Innovation to meet AI and HPC computing growth

Mark Papermaster

CTO and EVP

AMD

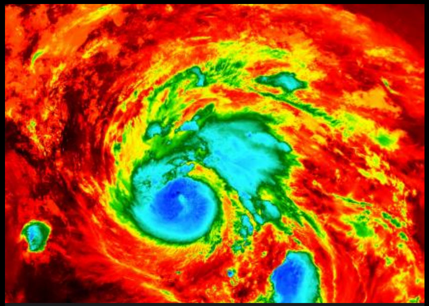
together we advance_

The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Fundamental inflection of AI powered computing

Benefits are key to solving the world’s most pressing problems

Education and Knowledge base



Climate Change

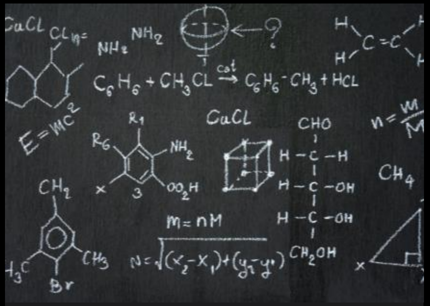
Chemical Sciences

Energy Solutions

Productivity

Healthcare

Content Creation

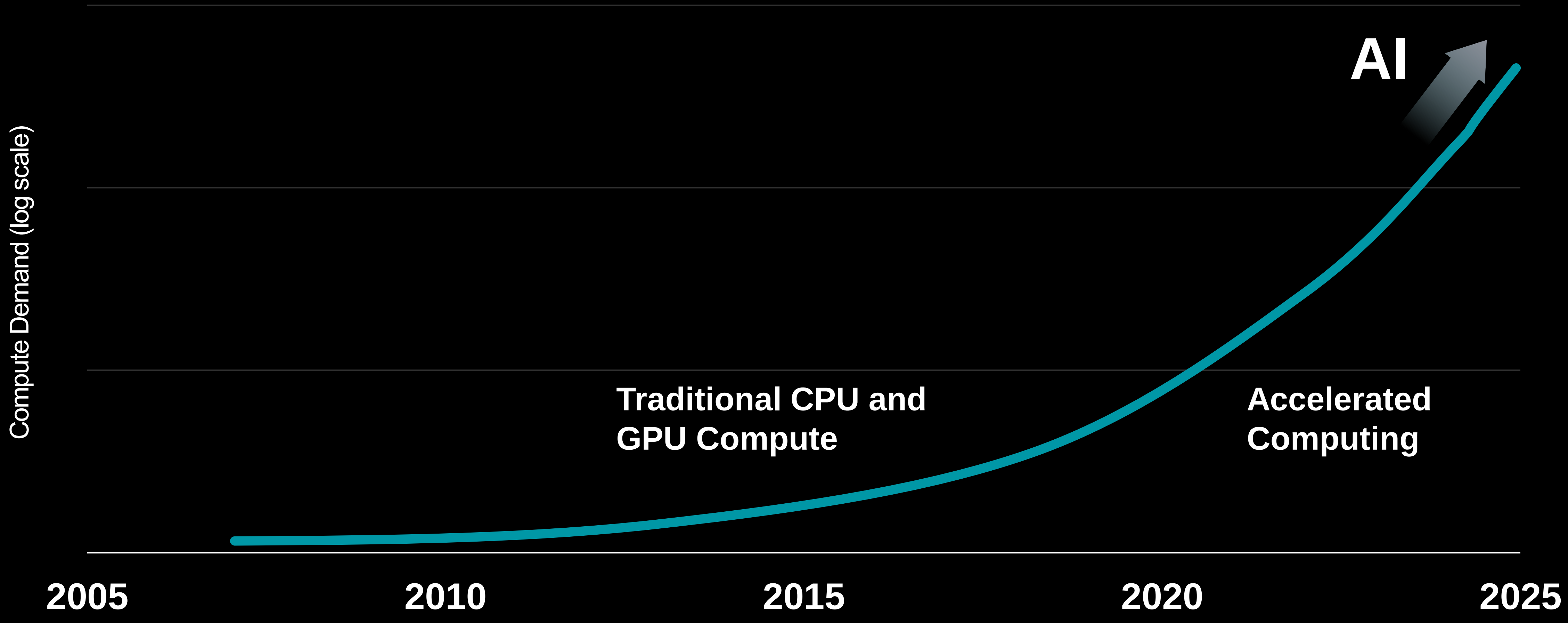


Research

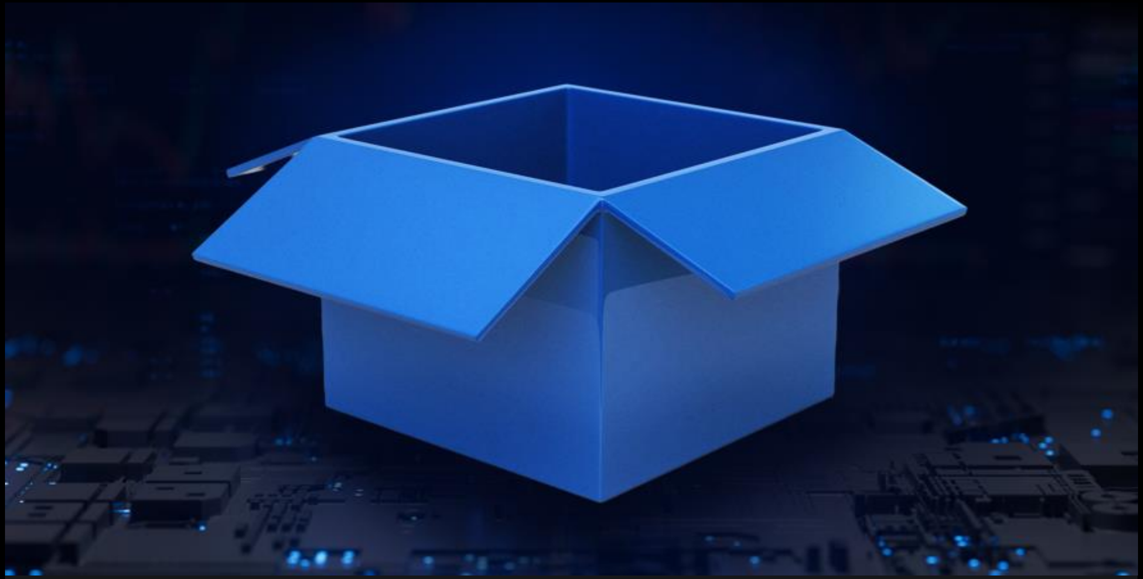
Gaming and Entertainment

Real-Time Simulation

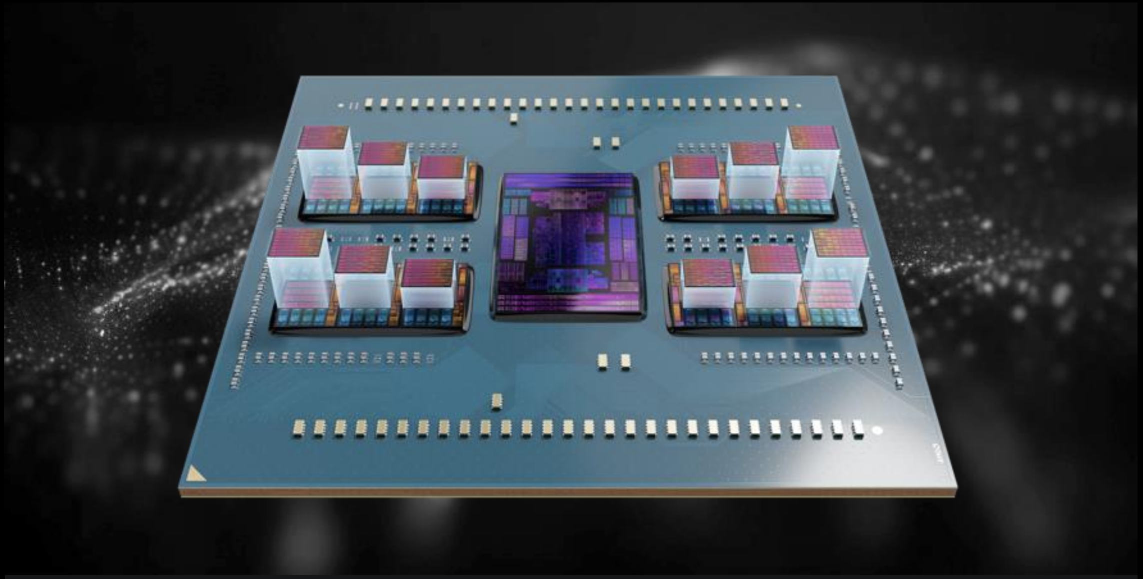
Insatiable demand for more compute



Constant innovation required in software and hardware



Open Ecosystem



