DELIVERING THE FUTURE OF HIGH-PERFORMANCE COMPUTING

DR. LISA SU

August 19, 2019

OUR HERITAGE

2000 — First to break the historic 1GHz barrier

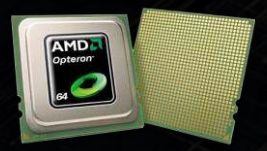
2006 — First to break teraflops performance barrier

2011 — World's first APU

2017 — First 32-Core x86 single socket server CPU

2018 — World's first 7nm datacenter GPU

2000 · 2003 · 2006 · 2009 · 2011 · 2013 · 2017 · 2018 · 2018 · 2019

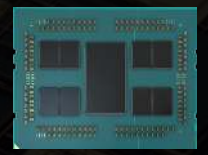


2003 — World's first x86-64 bit architecture

2009 — First to break 1GHz GPU barrier

2013 — Inside every major gaming console

2018 — World's first 32 core enthusiast desktop processor



2019 — First 7nm chiplet design

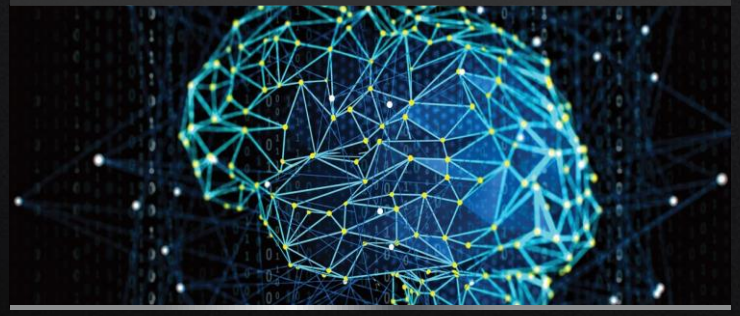
THE WORKLOADS OF THE FUTURE REQUIRE

INCREDIBLE AMOUNTS OF COMPUTE POWER

HIGH PERFORMANCE COMPUTING

CLOUD, HYPERSCALE & VIRTUALIZATION

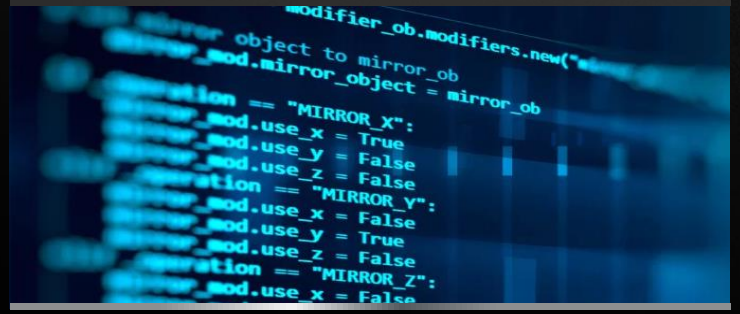
MACHINE INTELLIGENCE



BIG DATA ANALYTICS

IMMERSIVE & INSTINCTIVE COMPUTING

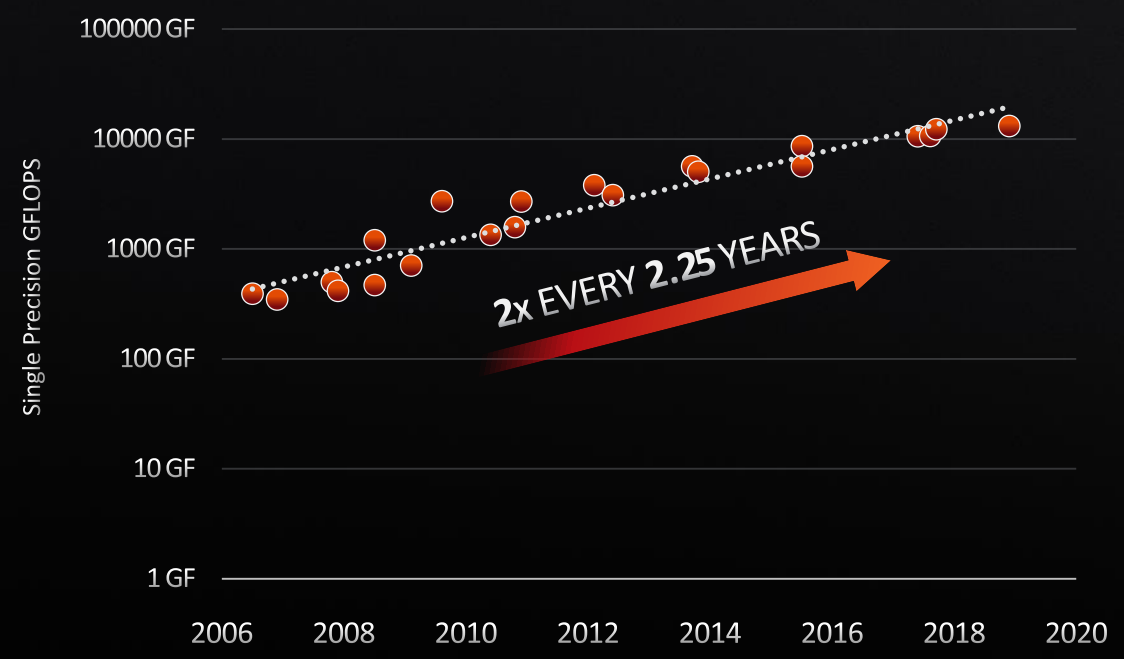
SOFTWARE-DEFINED STORAGE



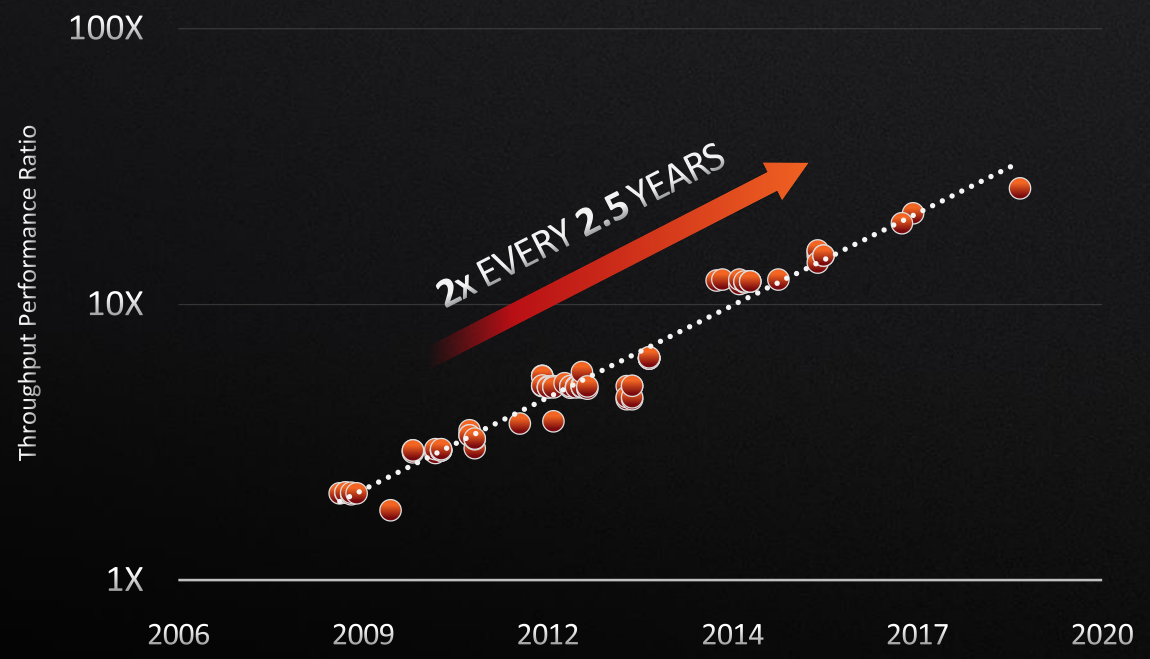
AMD

GPU AND CPU PERFORMANCE TRENDS

GPU Single Precision Floating Point Operations Per Second Trend



Specint®_rate2006 2P Server Performance Trend Over Time

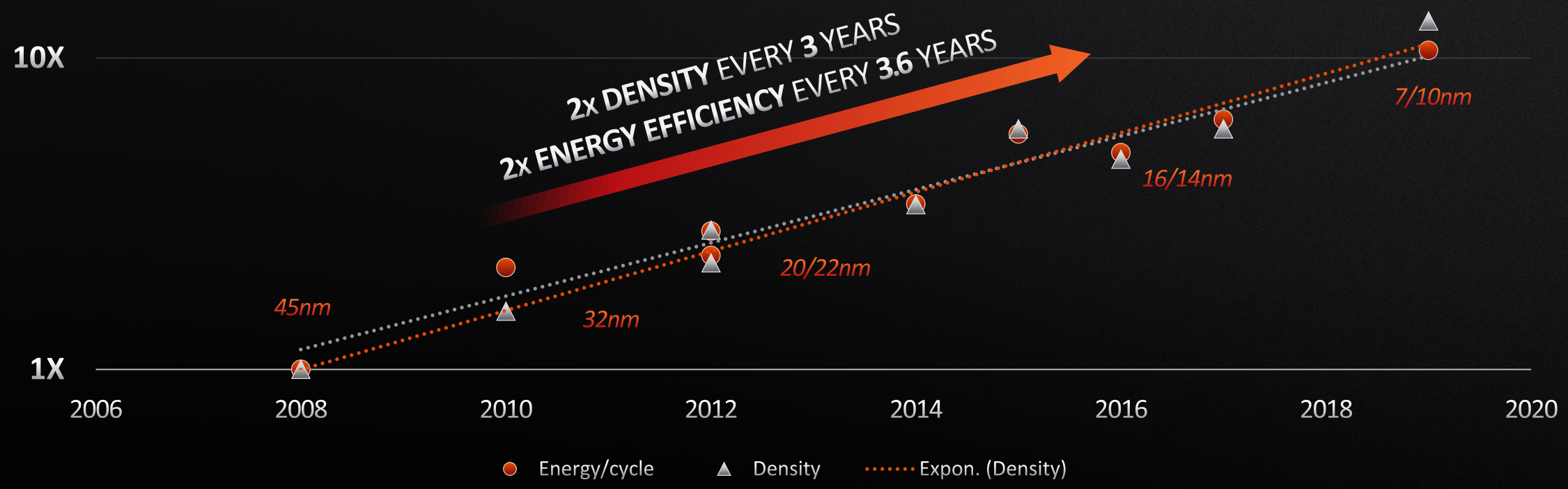


CONSISTENT AND EXPONENTIAL GPU/CPU PERFORMANCE GAINS

AMD

PROCESS TECHNOLOGY DELIVERED SIGNIFICANT GAINS

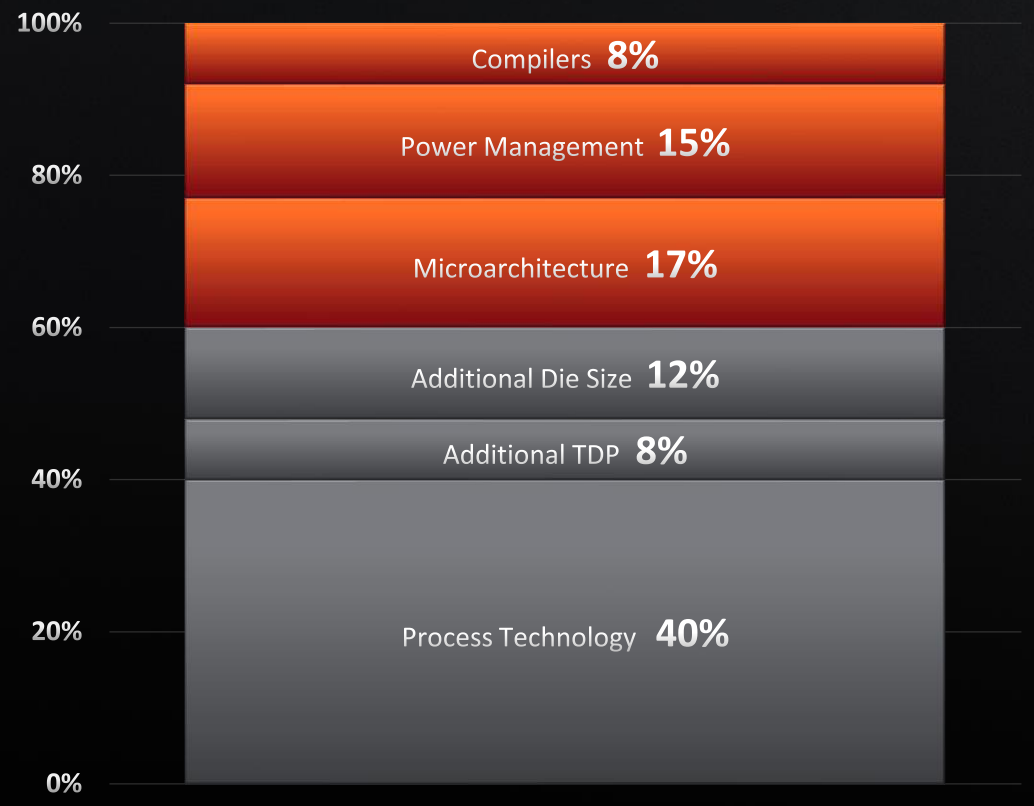
Technology Energy Efficiency and Density Across Process Nodes



