Lithography Technologies to Support High Performance Computing and Advancing AI

Toshi Hisamura, Principal Member of Technical Staff, Silicon Technology
May 24, 2024

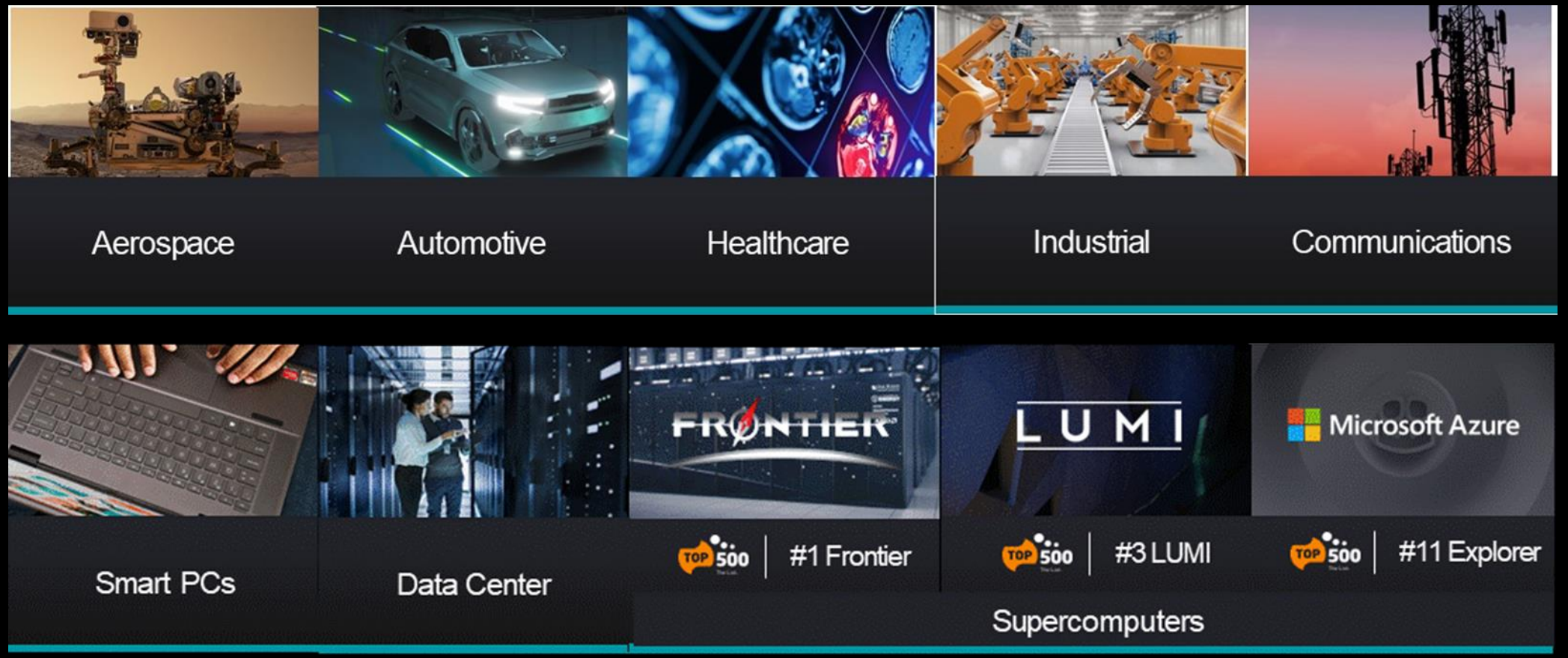
Lithography Technologies to Support High Performance Computing and Advancing AI

- Advancing AI with High Performance Computing
- Lithography technologies critical to High Performance Computing
- Manufacturing cost challenges
- Summary

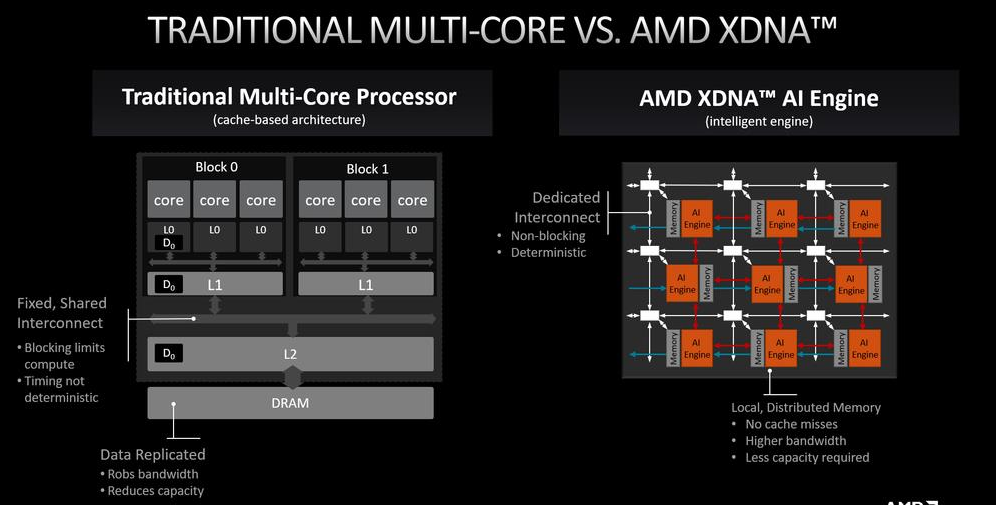
Lithography Technologies to Support High Performance Computing and Advancing AI

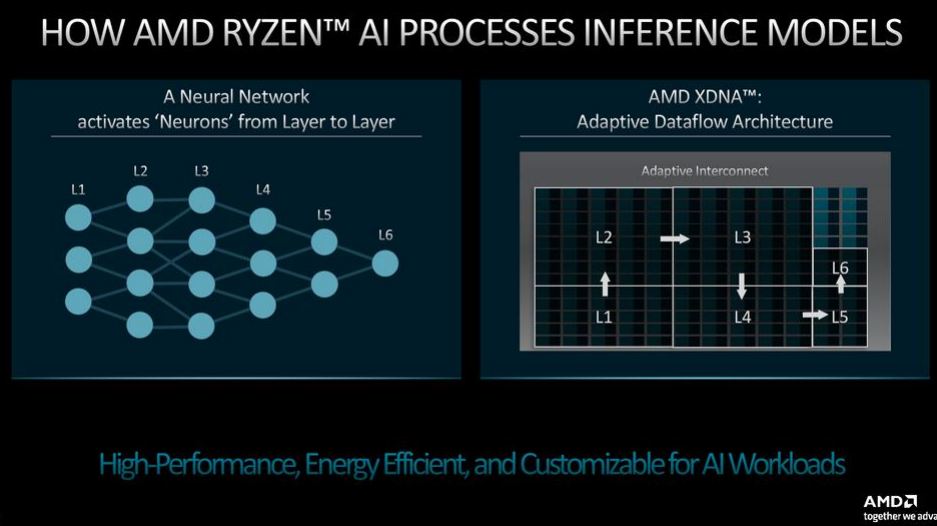
- Advancing AI with High Performance Computing
- Lithography technologies critical to High Performance Computing
- Manufacturing cost challenges
- Summary

