

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7282675号

(P7282675)

(45)発行日 令和5年5月29日(2023.5.29)

(24)登録日 令和5年5月19日(2023.5.19)

(51)国際特許分類

F I

G 0 6 F 12/0844(2016.01)

G 0 6 F 12/0844

G 0 6 F 9/34 (2018.01)

G 0 6 F 9/34 3 5 0 A

請求項の数 20 (全21頁)

(21)出願番号	特願2019-531774(P2019-531774)	(73)特許権者	591016172
(86)(22)出願日	平成29年11月21日(2017.11.21)		アドバンスト・マイクロ・デバイス
(65)公表番号	特表2020-513631(P2020-513631 A)		・インコーポレイテッド
(43)公表日	令和2年5月14日(2020.5.14)		ADVANCED MICRO DEVI
(86)国際出願番号	PCT/US2017/062889		CES INCORPORATED
(87)国際公開番号	WO2018/111515		アメリカ合衆国 9 5 0 5 4 カリフォル
(87)国際公開日	平成30年6月21日(2018.6.21)		ニア州、 サンタ クララ、 オーガスティ
審査請求日	令和2年11月10日(2020.11.10)		ン ドライブ 2 4 8 5
(31)優先権主張番号	15/377,998	(74)代理人	100108833
(32)優先日	平成28年12月13日(2016.12.13)		弁理士 早川 裕司
(33)優先権主張国・地域又は機関	米国(US)	(74)代理人	100111615
前置審査			弁理士 佐野 良太
		(74)代理人	100162156
			弁理士 村雨 圭介
		(72)発明者	ダニエル シュナイダー

最終頁に続く

(54)【発明の名称】 アウトオブオーダーキャッシュリターン

(57)【特許請求の範囲】

【請求項 1】

アウトオブオーダーキャッシュリターンを実行する方法であって、

キャッシュシステムが、前記キャッシュシステム内に存在する複数のリターンオーダーリングキューのうち第1リターンオーダーリングキューの先頭にある第1エントリが、単一命令複数データユニットで実行されるウェブフロントへのリターンに利用可能であると判別することであって、前記第1エントリは、第1キャッシュアクセス要求に対応し、前記第1リターンオーダーリングキューは、テクスチャサンブラ操作キャッシュアクセスタイプを含む第2キャッシュアクセスタイプではなく、前記テクスチャサンブラ操作キャッシュアクセスタイプを含まない第1キャッシュアクセスタイプのキャッシュアクセス要求のエントリを記憶し、前記複数のリターンオーダーリングキューのうち第2リターンオーダーリングキューは、第1キャッシュアクセスタイプではなく、第2キャッシュアクセスタイプのキャッシュアクセス要求のエントリを記憶する、ことと、

前記キャッシュシステムが、前記判別したことに応じて、前記第1キャッシュアクセス要求よりも古い前記第2リターンオーダーリングキューのエントリに対応するキャッシュアクセス要求が前記ウェブフロントへのリターンに利用可能になるのを待つことなく、前記第1エントリに対応するキャッシュリターンを前記ウェブフロントに送ることと、を含む、

方法。

【請求項 2】

前記第 1 キャッシュアクセスタイプは、読み出しタイプ及び書き込みタイプのうち何れかを含む、

請求項 1 の方法。

【請求項 3】

前記読み出しタイプは、メモリシステムに対してデータを要求し、応答としてデータを受信するアクセスタイプを含み、

前記書き込みタイプは、データを前記メモリシステムに書き込み、応答として肯定応答信号を受信するアクセスタイプを含み、

前記テクスチャサンプリング操作キャッシュアクセスタイプは、テクスチャ座標を介してテクスチャデータを要求し、応答として前記テクスチャデータを受信するアクセスタイプを含む、

10

請求項 2 の方法。

【請求項 4】

前記テクスチャサンプリング操作キャッシュアクセスタイプは、前記テクスチャ座標を 1 つ以上のメモリアドレスに変換することと、前記 1 つ以上のメモリアドレスからデータをフェッチすることと、前記フェッチされたデータを解凍することと、前記フェッチされたデータにフィルタリングを適用することと、のうち 1 つ以上を要求するアクセスタイプを含む、

請求項 3 の方法。

【請求項 5】

20

前記複数のリターンオーダーリングキューのモードを選択することであって、前記モードは、前記複数のリターンオーダーリングキュー内のリターンオーダーリングキューの数と、前記複数のリターンオーダーリングキューの各々に記憶され、オーダーリングされる 1 つ以上のキャッシュアクセスタイプと、を定義する、ことを含む、

請求項 1 の方法。

【請求項 6】

前記複数のリターンオーダーリングキューは、モノリシックメモリ内に記憶された仮想キューを含み、前記仮想キューは、前記選択されたモードに適應するようにサイズ変更可能である、

請求項 5 の方法。

30

【請求項 7】

各仮想キューの先頭から対応する物理キューの先頭にエントリをコピーすることを含み、前記第 1 エントリに対応するキャッシュリターンを前記ウェーブフロントに送ることは、前記第 1 リターンオーダーリングキューに対応する物理キューの先頭からエントリを除去することと、

前記物理キューの次に古いエントリを、前記物理キューの先頭になるように変更することと、

エントリを、前記第 1 リターンオーダーリングキューから、前記第 1 リターンオーダーリングキューに対応する前記物理キューにコピーすることと、を含む、

請求項 6 の方法。

40

【請求項 8】

前記ウェーブフロントにおいてキャッシュアクセスタイプに基づくバリア命令を実行することを含む、

請求項 1 の方法。

【請求項 9】

前記キャッシュアクセスタイプに基づくバリア命令を実行することは、

特定のキャッシュアクセスタイプの未処理のキャッシュアクセスが完了するまで前記ウェーブフロントをストールさせることを含む、

請求項 8 の方法。

【請求項 10】

50

アウトオブオーダーキャッシュリターンを実行するための計算ユニットであって、
ウェーブフロントを実行するように構成された単一命令複数データユニットと、
キャッシュシステムと、を備え、
前記キャッシュシステムは、

第1リターンオーダーリングキューと、第2リターンオーダーリングキューと、を含む複数の
リターンオーダーリングキューを記憶することであって、前記第1リターンオーダーリング
キューは、テクスチャサンブラ操作キャッシュアクセスタイプを含む第2キャッシュア
クセスタイプではなく、前記テクスチャサンブラ操作キャッシュアクセスタイプを含まない
第1キャッシュアクセスタイプのキャッシュアクセス要求のエントリを記憶し、前記第2
リターンオーダーリングキューは、第1キャッシュアクセスタイプではなく、第2キャッシ
ュアクセスタイプのキャッシュアクセス要求のエントリを記憶する、ことと、

10

前記第1リターンオーダーリングキューの先頭にある第1エントリが、前記ウェーブフ
ロントへのリターンに利用可能であると判別することであって、前記第1エントリは、第1
キャッシュアクセス要求に対応する、ことと、

前記判別したことに応じて、前記第1キャッシュアクセス要求よりも古い前記第2リタ
ーンオーダーリングキューのエントリに対応するキャッシュアクセス要求が前記ウェーブ
フロントへのリターンに利用可能になるのを待つことなく、前記第1エントリに対応する
キャッシュリターンを前記ウェーブフロントに送ることと、を行うように構成されている、
計算ユニット。

【請求項11】

20

前記第1キャッシュアクセスタイプは、読み出しタイプ及び書き込みタイプのうち何れ
かを含む、

請求項10の計算ユニット。

【請求項12】

前記読み出しタイプは、メモリシステムに対してデータを要求し、応答としてデータを
受信するアクセスタイプを含み、

前記書き込みタイプは、データを前記メモリシステムに書き込み、応答として肯定応答
信号を受信するアクセスタイプを含み、

前記テクスチャサンブラ操作キャッシュアクセスタイプは、テクスチャ座標を介してテ
クスチャデータを要求し、応答として前記テクスチャデータを受信するアクセスタイプを
含む、

30

請求項11の計算ユニット。

【請求項13】

前記テクスチャサンブラ操作キャッシュアクセスタイプは、前記テクスチャ座標を1つ
以上のメモリアドレスに変換することと、前記1つ以上のメモリアドレスからデータをフ
ェッチすることと、前記フェッチされたデータを解凍することと、前記フェッチされたデ
ータにフィルタリングを適用することと、のうち1つ以上を要求するアクセスタイプを含
む、

請求項12の計算ユニット。

【請求項14】

40

前記キャッシュシステムは、

前記複数のリターンオーダーリングキューのモードを選択することであって、前記モード
は、前記複数のリターンオーダーリングキュー内のリターンオーダーリングキューの数と、前
記複数のリターンオーダーリングキューの各々に記憶され、オーダーリングされる1つ以上の
キャッシュアクセスタイプと、を定義する、ことを行うように構成されている、

請求項10の計算ユニット。

【請求項15】

前記複数のリターンオーダーリングキューは、モノリシックメモリ内に記憶された仮想キ
ューを含み、前記仮想キューは、前記選択されたモードに適応するようにサイズ変更可能
である、

