

PCクラスタワークショップ in 大阪2023

第4世代AMD EPYC™プロセッサ
新ラインナップ (Genoa-X, Bergamo) の
技術概要

2023年6月21日
日本AMD株式会社

大原久樹 (Hisaki.Ohara@amd.com)

AMD 
together we advance_

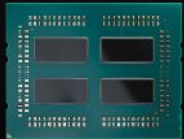
本日の内容

-
1. AMD EPYC™プロセッサのこれまでのロードマップ
 2. Bergamo アーキテクチャ
 3. Genoa-X アーキテクチャ
 4. AIへの取り組みについて

AMD EPYC™ JOURNEY

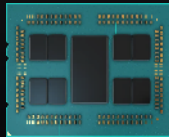
Four Generations of On-Time Execution

1st Gen EPYC™



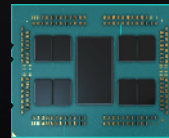
General Purpose
“Naples”

2nd Gen EPYC™

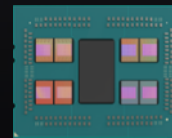


General Purpose
“Rome”

3rd Gen EPYC™

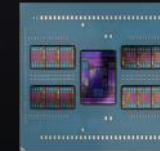


General Purpose
“Milan”



Technical
“Milan-X”

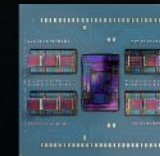
4th Gen EPYC™



General Purpose
“Genoa”



Cloud Native
“Bergamo”



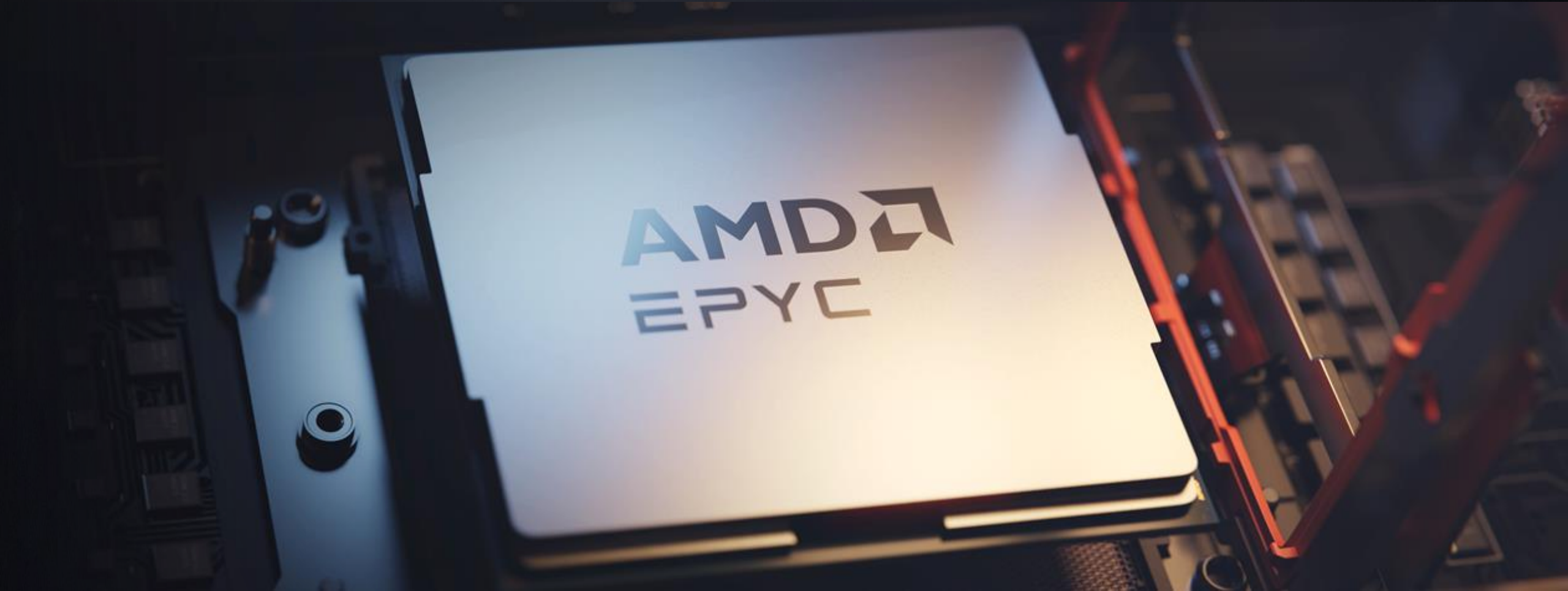
Technical
“Genoa-X”



Telco/Edge
“Siena”

2017

2023



AMD
EPYC

4th Gen AMD EPYC™ CPU “Genoa”



