

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局



(10) 国際公開番号

WO 2010/052753 A1

PCT

(43) 国際公開日
2010 年 5 月 14 日(14.05.2010)

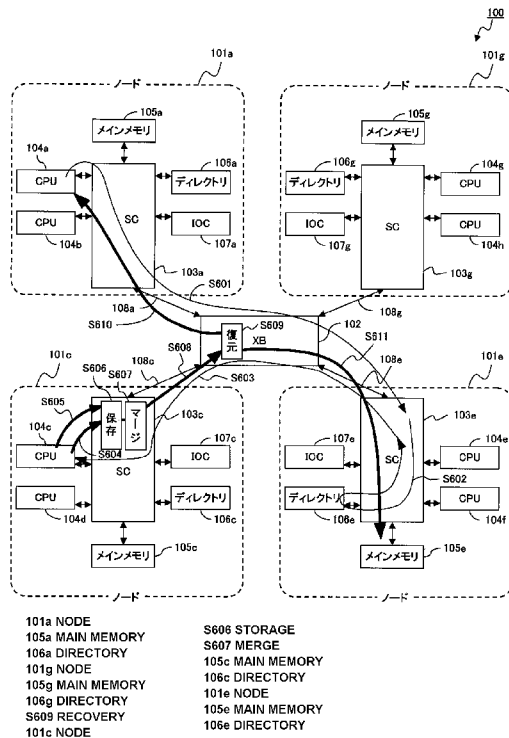
- (51) 国際特許分類:
G06F 15/173 (2006.01) G06F 12/08 (2006.01)
- (21) 国際出願番号: PCT/JP2008/003216
- (22) 国際出願日: 2008 年 11 月 6 日(06.11.2008)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (71) 出願人 (米国を除く全ての指定国について): 富士通株式会社 (FUJITSU LIMITED) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 Kanagawa (JP).
- (72) 発明者; および
- (75) 発明者/出願人 (米国についてのみ): 北原義博 (KITAHARA, Yoshihiro) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 富士通株式会社内 Kanagawa (JP).
- (74) 代理人: 大菅義之 (OSUGA, Yoshiyuki); 〒1020084 東京都千代田区二番町 8 番地 2 0 二番町ビル 3 F Tokyo (JP).
- (81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO,

[続葉有]

(54) Title: CONTROLLER, DATA TRANSFER UNIT, INFORMATION PROCESSOR, ARITHMETIC PROCESSING UNIT, AND INFORMATION PROCESSOR CONTROL METHOD

(54) 発明の名称: 制御装置、データ転送装置、情報処理装置、演算処理装置および情報処理装置の制御方法

[図14]



(57) Abstract: When an arithmetic processing unit outputs first data and second data to a controller, the controller compares the control information of the first data and that of the second data with each other to judge whether the first and second data have the same data main body section. The controller has a storage part for temporarily storing data inputted from the arithmetic processing unit until the data is outputted to a data transfer unit. When the first and second data have the same data main body section, the controller stores the first data in the storage part and stores a part of the control information on the second data. In this case, the controller may discard the second data. Furthermore, the controller may generate the third data from the first data stored in the storage part and the control information on the second data and output the third data to the data transfer unit

(57) 要約:

[続葉有]



PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, 添付公開書類:

CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

— 国際調査報告 (条約第 21 条(3))

演算処理装置が制御装置に第 1 のデータと第 2 のデータを出力すると、制御装置は、第 1 のデータと第 2 のデータの制御情報同士を比較することで、第 1 のデータと第 2 のデータが同じデータ本体部を有するか否かを判断する。また、制御装置は、演算処理装置から入力されたデータをデータ転送装置に出力するまで一時的に保存する保存部を有する。第 1 のデータと第 2 のデータが同じデータ本体部を有する場合、制御装置は、第 1 のデータを保存部に保存し、第 2 のデータの制御情報の一部を保存する。この場合、制御装置は第 2 のデータを破棄してもよい。また、制御装置は、保存部に保存した第 1 のデータと、第 2 のデータの制御情報から、第 3 のデータを生成し、第 3 のデータをデータ転送装置に出力してもよい。

明 細 書

制御装置、データ転送装置、情報処理装置、演算処理装置および情報処理装置の制御方法

技術分野

[0001] 本発明は、複数の演算処理装置を用いた情報処理技術に関する。

背景技術

[0002] 近年、マルチCPU（Central Processing Unit）システム（マルチプロセッサシステムともいう）の利用が一般化している。例えば、ノード間を相互に接続するノード相互接続スイッチにより結合されたノードを2つ以上有するNUMA（Non-Uniform Memory Access；不均等メモリアクセス）コンピュータシステムなどが開示されている。

[0003] 上記NUMAコンピュータシステムにおける各ノードは同一でもよい。各ノードは、ローカル相互接続に結合される処理ユニットとノード相互接続スイッチとの間に結合される。ノード制御装置は、ローカル相互接続上で受信される選択命令を、ノード相互接続スイッチを介して他のノードに伝送することにより、受信された選択命令をローカル相互接続上に伝送して、他のノードのためのローカル・エージェントとして機能する。

[0004] また、近年では多くのコンピュータシステムにおいてCPUがキャッシュメモリを備えている。そして、複数の装置が主記憶装置を共有する場合にキャッシュメモリ間のデータの整合性であるキャッシュコンシステンシ（Cache Consistency）（キャッシュコヒーレンシ（Cache Coherency）ともいう）を保つための様々な技術が知られている。

特許文献1：特開2003-44455号公報

特許文献2：特開2002-342162号公報

特許文献3：特開平9-237223号公報

発明の開示

[0005] 本発明は、同じデータ本体部を有する複数のデータが出力される場合に着

目して、データを保存するのに必要な容量を削減することを目的とする。

開示の技術の第 1 の態様によれば、演算処理装置およびデータ転送装置の間でのデータの転送を制御し、保存手段と判断手段と管理手段と制御手段を備える制御装置が提供される。

[0006] 前記保存手段は、前記演算処理装置から出力された第 1 のデータを保存する。

前記判断手段は、前記第 1 のデータの制御情報と、前記第 1 のデータより後に前記演算処理装置から受け取った第 2 のデータの制御情報を比較することで、前記第 1 のデータと前記第 2 のデータが同じデータ本体部を有する第 1 の場合であるか否かを判断する。

[0007] 前記管理手段は、前記第 1 の場合は、前記第 2 のデータの前記制御情報の一部を前記第 1 のデータと関連づけた複製情報として保存する。また、前記管理手段は、前記第 1 の場合以外である第 2 の場合は、前記保存手段に前記第 2 のデータを保存させる。

[0008] 前記制御手段は、前記保存手段に保存されたデータの前記データ転送装置への出力を制御する。

開示の技術の第 2 の態様によれば、第 1 の装置と第 2 の装置を含む複数の装置と接続され、前記複数の装置間でのデータの転送を行うデータ転送装置が提供される。前記データ転送装置は、送信手段と保存手段と判断手段と管理手段と制御手段を備える。

[0009] 前記送信手段は、前記複数の装置のうちの 1 つから出力されたデータを前記複数の装置のうちの他の 1 つへ送信する。

前記保存手段は、前記第 1 の装置から受け取った、前記第 2 の装置へ送信する第 1 のデータを前記第 2 の装置へ出力する前に保存しておく。

[0010] 前記判断手段は、前記第 1 の装置から受け取り、前記第 1 のデータより後に前記送信手段により送信した、前記第 2 の装置へ送信する第 2 のデータの制御情報を、前記第 1 のデータの制御情報と比較する。その比較により、前記判断手段は、前記第 1 のデータと前記第 2 のデータが同じデータ本体部を

有する第 1 の場合であるか否かを判断する。

[0011] 前記管理手段は、前記第 1 の場合は、前記第 2 のデータの前記制御情報の一部を前記第 1 のデータと関連づけた複製情報として保存する。また、前記管理手段は、前記第 1 の場合以外である第 2 の場合は、前記保存手段に前記第 2 のデータを保存させる。

[0012] 前記制御手段は、前記保存手段に保存されたデータの前記第 2 の装置への出力を制御する。

開示の技術の第 3 の態様によれば、複数の制御装置と、前記複数の制御装置間でのデータの転送を行うデータ転送装置を備える情報処理装置が提供される。複数の制御装置には、下記の生成手段をさらに備えた第 1 の態様と同様の第 1 の制御装置が含まれる。

[0013] 前記生成手段は、第 1 の態様に関して説明した前記第 1 の場合に、前記保存手段が保存する前記第 1 のデータと前記管理手段が保存する前記複製情報から、第 3 のデータを生成する。

[0014] そして、第 3 の態様において前記第 1 の制御装置の前記制御手段は、前記第 1 の場合には前記生成手段が生成した前記第 3 のデータを前記データ転送装置に出力するよう制御する。また前記第 1 の制御装置の前記制御手段は、前記第 2 の場合には前記保存手段が保存する前記第 1 のデータと前記第 2 のデータの双方を前記データ転送装置に出力するよう制御する。

[0015] また、第 3 の態様における前記データ転送装置は、前記第 1 の制御装置から受け取ったデータの制御情報に基づいて、前記第 1 の態様に関して説明した前記第 1 の場合か前記第 2 の場合かを判断する第 2 の判断手段を備える。

[0016] 前記データ転送装置はさらに、前記第 1 の場合に、前記第 3 のデータから前記第 1 のデータと前記第 2 のデータを復元する復元手段を備える。そして、前記データ転送装置は、前記第 2 の判断手段による判断に基づいて、前記第 3 のデータを送信するか、または前記復元手段が復元した前記第 1 のデータと前記第 2 のデータを送信する送信手段を備える。

[0017] 開示の技術の第 4 の態様によれば、記憶部と演算処理部と保存手段と判断

手段と管理手段と生成手段と制御手段を備えた演算処理装置が提供される。

第４の態様における前記保存手段、前記判断手段、前記管理手段、および前記制御手段は、第１の態様における前記制御装置のものと類似である。また、第４の態様における前記生成手段は、第３の態様における前記制御装置のものと類似である。

[0018] 開示の技術の第５の態様によれば、第３の態様の情報処理装置の制御方法が提供される。

したがって、開示の技術によれば、いずれの態様においても第１の場合に第２のデータを保存する必要がないため、保存手段に必要な記憶容量を削減することが可能である。

図面の簡単な説明

[0019] [図１]複数のＣＰＵを備えるシステムの例を示す構成図である。

[図２]第１のキャッシュプロトコルにしたがったサーバシステムの動作例を示す図である。

[図３]第２のキャッシュプロトコルにしたがったサーバシステムの動作例を示す図である。

[図４]第３のキャッシュプロトコルにしたがったサーバシステムの動作例を示す図である。

[図５]比較例におけるＸＢの構成図である。

[図６]比較例におけるＸＢの一部を抜粋して詳細を示した構成図である。

[図７]比較例における出力バッファ制御部の構成図である。

[図８]比較例におけるライトポインタ制御部の構成図である。

[図９]比較例におけるリードポインタ制御部の構成図である。

[図１０]比較例におけるＳＣの構成図である。

[図１１]比較例におけるディレクトリの例を示す図である。

[図１２]比較例においてＣＰＵがライトバックデータパケットをＳＣに出力したときの、ＳＣ内の出力バッファ制御部の動作を説明する図である。

[図１３]比較例においてＣＰＵが続いて転送データパケットをＳＣに出力した

ときの、S C内の出力バッファ制御部の動作を説明する図である。

[図14]第1実施形態によるサーバシステムの動作の概要を示す図である。

[図15]第1実施形態におけるS Cの構成図である。

[図16]第1実施形態におけるX Bの構成図である。

[図17]第1実施形態におけるS C内の出力バッファ制御部の構成図である。

[図18A]第1実施形態における管理テーブル部の構成図である。

[図18B]第1実施形態における比較部とライトポインタ制御部の構成図である。

。

[図19]第1実施形態における出力バッファ制御部の動作を説明する図である。

。

[図20]第1実施形態において先にライトバックデータパケットが入力されたときの出力バッファ制御部の動作の具体例を説明する図である。

[図21]第1実施形態においてライトバックデータパケットの後から共有型・転送データパケットが入力されたときの出力バッファ制御部の動作の具体例を説明する図である。

[図22]第1実施形態において共有型・転送データパケットが破棄された後にポート間アービタ部から出力許可が下りたときの出力バッファ制御部内のヘッダマージ部の動作の具体例を説明する図である。

[図23]第1実施形態における複製チェック部、データパケット復元部、セレクタ、およびルーティング部の構成を示す図である。

[図24]第1実施形態における複製チェック部、データパケット復元部、およびセレクタの動作を説明する図である。

[図25]第1実施形態におけるルーティングテーブル部の構成図である。

[図26]第1実施形態において複製情報を持ったライトバックデータパケットから元の2つのパケットを復元して2つのパケットをルーティングする動作の具体例を説明する図である。

[図27]第1実施形態における複製データパケット修正部の動作の具体例を説明する図である。

[図28] 第1実施形態におけるヘッダ復元部の動作の具体例を説明する図である。

[図29] 第1実施形態におけるデータパケット生成部の動作の具体例を説明する図である。

[図30] 第2実施形態におけるCPUの構成図である。

発明を実施するための最良の形態

[0020] 以下、実施形態について図面を参照しながら詳細に説明する。開示の技術の理解を助けるため、まずライトバック型のキャッシュメモリを用いるシステムについて説明する。

ライトバック型のキャッシュメモリを用いるシステムにおいては、キャッシュメモリに格納されたデータを主記憶装置に書き戻すライトバック動作が行われる。ライトバック動作が行われるときのシステム性能の向上に関する技術としては、以下のようなものが開示されている。

[0021] 例えば、ライトバック型キャッシュを有するCPUとホストブリッジとがCPUローカルバスを介してメインメモリを共有するシステムにおける次のような構成が開示されている。

[0022] ホストブリッジは、メインメモリへのリードアクセスアドレスを保持するアドレス保持回路と、アドレス保持回路に保持されたアドレスとCPUによるライトバック動作におけるラインアドレスとの一致を検出するアドレスヒット判定回路を備える。また、両アドレスが一致した場合に、ライトバック動作におけるライトデータを取り込み、且つ、メインメモリへのアクセスをキャンセルするメモリアクセス制御回路をホストブリッジが備える。以上の構成により、所望のデータを早期に得るとともに、ローカルバスのトラフィックを軽減することができる。

[0023] また、CPUローカルバスとIO（Input/Output）バスとをバスブリッジで連結してなるコンピュータシステムにおける次のような技術も開示されている。すなわち、IOバスに接続されるバスマスタ装置からIOバスを介し主記憶装置へのアクセス動作であるリード要求があった際、キャッシュ検査

部によりキャッシュへのライトバックが必要であることが通知された場合、次のようにして主記憶アクセス動作が高速化される。

[0024] 具体的には、書き戻し制御部は、主記憶制御部を介してライトバックデータを主記憶装置に書き戻すのと同時に、ライトバックデータをデータバイパス部にも転送する。すると、データバイパス部を介して、リード要求アドレスに応じた目的のデータが直接 I Oバス制御部に取り込まれ、I Oバスからバスマスタ装置へ転送される。以上により、主記憶装置に対する書き戻し処理の終了を待って主記憶装置からのデータ読み出し処理を行う必要がなくなる。

[0025] 以上のように、ライトバック動作が行われるときのシステム性能の向上に関するいくつかの技術が開示されている。しかしながら、複数のCPUを有するマルチCPUシステムにおいてライトバック型のキャッシュメモリが用いられると、キャッシュプロトコルによっては、システム内の制御装置およびデータ転送装置が比較的大きなサイズのバッファを必要とすることがある。

[0026] 以下では、ライトバック型のキャッシュメモリが用いられる場合に、システム内の制御装置およびデータ転送装置に必要なバッファサイズについて、具体的に説明する。

図1は複数のCPUを備えるシステムの例を示す構成図である。図1において、情報処理装置であるサーバシステム100は、複数のノード101a、101c、101e、101gとデータ転送装置であるクロスバススイッチ（以下「XB」と略す）102を備える。複数のノード101a、101c、101e、101gはXB102を介して互いに接続されている。図1のサーバシステム100のようにノード拡張された構成は、サーバシステムに好適である。

[0027] 図1においてノード101aは、バス108aによりXB102と接続されたシステムコントローラ（以下「SC」と略す）103aを備える。ノード101aはさらに、内部バスを介してSC103aに接続されたCPU1

04 a、CPU 104 b、メインメモリ 105 a、ディレクトリ 106 a、および I/O C (Input/Output Component) 107 a を備える。ノード 101 a 内の各部は、例えば、システムボードとも呼ばれる 1 枚のマザーボード上に実装されていてもよい。また、図 1 の例では、他のノード 101 c、101 e、および 101 g も、ノード 101 a と同様に構成されている。

[0028] サーバシステム 100 内の各コンポーネントについてさらに詳しく説明すれば下記のとおりである。

XB 102 は SC 103 a ~ 103 g 間でのデータ伝送を行うスイッチング機構である。後述のように、サーバシステム 100 内のコンポーネント間でのデータの授受はパケットの伝送により行われる。XB 102 は、ノード 101 a ~ 101 g 間でのパケットのルーティングを行うデータ転送装置の一例である。

[0029] SC 103 a は、ノード 101 a 内のコンポーネント間のデータ伝送を制御するとともに、ノード 101 a 内の各コンポーネントと他の SC 103 c ~ 103 g との間のデータ伝送を制御する制御装置である。また、SC 103 a はディレクトリ 106 a の参照や更新の制御も行う。SC 103 a ~ 103 g も SC 103 a と同様である。

[0030] CPU 104 a ~ 104 h のそれぞれは、演算や制御などを実行し、不図示のキャッシュメモリを備えている。キャッシュメモリは階層化されていてもよい。

メインメモリ 105 a は、例えば RAM (Random Access Memory) である。メインメモリ 105 a へのアクセスは、同じノード 101 a 内のコンポーネントからも行われ、他のノード 101 c ~ 101 g 内のコンポーネントからも行われる。メインメモリ 105 c ~ 105 g もメインメモリ 105 a と同様に、いずれのノード内のコンポーネントからもアクセス可能な RAM である。

[0031] ディレクトリ 106 a は、同じノード 101 a 内のメインメモリ 105 a に格納されたデータがどの CPU 内のキャッシュメモリにキャッシュされて

いるかを、例えばテーブル形式で格納する。つまり、ディレクトリ 106 a の参照や更新を行う SC 103 a は、データ伝送を制御するだけでなく、メインメモリ 105 a と CPU 内のキャッシュメモリ間のデータ整合性であるキャッシュコンシステンシの管理も行っている。ディレクトリ 106 c ~ 106 g もディレクトリ 106 a と同様に、同じノード内のメインメモリ 105 c ~ 105 g に格納されたデータについて管理する。

[0032] IOC 107 a は、不図示の PCI (Peripheral Component Interconnect) デバイスなどの様々な I/O デバイスと、IOC 107 a と SC 103 a の間のシステムバスとを接続するチップセットコンポーネントである。IOC 107 c ~ 107 g も IOC 107 a と同様である。

[0033] 以上のように構成されたサーバシステム 100 において、コンポーネント間の通信はパケットの送受信により行われる。パケットは、データの要求、データの通知、データの要求への応答、データの転送などのために送信される。パケットは、少なくともパケットヘッダを有し、パケットの種類 (タイプ) によってはデータペイロードをさらに有する。パケットは、さらに誤り検出用のトレイラを有していてもよい。

[0034] 以下では、パケットヘッダが以下の (a 1) ~ (a 3) の各フィールドを含むものとして説明する。

(a 1) パケットの宛先のコンポーネントを識別する宛先 ID (identifier)。

[0035] (a 2) パケットの要求元のコンポーネントを識別する要求元 ID。例えば、CPU 104 a が他のコンポーネントに何かを要求するときに送信する要求パケットの要求元 ID は、CPU 104 a の ID である。また、CPU 104 a からの要求パケットに対する応答として CPU 104 a へ送信されるパケットの要求元 ID も、CPU 104 a の ID である。

[0036] (a 3) 1 つの要求元からの異なる複数の要求 (リクエスト) を識別する要求 ID (以下「Req ID」と略す)。

図 1 の例において、各 CPU 104 a ~ 104 h が備えるキャッシュメモ

リがライトバック型のキャッシュメモリであるとする。すると、メインメモリ 105 a ~ 105 g のいずれかから読み出されてキャッシュメモリに格納されたデータが書き換えられた場合、ライトバック動作が必要である。しかし、具体的にどのタイミングでライトバック動作が行われるかは、キャッシュコンシステンシを保つためのキャッシュプロトコルに応じて様々である。

[0037] キャッシュプロトコルは、キャッシュラインの状態（ステート）の頭文字を使って、「MESI プロトコル」、「MOSI プロトコル」、「MOESI プロトコル」などと呼ばれる。ステートには以下の（b 1）～（b 5）のようなものがある。

[0038] （b 1）M（Modified）ステートは、キャッシュラインが1つのCPUのキャッシュメモリにだけ格納されており、かつキャッシュライン内の少なくとも一部のデータが、キャッシュメモリに格納された後で書き換えられた状態である。すなわち、Mステートのキャッシュラインはダーティ（dirty）である。

[0039] （b 2）E（Exclusive）ステートは、キャッシュラインが1つのCPUのキャッシュメモリにだけ格納されており、かつキャッシュラインがキャッシュメモリに格納された後で書き換えられていない状態である。すなわち、Eステートのキャッシュラインはクリーン（clean）である。

[0040] （b 3）S（Shared）ステートは、同じキャッシュラインが他のCPUのキャッシュメモリにも格納されている状態である。

（b 4）I（Invalid）ステートは、キャッシュラインが無効化された状態である。

[0041] （b 5）O（Owned）ステートは、メインメモリへのライトバックを行う責任を持った状態である。Oステートのキャッシュラインはダーティであり、同じキャッシュラインが他のCPUのキャッシュメモリにも格納されている。

[0042] キャッシュプロトコルによって、利用するステートやステート間の遷移の仕方が異なり、ライトバックの仕方も異なる。そして、CPUの仕様によっ

て、どのキャッシュプロトコルが採用されるかということが決まっている。

[0043] 例えば、CPU 104cがM状態で保持しているダーティなキャッシュラインのデータに対して共有型・データ転送要求が到来した場合のCPU 104cの動作には、CPUの仕様およびキャッシュプロトコルにより以下の(c1)～(c3)の3種類の動作をとりうる。

[0044] (c1) CPU 104cはライトバックデータパケットのみを出力する。データを要求した要求元への共有型・転送データパケットの送信は、ライトバックデータパケットを受け取り、メインメモリへデータをライトバックするSCが行う。なお、CPU 104cにおいて、共有型・データ転送要求により要求されたキャッシュラインは、キャッシュプロトコルに応じてIまたはS状態へ遷移する。

[0045] (c2) CPU 104cはデータを要求した要求元を宛先とする共有型・転送データパケットのみを出力する。また、CPU 104cにおいて、共有型・データ転送要求により要求されたキャッシュラインは、O状態へ遷移する。例えばMOESIプロトコルでは(c2)の動作が行われ、ライトバックはすぐには行われない。

[0046] (c3) CPU 104cはライトバックデータパケットと共有型・転送データパケットの双方を出力する。CPU 104cにおいて、共有型・データ転送要求により要求されたキャッシュラインは、キャッシュプロトコルに応じてIまたはS状態へ遷移する。

[0047] なお、以下の説明では、次の(d1)～(d4)に示すパケットタイプが定義されているものとして説明する。

(d1) CPUが、サーバシステム100内のいずれかのメインメモリに格納されているデータを要求するときに用いる「メモリアクセス要求パケット」がある。キャッシュプロトコルに応じて、メモリアクセス要求パケットには、排他型、共有型などの種類がある。

[0048] (d2) メモリアクセスが要求された対象データを、あるCPUがキャッシュしていたときに、メモリアクセスを要求した要求元へのデータの転送を

要求するために用いられる「データ転送要求パケット」がある。例えば、後述の図4の例のように、要求元以外のS Cがデータ転送要求パケットを出力する。キャッシュプロトコルに応じて、データ転送要求パケットにも、排他型、共有型などの種類がある。例えば、共有型のデータ転送要求パケットは、あるCPU内のキャッシュメモリにキャッシュされたデータを他のCPUに転送することを求めるパケットである。

[0049] (d3) CPUが、自分のキャッシュメモリにキャッシュしているデータ（すなわちMステートの）キャッシュラインのデータをメインメモリにライトバックするときに出力する「ライトバックデータパケット」がある。

[0050] (d4) データを要求した要求元へ、要求されたデータを転送するために用いられる「転送データパケット」がある。キャッシュプロトコルに応じて、転送データパケットにも、排他型、共有型、書き戻し責任付き共有型などの種類がある。

[0051] 続いて、上記(c1)～(c3)の場合について、図2～図4をそれぞれ参照してより具体的に説明する。

図2は、第1のキャッシュプロトコルにしたがったサーバシステム100の動作例を示す図である。

[0052] ステップS101において、CPU104aは、メインメモリ105eに格納されたデータを要求するため、SC103eを宛先とする共有型・メモリアクセス要求パケットを出力する。以下、説明の便宜上、ステップS101で出力される共有型・メモリアクセス要求パケットを識別するReqIDが5であると仮定する。

[0053] CPU104aが出力した共有型・メモリアクセス要求パケットは、SC103aとXB102を介して、メインメモリ105eと同じノード101e内のSC103eに送信される。

[0054] するとステップS102において、共有型・メモリアクセス要求パケットを受信したSC103eは、ディレクトリ106eを参照し、要求されたデータを含むキャッシュラインをキャッシュメモリに保持しているCPUを検

索する。

[0055] 以下では、ステップS 1 0 2におけるディレクトリ1 0 6 eの参照の結果、例えば、要求されたデータをCPU 1 0 4 cのみが保持していると判明したと仮定する。

すると、ステップS 1 0 3においてSC 1 0 3 eは、CPU 1 0 4 cを宛先とする共有型・データ転送要求パケットを出力する。SC 1 0 3 eが出力した共有型・データ転送要求パケットは、XB 1 0 2とSC 1 0 3 cを介して、CPU 1 0 4 cに送信される。

[0056] 上記(c 1)～(c 3)に関する前提として既に説明したように、図2の例においては、要求されたデータは、CPU 1 0 4 cがM状態で保持しているダーティなキャッシュライン内のデータである。よって、ステップS 1 0 4でCPU 1 0 4 cは、M状態のキャッシュラインのデータをメインメモリ1 0 5 eにライトバックするため、メインメモリ1 0 5 eと同じノード1 0 1 e内のSC 1 0 3 eを宛先とするライトバックデータパケットを出力する。出力されたライトバックデータパケットは、SC 1 0 3 cとXB 1 0 2を介してSC 1 0 3 eへと送信される。

[0057] また、上記(c 1)で述べたように、CPU 1 0 4 cは、要求されたデータを含むキャッシュラインのステートを、キャッシュプロトコルに応じてIまたはS状態に更新する。

[0058] そして、ライトバックデータパケットを受信したSC 1 0 3 eは、続くステップS 1 0 5において、ステップS 1 0 1でデータを要求した要求元であるCPU 1 0 4 aを宛先とする共有型・転送データパケットを出力する。共有型・転送データパケットのデータペイロードは、SC 1 0 3 eによって、ライトバックデータパケットのデータペイロードからコピーされたものである。

[0059] 共有型・転送データパケットは、XB 1 0 2とSC 1 0 3 aを介してCPU 1 0 4 aに送信される。

さらに、ステップS 1 0 6においてSC 1 0 3 eは、受信したライトバッ

クデータパケットに基づいて、CPU 104 cがM状態で保持していたキャッシュラインのデータをメインメモリ 105 eにライトバックする。

[0060] 以上の動作により、(c 1)の場合にCPU 104 cが有するキャッシュメモリとメインメモリ 105 e間のデータ整合性であるキャッシュコンシステンシが保たれる。

続いて、上記(c 2)の場合について図3を参照して説明する。図3は、第2のキャッシュプロトコルにしたがったサーバシステム100の動作例を示す図である。

[0061] 図3におけるステップS 201およびS 202は、図2におけるステップS 101およびS 102と同様であるので説明を省略する。

ステップS 203も、図2のステップS 103とほぼ同様であるが、以下の点においてステップS 103とは異なる。すなわち、上記(c 2)で述べたように、CPU 104 cは、要求されたデータを含むキャッシュラインのステートをOステートに変更する。

[0062] そして、続くステップS 204においてCPU 104 cは、データを要求した要求元であるCPU 104 aを宛先とする共有型・転送データパケットを出力する。出力された共有型・転送データパケットは、CPU 104 cからSC 103 c、XB 102、およびSC 103 aを介してCPU 104 aに送信される。

[0063] なお、図2の例とは異なり、図3の例では、ライトバックは後に行われる。具体的には、ステップS 203でOステートに変更したキャッシュラインをキャッシュメモリから追い出す入れ替え(リフィル)が後に生じたとき、CPU 104 cはライトバックデータパケットを出力してメインメモリ 105 eへのライトバックを行う。

[0064] 以上の動作により、(c 2)の場合にもCPU 104 cが有するキャッシュメモリとメインメモリ 105 e間のデータ整合性であるキャッシュコンシステンシが保たれる。

続いて、上記(c 3)のCPU 104がライトバックデータパケットと共

有型・転送データパケットの双方を出力する場合について図4を参照して説明する。図4は、第3のキャッシュプロトコルにしたがったサーバシステム100の動作例を示す図である。

[0065] 図4におけるステップS301～S304は、図2におけるステップS101～104と同様であるので説明を省略する。

図4の例ではライトバックデータパケットを出力するステップS304に続くステップS305において、CPU104cは、ステップS301でデータを要求した要求元であるCPU104aを宛先とする共有型・転送データパケットを出力する。共有型・転送データパケットは、SC103c、XB102、およびSC103aを介して送信され、CPU104aにおいて受信される。

[0066] なお、ステップS304とS305の実行順序は任意であり、両者が並列に実行されてもよい。また、CPU104cは、要求されたデータを含むキャッシュラインのステートを、キャッシュプロトコルに応じてIまたはSステートに更新する。

[0067] そして、ステップS306においてSC103eは、図2のステップS106と同様に、受信したライトバックデータパケットに基づいて、CPU104cがMステートで保持していたキャッシュラインのデータをメインメモリ105eにライトバックする。

[0068] 以上の動作により、(c3)の場合にもCPU104cが有するキャッシュメモリとメインメモリ105e間のデータ整合性であるキャッシュコンシステンシが保たれる。

以上、図2～図4を参照してキャッシュプロトコルに応じたサーバシステム100の動作例を説明したが、動作の違いに応じて、各SC103a～103gおよびXB102に必要なバッファの容量も異なる。具体的には、図4の例においては、図2および図3の例に比べて、より大きな容量のバッファが必要である。以下、その理由について説明する。

[0069] 上記のとおりXB102とSC103a～103gは、サーバシステム1

00内のコンポーネント間のデータ伝送を制御しており、具体的にはパケットのルーティングを行っている。そのため、XB102とSC103a~103gはいずれも、パケットを受け取る入力ポートと、入力されたパケットをルーティングするスイッチング機構と、ルーティング先のコンポーネントへとパケットを出力する出力ポートを備える。また、ルーティング先のコンポーネントにおけるパケットのオーバフローを避けるため、各SC103a~103gはXB102への出力ポートの前段にバッファを有し、XB102は各SC103a~103gへの出力ポートの前段にバッファを有する。

[0070] 例えば、SC103aは、XB102に出力するパケットをバッファに一時的に保存し、XB102から出力許可が得られると、バッファ内のパケットを出力ポートからXB102へと出力する。したがって、最大でどれくらいの量のパケットがバッファに滞留する可能性があるのかに応じて、必要なバッファの容量は異なる。

[0071] 図4の例では、キャッシュプロトコルから必然的に、ステップS304とS305という近接したタイミングで、2つのパケットがCPU104cからSC103cへ出力される。他方で、図2と図3の例では、2つのパケットがCPU104cからSC103cへ近接したタイミングで出力されるということはない。この違いのため、図4の例の方が、図2および図3の例よりも、必要なバッファの容量は大きい。

[0072] 例えばSC103cが同じノード101c内のCPU104c~104dへ出力するデータ転送要求パケットについて、応答を待たずに続けて発行できるパケットの最大数である最大アウトスタンディング（outstanding）数をPとして、SC103cに必要なバッファ容量について述べれば下記のごとくである。

[0073] 図2の例では、ステップS103でSC103cがCPU104cに1つの共有型・データ転送要求パケットを出力する。つまり、SC103cはXB102から受け取った共有型・データ転送要求パケットをCPU104cにルーティングする。

- [0074] すると、SC103cからCPU104cへの共有型・データ転送要求パケットの出力を契機として、ステップS104でCPU104cが1つのライトバックデータパケットをSC103cに出力する。したがって、SC103cは、少なくともP個のライトバックデータパケットを格納することが可能な容量のバッファを備える必要がある。
- [0075] 同様に、図3の例では、ステップS203でSC103cがCPU104cに1つの共有型・データ転送要求パケットを出力することを契機として、ステップS204でCPU104cは1つの共有型・転送データパケットを出力する。したがって、SC103cは、少なくともP個の共有型・転送データパケットを格納することが可能な容量のバッファを備える必要がある。
- [0076] 例えば、ライトバックデータパケットと共有型・転送データパケットのデータペイロードは、いずれも、要求された1つのキャッシュラインのデータであってもよい。したがって、図2と図3の例では、必要なバッファの容量は同様である。
- [0077] 他方で、図4の例では、ステップS303でSC103cがCPU104cに1つの共有型・データ転送要求パケットを出力することを契機として、ステップS304でCPU104cは1つのライトバックデータパケットをSC103cに出力する。そして、ステップS304とほぼ同時に行われるステップS305において、CPU104cは、1つの共有型・転送データパケットをSC103cに出力する。
- [0078] したがって、SC103cは、少なくともP個のライトバックデータパケットとP個の共有型・転送データパケットを格納することが可能な容量のバッファを備える必要がある。つまり、1つの共有型・データ転送要求パケットが契機となってCPU104cからSC103cへ出力されるデータパケットの数は、図2と図3の例では1個だけであるのに対し、図4の例では2個であるため、図4の例では必要なバッファの容量が大きい。
- [0079] もちろん、SC103cが備えるバッファの容量は、例えばIOC107cから受け付けるパケットの最大アウトスタンディング数など、他の要因に

も左右される。しかしながら、仕様により規定された様々な最大アウトスタンディング数が同じであれば、図2や図3の例よりも図4の例の方が、必要なバッファの容量が大きいことは上記の説明から明らかである。同様の理由から、XB102に必要なバッファの容量も、図2や図3の例よりも図4の例の方が大きい。

[0080] 続いて、開示の技術の理解を助けるため、図5～図13を参照して、比較例について詳細に説明する。その後、図14～図29を参照して、必要なバッファの容量が比較例と比べて約2分の1に削減される第1実施形態について詳細に説明する。また、必要なバッファの容量を削減するための別の構成を採用した第2実施形態について、図30を参照して説明する。

[0081] 以下、図5～図13において説明する比較例は、上記で説明した図1のサーバシステム100において図4に示したCPUがライトバックデータパケットと共有型・転送データパケットの双方を出力する動作を行う場合の例である。

[0082] 図5は、図1に示したXB102の比較例における構成図である。図1に示したように、XB102はバス108a～108gにより各SC103a～103gと接続されており、SC103a～103g間のデータ伝送を行う。図1では両向きの1本の矢印で表されている各バス108a～108gが、図5では2本の矢印により表されている。

[0083] XB102は、SC103aから出力されたパケットを受け取る入力ポート201aと、SC103aへパケットを出力する出力ポート202aを備える。他のSC103c、103e、および103gに関しても同様に、XB102は、入力ポート201c、201e、および201g、ならびに出力ポート202c、202e、および202gを備える。

[0084] 各SC103a～103gとの間のバス108a～108gの動作周波数とXB102内部の動作周波数が異なる場合、入力ポート201a～201gのそれぞれと、出力ポート202a～202gのそれぞれは、「周波数乗換え」と呼ばれる周波数の変換も行う。

[0085] X B 1 0 2 はさらに、ルーティング部 2 0 3 と、4 つの入力ポート 2 0 1 a ~ 2 0 1 g に対応する 4 つのルーティングテーブル部 2 0 4 a ~ 2 0 4 g を備える。

[0086] ルーティング部 2 0 3 は各入力ポート 2 0 1 a ~ 2 0 1 g と接続されており、入力ポート 2 0 1 a ~ 2 0 1 g からパケットを受け取る。また、ルーティングテーブル部 2 0 4 a ~ 2 0 4 g は、入力ポート 2 0 1 a ~ 2 0 1 g が受け取ったパケットをどの出力ポート 2 0 2 a ~ 2 0 2 g に振り分けるのかを決定するテーブル（後述する図 6 のテーブル 3 0 5）を有する。

[0087] ルーティング部 2 0 3 は、ルーティングテーブル部 2 0 4 a ~ 2 0 4 g の決定にしたがって、入力ポート 2 0 1 a ~ 2 0 1 g から受け取ったパケットを振り分ける。パケットの振り分けの詳細は図 6 とともに後述する。

[0088] また、X B 1 0 2 は、4 つの出力ポート 2 0 2 a ~ 2 0 2 g に対応した 4 組の回路を有する。

例えば、出力ポート 2 0 2 a に対応して出力ポート 2 0 2 a の前段に設けられた回路には、ポート間アービタ部 2 0 5 a と、ポート間セクタ 2 0 6 a と、3 つの出力バッファ制御部 2 0 7 c a、2 0 7 e a、2 0 7 g a が含まれる。各出力バッファ制御部 2 0 7 c a、2 0 7 e a、2 0 7 g a は、ルーティング部 2 0 3、ポート間アービタ部 2 0 5 a、およびポート間セクタ 2 0 6 a と接続されている。また、ポート間セクタ 2 0 6 a は、ポート間アービタ部 2 0 5 a および出力ポート 2 0 2 a と接続されている。

[0089] なお、出力バッファ制御部 2 0 7 c a は、添え字「c a」で示すように、S C 1 0 3 c から S C 1 0 3 a への送信（すなわち入力ポート 2 0 1 c から出力ポート 2 0 2 a への送信）に対応する。同様に、出力バッファ制御部 2 0 7 e a と 2 0 7 g a はそれぞれ、S C 1 0 3 e から S C 1 0 3 a への送信と S C 1 0 3 g から S C 1 0 3 a への送信に対応する。

[0090] 詳しくは図 7 とともに後述するが、3 つの出力バッファ制御部 2 0 7 c a、2 0 7 e a、および 2 0 7 g a は、出力ポート 2 0 2 a を介して S C 1 0 3 a へ出力するパケットを保存するバッファを有する。前述の説明において

、図2および図3の例に比べて図4の例ではバッファに必要な容量が大きいことを述べたが、当該バッファは具体的には図5に示した12個の出力バッファ制御部に含まれる。

[0091] 入力ポート201cがSC103cから受け取ったパケットが、ルーティング部203を介して出力バッファ制御部207caに入力されると、出力バッファ制御部207caは、ポート間アービタ部205aに対して出力リクエストを発行する。ここで「出力リクエスト」とは、出力バッファ制御部207caに入力されたパケットを、ポート間セクタ206aを介して出力ポート202aへと出力することを求めるリクエストである。

[0092] 同様に、入力ポート201eがSC103eから受け取ったパケットが、ルーティング部203を介して出力バッファ制御部207eaに入力されると、出力バッファ制御部207eaは、ポート間アービタ部205aに対して出力リクエストを発行する。また、出力バッファ制御部207gaも同様に、パケットが入力されるとポート間アービタ部205aに対して出力リクエストを発行する。

[0093] ポート間アービタ部205aは、出力バッファ制御部207ca、207ea、および207gaのそれぞれから発行された複数の出力リクエストを調停する。すなわち、ポート間アービタ部205aの「ポート間」とは、出力ポート202aを出力先とするパケットが入力された入力ポート201c、201e、および201gの間での調停を表している。

[0094] なお、調停のアルゴリズムは任意である。例えば、ポート間アービタ部205aは、LRU (Least Recently Used) アルゴリズム、ラウンドロビン (round robin) アルゴリズム、あるいはFIFO (First In First Out) アルゴリズムによる調停を行ってもよい。

[0095] 調停の結果、ポート間アービタ部205aは1つの出力バッファ制御部を選択する。

例えば、調停の結果、出力バッファ制御部207caが選択されたとする。すると、ポート間アービタ部205aは、出力バッファ制御部207ca

に出力許可を通知するとともに、出力バッファ制御部 207c a を選択したことをポート間セクタ 206 a に通知する。そして、ポート間セクタ 206 a は、ポート間アービタ部 205 a から通知された選択結果にしたがって、選択された出力バッファ制御部 207c a 内のバッファからパケットを受け取り、出力ポート 202 a へと出力する。

[0096] 以上のように、各出力バッファ制御部 207c a、207e a、および 207g a 内のバッファは、ポート間アービタ部 205 a から出力許可が通知されるまで（つまり、SC 103 a へのルーティングを待つ間）一時的にパケットを保存しておくための領域である。

[0097] 続いて、図 6 を参照して、SC 103 c から出力されたパケットを入力ポート 201 c が受け取ったときの XB 102 の動作について、ルーティングテーブル部 204 c の詳細とともに説明する。

[0098] 図 6 は、XB 102 の一部を抜粋して詳細を示した構成図である。図 6 には、入力ポート 201 c、ルーティング部 203 の一部、ルーティングテーブル部 204 c、ならびに 3 つの出力バッファ制御部 207c a、207c e、および 207c g が図示されている。入力ポート 201 c から入力されたパケットは、ルーティング部 203 を介して、3 つの出力バッファ制御部 207c a、207c e、および 207c g のいずれかに出力される。

[0099] XB 102 内の他のルーティングテーブル部 204 a、204 e、204 g、および他の入力ポート 201 a、201 e、201 g とルーティング部 203 との間関係も、図 6 と同様である。

[0100] 入力ポート 201 c が SC 103 c から受け取ったパケットを表す信号は、信号線 301 へ出力される。なお以下では、単に「パケットが信号線 301 へ出力される」のように表記する。

[0101] 信号線 301 は、信号線 302 と信号線 303 に分岐している。信号線 302 はルーティングテーブル部 204 c に接続されており、信号線 303 はルーティング部 203 に接続されている。分岐した信号線 302 を介してルーティングテーブル部 204 c へ送られるのは、信号線 301 に出力された

パケットのヘッダに含まれる宛先IDである。また、分岐した信号線303を介してルーティング部203へ送られるのは、信号線301に出力されたパケット全体である。また、ルーティングテーブル部204cは、信号線304によりルーティング部203と接続されている。

