

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4005083号
(P4005083)

(45) 発行日 平成19年11月7日(2007. 11. 7)

(24) 登録日 平成19年8月31日(2007. 8. 31)

(51) Int. Cl.

G06F 12/08 (2006.01)

F I

G06F 12/08 505Z

G06F 12/08 559D

請求項の数 16 (全 15 頁)

(21) 出願番号	特願2004-517262 (P2004-517262)	(73) 特許権者	395015319
(86) (22) 出願日	平成15年6月24日(2003. 6. 24)		株式会社ソニー・コンピュータエンタテインメント
(65) 公表番号	特表2005-531847 (P2005-531847A)		ンメント
(43) 公表日	平成17年10月20日(2005. 10. 20)		東京都港区南青山二丁目6番21号
(86) 国際出願番号	PCT/JP2003/007959	(74) 代理人	100099324
(87) 国際公開番号	W02004/003753		弁理士 鈴木 正剛
(87) 国際公開日	平成16年1月8日(2004. 1. 8)	(74) 代理人	100108604
審査請求日	平成17年3月4日(2005. 3. 4)		弁理士 村松 義人
(31) 優先権主張番号	10/187, 072	(74) 代理人	100111615
(32) 優先日	平成14年7月1日(2002. 7. 1)		弁理士 佐野 良太
(33) 優先権主張国	米国 (US)		

最終頁に続く

(54) 【発明の名称】 キャッシュメモリへのプリフェッチを制御するための方法ならびに装置

(57) 【特許請求の範囲】

【請求項 1】

複数の二次元のブロックに分割されているメインメモリを有し、各ブロックは1つ以上のデータバイトを有する複数のデータ単位を備え、各所定のブロックは複数のゾーンに分割されており、各ゾーンは前記データ単位を1つ以上備えるとともに前記ゾーンの少なくとも一部は、前記所定のブロックに隣接する対応するブロックに関連付けられているものであり、

前記メインメモリのデータバイトに対するアドレスを受け取ると共に、前記アドレスをデコードして、対応するデータバイトが存在する少なくとも一つのブロックのゾーンを特定するように動作可能なアドレスデコーダと、

前記アドレスデコーダが対応するゾーンに存在するデータバイトのアドレスを受け取ると、前記少なくとも一つのブロックがアクセス可能かつ前記対応するゾーンに関連付けられているブロックがプリフェッチ可能となるように、前記少なくとも一つのブロックの前記対応するゾーンに関連付けられているブロックをキャッシュメモリに更に追加してプリフェッチするように動作可能とされているプリフェッチアドレスジェネレータと、を有するメモリ制御装置。

【請求項 2】

前記各ブロックは、左の隣接ブロックと当該ブロックとで共有する左境界線、右の隣接ブロックと当該ブロックとで共有する右境界線、上の隣接ブロックと当該ブロックとで共有する上境界線、および下の隣接ブロックと当該ブロックとで共有する下境界線、のうち

の少なくとも1つを有し、

前記各ブロックは、前記左の隣接ブロックに関連付けられて前記左境界線に沿って配置されたデータ単位を少なくとも1つ有する左ゾーン、前記右の隣接ブロックに関連付けられて前記右境界線に沿って配置されたデータ単位を少なくとも1つ有する右ゾーン、前記上の隣接ブロックに関連付けられて前記上境界線に沿って配置されたデータ単位を少なくとも1つ有する上ゾーン、および前記下の隣接ブロックに関連付けられて前記下境界線に沿って配置されたデータ単位を少なくとも1つ有する下ゾーン、のうちの少なくとも1つを有する請求項1に記載の装置。

【請求項3】

前記左ゾーンは前記上境界線から前記下境界線まで延びる左分割線によって少なくともその一部が画定されており、前記左ゾーンは前記左境界線と前記左分割線との間に存在するデータ単位の少なくとも一部を有するものであり、

10

前記右ゾーンは前記上境界線から前記下境界線まで延びる右分割線によって少なくともその一部が画定されており、前記右ゾーンは前記右境界線と前記右分割線との間に存在するデータ単位の少なくとも一部を有するものであり、

前記上ゾーンは前記左境界線から前記右境界線まで延びる上分割線によって少なくともその一部が画定されており、前記上ゾーンは前記上境界線と前記上分割線との間に存在するデータ単位の少なくとも一部を有するものであり、

前記下ゾーンは前記左境界線から前記右境界線まで延びる下分割線によって少なくともその一部が画定されており、前記下ゾーンは前記下境界線と前記下分割線との間に存在するデータ単位の少なくとも一部を有するものである、請求項2に記載の装置。

20

【請求項4】

前記左分割線、前記右分割線、前記上分割線、および前記下分割線は、前記アドレスの特定のビットに対応付けられており、前記アドレスデコーダは、前記アドレスをデコードして、前記特定のビットに基づいて前記対応するデータバイトが存在するゾーンを特定するように動作可能である請求項3に記載の装置。

【請求項5】

各ブロックは、前記左分割線、前記右分割線、前記上分割線、および前記下分割線の間に存在する1つ以上のデータ単位を有する中央ゾーンを有し、前記中央ゾーンはどの隣接ブロックとも関連付けられていない請求項3に記載の装置。

30

【請求項6】

各ブロックは、

前記左境界線、前記上境界線、前記上分割線および前記左分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない左上ゾーンと、

前記左境界線、前記下境界線、前記下分割線および前記左分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない左下ゾーンと、

前記右境界線、前記上境界線、前記上分割線および前記右分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない右上ゾーンと、

前記右境界線、前記下境界線、前記下分割線および前記右分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない右下ゾーンと、のうちの少なくとも1つを有する請求項3に記載の装置。

40

【請求項7】

メインメモリを複数の二次元のブロックに分割すると共に、各ブロックをデータバイトを1つ以上含む複数のデータ単位に分割するステップと、

前記メインメモリの各ブロックを、前記データ単位を1つ以上有する複数のゾーンに分割するステップと、

前記ブロックのうちの少なくとも1つのブロックの少なくとも1つのゾーンを他のブロックに関連付けるステップと、

前記少なくとも1つのゾーンに存在する前記データ単位のうちの1つのデータバイトの1つ以上がアドレス指定された場合に、前記少なくとも一つのブロックがアクセス可能で

50

かつ前記少なくとも1つのブロックの少なくとも1つのゾーンに関連付けられた他のブロックがプリフェッチ可能となるように、前記他のブロックをキャッシュメモリに更に追加してプリフェッチするステップと、を有する方法。

