The bridge to possible

Unlocking the Engineering Secrets of UCS AMD M8 Servers

Pritish Nilangi, Product Marketing Manager AMD

BRKCOM-1722

cisco Live!

Cisco Webex App

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKCOM-1722

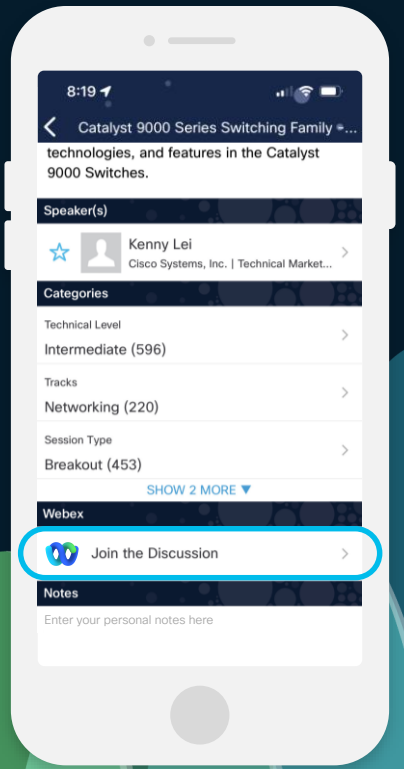
Questions?

Use Cisco Webex App to chat with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

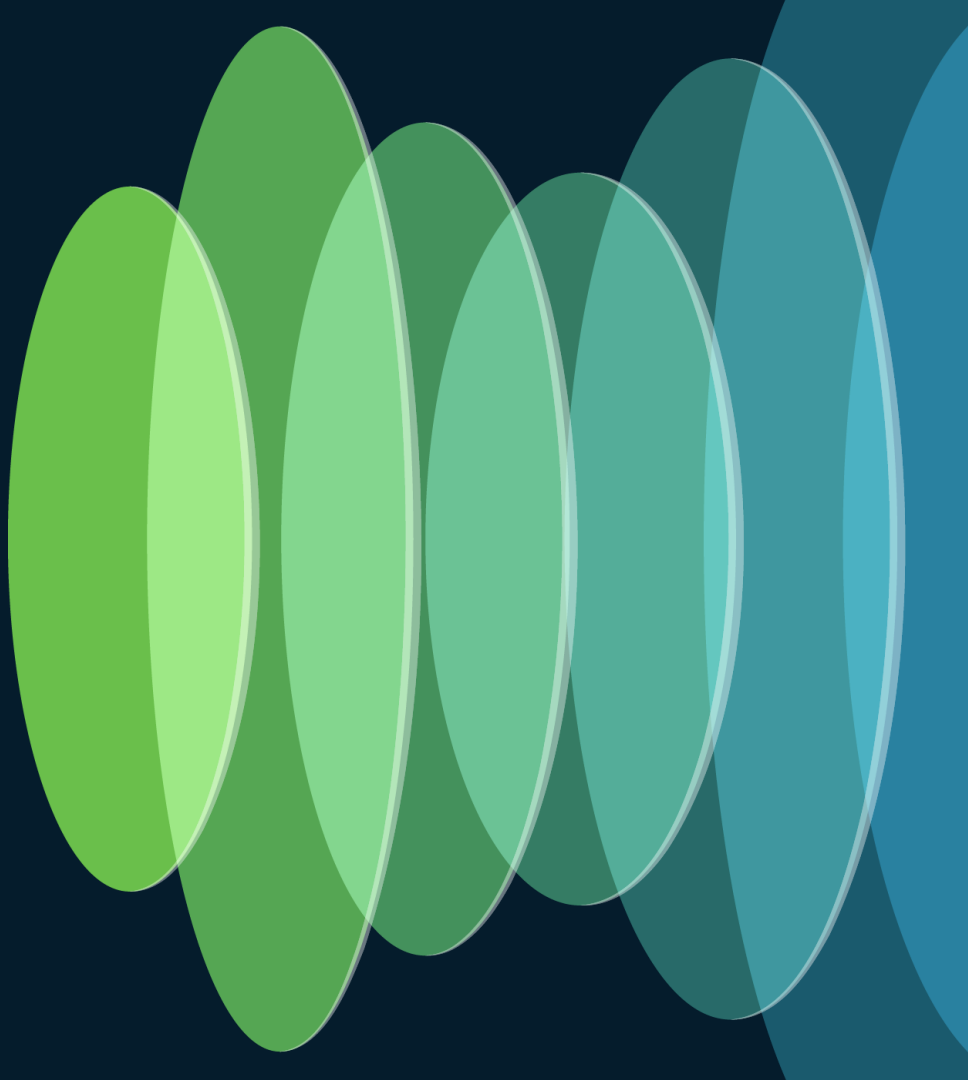
Webex spaces will be moderated by the speaker until June 7, 2024.



Agenda

- AMD EPYC CPU
- Cisco UCS and AMD

AMD EPYC CPU



INFORMATION TECHNOLOGY CHALLENGES

MARKET DISRUPTION

Changing Independent Software Vendor (ISV) Landscape

Increasing Power and Cooling Costs

AGING INFRASTRUCTURE

Costly to maintain

Can’t keep pace with business demands

Ever Increasing Security Threats

EXPANDING DEMANDS

Core Workloads

AI

Cloud Native

Digital Transformation

Hybrid Workforce

Security Enhancements

MODERN DATA CENTERS NEED
WORKLOAD-OPTIMIZED ENGINES

Server CPUs

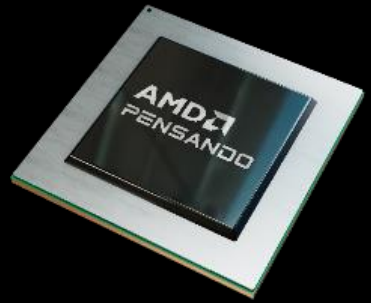
AMD EPYC

AI Accelerators

AMD INSTINCT AMD ALVEO

FPGAs and Adaptive SoCs

AMD ALVEO AMD VERSAL



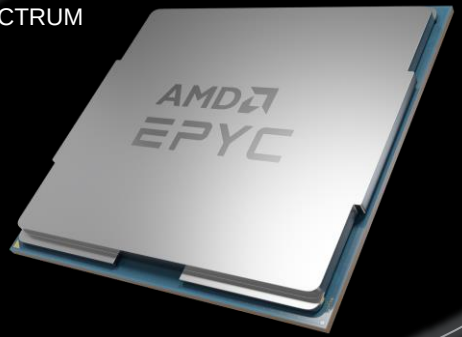
SmartNICs and DPUs

AMD ALVEO AMD PENSANDO

EXECUTING ON WORKLOAD OPTIMIZED PORTFOLIO

"GENOA"
Leading Performance & Performance per Core
WORKLOADS: BROAD SPECTRUM

"BERGAMO"
Highest Thread Density
WORKLOADS: CLOUD NATIVE & FLOP INTENSIVE



"GENOA-X"
Highest Cache
WORKLOADS: TECHNICAL COMPUTE

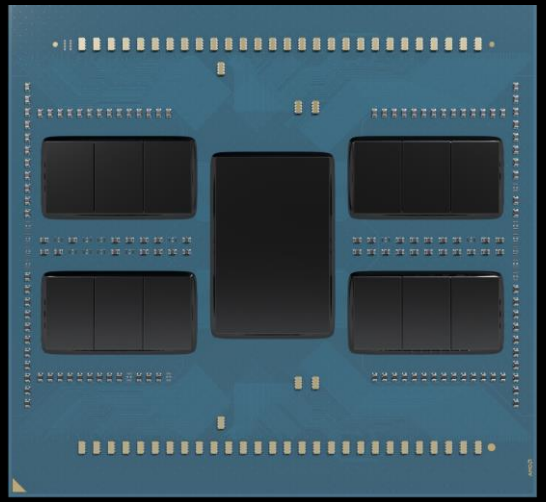
"SIENA"
Perf per Watt Optimized
WORKLOADS: LOW POWER FORM FACTORS

