

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5772911号

(P5772911)

(45) 発行日 平成27年9月2日 (2015.9.2)

(24) 登録日 平成27年7月10日 (2015.7.10)

(51) Int.Cl.

F I

G 0 6 F 11/18 (2006.01)

G 0 6 F 11/18 3 1 0 A

G 0 6 F 11/18 3 1 0 C

G 0 6 F 11/18 3 1 0 F

請求項の数 9 (全 34 頁)

(21) 出願番号	特願2013-200874 (P2013-200874)	(73) 特許権者	000004237
(22) 出願日	平成25年9月27日 (2013.9.27)		日本電気株式会社
(65) 公開番号	特開2015-69270 (P2015-69270A)		東京都港区芝五丁目7番1号
(43) 公開日	平成27年4月13日 (2015.4.13)	(74) 代理人	100124811
審査請求日	平成25年9月27日 (2013.9.27)		弁理士 馬場 資博
		(74) 代理人	100088959
			弁理士 境 廣巳
		(72) 発明者	白野 康之
			東京都港区芝五丁目7番1号 日本電気株式会社社内
		審査官	田中 幸雄

最終頁に続く

(54) 【発明の名称】 フォールトトレラントシステム

(57) 【特許請求の範囲】

【請求項1】

互いに同一のハードウェアで構成された複数のサブシステムを備えたフォールトトレラントシステムであって、

前記複数のサブシステムは、

前記複数のサブシステム間においてロックステップ同期の状態および非同期の状態で作可能なプロセッサ部と、

前記プロセッサ部に接続される入出力部と、

前記プロセッサ部および前記入出力部間に接続される制御部と、

前記制御部を介して前記複数のサブシステム間を互いに接続する複数の信号伝送路とを有し、

前記制御部は、自サブシステムおよび他サブシステムの前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、他サブシステムの前記プロセッサ部が停止し自サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行い、自サブシステムの前記プロセッ

10

20

サ部が停止し他サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行う

フォールトトレラントシステム。

【請求項 2】

前記制御部は、

自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされる前記データの受信時刻と前記データとからチェックサムを生成し、該生成した前記チェックサムを前記誤り検出用データとして出力する誤り検出用データ生成器を有する

10

請求項 1 に記載のフォールトトレラントシステム。

【請求項 3】

前記制御部は、

自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされる前記データが有する CRC を抽出し、該抽出した前記 CRC を前記誤り検出用データとして出力する誤り検出用データ生成器を有する

請求項 1 に記載のフォールトトレラントシステム。

【請求項 4】

前記制御部は、

自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされる前記データと前記誤り検出用データ生成器から出力された前記誤り検出用データとのうちの何れか一方を選択する第 1 のセレクタを有する

20

請求項 2 または 3 に記載のフォールトトレラントシステム。

【請求項 5】

前記制御部は、

自制御部の前記第 1 のセレクタの出力と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータとを同じ一つの前記信号伝送路へ送信する第 1 の送信部と、

自制御部の前記誤り検出用データ生成器から出力された前記誤り検出用データを一時的に記憶する第 1 のバッファと、

30

前記信号伝送路から受信した前記誤り検出用データを一時的に記憶する第 2 のバッファと、

自サブシステムの前記プロセッサ部から自サブシステムの前記入出力部へアクセスされる前記データを一時的に記憶する第 3 のバッファと、

前記信号伝送路から受信した他の前記サブシステムの前記プロセッサ部から自サブシステムの前記入出力部へアクセスされる前記データを一時的に記憶する第 4 のバッファと、

前記第 1 のバッファに記憶された前記誤り検出用データと前記第 2 のバッファに記憶された前記誤り検出用データとを比較する比較器と、

自サブシステムおよび他サブシステムの前記プロセッサ部が前記ロックステップ同期の状態にあるときは、前記比較器で一致が検出されることを条件に前記第 3 のバッファに記憶された前記データを自サブシステムの前記入出力部へ出力し、他サブシステムの前記プロセッサ部が停止し自サブシステムの前記プロセッサ部が前記ロックステップ非同期の状態にあるときは、前記第 3 のバッファに記憶された前記データを自サブシステムの前記入出力部へ出力し、自サブシステムの前記プロセッサ部が前記ロックステップ非同期の状態にあるときは、前記第 4 のバッファに記憶された前記データを自サブシステムの前記入出力部へ出力する第 2 のセレクタと

40

を有する

請求項 4 に記載のフォールトトレラントシステム。

【請求項 6】

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラン

50

トシステムの各サブシステムで用いる制御装置であって、

自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と前記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する複数の信号伝送路に接続され、

自サブシステムおよび他サブシステムの前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、他サブシステムの前記プロセッサ部が停止し自サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行い、自サブシステムの前記プロセッサ部が停止し他サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行うフォールトトレラントシステムで用いる制御装置。

10

【請求項 7】

互いに同一のハードウェアで構成された複数のサブシステムを備え、前記複数のサブシステムは、前記複数のサブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と、前記プロセッサ部に接続される入出力部と、前記プロセッサ部および前記入出力部間に接続される制御部と、前記制御部を介して前記複数のサブシステム間を互いに接続する複数の信号伝送路とを有するフォールトトレラントシステムの動作方法であって、

20

前記制御部が、

自サブシステムおよび他サブシステムの前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、他サブシステムの前記プロセッサ部が停止し自サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行い、自サブシステムの前記プロセッサ部が停止し他サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行うフォールトトレラントシステムの動作方法。

30

【請求項 8】

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いられ、自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と前記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する複数の信号伝送路に接続される制御装置の動作方法であって、

40

前記制御装置が、自サブシステムおよび他サブシステムの前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、他サブシステムの前記プロセッサ部が停止し自サブシステムの前記プロセッサ

50

部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行い、自サブシステムの前記プロセッサ部が停止し他サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行う

フォールトトレラントシステムで用いる制御装置の動作方法。

【請求項 9】

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いられ、自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と前記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する複数の信号伝送路に接続された制御装置を構成するコンピュータに、

自サブシステムおよび他サブシステムの前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行わせ、

他サブシステムの前記プロセッサ部が停止し自サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から他サブシステムの前記入出力部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行い、自サブシステムの前記プロセッサ部が停止し他サブシステムの前記プロセッサ部がロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のために使用する制御を行わせる

ためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、フォールトトレラントシステム、それに用いる制御装置、それらの動作方法、およびプログラムに関する。

【背景技術】

【0002】

可用性を高めた計算機システムとして、フォールトトレラントシステムが知られている。フォールトトレラントシステムでは、アプリケーションやOSが特別な処理を必要としないで、透過的にサービスを継続することができる。フォールトトレラントシステムを実現する方式は、ハードウェア方式とソフトウェア方式とに大別される。本発明は、ハードウェア方式のフォールトトレラントシステムの改良に関する。

【0003】

ハードウェア方式のフォールトトレラントシステムは、CPU、メモリ、ストレージなどの主要なハードウェアコンポーネントを冗長化し、あるコンポーネントに障害が発生した場合は、そのコンポーネントを切り離して動作を継続する。CPUやメモリを含むモジュールをCPUサブシステム、各種IOデバイスを含むモジュールをIOサブシステムと呼ぶ。コンポーネントを二重化する一般的なフォールトトレラントシステムでは、CPUサブシステムとIOサブシステムとで二重化の方式が異なる。CPUサブシステムは、クロック単位でハードウェアの動作を完全に一致させる。これをロックステップ同期と呼ぶ。二重化されたCPUサブシステムは、両方とも完全に同じ動作をしているので、障害発生時はその障害が発生したCPUサブシステムを論理的に切り離し、正常なCPUサブシ