50

請求項 14 の計算ユニット。

【請求項 16】

前記キャッシュシステムは、

各仮想キューの先頭から対応する物理キューの先頭にエントリをコピーすることを行うように構成されており、

前記第 1 エントリに対応するキャッシュリターンを前記ウェブフロントに送ることは、

前記第 1 リターンオーダリングキューに対応する物理キューの先頭からエントリを除去することと、

前記物理キューの次に古いエントリを、前記物理キューの先頭になるように変更することと、

エントリを、前記第 1 リターンオーダリングキューから、前記第 1 リターンオーダリングキューに対応する前記物理キューにコピーすることと、を含む、

請求項 15 の計算ユニット。

【請求項 17】

前記ウェブフロントは、

キャッシュアクセスタイプに基づくバリア命令を実行するように構成されている、

請求項 10 の計算ユニット。

【請求項 18】

前記ウェブフロントは、前記キャッシュアクセスタイプに基づくバリア命令を実行することに応じて、特定のキャッシュアクセスタイプの未処理のキャッシュアクセスが完了するまでストールされる、

請求項 17 の計算ユニット。

【請求項 19】

計算ユニットを含むアクセラレーテッド処理デバイスと、

前記計算ユニット内のウェブフロントを前記アクセラレーテッド処理デバイスに実行させるように構成されたプロセッサと、を備え、

前記計算ユニットは、

前記ウェブフロントを実行するように構成された単一命令複数データユニットと、

キャッシュシステムと、を備え、

前記キャッシュシステムは、

第 1 リターンオーダリングキューと、第 2 リターンオーダリングキューと、を含む複数のリターンオーダリングキューを記憶することであって、前記第 1 リターンオーダリングキューは、テクスチャサンブラ操作キャッシュアクセスタイプを含む第 2 キャッシュアクセスタイプではなく、前記テクスチャサンブラ操作キャッシュアクセスタイプを含まない第 1 キャッシュアクセスタイプのキャッシュアクセス要求のエントリを記憶し、前記第 2 リターンオーダリングキューは、第 1 キャッシュアクセスタイプではなく、第 2 キャッシュアクセスタイプのキャッシュアクセス要求のエントリを記憶する、ことと、

前記第 1 リターンオーダリングキューの先頭にある第 1 エントリが、前記ウェブフロントへのリターンに利用可能であると判別することであって、前記第 1 エントリは、第 1 キャッシュアクセス要求に対応する、ことと、

前記判別したことに応じて、前記第 1 キャッシュアクセス要求よりも古い前記第 2 リターンオーダリングキューのエントリに対応するキャッシュアクセス要求が前記ウェブフロントへのリターンに利用可能になるのを待つことなく、前記第 1 エントリに対応するキャッシュリターンを前記ウェブフロントに送ることと、を行うように構成されている、コンピュータシステム。

【請求項 20】

前記第 1 キャッシュアクセスタイプは、読み出しタイプ及び書き込みタイプのうち何れかを含む、

請求項 19 のコンピュータシステム。

【発明の詳細な説明】

10

20

30

40

50

【技術分野】

【0001】

(関連出願の相互参照)

本願は、2016年12月13日に出願された米国特許出願第15/377,998号の利益を主張するものであり、この内容は参照により本明細書に完全に記載されているものとして援用される。

【0002】

開示される実施形態は、概して、グラフィックス処理に関し、特に、アウトオブオーダーキャッシュリターンに関する。

【背景技術】

10

【0003】

3次元グラフィックスをレンダリングするためのハードウェアは高度に並列であり、このハードウェアは、メモリからのデータを要求し、当該データに対して計算を行い、処理されたデータをフレームバッファに提供してスクリーンに出力する多数の個別処理装置を含む。通常、メモリ内のデータにアクセスするには、多くのレイテンシが生じる。このレイテンシを短縮するためにキャッシュシステムが設けられている。しかしながら、レンダリング動作において一般的に処理される大量のデータのために、メモリアクセスレイテンシの更なる改善が望まれる。

【0004】

添付の図面と共に例として与えられる以下の説明から、より詳細な理解を得ることができる。

20

【図面の簡単な説明】

【0005】

【図1】1つ以上の開示された実施形態を実装し得る例示的なデバイスのブロック図である。

【図2】一例による、アクセラレーテッド処理デバイスを示すブロック図である。

【図3】一例による、グラフィックス処理パイプラインを示すブロック図である。

【図4】一例による、キャッシュシステムを示すブロック図である。

【図5】一例による、様々なタイプのキャッシュアクセス要求をオーダリング (ordering) するためのキューを示すブロック図である。

30

【図6】一例による、リターンオーダリングキューを示す図である。

【図7】一例による、アクセスタイプ毎のリターンオーダリングキューにキャッシュアクセス要求の指示を記録する方法のフロー図である。

【図8】一例による、アウトオブオーダーキャッシュリターンをワークグループに提供する方法のフロー図である。

【発明を実施するための形態】

【0006】

本開示は、キャッシュアクセスのリターンをアウトオブオーダーにすることを可能にする技術に関する。より具体的には、複数のキャッシュアクセスタイプ毎にリターンオーダリングキューが存在し、未処理のキャッシュアクセスを、それらのアクセスが行われた順に記憶する。キャッシュアクセスタイプには、読み出しタイプ、書き込みタイプ、テクスチャサンプリングタイプが含まれる。特定のタイプのキャッシュアクセス要求が当該タイプのリターンオーダリングキューの先頭にあり、キャッシュアクセスが当該アクセスを行ったウェーブフロント (wavefront) へのリターンに使用可能である場合、キャッシュシステムは、キャッシュアクセスをウェーブフロントに返す。このリターンは、使用可能なキャッシュアクセスを記憶するリターンオーダリングキューに関連するタイプ以外のタイプのキャッシュアクセスの順序 (オーダ) とは無関係に行われる。したがって、キャッシュアクセスは、異なるタイプのキャッシュアクセスに対してアウトオブオーダーで戻される場合がある。アウトオブオーダーリターンを可能にすると、例えば、比較的長いレイテンシのアクセスタイプ (例えば、テクスチャサンプリング操作) の後に比較的短いレイテンシのアクセス

40

50

タイプ（例えば、読み出し）が発行される状況等において、レイテンシを改善するのに効果的となり得る。

【 0 0 0 7 】

図 1 は、1 つ以上の開示された実施形態を実装し得る例示的なデバイス 1 0 0 のブロック図である。デバイス 1 0 0 は、例えば、コンピュータ、ゲーム機、ハンドヘルドデバイス、セットトップボックス、テレビ、携帯電話又はタブレットコンピュータ等を含む。デバイス 1 0 0 は、プロセッサ 1 0 2 と、メモリ 1 0 4 と、ストレージデバイス 1 0 6 と、1 つ以上の入力デバイス 1 0 8 と、1 つ以上の出力デバイス 1 1 0 と、を含む。また、デバイス 1 0 0 は、入力デバイス 1 0 8 及び出力デバイス 1 1 0 をそれぞれ駆動する入力ドライバ 1 1 2 及び出力ドライバ 1 1 4 も含む。デバイス 1 0 0 は、図 1 に示されていない追加のコンポーネントを含み得ることが理解されるであろう。

10

【 0 0 0 8 】

プロセッサ 1 0 2 は、中央処理装置（C P U）、グラフィックス処理装置（G P U）、同じダイ上に配置された C P U 及び G P U、又は、1 つ以上のプロセッサコアを含み、各プロセッサコアは、C P U 又は G P U であってもよい。メモリ 1 0 4 は、プロセッサ 1 0 2 と同じダイ上に配置されてもよいし、プロセッサ 1 0 2 とは別に配置されてもよい。メモリ 1 0 4 は、揮発性又は不揮発性メモリ（例えば、ランダムアクセスメモリ（R A M）、ダイナミック R A M、キャッシュ等）を含む。

【 0 0 0 9 】

ストレージデバイス 1 0 6 は、固定又は着脱可能なストレージデバイス（例えば、ハードディスクドライブ、ソリッドステートドライブ、光ディスク若しくはフラッシュドライブ等）を含む。入力デバイス 1 0 8 は、キーボード、キーパッド、タッチスクリーン、タッチパッド、検出器、マイクロフォン、加速度計、ジャイロスコープ、バイオメトリクススキャナ又はネットワーク接続（例えば、無線 I E E E 8 0 2 信号の送信及び／若しくは受信の無線ローカルエリアネットワークカード）を含む。出力デバイス 1 1 0 は、ディスプレイデバイス、スピーカ、プリンタ、触覚フィードバックデバイス、1 つ以上のライト、アンテナ又はネットワーク接続（例えば、無線 I E E E 8 0 2 信号の送信及び／若しくは受信の無線ローカルエリアネットワークカード）を含む。

20

【 0 0 1 0 】

入力ドライバ 1 1 2 は、プロセッサ 1 0 2 及び入力デバイス 1 0 8 と通信し、プロセッサ 1 0 2 が入力デバイス 1 0 8 から入力を受信するのを可能にする。出力ドライバ 1 1 4 は、プロセッサ 1 0 2 及び出力デバイス 1 1 0 と通信し、プロセッサ 1 0 2 が出力デバイス 1 1 0 に出力を送信するのを可能にする。出力ドライバ 1 1 4 は、ディスプレイデバイス 1 1 8 に接続されたアクセラレーテッド処理デバイス（A P D）1 1 6 を含む。A P D 1 1 6 は、プロセッサ 1 0 2 から計算コマンド及びグラフィックスレンダリングコマンドを受信し、計算コマンド及びグラフィックスレンダリングコマンドを処理して、表示のためにピクセル出力をディスプレイデバイス 1 1 8 に提供するように構成されている。

30

【 0 0 1 1 】

A P D 1 1 6 は、単一命令複数データ（S I M D）パラダイムに従って計算を実行するように構成された 1 つ以上の並列処理装置を含む。しかしながら、A P D 1 1 6 によって実行されるものとして説明した機能は、S I M D パラダイムに従ってデータを処理しない処理デバイスによって実行されてもよい。

