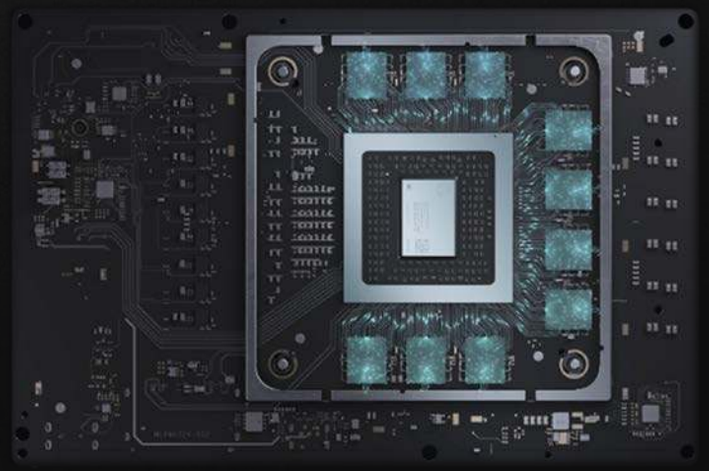
Xbox Series X System Architecture

Jeff Andrews
Mark Grossman
Azure Silicon Architecture Team
8/17/20

Xbox Series X | Next-Gen Console SOC HW Innovation

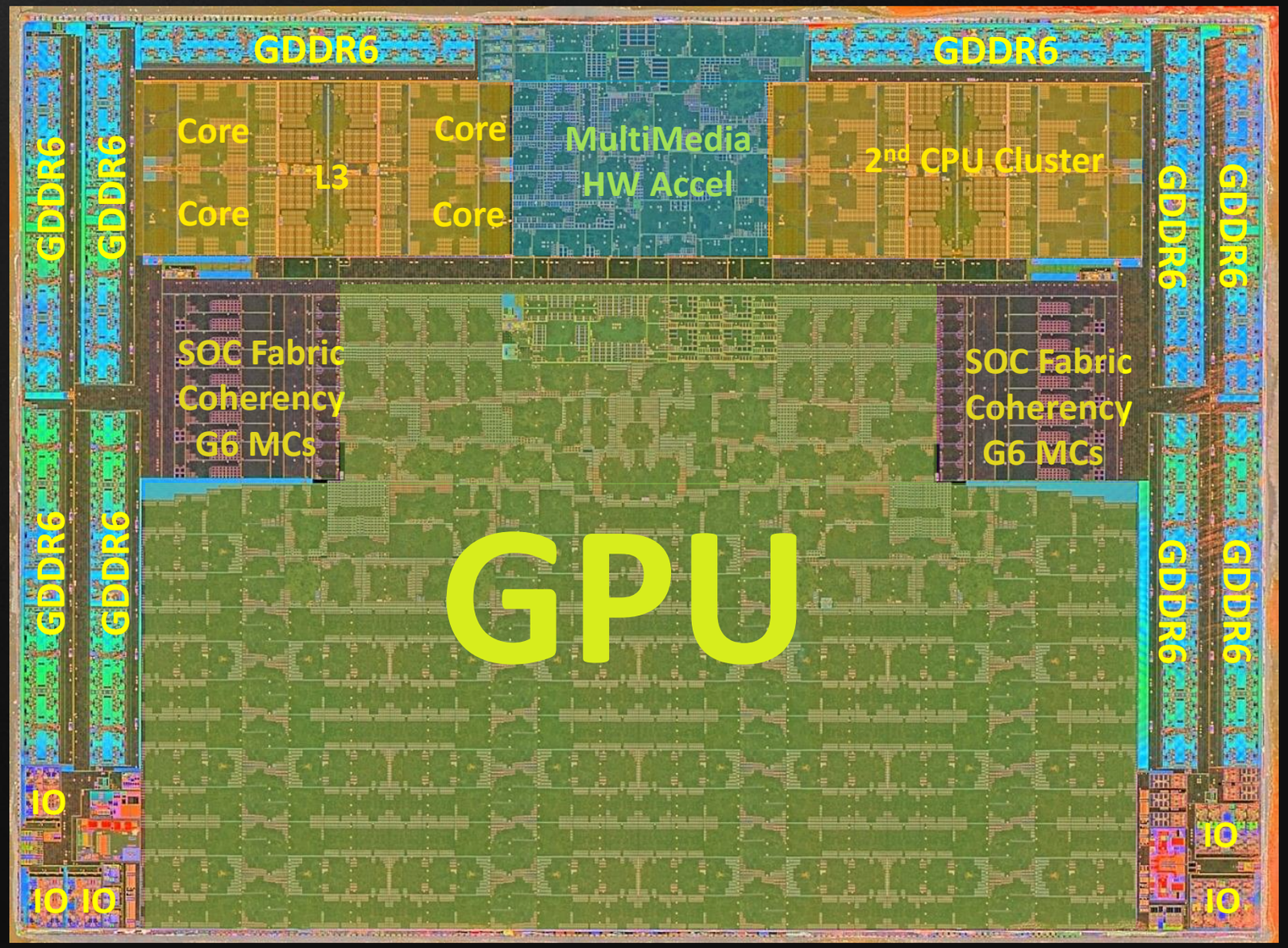
- 3.8Ghz Zen2 Server Class CPU Cores
- Sampler Feedback Streaming (SFS)
- DirectX Raytracing (DXR)
- Variable Rate Shading (VRS)
- Machine Learning Acceleration
- D3D Mesh Shading
- GDDR6 14Gbps, 320b, => 560GB/s