AI Wave Will Bring Another Boom to Semiconductors



Generative AI and On-Chip AI Integrated into PC Computing





High-Performance, Energy Efficient, and Customizable for AI Workloads

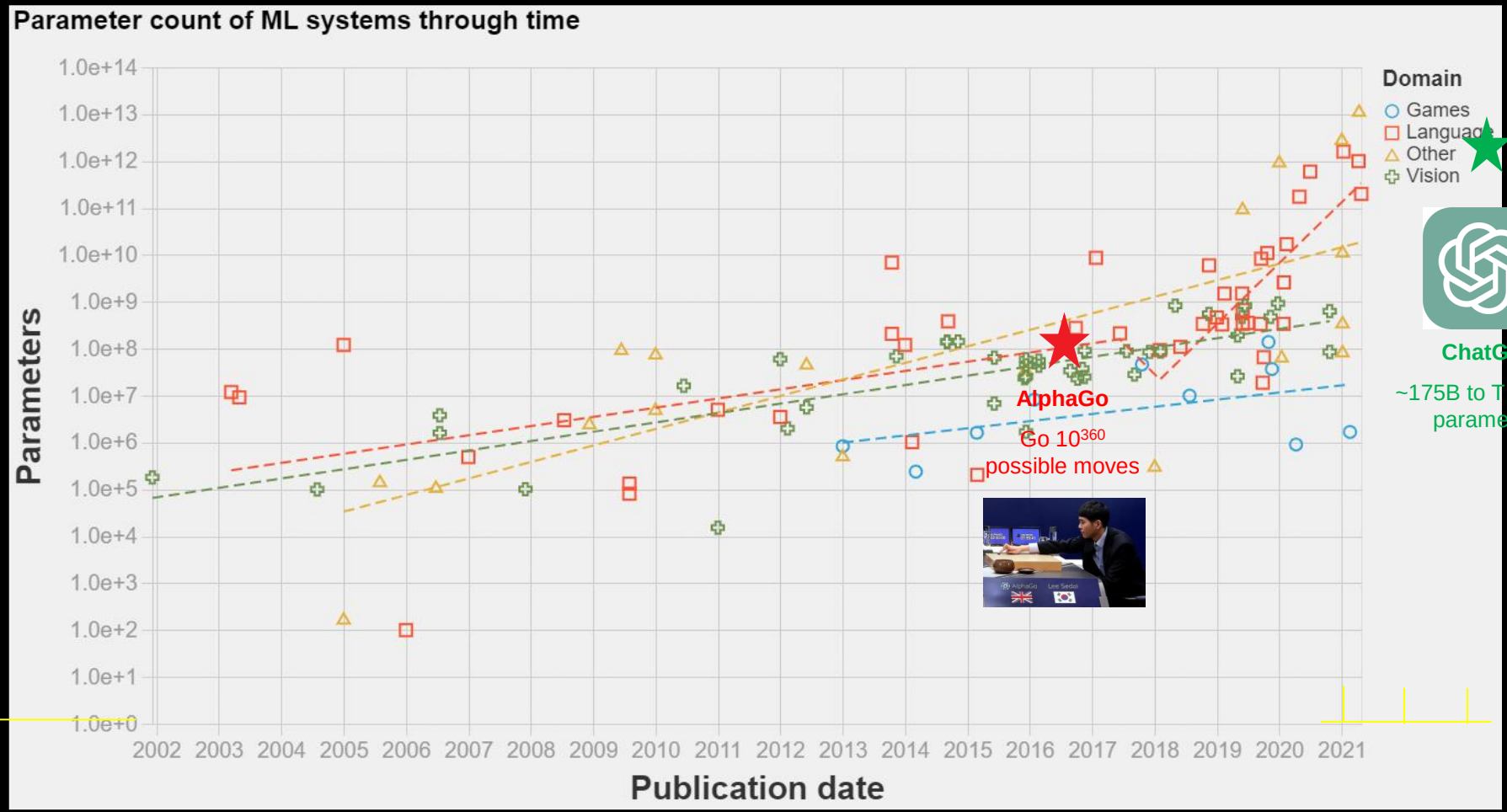
Rapid Pace of AI Development

Source: "AI, Chiplets, and Hetero-Integrations" Photomask Japan '24 Keynote Xin Wu, AMD

Deep Blue

Chess 10^{40} possible moves

1997



Graph: Jaime Sevilla, Towards Data Science, Jul. 2, 2021. Photos from various public domains

Delivering Winning Performance for AI

- **Maximize:**
  - ✓FLOP (need to double rate every 1.3 years)
  - ✓Memory Bandwidth (need to double every 2 years but HBM every 4 years)
  - ✓Networking Bandwidth (need to double every 2 years, Energy/bit constraint)

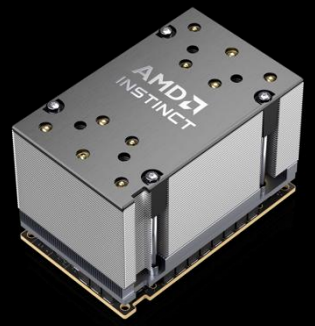
- **Within constraints:**
  - ✓Power (doubling every 2 years) -> primary limiter to future performance gains
  - ✓Silicon + Package Area and Cost (transistor counts not doubling every 2 years -> max reticle size limit)

