



Connect

# Make Room for AI – Through Datacenter Virtualization Density

Steve Bassett

*Server Business Unit, AMD*



Red Hat



# **AMD Introduction**

# AMD AI Platforms

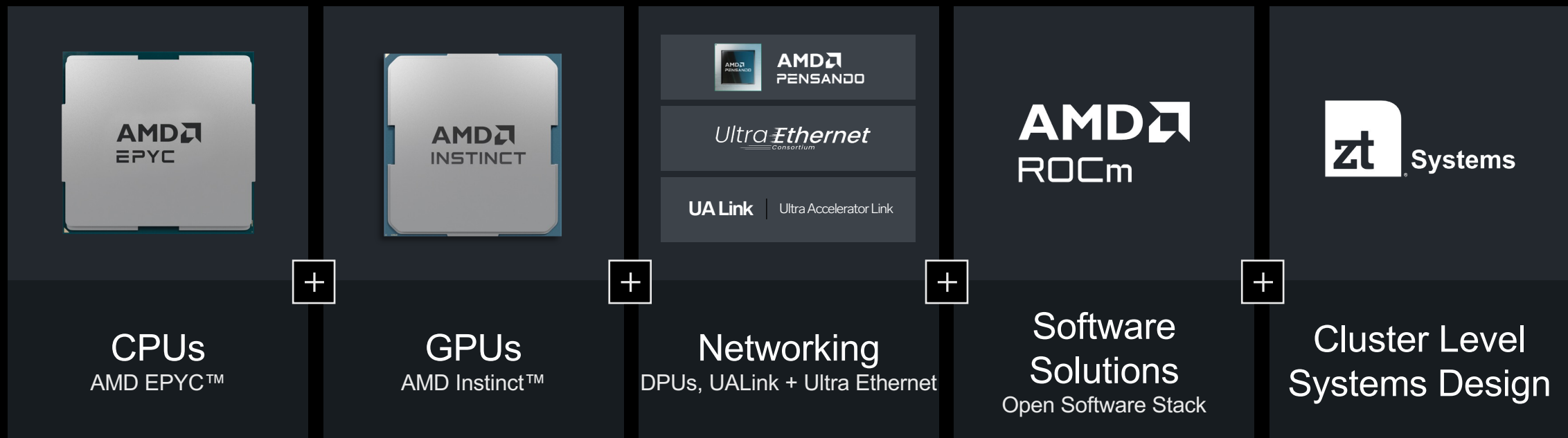
Unmatched  
portfolio of training  
and inference  
compute engines

Open software  
solutions

AI ecosystem  
with deep  
co-innovation

Cluster level  
systems design

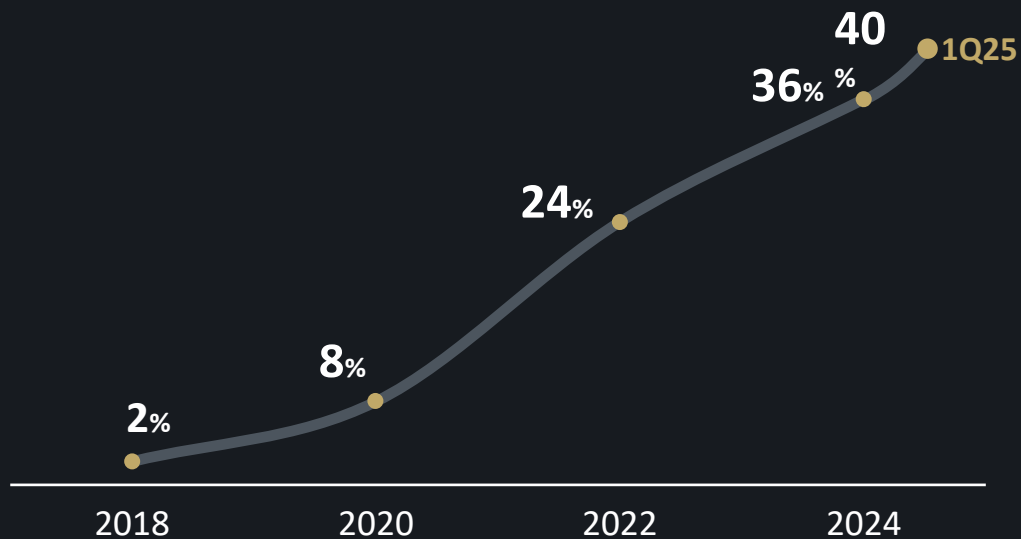
# AMD | Advancing the AI Data Center



# AMD EPYC™ Momentum Accelerates...

Oct. 6, 2025: AMD and OpenAI Announce Strategic Partnership  
to Deploy 6 Gigawatts of AMD GPUs

