SIGGRAPH
ASIA 2019
BRISBANE

Staged Metaprogramming for Shader System Development

Kerry A. Seitz, Jr.,* Tim Foley,†
Serban D. Porumbescu,* and John D. Owens*

sa2019.siggraph.org

Trinity University
Computer Science Colloquium
Fall 2020

*UCDAVIS
UNIVERSITY OF CALIFORNIA

†

This talk is based on a talk that I presented at SIGGRAPH Asia 2019, hence the branding. It was originally a ~20-minute talk for experts in the field of computer graphics, so I have expanded it to include additional background information for a wider experienced audience.

If I think back to when I was a student at Trinity, if I had heard this title "Staged Metaprogramming for Shader System Development," I think my reaction would be…

You might be thinking:
“I know some of those words.”

(That would have been my reaction when I was a student at Trinity.)

“I know some of those words.”

So let’s start by breaking down the title to see what this talk will be about.

Programming Languages

Staged Metaprogramming for Shader System Development

Computer Graphics

This talk is partly a talk on programming languages. I'll discuss what "staged metaprogramming" is, and also cover what I mean by "metaprogramming" more generally.

But this talk is primary about computer graphics, and that's where this thing called a "shader system" come in. We'll cover what that is a little later, but first, I want to dive into the field of computer graphics in the context of this talk. It's a very wide field, so I'm going to focus on the aspects at which this work is aimed, but keep in mind that it could be useful in other areas as well.

What do I mean by “computer graphics?”

Broadly, it means: “Generating images with the aid of computers”

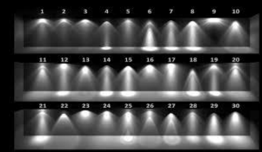
But specifically, I mean: “Using artist-authored content to generate images for animated movies, visual effects (VFX), video games, etc.”

Shape

Movement

Appearance

Illumination



Broadly, “computer graphics” means “generating images with the aid of computers.” This is extremely broad, and includes things like computational photography, image processing, computer vision, and many other things.

But in the context of this talk, what I mean by “computer graphics” is “using artist-authored content to generate images for things like animated movies, visual effects, and video games.”

By “artist-authored content,” I mean things like the 3D geometric models of objects (e.g., characters, buildings, furniture), movement or animations of those 3D models, the appearance of objects (Are they made of wood? Plastic? Metal? Are they reflective? Translucent? Is it skin? Or hair? Basically, what material is an object made of?), and finally illumination (What are the lights in a scene? How bright are they? Where are they located?).

All of this content is used together to generate the many images that make up a rendered scene.

I want to make one more distinction in our definition of computer graphics, and that’s to separate “real-time rendering” and “offline rendering.”

Real-Time Rendering vs. Offline Rendering

E.g., video games

~16 ms per frame
(= 60 frames per second)
or faster!

(The focus of this talk)

E.g., movies and visual effects

~3 hours per frame
(for example)

Real-time rendering refers to generating images (or frames) as quickly as possible in order to create a fluid interactive experience. The primary example of real-time rendering is video games, where you want the images displayed on screen to update quickly, for example, as you move the character around in the environment.

So in real-time rendering, performance is critical. Each new image must be generated in about 16 milliseconds (which corresponds to 60 frames per second). Maybe you're willing to go as high as ~33 milliseconds per frame (or 30 FPS), but any longer and the game wouldn't feel very fluid. Or maybe you want to generate frames even faster! This is especially important for virtual reality (VR) using head-mounted displayed. Because the screens are so close to your eyes, framerates of 90 FPS, 120 FPS, or even higher become more important.

This is in contrast to offline rendering, which is commonly used in movies and visual effects. In those applications, you don't need to instantly update the images displayed on screen, because users aren't interacting with the scene. Instead, artists create the scenes, and then ship off the rendering work to generate images to giant clusters of computers in a render farms, which might spend up to 3 hours (for example) rendering each frame of a movie. They can afford so much time per frame, because they are only going to generate that frame once, which will then be viewed by many people each time the movie is watched.

This work I'm talking about today focuses specifically on real-time rendering. However, the ideas and concepts can be useful in the offline rendering world too.

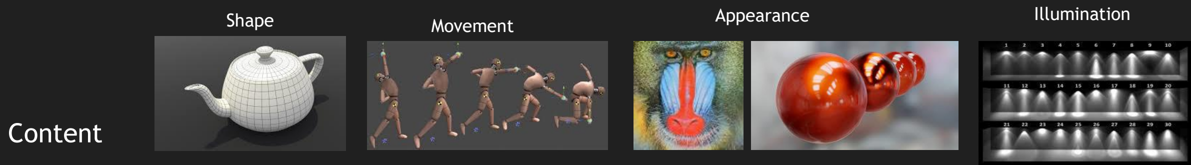
So now that we have our definition of computer graphics for this talk, how do we take artist-authored content and use it to generate images in real time?

Extra context for interested readers:

As an aside, between real-time and offline rendering is "interactive" rendering, where rendering a frame might take a few seconds or 1-10 FPS. So not quick enough to be "real-time." This is useful, for example, for movie artists to get a feel for how a scene will look and make adjustments along the way without spending 3 hours for a final render.

Content and Platform

Content



Platform (Hardware)

CPU: Threads | SIMD

GPU: VS | HS | DS | GS | PS | Compute

6

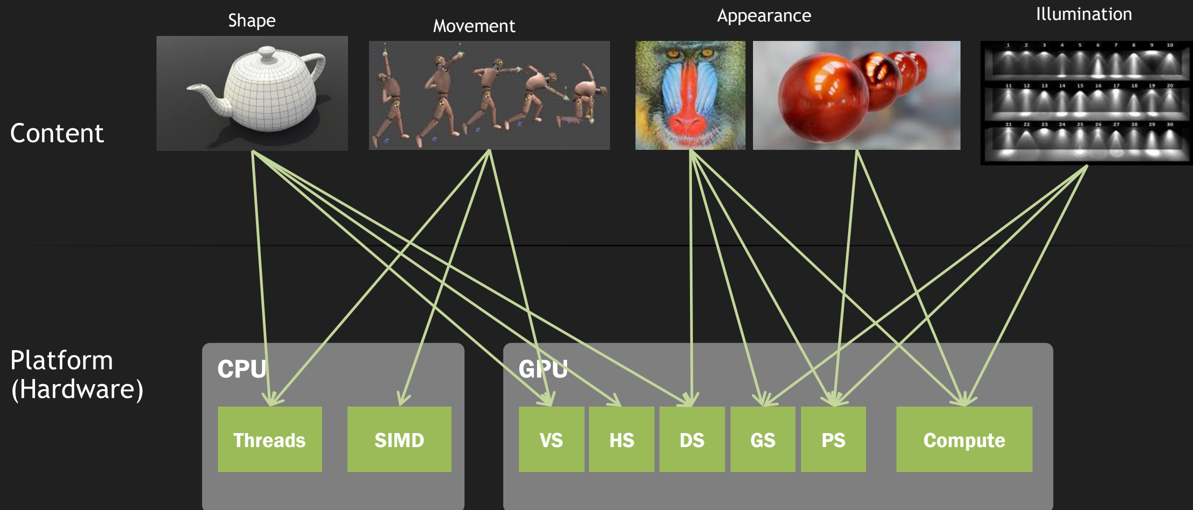
From “A Modern Programming Language for Real-Time Graphics: What is Needed?”
by Tim Foley (NVIDIA Research), Open Problems in Real-Time Rendering, SIGGRAPH 2016

(I’m borrowing these next few slides from a talk that my coauthor Tim Foley gave at SIGGRAPH a few years ago. Thanks Tim!)

That content must be mapped from the artist-authored concepts to the hardware platform that’s going to produce the images.

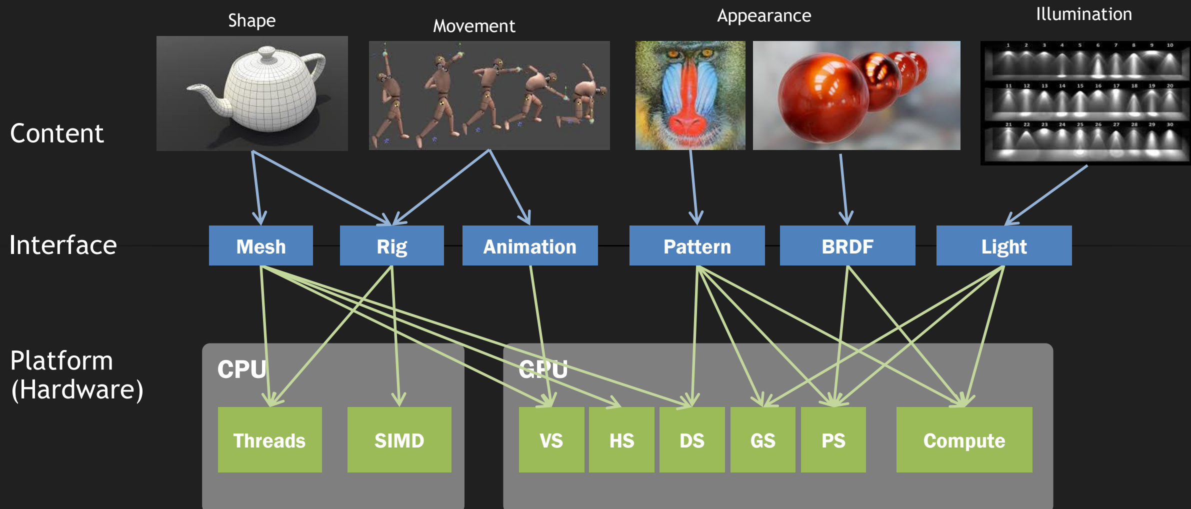
In real-time rendering, typically the hardware platform we’re talking about is CPU cores along with specialized GPU (or Graphics Processing Unit) hardware.

In order to get the stuff an artists produces up on the screen, we need to have a plan for how to map the concepts in the top row to the hardware platform on the bottom.



One (bad) way we could do that is by writing code on a per-asset, per-platform basis.

This obviously wouldn't scale, and so it isn't what we actually do.



What I claim we actually do in practice is this:

A graphics programmer defines an *interface* - a set of concepts that are specific to a particular engine or production.
This is how we will express an animation rig.
This is how we will represent materials (using reflectance functions).

Some of the representations in that interface might be pure data (e.g., for a mesh), and others might include code (e.g., if artists describe pattern generation as a "noodle graph").

Either way, the task is to:

- map the assets produced by artists so that they conform to the chosen concepts in the interface, and
- map those concepts efficiently to one or more target platforms

This thing I've referred to as a "shader system" helps to facilitate this process, helping graphics programmers define interfaces that enable artists to create the content for a game.

But before we get into what a "shader system" is, let's look at how we use a GPU to take this artist-authored content and generate an image from it using "shaders."

OpenGL / Direct3D Rasterization Pipeline (simplified)

OpenGL / Direct3D (part of DirectX) – APIs for 3D graphics

- They enable programmers utilize the GPU to achieve hardware-accelerated rendering

GPU Hardware Pipeline

- Series of stages executed in order
- Transforms data and performs calculations to generate final image

(Tons of handwaving and simplifying on the next few slides)

OpenGL and Direct3D (which is part of Microsoft’s DirectX framework) are APIs for 3D graphics that are used to interact with the GPU to achieve hardware-accelerated rendering.

GPU hardware has a series of stages that define a pipeline that I’ll refer to as the “OpenGL / Direct3D pipeline.” When rendering an object, these stages are executed sequentially to transform the data and perform the calculations necessary to generate the final image.

I’m handwaving heavily here, but let’s look at a simplified example to give you some idea about what the processes is.

OpenGL / Direct3D Rasterization Pipeline (simplified)

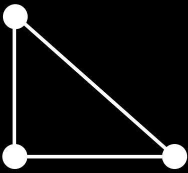
Let's say we want to draw this teapot, and we want to put it somewhere in a scene with other objects. In real-time rendering, geometric models are usually made up of a bunch of triangles. Let's follow one triangle of this teapot model through the pipeline to generate the final pixels we'll see on screen for that triangle.

(Again, I'm handwaving and over simplifying here in the interest of time.)

OpenGL / Direct3D Rasterization Pipeline (simplified)

GPU

Vertex Shade



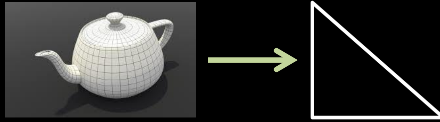
The first step in the OGL / D3D pipeline is the Vertex Shade stage, which runs a piece of code for each vertex of the triangle to perform some calculations.

Extra context for interested readers:

For example, generally this is where vertex coordinates transformed to their locations within the scene. (For example, a vertex that sits at the origin relative to the rest of the teapot probably doesn't sit at the origin of the scene, especially if the teapot moves around.)

Now that we know where the vertices are in the 3D scene, we need to determine where they will end up on the final 2D image and generate the appropriate number of pixels for that triangle.

