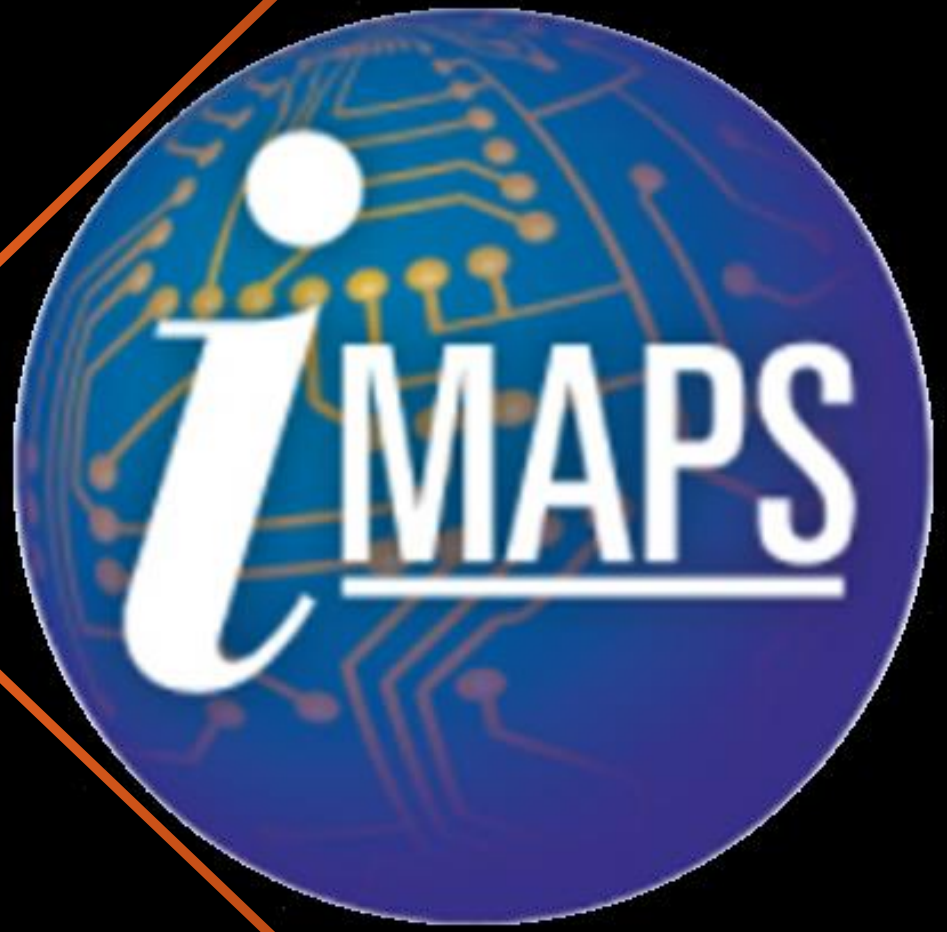
AMD

Enabling the Transition to Heterogeneous Integration

Bryan Black

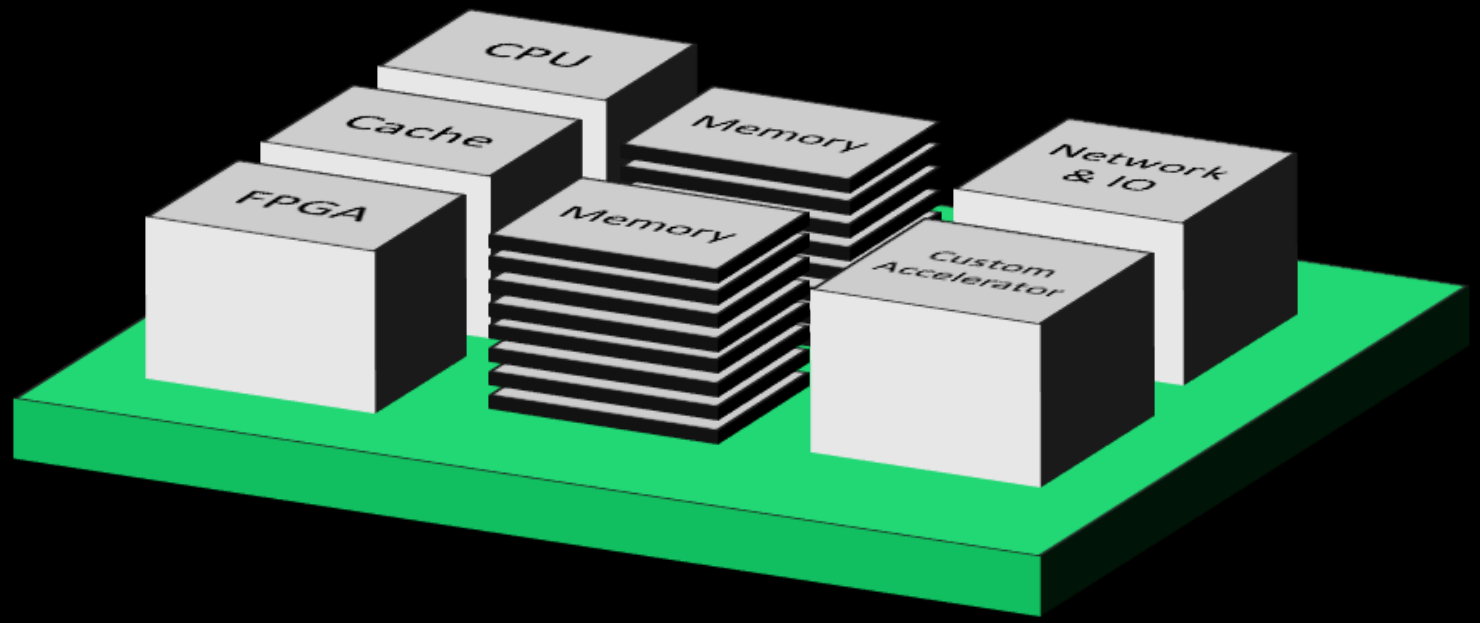


Cautionary Statement

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) including, but not limited to the features, functionality, availability, timing, expectations and expected benefits of AMD's products and AMD's product roadmaps, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's Quarterly Report on Form 10-K for the year ending on December 26, 2020.

Heterogeneous Integration

- Integrate many disparate dies from many different vendors into a single substrate
- New packaging technologies and on-die IP innovations are required
- Endless configurations mean endless complexity but needs to be straight forward for the design house



AMD

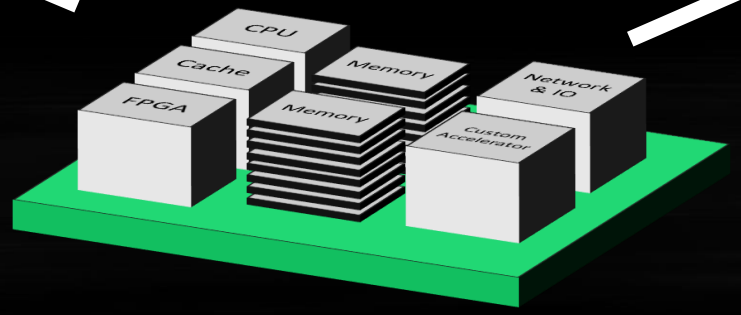
Why?

