AMD’s Next Generation GPU and High Bandwidth Memory Architecture: FURY

Joe Macri, AMD Corporate Fellow and Chief Technology Officer, AMD Products
AGENDA

- Overview of AMD Radeon R9 Fury
- Why HBM and Die Stacking
- The Journey to Fury
- Performance
- Form Factor Innovation
OVERVIEW

- First high-volume interposer
- First Through Silicon Vias (TSVs) and µBumps in the graphics industry
- Most discrete dies in a single package at 22
- Total 1011 sq. mm.

4GB High-Bandwidth Memory
4096-bit wide interface
512 Gb/s Memory Bandwidth

Graphics Core Next Architecture
64 Compute Units¹
4096 Stream Processors
596 sq. mm. Engine
Die stacking facilitates the integration of discrete dies.

8.5 years of technology development at AMD and its partners.
Platforms & devices must balance power usage between DRAM and logic chips

GDDR5 is entering the inefficient region of the power/performance curve

AMD anticipated this challenge seven years ago and began work on a solution
PROTOTYPING ACTIVITIES OVER TIME

First Time Out

CPU + D3 Mech.

dGPU + G3

Component reliability: TC | uHAST | HTS

2007 (100’s of samples) 2011 (<5000 of samples) Jul’14 (>5000 samples) 2014
PROTOTYPING ACTIVITIES OVER TIME

First Time Out

- CPU + D3 Mech.
- dGPU + G3

Primary Learning

- 345mm² ASIC
- 500mm² IP
- dGPU
- ESD | BLRT | Sort
- PwrCyc | uBump EM | TSV EM/SM
- Component reliability: TC | uHAST | HTS

2007 (100’s of samples)

2011 (<5000 of samples)

Jul’14 (>5000 samples)

2014
PROTOTYPING ACTIVITIES OVER TIME

First Time Out

- CPU + D3 Mech.
- dGPU + G3

Primary Learning

- 345mm² ASIC
- 500mm² IP

- dGPU

- Began migration to larger dGPUs
  - 502mm² ASIC w/ 818mm² interposer
  - Larger than reticle interposer
  - CPI of large system
  - Finalize TMP details

Component reliability: TC | uHAST | HTS

ESD | BLRT | Sort

Cost Down

PwrCyc | uBump EM | TSV EM/SM

2007 (100’s of samples) 2011 (<5000 of samples) Jul’14 (>5000 samples) 2014
PROTOTYPING ACTIVITIES OVER TIME

First Time Out
- CPU + D3 Mech.
- dGPU + G3

Primary Learning
- 345mm² ASIC
- 500mm² IP
dGPU
- dGPU + G3

Product Readiness
- 592mm² ASIC
- 1011mm² IP
Fiji Replica

Component reliability: TC | uHAST | HTS

Cost Down
- PwrCyc | uBump EM | TSV EM/SM

2007 (100's of samples)
2011 (<5000 of samples)
2014 (>5000 samples)
2014

2014

2007

2011
A new type of memory chip with low power consumption and an ultra-wide bus width

Many of those chips stacked vertically like floors in a skyscraper

New interconnects, called “through-silicon vias” (TSVs) and “µbumps”, connect one DRAM chip to the next

TSVs and µbumps also used to connect the SoC/GPU to the interposer

AMD and SK Hynix partnered to define and develop the first complete specification and prototype for HBM
**HBM: A DIFFERENT MEMORY FROM GDDR5**

<table>
<thead>
<tr>
<th>GDDR5</th>
<th>Per Package</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>32-bit</td>
<td>Bus Width</td>
<td>1024-bit</td>
</tr>
<tr>
<td>Up to 1750MHz (7GBps)</td>
<td>Clock Speed</td>
<td>Up to 500MHz (1GBps)</td>
</tr>
<tr>
<td>Up to 28GB/s per chip</td>
<td>Bandwidth</td>
<td>&gt;100GB/s per stack</td>
</tr>
<tr>
<td>1.5V</td>
<td>Voltage</td>
<td>1.3V</td>
</tr>
</tbody>
</table>
HBM & interposer give 60% more bandwidth than GDDR5 for 60% less power\(^2\)

HBM rebalances DRAM vs. logic power consumption to protect future GPU performance growth
EFFICIENT DESIGN

110mm

90mm

1.6X shorter

2X shorter

~3X reduction in PCB Footprint with HBM

PCB area occupied by ASIC + Memory (Radeon™ R9 290X)

PCB area occupied by ASIC with HBM
RADEON R9 FURY

- Configured for Gaming
  - 4 Prims/cycle
  - 64 Pix/cycle
  - 64CU
  - 4096 Ops/Cycle
    - 1/16 rate DPFP
  - 256 Texture/cycle
  - 2MB L2
  - Compute wave switch