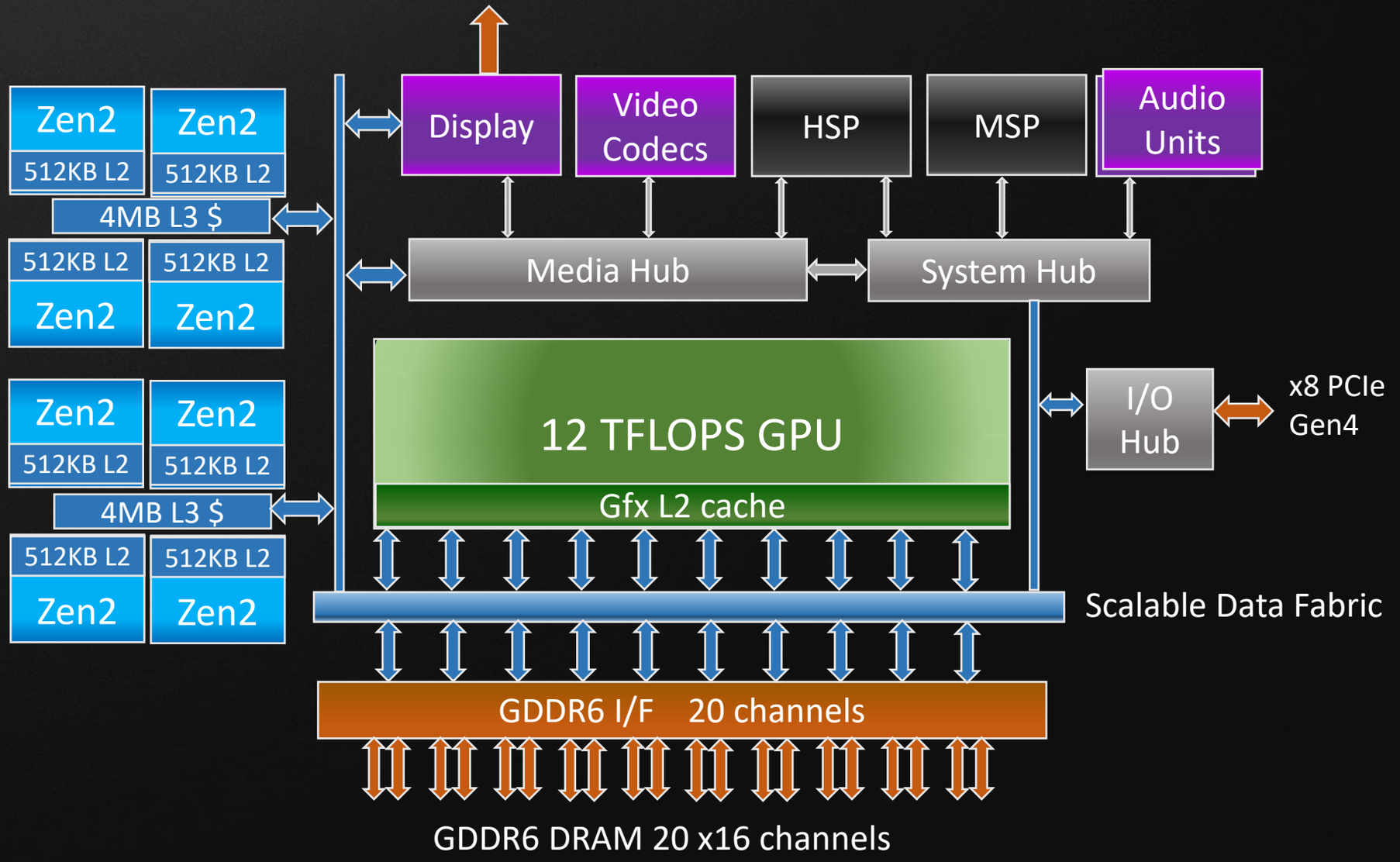
- Xbox Velocity Architecture
  - MSP Crypto/Decomp @ NVMe SSD BW
- Opus Audio Decode
- High quality Sample Rate Converter
- Project Acoustics acceleration
  - Convolution / FFT FP Audio Engine
- 8K Capable
- HDMI 2.1, 10Gbps FRL, ALLM, DSC
- Variable Refresh Rate (VRR)
- 120 Hz Support
- Linear Light Display Processing
- HSP/Pluton RoT, SHACK

Xbox Series X | SOC Specs – Physical View

- TSMC N7 Enhanced
- 360.4 mm^2
- 15.3 billion transistors
- 12-layer substrate (5-2-5)
- 52.5mm x 52.5mm BGA
- 0.80mm min pitch
- 2963 ball BGA
- Developed with AMD



Xbox Series X | SOC Block Diagram



Xbox Series X | SOC Specs – GPU, DRAM, CPU

- GPU:
  - 1.825 GHz, 52 CUs, 12 TFLOPS FP32, 3328 streaming processors
- DRAM:
  - 16 GB GDDR6, 10GB high memory interleave + 6GB low memory interleave
  - 20 channels of x16 GDDR6 @ 14Gbps -> 560 GB/s
  - 320 total DRAM banks
  - DRAM security: full BW crypto, integrity check regions
- CPU:
  - 8x Zen2 CPU cores @ 3.8 GHz, 3.6 GHz w/SMT
  - 32KB L1 I$, 32KB L1 D$, 512KB L2 per CPU core
  - 2x clusters, each cluster: 4C/8T, 4MB L3 -> L2+L3 LLC eff size ~12MB
  - 2x SIMD FP pipes/core: 2 MUL and 2 ADD AVX256 per clk -> 32x SPFP ops/clk
  - Xbox SPLEAP: HW to prevent Escalation Of Privilege attacks

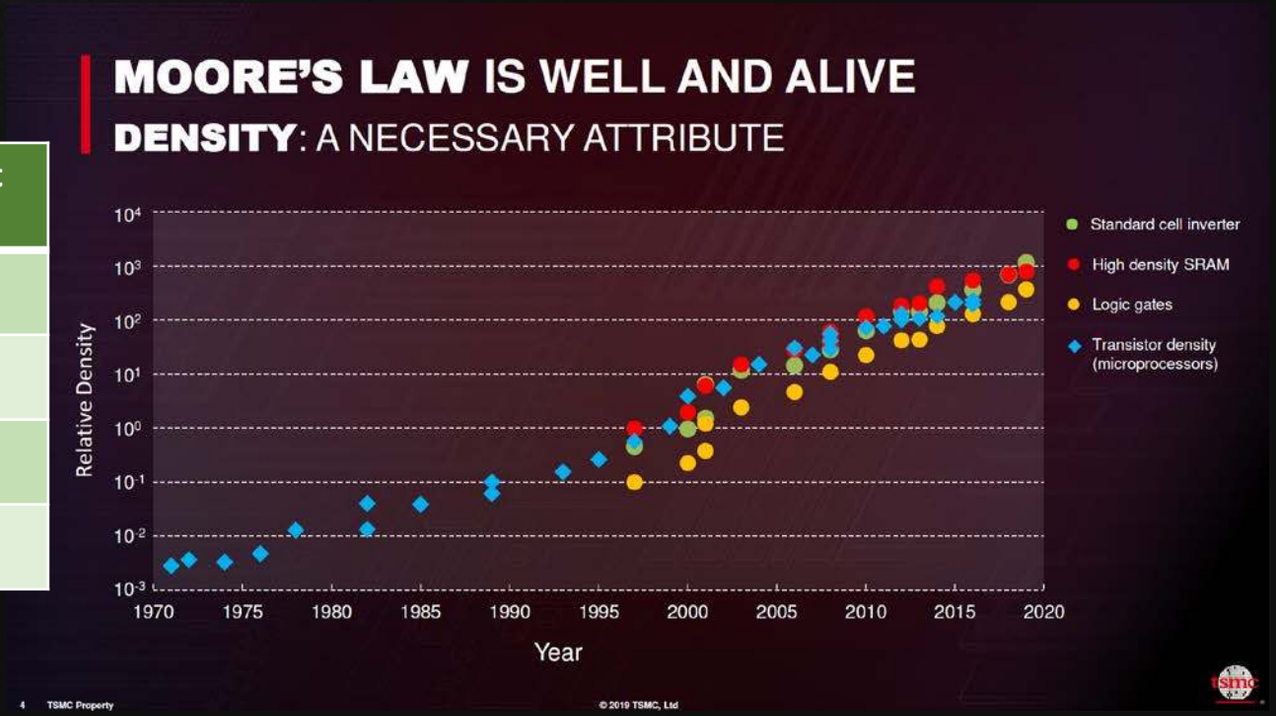
Xbox Series X | SOC Multi-Media, IO, Storage, misc.