Silicon Devices

Systems

Converging into Heterogeneous Integration

Cost, Reuse, and Flexibility managing Moore's Law

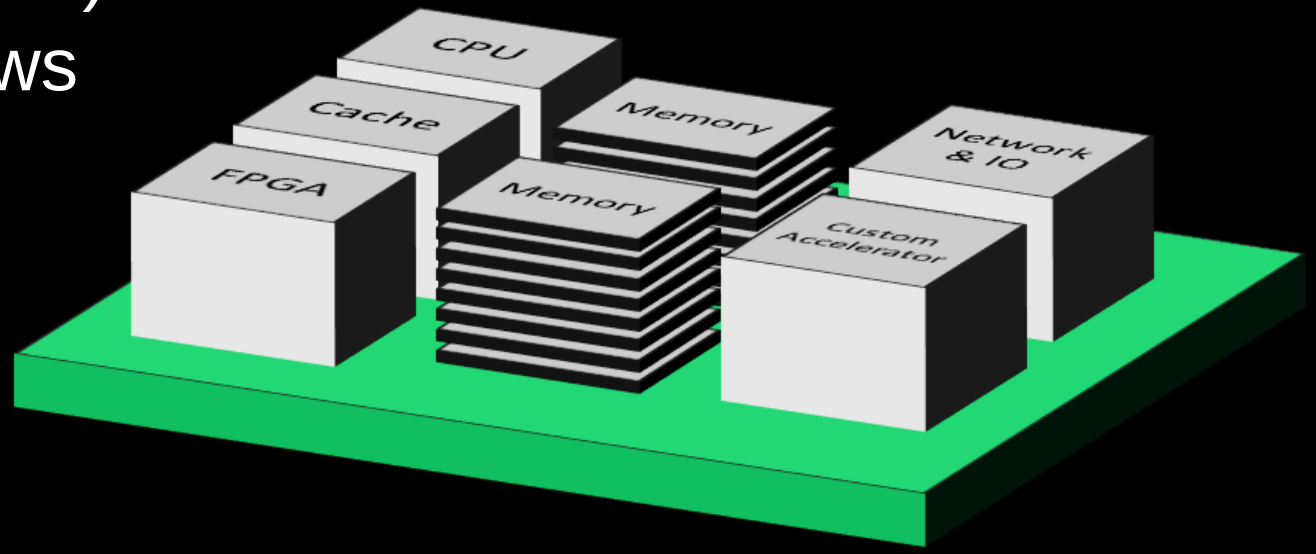


Ownership, Customization, and Differentiation

AMD

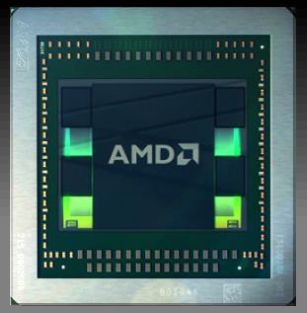
Heterogeneous Integration

- Two major challenges
  - Die source (internal, external)
  - Manufacturing and test flows



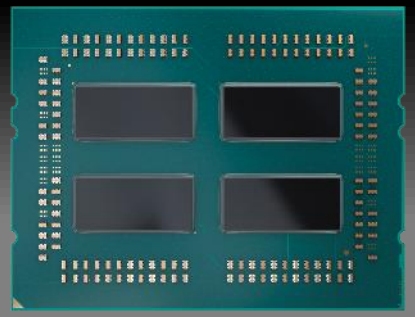
Cannot be achieved in a single step or with a single product

AMD Leadership Packaging



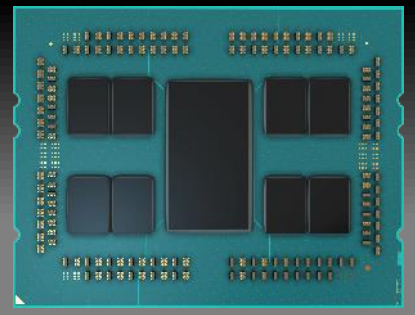
2.5D HBM

2015



MULTICHIP MODULE

2017

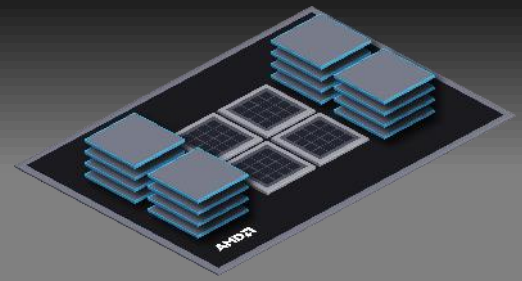


CHIPLETS

2019

>10x

BANDWIDTH DENSITY



3D CHIP DESIGN PACKAGING

(Hybrid 2.5 & 3D)

Future

Led Industry in 2.5D & Chiplet Architecture

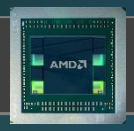
Aggressive Roadmap for Chiplet & 3D Integration

