(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7547526号
(P7547526)

(45)発行日　令和6年9月9日(2024.9.9)

(24)登録日　令和6年8月30日(2024.8.30)

(51)国際特許分類		FI		
G06F 12/0875(2016.01)		G06F	12/0875	106
G06F 9/38 (2018.01)		G06F	9/38	370X
G06F 12/0806(2016.01)		G06F	12/0806	100
G06F 13/38 (2006.01)		G06F	13/38	340C

請求項の数　6　外国語出願　（全10頁）

(21)出願番号　特願2023-33382(P2023-33382)
(22)出願日　令和5年3月6日(2023.3.6)
(62)分割の表示　特願2020-512508(P2020-512508)の分割
原出願日　平成30年8月30日(2018.8.30)
(65)公開番号　特開2023-78204(P2023-78204A)
(43)公開日　令和5年6月6日(2023.6.6)
審査請求日　令和5年3月15日(2023.3.15)
(31)優先権主張番号　62/552,475
(32)優先日　平成29年8月31日(2017.8.31)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　516291125
レール ビジョン リミテッド
ＲＡＩＬ ＶＩＳＩＯＮ ＬＴＤ
イスラエル国 ４３６６５１７ ラアナナ
ピー．オー．ボックス ２１５５ レホヴ
ハティダール １５
(74)代理人　100120891
弁理士　林 一好
(74)代理人　100165157
弁理士　芝 哲央
(74)代理人　100205659
弁理士　齋藤 拓也
(74)代理人　100126000
弁理士　岩池 満
(74)代理人　100185269

最終頁に続く

(54)【発明の名称】　複数計算における高スループットのためのシステムおよび方法

(57)【特許請求の範囲】

【請求項１】

データのストリームを受信するように構成され、前記データのストリームを、データ部分に分割し、前記データ部分の各々を、複数の先入れ先出し（ＦＩＦＯ）レジスタの１つを通過させるように構成された、複数の前記ＦＩＦＯレジスタを備えたアレイと、

前記データ部分を受信するように構成された、第１のアドバンスト・イクステンシブル・インタフェース（ＡＸＩ）ユニットと、

を備えたデータ・ストリーム・デバイダ・ユニット（ＤＳＤＵ）（３０４）と、

前記第１のＡＸＩユニットからデータ部分を受信するように構成された、第２のアドバンスト・イクステンシブル・インタフェース（ＡＸＩ）ユニットと、

それぞれのＦＩＦＯレジスタから各データ部分を受信するように構成され、前記受信したデータ部分を処理するように構成された、複数のストリーミングマルチプロセッサ（ＳＭ）と、

を備えたグラフィックプロセッシングユニット（ＧＰＵ）と、

を備えた、ＧＰＵでの処理がまだ施されていない未処理データを処理する回路。

【請求項２】

前記ＤＳＤＵ内のＦＩＦＯレジスタは、前記ＤＳＤＵ内の、前記ＦＩＦＯレジスタに割り当てられた第１のＡＸＩユニットと、前記ＧＰＵ内の、前記ＦＩＦＯレジスタに割り当てられた第２のＡＸＩユニットを介して前記ＧＰＵ内の、前記ＦＩＦＯレジスタに割り当てられたＳＭに接続される、請求項１に記載の回路。

