(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7650983号
(P7650983)

(45)発行日　令和7年3月25日(2025.3.25)

(24)登録日　令和7年3月14日(2025.3.14)

(51)国際特許分類

		FI		
G06T 15/00 (2011.01)		G06T	15/00	501
G06F 12/0875(2016.01)		G06F	12/0875	106
G06F 12/0895(2016.01)		G06F	12/0895	118
G06F 9/38 (2018.01)		G06F	9/38	310J

請求項の数　15　（全15頁）

(21)出願番号　特願2023-539265(P2023-539265)
(86)(22)出願日　令和3年12月22日(2021.12.22)
(65)公表番号　特表2024-501015(P2024-501015A)
(43)公表日　令和6年1月10日(2024.1.10)
(86)国際出願番号　PCT/US2021/064797
(87)国際公開番号　WO2022/146810
(87)国際公開日　令和4年7月7日(2022.7.7)
審査請求日　令和6年12月13日(2024.12.13)
(31)優先権主張番号　17/134,790
(32)優先日　令和2年12月28日(2020.12.28)
(33)優先権主張国・地域又は機関
米国(US)
早期審査対象出願

(73)特許権者　591016172
アドバンスト・マイクロ・ディバイシズ・インコーポレイテッド
ＡＤＶＡＮＣＥＤ ＭＩＣＲＯ ＤＥＶＩＣＥＳ ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国 ９５０５４ カリフォルニア州、 サンタ クララ、オーガスティン ドライブ ２４８５
(73)特許権者　503447036
サムスン エレクトロニクス カンパニー リミテッド
大韓民国・１６６７７・キョンギード・スウォン－シ・ヨントン－ク・サムスン－ロ・１２９
(74)代理人　100108833

最終頁に続く

(54)【発明の名称】　キャッシュラインに対するミス要求の選択的生成

(57)【特許請求の範囲】

【請求項１】

装置であって、

複数のサブセットに分割されたキャッシュラインを含むテクスチャキャッシュと、

グラフィックスパイプライン内の少なくとも１つの計算ユニットと、を備え、

前記計算ユニットは、前記テクスチャキャッシュ内のキャッシュラインの複数のサブセットのうち第１のサブセットに関連付けられたアドレスへのメモリアクセス要求に対するキャッシュミスと、色圧縮又は深度圧縮の何れが有効であるかを示す、前記キャッシュラインに記憶されたデータの特性と、に応じて、前記第１のサブセットに対するミス要求を選択的に生成するように構成されている、

装置。

【請求項２】

前記少なくとも１つの計算ユニットは、メモリアクセス要求に関連付けられたキャッシュミスが、前記複数のサブセットのうち前記第１のサブセットのみにマッピングするか、又は、前記複数のサブセットのうち前記第１のサブセットに追加された又は前記第１のサブセット以外の１つ以上のサブセットにマッピングするかを判定するように構成されている、

請求項１の装置。

【請求項３】

前記少なくとも１つの計算ユニットは、前記キャッシュミスが、前記複数のサブセット

のうち前記第１のサブセットに追加された又は前記第１のサブセット以外のサブセットにマッピングすることに応じて、フルキャッシュラインに対するミス要求を生成するように構成されている、
請求項２の装置。
【請求項４】
前記少なくとも１つの計算ユニットは、前記メモリアクセス要求が、前記第１のサブセットのみにマッピングすることに応じて、色圧縮及び深度圧縮のうち少なくとも１つがテクスチャデータに対して有効にされているか否かを判定するように構成されている、
請求項３の装置。
【請求項５】
前記少なくとも１つの計算ユニットは、色圧縮及び深度圧縮のうち少なくとも１つが前記テクスチャデータに対して有効にされていることに応じて、前記フルキャッシュラインに対するミス要求を生成するように構成されている、
請求項４の装置。
【請求項６】
前記少なくとも１つの計算ユニットは、色圧縮及び深度圧縮のうち少なくとも１つが前記テクスチャデータに対して有効にされていないことに応じて、前記キャッシュラインの前記第１のサブセットに対するミス要求を生成するように構成されている、
請求項４の装置。
【請求項７】
前記少なくとも１つの計算ユニットは、前記メモリアクセス要求の時間的局所性及び空間的局所性のうち少なくとも１つに基づいて、前記第１のサブセット又は前記複数のサブセットに対するミス要求を選択的に生成するように構成されている、
請求項１の装置。
【請求項８】
前記少なくとも１つの計算ユニットは、メモリアクセス要求シーケンスが、前記複数のサブセットにアクセスすることが予想されることに応じて、前記第１のサブセットにおけるキャッシュミスに応じて、前記複数のサブセットに対するミス要求を生成するように構成されている、
請求項７の装置。
【請求項９】
前記少なくとも１つの計算ユニットは、前記第１のサブセット内のキャッシュミスに応じて、且つ、メモリアクセス要求シーケンスが、閾値を上回る空間的局所性を有することに応じて、前記複数のサブセットに対するミス要求を生成するように構成されている、
請求項７の装置。
【請求項１０】
前記少なくとも１つの計算ユニットは、前記メモリアクセス要求が前記閾値を下回る空間的局所性を有することに応じて、前記第１のサブセットに対するミス要求を生成するように構成されている、
請求項９の装置。
【請求項１１】
方法であって、
複数のサブセットに分割されたキャッシュラインを含むテクスチャキャッシュ内のキャッシュラインに対するミス要求を検出することと、
前記キャッシュラインの第１のサブセットに関連付けられたアドレスに対するキャッシュミスと、色圧縮又は深度圧縮の何れが有効であるかを示す、前記キャッシュラインに記憶されたデータの特性と、に応じて、前記第１のサブセットに対するミス要求を選択的に生成することと、を含む、
方法。
【請求項１２】