[0102] 以下の各図においては、上記のようにある信号線（例えば信号線301）が複数の信号線（例えば信号線302と信号線303）に分岐しており、複数の信号線それぞれへ出力されるデータ同士が異なるとき、図6のように斜めに枝分かれする線によって分岐点を示す。なお、複数に分岐した信号線に同じデータが出力される場合は、後述の図8における信号線の分岐点のように、黒丸を用いて分岐を示す。

[0103] 斜めに枝分かれする線によって図示される分岐は、例えば、以下の（e1）～（e3）を含む様々な仕方で実装することができる。

（e1）第1と第2の回路素子は、A本の導線の組により実現される信号線（つまりビット幅がAビットの伝送路）により接続されている。また、A本の導線のうちB本は分岐して第3の回路素子に接続されている（ $A > B$ ）。

[0104] 例えば図6において、第1～第3の回路素子が、それぞれ入力ポート201c、ルーティング部203、ルーティングテーブル部204cであり、信号線301と信号線303のビット幅がAビット、信号線302のビット幅がBビットであってもよい。

[0105] （e2）第1の回路素子からは、A本の導線の組により実現される信号線にAビットの信号が出力される。A本の導線のうちB本は第2の回路素子に接続し、C本は第3の回路素子に接続している（ $A = B + C$ ）。

[0106] （e3）第1と第4の回路素子は、A本の導線の組により接続されている。また、第4の回路素子はB本の導線の組により第2の回路素子と接続され、C本の導線の組により第3の回路素子と接続されている（ $A \geq B$ かつ $A \geq C$ ）。

[0107] 第4の回路素子は、例えば、第1の回路素子から入力されたAビットの信

号のうちBビットを第2の回路素子に出力するとともに、AビットのうちCビットを第3の回路素子に出力する。第4の回路素子は、必要に応じて、Bビットの出力を遅延させてもよく、Cビットの出力を遅延させてもよい。

[0108] 例えば図6において、信号線302と信号線303が枝分かれする部分に不図示の第4の回路素子があってもよい。そして、第4の回路素子が、入力を遅延させることなく、ルーティングテーブル部204cへは信号線302を介して宛先IDを出力するとともに、ルーティング部203へは信号線303を介してパケットを出力してもよい。

[0109] 以上の(e1)～(e3)のように、斜めに枝分かれする線によって図示される分岐は、様々な実装が可能であるが、ある信号線へ出力されたデータの一部または全部が複数の信号線に分岐して出力されるという機能は、(e1)～(e3)で共通である。また、3本以上の分岐に関しても同様である。

[0110] ここで図6の説明に戻ると、ルーティングテーブル部204cは、テーブル305を備える。また、ルーティング部203は、入力ポート201cに対応した分岐部306を備えるとともに、さらに、他の3つの入力ポート201a、201e、201gにそれぞれ対応する不図示の3つの分岐部も備えている。

[0111] ルーティングテーブル部204cは、テーブル305を参照し、入力ポート201cが受け取ったパケットのルーティング先を、信号線302を介して入力された宛先IDに基づいて決定する。すなわち、ルーティングテーブル部204cは、入力ポート201cが受け取ったパケットを、出力ポート202a、202e、202gのいずれに出力するのかを決定する。そして、ルーティングテーブル部204cは、決定した出力ポートを識別する出力ポート情報を、信号線304を介して分岐部306に出力し、通知する。

[0112] 信号線303は、具体的にはルーティング部203内の分岐部306に接続されている。分岐部306は、例えばマルチプレクサでもよい。分岐部306は、信号線303を介して入力されるパケットを受け取り、通知された

出力ポート情報に基づいて、受け取ったパケットの出力先を次のように切り換える。

- [0113] ・出力ポート情報が出力ポート 202 a を示す場合、分岐部 306 は、出力ポート 202 a に対応して出力ポート 202 a の前段に設けられた出力バッファ制御部 207 c a へとパケットを出力する。
- [0114] ・出力ポート情報が出力ポート 202 e を示す場合、分岐部 306 は、出力ポート 202 e に対応して出力ポート 202 e の前段に設けられた出力バッファ制御部 207 c e へとパケットを出力する。
- [0115] ・出力ポート情報が出力ポート 202 g を示す場合、分岐部 306 は、出力ポート 202 g に対応して出力ポート 202 g の前段に設けられた出力バッファ制御部 207 c g へとパケットを出力する。
- [0116] このように、ルーティング部 203 は、各入力ポート 201 a ~ 201 g が受け取ったパケットを、適切な出力ポート 202 a ~ 202 g へ振り分けるためのスイッチング機構である。
- [0117] 続いて、上記のルーティングテーブル部 204 c の動作について、テーブル 305 の詳細とともに説明する。

テーブル 305 は、書き換え可能な記憶素子により実現されている。例えば、FF（フリップフロップ）によりテーブル 305 が実現される。テーブル 305 は、宛先 ID とルーティング先の出力ポート情報とを関連づけて記憶している。

- [0118] 具体的には、ノード 101 a 内の SC 103 a、CPU 104 a、CPU 104 b、および IOC 107 a それぞれを識別する ID は、ノード 101 a の SC 103 a と接続された出力ポート 202 a を識別する出力ポート情報と関連づけられている。同様に、ノード 101 e 内の SC 103 e、CPU 104 e、CPU 104 f、IOC 107 e の ID は、出力ポート 202 e を識別する出力ポート情報と関連づけられている。また、ノード 101 g 内の SC 103 g、CPU 104 g、CPU 104 h、IOC 107 g の ID は、出力ポート 202 g を識別する出力ポート情報と関連づけられている。

。

[0119] ルーティングテーブル部204cは、テーブル305のほかに、さらに比較器308a～308iと、論理和演算を行うOR回路309a～309cと、セクタ310を備えている。また、図6の例では、テーブル305を構成するFFは、具体的には、上記の12個の宛先IDをそれぞれ記憶するための12組のFF307a～307iと、3個の出力ポート情報をそれぞれ記憶するための3組のFF311a～311cを含む。なお、宛先IDおよび出力ポート情報を表すのに用いられるビット数に応じて、1組のFF（例えばFF307aやFF311a）を構成するFFの数が決められる。

[0120] 図6に示すように、テーブル305の左列に相当し、各宛先IDを記憶している12組のFF307a～307iの各々に対応して、12個の比較器308a～308iが設けられている。また、テーブル305の右列に相当し、ルーティング先の各出力ポート情報を記憶している3組のFF311a～311cの各々に対応して、3個のOR回路309a～309cが設けられている。

[0121] 信号線302からルーティングテーブル部204cに入力された宛先IDは、比較器308a～308iのそれぞれに入力される。

例えば、比較器308aは、比較器308aに対応するFF307aが保持している宛先IDと、入力された宛先IDとを比較する。比較器308aは、2つの宛先IDが一致すれば論理値1をOR回路309aに出力し、2つの宛先IDが一致しなければ論理値0をOR回路309aに出力する。他の比較器308b～308iに関しても同様である。

[0122] 図6の例では、OR回路309aは、比較器308a～308dからの入力の論理和をセクタ310に出力する。同様に、OR回路309bは、比較器308e～308hからの入力の論理和をセクタ310に出力し、OR回路309cは、比較器308i～308lからの入力の論理和をセクタ310に出力する。

[0123] したがって、セクタ310には、OR回路309a～309cから1ビ

ットずつ、合計 3 ビットの制御信号が入力される。この 3 ビットの制御信号は、1 ビットのみが論理値 1 になるワンホット (One-Hot) 信号である。セレクタ 310 は、論理値 1 のビットの位置に応じて、FF 311 a ~ 311 c が保持する出力ポート情報のいずれかを選択し、信号線 304 を介して分岐部 306 に出力する。すなわち、セレクタ 310 は次のように動作する。

[0124] ・ OR 回路 309 a から論理値 1 が出力され、OR 回路 309 b と 309 c から論理値 0 が出力された場合、セレクタ 310 は、OR 回路 309 a に対応する FF 311 a が保持する出力ポート情報を分岐部 306 に出力する。

[0125] ・ OR 回路 309 b から論理値 1 が出力され、OR 回路 309 a と 309 c から論理値 0 が出力された場合、セレクタ 310 は、OR 回路 309 b に対応する FF 311 b が保持する出力ポート情報を分岐部 306 に出力する。

[0126] ・ OR 回路 309 c から論理値 1 が出力され、OR 回路 309 a と 309 b から論理値 0 が出力された場合、セレクタ 310 は、OR 回路 309 c に対応する FF 311 c が保持する出力ポート情報を分岐部 306 に出力する。

[0127] 以上説明したようにルーティングテーブル部 204 c が動作することにより、入力ポート 201 c が受け取ったパケットは、ヘッダに指定された宛先 ID に応じて、出力バッファ制御部 207 c a、207 c e、207 c g のいずれかへと適切に振り分けられる。

[0128] 続いて、図 7 を参照して出力バッファ制御部の詳細を説明する。図 7 は出力バッファ制御部 207 c a の構成図であるが、図 5 に示した他の 11 個の出力バッファ制御部も同様の構成である。

[0129] 図 7 に示すように、出力バッファ制御部 207 c a は、制御部 400 と、データパケット保存部 410 と、要求パケット保存部 420 と、パケット間セレクタ 430 を備える。また、図 5 に関して説明したとおり、出力バッファ制御部 207 c a は、ルーティング部 203、ポート間アービタ部 205

a、およびポート間セクタ206aと接続されている。

[0130] 制御部400は、出力バッファ制御部207ca全体を制御しており、詳しくは後述するが、例えば下記(f1)～(f4)のような動作を行う。

(f1) データパケット保存部410および要求パケット保存部420へのパケットの書き込みの制御

(f2) データパケット保存部410および要求パケット保存部420からパケット間セクタ430へのパケットの出力の制御

(f3) データパケット保存部410と要求パケット保存部420の双方にパケットが保存されているときに、いずれに保存されたパケットをパケット間セクタ430へ出力するのか、という調停

(f4) ポート間アービタ部205aへの出力リクエストの発行の制御

また、制御部400は、上記(f3)の制御を行うパケット間アービタ部401を備える。パケット間アービタ部401は、パケット間アービタ402と、比較器403と、比較器404を含む。さらに、制御部400は、パケットタイプに応じたVALID信号を出力するエンコーダ405を備える。具体的には、エンコーダ405はデータパケットと要求パケットを判別してVALID信号を出力する。

[0131] なお、以下の説明において「データパケット」とはデータペイロードを有するパケットであり、「要求パケット」とはデータペイロードを持たないパケットである。例えば、上記(d3)のライトバックデータパケットと(d4)の転送データパケットは、データパケットの例である。また、(d1)のメモリアクセス要求パケットと(d2)のデータ転送要求パケットは、要求パケットの例である。

[0132] また、制御部400は、データパケット用のライトポインタ制御部406とリードポインタ制御部407を備え、要求パケット用のライトポインタ制御部408とリードポインタ制御部409を備える。

[0133] 「リードポインタ」および「ライトポインタ」とは、データパケット保存部410および要求パケット保存部420内の番地を示すポインタである。

ライトポインタは、次にパケットを保存する番地を表し、リードポインタは、次に読み出すパケットが保存されている番地を表す。

[0134] データパケット保存部 410 はデータパケットを一時的に保存するバッファであり、要求パケット保存部 420 は要求パケットを一時的に保存するバッファである。すなわち、上記の各図の説明における「バッファ」とは、データパケット保存部 410 と要求パケット保存部 420 であり、特に、上記の説明で注目したバッファの容量は、データペイロードを保存するだけの容量が要求されるデータパケット保存部 410 の容量である。データパケット保存部 410 と要求パケット保存部 420 には、例えば RAM が用いられる。

[0135] また、データパケット保存部 410 と要求パケット保存部 420 におけるパケットのリード制御およびライト制御は、FIFO アルゴリズムにしている。よって、詳しくは図 8 とともに後述するが、ライトポインタ制御部 406 と 408 は、FF (Flip-Flop) を使ったカウンタによってライトポインタを管理することができる。同様に、詳しくは図 9 とともに後述するが、リードポインタ制御部 407 と 409 も、FF を使ったカウンタによってリードポインタを管理することができる。

[0136] また、パケット間セクタ 430 は、制御部 400 の制御に仕掛けて、データパケット保存部 410 または要求パケット保存部 420 のいずれかを選択し、パケットをポート間セクタ 206a へと出力する。

[0137] なお、以下では、データパケット保存部 410 および要求パケット保存部 420 内について、次の (g1) ~ (g4) のように定める。

(g1) 1つのデータパケットを保存するのに必要な容量は最大 D バイトである。すなわち、データパケットは、D バイトという固定長であるか、または、最長で D バイトの可変長であることが仕様で決められている。

[0138] (g2) データパケット保存部 410 内の番地は、D バイトを 1つのブロックとして、ブロックの番号により管理される。つまり、ライトポインタ制御部 406 が制御するライトポインタの値、およびリードポインタ制御部 4

07が制御するリードポインタの値は、データパケット保存部410内のブロックの番号である。

[0139] (g3) 1つの要求パケットを保存するのに必要な容量は最大Rバイトである($R < D$)。すなわち、要求パケットは、Rバイトという固定長であるか、または、最長でRバイトの可変長であることが仕様で決められている。

[0140] (g4) 要求パケット保存部420内の番地は、Rバイトを1つのブロックとして、ブロックの番号により管理される。つまり、ライトポインタ制御部408が制御するライトポインタの値、およびリードポインタ制御部409が制御するリードポインタの値は、要求パケット保存部420内のブロックの番号である。

[0141] 以上、出力バッファ制御部207caの構成について説明した。続いて、ルーティング部203から出力バッファ制御部207caにパケットが出力されてから、出力バッファ制御部207caがパケットをポート間セクタ206aへ出力するまでの動作を説明する。

[0142] まず、ルーティング部203が信号線441を介してパケットを出力バッファ制御部207caに出力する。信号線441は、図7に示すように信号線442と信号線443に分岐しており、信号線442へはパケットが出力される。

[0143] また、ステップS401に示すように、信号線443へはパケットヘッダ内のパケットタイプフィールドが出力される。すると、信号線443の接続先であるエンコーダ405は、パケットタイプフィールドの値に基づいて、ルーティング部203から入力されたパケットがデータパケットであるか要求パケットであるかを判別する。

[0144] ステップS401においてパケットがデータパケットであると判明した場合の動作は下記ステップS402d~S406dのとおりである。

エンコーダ405は、ステップS402dに示すように、データパケット用のライトポインタ制御部406にVALID信号を出力する。以下、説明の便宜上VALID信号は正論理であるとする、ステップS402dにお

いてエンコーダ405は、ライトポインタ制御部406に論理値1の信号を出力する。なお、この場合、エンコーダ405は、要求パケット用のライトポインタ制御部408には、無効を示す論理値0を出力する。

- [0145] VALID信号を受け取ったライトポインタ制御部406は、次のステップS403dにおいて、データパケット保存部410へライト指示とライトポインタ値を出力し、ライトポインタ値を更新する。また、論理値0が入力されたライトポインタ制御部408は、ライトポインタ値を更新せず、従前のライトポインタ値を保持する。なお、データパケット保存部410への「ライト指示」とは、信号線442を介して入力されるパケットを、ライトポインタ値により示されるデータパケット保存部410内の番地に書き込むことの指示である。以下では、説明の便宜上、ライト指示も正論理であるとする。すなわち、論理値1はライト指示がアサートされていることを示し、論理値0はライト指示がネゲートされていることを示すものとする。

- [0146] ここで、ステップS403dの動作をより詳細に説明するため、図8を参照する。図8は、ライトポインタ制御部406の構成図である。

図8に示すように、ライトポインタ制御部406はライトポインタ値を格納するFF451を備える。FF451は、エンコーダ405から入力されるVALID信号をカウントイネーブル信号として用いるカウンタ回路である。FF451のビット数は、上記(g2)で説明したデータパケット保存部410のブロックの番号の最大値による。また、ライトポインタ制御部406はエンコーダ405から入力された信号そのものをデータパケット保存部410に出力する。

- [0147] すなわち、ステップS402dでエンコーダ405からVALID信号として論理値1が入力されると、ステップS403dでライトポインタ制御部406は、入力された論理値1をそのままライト指示としてデータパケット保存部410に出力する。また、ライトポインタ制御部406は、FF451が保持するライトポインタ値をデータパケット保存部410に出力する。その後、ライトポインタ制御部406は、カウントイネーブル信号としての

VALID信号にしたがって、FF451が保持する値を1増やす。

- [0148] 逆に、もしエンコーダ405から論理値0が入力されれば、ライトポインタ制御部406は論理値0をそのままデータパケット保存部410に出力し、FF451はカウントアップ動作を行わない。
- [0149] ここで図7の説明に戻ると、データパケット保存部410はステップS404dにおいて次のように動作する。すなわち、データパケット保存部410は、ステップS403dで受けたライト指示にしたがい、信号線442を介して入力されるデータパケットを、ステップS403dで入力されたライトポインタ値で示される番地に保存する。
- [0150] また、ステップS405dとして示すように、ライトポインタ制御部406は、ステップS403dで更新したライトポインタ値をパケット間アービタ部401内の比較器403に出力する。なお、ステップS404dとS405dは、同時に並行して行われてもよい。
- [0151] また、図7に示すようにリードポインタ制御部407も比較器403に接続されており、リードポインタ制御部407が保持するリードポインタが前回更新されたときに比較器403にはリードポインタ値が入力されている。
- [0152] そこで、ステップS406dにおいて比較器403は、ライトポインタ制御部406から入力されたライトポインタ値と、リードポインタ制御部407から入力されたリードポインタ値を比較する。
- [0153] 比較した結果、ライトポインタ値とリードポインタ値が一致した場合は、ポート間セクタ206aへの出力を待つためにデータパケット保存部410に保存されているデータパケットは存在しない。したがって、比較器403は、パケット間アービタ402へ出力リクエストを出力しない。換言すれば、比較器403は、出力リクエストをネゲート(negate)する。
- [0154] 逆に、ライトポインタ値とリードポインタ値が不一致だった場合は、ポート間セクタ206aへの出力を待つためにデータパケット保存部410に保存されているデータパケットが存在する。したがって、比較器403は、パケット間アービタ402へ出力リクエストを出力(すなわちアサート(assert))

ert)) する。

- [0155] 以上説明したステップS 4 0 2 d ~ S 4 0 6 d は、ステップS 4 0 1 においてパケットがデータペイロードを有するデータパケットであると判明した場合の動作である。逆に、ステップS 4 0 1 においてパケットがデータペイロードを有さない要求パケットであると判明した場合の動作は、次のステップS 4 0 2 r ~ S 4 0 6 r のとおりである。ステップS 4 0 2 r ~ S 4 0 6 r はステップS 4 0 2 d ~ S 4 0 6 d と類似なので、簡単に説明する。
- [0156] ステップS 4 0 1 においてパケットが要求パケットであると判明すると、エンコーダ 4 0 5 は、ステップS 4 0 2 r に示すように、要求パケット用のライトポインタ制御部 4 0 8 に V A L I D 信号を出力するとともに、データパケット用のライトポインタ制御部 4 0 6 には無効を示す論理値 0 を出力する。
- [0157] V A L I D 信号を受け取ったライトポインタ制御部 4 0 8 は、次のステップS 4 0 3 r において、要求パケット保存部 4 2 0 へライト指示とライトポインタ値を出力し、ライトポインタ値を更新する。ライトポインタ制御部 4 0 8 も、図 8 のライトポインタ制御部 4 0 6 と同様に構成されている。
- [0158] そして、次のステップS 4 0 4 r において、要求パケット保存部 4 2 0 は、ステップS 4 0 3 r で受けたライト指示にしたがい、信号線 4 4 2 を介して入力される要求パケットを、ステップS 4 0 3 r で入力されたライトポインタ値で示される番地に保存する。
- [0159] また、ステップS 4 0 5 r として示すように、ライトポインタ制御部 4 0 8 は、更新後のライトポインタ値を比較器 4 0 4 に出力する。ステップS 4 0 4 r と S 4 0 5 r は並行して行われてもよい。
- [0160] そして、ステップS 4 0 6 r において比較器 4 0 4 は、ライトポインタ制御部 4 0 8 から入力されたライトポインタ値と、リードポインタ制御部 4 0 9 から入力されたリードポインタ値を比較する。両者が一致した場合は、比較器 4 0 4 はパケット間アービタ 4 0 2 へ出力リクエストを出力せず、不一致の場合は、比較器 4 0 4 はパケット間アービタ 4 0 2 への出力リクエスト

を発行する。

- [0161] パケット間アービタ 402 は、ステップ S 406 d または S 406 r において出力リクエストを受けた場合、続くステップ S 407 においてポート間アービタ部 205 a へ出力リクエストを出力する。
- [0162] なお、比較器 403 と比較器 404 の双方から出力リクエストが出力されている場合、パケット間アービタ 402 は、LRU アルゴリズムやラウンドロビン アルゴリズムなどによる調停を行う。パケット間アービタ 402 は、調停の結果、どちらの出力リクエストを先にポート間アービタ部 205 a へ出力するかを決定し、出力リクエストをポート間アービタ部 205 a へ出力する。
- [0163] 図 5 に関して説明したように、ポート間アービタ部 205 a は、出力バッファ制御部 207 c a、207 e a、および 207 g a からの出力リクエストを調停する。調停の結果、ステップ S 408 でポート間アービタ部 205 a が出力バッファ制御部 207 c a に対して出力許可を通知する。
- [0164] すると、ステップ S 407 でデータパケット用の比較器 403 からの出力リクエストが選択されてポート間アービタ部 205 a に出力された場合は、出力バッファ制御部 207 c a は、以下のステップ S 409 d ~ S 412 d のように動作する。また、ステップ S 407 で要求パケット用の比較器 404 からの出力リクエストが選択されてポート間アービタ部 205 a に出力された場合は、出力バッファ制御部 207 c a は、以下のステップ S 409 r ~ S 412 r のように動作する。
- [0165] すなわち、ステップ S 409 d でパケット間アービタ 402 は、データパケットを出力するよう指示する出力パケットタイプ選択信号をパケット間セレクタ 430 へ出力する。例えば、出力パケットタイプ選択信号は 2 ビットで表され、当該 2 ビットがデコードされることにより、「10」はデータパケットを出力することを示し、「01」は要求パケットを出力することを示し、「00」はどちらの種類のパケットも出力しないことを示し、「11」はドントケア (don't care) として使われない。もちろん、出力パケットタ

イプ選択信号は他の形式でもよい。

[0166] また、ステップS 4 0 9 dでパケット間アービタ 4 0 2は、データパケット用のリードポインタ制御部 4 0 7に対してリードポインタ更新指示を出力する。なお、説明の便宜上、以下ではリードポインタ更新指示も正論理であるとする。

[0167] ここで、ステップS 4 0 9 dの動作をより詳細に説明するため、図 9 を参照する。図 9 はリードポインタ制御部 4 0 7の構成図である。

図 9 に示すように、リードポインタ制御部 4 0 7はリードポインタ値を格納する F F 4 5 2 を備える。F F 4 5 2 は、パケット間アービタ 4 0 2 から入力されるリードポインタ更新指示をカウティネーブル信号として用いるカウンタ回路である。F F 4 5 2 のビット数も、上記（g 2）で説明したデータパケット保存部 4 1 0 のブロックの番号の最大値を表現できるだけのビット数が必要とされる。また、リードポインタ制御部 4 0 7 は、パケット間アービタ 4 0 2 から入力された信号そのものをデータパケット保存部 4 1 0 に出力する。

[0168] 以下、図 9 と図 7 の双方を参照しながらステップS 4 0 9 d～S 4 1 1 dについて説明する。ステップS 4 0 9 dでリードポインタ更新指示として論理値 1 が入力されると、ステップS 4 1 0 dでリードポインタ制御部 4 0 7 は、入力された論理値 1 をそのままリード指示としてデータパケット保存部 4 1 0 に出力する。また、リードポインタ制御部 4 0 7 は、F F 4 5 2 が保持するリードポインタ値をデータパケット保存部 4 1 0 に出力する。その後、リードポインタ制御部 4 0 7 は、カウティネーブル信号としてのリードポインタ更新指示にしたがって、F F 4 5 2 が保持する値を 1 増やす。

[0169] 逆に、もしパケット間アービタ 4 0 2 から論理値 0 が入力されれば、リードポインタ制御部 4 0 7 は論理値 0 をそのままデータパケット保存部 4 1 0 に出力し、F F 4 5 2 はカウントアップ動作を行わない。

[0170] ステップS 4 1 0 dでリードポインタ値を更新した場合、リードポインタ制御部 4 0 7 は、続くステップS 4 1 1 dにおいて、更新したリードポイン

タ値をパケット間アービタ部401内の比較器403に出力する。さらに、ステップS412dにおいてデータパケット保存部410は、ステップS410dで受けたリード指示にしたがい、ステップS410dで入力されたリードポインタ値が示す番地に保存しているデータパケットをパケット間セクタ430に出力する。

[0171] なお、ステップS411dとS412dは、並行して行われてもよい。また、リードポインタ値が更新されなかった場合は、リード指示がアサートされないので、ステップS412dは行われない。

[0172] 以上説明したステップS409d～S412dは、ステップS407でデータパケット用の比較器403からの出力リクエストが選択された場合の動作である。逆に、ステップS407で要求パケット用の比較器404からの出力リクエストが選択された場合の動作は、次のステップS409r～S412rのとおりである。ステップS409r～S412rはステップS409d～S412dと類似なので、簡単に説明する。

[0173] ステップS409rでパケット間アービタ402は、要求パケットを出力するよう指示する出力パケットタイプ選択信号（上記に例示した形式では「01」）をパケット間セクタ430へ出力する。また、パケット間アービタ402は、要求パケット用のリードポインタ制御部409に対してリードポインタ更新指示を出力する。リードポインタ制御部409も図9のリードポインタ制御部407と同様に構成されている。

[0174] ステップS409rでリードポインタ更新指示として論理値1が入力されると、ステップS410rでリードポインタ制御部409は、入力された論理値1をそのままリード指示として要求パケット保存部420に出力する。また、リードポインタ制御部409は、リードポインタ値を要求パケット保存部420に出力し、リードポインタ値を1増やす。

[0175] 逆に、もしパケット間アービタ402から論理値0が入力されれば、リードポインタ制御部409は論理値0をそのまま要求パケット保存部420に出力し、リードポインタの更新は行わない。

[0176] ステップS 4 1 0 rでリードポインタ値を更新した場合、リードポインタ制御部4 0 9は、続くステップS 4 1 1 rにおいて、更新したリードポインタ値をパケット間アービタ部4 0 1内の比較器4 0 4に出力する。さらに、ステップS 4 1 2 rにおいて要求パケット保存部4 2 0は、ステップS 4 1 0 rで受けたリード指示にしたがい、ステップS 4 1 0 rで入力されたリードポインタ値が示す番地に保存している要求パケットをパケット間セクタ4 3 0に出力する。

[0177] なお、ステップS 4 1 1 rとS 4 1 2 rは、並行して行われてもよい。また、リードポインタ値が更新されなかった場合は、リード指示がアサートされないので、ステップS 4 1 2 rは行われぬ。

[0178] 以上のようにして、ステップS 4 0 9 d～S 4 1 2 dまたはステップS 4 0 9 r～S 4 1 2 rが実行されると、ステップS 4 1 3においてパケット間セクタ4 3 0は次のように動作する。すなわち、パケット間セクタ4 3 0は、ステップS 4 0 9 dまたはS 4 0 9 rで入力された出力パケットタイプ選択信号にしたがって、データパケット保存部4 1 0または要求パケット保存部4 2 0からの出力を選択し、ポート間セクタ2 0 6 aへとパケットを出力する。

[0179] 以上、図5～図9を参照して、比較例におけるX B 1 0 2の詳細について説明した。

なお、図7のデータパケット保存部4 1 0には、例えば次のような場合に、同じデータペイロードを有するライトバックデータパケットと共有型・転送データパケットが保存される場合がある。すなわち、図1のCPU 1 0 4 aがメインメモリ1 0 5 aに格納されたデータを要求するため、共有型・メモリアクセス要求パケットを出力し、要求されたデータがCPU 1 0 4 cにキャッシュされていると判明した場合である。

[0180] この場合、ライトバックデータパケットと共有型・転送データパケットの双方がCPU 1 0 4 cから出力され、同じノード1 0 1 aに属するCPU 1 0 4 aとSC 1 0 3 aにそれぞれルーティングされる。よって、図7の出力

バッファ制御部 207 c a 内のデータパケット保存部 410 は、同じデータペイロードを有する 2 つのパケットを保存する可能性がある。

[0181] 続いて、図 10 と図 11 を参照して、比較例における SC 103 c の詳細について説明する。

図 10 は、図 1 に示した SC 103 c の比較例における構成図である。図示は省略したが、他の SC 103 a、103 e、および 103 g も SC 103 c と同様の構成である。

[0182] 図 10 に示したように、SC 103 c は、バス 108 c を介して XB 102 と接続されている。さらに、SC 103 c は、ノード 101 c 内の CPU 104 c、CPU 104 d、メインメモリ 105 c、ディレクトリ 106 c、および IOC 107 c とともに、ノード 101 c 内のバスを介して接続されている。SC 103 c は、ノード 101 c 内のコンポーネントと、他の SC 103 a、103 e、および 103 g との間でのパケット形式のデータの伝送を制御する。

[0183] SC 103 c は、XB 102、CPU 104 c、CPU 104 d、メインメモリ 105 c、ディレクトリ 106 c、IOC 107 c それぞれからのパケットをそれぞれ受け取る入力ポート 501 x、501 c、501 d、501 m、501 r、および 501 i を備える。

[0184] また、SC 103 c は、XB 102、CPU 104 c、CPU 104 d、メインメモリ 105 c、ディレクトリ 106 c、IOC 107 c それぞれへとパケットを出力する出力ポート 502 x、502 c、502 d、502 m、502 r、および 502 i を備える。

[0185] 以上の入力ポート 501 x ~ 501 i と出力ポート 502 x ~ 502 i は、必要に応じて、各コンポーネントの動作周波数の違いを吸収するための周波数の変換処理を行う。

また、SC 103 c は、互いに接続されたメモリアクセス制御装置である MAC (Memory Access Controller) 503 と SC 内制御回路 504 を備え、メインメモリ 105 c へのアクセスを制御する。MAC 503 はさらに入

力ポート501mと出力ポート502mにも接続されている。また、SC内制御回路504は入力ポート501rと出力ポート502rにも接続されている。

[0186] SC内制御回路504は、メインメモリ105cへのメモリアクセス要求を受けた際のキャッシュコンシステンシ制御を行う。キャッシュコンシステンシ制御は、MESIプロトコルなどの予め定められたキャッシュプロトコルにしたがって行われる。また、SC内制御回路504は、キャッシュコンシステンシ制御のため、ディレクトリ106cの参照と更新も制御している。

[0187] なお、SC内制御回路504は、同じノード101c内のCPU104c、CPU104d、またはIOC107cからメモリアクセス要求を受けることもある。また、SC内制御回路504は、XB102を介して他のノード101a、101e、または101g内のコンポーネントから、メモリアクセス要求を受けることもある。

[0188] 図11は、比較例におけるディレクトリ106cの例を示す図である。ディレクトリ106cは、同じノード101c内のメインメモリ105cを管理するデータ構造であり、SC内制御回路504がメインメモリ105cのキャッシュコンシステンシ制御を行う際に用いられる。

[0189] 図11ではディレクトリ106cをテーブル形式で表している。また、ディレクトリ106cの各エントリは、アドレスと所有者とステートの組である。

図11の左列の「アドレス」は、メインメモリ105c内のアドレスである。

[0190] 図11の中列の「所有者」は、そのアドレスで示されるメインメモリ105c内のブロックのデータを、キャッシュラインとしてキャッシュメモリに保持しているCPUのIDである。

[0191] 図11の右列の「ステート」は、「所有者」が保持するメモリデータのステートである。キャッシュプロトコルによって、例えば上記(b1)～(b

５）に例示した各種ステートのいずれかが、「ステート」を表すのに用いられる。

[0192] ここで図１０の説明に戻ると、MAC５０３は、SC内制御回路５０４から出力されるメモリアクセス要求に基づいて、実際にメインメモリ１０５ｃへデータを書き込んだり、メインメモリ１０５ｃからデータを読み出したりする制御を行う。

[0193] また、SC１０３ｃはルーティング部５０５を備える。ルーティング部５０５は、図５および図６に示したXB１０２のルーティング部２０３と同様に、パケットの振り分けを行う。ルーティング部５０５は、入力ポート５０１ｘ、５０１ｃ、５０１ｄ、および５０１ｉ、ならびにSC内制御回路５０４と接続されている。

[0194] また、SC１０３ｃは、入力ポート５０１ｘ、５０１ｃ、５０１ｄ、および５０１ｉ、ならびにSC内制御回路５０４にそれぞれ接続されているルーティングテーブル部５０６ｘ、５０６ｃ、５０６ｄ、５０６ｉ、および５０６ｓを備える。ルーティングテーブル部５０６ｘ、５０６ｃ、５０６ｄ、５０６ｉ、および５０６ｓの各々は、ルーティング部５０５とも接続されている。各ルーティングテーブル部５０６ｘ、５０６ｃ、５０６ｄ、５０６ｉ、および５０６ｓの詳細は、図６のルーティングテーブル部２０４ｃと同様なので、詳しい説明は省略する。

[0195] また、SC１０３ｃは、出力ポート５０２ｘ、５０２ｃ、５０２ｄ、および５０２ｉ、ならびにSC内制御回路５０４にそれぞれ対応した５組の回路を備える。５組の回路は、５つのコンポーネント（すなわち、出力ポート５０２ｘ、５０２ｃ、５０２ｄ、および５０２ｉ、ならびにSC内制御回路５０４）の前段にそれぞれ設けられている。

[0196] 例えば、出力ポート５０２ｘの前段に設けられた１組の回路は、ポート間アービタ部５０７ｘと、ポート間セクタ５０８ｘと、出力バッファ制御部５０９ｃｘ、５０９ｄｘ、５０９ｉｘ、および５０９ｓｘを含む。他の４つのコンポーネントの前段に設けられた回路も、出力ポート５０２ｘの前段に

設けられた上記の回路と同様である。

- [0197] ポート間アービタ部 507x は、XB102 に備えられた図 5 のポート間アービタ部 205a と同様である。すなわち、ポート間アービタ部 507x は、出力バッファ制御部 509cx、509dx、509ix、および 509sx と接続され、これらの出力バッファ制御部 509cx、509dx、509ix、および 509sx からの出力リクエストを調停する。
- [0198] また、ポート間セクタ 508x は、XB102 に備えられた図 5 のポート間セクタ 206a と同様である。すなわち、ポート間セクタ 508x は、ポート間アービタ部 507x から通知される調停の結果にしたがい、選択された出力バッファ制御部 509cx、509dx、509ix、または 509sx から出力されるパケットを出力ポート 502x へと出力する。
- [0199] なお、出力バッファ制御部 509cx には、CPU104c から入力ポート 501c が受け取ったパケットが、ルーティング部 505 を介して入力される。同様に、出力バッファ制御部 509dx には、CPU104d から入力ポート 501d が受け取ったパケットが、ルーティング部 505 を介して入力される。
- [0200] また、出力バッファ制御部 509ix には、IOC107c から入力ポート 501i が受け取ったパケットが、ルーティング部 505 を介して入力される。そして、出力バッファ制御部 509sx には、メインメモリ 105c から入力ポート 501m、MAC503、SC 内制御回路 504、およびルーティング部 505 を介して到着したパケットが入力される。
- [0201] 以上の出力バッファ制御部 509cx、509dx、509ix、および 509sx の詳細は、図 7 に示した XB102 内の出力バッファ制御部 207ca と同様であるので説明を省略する。また、以下では、例えば出力バッファ制御部 509cx 内の各コンポーネントについても、図 7 の出力バッファ制御部 207ca 内の各コンポーネントと同じ参照符号を用いる。
- [0202] 以上、図 1 における各コンポーネントの内部について、図 5～図 11 を参照して詳細に説明した。

続いて、図4と同じキャッシュプロトコルが採用され、図4のステップS304とS305が行われるときの、比較例における出力バッファ制御部509cxの動作について図12と図13をそれぞれ参照して詳しく説明する。

[0203] 図4に関して説明したように、SC103cが同じノード101c内のCPU104c~104dへ出力するデータ転送要求パケットの最大アウトスタンディング数をPとする。すると、少なくとも2P個のデータパケットを保存する容量のバッファが出力バッファ制御部509cxと509dxのそれぞれに必要である。

[0204] 換言すれば、図7と同様に構成されている出力バッファ制御部509cxにおいて、データパケット保存部410は、上記(g2)で述べたDバイトのブロックを少なくとも2P個含む。図12と図13は、データパケット保存部410が少なくとも2P個のブロックを含む必要があることを説明している。

[0205] 図12は、本比較例において、CPU104cが図4のステップS304でライトバックデータパケットをSC103cに出力したときの、SC103c内の出力バッファ制御部509cxの動作を説明する図である。また、図13は、本比較例において、CPU104cが図4のステップS305で転送データパケットをSC103cに出力したときの、SC103c内の出力バッファ制御部509cxの動作を説明する図である。

[0206] なお、上記のように出力バッファ制御部509cxは図7の出力バッファ制御部207caと同様に構成されている。そこで、図12および図13においては、出力バッファ制御部509cx内部のコンポーネントのうち、ステップS304とS305に関連するもののみを抜粋して図示し、他のコンポーネントは省略している。

[0207] また、図12のステップS501~S507は、図7で出力バッファ制御部207cに関して説明したステップS401、S402d~S406d、およびS407にとそれぞれ同様である。

[0208] さて、図4のステップS304においてCPU104cは、SC103eを宛先とするライトバックデータパケット601をSC103cに出力する。すると、SC103cは図10の入力ポート501cでライトバックデータパケット601を受け取る。

[0209] ライトバックデータパケット601の宛先はSC103eであり、SC103eはSC103cとは別のノード101eにある。よって、図10のルーティングテーブル部506cは、XB102に対応する出力ポート502xを示す出力ポート情報をルーティング部505に通知する。通知にしたがい、ルーティング部505は、ライトバックデータパケット601を出力ポート502xの前段の出力バッファ制御部509cxに出力する。

[0210] すると、図12のステップS501では、信号線443を介してライトバックデータパケット601のパケットタイプフィールド602がエンコーダ405に出力される。

ライトバックデータパケット601はデータパケットの一種なので、エンコーダ405は、続くステップS502において、データパケット用のライトポインタ制御部406にVALID信号を出力する。

[0211] 以下、説明の便宜上、ライトバックデータパケット601が出力バッファ制御部509cxに出力された時点においてライトポインタ制御部406内のFF451が保持するライトポインタの値が「0000」とであるとする。同様に、説明の便宜上、ライトバックデータパケット601が出力バッファ制御部509cxに出力された時点においてリードポインタ制御部407内のFF452が保持するリードポインタの値も「0000」とであるとする。

[0212] VALID信号を受け取ったライトポインタ制御部406は、次のステップS503において、データパケット保存部410へライト指示と「0000」というライトポインタ値を出力し、ライトポインタ値を「0001」とへと更新する。

[0213] すると、ステップS504においてデータパケット保存部410は、信号線442を介して入力されるライトバックデータパケット601を、ライト

指示にしたがい、ステップS 5 0 3で入力された「0 0 0 0」という番地に保存する。

[0214] また、ステップS 5 0 5においてライトポインタ制御部4 0 6は、ステップS 5 0 3での更新後の新たな「0 0 0 1」というライトポインタ値を、パケット間アービタ部4 0 1内の比較器4 0 3に出力する。また、上記で仮定した「0 0 0 0」という値のリードポインタ値も、比較器4 0 3には出力されている。

[0215] そして、ステップS 5 0 6で比較器4 0 3は、データパケット用のライトポインタ値「0 0 0 1」とリードポインタ値「0 0 0 0」を比較する。比較した結果、2つの値は不一致であるため、比較器4 0 3はパケット間アービタ4 0 2へ出力リクエストを出力する。

[0216] そして、ステップS 5 0 7でパケット間アービタ4 0 2は、ポート間アービタ部5 0 7 xへと出力リクエストを出力する。図1 2の例では、すぐにはポート間アービタ部5 0 7 xから出力許可が下りず、しばらくパケット間アービタ4 0 2は、ポート間アービタ部5 0 7 xへ出力リクエストを出力した状態のまま、出力許可を待っているものとする。

[0217] 次に、図1 3を参照して、CPU 1 0 4 cが図4のステップS 3 0 5で転送データパケットをSC 1 0 3 cに出力したときの、SC 1 0 3 c内の出力バッファ制御部5 0 9 c xの動作を説明する。なお、図1 3におけるステップ5 0 8～S 5 1 4も、図7のステップS 4 0 1、S 4 0 2 d～S 4 0 6 d、およびS 4 0 7とそれぞれ同様である。

[0218] また、図1 2と図1 3に示した例においては、ステップS 5 0 8は、ステップS 5 0 1の後に実行されるが、必ずしもステップS 5 0 2～S 5 0 7の後に実行されるとは限らない。例えば、ステップS 5 0 2をエンコーダ4 0 5が実行するのと並行して、ステップS 5 0 8が実行されることもありうる。

[0219] さて、図4のステップS 3 0 5においてCPU 1 0 4 cは、CPU 1 0 4 aを宛先とする共有型・転送データパケット6 0 3をSC 1 0 3 cに出力す

る。すると、SC103cは図10の入力ポート501cで共有型・転送データパケット603を受け取る。

[0220] 共有型・転送データパケット603の宛先はCPU104aであり、CPU104aはSC103cとは別のノード101aにある。よって、図10のルーティングテーブル部506cは、XB102に対応する出力ポート502xを示す出力ポート情報をルーティング部505に通知する。通知にしたがい、ルーティング部505は、共有型・転送データパケット603を出力ポート502xの前段の出力バッファ制御部509cxに出力する。

[0221] すると、図13のステップS508では、信号線443を介して共有型・転送データパケット603のパケットタイプフィールド604がエンコーダ405に出力される。

共有型・転送データパケット603はデータパケットの一種なので、エンコーダ405は、続くステップS509において、データパケット用のライトポインタ制御部406にVALID信号を出力する。

[0222] ライトバックデータパケット601と共有型・転送データパケット603の間にCPU104cは他のパケットを出力しなかったものとする、ステップS509の時点でライトポインタ制御部406が管理するライトポインタ値は「0001」であり、リードポインタ制御部407が管理するリードポインタ値は「0000」である。

[0223] VALID信号を受け取ったライトポインタ制御部406は、次のステップS510において、データパケット保存部410へライト指示と「0001」というライトポインタ値を出力し、ライトポインタ値を「0002」へと更新する。

[0224] すると、ステップS511においてデータパケット保存部410は、信号線442を介して入力される共有型・転送データパケット603を、ライト指示にしたがい、ステップS510で入力された「0001」という番地に保存する。