40

【 0 0 1 2 】

図 2 は、一例による、アクセラレーテッド処理デバイス 1 1 6 のブロック図である。プロセッサ 1 0 2 は、プロセッサ 1 0 2 による実行のための 1 つ以上の制御論理モジュールをシステムメモリ 1 0 4 内に保持している。制御論理モジュールは、オペレーティングシステム 1 2 0 と、ドライバ 1 2 2 と、アプリケーション 1 2 6 と、を含む。これらの制御論理モジュールは、プロセッサ 1 0 2 及び A P D 1 1 6 の様々な態様の動作を制御する。例えば、オペレーティングシステム 1 2 0 は、ハードウェアと直接通信し、プロセッサ 1 0 2 で実行中の他のソフトウェア用のハードウェアへのインタフェースを提供する。ドラ

50

イバ１２２は、例えば、ＡＰＤ１１６の様々な機能にアクセスするために、プロセッサ１０２で実行中のソフトウェア（例えば、アプリケーション１２６）へのアプリケーションプログラミングインタフェース（ＡＰＩ）を提供することによって、ＡＰＤ１１６の動作を制御する。また、ドライバ１２２は、ＡＰＤ１１６の処理コンポーネント（例えば、以下でさらに詳細に説明するＳＩＭＤユニット１３８等）による実行のために、シェードプログラムをコンパイルするジャストインタイムコンパイラを含む。

【００１３】

ＡＰＤ１１６は、例えばグラフィックス処理及び非グラフィックス処理等のように、並列処理に適し得る選択された機能のためのコマンド及びプログラムを実行する。ＡＰＤ１１６は、例えばピクセル演算、幾何学的計算等のグラフィックスパイプライン処理を実行し、プロセッサ１０２から受信したコマンドに基づいて画像をディスプレイデバイス１１８にレンダリングするために使用することができる。また、ＡＰＤ１１６は、プロセッサ１０２から受信したコマンド、又は、グラフィックス処理パイプライン１３４の「正常な」情報フローの一部ではないコマンドに基づいて、例えばビデオ、物理シミュレーション、計算流体力学又は他のタスクに関係する動作等のようにグラフィックス処理に直接関連しない（若しくは関連していない）計算処理動作を実行する。

【００１４】

ＡＰＤ１１６は、ＳＩＭＤパラダイムに従って、プロセッサ１０２の要求に応じて並列に動作を実行するように構成された１つ以上のＳＩＭＤユニット１３８を有する計算ユニット１３２（本明細書では、まとめて「プログラマブル処理装置２０２」と呼ばれる）を含む。ＳＩＭＤパラダイムは、複数の処理要素が単一のプログラム制御フローユニット及びプログラムカウンタを共有することによって同一のプログラムを実行するが、異なるデータを用いて当該プログラムを実行することができるパラダイムである。一例では、各ＳＩＭＤユニット１３８は、１６個のレーンを含んでおり、各レーンは、ＳＩＭＤユニット１３８内の他のレーンと同時に同じ命令を実行するが、異なるデータを用いて当該命令を実行することができる。全てのレーンが所定の命令を実行する必要がない場合には、レーンを予測によってスイッチオフにすることができる。予測を使用して、分岐制御フローを有するプログラムを実行することもできる。より具体的には、制御フローが個々のレーンによって実行された計算に基づく条件付き分岐又は他の命令を有するプログラムの場合、現在実行されていない制御フローパスに対応するレーンの予測、及び、異なる制御フローパスの連続実行は、任意の制御フローに追従することを可能にする。計算ユニット１３２は、例えばＡＰＤ１１６内のＡＰＤメモリ１３９又はシステムメモリ１０４等のメモリから取得されたデータをキャッシュするキャッシュシステム１４０を含む。

【００１５】

計算ユニット１３２における実行の基本単位は、ワークアイテムである。各ワークアイテムは、特定のレーンで並列に実行されるプログラムの単一のインスタンス化（instantiation）を表す。ワークアイテムは、単一のＳＩＭＤユニット１３８上で「ウェーブフロント」として同時に実行することができる。複数のウェーブフロントは、同じプログラムを実行するように指定されたワークアイテムの集合を含む「ワークグループ」に含まれてもよい。ワークグループを構成するウェーブフロントの各々を実行することによって、ワークグループを実行することができる。ウェーブフロントは、単一のＳＩＭＤユニット１３８上で順次実行されてもよいし、異なるＳＩＭＤユニット１３８上で部分的に又は完全に並列に実行されてもよい。ウェーブフロントは、単一のＳＩＭＤユニット１３８上で同時に実行可能なワークアイテムの最大の集合と考えることができる。したがって、プロセッサ１０２から受信したコマンドが、特定のプログラムが単一のＳＩＭＤユニット１３８上で同時に実行できない程度に並列化されることを示す場合、当該プログラムは、２つ以上のＳＩＭＤユニット１３８で並列化され、又は、同じＳＩＭＤユニット１３８上で直列化（必要に応じて、並列化及び直列化の両方が行われる）されたウェーブフロントに分割される。スケジューラ１３６は、異なる計算ユニット１３２及びＳＩＭＤユニット１３８で様々なウェーブフロントをスケジュールすることに関連する動作を実行するように構成さ

10

20

30

40

50

れている。スケジューリングは、SIMDユニット138上で実行するためのウェーブフロントを割り当てること、ウェーブフロントがいつ終了したかを判別することと、ウェーブフロントがいつストールし、他のウェーブフロントとスワップアウトされるかを判別することと、他のスケジューリングタスクを実行すること、を含む。

【0016】

計算ユニット132によってもたらされる並列性は、例えば、ピクセル値計算、頂点変換及び他のグラフィックス動作等のグラフィックス関連動作に適している。プロセッサ102からグラフィックス処理コマンドを受信するグラフィックス処理パイプライン134は、並列実行のために計算タスクを計算ユニット132に提供する。

【0017】

また、計算ユニット132は、グラフィックスに関連しないか、グラフィックス処理パイプライン134の「通常」動作（例えば、グラフィックス処理パイプライン134の動作のために実行される処理を補足するのに実行されるカスタム動作）の一部として実行されない計算タスクを実行するために使用される。プロセッサ102で実行中のアプリケーション126又は他のソフトウェアは、係る計算タスクを定義するプログラム（「計算シェーダプログラム」とよく呼ばれる）を、実行のためにAPD116に送信する。

【0018】

図3は、図2に示すグラフィックス処理パイプライン134の更なる詳細を示すブロック図である。グラフィックス処理パイプライン134は、各々が特定の機能を実行するステージを含む。各ステージは、グラフィックス処理パイプライン134の機能の小区分を表す。各ステージは、プログラマブル処理装置202内で実行されるシェーダプログラムとして部分的若しくは完全に、又は、プログラマブル処理装置202の外部の固定機能である非プログラマブルハードウェアとして部分的若しくは完全に実装される。

【0019】

入力アセンブラステージ302は、ユーザが充填したバッファ（例えば、プロセッサ102によって実行されるソフトウェア（アプリケーション126等）の要求によって充填されたバッファ）からプリミティブデータを読み出し、パイプラインの残りの部分によって使用されるプリミティブに当該データをアセンブルする。入力アセンブラステージ302は、ユーザが充填したバッファに含まれるプリミティブデータに基づいて、異なるタイプのプリミティブを生成することができる。入力アセンブラステージ302は、残りのパイプラインによって使用されるために、アセンブルされたプリミティブをフォーマットする。

【0020】

頂点シェーダステージ304は、入力アセンブラステージ302によってアセンブルされたプリミティブの頂点を処理する。頂点シェーダステージ304は、変換、スキニング、モーフィング及び頂点毎のライティング等の様々な頂点毎の操作を実行する。変換操作は、頂点の座標を変換するための様々な操作を含むことができる。これらの操作には、モデリング変換、表示変換、投影変換、パースペクティブ分割及びビューポート変換のうち1つ以上が含まれてもよい。本明細書では、このような変換は、変換が実行される頂点の座標、つまり「位置」を変更するものとみなされる。頂点シェーダステージ304の他の操作によって、座標以外の属性が変更されてもよい。

【0021】

頂点シェーダステージ304は、1つ以上の計算ユニット132上で実行される頂点シェーダプログラムとして部分的又は完全に実装される。頂点シェーダプログラムは、プロセッサ102によって提供され、コンピュータプログラムによって事前に書かれたプログラムに基づいている。ドライバ122は、このようなコンピュータプログラムをコンパイルして、計算ユニット132内での実行に適したフォーマットを有する頂点シェーダプログラムを生成する。

【0022】

ハルシェーダステージ306、テッセレータステージ308及びドメインシェーダステ

10

20

30

40

50

ージ 3 1 0 は、連携してテッセレーションを実施し、プリミティブを更に分割することによって、単純なプリミティブをより複雑なプリミティブに変換する。ハルシェーダステージ 3 0 6 は、入力プリミティブに基づいてテッセレーションのためのパッチを生成する。テッセレータステージ 3 0 8 は、パッチのサンプルセットを生成する。ドメインシェーダステージ 3 1 0 は、パッチのサンプルに対応する頂点の頂点位置を計算する。ハルシェーダステージ 3 0 6 及びドメインシェーダステージ 3 1 0 は、プログラマブル処理装置 2 0 2 上で実行されるシェーダプログラムとして実装することができる。

