AMD STRATEGY IN EXASCALE SUPERCOMPUTING AND MACHINE INTELLIGENCE

TIMOUR PALTASHEV, D.SC.
SEPTEMBER 20, 2017

AGENDA

- Exascale Goals and Challenges
- AMD’s Vision and Technologies for Exascale Computing
- HPC Progress Towards Machine Intelligence
- Radeon Instinct and Radeon Open Compute (ROC) Initiatives
- AMD Radeon Instinct Accelerators and Naples server SoC for HPC and Machine Intelligence

DEPARTMENT OF ENERGY'S GOALS FOR EXASCALE COMPUTING SYSTEMS

- The Department of Energy (DOE) plans to deliver exascale supercomputers that provide a 50x improvement in application performance over their current highest-performance supercomputers by 2023
- System should provide a 50x performance improvement over today's fastest supercomputes with 20 MWatts of power while not requiring human intervention due to hardware or system faults more than once a week on average
- Important goals for exascale computing include
  - Enabling new engineering capabilities and scientific discoveries
  - Continuing U.S. leadership in science and engineering

https://asc.llnl.gov/pathforward/

http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf

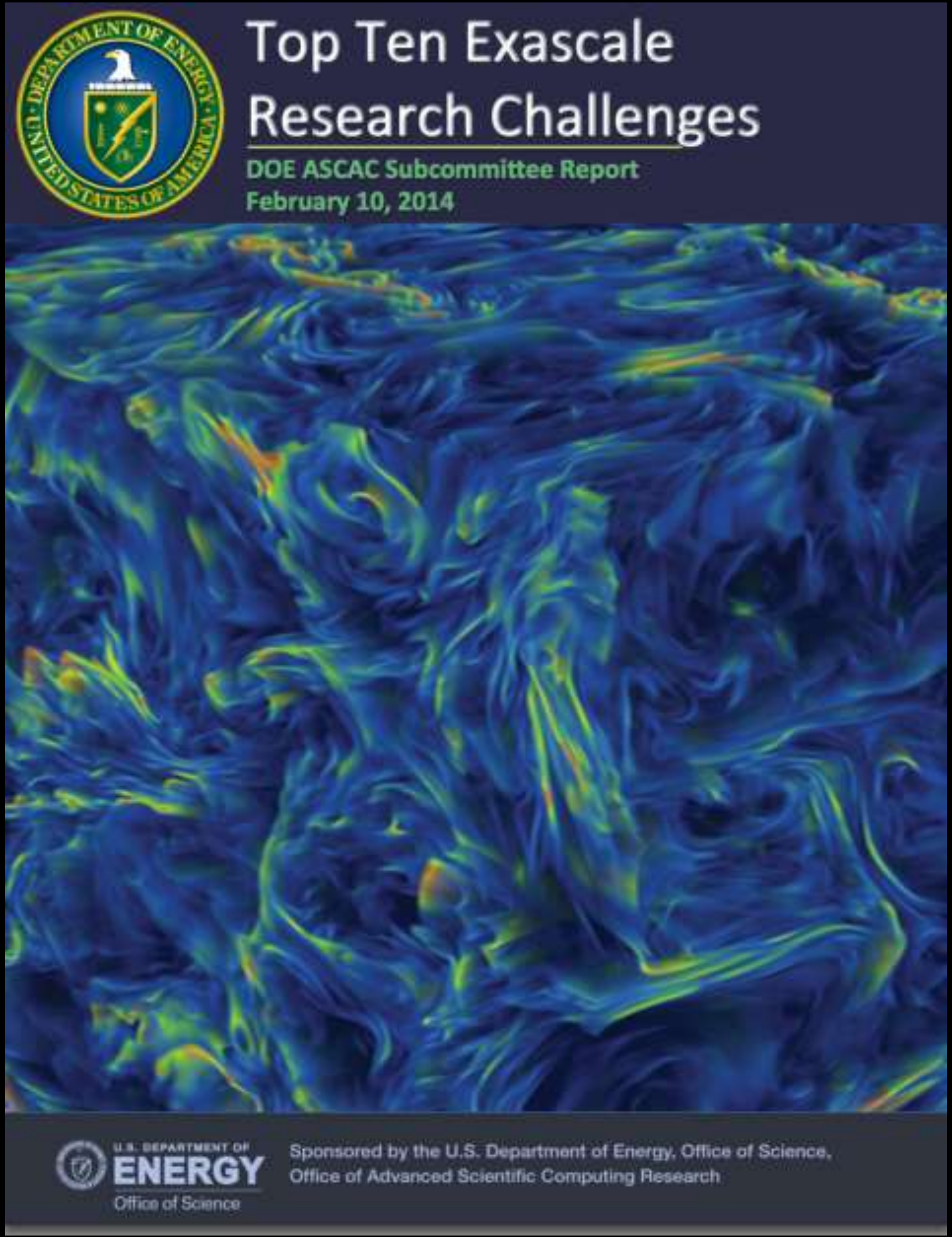
EXASCALE CHALLENGES

The Top Ten Exascale Research Challenges

1) Energy efficiency
2) Interconnect technology
3) Memory technology
4) Scalable system software
5) Programming systems
6) Data management
7) Exascale algorithms
8) Algorithms for discovery, design, and decision
9) Resilience and correctness
10) Scientific productivity

http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf

Requires significant advances in processors, memory, software, and system design



Top Ten Exascale Research Challenges

DOE ASCAC Subcommittee Report
February 10, 2014

U.S. DEPARTMENT OF ENERGY
Office of Science

Sponsored by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research

DOE EXASCALE TARGET REQUIREMENTS

- The DOE has aggressive goals and target requirements for exascale systems
  - Requires research and innovation in a variety of areas
- One of the most important goals is providing supercomputers that can be effectively utilized for important scientific discoveries
- Technologies explored for exascale can be applied to a wide variety of computing systems

Target Requirements	Target Value
System-Level Power Efficiency	50 GFLOPS/Watt
Compute Performance (per node)	10 TFLOPS
Memory Capacity (per node)	5TB
Memory Data Rate (per node)	4 TB/sec
Message per Second (per node)	500 million (MPI), 2 billion (PGAS)
Mean Time to Application Failure	7 days

- http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf

AMD'S VISION FOR SUPERCOMPUTING

EMBRACING HETEROGENEITY

CHAMPIONING OPEN SOLUTIONS

ENABLING LEADERSHIP SYSTEMS

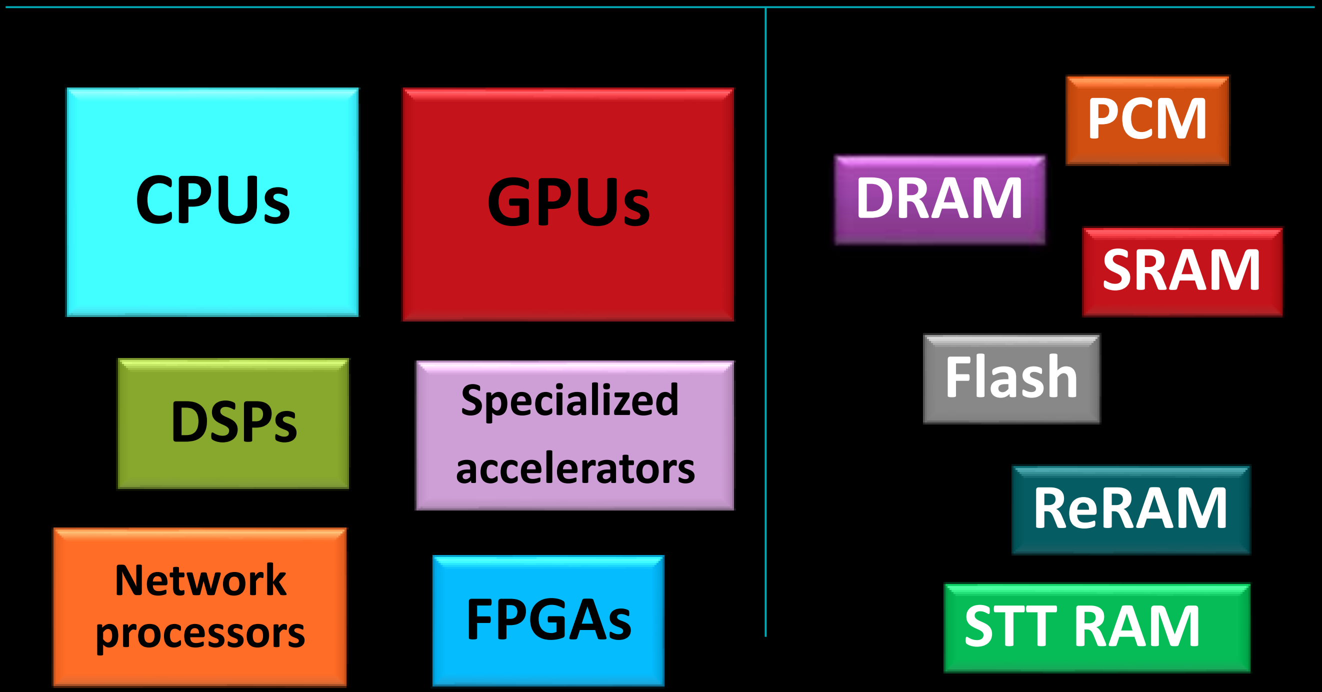
EMBRACING HETEROGENEITY

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

C/C++ FORTRAN Java

UPC/UPC++ python MPI

Kokkos/RAJA OpenMP OpenACC



CPUs

GPUs

DSPs

Specialized accelerators

Network processors

FPGAs

PCM

DRAM

SRAM

Flash

ReRAM

STT RAM

Heterogeneity Options

CHAMPIONING OPEN SOLUTIONS

AMD

- Harness the creativity and productivity of the entire industry
- Partner with best-in-class suppliers to enable leading solutions

- Multiple paths to open solutions
  - **Open standards**
  - **Open-source software**
  - **Open collaborations**

