



Rendering Roblox

Vulkan Optimisations on PowerVR

Presented By: Arseny Kapoulkine David Hamill



What is Roblox?

- Online multiplayer game creation platform
- All content is user generated
- Windows, macOS, iOS, Android, Xbox One
- 100M+ MAU, 2.5M+ CCU











Graphics scaling

- Substantial differences between devices
 - PowerVR SGX543: Apple iPad 2 (2011), Samsung Galaxy S4 (2013)
 - PowerVR GM 9446: Oppo Reno Z (2019)
- Developers make the content once
- It's our job to run it well*



Full spectrum design

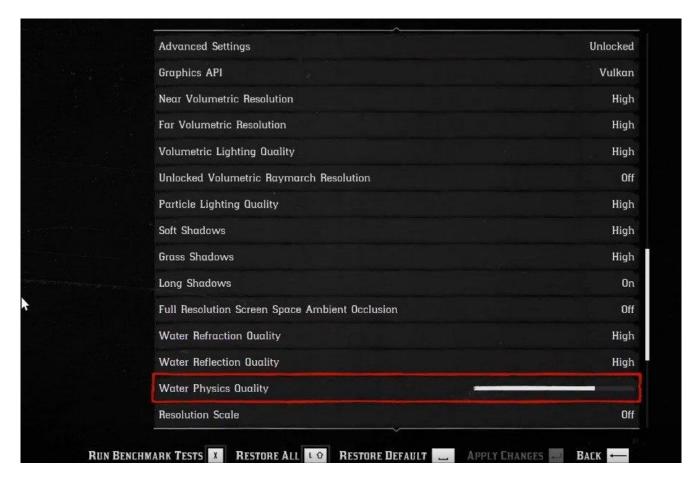
- How do you implement a graphics feature?
 - Build for the lowest common denominator and upscale?
 - Build for the highest end GPU and optimize?
- Build a feature for the lowest common denominator and upscale?
 - Pro: Consistent behavior
 - Con: Poor quality
- Build a feature for the highest end GPU and optimize?
 - Pro: Great quality on high-end
 - Con: Poor performance on low-end, inconsistent behavior, battery/thermal limits



Full spectrum design

- How do you implement a graphics feature?
 - Build for the lowest common denominator and upscale?
 - Build for the highest end GPU and optimize?
- Sometimes we pick one and iterate
- Prefer to design for the entire spectrum from the beginning







- Advanced settings give users lots of control! However...
- Suboptimal performance for users who can't tune these well
 - ... or know about them! Especially on mobile.
- Performance is often counter-intuitive
 - CPU vs GPU bottlenecks
 - Example: denser grass is faster
- Different games have different performance characteristics



- We have a single quality slider
 - ... arguably we made it too simple and need a separate "draw distance" slider
- By default, we automatically tune it based on observed performance
 - Do we have headroom on CPU vs GPU?
- All graphics effects are using this to balance quality vs performance
 - NB: Internally the slider is more granular than it is in UI
- Less control for the user, but easier for developers and graphics engineers





- Depending on the effect, scaling strategy can be different!
- Disable effect completely
 - Possible if not gameplay critical
 - Example: SSAO
- Geometry / shading LOD
 - Reduce computational cost while reducing quality
 - Important to keep the overall color balance!
- Temporal throttling
 - Visual pops due to late updates
 - Examples: lighting, terrain LOD



Lighting system: goals

- No baking. Every object can move at any point in time.
- Can't disable lighting altogether large visual and gameplay implications
- Want minimal "performance" annotations from content
- Want to run on low-end laptop/mobile (D3D9 / GL2 / GLES2 class)
- Want good looking lighting on high-end laptop/mobile



Lighting system: high level overview

Features

- Light sources: sun/moon, sky, local lights (point, spot, surface)
- Geometry and light sources are dynamic
- All light sources can cast shadows

Today

- Coarse voxel lighting: runs everywhere; efficient, soft lighting
- Shadow maps: runs on mid-end to high-end; high quality sun shadows

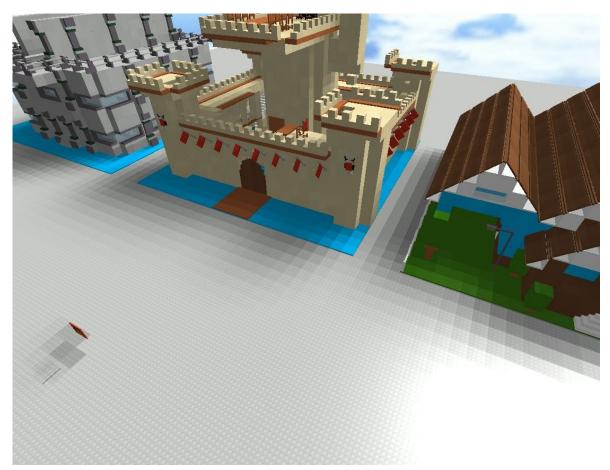
Future

- Forward+: high-end, high quality local lights
- Better voxel lighting: better skylight



Lighting system: phase 0, voxels

- Most of the work is done on CPU
- Voxelize all geometry dynamically
- Compute light impact per voxel
 - Sunlight
 - Skylight
 - Accumulated local light RGB
- Upload the information to GPU
 - RGB: light + skylight * skycolor
 - A: sunlight
- Final lighting in fragment shader





Lighting system: phase 0, voxels

- Managing CPU performance
 - Voxel grid is split into chunks
 - Only a handful of chunks update each frame
 - Fixed quality, but updates can be "stale"
 - Update kernels use hand-optimized SSE2 / NEON
- Managing GPU performance
 - A single 3D texture lookup with bilinear* filtering
 - GLES2: emulate 3D texture lookup with 2 2D texture atlas lookups
- Fully decoupled geometry vs light complexity



Lighting system: phase 1, better voxels

- Keep the overall system design
- HDR light color
 - Encoded using RGBM to save space
- Store skylight separately
 - Better integration of sky into BRDF
- Anisotropic occupancy
 - Voxelizer keeps 3 axial values per voxel
 - Critical for content compatibility
- Results are closer to shadow maps/Forward+





Lighting system: phase 1, performance

- CPU cost is larger but manageable
 - Anisotropic occupancy rasterization is more expensive
 - RGBM encoding is almost free
- GPU cost is still relatively small
 - Fetch two textures
 - RGBM decode (future: RGB9E5 / R11G11B10F?)
- CPU-GPU upload
 - We upload 1-4 chunks each frame
 - Each chunk is 32x32x16 volume (16K voxels) = 128KB per chunk
 - vkCmdCopyBufferToImage copies 128KB-512KB of data from a VkBuffer each frame
 - ... is this fast?







