AMD’s Radeon Next Generation GPU Architecture

2017

“Vega10”
AMD “VEGA10” SOC

14nm FinFET GPU
   Die Size: 19mm x 25.6mm
   Area: 486 sq mm2,
   Transistors: 12.5 Billion

2 Stack HBM2
   4, 8, or 16 GB Capacity
   Up to 484 GB/S with ECC
   2x HBM1 rate with ½ footprint

16x PCIE® Gen 3.0
   2nd Gen SR-IOV GPU Virtualization

Package
   47.5mm x 47.5 mm
   3.42 mm z-height
   Power Envelope:
      150W – 300W
      Idle: <2W
## GPU Architecture Comparison

**Fiji** to **“Vega10”**

<table>
<thead>
<tr>
<th>Metric</th>
<th>“Fiji” Architecture (eclk @ 1.05 GHz)</th>
<th>“Vega10” Architecture (eclk @ 1.677 GHz)</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32 Compute*</td>
<td>8.6 TFLOPS</td>
<td>13.7 TFLOPS</td>
<td>1.6x</td>
</tr>
<tr>
<td>FP16/Integer16*</td>
<td>8.6 TFLOPS</td>
<td>27.5 TFLOPS</td>
<td>3.2x</td>
</tr>
<tr>
<td>External Memory Bandwidth*</td>
<td>512 GB/sec</td>
<td>484 GB/sec</td>
<td>0.95x</td>
</tr>
<tr>
<td>Pixel Fill Rate*</td>
<td>67.2 GPixel/sec</td>
<td>108.8 GPixel/sec</td>
<td>1.6x</td>
</tr>
<tr>
<td>Texture Fill Rate*</td>
<td>269 GTexels/sec</td>
<td>435.2 GTexels/sec</td>
<td>1.6x</td>
</tr>
<tr>
<td>Die Area</td>
<td>596 mm2 (28 nm)</td>
<td>486 mm2 (14 nm)</td>
<td>0.8x</td>
</tr>
<tr>
<td>Transistors</td>
<td>8.9 billion</td>
<td>12.5 billion</td>
<td>1.4x</td>
</tr>
<tr>
<td>FP32 GFLOPS*/mm2</td>
<td>14.4 (28nm)</td>
<td>28.2 (14 nm)</td>
<td>1.96x</td>
</tr>
<tr>
<td>L2 Cache Capacity</td>
<td>2 MB</td>
<td>4 MB</td>
<td>2x</td>
</tr>
</tbody>
</table>

* (Up to) - theoretical peak at listed frequency
Memory System
HBM2
Efficient Memory with ECC

Compared to HBM1

- 2x bandwidth per pin
- 8x capacity / stack
- 3.5x more power efficient
- 75% smaller footprint

Compared to GDDR5

See endnotes for details
HBCC monitors GPU’s memory traffic
Memory pages are migrated across memory locations
Flexible programming model controls caching policies

See endnotes for details
Page-Based Memory Management

Removes the need for complicated memory management

Large resources are not required to remain complete in local memory

Active pages have prioritized residency in HBC

Inactive pages are marked for migration to slower memory

With HBCC

Without HBCC
Infinity Fabric
Infinity Fabric - Scalable Control & Data Fabric

**Scalable Control Fabric**

- GFX
- ACE
- DMA
- Video
- Display

**Scalable Data Fabric**

- HBM
- HBM

**Infinity Fabric Characteristics**

- Customized Topologies
- Low Latency
- QOS Capabilities
- Security Infrastructure
- Virtual Functions
- Power Monitoring
- Coherency Protocols
- Multi-Socket/Die Ready