HSA™ FOUNDATION

ROCM ON torch

OpenMP®

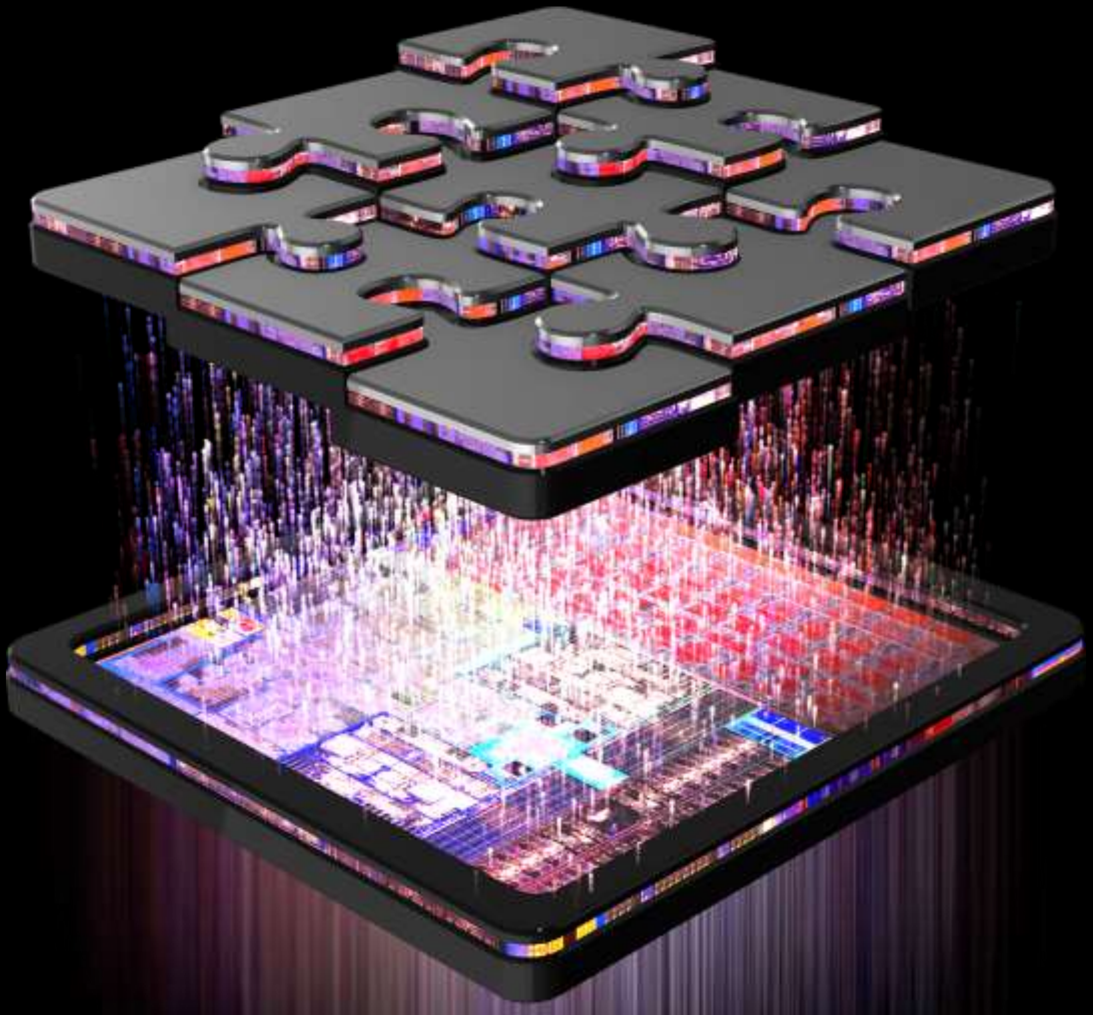
JEDEC

OPEN Compute Project

RADEON TECHNOLOGIES GROUP

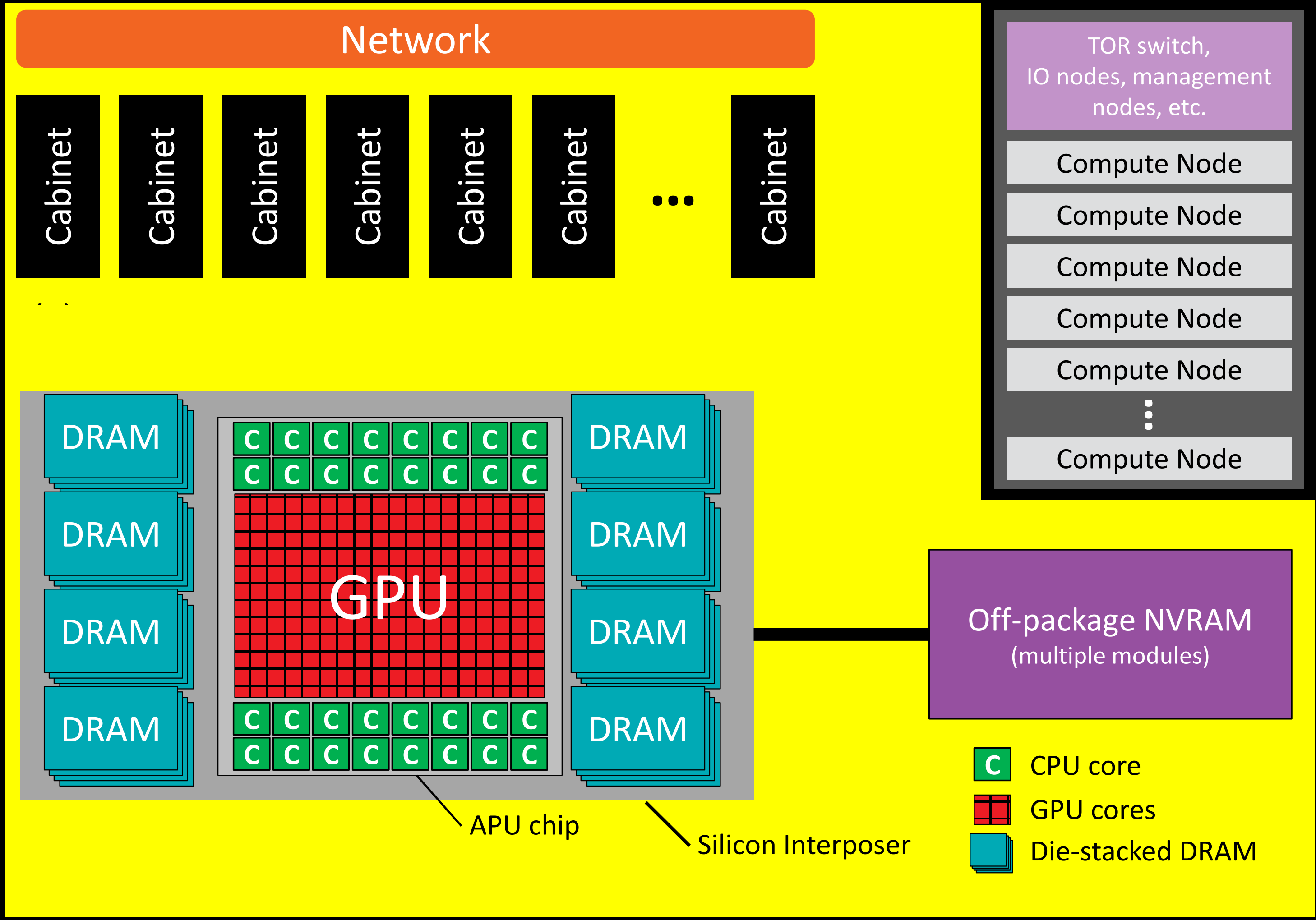
ENABLING LEADERSHIP SYSTEMS

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



FUTURE HIGH DENSITY COMPUTE CONFIGURATIONS

- Exascale systems require enhanced performance, power-efficiency, reliability, and programmer productivity
  - Significant advances are needed in multiple areas and technologies
- Exascale systems will be heterogeneous
  - Programming environments and runtimes should manage this heterogeneity
- New computing technologies provide a path to productive, power-efficient exascale systems



