

HIGH PERFORMANCE COMPUTING IS



第18回PCクラスタシンポジウム
秋葉原 UDX 2018年 12月

日本AMD株式会社
ソリューション・アーキテクト
中村正澄

1



2

2



デザイン

世界初 7nm
高性能 x86 CPU

CPU コア
実行性能強化

さらなるセキュリティ
機能強化

構成と製造性を
考慮した
モジュラーデザイン

3

3



DELIVERS UP TO 2x THROUGHPUT 最大で2倍のスループットを提供

実行
パイプラインの
向上

2倍の
浮動小数点演算
と
ロード・ストア

2倍の
コア密度

オペレーション
あたり半分の
エネルギー

4

4

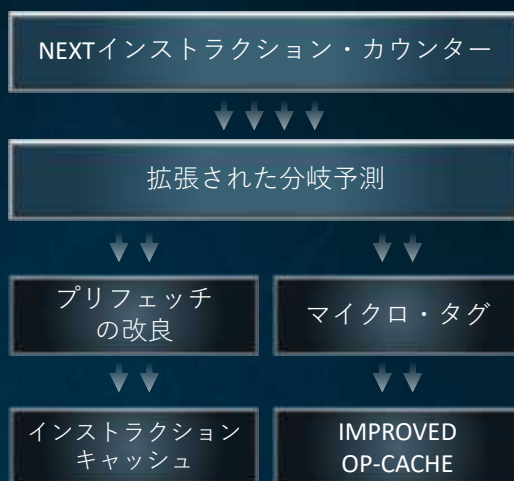
フロントエンド アドバンテージ

分岐予測の向上

命令プリフェッチの改良

命令キャッシュの再最適化

ラージ Op キャッシュ



5

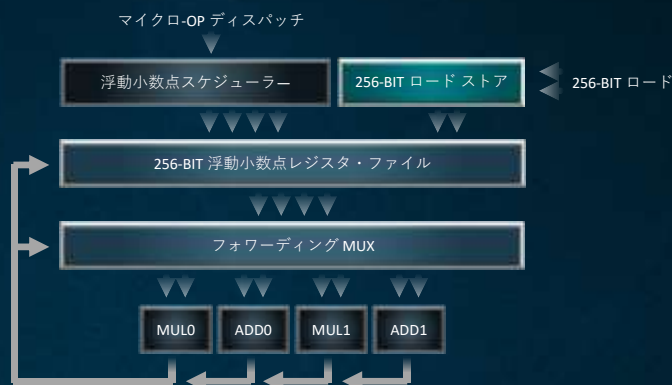
フローティングポイント アドバンテージ

浮動小数点幅を
256-Bitに拡張

ロード/ストアバンド幅
を2倍

ディスパッチ/リタイア
のバンド幅を増強

すべてのモードで高い
スループットを維持



6

“ZEN”

