

Aufbau einer modernen Rendering-Pipeline

Proseminar-Ausarbeitung von

Jonas Heinle

An der Fakultät für Informatik
Institut für Visualisierung und Datenanalyse,
Lehrstuhl für Computergrafik

27. Mai 2019

Inhaltsverzeichnis

1	Abstract	1
2	Einleitung	2
3	Moderne Rendering-Pipeline	3
3.1	Anwendung	3
3.2	Geometrie Stufe	3
3.2.1	Vertex Shader	3
3.2.2	Primitive Assembly	3
3.2.3	Tessellation	4
3.2.4	Geometry Shader	4
3.2.5	Clipping	5
3.2.6	Viewport Transform	5
3.3	Rasterisierung	5
3.4	Per Fragment Operations	5
3.4.1	Multisample Fragment Ops	5
3.4.2	Stencil Test	5
3.4.3	Occlusion Query	6
3.4.4	Blending	6
3.4.5	Logical Operations	6
3.5	Framebuffer und Buffer Objekte	6
3.6	GPU Memory	6
3.7	Compute Shader	6
4	Unterschied moderne und klassische Rendering-Pipeline	7
4.1	Freie Programmierung oder reines Konfigurieren	7
4.2	Vertex Arrays, Index Buffers,	7
5	Ausblick	8
5.1	Raytracing Unterstützung	8
5.2	Task-/Mesh Shaders	9
	Literaturverzeichnis	10

1. Abstract

In unserer heutigen hoch technologisierten Welt steigt stetig und rasant die Leistungsfähigkeit moderner Grafikhardware. Um heutzutage Grafik auf einem modernen Computer darzustellen sind mittlerweile eine Vielzahl von Zwischenschritten nötig. Diese Arbeit beschäftigt sich um die einzelnen Stufen dieser Kette, deren jeweilige Aufgabe, ihren Datentransfer untereinander, ihre Reihenfolge sowie ihren Arbeitskontext. Beim Arbeitskontext werden unter anderem die unterschiedlichen Koordinatensysteme untersucht. Dabei soll nicht nur konkret auf die Arbeitsweise der einzelnen Stufe eingegangen werden, sondern auch im Speziellen auf die Zusammenarbeit und Kommunikation. Exemplarische Fragestellungen die behandelt werden sind folgende: Können in dieser Art der Abarbeitung Flaschenhälse entstehen und wie werden sie umgangen bzw. bekämpft. Welchen Einfluss hat der Programmierer auf die Pipeline bzw. welche Schritte kann er selber implementieren und welche Schritte werden rein von der Hardware übernommen und können nicht von ihm modifiziert werden. Diese Arbeit macht den Aufbau einer modernen Rendering-Pipeline verständlich und vermittelt den Begriff Rasterisierung. Zusätzlich wird es bei dieser Abhandlung, dank des stetigen technologischen Fortschritts, ein Ausblick auf zukünftige Entwicklungen und Neuerungen gegeben, die zukünftig moderne Rendering-Pipelines beherrschen wird.

2. Einleitung

Wollen wir mit einer Anwendung 3.1 eine Szene von Objekten, z.B. eine Teekanne in einem Stadion, auf einem Computer darstellen, so liegt das Objekt zuerst in Form von vielen Eckpunkten(Dreiecken, Linien) in der Anwendung vor. Dabei befinden wir uns im "Model Space". Wollen wir die Teekanne innerhalb unserer Welt (das Stadion) bewegen, wenden wir auf jede Koordinate des Modells(gemeint sind alle mit dem Modell assoziierten Eckpunkte, Normalen) den "Model Transform" an und gehen somit in den "World Space". Die geschieht meist in der Anwendung bevor die Geometrie zur Geometriestufe 3.2 geschickt wird.

Als Anwender können wir nun zu Beginn der Geometriestufe mit einem eigenen Vertex Shader die Beleuchtungsberechnung etc. der Eckpunkte bestimmen. Dabei helfen die von der Anwendung mitgegebenen Attribute. Auch in welchen Koordinatensystem dies geschieht bleibt dem Programmierer überlassen. In der anschließenden Stufe werden diese Eckpunkte zu Primitiven zusammengesetzt und können weiter unterteilt werden 3.2.3. Die Primitive werden von dem Tessellation- zum Geometry Shader durchgereicht, welcher diese vervielfachen, entfernen oder beliebig verändern kann..

