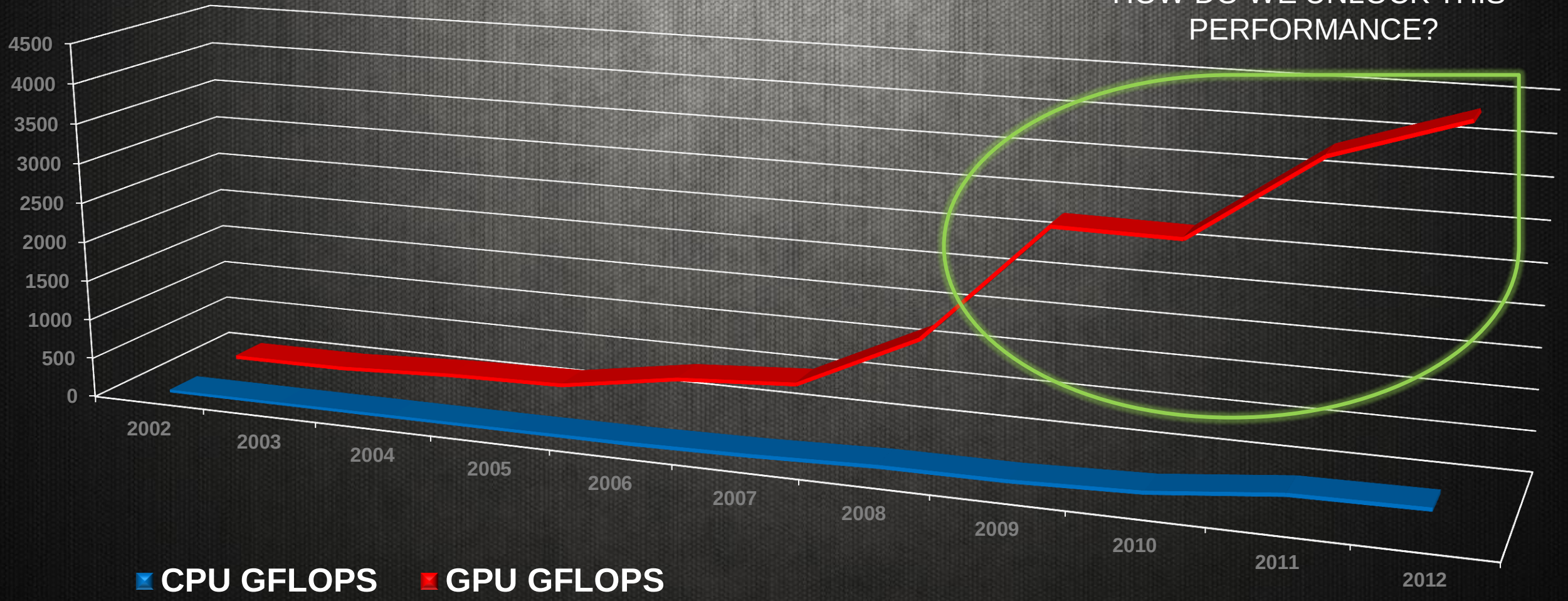
AMD heterogeneous Uniform Memory Access

HANJIN CHU,
DIRECTOR, GAME, SOFTWARE & HETEROGENEOUS SOLUTIONS

ABOUT HSA

GPU COMPUTE CAPABILITY IS MORE THAN 10X THAT OF THE CPU

HOW DO WE UNLOCK THIS PERFORMANCE?