**GRAPHICS TEXTURE CACHE WRITE LATENCY (Lower is Better)**

- “VEGA10” 1x
- Radeon™ R9 Fury X 3x

**GRAPHICS TEXTURE CACHE READ LATENCY (Lower is Better)**

- “VEGA10” 1x
- Radeon R9 Fury X 1.7x

**UP TO A 67% REDUCTION IN LATENCY**

See endnotes for details
"Vega10" Graphics

1. Graphics Engine
   - Flexible Geometry Engine
   - 4 Draw Stream Binning Rasterizers
   - 64 Pixels Units
   - 256 Texture Units
2. Core Asynchronous Compute Engine
   - Workload Manager
   - 64 Next Gen Compute Unit (NCU)
   - 4 MB L2
3. System DMA Units
4. UVD & VCE Video Engines

See endnotes for details
“VEGA10” 3D GRAPHICS ENHANCEMENTS

4 MB L2 - Double

Pixel Engine
- Draw Stream Binning Rasterizer
- Render Backends are L2 clients

Flexible Geometry Pipeline
- Improved Native Pipeline
- Next Generation Primitive Shader

Direct X 12.1 Features
- Conservative Rasterization
- Raster Ordered Views
- Standard Swizzle
- Axis Aligned Rectangular Primitives
Draw Stream Binning Rasterizer
Designed to improve performance and saves power

Fetch once enabled by smart primitive rasterization with on-chip bin cache

Shade once enabled by culling of pixels invisible to final scene

See endnotes for details
SPECviewperf 12 / energy-01

See endnotes for details
Bytes per frame savings due to DSBR

See endnotes for details
GAMING PERFORMANCE AND POWER GAINS DUE TO DSBR

Radeon™ Vega Frontier Edition XTX DSBR on/off comparisons

Relative performance gains

- Alien Vs Predator
- Battlefield Hardline
- Battlefield 4 (Airfield)
- Battlefield 4 (Naval)
- Metro Last Light Redux 2033: Very High
- Mirrors Edge Catalyst: Ultra
- Mass Effect Andromeda: Ultra (DX11)

Performance gains

Performance per watt gains

See endnotes for details
Single-Root I/O Virtualization

VCE (H.264) and UVD (H.265) encode hardware acceleration now included, decode capable

Supports 16 VM guest containers with native drivers

Auto-hardware scheduling for the three engine sets

See endnotes for details
Accelerated Compute Engine (ACE)

Hypervisor Agent (PCIe® SRIOV)
  VM Guest assignment

Hardware Scheduler
  OS/KMD Coordination
  Per process establishment
  User mode scheduling
  Policy Controls

Four ACE Core
  8 Accelerator Threads each
  Instruction based Preemption

Thread 0 Descriptor
Thread 1 Descriptor
Thread 7 Descriptor

Fetch Decode & Issue
Kernel Work Group Dispatcher
Wave/Fence Issue & Track Resource Alloc

Wave/Fence Resource Alloc

Shader System
NCU Array

Core Fabric

L2

Infinity Fabric
Next-Generation Compute Unit
“Vega10” NCU
Next-Generation Compute Unit

- Full rate IEEE compliant FMA32
- Cross-lane Data Parallel Ops (DPP)
- Shader Instruction Pre-Fetch

Extended ISA with 40 new instructions
ISA spec will be made public

- XAD_U32: hash / cryptocurrencies
- SAT_PK_U8_I16: video processing
- LSHL_ADD_U32
- ADD_LSHL_U32
- ADD3_U32
- LSHL_OR_B32
- AND_OR_B32
- OR3_B32
- addressing
“Vega10” NCU
Next-Generation Compute Unit

Rapid Packed Math
16 bit Math
256 - 16b ops per clock

IEEE compliant FMA
Register Footprint Reduction
Flexible Operand Source Swizzles
Mixed Precision MAD
Packed 16b Image/Buffer Data
16b Image Address Support
Supporting Software
SOFTWARE STACK

Applications
- Machine Learning

Frameworks
- Caffe
- TensorFlow
- Keras
- Caffe2
- MxNet
- Torch 7

Middleware & Libraries
- MIOpen
- BLAS, FFT, RNG
- NCCL
- Eigen
- C++ STL

ROCm
- HCC
- HIP
- OpenCL™
- Python

ROCM Platform

Open-source
- LLVM
- HSA
MIOpen

**High-performance deep learning primitives**

### Key Features
- Convolutions for Inference and Training
- “Inplace” Winograd Solver
- Optimized GEMM for Deep learning
- Pooling Forward & Backwards
- Softmax
- Activation
- Batch Normalization