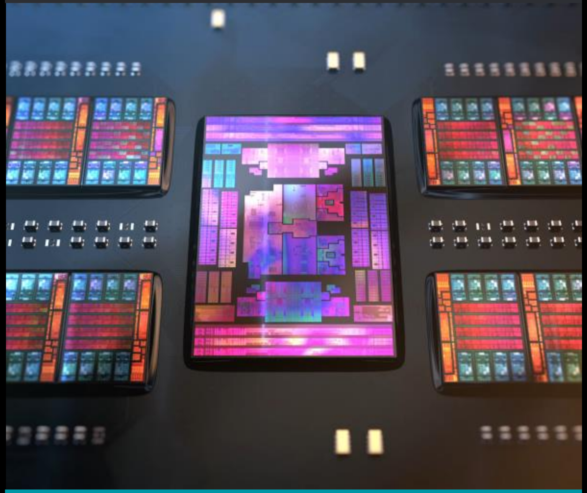
Flexible Chiplet Design

Modularity enables efficient and tailored solutions

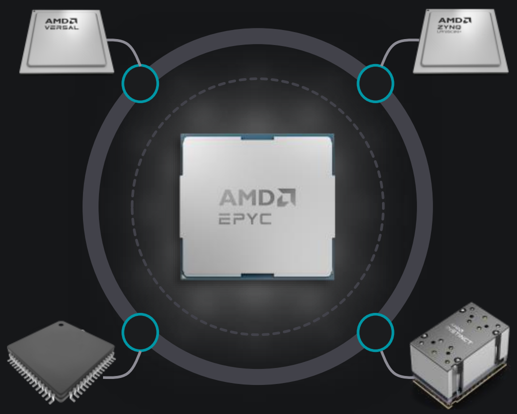
IP

8 compute units

Chiplets and 3D stacking



Component and system



Software

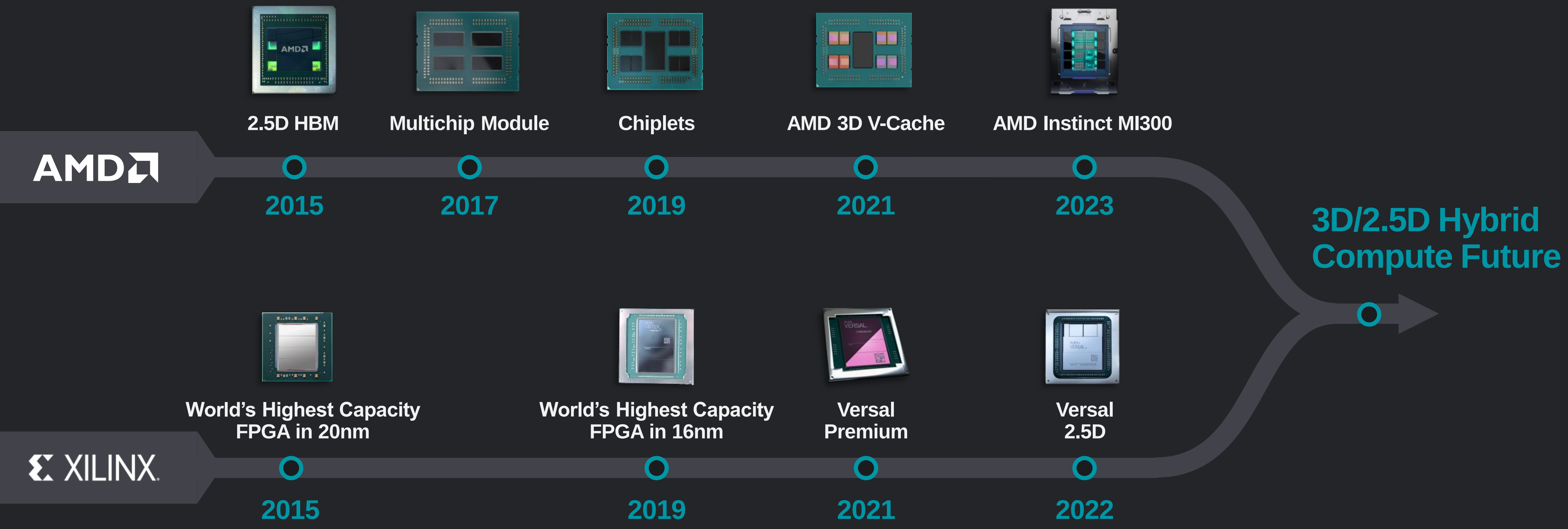
PyTorch OpenXLA

TensorFlow

Triton ONNX

deepspeed

Chiplet and packaging evolution



AMD

2.5D HBM — 2015