【 0 0 2 3 】

ジオメトリシェーダステージ 3 1 2 は、プリミティブ毎に頂点操作を実行する。ジオメトリシェーダステージ 3 1 2 によって、例えば、ポイントスプライト展開、動的パーティクルシステム操作、ファーフィン (fur-fin) 生成、シャドウボリューム生成、キューブマップへのシングルパスレンダリング、プリミティブ毎のマテリアルスワップ、プリミティブ毎のマテリアル設定等の操作を含む、様々な異なるタイプの操作を実行することができる。ジオメトリシェーダステージ 3 1 2 の操作は、プログラマブル処理装置 2 0 2 上で実行されるシェーダプログラムによって行われてもよい。

【 0 0 2 4 】

ラスタライザステージ 3 1 4 は、上流で生成された単純プリミティブを受け入れてラスタライズする。ラスタライズは、特定のプリミティブによってカバーされているスクリーンピクセル (又はサブピクセルサンプル) を判別することから構成されている。ラスタライズは、固定機能ハードウェアによって行われる。

【 0 0 2 5 】

ピクセルシェーダステージ 3 1 6 は、上流で生成されたプリミティブとラスタライズの結果とに基づいて、スクリーンピクセルの出力値を計算する。ピクセルシェーダステージ 3 1 6 は、テクスチャをテクスチャメモリから適用することができる。ピクセルシェーダステージ 3 1 6 の操作は、プログラマブル処理装置 2 0 2 上で実行されるシェーダプログラムによって行われる。

【 0 0 2 6 】

出力マージアステージ 3 1 8 は、ピクセルシェーダステージ 3 1 6 からの出力を受信し、これらの出力を結合 (マージ) し、z テスト及びアルファブレンディング等の操作を行うことによって、スクリーンピクセルの最終色を決定する。

【 0 0 2 7 】

テクスチャを定義するテクスチャデータは、テクスチャユニット 3 2 0 によって記憶及び/又はアクセスされる。テクスチャは、グラフィックス処理パイプライン 1 3 4 の様々な時点で使用されるビットマップ画像である。例えば、場合によっては、ピクセルシェーダステージ 3 1 6 は、テクスチャをピクセルに適用して、レンダリングされる頂点の数を増やすことなく、見かけ上のレンダリングの複雑さを向上させる (例えば、より「写實的」な外観を提供する)。

【 0 0 2 8 】

場合によっては、頂点シェーダステージ 3 0 4 は、テクスチャユニット 3 2 0 のテクスチャデータを使用して、例えば、美観を向上させるために頂点を生成又は変更することによって複雑さを高めるように、プリミティブを変更する。一例では、頂点シェーダステージ 3 0 4 は、テクスチャユニット 3 2 0 に記憶されている高さマップを使用して、頂点の変位を変更する。このタイプの技術は、例えば、水をレンダリングするのに使用される頂点の位置及び数を変更することによって、ピクセルシェーダステージ 3 1 6 のみで使われるテクスチャと比較してより現実的に見える水を生成するために使用することができる。場合によっては、ジオメトリシェーダステージ 3 1 2 は、テクスチャユニット 3 2 0 のテクスチャデータにアクセスする。

【 0 0 2 9 】

図 4 は、一例による、キャッシュシステム 1 4 0 を示す図である。キャッシュシステム 1 4 0 は、SIMD ユニット 1 3 8 で実行されているウェーブフロントからキャッシュア

10

20

30

40

50

クセスの要求を受信し、これらの要求を処理する。これらの要求を処理することの一部は、計算ユニット 1 3 2 内のキャッシュシステム 1 4 0 の 1 つ以上のキャッシュメモリ 4 0 4 において要求されたデータを検索することを含む。より具体的には、キャッシュシステム 1 4 0 は、キャッシュ階層の低レベル（第 1 レベル等）として機能する 1 つ以上のキャッシュメモリ 4 0 4 を有する。ウェブフロント 4 1 2 によってアクセス要求されたデータがキャッシュメモリ 4 0 4 に存在しない場合、キャッシュシステム 1 4 0 は、要求されたデータにアクセスするために、階層インタフェース 4 0 6 を介して、キャッシュ階層内の他のメモリ（例えば、より高レベルのキャッシュメモリ、A P D メモリ 1 3 9、及び/又は、システムメモリ 1 0 4）にアクセスする。

【 0 0 3 0 】

いくつかの例では、キャッシュアクセスの要求は、ベクトルベースの要求である。ベクトルベースの要求は、S I M D ユニット 1 3 8 での並列化された操作に従って、複数のメモリ位置からデータを要求する能力を有する。例えば、各ワークアイテムが異なるアドレスを指定する、単一のウェブフロント内の異なるワークアイテムによって実行されるロード命令等の単一の命令は、メモリ内の複数の位置からの読み出しをもたらし得る。ベクトルベースの要求は、通常、データを単一のメモリ位置に読み書きするスカラー要求とは対照的である。

【 0 0 3 1 】

キャッシュシステム 1 4 0 内のリターンオーダリングキュー 4 0 2 は、完了したキャッシュアクセスが順番にウェブフロント 4 1 2 に返されるように、完了したキャッシュアクセスをオーダリングする。アクセスを順番に返すことは、これらのアクセスが行われた順番で、アクセスを行ったウェブフロント 4 1 2 に対してリターンデータを提供することを意味する。いくつかの例では、キャッシュシステム 1 4 0 は、ウェブフロント毎にキャッシュアクセス要求をオーダリングする。これは、特定のウェブフロントによって行われたキャッシュアクセス要求の指標が、これらの要求を行った順序でウェブフロント 4 1 2 が記憶されるが、異なるウェブフロント 4 1 2 によって行われたキャッシュアクセス要求の順序が保持されないことを意味する。キャッシュアクセス要求のウェブフロント毎のオーダリングを達成するために、各ウェブフロント 4 1 2 は、キューメモリ空間の独自のセットに割り当てられて、キャッシュアクセス要求の順序を保持することができる。

【 0 0 3 2 】

リターンオーダリングキュー 4 0 2 は、各キャッシュアクセスの識別子を記憶し、この識別子は、ウェブフロント 4 1 2 によって各キャッシュアクセスが行われた順序で記憶される。新たなエントリ、すなわち、ウェブフロント 4 1 2 からの最新のキャッシュアクセス要求に対応するエントリは、リターンオーダリングキュー 4 0 2 の末尾に提供される。リターンオーダリングキューの先頭には、最も古い未処理のキャッシュアクセス要求、したがって、返される次のキャッシュアクセスの識別情報が記憶されている。リターンオーダリングキュー 4 0 2 の先頭におけるアクセスに対するデータが、キャッシュメモリ 4 0 4 において利用可能である場合（例えば、データの少なくとも一部が、より高レベルのキャッシュメモリ、システムメモリ 1 0 4 又は A P D メモリ 1 3 9 からキャッシュメモリ 4 0 4 にフェッチされている場合）、キャッシュシステム 1 4 0 は、当該データを要求元のウェブフロント 4 1 2 に提供する。

【 0 0 3 3 】

キャッシュリターンをオーダリングするための 1 つの技術は、キャッシュリターンの「タイプ」に関係なく、全てのキャッシュリターンをオーダリングすることを含む。本明細書において、キャッシュリターンの「タイプ」（これは、キャッシュアクセス要求のタイプでもある）は、読み出しタイプ、書き込みタイプ、テクスチャサンプリングタイプの何れかである。読み出しタイプアクセスは、メモリからのデータを要求し、要求されたデータがメモリから返されるアクセスである。読み出しタイプのアクセスでは、仮想アドレス又は物理アドレスによってメモリ位置が指定され、サンプリング操作とは異なり、テクスチャ座標

10

20

30

40

50

が指定されない。読み出しタイプのアクセスに対するリターンは、要求されたデータを、当該データを要求したウェーブフロント 4 1 2 に返すこと（例えば、要求されたデータを、ウェーブフロント 4 1 2 を実行する SIMD ユニット 1 3 8 のレジスタに配置することによって、SIMD ユニット 1 3 8 が当該データに依存する命令を実行できるようにすること）を意味する。読み出しタイプのアクセスの場合、リターンデータは、読み出しタイプのアクセス要求によって要求されたデータである。

【0034】

書き込みタイプのアクセスに対するリターンは、書き込みを要求したウェーブフロント 4 1 2 を実行する SIMD ユニット 1 3 8 に対して、「肯定応答」信号を返すことを意味する。「肯定応答」信号は、書き込みを要求した SIMD ユニット 1 3 8 に対して、要求された書き込みがメモリシステムによって承認されたことを示す信号である。SIMD ユニット 1 3 8 は、「肯定応答」信号を受信すると、「肯定応答」信号に依存する動作を進めることができるようになる。書き込みタイプのアクセスの場合、リターンデータは、「肯定応答」信号である。

10

【0035】

他の操作と同様に、（読み出し変更書き込み（read-modify-write）操作等の複雑な操作であってもよい）アトミック操作は、操作の完了に応じて返された信号のタイプに基づいて、読み出し又は書き込みとして分類される。より具体的には、データをウェーブフロントに返すアトミック操作は、読み出し操作として分類される。肯定応答信号をウェーブフロントに返すアトミック操作は、書き込み操作として分類される。