【請求項３】
前記ＤＳＤＵ内の、前記ＦＩＦＯレジスタの各々は、前記ＤＳＤＵ内の、第１の共通ＡＸＩユニット、および前記ＧＰＵ内の、共通ＡＸＩユニットを介して、前記ＧＰＵ内の、前記ＦＩＦＯレジスタに割り当てられたＳＭに接続される、請求項１に記載の回路。
【請求項４】
グラフィックプロセッサユニット（ＧＰＵ）での処理がまだ施されていない未処理データのストリームを受信するステップと、
前記ストリームを複数のデータ部分に分割するステップと、
データストリームデバイダユニット（ＤＳＤＵ）の特定のＦＩＦＯレジスタを介して各データ部分を通過させるステップと、
前記特定のＦＩＦＯレジスタからのデータ部分を、処理のためにグラフィックプロセッサユニット（ＧＰＵ）内の、前記特定のＦＩＦＯレジスタに割り当てられたストリーミングマルチプロセッサ（ＳＭ）に転送するステップと、
を備え、
前記データ部分は、前記ＤＳＤＵ内の、第１のアドバンスト・イクステンシブル・インタフェース（ＡＸＩ）ユニットと、前記ＧＰＵ内の、第２のアドバンスト・イクステンシブル・インタフェース（ＡＸＩ）ユニットを介して転送される、
大容量のデータを効率的に処理する方法。
【請求項５】
特定のＦＩＦＯレジスタから受信したデータ部分は、前記ＤＳＤＵ内の、前記ＦＩＦＯレジスタに割り当てられた第１のＡＸＩユニットと、前記ＧＰＵ内の、前記ＦＩＦＯレジスタに割り当てられた第２のＡＸＩユニットを介して前記ＧＰＵ内の、前記特定のＦＩＦＯレジスタに割り当てられたＳＭに転送される、請求項４に記載の方法。
【請求項６】
ＦＩＦＯレジスタから受信した前記データ部分の各々は、前記ＤＳＤＵ内の、共通の第１のＡＸＩユニットと、前記ＧＰＵ内の、共通の第２のＡＸＩユニットを介して、前記ＧＰＵ内の、前記特定のＦＩＦＯレジスタに割り当てられたＳＭに転送される、請求項４に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
膨大の量のグラフィックデータおよび計算を操作／処理するのに適したコンピューティングシステムは、典型的に、中央処理装置（ＣＰＵ）は別として、必要なデータを操作し、処理するために適合および指定された、ＧＰＵ、ＧＰＧＰＵ、ＤＳＰ、ＳＩＭＤベースプロセッシングユニット、ＶＬＩＷベース処理ユニットを備える。そのような計算システムの構造は、この分野においてよく知られている。この構造は、典型的にＣＰＵとＭＰＵとの間で計算タスクを分割するので、重たい計算は、ＭＰＵに割り当てられ、残りの計算タスクは、ＣＰＵが操作するように残される。
【０００２】
しかしながら、よく知られた構造は、大量の図形データが含まれる場合、ＣＰＵとＭＰＵとの間で双方向に（back-and-forth）図形データの転送を管理するのに必要な大量のハンドリングリソースにより、効率が低下する。いくつかの場合、ＣＰＵ－ＭＰＵ計算構造におけるデータ計算に使用可能なネット時間は、５％未満の可能性がある。たとえば、Ｎｖｉｄｉａ（登録商標）コンピューティング・ユニファイド・デバイス・アーキテクチャ（Computing Unified Device Architecture）（ＣＵＤＡ）並列コンピューティングプラットフォームと、アプリケーションプログラミングインターフェイスモデルの場合、グラフィックデータの処理に費やされる一般的な時間部分は、ＣＰＵ環境からＧＰＵ環境（ＣＵＤＡメモリなど）にグラフィックデータを転送する場合４９％、グラフィックデータをＧＰＵ環境からＣＰＵ環境（ＣＵＤＡメモリ）に返送する場合、４７％であり、グラフィック計算の場合には、４％未満である。このような非常に低いグラフィック計算効率

は、グラフィックデータが、プロセッサ間で転送される方法を定義する共通の構造から来る。

【０００３】

グラフィック計算に割り当てられた時間を、実質的に上昇させる、ＭＰＵ効率の実質的上昇を可能にする必要性がある。

【発明の概要】

【０００４】

ストリーミングマルチプロセッサを介したグラフィックデータソースと、グラフィック処理ユニット（ＧＰＵ）との間で交換されるグラフィックデータスループットを高める方法が開示される。ＧＰＵは、プロセッシングコアユニット（ＰＣＵ）、レジスタファイルユニット、複数のキャッシュユニット、シェアドメモリユニット、統合キャッシュユニットおよびインタフェースキャッシュユニットを備えることができる。この方法は、インタフェースキャッシュユニットを介して、および複数のキャッシュユニットを介して、および統合されたキャッシュユニットを介して、グラフィックデータのストリームを、レジスタファイルユニットに転送するステップと、レジスタファイルユニットからのグラフィックデータの第２のストリームを、プロセッシングコアユニットに転送するステップと、レジスタファイルユニットを介してデータの頻繁に使用される部分をシェアドメモリに記憶し、受信するステップとを備えることができる。