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

COMPUTING PROGRESS: MAINFRAMES

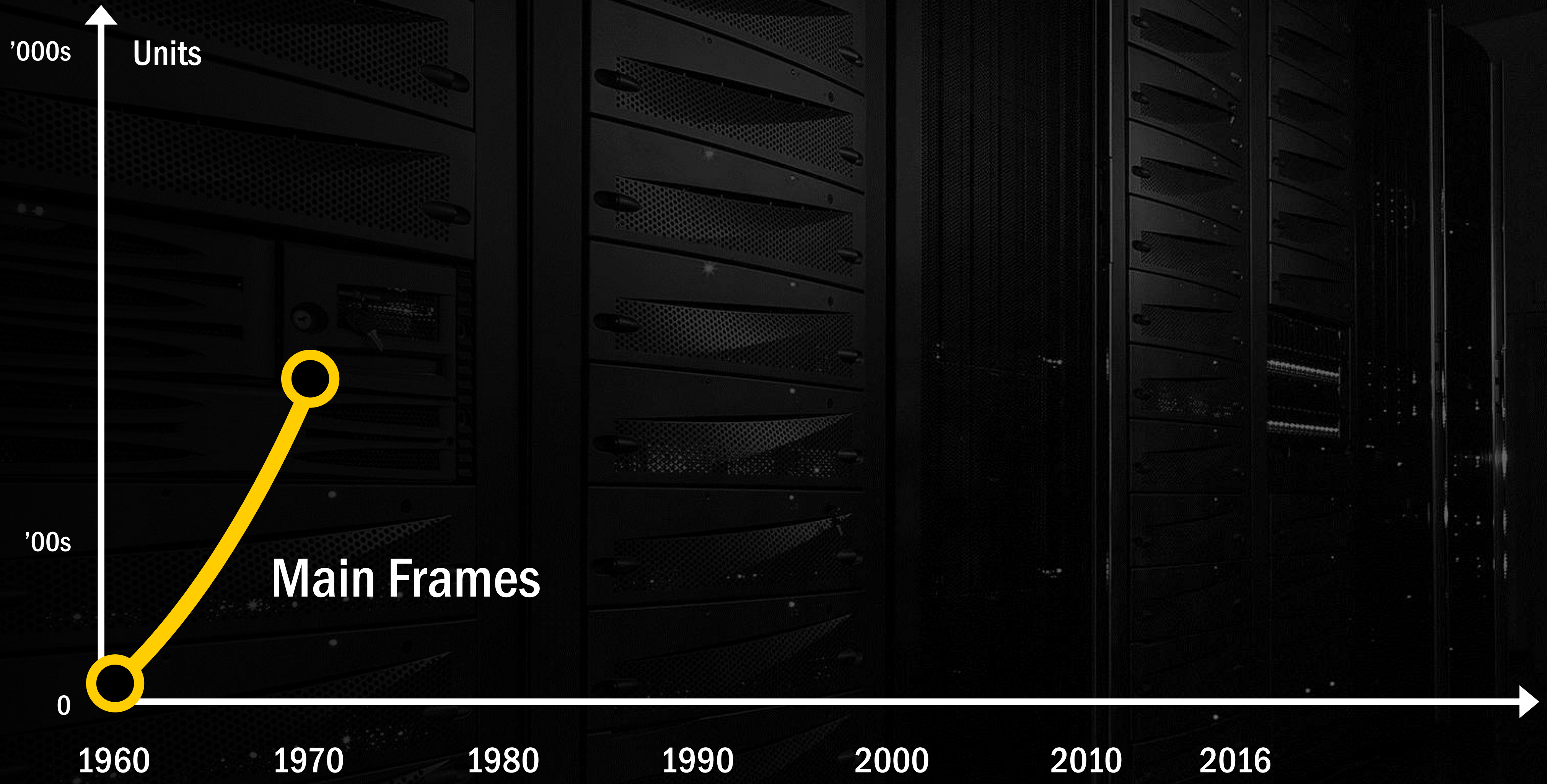


Chart for illustrative purposes

AMD

COMPUTING PROGRESS: CLIENT-SERVER

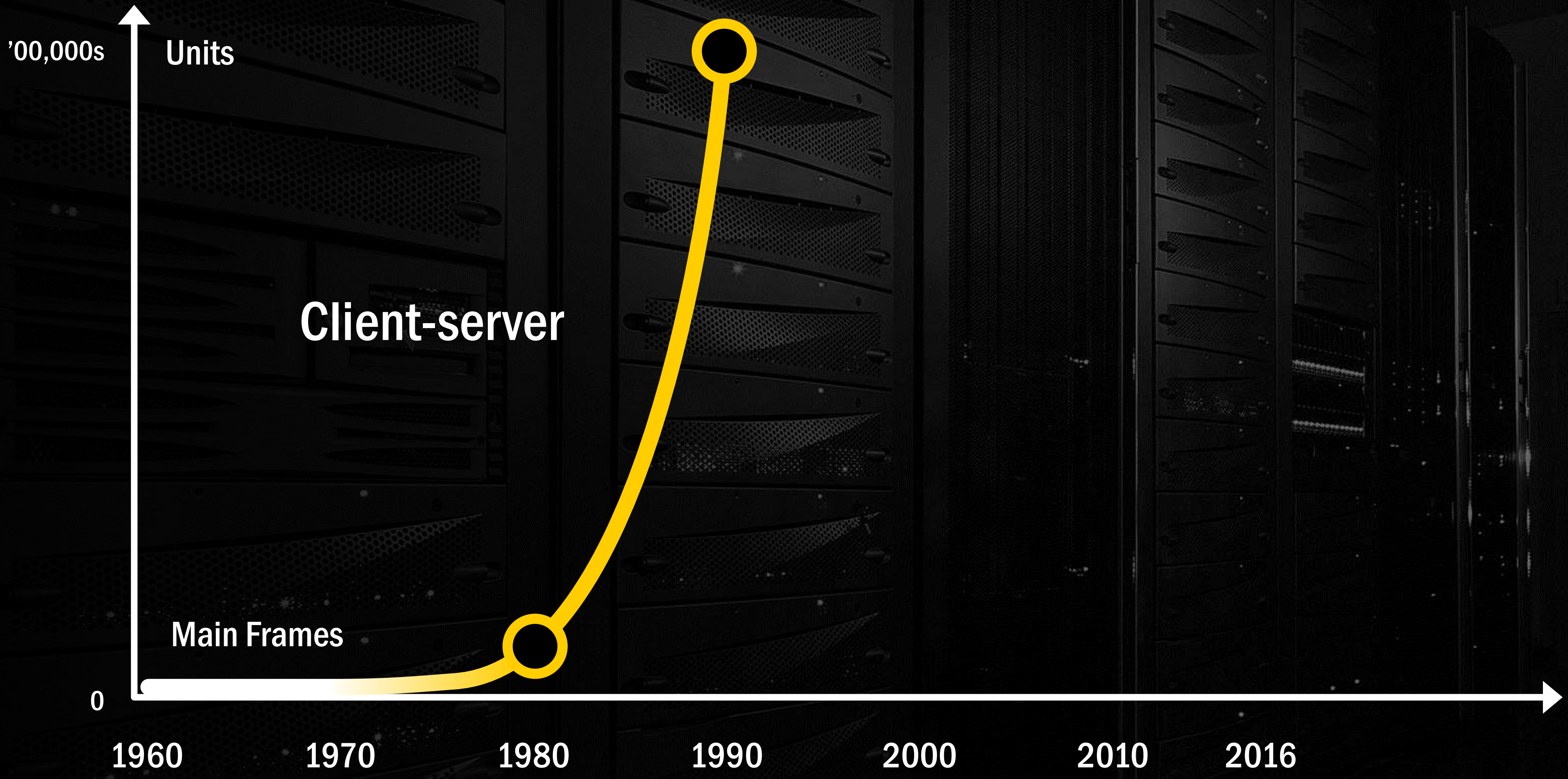


Chart for illustrative purposes

COMPUTING PROGRESS: CLOUD COMPUTING

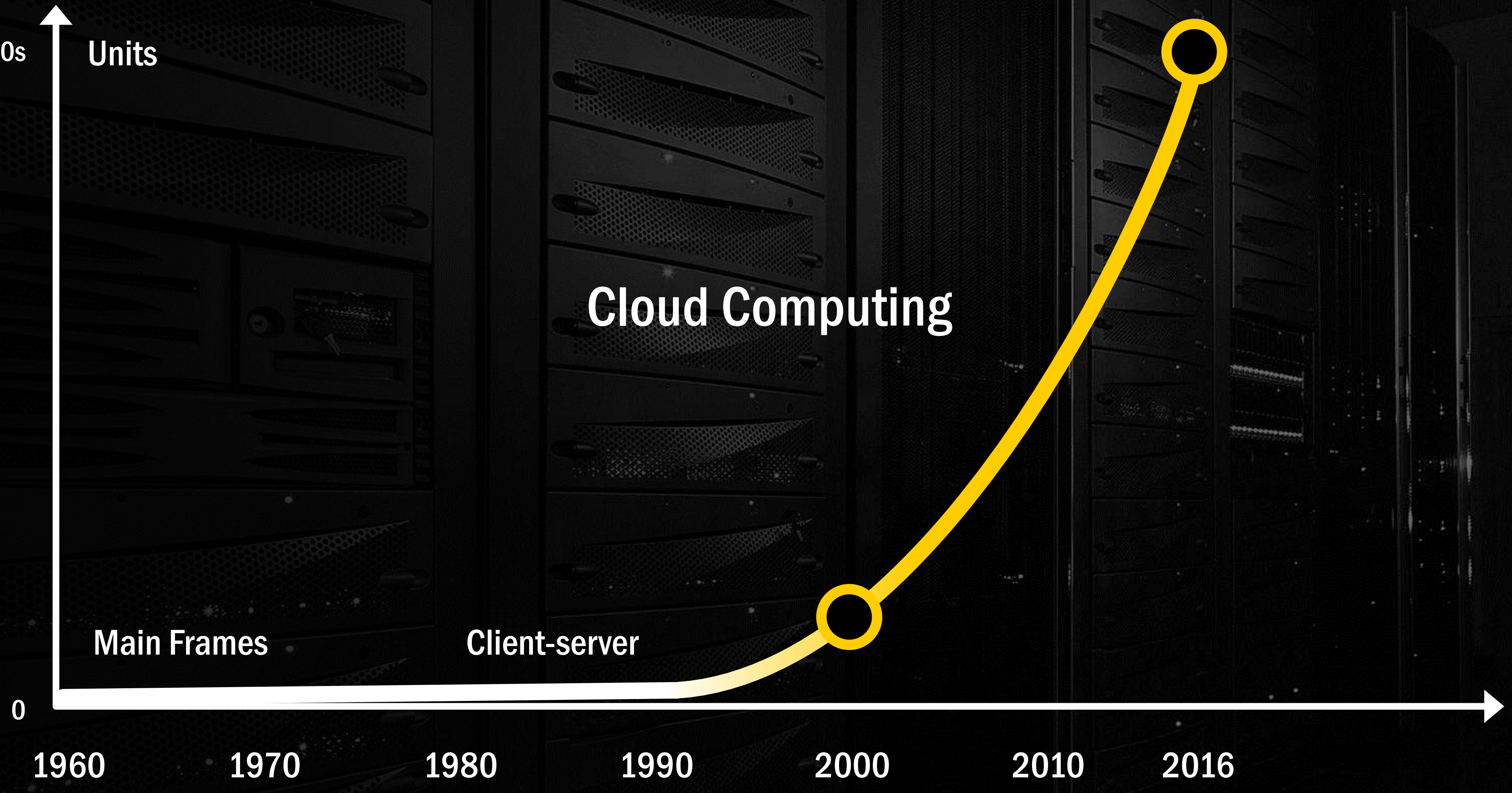


Chart for illustrative purposes

COMPUTING PROGRESS: MACHINE INTELLIGENCE ERA

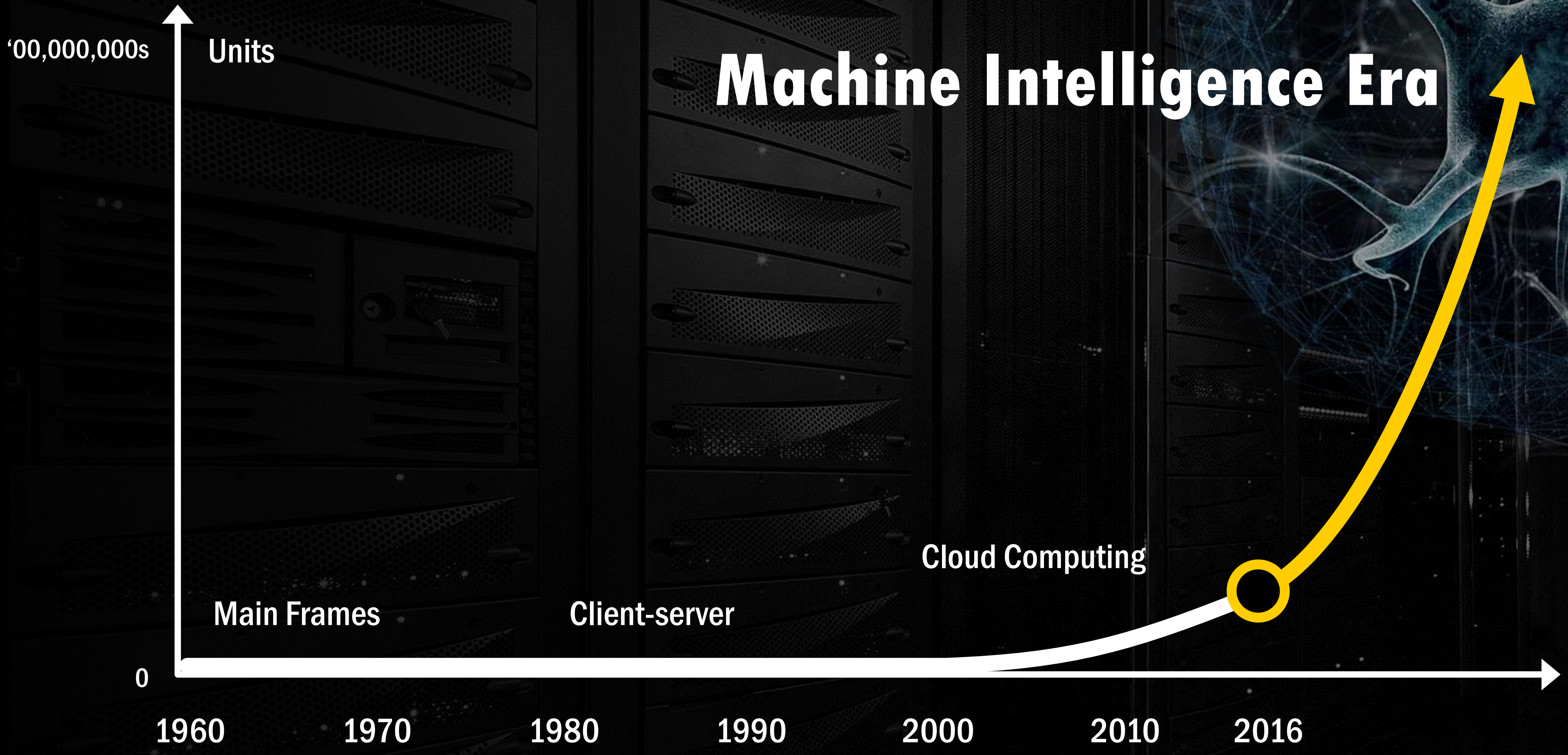


Chart for illustrative purposes

2.5 Quintillion Bytes of Data is Generated Every Day

500 million Tweets

4 million
hours of content
on Youtube

4.3 billion
Facebook entry

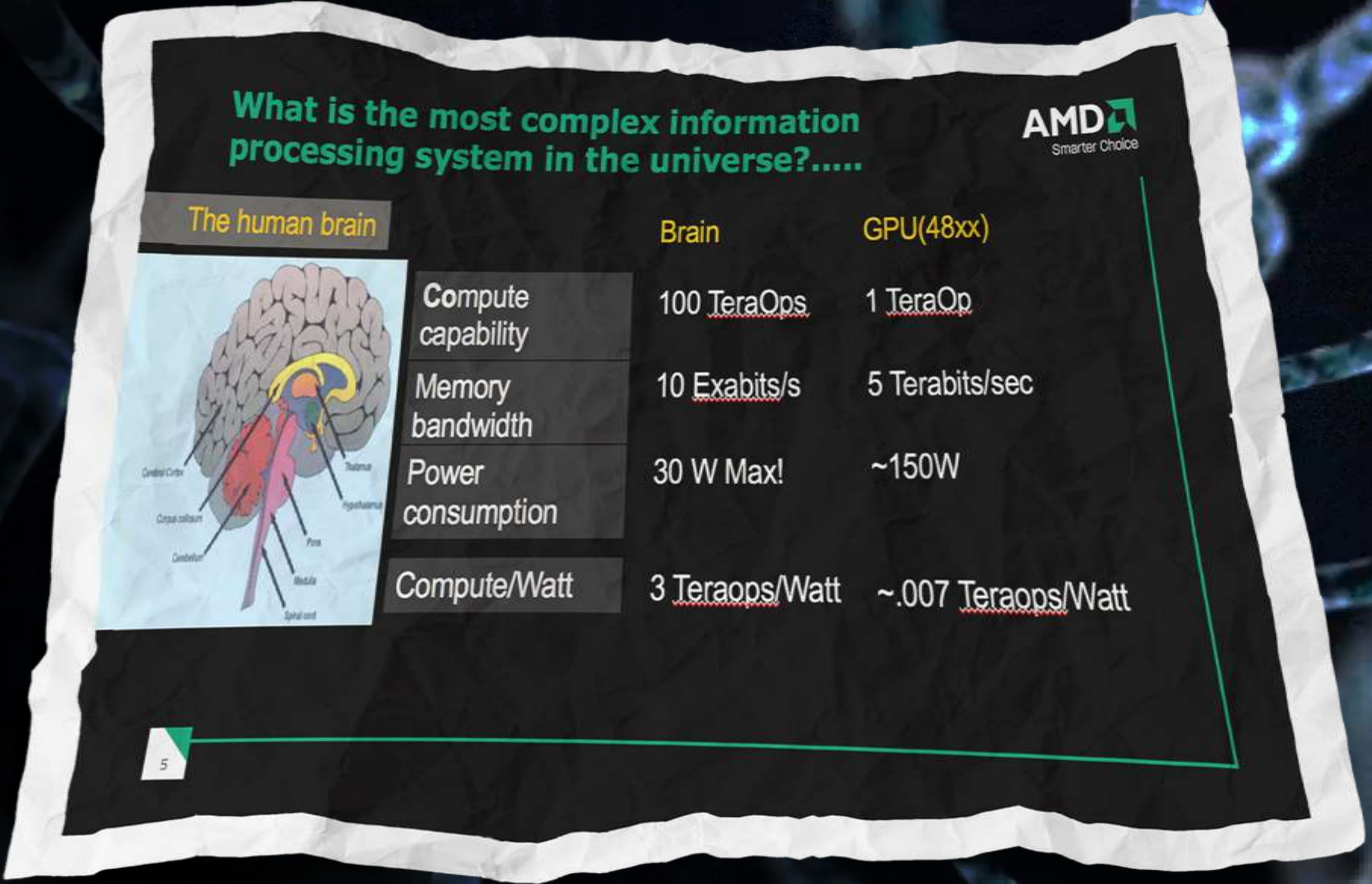
3.6 billion
Instagram

6 billion
Google searches

205 billion
emails

and many more...

Human Brain in your Hand



What is the most complex information processing system in the universe?.....

AMD Smarter Choice

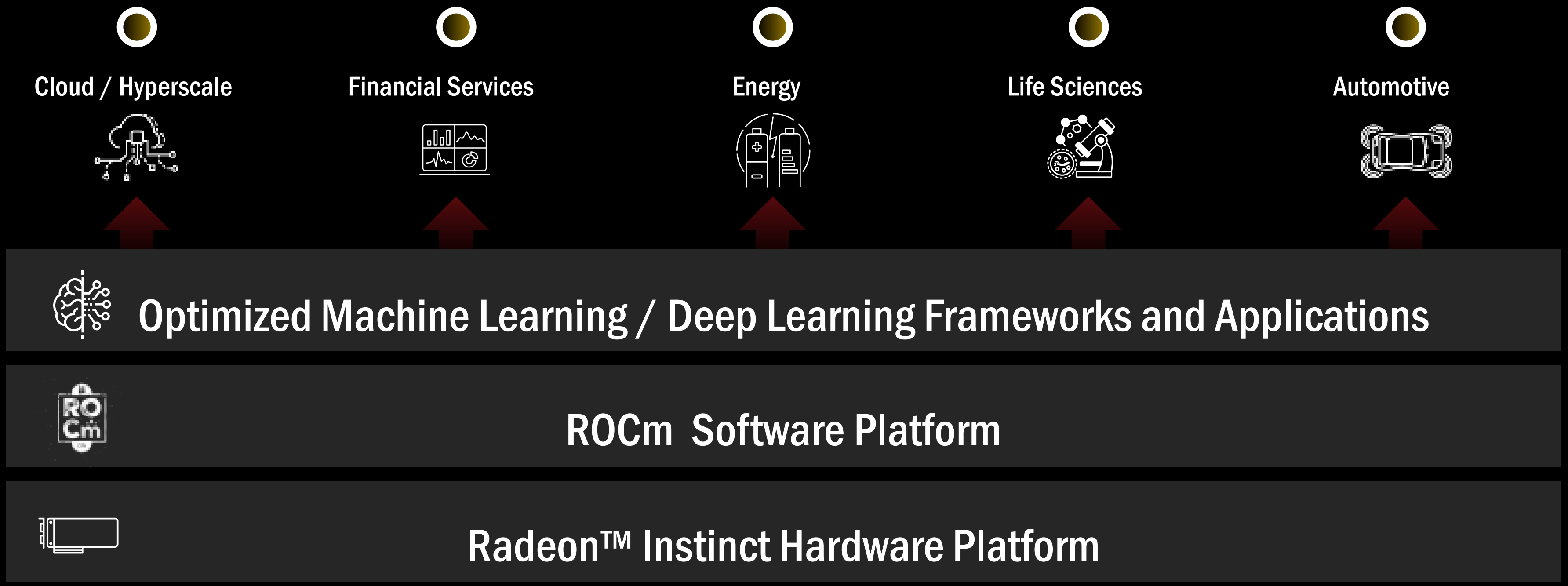
The human brain

Cerebral Cortex
Corpus callosum
Cerebellum
Thalamus
Hypothalamus
Pons
Medulla
Spinal cord

	Brain	GPU(48xx)
Compute capability	100 TeraOps	1 TeraOp