20

【0036】

テクスチャサンプラタイプのメモリ操作は、テクスチャデータの要求を受信することと、当該要求に対して処理を実行して、実際に必要とされる記憶データを判別することと、当該記憶データをフェッチすることと、任意に記憶データを解凍及び／又はフィルタリングして、ピクセル値又はサンプル値を取得することと、データを要求したウェーブフロント 4 1 2 を実行する SIMD ユニット 1 3 8 に対してピクセル値又はサンプル値を返すことと、を含む複雑な操作である。テクスチャサンプラタイプのアクセスに対するリターンは、データを要求したウェーブフロント 4 1 2 を実行する SIMD ユニット 1 3 8 に対して、要求されたデータを返すことを意味する。テクスチャサンプラタイプのアクセスの場合、リターンデータは、要求元の SIMD ユニット 1 3 8 に返されるピクセル値又はサンプル値である。

30

【0037】

テクスチャサンプラ操作には多数の操作が含まれるため、レイテンシが長くなる。例えば、テクスチャサンプラアクセスの要求は、通常、データが望まれるテクスチャビットマップ内の位置を識別するテクスチャ座標（ u, v, w ）を含む。また、テクスチャサンプラアクセスの要求は、スクリーン座標の変化率と比較してテクスチャ座標の変化率を指定する 1 つ以上の勾配値を含んでもよく、他のデータ（例えば、ミップマップレベルを識別するデータ、キューブマップ面を識別するデータ又は他のデータ等）を含んでもよい。キャッシュシステム 1 4 0 のテクスチャユニット 4 0 8 は、このデータに基づいて、要求されたデータが見つかるアドレスを識別し、メモリシステムから当該データをフェッチし（計算ユニット 1 3 2 内のキャッシュシステム 1 4 0 のキャッシュメモリ 4 0 4 にヒットするか、キャッシュメモリ 4 0 4 において欠落していることを検出し、欠落しているデータをキャッシュ階層内の他のメモリからキャッシュメモリ 4 0 4 にフェッチすることを含んでもよい）、テクスチャデータが圧縮されていることがあるので、オプションでデータを解凍し、オプションでフィルタリングスキーム（例えば、バイリニアフィルタリング、トリリニアフィルタリング、異方性フィルタリング）に従ってデータを処理し、テクスチャサンプラキャッシュアクセス要求を行ったウェーブフロント 4 1 2 を実行する SIMD ユニット 1 3 8 に対してデータを送る。これらの操作は、何百ものコンピュータクロックサイクルを必要とする可能性があり、これは長いレイテンシ（ここで、レイテンシとは、SIMD ユニット 1 3 8 がテクスチャサンプラ操作の実行を要求してから、テクスチャサ

40

50

ンプリング操作によって得られたデータがSIMDユニット138に提供されるまでの時間である)を表す。

【0038】

読み出し操作又は書き込み操作と比較してサンブラ操作のレイテンシが長いので、キャッシュリターンをオーダリングするための他のモードが本明細書で提供される。このような他のモードでは、全てのタイプのキャッシュリターンをオーダリングする単一のキューの代わりに、リターンオーダリングキュー402が、異なるタイプのキャッシュリターンの順序を保持する2つ以上の個別のキューを含む。いくつかのモードが可能である。異なるモード間の選択は、プロセッサ102からの要求、SIMDユニット138内で実行される命令、APD116内で実行されるファームウェアによって行われるアルゴリズム的

10

【0039】

或るモードでは、「サンブラ」、「読み出し」、及び、「書き込み」のタイプのキャッシュアクセス要求毎にキューが保持される。各キューは、各々のタイプのキャッシュアクセスの指標を記憶する。例えば、サンブラキューは、アクセスが行われた順で、サンブラタイプのキャッシュアクセスのアクセス指標を保持するが、読み出しタイプ又は書き込みタイプのキャッシュアクセスのアクセス指標を保持しない。同様に、読み出しキューは、読み出しタイプのキャッシュアクセスの順序付きの指標を保持するが、書き込みタイプ又はサンブラタイプのキャッシュアクセスの順序付きの指標を保持せず、書き込みキューは、書き込みタイプのキャッシュアクセスの順序付きの指標を保持するが、読み出しタイプ又はサンブラタイプのキャッシュアクセスの順序付きの指標を保持しない。

20

【0040】

各キューは、先頭と末尾とを有する。先頭は、そのキューに対応するタイプの最も古いキャッシュアクセス要求を表し、末尾は、そのキューに対応するタイプの最も新しいキャッシュアクセス要求を表す。特定のキューの先頭にあるキャッシュアクセス要求の全てのデータがキャッシュメモリ404に記憶されることによって「利用可能」になると、キャッシュシステム140は、当該データを更なる処理のためにSIMDユニット138に返す。複数キューの先頭にあるキャッシュアクセスが利用可能な場合には、キャッシュシステム140は、技術的に実現可能なアービトラジョンスキーム(例えば、ラウンドロビン)に従って、異なるキューの先頭のキャッシュアクセスが互いに発行された順序に関係なく、何れかのキューの先頭にある利用可能なキャッシュアクセスを選択することができる。

30

【0041】

別のモードでは、サンブラ操作及び読み出し操作を指示する1つのキューが保持され、書き込み操作を指示する別のキューが保持される。つまり、第1キューは、「読み出し」タイプのキャッシュアクセス要求と「サンブラ」タイプのキャッシュアクセスとの両方の指示を、これらのアクセス要求がウェブフロントから受信された順序で記憶し、第2キューは、「書き込み」タイプのキャッシュアクセスの指示を、これらのアクセス要求がウェブフロントから受信された順序で記憶する。したがって、読み出しタイプのキャッシュアクセスとサンブラタイプのキャッシュアクセスとは、互いに順序付けられているが、書き込みタイプのキャッシュアクセスについては順序付けられていない。書き込みタイプのキャッシュアクセスは、互いに順序付けられているが、読み出しタイプのキャッシュアクセス又はサンブラタイプのキャッシュアクセスについては順序付けられていない。

40

【0042】

読み出し及び書き込みをオーダリングするキューと、サンブラ操作をオーダリングする別のキューとを有する1つのモードや、書き込み及びサンブラ操作をオーダリングするキューと、読み出しをオーダリングする別のキューとを有する別のモード等の他のモードも可能である。

【0043】

図5は、一例による、様々なタイプのキャッシュアクセス要求をオーダリングするため

50

のキューを示すブロック図である。キューは、読み出しタイプのキャッシュアクセス要求を表すエントリ 5 1 4 を、これらの要求がウェーブフロント 4 1 2 によって行われる順序で記憶する読み出しキュー 5 0 2 と、書き込みタイプのキャッシュアクセス要求を表すエントリ 5 1 4 を、これらの要求がウェーブフロント 4 1 2 によって行われる順序で記憶する書き込みキューと、サンブラタイプのキャッシュアクセス要求を表すエントリ 5 1 4 を、これらの要求がウェーブフロント 4 1 2 によって行われる順序で記憶するサンブラキュー 5 0 6 と、を含む。各エントリ 5 1 4 は、1 つ以上のキャッシュアクセス要求に関するデータを記憶することができる。

【0044】

動作中、ウェーブフロント 4 1 2 は、キャッシュアクセス要求を含み得るシェーダプログラムの一部として実行される。このような要求が行われると、キャッシュシステム 1 4 0 は、当該要求を検出し、当該要求が読み出しタイプの要求であるか、書き込みタイプの要求であるか、又は、サンブラタイプの要求であるかを判別する。読み出しタイプ要求は、メモリからウェーブフロント 4 1 2 にデータを返すサンブラタイプ要求以外の要求である。書き込みタイプ要求は、データをメモリに記憶し、肯定応答信号をウェーブフロント 4 1 2 に返すが、メモリからウェーブフロント 4 1 2 にデータを返さない要求である。サンブラタイプ要求は、少なくともテクスチャ座標を受信し、当該座標及び他のデータを処理して、テクスチャデータを記憶するメモリ内の 1 つ以上のメモリ位置を識別し、テクスチャデータを取得し、テクスチャデータをウェーブフロント 4 1 2 に返す要求である。

【0045】

キャッシュシステム 1 4 0 は、キャッシュアクセス要求が何れのタイプであるかを判別したことに応じて、要求を表すエントリ 5 1 4 を適切なキュー（例えば、読み出しキュー 5 0 2、書き込みキュー 5 0 4 又はサンブラキュー 5 0 6）の末尾 5 1 0 に配置する。エントリ 5 1 4 は、ウェーブフロント 4 1 2 に返されるこれらのエントリ 5 1 4 に対するデータ又は書き込み肯定応答によって他のエントリがキューの先頭 5 1 2 から除去されるにつれて、各キューの先頭 5 1 2 に向かって移動する。要求を表すエントリ 5 1 4 が適切なキューにある間、キャッシュシステム 1 4 0 は、要求を満たすように動作する。

【0046】

読み出し要求の場合、キャッシュシステム 1 4 0 は、読み出されるように要求されたデータがキャッシュメモリ 4 0 4 に存在するかどうかを判別する。データがキャッシュメモリ 4 0 4 に存在する場合には、キャッシュシステム 1 4 0 は、対応するエントリ 5 1 4 を「利用可能」としてマークし、エントリ 5 1 4 が読み出しキュー 5 0 2 の先頭 5 1 2 にある場合に、適切なキャッシュリターンをウェーブフロント 4 1 2 に提供できることを示す。データが未だキャッシュメモリ 4 0 4 に存在しない場合には、対応するエントリ 5 1 4 は、利用可能として未だマークされていない。キャッシュシステム 1 4 0 は、階層インタフェース 4 0 6 を介して、キャッシュ階層の他のレベルから適切なデータをフェッチする。フェッチした後に全てのデータがキャッシュメモリ 4 0 4 で利用可能になった場合、キャッシュシステム 1 4 0 は、エントリ 5 1 4 を利用可能としてマークし、エントリ 5 1 4 が読み出しキュー 5 0 2 の先頭 5 1 2 にある場合にデータをウェーブフロント 4 1 2 に返す。