SIGNIFICANT PART OF ENERGY EFFICIENCY DERIVES FROM MOORE'S LAW

AMD

PERFORMANCE GAINS OVER THE PAST DECADE



- Integration of System Components
- Micro-Architectural Efficiency
- Power Management
- Software

- More Silicon Power
- Bigger Die

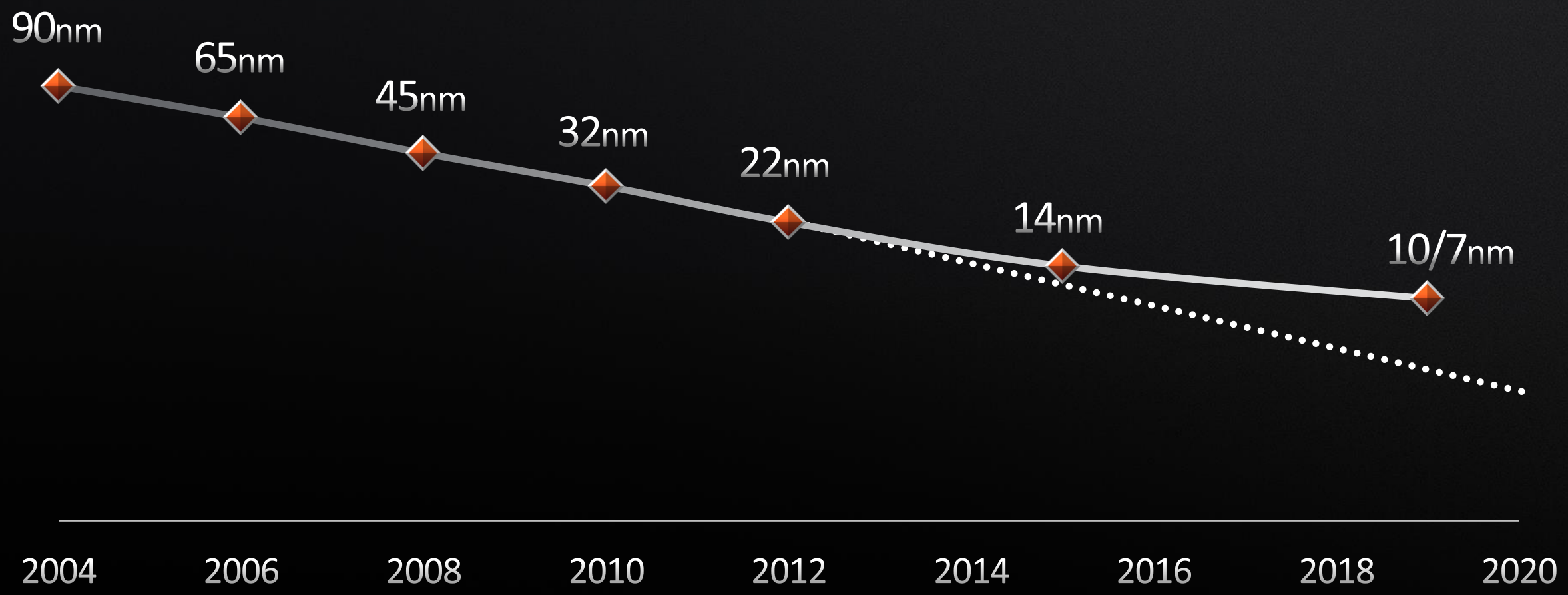
HIGHER PERFORMANCE, DENSER, LOWER POWER TRANSISTORS

ELEMENTS OF 2x IN 2.5 YEAR PERFORMANCE GAIN OVER THE PAST DECADE

AMD

BARRIERS TO CONTINUED PERFORMANCE IMPROVEMENT

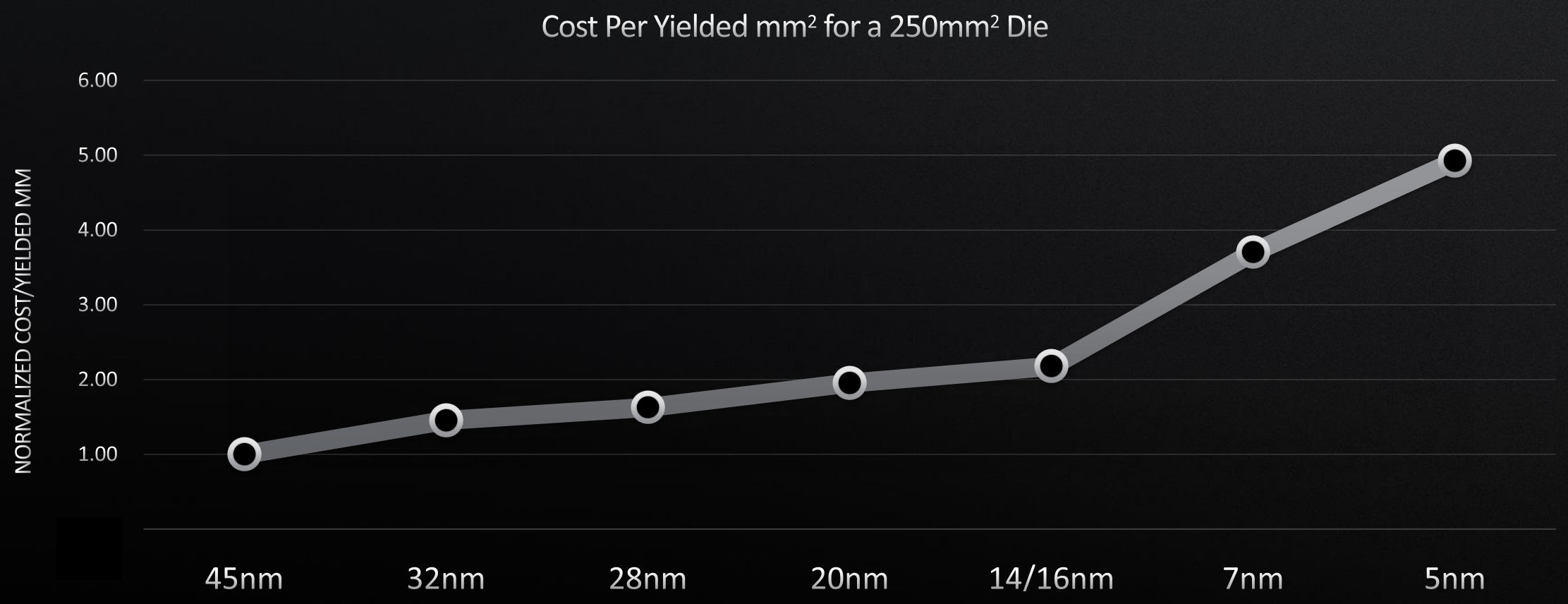
MOORE'S LAW KEEPS SLOWING



AMD Internal

AMD

WHILE COSTS CONTINUE TO INCREASE

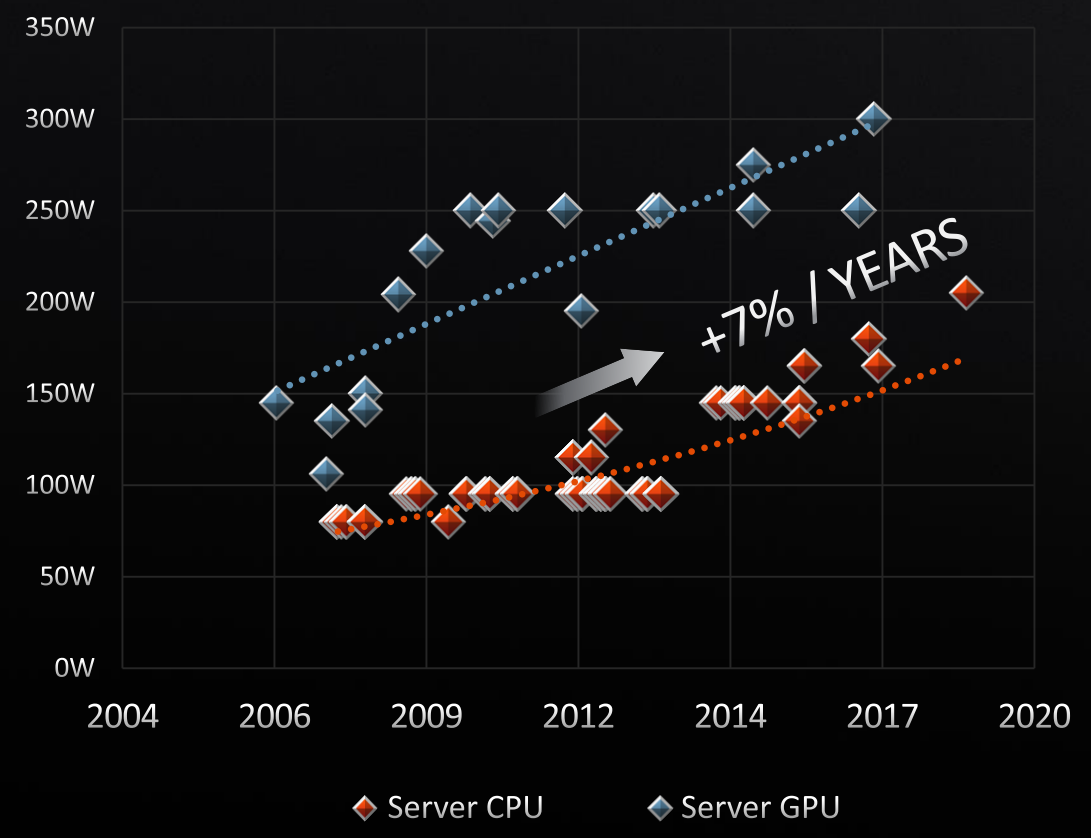


INCREASING DIE SIZES ARE ECONOMICALLY PROBLEMATIC