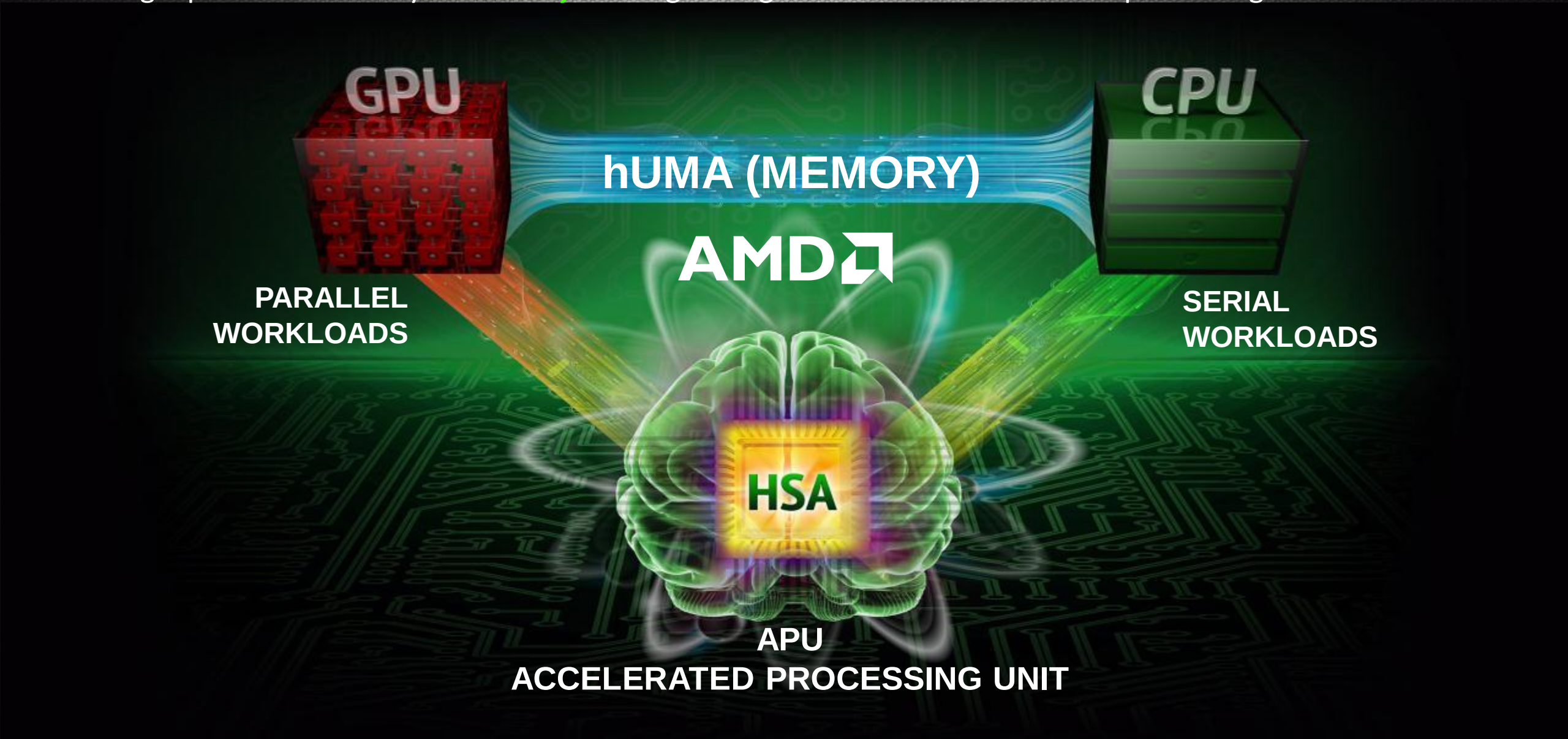
- See slide 24 for details

GFLOPS

Year	CPU	CPU GFLOPS	GPU (RADEON)	GPU GFLOPS
2002	Pentium 4 (Northwood)	12.24	9700 Pro	31.2
2003	Pentium 4 (Northwood)	12.8	9800 XT	36.48
2004	Pentium 4 (Prescott	15.2	X850 XT	103.68
2005		15.2	X1800 XT	134.4
2006	Core 2 Duo	23.44	X1950	375
2007	Core 2 Quad	48	HD 2900 XT	473.6
2008	Q9650	96	HD 4870	1200
2009	Core i7 960	102.4	HD 5870	2720
2010	Core i7 970	153.6	HD 6970	2703
2011	Core i7 3960X	316.8	HD7970	3789
2012	Core i7 3970X	336	HD 7970 GHz Edition	4301

WHAT IS NEXT CPU AND GPU FUSION – HAS – A REAL HETEROGENEOUS ARCH

An ***intelligent computing architecture*** that enables CPU, GPU and other processors to work in ***harmony*** on a single piece of silicon by ***seamlessly*** moving the right tasks to the best suited processing element



