



AMD Dives Deep On High Bandwidth Memory - What Will HBM Bring AMD?

163
Comments

by [Ryan Smith](#) on May 19, 2015 8:40 AM EST

Posted in [GPU S](#) [AMD](#) [Radeon](#) [HBM](#)

THE NET BENEFITS OF HBM & CLOSING THOUGHTS

The Net Benefits of HBM

Now that we’ve had a chance to talk about how HBM is constructed and the technical hurdles in building it, we can finally get to the subject of the performance and design benefits of HBM. HBM is of course first and foremost about further increasing memory bandwidth, but the combination of stacked DRAM and lower power consumption also opens up some additional possibilities that could not be pursued with GDDR5.

We’ll start with the bandwidth capabilities of HBM. The amount of bandwidth ultimately depends on the number of stacks in use along with the clockspeed of those stacks. HBM uses a DDR signaling interface, and while AMD is not disclosing final product specifications at this time, they have given us enough information to begin to build a complete picture.

GPU Memory Math			
	AMD Radeon R9 290X	NVIDIA GeForce GTX Titan X	Theoretical 4-Stack HBM1
Total Capacity	4GB	12GB	4GB
Bandwidth Per Pin	5Gbps	7Gbps	1Gbps
Number of Chips/Stacks	16	24	4
Bandwidth Per Chip/Stack	20GB/sec	14GB/sec	128GB/sec
Effective Bus Width	512-bit	384-bit	4096-bit
Total Bandwidth	320GB/sec	336GB/sec	512GB/sec
Estimated DRAM Power Consumption	30W	31.5W	14.6W

The first generation of HBM AMD is using allows for each stack to be clocked up to 500MHz, which after DDR signaling leads to 1Gbps per pin. For a 1024-bit stack this means a single stack can deliver up to 128GB/sec (1024b * 1G / 8b) of memory bandwidth. HBM in turn allows from 2 to 8 stacks to be used, with each stack carrying 1GB of DRAM. AMD’s example diagrams so far (along with NVIDIA’s Pascal test vehicle) have all been drawn with 4 stacks, in which case we’d be looking at 512GB/sec of memory bandwidth. This of course is quite a bit more than the 320GB/sec of memory bandwidth for the R9 290X or 336GB/sec for NVIDIA’s GTX titan X, working out to a 52-60% increase in memory bandwidth.

Sponsored Links

ヒーローウォーズの世界へ [プレイ開始]

Hero Wars

Advertisement

PIPELINE STORIES + SUBMIT NEWS

Sabrent Rocket nano V2 External SSD Review: Phison U18 in a Solid Offering

MediaTek to Add NVIDIA G-Sync Support to Monitor Scalars, Make G-Sync Displays More Accessible

Qualcomm Adds Snapdragon 7s Gen 3: Mid-Tier Snapdragon Gets Cortex-A720 Treatment

CXL Gathers Momentum at FMS 2024

Fadu's FC516l SSD Controller Breaks Cover in Western Digital's PCIe Gen5 Enterprise Drives

PCI-SIG Demonstrates PCIe 6.0 Interoperability at FMS 2024

DapuStor and Memblaze Target Global Expansion with State-of-the-Art Enterprise SSDs