96 Cores

5nm
Process Technology

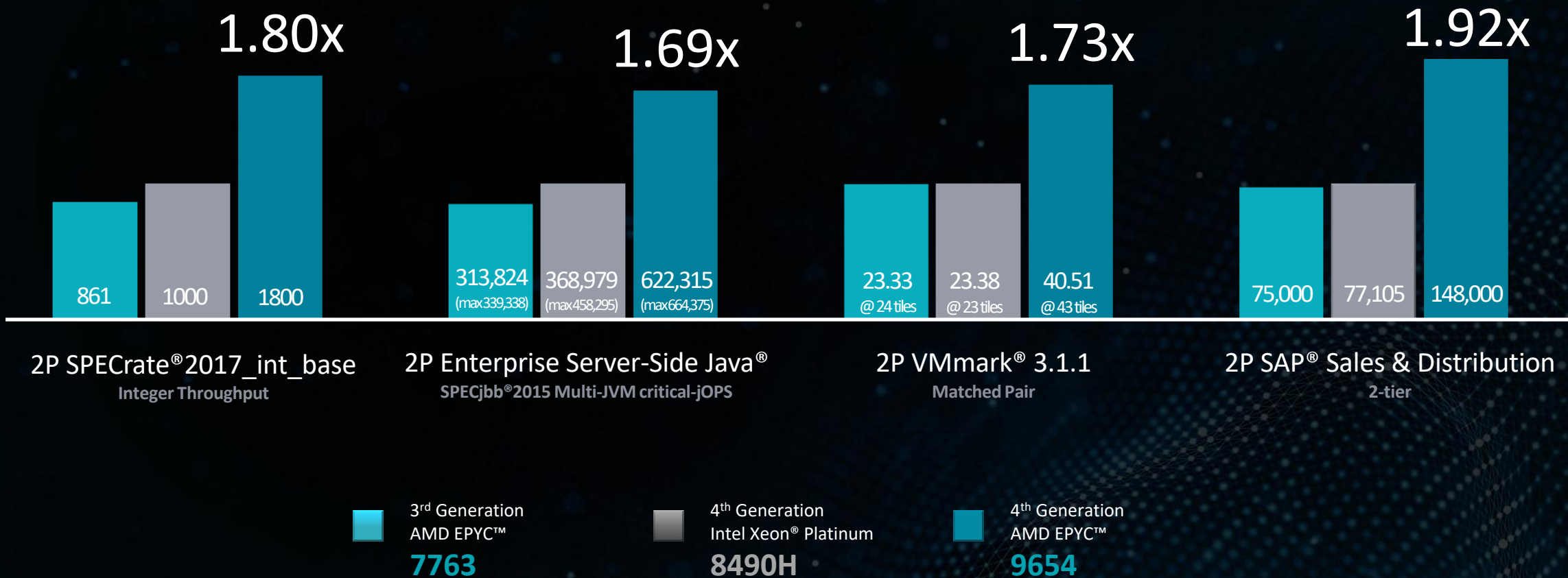
PCIe® 5.0

CXL™
Memory Expansion

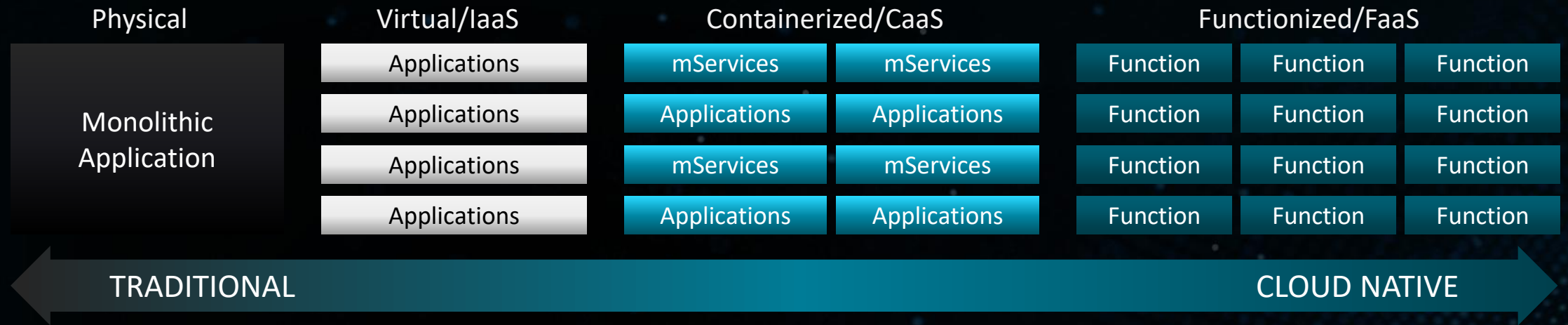
12 Channels
DDR5

第4世代AMD EPYC™ CPU

性能の優位性



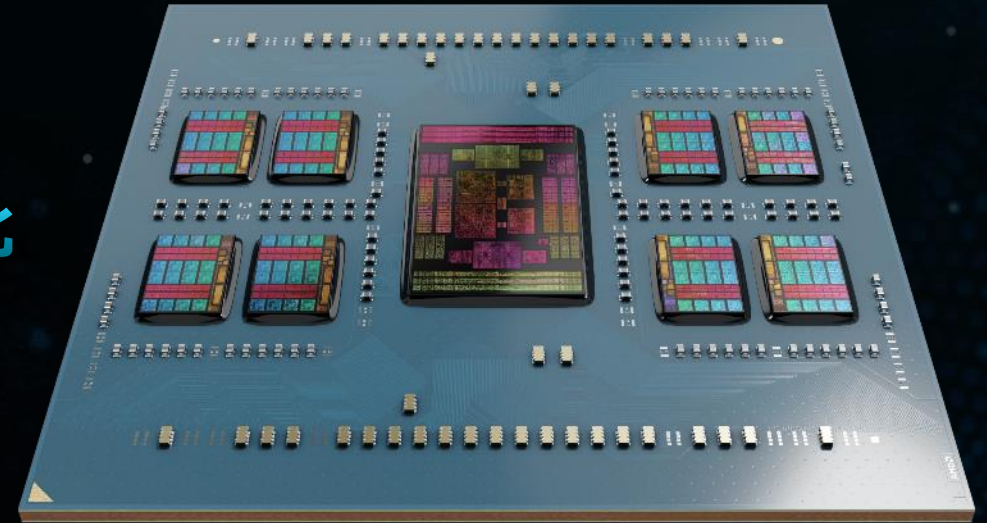
クラウドコンピューティングの特性



4TH Gen AMD EPYC™

97X4 CPU

クラウドネイティブなワークロードに最適化



Greatest vCPU
Density

Leadership Cloud
Performance

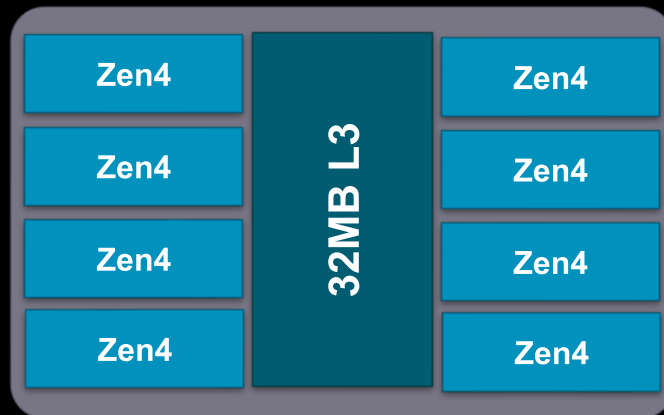