【請求項8】

メインメモリを複数の二次元のブロックに分割すると共に、各ブロックをデータバイトを1つ以上含む複数のデータ単位に分割するステップと、

前記メインメモリの各ブロックを、前記データ単位を1つ以上有する複数のゾーンに分割するステップと、

各ブロックの対応するゾーンの少なくとも一部を前記ブロックのうちの別のブロックに関連付けるステップと、

前記別のブロックのうちの1つのブロックに関連付けられているゾーンのいずれかにあるデータ単位のいずれかがアドレス指定された場合に、前記ブロックがアクセス可能でかつ前記対応するゾーンに関連付けられているブロックがプリフェッチ可能となるように、前記少なくとも一つのブロックの前記対応するゾーンに関連付けられているブロックをキャッシュメモリに更に追加してプリフェッチするステップと、を有する方法。

【請求項9】

各ブロックの前記少なくとも一部のゾーンのそれぞれは1つの隣接するブロックに関連付けられている請求項8に記載の方法。

【請求項10】

各ブロックの少なくとも1つのゾーンは、そのブロックと、関連付けられている隣接ブロックとの間の境界線に沿って配置されている請求項9に記載の方法。

【請求項11】

各ブロックは、そのブロックと、関連付けられている対応する隣接ブロックとの各境界線に沿って配置されている対応するゾーンを有する請求項10に記載の方法。

【請求項12】

メインメモリを複数の二次元の矩形ブロックに分割するステップであって、各ブロックは共通の境界線に沿って他のブロックのうちの少なくとも1つと隣接しており、各ブロックは、データバイトを1つ以上含む複数のデータ単位を有するステップと、

前記メインメモリの各ブロックを、前記データ単位を1つ以上有する複数のゾーンに分割するステップと、

各所定のブロックの対応するゾーンの少なくとも一部を前記所定のブロックに隣接するブロックの対応する別のブロックと関連付けるステップと、

前記別のブロックのうちのいずれかのブロックに関連付けられているゾーンのいずれかにあるデータ単位のいずれかがアドレス指定された場合に、

前記所定のブロックがアクセス可能でかつ前記別のブロックがプリフェッチ可能となるように、前記所定のブロックに隣接するブロックのうち前記対応する別のブロックをキャッシュメモリに更に追加してプリフェッチするステップと、を有する方法。

【請求項13】

前記各ブロックは、左の隣接ブロックと共通の左境界線、右の隣接ブロックと共通の右境界線、上の隣接ブロックと共通の上境界線、および下の隣接ブロックと共通の下境界線のうちの少なくとも1つを有し、

前記各ブロックは、前記左の隣接ブロックに関連付けられており、前記左境界線に沿って配置されたデータ単位を少なくとも1つ有する左ゾーン、前記右の隣接ブロックに関連付けられており、前記右境界線に沿って配置されたデータ単位を少なくとも1つ有する右ゾーン、前記上の隣接ブロックに関連付けられており、前記上境界線に沿って配置されたデータ単位を少なくとも1つ有する上ゾーン、および前記下の隣接ブロックに関連付けられており、前記下境界線に沿って配置されたデータ単位を少なくとも1つ有する下ゾーンのうちの少なくとも1つを有する請求項12に記載の方法。

【請求項14】

前記左ゾーンは前記上境界線から前記下境界線まで延びる左分割線によって少なくとも

10

20

30

40

50

一部画定されており、前記左ゾーンは前記左境界線と前記左分割線との間に存在するデータ単位の少なくとも一部を有し、

前記右ゾーンは前記上境界線から前記下境界線まで延びる右分割線によって少なくとも一部画定されており、前記右ゾーンは前記右境界線と前記右分割線との間に存在するデータ単位の少なくとも一部を有し、

前記上ゾーンは前記左境界線から前記右境界線まで延びる上分割線によって少なくとも一部画定されており、前記上ゾーンは前記上境界線と前記上分割線との間に存在するデータ単位の少なくとも一部を有し、

前記下ゾーンは前記左境界線から前記右境界線まで延びる下分割線によって少なくとも一部画定されており、前記下ゾーンは前記下境界線と前記下分割線との間に存在するデータ単位の少なくとも一部を有する請求項 1 3 に記載の方法。

10

【請求項 1 5】

各ブロックは、前記左分割線、前記右分割線、前記上分割線、および前記下分割線の間に存在する 1 つ以上のデータ単位を有する中央ゾーンを有し、前記中央ゾーンはどの隣接ブロックとも関連付けられていない請求項 1 4 に記載の方法。

【請求項 1 6】

各ブロックは、

前記左境界線、前記上境界線、前記上分割線および前記左分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない左上ゾーンと、

前記左境界線、前記下境界線、前記下分割線および前記左分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない左下ゾーンと、

20

前記右境界線、前記上境界線、前記上分割線および前記右分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない右上ゾーンと、

前記右境界線、前記下境界線、前記下分割線および前記右分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない右下ゾーンと、のうちの少なくとも 1 つを有する請求項 1 4 に記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、キャッシュメモリを制御するための方法ならびに装置に関し、より詳細にはキャッシュメモリにデータをプリフェッチするための方法ならびに装置に関する。

30

【背景技術】

【0002】

近年、最先端のコンピュータアプリケーションがますます複雑になっているため、コンピュータによるデータ処理のスループットの向上への要求がとどまることを知らず、マイクロプロセッシングシステムに対する要求が増加の一途をたどっている。とりわけ、グラフィックアプリケーションは、所望の視覚的效果を得るため、比較的短時間に膨大な数のデータアクセス、データ演算およびデータ操作を実行する必要があるため、マイクロプロセッシングシステムに対する要求が高いシステムの 1 つとなっている。従来のマイクロプロセッサは、サイクルタイム（マイクロプロセッサがデータを操作することができる時間の単位）が例えば 2 ナノ秒と非常に短い、メインメモリに記憶されているデータへのアクセスに必要な時間は、マイクロプロセッサのサイクルタイムに比べてかなり長い。例えば、ダイナミックランダムアクセスメモリ（DRAM）技術を利用して実装したメインメモリから 1 バイトのデータを取り出す際のアクセス時間は約 60 ナノ秒代である。