Phison Enterprise SSDs at FMS 2024: Pascari Branding and Accelerating AI

Intel Sells Its Arm Shares, Reduces Stakes in Other Companies

G.Skill Intros Low Latency DDR5 Memory Modules: CL30 at 6400 MT/s

Samsung's 128 TB-Class BM1743 Enterprise SSD Displayed at FMS 2024

Kioxia Demonstrates Optical Interface

Create effective marketing campaigns in minutes with Gen AI features

S Click

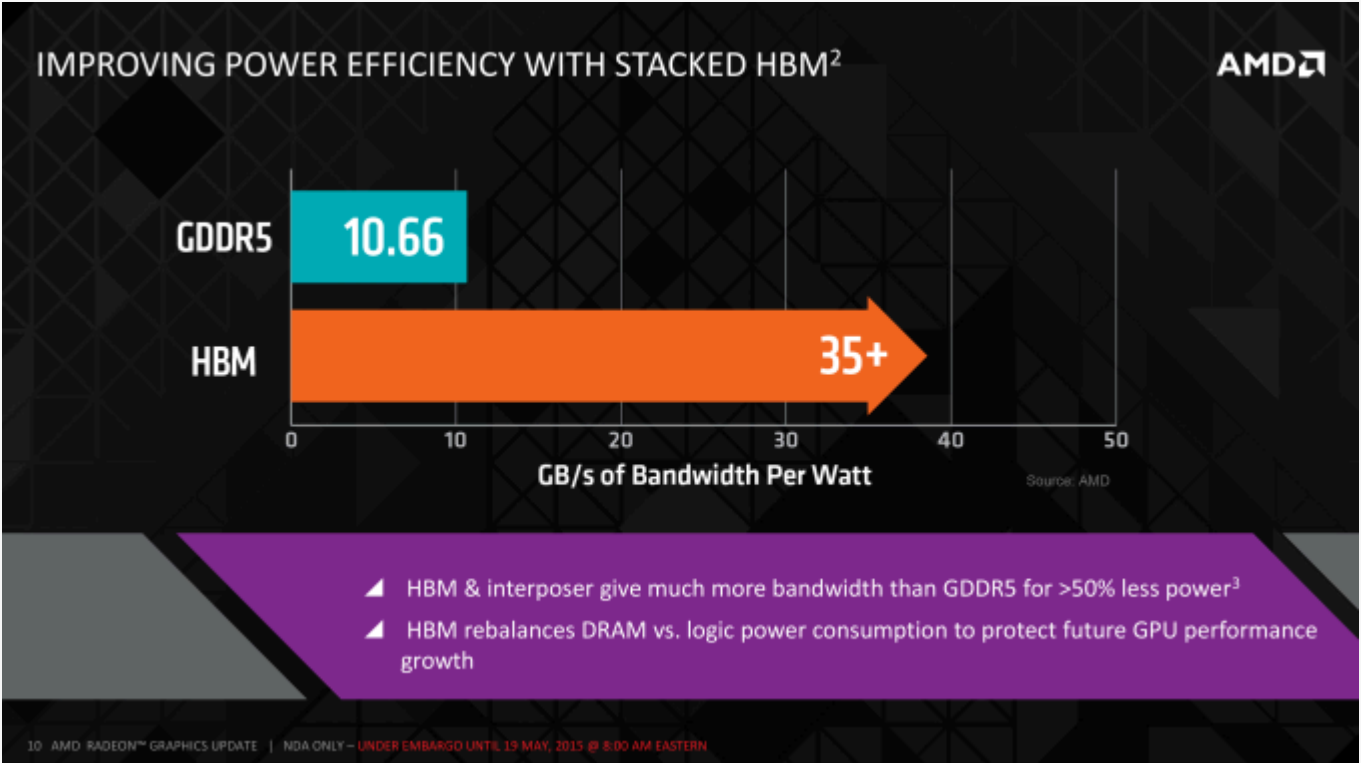
At the same time this also calls into question memory capacity – 4 1GB stacks is only 4GB of VRAM – though AMD seems to be saving that matter for the final product introduction later this quarter. Launching a new, high-end GPU with 4GB could be a big problem for AMD, but we'll see just what they have up their sleeves in due time.

©Getty Images

ANANDTECH

2. Integrated graphics or not?

If you don't plan to play games or you're on a tight budget, consider a CPU with integrated graphics.