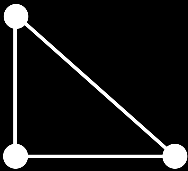
OpenGL / Direct3D Rasterization Pipeline (simplified)

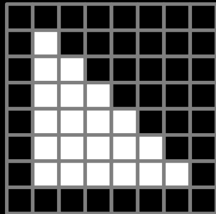


GPU

Vertex Shade

Rasterize





Then, the Rasterize stage chops up a triangle into pixels. Once you have those pixels…

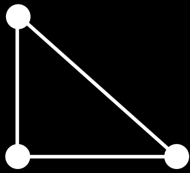
Extra context for interested readers:

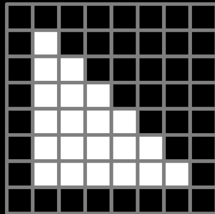
While the Vertex Shade stage could execute user-written code to perform calculations, the Rasterizer is a fixed-function piece of hardware. Its only job is to turn triangles into pixels, and it's extremely efficient at it.

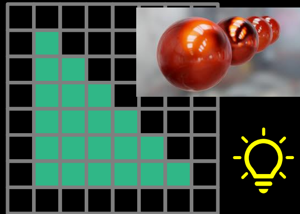
OpenGL / Direct3D Rasterization Pipeline (simplified)

GPU

Vertex Shade → Rasterize → Pixel Shade







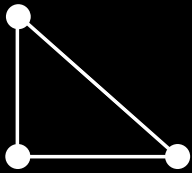
(Once you have those pixels…)

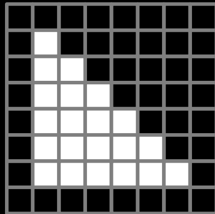
The next stage is the Pixel Shade stage. This stages run a piece of code on each pixel of the triangle to determine the color of that pixel, taking into account the artist-authored materials as well as the various lights in the scene.

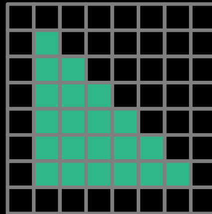
OpenGL / Direct3D Rasterization Pipeline (simplified)

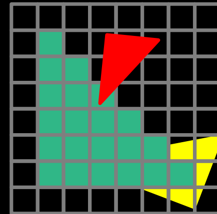
GPU

Vertex Shade → Rasterize → Pixel Shade → Depth Test







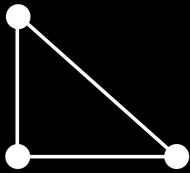


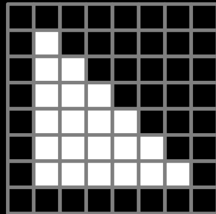
Then comes the Depth Test stage, where each pixel of the triangle is compared to the pixels from other triangles that overlap it. The pixels closer to the camera "win," and farther away pixels are discarded.

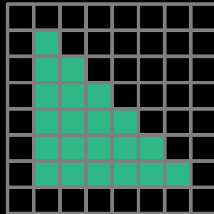
OpenGL / Direct3D Rasterization Pipeline (simplified)

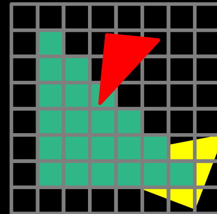
GPU

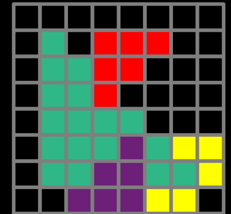
Vertex Shade → Rasterize → Pixel Shade → Depth Test → Composite







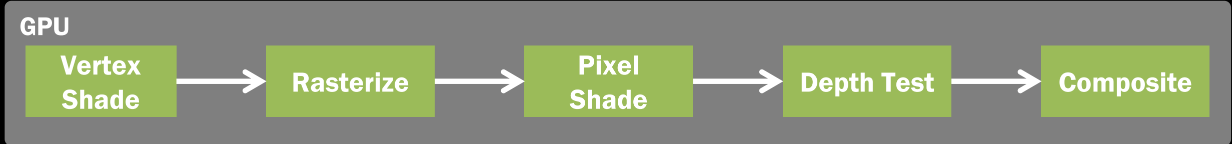




Finally, we have Composite, which is where the non-discarded pixels of the triangle are written to the framebuffer, which will be displayed on screen once all triangles for this frame have been processed.

This is the basic OpenGL / Direct3D rasterization pipeline.

OpenGL / Direct3D Rasterization Pipeline (simplified)

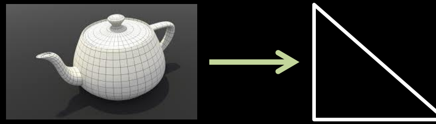


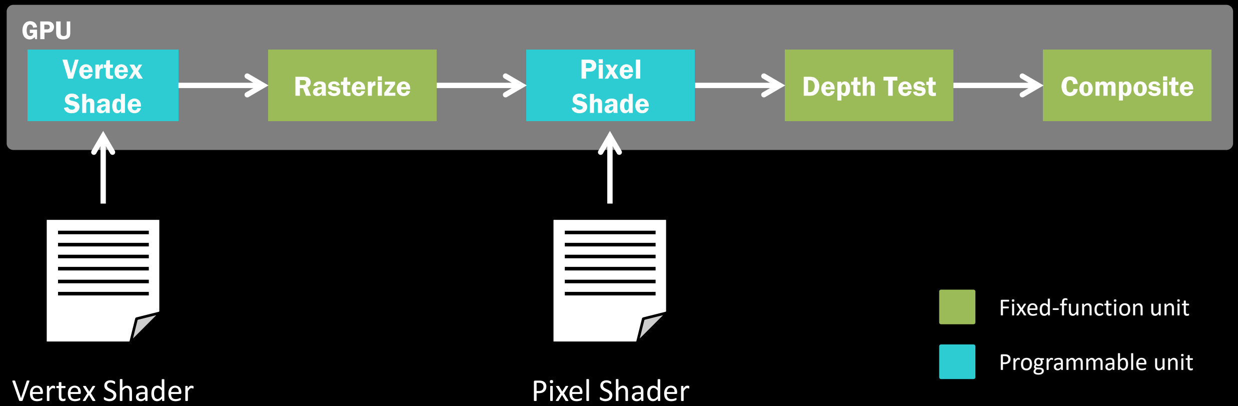
Fixed-function unit

In early GPU hardware, all of these stages were fixed-function pieces of hardware. They did have various degrees of user-configurable parameters to control their behavior, but they couldn't execute arbitrary pieces of code.

(But in modern GPUs…)

OpenGL / Direct3D Rasterization Pipeline (simplified)





But in modern GPUs, some stages are now programmable. Both the vertex and pixel shade stages allow you to run user-written code to perform calculations per-vertex and per-pixel, respectively. This gives graphics programmers much more flexibility to implement interesting techniques, while still achieving a high degree of performance to meet that 16 millisecond-per-image time budget.

The user-written code that run on the GPU during pipeline execution are called “shaders.”

Shaders

(In the context of this talk)

Code that executes on the GPU as part of the rendering process

- E.g., vertex shader, pixel shader, compute shader, and others

Shaders are written in HLSL (Direct3D) or GLSL (OpenGL)

- These are "shading languages"
- C-like languages
- Very simple / feature-poor – no object orientation, no templates/generics*

* Slang [He et al. 2018] adds generics and some other modern features to HLSL

The term "shader" is a bit overloaded, so I'm going to discuss it in the context of this talk.

A shader is a piece of code that executes on the GPU as part of the rendering process. As we saw on the previous slide, there are Vertex Shaders which run on a per-vertex basis and Pixel Shaders which run on a per-pixel basis. There are also Compute Shaders use a programming model similar to OpenCL and CUDA (if you're familiar with those).

Shaders are written in a "shading language" – either HLSL if for applications using Direct3D or GLSL for OpenGL.

HLSL and GLSL are C-like languages; they are very simple and low-level. They don't have many higher-level features that you might be accustomed to working with, such as object orientation or templates / generics. (As a side-note, I wanted to mention that Slang, which is a new shading language that my co-author Tim helped create, adds some more-modern programming language features to HLSL, including generics.)

Pixel Shader Example

HLSL shader code (GPU)

```
cbuffer LightData : register(b0) {
  float3 lightDirection;
};

// Object / Material Data goes here

float4 main() {
  return color * max(0, dot(normal,
    lightDirection));
}
```

Declare variables related to the light

Don't worry about all the details. I just want to highlight a few things.

First, we have declared some data related to the light. The float3 datatype is just three single-precision floating point numbers packed together. Similarly, the float4 returned by the function is four single-precision floats packed together.

I've omitted the object and material data, for the sake of brevity.

Extra context for interested readers:

The light data is declared in a "constant buffer," which basically means that this data is going to be the same for every pixel that we're currently shading, since the light doesn't change per pixel.

Pixel Shader Example

HLSL shader code (GPU)

```
cbuffer LightData : register(b0) {
  float3 lightDirection;
};

// Object / Material Data goes here

float4 main() {
  return color * max(0, dot(normal,
    lightDirection));
}
```

Compute the pixel's color

Then, we use the light data, along with the object and material data, to return a final color for whichever pixel we're shading.

In this case, the shader is pretty simple, but don't worry about the details of the math.

Extra context for interested readers:

Very very basically, in computer graphics, we treat typically triangles as if they are one-sided. If the light is on the back side of the triangle, then the dot product of the normal vector and the light direction will be negative, so we'll return 0 for the color of this pixel (because it's not receiving illumination from this light on its proper side). Otherwise, we'll modulate the object's color by how directly the light is shining on the surface and return the result.

Pixel Shader Example

HLSL shader code (GPU)

```
cbuffer LightData : register(b0) {
  float3 lightDirection;
};

// Object / Material Data goes here

float4 main() {
  return color * max(0, dot(normal,
    lightDirection));
}
```

C++ code to set up shader (CPU)

```
// Lots of gross code that I'm not going to
// show, except for:

MyLightData lightDataBuffer;
// Fill in light data
dxContext->PSSetConstantBuffer(0, 1,
  &lightDataBuffer);
```

Now, here's the corresponding CPU-side code that sets up this shader for execution as part of the graphics pipeline on the GPU. Well, ok actually I'm not showing most of it. It's a lot of complex and gross code, but I do want to show a glimpse of how you communicate data from CPU to GPU.

Somewhere, the CPU-side representation of our scene has information about the light that we need to communicate to the GPU shader code. In this case, we copy this data over to the GPU using the "PSSetConstantBuffer" function. There are two important things I want to highlight about this code.

Pixel Shader Example

HLSL shader code (GPU)

```
cbuffer LightData : register(b0) {
  float3 lig tDirection;
};

// Object / Material D    goes here

float4 main() {
  return color * max(0, do
    lightDirection));
}
```

C++ code to set up shader (CPU)

```
// Lots of gross code that I'm not going to
// show, except for:

MyLightData lightDataBuffer;
//  ill in light data
 Context->PSSetConstantBuffer(0, 1,
  &lightDataBuffer);
```

Completely different data structure on CPU and GPU!

First, the data structure we're using on the CPU is completely distinct from the data structure on the GPU. Neither HLSL nor the Direct3D API perform any checks to ensure that the layouts of these data structures match (GLSL / OpenGL aren't any better). This is a potential source for bugs!

Pixel Shader Example

HLSL shader code (GPU)

```
cbuffer LightData : register(b0) {
  float3 lightDirection;
};

// Object / Material Data goes

float4 main() {
  return color * max(0, do
    lightDirection));
}
```

C++ code to set up shader (CPU)

```
// Lots of gross code that I'm not going to
// show, except for:

MyLightData lightDataBuffer;
// Fill in light data
dxContext->PSSetConstantBuffer(0, 1,
  &lightDataBuffer);
```

Magic number! I hope they match in both CPU and GPU code!

23 Trinity University Computer Science Colloquium
Fall 2020 12 October 2020

Second, when we transfer the data from CPU to GPU, we tell the GPU to look for the data at a specific index. Here, I've declared the data to be at index 0 (using the `register(b0)` code). In the CPU code, I must then use that same index in the PSSetConstantBuffer function. Again, HLSL and Direct3D don't really help me here. If I get it wrong, then I'll have a bug in my code that might be hard to track down! (Again, GLSL/OpenGL aren't any better.)

Now, to be fair, graphics developers rarely write this kind of code directly. When you're first learning OpenGL or Direct3D, you might, but when you're developing a larger-scale graphics application, you'll probably develop tools to help with this or make use of tools someone else has written (like a major game engine like Unity or Unreal Engine). And that's exactly where this talk is headed. But there's one more bit of background information I need to cover first, and that's this thing called "specialization."

Shader Specialization

Three materials

- Standard Material
- Skin Material
- Cloth Material

Lots of other complex code to run along with material-specific code

- Need to slot in the material-specific code
- (We don't want to copy-paste the complex code for each material)

Remember: HLSL/GLSL don't have the features of higher-level languages you might be used to!

