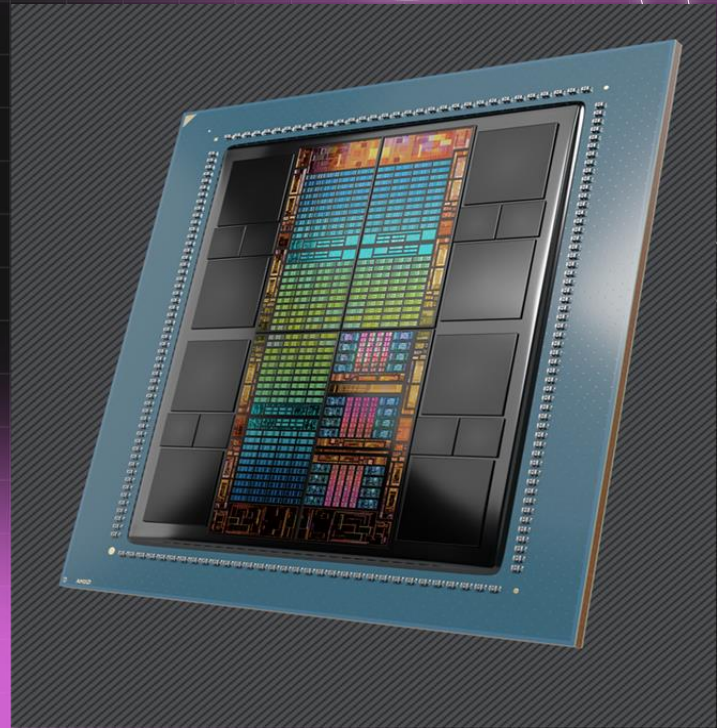
ИТМО

Платформы открытого кода ROCm для суперкомпьютеров и систем искусственного интеллекта



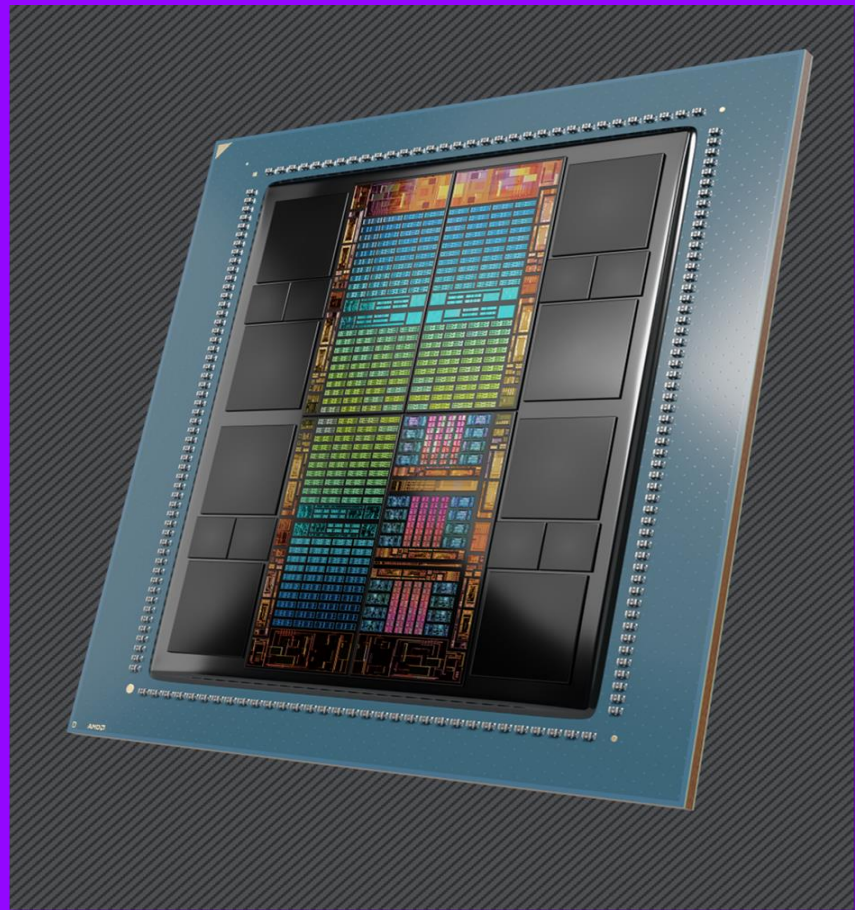
Введение в ROCm в учебном пособии по программированию с помощью HIP