Um eine (orthographische oder auch perspektivische) Projektion vorzubereiten, wenden wir einen "View Transform" an. Damit können wir in den "View Space" gelangen und vereinfachen die darauffolgende Projektion. Egal welche Art von Projektion wir angewendet haben, unsere Szene mit der Teekanne im Stadion liegt nun in einem Einheitswürfel.

Nun kann die Teekanne derart platziert sein, dass Teile außerhalb unseres Sichtfeldes(frustum) liegen. Aufgabe des Clipping 3.2.5 ist nun das "Frustum Culling", also das Abschneiden des nicht sichtbaren Bereichs, um weitere Berechnungen innerhalb darauffolgender Stufen zu beschleunigen. Der Entwickler hat keinen Einfluss auf das Clipping 3.2.5 (Abschneiden nicht relevanter Szenengeometrie). Ein Algorithmus der dieses Problem löst, ist der Cohen-Sutherland-Algorithmus.

Nach dem Beschneiden liegen nur noch die sichtbaren "Primitive" vor und gelangen in die nächste Pipelinestufe, dem Viewport Transform 3.2.6. Die x- und y-Komponente unserer Eckpunktpositionen werden auf die Bildschirmgröße skaliert und bilden die schlussendliche Position des Vertex. Dabei ist es wichtig anzumerken, dass wir die z-Komponente für weitere Berechnungen in der Rasterisierung 3.3 speichern.

Nun sind wir an dem Punkt angelangt, an dem wir unsere transformierten, projizierten, sichtbaren Eckpunkte mit ihren Shadinginformationen vorliegen haben (Außerdem auch Tiefenwert!). Ziel der Rasterisierung ist es nun jeden einzelnen Bildschirmpixel mit diesen Informationen einzufärben. Vor dem endgültigen Darstellen der Bildschirmpixel lassen sich auf den zuvor erzeugten Fragmenten eine Vielzahl von Operationen 3.4 ausführen. So haben wir nun das finale Bild produziert.

Besonders interessant sind die neuen Shader für Ray Tracing. Die neuen Einheiten bilden eine Ergänzung der bisherigen Rasterisierung 5.1.

3. Moderne Rendering-Pipeline

3.1 Anwendung

Zu rendernde Objekte werden zunächst von der Anwendung zur Grafikkarte geschickt. Dabei liegen die Daten über ein Objekt beispielhaft im OBJ-Dateiformat vor. Die Informationen über einen Vertex, das sind die Position, Normale und Texturkoordinaten sind für Berechnungen der nächsten Stufe, des Vertex Shaders 3.2.1, wichtig. Desweiteren werden Kamera- und Lichtposition in die Geometriestufe 3.2 übergeben. In der Anwendung können Informationen über ein Objekt geupdatet werden, z.B. bei Kollisionen von Objekten bei denen sich Positionen verändern. Allgemein werden hier Benutzereingaben gesammelt und verarbeitet. Als programmierbare Stufe dieser Pipeline kann der Anwender konkrete Verbesserungen in diese Stufe einbringen um z.B. mit einem BVH die Anzahl der Primitive für die Geometriestufe zu verringern.

```
4 newmtl Material
5 Ns 96.078431
6 Ka 1.000000 1.000000 1.000000
7 Kd 0.640000 0.640000 0.640000
8 Ks 0.500000 0.500000 0.500000
9 Ke 0.000000 0.000000 0.000000
10 Ni 1.000000
11 d 1.000000
12 illum 2
13
```

Abbildung 3.1: Definition beispielhaftes Material

```
1 # Blender v2.79 (sub 0) OBJ File: ''
2 # www.blender.org
3 mtllib Suzanne.mtl
4 o Suzanne
5 v 0.437500 0.164062 0.765625
6 v -0.437500 0.164062 0.765625
7 v 0.500000 0.093750 0.687500
8 v -0.500000 0.093750 0.687500
9 v 0.546875 0.054688 0.578125
```

