7NM “NAVI” GPU
A GPU BUILT FOR PERFORMANCE AND EFFICIENCY

MIKE MANTOR
RTG Chief GPU Architect
SETTING THE FOUNDATION FOR GREAT GAMING PRODUCTS: RDNA

- 7nm
  - 251 sqmm
  - 10.3 Billion Transistors

- X16 PCIe® Gen 4.0

- 8 GB GDDR6 256b @14 Gbps
  - 448 GB/S*

- 2560 Stream Processors
  - Up To 9.75 TFLOPs

*256 pin G6 * 14 Gbps * 1B/8b = 448 GBS
2019 “NAVI”

KEY TECHNOLOGY INFLECTIONS

PROCESS
7NM
Faster, Smaller, Lower Power Transistors

DRAM
GDDR6
Cost Effective 448 GB/S of Memory Bandwidth

INTERCONNECT BANDWIDTH
PCIE® 4.0 Support
Up to 2X Interconnect Bandwidth Of PCIE®Gen3

ARCHITECTURE
New GFX RDNA
A GPU Designed For Gaming Performance & Efficiency
"NAVI"

Radeon Display Engine
New High Resolution HDR Displays
New Levels of Compression

Radeon Multi-Media Engine
Seamless Streaming
Improved Encoding

New Graphics RDNA Architecture
New Compute Units
Multilevel Cache
Streamlined Engine
RADEON DISPLAY ENGINE

FEATURING

AMD RADEON FREESYNC™ TECHNOLOGY

- HDMI® 2.0 & DisplayPort 1.4 HDR
- Display Stream Compression 1.2a
- Direct Read of DCC Compressed Surfaces

Optimized for High Resolution HDR Displays

4K 240Hz | SINGLE CABLE | 8K 60Hz

Optimized for Head Mounted Displays

Single IO Connectivity

High Fidelity Internal Color Depth

30 bpp Color

Better Design For Power Efficiency

Multi Plane Overlay Protocol With Low Voltage Mode
# RaeDOn Multimedia Engine

**Seamless Streaming**

**Improved Encoding**
- **New HDR/WCG Encode (HEVC)**
- **8K Decode (HEVC & VP0)**
- **40% Encoder Speedups**

<table>
<thead>
<tr>
<th>Format</th>
<th>Encoder</th>
<th>Decode</th>
<th>Next Gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP9</td>
<td>8b/10b</td>
<td>YouTube</td>
<td>4K90</td>
</tr>
<tr>
<td>H.264</td>
<td>MPEG-4</td>
<td>twitch</td>
<td>1080p600</td>
</tr>
<tr>
<td>H.265</td>
<td>HEVC</td>
<td>NEXT</td>
<td>1080p360</td>
</tr>
</tbody>
</table>

- **YouTube**
  - 4K90
  - 8K24
- **twitch**
  - 1080p600
  - 4K150
- **NEXT GEN**
  - 1080p360
  - 4K90
  - 8K24
  - 4K60

Improved encoding based on AMD internal analysis between Navi and Vega, June 8, 2019.
NEW GRAPHICS ARCHITECTURE

RDNA

GPU Architecture Designed For Gaming Performance & Efficiency

AGENDA

Motivation
Radeon Architecture
New Compute Unit
Multi-Level Cache
Streamlined Engine
Results & Example
"NAVI" RDNA ARCHITECTURE
designed for the future of Radeon gaming

The motivation behind NAVI architecture

Energy efficient performance

- Improve computational efficiency for modern gaming workloads
- Lower latency, high bandwidth with reduced data movement
- Improve operating frequency while reducing power
RDNA ARCHITECTURE

FUNDAMENTAL CHANGES IN PROGRAMMABLE CORE

RADEON™ PRE GCN
- VLIW5/VLIW4
- Complex Compiler Technology
- Hard To Program For Performance
- Per Work-Item IPC = 1.25 Potential

RADEON GCN
- Wave64 on SIMD16 (4clk Issue)
- Standard Compiler Techniques
- Easy To Program For Performance
- Per Work-Item IPC = 0.25

