DirectCompute Use in Real-Time Rendering Products

Chas. Boyd
Microsoft Windows Graphics

Overview

- Current games using DirectCompute for visuals
- Examples of game rendering techniques:
  - Screen-space ambient occlusion
  - Optical effects (lens flare, depth-of-field)
  - Lighting
- Future projections

Current Compute-Based Techniques

- Screen-Space Ambient Occlusion
  - BattleForge
  - Colin McRae Dirt 2
- Depth of Field Effect
  - Metro 2033
  - Just Cause 2
- Post Processing Optical Effects
  - FutureMark 3DMark 11

DirectCompute Use in Real-Time Rendering Products

AMBIENT OCCLUSION

Screen-Space Ambient Occlusion

- Conventional technique uses pixel shaders
  - http://sites.google.com/site/perumaal/ao.pdf

- DirectCompute shaders enable more control of convolution filter cache

HDSSAO Off

Image credit:
Codemasters

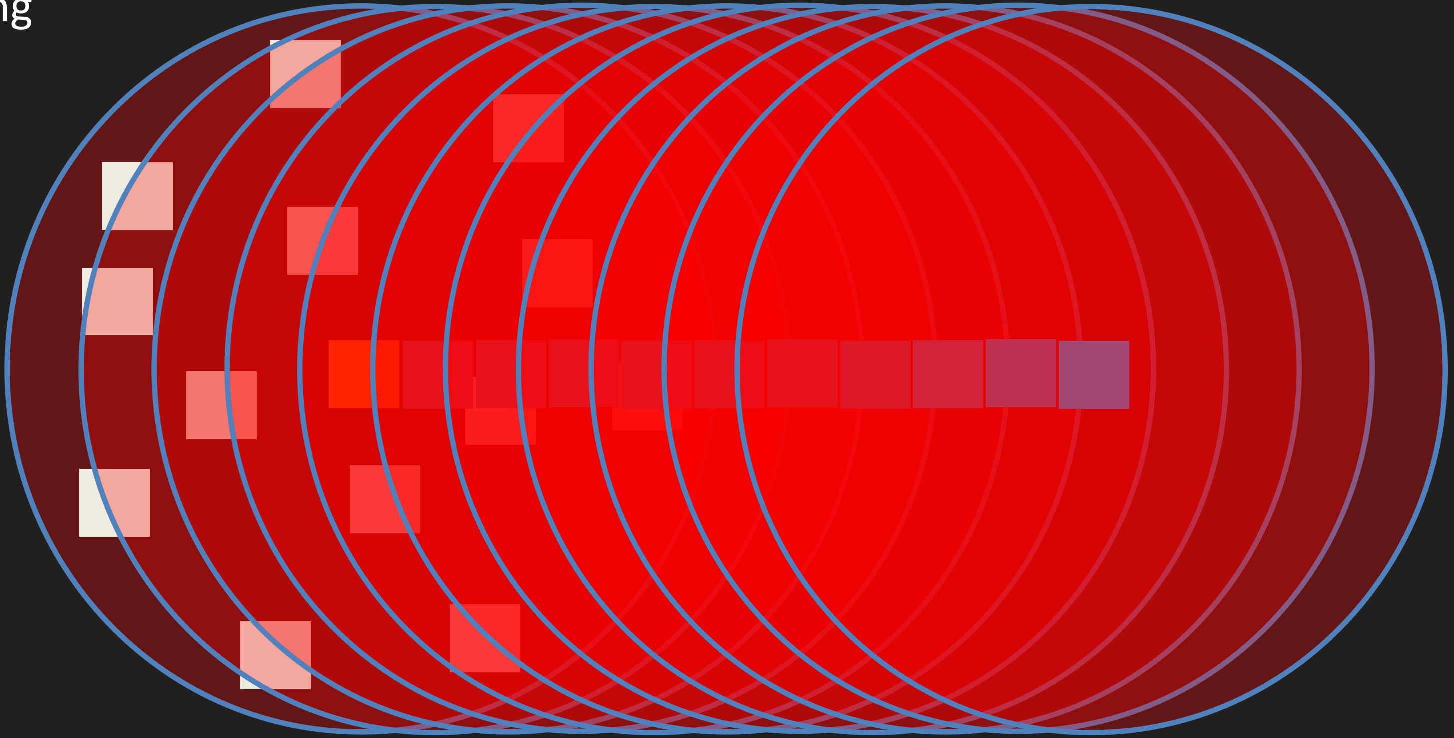
HDSSAO On

Image credit: Codemasters

Core Algorithm

- For each pixel in depth image
- Check surounding neighborhood to see if they form a concave region
  - Fit a cone, is it concave or convex
- Improved results if normals included in check

[illegible] a great deal of over sampling



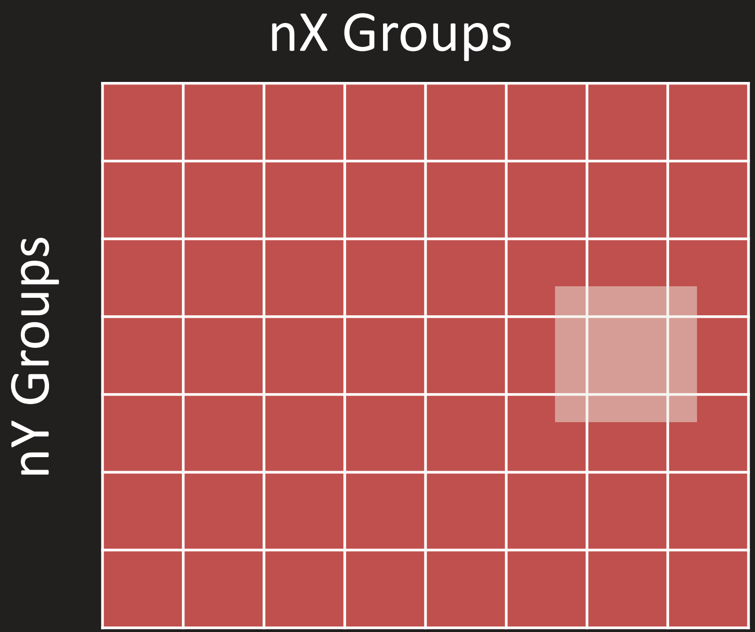
I/O Cost is large

- HDAO has an evenly distributed sparse sampling pattern
- The kernel size is 12 texels
- With 48 valleys the PS has to perform 193 samples
  - 1(Center)+(48(Depth)+48(Normals))*2(Valley)
- This gives rise to an ALU : TEX Ratio of 0.48

DirectCompute Alternative

- GPU has `groupshared` 'register' storage class
  - Corresponds to programmable cache
- Can store frequently read values
  - Beats performance of texture cache in this algorithm
- Aka "Local Data Store", "Local memory"