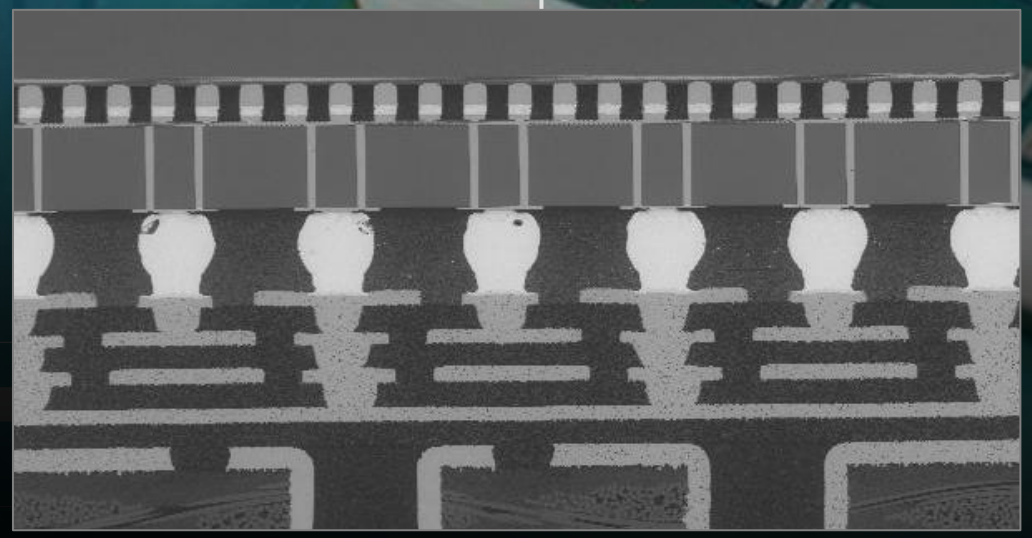
Future Bandwidth Number AMD Estimate

AMD

High-End Radeon™ Graphics with HBM in 2015

“Fiji” Chip

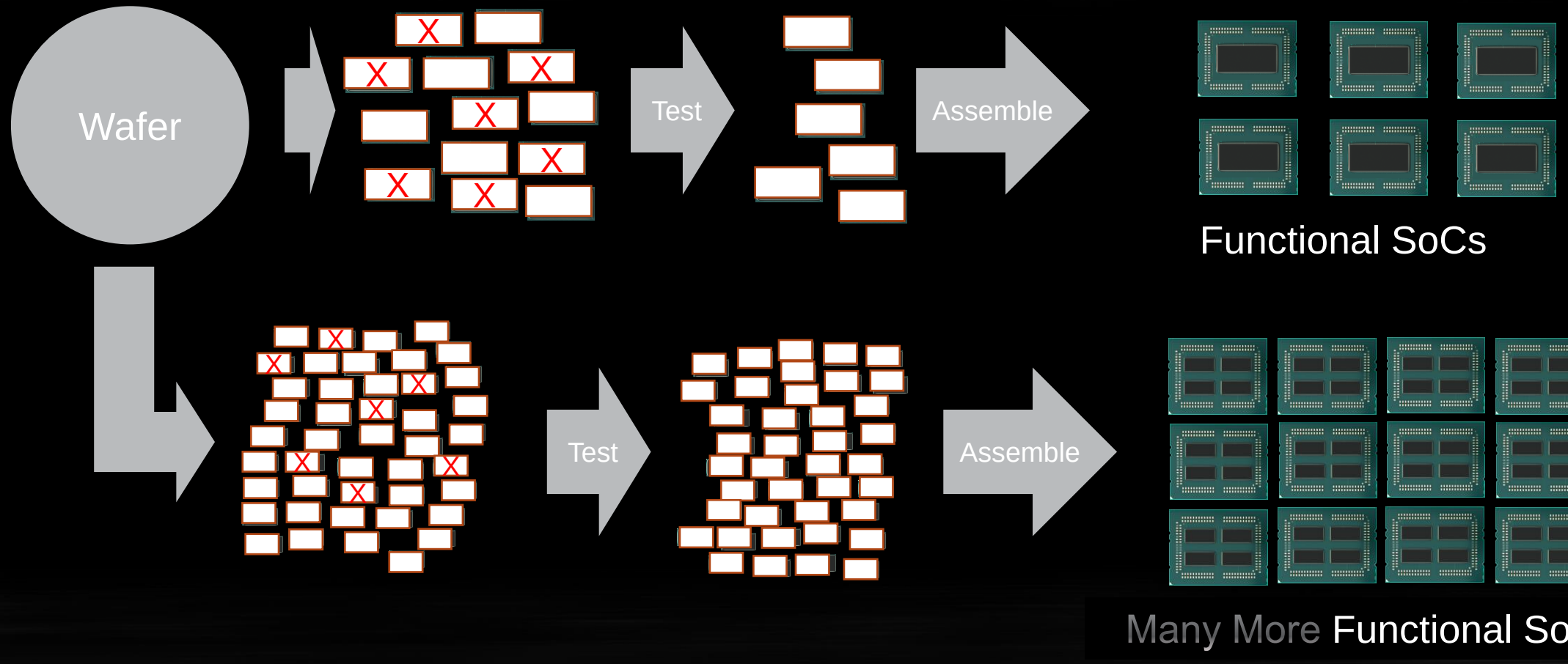




- 4GB High-Bandwidth Memory
- 4096-bit wide interface
- 512 GB/s Memory Bandwidth

- Graphics Core Next Architecture
- 64 Compute Units
- 4096 Stream Processors
- 596 sq. mm. Engine