40

【発明の開示】

【発明が解決しようとする課題】

【0003】

DRAM メモリのアクセス時間が比較的長いことによるボトルネックを改善するために、当業者はキャッシュメモリを利用してきた。キャッシュメモリは、DRAM メモリよりも格段に高速であり、メインメモリが備えるデータ記憶機能を強化する。例えば、L2 キ

50

キャッシュメモリをマイクロプロセッサの外部に結合したり、L1キャッシュメモリをマイクロプロセッサの内部に結合したりすることができ、これらのメモリはDRAM技術を利用して実装したメインメモリよりも格段に高速である。L2キャッシュメモリは、例えばスタティックランダムアクセスメモリ（SRAM）技術を利用して実装することができ、DRAM技術よりも約2～3倍高速である。L1キャッシュメモリは、L2キャッシュメモリよりもさらに高速である。

【0004】

キャッシュメモリは比較的高価であるため、通常はメインメモリよりも容量が大幅に小さい。このため、キャッシュメモリに記憶すべきデータを決定するために、従来のアルゴリズムが使用されている。例えば、このような従来のアルゴリズムは、「参照の局所性」の理論の概念に基づいており、この概念は、ある特定の時点において、マイクロプロセッサが膨大な実行可能プログラムの比較的限られた部分しか使用していないという事実を利用するものである。このため、参照の局所性の概念に従って、ある特定の時点において、全実行可能プログラムのうちの限られた部分のみがキャッシュメモリに記憶される。

10

【0005】

キャッシュメモリへのデータの記憶を制御するための、参照の局所性を利用する公知のアルゴリズムやその他の概念の詳細については、その数が多いためにここに記載することはできない。ただし、アプリケーションによってデータ処理の目的が大きく異なることがあるため、あらゆるアプリケーションに適するアルゴリズムは存在しないということを言及すれば十分であろう。

20

【0006】

グラフィックアプリケーションでは、例えば、メインメモリ内で記憶されているデータに強い局所性がみられることがあり、この局所性は、メインメモリ内のピクセルデータの座標の関数となっている。詳しく述べると、グラフィックアプリケーションに使用するメインメモリは、概念上、二次元平面に編成され得、画像内の各ピクセルに関する情報は、特定のXY座標に記憶される。グラフィックアプリケーションに使用するメインメモリは、さらに複数の矩形ブロックに編成され得、それぞれのブロックに、複数のピクセルに関する情報が格納される。特定のブロック内のピクセルは極めて線形的にアドレスを付与され得る。換言すれば、特定の矩形ブロック内で、ピクセル情報へのアクセスが極度に連続していることがある。

30

【0007】

このため、グラフィックアプリケーションでキャッシュメモリを制御するための従来のアルゴリズムは、通常、矩形ブロックの全体のピクセル情報を特定のキャッシュラインに記憶させる。キャッシュヒットが発生した場合（すなわち、特定のデータバイトに対するメモリアクセス要求が、キャッシュメモリにアクセスすることで満たされた場合）、その特定の矩形ブロック内のピクセル情報の線形性が高いため、次に行われるデータアクセス要求がキャッシュヒットとなる可能性が高い。一方、キャッシュミスが発生した場合（すなわち、キャッシュメモリにアクセスしても、特定のデータバイトに対するメモリアクセス要求が満たされない場合）、要求されたデータバイトが存在する矩形ブロック内の全ピクセル情報をメインメモリから読み出して、キャッシュメモリに記憶する。しかし、この従来のアルゴリズムは、グラフィックアプリケーションで使用するメモリの線形性と高い局所性とに付随する特性を有効に利用していない。実際、このようなアルゴリズムでは、かなりのキャッシュミスが発生し、この結果、メモリ制御の効率低下、処理のスループットの低下、およびグラフィックの画質の劣化が生じる。

40

【0008】

したがって、処理のスループットを向上させ、グラフィックの画質を向上させるために、グラフィックアプリケーションで使用するメモリにみられる線形性および高い局所性を利用するような、キャッシュメモリを含むメモリ全般を制御するための新しい方法ならびに装置が当業界で求められている。

【課題を解決するための手段】

50

【 0 0 0 9 】

本発明の少なくとも1つの態様によると、メモリ制御装置は、複数の二次元のブロックに分割されているメインメモリと、各ブロックは1つ以上のデータバイトを有する複数のデータ単位を有し、各所定のブロックは複数のゾーンに分割されており、各ゾーンは前記データ単位を1つ以上有し、前記ゾーンの少なくとも一部は、前記所定のブロックに隣接する対応するブロックに関連付けられており、前記メインメモリの前記データバイトに対するアドレスを受け取ると共に、前記アドレスをデコードして、対応するデータバイトが存在するゾーンを特定するように動作可能なアドレスデコーダと、前記アドレスデコーダが対応するゾーンに存在するデータバイトのアドレスを受け取ると、前記対応するゾーンに関連付けられているブロックをキャッシュメモリにプリフェッチするように動作可能なプリフェッチアドレスジェネレータと、を有する。

10

例えば、前記各ブロックは、左の隣接ブロックと共通の左境界線、右の隣接ブロックと共通の右境界線、上の隣接ブロックと共通の上境界線、および下の隣接ブロックと共通の下境界線のうちの少なくとも1つを有し得る。さらに、前記各ブロックは、前記左の隣接ブロックに関連付けられており、前記左境界線に沿って配置されたデータ単位を少なくとも1つ有する左ゾーン、前記右の隣接ブロックに関連付けられており、前記右境界線に沿って配置されたデータ単位を少なくとも1つ有する右ゾーン、前記上の隣接ブロックに関連付けられており、前記上境界線に沿って配置されたデータ単位を少なくとも1つ有する上ゾーン、および前記下の隣接ブロックに関連付けられており、前記下境界線に沿って配置されたデータ単位を少なくとも1つ有する下ゾーン少なくとも1つを有し得る。

20

【 0 0 1 0 】

このような分割を使用して、前記左ゾーンは前記上境界線から前記下境界線まで延びる左分割線によって少なくとも一部画定されてもよく、前記左ゾーンは前記左境界線と前記左分割線との間に存在するデータ単位の少なくとも一部を有してもよく、前記右ゾーンは前記上境界線から前記下境界線まで延びる右境界線によって少なくとも一部画定されてもよく、前記右ゾーンは前記右境界線と前記右分割線との間に存在するデータ単位の少なくとも一部を有してもよく、前記上ゾーンは前記左境界線から前記右境界線まで延びる上境界線によって少なくとも一部画定されてもよく、前記上ゾーンは前記上境界線と前記上分割線との間に存在するデータ単位の少なくとも一部を有してもよく、前記下ゾーンは前記左境界線から前記右境界線まで延びる下境界線によって少なくとも一部画定されてもよく、前記下ゾーンは前記下境界線と前記下分割線との間に存在するデータ単位の少なくとも一部を有してもよい。