What's perhaps more interesting is what happens to DRAM energy consumption with HBM. As we mentioned before, R9 290X spends 15-20% of its 250W power budget on DRAM, or roughly 38-50W of power on an absolute basis. Meanwhile by AMD's own reckoning, GDDR5 is good for 10.66GB/sec of **bandwidth per watt** of power, which works out to 30W+ via that calculation. HBM on the other hand delivers better than 35GB/sec of **bandwidth per watt**, an immediate 3x gain in energy efficiency per watt.

Of course AMD is then investing some of those gains back in to coming up with more memory bandwidth, so it's not as simple as saying that memory power consumption has been cut by 70%. Rather given our earlier bandwidth estimate of 512GB/sec of memory bandwidth for a 4 stack configuration, we would be looking at about 15W of power consumption for a 512GB/sec HBM solution, versus 30W+ for a 320GB/sec GDDR5 solution. The end result then points to DRAM power consumption being closer to halved, with AMD saving 15-20W of power.

What's the real-world advantage of a 15-20W reduction in DRAM power consumption? Besides being able to invest that in reducing overall video card power consumption, the other option is to invest it in increasing clockspeeds. With PowerTune putting a hard limit on power consumption, a larger GPU power budget would allow AMD to increase clockspeeds and/or run at the maximum GPU clockspeed more often, improving performance by a currently indeterminable amount. Now as fair warning here, higher GPU clockspeeds typically require higher voltages, which in turn leads to a rapid increase in GPU power consumption. So although having additional power headroom does help the GPU, it may not be good for quite as much of a clockspeed increase as one might hope.

Meanwhile the performance increase from the additional memory bandwidth is equally nebulous until AMD's new product is announced and benchmarked. As a rule of thumb GPUs are virtually always memory bandwidth bottlenecked – they are after all high-throughput processors capable of trillions of calculations per second working with only hundreds of billions of bytes of bandwidth – so there is no doubt that the higher memory bandwidths of HBM will improve performance. However memory bandwidth increases currently don't lead to 1:1 performance increases even on AMD's current cards, and it's unlikely to be any different on future products.

Throwing an extra wrinkle into matters, any new AMD product would be based on GCN 1.2 or newer, which