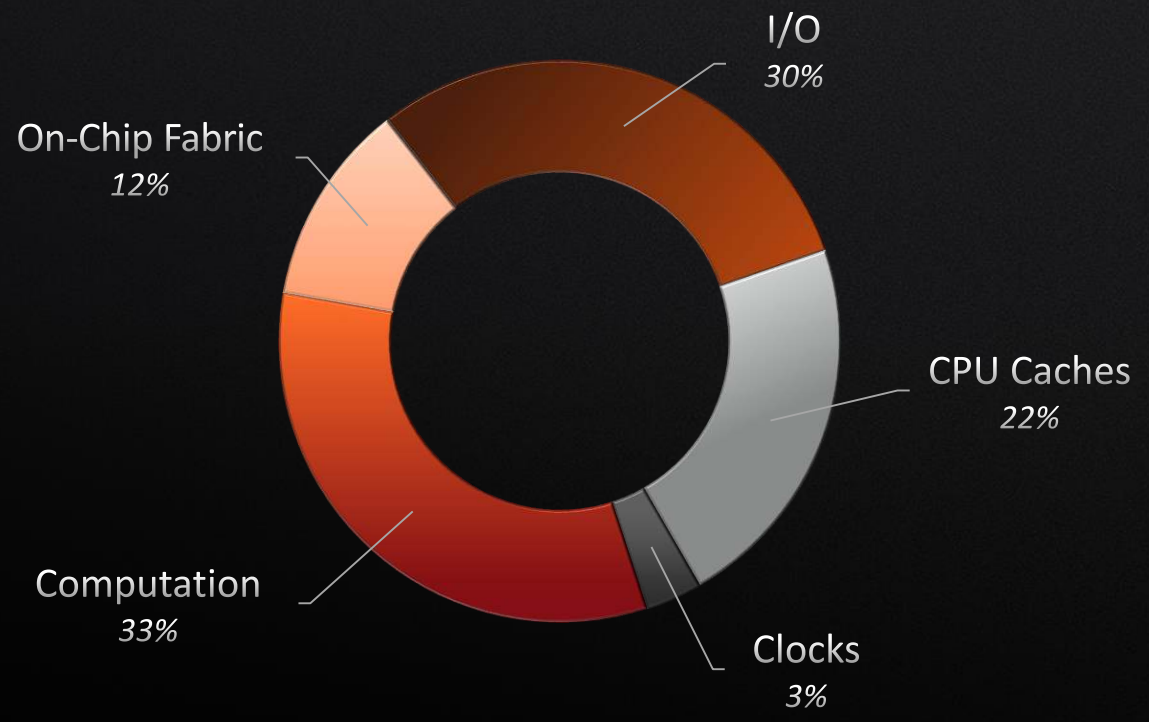
SOURCE: AMD

SOC POWER TREND

Thermal Design Power Over Time in Server CPUs and GPUs



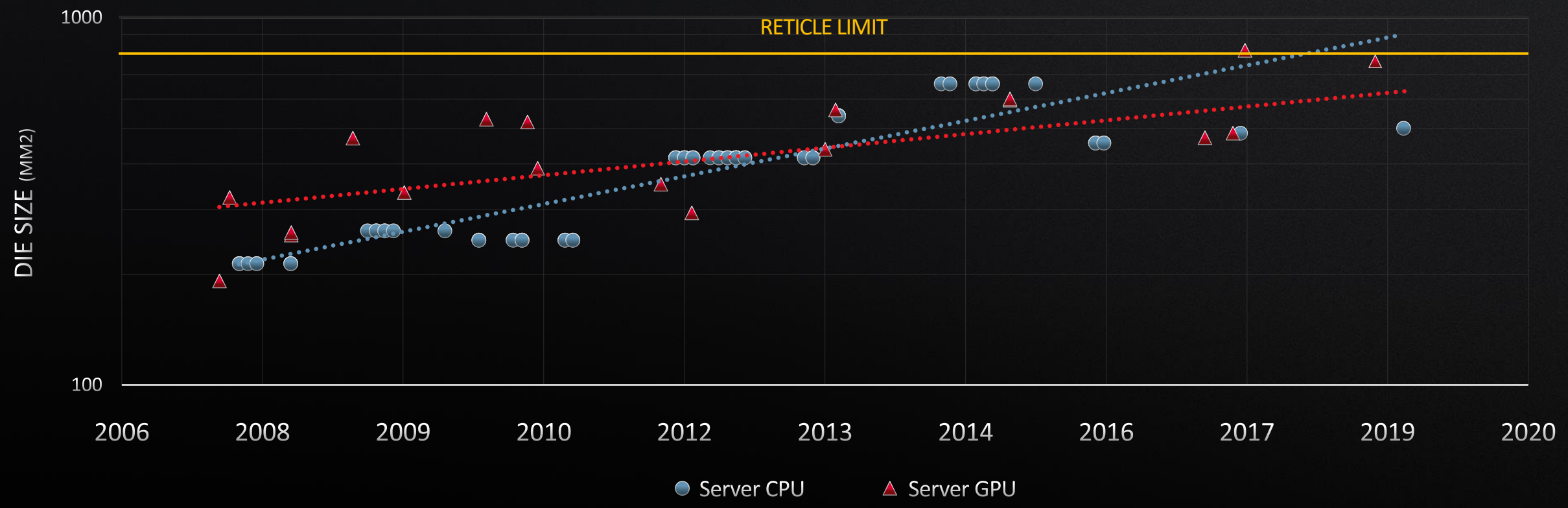
Estimated Typical Server Power Breakdown 2018



AMD

DIE SIZE TREND

Die Size Increases Over Time in Server CPUs and GPUs



DIE SIZES INCREASING AT AN UNSUSTAINABLE RATE

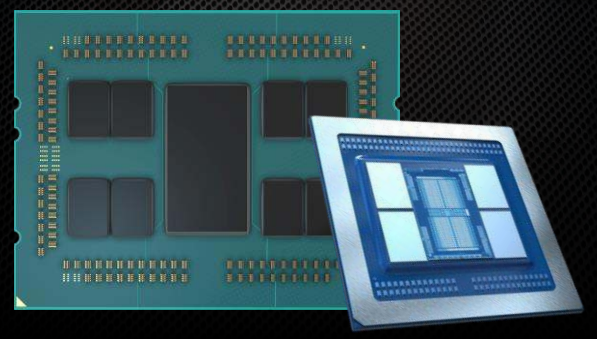
AMD

NEW APPROACHES TO EXTEND PERFORMANCE GAINS

AMD

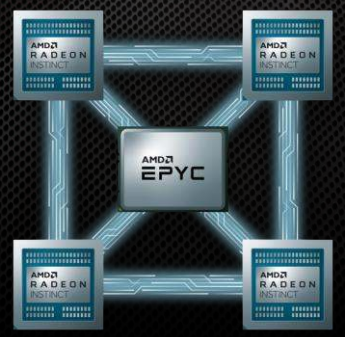
DELIVERING HIGH PERFORMANCE

SILICON



Efficient, Power Optimized Designs with Innovative Architecture

SYSTEM



Interconnects, Memory and Topologies Balanced to Address the Key Performance Bottlenecks