HSA ARCHITECTURE

Full Programming language support

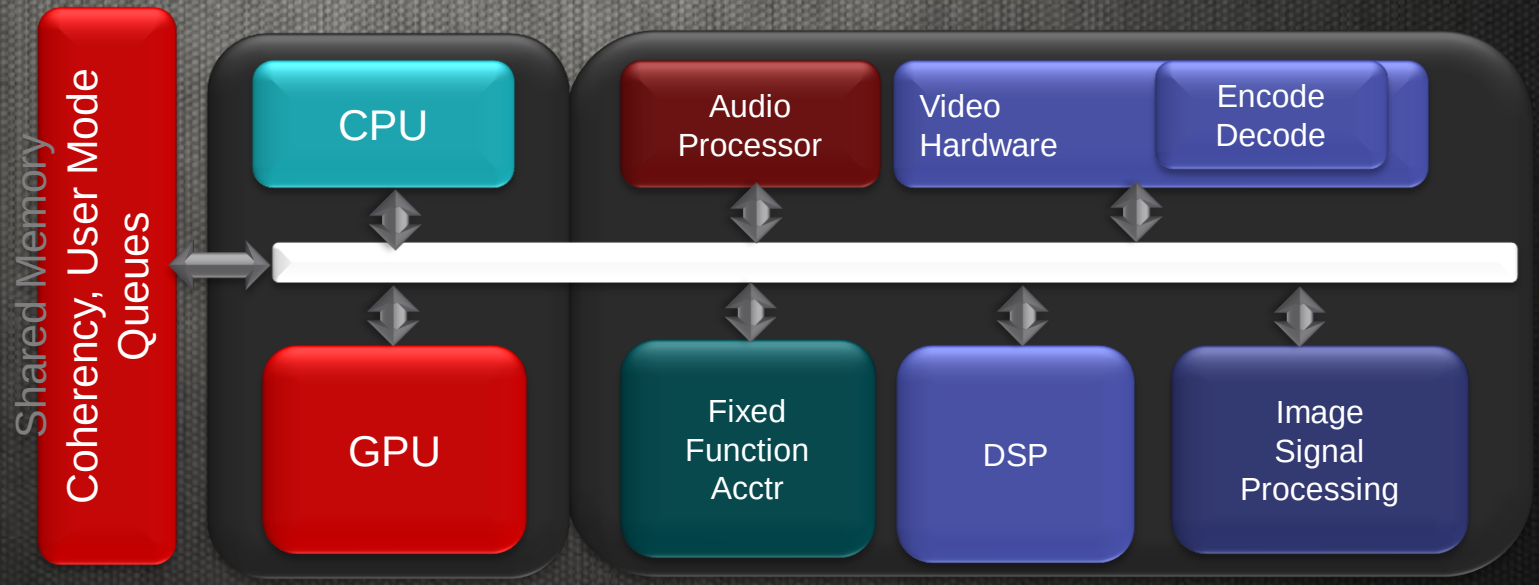
User Mode Queueing

Heterogeneous Unified Memory Access (hUMA)

Pageable Memory

Bidirectional coherency

Compute context switch and preemption



HSA EVOLUTION

AMD

Benefits

Unified power efficiency

Improved compute efficiency

Simplified data sharing

Capabilities

Integrate CPU and GPU in silicon

GPU can access CPU memory

Uniform memory access for CPU and GPU

WHAT IS hUMA?

heterogeneous
UNIFORM
MEMORY
ACCESS

UNDERSTANDING UMA

AMD

Original meaning of UMA is **Uniform Memory Access**

- *Refers to how processing cores in a system view and access memory*
- *All processing cores in a true UMA system share a single memory address space*

Introduction of GPU compute created systems with Non-Uniform Memory Access (NUMA)

- *Require data to be managed across multiple heaps with different address spaces*
- *Add programming complexity due to frequent copies, synchronization, and address translation*

HSA restores the GPU to Uniform memory Access

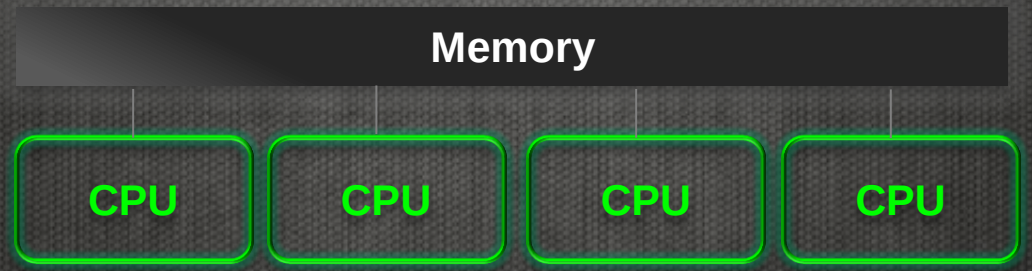
- *Heterogeneous computing replaces GPU Computing*

