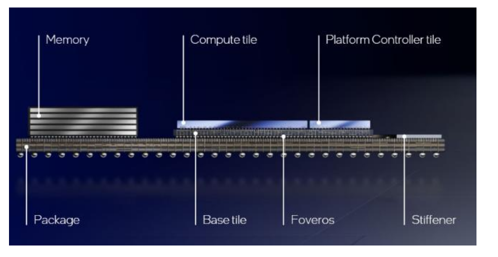
Apple iPhone 15—A17 Pro Bionic (~100 mm^2) & Package-On-Package Memory Stack



Ref: https://www.techinsights.com/blog/apple-iphone-15-pro-teardown

Intel Lunar Lake 3D with co-packaged Memory (~740mm^2)





Ref: Intel Newsroom

Each market drives a specific form factor (size) and affordability target.

AI drives a need for more memory capacity and higher bandwidth with lower latency even for mobile products.

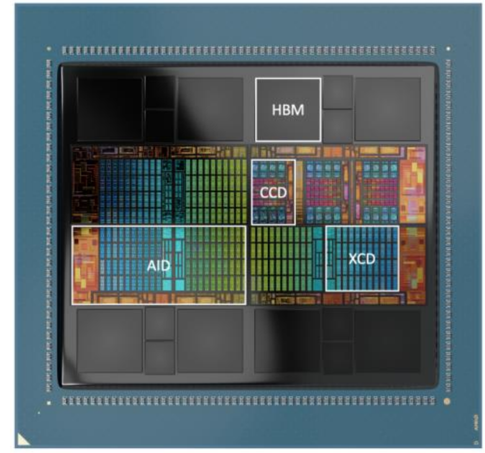
Domain Specific Compute System Trends (cont.)

Nvidia H200 GPU
2.5D + 3D (~3k mm^2)

Ref: https://www.nvidia.com/en-us/data-center/h100/

141 GB HBM3e
+ 4.8 TB/s Memory Bandwidth
+NVLink 900GB/s
+PCIe Gen 5: 128GB/s

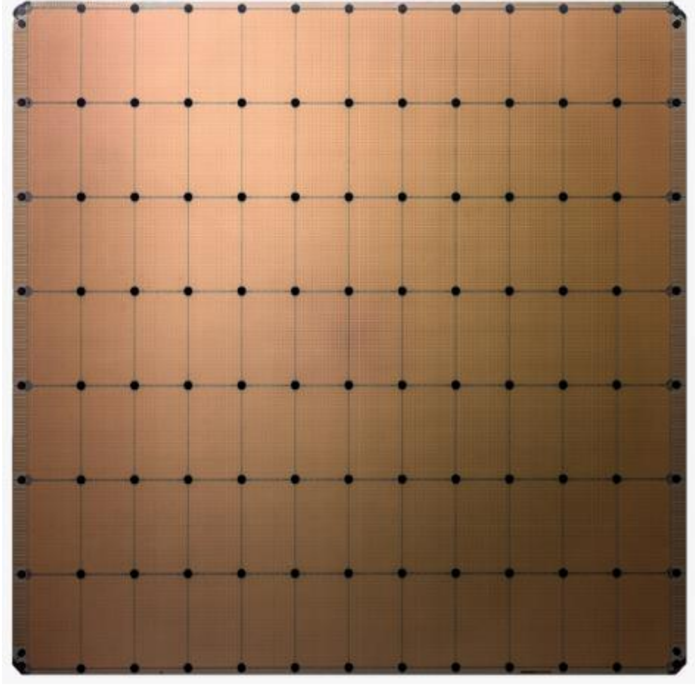
AMD MI300X Accelerator
Modular 2.5D + 3D (~6k mm^2)



Ref: A. Smith et al, ISSCC, 2024

192 GB HBM3e
+ 5.3 TB/s Memory Bandwidth
+ Infinity Fabric™ 896 GB/s
+PCIe Gen5: 128 GB/s

Cerebras Wafer Scale Engine
(~46k mm^2)



Ref: https://cerebras.ai/product-chip/

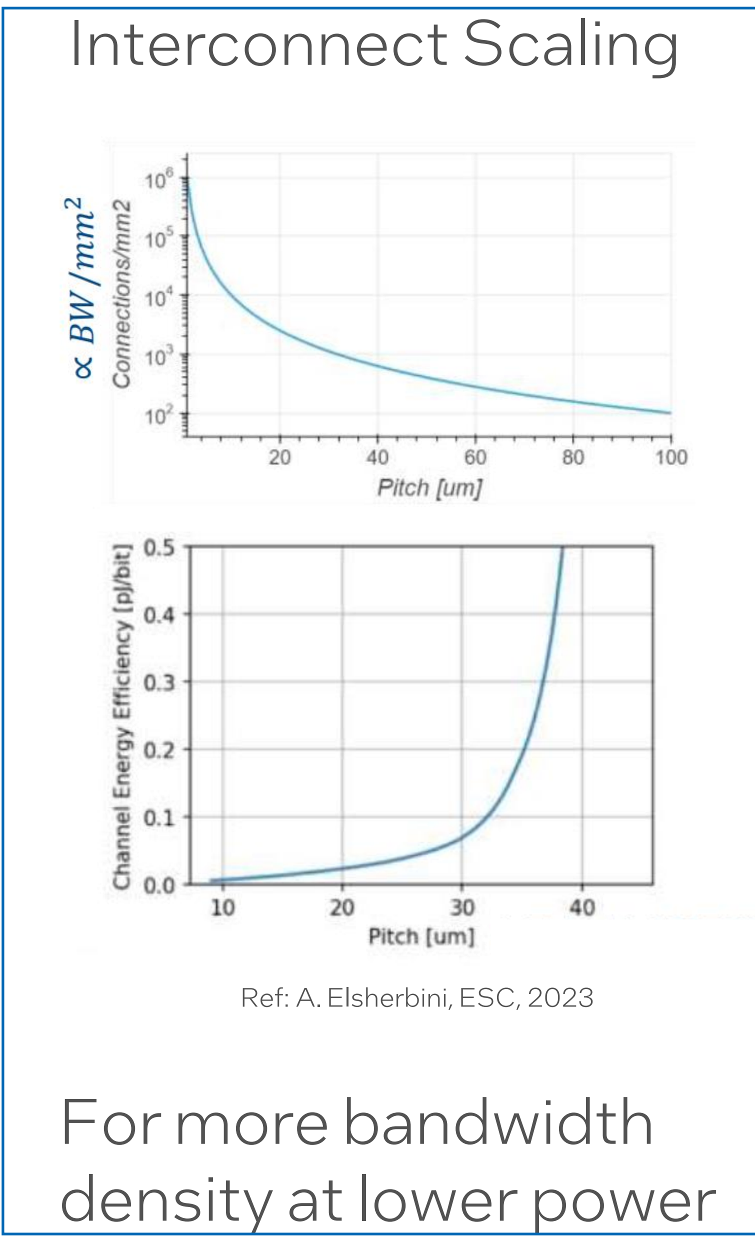
44GB on-chip memory
+ 21 PB/s Memory Bandwidth
+214 PB/s Fabric Bandwidth