- Video encoder/decoder:
  - Legacy 480p/1080p decoders + 4K/8K AVC and HEVC/VP9 HDR decode
  - AVC and HEVC HDR encode
- Display processor:
  - Full quality HDR/WCG, linear light HDR display processing with 3D LUT
  - HDMI 2.1 including ALLM, VRR, and 10Gb FRL with DSC
  - HDMI 2.1 10Gbps FRL with DSC enables HDR 444 YUV and RGB @ 8Kp60
- SERDES IO:
  - PCIe 8x5 Gen4
  - HDMI 2.1, 10Gbps FRL
- System Mass Storage:
  - Internal 1TB NVME SSD, x2 PCIe Gen4
  - External user accessible slot for 2nd NVME SSD, x2 PCIe Gen4
  - Blu-Ray 4K/UHD ODD
- Southbridge: PCIe attached, USB 3, SATA, system controller, SPI, I2C, etc.
- Networking: 1Gb Enet, LAN WiFi, MS Wireless Protocol Controller WiFi

Xbox Series X | Moore's Law Cost Issues -> Add HW Engines

- Moore's Law curve OK for effective logic density scaling, but cost is an issue
  - Diagram lower right, from: HC 31 2019, TSMC Keynote, Session 2.7, Page 4: see https://www.hotchips.org/hc31/
  - Power TSMC 16FF+ -> N7e @ historical full-node savings
  - N7e: added complex steps + complex structures + large heterogeneous IP SOC
  - **Higher wafer price AND lower yield => higher SOC die cost**
- ***=> Add improved HW, plus New HW accelerators to save die area and power***

SOC	Year	Process, Die area	Transistors	Die Cost
Xbox One	2013	TSMC 28HP, 375mm^2	4.8B	$
Xbox One S	2016	TSMC 16FF+, 237mm^2	5.4B	$-
Xbox One X	2017	TSMC 16FF+, 367mm^2	6.6B	$+
Xbox Series X	2020	TSMC N7e, 360mm^2	15.4B	$++



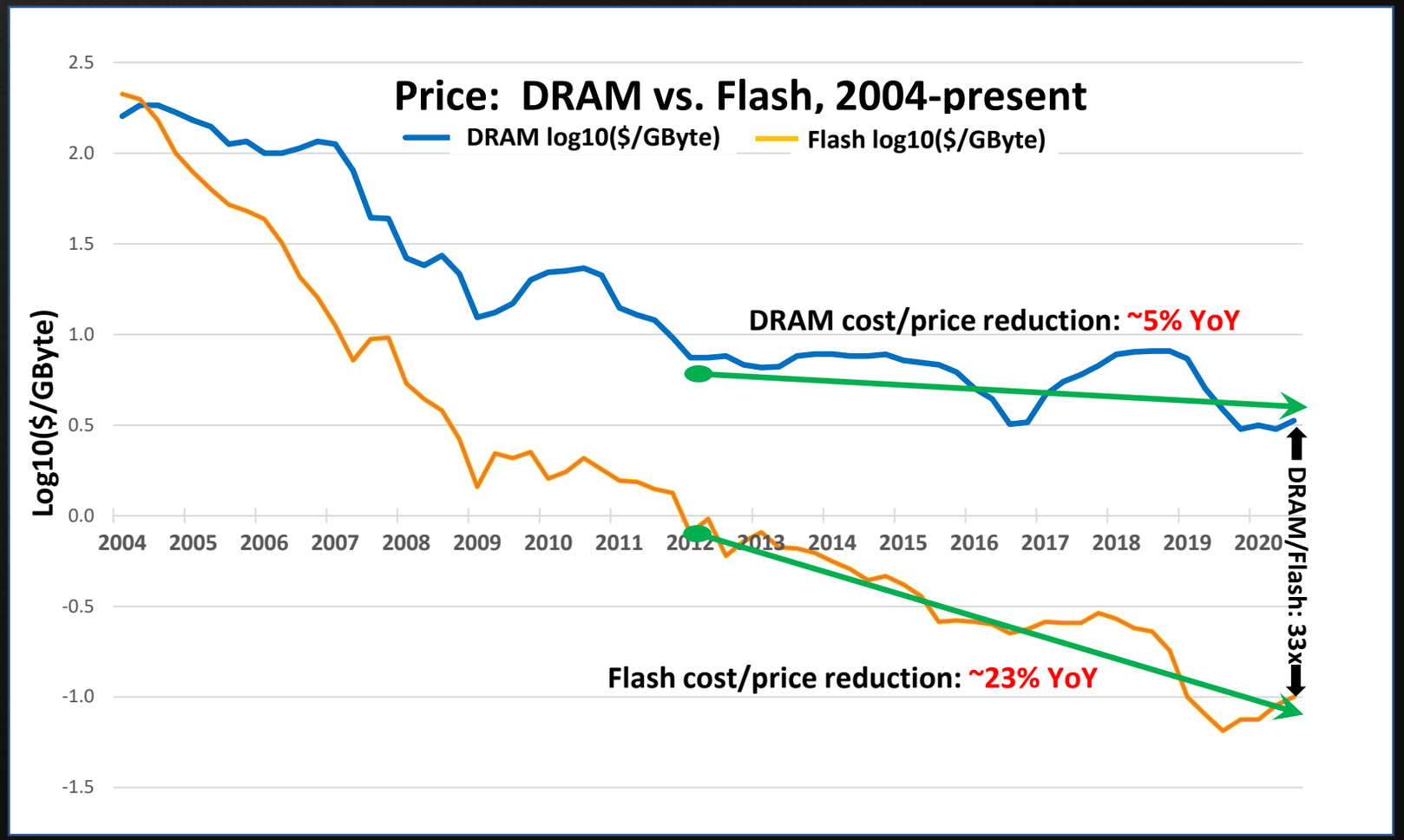
Xbox Series X | SOC Specs – MS Designed HW Engines

- Audio: greater SPFP HW math than ALL 8x CPUs in Xbox One X
  - CFPU2: 2x 4-way FP SIMD DSPs, 4x FP engines
    - Programmable, high throughput convolution, FFT, reverb, frequency domain audio
  - MOVAD: Opus real-time decoder, sample rate converter
    - >300x real-time channels of Opus HW decode
    - >100dB SNR SRC/pitching HW real-time perf matched to Opus decode
  - Logan: 4x DSP Cores, SRC, audio FX, XMA decode
    - >300x real-time channels of XMA HW decode

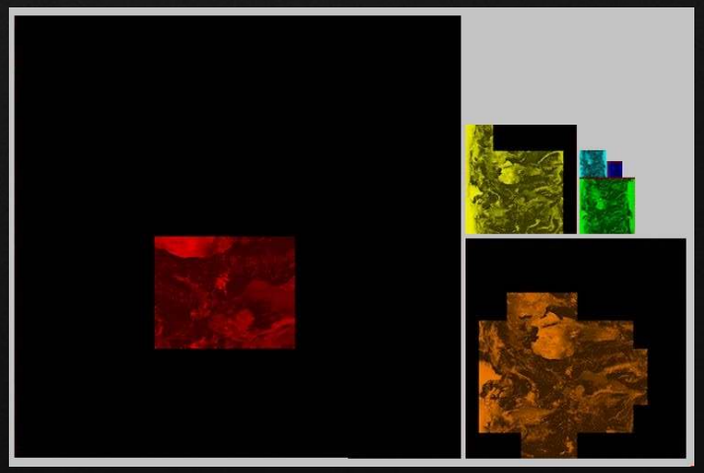
- Security and Decompression:
  - HSP/Pluton: Root of Trust, crypto, emCPUs, SHACK
    - SHACK: Secure HArdware Crypto Keys
  - MSP: crypto/hash/decomp @ NVMe SSD BW, saves 2+ Zen2 cores' workload
    - 4 high-BW HW crypto, total >5GB/s, plus 4 hash engines >5 GB/s
    - 2 general + texture decompression engines, total >6 GB/s output