[0225] また、ステップS512においてライトポインタ制御部406は、ステッ

プS 5 1 0での更新後の新たな「0 0 0 2」というライトポインタ値を、比較器4 0 3に出力する。

ステップS 5 1 3において比較器4 0 3は、データパケット用のライトポインタ値「0 0 0 2」とリードポインタ値「0 0 0 0」を比較し、2つの値が不一致であるため、新たな出力リクエストをパケット間アービタ4 0 2へ出力する。

[0226] ここで、図1 2のステップS 5 0 7の出力リクエストへの出力許可がまだ下りていないとすると、パケット間アービタ4 0 2は、ポート間アービタ部5 0 7 xへ出力リクエストの信号を出力したままである。よって、ステップS 5 1 4でもパケット間アービタ4 0 2は、出力リクエストをアサートした状態を保つ。

[0227] 以上、図1 2と図1 3を参照して説明したとおり、上記比較例においては、ライトバックデータパケット6 0 1と共有型・転送データパケット6 0 3の双方をデータパケット保存部4 1 0が保存する必要があることがある。

[0228] ここで、ライトバックデータパケット6 0 1と共有型・転送データパケット6 0 3のデータペイロードは、図4の説明から明らかなおおり、同じである。また、一般にはパケットのヘッダよりもデータペイロードの方が長い。つまり、図1 2と図1 3の例では、番地「0 0 0 0」と番地「0 0 0 1」に、大部分が重複する2つのパケットが保存されることになり、バッファの使用効率が悪い。

[0229] 具体的には、出力バッファ制御部5 0 9 c x内のパケット間アービタ4 0 2がポート間アービタ部5 0 7 xに出力リクエストを出力してから、ポート間アービタ部5 0 7 xが出力許可をアサートするまでの時間が長いと、上記のような重複した保存の必要がある。出力許可が得られるまでの時間は動的に決まるので、出力バッファ制御部5 0 9 c xの設計においては最悪のケースを想定する必要がある。つまり、出力バッファ制御部5 0 9 c xのデータパケット保存部4 1 0には、上記のとおり少なくともDバイト×2 Pブロックの容量が必要である。

- [0230] また、図4ではステップS301においてCPU104aがメインメモリ105eに格納されたデータを要求しているが、CPU104eがメインメモリ105eに格納されたデータを要求する場合もありうる。その場合、データをキャッシュメモリに保存しているCPU104cが、SC103eを宛先とするライトバックデータパケットと、CPU104eを宛先とする共有型・転送データパケットの双方を出力する。
- [0231] ここで、データペイロードが共通する2つのパケットの宛先は、いずれもノード101e内のコンポーネントである。したがって、ライトバックデータパケットと共有型・転送データパケットの双方は、CPU104cからSC103eへと至る同じ経路を通る。
- [0232] この経路上では、SC103c内においてXB102に接続された出力ポート502xの前段のポート間アービタ部507xや、XB102内においてSC103eに接続された出力ポート202eの前段のポート間アービタ部205eにおいて調停が行われる。調停の結果によっては、XB102においてもライトバックデータパケットと共有型・転送データパケットを同時にバッファリングする必要が生じる。
- [0233] つまり、SC103eに接続された出力ポート202eの前段の出力バッファ制御部207ce内のデータパケット保存部410が、ライトバックデータパケットと共有型・転送データパケットの双方を保存する必要が生じることがある。
- [0234] つまり、本比較例においては、SC103a～103gとXB102がそれぞれ備える出力バッファ制御部は、同じデータペイロードを有する2つのパケットを重複して保存する必要が生じる可能性がある。
- [0235] そこで、上記のようなライトバックデータパケット601と共有型・転送データパケット603の重複した保存の必要をなくせば、バッファ（つまりデータパケット保存部410）の実装面積や消費電力の削減が可能となる。その結果、サーバシステム100全体の性能も向上する。
- [0236] 以下では、SC103a～103gとXB102がそれぞれ備える出力バ

ッファ制御部に必要なバッファ容量を上記比較例よりも削減した第1実施形態について、図14～図29を参照して説明する。

[0237] 第1実施形態においても、上記比較例と同様、図1のサーバシステム100を例として説明する。また、上記比較例との共通点については、適宜説明を省略する。以下では第1実施形態について、まず図14を参照して図4と比較しながら動作の概要を説明した後、図15と図16を参照して、比較例と異なるコンポーネントの概要を説明する。その後、図17～図29を参照して第1実施形態の詳細について説明する。

[0238] 図14は、第1実施形態によるサーバシステム100の動作の概要を示す図である。図14のサーバシステム100も、図4と同様のキャッシュプロトコルにしたがって動作し、図14のステップS601～S603は図4のステップS301～S303と同様である。

[0239] すなわち、ステップS601において、CPU104aは、メインメモリ105eに格納されたデータを要求するため、SC103eを宛先とする共有型・メモリアクセス要求パケットを発行する。すると、共有型・メモリアクセス要求パケットは、SC103aとXB102を介して、SC103eに送信される。

[0240] 続くステップS602において、共有型・メモリアクセス要求パケットを受信したSC103eは、ディレクトリ106eを参照し、要求されたデータを含むキャッシュラインをキャッシュメモリに保持しているCPUを検索する。その結果、要求されたデータをCPU104cがE状態で保持していると判明したとする。

[0241] すると、ステップS603においてSC103eは、CPU104cを宛先とする共有型・データ転送要求パケットを出力する。SC103eが出力した共有型・データ転送要求パケットは、XB102とSC103cを介して、CPU104cに送信される。

[0242] そして、ステップS604においてCPU104cは、M状態のキャッシュラインのデータをメインメモリ105eにライトバックするため、メ

インメモリ 105 e と同じノード 101 e 内の SC 103 e を宛先とするライトバックデータパケットを出力する。また、CPU 104 c は、要求されたデータを含むキャッシュラインのステートを、キャッシュプロトコルに応じて I または S ステートに更新する。

[0243] 詳しくは図 15 および図 17 ~ 図 22 とともに後述するが、第 1 実施形態における SC 103 c は、比較例の図 10 に示した出力バッファ制御部 509 c x と類似の出力バッファ制御部 702 c x (図 14 には不図示) を備えている。そして、詳しくは図 15 に示すとおり、比較例と同様に第 1 実施形態における SC 103 c も、図 10 に示したのと同じポート間アービタ部 507 x とポート間セクタ 508 x を備える。

[0244] よって、ステップ S 604 で出力されたライトバックデータパケットは、SC 103 c 内の出力バッファ制御部 702 c x 内のバッファに保存され、出力バッファ制御部 702 c x はポート間アービタ部 507 x に出力リクエストを発行する。

[0245] また、ステップ S 605 において CPU 104 c は、ステップ S 601 でデータを要求した要求元である CPU 104 a を宛先とする共有型・転送データパケットを出力する。

ステップ S 604 と S 605 の実行順は、CPU 104 c の仕様によって決まり、逆順でもよいが、以下では説明の便宜上、ステップ S 604 がステップ S 605 よりも先に実行されるものとして説明する。なお、図 14 は、ポート間アービタ部 507 x によってライトバックデータパケットの XB 102 への出力が許可されるよりも前に、共有型・転送データパケットがステップ S 605 において SC 103 c の出力バッファ制御部 702 c x に出力された場合を示す図である。

[0246] この場合、第 1 実施形態では、次のステップ S 606 に示すように、出力バッファ制御部 702 c x は、ライトバックデータパケットを保存する。また、ステップ S 607 に示すように、出力バッファ制御部 702 c x は、共有型・転送データパケットのヘッダに含まれる情報をライトバックデータパ

ケットと関連づけて保存することで、2つのケットをマージする。すなわち、出力バッファ制御部702cxは、同じデータペイロードを有する2つのケットのうち、後から受け取った方の共有型・転送データケットを破棄する。

[0247] 同じデータペイロードを有するライトバックデータケットと共有型・転送データケットをステップS607において1つのケットにマージすることにより、出力バッファ制御部702cxは、ケットの保存に必要な記憶容量を削減することができる。

[0248] ステップS607で共有型・データ転送要求ケットがライトバックデータケットにマージされた後、ポート間アービタ部507xから出力許可が下りると、ステップS608において出力バッファ制御部702cxは、マージされたケットをXB102に出力する。

[0249] マージされたケットは、元は、SC103eを宛先とするライトバックデータケットと、CPU104aを宛先とする共有型・転送データケットであり、宛先が異なる。そこで、次のステップS609でXB102は、マージされたケットから元のライトバックデータケットと共有型・転送データケットを復元する。

[0250] そして、XB102は、ステップS610で共有型・転送データケットを宛先のCPU104aと同じノード101a内のSC103aに出力するとともに、ステップS611でライトバックデータケットを宛先のSC103eに出力する。詳しくは図23～図29とともに後述するが、ステップS610とS611は、同時に実行されてもよい。または、ステップS610が先に実行されてもよく、ステップS611が先に実行されてもよい。

[0251] 以上、図14に示したように動作するサーバシステム100においては、SC103c内の出力バッファ制御部702cxに必要なバッファの容量が削減される。もちろん、出力バッファ制御部702cx以外にも、各SC103a～103gにおいて同様に必要なバッファの容量が削減される。したがって、第1実施形態によれば、SC103a～103gを実現する半導体

チップの製造コストおよび実装面積を削減することができ、SC103a～103gの消費電力を抑えることもできる。

[0252] また、ライトバックデータパケットと共有型・転送データパケットの長さの合計は、マージされた1つのパケットの長さよりも長い。したがって、図4のようにSC103cがライトバックデータパケットと共有型・転送データパケットを個々にバス108cを介してXB102へと出力する場合に比べて、第1実施形態では、バス108cの占有時間が短い。したがって、比較例と比べて第1実施形態では、他のパケットが出力許可を待つ時間が削減され、実質的にSC103cからXB102へのパケット出力のスループットが向上する。つまり、サーバシステム100全体としての性能も向上する。

[0253] 続いて、図15と図16を参照して、図14の動作を可能とするサーバシステム100内のコンポーネントの構成について、比較例との違いを中心に説明する。

図15は、第1実施形態におけるSC103cの構成図である。図15は、比較例におけるSC103cの構成を示した図10と類似しているが、次の(h1)～(h6)の点において異なる。

[0254] (h1) ルーティングテーブル部506xの代わりにルーティングテーブル部701xが設けられている。ルーティングテーブル部701xの構成は、図25とともに後述する。

[0255] (h2) 出力バッファ制御部509cxの代わりに、図14で説明したパケットをマージする機能を備えた出力バッファ制御部702cxが設けられている。後述するように、パケットのマージは、主に出力バッファ制御部702cx内の管理テーブル部801とヘッダマージ部804により行われる。管理テーブル部801とヘッダマージ部804は、図15では紙面の都合上、図示を省略したが、詳しくは図17～図22とともに説明する。

[0256] (h3) (h2)と同様に、出力バッファ制御部509dxの代わりに出力バッファ制御部702dxが設けられている。

(h 4) ルーティング部 505 の代わりにルーティング部 703 が設けられている。ルーティング部 505 とルーティング部 703 の違いは、図 29 についての説明の後で述べる。

[0257] (h 5) 入力ポート 501 x からルーティングテーブル部 701 x を経由してルーティング部 703 へ至る経路上に、不図示の複製チェック部 1001 が設けられている。なお、複製チェック部 1001 の詳細は図 23、図 24、図 26 とともに後述する。

[0258] (h 6) 入力ポート 501 x からルーティング部 703 へ至る経路上に、不図示のデータパケット復元部 1002 が設けられている。なお、データパケット復元部 1002 の詳細は図 23、図 24、図 26～図 29 とともに後述する。

[0259] 図 16 は、第 1 実施形態における X B 102 の構成図である。図 16 は、比較例における X B 102 の構成を示した図 5 と類似しているが、次の (i 1) ～ (i 9) の点において異なる。

[0260] (i 1) ルーティングテーブル部 204 c の代わりにルーティングテーブル部 711 c が設けられている。ルーティングテーブル部 204 c とルーティングテーブル部 711 c の違いは、図 25 とともに後述する。

[0261] (i 2) (i 1) と同様に、ルーティングテーブル部 204 a、204 e、および 204 g の代わりにルーティングテーブル部 711 a、711 e、および 711 g が設けられている。

[0262] (i 3) 出力ポート 202 c の前段には、出力バッファ制御部 207 a c、207 e c、および 207 g c の代わりに、図 14 で説明したパケットをマージする機能を備えた出力バッファ制御部 712 a c、712 e c、および 712 g c が設けられている。上記 (h 2) と同様に、パケットのマージは、主に出力バッファ制御部 712 a c、712 e c、および 712 g c それぞれの内部の管理テーブル部 801 とヘッダマージ部 804 により行われる。管理テーブル部 801 とヘッダマージ部 804 は、図 16 では紙面の都合上、図示を省略した。

[0263] (i 4) (i 3)と同様に、他の出力ポート202a、202e、および202gの前段においても、それぞれ3個ずつ設けられた出力バッファ制御部の構成が比較例とは異なる。

[0264] (i 5) ルーティング部203の代わりにルーティング部713が設けられている。ルーティング部203とルーティング部713の違いは、図23とともに後述する。

(i 6) 入力ポート201cからルーティングテーブル部711cを経由してルーティング部713へ至る経路上に、不図示の複製チェック部1001が設けられている。

[0265] (i 7) 他の入力ポート201a、201e、および201gとルーティング部713の間の経路においても、同様に不図示の複製チェック部1001が設けられている。

(i 8) 入力ポート201cからルーティング部713へ至る経路上に、不図示のデータパケット復元部1002が設けられている。

[0266] (i 9) 他の入力ポート201a、201e、および201gとルーティング部713の間の経路においても、同様に不図示のデータパケット復元部1002が設けられている。

[0267] 以上、図15と図16を参照して、第1実施形態において比較例と異なるコンポーネントの概要を示した。

ところで、図14に示した動作シーケンスでは、SC103cにおいてパケットがマージされ、XB102においてパケットが復元される。しかし、図15と図16に示すように、第1実施形態では、さらにSC103a~103gにおけるパケットの復元や、XB102におけるパケットのマージも可能なように、サーバシステム100が構成されている。その理由は、図14とは異なる次のような場合がありうるからである。

[0268] 例えば、CPU104eが、同じノード101e内のメインメモリ105eに格納されたデータを要求するため、SC103eを宛先とする共有型・メモリアクセス要求パケットを発行する。

- [0269] すると、SC103eは、ディレクトリ106eを参照し、要求されたデータを含むキャッシュラインをキャッシュメモリに保持しているCPUを検索する。その結果、要求されたデータをCPU104cがE状態で保持していると判明したとする。
- [0270] この場合、図4や図14に関して説明したキャッシュプロトコルによれば、CPU104cは、CPU104eを宛先とする共有型・転送データパケットと、SC103eを宛先とするライトバックデータパケットを出力する。ここでは説明の便宜上、共有型・データ転送要求パケットが先に出力される場合を例にして説明する。
- [0271] もし、SC103c内の出力バッファ制御部509cxから共有型・転送データパケットがポート間セクタ508xへ出力される前にライトバックデータパケットが出力バッファ制御部509cxに入力されれば、図14と同様に2つのパケットはマージされる。
- [0272] しかし、SC103c内の出力バッファ制御部509cxから共有型・転送データパケットが出力された後でライトバックデータパケットが出力バッファ制御部509cxに入力されれば、2つのパケットはマージされない。以下の説明では、2つのパケットがマージされなかった場合に着目する。
- [0273] ここで、CPU104eを宛先とする共有型・転送データパケットは、XB102において、SC103eと接続された出力ポート202eへと出力されるパケットである。同様に、SC103eを宛先とするライトバックデータパケットは、XB102において、出力ポート202eへと出力されるパケットである。
- [0274] よって、マージされずに個別にSC103cからXB102へと出力された共有型・転送データパケットとライトバックデータパケットの双方は、ともにXB102内の出力ポート202eの前段に設けられた出力バッファ制御部712ceを経由する。
- [0275] したがって、場合によっては、出力バッファ制御部712ce内に共有型・転送データパケットが一時的に保存されている状況下で、出力バッファ制

御部 7 1 2 c e にライトバックデータパケットが入力されることがある。例えば、ノード 1 0 1 e 内のコンポーネントを宛先とする多くのパケットが集中する場合、出力バッファ制御部 7 1 2 c e から共有型・転送データパケットが出力される前に、出力バッファ制御部 7 1 2 c e へライトバックデータパケットが入力されることがある。

[0276] この場合、X B 1 0 2 内の出力バッファ制御部 7 1 2 c e において 2 つのパケットをマージして保存し、S C 1 0 3 e へと出力し、S C 1 0 3 e がパケットの復元を行う。それにより、共有型・転送データパケットとライトバックデータパケットにおいて重複する部分の容量が不要となるため、X B 1 0 2 内に必要なバッファ容量が削減される。また、結果として、バス 1 0 8 e の占有時間も削減され、実質的にはサーバシステム 1 0 0 のスループットが向上する。

[0277] そのため、第 1 実施形態では、図 1 6 に示すように X B 1 0 2 内の各出力バッファ制御部は、パケットのマージができるように管理テーブル部 8 0 1 やヘッダマージ部 8 0 4 などを備えている。また、図 1 5 に示すように各 S C 1 0 3 a ~ 1 0 3 g において、X B 1 0 2 と接続された入力ポート 5 0 1 x とルーティング部 7 0 3 の間には、パケットの復元ができるように複製チェック部 1 0 0 1 やデータパケット復元部 1 0 0 2 が備えられている。

[0278] 続いて、図 1 4 の動作シーケンスにそって、第 1 実施形態における各コンポーネントの構成と動作について詳細に説明する。

まず、図 1 7 と図 1 8 を参照して、図 1 5 の出力バッファ制御部 7 0 2 c x の詳細な構成を説明する。なお、以下では出力バッファ制御部 7 0 2 c x を例として説明するが、図 1 5 の出力バッファ制御部 7 0 2 d x および図 1 6 に示した X B 1 0 2 内の 1 2 個の出力バッファ制御部（7 1 2 c e など）も同様の構成である。

[0279] 図 1 7 は、第 1 実施形態における S C 1 0 3 c 内の出力バッファ制御部 7 0 2 c x の構成図である。図 1 5 に示すように、出力バッファ制御部 7 0 2 c x はルーティング部 7 0 3、ポート間アービタ部 5 0 7 x、およびポート

間セクタ 508x と接続されている。

[0280] また、XB102内の例えば出力バッファ制御部 712ce も図 17 と同様の構成である。出力バッファ制御部 712ce は、図 16 に示すように、ルーティング部 713、ポート間アービタ部 205e、およびポート間セクタ 206e と接続されている点で出力バッファ制御部 702cx とは異なる。しかし、出力バッファ制御部 712ce の内部の構成は、図 17 の出力バッファ制御部 702cx と同様である。

[0281] 図 17 に示した SC103c 内の出力バッファ制御部 702cx は、比較例において図 7 に示した XB102 内の出力バッファ制御部 207ca と類似の構成を有するので、以下では共通点についての説明を適宜省略する。

[0282] 出力バッファ制御部 702cx は、比較例の出力バッファ制御部 207ca と同様に、制御部 400、データパケット保存部 410、要求パケット保存部 420、およびパケット間セクタ 430 を備える。ただし、出力バッファ制御部 702cx においては、制御部 400 の内部の構造と、データパケット保存部 410 の容量が比較例とは異なる。

[0283] 制御部 400 は、パケット間アービタ部 401 を有する。パケット間アービタ部 401 はパケット間アービタ 402 と比較器 403 と比較器 404 を備え、比較例と同様である。また、制御部 400 は比較例と同様のエンコーダ 405 を備える。制御部 400 はさらに、比較例と同様に、データパケット用のリードポインタ制御部 407 を備え、要求パケット用のライトポインタ制御部 408 とリードポインタ制御部 409 を備える。

[0284] 第 1 実施形態の制御部 400 が比較例と異なるのは、管理テーブル部 801 を備える点、比較部 802 を備える点、および比較例におけるライトポインタ制御部 406 の代わりにライトポインタ制御部 803 を備える点である。また、第 1 実施形態においては、出力バッファ制御部 702cx がさらに、ヘッダマージ部 804 とセクタ 805 を備える。

[0285] 管理テーブル部 801 と比較部 802 は協働して、同じデータペイロードを有する 2 つのパケットの組を判別する。そのために、管理テーブル部 80

1は、データパケット保存部410に保存されているパケットのヘッダに含まれるヘッダ情報を管理している。

[0286] 比較部802は、出力バッファ制御部702cxに新たに入力されたパケットと同じデータペイロードを有するパケットがデータパケット保存部410に保存されているか否かをライトポインタ制御部803に通知する。

[0287] ライトポインタ制御部803は、比較部802からの通知の結果に基づいて、新たに入力されたパケットのデータパケット保存部410への保存、または新たに入力されたパケットの破棄を制御する。その制御のため、ライトポインタ制御部803は、比較例と同様にライト指示およびライトポインタ値を出力するだけでなく、「複製ライト指示」を出力する。詳細は後述するが、複製ライト指示は、既にデータパケット保存部410に保存されているデータパケットと同じデータペイロードを有するパケットのヘッダ情報を記録することを管理テーブル部801に指示するためのものである。

[0288] ヘッダマージ部804は、同じデータペイロードを有する2つのパケットをマージしたパケットを生成する。セクタ805は、ヘッダマージ部804によりマージされたパケットと、データパケット保存部410に保存されているデータパケットのいずれかを選択する。

[0289] 以上、概略を述べた出力バッファ制御部702cx内の各コンポーネント間の関係は、次のとおりである。

まず、信号線441を介してパケットがルーティング部703から出力バッファ制御部702cxへ出力される。信号線441は、パケットを出力するための信号線442とパケットヘッダ内に含まれるヘッダ情報を出力するための信号線811とに分岐している。比較例に関して具体的な実装の例(e1)～(e3)とともに説明したように、第1実施形態においても、斜めに枝分かれする線によって図示される分岐は、出力されるデータが分岐ごとに異なることを示す。

[0290] 信号線442を介してデータパケット保存部410へ出力されるのは、比較例と同じくパケットである。信号線811に出力されるヘッダ情報は、具

体的には、パケットタイプ、宛先 I D、要求元 I D、R e q I D、およびデータステートである。

[0291] パケットタイプは、前述の（d 1）～（d 4）などのタイプを示し、比較例においてもエンコーダ 4 0 5 に入力されている。宛先 I D、要求元 I D、および R e q I D については上記（a 1）～（a 3）で説明したとおりである。

[0292] また、データステートとしては、キャッシュプロトコルに応じて前述の（b 1）～（b 5）の少なくとも一部が指定可能である。以下では、少なくとも、データステートの値が（b 3）の S ステートにあたる「共有型」である場合と、データステートが指定されない場合を例に取り上げる。

[0293] 信号線 8 1 1 はさらに信号線 4 4 3 と信号線 8 1 2 に分岐し、信号線 8 1 2 は信号線 8 1 3 と信号線 8 1 4 に分岐する。

信号線 4 4 3 は比較例と同じくエンコーダ 4 0 5 に接続され、ヘッダ情報のうちパケットタイプが信号線 4 4 3 を介してエンコーダ 4 0 5 に入力される。また、信号線 8 1 3 は比較部 8 0 2 に接続され、ヘッダ情報のうち要求元 I D と R e q I D が信号線 8 1 3 を介して比較部 8 0 2 に入力される。そして、信号線 8 1 4 は管理テーブル部 8 0 1 に接続され、パケットタイプ、宛先 I D、要求元 I D、R e q I D、データステートのすべてが信号線 8 1 4 を介して管理テーブル部 8 0 1 に入力される。

[0294] また、管理テーブル部 8 0 1 は、さらにライトポインタ制御部 8 0 3 とリードポインタ制御部 4 0 7 からの入力を受け取り、後述の「複製情報」を信号線 8 1 5 に出力する。

管理テーブル部 8 0 1 は、ライトポインタ制御部 8 0 3 からの入力に基づいて、新たに入力されたパケットと同じデータペイロードを持つパケットがデータパケット保存部 4 1 0 に保存されている場合は、上記新たなパケットのヘッダ情報を記憶する。また、管理テーブル部 8 0 1 は、リードポインタ制御部 4 0 7 からの入力に基づいて、リードポインタ値で示されるブロックに保存されているパケットにマージするヘッダ情報を検索する。

- [0295] 複製情報は、出力しようとするパケットにマージするヘッダ情報があるか否かを示す。複製情報はさらに、出力しようとするパケットにマージするヘッダ情報がある場合は、当該ヘッダ情報を含む。
- [0296] 信号線 8 1 5 は信号線 8 1 6 と信号線 8 1 7 に分岐しており、信号線 8 1 6 はヘッダマージ部 8 0 4 へ、信号線 8 1 7 はセクタ 8 0 5 へ、それぞれ接続されている。信号線 8 1 5 に出力された複製情報は信号線 8 1 6 を介してヘッダマージ部 8 0 4 へ出力される。また、複製情報のうち、出力しようとするパケットにマージするヘッダ情報があるか否かを示す「複製データ V A L I D」フィールドのみが信号線 8 1 7 を介してセクタ 8 0 5 に出力される。
- [0297] さらに管理テーブル部 8 0 1 は比較部 8 0 2 とともに接続されているが、管理テーブル部 8 0 1 と比較部 8 0 2 との間のデータの入出力については図 1 8 A および図 1 8 B とともに後述する。
- [0298] 比較部 8 0 2 は、信号線 8 1 3 を介して入力された要求元 I D と R e q I D を管理テーブル部 8 0 1 から入力された情報と比較し、比較結果をライトポインタ制御部 8 0 3 に通知する。この比較は、新たに入力されたパケットと同じデータペイロードを持つパケットがデータパケット保存部 4 1 0 に保存されているか否かを判断するために行われる。
- [0299] データパケット用のライトポインタ制御部 8 0 3 は、比較例のライトポインタ制御部 4 0 6 と同様に、比較器 4 0 3 およびデータパケット保存部 4 1 0 と接続されている。ライトポインタ制御部 8 0 3 はさらに、上記のように比較部 8 0 2 とともに接続されており、また、管理テーブル部 8 0 1 とともに接続されている。
- [0300] 以上、図 1 7 を参照して出力バッファ制御部 7 0 2 c x の構成について説明した。
- 続いて、出力バッファ制御部 7 0 2 c x 内の管理テーブル部 8 0 1、比較部 8 0 2、およびライトポインタ制御部 8 0 3 の詳細を図 1 8 A および図 1 8 B を参照して説明する。

- [0301] 図18Aは、第1実施形態における管理テーブル部801の構成図であり、図18Bは第1実施形態における比較部802とライトポインタ制御部803の構成図である。紙幅の都合上、図18Aと図18Bは2枚に分かれているが、「A」～「M」なる符号で示したように図18Aの管理テーブル部801は、図18Bの比較部802およびライトポインタ制御部803と接続されている。
- [0302] 図18Aにおいて管理テーブル部801は、データパケット保存部410内の $(N+1)$ 個の保存番地0000～xxxxに対応する $(N+1)$ 個のエントリ901-0～901-Nを有する。例えば、 $N=63$ ならば、「xxxx」＝「0063」である。
- [0303] また、管理テーブル部801は、複数ビットで表され、リードポインタ制御部407から出力されたリードポインタ値を、各1ビットの $(N+1)$ 個の出力にデコードするデコーダ902を有する。管理テーブル部801はさらに、 $(N+1)$ 個のAND回路903-0～903-Nを有する。 $0 \leq j \leq N$ なる各jについて、AND回路903-jにはリードポインタ制御部407から出力されたリード指示と、デコーダ902の $(j+1)$ 番目の出力とが入力される。
- [0304] 例えば、「1」というリードポインタ値とリード指示がリードポインタ制御部407から出力された場合、デコーダ902の出力のうち保存番地0001に対応する2番目の出力のみが論理値1であり、その他のN個の出力は論理値0である。よって、 $(N+1)$ 個のAND回路903-0～903-Nのうち、AND回路903-1の出力のみが論理値1である。
- [0305] また、管理テーブル部801は、FF904とFF905を有する。FF904は、それぞれ1ビットを出力する $(N+1)$ 個のAND回路903-0～903-Nからの出力、すなわち $(N+1)$ ビットの出力を保持する。FF905もFF904と同様であり、FF904とFF905は同じデータを保持する。2つのFF904とFF905を用いる代わりに、1つのFF904だけを用い、FF904の出力を図18AにおけるFF905の出

力先へも供給するよう、信号線を適宜分岐させる実施形態も可能である。

[0306] さらに、管理テーブル部 801 は、複数ビットで表され、データパケット用のライトポインタ制御部 803 から出力されたライトポインタ値を、各 1 ビットの $(N+1)$ 個の出力にデコードするデコーダ 906 を有する。

[0307] FF904 が保持する $(N+1)$ ビットのデータは、エントリ 901-0 ~ 901-N に 1 ビットずつ入力される。FF905 が保持する $(N+1)$ ビットのデータも、エントリ 901-0 ~ 901-N に 1 ビットずつ入力される。また、デコーダ 906 の $(N+1)$ ビットの出力も、エントリ 901-0 ~ 901-N に 1 ビットずつ入力される。さらに、エントリ 901-0 ~ 901-N にはそれぞれライトポインタ制御部 803 からのライト指示と複製ライト指示が入力される。

[0308] エントリ 901-0 ~ 901-N の構造と動作はそれぞれ同様であり、図 18A では保存番地 0000 用エントリ 901-0 についてのみ詳細に示している。すなわち、エントリ 901-0 は、7 個の AND 回路 907 ~ 913、および AND 回路 907 ~ 913 にそれぞれ対応する以下の 7 個のフィールドを有する。

[0309] ・VALID フィールド 914
・要求元 ID フィールド 915
・Req ID フィールド 916
・複製データ VALID フィールド 917
・複製パケットタイプフィールド 918
・複製パケット宛先 ID フィールド 919
・複製データステートフィールド 920

上記 7 個のフィールドのそれぞれは、例えば FF などの記憶素子を用いて実現される。各フィールドとその入出力の詳細は、比較部 802 とライトポインタ制御部 803 の詳細を説明した後で説明する。なお、上記各フィールドを備えるエントリ 901-0 ~ 901-N の全体が、1 つの管理テーブルに相当する。

[0310] また、図 17 に関して説明したとおり、パケットのヘッダ内のパケットタイプ、要求元 ID、Req ID、宛先 ID、データステートの各フィールドのデータがルーティング部 703 から信号線 814 を介して、管理テーブル部 801 に入力される。図 17 と図 18 A に示すように信号線 812 から分岐した信号線 814 は、図 18 A に示すように、さらに次のように分岐して各エントリ 901-0 ~ 901-N に接続している。

[0311] ・ 保存番地 0000 用エントリ 901-0 にパケットタイプと宛先 ID とデータステートを入力するための信号線 931-0

・ 保存番地 0000 用エントリ 901-0 に要求元 ID と Req ID が入力される信号線 932-0

・

・ 保存番地 x x x x 用エントリ 901-N にパケットタイプと宛先 ID とデータステートが入力される信号線 931-N

・ 保存番地 x x x x 用エントリ 901-N に要求元 ID と Req ID が入力される信号線 932-N

また、信号線 812 からの他方の分岐は、信号線 813 として図 18 B の比較部 802 に入力されている。信号線 813 を介してルーティング部 703 から比較部 802 に入力されるのは、パケットのヘッダに含まれる要求元 ID と Req ID である。

[0312] 図 18 B のとおり、比較部 802 は、(N+1) 個の保存番地 0000 ~ x x x x に対応する (N+1) 個の比較器 941-0 ~ 941-N を備える。比較器 941-0 ~ 941-N の構成は同様であり、比較器 941-0 ~ 941-N それぞれの出力は、ライトポインタ制御部 803 内が備える OR 回路 951 に入力される。

[0313] 例えば、保存番地 0000 に対応する比較器 941-0 には、信号線 933 を介してルーティング部 703 から、図 17 の出力バッファ制御部 702 c x へ入力されるパケットのヘッダに含まれる要求元 ID と Req ID が入力される。また、比較器 941-0 には、保存番地 0000 用エントリ 90

１－０内の要求元ＩＤフィールド９１５とＲｅｑ ＩＤフィールド９１６から、要求元ＩＤとＲｅｑ ＩＤがそれぞれ入力される。つまり、比較器９４１－０には、要求元ＩＤとＲｅｑ ＩＤの組が２組入力される。

[0314] また、比較器９４１－０には、制御信号として、保存番地００００用エントリ９０１－０内のＶＡＬＩＤフィールド９１４の値、すなわち、データパケット保存部４１０の保存番地００００のブロックが有効か否かを示す値が入力される。

[0315] 制御信号として入力されたＶＡＬＩＤフィールド９１４の値が論理値０の場合、比較器９４１－０は、論理値０を、ＯＲ回路９５１および、保存番地００００用エントリ９０１－０内のＡＮＤ回路９１０に出力する。

[0316] 制御信号として入力されたＶＡＬＩＤフィールド９１４の値が論理値１の場合、比較器９４１－０は、入力された要求元ＩＤとＲｅｑ ＩＤの組同士を比較する。そして、入力された２組の要求元ＩＤとＲｅｑ ＩＤが一致すれば、比較器９４１－０は、論理値１をＯＲ回路９５１とエントリ９０１－０内のＡＮＤ回路９１０とに出力する。逆に、入力された２組の要求元ＩＤとＲｅｑ ＩＤが不一致の場合は、比較器９４１－０は、論理値０をＯＲ回路９５１とエントリ９０１－０内のＡＮＤ回路９１０とに出力する。

[0317] なお、比較器９４１－０が出力する論理値０は、データパケット保存部４１０の保存番地００００には、新たに入力されたデータパケットとデータペイロードが一致するパケットが保存されていないことを示す。また、比較器９４１－０が出力する論理値１は、データパケット保存部４１０の保存番地００００には、新たに入力されたデータパケットとデータペイロードが一致するパケットが保存されていることを示す。

[0318] つまり、第１実施形態では、要求ＩＤとＲｅｑ ＩＤが一致すればデータペイロードが等しいということを利用している。このことはパケットの仕様に依存しており、パケットの仕様に応じた様々な実施形態が可能である。いずれにせよ各比較器９４１－０～９４１－Ｎは、パケットの仕様に応じて適切なヘッダ情報同士を比較することによって、データペイロードの一致または

不一致を判断することができる。

- [0319] データパケット用のライトポインタ制御部 803 は、上記の OR 回路 951 のほかに、NOT 回路 952、AND 回路 953、AND 回路 954、および FF 955 を備える。

OR 回路 951 は (N+1) 入力の OR 回路であり、比較器 941-0 ~ 941-N からの入力がすべて論理値 0 である場合のみ論理値 0 を出力し、それ以外の場合は論理値 1 を出力する。つまり、新たに出力バッファ制御部 702cx に入力されたデータパケットと同じデータペイロードを有する別のデータパケットがデータパケット保存部 410 内のいずれかのブロックに保存されている場合のみ、OR 回路 951 は論理値 1 を出力する。

- [0320] OR 回路 951 の出力は NOT 回路 952 と AND 回路 953 に入力され、NOT 回路 952 の出力は AND 回路 954 に入力される。

ここで、エンコーダ 405 は比較例と同様に、入力されたパケットがデータパケットであるか否かを示す VALID 信号を出力する。エンコーダ 405 から出力された VALID 信号は、AND 回路 953 および 954 に入力される。

- [0321] したがって、AND 回路 953 は、出力バッファ制御部 702cx にデータパケットが入力され、かつ、入力されたデータパケットと同じデータペイロードを有する別のパケットがデータパケット保存部 410 に保存されている場合にのみ、論理値 1 を出力する。AND 回路 953 からの出力は、図 18A の各エントリ 901-0 ~ 901-N に、複製ライト指示としてそれぞれ入力される。

- [0322] また、AND 回路 954 は、出力バッファ制御部 702cx にデータパケットが入力され、かつ、入力されたデータパケットと同じデータペイロードを有する別のパケットがデータパケット保存部 410 に保存されていない場合にのみ、論理値 1 を出力する。AND 回路 954 からの出力は、図 18A の各エントリ 901-0 ~ 901-N に、ライト指示としてそれぞれ入力される。

- [0323] また、ライトポインタ制御部 803 は、図 8 に示した比較例のライトポインタ制御部 406 が FF451 を備えるのと同様に、FF955 を備える。FF955 は、FF451 と同様にライトポインタ値を保持する。また、AND 回路 954 の出力がカウントイネーブル信号として FF955 に入力される。そして、FF955 の出力であるライトポインタ値は、比較例と同様にデータパケット保存部 410 と比較器 403 に入力されるだけでなく、さらに、図 18A のデコーダ 906 にも入力される。
- [0324] 以上、図 18B を参照して比較部 802 とライトポインタ制御部 803 の詳細を説明した。そこで、ライトポインタ制御部 803 からの入力を受け取る図 18A の管理テーブル部 801 の説明に戻ると、例えば保存番地 0000 用エントリ 901-0 の内の各コンポーネントの構成と動作は次のとおりである。
- [0325] VALID フィールド 914 は、データパケット保存部 410 の保存番地 0000 のブロックが有効か否かを示す 1 ビットのフィールドである。論理値 1 が「有効」を示し、論理値 0 が「無効」を示す。
- [0326] VALID フィールド 914 の値は、保存番地 0000 に新たなパケットが保存されるときに論理値 1 となる。また、VALID フィールド 914 の値は、保存番地 0000 に保存されているパケットがヘッダマージ部 804 を介して、または介さずに、セクタ 805 へ出力され、さらにパケット間セクタ 430 へと出力されるときに、論理値 0 となる。具体的には、VALID フィールド 914 は次のように構成されている。
- [0327] VALID フィールド 914 は、例えば RS (Reset-Set) 型 FF により実現されてもよい。その場合、RS 型 FF の S (セット) 端子には AND 回路 907 の出力が与えられ、R (リセット) 端子には、(N+1) ビットを保持する FF 904 の最初のビットであるビット 0 が与えられる。
- [0328] ここで、AND 回路 907 の入力、デコーダ 906 からの出力の最初のビットであるビット 0 (すなわち (N+1) ビットの出力のうち、保存番地 0000 に相当するビット) と、ライトポインタ制御部 803 から出力され

たライト指示である。したがって、ライトポインタ制御部 803 が保存番地 0000 を指定してライト指示をアサートした場合にのみ、AND 回路 907 の出力は論理値 1 となり、VALID フィールド 914 の値が論理値 1 にセットされる。

[0329] また、リードポインタ制御部 407 が保存番地 0000 を示すリードポインタ値を出力するとともにリード指示をアサートした場合にのみ、FF904 の最初のビットであるビット 0（すなわち (N+1) ビットのうち、保存番地 0000 に相当するビット）は論理値 1 となる。そして、FF904 のビット 0 が論理値 1 となると、VALID フィールド 914 の値は論理値 0 にリセットされる。

[0330] AND 回路 907 の出力と FF904 のビット 0 がともに論理値 0 の場合は、VALID フィールド 914 の値は変わらない。

また、前述のとおり、VALID フィールド 914 の値は、保存番地 0000 に対応する比較器 941-0 へと出力される。

[0331] 次に要求元 ID フィールド 915 について説明する。要求元 ID フィールド 915 は、データパケット保存部 410 の保存番地 0000 に保存するデータパケットの要求元 ID を記憶する。要求元 ID フィールド 915 への値の設定は、AND 回路 908 により行われる。以下、要求元 ID が例えば M_1 ビットで表されるとして説明する。

[0332] AND 回路 908 への入力、ライトポインタ制御部 803 から出力されるライト指示と、デコーダ 906 の出力のうち、保存番地 0000 に相当するビット 0 と、信号線 932-0 から分岐した信号線を介して入力される M_1 ビットの要求元 ID である。AND 回路 908 は、ライト指示とデコーダ 906 のビット 0 の出力がともに論理値 1 のときのみ、入力された M_1 ビットの要求元 ID の値を要求元 ID フィールド 915 に上書きする。ライト指示とデコーダ 906 からの出力のビット 0 の少なくとも一方が論理値 0 のとき、AND 回路 908 は、要求元 ID フィールド 915 への上書きを行わず、要求元 ID フィールド 915 は現在の値を保持する。

[0333] 上記のように動作するAND回路908と要求元IDフィールド915の具体的な実装方法は様々である。

例えば、AND回路908は、ライト指示とデコーダ906からの出力のビット0との論理積を得るAND回路を含むとともに、入力された M_1 ビットの要求元IDを要求元IDフィールド915に通過させるよう構成されていてもよい。また、要求元IDフィールド915は、例えば M_1 ビットを格納するための M_1 組の記憶素子を備え、各記憶素子の制御端子にはライトイネーブル用の制御信号として上記AND回路の出力が与えられる。それにより、上記AND回路の出力が論理値0のときは、要求元IDフィールド915が現在の値を保持し、上記AND回路の出力が論理値1のときは、AND回路908を通過した要求元IDが要求元IDフィールド915に書き込まれる。

[0334] 例えば、要求元IDフィールド915が M_1 個のRS型FFを備えていてもよい。ここで、説明の便宜上、AND回路908内の上記AND回路の出力（すなわちライト指示とデコーダ906からの出力のビット0との論理積）を「X」とし、AND回路908を通過した要求元IDのビットjを「 Y_j 」とする（ $0 \leq j \leq M_1 - 1$ ）。要求元IDフィールド915は、ビットごとに2つのAND回路と1つのNOT回路を備え、ビットjに対応するRS型FFのR端子には $X \wedge Y_j$ が、S端子には $X \wedge \neg Y_j$ が、それぞれ入力されるように構成される。

[0335] すると、Xが論理値0の場合は各RS型FFが現状の値を保持し、Xが論理値1の場合は各RS型FFの値は Y_j （すなわち要求元IDのビットjの値）となる。したがって、AND回路908と要求元IDフィールド915の上記の動作が実現される。

[0336] なお、AND回路908および要求元IDフィールド915の具体的な構成が上記の例と異なってもよいことは無論である。

また、前述のとおり、要求元IDフィールド915の値は、保存番地0000に対応する図18Bの比較器941-0へと出力される。

[0337] 続いて、ReqIDフィールド916について説明する。ReqIDフィ

ールド916は、データパケット保存部410の保存番地0000に保存するデータパケットのReqIDを記憶する。ReqIDフィールド916への値の設定は、AND回路909により行われる。

[0338] AND回路909への入力、ライトポインタ制御部803から出力されるライト指示と、デコーダ906からの出力のうち保存番地0000に相当するビット0と、信号線932-0から分岐した信号線を介して入力されるReqIDである。AND回路908と要求元IDフィールド915について説明したのと同様に、AND回路909とReqIDフィールド916の具体的な実装は様々である。

[0339] 具体的にどのように実装されるにしろ、AND回路909とReqIDフィールド916は以下のように動作する。すなわち、AND回路909は、ライト指示とデコーダ906からの出力のビット0がともに論理値1のときのみ、入力されたReqIDの値をReqIDフィールド916に上書きする。他方、ライト指示とデコーダ906からの出力のビット0の少なくとも一方が論理値0のとき、AND回路909は、ReqIDフィールド916への上書きを行わず、ReqIDフィールド916は現在の値を保持する。