### Architecture
- HIP and OpenCL top-level APIs
- Kernels in high-level source and GCN asm
  - Documented ISA with open-source tools

### Benefits from “Vega10” include:
- Packed FP16 (>25 Tflops)
- Cross-lane “DPP” instructions
- LDS Scratchpad memory (>13 TB/s)
TensorFlow ImageNet Performance

ImageNet classification with “Googlenet” network forward+backward time. Vega10 Radeon Instinct Engineering Sample (1.63Ghz clock).

See endnotes for details.
Scalability
EPYC™ + MI25 – Optimized for Massive System Scalability

- **128 PCIe® links/CPU**
  - Removes PCIe switches
- **Full PCIe P2P support**
- **32c/CPU for I/O and compute balance**
- Provides strong I/O connectivity and bandwidth with single high-performance CPU
Radeon™ RX Vega 56
Radeon™ RX Vega 64
Radeon™ RX Vega 64 Liquid Cooled
<table>
<thead>
<tr>
<th>Next Gen Compute Units¹</th>
<th>64</th>
<th>64</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stream Processors</td>
<td>4096</td>
<td>4096</td>
<td>3584</td>
</tr>
<tr>
<td>Base GPU Clock</td>
<td>1406 MHz</td>
<td>1247 MHz</td>
<td>1156 MHz</td>
</tr>
<tr>
<td>Boost GPU Clock</td>
<td>1677 MHz</td>
<td>1546 MHz</td>
<td>1471 MHz</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>484 GB/s</td>
<td>484 GB/s</td>
<td>410 GB/s</td>
</tr>
<tr>
<td>Peak SP Performance</td>
<td>13.7 TFLOPS</td>
<td>12.66 TFLOPS</td>
<td>10.5 TFLOPS</td>
</tr>
<tr>
<td>Peak Half Precision Performance</td>
<td>27.5 TFLOPS</td>
<td>25.3 TFLOPS</td>
<td>21 TFLOPS</td>
</tr>
<tr>
<td>High Bandwidth Cache (HBM2)</td>
<td>8GB</td>
<td>8GB</td>
<td>8GB</td>
</tr>
<tr>
<td>Board Power</td>
<td>345W</td>
<td>295W</td>
<td>210W</td>
</tr>
</tbody>
</table>
# RADEON RX VEGA FAMILY

## PACKS

<table>
<thead>
<tr>
<th>Price</th>
<th>Pack Name</th>
<th>Graphics Card</th>
<th>Cooled Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>$699</td>
<td>Radeon Aqua Pack</td>
<td>Radeon RX Vega 64</td>
<td>Liquid Cooled</td>
</tr>
<tr>
<td>$599</td>
<td>Radeon Black Pack</td>
<td>Radeon RX Vega 64</td>
<td>Air Cooled</td>
</tr>
<tr>
<td>$499</td>
<td>Radeon Red Pack</td>
<td>Radeon RX Vega 56</td>
<td>Air Cooled</td>
</tr>
</tbody>
</table>

## GRAPHICS CARDS

<table>
<thead>
<tr>
<th>Price</th>
<th>Graphics Card</th>
<th>Cooled Type</th>
<th>Bundled Games</th>
</tr>
</thead>
<tbody>
<tr>
<td>$499</td>
<td>Radeon RX Vega 64</td>
<td>Air Cooled</td>
<td>No Bundled Games</td>
</tr>
<tr>
<td>$399</td>
<td>Radeon RX Vega 56</td>
<td>No Bundled Games</td>
<td></td>
</tr>
</tbody>
</table>
INTRODUCING RADEON™ PACKS