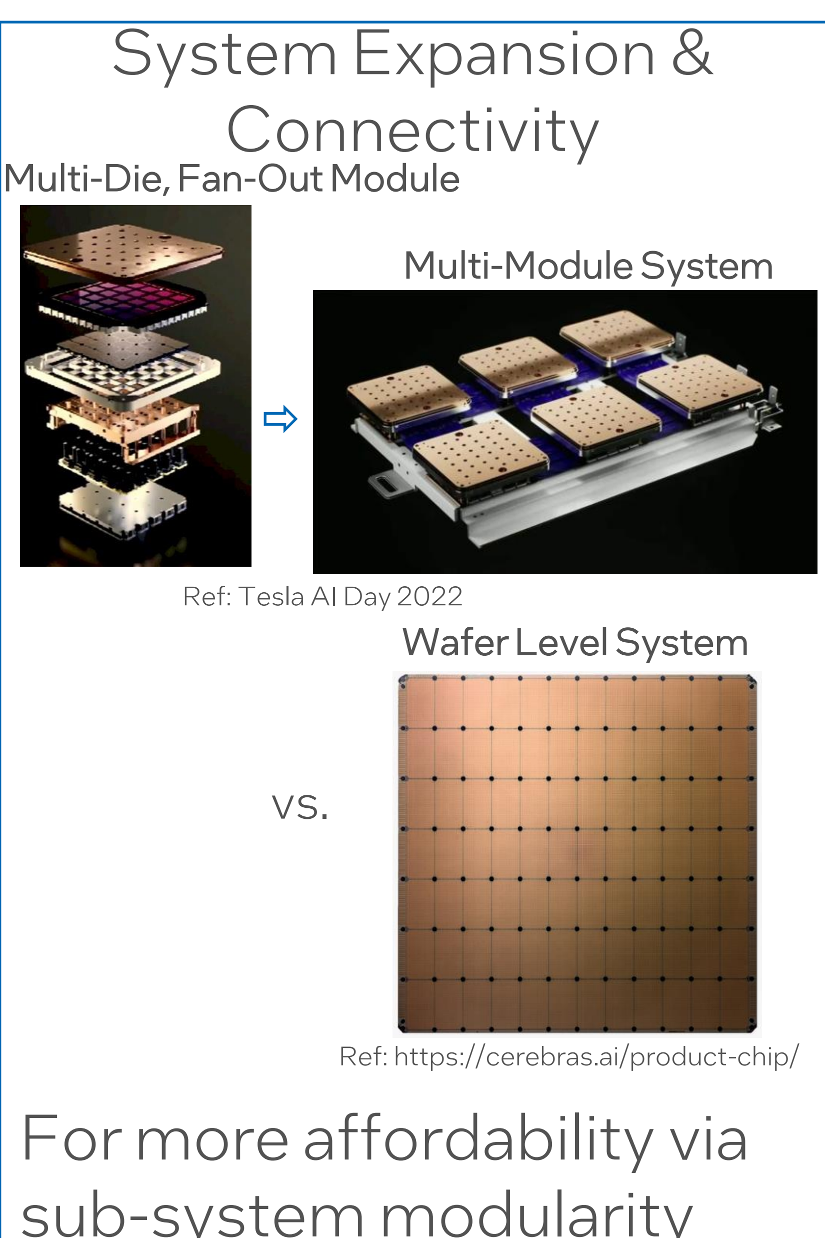
For high performance compute memory capacity & bandwidth as well as interconnect bandwidth are driving system architecture solutions

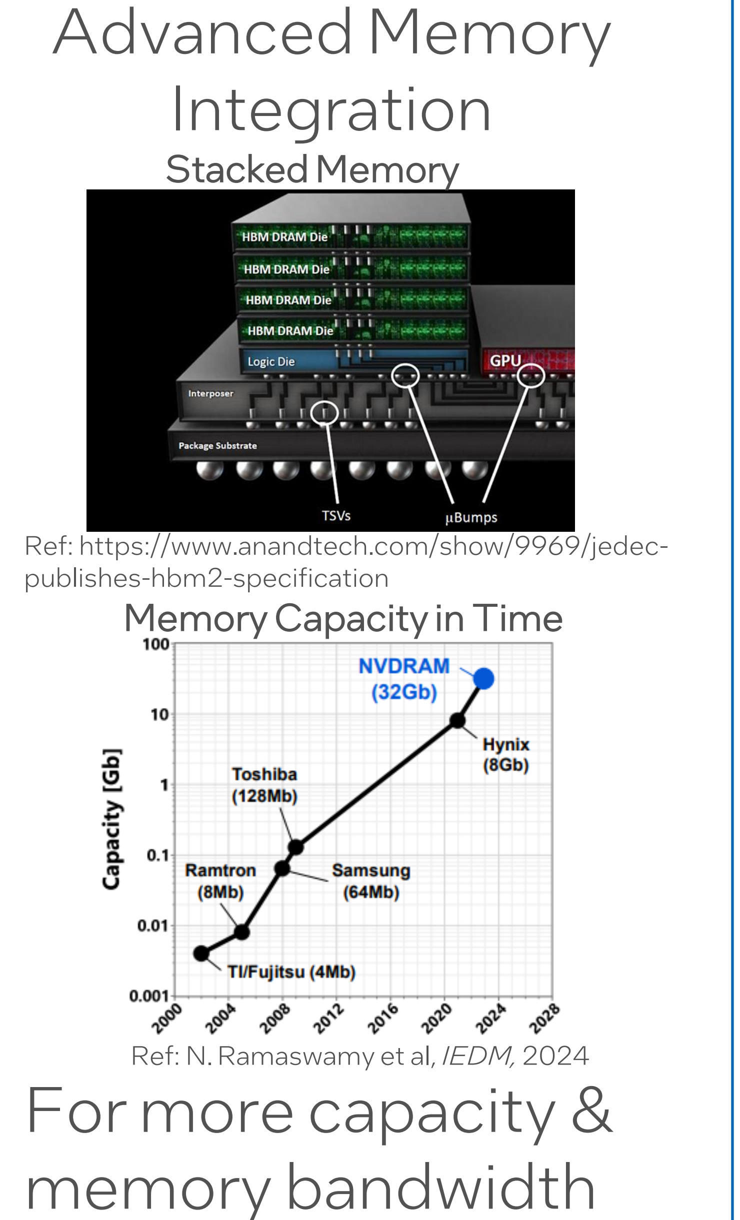
2.5 & 3D stacking is utilized yet still the package is growing in area

Modular chips with higher interconnect bandwidth are utilized to reduce cost

3 Key Thrusts are Evident







Outline

1. Today's Domain Specific Systems & Trends
   - Key thrusts to drive scale for high volume manufacturing
2. Interconnect Scaling
   - Solder vs. Hybrid Bonding
   - Implications & Interactions
   - Challenges & Opportunities
3. System Expansion & Connectivity
   - Methods to Expand the Package
   - Implications & Interactions
   - Challenges & Opportunities
4. Advanced Memory Integration
   - Stacked vs. Side-by-Side on Interposer or Package
   - Implications & Interactions
   - Challenges & Opportunities
5. Summary

Interconnect Scaling: Solder vs. Hybrid Bond (HBI) Assembly

Solder Assembly (Pitch >10 um)

Alignment

Alignment at Temperature

14.1 µm

Solder Collapse

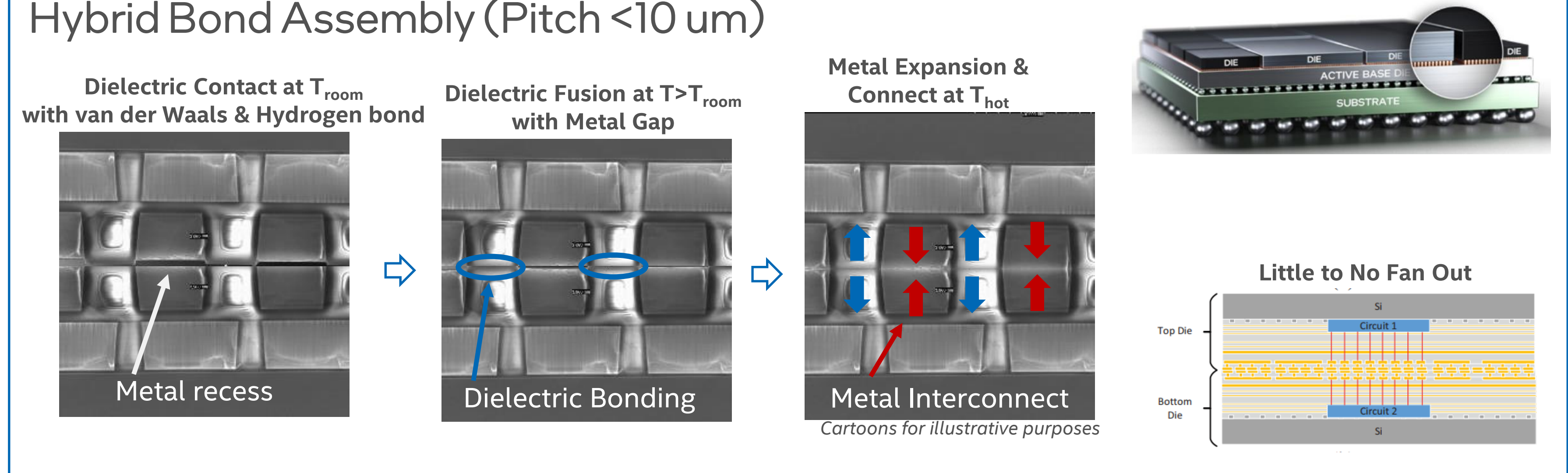
Temperature, Pressure & Displacement

Silica-Filled Polymer Reinforcement

DIE
ACTIVE OR PASSIVE
EMIB
SUBSTRATE

Fan Out

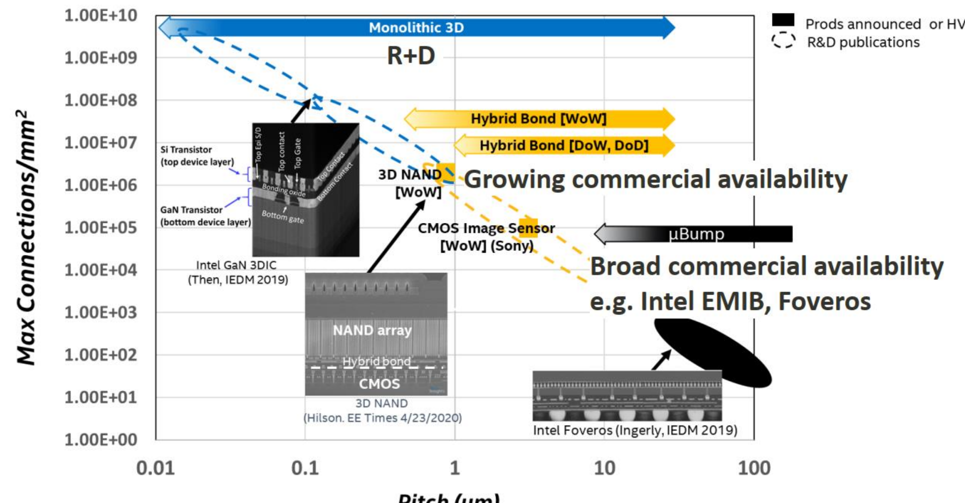
Si
Circuit 1
Top Die
Bottom Die
Circuit 2
Si