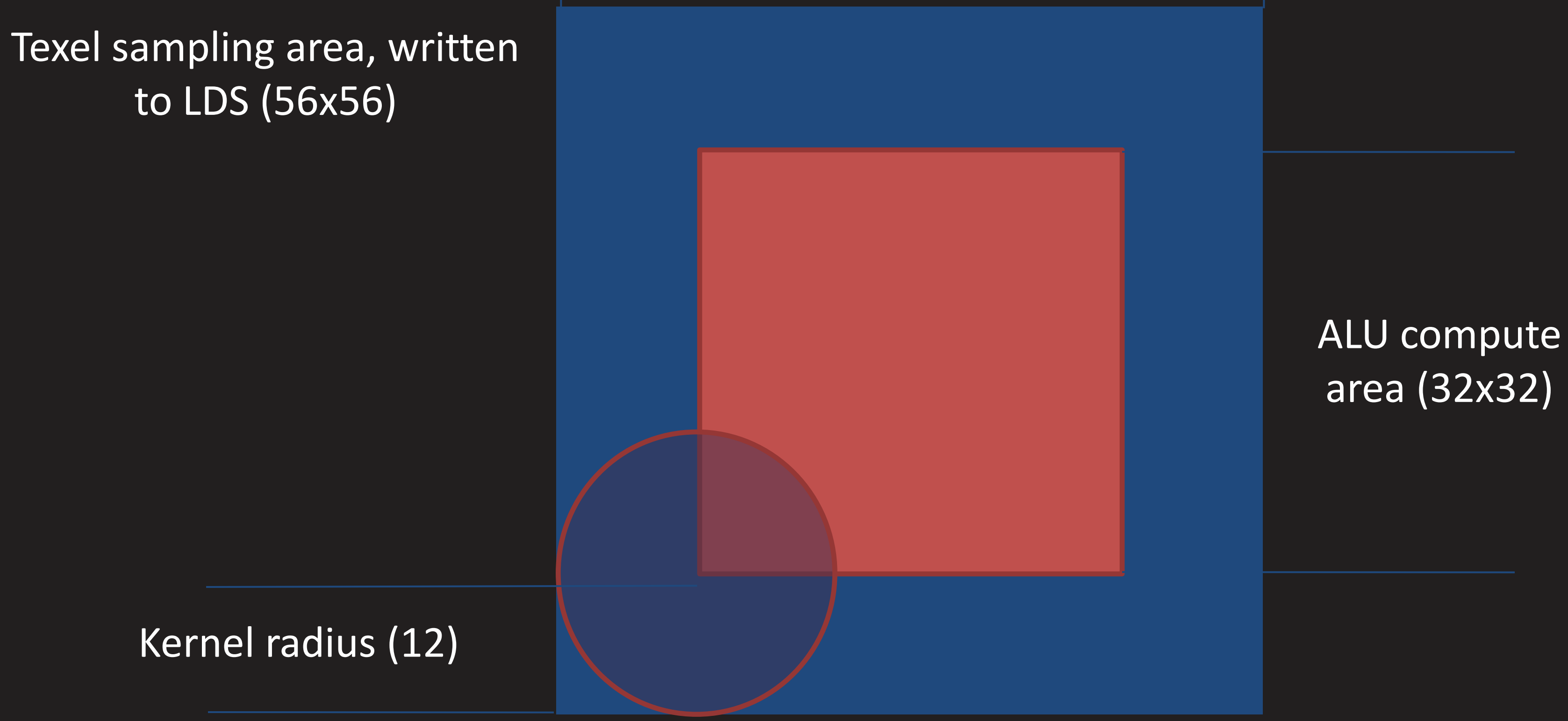
Structure Depth Buffer in Tiles



32x32 pixel tiles

(32x16 or 16x16 also work well)

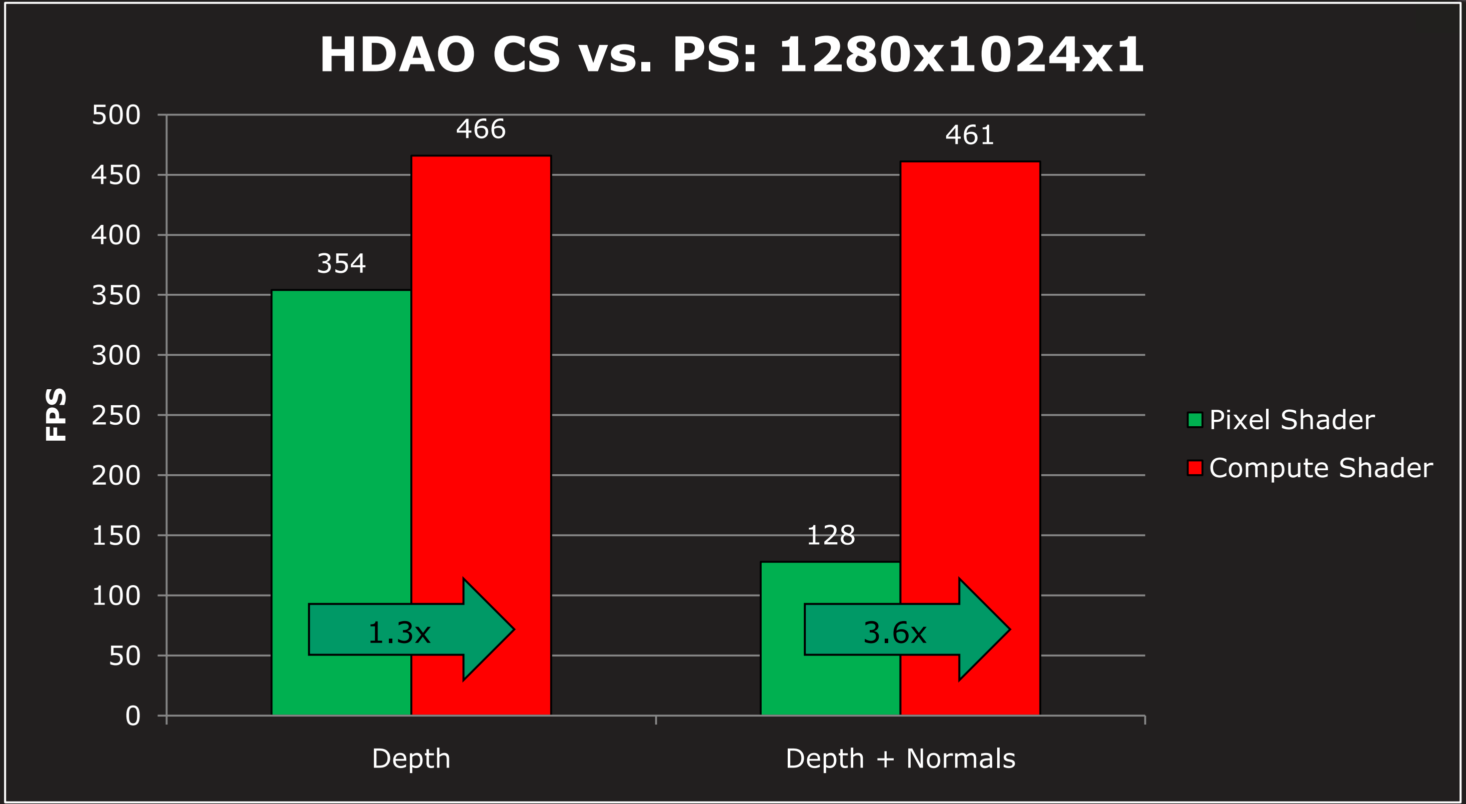
Tiles must overlap: Apron



Advantages

- Using the **`groupshared`** memory avoids an incredible amount of over-sampling
- It can be filled using the Gather instruction, which further reduces the number of TEX operations
- This gives rise to an ALU : TEX Ratio of 17.92
- The overlapping tile size is also critical to overall performance

Results


HDAO CS vs. PS: 1280x1024x1
FPS
500
450
400
350
300
250
200
150
100
50
0
466
461
354
128
1.3x
3.6x
Depth
Depth + Normals
Pixel Shader
Compute Shader

HDAO is not unique

- General Technique
  - Caching in **`groupshared`** memory is a performance win for many algorithms
- Aprons required for any kernel-based algorithm
  - Tile size with Apron is what needs to fit/align