30

【 0 0 1 1 】

好ましくは、前記左分割線、前記右分割線、前記上分割線、および前記下分割線は前記アドレスの特定のビットに対応付けられており、前記アドレスデコーダは、前記アドレスをデコードして、前記特定のビットに基づいて前記対応するデータバイトが存在するゾーンを特定するように動作可能である。

【 0 0 1 2 】

例えば、各ブロックは、前記左分割線、前記右分割線、前記上分割線、および前記下分割線の間に存在する1つ以上のデータ単位を有する中央ゾーンを有してもよく、前記中央ゾーンは好ましくはどの隣接ブロックとも関連付けられていない。各ブロックは、好ましくは、前記左境界線、前記上境界線、前記上分割線および前記左分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない左上ゾーンと、前記左境界線、前記下境界線、前記下分割線および前記左分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない左下ゾーンと、前記右境界線、前記上境界線、前記上分割線および前記右分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない右上ゾーンと、前記右境界線、前記下境界線、前記下分割線および前記右分割線の間に存在するデータ単位を有し、どの隣接ブロックとも関連付けられていない右下ゾーンと、のうちの少なくとも1つをさらに有する。

40

【 0 0 1 3 】

50

本発明の少なくとも１つの別の態様に係る方法は、メインメモリを複数の二次元のブロックに分割すると共に、各ブロックをデータバイトを１つ以上含む複数のデータ単位に分割するステップと、前記メインメモリの各ブロックを、前記データ単位を１つ以上有する複数のゾーンに分割するステップと、各ブロックの対応するゾーンの少なくとも一部を前記ブロックのうちの別のブロックに関連付けるステップと、前記別のブロックのうちの１つのブロックに関連付けられているゾーンのいずれかにあるデータ単位のいずれかがアドレス指定された場合に、そのブロックの少なくとも一部をキャッシュメモリにプリフェッチするステップと、を有する。

【００１４】

好ましくは、各ブロックの前記少なくとも一部のゾーンのそれぞれは１つの隣接ブロックに関連付けられている。さらに、各ブロックの少なくとも１つのゾーンは、好ましくはそのブロックと、関連付けられている隣接ブロックとの間の境界線に沿って配置されている。

【００１５】

添付の図面を参照しつつ、ここに記載する本発明の説明を読めば、他の態様、特徴および利点等は当業者に自明となるであろう。

【発明を実施するための最良の形態】

【００１６】

本発明を説明するために、現在の好ましい形態を図面の形で示すが、本発明は、図示したとおりの構成ならびに手段に限定されないことを理解されたい。

【００１７】

図面において、同じ参照符号は、同一の要素を示している。図１に、本発明による、メモリを制御するために適した方法および／または装置の態様を示す。簡潔を期すと共に明確になるように、図１のブロック図は装置１００の説明として本明細書において参照かつ記載される。しかし、この記載は同等の効力を有する方法のさまざまな態様に容易に適用できることを理解されたい。装置１００は、好ましくは、プロセッサ１０２、アドレスレジスタ１０４、アドレスデコーダ１０６、プリフェッチアドレスジェネレータ１０８、メモリコントローラ１１０、メインメモリ１１２、およびキャッシュメモリ１１４を備える。

【００１８】

プロセッサ１０２は、メインメモリおよび／またはキャッシュメモリ１１４からデータを要求し、このデータを操作して所望の結果を得ることができる公知の技術のいずれかを用いて実装することができる。例えば、プロセッサ１０２は、通常のマイクロプロセッサ、分散型マイクロプロセッサなど、ソフトウェアおよび／またはファームウェアを実行することができる公知のマイクロプロセッサのいずれかを利用して実装することができる。例えば、プロセッサ１０２は、グレースケール情報、色情報、テクスチャデータ、ポリゴン情報、ビデオフレーム情報等を含むピクセルデータなどのデータを要求して操作することができるグラフィックプロセッサであり得る。

【００１９】

プロセッサ１０２は、好ましくは、メモリアドレスなどのデータアクセス要求を、アドレスレジスタ１０４とメモリコントローラ１１０に対して与える。アドレスデコーダ１０６とプリフェッチアドレスジェネレータ１０８は、好ましくは、メインメモリ１１２に対してキャッシュメモリ１１４に記憶する情報の管理上望ましい結果を得るために、メモリコントローラ１１０にさらに情報を与えるように動作可能である。装置１００の各種機能ブロックのこの相互の関係をさらに詳しく説明するため、図２～５を参照する。

【００２０】

図２は、図１の機能ブロックに従って実行および／または実装され得る特定のアクション／機能を示すフローチャートである。アクション２００で、メインメモリ１１２（図１）が好ましくは複数の二次元のデータブロックに分割される。メインメモリ１１２が必ずしも物理的に分割されるわけではないという点で、メインメモリ１１２の分割は主として

10

20

30

40

50

概念的なものであり、メインメモリ 112 は、装置 100 のさまざまな機能ブロックが実行するデータの記憶およびデータへのアクセスに対して機能的に分割されるという点を理解されたい。

【0021】

より詳細には、図 3 を参照すると、メインメモリ 112 は、好ましくは、DRAM などの、データ記憶位置 112A を複数有する公知の技術のいずれかを利用して実装することができる。説明のために、各データ位置 112A は 1 データバイトを表し、各バイトは、例えば、8 ビット、16 ビット、32 ビット、64 ビット等を有するとする。データ記憶位置 112A は、16 進アドレス、すなわち 0000、0001、0002、...003F、...023F 等で示されている。この例では、図 3 のメインメモリ 112 は、1,024 のデータ記憶位置 112A を有し、それぞれの位置 112A に 1 データバイトを記憶することができるとする。このメインメモリ 112 のサイズは例に過ぎず、本発明の範囲を限定することを意図するものではないことを理解されたい。

