

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7441763号
(P7441763)

(45)発行日 令和6年3月1日(2024.3.1)

(24)登録日 令和6年2月21日(2024.2.21)

(51)国際特許分類

F I

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

B 2 5 J 9/22 (2006.01)

B 2 5 J 9/22

A

請求項の数 26 外国語出願 (全167頁)

(21)出願番号	特願2020-152796(P2020-152796)	(73)特許権者	501450960
(22)出願日	令和2年9月11日(2020.9.11)		エヌビディア コーポレーション
(65)公開番号	特開2021-77345(P2021-77345A)		アメリカ合衆国 9 5 0 5 1 カリフォル
(43)公開日	令和3年5月20日(2021.5.20)		ニア、サンタクララ、サン トーマス エ
審査請求日	令和5年3月2日(2023.3.2)		クスプレスウェイ 2 7 8 8
(31)優先権主張番号	62/900,226	(74)代理人	110000855
(32)優先日	令和1年9月13日(2019.9.13)		弁理士法人浅村特許事務所
(33)優先権主張国・地域又は機関	米国(US)	(72)発明者	ユー ー ウェイ チャオ
(31)優先権主張番号	16/931,211		アメリカ合衆国、カリフォルニア、サン
(32)優先日	令和2年7月16日(2020.7.16)		タクララ、サン トーマス エクスプレス
(33)優先権主張国・地域又は機関	米国(US)	(72)発明者	ウェイ 2 7 8 8
			デ ー アン ファン
			アメリカ合衆国、カリフォルニア、サン
			タクララ、サン トーマス エクスプレス
			ウェイ 2 7 8 8
			最終頁に続く

(54)【発明の名称】 模倣学習システム

(57)【特許請求の範囲】

【請求項1】

ビデオ・データを少なくとも第1のセグメント及び第2のセグメントにセグメント化するステップであって、前記第1のセグメントが、実演において操作される第1の物体の第1の軌道を表すビデオ・データを含み、前記第2のセグメントが、前記実演において操作される第2の物体の第2の軌道を表すビデオ・データを含む、セグメント化するステップと、

前記第1の軌道によって満足される動作述語を識別するステップであって、前記第1の軌道で前記第1の物体を動かすことにより、前記第2の軌道で前記第2の物体を動かすことが有効になったという判定に少なくとも部分的に基づき、前記動作述語が識別される、識別するステップと、

ドメイン定義において定義された論理条件を満足する前記第2の軌道に少なくとも部分的に基づき、前記第2の軌道によって満足されるタスク述語を識別するステップと、

前記タスク述語に少なくとも部分的に基づき、前記実演のゴールを識別するステップとを含むコンピュータ実装方法。

【請求項2】

前記実演において操作される物体の別の軌道を前記第2の軌道が有効にしないという判定に少なくとも部分的に基づき前記ゴールを識別するために、前記タスク述語を使用するステップをさらに含む、請求項1に記載のコンピュータ実装方法。

【請求項3】

前記ゴールの1つ又は複数の述語を満足するように1つ又は複数の操作を実行することにより、1つ又は複数のロボット操作デバイスに前記ゴールを達成させるステップをさらに含む、請求項1に記載のコンピュータ実装方法。

【請求項4】

前記実演が実行された環境以外の環境において、前記1つ又は複数のロボット操作デバイスによって前記ゴールが達成される、請求項3に記載のコンピュータ実装方法。

【請求項5】

前記実演における物体の少なくとも1つの操作が、前記ロボット操作デバイスによる前記1つ又は複数の操作から除外される、請求項3に記載のコンピュータ実装方法。

【請求項6】

前記第1の軌道によって満足される付加的タスク述語を除くタスク述語のセットから、前記ゴールが識別される、請求項1に記載のコンピュータ実装方法。

【請求項7】

前記第1の軌道で前記第1の物体を動かすことにより、前記第2の軌道で前記第2の物体を動かすことが有効になったという判定に少なくとも部分的に基づき、前記付加的タスク述語が除外される、請求項6に記載のコンピュータ実装方法。

【請求項8】

前記ドメイン定義が、物体のリストと、前記物体のリスト内の物体の操作によって満足可能なタスク述語のリストとを含む、請求項1に記載のコンピュータ実装方法。

【請求項9】

システムであって、

少なくとも1つのプロセッサと、

前記少なくとも1つのプロセッサによって実行された結果、前記システムに少なくとも、ビデオ・データを第1のセグメント及び第2のセグメントにセグメント化することであって、前記第1のセグメントが、実演において操作される第1の物体の第1の軌道を表すビデオ・データを含み、前記第2のセグメントが、前記実演において操作される第2の物体の第2の軌道を表すビデオ・データを含む、セグメント化すること、

前記第1の軌道に従って前記第1の物体を操作することにより満足される動作述語を識別することであって、前記第1の軌道に従って前記第1の物体を操作することにより、前記第2の軌道に従って前記第2の物体を動かすことが有効になったという判定に少なくとも部分的に基づき、前記動作述語が識別される、識別すること、

前記第2の軌道に従って前記第2の物体を操作することにより満足されるタスク述語を識別すること、及び

前記タスク述語に少なくとも部分的に基づき、前記実演のゴールを識別することを行わせる命令を含むメモリとを含むシステム。

【請求項10】

前記第2の軌道が別の軌道を有効にしないという判定に少なくとも部分的に基づき、前記タスク述語が識別される、請求項9に記載のシステム。

【請求項11】

前記少なくとも1つのプロセッサによって実行された結果、少なくとも、1つ又は複数のロボット操作デバイスに前記ゴールを達成させることを、前記システムに行わせるさらなる命令を前記メモリが含む、請求項9に記載のシステム。

【請求項12】

前記実演が実行された環境以外の環境において、前記ロボット操作デバイスによって前記ゴールが達成される、請求項11に記載のシステム。

【請求項13】

前記タスク述語の意図性の推定に少なくとも部分的に基づき、前記実演の前記ゴールが識別される、請求項9に記載のシステム。

【請求項14】

10

20

30

40

50

意図性の前記推定が、第 1 の軌道と前記第 2 の軌道との関係に少なくとも部分的に基づき、請求項 13 に記載のシステム。

【請求項 15】

前記第 1 の軌道によって満足される第 1 のタスク述語は、前記第 1 の軌道によって前記第 2 の軌道が有効になることに少なくとも部分的に基づき、意図性が低いと評価される、請求項 9 に記載のシステム。

【請求項 16】

コンピュータ・システムの 1 つ又は複数のプロセッサによって実行された結果、前記コンピュータ・システムに、

ビデオ・データを第 1 のセグメント及び第 2 のセグメントに分割することであって、前記第 1 のセグメントが、実演において操作される第 1 の物体の第 1 の軌道を表すビデオ・データを含み、前記第 2 のセグメントが、第 2 の物体の第 2 の軌道を表すビデオ・データを含む、分割すること、

前記第 1 の軌道によって満足される動作述語を識別することであって、前記第 1 の物体の操作により、前記第 2 の軌道が有効になったという判定に少なくとも部分的に基づき、前記動作述語が識別される、識別すること、

前記第 2 の軌道によって満足されるタスク述語を識別すること、及び

前記タスク述語に少なくとも部分的に基づき、前記実演のゴールを識別することを行わせる命令を含む、非一時的コンピュータ可読媒体。

【請求項 17】

前記第 2 の軌道が別の軌道を有効にしないという判定に少なくとも部分的に基づき前記ゴールを識別するために、前記タスク述語が使用される、請求項 16 に記載の非一時的コンピュータ可読媒体。

【請求項 18】

前記コンピュータ・システムの前記 1 つ又は複数のプロセッサによって実行された結果、物体を操作して前記ゴールのタスク述語を達成するようにロボット操作デバイスに命令することによって前記ゴールを達成することを、前記コンピュータ・システムに行わせるさらなる命令を含む、請求項 16 に記載の非一時的コンピュータ可読媒体。

【請求項 19】

前記タスク述語の推定される意図性に少なくとも部分的に基づき、前記実演の前記ゴールが識別される、請求項 16 に記載の非一時的コンピュータ可読媒体。

【請求項 20】

前記推定される意図性が、前記第 2 の軌道に対する前記第 1 の軌道の分析に少なくとも部分的に基づき、請求項 19 に記載の非一時的コンピュータ可読媒体。

【請求項 21】

ロボット・デバイスであって、

1 つ又は複数のロボット操作デバイスと、

少なくとも 1 つのプロセッサと、

前記少なくとも 1 つのプロセッサによって実行された結果、前記ロボット・デバイスに少なくとも、

ゴールを達成するために、1 つ又は複数のロボット・アームの 1 つ又は複数の操作を実行するための命令を取得させる命令を含むメモリとを備え、前記ゴールが、

実演のビデオ・データを第 1 のセグメント及び第 2 のセグメントにセグメント化することであって、前記第 1 のセグメントが、第 1 の物体の第 1 の軌道を表すビデオ・データを含み、前記第 2 のセグメントが、第 2 の物体の第 2 の軌道を表すビデオ・データを含む、セグメント化すること、

前記第 1 の軌道によって満足される動作述語を識別することであって、前記第 1 の軌道により前記第 2 の軌道が有効になったという判定に少なくとも部分的に基づき、前記動作述語が識別される、識別すること、

前記第 2 の軌道に従って前記第 2 の物体を操作することにより満足されるタスク述語を識別すること、及び

前記タスク述語に少なくとも部分的に基づき、前記実演のゴールを識別することによって判定される、ロボット・デバイス

【請求項 2 2】

前記第 2 の軌道が前記タスク述語の 1 つ又は複数の条件への適合をもたらし、且つ別の軌道を有効にしないという判定に少なくとも部分的に基づき、前記タスク述語が識別される、請求項 2 1 に記載のロボット・デバイス。

【請求項 2 3】

前記ゴールの前記識別された述語の 1 つ又は複数の条件を満足する操作を識別することにより、前記 1 つ又は複数の操作が判定される、請求項 2 1 に記載のロボット・デバイス。

【請求項 2 4】

前記ゴールを達成するための前記 1 つ又は複数の操作によって満足される 1 つ又は複数の述語が、前記実演において満足される 1 つ又は複数の述語とは異なる、請求項 2 3 に記載のロボット・デバイス。

【請求項 2 5】

前記ゴールを判定するために使用されるタスク述語が、閾値量を超える関連性の推定に関連付けられたタスク述語を含む、請求項 2 1 に記載のロボット・デバイス。

【請求項 2 6】

関連性の前記推定が、2 つの軌道の関係に少なくとも部分的に基づく、請求項 2 5 に記載のロボット・デバイス。

【発明の詳細な説明】

【技術分野】

【0001】

本出願は、2019年9月13日出願の米国仮特許出願第62／900，226号、名称「MOTION REASONING FOR GOAL-BASED IMITATION LEARNING」の利益を主張し、その内容全体を参照により本明細書に組み込む。

【0002】

少なくとも一実施例は、実演を観察することに基づきゴールを達成するようにロボットを訓練することに関する。たとえば、少なくとも一実施例は、実演を観察し、本明細書に記載の様々な新規の技法を使用して、その実演のゴールを識別するロボット・デバイスに関する。

【背景技術】

【0003】

支援を提供するようにロボットを訓練することは、様々な課題を呈する。たとえば、ユーザは、ユーザによるタスクの実演に基づき、そのタスクを実行するようにロボットを教育したいと望むことがある。

【発明の概要】

【発明が解決しようとする課題】

【0004】

しかし、観察に基づきロボットを訓練するための技法は、かなりの量のコンピューティング・リソースを必要とすることがあり、又はエラーを生じやすい傾向がある。

【図面の簡単な説明】

【0005】

【図 1】 少なくとも一実施例による、実演学習の実例を示す図である。

【図 2】 少なくとも一実施例による、ビデオ・データのセグメント化の実例を示す図である。

【図 3】 少なくとも一実施例による述語識別の実例を示す図である。

【図 4】 少なくとも一実施例による、軌道分析の実例を示す図である。

10

20

30

40

50

- 【図 5】少なくとも一実施例による、実演学習の実例の態様を示す図である。
- 【図 6】少なくとも一実施例による、実演学習におけるゴール識別の実例を示す図である。
- 【図 7】少なくとも一実施例による、ロボット操作デバイスによるゴール再現の実例を示す図である。
- 【図 8 A】少なくとも一実施例による、推論及び／又は訓練論理を示す図である。
- 【図 8 B】少なくとも一実施例による、推論及び／又は訓練論理を示す図である。
- 【図 9】少なくとも一実施例による、ニューラル・ネットワークの訓練及び展開を示す図である。
- 【図 10】少なくとも一実施例による、例示的データ・センタ・システムを示す図である。
- 【図 11 A】少なくとも一実施例による、自律車両の実例を示す図である。
- 【図 11 B】少なくとも一実施例による、図 11 A の自律車両のカメラのロケーション及び視野の実例を示す図である。
- 【図 11 C】少なくとも一実施例による、図 11 A の自律車両の例示的システム・アーキテクチャを示すブロック図である。
- 【図 11 D】少なくとも一実施例による、クラウド・ベースのサーバと図 11 A の自律車両との通信のためのシステムを示す図である。
- 【図 12】少なくとも一実施例による、コンピュータ・システムを示すブロック図である。
- 【図 13】少なくとも一実施例による、コンピュータ・システムを示すブロック図である。
- 【図 14】少なくとも一実施例による、コンピュータ・システムを示す図である。
- 【図 15】少なくとも一実施例による、コンピュータ・システムを示す図である。
- 【図 16 A】少なくとも一実施例による、コンピュータ・システムを示す図である。
- 【図 16 B】少なくとも一実施例による、コンピュータ・システムを示す図である。
- 【図 16 C】少なくとも一実施例による、コンピュータ・システムを示す図である。
- 【図 16 D】少なくとも一実施例による、コンピュータ・システムを示す図である。
- 【図 16 E】少なくとも一実施例による、共有プログラミング・モデルを示す図である。
- 【図 16 F】少なくとも一実施例による、共有プログラミング・モデルを示す図である。
- 【図 17】少なくとも一実施例による、例示的な集積回路及び関連するグラフィックス・プロセッサを示す図である。
- 【図 18 A】少なくとも一実施例による、例示的な集積回路及び関連グラフィックス・プロセッサを示す図である。
- 【図 18 B】少なくとも一実施例による、例示的な集積回路及び関連グラフィックス・プロセッサを示す図である。
- 【図 19 A】少なくとも一実施例による、追加の例示的グラフィックス・プロセッサ論理を示す図である。
- 【図 19 B】少なくとも一実施例による、追加の例示的グラフィックス・プロセッサ論理を示す図である。
- 【図 20】少なくとも一実施例による、コンピュータ・システムを示す図である。
- 【図 21 A】少なくとも一実施例による、並列プロセッサを示す図である。
- 【図 21 B】少なくとも一実施例による、パーティション・ユニットを示す図である。
- 【図 21 C】少なくとも一実施例による、処理クラスタを示す図である。
- 【図 21 D】少なくとも一実施例による、グラフィックス・マルチプロセッサを示す図である。
- 【図 22】少なくとも一実施例による、マルチ・グラフィックス・プロセッシング・ユニット (GPU) システムを示す図である。
- 【図 23】少なくとも一実施例による、グラフィックス・プロセッサを示す図である。
- 【図 24】少なくとも一実施例による、プロセッサ用のプロセッサ・マイクロ・アーキテクチャを示すブロック図である。
- 【図 25】少なくとも一実施例による、深層学習アプリケーション・プロセッサを示す図である。
- 【図 26】少なくとも一実施例による、例示的ニューロモーフィック・プロセッサを示す

10

20

30

40

50

ブロック図である。

【図 27】 1つ又は複数の実施例による、グラフィックス・プロセッサの少なくとも一部分を示す図である。

【図 28】 1つ又は複数の実施例による、グラフィックス・プロセッサの少なくとも一部分を示す図である。

【図 29】 1つ又は複数の実施例による、グラフィックス・プロセッサの少なくとも一部分を示す図である。

【図 30】 少なくとも一実施例によるグラフィックス・プロセッサのグラフィックス処理エンジンのブロック図である。

【図 31】 少なくとも一実施例による、グラフィックス・プロセッサ・コアの少なくとも一部分のブロック図である。

10

【図 32 A】 少なくとも一実施例による、グラフィックス・プロセッサ・コアの処理要素のアレイを含むスレッド実行論理を示す図である。

【図 32 B】 少なくとも一実施例による、グラフィックス・プロセッサ・コアの処理要素のアレイを含むスレッド実行論理を示す図である。

【図 33】 少なくとも一実施例による、並列処理ユニット（「PPU」）を示す図である。

【図 34】 少なくとも一実施例による、汎用処理クラスタ（「GPC」）を示す図である。

【図 35】 少なくとも一実施例による、並列処理ユニット（「PPU」）のメモリ・パーティション・ユニットを示す図である。

【図 36】 少なくとも一実施例による、ストリーミング・マルチプロセッサを示す図である。

20

【図 37】 少なくとも一実施例による、先進コンピューティング・パイプラインのための実例データ・フロー図である。

【図 38】 少なくとも一実施例による、先進コンピューティング・パイプラインにおいて機械学習モデルを訓練、適応、インスタンス化、及び導入するための実例システムのシステム図である。

【図 39】 少なくとも一実施例による、撮像データを処理するための先進コンピューティング・パイプライン 3810A の実例を含む図である。

【図 40 A】 少なくとも一実施例による、超音波デバイスをサポートする仮想器具の実例データ・フローを含む図である。

30

【図 40 B】 少なくとも一実施例による、CT スキャナをサポートする仮想器具の実例データ・フローを含む図である。

【図 41 A】 少なくとも一実施例による、機械学習モデルを訓練するためのプロセスのデータ・フロー図である。

【図 41 B】 少なくとも一実施例による、事前訓練済みアノテーション・モデルを用いてアノテーション・ツールを拡張するためのクライアント・サーバ・アーキテクチャの実例を示す図である。

【発明を実施するための形態】

【0006】

少なくとも一実施例では、ロボット・デバイスは、新規タスクの実行をビデオ実演に基づき学習する。いくつかの事例では、このシステムは、偶発的な又は実演のゴールに必須でないステップが実演に含まれる場合でも、わずか 1 回の実演で新規タスクを学習することができる。たとえば、少なくとも一実施例では、ロボットは、人間のユーザによるタスクの実演に基づき、そのタスクを実行するように訓練される。教師あり訓練又はメタ学習を利用することにより、ロボットは実演をまねることを学習し、その後、ロボットの所有者から独立して、タスクを実行することができる。しかし、実演に含まれるステップ、行動、又は細部の 1 つ 1 つをロボットが厳密にまねることは必要ない場合があり、且つ望ましくない場合がある。むしろ、少なくとも一実施例では、ロボットは、実演を実行する際の実演者のゴール又は意図を推論し、その後、識別されたゴールを考慮して、実演されたタスクを実行する。ロボットは、実演者によって実行されたすべての行動を再構築するこ

40

50

なく、これを行ってもよい。たとえば、実演の意図が、カウンターにボールを置くことであった場合、ボールがキャビネットから取られても食器洗い機から取られてもかまわず、又はボールを持ち上げるのに実演者の右手が使われても左手が使われてもかまわないことがある。したがって、少なくとも一実施例では、ロボットは、人間の実演者の意図を推論して、ロボットが学習することをより幅広い状況に汎化する。

【0007】

少なくとも一実施例では、実演のゴール又は意図が、観察するロボットにとって曖昧であったり、不明であったりする。より全般的に、実演のゴール又は意図は、観察するシステムによって推測的に知られることはなく、システムは、実演のゴールを決して明示的に知らされない場合がある。1つの課題は、ゴール推理を実世界の实演に適用するとき、観察された状況の最終状態から、又は観察された一連の高レベルな行動からは、ゴール又は意図を判定できない場合があることである。たとえば調理の実演において、ロボットは、クラッカーの箱が動かされることと、ポットがコンロに置かれることとを観察する場合がある。ロボットは、実演のゴールが、クラッカーの箱を保管し、ポットをコンロに置くことであったと推論するかもしれない。しかし、クラッカーの箱を動かしたのは意図的でなく、又は単に箱をよけることを意図していたにすぎない場合がある。少なくとも一実施例では、実演のゴールは、物体軌道の分析に少なくとも部分的に基づいて推論される。

【0008】

図1は、少なくとも一実施例による、実演学習の実例を示す。実例100では、コンピューティング・デバイス102がカメラ108に接続され、又はカメラ108を含む。コンピューティング・デバイス102は、少なくとも1つのプロセッサと、この少なくとも1つのプロセッサによって実行されたことに応答して、本明細書に記載の実施例によるゴール・ベースの模倣学習のためのステップ又は動作をコンピューティング・デバイス102に実行させる命令を記憶しているメモリとを含む。カメラ108は、ビデオ・カメラ、赤外線カメラ、多スペクトル撮像カメラなどを含め、ビデオ・データを捕捉できる様々なカメラのうちの任意のものとすることができる。

【0009】

少なくとも一実施例では、カメラ108は、実演が実行される区域130のビデオ・データを捕捉する。少なくとも一実施例では、実演は人間の实演者110によって実行される。実例100では、実演は調理に関する。しかし、この実例は例示であることを意図しており、したがってこの実例は、提供される具体的な実例を組み込む又は必要とする実施例だけに本開示の範囲を限定するようなやり方で解釈されるべきではないことが理解される。

【0010】

実例100では、箱112、缶114、及びポット116が、作業表面124に置かれている。実演者110は、缶の中身をポット116に入れ、このポット116をコンロ120の火口に移動させることにより、缶114の中身を調理することを意図している。このゴールは、コンピューティング・デバイス102に明示的に通信されない。少なくとも一実施例では、コンピューティング・デバイス102は、カメラ108によって捕捉される観察のビデオ・データに基づき、ゴールを推論する。その後、ロボット・デバイス106は、潜在的に異なる区域132において、異なるコンロ122及び作業表面126を用いて、ゴールを再度達成することができる。

【0011】

少なくとも一実施例では、推論されたゴールは、ロボット・デバイス106によって実行される。ロボット・デバイス106は、コンピューティング・デバイス102を含んでもよく、又はロボット・デバイス106とコンピューティング・デバイス102は別々であってもよい。少なくとも一実施例では、ロボット・デバイス106は、少なくともプロセッサと、プロセッサによって実行されたときに、推論されたゴールを達成する行動をロボット・デバイス106に実行させる命令を含むメモリとを備える。少なくとも一実施例では、ロボット・デバイス106は、物体を操作するためのロボット・アーム又は他の機

構を含んでもよい。こうしたデバイスは、ロボット操作デバイスと呼ばれてもよい。少なくとも一実施例では、ロボット・デバイス106は、コンピューティング・デバイス102とカメラ108とを備える統合デバイスとして提供される。

【0012】

図1に示すシステムなど少なくとも一実施例では、実演で物体を動かすことが、2つの異なる述語クラス、すなわちタスク述語と動作述語とを潜在的に満足するものとして分析される。たとえば、クラッカーの箱を動かすことは、`in_storage(cracker_box)`などのタスク述語、及び`out_of_the_way(cracker_box)`などの動作述語を満足してもよい。述語`in_storage(cracker_box)`は、潜在的なゴールの達成に関連するタスクとすることができるので、タスク述語として分類されてもよい。その一方で、述語`out_of_the_way(cracker_box)`は、他の物体の動作又は軌道を有効にすることに関連するので、動作述語として分類されてもよい。したがって動作述語の軌道は、タスク述語を満足するその後の軌道を有効にすることができる。これは、たとえば図2において確認することができる。この図は、クラッカーの箱がどかされて、ポットをコンロまで動かせるようになる様子を示す。少なくとも一実施例では、コンピューティング・デバイス102は、ゴール・ベースの模倣学習を実施するとき、述語を、タスク述語、動作述語、又はこの両方として分類しようとする。

【0013】

少なくとも一実施例では、コンピューティング・デバイス102は、ゴール・ベースの模倣学習を実施するとき、実演者が実演のゴールを協力的且つ明確に通信することを意図しているという仮定の下に、動作する。したがって、実演は、意図を表現する又は判読可能なものとして説明されてもよい。この意図を見分けるために、コンピューティング・デバイス102は、ゴールの判別可能性が、高レベル動作ではなく、少なくとも部分的に低レベルの物体軌道にあると仮定する。この仮定により、タスク述語と動作述語の分類を、逆方向計画(`inverse planning`)の問題として策定できるようになる。

【0014】

図1に示してあるように、実世界での実演のビデオ・データは、カメラ108によって取得され、コンピューティング・デバイス102によって処理されてもよい。ビデオ実演を前提として、少なくとも一実施例では、コンピューティング・デバイス102は、実演の象徴的なゴールを出力することを目指す。たとえば、実演 $D = [z_1, \dots, z_T]$ を前提として、実施例は、実演の意図するゴール G を識別してもよい。ここで z_t は、実演 D のビデオ・フレームに対応する。少なくとも一実施例では、実演は、その後同じタスクを実行することになるロボットと同じタスク計画ドメインの下で、又はより一般的には同じ知識の下で分析される。たとえば、ゴールの識別と、識別されたゴールを達成するためのタスクの実行との両方に、計画ドメイン定義言語(「`PDDL`」)ファイルが使用されてもよい。

【0015】

少なくとも一実施例では、ドメインは、`in_storage(cracker_box)`、`on_stove(pot)`などのグラウンド・アトム又は述語のリストを含む。このドメインはまた、グラウンデッド演算子(`grounded operator`)のセットも含んでもよい。少なくとも一実施例では、グラウンデッド演算子は、名前、引数のリスト、演算子に適用するための前提条件のリスト、及び演算子を適用したことによる状態変化効果の識別を含む。前提条件は、ドメインの述語のうちの1つ又は複数に対応してもよい。グラウンデッド演算子は、高レベル動作又は動作と呼ばれることもある。

【0016】

少なくとも一実施例では、ゴール $G = [g_1, \dots, g_N]$ は、満たされたときにゴール G を達成する述語 g_1 のリストを含む。サブゴール g_i もまた、タスク述語として記述されることがある。サブゴール g_i は、ドメインにおける述語のサブセットから選択されてもよい。

10

20

30

40

50

【0017】

少なくとも一実施例では、ロボットは、観察された実演を模倣し、それを異なる環境に適用することができる。たとえば、環境 e_1 におけるビデオ実演

【数1】

$$D_{e_1}$$

を前提として、ゴール G を識別することにより、タスク計画の問題を新規の環境 $e_2 \neq e_1$ において解くことができるようになる。次いでロボットは、

【数2】

$$D_{e_1}$$

から導かれたゴール G に基づき、 e_2 において出力計画を実行することができる。少なくとも一実施例では、タスク・ドメインの定義が e_1 と e_2 との間で共有されるとき、 G は、エージェントの状態及び動作からも独立している。たとえば、少なくとも一実施例では、実演が抽象環境又は人工環境において実行され、ロボットは、対応するゴールを実環境において達成する。

【0018】

図2は、少なくとも一実施例によるビデオ・データのセグメント化の実例を示す。実例200では、実演 $D = [z_1, \dots, z_T]$ のビデオ・データ202が、時間的に分離されてビデオ・セグメント $\{S_i\}$ 204、206、208になる。少なくとも一実施例では、各セグメント i は、単一の物体を操作することにより、タスク述語

【数3】

$$\bar{g}_i$$

のうちの1つを達成すると仮定される。ビデオ実演のセグメント化は、タスク述語を達成する物体軌道を識別するためのビデオ・データの分析に基づいてもよい。少なくとも一実施例では、行動セグメント化又は他の関連する、同様の、若しくは対応する方法を使用して、ビデオ・データがビデオ・セグメントにセグメント化される。

【0019】

少なくとも一実施例では、各ビデオ・セグメント204、206、208は、実演者によって操作される物体のビデオを含む。実例200では、第1のビデオ・セグメント204は、箱212の操作を示し、第2のビデオ・セグメント206は、缶214の操作を示し、第3のビデオ・セグメント208は、ポット216の操作を示す。次いでこれらのパーティションが分析されて、これらがタスク述語又は動作述語を満足するかどうかが判定される。

【0020】

図3は、少なくとも一実施例によるタスク述語の識別の実例を示す。少なくとも一実施例では、ビデオ・セグメントの分析は、物体を操作することによりビデオ・セグメントがタスク述語 g_i を達成するという仮定に基づく。また分析は、ビデオ・セグメントが動作述語も達成するという仮定に基づき進行してもよい。次いで逆方向計画が適用されて、ビデオ・セグメントがタスク述語 g_i 、動作述語、又はそれらの両方を達成する意図を示すかどうかを判定することができる。すべてのビデオ・セグメントにおいて達成されたタスク述語 g_i の判定が完了すると、ドメイン定義に基づきタスク述語 g_i がともにプールされて、最終的なゴール $G = \{g_i\}$ を識別することができる。

【0021】

図3の実例300では、図2のビデオ・セグメント204に対応する第1のビデオ・セグメント304が、箱312の操作を示す。これは、`in_storage(box)` な

10

20

30

40

50

どのタスク述語を満足するものとみなされてもよい。しかし、箱の軌道の分析により、箱の操作は、その後のビデオ・セグメント208においてポット316の軌道を有効にしたことも示される。結果的に、箱312の操作は、動作述語も満足する。

【0022】

引き続き実例では、図2のビデオ・セグメント206に対応する第2のビデオ・セグメント306が、缶314の操作を示す。特に、実例300は、缶314がポット316に向かって動かされ、缶314の中身がポット316に移される様子を示す。いくつかの実施例では、これは、複数の別々の軌道として処理されてもよい。しかし、この実例では、これは1つの軌道として扱われる。いずれの事例でも、缶314の操作は、input(food)などのタスク述語を満足する。

10

【0023】

図2の第3のビデオ・セグメント208に対応する第3のビデオ・セグメント308は、ポット316をコンロに置くためのポット316の操作を示す。この軌道は、タスク述語on_stove(pot)を満足する。上述したように、この軌道は、箱312を事前に操作したことにより有効になった。

【0024】

図4は、少なくとも一実施例による、軌道分析の実例を示す。少なくとも一実施例では、動作述語の識別、又は前の軌道により有効にされたタスク述語の識別は、図4に関連して記載する軌道分析に基づく。実例400では、ビデオ・セグメント404は、物体b₁の操作を示し、ビデオ・セグメント408は、物体b₂の操作を示す。実例を理解しやすくするために、この実例では、物体b₁が、図2及び図3の箱212、312に対応すると推定されてもよく、物体b₂が、ポット216、316に対応すると推定されてもよい。

20

【0025】

少なくとも一実施例では、タスク述語に基づき実演のビデオ・セグメント404、408を時間的にセグメント化した後、セグメント

【数4】

$$S_i = [z_1^i, \dots, z_t^i, \dots, z_{T_i}^i]$$

はそのセグメントにおける観察

30

【数5】

$$z_t^i$$

と、そのセグメントにおいて達成されたタスク述語

【数6】

$$\bar{g}_i$$

とを含む。タスク述語の引数に基づき、そのセグメント内で物体b_iが操作されていることもわかる。次いで、そのセグメントS_iにおいて達成されたタスク述語

40

【数7】

$$\bar{g}_i$$

が、実演の実際のゴールの一部であるかどうか判定されてもよい。ゴール認識の態様は、異なる述語間の関係に基づき実行されてもよい。しかし、いくつかの事例では、

【数8】

$$\bar{g}_i$$

50

のみに基づいて実演の意図を判定するのは不可能な場合がある。

【0026】

述語レベルのこうした曖昧さを解決するために、物体の動作又は軌道が分析されてもよい。たとえば、何らかのタスク述語

【数9】

$\bar{g}_i$

を満足することに加えて、物体の操作は何か他のものも達成することがある。たとえば、物体 b_i はここで新規の位置及び姿勢にある。タスク述語とは無関係に、 b_i のこの新規の位置及び姿勢は、その後の物体軌道を有効にすることもできる。図4を参照すると、物体 b_1 の動作は、タスク述語

【数10】

$\bar{g}_1$

を満足し、且つ動作述語 m_1 も満足することができる。

【0027】

少なくとも一実施例では、高レベルの行動ベースの手法は、ゴール認識のタスク計画の態様を考慮し、2次元軌道に基づく分析は、ゴール認識の動作計画の態様を考慮する。少なくとも一実施例では、これらの態様は、タスク計画の態様と動作計画の態様の両方を使用して、実演のゴールを理解する。たとえば、ボールをコンロに移動させることは、象徴的には前提条件として $in_hand(bowl)$ しか含まないかもしれないが、暗黙的には、ボールをコンロに移動できるようにする有効な経路という動作制約も存在する。有効な経路を生成するように b_i を移動させることによってこの制約を満足することは、動作述語 $m_i(b_i)$ を満足することと行うことができる。

【0028】

いくつかの事例では、物体を移動させることは、タスク述語を達成するが、他の物体にとって有効な経路を生成しない。少なくとも一実施例では、こうした事例は、タスク述語だけを実演することを意図していると解釈される。他の事例では、物体を移動させることは、いかなるタスク述語も達成しないが、新規の有効な経路を有効にし、この場合、 b_i を移動させることは、動作述語 $m_i(b_i)$ を満足することを教育すると仮定することが可能である。

【0029】

他の事例では、 b_i の移動は、タスク述語

【数11】

$\bar{g}_i$

と動作述語 $m_i(b_i)$ の両方を満足する。

【数12】

$\bar{g}_i$

も $m_i(b_i)$ も真の場合には、実演のゴールに関して、動作の意図がそれほど明確ではなくなる。たとえば、箱312を移動することが、タスク述語 $in_storage(cracker_box)$ を達成し、且つポット316をコンロに移動させるための有効な経路を生成した場合、 $in_storage(cracker_box)$ がゴールの一部であるかどうかを判定するのが困難になることがある。

【0030】

少なくとも一実施例では、合理的行動の原理が適用されて、動作述語又はタスク述語が、観察された軌道によって最も効率的に引き起こされるかどうかが評価される。少なくとも一実施例では、これはベイズ的な逆方向計画に従うものである。たとえば、 ξ_i を、 S_i における b_i の軌道とする。少なくとも一実施例では、次いで、 S_i の意図を

【数 1 3】

$$\arg \max_{g \in \{g_i, m_i(b_i)\}} P(g|\xi_i) = P(\xi_i|g)P(g)$$

10

及び

【数 1 4】

$$P(\xi_i|g) \propto \frac{\exp(-C(\xi_i) - C(\xi_{q \rightarrow g}^*))}{\exp(-C(\xi_{s_i \rightarrow g}^*))}$$

によって判定することができる。

【0 0 3 1】

上記の式を参照すると、少なくとも一実施例では、 s 及び q は、 ξ_i の開始ロケーション及び終了ロケーションであり、

20

【数 1 5】

$$\xi_{s \rightarrow g}^*$$

及び

【数 1 6】

$$\xi_{q \rightarrow g}^*$$

30

は、 s 及び q から g を達成するための最適な軌道であり、 $C(\cdot)$ は、軌道のコストを算出するための関数である。少なくとも一実施例では、 g は空間内の点に向かう動きに対応する。少なくとも一実施例では、 g は、空間内の領域を通る動きに対応する。少なくとも一実施例では、 g は、空間内の点又は領域に向かう動きによって満足することができるタスク述語又は動作熟語である。後者の事例では、空間を直接離散化してサーチ・アルゴリズムを走らせるのは不十分な場合がある。少なくとも一実施例では、高速探索ランダム・ツリー（「RRT」）又は RRT* などの経路計画アルゴリズムを s から q まで走らせて、

【数 1 7】

$$\bar{g}_i$$

40

、 $m_j(b_i)$ を達成するための最適な軌道が近似される。少なくとも一実施例では、他の物体 b_j $i \neq j$ が障害として扱われ、構成空間が算出される。

【数 1 8】

$$\bar{g}_i$$

が空間内の領域のとき、RRT* などの経路計画アルゴリズムにとっての成功条件は、領域

【数 1 9】

50

$\bar{g}_i$

に到達することである。これに関して、 $m_i(b_i)$ は、 b_i が b_{i+1} の軌道を邪魔しないことを意味する。少なくとも一実施例では、 b_{i+1} の軌道をカバーする凸包が見いだされ、RRT*又は何らかの他の経路計画アルゴリズムが使用されて、 b_i がこの凸包に交差しない最短経路が見いだされる。

【0032】

これを図4の例に適用するとき、 ξ_1 は、物体 b_1 の軌道を参照してもよく、動作推理は、

【数20】

$$P(m_1|\xi_1) > P(\bar{g}_1|\xi_1)$$

かどうかに関する判定を含んでもよい。したがって実施例は、そのセグメントにおける b_1 の操作が、動作述語 m_1 を満足することを意図した可能性が高いか、タスク述語

【数21】

 $\bar{g}_1$

を満足することを意図した能性が高いかを、軌道 ξ_1 に基づき判定してもよい。

【0033】

少なくとも一実施例では、動作述語の推理を使用して、対応する

【数22】

 $\bar{g}_i$

が意図的かどうか、セグメント S_i ごとに判定される。少なくとも一実施例では、動作述語の推理は、図4に関して記載した技法により実行される。

【0034】

少なくとも一実施例では、意図的であると判定されたタスク述語

【数23】

 $\bar{g}_i$

がプールされる。少なくとも一実施例では、意図的なタスク述語は、後のタスク述語の前提条件である特定のタスク述語を除去することによって、プールされる。たとえば、ポットがコンロに移動されてもよいが、これは単に、ポット内にあるものを調理するための前提条件であるかもしれない。その後、ポットがテーブルに戻されて提供される。したがって、ポットがコンロにあることは、実演の実際のゴールではないことが推論されてもよい。

【0035】

少なくとも一実施例では、物体軌道レベルでの推理は、3次元空間における物体の姿勢に基づく。少なくとも一実施例では、PoseCNNなどの畳み込みニューラル・ネットワーク(「CNN」)を使用して、物体の姿勢が初期化される。次いで、少なくとも一実施例では、物体姿勢の追跡アルゴリズムに対する初期化として、検出出力が使用されてもよい。たとえば、少なくとも一実施例では、PoseRBPFなどの Rao-Blackwellized 粒子フィルタリング(「RBPF」)に基づくアルゴリズムを使用して、物体の姿勢が追跡される。少なくとも一実施例では、姿勢追跡は、位置及び配向に関する情報を含み、6D姿勢追跡と呼ばれることがある。少なくとも一実施例では、複数のカメラが使用され、物体のビューは、最大特定スコアに基づき選択される。ビデオ・フレーム観察 z_t が、物体中心表現

【数24】

$$x_t = \phi(z_{1:t}) = [\hat{x}_t^1, \dots, \hat{x}_t^k, \dots, \hat{x}_t^K]$$

に変換されてもよい。ここで、 ϕ は、物体姿勢追跡パイプラインであり、 x_t は、 z_t の物体中心表現であり、

【数 2 5】

$$\hat{x}_t^k$$

は、 k 番目の物体の推定 6 D 姿勢である。少なくとも一実施例では、 x_t は、検出された手の中心を用いて拡張されて、手と物体の相互作用に基づきビデオ・セグメント化の頑健性を向上させる。

【0 0 3 6】

本明細書に記載するように、実演のビデオ・データをセグメント化して、他のステップにおける物体軌道について推理を可能にすることができる。少なくとも一実施例では、各セグメントは、単一の物体の操作を含む。少なくとも一実施例では、行動のセグメント化により収集されたデータに基づきビデオをセグメント化するために、ニューラル・ネットワークが実行される。

【0 0 3 7】

少なくとも一実施例では、実演者のビデオ・データのセグメント化は、動作推理のために抽出された物体姿勢の軌道に基づき行われる。時間ステップごとに、各物体について既知の姿勢

【数 2 6】

$$\hat{x}_t^k$$

が存在する。時間的に

【数 2 7】

$$\hat{x}_t^k$$

を比較することにより、物体 k が動かされる、又は他のやり方で操作されることが判定されてもよい。次いでこの情報を使用して、別々の物体軌道に応じてビデオ・データがセグメント化されてもよい。少なくとも一実施例では、手の検出を使用して、実演者の手が物体と相互作用するセグメントを識別することにより、セグメント化はより正確になる。

少なくとも一実施例では、ビデオ・データをセグメント化した後、推定された姿勢に基づき各フレームの述語が算出される。セグメントの開始時と終了時の述語を比較することにより、セグメントのタスク述語が判定されてもよい。

【0 0 3 8】

図 5 は、少なくとも一実施例による、実演学習の実例の態様を示す。図 5 の実例 5 0 0 では、実環境 5 0 2 は、作業表面 5 0 4 とコンロ 5 0 6 を含むことができる。少なくとも一実施例では、タスク推理及び動作推理は、ストレージ領域 5 1 2 と、作業空間 5 1 4 と、2 つの火口領域 5 1 6、5 1 8 とを含む抽象環境 5 1 0 を考慮して実行される。少なくとも一実施例では、動作及び／又は述語の推理は、これらの抽象領域を考慮して実行される。

【0 0 3 9】

少なくとも一実施例では、ドメイン定義は、以下のものなど、様々な述語を定義する。

$\text{pour}(X, Y) : \text{in_hand}(X), \text{on}(X, Y) \Rightarrow \text{in}(X, Y)$

$\text{cook}(X) : \text{in}(X, b), \text{on_burner}(b) \Rightarrow \text{cooked}(X)$

【0 0 4 0】

またドメイン定義は、入力に制約を課してもよい。たとえば上の述語に関して、ドメイン定義は、Yが容器であることによりpourが制約を受けることを述べてもよい。

【0041】

少なくとも一実施例では、ドメイン定義は、上に提供したものなどの述語定義を使用し、実行可能なタスクを定義する。ドメイン定義は、スープ、ビーンズ、クラッカーの箱、マスタードのボトルといったものなど、様々な物体を定義してもよく、これらの使われ方についての制約を、様々な述語を用いて示してもよい。タスクを定義する実例として、材料を調理することは、上述したpour及びcookのタスク述語を必要とすることがある。これらのタスク述語のうちの1つは、物体を邪魔することによって妨げられることがある。述語を有効にするために、邪魔な物体を異なる領域に移動しなくてはならないが、邪魔な物体を移動させることは、それ自体、ゴールの一部ではない場合がある。本明細書に記載のタスク推理と動作推理の両方を使用すると、たとえばタスクに関係しない物体がゴールの実行を邪魔しない場合、それらの物体を動かさないことにより、又は実演において操作された特定の物体が存在しないときでも、ロボットがゴールを達成できるようにすることにより、ロボットは、より高い効率及び信頼性で調理タスクを再現することができるようになる。

10

【0042】

しかし、提供された実例は例示であることを意図しており、したがってこの実例は、提供される具体的な実例を組み込む実施例だけに本開示を限定するようなやり方で解釈されるべきではないことが理解される。たとえば、調理タスクに関して特定の技法を本明細書に記載するが、これらの技法は多様な用途に適用可能である。これらは、自動車操作、自己運転車両、医療用途、エンタテインメント、工場自動化などを含むが、これらに限定されない。

20

【0043】

図6は、少なくとも一実施例による、実演学習におけるゴール識別の実例を示す。実例のプロセス600は一連の動作として示されるが、1つの動作からの入力が別の動作の出力に応じて変わるときなど、順番が明示的に述べてある又は論理的に示唆される場合を除き、実施例では、示される動作が様々なやり方で変更されてもよく、一部の動作が省略されても、並び替えられても、又は他の動作と並行して実行されてもよいことが理解される。

【0044】

30

図6に示す動作は、図1に示すコンピューティング・デバイス102などのシステムによって実行されてもよく、このコンピューティング・デバイス102は、少なくとも1つのプロセッサと、この少なくとも1つのプロセッサによって実行されたことに応答して、示した動作をシステムに実行させる命令を含むメモリとを備える。

【0045】

602において、実例のシステムは、実演のビデオ・データを取得する。少なくとも一実施例では、図1に示すコンピューティング・デバイス102などのコンピューティング・デバイスは、実演が行われることになる物理的領域について訓練された1つ又は複数のカメラを介して、ビデオ・データを取得する。ビデオ・データは、MPEG、WMV、MOV、AVI、又はFLVを含んでもよいがこれらに限定されない様々なフォーマットのいずれかとすることができる。少なくとも一実施例では、ビデオ・データは、JPEG、GIF、又はPNGを含んでもよいがこれらに限定されないフォーマットで画像として提供される。

40

【0046】

少なくとも一実施例では、実演は実演者によって実行される一連の行動を含む。実演者の行動により、環境内で物体が操作される。たとえば、人間の实演者は、箱を動かす、缶を空ける、ノブを回す、ポットをコンロに置くといったことなどの一連の行動を実行してもよい。人間の实演者は、何らかのゴールをどのように達成できるかを示すつもりでこれらのタスクを実行する。しかし、ゴールがシステムによってアプリアリに知られることはないことに留意すべきである。むしろ、システムは本明細書に記載の技法に基づき、ゴー

50

ルを推論する。

【0047】

604において、システムはビデオ・データを物体軌道によってセグメント化する。少なくとも一実施例では、ビデオ・データのセグメント化は、ビデオの一部分の開始点及び停止点を識別することを含む。少なくとも一実施例では、セグメント化は、各セグメントが物体の1回の移動に対応するように行われる。たとえば、セグメントは、物体が点Aにおいて持ち上げられ点Bに動かされるときに物体の移動を示してもよい。いくつかの事例では、セグメントは、回転、開放、又は他のやり方で物体に作用するものなど、軌道としても記述できる他の行動を示してもよい。いくつかの事例では、セグメントは、1つのこうした行動を示すビデオ・データを含んでもよい。

10

【0048】

少なくとも一実施例では、各セグメントが分析されてその動きによってどの述語が満足されるかを判定できるように、セグメント化が行われる。少なくとも一実施例では、ドメイン定義は、述語を識別するためにシステムによって使用可能な情報を提供する。ドメイン定義は、コード又は構成ファイルを介して提供されてもよい。ドメインは、調理、運転、医療処置といったものなど、特定のタイプの活動に関連してもよい。少なくとも一実施例では、ドメイン定義は、タスク述語を記述するが、動作述語は記述しない。

【0049】

606において、システムは第1の物体の操作によって満足される動作述語を識別する。たとえば、システムは後のビデオ・セグメントの軌道を調べてもよく、これらの軌道が第1の物体の操作によって有効になったかどうか、又はどの程度有効になったかを判定してもよい。

20

【0050】

たとえば図4に関して本明細書に記載するように、実施例は、タスク述語及び動作述語を満足する軌道に対して物体軌道を分析してもよい。少なくとも一実施例では、システムは、ドメイン定義の情報からタスク述語を識別する。タスク述語は、物体の観察された軌道によって満足される条件を識別する。たとえば、物体が作業表面の片側に動かされる場合、物体の軌道は、`in_storage(object)`などの述語を満足してもよく、ここで「`storage`」区域は、作業表面の片側である。同様の述語は、物体をキャビネットに置くことによって満足されてもよい。

30

【0051】

第1の物体の軌道は、動作述語を満足することに関しても分析されてよい。たとえば図4に関して本明細書に記載するように、軌道は、その軌道が後の軌道を有効にすることに基づき、動作述語として分類されてもよい。少なくとも一実施例では、この分類は、後の軌道がその動作によって有効になる程度、又は後の軌道を有効にすることに関して第1の物体の軌道が意図的であると推定される程度を示すパーセンテージの可能性又は他の値を割り当てることを含む。

【0052】

608において、システムは、第2の物体の操作によって満足されたタスク述語を識別する。少なくとも一実施例では、第1の物体の軌道が第2の物体の軌道に対して分析されて、実演者の意図がタスク述語を満足することであったか動作述語を満足することであったかが判定される。システムは、軌道が後の軌道を有効にすることを意図していた可能性が高いと判定してもよく、この軌道を、動作述語を満足するものとして分類してもよい。同様に、システムは、軌道が後の軌道を有効にすることを意図していた可能性が低いと判定した場合、システムは第1の軌道を、タスク述語を満足するものとして分類してもよい。少なくとも一実施例では、これらの分類は、それぞれの分類にパーセンテージ又は確率の値を割り当てることによって行われる。

40

【0053】

610において、システムは、意図的である又はゴールを実演するものであると判定されたタスク述語に基づき、実演のゴールを識別する。少なくとも一実施例では、軌道分析

50

に基づきゴールを識別するために、タスク述語が使用される。たとえば、タスク述語を満足した軌道が、後の軌道を有効にしないとき、又は有効にすることを意図しているように見えないときに、タスク述語を使用してゴールが識別される。少なくとも一実施例では、動作述語の推理を使用して、システムはタスク述語が意図的かどうかを決定し、次いですべての意図的なタスク述語を最終的なゴールにプールする。少なくとも一実施例では、後の意図的なタスク述語の前提条件である意図的なタスク述語は除去され、残りのタスク述語が最終的なゴールを構成する。

【0054】

図7は、少なくとも一実施例による、ロボット操作デバイスによるゴール再現の実例を示す。実例のプロセス700は一連の動作として示されるが、1つの動作からの入力がある別の動作の出力に応じて変わるときなど、順番が明示的に述べてある又は論理的に示唆される場合を除き、実施例では、示される動作が様々なやり方で変更されてもよく、一部の動作が省略されても、並び替えられても、又は他の動作と並行して実行されてもよいことが理解される。

【0055】

図7によって示される動作は、図1に示すロボット・デバイス106などのシステムによって実行されてもよい。少なくとも一実施例では、ロボット・デバイスは、ロボット・アームなどのロボット操作デバイスと、少なくとも1つのプロセッサと、この少なくとも1つのプロセッサによって実行可能な命令であって、実行されたときに、図7に関して記載する動作をロボット・デバイスに実行させる命令を含むメモリとを備える。概して、推論されたゴールを取得したとき、ロボット・デバイスは、推論されたゴールを新規の状況又は環境において達成しようとするよう命令されることが可能である。したがって、ロボット・デバイスは、ユーザが前に実演したタスクを、ユーザが実行するのを支援できるようになってもよい。

【0056】

702において、少なくとも一実施例では、ロボット・デバイスは実演のゴールを取得する。少なくとも一実施例では、これは図6に関して記載した技法を使用してロボット・デバイスによって実行されてもよい。いくつかの実施例では、図1に示すコンピューティング・デバイス102などの別のデバイスが、実演を観察し、そのゴールを判定する。少なくとも一実施例では、これも、図6に関して記載した技法により行われてもよい。少なくとも一実施例では、次いで判定されたゴールが、ロボット・デバイスに送信される。

【0057】

少なくとも一実施例では、ゴールは1つ又は複数の述語を含む。少なくとも一実施例では、これらの述語は、実演中に後の述語の前提条件と識別されなかった述語である。少なくとも一実施例では、ゴールの述語は、ユーザによって実演された最後の状態を表す。少なくとも一実施例では、ゴールは、述語に関連付けられた意図性 (*intentionality*) に少なくとも部分的に基づきゴールに含められた1つ又は複数の述語を含む。

【0058】

704において、ロボット・デバイスは、環境のビデオ・データを分析する。これは、実演が行われた環境とは異なってもよい。たとえば、調理タスクは1つのキッチンで実演されてもよく、次いでその後、異なるキッチンで実行されてもよい。

【0059】

少なくとも一実施例では、ビデオ・データの分析は、物体及び物体姿勢を検出することを含む。これは、たとえばドメイン定義に記述された物体を識別すること、及びビデオ・データによって表されるシーン内のどこにこれらの物体が位置するかを識別することを含んでもよい。

【0060】

706において、ロボット・デバイスは、識別された物体を考慮して、ゴールを完了するのに必要な述語を識別する。少なくとも一実施例では、これは、ゴールを定義する1つ又は複数の述語から始まって、ゴールを達成するための中間ステップ及び初期ステップを

10

20

30

40

50

識別するように逆算することによって行われる。

【0061】

708において、ロボット・デバイスは、識別された述語を満足することになる軌道を識別する。たとえば、ロボット・デバイスは、述語 `in_pot (ingredient)` を満足するためには、ポットに入れられる材料が、その現在の場所からポットの場所に動かされなくてはならないことを判定してもよい。いくつかの事例では、軌道が実行されるために、他の物体が動かされる必要がある。ロボット・デバイスは、こうした状況を識別し、これらに応答することができる。これは、実演中に実演者が物体を動かす状況と同様である。しかし、こうした軌道は、ゴール識別中に、タスク述語ではなく動作述語と意図されていると判定されてもよく、ゴールを達成するためのロボット・デバイスの計画から除外されてもよい。たとえば、物体が動作述語を満足するように実演中に動かされたが、別の軌道を邪魔していない場合には、ロボット・デバイスはこの物体を動かさない。同様に、ロボット・デバイスがゴールを達成するときに物体が存在しない場合には、ロボット・デバイスの計画アルゴリズムは、物体がないことによって中断されない。

10

【0062】

710において、ロボット・デバイスは、ロボット・アームなどのそのロボット操作デバイスに、ゴールを達成するために環境内で物体を操作させる。少なくとも一実施例では、ロボット・デバイスは、その操作デバイスのうちの1つ又は複数に軌道を実現させる命令を策定する。たとえば、述語 `in_pot (ingredient)` を満足するために、ロボット・デバイスは、「アームを x_1 、 y_1 、 z_1 に移動させる」、「物体を把持する」、「 x_2 、 y_2 、 z_2 」にアームを移動させるといったものなどの命令を策定してもよい。

20

【0063】

推論及び訓練の論理

図8Aは、1つ又は複数の実施例に関して推論及び／又は訓練の動作を実行するために使用される推論及び／又は訓練論理815を示す。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて以下に提供される。

【0064】

少なくとも一実施例では、推論及び／又は訓練論理815は、1つ又は複数の実施例の態様において推論するように訓練及び／若しくは使用されるニューラル・ネットワークのニューロン若しくは層を構成するための順伝播及び／若しくは出力の重み、及び／若しくは入力／出力データ、及び／若しくは他のパラメータを記憶するためのコード並びに／又はデータ・ストレージ801を、限定することなく含んでもよい。少なくとも一実施例では、訓練論理815は、タイミング及び／又は順序を制御するためのグラフ・コード又は他のソフトウェアを記憶するためのコード及び／又はデータ・ストレージ801を含んでもよく、又はそれに結合されてもよく、コード及び／又はデータ・ストレージ801には、重み及び／又は他のパラメータ情報がロードされて、整数及び／又は浮動小数点ユニット（総称して算術論理演算ユニット（ALU））を含む論理が構成される。少なくとも一実施例では、グラフ・コードなどのコードは、こうしたコードが対応するニューラル・ネットワークのアーキテクチャに基づき、重み又は他のパラメータ情報をプロセッサALUにロードする。少なくとも一実施例では、コード及び／又はデータ・ストレージ801は、1つ又は複数の実施例の態様を使用した訓練及び／又は推論中に、入力／出力データ及び／又は重みパラメータを順伝播する間に1つ又は複数の実施例と併せて訓練又は使用されるニューラル・ネットワークの各層の重みパラメータ及び／又は入力／出力データを記憶する。少なくとも一実施例では、コード及び／又はデータ・ストレージ801の任意の部分は、プロセッサのL1、L2、又はL3のキャッシュ、若しくはシステム・メモリを含む他のオン・チップ又はオフ・チップのデータ・ストレージとともに含められてもよい。

30

40

【0065】

少なくとも一実施例では、コード及び／又はデータ・ストレージ801の任意の部分は、1つ若しくは複数のプロセッサ、又は他のハードウェア論理デバイス若しくは回路の内

50

部にあっても外部にあってもよい。少なくとも一実施例では、コード及び／又はコード及び／又はデータ・ストレージ801は、キャッシュ・メモリ、ダイナミック・ランダム・アドレス可能メモリ（「DRAM」：dynamic randomly addressable memory）、スタティック・ランダム・アドレス可能メモリ（「SRAM」：static randomly addressable memory）、不揮発性メモリ（たとえば、フラッシュ・メモリ）、又は他のストレージであってもよい。少なくとも一実施例では、コード及び／又はコード及び／又はデータ・ストレージ801が、たとえばプロセッサの内部にあるか外部にあるかの選択、又はDRAM、SRAM、フラッシュ、若しくは何らか他のタイプのストレージを含むかの選択は、オン・チップ対オフ・チップで利用可能なストレージ、実行される訓練及び／又は推論の機能のレイテンシ要件、ニューラル・ネットワークの推論及び／又は訓練で使用するデータのバッチ・サイズ、又はこれらの要因の何からの組合せに応じて決められてもよい。

【0066】

少なくとも一実施例では、推論及び／又は訓練論理815は、1つ又は複数の実施例の態様において推論するために訓練及び／若しくは使用されるニューラル・ネットワークのニューロン若しくは層に対応した、逆伝播及び／若しくは出力の重み、及び／若しくは入力／出力データを記憶するためのコード並びに／又はデータ・ストレージ805を、限定することなく含んでもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ805は、1つ又は複数の実施例の態様を使用した訓練及び／又は推論中に、入力／出力データ及び／又は重みパラメータを逆伝播する間に1つ又は複数の実施例と併せて訓練又は使用されるニューラル・ネットワークの各層の重みパラメータ及び／又は入力／出力データを記憶する。少なくとも一実施例では、訓練論理815は、タイミング及び／又は順序を制御するためのグラフ・コード又は他のソフトウェアを記憶するためのコード及び／又はデータ・ストレージ805を含んでもよく、又はそれに結合されてもよく、コード及び／又はデータ・ストレージ805には、重み及び／又は他のパラメータ情報がロードされて、整数及び／又は浮動小数点ユニット（総称して算術論理演算ユニット（ALU））を含む論理が構成される。

【0067】

少なくとも一実施例では、グラフ・コードなどのコードは、こうしたコードが対応するニューラル・ネットワークのアーキテクチャに基づき、重み又は他のパラメータ情報をプロセッサALUにロードさせる。少なくとも一実施例では、コード及び／又はデータ・ストレージ805の任意の部分は、プロセッサのL1、L2、又はL3のキャッシュ、若しくはシステム・メモリを含む他のオン・チップ又はオフ・チップのデータ・ストレージとともに含められてもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ805の任意の部分は、1つ又は複数のプロセッサ、又は他のハードウェア論理デバイス若しくは回路の内部にあっても外部にあってもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ805は、キャッシュ・メモリ、DRAM、SRAM、不揮発性メモリ（たとえば、フラッシュ・メモリ）、又は他のストレージであってもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ805が、たとえばプロセッサの内部にあるか外部にあるかの選択、又はDRAM、SRAM、フラッシュ・メモリ、若しくは何らか他のタイプのストレージを含むかの選択は、オン・チップ対オフ・チップで利用可能なストレージ、実行される訓練及び／又は推論の機能のレイテンシ要件、ニューラル・ネットワークの推論及び／又は訓練で使用するデータのバッチ・サイズ、又はこれらの要因の何からの組合せに応じて決められてもよい。

【0068】

少なくとも一実施例では、コード及び／又はデータ・ストレージ801と、コード及び／又はデータ・ストレージ805は、別々のストレージ構造であってもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ801と、コード及び／又はデータ・ストレージ805は、組み合わされたストレージ構造であってもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ801と、コード及び／又はデータ・ス

10

20

30

40

50

ストレージ 805 は、部分的に組み合わせられ、部分的に別々であってもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ 801 と、コード及び／又はデータ・ストレージ 805 との任意の部分は、プロセッサの L1、L2、又は L3 のキャッシュ、若しくはシステム・メモリを含む他のオン・チップ又はオフ・チップのデータ・ストレージとともに含められてもよい。

【0069】

少なくとも一実施例では、推論及び／又は訓練論理 815 は、訓練及び／又は推論コード（たとえばグラフ・コード）に少なくとも部分的に基づく、又はそれにより示される論理演算及び／又は算術演算を実行するための、整数及び／又は浮動小数点ユニットを含む 1 つ又は複数の算術論理演算ユニット（「ALU」）810 を、限定することなく含んでもよく、その結果が、アクティブ化ストレージ 820 に記憶されるアクティブ化（たとえば、ニューラル・ネットワーク内の層若しくはニューロンからの出力値）を生成してもよく、これらは、コード及び／若しくはデータ・ストレージ 801、並びに／又はコード及び／若しくはデータ・ストレージ 805 に記憶される入力／出力及び／又は重みパラメータのデータの関数である。少なくとも一実施例では、アクティブ化ストレージ 820 に記憶されるアクティブ化は、命令又は他のコードを実行したことに応答して、ALU 810 によって実行される線形代数計算及び又は行列ベースの計算に従って生成され、ここでコード及び／又はデータ・ストレージ 805 及び／又はデータ 801 に記憶された重み値は、バイアス値、勾配情報、運動量値などの他の値、又は他のパラメータ若しくはハイパーパラメータとともにオペランドとして使用され、これらのいずれか又はすべてが、コード及び／若しくはデータ・ストレージ 805、又はコード及び／若しくはデータ・ストレージ 801、又はオン・チップ若しくはオフ・チップの別のストレージに記憶されてもよい。

【0070】

少なくとも一実施例では、ALU 810 は、1 つ若しくは複数のプロセッサ、又は他のハードウェア論理デバイス若しくは回路内に含まれるが、別の実施例では、ALU 810 は、それらを使用するプロセッサ又は他のハードウェア論理デバイス若しくは回路の外部にあってもよい（たとえばコプロセッサ）。少なくとも一実施例では、ALU 810 は、プロセッサの実行ユニット内に含まれてもよく、又は同じプロセッサ内にあるか異なるタイプの異なるプロセッサ（たとえば、中央処理装置、グラフィックス・プロセッシング・ユニット、固定機能ユニットなど）の間で分散されているかのいずれかであるプロセッサの実行ユニットによりアクセス可能な ALU バンク内に、他のやり方で含まれてもよい。少なくとも一実施例では、コード及び／又はデータ・ストレージ 801、コード及び／又はデータ・ストレージ 805、並びにアクティブ化ストレージ 820 は、プロセッサ又は他のハードウェア論理デバイス若しくは回路を共有してもよく、別の実施例では、それらは異なるプロセッサ又は他のハードウェア論理デバイス若しくは回路にあってもよく、或いは同じプロセッサ又は他のハードウェア論理デバイス若しくは回路と、異なるプロセッサ又は他のハードウェア論理デバイス若しくは回路との何らかの組合せにあってもよい。少なくとも一実施例では、アクティブ化ストレージ 820 の任意の部分は、プロセッサの L1、L2、又は L3 のキャッシュ、若しくはシステム・メモリを含む他のオン・チップ又はオフ・チップのデータ・ストレージとともに含められてもよい。さらに、推論及び／又は訓練コードが、プロセッサ又は他のハードウェア論理若しくは回路にアクセス可能な他のコードとともに記憶されてもよく、プロセッサのフェッチ、デコード、スケジューリング、実行、リタイア、及び／又は他の論理回路を使用してフェッチ及び／又は処理されてもよい。

【0071】

少なくとも一実施例では、アクティブ化ストレージ 820 は、キャッシュ・メモリ、DRAM、SRAM、不揮発性メモリ（たとえば、フラッシュ・メモリ）、又は他のストレージであってもよい。少なくとも一実施例では、アクティブ化ストレージ 820 は、完全に又は部分的に、1 つ若しくは複数のプロセッサ又は他の論理回路の内部にあってもよく、又は外部にあってもよい。少なくとも一実施例では、アクティブ化ストレージ 820 が

、たとえばプロセッサの内部にあるか外部にあるかの選択、又はDRAM、SRAM、フラッシュ・メモリ、若しくは何らか他のタイプのストレージを含むかの選択は、オン・チップ対オフ・チップの利用可能なストレージ、実行される訓練及び／又は推論機能のレイテンシ要件、ニューラル・ネットワークの推論及び／又は訓練で使用するデータのバッチ・サイズ、又はこれらの要因の何からの組合せに応じて決められてもよい。

【0072】

少なくとも一実施例では、図8Aに示す推論及び／又は訓練論理815は、グーグルからのTensorFlow（登録商標）処理ユニット、Graphcore（商標）からの推論処理ユニット（IPU：inference processing unit）、又はIntel CorpからのNervana（登録商標）（たとえば「Lake Crest」）プロセッサなどの特定用途向け集積回路（「ASIC：application-specific integrated circuit」）と併せて使用されてもよい。少なくとも一実施例では、図8Aに示す推論及び／又は訓練論理815は、中央処理装置（「CPU」：central processing unit）ハードウェア、グラフィックス・プロセッシング・ユニット（「GPU」：graphics processing unit）ハードウェア、又はフィールド・プログラマブル・ゲート・アレイ（「FPGA」：field programmable gate array）など他のハードウェアと併せて使用されてもよい。

【0073】

図8Bは、少なくとも1つの実施例による、推論及び／又は訓練論理815を示す。少なくとも一実施例では、推論及び／又は訓練論理815は、ハードウェア論理を限定することなく含んでもよく、このハードウェア論理では、計算リソースが、ニューラル・ネットワーク内のニューロンの1つ若しくは複数の層に対応する重み値又は他の情報の専用のものであるか、又は他のやり方でそれらと併せてしか使用されない。少なくとも一実施例では、図8Bに示す推論及び／又は訓練論理815は、グーグルからのTensorFlow（登録商標）処理ユニット、Graphcore（商標）からの推論処理ユニット（IPU）、又はインテルコーポレーションからのNervana（登録商標）（たとえば「Lake Crest」）プロセッサなどの特定用途向け集積回路（ASIC）と併せて使用されてもよい。少なくとも一実施例では、図8Bに示す推論及び／又は訓練論理815は、中央処理装置（CPU）ハードウェア、グラフィックス・プロセッシング・ユニット（「GPU」）ハードウェア、又はフィールド・プログラマブル・ゲート・アレイ（FPGA）など他のハードウェアと併せて使用されてもよい。少なくとも一実施例では、推論及び／又は訓練論理815は、限定することなく、コード及び／又はデータ・ストレージ801、並びにコード及び／又はデータ・ストレージ805を含み、これらを使用して、コード（たとえばグラフ・コード）、重み値、並びに／又はバイアス値、勾配情報、運動量値、及び／若しくは他のパラメータ若しくはハイパーパラメータ情報を含む他の情報を記憶してもよい。図8Bに示す少なくとも一実施例では、コード及び／又はデータ・ストレージ801並びにコード及び／又はデータ・ストレージ805のそれぞれは、それぞれ計算ハードウェア802及び計算ハードウェア806などの専用計算リソースに関連付けられる。少なくとも一実施例では、計算ハードウェア802及び計算ハードウェア806のそれぞれは、線形代数関数などの数学的関数を、それぞれコード及び／又はデータ・ストレージ801並びにコード及び／又はデータ・ストレージ805に記憶された情報に対してのみ実行する1つ又は複数のALUを備え、その結果は、アクティブ化ストレージ820に記憶される。

【0074】

少なくとも一実施例では、コード並びに／又はデータ・ストレージ801及び805のそれぞれ、並びに対応する計算ハードウェア802及び806は、ニューラル・ネットワークの異なる層にそれぞれ対応し、それにより、コード及び／又はデータ・ストレージ801並びに計算ハードウェア802との1つのストレージ／計算の対801／802から結果的に生じるアクティブ化は、ニューラル・ネットワークの概念的組織化を反映させる

ために、次のコード及び／又はデータ・ストレージ 805 並びに計算ハードウェア 806 とのストレージ／計算の対 805／806 への入力として提供される。少なくとも一実施例では、ストレージ／計算の対 801／802、及び 805／806 は、2つ以上のニューラル・ネットワークの層に対応してもよい。少なくとも一実施例では、ストレージ／計算の対 801／802、及び 805／806 の後に、又はそれと並列に、追加のストレージ／計算の対（図示せず）が、推論及び／又は訓練論理 815 に含まれてもよい。

【0075】

ニューラル・ネットワークの訓練及び導入

図 9 は、少なくとも一実施例による、ディープ・ニューラル・ネットワークの訓練及び導入を示す。少なくとも一実施例では、未訓練ニューラル・ネットワーク 906 が、訓練データ・セット 902 を使用して訓練される。少なくとも一実施例では、訓練フレームワーク 904 は、PyTorch フレームワークであり、一方他の実施例では、訓練フレームワーク 904 は、TensorFlow、Boost、Caffe、マイクロソフト Cognitive Toolkit／CNTK、MXNet、Chainer、Keras、DeepLearning4j、又は他の訓練フレームワークである。少なくとも一実施例では、訓練フレームワーク 904 は、未訓練ニューラル・ネットワーク 906 を訓練し、本明細書に記載の処理リソースを使用してそれが訓練されるのを可能にして、訓練済みニューラル・ネットワーク 908 を生成する。少なくとも一実施例では、重みは、ランダムに選択されてもよく、又はディープ・ビリーフ・ネットワークを使用した事前訓練によって選択されてもよい。少なくとも一実施例では、訓練は、教師あり、一部教師あり、又は教師なしのいずれかのやり方で実行されてもよい。

【0076】

少なくとも一実施例では、未訓練ニューラル・ネットワーク 906 は教師あり学習を使用して訓練され、ここで訓練データ・セット 902 は、入力に対する所望の出力と対になった入力を含み、又は訓練データ・セット 902 は、既知の出力を有する入力を含み、ニューラル・ネットワーク 906 の出力が手動で採点される。少なくとも一実施例では、未訓練ニューラル・ネットワーク 906 は教師ありのやり方で訓練され、訓練データ・セット 902 からの入力を処理し、結果として得られた出力を、予想の又は所望の出力のセットと比較する。少なくとも一実施例では、次いで、誤差が、未訓練ニューラル・ネットワーク 906 を通って逆伝播される。少なくとも一実施例では、訓練フレームワーク 904 は、未訓練ニューラル・ネットワーク 906 を制御する重みを調節する。少なくとも一実施例では、訓練フレームワーク 904 は、未訓練ニューラル・ネットワーク 906 が、新規データ・セット 912 などの入力データに基づき、結果 914 などにおいて正しい答えを生成するのに好適な訓練済みニューラル・ネットワーク 908 などのモデルに向かって、どれだけ良好に収束しているかを監視するツールを含む。少なくとも一実施例では、訓練フレームワーク 904 は、未訓練ニューラル・ネットワーク 906 を繰り返し訓練する一方、損失関数、及び確率的勾配降下法などの調整アルゴリズムを使用して、未訓練ニューラル・ネットワーク 906 の出力を精緻化するように重みを調整する。少なくとも一実施例では、訓練フレームワーク 904 は、未訓練ニューラル・ネットワーク 906 が所望の精度に到達するまで未訓練ニューラル・ネットワーク 906 を訓練する。少なくとも一実施例では、次いで訓練済みニューラル・ネットワーク 908 を、任意の数の機械学習動作を実装するように導入することができる。

【0077】

少なくとも一実施例では、未訓練ニューラル・ネットワーク 906 は、教師なし学習を使用して訓練され、ここで未訓練ニューラル・ネットワーク 906 は、ラベルなしデータを使用して自らを訓練しようとする。少なくとも一実施例では、教師なし学習の訓練データ・セット 902 は、いかなる関連出力データ又は「グラウンド・トゥルース」データもない入力データを含む。少なくとも一実施例では、未訓練ニューラル・ネットワーク 906 は、訓練データ・セット 902 内でグループ化を学習することができ、個々の入力が、未訓練データ・セット 902 にどのように関係しているかを判定することができる。少な

くとも一実施例では、新規データ・セット 912 の次元を低減するのに有用な動作を実行することができる訓練済みニューラル・ネットワーク 908 内で教師なし訓練を使用して、自己組織化マップを生成することができる。少なくとも一実施例では、教師なし訓練を使用して異常検出を実行することもでき、異常検出は、新規データ・セット 912 の通常のパターンから逸脱した、新規データ・セット 912 内のデータ点を識別できるようにする。

【0078】

少なくとも一実施例では、半教師あり学習が使用されてもよく、それは、ラベル付きデータとラベルなしデータが訓練データ・セット 902 に混在している技法である。少なくとも一実施例では、訓練フレームワーク 904 を使用して、伝達学習技法などによる漸次的学習が実行されてもよい。少なくとも一実施例では、漸次的学習により、訓練済みニューラル・ネットワーク 908 は、初期訓練中に訓練済みニューラル・ネットワーク 908 内に教え込まれた知識を忘れることなく、新規データ・セット 912 に適合できるようになる。

【0079】

データ・センタ

図 10 は、少なくとも一実施例が使用されてもよい例示的なデータ・センタ 1000 を示す。少なくとも一実施例では、データ・センタ 1000 は、データ・センタ・インフラストラクチャ層 1010、フレームワーク層 1020、ソフトウェア層 1030、及びアプリケーション層 1040 を含む。

【0080】

図 10 に示すように、少なくとも一実施例では、データセンタ・インフラストラクチャ層 1010 は、リソース・オーケストレータ 1012 と、グループ化済みコンピューティング・リソース 1014 と、ノード・コンピューティング・リソース（「ノード C. R.」） 1016（1）～1016（N）とを含んでもよく、ここで「N」は正の整数を表す（他の図で使用されるものとは異なる整数「N」であってもよい）。少なくとも一実施例では、ノード C. R. 1016（1）～1016（N）は、任意の数の中央処理装置（「CPU」）又は（アクセラレータ、フィールド・プログラマブル・ゲート・アレイ（FPGA）、グラフィックス・プロセッサなどを含む）他のプロセッサ、メモリ・ストレージ・デバイス 1018（1）～1018（N）（たとえば、ダイナミック読取り専用メモリ、半導体ストレージ・ドライブ又はディスク・ドライブ）、ネットワーク入力／出力（「NW I/O」：network input/output）デバイス、ネットワーク・スイッチ、仮想機械（「VM」：virtual machine）、電源モジュール、及び冷却モジュールを含んでもよいが、これらに限定されない。少なくとも一実施例では、ノード C. R. 1016（1）～1016（N）のうち 1 つ又は複数のノード C. R. は、上述したコンピューティング・リソースのうちの 1 つ又は複数を有するサーバであってもよい。

【0081】

少なくとも一実施例では、グループ化済みコンピューティング・リソース 1014 は、1 つ若しくは複数のラック（図示せず）内に収容されたノード C. R. の別々のグループ、又は様々なグラフィカル・ロケーション（同じく図示せず）においてデータ・センタに収容された多数のラックを含んでもよい。少なくとも一実施例では、グループ化済みコンピューティング・リソース 1014 内のノード C. R. の別々のグループは、1 つ若しくは複数のワークロードをサポートするように構成又は配分されてもよいグループ化済みのコンピュート・リソース、ネットワーク・リソース、メモリ・リソース、又はストレージ・リソースを含んでもよい。少なくとも一実施例では、CPU 又はプロセッサを含むいくつかのノード C. R. は、1 つ又は複数のラック内でグループ化されて、1 つ又は複数のワークロードをサポートするためのコンピュート・リソースが提供されてもよい。少なくとも一実施例では、1 つ又は複数のラックはまた、任意の数の電源モジュール、冷却モジュール、及びネットワーク・スイッチを任意の組合せで含んでもよい。

【0082】

少なくとも一実施例では、リソース・オーケストレータ1012は、1つ又は複数のノードC.R.1016(1)～1016(N)及び／若しくはグループ化済みコンピューティング・リソース1014を構成してもよく、又は他のやり方で制御してもよい。少なくとも一実施例では、リソース・オーケストレータ1012は、データ・センタ1000用のソフトウェア設計インフラストラクチャ(「SDI」：software design infrastructure)管理エンティティを含んでもよい。少なくとも一実施例では、リソース・オーケストレータ812は、ハードウェア、ソフトウェア、又はこれらの何らかの組合せを含んでもよい。

【0083】

図10に示す少なくとも一実施例では、フレームワーク層1020は、ジョブ・スケジューラ1022、構成マネージャ1024、リソース・マネージャ1026、及び分配ファイル・システム1028を含む。少なくとも一実施例では、フレームワーク層1020は、ソフトウェア層1030のソフトウェア1032、及び／又はアプリケーション層1040の1つ若しくは複数のアプリケーション1042をサポートするためのフレームワークを含んでもよい。少なくとも一実施例では、ソフトウェア1032又はアプリケーション1042はそれぞれ、アマゾン・ウェブ・サービス、グーグル・クラウド、及びマイクロソフト・アジュールによって提供されるものなど、ウェブ・ベースのサービス・ソフトウェア又はアプリケーションを含んでもよい。少なくとも一実施例では、フレームワーク層1020は、大規模なデータ処理(たとえば「ビック・データ」)のために分配ファイル・システム1028を使用することができるApache Spark(登録商標)(以下「Spark」)など、無料でオープン・ソースのソフトウェア・ウェブ・アプリケーション・フレームワークの一種であってもよいが、これに限定されない。少なくとも一実施例では、ジョブ・スケジューラ1032は、データ・センタ1000の様々な層によってサポートされるワークロードのスケジューリングを容易にするために、Sparkドライバを含んでもよい。少なくとも一実施例では、構成マネージャ1024は、ソフトウェア層1030、並びに大規模なデータ処理をサポートするためのSpark及び分配ファイル・システム1028を含むフレームワーク層1020などの異なる層を構成することが可能であってもよい。少なくとも一実施例では、リソース・マネージャ1026は、分配ファイル・システム1028及びジョブ・スケジューラ1022をサポートするようにマッピング若しくは配分されたクラスタ化済み又はグループ化済みのコンピューティング・リソースを管理することが可能であってもよい。少なくとも一実施例では、クラスタ化済み又はグループ化済みのコンピューティング・リソースは、データ・センタ・インフラストラクチャ層1010にあるグループ化済みコンピューティング・リソース1014を含んでもよい。少なくとも一実施例では、リソース・マネージャ1026は、リソース・オーケストレータ1012と連携して、これらのマッピング又は配分されたコンピューティング・リソースを管理してもよい。