Best Energy
Efficiency

Consistent x86
ISA

Up to 128
“Zen 4c” Cores

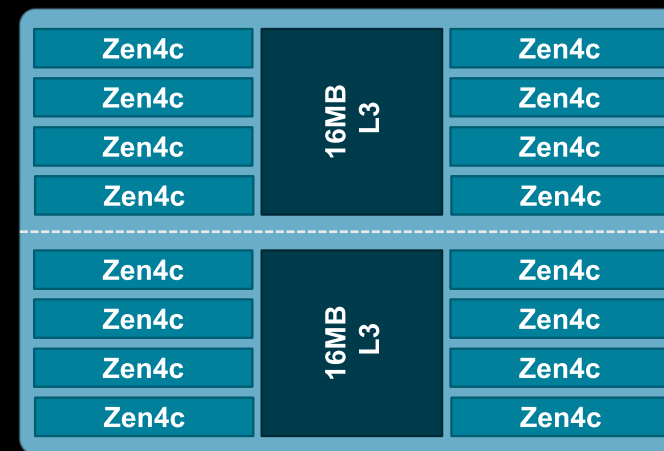
チップレットを活用した柔軟なアーキテクチャー

“Genoa” チップレット
Zen4



- コア性能最適化
- コアあたり1MB L2 キャッシュ
- 8コアで32MB L3キャッシュを共有

“Bergamo” チップレット
Zen4c



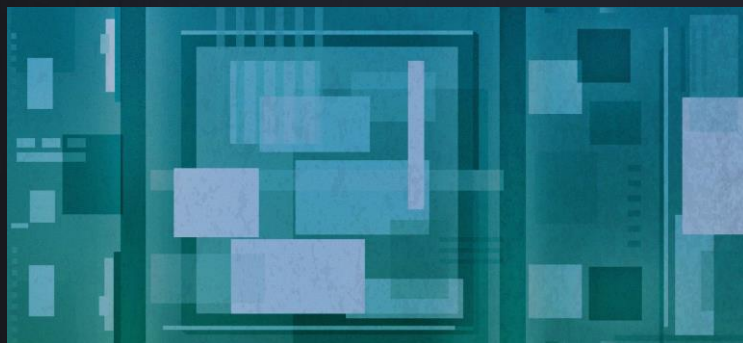
- 消費電力あたり性能に最適化
- コアあたり1MB L2 キャッシュ
- 8コアで16MB L3キャッシュを共有

論理的に同等(同じ命令セット、L1 & L2 キャッシュサイズ)