Let's say that we have three different materials that need to use different pieces of code to calculate their color. We have a standard material, that we use for most object. But we also have characters in our game that have skin. In order to make skin look good, we have to use special code that takes into account the properties of how light interacts with skin. We also have a cloth material that has special code too.

Along with the material-specific shader code, we also have a bunch of other complex code that we need to run in the shader in order to calculate the pixel's color. So we need to be able to slot in the material-specific code, at the appropriate place in the shader. Alternatively, we could copy-paste the other code to create three shaders – one for each material. But that would be a maintenance nightmare every time we need to change something.

If you're thinking how you might do this in a programming language you've been using, remember: HLSL and GLSL don't have many of the features you're used to using (both for historical reasons, but also for performance reasons).

Shader Specialization

Unspecialized Material Shader

```
float4 surfaceShader(…) {
…
if(isStandardMaterial(materialID)
  color = evalStandardMaterial(…);
else if(isSkinMaterial(materialID)
  color = evalSkinMaterial(…);
else if(isClothMaterial(materialID)
  color = evalClothMaterial(…);

…
}
```

One way we could do this is with a dynamic branch. At runtime in our shader, whenever we need to get an object's color, we look at the material ID of the triangle we're shading and choose which function to run accordingly. This totally works! We do have to manually maintain material IDs now, but that's not too hard.

But it turns out, this solution can be detrimental to performance! The code for this skin and cloth materials are much more complex than the code for the standard material. When you're shading cloth, for example, you of course have to pay the full runtime cost to run that code. But actually, even if you never execute that code path, writing the code this way means that this shader will still incur some of the resource allocation costs of the cloth code.

So, even if we're using this shader to shade an object that is just the standard material, we are still incurring some of the costs of the skin and cloth code, which hurts our performance! (Remember: we can only spend ~16 milliseconds per frame, so every performance penalty is a problem!)

How can we fix this? That's where shader specialization comes in.

Extra context for interested readers:

One of the reason why the code still incurs some of the costs of the cloth and skin code, even when not executing those code paths, is because the shader compiler must allocation enough resource (e.g., registers) for the worse-case code path. The cloth and skin material code uses many more registers than the standard material, meaning that when this shader runs the evalStandardMaterial() code path, those extra registers are effectively wasted (those registers could have been used by a different invocation of this shader working on a different pixel). A good search term to learn more is “register pressure,” especially in the context of GPU programming (both for graphics and GPGPU).

Shader Specialization

Unspecialized Material Shader

```
float4 surfaceShader(…
…
if(isStandardMaterial(
  color = evalStandard
else if(isSk  Material(materialID)
  color = evalSkin       l(…);
else if(isClothM
  color = evalCl

…
}
```

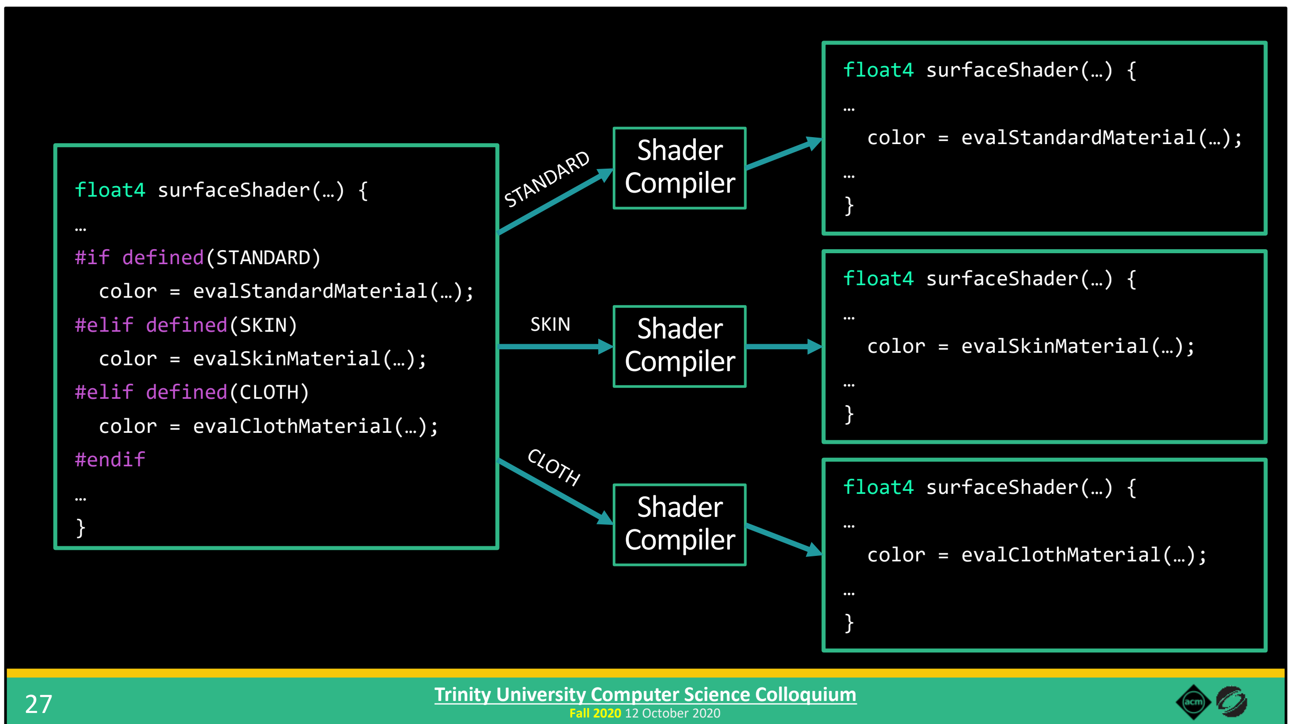
Statically specialized variants generated at shader compile time

Dynamically branch based on material ID at shader run time

Specialized Material Shader

```
float4 surfaceShader(…) {
…
#if defined(STANDARD)
  color = evalStandardMaterial(…);
#elif defined(SKIN)
  color = evalSkinMaterial(…);
#elif defined(CLOTH)
  color = evalClothMaterial(…);
#endif
…
}
```

Instead of using dynamic, runtime branches that are evaluated at shader run time, we can instead use static, compile-time branches. These will be evaluated when we compile the shader, depending on which material type we've defined.

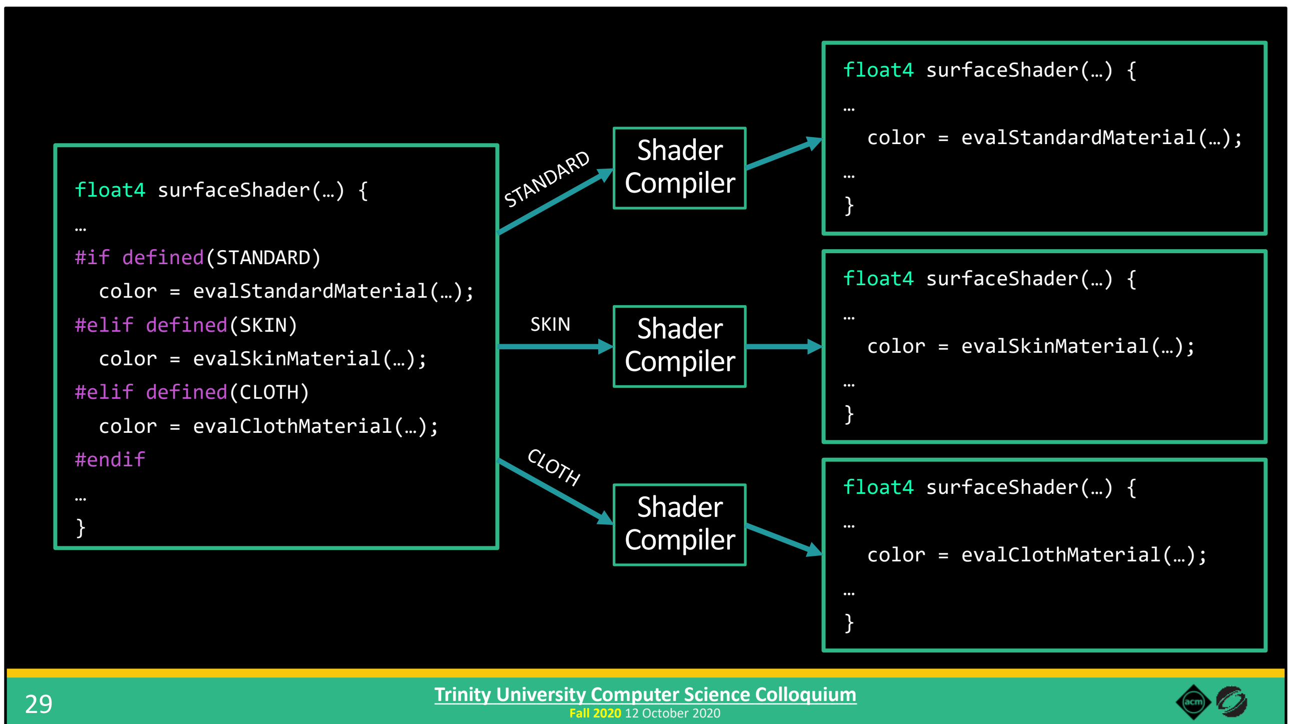


For example, we can generate a specialized shader variant just for the Standard material by defining STANDARD when we run the shader compiler. This results in a shader that just contains the Standard material code, stripping out the code for the other two.

We can do the same for the Skin and Cloth material, resulting in three total compiled shader variants, each specialized for exactly one material.

Then, when shading an object, we can use the shader variant specific to the material of that object. Generating specialized shader variants like these is very important in real-time computer graphics, not just for material, but for many different kinds of options and features.

This is a good time for questions!



We talked about how important these specialized shader variants are for performance, but writing code this way is kind of a mess. Maybe this simple example isn't too bad, but when you have more materials, more lights, and many many other types of shader features you might want variants for, it can get unwieldy fast!

Maybe there's a better way that we can write the code but still get optimized variants from it. If you're using plain ol' HLSL and D3D (or GLSL and OpenGL), there's not too much more you can do. But what about other tools? What if you use a game engine like Unity or Unreal? Do they help with this? Yes! They do so by providing a shader system!

What is a shader system?

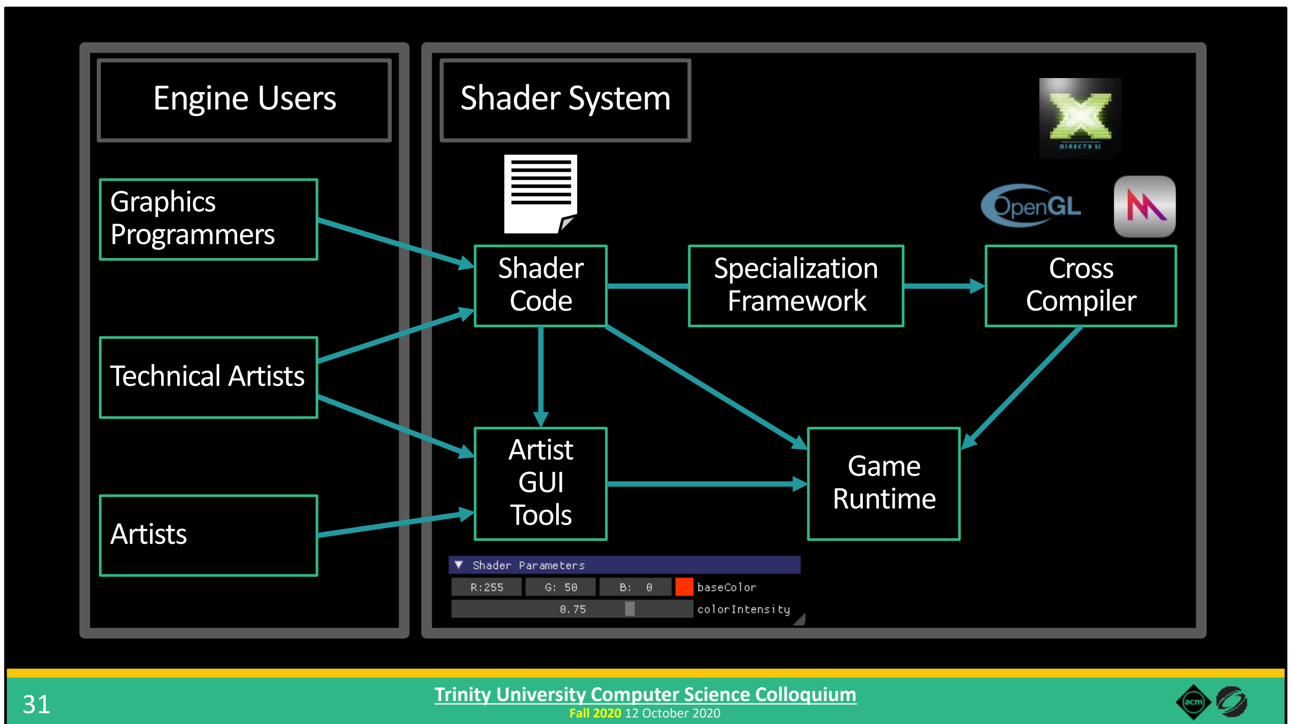
A game engine component that facilitates interacting with the rendering process

Let’s start by defining what I mean by a “shader system.”

