AMD EPYC Server Processor Architecture & Heterogeneous Computing Platforms Configurations

ПАВЕЛ СТАНАВОВ
PAVEL.STANAVOV@AMD.COM

ТИМУР ПОЛТАШЕВ
TIMOUR.PALTASHEV@AMD.COM
КОНКУРЕНЦИЯ ВОЗВРАЩАЕТСЯ В ДАТАЦЕНТРЫ

- X86 архитектура
- Производительность
- Безопасность, Простота, TCO
- Поддержка
Выдающаяся производительность помогает развивать бизнес

32 ядра на сокет

128 PCIe Gen 3 lanes с одного процессора

8 Каналов памяти

2ТБ RAM на сокет

Широкий выбор без потери функциональности

Максимальное количество возможных портов В/В

Самая высокая пропускная способность шины памяти

Самая высокая плотность памяти
<table>
<thead>
<tr>
<th>Модель EPYC™</th>
<th>Ядра/потоки</th>
<th>Частота базы/повышение</th>
<th>Мемориальная скорость</th>
<th>Мемориальная емкость</th>
<th>Количество каналов DDR4</th>
<th>128 Lanes PCIe®</th>
<th>Поддержка SMT и Turbo Boost</th>
<th>Поддержка HW Encrypted Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPYC™ 7601</td>
<td>32/64</td>
<td>2.2 / 3.2 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7551</td>
<td>32/64</td>
<td>2.0 / 3.0 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7501</td>
<td>32/64</td>
<td>2.0 / 3.0 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7451</td>
<td>24/48</td>
<td>2.3 / 3.2 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7401</td>
<td>24/48</td>
<td>2.0 / 3.0 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7371</td>
<td>16/32</td>
<td>3.1 / 3.6 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7351</td>
<td>16/32</td>
<td>2.4 / 2.9 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7301</td>
<td>16/32</td>
<td>2.2 / 2.7 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7281</td>
<td>16/32</td>
<td>2.1 / 2.7 GHz</td>
<td>8 Channels DDR4-2666</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7261</td>
<td>8/16</td>
<td>2.5 / 2.9 GHz</td>
<td>8 Channels DDR4-2400</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EPYC™ 7251</td>
<td>8/16</td>
<td>2.1 / 2.9 GHz</td>
<td>8 Channels DDR4-2400</td>
<td>2TB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Мировой рекорд BENCHMARKS!

EPYC 7601

SPECfp_rate2006
SPECfp_rate2017

Больше ядер

14%

Выше пропускная способность памяти

33%

Больше объем памяти

2.6x

EpYC 7601

UP TO 2.6x

Больше производительности / $
БЕСКОМПРОМИСНЫЙ ОДИН СОКЕТ

Intel
2x 5118

128 LANEs
8 DIMMS
4 MEMORY CHANNELS
HIGH SPEED I/O

EPYC 7551P

33% Больше ядер
33% Больше объем памяти
33% Выше производительность

~37% Меньше энергопотребление
11% Выше производительность
20% Выше производительность / $
ЕРУС – РАЗРАБОТАН С УЧЕТОМ ПОТРЕБНОСТЕЙ СОВРЕМЕННЫХ ДАТАЦЕНТРОВ

VIRTUALIZATION & VDI
• Больше ядер для пользователя
• Больше памяти – больше виртуальных машин на сервер
• Аппаратное шифрование на всех этапах
• Много портов ввода/вывода для масштабирования

SW-DEFINED STORAGE
• Direct SATA & NVMe поддержка
• Высокий уровень параллелизма
• Больше памяти для кэширования данных
• Шифрование памяти без потери производительности

HPC
• Оптимизация для графических ускорителей
• Много портов для создания кластеров
• Объем памяти для данных
• Прямое подключение большого количества NVMe
ЕРУС ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛЕНИЙ (НРС)

ОСОБЕННОСТИ

- Лидирующее количество ядер
  - До 32 на сокет
- Максимальное количество памяти на сокет
  - До 2 ТБ
- Высокая пропускная способность памяти
  - Свыше 290GB/s по тесту STREAM Triad
- Лидирующие I/O возможности
  - 128 PCIe Gen 3
- Встроенный процессор безопасности