INTRODUCING hUMA

AMD

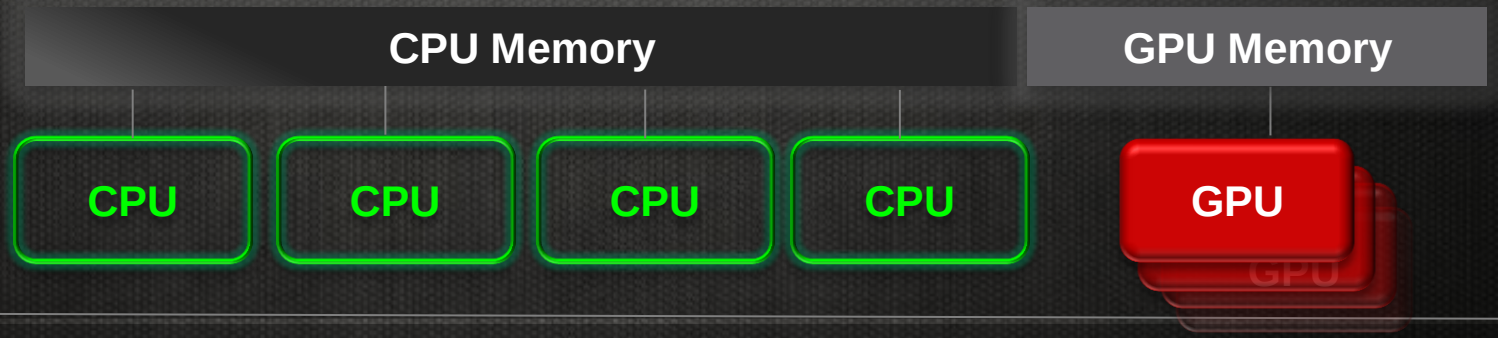
CPU

UMA



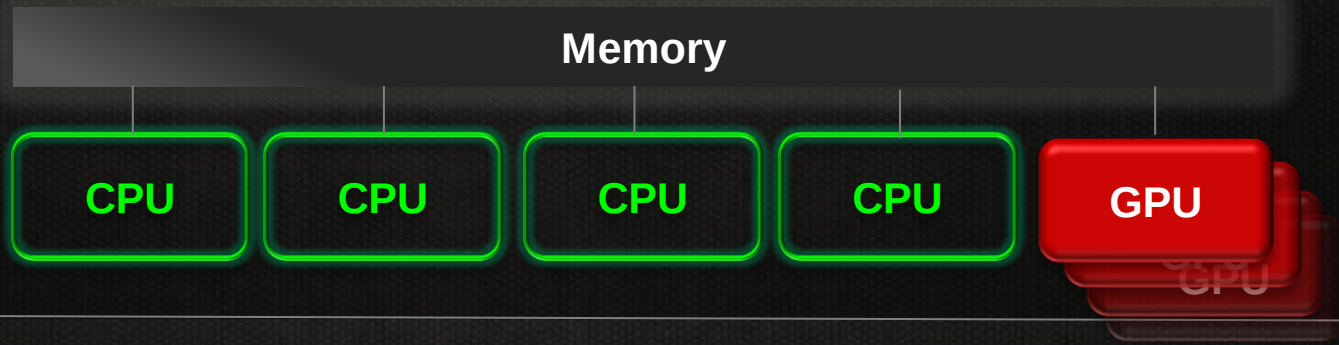
APU

NUMA



APU with HSA

hUMA



hUMA KEY FEATURES

BI-DIRECTIONAL COHERENT MEMORY

Any updates made by one processing element will be seen by all other processing elements - GPU or CPU