## >18x Server CPU Market Share Growth



## Industry Leaders Run on EPYC™

### Cloud

aws Microsoft Google ORACLE

### Digital

NETFLIX Uber ∞ Meta zoom

### Enterprise

BEST BUY IBM Emirates NBD NISSAN

### OEM

DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO CISCO

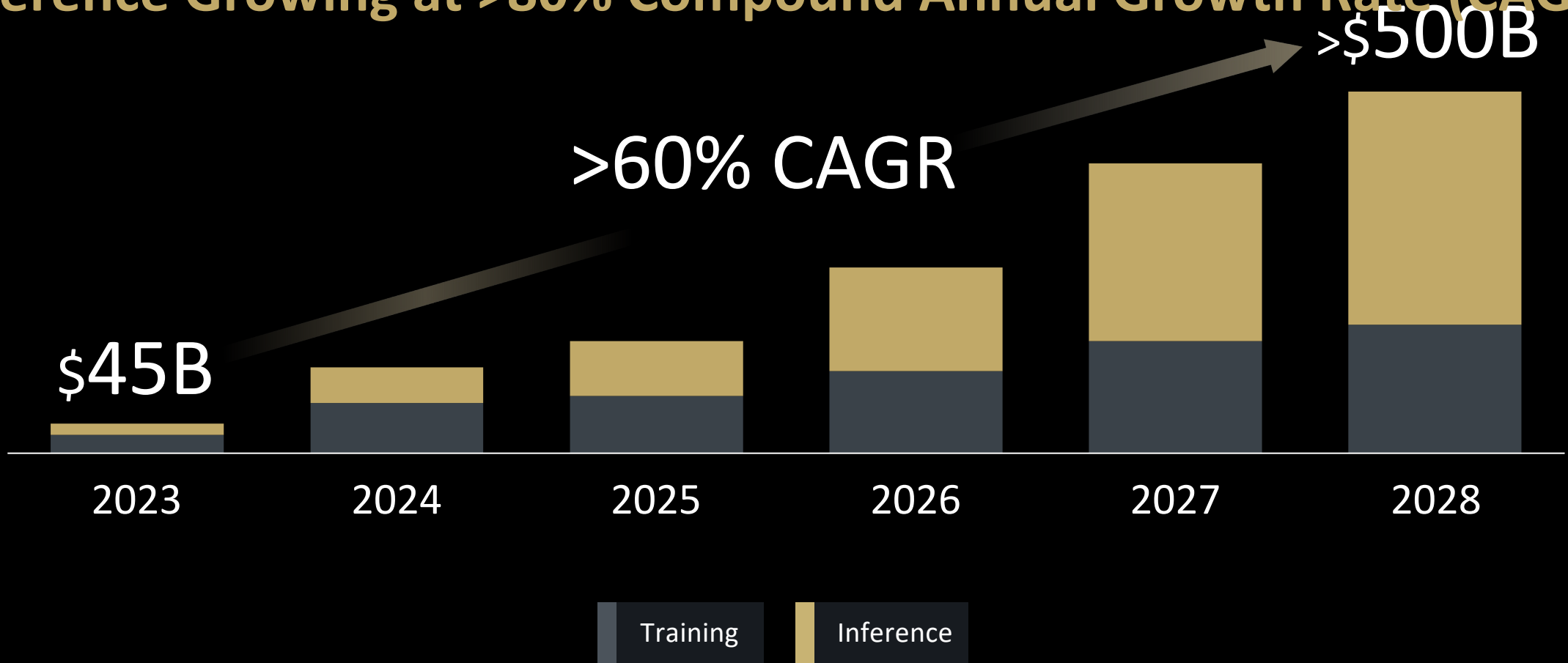
# The Challenge of AI

# Generative AI is the next technology revolution

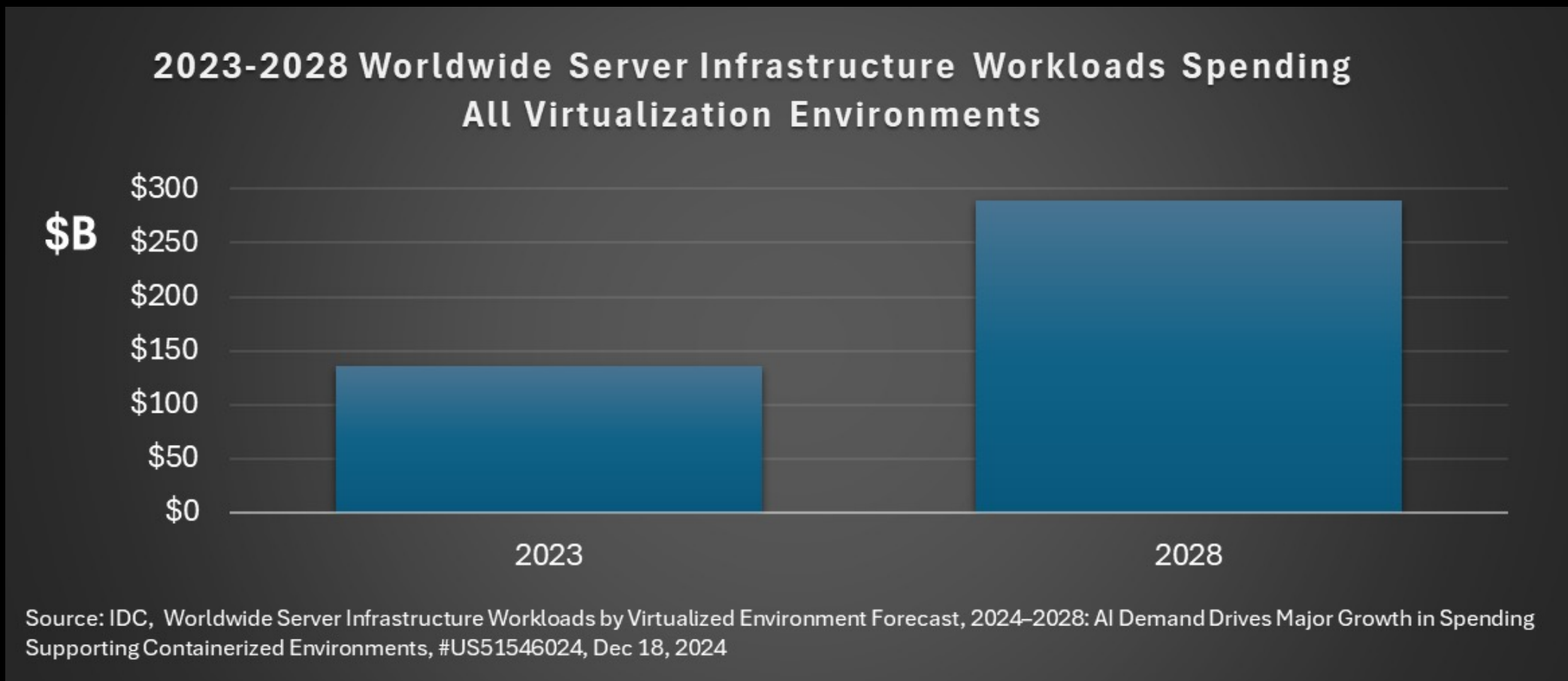
- AI is the enabler of the next generation of corporate productivity gains
- Today: experimental wave of public-cloud and hyperscaler-enabled generative AI
- Emerging: The next wave will be and is enterprise implementation for productivity