SOFTWARE

OpenMP

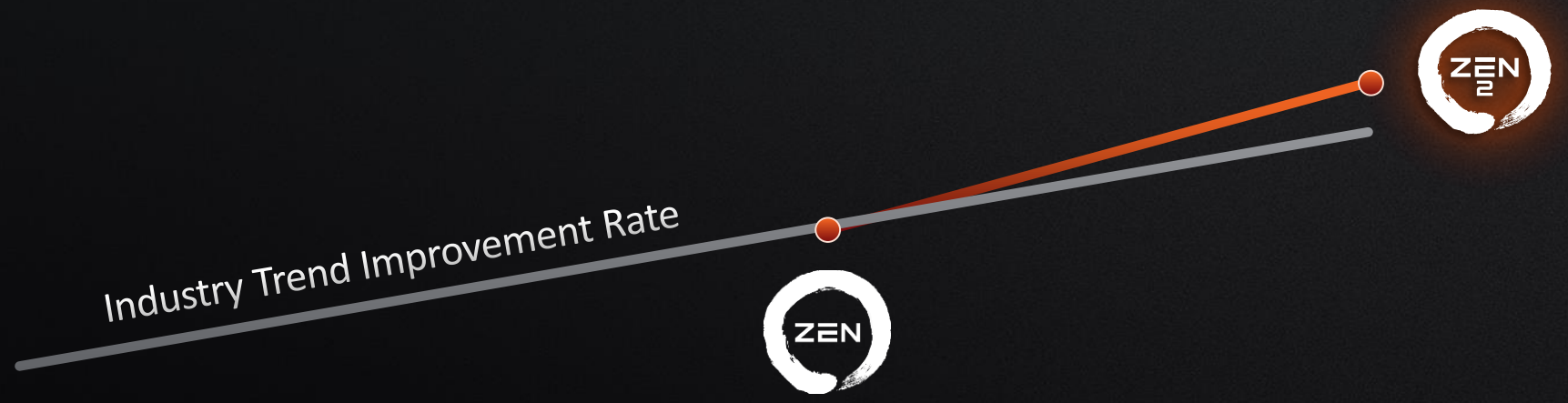
ROCm ON

Co-Optimized Software to Accelerate System Capabilities

AMD

SILICON ADVANCES

CORE DESIGN AND IPC



2013 2014 2015 2016 2017 2018 2019 2020

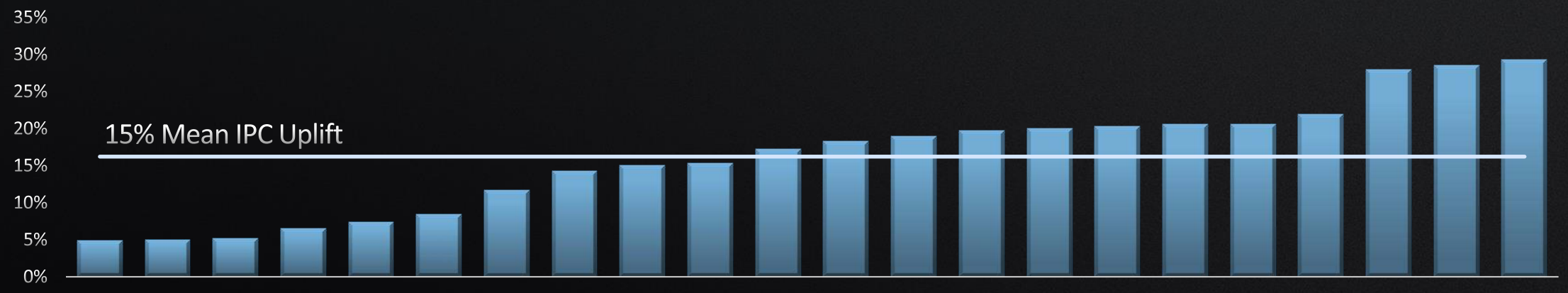
15% HIGHER INSTRUCTIONS PER CLOCK

See endnotes EPYC-07

AMD

HIGHER IPC FOR SERVER WORKLOADS

2ND GEN EPYC™ SINGLE THREAD IPC UPLIFT



Components of SPEC CPU®2017_int_ 2017 and 2006 at ISO Frequency

2ND GEN EPYC™ 32-CORE 64-THREAD IPC UPLIFT

Components of of SPEC CPU®2017_int_ 2017 and Selection of Server Workloads

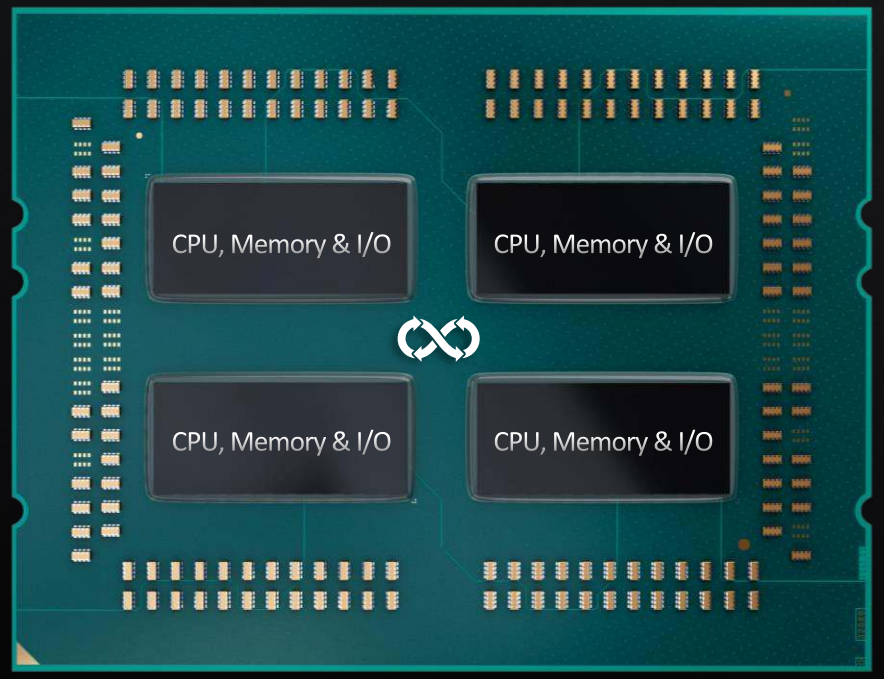
See endnotes EPYC-07, ROM-236

AMD

MULTI-CHIP ARCHITECTURES

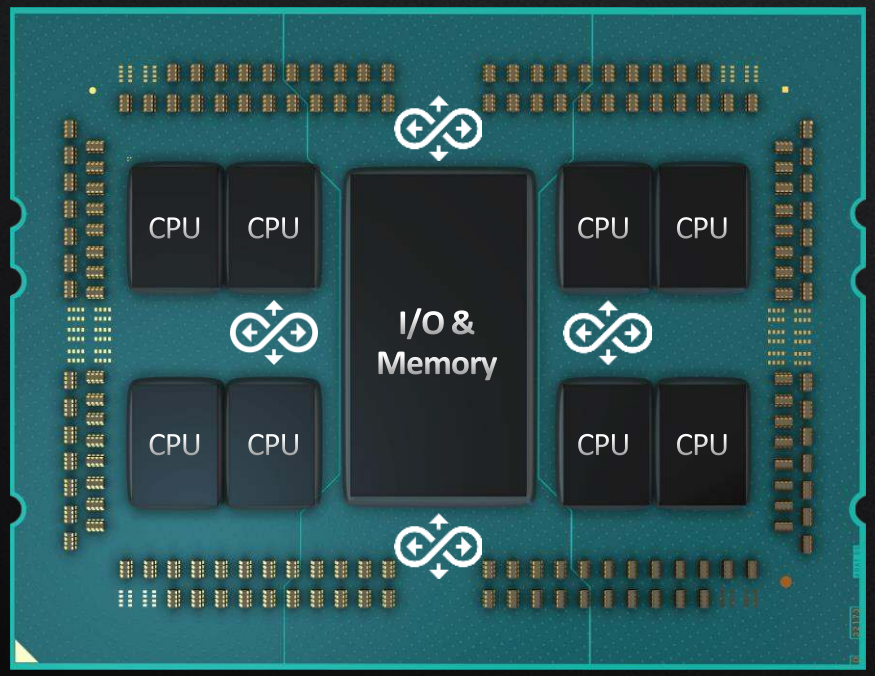
AMD

1ST GENERATION AMD EPYC



Four SOCs Interconnected
via 1st Gen AMD Infinity Architecture

2ND GENERATION AMD EPYC

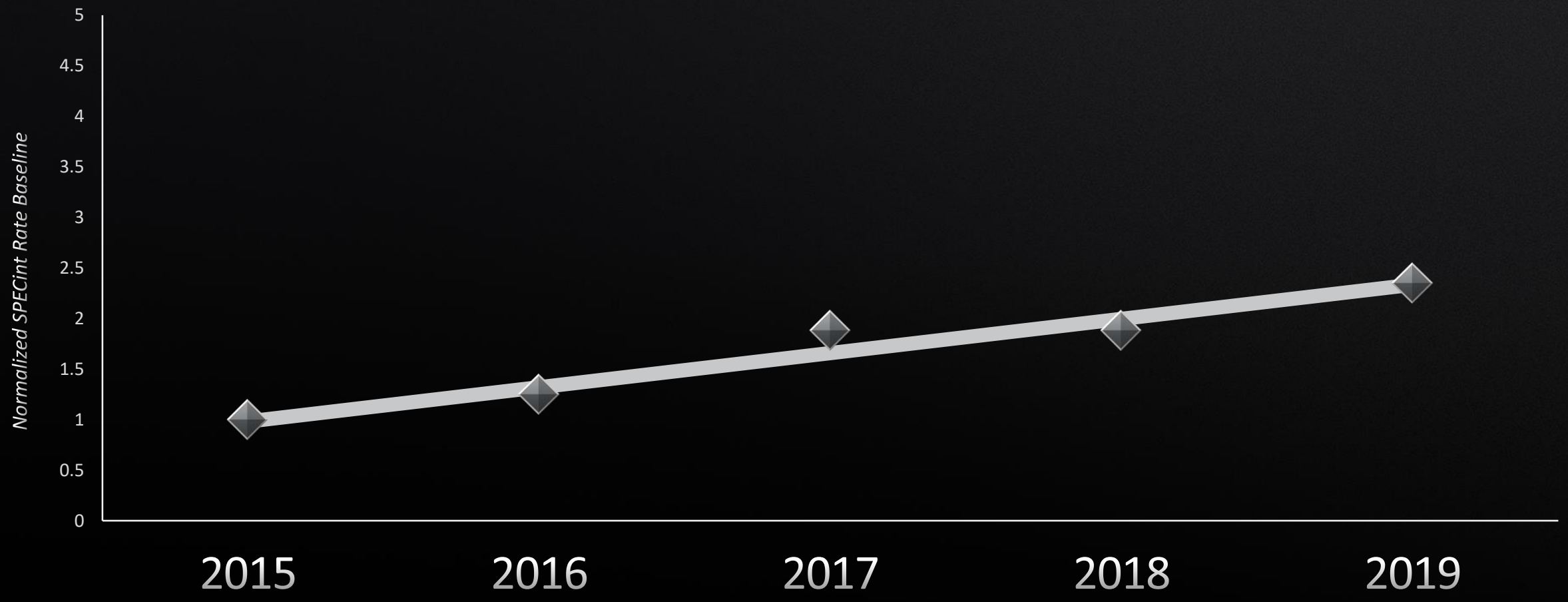


Eight 7nm Chiplet CPUs and One 12nm Chiplet I/O
Interconnected via 2nd Gen AMD Infinity Architecture

Each IP in its Optimal Process Technology	Distributed Control	I/O Die and CPU Die Optimizes Latency and Power	Flexible and More Unified Memory Architecture