A shader system is a game engine component that facilitates interacting with the rendering process.

Specifically, I’m talking about real-time 3D game engines like Unreal, Unity, and other in-house engines, which means that not only is performance critical, but so is enabling a wide variety of users to control different aspects of rendering.

Let’s look at an example of what I mean…



We have different types of engine users who need to use a shader system in different ways.

First, we have graphics programmers who need to be able to write shader code, in HLSL or GLSL for example.

Of course that shader code needs to be complied into executable kernels, and possibly cross-compiled if you're shipping on multiple platforms.

Then there's technical artists who also write shader code. Unlike graphics programmers, they are typically not experts in things like shader optimization. Maybe they'll use plain HLSL or GLSL too, or maybe an engine chooses to provide a custom Domain-Specific Language (or DSL).

That DSL might enable them to express which parameters to expose to a GUI that artists use to create and configure different materials.

Those configurations, along with the shader code and compiled kernels, need to be interfaced with the runtime engine code, which sets up and launches the rendering work.

And finally, shaders need to be specialized in order to achieve the best performance.

(Specialization involves taking a shader that includes code and parameters for multiple different feature options, and then generating many different variants from that shader, each corresponding to a different subset of those features. As a result, expensive features do not impact performance when they are not needed.)

Engine developers need to design these shader systems to both result in highly optimized final code while simultaneously providing the appropriate interfaces for each type of person involved in game development. But unfortunately…

Graphics APIs don't help with this task

Direct3D / OpenGL / etc. are singularly focused on providing robust, high-performance implementations on a wide range of hardware

In contrast, shader systems are ***multifaceted***

- They must provide a variety of interfaces for different users
- Engine devs are left to create these missing facets, layered on top of the APIs

Graphics APIs don't really help with this task. They are singularly focused on providing a robust, high-performance implementation on a wide range of hardware.

But as we've established, shader systems are multifaceted – they must provide a variety of interfaces for different users. Thus, engine developers are left to create layered implementations of these missing facets on top of the graphics APIs.

So how do they do that?

Current Methods

Let’s look at some current methods used to implement shader systems.

Four methods to implement shader systems

Plain C++ and HLSL*

- Preprocessor #ifdefs + #defines, shared headers for data structures, manually-authored C++ class for each shader

A simple, layered domain-specific language (DSL) with embedded HLSL*

- Unity's ShaderLab

A more complex DSL that manipulates and generates HLSL*

- Bungie's TFX language [Tatarchuk and Tchou 2017]

Modifying HLSL*

- Slang [He et al. 2018]

*or any modern shading language (e.g., GLSL or Metal Shading Language)

One is to just the facilities provided by plain C++ and HLSL. (e.g., We could use preprocessor #ifdefs and #defines in the shader to express specialization options, create shared headers for data structure, and manually author a C++ class for each shader to provide an interface for CPU engine code.)

Another is to implement a simple, layered DSL that contains embedded HLSL. Unity's ShaderLab is an example of this approach.

You could also create a more sophisticated DSL that manipulates and generates HLSL, such as Bungie's TFX language used in Destiny.

And finally, you could go so far as to modify HLSL to implement custom features, like the Slang shading language, which added some modern programming language features to HLSL.

In the paper, we go into details on all of these, but let's briefly look at one in a little more detail.

An example in Unity's ShaderLab

```
Shader "SurfaceShader" {
  Properties { lightDirection {"Light Direction", Vector} = (0,0,0) }
  …
  CGPROGRAM
  #pragma multi_compile STANDARD SUBSURFACE CLOTH

  float3 lightDirection;

  float4 surfaceShader(…) {
  …
  #if defined(STANDARD)
    color = evalStandardMaterial(shadingData);
  #elif defined(SUBSURFACE)
    color = evalSubsurfaceMaterial(shadingData);
  #elif defined(CLOTH)
    color = evalClothMaterial(shadingData);
  #endif
    return color * max(0, dot(shadingData.normal, lightDirection);
  }
  ENDCG
}
```

Express specialization options using #if

35

Here's an example shader written in Unity's ShaderLab DSL. I'm not going to discuss everything here, but I do want to highlight a few things.

First, specialization options are expressed using preprocessor #ifs, just like you would do if you were using plain HLSL.

One of the nice things that ShaderLab provides is…

Extra context for interested readers:

At first, expressing specialization options like this might seem fine, but what if (for some reason) you wanted to generate a specialized variant that contained both STANDARD and CLOTH material? You couldn't do that from this shader code. Why might you want to generate such a variant? Maybe you're using a deferred renderer, but more on that later.

An example in Unity's ShaderLab

```
Shader "SurfaceShader" {
  Properties { lightDirection {"Light Direction", Vector} = (0,0,0) }
  …
  CGPROGRAM
  #pragma multi_compile STANDARD SUBSURFACE CLOTH

  float3 lightDirection;

  float4 surfaceShader(…) {
  …
  #if defined(STANDARD)
    color = evalStandardMaterial(shadingData);
  #elif defined(SUBSURFACE)
    color = evalSubsurfaceMaterial(shadingData);
  #elif defined(CLOTH)
    color = evalClothMaterial(shadingData);
  #endif
    return color * max(0, dot(shadingData.normal, lightDirection);
  }
  ENDCG
}
```

Custom #pragma syntax to list specialization options

- Custom #pragma syntax enables ShaderLab compiler to automatically generate specialized variants

ShaderLab has a custom #pragma syntax to list specialization options.

This enables the ShaderLab compiler to automatically generate all specialized shader variants, rather than requiring users to manually generate them.

An example in Unity's ShaderLab

```
Shader "SurfaceShader" {
  Properties { lightDirection {"Light Direction", Vector} = (0,0,0) }
  …
  CGPROGRAM
  #pragma multi_compile STANDARD SUBSURFACE CLOTH

  float3 lightDirection;

  float4 surfaceShader(…) {
  …
  #if defined(STANDARD)
    color = evalStandardMaterial(shadingData);
  #elif defined(SUBSURFACE)
    color = evalSubsurfaceMaterial(shadingData);
  #elif defined(CLOTH)
    color = evalClothMaterial(shadingData);
  #endif
    return color * max(0, dot(shadingData.normal, lightDirection);
  }
  ENDCG
}
```

Double declaration of artist-configurable parameters

- Custom #pragma syntax enables ShaderLab compiler to automatically generate specialized variants

In order to expose artist-configurable parameters to a GUI, ShaderLab has a special "Properties" listing. But unfortunately, each of these parameters must be declared twice – once in the Properties and again in the embedded HLSL.

An example in Unity's ShaderLab

```
Shader "SurfaceShader" {
  Properties { lightDirection {"Light Direction", Vector} = (0,0,0) }
  …
  CGPROGRAM
  #pragma multi_compile STANDARD SUBSURFACE CLOTH

  float3 lightDirection;

  float4 surfaceShader(…) {
  …
  #if defined(STANDARD)
    color = evalStandardMaterial(shadingData);
  #elif defined(SUBSURFACE)
    color = evalSubsurfaceMaterial(shadingData);
  #elif defined(CLOTH)
    color = evalClothMaterial(shadingData);
  #endif
    return color * max(0, dot(shadingData.normal, lightDirection);
  }
  ENDCG
}
```

Bug!!!
lightDirection is a float3, not a float4

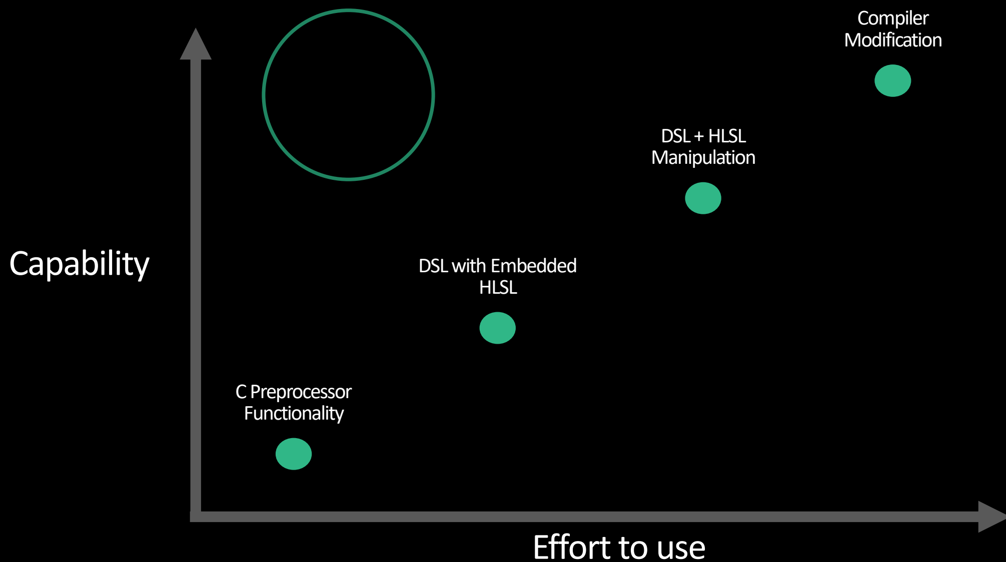
- Custom #pragma syntax enables ShaderLab compiler to automatically generate specialized variants
- Use a "stringly-typed" interface to set parameters:

```
Shader.SetVector("lightDirection",
  Vector4(1.0, 1.0, 1.0, 1.0);
```

38

Finally, runtime engine code sets parameters using a "stringly-typed" interface. This interface doesn't provide good error checking, which can lead to subtle bugs, such as here where I've accidentally used the wrong type when setting the lightDirection parameter.

Some of the other methods I mentioned improve upon these issues, but one important thing to note is that…



These methods are on an unfavorable continuum. Methods with greater capability require greater effort to use.

C preprocessor functionality is fairly simple to work with, but it's limited in the types of things you can implement with it.

At the other end of the spectrum, if you're willing to modify a compiler to add features to HLSL, you have a lot of flexibility but now you have a much larger codebase to maintain, especially as the core HLSL continues to evolve.

Engine developers today are faced with the problem of balancing between the benefits that new features might provide to users versus the effort required to implement those features.

What we'd really like is a technique that sidesteps this trade-off – one that provides lots of capabilities while requiring only a modest effort to use.

Based on this observation…

Design Goals

(… Based on this observation)

as well as the other issues we've seen in modern shader systems, we came up with a set of design goals to guide our work.

Design goals

Minimize implementation effort and maintenance costs

- E.g., we don't want to build or modify a compiler

Early error detection

- Unlike "stringly-typed" interfaces

Don't repeat yourself (DRY)

- Avoid repeat declarations of shader parameters, constant buffers, etc.

Each engine requires a unique shader system, customized to the engine's design and the needs of its users. If we can minimize the effort required to build and maintain shader systems, we can better enable developers to create robust, feature-rich implementations.

We'd like to be able to catch errors earlier, in contrast to "stringly-typed" interfaces in ShaderLab and in graphics APIs.

Programmers shouldn't have to declare the same shader parameter, constant buffer, etc. more than once.

Design goals (cont.)

Performance

- Minimize overheads to CPU and GPU code

Productivity for artists and technical artists

- Don't hinder their workflows

Support options for static and dynamic feature composition

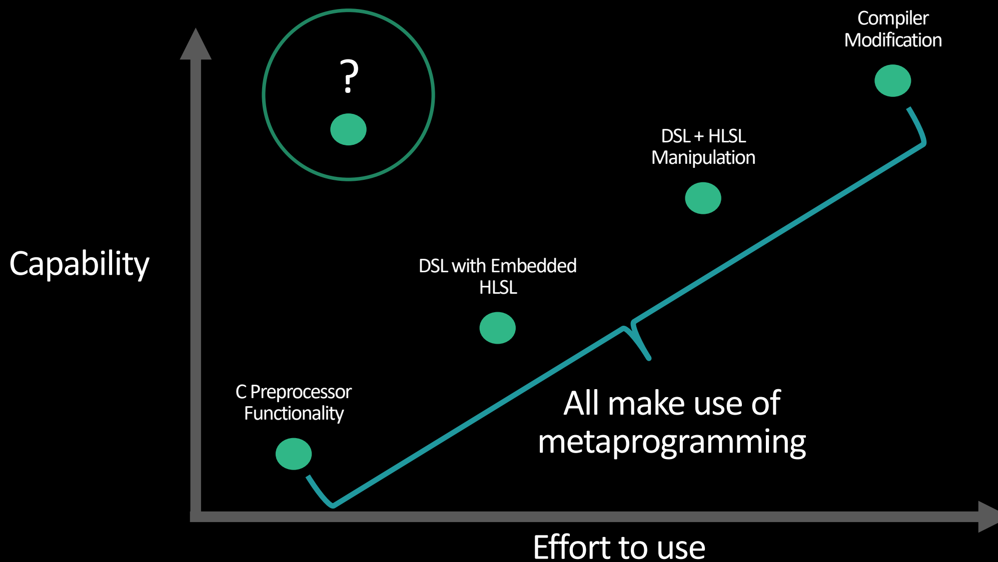
- To explore trade-offs between static and dynamic shader specialization

Performance is paramount in real-time graphics applications, so our system should strive to minimize overheads to GPU shader code and CPU engine code, as well as enable developers to explore opportunities to improve performance.