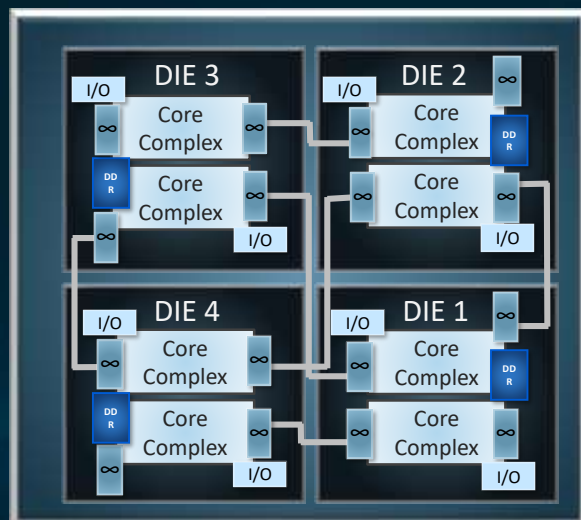
マルチチップ・アプローチによる最初の製品

すぐれた歩留まり

ピーク計算の向上

高いリコンフィグ性

“Zen” ベース EPYC™ プロセッサ



7



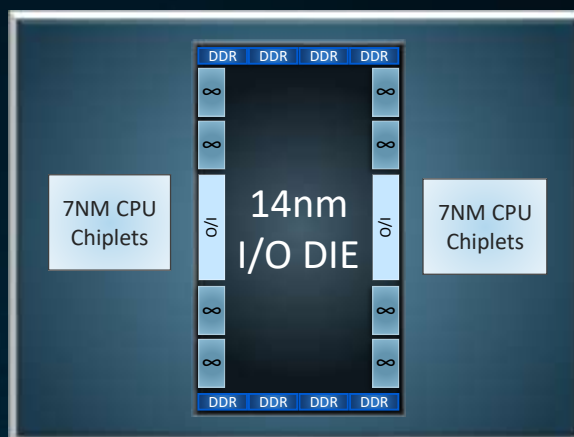
革命的 ROME アーキテクチャー

それぞれの IP に合ったテクノロジー

Infinity Fabric によるモジュラー化の実現

最適化された I/O ダイによりレイテンシーとパワーの向上

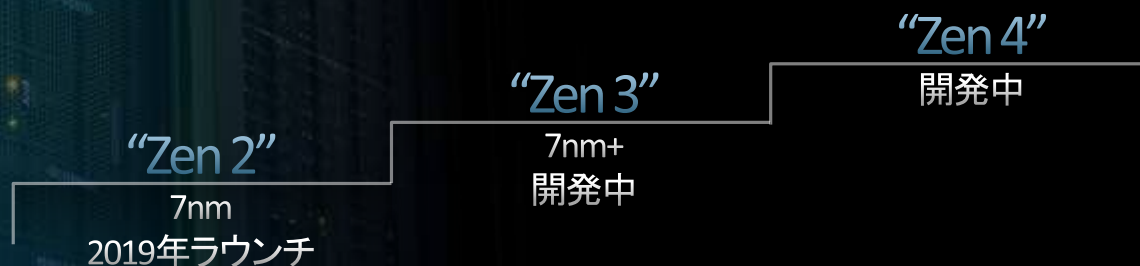
7nm テクノロジーによる
CPU 性能とパワー



“Zen 2” ベース EPYC プロセッサ

8

HIGH-PERFORMANCE MOMENTUM



リーダーシップをとるロードマップと実行 リーダーシップ プロダクト



“Rome” プロセッサー

注) 開発コード名

アーキテクチャー
7nm “Zen 2” cores
革命的 Rome アーキテクチャー

コンピューート
最大 64 コア
8 メモリ チャンネル

機能
128 レーン PCIe4
200Gbps HDR IB
200Gbit Ethernet
セキュリティ 機能

SC18でアナウンスされたROME採用のプロジェクト

H L R I S

High-Performance Computing Center of
the University of Stuttgart (HLRS)



Finnish IT Center for Science Ltd
CSC



NATIONAL ENERGY
RESEARCH SCIENTIFIC
COMPUTING CENTER

新製品AMD EPYC 7371 ご紹介



AMD EPYC 7351

16 コア / 32 スレッド
2.4GHz ベース
2.9GHz 全コア・ブースト
2.9GHz Max ブースト (8 cores)

24%
高周波数



AMD EPYC 7371

16 コア / 32 スレッド
3.1GHz ベース
3.6GHz 全コア・ブースト
3.8GHz Max ブースト (8 cores)
サンプル中 / 1Q19 量産