RADEON “NAVI”
- Wave32 On SIMD32 (1clk Issue)
- Enables New Compiler Techniques
- Easier Achieved Performance
- Per Work-Item IPC = 1 Potential
COMPUTE UNIT

IMPROVING SINGLE THREADED PERFORMANCE

2X INSTRUCTION RATE
Dual Schedulers
Dual Scalar Units
Dual SIMD32

SINGLE CYCLE ISSUE
Wave32 on SIMD32
ALU & LD/ST Unit
SFU Co-Execution

BYTES PER FLOP
128B Load/Store
64B Filter Rate

EXECUTION FLEXIBILITY
Wave64 Dual Issue
Cooperating CU Pair
“NAVI" SCALAR \ VECTOR UNIT

Vector Register File
128 KB

Scalar Register File 10 KB
16 KB Scalar Cache

Operand Gathering & xBAR

ALU Unit x32
DP Unit x2
Trans Unit x8

Destination Scheduling
**UP TO 20 WAVE CONTROLLERS**
- Wave32 or 64 Mode
- Improved Instruction Arbitration
- Multi-Lane Primitives

**32 WIDE SINGLE & DUAL HALF ALU**
- Single Cycle Issue
- Full Rate 32b FMA, Dual 16b FMA

**10KB SCALAR REGISTER FILE**
- 128 32b Register Per Wavefront
- Preload Up To 32 User SGPR

**COMPILER ASSIST SCHEDULING**
- W64 – Sub-Vector (1/2 Wave Mode)
- Compiler Driven Clause Support

**8 WIDE TRANSCENDENTAL ALU**
- Single Cycle Issue
- Multi-Cycle Co-Execution

**128 KB VECTOR REGISTER FILE**
- 1024 32b Physical VREGs Per Lane
Both Designs Utilize Multithreading of different waves to achieve throughput and engine utilization.
BENEFITS OF NEW WORK DISTRIBUTION & INSTRUCTION EXECUTION EXECUTION CHANGES
SINGLE THREADED PERFORMANCE IMPROVEMENT

WORK LOAD EXAMPLE: 64 WORK-ITEMS ALU INTENSIVE CODE

<table>
<thead>
<tr>
<th>GCN</th>
<th>RDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>SIMD 0</td>
<td>SIMD 0</td>
</tr>
<tr>
<td>SIMD 1</td>
<td>SIMD 1</td>
</tr>
<tr>
<td>SIMD 2</td>
<td>SIMD 2</td>
</tr>
<tr>
<td>SIMD 3</td>
<td>SIMD 3</td>
</tr>
<tr>
<td>SIMD 0</td>
<td>SIMD 0</td>
</tr>
<tr>
<td>SIMD 1</td>
<td>SIMD 1</td>
</tr>
<tr>
<td>SIMD 2</td>
<td>SIMD 2</td>
</tr>
<tr>
<td>SIMD 3</td>
<td>SIMD 3</td>
</tr>
</tbody>
</table>

1 Wave64 ➔ SIMD16
Instruction Issue ➔ 4 clock
CU ALU ➔ 25% Utilized
Effective Throughput

2 Wave 32 ➔ 2 SIMD32
Instruction Issue ➔ 1 clock
CU ALU ➔ 100% Utilized
ILP unlocks up to 4x faster focused execution

**RDNA MORE EFFECTIVELY UTILIZES THE MACHINE**

Engage Machine Quickly By Uniformly Distributing Work To All ALUs
Optimize Efficiency And Latency By Preferring Highest Priority/Oldest Work
Extract Program ILP And Scheduling To Benefit From Data Locality
Utilize Multi-Threading Of Waves To Hide Remaining Latencies For Throughput
WAVE EXECUTION

SHADERS MUST BE COMPILED SPECIFICALLY FOR WAVE32 OR WAVE64

“RDNA” SHADER BUILT FOR
WAVE32 EXECUTION

- INSTRUCTION ISSUES IN ONE CYCLE
- VECTORS MASKS (VCC, EXEC) ARE 32 BITS
- OPERATES IN SMALLER CACHE FOOTPRINT
- FEWER RESOURCES (GPRS & Cache)
- ENGAGE MORE SIMDS WITH SMALL WORKLOAD
- HIDE TOTAL LATENCY WITH LESS WORK