PAGEABLE MEMORY

GPU can take page faults, and is no longer restricted to page locked memory

ENTIRE MEMORY SPACE

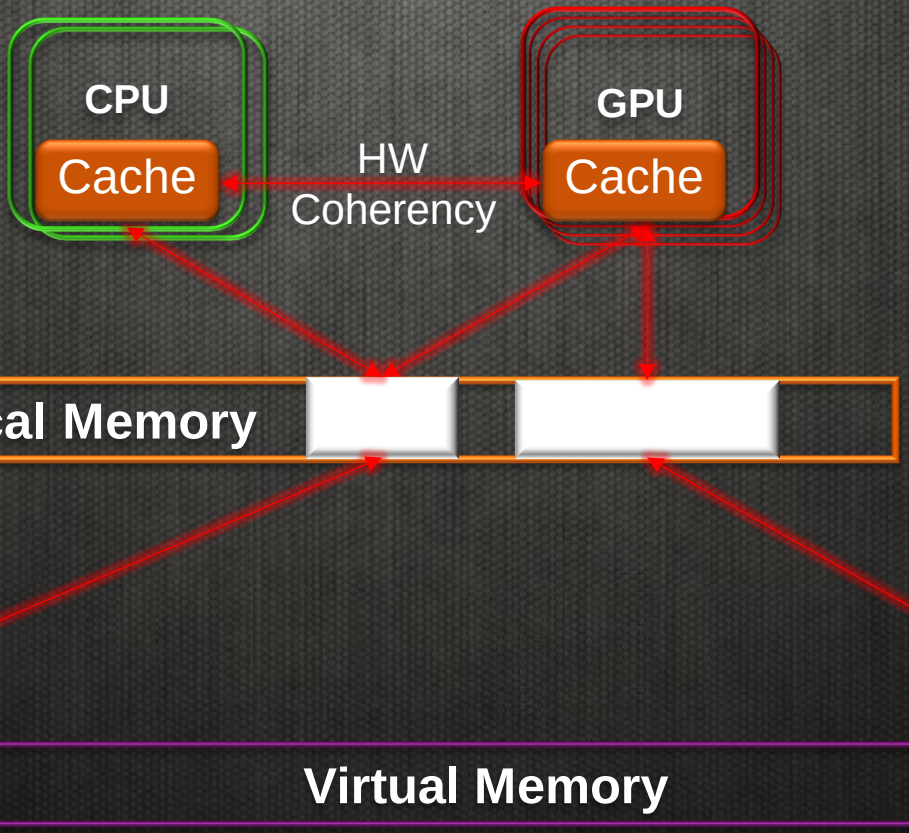
CPU and GPU processes can dynamically allocate memory from the entire memory space

hUMA KEY FEATURES

AMD

Coherent Memory:

Ensures CPU and GPU caches both see an up-to-date view of data



Pageable memory:

The GPU can seamlessly access virtual memory addresses that are not (yet) present in physical memory

Entire memory space:

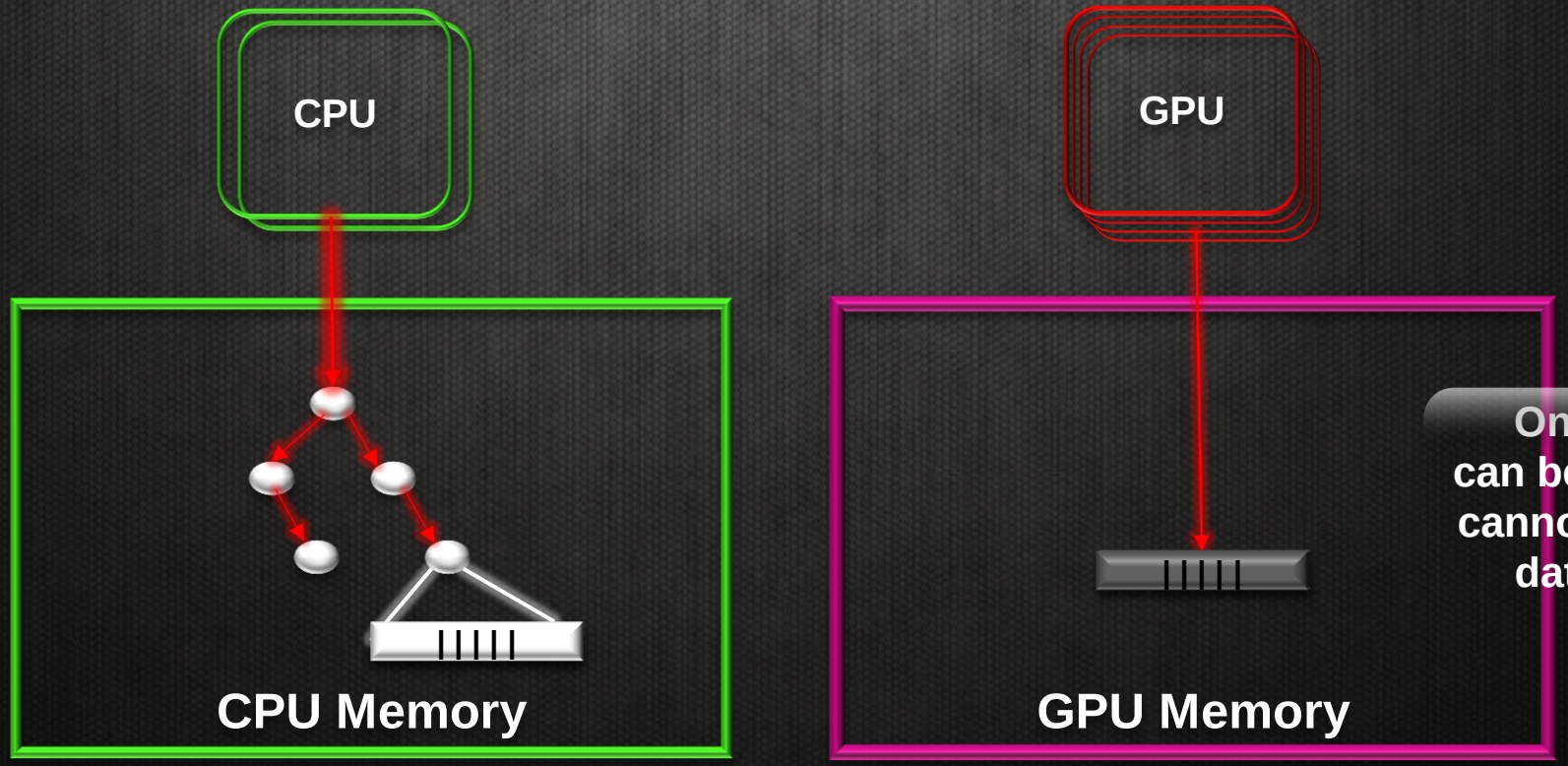
Both CPU and GPU can access and allocate any location in the system's virtual memory space