PR 2Q23

Leadership 5nm Process Node	High-Performance 'Zen4' Cores	Common ISA & software	Next Generation High-Speed I/O	Advanced Security Features

4th Gen AMD EPYC™ 9000 Series CPU

Extending Compute Leadership



High Performance Cores	Leadership 5nm Process Node	Leadership Memory Bandwidth and Capacity	AMD Infinity Guard	Leadership Energy Efficiency

AMD Infinity Guard features vary by EPYC™ Processor generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard. GD-183

AMD EPYC™ 9004 "GENOA" - AT A GLANCE

COMPUTE

- AMD "Zen4" x86 cores (Up to 12 CCDs / 96 cores / 192 threads)
- 1MB L2/Core, Up to 32MB L3/CCD
- ISA updates: BFLOAT16, VNNI, AVX-512 (256b data path)
- Memory addressability with 57b/52b Virtual/Physical Address

- Updated IOD and internal AMD Gen3 Infinity Fabric™ architecture with increased die-to-die bandwidth

- Target TDP range: Up to 400W (cTDP)

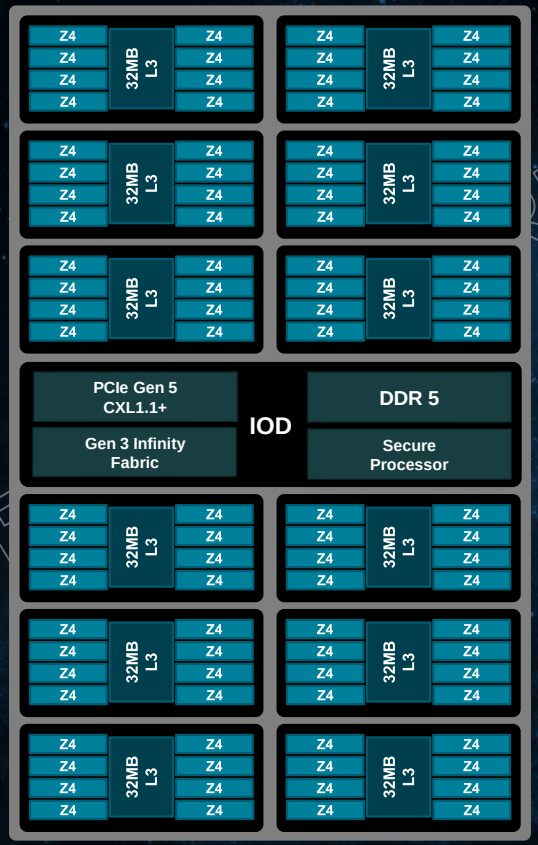
- Updated RAS

MEMORY

- 12 channel DDR5 with ECC up to 4800 MHz
- Option for 2,4,6, 8, 10, 12 channel memory interleaving[1]

- RDIMM, 3DS RDIMM

- Up to 2 DIMMs/channel capacity with up to 12TB in a 2 socket system (2DPC, 256GB 3DS RDIMMs)[1]



BLUE font indicates significant upgrades with EPYC 9004.

SP5 PLATFORM

- New socket, increased power delivery and VR
- Up to 4 links of Gen3 AMD Infinity Fabric™ with speeds of up to 32Gbps
- Flexible topology options

- Server Controller Hub (USB, UART, SPI, I2C, etc.)

INTEGRATED I/O – NO CHIPSET

Up to 160 IO lanes (2P) of PCIe® Gen5

- Speeds up to 32Gbps, bifurcations supported down to x1
- Up to 12 bonus PCIe Gen3 lanes in 2P config (8 lanes–1P)
- Up to 32 IO lanes for SATA
- SDCI (Smart Data Cache Injection) *

- 64 IO Lanes support for CXL1.1+ with bifurcations supported down to x4

SECURITY FEATURES

Dedicated Security Subsystem with enhancements

Secure Boot, Hardware Root-of-Trust

SME (Secure Memory Encryption)

SEV-ES (Secure Encrypted Virtualization & Register Encryption)

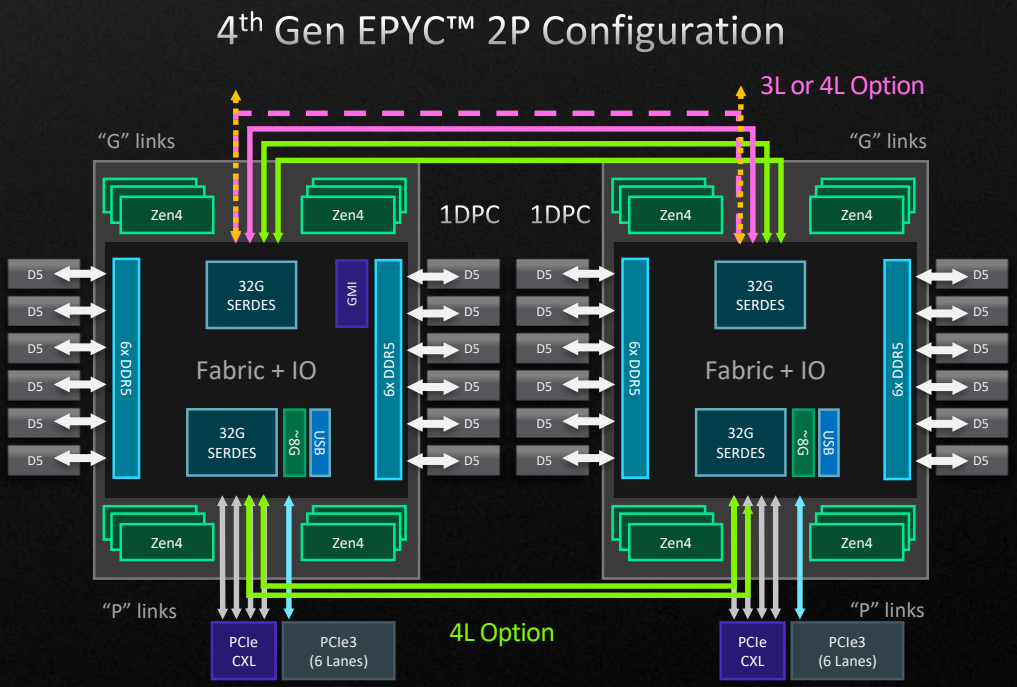
SEV-SNP (Secure Nested Paging), AES-256-XTS with more encrypted VMs

[1] With certain DIMM population rules.

* Post launch feature.