[0340] また、前述のとおり、ReqIDフィールド916の値は、保存番地0000に対応する比較器941-0へと出力される。

続いて、複製データVALIDフィールド917について説明する。複製データVALIDフィールド917は、複製パケットタイプフィールド918、複製パケット宛先IDフィールド919、および複製データステートフィールド920が有効か否かを示す1ビットのフィールドである。論理値1が「有効」を示し、論理値0が「無効」を示す。

[0341] 複製データVALIDフィールド917の値は、保存番地0000に第1のパケットが保存されている状態で、第1のパケットと同じデータペイロードを有する第2のパケットが出力バッファ制御部702cxに入力されたときに論理値1となる。このとき、第2のパケットのヘッダに含まれるパケットタイプ、宛先ID、データステートはそれぞれ複製パケットタイプフィー

ルド918、複製パケット宛先IDフィールド919、複製データステートフィールド920に記憶される。

[0342] また、複製データVALIDフィールド917の値は、保存番地0000に保存されている第1のパケットが、図17のヘッダマージ部804とセレクタ805を介してパケット間セレクタ430へと出力されるときに、論理値0となる。

[0343] 具体的には、複製データVALIDフィールド917は、例えばRS型FFにより実現されてもよい。その場合、S端子にはAND回路910の出力が与えられ、R端子には、(N+1)ビットを保持するFF905の最初のビットであるビット0が与えられる。

[0344] ここで、AND回路910の入力は、保存番地0000に対応して図18Bの比較部802に設けられた比較器941-0の出力と、ライトポイント制御部803から出力された複製ライト指示である。したがって、新たに入力された第2のパケットと同じデータペイロードを有する第1のパケットが既に保存番地0000に保存されているときのみ、AND回路910の2つの入力がともに論理値1となり、複製データVALIDフィールド917の値が論理値1にセットされる。

[0345] また、リードポイント制御部407が保存番地0000を示すリードポイント値を出力するとともにリード指示をアサートした場合にのみ、FF905のビット0が論理値1となり、複製データVALIDフィールド917の値が論理値0にリセットされる。

[0346] AND回路910の出力とFF905のビット0がともに論理値0の場合は、複製データVALIDフィールド917の値は変わらない。

次に、複製パケットタイプフィールド918について説明する。ここで、図17の出力バッファ制御部702cxに新たに入力された第2のパケットと同じデータペイロードを有する第1のパケットが、既にデータパケット保存部410の保存番地0000に保存されていたとする。このとき、複製パケットタイプフィールド918は、第2のパケットのヘッダに含まれるパケ

ットタイプを記憶する。また、複製パケットタイプフィールド 918 への値の設定は、AND 回路 911 により行われる。

[0347] AND 回路 911 への入力は、図 18B の比較器 941-0 の出力と、ライトポインタ制御部 803 から出力された複製ライト指示と、信号線 931-0 から分岐した信号線を介して入力されるパケットタイプである。AND 回路 908 と要求元 ID フィールド 915 について説明したのと同様に、AND 回路 911 と複製パケットタイプフィールド 918 の具体的な実装は様々である。

[0348] 具体的にどのように実装されるにしろ、AND 回路 911 と複製パケットタイプフィールド 918 は次のように動作する。すなわち、AND 回路 911 は、比較器 941-0 の出力と複製ライト指示がともに論理値 1 のときのみ、入力されたパケットタイプの値を複製パケットタイプフィールド 918 に上書きする。他方、比較器 941-0 の出力と複製ライト指示の少なくとも一方が論理値 0 のとき、AND 回路 911 は複製パケットタイプフィールド 918 への上書きを行わず、複製パケットタイプフィールド 918 は現在の値を保持する。

[0349] 次に、複製パケット宛先 ID フィールド 919 について説明する。複製パケット宛先 ID フィールド 919 は、上記第 2 のパケットのヘッダに含まれる宛先 ID を記憶する。また、複製パケット宛先 ID フィールド 919 への値の設定は AND 回路 912 により行われる。

[0350] AND 回路 912 への入力は、図 18B の比較器 941-0 の出力と、ライトポインタ制御部 803 から出力された複製ライト指示と、信号線 931-0 から分岐した信号線を介して入力される宛先 ID である。宛先 ID は、例えば要求元 ID と同じく M_1 ビットで表される。AND 回路 908 と要求元 ID フィールド 915 について説明したのと同様に、AND 回路 912 と複製パケット宛先 ID フィールド 919 の具体的な実装は様々である。

[0351] 具体的にどのように実装されるにしろ、AND 回路 912 と複製パケット宛先 ID フィールド 919 は次のように動作する。すなわち、AND 回路 9

12は、比較器941-0の出力と複製ライト指示がともに論理値1のときのみ、入力された宛先IDの値を複製パケット宛先IDフィールド919に上書きする。他方、比較器941-0の出力と複製ライト指示の少なくとも一方が論理値0のとき、AND回路912は複製パケット宛先IDフィールド919への上書きを行わず、複製パケット宛先IDフィールド919は現在の値を保持する。

[0352] 次に、複製データステートフィールド920について説明する。複製データステートフィールド920は、上記第2のパケットのヘッダに含まれるデータステートを記憶する。また、複製データステートフィールド920への値の設定はAND回路913により行われる。

[0353] AND回路913への入力、図18Bの比較器941-0の出力と、ライトポインタ制御部803から出力された複製ライト指示と、信号線931-0から分岐した信号線を介して入力されるデータステートである。AND回路908と要求元IDフィールド915について説明したのと同様に、AND回路913と複製データステートフィールド920の具体的な実装は様々である。

[0354] 具体的にどのように実装されるにしろ、AND回路913と複製データステートフィールド920は次のように動作する。すなわち、AND回路913は、比較器941-0の出力と複製ライト指示がともに論理値1のときのみ、入力されたデータステートの値を複製データステートフィールド920に上書きする。他方、比較器941-0の出力と複製ライト指示の少なくとも一方が論理値0のとき、AND回路913は複製データステートフィールド920への上書きを行わず、複製データステートフィールド920は現在の値を保持する。

[0355] そして、以上説明した複製データVALIDフィールド917、複製パケットタイプフィールド918、複製パケット宛先IDフィールド919、および複製データステートフィールド920の値は、管理テーブル部801内のセクタ921へと出力される。同様に、他のエントリ901-1~90

1-Nからもセクタ921へと4つの値の組がそれぞれ出力される。

[0356] また、セクタ921には、選択信号として、AND回路903-0~903-Nからの出力が入力される。AND回路903-0~903-Nの出力はそれぞれ1ビットであるから、セクタ921の選択信号は(N+1)ビットである。また、明らかに、選択信号の(N+1)ビットのうち論理値1となるのは、最大1ビットである。

[0357] 選択信号の全ビットが論理値0の場合、セクタ921は出力をネゲートする（すなわち、有効なデータを出ししない）。選択信号のビットj（ $0 \leq j \leq N$ ）が論理値1のとき、セクタ921はエントリ901-jから出力された4つの値の組を選択して信号線815へと出力する。例えば、「0」というリードポインタ値とリード指示がリードポインタ制御部407から出力されると、AND回路903-0の出力が論理値1となり、セクタ921は保存番地0000用エントリ901-0からの出力を選択する。

[0358] 信号線815は、ヘッダマージ部804へ接続された信号線816と、セクタ805へ接続された信号線817に分岐している。そして、セクタ921が出力する複製データVALIDフィールド917の値は、信号線815と信号線816、信号線817を介してそれぞれヘッダマージ部804、セクタ805に出力される。セクタ921が出力する複製パケットタイプフィールド918、複製パケット宛先IDフィールド919、複製データステートフィールド920の値は、信号線815と信号線816を介してヘッダマージ部804に出力される。

[0359] ヘッダマージ部804は、リードポインタ値に対応するエントリ（901-0~901-Nのいずれか）からセクタ921を介して入力された各フィールドの値を、リード指示に応じてデータパケット保存部410から出力されたパケットにマージする。そして、ヘッダマージ部804は、マージして得られたパケットをセクタ805に出力する。

[0360] 例えば、データパケット保存部410に保存されている第1のパケットと同じデータペイロードを有する第2のパケットのヘッダの内容がj番目のエ

ントリ 901-j に記憶されており、第 1 のパケットの保存番地を指定するリード指示が発行される。すると、ヘッダマージ部 804 は、エントリ 901-j からセクタ 921 を介して入力された内容（すなわち、第 2 のパケットのヘッダに相当する情報）を、データパケット保存部 410 から出力される第 1 のパケットにマージする。

[0361] 以上、図 18A と図 18B を参照して 702cx の構成を説明した。続いて、出力バッファ制御部 702cx の動作を説明する。

図 19 は、第 1 実施形態における出力バッファ制御部 702cx の動作を説明する図である。なお、要求パケットが入力された場合、出力バッファ制御部 702cx は比較例における図 7 の出力バッファ制御部 207ca と同様に動作するので、説明は省略する。以下では、データパケットが出力バッファ制御部 702cx に入力された場合について説明する。

[0362] ステップ S701 で、CPU 104c が出力したデータパケットがルーティング部 703 によりルーティングされ、出力バッファ制御部 702cx に入力される。

すると、ステップ S702 で、データパケットのヘッダ内のパケットタイプがエンコーダ 405 と管理テーブル部 801 に出力される。また、ヘッダ内の要求元 ID と Req ID が管理テーブル部 801 と比較部 802 に出力される。さらに、ヘッダ内の宛先 ID とデータステートは管理テーブル部 801 に出力される。

[0363] 次のステップ S703 では、ルーティング部 703 から入力されたのはデータパケットであるということを、エンコーダ 405 がパケットタイプの値から判断する。その結果、エンコーダ 405 は、データパケット用のライトポインタ制御部 803 に VALID 信号を出力する。

[0364] また、ステップ S704 では、比較部 802 が、ステップ S702 で受け取った要求元 ID と Req ID の組を、管理テーブル部 801 内の各エントリ 901-0 ~ 901-N から出力されている要求元 ID と Req ID の組と比較する。比較の結果、比較部 802 は、図 18B に関して説明したよう

に、 $(N+1)$ 個の出力を管理テーブル部 801 内の各エントリ 901-0 ~ 901-N とライトポインタ制御部 803 内の OR 回路 951 に出力する。ステップ S704 はステップ S703 と並行して行われてもよい。

[0365] なお、図 18A に関して説明したように、管理テーブル部 801 内の各エントリ 901-0 ~ 901-N 内からは VALID フィールド 914 の値が比較部 802 へ出力されている。したがって、ステップ S704 では、VALID フィールド 914 の値が「有効」を示す論理値 1 であるエントリ内の要求元 ID と Req ID の組と、ステップ S701 で入力されたデータパケットの要求元 ID と Req ID の組とが比較される。

[0366] ステップ S705 では、データパケット用のライトポインタ制御部 803 が、OR 回路 951 を用いて、比較部 802 からステップ S704 で出力された $(N+1)$ 個の出力の論理和を求める。そして、ステップ S703 でエンコーダ 405 から出力された VALID 信号と、求めた論理和とに基づいて、ライトポインタ制御部 803 は、次のように動作する。

[0367] エンコーダ 405 から VALID 信号が出力（アサート）されており、かつ、OR 回路 951 が出力する論理和が論理値 0 である場合、ステップ S701 で入力されたデータパケットと同じデータペイロードを有するパケットは保存されていない。よって、ステップ S701 で入力されたデータパケットは破棄してはならない。

[0368] したがって、この場合、ライトポインタ制御部 803 は、ステップ S701 で入力されたデータパケットを現在使われていないデータパケット保存部 410 内のブロックに保存する処理を行う。すなわち、ライトポインタ制御部 803 は、図 18B の FF955 が保持するライトポインタ値を出力し、ライトポインタ値を更新する。ライトポインタ値の出力先は、管理テーブル部 801、データパケット保存部 410、および比較器 403 である。また、ライトポインタ制御部 803 は、管理テーブル部 801 とデータパケット保存部 410 へライト指示を出力する。

[0369] こうしてライトポインタ値とライト指示が出力された後は、データパケッ

ト保存部 410 がパケットを保存し、パケット間アービタ 402 が調停の結果としてポート間アービタ部 507x へ出力リクエストを出力する。詳しくは、図 7 の比較例でステップ S404d ~ S407 として説明した動作と同様である。

[0370] また、ステップ S705 において、エンコーダ 405 から VALID 信号が出力されており、かつ、OR 回路 951 が出力する論理和が論理値 1 である場合、ステップ S701 で入力されたデータパケットと同じデータペイロードを有するパケットが保存されている。

[0371] したがって、この場合、ライトポインタ制御部 803 は、ステップ S701 で入力されたデータパケットのうち、後の復元で使うためのヘッダ情報のみを管理テーブル部 801 に記憶し、データパケットそのものは破棄する処理を行う。すなわち、ライトポインタ制御部 803 は、複製ライト指示を管理テーブル部 801 に出力する。

[0372] そして、ステップ S706 において管理テーブル部 801 は、ステップ S702 で入力された情報を適切なエントリ内に保存する。管理テーブル部 801 のステップ S706 における動作は、ライトポインタ制御部 803 から出力されるライト指示、複製ライト指示、およびライトポインタ値、ならびに比較部 802 から出力される (N+1) 個の比較結果に基づく。

[0373] 具体的には、ステップ S706 における管理テーブル部 801 の動作は、ステップ S705 でライト指示と複製ライト指示のいずれが出力 (アサート) されたかによって 2 通りに場合分けされる。図 18B の説明から明らかとおり、ライト指示と複製ライト指示は排他的であり、同時に論理値 1 になることはない。

[0374] ステップ S705 でライト指示が出力された場合、管理テーブル部 801 は、ライトポインタ値が示す番地用のエントリ内の要求元 ID フィールド 915 と Req ID フィールド 916 に、ステップ S702 で入力された要求元 ID と Req ID の値を記憶する。また、管理テーブル部 801 は、ライトポインタ値が示す番地に対応するエントリ内の VALID フィールド 91

4の値を論理値1にセットし、ライトポインタ値が示す番地に対応するエントリを有効にする。

[0375] ステップS705で複製ライト指示が出力された場合は、比較器941-0~941-Nのいずれか1つが、ステップS704の比較の結果、一致を示す論理値1を出力した場合である。つまり、ステップS701で入力されたデータパケットと同じデータペイロードを有するパケットが保存されている番地が比較部802により検出された場合に、複製ライト指示が出力される。

[0376] この場合、管理テーブル部801は、判別された保存番地用のエントリ内の複製パケットタイプフィールド918、複製パケット宛先IDフィールド919、および複製データステートフィールド920に、ステップS702で入力された値をそれぞれ記憶する。また、複製ライト指示が出力されると、管理テーブル部801は、比較部802により判別されたエントリ内の複製データVALIDIDフィールド917を論理値1にセットする。

[0377] このように、複製ライト指示が出力された場合は、ステップS701で入力されたパケットそのものはデータパケット保存部410に保存されず、破棄される。また、複製ライト指示が出力された場合は、図18Bに示すようにライトポインタ値が更新されない。したがって、パケット間アービタ部401において改めて調停が行われることもなく、ポート間アービタ部507xへの新たな出力リクエストをパケット間アービタ402が出力することもない。こうして、ステップS705で複製ライト指示が出力された場合は、ステップS701でのパケットの入力を契機とする一連の処理が、ステップS706で終了する。

[0378] 他方で、ステップS705でライト指示が出力された場合は、上記のように図7のステップS404d~S407と同様の動作が行われ、その後、ステップS707においてポート間アービタ部507xからデータパケットの出力許可が出力される。

[0379] するとステップS708でパケット間アービタ402は、図7のステップ

S 4 0 9 dと同様に、データパケットを出力するようパケット間セクタ 4 3 0に選択信号を出力し、データパケット用のリードポインタ制御部 4 0 7にリードポインタ更新指示を出力する。

[0380] すると、ステップS 7 0 9でリードポインタ制御部 4 0 7は、リード指示とリードポインタ値を管理テーブル部 8 0 1とデータパケット保存部 4 1 0に出力し、リードポインタ値を更新する。また、リードポインタ制御部 4 0 7は、更新後のリードポインタ値を比較器 4 0 3へ出力する。

[0381] そして、ステップS 7 1 0でデータパケット保存部 4 1 0は、リード指示にしたがって、リードポインタ値が示す番地に保存されているパケットをヘッダマージ部 8 0 4とセクタ 8 0 5に出力する。

[0382] また、ステップS 7 1 1に示すように管理テーブル部 8 0 1は、リード指示にしたがって、リードポインタ値が示す番地に対応するエントリ内のデータをヘッダマージ部 8 0 4とセクタ 8 0 5に出力する。

[0383] 具体的には、管理テーブル部 8 0 1は、複製データV A L I Dフィールド 9 1 7、複製パケットタイプフィールド 9 1 8、複製パケット宛先I Dフィールド 9 1 9、および複製データステートフィールド 9 2 0の値をヘッダマージ部 8 0 4に出力する。また、管理テーブル部 8 0 1は、複製データV A L I Dフィールド 9 1 7の値を選択信号としてセクタ 8 0 5に出力する。

[0384] 出力された複製データV A L I Dフィールド 9 1 7の値は、論理値0のとき、リードポインタ値により示されるデータパケット保存部 4 1 0内の保存番地から読み出されたパケットそのものをセクタ 8 0 5が選択することを示す。逆に、出力された複製データV A L I Dフィールド 9 1 7の値は、論理値1のとき、ヘッダマージ部 8 0 4で生成される複製情報を持ったパケットをセクタ 8 0 5が選択することを示す。

[0385] ステップS 7 1 1において管理テーブル部 8 0 1はさらに、図 1 8 AのF F 9 0 4とF F 9 0 5を用いて、リードポインタ値で示される保存番地用のエントリ内のV A L I Dフィールド 9 1 4と複製データV A L I Dフィールド 9 1 7の値を論理値0にリセットする。

[0386] なお、ステップS 7 1 0とS 7 1 1は並列に実行されてもよい。

そして、ステップS 7 1 2においてヘッダマージ部8 0 4は、ステップS 7 1 0で入力されたデータパケットのヘッダに、ステップS 7 1 1で入力された4つのフィールドの情報を追加したパケットを生成し、セクタ8 0 5に出力する。具体例は後述するが、ヘッダマージ部8 0 4は、例えば、ステップS 7 1 0で入力されたデータパケットのヘッダの所定の位置に、4つのフィールドの情報を挿入することにより、パケットを生成する。

[0387] するとステップS 7 1 3においてセクタ8 0 5は、ステップS 7 1 1で管理テーブル部8 0 1から出力された選択信号にしたがって、データパケット保存部4 1 0から入力されたパケットと、ヘッダマージ部8 0 4から入力されたパケットの一方を選択する。そして、セクタ8 0 5は、選択したパケットをパケット間セクタ4 3 0に出力する。

[0388] ここで、パケット間セクタ4 3 0には、図7のステップS 4 0 9 rおよびS 4 0 9 dに関して説明したように、要求パケットとデータパケットのいずれを選択するのかわかる選択信号がパケット間アービタ4 0 2から入力されている。選択信号が、データパケットを選択することを示す値になると、ステップS 7 1 4に示すように、パケット間セクタ4 3 0は、セクタ8 0 5から出力されたパケットを選択し、ポート間セクタ5 0 8 xへ出力する。

[0389] 以上、図19を参照して出力バッファ制御部7 0 2 c xの動作について説明した。なお、図19の説明におけるデータパケットは、例えば、ライトバックデータパケットでもよく、共有型・転送データパケットでもよい。

[0390] また、図16に示したXB 1 0 2内の出力バッファ制御部7 1 2 c eなども、図17～図19を参照して説明した出力バッファ制御部7 0 2 c xと同様に構成されている。例えば、図19は、ルーティング部7 0 3とポート間アービタ部5 0 7 xとポート間セクタ5 0 8 xを、それぞれルーティング部7 1 3とポート間アービタ部2 0 5 eとポート間セクタ2 0 6 eに置き換えれば、出力バッファ制御部7 1 2 c eを示す図となる。なお、以下の図

に関しても同様に、出力バッファ制御部 702c x を中心にして説明を行うが、他の出力バッファ制御部についても同様である。

[0391] 次に、より具体的な例として、図 14 においてステップ S604 と S605 が行われた場合の SC103c 内の出力バッファ制御部 702c x の動作について、図 20～図 22 を参照して説明する。

[0392] 図 20 は、第 1 実施形態において先にライトバックデータパケットが入力されたときの出力バッファ制御部 702c x の動作の具体例を説明する図である。

図 20 には、出力バッファ制御部 702c x 内のコンポーネントのうち、以下で説明するステップに関連するもののみを抜粋して示してある。また、図 20 には、データパケット保存部 410 の詳細を示した。さらに、図 20 では、管理テーブル部 801 内の各エントリ 901-0～901-N を表形式で示した。

[0393] なお、以下では説明の便宜上、図 20 の動作が開始される時点で、ライトポインタ制御部 803 のライトポインタ値とリードポインタ制御部 407 のリードポインタ値の初期値は、ともに「0000」とであると仮定する。また、図 20 の動作が開始される時点で、管理テーブル部 801 内のすべてのエントリ 901-0～901-N において、VALID フィールド 914 の初期値が、「無効」を示す論理値 0 であると仮定する。

[0394] ステップ S801 で、CPU104c が出力したライトバックデータパケット 601 が図 15 の SC103c 内のルーティング部 703 によりルーティングされ、XB102 と接続された出力ポート 502x の前段の出力バッファ制御部 702c x に入力される。

[0395] するとステップ S802 で、ライトバックデータパケット 601 のパケットヘッダ情報 605 がエンコーダ 405、管理テーブル部 801、および比較部 802 へ出力される。

図 17 に関して説明したように、具体的には、ライトバックデータパケット 601 のヘッダには、少なくとも、パケットタイプ、要求元 ID、Req

I D、宛先 I D、およびデータステートのフィールドがある。

[0396] 図 20 の例では、パケットタイプの値は、ライトバックデータパケットであることを示す値である。

そして、図 14 と同様に図 20 も、CPU 104 a がメインメモリ 105 e 内のデータを要求したことを契機として、要求されたデータをキャッシュしている CPU 104 c がライトバックデータパケット 601 を出力する例を示す。したがって、ライトバックデータパケット 601 の要求元 I D の値は CPU 104 a の I D である。また、宛先 I D の値は、メインメモリ 105 e と同じノード 101 e 内にあってメインメモリ 105 e を管理する SC 103 e の I D である。

[0397] なお、Req I D の値は、説明の便宜上、「5」であるとする。また、ライトバックデータパケット 601 ではデータステートが指定されないため、データステートの値は「なし」を示す値である。

[0398] ステップ S 802 では、パケットタイプはエンコーダ 405 と管理テーブル部 801 に出力され、要求元 I D と Req I D は管理テーブル部 801 と比較部 802 に出力され、宛先 I D とデータステートは管理テーブル部 801 に出力される。

[0399] そして、ステップ S 803 では、エンコーダ 405 が、受け取ったパケットタイプに基づいて、「ライトバックデータパケット 601 はデータパケットである」と判断し、データパケット用のライトポインタ制御部 803 に V A L I D 信号を出力する。

[0400] また、ステップ S 804 では、比較部 802 が、受け取った要求元 I D と Req I D の値を、管理テーブル部 801 内の各エントリ 901-0 ~ 901-N の要求元 I D フィールド 915 と Req I D フィールド 916 から出力される値とそれぞれ比較する。実際には、図 18 A ~ 図 19 に関して説明したように、V A L I D フィールド 914 の値が「有効」を示す論理値 1 のエントリのみが比較対象である。

[0401] したがって、上記のとおり「すべてのエントリ 901-0 ~ 901-N に

においてVALIDフィールド914の値は論理値0である」という場合には、比較部802からの(N+1)個の出力はすべて、「不一致」を示す論理値0である。図18Bに示すように、比較部802はこの比較結果を管理テーブル部801とライトポインタ制御部803に出力する。

[0402] すると、ステップS805においてデータパケット用のライトポインタ制御部803は、比較部802からの(N+1)個の出力の論理和を求める。求めた論理和は、ライトバックデータパケット601の要求元IDおよびReqIDと一致する要求元IDおよびReqIDを記憶する有効なエントリが管理テーブル部801に存在しないことを示す。また、上記ステップS803のとおり、エンコーダ405からはVALID信号が出力されている。

[0403] よって、ステップS805でライトポインタ制御部803は、管理テーブル部801とデータパケット保存部410にライト指示と「0000」というライトポインタ値を出力し、その後、ライトポインタ値を「0001」に更新する。

[0404] するとステップS806で管理テーブル部801は、ライト指示にしたがい、入力されたライトポインタ値が示す「0000」という番地用のエントリ901-0において、3つのフィールドの値をセットする。

[0405] すなわち、管理テーブル部801は、エントリ901-0のVALIDフィールド914を、「有効」を示す論理値1にセットする。また、管理テーブル部801は、エントリ901-0の要求元IDフィールド915に、ステップS802で入力されたCPU104aのIDを保存する。さらに、管理テーブル部801は、エントリ901-0のReqIDフィールド916に、ステップS802で入力された「5」というReqIDの値を保存する。

[0406] また、ステップS807でデータパケット保存部410は、ライト指示にしたがい、入力されたライトポインタ値が示す「0000」という番地に、ライトバックデータパケット601を保存する。ステップS806とS807の実行順は任意であり、ステップS806とS807は並列に実行されて

もよい。

- [0407] また、ステップS 8 0 8では、パケット間アービタ部4 0 1の比較器4 0 3が、データパケット用のライトポインタ値とリードポインタ値を比較する。ライトポインタ値は、ライトポインタ制御部8 0 3から受け取った更新後の「0 0 0 1」という値であり、リードポインタ値はリードポインタ制御部4 0 7から受け取った「0 0 0 0」という値である。したがって、比較結果は「不一致」となり、比較器4 0 3はパケット間アービタ部4 0 1に出力リクエストを出力する。
- [0408] すると、パケット間アービタ部4 0 1は、ステップS 8 0 8で受け取った出力リクエストと要求パケットに関する出力リクエストとの間の調停を必要に応じて行った後、ステップS 8 0 9で、ポート間アービタ部5 0 7 xへ出力リクエストを出力する。
- [0409] 続いて、ステップS 8 0 9の出力リクエストに対する出力許可が得られる前に、さらに共有型・転送データパケットが出力バッファ制御部7 0 2 c xに入力された場合について説明する。
- [0410] 図2 1は、第1実施形態においてライトバックデータパケットの後から共有型・転送データパケットが入力されたときの出力バッファ制御部7 0 2 c x内の管理テーブル部8 0 1の動作の具体例を説明する図である。なお、図2 1には、出力バッファ制御部7 0 2 c x内のコンポーネントのうち、以下で説明するステップに関連するもののみを抜粋して示してある。また、図2 1では、管理テーブル部8 0 1内の各エントリ9 0 1-0~9 0 1-Nを表形式で示した。
- [0411] ステップS 8 1 0で、図1 5のCPU 1 0 4 cが出力した共有型・転送データパケット6 0 3がSC 1 0 3 c内のルーティング部7 0 3によりルーティングされ、XB 1 0 2と接続された出力ポート5 0 2 xの前段の出力バッファ制御部7 0 2 c xに入力される。
- [0412] するとステップS 8 1 1で、共有型・転送データパケット6 0 3のパケットヘッダ情報6 0 6が、エンコーダ4 0 5、管理テーブル部8 0 1、および

比較部 802 へ出力される。

- [0413] 共有型・転送データパケット 603 は、CPU 104a がメインメモリ 105e 内のデータを要求したことを契機として、要求されたデータをキャッシュしている CPU 104c が出力したパケットである。したがって、パケットヘッダ情報 606 において、パケットタイプの値は転送データパケットであることを示す値であり、要求元 ID と宛先 ID の値は CPU 104a の ID であり、Req ID の値は図 20 での仮定のとおり「5」である。また、パケットヘッダ情報 606 においてデータステートの値は、キャッシュプロトコルにしたがって、「共有型」を示す値であるとする。
- [0414] ステップ S811 では、パケットタイプはエンコーダ 405 と管理テーブル部 801 に出力され、要求元 ID と Req ID は管理テーブル部 801 と比較部 802 に出力され、宛先 ID とデータステートは管理テーブル部 801 に出力される。
- [0415] 続いて、ステップ S812 でエンコーダ 405 が、受け取ったパケットタイプに基づいて、「共有型・転送データパケット 603 はデータパケットである」と判断し、データパケット用のライトポインタ制御部 803 に VALID ID 信号を出力する。
- [0416] また、ステップ S813 では比較部 802 が、受け取った要求元 ID と Req ID の値を、管理テーブル部 801 内の各エントリ 901-0 ~ 901-N の要求元 ID フィールド 915 と Req ID フィールド 916 から出力される値とそれぞれ比較する。実際には、図 20 の前提と図 20 の動作から、VALID ID フィールド 914 の値が論理値 1 であるエントリ 901-0 のみが比較対象である。
- [0417] 図 20 のステップ S806 の結果、エントリ 901-0 において要求元 ID フィールド 915 には CPU 104a の ID が保存され、Req ID フィールド 916 には「5」という値が保存されている。したがって、ステップ S813 では、比較部 802 は、エントリ 901-0 に対応する比較器 941-0 から「一致」を示す論理値 1 を出力し、他のエントリ 901-1 ~ 9

01-Nに対応する比較器941-1~941-Nからは「不一致」を示す論理値0を出力する。これら(N+1)個の比較結果の出力先は、管理テーブル部801とライトポインタ制御部803である。

[0418] すると、ステップS814においてデータパケット用のライトポインタ制御部803は、比較部802からの(N+1)個の出力の論理和を求める。求めた論理和は、共有型・転送データパケット603の要求元IDおよびReqIDと一致する要求元IDおよびReqIDを記憶する有効なエントリが管理テーブル部801に存在することを示す。また、上記ステップS812のとおり、エンコーダ405からはVALID信号が出力されている。

[0419] よって、ステップS814でライトポインタ制御部803は、管理テーブル部801に複製ライト指示を出力する。換言すれば、ステップS814ではライト指示がネゲートされており、ライト指示がデータパケット保存部410に出力されない。したがって共有型・転送データパケット603そのものは、データパケット保存部410に保存されず、破棄される。

[0420] そして、ステップS815で管理テーブル部801は、ライトポインタ制御部803から出力された複製ライト指示と比較部802からの(N+1)個の出力に基づいて、一致が検出されたエントリ901-0において、4つのフィールドの値をセットする。

[0421] すなわち、管理テーブル部801は、エントリ901-0の複製データVALIDフィールド917を、「有効」を示す論理値1にセットする。また、管理テーブル部801は、エントリ901-0の複製パケットタイプフィールド918に、ステップS811で入力された、転送データパケットを示す値を保存する。

[0422] さらに、管理テーブル部801は、エントリ901-0の複製パケット宛先IDフィールド919に、ステップS811で入力されたCPU104aのIDを保存する。また、管理テーブル部801は、エントリ901-0の複製データステートフィールド920に、ステップS811で入力された、「共有型」を示す値を保存する。

[0423] 以上、図20と図21を参照して説明したように、第1実施形態によれば、第1のデータパケットに対する出力許可が得られる前に第2のデータパケットが入力されたとき、データパケット保存部410内の1つのブロックのみが使われる。比較例では図13のとおり同じデータペイロードを持つ2つのパケットを保存するためにデータパケット保存部410の2つのブロックが使われるのと比較して、第1実施形態ではデータパケット保存部410に必要な記憶容量が2分の1に削減されている。

[0424] つまり、SC103cが同じノード101c内のCPU104c~104dへ出力する共有型・データ転送要求パケットの最大アウトスタンディング数がPのとき、比較例ではデータパケット保存部410に2Pブロックの容量が必要である。それに対し、第1実施形態では、データパケット保存部410にPブロックの容量が必要である。

[0425] なお、第1実施形態では、比較例には存在しない管理テーブル部801が設けられており、管理テーブル部801内にはFFなどの記憶素子により実現されるエントリ901-0~901-Nが含まれる。しかしながら、一般にパケットのヘッダはデータペイロードよりも十分に短い。よって、管理テーブル部801のために付加的に必要な記憶容量を考慮しても、第1実施形態の方が比較例よりも出力バッファ制御部に必要な記憶容量が大幅に少ない。よって、データパケット保存部410だけではなく出力バッファ制御部全体で必要な記憶容量で比較したとしても、第1実施形態は比較例の約2分の1の容量しか必要としない。

[0426] 続いて、図21のステップS814の後でポート間アービタ部507xから出力許可が下りたときの動作、すなわち図19のステップS707~S714の具体例を説明する。

図22は、第1実施形態において図21の処理により共有型・転送データパケットが破棄された後にポート間アービタ部507xから出力許可が下りたときの出力バッファ制御部内702cxのヘッダマージ部804の動作の具体例を説明する図である。図22には、出力バッファ制御部702cx内

のコンポーネントのうち、以下で説明するステップに関連するもののみを抜粋して示してある。また、図22にはヘッダマージ部804の動作を示す模式図を含む。

[0427] ステップS901で、ポート間アービタ部507xからパケット間アービタ402に対して、出力許可が出力される。

するとステップS902でパケット間アービタ402は、データパケットを出力するよう指示するパケット間セクタ430に選択信号を出力するとともに、データパケット用のリードポインタ制御部407へリードポインタ更新指示を出力する。

[0428] 続いて、ステップS903でリードポインタ制御部407は、リードポインタ更新指示にしたがって、まずリード指示と「0000」というリードポインタ値を管理テーブル部801とデータパケット保存部410に出力した後、リードポインタ値を更新する。そして、リードポインタ制御部407は、更新後のリードポインタ値「0001」を比較器403に出力する。

[0429] ここで、図20のステップS805においてライトポインタ値は「0001」に更新されているため、ステップS903の実行によりリードポインタ値とライトポインタ値が一致する。したがって、リードポインタの更新により、比較器403はパケット間アービタ402への出力リクエストをネゲートするようになる。つまり、比較器403からパケット間アービタ402へ、出力リクエストが出力されないようになる。

[0430] したがって、もし要求パケットに関して比較器404がパケット間アービタ402へ出力リクエストを出力していなければ、パケット間アービタ402からポート間アービタ部507xへの出力リクエストも出力されないようになる。

[0431] また、ステップS904でデータパケット保存部410は、ステップS903で入力されたリード指示とリードポインタ値にしたがい、リードポインタ値が示す番地「0000」に保存されているパケットをヘッダマージ部804とセクタ805に出力する。なお、図20に関して説明したように、

「００００」番地「００００」にはライトバックデータパケット６０１が保存されている。

[0432] また、ステップＳ９０５で管理テーブル部８０１は、ステップＳ９０３で入力されたリード指示にしたがい、リードポインタ値が示す番地「００００」用のエントリ９０１－０に保存されている４つのフィールドの値をヘッダマージ部８０４に出力する。

[0433] すなわち、管理テーブル部８０１は、エントリ９０１－０の複製データVALIDフィールド９１７、複製パケットタイプフィールド９１８、複製パケット宛先IDフィールド９１９、および複製データステートフィールド９２０の値をヘッダマージ部８０４に出力する。これらのフィールドの値は、図２１に関して説明したとおり、それぞれ、「有効」を示す論理値１、転送データパケットを示す値、CPU１０４aのID、「共有型」を示す値である。また、以上４つのフィールドの値を、以下ではまとめて複製情報６０７とよぶことがある。

[0434] また、ステップＳ９０５で管理テーブル部８０１はさらに、リードポインタ値が示す「００００」なる番地用のエントリ９０１－０の複製データVALIDフィールド９１７の値である論理値１をセクタ８０５に出力する。そして、図１８Ａに示したように管理テーブル部８０１は、FF９０５の出力をリセット信号として用いて、エントリ９０１－０の複製データVALIDフィールド９１７を、「無効」を示す論理値０にリセットする。

[0435] なお、ステップＳ９０４とＳ９０５の実行順は任意であり、ステップＳ９０４とＳ９０５が並列に実行されてもよい。

続いて、ステップＳ９０６でヘッダマージ部８０４は、ステップＳ９０４でデータパケット保存部４１０から出力されたライトバックデータパケット６０１のヘッダに、ステップＳ９０５で管理テーブル部８０１から出力された複製情報６０７を追加する。ヘッダマージ部８０４は、複製情報６０７の追加により、複製情報を持ったライトバックデータパケット６０８を生成し、セクタ８０５に出力する。

- [0436] 例えば、ライトバックデータパケット601のヘッダは、パケットタイプ、宛先ID、要求元ID、ReqID、データステート、アドレス、その他のフィールドを含む。また、ヘッダの末尾には1つ以上の予約(reserved)フィールドがあってもよい。なお、ヘッダ内のアドレスフィールドは、データペイロードに含まれるデータが記憶されているメインメモリのアドレスを示す。
- [0437] ヘッダマージ部804は、ライトバックデータパケット601のヘッダの所定の位置に複製情報607を挿入する。例えば、図22の例では「所定の位置」とはヘッダの末尾の予約フィールドの直前の位置だが、「所定の位置」は実施形態に応じて任意に定めることができる。
- [0438] 複製情報を持ったライトバックデータパケット608のヘッダでは、図22に示すように、ライトバックデータパケット601に元々含まれていたパケットタイプ、宛先ID、要求元ID、ReqID、データステート、アドレス、およびその他のフィールドの後に、複製情報607に相当する4つのフィールドが続く。そして最後に、ライトバックデータパケット601に元々含まれていた予約フィールドがある。複製情報を持ったライトバックデータパケット608のデータペイロードは、ライトバックデータパケット601のデータペイロードそのものである。
- [0439] 換言すれば、複製情報を持ったライトバックデータパケット608は、ライトバックデータパケット601のヘッダの情報と、共有型・転送データパケット603のヘッダの情報とを含む。さらに、複製情報を持ったライトバックデータパケット608は、ライトバックデータパケット601と共有型・転送データパケット603に共通のデータペイロードを含む。
- [0440] ここで、セクタ805には、データパケット保存部410からライトバックデータパケット601が入力され、ヘッダマージ部804から複製情報を持ったライトバックデータパケット608が入力されている。ステップS907では、セクタ805が、ステップS905で管理テーブル部801から出力された選択信号としての複製データVALIDフィールド917の

値にしたがって、複製情報を持ったライトバックデータパケット608を選択し、パケット間セクタ430に出力する。

[0441] そして、ステップS908においてパケット間セクタ430は、ステップS902で入力された選択信号に基づいて、データパケットを選択する。すなわち、パケット間セクタ430は、セクタ805からの入力である複製情報を持ったライトバックデータパケット608を選択してポート間セクタ508xへと出力する。

[0442] なお、ポート間アービタ部507xは、ステップS901で出力バッファ制御部702cxに対する出力許可を出力した時点で、出力バッファ制御部702cxを選択するよう指示する選択信号をポート間セクタ508xに出力している。したがって、ステップS908でパケット間セクタ430からポート間セクタ508xに出力された複製情報を持ったライトバックデータパケット608は、ポート間セクタ508xで選択され、図15の出力ポート502xへ出力される。そして、複製情報を持ったライトバックデータパケット608は、出力ポート502xからバス108cを介して、図16に示すXB102の入力ポート201cへと出力される。

[0443] ここで、比較例においては、図12と図13に示すように、データパケット保存部410に個々に保存されていたライトバックデータパケット601と共有型・転送データパケット603が個々にSC103cからバス108cを介してXB102へ出力される。それに対し、第1実施形態では、上記のとおり、1つの複製情報を持ったライトバックデータパケット608がSC103cからバス108cを介してXB102へ出力されるだけである。したがって、第1実施形態によれば、データパケットの出力によるバス108cの占有時間を比較例の約2分の1に短縮することができる。

[0444] なお、複製情報を持ったライトバックデータパケット608は、追加された複製情報607のぶん、ライトバックデータパケット601や共有型・転送データパケット603よりもパケット長が長い。しかし、一般には、ヘッダの長さがデータペイロードの長さより十分に短いようなパケット形式が採

用されることが多い。したがって、複製情報を持ったライトバックデータパケット608の長さは、ライトバックデータパケット601の長さと共有型・転送データパケット603の長さの和の、約2分の1である。

[0445] バス108cの占有時間が短縮されると、図15のSC103cにおいて各出力バッファ制御部702cx、702dx、509ix、および509sxがポート間アービタ部507xへ出力リクエストを出力してから出力許可が下りるまでの待ち時間が短縮される。したがって、第1実施形態によれば、比較例と比べて、サーバシステム100全体としてのレイテンシおよびスループットが改善される。つまり、第1実施形態によれば、サーバシステム100の性能が向上する。

[0446] 続いて、図22のようにしてXB102へ出力された、複製情報を持ったライトバックデータパケット608の、XB102における復元について、XB102内のコンポーネントの詳細とともに、図23～図29を参照して説明する。

[0447] 図23は、第1実施形態における複製チェック部1001、データパケット復元部1002、セレクタ1003、およびルーティング部713の構成を示す図である。図23は、図16に示すXB102においてSC103cに対応する部分を抜粋して示した図である。

[0448] つまり、図23は、図16においてSC103cと接続された入力ポート201cからルーティング部713へ至る経路上に設けられた複製チェック部1001とデータパケット復元部1002とセレクタ1003とを示している。また、図23には、ルーティング部713の一部（具体的には、入力ポート201cが受け取ったパケットのルーティングに関わる部分）を抜粋して示してある。

[0449] なお、図23と同様の複製チェック部1001とデータパケット復元部1002とセレクタ1003が、図16において他のSC103a、103e、および103gに対応して設けられている。また、ルーティング部713は、4つのSC103a～103gに関して対称的に構成されている。

[0450] さらに、図 23 と同様の複製チェック部 1001 とデータパケット復元部 1002 とセレクタ 1003 が、図 15 において入力ポート 501x からルーティング部 703 に至る経路上に設けられている。また、SC103c 内のルーティング部 703 の詳細の図示は省略したが、ルーティング部 713 の構成とともにルーティング部 703 の構成についても後述する。

[0451] 図 23 において、SC103c と接続された入力ポート 201c からルーティング部 713 へ至る経路上には、複製チェック部 1001 とデータパケット復元部 1002 とセレクタ 1003 とが設けられている。また、複製チェック部 1001 はルーティングテーブル部 711c と接続されている。ルーティングテーブル部 711c の詳細は、図 25 とともに後述する。

[0452] 複製チェック部 1001 は、パケットタイプチェック部 1004 と複製チェック部 1005 と比較器 1006 を備える。また、データパケット復元部 1002 は複製データパケット修正部 1007 とヘッダ復元部 1008 とデータパケット生成部 1009 を備える。