メモリアクセス要求に関連付けられたキャッシュミスが、前記複数のサブセットのうち前記第１のサブセットのみにマッピングするか、又は、前記複数のサブセットのうち前記第１のサブセットに追加された又は前記第１のサブセット以外の１つ以上のサブセットにマッピングするかを判定することを含む、
請求項１１の方法。
【請求項１３】
前記キャッシュミスが前記複数のサブセットにマッピングすることに応じて、フルキャッシュラインに対するミス要求を生成することを含む、
請求項１２の方法。
【請求項１４】
色圧縮及び深度圧縮のうち少なくとも１つがテクスチャデータに対して有効にされているか否かを判定することを含む、
請求項１３の方法。
【請求項１５】
前記ミス要求を選択的に生成することは、色圧縮及び深度圧縮のうち少なくとも１つが前記テクスチャデータに対して有効にされていることに応じて、前記複数のサブセットに対するミス要求を生成することを含む、
請求項１４の方法。
【発明の詳細な説明】
【背景技術】
【０００１】
グラフィックス処理ユニット（Graphics Processing Unit、ＧＰＵ）は、プログラマブルシェーダ及び固定機能ハードウェアブロックシーケンスで形成されるグラフィックスパイプラインを使用して三次元（three-dimensional、３Ｄ）グラフィックスを処理する。例えば、フレーム内で見えるオブジェクトの３Ｄモデルは、三角形、他の多角形又はパッチのセットによって表すことができ、これらはグラフィックスパイプラインで処理され、ユーザに表示するためのピクセルの値を生成する。三角形、他の多角形又はパッチは、まとめてプリミティブと呼ばれる。レンダリングプロセスは、テクスチャをプリミティブにマッピングして、プリミティブの解像度よりも高い解像度を有する視覚的詳細を組み込むことを含む。ＧＰＵは、グラフィックスパイプラインにおいて処理されているプリミティブにマッピングするためにテクスチャ値が利用可能であるように、テクスチャ値を記憶するために使用される専用メモリを含む。テクスチャは、ディスク上に記憶することができ、又は、グラフィックスパイプラインによって必要とされる場合に手続き的に生成することができる。専用ＧＰＵメモリに記憶されたテクスチャデータは、ディスクからテクスチャをロードすることによって又はデータを手続き的に生成することによってポピュレートされる。頻繁に使用されるテクスチャデータは、シェーダ又は固定機能ハードウェアブロックによってアクセスされる１つ以上のテクスチャキャッシュにキャッシュされる。
【０００２】
本開示は、添付の図面を参照することによってより良好に理解され、その多くの特徴及び利点が当業者に明らかになる。異なる図面における同じ符号の使用は、類似又は同一のアイテムを示す。
【図面の簡単な説明】
【０００３】
【図１】いくつかの実施形態による、キャッシュラインの部分に対するミス要求を選択的に生成する処理システムのブロック図である。
【図２】いくつかの実施形態による、高次ジオメトリプリミティブを処理して、所定の解像度で三次元（３Ｄ）シーンのラスタ化された画像を生成するように構成されたグラフィックスパイプラインを示す図である。
【図３】いくつかの実施形態による、第１の読み取りサイクルにおいて複数のセクタにわたって分散された要求と、第２の読み取りサイクルにおいて単一のセクタに制約された要

求と、を有するキャッシュラインのブロック図である。
【図４】いくつかの実施形態による、第１の読み取りサイクル及び第２の読み取りサイクル中に高度の時間的局所性を示さない要求を有するキャッシュラインのブロック図である。
【図５】いくつかの実施形態による、キャッシュラインの部分に対するミス要求を選択的に生成する方法のフロー図である。
【発明を実施するための形態】
【０００４】
テクスチャキャッシュ内のキャッシュラインは、通常、大量のデータを保持するように構成され、例えば、テクスチャキャッシュラインの幅は、１２８バイト又は１０２４（１Ｋ）ビット程度とすることができる。広いキャッシュラインは、グラフィックス処理の特性である大きな及び／又は可変サイズのデータブロックのキャッシュを容易にする。テクスチャデータは、４×４ピクセルフットプリント又は８×８ピクセルフットプリントを有するタイル等のタイルに記憶される。タイルのサイズは、テクスチャフォーマットにも依存し、テクスチャフォーマットは、８ビットフォーマット、３２ビットフォーマット、１２８ビットフォーマット等のように、各ピクセルを表すために使用されるビット数を示す。したがって、８×８ピクセルフットプリントを有するタイルは、テクスチャフォーマットに応じて、５２６ビット、２０４８ビット、８１９２ビット又は他のビット数で表すことができる。動作中、テクスチャキャッシュは、サイクルごとに最大Ｎ個のメモリアクセス要求（例えば、読み取り要求又は書き込み要求）を受信し、ここで、Ｎはベクトルのサイズであり（例えば、ベクトルサイズは６４、３２又は１６とすることができる）、各キャッシュミスは、より高レベルのキャッシュ又はメモリからキャッシュラインを取り出す要求を生成する。キャッシュラインのサイズが大きいと仮定すると、要求されたデータが複数のキャッシュラインにわたって分散される場合、キャッシュミス要求は、元のアクセス要求内のデータ量にかかわらず、かなりのメモリ帯域幅を消費する。更に、全ての要求サイクルについてフルキャッシュラインを有効にすることは、グラフィックスパイプラインによって使用されているデータを記憶するために必要とされないキャッシュの部分を無効にすることによって、電力節約の機会を制限する。
【０００５】
図１～図５は、キャッシュラインのサブセットに関連付けられたアドレスへのメモリアクセス要求に対するキャッシュミスに応じて、テクスチャキャッシュ内のキャッシュラインのサブセットに対するミス要求を選択的に生成することによって、テクスチャキャッシュとシステムメモリ（又はより高レベルのキャッシュ）との間のメモリ帯域幅を節約しながら、電力消費を潜在的に低減するためのシステム及び技術を開示する。いくつかの実施形態では、キャッシュラインは２つ以上のセクタに分割される。フルキャッシュラインに対するミス要求は、例えば、メモリアクセス要求内のアドレスに基づいて、キャッシュライン内の全てのセクタにマッピングするメモリアクセス要求（読み取り要求等）によるキャッシュミスに応じて生成される。メモリアクセス要求がキャッシュラインの単一のセクタにマッピングされる場合、ミス要求は、テクスチャデータの１つ以上のヒューリスティック又は特性の評価に基づいて、フルキャッシュライン又はキャッシュラインのセクタのうち何れかに対して選択的に生成される。例えば、テクスチャデータに対して色圧縮又は深度圧縮が有効にされている場合、ミス要求がフルキャッシュラインに対して生成される。テクスチャデータに対して圧縮が有効にされていない場合、ミス要求は、メモリアクセス要求によって示されるキャッシュラインのセクタに対してのみ生成される。また、ミス要求は、メモリアクセス要求の時間的局所性に基づいてキャッシュラインのサブセットに対して選択的に生成される。例えば、メモリアクセス要求シーケンスがキャッシュラインの異なるセクタにアクセスすると予想される場合、何れかのセクタにおけるキャッシュミスに応じて、フルキャッシュラインに対するミス要求が生成される。また、ミス要求は、メモリアクセス要求の空間的局所性に基づいてキャッシュラインのサブセットに対して選択的に生成される。例えば、メモリアクセス要求シーケンスが、隣接する、近接する又は近くのアドレスにアクセスすることが予想される場合、ミス要求は、何れかのセクタにお

けるミスに応じて、フルキャッシュラインについて生成される。対照的に、メモリアクセス要求のアドレスが分散しており、低い空間的局所性を有する場合、ミス要求は、キャッシュミスを含むキャッシュラインのセクタに対してのみ生成される。