# Data Center AI Accelerator Total Addressable Market (TAM)

Inference Growing at >80% Compound Annual Growth Rate (CAGR)



# IDC Server Infrastructure Software Platforms Spending Forecast



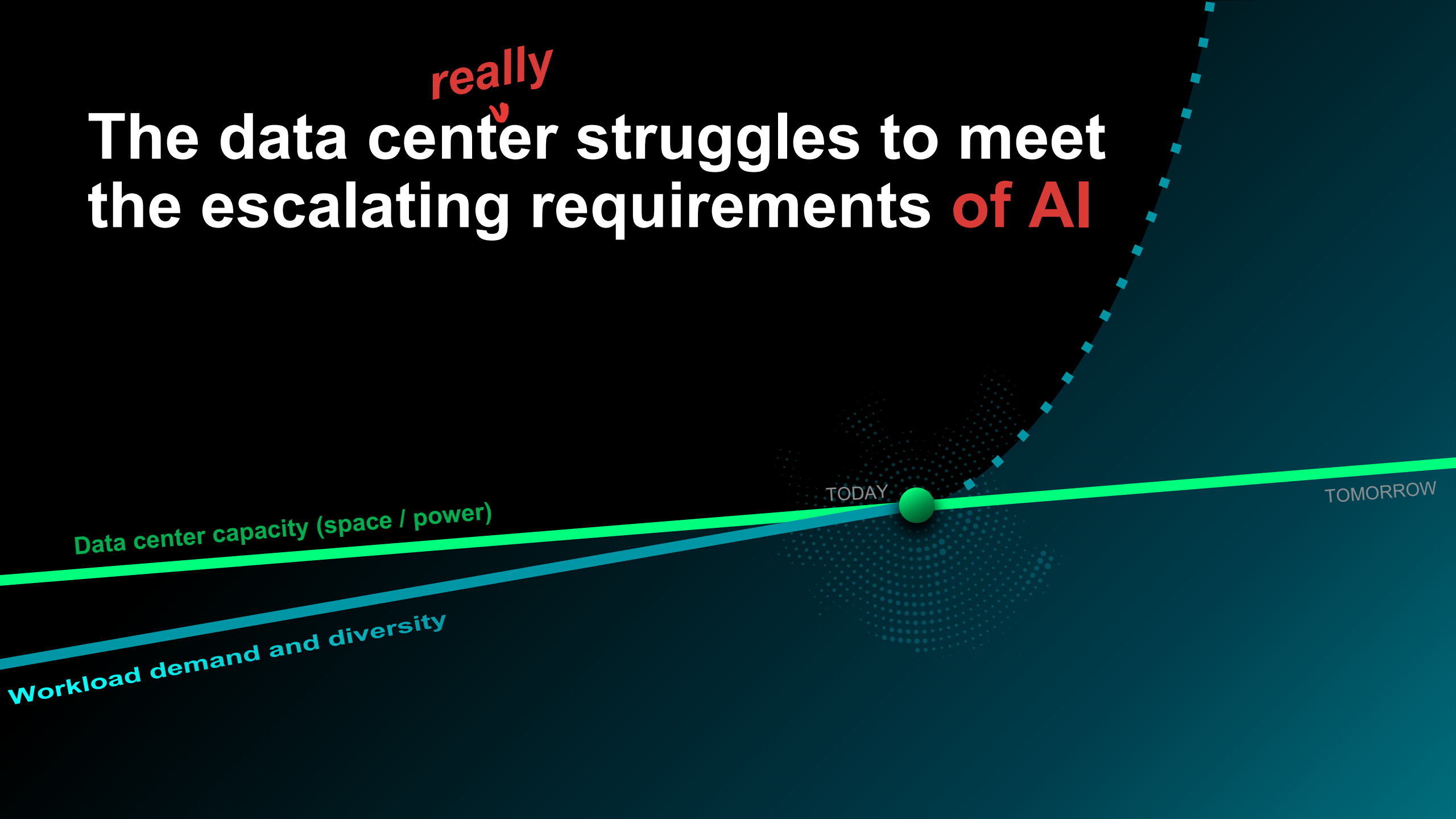
- Dominant applications driving growth across the platforms
  - Data management – AI lifecycle, Text and media analytics, Unstructured database
  - Business-specific and Engineering/technical applications

# **Making Room for AI**

# The data center struggles to meet the escalating requirements



# The data center <sup>really</sup> struggles to meet the escalating requirements <sup>of</sup> AI



# The solution to our challenge is mostly in today's data centers

An isometric illustration of a data center floor. The floor is covered with a grid of blue squares. Several rows of server racks are arranged in a perspective view, receding into the distance. Each rack is a dark blue rectangular block with many small, lighter blue squares on its front face, representing individual server units. The racks are arranged in a staggered pattern, with some rows being further back than others, creating a sense of depth.

The average data center size worldwide is 100,000 square feet.<sup>1</sup>

Much of it is dedicated to old, inefficient and hard-to-manage equipment.<sup>2</sup>

North American data center free capacity is down to <3%.<sup>3</sup>

1. <https://www.datacenters.com/news/and-the-title-of-the-largest-data-center-in-the-world-and-largest-data-center-in>

2. Analysis based on AMD internal data.

3. <https://www.networkworld.com/article/3695582/us-data-center-market-nears-full-capacity.html>