Analysing Robloxon Power VR g p u S



What is Imagination Technologies?

PowerVR mobile GPUs, Rogue architecture and new A-Series architecture

VideoLogic discrete graphics

Arcade machines,

Sega Dreamcast,

PS Vita

iPhones and iPads



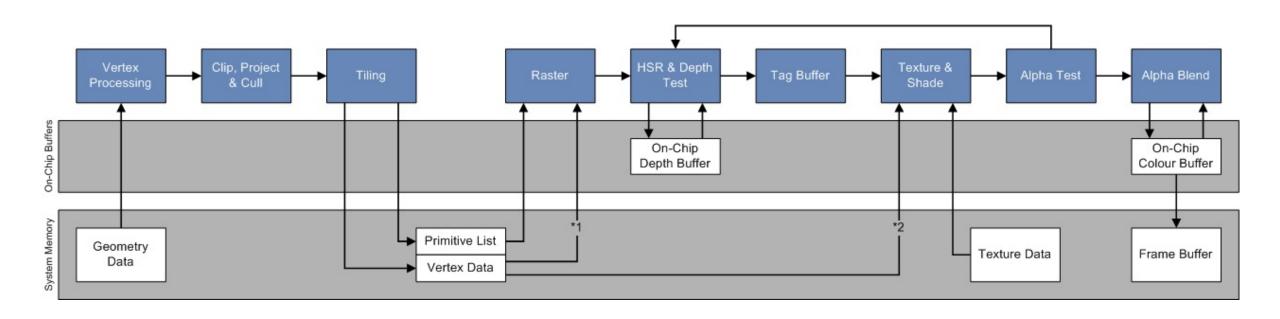
Now, neural network accelerators, networking chips, and in the near future hardware ray-tracing...