Productivity is key for artists, so a shader system must provide them with familiar workflows.

To achieve maximum performance, engines generate many specialized shader variants. However, complete static specialization can lead to additional overheads that decrease performance. So, we want our system to enable exploration of these trade-offs in the hopes of improving overall performance.

Given the landscape of existing solutions…



Our first goal of minimizing implementation effort seems at odds with some of our other goals. Certainly, we could achieve many of them by modifying a compiler and adding features to HLSL, but that requires a high implementation effort.

Is there a method that avoids this trade-off?

When examining these existing methods, we discovered that they all happen make extensive use of metaprogramming, whether they realize it or not.

So what is metaprogramming?

Metaprogramming

Writing code that manipulates other code

- Including reading, analyzing, transforming, or generating code

Examples

- Compilers
- C preprocessor functionality
  - (e.g., #if, #define, #include)
- Macros

```
float4 surfaceShader(…) {
…
#if defined(STANDARD)
  color = evalStandardMaterial(…);
#elif defined(SKIN)
  color = evalSkinMaterial(…);
#elif defined(CLOTH)
  color = evalClothMaterial(…);
#endif
…
}
```

We broadly define metaprogramming as writing code that manipulates other code, including reading, analyzing, transforming, or generating code.

What are some examples?

With this definition, a compiler is doing metaprogramming. A compiler reads in source code, performs analysis in order to find opportunities for optimization, performs those optimizations by transforming the code, and finally generates code that the hardware can execute. A compiler is a very complex form of metaprogramming.

At the other end of the spectrum, we have the functionality provided by the C preprocessor. Those #ifs that we saw earlier are metaprogramming. The #if lines here aren't going to be executed at runtime. Instead, they are going to be evaluated at compile time, which will result in different generated code depending on which of the material types has been #defined – it's going to generate one of the three evalMaterial() lines.

Macros are also metaprogramming. Macros are basically functions that are executed at compile time.

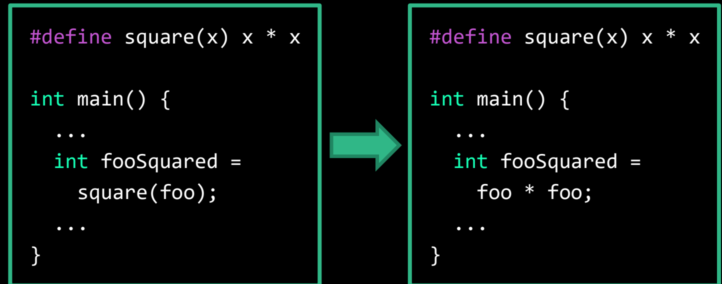
Metaprogramming

Writing code that manipulates other code

- Including reading, analyzing, transforming, or generating code

Examples

- Compilers
- C preprocessor functionality
  – (e.g., #if, #define, #include)
- Macros

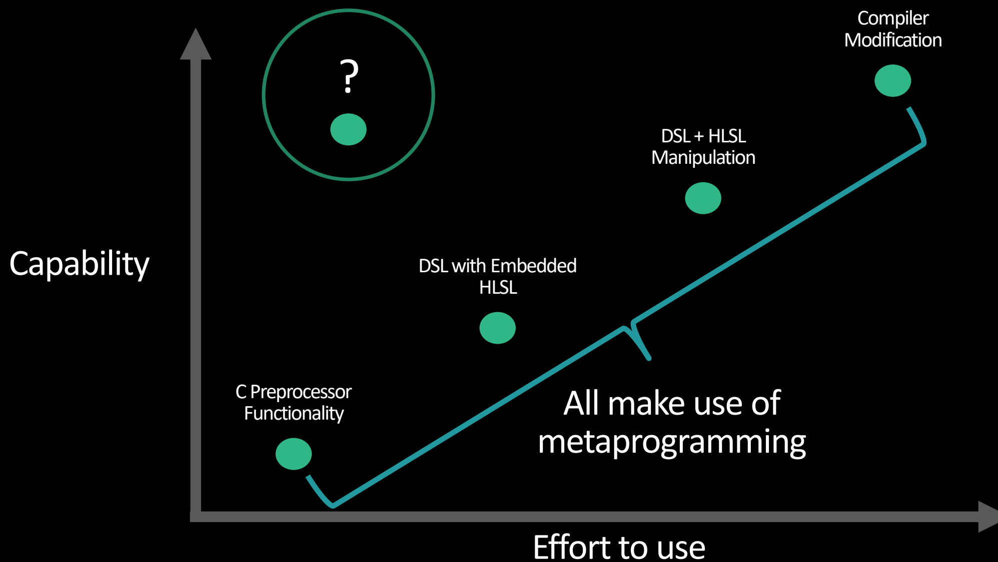


Here's a simple macro example. Many languages have macros with different degrees of functionality. This example is in C/C++, since C++ is the industry standard in game development.

In this example, I've defined a macro called "square" that will compute the square of a number.

When we use "square" in a function, we're not actually calling that macro at runtime. Instead, that macro is evaluated at compile time and code is generated as a result. In this case, the call to the "square" macro will be replaced with "foo * foo," as shown on the right.

That is broadly what metaprogramming is – writing code that manipulates other code. There are many techniques that involve metaprogramming…

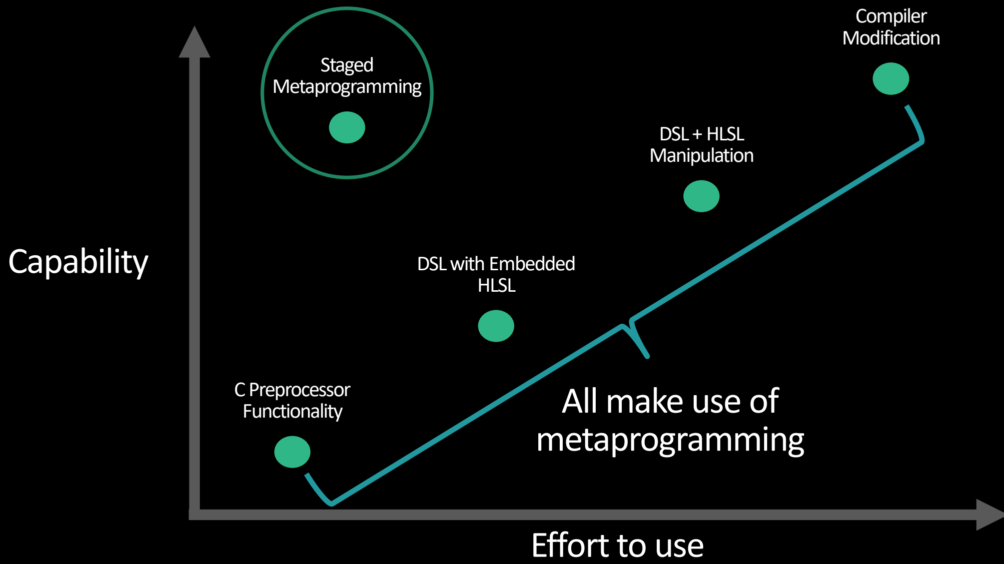


(There are many techniques that involve metaprogramming…)

Including these methods used to implement shader systems. And that was our key insight – all of these methods make extensive use of metaprogramming.

Using this key insight, we decided to make metaprogramming a fundamental design principle at the core of our shader system, and to find a metaprogramming technique that sidesteps this apparent trade-off between capability and complexity.

As I'm sure you can guess from the title of this talk…



The technique we identified is staged metaprogramming.

Staged Metaprogramming

So, what is staged metaprogramming…

Staged metaprogramming

Explicit stages of code execution

- E.g., compile-time stage / runtime stage

Code running in earlier stages can construct and manipulate code in later stages

Consistent with description of *multi-level languages* [Taha 1999]

- Also includes *multi-stage languages*

In staged metaprogramming, there are multiple explicit stages of code execution. For example, we could have a stage that conceptually executes at application compile time, versus a stage that executes at application run time.

Code running in an earlier stage of execution can construct and manipulate code that will run in a later stage.

Our definition of staged metaprogramming aligns with the description of a multi-level language and also includes multi-stage languages as well. If you're familiar with those terms, this might provide some extra context.

So what makes up a staged metaprogramming environment?

Key features of staged metaprogramming

Code is a first-class citizen

- Pass as arguments, return from functions, store in structs

Code is constructed using regular language syntax using *quasi-quote*

- myCode = quote var outColor = diffuse + specular end
- Quasi-quotes are hygienic and lexically scoped (but you can violate this)

Unquote inserts quasi-quoted code into the runtime application

- [myCode]

Quasi-quotes can be specialized to generate different version

Let's talk about the key features.

Code is a first-class citizen. Programs can operate on code in the same way that they can operate on other constructs, including passing code as arguments, returning code from functions, and storing code in data structures.

Code is constructed using regular language syntax using quasi-quote. In this example, we have the keywords "quote" and "end" to denote that we're constructing code. The code between these keywords is expressed using regular syntax, but this code is not executing here. Instead, it is stored in the "myCode" variable for use later.

These quotes are hygienic and lexically scoped, so the definition of "outColor" here would not conflict with another variable using the same name elsewhere (unlike preprocessor macros). However, you intentionally can violate this property when needed.

The unquote operator splices quoted code into the runtime application. Here we're taking the myCode variable and inserting its contents into the program.

Also, quotes can be specialized to generate different versions, similar to how shaders can be specialized into different variants.

If you're used to working in C++ or another popular systems language, you're probably not used to seeing these quote and unquote constructs. Well that's because…

Lua-Terra: a research substrate for staged metaprogramming

C++ and HLSL don't have the features of staged metaprogramming

So, we used Lua-Terra [DeVito et al. 2013] to explore our ideas

- Multi-stage language
- Uses Lua code in the first stage to manipulate next-stage Terra code

Lua – high-level scripting language
Terra – simple, low-level, C-like language

These languages don't have the features required for staged metaprogramming.

We used a language called Lua-Terra to demonstrate how staged metaprogramming is useful for shader systems. Lua-Terra is a multi-stage language that uses Lua code in the first stage to manipulate and generate next-stage Terra code

You might be familiar with Lua. It is a high-level scripting language commonly used in game development already.

In contrast, Terra is simple, low-level, C-like language. We chose Lua-Terra specifically because Terra models the lower-level systems language environment that is commonly used in engine development.

However, high-level scripting and code generation can be expensive operations, and as we know performance is critical for game engines…

Compile-time staged metaprogramming

Performance is important, so all metaprogramming occurs at application compile time

- Avoids overhead of generating code at runtime
- In contrast to prior work (e.g., Sh [McCool et al. 2002] and Vertigo [Elliott 2004])

All Lua code executes at compile time

Runtime application and shader code are written in Terra

So in our system, all metaprogramming occurs at application compile time. We aren't generating any code while the game is running, all of that happens beforehand. This is in contrast to prior work, which generates code at runtime.

On the last slide, I mentioned that the Lua code metaprogrammings the Terra code, and since all of the metaprogramming happens at compile time, all of the Lua code executes at compile time as well. We don't run any Lua code during game runtime (although we could if we wanted to add Lua scripting into the application).

What's left at runtime is just the C-like Terra code. The game runtime, as well as all shader code are written in Terra.

Now let's take a look at a shader in our system, which we call Selos.

Example shader in our system (called Selos)

```
shader SurfaceShader {
  ConfigurationOptions {
    MaterialType = MaterialSystem.MaterialTypeOption.new()
  }
  …
  uniform LightData {
    @UIType(Slider3) lightDirection : vec3
  }
  …
  fragment code
    …
    color = [MaterialType:eval()](shadingData)
    return color *
      max(0, dot(shadingData.normal, lightDirection);
  end
}
```

Single declaration of parameter

53

Again, I'm just going to highlight a few key points.

Unlike in ShaderLab, in Selos we can express GUI controls directly alongside the parameter declaration, avoiding the double-declaration problem.