Xbox Series X | Xbox Velocity Architecture Motivation

- Little DRAM cost reduction
  - -30% YoY traditionally
  - Only -5% YoY last ~8.5 years
- Flash still cost reducing well
  - ~23% YoY last ~8.5 years
- DRAM/Flash $/GB ratio:
  - **~33:1 today**
- **=> High BW NVMe SSD as backing store for DRAM game art content SW cache**
- Huge game SSD load time benefit Solves HDD areal density (^2) vs. linear bit-rate scaling issue
- Xbox team planning this transition to high BW flash to save DRAM footprint since 2007



Xbox Series X | XVA/SFS System – Shader Feedback Streaming



- Shader Feedback Streaming (SFS):
  - GPU SFS HW records active texture portions
  - Game prioritizes order of fetch/free
  - DirectStorage manages SSD, MSP crypto/decomp
- DRAM benefit: avg 2.5x game art capacity
- MSP enabled NVMe SSD storage savings:
  - Lossless MS XVA ~2:1 compression on BCn textures
    - MS XVA decomp supports higher RDO+lossy ratios
  - Opus audio compression
  - Deflate/Zlib general decompression

Upper Left: render with texture per-MIP level color overlay
Lower Left: SFS recording of active texture portions for globe

Xbox Series X | GPU

“We’re at a point where the technology is out of the way.” - Matt Booty

GPU goals: satisfy increased realism, screen resolution and frame rate

- Implement new algorithms in silicon w/o breaking the bank
- Efficient enhancements rather than isolated cores

Xbox Series X | GPU diagram

- 26 active Dual Compute Units
- Unified Geometry Engine
- Mesh Shading Geometry Engine
- Distributed Primitives & Rasterization
- Screen Tiled Color/Depth unit
- Multi-core Command Processor
- Three-level Cache, TLB

DirectX XII ULTIMATE

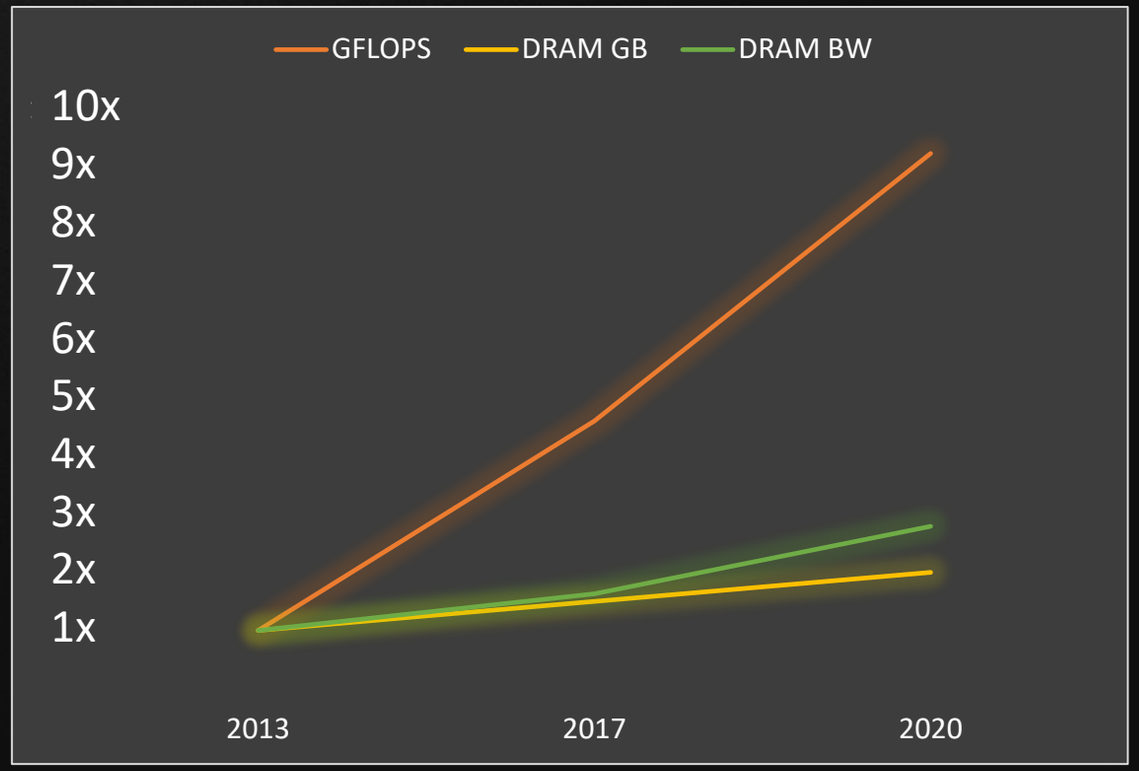
Windows | XBOX

Xbox Series X | Dual Compute Unit

- Four SIMDs plus 4 Scalar ALUs
- 32 Scalar FP32 FMAD per SIMD, 128 per Dual CU with data sharing
- Launch 7 instructions/clk per CU
  - 2 Vector ALU, 1 Vector Data, 2 Scalar, 2 Control
- 4 Texture or Ray ops/clk per CU
- Total per clk: 256 32b FP, 512 16b FP