DirectCompute Use in Real-Time Rendering Products

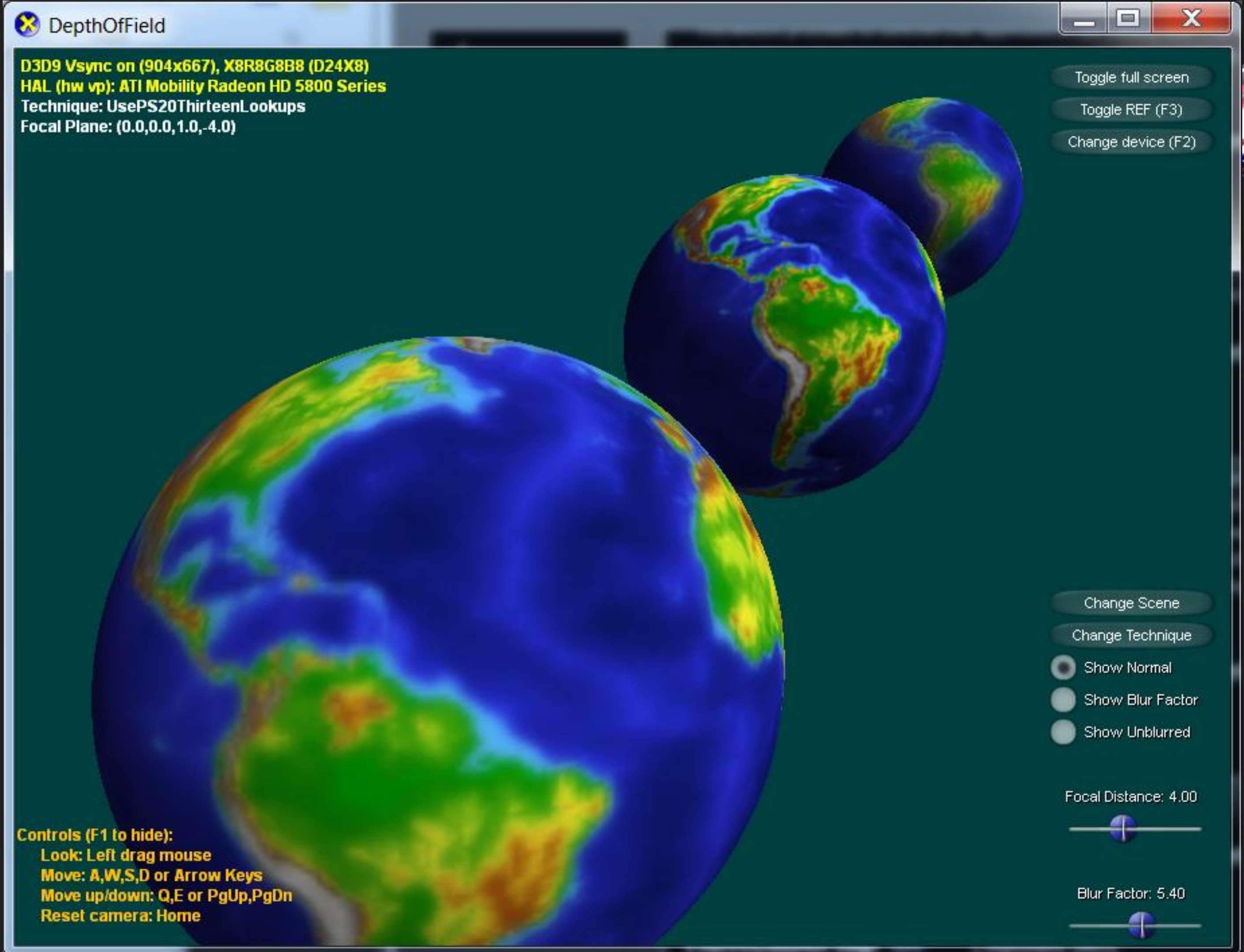
DEPTH OF FIELD

Depth of Field Effect (ala Metro 2033)

- Justification
  - Adds realism for a ‘you are there’ experience
  - Provides creator with a way to direct user’s attention (as heavily used in video/cinema)

Early Work

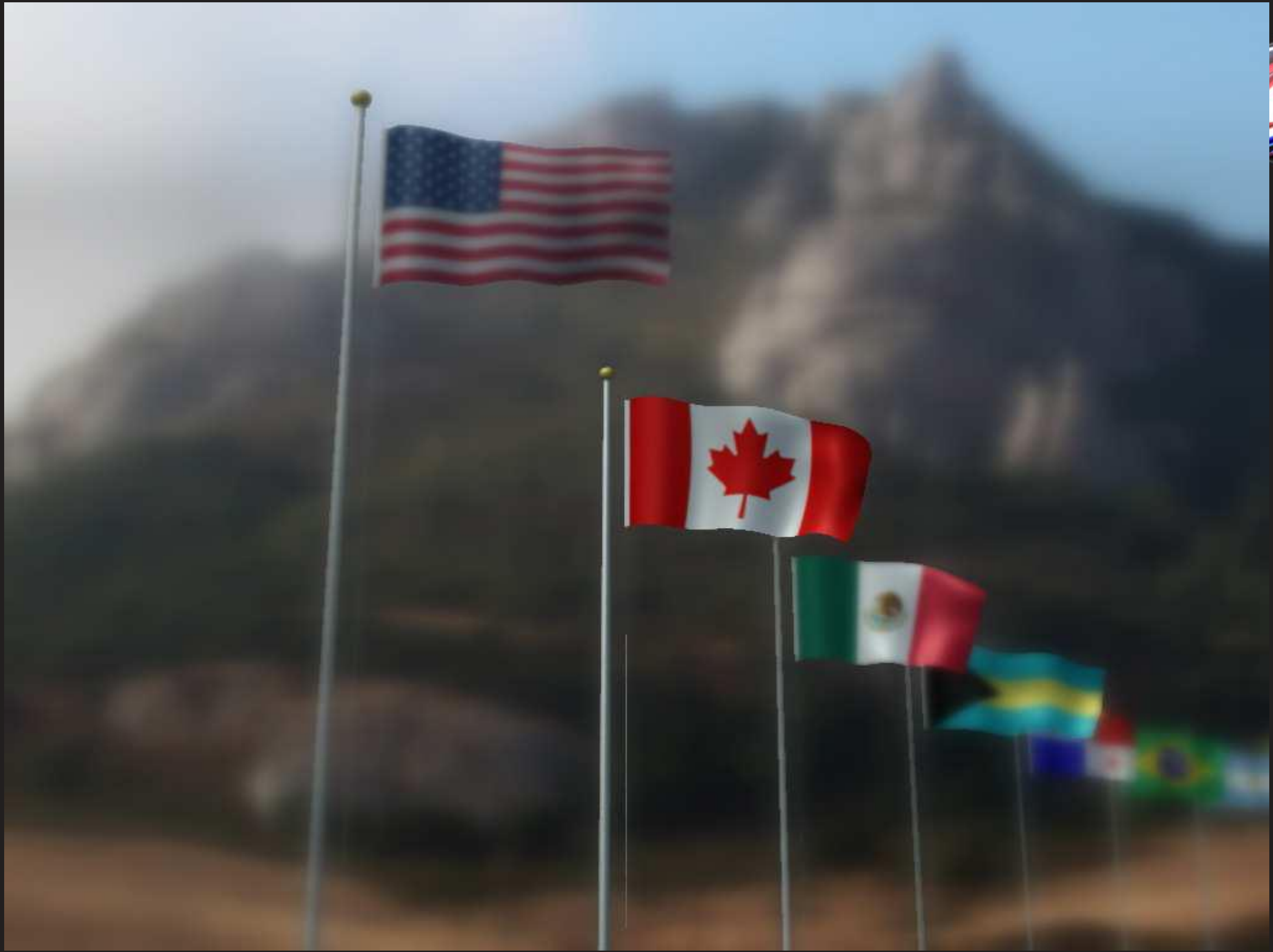
- Create a blurred (usually down-scaled) version of the frame
  - Blend between that and original (crisp) version based on distance from focal plane
- Issues:
  - Fixed blur radius across scene, no variation with depth
  - Artifacts along edges are very tricky to remove
    - Required significant logic for each kernel tap