【0084】

少なくとも一実施例では、ソフトウェア層1030に含まれるソフトウェア1032は、ノードC.R.1016(1)～1016(N)、グループ化済みコンピューティング・リソース1014、及び／又はフレームワーク層1020の分配ファイル・システム1028のうちの少なくとも一部分によって使用されるソフトウェアを含んでもよい。少なくとも一実施例では、1つ又は複数のタイプのソフトウェアは、インターネット・ウェブ・ページ検索ソフトウェア、電子メール・ウイルス・スキャン・ソフトウェア、データベース・ソフトウェア、及びストリーミング・ビデオ・コンテンツ・ソフトウェアを含んでもよいが、これらに限定されない。

【0085】

少なくとも一実施例では、アプリケーション層1040に含まれるアプリケーション1042は、ノードC.R.1016(1)～1016(N)、グループ化済みコンピューティング・リソース1014、及び／又はフレームワーク層1020の分配ファイル・シ

10

20

30

40

50

ステム 1028 のうちの少なくとも一部分によって使用される 1 つ若しくは複数のタイプのアプリケーションを含んでもよい。少なくとも一実施例では、1 つ若しくは複数のタイプのアプリケーションは、任意の数のゲノム学アプリケーション、認識コンピュート、並びに訓練若しくは推論のソフトウェア、機械学習フレームワーク・ソフトウェア（たとえば、PyTorch、TensorFlow、Caffe など）を含むアプリケーション及び機械学習アプリケーション、又は 1 つ若しくは複数の実施例と併せて使用される他の機械学習アプリケーションを含んでもよいが、これらに限定されない。

【0086】

少なくとも一実施例では、構成マネージャ 1024、リソース・マネージャ 1026、及びリソース・オーケストレータ 1012 のうちのいずれかは、任意の技術的に実行可能なやり方で取得された任意の量及びタイプのデータに基づき、任意の数及びタイプの自己修正措置を実装してもよい。少なくとも一実施例では、自己修正措置は、データ・センタ 1000 のデータ・センタ演算子が、不良の恐れのある構成を決定しないようにし、十分に利用されていない且つ／又は性能の低いデータ・センタの部分をなくせるようにしてもよい。

【0087】

少なくとも一実施例では、データ・センタ 1000 は、1 つ若しくは複数の機械学習モデルを訓練し、又は本明細書に記載の 1 つ若しくは複数の実施例による 1 つ若しくは複数の機械学習モデルを使用して情報を予測若しくは推論するためのツール、サービス、ソフトウェア、又は他のリソースを含んでもよい。たとえば、少なくとも一実施例では、機械学習モデルは、データ・センタ 1000 に関して上述したソフトウェア及びコンピューティング・リソースを使用して、ニューラル・ネットワーク・アーキテクチャに従って重みパラメータを計算することによって、訓練されてもよい。少なくとも一実施例では、1 つ又は複数のニューラル・ネットワークに対応する訓練済み機械学習モデルは、本明細書に記載の 1 つ又は複数の技法によって計算された重みパラメータを使用することにより、データ・センタ 1000 に関して上述したリソースを使用して、情報を推論又は予測するために使用されてもよい。

【0088】

少なくとも一実施例では、データ・センタは、上述したリソースを使用して訓練及び／又は推論を実行するために、CPU、特定用途向け集積回路（ASIC）、GPU、FPGA、又は他のハードウェアを使用してもよい。さらに、上述した 1 つ又は複数のソフトウェア及び／又はハードウェアのリソースは、画像認識、音声認識、又は他の人工知能サービスなどの情報の訓練又は推論の実行を、ユーザが行えるようにするためのサービスとして構成されてもよい。

【0089】

推論及び／又は訓練論理 815 を使用して、1 つ若しくは複数の実施例に関連する推論及び／又は訓練の動作が実行される。推論及び／又は訓練論理 815 に関する詳細事項は、図 8A 及び／又は図 8B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 815 は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図 10 のシステムにおいて使用されてもよい。

【0090】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照し

10

20

30

40

50

て適応される。

【0091】

自律車両

図11Aは、少なくとも一実施例による自律車両1100の例を示す。少なくとも一実施例では、自律車両1100（或いは、本明細書において「車両1100」と呼ばれる）は、限定することなく、車、トラック、バス、及び／又は1人若しくは複数の乗員を収容する別のタイプの車両などの乗用車とすることができる。少なくとも一実施例では、車両1100は、貨物運搬用のセミ・トラクタのトレーラ・トラックであってもよい。少なくとも一実施例では、車両1100は、航空機、ロボット車両、又は他の種類の車両であってもよい。

【0092】

自律車両は、米国運輸省の一部門である全米高速道路交通安全局（「NHTSA」：National Highway Traffic Safety Administration）、及び自動車技術者協会（「SAE」：Society of Automotive Engineers）の「自動車用運転自動化システムのレベル分類及び定義（Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles）」（たとえば、2018年6月15日発行の規格No. J3016-201806、2016年9月30日発行の規格No. J3016-201609、及びこの規格の旧版及び新版）により定義される自動化レベルという観点から説明されてもよい。少なくとも一実施例では、車両1100は、自律運転レベルのレベル1～レベル5のうちの1つ又は複数による機能性に対応可能であってもよい。たとえば、少なくとも一実施例では、車両1100は、実施例に応じて、条件付き自動化（レベル3）、高度自動化（レベル4）、及び／又は完全自動化（レベル5）に対応可能であってもよい。

【0093】

少なくとも一実施例では、車両1100は、限定することなく、シャシ、車両本体、ホイール（2本、4本、6本、8本、18本など）、タイヤ、車軸、及び車両の他の構成要素などの構成要素を含んでもよい。少なくとも一実施例では、車両1100は、限定することなく、内燃機関、ハイブリッド電力プラント、完全電気エンジン、及び／又は別のタイプの推進システムなどの推進システム1150を含んでもよい。少なくとも一実施例では、推進システム1150は、車両1100のドライブ・トレインに連結されてもよく、ドライブ・トレインは、限定することなく、車両1100の推進を可能にするためのトランスミッションを含んでもよい。少なくとも一実施例では、推進システム1150は、スロットル／アクセル1152からの信号を受信したことに応答して、制御されてもよい。

【0094】

少なくとも一実施例では、限定することなくハンドルを含んでもよい操縦システム1154は、推進システム1150が動作しているときに（たとえば、車両1100が動いているときに）車両1100を（たとえば所望の経路又はルートに沿って）操縦するために使用される。少なくとも一実施例では、操縦システム1154は、操縦アクチュエータ1156から信号を受信してもよい。少なくとも一実施例では、ハンドルは、完全自動化（レベル5）の機能性に関しては任意選択であってもよい。少なくとも一実施例では、ブレーキ・アクチュエータ1148及び／又はブレーキ・センサからの信号を受信したことに応答して車両ブレーキを動作させるために、ブレーキ・センサ・システム1146が使用されてもよい。

【0095】

少なくとも一実施例では、1つ又は複数のシステム・オン・チップ（「SoC」：system on chip）（図11Aには示さず）及び／若しくはグラフィックス・プロセッシング・ユニット（「GPU」）を限定することなく含んでもよいコントローラ1136は、車両1100の1つ又は複数の構成要素及び／若しくはシステムに（たとえば、

コマンドを表す) 信号を提供する。たとえば、少なくとも一実施例では、コントローラ 1136 は、ブレーキ・アクチュエータ 1148 を介して車両ブレーキを動作させるための信号、操縦アクチュエータ 1156 を介して操縦システム 1154 を動作させるための信号、スロットル／アクセル 1152 を介して推進システム 1150 を動作させるための信号を送信してもよい。少なくとも一実施例では、コントローラ 1136 は、自律運転を可能にし、且つ／又は運転車両 1100 において人間のドライバを支援するために、センサ信号を処理し、動作コマンド（たとえばコマンドを表す信号）を出力する 1 つ又は複数の搭載（たとえば一体型の）コンピューティング・デバイス（たとえば、スーパーコンピュータ）を含んでもよい。少なくとも一実施例では、コントローラ 1136 は、自律運転機能のための第 1 のコントローラ、機能的な安全機能のための第 2 のコントローラ、人工知能機能（たとえば、コンピュータ・ビジョン）のための第 3 のコントローラ、インフォテイメント機能のための第 4 のコントローラ、緊急事態における冗長性のための第 5 のコントローラ、及び／又は他のコントローラを含んでもよい。少なくとも一実施例では、単一のコントローラが、上記機能性のうちの 2 つ以上に対処してもよく、2 つ以上のコントローラが、単一の機能性に対処してもよく、且つ／又はこれらの何らかの組合せであってもよい。

【0096】

少なくとも一実施例では、コントローラ 1136 は、1 つ又は複数のセンサから受信したセンサ・データ（たとえば、センサ入力）に応答して、車両 1100 の 1 つ又は複数の構成要素及び／若しくはシステムを制御するための信号を提供する。少なくとも一実施例では、センサ・データは、たとえば限定することなく、全地球的航法衛星システム（「GNSS」：global navigation satellite system）センサ 1158（たとえば、全地球測位システム・センサ）、RADAR センサ 1160、超音波センサ 1162、LIDAR センサ 1164、慣性計測装置（「IMU」：inertial measurement unit）センサ 1166（たとえば、加速度計、ジャイロスコープ、磁気コンパス、磁力計など）、マイクロフォン 1196、ステレオ・カメラ 1168、広角カメラ 1170（たとえば、魚眼カメラ）、赤外線カメラ 1172、周囲カメラ 1174（たとえば、360 度カメラ）、長距離カメラ（図 11A には示さず）、中距離カメラ（図 11A には示さず）、（たとえば、車両 1100 のスピードを計測するための）スピード・センサ 1144、振動センサ 1142、操縦センサ 1140、（たとえば、ブレーキ・センサ・システム 1146 の一部分としての）ブレーキ・センサ、及び／又は他のタイプのセンサから、受信されてもよい。

【0097】

少なくとも一実施例では、コントローラ 1136 のうちの 1 つ又は複数のは、車両 1100 の計器クラスタ 1132 からの（たとえば入力データによって表される）入力を受信し、ヒューマン・マシン・インターフェース（「HMI」：human-machine interface）ディスプレイ 1134、可聴アナウンスエータ、拡声器を介して、且つ／又は車両 1100 の他の構成要素を介して、（たとえば、出力データ、ディスプレイ・データなどによって表される）出力を提供してもよい。少なくとも一実施例では、出力は、車両速度、スピード、時間、地図データ（たとえば、ハイ・デフィニション・マップ（図 11A には示さず）、ロケーション・データ（たとえば、地図上などの車両 1100 のロケーション）、方向、他車両のロケーション（たとえば、占有グリッド）、コントローラ 1136 が感知した物体及び物体の状態についての情報などの情報を含んでもよい。たとえば、少なくとも一実施例では、HMI ディスプレイ 1134 は、1 つ若しくは複数の物体（たとえば、道路標識、警告標識、信号の変化など）の存在についての情報、及び／又は車両が行った、行っている、又はこれから行う運転操作についての情報（たとえば、現在車線変更中、3.22 km（2 マイル）先の出口 34B を出る、など）を表示してもよい。

【0098】

少なくとも一実施例では、車両 1100 はさらにネットワーク・インターフェース 11

24を含み、このネットワーク・インターフェースは、1つ又は複数のネットワークを介して通信するためのワイヤレス・アンテナ1126及び／又はモデムを使用してもよい。たとえば、少なくとも一実施例では、ネットワーク・インターフェース1124は、ロング・ターム・エボリューション（「LTE」：Long-Term Evolution）、広帯域符号分割多元接続（「WCDMA（登録商標）」：Wideband Code Division Multiple Access）、ユニバーサル・モバイル・テレコミュニケーション・システム（「UMTS」：Universal Mobile Telecommunications System）、グローバル・システム・フォー・モバイル・コミュニケーションズ（「GSM」：Global System for Mobile communication）、IMT-CDMAマルチ・キャリア（「CDMA2000」）ネットワークなどを介して通信可能であってもよい。また、少なくとも一実施例では、ワイヤレス・アンテナ1126は、Bluetooth、Bluetooth Low Energy（「LE」：Low Energy）、Z-Wave、ZigBeeなどのローカル・エリア・ネットワーク、及び／又はLoRaWAN、SigFoxなどの低電力広域ネットワーク（「LPWAN」：low power wide-area network）を使用して、環境内の物体同士間（たとえば車両、モバイル・デバイスなど）での通信を可能にしてもよい。

【0099】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図11Aのシステムにおいて使用されてもよい。

【0100】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0101】

図11Bは、少なくとも一実施例による図11Aの自律車両1100についてカメラのロケーション及び視野の例を示す。少なくとも一実施例では、カメラ及びそれぞれの視野は、一例の実施例であり、限定するものではない。たとえば、少なくとも一実施例では、追加及び／又は代替のカメラが含まれてもよく、且つ／又はカメラが車両1100の異なるロケーションに位置付けられてもよい。

【0102】

少なくとも一実施例では、カメラのカメラ・タイプは、車両1100の構成要素及び／又はシステムとともに使用できるように適合されていてもよいデジタル・カメラを含んでもよいが、これに限定されない。少なくとも一実施例では、カメラは、自動車安全性要求レベル（「ASIL」：automotive safety integrity level）B及び／又は別のASILにおいて動作してもよい。少なくとも一実施例では、カメラ・タイプは、実施例に応じて、毎秒60フレーム（fps：frames per second）、1220fps、240fpsなど、任意の画像捕捉率に対応可能であってもよい。少なくとも一実施例では、カメラは、ロール・シャッタ、グローバル・シャ

10

20

30

40

50

ツタ、別のタイプのシャッタ、又はこれらの組合せを使用することが可能であってもよい。少なくとも一実施例では、カラー・フィルタ・アレイは、赤色、クリア、クリア、クリア（「RCCC」：red clear clear clear）のカラー・フィルタ・アレイ、赤色、クリア、クリア、青色（「RCCB」：red clear clear blue）のカラー・フィルタ・アレイ、赤色、青色、緑色、クリア（「RBGC」：red blue green clear）のカラー・フィルタ・アレイ、Foveon X3のカラー・フィルタ・アレイ、ベイヤー・センサ（RGGB）のカラー・フィルタ・アレイ、モノクロ・センサのカラー・フィルタ・アレイ、及び／又は別のタイプのカラー・フィルタ・アレイを含んでもよい。少なくとも一実施例では、光感度を上げるために、RCCC、RCCB、及び／又はRBGCのカラー・フィルタ・アレイを有するカメラなど、クリア・ピクセル・カメラが使用されてもよい。

10

【0103】

少なくとも一実施例では、カメラのうちの1つ又は複数を使用して、先進ドライバ支援システム（「ADAS」：advanced driver assistance systems）機能が（たとえば、冗長設計又はフェイル・セーフ設計の一部として）実行されてもよい。たとえば、少なくとも一実施例では、多機能モノ・カメラが設置されて、車線逸脱警告、交通標識支援、及びインテリジェント・ヘッドライト制御を含む機能が提供されてもよい。少なくとも一実施例では、カメラのうちの1つ又は複数（たとえばすべてのカメラ）は、画像データ（たとえばビデオ）の記録と提供を同時に行ってもよい。

【0104】

20

少なくとも一実施例では、1つ又は複数のカメラは、カメラの画像データ捕捉性能を妨げる恐れのある迷光及び車両1100内部からの反射（たとえば、ダッシュボードからフロントガラスに反射される反射）をなくすために、カスタム設計の（3次元（「3D」：three-dimensional）印刷された）アセンブリなどの取付けアセンブリに取り付けられてもよい。ドアミラー取付けアセンブリを参照すると、少なくとも一実施例では、ドアミラー・アセンブリは、カメラ取付けプレートがドアミラーの形の合うように、カスタムで3D印刷されてもよい。少なくとも一実施例では、カメラは、ドアミラーと一体であってもよい。少なくとも一実施例では、サイド・ビュー・カメラについて、カメラはこの場合もキャビンの各角にある4本のピラーに一体化されてもよい。

【0105】

30

少なくとも一実施例では、車両1100前方の環境の一部分を含む視野を有するカメラ（たとえば正面カメラ）は周囲のビューに対して使用されて、正面の経路及び障害物を識別しやすくするとともに、コントローラ1136及び／又は制御SOCのうちの1つ又は複数とともに使用されて、占有グリッドの生成及び／又は好ましい車両経路の判定に不可欠な情報の提供を補助してもよい。少なくとも一実施例では、正面カメラを使用して、緊急ブレーキ、歩行者検出、及び衝突回避を限定することなく含む、LIDARと同様のADAS機能のうちの多くが実行されてもよい。少なくとも一実施例では、正面カメラはまた、車線逸脱警告（「LDW」：Lane Departure Warnings）、自律クルーズ・コントロール（「ACC」：Autonomous Cruise Control）、及び／又は交通標識認識などの他の機能を限定することなく含むADASの機能及びシステムのために使用されてもよい。

40

【0106】

少なくとも一実施例では、たとえばCMOS：complementary metal oxide semiconductor（「相補型金属酸化膜半導体」）カラー撮像装置を含む単眼カメラのプラットフォームを含む様々なカメラが、正面構成で使用されてもよい。少なくとも一実施例では、周囲からビューに入ってくる物体（たとえば歩行者、クロス・トラフィック、又は自転車）を感知するために、広角カメラ1170が使用されてもよい。図11Bには1つの広角カメラ1170しか示していないが、他の実施例では、車両1100には（ゼロを含む）任意の数の広角カメラが存在してもよい。少なくとも一実施例では、特にニューラル・ネットワークがそれに対してまだ訓練されていない物体に

50

ついて、深度ベースの物体検出のために、任意の数の長距離カメラ 1198（たとえば、長距離ビューのステレオ・カメラの対）が使用されてもよい。少なくとも一実施例では、長距離カメラ 1198 はまた、物体検出及び分類、並びに基本的な物体追跡に使用されてもよい。

【0107】

少なくとも一実施例では、任意の数のステレオ・カメラ 1168 は、正面構成にも含まれてよい。少なくとも一実施例では、1つ又は複数のステレオ・カメラ 1168 は、拡張可能な処理ユニットを備えた一体型制御ユニットを含んでもよく、この制御ユニットは、一体型のコントローラ・エリア・ネットワーク（「CAN」：Controller Area Network）又はイーサネット（登録商標）・インターフェースを単一チップ上に有するプログラム可能論理（「FPGA」）及びマルチ・コア・マイクロプロセッサを提供してもよい。少なくとも一実施例では、こうしたユニットは、画像内のすべての点に対する距離推定を含め、車両 1100 の環境の 3D マップを生成するのに使用されてもよい。少なくとも一実施例では、ステレオ・カメラ 1168 のうちの 1つ又は複数の、限定することなくコンパクト・ステレオ・ビジョン・センサを含んでもよく、このセンサは、車両 1100 からターゲット物体までの距離を測定し、生成された情報（たとえば、メタデータ）を使用して自律緊急ブレーキ及び車線逸脱警告の機能をアクティブ化することができる 2つのカメラ・レンズ（左右に 1つずつ）及び画像処理チップを、限定することなく含んでもよい。少なくとも一実施例では、本明細書に記載のものに加えて、又はその代わりに、他のタイプのステレオ・カメラ 1168 が使用されてもよい。

【0108】

少なくとも一実施例では、車両 1100 の側方の環境の一部を含む視野を有するカメラ（たとえば、サイド・ビュー・カメラ）が、周囲のビューのために使用されて、占有グリッドの作製及び更新、並びに側面衝突警告の生成のために使用される情報を提供してもよい。たとえば、少なくとも一実施例では、周囲カメラ 1174（たとえば図 11B に示すように 4つの周囲カメラ）を、車両 1100 に配置することができる。少なくとも一実施例では、周囲カメラ 1174 は、限定することなく、任意の数及び組合せの広角カメラ、魚眼カメラ、360度カメラ及び／又は同様のカメラを含んでもよい。たとえば、少なくとも一実施例では、4つの魚眼カメラが、車両 1100 の前方、後方、及び側方に配置されてもよい。少なくとも一実施例では、車両 1100 は、3つの周囲カメラ 1174（たとえば、左、右、及び後方）を使用してもよく、第 4の周囲カメラとして、1つ又は複数の他のカメラ（たとえば正面カメラ）を活用してもよい。

【0109】

少なくとも一実施例では、車両 1100 後方の環境の一部を含む視野を有するカメラ（たとえば、リア・ビュー・カメラ）が、駐車支援、周囲のビュー、後方衝突警告のために使用されて、占有グリッドの作製及び更新がなされてもよい。少なくとも一実施例では、本明細書に記載の正面カメラとしても好適なカメラ（たとえば、長距離カメラ 1198、及び／又は中距離カメラ 1176、ステレオ・カメラ 1168）、赤外線カメラ 1172 など）を含むが、これらに限定されない多種多様なカメラが使用されてもよい。

【0110】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理 815 が使用される。推論及び／又は訓練論理 815 に関する詳細事項は、図 8A 及び／又は図 8B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 815 は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図 11B のシステムにおいて使用されてもよい。

【0111】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若

10

20

30

40

50

しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0112】

図11Cは、少なくとも一実施例による図11Aの自律車両1100の例示的システム・アーキテクチャを示すブロック図である。少なくとも一実施例では、図11Cの車両1100の構成要素、特徴、及びシステムのそれぞれは、バス1102を介して接続されるものとして示される。少なくとも一実施例では、バス1102は、限定することなく、CANデータ・インターフェース（或いは、本明細書において（CANバス）と呼ばれる）を含んでもよい。少なくとも一実施例では、CANは、ブレーキの作動、加速、ブレーキ制御、操縦、フロントガラス・ワイパなど、車両1100の様々な特徴及び機能の制御を補助するために使用される車両1100内部のネットワークであってもよい。少なくとも一実施例では、バス1102は、それぞれが独自の一意の識別子（たとえばCAN ID）をもつ数十又はさらには数百のノードを有するように構成されてもよい。少なくとも一実施例では、バス1102は、ハンドル角度、対地スピード、エンジンの毎分回転数（「RPM」：revolutions per minute）、ボタン位置、及び／又は他の車両状態インジケータを見いだすように読み取られてもよい。少なくとも一実施例では、バス1102は、ASILのBに準拠したCANバスであってもよい。

【0113】

少なくとも一実施例では、CANに加えて、又はその代わりに、FlexRay及び／又はEthernet（登録商標）プロトコルが使用されてもよい。少なくとも一実施例では、バス1102を形成する任意の数のバスが存在してもよく、これには、限定することなく、ゼロ以上のCANバス、ゼロ以上のFlexRayバス、ゼロ以上のEthernet（登録商標）バス、及び／又は他のプロトコルを使用するゼロ以上の他のタイプのバスが含まれてもよい。少なくとも一実施例では、2つ以上のバスを使用して異なる機能が実行されてもよく、且つ／又はそれらを使用して冗長性が与えられてもよい。たとえば、第1のバスが衝突回避機能のために使用され、第2のバスが作動制御のために使用されてもよい。少なくとも一実施例では、バス1102の各バスは、車両1100の構成要素のいずれかと通信してもよく、バス1102のうちの2つ以上のバスが対応する構成要素と通信してもよい。少なくとも一実施例では、任意の数のシステム・オン・チップ（「SoC」）1104（SoC1104（A）及びSoC1104（B）など）のそれぞれ、コントローラ1136のそれぞれ、及び／又は車両内の各コンピュータは、同じ入力データ（たとえば、車両1100のセンサからの入力）にアクセス可能であってもよく、CANバスなどの共通のバスに接続されてもよい。

【0114】

少なくとも一実施例では、車両1100は、図11Aに関して本明細書に記載するものなど、1つ又は複数のコントローラ1136を含んでもよい。少なくとも一実施例では、コントローラ1136は、様々な機能に使用されてもよい。少なくとも一実施例では、コントローラ1136は、車両1100の様々な他の構成要素及びシステムのうちの任意のものに結合されてもよく、車両1100、車両1100の人工知能、車両1100のインフォテインメント及び／又は他の機能の制御に使用されてもよい。

【0115】

少なくとも一実施例では、車両1100は、任意の数のSoC1104を含んでもよい。少なくとも一実施例では、SoC1104のそれぞれは、限定することなく、中央処理装置（「CPU」）1106、グラフィックス・プロセッシング・ユニット（「GPU」）1108、プロセッサ1110、キャッシュ1112、アクセラレータ1114、デー

10

20

30

40

50

タ・ストア1116、及び／又は図示していない他の構成要素及び特徴を含んでもよい。少なくとも一実施例では、車両1100を様々なプラットフォーム及びシステムにおいて制御するために、SoC1104が使用されてもよい。たとえば、少なくとも一実施例では、SoC1104は、1つ又は複数のサーバ（図11Cには示さず）からネットワーク・インターフェース1124を介して地図のリフレッシュ及び／又は更新を得ることができるハイ・デフィニション（「HD」：High Definition）マップ1122を有するシステム（たとえば車両1100のシステム）に組み込まれてもよい。

【0116】

少なくとも一実施例では、CPU1106は、CPUクラスタ、又はCPUコンプレックス（或いは本明細書において「CCPLEX」と呼ばれる）を含んでもよい。少なくとも一実施例では、CPU1106は、複数のコア及び／又はレベル2（「L2」）キャッシュを含んでもよい。たとえば、少なくとも一実施例では、CPU1106は、コヒーレントなマルチプロセッサ構成において8つのコアを含んでもよい。少なくとも一実施例では、CPU1106は、4つのデュアル・コア・クラスタを含んでもよく、ここで各クラスタは、専用のL2キャッシュ（たとえば、2メガバイト（MB）のL2キャッシュ）を有する。少なくとも一実施例では、CPU1106（たとえば、CCPLEX）は、CPU1106のクラスタの任意の組合せを、任意の所与の時間にアクティブ化できるようにする同時のクラスタ動作をサポートするように構成されてもよい。

【0117】

少なくとも一実施例では、CPU1106のうちの1つ又は複数のは、電力管理機能を実装してもよく、この機能は限定することなく、以下の特徴のうちの1つ又は複数を含む：個々のハードウェア・ブロックが、アイドル時に自動的にクロック・ゲート制御されて動的電力を節約することができる；割込み待ち（「WFI」：Wait for Interrupt）／イベント待ち（「WFE」：Wait for Event）命令の実行に起因してコアが能動的に命令を実行していないとき、各コア・クロックをゲート制御することができる；各コアを独立して電力ゲート制御することができる；すべてのコアがクロック・ゲート制御又は電力ゲート制御されるとき、各コア・クラスタを独立してクロック・ゲート制御することができる；且つ／又はすべてのコアが電力ゲート制御されるとき、各コア・クラスタを独立して電力ゲート制御することができる。少なくとも一実施例では、CPU1106はさらに、電力状態を管理するための拡張アルゴリズムを実装してもよく、ここで、許容された電力状態及び予想されるウェイクアップ時間が指定され、コア、クラスタ、及びCCPLEXが入るべき最良の電力状態はどれかを、ハードウェア／マイクロコードが判定する。少なくとも一実施例では、処理コアは、作業がマイクロコードにオフロードされた状態で、電力状態に入る簡単なシーケンスをソフトウェアにおいてサポートしてもよい。

【0118】

少なくとも一実施例では、GPU1108は、統合されたGPU（或いは、本明細書において「iGPU」と呼ばれる）を含んでもよい。少なくとも一実施例では、GPU1108は、プログラム可能であってもよく、並列なワークロードに対して効率的であってもよい。少なくとも一実施例では、GPU1108は、拡張テンソル命令セットを使用してもよい。少なくとも一実施例では、GPU1108は、1つ又は複数のストリーミング・マイクロプロセッサを含んでもよく、ここで各ストリーミング・マイクロプロセッサは、レベル1（「L1」）キャッシュ（たとえば少なくとも96KBのストレージ容量を有するL1キャッシュ）を含んでもよく、2つ以上のストリーミング・マイクロプロセッサは、L2キャッシュ（たとえば、512KBのストレージ容量を有するL2キャッシュ）を共有してもよい。少なくとも一実施例では、GPU1108は、少なくとも8つのストリーミング・マイクロプロセッサを含んでもよい。少なくとも一実施例では、GPU1108は、コンピュータ・アプリケーション・プログラミング・インターフェース（API：application programming interface）を使用してもよい。少なくとも一実施例では、GPU1108は、1つ又は複数の並列なコンピューティ

10

20

30

40

50

ング・プラットフォーム、及び／又はプログラミング・モジュール（たとえば、NVIDIAのCUDAモデル）を使用してもよい。

【0119】

少なくとも一実施例では、GPU1108のうちの1つ又は複数は、自動車用及び組み込まれたユース・ケースにおいて最良の性能になるように電力最適化されてもよい。たとえば、少なくとも一実施例では、GPU1108は、フィン電界効果トランジスタ（「FinFET」：Fin field-effect transistor）回路上で作製することができる。少なくとも一実施例では、各ストリーミング・マイクロプロセッサは、複数のブロックに区分けされた多数の混合精度の処理コアを組み込んでよい。たとえば、限定することなく64個のPF32コアと、32個のPF64コアを、4つの処理ブロックに区分けすることができる。少なくとも一実施例では、各処理ブロックに、16個のFP32コア、8個のFP64コア、16個のINT32コア、深層学習の行列演算用の2つの混合精度のNVIDIA Tensorコア、レベルゼロ（「L0」）命令キャッシュ、ワープ・スケジューラ、ディスパッチ・ユニット、及び／又は64KBのレジスタ・ファイルを配分することができる。少なくとも一実施例では、ストリーミング・マイクロプロセッサは、整数と浮動小数点の独立した並列のデータ経路を含み、コンピュータ処理とアドレッシング計算を混用することによってワークロードの効率的な実行を実現する。少なくとも一実施例では、ストリーミング・マイクロプロセッサは、独立したスレッド・スケジューリング機能を含み、並列スレッド間でよりきめ細かい同期及び連携を可能にしてもよい。少なくとも一実施例では、ストリーミング・マイクロプロセッサは、性能を向上させると同時にプログラミングを簡単にするために、L1データ・キャッシュと共有メモリ・ユニットの組合せを含んでもよい。

【0120】

少なくとも一実施例では、GPU1108のうちの1つ又は複数は、高帯域幅メモリ（「HBM」：high bandwidth memory）及び／又は16GBのHBM2メモリ・サブシステムを含み、いくつかの例では、約900GB／秒のピーク・メモリ帯域幅を提供してもよい。少なくとも一実施例では、HBMメモリに加えて、又はその代わりに、グラフィックス・ダブル・データ・レート・タイプの5つの同期ランダム・アクセス・メモリ（「GDDR5」：graphics double data rate type five）などの同期グラフィックス・ランダム・アクセス・メモリ（「SGRAM」：synchronous graphics random-access memory）が使用されてもよい。

【0121】

少なくとも一実施例では、GPU1108は、統合メモリ技術を含んでもよい。少なくとも一実施例では、アドレス・トランスレーション・サービス（「ATS」：address translation services）サポートを使用して、GPU1108が、CPU1106のページ・テーブルに直接アクセスできるようにしてもよい。少なくとも一実施例では、GPU1108メモリ管理ユニット（「MMU」：memory management unit）のGPUがミスに遭遇したときには、アドレス・トランスレーション要求が、CPU1106に送信されてもよい。少なくとも一実施例では、それに応答して、CPU1106のうちの2つのCPUは、自らのページ・テーブルで、仮想から物理へのアドレスのマッピングを探し、トランスレーションをGPU1108に送り返してもよい。少なくとも一実施例では、統合メモリ技術は、CPU1106とGPU1108の両方のメモリに対して単一の統合された仮想アドレス空間を与えることを可能にし、それにより、GPU1108のプログラミング、及びGPU1108へのアプリケーションの移植を簡単にする。

【0122】

少なくとも一実施例では、GPU1108は、他のプロセッサのメモリへのGPU1108のアクセス頻度を記録することができる任意の数のアクセス・カウンタを含んでもよい。少なくとも一実施例では、アクセス・カウンタは、最も頻繁にページにアクセスして

いるプロセッサの物理メモリに、メモリ・ページが確実に移動されるのを補助し、それにより、プロセッサ間で共有されるメモリ範囲の効率を向上させてもよい。

【0123】

少なくとも一実施例では、SoC1104のうちの1つ又は複数は、本明細書に記載のものを含む任意の数のキャッシュ1112を含んでもよい。たとえば、少なくとも一実施例では、キャッシュ1112は、CPU1106もGPU1108も利用可能な（たとえば、CPU1106とGPU1108の両方に接続された）レベル3（「L3」）キャッシュを含むことができる。少なくとも一実施例では、キャッシュ1112は、キャッシュ・コヒーレンス・プロトコルなど（たとえば、MEI、MESI、MSIなど）を使用することにより、線の状態を記録することができるライト・バック・キャッシュを含んでもよい。少なくとも一実施例では、L3キャッシュは、実施例に応じて4MBのメモリ以上を含んでもよいが、より小さいキャッシュ・サイズが使用されてもよい。

10

【0124】

少なくとも一実施例では、SoC1104のうちの1つ又は複数は、1つ又は複数のアクセラレータ1114（たとえば、ハードウェア・アクセラレータ、ソフトウェアアクセラレータ、又はこれらの組合せ）を含んでもよい。少なくとも一実施例では、SoC1104は、最適化されたハードウェア・アクセラレータ及び／又は大型のオン・チップ・メモリを含むことができるハードウェア加速クラスタを含んでもよい。少なくとも一実施例では、大型のオン・チップ・メモリ（たとえば、4MBのSRAM）は、ハードウェア加速クラスタが、ニューラル・ネットワーク及び他の計算を加速できるようにしてもよい。少なくとも一実施例では、ハードウェア加速クラスタを使用して、GPU1108を補完し、GPU1108のタスクのうちのいくつかをオフロードしてもよい（たとえば、他のタスクを実行できるようにGPU1108のサイクルをより多く解放してもよい）。少なくとも一実施例では、加速を受け入れるのに十分なほど安定している目的とするワークロード（たとえば、知覚、畳み込みニューラル・ネットワーク（「CNN」：convolutional neural network）、再帰ニューラル・ネットワーク（「RNN」：recurrent neural network）など）のために、アクセラレータ1114を使用することができる。少なくとも一実施例では、CNNは、領域ベースの、すなわち領域畳み込みニューラル・ネットワーク（「RCNN」：regional convolutional neural network）、及び（たとえば、物体検出に使用される）高速RCNN、又は他のタイプのCNNを含んでもよい。

20

30

【0125】

少なくとも一実施例では、アクセラレータ1114（たとえば、ハードウェア加速クラスタ）は、1つ又は複数の深層学習アクセラレータ（「DLA」：deep learning accelerator）を含んでもよい。少なくとも一実施例では、DLAは、限定することなく、1つ又は複数のTensor処理ユニット（「TPU」：Tensor processing units）を含んでもよく、このユニットは、深層学習アプリケーション及び推論のために、さらに毎秒10兆の演算を提供するように構成されてもよい。少なくとも一実施例では、TPUは、画像処理機能（たとえば、CNN、RCNNなど）を実行するように構成され、そのために最適化されたアクセラレータであってもよい。少なくとも一実施例では、DLAはさらに、ニューラル・ネットワーク・タイプと浮動小数点演算の特定のセット、並びに推論のために最適化されてもよい。少なくとも一実施例では、DLAの設計により、典型的な汎用GPUよりもミリメートル当たりの性能を向上させることができ、典型的には、CPUの性能を大いに上回る。少なくとも一実施例では、TPUは、たとえば特徴と重みの両方のためのINT8、INT16、及びFP16のデータ型をサポートする単一インスタンスの畳み込み関数、並びに後処理関数を含む、いくつかの関数を実行してもよい。少なくとも一実施例では、DLAは、たとえば、限定することなく、カメラ・センサからのデータを使用した物体識別及び検出のためのCNN、カメラ・センサからのデータを使用した距離推定のためのCNN、マイクロフォンからのデータを使用した緊急車両検出、及び識別、及び検出のためのCNN、カメラ・セン

40

50

サからのデータを使用した顔認識及び車両所有者識別のためのCNN、並びに／又はセキュリティ及び／若しくは安全に関するイベントのためのCNNを含め、様々な機能のうちのいずれかのための処理済み若しくは未処理のデータに対して、迅速且つ効率的にニューラル・ネットワーク、特にCNNを実行してもよい。

【0126】

少なくとも一実施例では、DLAは、GPU1108の任意の機能を実行してもよく、たとえば推論アクセラレータを使用することにより、設計者は、任意の機能のためにDLA又はGPU1108のいずれかをターゲットにしてもよい。たとえば、少なくとも一実施例では、設計者は、CNN及び浮動小数点演算の処理をDLAに集中させ、他の機能をGPU1108及び／又は他のアクセラレータ1114に任せてもよい。

10

【0127】

少なくとも一実施例では、アクセラレータ1114は、プログラマブル・ビジョン・アクセラレータ（「PVA」：programmable vision accelerator）を含んでもよく、このプログラマブル・ビジョン・アクセラレータは、本明細書において代替的にコンピュータ・ビジョン・アクセラレータと呼ばれてもよい。少なくとも一実施例では、PVAは、先進ドライバ支援システム（「ADAS」）1138、自律運転、拡張現実（「AR」：augmented reality）アプリケーション、及び／又は仮想現実（「VR」：virtual reality）アプリケーションのために、コンピュータ・ビジョン・アルゴリズムを加速するように設計及び構成されてもよい。少なくとも一実施例では、PVAにより、性能と融通性との均衡が保たれてもよい。たとえば、少なくとも一実施例では、各PVAは、たとえば限定することなく、任意の数の縮小命令セット・コンピュータ（「RISC」：reduced instruction set computer）コア、ダイレクト・メモリ・アクセス（「DMA」：direct memory access）、及び／又は任意の数のベクトル・プロセッサを含んでもよい。

20

【0128】

少なくとも一実施例では、RISCコアは、画像センサ（たとえば、本明細書に記載の任意のカメラの画像センサ）、画像信号プロセッサなどと相互作用してもよい。少なくとも一実施例では、各RISCコアは、任意の量のメモリを含んでもよい。少なくとも一実施例では、RISCコアは、実施例に応じて複数のプロトコルのうちの任意のものを使用してもよい。少なくとも一実施例では、RISCコアは、リアル・タイム・オペレーティング・システム（「RTOS」：real-time operating system）を実行してもよい。少なくとも一実施例では、RISCコアは、1つ又は複数の集積回路デバイス、特定用途向け集積回路（「ASIC」）、及び／又はメモリ・デバイスを使用して実装されてもよい。たとえば、少なくとも一実施例では、RISCコアは、命令キャッシュ及び／又は密結合RAMを含むことができる。

30

【0129】

少なくとも一実施例では、DMAは、PVAの構成要素がCPU1106とは無関係にシステム・メモリにアクセスできるようにしてもよい。少なくとも一実施例では、DMAは、多次元アドレッシング、及び／又はサーキュラ・アドレッシングを含むがこれらに限定されない、PVAに最適化を提供するために使用される任意の数の特徴をサポートしてもよい。少なくとも一実施例では、DMAは、6つ以上のアドレッシング次元までをサポートしてもよく、これには、限定することなく、ブロック幅、ブロック高さ、ブロック深度、水平ブロック・ステッピング、垂直ブロック・ステッピング、及び／又は深度ステッピングが含まれてもよい。

40

【0130】

少なくとも一実施例では、ベクトル・プロセッサは、コンピュータ・ビジョン・アルゴリズムのためのプログラミングを効率的でフレキシブルに実行するように設計されてもよいプログラム可能なプロセッサとすることができ、信号処理機能を提供する。少なくとも一実施例では、PVAは、PVAコアと、2つのベクトル処理サブシステム・パーティシ

50

ョンを含んでもよい。少なくとも一実施例では、PVAコアは、プロセッサ・サブシステム、DMAエンジン（たとえば2つのDMAエンジン）、及び／又は他の周辺装置を含んでもよい。少なくとも一実施例では、ベクトル処理サブシステムは、PVAの一次処理エンジンとして動作してもよく、ベクトル処理ユニット（「VPU」：vector processing unit）、命令キャッシュ、及び／又はベクトル・メモリ（たとえば、「VMEM」）を含んでもよい。少なくとも一実施例では、VPUは、たとえば単一命令複数データ（「SIMD」：single instruction, multiple data）、超長命令語（「VLIW」：very long instruction word）のデジタル信号プロセッサなどのデジタル信号プロセッサを含んでもよい。少なくとも一実施例では、SIMDとVLIWの組合せによって、スループット及びスピードが改善されてもよい。

10

【0131】

少なくとも一実施例では、ベクトル・プロセッサのそれぞれは、命令キャッシュを含んでもよく、専用のメモリに結合されてもよい。その結果、少なくとも一実施例では、ベクトル・プロセッサのそれぞれは、他のベクトル・プロセッサとは無関係に実行されるように構成されてもよい。少なくとも一実施例では、特定のPVAに含まれるベクトル・プロセッサは、データ並列処理を用いるように構成されてもよい。たとえば、少なくとも一実施例では、単一のPVAに含まれる複数のベクトル・プロセッサは、共通のコンピュータ・ビジョン・アルゴリズムを画像の異なる領域上で実行してもよい。少なくとも一実施例では、特定のPVAに含まれるベクトル・プロセッサは、異なるコンピュータ・ビジョン・アルゴリズムを1つの画像上で同時に実行してもよく、又はさらには、異なるアルゴリズムを連続した画像上で、若しくは画像の部分上で実行してもよい。少なくとも一実施例では、とりわけ、任意の数のPVAがハードウェア加速クラスタに含まれてもよく、任意の数のベクトル・プロセッサが各PVAに含まれてもよい。少なくとも一実施例では、PVAは、システムの全体的な安全性を強化するために、追加のエラー訂正コード（「ECC」：Error Correction Code）メモリを含んでもよい。

20

【0132】

少なくとも一実施例では、アクセラレータ1114は、オン・チップのコンピュータ・ビジョン・ネットワーク、及びスタティック・ランダム・アクセス・メモリ（「SRAM」）を含み、アクセラレータ1114のための高帯域幅、低レイテンシのSRAMを提供してもよい。少なくとも一実施例では、オン・チップ・メモリは、たとえば限定することなく、8つのフィールド設定可能なメモリ・ブロックを含む少なくとも4MBのSRAMを含んでもよく、これは、PVAからもDLAからもアクセス可能であってもよい。少なくとも一実施例では、メモリ・ブロックの各対は、アドバンスド・ペリフェラル・バス（「APB」：advanced peripheral bus）インターフェース、構成回路、コントローラ、及びマルチプレクサを含んでもよい。少なくとも一実施例では、任意のタイプのメモリが使用されてもよい。少なくとも一実施例では、PVA及びDLAは、メモリへの高速アクセスをPVA及びDLAに提供するバックボーンを介して、メモリにアクセスしてもよい。少なくとも一実施例では、バックボーンは、PVA及びDLAを（たとえばAPBを使用して）メモリに相互接続するオン・チップのコンピュータ・ビジョン・ネットワークを含んでもよい。

30

40

【0133】

少なくとも一実施例では、オン・チップのコンピュータ・ビジョン・ネットワークは、何らかの制御信号／アドレス／データを送信する前に、PVAとDLAの両方が準備信号及び有効信号を提供することを判定するインターフェースを含んでもよい。少なくとも一実施例では、インターフェースは、制御信号／アドレス／データを送信するための別々の位相及び別々のチャネル、並びに継続的なデータ転送のためのバースト型通信を提供してもよい。少なくとも一実施例では、インターフェースは、国際標準化機構（「ISO」：Organization for Standardization）26262又は国際電気標準会議（「IEC」：International Electrotechn

50

ical Commission) 61508の規格に準拠してもよいが、他の規格及びプロトコルが使用されてもよい。

【0134】

少なくとも一実施例では、SoC1104のうちの1つ又は複数は、リアル・タイム・レイ・トレーシングのハードウェア・アクセラレータを含んでもよい。少なくとも一実施例では、リアル・タイム・レイ・トレーシングのハードウェア・アクセラレータを使用して、物体の（たとえば世界モデル内での）位置及び範囲が迅速且つ効率的に判定されて、RADAR信号解釈のため、音伝播合成及び／若しくは分析のため、SONARシステムのシミュレーションのため、一般波形の伝播シミュレーションのため、ローカリゼーション及び／若しくは他の機能を目的としたLIDARデータとの比較のため、並びに／又は他の使用法のためのリアル・タイムの可視化シミュレーションが生成されてもよい。

10

【0135】

少なくとも一実施例では、アクセラレータ1114は、自律運転のための多様な使用法を有することができる。少なくとも一実施例では、PVAは、ADAS及び自律車両の主要な処理ステージに使用することができる。少なくとも一実施例では、PVAの性能は、低電力及び低レイテンシの予測可能な処理を必要とするアルゴリズム・ドメインに良好に適合する。言い換えれば、PVAは、低レイテンシ及び低電力の予測可能なラン・タイムを必要とするかもしれない半稠密(semi-dense)又は稠密な規則的計算に対して、データ・セットが小さくても良好に機能する。少なくとも一実施例では、車両1100内など、従来のコンピュータ・ビジョン・アルゴリズムを実行するようにPVAが設計され得るが、これは、それらが、物体検出及び整数数値の演算に有効となり得るからである。

20

【0136】

たとえば、技術の少なくとも一実施例によれば、PVAを使用して、コンピュータ・ステレオ・ビジョンが実行されてもよい。少なくとも一実施例では、いくつかの例においてセミ・グローバル・マッチングに基づくアルゴリズムが使用されてもよいが、これは限定するものではない。少なくとも一実施例では、レベル3～5の自律運転のためのアプリケーションは、動き推定／ステレオ・マッチング（たとえば、動きからの構造化、歩行者認識、車線検出など）をオン・ザ・フライで使用する。少なくとも一実施例では、PVAは、2つの単眼カメラからの入力に対して、コンピュータ・ステレオ・ビジョン機能を実行してもよい。

30

【0137】

少なくとも一実施例では、PVAを使用して、高密度オプティカル・フローが実行されてもよい。たとえば、少なくとも一実施例では、PVAは、未加工のRADARデータを（たとえば4Dの高速フーリエ変換を使用して）処理して、処理済みRADARデータを提供することができる。少なくとも一実施例では、PVAは、飛行時間の深度処理に使用され、たとえば未加工の飛行時間データを処理することにより、処理済みの飛行時間データが提供される。

【0138】

少なくとも一実施例では、たとえば限定することなく、物体検出ごとに信頼性の尺度を出力するニューラル・ネットワークを含む、制御及び運転の安全性を強化するための任意のタイプのネットワークを実行するために、DLAが使用されてもよい。少なくとも一実施例では、信頼性は、他の検出と比較した各検出の確率として、若しくはその相対的な「重み」を提供するものとして表されても、又は解釈されてもよい。少なくとも一実施例では、信頼性尺度によって、どの検出を誤検出ではなく正検出とみなすべきかに関して、システムがさらなる判定を下せるようになる。少なくとも一実施例では、システムは、信頼性に対して閾値を設定し、閾値を超える検出だけを正検出とみなしてもよい。自動緊急ブレーキ（「AEB」：automatic emergency braking）が使用される実施例では、誤検出によって車両は自動的に緊急ブレーキをかけることになり、これは明らかに望ましくない。少なくとも一実施例では、非常に信頼性の高い検出が、A E

40

50

Bのトリガとみなされてもよい。少なくとも一実施例では、DLAは、信頼値を回帰するようにニューラル・ネットワークを実行してもよい。少なくとも一実施例では、ニューラル・ネットワークは、とりわけ境界ボックスの次元、（たとえば別のサブシステムから）取得した地面推定、車両1100の配向と相関しているIMUセンサ1166からの出力、距離、ニューラル・ネットワーク及び／又は他のセンサ（たとえば、LIDARセンサ1164若しくはRADARセンサ1160）から取得した物体の3Dロケーション推定などのパラメータの少なくともいくつかのサブセットを、その入力として取ってもよい。

【0139】

少なくとも一実施例では、SOC1104のうちの1つ又は複数は、データ・ストア1116（たとえばメモリ）を含んでもよい。少なくとも一実施例では、データ・ストア1116は、SOC1104のオン・チップ・メモリであってもよく、このメモリは、GPU1108及び／又はDLA上で実行されるニューラル・ネットワークを記憶してもよい。少なくとも一実施例では、データ・ストア1116の容量は、冗長性及び安全性のためにニューラル・ネットワークの複数のインスタンスを記憶するのに十分なほど大きくてもよい。少なくとも一実施例では、データ・ストア1116は、L2又はL3のキャッシュを備えてもよい。

【0140】

少なくとも一実施例では、SOC1104のうちの1つ又は複数は、任意の数のプロセッサ1110（たとえば、組み込みプロセッサ）を含んでもよい。少なくとも一実施例では、プロセッサ1110は、ブート電力並びに管理機能及び関連するセキュリティ執行に対処するための専用プロセッサ及びサブシステムであってもよいブート及び電力管理プロセッサを含んでもよい。少なくとも一実施例では、ブート及び電力管理プロセッサは、SOC1104のブート・シーケンスの一部であってもよく、ラン・タイム電力管理サービスを提供してもよい。少なくとも一実施例では、ブート電力及び管理プロセッサは、クロックと電圧のプログラミング、システムの低電力状態への移行の支援、SOC1104の熱及び温度センサの管理、並びに／又はSOC1104の電力状態の管理を提供してもよい。少なくとも一実施例では、各温度センサは、その出力周波数が温度に比例するリング発振器として実装されてもよく、SOC1104は、リング発振器を使用して、CPU1106、GPU1108、及び／又はアクセラレータ1114の温度を検出してもよい。少なくとも一実施例では、温度が閾値を超えると判定された場合には、ブート及び電力管理プロセッサは、温度不良ルーチンに入り、SOC1104を低電力状態にし、且つ／又は車両1100を運転手－安全停止モードにしても（たとえば、車両1100を安全停止させる）よい。

【0141】

少なくとも一実施例では、プロセッサ1110はさらに、オーディオ処理エンジンとしての役割を果たすことができる組み込みプロセッサのセットを含んでもよく、これは、多重インターフェースを介した多チャネルのオーディオ、及び幅広くフレキシブルな様々なオーディオI/Oインターフェースのための、完全なハードウェア・サポートを可能にするオーディオ・サブシステムであってもよい。少なくとも一実施例では、オーディオ処理エンジンは、専用RAMのあるデジタル信号プロセッサを有する専用プロセッサ・コアである。

【0142】

少なくとも一実施例では、プロセッサ1110はさらに、低電力センサ管理及び立ち上げのユース・ケースをサポートするのに必要なハードウェア特徴を提供することができる常時オン・プロセッサ・エンジンを含んでもよい。少なくとも一実施例では、常時オン・プロセッサ・エンジンは、限定することなく、プロセッサ・コア、密結合RAM、サポート周辺装置（たとえば、タイマ、及び割込みコントローラ）、様々なI/Oコントローラ周辺装置、及びルーティング論理を含んでもよい。

【0143】

少なくとも一実施例では、プロセッサ1110はさらに安全クラスタ・エンジンを含ん

10

20

30

40

50

でもよく、このエンジンは限定することなく、自動車用途の安全管理に対処するための専用のプロセッサ・サブシステムを含む。少なくとも一実施例では、安全クラスタ・エンジンは、限定することなく、2つ以上のプロセッサ・コア、密結合RAM、サポート周辺装置（たとえば、タイマ、及び割込みコントローラなど）、及び／又はルーティング論理を含んでもよい。安全モードでは、少なくとも一実施例においてロックステップ・モードで2つ以上のコアが動作し、これらの動作間で何らかの差を検出するための比較論理を有する単一コアとして機能してもよい。少なくとも一実施例では、プロセッサ1110はさらにリアル・タイム・カメラ・エンジンを含んでもよく、このエンジンは限定することなく、リアル・タイムのカメラ管理に対処するための専用のプロセッサ・サブシステムを含んでもよい。少なくとも一実施例では、プロセッサ1110はさらに、高ダイナミック・レンジの信号プロセッサを含んでもよく、この信号プロセッサは、カメラ処理パイプラインの一部であるハードウェア・エンジンである画像信号プロセッサを限定することなく含んでもよい。

【0144】

少なくとも一実施例では、プロセッサ1110は、ビデオ画像合成器を含んでもよく、この合成器は、再生装置のウィンドウに最終画像を生成するのにビデオ再生アプリケーションが必要とするビデオ後処理機能を実装する（たとえばマイクロプロセッサに実装された）処理ブロックであってもよい。少なくとも一実施例では、ビデオ画像合成器は、広角カメラ1170、周囲カメラ1174、及び／又はキャビン内監視カメラ・センサに対して、レンズゆがみ補正を実行してもよい。少なくとも一実施例では、キャビン内監視カメラ・センサは、好ましくは、キャビン内のイベントを識別し、それに適宜応答するように構成された、SoC1104の別のインスタンスで実行されているニューラル・ネットワークによって監視される。少なくとも一実施例では、キャビン内システムは、セルラー・サービスをアクティブ化し、電話をかけたり、電子メールを書いたり、車両の行き先を変更したり、車両のインフォテイメント・システム及び設定をアクティブ化又は変更したり、音声作動式のウェブ・サーフィンを提供したりするために、限定することなく読唇を実行してもよい。少なくとも一実施例では、ある一定の機能は、車両が自律モードで動作しているときにドライバにとって利用可能になり、それ以外のときには使用不可になる。

【0145】

少なくとも一実施例では、ビデオ画像合成器は、空間と時間の両方のノイズ低減のための拡張された時間的ノイズ低減を含んでもよい。たとえば、少なくとも一実施例では、ビデオで動きが生じる場合には、ノイズ低減が空間情報に適切に重み付けして、隣接するフレームによって提供される情報の重みを軽くする。少なくとも一実施例では、画像又は画像の一部分が動きを含まない場合には、ビデオ画像合成器により実行される時間的ノイズ低減は、前の画像からの情報を使用して、現在の画像のノイズを低減してもよい。

【0146】

少なくとも一実施例では、ビデオ画像合成器はまた、入力されたステレオ・レンズ・フレームに対してステレオ平行化を実行するように構成されてもよい。少なくとも一実施例では、ビデオ画像合成器はさらに、オペレーティング・システムのデスクトップが使用中のときに、ユーザ・インターフェースを合成するために使用されてもよく、GPU1108は、新規の表面を継続的にレンダリングする必要がなくなる。少なくとも一実施例では、GPU1108の電源が入れられ、アクティブで3Dレンダリングを行っているとき、性能及び応答性を向上させるために、ビデオ画像合成器を使用してGPU1108をオフロードしてもよい。

【0147】

少なくとも一実施例では、SoC1104のうちの1つ又は複数のSoCはさらに、ビデオ及びカメラからの入力を受信するためのモバイル・インダストリ・プロセッサ・インターフェース（「MIPi」：mobile industry processor interface）のカメラ直列インターフェース、高速インターフェース、並びに／又はカメラ及び関連ピクセルの入力機能に使用されてもよいビデオ入力ブロックを含んでも

10

20

30

40

50

よい。少なくとも一実施例では、S o C 1 1 0 4のうちの1つ又は複数のS o Cはさらに、入力／出力コントローラを含んでもよく、このコントローラはソフトウェアによって制御されてもよく、特定の役割に縛られていないI／O信号を受信するために使用されてもよい。

【0148】

少なくとも一実施例では、S o C 1 1 0 4のうちの1つ又は複数のS o Cはさらに、周辺装置、オーディオ・エンコーダ／デコーダ（「コーデック」）、電力管理、及び／又は他の装置との通信を可能にするための幅広い周辺装置インターフェースを含んでもよい。少なくとも一実施例では、S o C 1 1 0 4は、（たとえば、ギガビット・マルチメディア・シリアル・リンク及びイーサネット（登録商標）チャネルを介して接続された）カメラからのデータ、センサ（たとえば、イーサネット（登録商標）チャネルを介して接続されてもよいL I D A Rセンサ1164、R A D A Rセンサ1160など）からのデータ、バス1102からのデータ（たとえば、車両1100のスピード、ハンドル位置など）、（たとえば、イーサネット（登録商標）バス又はC A Nバスを介して接続された）G N S Sセンサ1158からのデータなどを処理するために使用されてもよい。少なくとも一実施例では、S o C 1 1 0 4のうちの1つ又は複数のS o Cはさらに、専用の高性能大容量ストレージ・コントローラを含んでもよく、このコントローラは独自のD M Aエンジンを含んでもよく、ルーチンのデータ管理タスクからC P U 1 1 0 6を解放するために使用されてもよい。

【0149】

少なくとも一実施例では、S o C 1 1 0 4は、自動化レベル3～5に及ぶフレキシブルなアーキテクチャを有するエンドツーエンドのプラットフォームであってもよく、それにより、多様性及び冗長性を得るためにコンピュータ・ビジョン及びA D A S技法を活用し効率的に利用する包括的な機能的安全性アーキテクチャが提供され、フレキシブルで、信頼性の高い運転ソフトウェア・スタックが、深層学習ツールとともに提供される。少なくとも一実施例では、S o C 1 1 0 4は、従来のシステムより高速で、信頼性が高く、さらにはエネルギー効率及び空間効率が高い。たとえば、少なくとも一実施例では、アクセラレータ1114は、C P U 1 1 0 6、G P U 1 1 0 8、及びデータ・ストア1116と組み合わせられると、レベル3～5の自律車両のための高速で効率的なプラットフォームを実現することができる。

【0150】

少なくとも一実施例では、コンピュータ・ビジョン・アルゴリズムはC P U上で実行されてもよく、このアルゴリズムは、Cなどの高レベル・プログラミング言語を使用して構成されて、多様な視覚データにわたって多様な処理アルゴリズムを実行してもよい。しかし、少なくとも一実施例では、C P Uは、多くのコンピュータ・ビジョン・アプリケーションの性能要件、たとえば実行時間及び電力消費に関する要件などを満足できないことが多い。少なくとも一実施例では、多くのC P Uは、車両内のA D A Sアプリケーション及び現実的なレベル3～5の自律車両において使用される複雑な物体検出アルゴリズムを、リアル・タイムで実行することができない。

【0151】

本明細書に記載の実施例は、複数のニューラル・ネットワークを同時に且つ／又は順番に実行できるようにし、結果を組み合わせ、レベル3～5の自律運転機能を可能にすることができる。たとえば、少なくとも一実施例では、D L A又は個別G P U（たとえば、G P U 1 1 2 0）上で実行しているC N Nは、テキスト及び単語認識を含んでもよく、ニューラル・ネットワークがそれについて特に訓練されてこなかった標識を含む交通標識を読み、理解できるようにする。少なくとも一実施例では、D L Aはさらに、標識を識別し、解釈し、標識の意味的理解を提供することができ、その意味的理解を、C P Uコンプレックス上で実行されている経路計画モジュールに渡すことができるニューラル・ネットワークを含んでもよい。

【0152】

少なくとも一実施例では、レベル3、4、又は5の運転に関して、複数のニューラル・

10

20

30

40

50

ネットワークが同時に実行されてもよい。たとえば、少なくとも一実施例では、電光と併せて「注意：点滅時は凍結状態」と示される警告標識は、いくつかのニューラル・ネットワークによって別々に解釈されても、集合的に解釈されてもよい。少なくとも一実施例では、こうした警告標識自体は、第1の導入済みニューラル・ネットワーク（たとえば、訓練されてきたニューラル・ネットワーク）によって交通標識として識別されてもよく、「点滅時は凍結状態」という文字は、第2の導入済みニューラル・ネットワークによって解釈されてもよく、点滅光が検出された場合には、このニューラル・ネットワークが、凍結状態が存在することを車両の（好ましくはCPUコンプレックス上で実行している）経路計画ソフトウェアに通知する。少なくとも一実施例では、点滅光は、第3の導入済みニューラル・ネットワークを複数のフレームにわたって動作させることによって識別されてもよく、点滅光の存在（又は存在しないこと）が、車両の経路計画ソフトウェアに通知される。少なくとも一実施例では、3つすべてのニューラル・ネットワークが、DLA内及び／又はGPU1108上などで同時に実行されてもよい。

10

【0153】

少なくとも一実施例では、顔認識及び車両所有者識別のためのCNNは、カメラ・センサからのデータを使用して、車両1100の承認済みのドライバ及び／又は所有者の存在を識別してもよい。少なくとも一実施例では、常時オンのセンサ処理エンジンを使用して、所有者がドライバ用ドアに近づいてきたときに車両を解錠し、ライトを点灯させ、所有者がこうした車両から離れるときには、セキュリティ・モードでこうした車両を使用不可にしてもよい。こうして、SOC1104は、窃盗及び／又は自動車乗っ取りに対するセキュリティを実現する。

20

【0154】

少なくとも一実施例では、緊急車両の検出及び識別のためのCNNは、マイクロフォン1196からのデータを使用して、緊急車両のサイレンを検出及び識別してもよい。少なくとも一実施例では、SOC1104は、環境及び市街地の音を分類するとともに、視覚データを分類するためにCNNを使用する。少なくとも一実施例では、DLA上で実行されるCNNは、緊急車両が近づいてくる相対的なスピードを（たとえばドップラー効果を使用することによって）識別するように訓練される。少なくとも一実施例では、CNNはまた、GNSSセンサ1158によって識別される、車両が稼働している地域に特有の緊急車両を識別するように訓練されてもよい。少なくとも一実施例では、欧州で稼働している場合には、CNNは欧州のサイレンを検出しようとし、北米の場合には、北米のサイレンだけを識別しようとする。少なくとも一実施例では、緊急車両が検出されると、緊急車両安全ルーチンを実行するための制御プログラムを使用して、車両の速度を落とし、道路脇に寄せ、車両を停止させ、且つ／又は緊急車両が通過するまで、超音波センサ1162を併用して車両をアイドリングにしてもよい。

30

【0155】

少なくとも一実施例では、車両1100はCPU1118（たとえば、個別CPU又はdCPU）を含んでもよく、このCPUは高速相互接続（たとえば、PCIe）を介してSOC1104に結合されてもよい。少なくとも一実施例では、CPU1118は、たとえばX86プロセッサを含んでもよい。CPU1118は、たとえば、ADASセンサとSOC1104の間で潜在的に不整合な結果を調停すること、並びに／又はコントローラ1136及び／若しくはチップ上のインフォテイメント・システム（「インフォテイメントSOC」）1130の状態及び健全性を監視することを含め、様々な機能のうちの任意の機能を実行するために使用されてもよい。

40

【0156】

少なくとも一実施例では、車両1100はGPU1120（たとえば、個別GPU又はdGPU）を含んでもよく、このGPUは高速相互接続（たとえば、NVIDIAのNVLINKチャネル）を介してSOC1104に結合されてもよい。少なくとも一実施例では、GPU1120は、冗長な及び／又は異なるニューラル・ネットワークを実行することなどによって、追加の人工知能機能を提供してもよく、車両1100のセンサからの入

50

力（たとえば、センサ・データ）に少なくとも部分的に基づき、ニューラル・ネットワークを訓練及び／又は更新するために使用されてもよい。

【0157】

少なくとも一実施例では、車両1100はさらに、ネットワーク・インターフェース1124を含んでもよく、このインターフェースは限定することなく、ワイヤレス・アンテナ1126（たとえば、セルラー・アンテナ、Bluetoothアンテナなど、異なる通信プロトコル向けの1つ又は複数のワイヤレス・アンテナ）を含んでもよい。少なくとも一実施例では、他の車両、及び／又はコンピューティング・デバイス（たとえば、乗員のクライアント・デバイス）とのインターネット・クラウド・サービス（たとえば、サーバ及び／又は他のネットワーク・デバイス）へのワイヤレス接続を可能にするために、ネットワーク・インターフェース1124が使用されてもよい。少なくとも一実施例では、他の車両と通信するために、車両110と他の車両との間に直接リンクが確立されてもよく、且つ／又は（たとえば、ネットワークにわたって、且つインターネットを介して）間接リンクが確立されてもよい。少なくとも一実施例では、直接リンクは、車車間通信リンクを使用して提供されてもよい。少なくとも一実施例では、車車間通信リンクは、車両1100の近傍の車両（たとえば、車両1100の前方、側方、及び／又は後方の車両）についての情報を車両1100に提供してもよい。少なくとも一実施例では、こうした前述した機能は、車両1100の協調型アダプティブ・クルーズ・コントロール機能の一部であってよい。

【0158】

少なくとも一実施例では、ネットワーク・インターフェース1124は、変調及び復調の機能を提供し、コントローラ1136がワイヤレス・ネットワークを介して通信できるようにするSoCを含んでもよい。少なくとも一実施例では、ネットワーク・インターフェース1124は、ベースバンドから無線周波数へのアップ・コンバージョン、及び無線周波数からベースバンドへのダウン・コンバージョンのための無線周波数フロント・エンドを含んでもよい。少なくとも一実施例では、周波数変換は、任意の技術的に実行可能なやり方で実行されてもよい。たとえば、周波数変換は、よく知られたプロセスにより、且つ／又はスーパー・ヘテロダイン・プロセスを使用して実行することができる。少なくとも一実施例では、無線周波数フロント・エンド機能は、別個のチップによって提供されてもよい。少なくとも一実施例では、ネットワーク・インターフェースは、LTE、WCDMA、UMTS、GSM、CDMA2000、Bluetooth、Bluetooth LE、Wi-Fi、Z-Wave、ZigBee、LoRaWAN、及び／又は他のワイヤレス・プロトコルを介して通信するためのワイヤレス機能を含んでもよい。

【0159】

少なくとも一実施例では、車両1100はさらにデータ・ストア1128を含んでもよく、このデータ・ストアは限定することなく、オフ・チップ（たとえばSoC1104上にはない）ストレージを含んでもよい。少なくとも一実施例では、データ・ストア1128は、RAM、SRAM、ダイナミック・ランダム・アクセス・メモリ（「DRAM」）、ビデオ・ランダム・アクセス・メモリ（「VRAM」：video random-access memory）、フラッシュ・メモリ、ハード・ディスク、並びに／又は少なくとも1ビットのデータを記憶することができる他の構成要素及び／若しくはデバイスを含む1つ若しくは複数のストレージ要素を、限定することなく含んでもよい。

【0160】

少なくとも一実施例では、車両1100はさらに、マッピング、知覚、占有グリッド生成、及び／又は経路計画の機能を支援するためのGNSSセンサ1158（たとえば、GPS及び／又は補助GPSセンサ）を含んでもよい。少なくとも一実施例では、イーサネット（登録商標）からシリアル（たとえばRS-232）へのブリッジを有するUSBコネクタを使用するGPSをたとえば限定することなく含む任意の数のGNSSセンサ1158が使用されてもよい。

【0161】

少なくとも一実施例では、車両 1100 はさらに、RADAR センサ 1160 を含んでもよい。少なくとも一実施例では、RADAR センサ 1160 は、暗闇及び／又は厳しい気象条件の中でも、長距離の車両検出を行うために車両 1100 によって使用されてもよい。少なくとも一実施例では、RADAR の機能的安全性レベルは、ASIL B であってもよい。少なくとも一実施例では、RADAR センサ 1160 は、制御のために（たとえば、RADAR センサ 1160 によって生成されたデータを送信するために）、また物体追跡データにアクセスするために、CAN バス及び／又はバス 1102 を使用してもよく、いくつかの例では、未加工データにアクセスするためにイーサネット（登録商標）チャンネルにアクセスできる。少なくとも一実施例では、多様なタイプの RADAR センサが使用されてもよい。たとえば限定することなく、RADAR センサ 1160 は、前方、後方、及び側方の RADAR 使用に好適であってもよい。少なくとも一実施例では、RADAR センサ 1160 のうちの 1 つ又は複数のセンサは、パルス・ドップラ RADAR センサである。

【0162】

少なくとも一実施例では、RADAR センサ 1160 は、狭視野の長距離、広視野の短距離、側面を網羅する短距離など、異なる構成を含んでもよい。少なくとも一実施例では、長距離 RADAR は、アダプティブ・クルーズ・コントロール機能のために使用されてもよい。少なくとも一実施例では、長距離 RADAR システムは、2 つ以上の独立した走査によって実現される 250 m（メートル）の範囲内などの広視野を提供してもよい。少なくとも一実施例では、RADAR センサ 1160 は、静的物体と移動している物体とを区別しやすくしてもよく、緊急ブレーキ支援及び前方衝突警告を行うために ADAS システム 1138 によって使用されてもよい。少なくとも一実施例では、長距離 RADAR システムに含まれるセンサ 1160 は、複数の（たとえば 6 つ以上の）固定 RADAR アンテナ、並びに高速 CAN 及び Flex Ray インターフェースを有するモノスタティックのマルチモード RADAR を、限定することなく含んでもよい。少なくとも一実施例では、6 つのアンテナがある場合、中央の 4 つのアンテナは、隣接した車線からの干渉が最小の状態、より高速で車両 1100 の周囲を記録するように設計された集中したビーム・パターンを生成してもよい。少なくとも一実施例では、他の 2 つのアンテナは、視野を拡張してもよく、車両 1100 の車線に入る又はそこから出る車両を迅速に検出するのを可能にする。

【0163】

少なくとも一実施例では、中距離 RADAR システムは、一例として最大 160 m（前方）、又は 80 m（後方）の範囲、及び最大 42 度（前方）、又は 150 度（後方）の視野を含んでもよい。少なくとも一実施例では、短距離 RADAR システムは、限定することなく、後方バンパの両端部に設置されるように設計された任意の数の RADAR センサ 1160 を含んでもよい。後方バンパの両端部に設置されたとき、少なくとも一実施例では、RADAR センサ・システムは、後方向及び車両隣の死角を常に監視する 2 本のビームを生成してもよい。少なくとも一実施例では、短距離 RADAR システムは、死角検出及び／又は車線変更支援を行うために、ADAS システム 1138 において使用されてもよい。

【0164】

少なくとも一実施例では、車両 1100 はさらに、超音波センサ 1162 を含んでもよい。少なくとも一実施例では、超音波センサ 1162 は、車両 1100 の前方、後方、及び／又は側方位置に配置されてもよく、駐車支援のため、且つ／又は占有グリッドを生成し更新するために使用されてもよい。少なくとも一実施例では、多様な超音波センサ 1162 が使用されてもよく、異なる検出範囲（たとえば、2.5 m、4 m）には異なる超音波センサ 1162 が使用されてもよい。少なくとも一実施例では、超音波センサ 1162 は、機能的安全性レベル ASIL B で動作してもよい。

【0165】

少なくとも一実施例では、車両 1100 は、LIDAR センサ 1164 を含んでもよい

。少なくとも一実施例では、L I D A Rセンサ1164は、物体及び歩行者の検出、緊急ブレーキ、衝突回避、及び／又は他の機能のために使用されてもよい。少なくとも一実施例では、L I D A Rセンサ1164は、機能的安全性レベルA S I L Bで動作してもよい。少なくとも一実施例では、車両1100は、複数のL I D A Rセンサ1164（たとえば、2つ、4つ、6つなど）を含んでもよく、これらのセンサは、（たとえばデータをギガビット・イーサネット（登録商標）・スイッチに提供するために）イーサネット（登録商標）チャンネルを使用してもよい。

【0166】

少なくとも一実施例では、L I D A Rセンサ1164は、360度の視野について、物体及びそれらの距離のリストを提供可能であってもよい。少なくとも一実施例では、市販のL I D A Rセンサ1164は、たとえば宣伝された範囲がおおよそ100mであり、精度が2cm～3cmであり、100Mbpsのイーサネット（登録商標）接続をサポートしてもよい。少なくとも一実施例では、1つ又は複数の非突出型L I D A Rセンサが使用されてもよい。こうした実施例では、L I D A Rセンサ1164は、車両1100の前方、後方、側方、及び／又は角位置に組み込むことができる小さいデバイスを含んでもよい。少なくとも一実施例では、こうした実施例のL I D A Rセンサ1164は、最大120度の水平視野、及び35度の垂直視野を、低反射性の物体に対しても200mの範囲で提供してもよい。少なくとも一実施例では、前方に取り付けられたL I D A Rセンサ1164は、45度～135度の水平視野をもたらすように構成されてもよい。

【0167】

少なくとも一実施例では、3DフラッシュL I D A RなどのL I D A R技術も使用されてよい。少なくとも一実施例では、3DフラッシュL I D A Rは、レーザのフラッシュを送信源として使用して、車両1100の周囲を最大でおおよそ200mまで照射する。少なくとも一実施例では、フラッシュL I D A Rユニットは、限定することなくレセプタを含み、このレセプタは、レーザ・パルスの通過時間及び各ピクセルにおける反射光を記録し、それらは、車両1100から物体までの範囲に対応する。少なくとも一実施例では、フラッシュL I D A Rによって、非常に正確でゆがみのない周囲画像が、レーザのフラッシュごとに生成できるようになる。少なくとも一実施例では、4つのフラッシュL I D A Rが、車両1100の各側面に1つ導入されてもよい。少なくとも一実施例では、3DフラッシュL I D A Rシステムは、ファン以外に可動部品のない半導体3D凝視アレイ（s t a r i n g a r r a y）のL I D A Rカメラ（たとえば、非走査型L I D A Rデバイス）を、限定することなく含む。少なくとも一実施例では、フラッシュL I D A Rデバイスは、フレーム当たり5ナノ秒のクラスI（目に安全な）レーザ・パルスを使用してもよく、3D範囲の点群及び位置同期された（c o - r e g i s t e r e d）強度データとして反射レーザ光を捕捉してもよい。

【0168】

少なくとも一実施例では、車両1100はさらにIMUセンサ1166を含んでもよい。少なくとも一実施例では、IMUセンサ1166は、車両1100の後方車軸の中央に位置付けられてもよい。少なくとも一実施例では、IMUセンサ1166は、たとえば限定することなく、加速度計、磁力計、ジャイロスコープ、磁気コンパス、複数の磁気コンパス及び／又は他のタイプのセンサを含んでもよい。6軸の用途など少なくとも一実施例では、IMUセンサ1166は限定することなく、加速度計及びジャイロスコープを含んでもよい。9軸の用途など少なくとも一実施例では、IMUセンサ1166は限定することなく、加速度計、ジャイロスコープ、及び磁力計を含んでもよい。

【0169】

少なくとも一実施例では、IMUセンサ1166は、微小電気機械システム（「MEMS」：m i c r o - e l e c t r o - m e c h a n i c a l s y s t e m s）慣性センサ、高感度GPS受信機、及び先進のK a l m a nフィルタリング・アルゴリズムを組み合わせ、位置、速度、及び姿勢の推定値を提供する小型の高性能GPS補強型慣性航法システム（「GPS／INS」：G P S - A i d e d I n e r t i a l N a v i g a t

10

20

30

40

50

ion System) として実装されてもよい。少なくとも一実施例では、IMUセンサ1166により、車両1100は、速度変化を直接観察しそれをGPSからIMUセンサ1166に相関させることによって、磁気センサからの入力を必要とせずに車両1100の方位を推定できるようになる。少なくとも一実施例では、IMUセンサ1166及びGNSSセンサ1158は、単一の統合ユニットに組み合わされてもよい。

【0170】

少なくとも一実施例では、車両1100は、車両1100の中及び／又はその周りに設置されたマイクロフォン1196を含んでもよい。少なくとも一実施例では、マイクロフォン1196は、とりわけ緊急車両の検出及び識別のために使用されてもよい。

【0171】

少なくとも一実施例では、車両1100はさらに、ステレオ・カメラ1168、広角カメラ1170、赤外線カメラ1172、周囲カメラ1174、長距離カメラ1198、中距離カメラ1176、及び／又は他のカメラ・タイプを含む任意の数のカメラ・タイプを含んでもよい。少なくとも一実施例では、カメラは、車両1100の全周囲の周りで画像データを捕捉するために使用されてもよい。少なくとも一実施例では、どのタイプのカメラが使用されるかは、車両1100に応じて異なる。少なくとも一実施例では、車両1100の周りで必要な被写域を提供するために、カメラ・タイプの任意の組合せが使用されてもよい。少なくとも一実施例では、導入されるカメラの数は、実施例に応じて異なってもよい。たとえば、少なくとも一実施例では、車両1100は6台のカメラ、7台のカメラ、10台のカメラ、12台のカメラ、又は別の数のカメラを含むことができる。少なくとも一実施例では、カメラは、一例として限定することなく、ギガビット・マルチメディア・シリアル・リンク(「GMSL」: Gigabit Multimedia Serial Link)及び／又はギガビット・イーサネット(登録商標)通信をサポートしてもよい。少なくとも一実施例では、各カメラは、図11A及び図11Bに関して本明細書でさらに詳細に上で説明されているようであり得る。

【0172】

少なくとも一実施例では、車両1100はさらに、振動センサ1142を含んでもよい。少なくとも一実施例では、振動センサ1142は、車軸など、車両1100の構成要素の振動を測定してもよい。たとえば、少なくとも一実施例では、振動の変化は、路面の変化を示すことがある。少なくとも一実施例では、2つ以上の振動センサ1142が使用される場合には、路面の摩擦又はすべり量を判定するために振動の差が使用されてもよい(たとえば、動力により駆動される車軸と自由回転する車軸との間に振動差がある場合)。

【0173】

少なくとも一実施例では、車両1100は、ADASシステム1138を含んでもよい。少なくとも一実施例では、ADASシステム1138は、限定することなく、いくつかの例においてSoCを含んでもよい。少なくとも一実施例では、ADASシステム1138は、限定することなく、任意の数及び任意の組合せの、自律／アダプティブ／自動のクルーズ・コントロール(「ACC」: autonomous/adaptive/automatic cruise control)システム、協調型アダプティブ・クルーズ・コントロール(「CACC」: cooperative adaptive cruise control)システム、正面衝突警告(「FCW」: forward crash warning)システム、自動緊急ブレーキ(「AEB」: automatic emergency braking)システム、車線逸脱警告(「LDW」: lane departure warning)システム、車線維持支援(「LKA」: lane keep assist)システム、死角警告(「BSW」: blind spot warning)システム、後方クロス・トラフィック警告(「RCTW」: rear cross-traffic warning)システム、衝突警告(「CW」: collision warning)システム、車線センタリング(「LC」: lane centering)システム、並びに／又は他のシステム、特徴、及び／若しくは機能を含んでもよい。