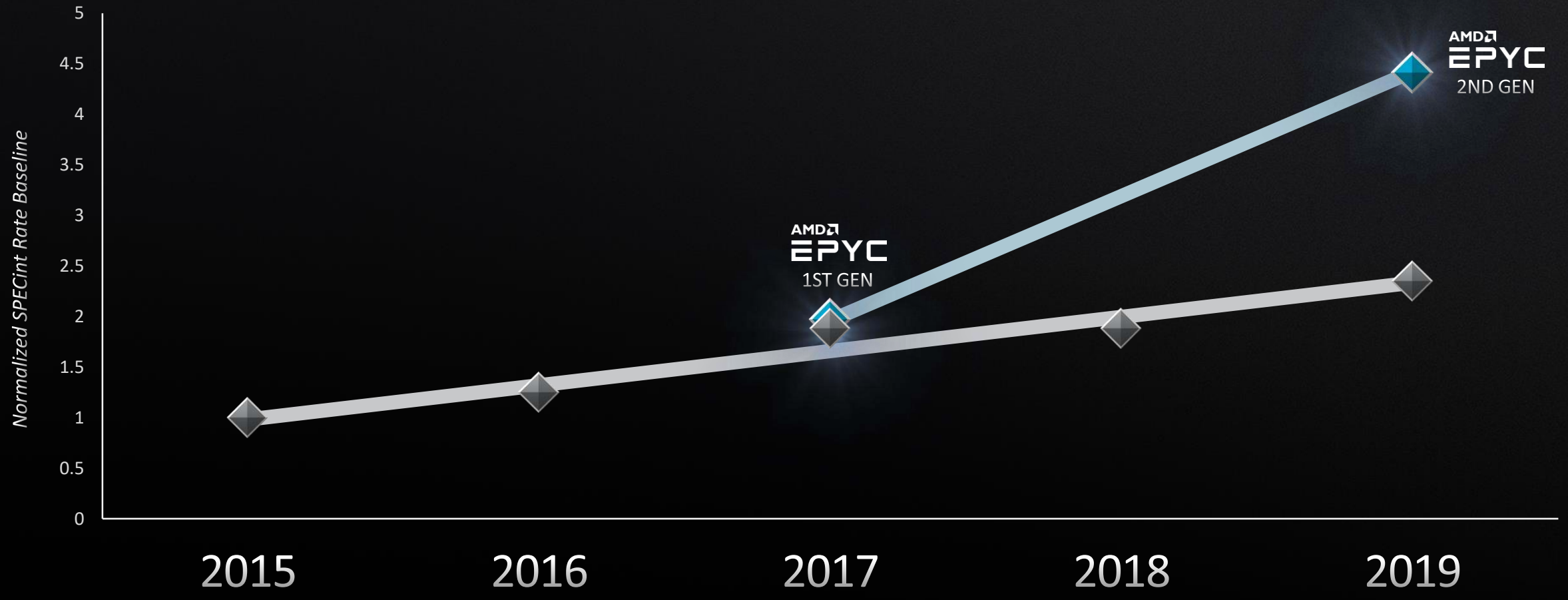
HISTORICAL DATACENTER PERFORMANCE

MAINSTREAM 2S SERVER PERFORMANCE



AMD

TAKING AN EPYC™ LEAP IN DATACENTER PERFORMANCE

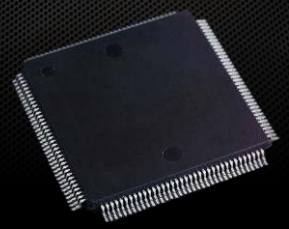


See endnotes NAP-170, ROM-92

SYSTEM ADVANCES

AMD

ACCELERATED COMPUTING PLATFORMS

CPUs	GPUs	FPGAs/Accelerators	High Speed Interconnects
			FABRIC

AMD

OPTIMIZING SYSTEM PERFORMANCE WITH ACCELERATED COMPUTING

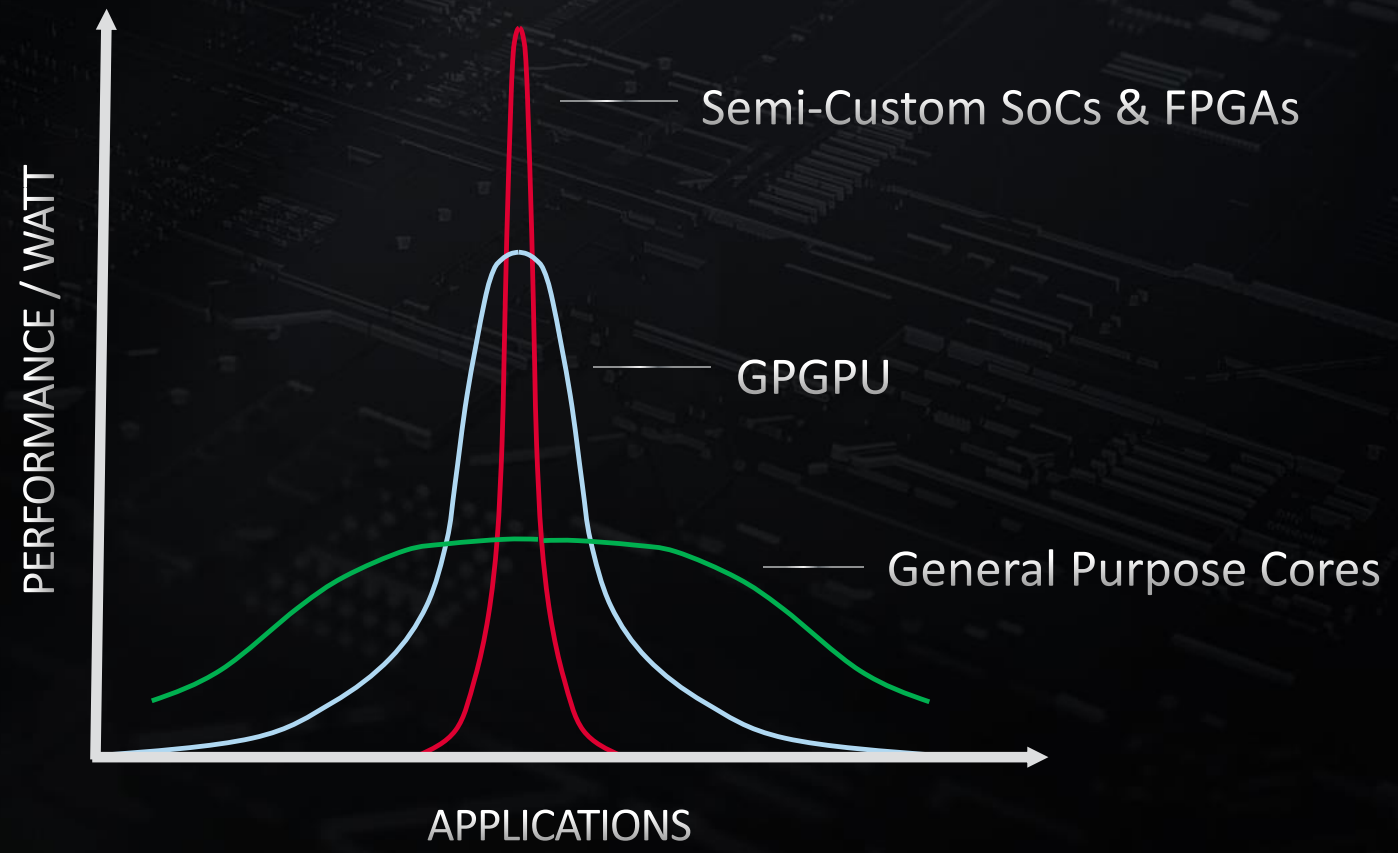
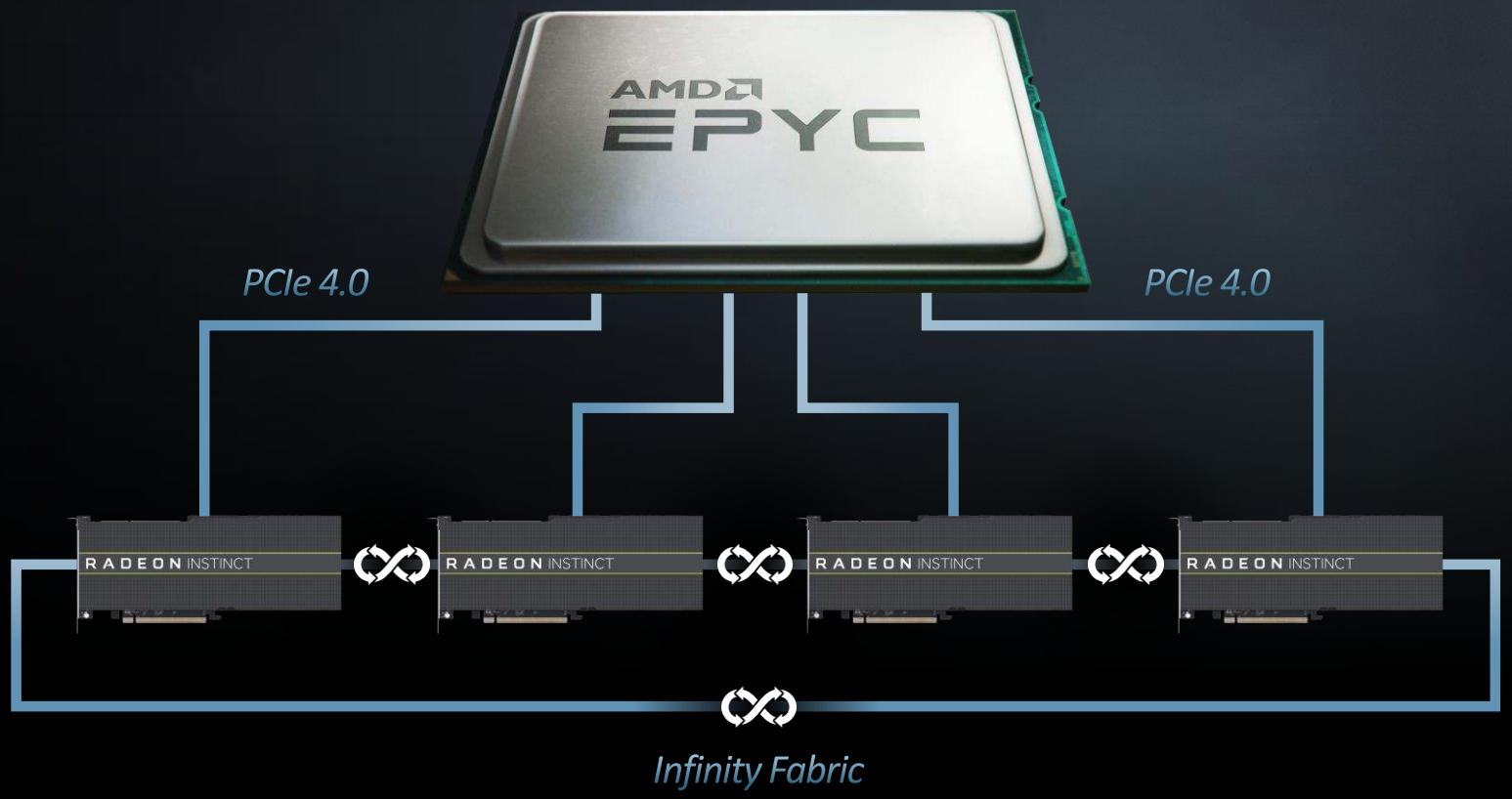


CHART FOR ILLUSTRATIVE PURPOSES

AMD

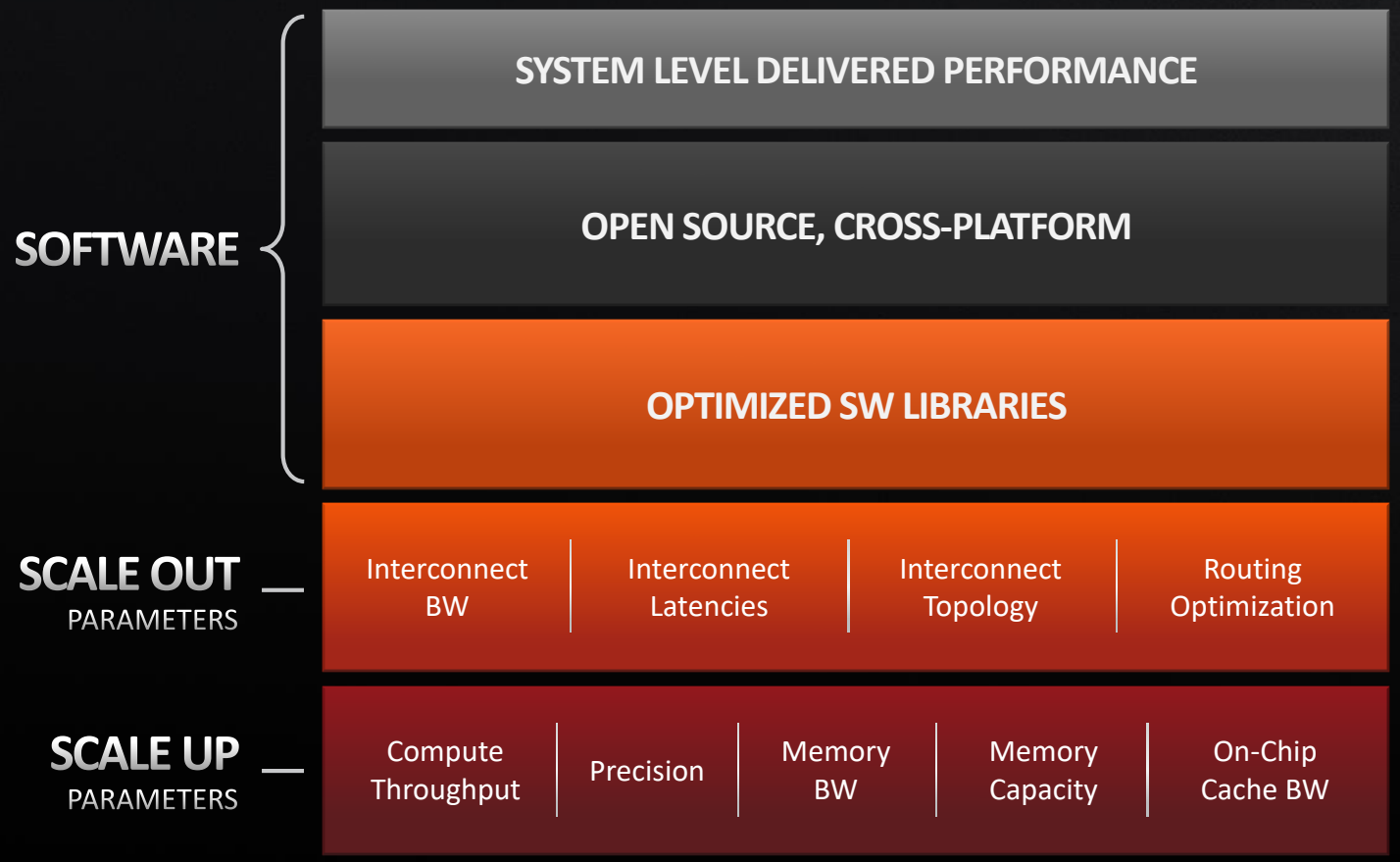
HIGH-PERFORMANCE HETEROGENEOUS PLATFORMS



SOFTWARE ADVANCES

AMD

SOFTWARE TIES IT TOGETHER



- Standard, open-source, cross-platform programming languages
- Library optimization
- Profilers and debuggers for development on AMD hardware