Example shader in our system (called Selos)

```
shader SurfaceShader {
  ConfigurationOptions {
    MaterialType = MaterialSystem.MaterialTypeOption.new()
  }
  …
  uniform LightData {
    @UIType(Slider3) lightDirection : vec3
  }
  …
  fragment code
    …
    color = [MaterialType:eval()](shadingData)
    return color *
      max(0, dot(shadingData.normal, lightDirection);
  end
}
```

- Statically-checked interface to shaders:

```
var myShader = SurfaceShader.new()
var lightData =
         myShader.LightData:map(…)
lightData.lightDirection = vec4(…)
```

Compile-time error:
lightDirection is a vec3

54

Our system generates a statically-checked interface for shaders, meaning that that bug from my ShaderLab code before is instead reported as a compile-time error in Selos.

Instead of using preprocessor #if…

Example shader in our system (called Selos)

```
shader SurfaceShader {
  ConfigurationOptions {
    MaterialType = MaterialSystem.MaterialTypeOption.new()
  }
  …
  uniform LightData {
    @UIType(Slider3)
  }
  …
  fragment code
    …
    color = [MaterialType:eval()](shadingData)
    return color *
      max(0, dot(shadingData.normal, lightDirection);
  end
}
```

Specialization expressed and controlled through ConfigurationOptions

- Statically-checked interface to shaders:

```
var myShader = SurfaceShader.new()
var lightData =
         myShader.LightData:map(…)
lightData.lightDirection = vec4(…)
```

- Automatically generate variant (like ShaderLab)
- Opportunity to explore more specialization options (we'll return to this)

55

Shader specialization is expressed and controlled through ConfigurationOptions.

This allows Selos to automatically generate all variants (like in ShaderLab).

And it also enables us to explore other options for shader specialization. I mentioned before that you couldn't generate a shader with both STANDARD and CLOTH materials from the ShaderLab code, but in ours we easily can. We'll return to this part later.

Staged metaprogramming is the principle design decision in our system…

Example shader in our system (called Selos)

```
shader SurfaceShader {
  ConfigurationOptions {
    MaterialType = MaterialSystem.MaterialTypeOption.new()
  }
  …
  uniform LightData {
    @UIType(Slider3) lightDirection : vec3
  }
  …
  fragment code
    …
    color = [MaterialType:eval()](shadingData)
    return color *
      max(0, dot(shadingData.normal, lightDirection);
  end
}
```

56

But looking at this example shader, notice that those quote and unquote mechanisms I described earlier don't show up in this code. In fact, this shader looks pretty similar to a shader written in GLSL or HLSL, and it doesn't really exhibit aspects of staged metaprogramming directly. This design is intentional.

While staged metaprogramming underlies our system's implementation, it also introduces some new and unfamiliar programming constructs. How do we cope with that?

Other Key Design Decisions

That leads me to our other key design decisions.

Other key design decisions

Write shader definitions using a domain-specific language (DSL)

- Present a familiar interface to technical artists
- Don't expose the metaprogramming directly

Represent shaders as compile-time Lua objects

- Consistent interface to manipulate shader code
- Compile-time only, so it doesn't add runtime overhead

First, shaders are written in a DSL that's similar to GLSL. This provides a familiar interface to technical artists, so that they can be productive without worrying about those new metaprogramming constructs. So how do we use staged metaprogramming then?

Internally in the system, we represent shaders as compile-time Lua objects. This provides a consistent interface to manipulate shader code, since we can store code directly in data structures. Because it only exists at compile time, this representation does not add overhead to the runtime application.

While shader-specific features are expressed through our DSL…

Other key design decisions (cont.)

Write shader logic and application code in the same language

- Share types and functions between CPU and GPU code

Generate runtime data structures for shaders

- Statically-checked interface – catches errors at application compile time
- Downside: must recompile application whenever a shader's interface changes
  - But we can hot reload if only the logic changes

Core shader logic as well as CPU-side code is written in Terra. This allows us to share types and functions between CPU and GPU code. This is something that you don't typically get when you're using C++ for the CPU code and HLSL for the GPU code. But you get that for free in our system.

As I mentioned earlier, we generate runtime data structures for shaders, which helps us catch more errors at compile time. However, this means that the game must be recompiled if the interface to a shader changes, like if you add a parameter for example. But if only the core logic changes, we can still hot reload shaders. It's a bit of a trade-off – the application needs to be recompiled more often, but it does provide better error checking.

I want to take a second to point out that…

Our implementation required only a modest effort

System Component	Language(s)	Lines of Code
Unity ShaderLab DSL	Flex/Bison/other	~2000
Slang Compiler	C++	~67,000
Selos (Ours)		
Core	Lua-Terra	~2300
HLSL/GLSL Backend	Lua-Terra	~2200

Improvements over ShaderLab, but with similar lines of code

Backends are reusable

Our implementation of these features required only a modest effort. The lines of code of the Selos core is comparable to that of Unity's ShaderLab implementation, while also improving on some of ShaderLab's issues. And both are much smaller than building or modifying a compiler.

We also had to implement backends to convert Terra to HLSL and GLSL, but we believe that these components are not engine-specific and could be shared across shader systems as an open source component.

This is a good time for questions!

In the previous design decisions, we recommend hiding the complexities of metaprogramming from many shader writers behind our shader DSL. But the power of raw metaprogramming provides the ability to implement some interesting features…

Other key design decisions (cont.)

Implement complex specialization options using raw metaprogramming

- Allows expert graphics programmers to explore the specialization design space

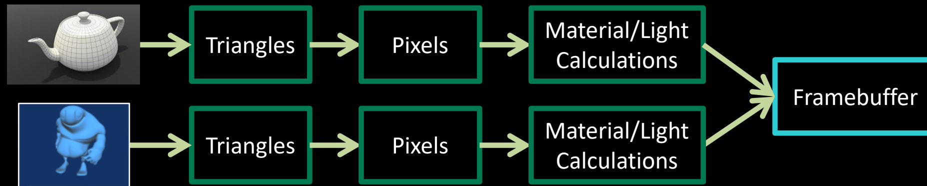
So we encourage expert graphics programming to use the features of staged metaprogramming directly when implementing specialization frameworks.

So let's look at a case study of something interesting we can do in Selos using staged metaprogramming…

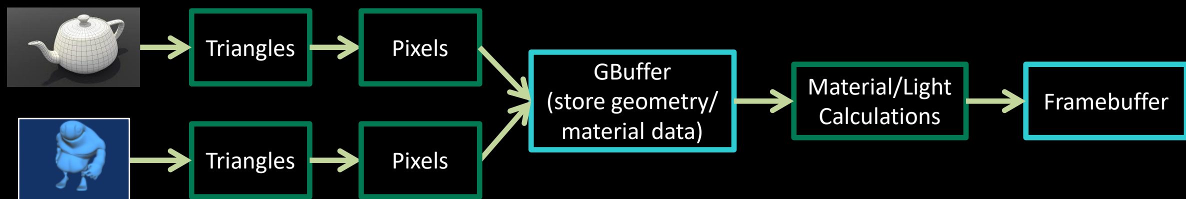
Case Study:
Exploring the Specialization Design Space

Which is to explore the shader specialization design space. Specifically, we're going to look at specialization in a deferred renderer.

Forward Rendering



Deferred Rendering



What we've talked about so far has been through the lens of forward rendering. You have an object that's composed of triangles, you chop those triangles up to generate pixels, you perform the material and lighting calculations to determine the pixel's color, and then you composite the pixels into the framebuffer to eventually display on screen. Rinse, repeat for every object in your scene, overwriting pixels in the framebuffer if newly generated pixels are closer to the camera.

In contrast, deferred rendering "defers" the material and lighting calculations until later. Instead, you take the triangles, rasterize them into pixels, and then store information about those pixels into a GBuffer (short for Geometry Buffer). This include geometric data as well as material data. The GBuffer stores one set of data per final on-screen pixel, so if another object produces a pixel that's closer to the camera, it's pixel data will overwrite whatever data is already stored in the GBuffer. Once you've processed all objects and stored the pixel data into the GBuffer, then you run the material and lighting calculations by fetching data from the GBuffer to calculate final pixel colors.

Deferred rendering can be significantly faster than forward rendering because it only performs the really expensive material and lighting calculations once per final, displayed pixel, whereas in forward rendering, you might perform that expensive calculation for one pixel, only for that pixel to be discarded because a pixel from a different object is closer to the camera. (I'm handwaving and oversimplifying heavily

here. There are other benefits of deferred rendering, and there are other ways to prevent “overdraw” in forward rendering.)

But one of the problems with deferred rendering is that we can’t do that complete specialization that I talked about previously.

Specialization in deferred rendering

Forward Lighting Shader

```
float4 surfaceShader(…) {
…
#if defined(STANDARD)
  color = evalStandardMate
#elif defined(SUBSURFACE)
  color = evalSubsurfaceMa
#elif defined(CLOTH)
  color = evalClothMateri
#endif
…
}
```

Statically specialized variants generated at shader compile time

Dynamically branch based on material ID at shader run time

Deferred Lighting Shader

```
float4 surfaceShader(…) {
…
if(isStandardMaterial(materialID)
  color = evalStandardMaterial(…);
else if(isSubsurfaceMaterial(materialID)
  color = evalSubsurfaceMaterial(…);
else if(isClothMaterial(materialID)
  color = evalClothMaterial(…);

…
}
```

65

Here's what a material shader might look like in a forward renderer. We're using preprocessor #ifs to denote different code paths based on the type of material we're rendering, and then we can generate statically specialized variants at shader compile time for each material, one variant per material.

However, when performing shading in a deferred renderer, different pixels in the GBuffer might require different material features. So, the shader must be able to dynamically enable or disable features per-pixel at shader run time, based on material ID in this case.

Even when complete static specialization is not feasible, some specialization can still be beneficial.

Deferred lighting specialization in Uncharted 4

Image from "Deferred Lighting in Uncharted 4" by Ramy El Garawany (Naughty Dog), Advances in Real-Time Rendering in Games: Part I, SIGGRAPH 2016

[El Garawany 2016]

For example, the deferred lighting pass in Uncharted 4 made use of partial specializations.

In the default deferred rendering setup, after generating the GBuffer, the renderer launches the material and lighting calculation shader on all pixels at once.

Deferred lighting specialization in Uncharted 4

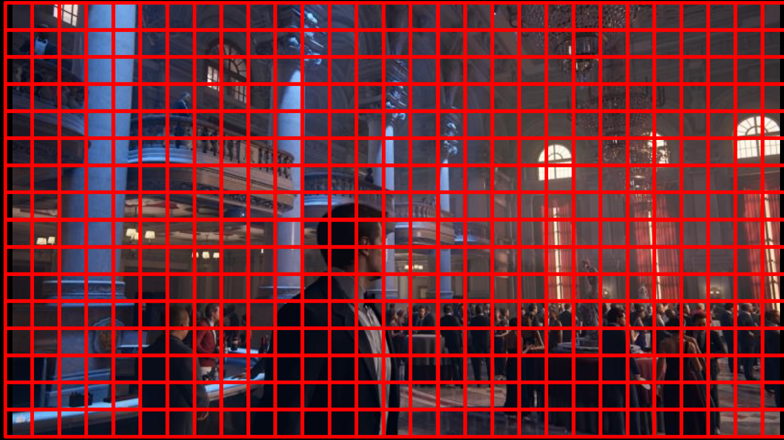


Image from "Deferred Lighting in Uncharted 4" by Ramy El Garawany (Naughty Dog),
Advances in Real-Time Rendering in Games: Part I, SIGGRAPH 2016

67 [El Garawany 2016]

(In the default deferred rendering setup, after generating the GBuffer, the renderer launches the material and lighting calculation shader on all pixels at once.)

Thus, the renderer must use a shader that contains code for all materials.