ИТМО

Гетерогенные процессоры больших систем

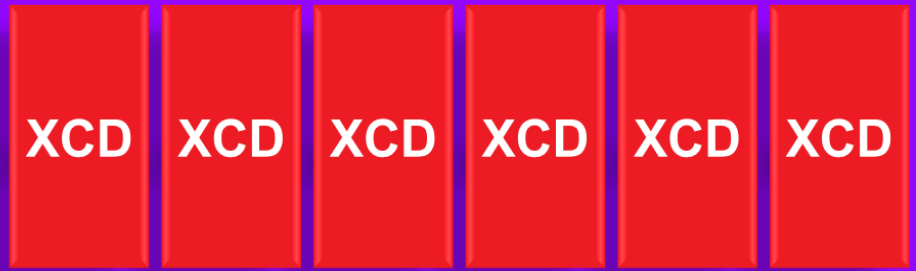
ИТМО

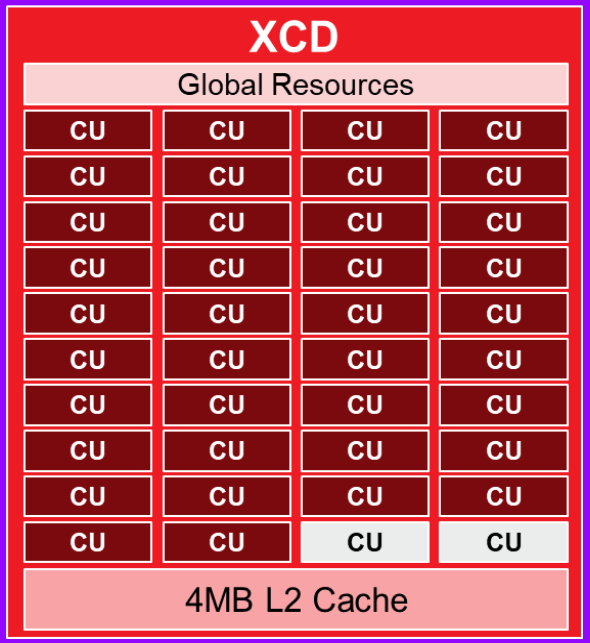
AMD Instinct™ MI300A APU и M300X GPU



Вычислительные ядра GPU в процессоре AMD Instinct™ MI300A

ИТМО

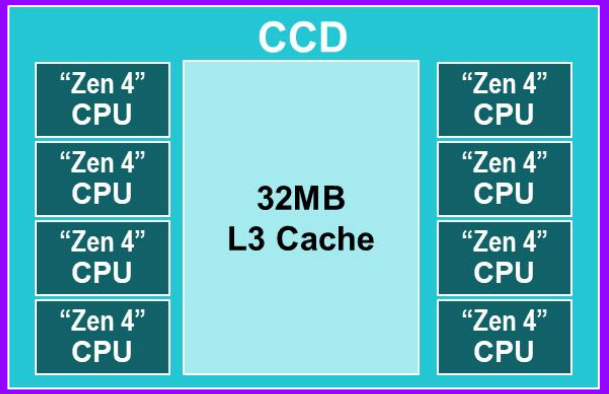




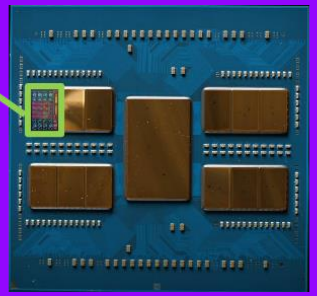
- XCD – Accelerator Complex Die
- 38 CUs per XCD, 228 total AMD CDNA™ 3 architecture

Вычислительные ядра CPU в процессоре AMD Instinct™ MI300A

ИТМО

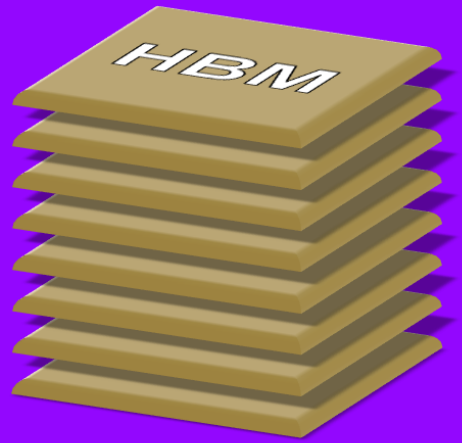


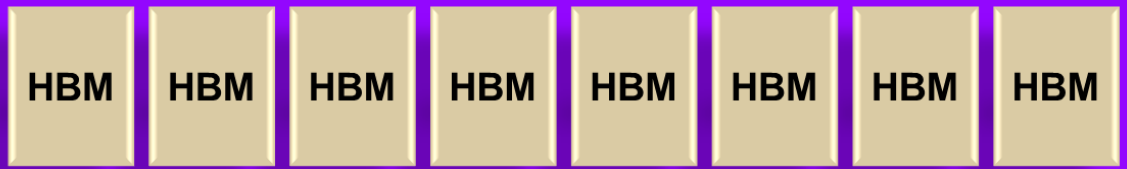
- CCD – CPU Complex Die
- 8 "Zen 4" cores per CCD, 24 total
- Leverage CCD from EPYC