- **Deliver the software and firmware to easily exploit it.**

AMD
AI Platforms

AI Hardware Portfolio

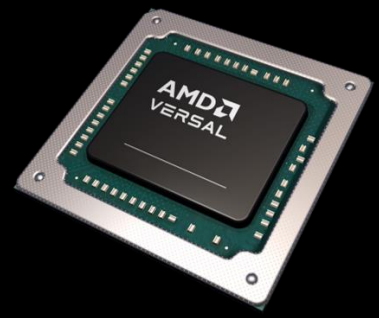
Data center | Edge | End point



AMD Instinct™
GPUs

AMD Alveo™
AI Accelerators

AMD EPYC™
Server Processors



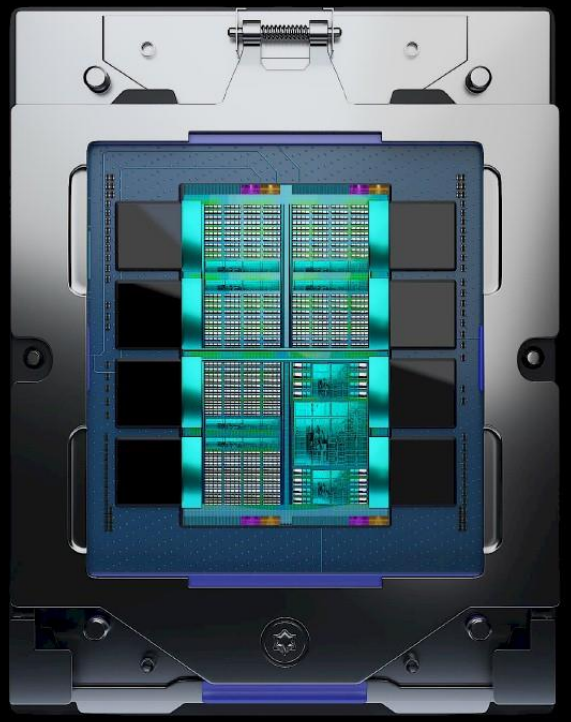
AMD Embedded
Versal™ AI Edge

AMD Ryzen™ AI
Mobile Processors

Architectures

CDNA [+Zen]	*XDNA*	*Zen*	*XDNA*	*Zen +RDNA +XDNA*

State-of-the-Art Technology Advancements Boosting AI

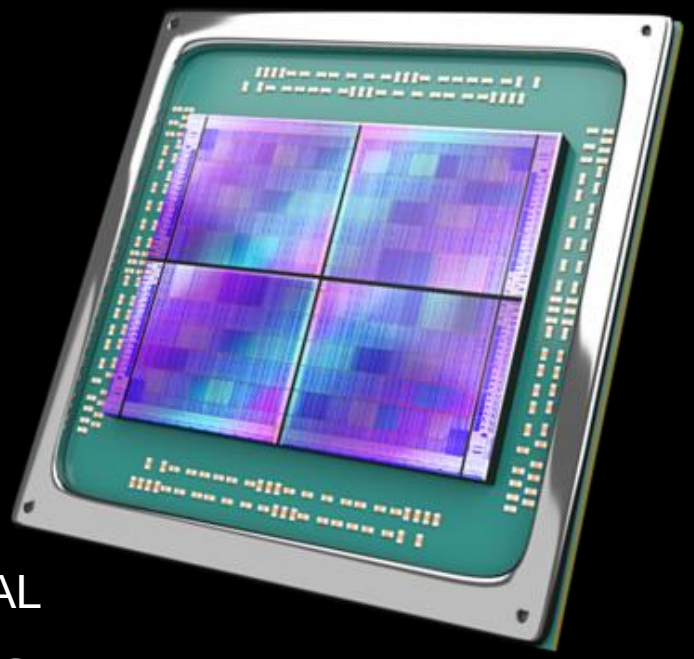


The world's first integrated data center CPU + GPU

AMD INSTINCT™

MI300

Breakthrough architecture to power the exascale AI era



AMD VERSAL

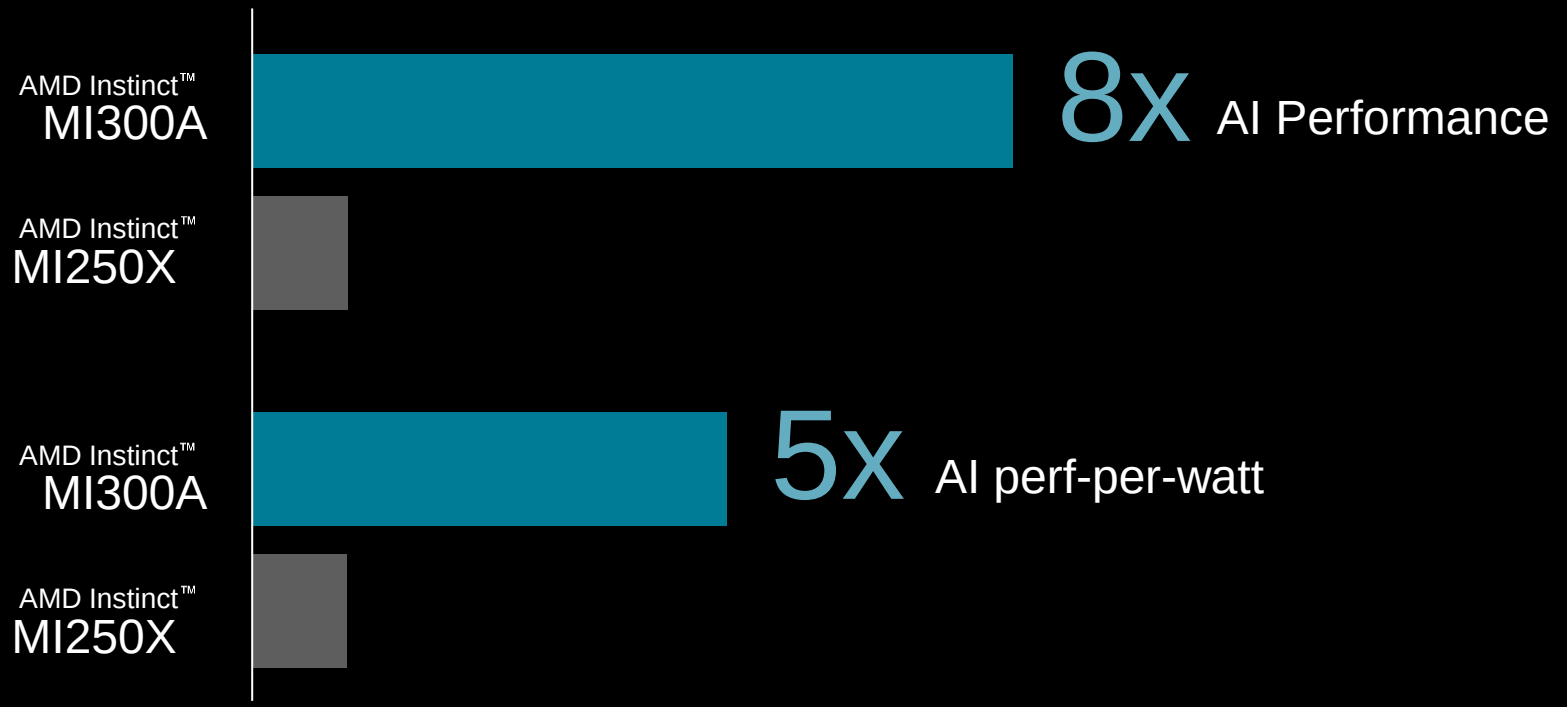
VP1902

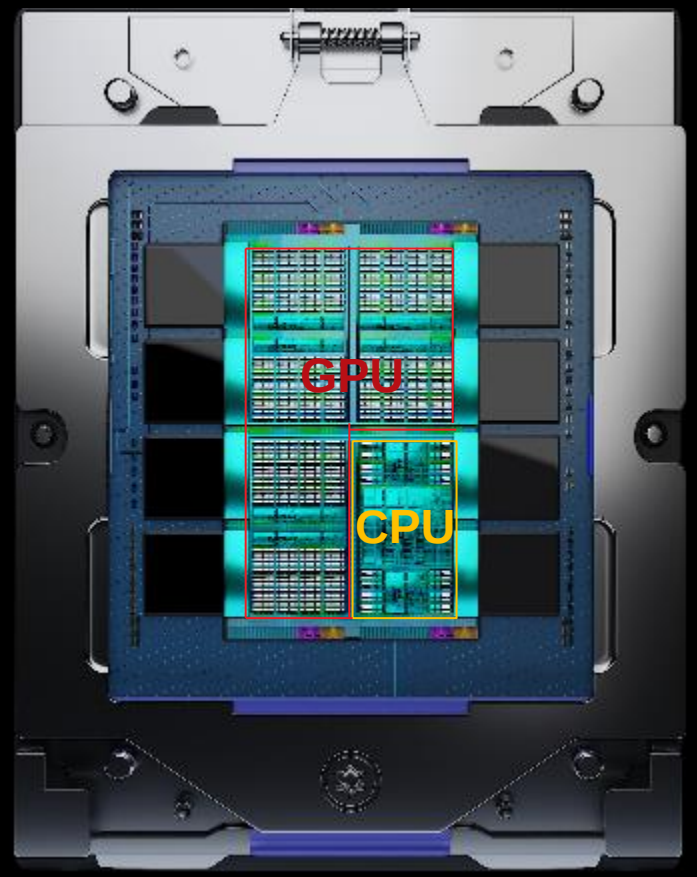
2.4x larger than previous largest FPGA (VU-19P)

18.5M logic cells

MI300A ARCHITECTURAL INNOVATION AT THE NEXT LEVEL

The world's first integrated data center CPU+GPU





Source: "The Next Generation of AI Architecture The Role of Advanced Packaging Technologies in Enabling Heterogeneous Chiplets" IEDM '23 Short Course Raja Swaminathan, AMD

Lithography Technologies to Support High Performance Computing and Advancing AI

- Advancing AI with High Performance Computing
- **Lithography technologies critical to High Performance Computing**
- Manufacturing cost challenges
- Summary

Source: “AI, Chiplets, and Hetero-Integrations” Photomask Japan ’24 Keynote Xin Wu, AMD

CMOS Continual Scaling at Slower Pace, and More Costly

- Standard-cell-based library design kept (almost kept) node scaling
- SRAM and customized designs scaling reduced to approximately half
- Economic benefit of Moore’s Law was reduced significantly (EUV, EUV multi-patterning, GAA, Si-Ge channels, BPDN, 3D CFETs, etc.)
- Cost of design increased from node to node as well

YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
LOGIC TECHNOLOGY ANCHORS						
Device technology inflection	Taller fin	LGAA	CFET-SRAM	Low-Temp Device	Low-Temp Device	Low-Temp Device
Patterning technology inflection for Mx interconnect	193i, EUV DP	193i, EUV DP	193i, High-NA EUV	193i, High-NA EUV	193i, High-NA EUV	193i, High-NA EUV
Beyond-CMOS as complimentary to platform CMOS	-	-	2D Device, FeFET	2D Device, FeFET	2D Device, FeFET	2D Device, FeFET
Channel material technology inflection	SiGe50%	SiGe60%	SiGe70%	SiGe70%, Ge	2D Mat	2D Mat
Local interconnect inflection	Self-Aligned Vias	Backside Rail	Backside Rail	Tier-to-tier Via	Tier-to-tier Via	Tier-to-tier Via
Process technology inflection	Channel, RMG	Lateral/AtomicEtch	P-over-N, N-over-P	3DVLSI	3DVLSI	3DVLSI
Stacking generation inflection	3D-stacking, Mem-on-Logic	3D-stacking, Mem-on-Logic	3D-stacking, CFET, Mem-on-Logic	3D-stacking, CFET, 3DVLSI	3D-stacking, CFET, 3DVLSI	3D-stacking, CFET, 3DVLSI

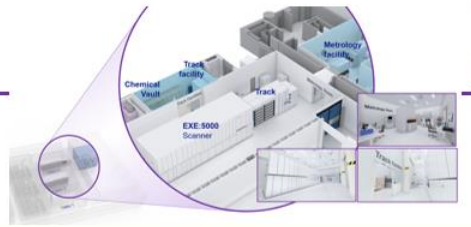
Source: International Roadmap of Devices and Systems, 2022, More Moore

EUV 0.33NA and 0.55NA Roadmap

High NA (0.55NA) is an evolutionary step in EUV technology

Source: Kaustuve B, et al, ASML, SPIE Bacus 2022

Wavelength	NA, Half pitch	2020	2021	2022	2023	2024	≥2025
EUV	0.33 NA, 13 nm	NXE:3400C 1.5 nm | 135 wph[2] / 145wph[3]		NXE:3600D 1.1 nm | 160wph		NXE:3800E <1.1 nm | >195wph / 220wph[3]	NXE:4000F <0.8 nm | >220wph
	Customer timing 0.55NA					Early Access ASML	Customer R&D / Customer HVM
	0.55NA, 8 nm					EXE:5000 <1.1 nm | 150wph[1]	EXE:5200B <0.8 nm | 220wph



wafers/hours (wph) are based on 30mJ/cm²
1) 185wph@20mJ/cm2
2) 170wph@20mJ/cm²
3) Throughput upgrade

Product: Matched Machine Overlay (nm)|Throughput(wph)

Product status Released | Development | Definition

High NA (0.55NA) added to the existing 0.33NA EUV portfolio to support HVM in 2025 - 2026 timeframe