Multichip Module — 2017

Chiplets — 2019

AMD 3D V-Cache — 2021

AMD Instinct MI300 — 2023

XILINX

World's Highest Capacity FPGA in 20nm — 2015

World's Highest Capacity FPGA in 16nm — 2019

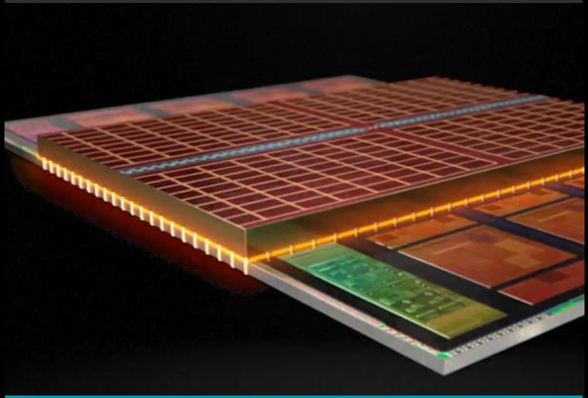
Versal Premium — 2021

Versal 2.5D — 2022

3D/2.5D Hybrid Compute Future

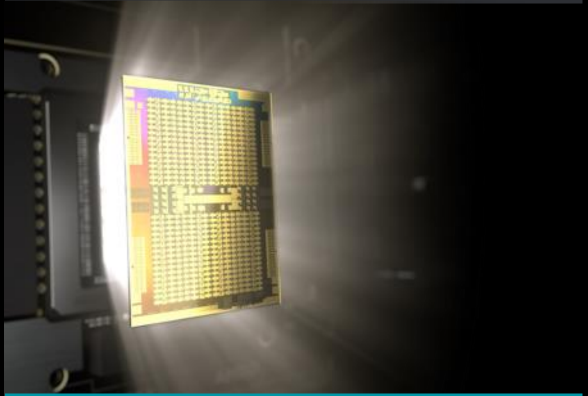
Enabling design innovations

