AMD ACCELERATING TECHNOLOGIES FOR EXASCALE COMPUTING

BILL.BRANTLEY@AMD.COM, FELLOW

3 OCTOBER 2016

AMD'S VISION FOR EXASCALE COMPUTING

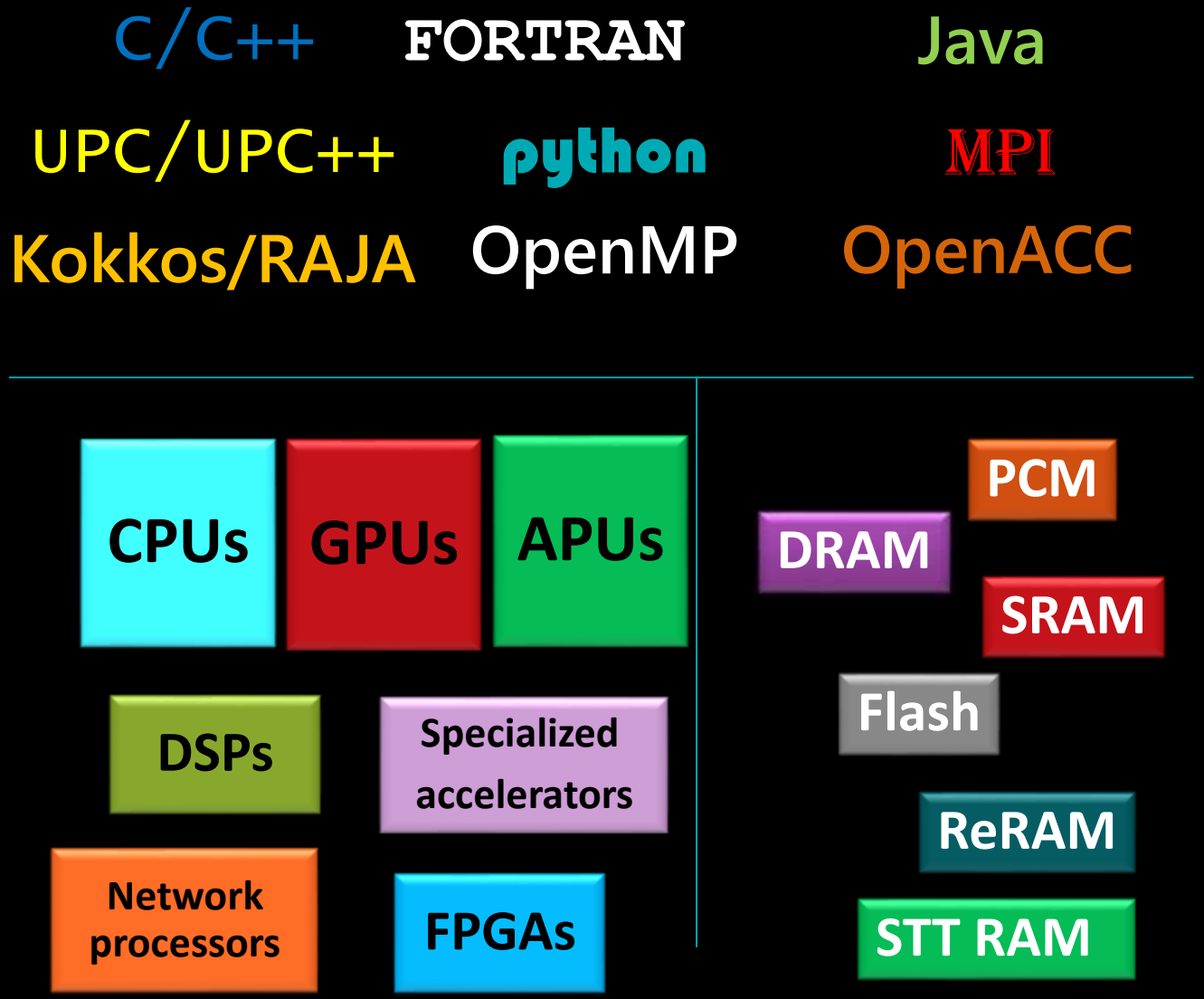
EMBRACING HETEROGENEITY

CHAMPIONING OPEN SOLUTIONS

ENABLING LEADERSHIP SYSTEMS

EMBRACING HETEROGENEITY

- Customers must be free to choose the technologies that suit their problems
  - Programming languages
  - Compute engines
  - Memory technologies
- Specialization is key to high performance and energy efficiency
- Heterogeneity should be managed by programming environments and runtimes
- The Heterogeneous System Architecture (HSA) and Radeon Open Compute Platform for GPUs (ROCm) provides:
  - A framework for heterogeneous computing
  - A platform for diverse programming languages