WAVE64 IS SUPPORTED

- SCALAR INSTRUCTION ARE ISSUED ONCE PER 64 WORK ITEMS
- VECTOR (ALU, MEMORY) INSTRUCTIONS ARE ISSUED TWICE
  - Once For Each Wave32 Half Of The Wave64
  - Either Half Can Be Skipped If EXEC==0
  - Wave-Shared VGPRs With Sub Vector Operating Mode
## INSTRUCTION ISSUE EXAMPLE

### EXAMPLE SHADER

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>s_add_i32</td>
</tr>
<tr>
<td>1</td>
<td>v_mul_f32</td>
</tr>
<tr>
<td>2</td>
<td>v_add_f32</td>
</tr>
<tr>
<td>3</td>
<td>v_sub_f32</td>
</tr>
<tr>
<td>4</td>
<td>v_mul_f32</td>
</tr>
<tr>
<td>5</td>
<td>v_add_f32</td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>v_add_f32</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>v_sub_f32</td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

### "VEGA" EXECUTION

- **Cycle 0**: `s_add_i32 s0, s1, s2`
- **Cycle 1**: ...
- **Cycle 2**: ...
- **Cycle 3**: ...
- **Cycle 4**: `v_mul_f32 v0, v1, s0`
- **Cycle 5**: `v_add_f32 v5, v4, v3`
- **Cycle 6**: ...
- **Cycle 7**: ...
- **Cycle 8**: `v_add_f32 v5, v4, v3`
- **Cycle 9**: ...
- **Cycle 10**: ...
- **Cycle 11**: ...
- **Cycle 12**: `v_sub_f32 v6, v7, v0`
- **Cycle 13**: ...
- **Cycle 14**: ...
- **Cycle 15**: ...

### "NAVI" WAVE32

- **Cycle 0**: `s_add_i32 s0, s1, s2`
- **Cycle 1**: ...
- **Cycle 2**: ...
- **Cycle 3**: ...
- **Cycle 4**: `v_mul_f32 v0, v1, s0`
- **Cycle 5**: `v_add_f32 v5, v4, v3`
- **Cycle 6**: ...
- **Cycle 7**: ...
- **Cycle 8**: `v_add_f32 v5, v4, v3`
- **Cycle 9**: ...
- **Cycle 10**: ...
- **Cycle 11**: ...
- **Cycle 12**: `v_sub_f32 v6, v7, v0`
- **Cycle 13**: ...
- **Cycle 14**: ...
- **Cycle 15**: ...

### "NAVI" WAVE64

- **Cycle 0**: `s_add_i32 s0, s1, s2`
- **Cycle 1**: ...
- **Cycle 2**: ...
- **Cycle 3**: ...
- **Cycle 4**: `v_mul_f32 v0, v1, s0`
- **Cycle 5**: `v_add_f32 v5, v4, v3`
- **Cycle 6**: ...
- **Cycle 7**: ...
- **Cycle 8**: `v_add_f32 v5, v4, v3`
- **Cycle 9**: ...
- **Cycle 10**: ...
- **Cycle 11**: ...
- **Cycle 12**: `v_sub_f32 v6, v7, v0`
- **Cycle 13**: ...
- **Cycle 14**: ...
- **Cycle 15**: ...