DepthOfField
D3D9 Vsync on (904x667), X8R8G8B8 (D24X8)
HAL (hw vp): ATI Mobility Radeon HD 5800 Series
Technique: UsePS20ThirteenLookups
Focal Plane: (0.0,0.0,1.0,-4.0)
Toggle full screen
Toggle REF (F3)
Change device (F2)
Change Scene
Change Technique
Show Normal
Show Blur Factor
Show Unblurred
Focal Distance: 4.00
Blur Factor: 5.40
Controls (F1 to hide):
Look: Left drag mouse
Move: A,W,S,D or Arrow Keys
Move up/down: Q,E or PgUp,PgDn
Reset camera: Home

Sophisticated Solution

- Model the problem as a heat-flow simulation
  - Blur radius analogous to heat distance traveled
  - Controlled by ‘thermal conductivity’ of image, which is defined by distance from focal plane
- Original work at Pixar
  - http://graphics.pixar.com/library/DepthOfField/paper.pdf
  - Michael Kass, Aaron Lefohn, John Owens

Image Credit: Pixar
Kass, Lefohn, Owens



Implementation in Metro 2033

- Alternating Direction Implicit solver
  - Decomposes problem into X and Y directions
  - Finite Difference scheme results in tri-diagonal systems in each axis (row/column)

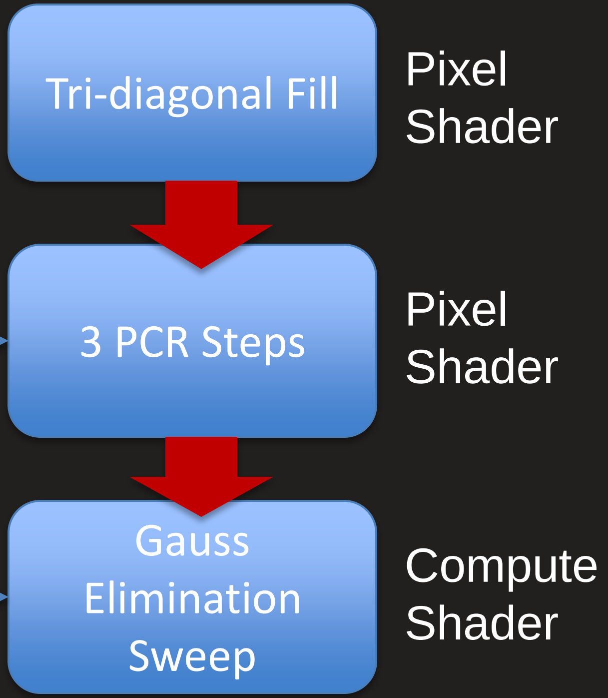
$$\begin{bmatrix} b_1 & c_1 & & & 0 \\ a_2 & b_2 & c_2 & & \\ & a_3 & b_3 & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ 0 & & & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_n \end{bmatrix}.$$

Hybrid Tri-diagonal Systems Solver

- 3 Steps of Parallel Cyclic Reduction
  - Each step reduces complexity of each system
  - And doubles the number of systems
  - Shifts problem to more data-parallel domain
- Gauss Elimination –using DirectCompute
  - Forward elimination -> Backward substitution
  - Complexity O(N)

Hybrid PCR/Gauss TriDiag Solver

Number of Systems	System Size
Height	Width
Height*8	Width/8



Metro 2033

- GDC 2010 Presentation by
  - OlesShishkovtsov, 4A Games
  - AshuRege, NVIDIA
- http://developer.nvidia.com/object/gdc-2010.html#metro

Image credit:
A4 Games

Advantages

- Blur size varies with distance to focal plane
  - Required for cinematic level of realism
- Higher image quality / fewer visible artifacts
  - Implicit techniques distribute errors more globally
    - Errors concentrated in one area == artifact
  - Accurate behavior near depth edges
    - Compared to earlier techniques

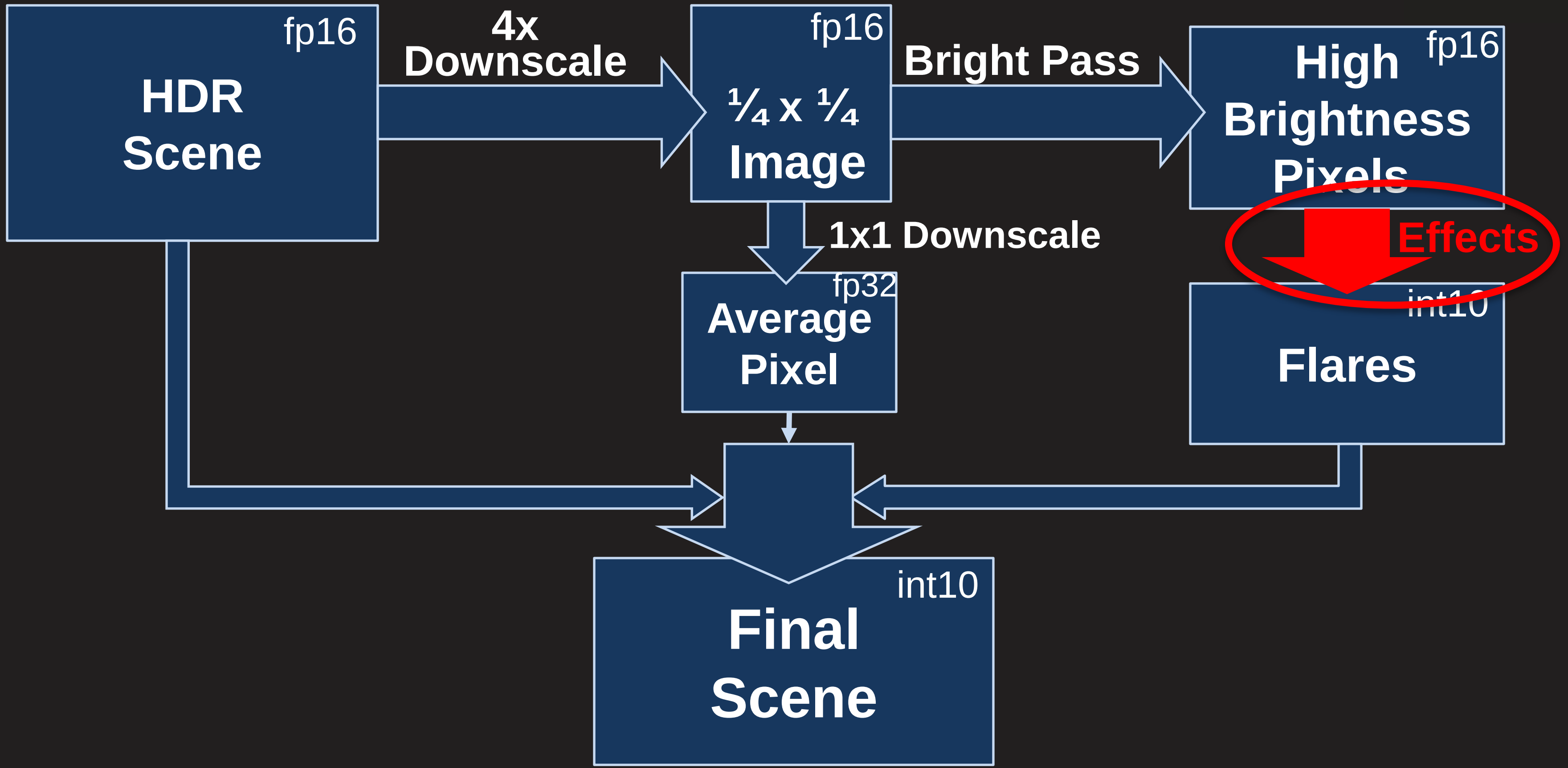
DirectCompute Use in Real-Time Rendering Products

