ECE 571 – Advanced Microprocessor-Based Design Lecture 36

Vince Weaver

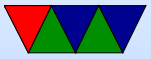
http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

4 December 2020

Announcements

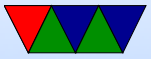
- Don't forget projects, presentations next week (Wed and Fri)
- Final writeup due last day of exams (18th)
- Will try to get homeworks graded soon.



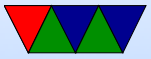
NVIDIA GPUs

Tesla	2006	90-40nm
Fermi	2010	40nm/28nm
Kepler	2012	28nm
Maxwell	2014	28nm
Pascal/Volta	2016	16nm/14nm
Turing	2018	12nm
Ampere	2020	8nm/7nm
Hopper	20??	??

- GeForce – Gaming



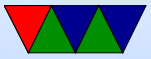
- Quadro – Workstation
- DataCenter



Also Read

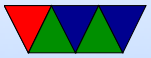
NVIDIA AMPERE GPU ARCHITECTURE blog post

`https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/`



A100 Whitepaper

- A100
- Price? From Dell: $15,434.81 (Free shipping)
- Ethernet and Infiniband (Mellanox) support?
- Asynchronous Copy
- HBM, ECC single-error correcting double-error detection (SECDED)



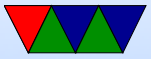
Homework Reading #1

NVIDIA Announces the GeForce RTX 30 Series: Ampere For Gaming, Starting With RTX 3080 & RTX 3090

https://www.anandtech.com/show/16057/

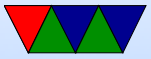
nvidia-announces-the-geforce-rtx-30-series-ampere-for-gaming-starting-with-rtx-3080-rtx-3090

September 2020 – by Ryan Smith



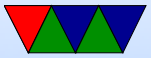
Background

- Ampere Architecture
- CUDA compute 8.0
- TSMC 7nm FINFET (A100)
- Samspun 8n, Geforce30

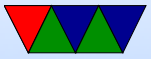


GeForce RTX 30

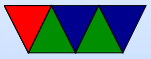
- Samsung 8nm process
- Gaming performance
- Comparison to RTX 20 (Turing based)
- RTX 3090
  - 10496 cores
  - 1.7GHz boost clock
  - 19.5 Gbps GDDR6X, 384 bit, 24GB
  - Single precision 35.7 TFLOP/s
  - Tensor (16-bit) 143 TFLOP/s



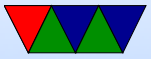
  - Ray perf 69 TFLOPs
  - 350W
  - 8nm Samsung (smallest non-EUV)
  - 28 billion transistors
  - $1500
- GA100 compute(?) TODO
- Third generation tensor cores
- Ray-tracing cores
- A lot of FP32 (shader?) cores
- PCIe 4.0 support (first bump 8 years, 32GB/s)
- SLI support



- What is DirectStorage API? GPU can read disk directly? Why might that be useful?
- 1.9x power efficiency? Is that possible? Might be comparing downclocked Ampere to Turing rather than vice-versa
- GDDR6X
  - NVidia and Micron?
  - multi-level singnaling?
  - can send 2 bits per clock (PAM4 vs NRZ). More complex but you can clock it slower. More power efficient. lower density though (8GB chips)

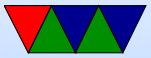


- HDMI2.1 support. More bandwidth, 48Gbps, 2.6x as much HDMI2.0. 4k at 165Hz or 8k. Variable refresh rate? Faster than displayport.
- No virtualink port (for VR headsets, didn't take off)
- NVDEC video coder, support for new AV1 codec. Good as HVEC but royalty free



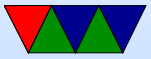
Tom's Hardware

- https://www.tomshardware.com/features/nvidia-am
- 7nm/8nm
- GA100 = 54 billion transistors, 826mm2 die area
- A100 is 400W
- 3090 requires a 12-pin power connector (perpendicular?) back-compat with old 8-pin connector. Also PCB shorter to fit better cooling
- A100 with 128SMs and six HBM2 stacks of 8GB each
- Card has CUDA cores that can do FP32+INT and some

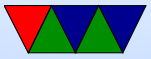


only FP32

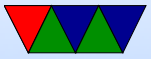
- NVIDIA say 1/3 workload is INT so target that ratio
- Improvements in memory compression?
- GA100 Tensor cores
  - 8x4x8 FMA matrix operation per clock
  - Sparsity (values can be ignored)
  - INT8, INT4 and binary operations
  - Support FP64 but much slower
  - BFLOAT16 (8-bit exponent 7-bit mantissa) (same as F32 less precision)
  - TF32 (tensor float 32 8-bit exponent 10-bit mantissa)



- Multi-instance GPU (MIG), partition GPU up into 7 sub-GPUs
- A100 L1-cache 192kB, L2 cache 40MB
- NVLINK – fast connect between cpu+cpu or cpu+gpu
- NVJPG – accelerated JPEG decoder
- NVDEC (video decoder)
- EDR – error detection and replay, can re-run things if error detected (maybe because of overclocking)
- Ray-tracing
- DLSS? deep-learning super-sampling? Uses AI to sharpen images?