**SHORTEST WAVE ISSUE LATENCY**

**44% REDUCTION IN ISSUE CYCLES**
“RDNA” SUB-VECTOR MODE

W64 SUBVECTOR MODE WITH HALF WAVE SHARED REGISTERS

- Wave64 will issue an instruction for each half of the wave64 as two wave32 instructions
- Sub-vector mode will issue a block of instructions for the first half of the wave, followed by the block of instruction for the other half

ADVANTAGES

- Memory operations contained to a smaller units of work and can cache better (multiple entries from strided buffer)
- Wave shared temporary GPRs can be used to reduce VGPRs pressures
  - W64 Assigns Two Work-Items To A Lane. These Two Work-Items Can Both Have Private And Shared VGPRs
  - Registers That Only Exist During Sub-Vector Code Block Can Use Shared VGPRs Instead Of Private

<table>
<thead>
<tr>
<th>SHADER PROGRAM</th>
<th>NORMAL EXECUTION SEQUENCE</th>
<th>SUB-VECTOR LOOP EXECUTION SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>inst0</td>
<td>inst0 - low</td>
<td>inst0 - low</td>
</tr>
<tr>
<td>inst1</td>
<td>inst0 - high</td>
<td>inst1 - low</td>
</tr>
<tr>
<td>inst2</td>
<td>inst1 - low</td>
<td>inst2 - low</td>
</tr>
<tr>
<td>inst3</td>
<td>inst1 - high</td>
<td>inst2 - high</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inst3 - low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inst3 - high</td>
</tr>
</tbody>
</table>
COMPUTE UNIT COOPERATION
WORKGROUP PROCESSOR

UP TO
2X ALUs
ACCESS TO
2X Registers
ACCESS TO
4X Cache Bandwidth
RDNA CACHE

HIERARCHY

NEW MULTI LEVEL SYSTEM

- New L1 Level Cache
- Improved Bandwidth Amplification
- Reduced Latency and Power
- Reduced Congestion at L2 Level
Unified LLC for GFX/ACE Pipes

Instruction Range Based Actions

OOO between R/W, L0, L1, L2, Mem

Reduced Latency and Power

Reduced Data Movement

MULTILEVEL CACHE HIERARCHY
LOW LATENCY, HIGH BANDWIDTH, LOW POWER
MULTILEVEL CACHE HIERARCHY
LOW LATENCY, HIGH BANDWIDTH, LOW POWER

Introduce L1 Cache Hierarchy

Double the Load Bandwidth from L0 to ALU

Relative Cache Latency

-21%
-23%
-8%

Reduce Latency at Each Level

Improve Effective Bandwidth

See RX-329 in Endnotes.
MULTILEVEL CACHE HIERARCHY
DELTA COLOR COMPRESSION EVERYWHERE

Improved Color Compression Algorithm

Shaders Can Read and Write to Compressed Color Data

Display Can Now Read Compressed Color Data

PCle® 4.0

Command Interfaces

Geometry

Shader Complex

Async Compute

Rasterizer & RBs

L1

L2

DISPLAY ENGINE

SOC Fabric

Compressed Data

Texture

GDDR6
STREAMLINED GRAPHICS ENGINE

Improved Architectural Efficiency for Performance

Hyper-Effective Clock Gating for Power Efficiency

Reduced Levels of Logic For Higher Frequency
RDNA DELIVERS

SAME POWER, SAME CONFIGURATION
PERFORMANCE GAINS

+50%

RDNA PERFORMANCE
CONTRIBUTORS

Design Frequency and
Power Improvement

7nm Process Gains

Performance per
Clock Enhancement

Delivered Performance

GCN  RDNA

0  20  40  60  80  100  120  140  160

0%  20%  40%  60%  80%  100%
RDNA

THE FOUNDATION FOR GREAT PRODUCTS

1.5X PERFORMANCE PER WATT

2.3X PERFORMANCE PER AREA

14 nm “Vega64”

495mm²

251mm²

7nm “Navi”
CONTRAST ADAPTIVE SHARPENING (CAS)

IMAGE BASED ALGORITHM SCALING IMAGE SIZE WHILE SHARPENING LOW CONTRAST DETAIL

- Compute Shader Wave32, Work Group 64 In CU Mode For Wide Distribution
- Algorithm Developed For Loop Unrolling To Keep Data Localized To L0 And L1 Caches
- Use Of Low Latency And High Throughput Load Path (Image Loads)
- Packed 16-bit Integer Math For Coordinate Generation And Shader Filtering Logic