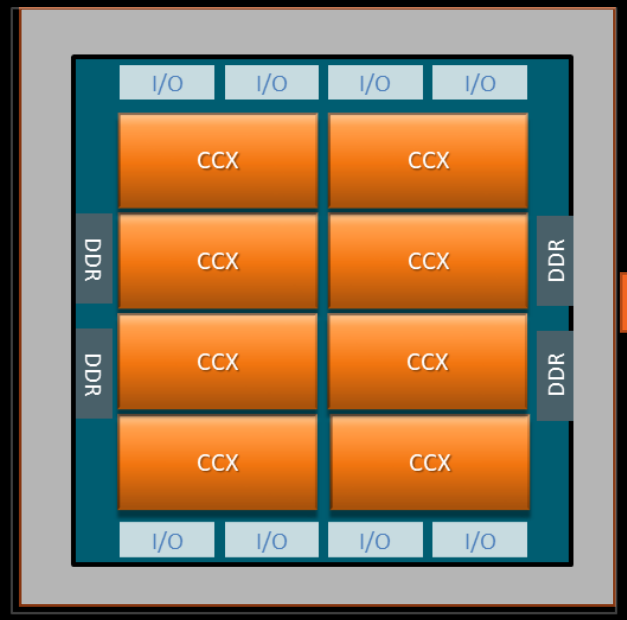
High-Level Approach to Chiplets



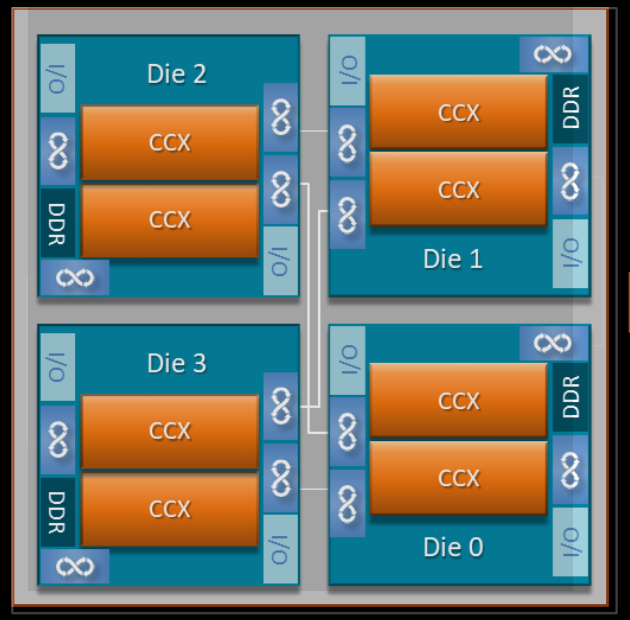
AMD

Chiplets Evolved – Hybrid Multi-die Architecture

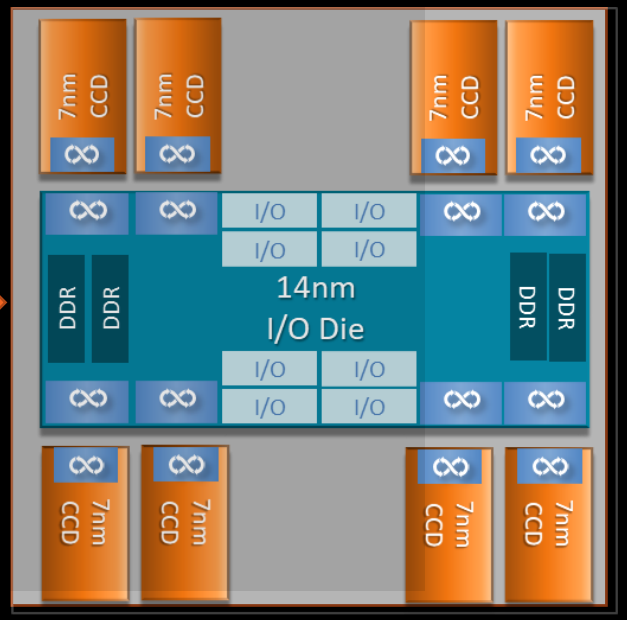
Traditional Monolithic



1st Gen EPYC™ CPU



2nd Gen EPYC™ CPU



Use an Advanced Technology Where it is Needed Most

Each IP in its Optimal Technology, 2nd Gen Infinity Fabric™ Connected

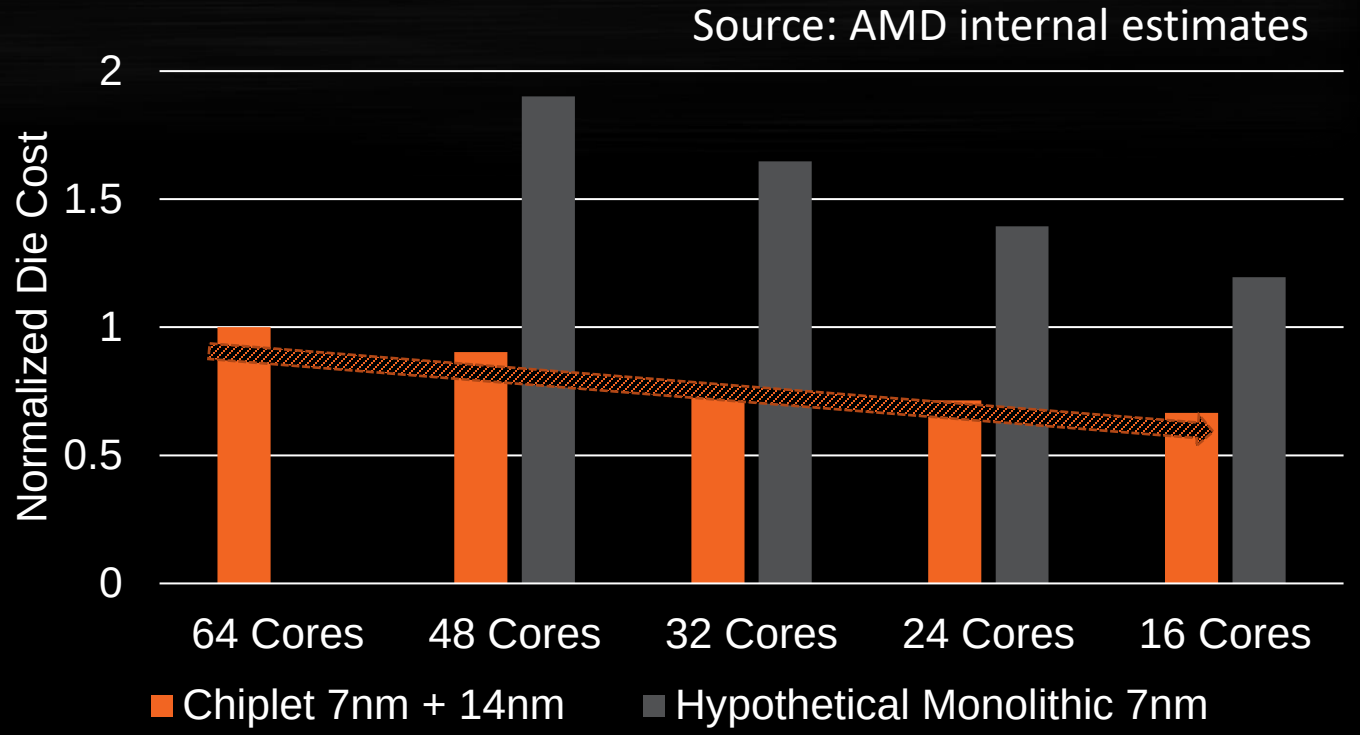
Centralized I/O Die Improves NUMA

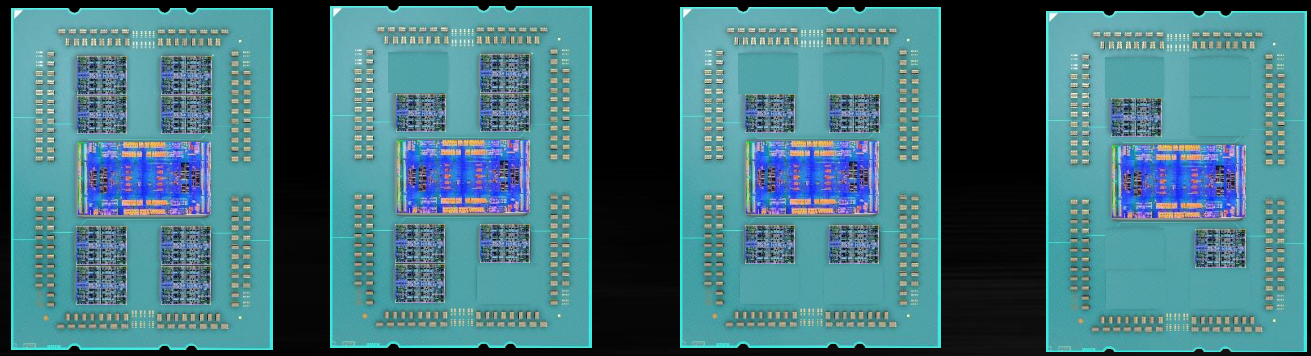
Superior Technology for CPU Performance and Power

AMD

2nd Gen AMD EPYC™ Chiplet Performance vs. Cost

- Higher core counts and performance than possible with a monolithic design
- Lower costs at core count/performance points in product line
- Cost scales down with performance by depopulating chiplets
- 14nm technology for IOD reduces fixed cost

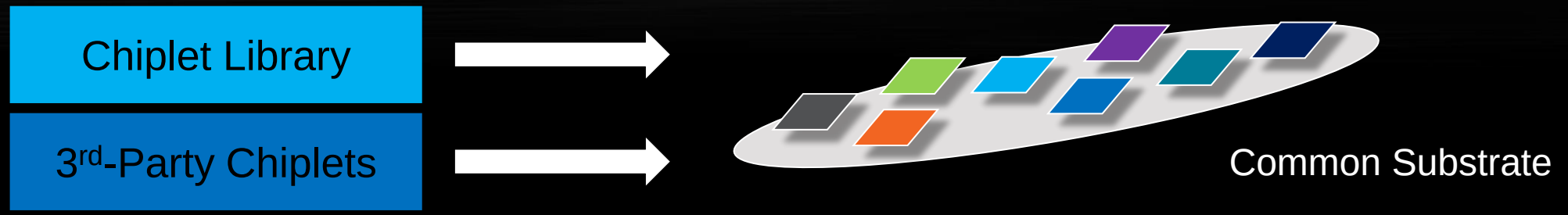




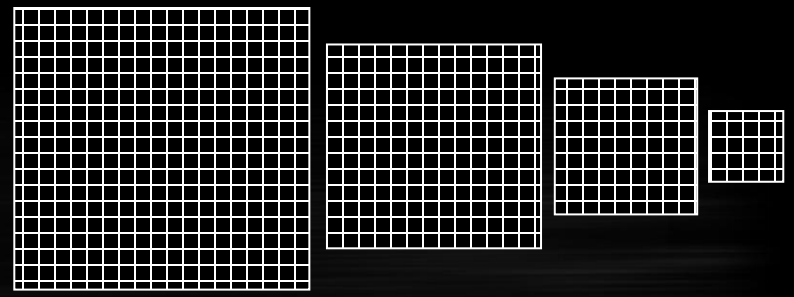
AMD

Challenges

Design For Reuse and Flexibility



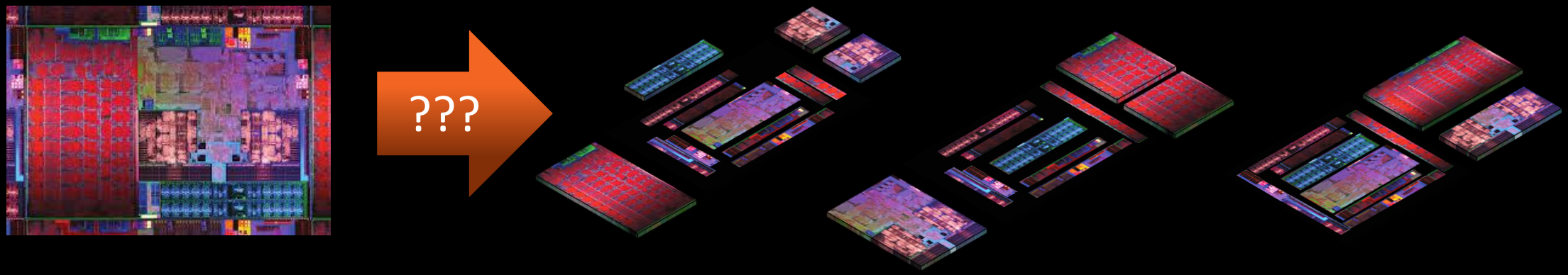
- Generality vs. Optimization
  - Interface widths and speeds, supported functionality/protocol(s), fixed pinouts
  - Memory options (e.g., not all systems need HBM)
  - Form factor/chiplet size (hard to support too many sizes, unnecessarily large silicon increases costs)
  - Power delivery: pinouts, supported voltage(s), voltage regulation, current draw, required decap
  - Thermal budgeting/allowance, cooling solutions
- Design for reuse in large-scale design
  - How to leverage wafer-scale systems across smaller scales?