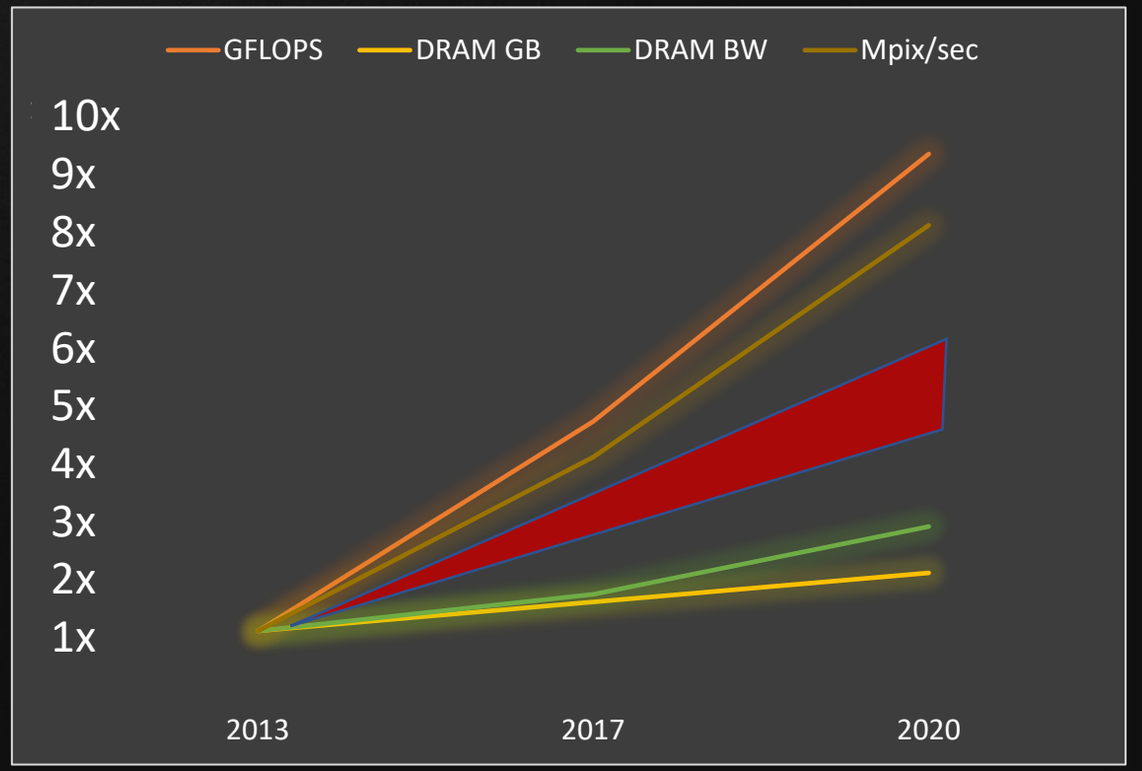
Xbox Series X | GPU Evolution

- **XBOX ONE**: 2013, 1080p display
  - 1.3 TFLOPS, 68+200 GB/sec, 1.6 Gtris/sec, 13 Gpix/sec
  - DX11.1+, megatexture, backward compatibility, scalar shader units
- **XBOX ONE X**: 2017, 4K display
  - 6 TFLOPS, 325 GB/sec, 4.4 Gtri/sec, 35 Gpix/sec
- **XBOX Series X**: 2020, 4K 120 Hz, 8K display
  - 12 TFLOPS, 560 GB/sec, 7.3 Gtri/sec, 116 Gpix/sec



Xbox Series X | GPU Evolution

- **XBOX ONE**: 2013, 1080p display
  - 1.3 TFLOPS, 68+200 GB/sec, 1.6 Gtris/sec, 13 Gpix/sec
  - DX11.1+, megatexture, backward compatibility, scalar shader units
- **XBOX ONE X**: 2017, 4K display
  - 6 TFLOPS, 325 GB/sec, 4.4 Gtri/sec, 35 Gpix/sec
- **XBOX Series X**: 2020, 4K 120 Hz, 8K display
  - 12 TFLOPS, 560 GB/sec, 7.3 Gtri/sec, 116 Gpix/sec

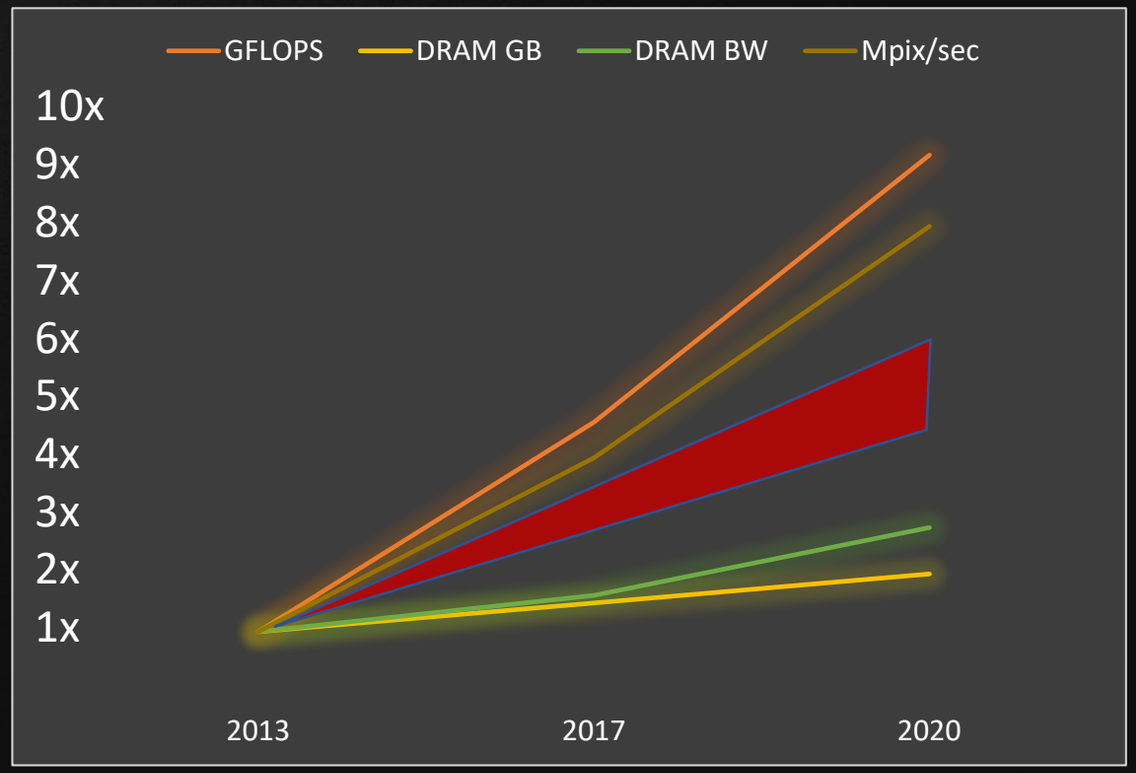


Xbox Series X | GPU Evolution

- **XBOX ONE**: 2013, 1080p display
  - 1.3 TFLOPS, 68+200 GB/sec, 1.6 Gtris/sec, 13 Gpix/sec
  - DX11.1+, megatexture, backward compatibility, scalar shader units
- **XBOX ONE X**: 2017, 4K display
  - 6 TFLOPS, 325 GB/sec, 4.4 Gtri/sec, 35 Gpix/sec
- **XBOX Series X**: 2020, 4K 120 Hz, 8K display
  - 12 TFLOPS, 560 GB/sec, 7.3 Gtri/sec, 116 Gpix/sec

How to fill 10x more pixels with 4-6x raw GPU and 1x the power consumption?

Answer: **patented innovations**



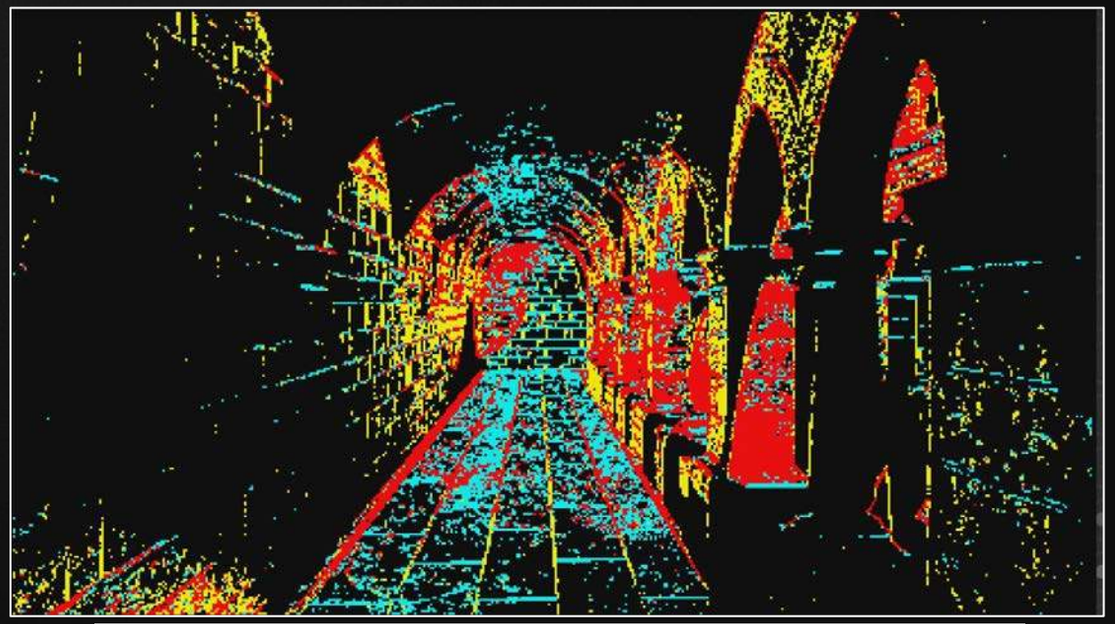
Xbox Series X | Variable Rate Shading

Shade fragments via finely controlled density bias
- Rate per x,y axis per 8x8 tile
- Coarse rates: 1x2, 2x1, or 2x2 pixels per color
- Fine rates: 1xAA to 8xAA