Advertisement

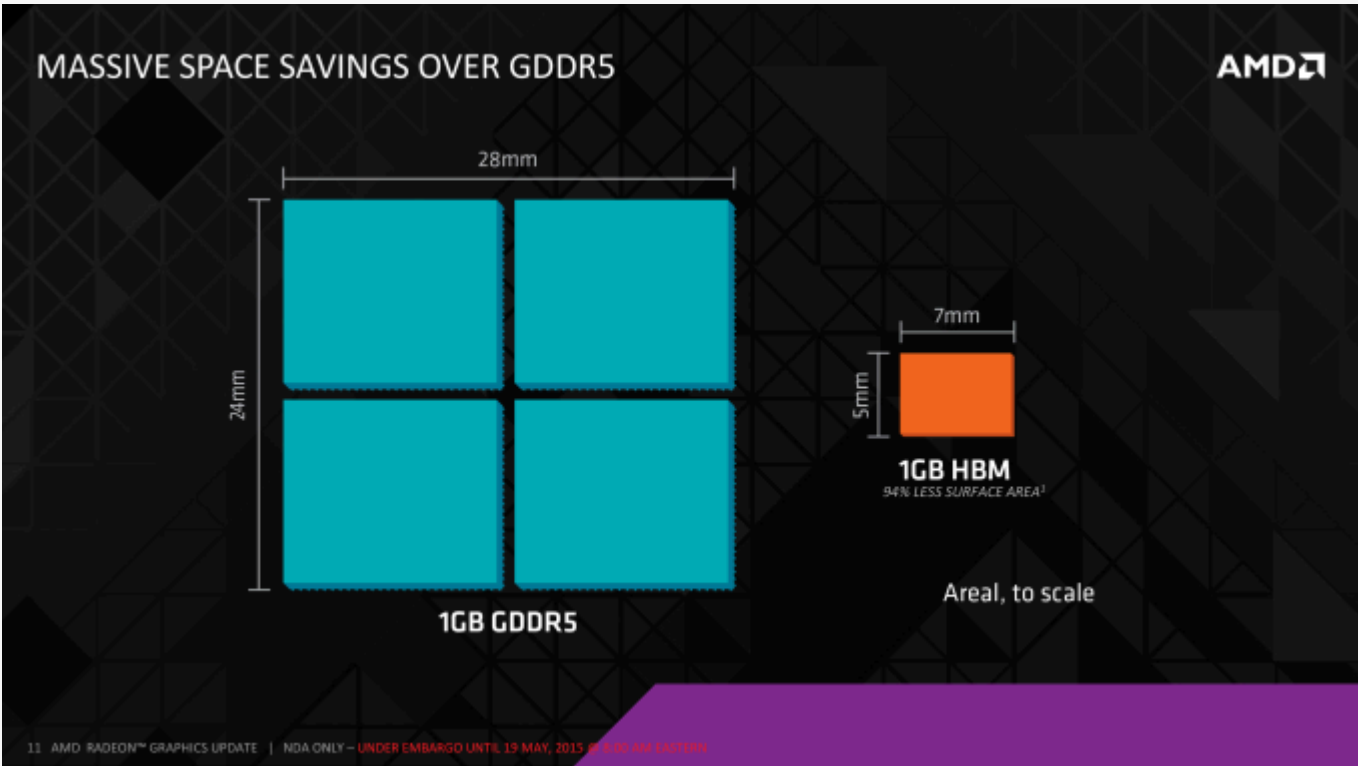
Advertisement

Advertisement

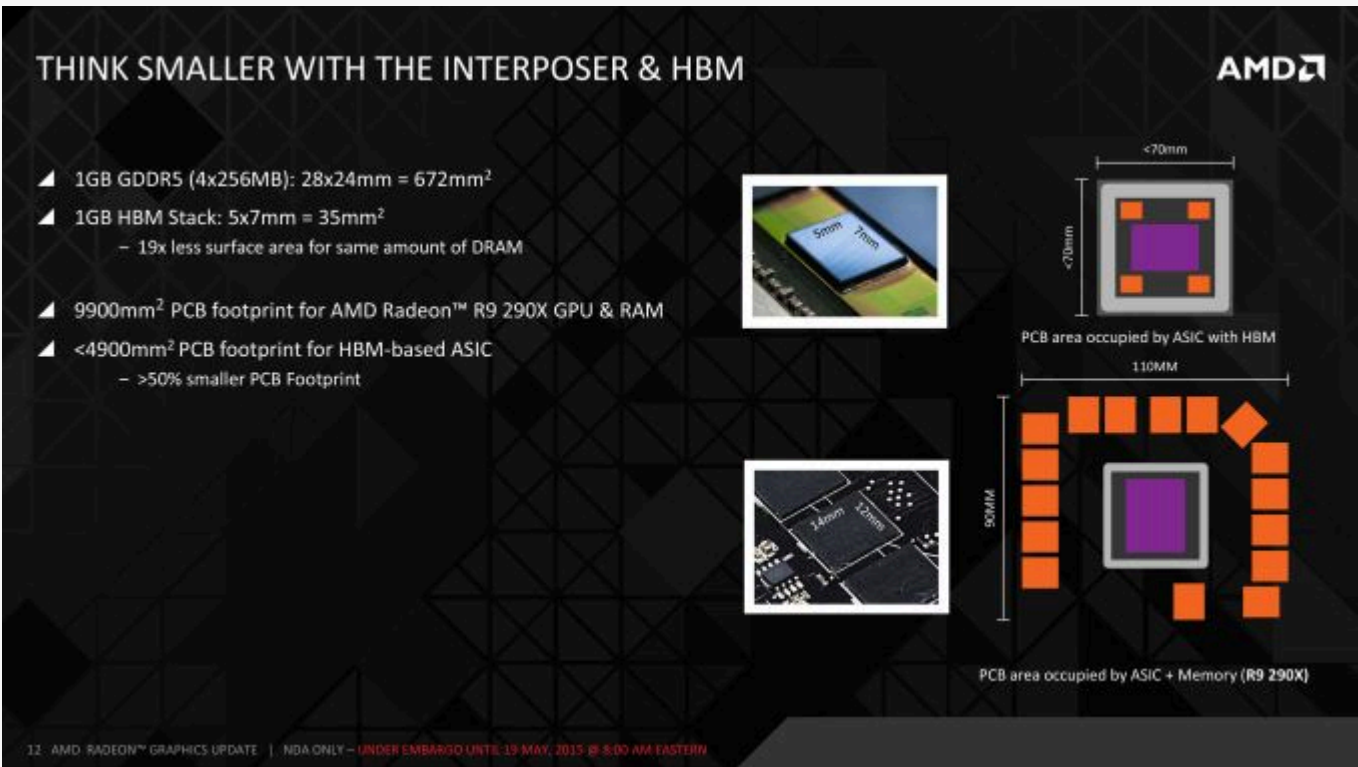


very handy for high resolutions – but it also makes it impossible to predict what the final performance impact might be. Still, it will be interesting to see what AMD can do with a 2x+ increase in effective memory bandwidth for graphics workloads.

The final major benefit AMD is looking at taking advantage of with HBM – and that this point they’re not even being subtle about – is new form factor designs from the denser designs enabled by HBM. With the large GDDR5 memory chips replaced with much narrower HBM stacks, AMD is telling us that the resulting ASIC + RAM setups can be much smaller.



How much smaller? Well 1GB of GDDR5, composed of 2Gbit modules (the standard module size for R9 290X) would take up 672mm², versus just 35mm² for the same 1GB of DRAM as an HBM stack. Even if we refactor this calculation for 4Gbit modules – the largest modules used in currently shipping video cards – then we still end up with 336mm² versus 35mm², which is still a savings of 89% for 1GB of DRAM. Ultimately the HBM stack itself is composed of multiple DRAM dies, so there’s still quite a bit of silicon in play, however its 2D footprint is reduced significantly thanks to stacking.



By AMD’s own estimate, a single HBM-equipped GPU package would be less than 70mm X 70mm (4900mm²), versus 110mm X 90mm (9900mm²) for R9 290X. Throw in additional space savings from the fact that HBM stacks don’t require quite as complex power delivery circuitry, and the card space savings could be significant. By our reckoning the total card size will still be fairly big – all of those VRMs and connectors need to go somewhere – but there is potential for significant savings. What AMD intends to do with those savings remains to be seen, but with apologies to AMD on this one, NVIDIA has already shown off their Pascal test vehicle for their mezzanine connector design, and it goes without saying that such a form factor opens up some very interesting possibilities.