AMD

Architectural Partitioning

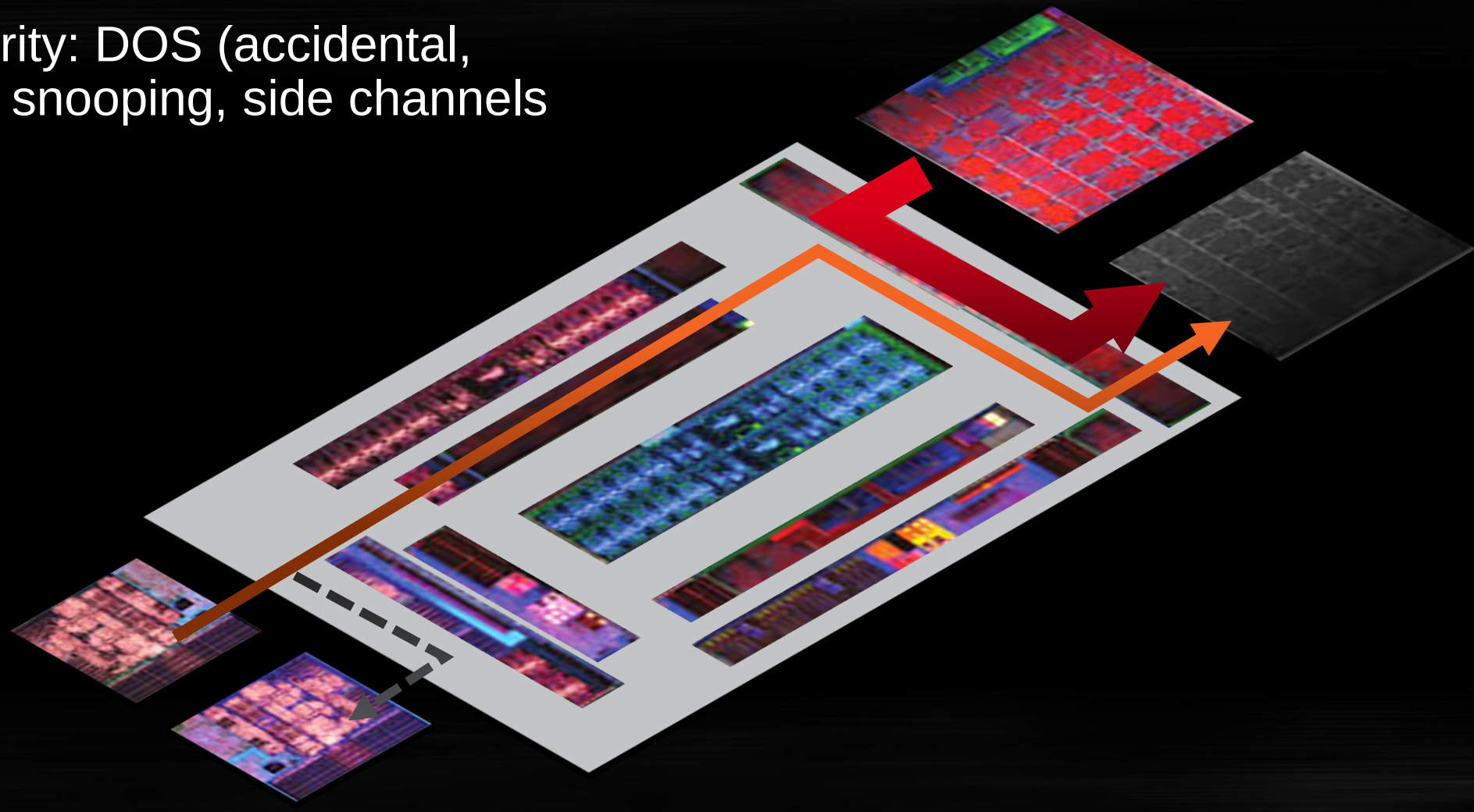
- Decomposition for **a** system



- Multi-objective technology selection and optimization
- Decomposition for N arbitrary, undesigned, unimagined systems
  - Interfaces
    - Datapath interfaces: physical, protocol
    - Control interfaces: power management, debug, profiling, security, system alerts (e.g., thermal emergency)

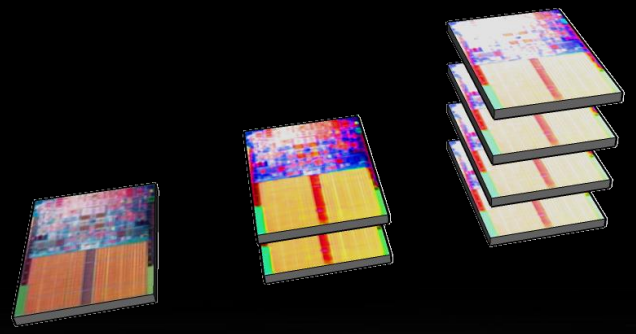
Security, Untrusted Chiplets, etc.

- QoS, Security: DOS (accidental, malicious), snooping, side channels



Parametric Variations

- Slow Die + Fast Die challenge
  - Do not want to sell as all-slow parts
    - Perform pre-assembly binning (how?)
  - Restrict frequency-boosting to fast core(s)?
- Clock distribution
  - Skew/jitter across dies

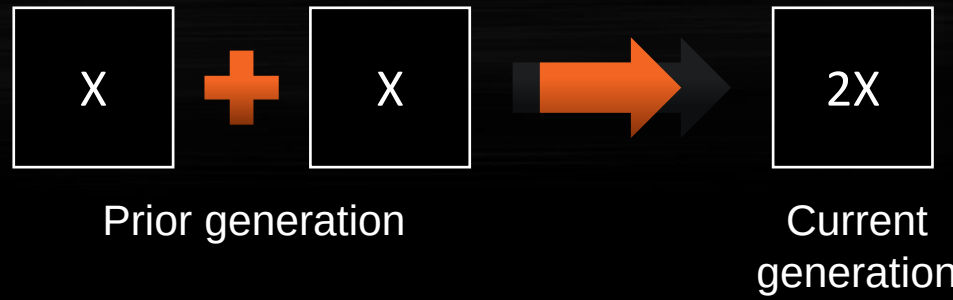


[1] WikiMedia, Creative Commons (link, license)
[2] amd.com

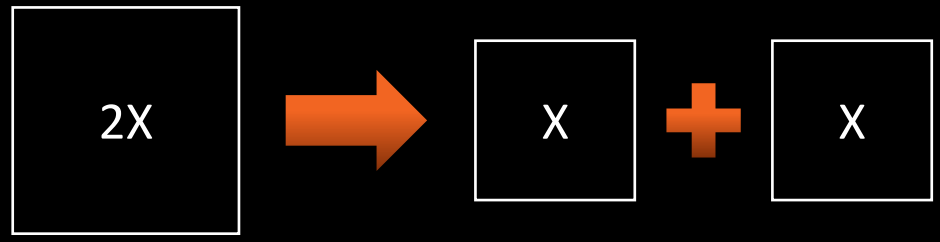
AMD

Multiple Die Are Not Free

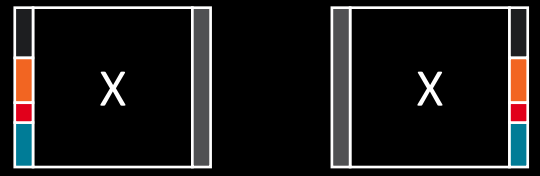
- Historically, Moore's Law enabled 2X the transistors each generation to meet compute needs