FFT LENS EFFECTS

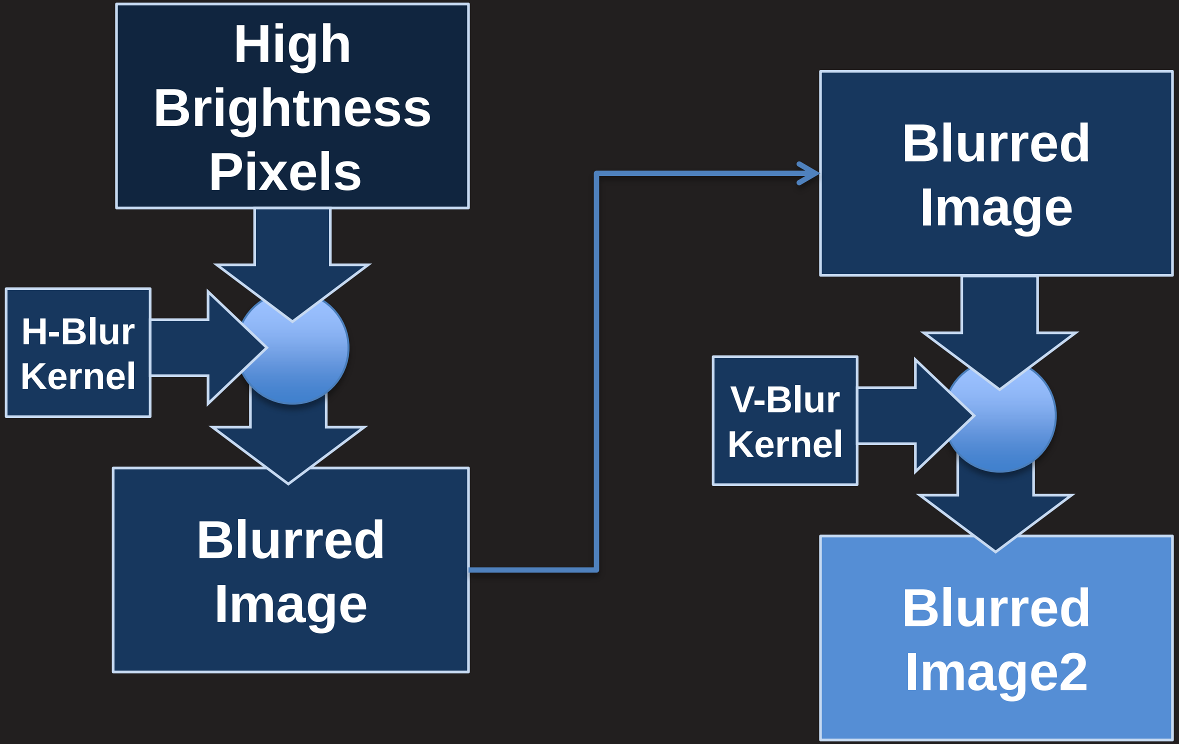
3DMark is not a game

- Futuremark produces 3DMark as a benchmark
  - Intended as a ‘stress test’ for DirectX11 hardware and drivers
- Designed to be as close to game as possible
  - but capable of simulating a very heavy workload
- Utilizes FFT implemented in compute shaders
  - For post-process lens effects

Post Processing Pipeline

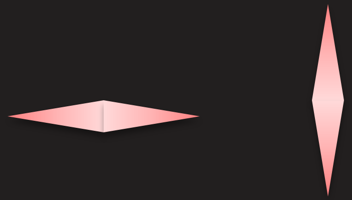


Conventional Post Processing Effects

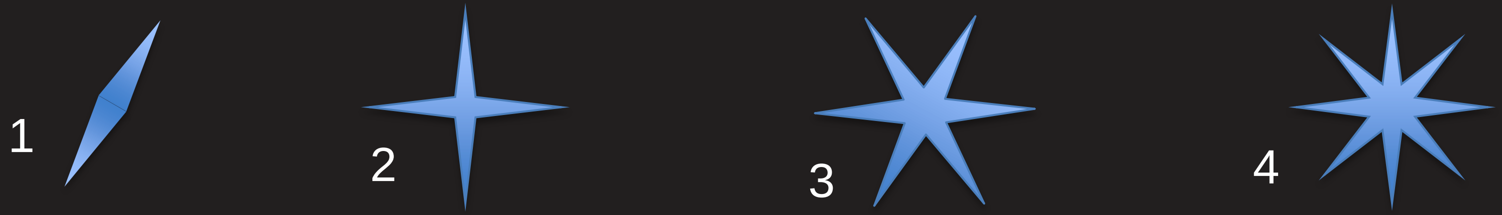


Post Effects Require Multiple Passes

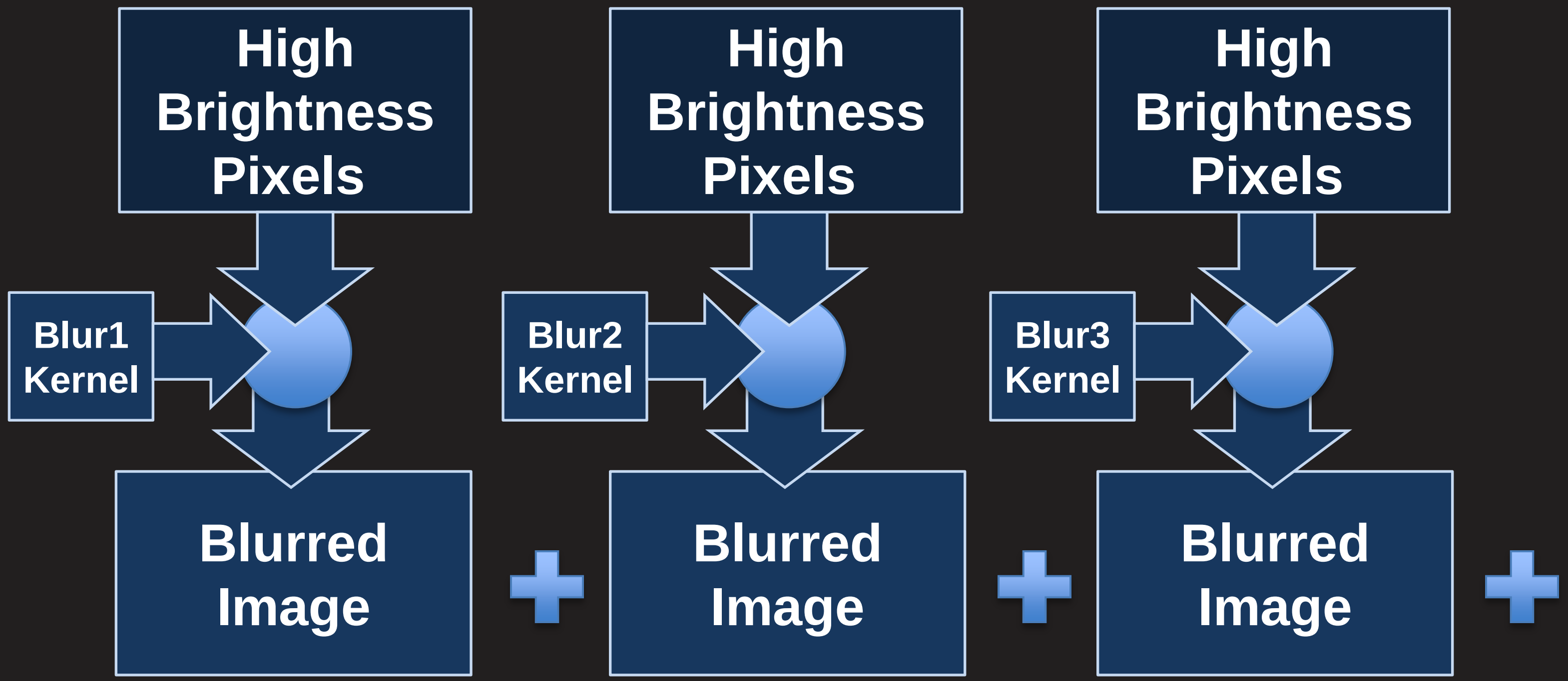
- Blur/Halation Effect
  - 2 passes of separable kernel (0.5-1ms)