Multichiplet SoCs



Power delivery

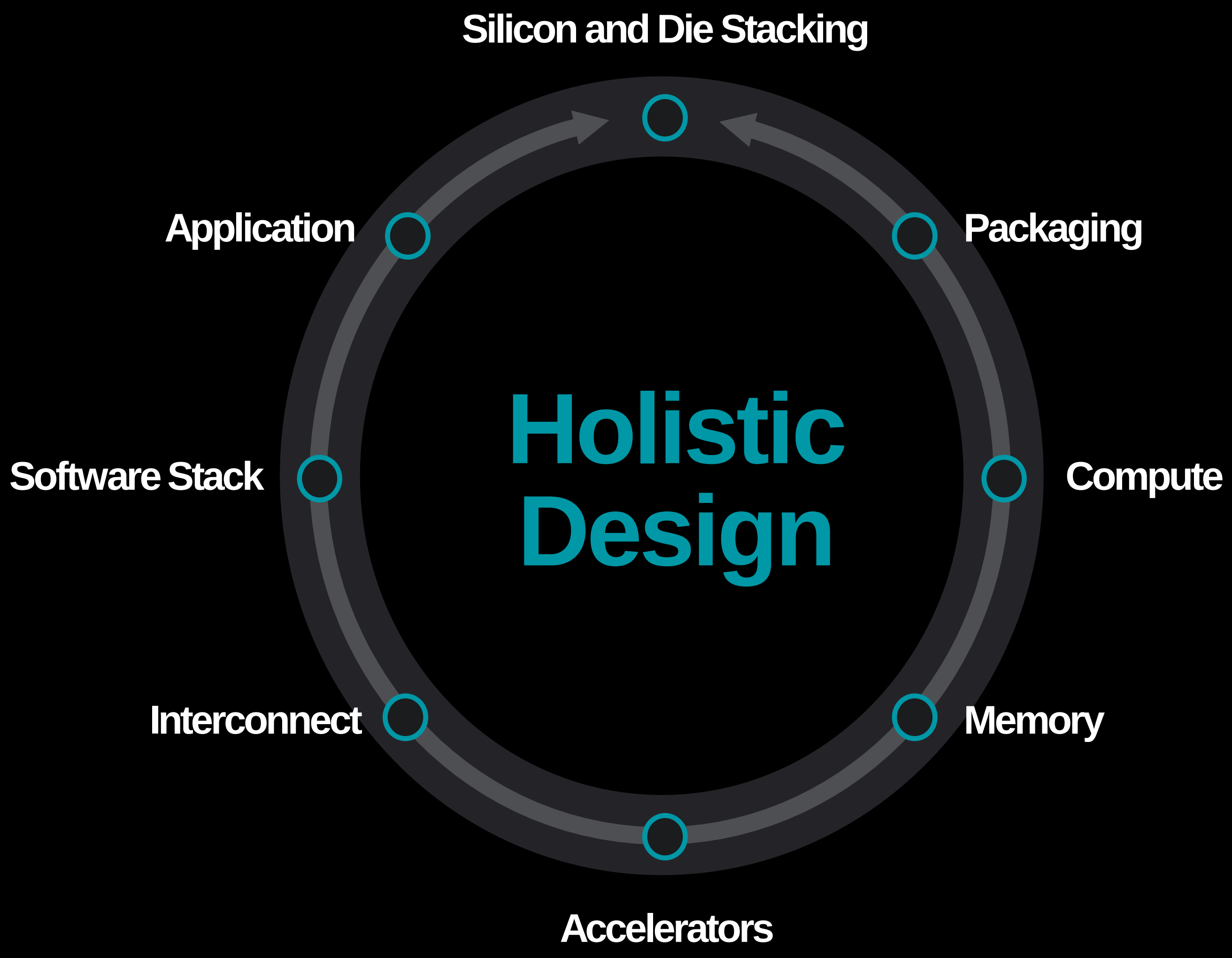
Heterogeneous tailored compute



SW and HW system level integration

A holistic design approach is required

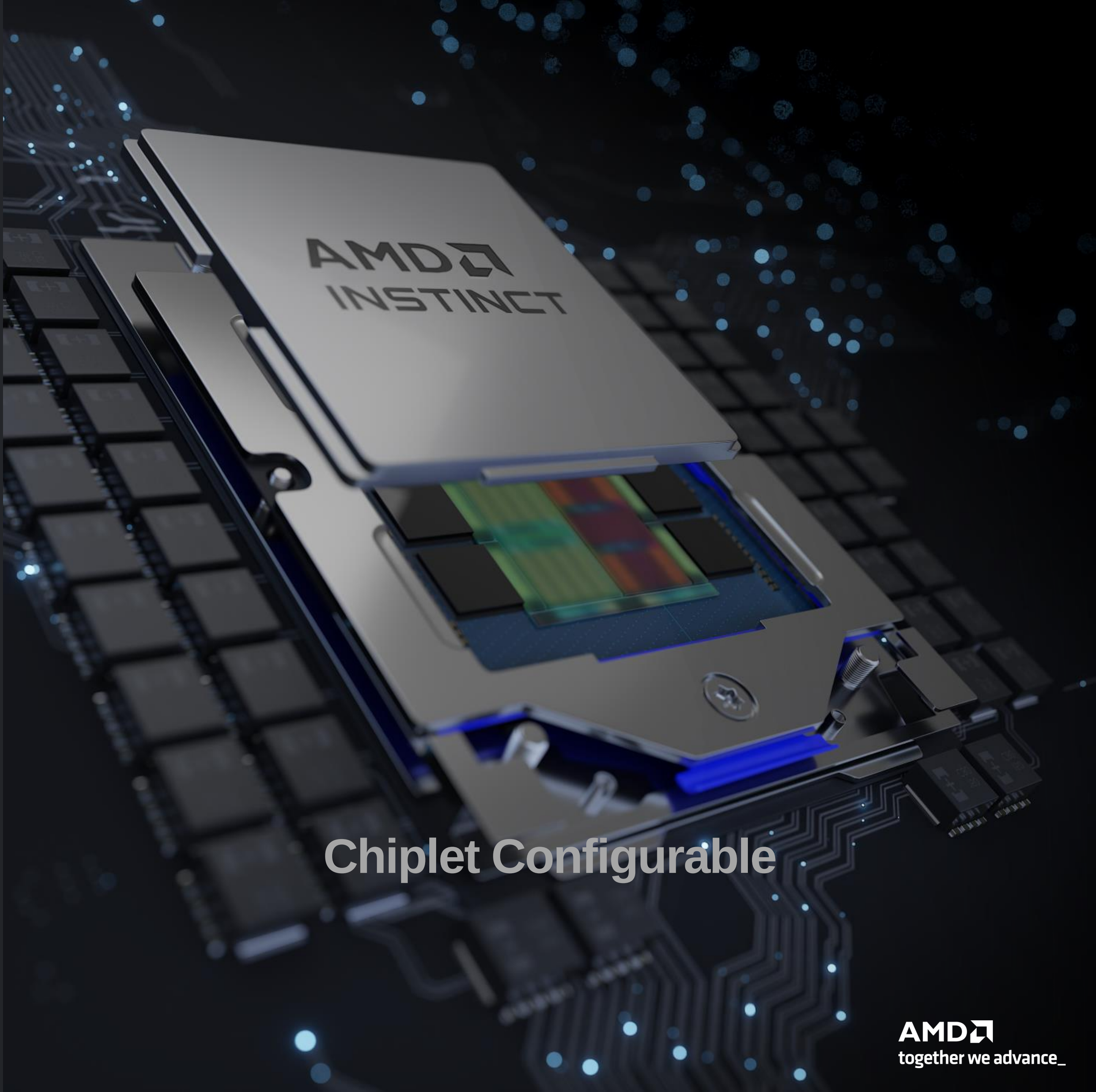
- System-level optimizations
- Domain-specific heterogeneous architecture
- Tight integration of processors, packaging and interconnect
- Leveraging AI holistically