【0174】

少なくとも一実施例では、ACCシステムは、RADARセンサ1160、LIDARセンサ1164、及び／又は任意の数のカメラを使用してもよい。少なくとも一実施例では、ACCシステムは、縦方向ACCシステム及び／又は横方向ACCシステムを含んでもよい。少なくとも一実施例では、縦方向ACCシステムは、車両1100の直前の別の車両までの距離を監視及び制御し、車両1100のスピードを自動的に調節して、前の車両からの安全な距離を維持する。少なくとも一実施例では、横方向ACCシステムは、距離の維持を実行し、必要なときに車線変更するよう車両1100に通知する。少なくとも一実施例では、横方向ACCは、LC及びCWなどの他のADAS用途に関係する。

【0175】

少なくとも一実施例では、CACCシステムは、他の車両からの情報を使用し、この情報は、ワイヤレス・リンクにより、又は間接的にネットワーク接続を介して（たとえばインターネットを介して）、他の車両からネットワーク・インターフェース1124及び／又はワイヤレス・アンテナ1126により受信されてもよい。少なくとも一実施例では、車車間（「V2V」：vehicle-to-vehicle）通信リンクによって直接リンクが提供されてもよく、一方インフラストラクチャ車間（「I2V」：infrastructure-to-vehicle）通信リンクによって間接リンクが提供されてもよい。一般に、V2V通信は、すぐ前の先行車両（たとえば、車両1100のすぐ前で同じ車線にいる車両）についての情報を提供し、I2V通信は、さらにその前の交通についての情報を提供する。少なくとも一実施例では、CACCシステムは、I2VとV2Vの情報源のいずれか又は両方を含んでもよい。少なくとも一実施例では、車両1100の前の車両についての情報があれば、CACCシステムは信頼性をさらに高めることができ、交通の流れをより円滑にし、路上での渋滞を低減できる可能性を有する。

【0176】

少なくとも一実施例では、FCWシステムは、危険物に対してドライバに忠告するように設計され、それによりこうしたドライバは修正措置を取ることができる。少なくとも一実施例では、FCWシステムは正面カメラ及び／又はRADARセンサ1160を使用し、これらは、ディスプレイ、スピーカ、及び／又は振動構成要素などのドライバへのフィードバックを提供するように電氣的に結合されている専用のプロセッサ、DSP、FPGA、及び／又はASICに結合されている。少なくとも一実施例では、FCWシステムは、音、視覚的警告、振動、及び／又はクイック・ブレーキ・パルスなどの形で警告を提供してもよい。

【0177】

少なくとも一実施例では、AEBシステムは、別の車両又は他の物体との差し迫った正面衝突を検出し、指定された時間内又は距離パラメータ内にドライバが修正措置を取らない場合には、自動でブレーキをかけてもよい。少なくとも一実施例では、AEBシステムは、専用のプロセッサ、DSP、FPGA、及び／又はASICに結合された正面カメラ及び／又はRADARセンサ1160を使用してもよい。少なくとも一実施例では、AEBシステムが危険物を検出したとき、AEBシステムは通常、修正措置を取って衝突を避けるよう最初にドライバに忠告し、ドライバが修正措置を取らない場合には、AEBシステムは、予測される衝突を防ぐ又は少なくともその衝撃を軽減するために自動的にブレーキをかけてもよい。少なくとも一実施例では、AEBシステムは、ダイナミック・ブレーキ・サポート及び／又は衝突直前ブレーキなどの技法を含んでもよい。

【0178】

少なくとも一実施例では、LDWシステムは、車両1100が車線の目印に交差したときにドライバに忠告するために、ハンドル又は座席の振動など、視覚的、聴覚的、及び／又は触覚的な警告を提供する。少なくとも一実施例では、ドライバが方向指示器を作動させることなどによって意図的な車線逸脱を示す場合には、LDWシステムは作動しない。少なくとも一実施例では、LDWシステムは、正面カメラを使用してもよく、これは、ディスプレイ、スピーカ、及び／又は振動構成要素などのドライバへのフィードバックを提供するように電氣的に結合することができる専用のプロセッサ、DSP、FPGA、及び

／又はA S I Cに結合されている。少なくとも一実施例では、L K Aシステムは、L D Wシステムの変形形態である。少なくとも一実施例では、L K Aシステムは、車両1100が車両1100の車線からはみ出し始めた場合に、車両1100を修正するように操縦入力又はブレーキ制御を提供する。

【0179】

少なくとも一実施例では、B S Wシステムは、自動車の死角にある車両を検出し、ドライバに警告する。少なくとも一実施例では、B S Wシステムは、視覚的、聴覚的、及び／又は触覚的なアラートを提供して、合流又は車線変更が安全ではないことを示してもよい。少なくとも一実施例では、B S Wシステムは、ドライバが方向指示器を使用したときに追加の警告を提供してもよい。少なくとも一実施例では、B S Wシステムは、専用のプロセッサ、D S P、F P G A、及び／又はA S I Cに結合された背面カメラ及び／又はR A D A Rセンサ1160を使用してもよく、これらの専用のプロセッサ、D S P、F P G A、及び／又はA S I Cは、ディスプレイ、スピーカ、及び／又は振動構成要素などのドライバへのフィードバックに電氣的に結合されている。

【0180】

少なくとも一実施例では、R C T Wシステムは、車両1100の後退時に、後方カメラの範囲外に物体が検出されたときに、視覚的、聴覚的、及び／又は触覚的な通知を提供してもよい。少なくとも一実施例では、R C T Wシステムは、衝突を回避するために確実に車両ブレーキがかけられるように、A E Bシステムを含む。少なくとも一実施例では、R C T Wシステムは、1つ又は複数の背面R A D A Rセンサ1160を使用してもよく、これはディスプレイ、スピーカ、及び／又は振動構成要素などのドライバへのフィードバックを提供するように電氣的に結合された専用のプロセッサ、D S P、F P G A、及び／又はA S I Cに結合されている。

【0181】

少なくとも一実施例では、従来のA D A Sシステムは、誤検出結果を出しがちなことがあり、これはドライバにとっては迷惑で気が散るものであり得るが、通常は大したことはない。なぜなら、従来のA D A Sシステムは、ドライバに忠告し、安全を要する状態が本当に存在し、それに適宜対応するかどうかを、ドライバが判断できるようにするからである。少なくとも一実施例では、結果が矛盾する場合、一次コンピュータ（たとえば、コントローラ1136の第1のコントローラ）からの結果に従うか、又は二次コンピュータ（たとえば、コントローラ1136の第2のコントローラ）からの結果に従うかどうかを、車両1100自体が判断する。たとえば、少なくとも一実施例では、A D A Sシステム1138は、バックアップ・コンピュータの合理性モジュールに知覚情報を抵抗するための、バックアップ及び／又は二次コンピュータであってもよい。少なくとも一実施例では、バックアップ・コンピュータの合理性モニタが、ハードウェア構成要素上の冗長性の多様なソフトウェアを実行して、知覚の誤り及び動的な運転タスクを検出してもよい。少なくとも一実施例では、A D A Sシステム1138からの出力は、監視M C Uに提供されてもよい。少なくとも一実施例では、一次コンピュータからの出力と二次コンピュータからの出力とが矛盾する場合には、監視M C Uが、安全な動作を確保するために矛盾をどのように調和させるかを判定する。

【0182】

少なくとも一実施例では、一次コンピュータは、一次コンピュータの選択した結果の信頼性を示す信頼性スコアを、監視M C Uに提供するように構成されてもよい。少なくとも一実施例では、信頼性スコアが閾値を超える場合には、二次コンピュータが矛盾する又は一貫性のない結果を提供しているかどうかに関わらず、監視M C Uは一次コンピュータの指示に従ってもよい。少なくとも一実施例では、信頼性スコアが閾値を満足せず、一次コンピュータと二次コンピュータが異なる結果（たとえば、矛盾）を示す場合には、監視M C Uは、コンピュータ同士を調停して、適切な結果を判定してもよい。

【0183】

少なくとも一実施例では、二次コンピュータが誤アラームを提供する条件を、一次コン

コンピュータからの出力と二次コンピュータからの出力とに少なくとも部分的に基づき判定するように訓練及び構成されたニューラル・ネットワークを、監視MCUが実行するように構成されてもよい。少なくとも一実施例では、監視MCUのニューラル・ネットワークは、二次コンピュータの出力が信用されてもよいときと、信用できないときとを学習してもよい。たとえば、少なくとも一実施例では、二次コンピュータがRADARベースのFCWシステムである場合、監視MCUのニューラル・ネットワークは、アラームをトリガする排水溝の格子又はマンホール・カバーなど、実際には危険物ではない金属物体をFCWシステムが識別するときを学習してもよい。少なくとも一実施例では、二次コンピュータがカメラ・ベースのLDWシステムである場合、自転車や歩行者が存在し、車線逸脱が実際には最も安全な操作であるときに、監視MCUのニューラル・ネットワークはLDWを無効にするように学習してもよい。少なくとも一実施例では、監視MCUは、ニューラル・ネットワークに関連するメモリとともに実行するのに好適なDLA又はGPUのうちの少なくとも1つを含んでもよい。少なくとも一実施例では、監視MCUは、SoC1104の構成要素を備えても、且つ／又はその構成要素として含まれてもよい。

【0184】

少なくとも一実施例では、ADASシステム1138は、コンピュータ・ビジョンの従来のルールを使用してADAS機能を実行する二次コンピュータを含んでもよい。少なくとも一実施例では、二次コンピュータは、従来のコンピュータ・ビジョン・ルール（if-thenルール）を使用してもよく、ニューラル・ネットワークが監視MCUに存在することによって、信頼性、安全性、及び性能が向上してもよい。たとえば、少なくとも一実施例では、多様な実装及び意図的な非同一性により、特にソフトウェア（又はソフトウェアとハードウェアのインターフェース）の機能によって生じる誤りに対し、システム全体の誤り耐性が高まる。たとえば、少なくとも一実施例では、一次コンピュータ上で実行中のソフトウェアにバグ又はエラーがあり、二次コンピュータ上で実行中の非同一のソフトウェア・コードが、全体的に一貫性のある結果を提供する場合には、監視MCUは、全体的な結果が正しく、一次コンピュータ上のソフトウェア又はハードウェアのバグが重大なエラーを引き起こしていないという、より高い信頼性を有してもよい。

【0185】

少なくとも一実施例では、ADASシステム1138の出力は、一次コンピュータの知覚ブロック、及び／又は一次コンピュータの動的運転タスクブロックに供給されてもよい。たとえば、少なくとも一実施例では、ADASシステム1138が、直前の物体に起因して正面衝突警告を示している場合には、知覚ブロックは、物体を識別するときにこの情報を使用してもよい。少なくとも一実施例では、二次コンピュータは、本明細書に記載するように、訓練済みの、したがって誤検出のリスクを低減する独自のニューラル・ネットワークを有してもよい。

【0186】

少なくとも一実施例では、車両1100はさらに、インフォテインメントSoC1130（たとえば、車両内インフォテインメント・システム（IVI）：in-vehicle infotainment system）を含んでもよい。インフォテインメント・システム1130はSoCとして図示及び説明されるが、少なくとも一実施例では、SoCではなくてもよく、限定することなく2つ以上の個別の構成要素を含んでもよい。少なくとも一実施例では、インフォテインメントSoC1130は、限定することなく、ハードウェアとソフトウェアの組合せを含んでもよく、この組合せを使用して、オーディオ（たとえば、音楽、パーソナル・デジタル・アシスタント、ナビゲーション命令、ニュース、ラジオなど）、ビデオ（たとえば、TV、映画、ストリーミングなど）、電話（たとえば、ハンズフリー通話）、ネットワーク接続（たとえば、LTE、Wi-Fiなど）、及び／又は情報サービス（たとえば、ナビゲーション・システム、後方駐車支援、無線データ・システム、車両関連情報、たとえば燃料レベル、合計走行距離、ブレーキ燃料レベル、オイル・レベル、ドアの開閉、空気フィルタ情報など）を車両1100に提供してもよい。たとえば、インフォテインメントSoC1130は、ラジオ、ディスク再生装置、ナビゲーシ

10

20

30

40

50

ョン・システム、ビデオ再生装置、USB及びBluetooth接続、カーピュータ、車内エンタテインメント、Wi-Fi、ハンドル・オーディオ制御、ハンズフリー音声制御、ヘッド・アップ・ディスプレイ（「HUD」：heads-up display）、HMIディスプレイ1134、テレマテックス・デバイス、（たとえば、様々な構成要素、特徴、及び／若しくはシステムを制御及び／若しくは相互作用するための）制御パネル、並びに／又は他の構成要素を含むことができる。少なくとも一実施例では、さらにインフォテインメントSOC1130を使用して、ADASシステム1138からの情報、車両操作計画、軌道などの自律運転情報、周囲環境情報（たとえば、交差点情報、車両情報、道路情報など）、及び／又は他の情報などの（たとえば、視覚的及び／又は聴覚的な）情報が、車両1100のユーザに提供されてもよい。

10

【0187】

少なくとも一実施例では、インフォテインメントSOC1130は、任意の量及びタイプのGPU機能を含んでもよい。少なくとも一実施例では、インフォテインメントSOC1130は、バス1102を介して、車両1100の他のデバイス、システム、及び／又は構成要素と通信してもよい。少なくとも一実施例では、インフォテインメントSOC1130は監視MCUに結合されてもよく、それにより、一次コントローラ1136（たとえば、車両1100の一次及び／又はバックアップのコンピュータ）が故障したときに、インフォテインメント・システムのGPUが、一部の自己運転機能を実行してもよい。少なくとも一実施例では、インフォテインメントSOC1130は、本明細書に記載するように、車両1100を運転手－安全停止モードにしてもよい。

20

【0188】

少なくとも一実施例では、車両1100はさらに、計器クラスタ1132（たとえば、デジタル・ダッシュボード、電子計器クラスタ、デジタル計器パネルなど）を含んでもよい。少なくとも一実施例では、計器クラスタ1132は、限定することなく、コントローラ、及び／又はスーパーコンピュータ（たとえば、個別のコントローラ又はスーパーコンピュータ）を含んでもよい。少なくとも一実施例では、計器クラスタ1132は、限定することなく、スピード・メータ、燃料レベル、油圧、タコメータ、オドメータ、方向指示器、シフトレバー位置インジケータ、シート・ベルト警告灯、パーキング・ブレーキ警告灯、エンジン故障灯、補助拘束システム（たとえば、エアバッグ）情報、ライト制御、安全システム制御、ナビゲーション情報など、任意の数及び組合せの計器セットを含んでもよい。いくつかの例では、インフォテインメントSOC1130と計器クラスタ1132との間で、情報が表示及び／又は共有されてもよい。少なくとも一実施例では、計器クラスタ1132は、インフォテインメントSOC1130の一部として含まれてもよく、又はその逆であってもよい。

30

【0189】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図11Cのシステムにおいて使用されてもよい。

40

【0190】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によっ

50

てゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0191】

図11Dは、少なくとも一実施例による、クラウド・ベースのサーバと図11Aの自律車両1100との間で通信するためのシステム1176の図である。少なくとも一実施例では、システム1176は、限定することなく、サーバ1178、ネットワーク1190、並びに車両1100を含む任意の数及びタイプの車両を含んでもよい。少なくとも一実施例では、サーバ1178は、限定することなく、複数のGPU1184(A)～1184(H)（本明細書ではまとめてGPU1184と呼ぶ）、PCIeスイッチ1182(A)～1182(D)（本明細書ではまとめてPCIeスイッチ1182と呼ぶ）、及び／又はCPU1180(A)～1180(B)（本明細書ではまとめてCPU1180と呼ぶ）を含んでもよい。少なくとも一実施例では、GPU1184、CPU1180、及びPCIeスイッチ1182は、たとえば限定することなく、NVIDIAにより開発されたNVLinkインターフェース1188、及び／又はPCIe接続1186などの高速相互接続によって、相互接続されてもよい。少なくとも一実施例では、GPU1184同士は、NVLink及び／又はNVSwitchを介して接続され、GPU1184とPCIeスイッチ1182は、PCIe相互接続を介して接続される。8個のGPU1184、2個のCPU1180、及び4個のPCIeスイッチ1182が図示してあるが、これは限定するものではない。少なくとも一実施例では、サーバ1178のそれぞれは、限定することなく、任意の数のGPU1184、CPU1180、及び／又はPCIeスイッチ1182を任意の組合せで含んでもよい。たとえば、少なくとも一実施例では、サーバ1178は、それぞれが8個、16個、32個、及び／又はそれ以上のGPU1184を含むことができる。

【0192】

少なくとも一実施例では、サーバ1178は、最近始まった道路工事などの予想外の又は変更された道路状態を示す画像を表す画像データを、ネットワーク1190を介して車両から受信してもよい。少なくとも一実施例では、サーバ1178は、更新済み若しくはそうではないニューラル・ネットワーク1192及び／又は、限定することなく交通状態及び道路状態に関する情報を含む地図情報1194を、ネットワーク1190を介して車両に送信してもよい。少なくとも一実施例では、地図情報1194の更新は、建築現場、穴、迂回路、洪水、及び／又は他の障害物に関する情報など、HDマップ1122に対する更新を、限定することなく含んでもよい。少なくとも一実施例では、ニューラル・ネットワーク1192及び／又は地図情報1194は、環境内の任意の数の車両から受信したデータに表された新しい訓練及び／又は経験から得られたものであってもよく、且つ／又は、データ・センタにおいて（たとえば、サーバ1178及び／又は他のサーバを使用し）実行された訓練に少なくとも部分的に基づき、得られたものであってもよい。

【0193】

少なくとも一実施例では、サーバ1178を使用して、訓練データに少なくとも部分的に基づき、機械学習モデル（たとえば、ニューラル・ネットワーク）が訓練されてもよい。少なくとも一実施例では、訓練データは車両によって生成されてもよく、且つ／又はシミュレーションで（たとえば、ゲーム・エンジンを使用して）生成されてもよい。少なくとも一実施例では、（たとえば、関連するニューラル・ネットワークが教師あり学習により恩恵を受ける場合には）任意の量の訓練データがタグ付けされ、且つ／又は他の前処理を受ける。少なくとも一実施例では、（たとえば、関連するニューラル・ネットワークが教師あり学習を必要としない場合には）任意の量の訓練データはタグ付け及び／又は前処理されない。少なくとも一実施例では、機械学習モデルが訓練されると、機械学習モデルは車両によって使用されてもよく（たとえば、ネットワーク1190を介して車両に送信されてもよく、且つ／又は機械学習モデルは、車両を遠隔監視するためにサーバ1178によって使用されてもよい。

【0194】

10

20

30

40

50

少なくとも一実施例では、サーバ1178は車両からデータを受信し、リアル・タイムの知的推論ができるように、最新のリアル・タイムのニューラル・ネットワークにデータを適用してもよい。少なくとも一実施例では、サーバ1178は、NVIDIAによって開発されたDGX及びDGXステーション・マシンなど、GPU1184によって動く深層学習スーパーコンピュータ及び／又は専用AIコンピュータを含んでもよい。しかし、少なくとも一実施例では、サーバ1178は、CPUにより動くデータ・センタを使用する深層学習インフラストラクチャを含んでもよい。

【0195】

少なくとも一実施例では、サーバ1178の深層学習インフラストラクチャは、高速のリアル・タイムの推論が可能であってもよく、その機能を使用して、車両1100のプロセッサ、ソフトウェア、及び／又は関連ハードウェアの健全性を評価及び確認してもよい。たとえば、少なくとも一実施例では、深層学習インフラストラクチャは、一連の画像、及び／又はその一連の画像において（たとえば、コンピュータ・ビジョン及び／又は他の機械学習の物体分類技法により）車両1100が位置特定した物体など、周期的な更新を車両1100から受信してもよい。少なくとも一実施例では、深層学習インフラストラクチャは、独自のニューラル・ネットワークを実行して物体を識別し、それを車両1100によって識別された物体と比較してもよく、結果が一致せず、車両1100のAIが故障していると深層学習インフラストラクチャが結論づけた場合には、サーバ1178は、車両1100のフェイル・セーフ・コンピュータに制御を掌握し、乗員に通知し、安全な停車操作を完了するよう命じる信号を車両1100に送信してもよい。

【0196】

少なくとも一実施例では、サーバ1178は、GPU1184、及び1つ又は複数のプログラム可能な推論アクセラレータ（たとえば、NVIDIAのTensorRT3デバイス）を含んでもよい。少なくとも一実施例では、GPUにより動くサーバと、推論の加速とを組み合わせることによって、リアル・タイムの応答を可能にすることができる。性能がそれほど重要ではない場合など、少なくとも一実施例では、CPU、FPGA、及び他のプロセッサにより動くサーバが、推論に使用されてもよい。少なくとも一実施例では、1つ又は複数の実施例を実行するために、ハードウェア構造体815が使用される。ハードウェア構造体815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。

【0197】

コンピュータ・システム

図12は、例示的なコンピュータ・システムを示すブロック図であり、このコンピュータ・システムは、少なくとも一実施例による、命令を実行するための実行ユニットを含んでもよいプロセッサとともに形成された、相互接続されたデバイス及び構成要素、システム・オン・チップ（SoC）、又はこれらの何らかの組合せを有するシステムであってもよい。少なくとも一実施例では、コンピュータ・システム1200は、本明細書に記載の実施例などにおいて本開示に従ってデータを処理するためのアルゴリズムを実行する論理を含む実行ユニットを使用するための、プロセッサ1202などの構成要素を、限定することなく含んでもよい。少なくとも一実施例では、コンピュータ・システム1200は、カリフォルニア州サンタクララのインテルコーポレーションから入手可能なPENTIUM（登録商標）プロセッサ・ファミリー、Xeon™、Itanium（登録商標）、XScale™及び／又はStrongARM™、Intel（登録商標）Core（商標）、又はIntel（登録商標）Nervana（商標）マイクロプロセッサなどのプロセッサを含んでもよいが、（他のマイクロプロセッサ、エンジニアリング・ワークステーション、セット・トップ・ボックスなどを有するPCを含め）他のシステムが使用されてもよい。少なくとも一実施例では、コンピュータ・システム1200は、ワシントン州、レッドモンドのマイクロソフトコーポレーションから入手可能なWINDOWS（登録商標）のオペレーティング・システムのあるバージョンを実行してもよいが、他のオペレーティング・システム（たとえば、UNIX（登録商標）及びLinux（登録商標）

）、組み込みソフトウェア、及び／又はグラフィカル・ユーザ・インターフェースが使用されてもよい。

【0198】

実施例は、携帯型デバイス及び組み込みアプリケーションなど、他のデバイスで使用されてもよい。携帯型デバイスのいくつかの例は、セルラー・フォン、インターネット・プロトコル・デバイス、デジタル・カメラ、パーソナル・デジタル・アシスタント（「PDA」：personal digital assistants）、及び携帯型PCを含む。少なくとも一実施例では、組み込みアプリケーションは、マイクロコントローラ、デジタル信号プロセッサ（「DSP」：digital signal processor）、システム・オン・チップ、ネットワーク・コンピュータ（「NetPC」：network computers）、セット・トップ・ボックス、ネットワーク・ハブ、広域ネットワーク（「WAN」：wide area network）スイッチ、又は少なくとも一実施例による1つ又は複数の命令を実行することができる任意の他のシステムを含んでもよい。

【0199】

少なくとも一実施例では、コンピュータ・システム1200は、限定することなくプロセッサ1202を含んでもよく、このプロセッサ1202は限定することなく、本明細書に記載の技法による機械学習モデルの訓練及び／又は推論を実行するための1つ又は複数の実行ユニット1208を含んでもよい。少なくとも一実施例では、コンピュータ・システム1200は、シングル・プロセッサのデスクトップ又はサーバ・システムであるが、別の実施例では、コンピュータ・システム1200はマルチプロセッサ・システムであってもよい。少なくとも一実施例では、プロセッサ1202は、限定することなく、複合命令セット・コンピュータ（「CISC」：complex instruction set computer）マイクロプロセッサ、縮小命令セット・コンピューティング（「RISC」）マイクロプロセッサ、超長命令語（「VLIW」）マイクロプロセッサ、命令セットの組合せを実装するプロセッサ、又は任意の他のプロセッサ・デバイス、たとえばデジタル信号プロセッサなどを含んでもよい。少なくとも一実施例では、プロセッサ1202は、プロセッサ・バス1210に結合されてもよく、このプロセッサ・バスは、プロセッサ1202とコンピュータ・システム1200内の他の構成要素との間でデジタル信号を送信してもよい。

【0200】

少なくとも一実施例では、プロセッサ1202は、限定することなく、レベル1（「L1」）の内部キャッシュ・メモリ（「キャッシュ」）1204を含んでもよい。少なくとも一実施例では、プロセッサ1202は、単一の内部キャッシュ又は複数レベルの内部キャッシュを有してもよい。少なくとも一実施例では、キャッシュ・メモリは、プロセッサ1202の外部にあってもよい。他の実施例は、特定の実装形態及び必要性に応じて、内部キャッシュと外部キャッシュの両方の組合せも含んでもよい。少なくとも一実施例では、レジスタ・ファイル1206は、整数レジスタ、浮動小数点レジスタ、状態レジスタ、及び命令ポインタ・レジスタを限定することなく含む様々なレジスタに、異なるタイプのデータを記憶してもよい。

【0201】

少なくとも一実施例では、整数及び浮動小数点の演算を実行するための論理を限定することなく含む実行ユニット1208も、プロセッサ1202にある。少なくとも一実施例では、プロセッサ1202は、ある一定のマクロ命令のためのマイクロコードを記憶するマイクロコード（「uコード」）読取り専用メモリ（「ROM」：read only memory）も含んでよい。少なくとも一実施例では、実行ユニット1208は、パック命令セット1209に対処する論理を含んでもよい。少なくとも一実施例では、パック命令セット1209を、命令を実行する関連回路とともに汎用プロセッサの命令セットに含めることにより、多くのマルチメディア・アプリケーションによって使用される演算を、プロセッサ1202のパック・データを使用して実行することができる。少なくとも一実施例では、プロセッサのデータ・バスの全幅を使用してパック・データの演算を実行する

ことによって、多くのマルチメディア・アプリケーションを加速し、より効率的に実行することができ、これにより、1度に1つのデータ要素に対して1つ又は複数の演算を実行するためにプロセッサのデータ・バス間でより小さい単位のデータを転送する必要をなくすることができる。

【0202】

少なくとも一実施例では、実行ユニット1208はまた、マイクロコントローラ、組み込みプロセッサ、グラフィックス・デバイス、DSP、及び他のタイプの論理回路において使用されてもよい。少なくとも一実施例では、コンピュータ・システム1200は、限定することなくメモリ1220を含んでもよい。少なくとも一実施例では、メモリ1220は、ダイナミック・ランダム・アクセス・メモリ（「DRAM」）デバイス、スタティック・ランダム・アクセス・メモリ（「SRAM」）デバイス、フラッシュ・メモリ・デバイス、又は他のメモリ・デバイスであってもよい。少なくとも一実施例では、メモリ1220は、プロセッサ1202によって実行されてもよいデータ信号によって表される命令1219、及び／又はデータ1221を記憶してもよい。

【0203】

少なくとも一実施例では、システム論理チップが、プロセッサ・バス1210及びメモリ1220に結合されてもよい。少なくとも一実施例では、システム論理チップは、限定することなく、メモリ・コントローラ・ハブ（「MCH」：memory controller hub）1216を含んでもよく、プロセッサ1202は、プロセッサ・バス1210を介してMCH1216と通信してもよい。少なくとも一実施例では、MCH1216は、命令及びデータを記憶するため、及びグラフィックス・コマンド、データ、及びテクスチャを記憶するために、高帯域幅メモリ経路1218をメモリ1220に提供してもよい。少なくとも一実施例では、MCH1216は、プロセッサ1202と、メモリ1220と、コンピュータ・システム1200の他の構成要素との間でデータ信号を導き、プロセッサ・バス1210と、メモリ1220と、システムI/Oインターフェース1222との間でデータ信号をブリッジしてもよい。少なくとも一実施例では、システム論理チップは、グラフィックス・コントローラに結合するためのグラフィックス・ポートを提供してもよい。少なくとも一実施例では、MCH1216は、高帯域幅メモリ経路1218を介してメモリ1220に結合されてもよく、グラフィックス／ビデオカード1212は、アクセラレーテッド・グラフィックス・ポート（「AGP」：Accelerated Graphics Port）相互接続1214を介してMCH1216に結合されてもよい。

【0204】

少なくとも一実施例では、コンピュータ・システム1200は、MCH1216をI/Oコントローラ・ハブ（「ICH」：I/O controller hub）1230に結合するためのプロプライエタリ・ハブ・インターフェース・バスとしてシステムI/Oインターフェース1222を使用してもよい。少なくとも一実施例では、ICH1230は、ローカルのI/Oバスを介していくつかのI/Oデバイスに直接接続を提供してもよい。少なくとも一実施例では、ローカルI/Oバスは、周辺装置をメモリ1220、チップセット、及びプロセッサ1202に接続するための高速I/Oバスを、限定することなく含んでもよい。例としては、オーディオ・コントローラ1229、ファームウェア・ハブ（「フラッシュBIOS」）1228、ワイヤレス・トランシーバ1226、データ・ストレージ1224、ユーザ入力及びキーボードのインターフェース1225を含むレガシーI/Oコントローラ1223、ユニバーサル・シリアル・バス（「USB」：Universal Serial Bus）ポートなどのシリアル拡張ポート1227、及びネットワーク・コントローラ1234が、限定することなく含まれてもよい。少なくとも一実施例では、データ・ストレージ1224は、ハード・ディスク・ドライブ、フロッピー（登録商標）・ディスク・ドライブ、CD-ROMデバイス、フラッシュ・メモリ・デバイス、又は他の大容量ストレージ・デバイスを備えてもよい。

【0205】

少なくとも一実施例では、図 12 は、相互接続されたハードウェア・デバイス又は「チップ」を含むシステムを示すが、一方他の実施例では、図 12 は例示的な SoC を示してもよい。少なくとも一実施例では、図 12 で示すデバイスは、プロプライエタリ相互接続、標準相互接続（たとえば、PCIe）、又はこれらの何らかの組合せで相互接続されてもよい。少なくとも一実施例では、コンピュータ・システム 1200 の 1 つ又は複数の構成要素は、コンピュータ・エクスプレス・リンク（CXL: compute express link）相互接続を使用して相互接続されてもよい。

【0206】

1 つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理 815 が使用される。推論及び／又は訓練論理 815 に関する詳細事項は、図 8A 及び／又は図 8B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 815 は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図 12 のシステムにおいて使用されてもよい。

【0207】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0208】

図 13 は、少なくとも一実施例による、プロセッサ 1310 を利用するための電子デバイス 1300 を示すブロック図である。少なくとも一実施例では、電子デバイス 1300 は、たとえば限定することなく、ノートブック、タワー・サーバ、ラック・サーバ、ブレード・サーバ、ラップトップ、デスクトップ、タブレット、モバイル・デバイス、電話、組み込みコンピュータ、又は任意の他の好適な電子デバイスであってもよい。

【0209】

少なくとも一実施例では、電子デバイス 1300 は、任意の好適な数又は種類の構成要素、周辺装置、モジュール、若しくはデバイスに通信可能に結合されたプロセッサ 1310 を、限定することなく含んでもよい。少なくとも一実施例では、プロセッサ 1310 は、I²C バス、システム・マネージメント・バス（「SMBus」: System Management Bus）、ロー・ピン・カウント（LPC: Low Pin Count）バス、シリアル・ペリフェラル・インターフェース（「SPI」: Serial Peripheral Interface）、ハイ・デフィニション・オーディオ（「HDA」: High Definition Audio）バス、シリアル・アドバンス・テクノロジー・アタッチメント（「SATA」: Serial Advance Technology Attachment）バス、ユニバーサル・シリアル・バス（「USB」）（バージョン 1、2、3 など）、又はユニバーサル非同期レシーバ／トランスミッタ（「UART」: Universal Asynchronous Receiver/Transmitter）バスなどのバス若しくはインターフェースを使用して結合される。少なくとも一実施例では、図 13 は、相互接続されたハードウェア・デバイス又は「チップ」を含むシステムを示すが、一方他の実施例では、図 13 は例示的な SoC を示してもよい。少なくとも一実施例では、図 13 に示すデバイスは、プロプライエタリ相互接続、標準相互接続（たとえば、PCIe）、又はこれらの何らかの組合せで相互接続されてもよい。少なくとも一実施例では、図 13 の 1 つ又は複数の構成要素は、コンピュータ・エクスプレス・リンク（CXL）相互接続を使用して相互接続されてもよい。

10

20

30

40

50

【0210】

少なくとも一実施例では、図13は、ディスプレイ1324、タッチ画面1325、タッチ・パッド1330、近距離無線通信ユニット（「NFC」：Near Field Communications unit）1345、センサ・ハブ1340、熱センサ1346、エクスプレス・チップセット（「EC」：Express Chipset）1335、トラステッド・プラットフォーム・モジュール（「TPM」：Trusted Platform Module）1338、BIOS／ファームウェア／フラッシュ・メモリ（「BIOS、FWフラッシュ」：BIOS／firmware／flash memory）1322、DSP1360、ソリッド・ステート・ディスク（「SSD」：Solid State Disk）若しくはハード・ディスク・ドライブ（「HDD」：Hard Disk Drive）などのドライブ1320、ワイヤレス・ローカル・エリア・ネットワーク・ユニット（「WLAN」：wireless local area network unit）1350、Bluetoothユニット1352、ワイヤレス広域ネットワーク・ユニット（「WWAN」：Wireless Wide Area Network unit）1356、全地球測位システム（GPS：Global Positioning System）ユニット1355、USB3.0カメラなどのカメラ（「USB3.0カメラ」）1354、及び／又は、たとえばLPDDR3規格に実装された低電力ダブル・データ・レート（「LPDDR」：Low Power Double Data Rate）メモリ・ユニット（「LPDDR3」）1315を含んでもよい。これらの構成要素は、それぞれ任意の好適なやり方で実装されてもよい。

10

20

【0211】

少なくとも一実施例では、上述した構成要素を介して、他の構成要素がプロセッサ1310に通信可能に結合されてもよい。少なくとも一実施例では、加速度計1341、周囲光センサ（「ALS」：Ambient Light Sensor）1342、コンパス1343、及びジャイロスコープ1344が、センサ・ハブ1340に通信可能に結合されてもよい。少なくとも一実施例では、熱センサ1339、ファン1337、キーボード1336、及びタッチ・パッド1330が、EC1335に通信可能に結合されてもよい。少なくとも一実施例では、スピーカ1363、ヘッドフォン1364、及びマイクロフォン（「mic」）1365が、オーディオ・ユニット（オーディオ・コーデック及びクラスDアンプ）1362に通信可能に結合されてもよく、このオーディオ・ユニットが、DSP1360に通信可能に結合されてもよい。少なくとも一実施例では、オーディオ・ユニット1362は、たとえば限定することなく、オーディオ・コーダ／デコーダ（「コーデック」）及びクラスDアンプリファイアを含んでもよい。少なくとも一実施例では、SIMカード（「SIM」）1357は、WWANユニット1356に通信可能に結合されてもよい。少なくとも一実施例では、WLANユニット1350及びBluetoothユニット1352などの構成要素、並びにWWAN1356は、次世代フォーム・ファクタ（「NGFF」：Next Generation Form Factor）に実装されてもよい。

30

【0212】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図13のシステムにおいて使用されてもよい。

40

【0213】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なく

50

とも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0214】

図14は、少なくとも一実施例による、コンピュータ・システム1400を示す。少なくとも一実施例では、コンピュータ・システム1400は、本開示全体を通して説明する様々なプロセス及び方法を実装するように構成される。

【0215】

少なくとも一実施例では、コンピュータ・システム1400は、限定することなく、少なくとも1つの中央処理装置（「CPU」）1402を含み、この処理装置は、PCI：Peripheral Component Interconnect（「ペリフェラル・コンポーネント・インターコネクト」）、ペリフェラル・コンポーネント・インターコネクト・エクスプレス（「PCI-Express」：peripheral component interconnect express）、AGP：Accelerated Graphics Port（「アクセラレーテッド・グラフィックス・ポート」）、ハイパートランスポート、又は任意の他のバス若しくはポイントツーポイントの通信プロトコルなど、任意の好適なプロトコルを使用して実装された通信バス1410に接続される。少なくとも一実施例では、コンピュータ・システム1400は、メイン・メモリ1404、及び（たとえば、ハードウェア、ソフトウェア、又はこれらの組合せとして実装される）制御論理を限定することなく含み、データは、ランダム・アクセス・メモリ（「RAM」：random access memory）の形をとってもよいメイン・メモリ1404に記憶される。少なくとも一実施例では、ネットワーク・インターフェース・サブシステム（「ネットワーク・インターフェース」）1422は、コンピュータ・システム1400を有する他のシステムからデータを受信し、コンピュータ・システム1400を有する他のシステムにデータを送信するための他のコンピューティング・デバイス及びネットワークとのインターフェースを提供する。

【0216】

少なくとも一実施例では、コンピュータ・システム1400は、少なくとも一実施例では、限定することなく、入力デバイス1408、パラレル処理システム1412、及びディスプレイ・デバイス1406を含み、このディスプレイ・デバイスは、従来の陰極線管（「CRT」：cathode ray tube）、液晶ディスプレイ（「LCD」：liquid crystal display）、発光ダイオード（「LED」：light emitting diode）ディスプレイ、プラズマ・ディスプレイ、又は他の好適なディスプレイ技術を使用して実装することができる。少なくとも一実施例では、ユーザ入力は、キーボード、マウス、タッチ・パッド、マイクロフォンなどの入力デバイス1408から受け取る。少なくとも一実施例では、本明細書に記載の各モジュールを単一の半導体プラットフォームに置いて、処理システムを形成することができる。

【0217】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図14のシステムにおいて使用されてもよい。

【0218】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくと

10

20

30

40

50

も部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0219】

図15は、少なくとも一実施例によるコンピュータ・システム1500を示す。少なくとも一実施例では、コンピュータ・システム1500は、限定することなく、コンピュータ1510及びUSBスティック1520を含んでもよい。少なくとも一実施例では、コンピュータ・システム1510は、限定することなく、任意の数及びタイプのプロセッサ（図示せず）、並びにメモリを含んでもよい。少なくとも一実施例では、コンピュータ1510は、限定することなく、サーバ、クラウド・インスタンス、ラップトップ、及びデスクトップ・コンピュータを含む。

【0220】

少なくとも一実施例では、USBスティック1520は、限定することなく、処理ユニット1530、USBインターフェース1540、及びUSBインターフェース論理1550を含む。少なくとも一実施例では、処理ユニット1530は、命令を実行することができる任意の命令実行システム、装置、又はデバイスであってもよい。少なくとも一実施例では、処理ユニット1530は、限定することなく、任意の数及びタイプの処理コア（図示せず）を含んでもよい。少なくとも一実施例では、処理ユニット1530は、機械学習に関連する任意の量及びタイプの演算を実行するように最適化された特定用途向け集積回路（「ASIC」）を備える。たとえば、少なくとも一実施例では、処理ユニット1530は、機械学習の推論演算を実行するように最適化されたテンソル処理ユニット（「TPC」：tensor processing unit）である。少なくとも一実施例では、処理ユニット1530は、機械視覚及び機械学習の推論演算を実行するように最適化された視覚処理ユニット（「VPU」）である。

【0221】

少なくとも一実施例では、USBインターフェース1540は、任意のタイプのUSBコネクタ又はUSBソケットであってもよい。たとえば、少なくとも一実施例では、USBインターフェース1540は、データ及び電源用のUSB 3.0 Type-Cのソケットである。少なくとも一実施例では、USBインターフェース1540は、USB 3.0 Type-Aのコネクタである。少なくとも一実施例では、USBインターフェース論理1550は、処理ユニット1530がUSBコネクタ1540を介してデバイス（たとえばコンピュータ1510）と又はインターフェースをとることを可能にする任意の量及びタイプの論理を含んでもよい。

【0222】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図15のシステムにおいて使用されてもよい。

【0223】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的

に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0224】

図16Aは、複数のGPU1610(1)～1610(N)が、高速リンク1640(1)～1640(N)(たとえば、バス、ポイントツーポイント相互接続など)を介して複数のマルチ・コア・プロセッサ1605(1)～1605(M)に通信可能に結合されている例示的なアーキテクチャを示す。少なくとも一実施例では、高速リンク1640(1)～1640(N)は、4GB/秒、30GB/秒、80GB/秒、又はそれ以上の通信スループットをサポートする。1640(1)～1640(N)PCIe4.0又は5.0、及びNVLink2.0を含むがこれらに限定されない様々な相互接続プロトコルが使用されてもよい。様々な図において、「N」及び「M」は、正の整数を表し、その値は図ごとに異なってもよい。

10

【0225】

さらに、少なくとも一実施例では、GPU1610のうちの2つ以上は高速リンク1629(1)～1629(2)を介して相互接続され、これらは、高速リンク1640(1)～1640(N)に使用されたものと同様の又は異なるプロトコル/リンクを使用して実装されてもよい。同様に、マルチ・コア・プロセッサ1605のうちの2つ以上は、高速リンク1628を介して接続されてもよく、この高速リンク1628は、20GB/秒、30GB/秒、120GB/秒、又はそれ以上で動作する対称型マルチプロセッサ(SMP)バスとすることができる。或いは、図16Aに示す様々なシステム構成要素間のすべての通信は、同様のプロトコル/リンクを使用して(たとえば、共通の相互接続ファブリックを介して)実現されてもよい。

20

【0226】

少なくとも一実施例では、各マルチ・コア・プロセッサ1605は、それぞれメモリ相互接続1626(1)～1626(M)を介してプロセッサ・メモリ1601(1)～1601(M)に通信可能に結合され、各GPU1610(1)～1610(N)は、それぞれGPUメモリ・相互接続1650(1)～1650(N)を介してGPUメモリ1620(1)～1620(N)に通信可能に結合される。少なくとも一実施例では、メモリ相互接続1626及び1650は、同様の又は異なるメモリ・アクセス技術を利用してもよい。例として、限定ではなく、プロセッサ・メモリ1601(1)～1601(M)及びGPUメモリ1620は、(積層DRAMを含む)ダイナミック・ランダム・アクセス・メモリ(DRAM)、グラフィックスDDR SDRAM(GDDR)(たとえば、GDDR5、GDDR6)、又は高帯域幅メモリ(HBM)などの揮発性メモリであってもよく、且つ/又は3D XPoint又はNano-Ramなどの不揮発性メモリであってもよい。少なくとも一実施例では、(たとえば、2レベルのメモリ(2LM)階層を使用して)、プロセッサ・メモリ1601のいくつかの部分は揮発性メモリであってもよく、別の部分は不揮発性メモリであってもよい。

30

【0227】

本明細書に記載するように、様々なマルチ・コア・プロセッサ1605及びGPU1610は、それぞれ特定のメモリ1601、1620に物理的に結合されてもよい、及び/又は仮想システムのアドレス空間(「実効アドレス」空間とも呼ぶ)が様々な物理メモリ間に分配されている統合されたメモリ・アーキテクチャが実装されてもよい。たとえば、プロセッサ・メモリ1601(1)～1601(M)はそれぞれ、64GBのシステム・メモリ・アドレス空間を備えてもよく、GPUメモリ1620(1)～1620(N)はそれぞれ、32GBのシステム・メモリ・アドレス空間を備えてもよく、M=2でN=4の場合、合計256GBのアドレス指定可能メモリが得られる。N及びMについて他の値が考えられる。

40

【0228】

図16Bは、1つの例示的な実施例によるマルチ・コア・プロセッサ1607とグラフ

50

ィックス加速モジュール1646との相互接続のさらなる詳細事項を示す。少なくとも一実施例では、グラフィックス加速モジュール1646は、高速リンク1640（たとえば、PCIeバス、NVLinkなど）を介してプロセッサ1607に結合されるライン・カードに集積された1つ又は複数のGPUチップを含んでもよい。少なくとも一実施例では、或いは、グラフィックス加速モジュール1646は、プロセッサ1607を有するパッケージ又はチップに集積されてもよい。

【0229】

少なくとも一実施例では、プロセッサ1607は、複数のコア1660A～1660Dを含み、それぞれのコアが、トランスレーション・ルックアサイド・バッファ（「TLB」）1661A～1661Dと、1つ又は複数のキャッシュ1662A～1662Dとを有する。少なくとも一実施例では、コア1660A～1660Dは、命令を実行しデータを処理するための、図示していない様々な他の構成要素を含んでもよい。少なくとも一実施例では、キャッシュ1662A～1662Dは、レベル1（L1）及びレベル2（L2）のキャッシュを備えてもよい。さらに、1つ又は複数の共有キャッシュ1656が、キャッシュ1662A～1662Dに含まれ、コア1660A～1660Dのセットによって共有されてもよい。たとえば、プロセッサ1607の一実施例は、24個のコアを含み、各コアが、独自のL1キャッシュ、12個の共有L2キャッシュ、及び12個の共有L3キャッシュを有する。この実施例では、1つ又は複数のL2及びL3のキャッシュが、2つの隣接するコアによって共有される。少なくとも一実施例では、プロセッサ1607及びグラフィックス加速モジュール1646は、システム・メモリ1614に接続されており、このシステム・メモリは、図16Aのプロセッサ・メモリ1601（1）～1601（M）を含んでもよい。

【0230】

少なくとも一実施例では、様々なキャッシュ1662A～1662D、1656、及びシステム・メモリ1614に記憶されたデータ及び命令については、コヒーレンス・バス1664を介したコア間通信によって、コヒーレンスが維持される。少なくとも一実施例では、たとえば、各キャッシュは、特定のキャッシュ・ラインに対する読取り又は書込みを検出したことに応答して、コヒーレンス・バス1664を介して通信するために、それに関連するキャッシュ・コヒーレンス論理／回路を有してもよい。少なくとも一実施例では、キャッシュ・アクセスを監視するために、コヒーレンス・バス1664を介してキャッシュ・スヌーピング・プロトコルが実装される。

【0231】

少なくとも一実施例では、プロキシ回路1625が、グラフィックス加速モジュール1646をコヒーレンス・バス1664に通信可能に結合して、グラフィックス加速モジュール1646がコア1660A～1660Dのピアとしてキャッシュ・コヒーレンス・プロトコルに参加できるようにする。特に、少なくとも一実施例では、インターフェース1635は、高速リンク1640を介してプロキシ回路1625への接続を提供し、インターフェース1637は、グラフィックス加速モジュール1646を高速リンク1640に接続する。

【0232】

少なくとも一実施例では、アクセラレータ統合回路1636は、グラフィックス加速モジュール1646の複数のグラフィックス処理エンジン1631（1）～1631（N）の代わりに、キャッシュ管理、メモリ・アクセス、コンテンツ管理、及び割込み管理のサービスを提供する。少なくとも一実施例では、グラフィックス処理エンジン1631（1）～1631（N）はそれぞれ、別個のグラフィックス・プロセッシング・ユニット（GPU）を備えてもよい。少なくとも一実施例では、或いは、グラフィックス処理エンジン1631（1）～1631（N）は、GPUの中に、グラフィックス実行ユニット、メディア処理エンジン（たとえば、ビデオ・エンコーダ／デコーダ）、サンブラ、及びブリット・エンジンなど、異なるタイプのグラフィックス処理エンジンを備えてもよい。少なくとも一実施例では、グラフィックス加速モジュール1646は、複数のグラフィックス処

理エンジン1631(1)~1631(N)を有するGPUであってもよく、又はグラフィックス処理エンジン1631(1)~1631(N)は、共通のパッケージ、ライン・カード、若しくはチップに集積された個々のGPUであってもよい。

【0233】

少なくとも一実施例では、アクセラレータ統合回路1636は、仮想から物理のメモリ・トランスレーション(実効から実(effective-to-real)のメモリ・トランスレーションとも呼ばれる)など、様々なメモリ管理機能を実行するためのメモリ管理ユニット(MMU)1639、及びシステム・メモリ1614にアクセスするためのメモリ・アクセス・プロトコルを含む。少なくとも一実施例では、MMU1639は、仮想/実効から物理/実へのアドレス・トランスレーションをキャッシュするためのトランスレーション・ルックアサイド・バッファ(TLB)(図示せず)も含むことができる。少なくとも一実施例では、キャッシュ1638は、グラフィックス処理エンジン1631(1)~1631(N)から効率的にアクセスできるように、コマンド及びデータを記憶することができる。少なくとも一実施例では、キャッシュ1638及びグラフィックス・メモリ1633(1)~1633(M)に記憶されたデータは、場合によりフェッチ・ユニット1644を使用して、コア・キャッシュ1662A~1662D、1656、及びシステム・メモリ1614とコヒーレントに保たれる。述べたように、これは、キャッシュ1638及びメモリ1633(1)~1633(M)の代わりにプロキシ回路1625を介して(たとえば、プロセッサ・キャッシュ1662A~1662D、1656におけるキャッシュ・ラインの修正/アクセスに関するアップデートをキャッシュ1638に送り、キャッシュ1638からのアップデートを受け取って)実現されてもよい。

【0234】

少なくとも一実施例では、レジスタ1645のセットが、グラフィックス処理エンジン1631(1)~1631(N)によって実行されるスレッドのためのコンテキスト・データを記憶し、コンテキスト管理回路1648が、スレッド・コンテキストを管理する。たとえば、コンテキスト管理回路1648は、コンテキスト・スイッチ中に様々なスレッドのコンテキストを保存及び復元するために、保存及び復元の動作を実行してもよい(たとえば、ここで、第2のスレッドをグラフィックス処理エンジンによって実行できるように、第1のスレッドが保存され、第2のスレッドが記憶される)。たとえば、コンテキスト・スイッチ時に、コンテキスト管理回路1648は、現在のレジスタ値を(たとえば、コンテキスト・ポインタによって識別された)メモリの指定領域に記憶してもよい。次いで、コンテキストに戻るときに、コンテキスト管理回路1648がレジスタ値を復元してもよい。少なくとも一実施例では、割込み管理回路1647は、システム・デバイスから受け取った割込みを受け取り、処理する。

【0235】

少なくとも一実施例では、グラフィックス処理エンジン1631からの仮想/実効アドレスは、MMU1639によってシステム・メモリ1614の実/物理アドレスにトランスレートされる。少なくとも一実施例では、アクセラレータ統合回路1636の一実施例は、複数(たとえば、4個、8個、16個)のグラフィックス・アクセラレータ・モジュール1646、及び/又は他のアクセラレータ・デバイスをサポートする。少なくとも一実施例では、グラフィックス・アクセラレータ・モジュール1646は、プロセッサ1607上で実行される単一のアプリケーション専用のものであってもよく、又は複数のアプリケーション間で共有されてもよい。少なくとも一実施例では、グラフィックス処理エンジン1631(1)~1631(N)のリソースが複数のアプリケーション又は仮想機械(VM)と共有される仮想化グラフィックス実行環境が存在する。少なくとも一実施例では、リソースは、「スライス」に細分化されてもよく、このスライスが、処理要件、並びにVM及び/又はアプリケーションに関連付けられた優先度に基づき、異なるVM及び/又はアプリケーションに割り振られる。

【0236】

少なくとも一実施例では、アクセラレータ統合回路1636は、グラフィックス加速モ

ジュール 1 6 4 6 のためのシステムへのブリッジとして機能し、アドレス・トランスレーション及びシステム・メモリのキャッシュ・サービスを提供する。さらに、少なくとも一実施例では、アクセラレータ統合回路 1 6 3 6 は、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) の仮想化、割込み、及びメモリ管理をホスト・プロセッサが管理するための仮想化設備を提供してもよい。

【0 2 3 7】

少なくとも一実施例では、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) のハードウェア・リソースは、ホスト・プロセッサ 1 6 0 7 が見る実アドレス空間に明示的にマッピングされるので、いかなるホスト・プロセッサも、実効アドレス値を使用して、これらのリソースに直接アドレス指定することができる。少なくとも一実施例では、アクセラレータ統合回路 1 6 3 6 の 1 つの機能は、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) を、システムにとって独立したユニットに見えるように物理的に分離することである。

【0 2 3 8】

少なくとも一実施例では、1 つ又は複数のグラフィックス・メモリ 1 6 3 3 (1) ~ 1 6 3 3 (M) はそれぞれ、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) のそれぞれに結合され、 $N = M$ である。少なくとも一実施例では、グラフィックス・メモリ 1 6 3 3 (1) ~ 1 6 3 3 (M) は、それぞれのグラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) によって処理される命令及びデータを記憶する。少なくとも一実施例では、グラフィックス・メモリ 1 6 3 3 (1) ~ 1 6 3 3 (M) は、(積層 DRAM を含む) DRAM、GDDR メモリ、(たとえば、GDDR5、GDDR6)、又は HBM などの揮発性メモリであってもよく、且つ／又は 3D XPoint 又は Nano-Ram などの不揮発性メモリであってもよい。

【0 2 3 9】

少なくとも一実施例では、高速リンク 1 6 4 0 を介したデータ・トラフィックを低減するために、グラフィックス・メモリ 1 6 3 3 (1) ~ 1 6 3 3 (M) に記憶されるデータが、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) によって最も頻繁に使用されることになるデータであるようにし、好ましくはコア 1 6 6 0 A ~ 1 6 6 0 D によっては使用されない(少なくとも頻繁には使用されない)データであるようにするためのバイアス技法が使用され得る。同様に、少なくとも一実施例では、バイアス機構は、コアが必要とする(したがって、好ましくはグラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) は必要としない)データを、コアのキャッシュ 1 6 6 2 A ~ 1 6 6 2 D、1 6 5 6、及びシステム・メモリ 1 6 1 4 の中に保つよう試みる。

【0 2 4 0】

図 1 6 C は、アクセラレータ統合回路 1 6 3 6 がプロセッサ 1 6 0 7 内に一体化されている別の例示的な実施例を示す。少なくともこの実施例では、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) は、インターフェース 1 6 3 7 及びインターフェース 1 6 3 5 により、高速リンク 1 6 4 0 を介して直接アクセラレータ統合回路 1 6 3 6 と通信する(この場合も任意の形のバス又はインターフェース・プロトコルであり得る)。少なくとも一実施例では、アクセラレータ統合回路 1 6 3 6 は、図 1 6 B に関して説明したのと同様の動作を実行してもよいが、コヒーレンス・バス 1 6 6 4 及びキャッシュ 1 6 6 2 A ~ 1 6 6 2 D、1 6 5 6 に近接していることを考えると、潜在的には、より高いスループットで動作してもよい。少なくとも一実施例は、アクセラレータ統合回路は、(グラフィックス加速モジュールの仮想化のない)専用プロセスのプログラミング・モデルと、(仮想化のある)共有プログラミング・モデルとを含む異なるプログラミング・モデルをサポートし、これらは、アクセラレータ統合回路 1 6 3 6 によって制御されるプログラミング・モデルと、グラフィックス加速モジュール 1 6 4 6 によって制御されるプログラミング・モデルとを含んでもよい。

【0 2 4 1】

少なくとも一実施例では、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N)

）は、単一のオペレーティング・システムの下で単一のアプリケーション又はプロセスに専用のものである。少なくとも一実施例では、単一のアプリケーションは、他のアプリケーション要求をグラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) に集中させて、VM / パーティション内で仮想化を実現することができる。

【 0 2 4 2 】

少なくとも一実施例では、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) は、複数の VM / アプリケーション・パーティションによって共有されてもよい。少なくとも一実施例では、共有モデルはシステム・ハイパーバイザを使用して、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) を仮想化して、各オペレーティング・システムによるアクセスを可能にしてもよい。少なくとも一実施例では、ハイパーバイザのない単一パーティションのシステムでは、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) は、オペレーティング・システムによって所有される。少なくとも一実施例では、オペレーティング・システムは、グラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) を仮想化して、各プロセス又はアプリケーションへのアクセスを提供することができる。

【 0 2 4 3 】

少なくとも一実施例では、グラフィックス加速モジュール 1 6 4 6 又は個々のグラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) は、プロセス・ハンドルを使用して、プロセス要素を選択する。少なくとも一実施例では、プロセス要素は、システム・メモリ 1 6 1 4 に記憶されており、本明細書に記載の実効アドレスから実アドレスへのトランスレーション技法を使用してアドレス指定可能である。少なくとも一実施例では、プロセス・ハンドルは、ホスト・プロセスのコンテキストをグラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) に登録する（すなわち、プロセス要素リンク・リストにプロセス要素を追加するためのシステム・ソフトウェアをコールする）ときに、ホスト・プロセスに提供される実装固有の値であってもよい。少なくとも一実施例では、プロセス・ハンドルの下位 1 6 ビットは、プロセス要素リンク・リスト内のプロセス要素のオフセットであってもよい。

【 0 2 4 4 】

図 1 6 D は、例示的なアクセラレータ統合スライス 1 6 9 0 を示す。少なくとも一実施例では、「スライス」は、アクセラレータ統合回路 1 6 3 6 の処理リソースの指定部分を備える。少なくとも一実施例では、システム・メモリ 1 6 1 4 内のアプリケーション実効アドレス空間 1 6 8 2 は、プロセス要素 1 6 8 3 を記憶する。少なくとも一実施例では、プロセス要素 1 6 8 3 は、プロセッサ 1 6 0 7 上で実行されているアプリケーション 1 6 8 0 からの GPU 呼出し 1 6 8 1 に応答して、記憶される。少なくとも一実施例では、プロセス要素 1 6 8 3 は、対応するアプリケーション 1 6 8 0 のプロセス状態を収容する。少なくとも一実施例では、プロセス要素 1 6 8 3 に収容されたワーク記述子 (WD) 1 6 8 4 は、アプリケーションによって要求される単一のジョブとすることができ、又はジョブのキューに対するポインタを収容してもよい。少なくとも一実施例では、WD 1 6 8 4 は、アプリケーションの実効アドレス空間 1 6 8 2 におけるジョブ要求キューに対するポインタである。

【 0 2 4 5 】

少なくとも一実施例では、グラフィックス加速モジュール 1 6 4 6 及び／又は個々のグラフィックス処理エンジン 1 6 3 1 (1) ~ 1 6 3 1 (N) は、システム内のプロセスのすべて又はサブセットによって共有されることが可能である。少なくとも一実施例では、プロセス状態を設定し、WD 1 6 8 4 をグラフィックス加速モジュール 1 6 4 6 に送信して、仮想化環境においてジョブを開始するためのインフラストラクチャが、含められてもよい。

【 0 2 4 6 】

少なくとも一実施例では、専用のプロセス・プログラミング・モデルは、実装固有である。少なくとも一実施例では、このモデルでは、単一のプロセスが、グラフィックス加速

10

20

30

40

50

モジュール 1 6 4 6 又は個々のグラフィックス処理エンジン 1 6 3 1 を所有する。少なくとも一実施例では、グラフィックス加速モジュール 1 6 4 6 が単一のプロセスによって所有されるとき、グラフィックス加速モジュール 1 6 4 6 が割り当てられたときに、ハイパーバイザは、所有パーティションについてアクセラレータ統合回路 1 6 3 6 を初期化し、オペレーティング・システムは、所有プロセスについてアクセラレータ統合回路 1 6 3 6 を初期化する。

【0 2 4 7】

少なくとも一実施例では、動作時、アクセラレータ統合スライス 1 6 9 0 内の WD フェッチ・ユニット 1 6 9 1 は、グラフィックス加速モジュールの 1 つ又は複数のグラフィックス処理エンジンによって行われることになるワークの表示を含む次の WD 1 6 8 4 をフェッチする。少なくとも一実施例では、図示してあるように、WD 1 6 8 4 からのデータは、レジスタ 1 6 4 5 に記憶され、MMU 1 6 3 9、割込み管理回路 1 6 4 7、及び／又はコンテキスト管理回路 1 6 4 8 によって使用されてもよい。たとえば、MMU 1 6 3 9 の一実施例は、OS 仮想アドレス空間 1 6 8 5 内のセグメント／ページ・テーブル 1 6 8 6 にアクセスするためのセグメント／ページ・ウォーク回路を含む。少なくとも一実施例では、割込み管理回路 1 6 4 7 は、グラフィックス加速モジュール 1 6 4 6 から受け取った割込みイベント 1 6 9 2 を処理してもよい。少なくとも一実施例では、グラフィックス動作を実行するとき、グラフィックス処理エンジン 1 6 3 1 (1) ～ 1 6 3 1 (N) によって生成された実効アドレス 1 6 9 3 は、MMU 1 6 3 9 によって実アドレスにトランスレートされる。

【0 2 4 8】

少なくとも一実施例では、レジスタ 1 6 4 5 が、各グラフィックス処理エンジン 1 6 3 1 (1) ～ 1 6 3 1 (N)、及び／又はグラフィックス加速モジュール 1 6 4 6 について複製され、ハイパーバイザ又はオペレーティング・システムによって初期化されてもよい。少なくとも一実施例では、これらの複製されたレジスタのそれぞれは、アクセラレータ統合スライス 1 6 9 0 に含まれてもよい。ハイパーバイザによって初期化されてもよい例示的なレジスタを、表 1 に示す。

【表 1】

表 1 ハイパーバイザ初期化レジスタ

レジスタ #	説明
1	スライス制御レジスタ
2	実アドレス (RA) スケジュール済みプロセス・エリア・ポインタ
3	権限マスク・オーバーライド・レジスタ
4	割込みベクトル・テーブル・エントリ・オフセット
5	割込みベクトル・テーブル・エントリ制限
6	状態レジスタ
7	論理パーティション ID
8	実アドレス (RA) ハイパーバイザ・アクセラレータ利用記録ポインタ
9	ストレージ記述レジスタ

【0 2 4 9】

オペレーティング・システムによって初期化されてもよい例示的なレジスタを、表 2 に示す。

10

20

30

40

50

【表 2】

表 2 オペレーティング・システム初期化レジスタ

レジスタ #	説明
1	プロセス及びスレッド識別
2	実効アドレス (E A) コンテキスト保存／復元ポインタ
3	仮想アドレス (V A) アクセラレータ利用記録ポインタ
4	仮想アドレス (V A) ストレージ・セグメント・テーブル・ポインタ
5	権限マスク
6	ワーク記述子

10

【0 2 5 0】

少なくとも一実施例では、各WD 1 6 8 4は、特定のグラフィックス加速モジュール 1 6 4 6 及び／又はグラフィックス処理エンジン 1 6 3 1 (1) ～ 1 6 3 1 (N) に固有のものである。少なくとも一実施例では、WD 1 6 8 4は、グラフィックス処理エンジン 1 6 3 1 (1) ～ 1 6 3 1 (N) がワークを行うために必要とするすべての情報を収容し、又は完了すべきワークのコマンド・キューをアプリケーションがセットアップした場所であるメモリ・ロケーションを指すポインタとすることができる。

【0 2 5 1】

図 1 6 E は、共有モデルの例示的な一実施例のさらなる詳細事項を示す。この実施例は、プロセス要素リスト 1 6 9 9 が記憶されているハイパーバイザ実アドレス空間 1 6 9 8 を含む。少なくとも一実施例では、ハイパーバイザ実アドレス空間 1 6 9 8 は、オペレーティング・システム 1 6 9 5 のグラフィックス加速モジュール・エンジンを仮想化するハイパーバイザ 1 6 9 6 を介してアクセス可能である。

20

【0 2 5 2】

少なくとも一実施例では、共有プログラミング・モデルは、システム内のすべて又はサブセットのパーティションからのすべて又はサブセットのプロセスが、グラフィックス加速モジュール 1 6 4 6 を使用できるようにする。少なくとも一実施例では、グラフィックス加速モジュール 1 6 4 6 が複数のプロセス及びパーティションによって共有されるプログラミング・モデルが、2 つ、つまり時間スライス共有及びグラフィックス指定共有 (g r a p h i c s - d i r e c t e d s h a r e d) が存在する。

30

【0 2 5 3】

少なくとも一実施例では、このモデルでは、システム・ハイパーバイザ 1 6 9 6 がグラフィックス加速モジュール 1 6 4 6 を所有しており、その機能をすべてのオペレーティング・システム 1 6 9 5 にとって利用可能にする。少なくとも一実施例では、システム・ハイパーバイザ 1 6 9 6 による仮想化をグラフィックス加速モジュール 1 6 4 6 がサポートするために、グラフィックス加速モジュール 1 6 4 6 は、1) アプリケーションのジョブ要求は自律でなくてはならず（すなわち、ジョブ間で状態を維持する必要はなく）、又はグラフィックス加速モジュール 1 6 4 6 が、コンテキストの保存及び復元の機構を提供しなくてはならない、2) アプリケーションのジョブ要求は、あらゆるトランスレーション誤りも含めて指定された時間量で完了するようグラフィックス加速モジュール 1 6 4 6 によって保証され、又はグラフィックス加速モジュール 1 6 4 6 が、ジョブの処理をプリエンプションする機能を提供する、及び 3) グラフィックス加速モジュール 1 6 4 6 は、指定の共有プログラミング・モデルで動作しているとき、プロセス間で公平性が保証されなくてはならないなどのいくつかの要件に準拠してもよい。

40

【0 2 5 4】

少なくとも一実施例では、アプリケーション 1 6 8 0 は、グラフィックス加速モジュールのタイプ、ワーク記述子 (W D)、権限マスク・レジスタ (A M R) 値、及びコンテキスト保存／復元エリア・ポインタ (C S R P) を伴って、オペレーティング・システム 1 6 9 5 のシステム・コールを行う必要がある。少なくとも一実施例では、グラフィックス

50

加速モジュールのタイプは、システム・コールで目的とする加速機能を記述している。少なくとも一実施例では、グラフィックス加速モジュールのタイプは、システム固有値であってもよい。少なくとも一実施例では、WDは、グラフィックス加速モジュール1646のために特にフォーマット化されており、グラフィックス加速モジュール1646のコマンド、ユーザ定義の構造を指す実効アドレス・ポインタ、コマンドのキューを指す実効アドレス・ポインタ、又はグラフィックス加速モジュール1646によって行われるワークを記述するための任意の他のデータ構造の形とすることができる。

【0255】

少なくとも一実施例では、AMR値は、現在のプロセスに使用するためのAMR状態である。少なくとも一実施例では、オペレーティング・システムに渡される値は、AMRをセッティングするアプリケーションと同様である。少なくとも一実施例では、アクセラレータ統合回路1636（図示せず）及びグラフィックス加速モジュール1646の実装形態が、ユーザ権限マスク・オーバーライド・レジスタ（UAMOR）をサポートしていない場合、オペレーティング・システムは、AMR値に現在のUAMOR値を適用してから、ハイパーバイザ・コールにAMRを渡してもよい。少なくとも一実施例では、ハイパーバイザ1696は、任意選択で、現在の権限マスク・オーバーライド・レジスタ（AMOR）値を適用してから、AMRをプロセス要素1683に入れてもよい。少なくとも一実施例では、CSRPは、グラフィックス加速モジュール1646がコンテキスト状態を保存及び復元するためのアプリケーションの実効アドレス空間1682内のエリアの実効アドレスを収容するレジスタ1645のうちの1つである。少なくとも一実施例では、ジョブ間で、又はジョブがプリエンプションされるときに、いかなる状態も保存する必要のない場合は、このポインタは任意選択である。少なくとも一実施例では、コンテキスト保存／復元エリアは、ピン留めされたシステム・メモリであってもよい。

【0256】

システム・コールを受け取ると、オペレーティング・システム1695は、アプリケーション1680が登録済みであり、グラフィックス加速モジュール1646を使用する権限が与えられていることを検証してもよい。少なくとも一実施例では、次いで、オペレーティング・システム1695は、表3に示す情報を伴ってハイパーバイザ1696にコールする。

【表3】

表3 OSからハイパーバイザへのコール・パラメータ

パラメータ #	説明
1	ワーク記述子 (WD)
2	権限マスク・レジスタ (AMR) 値 (おそらくマスクされている)
3	実効ドレス (EA) コンテキスト保存／復元エリア・ポインタ (CSRP)
4	プロセスID (PID) 及び任意選択のスレッドID (TID)
5	仮想アドレス (VA) アクセラレータ利用記録ポインタ (AURP)
6	ストレージ・セグメント・テーブル・ポインタ (SSTP) の仮想アドレス
7	論理割込みサービス番号 (LISN)

【0257】

少なくとも一実施例では、ハイパーバイザ・コールを受け取ると、ハイパーバイザ1696は、オペレーティング・システム1695が登録済みであり、グラフィックス加速モジュール1646を使用する権限が与えられていることを検証する。少なくとも一実施例では、次いでハイパーバイザ1696は、プロセス要素1683を、対応するグラフィックス加速モジュール1646のタイプのプロセス要素リンク・リストに入れる。少なくとも一実施例では、プロセス要素は、表4に示す情報を含んでもよい。

10

20

30

40

50

【表 4】

表 4 プロセス要素情報

要素 #	説明
1	ワーク記述子 (WD)
2	権限マスク・レジスタ (AMR) 値 (おそらくマスクされている)
3	実効ドレス (EA) コンテキスト保存／復元エリア・ポインタ (CSR P)
4	プロセス ID (PID) 及び任意選択のスレッド ID (TID)
5	仮想アドレス (VA) アクセラレータ利用記録ポインタ (AUR P)
6	ストレージ・セグメント・テーブル・ポインタ (SST P) の仮想アドレス
7	理割込みサービス番号 (LIS N)
8	ハイパーバイザ・コール・パラメータから導かれる割込みベクトル・テーブル
9	状態レジスタ (SR) 値
10	論理パーティション ID (LPID)
11	実アドレス (RA) ハイパーバイザ・アクセラレータ利用記録ポインタ
12	ストレージ記述子レジスタ (SDR)

10

【0258】

少なくとも一実施例では、ハイパーバイザは、複数のアクセラレータ統合スライス 1690 のレジスタ 1645 を初期化する。

【0259】

20

図 16 F に示すように、少なくとも一実施例では、物理プロセッサ・メモリ 1601 (1) ~ 1601 (N) 及び GPU メモリ 1620 (1) ~ 1620 (N) にアクセスするために使用される共通の仮想メモリ・アドレス空間を介してアドレス指定可能である統合メモリが使用される。この実装形態では、GPU 1610 (1) ~ 1610 (N) で実行される動作は、プロセッサ・メモリ 1601 (1) ~ 1601 (N) にアクセスするのと同じ仮想／実効メモリ・アドレス空間を利用し、且つその逆も同様であり、それによりプログラマビリティが簡単になる。少なくとも一実施例では、仮想／実効アドレス空間の第 1 の部分はプロセッサ・メモリ 1601 (1) に割り振られ、第 2 の部分は第 2 のプロセッサ・メモリ 1601 (N) に割り振られ、第 3 の部分は GPU メモリ 1620 (1) に割り振られるというように続く。少なくとも一実施例では、仮想／実効メモリ空間全体 (実効アドレス空間と呼ばれることもある) は、これによりプロセッサ・メモリ 1601 及び GPU メモリ 1620 のそれぞれにわたって分配されて、仮想アドレスが物理メモリにマッピングされた状態で、いずれかのプロセッサ又は GPU が、いずれかの物理メモリにアクセスできるようになる。

30

【0260】

少なくとも一実施例では、MMU 1639A ~ 1639E のうちの 1 つ又は複数の中のバイアス／コヒーレンス管理回路 1694A ~ 1694E は、1 つ又は複数のホスト・プロセッサ (たとえば、1605) のキャッシュと GPU 1610 のキャッシュとの間でキャッシュ・コヒーレンスを確保し、バイアス技法を実装して、ある特定のタイプのデータが記憶されるべき物理メモリを示す。少なくとも一実施例では、バイアス／コヒーレンス管理回路 1694A ~ 1694E の複数のインスタンスが図 16 F に示されるが、バイアス／コヒーレンス回路は、1 つ又は複数のホスト・プロセッサ 1605 の MMU 内に実装されてもよく、且つ／又はアクセラレータ統合回路 1636 内に実装されてもよい。

40

【0261】

一実施例は、GPU メモリ 1620 をシステム・メモリの一部としてマッピングできるようにし、共有仮想メモリ (SVM) 技法を使用してアクセス可能にすることができるが、完全なシステム・キャッシュ・コヒーレンスに関連する性能の低下が生じることはない。少なくとも一実施例では、GPU メモリ 1620 が、面倒なキャッシュ・コヒーレンス・オーバーヘッドなく、システム・メモリとしてアクセス可能であることにより、GPU オフロードのための有益な動作環境が提供される。少なくとも一実施例では、この構成に

50

よって、従来のI/O DMAデータ・コピーのオーバーヘッドがなくとも、ホスト・プロセッサ1605ソフトウェアがオペランドを設定し、計算結果にアクセスすることが可能になる。少なくとも一実施例では、こうした従来のコピーは、ドライバ・コール、割込み、及びメモリ・マップドI/O (MMIO) アクセスを必要とし、これらはすべて、単純なメモリ・アクセスより非効率的である。少なくとも一実施例では、キャッシュ・コヒーレンス・オーバーヘッドなしでGPUメモリ1620にアクセスできることが、オフロードされた計算の実行時間に不可欠であり得る。少なくとも一実施例では、たとえば、かなりのストリーミング書込みメモリ・トラフィックがある場合には、キャッシュ・コヒーレンス・オーバーヘッドは、GPU1610が見る有効な書込み帯域幅を大幅に低減することある。少なくとも一実施例では、オペランド設定の効率、結果へのアクセスの効率、及びGPU計算の効率は、GPUオフロードの有効性を判定する際に役立つことがある。

10

【0262】

少なくとも一実施例では、GPUバイアス及びホスト・プロセッサ・バイアスの選択は、バイアス・トラッカー・データ構造によって決められる。少なくとも一実施例では、たとえばバイアス・テーブルが使用されてもよく、このテーブルは、GPU付きメモリ・ページ当たり1ビット又は2ビットを含むページ粒度構造であってもよい（たとえば、メモリ・ページの粒度で制御されてもよい）。少なくとも一実施例では、バイアス・テーブルは、（たとえば、バイアス・テーブルの頻繁に使用された／最近使用されたエントリをキャッシュするための）バイアス・キャッシュがGPU1610にある状態又はない状態で、1つ又は複数のGPUメモリ1620の奪われたメモリ範囲（*stolen memory range*）において実装されてもよい。或いは、少なくとも一実施例では、バイアス・テーブル全体が、GPU内に維持されてもよい。

20

【0263】

少なくとも一実施例では、GPU付きメモリ1620への各アクセスに関連付けられたバイアス・テーブルのエントリが、GPUメモリへの実際のアクセスより先にアクセスされて、以下の動作を生じさせる。少なくとも一実施例では、GPUバイアス内での自らのページを見いだすGPU1610からのローカル要求が、対応するGPUメモリ1620に直接転送される。少なくとも一実施例では、ホスト・バイアスにおいて自らのページを見いだすGPUからのローカル要求は、（たとえば、上述した高速リンクを介して）プロセッサ1605に転送される。少なくとも一実施例では、要求されたページをホスト・プロセッサ・バイアスにおいて見いだすプロセッサ1605からの要求は、通常のメモリ読取りと同様に要求を完了させる。或いは、GPUバイアス化ページに向けられた要求は、GPU1610に転送されてもよい。少なくとも一実施例では、次いでGPUは、現在ページを使用していない場合、ホスト・プロセッサ・バイアスにページを移行してもよい。少なくとも一実施例では、ページのバイアス状態は、ソフトウェア・ベースの機構、ハードウェア支援型ソフトウェア・ベースの機構のいずれかによって、又は限られた事例のセットについては、単にハードウェア・ベースの機構によって、変更することができる。

30

【0264】

少なくとも一実施例では、バイアス状態を変更するための1つの機構は、APIコール（たとえば、*OpenCL*）を利用し、このAPIコールが、GPUのデバイス・ドライバをコールし、このデバイス・ドライバが、GPUにメッセージを送って（又はコマンド記述子をキューに加えて）、バイアス状態を変更し、一部の移行については、ホストにおいてキャッシュ・フラッシング動作を実行するよう、GPUを導く。少なくとも一実施例では、キャッシュ・フラッシング動作は、ホスト・プロセッサ1605のバイアスからGPUバイアスへの移行のために使用されるが、反対向きの移行には使用されない。

40

【0265】

少なくとも一実施例では、キャッシュ・コヒーレンスは、ホスト・プロセッサ1605によってキャッシュできないGPUバイアス化ページを一時的にレンダリングすることによって、維持される。少なくとも一実施例では、これらのページにアクセスするために、プロセッサ1605は、GPU1610からのアクセスを要求してもよく、GPU161

50

0 は、すぐにアクセスを許可してもよく、又は許可しなくてもよい。少なくとも一実施例では、したがって、プロセッサ 1605 と GPU 1610 との間の通信を低減するために、GPU バイアス化ページが、GPU によって要求されるが、ホスト・プロセッサ 1605 によっては要求されないようにすること、又はその逆にすることが有益である。

【0266】

1 つ又は複数の実施例を実行するために、ハードウェア構造体 815 が使用される。ハードウェア構造体 815 に関する詳細事項は、図 8 A 及び／又は図 8 B と併せて本明細書に提供され得る。

【0267】

図 17 は、本明細書に記載の様々な実施例による、1 つ又は複数の IP コアを使用して作製することができる例示的な集積回路及び関連グラフィックス・プロセッサを示す。図示してあるものに加えて、少なくとも一実施例では、追加のグラフィックス・プロセッサ／コア、周辺装置インターフェース・コントローラ、若しくは汎用プロセッサ・コアを含む他の論理及び回路が含まれてもよい。