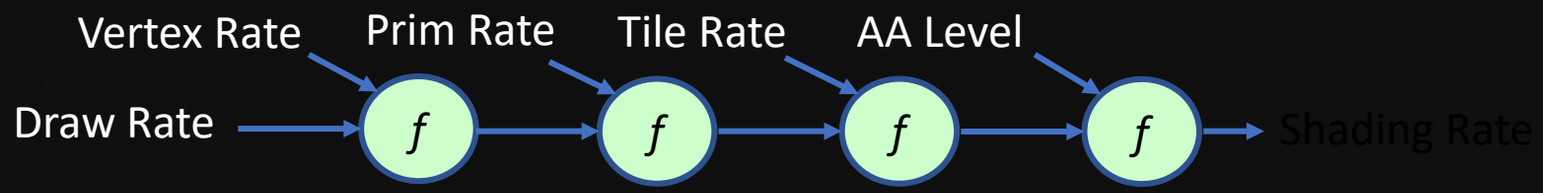
Modify per combination of draw, vertex, primitive, screen tile, and AA level

Full edge detail. No holes, warping, or checkerboard artifacts
Compatible with Temporal Anti-Aliasing techniques

Tiny area cost for 10-30% performance gain



2x2 2x1 1x2 1x1 0.5 fragments/pixel



Xbox Series X | Sampler Feedback Streaming

XBOX Velocity Architecture GPU Support

Old style PRT: virtualized texture via soft page faults

Burden of shader enlightenment

Serialization through driver memory manager

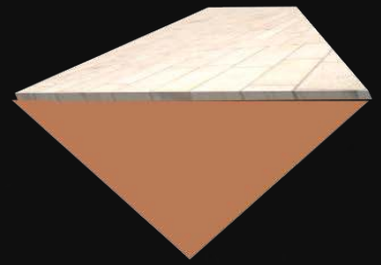
New HW: Two new MIP Mapping structures

LOD Tile Residency Map