AMD Infinity Fabric™ Platform Capability

- Up to 32Gbps performance
- 3Link or 4Link Infinity Fabric platform options (“3G” or “4G” option)
  - 3Link: 160L + 12L / platform
  - 4Link: 128L + 12L / platform
- Additional 4L option with front/back connectivity (“2P + 2G” option)
- Platform BW scaling and flexibility for platform innovation



I/O Performance

Adv. Virtual Interrupt Controller (AVIC)

- Improve virtualized interrupt performance
- Inter-processor interrupts (IPI) and endpoint devices
- Can increase virtualized system performance
- Reduces virtualization overhead leaving more resources available for Guest usage

Improved interrupt processing throughput

- Within the CPU cores
- SOC-level throughput

>90% efficiency out-of-box[1] for 200Gbps NICs

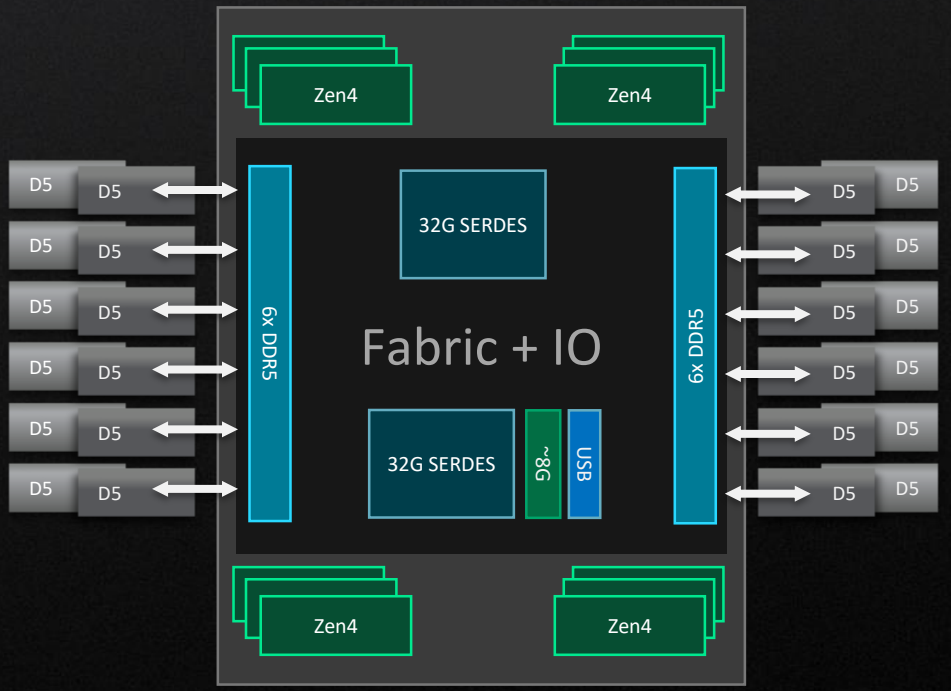
>90% efficiency for PCIe5 400Gbps InfiniBand

TEST		CONFIG	Uni-Directional (Gbps)		Bi-Directional (Gbps)	
			4th Gen EPYC™	Eff.	4th Gen EPYC™	Eff.
x16 Gen4 200 GbE	IOMMU Off	Out of Box	188	94%	375	94%
	IOMMU On	Out of Box	188	94%	367	92%
x16 Gen5 400 Gb InfiniBand	IOMMU Off	Std	396	99%	790	98%

1: “out-of-box”: MTU=1500, power mgmt. enabled, No PCIe Relaxed Ordering, no thread/IRQ affinity, NPS=1

4th Gen EPYC™ CPU Memory Capabilities

- 12ch DDR5/CPU; up-to-DDR4800
- 460GB/s peak theoretical BW (12ch * 8B * 4.8GTs)
- 1DIMM/ch and 2DIMM/ch capability
- x80 and x72 DIMMs
- RDIMM and 3DS RDIMM
- Up-to 6TB/socket capacity [1]
- AMD-C (x4 DRAMs) and “Bounded Fault” DRAM ECC
- Read UECC retry capability
- High BW and efficiency in both dual-rank and single-rank for DRAM capacity and system TCO optimization



1: 2x8R 3DS-RDIMM with 16Gbx4 Devices

Selected EPYC™ 9004 Power Management Features

Server systems and environments vary (Thermal headroom etc.) & Silicon varies

- **Faster / higher leakage parts**
- **Slower / lower leakage parts**

Some customers/environments desire deterministic performance

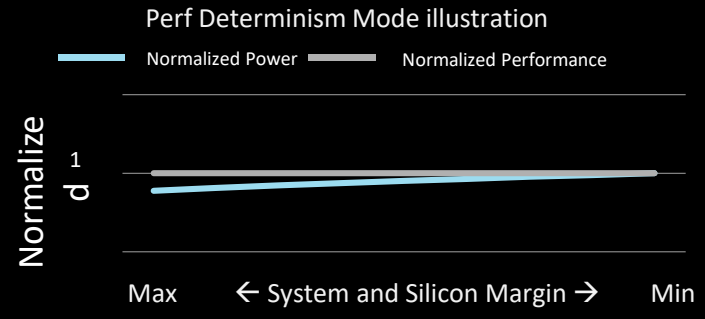
- **“Performance Determinism” mode: power consumption will vary part-to-part and based on environment**

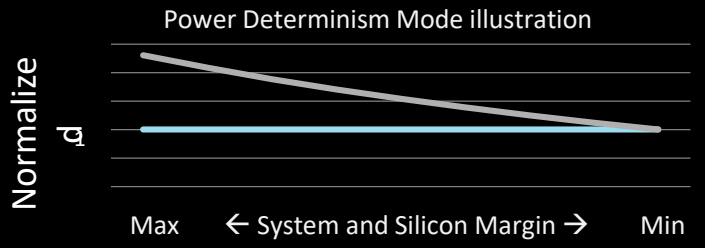
Some customers/environments desire max performance within platform limits

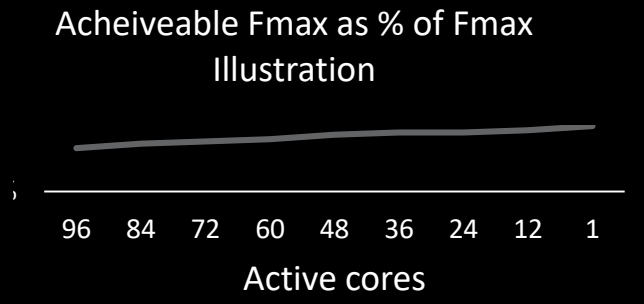
- **“Power Determinism” mode: performance varies**
- **Performance can be greater than or equal to “Performance Determinism”**
- **EPYC™ CPUs allow boot-time choice, commonly set to Performance Determinism**
- **EPYC™ also allows configurable TDP (BIOS) for TCO and peak performance tuning**

Peak Boost Behavior

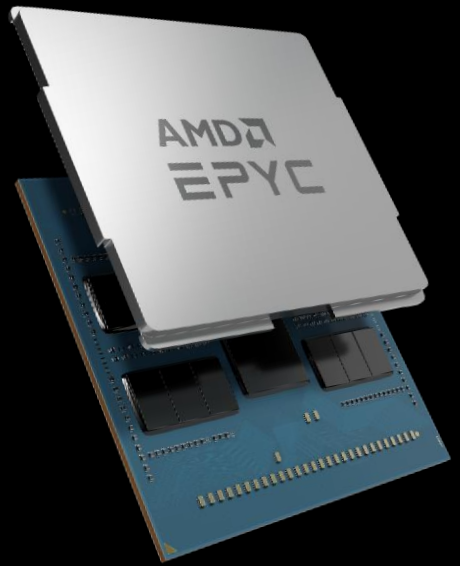
- **Maximum frequencies limited by active cores, cooling, silicon reliability**
- **Not limited based on instruction mix**
- **Internal hardware management keeps CPU within platform average and peak power delivery constraints**
- **Actual frequency varies with workload activity/performance heuristics**







AMD EPYC™ 9004 CPU MODELS



ALL-IN FEATURE SET INCLUDES

- 12 Channels of DDR5-4800
- 6TB memory capacity
- 128 lanes PCIe®5
- 64 I/O Lanes Support CXL™ 1.1
- 32gbps AMD Gen 3 Infinity Fabric™
- Flexible Topology Options
- Secure Memory Encryption*
- Secure Encrypted Virtualization*

CORES	AMD EPYC
96 CORES	9654/P
84 CORES	9634
64 CORES	9554/P 9534
48 CORES	9474F 9454/P
32 CORES	9374F 9354/P 9334
24 CORES	9274F 9254 9224
16 CORES	9174F 9124

"F" Performance Per Core Optimized

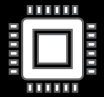
See endnotes: GD-183

EPYC™ 9004 Series CPU Positioning

	Processor Groups	
Core Performance High frequency with large cache/core ratio	9474F (48C-360W)	9374F (32C-320W)
	9274F (24C-320W)	9174F (16C-320W)
Core Density Highest core and thread count	9654/P (96C-360W)	9634 (84C-290W)
	9554/P (64C-360W)	9534 (64C-280W)
	9454/P (48C-290W)	
Balanced and Optimized Performance and TCO	9354/P (32C-280W)	9334 (32C-210W)
	9254 (24C-200W)	9224 (24C-200W)
	9124 (16C-200W)	

AMD INFINITY GUARD

HELPS MINIMIZE POTENTIAL ATTACK SURFACES AS SOFTWARE IS BOOTED AND EXECUTED AND PROCESSES YOUR CRITICAL DATA



AMD Secure Processor

A hardware root of trust which helps protect confidentiality and integrity of data with minor impact to system performance

Secure Memory Encryption

Industry's first full memory encryption.
AMD innovative technology helps defend data against certain cold boot and even physical attacks

Secure Encrypted Virtualization

Set of AMD technologies that help protect virtual machines with one of up to 509 unique encryption keys known only to the processor. Only available on AMD processors.