【０００５】

いくつかの実施形態において、レジスタファイルユニットは、ＰＣＵにより処理されたデータを、そのデータを使用する頻度のレベルに応じて、さらにデータを受信することができる限り、シェアドメモリに仕向けるように構成される。いくつかの実施形態において、使用する頻度のレベルは、ＰＣＵにより決定される。

【０００６】

グラフィックデータを処理するように構成されるプロセッシングコアユニット（ＰＣＵ）と、ＰＣＵからグラフィックデータを供給し、ＰＣＵからの処理したグラフィックデータを受信して一時的に記憶するように構成された、レジスタファイルユニットと、レジスタファイルユニットからグラフィックデータを供給し、前記レジスタファイルユニットからの処理されたグラフィカルデータを受信し、一時的に記憶するように構成された、複数のキャッシュユニットと、レジスタファイルユニットからグラフィックデータを供給し、レジスタファイルユニットから処理されたグラフィックデータを受信し一時的に記憶するように構成されたシェアドメモリユニットと、レジスタファイルユニットからグラフィックデータを供給し、レジスタファイルユニットからの処理されたグラフィックデータを受信し一時的に記憶するように構成された、統合キャッシュユニットと、そして、高速でグラフィック処理のためのグラフィックデータを受信し、共有メモリユニットと統合キャッシュユニットの少なくとも一方に、グラフィックデータを提供し、統合キャッシュユニットから処理済みグラフィックデータを受信し、 処理されたグラフィックデータを外部処理ユニットに供給する。

【０００７】

いくつかの実施形態では、グラフィックデータ要素の少なくともいくつかは、ＰＣＵによる近い呼び出し（close call）の確率に関連付けられた優先度に基づいて、共有メモリユニット内のＰＣＵによる処理の前および／または後に格納される。いくつかの実施形態において、優先度は確率が高くなるにつれ高くなる。

【０００８】

データストリーム分割ユニット（ＤＳＤＵ）およびグラフィクス処理ユニット（ＧＰＵ）を備えた、未処理データを処理する回路が開示される。ＤＳＤＵは、データのストリームを受信し、データの一部に分割し、データの一部の各々を、複数の先入れ先出し（ＦＩＦＯ）レジスタの１つを介して通過させるように構成された、複数のＦＩＦＯレジスタと、前記データ部分を受信するように構成された、アドバンス・トイクステンシブル・インタフェース（Advanced Extensible Interface）（ＡＸＩ）とを備える。ＧＰＵは、第

１のＡＸＩユニットからデータ部分を受信するように構成された、第２のアドバンスト・イクステンシブル・インタフェース（ＡＸＩ）ユニットと、それぞれのＦＩＦＯレジスタから各データ部分を受信するように構成され、受信したデータ部分を処理する複数のストリーミングマルチプロセッサ（ＳＭ）を備える。

【０００９】

いくつかの実施形態において、ＤＳＤＵ内の特定のＦＩＦＯレジスタは、ＤＳＤＵ内に割り当てられた第１のＡＸＩユニットを介して、およびＧＰＵ内に割り当てられた第２のＡＸＩユニットを介して、ＧＰＵ内に割り当てられたＳＭに接続される。いくつかの実施形態において、ＤＳＤＵ内のＦＩＦＯレジスタの各々は、ＤＳＤＵ内の第１のＡＸＩユニット、およびＧＰＵ内の共通ＡＸＩユニットを介してＧＰＵ内に割り当てられたＳＭに接続される。

【００１０】

未処理データのストリームを受信するステップと、前記ストリームを複数のデータ部分に分割するステップと、各データ部分をデータストリームデバイダユニット（ＤＳＤＵ）内の特定のＦＩＦＯレジスタを介して通過させるステップと、前記特定のＦＩＦＯレジスタからの前記データ部分を、処理のために、グラフィクスプロセッサユニット（ＧＰＵ）内に割り当てられたストリーミングマルチプロセッサ（ＳＭ）に転送するステップと、を備えた、大量のデータを効率よく処理するための方法が開示される。いくつかの実施形態において、データ部分は、ＤＳＤＵ内の第１の特定のアドバンストイクステンシブルインタフェース（ＡＸＩ）と、ＧＰＵ内の第２の特定のアドバンストイクステンシブルインタフェース（ＡＸＩ）を介して転送される。