電子設計自動化ツール(EDA)、ビデオ・トランスコード、解析サーチなどに最適

HPC ISV エコシステム



13 | AMD EPYC | PC CLUSTER CONSORTIUM DEC 2019

13

AMD

AMD 機械学習ソリューション のご紹介

AMD
RADEON INSTINCT



14

クラウド環境向けにデザイン AMD RADEON INSTINCT™ MI-50/60

世界初
7nm GPU

高性能でフレキシブルな
Vega アーキテクチャ

最先端のメモリ技術
とスケーラビリティ

業界で唯一のハードウェア仮想化 GPU

15

世界初の 7 nm GPUs



MI25



MI60

最大
2倍の密度
13.2Bトランジスタ, 331 mm²

1.25倍以上
の性能
同じ消費電力において

50% 低い
消費電力
同じ周波数において

高い 消費電力あたりの性能

16

AMD RADEON INSTINCT™ MI60 AND MI50



PRODUCT DETAILS



	MI60	MI50
Compute Units	64	60
Stream Processors	4096	3840
Peak INT4 Performance (TOPS)	Up to 118	Up to 107
Peak INT8 Performance (TOPS)	Up to 59.0	Up to 53.6
Peak FP16 Performance (TFLOPS)	Up to 29.5	Up to 26.8
Peak FP32 Performance (TFLOPS)	Up to 14.7	Up to 13.4
Peak FP64 Performance (TFLOPS)	Up to 7.4	Up to 6.7
Memory Size	32GB HBM2	16GB HBM2
Memory Bandwidth	1 TB/s	1 TB/s
Memory Interface	4096 Bit	4096 Bit
ECC (Full-chip)	Yes	Yes
RAS Support	Yes	Yes
PCIe® Gen 4 Capable**	Yes	Yes
Infinity Fabric™ Link	Yes – 2 links	Yes – 2 links
MxGPU (HW SR-IOV)	Yes	Yes
Max Power	300W TDP	300W TDP

*See Endnotes
**Pending

17

AI ワークロードのための超高速な演算性能



Up to

29.5
TFLOPS FP16

Up to

59
TOPS INT8

Up to

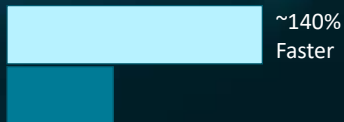
118
TOPS INT4

MI60

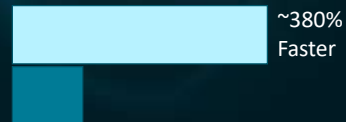


MI25

~20% Faster



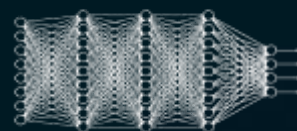
~140% Faster



~380% Faster

新しい演算命令:

- ▶ V_DOT2_F32_F16
- ▶ V_DOT2_I32_I16
- ▶ V_DOT2_U32_U16
- ▶ V_DOT4_I32_I8
- ▶ V_DOT4_U32_U8
- ▶ V_DOT8_I32_I4
- ▶ V_DOT8_U32_U4