Sampler instruction fetches clamped LOD

LOD Tile Request Map

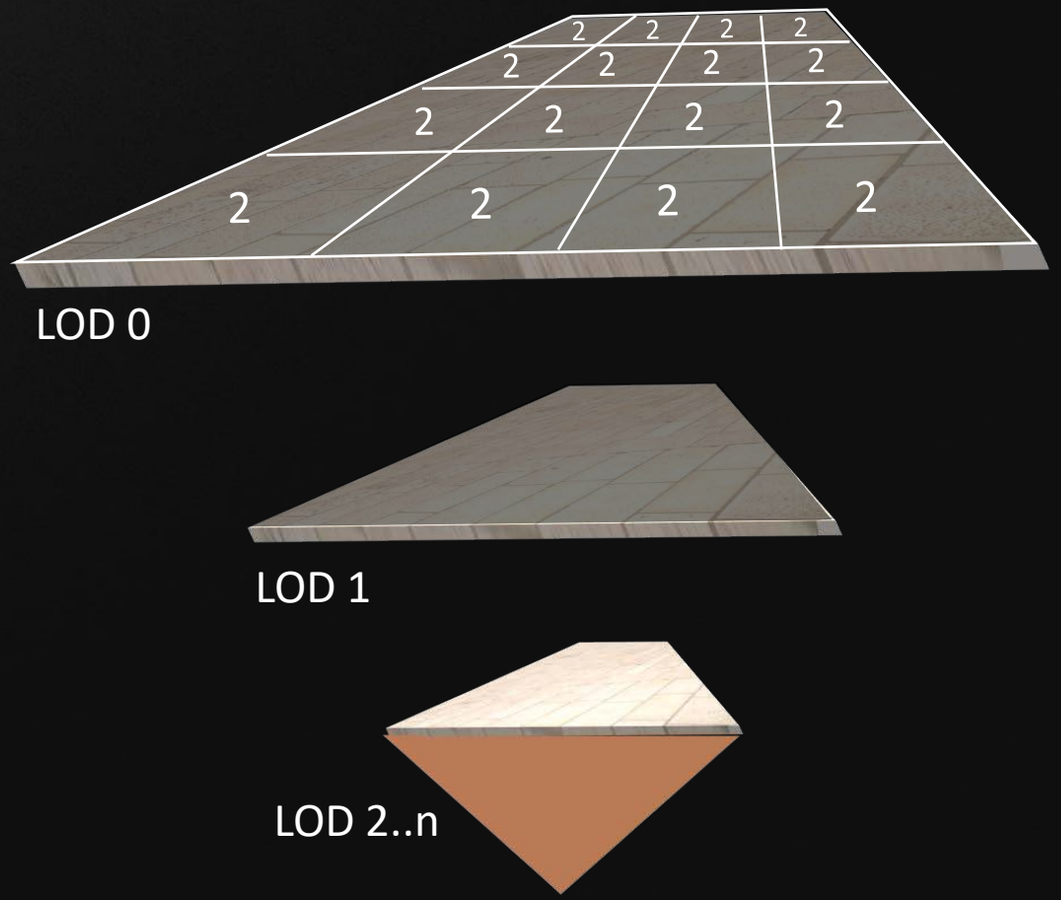
Accumulates min LOD that was needed



Simplified shader model

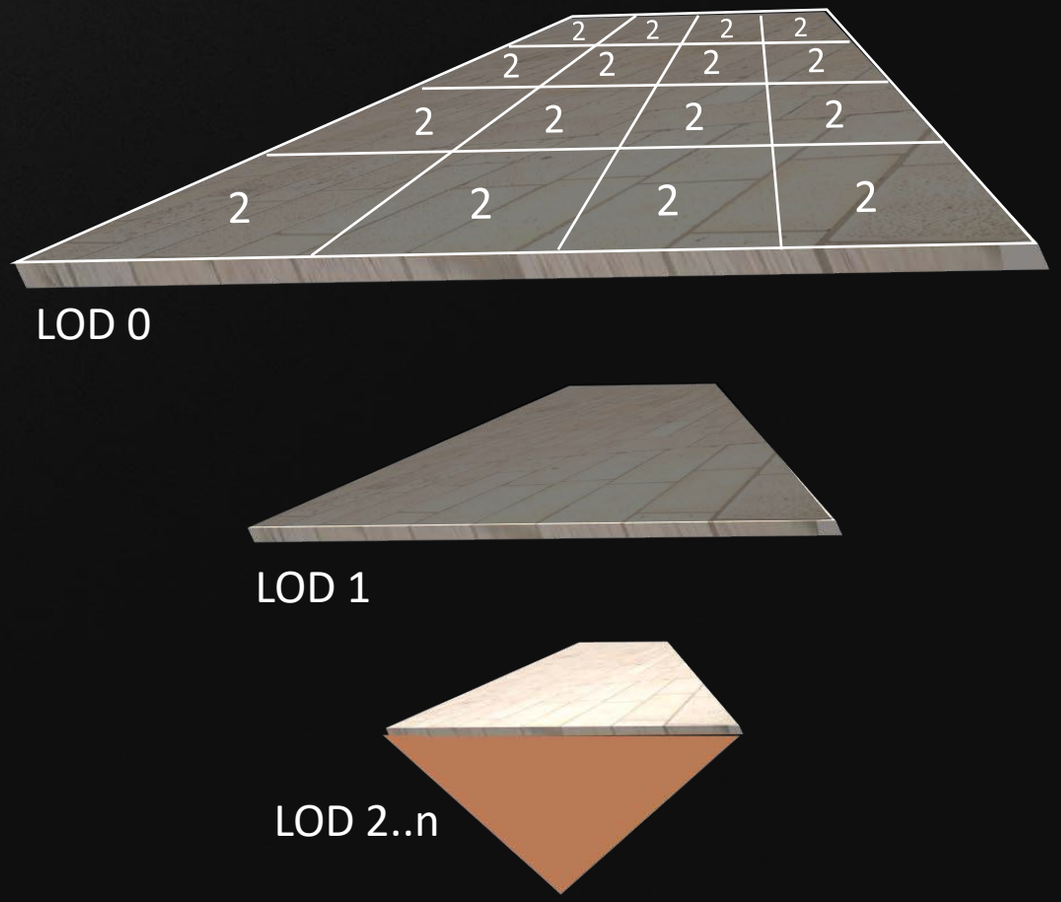
Xbox Series X | Sampler Feedback Streaming

1. Allocate entire texture, populate the small LODs (e.g. 2 and up), mark resident



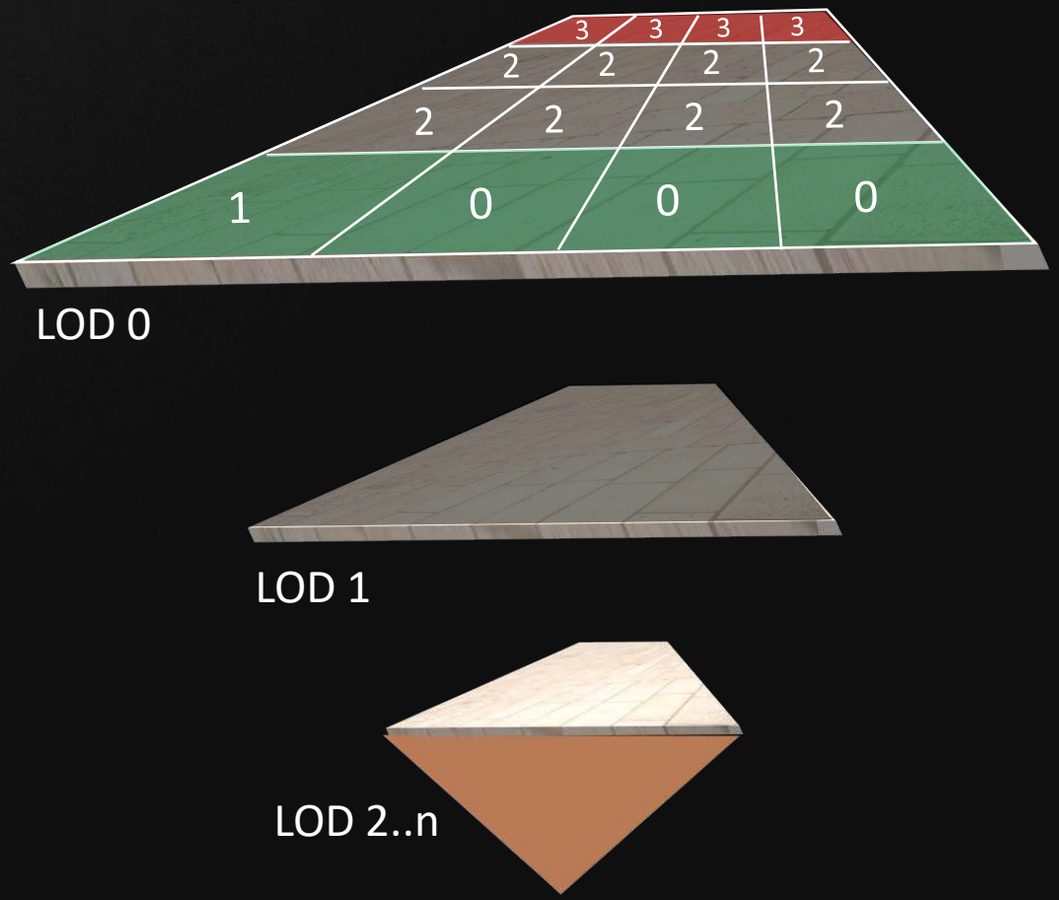
Xbox Series X | Sampler Feedback Streaming

1. Allocate entire texture, populate the small LODs (e.g. 2 and up), mark resident
2. Render frame:
   a) Sample (residency map, resident texture)



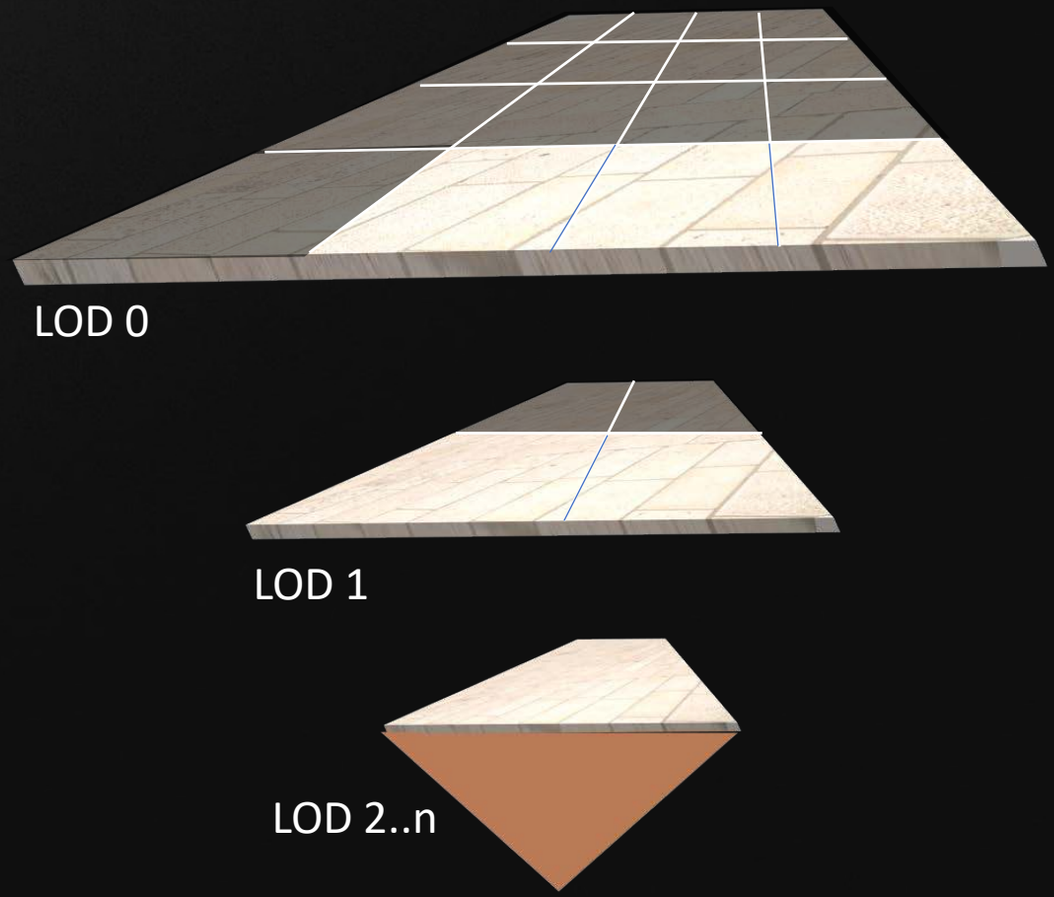
Xbox Series X | Sampler Feedback Streaming