AMD Shadow Stack

Provides hardware-enforced stack protection capabilities to help guard against malware attacks.

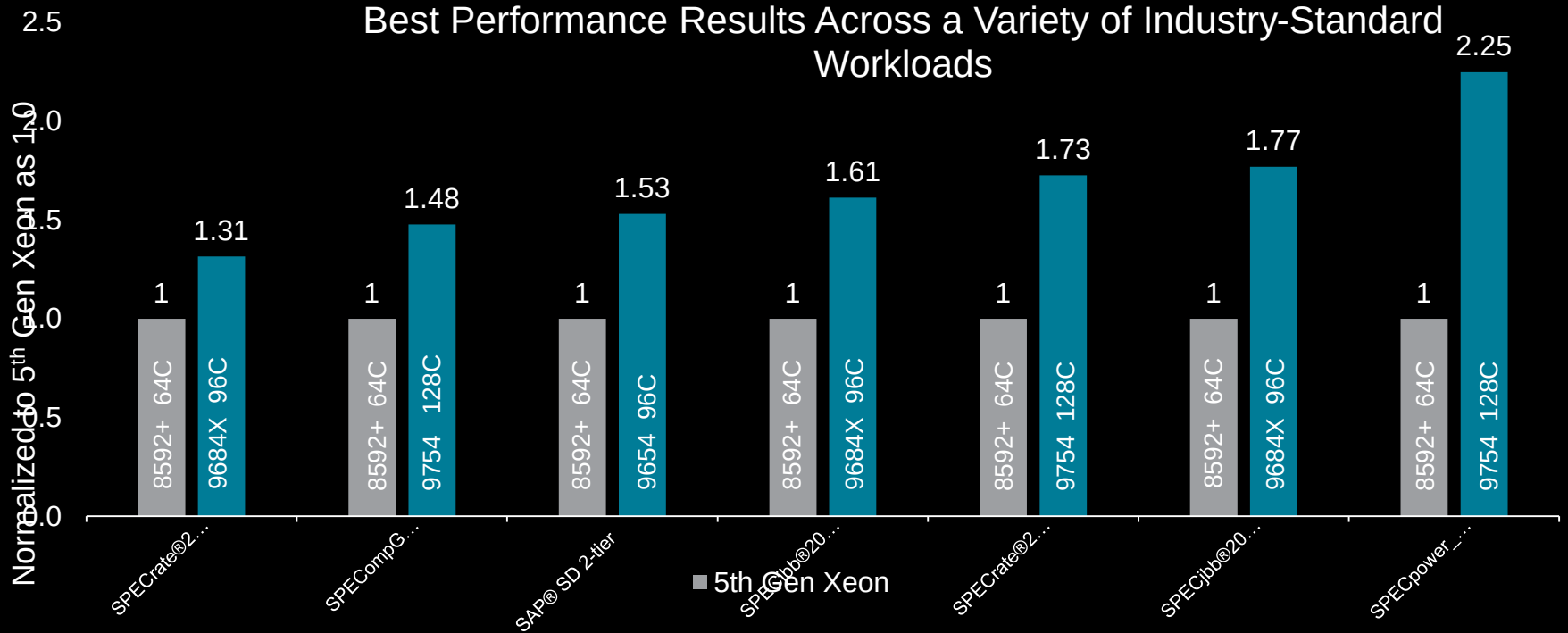
Help your organization ***TAKE CONTROL*** of security and ***DECREASE RISKS*** to your most important assets – Your Data

AMD Infinity Guard features vary by EPYC™ Processor generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard . GD-183

AMD | NUTANIX

4th Gen AMD EPYC™ CPUs

up to ***2.25x the performance*** of 5th Gen Intel® Xeon® CPUs



As of 1/4/2024. See endnotes SP5-011F, SP5-012E, SP5-056C, SP5-104B, SP5-064C, SP5-175A, SP5-235

AMD Tooling - Easy VMware Migration From Intel

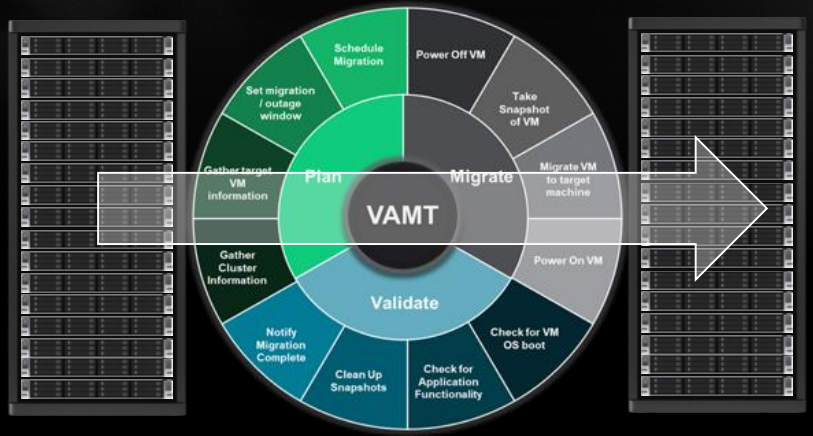
Fast Automated Cold Migration (VAMT)

- Automates cold migration; open sourced coding
- Co-developed with VMware professional services

+

Fully Supported Live Migration Across AMD Clusters

- Once workloads are on AMD clusters, live migration features are all fully supported and on par with Intel