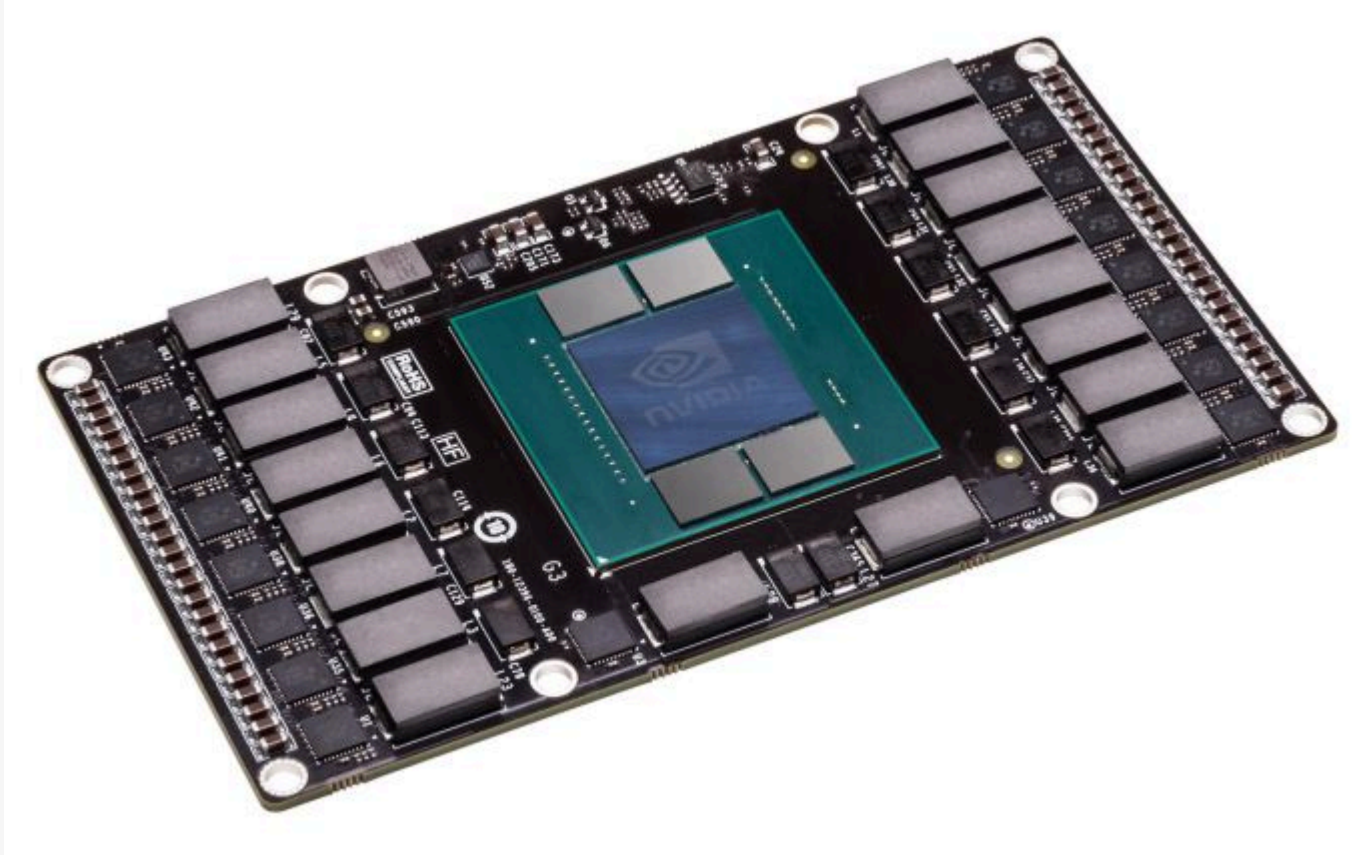
Advertisement

AT&T

That's up to
\$1,700
in savings

Learn more

Offer details >



With apologies to AMD: NVIDIA's Pascal Test Vehicle, An Example Of A Smaller, Non-Traditional Video Card Design

Finally, aftermarket enthusiasts may or may not enjoy one final benefit from the use of HBM. Because the DRAM and GPU are now on the same package, AMD is going to be capping the package with an integrated heat spreader (IHS) to compensate for any differences in height between the HBM stacks and GPU die, to protect the HBM stacks, and to supply the HBM stacks with sufficient cooling. High-end GPU dies have been bare for some time now, so an IHS brings with it the same kind of protection for the die that IHSs brought to CPUs. At the same time however this means it's no longer possible to make direct contact with the GPU, so extreme overclockers may come away disappointed. We'll have to see what the shipping products are like and whether in those cases it's viable to remove the IHS.