ПРЕИМУЩЕСТВА для НРС

- Параллелизм для перемалывания больших объемов данных
- Быстрый доступ к данным
- Возможность быстро загружать в память данные для обработки
- Быстрый прямой доступ к устройствам
- Безопасность данных
EPYC Delivers up to 77% Better Performance than Broadwell

- EPYC 7601, Weather Research Forecast: Up to 77%
- EPYC 7601, ANSYS Fluent: Up to 76%
- EPYC 7601, SPEC OMP: Up to 69%
- Intel Xeon E5-2699 v4

WHY
- More Memory Bandwidth
- More High-Performance Cores
- Balanced Architecture

IDEAL FOR MEMORY BOUND HPC
- Computational Fluid Dynamics
- Atmospheric Research and Weather Modeling
- Oil and Gas Exploration
- Crash Simulation

- Mesoscale numerical weather prediction engine based on Eulerian Mass solver
- Simulating the turbulent airflow around an aircraft
- 14 million cell problem
- 14 scientific and engineering application codes
- molecular dynamics, pattern matching, CFD, linear algebra, image manipulation, finite elements
FLOATING POINT ЛИДЕРСТВО В ТЕСТАХ ПАРТНЕРОВ
VS XEON PLATINUM 8176

**NAMD Molecular Dynamics 2.10**
- EPYC 7601; 5,8
- Intel Xeon Platinum 8176; 4,1
- Intel Xeon E5-2699 v4; 3,9

**C-ray rendering at 3840x2160**
- EPYC 7601; 526
- Intel Xeon Platinum 8176; 340
- Intel Xeon E5-2699 v4; 263

**POV-ray**
- EPYC 7601; 249
- Intel Xeon Platinum 8176; 214
- Intel Xeon E5-2699 v4; 188

**+41%**
Measures: FP performance, with frequent use of memory subsystem

**+55%**
Measures: FP performance, primarily out of L1 cache

**+16%**
Measures: FP performance, primarily out of L2 cache

- Больше ядер
- Эффективный FP блок
- Высокая пропускная способность памяти
<table>
<thead>
<tr>
<th>ДОСТОИНСТВА</th>
<th>Лидер по набору характеристик</th>
<th>Плотность ядер</th>
<th>Объем памяти</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Лидерство в двухпроцессорных решениях</td>
<td>Количество каналов памяти</td>
<td>Порты ввода/вывода</td>
</tr>
<tr>
<td></td>
<td>Бескомпромиссное односокетное решение</td>
<td>Масштабируемость по ядрам</td>
<td>Пропускная способность памяти</td>
</tr>
<tr>
<td></td>
<td>Безопасность</td>
<td>Плотность виртуальных машин</td>
<td>Производительность на $</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Выгодное предложение</td>
<td>Объем памяти</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Набор портов</td>
<td>Производительность/$/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TCO вне конкуренции</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Безопасная загрузка</td>
<td>AMD Шифрование виртуальных машин (SEV)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AMD Шифрование данных в памяти (SME)</td>
<td></td>
</tr>
</tbody>
</table>
SECURITY BEYOND THE STATUS QUO

SECURE ROOT-OF-TRUST TECHNOLOGY

Boot from a secure root of Trust

SECURE MEMORY ENCRYPTION (SME)

Helps protect from memory hacks and scrapes

SECURE ENCRYPTED VIRTUALIZATION (SEV)

Encrypt and Isolate Virtual Machines

SECURITY BEYOND THE STATUS QUO
ГЛОБАЛЬНЫЕ КЛИЕНТЫ

Microsoft Azure
Baidu 百度
Tencent Cloud
Dropbox
Yahoo Japan
Packet
HiveVelocity

NПC, АНАЛИТИКА И SDS
Technological University of Denmark

ВИРТУАЛИЗАЦИЯ
Непрерывные инновации

Лидерство в производительности

“Naples”
- “Zen”
- 14nm

“Rome”
- “Zen 2”
- 7nm

“Milan”
- “Zen 3”
- 7nm+
Вопросы?
WORLD’S FIRST 7NM GPU
ULTRA FAST, PERFECTLY BALANCED
MAKING A BETTER ACCELERATOR

SPEED

AMD Radeon Instinct™ MI25

AMD Radeon Instinct™ MI50 and MI60

ACCURACY

INT8: INference
FP16: TRAINING
FP32: VISUALIZATION
FP64: HPC
<table>
<thead>
<tr>
<th>AMD RADEON INSTINCT™ MI60 AND MI50</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PRODUCT DETAILS</strong></td>
</tr>
</tbody>
</table>