[0453] そして、ルーティング部 713 は、比較例における図 6 のルーティング部 203 と同様の分岐部 306 を備え、さらに、分岐部 1010、OR 回路 1011、OR 回路 1012、および OR 回路 1013 を備える。ただし、図 6 の分岐部 306 は出力バッファ制御部 207ca、207ce、および 207cg に接続されているが、分岐部 306 は OR 回路 1011、1012、および 1013 に接続されている。そして、OR 回路 1011、1012、および 1013 がそれぞれに対応する出力バッファ制御部 712ca、712ce、および 712cg に接続されている。

[0454] 入力ポート 201c から入力されたパケットの全部または一部は、以下のようにして各部に入力される。

入力ポート 201c に接続された信号線 1021 には、入力ポート 201c が SC103c から受け取ったパケットが出力される。信号線 1021 は信号線 1022 と信号線 1023 に分岐している。信号線 1022 は分岐してセレクタ 1003 と複製データパケット修正部 1007 に接続されており

、信号線 1022 を介してパケットがセクタ 1003 と複製データパケット修正部 1007 に出力される。

[0455] また、信号線 1021 から分岐した信号線 1023 を介して、複製チェック部 1001 にはパケットに含まれるヘッダ情報が出力される。信号線 1023 はさらに信号線 1024 と信号線 1025 に分岐し、信号線 1025 は信号線 1026 と信号線 1027 に分岐する。信号線 1024 を介してヘッダ情報のうちパケットタイプフィールドの値がパケットタイプチェック部 1004 に出力される。

[0456] また、信号線 1026 を介してヘッダ情報のうち複製データ VALID フィールドの値が複製チェック部 1005 に出力され、信号線 1027 を介してヘッダ情報のうち宛先 ID フィールドと複製宛先フィールドの値がルーティングテーブル部 711c に出力される。

[0457] なお、入力ポート 201c が受け取ったパケットは、例えば要求パケットである場合などでは、複製データ VALID フィールドや複製パケット宛先 ID フィールドを持たない。その場合、信号線 1026 を介して複製チェック部 1005 には、ダミーの値として論理値 0 が出力される。また、信号線 1027 を介してルーティングテーブル部 711c には、例えば、宛先 ID として存在しないダミーの値が出力されてもよい。例えば、すべてのビットが論理値 0 の値を宛先 ID として用いないことが予め決められていれば、ダミーの値としてすべてのビットが論理値 0 の値を用いることができる。

[0458] このようなダミーの値の出力は、パケットの形式を定める仕様に応じて、適宜実現可能である。

例えば、任意の用途に使用してよいオプション領域がヘッダに含まれる形式のパケットが、サーバシステム 100 において用いられてもよい。そして、図 22 に示した複製データ VALID、複製パケットタイプ、複製パケット宛先 ID、および複製データステートの各フィールドは、オプション領域を使ったものであってもよい。例えば、オプション領域は、未使用の場合には「0」で埋められると仕様で決まってもよい。

- [0459] この場合、例えば、信号線 1023 が信号線 1021 から分岐する点が例えば上記（e2）のように実現されていれば、上記のような必要に応じたダミーの値の出力が実現される。
- [0460] あるいは、ヘッダ長を示すフィールドを含む形式の packets が用いられ、例えば上記（e3）で説明したように、信号線 1023 が信号線 1021 から分岐する点に不図示の回路素子があってもよい。当該回路素子は、ヘッダ長を示すフィールドの値に応じて、入力ポート 201c から出力された packets に複製データ VALID フィールドや複製 packet 宛先 ID フィールドがあるか否かを判断し、必要に応じてダミーのデータを出力してもよい。
- [0461] また、分岐した信号線 1022 は、上記のとおり複製データ packet 修正部 1007 に接続するだけでなく、さらに信号線 1028 と信号線 1029 へと分岐し、それぞれヘッダ復元部 1008 とデータ packet 生成部 1009 へと接続している。信号線 1022 を介して出力される packets のうち、ヘッダが信号線 1028 を介してヘッダ復元部 1008 へと出力され、データペイロードがデータ packet 生成部 1009 へと出力される。
- [0462] 以上、複製チェック部 1001 とデータ packet 復元部 1002 とセレクタ 1003 とルーティング部 713 の構成の概要を説明するとともに、図 23 において入力ポート 201c と直接接続されている各コンポーネントへの入力について説明した。続いて、図 23 に示した各コンポーネントの動作について、図 24 を参照して詳しく説明する。
- [0463] 図 24 は、第 1 実施形態における複製チェック部、データ packet 復元部、およびセレクタの動作を説明する図である。
- まず、メモリアクセス要求 packet や、データ転送要求 packet などの要求 packet（すなわち、データペイロードを持たない packet）が入力ポート 201c に入力された場合の動作を説明する。要求 packet を処理する動作の流れは、図 24 において実線矢印により示される。
- [0464] ステップ S1001r において、入力ポート 201c が要求 packet を出力する。

するとステップS 1 0 0 2 rにおいて、要求パケットのヘッダ内の宛先 I Dフィールドがルーティングテーブル部 7 1 1 cに出力されるとともに、ヘッダ内のパケットタイプフィールドがパケットタイプチェック部 1 0 0 4に出力される。また、上記のように複製データ V A L I Dフィールドと複製パケット宛先 I Dフィールドのダミーの値が出力されてもよい。

[0465] ここで、パケットタイプチェック部 1 0 0 4は、図示したように、入力されたパケットタイプの値がデータパケットを示す値であるか否かを判断する判断部 1 1 0 1を備えている。ステップS 1 0 0 3 rにおいて判断部 1 1 0 1は、入力されたパケットタイプの値をエンコードすることで、入力されたパケットがデータパケットであると判断する。

[0466] するとステップS 1 0 0 4 rで判断部 1 1 0 1は、セクタ 1 0 0 3に対し、複製データパケット修正部 1 0 0 7からの入力ではなく、図 2 3の信号線 1 0 2 2を介して入力ポート 2 0 1 cから入力されたパケットをそのまま出力するよう命令する選択信号を出力する。選択信号にしたがい、セクタ 1 0 0 3は入力ポート 2 0 1 cから入力されたパケットをルーティング部 7 1 3の分岐部 3 0 6に出力する。

[0467] また、入力されたパケットタイプの値が要求パケットを示すので、ステップS 1 0 0 4 rで判断部 1 1 0 1はさらに、複製チェック部 1 0 0 5へ出力する V A L I D信号をネゲートする。

[0468] また、ステップS 1 0 0 5 rでルーティングテーブル部 7 1 1 cは、図 2 3の信号線 1 0 2 7を介して入力された宛先 I Dに基づいて、入力されたパケットのルーティング先を決定し、出力ポート情報をルーティング部 7 1 3の分岐部 3 0 6に出力する。ステップS 1 0 0 5 rは、ステップS 1 0 0 3 r～S 1 0 0 4 rと並列に実行されてもよい。

[0469] そして、分岐部 3 0 6は、出力ポート情報に基づいて、入力された要求パケットをO R回路 1 0 1 1、1 0 1 2、または1 0 1 3に出力する。

O R回路 1 0 1 1、1 0 1 2、または1 0 1 3を介してのパケットのルーティングの詳細は図 2 6とともに後述するが、概略を述べると次のとおりで

ある。例えば、出力先が図 16 の出力ポート 202c である要求パケットは、ステップ S1005r で決定された出力ポート情報にしたがい、出力ポート 202c の前段の出力バッファ制御部 712ca と接続された OR 回路 1011 へと出力される。このとき、後述するように分岐部 1010 から OR 回路 1011 へは論理値 0 で埋められたデータが出力される。したがって、分岐部 306 が出力した要求パケットが OR 回路 1011 を介して出力バッファ制御部 712ca へと出力される。

[0470] また、宛先 ID から出力ポート情報を取得するためのルーティングテーブル部 711c の内部の構成は、図 6 のルーティングテーブル部 204c と類似である。ルーティングテーブル部 711c の構成の詳細については、図 25 とともに後述する。

[0471] 以上のとおり、入力ポート 201c から要求パケットが出力された場合は、データパケット復元部 1002 を使わずにルーティングが行われる。

続いて、ライトバックデータパケットや転送データパケットなどのデータパケット（すなわち、データペイロードを持つパケット）が入力ポート 201c に入力された場合の動作を説明する。データパケットを処理する動作の流れは、図 24 において 1 点鎖線矢印により示される。

[0472] ステップ S1001d において、入力ポート 201c がデータパケットを出力する。

するとステップ S1002d において、データパケットのヘッダ内の宛先 ID フィールドと複製パケット宛先 ID フィールドがルーティングテーブル部 711c に出力される。また、ヘッダ内のパケットタイプフィールドがパケットタイプチェック部 1004 に出力される。さらに、複製データ VALID フィールドが複製チェック部 1005 に出力される。なお、上記のように、入力されたデータパケットが複製データ VALID フィールドや複製パケット宛先 ID フィールドを持たない場合は、ダミーの値が使われてもよい。

[0473] そして、ステップ S1003d ではパケットタイプチェック部 1004 内

の判断部 1101 が、入力されたパケットがデータパケットであると判断する。

するとステップ S1004d で判断部 1101 は、入力されたパケットがデータパケットであることを示す VALID 信号を複製チェック部 1005 に出力（アサート）する。第 1 実施形態では、VALID 信号は正論理の信号なので、ステップ S1004d では論理値 1 が複製チェック部 1005 に出力される。なお、データパケットが入力された場合のステップ S1004d では、判断部 1101 は、セクタ 1003 への選択信号を無効な値に設定する。それにより、判断部 1101 は、後述する比較器 1006 からの選択信号のみにしたがって動作するよう、セクタ 1003 に命令する。

[0474] そして、ステップ S1005d で複製チェック部 1005 は、「入力されたデータパケットからの復元が必要か否かを判断するための比較を、比較器 1006 が行う必要があるか否か」を判断する。ここで、ステップ S1005d の判断について説明するため、データパケットが入力される場合を次の（j1）～（j3）に場合分けする。

[0475] （j1）入力されたパケットが有効な複製情報を持たないパケットである場合。

例えば、SC103c 内の出力バッファ制御部 702cx において、図 20 の処理の後、かつ、図 21 と同様の共有型・転送データパケット 603 が入力される前に、ポート間アービタ部 507x から出力許可が出力されたとする。すると、図 22 のような複製情報 607 の追加が行われないので、XB102 内の入力ポート 201c に入力されたデータパケットは、有効な複製情報を持たない。

[0476] （j2）入力されたパケットが有効な複製情報を持ち、かつ、宛先 ID フィールドに対応する出力ポートと、複製宛先 ID フィールドに対応する出力ポートが等しい場合。

例えば、CPU104e を宛先とする共有型・転送データパケットと SC103e を宛先とするライトバックデータパケットから、図 20～図 22 と

類似の処理により、複製情報を持ったデータパケットが生成されることがある。こうして生成されたパケットが入力ポート 201c に入力された場合、宛先 ID フィールドに対応する出力ポートと複製宛先 ID フィールドに対応する出力ポートは、ともに図 16 の出力ポート 202e であり、等しい。

[0477] (j3) 入力されたパケットが有効な複製情報を持ち、かつ、宛先 ID フィールドに対応する出力ポートと、複製宛先 ID フィールドに対応する出力ポートが異なる場合。

例えば、図 20～図 22 の処理により生成された、図 22 の複製情報を持ったライトバックデータパケット 608 が、入力ポート 201c に入力された場合が、(j3) には含まれる。

[0478] 第 1 実施形態では、上記 (j1) と (j2) の場合には、入力されたパケットそのものをルーティング部 713 がルーティングする。なぜなら、(j1) の場合はそもそも復元の必要がないためである。また、(j2) の場合は、より後の時点で復元する方が、XB102 の内部バスや、XB102 と SC との間のバス（上記の例では SC103e との間のバス 108e）の占有時間を短縮することができるためである。

[0479] 逆に、(j3) の場合には、入力されたパケットからデータパケット復元部 1002 が 2 つのパケットを復元し、復元された 2 つのパケットをルーティング部 713 がルーティングする。つまり、(j3) の場合にのみ復元が必要である。

[0480] ここで、入力されたパケットの複製 VALID フィールドの値が「無効」を示す場合は、上記 (j1) に相当し、そもそもパケットの復元とは無関係である。よって、比較器 1006 による比較を行うまでもなく、復元は不要であると直ちに判明する。

[0481] しかし、入力されたパケットの複製 VALID フィールドの値が「有効」を示す場合は、上記 (j2) と (j3) のいずれであるのかによって、復元の要否が異なる。そして、上記 (j2) と (j3) を弁別するには、宛先 ID フィールドに対応する出力ポートと、複製宛先 ID フィールドに対応する

出力ポートを比較する必要がある。つまり、入力されたパケットの復元が必要か否かを判断するためには、比較器 1006 による比較が必要である。

[0482] 比較器 1006 による比較が必要であるか否かを決定するため、複製チェック部 1005 は、図 24 に示すように AND 回路 1102 を備えている。AND 回路 1102 への入力、図 23 の信号線 1026 を介して入力される複製 VALID フィールドの値と、パケットタイプチェック部 1004 の判断部 1101 が出力する VALID 信号である。なお、要求パケットが入力された場合は、比較器 1006 による比較は当然不要である。

[0483] ステップ S1005d において、AND 回路 1102 の出力は、上記 (j1) の場合に論理値 0 となり、上記 (j2) または (j3) の場合に論理値 1 となる。AND 回路 1102 の出力は比較器 1006 に入力される。

[0484] すなわち、比較器 1006 による比較が必要なとき、複製チェック部 1005 の AND 回路 1102 は、正論理の VALID 信号を比較器 1006 に出力する。

また、ステップ S1006d に示すように、ルーティングテーブル部 711c は、ステップ S1002d で入力された宛先 ID フィールドと複製宛先 ID フィールドの値に基づいて、出力ポート情報と複製データパケット出力ポート情報を出力する。

[0485] なお、「出力ポート情報」とは、比較例と同様に、宛先 ID フィールドで指定された宛先に対応する出力ポートを識別する情報である。また、「複製データパケット出力ポート情報」とは、複製パケット宛先 ID フィールドで指定された宛先に対応する出力ポートを識別する情報である。

[0486] ここで、ルーティングテーブル部 711c の詳細な構成を示した図 25 を参照して、ステップ S1006d におけるルーティングテーブル部 711c の動作を説明する。

図 25 は、第 1 実施形態におけるルーティングテーブル部 711c の構成図である。ルーティングテーブル部 711c は、図 16 および図 23 に示すように、SC103c から他の SC へのルーティングを管理するためのコン

ポーネントである。

[0487] ルーティングテーブル部 711c は、比較例において図 6 に示したルーティングテーブル部 204c と同様の FF307a ~ 307i と、比較器 308a ~ 308i と、OR 回路 309a ~ 309c と、セクタ 310 と、FF311a ~ 311c を備える。比較例と同様に、FF307a ~ 307i はそれぞれパケットの宛先となりうる CPU104a などのコンポーネントの ID を格納しており、FF311a ~ 311c は出力ポート 202a、202e、202g を識別する情報を格納している。

[0488] したがって、ルーティングテーブル部 711c は、入力された宛先 ID フィールドの値に応じて、比較例のルーティングテーブル部 204c と同様に、セクタ 310 から出力ポート情報を出力する。

[0489] また、ルーティングテーブル部 711c はさらに、12 個の比較器 1201a ~ 1201i と、3 個の OR 回路 1202a ~ 1202c と、セクタ 1203 を備える。

12 個の比較器 1201a ~ 1201i には、12 組の FF307a ~ 307i が保持する値がそれぞれ入力される。また、12 個の比較器 1201a ~ 1201i の各々には、複製宛先 ID フィールドの値が入力される。比較器 1201a ~ 1201i は、それぞれ 2 つの入力を比較し、2 つの入力が一致すれば論理値 1 を出力し、2 つの入力が一致しなければ論理値 0 を出力する。

[0490] OR 回路 1202a は、比較器 1201a ~ 1201d の出力の論理和をセクタ 1203 に出力する。OR 回路 1202b は、比較器 1201e ~ 1201h の出力の論理和をセクタ 1203 に出力する。OR 回路 1202c は、比較器 1201i ~ 1201l の出力の論理和をセクタ 1203 に出力する。

[0491] したがって、セクタ 1203 には、最大 1 ビットが論理値 1 となる合計 3 ビットの制御信号が入力される。セクタ 1203 は、論理値 1 のビットの位置に応じて、FF311a ~ 311c が保持する出力ポート情報のいず

れかを選択し、複製データパケット出力ポート情報として出力する。

[0492] すなわち、セクタ 1203 は、OR 回路 1202 a の出力が論理値 1 のとき、FF311a が保持する値を選択する。また、セクタ 1203 は、OR 回路 1202 b の出力が論理値 1 のとき、FF311b が保持する値を選択し、OR 回路 1202 c の出力が論理値 1 のとき、FF311c が保持する値を選択する。

[0493] このように、図 25 のルーティングテーブル部 711c は、比較例と同様に宛先 ID から出力ポート情報を取得する回路と、複製宛先 ID から複製データパケット出力ポート情報を取得する回路とを有し、後者の構成は前者と同様である。また、前者と後者の回路は、宛先のコンポーネントを識別する情報を保持する FF307a などの記憶素子と、ポートを識別する情報を保持する FF311a などの記憶素子を共有している。図 15 のルーティングテーブル部 701x も図 25 と同様に構成されている。

[0494] ここで図 24 の説明に戻ると、ステップ S1006d でルーティングテーブル部 711c は、出力ポート情報を比較器 1006 と分岐部 306 に出力するとともに、複製データパケット出力ポート情報を比較器 1006 と分岐部 1010 に出力する。なお、ステップ S1006d は、ステップ S1004d ~ S1005d と並列に実行されてもよい。

[0495] 続いて、ステップ S1007d で比較器 1006 は、複製チェック部 1005 からの VALID 信号、ならびにルーティングテーブル部 711c からの出力ポート情報および複製データパケット出力ポート情報に基づいて、比較操作を実行する。比較器 1006 は、比較結果を選択信号としてセクタ 1003 と分岐部 1010 に出力する。

[0496] つまり、複製チェック部 1005 から VALID 信号がアサートされており、かつ、出力ポート情報と複製データパケット出力ポート情報が一致しない場合、比較器 1006 は、データパケット復元部 1002 からの出力を選択するよう指示する選択信号を出力する。それ以外の場合、比較器 1006 は、データパケット復元部 1002 からの出力を選択しないよう指示する選

択信号を出力する。

[0497] 換言すれば、比較器 1006 は、上記（j 3）の場合、データパケット復元部 1002 からの出力を選択するよう指示する選択信号を出力する。また、比較器 1006 は、上記（j 1）または（j 2）の場合、データパケット復元部 1002 からの出力を選択しないよう指示する選択信号を出力する。

[0498] なお、データパケットが入力された場合、パケットタイプチェック部 1004 からセクタ 1003 への出力は上記のとおり無効な値なのでセクタ 1003 では無視される。よって、セクタ 1003 は比較器 1006 からの選択信号のみにしたがって、入力ポート 201c から図 23 の信号線 1022 を介して入力されたパケットまたはデータパケット復元部 1002 内の複製データパケット修正部 1007 から入力されたパケットを選択する。そして、セクタ 1003 は、選択したパケットを分岐部 306 に出力する。

[0499] また、分岐部 1010 は、データパケット復元部 1002 からの出力を選択しないよう指示された場合、すべてのビットを論理値 0 としたデータを出力する。

続いて、上記ステップ S1002d ~ S1007d と並行してデータパケット復元部 1002 で実行されるステップ S1008d ~ S1010d について説明する。なお、以下で説明するステップ S1008d ~ S1010d は、上記（j 3）の場合の動作である。上記（j 3）の場合以外では、下記の理由から、データパケット復元部 1002 の動作は後段のセクタ 1003 およびルーティング部 713 の動作に影響しない。

[0500] 要求パケットが入力された場合または上記（j 1）の場合には、セクタ 1003 でデータパケット復元部 1002 からの出力が選択されることがない。そして、データパケット復元部 1002 からの出力を選択しないよう指示する選択信号を比較器 1006 が分岐部 1010 に出力している。

[0501] また、上記（j 2）の場合は、下記と同様にデータパケット復元部 1002 での復元が行われる。しかし、上記のとおり分岐部 1010 の出力の各ビットは論理値 0 であるため、データパケット復元部 1002 からの出力は、

分岐部 1010 を介して OR 回路 1011、1012、あるいは 1013 へ伝達されることはない。

[0502] よって、以下では (j 3) の場合に注目してデータパケット復元部 1002 の動作を説明する。

上記 (j 3) の場合、ステップ S 1001 d で入力ポート 201 c から出力されるのは、複製情報を持ったデータパケットである。そこで、ステップ S 1008 d では、複製データパケット修正部 1007 に、入力ポート 201 c から出力されたデータパケットが出力される。また、ヘッダ復元部 1008 には、入力ポート 201 c から出力されたデータパケットのヘッダが出力される。そして、データパケット生成部 1009 には、入力ポート 201 c から出力されたデータパケットのデータペイロードが出力される。

[0503] すると、ステップ S 1009 d で複製データパケット修正部 1007 は、入力されたパケットのヘッダから複製情報を削除したパケットを生成してセレクタ 1003 に出力する。つまり、複製データパケット修正部 1007 は、図 17 のヘッダマージ部 804 でマージされる前の元の 2 つのパケットのうち、先に出力バッファ制御部（例えば 702 c x）のデータパケット保存部 410 に保存されていた方のパケットを復元する。

[0504] また、ステップ S 1010 d でヘッダ復元部 1008 は、入力されたヘッダから、復元しようとする元のデータパケットのヘッダを生成して復元する。つまり、ヘッダ復元部 1008 は、元の 2 つのパケットのうち、後から出力バッファ制御部（例えば 702 c x）に入力されて破棄された方のパケットのヘッダを復元する。ヘッダ復元部 1008 は、復元したヘッダをデータパケット生成部 1009 に出力する。なお、ステップ S 1009 d と S 1010 d は並列に実行されてもよい。

[0505] そして、ステップ S 1011 d でデータパケット生成部 1009 は、ヘッダ復元部 1008 から入力されたヘッダと、図 23 の信号線 1029 を介して入力されたデータペイロードとをマージする（すなわち連結する）ことによりパケットを生成する。すなわち、データパケット生成部 1009 は、元

の2つのパケットのうち、後から出力バッファ制御部（例えば702c x）に入力されて破棄された方のパケットを復元する。データパケット生成部1009は、復元したパケットをルーティング部713の分岐部1010に出力する。

[0506] 以上から、上記（j 3）の場合は、複製データパケット修正部1007からの出力がセクタ1003と分岐部306を介してルーティングされ、データパケット生成部1009からの出力が分岐部1010を介してルーティングされる。

[0507] また、上記（j 3）の場合には、分岐部306におけるルーティング先と分岐部1010におけるルーティング先が異なる。よって、復元された2つのパケットがともに同じ1つのOR回路（1011～1013のいずれか）に入力されることはありえない。そのため、第1実施形態では、セクタではなくより構造が単純なOR回路を用いてルーティング部713を構成することが可能となっている。

[0508] 続いて、図20～図22の処理によりSC103c内の出力バッファ制御部702c xから出力された、複製情報を持ったライトバックデータパケット608が、XB102の入力ポート201cに入力されたという具体例について、図26を参照して説明する。

[0509] 図26は、第1実施形態において複製情報を持ったライトバックデータパケット608から元の2つのパケットを復元して2つのパケットをルーティングする動作の具体例を説明する図である。

[0510] 図22の処理によって生成された、複製情報を持ったライトバックデータパケット608が、ステップS1101でXB102の入力ポート201cから出力される。

すると、複製情報を持ったライトバックデータパケット608の全体が複製データパケット修正部1007へ出力される。また、複製情報を持ったライトバックデータパケット608のヘッダのみがヘッダ復元部1008へ出力される。そして、複製情報を持ったライトバックデータパケット608の

データペイロードがデータパケット生成部 1009 へ出力される。

[0511] また、ステップ S 1102 では、複製情報を持ったライトバックデータパケット 608 のヘッダ内のいくつかのフィールドの値が下記のように出力される。

- ・宛先 ID フィールドの値と複製パケット宛先 ID フィールドの値がルーティングテーブル部 711c に出力される。図 22 に示すとおり、宛先 ID フィールドは SC 103e の ID を示し、複製パケット宛先 ID フィールドは CPU 104a の ID を示す。

[0512] ・パケットタイプフィールドの値がパケットタイプチェック部 1004 に出力される。図 22 に示すとおり、パケットタイプフィールドは、ライトバックデータパケットであることを示す。

[0513] ・複製データ VAL ID フィールドの値が複製チェック部 1005 に出力される。図 22 に示すとおり、複製データ VAL ID フィールドの値は、「有効」を示す論理値 1 である。

[0514] すると、ステップ S 1103 でパケットタイプチェック部 1004 は、入力されたパケットタイプをエンコードし、その結果、ステップ S 1101 で入力された複製情報を持ったライトバックデータパケット 608 がデータパケットであると判断する。よって、パケットタイプチェック部 1004 は、複製チェック部 1005 に VAL ID 信号を出力する。

[0515] すると、ステップ S 1104 で複製チェック部 1005 は、パケットタイプチェック部 1004 から入力された VAL ID 信号と、ステップ S 1102 で入力された複製データ VAL ID の論理積を求め、比較器 1006 に出力する。換言すれば、この例では求めた論理積の値は論理値 1 であるから、複製チェック部 1005 は比較器 1006 に VAL ID 信号を出力（アサート）する。

[0516] また、ステップ S 1105 に示すようにルーティングテーブル部 711c は、入力された宛先 ID（すなわち SC 103e の ID）から、出力ポート 202e を識別する情報を出力ポート情報として取得し、出力ポート情報を

分岐部 306 に出力する。さらに、ルーティングテーブル部 711c は、入力された複製パケット宛先 ID（すなわち CPU 104a の ID）から、出力ポート 202a を識別する情報を複製データパケット出力ポート情報として取得し、分岐部 1010 に出力する。

[0517] また、ステップ S 1106 で比較器 1006 は、ルーティングテーブル部 711c から出力された出力ポート情報と複製データパケット出力ポート情報を比較する。その結果、出力ポート情報と複製データパケット出力ポート情報が不一致であるため、比較器 1006 は、データパケット復元部 1002 から出力されるパケットを選択するようにセレクタ 1003 と分岐部 1010 に選択信号を出力する。

[0518] また、ステップ S 1107 で複製データパケット修正部 1007 は、図 22 に示した複製情報を持ったライトバックデータパケット 608 のヘッダから、追加された複製情報 607 を削除することにより、図 20 のライトバックデータパケット 601 を復元する。以下では元のパケットと復元されたパケットを区別するため、復元後のライトバックデータパケットの参照符号を「609」とする。

[0519] 複製データパケット修正部 1007 は、復元したライトバックデータパケット 609 をセレクタ 1003 に出力する。

ステップ S 1107 の詳細を図 27 に示す。すなわち、図 27 は、第 1 実施形態における複製データパケット修正部 1007 の動作の具体例を説明する図である。

[0520] 図 22 に示したように、複製情報を持ったライトバックデータパケット 608 のヘッダの所定の位置（具体的には、ヘッダ末尾の予約フィールドの直前の位置）には、複製情報 607 が含まれる。複製情報 607 は、具体的には、複製データ VALID、複製パケットタイプ、複製パケット宛先 ID、複製データステートという 4 個のフィールドを有する。ステップ S 1107 において複製データパケット修正部 1007 は、複製情報を持ったライトバックデータパケット 608 からヘッダ内の複製情報 607 を削除することに

よってライトバックデータパケット609を生成する。

[0521] ここで図26の説明に戻ると、ステップS1108でヘッダ復元部1008は、入力されたヘッダ内から、図21に示した元の共有型・転送データパケット603のヘッダを生成し、データパケット生成部1009に生成したヘッダを出力する。

[0522] ステップS1108の詳細を図28に示す。図28は、第1実施形態におけるヘッダ復元部1008の動作の具体例を説明する図である。

図28に示すように、ヘッダ復元部1008は、複製情報を持ったライトバックデータパケット608のヘッダ内のフィールドの一部を取り出して適切に並べ替えることで、図21の元の共有型・転送データパケット603のヘッダを復元する。

[0523] 具体的には、ヘッダ復元部1008は、複製情報を持ったライトバックデータパケット608のヘッダから、下記の各フィールドを取り出し、下記のリスト順に並べることで、元の共有型・転送データパケット603のヘッダを生成し、復元する。なお、下記リストにおいて括弧内に値を示した。

- [0524] ・複製パケットタイプ（転送データパケットであることを示す値）
・複製パケット宛先ID（CPU104aのID）
・要求元ID（CPU104aのID）
・ReqID（5）
・複製データステート（「共有型」を示す値）
・アドレスその他の不図示のフィールド
・末尾の予約フィールド

ここで、複製情報を持ったライトバックデータパケット608の複製パケットタイプ、複製パケット宛先ID、および複製データステートは、元の共有型・転送データパケット603のパケットタイプ、宛先ID、およびデータステートを複製したものである。

[0525] また、図20～図22に示したように、元のライトバックデータパケット601と共有型・転送データパケット603は、同じ1つの転送データ要求

パケットに起因してCPU 104cにより出力される。したがって、両者において必然的に、要求元ID、ReqID、およびアドレスという3つのフィールドの値は等しい。また、第1実施形態では、その他の予約フィールドや特に図示しないフィールドの値も、ライトバックデータパケット601と共有型・転送データパケット603で共通であるとする。

[0526] したがって、ヘッダ復元部1008は、上記のように複製情報を持ったライトバックデータパケット608のヘッダ内のフィールドの一部を取り出して必要に応じて並べ替えることで、共有型・転送データパケット603のヘッダを復元することができる。ヘッダ復元部1008は、復元したヘッダ611をデータパケット生成部1009に出力する。

[0527] ここで図26の説明に戻ると、ステップS1109でデータパケット生成部1009は、ヘッダ復元部1008から出力された、復元したヘッダ611と、入力ポート201cから出力されたデータペイロードとをマージする。それにより、データパケット生成部1009は、元の共有型・転送データパケット603を復元する。以下では元のパケットと復元されたパケットを区別するため、復元後の共有型・転送データパケットの参照符号を「613」とする。

[0528] ステップS1109の詳細を図29に示す。図29は、第1実施形態におけるデータパケット生成部1009の動作の具体例を説明する図である。

図29に示すように、データパケット生成部1009には、復元したヘッダ611とデータペイロード612が入力される。データパケット生成部1009は、復元したヘッダ611の後にデータペイロード612を連結することで共有型・転送データパケット613を生成し、共有型・転送データパケット613を分岐部1010に出力する。

[0529] ここで図26の説明に戻る。ステップS1110でセクタ1003は、ステップS1006で入力された選択信号にしたがい、ステップS1107でデータパケット復元部1002の複製データパケット修正部1007から入力されたライトバックデータパケット609を選択する。セクタ100

3は、選択したライトバックデータパケット609をルーティング部713の分岐部306に出力する。

[0530] すると、ステップS1111で分岐部306は、ステップS1105で入力された出力ポート情報にしたがい、入力されたライトバックデータパケット609を、OR回路1012に出力する。

[0531] 上記のとおり、複製情報を持ったライトバックデータパケット608の宛先IDであるSC103eのIDから得られた出力ポート情報は、SC103eと接続された出力ポート202eを識別する情報である。したがって、分岐部306は、入力されたライトバックデータパケット609を、出力ポート202eの前段に設けられた出力バッファ制御部712ceと接続されたOR回路1012に出力する。

[0532] なお、分岐部306は、ルーティングテーブル部711cからの出力ポート情報により選択しなかったOR回路へは、パケットのビット数分の論理値0を出力する。また、分岐部1010は、データパケット復元部1002からの出力を選択するよう比較器1006から命令された場合、複製データパケット出力ポート情報により選択しなかったOR回路へは、パケットのビット数分の論理値0を出力する。そして、上記のとおり、(j3)の場合、出力ポート情報と複製データパケット出力ポート情報は排他的である。

[0533] したがって、OR回路1012は、分岐部306から出力されたライトバックデータパケット609の各ビットと論理値0との論理和を求めることにより、実質的にはライトバックデータパケット609を出力バッファ制御部712ceへルーティングする。

[0534] その結果、復元されたライトバックデータパケット609が出力バッファ制御部712ceを介して出力ポート202eへとさらにルーティングされ、SC103eへ出力される。

[0535] また、ステップS1112で分岐部1010は、ステップS1105で入力された複製データパケット出力ポート情報にしたがい、入力された共有型・転送データパケット613を、OR回路1011に出力する。

- [0536] 上記のとおり、複製情報を持ったライトバックデータパケット608の複製パケット宛先IDは、CPU104aのIDである。よって、CPU104aのIDから得られた複製データパケット出力ポート情報は、CPU104aと同じノード101aのSC103aと接続された、出力ポート202aを識別する情報である。
- [0537] したがって、分岐部1010は、入力された共有型・転送データパケット613を、出力ポート202aの前段に設けられた出力バッファ制御部712caと接続されたOR回路1011へ出力する。
- [0538] ここで、上記のとおり分岐部306は、選択しなかったOR回路1011には論理値0で埋めたデータを出力している。よって、OR回路1011は、分岐部1010から出力された共有型・転送データパケット613の各ビットと論理値0との論理和を求めることにより、実質的には共有型・転送データパケット613を出力バッファ制御部712caへルーティングする。
- [0539] その結果、復元された共有型・転送データパケット613が出力バッファ制御部712caを介して出力ポート202aへとさらにルーティングされ、SC103aへ出力される。
- [0540] 以上、図26～図29を参照して、複製情報を持ったライトバックデータパケット608からライトバックデータパケット609と共有型・転送データパケット613を復元してそれぞれルーティングする処理について説明した。
- [0541] なお、図26においては、いくつかのステップについての処理が並列に実行されてもよい。例えば、複製チェック部1001とルーティングテーブル部711cにおける処理は、データパケット復元部1002における処理と並列に実行されてもよい。また、ステップS1111とS1112のルーティングは並列に実行されてもよい。
- [0542] また、図15に示したように、各SC103a～103gでXB102と接続された入力ポート501xからルーティング部703へ至る経路上にも、上記と同様の複製チェック部1001、データパケット復元部1002お

よびセレクタ 1003 が設けられている。また、例えば図 15 の SC103 c ルーティング部 703 は、比較例と比べて、入力ポート 501 x に対応した部分が異なる。

[0543] 具体的には、図 23 と同様に、ルーティング部 703 において、比較例と同様の分岐部 306 と各出力バッファ制御部 509 x c、509 x d、509 x i、および 509 x s との間に、それぞれ OR 回路が設けられる。つまり、比較例と比べ、新たに 4 つの OR 回路が追加されている。

[0544] また、ルーティング部 703 は、図 23 の分岐部 1010 と類似の新たな分岐部を有する。この新たな分岐部は、入力ポート 501 x に対応して設けられたデータパケット復元部 1002 から復元されたパケットを受け取るとともに、複製チェック部 1001 およびルーティングテーブル部 701 x から制御信号を受け取る。また、この新たな分岐部は、図 23 と同様に、上記 4 つの OR 回路とそれぞれ接続されている。

[0545] しかし、ルーティング部 703 は、入力ポート 501 c、501 d、および 501 i、ならびに SC 内制御回路 504 に対応した部分は、比較例と同様である。

以上、図 14 ～図 29 を参照して第 1 実施形態について説明した。

[0546] なお、CPU 104 c などの演算処理装置と XB102 などのデータ転送装置の間でのデータ（例えばパケット形式のデータ）の転送を制御する SC103 c などの制御装置に注目して第 1 実施形態を概観すれば下記のとおりである。

[0547] すなわち、上記制御装置は、演算処理装置から出力された第 1 のパケットを保存するデータパケット保存部 410 などの保存手段を有する。第 1 実施形態では、保存手段としてのデータパケット保存部 410 に必要な容量が、上記のとおり比較例の約 2 分の 1 で済む。

[0548] また、上記制御装置は、第 1 の判断手段を実現するコンポーネントとして、例えば図 17 ～図 18 B に示す管理テーブル部 801、比較部 802、およびライトポインタ制御部 803 を備える。第 1 の判断手段は、第 1 のパケ

ットのヘッダと、第１の packets より後に演算処理装置から受け取った第２の packets のヘッダを比較する。packets はデータの具体例であり、ヘッダはデータの制御情報の具体例である。上記の比較により、第１の判断手段は、第１の packets と第２の packets が同じデータペイロードを有する第１の場合であるか否かを判断する。データペイロードはデータ本体部の具体例である。

[0549] また、管理テーブル部 801、比較部 802、およびライトポインタ制御部 803 は、協働して次のような管理手段としても機能する。すなわち、管理手段は、第１の場合は、第２の packets のヘッダの一部を第１の packets と関連づけて複製情報として保存する。また、管理手段は、第１の場合に第２の packets を破棄してもよい。また、管理手段は、第１の場合以外である第２の場合は、保存手段としてのデータ packets 保存部 410 に第２の packets を保存させる。

[0550] そして、保存手段に保存された packets のデータ転送装置（例えば XB102）への出力を制御する制御手段として、第１実施形態では、例えば以下のコンポーネントが協働している。すなわち、XB102 への packets の出力を制御する制御手段として、少なくとも図 17 のポート間アービタ部 507x、ポート間セレクタ 508x、制御部 400、および packets 間セレクタ 430 が協働している。また、第１実施形態ではさらにセレクタ 805 も制御手段の一部として機能している。データ転送装置としての XB102 へのデータ（例えば packets）の出力は、データ転送装置へのデータの出力を許可する出力許可に応じて制御されてもよい。

[0551] なお、第１実施形態では、ヘッダマージ部 804 が、図 22 の複製情報を持ったライトバックデータ packets 608 のような第３の packets を上記第１の場合に生成する生成手段としても機能する。

[0552] そして、上記制御手段として機能する制御部 400 内の管理テーブル部 801 およびセレクタ 805 は、上記第１の場合には生成手段としてのヘッダマージ部 804 が生成した第３の packets をデータ転送装置に出力するよう

制御する。また、上記制御手段としての管理テーブル部 801 およびセレクタ 805 は、上記第 2 の場合には保存手段としてのデータパケット保存部 410 が保存する上記第 1 と第 2 のパケットの双方をデータ転送装置に出力するように制御する。

[0553] また、第 1 実施形態では、第 1 の判断手段としての管理テーブル部 801、比較部 802、およびライトポインタ制御部 803 では、要求元 ID と Req ID に基づく判断が行われる。要求元 ID は、パケットの出力の起因となる要求を出力した要求元装置を識別する要求元識別情報の一例である。また、Req ID は、要求元識別情報により識別される 1 つの要求元装置における複数の要求を識別する要求識別情報の一例である。

[0554] また、上記管理手段の一部として機能する管理テーブル部 801 は、例えば図 18A のエントリ 901-0 ~ 901-N により実現される管理記憶部を備える。そして、上記第 1 の場合には、上記第 1 のパケットが保存されたデータパケット保存部 410 上のアドレスと関連づけられた管理記憶部のエントリに複製情報を保存することにより、管理手段としての管理テーブル部 801 は、複製情報を第 1 のパケットと関連づける。

[0555] また、SC103c などの制御装置は、図 15 に示すように、さらに第 2 の判断手段としての複製チェック部 1001 と、復元手段としてのデータパケット復元部 1002 を備えてもよい。

[0556] また、XB102 などのデータ転送装置に注目して第 1 実施形態を概観すれば下記のとおりである。

すなわち、XB102 などのデータ転送装置は、例えば SC103a ~ 103g などの複数の装置と接続され、それら複数の装置間でのデータの転送を行う。

[0557] データ転送装置は、前記複数の装置のうちの 1 つから出力されたパケットを前記複数の装置のうちの他の 1 つへ送信する（すなわちルーティングする）ための送信手段として、図 16 や図 23 に示したルーティング部 713 を備える。

[0558] また、図 16 に示した X B 1 0 2 内の各出力バッファ制御部には、上記の S C に注目した観点で説明したのと同様の意味で、保存手段、判断手段、管理手段、および制御手段として機能する各コンポーネントが含まれる。さらに、第 1 実施形態では、生成手段としてのヘッダマージ部 8 0 4 も、X B 1 0 2 内の各出力バッファ制御部が備えている。

[0559] また、図 16 と図 23 に示すように、X B 1 0 2 は、入力されたパケットが同じデータペイロードを有する 2 つのパケットから生成された複製データパケットであるか否かを判断する判断手段としての複製チェック部 1 0 0 1 を備える。また、X B 1 0 2 は、復元手段としてのデータパケット復元部 1 0 0 2 も備える。そして、送信手段としてのルーティング部 7 1 3 は、複製チェック部 1 0 0 1 の判断に基づいてデータの送信（具体的にはパケットのルーティング）を行う。

[0560] 具体的には、第 1 実施形態では、判断手段としての複製チェック部 1 0 0 1 は、入力されたパケットが複製データパケットであるか否かだけでなく、元の 2 つのパケットは同じ装置に送信するものであるか否かについても判断する。

[0561] そして、元の 2 つのパケットは同じ装置に送信する（つまりルーティングする）ものであると複製チェック部 1 0 0 1 が判断した場合、送信手段としてのルーティング部 7 1 3 は入力された複製データパケットそのものを送信する。逆に、元の 2 つのパケットは同じ装置に送信するものではないと複製チェック部 1 0 0 1 が判断した場合は、送信手段としてのルーティング部 7 1 3 は、復元された元の 2 つのパケットをそれぞれの送信先に送信する。

[0562] 以上、制御装置としての S C 1 0 3 a ~ 1 0 3 g に注目した観点と、データ転送装置としての X B 1 0 2 に注目した観点から第 1 実施形態を概観した。

続いて、図 30 を参照して、第 2 実施形態について説明する。第 2 実施形態も第 1 実施形態と同様に、例えば図 1 のように構成されたサーバシステム 1 0 0 に適用される。

- [0563] 図30は、第2実施形態におけるCPU104cの構成図である。説明の簡略化のため、第2実施形態では、他のCPU104a、104b、104d～104hも図30と同様に構成されているものとする。
- [0564] 第2実施形態におけるCPU104cはマルチコアCPUであり、具体的にはデュアルコアCPUである。すなわち、演算処理装置であるCPU104cは、演算処理部であるCPUコア1301aと1301bを含む。CPUコア1301aと1301bはそれぞれ、不図示の命令フェッチ制御部、分岐予測部、命令デコード部、ALU (Arithmetic Logic Unit) などの命令実行部、およびレジスタなどを備える。アーキテクチャによっては、CPUコア1301aと1301bは、さらにアウトオブオーダー実行のためのリザーベーションステーションなどを備えていてもよい。
- [0565] CPUコア1301aと1301bはそれぞれ、不図示のキャッシュメモリを備える。キャッシュメモリは階層化されていてもよい。例えば、L1 (レベル1) キャッシュが各CPUコア1301aと1301bに個別に設けられ、L2 (レベル2) キャッシュが両CPUコア1301aと1301bで共有されていてもよい。L2キャッシュを各CPUコア1301aと1301bに個別に設ける構成も可能である。なお、データキャッシュと命令キャッシュが分かれていてもよい。
- [0566] また、CPU104cは、同じノード101c内のSC103cとのインタフェースとして、入力ポート1302と出力ポート1303を備える。入力ポート1302と出力ポート1303は、CPU104cとSC103cの動作周波数の違いに応じた「周波数乗り換え」と呼ばれる周波数の変換などを行ってもよい。
- [0567] 第1実施形態では、パケットの宛先IDや要求元IDとしてCPU104cのIDが使われるが、第2実施形態では、CPU104cのIDではなく、CPUコア1301aまたはCPUコア1301bのIDが宛先IDや要求元IDとして使われる。したがって、CPU104cは、SC103cから入力ポート1302を介して入力されるパケットを、宛先IDに応じてC