10

20

30

40

50

システムに瞬時に処理を切り換えて動作を継続させる。一方、I Oサブシステムは、ロックステップ同期こそしていないが、障害が発生した場合、直ちに他方のI Oサブシステムに切り換える。

【0004】

ロックステップ同期の状態において、CPUサブシステムの不正な動作を検出する方法として、各CPUサブシステムからI Oサブシステムへアクセスするデータを比較する方法がある。その際、各サブシステムが、自サブシステムのCPUで生成したアクセスデータからチェックサムを生成してクロスリンクを介して他サブシステムへ送信する一方、上記生成したチェックサムとクロスリンクを介して他サブシステムから受信したチェックサムとを比較してサブシステム間の処理の不一致を検知することが、本発明に関連する第1の関連技術として提案されている（例えば特許文献1参照）。

10

【0005】

その他、フォールトトレラントシステムに関して、以下のような先行技術がある。

【0006】

まず、自他システム間でクロックステップ同期により同じタイミングで動作するCPUサブシステムと、これに接続されるI Oサブシステムと、両者間に接続されるFTコントローラとを有する2つのシステムをクロスリンクによって互いに接続したフォールトトレラントシステムにおいて、FTコントローラが、両システムによるフォールトトレラント用のエラー処理、二重化処理、および再同期化処理を行うための複数のシステム動作をこれに対応する複数ステートとして所定のイベント信号に関連付けて管理し、イベント信号に応じて、システム毎にステートを遷移させながら複数のシステム動作を選択してCPUサブシステムに実行させることが、本発明に関連する第2の関連技術として提案されている（例えば特許文献2参照）。

20

【0007】

また、上記フォールトトレラントシステムにおいて、CPUサブシステムからI Oサブシステムへのアクセスパケットに、アクセス元およびアクセス先のIDコードと、同期アクセスか否かの同期情報とを含むタグ情報を付与し、各々のシステムのアクセス比較部が、アクセスパケットに付与されたタグ情報に基づいて、複数のCPUサブシステムがロックステップ同期状態にある場合のアクセス処理を行うか、非同期状態に応じたアクセス処理を行うかを決定することが、本発明に関連する第3の関連技術として提案されている（例えば特許文献3参照）。

30

【先行技術文献】

【特許文献】

【0008】

【特許文献1】特開2010-218370号公報

【特許文献2】特開2006-178616号公報

【特許文献3】特開2006-178615号公報

【特許文献4】特開2006-178618号公報

【発明の概要】

【発明が解決しようとする課題】

40

【0009】

本発明に関連する第1の関連技術では、パケットそのものでなくパケットから生成したチェックサムをクロスリンクを介して他サブシステムへ送信する一方、上記生成したチェックサムとクロスリンクを介して他サブシステムから受信したチェックサムとを比較してサブシステム間の処理の不一致を検知している。これにより、パケットそのものをクロスリンクを介して他サブシステムへ送信する場合に比べて、クロスリンクを流れるデータ量を低減することができる。しかしながら、パケットそのものでなくパケットから生成したチェックサムをクロスリンクを介して他サブシステムへ送信する上記第1の関連技術では、フォールトトレラントシステムをロックステップ非同期の状態で作動作させるのは困難である。その理由は、パケットから生成したチェックサムをクロスリンクを介して他サブシ

50

システムへ送信しても、他サブシステム側にその元となるパケットと同じパケットが存在しないロックステップ非同期の状態では他サブシステムの入出力部をアクセスできないためである。他方、本発明に関連する第3の関連技術では、フォールトトレラントシステムをロックステップ同期の状態およびロックステップ非同期の状態の双方の状態で作動させることが可能であるが、パケットそのものをクロスリンクを介して送信するため、クロスリンクを流れるデータ量を低減することは困難である。

【0010】

本発明の目的は、上述した課題、すなわち、サブシステム間を接続する信号伝送路を流れるデータ量を低減しつつ、フォールトトレラントシステムをロックステップ同期の状態およびロックステップ非同期の状態の双方の状態で作動させるのは困難である、という課題を解決するフォールトトレラントシステムを提供することにある。

10

【課題を解決するための手段】

【0011】

本発明の第1の観点に係るフォールトトレラントシステムは、
互いに同一のハードウェアで構成された複数のサブシステムを備えたフォールトトレラントシステムであって、

上記複数のサブシステムは、

上記複数のサブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と、

上記プロセッサ部に接続される入出力部と、

20

上記プロセッサ部および上記入出力部に接続される制御部と、

上記制御部を介して上記複数のサブシステム間を互いに接続する信号伝送路とを有し、

上記制御部は、上記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータの他サブシステムへの送信を行う。

【0012】

本発明の第2の観点に係るフォールトトレラントシステムで用いる制御装置は、

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いる制御装置であって、

30

自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と上記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する信号伝送路に接続され、

上記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータの他サブシステムへの送信を行う。

【0013】

40

本発明の第3の観点に係るフォールトトレラントシステムの動作方法は、

互いに同一のハードウェアで構成された複数のサブシステムを備え、上記複数のサブシステムは、上記複数のサブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と、上記プロセッサ部に接続される入出力部と、上記プロセッサ部および上記入出力部に接続される制御部と、上記制御部を介して上記複数のサブシステム間を互いに接続する信号伝送路とを有するフォールトトレラントシステムの動作方法であって、

上記制御部が、

上記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータから生成した誤り検出用データの他

50

サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータの他サブシステムへの送信を行う。

【0014】

本発明の第4の観点に係るフォールトトレラントシステムで用いる制御装置の動作方法は、

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いられ、自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と上記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する信号伝送路に接続される制御装置の動作方法であって、

10

上記制御装置が、上記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータの他サブシステムへの送信を行う。

【0015】

本発明の第4の観点に係るプログラムは、

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いられ、自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と上記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する信号伝送路に接続された制御装置を構成するコンピュータに、

20

上記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行わせ、

ロックステップ非同期の状態にあるときは、自サブシステムの上記プロセッサ部から上記入出力部へアクセスされるデータの他サブシステムへの送信を行わせる。

【発明の効果】

【0016】

30

本発明は上述した構成を有するため、サブシステム間を接続する信号伝送路を流れるデータ量を低減しつつ、フォールトトレラントシステムをロックステップ同期の状態およびロックステップ非同期の状態の双方の状態で作動させることができる。