【0022】

図 2 のアクション 200 に説明を戻すと、メインメモリ 112 は二次元のデータブロック 120 に分割されており、さらに図 3 を参照すると、メインメモリ 112 のデータ記憶位置 112A は、好ましくはデータブロック 120 に編成されており、例えばアドレス 0000 ~ 003F に記憶されている全ピクセル情報（すなわちデータバイト）を格納している。説明のために、画像の 1 ピクセルを表すための情報が、4 データバイト（例えば、グレースケール情報に 1 バイト、色情報に 1 バイトなど）を必要とするとさらに想定する。このため、この例では、図 3 のメモリ 112 は、144 個のピクセルに関する情報を記憶することができる。各データブロック 120 は、メインメモリ 112 に記憶されている 144 個のピクセルからなるサブセットに関する情報を格納している。各データブロック 120 は、好ましくは二次元の矩形ブロックであり、それぞれのピクセルのサブセットに関する情報が X Y デカルト座標系に配置される。

【0023】

図 2, 3 を参照すると、アクション 202（図 2）で、各データブロック 120 が好ましくは複数のデータ単位 122（図 3）に分割され、このとき、各データ単位 122 はピクセルのそれぞれ 1 つに関する情報を格納している。上に挙げた例では、図 3 の各データ単位 122 は 4 データバイトを有する。本発明に従ってデータ単位 122 を編成する方法は数多くあるが、ジグザグパターンで直線状に並ぶように、データ単位 122 を左から右、上から下の順で編成することが好ましい。この場合、データバイトのアドレスは、順に（先頭のデータ単位 122 の）0000、0001、0002、0003、（次のデータ単位 122 の）0004、0005、0006、0007、...と続く。データブロック 120 の全体について、後続のデータ単位 122 の後続のデータバイトが、左から右（X 方向）に、上から下に（Y 方向）に並んでいく。

【0024】

図 4 に示すように、メインメモリ 112 内の別のピクセル群に関連する情報は、好ましくは別のデータブロック 120 に分割されており、このデータブロック 120 は二次元の X、Y の格子に配置されている。上に挙げた例では、これにより、中央ブロック（名称表示なし）、左上データブロック、上データブロック、右上データブロック、左データブロック、右データブロック、左下データブロック、下データブロック、および右下データブロックの 9 個のデータブロック 120 が形成されている。左上データブロック 120（すなわち図 3 のデータブロック 120）がアドレス 0000 ~ 003F に記憶されているデータバイトに関連する情報を表すと想定すると、図 4 の中央ブロックは、メインメモリ 112 のアドレス 0100 ~ 013F に記憶されているピクセル情報を格納している。

【0025】

メモリコントローラ 110 に採用されているアルゴリズムによって、キャッシュメモリ 114 内の特定のキャッシュラインに、データブロック 120 全体のピクセル情報が記憶されていると想定すると、プロセッサ 102 が、例えば（図 4 の中央ブロック内の）アド

10

20

30

40

50

レス 0 1 1 4 に記憶されているデータバイトを要求したときに、キャッシュメモリ 1 1 4 が中央ブロックに関連するピクセル情報を格納していなければ、キャッシュミスが発生する。この結果、メモリコントローラ 1 1 0 は好ましくは、メインメモリ 1 1 2 から、アドレス 0 1 0 0 ~ 0 1 3 F に記憶されているピクセル情報を読み込んで、この情報をキャッシュメモリ 1 1 4 に記憶する。

【 0 0 2 6 】

上に挙げた例では、次に行われるデータアクセス要求が、メモリ位置 0 1 1 5 に記憶されているデータバイトに対するものである可能性が高く、このデータバイトは既にキャッシュメモリ 1 1 4 に記憶されているため、キャッシュヒットとなる。さらに、プロセッサ 1 0 2 がアドレス 0 1 1 4 ~ 0 1 1 7 (同じデータ単位 1 2 2 内のアドレス) に記憶されているデータバイトを既に全て要求している場合には、次に行われるデータアクセス要求の対象が、中央データブロック 1 2 0 に存在する可能性もあれば、存在しない可能性もある。実際、次に行われるデータアクセス要求が、隣接するデータブロック 1 2 0 (例えば左上データブロック 1 2 0、上データブロック 1 2 0、右上データブロック 1 2 0 など) に存在するデータバイトに対するものである可能性がある。事実、次に要求されたデータバイトが、キャッシュメモリ 1 1 4 に記憶していない隣接データブロック 1 2 0 に存在する場合にはキャッシュミスが発生する。

【 0 0 2 7 】

本発明は、別のデータブロックにあるデータに対する要求に基づいて、1つ以上の隣接データブロック 1 2 0 をキャッシュメモリ 1 1 4 にプリフェッチすることによって、キャッシュヒット率を向上させる方法ならびに装置を得ることを意図するものである。これを実現するため、アクション 2 0 4 (図 2) で、各データブロック 1 2 0 が好ましくは複数のゾーンに分割され、各ゾーンは1つ以上のデータ単位を含む。

【 0 0 2 8 】

図 5 を参照して例を挙げると、中央データブロック 1 2 0 が、ゾーン 1 ~ ゾーン 9 と記された 9 個のゾーンに分割される。図 5 では、ゾーンに分割されているのは中央データブロック 1 2 0 だけであるが、本発明では、ゾーンに分割するデータブロック 1 2 0 の数は幾つであってもよく、好ましくは、各データブロック 1 2 0 がこのように分割される。この例では、ゾーン 1、3、7、9 はそれぞれデータ単位 1 2 2 を 1 つ含み、ゾーン 2、4、6、8 はそれぞれデータ単位 1 2 2 を 2 つ含み、ゾーン 5 はデータ単位 1 2 2 を 4 つ含む。

【 0 0 2 9 】

アクション 2 0 6 (図 2) で、図 5 の中央データブロック 1 2 0 などの各データブロック 1 2 0 内の少なくとも一部のゾーンのそれぞれが、中央データブロック 1 2 0 に隣接しているデータブロック 1 2 0 に関連付けられる。例えば、本発明の一態様によれば、ゾーン 2 は好ましくは上データブロック 1 2 0 に関連付けられ、ゾーン 4 は好ましくは左データブロック 1 2 0 に関連付けられ、ゾーン 6 は好ましくは右データブロック 1 2 0 に関連付けられ、ゾーン 8 は好ましくは下データブロック 1 2 0 に関連付けられる。中央データブロック 1 2 0 のゾーン 1、3、7、9 は、好ましくはどの隣接データブロック 1 2 0 とも関連付けられない。同様に、中央データブロック 1 2 0 のゾーン 5 は、好ましくはどの隣接データブロック 1 2 0 とも関連付けられない。このように関連付けをする理由、および/または関連付けをしない理由は、本明細書で詳細に後述する。