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

*See Endnotes

**Pending
AMD RADEON INSTINCT™ MI60 AND MI50
POWERED BY THE "VEGA" ARCHITECTURE 7NM GPU

3.2B Transistors, 331 mm²

Flexible, High-performance Compute Engine

Scalable Interconnect & Communications

End-to-End ECC* Enablement

*See Endnotes
ULTRA FAST FOR AI WORKLOADS

Up to 29.5 TFOPS FP16

Up to 59 TOPS INT8

Up to 118 TOPS INT4

NEW INSTRUCTIONS:
- 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

MI60 ~20% Faster

MI25 ~140% Faster

~380% Faster
THE WORLD’S FIRST 1TB/S GPU

Up to 32GB 2x
2nd Gen High Bandwidth Memory Memory Bandwidth of MI25

Optimized for Huge Datasets
Built-in Error Detection and Correction with ECC*

*See Endnotes
OPTIMIZED ARCHITECTURE, HIGHER PERFORMANCE

Matrix Multiplication – FP16

- Instinct MI25
- Instinct MI60

>1.25X MORE

Resnet-50 FP16

- Instinct MI25
- Instinct MI60

UP TO 2.8X HIGHER

*See Endnotes
GENERATIONAL GAINS VS. INSTINCT MI25 AND STRONG MULTI-GPU SCALING

Generational Gains vs. MI25

Near Linear Scaling with Resnet-50

*See Endnotes
WIDER APPLICABILITY FOR HPC

Error Correcting Code*

Memory before

1 0 1 0 1

Memory after

1 0 1 1 1

Lowering voltage
Increasing heat
Cosmic radiation

Memory Error Corrected

Double Precision Performance

Exponent (11 bit)
Sign
Fraction (52 bit)

FP64

Reliability, Availability, Serviceability

Cosmic radiation

Lowering voltage
Increasing heat

Memory Error Corrected

FP32

*See Endnotes
SCALABILITY WITH PCIe® GEN4 CAPABLE GPU*

64GB/s
Bi-directional CPU-to-GPU Bandwidth

2x Faster
Than PCIe® Gen3

*Pending. See Endnotes
HIGHLY SCALABLE WITH INFINITY FABRIC™ LINK

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

Up to 6x
Speed of PCIe Gen 3

High speed Peer-to-Peer GPU Communications on a PCIe® Form- Factor
ROCm PROVIDES A COMPETITIVE OPEN SOURCE FOUNDATION

### Radeon Open Ecosystem

<table>
<thead>
<tr>
<th>Applications</th>
<th>Machine Learning Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frameworks</td>
<td>Caffe 2</td>
</tr>
<tr>
<td>Middleware and Libraries</td>
<td>MIOpen</td>
</tr>
</tbody>
</table>

| ROCm | OpenMP | HIP | OpenCL™ | Python |

| Devices | GPU | CPU | APU | DLA |

Supports Latest Machine Learning Frameworks

Optimized Deep Learning & Math Libraries

Support for Dockers and Kubernetes

ROCm Up-Streamed for Linux Kernel Distributions
Programming Models

- HIP
- OpenMP
- Python
- OpenCL

TVM: https://github.com/dmlc/tvm/tree/master/src/runtime/rocm
LLVM: https://llvm.org/docs/AMDGPUUsage.html
CLANG HIP: https://clang.llvm.org/doxygen/HIP_8h_source.html

MI Graph Compilers

- XLA
- TVM

TVM: https://github.com/dmlc/tvm/tree/master/src/runtime/rocm
LLVM: https://llvm.org/docs/AMDGPUUsage.html