WITHOUT POINTERS* AND DATA SHARING

AMD

Without hUMA:

- CPU explicitly copies data to GPU memory
- GPU completes computation
- CPU explicitly copies result back to CPU memory



Only the data array can be copied since GPU cannot follow embedded data-structure links

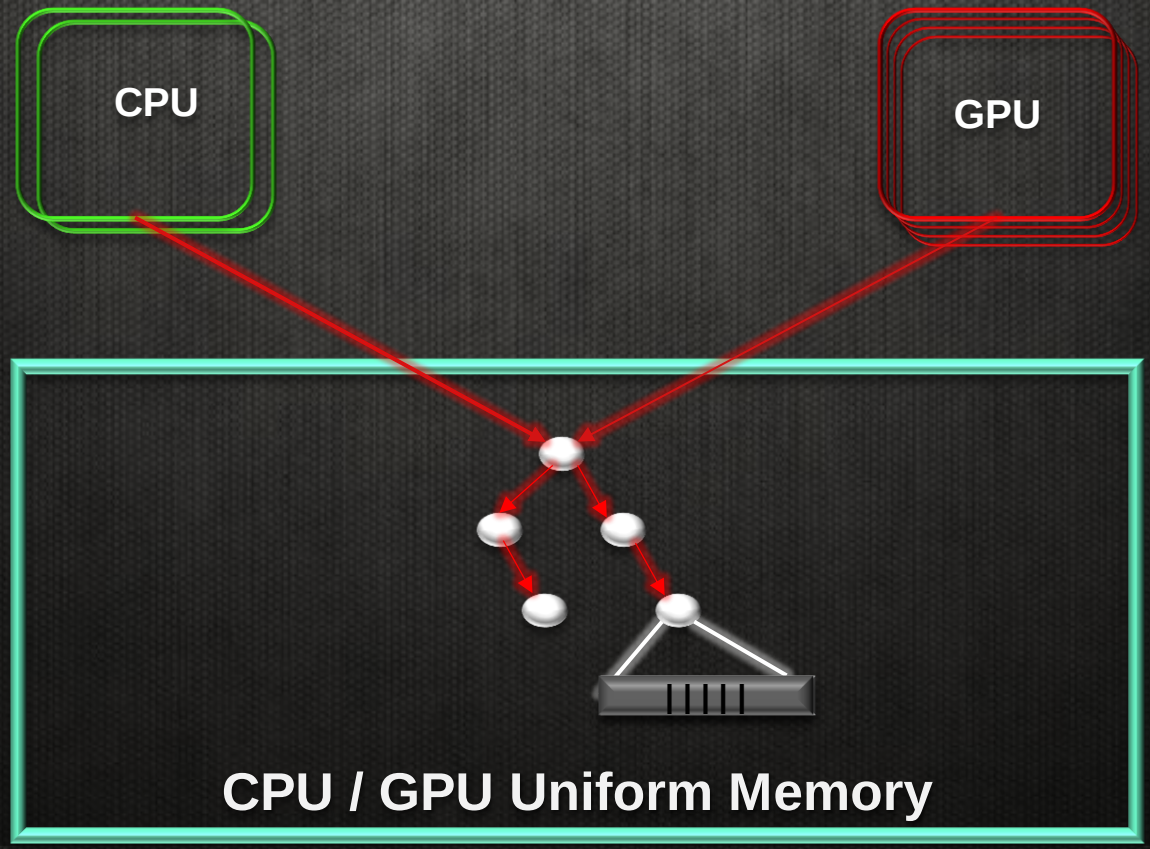
*A Pointer is a named variable that holds a memory address. It makes it easy to reference data or code segments by a name and eliminates the need for the developer to know the actual address in memory. Pointers can be manipulated by the same expressions used to operate on any other variable

WITH POINTERS* AND DATA SHARING

AMD

With hUMA:

- CPU simply passes a pointer to GPU
- GPU completes computation
- CPU can read the result directly – no copying needed!