“Zen 4c” cloud native core architecture

Designed for **density** and **power efficiency**

“Zen 4” core



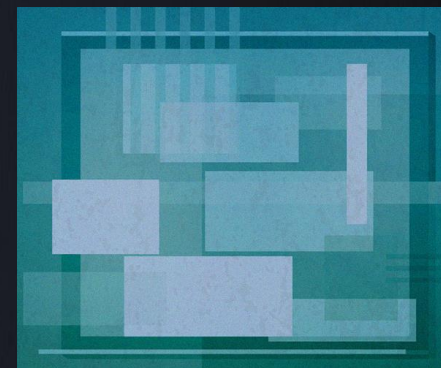
Node

TSMC 5nm

Core + L2 Area

3.84 mm²

“Zen 4c” core



Node

TSMC 5nm

Core + L2 Area

2.48 mm²

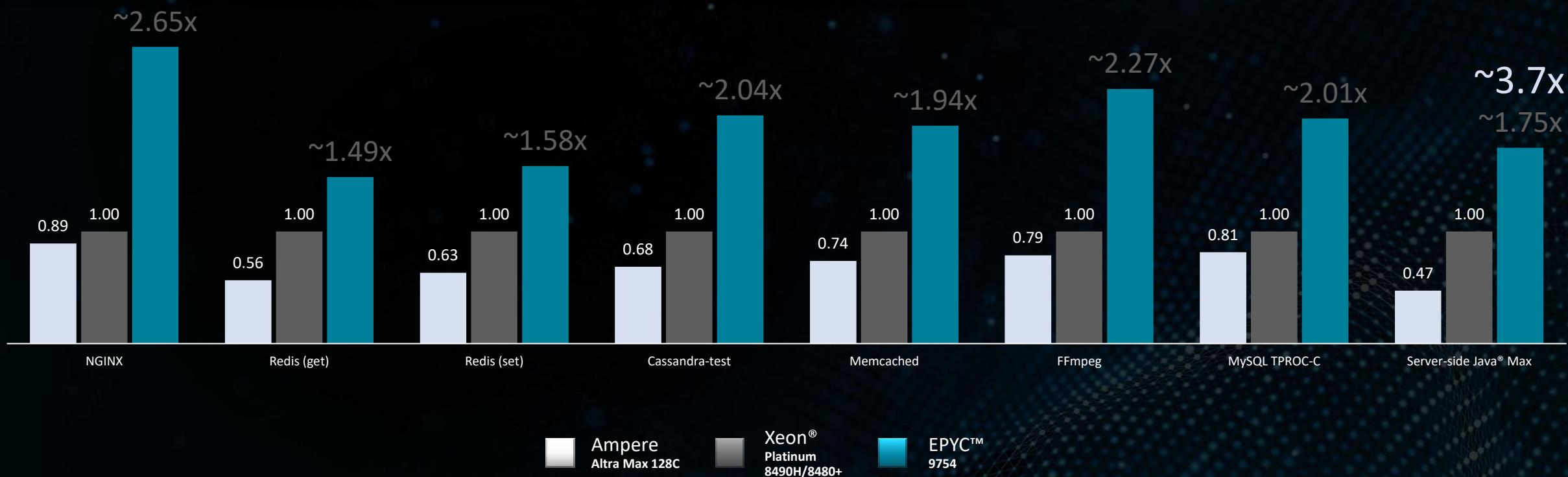
~**35%** smaller core

クラウドネイティブ・ワークロードの性能比較

最大

3.7x
vs ampere

クラウドネイティブな様々なワークロードにおいて
スループット性能 (平均 ~2.9x)



2P servers: 128C AMD EPYC™ 9754 vs. Ampere Altra® Max M128-30 and 56C/60C Intel Xeon Platinum 8480+/8490H

Results may vary due to factors including system configurations, software versions and BIOS settings. As of 6/13/2023, see Cloud Native Workloads <https://www.amd.com/system/files/documents/amd-epyc-9004-pb-cloud-native-workloads.pdf>.

AMD
together we advance.

配置可能なコンテナ数の比較

最大

3x 配置可能なコンテナ数

Each Container Delivering ~25K JOPS/sec meeting SLAs using Kubernetes

10 containers

14 containers

30 containers

3x

vs. Altra Max M128-30

~2.1x

vs. Xeon Platinum 8490H

Equivalent Java® Ops Per Sec Per Container

Ampere
Altra Max 128C

Xeon®
Platinum 8490H

EPYC™
9754

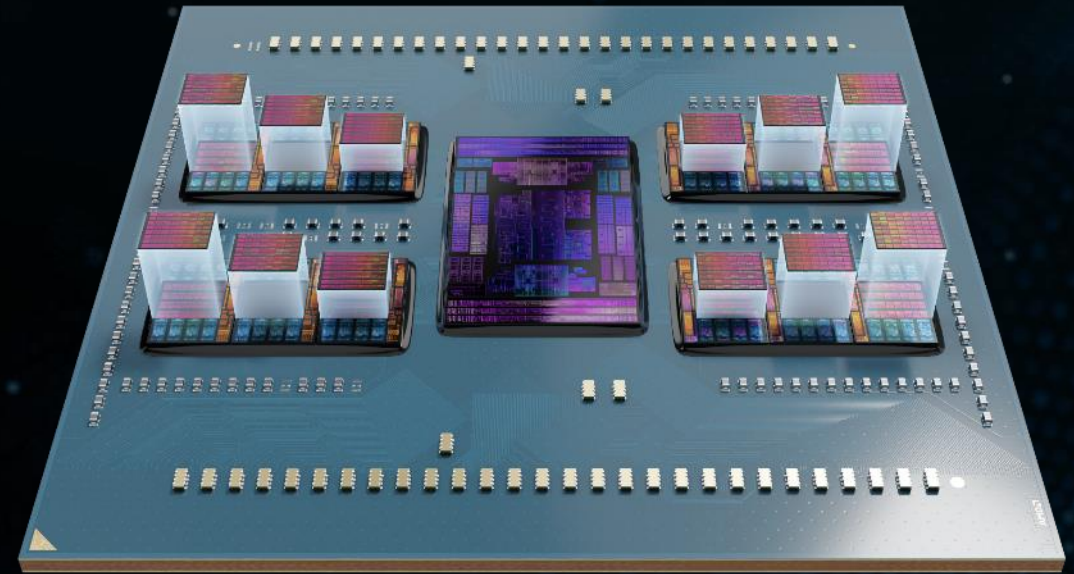
2P servers: 128C AMD EPYC 9754 vs Ampere Altra Max M128-30 and 60C Intel Xeon Platinum 8490H