AMD

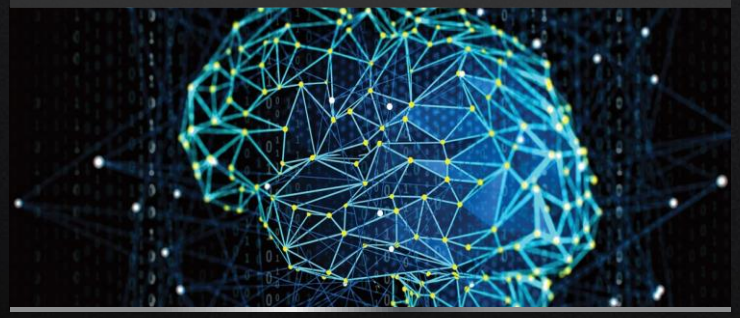
THE WORKLOADS OF THE FUTURE REQUIRE

INCREDIBLE AMOUNTS OF COMPUTE POWER

HIGH PERFORMANCE COMPUTING

CLOUD, HYPERSCALE & VIRTUALIZATION

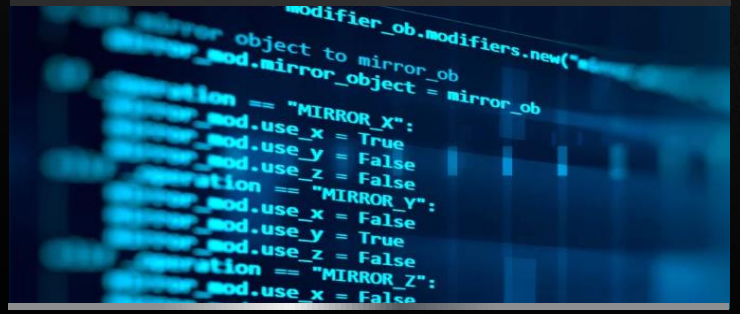
MACHINE INTELLIGENCE



BIG DATA ANALYTICS

IMMERSIVE & INSTINCTIVE COMPUTING

SOFTWARE-DEFINED STORAGE

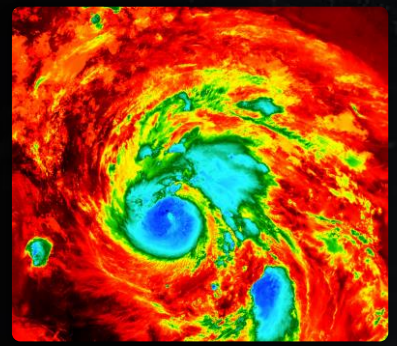


AMD

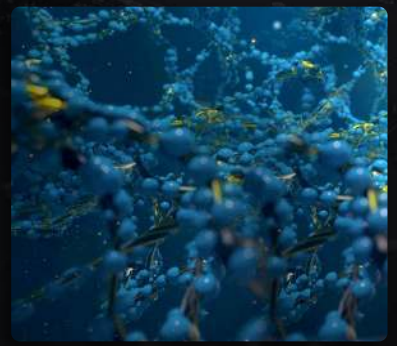
HIGH-PERFORMANCE COMPUTING

APPLICATIONS

SPACE EXPLORATION

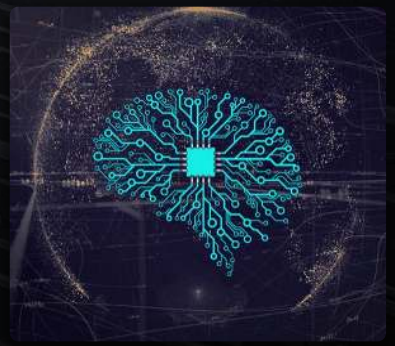


CLIMATE CHANGE

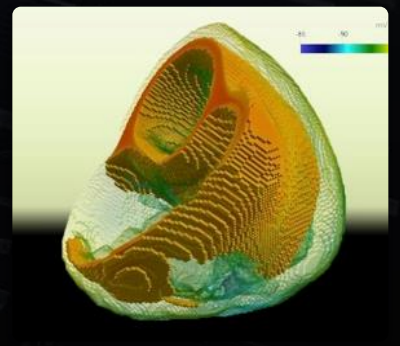


CHEMICAL SCIENCES

ENERGY SOLUTIONS



MACHINE LEARNING

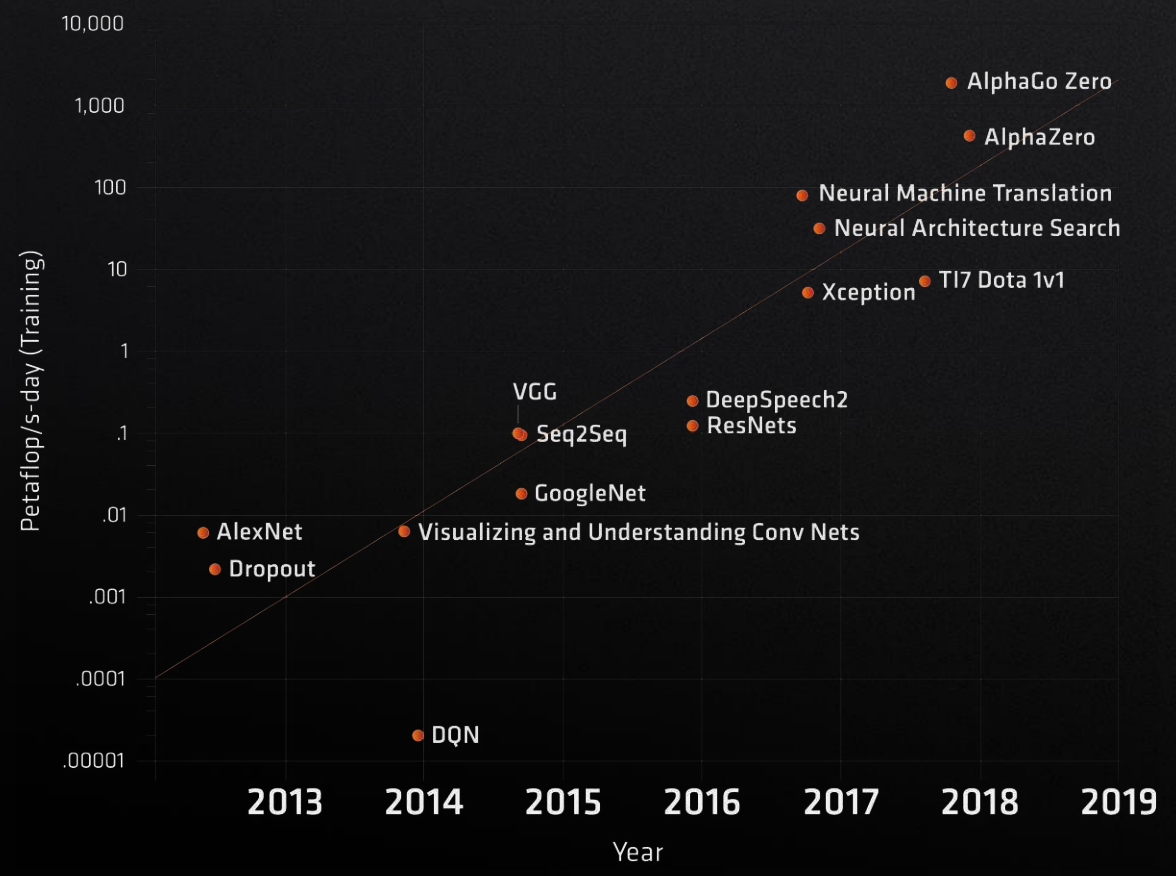


REAL TIME SIMULATION

AMD

THE RELENTLESS DEMAND FOR MORE TRAINING

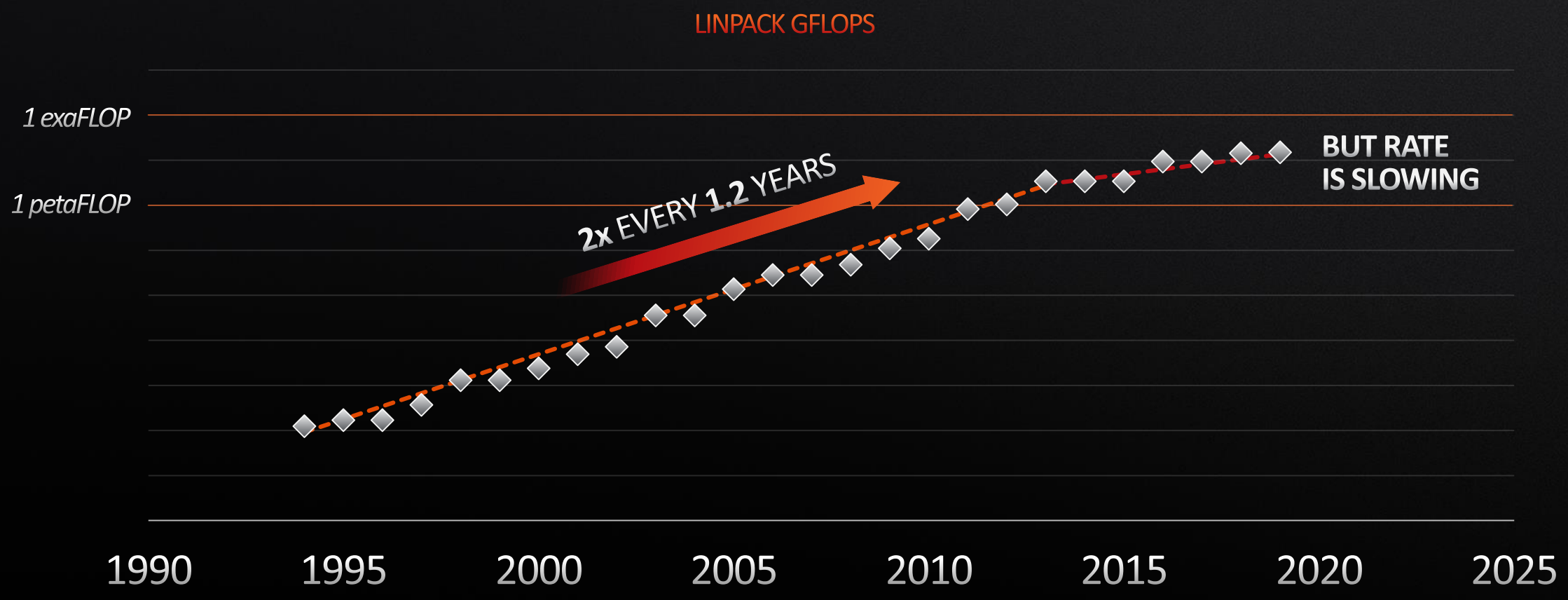
The advent of machine learning has driven a doubling of compute consumption every 3.5 months