Memory bandwidth	10 Exabits/s	5 Terabits/sec
Power consumption	30 W Max!	~150W
Compute/Watt	3 Teraops/Watt	~.007 Teraops/Watt

RADEON INSTINCT

Radeon Instinct Initiative



Cloud / Hyperscale

Financial Services

Energy

Life Sciences

Automotive

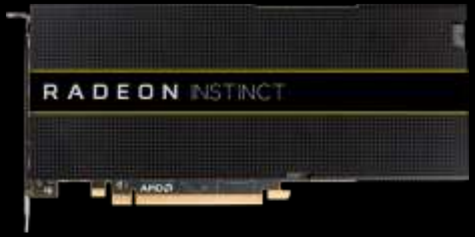
Optimized Machine Learning / Deep Learning Frameworks and Applications

ROCm Software Platform

Radeon™ Instinct Hardware Platform

Address market verticals that use a common infrastructure to leverage the investments and scale fast across multiple industries

Accelerators | RADEON INSTINCT



MI6

Passively Cooled Inference Accelerator

5.70 TFLOPS

224 GB/s Memory Bandwidth

<150W

MI8

Small Form Factor Accelerator

8.2 TFLOPS

512 GB/s Memory Bandwidth

<175W

MI25 Vega with NCU

Passively cooled Training Accelerator

2X Packed Math

High Bandwidth Cache
and Controller

<300W

ROCm PROGRAMMING MODEL OPTIONS

HIP

Convert CUDA to portable C++

- Single-source Host+Kernel
- C++ Kernel Language
- C Runtime
- Platforms: AMD GPU, NVIDIA (Designed to have the same or better perf as native CUDA)

When to use it?