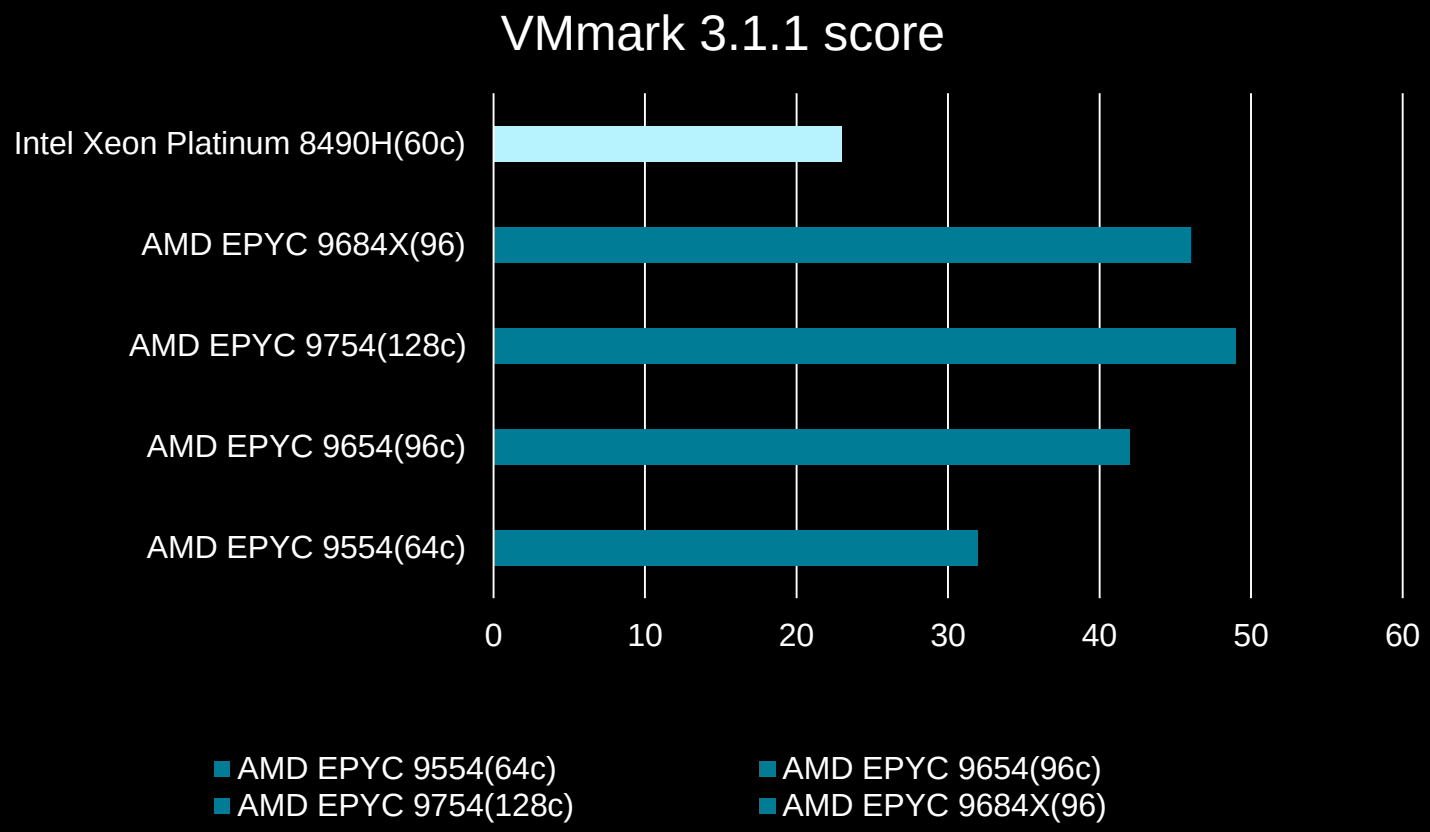


VMware Automated Migration Tool

Live Migration with-in AMD server cluster

All new x86 CPUs require cold migration (including Icelake to Sapphire) – AMD makes it easier

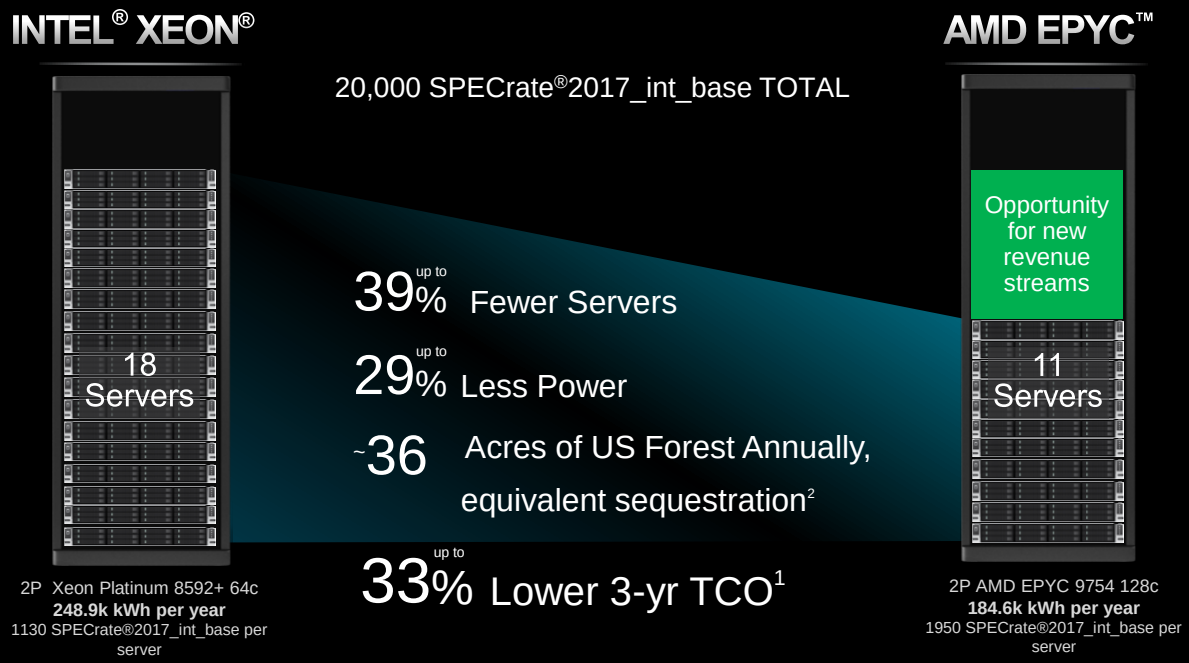
BEST PERFORMANCE FOR VIRTUALIZED APPLICATIONS



See: SP5-049D

AMD EPYC™ CPU Data Center Savings est

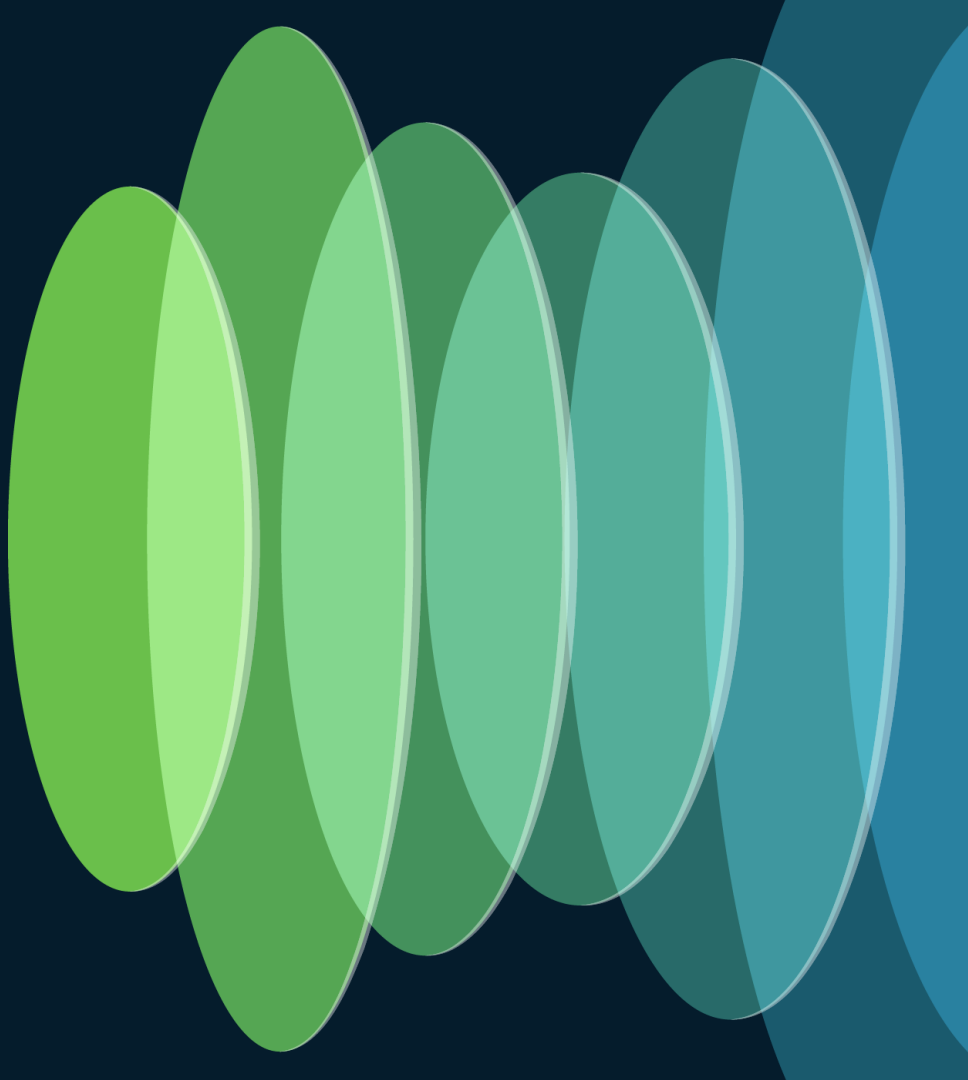
FEWER SERVERS, LESS POWER → LOWER EMISSIONS