Deferred lighting specialization in Uncharted 4

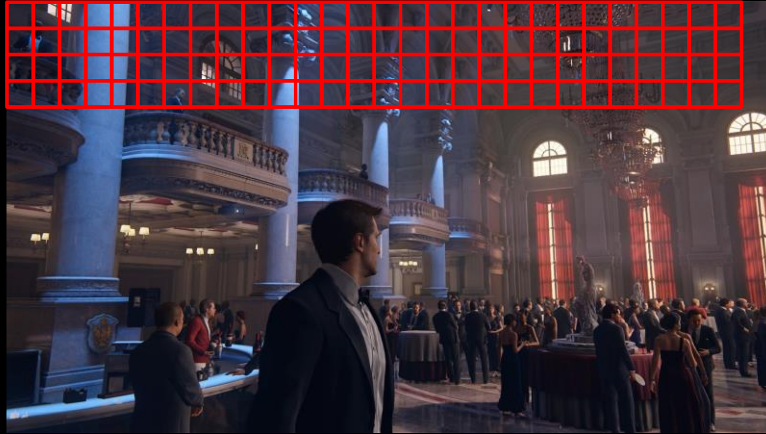


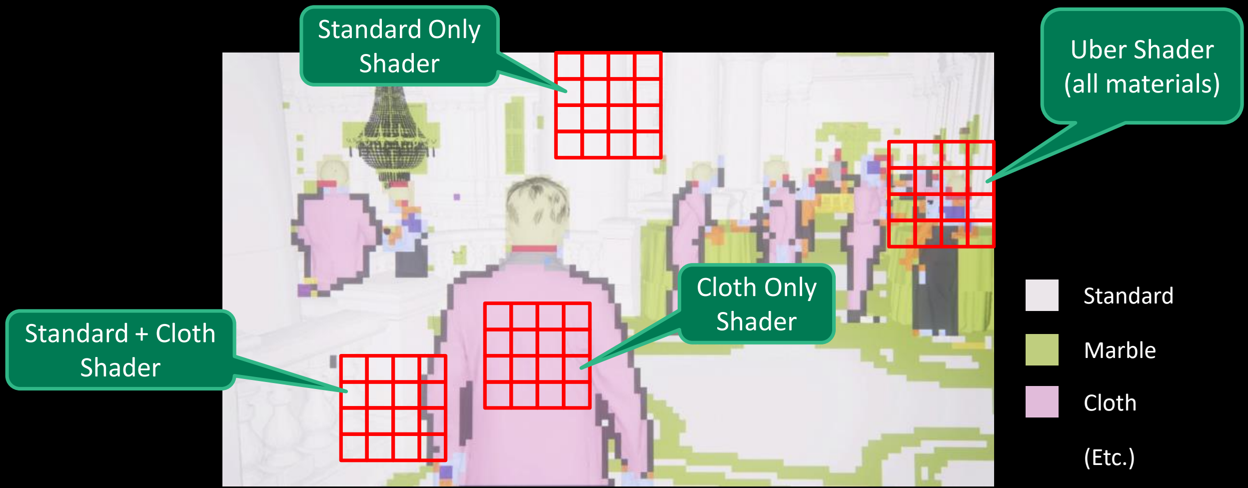
Image from "Deferred Lighting in Uncharted 4" by Ramy El Garawany (Naughty Dog), Advances in Real-Time Rendering in Games: Part I, SIGGRAPH 2016

[El Garawany 2016]

68

Instead of launching all pixels at the same time, we can instead dispatch pixels on a per-tile basis.

What benefit does this give us?



If we look at the different materials in this scene, we can see that pixels near each other tend to be the same material type.

So, when we dispatch a tile whose pixels are all the standard material, we can use a shader that only has standard material code.

Similarly, for this tile, we can use a cloth-only shader.

This tile overlaps both standard and cloth pixels, so we need to use a shader that has code for both.

And sometimes, we might have to fall back to that default uber shader with all materials.

This design significantly improved performance for Uncharted 4. But, as you might realize, this results in many shader variants for all possible combinations of materials, and it turns out that…

Deferred lighting specialization in Uncharted 4

Generated per-tile bitmask of features needed in that tile

- E.g., This 16x16 tile contains metal and fabric

Dispatch tiles using shaders variants specialized for different feature combinations

- E.g., Render this tile with a shader that just has metal and fabric code

If all pixels in a tile are the same material, use a “branchless” variant

- E.g., This tile is all fabric, so dispatch using a fabric-only shader that omits checking materialID

70

[El Garawany 2016]

(Skipping this slide during the talk, but leaving it here for interested readers.)

Extra context for interested readers:

The deferred lighting pass in Uncharted 4 made use of partial specializations.

First, it splits the screen space into 16 by 16 pixel tiles and generates a per-tile bitmask of all of the material features present in that tile. For example, let’s say a given tile contains some pixels that should be rendered as metal and others that should be rendered as fabric.

Then, it dispatches tiles using different shaders, specialized for particular feature combinations. So that tile containing both metal and fabric would be rendered using a shader that just contains metal and fabric code. Code for other types of materials would be striped away from that variant.

As an optimization, if all pixels in a given tile are the same material, they dispatch it using a “branchless” variant. So if a tile is only fabric, for example, the shader can skip checking the material ID.

This design significantly improved performance for Uncharted 4. But, as you might realize, this results in many shader variants, and it turns out that…

Overspecialization can hurt performance!

Lots of shader variants!

- Increases shader switching overhead, dispatch overhead, game load time, etc.

Do we need all of these specializations?

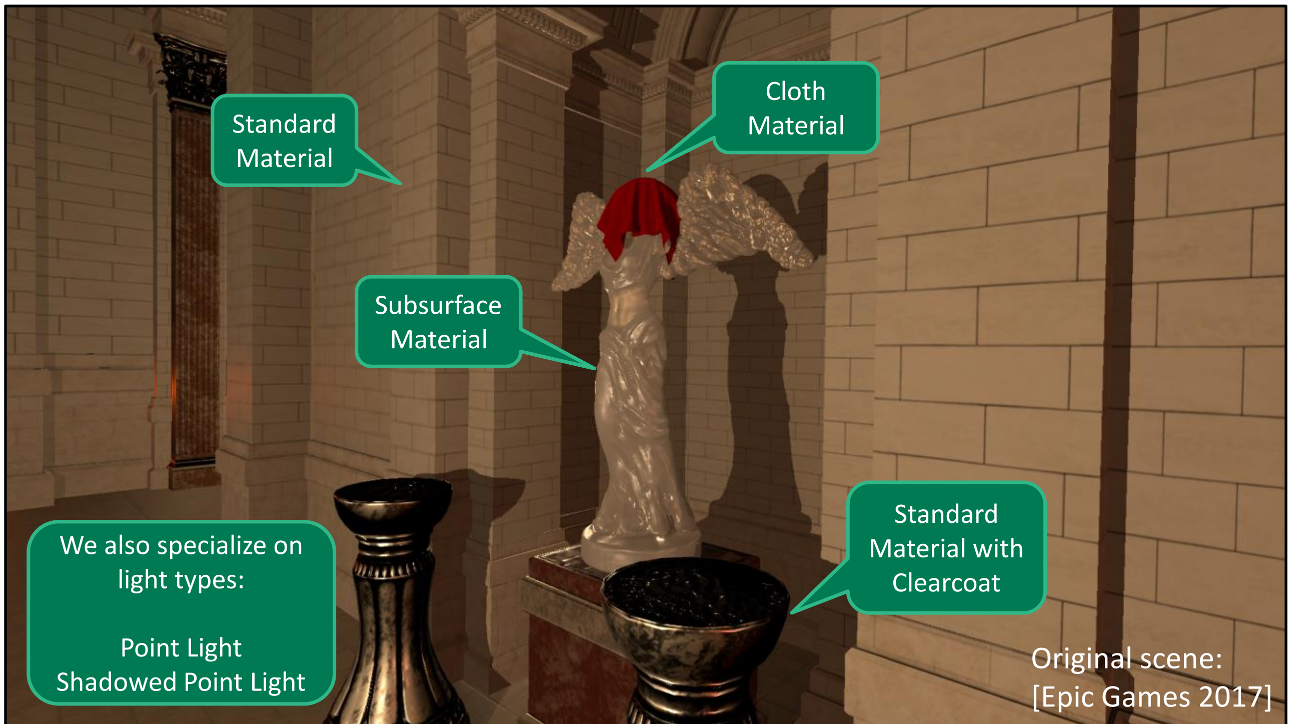
Staged metaprogramming enables exploration of this design space

Overspecialization can actually hurt performance!

Having lots of variants increases shader switching overhead, dispatch overhead, and game load time.

Do we really need all of these specialized variants? It's likely that some materials are more important to specialize than others.

We can explore this design space in Selos using staged metaprogramming.



We implemented a deferred renderer and also implemented specialization similar to Uncharted 4's for our deferred lighting shader.

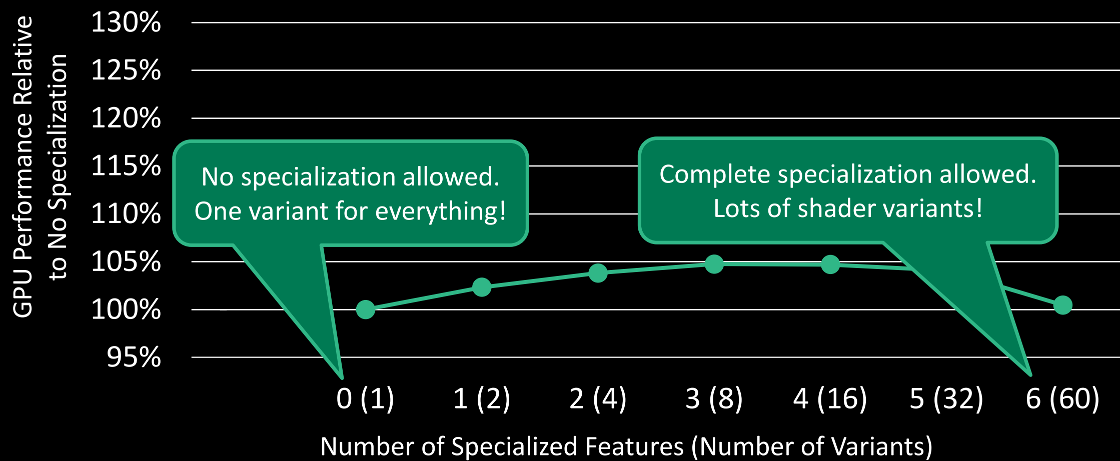
We used our system to render the Sun Temple scene from the ORCA repository, but we had to make a few modifications.

Like most other widely available scenes, this scene does not specific what type of BRDF to use for each material. So we chose to render certain objects with different BRDF types in order add some material variation. We also added some cloth geometry, which isn't in the original scene.

We also specialize based on light types by doing light culling, and then determining whether a tile does or does not contain at least one of a given light type.

This gives us six different features that we can specialize – four material features and two light features.

So we generated all of the possible variants, but then we also restricted our system to only specialize some of the material and light features.

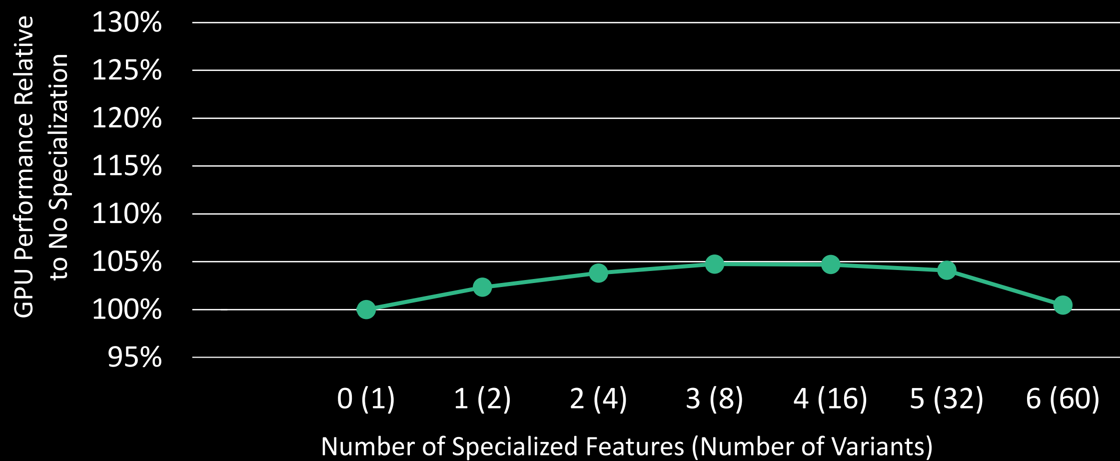


What we found is that partial specialization achieves the best performance.

This graph measures the GPU performance of our deferred lighting pass.

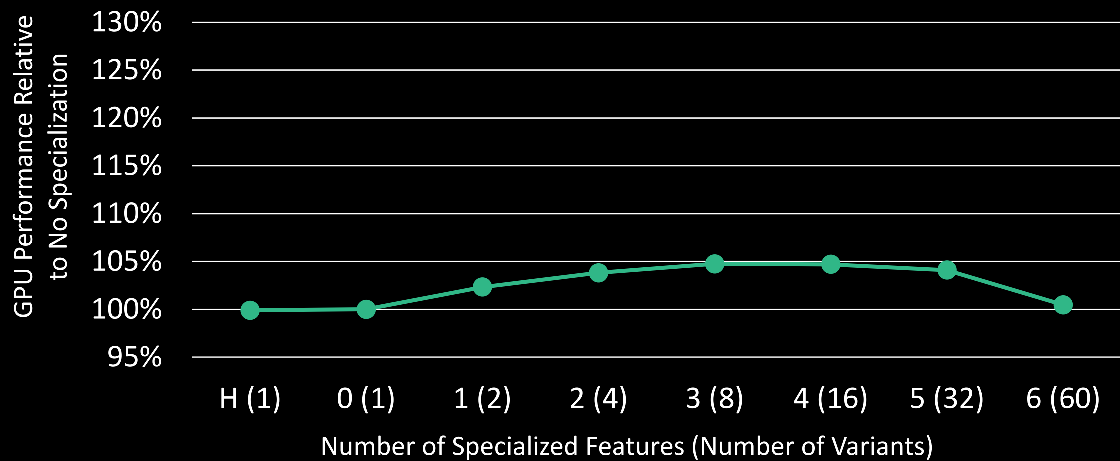
On the X axis, we have the number of material and light features we're allowing to be specialized, with the total number of generated variants in parentheses. For example, the zero case mean we're not allowing any features to be specialized, so the only shader variant is the typical deferred lighting shader, which uses dynamic branches for all feature selection. On the other end, we allow specialization for all features, which generates variants for all possible combinations of features, resulting in 60 total variants.