ＰＵコア１３０１ aまたは１３０１ bに振り分けるための回路を有する。

[0568] ところで、第１実施形態では、同じデータペイロードを有するデータパケットの組として、ＳＣの１つを宛先とするライトバックデータパケットと、ＣＰＵの１つを宛先とする共有型・転送データパケットを想定している。サーバシステム１００の仕様により、それ以外の組み合わせで、同じデータペイロードを有するデータパケットの組が存在する場合にも第１実施形態は適用可能である。同様に、第２実施形態は、２つの異なるＣＰＵコアを宛先とし、互いに同じデータペイロードを有するデータパケットの組が存在する場合にも適用可能である。

[0569] なぜなら、第２実施形態では、上記のパケットを振り分けるための回路が、以下の各コンポーネントを含むからである。

- ・図２３と同様の複製チェック部１００１、データパケット復元部１００２、およびセレクタ１００３

- ・図２５のルーティングテーブル部７１１ cと類似の構成を有し、出力ポートの代わりに出力先のＣＰＵコア（１３０１ aまたは１３０１ b）を指定する情報を、宛先ＩＤと複製宛先ＩＤにそれぞれ対応して取得する情報ルーティングテーブル部１３０４

- ・図２３の分岐部１０１０と同様の分岐部１３０５ a

- ・図２３の分岐部３０６と同様の分岐部１３０５ b

- ・図２３のＯＲ回路１０１１～１０１３と同様のＯＲ回路１３０６ a～１３０６ b

したがって、入力ポート１３０２から入力されたパケットが複製情報を持っていなければ、宛先ＩＤにしたがってＣＰＵコア１３０１ aまたは１３０１ bに出力される。また、入力ポート１３０２から入力されたパケットが複製情報を持っていれば、元の２つのパケットが復元されて、それぞれＣＰＵコア１３０１ aと１３０１ bに出力される。

[0570] また、ＣＰＵ１０４ cは、出力ポート１３０３の前段に、ＣＰＵコア１３０１ a用の出力バッファ制御部１３０７ aと、ＣＰＵコア１３０１ b用の出

カバッファ制御部 1307b と、アービタ部 1308 と、コア間セクタ 1309 とを備える。

[0571] 出力バッファ制御部 1307a は、CPU コア 1301a から出力されたパケットを一時的に保存するバッファを含み、具体的には、第 1 実施形態の図 17 の出力バッファ制御部 702cx と同様の構成を有する。出力バッファ制御部 1307b も同様である。

[0572] また、アービタ部 1308 は、出力バッファ制御部 1307a と 1307b の間の調停を行う。アービタ部 1308 は、例えば第 1 実施形態における図 15 のポート間アービタ部 507x や図 16 のポート間アービタ部 205e と同様である。

[0573] コア間セクタ 1309 は、アービタ部 1308 の調停の結果にしたがって、出力バッファ制御部 1307a と 1307b からの出力の一方を選択して出力ポート 1303 に出力する。コア間セクタ 1309 は、例えば第 1 実施形態における図 15 のポート間セクタ 508x や図 16 のポート間セクタ 206e と同様である。

[0574] また、第 2 実施形態における XB102 は、第 1 実施形態における図 16 の XB102 と同様に構成されている。

そして、第 2 実施形態における SC103c は、下記の点において第 1 実施形態と異なる。すなわち、図 30 の CPU104c と接続された SC103c は、図 15 における入力ポート 501c とルーティング部 703 の間に、図 23 と同様の複製チェック部 1001、データパケット復元部 1002、およびセクタ 1003 を備える。そして、CPU104c に対応するルーティングテーブル部 506c は、ルーティングテーブル部 701x と同様の構成のものに置き換えられる。

[0575] なお、CPU104c 以外の他の CPU も図 30 と同様に構成されていてもよく、それに応じて各 SC 内のコンポーネントが適宜変形されてもよい。

上記のような第 2 実施形態によっても、サーバシステム 100 内でのデータの授受のために用いられるパケットを一時的に保存するためのバッファの

容量を削減し、サーバシステム 100 内のバスの占有時間を短縮するという効果が得られる。

[0576] なお、本発明は上記の実施形態に限られるものではなく、例えば下記のように様々な変形して実施することが可能である。

第 1 および第 2 実施形態は、図 1 に示した構成のサーバシステム 100 への適用例である。しかし、サーバシステム 100 に含まれるノードの数や、各ノード内の CPU の数は任意である。

[0577] また、サーバシステム 100 は c c N U M A (Cache Coherent Non-Uniform Memory Access) システムの一例だが、S M P (Symmetric Multiple Processor) システム (つまり U M A (Uniform Memory Access) システム) に、第 1 および第 2 実施形態と同様の構成を適用することもできる。

[0578] また、第 1 実施形態では、先に受け取ったデータパケットのヘッダの情報を、図 18 A のエントリ 901-0 ~ 901-N のいずれかの中の要求元 I D フィールド 915 と R e q I D フィールド 916 に記憶している。しかし、要求元 I D フィールド 915 と R e q I D フィールド 916 を省略し、代わりに、データパケット保存部 410 から図 18 B の比較部 802 が要求元 I D と R e q I D を読み出す実施形態も可能である。ただし、R A M を用いてデータパケット保存部 410 を実現する場合は、F F を用いて管理テーブル部 801 内の各エントリ 901-0 ~ 901-N に要求元 I D フィールド 915 と R e q I D フィールド 916 を設ける方が、処理速度の点で好ましい。

[0579] また、バスの占有時間が比較例と同等でも問題がない場合、例えば図 17 の出力バッファ制御部 702 c x を次のように変形してもよい。すなわち、出力バッファ制御部 702 c x は、ポート間アービタ部 507 x からの出力許可が得られたら、図 17 のヘッダマージ部 804 で複製情報を持ったデータパケットを生成して出力する代わりに、元の 2 つのパケットを連続して出力してもよい。

[0580] 例えば、出力バッファ制御部 702 c x は、データパケット保存部 410

のリードポインタ値で指示される番地に保存されている第1の packets をまず出力する。さらに、出力バッファ制御部702cは、第1の packets の番地に対応する管理テーブル部801内のエントリから複製情報を読み出す。そして、出力バッファ制御部702cは、読み出した複製情報と第1の packets から、第1の packets と同じデータペイロードを持つために破棄された第2の packets を復元し、出力する。

[0581] 第2の packets の復元は、例えば図23のデータ packets 生成部1009と類似の新たなコンポーネントをヘッダマージ部804の代わりに設けることで実現することができる。また、例えば図17において管理テーブル部801は、セレクタ805に対して、データ packets 保存部410から第1の packets を出力するよう指示した後、必要な時間の経過後、上記新たなコンポーネントから第2の packets を出力するよう指示を切り換える。

[0582] 上記のような変形は、SC103c内の出力バッファ制御部702cに限らず、XB102内の出力バッファ制御部においても採用することができる。

また、上記のように出力バッファ制御部からの出力時に packets を復元する場合は、図15および図16において複製チェック部1001、データ packets 復元部1002、およびセレクタ1003を設ける必要はない。また、図15のルーティング部703の代わりに比較例の図10のルーティング部505を用いることができ、図16のルーティング部713の代わりに比較例の図5のルーティング部203を用いることができる。

[0583] このように、出力バッファ制御部からの出力時に packets を復元する変形例においても、出力バッファ制御部内に必要な記憶容量が比較例の約2分の1であり、そのため消費電力や半導体チップ上の実装面積を節約することができるという効果が得られる。

[0584] また、図30の第2実施形態では、デュアルコアCPUを例示したが、コアの数が3以上のマルチコアCPUに合わせて第2実施形態を変形した実施形態も可能であるし、シングルコアCPUが図30と同様に出力バッファ制

御部を含む実施形態も可能である。

- [0585] シングルコアCPUの内部に出力バッファ制御部がある場合、CPU内部には、図30のようなアービタ部1307およびコア間セクタ1308は不要であり、CPU内の出力バッファ制御部は直接CPUの出力ポートと接続されてもよい。また、シングルコアCPUでは、入力ポート1302が直接コアと接続されていてもよい。
- [0586] また、マルチコアCPUの場合でも、複数のコアを宛先として同じデータペイロードを有するパケットが出力されることがないならば、図30の第2実施形態を次のように変形してもよい。
- [0587] すなわち、入力ポート1302が分岐部1305bと直接接続され、分岐部1305bがCPUコア1301aと1301bに直接接続されていてもよい。この場合、分岐部1305bは、1入力2出力のデマルチプレクサでもよい。
- [0588] また、この場合、図30の複製チェック部1001、データパケット復元部1002、セクタ1003、分岐部1305a、OR回路1306a、およびOR回路1306bは省略可能である。さらに、ルーティングテーブル部1304は、出力ポート情報の取得のみを行うものに置き換えることができる。

請求の範囲

- [1] 第1の演算処理装置およびデータ転送装置の間でのデータの転送を制御する制御装置であって、
- 前記第1の演算処理装置から出力された第1のデータを保存する保存手段と、
- 前記第1のデータの制御情報と、前記第1のデータより後に前記第1の演算処理装置から受け取った第2のデータの制御情報を比較することで、前記第1のデータと前記第2のデータが同じデータ本体部を有する第1の場合であるか否かを判断する第1の判断手段と、
- 前記第1の場合は、前記第2のデータの前記制御情報の一部を前記第1のデータと関連づけた複製情報として保存し、前記第1の場合以外である第2の場合は、前記保存手段に前記第2のデータを保存させる管理手段と、
- 前記保存手段に保存されたデータの前記データ転送装置への出力を制御する制御手段と、
- を備えることを特徴とする制御装置。
- [2] 前記第1のデータと前記第2のデータの一方が、前記第1の演算処理装置が備える記憶部に保存されたデータを前記制御装置に接続された記憶装置に書き込むデータであり、前記第1のデータと前記第2のデータの他方が、第2の演算処理装置が要求した前記記憶部に保存されたデータを転送する転送データであるとき、
- 前記判断手段は、前記第1の場合であると判断する、
- ことを特徴とする請求項1に記載の制御装置。
- [3] 前記第1の場合に、前記保存手段が保存する前記第1のデータと前記管理手段が保存する前記複製情報から、第3のデータを生成する生成手段をさらに備え、
- 前記制御手段は、前記第1の場合には前記生成手段が生成した前記第3のデータを前記データ転送装置に出力するよう制御するとともに、前記第2の場合には前記保存手段が保存する前記第1のデータと前記第2のデータの双

方を前記データ転送装置に出力するよう制御する、

ことを特徴とする請求項 1 に記載の制御装置。

- [4] 前記第 1 のデータと前記第 2 のデータそれぞれが有する制御情報は、前記第 1 の演算処理装置によるデータの出力の起因となる要求を出力した要求元装置を識別する要求元識別情報と、前記要求元識別情報により識別される 1 つの前記要求元装置における複数の要求を識別する要求識別情報とを含み、
- 前記判断手段は、前記第 1 のデータと前記第 2 のデータの前記制御情報において前記要求元識別情報と前記要求識別情報がともに等しければ、前記第 1 のデータと前記第 2 のデータの前記データ本体部が同じであると判断する、

ことを特徴とする請求項 1 に記載の制御装置。

- [5] 前記管理手段は、前記複製情報を保存する管理記憶部を備え、
- 前記管理手段は、前記第 1 の場合において、前記第 1 のデータが保存された前記保存手段上のアドレスと関連づけられた前記管理記憶部のエントリに前記複製情報を保存する、

ことを特徴とする請求項 1 に記載の制御装置。

- [6] 前記第 1 のデータが前記第 1 の演算処理装置から出力されたとき、前記管理手段はさらに前記第 1 のデータの前記制御情報の一部を保存し、
- 前記判断手段は、前記管理手段が保存した前記第 1 のデータの前記制御情報の前記一部を参照し、前記第 2 のデータの前記制御情報との比較に用いる、

ことを特徴とする請求項 1 に記載の制御装置。

- [7] 前記データ転送装置から前記制御装置に出力された第 4 のデータが、同じデータ本体部を有する第 5 のデータと第 6 のデータに基づいて生成されたデータであるか否かを、前記第 4 のデータの制御情報から判断する第 2 の判断手段と、

前記第 5 のデータと前記第 6 のデータに基づいて前記第 4 のデータが生成されたと前記第 2 の判断手段が判断したとき、前記第 4 のデータから前記第

5と前記第6のデータを復元する復元手段と、

前記復元手段が復元した前記第5のデータと前記第6のデータをそれぞれ送信する送信手段と、

をさらに備えることを特徴とする請求項1に記載の制御装置。

[8] 前記管理手段は、前記第1の場合にさらに前記第2のデータを破棄することを特徴とする請求項1に記載の制御装置。

[9] 第1の装置と第2の装置を含む複数の装置と接続され、前記複数の装置間でのデータの転送を行うデータ転送装置であって、

前記複数の装置のうちの1つから出力されたデータを前記複数の装置のうちの他の1つへ送信する送信手段と、

前記第1の装置から受け取った、前記第2の装置へ送信する第1のデータを前記第2の装置へ出力する前に保存しておく保存手段と、

前記第1の装置から受け取り、前記第1のデータより後に前記送信手段により送信した、前記第2の装置へ送信する第2のデータの制御情報を、前記第1のデータの制御情報と比較することで、前記第1のデータと前記第2のデータが同じデータ本体部を有する第1の場合であるか否かを判断する第1の判断手段と、

前記第1の場合は、前記第2のデータの制御情報の一部を前記第1のデータと関連づけた複製情報として保存し、前記第1の場合以外である第2の場合は、前記保存手段に前記第2のデータを保存させる管理手段と、

前記保存手段に保存されたデータの制御情報の前記第2の装置への出力を制御する制御手段と、

を備えることを特徴とするデータ転送装置。

[10] 前記第1の場合に、前記保存手段が保存する前記第1のデータと前記管理手段が保存する前記複製情報から、第3のデータを生成する生成手段をさらに備え、

前記制御手段は、前記第1の場合には前記生成手段が生成した前記第3のデータを前記第2の装置に出力するよう制御するとともに、前記第2の場合

には前記保存手段が保存する前記第 1 のデータと前記第 2 のデータの双方を前記第 2 の装置に出力するよう制御する、

ことを特徴とする請求項 9 に記載のデータ転送装置。

- [11] 前記第 1 のデータが、前記第 2 の装置に送信する第 4 のデータと、前記第 4 のデータと同じデータ本体部を有する第 5 のデータから生成された複製データであるか否かを前記第 1 のデータの前記制御情報から判断する第 2 の判断手段と、

前記第 1 のデータが複製データであるとき、前記第 1 のデータから前記第 4 のデータと前記第 5 のデータを復元する復元手段とをさらに備え、

前記送信手段は、前記第 2 の判断手段による判断に基づいて、前記復元手段が復元した前記第 4 のデータと前記第 5 のデータを送信するか、または前記第 1 のデータを送信する、

ことを特徴とする請求項 9 に記載のデータ転送装置。

- [12] 前記第 2 の判断手段は、前記第 1 のデータが前記複製データである場合はさらに、前記第 5 のデータを前記第 2 の装置に送信する第 3 の場合か否かを判断し、

前記第 2 の判断手段が前記第 3 の場合であると判断したときは、前記送信手段は、前記第 1 のデータを前記第 2 の装置に送信し、

前記第 1 のデータが前記複製データであり、かつ、前記第 2 の判断手段が前記第 3 の場合ではないと判断したときは、前記送信手段は、前記復元手段が復元した前記第 4 のデータと前記第 5 のデータをそれぞれの送信先に送信する、

ことを特徴とする請求項 11 に記載のデータ転送装置。

- [13] 前記復元手段は、

前記第 1 のデータの前記制御情報の一部を削除することによって前記第 1 のデータから前記第 4 のデータを復元する複製データ修正手段と、

前記第 1 のデータの前記制御情報から前記第 5 のデータの制御情報を復元する制御情報復元手段と、

前記制御情報復元手段が復元した前記第 5 のデータのデータの前記制御情報と前記第 1 のデータのデータ本体部を連結することで前記第 5 のデータを復元するデータ生成手段と、

を備えることを特徴とする請求項 11 に記載のデータ転送装置。

[14] 前記複数の装置にはさらに第 3 の装置が含まれ、

前記第 5 のデータが、前記第 3 の装置に接続された記憶装置に、前記第 1 の装置に接続された第 1 の演算処理装置が備える記憶部に保存されたデータを書き込むデータであり、かつ前記第 4 のデータが、前記第 2 の装置に接続されており前記記憶部に保存された前記データを要求した第 2 の演算処理装置に、前記記憶部に保存された前記データを転送する転送データであるか、または、

前記第 4 のデータが、前記第 2 の装置に接続された記憶装置に前記記憶部に保存された前記データを書き込むデータであり、かつ前記第 5 のデータが、前記第 3 の装置に接続されており前記記憶部に保存された前記データを要求した第 3 の演算処理装置に、前記記憶部に保存された前記データを転送する転送データである、

ことを特徴とする請求項 11 に記載のデータ転送装置。

[15] 第 1 の制御装置および第 2 の制御装置を含む複数の制御装置と、前記複数の制御装置間でのデータの転送を行うデータ転送装置を備える情報処理装置であって、

前記第 1 の制御装置は、

前記第 1 の制御装置に接続された第 1 の演算処理装置から出力され、前記第 2 の制御装置へと送信する第 1 のデータを保存する保存手段と、

前記第 1 のデータの制御情報と、前記第 1 のパケットより後に前記第 1 の演算処理装置から受け取った第 2 のデータの制御情報とを比較することにより、前記第 1 のデータと前記第 2 のデータが同じデータ本体部を有する第 1 の場合であるか否かを判断する第 1 の判断手段と、

前記第 1 の場合は、前記第 2 のデータのデータの前記制御情報の一部を前記第 1 の

データと関連づけて複製情報として保存し、前記第 1 の場合以外の第 2 の場合は、前記保存手段に前記第 2 のデータを保存させる管理手段と、

前記第 1 の場合に、前記保存手段が保存する前記第 1 のデータと前記管理手段が保存する前記複製情報から、第 3 のデータを生成する生成手段と、

前記第 1 の場合には前記生成手段が生成した前記第 3 のデータを前記データ転送装置に出力するよう制御するとともに、前記第 2 の場合には前記保存手段が保存する前記第 1 のデータと前記第 2 のデータの双方を前記データ転送装置に出力するよう制御する第 1 の制御手段を備え、

前記データ転送装置は、

前記第 1 の制御装置から受け取ったデータの制御情報に基づいて、前記第 1 の場合か前記第 2 の場合かを判断する第 2 の判断手段と、

前記第 1 の場合に、前記第 3 のデータから前記第 1 のデータと前記第 2 のデータを復元する復元手段と、

前記第 2 の判断手段による判断に基づいて、前記第 3 のデータを送信するか、または前記復元手段が復元した前記第 1 のデータと前記第 2 のデータを送信する送信手段を備える、

ことを特徴とする情報処理装置。

[16] 記憶部と演算処理部を備えた演算処理装置であって、

前記演算処理部から出力された第 1 のデータを保存する保存手段と、

前記第 1 のデータの制御情報と、前記第 1 のデータより後に前記演算処理部から出力された第 2 のデータの制御情報とを比較することにより、前記第 1 のデータと前記第 2 のデータが同じデータ本体部を有するか否かを判断する判断手段と、

前記第 1 の場合は、前記第 2 のデータの前記制御情報の一部を前記第 1 のデータと関連づけて複製情報として保存し、前記第 1 の場合以外の第 2 の場合は、前記保存手段に前記第 2 のデータを保存させる管理手段と、

前記第 1 の場合に、前記保存手段が保存する前記第 1 のデータと前記管理手段が保存する前記複製情報から、第 3 のデータを生成する生成手段と、

前記第 1 の場合には前記生成手段が生成した前記第 3 のデータを出力するよう制御するとともに、前記第 2 の場合には前記保存手段が保存する前記第 1 のデータと前記第 2 のデータの双方を出力するよう制御する制御手段と、を備えることを特徴とする演算処理装置。

[17] 第 1 の制御装置と、前記第 1 の制御装置に接続された第 1 の演算処理装置と、前記第 1 の制御装置を含む複数の制御装置の間のデータの転送を行うデータ転送装置を有する情報処理装置の制御方法において、

前記第 1 の演算処理装置が、第 1 のデータと第 2 のデータを前記第 1 の制御装置に出力するステップと、

前記第 1 の制御装置が、前記第 1 のデータを受け取って保存し、前記第 1 のデータを受け取った後で前記第 2 のデータを受け取り、前記第 1 のデータの制御情報と前記第 2 のデータの制御情報を比較することにより、前記第 1 のデータと前記第 2 のデータが同じデータ本体部を有する第 1 の場合であるか否かを判断するステップと、

前記第 1 の場合だと前記第 1 の制御装置が判断した第 1 の場合は、

前記第 1 の制御装置は、前記第 2 のデータの前記制御情報の一部を前記第 1 のデータと関連づけて複製情報として保存し、保存している前記第 1 のデータと前記複製情報から第 3 のデータを生成し、前記データ転送装置に前記第 3 のデータを出力するステップと、

前記データ転送装置が、前記第 3 のデータの制御情報に基づいて前記第 1 の場合であると判断するとともに、前記第 3 のデータの前記制御情報に基づいて前記第 1 のデータと前記第 2 のデータの送信先が同じか否かを判断するステップと、

前記第 1 のデータと前記第 2 のデータの送信先が同じならば、前記データ転送装置は、前記第 3 のデータを前記同じ送信先に送信するステップと、

前記第 1 のデータと前記第 2 のデータの送信先が異なれば、前記データ転送装置は、前記第 3 のデータの前記制御情報に基づいて、前記第 3 のデータから前記第 1 のデータと前記第 2 のデータを復元してそれぞれを前記複数

の制御装置のいずれかへ送信するステップと、

前記第 1 の場合以外の第 2 の場合は、

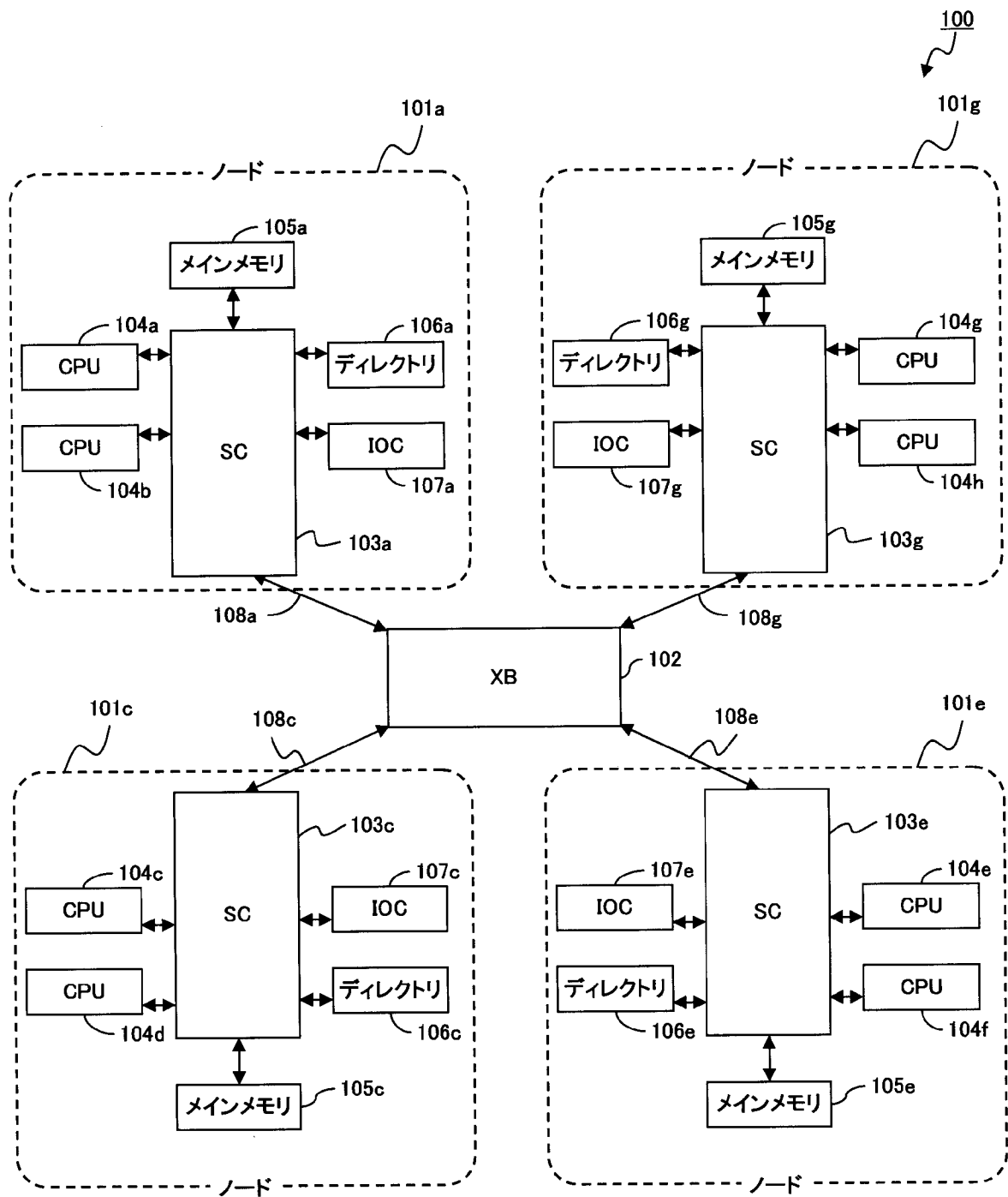
前記第 1 の制御装置が、さらに前記第 2 のデータを保存し、保存している前記第 1 のデータと前記第 2 のデータの双方を前記データ転送装置に出力するステップと、

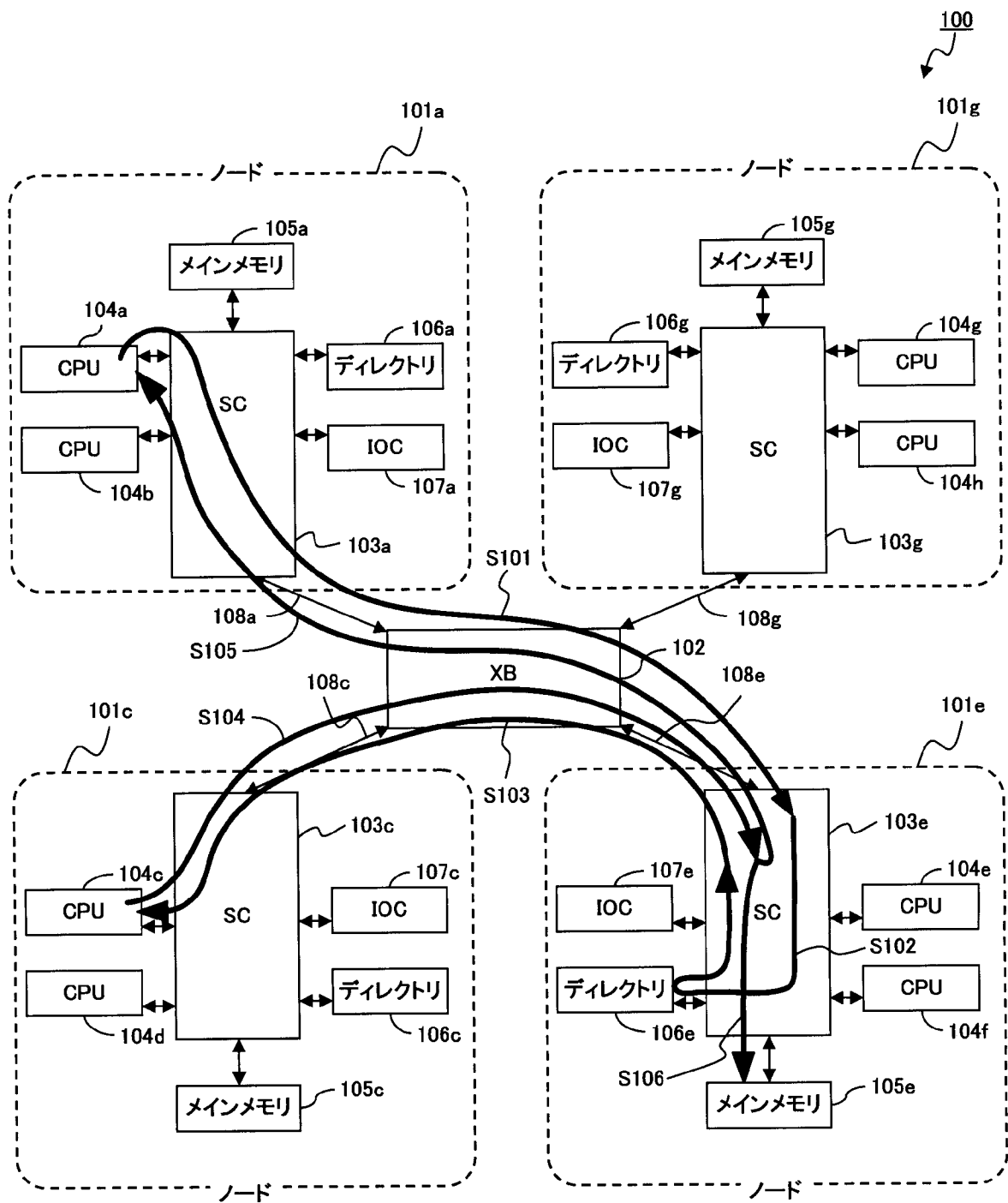
前記データ転送装置が、前記第 1 のデータの前記制御情報に基づいて、前記複数の制御装置のいずれかへ前記第 1 のデータを送信し、前記第 2 のデータの前記制御情報に基づいて、前記複数の制御装置のいずれかへ前記第 2 のデータを送信するステップと、

を含むことを特徴とする情報処理装置の制御方法。

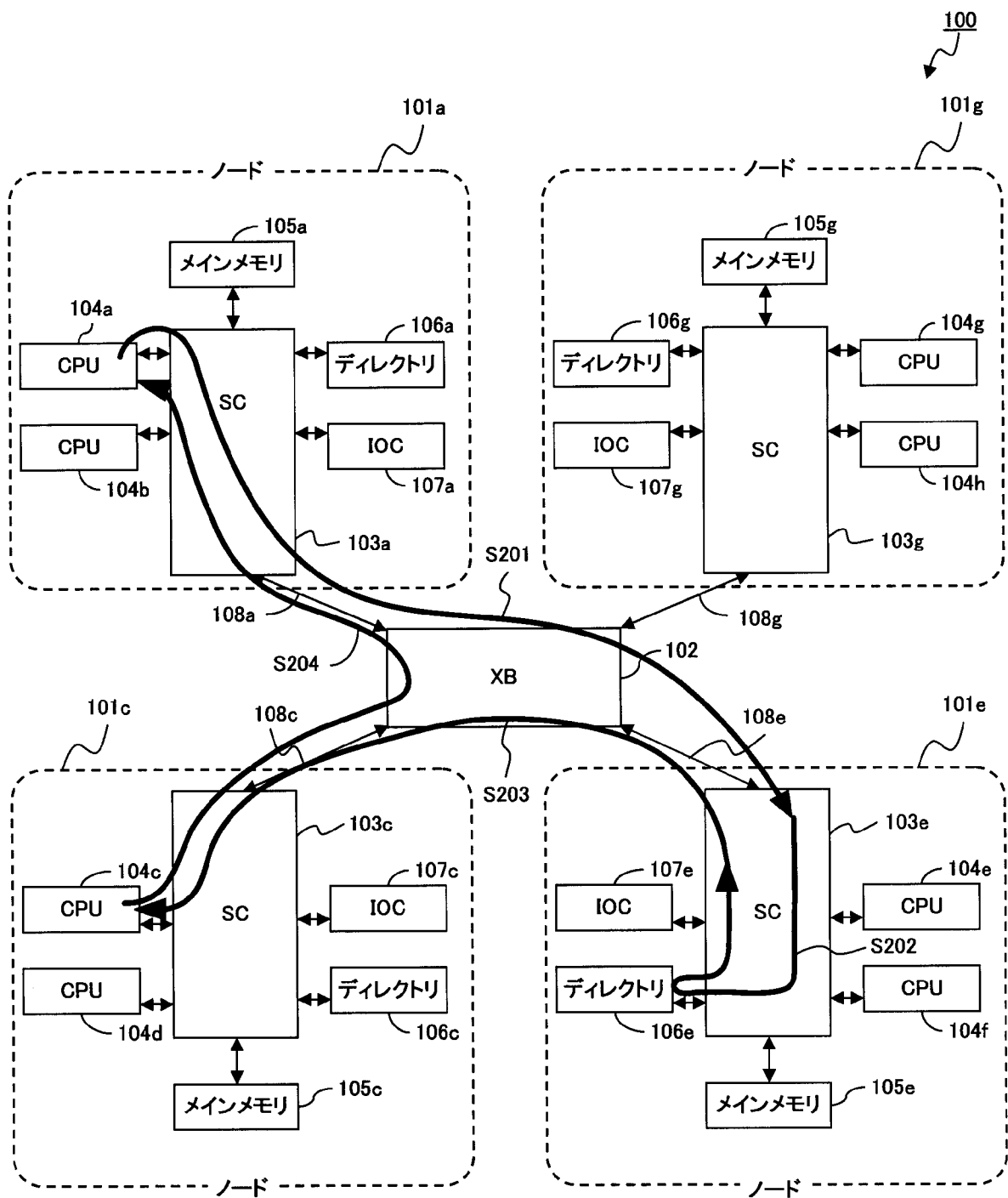
- [18] 前記第 1 の場合に前記第 1 の制御装置がさらに前記第 2 のデータを破棄することを特徴とする請求項 17 に記載の情報処理装置の制御方法。

[図1]

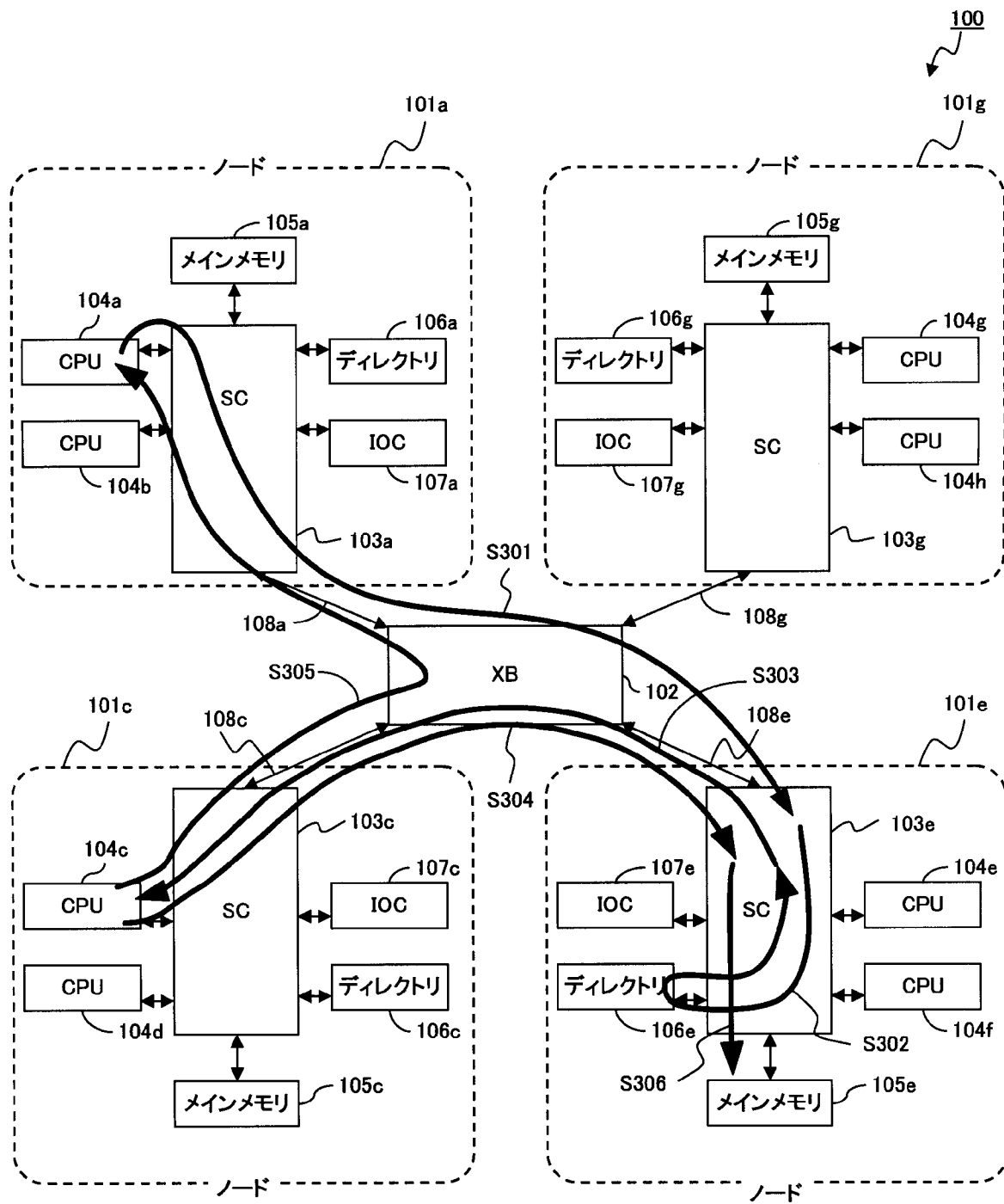




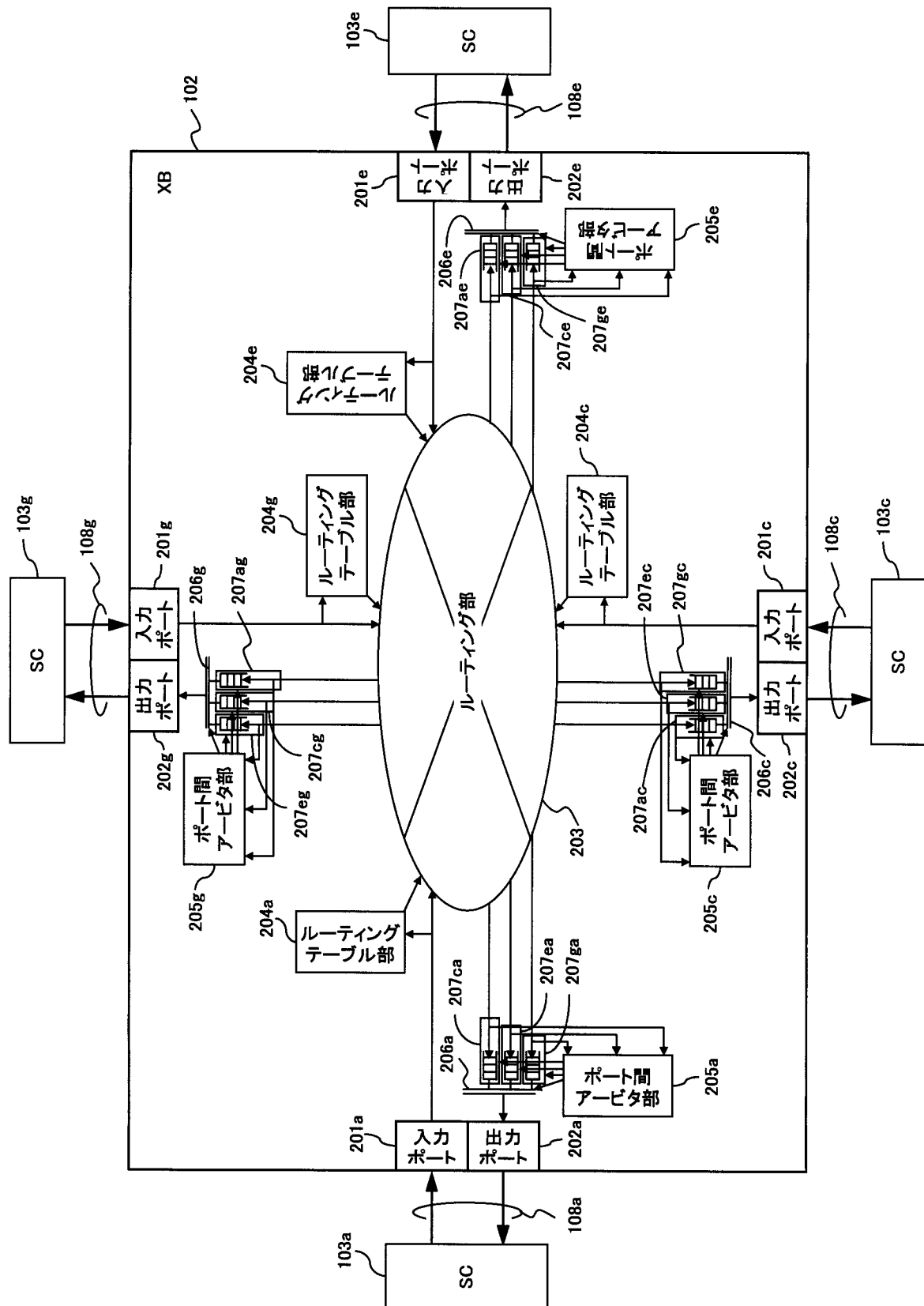
[図3]