- Port existing CUDA code
- Developers familiar with CUDA
- New project that needs portability to AMD and NVIDIA

HCC

True single-source C++ accelerator language

- Single-source Host+Kernel
- C++ Kernel Language
- C++ Runtime
- Platforms: AMD GPU

When to use it?

- New projects where true C++ language preferred
- Use features from latest ISO C++ standards

OpenCL

Khronos Industry Standard accelerator language

- Split Host/Kernel
- C99-based Kernel Language
- C Runtime
- Platforms: CPU, GPU, FPGA

When to use it?

- Port existing OpenCL code
- New project that needs portability to CPU,GPU,FPGA

INTRODUCING ROCm SOFTWARE PLATFORM

A new, fully “Open Source” foundation for Hyper Scale and HPC-class GPU computing

Graphics Core Next Headless Linux® 64-bit Driver

- 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

HSA™ FOUNDATION

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
- Full documentation to GCN ISA

“Open Source” Tools and Libraries

- Rich Set of “Open Source” math libraries
- Tuned “Deep Learning” frameworks
- Optimized parallel programing frameworks
- CodeXL profiler and GDB debugging

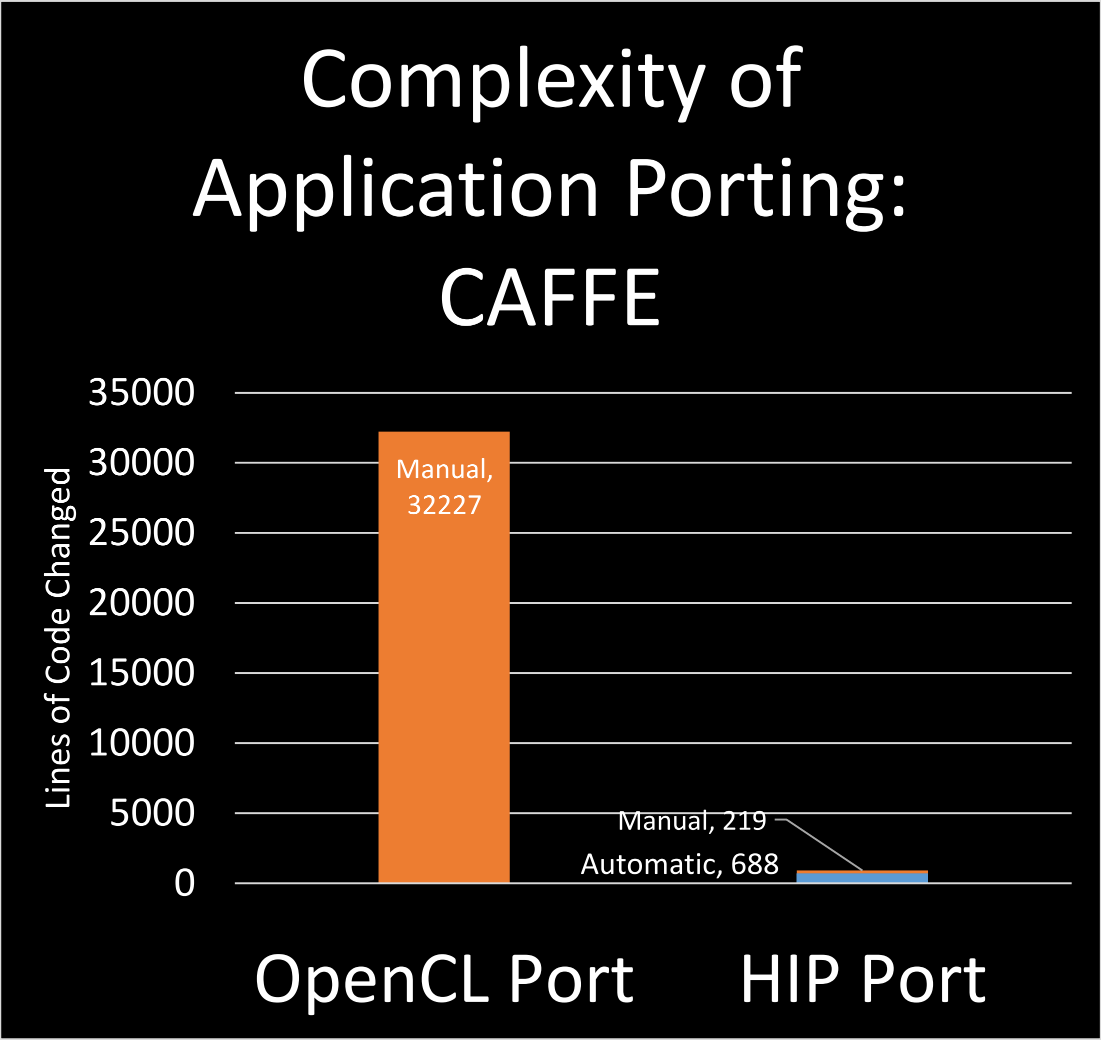
GPUOpen

ROCm : DEEP LEARNING GETS HIP

Bringing a faster path to bring deep learning application to AMD GPUs

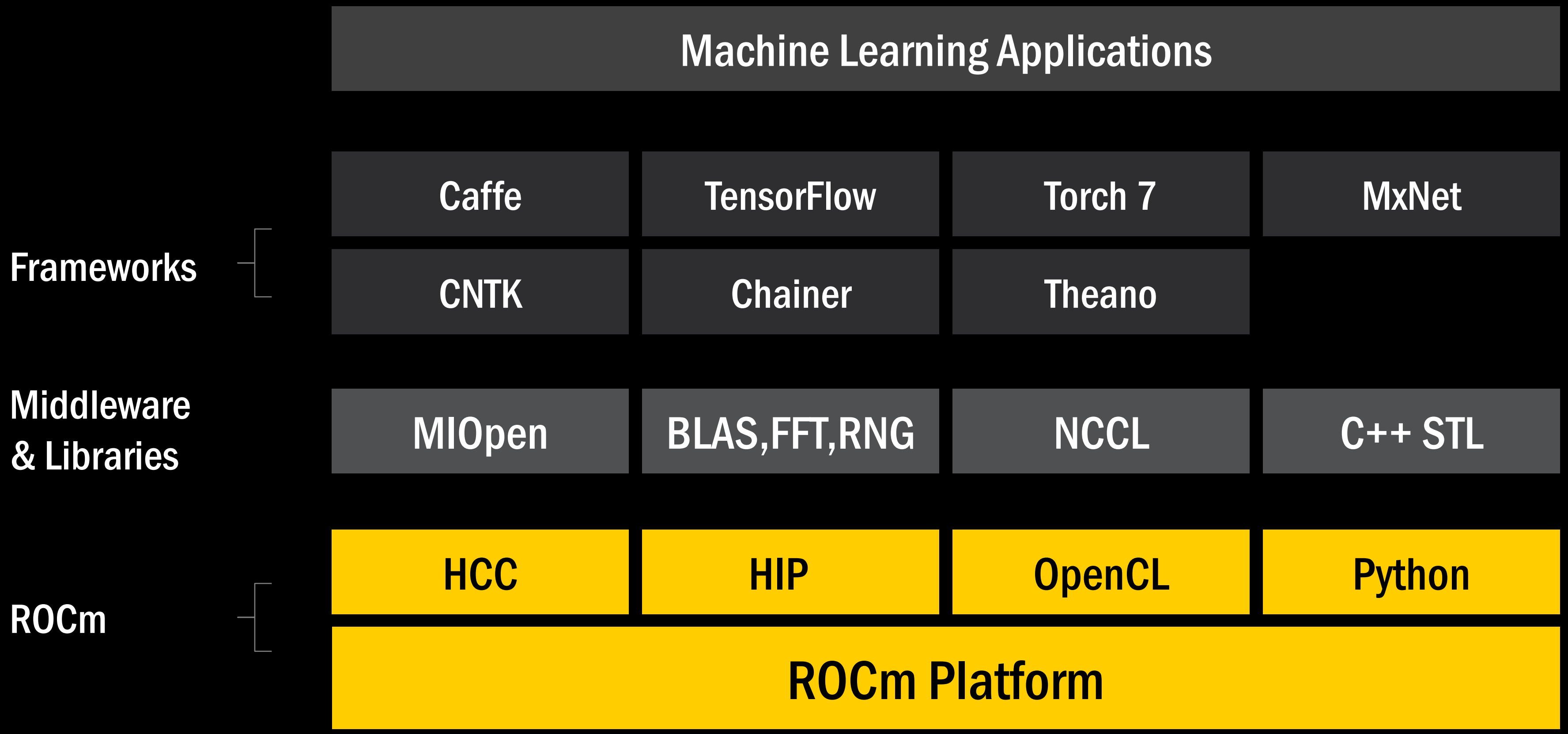
- The Challenge: CAFFE
  - Popular machine-learning framework
  - Tip version on GitHub has 55000+ lines-of-code
  - GPU-accelerated with CUDA

- Results:
  - 99.6% of code unmodified or automatically converted
  - Port required less than 1 week developer time
  - Supports all CAFFE features (multi-gpu, P2P, FFT filters)
  - HIPCAFFE is the fastest CAFFE on AMD hardware – 1.8X faster than CAFFE/OpenCL