Память высокой пропускной способности в процессоре AMD Instinct™ MI300A

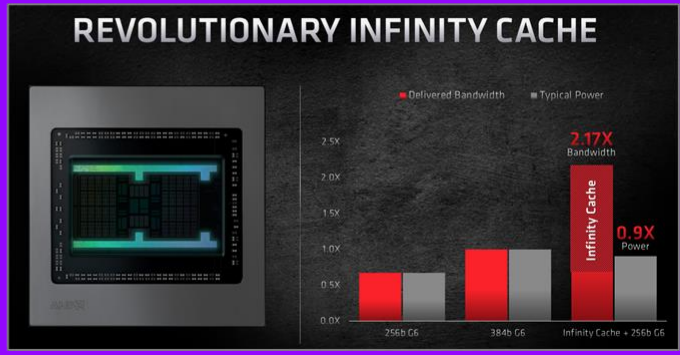
ИТМО

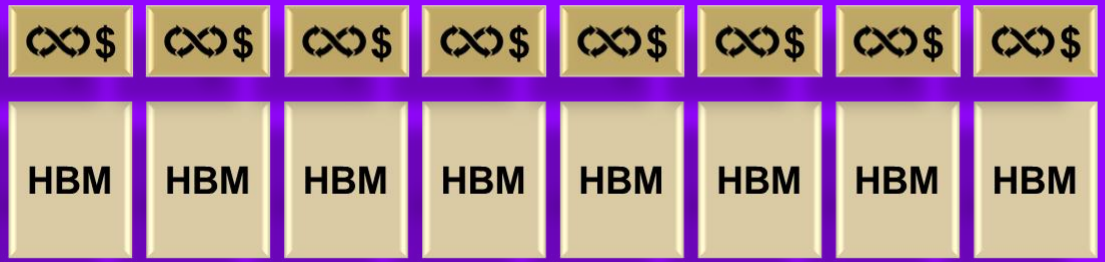




- HBM – High Bandwidth Memory
- HBM gen 3, 16GB (128 GB total)
- 665 GB/s/stack (5.3 TB/s total)
- 128 total memory channels

Многоканальный кэш для доступа к памяти в процессоре AMD Instinct™ MI300A

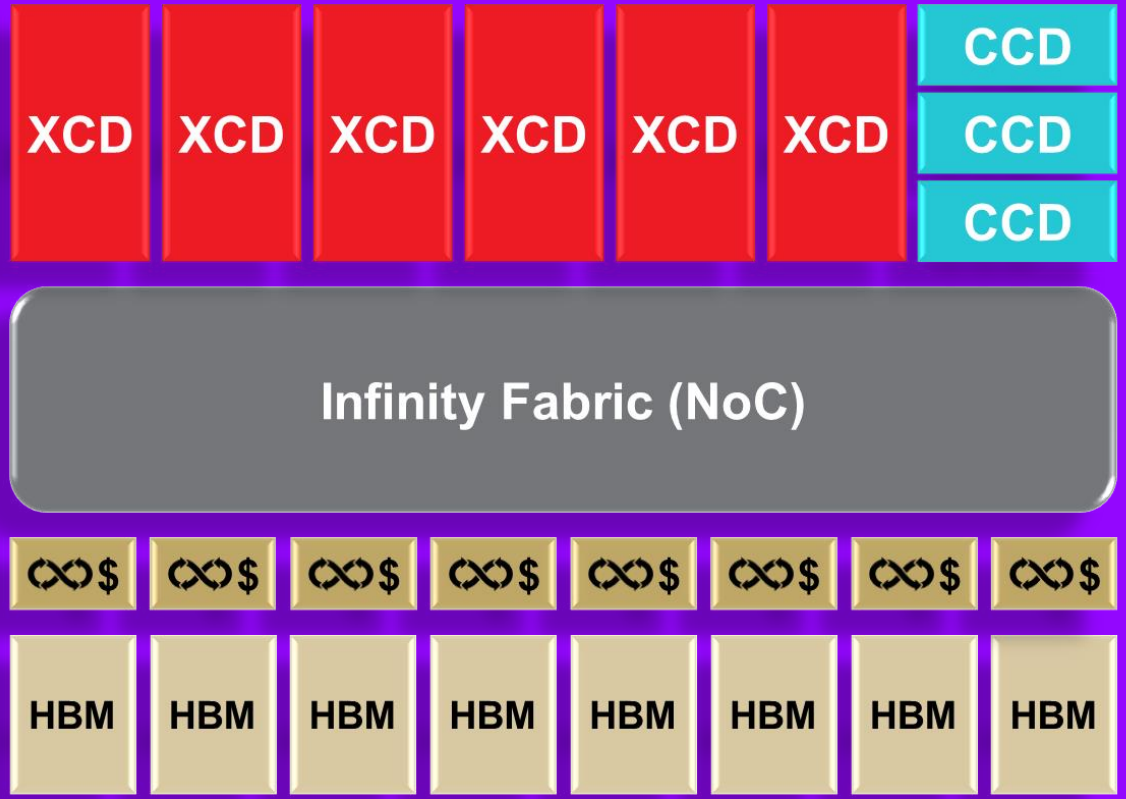


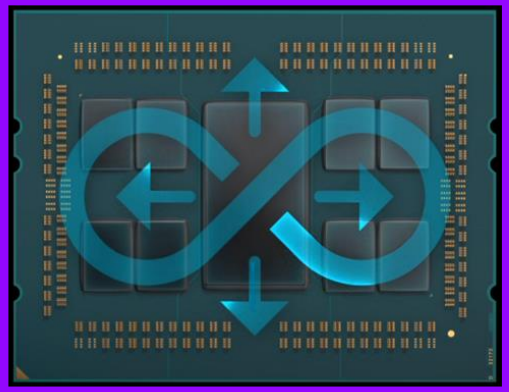


- Memory-side Infinity Cache
- 2MB/channel (256 MB total)
- BW amplification (up to 17 TB/s)