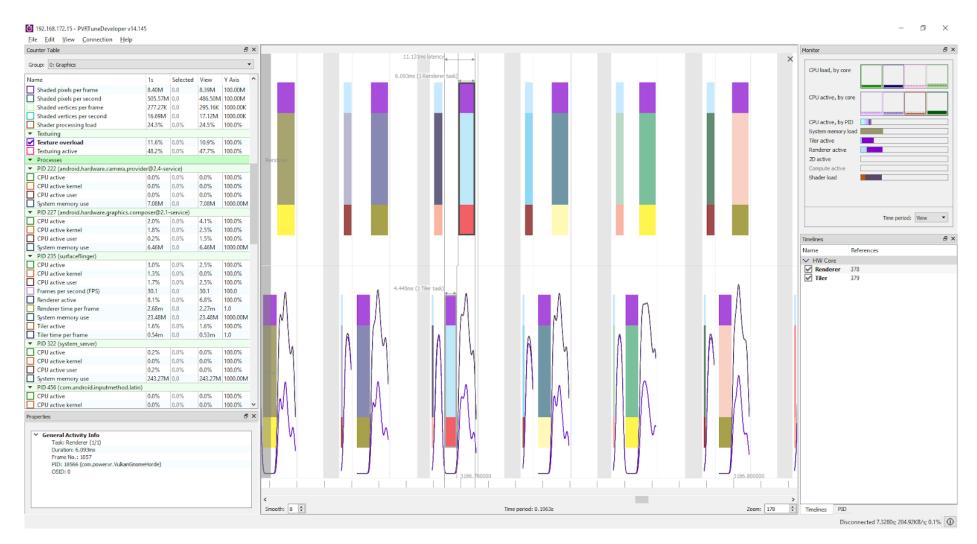




PowerVR Tile based deferred rendering

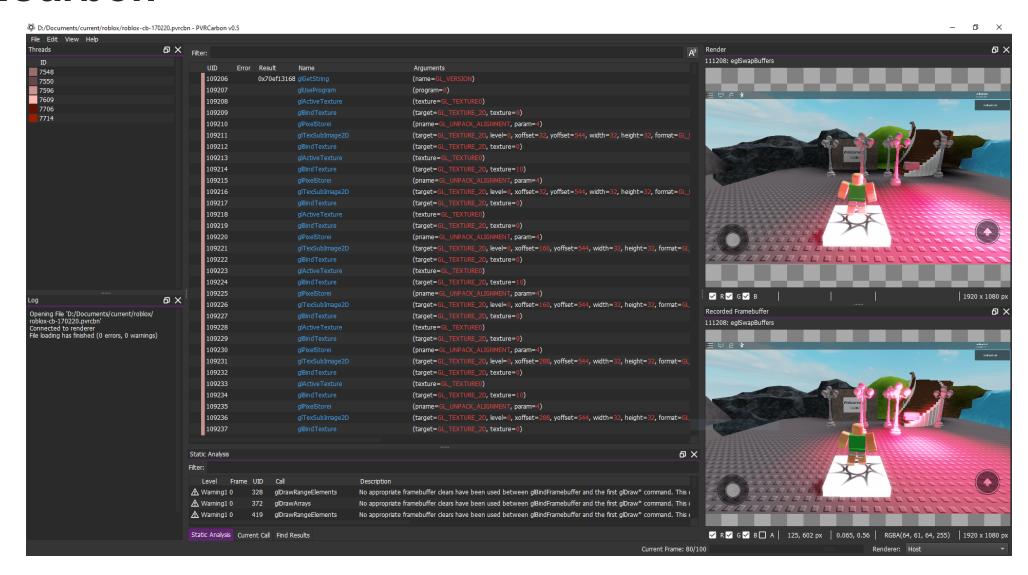


PVRTune



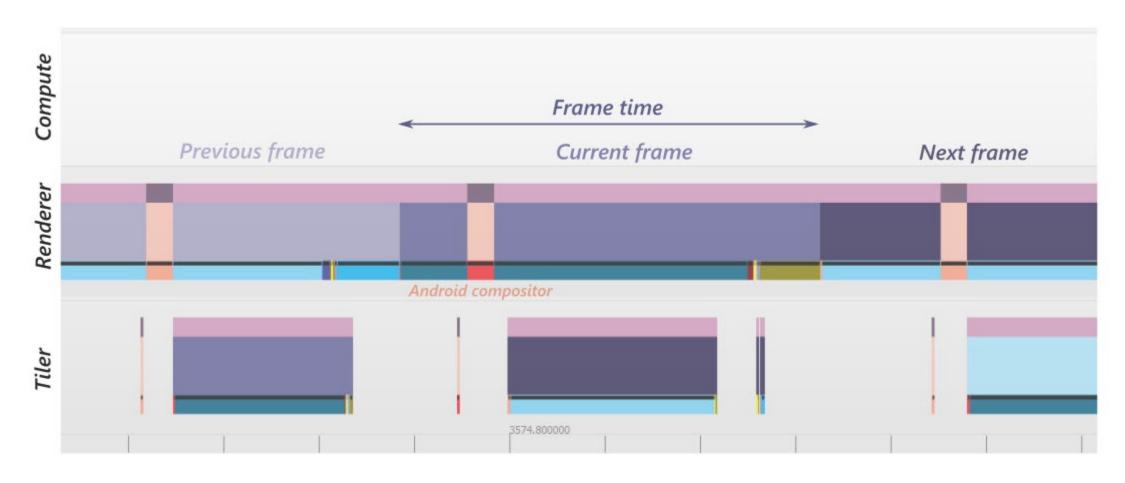


PVRCarbon





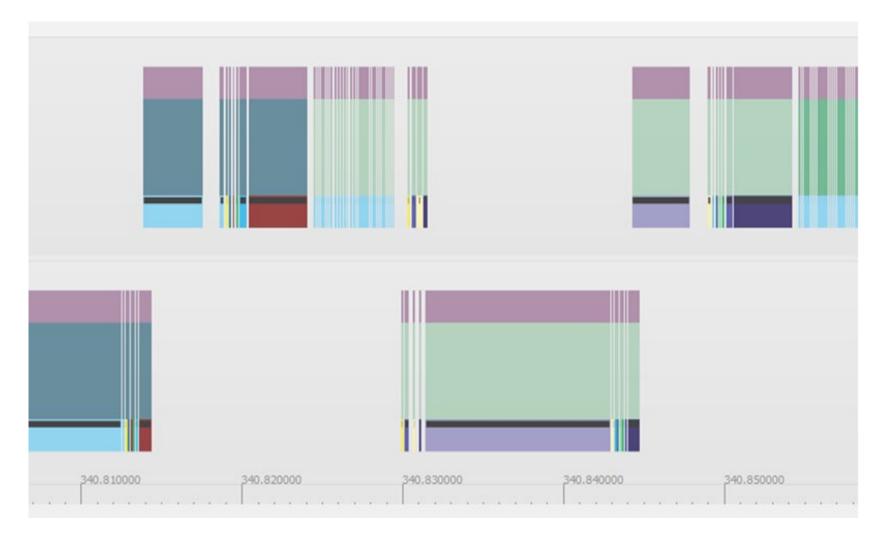
PVRTune timeline



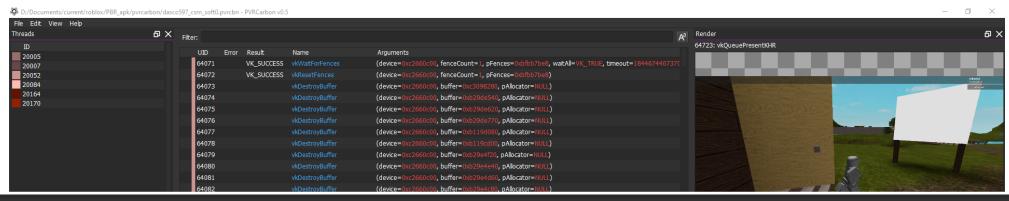
Vulkan – use double/triple buffering please



Analysing Roblox on PowerVR

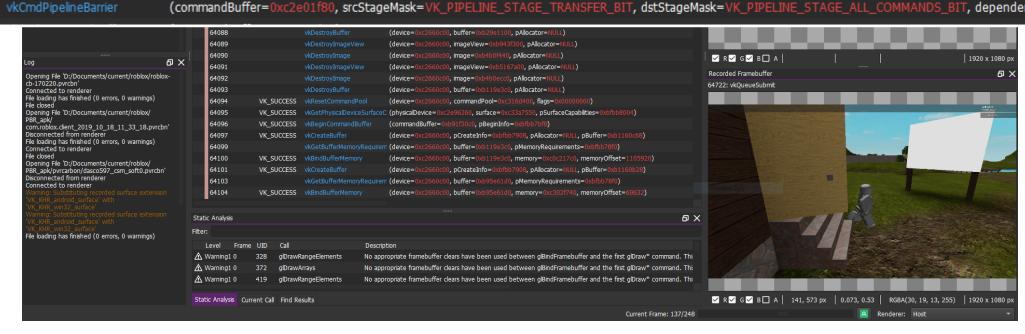


PVRCarbon analysis



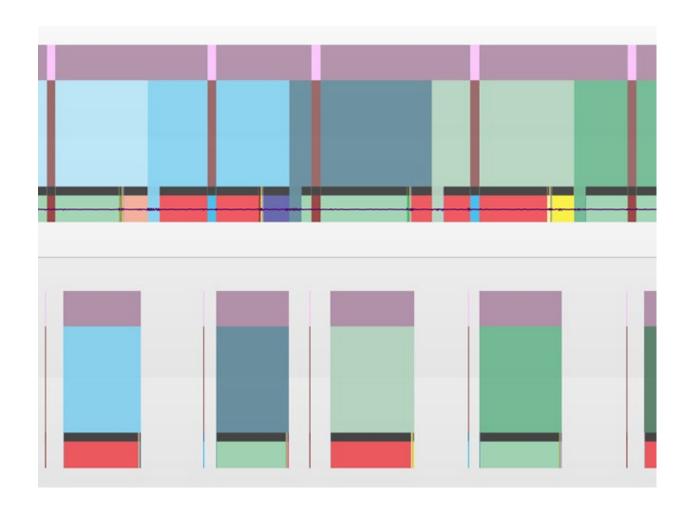
97371 vkCmdCopyBufferToImage 97372 vkCmdPipelineBarrier

(commandBuffer=0xc2e01f80, srcBuffer=0xb92870d0, dstImage=0xc0b10140, dstImageLayout=VK_IMAGE_LAYOUT_TRANSFER_DST_OPTIMAL, regionCount=1, pRegic (commandBuffer=0xc2e01f80, srcStageMask=VK_PIPELINE_STAGE_TRANSFER_BIT, dstStageMask=VK_PIPELINE_STAGE_ALL_COMMANDS_BIT, dependencyFlags=0x00





New tune recording



Glorious task packing!