ASML

Jan van Schoot et al, MNC Conference, Tokushima, November 2022

EUVL Adoption Enabled 7nm Successful Production Ramping

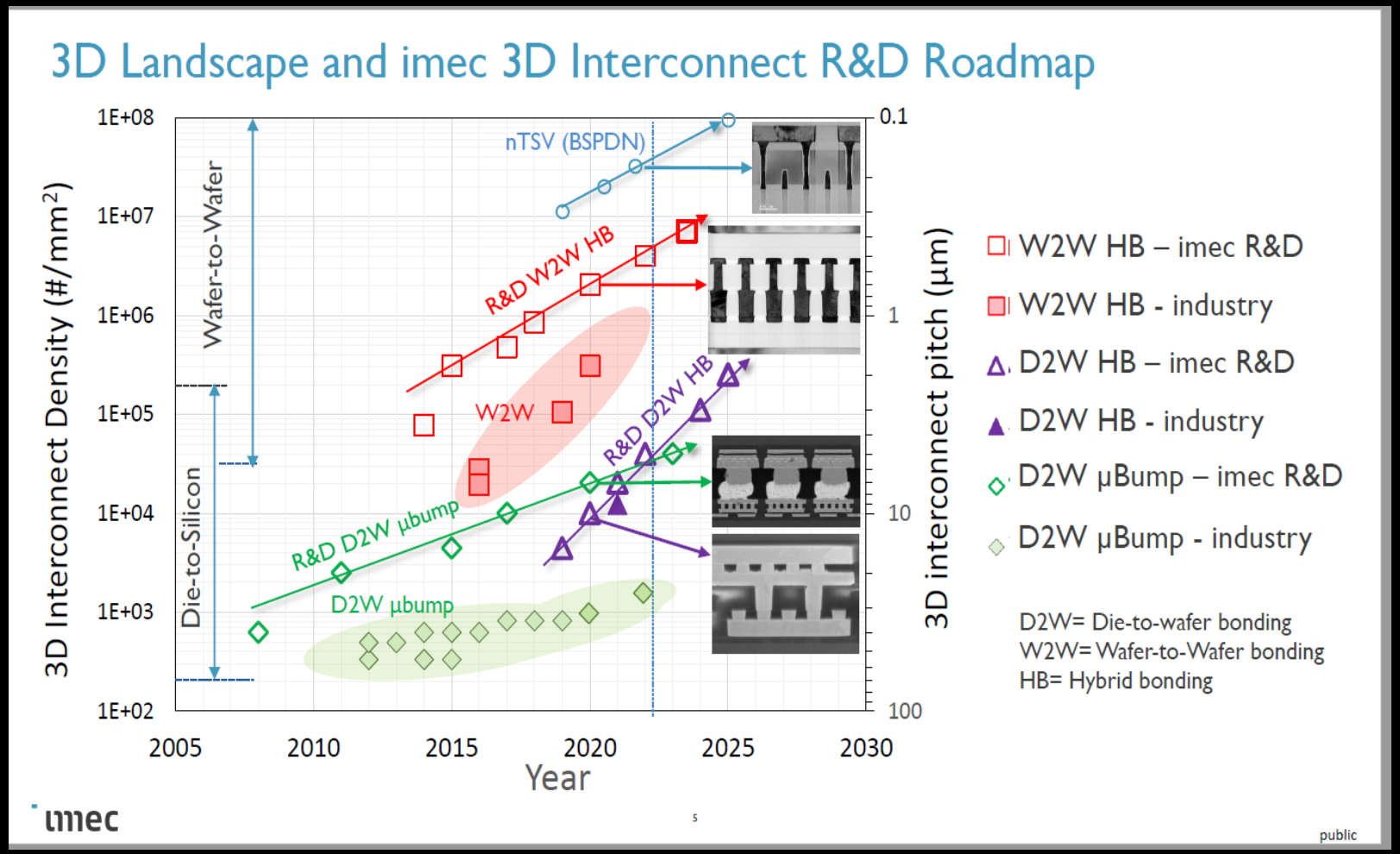
- Performance improvement (Contact & Metal Sheet Resistance Reduction)
- Yield Improvement (Defect Density)
- Wafer Cycle-time Reduction

GPU Accelerated Computational Lithography

- ML (Machine Learning) based DFM (“Feature-based” lithography hot spot simulation and the auto-fix (prior to tape-out)
- OPC modeling and the implementation (pixel-based OPC w/curvilinear shapes)
- Post-OPC Rule Checks (pixel-based OPC w/curvilinear shapes)
- Mask inspection (scanned image computation)

Source: "AI, Chiplets, and Hetero-Integrations" Photomask Japan '24 Keynote Xin Wu, AMD

Hetero-Integration (HI) Provides New Opportunities



- HI is not another Moore's Law – they fit different applications differently and come with a cost
- Successful HI products require multi-discipline development

Source: IMEC public (https://www.imec-int.com/en/articles/view-3d-technology-landscape)

AMD

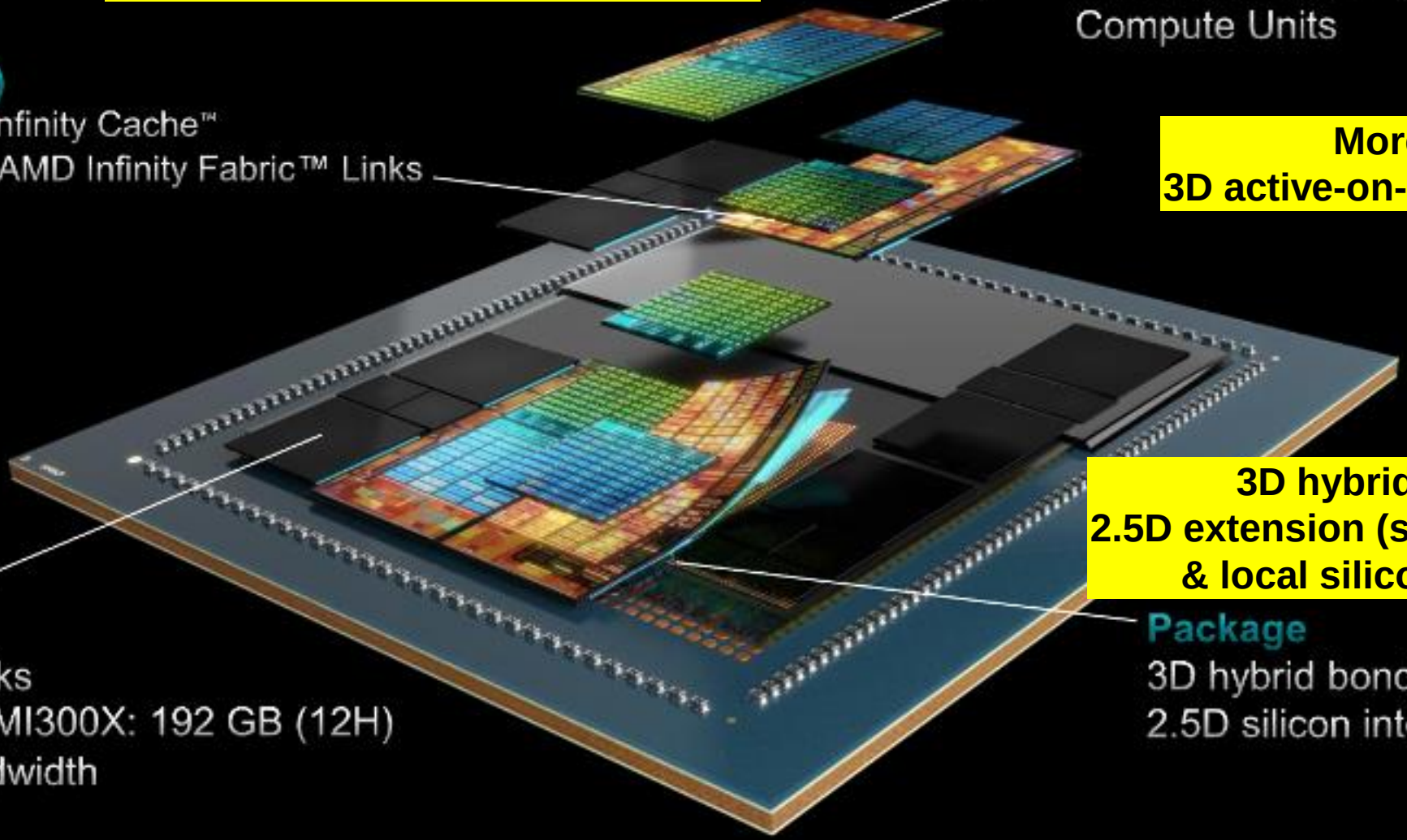
AMD Instinct™ MI300X Accelerator

PPA improvement
by continuous dimensional scaling

Accelerator Complex Die (XCD)
304 AMD CDNA™ 3 Architecture
Compute Units

I/O Die (IOD)
256MB AMD Infinity Cache™
7 x16 4th Gen AMD Infinity Fabric™ Links

More # of
3D active-on-active stacking



3D hybrid bonded
2.5D extension (silicon interposer
& local silicon bridging)

HBM3
8 physical stacks
AMD Instinct™ MI300X: 192 GB (12H)
~5.3 TB/s Bandwidth

Package
3D hybrid bonded
2.5D silicon interposer

AMD

Key Challenges of Hetero Integration (HI) Product

- 3D stack technology itself
- Architecture and applications that benefited from HI connections and can afford its cost
- HI design, CAD, simulation flow and PDKs (including 3D modeling)
- Software
- Pkg (I/Os) routability, signal and power integrity (SIPI), and power delivery
- Thermal solutions
- Multi-chip production flow, testing, supply-chain and logistics
- Reliability
- And many more

Package Scaling being Enabled!