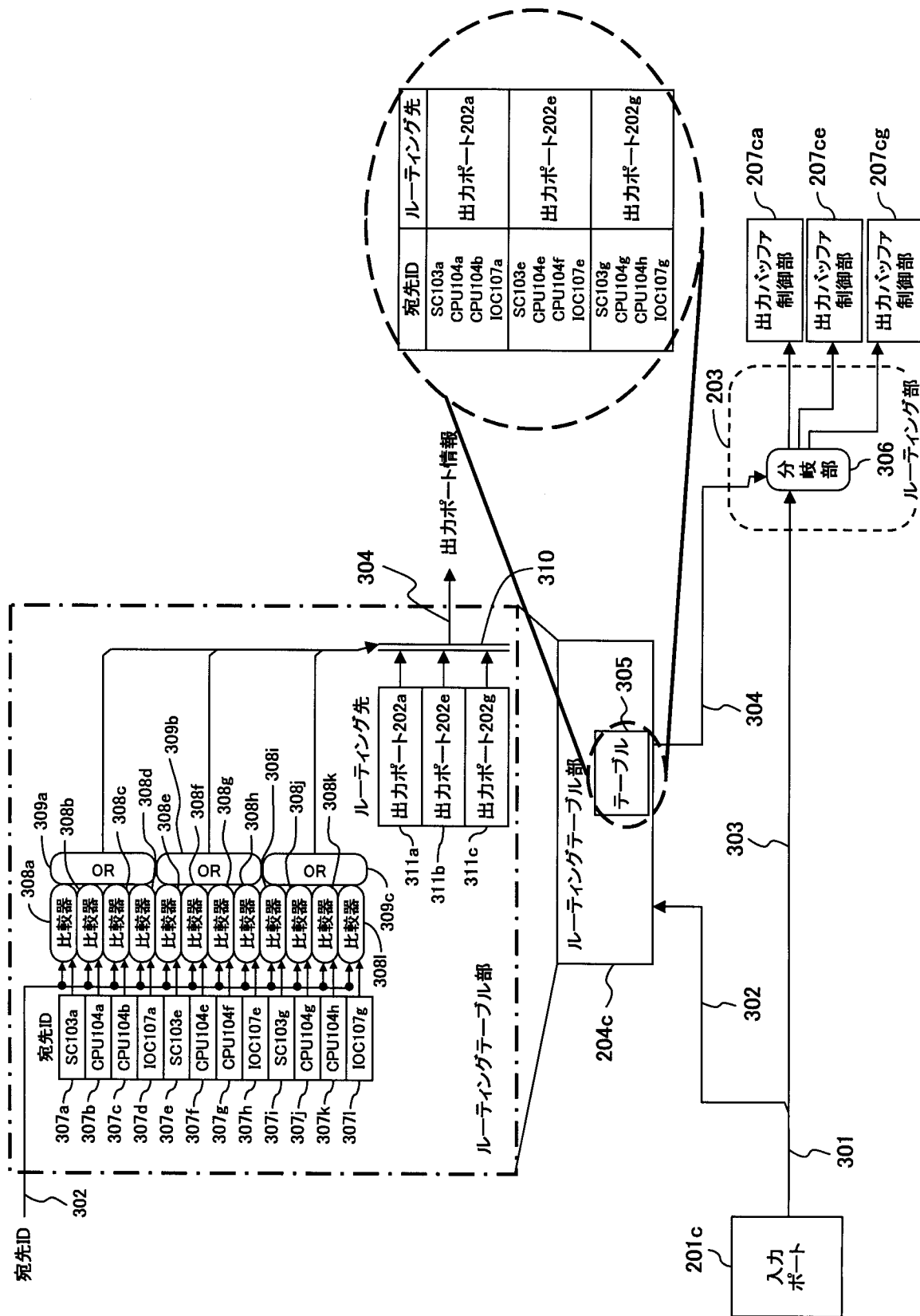
[図4]



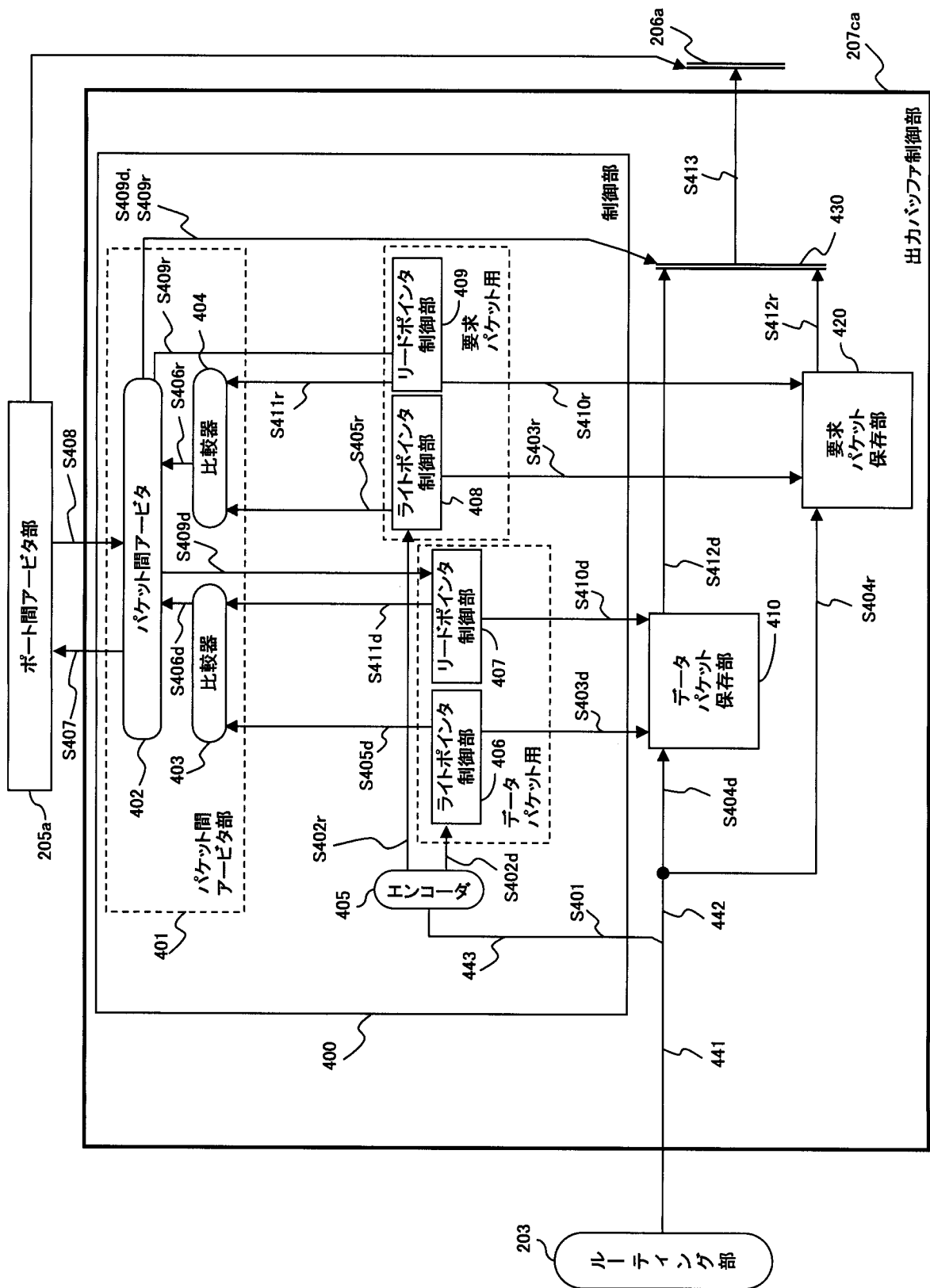
[図5]



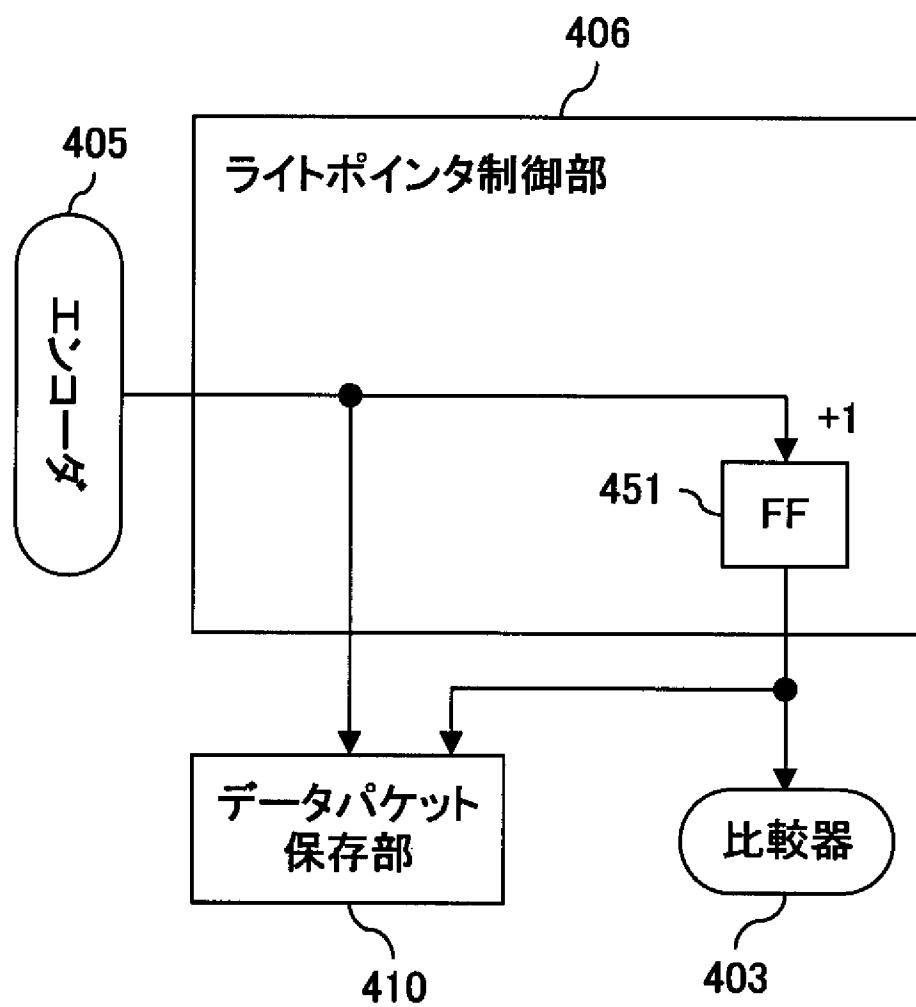
[図6]



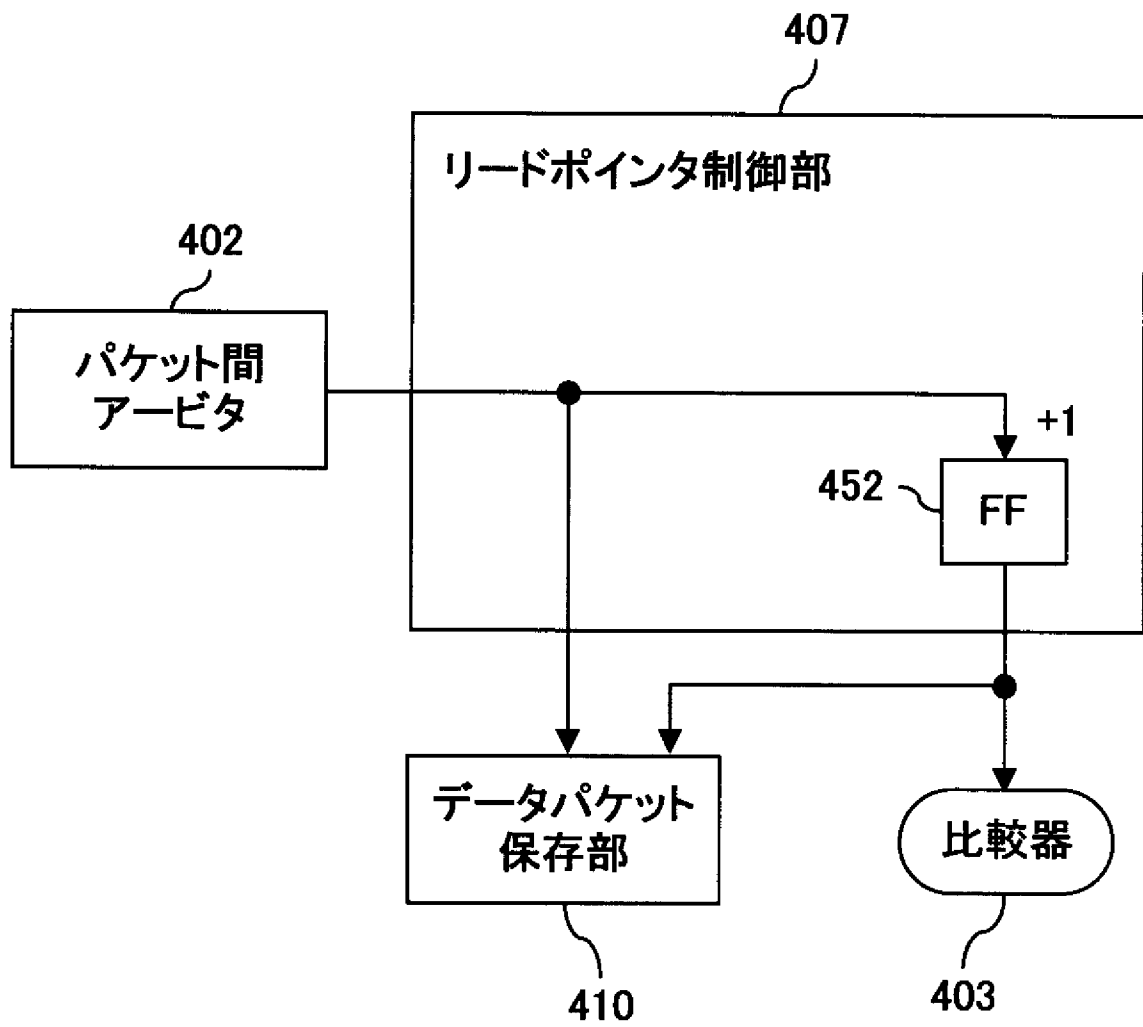
[図7]



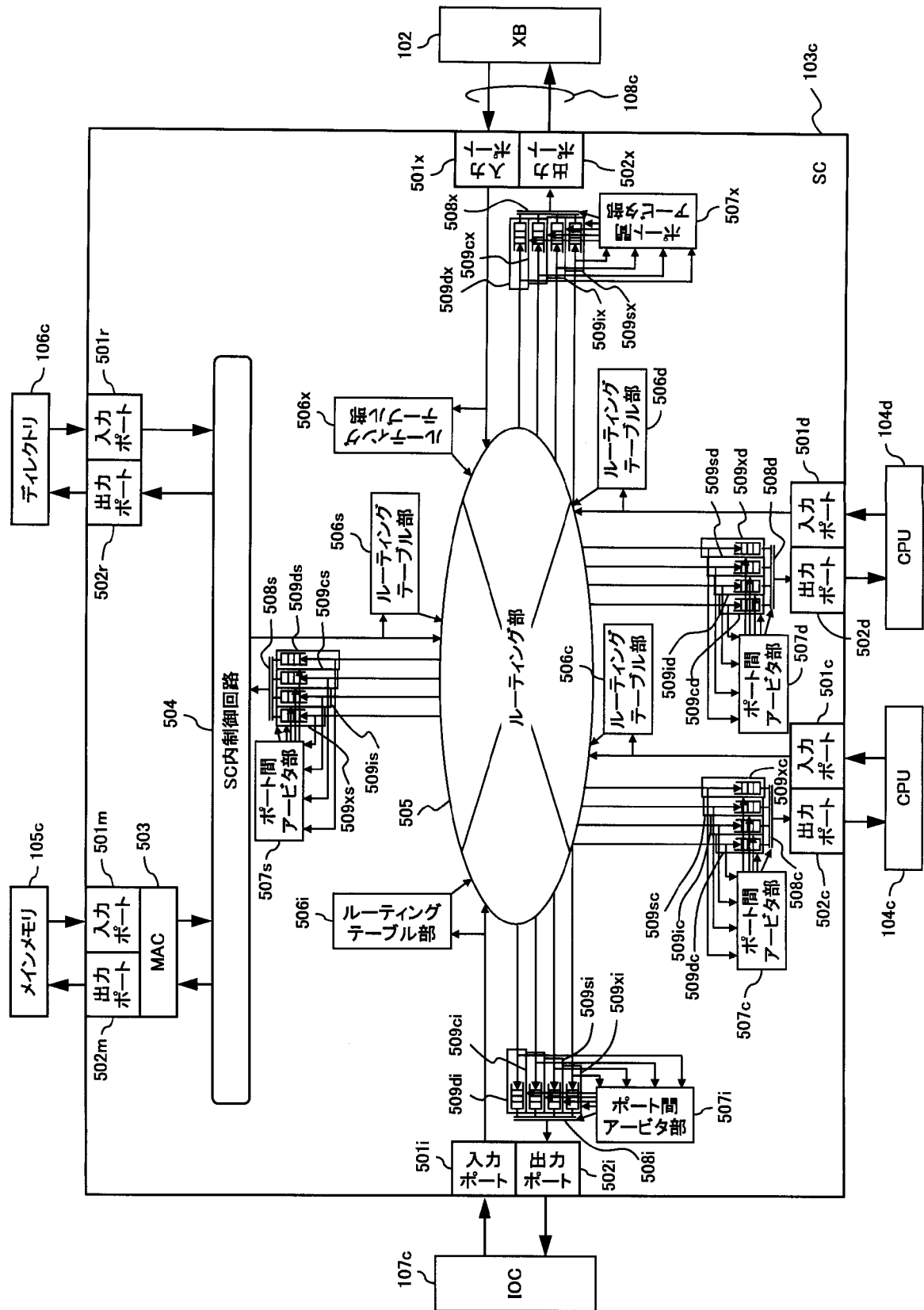
[図8]



[図9]



[図10]



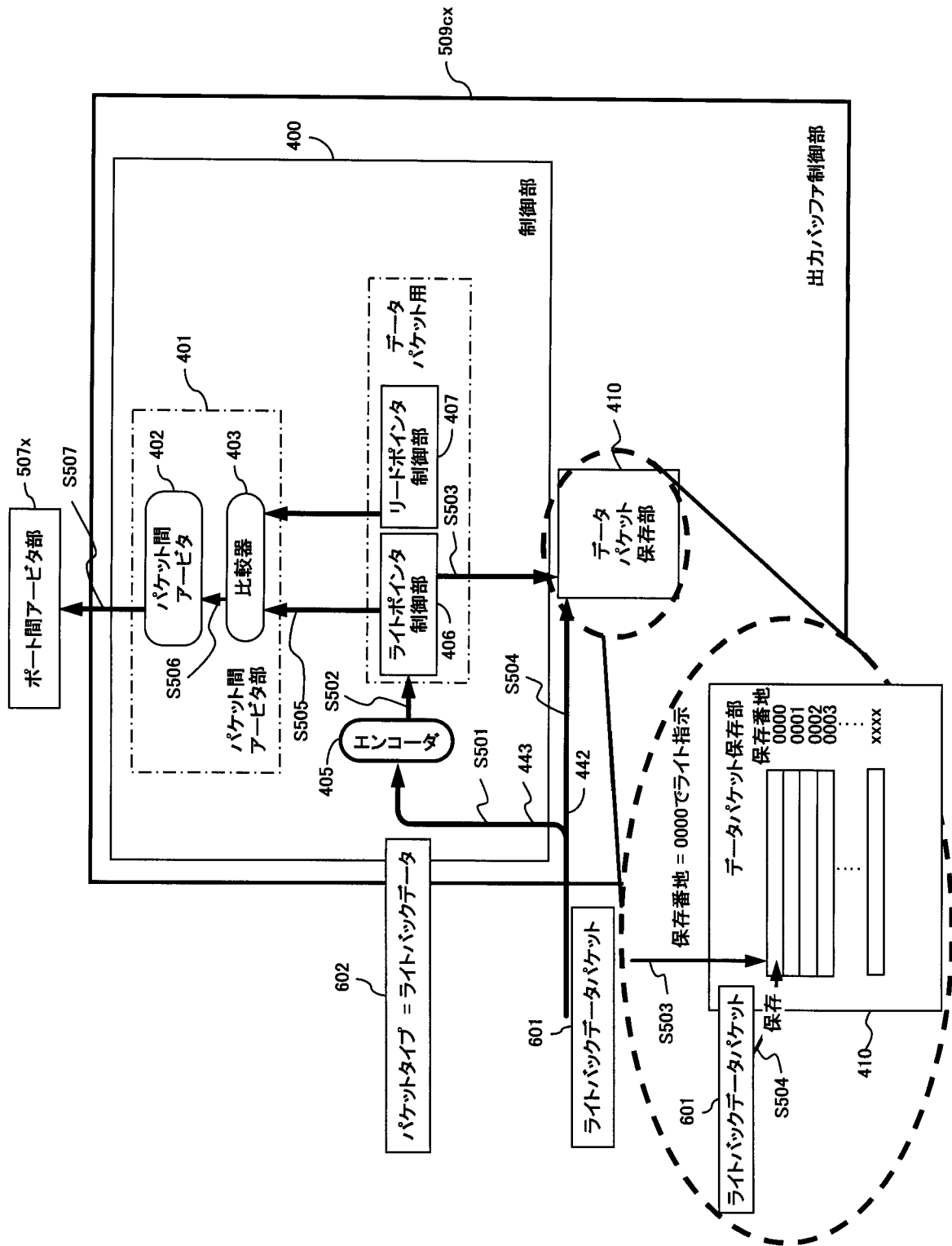
[図11]

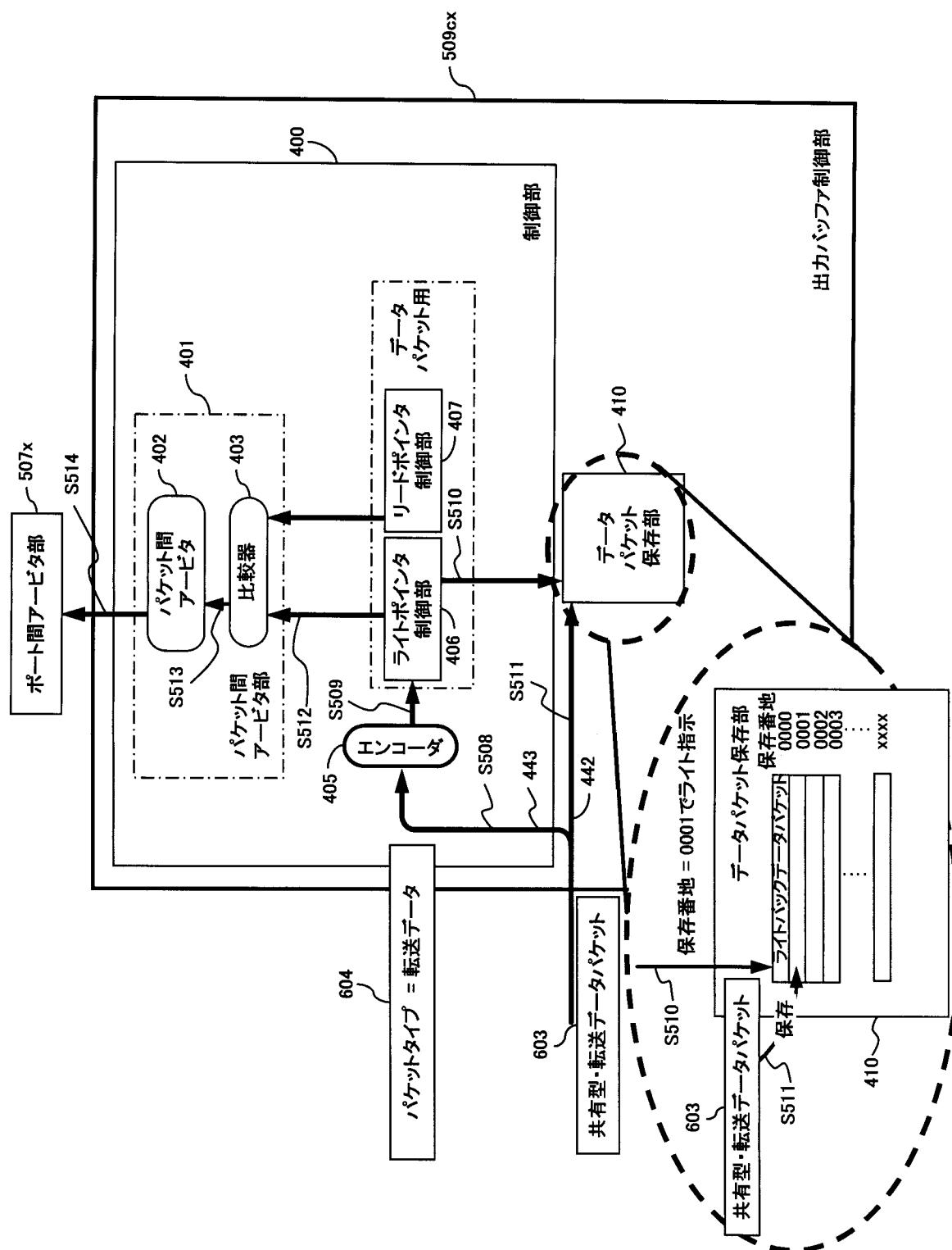
アドレス	所有者	ステート
xxxx	CPU104b	exclusive
xxxx	CPU104a	shared
:	:	:
:	:	:
xxxx	CPU104a	shared

ディレクトリ

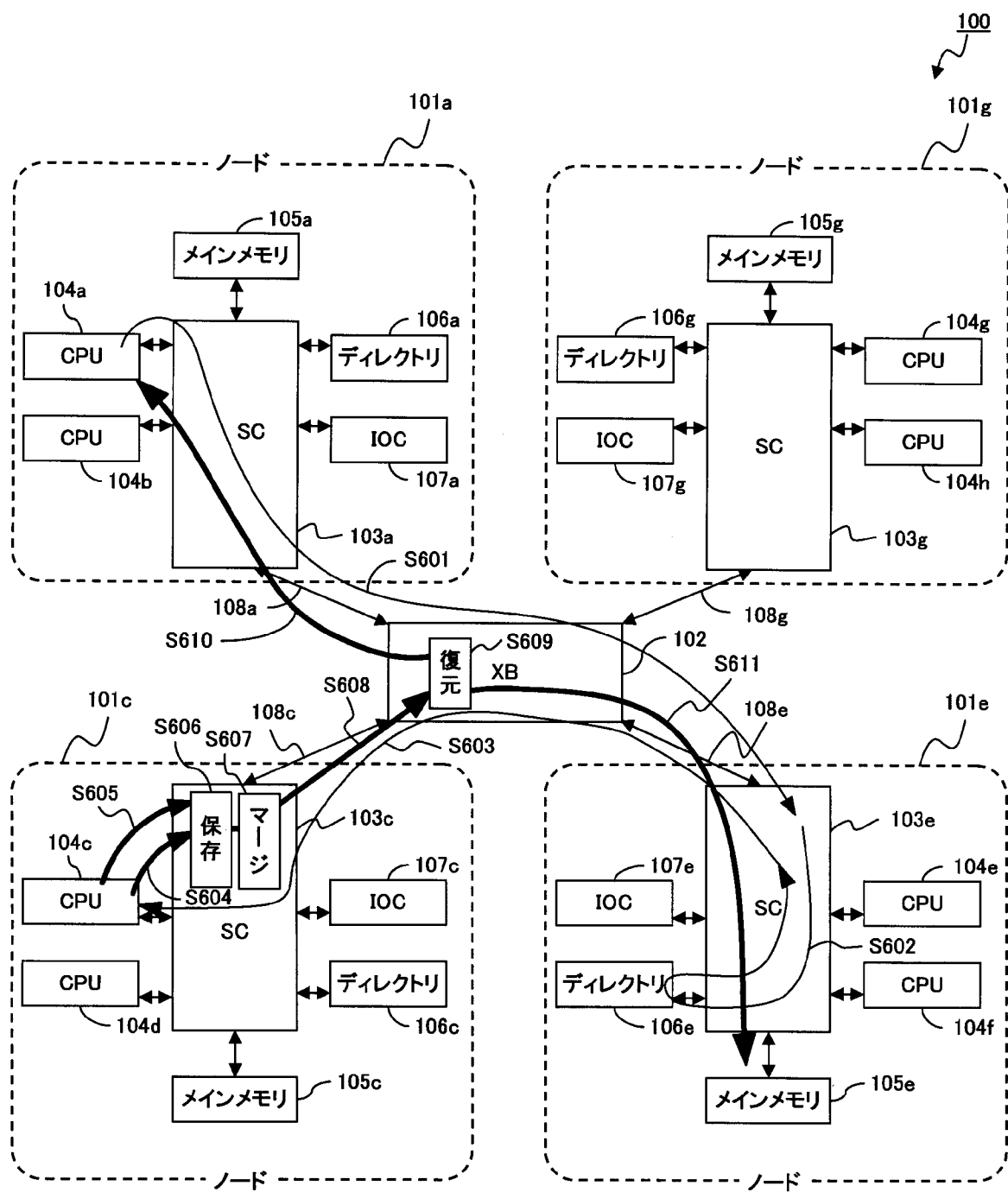
106c

[図12]

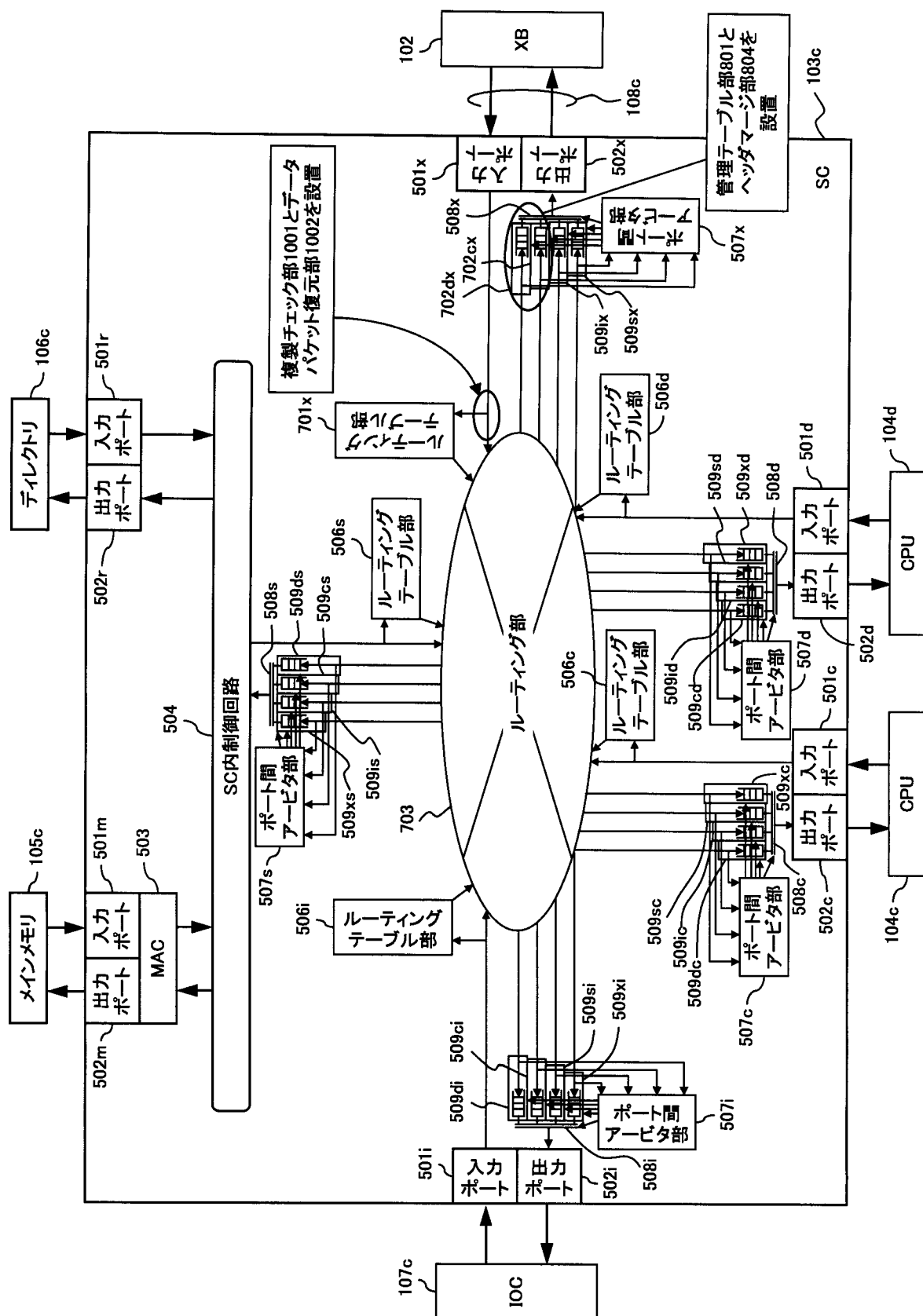




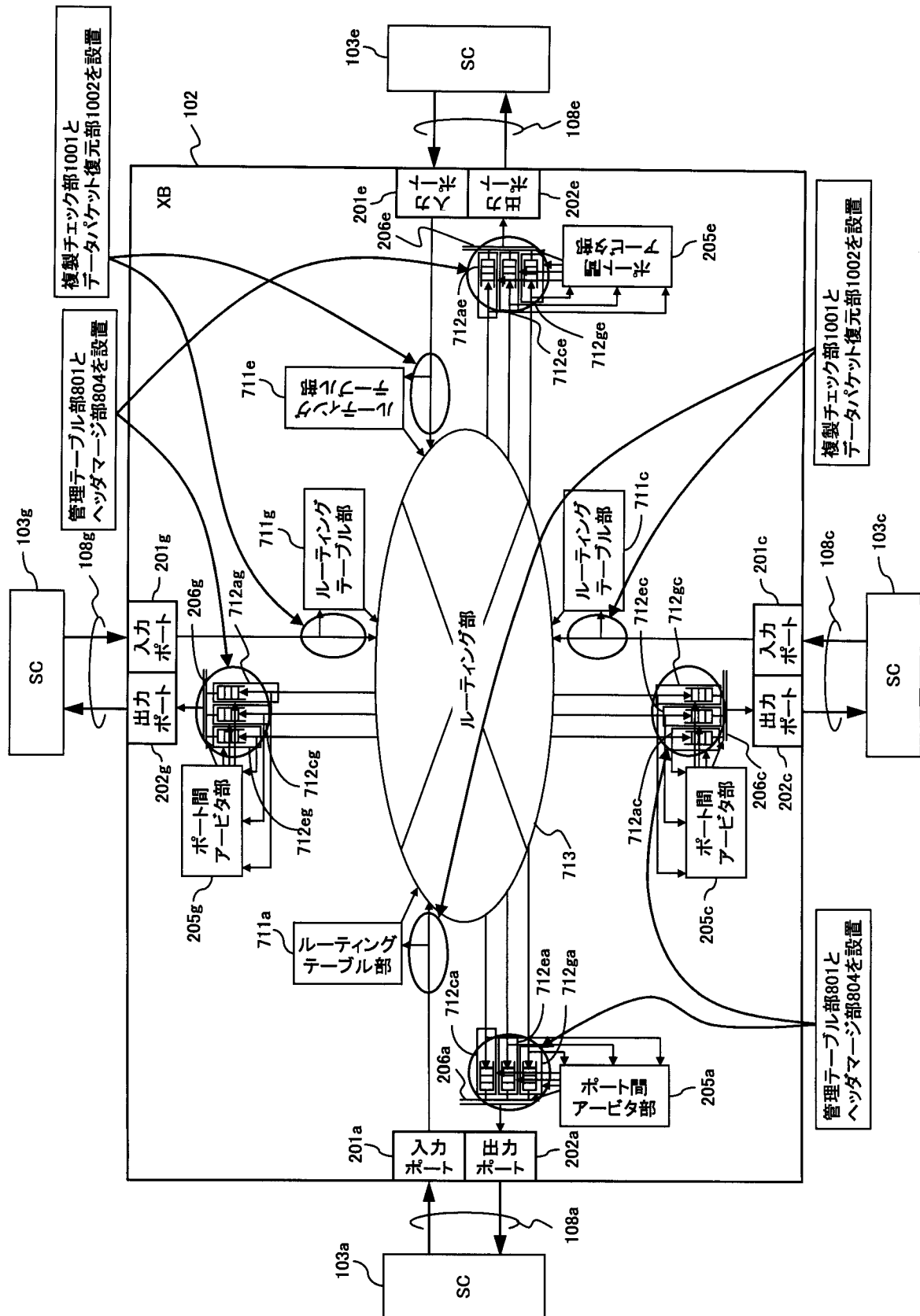
[図14]



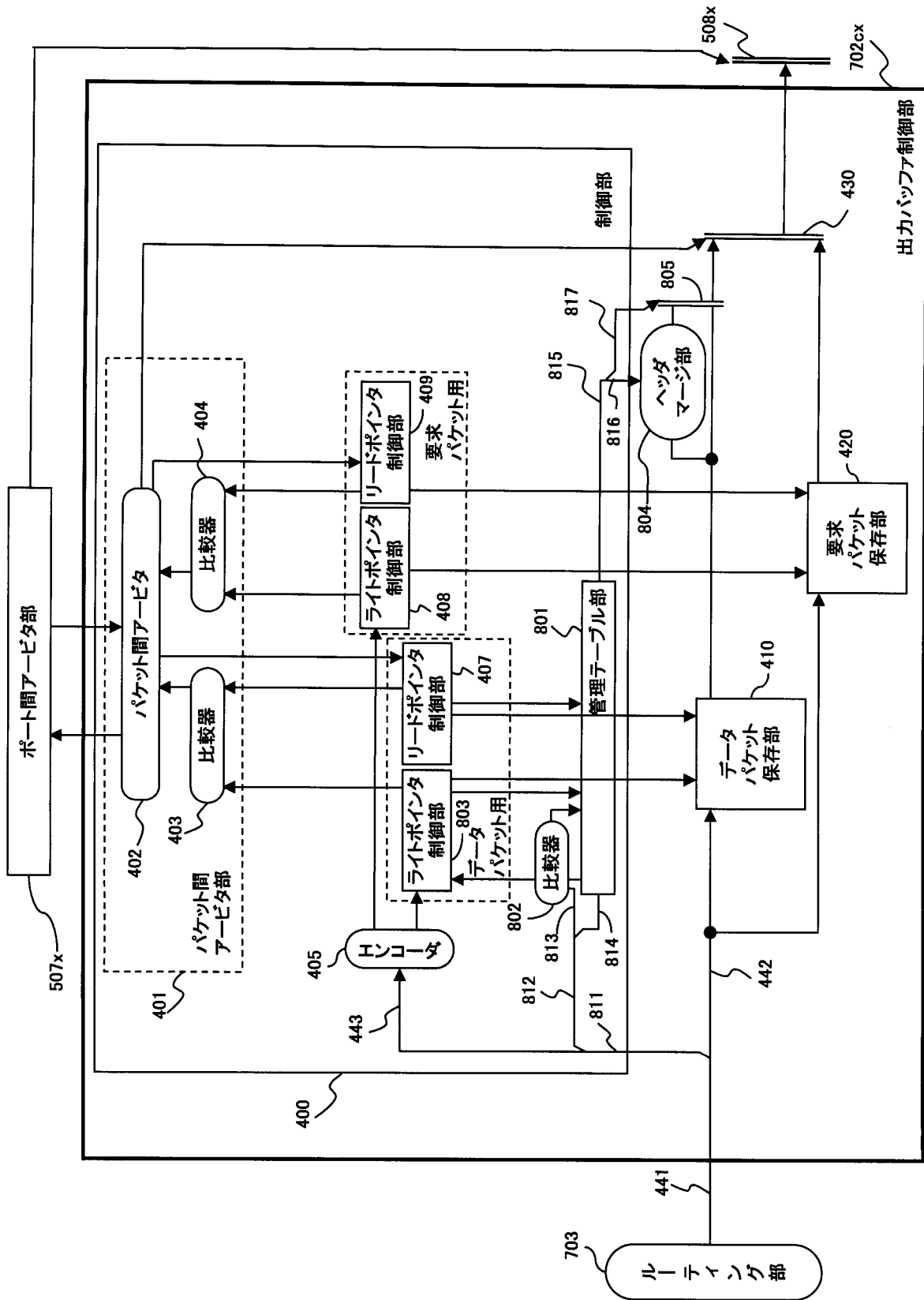
[図15]



[図16]

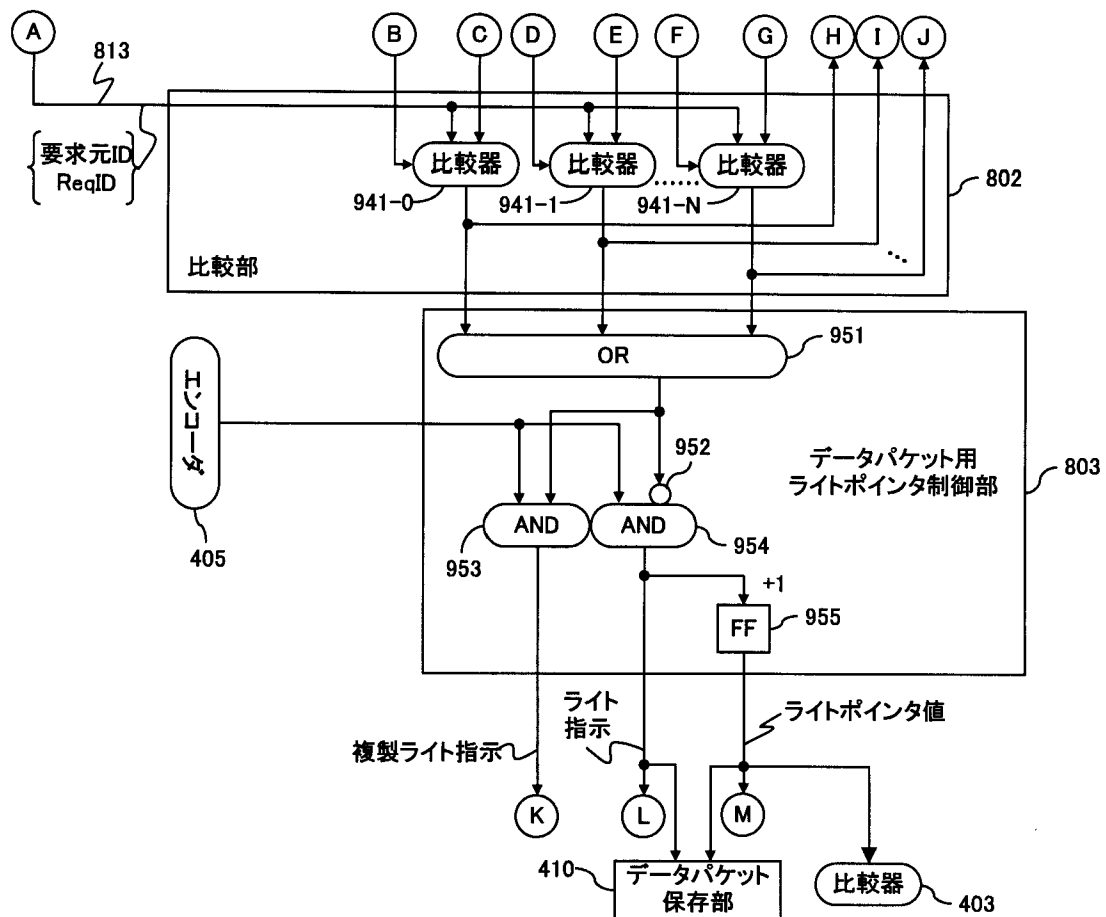


[図17]