FOUNDATION: COMMON COMPILER PROVIDES MANY PROGRAMMING OPTIONS FOR HPC AND ML

FOUNDATION: COMMON COMPILER PROVIDES MANY PROGRAMMING OPTIONS FOR HPC AND ML
FOUNDATION: UPSTREAM LINUX KERNEL SUPPORT FACILITATES ROCM BASED APPLICATIONS

Linux Kernel 4.17

245+ upstream ROCm driver commits since 4.12 kernel

https://github.com/RadeonOpenCompute/ROCK-Kernel-Driver
THANK YOU
Slide 8:
As of Oct 22, 2018. The results calculated for Radeon Instinct MI60 designed with Vega 7nm FinFET process technology resulted in 118 TOPS INT4, 59 TOPS INT8, 29.5 TFLOPS half precision (FP16), 14.8 TFLOPS single precision (FP32) and 7.4 TFLOPS double precision (FP64) peak theoretical floating-point performance. The results calculated for Radeon Instinct MI50 designed with Vega 7nm FinFET process technology resulted in 107 TOPS INT4, 53.6 TOPS INT8, 26.8 TFLOPS peak half precision (FP16), 13.4 TFLOPS peak single precision (FP32) and 6.7 TFLOPS peak double precision (FP64) floating-point performance. The results calculated for Radeon Instinct MI25 GPU based on the “Vega10” architecture resulted in 24.6 TFLOPS peak half precision (FP16), 12.3 TFLOPS peak single precision (FP32) and 768 GFLOPS peak double precision (FP64) floating-point performance. AMD TFLOPS calculations conducted with the following equation for Radeon Instinct MI25, MI50, and MI60 GPUs: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per GPU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, that number is multiplied by 2 FLOPS per clock for FP32, 4 FLOPS per clock for FP16, 8 FLOPS per clock for INT8, and 16 FLOPS per clock for INT8 to determine TFLOPS or TOPS. The FP64 TFLOPS rate or MI50 and MI60 is calculated using 1/2 rate. The FP64 TFLOPS rate for MI25 is calculated using 1/16th rate. RIV-7

Slide 10:

Chart 2: Testing Conducted by AMD performance labs on October 31, 2018, on a system comprising of Dual Intel Xeon Gold 6132, 256GB DDR4 system memory, Ubuntu 16.04.5 LTS, AMD Radeon Instinct MI25 graphics, AMD Radeon Instinct MI60 graphics, ROCm 19.224 driver, TensorFlow 1.11. Benchmark application: Resnet50 FP16 batch size 256. AMD Radeon Instinct MI25 = 179 images/s. AMD Radeon Instinct MI60 = max 498.99 images/s. Performance differential: 498.99/179 = up to 2.8x more performance than Radeon Instinct MI25. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations RIV-8
Testing Conducted by Keith Vanderlinde/CHIME on October 30, 2018, on a system comprising of Dual Socket Intel Xeon Gold 6132, 256GB DDR4 system memory, Ubuntu 16.04.5 LTS, AMD Radeon Instinct MI60 graphics, ROCm 19.211 driver. Benchmark application: Internal application – Simple OpenCL kernel data processing using SDOT8. AMD FirePro S9300x2 = 70ms. AMD Radeon Instinct MI60 = 14ms. Performance differential: 70/14 = 500% improvement over AMD FirePro S9300x2. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. RIV-9

Testing Conducted by AMD performance labs on October 15, 2018, on a system comprising of Dual Intel E5-2640 v4, GB DDR4 system memory, Ubuntu 16.04.4 LTS, AMD Radeon Instinct MI25 graphics, AMD Radeon Instinct MI60 graphics running at 1500e/1000m, ROCm 1.9.211 driver, TensorFlow 1.10 FP32. Benchmark application: Resnet50 batch size 64. AMD Radeon Instinct MI25 = 171.54 images/s. AMD Radeon Instinct MI60 = 257.28 images/s. Performance differential: 257.28/171.54 = 49.9% more performance than Radeon Instinct MI25. Benchmark application: Inception4 batch size 32. AMD Radeon Instinct MI25 = 37.03 images/s. AMD Radeon Instinct MI60 = 56.34 images/s. Performance differential: 56.34/37.03 = 52.1% more performance than Radeon Instinct MI25. Benchmark application: VGG16 batch size 64. AMD Radeon Instinct MI25 = 90.05 images/s. AMD Radeon Instinct MI60 = 131.78 images/s. Performance differential: 131.78/90.05 = 46.3% more performance than Radeon Instinct MI25. Benchmark application: Alexnet batch size 128. AMD Radeon Instinct MI25 = 1547.99 images/s. AMD Radeon Instinct MI60 = 1923.58 images/s. Performance differential: 1923.58/1547.99 = 24.3% more performance than Radeon Instinct MI25. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. RIV-10