Results may vary due to factors including system configurations, software versions and BIOS settings. As of 6/13/2023, see SP5-149.
<https://www.amd.com/en/claims/epyc4>

AMD
together we advance_

4TH GEN AMD EPYC

With AMD 3D V-Cache[®] Technology



Leadership 5nm
Process Node

High Performance
“Zen 4” cores

Up to 1.1 GB
of L3 Cache

AMD
Infinity Guard

Rich Ecosystem
of Solutions

World's highest performance x86 server CPU for technical computing

AMD EPYC™ 9004 シリーズ・プロセッサ

すべての製品で同じ機能が提供されます

- 12 チャンネルの DDR5-4800
- 最大 6TB DDR5 メモリー容量
- 128 レーンの PCIe Gen5
- 64 レーンの CXL™ 1.1+
- AVX-512 ISA、SMT & コア周波数ブースト
- AMD Infinity Fabric™
- AMD Infinity Guard²

コア数	AMD EPYC	ベース/ブースト ¹ (最大 GHz)	デフォルト TDP (w)	cTDP (w)
128 コア	9754 /S	2.25/3.10	360w	320-400w
112 コア	9734	2.20/3.00	340w	320-400w
96 コア	9684X	2.55/3.70	400W	320-400w
96 コア	9654/P	2.40/3.70	360w	320-400w
84 コア	9634	2.25/3.70	290w	240-300w
64 コア	9554/P	3.10/3.75	360w	320-400w
	9534	2.45/3.70	280W	240-300w
48 コア	9474F	3.60/4.10	360w	320-400w
	9454/P	2.75/3.80	290w	240-300w
	9384X	3.10/3.90	320w	320-400w
32 コア	9374F	3.85/4.30	320w	320-400w
	9354/P	3.25/3.80	280W	240-300w
	9334	2.70/3.90	210w	200-240w
24 コア	9274F	4.05/4.30	320w	320-400w
	9254	2.90/4.15	200W	200-240w
	9224	2.50/3.70	200W	200-240w
	9184X	3.55/4.20	320W	320-400w
16 コア	9174F	4.10/4.40	320w	320-400w
	9124	3.00/3.70	200W	200-240w

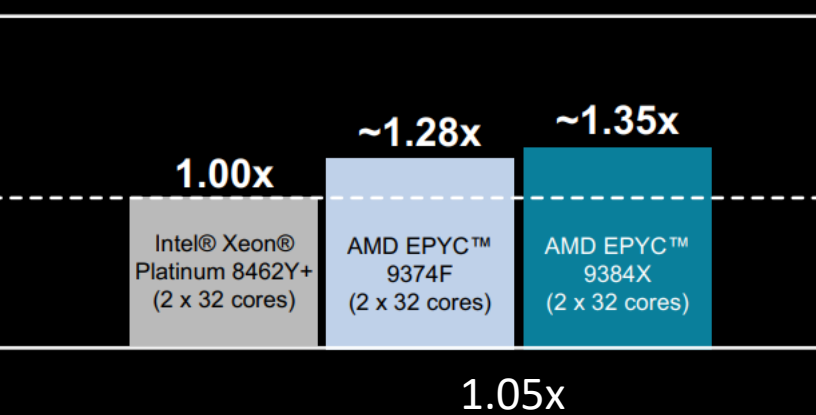
¹ 脚注 EPYC-18 を参照。AMD EPYC プロセッサの最大ブーストとは、サーバー・システムの通常の動作条件下において、プロセッサのシングル・コアで達成可能な最大周波数を指しています。EPYC-018

² AMD Infinity Guard の機能は、EPYC プロセッサの世代によって異なります。AMD EPYC プロセッサの Infinity Guard のセキュリティ機能は、サーバー OEM および/またはクラウド・サービス・プロバイダーにより有効化されている必要があります。これらの機能のサポートについては、OEM またはプロバイダーに確認してください。Infinity Guard の詳細については、<https://www.amd.com/en/technologies/infinity-guard> をご覧ください。GD-183

GenoaとGenoa-Xの同コア数での比較

Ansyes® Fluent®

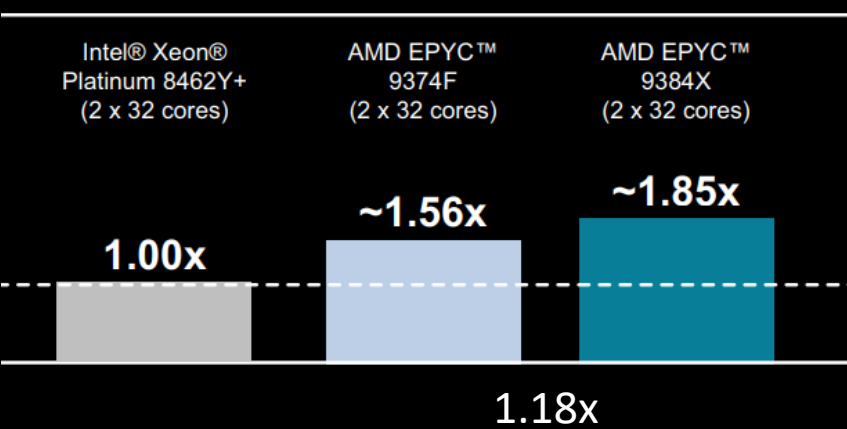
2P AMD EPYC™ 9384X vs. 2P INTEL XEON PLATINUM 8462Y+
(32-CORE - COMPOSITE AVERAGE, NORMALIZED TO 8462Y+)