Abbildung 3.2: Vertex Daten im OBJ-Dateiformat

3.2 Geometrie Stufe

3.2.1 Vertex Shader

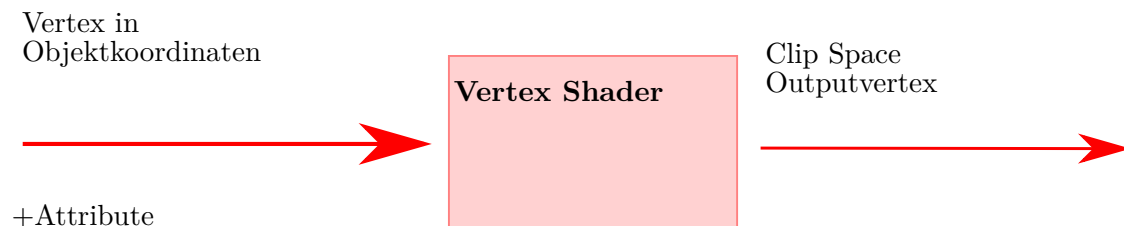


Abbildung 3.3: Funktionsweise Vertex Shader

Als freiprogrammierbare Stufe der Geometry Stage werden hier die ankommenden Vertices mit ihren Attributen (jedoch keine Erzeugung oder Löschung!!) und mit Hilfe von übergebenen Daten der Anwendung (Lichtposition, Transformationsmatrix) bearbeitet. Somit können Teile der Leuchtungsrechnung ("Gouraud shading") hier angewandt werden.

3.2.2 Primitive Assembly

Aufgabe ist das Zusammensetzen der Eckpunkte, welche in einem Strom aus dem Vertex Shader kommen. Auf diesen Primitiven (*Patches*, also Punkte, Linien, Drei- oder Vierecke) arbeiten die darauffolgenden Tessellation und Geometry Shader.

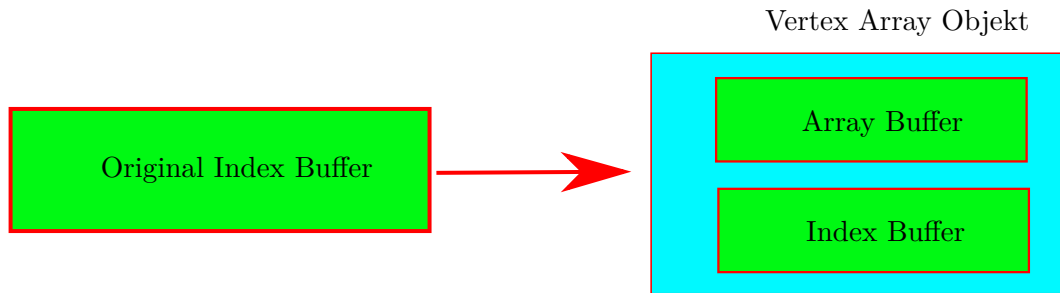


Abbildung 3.4: Buffer Objekte

3.2.3 Tessellation

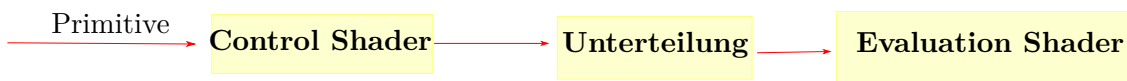


Abbildung 3.5: Funktionsweise Tessellation Shader

Ist eine optionale teil-programmierbare Stufe zur Geometriervervielfachung, d.h. eingehendes Primitiv wird weiter unterteilt. Der Tessellation Control Shader ist voll programmierbar und damit ist die Anzahl der entstehenden Primitive direkt beeinflussbar. Dabei lassen sich zwei jeweils benachbarte patches kontrolliert unterteilen. Der Fixed Function Tessellator nimmt die patches entgegen und unterteilt Sie auf die Art und Weise wie im Control Shader vorgegeben. Mit dem Tessellation Evaluation Shader kann man jeden Vertex der vom Unterteiler kommenden patches nachbearbeiten.