【０００６】

図１は、いくつかの実施形態による、キャッシュラインの部分に対するミス要求を選択的に生成する処理システム１００のブロック図である。処理システム１００は、ダイナミックランダムアクセスメモリ（Dynamic Random-Access Memory、ＤＲＡＭ）等の非一時的なコンピュータ可読記憶媒体を使用して実装されるメモリ１０５又は他の記憶コンポーネントを含むか又はそれらへのアクセスを有する。しかしながら、場合によっては、メモリ１０５は、スタティックランダムアクセスメモリ（Static Random-Access Memory、ＳＲＡＭ）、不揮発性ＲＡＭ等を含む他のタイプのメモリを使用して実装することもできる。メモリ１０５は、処理システム１００において実装される処理ユニットの外部に実装されるために外部メモリと呼ばれる。また、処理システム１００は、メモリ１０５等のように、処理システム１００において実装されるエンティティ間の通信をサポートするためのバス１１０を含む。処理システム１００のいくつかの実施形態は、他のバス、ブリッジ、スイッチ、ルータ等を含むが、これらは明確にするために図１には示されていない。

【０００７】

本明細書で説明される技術は、様々な実施形態では、様々な並列プロセッサ、例えば、ベクトルプロセッサ、グラフィックス処理ユニット（ＧＰＵ）、汎用ＧＰＵ（ＧＰＧＰＵ）、非スカラプロセッサ、高並列プロセッサ、人工知能（ＡＩ）プロセッサ、推論エンジン、機械学習プロセッサ、他のマルチスレッド処理ユニット等の何れかで利用される。図１は、いくつかの実施形態による、並列プロセッサ、特に、グラフィックス処理ユニット（ＧＰＵ）１１５の一例を示す。グラフィックス処理ユニット（ＧＰＵ）１１５は、ディスプレイ１２０上に提示するための画像をレンダリングする。例えば、ＧＰＵ１１５は、オブジェクトをレンダリングして、ディスプレイ１２０に提供されるピクセルの値を生成し、ディスプレイ１２０は、ピクセル値を使用して、レンダリングされたオブジェクトを表す画像を表示する。ＧＰＵ１１５は、命令を同時に又は並列に実行する複数の計算ユニット（ＣＵ）１２１、１２２、１２３（本明細書ではまとめて「計算ユニット１２１～１２３」と呼ぶ）を実装する。いくつかの実施形態では、計算ユニット１２１～１２３は、１つ以上の単一命令複数データ（ＳＩＭＤ）ユニットを含み、計算ユニット１２１～１２３は、ワークグループプロセッサ、シェーダアレイ、シェーダエンジン等に集約される。ＧＰＵ１１５において実装される計算ユニット１２１～１２３の数は、設計上の選択の問題であり、ＧＰＵ１１５のいくつかの実施形態は、図１に示されるよりも多い又は少ない計算ユニットを含む。計算ユニット１２１～１２３は、本明細書で説明するように、グラフィックスパイプラインを実装するために使用することができる。ＧＰＵ１１５のいくつかの実施形態は、汎用コンピューティングのために使用される。ＧＰＵ１１５は、メモリ１０５に記憶されたプログラムコード１２５等の命令を実行し、ＧＰＵ１１５は、実行された命令の結果等の情報をメモリ１０５に記憶する。

【０００８】

また、処理システム１００は、バス１１０に接続され、したがってバス１１０を介してＧＰＵ１１５及びメモリ１０５と通信する中央処理装置（Central Processing Unit、ＣＰＵ）１３０を含む。ＣＰＵ１３０は、命令を同時に又は並列に実行する複数のプロセッサコア１３１、１３２、１３３（本明細書ではまとめて「プロセッサコア１３１～１３３」と呼ぶ）を実装する。ＣＰＵ１３０において実装されるプロセッサコア１３１～１３３の数は、設計上の選択の問題であり、いくつかの実施形態は、図１に示されるよりも多い又は少ないプロセッサコアを含む。プロセッサコア１３１～１３３は、メモリ１０５に記憶されたプログラムコード１３５等の命令を実行し、ＣＰＵ１３０は、実行された命令の結果等の情報をメモリ１０５に記憶する。また、ＣＰＵ１３０は、ＧＰＵ１１５にドローコールを発行することによって、グラフィックス処理を開始することができる。ＣＰＵ１３０のいくつかの実施形態は、同時に又は並列に命令を独立して実行する複数のプロセッ

サコア（明確化のために図１には示さず）を含む。
【０００９】
入力／出力（Input/Output、Ｉ／Ｏ）エンジン１４５は、ディスプレイ１２０と関連付けられた入力又は出力動作、及び、キーボード、マウス、プリンタ、外部ディスク等のような処理システム１００の他の要素を扱う。Ｉ／Ｏエンジン１４５は、Ｉ／Ｏエンジン１４５がメモリ１０５、ＧＰＵ１１５又はＣＰＵ１３０と通信するようにバス１１０に結合される。図示される実施形態では、Ｉ／Ｏエンジン１４５は、コンパクトディスク（Compact Disk、ＣＤ）、デジタルビデオディスク（Digital Video Disc、ＤＶＤ）等の非一時的なコンピュータ可読記憶媒体を使用して実装される、外部記憶コンポーネント１５０上に記憶される情報を読み取る。また、Ｉ／Ｏエンジン１４５は、ＧＰＵ１１５又はＣＰＵ１３０による処理の結果等の情報を外部記憶コンポーネント１５０に書き込むことができる。
【００１０】
図示した実施形態では、ＧＰＵ１１５内の計算ユニット１２１～１２３は、本明細書では集合的に「キャッシュ１５１～１５２」と呼ばれる１つ以上のキャッシュ１５１、１５３、１５３を含む（又はそれらに関連付けられる）。キャッシュ１５１～１５３は、Ｌ１キャッシュ、Ｌ２キャッシュ、Ｌ３キャッシュ、又は、キャッシュ階層内の他のキャッシュを含むことができる。キャッシュ１５１～１５３の部分は、計算ユニット１２１～１２３上で実行されるグラフィックスパイプラインのためのテクスチャキャッシュを実装するために使用される。キャッシュ１５１～１５３内のキャッシュラインは、キャッシュラインの１つ以上のセクタ等のサブセットに分割される。グラフィックスパイプラインは、キャッシュラインのサブセットに関連付けられたアドレスへのメモリアクセス要求に対するキャッシュミスに応じて、テクスチャキャッシュ内のキャッシュラインのサブセットに対するミス要求を選択的に生成する。いくつかの実施形態では、キャッシュラインは、第１のセクタ及び第２のセクタに分割される。本明細書で説明するように、要求サイクル中に受信されたメモリアクセス要求に対するキャッシュミスが第１のセクタ内に（排他的に又は主に）あることに応じて、ミス要求が第１のセクタに対して生成され、第２のセクタに対するミス要求の生成がバイパスされる。
【００１１】
図２は、いくつかの実施形態による、高次ジオメトリプリミティブを処理して、所定の解像度で三次元（３Ｄ）シーンのラスタ化された画像を生成するように構成されたグラフィックスパイプライン２００を示す。グラフィックスパイプライン２００は、図１に示される処理システム１００のいくつかの実施形態で実施される。グラフィックスパイプライン２００の図示された実施形態は、ＤＸ１１仕様に従って実装される。グラフィックスパイプライン２００の他の実施形態は、Ｖｕｌｋａｎ、Ｍｅｔａｌ、ＤＸ１２等の他のアプリケーションプログラミングインターフェース（Application Programming Interfaces、ＡＰＩ）に従って実装される。グラフィックスパイプライン２００は、ラスタ化前のグラフィックスパイプライン２００の部分を含むジオメトリ部２０１と、ラスタ化後のグラフィックスパイプライン２００の部分を含むピクセル処理部２０２と、に細分される。
【００１２】
グラフィックスパイプライン２００は、バッファを実装し、頂点データ、テクスチャデータ等を記憶するために使用される１つ以上のメモリ又はキャッシュの階層等のストレージリソース２０５へのアクセスを有する。図示される実施形態では、ストレージリソース２０５は、データを記憶するために使用されるローカルデータストア（ＬＤＳ）２０６回路と、グラフィックスパイプライン２００によるレンダリング中に頻繁に使用されるデータをキャッシュするために使用されるキャッシュ２０７と、を含む。ストレージリソース２０５は、図１に示されるメモリ１０５のいくつかの実施形態を使用して実装され得る。
【００１３】
入力アセンブラ２１０は、シーンのモデルの部分を表すオブジェクトを定義するために使用される、ストレージリソース２０５から情報にアクセスする。プリミティブの一例が