1. Allocate entire texture, populate the small LODs (e.g. 2 and up), mark resident
2. Render frame:
   a) Sample (residency map, resident texture)
   b) Record requested LODs



Xbox Series X | Sampler Feedback Streaming

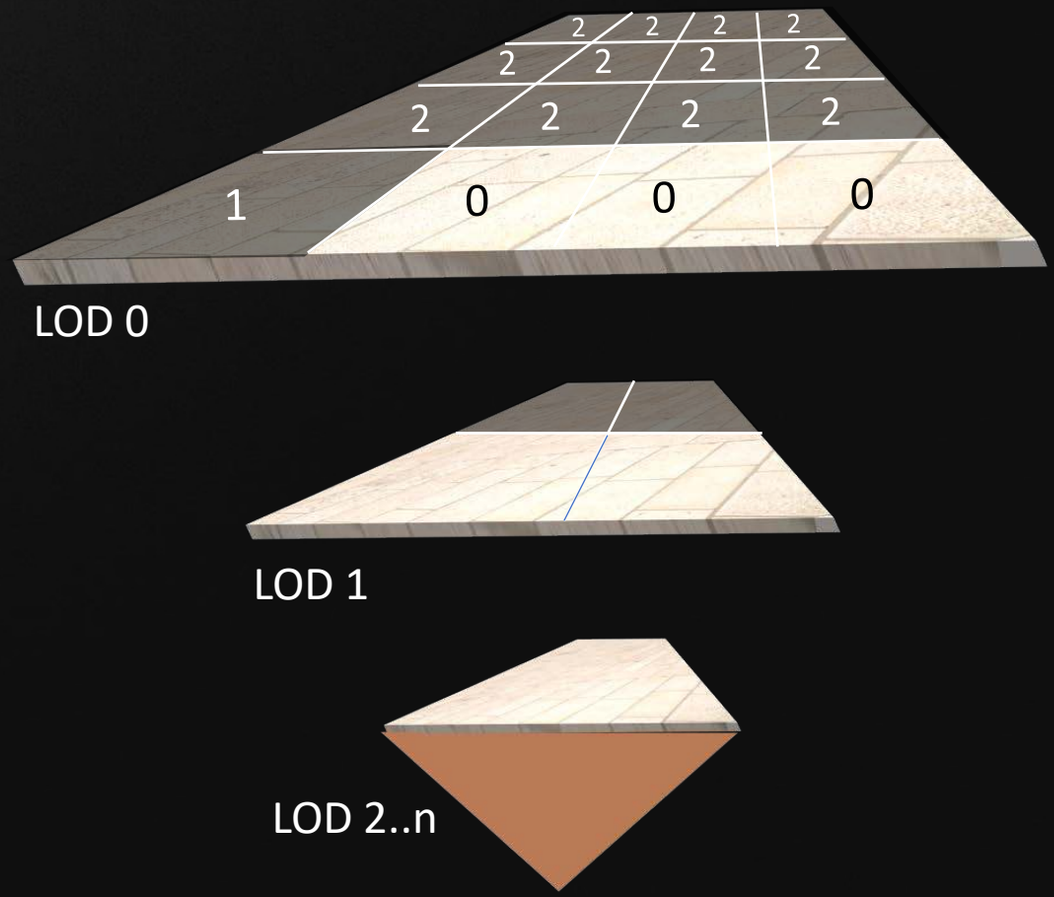
1. Allocate entire texture, populate the small LODs (e.g. 2 and up), mark resident
2. Render frame:
   a) Sample (residency map, resident texture)
   b) Record requested LODs
3. App inspects LOD record, evicts unused tiles, loads requested tiles



Xbox Series X | Sampler Feedback Streaming

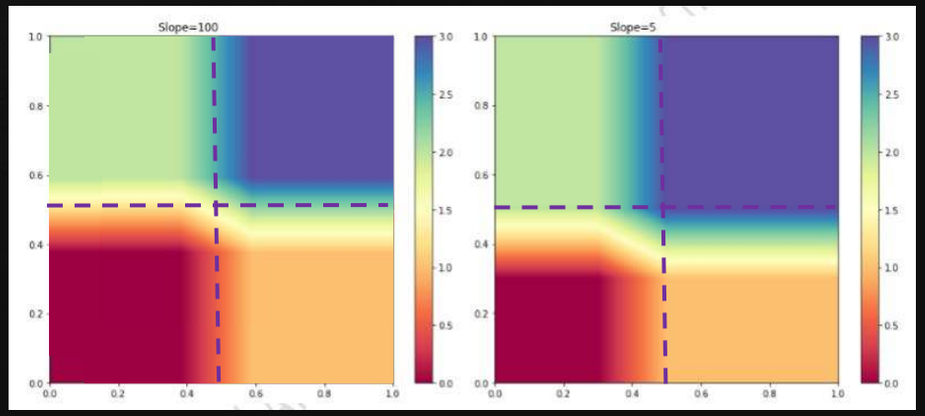
1. Allocate entire texture, populate the small LODs (e.g. 2 and up), mark resident
2. Render frame:
   a) Sample (residency map, resident texture)
   b) Record requested LODs
3. App inspects LOD record, evicts unused tiles, loads requested tiles
4. Update residency map

LOD 0



Xbox Series X | HW Custom Residency Map Filtering

Transition function allows just one map “texel” per texture tile
Smoothing minimizes transition artifacts
Traditional bilinear produces wrong answers
New filter: transitions always in coarser-LOD regions



Bilinear filtering New XSX filtering

Sampler Feedback Streaming:

Very small area cost, up to 60% I/O and memory savings

Xbox Series X | DirectX Ray Tracing Acceleration

Economical upgrade to traditional rendering
Not a complete replacement

Xbox Series X | DirectX Ray Tracing Acceleration

Custom ray-box and ray-triangle units

- 380G/sec ray-box peak, 95G/sec ray-tri peak
- Net perf depends on bandwidth, number of nodes/tris visited per ray
- Shader can run in parallel for BVH traversal, material shading, etc.
- Minor area cost for 3-10x acceleration

Xbox Series X | Other tricks

- ML inference acceleration for game (character behavior, resolution scaling)
  - Very small area cost, 3-10x perf improvement
- Two independent virtualized command streams
  - Different DX API levels
  - Separate virtual machines for main title OS, system OS
- 32b HDR format for rendering, blending, display
  - 999E5 -- 9-bit mantissas, 5-bit shared exponent
  - Better quality than 11:11:10; 50% BW and blend time vs. 64b

Xbox Series X | Optimized Games

Thank you!

Microsoft