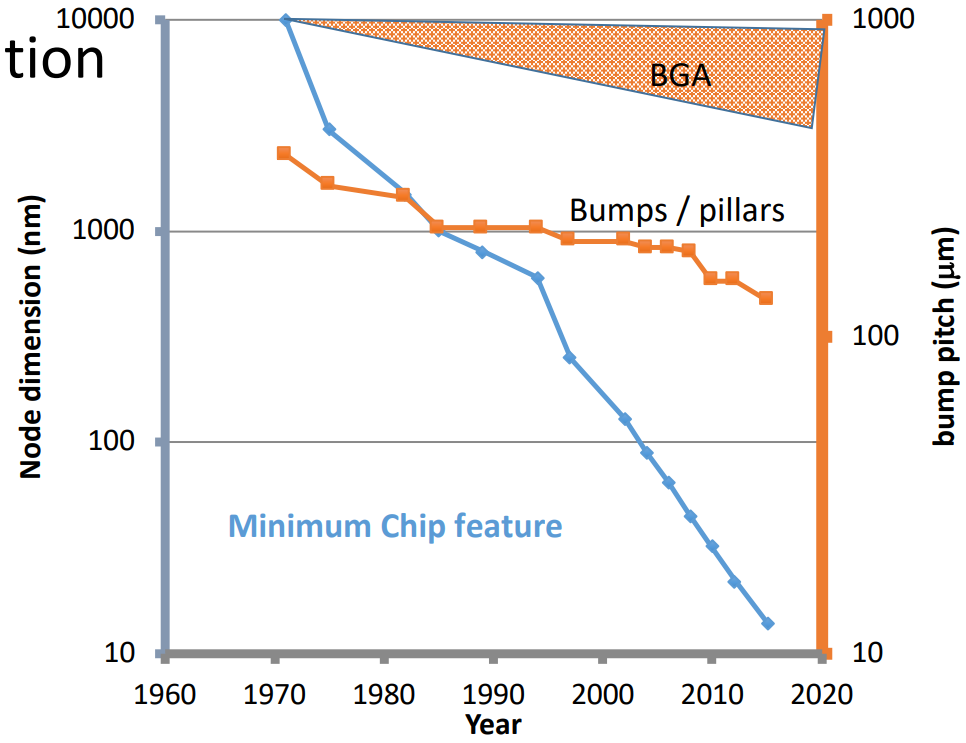
Source: "Advanced Packaging Tutorial" IEDM '2018 S.S. Iyer

What has limited package scaling (vs Si) ?

- Motivation - value realization
- Investment
- Material Limitation
- Processing techniques
- Equipment

For most part packaging has been an afterthought

But all that is changing now



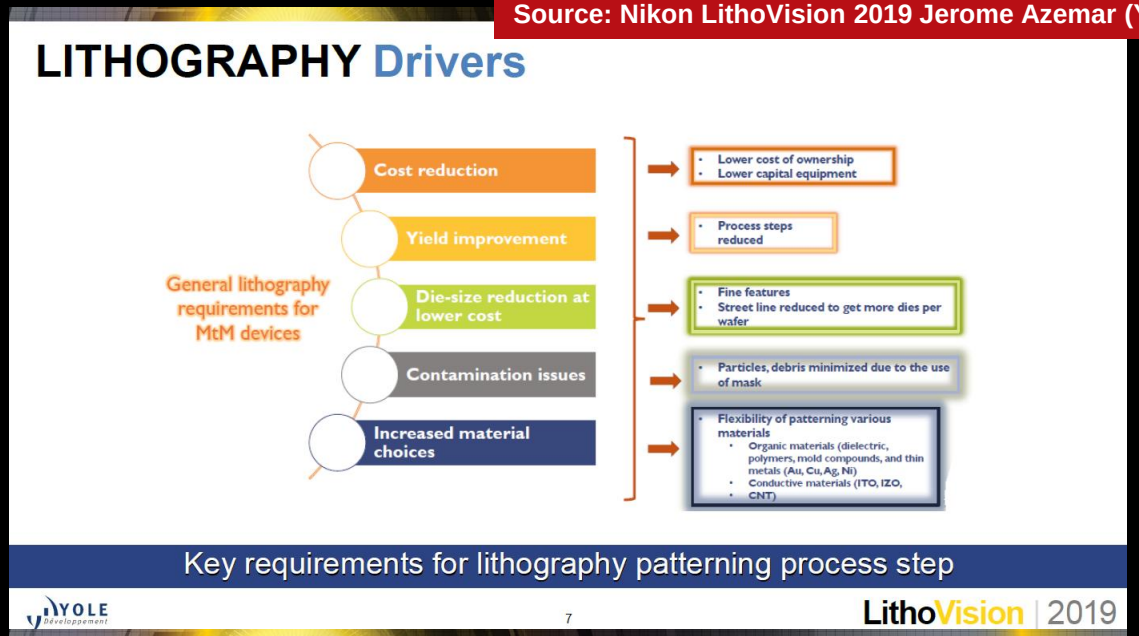
Iyer MRS Bull (2015)

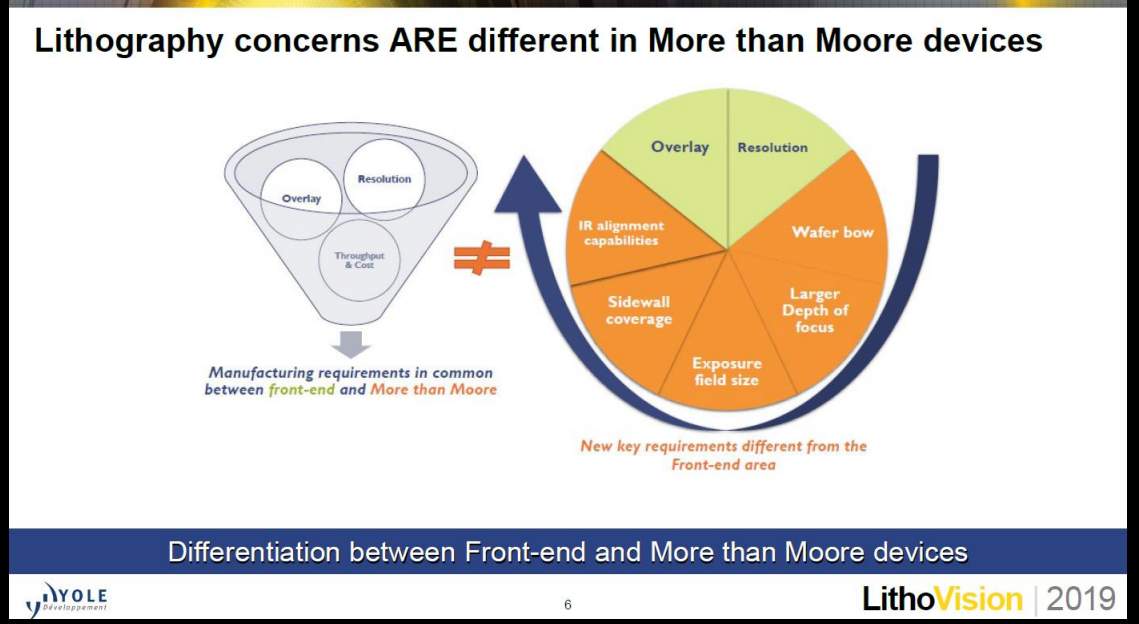
Lithography Requirements for Advanced Packaging

- Throughput
- Patterning Cost
- Process Control
- Resolution
- CD Control and Overlay
- **Exposure Field size**
- **Thick Photo Resist**
- **Warpage control**
- **Depth of Focus Enhancement**
- **Backside Alignment (IR)**

Need to prioritize the requirements for each package technology (cannot satisfy all)