AMD
together we advance_

MI300: Next Gen Chiplet Design

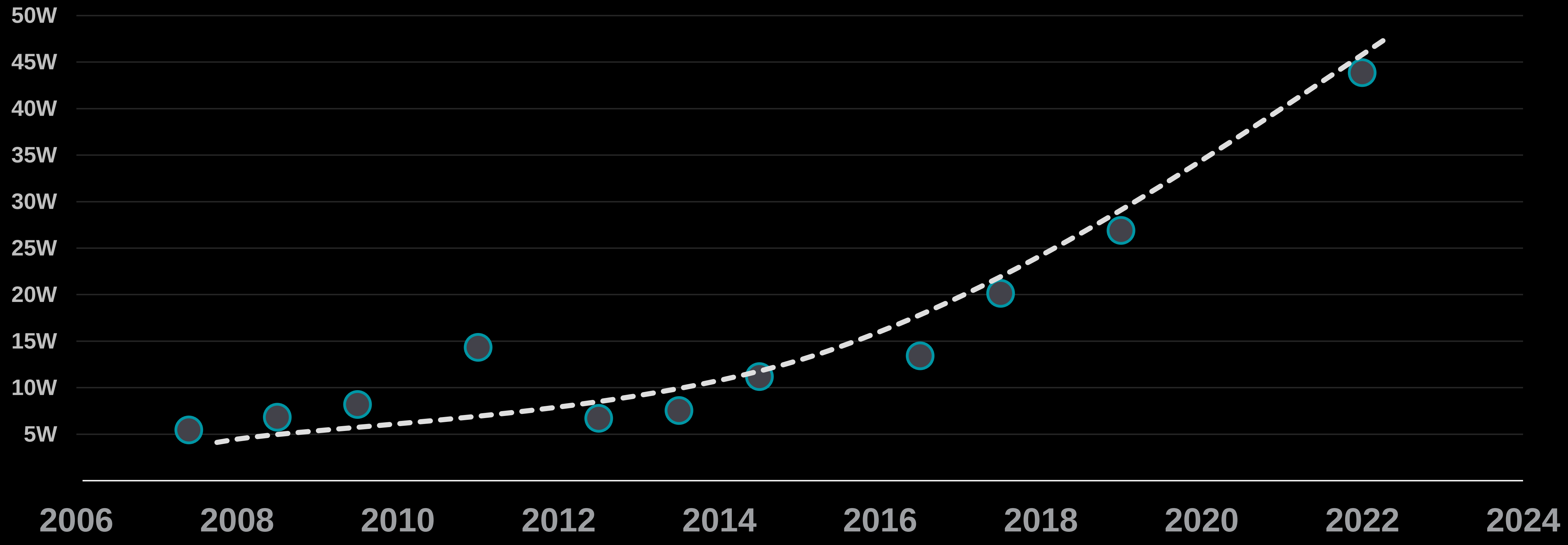
- Leadership HPC and Generative AI **accelerator** based on configuration
- 5nm process technology with 3D stacking
- Up to 5.2 TB/s memory bandwidth
- Up to 153 billion transistors
- Frameworks and open models fully supported
  PyTorch, TensorFlow, ONNX, Hugging Face, others
- Easy migration path from CUDA



Many more challenges to solve

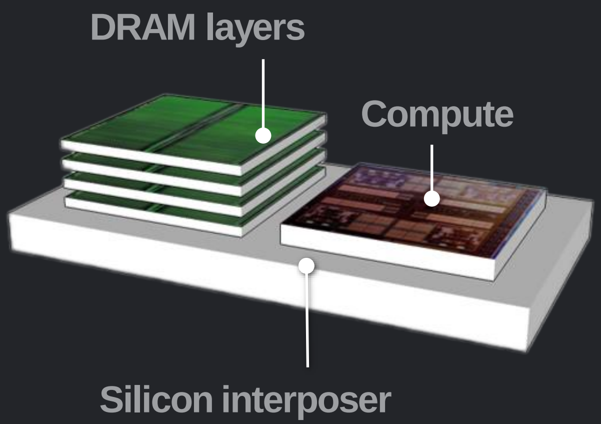
Bandwidth demands accelerating power consumption

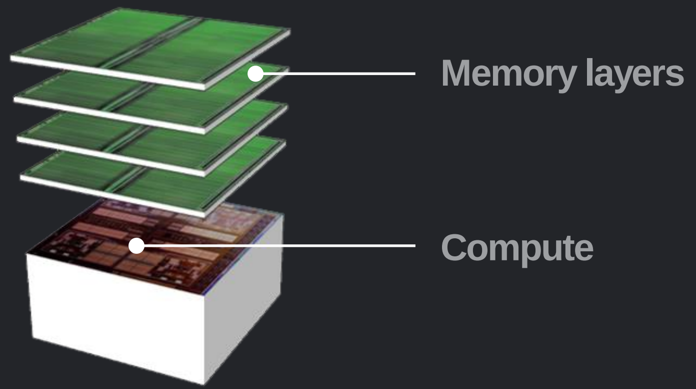
Server memory interface power



Based on AMD internal data

Even tighter integration of compute and memory





Integration enables higher bandwidth at lower power

	DIMMS	2.5D Micro-bumps (HBM)	3D Hybrid Bond
pj/bit	~12	~3.5	~0.2

Image source: https://commons.wikimedia.org/wiki/File:SDRAM-Modul.jpg, Creative Commons 4.0.