Heterogeneity Options

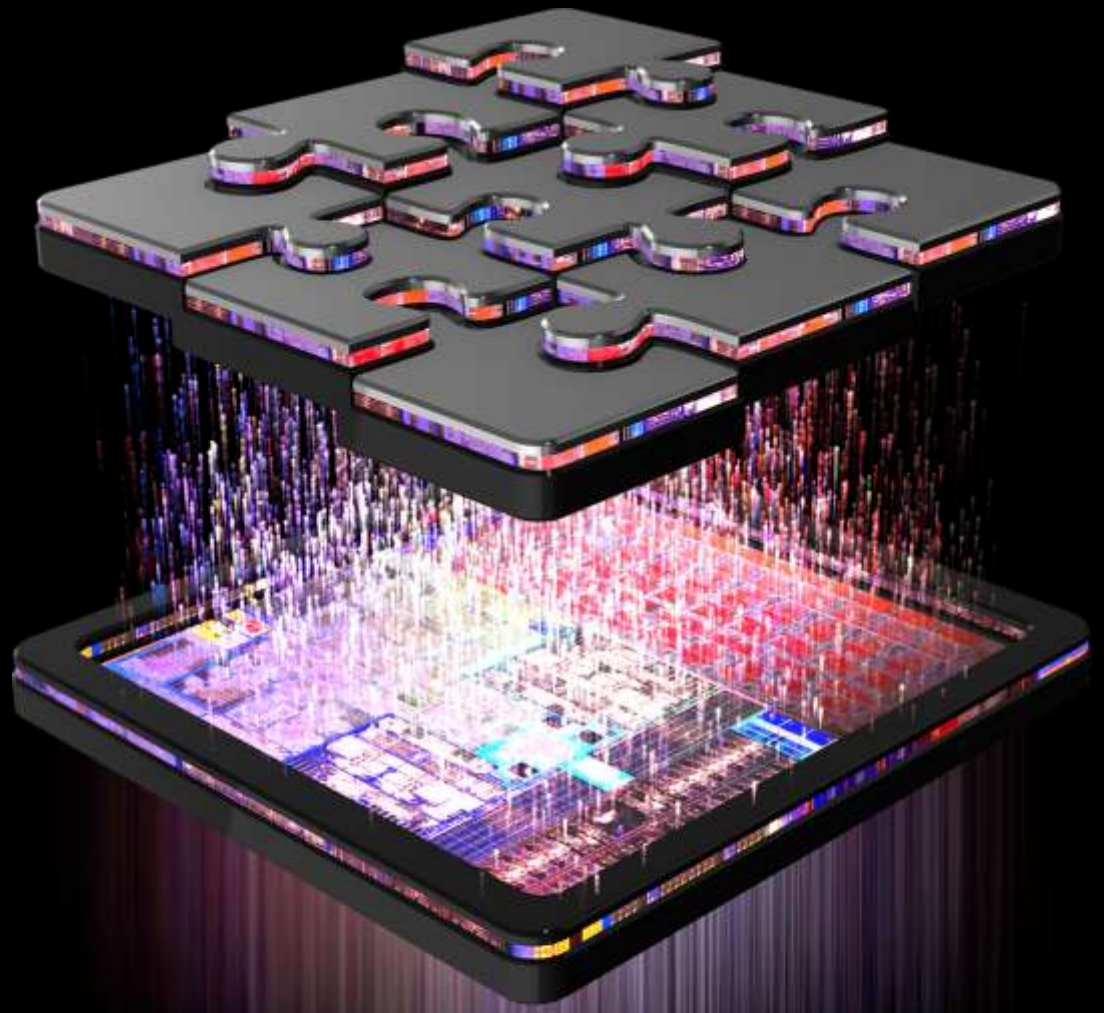
CHAMPIONING OPEN SOLUTIONS

- Harness the creativity and productivity of the entire industry
- Partner with best-in-class suppliers to enable leading solutions
  - Memory and interconnect technology
  - Software tools
  - System integration
- Multiple paths
  - Open standards
  - Open-source software
  - Open collaborations across industry, academia, and government agencies

torch

ENABLING LEADERSHIP SYSTEMS

- **Re-usable, high-performance technology building blocks**
- **High-performance network on chip**
- **Modular engineering methodology and tools**
- **Software tools and programming environments**