$200 USD OFF

Radeon™ FreeSync Enabled Monitor

$100 USD OFF

Select AMD Ryzen™ 7 CPU & Motherboard Combo

$120 USD VALUE

2 Free Games (Varies by Region)
RADEON PRO
WX 9100

Up to 12.3 TFLOPS FP32
6, 4K DisplayPort 1.4 HDR Ready Displays*
16GB HBC Memory with ECC*
484 GB/s Memory Bandwidth
$2199 MSRP

*See Endnotes
RADEON PRO SSG

- **Read Performance** from SSG: Up to 8 GB/s
- **Write Performance** to SSG: Up to 6 GB/s
- **Onboard SSG Memory**: 2 TB
- **Real-Time**: 8K
- **MSRP**: $6999
## Product Comparison Table

<table>
<thead>
<tr>
<th></th>
<th>Radeon™ Vega Frontier Edition</th>
<th>Radeon™ Pro WX 9100</th>
<th>Radeon™ Pro SSG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU Architecture</strong></td>
<td>“Vega”</td>
<td>“Vega”</td>
<td>“Vega”</td>
</tr>
<tr>
<td><strong>Peak Compute (FP32)</strong></td>
<td>Up to 13.1 TFLOPS</td>
<td>Up to 12.3 TFLOPS</td>
<td>Up to 12.3 TFLOPS</td>
</tr>
<tr>
<td><strong>Peak Compute (FP16)</strong></td>
<td>Up to 26.2 TFLOPS</td>
<td>Up to 24.6 TFLOPS</td>
<td>Up to 24.6 TFLOPS</td>
</tr>
<tr>
<td><strong>Native Display Outputs</strong></td>
<td>3x DisplayPort™ 1.4 HDR Ready* 1x HDMI™ 4K60</td>
<td>6x DisplayPort™ 1.4 HDR Ready*</td>
<td>6x DisplayPort™ 1.4 HDR Ready*</td>
</tr>
<tr>
<td><strong>Total Board Power</strong></td>
<td>&lt;300W (Air) &lt;350W (Liquid)</td>
<td>&lt;250W</td>
<td>&lt;300W</td>
</tr>
<tr>
<td><strong>Total Onboard Memory</strong></td>
<td>16GB HBC</td>
<td>16GB HBC</td>
<td>16GB HBC + 2TB SSG</td>
</tr>
<tr>
<td><strong>ECC</strong></td>
<td>No</td>
<td>Yes*</td>
<td>Yes*</td>
</tr>
<tr>
<td><strong>ISV Certification</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Warranty</strong>*</td>
<td>1 Year Limited Warranty</td>
<td>3 Year Limited + Optional 7 Year Extended Warranty</td>
<td>2 Year Limited Warranty</td>
</tr>
<tr>
<td><strong>MSRP</strong></td>
<td>$999 (Air) $1499 (Liquid)</td>
<td>$2199</td>
<td>$6999</td>
</tr>
</tbody>
</table>

*See Endnotes
1 Petaflop Single Precision
2 Petaflops Half Precision
30 Gigaflops/Watt (Single Precision)
20X AMD EPYC 7601 CPU
20X Mellanox 100G IB Cards + 1 Switch
80X Radeon Instinct MI25 Accelerators
Slide 5
75% smaller footprint is based on Vega 10 package size with HBM2 (47.5 mm x 47.5 mm) vs. total PCB footprint of R9 290X GPU package + GDDR5 memory devices and interconnects (110 mm x 90 mm).
8x capacity per stack is based on maximum of 8 GB per stack for HBM2 vs. 1 GB per stack for GDDR5.
3.5x power efficiency is based on measured memory device + interface power consumption for R9 390X (GDDR5) vs. RX Vega 64 (HBM2).

Slide 6
Based on AMD Internal testing of an early Vega sample using an AMD Summit Ridge pre-release CPU with 8GB DDR4 RAM, Vega GPU, Windows 10 64 bit, AMD test driver as of Dec 5, 2016. Results may vary for final product, and performance may vary based on use of latest available drivers. VG-4

Slide 7
This feature (Inclusive Cache Model) is still in development and may be better utilized in future releases of Radeon Software, SDKs available via GPUOpen, or updates from the owners of 3D graphics APIs.