【００１１】

いくつかの実施形態において、特定のＦＩＦＯレジスタから受信したデータ部分は、ＤＳＤＵ内に割当てられた第１のＡＸＩユニットと、ＧＰＵ内に割り当てられた第２のＡＸＩユニットに転送される。いくつかの実施形態において、ＤＳＤＵ内のＦＩＦＯレジスタから受信したデータ部分の各々は、ＤＳＤＵ内の共通の第１のＡＸＩユニット、およびＧＰＵ内の共通の第２のＡＸＩユニットを介して、ＧＰＵに割り当てられたＳＭに転送される。この発明としてみなされる主題は、明細書の結論部分で特に指摘され、明確に特許請求の範囲に記載される。しかしながら、この発明は、動作の組織および方法の両方に関して、それらのオブジェクト、特徴および利点と一緒に、添付した図面と共に読むとき以下の詳細な記載によって最もよく理解することができる。

【図面の簡単な説明】

【００１２】

【図１】図１は、ＧＰＵを用いたコンピューティングユニットのデータフローを概略的に説明する。

【図２】図２は、ＧＰＵユニット内の典型的なストリーミングマルチプロセッサ（ＳＭ）の概略ブロック図である。

【図３Ａ】図３Ａは、この発明の実施形態に従って構成され、動作可能な未処理データ（ＵＰＤ）処理ユニット（ＵＰＤＨＵ）３００を描画する概略ブロック図である。

【図３Ｂ】図３Ｂは、図３ＡのＵＰＤＨＵ３００のような、ＵＰＤＨＵの一実施形態の概略ブロック図である。

【図３Ｃ】図３Ｃは、図３ＡのＵＰＤＨＵ３００のような、ＵＰＤＨＵの他の実施形態の概略ブロック図である。説明の簡便かつ明瞭さのために、図に示されるエレメントは必ずしも縮尺通りではない。例えば、エレメントのいつかの寸法は、明瞭さのためにエレメントに対して誇張されている可能性がある。さらに、適切に考慮する場合、参照符号は、図面間で対応するまたは類似するエレメントを示すために反復可能である。

【発明を実施するための形態】

【００１３】

以下の詳細な記載において、この発明の完全な理解を提供するために、多くの特定の詳細が述べられる。しかしながら、当業者には、この発明は、これらの特定の詳細なしに実

施可能であることが理解されるであろう。他のインスタンスにおいて、よく知られた方法、手続およびコンポーネントは、この発明を不明瞭にしないように詳細に記載していない。既知のコンピューティングシステムにおけるＣＰＵ－ＧＰＵ相互動作のボトルネックは、大部分が、ＣＰＵによってグラフィック関連データをＧＰＵに仕向け、および処理したグラフィックデータをＧＰＵから受信するのに使用される、データ転送チャネルにある。典型的に、ＣＰＵとＧＰＵプロセッサは、標準のコンピューティング環境で動作し通信する。

【００１４】

図１を参照すると、ＧＰＵを用いてコンピューティングユニット１００のデータフローを概略的に説明する。コンピューティングユニット１００は、ＣＰＵ１１１、ＣＰＵダイナミックＲＡＭ（ＤＲＡＭ）１１１Ａ、（メインボードチップセット）のようなコンピューティングユニット周辺制御ユニット１１２を備える。ユニット１００はさらに、ユニット１１２を介してＣＰＵとデータを通信するＧＰＵユニット１５０を備える。ＧＰＵユニット１５０は、典型的に、ユニット１１２とＧＰＵプロセッサとの間でデータをインタフェースするＧＰＵ　ＤＲＡＭユニット１５４、ＧＰＵ処理ユニットのためのデータをキャッシュするように適合したＧＰＵキャッシュユニット１５６（例えば、Ｌ２キャッシュユニット）、およびＧＰＵ処理ユニット１５８（例えば、ストリーミングマルチプロセッサ／ＳＭ）を備える。