GLSL optimisations

```
Group scalar arithmetic
            vec 1 = sqrt(vec 2);
                                           vec_1 = vec_2 * inversesqrt(vec_2);
      float_0 = max(float_1, 0.0);
                                           float_0 = clamp(float_1, 0.0, 1.0);
                   abs() neg() and clamp(..., 0.0, 1.0) can be free
   vec_0 = vec_1 * vec_2 + vec_3; \longleftrightarrow vec_0 = clamp(vec_1 * vec_2 + vec_3, 0.0, 1.0);
```

PowerVR developer documents: https://docs.imgtec.com/



GLSL optimisations - utilisation

Too many registers being used

- -> less threads can be processed in a cluster at a time
- -> reduced utilisation

Mediump allows PowerVR to use 16 bit floats

-> reduce register pressure, increase occupancy (up to 100%)



GLSL optimisations

PBR shader:

- A-Series: cycle count decreased by 9%
- Rogue (Oppo Reno): cycle count decreased by 12% and utilisation improved by 33%

PBR+IBL shader:

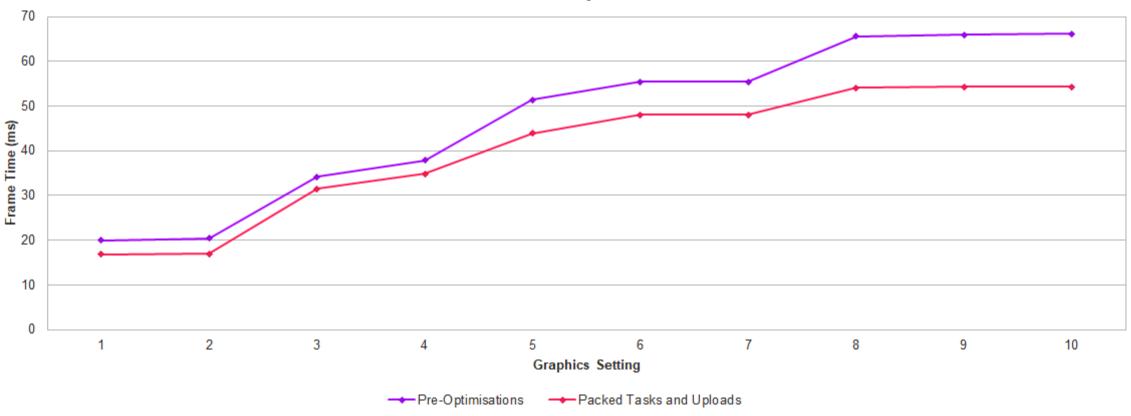
- A-Series: cycle count decreased by 9%
- Rogue (Oppo Reno): cycle count decreased by 20% and utilisation improved by 36%.

Use PVRShaderEditor to analyse cycle count and disassembly



Analysing Roblox on PowerVR

Performance improvements







Lighting Shadows Roblox

and in

Lighting system: phase 2, shadow maps

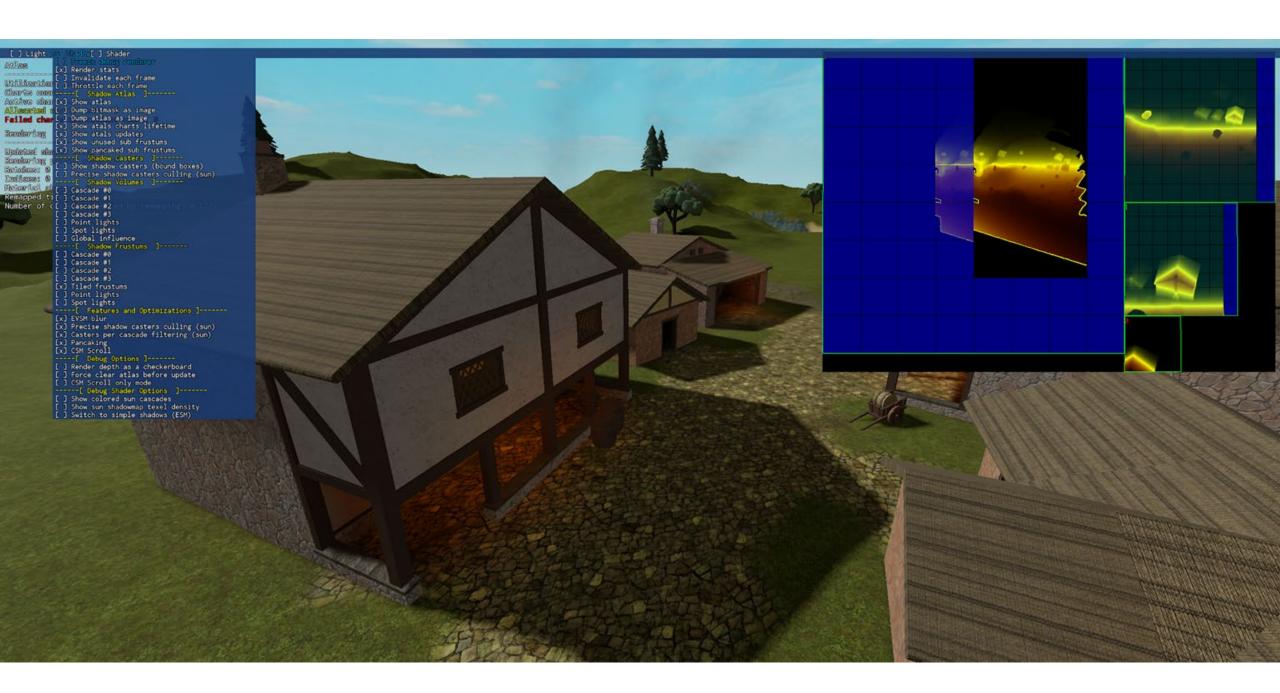
- Voxel shadows are too coarse
- Shadow maps to the rescue!
- Excellent quality
- Challenge: "soft" shadows
- Challenge: high cost to re-render
- Challenge: many shadow casting lights
- Only released sun shadows to simplify