三角形２１１として図２に示されているが、グラフィックスパイプライン２００のいくつかの実施形態では、他のタイプのプリミティブが処理される。三角形２１１は、１つ以上の辺２１４によって接続された１つ以上の頂点２１２を含む（明確にするために、図２には各々の１つのみが示されている）。頂点２１２は、グラフィックスパイプライン２００のジオメトリ処理部２０１中にシェーディングされる。

【００１４】

頂点シェーダ２１５は、図示される実施形態ではソフトウェアで実装されており、プリミティブの単一の頂点２１２を入力として論理的に受信し、単一の頂点を出力する。頂点シェーダ２１５等のシェーダのいくつかの実施形態は、複数の頂点が同時に処理されるように、単一命令－複数データ（ＳＩＭＤ）処理を実装する。グラフィックスパイプライン２００は、グラフィックスパイプライン２００に含まれる全てのシェーダが、共有大規模ＳＩＭＤ計算ユニット上に同じ実行プラットフォームを有するように、統一されたシェーダモデルを実装する。したがって、頂点シェーダ２１５を含むシェーダは、本明細書では統一されたシェーダプール２１６と呼ばれるリソースの共通セットを使用して実装される。

【００１５】

ハルシェーダ２１８は、入力パッチを定義するために使用される入力高次パッチ又は制御ポイント上で動作する。ハルシェーダ２１８は、テッセレーション係数及び他のパッチデータを出力する。いくつかの実施形態では、ハルシェーダ２１８によって生成されたプリミティブは、テッセレータ２２０に提供される。テッセレータ２２０は、ハルシェーダ２１８からオブジェクト（パッチ等）を受信し、例えば、ハルシェーダ２１８によってテッセレータ２２０に提供されたテッセレーション係数に基づいて、入力オブジェクトをテッセレーションすることにより、入力オブジェクトに対応するプリミティブを識別する情報を生成する。テッセレーションは、例えば、テッセレーションプロセスによって生成されたプリミティブの粒度を指定するテッセレーション係数によって示されるように、パッチ等の入力高次プリミティブを、より細かいレベルの詳細を表す低次出力プリミティブのセットに細分する。したがって、シーンのモデルは、（メモリ又は帯域幅を節約するため）より少数の高次プリミティブによって表され、追加の詳細は、高次プリミティブをテッセレーションすることによって追加される。

【００１６】

ドメインシェーダ２２４は、ドメインの場所及び（任意選択的に）他のパッチデータを入力する。ドメインシェーダ２２４は、提供された情報で動作し、入力ドメインの場所及び他の情報に基づいて、出力のための単一の頂点を生成する。図示した実施形態では、ドメインシェーダ２２４は、三角形２１１及びテッセレーション係数に基づいてプリミティブ２２２を生成する。ジオメトリシェーダ２２６は、入力プリミティブを受信し、入力プリミティブに基づいてジオメトリシェーダ２２６によって生成される最大４つのプリミティブを出力する。図示した実施形態では、ジオメトリシェーダ２２６は、テッセレートされたプリミティブ２２２に基づいて出力プリミティブ２２８を生成する。

【００１７】

プリミティブの１つのストリームが１つ以上のスキャンコンバータ２３０に提供され、いくつかの実施形態では、プリミティブの最大４つのストリームは、ストレージリソース２０５内のバッファに連結される。スキャンコンバータ２３０は、シェーディング動作、クリッピング、透視分割、切断及びビューポート選択等の他の動作を実行する。スキャンコンバータ２３０は、グラフィックスパイプライン２００のピクセル処理部２０２において後で処理されるピクセルのセット２３２を生成する。

【００１８】

図示された実施形態では、ピクセルシェーダ２３４は、ピクセルフロー（例えば、ピクセルのセット２３２を含む）を入力し、入力ピクセルフローに応じて０又は別のピクセルフローを出力する。出力マージャブロック２３６は、ピクセルシェーダ２３４から受信したピクセルに対してブレンド、深度、ステンシル又は他の動作を実行する。

【００１９】

グラフィックスパイプライン２００内のシェーダの一部又は全部は、ストレージリソース２０５に記憶されたテクスチャデータを使用してテクスチャマッピングを実行する。例えば、ピクセルシェーダ２３４は、ストレージリソース２０５からテクスチャデータを読み取り、テクスチャデータを使用して１つ以上のピクセルをシェーディングすることができる。次いで、シェーディングされたピクセルは、ユーザに提示するためにディスプレイに提供される。本明細書で説明するように、グラフィックスパイプライン２００内のシェーダによって使用されるテクスチャデータは、キャッシュ２０７を使用してキャッシュされる。ミス要求は、キャッシュ２０７におけるキャッシュミスに応じて、例えばキャッシュ２０７のキャッシュラインの部分におけるアドレスの位置、要求又はデータのヒューリスティック又は特性、時間的局所性、空間的局所性等に基づいて、選択的に生成される。