Solder & Hybrid Bond Assembly may be utilized hierarchically to make composite packages.

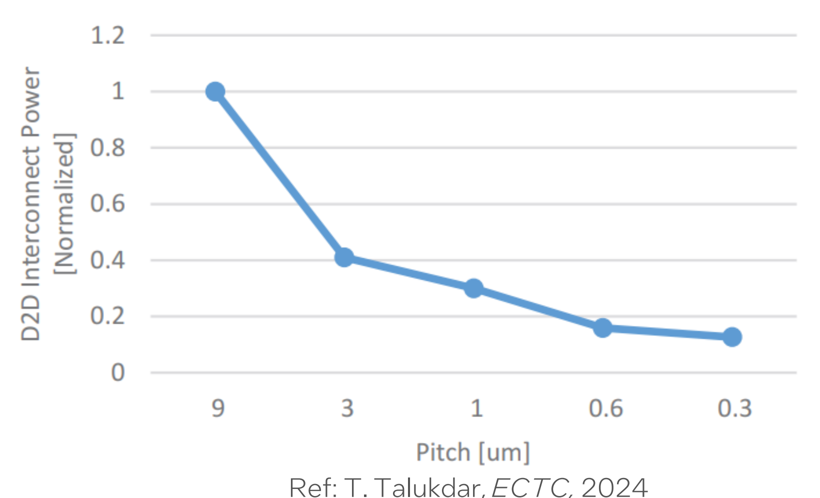
Interconnect Scaling Implications & Interactions

IO Density Scales With Interconnect Pitch



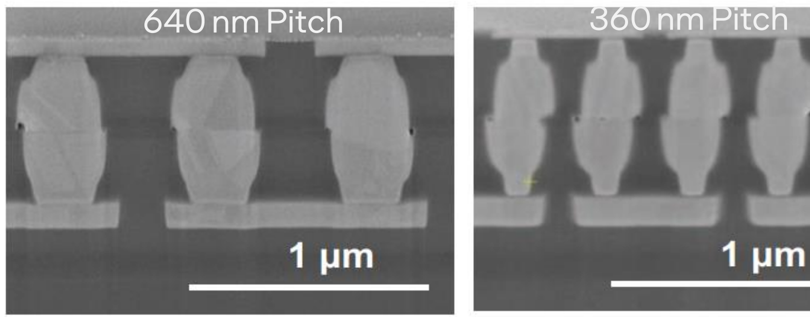
Ref: P. Fisher & F. Sheikh, *DARPA ERI Summit,* Aug 2020

Die to Die Power Decreases With Pitch



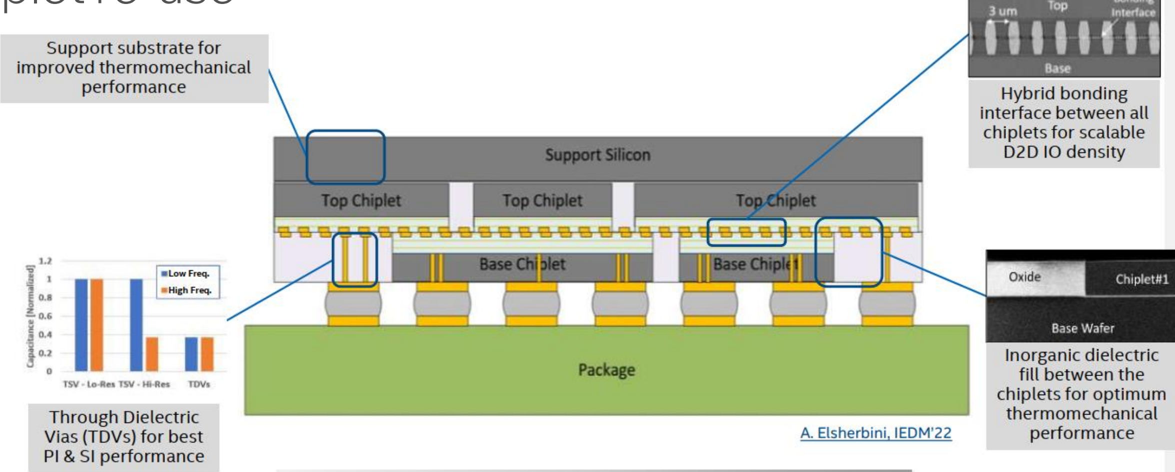
Ref: T. Talukdar, *ECTC,* 2024

Potential to enable assembled interconnect densities at >7M connections/mm^2



Ref: T. Talukdar, *ECTC,* 2024

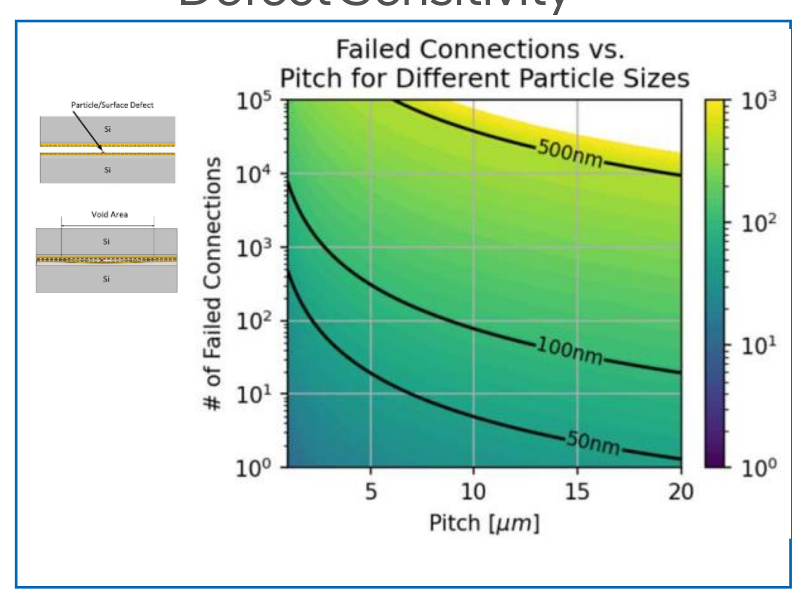
By utilizing quasi-monolithic integration features and materials may be selected for maximal performance & chiplet re-use