【0268】

図 17 は、少なくとも一実施例による 1 つ又は複数の IP コアを使用して作製することができる例示的なシステム・オン・チップ集積回路 1700 を示すブロック図である。少なくとも一実施例では、集積回路 1700 は、1 つ又は複数のアプリケーション・プロセッサ 1705（たとえば、CPU）、少なくとも 1 つのグラフィックス・プロセッサ 1710 を含み、さらに、画像プロセッサ 1715 及び／又はビデオ・プロセッサ 1720 を含んでもよく、これらのいずれもが、モジュール式 IP コアであってもよい。少なくとも一実施例では、集積回路 1700 は、USB コントローラ 1725、UART コントローラ 1730、SPI／SDIO コントローラ 1735、及び I²S／I²C コントローラ 1740 を含む周辺装置又はバス論理を含む。少なくとも一実施例では、集積回路 1700 は、ハイ・デフィニション・マルチメディア・インターフェース（HDMI（登録商標）：high-definition multimedia interface（登録商標））コントローラ 1750 及びモバイル・インダストリ・プロセッサ・インターフェース（MIPI）ディスプレイ・インターフェース 1755 のうちの 1 つ又は複数の結合されるディスプレイ・デバイス 1745 を含むことができる。少なくとも一実施例では、フラッシュ・メモリ及びフラッシュ・メモリ・コントローラを含むフラッシュ・メモリ・サブシステム 1760 によって、ストレージが提供されてもよい。少なくとも一実施例では、SDRAM 又は SRAM メモリ・デバイスにアクセスするために、メモリ・コントローラ 1765 を介してメモリ・インターフェースが提供されてもよい。少なくとも一実施例では、いくつかの集積回路はさらに、組み込みセキュリティ・エンジン 1770 を含む。

【0269】

1 つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理 815 が使用される。推論及び／又は訓練論理 815 に関する詳細事項は、図 8 A 及び／又は図 8 B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 815 は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために集積回路 1700 において使用されてもよい。

【0270】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によっ

てゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0271】

図18A～図18Bは、本明細書に記載の様々な実施例による、1つ又は複数のIPコアを使用して作製することができる例示的な集積回路及び関連グラフィックス・プロセッサを示す。図示してあるものに加えて、少なくとも一実施例では、追加のグラフィックス・プロセッサ／コア、周辺装置インターフェース・コントローラ、又は汎用プロセッサ・コアを含む他の論理及び回路が含まれてもよい。

【0272】

図18A及び図18Bは、本明細書に記載の実施例による、SoC内で使用するための例示的なグラフィックス・プロセッサを示すブロック図である。図18Aは、少なくとも一実施例による、1つ又は複数のIPコアを使用して作製することができるシステム・オン・チップ集積回路の例示的なグラフィックス・プロセッサ1810を示す。図18Bは、少なくとも一実施例による、1つ又は複数のIPコアを使用して作製することができるシステム・オン・チップ集積回路のさらなる例示的なグラフィックス・プロセッサ1840を示す。少なくとも一実施例では、図18Aのグラフィックス・プロセッサ1810は、低電力グラフィックス・プロセッサ・コアである。少なくとも一実施例では、図18Bのグラフィックス・プロセッサ1840は、高性能グラフィックス・プロセッサ・コアである。少なくとも一実施例では、グラフィックス・プロセッサ1810、1840のそれぞれは、図17のグラフィックス・プロセッサ1710の変形形態とすることができる。

【0273】

少なくとも一実施例では、グラフィックス・プロセッサ1810は、頂点プロセッサ1805と、1つ又は複数のフラグメント・プロセッサ1815A～1815N（たとえば、1815A、1815B、1815C、1815D～1815N-1、及び1815N）とを含む。少なくとも一実施例では、グラフィックス・プロセッサ1810は、別個の論理を介して異なるシェーダ・プログラムを実行することができ、それにより、頂点プロセッサ1805は、頂点シェーダ・プログラムのための動作を実行するように最適化され、一方、1つ又は複数のフラグメント・プロセッサ1815A～1815Nは、フラグメント又はピクセルのシェーダ・プログラムのためのフラグメント（たとえば、ピクセル）シェーディング動作を実行する。少なくとも一実施例では、頂点プロセッサ1805は、3Dグラフィックス・パイプラインの頂点処理ステージを実行し、プリミティブ及び頂点データを生成する。少なくとも一実施例では、フラグメント・プロセッサ1815A～1815Nは、頂点プロセッサ1805によって生成されたプリミティブ及び頂点データを使用して、ディスプレイ・デバイスに表示されるフレーム・バッファを生成する。少なくとも一実施例では、フラグメント・プロセッサ1815A～1815Nは、OpenGLのAPIにおいて提供されるフラグメント・シェーダ・プログラムを実行するように最適化され、OpenGLのAPIは、Direct 3D APIにおいて提供されるピクセル・シェーダ・プログラムと同様の動作を実行するために使用されてもよい。

【0274】

少なくとも一実施例では、グラフィックス・プロセッサ1810はさらに、1つ又は複数のメモリ管理ユニット（MMU）1820A～1820B、キャッシュ1825A～1825B、及び回路相互接続1830A～1830Bを含む。少なくとも一実施例では、1つ又は複数のMMU1820A～1820Bは、頂点プロセッサ1805及び／又はフラグメント・プロセッサ1815A～1815Nを含め、グラフィックス・プロセッサ1810のための仮想から物理のアドレス・マッピングを提供し、それらは、1つ又は複数のキャッシュ1825A～1825Bに記憶された頂点又は画像／テクスチャのデータに加えて、メモリに記憶された頂点又は画像／テキストのデータを参照してもよい。少なくとも一実施例では、1つ又は複数のMMU1820A～1820Bは、図17の1つ若しくは複数のアプリケーション・プロセッサ1705、画像プロセッサ1715、及び／又はビデオ・プロセッサ1720に関連付けられた1つ若しくは複数のMMUを含む、シス

テム内の他のMMUと同期されてもよく、それにより各プロセッサ1705～1720は、共有の又は統合された仮想メモリ・システムに参加することができる。少なくとも一実施例では、1つ又は複数の回路相互接続1830A～1830Bは、グラフィックス・プロセッサ1810が、SoCの内部バスを介して、又は直接接続を介して、SoC内の他のIPコアとインターフェースをとることができるようにする。

【0275】

少なくとも一実施例では、図18Bに示すように、グラフィックス・プロセッサ1840は、1つ又は複数のシェーダ・コア1855A～1855N（たとえば、1855A、1855B、1855C、1855D、1855E、1855F～1855N-1、及び1855N）を含み、このシェーダ・コアは、単一のコア、又はタイプ、又はコアが、頂点シェーダ、フラグメント・シェーダ、及び／又はコンピュート・シェーダを実装するためのシェーダ・プログラム・コードを含むすべてのタイプのプログラム可能なシェーダ・コードを実行することができる統合されたシェーダ・コア・アーキテクチャを提供する。少なくとも一実施例では、シェーダ・コアの数は変えることができる。少なくとも一実施例では、グラフィックス・プロセッサ1840は、1つ又は複数のシェーダ・コア1855A～1855Nに実行スレッドをディスパッチするためのスレッド・ディスパッチャとして作用するコア間タスク・マネージャ1845と、たとえばシーン内のローカル空間コヒーレンスを利用するため、又は内部キャッシュの使用を最適化するために、シーンのレンダリング動作が画像空間において細分化される、タイル・ベースのレンダリングのためのタイリング動作を加速するためのタイリング・ユニット1858とを含む。

【0276】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために集積回路18A及び／又は18Bにおいて使用されてもよい。

【0277】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0278】

図19A～図19Bは、本明細書に記載の実施例による、さらなる例示的なグラフィックス・プロセッサ論理を示す。図19Aは、グラフィックス・コア1900を示し、このグラフィックス・コア1900は、少なくとも一実施例では図17のグラフィックス・プロセッサ1710に含められてもよく、少なくとも一実施例では図18Bのように、統合されたシェーダ・コア1855A～1855Nであってもよい。図19Bは、少なくとも一実施例におけるマルチ・チップ・モジュールに導入するのに適した高並列の汎用グラフィックス・プロセッシング・ユニット（「GPGPU」）1930を示す。

【0279】

少なくとも一実施例では、グラフィックス・コア1900は、共有命令キャッシュ1902、テクスチャ・ユニット1918、及びキャッシュ／共有メモリ1920を含み、これらは、グラフィックス・コア1900内の実行リソースに共通である。少なくとも一実

施例では、グラフィックス・コア1900は、複数のスライス1901A～1901N、又はコアごとのパーティションを含むことができ、グラフィックス・プロセッサは、グラフィックス・コア1900の複数のインスタンスを含むことができる。少なくとも一実施例では、スライス1901A～1901Nは、ローカル命令キャッシュ1904A～1904N、スレッド・スケジューラ1906A～1906N、スレッド・ディスパッチャ1908A～1908N、及びレジスタのセット1910A～1910Nを含むサポート論理を含むことができる。少なくとも一実施例では、スライス1901A～1901Nは、追加機能ユニット（AFU1912A～1912N）、浮動小数点ユニット（FPU1914A～1914N）、整数算術論理演算ユニット（ALU1916～1916N）、アドレス計算ユニット（ACU1913A～1913N）、倍精度浮動小数点ユニット（DPFPU1915A～1915N）、及び行列処理ユニット（MPU1917A～1917N）のセットを含むことができる。

【0280】

少なくとも一実施例では、FPU1914A～1914Nは、単精度（32ビット）及び半精度（16ビット）の浮動小数点演算を実行することができ、DPFPU1915A～1915Nは、倍精度（64ビット）の浮動小数点演算を実行する。少なくとも一実施例では、ALU1916A～1916Nは、8ビット、16ビット、及び32ビットの精度で可変精度の整数演算を実行することができ、混合精度の演算ができるように構成されることが可能である。少なくとも一実施例では、MPU1917A～1917Nも、半精度浮動小数点及び8ビット整数演算を含む混合精度の行列演算ができるように構成されることが可能である。少なくとも一実施例では、MPU1917A～1917Nは、汎用行列－行列乗算（GEMM）の加速をサポートできるようにすることを含め、機械学習アプリケーション・フレームワークを加速するための様々な行列演算を実行することができる。少なくとも一実施例では、AFU1912A～1912Nは、三角関数演算（たとえば、サイン、コサインなど）を含む、浮動小数点ユニット又は整数ユニットにサポートされていない追加の論理演算を実行することができる。

【0281】

1つ又は複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のためにグラフィックス・コア1900において使用されてもよい。

【0282】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0283】

図19Bは、汎用処理ユニット（GPGPU）1930を示し、この処理ユニットは、少なくとも一実施例において、グラフィックス・プロセッシング・ユニットのアレイによる高並列の計算動作を実行可能にするように構成されることが可能である。少なくとも一実施例では、GPGPU1930は、GPGPU1930の他のインスタンスに直接リンクされて、ディープ・ニューラル・ネットワークの訓練スピードを向上させるために複数

10

20

30

40

50

のGPUクラスタを生成することができる。少なくとも一実施例では、GPGPU1930は、ホスト・プロセッサとの接続を可能にするためのホスト・インターフェース1932を含む。少なくとも一実施例では、ホスト・インターフェース1932は、PCIエクスプレス・インターフェースである。少なくとも一実施例では、ホスト・インターフェース1932は、ベンダー固有の通信インターフェース又は通信ファブリックとすることができる。少なくとも一実施例では、GPGPU1930は、ホスト・プロセッサからコマンドを受け取り、グローバル・スケジューラ1934を使用して、これらのコマンドに関連付けられた実行スレッドを、コンピュータ・クラスタ1936A～1936Hのセットに分配する。少なくとも一実施例では、コンピュータ・クラスタ1936A～1936Hは、キャッシュ・メモリ1938を共有する。少なくとも一実施例では、キャッシュ・メモリ1938は、コンピュータ・クラスタ1936A～1936H内のキャッシュ・メモリ用の高レベル・キャッシュとして作用することができる。

10

【0284】

少なくとも一実施例では、GPGPU1930は、メモリ・コントローラ1942A～1942Bのセットを介して、コンピュータ・クラスタ1936A～1936Hに結合されたメモリ1944A～1944Bを含む。少なくとも一実施例では、メモリ1944A～1944Bは、グラフィックス・ダブル・データ・レート（GDDR：graphic double data rate）メモリを含む同期グラフィックス・ランダム・アクセス・メモリ（SGRAM）など、ダイナミック・ランダム・アクセス・メモリ（DRAM）又はグラフィックス・ランダム・アクセス・メモリを含む、様々なタイプのメモリ・デバイスを含むことができる。

20

【0285】

少なくとも一実施例では、コンピュータ・クラスタ1936A～1936Hはそれぞれ、図19Aのグラフィックス・コア1900などのグラフィックス・コアのセットを含み、このグラフィックス・コアのセットは、機械学習計算に適したものを含め、様々な精度で計算動作を実行することができる複数のタイプの整数及び浮動小数点の論理ユニットを含むことができる。たとえば、少なくとも一実施例では、コンピュータ・クラスタ1936A～1936Hのそれぞれにおける浮動小数点ユニットの少なくともサブセットは、16ビット又は32ビットの浮動小数点演算を実行するように構成されることが可能であり、一方、浮動小数点ユニットの別のサブセットは、64ビットの浮動小数点演算を実行するように構成されることが可能である。

30

【0286】

少なくとも一実施例では、GPGPU1930の複数のインスタンスは、コンピュータ・クラスタとして動作するように構成されることが可能である。少なくとも一実施例では、コンピュータ・クラスタ1936A～1936Hにより同期及びデータ交換のために使用される通信は、実施例にわたって異なる。少なくとも一実施例では、GPGPU1930の複数のインスタンスは、ホスト・インターフェース1932を介して通信する。少なくとも一実施例では、GPGPU1930は、I/Oハブ1939を含み、このハブは、GPGPU1930の他のインスタンスへの直接接続を可能にするGPUリンク1940に、GPGPU1930を結合する。少なくとも一実施例では、GPUリンク1940は、GPGPU1930の複数のインスタンス間での通信及び同期を可能にするGPUからGPUへの専用のブリッジに結合される。少なくとも一実施例では、GPUリンク1940は、他のGPGPU又は並列プロセッサにデータを送受信するための高速相互接続に結合される。少なくとも一実施例では、GPGPU1930の複数のインスタンスは、別々のデータ処理システムに位置付けられ、ホスト・インターフェース1932を介してアクセス可能なネットワーク・デバイスを介して通信する。少なくとも一実施例では、GPUリンク1940は、ホスト・インターフェース1932に加えて、又はその代わりに、ホスト・プロセッサへの接続を可能にするように構成することができる。

40

【0287】

少なくとも一実施例では、GPGPU1930は、ニューラル・ネットワークを訓練す

50

るように構成されることが可能である。少なくとも一実施例では、G P G P U 1 9 3 0 は、推論プラットフォーム内で使用することができる。G P G P U 1 9 3 0 が推論のために使用される少なくとも一実施例では、G P G P U 1 9 3 0 は、G P G P U 1 9 3 0 がニューラル・ネットワークの訓練に使用されるときよりも少数のコンピュータ・クラスタ 1 9 3 6 A ~ 1 9 3 6 H を含んでもよい。少なくとも一実施例では、メモリ 1 9 4 4 A ~ 1 9 4 4 B に関連するメモリ技術は、推論の構成と訓練の構成とで異なってもよく、高帯域幅のメモリ技術が、訓練構成に当てられる。少なくとも一実施例では、G P G P U 1 9 3 0 の推論構成は、推論固有の命令をサポートすることができる。たとえば、少なくとも一実施例では、推論構成は、1 つ又は複数の 8 ビットの整数のドット積命令をサポートすることができ、これは、導入済みニューラル・ネットワークの推論動作中に使用されてもよい。

10

【0288】

1 つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理 8 1 5 が使用される。推論及び／又は訓練論理 8 1 5 に関する詳細事項は、図 8 A 及び／又は図 8 B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 8 1 5 は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために G P G P U 1 9 3 0 において使用されてもよい。

【0289】

20

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ~ 図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ~ 図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0290】

図 2 0 は、少なくとも一実施例によるコンピューティング・システム 2 0 0 0 を示すブロック図である。少なくとも一実施例では、コンピューティング・システム 2 0 0 0 は、メモリ・ハブ 2 0 0 5 を含んでもよい相互接続経路を介して通信する 1 つ又は複数のプロセッサ 2 0 0 2 とシステム・メモリ 2 0 0 4 とを有する処理サブシステム 2 0 0 1 を含む。少なくとも一実施例では、メモリ・ハブ 2 0 0 5 は、チップセット構成要素内の別個の構成要素であってもよく、又は 1 つ若しくは複数のプロセッサ 2 0 0 2 内に一体化されていてもよい。少なくとも一実施例では、メモリ・ハブ 2 0 0 5 は、通信リンク 2 0 0 6 を介して I / O サブシステム 2 0 1 1 に結合される。少なくとも一実施例では、I / O サブシステム 2 0 1 1 は、コンピューティング・システム 2 0 0 0 が 1 つ又は複数の入力デバイス 2 0 0 8 からの入力を受け取れるようにすることができる I / O ハブ 2 0 0 7 を含む。少なくとも一実施例では、I / O ハブ 2 0 0 7 は、ディスプレイ・コントローラを有効にすることができ、このディスプレイ・コントローラは、1 つ又は複数のプロセッサ 2 0 0 2 に含まれて、1 つ又は複数のディスプレイ・デバイス 2 0 1 0 A に出力を提供してもよい。少なくとも一実施例では、I / O ハブ 2 0 0 7 に結合された 1 つ又は複数のディスプレイ・デバイス 2 0 1 0 A は、ローカルの、内部の、又は組み込まれたディスプレイ・デバイスを含むことができる。

30

40

【0291】

少なくとも一実施例では、処理サブシステム 2 0 0 1 は、バス又は他の通信リンク 2 0 1 3 を介してメモリ・ハブ 2 0 0 5 に結合された 1 つ又は複数の並列プロセッサ 2 0 1 2 を含む。少なくとも一実施例では、通信リンク 2 0 1 3 は、P C I エクスプレスなどであるがこれに限定されない任意の数の規格に基づく通信リンク技術若しくはプロトコルのう

50

ちの1つを使用することができ、又はベンダー固有の通信インターフェース若しくは通信ファブリックであってもよい。少なくとも一実施例では、1つ又は複数の並列プロセッサ2012は、メニー・インテグレートッド・コア（MIC：many integrated core）プロセッサなど、多数の処理コア及び／又は処理クラスタを含むことのできる、計算に集中した並列又はベクトルの処理システムを形成する。少なくとも一実施例では、いくつか又はすべての並列プロセッサ2012は、グラフィックス処理サブシステムを形成し、このサブシステムは、I／Oハブ2007を介して結合された1つ又は複数のディスプレイ・デバイス2010Aのうちの1つに、ピクセルを出力することができる。少なくとも一実施例では、並列プロセッサ2012はまた、1つ又は複数のディスプレイ・デバイス2010Bへの直接接続を可能にするディスプレイ・コントローラ及びディスプレイ・インターフェース（図示せず）を含むことができる。

10

【0292】

少なくとも一実施例では、システム・ストレージ・ユニット2014は、I／Oハブ2007に接続されて、コンピューティング・システム2000のためのストレージ機構を提供することができる。少なくとも一実施例では、I／Oスイッチ2016を使用して、I／Oハブ2007と、プラットフォームに一体化されてもよいネットワーク・アダプタ2018及び／又はワイヤレス・ネットワーク・アダプタ2019などの他の構成要素、並びに1つ又は複数のアドイン・デバイス2020を介して加えることができる様々な他のデバイスとの通信を可能にするためのインターフェース機構を提供することができる。少なくとも一実施例では、ネットワーク・アダプタ2018は、イーサネット（登録商標）・アダプタ、又は別の有線ネットワーク・アダプタとすることができる。少なくとも一実施例では、ワイヤレス・ネットワーク・アダプタ2019は、Wi-Fi、Bluetooth、近距離無線通信（NFC）、又は1つ若しくは複数のワイヤレス無線を含む他のネットワーク・デバイスのうちの1つ又は複数を含むことができる。

20

【0293】

少なくとも一実施例では、コンピューティング・システム2000は、USB又は他のポート接続、光学ストレージ・ドライブ、ビデオ捕捉デバイスなどを含む明示されていない他の構成要素を含むことができ、これらもI／Oハブ2007に接続されてもよい。少なくとも一実施例では、図20の様々な構成要素を相互接続する通信経路が、PCI（ペリフェラル・コンポーネント・インターコネクト）ベースのプロトコル（たとえば、PCI-Express）などの任意の好適なプロトコル、又はNV-Link高速相互接続などの他のバス若しくはポイントツーポイント通信インターフェース、又は他の相互接続プロトコルを使用して、実装されてもよい。

30

【0294】

少なくとも一実施例では、並列プロセッサ2012は、たとえばビデオ出力回路を含むグラフィックス及びビデオの処理に最適化された回路を組み込んでおり、グラフィックス・プロセッシング・ユニット（GPU）を構成する。少なくとも一実施例では、並列プロセッサ2012は、汎用処理に最適化された回路を組み込んでいる。少なくとも一実施例では、コンピューティング・システム2000の構成要素は、単一の集積回路上の1つ又は複数の他のシステム要素と一体化されてもよい。たとえば、少なくとも一実施例では、並列プロセッサ2012、メモリ・ハブ2005、プロセッサ2002、及びI／Oハブ2007を、システム・オン・チップ（SoC）集積回路に一体化することができる。少なくとも一実施例では、コンピューティング・システム2000の構成要素は、単一のパッケージに一体化されて、システム・イン・パッケージ（SIP：system in package）構成を形成することができる。少なくとも一実施例では、コンピューティング・システム2000の構成要素の少なくとも一部分を、マルチ・チップ・モジュール（MCM：multi-chip module）に一体化することができ、このモジュールを、他のマルチ・チップ・モジュールと相互接続して、モジュール式コンピューティング・システムにすることができる。

40

【0295】

50

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理 8 1 5 が使用される。推論及び／又は訓練論理 8 1 5 に関する詳細事項は、図 8 A 及び／又は図 8 B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 8 1 5 は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のために図 2 0 0 0 のシステムにおいて使用されてもよい。

【0296】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0297】

プロセッサ

図 2 1 A は、少なくとも一実施例による並列プロセッサ 2 1 0 0 を示す。少なくとも一実施例では、並列プロセッサ 2 1 0 0 の様々な構成要素は、プログラム可能なプロセッサ、特定用途向け集積回路 (ASIC)、又はフィールド・プログラマブル・ゲート・アレイ (FPGA) などの 1 つ又は複数の集積回路デバイスを使用して実装されてもよい。少なくとも一実施例では、図示してある並列プロセッサ 2 1 0 0 は、例示的な実施例による図 2 0 に示す 1 つ又は複数の並列プロセッサ 2 0 1 2 の変形形態である。

【0298】

少なくとも一実施例では、並列プロセッサ 2 1 0 0 は並列処理ユニット 2 1 0 2 を含む。少なくとも一実施例では、並列処理ユニット 2 1 0 2 は、並列処理ユニット 2 1 0 2 の他のインスタンスを含む他のデバイスとの通信を可能にする I/O ユニット 2 1 0 4 を含む。少なくとも一実施例では、I/O ユニット 2 1 0 4 は、他のデバイスに直接接続されてもよい。少なくとも一実施例では、I/O ユニット 2 1 0 4 は、メモリ・ハブ 2 1 0 5 などのハブ又はスイッチ・インターフェースの使用を介して、他のデバイスと接続される。少なくとも一実施例では、メモリ・ハブ 2 1 0 5 と I/O ユニット 2 1 0 4 との間の接続は、通信リンク 2 1 1 3 を形成する。少なくとも一実施例では、I/O ユニット 2 1 0 4 は、ホスト・インターフェース 2 1 0 6 及びメモリ・クロスバー 2 1 1 6 に接続され、ここでホスト・インターフェース 2 1 0 6 は、処理動作の実行を対象とするコマンドを受け取り、メモリ・クロスバー 2 1 1 6 は、メモリ動作の実行を対象とするコマンドを受け取る。

【0299】

少なくとも一実施例では、ホスト・インターフェース 2 1 0 6 が、I/O ユニット 2 1 0 4 を介してコマンド・バッファを受け取るとき、ホスト・インターフェース 2 1 0 6 は、これらのコマンドを実行するためのワーク動作をフロント・エンド 2 1 0 8 に向けることができる。少なくとも一実施例では、フロント・エンド 2 1 0 8 はスケジューラ 2 1 1 0 に結合され、このスケジューラは、コマンド又は他のワーク・アイテムを処理クラスタ・アレイ 2 1 1 2 に分配するように構成される。少なくとも一実施例では、スケジューラ 2 1 1 0 は、処理クラスタ・アレイ 2 1 1 2 のクラスタにタスクが分配される前に、処理クラスタ・アレイ 2 1 1 2 が適切に構成され、有効な状態にあることを確実にする。少なくとも一実施例では、スケジューラ 2 1 1 0 は、マイクロコントローラで実行しているファームウェア論理を介して実装される。少なくとも一実施例では、マイクロコントローラ実装スケジューラ 2 1 1 0 は、複雑なスケジューリング及びワーク分配動作を、粗い粒度

と細かい粒度で実行するように構成可能であり、処理アレイ 2 1 1 2 で実行しているスレッドの迅速なプリエンプション及びコンテキストのスイッチングを可能にする。少なくとも一実施例では、ホスト・ソフトウェアは、処理クラスタ・アレイ 2 1 1 2 でのスケジューリングのワークロードを、複数のグラフィックス処理の経路のうちの 1 つを介して証明することができる。少なくとも一実施例では、次いで、スケジューラ 2 1 1 0 を含むマイクロコントローラ内のスケジューラ 2 1 1 0 論理によって、ワークロードを自動的に処理クラスタ・アレイ 2 1 1 2 全体に分配することができる。

【0300】

少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 は、最大「N 個」の処理クラスタ（たとえば、クラスタ 2 1 1 4 A、クラスタ 2 1 1 4 B～クラスタ 2 1 1 4 N）を含むことができ、ここで「N」は、正の整数を表す（他の図で使用されるものとは異なる整数「N」であってもよい）。少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 の各クラスタ 2 1 1 4 A～2 1 1 4 N は、大量の同時スレッドを実行することができる。少なくとも一実施例では、スケジューラ 2 1 1 0 は、様々なスケジューリング及び／又はワーク分配のアルゴリズムを使用して、処理クラスタ・アレイ 2 1 1 2 のクラスタ 2 1 1 4 A～2 1 1 4 N にワークを配分することができ、これらのアルゴリズムは、プログラム又は計算のタイプごとに生じるワークロードに応じて、異なってもよい。少なくとも一実施例では、スケジューリングは、スケジューラ 2 1 1 0 によって動的に対処されてもよく、又は処理クラスタ・アレイ 2 1 1 2 によって実行されるように構成されたプログラム論理のコンパイル中に、コンパイラ論理によって部分的に支援されてもよい。少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 の異なるクラスタ 2 1 1 4 A～2 1 1 4 N は、異なるタイプのプログラムを処理するように、又は異なるタイプの計算を実行するように配分されることが可能である。

【0301】

少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 は、様々なタイプの並列処理動作を実行するように構成されることが可能である。少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 は、汎用の並列コンピュータ動作を実行するように構成される。たとえば、少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 は、ビデオ及び／又はオーディオ・データのフィルタリング、物理動作を含むモデリング動作の実行、及びデータ変換の実行を含む処理タスクを実行するための論理を含むことができる。

【0302】

少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 は、並列グラフィックス処理動作を実行するように構成される。少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 は、テクスチャ動作を実行するためのテクスチャ・サンプリング論理、並びにモザイク論理、及び他の頂点処理論理を含むがこれらに限定されないこうしたグラフィックス処理動作の実行をサポートするための追加の論理を含むことができる。少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 は、頂点シェーダ、モザイク・シェーダ、ジオメトリ・シェーダ、及びピクセル・シェーダなどであるが、これらに限定されないグラフィックス処理関連のシェーダ・プログラムを実行するように構成されることが可能である。少なくとも一実施例では、並列処理ユニット 2 1 0 2 は、処理できるようにデータをシステム・メモリから I/O ユニット 2 1 0 4 を介して転送することができる。少なくとも一実施例では、処理中、転送されたデータを、処理中にオン・チップ・メモリ（たとえば、並列プロセッサ・メモリ 2 1 2 2）に記憶し、次いでシステム・メモリに書き戻すことができる。

【0303】

少なくとも一実施例では、並列処理ユニット 2 1 0 2 を使用してグラフィックス処理が実行される場合には、処理クラスタ・アレイ 2 1 1 2 の複数のクラスタ 2 1 1 4 A～2 1 1 4 N にグラフィックス処理動作をよりうまく分配できるようにするため、処理ワークロードをおおよそ等しい大きさのタスクに分割するようにスケジューラ 2 1 1 0 を構成することができる。少なくとも一実施例では、処理クラスタ・アレイ 2 1 1 2 の一部分は、異

なるタイプの処理を実行するように構成されることが可能である。たとえば、少なくとも一実施例では、レンダリング画像を生成して表示するために、第1の部分は、頂点シェーディング及びトポロジ生成を実行するように構成されてもよく、第2の部分は、モザイク及びジオメトリのシェーディングを実行するように構成されてもよく、第3の部分は、ピクセル・シェーディング又は他の画面空間動作を実行するように構成されてもよい。少なくとも一実施例では、クラスタ2114A~2114Nのうちの1つ又は複数によって生成される中間データをバッファに記憶して、さらなる処理ができるようにクラスタ2114A~2114Nの間で中間データを送信できるようにしてもよい。

【0304】

少なくとも一実施例では、処理クラスタ・アレイ2112は、実行される処理タスクをスケジューラ2110を介して受け取ることができ、スケジューラ2110は、処理タスクを定義するコマンドをフロント・エンド2108から受け取る。少なくとも一実施例では、処理タスクは、処理されるデータのインデックス、たとえば、表面（パッチ）データ、プリミティブ・データ、頂点データ、及び／又はピクセル・データ、並びに状態パラメータ、及びデータをどのように処理すべきかを定義するコマンド（たとえば、どのプログラムを実行すべきか）を含むことができる。少なくとも一実施例では、スケジューラ2110は、タスクに対応するインデックスをフェッチするように構成されてもよく、又はフロント・エンド2108からインデックスを受け取ってもよい。少なくとも一実施例では、フロント・エンド2108は、入ってくるコマンド・バッファ（たとえば、バッチ・バッファ、プッシュ・バッファなど）によって指定されるワークロードが開始される前に、処理クラスタ・アレイ2112が有効な状態に構成されていることを保証するように構成されることが可能である。

【0305】

少なくとも一実施例では、並列処理ユニット2102の1つ又は複数のインスタンスのそれぞれは、並列プロセッサ・メモリ2122と結合することができる。少なくとも一実施例では、並列プロセッサ・メモリ2122には、メモリ・クロスバー2116を介してアクセスすることができ、メモリ・クロスバー2116は、処理クラスタ・アレイ2112並びにI/Oユニット2104からメモリ要求を受け取ることができる。少なくとも一実施例では、メモリ・クロスバー2116は、メモリ・インターフェース2118を介して並列プロセッサ・メモリ2122にアクセスすることができる。少なくとも一実施例では、メモリ・インターフェース2118は、複数のパーティション・ユニット（たとえば、パーティション・ユニット2120A、パーティション・ユニット2120B~パーティション・ユニット2120N）を含むことができ、これらのユニットはそれぞれ、並列プロセッサ・メモリ2122の一部（たとえば、メモリ・ユニット）に結合することができる。少なくとも一実施例では、パーティション・ユニット2120A~2120Nの数は、メモリ・ユニットの数と等しくなるように構成され、それにより、第1のパーティション・ユニット2120Aは、対応する第1のメモリ・ユニット2124Aを有し、第2のパーティション・ユニット2120Bは、対応するメモリ・ユニット2124Bを有し、N番目のパーティション・ユニット2120Nは、対応するN番目のメモリ・ユニット2124Nを有する。少なくとも一実施例では、パーティション・ユニット2120A~2120Nの数は、メモリ・デバイスの数に等しくなくてもよい。

【0306】

少なくとも一実施例では、メモリ・ユニット2124A~2124Nは、グラフィックス・ダブル・データ・レート（GDDR）メモリを含む同期グラフィックス・ランダム・アクセス・メモリ（SGRAM）など、ダイナミック・ランダム・アクセス・メモリ（DRAM）又はグラフィックス・ランダム・アクセス・メモリを含む、様々なタイプのメモリ・デバイスを含むことができる。少なくとも一実施例では、またメモリ・ユニット2124A~2124Nはまた、高帯域幅メモリ（HBM）を含むがこれに限定されない3D積層メモリを含んでもよい。少なくとも一実施例では、並列プロセッサ・メモリ2122の利用可能な帯域幅を効率的に使用するために、フレーム・バッファ又はテクスチャ・マ

10

20

30

40

50

ップなどのレンダー・ターゲットが、メモリ・ユニット2124A～2124Nにわたって記憶されて、パーティション・ユニット2120A～2120Nが、各レンダー・ターゲットの部分を並列に書き込みできるようにしてもよい。少なくとも一実施例では、システム・メモリとローカル・キャッシュ・メモリを併用する統合メモリ設計に有利なように、並列プロセッサ・メモリ2122のローカル・インスタンスは除外されてもよい。

【0307】

少なくとも一実施例では、処理クラスタ・アレイ2112のクラスタ2114A～2114Nのうちのいずれか1つは、並列プロセッサ・メモリ2122内のメモリ・ユニット2124A～2124Nのいずれかに書き込まれることになるデータを処理することができる。少なくとも一実施例では、メモリ・クロスバー2116は、各クラスタ2114A～2114Nの出力を、出力に対してさらなる処理動作を実行することができる任意のパーティション・ユニット2120A～2120N、又は別のクラスタ2114A～2114Nに転送するように構成されることが可能である。少なくとも一実施例では、各クラスタ2114A～2114Nは、メモリ・クロスバー2116を通してメモリ・インターフェース2118と通信して、様々な外部メモリ・デバイスからの読取り、又はそれへの書込みを行うことができる。少なくとも一実施例では、メモリ・クロスバー2116は、I/Oユニット2104と通信するためのメモリ・インターフェース2118への接続部、並びに並列プロセッサ・メモリ2122のローカル・インスタンスへの接続部を有して、異なる処理クラスタ2114A～2114N内の処理ユニットが、システム・メモリ、又は並列処理ユニット2102のローカルにない他のメモリと通信できるようにする。少なくとも一実施例では、メモリ・クロスバー2116は、仮想チャネルを使用して、クラスタ2114A～2114Nと、パーティション・ユニット2120A～2120Nとの間でトラフィック・ストリームを分離することができる。

【0308】

少なくとも一実施例では、並列処理ユニット2102の複数のインスタンスは、単一のアドイン・カードに提供されてもよく、又は複数のアドイン・カードが相互接続されてもよい。少なくとも一実施例では、異なるインスタンスが異なる数の処理コア、異なる量のローカル並列プロセッサ・メモリ、及び／又は他の異なる構成を有する場合でも、並列処理ユニット2102の異なるインスタンスは相互動作するように構成されることが可能である。たとえば、少なくとも一実施例では、並列処理ユニット2102のいくつかのインスタンスは、他のインスタンスに比べて高い精度の浮動小数点ユニットを含むことができる。少なくとも一実施例では、並列処理ユニット2102又は並列プロセッサ2100のうちの1つ又は複数のインスタンスを組み込んだシステムは、デスクトップ、ラップトップ、若しくは携帯型のパーソナル・コンピュータ、サーバ、ワークステーション、ゲーム・コンソール、及び／又は組み込みシステムを含むが、これらに限定されない様々な構成及びフォーム・ファクタで実装することができる。

【0309】

図21Bは、少なくとも一実施例によるパーティション・ユニット2120のブロック図である。少なくとも一実施例では、パーティション・ユニット2120は、図21Aのパーティション・ユニット2120A～2120Nのうちの1つのパーティション・ユニットのインスタンスである。少なくとも一実施例では、パーティション・ユニット2120は、L2キャッシュ2121、フレーム・バッファ・インターフェース2125、及びROP: raster operations unit 2126 (ラスタ演算ユニット)を含む。少なくとも一実施例では、L2キャッシュ2121は、メモリ・クロスバー2116及びROP 2126から受け取ったロード及びストアの動作を実行するように構成された読取り／書込みキャッシュである。少なくとも一実施例では、読取りミス及び至急の書戻し要求が、処理されるようにL2キャッシュ2121によってフレーム・バッファ・インターフェース2125に出力される。少なくとも一実施例では、更新も、処理されるようにフレーム・バッファ・インターフェース2125を介してフレームに送られる。少なくとも一実施例では、フレーム・バッファ・インターフェース2125は、図21の(

たとえば並列プロセッサ・メモリ 2122 内の) メモリ・ユニット 2124A~2124N など、並列プロセッサ・メモリのメモリ・ユニットのうちの 1 つとインターフェースをとる。

【0310】

少なくとも一実施例では、ROP 2126 は、ステンシル、z テスト、ブレンディングなどのラスタ演算を実行する処理ユニットである。少なくとも一実施例では、次いで ROP 2126 は、グラフィックス・メモリに記憶された処理済みグラフィックス・データを出力する。少なくとも一実施例では、ROP 2126 は、メモリに書き込まれる深度又は色データを圧縮し、メモリから読み取られた深度又は色データを解凍するための圧縮論理を含む。少なくとも一実施例では、圧縮論理は、複数の圧縮アルゴリズムのうちの 1 つ又は複数を利用するロスレス圧縮論理とすることができる。少なくとも一実施例では、ROP 2126 によって実行される圧縮のタイプは、圧縮されるデータの統計的特徴に基づき変更することができる。たとえば、少なくとも一実施例では、深度及び色データに対してはタイルごとにデルタ色圧縮が実行される。

【0311】

少なくとも一実施例では、ROP 2126 は、パーティション・ユニット 2120 内ではなく、各処理クラスタ内（たとえば、図 21A のクラスタ 2114A~2114N）に含まれる。少なくとも一実施例では、ピクセル・フラグメント・データではなく、ピクセル・データの読取り及び書き込み要求が、メモリ・クロスバー 2116 を介して送信される。少なくとも一実施例では、処理済みグラフィックス・データは、図 20 の 1 つ又は複数のディスプレイ・デバイス 2010 のうちの 1 つなどのディスプレイ・デバイスに表示されてもよく、プロセッサ 2002 によってさらに処理できるようにルーティングされてもよく、又は図 21A の並列プロセッサ 2100 内の処理エンティティのうちの 1 つによってさらに処理できるようにルーティングされてもよい。

【0312】

図 21C は、少なくとも一実施例による並列処理ユニット内の処理クラスタ 2114 のブロック図である。少なくとも一実施例では、処理クラスタは、図 21A の処理クラスタ 2114A~2114N のうちの 1 つの処理クラスタのインスタンスである。少なくとも一実施例では、処理クラスタ 2114 は、多数のスレッドを並列で実行するように構成されてもよく、ここで「スレッド」とは、入力データの特定のセットに対して実行している特定のプログラムのインスタンスを指す。少なくとも一実施例では、複数の独立した命令ユニットを提供することなく、多数のスレッドの並列実行をサポートするために、単一命令複数データ (SIMD) の命令発行技法が使用される。少なくとも一実施例では、それぞれの処理クラスタ内の処理エンジンのセットに命令を発行するように構成された共通の命令ユニットを使用して、全体的に同期された多数のスレッドの並列実行をサポートするために、単一命令複数スレッド (SIMT: single-instruction, multiple-thread) の技法が使用される。

【0313】

少なくとも一実施例では、処理クラスタ 2114 の動作は、SIMT 並列プロセッサに処理タスクを分配するパイプライン・マネージャ 2132 を介して制御することができる。少なくとも一実施例では、パイプライン・マネージャ 2132 は、図 21A のスケジューラ 2110 から命令を受け取り、グラフィックス・マルチプロセッサ 2134 及び／又はテクスチャ・ユニット 2136 を介してこれらの命令の実行を管理する。少なくとも一実施例では、グラフィックス・マルチプロセッサ 2134 は、SIMT 並列プロセッサの例示的なインスタンスである。しかし、少なくとも一実施例では、アーキテクチャの異なる様々なタイプの SIMT 並列プロセッサが、処理クラスタ 2114 内に含まれてもよい。少なくとも一実施例では、グラフィックス・マルチプロセッサ 2134 の 1 つ又は複数のインスタンスは、処理クラスタ 2114 内に含めることができる。少なくとも一実施例では、グラフィックス・マルチプロセッサ 2134 はデータを処理することができ、処理済みデータを、他のシェーダ・ユニットを含む複数の可能な宛先のうちの 1 つに分配する

ためにデータ・クロスバー 2 1 4 0 が使用されてもよい。少なくとも一実施例では、パイプライン・マネージャ 2 1 3 2 は、データ・クロスバー 2 1 4 0 を通して分配されることになる処理済みデータの宛先を指定することによって、処理済みデータの分配を容易にすることができる。

【0 3 1 4】

少なくとも一実施例では、処理クラスタ 2 1 1 4 内の各グラフィックス・マルチプロセッサ 2 1 3 4 は、関数実行論理（たとえば、算術論理演算ユニット、ロード・ストア・ユニットなど）の同一のセットを含むことができる。少なくとも一実施例では、関数実行論理は、前の命令が完了する前に新規の命令を発行することができるパイプライン式に構成されることが可能である。少なくとも一実施例では、関数実行論理は、整数及び浮動小数点の算術、比較演算、ブール演算、ビット・シフト、及び様々な代数関数の計算を含む様々な演算をサポートする。少なくとも一実施例では、同じ関数ユニットのハードウェアを活用して、異なる演算を実行することができ、関数ユニットの任意の組合せが存在してもよい。

【0 3 1 5】

少なくとも一実施例では、処理クラスタ 2 1 1 4 に送信される命令がスレッドを構成する。少なくとも一実施例では、並列処理エンジンのセットにわたって実行されているスレッドのセットが、スレッド・グループである。少なくとも一実施例では、スレッド・グループは、異なる入力データに対して共通のプログラムを実行する。少なくとも一実施例では、スレッド・グループ内の各スレッドを、グラフィックス・マルチプロセッサ 2 1 3 4 内の異なる処理エンジンに割り当てることができる。少なくとも一実施例では、スレッド・グループは、グラフィックス・マルチプロセッサ 2 1 3 4 内の処理エンジンの数よりも少ないスレッドを含んでもよい。少なくとも一実施例では、スレッド・グループが処理エンジンの数よりも少ないスレッドを含む場合、処理エンジンのうちの 1 つ又は複数は、そのスレッド・グループが処理されているサイクル中にはアイドルであってもよい。少なくとも一実施例では、スレッド・グループはまた、グラフィックス・マルチプロセッサ 2 1 3 4 内の処理エンジンの数よりも多いスレッドを含んでもよい。少なくとも一実施例では、スレッド・グループがグラフィックス・マルチプロセッサ 2 1 3 4 内の処理エンジンの数より多くのスレッドを含む場合には、連続したクロック・サイクルにわたって処理を実行することができる。少なくとも一実施例では、複数のスレッド・グループを、グラフィックス・マルチプロセッサ 2 1 3 4 上で同時に実行することができる。

【0 3 1 6】

少なくとも一実施例では、グラフィックス・マルチプロセッサ 2 1 3 4 は、ロード及びストアの動作を実行するための内部キャッシュ・メモリを含む。少なくとも一実施例では、グラフィックス・マルチプロセッサ 2 1 3 4 は、内部キャッシュをやめて、処理クラスタ 2 1 1 4 内のキャッシュ・メモリ（たとえば、L 1 キャッシュ 2 1 4 8）を使用することができる。少なくとも一実施例では、各グラフィックス・マルチプロセッサ 2 1 3 4 は、パーティション・ユニット（たとえば、図 2 1 A のパーティション・ユニット 2 1 2 0 A ~ 2 1 2 0 N）内の L 2 キャッシュにもアクセスすることができ、これらのキャッシュが、すべての処理クラスタ 2 1 1 4 間で共有され、スレッド間でデータを転送するために使用されてもよい。少なくとも一実施例では、グラフィックス・マルチプロセッサ 2 1 3 4 は、オフ・チップのグローバル・メモリにもアクセスすることができ、このメモリは、ローカル並列プロセッサ・メモリ及び／又はシステム・メモリのうちの 1 つ又は複数を含むことができる。少なくとも一実施例では、並列処理ユニット 2 1 0 2 の外部にある任意のメモリが、グローバル・メモリとして使用されてもよい。少なくとも一実施例では、処理クラスタ 2 1 1 4 は、グラフィックス・マルチプロセッサ 2 1 3 4 の複数のインスタンスを含み、共通の命令及びデータを共有することができ、これらは L 1 キャッシュ 2 1 4 8 に記憶されてもよい。

【0 3 1 7】

少なくとも一実施例では、各処理クラスタ 2 1 1 4 は、仮想アドレスを物理アドレスに

マッピングするように構成されたMMU 2145（メモリ管理ユニット）を含んでもよい。少なくとも一実施例では、MMU 2145の1つ又は複数のインスタンスは、図21Aのメモリ・インターフェース2118内にあってもよい。少なくとも一実施例では、MMU 2145は、仮想アドレスを、タイル及び任意選択でキャッシュ・ライン・インデックスの物理アドレスにマッピングするために使用されるページ・テーブル・エントリ（PTE）のセットを含む。少なくとも一実施例では、MMU 2145は、アドレスのトランスレーション・ルックアサイド・バッファ（TLB: translation look aside buffer）又はキャッシュを含んでもよく、これらは、グラフィックス・マルチプロセッサ2134若しくはL1 2148キャッシュ、又は処理クラスタ2114内にあってもよい。少なくとも一実施例では、表面データ・アクセスをローカルに分散するように物理アドレスを処理して、パーティション・ユニット間で要求の効率的なインターリーブが可能になる。少なくとも一実施例では、キャッシュ・ライン・インデックスを使用して、キャッシュ・ラインの要求がヒットかミスかが判定されてもよい。

【0318】

少なくとも一実施例では、各グラフィックス・マルチプロセッサ2134がテクスチャ・ユニット2136に結合されて、テクスチャ・マッピング動作、たとえば、テクスチャ・サンプル位置の判定、テクスチャ・データの読取り、及びテクスチャ・データのフィルタリングが実行されるように、処理クラスタ2114が構成されてもよい。少なくとも一実施例では、テクスチャ・データは、内部テクスチャL1キャッシュ（図示せず）から、又はグラフィックス・マルチプロセッサ2134内のL1キャッシュから読み取られ、必要に応じて、L2キャッシュ、ローカル並列プロセッサ・メモリ、又はシステム・メモリからフェッチされる。少なくとも一実施例では、各グラフィックス・マルチプロセッサ2134は、処理済みタスクをデータ・クロスバー2140に出力して、さらなる処理ができるように別の処理クラスタ2114に処理済みタスクを提供し、又はメモリ・クロスバー2116を介して、L2キャッシュ、ローカル並列プロセッサ・メモリ、又はシステム・メモリに処理済みタスクを記憶する。少なくとも一実施例では、プレROP 2142（プレ・ラスタ演算ユニット）は、グラフィックス・マルチプロセッサ2134からデータを受け取り、ROPユニットにデータを仕向けるように構成されており、ROPユニットは、本明細書に記載のするように、パーティション・ユニット（たとえば、図21Aのパーティション・ユニット2120A～2120N）内に位置付けられてもよい。少なくとも一実施例では、プレROP 2142ユニットは、色ブレンディングの最適化を実行し、ピクセル色データを組織化し、アドレス・トランスレーションを実行することができる。

【0319】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のためにグラフィックス処理クラスタ2114において使用されてもよい。

【0320】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0321】

図21Dは、少なくとも一実施例によるグラフィックス・マルチプロセッサ2134を示す。少なくとも一実施例では、グラフィックス・マルチプロセッサ2134は、処理クラスタ2114のパイプライン・マネージャ2132と結合する。少なくとも一実施例では、グラフィックス・マルチプロセッサ2134は、命令キャッシュ2152、命令ユニット2154、アドレス・マッピング・ユニット2156、レジスタ・ファイル2158、1つ又は複数の汎用グラフィックス・プロセッシング・ユニット（GP GPU）コア2162、及び1つ又は複数のロード／ストア・ユニット2166を含むがこれらに限定されない実行パイプラインを有する。少なくとも一実施例では、GP GPUコア2162、及びロード／ストア・ユニット2166は、メモリ及びキャッシュ相互接続2168を介して、キャッシュ・メモリ2172及び共有メモリ2170に結合される。

10

【0322】

少なくとも一実施例では、命令キャッシュ2152は、実行すべき命令のストリームをパイプライン・マネージャ2132から受け取る。少なくとも一実施例では、命令は、命令キャッシュ2152にキャッシュされ、命令ユニット2154により実行されるようにディスパッチされる。少なくとも一実施例では、命令ユニット2154は、命令をスレッド・グループ（たとえば、ワーブ）としてディスパッチすることができ、アスレッド・グループの各スレッドは、GP GPUコア2162内の異なる実行ユニットに割り当てられる。少なくとも一実施例では、命令は、統一アドレス空間内のアドレスを指定することによって、ローカル、共有、又はグローバルのアドレス空間のいずれかにアクセスすることができる。少なくとも一実施例では、アドレス・マッピング・ユニット2156を使用して、統一アドレス空間のアドレスを、ロード／ストア・ユニット2166がアクセスできる個別メモリ・アドレスにトランスレーションすることができる。

20

【0323】

少なくとも一実施例では、レジスタ・ファイル2158は、グラフィックス・マルチプロセッサ2134の機能ユニットにレジスタのセットを提供する。少なくとも一実施例では、レジスタ・ファイル2158は、グラフィックス・マルチプロセッサ2134の機能ユニット（たとえばGP GPUコア2162、ロード／ストア・ユニット2166）のデータ経路に接続された、オペランドのための一時的なストレージを提供する。少なくとも一実施例では、レジスタ・ファイル2158は、レジスタ・ファイル2158の専用部分に各機能ユニットが配分されるように、それぞれの機能ユニット間で分割される。一実施例では、レジスタ・ファイル2158は、グラフィックス・マルチプロセッサ2134によって実行されている異なるワーブ間で分割される。

30

【0324】

少なくとも一実施例では、GP GPUコア2162はそれぞれ、グラフィックス・マルチプロセッサ2134の命令を実行するために使用される浮動小数点ユニット（FPU）及び／又は整数算術論理演算ユニット（ALU）を含むことができる。少なくとも一実施例では、GP GPUコア2162同士は、同様のアーキテクチャであってもよく、又は異なるアーキテクチャであってもよい。少なくとも一実施例では、GP GPUコア2162の第1の部分は、単精度FPU及び整数ALUを含み、GP GPUコアの第2の部分は、倍精度FPUを含む。少なくとも一実施例では、FPUは、IEEE 754-2008規格浮動小数点演算を実装することができ、又は、可変精度の浮動小数点演算を有効にすることができる。少なくとも一実施例では、グラフィックス・マルチプロセッサ2134はさらに、矩形コピー又はピクセル・ブレンディングの動作などの特定の機能を実行するための、1つ若しくは複数の固定機能ユニット又は特別機能ユニットをさらに含むことができる。少なくとも一実施例では、GP GPUコア2162の1つ又は複数は、固定の又は特別な機能論理も含むことができる。

40

【0325】

少なくとも一実施例では、GP GPUコア2162は、複数のデータ・セットに対して単一の命令を実行することができるSIMD論理を含む。少なくとも一実施例では、GP

50

GPUコア2162は、SIMD4、SIMD8、及びSIMD16の命令を物理的に実行することができ、SIMD1、SIMD2、及びSIMD32の命令を論理的に実行することができる。少なくとも一実施例では、GPGPUコアのためのSIMD命令は、シェーダ・コンパイラによるコンパイル時に生成されてもよく、又は単一プログラム複数データ(SPMD: single program multiple data)又はSIMTのアーキテクチャ向けに書かれコンパイルされたプログラムを実行しているときに、自動的に生成されてもよい。少なくとも一実施例では、SIMT実行モデルのために構成されたプログラムの複数のスレッドは、単一のSIMD命令を介して実行することができる。たとえば、少なくとも一実施例では、同じ又は同様の動作を実行する8個のSIMTスレッドを、単一のSIMD8の論理ユニットを介して並列に実行することができる。

10

【0326】

少なくとも一実施例では、メモリ及びキャッシュ相互接続2168は、グラフィックス・マルチプロセッサ2134の各機能ユニットをレジスタ・ファイル2158及び共有メモリ2170に接続する相互接続ネットワークである。少なくとも一実施例では、メモリ及びキャッシュ相互接続2168は、ロード／ストア・ユニット2166が、共有メモリ2170とレジスタ・ファイル2158の間でロード及びストアの動作を実装できるようにするクロスバー相互接続である。少なくとも一実施例では、レジスタ・ファイル2158は、GPGPUコア2162と同じ周波数で動作することができ、したがって、GPGPUコア2162とレジスタ・ファイル2158の間のデータ転送は非常に低レイテンシを有し得る。少なくとも一実施例では、共有メモリ2170を使用して、グラフィックス・マルチプロセッサ2134内の機能ユニットで実行されるスレッド間の通信を可能にすることができる。少なくとも一実施例では、キャッシュ・メモリ2172を、たとえばデータ・キャッシュとして使用して、機能ユニットとテクスチャ・ユニット2136の間で通信されるテクスチャ・データをキャッシュすることができる。少なくとも一実施例では、共有メモリ2170は、プログラム管理キャッシュとしても使用することができる。少なくとも一実施例では、GPGPUコア2162で実行されているスレッドは、キャッシュ・メモリ2172内に記憶される自動キャッシュ・データに加えて、共有メモリ内にプログラムのデータを記憶することができる。

20

【0327】

少なくとも一実施例では、本明細書に記載の並列プロセッサ又はGPGPUは、ホスト／プロセッサ・コアに通信可能に結合されて、グラフィックス動作、機械学習動作、パターン分析動作、及び様々な汎用GPU(GPGPU)機能を加速する。少なくとも一実施例では、GPUは、バス又は他の相互接続(たとえば、PCIe又はNVLinkなどの高速相互接続)を介してホスト・プロセッサ／コアに通信可能に結合されてもよい。少なくとも一実施例では、GPUは、コアとしてパッケージ又はチップに一体化されてもよく、パッケージ又はチップの内部の内部プロセッサ・バス／相互接続を介してコアに通信可能に結合されてもよい。少なくとも一実施例では、GPUの接続方法に関わらず、プロセッサ・コアは、ワーク記述子に含まれたコマンド／命令のシーケンスの形でワークをこうしたGPUに配分してもよい。少なくとも一実施例では、次いでGPUは、これらのコマンド／命令を効率的に処理するために専用の回路／論理を使用する。

30

40

【0328】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のためにグラフィックス・マルチプロセッサ2134において使用されてもよい。

【0329】

50

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0330】

図22は、少なくとも一実施例による、マルチGPUコンピューティング・システム2200を示す。少なくとも一実施例では、マルチGPUコンピューティング・システム2200は、ホスト・インターフェース・スイッチ2204を介して複数の汎用グラフィックス・プロセッシング・ユニット(GPGPU)2206A～Dに結合されたプロセッサ2202を含むことができる。少なくとも一実施例では、ホスト・インターフェース・スイッチ2204は、プロセッサ2202をPCIエクスプレス・バスに結合するPCIエクスプレス・スイッチ・デバイスであり、このPCIエクスプレス・バスを介して、プロセッサ2202は、GPGPU2206A～Dと通信することができる。少なくとも一実施例では、GPGPU2206A～Dは、高速ポイントツーポイントGPUツーGPUリンク2216のセットを介して相互接続することができる。少なくとも一実施例では、GPUツーGPUリンク2216は、専用GPUリンクを介して、GPGPU2206A～Dのそれぞれに接続される。少なくとも一実施例では、P2PのGPUリンク2216は、プロセッサ2202が接続されているホスト・インターフェース・バス2204を介した通信を必要とせずに、GPGPU2206A～Dのそれぞれの間で直接通信を可能にする。少なくとも一実施例では、P2PのGPUリンク2216に仕向けられたGPUツーGPUトラフィックがあると、ホスト・インターフェース・バス2204は、システム・メモリへのアクセスができるように、又はたとえば1つ又は複数のネットワーク・デバイスを介して、マルチGPUコンピューティング・システム2200の他のインスタンスと通信するために、利用可能な状態に保たれる。少なくとも一実施例では、GPGPU2206A～Dは、ホスト・インターフェース・スイッチ2204を介してプロセッサ2202に接続され、少なくとも一実施例では、プロセッサ2202は、P2PのGPUリンク2216のための直接サポートを含み、GPGPU2206A～Dに直接接続することができる。

【0331】

1つ又は複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のためにマルチGPUコンピューティング・システム2200において使用されてもよい。

【0332】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

10

20

30

40

50

【0333】

図23は、少なくとも一実施例によるグラフィックス・プロセッサ2300のブロック図である。少なくとも一実施例では、グラフィックス・プロセッサ2300は、リング相互接続2302、パイプライン・フロント・エンド2304、メディア・エンジン2337、及びグラフィックス・コア2380A～2380Nを含む。少なくとも一実施例では、リング相互接続2302は、グラフィックス・プロセッサ2300を、他のグラフィックス・プロセッサ又は1つ又は複数の汎用プロセッサ・コアを含む他の処理ユニットに結合する。少なくとも一実施例では、グラフィックス・プロセッサ2300は、マルチ・コア処理システム内に一体化された多数のプロセッサのうちの1つである。

【0334】

少なくとも一実施例では、グラフィックス・プロセッサ2300は、リング相互接続2302を介してコマンドのバッチを受け取る。少なくとも一実施例では、入ってくるコマンドは、パイプライン・フロント・エンド2304のコマンド・ストリーマ2303によって解釈される。少なくとも一実施例では、グラフィックス・プロセッサ2300は、グラフィックス・コア2380A～2380Nを介して3Dジオメトリ処理及びメディア処理を実行するためのスケラブルな実行論理を含む。少なくとも一実施例では、3Dジオメトリ処理コマンドについては、コマンド・ストリーマ2303はコマンドをジオメトリ・パイプライン2336に供給する。少なくとも一実施例では、少なくとも一部のメディア処理コマンドについては、コマンド・ストリーマ2303はコマンドをビデオ・フロント・エンド2334に供給し、ビデオ・フロント・エンド2334はメディア・エンジン2337に結合される。少なくとも一実施例では、メディア・エンジン2337は、ビデオ及び画像の後処理のためのVideo Quality Engine (VQE) 2330と、ハードウェア加速されたメディア・データのエンコード及びデコードを提供するマルチ・フォーマット・エンコード／デコード(MFX) 2333エンジンとを含む。少なくとも一実施例では、ジオメトリ・パイプライン2336及びメディア・エンジン2337はそれぞれ、少なくとも1つのグラフィックス・コア2380によって提供されるスレッド実行リソースのための実行スレッドを生成する。

【0335】

少なくとも一実施例では、グラフィックス・プロセッサ2300は、グラフィックス・コア2380A～2380N（モジュール式であり得、コア・スライスと呼ばれることもある）を特徴とするスケラブルなスレッド実行リソースを含み、それぞれのグラフィックス・コア2380A～2380Nは、複数のサブ・コア2350A～50N、2360A～2360N（コア・サブ・スライスと呼ばれることもある）を有する。少なくとも一実施例では、グラフィックス・プロセッサ2300は、任意の数のグラフィックス・コア2380Aを有することができる。少なくとも一実施例では、グラフィックス・プロセッサ2300は、少なくとも第1のサブ・コア2350A及び第2のサブ・コア2360Aを有するグラフィックス・コア2380Aを含む。少なくとも一実施例では、グラフィックス・プロセッサ2300は、単一のサブ・コア（たとえば、2350A）を有する低電力プロセッサである。少なくとも一実施例では、グラフィックス・プロセッサ2300は、複数のグラフィックス・コア2380A～2380Nを含み、このそれぞれが、第1のサブ・コア2350A～2350Nのセット、及び第2のサブ・コア2360A～2360Nのセットを含む。少なくとも一実施例では、第1のサブ・コア2350A～2350Nの各サブ・コアは、少なくとも、実行ユニット2352A～2352Nとメディア／テクスチャ・サンプラ2354A～2354Nの第1のセットを含む。少なくとも一実施例では、第2のサブ・コア2360A～2360Nの各サブ・コアは、少なくとも、実行ユニット2362A～2362Nとサンプラ2364A～2364Nの第2のセットを含む。少なくとも一実施例では、各サブ・コア2350A～2350N、2360A～2360Nは、共有リソース2370A～2370Nのセットを共有する。少なくとも一実施例では、共有リソースは、共有キャッシュ・メモリ及びピクセル動作論理を含む。

【0336】

1つ又は複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815は、本明細書に記載のニューラル・ネットワークの訓練動作、ニューラル・ネットワークの機能及び／若しくはアーキテクチャ、又はニューラル・ネットワークのユース・ケースを使用して計算された重みパラメータに少なくとも部分的に基づき、推論又は予測の動作のためにグラフィックス・プロセッサ2300において使用されてもよい。

【0337】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0338】

図24は、少なくとも一実施例による、命令を実行するための論理回路を含んでもよいプロセッサ2400のマイクロ・アーキテクチャを示すブロック図である。少なくとも一実施例では、プロセッサ2400は、x86命令、AMR命令、特定用途向け集積回路（ASIC）用の特別命令などを含む命令を実行してもよい。少なくとも一実施例では、プロセッサ2400は、カリフォルニア州サンタクララのインテルコーポレーションによる、MMX技術で有効化されたマイクロプロセッサ内の64ビット幅MMXTMレジスタなど、パック・データを記憶するためのレジスタを含んでもよい。少なくとも一実施例では、整数形式と浮動小数点形式の両方で利用可能なMMXレジスタは、単一命令複数データ（「SIMD」）及びストリーミングSIMD拡張（「SSE」：streaming SIMD extensions）命令を伴うパック・データ要素で動作してもよい。少なくとも一実施例では、SSE2、SSE3、SSE4、AVX、又はそれ以上（総称して「SSEx」と呼ばれる）の技術に関する128ビット幅のXMMレジスタは、こうしたパック・データのオペランドを保持してもよい。少なくとも一実施例では、プロセッサ2400は、機械学習若しくは深層学習のアルゴリズム、訓練、又は推論を加速するために命令を実行してもよい。

【0339】

少なくとも一実施例では、プロセッサ2400は、実行すべき命令をフェッチし、プロセッサ・パイプラインで後に使用すべき命令を準備するイン・オーダー・フロント・エンド（「フロント・エンド」）2401を含む。少なくとも一実施例では、フロント・エンド2401は、いくつかのユニットを含んでもよい。少なくとも一実施例では、命令プリフェッチャ2426が、メモリから命令をフェッチし、命令デコーダ2428に命令を供給し、命令デコーダが、命令をデコード又は解釈する。たとえば、少なくとも一実施例では、命令デコーダ2428は、受け取った命令を、機械が実行することのできる「マイクロ命令」又は「マイクロ・オペレーション」と呼ばれる（「マイクロ・オプス」又は「uops」とも呼ばれる）1つ又は複数のオペレーションにデコードする。少なくとも一実施例では、命令デコーダ2428は、命令を、オPCODE及び対応するデータ、並びに制御フィールドに構文解析して、これらがマイクロ・アーキテクチャによって使用されて、少なくとも一実施例による動作が実行されてもよい。少なくとも一実施例では、トレース・キャッシュ2430は、デコードされたuopsを、実行できるようにuopキュー2434においてプログラム順のシーケンス又はトレースにアセンブルしてもよい。少なくとも一実施例では、トレース・キャッシュ2430が複雑な命令に遭遇すると、マイクロコードROM2432が、動作の完了に必要なuopsを提供する。

10

20

30

40

50

【0340】

少なくとも一実施例では、単一のマイクロ・オプスに変換できる命令もあれば、全動作を完了するためにいくつかのマイクロ・オプスを必要とする命令もある。少なくとも一実施例では、命令を完了するために5つ以上のマイクロ・オプスが必要な場合、命令デコーダ2428は、マイクロコードROM2432にアクセスして、命令を実行してもよい。少なくとも一実施例では、命令は、命令デコーダ2428において処理できるように、少数のマイクロ・オプスにデコードされてもよい。少なくとも一実施例では、こうした動作を完了するのに多数のマイクロ・オプスが必要な場合には、命令は、マイクロコードROM2432に記憶されてもよい。少なくとも一実施例では、トレース・キャッシュ2430は、少なくとも一実施例によるマイクロコードROM2432からの1つ又は複数の命令を完了するために、エントリ・ポイント・プログラマブル論理アレイ（「PLA」：programmable logic array）を参照して、マイクロコード・シーケンスを読み取るための正しいマイクロ命令ポインタを判定する。少なくとも一実施例では、マイクロコードROM2432が命令のためのマイクロ・オプスのシーケンシングを終了した後、機械のフロント・エンド2401は、トレース・キャッシュ2430からマイクロ・オプスのフェッチを再開してもよい。

10

【0341】

少なくとも一実施例では、アウト・オブ・オーダー実行エンジン（「アウト・オブ・オーダー・エンジン」）2403は、実行できるように命令を準備してもよい。少なくとも一実施例では、アウト・オブ・オーダー実行論理は、命令のフローをなめらかにし、その順序を変更するために多数バッファを有し、命令がパイプラインを下り、実行されるようにスケジューリングされるときに性能を最適化する。少なくとも一実施例では、アウト・オブ・オーダー実行エンジン2403は、限定することなく、アロケータ／レジスタ・リネーマ2440、メモリuopキュー2442、整数／浮動小数点uopキュー2444、メモリ・スケジューラ2446、高速スケジューラ2402、低速／汎用浮動小数点スケジューラ（「低速／汎用FP：floating pointスケジューラ」）2404、及び単純浮動小数点スケジューラ（「単純FPスケジューラ」）2406を含む。少なくとも一実施例では、高速スケジューラ2402、低速／汎用浮動小数点スケジューラ2404、及び単純浮動小数点スケジューラ2406は、本明細書において集合的に「uopスケジューラ2402、2404、2406」とも呼ばれる。少なくとも一実施例では、アロケータ／レジスタ・リネーマ2440は、実行するために各uopが必要とする機械バッファ及びリソースを配分する。少なくとも一実施例では、アロケータ／レジスタ・リネーマ2440は、レジスタ・ファイルへのエントリ時に論理レジスタの名前を変更する。少なくとも一実施例では、アロケータ／レジスタ・リネーマ2440はまた、メモリ・スケジューラ2446及びuopスケジューラ2402、2404、2406の前の、2つのuopキュー、すなわちメモリ動作のためのメモリuopキュー2442と非メモリ動作のための整数／浮動小数点uopキュー2444のうちの1つに、各uopのエントリを配分する。少なくとも一実施例では、uopスケジューラ2402、2404、2406は、uopsがいつ実行準備されるかを、それらの従属入力レジスタ・オペランドのソースが準備されていること、及びそれらの動作を完了するためにuopが必要とする実行リソースが利用可能であることに基づき、判定する。少なくとも一実施例では、高速スケジューラ2402は、メイン・クロック・サイクルの半分ごとにスケジューリングしてもよく、低速／汎用浮動小数点スケジューラ2404及び単純浮動小数点スケジューラ2406は、メイン・プロセッサのクロック・サイクルあたりに1回スケジューリングしてもよい。少なくとも一実施例では、uopスケジューラ2402、2404、2406は、実行できるようにuopsをスケジューリングするためにディスパッチ・ポートを調停する。

20

30

40

【0342】

少なくとも一実施例では、実行ブロック2411は、限定することなく、整数レジスタ・ファイル／バイパス・ネットワーク2408、浮動小数点レジスタ・ファイル／バイパ

50

ス・ネットワーク（「F Pレジスタ・ファイル／バイパス・ネットワーク」）2410、アドレス生成ユニット（「AGU」：address generation units）2412及び2414、高速算術論理演算ユニット（ALU）（「高速ALU」）2416及び2418、低速算術論理演算ユニット（「低速ALU」）2420、浮動小数点ALU（「FP」）2422、並びに浮動小数点移動ユニット（「FP移動」）2424を含む。少なくとも一実施例では、整数レジスタ・ファイル／バイパス・ネットワーク2408及び浮動小数点レジスタ・ファイル／バイパス・ネットワーク2410は、本明細書において「レジスタ・ファイル2408、2410」とも呼ばれる。少なくとも一実施例では、AGU2412及び2414、高速ALU2416及び2418、低速ALU2420、浮動小数点ALU2422、及び浮動小数点移動ユニット2424は、本明細書において「実行ユニット2412、2414、2416、2418、2420、2422、及び2424」とも呼ばれる。少なくとも一実施例では、実行ブロック2411は、限定することなく、（ゼロを含む）任意の数及びタイプのレジスタ・ファイル、バイパス・ネットワーク、アドレス生成ユニット、及び実行ユニットを、任意の組合せで含んでもよい。

【0343】

少なくとも一実施例では、レジスタ・ネットワーク2408、2410は、uopスケジューラ2402、2404、2406と、実行ユニット2412、2414、2416、2418、2420、2422、及び2424との間に配置されてもよい。少なくとも一実施例では、整数レジスタ・ファイル／バイパス・ネットワーク2408は、整数演算を実行する。少なくとも一実施例では、浮動小数点レジスタ・ファイル／バイパス・ネットワーク2410は、浮動小数点演算を実行する。少なくとも一実施例では、レジスタ・ネットワーク2408、2410のそれぞれは、限定することなく、バイパス・ネットワークを含んでもよく、このバイパス・ネットワークは、レジスタ・ファイルにまだ書き込まれていない完了したばかりの結果を、新しい従属uopsにバイパス又は転送してもよい。少なくとも一実施例では、レジスタ・ネットワーク2408、2410は、互いにデータを通信してもよい。少なくとも一実施例では、整数レジスタ・ファイル／バイパス・ネットワーク2408は、限定することなく、2つの別々のレジスタ・ファイル、すなわち低次32ビットのデータ用の1つのレジスタ・ファイル、及び高次32ビットのデータ用の第2のレジスタ・ファイルを含んでもよい。少なくとも一実施例では、浮動小数点命令は、通常、64～128ビット幅のオペランドを有することから、浮動小数点レジスタ・ファイル／バイパス・ネットワーク2410は、限定することなく、128ビット幅のエントリを含んでもよい。

【0344】

少なくとも一実施例では、実行ユニット2412、2414、2416、2418、2420、2422、2424は、命令を実行してもよい。少なくとも一実施例では、レジスタ・ネットワーク2408、2410は、マイクロ命令が実行する必要がある整数及び浮動小数点のデータのオペランド値を記憶する。少なくとも一実施例では、プロセッサ2400は、限定することなく、任意の数及び組合せの実行ユニット2412、2414、2416、2418、2420、2422、2424を含んでよい。少なくとも一実施例では、浮動小数点ALU2422及び浮動小数点移動ユニット2424は、浮動小数点、MMX、SIMD、AVX、及びSEE、又は特別な機械学習命令を含む他の演算を実行してもよい。少なくとも一実施例では、浮動小数点ALU2422は、限定することなく、64ビットずつの浮動小数点デバイダを含み、除算、平方根、及び残りのマイクロ・オプスを実行してもよい。少なくとも一実施例では、浮動小数点値を含む命令は、浮動小数点ハードウェアによって対処されてもよい。少なくとも一実施例では、ALU演算は、高速ALU2416、2418に渡されてもよい。少なくとも一実施例では、高速ALU2416、2418は、クロック・サイクルの半分の実効レイテンシで高速演算を実行してもよい。少なくとも一実施例では、低速ALU2420は、乗数、シフト、フラグ論理、及びブランチ処理などの長レイテンシ・タイプの演算のための整数実行ハードウェアを、

限定することなく含んでもよいことから、ほとんどの複雑な整数演算は低速 A L U 2 4 2 0 に進む。少なくとも一実施例では、メモリのロード／ストア動作は、A G U S 2 4 1 2、2 4 1 4 によって実行されてもよい。少なくとも一実施例では、高速 A L U 2 4 1 6、高速 A L U 2 4 1 8、及び低速 A L U 2 4 2 0 は、6 4 ビットのデータ・オペランドで整数演算を実行してもよい。少なくとも一実施例では、高速 A L U 2 4 1 6、高速 A L U 2 4 1 8、及び低速 A L U 2 4 2 0 は、1 6、3 2、1 2 8、2 5 6 などを含む様々なデータ・ビット・サイズをサポートするように実装されてもよい。少なくとも一実施例では、浮動小数点 A L U 2 4 2 2 及び浮動小数点移動ユニット 2 4 2 4 は、S I M D 及びマルチメディア命令と併せた 1 2 8 ビット幅のパック・データ・オペランドなど様々なビット幅を有する幅広いオペランドをサポートするように実装されてもよい。

10

【0345】

少なくとも一実施例では、u o p スケジューラ 2 4 0 2、2 4 0 4、2 4 0 6 は、親ロードが実行を終了する前に、従属演算をディスパッチする。少なくとも一実施例では、u o p s は、プロセッサ 2 4 0 0 において投機的にスケジューリング及び実行されてもよいので、プロセッサ 2 4 0 0 は、メモリ・ミスに対処するための論理も含んでよい。少なくとも一実施例では、データ・キャッシュにおいてデータ・ロードがミスした場合、一時的に不正確なデータを有するスケジューラを通り過ぎたパイプラインに、進行中の従属演算が存在してもよい。少なくとも一実施例では、リプレイ機構が、不正確なデータを使用する命令を追跡及び再実行する。少なくとも一実施例では、従属演算は、リプレイされる必要があってもよく、独立した演算は、完了が許容されてもよい。少なくとも一実施例では、プロセッサの少なくとも一実施例のスケジューラ及びリプレイ機構はまた、テキスト・ストリング比較演算のための命令シーケンスを捕捉するように設計されてもよい。

20

【0346】

少なくとも一実施例では、「レジスタ」は、オペランドを識別するための命令の一部として使用することができるオンボード・プロセッサのストレージ・ロケーションを指してもよい。少なくとも一実施例では、レジスタは、(プログラマの視点から見て)プロセッサの外部から使用可能であり得るものであってもよい。少なくとも一実施例では、レジスタは、特定のタイプの回路に限定されなくてもよい。むしろ、少なくとも一実施例では、レジスタは、データを記憶し、データを提供し、本明細書に記載の機能を実行してもよい。少なくとも一実施例では、本明細書に記載のレジスタは、専用物理レジスタ、レジスタ・リネーミングを使用して動的に配分される物理レジスタ、専用物理レジスタと動的に配分される物理レジスタとの組合せなど、任意の数の異なる技法を使用して、プロセッサ内の回路によって実装されてもよい。少なくとも一実施例では、整数レジスタは、3 2 ビットの整数データを記憶する。少なくとも一実施例のレジスタ・ファイルは、パック・データのための 8 つのマルチメディア S I M D レジスタも含む。

30

【0347】

1 つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理 8 1 5 が使用される。推論及び／又は訓練論理 8 1 5 に関する詳細事項は、図 8 A 及び／又は図 8 B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 8 1 5 の一部又はすべてが、実行ブロック 2 4 1 1、及び図示してある若しくは図示していない他のメモリ又はレジスタに組み込まれてもよい。たとえば、少なくとも一実施例では、本明細書に記載の訓練及び／又は推論の技法は、実行ブロック 2 4 1 1 に示す A L U のうちの 1 つ又は複数を使用してもよい。さらに、重みパラメータは、本明細書に記載の 1 つ又は複数の機械学習アルゴリズム、ニューラル・ネットワーク・アーキテクチャ、ユース・ケース、又は訓練技法を実行するための実行ブロック 2 4 1 1 の A L U を構成するオン・チップ若しくはオフ・チップのメモリ及び／又はレジスタ(図示する又は図示せず)に記憶されてもよい。

40

【0348】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ~ 図 7 に関連して上に記載した技法に少なくと

50

も部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0349】

図25は、少なくとも一実施例による深層学習アプリケーション・プロセッサ2500を示す。少なくとも一実施例では、深層学習アプリケーション・プロセッサ2500は、深層学習アプリケーション・プロセッサ2500によって実行される場合に、本開示全体を通して記載するプロセス及び技法の一部又はすべてを、深層学習アプリケーション・プロセッサ2500に実行させる命令を使用する。少なくとも一実施例では、深層学習アプリケーション・プロセッサ2500は、特定用途向け集積回路（ASIC）である。少なくとも一実施例では、アプリケーション・プロセッサ2500は、1つ若しくは複数の命令又は両方を実行した結果としていずれもハードウェアに「ハード・ワイヤード」された行列乗算演算を実行する。少なくとも一実施例では、深層学習アプリケーション・プロセッサ2500は、限定することなく、処理クラスタ2510（1）～2510（12）、チップ間リンク（「ICL」）2520（1）～2520（12）、チップ間コントローラ（「ICC」）2530（1）～2530（2）、高帯域幅メモリ第2世代（「HBM2」）2540（1）～2540（4）、メモリ・コントローラ（「Mem Ctrlrs」）2542（1）～2542（4）、高帯域幅メモリ物理層（「HBM PHY」）2544（1）～2544（4）、管理コントローラ中央処理装置（「管理コントローラCPU」）2550、シリアル・ペリフェラル・インターフェース、集積回路間、及び汎用入力／出力ブロック（「SPI、I2C、GPIO」）2560、周辺構成要素相互接続エクスプレス・コントローラ及びダイレクト・メモリ・アクセス・ブロック（「PCIeコントローラ及びDMA」）2570、並びに16レーン周辺構成要素相互接続エクスプレス・ポート（「PCI Express x16」）2580を含む。