AMD Internal Data

ROCm SOFTWARE



Machine Learning Applications

Frameworks

Caffe | TensorFlow | Torch 7 | MxNet

CNTK | Chainer | Theano

Middleware & Libraries

MIOpen | BLAS,FFT,RNG | NCCL | C++ STL

ROCm

HCC | HIP | OpenCL | Python

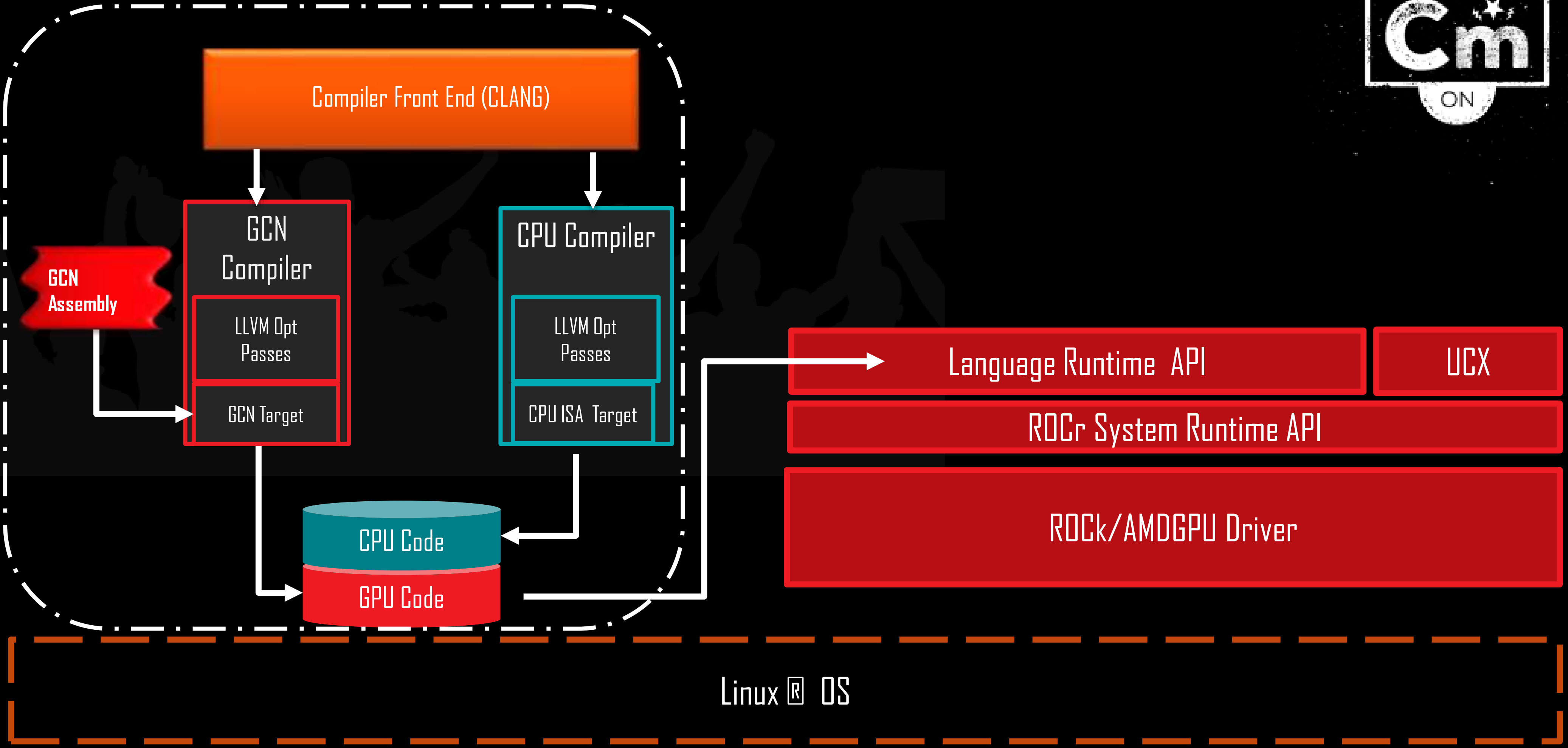
ROCm Platform

DELIVERING AN OPEN PLATFORM FOR GPU COMPUTING

Language neutral solution to match developer needs as heterogeneous programing models evolve

GCN Compiler

- Direct-to-ISA
- GCN Docs
- CLANG/LLVM
- GCN Assembler
- *Open-source*



EXTENDING SUPPORT TO A BROADER HARDWARE ECOSYSTEM

AMD

ROCm “Open Source” foundation brings a rich foundation to these new ecosystems

AMD64 Support
AMD ZEN

Intel Xeon E5 v3 v4

ARM AArch64 Support
Cavium Thunder X

IBM OpenPower Support
- IBM Power 8

POWER8

ROCm is being built to support next generation I/O Interfaces

GenZ Founding Member

GEN Z

CCIX Founding Member

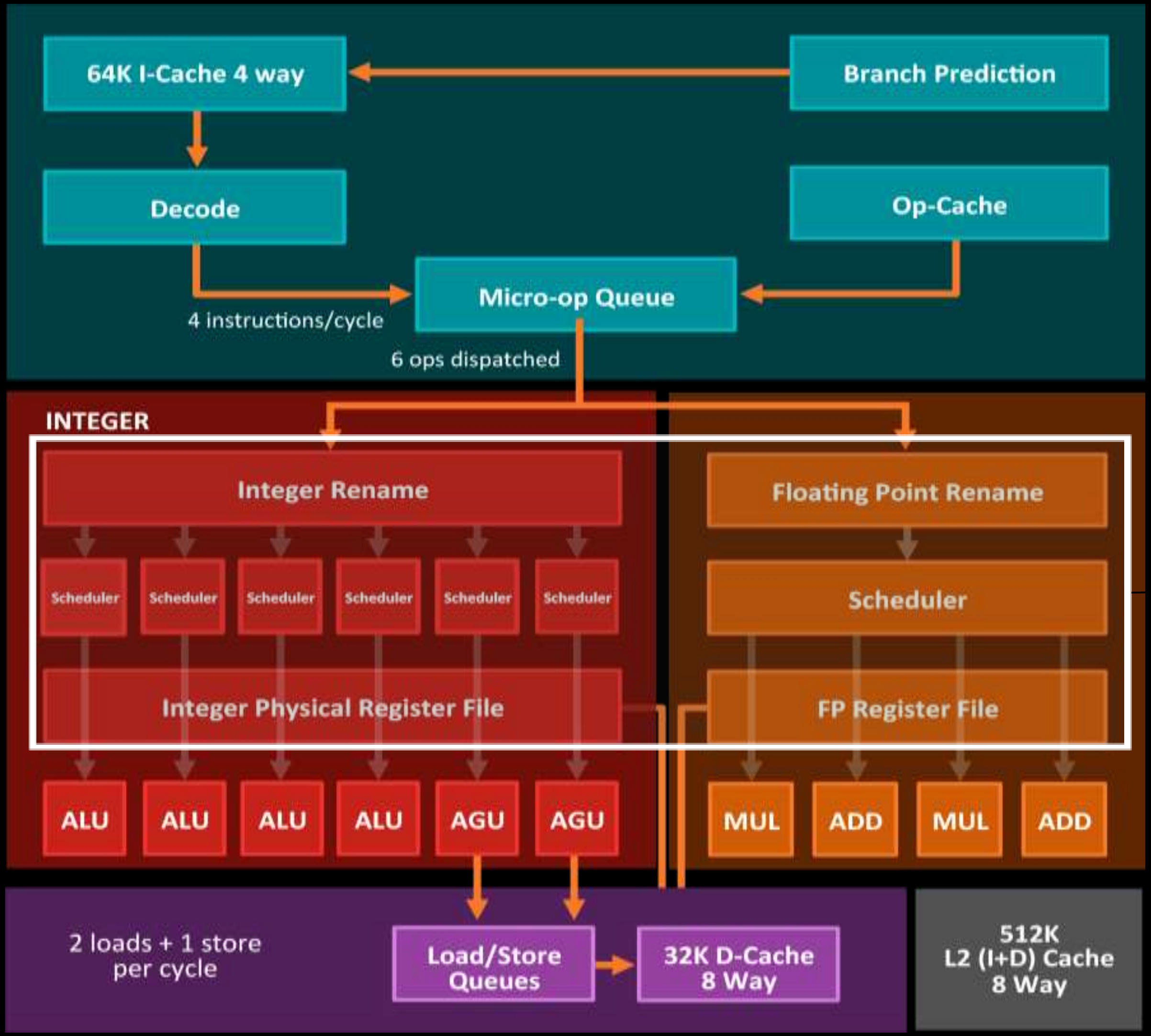
CCIX Cache Coherent Interconnect for Accelerators

OpenCAPI Founding Member

OpenCAPI™

RADEON TECHNOLOGIES GROUP

ZEN CPU CORE: PERFORMANCE AND THROUGHPUT



QUANTUM LEAP IN CORE EXECUTION CAPABILITY

- ❑ Enhanced branch prediction to select the right instructions
- ❑ Micro-op cache for efficient ops issue
- ❑ 1.75X instruction scheduler window*
- ❑ 1.5X issue width and execution resources*

Result: instruction level parallelism designed for dramatic gains in single-threaded performance

*Compared to predecessor

NEW ZEN CPU CORE IN DESKTOPS/WORKSTATIONS

"RYZEN" aka "SUMMIT RIDGE"