18

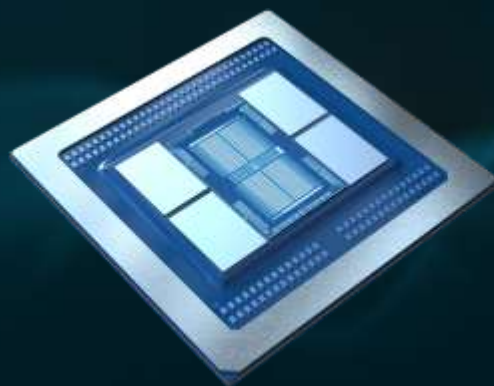
高い効率のメモリサブシステム 1 TB/S MEMORY BANDWIDTH

32 GBの広帯域
メモリー (HBM2)

ECCによるエラー検出・
訂正機能を内蔵

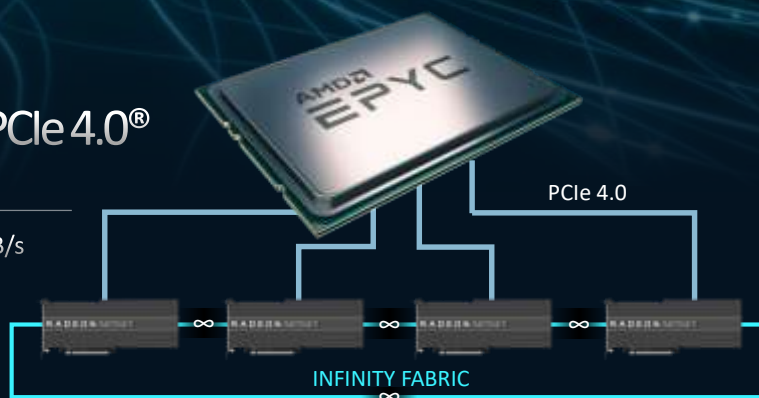
エンタープライズクラスの信頼性で

大規模なデータセットをサポート



業界初のPCIe 4.0® 対応GPU

双方向帯域 64 GB/s
のCPU-GPU間接続



INFINITY FABRIC LINKS

リンク当たり帯域
100 GB/s のGPU-
GPU間接続

先進のインターコネクト

INFINITY FABRIC™ リンクによる高いスケーラビリティ



Up to **200 GB/s**
GPU peer-to-peer

Up to **6x**
Speed of PCIe Gen 3

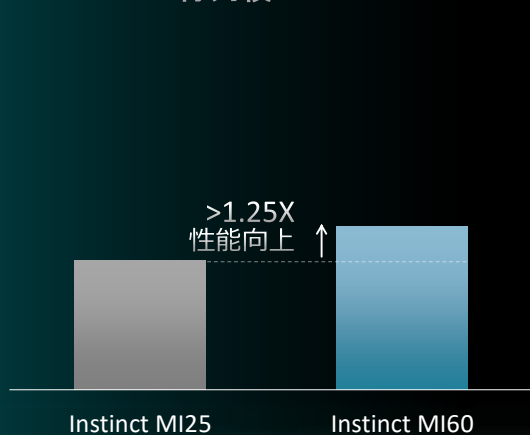
PCIe® フォームファクタでの高性能P2PGPU通信

21 | AMD EPYC | PC CLUSTER CONSORTIUM DEC 2018

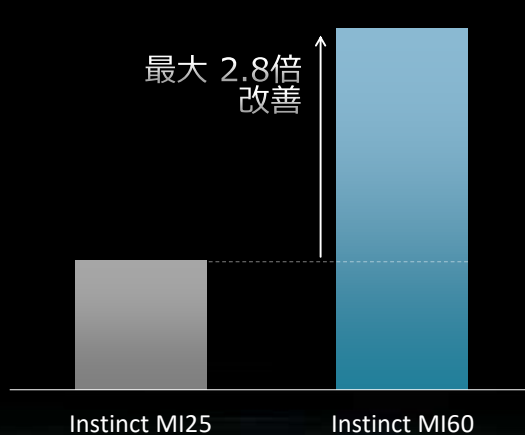
21

最適化されたアーキテクチャによる高いパフォーマンス

行列積 – FP16



Resnet-50 FP16



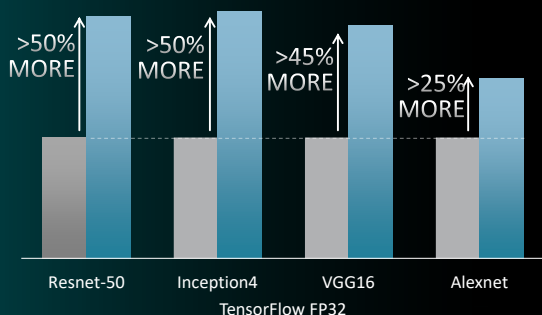
22 | AMD EPYC | PC CLUSTER CONSORTIUM DEC 2018