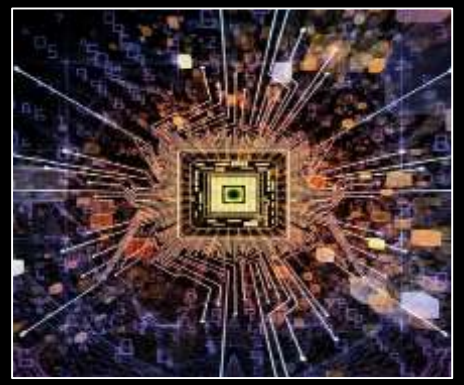


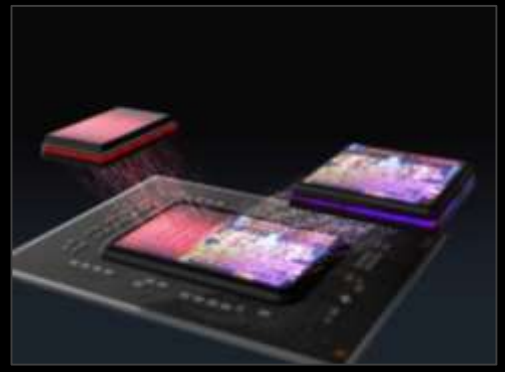
TECHNOLOGIES FOR EXASCALE COMPUTING

AMD TECHNOLOGIES: INVESTING IN THE FUTURE

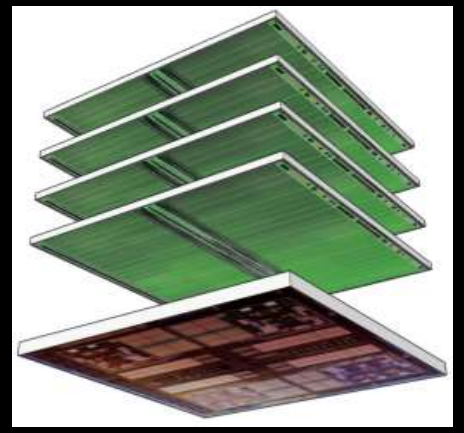
AMD

CPU CORES	GRAPHICS PROCESSORS	ACCELERATED PROCESSING UNITS





NODE TECHNOLOGIES	ENERGY EFFICIENCY	SOFTWARE



INTRODUCING ROCM SOFTWARE PLATFORM

A New Fully Open Source Foundation for HPC Class GPU computing

Graphics Core Next Headless Linux® 64-bit Driver

- Multi-GPU Shared Virtual Memory
- Large Memory Single Allocation
- Peer-to-Peer Multi-GPU
- Peer-to-Peer with RDMA
- Systems Management API and Tools

HSA drives rich capabilities into the ROCm hardware and software

- User Mode Queues
- Architected Queuing Language
- Flat memory Addressing
- Atomic Memory Transactions
- Process Concurrency & Preemption