【００２０】

図３は、いくつかの実施形態による、第１の読み取りサイクル３０１において複数のセクタにわたって分散された要求と、第２の読み取りサイクル３０２において単一のセクタに制約された要求と、を有するキャッシュラインのブロック図である。キャッシュライン３００、３０５は、図１に示されるキャッシュ１５１～１５３のいくつかの実施形態及び図２に示されるキャッシュ２０７のいくつかの実施形態におけるキャッシュラインを表す。キャッシュライン３００、３０５は、グラフィックス処理のためのテクスチャを記憶するために使用され、したがって、キャッシュラインは比較的大きい。例えば、キャッシュライン３００、３０５の各々は、対応する計算ユニット又は他のプロセッサ、プロセッサコア、処理要素等によるアクセスのために１２８バイト、すなわち、１Ｋビットのデータを記憶することができる。図示された実施形態では、キャッシュライン３００、３０５は２つのセクタに分割される。しかしながら、いくつかの実施形態では、キャッシュライン３００、３０５は、３つ以上のセクタに分割される。

【００２１】

第１の読み取りサイクル３０１中に、キャッシュライン３００は、キャッシュライン３００によって記憶されたバイトのサブセットを保持する位置３１０を示すアドレスへの読み取り要求を受信する。位置３１０は、キャッシュライン３００の第１のセクタ３１１内にある。また、キャッシュライン３００は、キャッシュライン３００によって記憶されたバイトの別のサブセットを保持する位置３１５を示すアドレスへの読み取り要求を受信する。位置３１５は、キャッシュライン３００の第２のセクタ３１２内にある。図示した実施形態では、位置３１５及び位置３１０への読み取り要求は、キャッシュライン３００においてミスする。

【００２２】

第２の読み取りサイクル３０２中に、キャッシュライン３０５は、キャッシュライン３０５によって記憶されたバイトのサブセットを保持する位置３２０を示すアドレスへの読み取り要求を受信する。位置３２０は、キャッシュライン３０５の第１のセクタ３２１内にあり、位置３２０の何れもキャッシュライン３０５の第２のセクタ３２２内にない。図示した実施形態では、位置３２０への読み取り要求は、キャッシュライン３０５においてミスする。

【００２３】

ミス要求は、キャッシュライン３００、３０５におけるキャッシュミスの位置に基づいて、第１のセクタ３１１、３２１、第２のセクタ３１２、３２２、又は、両方のセクタ３１１、３１２、３２１、３２２（例えば、フルキャッシュライン３００及び３０５）に対して選択的に生成される。いくつかの実施形態では、キャッシュミスの他のヒューリスティック又は特性も、本明細書で説明されるように、ミス要求が第１のセクタ３１１、３２１、第２のセクタ３１２、３２２、又は、両方のセクタ３１１，３１２、３２１、３２２に対して生成されるか否かを判定するために使用される。例えば、キャッシュミスの測定、予想又は予測された空間的局所性又は時間的局所性を使用して、ミス要求がどのように生成されるかを判定することができる。図示した実施形態では、第１のセクタ３１１内の位置３１０及び第２のセクタ３１２内の位置３１５を含む第１の読み取りサイクル３０１

内のキャッシュミスに応じて、フルキャッシュライン３００（例えば、セクタ３１１及び３１２）に対するミス要求が生成される。第２の読み取りサイクル３０２中に、第２の読み取りサイクル３０２内のキャッシュミスが第１のセクタ３２１内のみにある位置３２０に対するものであることに応じて、ミス要求が第１のセクタ３２１に対してのみ生成される（第２のセクタ３２２に対するミス要求の生成はバイパスされる）。

【００２４】

図４は、いくつかの実施形態による、第１の読み取りサイクル４０１及び第２の読み取りサイクル４０２に高度の時間的局所性を示さない要求を有するキャッシュライン４００のブロック図である。キャッシュライン４００は、図１に示されるキャッシュ１５１～１５３のいくつかの実施形態及び図２に示されるキャッシュ２０７のいくつかの実施形態におけるキャッシュラインを表す。キャッシュライン４００は、グラフィックス処理のためのテクスチャを記憶するために使用され、したがって、キャッシュラインは、比較的大きく、例えば、１２８バイトのデータである。図示された実施形態では、キャッシュライン４００はセクタ４１１、４１２に分割される。しかしながら、いくつかの実施形態では、キャッシュライン４００は、３つ以上のセクタに分割される。

【００２５】

第１の読み取りサイクル４０１中に、キャッシュライン４００は、キャッシュライン４００によって記憶されたバイトのサブセットを保持する位置４１５を示すアドレスへの読み取り要求を受信する。位置４１５は、全て、第１のセクタ４１１内に見出される。第２の読み取りサイクル４０２中に、キャッシュライン４００は、キャッシュライン４００によって記憶されたバイトのサブセットを保持する位置４２０を示すアドレスへの読み取り要求を受信する。位置４２０は、全て、キャッシュライン４００の第２のセクタ４１２内にある。図示した実施形態では、ミス要求は、キャッシュミスの実際の又は予測された時間的局所性に少なくとも部分的に基づいて、第１のセクタ４１１又は第２のセクタ４１２に対して選択的に生成される。したがって、連続する読み取りサイクル（例えば、第１の読み取りサイクル４０１及び第２の読み取りサイクル４０２）におけるキャッシュミスが第１のセクタ４１１及び第２のセクタ４１２にわたって分散されるので、ミス要求がフルキャッシュライン（例えば、第１のセクタ４１１及び第２のセクタ４１２）に対して生成される。対照的に、キャッシュライン４００への読み取り要求が高度の時間的局所性を示す場合、例えば、複数のサイクル中の読み取り要求が、セクタ４１１、４１２のうち何れかのみに位置するアドレスに対するものであると予想又は予測される場合、ミス要求は、対応するセクタに対してのみ生成される。

【００２６】

また、ミス要求は、読み取り要求の予測された空間的局所性に基づいて、キャッシュライン４００の異なる部分に対して生成される。例えば、ピクセルシェーダがスクリーンにわたってスキャンしている場合、読み取り要求シーケンスは、隣接する又は近接するピクセル位置に関連付けられたローカルアドレスに対するものである可能性が高い。したがって、メモリアクセスシステムの効率及び性能は、後続の読み取り要求がキャッシュラインの他のセクタ内にあり得る近くのアドレスに対するものである可能性が高いので、現在の読み取りサイクル中のキャッシュミスが単一のセクタ内のみ（又は主に）にある場合でも、予測された空間的局所性が高い（例えば、閾値を上回る）場合、フルキャッシュラインをフェッチすることによって改善される可能性が高い。対照的に、予測された空間的局所性が低い（例えば、閾値を下回る）場合、現在の読み取りサイクル中のキャッシュミスが単一のセクタ内の位置に対するものである場合、ミス要求は単一のセクタに対してのみ生成され得る。いくつかの実施形態では、読み取り要求又はミス要求に関連付けられた情報は、ヒステリシスウィンドウのために保持され、ヒステリシスウィンドウ内の情報は、読み取り要求又はミス要求の時間的局所性又は空間的局所性を判定又は予測するために使用される。