【図面の簡単な説明】

【0017】

【図1】本発明の第1の実施形態のブロック図である。

【図2】本発明の第1の実施形態におけるロックステップ同期状態時の動作の一例を示す図である。

【図3】本発明の第1の実施形態におけるロックステップ同期状態時の動作の他の例を示す図である。

40

【図4】本発明の第1の実施形態におけるロックステップ非同期状態時の動作の一例を示す図である。

【図5】本発明の第1の実施形態におけるロックステップ非同期状態時の動作の他の例を示す図である。

【図6】本発明の第2の実施形態のブロック図である。

【図7】本発明の第2の実施形態におけるロックステップ同期状態時の動作の一例を示す図である。

【図8】本発明の第2の実施形態におけるロックステップ同期状態時の動作の他の例を示す図である。

【図9】本発明の第2の実施形態におけるロックステップ非同期状態時の動作の一例を示

50

す図である。

【図10】本発明の第2の実施形態におけるロックステップ非同期状態時の動作の他の例を示す図である。

【図11】本発明の第3の実施形態のブロック図である。

【図12】本発明の第4の実施形態のブロック図である。

【図13】本発明に関連する第1のフォールトトレラントシステムにおけるロックステップ同期状態時の動作の一例を示す図である。

【図14】本発明に関連する第1のフォールトトレラントシステムにおけるロックステップ同期状態時の動作の他の例を示す図である。

【図15】本発明に関連する第1のフォールトトレラントシステムにおけるロックステップ非同期状態時の動作の一例を示す図である。

10

【図16】本発明に関連する第1のフォールトトレラントシステムにおけるロックステップ非同期状態時の動作の他の例を示す図である。

【図17】本発明に関連する第2のフォールトトレラントシステムのブロック図である。

【発明を実施するための形態】

【0018】

次に本発明の実施の形態について図面を参照して詳細に説明する。

【0019】

[第1の実施形態]

図1を参照すると、本発明の第1の実施形態にかかるフォールトトレラントシステム10000は、互いに同一のハードウェアで構成された2つのサブシステム11000、12000を備えている。

20

【0020】

サブシステム11000は、プロセッサ部11100、このプロセッサ部11100に接続される入出力部11200、および、プロセッサ部11100と入出力部11200間に接続される制御部11300を有する。また、サブシステム12000は、プロセッサ部12100、このプロセッサ部12100に接続される入出力部12200、および、プロセッサ部12100と入出力部12200間に接続される制御部12300を有する。さらに、制御部11300、12300を介してサブシステム11000とサブシステム12000間を互いに接続する信号伝送路13000が設けられている。

30

【0021】

プロセッサ部11100とプロセッサ部12100は、サブシステム11000とサブシステム12000とがロックステップ動作している状態（ロックステップ同期の状態）で動作可能であると共に、サブシステム11000とサブシステム12000とがロックステップ動作していない状態（ロックステップ非同期の状態）で動作可能である。

【0022】

制御部11300、12300は、以下のような機能を有する。

【0023】

まず、制御部11300、12300は、プロセッサ部11100、12100がロックステップ同期の状態にあるときは、自サブシステムのプロセッサ部から入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行う。ここで、誤り検出用データの種別およびその生成方法は、任意である。例えば、プロセッサ部から入出力部へアクセスされるデータが本来のデータ以外にCRC (Cyclic Redundancy Code) を有する場合、このCRCを上記誤り検出用データとして使用して良い。また、プロセッサ部から入出力部へアクセスされるデータに基づいてチェックサムを算出し、この算出したチェックサムを上記誤り検出用データとして使用して良い。

40

【0024】

他方、プロセッサ部11100、12100がロックステップ非同期の状態にあるときは、制御部11300、12300は、自サブシステムのプロセッサ部から入出力部へア

50

アクセスされるデータの他サブシステムへの送信を行う。すなわち、プロセッサ部11100、12100がロックステップ非同期の状態にあるときは、主として、何れか一方のプロセッサ部が停止している。仮にプロセッサ部12100が停止しているとする、入出力部11200、12200へのアクセスはプロセッサ部11100が担うことになる。このため、制御部11300は、プロセッサ部11100から入出力部12200へアクセスされるデータそのものの他サブシステムへの送信を行う。

【0025】

次に本実施形態に係るフォールトトレラントシステム10000の動作を説明する。

【0026】

まず、プロセッサ部11100、12100がロックステップ同期の状態にあるときの動作を説明する。

【0027】

図2は、プロセッサ部11100とプロセッサ部12100とが入出力部11200内の入出力デバイス11210をアクセスするデータを生成したとき、入出力デバイス11210がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ同期の状態にあるので、制御部11300は、入出力部11200内の入出力デバイス11210宛のデータをプロセッサ部11100から受信すると、そのデータから誤り検出用データを作成し、元のデータと上記作成した誤り検出用データとを制御部11300内部のバッファに蓄積して一時的に保存する。他方、制御部12300は、入出力デバイス11210宛のデータをプロセッサ部12100から受信すると、そのデータから誤り検出用データを作成し、信号伝送路13000を介して制御部11300へ送信する。制御部11300は、バッファに蓄積されている誤り検出用データと信号伝送路13000を介して制御部12300から受信した誤り検出用データとを比較する。そして、双方の誤り検出用データが一致した場合に限り、バッファに保持している入出力デバイス11210宛のデータを入出力部11200の入出力デバイス11210へ送信する。これにより、双方の誤り検出用データが一致しないような不正な状況の下で入出力デバイス11210がアクセスされる事態が回避される。

【0028】

入出力デバイス11210は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部11300へ送信する。ロックステップ同期の状態にあるので、制御部11300は、入出力デバイス11210からのデータを受信すると、プロセッサ部11100へ転送すると共に、信号伝送路13000を介して制御部12300へ転送する。制御部12300は、この転送されてきたデータをプロセッサ部12100へさらに転送する。

【0029】

このように、プロセッサ部11100、12100がロックステップ同期の状態にある場合、プロセッサ部12100から入出力デバイス11210宛のデータは信号伝送路13000に送信されず、そのデータから生成された誤り検出用データが信号伝送路13000に送信される。その結果、信号伝送路13000上のデータ量を削減することができる。

【0030】

図3は、プロセッサ部11100とプロセッサ部12100とが入出力部12200内の入出力デバイス12210をアクセスするデータを生成したとき、入出力デバイス12210がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ同期の状態にあるので、制御部12300は、入出力部12200内の入出力デバイス12210宛のデータをプロセッサ部12100から受信すると、そのデータから誤り検出用データを作成し、元のデータと上記作成した誤り検出用データとを制御部12300内部のバッファに蓄積して一時的に保存する。他方、制御部11300は、入出力デバイス12210宛のデータをプロセッサ部11100から受信すると、そのデータから誤り検出用データを作成し、信号伝送路13000を介して制御部12

300へ送信する。制御部12300は、バッファに蓄積されている誤り検出用データと信号伝送路13000を介して制御部11300から受信した誤り検出用データとを比較する。そして、双方の誤り検出用データが一致した場合に限り、バッファに保持している入出力デバイス12210宛のデータを入出力部12200の入出力デバイス12210へ送信する。これにより、双方の誤り検出用データが一致しないような不正な状況の下で入出力デバイス12210がアクセスされる事態が回避される。

【0031】

入出力デバイス12210は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部12300へ送信する。ロックステップ同期の状態にあるので、制御部12300は、入出力デバイス12210からのデータを
10
受信すると、プロセッサ部12300へ転送すると共に、信号伝送路13000を介して制御部11300へ転送する。制御部11300は、この転送されてきたデータをプロセッサ部11100へさらに転送する。