Step 1 –



Consolidate VMs and Containers

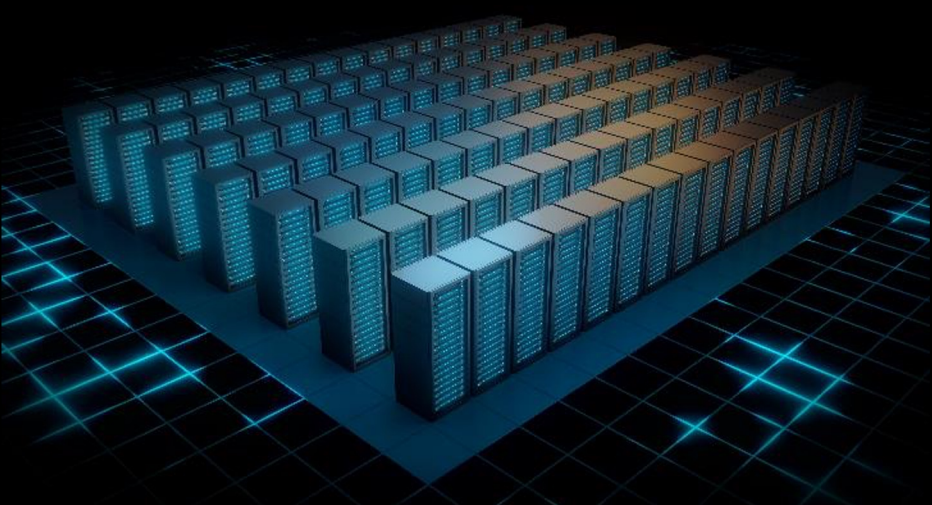
Free up datacenter space and power

# 5<sup>th</sup> Gen AMD EPYC™ CPUs

Refresh and Modernize your data center – Add more capacity for your compute needs

**1000 Legacy Servers**

2P Intel® Xeon® Platinum 8280 servers



**127 Modern Servers**

2P AMD EPYC™ 9965 servers



Easy to migrate to AMD

- x86 architecture
- Mature ecosystem
- Robust tools

Up to **69%**  
Less power

Up to **87%**  
Fewer Servers

Up to **79%**  
Lower 5-yr TCO

Servers required to achieve a total of 391,000 SPECrate®2017\_int\_base performance score.  
See endnotes 9xx5TCO-005

# DBS TRANSFORMS ITS DATA CENTER WITH AMD EPYC™ CPUS

“When we moved from our traditional infrastructure to the new virtualized commodity server-based one, we reduced the cost by 75 percent.”

“[With AMD EPYC servers], our power consumption reduced by 50 percent. But we had ten times the capacity to grow.”

Choon Boon Tan, Managing Director and Head of Cloud Engineering & Services at DBS



## CHALLENGES

DBS Bank Ltd wanted to accelerate its digital transformation with technology at the core, to provide greater resiliency, improve sustainability, enable faster release cadences and lower costs. To enable this required a switch from monolithic systems to wholesale virtualization across most workloads, particularly the next generation of cloud, machine learning, and AI.

## SOLUTION

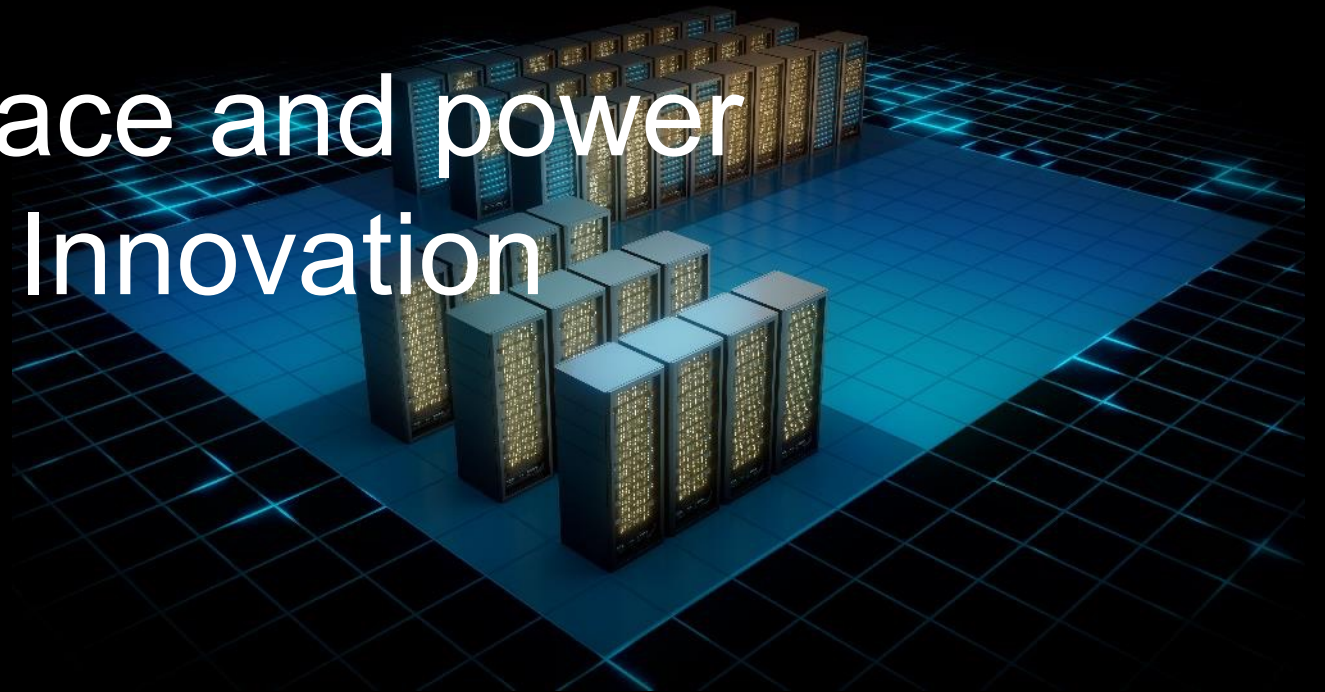
The DBS computing infrastructure was transferred from premium highly resilient systems to technologies such as the Dell PowerEdge R7425 and R6525 servers powered by **AMD EPYC™ 7542, 7642 and 7742 processors**, running VMware virtualization, open source software, and aggressive automation from Day 0 provisioning to Day 2 operations.