CPU can pass a pointer to entire data structure since the GPU can now follow embedded links

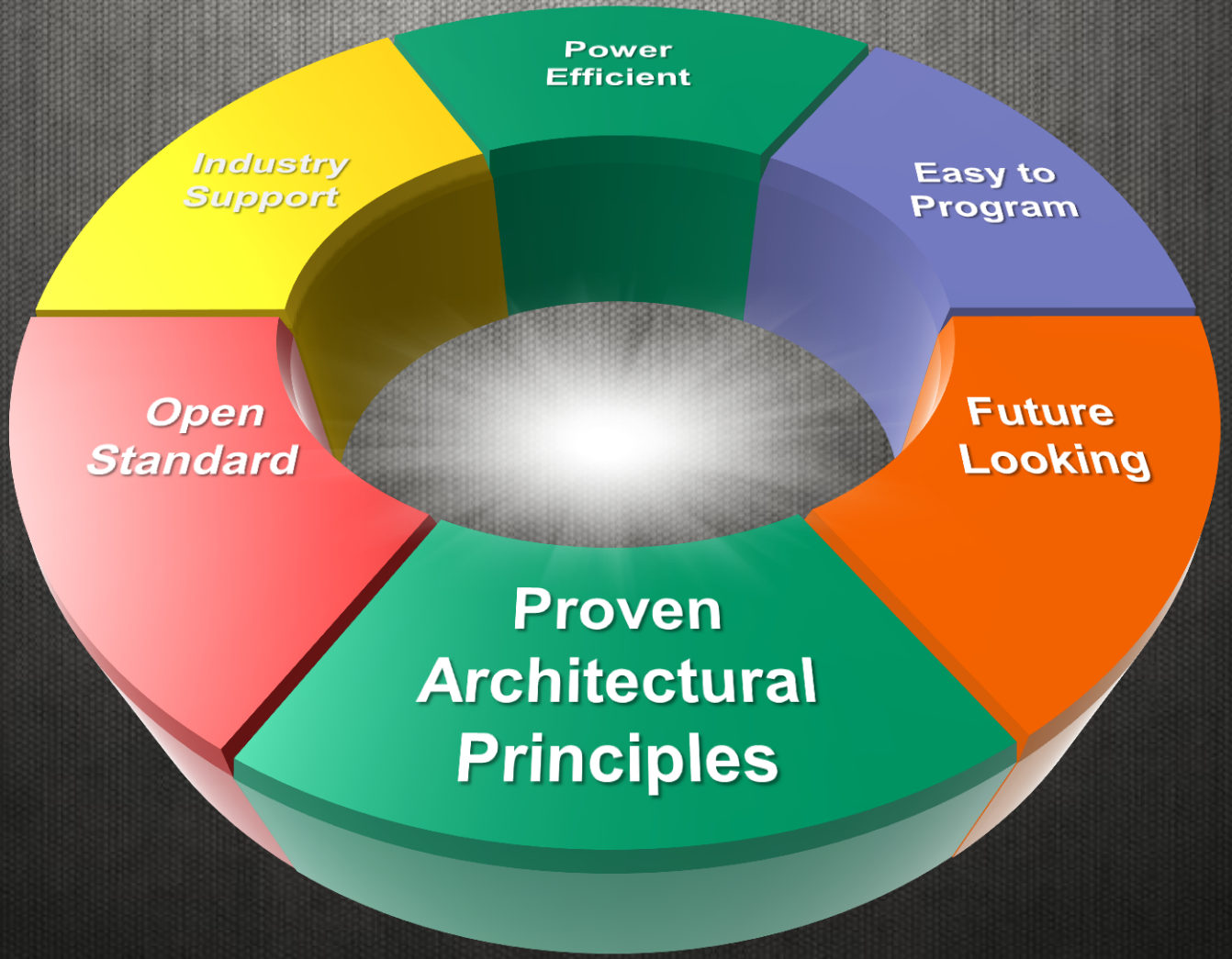
*A Pointer is a named variable that holds a memory address. It makes it easy to reference data or code segments by a name and eliminates the need for the developer to know the actual address in memory. Pointers can be manipulated by the same expressions used to operate on any other variable