【0350】

少なくとも一実施例では、処理クラスタ2510は、本明細書に記載の技法を含む1つ又は複数の訓練技法を使用して計算された重みパラメータに基づき、推論又は予測の演算を含む深層学習演算を実行してもよい。少なくとも一実施例では、各処理クラスタ2510は、限定することなく、任意の数及びタイプのプロセッサを含んでもよい。少なくとも一実施例では、深層学習アプリケーション・プロセッサ2500は、任意の数及びタイプの処理クラスタ2500を含んでもよい。少なくとも一実施例では、チップ間リンク2520は、双方向性である。少なくとも一実施例では、チップ間リンク2520及びチップ間コントローラ2530は、1つ又は複数のニューラル・ネットワークに具体化された1つ又は複数の機械学習アルゴリズムを実行した結果得られるアクティブ化情報を含む情報を、複数の深層学習アプリケーション・プロセッサ2500が交換できるようにする。少なくとも一実施例では、深層学習アプリケーション・プロセッサ2500は、（ゼロを含む）任意の数及びタイプのICL2520及びICC2530を含んでもよい。

【0351】

少なくとも一実施例では、HBM2 2540は、合計32ギガバイト（GB：Gigabyte）のメモリを提供する。少なくとも一実施例では、HBM2 2540（i）は、メモリ・コントローラ2542（i）とHBM PHY2544（i）の両方に関連付けられ、ここで「i」は任意の整数である。少なくとも一実施例では、任意の数のHBM2 2540が、任意のタイプ及び合計量の高帯域幅メモリを提供してもよく、（ゼロを含む）任意の数及びタイプのメモリ・コントローラ2542及びHBM PHY2544に関連付けられてもよい。少なくとも一実施例では、SPI、I2C、GPIO2560、PCIeコントローラ及びDMA2570、並びに／又はPCIe2580は、任意の技術的に実行可能なやり方で任意の数及びタイプの通信規格を有効にする任意の数及び

10

20

30

40

50

タイプのブロックに置き換えられてもよい。

【0352】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、深層学習アプリケーション・プロセッサは、深層学習アプリケーション・プロセッサ2500に提供される情報を予測又は推論するようにニューラル・ネットワークなどの機械学習モデルを訓練するために使用される。少なくとも一実施例では、深層学習アプリケーション・プロセッサ2500は、別のプロセッサ若しくはシステムによって、又は深層学習アプリケーション・プロセッサ2500によって訓練されてきた訓練済み機械学習モデル（たとえば、ニューラル・ネットワーク）に基づき、情報を推論又は予測するために使用される。少なくとも一実施例では、プロセッサ2500は、本明細書に記載の1つ又は複数のニューラル・ネットワークのユース・ケースを実行するために使用されてもよい。

10

【0353】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

20

【0354】

図26は、少なくとも一実施例による、ニューロモーフィック・プロセッサ2600のブロック図である。少なくとも一実施例では、ニューロモーフィック・プロセッサ2600は、ニューロモーフィック・プロセッサ2600の外部のソースから1つ又は複数の入力を受信する。少なくとも一実施例では、これらの入力は、ニューロモーフィック・プロセッサ2600内の1つ又は複数のニューロン2602に送信されてもよい。少なくとも一実施例では、ニューロン2602及びその構成要素は、1つ又は複数の算術論理演算ユニット（ALU）を含む回路又は論理を使用して、実装されてもよい。少なくとも一実施例では、ニューロモーフィック・プロセッサ2600は、限定することなく、ニューロン2602の数千又は数百万のインスタンスを含んでもよいが、任意の好適な数のニューロン2602が使用されてもよい。少なくとも一実施例では、ニューロン2602の各インスタンスは、ニューロン入力2604及びニューロン出力2606を含んでもよい。少なくとも一実施例では、ニューロン2602は出力を生成してもよく、この出力は、ニューロン2602の他のインスタンスの入力に送信されてもよい。たとえば、少なくとも一実施例では、ニューロン入力2604及びニューロン出力2606は、シナプス2608を介して相互接続されてもよい。

30

【0355】

少なくとも一実施例では、ニューロン2602とシナプス2608は、ニューロモーフィック・プロセッサ2600が受信した情報をニューロモーフィック・プロセッサ2600が動作して処理又は分析するように、相互接続されてもよい。少なくとも一実施例では、ニューロン2602は、ニューロン入力2604を介して受信した入力が、閾値を超えているとき、出力パルス（又は「発火」若しくは「スパイク」）を送信してもよい。少なくとも一実施例では、ニューロン2602は、ニューロン入力2604において受信した信号を合計又は積分してもよい。たとえば、少なくとも一実施例では、ニューロン2602は、漏れ積分発火ニューロン（leaky integrate-and-fire neuron）として実装されてもよく、ここで、合計（「膜電位」と呼ばれる）が閾値を超える場合には、ニューロン2602は、シグモイド関数又は閾値関数などの伝達関数を使用して、出力（又は「発火」）を生成してもよい。少なくとも一実施例では、漏れ積分

40

50

発火ニューロンは、ニューロン入力2604で受信した信号を合計して膜電位にしてもよく、また、崩壊因子（又は漏れ）を適用して膜電位を低減してもよい。少なくとも一実施例では、複数の入力信号が、閾値を超えるほど十分に素早く（すなわち、膜電位の崩壊が少なすぎて発火できなくなる前に）ニューロン入力2604において受信された場合には、漏れ積分発火ニューロンが発火してもよい。少なくとも一実施例では、ニューロン2602は、入力を受信し、入力を積分して膜電位にし、膜電位を崩壊させる回路又は論理を使用して、実装されてもよい。少なくとも一実施例では、入力は平均化されてもよく、又は任意の他の好適な伝達関数を使用されてもよい。さらに、少なくとも一実施例では、ニューロン2602は、ニューロン2604に伝達関数を適用した結果が閾値を超えるとき、ニューロン2606において出力スパイクを生成するコンパレータ回路又は論理を、限定することなく含んでもよい。少なくとも一実施例では、ニューロン2602が発火すると、前に受信した入力情報を、たとえば膜電位を0又は他の好適なデフォルト値に再設定することによって、無視してもよい。少なくとも一実施例では、膜電位が0にリセットされると、ニューロン2602は、好適な期間（又は不応期）の後に通常の動作を再開してもよい。

【0356】

少なくとも一実施例では、ニューロン2602は、シナプス2608を通して相互接続されてもよい。少なくとも一実施例では、シナプス2608は、第1のニューロン2602の出力から第2のニューロン2602の入力に信号を送信するように動作してもよい。少なくとも一実施例では、ニューロン2602は、シナプス2608の2つ以上のインスタンスを介して情報を送信してもよい。少なくとも一実施例では、ニューロン出力2606の1つ又は複数のインスタンスは、シナプス2608のインスタンスを介して、同じニューロン2602のニューロン入力2604のインスタンスに接続されてもよい。少なくとも一実施例では、シナプス2608のインスタンスを介して送信されることになる出力を生成するニューロン2602のインスタンスは、シナプス2608のそのインスタンスに対して「シナプス前ニューロン」と呼ばれてもよい。少なくとも一実施例では、シナプス2608のインスタンスを介して送信されることになる入力を受信するニューロン2602のインスタンスは、シナプス2608のそのインスタンスに対して「シナプス後ニューロン」と呼ばれてもよい。少なくとも一実施例では、ニューロン2602のインスタンスは、シナプス2608の1つ又は複数のインスタンスから入力を受信してもよく、また、シナプス2608の1つ又は複数のインスタンスを介して出力を送信してもよいので、ニューロン2602の単一のインスタンスは、したがって、シナプス2608の様々なインスタンスに対して「シナプス前ニューロン」と「シナプス後ニューロン」の両方であってもよい。

【0357】

少なくとも一実施例では、ニューロン2602は、1つ又は複数の層に組織化されてもよい。少なくとも一実施例では、ニューロン2602の各インスタンスは、1つ又は複数のシナプス2608を通して1つ又は複数のニューロン入力2604にファン・アウトすることができる1つのニューロン出力2606を有してもよい。少なくとも一実施例では、第1の層2610のニューロン2602のニューロン出力2606は、第2の層2612のニューロン2602のニューロン入力2604に接続されてもよい。少なくとも一実施例では、層2610は、「フィード・フォワード」層と呼ばれてもよい。少なくとも一実施例では、第1の層2610のインスタンスにおけるニューロン2602の各インスタンスは、第2の層2612におけるニューロン2602の各インスタンスにファン・アウトしてもよい。少なくとも一実施例では、第1の層2610は、「完全に接続されたフィード・フォワード層」と呼ばれてもよい。少なくとも一実施例では、第2の層2612のインスタンスにおけるニューロン2602の各インスタンスは、第3の層2614におけるニューロン2602の全インスタンスより少ないインスタンスにファン・アウトしてもよい。少なくとも一実施例では、第2の層2612は、「疎に接続されたフィード・フォワード層」と呼ばれてもよい。少なくとも一実施例では、第2の層2612のニューロン

10

20

30

40

50

2602は、第2の層2612におけるニューロン2602を含め、複数の他の層のニューロン2602にファン・アウトしてもよい。少なくとも一実施例では、第2の層2612は、「回帰層」と呼ばれてもよい。少なくとも一実施例では、ニューロモーフィック・プロセッサ2600は、疎に接続されたフィード・フォワード層と完全に接続されたフィード・フォワード層の両方を限定することなく含む、回帰層とフィード・フォワード層の任意の好適な組合せを限定することなく含んでもよい。

【0358】

少なくとも一実施例では、ニューロモーフィック・プロセッサ2600は、シナプス2608をニューロン2602に接続するための再構成可能相互接続アーキテクチャ、又は専用ハード・ワイヤード相互接続を、限定することなく含んでもよい。少なくとも一実施例では、ニューロモーフィック・プロセッサ2600は、ニューラル・ネットワーク・トポロジ、及びニューロンのファン・イン／ファン・アウトに基づき、必要に応じてシナプスを異なるニューロン2602に配分できるようにする回路又は論理を、限定することなく含んでもよい。たとえば、少なくとも一実施例では、シナプス2608は、ネットワーク・オン・チップなどの相互接続ファブリックを使用して、又は専用の接続を用いて、ニューロン2602に接続されてもよい。少なくとも一実施例では、シナプス相互接続及びその構成要素は、回路又は論理を使用して実装されてもよい。

【0359】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0360】

図27は、少なくとも一実施例による処理システムのブロック図である。少なくとも一実施例では、システム2700は、1つ又は複数のプロセッサ2702、及び1つ又は複数のグラフィックス・プロセッサ2708を含み、単一プロセッサのデスクトップ・システム、マルチプロセッサのワークステーション・システム、又は多数のプロセッサ2702若しくはプロセッサ・コア2707を有するサーバ・システムであってもよい。少なくとも一実施例では、システム2700は、モバイル・デバイス、携帯型デバイス、又は組み込みデバイスで使用するためのシステム・オン・チップ（SoC）集積回路内に組み込まれた処理プラットフォームである。

【0361】

少なくとも一実施例では、システム2700は、サーバ・ベースのゲーミング・プラットフォーム、ゲーム及びメディアのコンソールを含むゲーム・コンソール、モバイル・ゲーミング・コンソール、携帯型ゲーム・コンソール、若しくはオンライン・ゲーム・コンソールを含んでもよく、又はそれらに組み込まれてもよい。少なくとも一実施例では、システム2700は、モバイル・フォン、スマート・フォン、タブレット・コンピューティング・デバイス、又はモバイル・インターネット・デバイスである。少なくとも一実施例では、処理システム2700はまた、スマート・ウォッチ・ウェアラブル・デバイス、スマート・アイウェア・デバイス、拡張現実デバイス、若しくは仮想現実デバイスなどのウェアラブル・デバイスを含んでもよく、それらに結合されてもよく、又はそれらの中に一体化されてもよい。少なくとも一実施例では、処理システム2700は、1つ又は複数のプロセッサ2702と、1つ又は複数のグラフィックス・プロセッサ2708によって生成されるグラフィカル・インターフェースとを有するテレビ又はセット・トップ・ボックス・デバイスである。

【0362】

少なくとも一実施例では、1つ又は複数のプロセッサ2702はそれぞれ、実行されたときにシステム及びユーザ・ソフトウェアのための動作を実行する命令を処理するための1つ又は複数のプロセッサ・コア2707を含む。少なくとも一実施例では、1つ又は複数のプロセッサ・コア2707のそれぞれは、特定の命令シーケンス2709を処理するように構成される。少なくとも一実施例では、命令シーケンス2709は、複合命令セット・コンピューティング(CISC)、縮小命令セット・コンピューティング(RISC)、又は超長命令語(VLIW)を介したコンピューティングを容易にしてもよい。少なくとも一実施例では、プロセッサ・コア2707はそれぞれ、異なる命令シーケンス2709を処理してもよく、この命令セットは、他の命令シーケンスのエミュレーションを容易にする命令を含んでもよい。少なくとも一実施例では、プロセッサ・コア2707はまた、デジタル信号プロセッサ(DSP)などの他の処理デバイスを含んでもよい。

【0363】

少なくとも一実施例では、プロセッサ2702はキャッシュ・メモリ2704を含む。少なくとも一実施例では、プロセッサ2702は、単一の内部キャッシュ又は複数レベルの内部キャッシュを有してもよい。少なくとも一実施例では、キャッシュ・メモリは、プロセッサ2702の様々な構成要素間で共有される。少なくとも一実施例では、プロセッサ2702はまた、外部キャッシュ(たとえば、レベル3(L3)キャッシュ又はラスト・レベル・キャッシュ(LLC)) (図示せず)を使用し、このキャッシュは、知られているキャッシュ・コヒーレンス技法を使用して、プロセッサ・コア2707間で共有されてもよい。少なくとも一実施例では、さらにレジスタ・ファイル2706がプロセッサ2702に含まれ、このレジスタ・ファイルは、異なるタイプのデータを記憶するための異なるタイプのレジスタ(たとえば、整数レジスタ、浮動小数点レジスタ、状態レジスタ、及び命令ポインタ・レジスタ)を含んでもよい。少なくとも一実施例では、レジスタ・ファイル2706は、汎用レジスタ又は他のレジスタを含んでもよい。

【0364】

少なくとも一実施例では、1つ又は複数のプロセッサ2702は、1つ又は複数のインターフェース・バス2710に結合されて、アドレス、データ、又は制御信号などの通信信号を、プロセッサ2702とシステム2700内の他の構成要素との間で送信する。少なくとも一実施例では、インターフェース・バス2710は、一実施例では、ダイレクト・メディア・インターフェース(DMI)バスのバージョンなどのプロセッサ・バスとすることができる。少なくとも一実施例では、インターフェース2710は、DMIバスに限定されず、1つ又は複数のペリフェラル・コンポーネント・インターコネクト・バス(たとえば、PCI、PCIエクスプレス)、メモリ・バス、又は他のタイプのインターフェース・バスを含んでもよい。少なくとも一実施例では、プロセッサ2702は、統合メモリ・コントローラ2716、及びプラットフォーム・コントローラ・ハブ2730を含む。少なくとも一実施例では、メモリ・コントローラ2716は、メモリ・デバイスとシステム2700の他の構成要素との間の通信を容易にし、一方でプラットフォーム・コントローラ・ハブ(PCIH)2730は、ローカルI/Oバスを介してI/Oデバイスへの接続を提供する。

【0365】

少なくとも一実施例では、メモリ・デバイス2720は、ダイナミック・ランダム・アクセス・メモリ(DRAM)デバイス、スタティック・ランダム・アクセス・メモリ(SRAM)デバイス、フラッシュ・メモリ・デバイス、相変化メモリ・デバイス、又はプロセッサ・メモリとしての役割を果たすのに好適な性能を有する何らかの他のメモリ・デバイスとすることができる。少なくとも一実施例では、メモリ・デバイス2720は、システム2700のためのシステム・メモリとして動作して、1つ又は複数のプロセッサ2702がアプリケーション若しくはプロセスを実行するときに使用するためのデータ2722及び命令2721を記憶することができる。少なくとも一実施例では、メモリ・コントローラ2716はまた、任意選択の外部グラフィックス・プロセッサ2712と結合しており、このグラフィックス・プロセッサは、プロセッサ2702内の1つ又は複数のグラフ

ックス・プロセッサ2708と通信して、グラフィックス及びメディアの動作を実行してもよい。少なくとも一実施例では、ディスプレイ・デバイス2711は、プロセッサ2702に接続することができる。少なくとも一実施例では、ディスプレイ・デバイス2711は、モバイル電子デバイス又はラップトップ・デバイスのような内部ディスプレイ・デバイス、又はディスプレイ・インターフェース（たとえば、ディスプレイ・ポートなど）を介して取り付けられる外部ディスプレイ・デバイスのうちの1つ又は複数を含むことができる。少なくとも一実施例では、ディスプレイ・デバイス2711は、仮想現実（VR）アプリケーション又は拡張現実（AR）アプリケーションで使用するための立体ディスプレイ・デバイスなどの頭部装着型ディスプレイ（HMD）を含むことができる。

【0366】

少なくとも一実施例では、プラットフォーム・コントローラ・ハブ2730は、周辺装置が高速I/Oバスを介してメモリ・デバイス2720及びプロセッサ2702に接続できるようにする。少なくとも一実施例では、I/O周辺装置は、オーディオ・コントローラ2746、ネットワーク・コントローラ2734、ファームウェア・インターフェース2728、ワイヤレス・トランシーバ2726、タッチ・センサ2725、データ・ストレージ・デバイス2724（たとえば、ハード・ディスク・ドライブ、フラッシュ・メモリなど）を含むが、これらに限定されない。少なくとも一実施例では、データ・ストレージ・デバイス2724は、ストレージ・インターフェース（たとえば、SATA）を介して、又はペリフェラル・コンポーネント・インターコネクト・バス（たとえば、PCI、PCIエクスプレス）などのペリフェラル・バスを介して、接続することができる。少なくとも一実施例では、タッチ・センサ2725は、タッチ画面センサ、圧力センサ、又は指紋センサを含むことができる。少なくとも一実施例では、ワイヤレス・トランシーバ2726は、WiFiトランシーバ、Bluetoothトランシーバ、又は3G、4G、若しくはLong Term Evolution（LTE）トランシーバなどのモバイル・ネットワーク・トランシーバとすることができる。少なくとも一実施例では、ファームウェア・インターフェース2728は、システム・ファームウェアとの通信を可能にし、たとえば、ユニファイド・エクステンシブル・ファームウェア・インターフェース（UEFI）とすることができる。少なくとも一実施例では、ネットワーク・コントローラ2734は、有線ネットワークへのネットワーク接続を可能にすることができる。少なくとも一実施例では、高性能ネットワーク・コントローラ（図示せず）は、インターフェース・バス2710と結合する。少なくとも一実施例では、オーディオ・コントローラ2746は、多チャンネル・ハイ・デフィニション・オーディオ・コントローラである。少なくとも一実施例では、システム2700は、レガシー（たとえば、パーソナル・システム2（PS/2））デバイスをシステム2700に結合するための任意選択のレガシーI/Oコントローラ2740を含む。少なくとも一実施例では、プラットフォーム・コントローラ・ハブ2730は、キーボードとマウス2743の組合せ、カメラ2744、又は他のUSB入力デバイスなど、1つ又は複数のユニバーサル・シリアル・バス（USB）コントローラ2742の接続入力デバイスにも接続することができる。

【0367】

少なくとも一実施例では、メモリ・コントローラ2716及びプラットフォーム・コントローラ・ハブ2730のインスタンスは、外部グラフィックス・プロセッサ2712などの個別の外部グラフィックス・プロセッサに一体化されてもよい。少なくとも一実施例では、プラットフォーム・コントローラ・ハブ2730及び／又はメモリ・コントローラ2716は、1つ又は複数のプロセッサ2702の外部にあってもよい。たとえば、少なくとも一実施例では、システム2700は、外部のメモリ・コントローラ2716及びプラットフォーム・コントローラ・ハブ2730を含むことができ、これらは、プロセッサ2702と通信するシステム・チップセット内のメモリ・コントローラ・ハブ及び周辺装置コントローラ・ハブとして構成されてもよい。

【0368】

1つ又は複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及

10

20

30

40

50

び／又は訓練論理 8 1 5 が使用される。推論及び／又は訓練論理 8 1 5 に関する詳細事項は、図 8 A 及び／又は図 8 B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 8 1 5 の一部又はすべてが、グラフィックス・プロセッサ 2 7 0 0 に組み込まれてもよい。たとえば、少なくとも一実施例では、本明細書に記載の訓練及び／又は推論の技法は、3 D パイプラインに具体化された A L U のうちの 1 つ又は複数を使用してもよい。さらに、少なくとも一実施例では、本明細書に記載の推論及び／又は訓練の動作は、図 8 A 又は図 8 B に示す論理以外の論理を使用して行われてもよい。少なくとも一実施例では、重みパラメータは、本明細書に記載の 1 つ又は複数の機械学習アルゴリズム、ニューラル・ネットワーク・アーキテクチャ、ユース・ケース、又は訓練技法を実行するためのグラフィックス・プロセッサ 2 7 0 0 の A L U を構成するオン・チップ若しくはオフ・チップのメモリ及び／又はレジスタ（図示している又は図示せず）に記憶されてもよい。

【0369】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0370】

図 2 8 は、少なくとも一実施例による、1 つ又は複数のプロセッサ・コア 2 8 0 2 A ～ 2 8 0 2 N、統合メモリ・コントローラ 2 8 1 4、及び統合グラフィックス・プロセッサ 2 8 0 8 を有するプロセッサ 2 8 0 0 のブロック図である。少なくとも一実施例では、プロセッサ 2 8 0 0 は、破線の四角によって表される追加コア 2 8 0 2 N を含むそれ以下の数の追加コアを含むことができる。少なくとも一実施例では、プロセッサ・コア 2 8 0 2 A ～ 2 8 0 2 N のそれぞれは、1 つ又は複数の内部キャッシュ・ユニット 2 8 0 4 A ～ 2 8 0 4 N を含む。少なくとも一実施例では、各プロセッサ・コアはまた、1 つ又は複数の共有キャッシュ・ユニット 2 8 0 6 にアクセスできる。

【0371】

少なくとも一実施例では、内部キャッシュ・ユニット 2 8 0 4 A ～ 2 8 0 4 N、及び共有キャッシュ・ユニット 2 8 0 6 は、プロセッサ 2 8 0 0 内のキャッシュ・メモリ階層を表す。少なくとも一実施例では、キャッシュ・メモリ・ユニット 2 8 0 4 A ～ 2 8 0 4 N は、各プロセッサ・コア内の命令及びデータのキャッシュの少なくとも 1 つのレベル、並びにレベル 2（L 2）、レベル 3（L 3）、レベル 4（L 4）などの共有中間レベル・キャッシュの 1 つ又は複数のレベル、又はキャッシュの他のレベルを含んでもよく、ここで外部メモリの前の最高レベルのキャッシュは、L L C として分類される。少なくとも一実施例では、キャッシュ・コヒーレンス論理は、様々なキャッシュ・ユニット 2 8 0 6 及び 2 8 0 4 A ～ 2 8 0 4 N 間でコヒーレンスを維持する。

【0372】

少なくとも一実施例では、プロセッサ 2 8 0 0 はまた、1 つ又は複数のバス・コントローラ・ユニット 2 8 1 6 とシステム・エージェント・コア 2 8 1 0 のセットを含んでもよい。少なくとも一実施例では、バス・コントローラ・ユニット 2 8 1 6 は、1 つ又は複数の P C I 若しくは P C I エクスプレス・バスなどのペリフェラル・バスのセットを管理する。少なくとも一実施例では、システム・エージェント・コア 2 8 1 0 は、様々なプロセッサ構成要素のための管理機能を提供する。少なくとも一実施例では、システム・エージェント・コア 2 8 1 0 は、様々な外部メモリ・デバイス（図示せず）へのアクセスを管理するための 1 つ又は複数の統合メモリ・コントローラ 2 8 1 4 を含む。

【0373】

少なくとも一実施例では、プロセッサ・コア 2802A~2802N の 1 つ又は複数は、同時マルチスレッディングのサポートを含む。少なくとも一実施例では、システム・エージェント・コア 2810 は、マルチスレッドの処理中にコア 2802A~2802N を調整し動作させるための構成要素を含む。少なくとも一実施例では、システム・エージェント・コア 2810 はさらに、電力制御ユニット (PCU) を含んでもよく、このユニットは、プロセッサ・コア 2802A~2802N 及びグラフィックス・プロセッサ 2808 の 1 つ又は複数の電力状態を調整するための論理及び構成要素を含む。

【0374】

少なくとも一実施例では、プロセッサ 2800 はさらに、グラフィックス処理動作を実行するためのグラフィックス・プロセッサ 2808 を含む。少なくとも一実施例では、グラフィックス・プロセッサ 2808 は、共有キャッシュ・ユニット 2806 と、1 つ又は複数の統合メモリ・コントローラ 2814 を含むシステム・エージェント・コア 2810 とに結合する。少なくとも一実施例では、システム・エージェント・コア 2810 はまた、1 つ又は複数の結合されたディスプレイに対してグラフィックス・プロセッサの出力を行わせるためのディスプレイ・コントローラ 2811 を含む。少なくとも一実施例では、ディスプレイ・コントローラ 2811 はまた、少なくとも 1 つの相互接続を介してグラフィックス・プロセッサ 2808 に結合された別個のモジュールであってもよく、又はグラフィックス・プロセッサ 2808 内に一体化されていてもよい。

【0375】

少なくとも一実施例では、プロセッサ 2800 の内部構成要素を結合するために、リング・ベースの相互接続ユニット 2812 が使用される。少なくとも一実施例では、ポイントツーポイント相互接続、スイッチ相互接続、又は他の技法などの代替的な相互接続ユニットが使用されてもよい。少なくとも一実施例では、グラフィックス・プロセッサ 2808 は、I/O リンク 2813 を介してリング相互接続 2812 と結合する。

【0376】

少なくとも一実施例では、I/O リンク 2813 は、様々なプロセッサ構成要素と、eDRAM モジュールなどの高性能組み込みメモリ・モジュール 2818 との間の通信を容易にするオン・パッケージ I/O 相互接続を含む多様な I/O 相互接続のうちの少なくとも 1 つを表す。少なくとも一実施例では、プロセッサ・コア 2802A~2802N のそれぞれ及びグラフィックス・プロセッサ 2808 は、共有ラスト・レベル・キャッシュとして組み込みメモリ・モジュール 2818 を使用する。

【0377】

少なくとも一実施例では、プロセッサ・コア 2802A~2802N は、共通の命令セット・アーキテクチャを実行する同種のコアである。少なくとも一実施例では、プロセッサ・コア 2802A~2802N は、命令セット・アーキテクチャ (ISA) の観点から見れば異種であり、ここでプロセッサ・コア 2802A~2802N のうちの 1 つ又は複数は、共通の命令セットを実行するが、プロセッサ・コア 2802A~2802N のうちの 1 つ又は複数の他のコアは、共通の命令セットのサブセット、又は異なる命令セットを実行する。少なくとも一実施例では、プロセッサ・コア 2802A~2802N は、マイクロ・アーキテクチャの観点から見れば異種であり、ここで電力消費量が相対的に高い 1 つ又は複数のコアは、電力消費量がより低い 1 つ又は複数のコアと結合する。少なくとも一実施例では、プロセッサ 2800 は、1 つ又は複数のチップ上に、又は SOC 集積回路として実装することができる。

【0378】

1 つ又は複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理 815 が使用される。推論及び／又は訓練論理 815 に関する詳細事項は、図 8A 及び／又は図 8B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 815 の一部又はすべてが、グラフィックス・プロセッサ 2810 に組み込まれてもよい。たとえば、少なくとも一実施例では、本明細書に記載の訓練及び／又は推論の技法は、3D パイプライン、グラフィックス・コア 2802、共有機能論

10

20

30

40

50

理、又は図 28 の他の論理に具体化された A L U のうちの 1 つ又は複数を使用してもよい。さらに、少なくとも一実施例では、本明細書に記載の推論及び／又は訓練の動作は、図 8 A 又は図 8 B に示す論理以外の論理を使用して行われてもよい。少なくとも一実施例では、重みパラメータは、本明細書に記載の 1 つ又は複数の機械学習アルゴリズム、ニューラル・ネットワーク・アーキテクチャ、ユース・ケース、又は訓練技法を実行するためのプロセッサ 2800 の A L U を構成するオン・チップ若しくはオフ・チップのメモリ及び／又はレジスタ（図示している又は図示せず）に記憶されてもよい。

【0379】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0380】

図 29 は、グラフィックス・プロセッサ 2900 のブロック図であり、これは、個別グラフィックス・プロセッシング・ユニットであってもよく、又は複数の処理コアと統合されたグラフィックス・プロセッサであってもよい。少なくとも一実施例では、グラフィックス・プロセッサ 2900 は、メモリにマッピングされた I / O インターフェースを介して、メモリに入れられたコマンドを用いて、グラフィックス・プロセッサ 2900 のレジスタと通信する。少なくとも一実施例では、グラフィックス・プロセッサ 2900 は、メモリにアクセスするためのメモリ・インターフェース 2914 を含む。少なくとも一実施例では、メモリ・インターフェース 2914 は、ローカル・メモリ、1 つ若しくは複数の内部キャッシュ、1 つ若しくは複数の共有外部キャッシュ、及び／又はシステム・メモリへのインターフェースである。

【0381】

少なくとも一実施例では、グラフィックス・プロセッサ 2900 はまた、ディスプレイ出力データをディスプレイ・デバイス 2920 に向けて駆動するためのディスプレイ・コントローラ 2902 も含む。少なくとも一実施例では、ディスプレイ・コントローラ 2902 は、ディスプレイ・デバイス 2920 用の 1 つ又は複数の重なり平面、及び多層のビデオ若しくはユーザ・インターフェース要素の合成のためのハードウェアを含む。少なくとも一実施例では、ディスプレイ・デバイス 2920 は、内部又は外部のディスプレイ・デバイスとすることができる。少なくとも一実施例では、ディスプレイ・デバイス 2920 は、仮想現実（VR）ディスプレイ・デバイス又は拡張現実（AR）ディスプレイ・デバイスなどの頭部装着型ディスプレイ・デバイスである。少なくとも一実施例では、グラフィックス・プロセッサ 2900 は、MPEG-2 などの動画エキスパート・グループ（MPEG）フォーマット、H.264 / MPEG-4 AVC などのアドバンスド・ビデオ・コーディング（AVC）フォーマット、並びに映画テレビ技術者協会（SMPTE）421M / VC-1、及び JPEG などのジョイント・フォトグラフィック・エキスパート・グループ（JPEG）フォーマット、及びモーション JPEG（MJPEG）フォーマットを含むがこれらに限定されない 1 つ又は複数のメディア符号化フォーマットに、それらのフォーマットから、又はそれらのフォーマット間で、メディアをエンコード、デコード、又はコード変換するためのビデオ・コーデック・エンジン 2906 を含む。

【0382】

少なくとも一実施例では、グラフィックス・プロセッサ 2900 は、たとえばビット境界ブロック転送を含む 2 次元（2D）ラスタライザ動作を実行するためのブロック画像転送（BLIT）エンジン 2904 を含む。しかし、少なくとも一実施例では、2D グラフィックス動作は、グラフィックス処理エンジン（GPE）2910 の 1 つ又は複数の構

10

20

30

40

50

成要素を使用して実行される。少なくとも一実施例では、G P E 2 9 1 0 は、3 次元（3 D）グラフィックス動作及びメディア動作を含むグラフィックス動作を実行するためのコンピュータ・エンジンである。

【0383】

少なくとも一実施例では、G P E 2 9 1 0 は、3 D のプリミティブ形状（たとえば、矩形、三角形など）に作用する処理関数を使用して、3 次元画像及びシーンをレンダリングするなど、3 D 動作を実行するための3 D パイプライン2 9 1 2 を含む。少なくとも一実施例では、3 D パイプライン2 9 1 2 は、プログラム可能で固定された関数要素を含み、これは、3 D / メディア・サブシステム2 9 1 5 に対して様々なタスクを実行し、且つ／又は実行スレッドをスポンする。3 D パイプライン2 9 1 2 を使用してメディア動作を実行できるが、少なくとも一実施例では、G P E 2 9 1 0 は、ビデオの後処理及び画像強調などのメディア動作を実行するために使用されるメディア・パイプライン2 9 1 6 も含む。

10

【0384】

少なくとも一実施例では、メディア・パイプライン2 9 1 6 は、ビデオ・コーデック・エンジン2 9 0 6 の代わりに、又はそれを代表して、ビデオ・デコード加速、ビデオ・インターレース解除、及びエンコード加速などの1 つ又は複数の特別なメディア動作を実行するための固定機能又はプログラム可能論理ユニットを含む。少なくとも一実施例では、メディア・パイプライン2 9 1 6 は、3 D / メディア・サブシステム2 9 1 5 で実行するためのスレッドをスポンするためのスレッド・スポーニング・ユニットをさらに含む。少なくとも一実施例では、スポンされたスレッドは、3 D / メディア・サブシステム2 9 1 5 に含まれた1 つ又は複数のグラフィックス実行ユニット上で、メディア動作のための計算を実行する。

20

【0385】

少なくとも一実施例では、3 D / メディア・サブシステム2 9 1 5 は、3 D パイプライン2 9 1 2 及びメディア・パイプライン2 9 1 6 によってスポンされたスレッドを実行するための論理を含む。少なくとも一実施例では、3 D パイプライン2 9 1 2 及びメディア・パイプライン2 9 1 6 は、スレッド実行要求を3 D / メディア・サブシステム2 9 1 5 に送信し、この3 D / メディア・サブシステム2 9 1 5 は、様々な要求を調停し、利用可能なスレッド実行リソースにディスパッチするためのスレッド・ディスパッチ論理を含む。少なくとも一実施例では、実行リソースは、3 D 及びメディア・スレッドを処理するためのグラフィックス実行ユニットのアレイを含む。少なくとも一実施例では、3 D / メディア・サブシステム2 9 1 5 は、スレッド命令及びデータのための1 つ又は複数の内部キャッシュを含む。少なくとも一実施例では、サブシステム2 9 1 5 はまた、スレッド間でデータを共有し、出力データを記憶するための、レジスタ及びアドレス可能メモリを含む共有メモリも含む。

30

【0386】

1 つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理8 1 5 が使用される。推論及び／又は訓練論理8 1 5 に関する詳細事項は、図8 A 及び／又は図8 B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理8 1 5 の一部又はすべてが、グラフィックス・プロセッサ2 9 0 0 に組み込まれてもよい。たとえば、少なくとも一実施例では、本明細書に記載の訓練及び／又は推論の技法は、3 D パイプライン2 9 1 2 に具体化されたA L U のうちの1 つ又は複数を使用してもよい。さらに、少なくとも一実施例では、本明細書に記載の推論及び／又は訓練の動作は、図8 A 又は図8 B に示す論理以外の論理を使用して行われてもよい。少なくとも一実施例では、重みパラメータは、本明細書に記載の1 つ又は複数の機械学習アルゴリズム、ニューラル・ネットワーク・アーキテクチャ、ユース・ケース、又は訓練技法を実行するためのグラフィックス・プロセッサ2 9 0 0 のA L U を構成するオン・チップ若しくはオフ・チップのメモリ及び／又はレジスタ（図示する又は図示せず）に記憶されてもよい。

40

50

【0387】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0388】

図30は、少なくとも一実施例によるグラフィックス・プロセッサのグラフィックス処理エンジン3010のブロック図である。少なくとも一実施例では、グラフィックス処理エンジン(GPE)3010は、図29に示すGPE2910の1つのバージョンである。少なくとも一実施例では、メディア・パイプライン3016は任意選択であり、GPE3010内に明示的に含まれなくてもよい。少なくとも一実施例では、別個のメディア及び／又は画像のプロセッサが、GPE3010に結合される。

【0389】

少なくとも一実施例では、GPE3010は、コマンド・ストリーマ3003に結合され、又はそれを含み、このコマンド・ストリーマは、3Dパイプライン3012及び／又はメディア・パイプライン3016にコマンド・ストリームを提供する。少なくとも一実施例では、コマンド・ストリーマ3003はメモリに結合され、このメモリは、システム・メモリであってもよく、又は内部キャッシュ・メモリ及び共有キャッシュ・メモリのうちの1つ若しくは複数であってもよい。少なくとも一実施例では、コマンド・ストリーマ3003は、メモリからコマンドを受信し、3Dパイプライン3012及び／又はメディア・パイプライン3016にコマンドを送信する。少なくとも一実施例では、コマンドは、リング・バッファからフェッチされる命令、プリミティブ、又はマイクロ・オペレーションであり、このリング・バッファは、3Dパイプライン3012及びメディア・パイプライン3016のためのコマンドを記憶する。少なくとも一実施例では、リング・バッファはさらに、複数のコマンドのバッチを記憶するバッチ・コマンド・バッファを含むことができる。少なくとも一実施例では、3Dパイプライン3012用のコマンドはまた、3Dパイプライン3012用の頂点及び形状のデータ、並びに／又はメディア・パイプライン3016用の画像データ及びメモリ・オブジェクトなどであるがこれらに限定されないメモリに記憶されたデータへの参照も含むことができる。少なくとも一実施例では、3Dパイプライン3012及びメディア・パイプライン3016は、演算を実行することにより、又は1つ若しくは複数の実行スレッドをグラフィックス・コア・アレイ3014にディスパッチすることにより、コマンド及びデータを処理する。少なくとも一実施例では、グラフィックス・コア・アレイ3014は、グラフィックス・コア(たとえば、グラフィックス・コア3015A、グラフィックス・コア3015B)の1つ又は複数のブロックを含み、各ブロックは、1つ又は複数のグラフィックス・コアを含む。少なくとも一実施例では、各グラフィックス・コアは、グラフィックス及びコンピュータの動作を実行するための汎用及びグラフィックス専用の実行論理、並びに、図8A及び図8Bの推論及び／又は訓練論理815を含め、固定機能のテクスチャ処理及び／又は機械学習、及び人工知能の加速論理を含むグラフィックス実行リソースのセットを含む。

【0390】

少なくとも一実施例では、3Dパイプライン3012は、命令を処理し、実行スレッドをグラフィックス・コア・アレイ3014にディスパッチすることにより、頂点シェーダ、ジオメトリ・シェーダ、ピクセル・シェーダ、フラグメント・シェーダ、コンピュータ・シェーダ、又は他のシェーダ・プログラムなどの1つ又は複数のシェーダ・プログラムを処理するための固定機能及びプログラム可能論理を含む。少なくとも一実施例では、グラフィックス・コア・アレイ3014は、シェーダ・プログラムを処理する際に使用する

10

20

30

40

50

ための実行リソースの統合ブロックを提供する。少なくとも一実施例では、グラフィックス・コア・アレイ 3014 のグラフィックス・コア 3015 A ~ 3015 B 内の多目的の実行論理（たとえば、実行ユニット）は、様々な 3D の API シェーダ言語のサポートを含み、複数のシェーダに関連付けられた複数の同時実行スレッドを実行することができる。

【0391】

少なくとも一実施例では、グラフィックス・コア・アレイ 3014 はまた、ビデオ及び／又は画像の処理など、メディア機能を実行するための実行論理も含む。少なくとも一実施例では、実行ユニットはさらに、グラフィックス処理動作に加えて並列の汎用計算動作を実行するようにプログラム可能な汎用論理を含む。

【0392】

少なくとも一実施例では、グラフィックス・コア・アレイ 3014 上で実行しているスレッドにより生成される出力データは、統合リターン・バッファ（URB）3018 のメモリにデータを出力することができる。少なくとも一実施例では、URB 3018 は、複数のスレッド用のデータを記憶することができる。少なくとも一実施例では、グラフィックス・コア・アレイ 3014 上で実行している異なるスレッド間でデータを送信するために、URB 3018 を使用してもよい。少なくとも一実施例では、グラフィックス・コア・アレイ 3014 上のスレッドと、共有機能論理 3020 内の固定機能論理との間の同期のために、URB 3018 がさらに使用されてもよい。

【0393】

少なくとも一実施例では、グラフィックス・コア・アレイ 3014 はスケーラブルであり、それにより、グラフィックス・コア・アレイ 3014 は、可変数のグラフィックス・コアを含み、それぞれのグラフィックス・コアが、GPE 3010 の目的とする電力及び性能のレベルに基づき可変数の実行ユニットを有する。少なくとも一実施例では、実行リソースは動的にスケーラブルであり、それにより実行リソースは、必要に応じて有効化又は無効化されてもよい。

【0394】

少なくとも一実施例では、グラフィックス・コア・アレイ 3014 は、グラフィックス・コア・アレイ 3014 のグラフィックス・コア間で共有される複数のリソースを含む共有機能論理 3020 に結合される。少なくとも一実施例では、共有機能論理 3020 によって実行される共有機能は、専用の補足機能をグラフィックス・コア・アレイ 3014 に提供するハードウェア論理ユニットに具体化される。少なくとも一実施例では、共有機能論理 3020 は、サンプリングユニット 3021、数値ユニット 3022、及びスレッド間通信（ITC）論理 3023 を含むが、これらに限定されない。少なくとも一実施例では、1 つ又は複数のキャッシュ 3025 が、共有機能論理 3020 に含まれ、又はそれに結合される。

【0395】

少なくとも一実施例では、専用機能の需要が不十分でグラフィックス・コア・アレイ 3014 内に含められない場合に、共有機能が使用される。少なくとも一実施例では、専用機能を 1 つにインスタンス化したものが、共有機能論理 3020 において使用され、グラフィックス・コア・アレイ 3014 内の他の実行リソース間で共有される。少なくとも一実施例では、共有機能論理 3020 内の、グラフィックス・コア・アレイ 3014 によってのみ使用される特定の共有機能は、グラフィックス・コア・アレイ 3014 内の共有機能論理 3016 内に含まれてもよい。少なくとも一実施例では、グラフィックス・コア・アレイ 3014 内の共有機能論理 3016 は、共有機能論理 3020 内の一部又はすべての論理を含むことができる。少なくとも一実施例では、共有機能論理 3020 内のすべての論理要素は、グラフィックス・コア・アレイ 3014 の共有機能論理 3026 内で複製されてもよい。少なくとも一実施例では、共有機能論理 3020 は、グラフィックス・コア・アレイ 3014 内の共有機能論理 3026 に有利なように除外される。

【0396】

1 つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推

10

20

30

40

50

論及び／又は訓練論理 8 1 5 が使用される。推論及び／又は訓練論理 8 1 5 に関する詳細事項は、図 8 A 及び／又は図 8 B と併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理 8 1 5 の一部又はすべてが、グラフィックス・プロセッサ 3 0 1 0 に組み込まれてもよい。たとえば、少なくとも一実施例では、本明細書に記載の訓練及び／又は推論の技法は、3 D パイプライン 3 0 1 2、グラフィックス・コア 3 0 1 5 A、共有機能論理 3 0 2 6、共有機能論理 3 0 2 0、又は図 3 0 の他の論理に具体化された A L U のうちの 1 つ又は複数を使用してもよい。さらに、少なくとも一実施例では、本明細書に記載の推論及び／又は訓練の動作は、図 8 A 又は図 8 B に示す論理以外の論理を使用して行われてもよい。少なくとも一実施例では、重みパラメータは、本明細書に記載の 1 つ又は複数の機械学習アルゴリズム、ニューラル・ネットワーク・アーキテクチャ、ユース・ケース、又は訓練技法を実行するためのグラフィックス・プロセッサ 3 0 1 0 の A L U を構成するオン・チップ若しくはオフ・チップのメモリ及び／又はレジスタ（図示する又は図示せず）に記憶されてもよい。

【0397】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0398】

図 3 1 は、本明細書に記載の少なくとも一実施例によるグラフィックス・プロセッサ・コア 3 1 0 0 のハードウェア論理のブロック図である。少なくとも一実施例では、グラフィックス・プロセッサ・コア 3 1 0 0 は、グラフィックス・コア・アレイ内に含まれる。少なくとも一実施例では、コア・スライスと呼ばれることもあるグラフィックス・プロセッサ・コア 3 1 0 0 は、モジュール式グラフィックス・プロセッサ内の 1 つ又は複数のグラフィックス・コアとすることができる。少なくとも一実施例では、グラフィックス・プロセッサ・コア 3 1 0 0 は、1 つのグラフィックス・コア・スライスの例示であり、本明細書に記載のグラフィックス・プロセッサは、目的の電力及び性能のエンベロップに基づき、複数のグラフィックス・コア・スライスを含んでもよい。少なくとも一実施例では、各グラフィックス・コア 3 1 0 0 は、汎用及び固定の機能論理のモジュール式ブロックを含むサブ・スライスとも呼ばれる複数のサブ・コア 3 1 0 1 A ～ 3 1 0 1 F に結合された固定機能ブロック 3 1 3 0 を含むことができる。

【0399】

少なくとも一実施例では、固定機能ブロック 3 1 3 0 は、たとえば低性能及び／又は低電力のグラフィックス・プロセッサ実装形態において、グラフィックス・プロセッサ 3 1 0 0 内のすべてのサブ・コアが共有できるジオメトリ及び固定機能パイプライン 3 1 3 6 を含む。少なくとも一実施例では、ジオメトリ及び固定機能パイプライン 3 1 3 6 は、3 D 固定機能パイプライン、ビデオ・フロント・エンド・ユニット、スレッド・スポナー（*spawner*）及びスレッド・ディスパッチャ、並びに統合リターン・バッファを管理する統合リターン・バッファ・マネージャを含む。

【0400】

少なくとも一実施例では、固定機能ブロック 3 1 3 0 はまた、グラフィックス S o C インターフェース 3 1 3 7、グラフィックス・マイクロコントローラ 3 1 3 8、及びメディア・パイプライン 3 1 3 9 を含む。少なくとも一実施例では、グラフィックス S o C インターフェース 3 1 3 7 は、グラフィックス・コア 3 1 0 0 と、システム・オン・チップ集積回路内の他のプロセッサ・コアとのインターフェースを提供する。少なくとも一実施例では、グラフィックス・マイクロコントローラ 3 1 3 8 は、スレッド・ディスパッチ、ス

ケジューリング、及びプリエンブションを含め、グラフィックス・プロセッサ 3100 の様々な機能を管理するように構成可能なプログラム可能サブ・プロセッサである。少なくとも一実施例では、メディア・パイプライン 3139 は、画像及びビデオのデータを含むマルチメディア・データのデコーディング、エンコーディング、前処理、及び／又は後処理を容易にする論理を含む。少なくとも一実施例では、メディア・パイプライン 3139 は、サブ・コア 3101～3101F 内のコンピュート論理又はサンプリング論理への要求を介して、メディア動作を実装する。

【0401】

少なくとも一実施例では、SoC インターフェース 3137 は、汎用アプリケーション・プロセッサ・コア（たとえば、CPU）、及び／又は SoC 内の他の構成要素と、グラフィックス・コア 3100 が通信できるようにし、SoC 内の他の構成要素には、共有ラスト・レベル・キャッシュ・メモリ、システム RAM、及び／又は組み込みオン・チップ若しくはオン・パッケージの DRAM などのメモリ階層要素が含まれる。少なくとも一実施例では、SoC インターフェース 3137 はまた、カメラ・イメージング・パイプラインなど、SoC 内の固定機能デバイスとの通信を可能にし、グラフィックス・コア 3100 と SoC 内の CPU との間で共有することができるグローバル・メモリ・アトミックの使用を可能にし、且つ／又はそれを実装する。少なくとも一実施例では、グラフィックス SoC インターフェース 3137 はまた、グラフィックス・プロセッサ・コア 3100 の電力管理制御を実装することができ、グラフィックス・プロセッサ・コア 3100 のクロック・ドメインと、SoC 内の他のクロック・ドメインとの間でインターフェースをとれるようにする。少なくとも一実施例では、SoC インターフェース 3137 は、グラフィックス・プロセッサ内の 1 つ又は複数のグラフィックス・コアのそれぞれにコマンド及び命令を提供するように構成されたコマンド・ストリーマ及びグローバル・スレッド・ディスパッチャから、コマンド・バッファを受信できるようにする。少なくとも一実施例では、コマンド及び命令は、メディア動作が実行されるときにはメディア・パイプライン 3139 にディスパッチされることが可能であり、又はグラフィックス処理動作が実行されるときには、ジオメトリ及び固定機能パイプライン（たとえば、ジオメトリ及び固定機能パイプライン 3136 及び／又はジオメトリ及び固定機能パイプライン 3114）にディスパッチされることが可能である。

【0402】

少なくとも一実施例では、グラフィックス・マイクロコントローラ 3138 は、グラフィックス・コア 3100 のための様々なスケジューリング及び管理タスクを実行するように構成されることが可能である。少なくとも一実施例では、グラフィックス・マイクロコントローラ 3138 は、サブ・コア 3101A～3101F 内の実行ユニット（EU: execution unit）アレイ 3102A～3102F、3104A～3104F 内の様々なグラフィックス並列エンジンで、グラフィックスを実行し、且つ／又はワークロードのスケジューリングをコンピュートすることができる。少なくとも一実施例では、グラフィックス・コア 3100 を含む SoC の CPU コア上で実行されているホスト・ソフトウェアは、複数のグラフィックス・プロセッサ経路のうちの 1 つにワークロードを送出することができ、この経路が、適切なグラフィックス・エンジンに対するスケジューリング動作を呼び出す。少なくとも一実施例では、スケジューリング動作は、どのワークロードを次に実行すべきかを判定すること、コマンド・ストリーマにワークロードを送出すること、エンジン上で実行されている既存のワークロードをプリエンブションすること、ワークロードの進行を管理すること、及びワークロードが完了したときにホスト・ソフトウェアに通知することを含む。少なくとも一実施例では、グラフィックス・マイクロコントローラ 3138 はまた、グラフィックス・コア 3100 の低電力又はアイドル状態を促進して、オペレーティング・システム及び／又はシステム上のグラフィックス・ドライバ・ソフトウェアとは無関係に、低電力状態の移行全体にわたってグラフィックス・コア 3100 内のレジスタを保存及び復元する機能をグラフィックス・コア 3100 に提供することができる。

10

20

30

40

50

【0403】

少なくとも一実施例では、グラフィックス・コア3100は、図示してあるサブ・コア3101A～3101Fより多くの、又はそれより少ない、N個までのモジュール式サブ・コアを有してもよい。N個のサブ・コアのセットごとに、少なくとも一実施例では、グラフィックス・コア3100はまた、共有機能論理3110、共有及び／又はキャッシュ・メモリ3112、ジオメトリ／固定機能パイプライン3114、並びに様々なグラフィックスを加速し、処理動作をコンピュータするための追加の固定機能論理3116を含むことができる。少なくとも一実施例では、共有機能論理3110は、グラフィックス・コア3100内の各N個のサブ・コアが共有できる論理ユニット（たとえば、サンブラ、数理、及び／又はスレッド間通信の論理）を含むことができる。少なくとも一実施例では、共有の、及び／又はキャッシュのメモリ3112は、グラフィックス・コア3100内のN個のサブ・コア3101A～3101Fのためのラスト・レベル・キャッシュとすることができ、また、複数のサブ・コアがアクセスできる共有メモリとしての役割も果たすことができる。少なくとも一実施例では、ジオメトリ／固定機能パイプライン3114は、固定機能ブロック3130内のジオメトリ／固定機能パイプライン3136の代わりに含まれてもよく、同様の論理ユニットを含むことができる。

10

【0404】

少なくとも一実施例では、グラフィックス・コア3100は、グラフィックス・コア3100が使用するための様々な固定機能加速論理を含むことができる追加の固定機能論理3116を含む。少なくとも一実施例では、追加の固定機能論理3116は、位置限定シェーディング（*position only shading*）に使用するための追加のジオメトリ・パイプラインを含む。位置限定シェーディングでは、少なくとも2つのジオメトリ・パイプラインが存在しているが、ジオメトリ及び固定機能パイプライン3114、3136内の完全ジオメトリ・パイプラインと選別パイプライン（*cull pipeline*）においてであり、この選別パイプラインは、追加の固定機能論理3116内に含まれてもよい追加のジオメトリ・パイプラインである。少なくとも一実施例では、選別パイプラインは、完全ジオメトリ・パイプラインの縮小版である。少なくとも一実施例では、完全パイプライン及び選別パイプラインは、アプリケーションの異なるインスタンスを実行することができ、各インスタンスは別個のコンテキストを有する。少なくとも一実施例では、位置限定シェーディングは、切り捨てられた三角形の長い選別ランを隠すことができ、いくつかのインスタンスにおいてシェーディングを早く完了させることができる。たとえば、少なくとも一実施例では、選別パイプラインは、ピクセルをフレーム・バッファにラスタ化及びレンダリングすることなく、頂点の位置属性をフェッチしシェーディングするので、追加の固定機能論理3116内の選別パイプライン論理は、メイン・アプリケーションと並列で位置シェーダを実行することができ、完全パイプラインよりも全体的に早く臨界結果（*critical result*）を生成する。少なくとも一実施例では、選別パイプラインは、生成された臨界結果を使用して、すべての三角形について、これらの三角形が選別されているかどうかに関わらず、可視性情報をコンピュータすることができる。少なくとも一実施例では、（このインスタンスではリプレイ・パイプラインと呼ばれてもよい）完全パイプラインは、可視性情報を消費して、選別された三角形を飛ばして可視三角形だけをシェーディングすることができ、この可視性三角形が、最終的にラスタ化フェーズに渡される。

20

30

40

【0405】

少なくとも一実施例では、追加の固定機能論理3116はまた、機械学習の訓練又は推論の最適化を含む実装形態のために、固定機能の行列乗算論理など、機械学習の加速論理を含むことができる。

【0406】

少なくとも一実施例では、各グラフィックス・サブ・コア3101A～3101F内において、実行リソースのセットを含み、このセットは、グラフィックス・パイプライン、メディア・パイプライン、又はシェーダ・プログラムからの要求に応答して、グラフィッ

50

クス動作、メディア動作、及びコンピュータ動作を実行するために使用されてもよい。少なくとも一実施例では、グラフィックス・サブ・コア3101A~3101Fは、複数のEUアレイ3102A~3102F、3104A~3104F、スレッド・ディスパッチ及びスレッド間通信(TD/IC: thread dispatch and inter-thread communication)論理3103A~3103F、3D(たとえば、テクスチャ)サンプラ3105A~3105F、メディア・サンプラ3106A~3106F、シェーダ・プロセッサ3107A~3107F、及び共有ローカル・メモリ(SLM: shared local memory)3108A~3108Fを含む。少なくとも一実施例では、EUアレイ3102A~3102F、3104A~3104Fはそれぞれ、複数の実行ユニットを含み、これらは、グラフィックス、メディア、又はコンピュータ・シェーダ・プログラムを含むグラフィックス動作、メディア動作、又はコンピュータ動作のサービスにおいて浮動小数点及び整数/固定小数点の論理演算を実行することができる汎用グラフィックス・プロセッシング・ユニットである。少なくとも一実施例では、TD/IC論理3103A~3103Fは、サブ・コア内の実行ユニットのためのローカル・スレッド・ディスパッチ及びスレッド制御動作を実行し、サブ・コアの実行ユニット上で実行されているスレッド間の通信を容易にする。少なくとも一実施例では、3Dサンプラ3105A~3105Fは、テクスチャ又は他の3Dグラフィックス関連のデータをメモリに読み取ることができる。少なくとも一実施例では、3Dサンプラは、所与のテクスチャに関連付けられた構成済みサンプル状態及びテクスチャ・フォーマットに基づき、テクスチャ・データを異なるやり方で読み取ることができる。少なくとも一実施例では、メディア・サンプラ3106A~3106Fは、メディア・データに関連付けられたタイプ及びフォーマットに基づき、同様の読取り動作を実行することができる。少なくとも一実施例では、各グラフィックス・サブ・コア3101A~3101Fは、代替的に3Dとメディアの統合サンプラを含むことができる。少なくとも一実施例では、各サブ・コア3101A~3101F内の実行ユニット上で実行しているスレッドは、スレッド・グループ内で実行しているスレッドが、オン・チップ・メモリの共通プールを使用して実行できるようにするために、各サブ・コア内の共有ローカル・メモリ3108A~3108Fを利用することができる。

【0407】

1つ若しくは複数の実施例に関連する推論及び/又は訓練の動作を実行するために、推論及び/又は訓練論理815が使用される。推論及び/又は訓練論理815に関する詳細事項は、図8A及び/又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び/又は訓練論理815の一部又はすべてが、グラフィックス・プロセッサ3110に組み込まれてもよい。たとえば、少なくとも一実施例では、本明細書に記載の訓練及び/又は推論の技法は、3Dパイプライン、グラフィックス・マイクロコントローラ3138、ジオメトリ及び固定機能パイプライン3114及び3136、又は図31の他の論理に具体化されたALUのうちの1つ又は複数を使用してもよい。さらに、少なくとも一実施例では、本明細書に記載の推論及び/又は訓練の動作は、図8A又は図8Bに示す論理以外の論理を使用して行われてもよい。少なくとも一実施例では、重みパラメータは、本明細書に記載の1つ又は複数の機械学習アルゴリズム、ニューラル・ネットワーク・アーキテクチャ、ユース・ケース、又は訓練技法を実行するためのグラフィックス・プロセッサ3100のALUを構成するオン・チップ若しくはオフ・チップのメモリ及び/又はレジスタ(図示する又は図示せず)に記憶されてもよい。

【0408】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1~図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1~図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によっ

てゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0409】

図32A及び図32Bは、少なくとも一実施例による、グラフィックス・プロセッサ・コアの処理要素のアレイを含むスレッド実行論理3200を示す。図32Aは、スレッド実行論理3200が使用される少なくとも一実施例を示す。図32Bは、少なくとも一実施例による、グラフィックス実行ユニット3208の例示的な内側細部を示す図である。

【0410】

図32Aに示すように、少なくとも一実施例では、スレッド実行論理3200は、シェーダ・プロセッサ3202、スレッド・ディスパッチャ3204、命令キャッシュ3206、複数の実行ユニット3207A～3207N、3208A～3208Nを含むスケラブル実行ユニット・アレイ、サンプラ3210、データ・キャッシュ3212、及びデータ・ポート3214を含む。少なくとも一実施例では、スケラブル実行ユニット・アレイは、1つ又は複数の実行ユニット（たとえば、実行ユニット3208A～N又は3207A～Nのうちのいずれか）を、たとえばワークロードの計算要件に基づき有効又は無効にすることによって、動的に拡大縮小することができる。少なくとも一実施例では、スケラブル実行ユニットは、実行ユニットのそれぞれにリンクされる相互接続ファブリックを介して相互接続される。少なくとも一実施例では、スレッド実行論理3200は、命令キャッシュ3206、データ・ポート3214、サンプラ3210、及び実行ユニット3207又は3208のうちの1つ又は複数を介した、システム・メモリ又はキャッシュ・メモリなどのメモリへの1つ又は複数の接続を含む。少なくとも一実施例では、各実行ユニット（たとえば、3207A）は、スレッドごとに複数のデータ要素を並列で処理しながら、複数の同時のハードウェア・スレッドを実行することができるスタンドアロンのプログラム可能な汎用計算ユニットである。少なくとも一実施例では、実行ユニット3207及び／又は3208のアレイは、任意の数の個々の実行ユニットを含むように拡大縮小可能である。

【0411】

少なくとも一実施例では、実行ユニット3207及び／又は3208は、シェーダ・プログラムを実行するために主に使用される。少なくとも一実施例では、シェーダ・プロセッサ3202は、様々なシェーダ・プログラムを処理し、シェーダ・プログラムに関連付けられた実行スレッドを、スレッド・ディスパッチャ3204を介してディスパッチすることができる。少なくとも一実施例では、スレッド・ディスパッチャ3204は、グラフィックス及びメディア・パイプラインからのスレッド開始要求を調停し、要求されたスレッドを、実行ユニット3207及び／又は3208の1つ又は複数の実行ユニット上でインスタンス化するための論理を含む。たとえば、少なくとも一実施例では、ジオメトリ・パイプラインは、頂点シェーダ、モザイク・シェーダ、又はジオメトリ・シェーダを、処理できるようにスレッド実行論理にディスパッチすることができる。少なくとも一実施例では、スレッド・ディスパッチャ3204はまた、実行しているシェーダ・プログラムからのラン・タイム・スレッド・スポーニング要求（`spawning request`）を処理することができる。

【0412】

少なくとも一実施例では、実行ユニット3207及び／又は3208は、多くの標準的な3Dグラフィックス・シェーダ命令のネイティブ・サポートを含む命令セットをサポートし、それにより、グラフィックス・ライブラリ（たとえば、`Direct3D`及び`OpenGL`）からのシェーダ・プログラムが、最小のトランスレーションで実行される。少なくとも一実施例では、実行ユニットは、頂点及びジオメトリの処理（たとえば、頂点プログラム、ジオメトリ・プログラム、及び／又は頂点シェーダ）、ピクセル処理（たとえば、ピクセル・シェーダ、フラグメント・シェーダ）、及び汎用処理（たとえば、コンピュータ及びメディアのシェーダ）をサポートする。少なくとも一実施例では、1つ又は複数の算術論理演算ユニット（`ALU`）を含む各実行ユニット3207及び／又は3208

のそれぞれは、単一命令複数データ（SIMD）の実行を複数発行することができ、マルチスレッド化された動作によって、メモリ・アクセスのレイテンシが高いにもかかわらず、効率的な実行環境が可能になる。少なくとも一実施例では、各実行ユニット内の各ハードウェア・スレッドは、専用の高帯域幅レジスタ・ファイル及び関連する独立したスレッド状態を有する。少なくとも一実施例では、実行は、整数演算、単精度及び倍精度の浮動小数点演算、SIMDブランチ性能、論理演算、超越演算、及び他の種々の演算を行うことができるパイプラインに対して、クロック当たり複数発行される。少なくとも一実施例では、メモリ、又は共有機能のうちの1つからのデータを待機している間に、実行ユニット3207及び／又は3208内の従属論理は、要求したデータが戻されるまで、待機スレッドをスリープ状態にする。少なくとも一実施例では、待機スレッドがスリープ状態の間に、ハードウェア・リソースは他のスレッドの処理に専念してもよい。たとえば、少なくとも一実施例では、頂点シェーダ動作に関連する遅延中に、実行ユニットは、ピクセル・シェーダ、フラグメント・シェーダ、又は異なる頂点シェーダを含む別のタイプのシェーダ・プログラムを実行することができる。

【0413】

少なくとも一実施例では、実行ユニット3207及び／又は3208の各実行ユニットは、データ要素のアレイに対して動作する。少なくとも一実施例では、データ要素の数は「実行サイズ」であり、又は命令に対するチャンネルの数である。少なくとも一実施例では、実行チャンネルは、データ要素のアクセス、マスキング、及び命令内のフロー制御に関する実行の論理ユニットである。少なくとも一実施例では、チャンネルの数は、特定のグラフィックス・プロセッサのための物理的な算術論理演算ユニット（ALU）又は浮動小数点ユニット（FPU）の数とは無関係であってもよい。少なくとも一実施例では、実行ユニット3207及び／又は3208は、整数及び浮動小数点のデータ・タイプをサポートしてもよい。

【0414】

少なくとも一実施例では、実行ユニット命令セットは、SIMD命令を含む。少なくとも一実施例では、様々なデータ要素が、パック・データ・タイプとしてレジスタに記憶されてもよく、実行ユニットは、要素のデータ・サイズに基づき様々な要素を処理する。たとえば、少なくとも一実施例では、256ビット幅ベクトルで動作しているとき、ベクトルの256ビットがレジスタに記憶され、実行ユニットは、4個の別々の64ビット・パック・データ要素（クワッド・ワード（QW：Quad-Word）サイズのデータ要素）、8個の別々の32ビット・パック・データ要素（ダブル・ワード（DW：Double Word）サイズのデータ要素）、16個の別々の16ビット・パック・データ要素（ワード（W：Word）サイズのデータ要素）、又は32個の別々の8ビット・データ要素（バイト（B：byte）サイズのデータ要素）としてベクトル上で動作する。しかし少なくとも一実施例では、異なるベクトル幅及びレジスタサイズが考えられる。

【0415】

少なくとも一実施例では、実行ユニット3207Aを実行ユニット3208Aと融合して融合実行ユニット3209Aにするなど、1つ又は複数の実行ユニットを組み合わせ、融合EUに共通のスレッド制御論理（3211A～3211N）を有する融合実行ユニット3209A～3209Nにすることができる。少なくとも一実施例では、複数のEUを融合して、EUグループにすることができる。少なくとも一実施例では、融合EUグループの各EUは、融合EUグループのEUの数が、様々な実施例に応じて異なる可能性がある状態で、別々のSIMDハードウェア・スレッドを実行するように構成されることが可能である。少なくとも一実施例では、SIMD8、SIMD16、及びSIMD32を含むがこれに限定されない様々なSIMD幅を、EUごとに実行することができる。少なくとも一実施例では、各融合グラフィックス実行ユニット3209A～3209Nは、少なくとも2つの実行ユニットを含む。たとえば、少なくとも一実施例では、融合実行ユニット3209Aは、第1のEU3207A、第2のEU3208B、及び第1のEU3207Aと第2のEU3208Aに共通のスレッド制御論理3211Aを含む。少な

10

20

30

40

50

くとも一実施例では、スレッド制御論理 3 2 1 1 A は、融合グラフィックス実行ユニット 3 2 0 9 A で実行されているスレッドを制御して、融合実行ユニット 3 2 0 9 A ~ 3 2 0 9 N 内の各 E U を、共通の命令ポインタ・レジスタを使用して実行できるようにする。

【0 4 1 6】

少なくとも一実施例では、1 つ又は複数の内部命令キャッシュ（たとえば、3 2 0 6）は、実行ユニットに対するスレッド命令をキャッシュするためにスレッド実行論理 3 2 0 0 に含まれる。少なくとも一実施例では、1 つ又は複数のデータ・キャッシュ（たとえば、3 2 1 2）は、スレッド実行中にスレッド・データをキャッシュするために含まれる。少なくとも一実施例では、サンブラ 3 2 1 0 は、3 D 動作のためのテクスチャ・サンプリング、及びメディア動作のためのメディア・サンプリングを実行するために含まれる。少なくとも一実施例では、サンブラ 3 2 1 0 は、特別なテクスチャ又はメディア・サンプリング機能を含み、サンプリングされたデータを実行ユニットに提供する前に、サンプリング処理中にテクスチャ又はメディアのデータを処理する。

【0 4 1 7】

実行中、少なくとも一実施例では、グラフィックス及びメディア・パイプラインは、スレッド開始要求を、スレッド・スポーニング及びディスパッチ論理を介してスレッド実行論理 3 2 0 0 に送る。少なくとも一実施例では、幾何学的物体のグループが処理され、ピクセル・データにラスタ化されたら、シェーダ・プロセッサ 3 2 0 2 内のピクセル・プロセッサ論理（たとえば、ピクセル・シェーダ論理、フラグメント・シェーダ論理など）が呼び出されて、出力情報をさらにコンピュートし、結果を出力面（たとえば、色バッファ、深度バッファ、ステンシル・バッファなど）に書き込ませる。少なくとも一実施例では、ピクセル・シェーダ又はフラグメント・シェーダは、ラスタ化された物体間で補間されることになる様々な頂点属性の値を計算する。少なくとも一実施例では、次いで、シェーダ・プロセッサ 3 2 0 2 内のピクセル・プロセッサ論理が、アプリケーション・プログラミング・インターフェース（A P I）付きのピクセル・シェーダ・プログラム又はフラグメント・シェーダ・プログラムを実行する。少なくとも一実施例では、シェーダ・プログラムを実行するために、シェーダ・プロセッサ 3 2 0 2 は、スレッド・ディスパッチャ 3 2 0 4 を介してスレッドを実行ユニット（たとえば、3 2 0 8 A）にディスパッチする。少なくとも一実施例では、シェーダ・プロセッサ 3 2 0 2 は、サンブラ 3 2 1 0 のテクスチャ・サンプリング論理を使用して、メモリに記憶されたテクスチャ・マップのテクスチャ・データにアクセスする。少なくとも一実施例では、テクスチャ・データ及び入力ジオメトリ・データに対する算術演算によって、各ジオメトリ・フラグメントのピクセル色データがコンピュートされ、又はさらに処理されないように 1 つ又は複数のピクセルが切り捨てられる。

【0 4 1 8】

少なくとも一実施例では、データ・ポート 3 2 1 4 は、スレッド実行論理 3 2 0 0 のためのメモリ・アクセス機構を提供して、処理済みデータを、グラフィックス・プロセッサ出力パイプラインでさらに処理できるようにメモリに出力する。少なくとも一実施例では、データ・ポート 3 2 1 4 は、1 つ又は複数のキャッシュ・メモリ（たとえば、データ・キャッシュ 3 2 1 2）を含み、又はそれに結合されて、データ・ポートを介したメモリ・アクセスのためのデータをキャッシュする。

【0 4 1 9】

図 3 2 B に示してあるように、少なくとも一実施例では、グラフィック実行ユニット 3 2 0 8 は、命令フェッチ・ユニット 3 2 3 7、汎用レジスタ・ファイル・アレイ（G R F : g e n e r a l r e g i s t e r f i l e a r r a y）3 2 2 4、アーキテクチャ・レジスタ・ファイル・アレイ（A R F）3 2 2 6、スレッド調停装置（a r b i t e r）3 2 2 2、送信ユニット 3 2 3 0、ブランチ・ユニット 3 2 3 2、S I M D 浮動小数点ユニット（F P U）3 2 3 4 のセット、及び専用整数 S I M D A L U 3 2 3 5 のセットを含むことができる。少なくとも一実施例では、G R F 3 2 2 4 及び A R F 3 2 2 6 は、各同時ハードウェア・スレッドに関連付けられた汎用レジスタ・ファイルとアーキテクチ

ャ・レジスタ・ファイルのセットを含み、このハードウェア・スレッドは、グラフィックス実行ユニット3208においてアクティブであってもよい。少なくとも一実施例では、スレッドごとのアーキテクチャ状態が、ARF3226において維持され、スレッド実行中に使用されるデータが、GRF3224に記憶される。少なくとも一実施例では、各スレッドに対する命令ポインタを含む各スレッドの実行状態は、ARF3226のスレッド専用レジスタに保持することが可能である。

【0420】

少なくとも一実施例では、グラフィックス実行ユニット3208は、同時マルチスレッディング(SMT: Simultaneous Multi-Threading)と微細化インターリーブ・マルチスレッディング(IMT: Interleaved Multi-Threading)の組合せであるアーキテクチャを有する。少なくとも一実施例では、アーキテクチャは、実行ユニット当たりの同時スレッドのターゲット数及びレジスタ数に基づき設計時に微調整することができるモジュール式構成を有し、ここで実行ユニットのリソースは、複数の同時スレッドを実行するために使用される論理にわたって分割される。

【0421】

少なくとも一実施例では、グラフィックス実行ユニット3208は複数の命令を共同発行することができ、この命令は、それぞれ異なる命令であってもよい。少なくとも一実施例では、グラフィックス実行ユニット・スレッド3208のスレッド調停装置3222は、送信ユニット3230、ブランチ・ユニット3232、又はSIMD FPU3234のうちの1つに命令をディスパッチして実行できるようにすることができる。少なくとも一実施例では、各実行スレッドは、GRF3224内の128個の汎用レジスタにアクセスすることができ、ここで各レジスタは、32ビットのデータ要素のSIMD8要素のベクトルとしてアクセス可能な32バイトを記憶することができる。少なくとも一実施例では、各実行ユニット・スレッドは、GRF3224内の4キロバイトにアクセスすることができるが、実施例はこのように限定されず、他の実施例ではより多くの、又はより少ないリソースが提供されてもよい。少なくとも一実施例では、最大7個のスレッドを同時に実行できるが、実行ユニット当たりのスレッド数も、実施例に応じて変えることができる。7個のスレッドが4キロバイトにアクセスできる少なくとも一実施例では、GRF3224は、合計28キロバイトを記憶することができる。少なくとも一実施例では、フレキシブルなアドレッシング・モードにより、複数のレジスタがともにアドレスされてより幅広いレジスタを構築したり、ストライド設定された矩形ブロック・データ構造を表したりできるようにすることができる。

【0422】

少なくとも一実施例では、メモリ動作、サンブラ動作、及び他のレイテンシの長いシステム通信は、メッセージ引渡し送信ユニット3230によって実行される「送信」命令を介してディスパッチされる。少なくとも一実施例では、ブランチ命令は、SIMDの発散及び最終的な収束を容易にするために、ブランチ・ユニット3232にディスパッチされる。

【0423】

少なくとも一実施例では、グラフィックス実行ユニット3208は、浮動小数点演算を実行するための1つ又は複数のSIMD浮動小数点ユニット(FPU)3234を含む。少なくとも一実施例では、FPU3234は、整数計算もサポートする。少なくとも一実施例ではFPU3234は、最大M個の32ビット浮動小数点(若しくは整数)演算をSIMDで実行し、又は最大で2M個の16ビット整数演算、若しくは16ビット浮動小数点演算をSIMDで実行することができる。少なくとも一実施例では、少なくとも1つのFPUは、拡張数理機能を提供して、高スループットの超越数理関数、及び倍精度の64ビット浮動小数点をサポートする。少なくとも一実施例では、8ビットの整数SIMD ALU3235のセットも存在し、機械学習計算に関連する動作を実行するように特に最適化されてもよい。

10

20

30

40

50

【0424】

少なくとも一実施例では、グラフィックス実行ユニット3208の複数のインスタンスのアレイが、グラフィックス・サブ・コア・グループ（たとえば、サブ・スライス）においてインスタンス化されてもよい。少なくとも一実施例では、実行ユニット3208は、複数の実行チャンネルにわたって命令を実行することができる。少なくとも一実施例では、グラフィックス実行ユニット3208で実行される各スレッドは、異なるチャンネルで実行される。

【0425】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、推論及び／又は訓練論理815の一部又はすべてが、スレッド実行論理3200に組み込まれてもよい。さらに、少なくとも一実施例では、本明細書に記載の推論及び／又は訓練の動作は、図8A又は図8Bに示す論理以外の論理を使用して行われてもよい。少なくとも一実施例では、重みパラメータは、本明細書に記載の1つ又は複数の機械学習アルゴリズム、ニューラル・ネットワーク・アーキテクチャ、ユース・ケース、又は訓練技法を実行するための実行論理3200のALUSレッドを構成するオン・チップ若しくはオフ・チップのメモリ及び／又はレジスタ（図示する又は図示せず）に記憶されてもよい。

【0426】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0427】

図33は、少なくとも一実施例による並列処理ユニット（「PPU」）3300を示す。少なくとも一実施例では、PPU3300は、PPU3300によって実行された場合に、本開示全体を通して記載するプロセス及び技法の一部又はすべてを、PPU3300に実行させる機械可読コードで構成される。少なくとも一実施例では、PPU3300はマルチスレッド・プロセッサであり、このプロセッサは、1つ又は複数の集積回路デバイスに実装され、コンピュータ可読命令（機械可読命令若しくは単に命令とも呼ばれる）を、複数のスレッドで並列に処理するように設計されたレイテンシ隠蔽技法としてマルチスレッディングを利用する。少なくとも一実施例では、スレッドとは、実行スレッドを指し、PPU3300によって実行されるように構成された命令のセットをインスタンス化したものである。少なくとも一実施例では、PPU3300は、液晶ディスプレイ（「LCD」）デバイスなどのディスプレイ・デバイスに表示できるように2次元（「2D」）画像データを生成するために、3次元（「3D」）グラフィックス・データを処理するためのグラフィックス・レンダリング・パイプラインを実装するように構成されたグラフィックス・プロセッシング・ユニット（「GPU」）である。少なくとも一実施例では、PPU3300を利用して、線形代数演算及び機械学習演算などの計算が実行される。図33は、単に例示を目的とした例示的な並列プロセッサを示しており、本開示の範囲内で企図されるプロセッサ・アーキテクチャの非限定的な例として解釈されるべきであり、同プロセッサに追加するため、且つ／又はそれを置き換えるために、任意の好適なプロセッサが利用されてもよいことが解釈されるべきである。

【0428】

少なくとも一実施例では、1つ又は複数のPPU3300は、高性能コンピューティング（「HPC」：High Performance Computing）、データ・セ

10

20

30

40

50

ンタ、及び機械学習のアプリケーションを加速するように構成される。少なくとも一実施例では、PPU3300は、以下の非限定的な例を含む深層学習システム及びアプリケーションを加速するように構成される：自律車両プラットフォーム、深層学習、高精度音声、画像、テキスト認識システム、インテリジェント・ビデオ分析、分子シミュレーション、創薬、病気診断、天気予報、ビッグ・データ分析、天文学、分子動態シミュレーション、金融モデリング、ロボット工学、工場自動化、リアル・タイム言語翻訳、オンライン検索最適化、及び個別化ユーザ推奨など。

【0429】

少なくとも一実施例では、PPU3300は、限定することなく、入力／出力（「I/O」）ユニット3306、フロント・エンド・ユニット3310、スケジューラ・ユニット3312、ワーク分配ユニット3314、ハブ3316、クロスバー（「Xbar」：crossbar）3320、1つ又は複数の汎用処理クラスタ（「GPC」：general processing cluster）3318、及び1つ又は複数のパーティション・ユニット（「メモリ・パーティション・ユニット」）3322を含む。少なくとも一実施例では、PPU3300は、1つ又は複数の高速GPU相互接続（「GPU相互接続」）3308を介してホスト・プロセッサ又は他のPPU3300に接続される。少なくとも一実施例では、PPU3300は、システム・バス3302を介してホスト・プロセッサ又は他の周辺デバイスに接続される。少なくとも一実施例では、PPU3300は、1つ又は複数のメモリ・デバイス（「メモリ」）3304を備えるローカル・メモリに接続される。少なくとも一実施例では、メモリ・デバイス3304は、限定することなく、1つ又は複数のダイナミック・ランダム・アクセス・メモリ（「DRAM」）デバイスを含む。少なくとも一実施例では、1つ又は複数のDRAMデバイスは、複数のDRAMダイが各デバイス内で積層された高帯域幅メモリ（「HBM」）サブシステムとして構成されても、且つ／又は構成可能であってもよい。