【0032】

このように、プロセッサ部11100、12100がロックステップ同期の状態にある場合、プロセッサ部11100から入出力デバイス12210宛のデータは信号伝送路13000に送信されず、そのデータから生成された誤り検出用データが信号伝送路13000に送信される。その結果、信号伝送路13000上のデータ量を削減することができる。

【0033】

次に、プロセッサ部11100、12100がロックステップ非同期の状態にあるときの動作を説明する。例として、プロセッサ部11100が障害など何らかの理由で稼働を停止し、プロセッサ部12100が稼働している状態を想定する。このとき、入出力部11200および入出力部12200をアクセスするデータを生成するのは、プロセッサ部12100のみである。
20

【0034】

図4は、プロセッサ部12100が入出力部11200内の入出力デバイス11210をアクセスするデータを生成したときと、入出力デバイス11210がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ非同期の状態にあるので、制御部12300は、入出力部11200内の入出力デバイス11210宛のデータをプロセッサ部12100から受信すると、受信したデータを信号伝送路13000を介して制御部11300へ送信する。また制御部11300は、受信したデータを入出力部11200の入出力デバイス11210へ送信する。
30

【0035】

入出力デバイス11210は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部11300へ送信する。プロセッサ部11100が稼働していないロックステップ非同期の状態にあるので、制御部11300は、入出力デバイス11210からのデータを受信すると、信号伝送路13000を介して制御部12300へ転送する。制御部12300は、この転送されてきたデータをプロセッサ部12100へさらに転送する。
40

【0036】

このように、プロセッサ部11100、12100がロックステップ非同期の状態にある場合、プロセッサ部12100から入出力デバイス11210宛のデータそのものが信号伝送路13000に送信される。

【0037】

図5は、プロセッサ部12100が入出力部12200内の入出力デバイス12210をアクセスするデータを生成したときと、入出力デバイス12210がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。制御部12300は、入出力部12200内の入出力デバイス12210宛のデータをプロセッサ部12100から受信すると、そのデータを入出力部12200の入出力デバイス12210へ送信
50

する。プロセッサ部 1 1 1 0 0 が稼働していないロックステップ非同期の状態にあるので、制御部 1 2 3 0 0 は信号伝送路 1 3 0 0 0 への送信は行わない。

【0 0 3 8】

入出力デバイス 1 2 2 1 0 は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部 1 2 3 0 0 へ送信する。制御部 1 2 3 0 0 は、入出力デバイス 1 2 2 1 0 からのデータを受信すると、プロセッサ部 1 2 1 0 0 へ転送する。プロセッサ部 1 1 1 0 0 が稼働していないロックステップ非同期の状態にあるので、制御部 1 2 3 0 0 は信号伝送路 1 3 0 0 0 への送信は行わない。

【0 0 3 9】

以上説明したように本実施形態によれば、サブシステム間を接続する信号伝送路 1 3 0 0 0 を流れるデータ量を低減しつつ、フォールトトレラントシステムをロックステップ同期の状態およびロックステップ非同期の状態の双方の状態で作動させることができる。

【0 0 4 0】

その理由は、制御部 1 1 3 0 0、1 2 3 0 0 は、プロセッサ部 1 1 1 0 0、1 2 1 0 0 がロックステップ同期の状態にあるときは、自サブシステムのプロセッサ部から入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、他方、プロセッサ部 1 1 1 0 0、1 2 1 0 0 がロックステップ非同期の状態にあるときは、自サブシステムのプロセッサ部から入出力部へアクセスされるデータの他サブシステムへの送信を行うためである。

[第 2 の実施形態]

次に本発明の第 2 の実施形態について説明する。まず、本実施形態が解決しようとする課題について説明する。

【0 0 4 1】

特許文献 1 ～ 3 に詳細は記載されていないが、例えば特許文献 4 の図 1 等 に示されるように、フォールトトレラントシステムを構成する冗長化されたサブシステム間を接続するクロスリンクは、例えば本発明に関連する第 2 のフォールトトレラントシステムの構成図である図 1 7 に示すように、以下の 4 つの信号伝送路（リンク）で構成されている。

【0 0 4 2】

- (1) モジュール 1 の CPU サブシステム 1 0 からモジュール 2 の IO サブシステム 2 1 へアクセスするデータを転送するための信号伝送路 L 1
- (2) モジュール 1 の IO サブシステム 1 1 からモジュール 2 の CPU サブシステム 2 0 へアクセスするデータを転送するための信号伝送路 L 2
- (3) モジュール 2 の CPU サブシステム 2 0 からモジュール 1 の IO サブシステム 1 1 へアクセスするデータを転送するための信号伝送路 L 3
- (4) モジュール 2 の IO サブシステム 2 1 からモジュール 1 の CPU サブシステム 1 0 へアクセスするデータを転送するための信号伝送路 L 4

【0 0 4 3】

CPU サブシステム 1 0、2 0 がロックステップ動作している場合、信号伝送路 L 1、L 3 には、プロセッサ 1 0 0、2 0 0 から f t チップセット 1 1 0、2 1 0 へ送信されるパケット量と同じ量のパケットが流れ、信号伝送路 L 2、L 4 には、IO デバイス 1 3 0、2 3 0 から f t チップセット 1 1 0、2 1 0 を通じてプロセッサ 1 0 0、2 0 0 へ送信されるパケット量と同じ量のパケットが流れる。即ち、モジュール間リンクの帯域は、プロセッサと f t チップセット間のリンクの帯域と同等になるように設計されている。例えば、プロセッサと f t チップセットとの間が P C I - e x p r e s s 5 . 0 G T / s x 4 で接続されている場合、自系の CPU サブシステムと他系の IO サブシステムの間も、P C I - e x p r e s s 5 . 0 G T / s x 4 と同等のデータ転送帯域を持つリンクで接続されている。

【0 0 4 4】

近年、より高い IO 性能が要求されるのに従い、プロセッサと f t チップセット間のリンクに必要な帯域も増加している。その結果、モジュール間リンクにもより広い帯域

10

20

30

40

50

が必要となり、そのためには、モジュール間リンクの転送速度を高速化するか、リンク幅を広げる必要がある。しかし、モジュール間リンクの転送速度を高速化するとシグナルインテグリティの検証のためのコストが増加し、リンク幅を広げるとメモリチップセットのコストが大きくなる。そのため、コストの増加を抑えるために、モジュール間リンクに必要な帯域を広げることなく、より高いIO性能を実現する方法が必要となる。

【0045】

本発明に関連する第1の関連技術では、パケットそのものでなくパケットから生成したチェックサムをクロスリンクを介して他サブシステムへ送信する一方、上記生成したチェックサムとクロスリンクを介して他サブシステムから受信したチェックサムとを比較してサブシステム間の処理の不一致を検知している。従って、上記第1の関連技術を図17の本発明に関連する第2のフォールトトレラントシステムに適用すれば、CPUサブシステム10、20がロックステップ動作している際に信号伝送路L1、L3を流れるデータ量を、パケットそのものを送信する場合に比べて低減することが可能である。しかしながら、信号伝送路L1、L3は、CPUサブシステム10、20がロックステップ動作している場合に使用される以外に、ロックステップ動作していない場合にも使用される。例えば信号伝送路L1は、CPUサブシステム20が停止している状態で、CPUサブシステム10から他系のIOサブシステム230を制御するために、プロセッサ100で生成したパケットそのものを流すために使用される。また信号伝送路L3は、CPUサブシステム10が停止している状態で、CPUサブシステム20から他系のIOサブシステム130を制御するために、プロセッサ200で生成したパケットそのものを流すために使用される。従って、ロックステップ動作時の必要帯域を狭くしただけでは、信号伝送路L1、L3に必要な帯域を削減することはできない。