## RESULTS

The footprint of a DBS data center was reduced to a quarter of its size in square feet, **consuming half the power and providing a tenfold increase in room for growth**. Coupled with wide adoption of open-source software and aggressive automation, cloud infrastructure services can now be provisioned in a matter of minutes instead of months.

Step 1 –  
Consolidate VMs and Containers  
Free up datacenter space and power

Step 2 - Invest space and power  
in AI and Innovation



# **Datacenter consolidation**

**Making room for AI with Red Hat® OpenShift®  
on Servers using AMD EPYC™ CPUs**

The Datacenter Virtualization and Consolidation Solution

# 5th Gen AMD EPYC™

World's best CPU for Cloud, Enterprise and AI



3nm  
4nm

150 billion  
transistors

Up to 192 cores  
384  
threads

17% IPC uplift\*  
Full AVX512

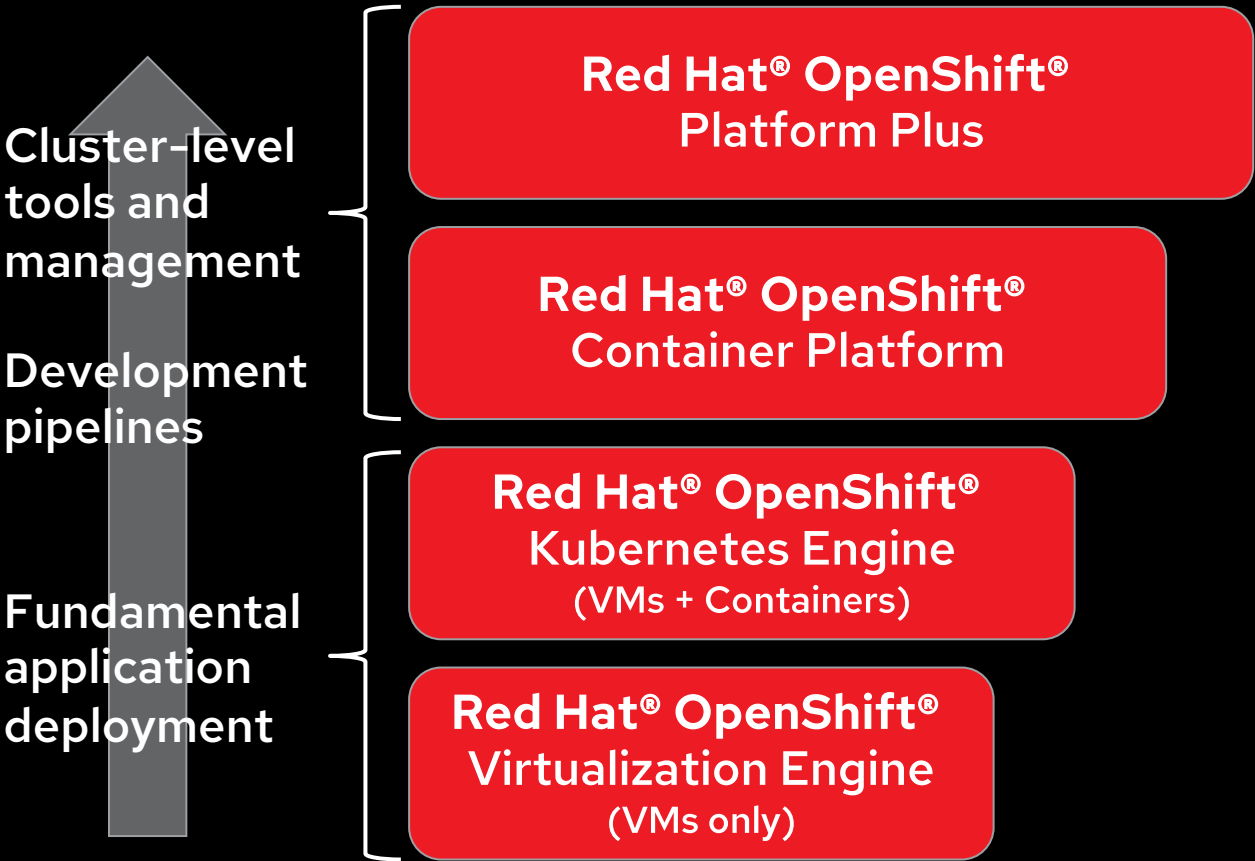
Up to 5  
GHz

\*~17% Across 36 cloud and enterprise workloads see endnote 9xx5-001

As of 10/1/2024. See endnotes EPYC-029C

# Red Hat® OpenShift® platforms

- Enterprise-grade, secure, commercial Kubernetes distributions
- Containers, VMs, and unified application deployment
- Commercial, multi-cluster, hybrid-cloud management tools