【0430】

少なくとも一実施例では、高速GPU相互接続3308は、有線ベースのマルチ・レーン通信リンクを指してもよく、このリンクは、拡張縮小するためにシステムによって使用され、1つ又は複数の中央処理装置（「CPU」）と組み合わせられた1つ又は複数のPPU3300を含み、PPU3300とCPUとの間のキャッシュ・コヒーレンス、及びCPUマスタリングをサポートする。少なくとも一実施例では、データ及び／又はコマンドは、高速GPU相互接続3308により、ハブ3316を介して、1つ又は複数のコピー・エンジン、ビデオ・エンコーダ、ビデオ・デコーダ、電力管理ユニット、及び図33に明示されていないこともある他の構成要素などのPPU3300の別のユニットに／から送信される。

【0431】

少なくとも一実施例では、I/Oユニット3306は、システム・バス3302を介してホスト・プロセッサ（図33には示さず）から通信（たとえば、コマンド、データ）を送受信するように構成される。少なくとも一実施例では、I/Oユニット3306は、システム・バス3302を介して直接、又は1つ若しくは複数の、メモリ・ブリッジなどの中間デバイスを介して、ホスト・プロセッサと通信する。少なくとも一実施例では、I/Oユニット3306は、システム・バス3302を介してPPU3300のうちの1つ又は複数などの1つ又は複数の他のプロセッサと通信してもよい。少なくとも一実施例では、I/Oユニット3306は、ペリフェラル・コンポーネント・インターコネクト・エクスプレス（「PCIe」）インターフェースを実装して、PCIeバスを介して通信できるようにする。少なくとも一実施例では、I/Oユニット3306は、外部デバイスと通信するためのインターフェースを実装する。

【0432】

少なくとも一実施例では、I/Oユニット3306は、システム・バス3302を介して受信したパケットをデコードする。少なくとも一実施例では、少なくともいくつかのパケットは、PPU3300に様々な動作を実行させるように構成されたコマンドを表す。

少なくとも一実施例では、I/Oユニット3306は、デコードされたコマンドを、コマンドによって指定されるPPU3300の様々な他のユニットに送信する。少なくとも一実施例では、コマンドは、フロント・エンド・ユニット3310に送信され、且つ／又はハブ3316、若しくは（図33には明示していない）1つ若しくは複数のコピー・エンジン、ビデオ・エンコーダ、ビデオ・デコーダ、電力管理ユニットなどのPPU3300の他のユニットに送信される。少なくとも一実施例では、I/Oユニット3306はPPU3300の様々な論理ユニット間で、通信をルーティングするように構成される。

【0433】

少なくとも一実施例では、ホスト・プロセッサによって実行されるプログラムは、ワークロードをPPU3300に提供して処理できるようにするバッファにおいて、コマンド・ストリームをエンコードする。少なくとも一実施例では、ワークロードは、命令と、これらの命令によって処理されることになるデータとを含む。少なくとも一実施例では、バッファは、ホスト・プロセッサとPPU3300の両方がアクセス（たとえば、書込み／読取り）可能なメモリ内の領域であり、ホスト・インターフェース・ユニットは、I/Oユニット3306によってシステム・バス3302を介して送信されるメモリ要求を介して、システム・バス3302に接続されたシステム・メモリ内のバッファにアクセスするように構成されてもよい。少なくとも一実施例では、ホスト・プロセッサは、バッファにコマンド・ストリームを書き込み、次いでコマンド・ストリームの開始点を指すポインタをPPU3300に送信し、それによりフロント・エンド・ユニット3310は、1つ又は複数のコマンド・ストリームを指すポインタを受信し、1つ又は複数のコマンド・ストリームを管理して、コマンド・ストリームからコマンドを読み取り、コマンドをPPU3300の様々なユニットに転送する。

【0434】

少なくとも一実施例では、フロント・エンド・ユニット3310は、1つ又は複数のコマンド・ストリームによって定義されるタスクを処理するように様々なGPC3318を構成するスケジューラ・ユニット3312に結合される。少なくとも一実施例では、スケジューラ・ユニット3312は、スケジューラ・ユニット3312によって管理される様々なタスクに関連する状態情報を追跡するように構成され、ここで状態情報は、どのGPC3318にタスクが割り当てられるか、タスクがアクティブか非アクティブか、タスクに関連付けられた優先レベルなどを示してもよい。少なくとも一実施例では、スケジューラ・ユニット3312は、GPC3318のうちの1つ又は複数において、複数のタスクの実行を管理する。

【0435】

少なくとも一実施例では、スケジューラ・ユニット3312は、GPC3318で実行するためのタスクをディスパッチするように構成されたワーク分配ユニット3314に結合される。少なくとも一実施例では、ワーク分配ユニット3314は、スケジューラ・ユニット3312から受信したスケジュール済みタスクの数を追跡し、ワーク分配ユニット3314は、GPC3318のそれぞれについて、ペンディング・タスク・プール、及びアクティブ・タスク・プールを管理する。少なくとも一実施例では、ペンディング・タスク・プールは、特定のGPC3318によって処理されるように割り当てられたタスクを含むいくつかのロット（たとえば、32ロット）を備え、アクティブ・タスク・プールは、GPC3318によりアクティブに処理されているタスクのためのいくつかのロット（たとえば、4ロット）を備え、それにより、GPC3318のうちの1つがタスクの実行を完了すると、GPC3318のアクティブ・タスク・プールからそのタスクが排除され、ペンディング・タスク・プールからの他のタスクが選択され、GPC3318で実行されるようにスケジューリングされる。少なくとも一実施例では、データ依存性が解決されるのを待機している間など、アクティブ・タスクがGPC3318上でアイドルである場合には、アクティブ・タスクがGPC3318から排除され、ペンディング・タスク・プールに戻され、その間に、ペンディング・タスク・プールの別のタスクが選択され、GPC3318で実行されるようにスケジューリングされる。

10

20

30

40

50

【0436】

少なくとも一実施例では、ワーク分配ユニット3314は、Xバー3320を介して1つ又は複数のGPC3318と通信する。少なくとも一実施例では、Xバー3320は、PPU3300のユニットのうちの多くを、PPU3300の別のユニットに結合する相互接続ネットワークであり、ワーク分配ユニット3314を特定のGPC3318に結合するように構成されることが可能である。少なくとも一実施例では、PPU3300の1つ又は複数の他のユニットも、ハブ3316を介してXバー3320に接続されてもよい。

【0437】

少なくとも一実施例では、タスクはスケジューラ・ユニット3312によって管理され、ワーク分配ユニット3314によってGPC3318のうちの1つにディスパッチされる。少なくとも一実施例では、GPC3318は、タスクを処理し、結果を生成するように構成される。少なくとも一実施例では、結果は、GPC3318内の他のタスクによって消費されてもよく、Xバー3320を介して異なるGPC3318にルーティングされてもよく、又はメモリ3304に記憶されてもよい。少なくとも一実施例では、結果を、パーティション・ユニット3322を介してメモリ3304に書き込むことができ、パーティション・ユニット3322は、メモリ3304への／からのデータの読取り及び書込みを行うためのメモリ・インターフェースを実装する。少なくとも一実施例では、結果を、高速GPU相互接続3308を介して別のPPU3304又はCPUに送信することができる。少なくとも一実施例では、図35と併せて本明細書でさらに詳細に説明されるように、PPU3300は、PPU3300に結合された別々の個別メモリ・デバイス3304の数に等しいU個のパーティション・ユニット3322を、限定することなく含む。

【0438】

少なくとも一実施例では、ホスト・プロセッサはドライバ・カーネルを実行し、このカーネルは、ホスト・プロセッサで実行されている1つ又は複数のアプリケーションがPPU3300で実行するための動作をスケジューリングできるようにするアプリケーション・プログラミング・インターフェース(API)を実装している。少なくとも一実施例では、複数のコンピュート・アプリケーションが、PPU3300によって同時に実行され、PPU3300は、複数のコンピュート・アプリケーションに対して、隔離、サービス品質(「QoS」：quality of service)、及び独立したアドレス空間を提供する。少なくとも一実施例では、アプリケーションは、PPU3300によって実行するための1つ又は複数のタスクをドライバ・カーネルに生成させる(たとえば、APIコールの形の)命令を生成し、ドライバ・カーネルは、PPU3300によって処理されている1つ又は複数のストリームにタスクを出力する。少なくとも一実施例では、各タスクは、ワーブと呼ばれてもよい関連スレッドの1つ又は複数のグループを備える。少なくとも一実施例では、ワーブは、並列に実行することができる複数の関連スレッド(たとえば、32個のスレッド)を備える。少なくとも一実施例では、連動スレッドとは、タスクを実行するための命令を含み、共有メモリを介してデータを交換する複数のスレッドを指してもよい。少なくとも一実施例では、スレッド及び連動スレッドは、図35と併せてさらに詳細に説明される。

【0439】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、深層学習アプリケーション・プロセッサは、PPU3300に提供される情報を予測又は推論するようにニューラル・ネットワークなどの機械学習モデルを訓練するために使用される。少なくとも一実施例では、深層学習アプリケーション・プロセッサ3300は、別のプロセッサ若しくはシステムによって、又はPPU3300によって訓練されてきた訓練済み機械学習モデル(たとえば、ニューラル・ネットワーク)に基づき、情報を推論又は予測するために使用される。少なくとも一実施例では、PPU3300は、本明細書に記載の1つ又は複数のニューラル・ネットワークのユース・ケースを実行するために

10

20

30

40

50

使用されてもよい。

【0440】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

10

【0441】

図34は、少なくとも一実施例による汎用処理クラスタ（「GPC」）3400を示す。少なくとも一実施例では、GPC3400は、図33のGPC3318である。少なくとも一実施例では、各GPC3400は、限定することなく、タスクを処理するためのいくつかのハードウェア・ユニットを含み、各GPC3400は、限定することなく、パイプライン・マネージャ3402、プレ・ラスト演算ユニット（「preROP」：pre-raster operations unit）3404、ラスト・エンジン3408、ワーク分配クロスバー（「WDX」：work distribution cross bar）3416、メモリ管理ユニット（「MMU」）3418、1つ又は複数のデータ処理クラスタ（「DPC」：Data Processing Clusters）3406、及びパーツの任意の好適な組合せを含む。

20

【0442】

少なくとも一実施例では、GPC3400の動作は、パイプライン・マネージャ3402によって制御される。少なくとも一実施例では、パイプライン・マネージャ3402は、GPC3400に配分されたタスクを処理するために1つ又は複数のDPC3406の構成を管理する。少なくとも一実施例では、パイプライン・マネージャ3402は、グラフィックス・レンダリング・パイプラインの少なくとも一部分を実装するように、1つ又は複数のDPC3406のうちの少なくとも1つを構成する。少なくとも一実施例では、DPC3406は、プログラム可能なストリーミング・マルチプロセッサ（「SM」：streaming multi-processor）3414で頂点シェーダ・プログラムを実行するように構成される。少なくとも一実施例では、パイプライン・マネージャ3402は、少なくとも一実施例では、ワーク分配ユニットから受信したパケットを、GPC3400内の適切な論理ユニットにルーティングするように構成され、いくつかのパケットは、preROP3404の固定機能ハードウェア・ユニット及び／又はラスト・エンジン3408にルーティングされてもよく、他のパケットは、プリミティブ・エンジン3412又はSM3414によって処理されるようにDPC3406にルーティングされてもよい。少なくとも一実施例では、パイプライン・マネージャ3402は、ニューラル・ネットワーク・モデル及び／又はコンピューティング・パイプラインを実装するように、DPC3406のうちの少なくとも1つを構成する。

30

【0443】

少なくとも一実施例では、preROPユニット3404は、少なくとも一実施例では、ラスト・エンジン3408及びDPC3406によって生成されたデータを、図33と併せて上でより詳細に説明したパーティション・ユニット3322のラスト動作（ROP）ユニットにルーティングするように構成される。少なくとも一実施例では、preROPユニット3404は、色ブレンディングの最適化を実行し、ピクセル・データを組織化し、アドレス・トランスレーションを実行し、その他の動作を行うように構成される。少なくとも一実施例では、ラスト・エンジン3408は、少なくとも一実施例では様々なラスト動作を実行するように構成されたいくつかの固定機能ハードウェア・ユニットを、限定することなく含み、ラスト・エンジン3408は、限定することなく、セットアップ・エンジン、粗いラスト・エンジン、選別エンジン、クリッピング・エンジン、細かいラス

40

50

タ・エンジン、タイル合体エンジン、及びこれらの任意の好適な組合せを含む。少なくとも一実施例では、セットアップ・エンジンは、変換された頂点を受信し、頂点によって定義された幾何プリミティブに関連付けられた平面方程式を生成し、平面方程式が、粗いラスタ・エンジンに送信されて、プリミティブに対するカバレッジ情報（たとえば、タイルのx、yカバレッジ・マスク）が生成され、粗いラスタ・エンジンの出力が、選別エンジンに送信され、ここでzテストに落ちたプリミティブに関連付けられたフラグメントが選別され、クリッピング・エンジンに送信され、ここで視錐台の外側にあるフラグメントがクリップされる。少なくとも一実施例では、クリッピング及び選別を通過したフラグメントは、細かいラスタ・エンジンに渡されて、セットアップ・エンジンによって生成された平面方程式に基づき、ピクセル・フラグメントに対する属性が生成される。少なくとも一実施例では、ラスタ・エンジン3408の出力は、DPC3406内に実装されたフラグメント・シェーダによってなど任意の好適なエンティティによって処理されることになるフラグメントを含む。

【0444】

少なくとも一実施例では、GPC3400に含まれる各DPC3406は、限定することなく、Mパイプ・コントローラ（「MPC」：M-Pipe Controller）3410、プリミティブ・エンジン3412、1つ又は複数のSM3414、及びこれらの任意の好適な組合せを含む。少なくとも一実施例では、MPC3410は、DPC3406の動作を制御して、パイプライン・マネージャ3402から受信したパケットを、DPC3406内の適切なユニットにルーティングする。少なくとも一実施例では、頂点に関連付けられたパケットは、頂点に関連付けられた頂点属性をメモリからフェッチするように構成されたプリミティブ・エンジン3412にルーティングされ、対照的に、シェーダ・プログラムに関連付けられたパケットは、SM3414に送信されてもよい。

【0445】

少なくとも一実施例では、SM3414は、いくつかのスレッドにより表されたタスクを処理するように構成されたプログラム可能なストリーミング・プロセッサを、限定することなく含む。少なくとも一実施例では、SM3414はマルチスレッド化されており、スレッドの特定のグループからの複数のスレッド（たとえば、32個のスレッド）を同時に実行するように構成され、単一命令複数データ（SIMD）アーキテクチャを実装し、ここでスレッドのグループ（ワーブ）内の各スレッドは、同じ命令セットに基づき、異なるデータ・セットを処理するように構成される。少なくとも一実施例では、スレッド・グループ内のすべてのスレッドが命令の共通のセットを実行する。少なくとも一実施例では、SM3414は、単一命令複数スレッド（SIMT）アーキテクチャを実装し、ここで、スレッド・グループの各スレッドは、命令の共通セットに基づき、異なるデータ・セットを処理するように構成されるが、スレッド・グループ内の個々のスレッドは、実行中に発散することが許容される。少なくとも一実施例では、プログラム・カウンタ、コール・スタック、及び実行状態がワーブごとに維持されて、ワーブ内のスレッドが発散するときに、ワーブ間の同時処理、及びワーブ内での直列実行が可能になる。別の実施例では、プログラム・カウンタ、コール・スタック、及び実行状態が個々のスレッドごとに維持されて、すべてのスレッド間、ワーブ内、及びワーブ間で等しい同時処理が可能になる。少なくとも一実施例では、実行状態が個々のスレッドごとに維持され、共通の命令を実行しているスレッドが、より効率的になるように収束され並列に実行されてもよい。SM3414の少なくとも一実施例は、本明細書でさらに詳細に説明される。

【0446】

少なくとも一実施例では、MMU3418は、GPC3400とメモリ・パーティション・ユニット（たとえば、図33のパーティション・ユニット3322）との間でインターフェースを提供し、MMU3418は、仮想アドレスから物理アドレスへのトランスレーション、メモリ保護、及びメモリ要求の調停を提供する。少なくとも一実施例では、MMU3418は、仮想アドレスからメモリの物理アドレスへのトランスレーションを実行するための1つ又は複数のトランスレーション・ルックアサイド・バッファ（「TLB」

10

20

30

40

50

)を提供する。

【0447】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、深層学習アプリケーション・プロセッサは、GPC3400に提供される情報を予測又は推論するようにニューラル・ネットワークなどの機械学習モデルを訓練するために使用される。少なくとも一実施例では、GPC3400は、別のプロセッサ若しくはシステムによって、又はGPC3400によって訓練されてきた訓練済み機械学習モデル（たとえば、ニューラル・ネットワーク）に基づき、情報を推論又は予測するために使用される。少なくとも一実施例では、GPC3400は、本明細書に記載の1つ又は複数のニューラル・ネットワークのユース・ケースを実行するために使用されてもよい。

10

【0448】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

20

【0449】

図35は、少なくとも一実施例による並列処理ユニット（「PPU」）のメモリ・パーティション・ユニット3500を示す。少なくとも一実施例では、パーティション・ユニット3500は、限定することなく、ラスタ演算（「ROP」）ユニット3502、レベル2（「L2」）キャッシュ3504、メモリ・インターフェース3506、及びそれらの任意の好適な組合せを含む。少なくとも一実施例では、メモリ・インターフェース3506はメモリに結合される。少なくとも一実施例では、メモリ・インターフェース3506は、高速データ転送のために、32、64、128、1024ビットのデータ・バスなどを実装してもよい。少なくとも一実施例では、PPUは、Uが正の整数である、U個のメモリ・インターフェース3506をパーティション・ユニット3500の対当たり1つのメモリ・インターフェース3506に組み込んでおり、ここでパーティション・ユニット3500の各対は、対応するメモリ・デバイスに接続される。たとえば、少なくとも一実施例では、PPUは、高帯域幅メモリ・スタック、又はグラフィックス・ダブル・データ・レート、バージョン5、同期ダイナミック・ランダム・アクセス・メモリ（「GDDR5 SDRAM」）など、最大Y個のメモリ・デバイスに接続されてもよい。

30

【0450】

少なくとも一実施例では、メモリ・インターフェース3506は、高帯域幅メモリの第2世代（「HBM2」：high bandwidth memory second generation）メモリ・インターフェースを実装し、YはUの半分に等しい。少なくとも一実施例では、HBM2メモリ・スタックは、PPUを有する物理パッケージに位置付けられて、従来のGDDR5 SDRAMシステムに比べて実質的な電力と面積の節約を実現する。少なくとも一実施例では、各HBM2スタックは、限定することなく4個のメモリ・ダイを含み、Y=4であり、各HBM2スタックは、1つのダイ当たり2つの128ビット・チャネルの合計8チャネル、及び1024ビットのデータ・バス幅を含む。少なくとも一実施例では、メモリは、1ビット・エラー訂正2ビット・エラー検出（「SECDED」：Single-Error Correcting Double-Error Detecting）エラー訂正コード（「ECC」）をサポートしてデータを保護する。少なくとも一実施例では、ECCは、データ破損を受けやすいコンピュータ・アプリケーションに、より高い信頼性を提供し得る。

40

50

【0451】

少なくとも一実施例では、PPUは、マルチ・レベルのメモリ階層を実装する。少なくとも一実施例では、メモリ・パーティション・ユニット3500は、統合されたメモリをサポートして、中央処理装置（「CPU」）及びPPUメモリに単一の統合された仮想アドレス空間を提供し、仮想メモリ・システム間でのデータの共有を可能にする。少なくとも一実施例では、他のプロセッサに位置付けられたメモリにPPUがアクセスする頻度を追跡して、より頻繁にページにアクセスしているPPUの物理メモリに、メモリ・ページが確実に移動されるようにする。少なくとも一実施例では、高速GPU相互接続3308は、アドレス・トランスレーション・サービスをサポートして、PPUが直接CPUのページ・テーブルにアクセスできるようにし、PPUによるCPUメモリへのフル・アクセスを実現する。

10

【0452】

少なくとも一実施例では、コピー・エンジンは、複数のPPU間、又はPPUとCPUの間で、データを転送する。少なくとも一実施例では、コピー・エンジンは、ページ・テーブルにマッピングされていないアドレスについてページ誤りを生成することができ、次いでメモリ・パーティション・ユニット3500がページ誤りに対応して、アドレスをページ・テーブルにマッピングし、その後で、コピー・エンジンが転送を実行する。少なくとも一実施例では、メモリは、複数のプロセッサ間でコピー・エンジンの複数の動作についてピン留めされて（たとえば、ページ移動不可能にされて）、実質的に利用可能なメモリを低減させる。少なくとも一実施例では、ハードウェアのページ誤りがある場合、メモリ・ページが常駐であるかどうかに関わらず、アドレスをコピー・エンジンに渡すことができ、コピー・プロセスは透過的である。

20

【0453】

少なくとも一実施例によれば、図33のメモリ3304又は他のシステム・メモリからのデータは、メモリ・パーティション・ユニット3500によってフェッチされ、L2キャッシュ3504に記憶され、このL2キャッシュは、オン・チップに位置付けられ、様々なGPC間で共有される。少なくとも一実施例では、各メモリ・パーティション・ユニット3500は、対応するメモリ・デバイスに関連付けられたL2キャッシュの少なくとも一部分を、限定することなく含む。少なくとも一実施例では、より低いレベルのキャッシュが、GPC内の様々なユニットに実装される。少なくとも一実施例では、図34のSM3414のそれぞれは、レベル1（「L1」）キャッシュを実装してもよく、ここでL1キャッシュは、特定のSM3414専用のプライベート・メモリであり、L2キャッシュ3504からのデータは、SM3414の機能ユニットで処理するために、L1キャッシュのそれぞれにフェッチされ記憶される。少なくとも一実施例では、L2キャッシュ3504は、メモリ・インターフェース3506及び図33に示されるXバー3320に結合される。

30

【0454】

少なくとも一実施例では、ROPユニット3502は、色圧縮、ピクセル・ブレンディングなど、ピクセル色に関係するグラフィックス・ラスタ演算を実行する。ROPユニット3502は、少なくとも一実施例では、ラスタ・エンジン3408と併せて深度テストを実装して、ピクセル・フラグメントに関連付けられたサンプル・ロケーションの深度を、ラスタ・エンジン3408の選別エンジンから受信する。少なくとも一実施例では、深度は、フラグメントに関連付けられたサンプル・ロケーションの深度バッファにおける対応する深度と比べてテストされる。少なくとも一実施例では、フラグメントが、サンプル・ロケーションの深度テストを通過すると、ROPユニット3502は、深度バッファを更新し、深度テストの結果をラスタ・エンジン3408に送信する。パーティション・ユニット3500の数はGPCの数とは異なってもよく、したがって、各ROPユニット3502は、少なくとも一実施例では、GPCのそれぞれに結合されてもよいことが理解されよう。少なくとも一実施例では、ROPユニット3502は、異なるGPCから受信したパケットを追跡し、ROPユニット3502によって生成された結果が、Xバー332

40

50

0を通してルーティングされることになるかを判定する。

【0455】

図36は、少なくとも一実施例による、ストリーミング・マルチプロセッサ（「SM」）3600を示す。少なくとも一実施例では、SM3600は、図34のSMである。少なくとも一実施例では、SM3600は、限定することなく、命令キャッシュ3602、1つ又は複数のスケジューラ・ユニット3604、レジスタ・ファイル3608、1つ又は複数の処理コア（「コア」）3610、1つ又は複数の特殊機能ユニット（「SFU」：special function unit）3612、1つ又は複数のロード／ストア・ユニット（「LSU」load/store unit）3614、相互接続ネットワーク3616、共有メモリ／レベル1（「L1」）キャッシュ3618、及び／又はこれらの任意の好適な組合せを含む。

10

【0456】

少なくとも一実施例では、ワーク分配ユニットは、並列処理ユニット（「PPU」）の汎用処理クラスタ（「GPC」）で実行するためにタスクをディスパッチし、各タスクは、GPC内の特定のデータ処理クラスタ（「DPC」）に配分され、タスクがシェーダ・プログラムに関連する場合には、タスクはSM3600のうちの1つに配分される。少なくとも一実施例では、スケジューラ・ユニット3604は、ワーク分配ユニットからタスクを受信し、SM3600に割り当てられた1つ又は複数のスレッド・ブロックについて命令スケジューリングを管理する。少なくとも一実施例では、スケジューラ・ユニット3604は、並列スレッドのワープとして実行できるようにスレッド・ブロックをスケジューリングし、ここで各スレッド・ブロックは、少なくとも1つのワープに配分される。少なくとも一実施例では、各ワープは、スレッドを実行する。少なくとも一実施例では、スケジューラ・ユニット3604は、複数の異なるスレッド・ブロックを管理して、異なるスレッド・ブロックにワープを配分し、次いで複数の異なる連動グループからの命令を、各クロック・サイクル中に様々な機能ユニット（たとえば、処理コア3610、SFU3612、及びLSU3614）にディスパッチする。

20

【0457】

少なくとも一実施例では、連動グループとは、通信するスレッドのグループを組織化するためのプログラミング・モデルを指し、このモデルは、スレッドが通信する粒度をデベロッパが表せるようにして、より豊富でより効率的な並列分解の表現を可能にする。少なくとも一実施例では、連動した起動APIは、並列アルゴリズムを実行できるようにスレッド・ブロック間の同期をサポートする。少なくとも一実施例では、従来のプログラミング・モデルのアプリケーションは、連動スレッドを同期するための単一の簡単な構造、すなわちスレッド・ブロックのすべてのスレッドにわたるバリア（たとえば、sync threads（）関数）を提供する。しかし、少なくとも一実施例では、プログラマは、スレッド・ブロックの粒度よりも小さいスレッド・グループを定義し、定義されたグループ内で同期して、集合的なグループ全般にわたる機能インターフェースの形で、より高い性能、設計の融通性、及びソフトウェア再利用を可能にしてもよい。少なくとも一実施例では、連動グループによって、プログラマは、サブ・ブロック（すなわち、単一スレッドと同じ大きさ）の粒度及びマルチ・ブロックの粒度において、スレッドのグループを明示的に定義し、連動グループ内のスレッドに対する同期などの集合的な動作を実行できるようになる。少なくとも一実施例では、プログラミング・モデルは、ソフトウェア境界を横切るクリーンな合成をサポートし、それにより、ライブラリ及びユーティリティ関数を、収束について仮定する必要なくそれらのローカルなコンテキスト内で安全に同期することができる。少なくとも一実施例では、連動グループのプリミティブは、プロデューサー・コンシューマ並列性、日和見並列性（opportunistic parallelism）、及びスレッド・ブロックのグリッド全体にわたるグローバルな同期を限定することなく含む新しいパターンの連動並列性を可能にする。

30

40

【0458】

少なくとも一実施例では、ディスパッチ・ユニット3606は、1つ又は複数の機能ユ

50

ニットに命令を送信するように構成され、スケジューラ・ユニット 3604 は、共通のワーブからの 2 つの異なる命令を、各クロック・サイクル中にディスパッチできるようにする 2 つのディスパッチ・ユニット 3606 を限定することなく含む。少なくとも一実施例では、各スケジューラ・ユニット 3604 は、単一のディスパッチ・ユニット 3606 又は追加のディスパッチ・ユニット 3606 を含む。

【0459】

少なくとも一実施例では、各 SM 3600 は、少なくとも一実施例では、SM 3600 の機能ユニットにレジスタのセットを提供するレジスタ・ファイル 3608 を限定することなく含む。少なくとも一実施例では、レジスタ・ファイル 3608 は、各機能ユニットがレジスタ・ファイル 3608 の専用部分に配分されるように、各機能ユニット間で分割される。少なくとも一実施例では、レジスタ・ファイル 3608 は、SM 3600 によって実行されている異なるワーブ間で分割され、レジスタ・ファイル 3608 は、機能ユニットのデータ経路に接続されたオペランド用の一時的なストレージを提供する。少なくとも一実施例では、各 SM 3600 は、限定することなく複数の L 処理コア 3610 を含み、ここで L は正の整数である。少なくとも一実施例では、各 SM 3600 は、限定することなく、多数の（たとえば、128 個以上の）個別の処理コア 3610 を含む。少なくとも一実施例では、各処理コア 3610 は、浮動小数点算術論理演算ユニット及び整数算術論理演算ユニットを限定することなく含む完全にパイプライン化された、単精度の、倍精度の、及び／又は混合精度の処理ユニットを限定することなく含む。少なくとも一実施例では、浮動小数点算術論理演算ユニットは、浮動小数点演算のための IEEE 754-2008 規格を実装する。少なくとも一実施例では、処理コア 3610 は、限定することなく、64 個の単精度（32 ビット）浮動小数点コア、64 個の整数コア、32 個の倍精度（64 ビット）浮動小数点コア、及び 8 個のテンソル・コアを含む。

【0460】

テンソル・コアは、少なくとも一実施例による行列演算を実行するように構成される。少なくとも一実施例では、1 つ又は複数のテンソル・コアは、処理コア 3610 に含まれる。少なくとも一実施例では、テンソル・コアは、ニューラル・ネットワークの訓練及び推論のための畳み込み演算など、深層学習の行列演算を実行するように構成される。少なくとも一実施例では、各テンソル・コアは、 4×4 の行列で動作し、行列の積和演算（matrix multiply and accumulate operation） $D = A \times B + C$ を実行し、ここで A、B、C、及び D は 4×4 の行列である。

【0461】

少なくとも一実施例では、行列乗算の入力 A 及び B は、16 ビットの浮動小数点行列であり、和の行列 C 及び D は、16 ビットの浮動小数点又は 32 ビットの浮動小数点行列である。少なくとも一実施例では、テンソル・コアは、32 ビットの浮動小数点の和を有する 16 ビットの浮動小数点入力データで動作する。少なくとも一実施例では、16 ビットの浮動小数点乗算は、64 個の演算を使用し、結果的に完全精度の積をもたらし、次いでその積が、 $4 \times 4 \times 4$ の行列乗算の他の中間積との 32 ビット浮動小数点加算を使用して加算される。テンソル・コアを使用して、少なくとも一実施例では、これらの小さい要素から構築される、はるかに大きい 2 次元又はさらに高次元の行列演算が実行される。少なくとも一実施例では、CUDA 9 C++ API などの API は、CUDA-C++ プログラムからテンソル・コアを効率的に使用するために、特殊な行列ロード演算、行列積和演算、及び行列ストア演算を公開している。少なくとも一実施例では、CUDA レベルにおいて、ワーブ・レベル・インターフェースは、ワーブの 32 スレッドすべてにわたる 16×16 のサイズの行列を仮定している。

【0462】

少なくとも一実施例では、各 SM 3600 は、特殊関数（たとえば、属性評価、逆数平方根など）を実行する M 個の SFU 3612 を、限定することなく含む。少なくとも一実施例では、SFU 3612 は、限定することなく、階層ツリー・データ構造をトラバースするように構成されたツリー・トラバーサル・ユニットを含む。少なくとも一実施例では

、SFU3612は、テクスチャ・マップのフィルタリング動作を実行するように構成されたテクスチャ・ユニットを、限定することなく含む。少なくとも一実施例では、テクスチャ・ユニットは、メモリ及びサンプル・テクスチャ・マップからテクスチャ・マップ（たとえば、テクセルの2Dアレイ）をロードして、SM3600により実行されるシェーダ・プログラムで使用するためのサンプリングされたテクスチャ値を生成するように構成される。少なくとも一実施例では、テクスチャ・マップは、共有メモリ／レベル1キャッシュ3618に記憶される。少なくとも一実施例では、テクスチャ・ユニットは、少なくとも一実施例によれば、ミップ・マップ（たとえば、詳細さのレベルが異なるテクスチャ・マップ）を使用したフィルタリング動作などのテクスチャ動作を実装する。少なくとも一実施例では、各SM3600は、限定することなく、2つのテクスチャ・ユニットを含む。

10

【0463】

各SM3600は、少なくとも一実施例では、共有メモリ／L1キャッシュ3618とレジスタ・ファイル3608の間でロード及びストア動作を実装するN個のLSU3614を、限定することなく含む。少なくとも一実施例では、相互接続ネットワーク3616は、各機能ユニットをレジスタ・ファイル3608に接続し、LSU3614をレジスタ・ファイル3608及び共有メモリ／L1キャッシュ3618に接続する。少なくとも一実施例では、相互接続ネットワーク3616はクロスバーであり、このクロスバーは、任意の機能ユニットをレジスタ・ファイル3608の任意のレジスタに接続し、LSU3614をレジスタ・ファイル3608と共有メモリ／L1キャッシュ3618のメモリ・ロケーションとに接続するように構成されてもよい。

20

【0464】

少なくとも一実施例では、共有メモリ／L1キャッシュ3618は、少なくとも一実施例では、SM3600とプリミティブ・エンジンの間、及びSM3600のスレッド間でデータ・ストレージ及び通信を可能にするオン・チップ・メモリのアレイである。少なくとも一実施例では、共有メモリ／L1キャッシュ3618は、限定することなく、128KBのストレージ容量を備え、SM3600からパーティション・ユニットに向かう経路にある。少なくとも一実施例では、共有メモリ／L1キャッシュ3618は、少なくとも一実施例では、読取り及び書込みをキャッシュするために使用される。少なくとも一実施例では、共有メモリ／L1キャッシュ3618、L2キャッシュ、及びメモリのうちの1つ又は複数は、補助ストレージである。

30

【0465】

少なくとも一実施例では、データ・キャッシュと共有メモリ機能とを単一のメモリ・ブロックに組み合わせることによって、両方のタイプのメモリ・アクセスについて性能が向上する。少なくとも一実施例では、容量は、共有メモリを使用しないプログラムによってキャッシュとして使用され、又は使用可能であり、それにより、共有メモリが容量の半分を使用するように構成されている場合、テクスチャ及びロード／ストア動作が、残りの容量を使用することができる。少なくとも一実施例によれば、共有メモリ／L1キャッシュ3618内に統合することによって、共有メモリ／L1キャッシュ3618が、データをストリームするための高スループットの管として機能しながら、同時に高帯域幅及び低レイテンシのアクセスを、頻繁に再使用されるデータに提供できるようになる。少なくとも一実施例では、汎用並列計算向けに構成されるときには、グラフィックス処理と比べてより簡単な構成を使用することができる。少なくとも一実施例では、固定機能のグラフィックス・プロセッシング・ユニットがバイパスされて、はるかに簡単なプログラミング・モデルが作製される。汎用並列計算の構成では、ワーク分配ユニットは、少なくとも一実施例においてスレッド・ブロックを直接DPCに割当て及び分配する。少なくとも一実施例では、ブロック内のスレッドは、各スレッドが確実に一意の結果を生成するように、計算において一意のスレッドIDを使用して共通のプログラムを実行し、SM3600を使用して、プログラムを実行し計算を行い、共有メモリ／L1キャッシュ3618を使用してスレッド間で通信し、LSU3614を使用して、共有メモリ／L1キャッシュ3618

40

50

及びメモリ・パーティション・ユニットを介してグローバル・メモリを読み取り、書き込む。少なくとも一実施例では、汎用並列計算向けに構成されるときには、SM3600は、DCP上で新規のワークを起動するためにスケジューラ・ユニット3604が使用できるコマンドを書き込む。

【0466】

少なくとも一実施例では、PPUは、デスクトップ・コンピュータ、ラップトップ・コンピュータ、タブレット・コンピュータ、サーバ、スーパーコンピュータ、スマート・フォン（たとえば、ワイヤレスの携帯型デバイス）、パーソナル・デジタル・アシスタント（「PDA」）、デジタル・カメラ、車両、頭装着型ディスプレイ、携帯型電子デバイスなどに含まれ、又はこれらに結合される。少なくとも一実施例では、PPUは、単一の半導体基板に具体化される。少なくとも一実施例では、PPUは、追加のPPU、メモリ、縮小命令セット・コンピュータ（「RISC」）CPU、メモリ管理ユニット（「MMU」）、デジタル・アナログ変換器（「DAC」：digital-to-analog converter）などの1つ又は複数の他のデバイスとともにシステム・オン・チップ（「SoC」）に含まれる。

【0467】

少なくとも一実施例では、PPUは、1つ又は複数のメモリ・デバイスを含むグラフィックス・カードに含まれてもよい。少なくとも一実施例では、グラフィックス・カードは、デスクトップ・コンピュータのマザーボード上のPCIeスロットとインターフェースをとるように構成されてもよい。少なくとも一実施例では、PPUは、マザーボードのチップセットに含まれる統合グラフィックス・プロセッシング・ユニット（「iGPU」：integrated graphics processing unit）であってもよい。

【0468】

1つ若しくは複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細事項は、図8A及び／又は図8Bと併せて本明細書に提供される。少なくとも一実施例では、深層学習アプリケーション・プロセッサは、SM3600に提供される情報を予測又は推論するようにニューラル・ネットワークなどの機械学習モデルを訓練するために使用される。少なくとも一実施例では、SM3600は、別のプロセッサ若しくはシステムによって、又はSM3600によって訓練されてきた訓練済み機械学習モデル（たとえば、ニューラル・ネットワーク）に基づき、情報を推論又は予測するために使用される。少なくとも一実施例では、SM3600は、本明細書に記載の1つ又は複数のニューラル・ネットワークのユース・ケースを実行するために使用されてもよい。

【0469】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0470】

医療用途における画像推論及び画像処理など、先進コンピューティングのための仮想化コンピューティング・プラットフォームに関する実施例が開示される。限定することなく、実施例は、放射線写真撮影、磁気共鳴撮像（MRI）、核医学、超音波、ソノグラフィ、エラストグラフィ、光音響撮像、トモグラフィ、心エコー、機能的近赤外分光法、及び磁性粒子撮像、又はこれらの組合せを含んでもよい。少なくとも一実施例では、本明細書に記載の仮想化コンピューティング・プラットフォーム、及び関連するプロセスは、限定

10

20

30

40

50

することなく、法科学分析、表面下検出及び撮像（たとえば、石油探査、考古学、古生物学など）、地形学、海洋学、地質学、骨学、気象学、知能分野、又は物体追跡及び監視、センサ・データ処理（たとえば、RADAR、SONAR、LIDARなど）、並びに／又はゲノミクス及び遺伝子配列において、追加的又は代替的に使用されてもよい。

【0471】

図37を参照すると、図37は、少なくとも一実施例による、画像処理及び推論のパイプラインを生成及び導入するプロセス3700の実例データ・フロー図である。少なくとも一実施例では、プロセス3700は、医療施設、病院、ヘルスケア機関、クリニック、リサーチ若しくは診断の研究所などの1つ又は複数の施設3702において、撮像デバイス、処理デバイス、ゲノミクス・デバイス、遺伝子配列デバイス、放射線デバイス、及び／又は他のタイプのデバイスとともに使用するために導入されてもよい。少なくとも一実施例では、プロセッサ3700は、シーケンシング・データについてゲノミクスの分析及び推論を実行するために導入されてもよい。本明細書に記載のシステム及びプロセスを使用して実行することができるゲノム分析の実例は、限定することなく、バリエーション・コール、変異検出、及び遺伝子発現の定量化を含む。

【0472】

少なくとも一実施例では、プロセス3700は、訓練システム3704内及び／又は導入システム3706内で実行されてもよい。少なくとも一実施例では、訓練システム3704を使用して、導入システム3706で使用するための機械学習モデル（たとえば、ニューラル・ネットワーク、物体検出アルゴリズム、コンピュータ・ビジョン・アルゴリズムなど）の訓練、導入、及び実装が実行されてもよい。少なくとも一実施例では、導入システム3706は、処理及び計算のリソースを分散コンピューティング環境間でオフロードするように構成されて、施設3702におけるインフラストラクチャ要件を低減してもよい。少なくとも一実施例では、導入システム3706は、施設3702において撮像デバイス（たとえば、MRI、CTスキャン、X線、超音波など）又はシーケンシング・デバイスとともに使用するための仮想機器を選択し、カスタマイズし、実装するための合理化されたプラットフォームを提供してもよい。少なくとも一実施例では、仮想機器は、撮像デバイス、シーケンシング・デバイス、放射線デバイス、及び／又は他のタイプのデバイスによって生成された撮像データに対して1つ又は複数の処理動作を実行するためのソフトウェア定義アプリケーションを含んでもよい。少なくとも一実施例では、パイプライン内の1つ又は複数のアプリケーションは、アプリケーションの実行中に導入システム3706のサービス（たとえば、推論、仮想化、計算、AIなど）を使用又はコールしてもよい。

【0473】

少なくとも一実施例では、先進処理及び推論パイプラインで使用されるアプリケーションのいくつかは、1つ又は複数の処理ステップを実行するために機械学習モデル又は他のAIを使用してもよい。少なくとも一実施例では、機械学習モデルは、施設3702で生成された（且つ、施設3702において1つ若しくは複数の画像アーカイブ及び通信システム（PACS）サーバに記憶された）（撮像データなどの）データ3708を使用して、施設3702において訓練されてもよく、1つ又は複数の別の施設（たとえば、異なる病院、研究所、クリニックなど）からの撮像若しくはシーケンシングのデータ3708を使用して訓練されてもよく、又はそれらの組合せであってもよい。少なくとも一実施例では、訓練システム3704を使用して、導入システム3706向けの実用的で導入可能な機械学習モデルを生成するためのアプリケーション、サービス、及び／又は他のリソースが提供されてもよい。

【0474】

少なくとも一実施例では、モデル・レジストリ3724は、バージョン管理及び物体メタデータをサポートすることができる物体ストレージによってバックアップされてもよい。少なくとも一実施例では、物体ストレージには、たとえば、クラウド・プラットフォーム内から、クラウド・ストレージ（たとえば、図38のクラウド3826）の互換性アプ

リケーション・プログラミング・インターフェース（API）を介してアクセス可能であってもよい。少なくとも一実施例では、モデル・レジストリ3724内の機械学習モデルは、システムの開発者又はパートナーがAPIと対話することによって、アップロード、リスト化、修正、又は削除されてもよい。少なくとも一実施例では、APIは、適切な資格を有するユーザがモデルをアプリケーションに関連付けできるようにする方法へのアクセスを提供してもよく、それによりアプリケーションのコンテナ化されたインスタンスを実行することの一部として、モデルを実行できるようになる。

【0475】

少なくとも一実施例では、訓練パイプライン3804（図38）は、施設3702が独自の機械学習モデルを訓練している状況、又は最適化若しくは更新される必要がある既存の機械学習モデルを有している状況を含んでもよい。少なくとも一実施例では、撮像デバイス、シーケンシング・デバイス、及び／又は他のタイプのデバイスによって生成された撮像データ3708が受信されてもよい。少なくとも一実施例では、撮像データ3708が受信されると、機械学習モデルのグラウンド・トゥールズ・データとして使用されることになる撮像データ3708に対応するアノテーションの生成を支援するために、AI支援アノテーション3710が使用されてもよい。少なくとも一実施例では、AI支援アノテーション3710は、1つ又は複数の機械学習モデル（たとえば、畳み込みニューラル・ネットワーク（CNN））を含んでもよく、これは（たとえば特定のデバイスからの）特定のタイプの撮像データ3708、及び／又は撮像データ3708内の特定のタイプの異常に対応するアノテーションを生成するように訓練されてもよい。少なくとも一実施例では、次いでAI支援アノテーション3710は、グラウンド・トゥールズ・データを生成するために直接使用されてもよく、又は（たとえば、研究者、臨床医、医師、科学者などによって）アノテーション・ツールを使用して調節若しくは微調整されてもよい。少なくとも一実施例では、いくつかの実例において、ラベル付けされたクリニック・データ3712（たとえば、臨床医、医師、科学者、技術者などによって提供されたアノテーション）が、機械学習モデルを訓練するためのグラウンド・トゥールズ・データとして使用されてもよい。少なくとも一実施例では、AI支援アノテーション3710、ラベル付けされたクリニック・データ3712、又はこれらの組合せが、機械学習モデルを訓練するためのグラウンド・トゥールズ・データとして使用されてもよい。少なくとも一実施例では、訓練済み機械学習モデルは出力モデル3716と呼ばれてもよく、本明細書に記載の導入システム3706によって使用されてもよい。

【0476】

少なくとも一実施例では、訓練パイプライン3804（図38）は、施設3702が、導入システム3706内の1つ又は複数のアプリケーションのための1つ又は複数の処理タスクを実行する際に使用する機械学習モデルを必要としているが、施設3702は現在そのような機械学習モデルを有していないかもしれない（又はそうした目的のために最適化された、効率よい、若しくは有効なモデルを有していないかもしれない）という状況を含んでもよい。少なくとも一実施例では、既存の機械学習モデルが、モデル・レジストリ3724から選択されてもよい。少なくとも一実施例では、モデル・レジストリ3724は、撮像データに対して様々な異なる推論タスクを実行するように訓練された機械学習モデルを含んでもよい。少なくとも一実施例では、モデル・レジストリ3724の機械学習モデルは、施設3702とは異なる施設（たとえば、離れた場所にある施設）からの撮像データについて訓練されたものであってもよい。少なくとも一実施例では、機械学習モデルは、1つの場所、2つの場所、又は任意の数の場所からの撮像データについて訓練されたものであってもよい。少なくとも一実施例では、特定の場所からの撮像データについて訓練されるとき、訓練は、その場所で行われてもよく、又は少なくとも、撮像データの機密性を保護するようなやり方で、若しくは撮像データが構外へ転送されるのを制限するようなやり方で（たとえば、HIPPA規定、プライバシー規定に準拠するように）行われてもよい。少なくとも一実施例では、1つの場所においてモデルが訓練されると、又は部分的に訓練されると、機械学習モデルはモデル・レジストリ3724に加えられてもよい

10

20

30

40

50

。少なくとも一実施例では、次いで機械学習モデルは、任意の数の他の施設において再訓練又は更新されてもよく、再訓練又は更新されたモデルが、モデル・レジストリ 3724 において利用可能にされてもよい。少なくとも一実施例では、次いで機械学習モデルは、モデル・レジストリ 3724 から選択されてもよく、出力モデル 3716 と呼ばれてもよく、導入システム 3706 において使用されて、導入システムの 1 つ又は複数のアプリケーションのための 1 つ又は複数の処理タスクを実行してもよい。

【0477】

少なくとも一実施例では、訓練パイプライン 3804 (図 38) は、施設 3702 が、導入システム 3706 内の 1 つ又は複数のアプリケーションのための 1 つ又は複数の処理タスクを実行する際に使用する機械学習モデルを必要としているが、施設 3702 は現在そのような機械学習モデルを有していないかもしれない (又はそうした目的のために最適化された、効率よい、若しくは有効なモデルを有していないかもしれない) という状況を含むシナリオで使用する事ができる。少なくとも一実施例では、モデル・レジストリ 3724 から選択された機械学習モデルは、母集団、遺伝的差異、機械学習モデルを訓練するために使用される訓練データの頑健性、訓練データの異常の多様性、及び／又は訓練データに伴う他の問題に違いがあることから、施設 3702 において生成される撮像データ 3708 向けに微調整又は最適化されていないことがある。少なくとも一実施例では、機械学習モデルを再訓練又は更新するためのグラウンド・トゥールズ・データとして使用されることになる撮像データ 3708 に対応するアノテーションの生成を支援するために、AI 支援アノテーション 3710 が使用されてもよい。少なくとも一実施例では、ラベル付けされたクリニック・データ 3712 (たとえば、臨床医、医師、科学者、技術者などによって提供されたアノテーション) が、機械学習モデルを訓練するためのグラウンド・トゥールズ・データとして使用されてもよい。少なくとも一実施例では、機械学習モデルを再訓練又は更新することは、モデル訓練 3714 と呼ばれてもよい。少なくとも一実施例では、モデル訓練 3714、たとえば AI 支援アノテーション 3710、ラベル付けされたクリニック・データ 3712、又はこれらの組合せは、機械学習モデルを再訓練若しくは更新するためのグラウンド・トゥールズ・データとして使用されてもよい。

【0478】

少なくとも一実施例では、導入システム 3706 は、ソフトウェア 3718、サービス 3720、ハードウェア 3722、並びに／又は他の構成要素、特徴、及び機能を含んでもよい。少なくとも一実施例では、導入システム 3706 は、ソフトウェア「スタック」を含んでもよく、それによりソフトウェア 3718 は、サービス 3720 の上に構築されてもよく、サービス 3720 を使用して一部若しくはすべての処理タスクを実行してもよく、サービス 3720 及びソフトウェア 3718 は、ハードウェア 3722 の上に構築され、ハードウェア 3722 を使用して、導入システム 3706 の処理、ストレージ、及び／又は他の計算のタスクを実行してもよい。

【0479】

少なくとも一実施例では、ソフトウェア 3718 は、任意の数の異なるコンテナを含んでもよく、ここで各コンテナは、アプリケーションのインスタンス化を実行してもよい。少なくとも一実施例では、各アプリケーションは、先進処理及び推論パイプラインの 1 つ又は複数の処理タスク (たとえば、推論、物体検出、特徴検出、セグメント化、画像強調、キャリブレーションなど) を実行してもよい。少なくとも一実施例では、撮像デバイス (たとえば、CT、MRI、X 線、超音波、ソノグラフィ、心エコーなど)、シーケンシング・デバイス、放射線デバイス、ゲノミクス・デバイスなどのタイプごとに、デバイスによって生成された撮像データ 3708 (又は、本明細書に記載のものなどの他のタイプのデータ) に対してデータ処理タスクを実行できる任意の数のコンテナが存在してもよい。少なくとも一実施例では、先進処理及び推論パイプラインは、(たとえば、医用におけるデジタル画像と通信 (digital imaging and communications in medicine: DICOM) データ、放射線医学情報システム (RIS) データ、臨床情報システム (CIS) データ、リモート・プロシージャ・コール (R

10

20

30

40

50

PC) データ、表現状態転送 (REST) インターフェースに実質的に準拠したデータ、ファイルベースのインターフェースに実質的に準拠したデータ、及び／又は生のデータなどの使用可能なタイプのデータに出力を再変換して、施設 3702 において記憶及び表示するように) パイプラインを通して処理した後に、各コンテナによって使用される、且つ／又は施設 3702 によって使用される撮像データを受信及び構成するコンテナに加えて、撮像データ 3708 を処理するのに所望される又は必要とされる異なるコンテナの選択に基づき定義されてもよい。少なくとも一実施例では、(たとえばパイプラインを構成する) ソフトウェア 3718 内のコンテナの組合せは、(本明細書においてより詳細に記載する) 仮想機器と呼ばれてもよく、仮想機器は、サービス 3720 及びハードウェア 3722 を利用して、コンテナにおいてインスタンス化されたアプリケーションの一部又はすべての処理タスクを実行してもよい。

10

【0480】

少なくとも一実施例では、データ処理パイプラインは、推論要求(たとえば、臨床医、医師、放射線医など、導入システム 3706 のユーザからの要求)に応答して、DICOM、RIS、CIS、REST 準拠、RPC、生、及び／又は他のフォーマットで入力データ(たとえば、撮像データ 3708)を受け取ってもよい。少なくとも一実施例では、入力データは、1つ又は複数の撮像デバイス、シーケンシング・デバイス、放射線デバイス、ゲノミクス・デバイス、及び／又は他のタイプのデバイスによって生成される1つ又は複数の画像、ビデオ、及び／又は他のデータ表現を表してもよい。少なくとも一実施例では、データは、データ処理パイプラインの一部としての事前処理を受けて、1つ又は複数のアプリケーションによって処理できるようにデータが準備されてもよい。少なくとも一実施例では、パイプラインの1つ若しくは複数の推論タスク又は他の処理タスクの出力に対して後処理が実行されて、次のアプリケーション用に出力データが準備されてもよく、且つ／又は送信及び／若しくはユーザによる使用のために(たとえば、推論要求への応答として)出力データが準備されてもよい。少なくとも一実施例では、推論タスクは、訓練済み若しくは導入済みのニューラル・ネットワークなど、1つ又は複数の機械学習モデルによって実行されてもよく、このモデルは、訓練システム 3704 の出力モデル 3716 を含んでもよい。

20

【0481】

少なくとも一実施例では、データ処理パイプラインのタスクはコンテナにカプセル化されてもよく、コンテナはそれぞれ、アプリケーションの個別の完全に機能的なインスタンス化、及び機械学習モデルを参照できる仮想化コンピューティング環境を表す少なくとも一実施例では、コンテナ又はアプリケーションは、(本明細書においてより詳細に記載する) コンテナ・レジストリのプライベート(たとえば、アクセスの制限された)区域に発行されてもよく、訓練済み又は導入済みのモデルは、モデル・レジストリ 3724 に記憶され、1つ又は複数のアプリケーションに関連付けられてもよい。少なくとも一実施例では、アプリケーションの画像(たとえば、コンテナの画像)は、コンテナ・レジストリにおいて入手可能であってもよく、パイプラインに導入するためにユーザによってコンテナ・レジストリから選択されると、画像は、ユーザのシステムで使用できるようにアプリケーションをインスタンス化するためのコンテナを生成するために使用されてもよい。

30

40

【0482】

少なくとも一実施例では、開発者(たとえば、ソフトウェア開発者、臨床医、医師など)は、供給されたデータに対して画像処理及び／又は推論を実行するために、アプリケーションを(たとえばコンテナとして)開発、公開、及び記憶してもよい。少なくとも一実施例では、開発、公開、及び／又は記憶は、(たとえば、開発されたアプリケーション及び／又はコンテナが、確実にシステムに準拠するように、又はシステムと互換性があるようにするために)システムに関連付けられたソフトウェア開発キット(SDK)を使用して実行されてもよい。少なくとも一実施例では、開発されたアプリケーションは、システム(たとえば図38のシステム3800)としてサービス3720の少なくとも一部をサポートすることができるSDKを用いて、ローカルに(たとえば第1の施設において、第

50

1の施設からのデータについて)テストされてもよい。少なくとも一実施例では、D I C O M物体は、1個から数百個にわたる画像又は他のタイプのデータをどこにでも含むことができるうえに、データのバリエーションがあることから、開発者は、入力されるD I C O Mデータの抽出及び準備を管理する(たとえば、アプリケーション用の構成を設定する、事前処理をアプリケーションに構築するなどの)責任を負うことがある。少なくとも一実施例では、システム3800によって(たとえば、精度、安全性、患者のプライバシーなどが)検証されると、アプリケーションは、ユーザ(たとえば、病院、クリニック、研究所、ヘルスケア提供者など)によって選択及び/又は実装できるようにコンテナ・レジストリにおいて利用可能にされて、ユーザの施設(たとえば、第2の施設)におけるデータに対して1つ又は複数の処理タスクが実行されてもよい。

10

【0483】

少なくとも一実施例では、次いで開発者は、アプリケーション又はコンテナを、システム(たとえば、図38のシステム3800)のユーザによってアクセス及び使用できるようにネットワークを通して共有してもよい。少なくとも一実施例では、完成し検証されたアプリケーション又はコンテナは、コンテナ・レジストリに記憶されてもよく、関連する機械学習モデルは、モデル・レジストリ3724に記憶されてもよい。少なくとも一実施例では、推論又は画像処理の要求を出す要求元エンティティ(たとえば、医療施設のユーザ)は、コンテナ・レジストリ及び/又はモデル・レジストリ3724をブラウジングしてアプリケーション、コンテナ、データセット、機械学習モデルなどを探し、データ処理パイプラインに含めるための要素の所望の組合せを選択し、撮像処理要求を送出してもよい。少なくとも一実施例では、要求は、要求を実行するために必要な入力データ(及びいくつかの実例では、関連する患者データ)を含んでもよく、且つ/又は要求を処理する際に実行されることになるアプリケーション及び/又は機械学習モデルの選択を含んでもよい。少なくとも一実施例では、次いで要求は、導入システム3706(たとえばクラウド)の1つ又は複数の構成要素に渡されて、データ処理パイプラインの処理が実行されてもよい。少なくとも一実施例では、導入システム3706による処理は、コンテナ・レジストリ及び/又はモデル・レジストリ3724から選択された要素(たとえば、アプリケーション、コンテナ、モデルなど)を参照することを含んでもよい。少なくとも一実施例では、パイプラインによって結果が生成されると、結果がユーザに返されて参照されてもよい(たとえば、ローカルで、構内のワークステーション又は端末で実行している視聴アプリケーション・スイートで視聴されてもよい)。少なくとも一実施例では、放射線医は、任意の数のアプリケーション及び/又はコンテナを含むデータ処理パイプラインから結果を受信してもよく、ここで結果は、X線、CTスキャン、MRIなどにおける異常検出を含んでもよい。

20

30

【0484】

少なくとも一実施例では、パイプラインにおけるアプリケーション又はコンテナの処理又は実行を支援するために、サービス3720が利用されてもよい。少なくとも一実施例では、サービス3720は、計算サービス、人工知能(AI)サービス、視覚化サービス、及び/又は他のタイプのサービスを含んでもよい。少なくとも一実施例では、サービス3720は、ソフトウェア3718の1つ又は複数のアプリケーションに共通の機能を提供してもよく、それにより機能は、アプリケーションによってコール又は利用されることが可能なサービスに対して抽象化されてもよい。少なくとも一実施例では、サービス3720によって提供される機能は、動的でより効率的に実行されてもよく、それと同時に、(たとえば、並列コンピューティング・プラットフォーム3830(図38)を使用して)アプリケーションが並列にデータを処理できるようにすることにより、良好にスケーリングされてもよい。少なくとも一実施例では、サービス3720により提供される同じ機能を共有する各アプリケーションに、サービス3720のそれぞれのインスタンスを有するよう要求するのではなく、サービス3720が、様々なアプリケーション間で共有されてもよい。少なくとも一実施例では、サービスは、非限定的な実例として、検出又はセグメント化のタスクを実行するために使用されてもよい推論のサーバ又はエンジンを含んで

40

50

もよい。少なくとも一実施例では、機械学習モデルの訓練及び／又は再訓練の機能を提供することができるモデル訓練サービスが含まれてもよい。少なくとも一実施例では、GPU加速化データ（たとえば、DICOM、RIS、CIS、REST準拠、RPC、生など）の抽出、リサイズ、スケーリング、及び／又は他の拡張を提供することができるデータ拡張サービスがさらに含まれてもよい。少なくとも一実施例では、レイ・トレーシング、ラスタ化、ノイズ除去、鮮鋭化などの画像レンダリング効果を加えることができる視覚化サービスが使用されて、2次元（2D）及び／又は3次元（3D）のモデルにリアル感が付加されてもよい。少なくとも一実施例では、仮想機器のパイプライン内の他のアプリケーションについてビーム形成、セグメント化、推論、撮像、及び／又はサポートを実現する仮想機器サービスが含まれてもよい。

10

【0485】

少なくとも一実施例では、サービス3720がAIサービス（たとえば、推論サービス）を含む場合、異常検出（たとえば、腫瘍、発育異常、瘢痕化など）のためのアプリケーションに関連付けられた1つ又は複数の機械学習モデルは、機械学習モデル、又はその処理を、アプリケーション実行の一部として実行するように推論サービス（たとえば、推論サーバ）に（APIコールとして）コールすることによって、実行されてもよい。少なくとも一実施例では、セグメント化タスクのための1つ又は複数の機械学習モデルを別のアプリケーションを含む場合、セグメント化タスクに関連付けられた処理動作のうちの1つ又は複数を実行するための機械学習モデルを実行するように、アプリケーションは推論サービスをコールしてもよい。少なくとも一実施例では、セグメント化アプリケーション及び異常検出アプリケーションを含む先進処理及び推論パイプラインを実装するソフトウェア3718は、1つ又は複数の推論タスクを実行するためにそれぞれのアプリケーションが同じ推論サービスをコールすることがあるので、合理化されてもよい。

20

【0486】

少なくとも一実施例では、ハードウェア3722は、GPU、CPU、グラフィックス・カード、AI／深層学習システム（たとえば、NVIDIAのDGXスーパーコンピュータ・システムなどのAIスーパーコンピュータ）、クラウド・プラットフォーム、又はそれらの組合せを含んでもよい。少なくとも一実施例では、異なるタイプのハードウェア3722を使用して、導入システム3706のソフトウェア3718及びサービス3720のための効率的で専用のサポートが提供されてもよい。少なくとも一実施例では、画像処理、画像再構築、セグメント化、MRI検査、脳卒中又は心臓発作の（たとえばリアルタイムの）検出、レンダリングの画像品質などの効率、精度、及び有効性を向上させるために、AI／深層学習システム内、クラウド・システム、及び／又は導入システム3706の他の処理構成要素において、ローカルで（たとえば、施設3702で）処理をおこなうためのGPU処理の使用が実装されてもよい。少なくとも一実施例では、施設は、撮像デバイス、ゲノミクス・デバイス、シーケンシング・デバイス、及び／又は他のタイプのデバイスを構内に含んでもよく、これらは、GPUを利用して、対象者の解剖学的組織を表す撮像データを生成してもよい。

30

【0487】

少なくとも一実施例では、ソフトウェア3718及び／又はサービス3720は、非限定的な実例として深層学習、機械学習、及び／又は高性能コンピューティングに関するGPU処理のために最適化されてもよい。少なくとも一実施例では、導入システム3706及び／又は訓練システム3704のコンピューティング環境のうちの少なくとも一部は、データセンタの1つ若しくは複数のスーパーコンピュータ、又は高性能コンピューティング・システムにおいて、GPU最適化ソフトウェア（たとえば、NVIDIAのDGXシステムのハードウェアとソフトウェアの組合せ）を用いて実行されてもよい。少なくとも一実施例では、データセンサは、HIPAAの条項に準拠してもよく、したがって、撮像データ及び／又は他の患者データの受信、処理、及び送信は、患者データのプライバシーに関して安全に取り扱われる。少なくとも一実施例では、ハードウェア3722は、任意の数のGPUを含んでもよく、これらのGPUは、本明細書に記載するように、データの

40

50

並列処理を実行するためにコールされてもよい。少なくとも一実施例では、クラウド・プラットフォームはさらに、深層学習タスク、機械学習タスク、又は他のコンピューティング・タスクのGPU最適化された実行のためのGPU処理を含んでもよい。少なくとも一実施例では、クラウド・プラットフォーム（たとえば、NVIDIAのNGC）は、（たとえば、NVIDIAのDGXシステムによって提供される）AI／深層学習スーパーコンピュータ、及び／又はGPU最適化ソフトウェアをハードウェア抽象化及びスケーリングのプラットフォームとして使用して、実行されてもよい。少なくとも一実施例では、クラウド・プラットフォームは、シームレスなスケーリング及びロード・バランシングを可能にするために、複数のGPUに対するアプリケーション・コンテナ・クラスターリング・システム又はオーケストレーション・システム（たとえば、KUBERNETES）を統合してもよい。

10

【0488】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

20

【0489】

図38は、少なくとも一実施例による撮像導入パイプラインを生成及び導入するための実例システム3800を示すシステム図である。少なくとも一実施例では、システム3800は、図37のプロセス3700、及び／又は先進処理及び推論パイプラインを含む他のプロセスを実装するために使用されてもよい。少なくとも一実施例では、システム3800は、訓練システム3704及び導入システム3706を含んでもよい。少なくとも一実施例では、訓練システム3704及び導入システム3706は、本明細書に記載するように、ソフトウェア3718、サービス3720、及び／又はハードウェア3722を使用して実装されてもよい。

【0490】

少なくとも一実施例では、システム3800（たとえば、訓練システム3704及び／又は導入システム3706）は、クラウド・コンピューティング環境（たとえば、クラウド3826）において実装されてもよい。少なくとも一実施例では、システム3800は、ヘルスケア・サービス施設に関してローカルに実装されてもよく、又はクラウドとローカル・コンピューティング・リソースとの組合せとして実装されてもよい。少なくとも一実施例では、クラウド・コンピューティングが実装される実施例では、HIPAA並びに／又は他のデータ取扱い及びプライバシーの規定若しくは法律に準拠していない処理を提供するシステム3800の1つ又は複数の構成要素から、患者データは分離されてもよく、又はそれらによって処理されなくてもよい。少なくとも一実施例では、クラウド3826のAPIへのアクセスは、制定されたセキュリティ対策又はプロトコルを介して許可されたユーザに限定されてもよい。少なくとも一実施例では、セキュリティ・プロトコルはウェブ・トークンを含んでもよく、このウェブ・トークンは、認証（たとえば、AuthN、AuthZ、Glueconなど）のサービスによって署名されてもよく、適切な許可を持っていてもよい。少なくとも一実施例では、（本明細書に記載の）仮想機器のAPI、又はシステム3800の他のインスタンス化は、対話について検査済み又は許可済みのパブリックIPのセットに限定されてもよい。

30

40

【0491】

少なくとも一実施例では、システム3800の様々な構成要素は、有線及び／又は無線の通信プロトコルを介して、ローカル・エリア・ネットワーク（LAN）及び／又は広域ネットワーク（WAN）を含むがこれらに限定されない様々な異なるタイプのネットワー

50

クのうちの任意のものを使用して、相互に通信してもよい。少なくとも一実施例では、（たとえば推論要求を送信するため、推論要求の結果を受信するためなど）施設とシステム 3800 の構成要素との間の通信は、1 つ又は複数のデータ・バス、無線データ・プロトコル（W i - F i）、有線データ・プロトコル（たとえば、イーサネット（登録商標））などを介して通信されてもよい。

【0492】

少なくとも一実施例では、訓練システム 3704 は、図 37 に関して本明細書に記載したものと同様の訓練パイプライン 3804 を実行してもよい。少なくとも一実施例では、1 つ又は複数の機械学習モデルが導入システム 3706 により導入パイプライン 3810 において使用されることになる場合、訓練パイプライン 3804 を使用して、1 つ又は複数の（たとえば、事前訓練された）モデルが訓練若しくは再訓練されてもよく、且つ／又は事前訓練されたモデル 3806 のうちの 1 つ又は複数が（たとえば再訓練若しくは更新を必要とせずに）実装されてもよい。少なくとも一実施例では、訓練パイプライン 3804 の結果として、出力モデル 3716 が生成されてもよい。少なくとも一実施例では、訓練パイプライン 3804 は、（たとえば D I C O M 画像を、それぞれの機械学習モデルによって処理するのに適した別のフォーマット、たとえば Neuro imaging Informatics Technology Initiative (N I f T I) フォーマットなどに変換するための D I C O M アダプタ 3802 A を使用した）撮像データ（若しくは他の入力データ）の変換若しくは適合、A I 支援アノテーション 3710、ラベル付きクリニック・データ 3712 を生成するための撮像データ 3708 のラベル付け又はアノテーション付け、モデル・レジストリからのモデル選択、モデル訓練 3714、モデルの訓練、再訓練、若しくは更新、及び／又は他の処理ステップなどであるがこれらに限定されない任意の数の処理ステップを含んでもよい。少なくとも一実施例では、導入システム 3706 によって使用される異なる機械学習モデルについて、異なる訓練パイプライン 3804 が使用されてもよい。少なくとも一実施例では、図 37 に関して記載した第 1 の実例と同様の訓練パイプライン 3804 は、第 1 の機械学習モデルに使用されてもよく、図 37 に関して記載した第 2 の実例と同様の訓練パイプライン 3804 は、第 2 の機械学習モデルに使用されてもよく、図 37 に関して記載した第 3 の実例と同様の訓練パイプライン 3804 は、第 3 の機械学習モデルに使用されてもよい。少なくとも一実施例では、それぞれの各機械学習モデルに要求されるものに依じて、訓練システム 3704 内のタスクの任意の組合せが使用されてもよい。少なくとも一実施例では、機械学習モデルのうちの 1 つ又は複数は、すでに訓練済みで導入の準備が整っていてもよく、それにより機械学習モデルは、訓練システム 3704 によるいかなる処理も受けなくてもよく、導入システム 3706 によって実装されてもよい。

【0493】

少なくとも一実施例では、出力モデル 3716 及び／又は事前訓練されたモデル 3806 は、実装形態又は実施例に応じて任意のタイプの機械学習モデルを含んでもよい。少なくとも一実施例では、限定することなく、システム 3800 によって使用される機械学習モデルは、線形回帰、ロジスティック回帰、決定木、サポート・ベクター・マシン（S V M）、ナイーブ・ベイズ、k 近傍法（k - n e a r e s t n e i g h b o r : K n n）、k 平均クラスタリング、ランダム・フォレスト、次元縮小アルゴリズム、勾配ブースティング・アルゴリズム、ニューラル・ネットワーク（たとえば、オート・エンコーダ、畳み込み、再帰、パーセプトロン、長／短期メモリ（L S T M）、ホップフィールド、ボルツマン、ディープ・ビリーフ、逆畳み込み、敵対的生成、液体状態マシンなど）を使用する機械学習モデル、及び／又は他のタイプの機械学習モデルを含んでもよい。

【0494】

少なくとも一実施例では、訓練パイプライン 3804 は、少なくとも図 41B に関して、より詳細に本明細書に記載する A I 支援アノテーションを含んでもよい。少なくとも一実施例では、ラベル付きクリニック・データ 3712（たとえば、従来のアノテーション）は、任意の数の技法によって生成されてもよい。少なくとも一実施例では、ラベル又は

他のアノテーションは、描画プログラム（たとえば、アノテーション・プログラム）、コンピュータ支援設計（ＣＡＤ）プログラム、ラベル付けプログラム、グラウンド・トゥールース用のアノテーション若しくはラベルの生成に適した別のタイプのプログラム内で生成されてもよく、且つ／又はいくつかの実例では、手書きされてもよい。少なくとも一実施例では、グラウンド・トゥールース・データは、合成により生成されてもよく（たとえば、コンピュータ・モデル又はレンダリングから生成されてもよく）、現実的に生成されてもよく（たとえば、実世界のデータから設計及び生成されてもよく）、機械自動化されてもよく（たとえば、特徴の分析及び学習を使用して、データから特徴を抽出し、次いでラベルを生成してもよく）、人間によりアノテーション付けされてもよく（たとえば、ラベラ、又はアノテーション専門家がラベルのロケーションを定義してもよく）、且つ／又はこれらの組合せであってもよい。少なくとも一実施例では、撮像データ３７０８のインスタンス（又は機械学習モデルによって使用される他のタイプのデータ）ごとに、訓練システム３７０４によって生成される対応するグラウンド・トゥールース・データが存在してもよい。少なくとも一実施例では、訓練パイプライン３８０４に含まれるＡＩ支援アノテーションに加えて又はその代わりに、導入パイプライン３８１０の一部としてＡＩ支援アノテーションが実行されてもよい。少なくとも一実施例では、システム３８００は多層プラットフォームを含んでもよく、このプラットフォームは、１つ又は複数の医療用撮像及び診断の機能を実行することができる診断アプリケーション（又は他のタイプのアプリケーション）のソフトウェア層（たとえば、ソフトウェア３７１８）を含んでもよい。少なくとも一実施例では、システム３８００は、１つ又は複数の施設のＰＡＣＳサーバ・ネットワークに、（たとえば、暗号化リンクを介して）通信可能に結合されてもよい。少なくとも一実施例では、システム３８００は、ＰＡＣＳサーバからのデータ（たとえば、ＤＩＣＯＭデータ、ＲＩＳデータ、生データ、ＣＩＳデータ、ＲＥＳＴ準拠データ、ＲＰＣデータ、生データなど）に（たとえば、ＤＩＣＯＭアダプタ３８０２、又はＲＩＳ、ＣＩＳ、ＲＥＳＴ準拠、ＲＰＣ、生など別のタイプのデータ・アダプタを介して）アクセスし、それを参照するように構成されて、機械学習モデルの訓練、機械学習モデルの導入、画像処理、推論、及び／又は他の動作などの動作を実行してもよい。

【０４９５】

少なくとも一実施例では、ソフトウェア層は、セキュアな、暗号化された、且つ／又は認証されたＡＰＩとして実装されてもよく、これを介して、アプリケーション又はコンテナが、外部環境（たとえば、施設３７０２）から呼び出し（たとえばコール）されてもよい。少なくとも一実施例では、次いでアプリケーションは、それぞれのアプリケーションに関連付けられた計算、ＡＩ、又は視覚化のタスクを実行するために１つ又は複数のサービス３７２０をコール又は実行してもよく、ソフトウェア３７１８及び／又はサービス３７２０は、ハードウェア３７２２を利用して、処理タスクを有効且つ効率的なやり方で実行してもよい。

【０４９６】

少なくとも一実施例では、導入システム３７０６は、導入パイプライン３８１０を実行してもよい。少なくとも一実施例では、導入パイプライン３８１０は任意の数のアプリケーションを含んでもよく、これらは、上に記載のＡＩ支援アノテーションを含め、撮像デバイス、シーケンシング・デバイス、ゲノミクス・デバイスなどによって生成された撮像データ（及び／又は他のタイプのデータ）に連続的に、非連続的に、又は他のやり方で適用されてもよい。少なくとも一実施例では、本明細書に記載するように、個々のデバイス用の導入パイプライン３８１０は、デバイス用の仮想機器（たとえば、仮想超音波機器、仮想ＣＴスキャン機器、仮想シーケンシング機器など）と呼ばれてもよい。少なくとも一実施例では、デバイスによって生成されるデータに必要な情報に応じて、１つのデバイスにつき２つ以上の導入パイプライン３８１０が存在してもよい。少なくとも一実施例では、異常検出がＭＲＩマシンに必要とされる場合、第１の導入パイプライン３８１０が存在してもよく、画像強調がＭＲＩマシンの出力に必要とされる場合、第２の導入パイプライン３８１０が存在してもよい。

10

20

30

40

50

【0497】

少なくとも一実施例では、導入パイプライン3810にとって利用可能なアプリケーションは、デバイスからの撮像データ又は他のデータに対して処理タスクを実行するために使用することができる任意のアプリケーションを含んでもよい。少なくとも一実施例では、画像強調、セグメント化、再構築、異常検出、物体検出、特徴検出、処置計画、線量測定、ビーム計画（又は他の放射線処置手順）、及び／又は他の分析、画像処理、又は推論のタスクを、異なるアプリケーションが担当してもよい。少なくとも一実施例では、導入システム3706は、それぞれのアプリケーションの構造を定義してもよく、それにより導入システム3706のユーザ（たとえば、医療施設、研修所、クリニックなど）は、構造を理解し、自らのそれぞれの施設内で実装できるようにアプリケーションを適応させてもよい。少なくとも一実施例では、導入パイプライン3810に含めるために、画像再構築用のアプリケーションが選択されてもよいが、撮像デバイスによって生成されるデータのタイプは、アプリケーション内で使用されるデータのタイプとは異なってもよい。少なくとも一実施例では、DICOMアダプタ3802B（及び／又はDICOMリーダ）は、又は別のタイプのデータ・アダプタ若しくはリーダ（たとえば、RIS、CIS、REST準拠、RPC、生など）が導入パイプライン3810内で使用されて、導入システム3706内のアプリケーションによって使用可能な形にデータを変換してもよい。少なくとも一実施例では、DICOM、RIS、CIS、REST準拠、RPC、生、及び／又は他のタイプのデータ・ライブラリへのアクセスは、データに対する任意の畳み込み、色補正、鮮明度、ガンマ、及び／又は他の拡張を、デコード、抽出、及び／又は実行することを含め、累積され、事前処理されてもよい。少なくとも一実施例では、DICOM、RIS、CIS、REST準拠、RPC、及び／又は生データは、順序なしであってもよく、収集されたデータを整理しソートするために、事前パスが実行されてもよい。少なくとも一実施例では、様々なアプリケーションは共通の画像動作を共有することができるので、いくつかの実施例では、（たとえば、サービス3720の1つとして）データ拡張ライブラリを使用して、これらの動作が加速化されてもよい。少なくとも一実施例では、CPU処理に依存する従来の処理手法のボトルネックを回避するために、並列コンピューティング・プラットフォーム3830を使用して、これらの処理タスクがGPU加速化されてもよい。

【0498】

少なくとも一実施例では、画像再構築アプリケーションは、機械学習モデルの使用を含む処理タスクを含んでもよい。少なくとも一実施例では、ユーザは、独自の機械学習モデルを使用すること、又はモデル・レジストリ3724から機械学習モデルを選択することを望む場合がある。少なくとも一実施例では、ユーザは、処理タスクを実行するために、独自の機械学習モデルを実装してもよく、又は機械学習モデルを選択してアプリケーションに含めてもよい。少なくとも一実施例では、アプリケーションは選択可能及びカスタマイズ可能であってもよく、アプリケーションの構造を定義することにより、特定のユーザ向けのアプリケーションの導入及び実装が、よりシームレスなユーザ・エクスペリエンスとして提示される。少なくとも一実施例では、システム3800の他の特徴、たとえばサービス3720及びハードウェア3722などを利用することにより、導入パイプライン3810は、さらによりユーザ・フレンドリになることができ、より容易な統合を実現でき、より正確で、効率的で、タイムリーな結果を生み出すことができる。

【0499】

少なくとも一実施例では、導入システム3706はユーザ・インターフェース3814（たとえば、グラフィカル・ユーザ・インターフェース、ウェブ・インターフェースなど）を含んでもよく、これらは、アプリケーションを選択して導入パイプライン3810に含める、アプリケーションを構成する、アプリケーション又はそのパラメータ若しくは構造を修正又は変更する、セットアップ及び／又は導入中に導入パイプライン3810を使用しそれと対話する、且つ／又は他のやり方で導入システム3706と対話するために使用されてもよい。少なくとも一実施例では、訓練システム3704に関して図示されて

いないが、ユーザ・インターフェース 3814（又は異なるユーザ・インターフェース）は、導入システム 3706 で使用するモデルを選択するため、訓練システム 3704 において訓練若しくは再訓練するモデルを選択するため、且つ／又は他のやり方で訓練システム 3704 と対話するために使用されてもよい。