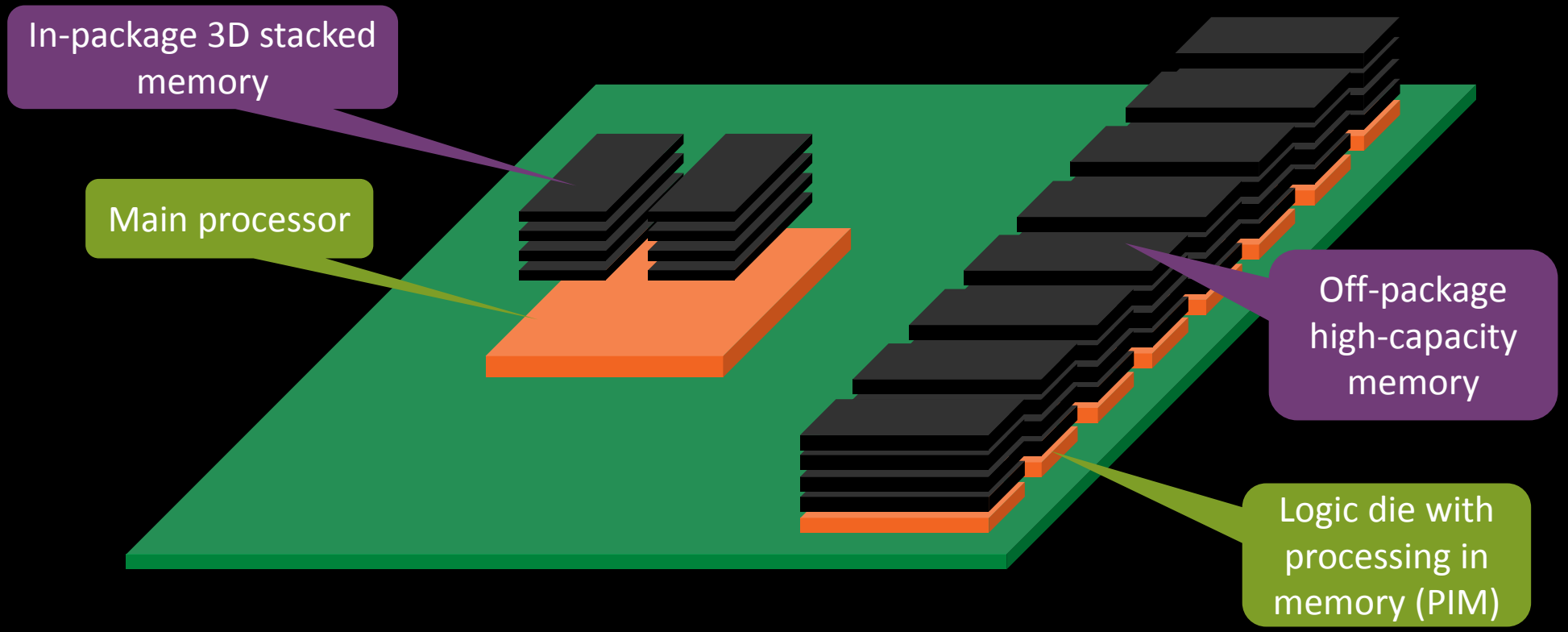
Rich Compiler Foundation for HPC Developer

- LLVM Native GCN ISA Code Generation
- Offline Compilation Support
- Standardized loader and Code Object Format
- GCN ISA Assembler and Disassembler

Open Source Tools and Libraries

- Rich Set of Open Source Math Libraries
- Tuned Deep Learning Library
- Optimized Parallel Programing Frameworks
- CodeXL Profiler and GDB Debugging
- Open CUDA porting tool, HIP

HETEROGENEOUS MEMORY SYSTEMS

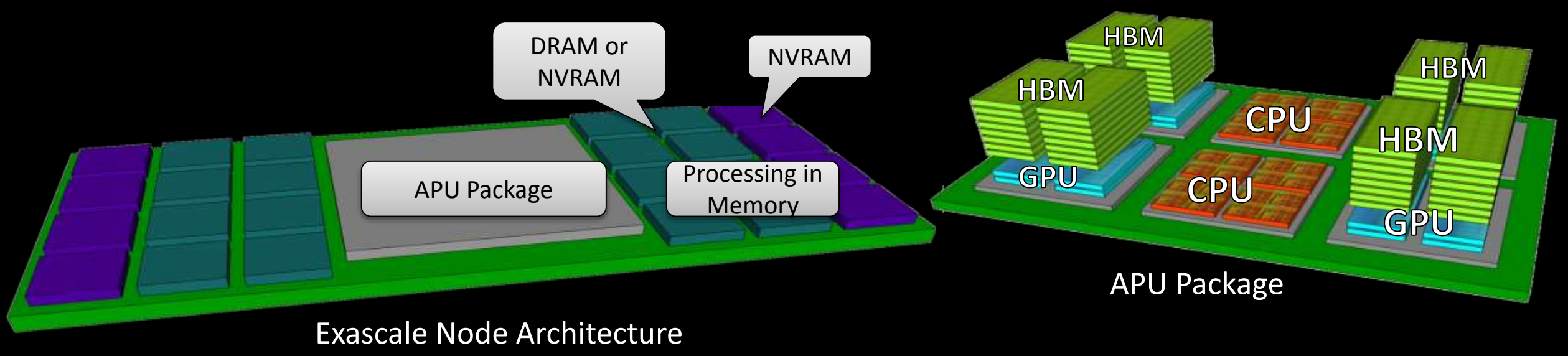


- Leverage memory stacking, non-volatile memory, and processing-in-memory (PIM) to provide very high memory bandwidth and capacity
- Data management is critical to exploit locality and limit data movement
- Opportunities to optimize processors and software for near-memory accesses