# The tools for datacenter transformation are here today

HPE  
ProLiant



Dell  
PowerEdge



Lenovo  
ThinkSystem



Supermicro A+



Cisco UCS



Red Hat® OpenShift®

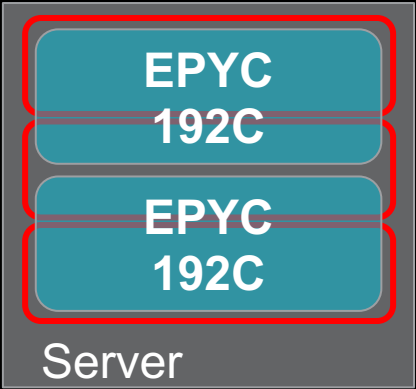
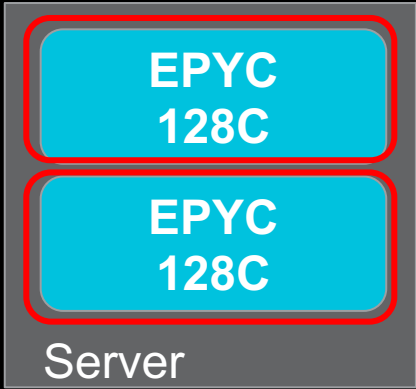
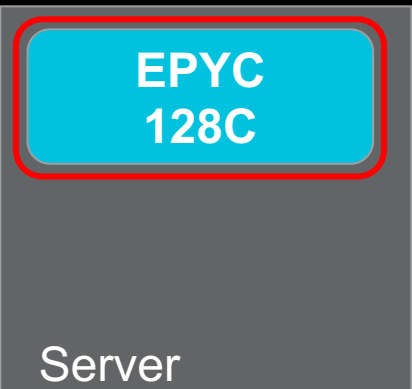
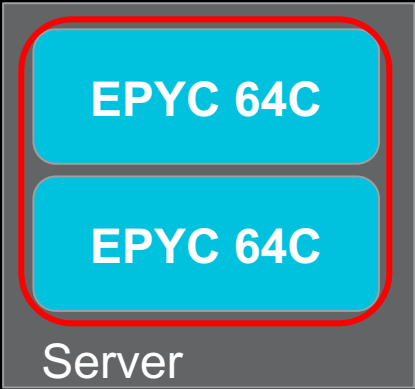
4th and 5th Gen AMD EPYC™ Series

# Aligning Red Hat® OpenShift® Bare-Metal Subscriptions with AMD EPYC™ CPU-based Server Configurations

*High-Performance  
VMs*

*Containers and VMs  
High Density      - - - -      Ultra Density*

Configuration  
Cases



Stacking  
Red Hat OpenShift  
Subscriptions

1

1

2

3

*Each subscription is fully utilized in 128-core increments*

# Consolidation with 5<sup>th</sup> Gen AMD EPYC™ CPUs

Modernize your data center – Focus on VM migration with Red Hat® OpenShift® Virtualization Engine

## 1000 Legacy Servers

2P Intel® Xeon® Gold 6226R servers  
VMware® vSphere Standard per core

Refresh

## 150 Modern Servers

2P AMD EPYC™ 9555 servers  
Red Hat® OpenShift® Virtualization Engine per node

6 to 1

Easy to migrate to AMD

- x86 architecture
- Mature ecosystem
- Proven support

Up to **85%**  
Fewer Servers

Up to **64%**  
Less power

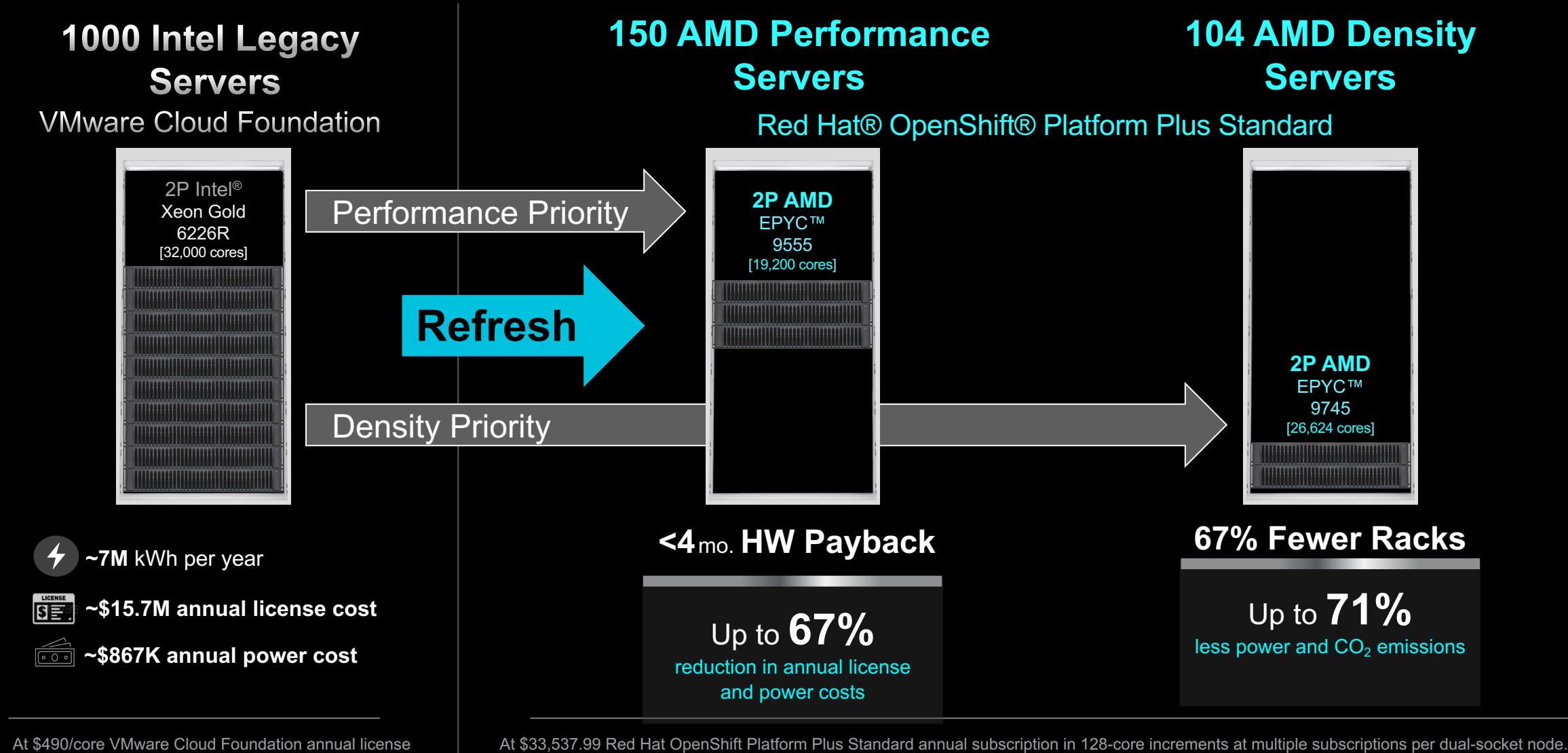
Up to **~80%**  
Lower licensing cost

Up to **60%**  
Lower 5-yr TCO

Servers required to achieve a total of 252,000 SPECrate®2017\_int\_base performance score. VMware vSphere Standard modeled at \$90/core for a 1-year license. Red Hat OpenShift Virtualization Engine Premium subscriptions modeled at \$3,294 per 128-core increment per node per one year. See endnotes 9xx5TCO-016