<https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-fluent.pdf>

Ansyes® CFX®

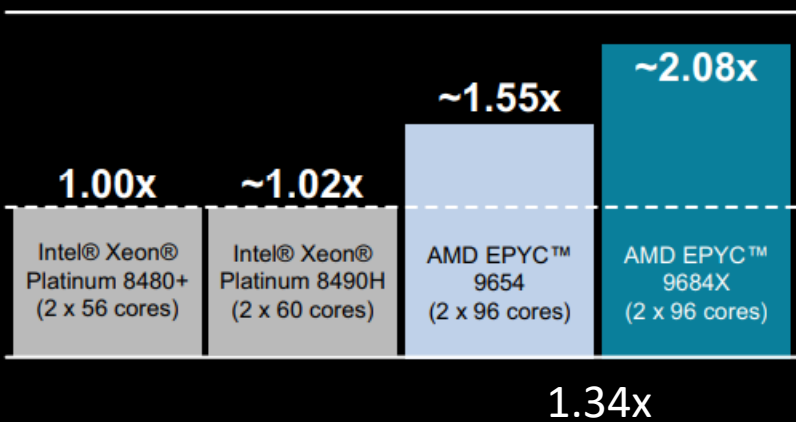
2P AMD EPYC™ 9384X vs. 2P INTEL XEON PLATINUM 8462Y+
(32-CORE - COMPOSITE AVERAGE, NORMALIZED TO 8462Y+)



<https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-cfx.pdf>

OpenFOAM®

2P AMD EPYC™ 9684X vs. 2P INTEL XEON PLATINUM 8480+
(TOP OF STACK - COMPOSITE AVERAGE, NORMALIZED TO 8480+)



<https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-openfoam.pdf>

AMD EPYC™ Technical Whitepapers and Briefs
<https://www.amd.com/en/processors/server-tech-docs/search>