- HBM
- Delta Color Compression
- Tessellation Redistribution
- SRIOV Virtualization
- 4 core Async Compute
- Dispatch Draw
- TSMC 28nm HPX
- Improved clock gating
- 1.5x Perf/Watt over Hawaii
<table>
<thead>
<tr>
<th></th>
<th>Radeon R9 Fury X</th>
<th>Radeon R9-290X</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPFP Compute</td>
<td>&gt; 8 TFLOPS</td>
<td>5.6 TFLOPS</td>
<td>&gt; 1.4x</td>
</tr>
<tr>
<td>Peak Bandwidth</td>
<td>512 GB/sec</td>
<td>320 GB/sec</td>
<td>1.6x</td>
</tr>
<tr>
<td>Texture fill rate</td>
<td>&gt; 256 Gtxels/sec</td>
<td>176 Gtxels/sec</td>
<td>&gt; 1.45x</td>
</tr>
<tr>
<td>Die area</td>
<td>593 mm²</td>
<td>438 mm²</td>
<td>1.35x</td>
</tr>
<tr>
<td>Peak SP GFLOPS/mm²</td>
<td>&gt; 13.5</td>
<td>12.8</td>
<td>&gt; 1.05x</td>
</tr>
<tr>
<td>L2 Cache Capacity</td>
<td>2 MB</td>
<td>1 MB</td>
<td>2x</td>
</tr>
</tbody>
</table>
AMD Radeon™ R9 Fury X Graphics Card

SMALL SIZE, SIGNIFICANT IMPACT

Incredibly compact graphics cards for small form factor PCs

7.5” PCB

30%
Shorter than the Radeon™ R9 290X (11.5”)

Board shot shown for illustration purposes only. Final board design may differ.
AMD Radeon™ R9 Fury X Graphics Card

COOL & QUIET OPERATION

Closed Loop Liquid Cooling Solution

Typical gaming scenario GPU temperature

~50°C

High quality 120mm fan

Radeon™ R9 Fury X < 32 dBA

GTX Titan X 45 dBA

Measured Acoustics during typical load (dBA - Lower is better)

Board shot shown for illustration purposes only. Final board design may differ.
AMD Radeon™ R9 Fury X Graphics Card

4K ULTRA SETTINGS SMOOTH GAMEPLAY

FAR CRY 4
4K ULTRA SETTINGS

54 fps
AVERAGE FPS

43 fps
MINIMUM FPS

FPS

Time
AMD Radeon™ R9 Nano Graphics Card

FORM FACTOR INNOVATION

175W SINGLE 8-PIN PCIe® CONNECTOR

UP TO 2X* PERFORMANCE DENSITY

UP TO 2X* PERFORMANCE PER WATT

*Compared to Radeon™ R9 290X graphics card

Board shot shown for illustration purposes only. Final board design may differ.
PERFORMANCE RESULTS

Gaming Performance (Frames Per Second)

Performance Per Watt

Fury X  Fury Nano  "Hawaii"
A NEW PC FORM FACTOR

- A unique approach to combine powerful hardware within a small form factor without compromises to thermals or acoustics
- Designed to deliver the best possible VR experiences with AMD LiquidVR™ technology
A NEW PC FORM FACTOR

- Powered by up to two Radeon™ R9 “Fiji” GPUs
- Fully liquid cooled system
FURY

First with HBM

Enables Innovative Form Factors

Up to 2x Performance/Watt increase
Thank You
FOOTNOTES

1. Discrete AMD Radeon™ GPUs and AMD 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 in unison. GRT-5

2. Testing conducted by AMD engineering on the AMD Radeon™ R9 290X GPU vs. an HBM-based device. Data obtained through isolated direct measurement of GDDR5 and HBM power delivery rails at full memory utilization. Power efficiency calculated as GB/s of bandwidth delivered per watt of power consumed. AMD Radeon™ R9 290X (10.66 GB/s bandwidth per watt) and HBM-based device (35+ GB/s bandwidth per watt), AMD FX-8350, Gigabyte GA-990FX-UD5, 8GB DDR3-1866, Windows 8.1 x64 Professional, AMD Catalyst™ 15.20 Beta. HBM-1

3. Testing conducted by AMD engineering on the AMD Radeon™ R9 290X GPU vs. the AMD Radeon™ R9 Fury X GPU. Measured performance and power on Far Cry 4. System Configuration: Core i7-5960X (3001 MHz), AMD Catalyst™ 15.20 Beta.