3.2.4 Geometry Shader

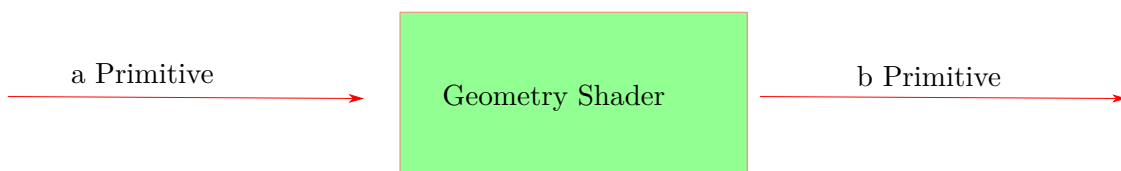


Abbildung 3.6: Funktionsweise Geometry Shader

Ist wiederum eine optionale frei-programmierbare Einheit die auf den zusammengesetzten Primitiven arbeitet. Eine Vervielfachung ist theoretisch möglich, sollte aber als Aufgabe im Tessellation Shader erbracht werden. Vielmehr können wir diesen Shader einsetzen, um eine Geometrie gleichzeitig auf mehrer rendertargets zu rendern oder wir machen transform Feedback: Buffer Objects dienen zur Speicherung von Informationen von eingehenden Primitiven. Diese können in weiteren Durchgängen wieder benutzt werden.

3.2.5 Clipping

Um Objekte bzw. Objektausschnitte, welche außerhalb des Sichtfensters liegen, für die Bildsynthese zu verwerfen kommt nun das Abschneiden. Ein Algorithmus der dies leistet ist von Sutherland-Hodgman. Außerdem gehen wir vom Clip-Space zum Window-Space.

3.2.6 Viewport Transform

Der auf die sichtbare Szenengeometrie reduzierte Einheitswürfel wird in diesen Schritt auf die Fenstergröße skaliert und zur Rasterisierung weitergegeben.

3.3 Rasterisierung

Zu Beginn dieser Stufe haben wir nun die Eckpunkte der bereits verarbeiteten, transformierten, projizierten Geometrie mit möglichen Beleuchtungsinformationen aus den vorherigen Stufen vorliegen. Zu den Window Coordinates wurde auch der Tiefenwert gespeichert. Mit der Rasterisierung wird nun die Farbe jedes einzelnen Pixels bestimmt. Es ist also die Aufgabe dieser Pipelinestufe herauszufinden welche Geometrie welchen Pixel zu welchen Anteil bedeckt und wie die Shading Informationen zur Farbgebung des Pixels beitragen.

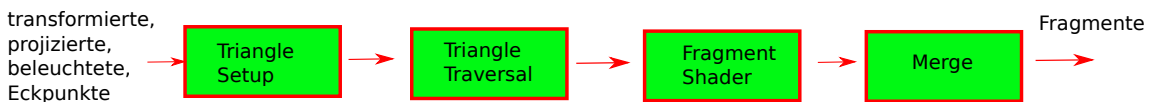


Abbildung 3.7: Ablauf der Rasterisierung

Zuallererst befindet sich die Geometrie im *Triangle Setup*. Unbeeinflussbar vom Programmierer werden hier Daten berechnet, welche zur Pixeleinfärbung benötigt werden. So werden viele zuvor per Vertex berechnete Werte interpoliert (Beleuchtung, Tiefe). Beim darauffolgenden *Triangle Traversal* werden die wichtigen Fragmente erzeugt. Dieser Schritt bestimmt diejenigen Pixel, welche innerhalb des Dreiecks liegen und erzeugt darauf hin die Fragmente für dieses Dreieck anhand der zuvor berechneten/interpolierten per Dreieck Informationen. Als freiprogrammierbare Shadereinheit können im *Fragment Shader* vom Programmierer weitere Berechnungen vorgenommen werden. Dazu zählt eine pro Pixel Beleuchtungsberechnung. Will man Texturen verwenden, so passiert das nun hier. Das abschließende *Merging* hat eine besondere Aufgabe beim Abspeichern der Farbe für jeden Pixel im color buffer. Da der buffer bereits von dem vorherigen Durchlauf Farbwerte beinhaltet ist es nun daran eine Kombination aus aktuellen Inhalt, sowie einkommende Farbe zu finden, welche wiederum im color buffer abgespeichert wird. Obwohl dieses Zusammenfügen nicht komplett freiprogrammierbar ist, hat der Programmierer dennoch etliche Konfigurationsmöglichkeiten die er vornehmen kann.

3.4 Per Fragment Operations

Auf den durch Rasterisierung entstandenen Fragmenten können noch einige Operationen vorgenommen werden.

3.4.1 Multisample Fragment Ops

3.4.2 Stencil Test

Damit können wir schlussendlich entscheiden, ob ein Fragment gezeichnet werden soll oder nicht durch den Vergleich mit einem Referenzwert und Stencil Buffer. (im Framebuffer).