Optimized AI software stack

AI Models and Algorithms

 PyTorch

 TensorFlow

 ONNX



AI Ecosystem optimized for AMD

Libraries

Compilers and Tools

Runtime



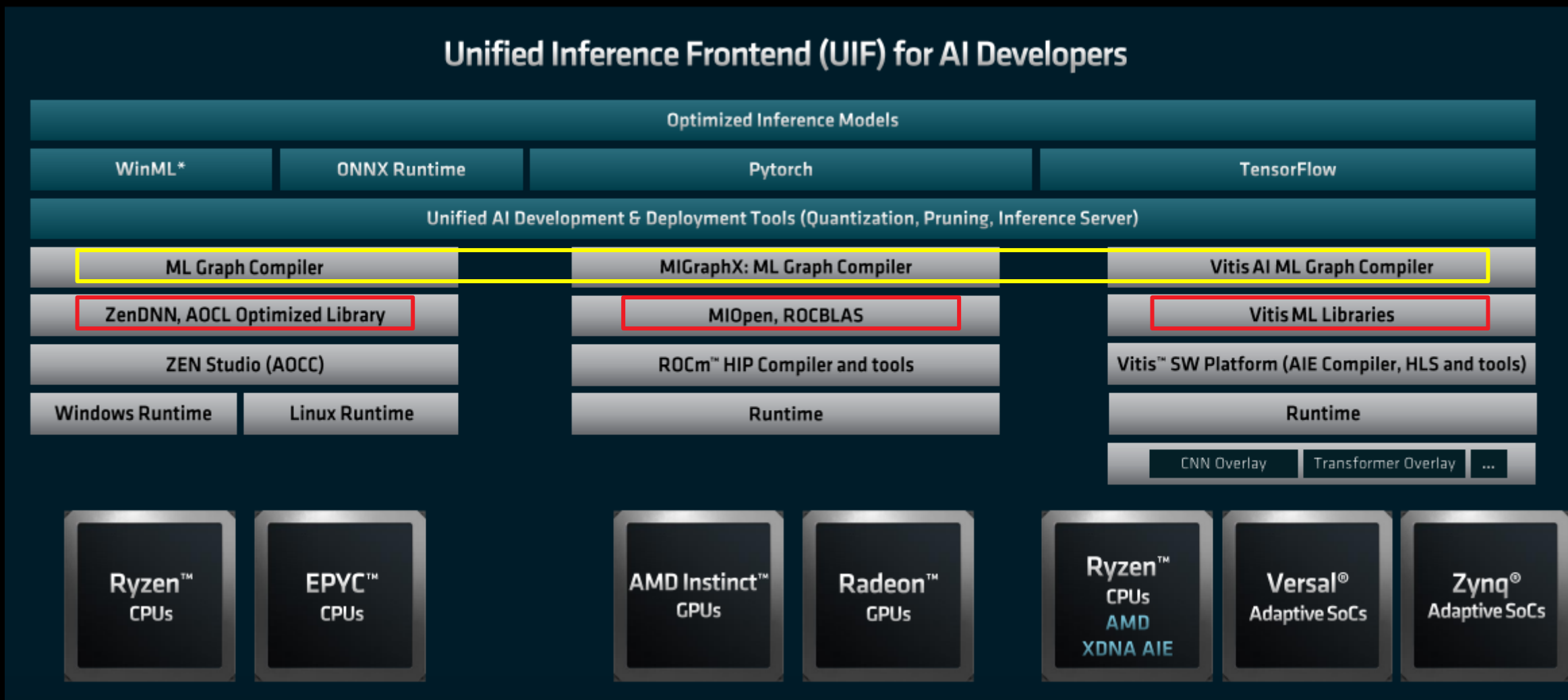
A proven software stack

AMD Instinct™ GPU



Leadership performance

AMD Unified Inference Frontend



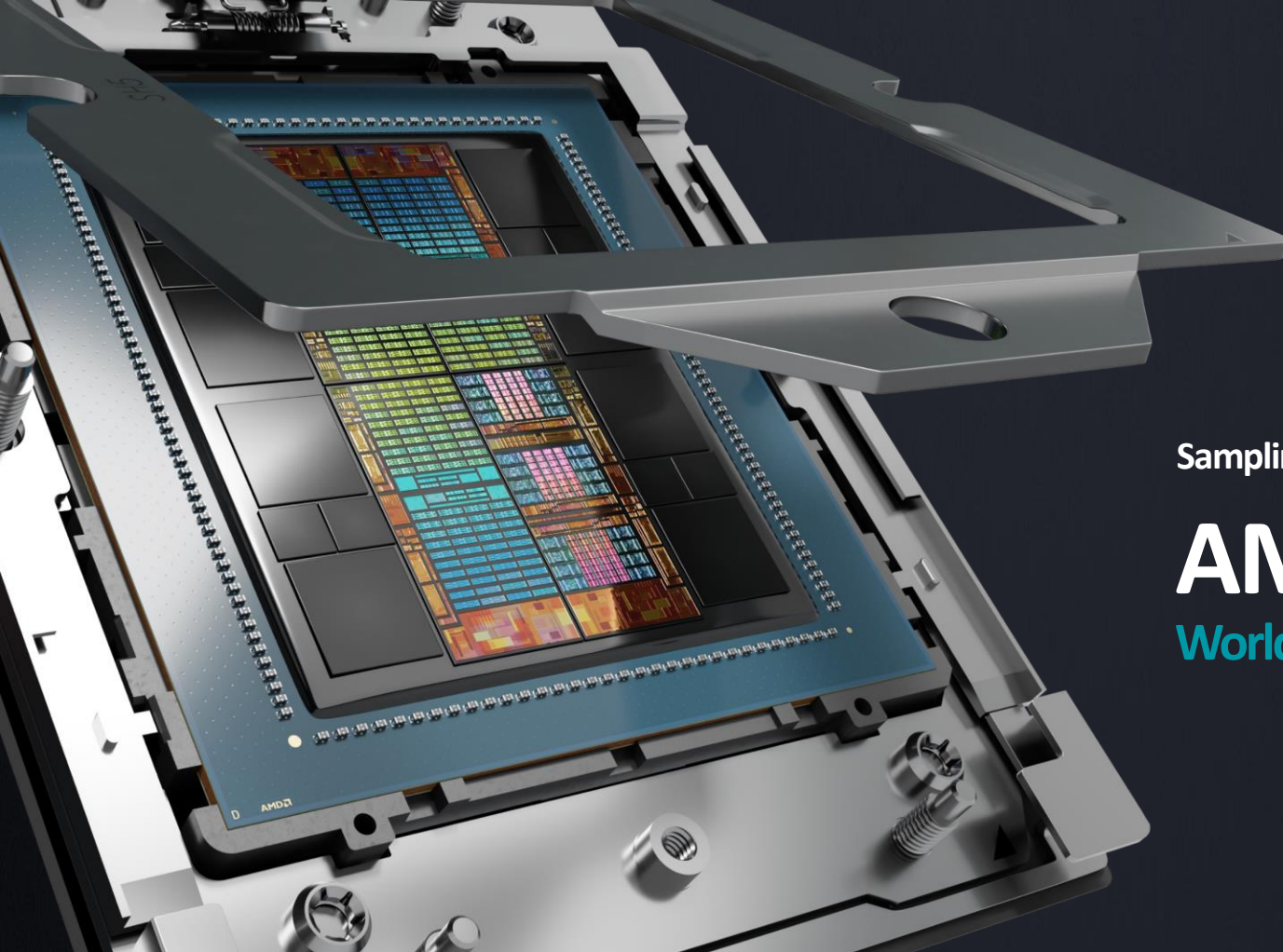


Next-gen AI accelerator architecture

Dedicated accelerator
engines for AI and HPC

3D packaging with 4th Gen AMD
Infinity architecture

Optimized for performance
and power efficiency



Sampling now

AMD Instinct™ MI300A

World's first APU accelerator for AI and HPC



Next-Gen
Accelerator
Architecture

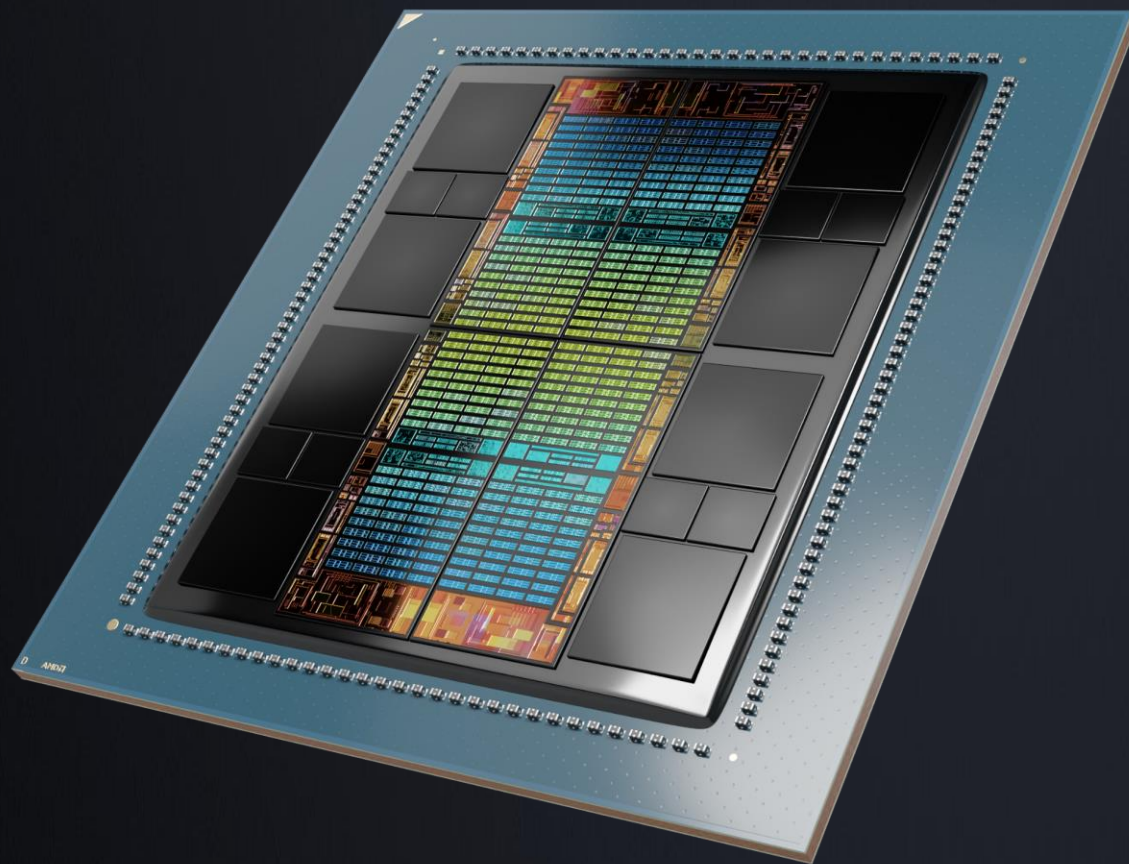


24 CPU
Cores

128 GB
HBM3

5nm and 6nm
Process Technology

Shared Memory
CPU + GPU



Introducing today

AMD Instinct™ MI300X

Leadership generative AI accelerator