EXASCALE RESEARCH AND DEVELOPMENT

FASTFORWARD 2 NODE ARCHITECTURE

- Node Architecture Design, Integration, and Evaluation
- Parallel Programming Environments and Applications
- Power Efficiency and Reliability
- APU and GPU Microarchitecture
- Advanced Memory Architectures and Data Movement
- Extensive Evaluation via Test Chips and an Exascale Node Architecture Testbed

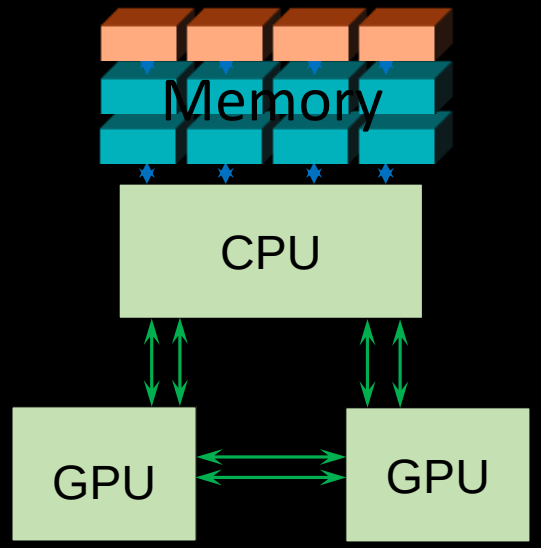


Exascale Node Architecture

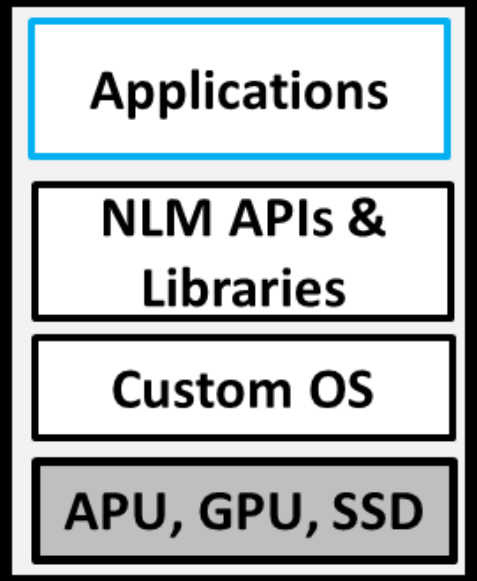
FASTFORWARD 2 MEMORY TECHNOLOGIES

AMD

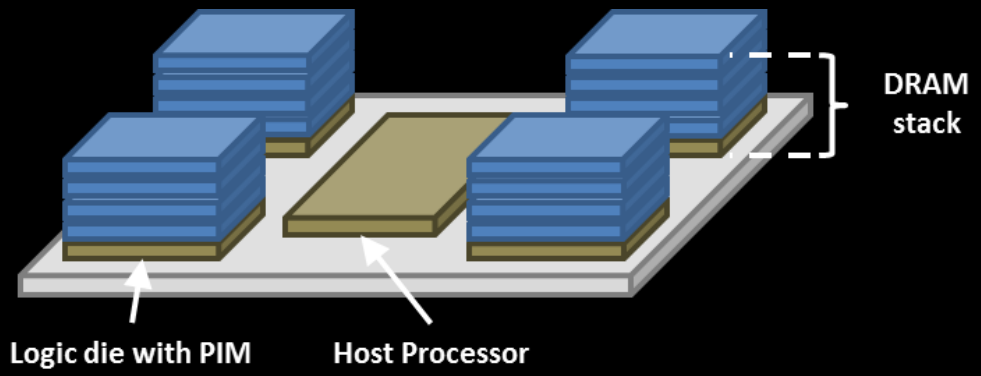
- **New Memory Interface** (NMI) – develop and propose a NMI standard (NVRAM, PIM, accelerators)
- **N-Level Memory** (NLM) – enable & demonstrate NLM architectures, libraries, APIs, and software tools
- **Processing-in-Memory** (PIM) – investigate PIM architectures, APIs, and programming abstractions
- **PIM Test Bed** – FPGA-based hardware test bed
  – Demonstration vehicle and software development platform