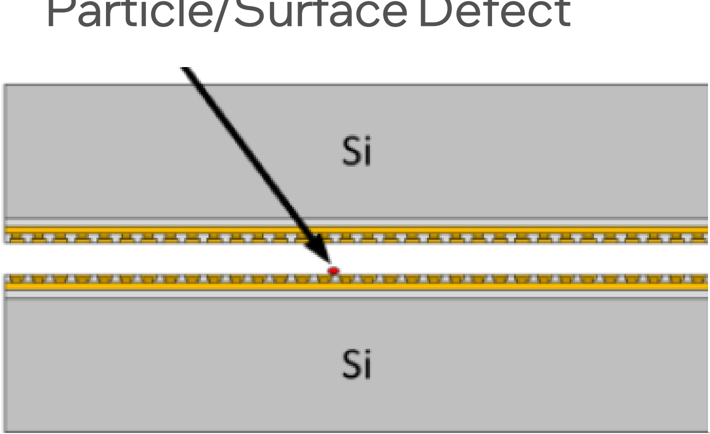


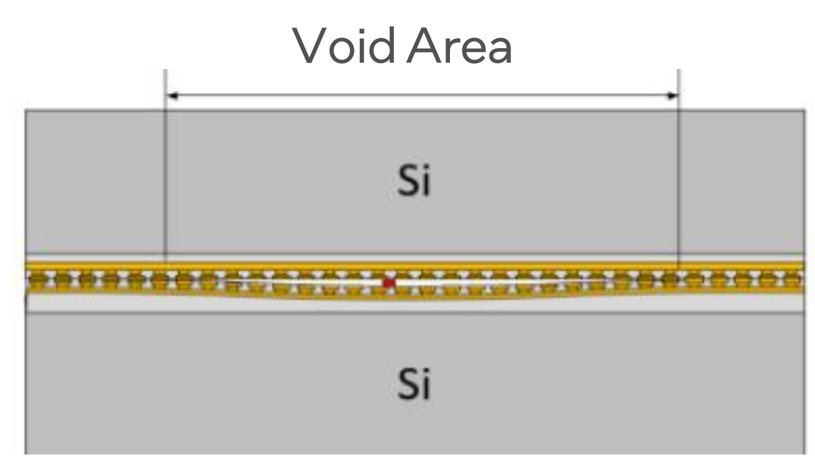
Ref: A. Elsherbini, *ECS,* 2023

Interconnect Challenges & Opportunities

Defect Sensitivity







Ref: A. Elsherbini, *IEDM,* 2021

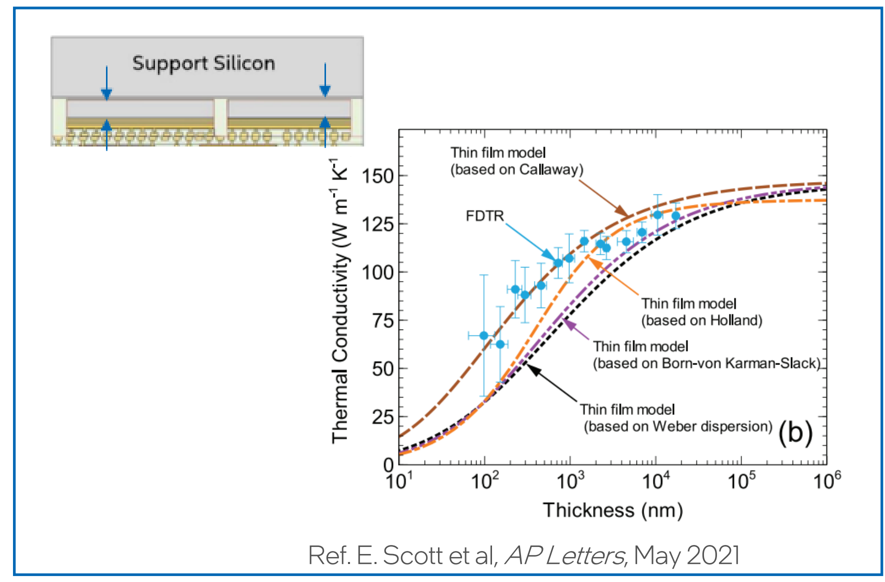
Hybrid bonding drives significant cost to enable, i.e.

- Planarization
- Cleanliness
- Precision die placement

Opportunity: Lower-Cost, Pitch Scaled Assembly that is more defect tolerant.

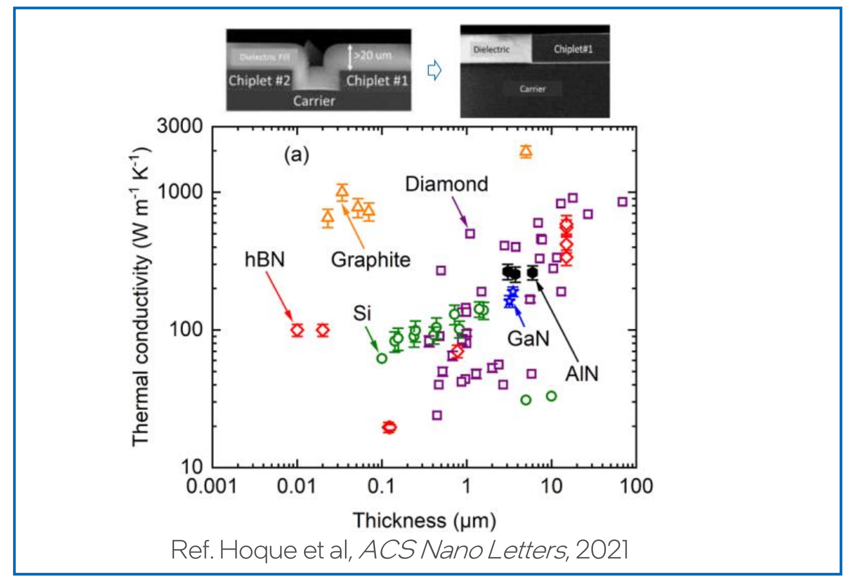
Interconnect Challenges & Opportunities

Thin Die Integration



Ref. E. Scott et al, *AP Letters*, May 2021

Dielectric Thermal Conductivity



Ref. Hoque et al, *ACS Nano Letters*, 2021

Today's systems rely on thin die for stacked integration and encapsulated with SiO_2

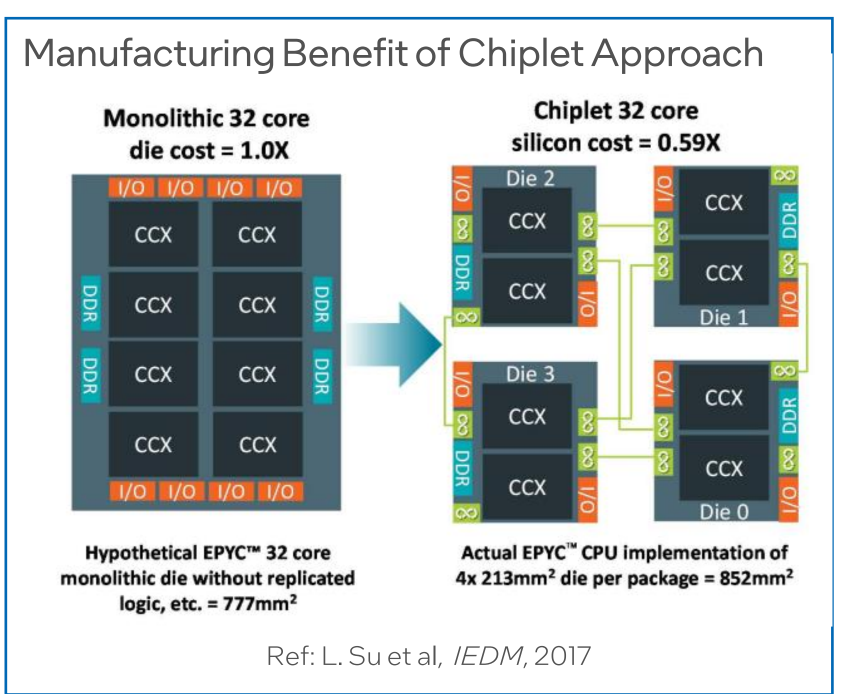
- Thermal Conductivity decays when silicon thickness < 10 um
- Dielectric Constant ~3.9
- Thermal Conductivity ~1.3 W/mK