Когерентный интерфейс в процессоре AMD Instinct™ MI300A

ИТМО



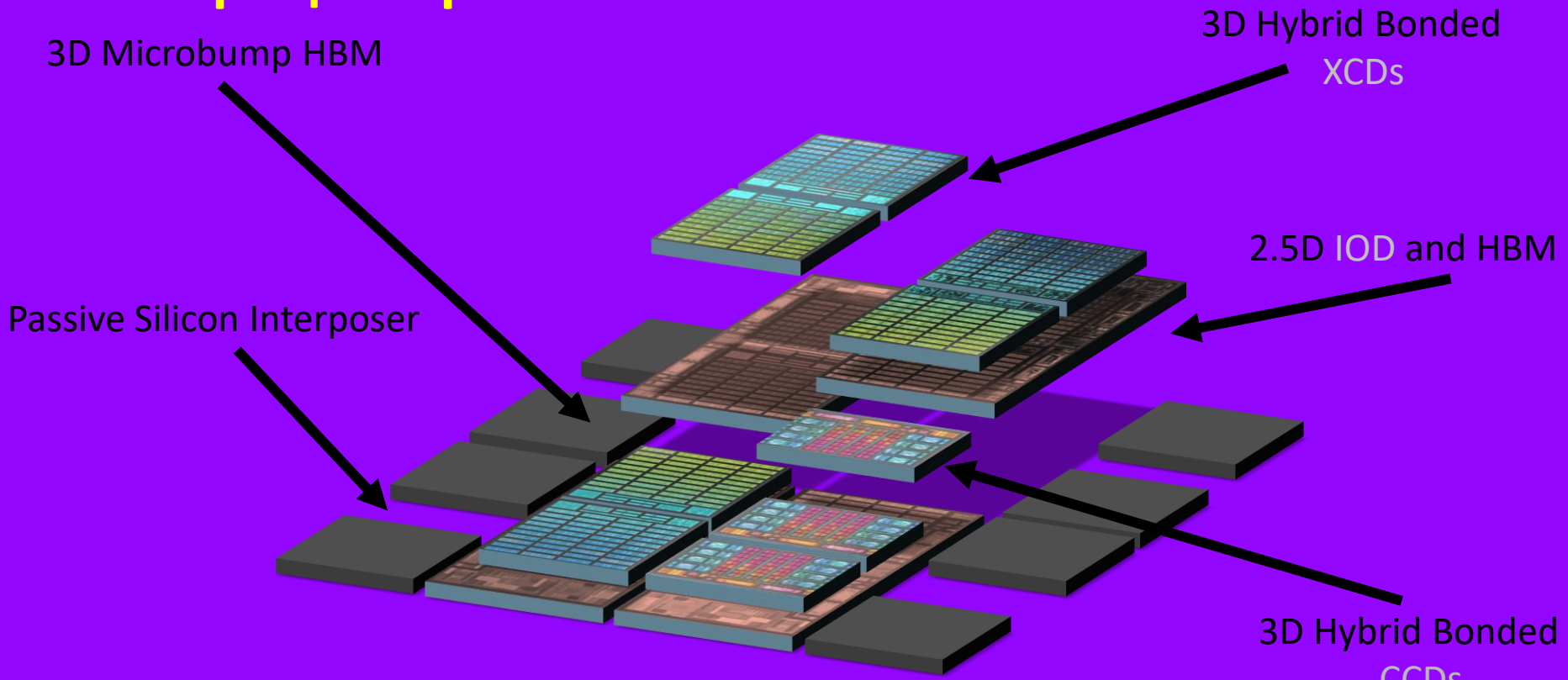


- IF – Infinity Fabric
- Fully-coherent fabric (CPU+GPU)
- Provides I/O connectivity
  - Four x16 links to other MI300A (IF)
  - Four x16 links IF or PCIe® gen5
  - Each link at 64 GB/s/dir

Основные компоненты процессора AMD Instinct™ MI300A

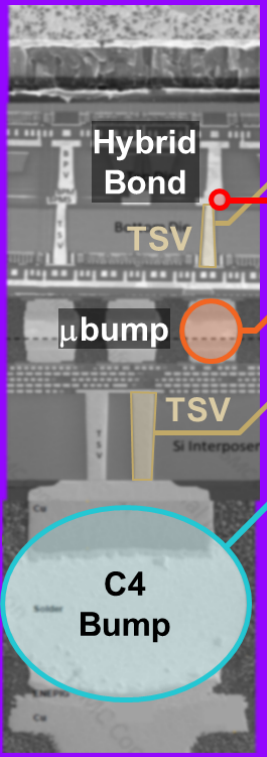
Трехмерная сборка основных компонент процессора AMD Instinct™ MI300A