See Endnotes

THE RELENTLESS DEMAND FOR MORE COMPUTE

The World's Fastest Supercomputers

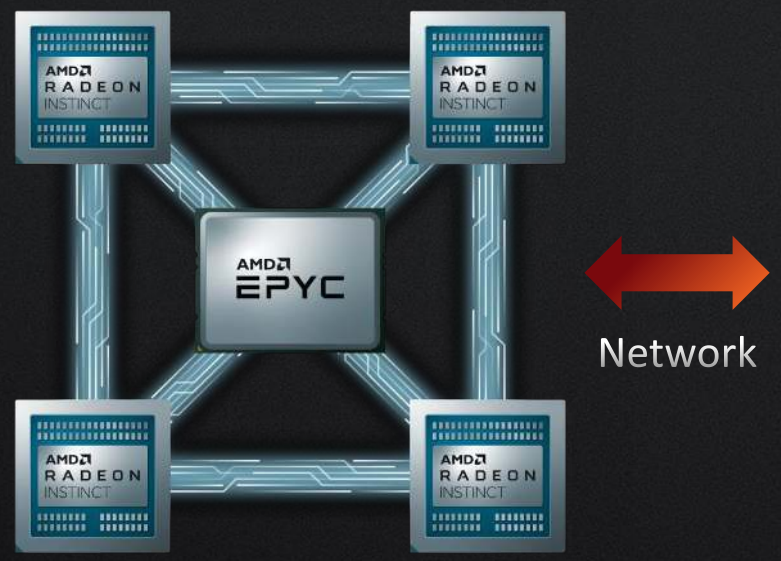


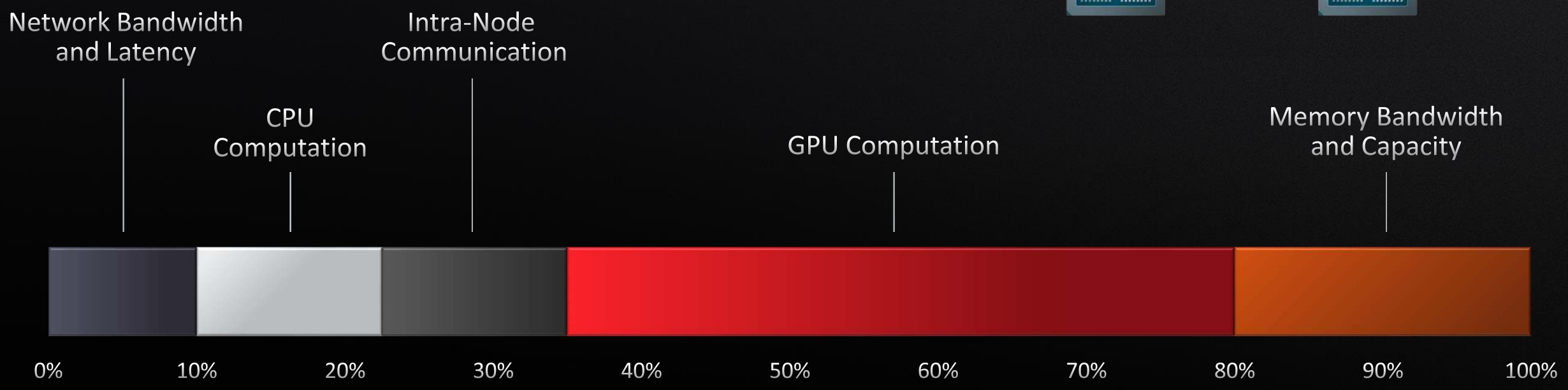
https://www.top500.org/

AMD

PRIMARY ELEMENTS OF HPC SYSTEM PERFORMANCE

Canonical Workload Runtime





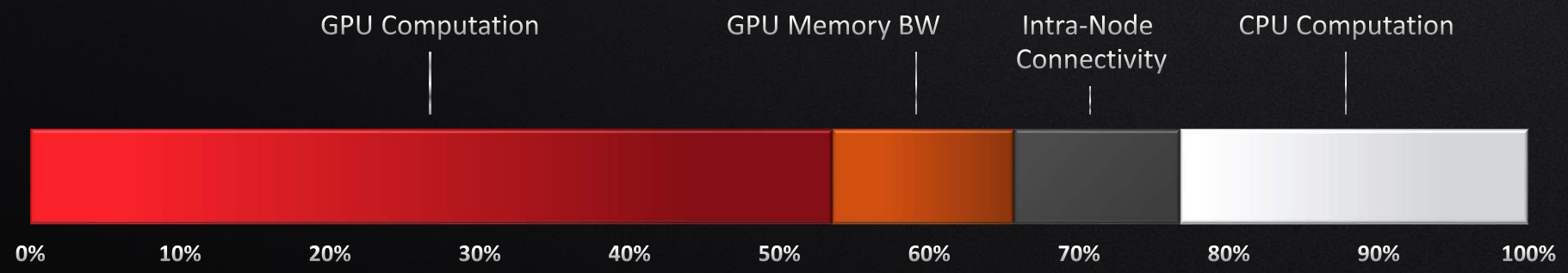
See endnote

IMPROVING EACH PERFORMANCE LIMITING ELEMENT

Common Workload Runtime Components

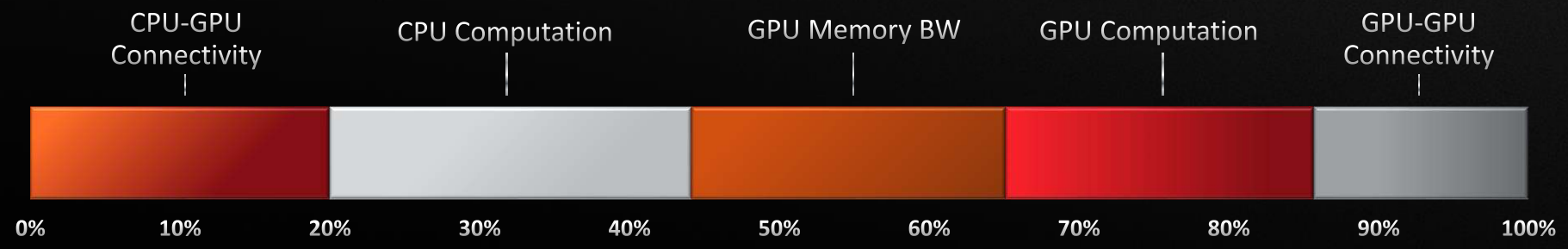
NAMD

Key application representative of workloads in molecular dynamics



TRANSFORMER

Google Brain Machine Learning algorithm applied to language translation



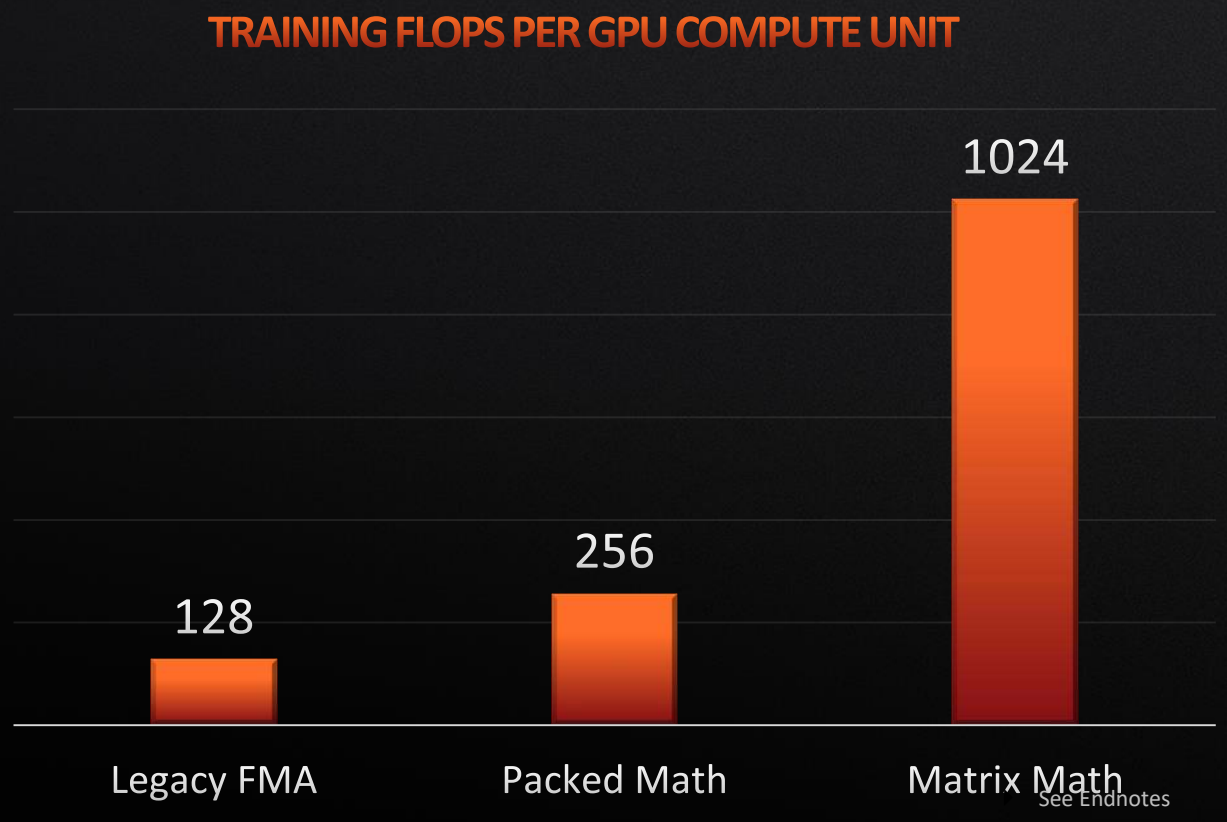
See endnote

AMD

OPTIMIZING EACH ELEMENT

GPU COMPUTATION

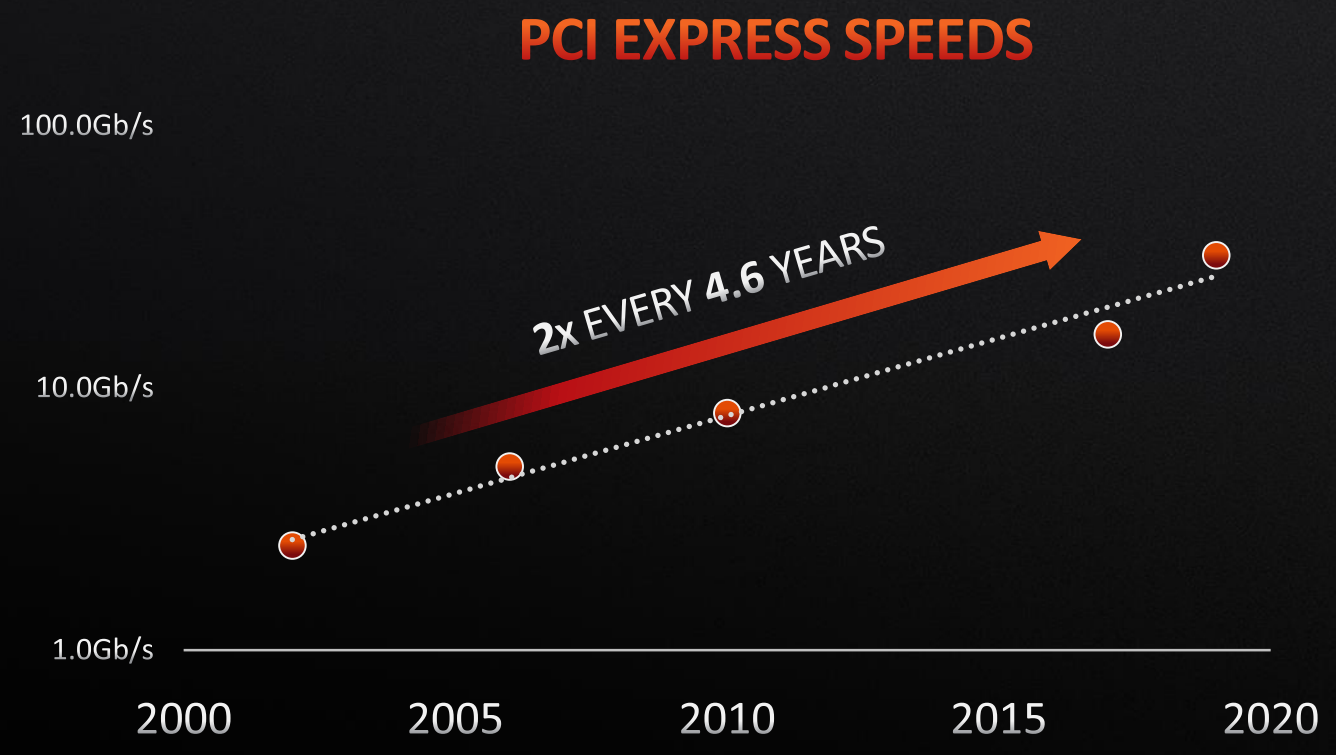
- GPU compute FLOPS/Watt is a significant contributor to performance
- Improvement in GPU compute FLOPS/Watt is under pressure from slowdown in Moore's Law
- Architectural innovations are required to maintain historical trend



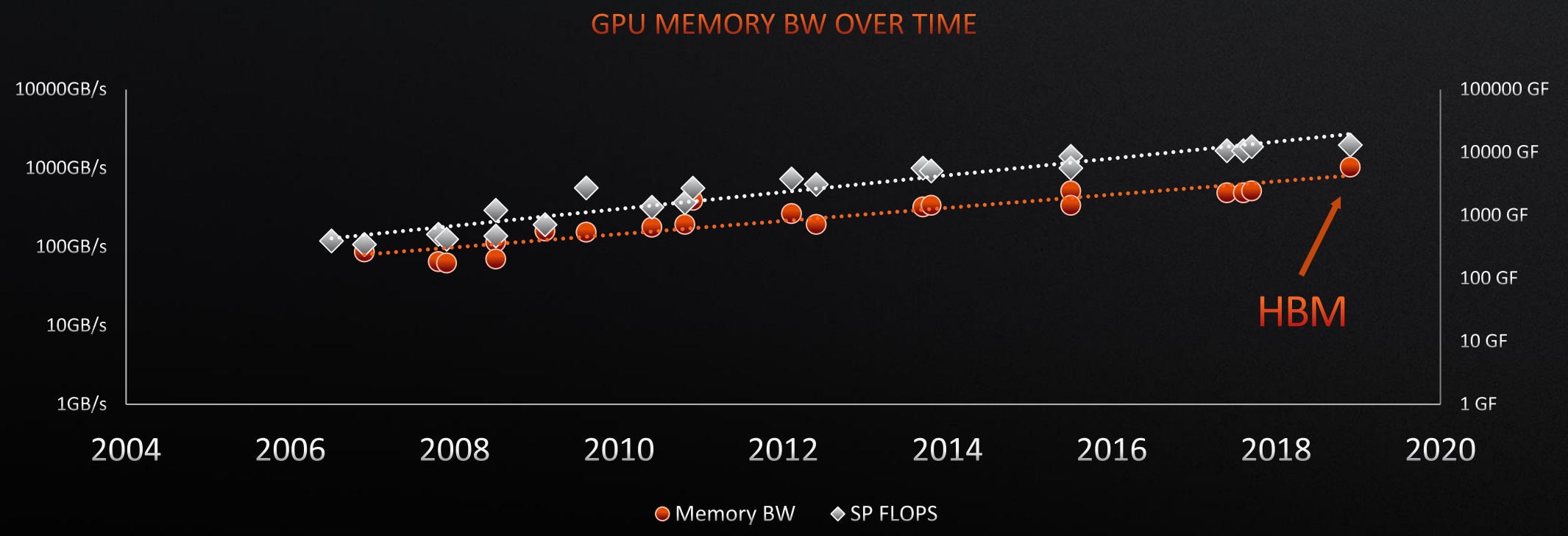
See Endnotes

OPTIMIZING EACH ELEMENT
CONNECTIVITY CPU-GPU

- High speed interfaces have been improving at a much slower rate than improvements in compute
- To support compute growth we must get more out of each bit of bandwidth