【００２７】

図５は、いくつかの実施形態による、キャッシュラインの部分に対するミス要求を選択

的に生成する方法５００のフロー図である。方法５００は、図１に示される処理システム１００及び図２に示されるグラフィックスパイプライン２００のいくつかの実施形態で実施される。キャッシュラインは、２つのセクタ（又は部分若しくは半分）を含むが、いくつかの実施形態では、キャッシュラインは、より多くのセクタを含む。

【００２８】

ブロック５０５において、キャッシュは、要求サイクル中にスレッドからキャッシュラインへの読み取り要求を受信する。読み取り要求は、キャッシュライン及び対応するメモリ内の位置を示すアドレスを含む。図示した実施形態では、読み取り要求はキャッシュライン内でミスし、これが、要求されたデータをバッキングメモリ又はより高レベルのキャッシュからフェッチするためのミス要求の選択的生成をトリガする。

【００２９】

判定ブロック５１０において、キャッシュは、キャッシュミスを生成したスレッドが現在の要求サイクル中にキャッシュラインの両方のセクタ内の位置にマッピングする否かを判定する。マッピングする場合、方法５００はブロック５２０に進む。マッピングしない場合、方法５００は判定ブロック５１５に進む。

【００３０】

判定ブロック５１５において、キャッシュは、特定のヒューリスティックに基づいて空間的局所性又は時間的局所性の尤度を判定する。いくつかの実施形態では、空間的局所性又は時間的局所性の尤度は、固定ヒューリスティック又はプログラマブルヒューリスティックとの一致があるか否かに基づいて判定される。例えば、キャッシュライン内の情報が色圧縮又は深度圧縮を使用して生成されるか否かにより、関連するデータが高度の空間的局所性を有し、後続の読み取り要求がキャッシュラインの両方のセクタ内のアドレスを含む可能性が高いことを示す。したがって、情報が高度の局所性を有すると予想される場合、方法５００はブロック５２０に進む。そうでない場合、方法５００はブロック５２５に進む。

【００３１】

ブロック５２０において、キャッシュは、フルキャッシュラインに対するミス要求を生成する、すなわち、キャッシュは、キャッシュラインの全てのセクタを含むミス要求を生成する。ブロック５２５において、キャッシュは、要求サイクル中のスレッドに対するキャッシュミスにおけるアドレスに対応する位置を含むキャッシュラインのセクタ（又は部分若しくは半分）に対するミス要求を生成する。

【００３２】

本明細書で開示するように、いくつかの実施形態では、装置は、複数のサブセットに分割されるキャッシュラインを含むテクスチャキャッシュと、グラフィックスパイプライン内の少なくとも１つの計算ユニットと、を含み、プロセッサは、キャッシュラインの第１のサブセットに関連付けられたアドレスへのメモリアクセス要求に対するキャッシュミスに応じて、テクスチャキャッシュ内のキャッシュラインの複数のサブセットのうち第１のサブセットに対するミス要求を選択的に生成するように構成されている。一態様では、少なくとも１つの計算ユニットは、メモリアクセス要求に関連付けられたキャッシュミスが、複数のサブセットのうち第１のサブセットのみにマッピングするのか、又は、複数のサブセットのうち第１のサブセットに追加の又は第１のサブセット以外の１つ以上のサブセットにマッピングするかを判定するように構成されている。別の態様では、少なくとも１つの計算ユニットは、複数のサブセットのうち第１のサブセットに追加の又は第１のサブセット以外のサブセットへのキャッシュミスのマッピングに応じて、フルキャッシュラインに対するミス要求を生成するように構成されている。更に別の態様では、少なくとも１つの計算ユニットは、メモリアクセス要求が第１のサブセットのみにマッピングすることに応じて、色圧縮及び深度圧縮のうち少なくとも１つがテクスチャデータに対して有効にされているか否かを判定するように構成されている。

【００３３】

一態様では、少なくとも１つの計算ユニットは、色圧縮及び深度圧縮のうち少なくとも

１つがテクスチャデータに対して有効にされていることに応じて、フルキャッシュラインに対するミス要求を生成するように構成されている。別の態様では、少なくとも１つの計算ユニットは、色圧縮及び深度圧縮のうち少なくとも１つがテクスチャデータに対して有効にされていないことに応じて、キャッシュラインの第１のサブセットに対するミス要求を生成するように構成されている。更に別の態様では、少なくとも１つの計算ユニットは、メモリアクセス要求の時間的局所性及び空間的局所性のうち少なくとも１つに基づいて、第１のサブセット又は複数のサブセットへのミス要求を選択的に生成するように構成されている。

【００３４】

一態様では、少なくとも１つの計算ユニットは、メモリアクセス要求シーケンスが複数のサブセットにアクセスすることが予想されることに応じて、第１のサブセットにおけるキャッシュミスに応じて、複数のサブセットに対するミス要求を生成するように構成されている。別の態様では、少なくとも１つの計算ユニットは、第１のサブセット内のキャッシュミスに応じて、且つ、閾値を上回る空間的局所性を有するメモリアクセス要求シーケンスに応じて、複数のサブセットに対するミス要求を生成するように構成されている。更に別の態様では、少なくとも１つの計算ユニットは、メモリアクセス要求が閾値を下回る空間的局所性を有することに応じて、第１のサブセットに対するミス要求を生成するように構成されている。

【００３５】

いくつかの実施形態では、方法は、複数のサブセットに分割されるキャッシュラインを含むテクスチャキャッシュ内のキャッシュラインに対するミス要求を検出することと、キャッシュミスがキャッシュラインの第１のサブセットに関連付けられたアドレスに対するものであることに応じて、テクスチャキャッシュ内のキャッシュラインの複数のサブセットのうち第１のサブセットに対するミス要求を選択的に生成することと、を含む。一態様では、本方法は、メモリアクセス要求に関連付けられたキャッシュミスが、複数のサブセットのうち第１のサブセットのみにマッピングするのか、又は、複数のサブセットのうち第１のサブセットに追加の又は第１のサブセット以外の１つ以上のサブセットにマッピングするかを判定することを含む。別の態様では、本方法は、複数のサブセットへのキャッシュミスのマッピングに応じて、フルキャッシュラインに対するミス要求を生成することを含む。更に別の態様では、本方法は、色圧縮及び深度圧縮のうち少なくとも１つがテクスチャデータに対して有効にされているか否かを判定することを含む。