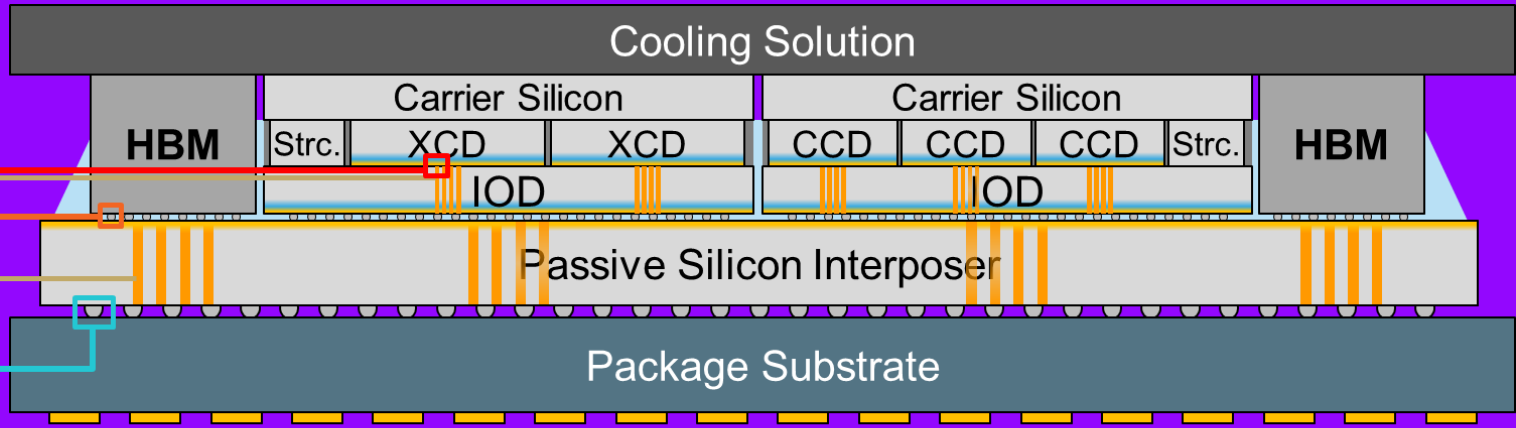
ИТМО



Передовая технология 3D-упаковки гетерогенных модулей и чиплетов

ИТМО





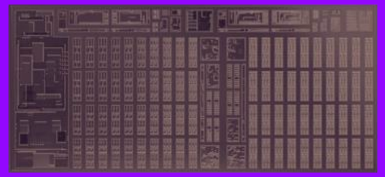
- 8 stacks of HBM
- 6 XCDs
- 3 CCDs
- 4 IODs
  - IOD – I/O Die

- 3D hybrid bonding
- 2.5D silicon interposer
  - IOD-IOD links
  - IOD-HBM links

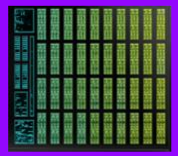
Модульная конструкция с многократным использованием чиплетов