1. As of 1/31/2024, see endnote SP5TCO-072. Analysis based on the AMD EPYC™ Bare Metal Server & Greenhouse Gas Emission TCO Estimation Tool - version 10.2. AMD processor pricing based on 1KU price as of Jan 2024. Intel® Xeon® Scalable CPU data and pricing from https://ark.intel.com as of Jan 2024. All pricing is in USD.
2. Values are for USA.

UCS and AMD

Joost van der Made
Compute TME



cisco Live!

UCS- Unified Compute and Core density

B200 M1/M2
Up to two Intel Xeon 5600
12 DIMMs, up to 192GB

B200 M3
Up to two Intel Xeon E5-2600
v3 24 DIMMs, up to 768 GB

B200 M4
Up to two Intel Xeon E5-2600
v3/v4 24 DIMMs, up to 1.5TB

B200 M5
Up to two Intel Xeon Scalable
24 DIMMs, up to 9TB

B200 M6
Up to two Intel Whitley Ice Lake
CPUs 32 DIMMs, up to 10TB

B200 M2 FCS: March 2010 EOL: March 2013

B200 M4: FCS: Oct 2014 EOL: Feb 2019

B200 M6: FCS: Apr 2021

2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024

B200 M1/M2

8-12 cores

B200 M3: FCS: Mar 2012, EOL: Dec 2016

B200 M3

24 Cores

B200 M4

44 Cores

B200 M5: FCS: Jul 2017

B200 M5

56 Cores

B200 M6

80 Cores

Single Chassis

8x B200M1:
64 Cores

UCS domain

20x chassis and 8x
B200M1: 1280 Cores

Cisco UCS X-Series Momentum

~$1.5B

Success

Modernization
Reduces total power consumption of M4 servers by up to 31%

51% Blade Market Share
Continued market share lead

2000+
Unique customers

+115%
Y/Y Growth

56,000+
Servers

54%
M7 X-Series Mix of Total UCS

25%
New/dormant customers

5%
High-capacity local storage

1300+
Production customers

3,362
Offer Registrations

34,000+
Servers on-line

Innovations

X-Series Sustainability Benefits
~50% fewer raw materials vs 3 generations of rack servers

5th Gen Fabric
Strong 100G adoption (60% Mix on X-series)

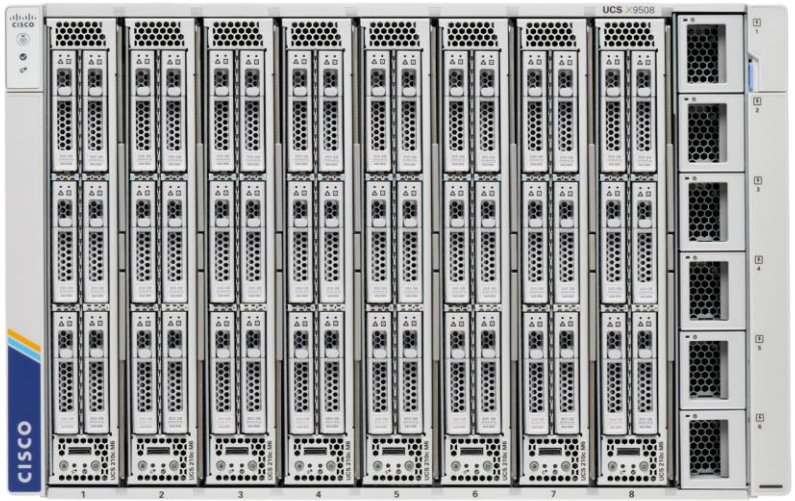
GPU on X-Series
Wide range of GPUs being adopted for rack W/L

X-Direct and AMD Blade
Demo at Clive, orderability, EFT

UCS X-Series Chassis with 8x AMD Node M8

UCS X-Series with AMD M8

AMD



Up to 2048
Cores per Chassis (M8)

2030+

- Liquid-cooling
- Silicon photonics
- Future fabrics ready
- Power-hungry processing

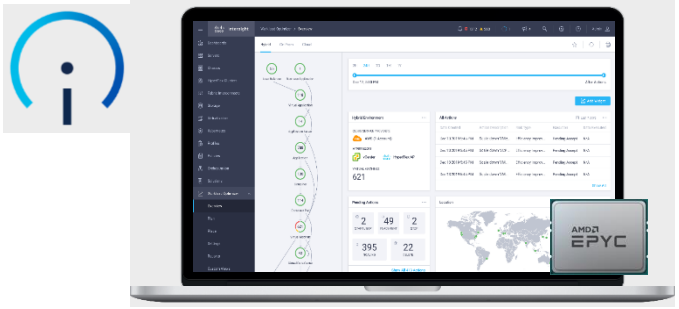
2021

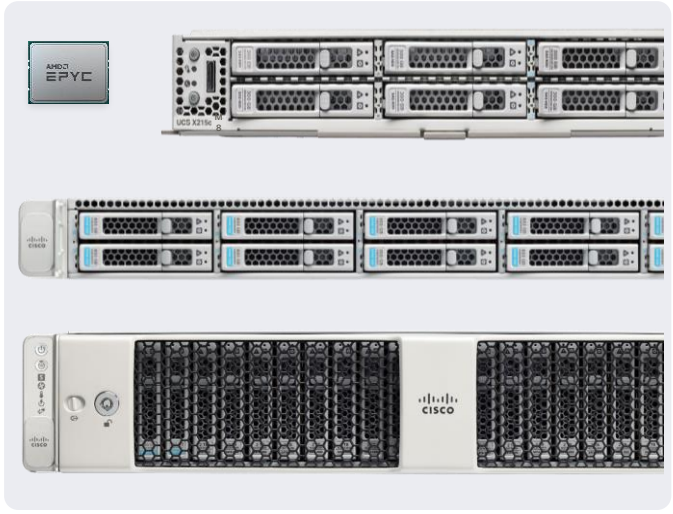
UCS and AMD Intersight Features

Every AMD CPU feature is easily enabled and configured through policy-based service profiles with Intersight. Examples include:

- Secure Memory Encryption
- Secure Encrypted Virtualization
- Visibility to all RAS diagnostics
- Compatible with multi-processor and multi-vendor environments

Cisco Intersight





UCS X215C M8

UCS 225 M6/M8

UCS 245 M6/M8

X-Series portfolio

Compute

X210c Compute Node
- 2- Socket, single slot servers
- Two Generations: M6 and M7
- Intel 3rd Gen. (Ice Lake) and 4th Gen & 5th gen Xeon CPUs

X410c Compute Node
- 4- Socket, dual slot servers
- Intel 4th Gen Xeon CPU
- Up to 64 DDR5 DIMMs

X215c Compute Node
- 2- Socket, single slot servers
- M8 with AMD 4th gen EPYC CPU

Fabric

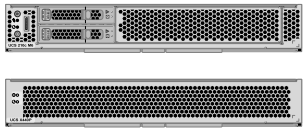
4th and 5th Gen FI
- 25/100G ports
- unified ports – up to 16x 32G FC ports (6536)
- Supports VIC 1400, 14000 and 15000 series

25/100G IFM
8 x 25/100G connectivity

4th and 5th Gen VIC
25/100G connectivity for both blades and racks.