- Lens Flare Effect
  - At least 1 pass (0.3ms or 3fps) per flare ‘point’



Conventional Post Processing Effects



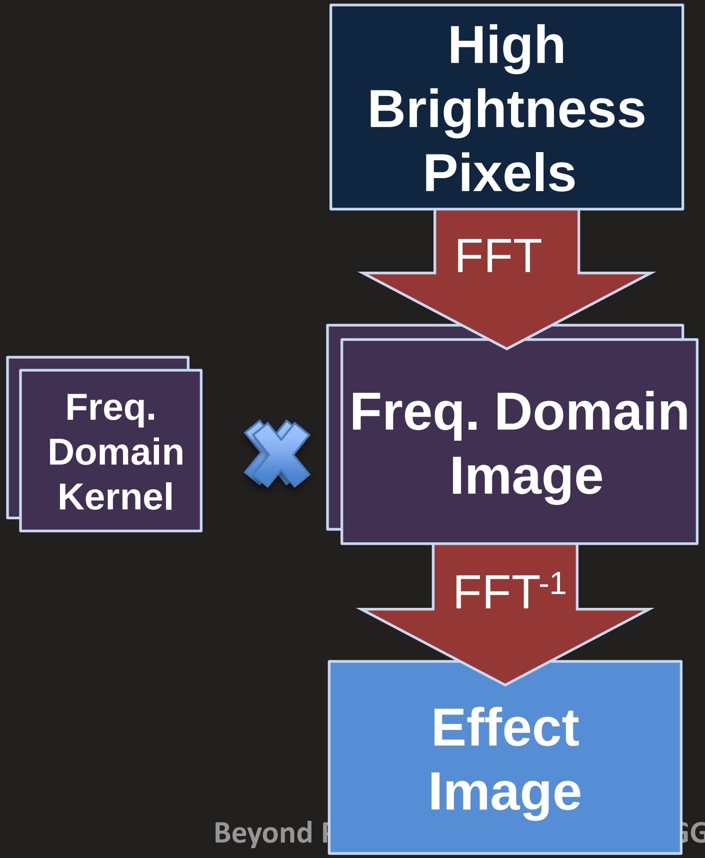
Kernel Size

- Entertainment apps must be resolution-independent
- Post effects must scale with screen resolution
  - Not a fixed n-tap kernel
  - Kernel size is proportional to image size
  - Number of Ops = $M*N*k = M*N*b*N = O(N^3)$
- May require more passes for larger screen sizes

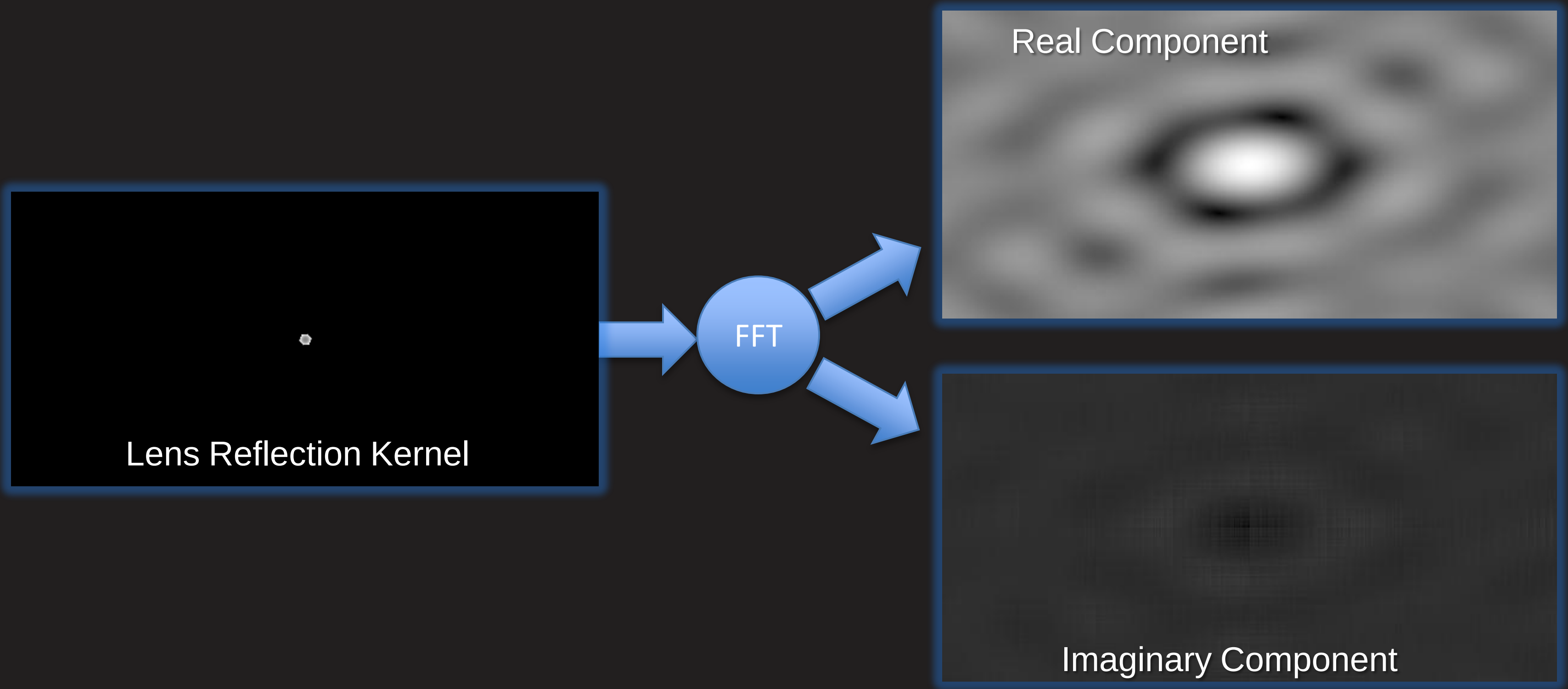
FFT Approach

- Compute interesting kernel shapes for glare/flare
  - Procedural algorithm to control nr of streaks
- Ideally HDR level of precision
  - Could paint in photoshop?
- Convert kernel into Frequency domain (re+im)
  - Requires 2 passes