ИТМО

MI300 **IOD**



CDNA™ 3 **XCD**



4th-gen EPYC™ “Genoa” **CCD**

4th-gen EPYC™ “Genoa” **IOD**

4x IOD

4x IOD

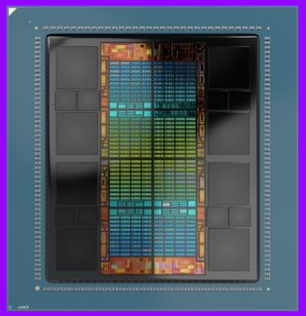
8x XCD

6x XCD

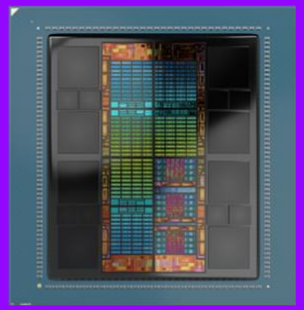
3x CCD

up to 12x CCD

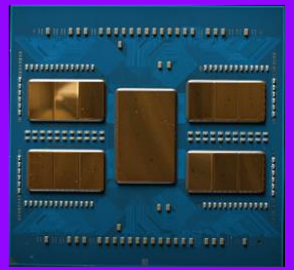
1x IOD



AMD Instinct™ MI300X Accelerator

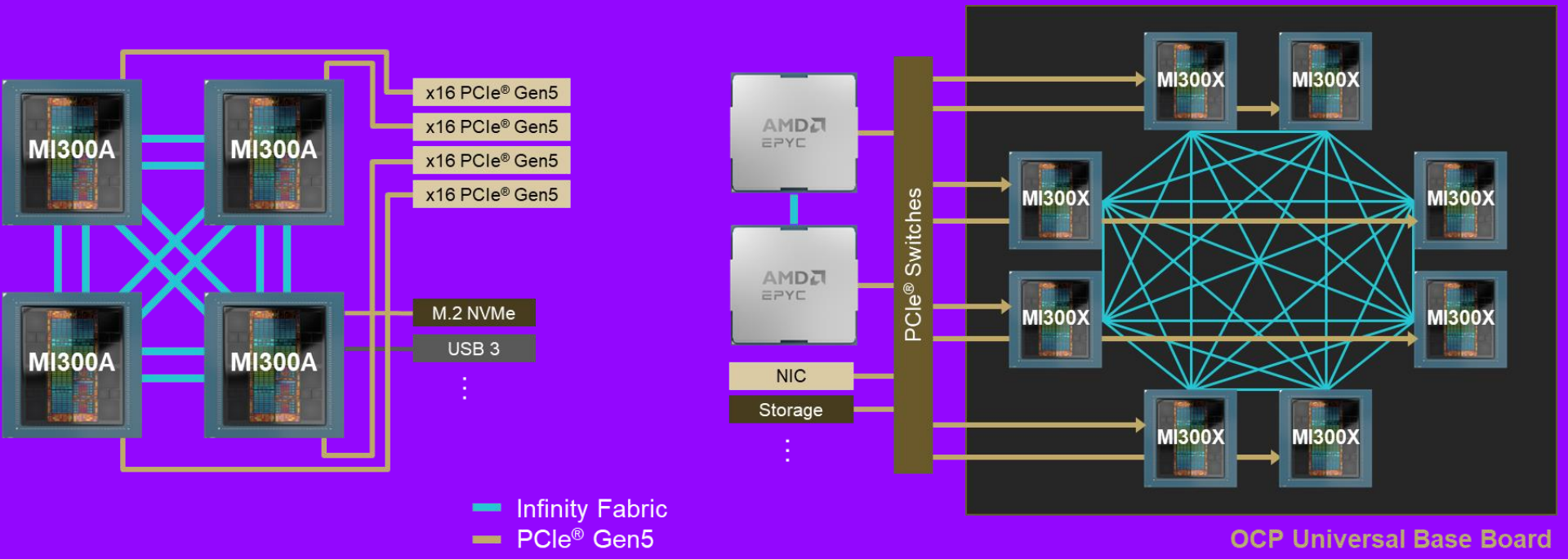


AMD Instinct™ MI300A APU



4th-gen AMD EPYC™ CPU

Модульный ввод/вывод обеспечивает гибкость масштабирования



Разница между поколениями AMD Instinct™ MI200 и MI300

ИТМО

		AMD Instinct MI250X (Up to)	AMD Instinct MI300A (Up to)	AMD Instinct MI300X (Up to)
Hardware Specs	Memory Capacity	128 GB HBM2e	128 GB HBM3	192 GB HBM3
	Memory Bandwidth (Peak Theoretical)	~3.2 TB/s	~5.3 TB/s	~5.3 TB/s
	Scale-Out (Back-end) Network Bandwidth	200 Gb/s Ethernet/IB	400 Gb/s Ethernet/IB	400 Gb/s Ethernet/IB
	Max TDP/TBP	560 W	760 W	760 W
	Heterogeneous Integration	2D IF, 2.5D EFB, 3D μbump HBM	2.5D passive silicon interposer, 3D HB active interposer and multiple chiplets, 3D μbump HBM	
HPC Peak Perf. (Peak)	FP64 Vector (TFLOPS)	47.9	61.3	81.7
	FP32 Vector (TFLOPS)	47.9	122.6	163.4
	FP64 Matrix (TFLOPS)	95.7	122.6	163.4
	FP32 Matrix (TFLOPS)	95.7	122.6	163.4
AI Peak Perf. (Peak)	TF32* [TF32 Sparsity] (Matrix)	Not Supported	490.3 [980.6]	653.7 [1307.4]
	FP16 [FP16 Sparsity] (TFLOPS)	383.0 [Not Supported]	980.6 [1961.2]	1307.4 [2614.9]
	BFLOAT16 [BF16 Sparsity] (TFLOPS)	383.0 [Not Supported]	980.6 [1961.2]	1307.4 [2614.9]
	FP8* [FP8 Sparsity] (TFLOPS)	Not Supported	1961.2 [3922.3]	2614.9 [5229.8]
	INT8 [INT8 Sparsity] (TOPS)	383.0 [Not Supported]	1961.2 [3922.3]	2614.9 [5229.8]

Ссылка на статью по AMD Instinct™ MI300A со всеми деталями

ИТМО

Smith, A., Loh, G.H., Schulte, M.J., Ignatowski, M., Naffziger, S., Mantor, M., Kalyanasundharam, M.F., Alla, V., Malaya, N., Greathouse, J.L., Chapman, E., & Swaminathan, R. (2024).