Closing Thoughts

Bringing this deep dive to a close, as the first GPU manufacturer to be shipping an HBM solution – in fact AMD expects to be the only vendor to ship an HBM1 solution – AMD has set into motion some very aggressive product goals thanks to the gains from HBM. Until we know more about AMD's forthcoming video card I find it prudent to keep expectations in check here, as HBM is just one piece of the complete puzzle that is a GPU. But at the same time let's be clear here: HBM is the future memory technology of GPUs, there is potential for significant performance increases thanks to the massive increase in memory bandwidth offers, and for roughly the next year AMD is going to be the only GPU vendor offering this technology.

AMD for their part is looking to take as much of an advantage of their lead as they can, both at the technical level and the consumer level. At the technical level AMD has said very little about performance so far, so we'll have to wait and see just what their new product brings. But AMD is being far more open about their plans to exploit the size advantage of HBM, so we should expect to see some non-traditional designs for high-end GPUs. Meanwhile at the consumer level, expect to see HBM enter the technology lexicon as the latest buzzword for high-performance products – almost certainly to be stamped on video card boxes today just as GDDR5 has been for years – as AMD looks to let everyone know about their advantage.

HBM WITH INTERPOSER: SPEED, POWER & SMALL FORM FACTORS

A REVOLUTION IN CHIP DESIGN

HIGH BANDWIDTH
Performance well beyond DDR4/GDDR5/LPDDR4

POWER EFFICIENCY
>3X the performance per watt of GDDR5²

SMALL FORM FACTORS
94% less PCB surface area than GDDR5¹

INNOVATION
New interconnects, interposer & DRAM type designed by AMD

AMD

13 AMD RADEON™ GRAPHICS UPDATE | NDA ONLY – UNDER EMBARGO UNTIL 19 MAY, 2015 @ 5:00 AM EASTERN

Meanwhile shifting gears towards the long term, high-end GPUs are just the first of what AMD expects to be a wider rollout for HBM. Though AMD is not committing to any other products at this time, as production ramps up and costs come down, HBM is expected to become financially viable in a wider range, including lower-end GPUs, HPC products (e.g. FirePro S and AMD's forthcoming HPC APU), high-end communications gear, and

Advertisement



PRINT THIS ARTICLE

You May Like

Sponsored Links by Taboola

ヒーローウォーズがTokyoに登場 [今すぐ]

Hero Wars

すべてを変える無料ゲーム

レイドシャドウレジェンド

誰もがロボットに夢中になるゲーム

メックアリーナ

She Was Everyone's Dream Girl In 90's, This Is Her Recently.

Investructor

JSOLは、未来志向のアプローチで迅速に対応し、お客さまのビジネスを確実に成功へ導きます。

株式会社JSOL

If you have a mouse, this game will keep you awake all night long.

CombatSiege

GW・夏休み対象、モルディブお子様宿泊無料

クラブメッド

If You Have A Computer, This Adventure Game Is A Must-Play.

Sunrise Village

ものづくり補助金、IT導入補助金他

ダッソー・システムズ

40歳以上なら、このゲームは必ずプレイしてください！

Taonga Farm

27 Photos Of Country Girls Leave World In Awe

True Edition

コスモエネルギーHD

コスモエネルギーHD

Comments Locked

163 Comments

View All Comments

HighTech4US - Tuesday, May 19, 2015 - link

So Fiji is really limited to 4 GB VRAM.

chizow - Tuesday, May 19, 2015 - link



last year when Tonga launched that it would be the launch vehicle for HBM? I guess it does support HBM, it just wasn't ready yet. Would also make sense as we have yet to see a fully-enabled Tonga ASIC; even though the Apple M295X has the full complement of 2048 SP, it doesn't have all memory controllers.