EFFICIENT EXECUTION ACHIEVED ON RDNA HARDWARE

- Focused Single Threaded Execution
- Heavy VALU Utilization During Steady State Execution

One SIMD Instruction trace of oldest wave (12), next to oldest wave (13), etc
THE ALL NEW

“NAVI” GPU FAMILY

- **RDNA** Architecture
- **7nm** Process
- **GDDR6** Memory
- **PCIe® 4.0** Support
- **RADEON™** Media Engine
- **RADEON** Display Engine
ENDNOTES

RX-325
Testing done by AMD performance labs 5/23/19, using the Division 2 @ 25x14 Ultra settings. Performance may vary based on use of latest drivers. RX-325

RX-327
Testing done by AMD performance labs 5/23/19, showing a geomean of 1.25x per-clock across 30 different games @ 4K Ultra, 4xAA settings. Performance may vary based on use of latest drivers. RX-327

RX-329
Testing conducted by AMD Performance Labs as of 05/30/2019 on Radeon RX 5700XT with AMD Driver 19.10 (1902270946) on Intel i7-6800k, and on Radeon Vega Frontier Edition with AMD Driver 19.30 (1904231814) on Intel i7-5960x. Both systems used 2x8GB DDR4 2133MHz RAM, Asus ROG Rampage V Edition Motherboard, and Windows 10 Enterprise. Performance may vary. RX-329.

RX-358
Testing done by AMD performance labs on June 4, 2019. Systems were tested with: Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz (6 core) with 16GB DDR4 @ 2133 MHz using a Asus X99-E Motherboard running Windows 10 Enterprise 64-bit (Ver. 1809, build 17763.053). Using the following graphics cards: Novi (Driver 19.30_19055611344 [C1874070]) with 40 compute units, versus a Vega 64 (Driver 19.4.1) with 40 compute units enabled. Running 3D Mark 11 GT1 (1280 x 720), 3D Mark 11 GT2 (1280 x 720), 3D Mark Firestrike GT1 (2560 x 1440), 3D Mark Firestrike GT2 (2560 x 1440), Unigine Heaven (1080 x 1080), the Novi (with a die size of 251mm^2) achieved an average FPS score of 140, 136, 49, 37, and 84 respectively. Compared to the Vega 56 (with a die size of 486mm^2) which achieved 103, 113, 41, 32, and 72 respectively. RX-358

RX-362
Testing done by AMD performance labs on June 4, 2019. Systems were tested with: Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz (6 core) with 16GB DDR4 @ 2133 MHz using a Asus X99-E Motherboard running Windows 10 Enterprise 64-bit (Ver. 1809, build 17763.053). Using the following graphics cards: Novi 10 (Driver 19.30_19055611344 [C1874070]) with 40 compute units, versus a Vega 64 (Driver 19.4.1) with 40 compute units enabled. Breakdown based on AMD internal data June 4, 2019. Performance may vary. RX-362

RX-365
Testing done by AMD performance labs on June 4, 2019. Systems were tested with: Intel(R) Core(TM) i7-5960X with 16GB DDR4 @ 2133 MHz running Windows 10. Using the following graphics cards: Radeon RX 5700 XT (Driver 19.30) with 40 compute units, versus a Vega 64 (Driver 19.4.1) with 64 compute units enabled. On average across 1080p, 1440p, and 2160p resolutions, the Radeon RX 5700 XT averages 14% higher. RX 365

GPU-38
HEVC (H.265), H.264, and VP9 acceleration are subject to and not operable without inclusion/installation of compatible HEVC players. GD-81

GPU-122
Radeon Freesync technology requires a monitor and AMD Radeon™ graphics, both with Freesync support. See www.amd.com/freesync for complete details. Confirm capability with your system manufacturer before purchase. GD-127

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 warranties or representations 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 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 products are set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale.

ATTRIBUTION
© 2019 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Radeon, Freesync, and combinations thereof are trademarks of Advanced Micro Devices, Inc., in the United States and/or other jurisdictions. HDMI, the HDMI logo and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing, LLC in the United States and other countries. Other names are for informational purposes only and may be trademarks of their respective owners.