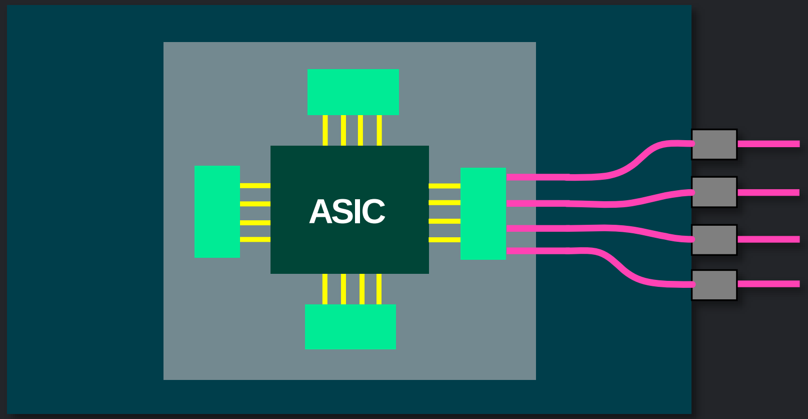
Optical communication for energy efficient connectivity

Co-packaged optics provide compelling efficiency gains

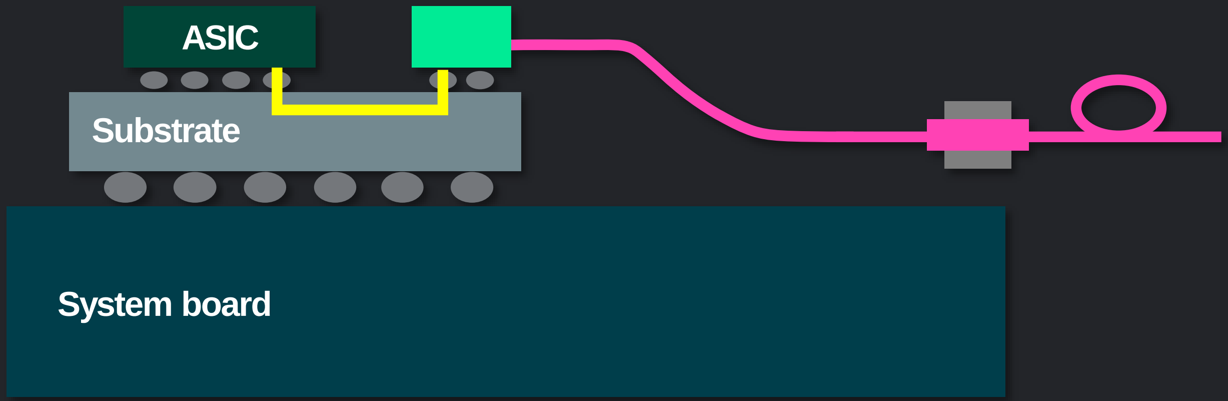
Single mode, enabling 10m up to 2km reach

Energy efficient at <1pJ/bit receive energy

Tight integration of optical transceivers to compute die is the key to efficiency