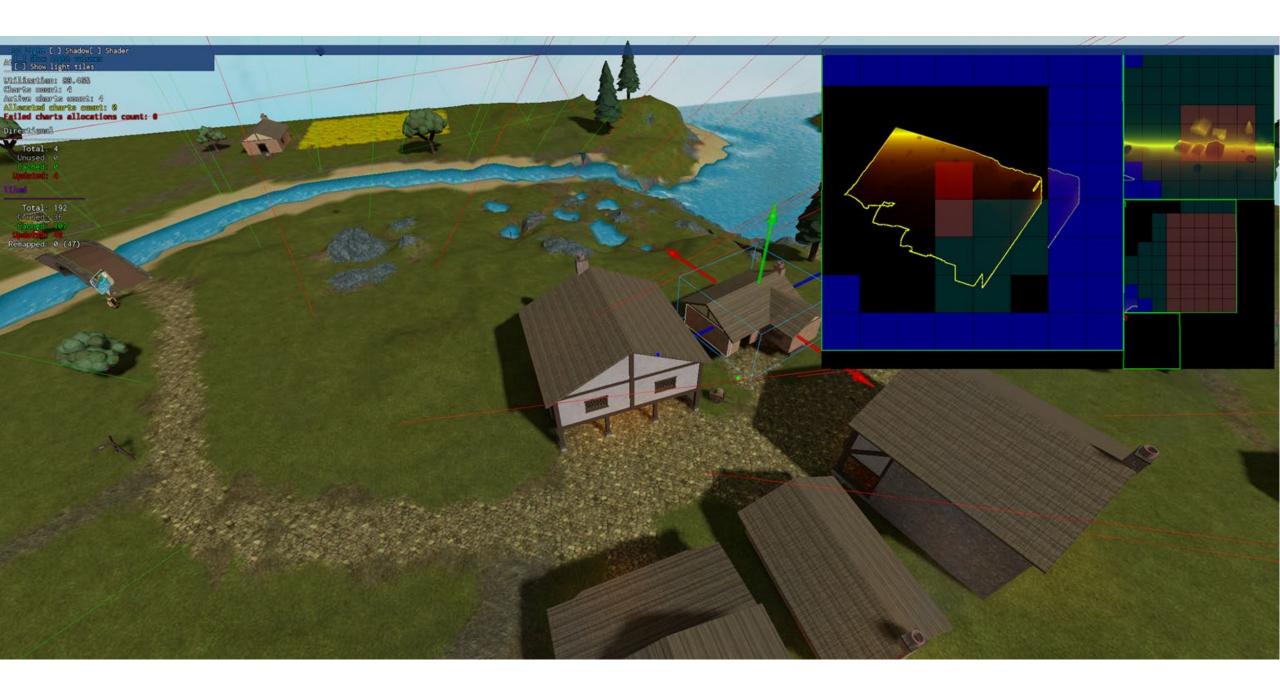


Lighting system: rendering shadow maps

- Cascaded tiled shadow maps
 - 1-4 cascades depending on the quality levels
 - Far cascades are split into tiles to be able to throttle shadow updates
- CSM Scrolling (Insomniac @ SIGGRAPH 2012)
 - CSM Scrolling: An acceleration technique for the rendering of cascaded shadow maps
- When we need to update a cascade, it's done in single pass
 - If multiple tiles are marked as dirty, we re-render geometry in these tiles
 - CPU does per-tile frustum culling
- Light direction changes invalidate caching:(
 - Haven't found a good solution to this



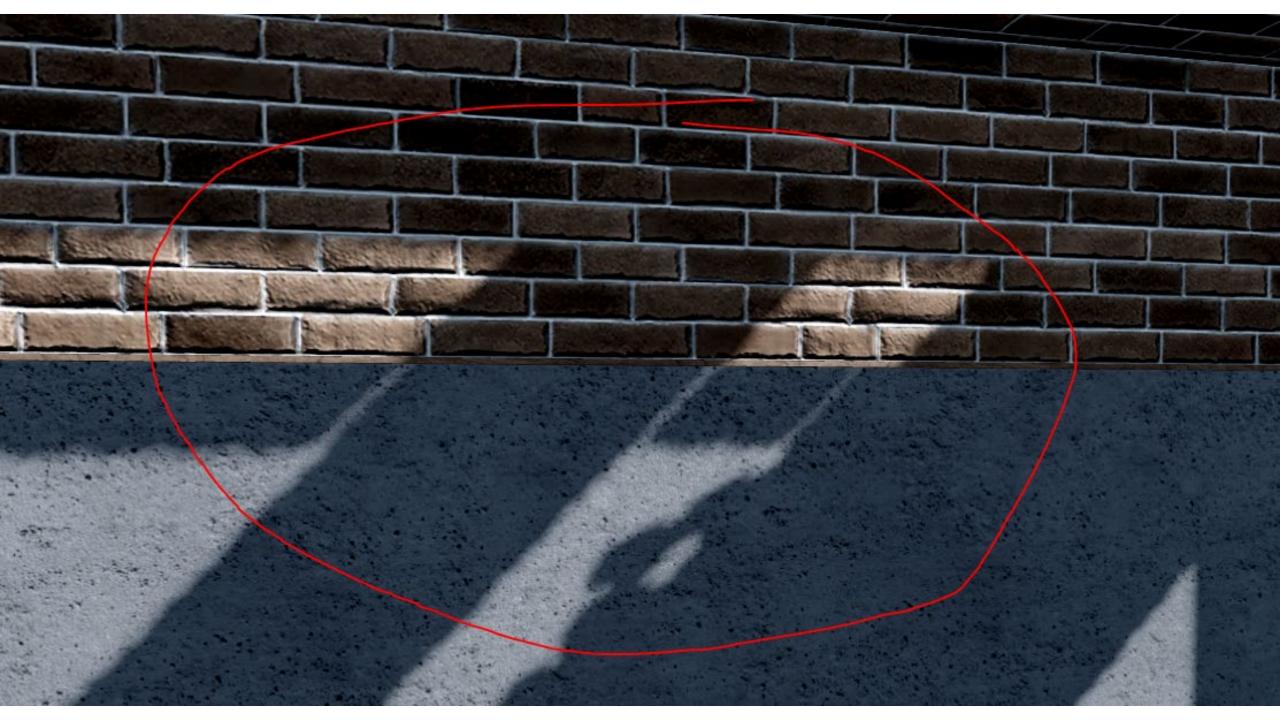




Lighting system: shadow softness

- Want a wide range of shadow map "softness"
 - Wide PCF kernel expensive at high resolutions
 - Screen-space filtering impractical due to transparent geometry and tiler performance
- EVSM to the rescue!
 - Fully decouples shadow rendering from sampling
 - Light leaks...