X-Fabric and PCIe node

X-Fabric
- Based on native PCIe Gen. 4
- Provides GPU acceleration to enterprise application
- No backplane or cables = Easy upgrades



GPU Node and Front Mezz GPUs
Nvidia A16, Nvidia L40, Nvidia L4 and Nvidia H100 GPUs today in various configurations.

UCS X215c M8 X-series Server Specs Overview

Use Cases	Enterprise-class mainstream blade server with high performance for compute-intensive workloads.	
	• Virtualization • Hyperconverged • Mixed Workload Standardization	• Database and Analytics • Virtual Desktop Infrastructure • AI/ML
Core Platform	Dual Socket Platform Support for 4th Gen AMD EPYC CPU Support for 24xDDR5 DIMMs (1DPC, 12 channel) PCIe Gen 5 support and CXL 1.1+ support	
X Fabric/PCIe Node	Connect to PCIe nodes, only Nvidia GPUs	
VICS	200G aggregate/100G per fabric, 100G5th gen VIC mLOM or 25G 5th gen VIC mLOM, 25G 5th Gen VIC Mezz	
Drives/GPU	Front mezz options | 6 SAS/SATA with HW RAID, 6 NVMe PCIe Gen5 x4, up to 2x GPUs Internal | 2x M.2 SATA HW RAID1, 2x M.2 NVMe	
Management	Intersight Management and UCSM	

UCS X215c M8 2S Compute Node – Key features

CPU

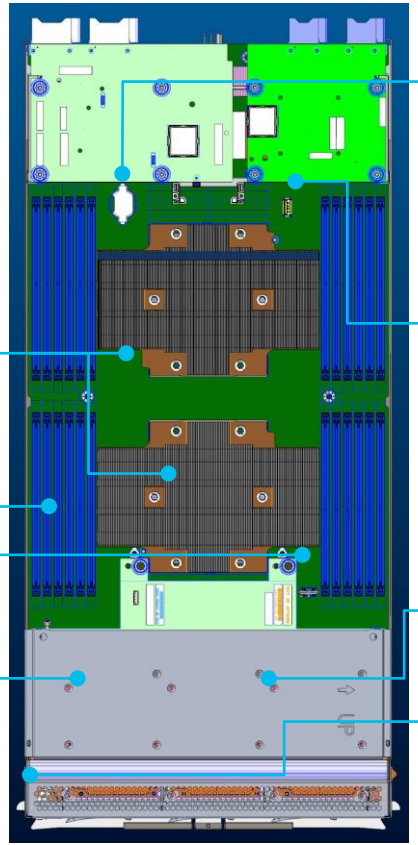
Up to 2 AMD 4th Gen EPYC CPUs

Memory

Up to 24x DDR5 4800 DIMMs

GPU

Up to 2 x PCIe GPU
+
Up to 2 x NVMe PCIe Drives



X-Fabric Mezz

Connect to PCIe Node

In future, to Memory/Drive Nodes

VIC

Up to 2 VICs

15230 100G Cisco mLoM*

15420 25G Cisco mLoM*

15422 25G Cisco VIC Mezz*

Agg NW 200G per node

*With Secure Boot FW

Drives

Up to 6 x SAS/SATA/NVMe

- RAID Controller for SAS/SATA
- 15TB NVMe drives (800K IOPs, 6GB/s)
- Up to 90TB (15TB x 6) per Blade
- 2 x M.2 SATA HW Raid1/ 2 x M.2 NVMe

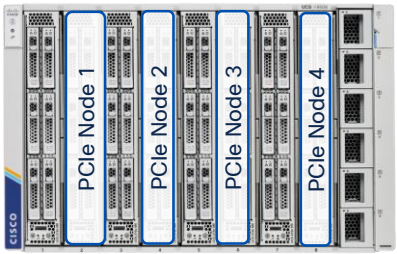
UCS X-Fabric Technology and PCIe Nodes with GPU

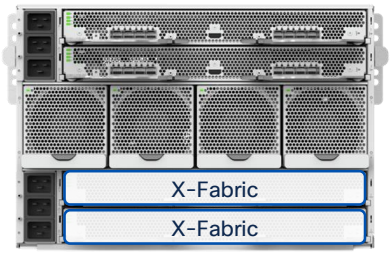
PCIe node supports up to:

2x
Nvidia A16

4x
Nvidia L4

2x
Nvidia H100
Nvidia L40
Nvidia L40S
Nvidia H100 NVL *





X-Fabric decouples the lifecycles of CPU, GPU, memory, storage and fabrics - providing a perpetual architecture that efficiently brings you the latest innovations.

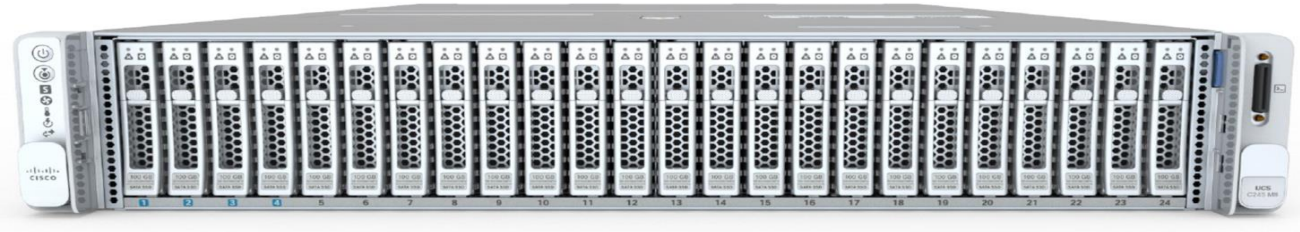
- Cloud-powered composability with Cisco Intersight
- Flexible GPU acceleration across server nodes
- No backplane or cables = easily upgrades

* Will be available Q4 CY24

Announcing the Cisco UCS C245 M8 Rack Server

Exceptional performance for enterprise workloads, including big data analytics, collaboration, databases, virtualization, and high-performance applications

Up to 28 SFF HDD/SSD/NVMe drive support
Up to 24 drives in the front and
optional 4 drives in rear



Cisco Intersight
and IMM support

PCIe 5.0
OCP 3.0 NIC
VIC 15000 Series

Support for all 4th Gen
EPYC CPU SKU's
Up to 256 Cores

DDR5 4800 MHz DIMMs
24 DIMM slots
Up to 6TB of Memory

Announcing the Cisco UCS C225 M8 Rack Server

This high-density, 1RU, single-socket rack server delivers industry-leading performance and efficiency for a wide range of workloads, including virtualization, collaboration, and bare-metal applications.