Top view



Side view

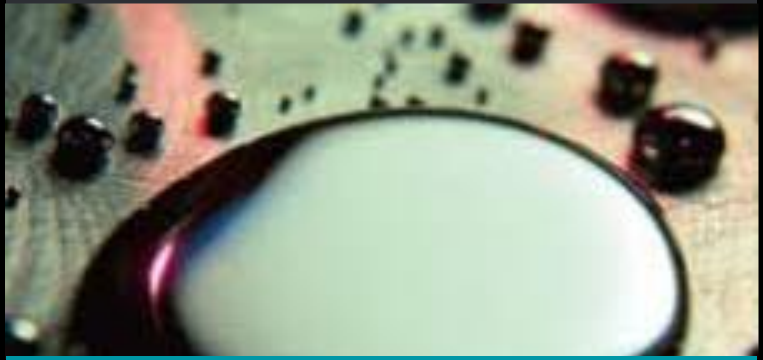
AMD
together we advance_

Heat and power density

Managing thermal density remains a priority

Thermal interface and lid materials

Improving thermal conductivity from chip to lid

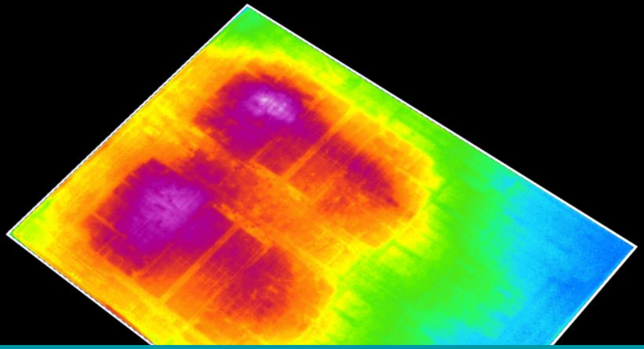


System cooling techniques

Improved ambient temperature control in system

Power management design

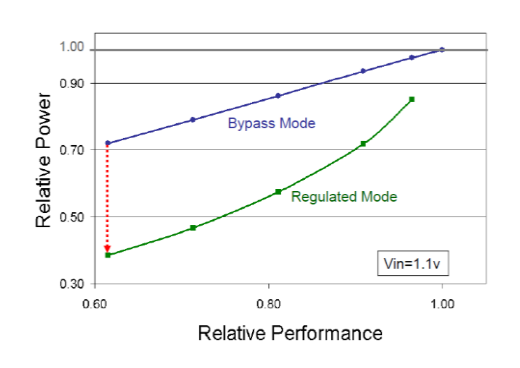
More interactive thermal management



Improved power efficiency and regulation

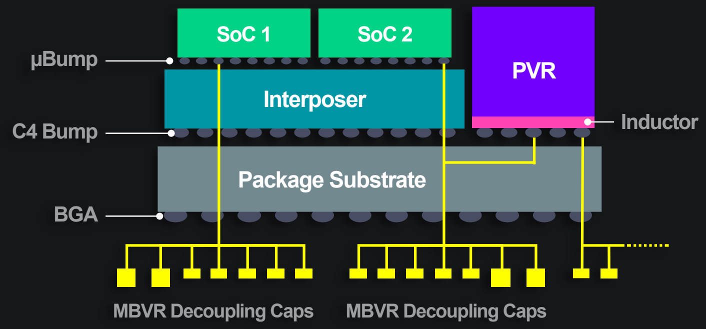
Per core regulation

Up to 19% power reduction (64-cores)



Heterogenous optimized

Optimized PDN serving multichip HPC SoC



SoC 1

SoC 2

PVR

µBump

Interposer

C4 Bump

Inductor

Package Substrate

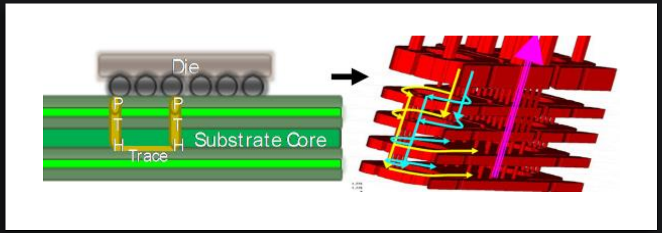
BGA

MBVR Decoupling Caps

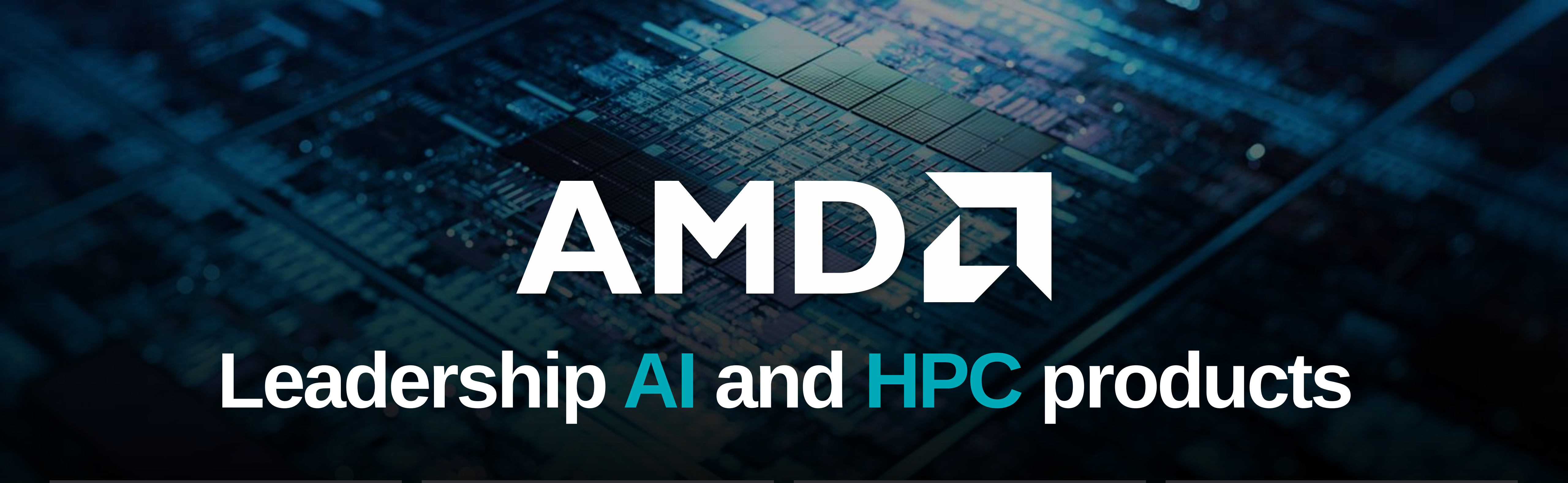
MBVR Decoupling Caps

Package integrated VRs

On-package voltage regulation



Sources: See End Notes.



AMD

Leadership AI and HPC products

- Energy efficient high-performance
- Holistic design
- Open solutions
- Public and private sector partnerships

AMD

together we advance_

ENDNOTES

Slide 15.

The source for per core regulation image is [Friedrich14]. J. Friedrich, et. al., “POWER8: A 12-Core Server-Class Processor in 22nm SOI with 7.6Tb/s Off-Chip Bandwidth,” ISSCC, Feb. 2014. The source for package integrated VRs image is [Kurd14] N. Kurd, et. al., “Haswell: A Family of IA 22nm Processors,” ISSCC, Feb. 2014.

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. 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