Lighting system: making EVSM practical

- Light leaks
 - Reduce Z range!
 - Tightly fit shadow frustum to receiving geometry
 - "Pancaking" when rendering caster geometry
 - VkPipelineRasterizationStateCreateInfo::depthClampEnable
 - Over-darkening when sampling EVSM
- Performance
 - Use two moments (RG16F) on mobile, four (RGBA16F) on desktop
 - Render depth only, reconstruct exponentially warped moments during first blur pass
- Integrating with tiling...
 - Blurring is per tile to make it fully incremental, we cache blurred EVSM contents
 - CSM scrolling might require adjusting the cascade depth
 - exp(z + offset) = exp(z) * exp(offset)



Lighting system: blurring shadow maps

- Standard two-pass blur (horizontal & vertical)
- First pass reads directly from depth texture (D16/D24)
 - Need to reconstruct EVSM moments from depth
 - Can't use bilinear filter optimizations :(
- Second pass reads from EVSM texture
 - Fewer taps!
 - See "Efficient Gaussian blur with linear sampling"
- Filter width varies from 3x3 to 17x17
 - Depends on developer-configured parameter, "ShadowSoftness"





Lighting system: shadow MSAA

- Wide blurs are pretty expensive
 - Render depth at full resolution, two pass blur
 - Three full stores of tile contents to memory
- Since shadow moments are filterable, we can use MSAA!
 - Use 4x MSAA to reduce shadow resolution by 2x2
 - Fragment shader directly outputs shadow moments (render pass isn't depth only anymore)
 - Subsequent filters are smaller (2x) and can use bilinear optimizations
- On-chip resolve is crucial
 - *Don't* use vkCmdResolveImage!
 - Instead, use pResolveAttachments in render pass
- MSAA becomes an almost free way to increase shadow quality



Lighting system: sampling shadow maps

- Find a single cascade per pixel
 - Some ALU to do point-sphere tests per cascade
 - Because of throttling & quality concerns can't use pixel depth
- Sample EVSM texture
 - Requires adjusting UV coordinates
 - All shadow maps are stored in a single shadow atlas
- Compute shadow factor
 - Exponential depth warping
 - 2 or 4 moment VSM
 - Apply over-darkening to reduce leaks





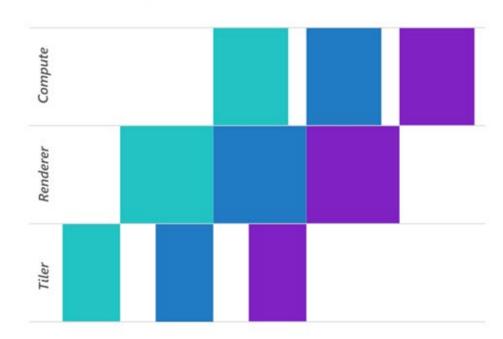


Compute Optimisations for PowerVR

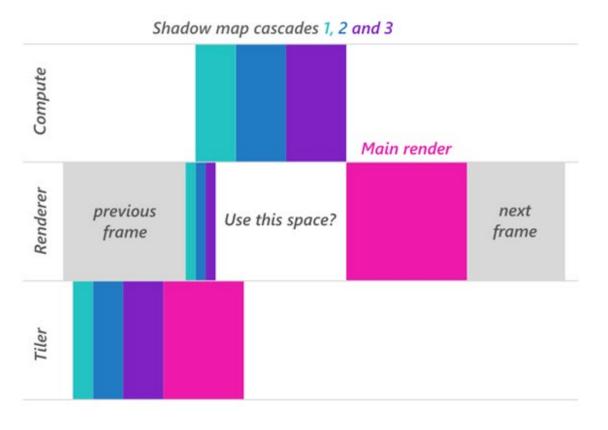


Move gaussian blur to compute?

Compute post-processing allows more scheduling flexibility



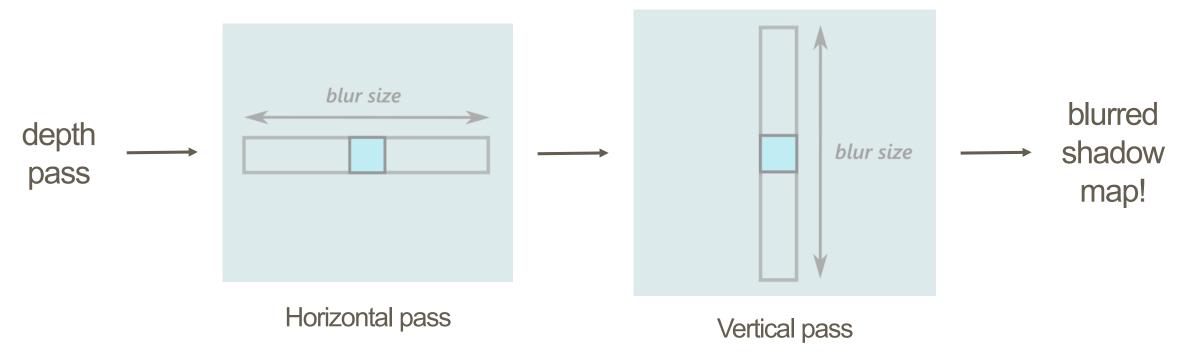
What about work in the middle of the frame?





Separable-kernel gaussian blur

- Perform 2D gaussian blur with two 1D passes
- Mathematically equivalent (rank 1 matrix)
- 2n texture fetches instead of n²



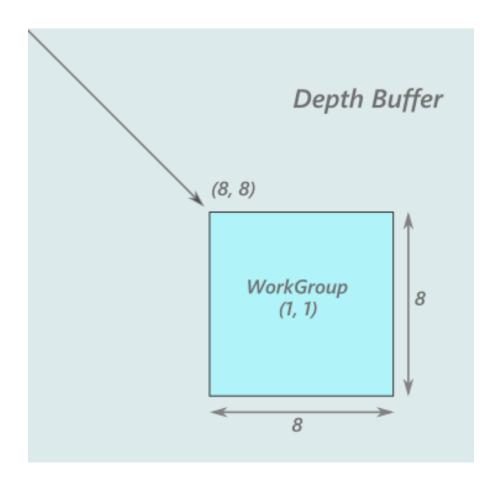


Compute gaussian blur gather algorithm

Optimal work-group size on PowerVR is 32

8x4 work group size processing an 8x8 area

Experiment with looping shader multiple times (2 texels per thread here)