3.4.3 Occlusion Query

3.4.4 Blending

Wir mischen Farbwerte die bereits in den Framebuffer geschrieben wurden mit den Farbwerten, welche in den erzeugten Fragmenten vorliegen. Die Art in der die beiden Farben gemischt werden bestimmt eine Funktion z.B Multiplikativ, Additiv...

3.4.5 Logical Operations

3.5 Framebuffer und Buffer Objekte

3.6 GPU Memory

3.7 Compute Shader

Losgelöst von der Grafikpipeline steht der Compute Shader. Als freiprogrammierbare Einheit lässt sie sich nicht nur für graphische Berechnungen in der Pipeline, sondern auch für stark parallelisierbare Probleme im Allgemeinen einsetzen. Im Folgenden wird auf die DirectX12-API [Dir19,] eingegangen. Der Compute Shader arbeitet auf einer Threadgruppe (≤ 1024). In einem dreidimensionalen Array organisiert, teilen sich die Threads einer Gruppe Speicher, über den Sie kommunizieren und im Gesamten Fortschritt machen können, d.h. alle Threads laufen nebenläufig. Mit der zusätzlichen Eigenschaft auf Buffer Objekte (Daten auf der GPU) zugreifen zu können, haben die Compute Shader beispielsweise im Post-Processing 3.4 Bedeutung. Als Threadgruppe greifen sie jeweils auf einen bestimmten Bereich des Framebuffers zu und können die nötigen Nachbarschaftsinformationen austauschen.

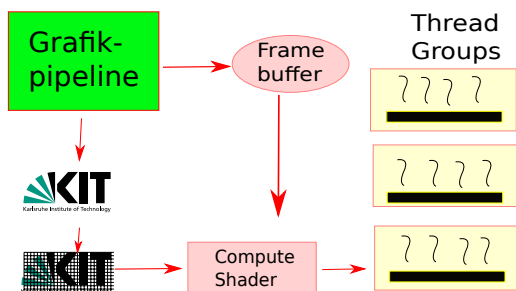


Abbildung 3.8: Funktionsweise Compute Shader

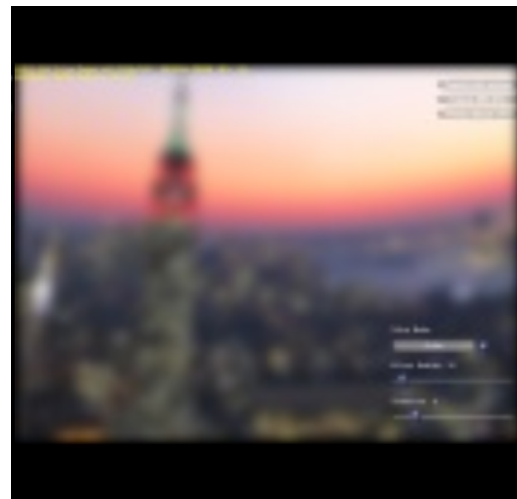


Abbildung 3.9: Postprocessing Blur mit Compute Shader [11]

Die erste Abbildung zeigt nochmal deutlich, dass der Compute Shader für jedwede Art der Berechnung eingesetzt werden kann, bei dem *stream processors* zu tragen kommen. Eine Beobachtung war die effiziente Verlegung der Mesh Verarbeitung auf den Compute Shader. Dies führte zu der Einführung von Mesh Shadern 5.2, die stark an dem Aufbau von Compute Shadern orientiert sind.

4. Unterschied moderne und klassische Rendering-Pipeline

4.1 Freie Programmierung oder reines Konfigurieren

Abbildung 4.1: OpenGL \leq 2.0

Abbildung 4.2: Modernes OpenGL

4.2 Vertex Arrays, Index Buffers, ..

5. Ausblick

5.1 Raytracing Unterstützung

In heutigen modernen Grafikprogrammierschnittstellen(Vulkan, DirectX) befindet sich Raytracing-Funktionalität. Raytracing Ansätze gehen von der objektbasierten (siehe Rasterisierung) zu der Image-ordered Bilderstellung über. Hiermit werden nicht nur Primärsondern auch Sekundärstrahlen(etc.) betrachtet und somit unter Anderem Schatten und Spiegelungen. Wir können diese neuen Shader wie Compute Shader 3.7 programmieren.