Testing Conducted by AMD performance labs on October 31, 2018, on a system comprising of Dual Intel Xeon Gold 6132, 256GB DDR4 system memory, Ubuntu 16.04.5 LTS, AMD Radeon Instinct MI60 graphics running at 1600e/500m, ROCm 19.224 driver, TensorFlow 1.11. Benchmark application: Resnet50 FP32 batch size 256. 1x AMD Radeon Instinct MI60 = 278.63 images/s, 2x Radeon Instinct MI60 = 553.98 images/s. Performance differential: 553.98/278.63 = 1.99x times more performance than 1x Radeon Instinct MI60. 4x Radeon Instinct MI60 = 1109.24 images/s. Performance differential: 1109.24/278.63 = 3.98x times more performance than 1x Radeon Instinct MI60. 8x Radeon Instinct MI60 = 2128.33 images/s. Performance differential: 2128.33/278.63 = 7.64x times more performance than 1x Radeon Instinct MI60. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. RIV-11
ENDNOTES

Slide 14:
Testing Conducted by AMD performance labs on October 31, 2018, on a system comprising of Dual Socket Intel Xeon Gold 6130, 256GB DDR4 system memory, Ubuntu 16.04.5 LTS, AMD Radeon Instinct MI25 graphics, AMD Radeon Instinct MI60 graphics, ROCm 19.211 driver. Benchmark application: rocBLAS DEGEMM N=M=K=5760. AMD Radeon Instinct MI25 GFLOPS = 763. AMD Radeon Instinct MI60 TFLOPS = 6.717. Performance differential: 6717/763 = 8.8x more performance than Radeon Instinct MI25. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

Calculated on Oct 22, 2018, the Radeon Instinct MI60 GPU resulted in 7.4 TFLOPS peak theoretical double precision floating-point (FP64) performance. AMD TFLOPS calculations conducted with the following equation: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per GPU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, that number is multiplied by 1/2 FLOPS per clock for FP64. TFLOP calculations for MI60 can be found at https://www.amd.com/en/products/professional-graphics/instinct-mi60

Slide 15:
Chart 1 : Testing Conducted by AMD performance labs on October 31, 2018, on a system comprising of Dual Socket Intel Xeon Gold 6130, 256GB DDR4 system memory, Ubuntu 16.04.5 LTS, NVIDIA Tesla V100 PCIe with CUDA 10.0.130 and CUDNN 7.3, AMD Radeon Instinct MI60 graphics, ROCm 19.224 driver, TensorFlow 1.11. Benchmark application: Resnet50 FP32 batch size 256. NVIDIA Tesla V100 PCIe = 357 images/s. AMD Radeon Instinct MI60 = 334 images/s. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

Chart 2: Testing Conducted by AMD performance labs on October 31, 2018, on a system comprising of Dual Socket Intel Xeon Gold 6130, 256GB DDR4 system memory, Ubuntu 16.04.5 LTS, NVIDIA Tesla V100 PCIe with CUDA 10.0.130 and CUDNN 7.3, AMD Radeon Instinct MI60 graphics with ROCm 19.224 driver. Benchmark application: cuBLAS/rocBLAS SGEMM N=M=K=5760. AMD NVIDIA Tesla V100 PCIe TFLOPS = 13.106. AMD Radeon Instinct MI60 TFLOPS = 14.047. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