【0047】

サンブラ要求の場合、キャッシュシステム 1 4 0 は、サンブラ要求に対して適切な処理を実行して、サンブラ要求を満たす 1 つ以上のメモリアドレスを識別する。サンブラ要求のメモリアドレスを識別する処理が完了した後に、キャッシュシステム 1 4 0 は、サンブラ要求のデータがキャッシュメモリ 4 0 4 に存在するかどうかを判別する。データがキャッシュメモリ 4 0 4 に存在する場合には、キャッシュシステム 1 4 0 は、対応するエントリ 5 1 4 を利用可能としてマークする。データがキャッシュメモリ 4 0 4 に存在しない場合には、対応するエントリ 5 1 4 が利用可能としてマークされず、キャッシュシステム 1 4 0 は、階層インタフェース 4 0 6 を介してキャッシュ階層からデータをフェッチする。サンブラ要求のデータがキャッシュメモリ 4 0 4 に存在することに依拠して、キャッシュシ

10

20

30

40

50

ステム 140 は、対応するエントリ 514 を利用可能としてマークする。対応するエントリ 514 が利用可能であるとマークされ、当該エントリ 514 がサンブラキュー 506 の先頭 512 にある場合、キャッシュシステム 140 は、サンブラ要求に対するデータをウェーブフロント 412 に返す。

【0048】

書き込み要求の場合、キャッシュシステム 140 は、データをメモリシステムに書き込む。具体的には、キャッシュシステム 140 は、データをキャッシュメモリ 404 に書き込む。その後、データは、階層インタフェース 406 を介してキャッシュメモリ階層内の他のメモリに書き込まれる。データが、キャッシュメモリ 404、又は、階層インタフェース 406 を介して到達可能な外部キャッシュメモリ内の 1 つ以上の他のメモリに書き込まれると（例えば、データが、キャッシュメモリ 404 からキャッシュ階層の 1 つ上位のレベルのキャッシュメモリに書き込まれると）、キャッシュシステム 140 は、書き込まれたデータに対応するエントリ 514 を利用可能としてマークする。利用可能なエントリ 514 が書き込みキュー 504 の先頭 512 にあることに応じて、キャッシュシステム 140 は、当該エントリ 514 を書き込みキュー 504 から除去し、肯定応答信号をウェーブフロント 412 に返す。

【0049】

図 6 は、一例による、リターンオーダリングキュー 600 を示す図である。リターンオーダリングキュー 600 は、図 4 のリターンオーダリングキュー 402 の一例である。リターンオーダリングキュー 600 は、読み出しキュー 502 と、書き込みキュー 504 と、サンブラキュー 506 と、を含む。読み出しキュー 502、書き込みキュー 504 及びサンブラキュー 506 の各々は、モノリシック (monolithic) メモリ 602 内の仮想キューとして実装される。「仮想キュー」という用語は、読み出しキュー 502、書き込みキュー 504 及びサンブラキュー 506 の各々が、モノリシックメモリ 602 内の固定位置又は長さを有さず、モノリシックメモリ 602 内の異なるスロット 604 (「スロット 604」という用語は、1 つのエントリ 514 を記憶するように区分されたモノリシックメモリ 602 の最小単位を指す) で開始又は終了し得ることを意味している。異なるキューが背中合わせに示されているが（例えば、先頭 512 が、末尾 510 に隣接するスロット 604 にある）、キューの先頭 512 は、他のキューの末尾 510 に隣接していない場合がある。一例では、読み出しキュー 502 には、モノリシックメモリ 602 内の一定数のスロット 604 が割り当てられているが、読み出しキュー 502 の先頭 512 (1) は、割り当てられたスペースの最後まで延びていない。このような例では、読み出しキュー 502 の先頭 512 (1) は、書き込みキュー 504 の末尾 510 (2) に隣接しないことになる。別の例では、末尾 510 は、特定のキューの先頭に存在しない。キャッシュシステム 140 は、必要に応じて各仮想キューのサイズを変更することができ、特定の状況では、1 つ以上のキューに対してゼロスロット 604 を割り当てることができる。一例では、読み出し及びサンブラ操作が互いにオーダリングされ、書き込みが互いにオーダリングされるが、読み出し及びサンブラ操作が書き込みに対してオーダリングされないモードでは、キャッシュシステム 140 は、2 つのキュー (1 つは読み出し及びサンブラ操作用、もう 1 つは書き込み用) に対して 0 より多いスロット 604 を割り当て、第 3 キューに対してゼロスロット 604 を割り当てる。

【0050】

先頭抽出器 (head extractor) 610 は、現在のモードに対してアクティブである各キューの先頭にあるデータを抽出し、当該データを各々の物理キューに配置する。物理読み出しキュー 604 と、物理書き込みキュー 606 と、物理サンブラキュー 608 と、を含む物理キューは、アービタ 612 が、特定のキューの先頭 512 を容易に読み出し、キャッシュシステム 140 の他の部分に提供してウェーブフロント 412 に返すことを可能にする。例えば、物理キューがない場合、アービタは、まず、キューの先頭 512 のアドレスを検索し、次に、当該先頭 512 に対応するエントリ 514 内のデータを取得する。先頭抽出器 610 が各キューの先頭にあるエントリ 514 を抽出すると、アービタ 612 は

10

20

30

40

50

、固定されたメモリ位置（物理キューの先頭）を調べて、次の利用可能なキャッシュメモリアクセスのためのデータを取得する。物理キューは、エントリ 6 1 4 を記憶するための 1 つ以上のスロット 6 1 3 を含む。物理キュー内のエントリ 6 1 4 は、仮想キュー（例えば、モノリシックメモリ 6 0 2 内の読み出しキュー 5 0 2、書き込みキュー 5 0 4 及びサンブラキュー 5 0 6）内のエントリ 5 1 4 と同一又は類似のデータを含む。各物理キューの先頭 6 2 2 は、対応する仮想キュー内の対応する先頭 5 1 2 のエントリ 5 1 4 のデータを有するエントリ 6 1 4 を記憶する。

【 0 0 5 1 】

アービタ 6 1 2 は、エントリ 6 1 4 が利用可能としてマークされている場合に、物理キューの先頭 6 2 2 からエントリ 6 1 4 を選択する。異なる物理キューの先頭 6 2 2 にある複数のエントリ 6 1 4 が利用可能であるとマークされている場合には、アービタ 6 1 2 は、技術的に実現可能なアービトレーションスキーム（例えば、ラウンドロビン等）を適用して、ウェーブフロント 4 1 2 へのキャッシュリターンのために、先頭 6 2 2 にあるエントリ 6 1 4 のうち 1 つのエントリを選択する。

10

【 0 0 5 2 】

アービタ 6 1 2 が物理キューの先頭 6 2 2 からエントリ 6 1 4 を読み出すと、物理キューは他のエントリ 6 1 4 を先頭 6 2 2 に向けて移動させ、先頭抽出器 6 1 0 は、モノリシックメモリ 6 0 2 内の対応する仮想キュー内の対応するスロット 6 0 4 から別のエントリ 5 1 4 を読み出し、先頭抽出器 6 1 0 によって読み出されたエントリ 5 1 4 のデータを含む新たなエントリ 6 1 4 を、適切な物理キュー内の新たに空いたスロットに配置する。

20

【 0 0 5 3 】

図 7 は、一例による、アクセスタイプ毎のリターンオーダリングキューにキャッシュアクセス要求の指示を記録する方法 7 0 0 のフロー図である。図 1 ~ 図 6 に関して示し、説明したシステムについて説明するが、技術的に実行可能な任意の順序でこの方法を実行するように構成された何れのシステムも、本発明の範囲に含まれることを理解されたい。

【 0 0 5 4 】

図示するように、方法 7 0 0 は、ステップ 7 0 2 で開始し、キャッシュシステム 1 4 0 は、ウェーブフロント 4 1 2 からキャッシュアクセス要求を受信する。キャッシュアクセス要求は、キャッシュシステム 1 4 0 を介してメモリにアクセスするための要求である。ステップ 7 0 4 において、キャッシュシステム 1 4 0 は、キャッシュアクセス要求が読み出しタイプ、書き込みタイプ、テクスチャサンブラタイプのうち何れであるかを判別する。読み出しタイプは、1 つ以上の位置からデータを読み出し、当該データを、そのアクセスを行ったウェーブフロント 4 1 2 に返すタイプのアクセスである。書き込みタイプは、データを特定の位置に書き込み、そのアクセスを行ったウェーブフロント 4 1 2 に対して「肯定応答」信号を返すタイプのアクセスである。テクスチャサンブラタイプは、テクスチャ座標と、場合によっては他のテクスチャ関連の入力と、を処理し、テクスチャ座標と、場合によっては他のテクスチャ関連入力と、に基づいて 1 つ以上のメモリアドレスを識別し、1 つ以上のメモリアドレスからデータをフェッチし、オプションでデータを処理して当該データを解凍及び/又はデータをフィルタリング（例えば、バイリニアフィルタリング、トリリニアフィルタリング、異方性フィルタリング）し、フェッチされ処理されたテクスチャデータを、アクセスを行ったウェーブフロント 4 1 2 に提供するタイプのアクセスである。