New Memory Interface Design



N-Level Memory Test Bed

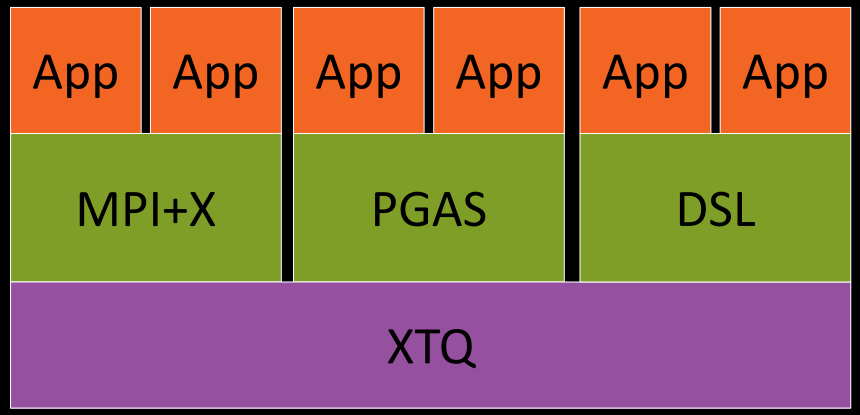


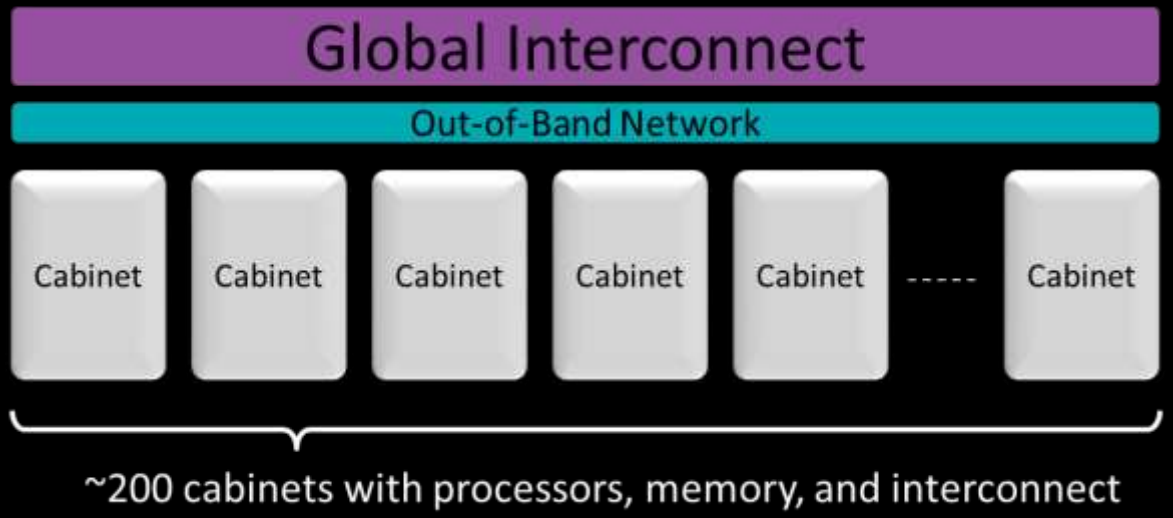
PIM-based Architecture

DESIGNFORWARD AND DESIGNFORWARD 2

- DesignForward explores extending key HSA capabilities to multi-node systems
  - Builds on the HSA features of user-level queuing and shared virtual addressing
  - Develops an eXtended Task Queuing (XTQ) architecture for inter-node tasking and communication
  - Provides support for high-level parallel programming environments