Slide 10
Testing conducted by AMD Engineering as of December 5, 2016 on a test system comprising Intel Core i7 6700K at 8GB DDR4 memory at 2667Mhz using a Radeon Fury X and an early sample of Vega. Measuring graphics to texture cache read latency, the Fury X took 201ns and the Vega took 118ns. Measuring graphics to texture cache write latency, the Fury X took 201ns and the Vega took 67ns. Results may vary for final product, and performance may vary based on use of latest available drivers. VG-1

Slide 12
Discrete AMD Radeon™ and FirePro™ GPUs based on the Graphics Core Next architecture consist of multiple discrete execution engines known as a Compute Unit (“CU”). Each CU contains 64 shaders (“Stream Processors”) working together. GD-78

Slide 14
DSBR can reduce the bandwidth or pixel shading required for content that has sequential opaque depth complexity. Results of bandwidth and power savings is illustrated on slide 15, 16, 17

Slide 15
SPECviewperf performance for DSBR: Data based on AMD Internal testing of an early Radeon™ Pro WX 9100 sample using an Intel Xeon E5-1650 v3 CPU with 16 GB DDR3 RAM, Windows® 10 64 bit, AMD Radeon Software driver 17.30. Using SPECviewperf 12.1.1 energy-01 subtest, the scores were 8.80 with DSBR off and 18.96 with DSBR on. Results may vary for final product, and performance may vary based on use of latest available drivers.
Slide 16 & 17
Bytes per frame savings for DSBR & Gaming Performance and power gains from DSBR: Data based on AMD Internal testing of the Radeon Vega Frontier Edition using an Intel Core i7-5960X CPU with 16 GB DDR4 RAM, Windows 10 64 bit, AMD Radeon Software driver 17.20. Results may vary for final product and performance may vary based on use of the latest available drivers.

Slide 18
Inclusion of hardware virtualization of UVD decode requires firmware update

Slide 26
Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz with 128GB memory and Radeon Fiji Radeon R9 FURY / NANO Series 985Mhz.
AMD(R) “Threadripper” AMD Ryzen Threadripper 1950X 16-Core Processor 2200Mz with Radeon Vega10 Engineering Sample 1630 Mhz
Results may vary for final product, and performance may vary based on use of latest available ROCm drivers, MIOpen libraries, and TensorFlow Frameworks
The data was collected using Ubuntu 16.04 ROCm 1.6.3 plus development versions of MIOpen and TensorFlow.
The benchmark is the “tensorflow/bench_googlenet.py” test from https://github.com/soumith/convnet-benchmarks.git

Slide 35
As of June 2017. Product is based on the DisplayPort 1.4 Specification published February 23, 2016, and has passed VESA's compliance testing process (excluding HDR) in June 2017. GD-123
ECC support is limited to the HBM2 memory and ECC protection is not provided for internal GPU structures.

Slide 37
As of June 2017. Product is based on the DisplayPort 1.4 Specification published February 23, 2016, and has passed VESA's compliance testing process (excluding HDR) in June 2017. GD-123
ECC support is limited to the HBM2 memory and ECC protection is not provided for internal GPU structures.

For warranty information, visit www.amd.com/warranty, for extended warranty information, visit www.amd.com/extendedwarranty
DISCLAIMERS AND ATTRIBUTIONS

DISCLAIMER
The information contained herein is for informational purposes only, and is subject to change without notice. Timelines, roadmaps, and/or product release dates shown in these slides are plans only and subject to change. “Vega” is a codename for AMD architectures, and are not product names.

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 non-infringement, 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.

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

SPEC® and the benchmark named SPECint® and SPEC CPU and are registered trademarks of the Standard Performance Evaluation Corporation. For more information about SPECint and SPEC CPU see www.spec.org.

©2017 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Epyc, Radeon, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Microsoft and DirectX are trademarks or registered trademarks of Microsoft Corporation in the US and other jurisdictions. OpenCL is a trademark of Apple Inc. used by permission by Khronos. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.