On the Y axis, we have the relative GPU performance compared against the baseline deferred lighting shader, which again has no specialization.



What we observe is that increasing the amount of specialization increases performance, but only to a point. Then, performance starts to degrade. So overspecialization decreases performance, and the sweet spot is in the middle.

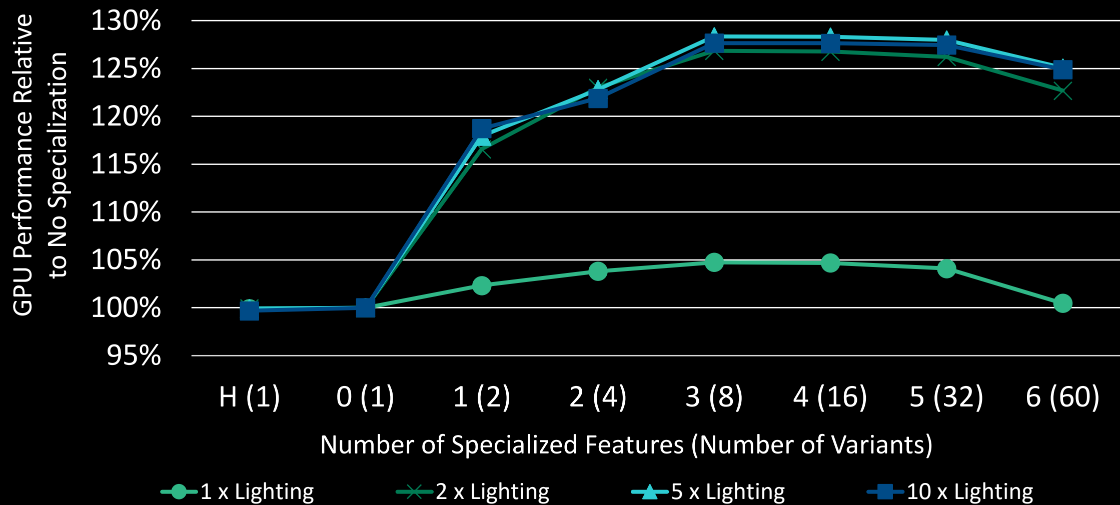
As a sanity check, we also handwrote an HLSL shader for the typical deferred shader case, to make sure our abstractions weren't adding overhead.



(As a sanity check, we also handwrote an HLSL shader for the typical deferred shader case, to make sure our abstractions weren't adding overhead.)

As you can see, it performs similarly to the version generated by our system.

Our test scene only has 14 lights, whereas game often have many more. We wanted to see how performance changes as we increases the amount of lighting computations to be more in line with modern 3D games…

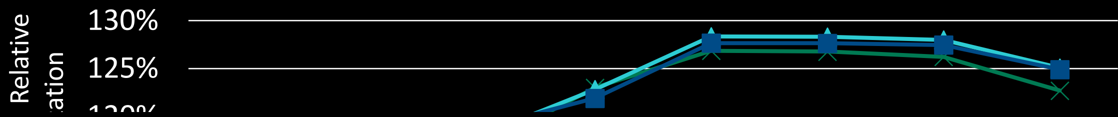


What we found is that as the amount of lighting work increases, the effects of specialization is even more pronounced. And still, partial specialization produced the best results.

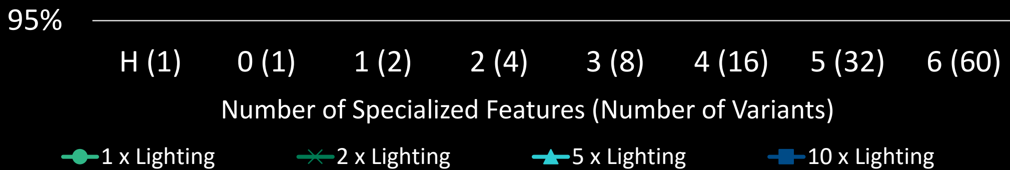
This and other types of exploration is something a shader system should help you with, and…

// Note: The reason why there are only 60 variants in the last case, rather than 64 – When enabling specialization for all 6 features, the system generates variants for all possible combinations of material and lighting features. For each feature, there's basically a choice of whether to include it in the shader or to omit it, hence there will be 2^6 = 64 variants. However, in the variants where all material features are omitted from the shader, there's no material to shade, so those shaders are effectively invalid. There are four such cases – when all of the material and light features are omitted, when only both light features are enabled, or when either one or the other light feature is enabled. So we're left with 60 total valid variants.

Partial specialization achieves the best performance



Staged metaprogramming enabled this exploration in our system



77

We could easily perform this exploration in our system because of our principled use of staged metaprogramming.

Wrapping Up

Summary

Staged metaprogramming is a key methodology for shader system development

Using it, we build the Selos shader system

- Same language for CPU & GPU code
- Statically-checked shader interface
- Performance improvements through design space exploration

We build everything in user-space code, using off-the-shelf Lua-Terra

We identified staged metaprogramming as a key methodology to aid in shader system development.

We used it to build a shader system that uses the same language for both CPU and GPU code, provides statically-checked interfaces to shaders to catch more errors earlier. And we were able to improve performance of our deferred renderer by exploring the shader specialization design space.

I want to emphasize that we build everything in user-space code, using off-the-shelf Lua-Terra. We didn't modify the Lua-Terra compiler at all.

Unfortunately, popular systems languages today don't have all the features required for staged metaprogramming, but…

The future of metaprogramming

Systems languages are trending towards better metaprogramming facilities

- Rust
- Various C++ proposals (e.g., [Sutter 2018], [Chochlik et al. 2018])
- Circle compiler for C++ [Baxter 2019]

They are trending in the right direction. The Rust language has some interesting features. There's various proposals about metaprogramming to the C++ committee, such as metaclasses and better support for compile-time reflection. And there's also the Circle compiler, which adds new introspection, reflection, and compile-time execution features to C++.

So hopefully in the future, the features of staged metaprogramming will be available in modern systems languages.

But beyond shader systems…

Heterogeneous programming for graphics

NVIDIA's CUDA language gives GPU compute code first-class, heterogeneous treatment in C++

How can we achieve the same for GPU graphics code?

- Many additional challenges!

I'm interested in thinking more broadly about about heterogeneous programming for graphics.

NVIDIA's CUDA language gives GPU compute code (General Purpose GPU Computing, or GPGPU) first-class, heterogeneous treatment in C++. How can we achieve the same for GPU graphics code? Graphics has many additional challenges!

Our work is a step in the right direction. We can use the same language for CPU and GPU code and provide some nice shader interfaces, but it's far from achieving true heterogeneity.

And furthermore...

Heterogeneous programming for graphics … and other domains?

Graphics is a challenging domain!

- Can we apply the lessons to other areas?

Potentially many future processor types

- We need programming models to support them

Staged metaprogramming allowed us to add support for a different processor type (GPU) purely as library code

- No fundamental language changes → don't have to involve standard bodies

What about other domains?

Graphics provides a complex and well-explored area in which to investigate the broader concept of heterogeneity. Can we apply the lessons we learn about graphics programming in other domains?

In a future with potentially many different processor types, we need programming models to support them.

What I think is really interesting is that staged metaprogramming allowed us to add support for a different processor type purely as library code. We didn't have to modify Terra at all. And I think that's a very powerful property of staged metaprogramming.

Acknowledgments

Discussions and advice

- Anjul Patney, Ahmed Mahmoud, Alex Kennedy, Angelo Pesce, Aras Pranckevičius, Brett Lajzer, Brian Karis, Chuck Lingle, Dave Shreiner, Hugues Labbe, Joe Forte, Michael Vance, Padraic Hennessy, Paul Lalonde, Wade Brainerd, Zachary DeVito, and the reviewers

Feedback on the presentation

- Owens Group Members

Early code contributions

- Francois Demoullin

Financial Support

- Intel Corporation, National Science Foundation Graduate Research Fellowship Program, NVIDIA

There's many people I would like to thank for discussions and advice, feedback on the presentation, early code contributions, and financial support.

SIGGRAPH
ASIA 2019
BRISBANE

Thank you

Kerry A. Seitz, Jr.
kaseitz@ucdavis.edu

Code: github.com/kseitz/selos **Paper & Slides:** seitz.tech

sa2019.siggraph.org

And also the source code is available on GitHub, and you can find the full paper and slides on my website.

Thank you for your attention!

This is a good time for questions!

References

Sean Baxter. 2019. Circle. https://github.com/seanbaxter/circle

Matus Chochlik, Axel Naumann, and David Sankel. 2018. Static reflection. C++ Standards Committee Papers. http://wg21.link/p0194

Zachary DeVito, James Hegarty, Alex Aiken, Pat Hanrahan, and Jan Vitek. 2013. Terra: A Multi-Stage Language for High-Performance Computing. In Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI ’13). 105–116. https://doi.org/10.1145/2491956.2462166

Ramy El Garawany. 2016. Advances in Real-time Rendering, Part I: Deferred Lighting in Uncharted 4. In ACM SIGGRAPH 2016 Courses (SIGGRAPH ’16). http://advances.realtimerendering.com/s2016/index.html

References (cont.)

Conal Elliott. 2004. Programming Graphics Processors Functionally. In Proceedings of the 2004 ACM SIGPLAN Workshop on Haskell (Haskell ’04). 45–56. https://doi.org/10.1145/1017472.1017482

Epic Games. 2017. Unreal Engine Sun Temple, Open Research Content Archive (ORCA). https://developer.nvidia.com/ue4-sun-temple

Yong He, Kayvon Fatahalian, and Tim Foley. 2018. Slang: Language Mechanisms for Extensible Real-time Shading Systems. ACM Transactions on Graphics 37, 4, Article 141 (July 2018), 13 pages. https://doi.org/10.1145/3197517.3201380

Michael D. McCool, Zheng Qin, and Tiberiu S. Popa. 2002. Shader Metaprogramming. In Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware (HWWS ’02). 57–68. http://dl.acm.org/citation.cfm?id=569046.569055

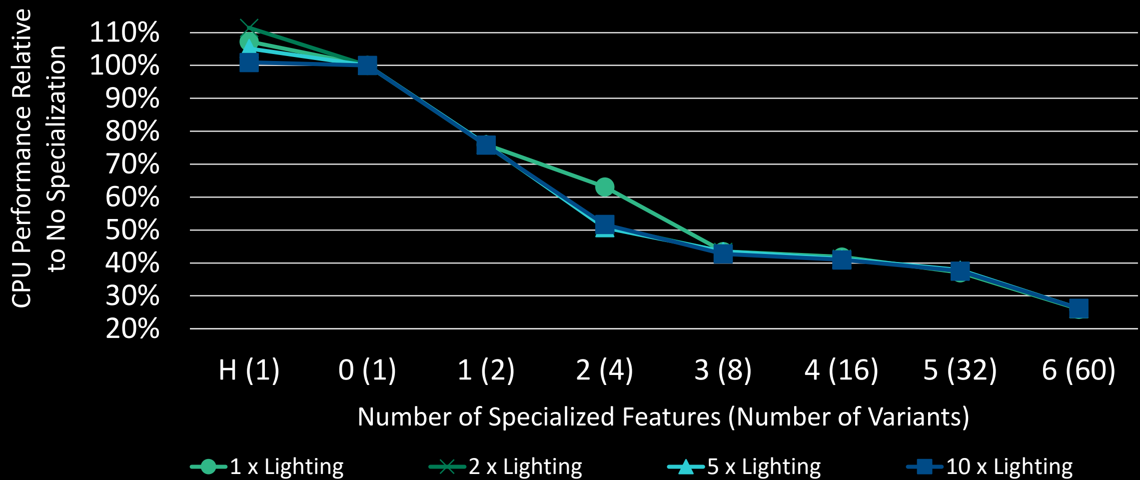
References (cont.)

Herb Sutter. 2018. Metaclasses: Generative C++. C++ Standards Committee Papers. https://wg21.link/P0707

Walid Mohamed Taha. 1999. Multistage Programming: Its Theory and Applications. Ph.D. Dissertation. Oregon Graduate Institute of Science and Technology.

Natalya Tatarchuk and Chris Tchou. 2017. Destiny Shader Pipeline. Game Developers Conference 2017. http://advances.realtimerendering.com/destiny/gdc_2017/

Extra Slides



More specialization means more variants. So, as we expected, there is more CPU overhead needed to bind the variants and dispatch tiles. So, CPU performance decreases with increased specialization.