- Splitting compute into two dies is not sufficient



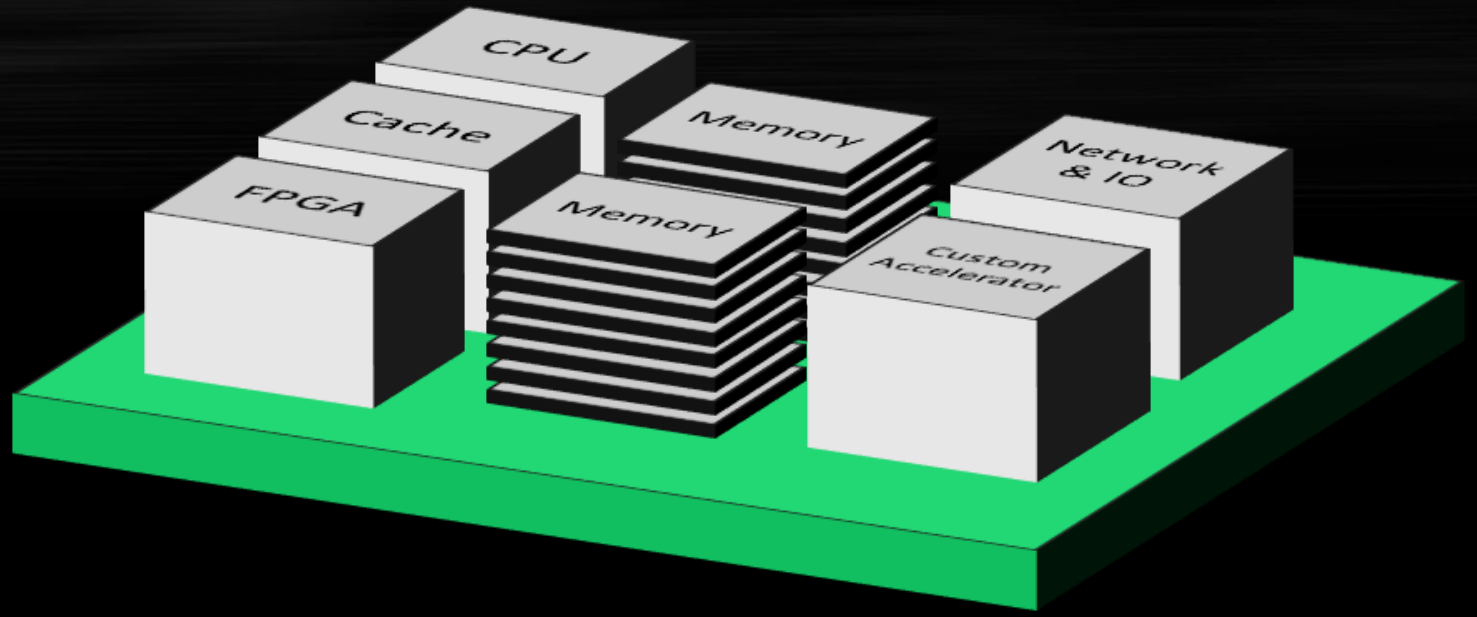
- Chiplets are not free
  - Additional area for interfaces, replicated logic
  - Higher packaging costs
  - Additional design effort, complexity
  - Past methodologies less suited for chiplets



2X device functionality costs > 2X silicon area

AMD

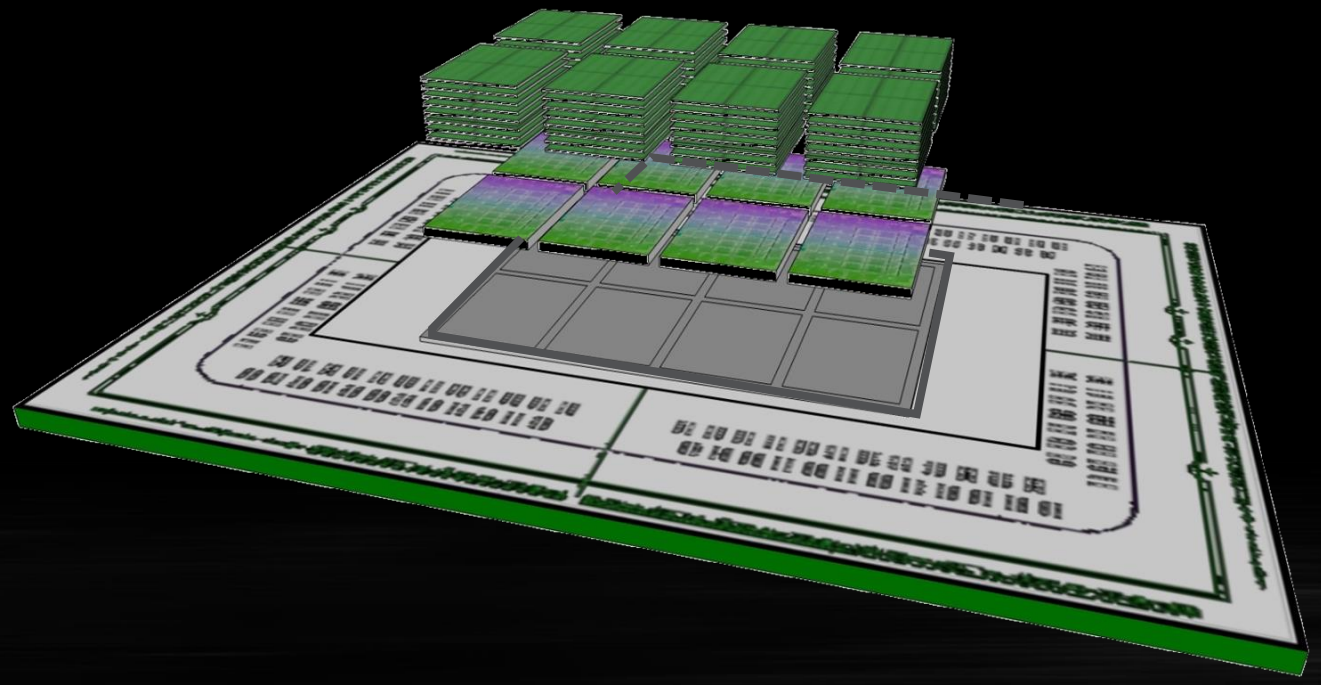
Reticle Limit



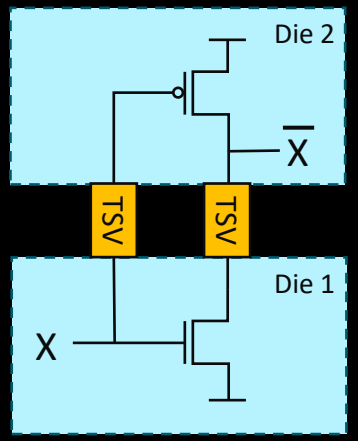
- When placing many die side by side in an array such as this, X-Y dimensions can be quite large
- Traditional die reticles in a silicon wafer process are ~25mmx30mm
- With interposer and fan-out solutions offering multiple stitched reticles today, there isn't much room to scale
- Need to reduce the X-Y size growth trend and find physical interconnect with larger reticle sizes

Reduce X-Y with More Stacking and Denser Interconnect

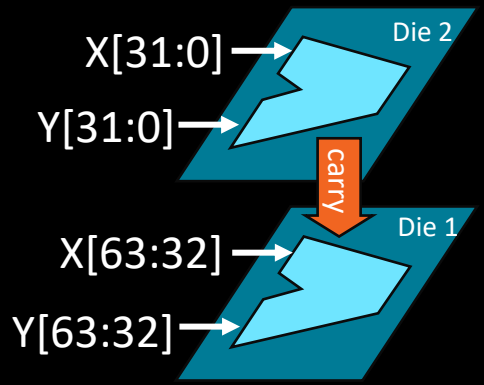
- Reduce chiplet connection overhead with interposers and denser interconnect
- Memory stacking directly on compute is possible
- True 3D stacking beyond memory is possible
  - HBM already doing memory-on-logic of sorts
  - Expand to heterogeneous components