【0500】

少なくとも一実施例では、アプリケーション・オーケストレーション・システム 3828 に加えてパイプライン・マネージャ 3812 を使用して、導入パイプライン 3810 のアプリケーション又はコンテナと、サービス 3720 及び／又はハードウェア 3722 との間で対話が管理されてもよい。少なくとも一実施例では、パイプライン・マネージャ 3812 は、アプリケーションからアプリケーションへの対話、アプリケーションからサービス 3720 への対話、及び／又はアプリケーション若しくはサービスからハードウェア 3722 への対話を容易にするように構成されてもよい。少なくとも一実施例では、ソフトウェア 3718 に含まれるように図示してあるが、これは限定を意図しておらず、（たとえば、図 39 に示すものなど）いくつかの事例では、パイプライン・マネージャ 3812 は、サービス 3720 に含まれてもよい。少なくとも一実施例では、アプリケーション・オーケストレーション・システム 3828（たとえば、Kubernetes、DOCKER など）は、コンテナ・オーケストレーション・システムを含んでもよく、このシステムは、アプリケーションを、調整、管理、スケリング、及び導入のための論理ユニットとして、コンテナにグループ化することができる。少なくとも一実施例では、導入パイプライン 3810 からのアプリケーション（たとえば、再構築アプリケーション、セグメント化アプリケーションなど）を個々のコンテナに関連付けることより、各アプリケーションは自己完結型環境内（たとえば、カーネル・レベル）で実行して、スピード及び効率を向上させることができる。

【0501】

少なくとも一実施例では、各アプリケーション及び／又はコンテナ（又はその画像）は、個々に開発、修正、及び導入されてもよく（たとえば、第 1 のユーザ又は開発者が、第 1 のアプリケーションを開発、修正、及び導入し、第 2 のユーザ又は開発者が、第 1 のユーザ又は開発者とは別に第 2 のアプリケーションを開発、修正、及び導入してもよく）、これにより、別のアプリケーション又はコンテナのタスクに邪魔されることなく、1 つのアプリケーション及び／又はコンテナのタスクに集中し、注意を払うことが可能になる。少なくとも一実施例では、異なるコンテナ間又はアプリケーション間の通信、及び協調が、パイプライン・マネージャ 3812 及びアプリケーション・オーケストレーション・システム 3828 によって支援されてもよい。少なくとも一実施例では、各コンテナ又はアプリケーションの予測される入力及び／又は出力が、（たとえば、アプリケーション又はコンテナの構造に基づき）システムによって知られている限り、アプリケーション・オーケストレーション・システム 3828 及び／又はパイプライン・マネージャ 3812 は、アプリケーション又はコンテナのそれぞれ間の通信、及びそれらの間でのリソースの共有を容易にすることができる。少なくとも一実施例では、導入パイプライン 3810 のアプリケーション又はコンテナのうちの 1 つ又は複数は、同じサービス及びリソースを共有することができるので、アプリケーション・オーケストレーション・システム 3828 は、様々なアプリケーション間又はコンテナ間でサービス又はリソースをオーケストレートし、ロード・バランシングをおこない、共有を決定してもよい。少なくとも一実施例では、スケジューラを使用して、アプリケーション又はコンテナのリソース要件、これらのリソースの現在の使用量又は計画された使用量、及びリソースの利用可能性が追跡されてもよい。少なくとも一実施例では、こうしてスケジューラは、異なるアプリケーションにリソースを配分し、システムの要件及び利用可能性を考慮してアプリケーション間でリソースを分配してもよい。いくつかの実例では、スケジューラ（及び／又はアプリケーション・オーケストレーション・システム 3828 の他の構成要素）は、サービスの品質（QoS）、（たとえば、リアルタイム処理を実行するか、遅延処理を実行するかを決定するための）データ出力を必要とする緊急度など、システムに課される制約（たとえば、ユーザ制

10

20

30

40

50

約)に基づき、リソースの利用可能性及び分配を決定してもよい。

【0502】

少なくとも一実施例では、導入システム3706のアプリケーション又はコンテナによって利用及び共有されるサービス3720は、計算サービス3816、AIサービス3818、視覚化サービス3820、及び／又は他のタイプのサービスを含んでもよい。少なくとも一実施例では、アプリケーションは、サービス3720のうちの1つ又は複数をコール(たとえば実行)して、アプリケーションのための処理動作を実行してもよい。少なくとも一実施例では、計算サービス3816は、スーパーコンピューティング又は他の高性能コンピューティング(HPC)のタスクを実行するために、アプリケーションによって利用されてもよい。少なくとも一実施例では、アプリケーションのうちの1つ又は複数を介してデータを実質的に同時に処理するため、且つ／又は1つのアプリケーションの1つ又は複数のタスクを実質的に同時に処理するために、計算サービス3816を利用して(たとえば、並列コンピューティング・プラットフォーム3830を使用して)並列処理が実行されてもよい。少なくとも一実施例では、並列コンピューティング・プラットフォーム3830(たとえば、NVIDIAのCUDA)は、GPU(たとえば、GPU3822)上での汎用コンピューティング(GPGPU)を可能にしてもよい。少なくとも一実施例では、並列コンピューティング・プラットフォーム3830のソフトウェア層は、計算カーネルを実行するために仮想命令セット及びGPUの並列計算要素へのアクセスを提供してもよい。少なくとも一実施例では、並列コンピューティング・プラットフォーム3830はメモリを含んでもよく、いくつかの実施例では、メモリは、複数のコンテナ間で、且つ／又は1つのコンテナ内の異なる処理タスク間で共有されてもよい。少なくとも一実施例では、複数のコンテナ、及び／又はコンテナ内の複数のプロセスが、並列コンピューティング・プラットフォーム3830のメモリの共有セグメントからの同じデータを使用するために(たとえば、アプリケーションの複数の異なるステージ、又は複数のアプリケーションが、同じ情報を処理する場合)、プロセス間通信(IPC)コールが生成されてもよい。少なくとも一実施例では、データのコピーを作成し、データをメモリの異なるロケーションに移動(たとえば、読取り／書込みの動作)させるのではなく、メモリの同じロケーションの同じデータが、任意の数の処理タスクに(たとえば、同じ時間、異なる時間などに)使用されてもよい。少なくとも一実施例では、データが使用されて、処理の結果として新規データが生成されるとき、データの新しいロケーションのこの情報は、様々なアプリケーションに記憶され、それらの間で共有されてもよい。少なくとも一実施例では、データのロケーション及び更新済み又は修正済みのデータのロケーションは、コンテナ内でペイロードがどのように理解されるかという定義の一部であってもよい。

【0503】

少なくとも一実施例では、AIサービス3818は、アプリケーションに関連付けられた(たとえば、アプリケーションの1つ又は複数の処理タスクを実行する役割を課された)機械学習モデルを実行するための推論サービスを実行するために利用されてもよい。少なくとも一実施例では、AIサービス3818は、セグメント化、再構築、物体検出、特徴検出、分類、及び／又は他の推論タスクのための機械学習モデル(たとえば、CNNなどのニューラル・ネットワーク)を実行するために、AIシステム3824を利用して、導入パイプライン3810のアプリケーションは、訓練システム3704からの出力モデル3716及び／又はアプリケーションの他のモデルのうちの1つ又は複数を使用して、撮像データ(たとえば、DICOMデータ、RISデータ、CISデータ、REST準拠データ、RPCデータ、生データなど)について推論を実行してもよい。少なくとも一実施例では、アプリケーション・オーケストレーション・システム3828(たとえば、スケジューラ)を使用する推論の2つ以上の実例が利用可能であってもよい。少なくとも一実施例では、第1のカテゴリは、緊急時の緊急要求について推論を実行するため、又は診断時の放射線医のためなど、より高いサービス・レベル合意を達成できる高優先順位／低レイテンシの経路を含むことができる。少なくとも一実施例では、第2のカテゴリは、緊急ではない要求のため、又は分析が後で実行されてもよい

10

20

30

40

50

場合に使用することができる標準優先順位の経路を含んでもよい。少なくとも一実施例では、アプリケーション・オーケストレーション・システム 3828 は、A I サービス 3818 の異なる推論タスク向けの優先順位経路に基づき、リソース（たとえば、サービス 3720 及び／又はハードウェア 3722）を分配してもよい。

【0504】

少なくとも一実施例では、共有ストレージが、システム 3800 内で A I サービス 3818 に取り付けられてもよい。少なくとも一実施例では、共有ストレージは、キャッシュ（又は他のタイプのストレージ・デバイス）として動作してもよく、アプリケーションからの推論要求を処理するために使用されてもよい。少なくとも一実施例では、推論要求が送出されたとき、要求は、導入システム 3706 の A P I インスタンスのセットによって受信されてもよく、1 つ又は複数のインスタンスが（たとえば、最良な適合のため、ロード・バランシングのためなどに）選択されて、要求が処理されてもよい。少なくとも一実施例では、要求を処理するために、要求がデータベースに入れられてもよく、機械学習モデルは、まだキャッシュにない場合には、モデル・レジストリ 3724 から特定されてもよく、検証ステップは、適切な機械学習モデルがキャッシュ（たとえば、共有ストレージ）に確実にロードされるようにしてもよく、且つ／又はモデルのコピーがキャッシュに保存されてもよい。少なくとも一実施例では、アプリケーションがまだ実行されていない場合、又はアプリケーションの十分なインスタンスが存在しない場合には、スケジューラ（たとえば、パイプライン・マネージャ 3812）を使用して、要求において参照されたアプリケーションが起動されてもよい。少なくとも一実施例では、モデルを実行するための推論サーバがまだ起動されていない場合には、推論サーバが起動されてもよい。少なくとも一実施例では、任意の数の推論サーバがモデルごとに起動されてもよい。少なくとも一実施例では、推論サーバがクラスタ化済みであるプル・モデルでは、ロード・バランシングが有利な場合にはいつでもモデルがキャッシュされてもよい。少なくとも一実施例では、推論サーバは、対応する分散サーバに静的にロードされてもよい。

【0505】

少なくとも一実施例では、推論は、コンテナ内で実行される推論サーバを使用して実行されてもよい。少なくとも一実施例では、推論サーバのインスタンスは、モデルに（任意選択でモデルの複数のバージョンに）関連付けられてもよい。少なくとも一実施例では、モデルに対して推論を実行する要求が受信されたとき、推論サーバのインスタンスが存在しない場合には、新規のインスタンスがロードされてもよい。少なくとも一実施例では、推論サーバをスタートするとき、モデルが推論サーバに渡されてもよく、それにより、推論サーバが異なるインスタンスとして実行されている限り、同じコンテナを使用して異なるモデルにサービス提供されてもよい。

【0506】

少なくとも一実施例では、アプリケーションの実行中、所与のアプリケーションについて推論要求が受信されてもよく、（たとえば、推論サーバのインスタンスをホストする）コンテナが（まだロードされていなければ）ロードされてもよく、開始プロシージャがコールされてもよい。少なくとも一実施例では、コンテナの事前処理論理が、（たとえば C P U 及び／又は G P U を使用して）入力データに対する任意の追加的な事前処理をロード、デコード、及び／又は実行してもよい。少なくとも一実施例では、推論のためにデータが準備されると、コンテナは、必要に応じてデータに推論を実行してもよい。少なくとも一実施例では、これは 1 つの画像（たとえば手の X 線）に対する単一の推論コールを含んでもよく、又は何百もの画像（たとえば胸の C T）について推論を要求してもよい。少なくとも一実施例では、アプリケーションは、完了前に結果を要約してもよく、これは限定することなく、単一の信頼性スコア、ピクセル・レベルのセグメント化、ボクセル・レベルのセグメント化、視覚化の生成、又は所見を要約するためのテキストの生成を含んでもよい。少なくとも一実施例では、異なるモデル又はアプリケーションには、異なる優先順位が割り当てられてもよい。たとえば、リアルタイム（T A T 1 分未満）の優先順位を有するモデルもあれば、低優先順位（たとえば、T A T 10 分未満）を有するモデルもある

10

20

30

40

50

。少なくとも一実施例では、モデル実行時間は、要求元の施設又はエンティティから測定されてもよく、推論サービスに対する実行に加えてパートナー・ネットワーク横断時間を含んでもよい。

【0507】

少なくとも一実施例では、サービス3720と推論アプリケーションの間での要求の移行は、ソフトウェア開発キット（SDK）の後ろに隠されてもよく、キューを通して頑健な移送が提供されてもよい。少なくとも一実施例では、個々のアプリケーション／テナントIDの組合せを求めて、要求がAPIを介してキューに入れられ、SDKは、キューから要求を引き出し、要求をアプリケーションに与える。少なくとも一実施例では、SDKが要求をピックアップする環境において、キューの名称が提供されてもよい。少なくとも一実施例では、キューを介した非同期の通信は、その通信が利用可能になったときに、その通信によって、アプリケーションの任意のインスタンスがワークをピックアップできるようになるので、有用な場合がある。少なくとも一実施例では、結果はキューを介して返送されて、データが失われないようにしてもよい。少なくとも一実施例では、最高優先順位のワークは、アプリケーションのほとんどのインスタンスがキューに接続された状態のキューに進むことができ、一方で最低優先順位のワークは、1つのインスタンスがキューに接続された状態の、受信した順番にタスクを処理するキューに進むことができるので、キューは、ワークをセグメント化する機能も提供することができる。少なくとも一実施例では、アプリケーションは、クラウド3826に生成されたGPU加速インスタンス上で実行されてもよく、推論サービスは、GPU上で推論を実行してもよい。

【0508】

少なくとも一実施例では、視覚化サービス3820を利用して、アプリケーション及び／又は導入パイプライン3810の出力を見るための視覚化が生成されてもよい。少なくとも一実施例では、視覚化を生成するために、視覚化サービス3820によってGPU3822が利用されてもよい。少なくとも一実施例では、レイ・トレーシングなどのレンダリング効果が、視覚化サービス3820によって実装されて、より高品質の視覚化が生成されてもよい。少なくとも一実施例では、視覚化は、2D画像のレンダリング、3Dボリウムのレンダリング、3Dボリウムの再構築、2Dトモグラフィ・スライス、仮想現実表示、拡張現実表示などを、限定することなく含んでもよい。少なくとも一実施例では、仮想化された環境を使用して、システムのユーザが対話するための仮想のインタラクティブ表示又はインタラクティブ環境（たとえば、仮想環境）が生成されてもよい。少なくとも一実施例では、視覚化サービス3820は、内部ビジュアライザ、シネマティクス、及び／又は他のレンダリング若しくは画像処理の能力若しくは機能（たとえば、レイ・トレーシング、ラスタ化、内部光学など）を含んでもよい。

【0509】

少なくとも一実施例では、ハードウェア3722は、GPU3822、AIシステム3824、クラウド3826、並びに／又は訓練システム3704及び／若しくは導入システム3706を実行するために使用される任意の他のハードウェアを含んでもよい。少なくとも一実施例では、GPU3822（たとえば、NVIDIAのTESLA及び／又はQUADROのGPU）は、任意の数のGPUを含んでもよく、これらは、計算サービス3816、AIサービス3818、視覚化サービス3820、他のサービス、及び／又はソフトウェア3718の任意の特徴若しくは機能の処理タスクを実行するために使用されてもよい。たとえば、AIサービス3818に関して、GPU3822を使用して、撮像データ（又は機械学習モデルによって使用される他のタイプのデータ）に対して事前処理が実行されてもよく、機械学習モデルの出力に対して事後処理が実行されてもよく、且つ／又は推論が実行されてもよい（たとえば、機械学習モデルが実行されてもよい）。少なくとも一実施例では、クラウド3826、AIシステム3824、及び／又はシステム3800の他の構成要素は、GPU3822を使用してもよい。少なくとも一実施例では、クラウド3826は、深層学習タスクのためにGPU最適化されたプラットフォームを含んでもよい。少なくとも一実施例では、AIシステム3824は、GPUを使用してもよ

く、クラウド3826、又は深層学習若しくは推論の役割を課された少なくとも一部分は、1つ又は複数のAIシステム3824を使用して実行されてもよい。したがって、ハードウェア3722は、個別構成要素として示されているが、これは限定を意図したものではなく、ハードウェア3722の任意の構成要素が、ハードウェア3722の任意の他の構成要素と組み合わせられてもよく、それらによって利用されてもよい。

【0510】

少なくとも一実施例では、AIシステム3824は、推論、深層学習、機械学習、及び／又は他の人工知能タスク向けに構成された専用のコンピューティング・システム（たとえば、スーパーコンピュータ又はHPC）を含んでもよい。少なくとも一実施例では、AIシステム3824（たとえば、NVIDIAのDGX）は、GPU最適化されたソフトウェア（たとえば、ソフトウェア・スタック）を含んでもよく、これは、CPU、RAM、ストレージ、及び／又は他の構成要素、特徴、若しくは機能に加えて、複数のGPU3822を使用して実行されてもよい。少なくとも一実施例では、1つ又は複数のAIシステム3824は、システム3800の一部又はすべてのAIベースの処理タスクを実行するために、（たとえば、データセンタにおいて）クラウド3826に実装されてもよい。

【0511】

少なくとも一実施例では、クラウド3826は、GPU加速化インフラストラクチャ（たとえば、NVIDIAのNGC）を含んでもよく、これは、システム3800の処理タスクを実行するためのGPU最適化されたプラットフォームを提供してもよい。少なくとも一実施例では、クラウド3826は、システム3800のAIベースのタスクのうちの1つ又は複数を実行するためのAIシステム3824を（たとえば、ハードウェア抽象化及びスケーリングのプラットフォームとして）含んでもよい。少なくとも一実施例では、クラウド3826は、複数のGPUを利用してアプリケーション・オーケストレーション・システム3828と統合されて、アプリケーションとサービス3720の間でシームレスなスケーリング及びロード・バランシングを可能にしてもよい。少なくとも一実施例では、クラウド3826は、本明細書に記載する計算サービス3816、AIサービス3818、及び／又は視覚化サービス3820を含むシステム3800のサービス3720の少なくとも一部を実行する役割を課されてもよい。少なくとも一実施例では、クラウド3826は、大小のバッチ推論（たとえば、NVIDIAのテンソルRTの実行）を実行してもよく、加速化された並列コンピューティングのAPI及びプラットフォーム3830（たとえば、NVIDIAのCUDA）を提供してもよく、アプリケーション・オーケストレーション・システム3828（たとえば、KUBERNETES）を実行してもよく、グラフィックス・レンダリングのAPI及びプラットフォーム（たとえば、高品質のシネマティクスを生成するためのレイ・トレーシング、2Dグラフィックス、3Dグラフィックス、及び／又は他のレンダリング技法）を提供してもよく、且つ／又はシステム3800のための他の機能を提供してもよい。

【0512】

少なくとも一実施例では、患者の機密性を保護するために（たとえば、患者のデータ又は記録が構外で使用されることになる場合）、クラウド3826は、深層学習コンテナ・レジストリなどのレジストリを含んでもよい。少なくとも一実施例では、レジストリは、患者データに対する事前処理、事後処理、又は他の処理タスクを実行できるアプリケーションのインスタンス化のためのコンテナを記憶してもよい。少なくとも一実施例では、クラウド3826は、患者データ並びにセンサ・データをコンテナに含むデータを受信してもよく、これらのコンテナにおいてセンサ・データについてのみ要求された処理を実行してもよく、次いで、いずれも患者データを抽出、記憶、又は他のやり方でそれにアクセスする必要なしに、結果の出力及び／又は視覚化を適切なパーティ及び／又はデバイス（たとえば、視覚化又は診断に使用される構内の医療デバイス）に転送してもよい。少なくとも一実施例では、患者データの機密性は、HIPAA及び／又は他のデータ規定に準拠して保護される。

【0513】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0514】

図39は、少なくとも一実施例による、撮像データを処理するための導入パイプライン3810Aの実例の図を含む。少なくとも一実施例では、システム3800、具体的には導入システム3706は、導入パイプライン3810Aをカスタマイズ、更新、及び／又は統合して1つ若しくは複数の生成環境にするために使用されてもよい。少なくとも一実施例では、図39の導入パイプライン3810Aは、施設（たとえば、病院、クリニック、研究所、リサーチ環境など）において特定のユーザ（又はユーザのチーム）によってカスタム定義できる導入パイプライン3810Aの非限定的な実例を含む。少なくとも一実施例では、CTスキャナ3902用に導入パイプライン3810Aを定義するために、ユーザは、CTスキャナ3902によって生成される撮像データに対して特定の機能又はタスクを実行する1つ又は複数のアプリケーションを、たとえばコンテナ・レジストリから選択してもよい。少なくとも一実施例では、アプリケーションは、システム3800のサービス3720及び／又はハードウェア3722を利用できるコンテナとして、導入パイプライン3810Aに適用されてもよい。さらに、導入パイプライン3810Aは、アプリケーションによって使用されるデータを準備するように実装することができる追加の処理タスク又はアプリケーションを含んでもよい（たとえば、DICOMアダプタ3802B及びDICOMリーダ3906が、導入パイプライン3810Aにおいて使用されて、CT再構築3908、器官セグメント化3910などによって使用されるデータを準備してもよい）。少なくとも一実施例では、導入パイプライン3810Aは、一貫性のある導入、1回限りの使用、又は別の頻度若しくは間隔に合わせてカスタマイズ又は選択されてもよい。少なくとも一実施例では、ユーザは、特定の間隔で何人かの対象者について、CT再構築3908及び器官セグメント化3910をおこないたいと思うことがあり、したがって、その期間にわたってパイプライン3810Aを導入してもよい。少なくとも一実施例では、ユーザはシステム3800からの要求ごとに、その要求のためにそのデータに対してユーザが実行したい処理のアプリケーションを選択してもよい。少なくとも一実施例では、導入パイプライン3810Aは、任意の間隔で調節されてもよく、システム3800内のコンテナ構造は適応性及びスケーラビリティがあるので、これはシームレスなプロセスとすることができる。

【0515】

少なくとも一実施例では、図39の導入パイプライン3810Aは、患者又は対象者の撮像データを生成するCTスキャナ3902を含んでもよい。少なくとも一実施例では、CTスキャナ3902からの撮像データは、CTスキャナ3902を収容する施設に関連付けられたPACSサーバ3904に記憶されてもよい。少なくとも一実施例では、PACSサーバ3904は、ソフトウェア及び／又はハードウェアの構成要素を含んでもよく、これらは施設において撮像モダリティ（たとえば、CTスキャナ3902）と直接インターフェースをとってもよい。少なくとも一実施例では、DICOMアダプタ3802Bは、DICOMプロトコルを使用してDICOM物体の送信及び受信を可能にしてもよい。少なくとも一実施例では、DICOMアダプタ3802Bは、導入パイプライン3810Aによって使用するために、PACSサーバからのDICOMデータの準備又は構成を支援してもよい。少なくとも一実施例では、DICOMデータがDICOMアダプタ3802Bを介して処理されると、パイプライン・マネージャ3812は、導入パイプライン3810Aを通るようにデータをルーティングしてもよい。少なくとも一実施例では、D

10

20

30

40

50

ＩＣＯＭリーダ３９０６は、画像ファイル及び関連する任意のメタデータをＤＩＣＯＭデータ（たとえば、視覚化３９１６Ａに示す生のシノグラム・データ）から抽出してもよい。少なくとも一実施例では、抽出された作業ファイルは、導入パイプライン３８１０Ａの他のアプリケーションによってより高速に処理できるようにキャッシュに記憶されてもよい。少なくとも一実施例では、ＤＩＣＯＭリーダ３９０６がデータの抽出及び／又は記憶を終了したら、完了信号がパイプライン・マネージャ３８１２に通信されてもよい。少なくとも一実施例では、次いでパイプライン・マネージャ３８１２は、導入パイプライン３８１０Ａ内の１つ又は複数の他のアプリケーション若しくはコンテナを開始してもよく、又はそれをコールしてもよい。

【０５１６】

少なくとも一実施例では、ＣＴ再構築３９０８のアプリケーションによる処理のためにデータ（たとえば、生のシノグラム・データ）が利用可能になると、ＣＴ再構築３９０８のアプリケーション及び／又はコンテナが実行されてもよい。少なくとも一実施例では、ＣＴ再構築３９０８は、生のシノグラム・データをキャッシュから読み取り、生のシノグラム・データから（たとえば、視覚化３９１６Ｂに示す）画像ファイルを再構築し、結果として生じる画像ファイルをキャッシュに記憶してもよい。少なくとも一実施例では、再構築の完了時、パイプライン・マネージャ３８１２は、再構築タスクが完了したことを通知されてもよい。少なくとも一実施例では、再構築が完了し、再構築された画像ファイルがキャッシュ（又は他のストレージ・デバイス）に記憶されると、器官セグメント化３９１０のアプリケーション及び／又はコンテナが、パイプライン・マネージャ３８１２によってトリガされてもよい。少なくとも一実施例では、器官セグメント化３９１０のアプリケーション及び／又はコンテナは、キャッシュから画像ファイルを読み取り、画像ファイルを推論に適したフォーマットに正規化又は変換し（たとえば、画像ファイルを機械学習モデルの入力解像度に変換し）、正規化された画像に対して推論を実行してもよい。少なくとも一実施例では、正規化された画像に対して推論を実行するために、器官セグメント化３９１０のアプリケーション及び／又はコンテナは、サービス３７２０に依存してもよく、パイプライン・マネージャ３８１２及び／又はアプリケーション・オーケストレーション・システム３８２８は、器官セグメント化３９１０のアプリケーション及び／又はコンテナによるサービス３７２０の使用を容易にしてもよい。少なくとも一実施例では、たとえば器官セグメント化３９１０のアプリケーション及び／又はコンテナは、ＡＩサービス３８１８を利用して正規化画像に対して推論を実行してもよく、ＡＩサービス３８１８はハードウェア３７２２（たとえば、ＡＩシステム３８２４）を利用してＡＩサービス３８１８を実行してもよい。少なくとも一実施例では、推論の結果は（たとえば、視覚化３９１６Ｃに示す）マスク・ファイルであってもよく、このファイルはキャッシュ（又は他のストレージ・デバイス）に記憶されてもよい。

【０５１７】

少なくとも一実施例では、ＤＩＣＯＭデータ及び／又はＤＩＣＯＭデータから抽出されたデータを処理するアプリケーションが処理を完了したら、パイプライン・マネージャ３８１２向けに信号が生成されてもよい。少なくとも一実施例では、次いでパイプライン・マネージャ３８１２は、ＤＩＣＯＭライタ３９１２を実行してキャッシュ（又は他のストレージ・デバイス）から結果を読み取り、要求を生成した施設のユーザによる使用のため、結果をＤＩＣＯＭフォーマットに（たとえば、ＤＩＣＯＭ出力３９１４として）パッケージ化してもよい。少なくとも一実施例では、次いでＤＩＣＯＭ出力３９１４は、ＤＩＣＯＭアダプタ３８０２Ｂに送信されて、（たとえば、施設のＤＩＣＯＭビューワによる視聴のために）ＰＡＣＳサーバ３９０４に記憶するようにＤＩＣＯＭ出力３９１４が準備されてもよい。少なくとも一実施例では、再構築及びセグメント化の要求に応答して、視覚化３９１６Ｂ及び３９１６Ｃが生成され、診断、リサーチ、及び／又は他の目的のためにユーザにとって利用可能にされてもよい。

【０５１８】

導入パイプライン３８１０Ａでは連続したアプリケーションとして示されているが、Ｃ

10

20

30

40

50

T再構築3908及び器官セグメント化3910のアプリケーションは、少なくとも一実施例において並列で処理されてもよい。少なくとも一実施例では、アプリケーションが互いに依存性を有しておらず、（たとえば、DICOMリーダ3906がデータを抽出した後に）アプリケーションごとにデータが利用可能である場合、アプリケーションは同時に、実質的に同時に、又は一部が重なって実行されてもよい。少なくとも一実施例では、2つ以上のアプリケーションが同様のサービス3720を要求する場合、システム3800のスケジューラを使用して、様々なアプリケーション間で計算又は処理のリソースのロード・バランシングが行われ、それらを分散させてもよい。少なくとも一実施例では、いくつかの実施例において、導入パイプライン3810Aのランタイムを短縮してリアルタイムの結果を提供するために、並列コンピューティング・プラットフォーム3830を使用してアプリケーションのための並列処理が実行されてもよい。

10

【0519】

少なくとも一実施例では、図40A～図40Bを参照すると、導入システム3706は、画像処理、セグメント化、強調、AI、視覚化、及び推論などの異なる機能を、撮像デバイス（たとえば、CTスキャナ、X線機械、MRI機械など）、シーケンシング・デバイス、ゲノミクス・デバイス、及び／又は他のタイプのデバイスを用いて実行するための1つ又は複数の仮想機器として実装されてもよい。少なくとも一実施例では、システム3800は、ソフトウェア定義された導入パイプライン3810を含むことができる仮想機器の生成及び提供を可能にしてもよく、この導入パイプライン3810は、デバイスによって生成された生／未処理の入力データを受信し、処理済み／再構築済みのデータを出力してもよい。少なくとも一実施例では、仮想機器を表す導入パイプライン3810（たとえば、3810A及び3810B）は、機械学習モデルを利用することなどにより、知能をパイプラインに実装して、コンテナ化された推論サポートをシステムに提供してもよい。少なくとも一実施例では、仮想機器は、アプリケーションのインスタンスをそれぞれが含む任意の数のコンテナを実行してもよい。少なくとも一実施例では、リアルタイムの処理が望ましい場合などでは、仮想機器を表す導入パイプライン3810は静的であってもよく（たとえば、コンテナ及び／又はアプリケーションが設定されていてもよく）、一方他の実施例では、仮想機器用のコンテナ及び／又はアプリケーションが、アプリケーション又はリソースのプール（たとえばコンテナ・レジストリ内）から（たとえば要求ごとに）選択されてもよい。

20

30

【0520】

少なくとも一実施例では、システム3800は、たとえば施設の放射線機械、撮像デバイス、及び／又は別のタイプのデバイスに隣接して導入された、又は他のやり方でそれらと通信するコンピューティング・システムにおいて、施設構内の1つ又は複数の仮想機器としてインスタンス化されてもよく、又は実行されてもよい。しかし少なくとも一実施例では、構内でのインストールは、デバイス自体（たとえば、撮像デバイスに統合されたコンピューティング・システム）内で、ローカルなデータセンタ（たとえば、構内のデータセンタ）で、及び／又はクラウド環境（たとえば、クラウド3826）でインスタンス化又は実行されてもよい。少なくとも一実施例では、仮想機器として動作する導入システム3706は、いくつかの実例においてスーパーコンピュータ又は他のHPCシステムによってインスタンス化されてもよい。少なくとも一実施例では、構内でのインストールにより、リアルタイムの処理のために（たとえば、イーサネット（登録商標）を介したRFなど、高スループットのローカル通信インターフェースを介した）広帯域の使用が可能になる。少なくとも一実施例では、リアルタイム又はほぼリアルタイムの処理は、正確な診断及び分析のために即時の視覚化が期待され必要とされる超音波デバイス又は他の撮像モダリティを仮想機器がサポートする場合に、特に有用なことがある。少なくとも一実施例では、ローカルな要求が構内の容量又は能力を超過するとき、クラウド・コンピューティング・アーキテクチャは、クラウド・コンピューティングのサービス・プロバイダ又は他の計算クラスタに対する動的バーストを実行することができる。少なくとも一実施例では、クラウド・アーキテクチャは、実装されると、訓練システム3704に関して本明細書に

40

50

記載のニューラル・ネットワーク又は他の機械学習モデルを訓練するように調整されてもよい。少なくとも一実施例では、訓練パイプラインが定位置にあるとき、機械学習モデルは、それがサポートするデバイスからの追加情報を処理するとき、継続的に学習及び改善されてもよい。少なくとも一実施例では、仮想機器は、追加データ、新規データ、既存の機械学習モデル、及び／又は新規若しくは更新済みの機械学習モデルを使用して、継続的に改善されてもよい。

【0521】

少なくとも一実施例では、コンピューティング・システムは、本明細書に記載のハードウェア3722の一部又はすべてを含んでもよく、ハードウェア3722は、デバイス内、デバイスに結合され近位に位置するコンピューティング・デバイスの一部として、施設のローカル・データセンタ内、及び／又はクラウド3826内を含む複数のやり方のうちの任意のやり方で分散されてもよい。少なくとも一実施例では、導入システム3706及び関連付けられたアプリケーション又はコンテナは、ソフトウェアに（たとえば、アプリケーションの個別のコンテナ化インスタンスとして）生成されるので、仮想機器の挙動、動作、及び構成、並びに仮想機器によって生成される出力は、仮想機器がサポートするデバイスの生の出力を変える又は変更する必要なしに、望み通りに修正又はカスタマイズされることが可能である。

【0522】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0523】

図40Aは、少なくとも一実施例による、超音波デバイスをサポートする仮想器具の実例データ・フロー図を含む。少なくとも一実施例では、導入パイプライン3810Bは、システム3800のサービス3720のうちの1つ又は複数を利用してもよい。少なくとも一実施例では、導入パイプライン3810B及びサービス3720は、ローカルクラウド3826のいずれかにおいて、システムのハードウェア3722を利用してもよい。少なくとも一実施例では、図示していないが、プロセス4000は、パイプライン・マネージャ3812、アプリケーション・オーケストレーション・システム3828、及び／又は並列コンピューティング・プラットフォーム3830によって促進されてもよい。

【0524】

少なくとも一実施例では、プロセス4000は、超音波デバイス4002から撮像データを受信することを含んでもよい。少なくとも一実施例では、撮像データは、DICOMフォーマット（又はRIS、CIS、REST準拠、RPC、生などの他のフォーマット）でPACSサーバに記憶されてもよく、超音波デバイス4002用の仮想機器（たとえば仮想超音波）として選択又はカスタマイズされた導入パイプライン3810を通して処理するために、システム3800によって受信されてもよい。少なくとも一実施例では、撮像データは、撮像デバイス（たとえば、超音波デバイス4002）から直接受信されてもよく、仮想機器によって処理されてもよい。少なくとも一実施例では、撮像デバイスと仮想機器の間に通信可能に結合されたトランスデューサ又は他の信号変換器は、撮像デバイスによって生成された信号データを、仮想機器によって処理することができる画像データに変換してもよい。少なくとも一実施例では、生データ及び／又は画像データは、導入パイプライン3810Bのアプリケーション又はコンテナによって使用されるデータを抽出するために、DICOMリーダ3906に適用されてもよい。少なくとも一実施例では、DICOMリーダ3906は、アプリケーション又はコンテナによって使用されるデー

10

20

30

40

50

タを抽出、リサイズ、リスケーリング、及び／又は他のやり方で準備するためのサービス 3720 として（たとえば、計算サービス 3816 のうちの 1 つとして）、データ拡張ライブラリ 4014（たとえば、NVIDIA の DALI（登録商標））を利用してもよい。
【0525】

少なくとも一実施例では、データが準備されると、再構築 4006 のアプリケーション及び／又はコンテナが実行されて、超音波デバイス 4002 からのデータが画像ファイルに再構築されてもよい。少なくとも一実施例では、再構築 4006 の後、又は再構築 4006 と同時に、検出 4008 のアプリケーション及び／又はコンテナが、異常検出、物体検出、特徴検出、及び／又はデータに関する他の検出タスクのために実行されてもよい。少なくとも一実施例では、再構築 4006 中に生成された画像ファイルは、検出 4008 中に使用されて、異常、物体、特徴などが識別されてもよい。少なくとも一実施例では、検出 4008 のアプリケーションは、推論エンジン 4016 を（たとえば、AI サービス 3818 のうちの 1 つとして）利用して、データについて推論を実行して検出を生成してもよい。少なくとも一実施例では、（たとえば、訓練システム 3704 からの）1 つ又は複数の機械学習モデルは、検出 4008 のアプリケーションによって実行又はコールされてもよい。

10

【0526】

少なくとも一実施例では、再構築 4006 及び／又は検出 4008 が完了すると、これらのアプリケーション及び／又はコンテナからのデータ出力を使用して、ワークステーション又はディスプレイ端末に表示される視覚化 4012（たとえば、グレースケール出力）などの視覚化 4010 が生成されてもよい。少なくとも一実施例では、視覚化により、超音波デバイス 4002 に対する導入パイプライン 3810B の結果を、技術者又は他のユーザが視覚化できるようになる。少なくとも一実施例では、視覚化 4010 は、システム 3800 のレンダリング構成要素 4018（たとえば、視覚化サービス 3820 のうちの 1 つ）を利用することによって、実行されてもよい。少なくとも一実施例では、レンダリング構成要素 4018 は、2D、OpenGL、又はレイ・トレーシングのサービスを実行して、視覚化 4012 を生成してもよい。

20

【0527】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

30

【0528】

図 40B は、少なくとも一実施例による、CT スキャナをサポートする仮想器具の実例データ・フロー図を含む。少なくとも一実施例では、導入パイプライン 3810C は、システム 3800 のサービス 3720 のうちの 1 つ又は複数を利用してもよい。少なくとも一実施例では、導入パイプライン 3810C 及びサービス 3720 は、ローカルかクラウド 3826 のいずれかにおいて、システムのハードウェア 3722 を利用してもよい。少なくとも一実施例では、図示していないが、プロセス 4020 は、パイプライン・マネージャ 3812、アプリケーション・オーケストレーション・システム 3828、及び／又は並列コンピューティング・プラットフォーム 3830 によって促進されてもよい。

40

【0529】

少なくとも一実施例では、プロセス 4020 は、生データを生成する CT スキャナ 4022 を含んでもよく、この生データは、DICOM リーダ 3906 によって（たとえば、直接、PACS サーバ 3904 を介して、処理後になど）受信されてもよい。少なくとも一実施例では、（導入パイプライン 3810C によってインスタンス化された）仮想 CT

50

は、患者を監視するため（たとえば、患者動き検出 A I 4 0 2 6）且つ／又は C T スキャナ 4 0 2 2 の露出を（たとえば、露出制御 A I 4 0 2 4 を使用して）調節又は最適化するための、第 1 のリアルタイム・パイプラインを含んでもよい。少なくとも一実施例では、アプリケーションのうちの 1 つ又は複数（たとえば、4 0 2 4 及び 4 0 2 6）は、A I サービス 3 8 1 8 などのサービス 3 7 2 0 を利用してもよい。少なくとも一実施例では、露出制御 A I 4 0 2 4 のアプリケーション（又はコンテナ）及び／又は患者動き検出 A I 4 0 2 6 のアプリケーション（コンテナ）の出力は、C T スキャナ 4 0 2 2 及び／又は技術者に対するフィードバックとして使用されて、露出（又は C T スキャナ 4 0 2 2 の他の設定）が調節されてもよく、且つ／又は患者にあまり動かないように伝えられてもよい。

【0530】

少なくとも一実施例では、導入パイプライン 3 8 1 0 C は、C T スキャナ 4 0 2 2 によって生成されるデータを分析するための非リアルタイム・パイプラインを含んでもよい。少なくとも一実施例では、第 2 のパイプラインは、C T 再構築 3 9 0 8 のアプリケーション及び／又はコンテナ、粗検出 A I 4 0 2 8 のアプリケーション及び／又はコンテナ、（たとえば、粗検出 A I 4 0 2 8 によってある特定の結果が検出された場合の）精検出 A I 4 0 3 2 のアプリケーション及び／又はコンテナ、視覚化 4 0 3 0 のアプリケーション及び／又はコンテナ、及び D I C O M ライタ 3 9 1 2（及び／又は R I S、C I S、R E S T 準拠、R P C、生など他のタイプのデータ・ライタ）のアプリケーション及び／又はコンテナを含んでもよい。少なくとも一実施例では、C T スキャナ 4 0 2 2 によって生成された生データは、（仮想 C T 機器としてインスタンス化された）導入パイプライン 3 8 1 0 C のパイプラインに通されて、結果が生成されてもよい。少なくとも一実施例では、D I C O M ライタ 3 9 1 2 からの結果は、表示のために送信されてもよく、且つ／又は技術者、開業医、若しくは他のユーザによって後で検索、分析、又は表示できるように P A C S サーバ 3 9 0 4 に記憶されてもよい。

【0531】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

【0532】

図 4 1 A は、少なくとも一実施例による、機械学習モデルを訓練、再訓練、又は更新するためのプロセス 4 1 0 0 のデータ・フロー図を示す。少なくとも一実施例では、プロセス 4 1 0 0 は、図 3 8 のシステム 3 8 0 0 を非限定的な実例として使用して、実行されてもよい。少なくとも一実施例では、プロセス 4 1 0 0 は、本明細書に記載のシステム 3 8 0 0 のサービス 3 7 2 0 及び／又はハードウェア 3 7 2 2 を利用してもよい。少なくとも一実施例では、プロセス 4 1 0 0 によって生成される精緻化モデル 4 1 1 2 は、導入パイプライン 3 8 1 0 内の 1 つ又は複数のコンテナ化アプリケーションのために、導入システム 3 7 0 6 によって実行されてもよい。

【0533】

少なくとも一実施例では、モデル訓練 3 7 1 4 は、新規訓練データ（たとえば、顧客データセット 4 1 0 6、及び／又は入力データに関連付けられた新規グラウンド・トゥールース・データなどの新規入力データ）を使用して、初期モデル 4 1 0 4（たとえば、事前訓練済みモデル）を再訓練又は更新することを含んでもよい。少なくとも一実施例では、初期モデル 4 1 0 4 を再訓練又は更新するために、初期モデル 4 1 0 4 の出力又は損失層がリセットされてもよく、削除されてもよく、且つ／又は更新済み若しくは新規の出力若しくは損失層と置換されてもよい。少なくとも一実施例では、初期モデル 4 1 0 4 は、以前

に微調整された、前の訓練から残っているパラメータ（たとえば、重み及び／又はバイアス）を有してもよく、それにより、訓練又は再訓練 3714 は、最初からモデルを訓練するほど長い時間がかからず、又は多くの処理を必要としなくても済む。少なくとも一実施例では、モデル訓練 3714 の間に、初期モデル 4104 のリセット又は置換された出力又は損失層を有することにより、パラメータは、新規の顧客データセット 4106（たとえば、図 37 の画像データ 3708）について予測を生成する際の出力又は損失層の精度に関連付けられた損失計算に基づき、新規データセットのために更新又は再調整されてもよい。

【0534】

少なくとも一実施例では、事前訓練済みモデル 3806 は、データストア又はレジストリ（たとえば、図 37 のモデル・レジストリ 3724）に記憶されてもよい。少なくとも一実施例では、事前訓練済みモデル 3806 は、少なくとも部分的に、プロセス 4100 を実行する施設とは異なる 1 つ又は複数の施設において訓練済みであってもよい。少なくとも一実施例では、異なる施設の患者、対象者、顧客のプライバシー及び権利を保護するために、事前訓練済みモデル 3806 は、構内で生成された顧客又は患者のデータを使用して、構内で訓練されたものであってもよい。少なくとも一実施例では、事前訓練済みモデル 3806 は、クラウド 3826 及び／又は他のハードウェア 3722 を使用して訓練されてもよいが、プライバシー保護された機密の患者データは、クラウド 3826（又は他の構外のハードウェア）の任意の構成要素に転送できず、それらの構成要素によって使用されず、又はアクセス不可能であってもよい。少なくとも一実施例では、事前訓練済みモデル 3806 が 2 つ以上の施設からの患者データを使用して訓練される場合、事前訓練済みモデル 3806 は、各施設について個々に訓練されてから、別の施設からの患者若しくは顧客のデータについて訓練されてもよい。少なくとも一実施例では、顧客又は患者のデータが（たとえば、実験での使用を目的とした権利放棄などによって）プライバシー問題から解放されている場合、又は顧客若しくは患者のデータがパブリック・データセットに含まれる場合などには、任意の数の施設からの顧客又は患者のデータを使用して、データセンタ又は他のクラウド・コンピューティング・インフラストラクチャなど、構内及び／又は構外で事前訓練済みモデル 3806 が訓練されてもよい。

【0535】

少なくとも一実施例では、導入パイプライン 3810 で使用するアプリケーションを選択するとき、ユーザは、特定のアプリケーションで使用するようになる機械学習モデルも選択することができる。少なくとも一実施例では、ユーザは、使用するモデルを有していないことがあり、したがって、ユーザはアプリケーションとともに使用する事前訓練済みモデル 3806 を選択してもよい。少なくとも一実施例では、訓練済みモデル 3806 は、（たとえば、患者の多様性、人口統計、使用される医療用撮像デバイスのタイプなどに基づき）ユーザの施設の顧客データセット 4106 について正確な結果を生成するように最適化されてもよい。少なくとも一実施例では、事前訓練済みモデル 3806 を、アプリケーションとともに使用するために導入パイプライン 3810 に導入する前に、事前訓練済みモデル 3806 は、それぞれの施設において使用するために更新、再訓練、及び／又は微調整されてもよい。

【0536】

少なくとも一実施例では、ユーザは、更新、再訓練、及び／又は微調整されることになる事前訓練済みモデル 3806 を選択してもよく、事前訓練済みモデル 3806 は、プロセス 4100 内でシステム 3704 を訓練するための初期モデル 4104 と呼ばれてもよい。少なくとも一実施例では、顧客データセット 4106（たとえば、施設のデバイスによって生成された撮像データ、ゲノミクス・データ、シーケンシング・データ、又は他のタイプのデータ）を使用して、初期モデル 4104 について（限定することなく転送学習（transfer learning）を含んでもよい）モデル訓練 3714 が実行されて、精緻化モデル 4112 が生成されてもよい。少なくとも一実施例では、顧客データセット 4106 に対応するグラウンド・トゥールズ・データが、訓練システム 3704 に

よって生成されてもよい。少なくとも一実施例では、グラウンド・トゥールス・データは、（たとえば、図 3 7 のラベル付けされたクリニック・データ 3 7 1 2 として）施設において臨床医、科学者、医師、開業医によって、少なくとも部分的に生成されてもよい。

【0 5 3 7】

少なくとも一実施例では、A I 支援アノテーション 3 7 1 0 がいくつかの実例において使用されて、グラウンド・トゥールス・データが生成されてもよい。少なくとも一実施例では、（たとえば、A I 支援アノテーション S D K を使用して実装された）A I 支援アノテーション 3 7 1 0 は、機械学習モデル（たとえば、ニューラル・ネットワーク）を利用して、顧客データセットについて示唆又は予測されるグラウンド・トゥールス・データを生成してもよい。少なくとも一実施例では、ユーザ 4 1 1 0 は、コンピューティング・デバイス 4 1 0 8 上のユーザ・インターフェース（グラフィカル・ユーザ・インターフェース（G U I））内でアノテーション・ツールを使用してもよい。

10

【0 5 3 8】

少なくとも一実施例では、ユーザ 4 1 1 0 は、コンピューティング・デバイス 4 1 0 8 を介して G U I と対話して、アノテーション又は自動アノテーションを編集又は微調整してもよい。少なくとも一実施例では、ポリゴン編集特徴を使用して、ポリゴンの頂点をより正確なロケーション又は微調整されたロケーションに移動させてもよい。

【0 5 3 9】

少なくとも一実施例では、顧客データセット 4 1 0 6 が、関連付けられたグラウンド・トゥールス・データを得ると、（たとえば、A I 支援アノテーション、手動ラベリングなどからの）グラウンド・トゥールス・データが、モデル訓練 3 7 1 4 中に使用されて、精緻化モデル 4 1 1 2 が生成されてもよい。少なくとも一実施例では、顧客データセット 4 1 0 6 は、初期モデル 4 1 0 4 に任意の回数、適用されてもよく、グラウンド・トゥールス・データは、精緻化モデル 4 1 1 2 について許容可能なレベルの精度が達成されるまで、初期モデル 4 1 0 4 のパラメータを更新するために使用されてもよい。少なくとも一実施例では、精緻化モデル 4 1 1 2 が生成されると、精緻化モデル 4 1 1 2 は、医療用撮像データに対して 1 つ又は複数の処理タスクを実行するために、施設において 1 つ又は複数の導入パイプライン 3 8 1 0 内に導入されてもよい。

20

【0 5 4 0】

少なくとも一実施例では、精緻化モデル 4 1 1 2 は、別の施設によって選択されることになるモデル・レジストリ 3 7 2 4 の事前訓練済みモデル 3 8 0 6 にアップロードされてもよい。少なくとも一実施例では、このプロセスは任意の数の施設において完了されてもよく、それにより精緻化モデル 4 1 1 2 は、新規データセットについて任意の回数さらに精緻化されて、より普遍的なモデルが生成されてもよい。

30

【0 5 4 1】

少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1 つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図 1 ～図 7 に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

40

【0 5 4 2】

図 4 1 B は、少なくとも一実施例による、事前訓練済みのアノテーション・モデルを用いてアノテーション・ツールを強化するためのクライアント・サーバのアーキテクチャの実例の図である。少なくとも一実施例では、A I 支援アノテーション・ツール 4 1 3 6 は、クライアント・サーバのアーキテクチャ 4 1 3 2 に基づきインスタンス化されてもよい。少なくとも一実施例では、撮像アプリケーションのアノテーション・ツール 4 1 3 6 は、たとえば放射線医が器官及び異常を識別するのを支援してもよい。 少なくとも一実施

50

例では、撮像アプリケーションは、非限定的な実例として（たとえば、3DのMRI又はCRスキャンの）生画像4134において、特定の対象器官上の数少ない極値点をユーザ4110が識別するのを援助し、特定の器官の2Dスライスすべてについて自動アノテーション付けされた結果を受信するソフトウェア・ツールを含んでもよい。少なくとも一実施例では、結果は、訓練データ4138としてデータストアに記憶されてもよく、（たとえば、限定することなく）訓練用のグラウンド・トゥルース・データとして使用されてもよい。少なくとも一実施例では、コンピューティング・デバイス4108が、AI支援アノテーション3710のために極値点を送るとき、たとえば深層学習モデルがこのデータを入力として受信してもよく、セグメント化された器官又は異常の推論結果を返してもよい。少なくとも一実施例では、図41BのAI支援アノテーション・ツール4136Bなどの事前インスタンス化されたアノテーション・ツールは、たとえばアノテーション・モデル・レジストリに記憶された事前訓練済みモデル4142のセットを含むことができるアノテーション支援サーバ4140などのサーバに、APIコール（たとえば、APIコール4144）をおこなうことによって、拡張されてもよい。少なくとも一実施例では、アノテーション・モデル・レジストリは、特定の器官又は異常に対してAI支援アノテーションを実行するように事前訓練された事前訓練済みモデル4142（たとえば、深層学習モデルなどの機械学習モデル）を記憶してもよい。少なくとも一実施例では、これらのモデルは、訓練パイプライン3804を使用することにより、さらに更新されてもよい。少なくとも一実施例では、事前インストールされたアノテーション・ツールは、ラベル付けされた新規クリニック・データ3712が加えられるにつれて、経時的に改善されてもよい。

10

20

【0543】

1つ又は複数の実施例に関連する推論及び／又は訓練の動作を実行するために、推論及び／又は訓練論理815が使用される。推論及び／又は訓練論理815に関する詳細は、図8A及び／又は図8Bと併せて本明細書において提供される。

【0544】

少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき実演のゴールを識別するように、上の図を参照して適応される。少なくとも一実施例では、1つ又は複数の回路、プロセッサ、システム、ロボット、若しくは他のデバイス若しくは技法は、図1～図7に関連して上に記載した技法に少なくとも部分的に基づき、実演を観察すること、実演のゴールを識別すること、及びロボット操作によってゴールを達成することができるロボット・デバイスを実装するように、上の図を参照して適応される。

30

【0545】

本開示の少なくとも一実施例は、以下の条項を考慮して説明されてもよい。

1.

ビデオ・データを少なくとも第1のセグメント及び第2のセグメントにセグメント化するステップであって、第1のセグメントが、実演において操作される第1の物体の第1の軌道を表すビデオ・データを含み、第2のセグメントが、実演において操作される第2の物体の第2の軌道を表すビデオ・データを含む、セグメント化するステップと、

40

第1の軌道によって満足される動作述語を識別するステップであって、第1の軌道で第1の物体を動かすことにより、第2の軌道で第2の物体を動かすことが有効になったという判定に少なくとも部分的に基づき、動作述語が識別される、識別するステップと、

ドメイン定義において定義された論理条件を満足する第2の軌道に少なくとも部分的に基づき、第2の軌道によって満足されるタスク述語を識別するステップと、

タスク述語に少なくとも部分的に基づき、実演のゴールを判定するステップとを含むコンピュータ実装方法。

2. 実演において操作される物体の別の軌道を第2の軌道が有効にしないという判定に少なくとも部分的に基づきゴールを識別するために、タスク述語を使用するステップをさら

50

に含む、条項 1 に記載のコンピュータ実装方法。

3. ゴールの 1 つ又は複数の述語を満足するように 1 つ又は複数の操作を実行することにより、1 つ又は複数のロボット操作デバイスにゴールを達成させるステップをさらに含む、条項 1 又は 2 に記載のコンピュータ実装方法。

4. 実演が実行された環境以外の環境において、1 つ又は複数のロボット操作デバイスによってゴールが達成される、条項 3 に記載のコンピュータ実装方法。

5. 実演における物体の少なくとも 1 つの操作が、ロボット操作デバイスによる 1 つ又は複数の操作から除外される、条項 3 又は 4 に記載のコンピュータ実装方法。

6. 第 1 の軌道によって満足される付加的タスク述語を除くタスク述語のセットから、ゴールが識別される、条項 1 から 5 までのいずれか一項に記載のコンピュータ実装方法。

7. 第 1 の軌道で第 1 の物体を動かすことにより、第 2 の軌道で第 2 の物体を動かすことが有効になったという判定に少なくとも部分的に基づき、付加的タスク述語が除外される、条項 6 に記載のコンピュータ実装方法。

8. ドメイン定義が、物体のリストと、物体のリスト内の物体の操作によって満足可能なタスク述語のリストとを含む、条項 1 から 7 までのいずれか一項に記載のコンピュータ実装方法。

9. システムであって、

少なくとも 1 つのプロセッサと、

少なくとも 1 つのプロセッサによって実行された結果、システムに少なくとも

ビデオ・データを第 1 のセグメント及び第 2 のセグメントにセグメント化することであって、第 1 のセグメントが、実演において操作される第 1 の物体の第 1 の軌道を表すビデオ・データを含み、第 2 のセグメントが、実演において操作される第 2 の物体の第 2 の軌道を表すビデオ・データを含む、セグメント化すること、

第 1 の軌道に従って第 1 の物体を操作することにより満足される動作述語を識別することであって、第 1 の軌道に従って第 1 の物体を操作することにより、第 2 の軌道に従って第 2 の物体を動かすことが有効になったという判定に少なくとも部分的に基づき、動作述語が識別される、識別すること、

第 2 の軌道に従って第 2 の物体を操作することにより満足されるタスク述語を識別すること、及び

タスク述語に少なくとも部分的に基づき、実演のゴールを識別することを行わせる命令を含むメモリとを含むシステム。

10. 第 2 の軌道が別の軌道を有効にしないという判定に少なくとも部分的に基づき、タスク述語が識別される、条項 9 に記載のシステム。

11. 少なくとも 1 つのプロセッサによって実行された結果、少なくとも、1 つ又は複数のロボット操作デバイスにゴールを達成させることを、システムに行わせるさらなる命令をメモリが含む、条項 9 又は 10 に記載のシステム。

12. 実演が実行された環境以外の環境において、ロボット操作デバイスによってゴールが達成される、条項 11 に記載のシステム。

13. タスク述語の意図性の推定に少なくとも部分的に基づき、実演のゴールが識別される、条項 9 から 12 までのいずれか一項に記載のシステム。

14. 意図性の推定が、第 1 の軌道と第 2 の軌道との関係に少なくとも部分的に基づく、条項 13 に記載のシステム。

15. 第 1 の軌道によって満足される第 1 のタスク述語は、第 1 の軌道によって第 2 の軌道が有効になることに少なくとも部分的に基づき、意図性が低いと評価される、条項 9 から 14 までのいずれか一項に記載のシステム。

16. コンピュータ・システムの 1 つ又は複数のプロセッサによって実行された結果、コンピュータ・システムに

ビデオ・データを第 1 のセグメント及び第 2 のセグメントに分割することであって、第 1 のセグメントが、実演において操作される第 1 の物体の第 1 の軌道を表すビデオ・デー

10

20

30

40

50

タを含み、第2のセグメントが、第2の物体の第2の軌道を表すビデオ・データを含む、分割すること、

第1の軌道によって満足される動作述語を識別することであって、第1の物体の操作により、第2の軌道が有効になったという判定に少なくとも部分的に基づき、動作述語が識別される、識別すること、

第2の軌道によって満足されるタスク述語を識別すること、及び

タスク述語に少なくとも部分的に基づき、実演のゴールを識別すること
を行わせる命令を含む、非一時的コンピュータ可読媒体。

17. 第2の軌道が別の軌道を有効にしないという判定に少なくとも部分的に基づきゴールを識別するために、タスク述語が使用される、条項16に記載の非一時的コンピュータ可読媒体。

18. コンピュータ・システムの1つ又は複数のプロセッサによって実行された結果、物体を操作してゴールのタスク述語を達成するようにロボット操作デバイスに命令することによってゴールを達成することを、コンピュータ・システムに行わせるさらなる命令を含む、条項16又は17に記載の非一時的コンピュータ可読媒体。

19. タスク述語の推定される意図性に少なくとも部分的に基づき、実演のゴールが識別される、条項16から18までのいずれか一項に記載の非一時的コンピュータ可読媒体。

20. 推定される意図性が、第2の軌道に対する第1の軌道の分析に少なくとも部分的に基づく、条項19に記載の非一時的コンピュータ可読媒体。

21. ロボット・デバイスであって、

1つ又は複数のロボット操作デバイスと、
少なくとも1つのプロセッサと、

少なくとも1つのプロセッサによって実行された結果、ロボット・デバイスに少なくともゴールを達成するために1つ又は複数のロボット・アームの1つ又は複数の操作を実行するための命令を取得することを行わせる命令を含むメモリと
を備え、ゴールが、

実演のビデオ・データを第1のセグメント及び第2のセグメントにセグメント化することであって、第1のセグメントが、第1の物体の第1の軌道を表すビデオ・データを含み、第2のセグメントが、第2の物体の第2の軌道を表すビデオ・データを含む、セグメント化すること、

第1の軌道によって満足される動作述語を識別することであって、第1の軌道により第2の軌道が有効になったという判定に少なくとも部分的に基づき、動作述語が識別される、識別すること、

第2の軌道に従って第2の物体を操作することにより満足されるタスク述語を識別すること、及び

タスク述語に少なくとも部分的に基づき、実演のゴールを識別すること
によって判定される、ロボット・デバイス。

22. 第2の軌道がタスク述語の1つ又は複数の条件への適合をもたらし、且つ別の軌道を有効にしないという判定に少なくとも部分的に基づき、タスク述語が識別される、条項21に記載のロボット・デバイス。

23. ゴールの識別された述語の1つ又は複数の条件を満足する操作を識別することにより、1つ又は複数の操作が判定される、条項21又は22に記載のロボット・デバイス。

24. ゴールを達成するための1つ又は複数の操作によって満足される1つ又は複数の述語が、実演において満足される1つ又は複数の述語とは異なる、条項23に記載のロボット・デバイス。

25. ゴールを判定するために使用されるタスク述語が、閾値量を超える関連性の推定に関連付けられたタスク述語を含む、条項21から24までのいずれか一項に記載のロボット・デバイス。

26. 関連性の推定が、2つの軌道の関係に少なくとも部分的に基づく、条項25に記載のロボット・デバイス。

10

20

30

40

50

【0546】

少なくとも一実施例では、単一の半導体プラットフォームとは、単独で単体の半導体ベースの集積回路又はチップを指してもよい。少なくとも一実施例では、マルチ・チップ・モジュールは、オン・チップ動作をシミュレートする接続性が向上した状態で使用されてもよく、従来の中央処理装置（「CPU」）及びバスの実装形態の利用を大幅に改善する。少なくとも一実施例では、ユーザの希望に応じて、半導体プラットフォームとは別々に、又は半導体プラットフォームとの様々な組合せで、様々なモジュールがさらに設置されてもよい。

【0547】

少なくとも一実施例では、図14に戻って参照すると、機械読取り可能で実行可能なコード若しくはコンピュータ制御論理アルゴリズムの形のコンピュータ・プログラムが、メイン・メモリ1404及び／又は二次ストレージに記憶される。コンピュータ・プログラムは、1つ又は複数のプロセッサによって実行された場合に、少なくとも一実施例による様々な機能をシステム1400が実行できるようにする。少なくとも一実施例では、メモリ1404、ストレージ、及び／又は任意の他のストレージが、コンピュータ読取り可能媒体の考えられる例である。少なくとも一実施例では、二次ストレージとは、フロッピー（登録商標）・ディスク・ドライブ、磁気テープ・ドライブ、コンパクト・ディスク・ドライブ、デジタル多用途ディスク（「DVD」：digital versatile disk）ドライブ、記録デバイス、ユニバーサル・シリアル・バス（「USB」）フラッシュ・メモリなどを表すハード・ディスク・ドライブ及び／若しくはリムーバブル・ストレージ・ドライブなどの任意の好適なストレージ・デバイス又はシステムを指してもよい。少なくとも一実施例では、様々な先の図面のアーキテクチャ及び／又は機能は、CPU1402、並列処理システム1412、CPU1402と並列処理システム1412の両方の機能の少なくとも一部分を実現可能な集積回路、チップセット（たとえば、関連機能を実行するためのユニットとして機能し、販売されるように設計された集積回路のグループなど）、及び／又は集積回路の任意の好適な組合せの文脈において実装される。

【0548】

少なくとも一実施例では、様々な先の図面のアーキテクチャ及び／又は機能は、汎用コンピュータ・システム、回路板システム、エンタテインメント目的専用のゲーム・コンソール・システム、及び特定用途システムなどの文脈において実装される。少なくとも一実施例では、コンピュータ・システム1400は、デスクトップ・コンピュータ、ラップトップ・コンピュータ、タブレット・コンピュータ、サーバ、スーパーコンピュータ、スマート・フォン（たとえば、ワイヤレスの携帯型デバイス）、パーソナル・デジタル・アシスタント（「PDA」）、デジタル・カメラ、車両、頭装着型ディスプレイ、携帯型電子デバイス、モバイル・フォン・デバイス、テレビ、ワークステーション、ゲーム・コンソール、組み込みシステム、及び／又は任意の他のタイプの論理の形をとってもよい。

【0549】

少なくとも一実施例では、並列処理システム1412は、限定することなく、複数の並列処理ユニット（「PPU」）1414、及び関連メモリ1416を含む。少なくとも一実施例では、PPU1414は、相互接続1418及びスイッチ1420又はマルチプレクサを介してホスト・プロセッサ又は他の周辺デバイスに接続される。少なくとも一実施例では、並列処理システム1412は、計算タスクをPPU1414にわたって分配し、これは、たとえば複数のグラフィックス・プロセッシング・ユニット（「GPU」）のスレッド・ブロックにわたる計算タスクの分配の一部として、並列化可能とすることができる。少なくとも一実施例では、メモリは、PPU1414の一部又は全部にわたって共有され、（たとえば、読取り及び／又は書込みアクセスのために）アクセス可能であるが、こうした共有メモリは、PPU1414に常駐しているローカル・メモリ及びレジスタの使用に対して、性能に不利益をもたらすことがある。少なくとも一実施例では、PPU1414の動作は、`_sync_threads()`などのコマンドを使用することによって同期され、ここで（たとえば、複数のPPU1414にわたって動作している）ブロック

10

20

30

40

50

内のすべてのスレッドが、進行前にコードのある一定の実行ポイントに到達する。

【0550】

他の変形形態は、本開示の範囲内にある。したがって、開示した技法は、様々な修正及び代替的な構成が可能であるが、それらのうち一定の例示的な実施例が図面に示され、上で詳細に説明されてきた。しかし、特定の1つ又は複数の開示された形に本開示を限定する意図はなく、その反対に、特許請求の範囲に定義される開示の趣旨及び範囲に入るすべての修正形態、代替的な構成、及び等価物を網羅することを意図している。

【0551】

開示される実施例を説明する文脈において（特に、以下の特許請求の範囲の文脈において）「a」及び「an」及び「the」という用語、並びに同様の指示語を使用することは、本明細書に別段の記載のない限り、又は文脈によって明らかに否定されない限り、単数と複数の両方を網羅すると解釈されるべきであり、用語の定義であると解釈されるべきではない。「備える (comprising)」、「有する (having)」、「含む (including)」、「収容する (containing)」という用語は、別段の記載のない限り、オープンエンドの用語（「含むが、これに限定されない」を意味する）と解釈される。「接続される」は、修飾されずに物理的接続を指している場合には、何か介在するものがあっても、部分的に又は完全に中に収容される、取り付けられる、又は互いに接合されるものとして解釈される。本明細書において値の範囲を詳述することは、本明細書において別段の記載がない限り、またそれぞれ別々の値が、本明細書に個々に詳述されているかのように明細書に組み込まれていない限り、範囲内に含まれるそれぞれ別々の値を個々に参照する簡潔な方法として機能することを単に意図しているにすぎない。少なくとも一実施例では、「セット」（たとえば、「アイテムのセット」）又は「サブセット」という用語の使用は、文脈によって別段の記載がない、又は否定されていない限り、1つ又は複数の部材を備える空ではない集合として解釈されるべきである。さらに、文脈によって別段の記載がない、又は否定されていない限り、対応するセットの「サブセット」という用語は、対応するセットの厳密なサブセットを必ずしも指すのではなく、サブセットと対応するセットは等しくてもよい。

【0552】

「A、B、及びCのうちの少なくとも1つ」又は「A、B、及びCのうちの少なくとも1つ」という形の言い回しなどの結合語は、別段の具体的な記載のない限り、又は文脈によって明確に否定されていない限り、項目、用語などが、AかBかCである、又はAとBとCのセットのいずれかの空でないサブセットであることを提示するために一般に使用される文脈で理解される。たとえば、3つの部材を有するセットの説明的な例では、「A、B、及びCのうちの少なくとも1つ」並びに「A、B、及びCのうちの少なくとも1つ」という結合句は、次のセットのうちのいずれかを指す：{A}、{B}、{C}、{A、B}、{A、C}、{B、C}、{A、B、C}。したがって、こうした結合語は、ある一定の実施例が、少なくとも1つのA、少なくとも1つのB、及び少なくとも1つのCのそれぞれの存在を必要とすることを全体的に暗示するものではない。さらに、別段の記載のない、又は文脈によって否定されていない限り、「複数」という用語は、複数である状態を示す（たとえば、「複数の項目 (a plurality of items)」は複数の項目 (multiple items) を示す）。少なくとも一実施例では、複数である項目の数は、少なくとも2つであるが、明示的に、又は文脈によって示されている場合にはそれより多くてもよい。さらに、別段の記載のない、又は文脈からそうでないことが明らかでない限り、「～に基づく」という言い回しは、「少なくとも部分的に～に基づく」を意味し、「～だけにに基づく」を意味しない。

【0553】

本明細書に記載のプロセスの動作は、本明細書に別段の記載のない、又は文脈によって明確に否定されない限り、任意の好適な順序で実行することができる。少なくとも一実施例では、本明細書に記載のプロセス（又はその変形及び／又は組合せ）などのプロセスは、実行可能命令で構成された1つ又は複数のコンピュータ・システムの制御下で実行され

10

20

30

40

50

、1つ又は複数のプロセッサ上で、ハードウェアによって、又はそれらの組合せによって集合的に実行されるコード（たとえば、実行可能な命令、1つ若しくは複数のコンピュータ・プログラム、又は1つ若しくは複数のアプリケーション）として実装される。少なくとも一実施例では、コードは、たとえば1つ又は複数のプロセッサによって実行可能な複数の命令を備えるコンピュータ・プログラムの形で、コンピュータ読取り可能ストレージ媒体に記憶される。少なくとも一実施例では、コンピュータ読取り可能ストレージ媒体は、一時的な信号（たとえば、伝播する一時的な電気若しくは電磁送信）を除外するが、一時的な信号のトランシーバ内の非一時的なデータ・ストレージ回路（たとえば、バッファ、キャッシュ、及びキュー）を含む非一時的なコンピュータ読取り可能ストレージ媒体である。少なくとも一実施例では、コード（たとえば、実行可能コード又はソース・コード）は、1つ又は複数の非一時的なコンピュータ読取り可能ストレージ媒体のセットに記憶され、このストレージ媒体には、コンピュータ・システムの1つ又は複数のプロセッサによって実行されたときに（すなわち、実行された結果として）、コンピュータ・システムに本明細書に記載の動作を実行させる実行可能命令が記憶されている（又は、実行可能命令を記憶するための他のメモリを有する）。少なくとも一実施例では、非一時的なコンピュータ読取り可能ストレージ媒体のセットは、複数の非一時的なコンピュータ読取り可能ストレージ媒体を備え、複数の非一時的なコンピュータ読取り可能ストレージ媒体の個々の非一時的なストレージ媒体のうちの1つ又は複数には、すべてのコードがないが、複数の非一時的なコンピュータ読取り可能ストレージ媒体は、集合的にすべてのコードを記憶している。少なくとも一実施例では、実行可能命令は、異なる命令が異なるプロセッサによって実行されるように実行され、たとえば、非一時的なコンピュータ読取り可能ストレージ媒体は命令を記憶し、メインの中央処理装置（「CPU」）は一部の命令を実行し、グラフィックス・プロセッシング・ユニット（「GPU」）は他の命令を実行する。少なくとも一実施例では、コンピュータ・システムの異なる構成要素は、別々のプロセッサを有し、異なるプロセッサは、命令の異なるサブセットを実行する。

【0554】

したがって、少なくとも一実施例では、コンピュータ・システムは、本明細書に記載のプロセスの動作を単独で又は集合的に実行する1つ又は複数のサービスを実装するように構成され、こうしたコンピュータ・システムは、動作の実行を可能にする適用可能なハードウェア及び／又はソフトウェアで構成される。さらに、本開示の少なくとも一実施例を実装するコンピュータ・システムは、単一のデバイスであり、別の実施例では、異なるやり方で動作する複数のデバイスを備える分散型のコンピュータ・システムであり、それにより単一のデバイスがすべての動作を実行しないように分散型のコンピュータ・システムが本明細書に記載の動作を実行する。

【0555】

本明細書に提供されるあらゆる例、又は例示的な言葉（たとえば、「など」）の使用は、本開示の実施例をより明らかにすることだけを意図しており、別段の主張のない限り、本開示の範囲に制限を加えるものではない。本明細書のいかなる言葉も、特許請求されていない任意の要素を、本開示の実践に不可欠なものとして示すと解釈されるべきではない。

【0556】

本明細書に引用される出版物、特許出願、及び特許を含むすべての参考文献は、各参考文献が参照により組み込まれることがあたかも個別に明確に示され、その全体が本明細書に記載されたかのように、それと同程度まで参照により本明細書に組み込まれる。

【0557】

明細書及び特許請求の範囲において、「結合される」及び「接続される」という用語が、その派生語とともに使用されてもよい。これらの用語は、互いに同義語として意図されていない場合があることを理解すべきである。むしろ、特定の例では、「接続される」又は「結合される」は、2つ以上の要素が物理的又は電氣的に互いに直接又は間接的に接触していることを示すために使用されてもよい。また「結合される」は、2つ以上の要素が直接互いに接触していないが、なお互いに連動又は相互作用することを意味してもよい。

【0558】

別段の具体的な記載のない限り、明細書全体を通して「処理する」、「コンピューティング」、「計算する」、又は「判定する」などの用語は、コンピューティング・システムのレジスタ及び／又はメモリ内の、電子的などの物理的な量として表されるデータをコンピューティング・システムのメモリ、レジスタ、又は他のそのような情報ストレージ・デバイス、送信デバイス、若しくはディスプレイ・デバイス内の物理的な量として同様に表される他のデータになるよう操作及び／又は変換するコンピュータ若しくはコンピューティング・システム、又は同様の電子コンピューティング・デバイスの行為及び／又はプロセスを指す。

【0559】

同様に、「プロセッサ」という用語は、レジスタ及び／又はメモリからの電子データを処理し、その電子データを、レジスタ及び／又はメモリに記憶することができる他の電子データに変換する任意のデバイス、又はデバイスの一部分を指してもよい。非限定的な例として、「プロセッサ」は、CPU又はGPUであってもよい。「コンピューティング・プラットフォーム」は、1つ又は複数のプロセッサを備えてもよい。本明細書で使用する「ソフトウェア」プロセスは、たとえば、タスク、スレッド、及び知的エージェントなど、経時的にワークを実行するソフトウェア及び／又はハードウェアのエンティティを含んでもよい。また、各プロセスは、命令を直列で又は並列で連続的に又は断続的に実行するための複数のプロセスを指してもよい。少なくとも一実施例では、「システム」及び「方法」は、1つ又は複数の方法をシステムが具体化することができ、方法がシステムと考えられてもよい場合に限り、本明細書において交換可能に使用される。

【0560】

本明細書では、アナログ・データ又はデジタル・データを得る、取得する、受信する、又はそれらをサブシステム、コンピュータ・システム、又はコンピュータ実装機械に入力することに言及することができる。少なくとも一実施例では、アナログ・データ又はデジタル・データを得る、取得する、受信する、又は入力するプロセスは、関数呼出し、又はアプリケーション・プログラミング・インターフェースへの呼出しのパラメータとしてデータを受信するなど、様々なやり方で実現することができる。少なくとも一実施例では、アナログ・データ又はデジタル・データを得る、取得する、受信する、又は入力するプロセスは、直列又は並列のインターフェースを介してデータを転送することによって実現することができる。少なくとも一実施例では、アナログ・データ又はデジタル・データを得る、取得する、受信する、又は入力するプロセスは、提供するエンティティから取得するエンティティにコンピュータ・ネットワークを介してデータを転送することによって実現することができる。また、少なくとも一実施例では、アナログ・データ又はデジタル・データを提供する、出力する、送信する、送る、又は提示することにも言及することができる。様々な例では、アナログ・データ又はデジタル・データを提供する、出力する、送信する、送る、又は提示するプロセスは、関数呼出しの入力又は出力のパラメータ、アプリケーション・プログラミング・インターフェース若しくはプロセス間通信機構のパラメータとしてデータを転送することによって実現することができる。

【0561】

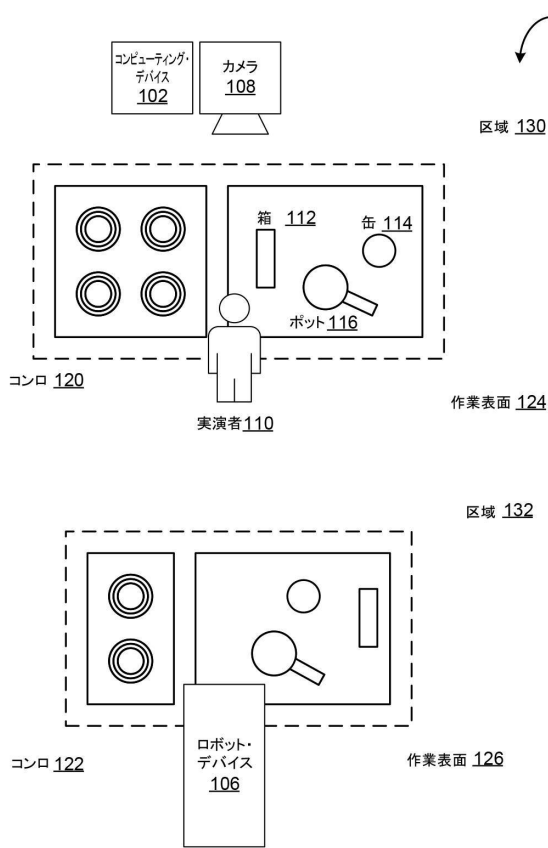
本明細書の記載は、記載した技法の例示的な実装形態について述べているが、記載した機能を実装するために他のアーキテクチャが使用されてもよく、この他のアーキテクチャは、本開示の範囲内にあることが意図される。さらに、説明を目的として、役割の具体的な分配が定義され得るが、様々な機能及び役割は、状況に応じて異なるやり方で分配及び分割されてもよい。

【0562】

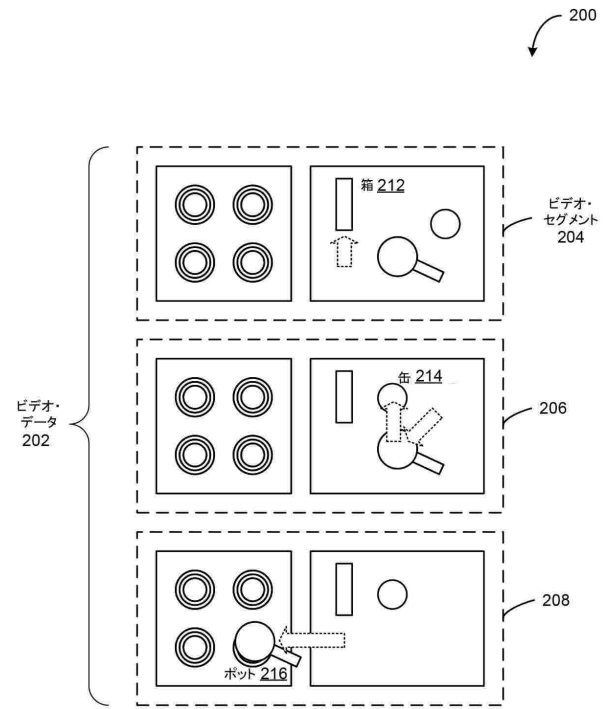
さらに、主題は、構造的特徴及び／又は方法論的動作に特有の言語で説明されてきたが、添付の特許請求の範囲で特許請求される主題は、説明した特有の特徴又は動作に必ずしも限定されないことが理解されるべきである。むしろ、特有の特徴及び動作は、特許請求の範囲を実装する例示的な形として開示されている。