【００１５】

処理ユニット１００に入力され、ＧＰＵ１５０により処理されるように意図されるグラフィックデータのフローは、データフロー（ＤＦ）の矢印により記載される。第１のデータフローＤＦ１は、コンピューティングユニット１００へのデータのフローを描画し、ＤＰＵ１１１は、ＤＦ２のフローを、周辺制御ユニット（ＰＣＵ）１１２を介してＤＲＡＭ１１１Ａに仕向け、そこから戻って、ＤＦ３－ＰＣＵ１１２－ＤＦ４を介してＧＰＵ１５０に仕向ける。ＧＰＵ１５０において、データは、ＤＲＡＭユニット１５４を通り、キャッシュユニット１５６を通って、複数のストリーミングマルチプロセッサ（ＳＭｓ）ユニット１５８に到達し、そこで、グラフィック処理が行われる。出来るだけデータフローボトルネックを無くすことは、この発明に従う方法と構造のターゲットである。

【００１６】

図２を参照すると、ＧＰＵユニット内の典型的ストリーミングマルチプロセッサの概略ブロック図である。ＳＭ２００は、プロセッシングコアユニット２１０（コンピュート・ユニファイド・デバイス・アーキテクチャ（ＣＵＤＡ）コアと呼ばれるときもある）と、コア２０１とキャッシュユニット２３０（コンスタントキャッシュ）、２５０（ユニファイドキャッシュ）、およびシェアドメモリ２４０を備える。ＳＭ２００に入ってくるデータおよびそこから出ていくデータは、図１のＧＰＵキャッシュユニット２５６（例えば、キャッシュユニット１５６（Ｌ２））を用いて交換される。グラフィック処理が既知の方法で実行されると、ＧＰＵユニットは、処理されるデータ量全体が、グラフィック処理が開始される前にいくつかのＳＭ２００ユニットのメモリユニットにロードされるまで待つであろう。

【００１７】

データ転送時間を低減する１つの方法は、データ転送を最小に低減することである。例えば、コア２１０により計算された中間結果は、ＤＲＡＭに記憶する代わりにレジスタファイル２２０に記憶することができる。さらに、シェアドメモリ２４０は、通常行われるようにアウトバウンド（outbound）で循環させる代わりに、ＳＭ２００内で頻繁に使用されるデータを記憶するために使用することができる。いくつかの実施形態において、使用頻度のレベルは、ＰＣＵにより決定される。さらに、コンスタントメモリユニットおよび／またはキャッシュメモリユニットは、ＳＭ２１０で定義することができる。この発明のさらなる実施形態によれば、ＣＰＵコンピューティング環境とＧＰＵコンピューティングとの間のデータフローボトルネックは、ＣＰＵを、グラフィック関連データをすべて処理するように特に構成されたコンピューティングユニットと置き換えることにより、低減