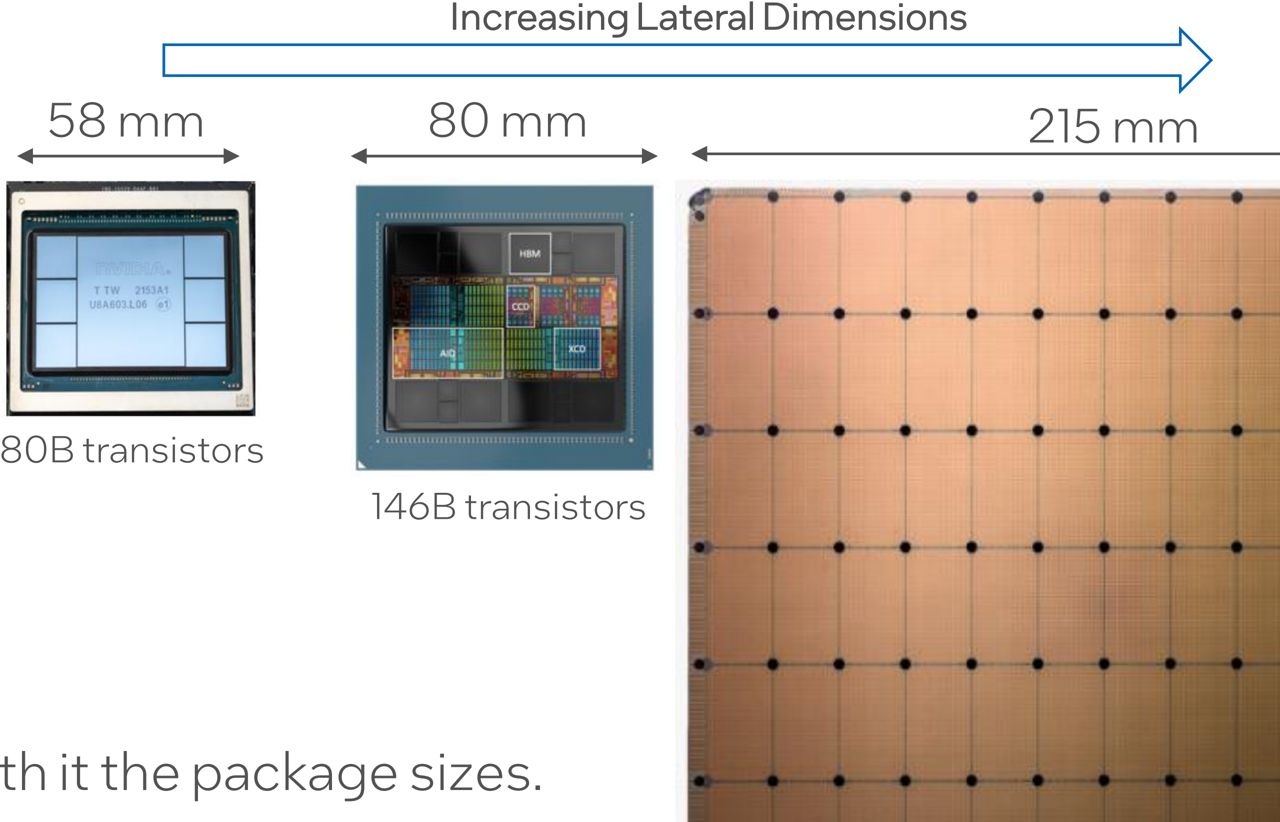
Opportunity: Enable higher thermal conductivity integration schemes with low dielectric constant materials

Outline

1. Today's Domain Specific Systems & Trends
   - Key thrusts to drive scale for high volume manufacturing
2. Interconnect Scaling
   - Solder vs. Hybrid Bonding
   - Implications & Interactions
   - Challenges & Opportunities
3. System Expansion & Connectivity
   - Methods to Expand the Package
   - Implications & Interactions
   - Challenges & Opportunities
4. Advanced Memory Integration
   - Stacked vs. Side-by-Side on Interposer or Package
   - Implications & Interactions
   - Challenges & Opportunities
5. Summary

Systems are Expanding in Dimension



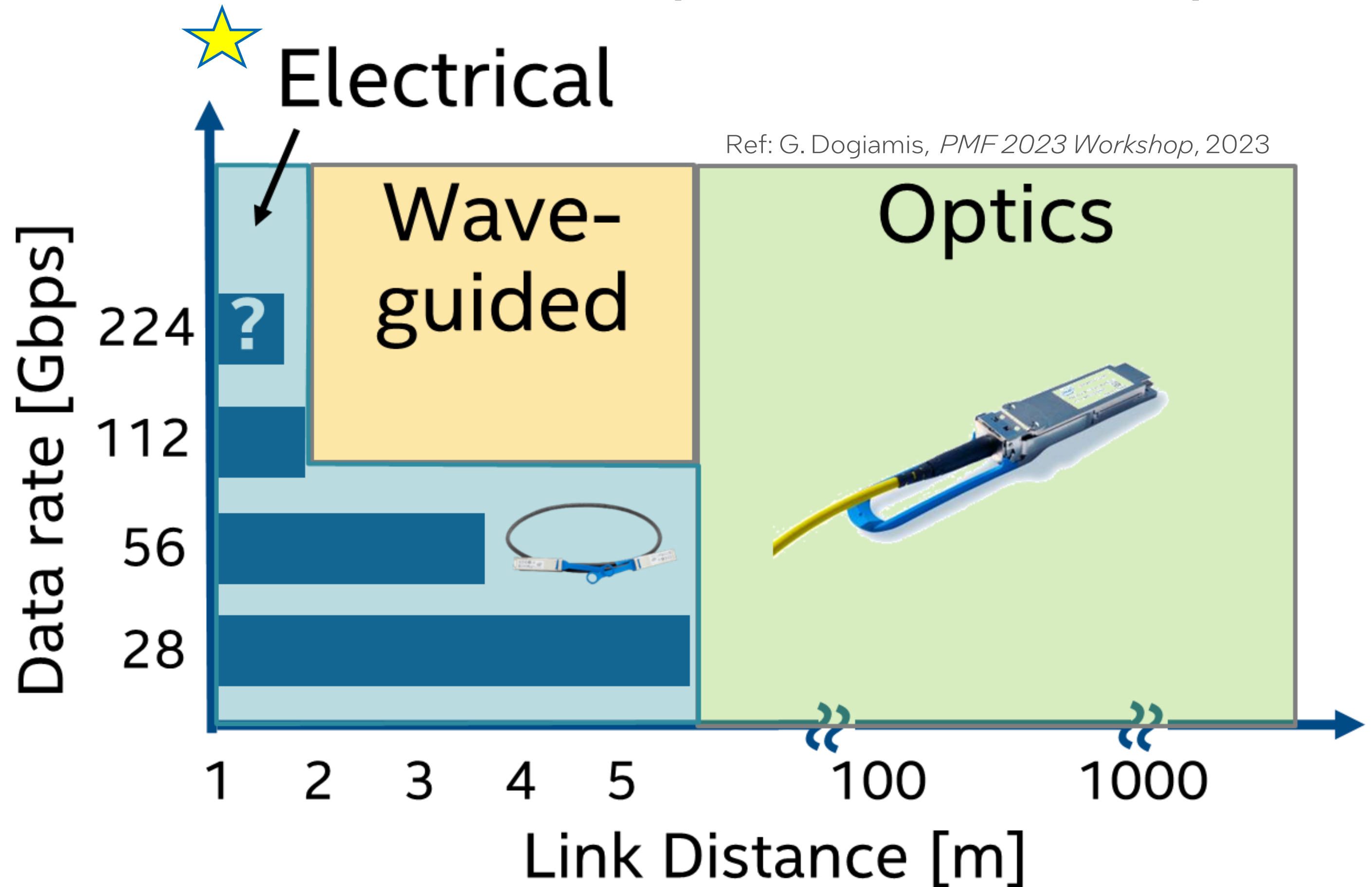


AI is driving system expansion and with it the package sizes.

Solutions include either

- Transition from Wafer to Panel level fan out "packages" or
- Modular connectorized systems across large areas

Methods to Expand the System (Not the Package)

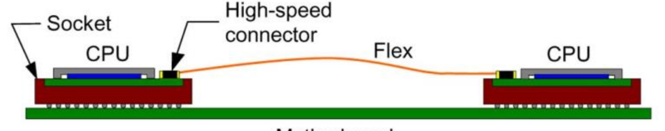


As data rates increase—the reach may be penalized by power and cost overheads.

System co-optimization for bandwidth, power efficiency, bit error rate, physical fit and cost provide a means to modularize solutions

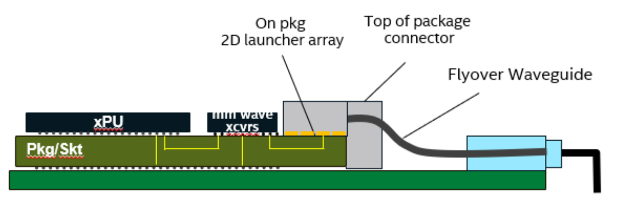
Connector Solutions

Electrical Top Side Connector Solutions



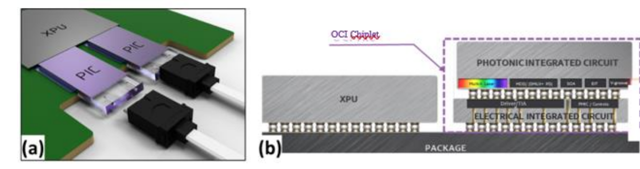
Ref: H. Braunisch et al, *IEEE Trans on Adv. Packaging*, 2008

mm-wave Guided Interconnect Solutions



Ref: G. Dogiamis, *PMF 2023 Workshop*, 2023

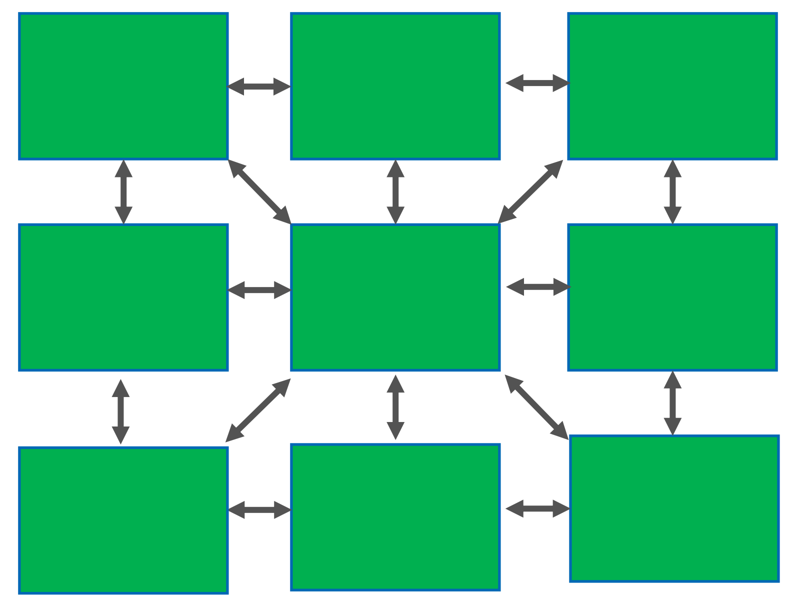
Co-Packaged Photonics Connector Solutions