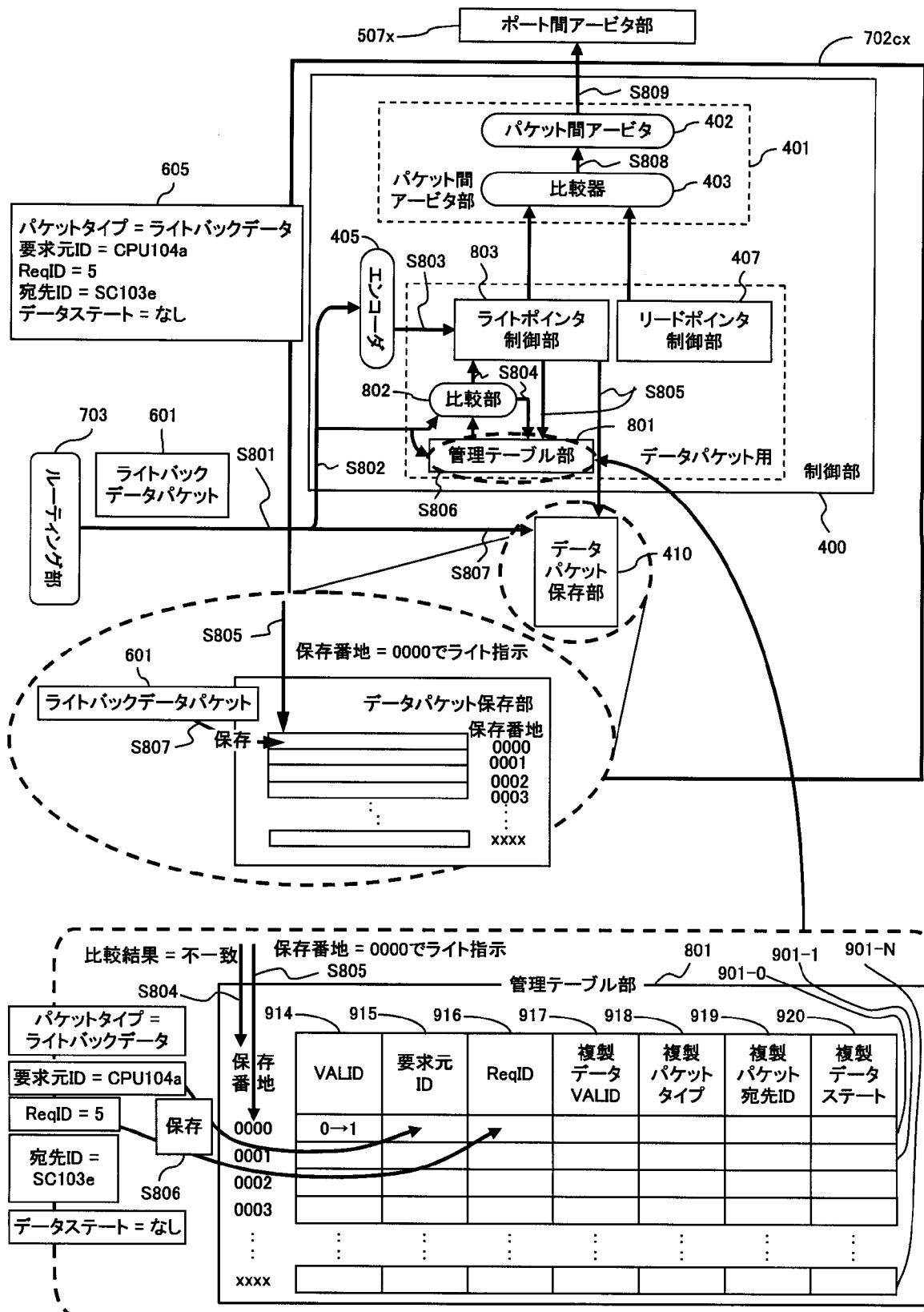


[illegible]

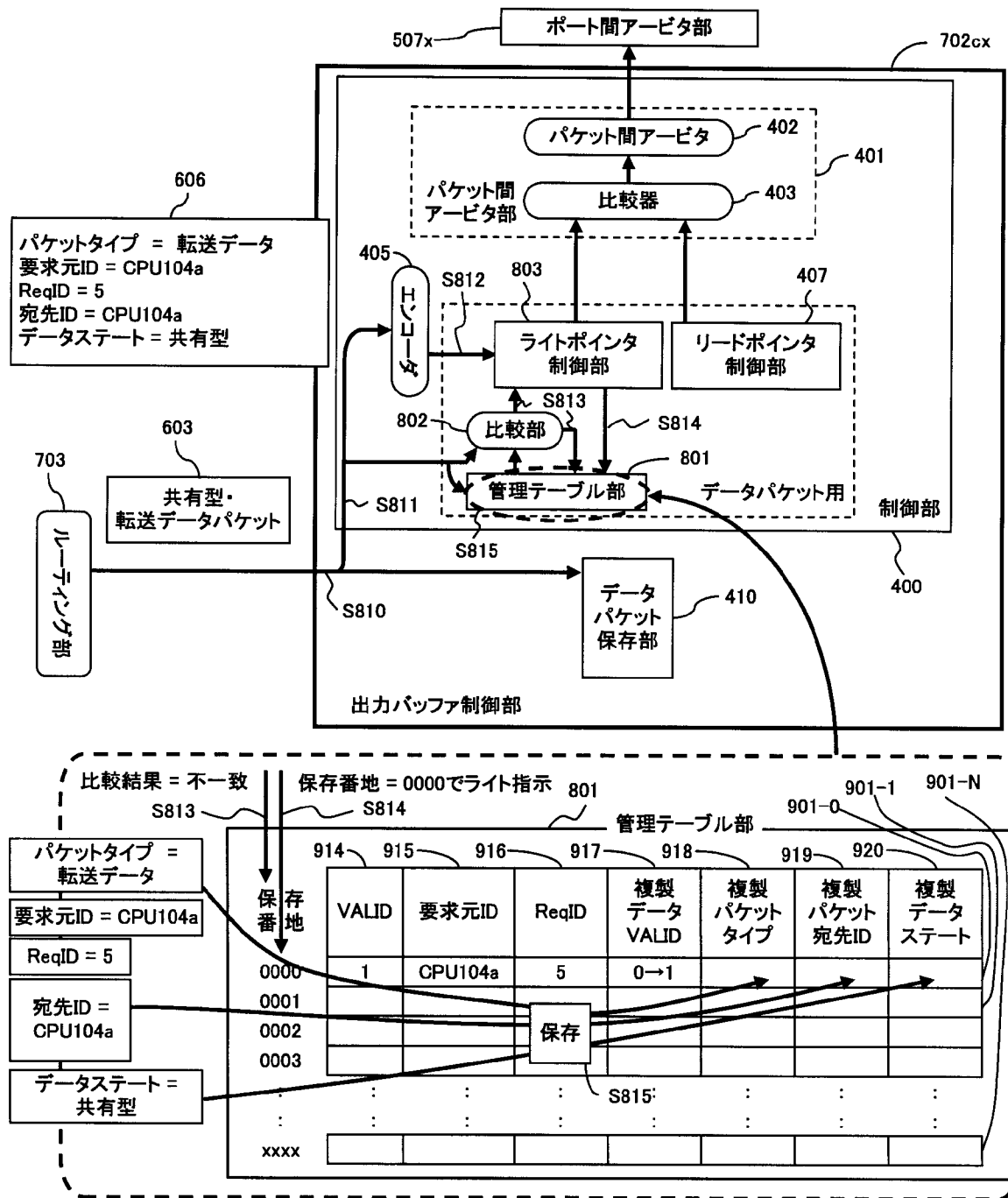
[図18B]



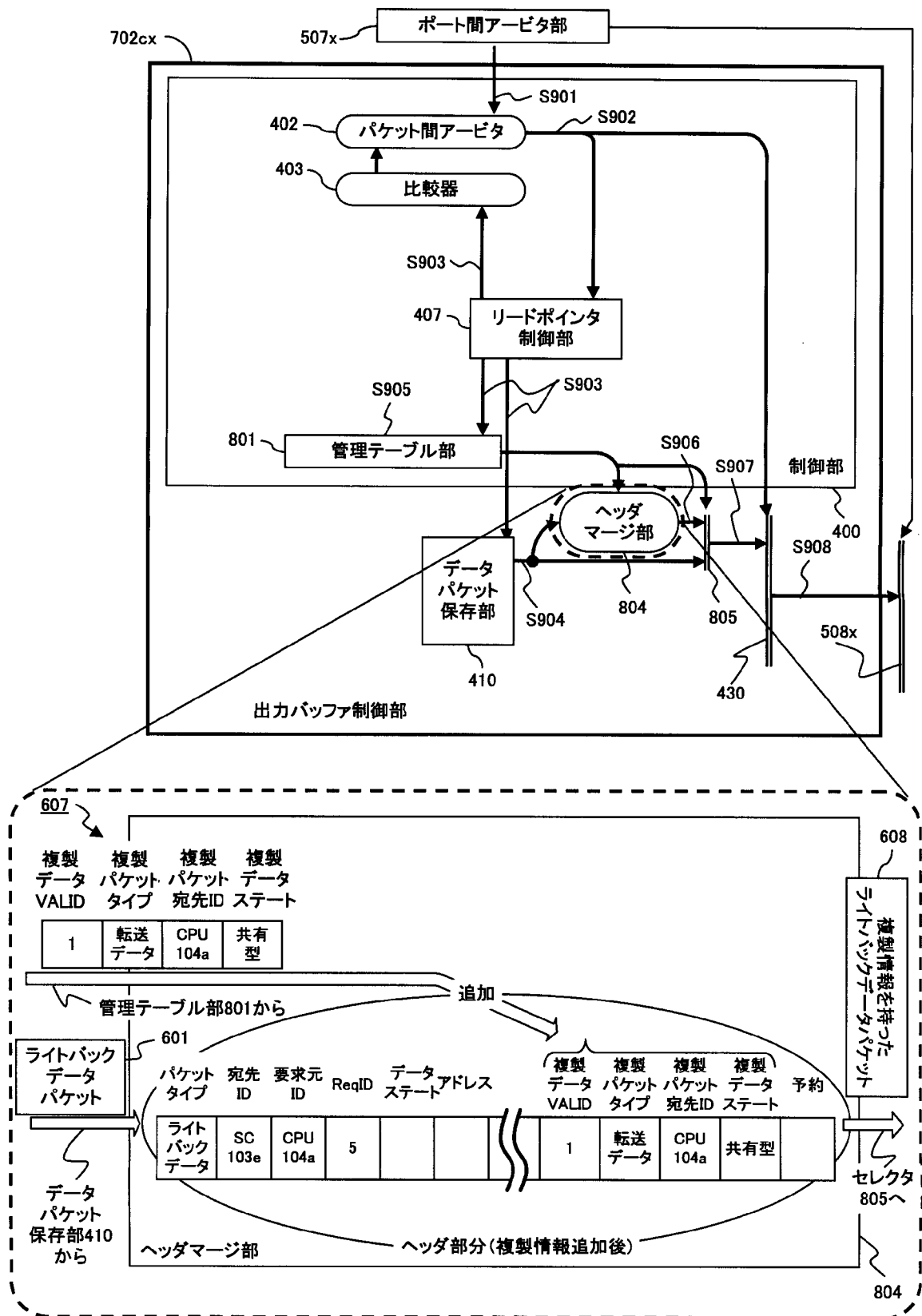
[図20]



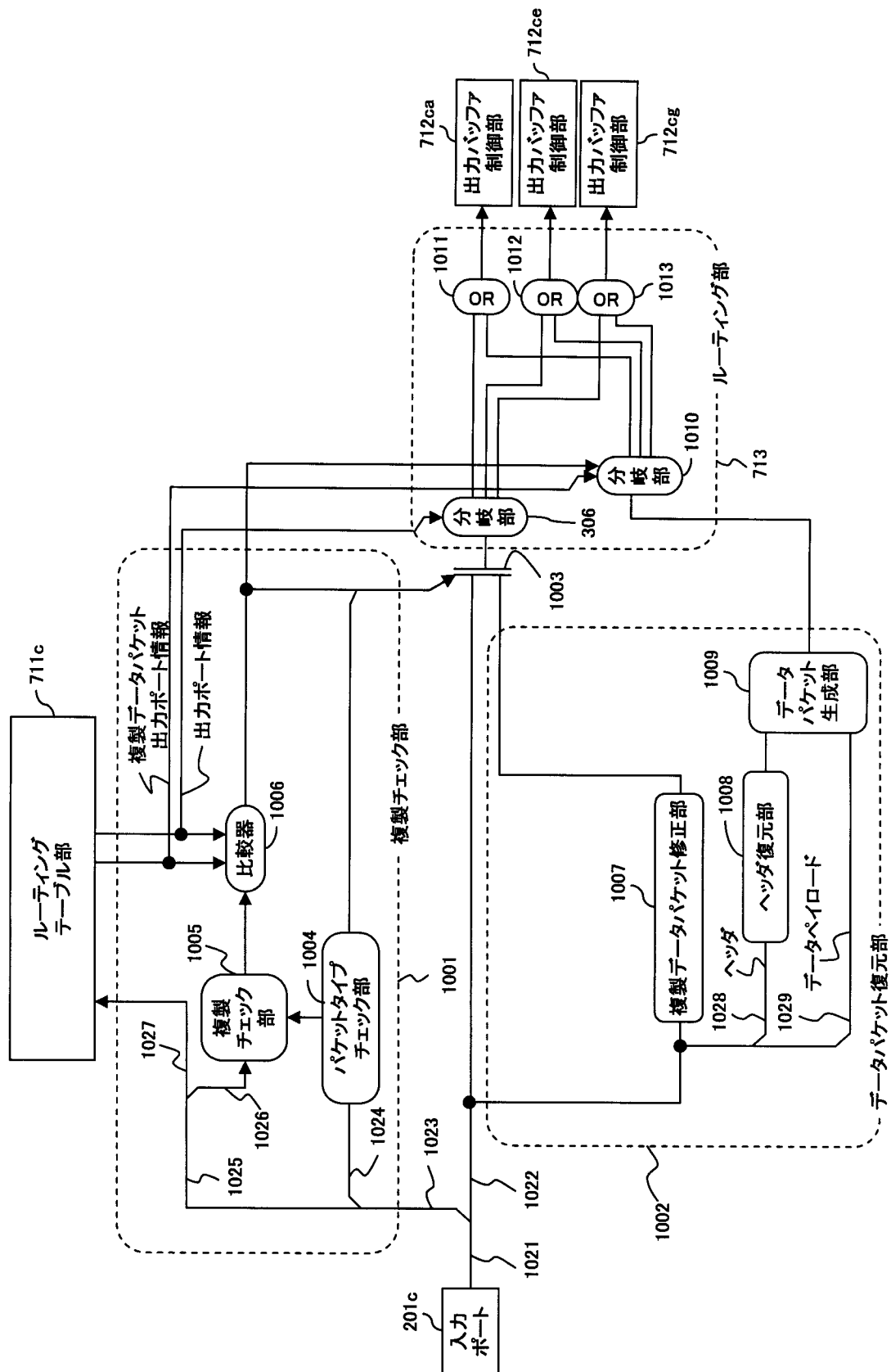
[図21]



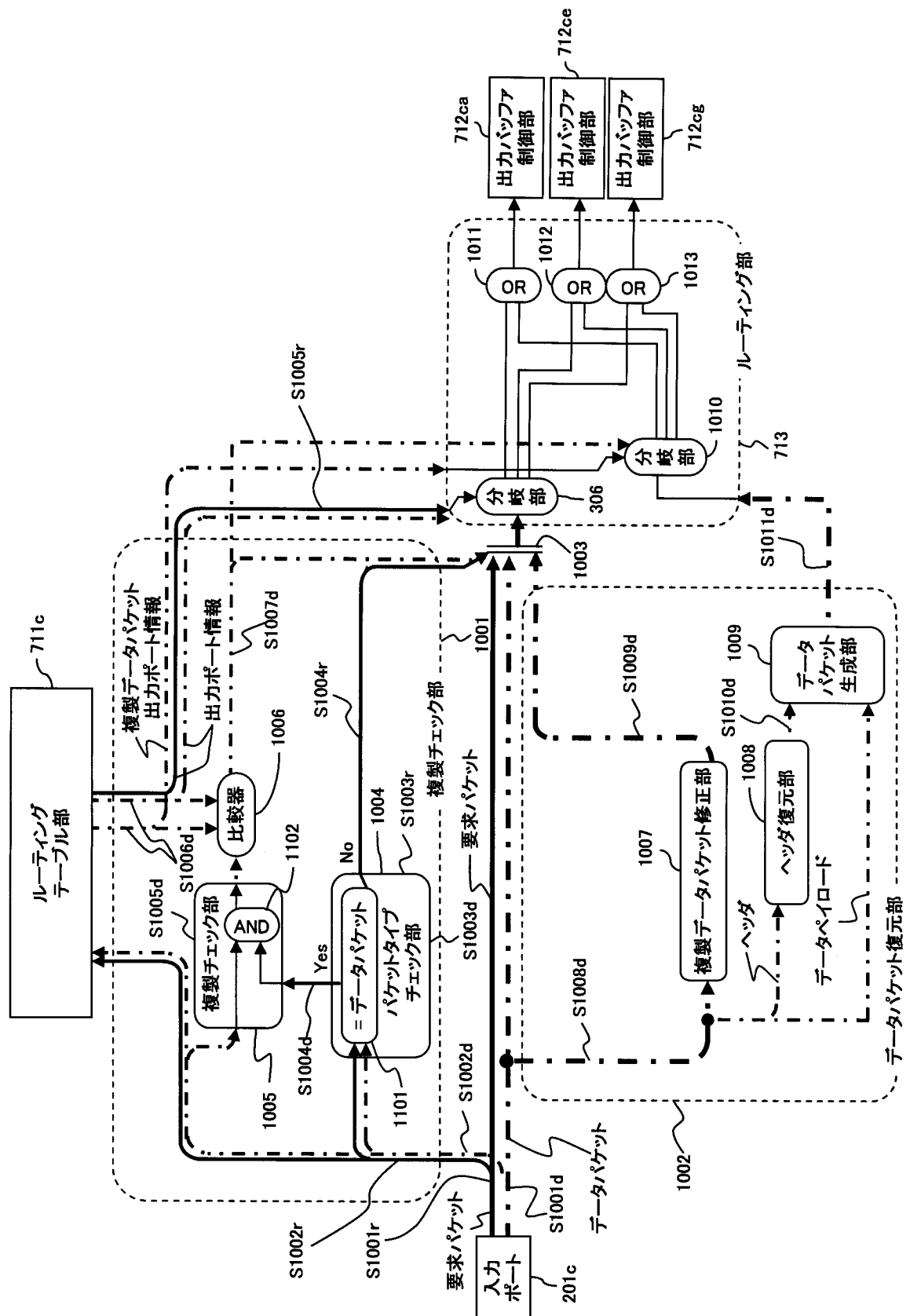
[図22]



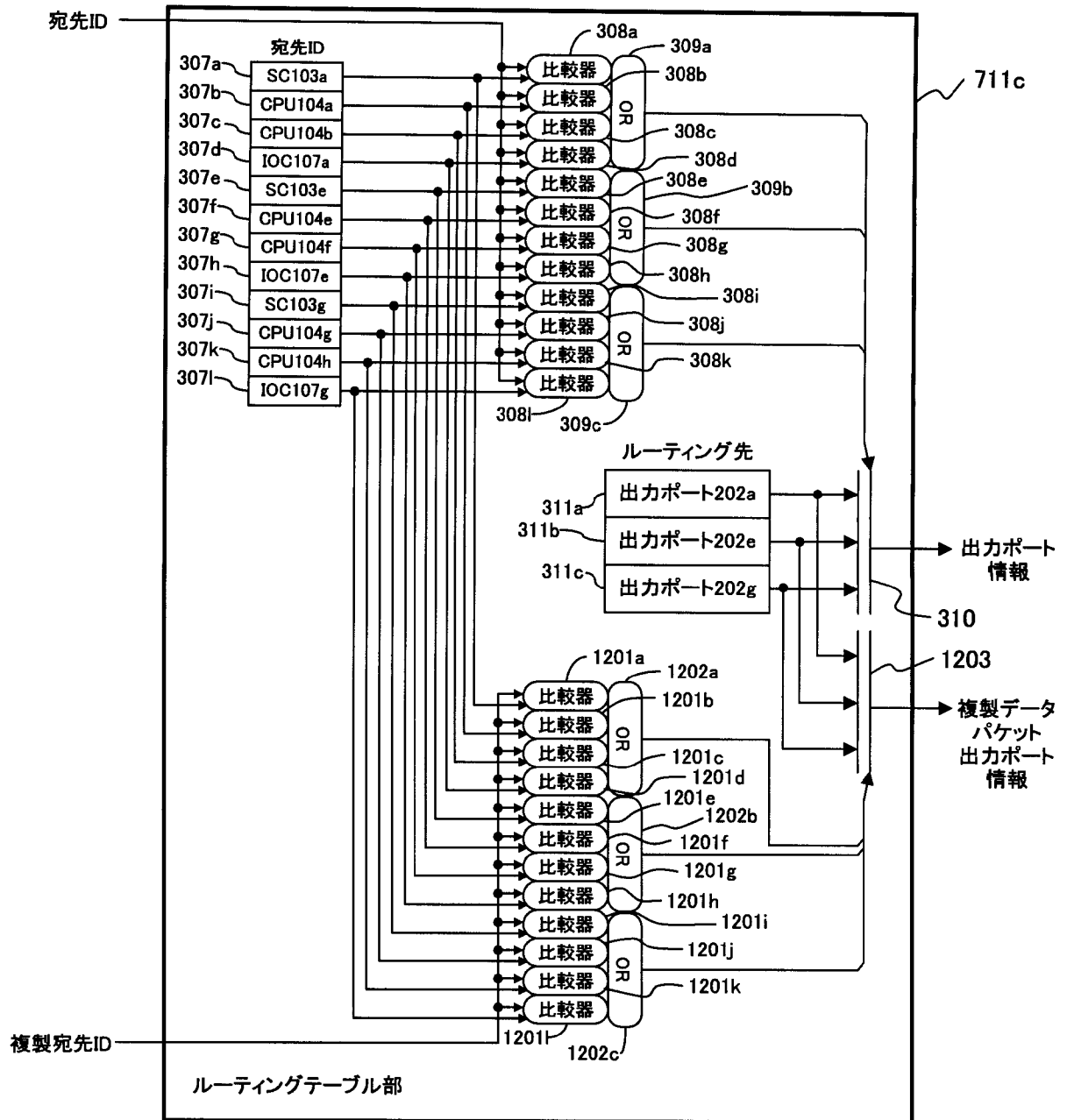
[図23]



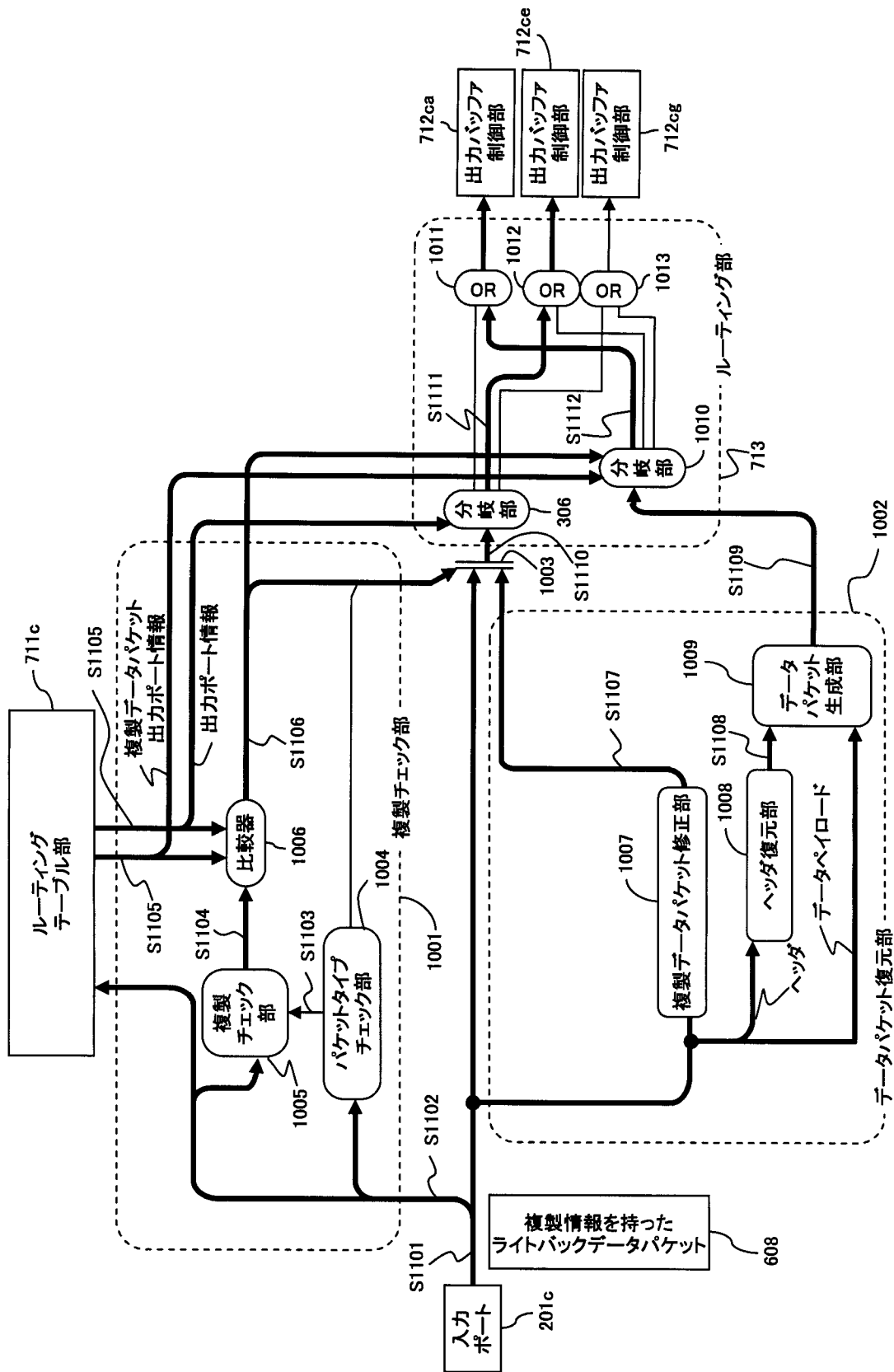
[図24]



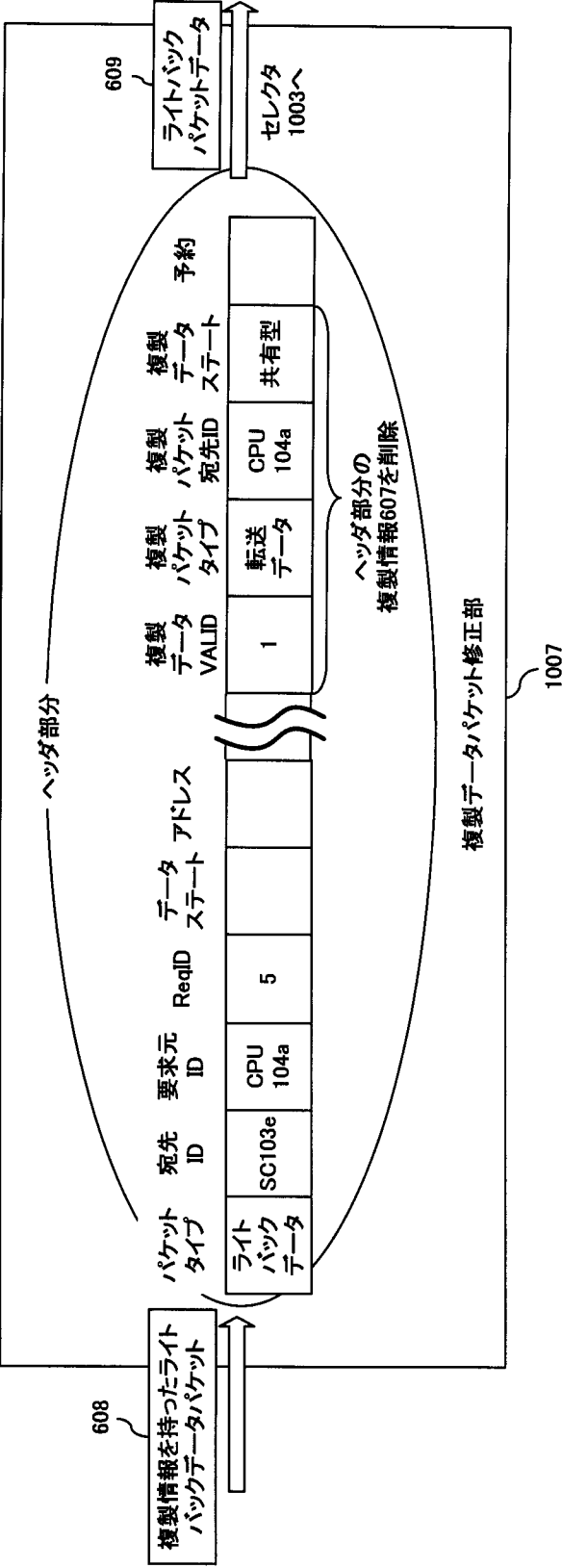
[図25]



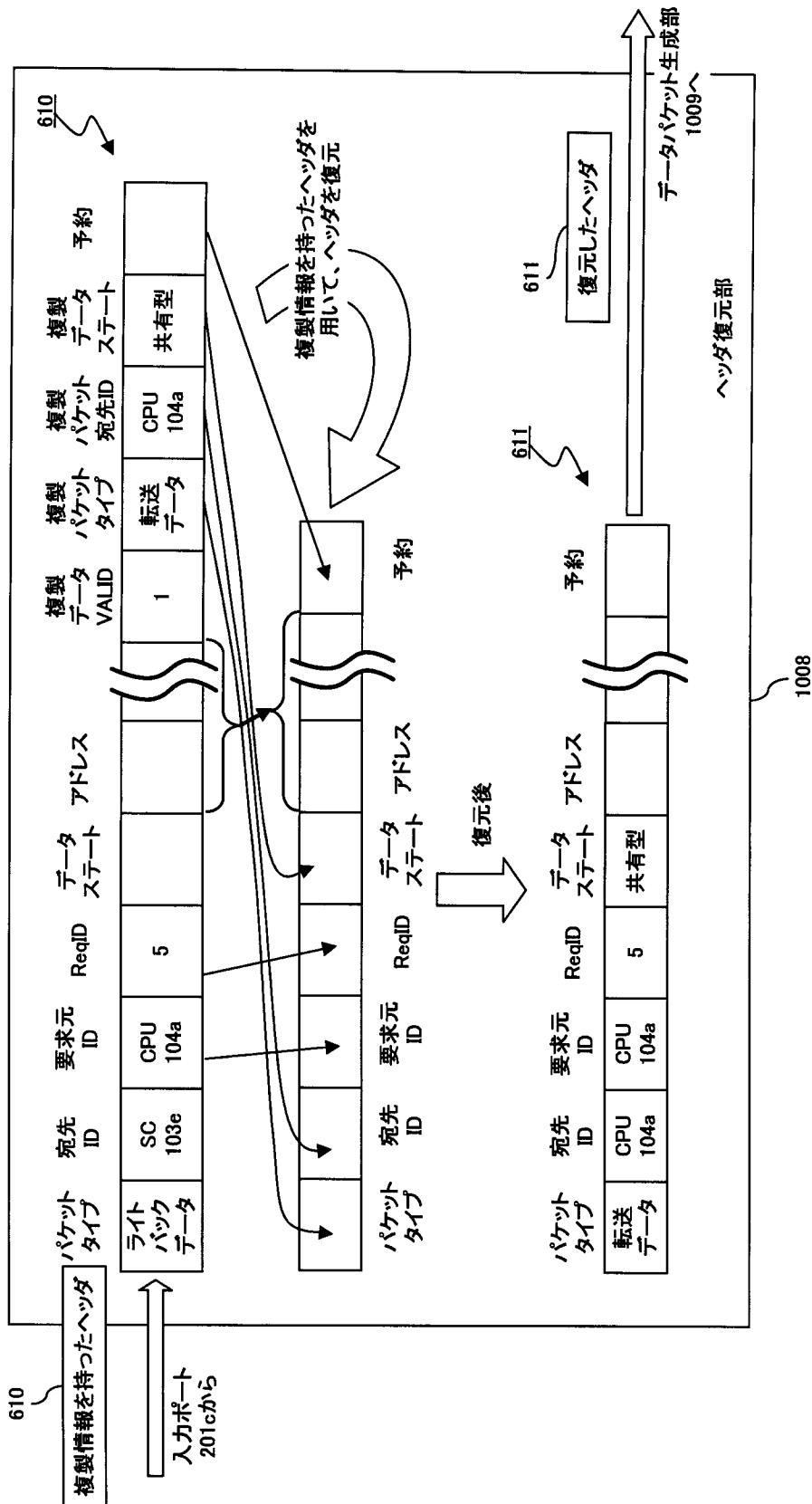
[図26]



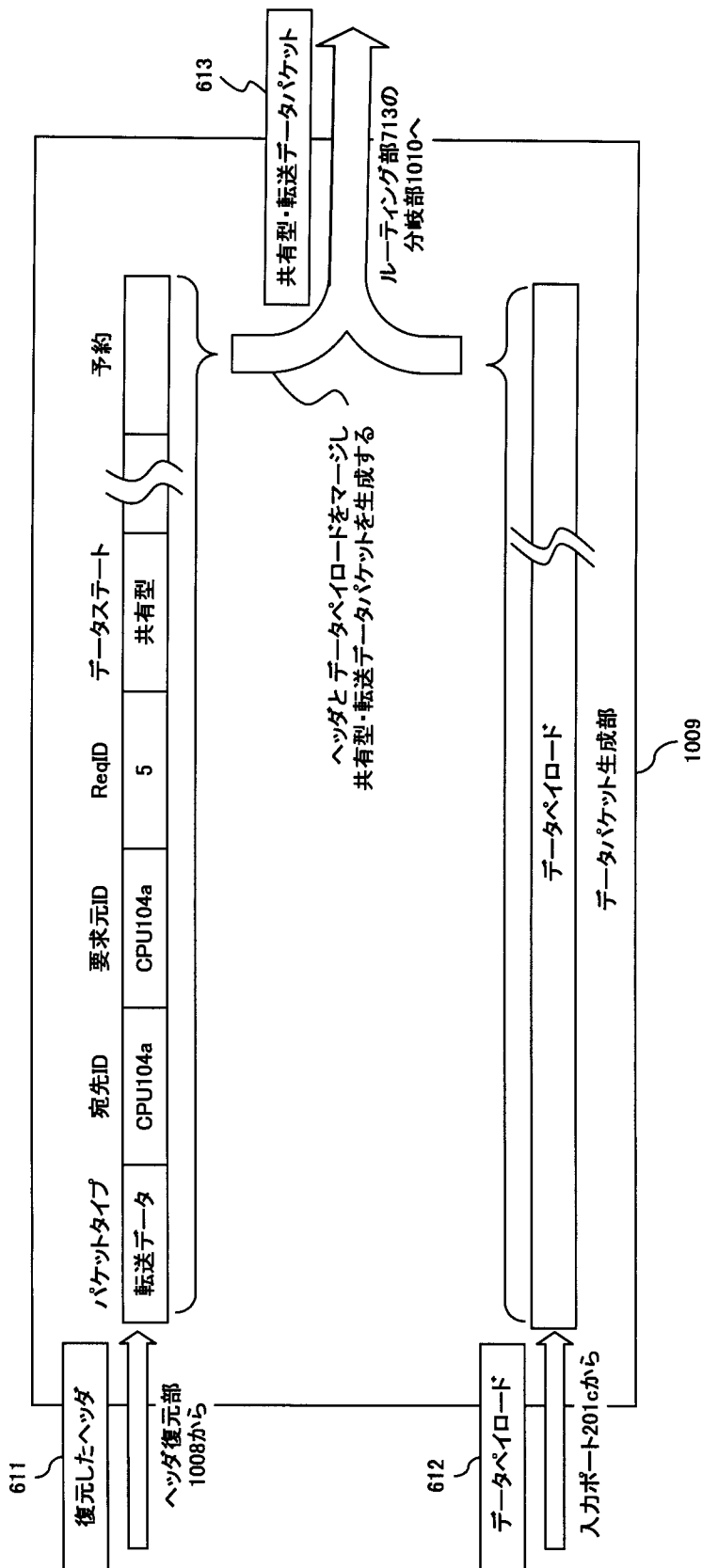
[図27]



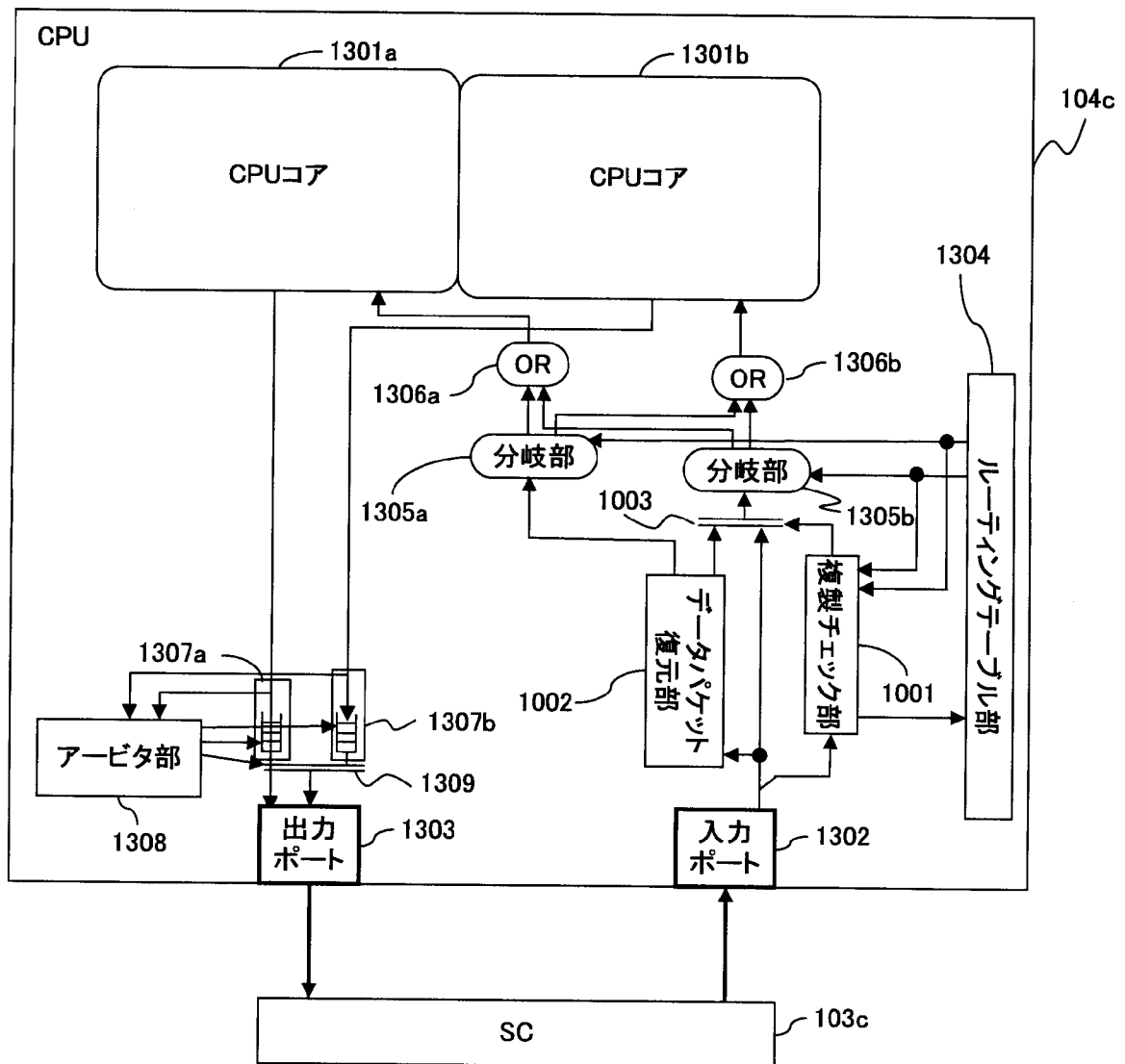
[図28]



[図29]



[図30]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2008/003216

A. CLASSIFICATION OF SUBJECT MATTER

G06F15/173 (2006.01) i, G06F12/08 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F15/173, G06F12/08

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2008
Kokai Jitsuyo Shinan Koho	1971-2008	Toroku Jitsuyo Shinan Koho	1994-2008

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 8-235141 A (NEC Kofu Ltd.), 13 September, 1996 (13.09.96), Full text; all drawings & CA 2170458 C & DE 19607531 A1 & US 5890007 A	1-18
A	JP 9-138782 A (Hitachi, Ltd.), 27 May, 1997 (27.05.97), Full text; all drawings & US 6011791 A & US 6728258 B1	1-18
A	JP 2003-132037 A (Fujitsu Ltd.), 09 May, 2003 (09.05.03), Full text; all drawings & US 2003/0023666 A1	1-18

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
25 December, 2008 (25.12.08)

Date of mailing of the international search report
13 January, 2009 (13.01.09)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

A. 発明の属する分野の分類（国際特許分類（I P C））
Int.Cl. G06F15/173(2006.01)i, G06F12/08(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F15/173, G06F12/08

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 0 8 年
日本国実用新案登録公報	1 9 9 6 - 2 0 0 8 年
日本国登録実用新案公報	1 9 9 4 - 2 0 0 8 年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	J P 8 - 2 3 5 1 4 1 A（甲府日本電気株式会社） 1 9 9 6 . 0 9 . 1 3, 全文, 全図 & C A 2 1 7 0 4 5 8 C & D E 1 9 6 0 7 5 3 1 A 1 & U S 5 8 9 0 0 0 7 A	1 - 1 8

☒ C 欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日

2 5 . 1 2 . 2 0 0 8

国際調査報告の発送日

1 3 . 0 1 . 2 0 0 9

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号 1 0 0 - 8 9 1 5

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

久保 正典

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 4 5

5 B

9 6 4 2

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	JP 9-138782 A (株式会社日立製作所) 1997. 05. 27, 全文, 全図 & US 6011791 A & US 6728258 B1	1-18
A	JP 2003-132037 A (富士通株式会社) 2003. 05. 09, 全文, 全図 & US 2003/0023666 A1	1-18