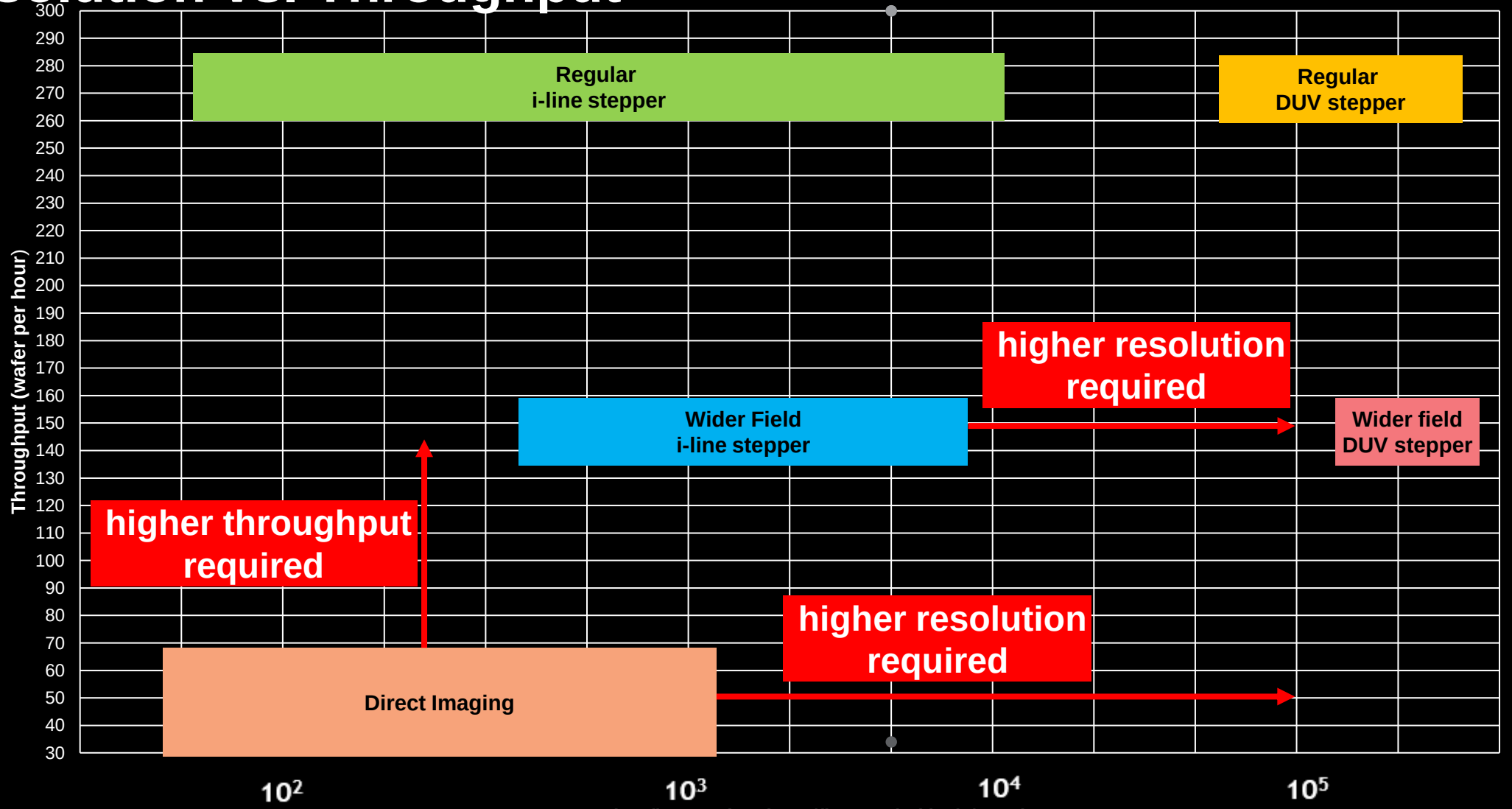
Source: Nikon LithoVision 2019 Jerome Azemar (Yole)

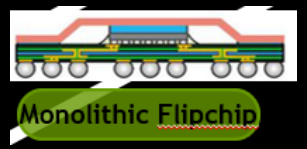


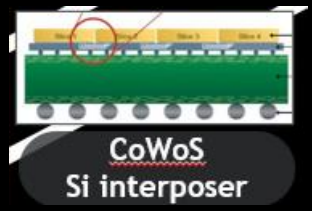


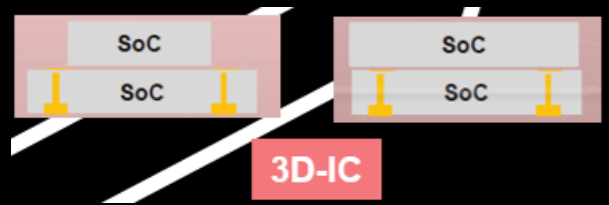
20

Resolution vs. Throughput



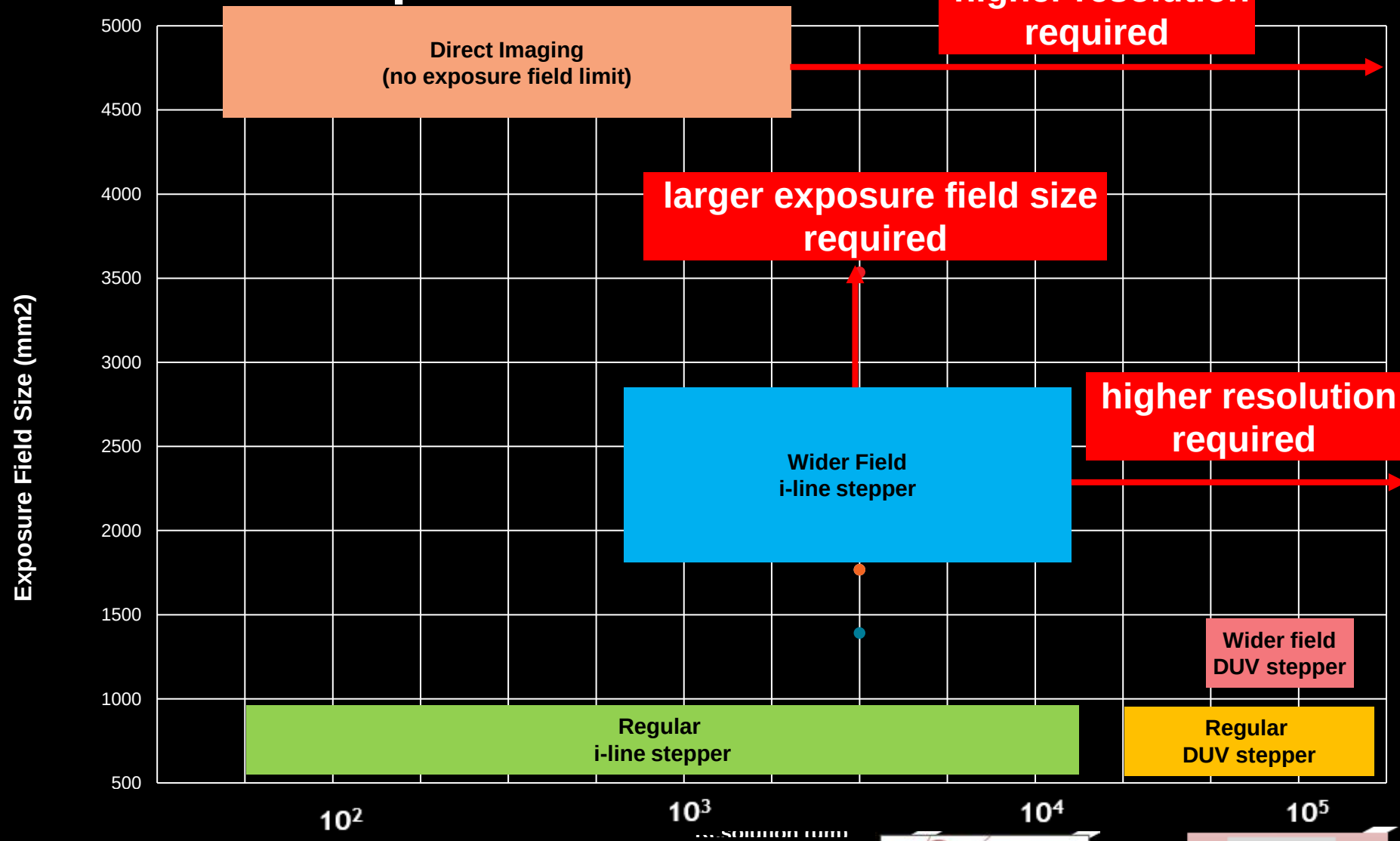


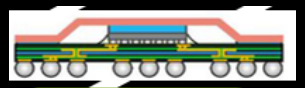


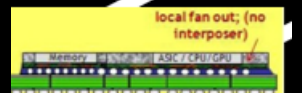


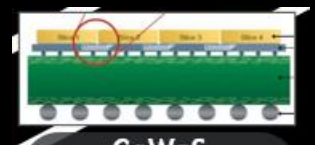
21

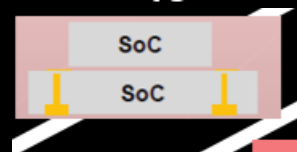
Resolution vs. Exposure Field Size

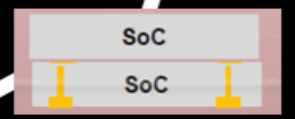












22

Packaging Evolution by Lithography Enhancement

1) **Patterning tool with wider exposure field size**
    1) Reduction or elimination of amount of stitching
    2) Extendibility even to Panel-level package
2) **Higher Resolution Stepper for finer pitch bonding (u-bump -> hybrid bond)**
3) **Lithography process optimization for heterogeneous integration**
4) **Warpage management and OVL control (pre-bond/post-bond wafer measurement and feed-forward/feedback to bonding process including correction during the litho process)**
5) **Backside Alignment Methodology**