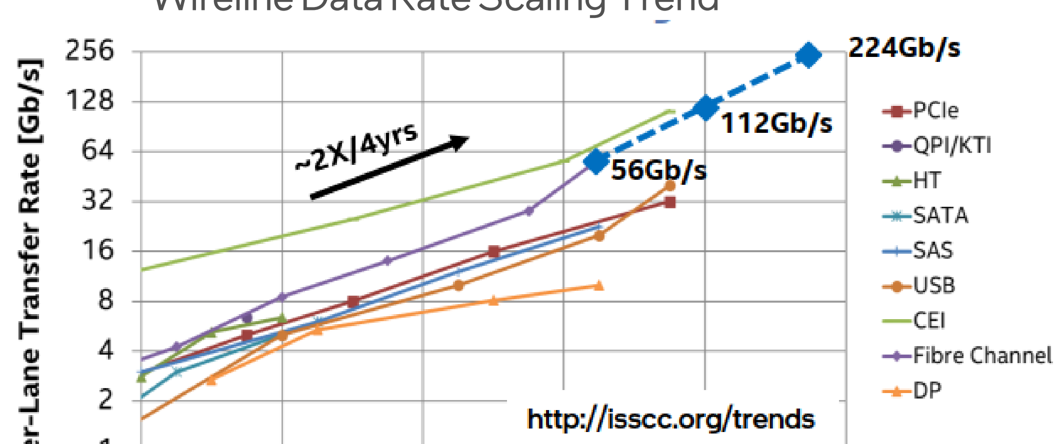
Ref: S. Fathololoumi et al, *Hot Chips*, 2024

System Expansion Challenges & Opportunities

System Expansion Via Multiple Interconnected Packages



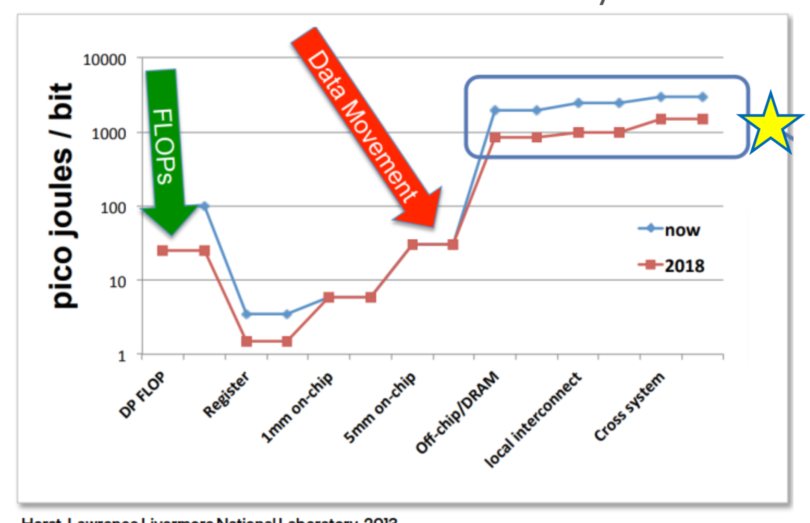
Wireline Data Rate Scaling Trend



Challenges:

- Shoreline restrictions limiting fit & on-to-off package transition
- Data rate scaling with power efficiency improvements
- Extremely low loss, reworkable, cost-affordable connectors

Data Movement Power Efficiency Trend



Horst, Lawrence Livermore National Laboratory, 2013

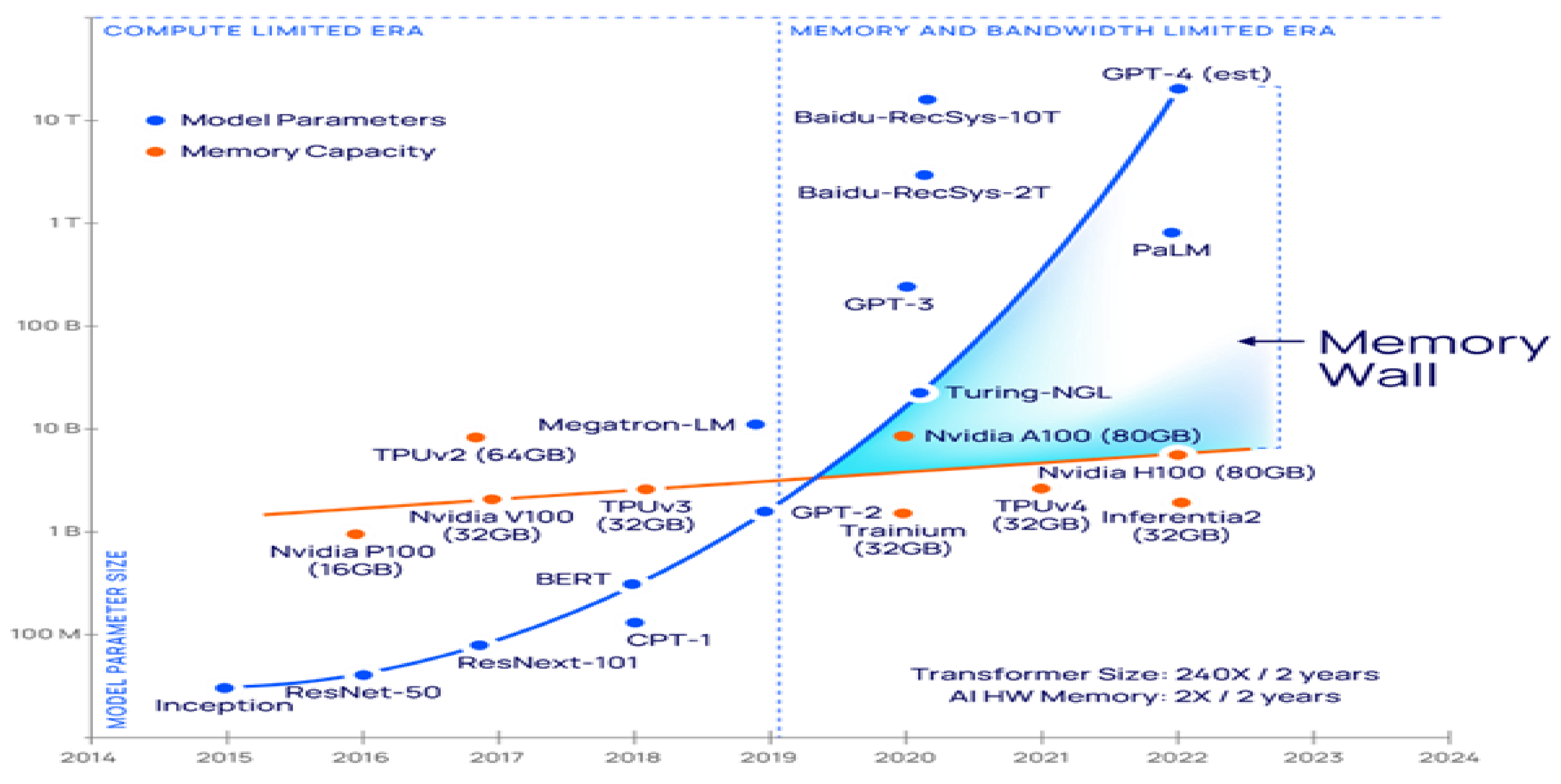
Opportunity: to modularize packages for affordability