【００３６】

一態様では、ミス要求を選択的に生成することは、色圧縮及び深度圧縮のうち少なくとも１つがテクスチャデータに対して有効にされていることに応じて、複数のサブセットに対するミス要求を生成することを含む。別の態様では、ミス要求を選択的に生成することは、色圧縮及び深度圧縮のうち少なくとも１つがテクスチャデータに対して有効にされていないことに応じて、キャッシュラインの第１のサブセットに対するミス要求を生成することを含む。別の態様では、ミス要求を選択的に生成することは、第１のサブセットにおけるキャッシュミスに応じて、且つ、メモリアクセス要求シーケンスがキャッシュラインの異なるセクタにアクセスすることが予想されることに応じて、複数のサブセットに対するミス要求を生成することを含む。

【００３７】

一態様では、ミス要求を選択的に生成することは、第１のサブセットにおけるキャッシュミスに応じて、且つ、閾値を上回る空間的局所性を有するメモリアクセス要求シーケンスに応じて、複数のサブセットに対するミス要求を生成することを含む。別の態様では、ミス要求を選択的に生成することは、メモリアクセス要求が閾値を下回る空間的局所性を有することに応じて、第１のサブセットに対するミス要求を生成することを含む。

【００３８】

いくつかの実施形態では、装置は、第１のセクタ及び第２のセクタに分割されるキャッシュラインを含むテクスチャキャッシュと、グラフィックスパイプライン内の少なくとも

１つの計算ユニットと、を含み、少なくとも１つの計算ユニットは、要求サイクル中に受信されたメモリアクセス要求に対するキャッシュミスが第１のセクタ中にあることに応じて、第１のセクタに対するミス要求を生成し、第２のセクタに対するミス要求の生成をバイパスするように構成されている。一態様では、少なくとも１つの計算ユニットは、要求サイクル中に受信されたメモリアクセス要求に対するキャッシュミスが第１のセクタ及び第２のセクタ内にあることに応じて、第１のセクタ及び第２のセクタに対するミス要求を生成するように構成されている。

【００３９】

コンピュータ可読記憶媒体は、命令及び／又はデータをコンピュータシステムに提供するために、使用中にコンピュータシステムによってアクセス可能な任意の非一時的な記憶媒体又は非一時的な記憶媒体の組み合わせを含む。このような記憶媒体には、限定されないが、光学媒体（例えば、コンパクトディスク（ＣＤ）、デジタル多用途ディスク（ＤＶＤ）、ブルーレイ（登録商標）ディスク）、磁気媒体（例えば、フロッピー（登録商標）ディスク、磁気テープ、磁気ハードドライブ）、揮発性メモリ（例えば、ランダムアクセスメモリ（ＲＡＭ）若しくはキャッシュ）、不揮発性メモリ（例えば、読取専用メモリ（ＲＯＭ）若しくはフラッシュメモリ）、又は、微小電気機械システム（ＭＥＭＳ）ベースの記憶媒体が含まれ得る。コンピュータ可読記憶媒体（例えば、システムＲＡＭ又はＲＯＭ）はコンピューティングシステムに内蔵されてもよいし、コンピュータ可読記憶媒体（例えば、磁気ハードドライブ）はコンピューティングシステムに固定的に取り付けられてもよいし、コンピュータ可読記憶媒体（例えば、光学ディスク又はユニバーサルシリアルバス（ＵＳＢ）ベースのフラッシュメモリ）はコンピューティングシステムに着脱可能に取り付けられてもよいし、コンピュータ可読記憶媒体（例えば、ネットワークアクセス可能ストレージ（ＮＡＳ））は有線又は無線ネットワークを介してコンピュータシステムに結合されてもよい。

【００４０】

いくつかの実施形態では、上述した技術の特定の態様は、ソフトウェアを実行する処理システムの１つ以上のプロセッサによって実装される。ソフトウェアは、非一時的なコンピュータ可読記憶媒体に記憶されるか、別の方法で明確に具体化された実行可能命令の１つ以上のセットを含む。ソフトウェアは、命令及び特定のデータを含んでもよく、当該命令及び特定のデータは、１つ以上のプロセッサによって実行されると、上述した技術の１つ以上の態様を実行するように１つ以上のプロセッサを操作する。非一時的なコンピュータ可読記憶媒体は、例えば、磁気又は光ディスク記憶デバイス、フラッシュメモリ等のソリッドステート記憶デバイス、キャッシュ、ランダムアクセスメモリ（ＲＡＭ）、又は、他の不揮発性メモリデバイス（単数又は複数）等を含み得る。非一時的なコンピュータ可読記憶媒体に記憶された実行可能命令は、ソースコード、アセンブリ言語コード、オブジェクトコード、又は、１つ以上のプロセッサによって解釈され若しくは別の方法で実行可能な他の命令形式で実装可能である。

【００４１】

上述したものに加えて、概要説明において説明した全てのアクティビティ又は要素が必要とされているわけではなく、特定のアクティビティ又はデバイスの一部が必要とされない場合があり、１つ以上のさらなるアクティビティが実行される場合があり、１つ以上のさらなる要素が含まれる場合があることに留意されたい。さらに、アクティビティが列挙された順序は、必ずしもそれらが実行される順序ではない。また、概念は、特定の実施形態を参照して説明された。しかしながら、当業者であれば、特許請求の範囲に記載されているような本発明の範囲から逸脱することなく、様々な変更及び変形を行うことができるのを理解するであろう。したがって、明細書及び図面は、限定的な意味ではなく例示的な意味で考慮されるべきであり、これらの変更形態の全ては、本発明の範囲内に含まれることが意図される。