【図面】

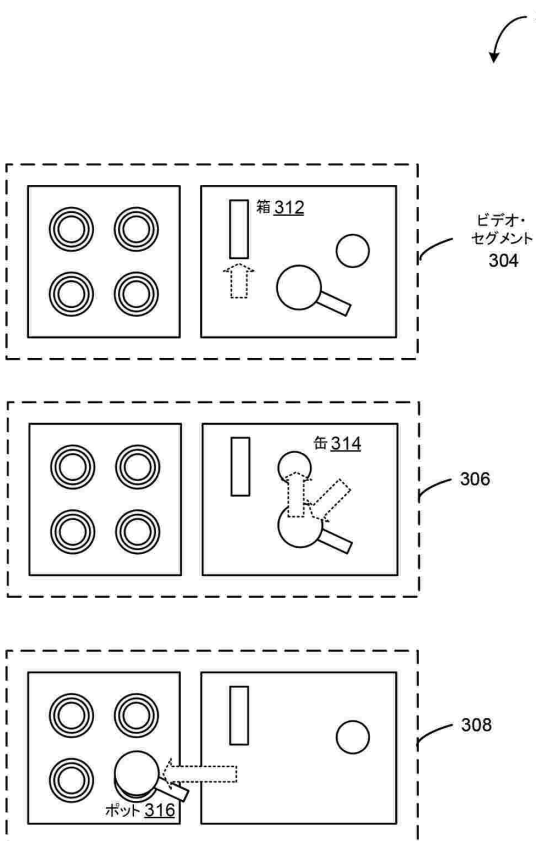
【図 1】



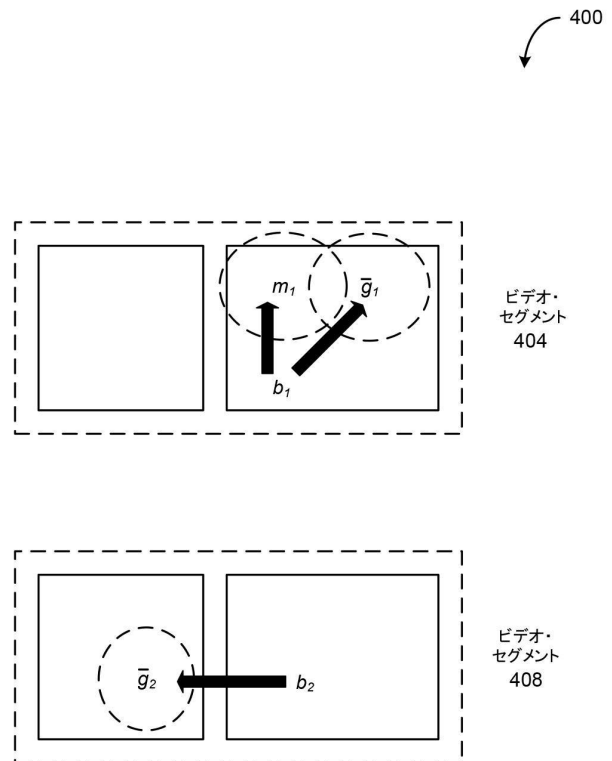
【図 2】



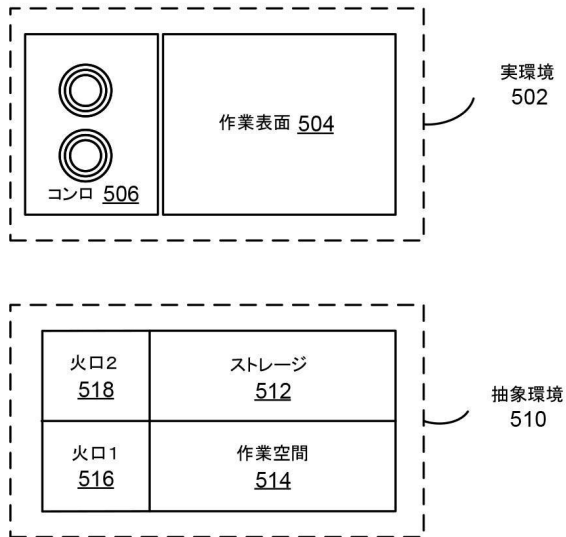
【図 3】



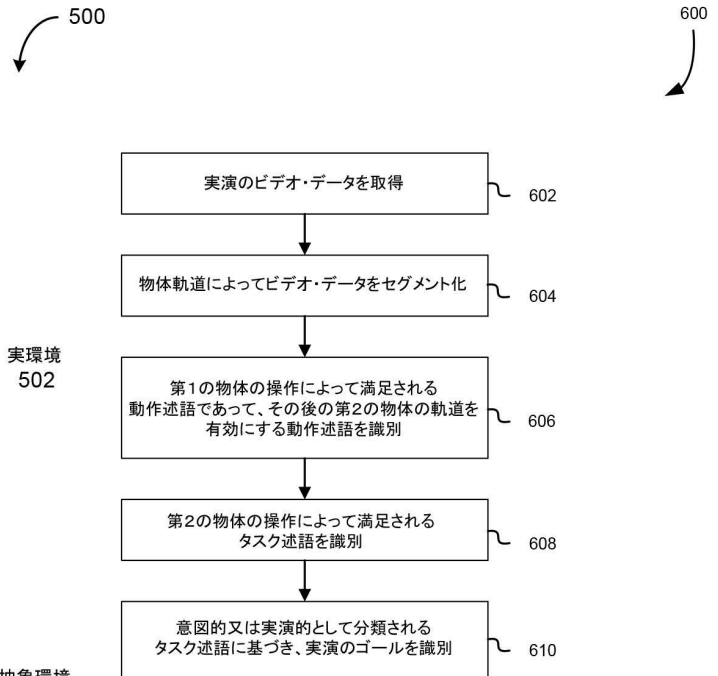
【図 4】



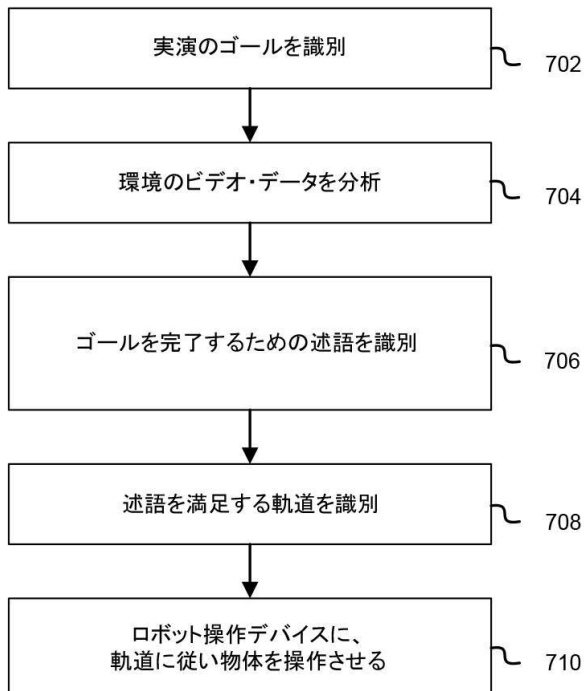
【図 5】



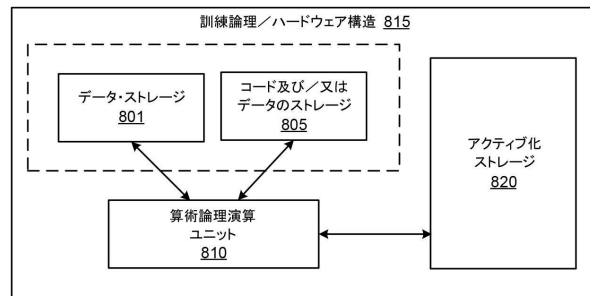
【図 6】



【図 7】



【図 8 A】



10

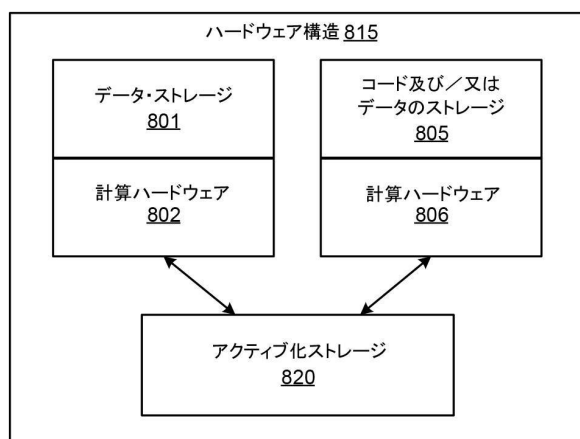
20

30

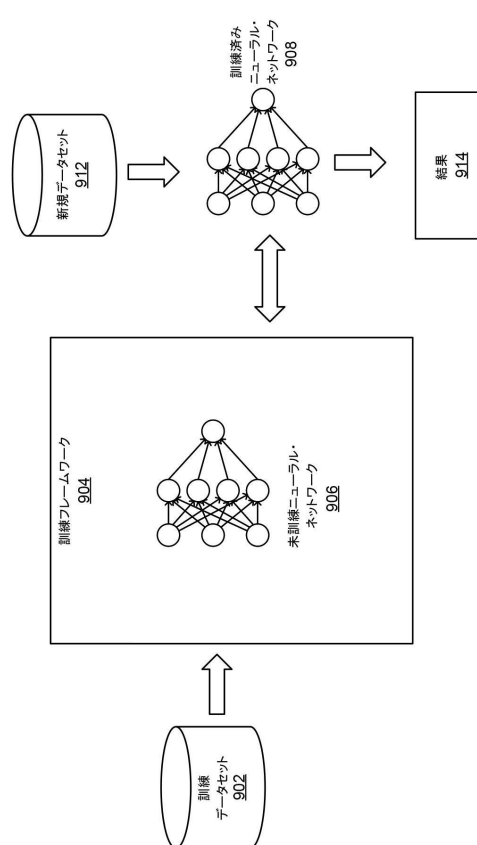
40

50

【図 8 B】



【圖 9】

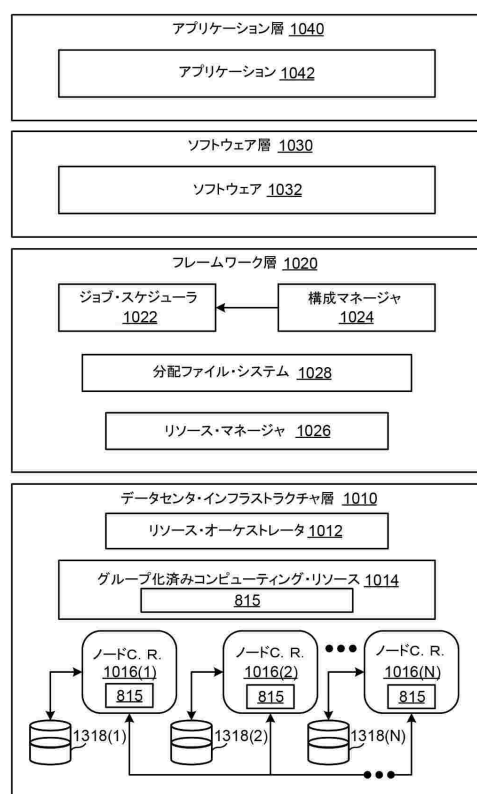


10

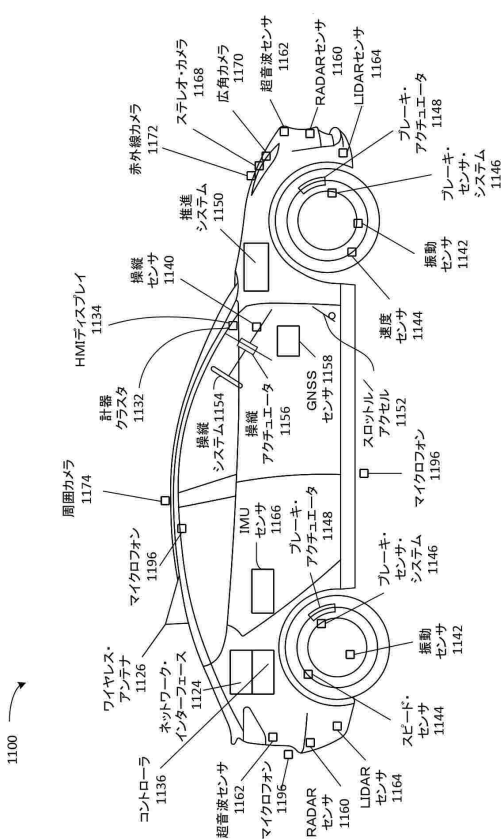
20

【図 10】

データセンタ
1000



【図 1 1 A】

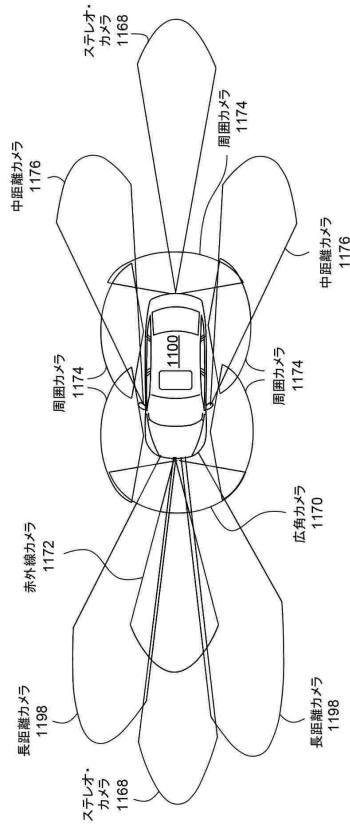


30

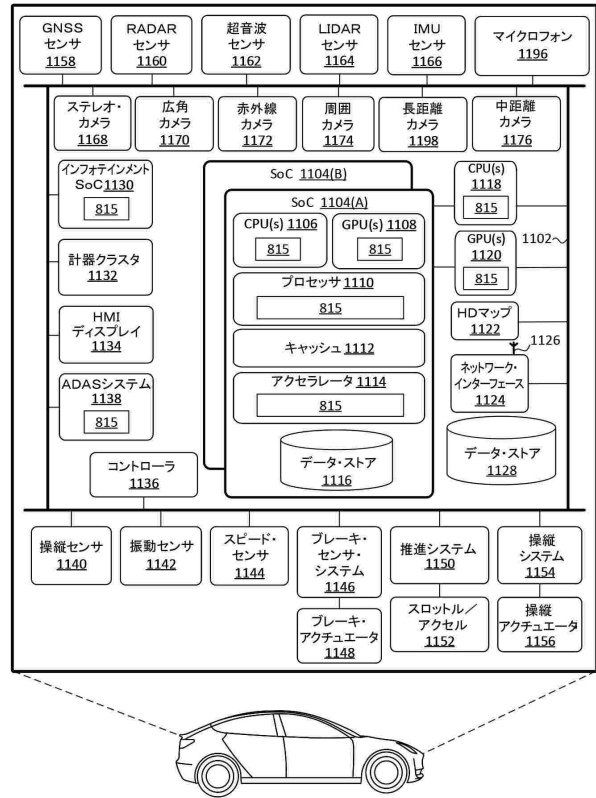
40

50

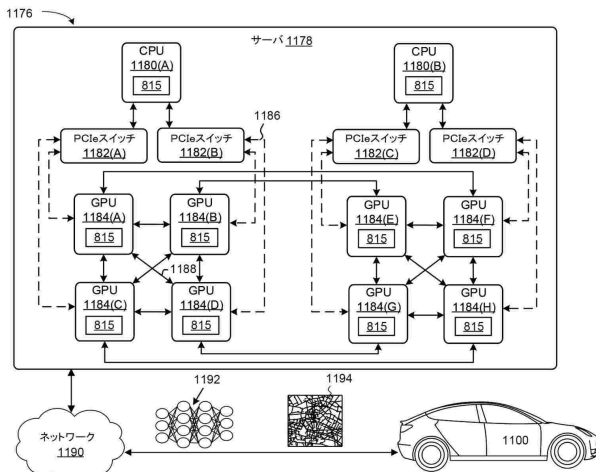
【図 1 1 B】



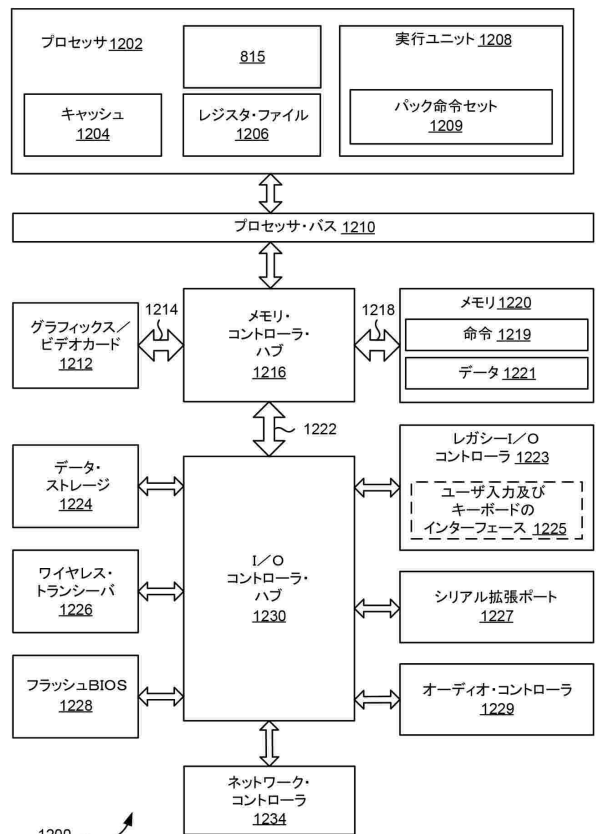
【図 1 1 C】



【図 1 1 D】



【図 1 2】



10

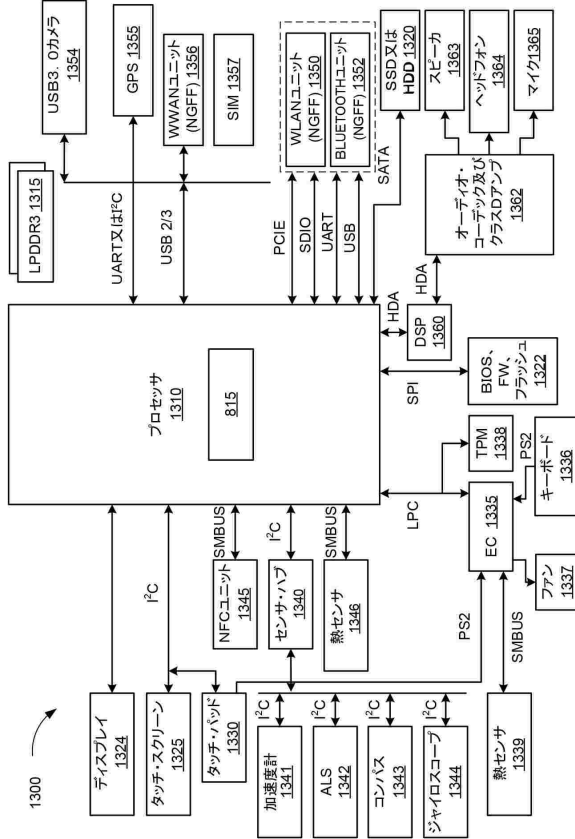
20

30

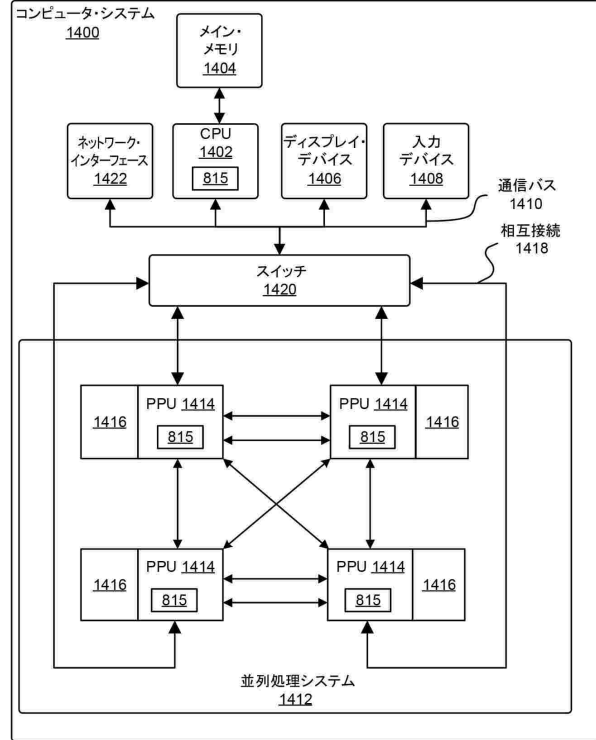
40

50

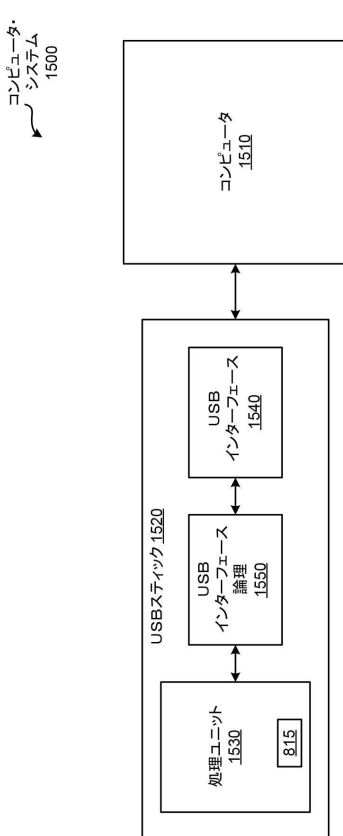
【図 13】



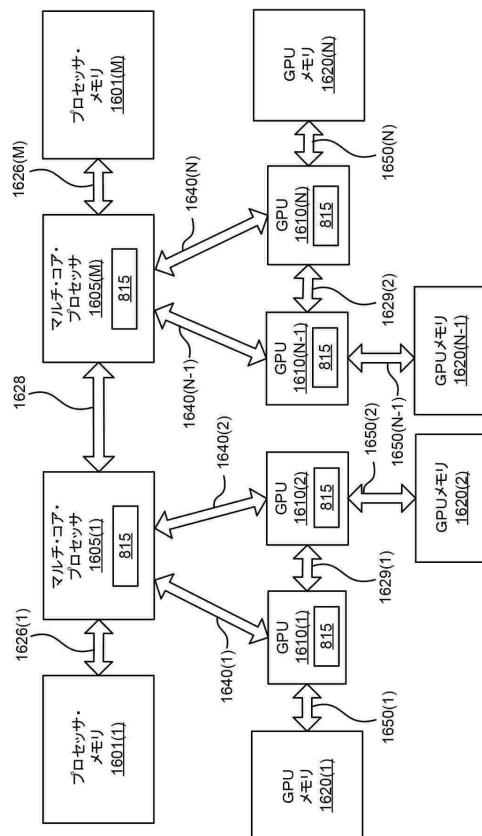
【図 14】



【図 15】



【図 16 A】



10

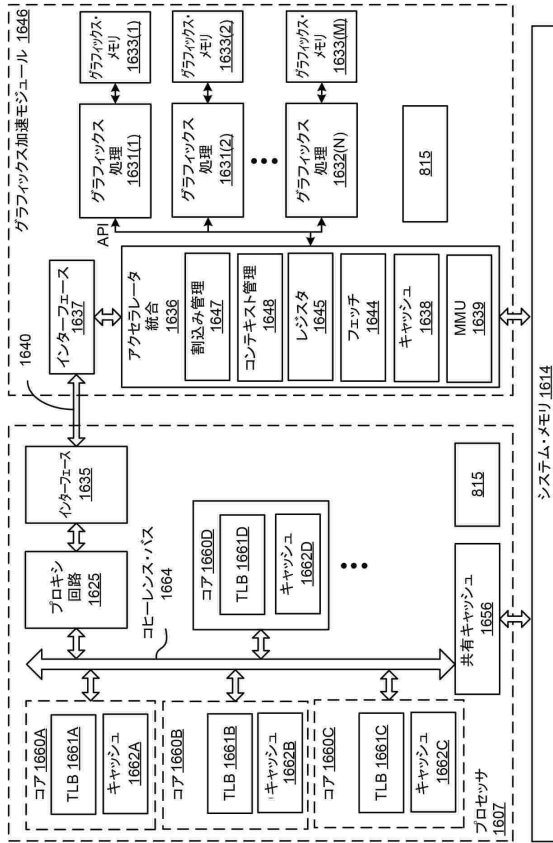
20

30

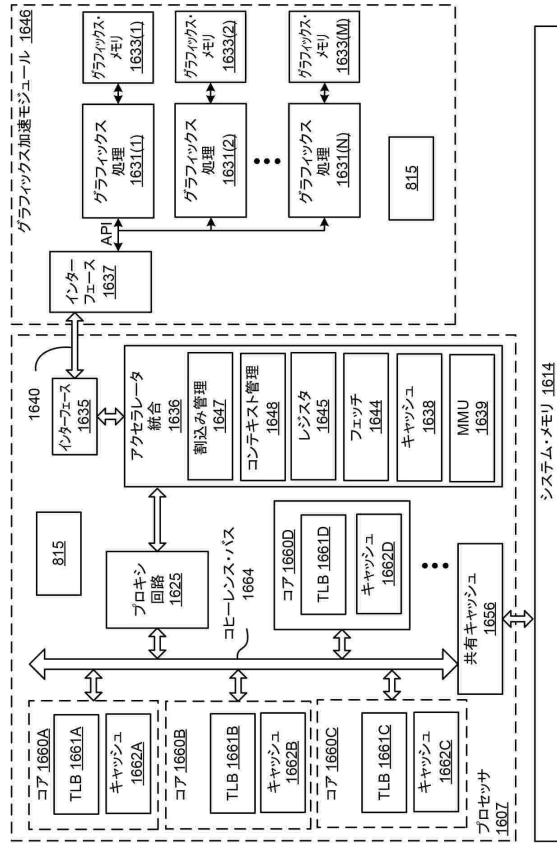
40

50

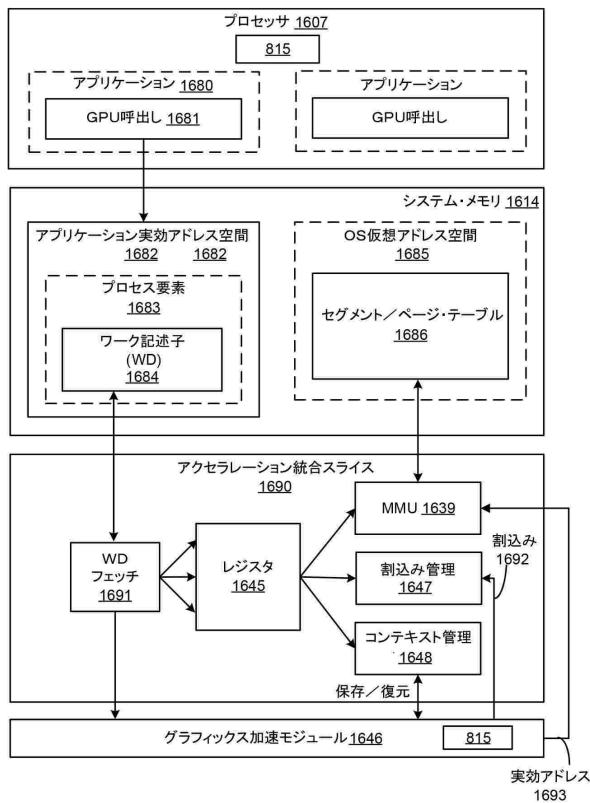
【図 16 B】



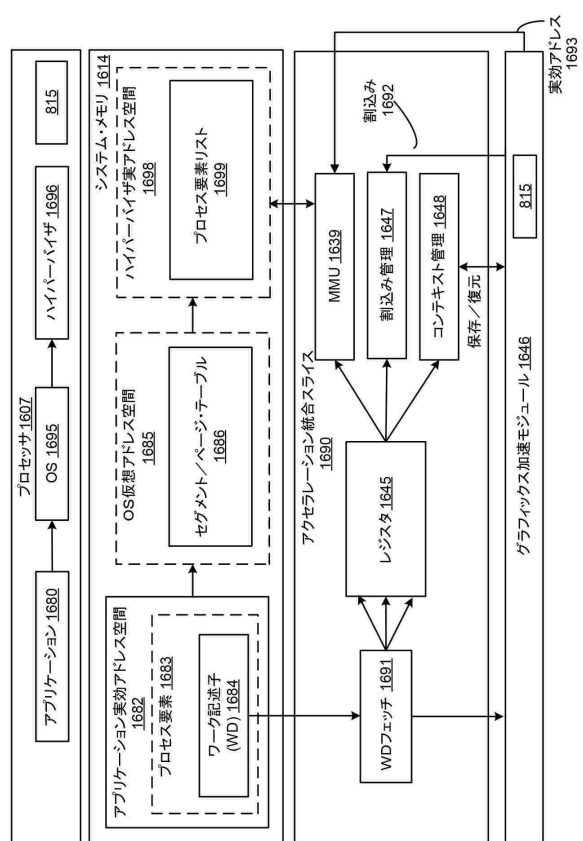
【図 16 C】



【図 16 D】



【図 16 E】



10

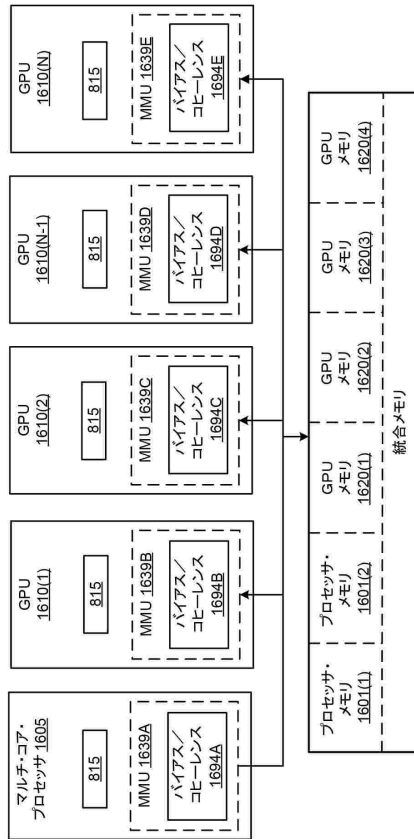
20

30

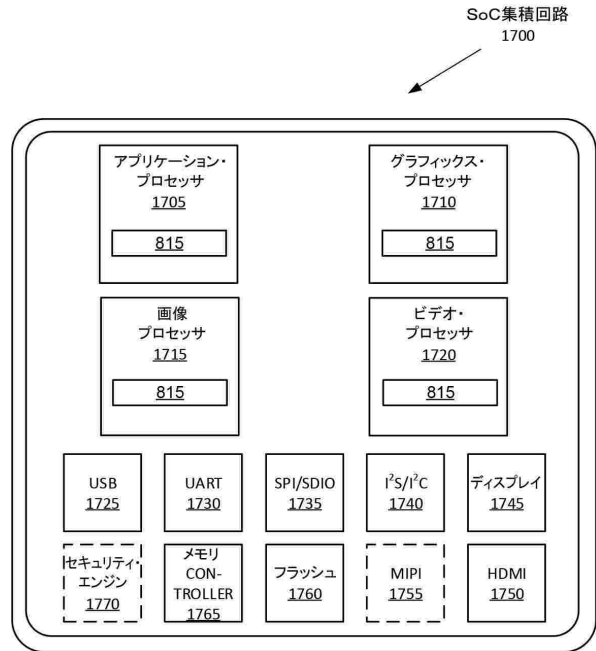
40

50

【図 16 F】



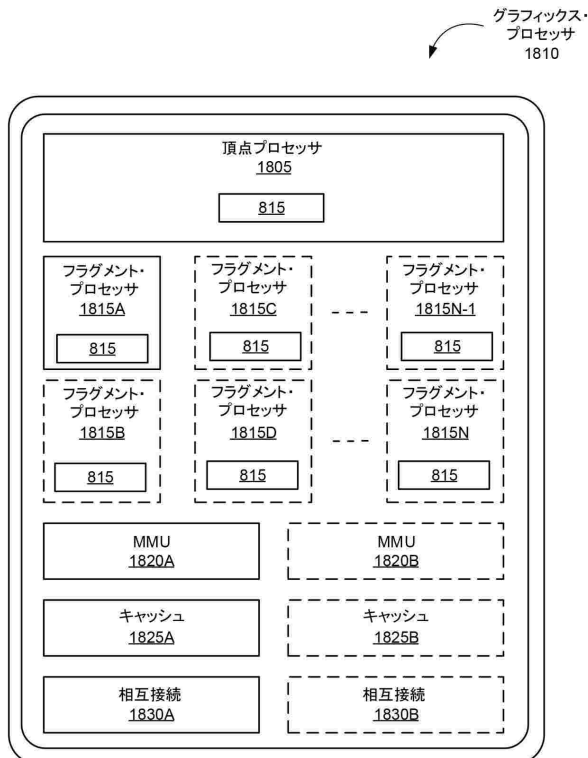
【図 17】



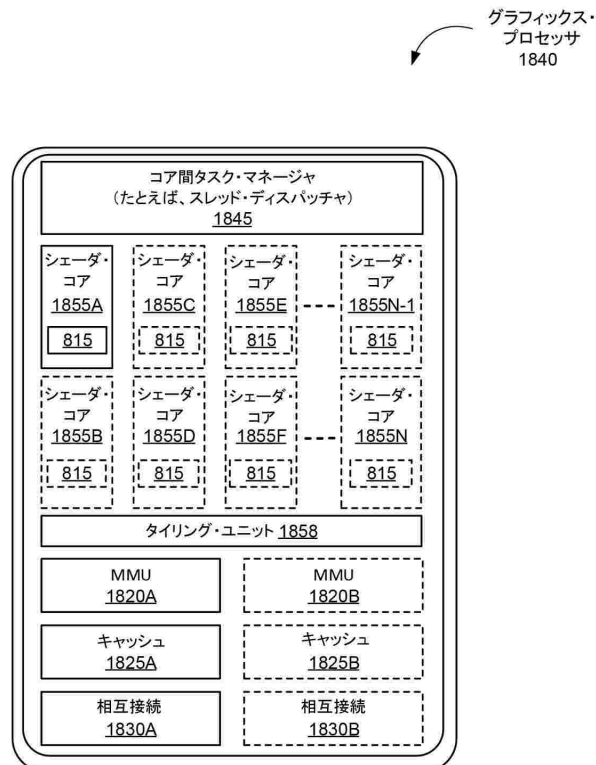
10

20

【図 18 A】



【図 18 B】

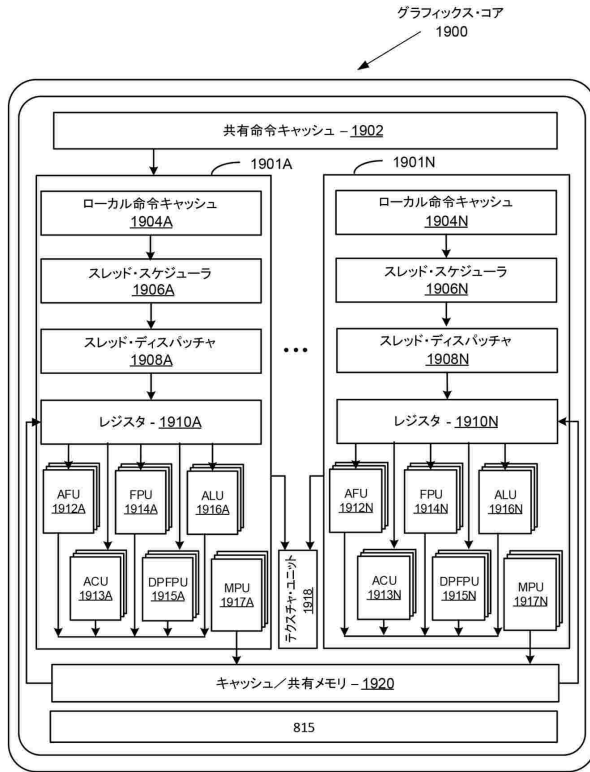


30

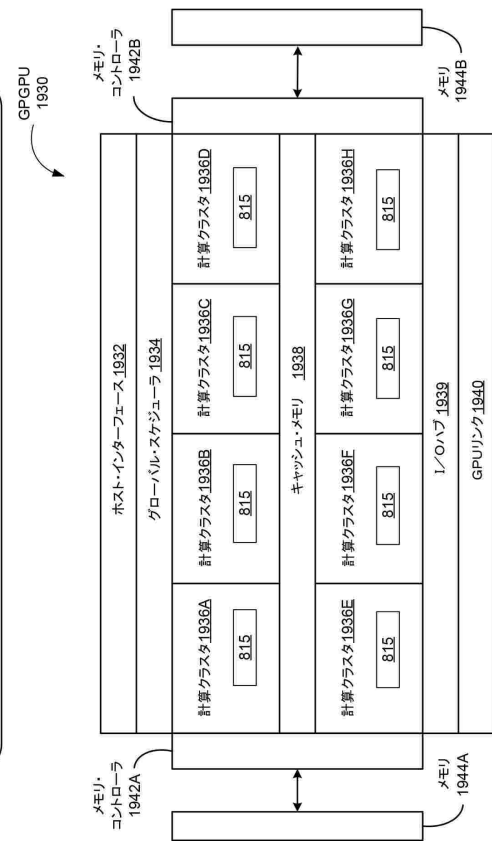
40

50

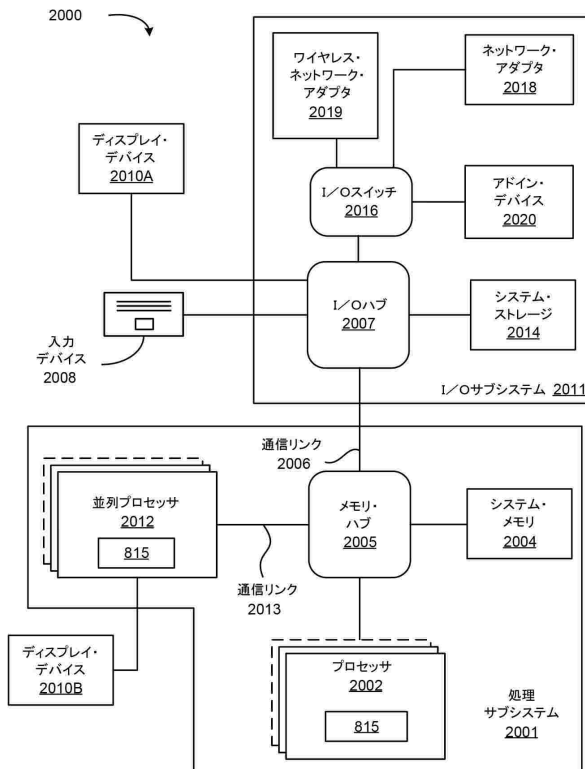
【図 19 A】



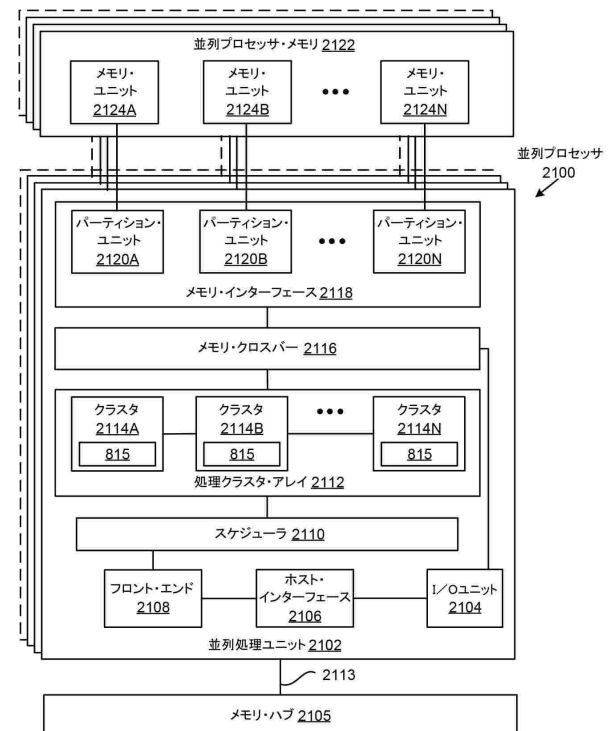
【図 19 B】



【図 20】



【図 21 A】



10

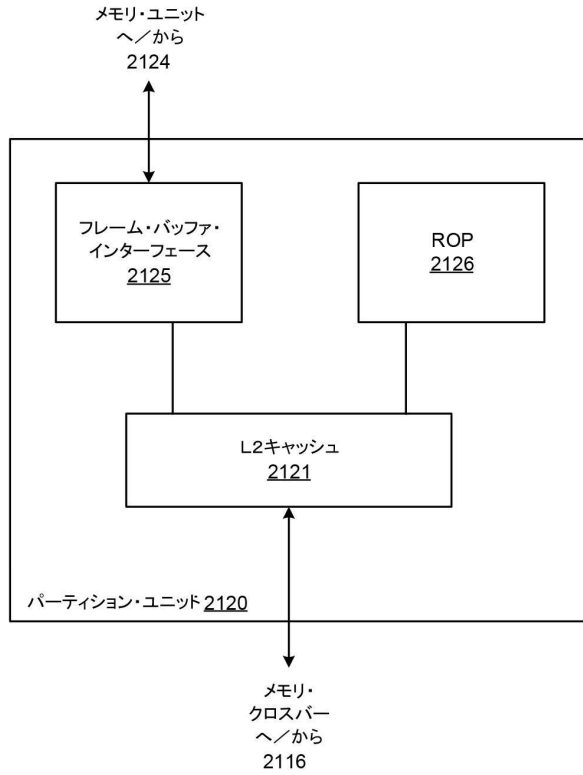
20

30

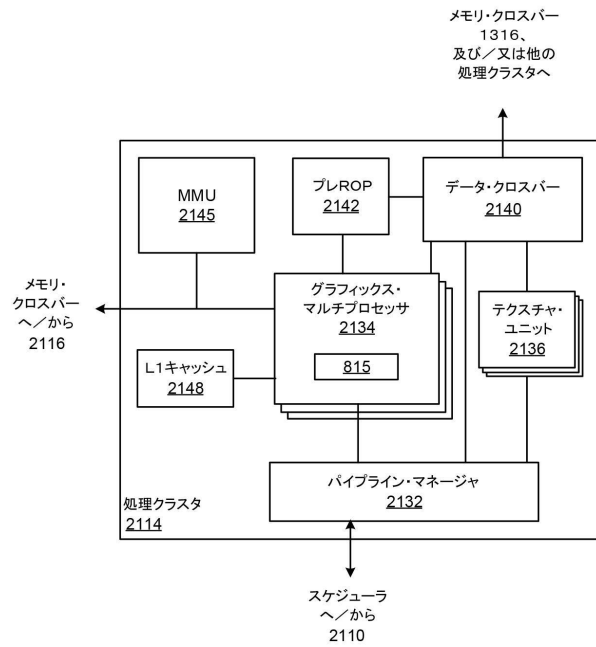
40

50

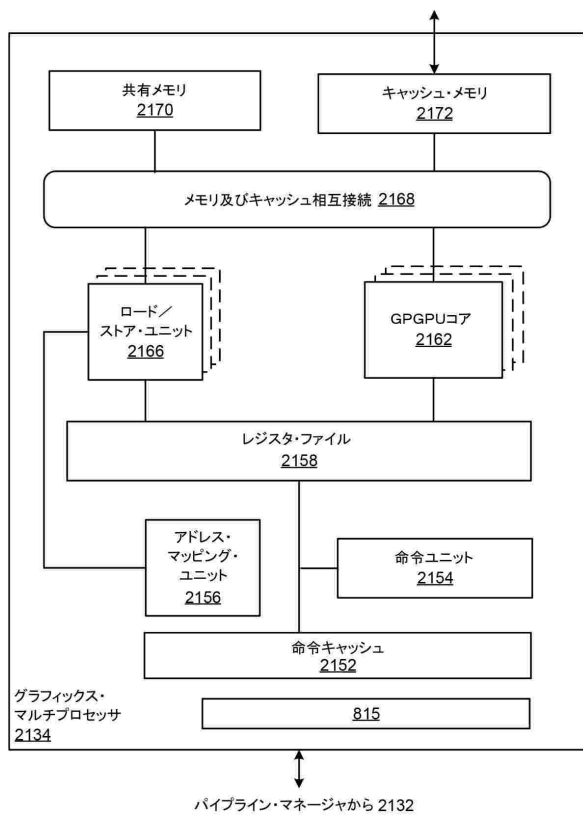
【図 2 1 B】



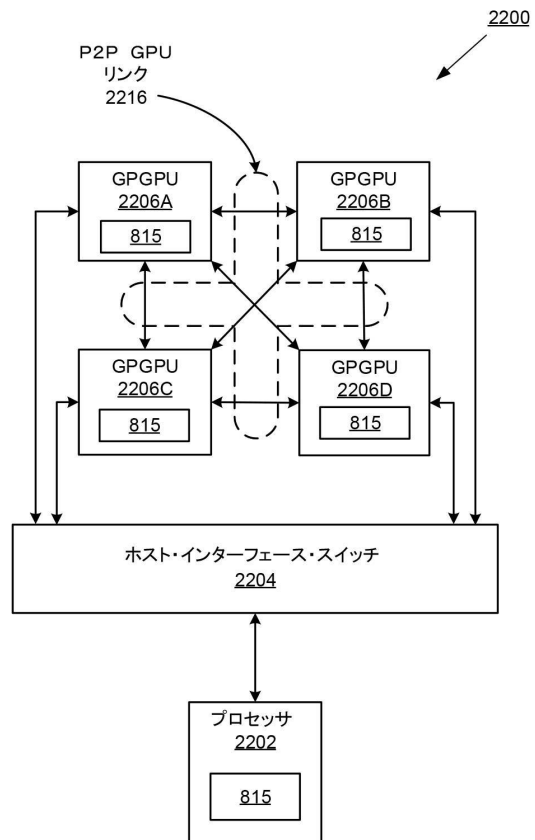
【図 2 1 C】



【図 2 1 D】



【図 2 2】



10

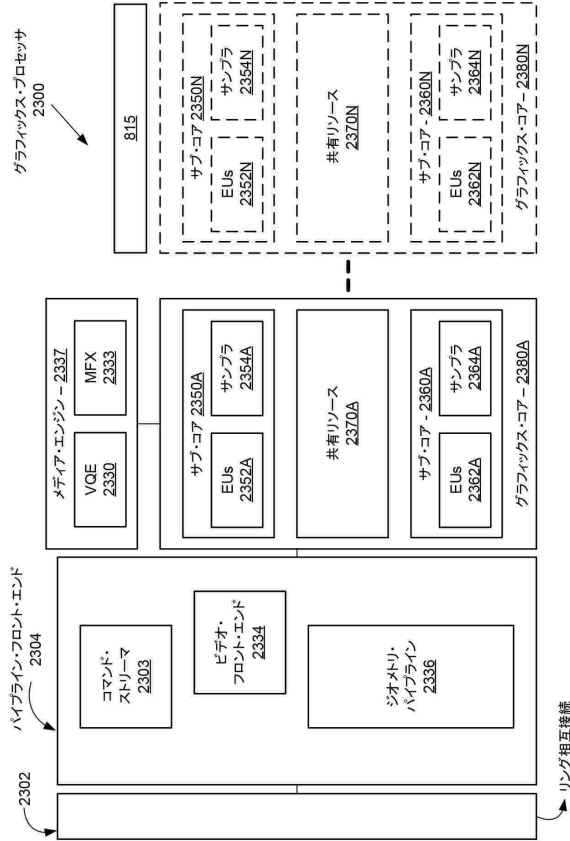
20

30

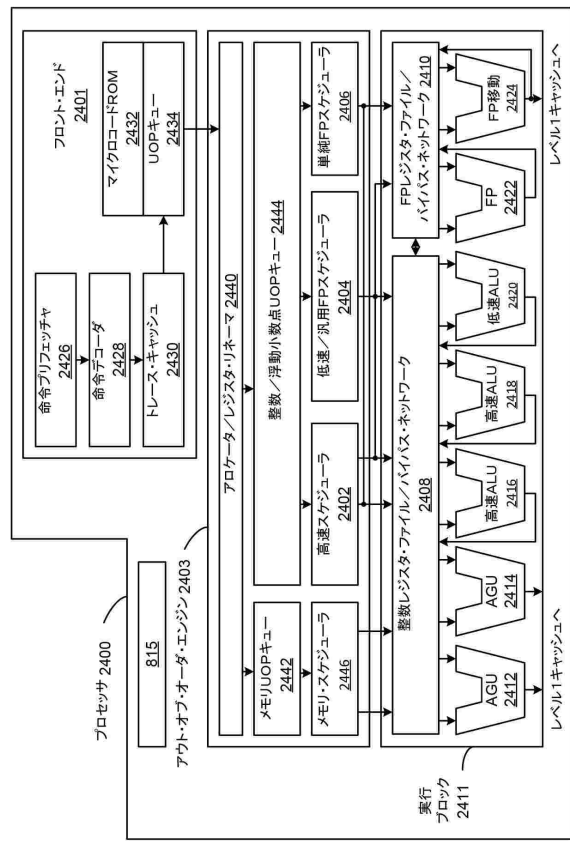
40

50

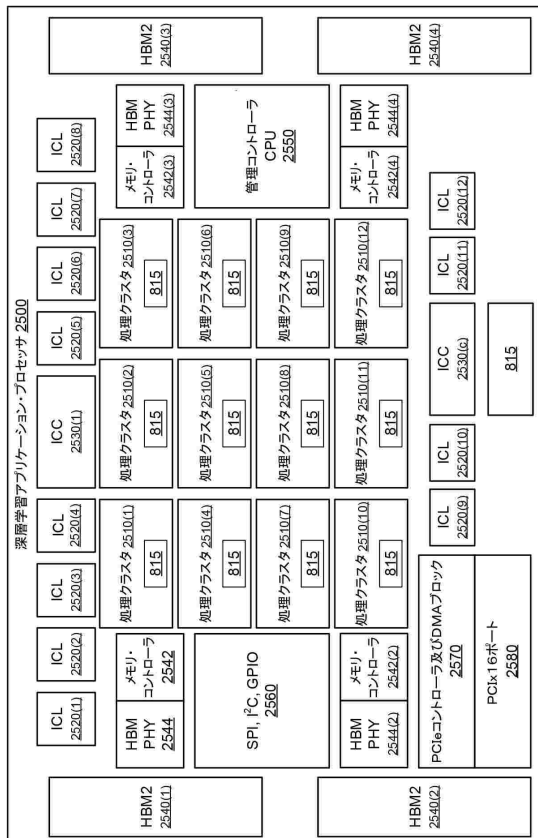
【図 23】



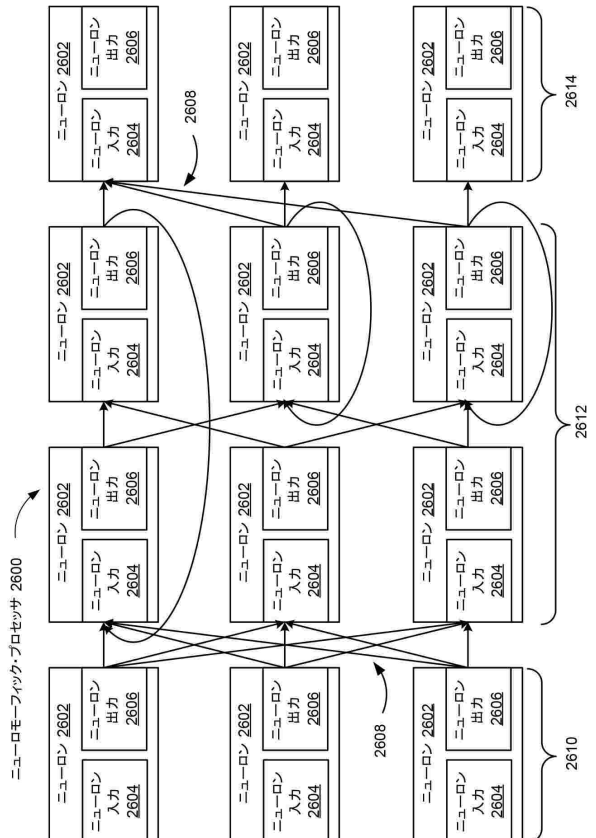
【図 24】



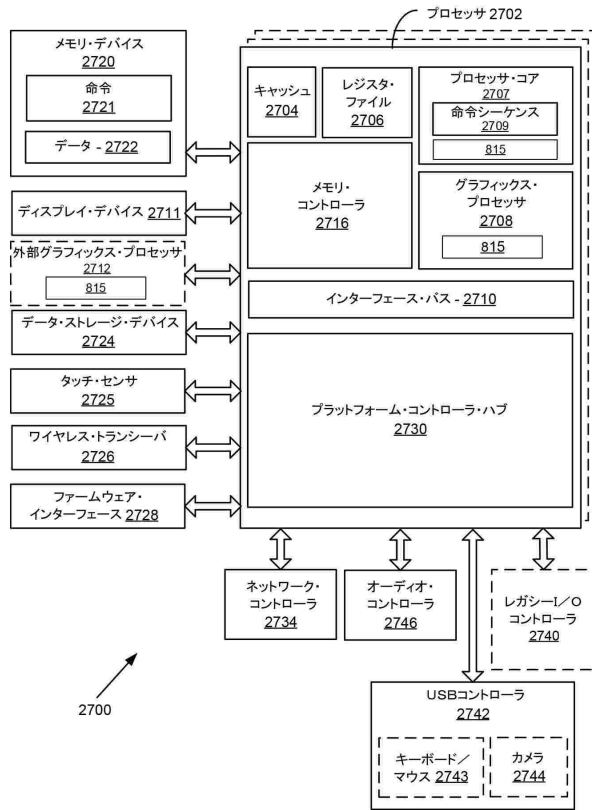
【図 25】



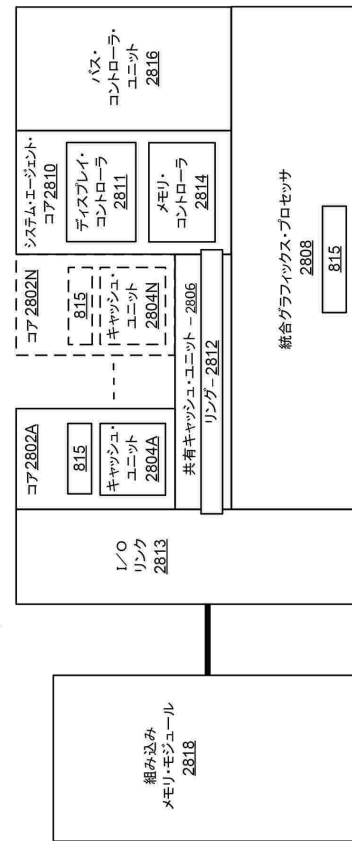
【図 26】



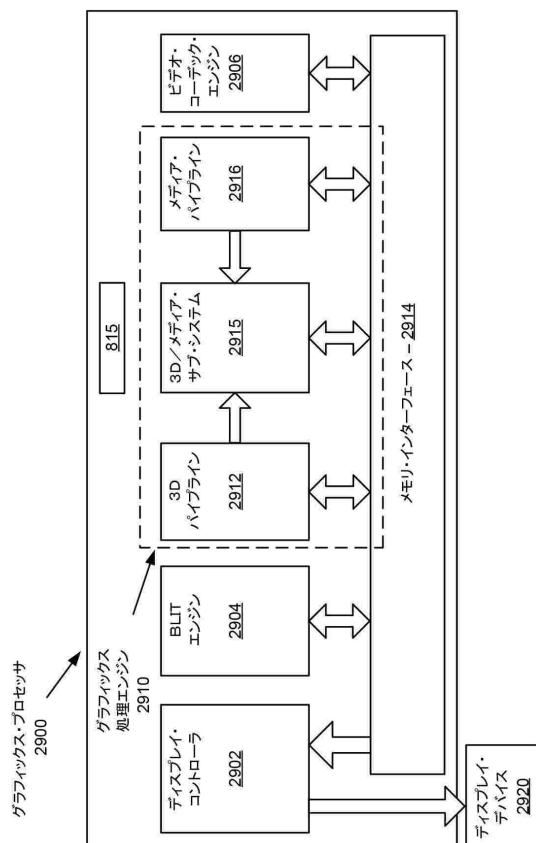
【図 27】



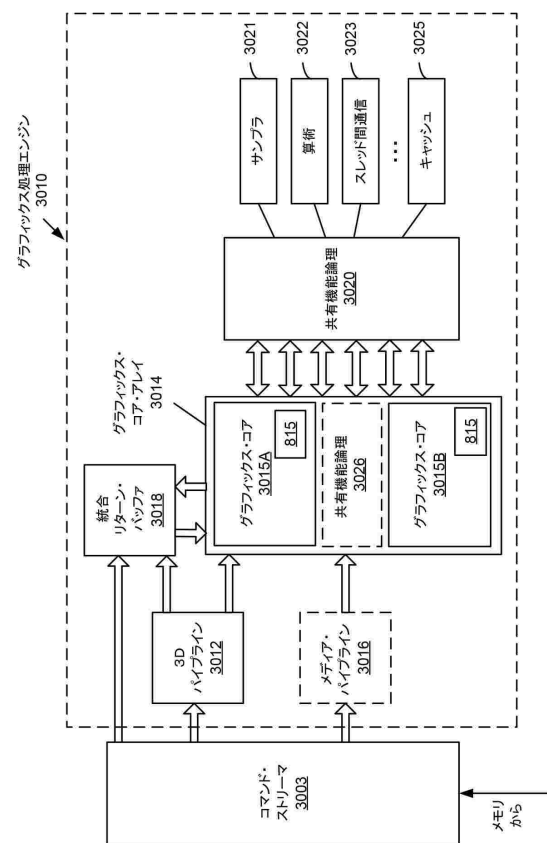
【図 28】



【図 29】



【図 30】



10

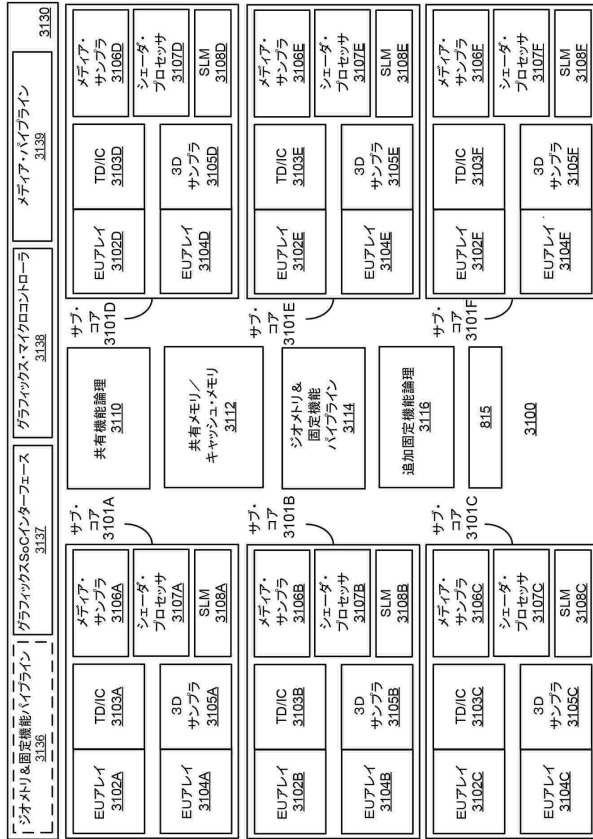
20

30

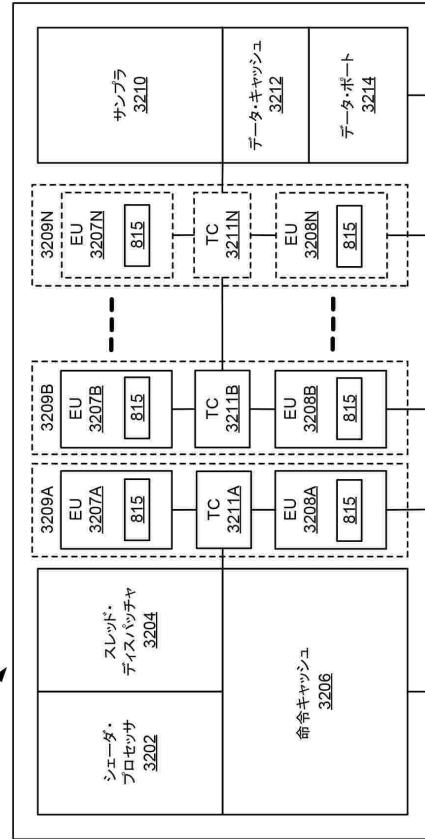
40

50

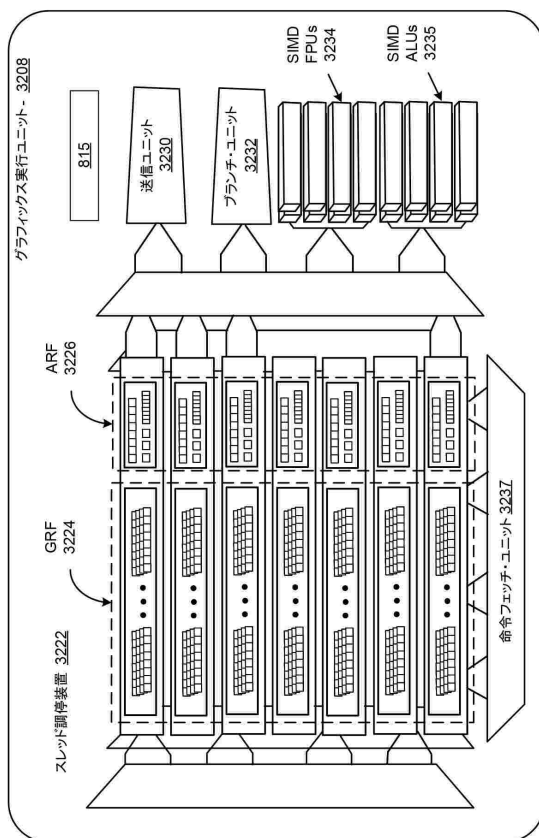
【図 3 1】



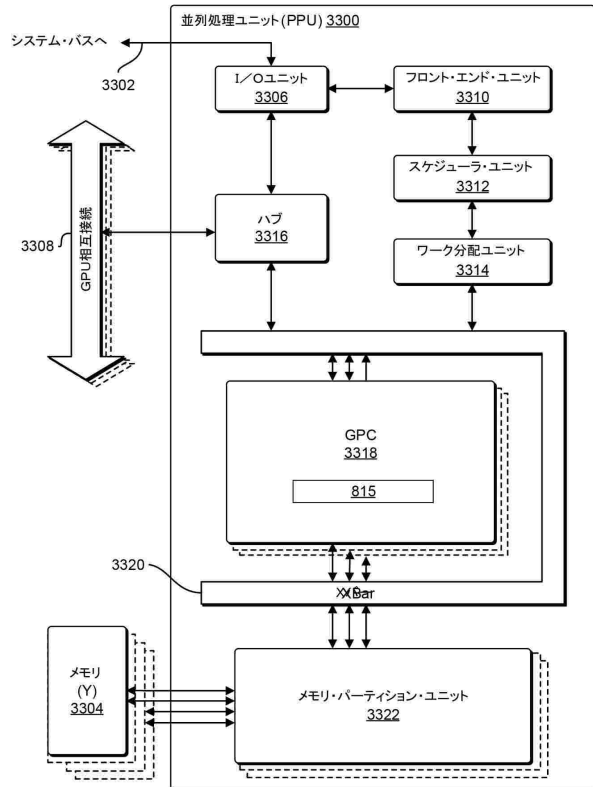
【図 3 2 A】



【図 3 2 B】



【図 3 3】



10

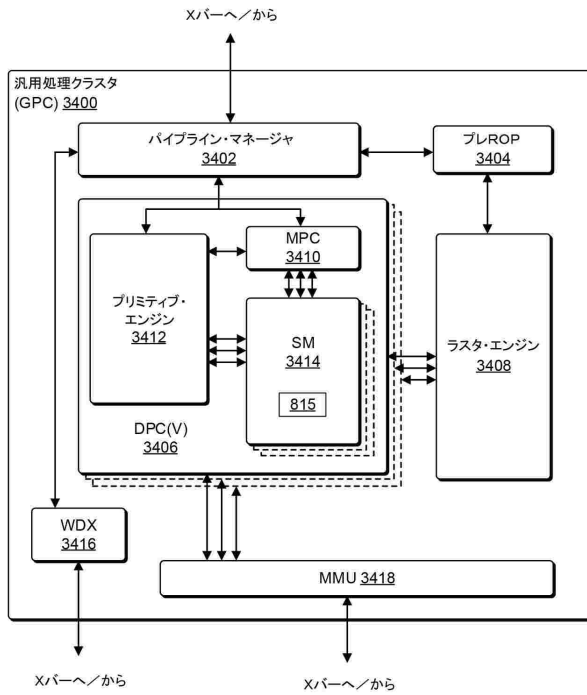
20

30

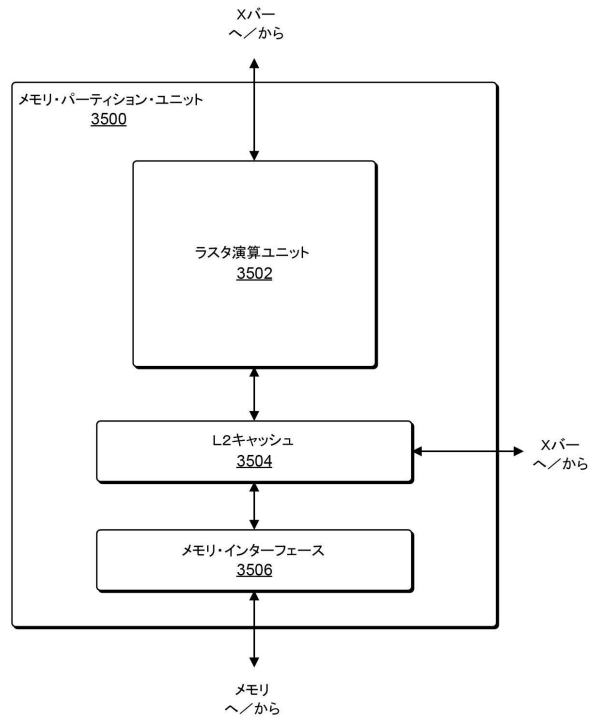
40

50

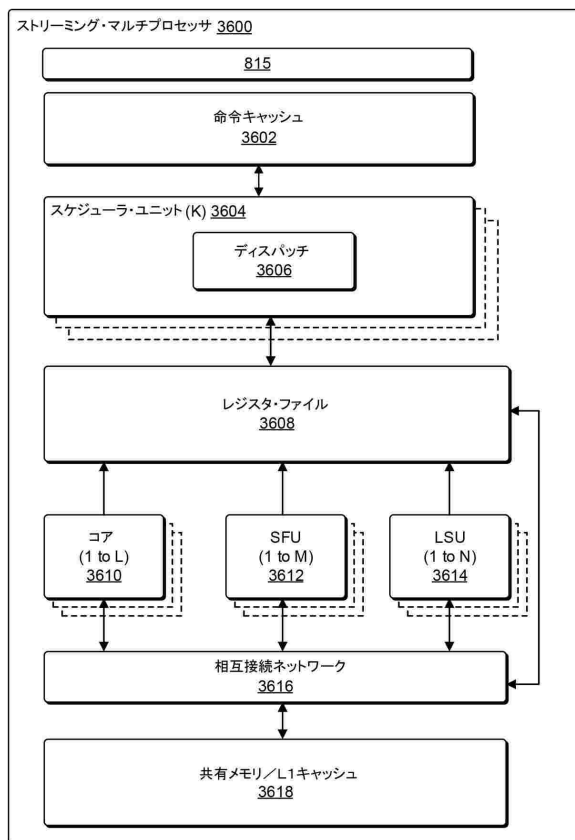
【図 3 4】



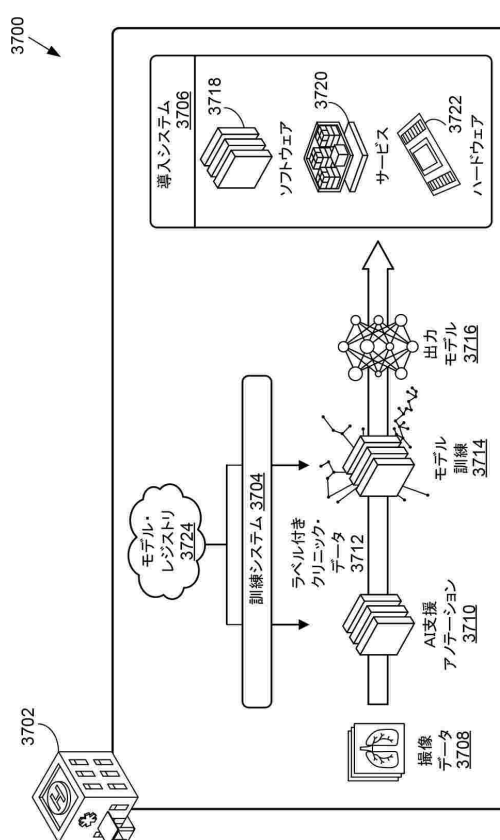
【図 3 5】



【図 3 6】



【図 3 7】



10

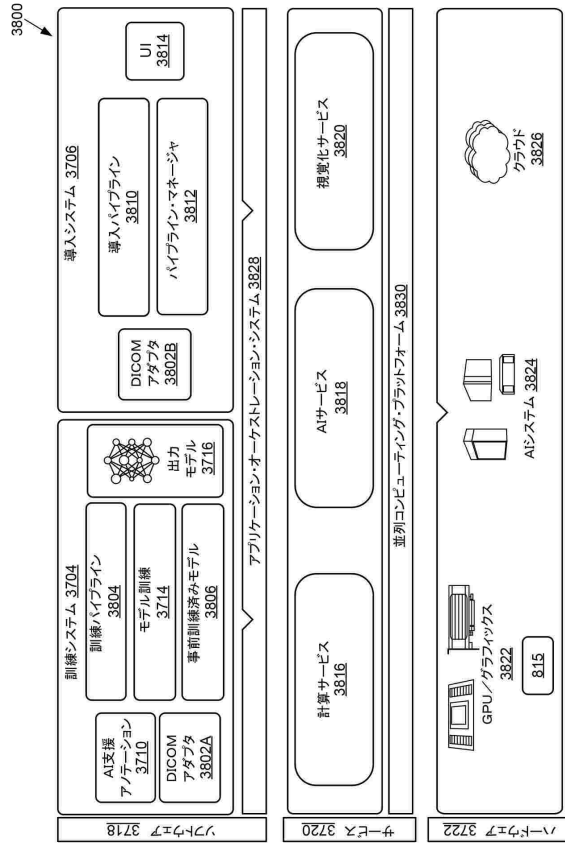
20

30

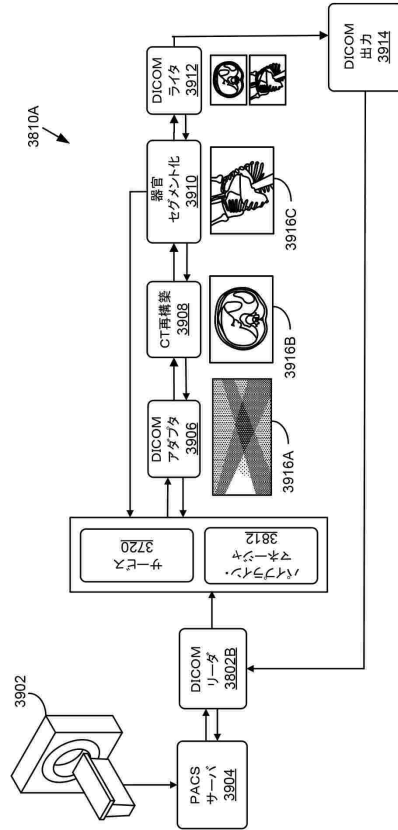
40

50

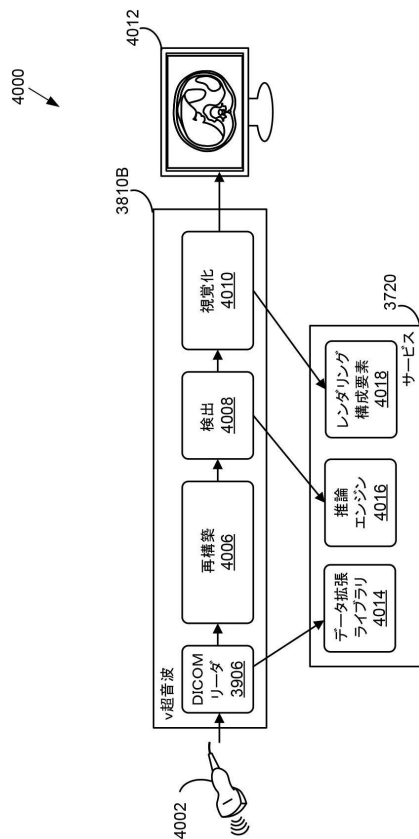
【図 38】



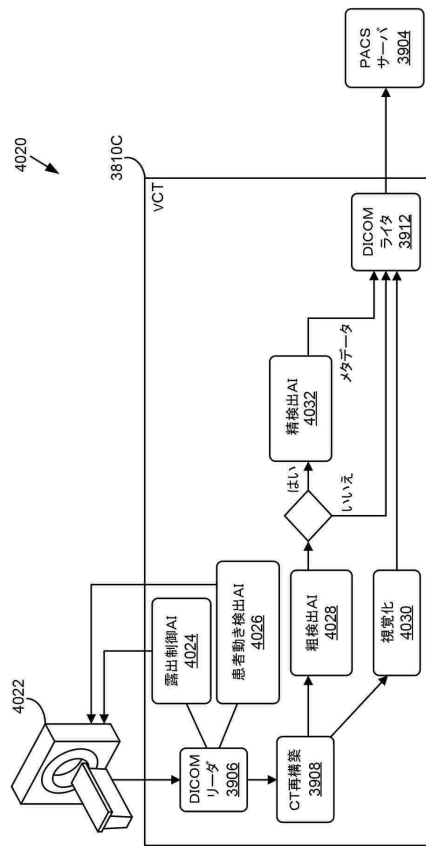
【図 39】



【図 40 A】



【図 40 B】



10

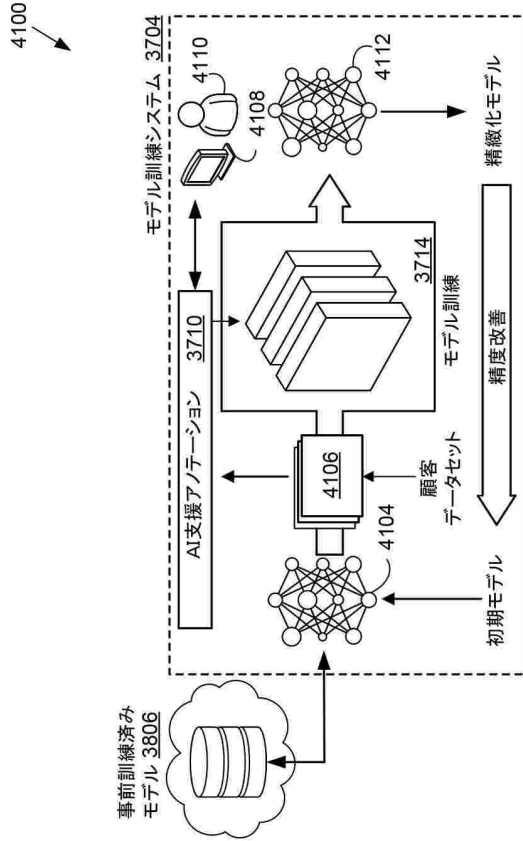
20

30

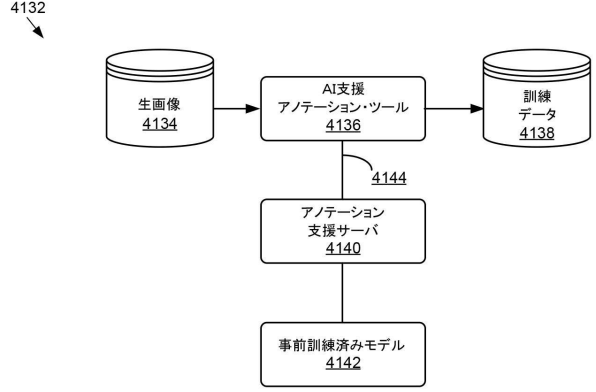
40

50

【図 4 1 A】



【図 4 1 B】



10

20

30

40

50

フロントページの続き

- (72)発明者 クリストファー ジェイソン パクストン
 アメリカ合衆国、カリフォルニア、サンタクララ、サン トーマス エクスプレスウェイ 2788
- (72)発明者 アニメシュ ガルグ
 アメリカ合衆国、カリフォルニア、サンタクララ、サン トーマス エクスプレスウェイ 2788
- (72)発明者 ディーター フォックス
 アメリカ合衆国、カリフォルニア、サンタクララ、サン トーマス エクスプレスウェイ 2788
- 審査官 武田 広太郎
- (56)参考文献 欧州特許出願公開第02366502 (EP, A1)
 特開2016-058074 (JP, A)
- (58)調査した分野 (Int.Cl., DB名)
 G06N 20/00
 B25J 9/22