Ref: E. Karl, *CICC*, 2022

Outline

1. Today's Domain Specific Systems & Trends
   - Key thrusts to drive scale for high volume manufacturing
2. Interconnect Scaling
   - Solder vs. Hybrid Bonding
   - Implications & Interactions
   - Challenges & Opportunities
3. System Expansion & Connectivity
   - Methods to Expand the Package
   - Implications & Interactions
   - Challenges & Opportunities
4. Advanced Memory Integration
   - Stacked vs. Side-by-Side on Interposer or Package
   - Implications & Interactions
   - Challenges & Opportunities
5. Summary

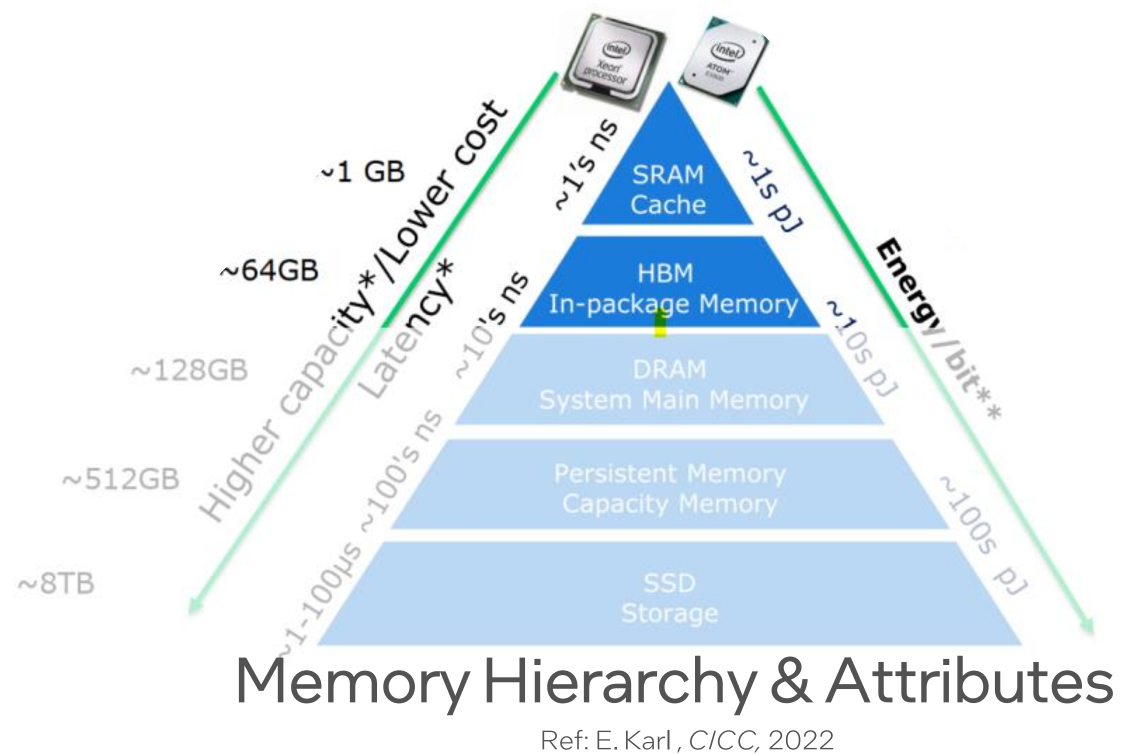
Memory Integration is even more challenging for AI LLM



AI compute and memory demands are growing exponentially.
[Training Compute-Optimal LLMs (Mar '22), Celestial.ai]

Today memory capacity & interconnect bandwidth limit AI Large Language Models

Memory Integration Key Challenge



Memory Hierarchy & Attributes

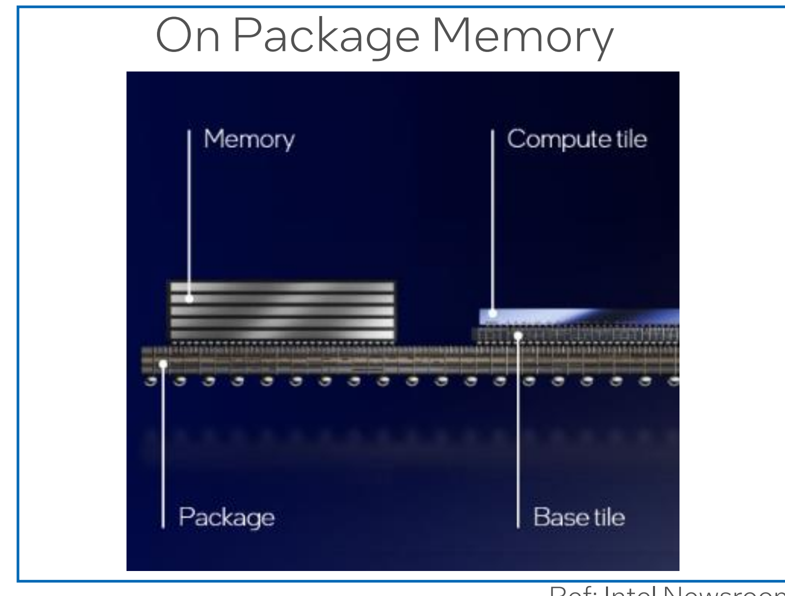
Ref: E. Karl , *CICC*, 2022

More memory capacity & bandwidth is needed, driving

- In/Near Memory Compute Advancements and/or
- More in-package memory with hierarchical caching & higher density interconnects