OPTIMIZING EACH ELEMENT
MEMORY BANDWIDTH



Memory BW is a Key Bottleneck

High Bandwidth 2.5D Memory is Key Innovation

On-Die Cache Hierarchy Must be Optimized

See Endnotes

AMD

POWERING THE EXASCALE ERA

EPYC CPUs and Radeon Instinct GPUs to Deliver >1.5 exaFLOPS

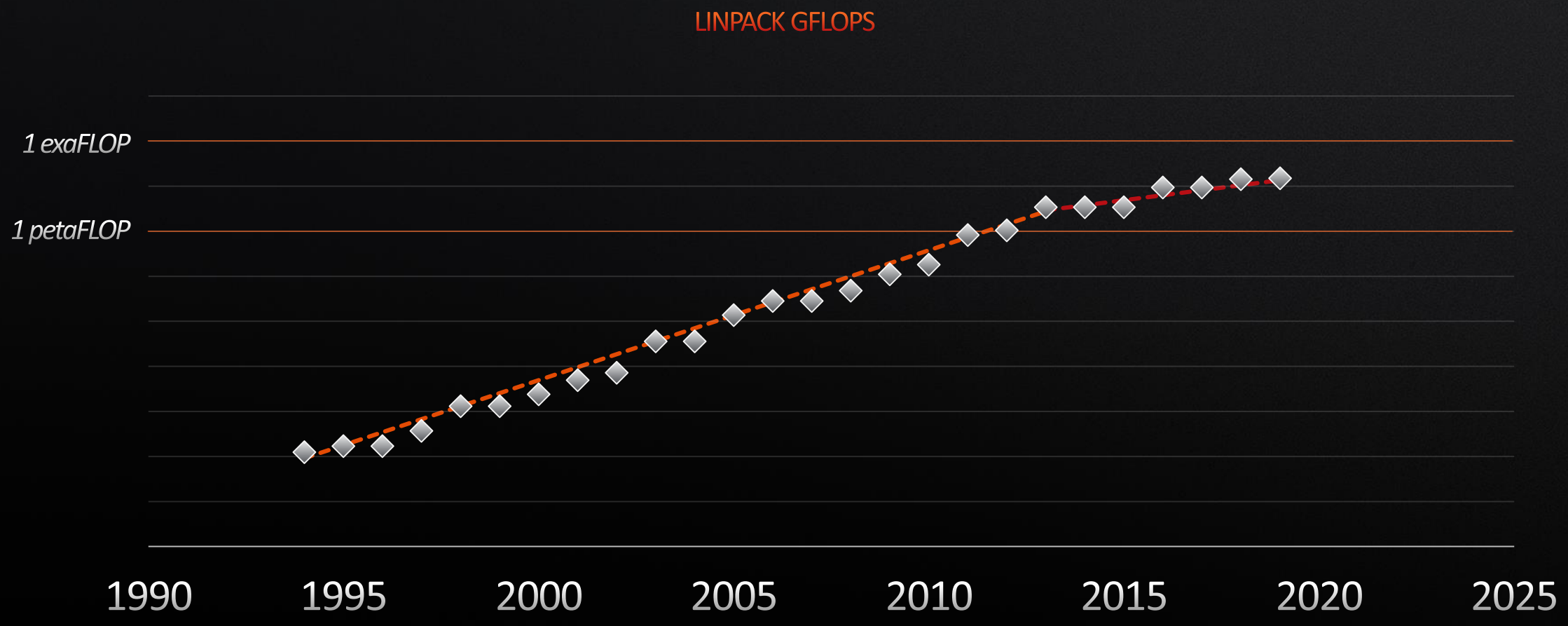
Fully Optimized CPU and GPU Design for Supercomputing

Leadership Interconnects for System Performance

Open Software Tools to Unlock the Performance

ACCELERATING HPC GAINS

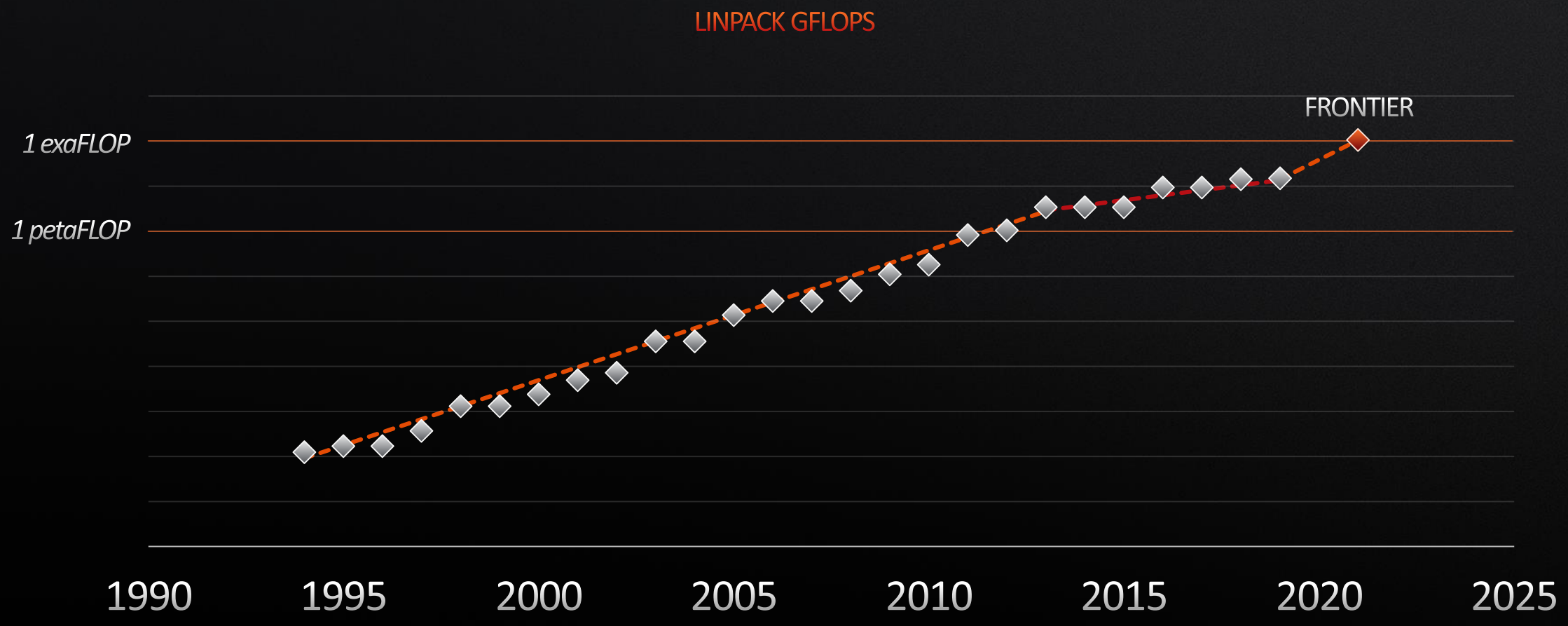
WITH SYSTEM, SOFTWARE AND SILICON CO-OPTIMIZATION



See Endnotes

ACCELERATING HPC GAINS

WITH SYSTEM, SOFTWARE AND SILICON CO-OPTIMIZATION



See Endnotes

DELIVERING HIGH PERFORMANCE COMPUTING FOR THE NEXT DECADE

Accelerated Core IP	Chiplet Architecture	High-Speed Coherent Interconnects	System & Software Co-Optimization	Continued Technology Scaling

AMD

ACKNOWLEDGEMENTS

Dr. Lisa Su gratefully acknowledges the contributions of Vineet Goel, Frank Helms, Nick Malaya, Sam Naffziger, Allen Rush, Ben Sander, Swapnil Sakharshete, Mike Schulte in the development of this presentation.

AMD

ENDNOTES

Slides 4,5,6,9,10,11

Lisa T. Su, Samuel Naffziger, and Mark Papermaster, “Multi-Chip Technologies to Unleash Computing Performance Gains over the Next Decade,” IEDM Conference 2017.

Slide 11:

Original data up to the year 2010 collected and plotted by M. Horowitz, F Labonte O.Shacham, K. Olukotun, L. Hammond, and C. Batten.
New plot and data collected for 2010-2015 by K. Rupp. https://www.karlrupp.net/2015/06/40-years-of-microprocessor-trend-data/

Slide 17:

Testing by AMD Performance Labs as of 06/03/2019 utilizing 3rd Gen AMD Ryzen™ Processors: 3900X, 3800X, 3700X, 3600X, 3600 and Ryzen™ 7 2700X in Cinebench R20 1T. Results may vary. RZ3-25
Based on June 8, 2018 AMD internal testing of same-architecture product ported from 14 to 7 nm technology with similar implementation flow/methodology, using performance from SGEMM. EPYC-07
Based on AMD internal testing, average per thread performance improvement at ISO-frequency on a 32-core, 64-thread, 2nd generation AMD EPYC™ platform as compared to 32-core 64-thread 1st generation AMD EPYC™ platform measured on a selected set of workloads including sub-components of SPEC CPU® 2017_int and representative server workloads. ROM-236

Slide 20

The comparison is based on the highest performing results for two-processor servers using AMD EPYC 7601 processors and Intel Xeon Gold 6248 processors published on www.spec.org as of April 27, 2019.

- • Score of 234 using 2 x Intel Xeon Gold 6248 processors.
- https://www.spec.org/cpu2017/results/res2019q2/cpu2017-20190318-11225.html
- • Score of 301 using 2 x AMD EPYC™ processor model 7601.
- https://www.spec.org/cpu2017/results/res2019q1/cpu2017-20190304-11124.html
- SPEC® and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. Learn more at www.spec.org. NAP-170

ENDNOTES

Slides 20

A 2P EPYC™ 7742 processor powered server has SPECrate®2017_int_base score of 682, https://spec.org/cpu2017/results/res2019q3/cpu2017-20190722-16242.html as of August 7, 2019 The next highest base score is a 2P Intel Platinum 9282 server with a score of 643, http://spec.org/cpu2017/results/res2019q3/cpu2017-20190624-15369.pdf as of July 28, 2019. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. ROM-92

Slide 29

Dario Amodei and Danny Hernandez. "AI and Compute." https://openai.com/blog/ai-and-compute/.

Slide 31, 32 , 35 AMD internal performance modeling and analysis
Slide 33 AMD Internal performance modeling and analysis
https://en.wikipedia.org/wiki/Graphics_Core_Next

Slide 38
https://www.top500.org/ and ORNL performance estimate.