30

40

【 0 0 5 5 】

ステップ 7 0 6 において、キャッシュシステム 1 4 0 は、キャッシュアクセス要求の指標を、そのタイプに対応するタイプ毎のリターンオーダリングキューに入れる。例えば、キャッシュアクセス要求が書き込みタイプである場合、キャッシュシステム 1 4 0 は、そのキャッシュアクセス要求の指標を、書き込みタイプに対応するリターンオーダリングキューに入れる。各キューは、キャッシュアクセス要求を受信した順序でキャッシュアクセス要求を保持する。最も古いキャッシュアクセス要求は、キューの先頭にあり、最も新しいキャッシュアクセス要求は、キューの末尾にある。各キューは、キャッシュシステム 1

50

40に設定されたモードに応じて、1種類、2種類又は3種類の要求をオーダリングすることができる。

【0056】

図8は、一例による、アウトオブオーダキャッシュリターンをワークグループに提供する方法800のフローチャートである。図1～図6に関連して示し、説明したシステムについて説明するが、技術的に実行可能な任意の順序でこの方法を実行するように構成された何れのシステムも、本発明の範囲に含まれることを理解されたい。

【0057】

図示するように、方法800は、ステップ802で開始し、キャッシュシステム140は、何れかのタイプ別のオーダリングキューの先頭のエントリを、関連するキャッシュアクセス要求を行ったウェーブフロントに返す準備ができてい

10

かどうかを判別する。読み出し又はサンブラ操作によって要求されたデータがキャッシュシステムのキャッシュメモリ404に記憶されている場合には、読み出し又はサンブラ操作の準備ができてい

20

いくつかの例では、書き込みが行われるデータが他のメモリトランザクションに関して適切にオーダリングされていることが保証されていることを示す肯定応答を受信した場合、その時点で他のメモリトランザクションは、その書き込みの効果を「見る」ことができるので、書き込みは準備完了である。他の例では、書き込みが行われるデータがキャッシュ階層内のキャッシュシステム140のキャッシュメモリ404の少なくとも1つ上位のレベルにある1つ以上のメモリに記憶されていることを示す肯定応答を受信した場合、書き込みは準備完了である。

【0058】

ステップ804において、キャッシュシステム140は、ウェーブフロント412に返す準備ができてい

キューの先頭512にあるエントリ514を選択する。複数のキューが準備のできたエントリ514を有する先頭512を有している場合には、キャッシュシステム140は、エントリ514のうち1つのエントリをリターンのために選択する。この選択は、技術的に実現可能なアービトレーションスキーム（例えば、ラウンドロビン）に基づいて行うことができる。

【0059】

ステップ806において、キャッシュシステム140は、選択されたエントリ514に対応するキャッシュリターンを、キャッシュアクセスを要求したウェーブフロント412に提供する。読み出し及びサンブラ操作の場合、リターンは、要求されたデータである。書き込みの場合、リターンは、肯定応答信号である。ステップ808において、キャッシュシステム140は、選択されたエントリ514を含むキューから選択されたエントリ514を除去し、先頭512が次に古いエントリ514を指すようにキューを変更する。

30

【0060】

本明細書で説明した技術は、ウェーブフロント412がアクセスタイプ毎にメモリバリア操作を実行することを可能にする。メモリバリア操作は、バリア操作までの全てのメモリ操作についてリターンを受信するまで、ウェーブフロント412をストールさせる操作である。アクセスタイプ毎ベースのキュー及びオーダリングは、ウェーブフロント412がアクセスタイプ毎のバリア命令を実行することを可能にする。例えば、読み出しバリアは、未処理の書き込み操作又はサンブラ操作があるかどうかに関係なく、処理を進める前に、全ての未処理の読み出し操作ののリターンを受信するのをウェーブフロント412に待機させる。同様に、書き込みバリア又はサンブラバリアは、処理を進める前に、全ての未処理の書き込み又は未処理のサンブラを受信するのをウェーブフロント412に待機させる。

40

【0061】

本明細書で説明した技術は、メモリアクセス要求がタイプ毎にオーダリングされているが、キャッシュアクセスタイプ間でアウトオブオーダでリターンされ得るモードを提供することによって、メモリレイテンシを改善する。したがって、或るタイプ（例えば、読み出し）のキャッシュアクセスは、別のタイプ（例えば、テクスチャサンブラ操作）のキャ

50

ッシュアクセスを待つ必要がない。このオーダリングの緩和によって、例えば、或るタイプのレイテンシが別のタイプのレイテンシよりも長い場合に、キャッシュアクセスのレイテンシを短縮することができる。一例では、読み出し操作は、比較的短いレイテンシを有し、テクスチャサンブラ操作は、比較的長いレイテンシを有する。キャッシュアクセスが古いテクスチャサンブラ操作のためにリターンする前に、新たな読み出し操作のためのキャッシュリターンを許可することによって、読み出し操作のレイテンシは、新たな読み出し操作が古いテクスチャサンブラ操作を待たなければならない状況と比較して、短縮する。

【 0 0 6 2 】

アウトオブオーダキャッシュリターンを実行する方法が提供される。方法は、複数のリターンオーダリングキューのうち第1リターンオーダリングキューの先頭にある第1エン
10
トリが、ウェーブフロントへのリターンに利用可能であると判断することによって、第1エン
トリは、第1キャッシュアクセス要求に対応する、ことを含む。第1リターンオー
ダリングキューは、第2キャッシュアクセスタイプではなく、第1キャッシュアクセスタイ
プのキャッシュアクセス要求のエントリを記憶し、複数のリターンオーダリングキューの
うち第2リターンオーダリングキューは、第1キャッシュアクセスタイプではなく、第2
キャッシュアクセスタイプのキャッシュアクセス要求のエントリを記憶する。また、方法
は、判断したことに応じて、第1キャッシュアクセス要求よりも古い第2リターンオー
ダリングキューのエントリに対応するキャッシュアクセス要求がウェーブフロントへのリ
ターンに利用可能になるのを待つことなく、第1エントリに対応するキャッシュリターンを
20
ウェーブフロントに送ることを含む。

【 0 0 6 3 】

アウトオブオーダキャッシュリターンを実行するための計算ユニットが提供される。計
算ユニットは、ウェーブフロントを実行するように構成された単一命令複数データユニ
ットと、キャッシュシステムと、を含む。キャッシュシステムは、第1リターンオーダリ
ングキューと、第2リターンオーダリングキューと、を含む複数のリターンオーダリ
ングキューを記憶することによって、第1リターンオーダリングキューは、第2キャ
ッシュアクセスタイプではなく、第1キャッシュアクセスタイプのキャッシュアクセス
要求のエントリを記憶し、第2リターンオーダリングキューは、第1キャッシュア
クセスタイプではなく、第2キャッシュアクセスタイプのキャッシュアクセス要求の
エントリを記憶すること
と、第1リターンオーダリングキューの先頭にある第1エントリが、ウェーブフ
ロントへのリターンに利用可能であると判断することによって、第1エントリは、第1
キャッシュアクセス要求に対応することと、判断したことに応じて、第1キャ
ッシュアクセス要求よりも古い第2リターンオーダリングキューのエントリに対
応するキャッシュアクセス要求がウェーブフロントへのリターンに利用可能になる
のを待つことなく、第1エントリに対応するキャッシュリターンをウェーブフ
ロントに送ることと、を行うように構成されている。
30

【 0 0 6 4 】

また、コンピュータシステムが提供される。コンピュータシステムは、計算ユニ
ットを含むアクセラレーテッド処理デバイスと、計算ユニット内のウェーブフ
ロントをアクセラレーテッド処理デバイスに実行させるように構成されたプロ
セッサと、を含む。計算ユニ
ットは、ウェーブフロントを実行するように構成された単一命令複数データユニ
ットと、キャッシュシステムと、を含む。キャッシュシステムは、第1リターンオー
ダリングキューと、第2リターンオーダリングキューと、を含む複数のリターン
オーダリングキューを記憶することによって、第1リターンオーダリングキュー
は、第2キャッシュアクセスタイプではなく、第1キャッシュアクセスタイプの
キャッシュアクセス要求のエントリを記憶し、第2リターンオーダリング
キューは、第1キャッシュアクセスタイプではなく、第2キャッシュア
クセスタイプのキャッシュアクセス要求のエントリを記憶することと、第
1リターンオーダリングキューの先頭にある第1エントリが、ウェーブフ
ロントへのリターンに利用可能であると判断することによって、第1エントリは、第1
キャッシュアクセス要求に対応することと、判断したことに応じて、第1
キャッシュアクセス要求よりも古
40
50

い第2リターンオーダリングキューのエントリに対応するキャッシュアクセス要求がウェブフロントへのリターンに利用可能になるのを待つことなく、第1エントリに対応するキャッシュリターンをウェブフロントに送ることと、を行うように構成されている。

【0065】

本明細書の開示に基づいて多くの変形が可能であることを理解されたい。機能及び要素を特定の組み合わせで上記のように説明されているが、各機能又は要素は、他の機能や要素なしに単独で使用されてもよいし、他の機能や要素を伴って若しくは伴わずに様々な組み合わせで使用されてもよい。