【0046】

本実施形態の目的は、上述した課題、すなわち、サブシステム間を相互に接続する信号伝送路に必要な帯域を削減することは困難である、という課題を解決するフォールトトレラントシステムを提供することにある。

【0047】

図6を参照すると、本発明の第2の実施形態にかかるフォールトトレラントシステム1000は、互いに同一のハードウェアで構成された2つのサブシステム1100、1200を備えている。

【0048】

サブシステム1100は、プロセッサ部1110、このプロセッサ部1110に接続される入出力部1120、および、プロセッサ部1110と入出力部1120間に接続される制御部1130を有する。また、サブシステム1200は、プロセッサ部1210、このプロセッサ部1210に接続される入出力部1220、および、プロセッサ部1210と入出力部1220間に接続される制御部1230を有する。さらに、制御部1130、1230を介してサブシステム1100とサブシステム1200間を互いに接続する2本の信号伝送路1300、1400が設けられている。

【0049】

プロセッサ部1110とプロセッサ部1210は、サブシステム1100とサブシステム1200とがロックステップ動作している状態（ロックステップ同期の状態）で動作可能であると共に、サブシステム1100とサブシステム1200とがロックステップ動作していない状態（ロックステップ非同期の状態）で動作可能である。

【0050】

制御部1130、1230は、以下のような機能を有する。

【0051】

まず、制御部1130、1230は、プロセッサ部1110、1210がロックステップ同期の状態にあるときは、制御部1130、1230を介してサブシステム1100のプロセッサ部1110からサブシステム1200の入出力部1220へアクセスされるデータから制御部1130によって生成された誤り検出用データの送信と、サブシステム1

100の入出力部1120からサブシステム1200のプロセッサ部1210へアクセスされるデータの送信とで、信号伝送路1300を共用する制御を行う。同じくプロセッサ部1110、1210がロックステップ同期の状態にあるときは、制御部1130、1230を介してサブシステム1200のプロセッサ部1210からサブシステム1100の入出力部1120へアクセスされるデータから制御部1230によって生成された誤り検出用データの送信と、サブシステム1200の入出力部1220からサブシステム1100のプロセッサ部1110へアクセスされるデータの送信とで信号伝送路1400を共用する制御を行う。

【0052】

ここで、誤り検出用データの種類およびその生成方法は、任意である。例えば、プロセッサ部から入出力部へアクセスされるデータが本来のデータ以外にCRC (Cyclic Redundancy Code) を有する場合、このCRCを上記誤り検出用データとして使用して良い。また、プロセッサ部から入出力部へアクセスされるデータに基づいてチェックサムを算出し、この算出したチェックサムを上記誤り検出用データとして使用して良い。

【0053】

他方、プロセッサ部1110、1210がロックステップ非同期の状態にあるときは、制御部1130、1230は、制御部1130、1230を介してサブシステム1100のプロセッサ部1110からサブシステム1200の入出力部1220へアクセスされるデータの送信と、サブシステム1100の入出力部1120からサブシステム1200のプロセッサ部1210へアクセスされるデータの送信とのうちの何れか一方で、信号伝送路1300を使用する制御を行う。同じくプロセッサ部1110、1210がロックステップ非同期の状態にあるときは、制御部1130、1230を介してサブシステム1200のプロセッサ部1210からサブシステム1100の入出力部1120へアクセスされるデータの送信と、サブシステム1200の入出力部1220からサブシステム1100のプロセッサ部1110へアクセスされるデータの送信とのうちの何れか一方で、信号伝送路1400を使用する制御を行う。

【0054】

すなわち、プロセッサ部1110、1210がロックステップ非同期の状態にあるときは、主として、何れか一方のプロセッサ部が停止している。仮にプロセッサ部1210が停止しているとする、入出力部1120、1220へのアクセスはプロセッサ部1110が担うことになる。このため、信号伝送路1300は、プロセッサ部1110から入出力部1220へアクセスするデータの送信に使用され、信号伝送路1400は、入出力部1220からプロセッサ部1110へアクセスするデータの送信に使用される。プロセッサ部1210は停止しているので、プロセッサ部1210から信号伝送路1400を通じて入出力部1120へアクセスするデータを送信することはない、また入出力部1120から信号伝送路1300を通じてプロセッサ部1210へアクセスするデータを送信することはない。

【0055】

次に本実施形態に係るフォールトトレラントシステム1000の動作を説明する。

【0056】

まず、プロセッサ部1110、1210がロックステップ同期の状態にあるときの動作を説明する。

【0057】

図7は、プロセッサ部1110とプロセッサ部1210とが入出力部1120内の入出力デバイス1121をアクセスするデータを生成したときと、入出力デバイス1121がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ同期の状態にあるので、制御部1130は、入出力部1120内の入出力デバイス1121宛のデータをプロセッサ部1110から受信すると、そのデータから誤り検出用データを作成し、元のデータと上記作成した誤り検出用データとを制御部1130

10

20

30

40

50

内部のバッファに蓄積して一時的に保存する。他方、制御部1230は、入出力デバイス1121宛のデータをプロセッサ部1210から受信すると、そのデータから誤り検出用データを作成し、信号伝送路1400を介して制御部1130へ送信する。制御部1130は、バッファに蓄積されている誤り検出用データと信号伝送路1400を介して制御部1230から受信した誤り検出用データとを比較する。そして、双方の誤り検出用データが一致した場合に限り、バッファに保持している入出力デバイス1121宛のデータを入出力部1120の入出力デバイス1121へ送信する。これにより、双方の誤り検出用データが一致しないような不正な状況の下で入出力デバイス1121がアクセスされる事態が回避される。

【0058】

10

入出力デバイス1121は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部1130へ送信する。ロックステップ同期の状態にあるので、制御部1130は、入出力デバイス1121からのデータを受信すると、プロセッサ部1110へ転送すると共に、信号伝送路1300を介して制御部1230へ転送する。制御部1230は、この転送されてきたデータをプロセッサ部1210へさらに転送する。

【0059】

このように、プロセッサ部1110、1210がロックステップ同期の状態にある場合、プロセッサ部1210から入出力デバイス1121宛のデータは信号伝送路1400に送信されず、そのデータから生成された誤り検出用データが信号伝送路1400に送信される。他方、信号伝送路1300には、入出力デバイス1121が生成したデータが送信される。

20

【0060】

図8は、プロセッサ部1110とプロセッサ部1210とが入出力部1220内の入出力デバイス1221をアクセスするデータを生成したときと、入出力デバイス1221がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ同期の状態にあるので、制御部1230は、入出力部1220内の入出力デバイス1221宛のデータをプロセッサ部1210から受信すると、そのデータから誤り検出用データを作成し、元のデータと上記作成した誤り検出用データとを制御部1230内部のバッファに蓄積して一時的に保存する。他方、制御部1130は、入出力デバイス1221宛のデータをプロセッサ部1110から受信すると、そのデータから誤り検出用データを作成し、信号伝送路1300を介して制御部1230へ送信する。制御部1230は、バッファに蓄積されている誤り検出用データと信号伝送路1300を介して制御部1130から受信した誤り検出用データとを比較する。そして、双方の誤り検出用データが一致した場合に限り、バッファに保持している入出力デバイス1221宛のデータを入出力部1220の入出力デバイス1221へ送信する。これにより、双方の誤り検出用データが一致しないような不正な状況の下で入出力デバイス1221がアクセスされる事態が回避される。