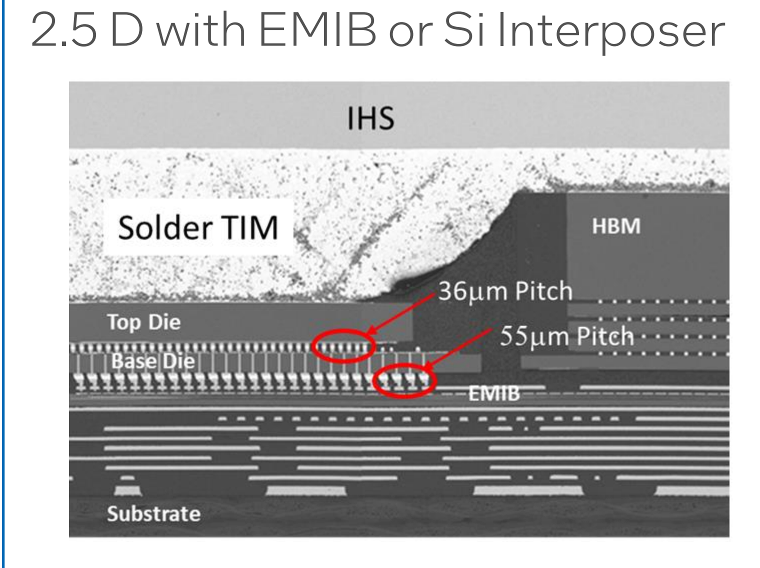
Memory Implementations Today

On Package Memory



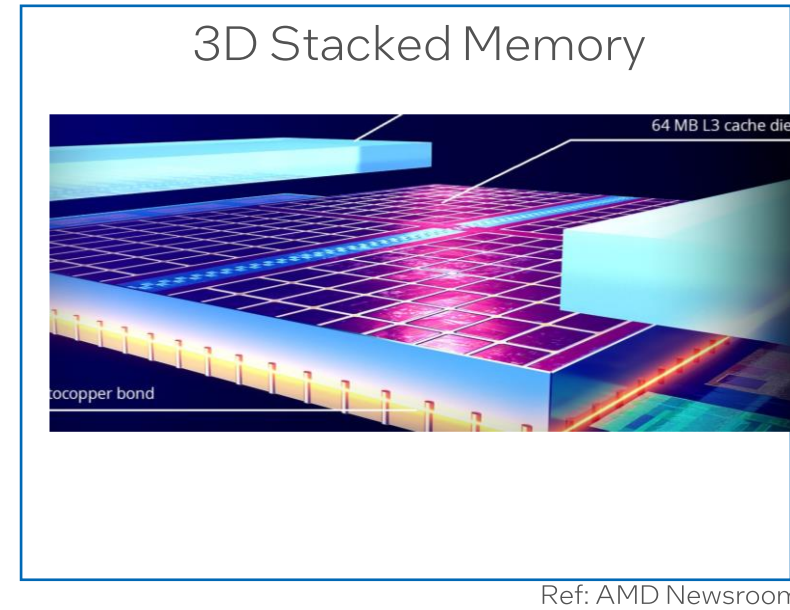
Ref: Intel Newsroom

2.5 D with EMIB or Si Interposer



Ref: W. Gomes, ISSCC, 2020

3D Stacked Memory



Ref: AMD Newsroom

Bump pitch ~100um
Bump density ~100/mm^2
Power ~0.5 pJ/bit

Memory soldered to package
Normalized latency

Bump pitch ~55-36um
Bump density ~330-772/mm^2
Power ~0.25 pJ/bit

Memory soldered to passive silicon
Relatively Lower latency

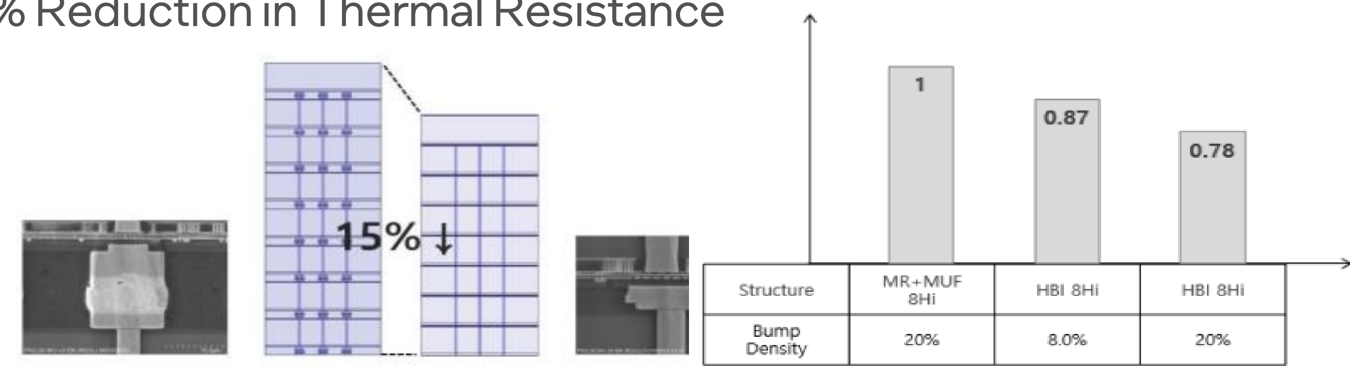
Bump pitch ~15-9um
Bump density ~ 4300-12000 mm^2
Power ~0.05 pJ/bit

Cache stacked Hybrid Bond
~8x lower latency than 55um

Demand for memory stacking is at an all time high

Memory Integration Key Challenge Thermal Management

Utilizing Hybrid Bonding Instead of Solder in HBM stacks Can Drive +22% Reduction in Thermal Resistance



Structure	MR+MUF 8Hi	HBI 8Hi	HBI 8Hi
Bump Density	20%	8.0%	20%

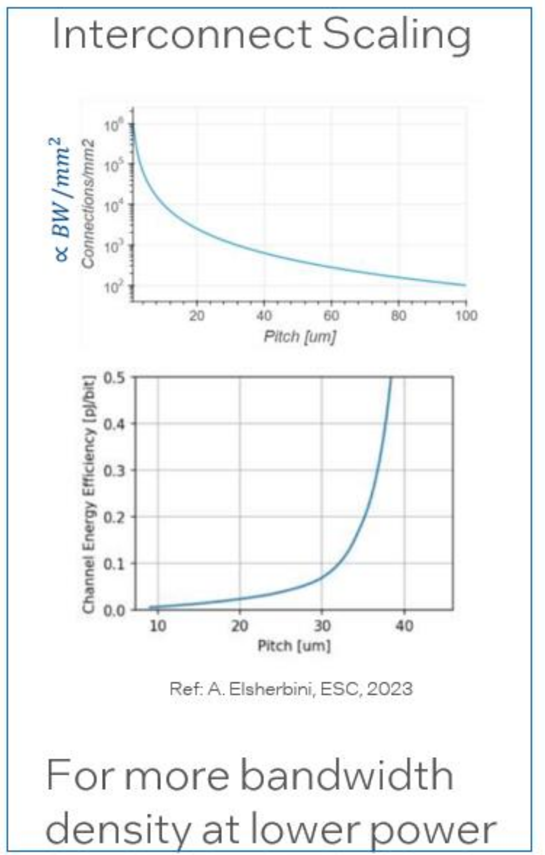
Ref: K. Kim et al, *ECTC,* 2023

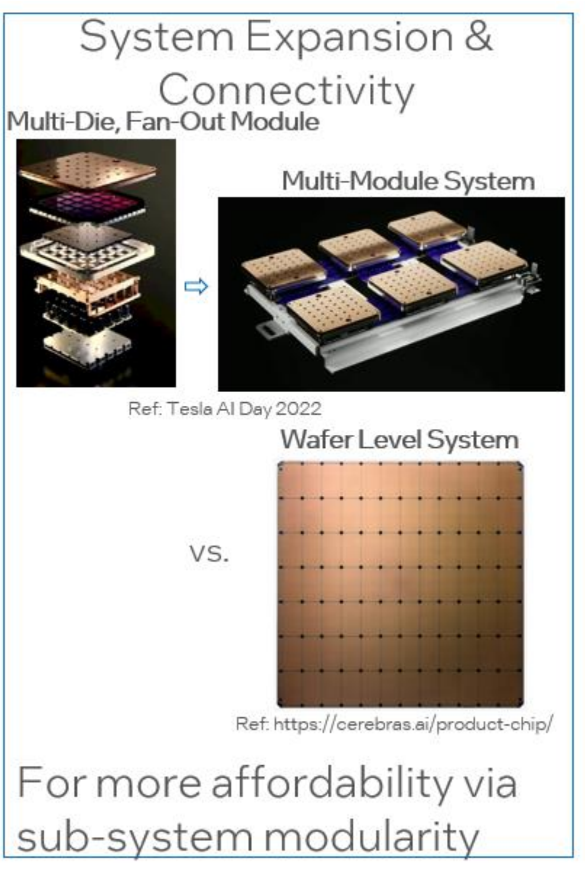
Thermal management is a challenge, despite hybrid bond assembly providing a thermal conductivity advantage.

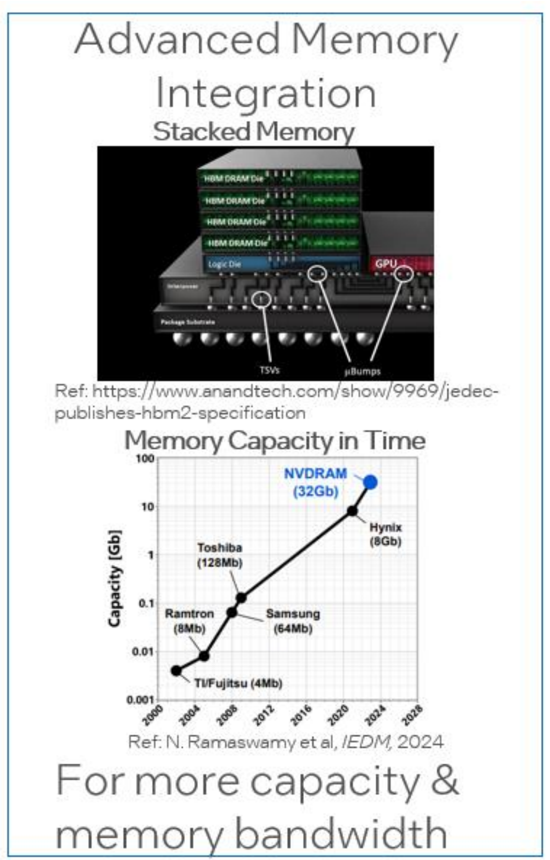
- Memory on logic translates to even higher power density
- Tighter coupling leads to more thermal cross-talk
- Available thermal dissipative power (TDP) is relatively fixed per form factor

Opportunities: Unified control of both processor and memory power in logic to memory stack & disruptive TDP enablers.

Summary







From augmented reality to high performance compute, there is a demand for more memory capacity and higher interconnect bandwidth with lower latency.

- Technology which can be modularized and scaled out enables affordable, ubiquitous compute system solutions
- Three key vectors for addressing this vision of modularized solutions include: a) interconnect scaling, b) system expansion & connectivity, and c) advanced memory integration
- For each market, co-optimization across these thrusts are needed to meet the form factor & performance target.

All product and service plans, and roadmaps are subject to change without notice. Any forecasts of products, services or technologies needed for Intel's operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document. Code names are often used by Intel to identify products, services or technologies that are in development and usage may change over time. Product, service and technology performance varies by use, configuration and other factors. No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Learn more at www.Intel.com/PerformanceIndex and www.Intel.com/ProcessInnovation.

Reference to research results, including comparisons to products, services or technology performance are estimates and do not imply availability. The products and services described may contain defects or errors which may cause deviation from published specifications. Current characterized errata are available on request. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade. Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings releases, annual report on form 10-K and other SEC filings at www.intc.com.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others. This document contains information on products and technologies in development.