# Red Hat® OpenShift® on AMD EPYC™ CPUs

Unifying VMs with Containers when Refreshing and Modernizing a Datacenter



Clusters at equivalent SPEC CPU® 2017 Integer Rate base aggregate performance.  
See Endnotes 9xx5TCO-014, 9xx5TCO-015

# DEFENSE AGENCIES MODERNIZE INFRASTRUCTURE WITH AMD EPYC™ CPUS

“All of the defense agencies have noted the exceptional performance and unwavering reliability of Cisco UCS C-Series servers [powered by AMD EPYC processors].”

Cisco



## CHALLENGES


A defense agency needed to modernize its information technology provision to reduce server and cabling footprints, alongside simplifying infrastructure management. Another agency also needed an easier path to infrastructure scaling while minimizing operating costs inside a limited physical data center footprint.

## SOLUTION

The defense agencies deployed Cisco® Unified Computing System™ (Cisco UCS®) C-Series Multinode Rack Servers, powered by AMD EPYC™ processors to maximize core density with Cisco UCS Manager to simplify administration. Cisco Intersight™ Private Virtual Appliance was deployed by another defense agency to simplify and centralize server management.

## RESULTS

One defense agency was able to consolidate 20 racks of gear down to just one single rack vastly reducing power consumption, while another expanded 24 nodes to 40 in a single rack, with improved performance allowing reduced hypervisor licensing costs. This in turn enabled a reduction from 200 cables down to 80 cables.



From  
A Legacy Datacenter  
With No Space or Power Left

A 3D digital illustration of a server room. In the center, several rows of server racks are arranged, each glowing with a warm, golden-yellow light. The racks are set on a large, flat surface that features a glowing blue grid pattern, resembling a floor or a data plane. The background is dark, with some faint, out-of-focus light spots, giving the impression of a vast, high-tech environment. The overall aesthetic is futuristic and technological.

# To Room and Power for AI



**Red Hat® OpenShift®**

**Red Hat® OpenShift® AI**

Leading OEM Datacenter and Edge Server Products

**5<sup>th</sup> Gen AMD EPYC™ CPUs**

**AMD Instinct™ Accelerators**





Connect

# Grazie



[linkedin.com/company/red-hat](https://linkedin.com/company/red-hat)



[facebook.com/redhatinc](https://facebook.com/redhatinc)



[youtube.com/user/RedHatVideos](https://youtube.com/user/RedHatVideos)



[twitter.com/RedHat](https://twitter.com/RedHat)



# Endnotes

SANTA CLARA, Calif., Oct. 06, 2025 (GLOBE NEWSWIRE) – AMD (NASDAQ: AMD) and OpenAI today announced a 6 gigawatt agreement to power OpenAI’s next-generation AI infrastructure across multiple generations of AMD Instinct GPUs. The first 1 gigawatt deployment of AMD Instinct MI450 GPUs is set to begin in the second half of 2026.. <https://www.amd.com/en/newsroom/press-releases/2025-10-6-amd-and-openai-announce-strategic-partnership-to-d.html>

9xx5-001: Based on AMD internal testing as of 9/10/2024, geomean performance improvement (IPC) at fixed-frequency. 5th Gen EPYC CPU Enterprise and Cloud Server Workloads generational IPC Uplift of 1.170x (geomean) using a select set of 36 workloads and is the geomean of estimated scores for total and all subsets of SPECrate@2017\_int\_base (geomean), estimated scores for total and all subsets of SPECrate@2017\_fp\_base (geomean), scores for Server Side Java multi instance max ops/sec, representative Cloud Server workloads (geomean), and representative Enterprise server workloads (geomean). “Genoa” Config (all NPS1): EPYC 9654 BIOS TQZ1005D 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-4800 (2Rx4 64GB), 32Gbps xGMI; “Turin” config (all NPS1): EPYC 9V45 BIOS RVOT1000F 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-6000 (2Rx4 64GB), 32Gbps xGMI Utilizing Performance Determinism and the Performance governor on Ubuntu@ 22.04 w/ 6.8.0-40-generic kernel OS for all workloads. 5th Gen EPYC generational ML/HPC Server Workloads IPC Uplift of 1.369x (geomean) using a select set of 24 workloads and is the geomean of representative ML Server Workloads (geomean), and representative HPC Server Workloads (geomean). “Genoa” Config (all NPS1) “Genoa” config: EPYC 9654 BIOS TQZ1005D 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-4800 (2Rx4 64GB), 32Gbps xGMI; “Turin” config (all NPS1): EPYC 9V45 BIOS RVOT1000F 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-6000 (2Rx4 64GB), 32Gbps xGMI Utilizing Performance Determinism and the Performance governor on Ubuntu 22.04 w/ 6.8.0-40-generic kernel OS for all workloads except LAMMPS, HPCG, NAMD, OpenFOAM, Gromacs which utilize 24.04 w/ 6.8.0-40-generic kernel. SPEC® and SPECrate® are registered trademarks for Standard Performance Evaluation Corporation. Learn more at [spec.org](https://spec.org).

9xx5TCO-005 This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The AMD Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool - version 1.3, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL\_PERFORMANCE of 391000 units of SPECrate@2017\_int\_base performance as of November 21, 2024. This estimation compares upgrading from a legacy 2P Intel Xeon 28 core Platinum\_8280 based server with a score of 391 (<https://spec.org/cpu2017/results/res2020q3/cpu2017-20200915-23984.pdf>) versus 2P EPYC 9965 (192C) powered server with a score of 3100 (<https://spec.org/cpu2017/results/res2024q4/cpu2017-20241004-44979.pdf>) Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from Country Specific Electricity Factors - 2024, and the United States Environmental Protection Agency Greenhouse Gas Equivalencies Calculator.