30

【0061】

入出力デバイス1221は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部1230へ送信する。ロックステップ同期の状態にあるので、制御部1230は、入出力デバイス1221からのデータを受信すると、プロセッサ部1210へ転送すると共に、信号伝送路1400を介して制御部1130へ転送する。制御部1130は、この転送されてきたデータをプロセッサ部1110へさらに転送する。

40

【0062】

このように、プロセッサ部1110、1210がロックステップ同期の状態にある場合、プロセッサ部1110から入出力デバイス1221宛のデータは信号伝送路1300に送信されず、そのデータから生成された誤り検出用データが信号伝送路1300に送信される。他方、信号伝送路1400には、入出力デバイス1221が生成したデータが送信

50

される。

【0063】

現用で使用する入出力デバイスが入出力部1120と入出力部1220とに分散している場合、図7に示したような動作と図8に示したような動作とが並行して行われる。その結果、信号伝送路1300、1400に流れる誤り検出用データのデータ量の最大値をX、入出力部で生成されるデータ量の最大値をYとすると、信号伝送路1300、1400に流れるデータ量の最大値はX+Yとなる。

【0064】

次に、プロセッサ部1110、1210がロックステップ非同期の状態にあるときの動作を説明する。例として、プロセッサ部1110が障害など何らかの理由で稼働を停止し、プロセッサ部1210が稼働している状態を想定する。このとき、入出力部1120および入出力部1220をアクセスするデータを生成するのは、プロセッサ部1210のみである。

10

【0065】

図9は、プロセッサ部1210が入出力部1120内の入出力デバイス1121をアクセスするデータを生成したときと、入出力デバイス1121がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ非同期の状態にあるので、制御部1230は、入出力部1120内の入出力デバイス1121宛のデータをプロセッサ部1210から受信すると、受信したデータを信号伝送路1400を介して制御部1130へ送信する。また制御部1130は、受信したデータを入出力部1120の入出力デバイス1121へ送信する。

20

【0066】

入出力デバイス1121は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部1130へ送信する。プロセッサ部1110が稼働していないロックステップ非同期の状態にあるので、制御部1130は、入出力デバイス1121からのデータを受信すると、信号伝送路1300を介して制御部1230へ転送する。制御部1230は、この転送されてきたデータをプロセッサ部1210へさらに転送する。

【0067】

このように、プロセッサ部1110、1210がロックステップ非同期の状態にある場合、プロセッサ部1210から入出力デバイス1121宛のデータそのものが信号伝送路1400に送信される。他方、信号伝送路1300には、入出力デバイス1121が生成したデータが送信される。

30

【0068】

図10は、プロセッサ部1210が入出力部1220内の入出力デバイス1221をアクセスするデータを生成したときと、入出力デバイス1221がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。制御部1230は、入出力部1220内の入出力デバイス1221宛のデータをプロセッサ部1210から受信すると、そのデータを入出力部1220の入出力デバイス1221へ送信する。プロセッサ部1110が稼働していないロックステップ非同期の状態にあるので、制御部1230は信号伝送路1400への送信は行わない。

40

【0069】

入出力デバイス1221は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部1230へ送信する。制御部1230は、入出力デバイス1221からのデータを受信すると、プロセッサ部1210へ転送する。プロセッサ部1110が稼働していないロックステップ非同期の状態にあるので、制御部1230は信号伝送路1400への送信は行わない。

【0070】

このように、プロセッサ部1110、1210がロックステップ非同期の状態にある場合、プロセッサ部1210から入出力デバイス1221をアクセスする際には信号伝送路

50

1400は使用されない。また入出力デバイス1221がデータを生成した際も、信号伝送路1400は使用されない。

【0071】

現用で使用する入出力デバイスが入出力部1120と入出力部1220とに分散している場合、図9に示したような動作と図10に示したような動作とが並行して行われる。プロセッサ部から入出力部をアクセスするデータのデータ量の最大値は、入出力部で生成されるデータ量の最大値とほぼ同じである。従って、信号伝送路1300、1400に流れるデータ量の最大値はYとなる。

【0072】

次に、比較のために、図13～図16に示すように、サブシステム1100、1200間を以下のような4つの信号伝送路で接続した本発明に関連する第1のフォールトトレラントシステムについて説明する。

【0073】

(1) サブシステム1100のプロセッサ部1110からサブシステム1200の入出力部1220へアクセスするデータを転送するための信号伝送路1310

(2) サブシステム1100の入出力部1120からサブシステム1200のプロセッサ部1210へアクセスするデータを転送するための信号伝送路1320

(3) サブシステム1200のプロセッサ部1210からサブシステム1100の入出力部1120へアクセスするデータを転送するための信号伝送路1410

(4) サブシステム1200の入出力部1220からサブシステム1100のプロセッサ部1110へアクセスするデータを転送するための信号伝送路1420

【0074】

まず、プロセッサ部1110、1210がロックステップ同期の状態にあるときの動作を説明する。

【0075】

図13は、プロセッサ部1110とプロセッサ部1210とが入出力部1120内の入出力デバイス1121をアクセスするデータを生成したときと、入出力デバイス1121がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ同期の状態にあるので、制御部1130は、入出力部1120内の入出力デバイス1121宛のデータをプロセッサ部1110から受信すると、そのデータからチェックサムを作成し、元のデータと上記作成したチェックサムとを制御部1130内部のバッファに蓄積して一時的に保存する。他方、制御部1230は、入出力デバイス1121宛のデータをプロセッサ部1210から受信すると、そのデータからチェックサムを作成し、信号伝送路1410を介して制御部1130へ送信する。制御部1130は、バッファに蓄積されているチェックサムと信号伝送路1410を介して制御部1230から受信したチェックサムとを比較する。そして、双方のチェックサムが一致した場合に限り、バッファに保持している入出力デバイス1121宛のデータを入出力部1120の入出力デバイス1121へ送信する。

【0076】

入出力デバイス1121は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部1130へ送信する。ロックステップ同期の状態にあるので、制御部1130は、入出力デバイス1121からのデータを受信すると、プロセッサ部1110へ転送すると共に、信号伝送路1320を介して制御部1230へ転送する。制御部1230は、この転送されてきたデータをプロセッサ部1210へさらに転送する。

【0077】

図14は、プロセッサ部1110とプロセッサ部1210とが入出力部1220内の入出力デバイス1221をアクセスするデータを生成したときと、入出力デバイス1221がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ同期の状態にあるので、制御部1230は、入出力部1220内の入出力