*See End of AMD

22

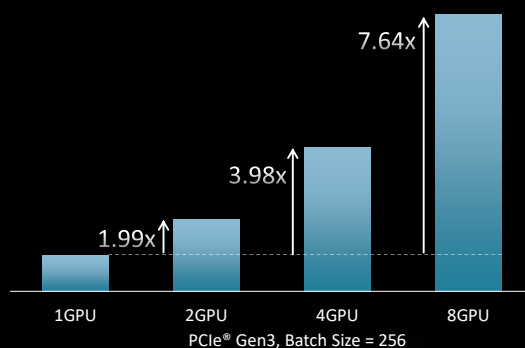
MI25との性能比較とマルチGPUにおける優れたスケーリング

MI25に対する性能向上



リニアな性能スケーリング

Resnet-50



■ Instinct MI25

■ Instinct MI60

23 | AMD EPYC | PC CLUSTER CONSORTIUM DEC 2018

*See End

23

ROCm はオープンソースにもとづいた環境を提供

最新の機械学習フレームワークをサポート

最適化されたディープラーニングと算術用ライブラリ

DockerとKubernetesのサポート

Linux Kernel への ROCm のアップストリーム



RADEON OPEN ECOSYSTEM

Applications

Machine Learning Apps

Frameworks

Caffe 2

MXnet

PyTorch

TensorFlow

Middleware and Libraries

MIOpen

BLAS, FFT, RNG

RCCL

Eigen

ROCm

OpenMP

HIP

OpenCL™

Python

Fully Open Source ROCm Platform

Devices

GPU

CPU

APU

DLA

24 | AMD EPYC | PC CLUSTER CONSORTIUM DEC 2018

AMD

24

機械学習フレームワークへの対応

TensorFlow

v1.11 – Available Today -- <https://github.com/tensorflow/tensorflow>

PYTORCH

Caffe2

Available Today -- <https://github.com/ROCmSoftwarePlatform/pytorch>

mxnet

Caffe

PaddlePaddle

Available Today -- <https://github.com/ROCmSoftwarePlatform/>

Applications	Machine Learning Apps			
Frameworks	Caffe 2	MXnet	PyTorch	TensorFlow
Middleware and Libraries	MIDopen	BLAS/PT/ANN	RCCL	Eigen
ROCm	OpenCL™	HIP	OpenCL™	Python
	Fully Open Source ROCm Platform			
Devices	CPU	CPU	APU	GPU

25 | AMD EPYC | PC CLUSTER CONSORTIUM DEC 2018

AMD

25

ROCMはスケーラブルですぐにデプロイ可能

1 時間以内に新規のサーバーで実行可能です



hub.docker.com/r/rocm/

kubernetes

<https://github.com/RadeonOpenCompute/k8s-device-plugin>
<https://hub.docker.com/r/rocm/k8s-device-plugin/>

26 | AMD EPYC | PC CLUSTER CONSORTIUM DEC 2018

AMD

26

将来を見据えた選択とイノベーション データセンター GPU ロードマップ

AMD Radeon Instinct™ MI25

14nm GPU
PCIe® 3.0
16 GB of HBM2 Memory

AMD Radeon Instinct™ MI60

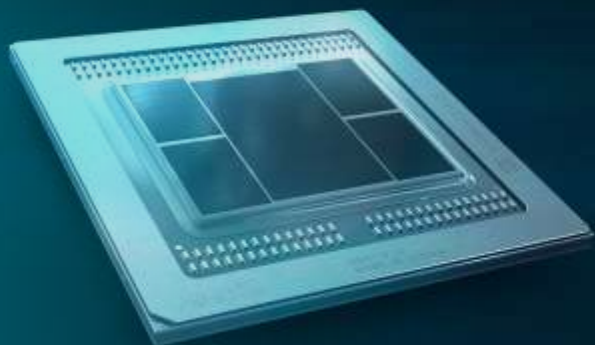
First 7nm GPU
First PCIe® 4.0 Capable GPU
32 GB of HBM2 Memory
Infinity Fabric Links

“MI-NEXT”

Higher Performance
Increased Connectivity
Software Compatibility

世代ごとのパフォーマンス向上を伴う予測可能な製品ロードマップ

AMDは、データセンターCPUとGPU アーキテクチャー両方を設計



世界初 7nm GPU

AMD RADEON INSTINCT™ MI50/60



世界初 7nm データセンター CPU

“Rome”



29

DISCLAIMER

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's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

©2018 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, EPYC™ and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

30