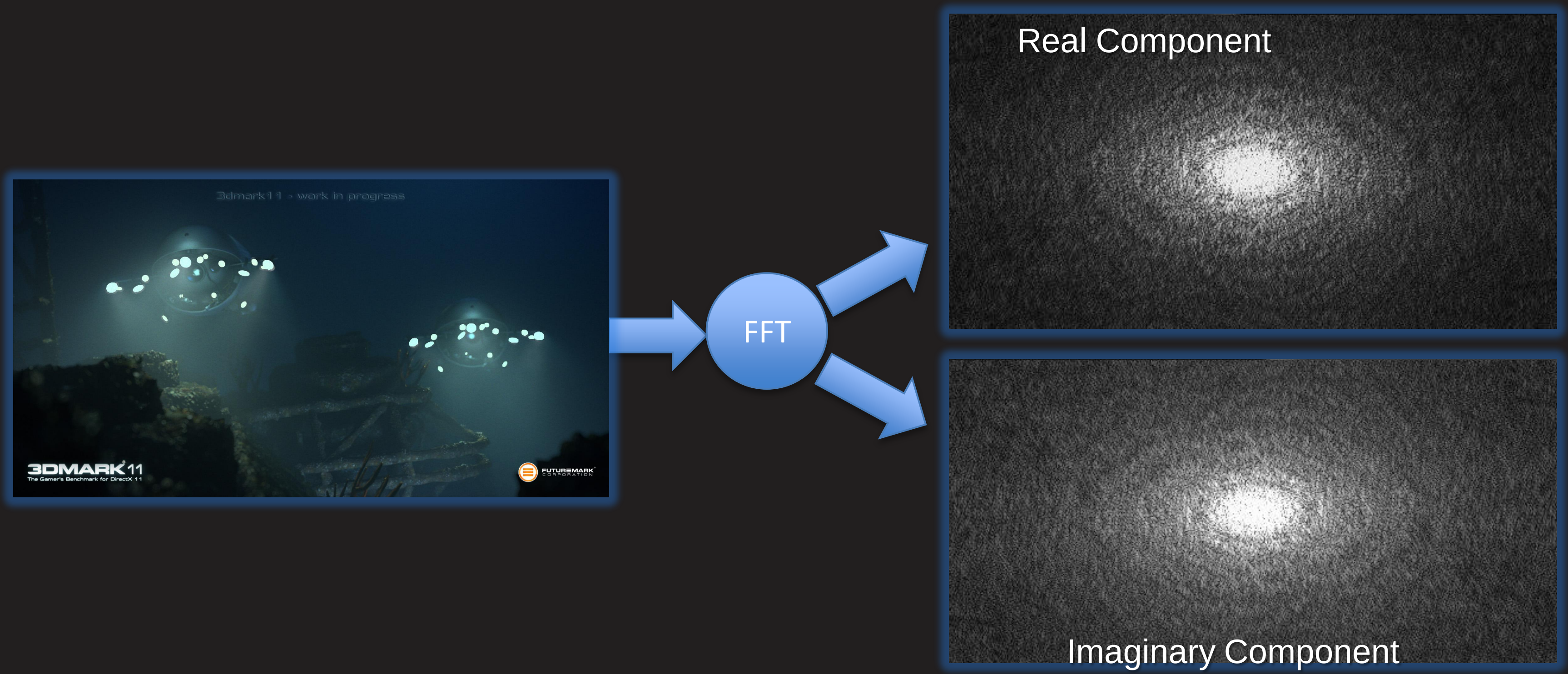
FFT Post Processing Effects



Lens Reflection Kernel



Convert Image into Frequency Domain



Complex Multiply

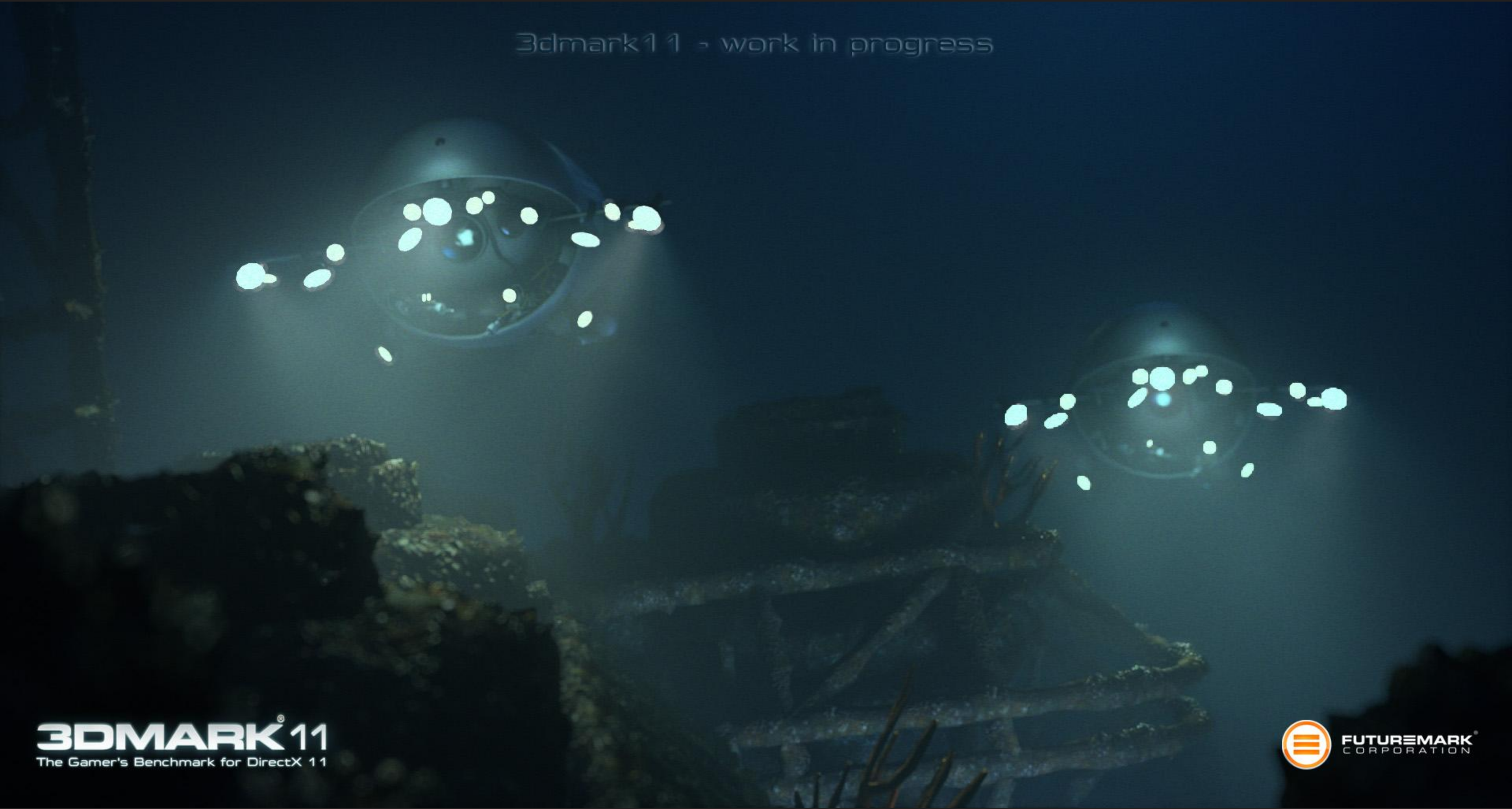
- Usual Process:
  - `Result.re = A.re*B.re – A.im*B.im`
  - `Result.im = A.im*B.re + A.re*B.im`
- This can be a single pass
  - Replaces n-passes of conventional technique