デバイス１２２１宛のデータをプロセッサ部１２１０から受信すると、そのデータからチェックサムを作成し、元のデータと上記作成したチェックサムとを制御部１２３０内部のバッファに蓄積して一時的に保存する。他方、制御部１１３０は、入出力デバイス１２２１宛のデータをプロセッサ部１１１０から受信すると、そのデータからチェックサムを作成し、信号伝送路１３１０を介して制御部１２３０へ送信する。制御部１２３０は、バッファに蓄積されているチェックサムと信号伝送路１３１０を介して制御部１１３０から受信したチェックサムとを比較する。そして、双方のチェックサムが一致した場合に限り、バッファに保持している入出力デバイス１２２１宛のデータを入出力部１２２０の入出力デバイス１２２１へ送信する。

【００７８】

10

入出力デバイス１２２１は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部１２３０へ送信する。ロックステップ同期の状態にあるので、制御部１２３０は、入出力デバイス１２２１からのデータを受信すると、プロセッサ部１２１０へ転送すると共に、信号伝送路１４２０を介して制御部１１３０へ転送する。制御部１１３０は、この転送されてきたデータをプロセッサ部１１１０へさらに転送する。

【００７９】

現用で使用する入出力デバイスが入出力部１１２０と入出力部１２２０とに分散している場合、図１３に示したような動作と図１４に示したような動作とが並行して行われる。その結果、信号伝送路１３１０、１３２０、１４１０、１４２０に流れるデータ量の最大値はＹとなる。

20

【００８０】

次に、プロセッサ部１１１０、１２１０がロックステップ非同期の状態にあるときの動作を説明する。例として、プロセッサ部１１１０が障害など何らかの理由で稼働を停止し、プロセッサ部１２１０が稼働している状態を想定する。このとき、入出力部１１２０および入出力部１２２０をアクセスするデータを生成するのは、プロセッサ部１２１０のみである。

【００８１】

図１５は、プロセッサ部１２１０が入出力部１１２０内の入出力デバイス１１２１をアクセスするデータを生成したときと、入出力デバイス１１２１がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。ロックステップ非同期の状態にあるので、制御部１２３０は、入出力部１１２０内の入出力デバイス１１２１宛のデータをプロセッサ部１２１０から受信すると、受信したデータを信号伝送路１４１０を介して制御部１１３０へ送信する。また制御部１１３０は、受信したデータを入出力部１１２０の入出力デバイス１１２１へ送信する。

30

【００８２】

入出力デバイス１１２１は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部１１３０へ送信する。プロセッサ部１１１０が稼働していないロックステップ非同期の状態にあるので、制御部１１３０は、入出力デバイス１１２１からのデータを受信すると、信号伝送路１３２０を介して制御部１２３０へ転送する。制御部１２３０は、この転送されてきたデータをプロセッサ部１２１０へさらに転送する。

40

【００８３】

図１６は、プロセッサ部１２１０が入出力部１２２０内の入出力デバイス１２２１をアクセスするデータを生成したときと、入出力デバイス１２２１がそのアクセスデータに対する応答データを生成したときのデータの流れを示している。制御部１２３０は、入出力部１２２０内の入出力デバイス１２２１宛のデータをプロセッサ部１２１０から受信すると、そのデータを入出力部１２２０の入出力デバイス１２２１へ送信する。プロセッサ部１１１０が稼働していないロックステップ非同期の状態にあるので、制御部１２３０は信号伝送路１４１０への送信は行わない。

50

【0084】

入出力デバイス1221は、上記データを受信し、上記データに応じた処理を行い、その処理結果を応答データとして生成し、制御部1230へ送信する。制御部1230は、入出力デバイス1221からのデータを受信すると、プロセッサ部1210へ転送する。プロセッサ部1110が稼働していないロックステップ非同期の状態にあるので、制御部1230は信号伝送路1420への送信は行わない。

【0085】

このように、プロセッサ部1110、1210がロックステップ非同期の状態にある場合、プロセッサ部1210から入出力デバイス1221をアクセスする際には信号伝送路1410は使用されない。また入出力デバイス1221がデータを生成した際も、信号伝送路1420は使用されない。

10

【0086】

現用で使用する入出力デバイスが入出力部1120と入出力部1220とに分散している場合、図15に示したような動作と図16に示したような動作とが並行して行われる。従って、信号伝送路1320、1410に流れるデータ量の最大値はYとなる。同様に、プロセッサ部1110でなくプロセッサ部1210が障害等により停止した場合には、信号伝送路1310、1420に流れるデータ量の最大値はYとなる。

【0087】

以上の結果、サブシステム1100、1200間を図13～図16に示すように4本の信号伝送路で接続したフォールトトレラントシステムにおいては、各信号伝送路に必要な帯域はYとなる。従って、サブシステム全体では、サブシステム1100からサブシステム1200へデータを送信するために必要となる帯域、および、サブシステム1200からサブシステム1100へデータを送信するために必要となる帯域は、それぞれ2Yとなる。

20

【0088】

これに対して、サブシステム1100、1200間を図7～図10に示すように2本の信号伝送路1300、1400で接続し、一方のサブシステムのプロセッサ部から他方のサブシステムの入出力部へアクセスするデータと一方のサブシステムの入出力部から他方のサブシステムのプロセッサ部へアクセスするデータとで同じ信号伝送路を共有する本実施形態に係るフォールトトレラントシステムでは、サブシステム1100からサブシステム1200へデータを送信するために必要となる帯域、および、サブシステム1200からサブシステム1100へデータを送信するために必要となる帯域は、それぞれX+Yとなる。ここで、XはYより十分に小さい。従って、本実施形態によれば、サブシステム間を相互に接続する信号伝送路に必要な帯域を大幅に削減することができる。

30

【0089】

なお、プロセッサ部1110、1210がロックステップ非同期の状態にあるときに、プロセッサ部1110、1210の双方を独立して稼働させ、プロセッサ部1110に他系の入出力部1220との間でデータの授受を行わせ、且つ、プロセッサ部1210に他系の入出力部1120との間でデータの授受を行わせることは、理論的には可能である。そして、その際には信号伝送路1300、1400に必要な帯域は2Yとなる。従って、信号伝送路1300、1400の帯域をX+Yに設定した場合、上述したような動作は信号伝送路の帯域不足によって実現困難になるが、実用面では全く問題とはならない。何故ならば、上述したような動作形態は実際の運用では必要性がないためである。

40

【0090】

[第3の実施形態]

次に本発明の第3の実施形態について詳細に説明する。

[本実施形態が解決しようとする課題]

まず、図17を参照して本実施形態が解決しようとする課題について説明する。図17は、2つのモジュールをロックステップ動作させる本発明に関連する第2のフォールトトレラントシステムの構成を示している。このフォールトトレラントシステムでは、自モジ

50

ジュールのCPUサブシステムと他モジュールのIOサブシステムとの間、また自モジュールのIOサブシステムと他モジュールのCPUサブシステムの間が、モジュール間リンクにより双方向接続されている。このとき、モジュール間リンクの帯域は、プロセッサとf tチップセット間のリンクの帯域と同等になるように設計されている。例えば、プロセッサとf tチップセットとの間がP C I - e x p r e s s 5 . 0 G T / s x 4で接続されている場合、自系のCPUサブシステムと他系のIOサブシステムの間も、P C I - e x p r e s s 5 . 0 G T / s x 4と同等のデータ転送帯域を持つリンクで接続されている。