することができるか、または消去することができる。
【００１８】
図３Ａを参照すると、この発明の実施形態に従って、構成され、動作可能な未処理データ（ＵＰＤ）ハンドリングユニット（ＵＰＤＨＵ）３００を描画する概略ブロック図であり、図３Ｂおよび図３Ｃは、図３ＡのＵＰＤＨＵ３００のようなＵＰＤＨＵの２つの異なる実施形態３５０および３８０の概略ブロック図である。ここで使用される「未処理データ」という用語は、処理しようとしているデータの大きなストリームに関連し、典型的には、大きな計算容量を必要とし、例えば、仮想的に「リアルタイム」で（すなわち、できるだけ小さい待ち時間で）処理する必要がある、グラフィックデータの高速ストリーム（例えば、４Ｋビデオカメラから受信された）である。図３Ａに描画されるＵＰＤＨＵ３００のアーキテクチャは、ＣＰＵ－ＧＰＵの既知のアーキテクチャに典型的な、データストリームの固有のボトルネックを克服するように設計され、データ獲得の入力ストリームは、最初に、ＣＰＵにより処理され、次に、一時的にＣＰＵに関連付けられたＣＰＵメモリ、および／またはＲＡＭに記憶され、次に、（例えば、周辺コンポーネント相互接続イクスプレス（ＰＣＩｅ）バスを介して）ＧＰＵに再び転送され、ＧＰＵの一部である複数のストリーミングプロセッサに送信される前に、ＧＰＵプロセッサにより再び処理される。
【００１９】
図３Ａに関して、ここで記載された例は、アドバンスト・イクステンシブル・インタフェース（ＡＸＩ）に従って、動作するようにプログラムされたフィールドプログラマブルゲートアレイ（ＦＰＧＡ）の使用を示しているが、当業者には、ここに記載した動作の方法は、それぞれのＧＰＵとインタフェースするように適合し、および高いスループットで、グラフィック関連データを大量に転送するのに適した、他のコンピューティングユニットを用いて具現化することができる。
【００２０】
この発明の実施形態によれば、データストリーム・デバイダ・ユニット（ＤＳＤＵ）３０４は、例えば、大量のストリーミングＵＰＤ、例えばカメラからのビデオストリームを受信し、それを複数のより小さなストリームに分散し、ＧＰＵのＳＭｓに転送するようにプログラムされた、ＦＰＧＡを用いて具現化することができる。ＦＰＧＡとＧＰＵは、さらに、ＧＰＵの複数のＳＭｓの、少なくとも１つのＳＭが完全にロードされるやいなや、ＧＰＵが、転送されたグラフィックデータの処理を開始するように動作するようにプログラムすることができる。ほとんどの場合、完全にロードされたＳＭｓは、フルデータファイルよりも小さいデータ量を保持し、それゆえ、ＧＰＵによる処理は、この実施形態に従って開始され、データファイル全体がＧＰＵにロードされた後で、処理を開始する一般的に知られている実施形態に比べて、はるかに速い。
【００２１】
一例示実施形態において、ＵＰＤＨＵ３００は、複数の先入れ先出し（ＦＩＦＯ）レジスタ／ストレージユニットアレイ３０４Ａ（ＦＩＦＯユニットは、個別に示されていない）を備えたＤＳＤＵ３０４を備え、そのうちの１つのＦＩＦＯユニットを、ＧＰＵ３２０のＳＭｓ３１８の各々に割当てることができる。いくつかの実施形態において、ＤＳＤＵ３０４により受信されるＵＰＤストリームは、複数のデータユニットに分割され、ＦＩＦＯユニット３０４Ａを介して、ＧＰＵ３２０に転送することができ、ＡＸＩインタフェースのようなインタフェースユニットを介して、ＧＰＵにブロードキャストすることができ、それにより、各ＦＩＦＯ３０４Ａのデータユニットが、関連するＳＭ３１８に転送され、それにより、例えば、シングルアクション・マルチデータ（ＳＥＶＩＤ）コンピューティングを可能にする。ＧＰＵ３２０の各（単一でも）ＳＭ３１８は、ＡＸＩインタフェースを介して、関連するＦＩＦＯ３０４Ａから受信した未処理データのそれぞれの部分がロードされるとき、ＧＰＵ３２０は、処理を開始することができ、ＵＰＤファイル全体がロードされるまで待つ必要はない。
【００２２】
ＭＳＵ３１０は、長いストリームのグラフィックデータを受信するように構成された、

未処理データインタフェースユニット３０２を備えることができる。インタフェースユニット３０２を介して受信した大量の未処理データは、より小さなサイズの複数の数のデータユニットに分割され、ＦＩＦＯユニット３０４Ａ内の割り当てられたＦＩＦＯユニットを介して転送され、次に、ＡＸＩチャネル３１５で、ＧＰＵ　ＡＸＩインタフェースを介して、ＧＰＵ３２０の割り当てられたＳＭ３１８に転送される。ＳＭｓ３１８のそれぞれのＳＭにより処理されたデータユニットは、次に、ＡＸＩ接続を介して、ＭＳＵに転送することができる。上記のように、上述した実施形態では、ＣＰＵ－ＧＰＵアーキテクチャにおいて、典型的に大きなオーバヘッドが節約される。