- 8 CORES, 16 THREADS
- AM4 Platform
  - DDR4
  - PCI EXPRESS® GEN 3
  - NEXT-GEN I/O

- https://www.amd.com/en/ryzen?&gclid=CL7W9ZyX-tICFUOXfgodGt8BPg

“EPYC” SERVER SOC

AMD

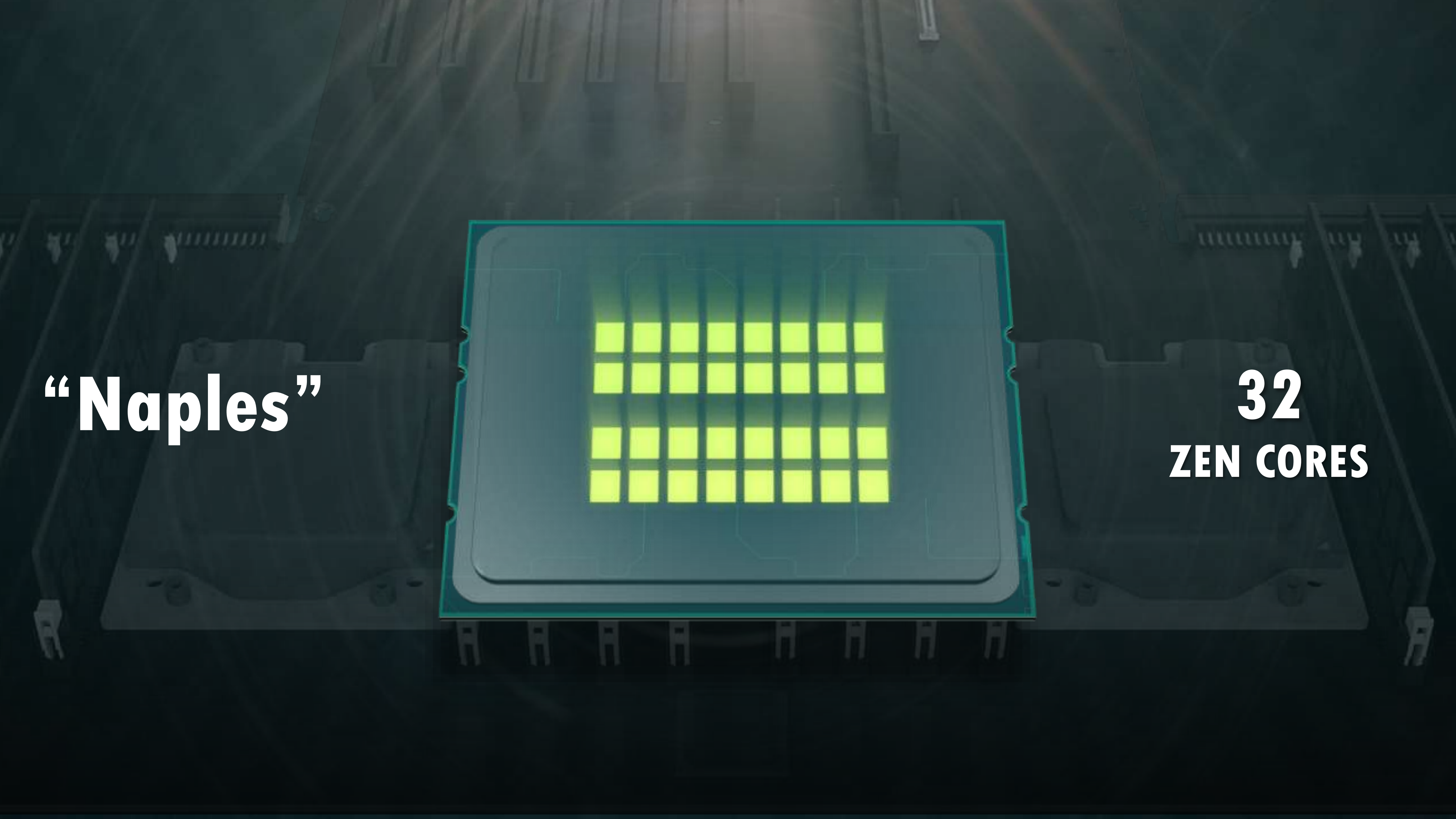
ZEN

32-Core,
64-Thread

1st Public Demo,
August 2016,
San Francisco

Microsoft ubuntu
Windows Server SUSE
Linux vmware
CANONICAL
redhat CITRIX

Enabling Industry
Software Support

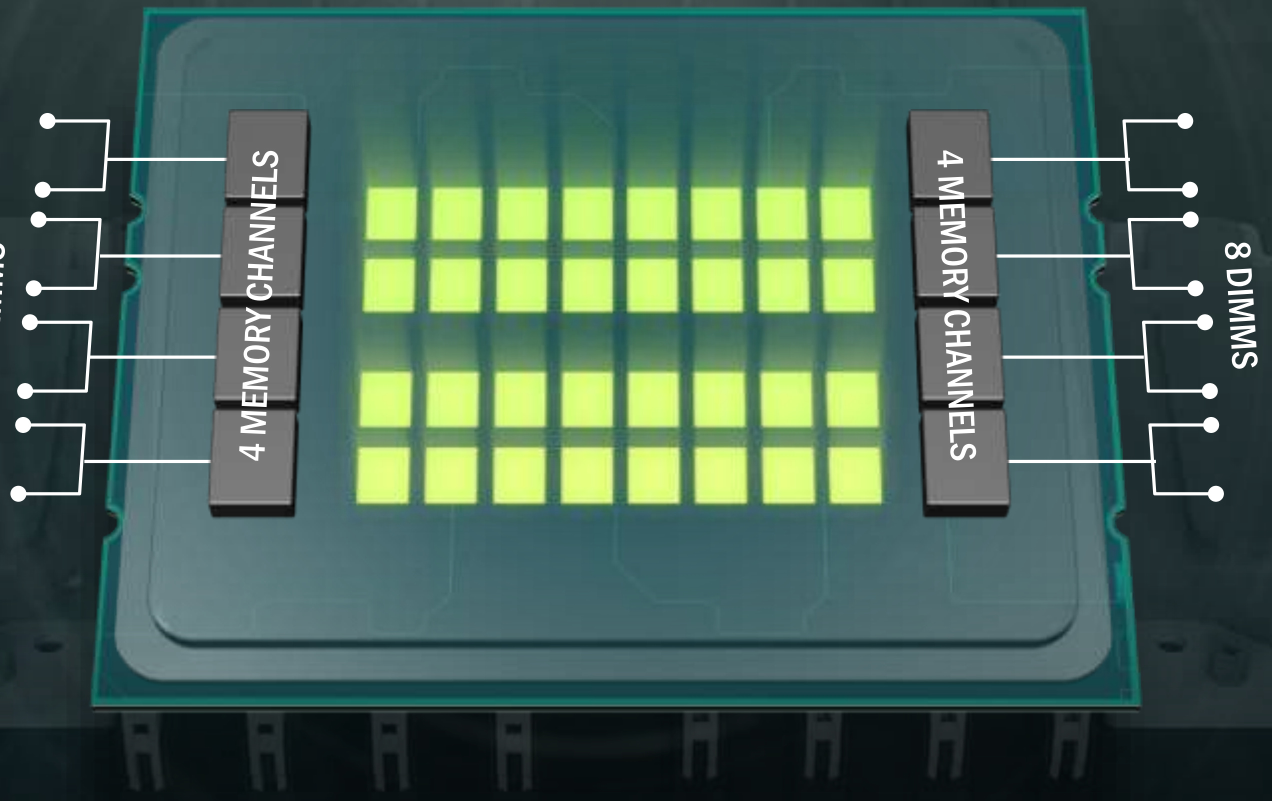


“Naples”

32

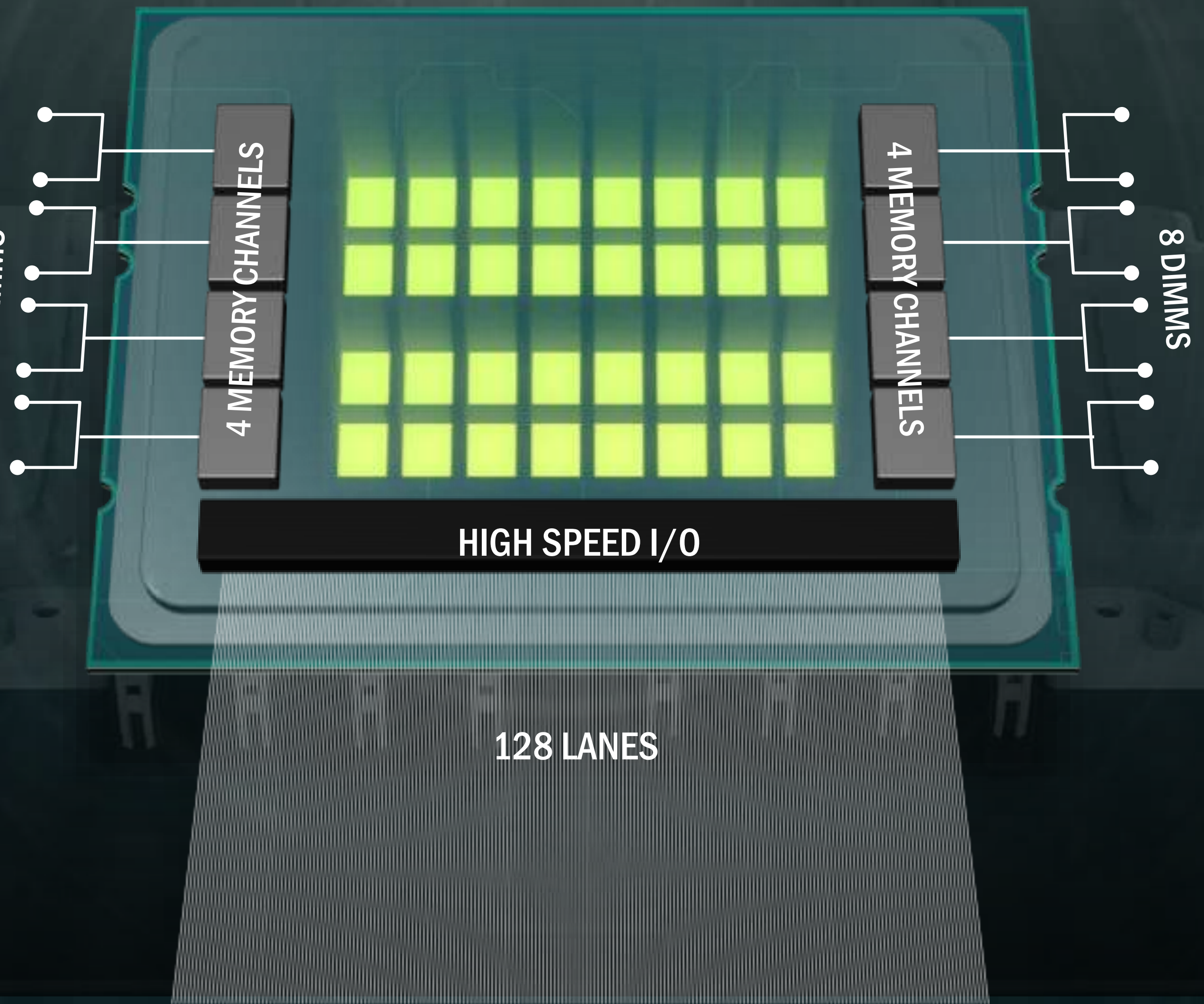
ZEN CORES

“EPYC”