【 0 0 3 0 】

アクション 2 0 8 (図 2) で、プロセッサ 1 0 2 は、好ましくはアクセス要求 (1 つ以上のデータバイトのアドレスなど) を生成する。上記したように、このデータバイトは、データブロック 1 2 0 のゾーンのデータ単位 1 2 2 に存在する。このアクセス要求は、好ましくはアドレスレジスタ 1 0 4 とメモリコントローラ 1 1 0 に与えられる (アクション 2 0 8)。メモリコントローラ 1 1 0 は、好ましくは、このアクセス要求がキャッシュメモリ 1 1 4 のキャッシュラインに関連付けられている複数のキャッシュタグのうちのいずれか 1 つと一致するかどうかを判定する。キャッシュヒットの場合、すなわち、アクセス

10

20

30

40

50

要求がキャッシュメモリ 114 に記憶されているデータと一致した場合、要求されたデータバイトがキャッシュメモリ 114 から読み出されて、プロセッサ 102（および/または他の任意の適切なデバイスまたはプロセス）に提供される。

【0031】

本発明の 1 つ以上の別の態様よれば、アドレスレジスタ 104 は、好ましくはこのアクセス要求（すなわち要求されたデータバイトのアドレス）をアドレスデコーダ 106 に与える。アドレスデコーダ 106 は、好ましくはこのアクセス要求をデコード、すなわち要求されたデータバイトのアドレスをデコードして、そのデータバイトが存在するゾーンを特定するように動作可能である（アクション 210）。例えば、アドレスデコーダ 106 は、要求されたデータバイトが存在するゾーンを特定する際に、好ましくはそのアドレスの特定のビットを分析するように動作可能である。この点について、図 6 を参照してさらに詳細に説明する。中央データブロック 120 は、9 個のゾーンに分割されており、これらのゾーンは、左境界線 130（隣接する左データブロック 120 と共通）、右境界線 132（隣接する右データブロック 120 と共通）、上境界線 134（隣接する上データブロック 120 と共通）、および下境界線 136（隣接する下データブロック 120 と共通）によって画定されている。これらのゾーンは、上境界線 134 から下境界線 136 まで延びる左分割線 140、同様に上境界線 134 から下境界線 136 まで延びる右境界線、左境界線 130 から右境界線 132 まで延びる上分割線 144、および同様に左境界線 130 から右境界線 132 まで延びる下分割線 146 によってさらに画定されている。

【0032】

この境界線および分割線の例によると、ゾーン 1（左上ゾーン）は、左境界線 130、上境界線 134、上分割線 144、および左分割線 140 の間に存在する 1 つのデータ単位 122 を有する。ゾーン 2（上ゾーン）は、上境界線 134 と上分割線 144 との間に存在する 2 つのデータ単位 122 を有する。ゾーン 3（右上ゾーン）は、右境界線 132、上境界線 134、上分割線 144、および右分割線 142 の間に存在する 1 つのデータ単位 122 を有する。ゾーン 4（左ゾーン）は、左境界線 130 と左分割線 140 との間に存在する 2 つのデータ単位 122 を有する。ゾーン 5（中央ゾーン）は、左分割線 140、右分割線 142、上分割線 144、および下分割線 146 の間に存在する 4 つのデータ単位 122 を有する。ゾーン 6（右ゾーン）は、右境界線 132 と右分割線 142 との間に存在する 2 つのデータ単位 122 を有する。ゾーン 7（左下ゾーン）は、左境界線 130、下境界線 136、下分割線 146、および左分割線 140 の間に存在する 1 つのデータ単位 122 を有する。ゾーン 8（下ゾーン）は下境界線 136 と下分割線 146 との間に存在する 2 つのデータ単位 122 を有する。最後に、ゾーン 9（右下ゾーン）は、右境界線 132、下境界線 136、下分割線 146、および右分割線 142 の間に存在する 1 つのデータ単位 122 を有する。

【0033】

上述の例では、左分割線 140、右分割線 142、上分割線 144、および下分割線 146 は、中央データブロック 120（さらには任意のデータブロック 120）に記憶されているデータバイトのアドレスの特定のビットによって定義することができる。実際、左分割線 140 と右分割線 142 は、アクセス要求アドレスの第 3 ビットと第 2 ビットによって定義することができる（例えば、これらからデコードまたは確認されても、この両方が行われてもよい）。同様に、上分割線 144 と下分割線 146 は、アクセス要求アドレスの第 5 ビットと第 4 ビットによって定義される。当業者であれば、それぞれのゾーンの境界を定義またはデコードするために使用するアドレスの特定のビットは、データブロック 120 のサイズ、データ単位 122 のサイズ、バイトのサイズ、および境界線と分割線の位置や経路によって決まることを理解されよう。

【0034】

上述の例では、アドレスデコーダ 106 は、好ましくはアクセス要求アドレスのビット 5、4、3、2 を分析して、要求されたデータバイトが存在するゾーンを特定する（図 2 のアクション 210）。アクション 212 で、特定されたゾーンが隣接データブロック 1

20に関連付けられているかどうかを判定する。プリフェッチアドレスジェネレータ108は、好ましくはアクション206で定義した関連付けに基づいてこの判定を行うように動作可能である。例えば、アクセス要求アドレスが011Cの場合(図5)、アドレスデコード106は、好ましくはアクセス要求アドレスの上記のビットを分析して、要求されたデータバイトがゾーン6に存在することを特定する。プリフェッチアドレスジェネレータ108は、好ましくはこの情報を受け取って、ゾーン6が隣接データブロック120に関連付けられているかどうかを判定する。上に挙げた例では、ゾーン6は、右データブロック120に関連付けられている。これにより、プリフェッチアドレスジェネレータ108は、好ましくは、メインメモリ112から右データブロック120をキャッシュメモリ114にプリフェッチするよう求める要求を生成する(アクション214)。アドレスデコード106によって特定されたゾーンが隣接データブロック120に関連付けられていない場合(アクション212)、手順フローは好ましくはアクション208に分岐して、プロセッサ102から次のアクセス要求を受け取る。