【００２３】

図３Ｂおよび３Ｃは、この発明の実施形態に従う、図３ＡのＭＳＵ３１０を具現化する２つのオプションのアーキテクチャの概略ブロック図を描画する。図３Ｂは、ＭＳＵ３５０がＦＩＵ３５６とＧＰＵ３５８を備えることを描画する。ＦＩＵ３５６は、複数のＦＩＦＯユニット（集合的に３５６Ａと名前が付けられている）－ＦＩＦＯ０、ＦＩＦＯｉ・・・ＦＩＦＯｎを備える。各ＦＩＦＯユニットは、割り当てられたＦＰＧＡ　ＡＸＩ（Ｆ－ＡＸＩ）ユニット－Ｆ－ＡＸＩ０、Ｆ－ＡＸＩｉ・・・Ｆ－ＡＸＩｎ（集合的に３５６Ｂと名前が付けられている）とアクティブ通信中であり得る。別個のＦ－ＡＸＩユニットの各々は、間接的に、割り当てられたＧＰＵ　ＡＸＩ（Ｇ－ＡＸＩ）ユニット－Ｇ－ＡＸＩ０、Ｇ－ＡＸＩｉ・・・Ｇ－ＡＸＩｎとダイレクトに接続し得る。Ｇ－ＡＸＩインタフェースユニットの各々は、割り当てられたＳＭ－ＳＭ０、ＳＭｉ・・・ＳＭｎとアクティブに接続することができ、データを供給することができる。

【００２４】

さらに他の実施形態によれば、図３Ｃに示すように、ＭＳＵ３８０は、ＦＩＵ３８６とＧＰＵ３８８を備える。ＦＩＵ３８６は、複数のＦＩＦＯユニット（集合的に３８６Ａの名前がつけられている）－ＦＩＦＯ０、ＦＩＦＯｉ・・・ＦＩＦＯｎを備えることができる。各ＦＩＦＯユニットは、複数のＦＩＦＯユニットから単一のＡＸＩストリームへのデータストリームの管理を制御するように構成することが出来るＦＰＧＡ　ＡＸＩ（Ｆ－ＡＸＩ）ユニットとアクティブに通信することができる。ＡＸＩストリームは、ＧＰＵ３８８のＡＸＩインタフェースに送信して、次にそれぞれのＳＭｓユニット－ＳＭ０、ＳＭｉ、・・・ＳＭｎに分割することができる。図３Ｂに描画されるアーキテクチャは、より高速な全体性能を提供することができるが、（記載した回路を具現化する集積回路（ＩＣ）のために）より多くのピンと、より多くのワイヤ／導管（conduits）を必要とする場合がある。図３Ｃに描画される、アーキテクチャは、相対的により遅い全体性能を提供するが、（記載した回路を具現化する集積回路（ＩＣ）のために）より少ないピンと、より少ないワイヤ／導管（conduits）しか必要としない。

【００２５】

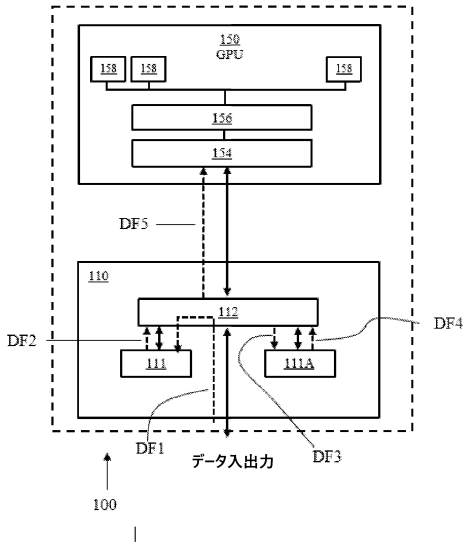
上述した、デバイス、構造、および方法は、既知のアーキテクチャおよび方法に比べて、大量の未処理データの処理を加速することができる。例えば、既知の実施形態において、処理／アルゴリズムがＨＰＵ上で開始できる前に、画像全体を転送する必要がある。画像サイズが１ＧＢの場合、ＧＰＵへデータを転送するＰＣＩ－Ｅバスの理論上のスループットは、３２ＧＢ／ｓであり、待ち時間は、１ＧＢ／（３２ＧＢ／ｓ）＝１／３２ｓ＝３１．１２５ｍｓ　３１．３ｍｓである。それに反して、この発明の実施形態に従うＦＰＧＡを用いると、すべてのＳＭユニットをフルにロードする必要があるだけである。たとえば、Ｔｅｓｌａ　Ｐ１００　ＧＰＵには、５６個のＳＭユニットがあり、各ＳＭには、３２ビット（単精度モード）をサポートする６４コアまたは６４ビット（拡張精度モード）をサポートする３２コアがあるため、完全にロードされたＧＰＵのデータサイズ（単精度モードまたは拡張精度モードで同じ結果）は、５６*３２*６４=１１４６８８ビット=１４．３３６Ｍ Ｍバイトである。ＦＰＧＡからＧＰＵ ＡＸＩストリームへの理論的スループットは、８９６ＭＢ／Ｓ（５６レーンの場合）であり、待ち時間は、１４．３３６ＭＢ（８９６ＭＢ／ｓ）＝１４．３３６／８９６ｓ＝１６ｍｓであり、これは実質半分の待ち時間である。