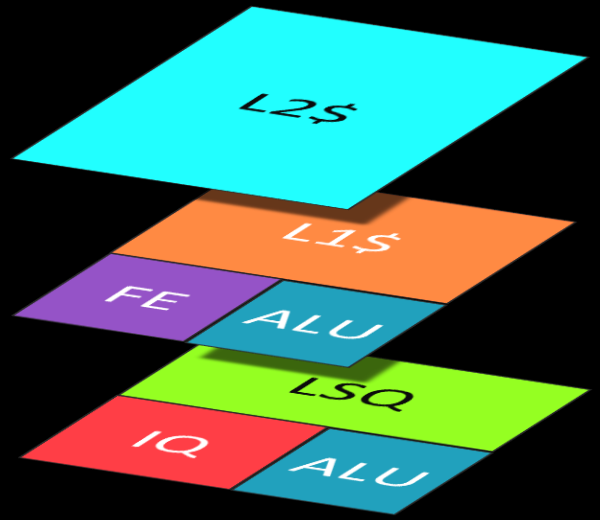
3D stacking options vs. TSV pitch



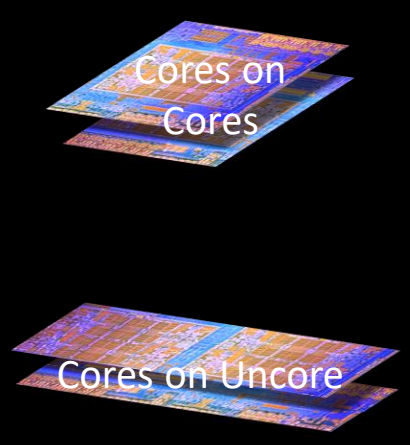
Circuit Slicing



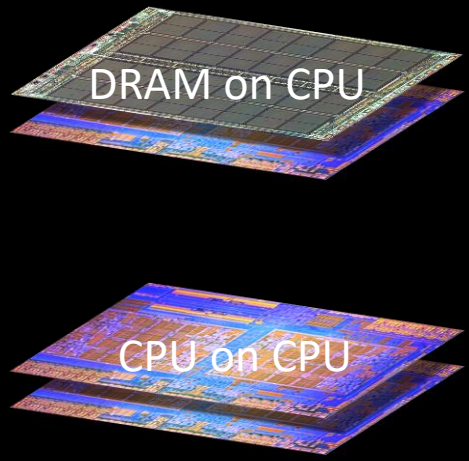
IP Folding/Splitting



Macro on Macro



IP on IP

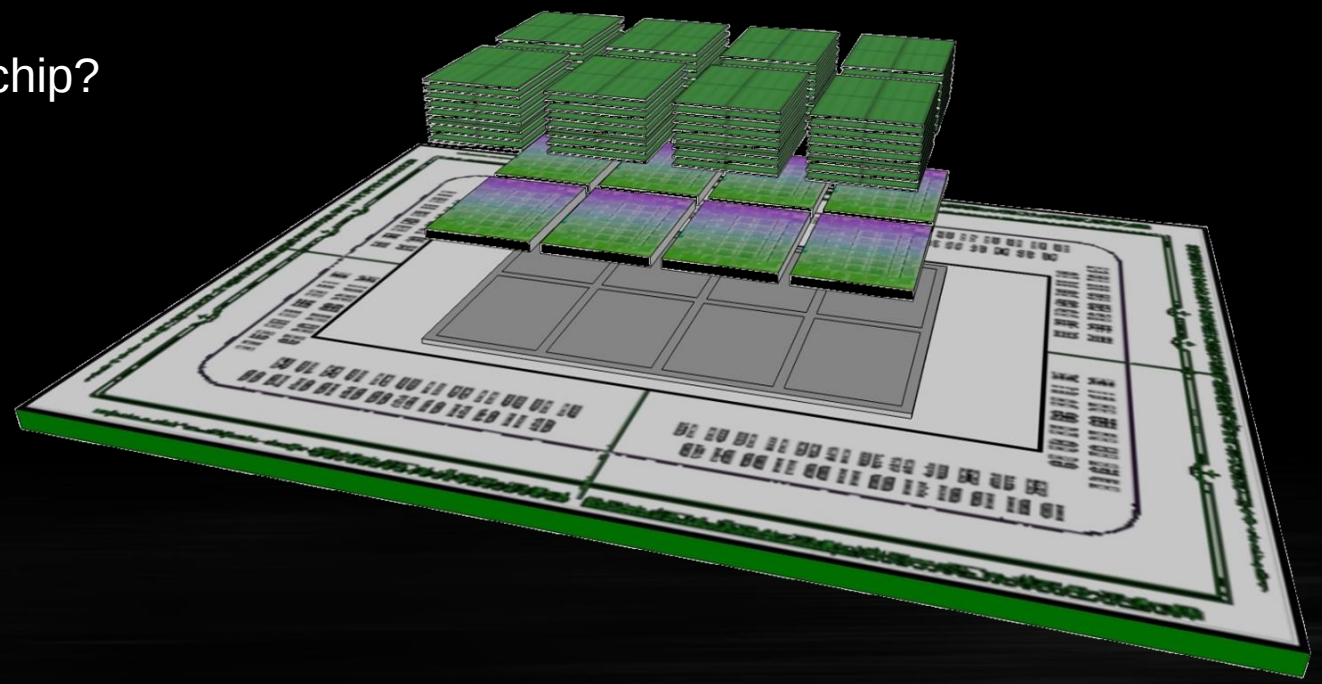


Full die-to-die

AMD

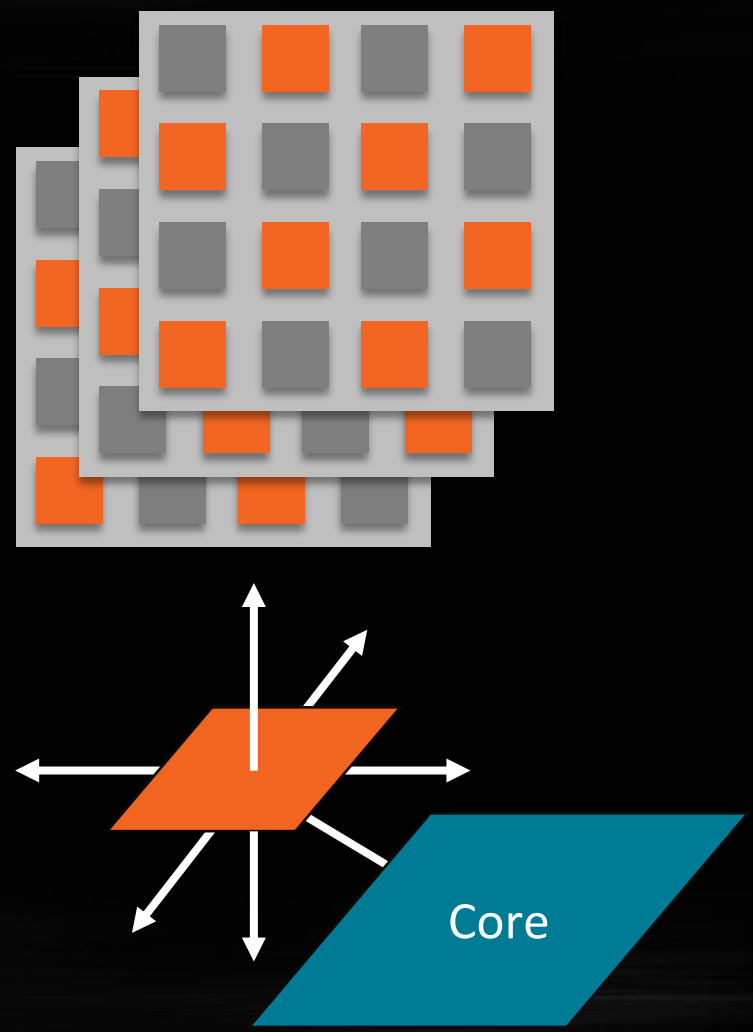
Known Good Die Testing

- One faulty die in a system can ruin the entire construction (very expensive)
- Need to test functionality (at least partially) before assembly
- Challenges
  - Limited connectivity (few pads, maybe none? Only TSVs/uBumps?)
  - Limited ability to supply power
  - Incomplete functionality
    - What if PLL/clocking is on a different chip?