【0035】

図5を参照すると、ゾーン2、ゾーン4、ゾーン6、およびゾーン8は、ある一定の仮定に基づいて、上データブロック120、左データブロック120、右データブロック120、および下データブロック120にそれぞれ関連付けされた点に留意されたい。例えば、現在のアクセス要求が中央データブロック120のゾーン2に存在するデータバイトに対するものである場合、次のアクセス要求は、上データブロック120に対して行われる可能性が高いと仮定した。ゾーン4と左データブロック120、ゾーン6と右データブロック120、およびゾーン8と下データブロック120の間の関係についても同じように仮定した。

【0036】

さらに、現在のアクセス要求がゾーン5に存在するデータバイトに対するものである場合には、次のアクセス要求が、中央データブロック12以外に対して行われる可能性は低いと仮定した。ゾーン1、ゾーン3、ゾーン7、およびゾーン9については、次のアクセス要求が隣接データブロック120に対するものである可能性があるものの、この隣接データブロック120が左データブロックである可能性、左上データブロックである可能性、および上データブロックである可能性は同程度であると仮定した。このため、例えばゾーン1とこれらの隣接データブロック120のうちの特定の1つとを関連付けたからといって、キャッシュヒット率が必ずしも向上するわけでない。当然、状況の緊急度が高い場合には、ゾーン1、ゾーン3、ゾーン7、および/またはゾーン9と、隣接データブロック120の1つ以上とを関連付けることが有用なこともあり、このため、本発明のさらに別の態様に従ってこの点が考察される。

【0037】

図5, 図6に示すゾーン構成は、本発明に必須というわけではない。図7A, 7Bに別法によるゾーン構成を示す。これらのゾーン構成についても例に過ぎず、本発明の範囲を限定することを意図するものではない。

【0038】

本発明の少なくとも1つの別の態様によれば、メインメモリ112とキャッシュメモリ114とを制御するための方法ならびに装置は、図1に示すような適切なハードウェアを用いて実現することができる。この種のハードウェアは、公知の技術のいずれかを用いて実装することができる。この例には、標準的なデジタル回路、ソフトウェアプログラムおよび/またはファームウェアプログラムを実行するように動作可能な公知のプロセッサのいずれか、プログラマブル読出し専用メモリ(PROM)、プログラマブルアレイロジックデバイス(PAL)などの、1つ以上のプログラム可能なデジタル装置またはシステムなどがある。さらに、図1に示した装置100は、特定の機能ブロックに分けて図示しているが、これらのブロックは、別個の回路によって実施しても、1つ以上の機能単位を結合しても、この両方を行ってもよい。さらに、本発明は、持ち運んだり配布したり、あるいはこの両方を行えるように、適切な記憶媒体(フロッピーディスク、メモリチップ

10

20

30

40

50

など)に記憶可能なソフトウェアプログラムおよび/またはファームウェアプログラムによって実装されてもよい。

【0039】

本明細書に記載し、特許を請求するメモリ制御のための方法ならびに装置は、例えばグラフィックアプリケーションで使用されるメモリの線形性と高い局所性との付随する特性を有利に利用する。キャッシュヒットが大幅に向上し、このため、メモリ効率、処理のスループット、およびグラフィックの画質の向上が達成される。

【0040】

本明細書において、具体的な実施形態を用いて本発明を記載したが、これらの実施形態は、本発明の原理および応用の例を示すものに過ぎないことを理解されたい。このため、添付の請求の範囲に記載した本発明の趣旨および範囲から逸脱することなく、これら例示的な実施形態を種々に変更したり、上記以外の構成を考案し得ることが理解されよう。

【図面の簡単な説明】

【0041】

【図1】本発明の1つ以上態様に従ってメモリを制御するために適した装置の態様(および/または方法の態様)を示すブロック図である。

【図2】図1に示したような本発明の1つ以上の態様に従って実行および/または実装され得る特定のアクション/機能を示すフローチャートである。

【図3】図1に示した装置(および/または方法の態様)と共に使用され得るメモリの特定の特徴を示す概念ブロック図である。

【図4】図3のメモリの概念を詳細に示す別の概念ブロック図である。

【図5】図3のメモリの概念を詳細に示すさらに別の概念ブロック図である。

【図6】図3のメモリのデータブロックの特定の態様を詳細に示す概念ブロック図である。

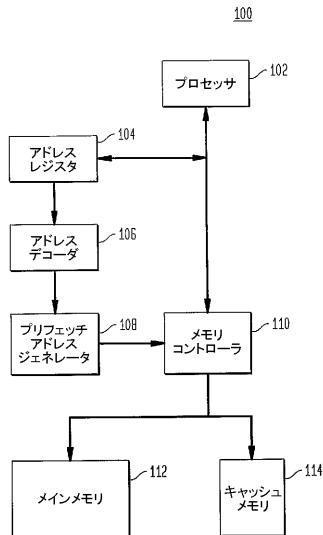
【図7A】本発明の1つ以上の別の態様による、メモリの制御および/または実装に関し、図5に示した概念の別例を詳細に示す概念ブロック図である。