For additional details, see <https://www.amd.com/en/claims/epyc.html#q=9xx5TCO-005>.

9xx5TCO-014: As of May 13, 2025, this scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a Target Performance Metric of ~252000 units of integer performance based on the published scores (or estimated if indicated by an asterisk) for Intel Xeon and AMD EPYC CPU based servers. This estimation reflects a 5 year time frame. Only power and software virtualization costs contribute to OPEX. This analysis compares a 2P 128 core AMD EPYC\_9745 Server with a SPECrate2017\_int\_base score of 2440, <https://spec.org/cpu2017/results/res2025q1/cpu2017-20241230-45854.pdf>; compared to a 2P 16 core legacy Intel Gold\_6226R Server with a SPECrate2017\_int\_base score of 252, <https://spec.org/cpu2017/results/res2020q3/cpu2017-20200731-23591.pdf> For additional details, see [9xx5TCO-014](#)

9xx5TCO-015 – As of May 13, 2025, this scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a Target Performance Metric of ~252000 units of integer performance based on the published scores (or estimated if indicated by an asterisk) for Intel Xeon and AMD EPYC CPU based servers. This estimation reflects a 5 year time frame. Only power costs contribute to OPEX. This analysis compares a 2P 64 core AMD EPYC\_9555 Server with a SPECrate2017\_int\_base score of 1690, <https://spec.org/cpu2017/results/res2025q1/cpu2017-20250205-46214.pdf>; compared to a 2P 16 core Intel Gold\_6226R Server with a SPECrate2017\_int\_base score of 252, <https://spec.org/cpu2017/results/res2020q3/cpu2017-20200731-23591.pdf> For additional details, see [9xx5TCO-015](#)

# Endnotes

9xx5-015: Llama3.1-8B (BF16, max sequence length 1024) training testing results based on AMD internal testing as of 09/05/2024. Llama3.1-8B configurations: Max Sequence length 1024, BF16, Docker: huggingface/transformers-pytorch-gpu:latest 2P AMD EPYC 9575F (128 Total Cores ) with 8x NVIDIA H100 80GB HBM3, 1.5TB 24x64GB DDR5-6000, 1.0 Gbps 3TB Micron\_9300\_MTFDHAL3T8TDP NVMe®, BIOS T20240805173113 (Determinism=Power,SR-IOV=On), Ubuntu 22.04.3 LTS, kernel=5.15.0-117-generic (mitigations=off, cpupower frequency-set -g performance, cpupower idle-set -d 2, echo 3> /proc/sys/vm/drop\_caches) , For 31.79 Train Samples/Second2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1TB 16x64GB DDR5-5600, 3.2TB Dell Ent NVMe® PM1735a MU, Ubuntu 22.04.3 LTS, kernel-5.15.0-118-generic, (processor.max\_cstate=1, intel\_idle.max\_cstate=0 mitigations=off, cpupower frequency-set -g performance ) , BIOS 2.1, (Maximum performance, SR-IOV=On), For 27.74 Train Samples/SecondFor average throughput increase of 1.146. Results may vary due to factors including system configurations, software versions and BIOS settings.

9xx5-059A: Stable Diffusion XL v2 training results based on AMD internal testing as of 10/10/2024. SDXL configurations: DeepSpeed 0.14.0, TP8 Parallel, FP8, batch size 24, results in seconds 2P AMD EPYC 9575F (128 Total Cores) with 8x AMD Instinct MI300X-NPS1-SPX-192GB-750W, GPU Interconnectivity XGMI, ROCm™ 6.2.0-66, 2304GB 24x96GB DDR5-6000, BIOS 1.0 (power determinism = off), Ubuntu® 22.04.4 LTS, kernel 5.15.0-72-generic, 334.80 seconds 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x AMD Instinct MI300X-NPS1-SPX-192GB-750, GPU Interconnectivity XGMI, ROCm 6.2.0-66, 2048GB 32x64GB DDR5-4400, BIOS 2.0.4, (power determinism = off), Ubuntu 22.04.4 LTS, kernel 5.15.0-72-generic, 400.43 seconds For 19.600% training performance increase. Results may vary due to factors including system configurations, software versions and BIOS settings.

EPYC-029D: Comparison based on thread density, performance, features, process technology and built-in security features of currently shipping servers as of 10/10/2024. EPYC 9005 series CPUs offer the highest thread density, leads the industry with 500+ performance world records including world record enterprise leadership Java® ops/sec performance, top HPC leadership with floating-point throughput performance, AI end-to-end performance with TPCx-AI performance and highest energy efficiency scores. Compared to 5th Gen Xeon, the 5th Gen EPYC series also has more DDR5 memory channels with more memory bandwidth and supports more PCIe® Gen5 lanes for I/O throughput, and has up to 5x the L3 cache/core for faster data access. The EPYC 9005 series uses advanced 3-4nm technology, and offers Secure Memory Encryption + Secure Encrypted Virtualization (SEV) + SEV Encrypted State + SEV-Secure Nested Paging security features. For additional details, see <https://www.amd.com/en/legal/claims/epyc.html#q=epyc5#EPYC-029D>

# Disclaimers and attributions

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u

©2025 Advanced Micro Devices, Inc. all rights reserved. AMD, the AMD arrow, EPYC, and combinations thereof are trademarks of Advanced Micro Devices, Inc Intel, the Intel logo and Xeon are trademarks of Intel Corporation or its subsidiaries. NVIDIA, the NVIDIA logo, are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. PCIe is a registered trademark of PCI-SIG Corporation. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See [www.spec.org](http://www.spec.org) for more information. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

Red Hat, Red Hat Enterprise Linux and Red Hat OpenShift are registered trademarks of Red Hat, Inc.