Shader	Beschreibung
Any Hit	Kann ausgeführt werden wenn wir einen Schnitt mit unserem Strahl gefunden haben. Soll bei transparenten Objekten dazu führen, früher abbrechen zu können.
Callable	Kann von einem anderen Shader aufgerufen werden.
Closest Hit	Wird ausgeführt wenn ein Standardstrahl von einem Ray Generation Shader verschossen wurde, und ein "nächster Schnittpunkt"gefunden wurde. Hier kann der Programmierer die Methode <i>shade()</i> implementieren. Dabei sollte auf Schatten getestet, weitere Reflektions- und Transmission miteinberechnet werden
Intersection	Wird aufgerufen, falls wir eine Bounding Box unser Beschleunigungsstruktur treffen.
Miss	Wird aufgerufen, falls keine Szenengeometrie getroffen wurde.
Ray Generation	Mit <i>TraceRay()</i> können wir Strahlen verschießen. Normalerweise versendet man einen Strahl pro Pixel. Wir können jedwedigen Typ von Strahl damit verschießen(Schatten, Primär, Transmit)

Tabelle 5.1: Shadereinheiten nach DirectX12.

Hierbei führt ein verschossener Strahl eine Datenstruktur("*payload*") mit sich, die eine variable Menge an Informationen speichern kann. Als wichtigste zu nennen ist die Distanz bei einem Hit.

Aktuelle Bemühungen gehen nun daran Raytracing und Rasterisierung zu kombinieren. Barré-Brisebois [BBHW⁺19] stellte mit dem Spiel *PICA PICA* eine solche Rendering-Pipeline vor, welche mithilfe von Path Tracing arbeitet. Dabei wird der G-Buffer (Texturen die Position, Normalen, Belichtung eines Frames speichern) noch über Rasterisierung berechnet. Direkten Schatten kann man rastern oder raytracen. Diese Option verspricht eine Anpassungsfähigkeit der Pipeline nach Leistungsfähigkeit. Ähnlich können nun Reflektionen, Global Illumination, Ambient Occlusion und Transmission geraytraced oder auf Compute Shader ausgeführt werden.(Wieder je nach Hardwareleistung). Einzig direkte Beleuchtung sowie Post-Processing Effekte laufen nur über Compute-Shader.

5.2 Task-/Mesh Shaders

Der aktuelle Trend bei der Modellerstellung geht über zu immer komplexeren Modellen. Dabei stößt die traditionelle Pipeline mit der enormen Anzahl von Dreiecken an ihre Grenzen. Die damalige Einführung des Tessellation 3.2.3- und des Geometryshaders 3.2.4 (vor allem aber Tessellation!) war aus selbigen Grund und brachte eine erwünschte Weitung des Flaschenhalses und bessere Arbeitsaufteilung innerhalb der Pipeline. Das obige besprochene Primitive Assembly 3.2.2 erweist sich als *Fixed Function* innerhalb der Pipeline als zu star und unflexibel. Für starre Szenen wird hohe Bandbreite veranschlagt, da wir unseren Index Buffer jedesmal berechnen müssen. Anstatt wie beim Primitive Assembly 3.2.2 von der Hardware das Vertex Array Objekt aufs neue Berechnen zu lassen, teilt man nun die Geometrie selbst in *Meshlets* auf und speichert diese in einen Buffer (*Meshlet Desc Buffer*), wobei jedes *Meshlet* für sich auf einen VBO zeigt (Man hält den Buffer *On-Chip*; ähnlich Compute Shader Model 3.7). Diese Aufteilung gibt uns ein besseres Wiederverwenden der Eckpunkte und daher wird Bandbreite gespart! Die Mesh Shader kommen nach dem Compute Shader Model 3.7 und lassen eindeutig den Trend von fixen Vorgängen innerhalb der Pipeline zu mehr Flexibleren wiedererkennen. Durch diese freie Programmiermöglichkeit ergeben sich vielfältige Möglichkeiten, so z.B. für *Mesh compression* da sich die Buffer selber beschreiben lassen.