hUMA FEATURES

- Access to Entire Memory Space
- Pageable memory
- Bi-directional Coherency
- Fast GPU access to system memory
- Dynamic Memory Allocation

hUMA BENEFITS

BENEFITS OF HSA



UNIFORM MEMORY BENEFITS TO DEVELOPERS

EASE AND SIMPLICITY OF PROGRAMMING

Single, standard computing environments

SUPPORT FOR MAINSTREAM PROGRAMING LANGUAGES

Python, C++, Java

LOWER DEVELOPMENT COST

More efficient architecture enables less people to do the same work

BENEFITS TO CONSUMERS

AMD

BETTER EXPERIENCES
Radically different user experiences

MORE PERFORMANCE
Getting more performance from the same form factor

LONGER BATTERY LIFE
Less power at the same performance

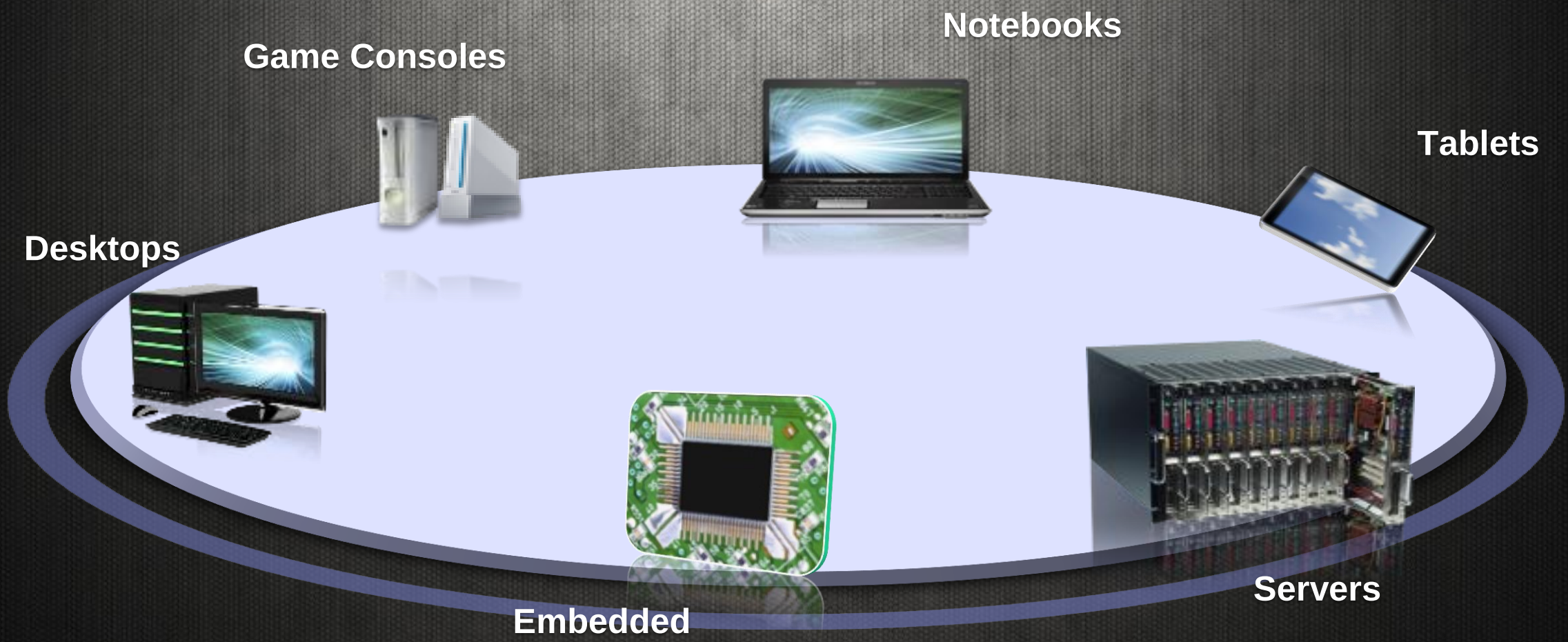
SUPPORT FROM MAJOR INDUSTRY PLAYERS

AMD

AMD
ARM
Imagination
QUALCOMM
MEDIATEK
SAMSUNG
Texas Instruments

- For more information go to: http://hsafoundation.com/
- Source http://pinterest.com/pin/193021534001931884/

POTENTIAL MARKET IS HUGE



AMD | APU13 DEVELOPER SUMMIT

HSA

Nov 11 – 14, 2013
San Jose
McEnery Convention Center

14 Different Tracks with over 140 Individual Presentations


AMD
THANK YOU

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