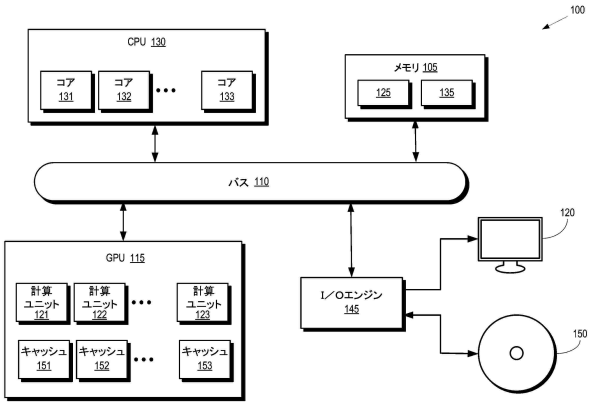
【００４２】

利益、他の利点及び問題に対する解決手段を、特定の実施形態に関して上述した。しか

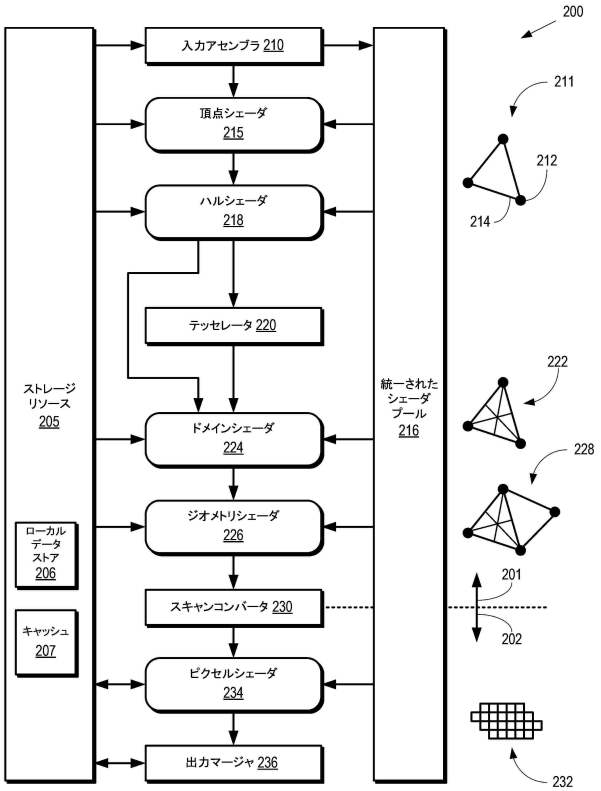
し、利益、利点、問題に対する解決手段、及び、何かしらの利益、利点若しくは解決手段が発生又は顕在化する可能性のある特徴は、何れか若しくは全ての請求項に重要な、必須の、又は、不可欠な特徴と解釈されない。さらに、開示された発明は、本明細書の教示の利益を有する当業者には明らかな方法であって、異なっているが同様の方法で修正され実施され得ることから、上述した特定の実施形態は例示にすぎない。添付の特許請求の範囲に記載されている以外に本明細書に示されている構成又は設計の詳細については限定がない。したがって、上述した特定の実施形態は、変更又は修正されてもよく、かかる変更形態の全ては、開示された発明の範囲内にあると考えられることが明らかである。したがって、ここで要求される保護は、添付の特許請求の範囲に記載されている。

【図面】

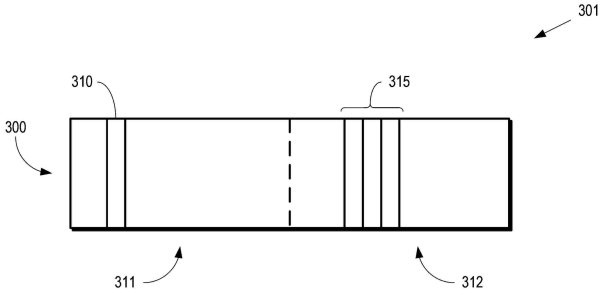
【図１】



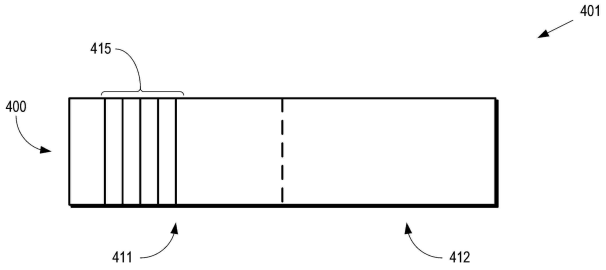
【図２】

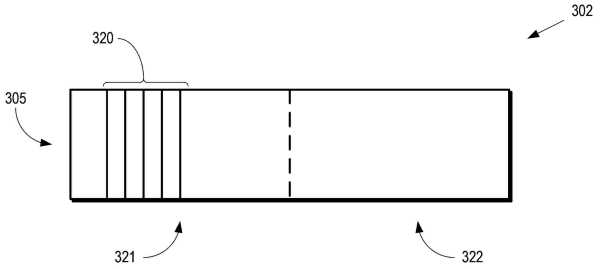


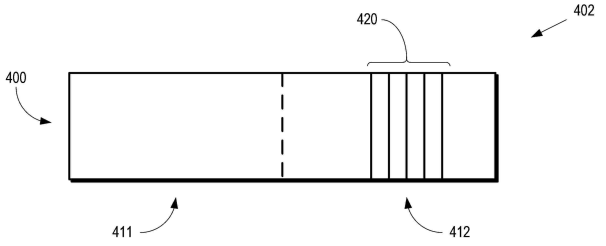
【図３】


301
310
315
300
311
312

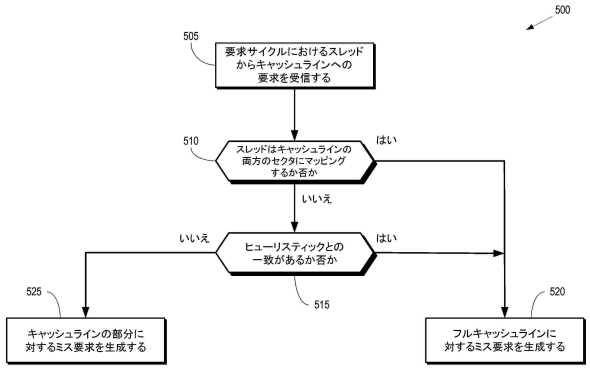
【図４】


401
415
400
411
412


302
320
305
321
322


402
420
400
411
412

【図５】


500
505
要求サイクルにおけるスレッドからキャッシュラインへの要求を受信する
510
スレッドはキャッシュラインの両方のセクタにマッピングするか否か
はい
いいえ
いいえ
ヒューリスティックとの一致があるか否か
はい
515
525
キャッシュラインの部分に対するミス要求を生成する
520
フルキャッシュラインに対するミス要求を生成する

フロントページの続き

弁理士　早川 裕司
(74)代理人　100111615
弁理士　佐野 良太
(74)代理人　100162156
弁理士　村雨 圭介
(72)発明者　ファタネー エフ． ゴッドラット
アメリカ合衆国 ９５０５４ カリフォルニア州、サンタ クララ、オーガスティン ドライブ ２４８５
(72)発明者　スティーブン ダブリュ． ソモギ
アメリカ合衆国 ９５０５４ カリフォルニア州、サンタ クララ、オーガスティン ドライブ ２４８５
(72)発明者　チェンホン リウ
大韓民国 １６６７７ ギョンギ－ド、スウォン－シ、ヨントン－ク、サムスン－ロ、１２９

審査官　渡部 幸和
(56)参考文献　米国特許出願公開第２００６／０２５０４０８（ＵＳ，Ａ１）
米国特許第０６７２４３９１（ＵＳ，Ｂ１）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　１５／００
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　９／３８