【図7B】本発明の1つ以上の別の態様による、メモリの制御および/または実装に関し、図5に示した概念の別例を詳細に示す概念ブロック図である。

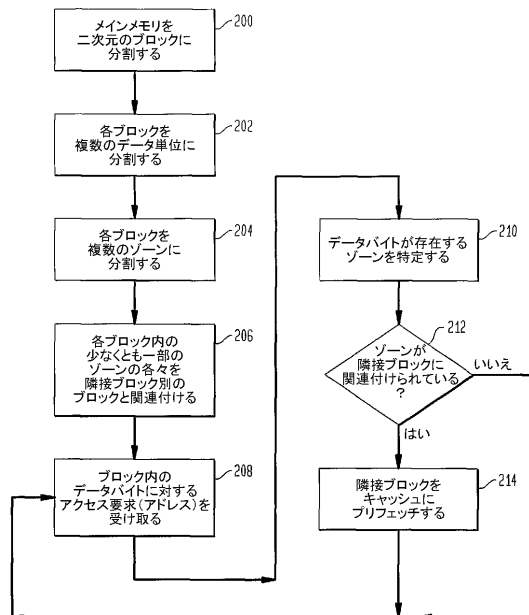
10

20

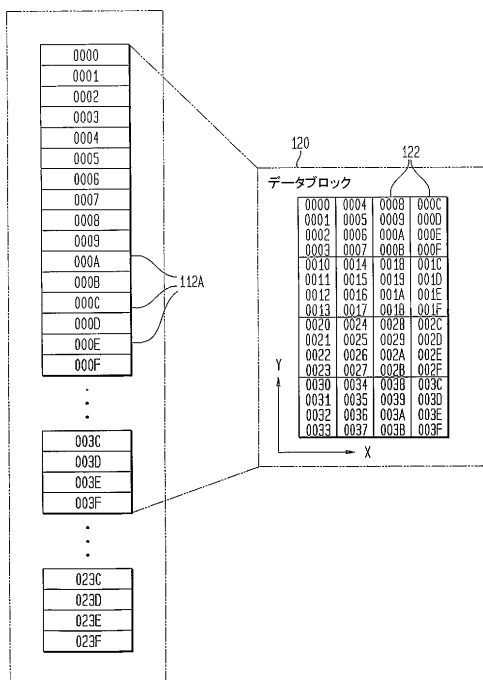
【図 1】



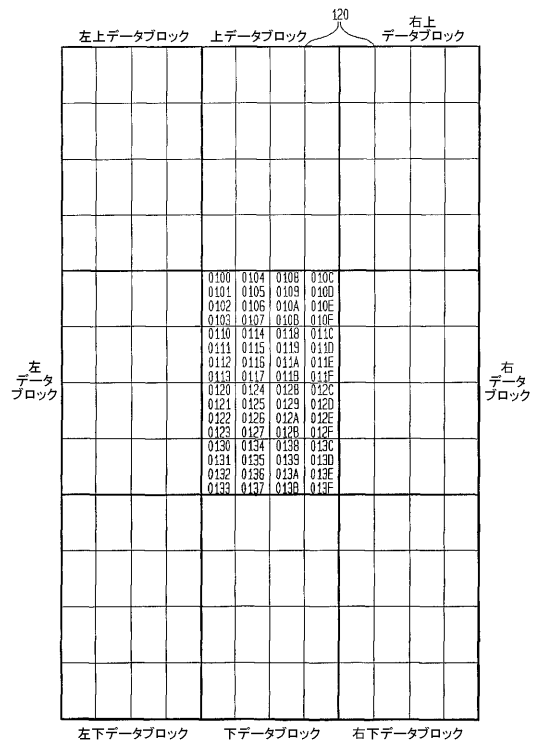
【図 2】



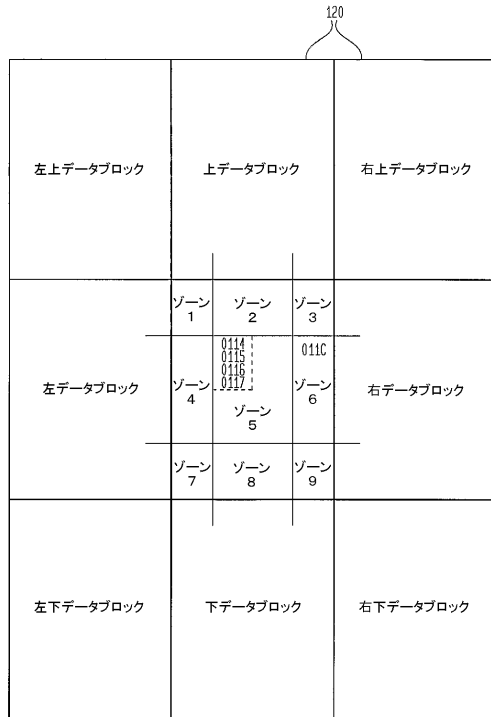
【図 3】



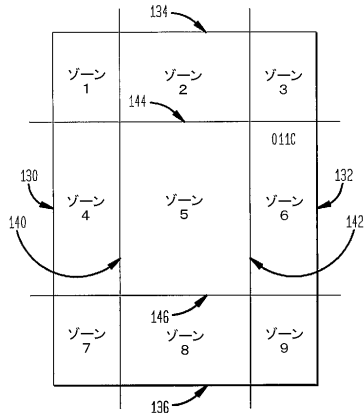
【図 4】



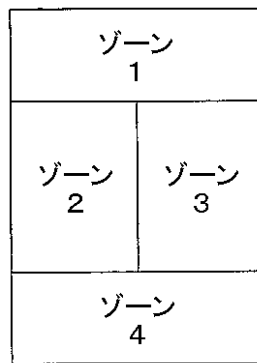
【図 5】



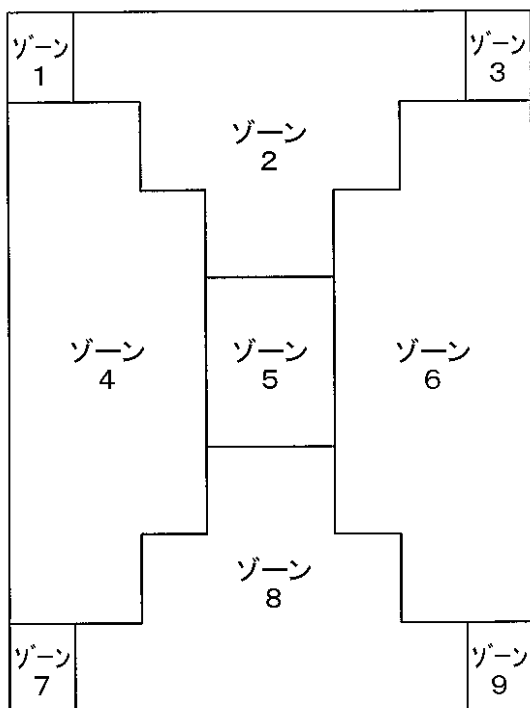
【図 6】



【図 7 A】



【図 7 B】



フロントページの続き

(72)発明者 馬越 英尚

アメリカ合衆国、カリフォルニア州 94404、フォスター シティ、セカンド フロア、イ
ースト ヒルスデイル ブルバード 919 ソニー・コンピュータ・エンタテインメント・アメ
リカ・インク内

審査官 清木 泰

(56)参考文献 特開平03 - 154977 (JP, A)

特開平05 - 053909 (JP, A)

特開平11 - 215509 (JP, A)

特開平09 - 050399 (JP, A)

特開平06 - 180669 (JP, A)

特開平04 - 153753 (JP, A)

特開平02 - 091776 (JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F12/08-12/12

G06T 1/60

H04N 1/00-17/00

G09G 1/00- 5/399