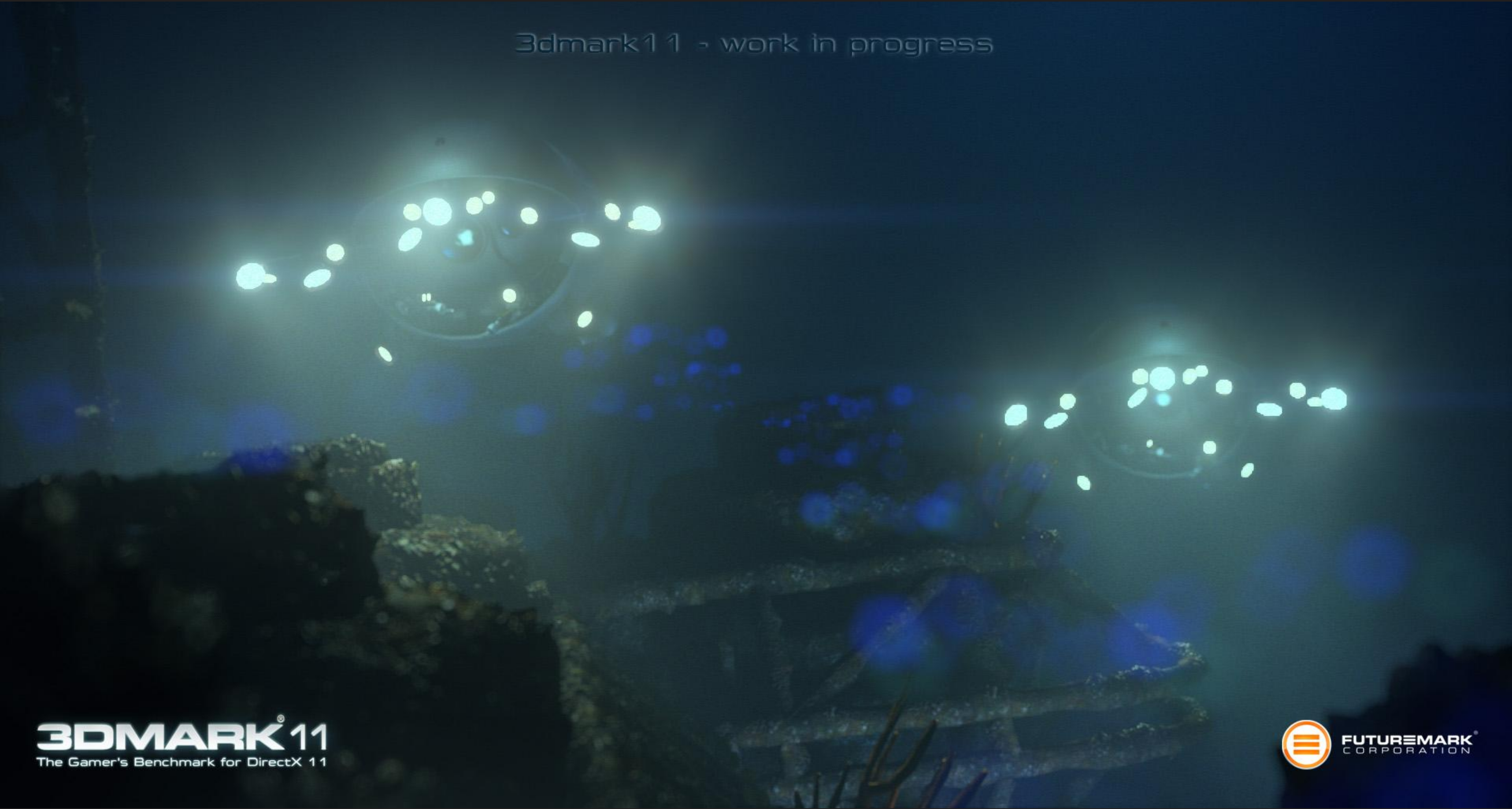
Inverse FFT

- Inverse FFT the result
  - Produces image of streaks
  - Requires 2 passes

- Add to original image


3dmark11 - work


3dmark11 - work in progress
3DMARK®11
The Gamer's Benchmark for DirectX 11
FUTUREMARK®
CORPORATION


3dmark11 - work in progress
3DMARK®11
The Gamer's Benchmark for DirectX 11
FUTUREMARK®
CORPORATION

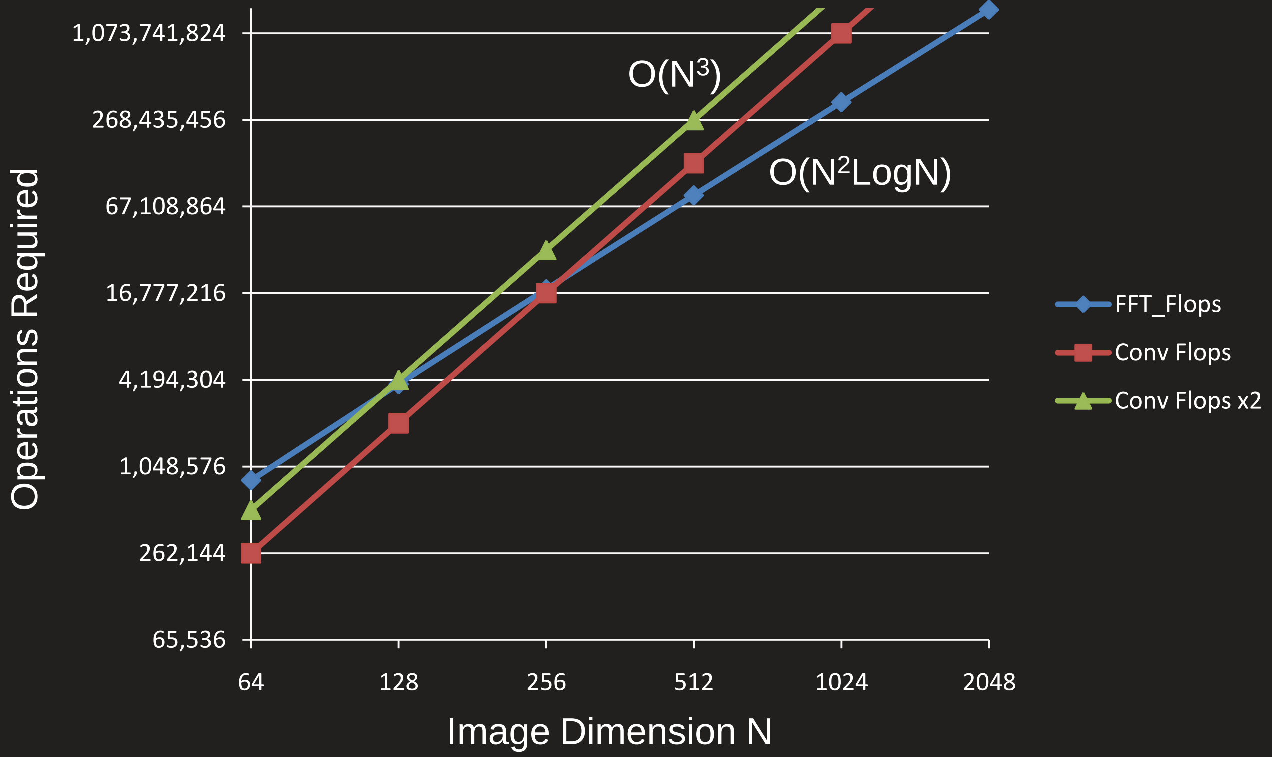
Pixel Shader Convolution Performance

- Op count grows as $O(N^3)$
- So does I/O bandwidth
  - Into texture unit at least,
  - Maybe not from memory

FFT Performance

- 1-D FFT Operations count
  - $N*Log_2(N)$
- 2-D FFT Operations count
  - $5 * (M*N*log_2(N) + N*M*log_2(M)) \sim O(N^2)Log_2(N)$
- 2-D FFT I/O count
  - 2 reads and writes per pixel ~O(M*N)

Performance (ops)



Conclusion

- FFT enables more post effects in fewer passes
  - May also enable new types of effects
  - Performance improves with effect complexity
- Challenges
  - 2 passes required for FFT in each direction
  - Pixel Shaders are faster for less complex effects

FFT Future Work

- Compute APIs currently target 32-bit float data
  - Reduced precision may be adequate for image uses
  - Consider float16 precision for intermediate surfaces
- Most current implementations assume scalars
  - Multiple color channels could be better optimized
- What other effects can benefit from freq domain?
  - SSAO? Depth-of-Field? Motion Blur?

Summary & Forecast

- Adoption of DirectCompute technologies is increasing steadily:
  - as new techniques are identified
  - new solutions to existing problems found
  - cross-ecosystem collaboration in operation
- Developers continuously encounter situations that require increased generality to implement

Acknowledgements

- Team Dirt 2, Codemasters
- Jonathan Story, AMD

- Oles Shishkovtsov, 4A Games
- Jim Black, Ashu Regde, nVidia

- Juha Sjoholm, Ilkka Koho, FutureMark