Realizing the AMD Exascale Heterogeneous Processor Vision : Industry Product. 2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA), 876-889.

https://drive.google.com/file/d/1J9tLeVbFRtarzIzkVrULiEBsRsEMNmEO

Стек программного обеспечения ROCm™

Фреймворки
TensorFlow, PyTorch, Kokkos

Библиотеки
MIOpen, библиотеки roc*

Модели программирования
HIP, OpenCL

Промежуточные среды выполнения/компиляторы
Clang на основе LLVM (HIP-Clang)

Среды выполнения
ROCm

Программные и системные инструменты
- отладка
- профилирование

Общий подход к использованию AMD Instinct™

ИТМО

AMD предлагает набор программ, упрощающих переход с **Nvidia GPU**

1 УПРОЩЕНИЕ ПРОГРАММИРУЕМОСТИ И ПЕРЕНОСИМОСТИ КОДА

- Поддержка заказчиков в режиме Drop-in для облегчения миграции ведущих AI фреймворков and HPC моделей программирования
- Облегчение переносимости, упрощение программируемости и удобства пользования для ускорения получения конечных результатов
- Богатый набор высокопроизводительных и эквивалентных библиотек, а также программных инструментов для повышения производительности

2 РАСТУЩАЯ ЭКОСИСТЕМА ОТКРЫТОГО КОДА ДЛЯ HPC & AI

- Растущий репозитарий портированных и оптимизированных приложений через углубленное партнерство с ведущими клиентами в доменах HPC & AI
- Стек открытого кода ROCm и модель кросс-программирования HIP
- Активные инвестиции и участие в инициативах по разработке кросс-платформенного открытого кода (OpenXLA, Triton, MLIR)

Стек программного обеспечения ROCm 6.2

MIT/BSD License
Apache License
GPL License

AI Solutions & Services
- Business Services
- Fine Tunes Models

Ecosystem
- mosaicML
- Hugging Face
- PyTorch
- TensorFlow
- ONNX
- Triton
- deepspeed
- OpenXLA

Open Software Platform
- AMD ROCm

GPU
- AMD INSTINCT

Accelerated AI Math & Communication Libraries

HPC Focused
- rocALUTION
- rocSOLVER
- rocPRIM
- rocSPARSE
- rocThrust

AI Focused
- MIOpen
- MIGraphX
- rocWMMA
- rocBLASLt
- Composable Kernel

- rocBLAS
- rocRAND
- rocFFT
- RCCL

Compilers & Tools
- Compiler
- Profile/Tracer
- Debugger
- HIPIFY

Drivers Runtimes
- RedHat, SLES & Ubuntu Device Drivers and Run-time

Эквивалентные и оптимизированные библиотеки для повышения производительности широко используемых HPC и AI функций

	NVIDIA	AMD
Math Libraries	cuSparseLT	hipSparseLT [Alpha] (ROCm 6.0)
	cuTensor	hipTensor [Alpha] (ROCm 6.0)
	cuBLAS	rocBLAS
	cuBLASLt	hipBLASLt
	cuFFT	rocFFT
	cuSOLVER	rocSOLVER
	cuSPARSE	rocSPARSE
	cuRAND	rocRAND
Communication Library	NCCL	RCCL
C++ Core library	Thrust	rocThrust
	CUB	hipCUB
Wave Matrix Multiply Accumulate Library	WMMA	rocWMMA
Deep Learning / Machine Learning primitives	cuDNN	MIOpen
C++ templates abstraction for GEMMs	CUTLASS	Composable Kernel (CK)

Портал AMD INFINITY HUB

Готовые к использованию HPC/ML контейнеры

ИТМО

Поддержка AMD Instinct™ GPU

Уже более 30 приложений с постоянным ростом общего числа

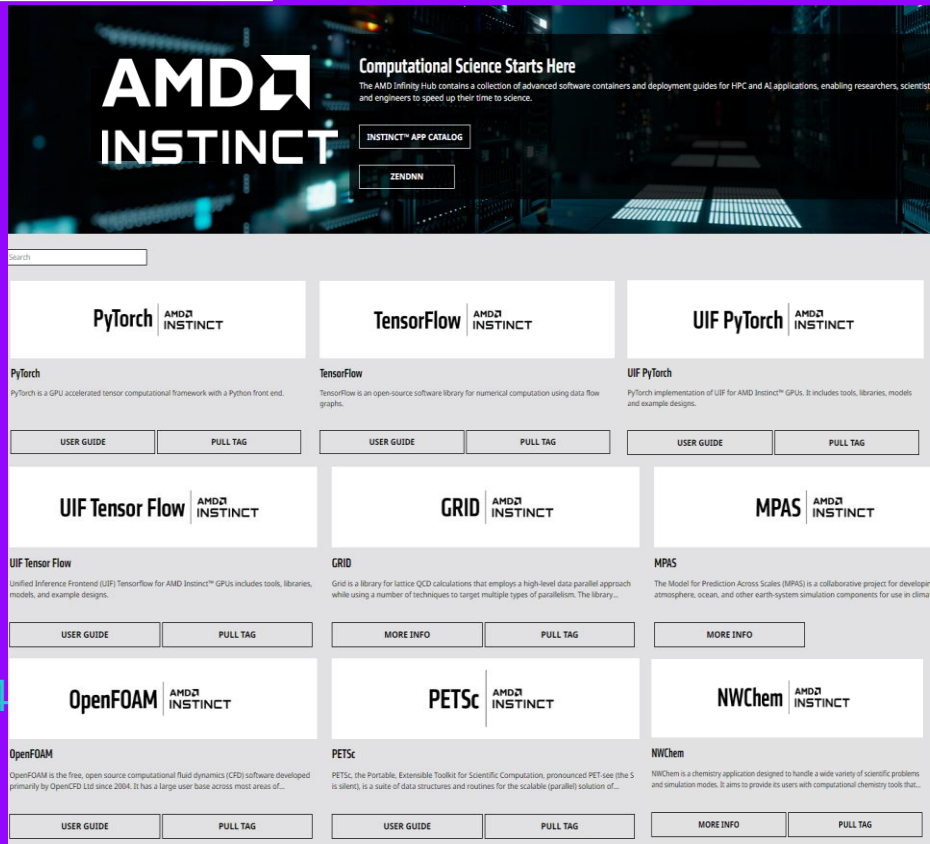
AI – фреймворки, AI – модели, Молекулярная динамика, Квантовая химия, Нефте-газовые задачи