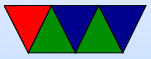


- Data compression
- Direct storage
- ROP raster operations
- Enough 7nm wafers?



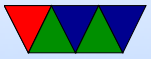
A100 80GB

- `https://www.anandtech.com/show/16250/nvidia-ann`
- SC'20, 80GB HBM2e memory
- DGX system with AMD processor in it



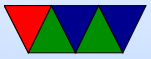
How does Real Hardware

- Biggest impediment was memory size/bandwidth
- 3D Polygon Rendering/Rasterizing – vertices together make polygons, all shapes broken up into smaller amount of polygons. Textures applied, lighting calculated based on normal to polygon.
- Scanline Rendering – no framebuffer, triangles processed and what line they start/stop are stored. One line drawn at a time
- Tiled rendering, similar, but tiles rather than lines



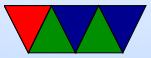
as small as 16x16 or 32x32 (used on various boards, including Videocore IV on Pi)

- 3D Polygon rendering has problems. Lots of things, such as shadows, relections, and transparency, are effectively faked with hacks.



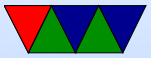
Ray Casting vs Ray Tracing

- Complex, not really good definitions
- Ray casting in general is when you cast rays from eye into scene, but stop when hit first object.
- Ray tracing cast rays from eye into scene, but reflect/refract off of objects until hit light source



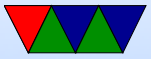
Ray Tracing Hardware

- Anything other than simple ray-casting requires recursion (Each time you hit an object) as well as random-access to the entire 3d-space
- NVIDIA RTX
- Hybrid raytracing – traditional casting+rasterization used for visibility, raytracing for shadows
- Only continue tracing rays of surface has more than threshold of reflectability



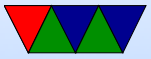
Turing Streaming Multiprocessor (SM)

- Each SM has 64 INT32 and 64 FP32 cores, partitioned into 4 blocks
- Integer units separated out instead of being part of FPU
  Can run int and fp in parallel
- "Unified Cache Architecture"
  Texture, shared memory, load caching in one unit (2x as fast?)
- Variable Rate Shading – different parts of screen with different details



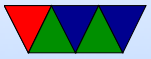
Tensor Cores

- Deep Learning accelerator
- Matrix calculations
- INT8 and INT4 modes



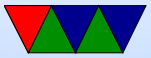
Shading Advancements

- Mesh shading
- Variable Rate Shading (VRS) – spend less time shading if distant, or fast moving, if you're not looking at it (Foveated Rendering, VR)
- Texture-space shading
- Multi-view Rendering (MVR) (two scenes slightly offset, for 3d)



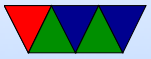
NGX – Neural Graphics Acceleration

- Deep learning graphics framework
- Use AI to speed up rendering
  - Deep Learning Super-Sampling (DLSS)
  - Inpainting (fix missing pixels)
  - AI Slow-mo
  - AI Super res/zoom



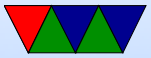
Hybrid Rendering and Neural Networking

- Ray-tracing processor
  - 10-billion giga-rays a second
  - 25x performance over Pascal
- Enhanced tensor cores
  - Use AI to denoise to improve ray-trace performance
- More precisions
  - Volta had FP16
  - Now INT8 and INT4



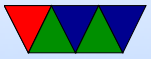
GDDR6 Support

- 16Gbps per pin
  - 2x GDDR5
  - 40% more than GDDR5X
- Low voltage 1.35V
- Higher bandwidth per pin than HBM2 but HBM much wider.
- Samsung, clamshell mode?