Chart 3: Testing Conducted by AMD performance labs on October 31, 2018, on a system comprising of Dual Socket Intel Xeon Gold 6130, 256GB DDR4 system memory, Ubuntu 16.04.5 LTS, NVIDIA Tesla V100 PCIe with CUDA 10.0.130 and CUDNN 7.3, AMD Radeon Instinct MI60 graphics with ROCm 19.224 driver. Benchmark application: cuBLAS/rocBLAS DGEMM N=M=K=5760. AMD NVIDIA Tesla V100 PCIe TFLOPS = 6.627. AMD Radeon Instinct MI60 TFLOPS = 6.717. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations
ENDNOTES

Slide 5, 9, 16 & 17:
As of October 22, 2018. Radeon Instinct™ MI50 and MI60 “Vega 7nm” technology-based accelerators are PCIe Gen 4.0* capable providing up to 64 GB/s Peak bandwidth per GPU card with PCIe Gen 4.0 x16 certified servers. Peak theoretical transport rate performance guidelines are estimated only and may vary. Previous Gen Radeon Instinct compute GPU cards are based on PCIe Gen 3.0 providing up to 32 GB/s peak theoretical transport rate bandwidth performance.

Peak theoretical transport rate performance is calculated by Baud Rate * width in bytes * # directions = GB/s

PCle Gen 3: 8 * 2 * 2 = 32 GB/s
PCle Gen 4: 16 * 2 * 2 = 64 GB/s
Infinity Fabric Link (xGMI) = 25 * 2 * 2 = 100 GB/s * 2 links per GPU = 200 GB/s

PCle Gen 4 vs. Pcie Gen 3: 64/32 = 2x faster
Infinity Fabric Link (xGMI) vs. PCIe Gen 3 = 200/32 = 6.25x faster

Refer to server manufacture PCIe Gen 4.0 compatibility and performance guidelines for potential peak performance of the specified server models. Server manufacturers may vary configuration offerings yielding different results.

https://pcisig.com/
https://www.chipestimate.com/PCI-Express-Gen-4-a-Big-Pipe-for-Big-Data/Cadence/Technical-Article/2014/04/15

AMD has not independently tested or verified external/third party results/data and bears no responsibility for any errors or omissions therein.

Slide 5, 6, 7, 9 & 13:
ECC support on 2nd Gen Radeon Instinct™ GPU cards, based on the “Vega 7nm” technology has been extended to full-chip ECC including HBM2 memory and internal GPU structures.
DISCLAIMER AND ATTRIBUTION

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, FirePro, Instinct, Radeon, 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.
# AMD EPYC AND INTEL XEON X86 INSTRUCTION SET COMPATIBILITY

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Description</th>
<th>EPYC</th>
<th>Broadwell</th>
<th>Skylake</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADX</td>
<td>Extended multi-precision arithmetic support</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>AVX 1.0</td>
<td>New vector, 256 bit instructions</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>AVX 2.0</td>
<td>Additional vector, 512 bit instructions</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>BMI1</td>
<td>Bit manipulation instructions</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>BMI2</td>
<td>Bit manipulation instructions</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CLFLUSHOPT</td>
<td>CLFLUSH ordered by SFENCE</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLZERO</td>
<td>Clear cache line</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMA3</td>
<td>3 operand fused-multiply-accumulate</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FMA4</td>
<td>4 operand fused-multiply-accumulate</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSGBASE</td>
<td>Read or write the FS or GS BASE registers</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>F16C</td>
<td>16 bit floating point conversion</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MOVBE</td>
<td>Useful for big-endian to little-endian swaps</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RDRAND</td>
<td>Random number generation</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RDSEED</td>
<td>Complement to RDRAND random number generation</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SHA1 / SHA256</td>
<td>Secure hash instructions</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMAP</td>
<td>Supervisor Mode Access Prevention</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SMEE</td>
<td>Secure Mode Execution Protection</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SSE4.1, 4.2</td>
<td>Additional SSE instructions</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>XSAVE</td>
<td>Saves YMM (256 bit) register state</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>XSAVEC/S, XRSTORS</td>
<td>New Compact and Supervisor Save/Restore</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XSAVEOPT</td>
<td>Speeds XSAVE during context switch</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TSX</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPX</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>