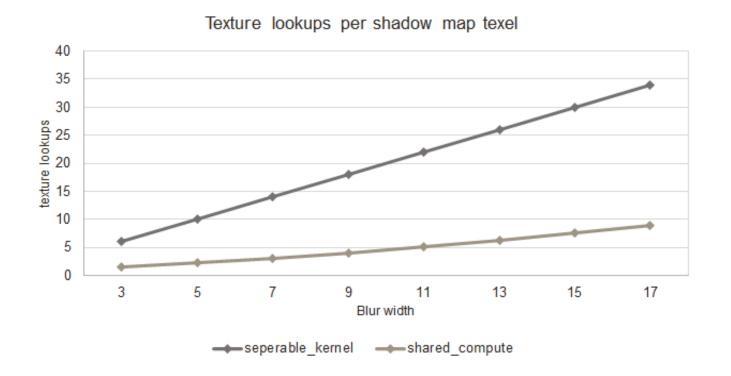


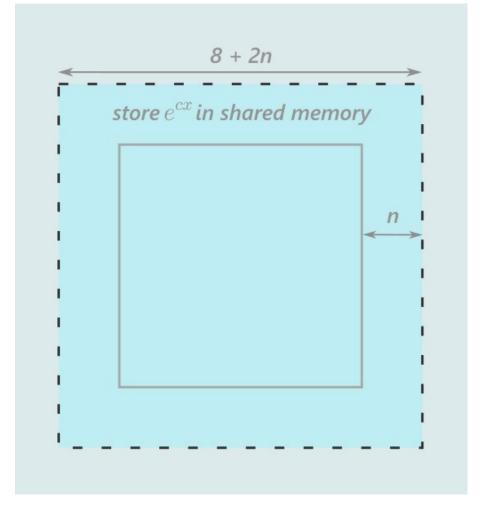


Compute gaussian blur gather algorithm

Read depth values including surrounding area into shared memory

Reduces the number of texture fetches

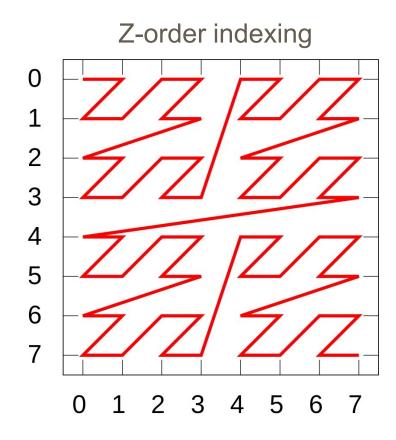


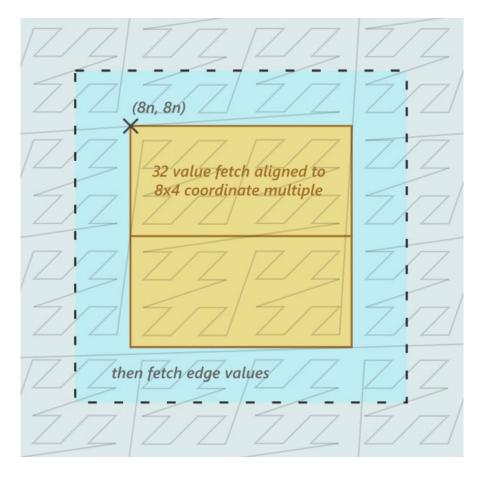




Morton ordering

Images created with VK_IMAGE_TILING_OPTIMAL addressed with Morton ordering Improve cache efficiency by loading an aligned area of texels (512 bit cache line)







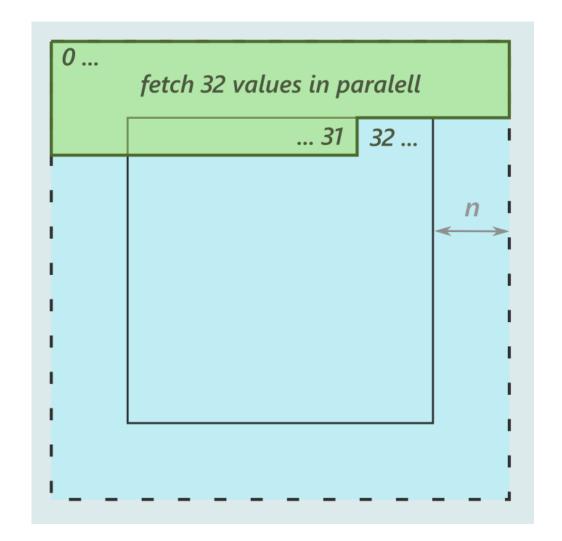
Linear indexing

	Bank 1				Bank 2				Bank 3				Bank 4			
dword	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
О																
16																
32																
4080	L												L			

Shared memory:

- Read up to 16 adjacent floats 1 cycle
- Read multiple rows in a bank in n cycles
- Read = 1 cycle, write = 4 cycles

```
// our 32 threads write to adjacent memory simultaneously
float warped_depth = warp(texelFetch(depth_buffer,
    gl_LocalInvocationID.xy + offset).x);
shared_storage[gl_LocalInvocationIndex + n*32].x =
    warped_depth;
```





Horizontal pass

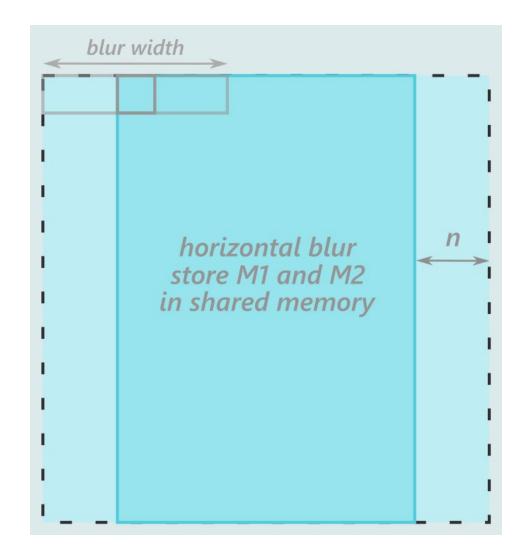
Write moments M1 and M2 to shared memory

$$M_1 = E(x) = \int_{-\infty}^{\infty} x p(x) dx$$

$$M_2 = E(x^2) = \int_{-\infty}^{\infty} x^2 p(x) dx$$

VSM algorithm:

http://www.punkuser.net/vsm/vsm_paper.pdf





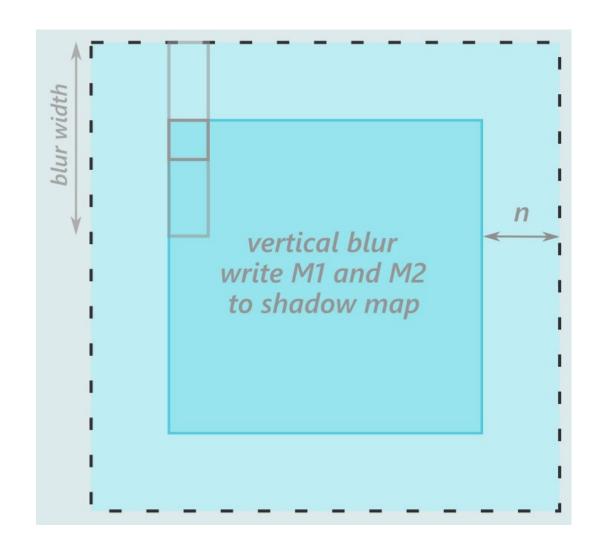
Vertical pass

Write moments M1 and M2 to the shadow map

```
vec2 M1M2 = vec2(0.0);
memoryBarrierShared();

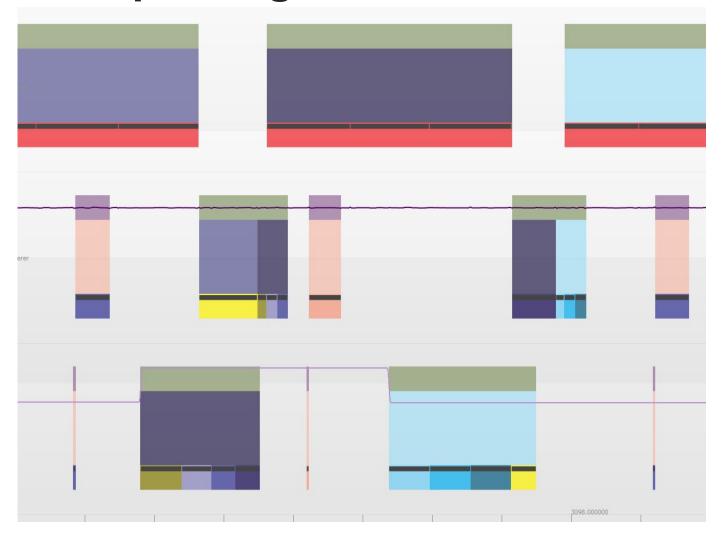
for (uint y = 0u; y < blur_width; y++) {
   uint shared_index = get_shared_index(y,
        gl_LocalInvocationID);
   vec2 M1M2_val = shared_storage[shared_index];
   M1M2 += M1M2_val * get_gaussian_factor(y);
}

imageStore(shadow_map, global_coord, vec4(M1M2,
   0.0f, 0.0f));</pre>
```





Task packing



Vulkan spec section 6.2:

"The initial order is determined by the order in which vkQueueSubmit commands are executed on the host, for a single queue, from first to last."



Task packing

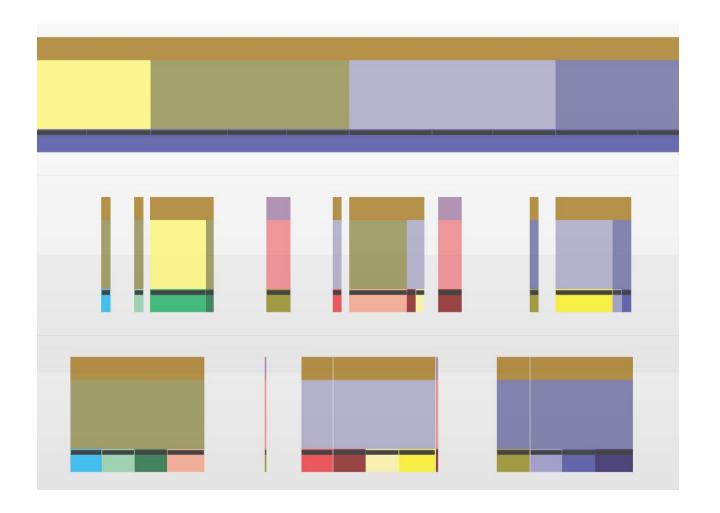
Alternate between two queues each frame for submitting commands.

Allows frame independent scheduling and increases task packing!

May require alternating between two sets of resources.

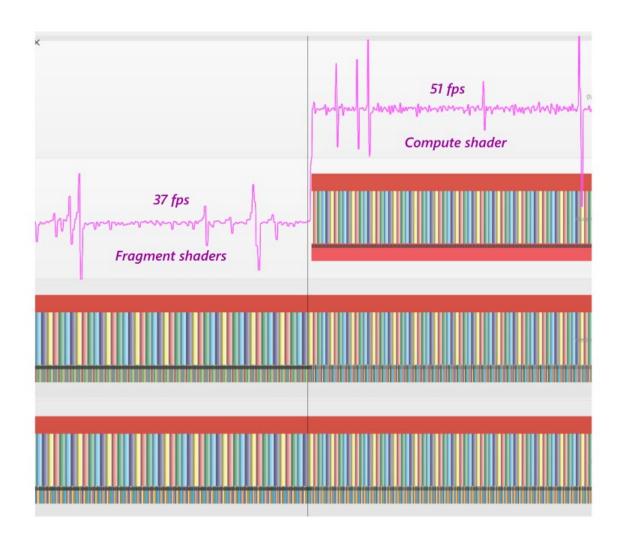
More developer recommendations!

https://docs.imgtec.com/





Compute improvements



40% fps increase for 5x5 blur!

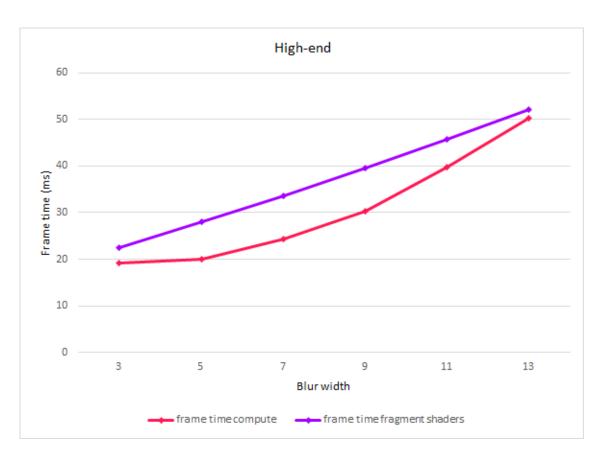
Meizu Pro 7 Plus - PowerVR 7XTP



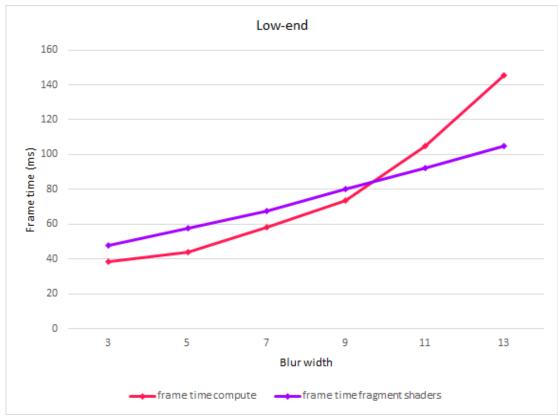


Compute improvements

Frame times for high-end PowerVR GPUs



Low-end PowerVR GPUs







Thank you!