Up to 10 SFF HDD/SSD/NVMe drive support

Cisco Intersight
and IMM support

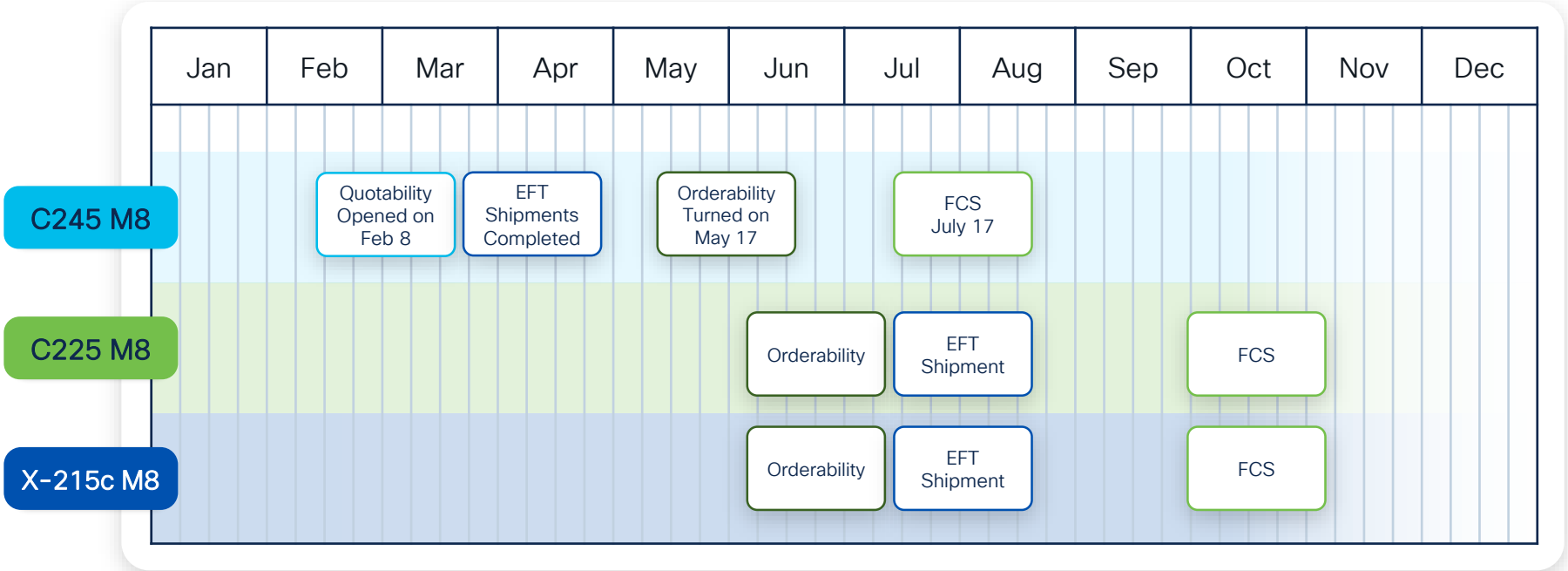
PCIe 5.0
OCP 3.0 NIC
VIC 15000 Series

Support for all 4th Gen
EPYC CPU SKU's
Up to 128 Cores

DDR5 4800 MHz DIMMs
12 DIMM slots
Up to 3TB of Memory

AMD M8 Schedule

CY2024 View



Key Takeaway

High Performance

AMD drives high performance and high density leading to better out comes for consolidation, modern workload demand etc. including AI with the advantage for precision computing usecases on 128+ core cpu

UCS X-Series and AMD

There is not better system to adopt new AMD EYPC CPUs than Cisco UCSX-series. Best-in-class density +2000 cores in 7RU, 40,000 cores in a single domain. Intersight standard operating model, enact all AMD features like infinity guard from Intersight.

NOW

The time is now. We are launching orderability and have a strong supply chain. We have opened the door to unlock all the advantages of AMD EPYC with simplicity, sustainability, and performance

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to **win 1 of 5 full conference passes** to Cisco Live 2025.

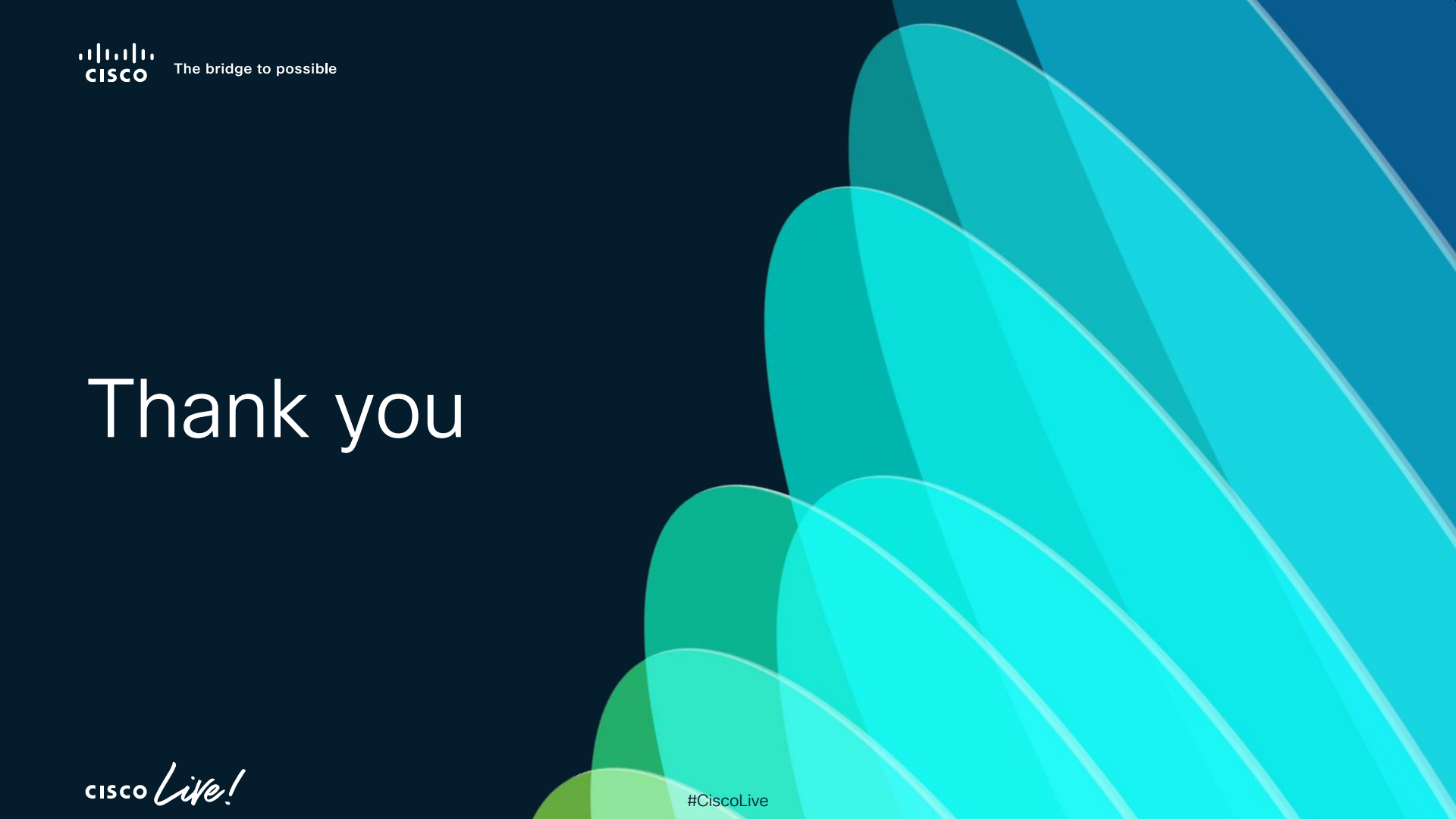
Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn **exclusive prizes!**

Complete your surveys in the **Cisco Live mobile app.**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand


The bridge to possible
Thank you
cisco Live!
#CiscoLive

AI Technologies Bring Broad Industry Impact

Extending and enriching industry workloads and activities

Machine Learning

- Personalization and pricing optimization,
- Improve fraud and cyber threat detection
- Process and operations automation/optimization

Generative AI

- Accelerate research and time to insights
- Frictionless human-machine communication
- Automated transcription and summarization

Recommendation Systems

- Offer targeted product suggestions based on user patterns
- Combine media to optimize consumer interaction and retention
- Enhance customer service responsiveness/accuracy

Streaming and Gaming

Public Safety

Manufacturing

Retail

Financial Services

Medical

Service Automation