Руководства пользователя и рецепты построения

Новые версии доступны на GitHub

Больше приложений будет на GitHub в 2024

AMD.com/InfinityHub



AMD GITHUB – Готовые рецепты для HPC

Открытый код + Рецепты → Быстрое освоение экосистемы

ИТМО

- AMD – единственный поставщик открытого кода библиотек HPL и HPCG

2023									
Jan-Feb	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov-Dec**
	PETSc **HPCG** GROMACS **ROCHPL**	**Base** **ROCm** LAMMPS	OpenMM OpenFOAM	PyFR SPECFEM3D	CHOLLA GRID	MILC CHROMA	Kokkos	PIConGPU QUDA	HPCToolkit* MPAS Refresh for MI300

- Заказчики могут создавать собственные контейнеры согласно инструкций

To run a custom configuration, include one or more customized build-arg

DISCLAIMER: This Docker build has only been validated using the default values. Using a different base image or branch may result in build failures or poor performance.

```
docker build \
    -t mycontainer/rochpcg \
    -f /path/to/Dockerfile \
    --build-arg IMAGE=rocm/dev-ubuntu-20.04:5.2.3-complete \
    --build-arg ROCHPCG_BRANCH=devel \
    --build-arg UCX_BRANCH=master \
    --build-arg OMPI_BRANCH=main \
    .
```

Каталог научно-технических приложений

Масштабная адаптация таких приложений на архитектуру AMD Instinct™ MI300A

- Ported/Optimized
- 1H '24
- 2H '24
- 1H '25

ИТМО

Life Science

GROMACS
RELION
AMBER
LAMMPS
AlphaFold
NAMD

Physics

HACC
M-AIA
CHROMA
MILC
QUEST
Quicksilver
Open-GADGET

CFD

OpenFOAM
Parthenon
NEKBONE
LAGHOS
PELE-LM
NS3D-NEO
Athena-PK

Chemistry

QMCPACK
CP2K
Open Catalyst
VASP
QUANTUM ESPRESSO

Benchmarks

BabelStream
OSU
RAJAPerf
HPCG
HPL || HPL-MxP
Pennant
Kripke

Earth Science

FastEddy
TOAST3
Octopus
COSMOFLOW
DeepCAM

Libraries

HYPRE
TRILINOS
Kokkos Kernels
RAJA
OCCA
ELPA
AMG

ISV

ANSYS FLUENT
ANSYS MECHANICAL
CADENCE CharLES
SIEMENS StarCCM+

Инструменты отладки, профилирования и анализа в экосистеме ROCm™

ИТМО

Инструментирование приложений	HPC Apps				ML Frameworks				
Визуализация и анализ	*Vampir	Trace Compass	*Apprentice 2	Babel Trace	Chromium	Perfetto	Grafana	Scalasca	Periscope
Инструменты сторонних разработчиков	*Linaro Forge MAP/DDT	*Perforce TotalView	*HPE CrayPAT & gdb4hpc	Argonne THAPI	LLNL Caliper	TAU + ParaProf	Score-P	HPCToolkit + hpcviewer	Theia IDE
Инструментарий и API интегрированный с ROCm	ExaPAPI	LIKWID	OmniTrace	OmniPerf	LTTng ExaTracer				
Инструменты и API в ROCm	ROCProfiler SDK	ROCgdb & ROCdbgapi	OMPT	ROCm™ SMI	ROCm™ Validation Suite				
Среды выполнения API	HIP Runtime	OpenMP Runtime	OpenCL™ Runtime	Python Runtime					
	ROCr HSA Runtime								
Системный уровень	Linux Core	KFD Kernel Driver	LTTng™	MPI					
Апппаратная A+A платформа	**CPU + GPU + Memory Hierarchy**								

Спасибо
за внимание!

ITMO more than a UNIVERSITY