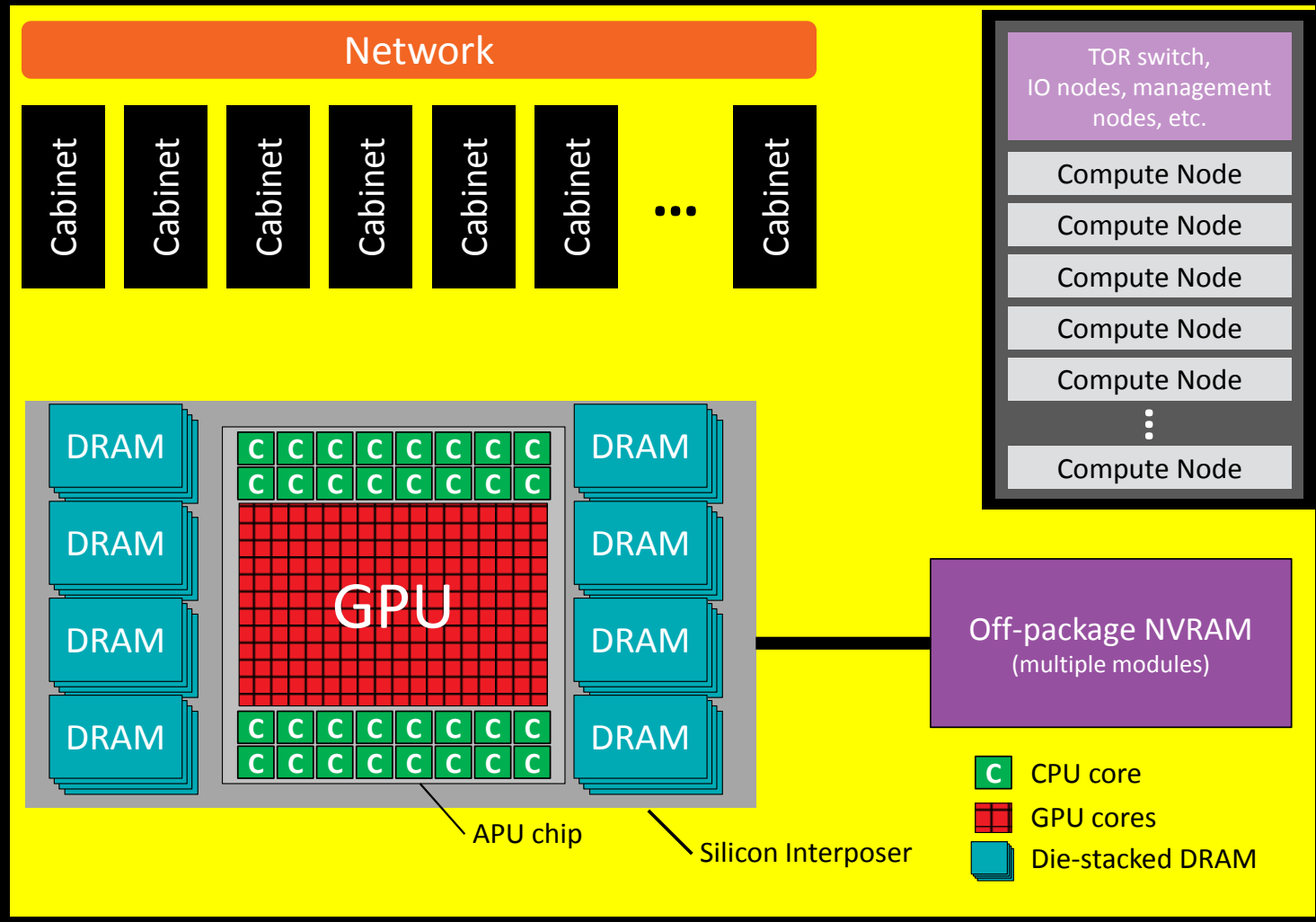
- DesignForward 2 develops a conceptual system design and execution model for exascale computing
  - Analyzes the impact of the conceptual system design and execution model on key exascale challenges
  - Conducts an analysis of various component technology options
  - Explores the impact of design trade-offs on HPC applications and workflows





CONCLUSIONS

- Exascale systems require enhanced performance, power-efficiency, reliability, scalability, and programmer productivity
  - Significant advances are needed in multiple areas and technologies
- Exascale systems will be heterogeneous
  - Programming environments and runtimes should manage heterogeneity
- AMD's technologies provide a path to productive, power-efficient exascale systems
- Technology transfer and co-design will help ensure these technologies are available for use in future for HPC and data-centric systems



For further details see: "Achieving Exascale Capabilities through Heterogeneous Computing," IEEE Micro, July/August 2015.

Thank You!

DISCLAIMER & 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