MESHLETS

TRADITIONAL PIPELINE



TASK/MESH PIPELINE

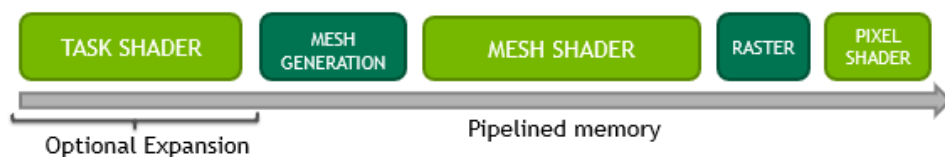


Abbildung 5.1: Vergleich bisherige Vertex, Tessellation, Geometry Pipeline <-> stream processor artig Mesh Shader Pipeline

Dabei arbeitet der Task Shader ähnlich der Tessellation Control Einheit 3.2.3 und verteilt zunächst das Mesh in *workgroups*. Jedoch ist anzumerken, dass der Task Shader mit einem kooperativen Threadmodell arbeitet und dessen In-/Output komplett frei programmierbar ist. Man ist nicht mehr auf ein ankommendes Primitiv und eine ausgehende Tessellation Entscheidung beschränkt. Damit lassen sich Level Of Detail Entscheidungen treffen, also wie fein unterteile ich mein einkommendes Primitiv. Oder man entscheidet sogar eine Gruppe gar nicht zu zeichnen (Culling) und schickt keine weitere Arbeit weiter. Der Mesh Shader nimmt lediglich die *workgroup id* entgegen und arbeitet ähnlich dem Compute Shader 3.7 auf einer Threadgruppe (hier langt jedoch ein eindimensionales Array). Es gibt hier jedoch nicht nur einen gemeinsamen Speicher, sondern auch einen Output, welcher frei-programmierbar ist. Zum Geometry Shader, der keine Threadgruppe besitzt, ist dies ein großer Schritt.

Literaturverzeichnis

- [11] NVIDIA SDK 11. <https://developer.nvidia.com/dx11-samples>.
- [ABB18] Johan Andersson und Colin Barré-Brisebois: *Shiny Pixels and Beyond: Real-Time Raytracing at SEED*. In: *NVIDIA Sponsored Session, Game Developers Conference*, 2018.
- [AMHH18] Tomas Akenine-Moller, Eric Haines und Naty Hoffman: *Real-time rendering*. AK Peters/CRC Press, 2018.
- [BBHW⁺19] Colin Barré-Brisebois, Henrik Halén, Graham Wihlidal, Andrew Lauritzen, Jasper Bekkers, Tomasz Stachowiak und Johan Andersson: *Hybrid Rendering for Real-Time Ray Tracing*, Seiten 437–473. Apress, Berkeley, CA, 2019, ISBN 978-1-4842-4427-2. https://doi.org/10.1007/978-1-4842-4427-2_25.
- [Boy08] Chas Boyd: *The DirectX 11 compute shader*. ACM SIGGRAPH. Cité page, 25, 2008. <http://s08.idav.ucdavis.edu/boyd-dx11-compute-shader.pdf>.
- [Dac18] Prof. Dr. Ing. Carsten Dachsbacher: *Vorlesung*. Internet, 2018. https://cg.ivd.kit.edu/lehre/ws2019/index_2114.php.
- [Dir19] *DirectX12API*. Internet, May 2019. <https://docs.microsoft.com/en-us/windows/desktop/direct3d12/direct3d-12-raytracing>.
- [F⁺04] Randima Fernando *et al.*: *GPU gems: programming techniques, tips, and tricks for real-time graphics*, Band 590. Addison-Wesley Reading, 2004.
- [KWM16] David B Kirk und W Hwu Wen-Mei: *Programming massively parallel processors: a hands-on approach*. Morgan kaufmann, 2016.
- [MS15] Steve Marschner und Peter Shirley: *Fundamentals of computer graphics*. CRC Press, 2015.
- [Ni09] Tianyun Ni: *Direct Compute: Bring GPU computing to the mainstream*. In: *GPU Technology Conference*, Seite 23, 2009. <http://on-demand.gputechconf.com/gtc/2009/presentations/1015-Features-Advantages-DirectCompute.pdf>.
- [tas19] *Mesh Shader*. Internet, May 2019. <https://devblogs.nvidia.com/introduction-turing-mesh-shaders/>.
- [Vul19] *VulkanAPI*. Internet, May 2019. <https://www.khronos.org/vulkan/>.

Erklärung

Ich versichere, dass ich die Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet habe. Die Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt und von dieser als Teil einer Prüfungsleistung angenommen.

Karlsruhe, den 27. Mai 2019

(Jonas Heinle)