Kevin G - Tuesday, May 19, 2015 - link

The 1024 bit wide bus of an HBM stack is composed of eight 128 bit wide channels. Perhaps only half of the channels need to be populated allowing for twice the number of stacks to reach 8 GB without changing the Fiji chip itself?

akamateau - Thursday, May 28, 2015 - link

Electrical path latency is cut to ZERO. EP Latency is how many clocks are used moving the data over the length of the electrical path. That latecy is about a one clock.

testbug00 - Tuesday, May 19, 2015 - link

M295X isn't only in Apple... I think Alienware has one to! XD

Yeah. Interesting, even Charlie points that out a lot. He also claims that developers laugh at needing over 4GB, which, may be true in some games... GTA V and also quite a few (very poor) games show otherwise.

Of course, how much you need to have ~60+ FPS, I don't know. I believe at 1440/1600p, GTA V at max doesn't get over 4GB. Dunno how lowering settings changes the VRAM there. So, to hit 60FPS at a higher res might require turning down the settings of CURRENT GAMES (not future games, those are the problem!) probably would fit *MOST* of them inside 4GB. I still highly doubt that GTA V and some others would fit, however. *grumble grumble*

Hope AMD is pulling wool over everyone's eyes, however, their presentation does indeed seem to limit it to 4GB.

xthetenth - Tuesday, May 19, 2015 - link

GTA V taking over 4 GB if available and GTA V needing over 4 GB are two very different things. If it needed that memory then 980 SLI and 290X CF/the 295X would choke and die. They don't.

hansmuff - Tuesday, May 19, 2015 - link

The don't choke and die, but they also can't deliver 4K at max detail and that is *in part* because of 4GB memory. http://www.hardocp.com/article/2015/05/04/grand_th...

The 3.5GB 970 chokes early on 4K and needs feature reduction, the 980 allows more features, the Titan yet more features, in large part due to memory config.

Yeah it will be interesting how compression or new AA approaches lower memory usage but I will not buy a 4GB high end card now or in the future and depend on even more driver trickery to lower memory usage for demanding titles.

hansmuff - Tuesday, May 19, 2015 - link

To substantiate my comment about driver trickery, this is a quote from TechReport's HBM article:

"When I asked Macri about this issue, he expressed confidence in AMD's ability to work around this capacity constraint. In fact, he said that current GPUs aren't terribly efficient with their memory capacity simply because GDDR5's architecture required ever-larger memory capacities in order to extract more bandwidth. As a result, AMD "never bothered to put a single engineer on using frame buffer memory better," because memory capacities kept growing. Essentially, that capacity was free, while engineers were not. Macri classified the utilization of memory capacity in current Radeon operation as "exceedingly poor" and said the "amount of data that gets touched sitting in there is embarrassing."

Strong words, indeed.

With HBM, he said, "we threw a couple of engineers at that problem," which will be addressed solely via the operating system and Radeon driver software. "We're not asking anybody to change their games.""

I don't trust them to deliver that on time and consistently.

chizow - Tuesday, May 19, 2015 - link

lol yeah, hopefully they didn't just throw the same couple of engineers who threw together the original FreeSync demos together on that laptop, or the ones who are tasked with fixing the FreeSync ghosting/overdrive issues, or the FreeSync CrossFire issues, or the Project Cars/TW3 driver updates. You get the point hehe, those couple engineers are probably pretty busy, I am sure they are thrilled to have one more promise added to their plates. :)

dew111 - Tuesday, May 19, 2015 - link

Making something that is inefficient more efficient isn't "trickery," it's good engineering. And when the product comes out, we will be able to test it, so your trust is not required.



LINKS

- Home
- About
- Forums
- RSS
- Pipeline News
- Bench
- Terms of Use
- Contact Us
- Accessibility Statement

TOPICS

- CPUs
- Motherboards
- SSD/HDD
- GPUs
- Mobile
- Enterprise & IT
- Smartphones
- Memory
- Cases/Cooling/PSU(s)

FOLLOW

-  Facebook
-  Twitter
-  RSS

The Most Trusted in Tech Since 1997

[About](#) [Advertising](#) [Privacy Policy](#)



[Visit Our Corporate Site](#)

COPYRIGHT © 2024. ALL RIGHTS RESERVED.