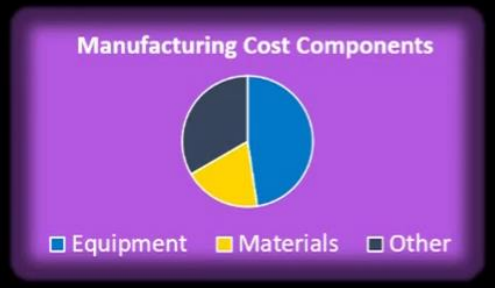
23

Lithography Technologies to Support High Performance Computing and Advancing AI

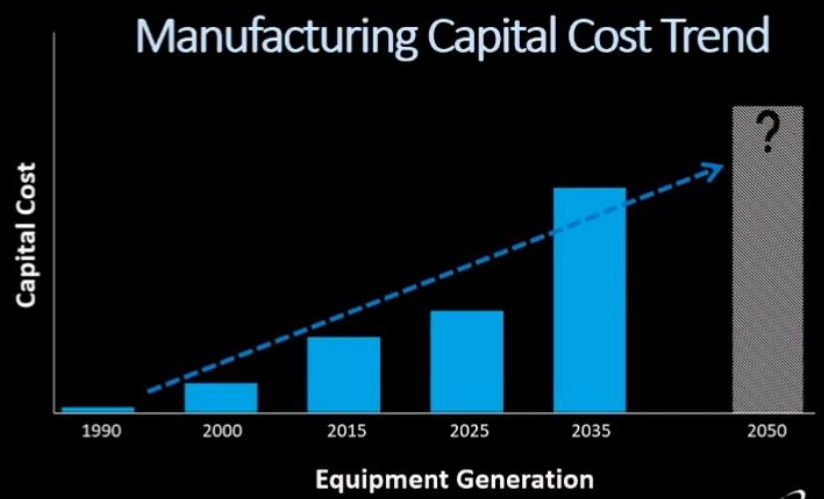
- Advancing AI with High Performance Computing
- Lithography technologies critical to High Performance Computing
- **Manufacturing cost challenges**
- Summary

Source: “Making Memory Magic and the Economics Beyond Moore’s Law” IEDM ’23 Plenary #2 Thy Tran/Micron

Manufacturing cost a major factor of “PPAC”



Capital cost trend increasing (also install cost), materials, spares, are top contributors to total manufacturing cosgts



Micron

EUVL Challenges for HVM

- **Throughput**
  - Scanner source power
  - Resist sensitivity
  - Light transmission through pellicle
- **Scanner capacity/availability**
  - Timing to get sufficient numbers of EUV scanners
  - Tin droplet/collector mirror life time and the replacement TAT
- **Stochastic defect**
  - Understanding of physical/chemical reaction mechanism (how many photons absorbed in the resist, how many electrons created?)
  - Metrology (optical, e-beam, X-ray?)
- **Mask/wafer inspection**
  - Actinic Patterned Mask Inspection productivity (inspection time & inspection capacity)
  - “through-pellicle” inspection
  - Periodic mask inspection@wafer fab
  - Mask maintenance@wafer fab (mask re-Qual)
- **Overall Cost and Cycle-time**

Higher-NA EUVL Cost Benefit?

Source: "Analysis of advanced technology nodes and h-NA EUV introduction: a cost perspective" 2021 SPIE EUVL, imec

h-NA introduction: assumptions and case studies

Layer	Pitch (nm)	EUV reference – 2 nm	h-NA Limited adoption – 2+ nm	h-NA - Full usage – 2++ nm
M0A	42	EUV LE + Cut	hNA-EUV LE	hNA-EUV LE
Mint	18	EUV SALELE + EUV SAB2	EUV SALELE + EUV SAB2	
M1	28	EUV SALELE + 193i SAB2	EUV SALELE + 193i SAB2	
M2	18	EUV SALELE + EUV SAB2	EUV SALELE + EUV SAB2	
V0 / Vint_A / V1	>30	EUV LELE	hNA-EUV LE	

Lithography tool	Tool cost	Throughput @ 30mJ/cm2	Throughput @ 50mJ/cm2
EUV (Reference)	1.0X	220	160
h-NA	1.5X	198	146

- 2 nm technology node will be used as candidate for benchmarking h-NA EUV lithography
- 10% lower throughput because of anamorphic imaging in h-NA EUV lithography*
- Tool cost considered 1.5 times higher with respect of the EUV reference

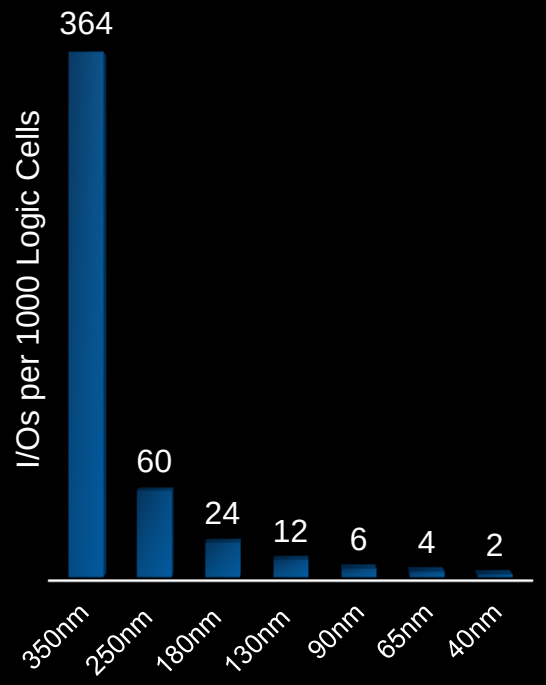
*This is just an initial assumption and the actual difference in throughout will depend on the actual die size and application

- **Scanner throughput just 10% lower even with the reduced exposure field size?**
- **New material cost (new resist, new mask absorber, new pellicle)**

Yield Risk Becoming Greater for High Density Monolithic Devices

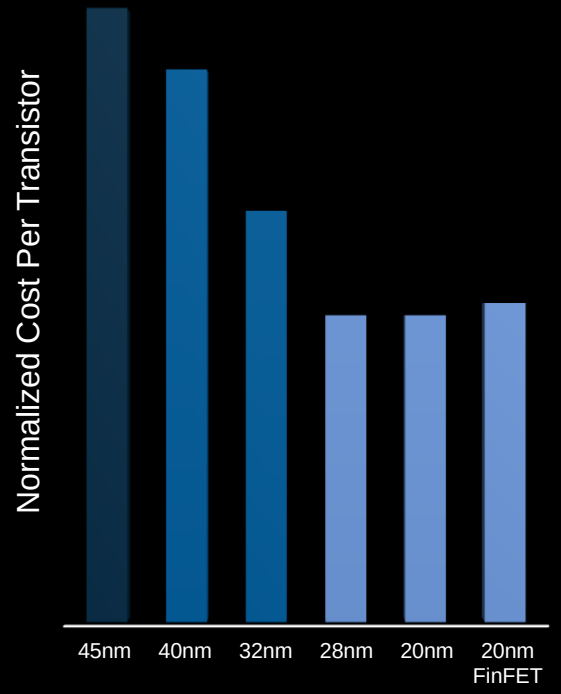
Source: Henley Liu, Semicon Taiwan 2018

On-Die I/Os Not Scaling



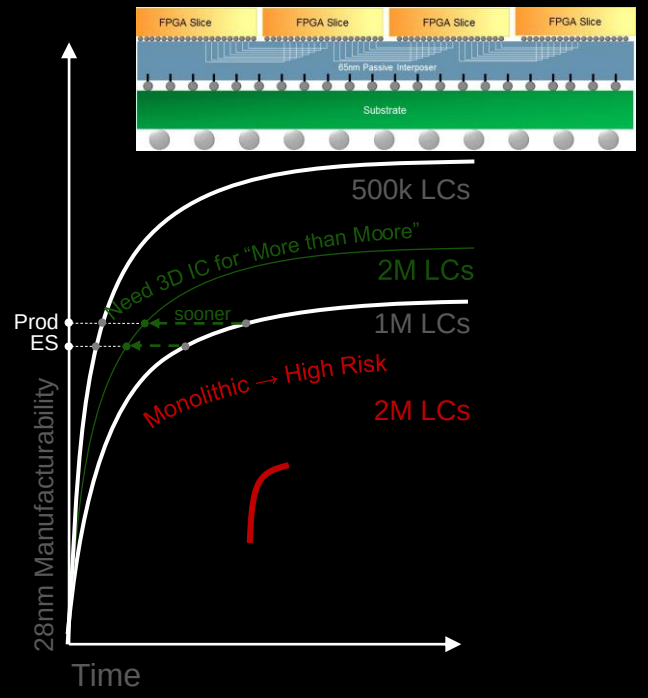
(Source: Xilinx)