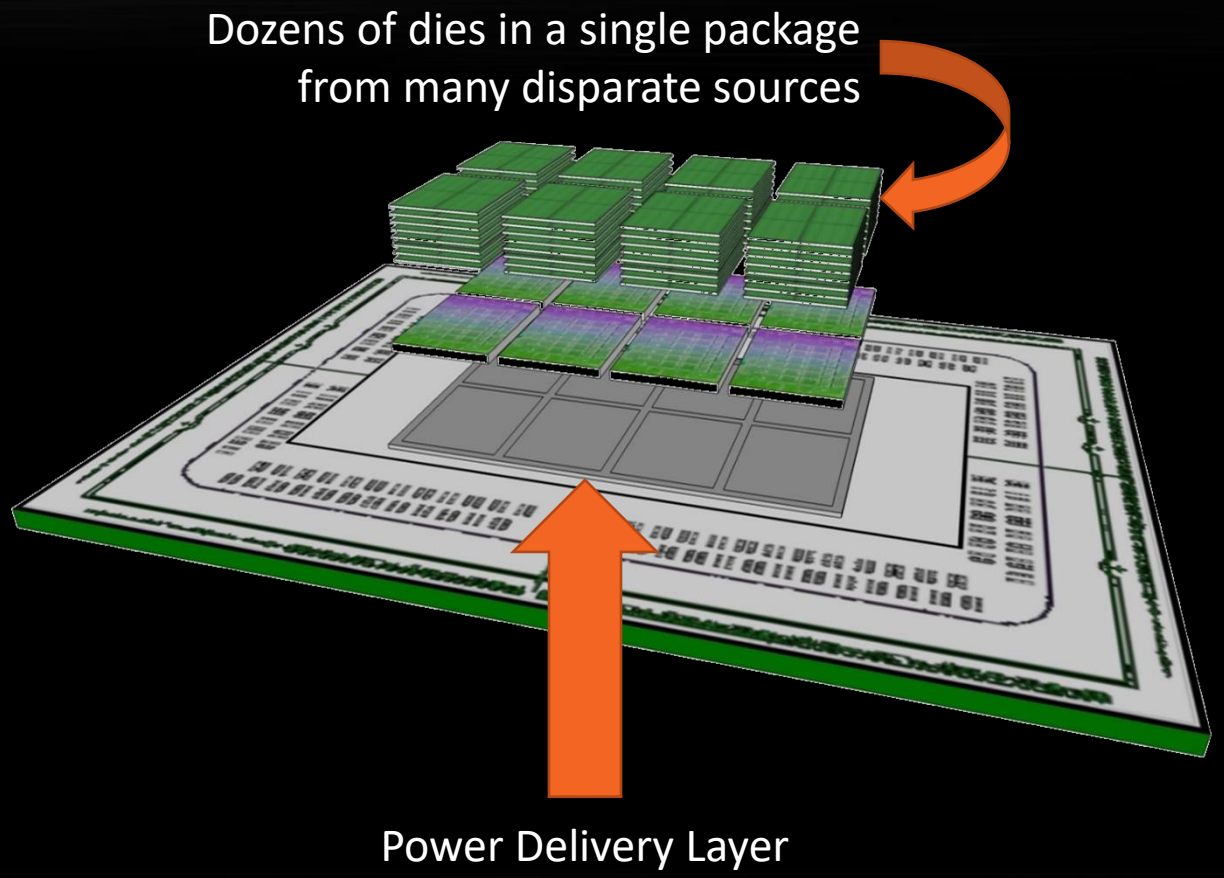
Thermal requirements

- 3D stacking increases power density
- More dies in the package require more power
- Power management must become more sophisticated
- Design for thermals is important



Power Delivery is Important

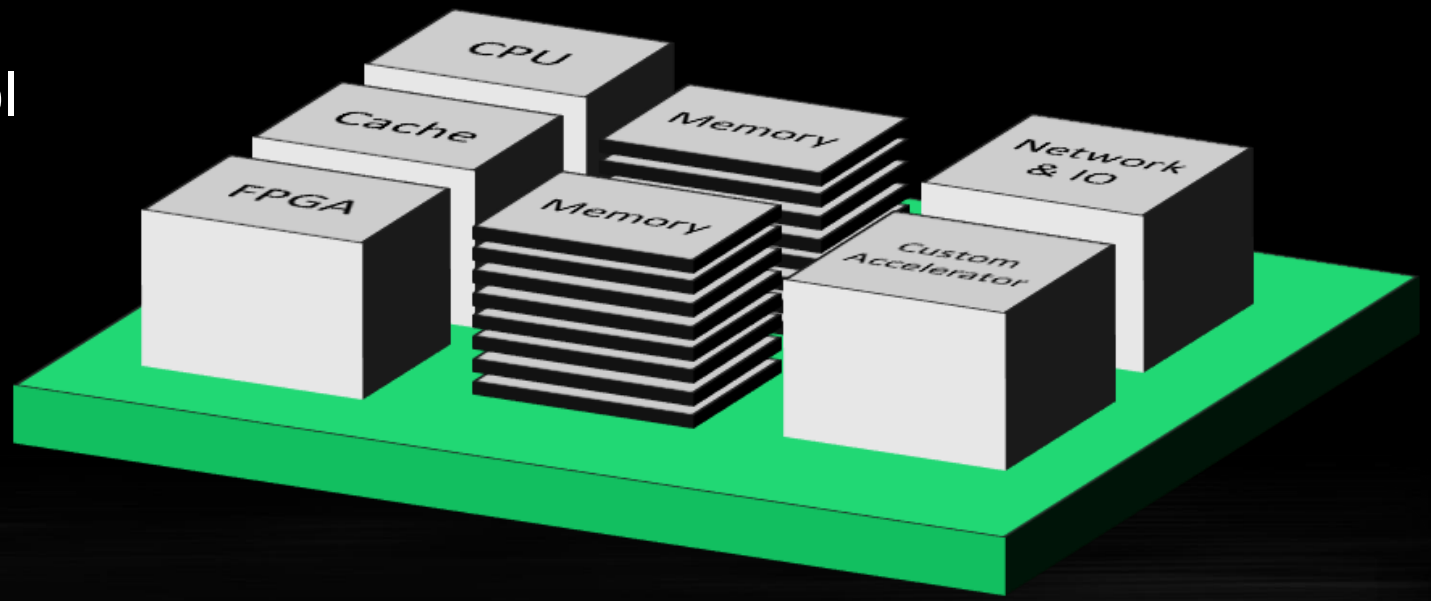
- Board level power solutions are challenged to support increasing current demand as Moore's Law scales as well as meet the demands of a more diverse set of dies in a single package



AMD

Summary

- There are many technology challenges and microarchitectural challenges
  - How to partition the dies in the system
  - Reticle size limits with interposer and fan-out
  - Power delivery limitations to total power and efficiency
  - Interconnect speed and BW scaling
  - Partitioning overhead
  - Global clocking
  - Security across different dies
  - Total heat and hotspot control



AMD

Acknowledgments

Special thanks to Sam Naffziger, Gabe Loh, Kevin Lepak, Mahesh Subramony for their contributions to this material, and of course the talented AMD design teams from around the world for the AMD EPYC™ server processor engineering achievements

Disclaimer

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. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. 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.

THIS INFORMATION IS PROVIDED ‘AS IS.” 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 NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, 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.

AMD


AMD