【００２６】

本明細書では、本発明の特定の特徴を例示および説明してきたが、多くの修正、置換、変更、および同等物が当業者に思い浮かぶであろう。したがって、添付の特許請求の範囲は、本発明の真の精神の範囲内に、あるそのようなすべての修正および変更を網羅することを意図していることを理解されたい。

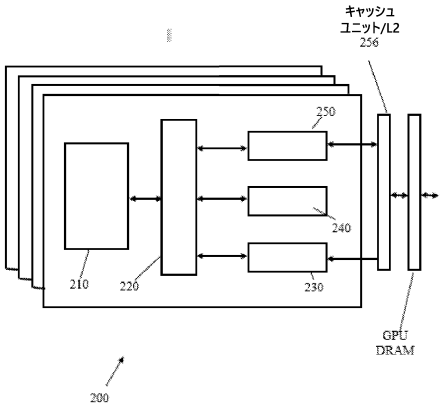
【図面】

【図１】

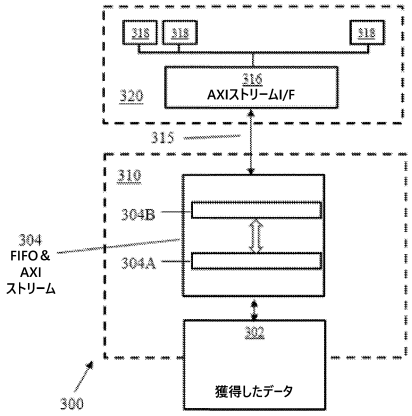


従来技術

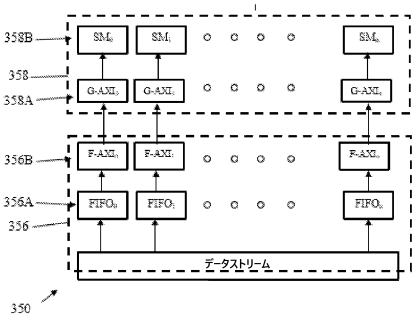
【図２】



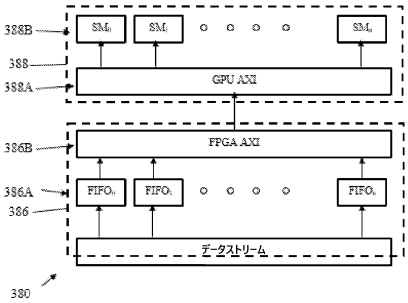
【図３Ａ】



【図３Ｂ】



【図３Ｃ】



フロントページの続き

弁理士　小菅 一弘
(72)発明者　ハニア シャハール
イスラエル国 ４４８５６００ ケドゥミム アリエル ストリート ７
(72)発明者　ゼルツァー ハナン
イスラエル国 ４５２６５１４ ホド ハシャロン ブネイ ブリット ストリート １２
審査官　北村 学
(56)参考文献　特表２００３－５０２９５８（ＪＰ，Ａ）
特開２０１２－１４６３２３（ＪＰ，Ａ）
特開平０６－３２４９９７（ＪＰ，Ａ）
特開２００６－０９１９８２（ＪＰ，Ａ）
特開２０１２－１５０７３５（ＪＰ，Ａ）
特開２０１５－１５６６４５（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
ＩＰＣ Ｇ０６Ｆ １２／０８ － １２／１２８
Ｇ０６Ｆ　９／３８
Ｇ０６Ｆ １３／３８ － １３／４２