Cost per Transistor Not Scaling



(Source: Nvidia at and ISSCC 2013)

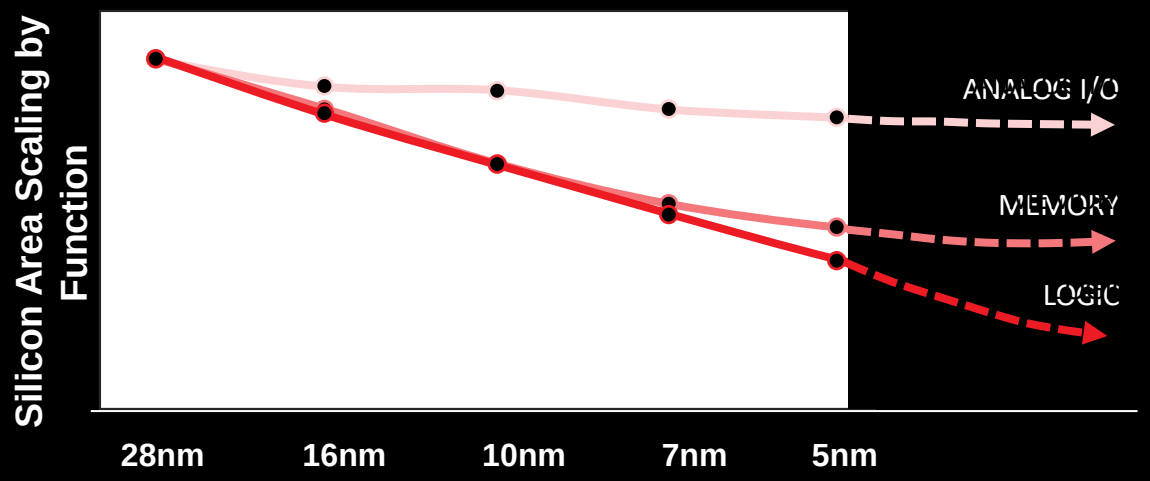
Yield Risk Becoming Greater for High Density Monolithic FPGAs



We Set Out to Deliver Breakthrough Integration, and Do It with the Highest Yield & Reliability

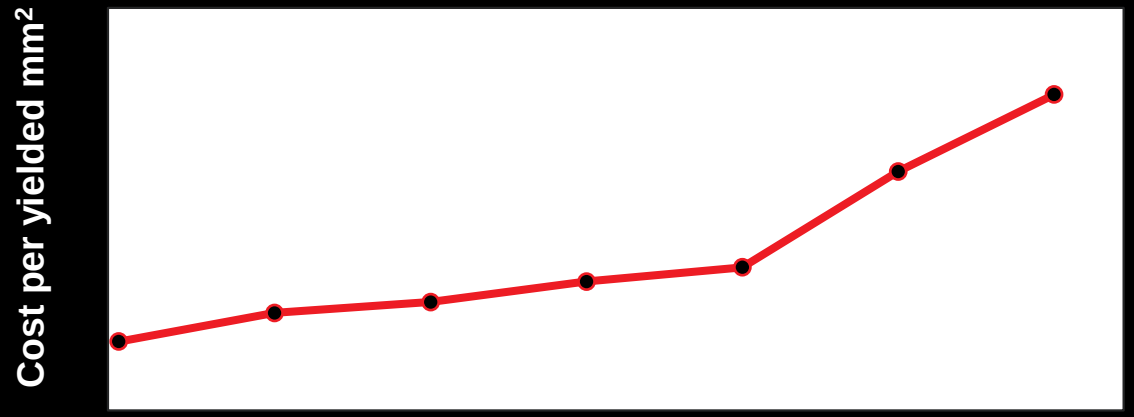
CHIPLET TECHNOLOGY
Our Motivation

Limited and Divergent Scale Factors



Density gains diminishing

Increasing Costs[1]



Costs Increasing

Drivers for chiplet architecture:

- Yield/Cost – Building monolithic SOCs on leading process technology is not cost effective
- Physical reticle size limitations – Data Center chips just do not fit with reticle limits
- Modularity – Chiplets enable design reuse – monolithic SOCs require custom design for every product

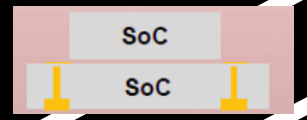
Source: "The Next Generation of AI Architecture The Role of Advanced Packaging Technologies in Enabling Heterogeneous Chiplets" IEDM '23 Short Course Raja Swaminathan, AMD

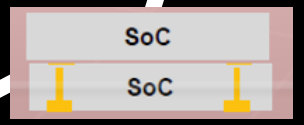
Adaptive Package of Choices (Chiplets & Heterogeneous Integration)

Consumer

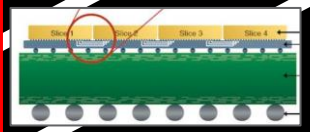
Video & Broadcast

Automotive



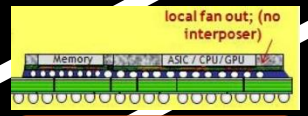


3D-IC



CoWoS
Si interposer

Emulation



Wafer Level Fan-out

Monolithic Flipchip

HPC
Data Center

Aerospace &
Defense

Connection (bump) density:
1/(bump pitch)^2 (1/mm^2)

10^5

10^4

10^3

10^2

0.1

1.0

10

10^2

Routing density: 1/(line pitch) (1/um)

AMD

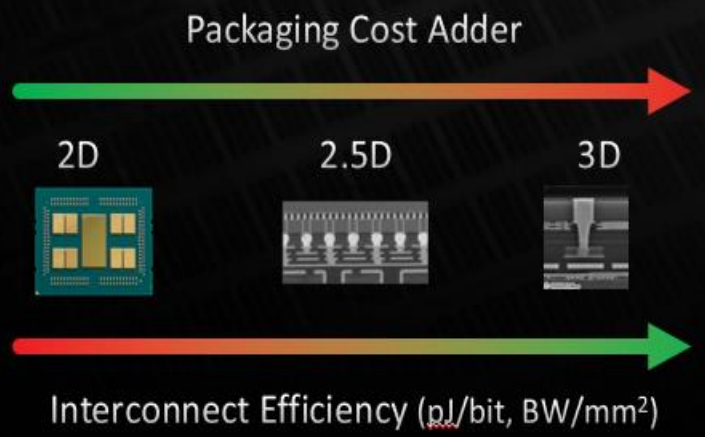
Source: "The Next Generation of AI Architecture The Role of Advanced Packaging Technologies in Enabling Heterogeneous Chiplets" IEDM '23 Short Course Raja Swaminathan, AMD

HOW DO WE SELECT PACKAGING TECHNOLOGY?

PPAC Optimization:

- Chiplet architecture should be cost effective to compared to monolithic SOC ($Perf/$, $Perf/Watt)

Chiplet package architecture selection requires balancing a complex equation…



[Value of modular product]
+
[Chiplet yield and technology cost benefit]

>

[Chiplet die overhead adder]
+
[Packaging cost adder]
+
[Power cost of interconnect]
+
[Engineering complexity of solution]

Architectural need for bandwidth, die partition options and package technology create a multi-disciplinary optimization equation

AMD

Lithography Technologies to Support High Performance Computing and Advancing AI

- Advancing AI with High Performance Computing
- Lithography technologies critical to High Performance Computing
- Manufacturing cost challenges
- **Summary**

Summary

- AI wave is bringing another boom to semiconductors.
- State-of-the-Art Technology Advancements Boosting AI.
  - AMD Instinct – MI300X accelerator - The world's first integrated data center CPU+GPU
- Monolithic CMOS scaling to continue but slowing down -> need new innovation
  - EUV enables continuous chip-level scaling. (resolution vs. cost/throughput challenge)
  - GPU accelerated computational lithography
- Chiplets + SoIC enables system-level scaling.
- Lithography technology evolution critical to advanced packaging technologies.
- Cost challenges
  - EUV productivity (Higher NA insertion timing)
  - Advanced packaging cost adder vs. performance & power gain

Disclaimer and Attribution

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION


AMD