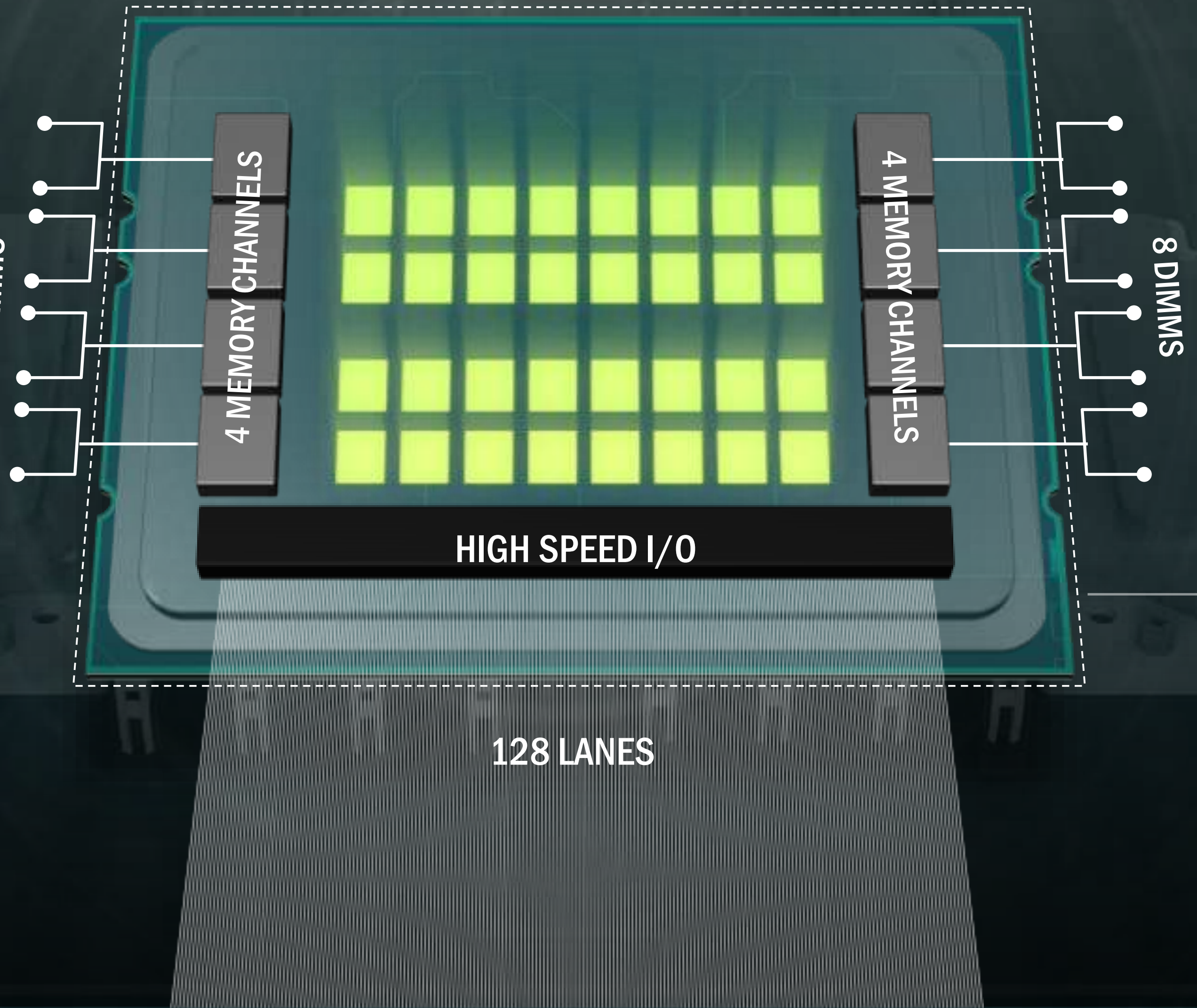
8 CHANNELS/ SOCKET

"EPYC"



MASSIVE I/O

“EPYC”



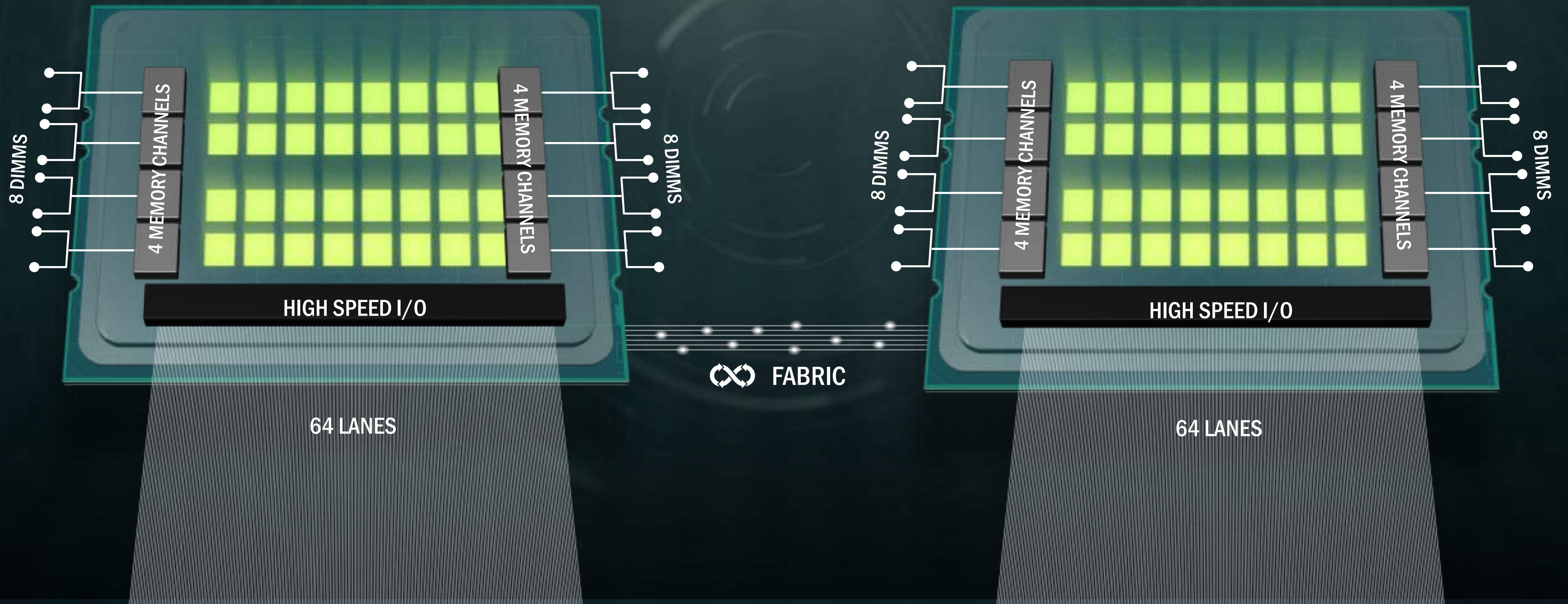
BALANCED ARCHITECTURE

SYSTEM ON A CHIP

2 SOCKET

64 CORE

128 THREAD



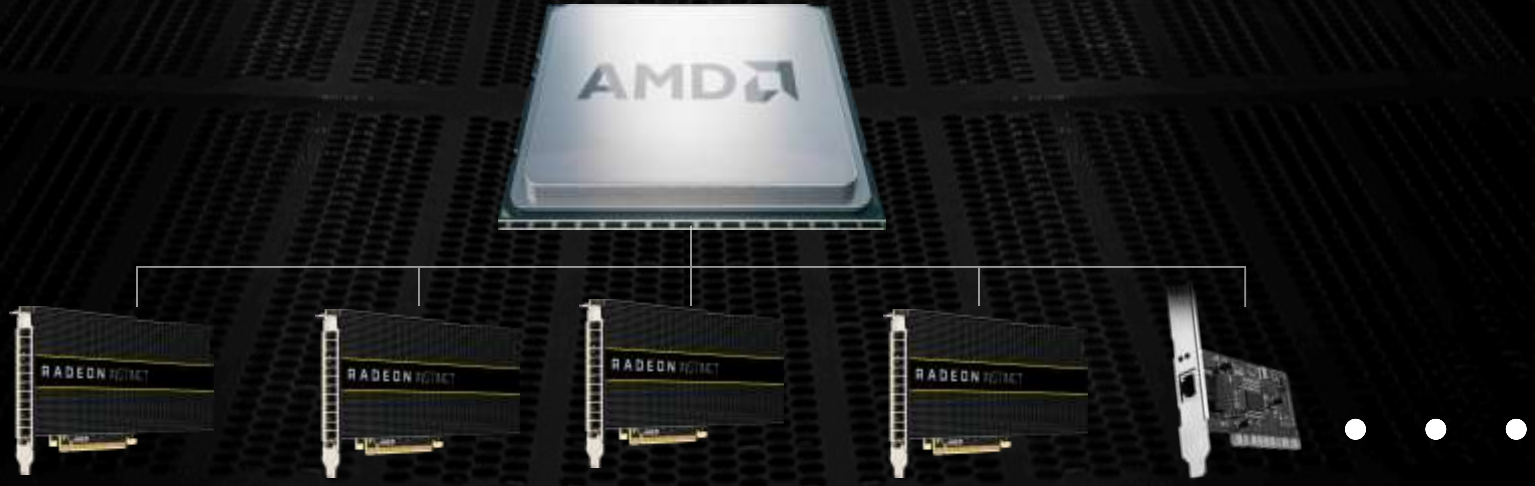
DEMO SETUP: EPYC VS. FASTEST INTEL 2-SOCKET SERVER

Both systems AMD and INTEL have the following features:

Component	AMD	INTEL
CPU model	“EPYC”	E5-2699A V4
Total CPUS	2	2
Total cores (SMT/HT on)	128	88
Total memory channels	16	8
Total memory capacity (16 GB DIMMS)	512	384
Memory frequency	2400	1866
Total PCIE gen3 lanes to CPUs	8x16=128	2x40=80

- Intel server is a standard, commercially available server from a major OEM

Radeon Instinct with Zen “EPYC” Platform



High-speed Network Fabric

Optimized for GPU and Accelerator Throughput computing

- Lower System Cost
- Lower Latency Architecture
- Peer to Peer Communication
- High Density Footprint

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

Backup slides

THE HETEROGENEOUS SYSTEM ARCHITECTURE (HSA)

AMD

- HSA is a platform architecture and software environment for simplified efficient parallel programming of heterogeneous systems, targeting:
  - Single-source language support:
    - Mainstream languages: C, C++, Fortran, Python, OpenMP
    - Task-based, domain-specific, and PGAS languages
  - Extensibility to a variety of accelerators
    - GPUs, DSPs, FPGAs,, etc.

HSA FOUNDATION

- The HSA Foundation promotes HSA via:
  - Open, royalty-free, multi-vendor specifications
  - Open-source software stack and tools
    - Runtime stack
    - Compilers, debuggers, and profilers
- See http://www.hsafoundation.com and http://github.com/hsafoundation

Board
AMD
ARM
SAMSUNG
LG Electronics
QUALCOMM
Imagination
MEDIATEK
Supporters
Arteris
FABRICENGINE
allinea
Stream Computing
Oak Ridge National Laboratory
Argonne National Laboratory
codeplay
MULTICORE WARE
Linaro
SUSE
Lawrence Livermore National Laboratory
Sandia National Laboratories
Contributors
Texas Instruments
ANALOG DEVICES
apical
BROADCOM
CANONICAL
CEVA
ETRI
ITRI
MARVELL
mobica
omp
ubuntu
SONICS
SONY make.believe
Swarm64
SYNOPSYS
tensilica
TOSHIBA Leading Innovation
VIA
S3
HUAWEI
Academic
NTHU Programming Language Lab
NTHU System Software Lab
University of BRISTOL
ILLINOIS
informatics