【0066】

提供された方法には、汎用コンピュータ、プロセッサ又はプロセッサコアにおける実施が含まれる。適切なプロセッサには、例として、汎用プロセッサ、専用プロセッサ、従来のプロセッサ、デジタル信号プロセッサ(DSP)、複数のマイクロプロセッサ、DSPコアに関連する1つ以上のマイクロプロセッサ、コントローラ、マイクロコントローラ、特定用途向け集積回路(ASIC)、フィールドプログラマブルゲートアレイ(FPGA)回路、任意の他のタイプの集積回路(IC)、及び/又は、ステートマシンが含まれる。このようなプロセッサは、処理されたハードウェア記述言語(HDL)命令、及び、ネットリストを含む他の中間データ(このような命令はコンピュータ可読媒体に記憶され得る)の結果を使用して製造プロセスを構成することによって製造されてもよい。このような処理の結果は、本実施形態の態様を実施するプロセッサを製造するために半導体製造プロセスにおいて使用されるマスクワークであってもよい。

【0067】

本明細書で提供される方法又はフロー図は、汎用コンピュータ若しくはプロセッサによる実行のために非一時的なコンピュータ可読記憶媒体に組み込まれたコンピュータプログラム、ソフトウェア、ファームウェアにおいて実施されてもよい。非一時的なコンピュータ可読記憶媒体の例には、ROM、ランダムアクセスメモリ(RAM)、レジスタ、キャッシュメモリ、半導体メモリデバイス、内蔵ハードディスク及びリムーバブルディスク等の磁気媒体、光磁気媒体、光学媒体(CD-ROMディスク等)、デジタル多用途ディスク(DVD)等が含まれる。

10

20

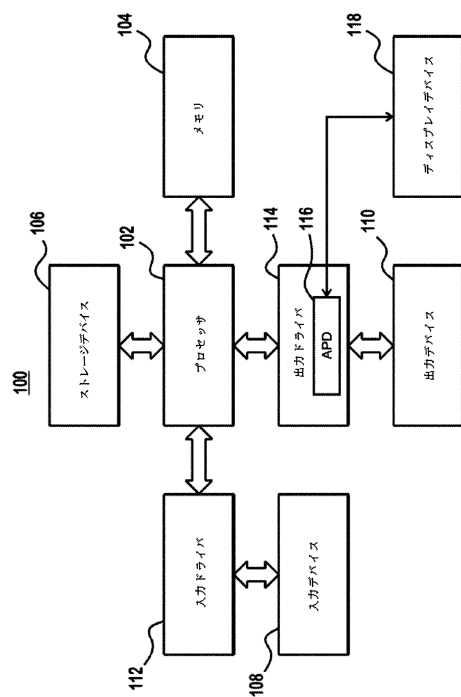
30

40

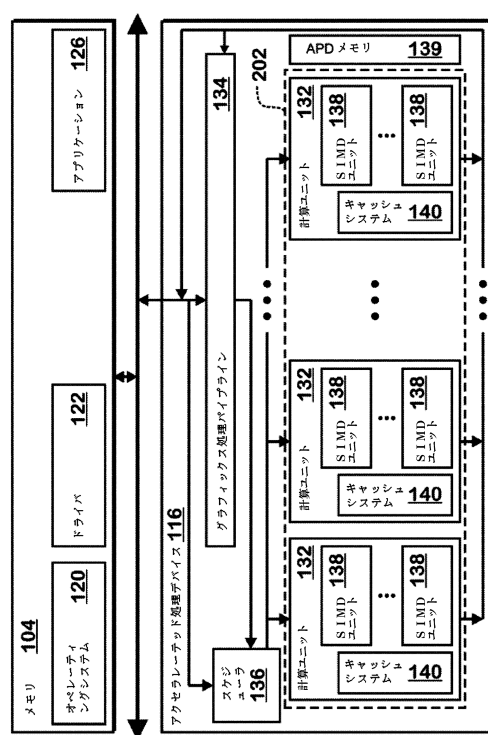
50

【図面】

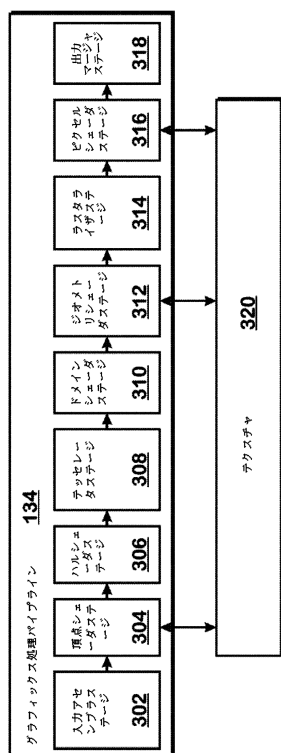
【 図 1 】



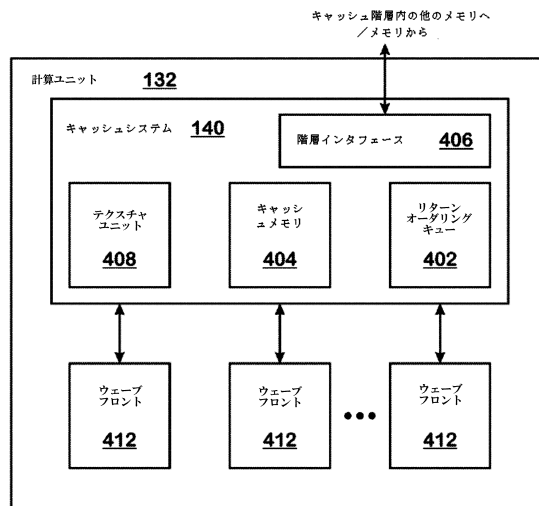
【圖 2】



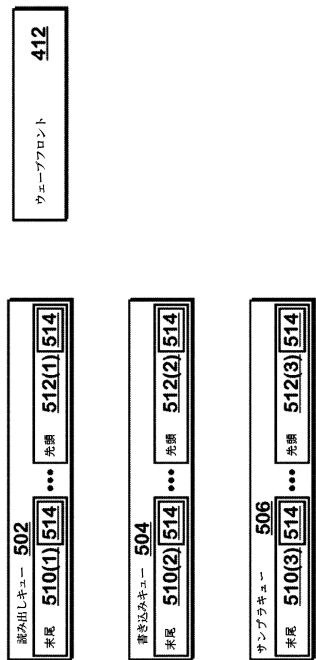
【 図 3 】



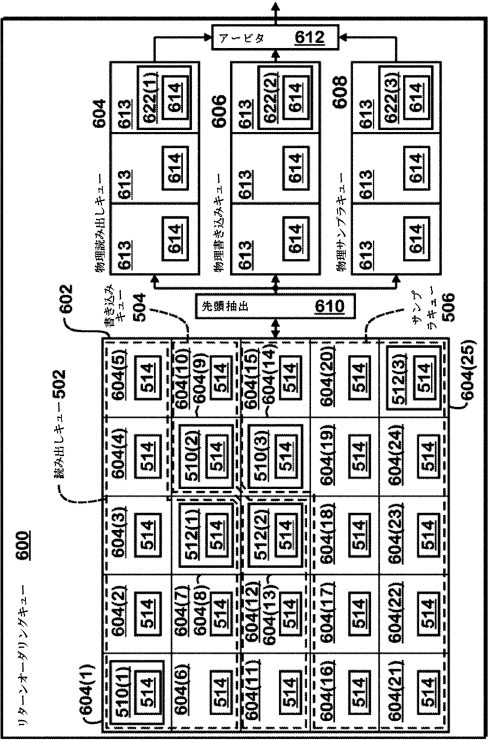
【圖 4】



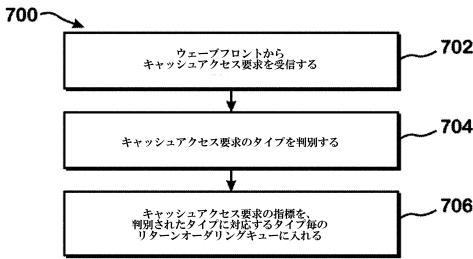
【図 5】



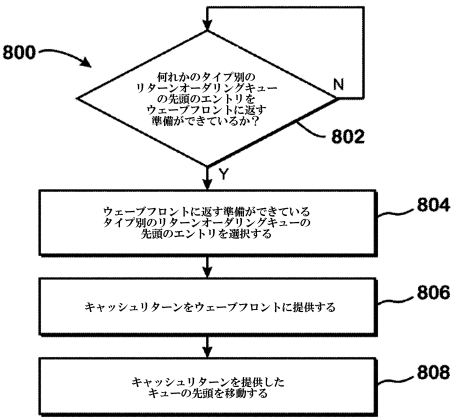
【図 6】



【図 7】



【図 8】



10

20

30

40

50

フロントページの続き

アメリカ合衆国 9 5 0 5 4 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 2 4
8 5、アドバンスト・マイクロ・デバイス・インコーポレイテッド内

(72)発明者 ファタネー ゴッドラット

アメリカ合衆国 0 1 7 1 9 マサチューセッツ州、ボックスボロー、セントラル ストリート 9
0、1、2 & 3 フロア、アドバンスト・マイクロ・デバイス・インコーポレイテッド内

審査官 北村 学

(56)参考文献

特開 2 0 0 4 - 1 4 5 8 3 8 (J P , A)

特開 2 0 0 7 - 3 1 6 8 5 9 (J P , A)

特表 2 0 0 2 - 5 3 9 5 0 9 (J P , A)

米国特許出願公開第 2 0 1 3 / 0 2 6 2 7 7 6 (U S , A 1)

米国特許出願公開第 2 0 1 4 / 0 2 2 3 1 1 1 (U S , A 1)

米国特許第 0 6 2 3 7 0 6 7 (U S , B 1)

(58)調査した分野 (Int.Cl. , D B 名)

G 0 6 F 1 2 / 0 8 4 4

G 0 6 F 9 / 3 4