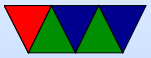
Caches, Memory

- Larger L2 cache (6MB)
- Render Output Unit (ROP)
- Memory compression



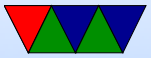
NVLink, VirtualLink, 8K HVEC

- NVLink
- VirtualLink
  - VR – can drive vrtual reality display
  - USB-C alternative?
  - 15W+ power
  - 10Gbps USB3.1
  - 4-lanes of display port
- HVEC – hardware video encoding – 8k support, 25% lower bitrate

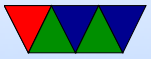


Performance

- RTX 8000 – $6000 – limit 5 per customer
  - High end with 4608 CUDA cores, 576 tensor cores
  - 48GB ECC
  - 295W total board power
  - PCI express 3.0x16
  - DP 1.4 (4), VirtualLink (1)
  - HB Bridge to connect two cards
- 500 Trillian tensor ops/second
- NVidia quoting specs on INT4 (4x faster than FP16)

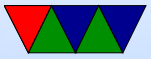


- CUDA cores, 16 TFLOPS
- $754mm^2$



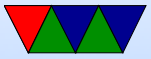
RDNA Whitepaper – Eras

- R100 – pre 2000 – Fixed Function
- R200-R500 – 2001-2007 – Simple VS/PS Shaders
- R600 – 2008-2011 – Terascale, Unified Shaders with VLIW
- Southern Islands – 2012-2018 – GCN (Graphics Core Next)
- RDNA – Navi



AMD GPUs

Caribbean Islands			Fiji
Sea Islands			
Volcanic Islands			
Polaris RX400	2016	28/14nm	
Polaris / RX500	2017	14/12nm	
Vega	2017	14/7nm	GCN5
Navi / RX5000	2019	7nm	RDNA
Navi2x / RX6000	2020	7nm	RDNA2

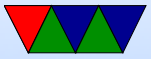


HW Reading #2

AMD Announces Radeon RX 5700 XT & RX 5700: The Next Gen of AMD Video Cards Starts on July 7th At $449/$379

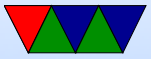
by Ryan Smith, 10 June 2019

https://www.anandtech.com/show/14528/amd-announces-radeon-rx-5700-xt-rx-5700-series

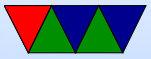


RDNA Whitepaper

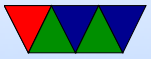
- Backwards compat with GCN
- seven basic instruction types:
  scalar compute, scalar memory, vector compute, vector memory, branches, export, and messages
- GCN had 64-wide wavefront SIMD
- Navi down to 32-wide
- Infinity Fabric
- PCIe4
- Virtualizable, can share GPU between operating systems



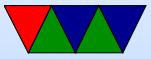
- Asynchronous Compute Tunneling – compute and graphics workloads need to be co-scheduled. Which has priority? This allows some compute tasks to be suspended if higher priority comes in
- Sub-vector mode, split calculations in two, take same number of cycles but can free registers faster
- Can issue instructions every cycle, instead of round-robin (4x faster)
- Accelerated mixed-precision dot product (for machine learning)
- Separate execution pipeline for double-precision data



- Transcendental units
- Crossbars and swizzles
- L0/L1/L2 caches
- Vector caches, interact with texture unit
- Local data share vs global data share, exports
- At 7nm, wires are really long and transmitting data takes time
- Output, support, compression VSC
- Video decoding
- Audio decoding?
- Contrast Adaptice Sharpening

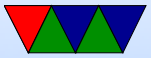


- Radeon RX 5700 XT: FP32 9.75TFLOP/s, 251mm^2, 225W

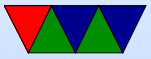


New RDNA Architecture

- Excrutiating detail at a later time? Not yet?
- Navi codename
- Massive redesign
- RX 5700 XT – $400
  - 2560 processors (40CPUs)
  - 1.9GHz boost clock
  - 8GB 14Gbps GDDR6
  - 7nm
  - 225W

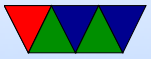


- “Game Clock” – how fast will run during average game
- 9TFLOPs 32-bit FP operation
- Data color compression
- New Cache Hierarchy



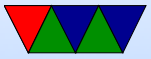
Architecture and Features

- Drop size of wavefront from 64 to 32 threads wide
- SIMD up from 16-slots to 32-slots
- Primitive Shader
  break difference between vertex and geometry shaders?
- PCIe 4.0 (works well with Zen2)
- Display port display stream compression



Intel X^e

- New ISA, Tiger Lake, Q1 2021
- Discrete GPU for first time in years?
- 3D die stacking
- Already available in China? Compete with Geforce MX
- 4GB LPDDR4X, 2.46 FP32 TFLOPS, 25W
- Can split work with integreated GPU



Embedded System GPUs

- VideocodeIV in Pi
- NVIDIA Jetson
- Others