AMD
CDNA 3

192 GB
HBM3

5.2 TB/s
Memory Bandwidth

896 GB/s
Infinity Fabric™ Bandwidth

153 B
Transistors



注意事項

本プレゼンテーションには、Advanced Micro Devices, Inc. (AMD) データセンターの製品の時期・可用性・機能・特徴・期待される利点、データセンター市場の傾向とTAM、AMDデータセンターのロードマップを含むがこれらに限定されないAMDの将来に関する記述を含んでおり、1995年私募証券訴訟改革法 (Private Securities Litigation Reform Act) の免責条項に従って作成されました。将来に関する記述は一般的に、「予想」、「可能性」、「予測」、「思われる」、「計画」、「意図」、「プロジェクト」、その他同様の意味を持つ言葉で識別されます。投資家の皆様は、本プレゼンテーションにおける将来に関する記述は、現在の考え、前提条件、期待に基づいて作成されたものであり、本プレゼンテーションの日付時点における内容であり、実際の結果が現在の予想と大きく異なるリスクや不確実性を含むことにご注意ください。将来に関する記述は、既知および未知のリスクや不確実性の影響を受けるものです。その多くは予測が困難であり、AMDが制御できる範囲を超えています。そのため、実際の結果やその他の将来の事象は、将来に関する情報や記述で明示・暗示・予測されている内容とは大きく異なる可能性があります。投資家の皆様は、形式10-K及び10-QでのAMDの最近の報告書を含むがこれに限定されない、AMDの証券取引委員会の書類におけるリスクと不確実性を詳細に検討することが推奨されます。

AMDは、法律によって要求される場合を除き、本プレゼンテーションで作成された将来に関する記述を更新する義務を一切負わず、放棄します。