【0091】

近年、より高いIO性能が要求されるのに従い、プロセッサとf tチップセット間のリンクに必要となる帯域も増加している。その結果、モジュール間リンクにもより高い帯域が必要となるが、帯域を高めるためには、転送速度を高速化させるか、リンク幅を増やす必要がある。しかし、転送速度を高速化させるとシグナルインテグリティの検証の為にコストが増加し、リンク幅を増やすとf tチップセットのコストが大きくなる。その為、コストの増加を抑えるために、モジュール間リンクに必要となる帯域を増やすことなく、より高いIO性能を実現する方法が必要となる。

【0092】

[本実施形態の概要]

本実施形態では、以下の2つの手段により上記課題を解決する。

【0093】

1つ目の手段として、モジュール間リンクを共用化する。本発明に関連する第2のフォールトトレラントシステムでは、モジュール間リンクはモジュール1のCPUサブシステムからモジュール2のIOサブシステムへのリンク、モジュール1のIOサブシステムからモジュール2のCPUサブシステムへのリンク、モジュール2のCPUサブシステムからモジュール1のIOサブシステムへのリンク、モジュール2のIOサブシステムからモジュール1のCPUサブシステムへのリンクの、4つのリンクで構成されていた。一方、本実施形態による構成では、モジュール間リンクをモジュール1からモジュール2へのリンク、モジュール2からモジュール1へのリンクの2つのリンクに再構成し、モジュール1のCPUサブシステムからモジュール2のIOサブシステムへのデータ転送と、モジュール1のIOサブシステムからモジュール2のCPUサブシステムへのデータ転送でモジュール間リンクを共用する。

【0094】

2つ目の手段として、CPUサブシステムがロックステップ動作している場合に、反対側のモジュールのIOサブシステムへのデータ転送量を削減する。本発明に関連する第2のフォールトトレラントシステムでは、CPUサブシステムはプロセッサから受信したパケットの宛先を確認し、モジュール1のIOサブシステム、またはモジュール2のIOサブシステムにパケット全体を転送していた。IOサブシステムはモジュール1のCPUサブシステムから受信したパケットと、モジュール2のCPUサブシステムから受信したパケットが一致していることを確認し、IOデバイスにそのパケットを転送していた。一方、本実施形態による構成では、CPUサブシステムはプロセッサから受信したパケットの宛先を確認し、宛先が反対側のモジュールのIOデバイスであった場合、かつCPUサブシステムがロックステップ動作している場合には、パケット全体を送信せずに、パケットの一部（または後述する実施形態のようにパケット情報やパケットを受信した時刻情報から生成したチェックサム等であってもよい）を送信する。一例としてパケットに含まれるCRCのみを送信する場合、IOサブシステムは、モジュール1のCPUサブシステムから受信したCRCと、モジュール2のCPUサブシステムから受信したCRCとを比較し、一致した場合には、自系のCPUサブシステムから受信したパケットをIOデバイスに転送する。一方、パケットの宛先が自系のIOサブシステムであった場合、またはCPUサブシステムがロックステップ動作していない場合には、パケット全体をIOサブシステムに転送する。

【0095】

[本実施形態の構成]

図11を参照すると、本発明の第3の実施形態として、モジュール1とモジュール2から構成されるフォールトトレラントシステムが示されている。モジュール1とモジュール2とは、モジュール間リンク30によって相互に接続されている。モジュール間リンク30は、モジュール1からモジュール2への送信に使用する信号伝送路31と、その逆にモジュール2からモジュール1への送信に使用する信号伝送路32との2つの信号伝送路で構成されている。

【0096】

モジュール1は、プロセッサ100、f tチップセット110、I Oデバイス130で構成される。プロセッサ100は、図示しないC P Uやメモリ等に加えて、パケット送信部101とパケット受信部102を有する。I Oデバイス130は、図示しないI Oデバイス等に加えて、パケット送信部131とパケット受信部132を有する。f tチップセット110は、パケット受信部111、パケット送信部112、セクタ113、C R C抽出器115、送信部116、受信部117、バッファ118、バッファ119、バッファ120、バッファ121、比較器122、セクタ123、パケット送信部124、パケット受信部125から構成される。また、f tチップセット110は、モジュール間リンク30に接続されている。

10

【0097】

モジュール2は、モジュール1と同じ構成を有する。すなわち、モジュール2は、プロセッサ200、f tチップセット210、I Oデバイス230で構成される。プロセッサ200は、図示しないC P Uやメモリ等に加えて、パケット送信部201とパケット受信部202を有する。I Oデバイス230は、図示しないI Oデバイス等に加えて、パケット送信部231とパケット受信部232を有する。f tチップセット210は、パケット受信部211、パケット送信部212、セクタ213、C R C抽出器215、送信部216、受信部217、バッファ218、バッファ219、バッファ220、バッファ221、比較器222、セクタ223、パケット送信部224、パケット受信部225から構成される。また、f tチップセット210は、モジュール間リンク30に接続されている。

20

【0098】

モジュール1とモジュール2は同じ構成なので、以下では、主にモジュール1について説明する。

30

【0099】

プロセッサ100は、I Oデバイス130、I Oデバイス230に対するパケットを生成する。パケットは、例えば、宛先の情報、送信元の情報、データ、C R C、およびその他の情報を有する。プロセッサ100が生成したパケットは、パケット送信部101により、パケットの中継を行うf tチップセット110に送信される。また、パケット受信部102は、f tチップセット110からパケットを受信する。

【0100】

I Oデバイス130は、プロセッサ100、プロセッサ200に対するパケットを生成する。生成されたパケットは、パケット受信部132により、パケットの中継を行うf tチップセット110に送信される。また、パケット受信部131は、f tチップセット110からパケットを受信する。

40

【0101】

f tチップセット110は、プロセッサ100、プロセッサ200、I Oデバイス130、I Oデバイス230から発行されたパケットの中継を行う。

【0102】

パケット受信部111は、プロセッサ100から受信したパケットを出力する。またパケット受信部111は、プロセッサ100から受信したパケットの宛先を確認し、I Oデバイス130宛てのパケットであった場合にはバッファ121に登録する。

【0103】

50

パケット送信部 112 は、受信部 117、またはパケット受信部 125 から受信したパケットを、プロセッサ 100 に送信する。

【0104】

セクタ 113 は、CPU サブシステムがロックステップ動作している場合には、CRC 抽出器 115 の出力を選択し、CPU サブシステムがロックステップ動作していない場合には、パケット受信部 111 からの出力を選択する。

【0105】

CRC 抽出器 115 は、パケット受信部 111 から出力されるパケットから CRC を抽出し、セクタ 113 に出力する。また、CRC をバッファ 118 に登録する。

【0106】

送信部 116 は、セクタ 113 から出力される CRC またはパケットを、モジュール間リンク 30 の信号伝送路 31 を介して、モジュール 2 の IO サブシステム 21 に送信する。また、送信部 116 は、パケット受信部 125 から受信したパケットも、信号伝送路 31 を介して、モジュール 2 の IO サブシステム 21 に送信する。

【0107】

受信部 117 は、モジュール間リンク 30 の信号伝送路 32 を介してモジュール 2 の CPU サブシステム 20 から受信した CRC をバッファ 119 に登録する。また、受信部 117 は、信号伝送路 32 を介してモジュール 2 の CPU サブシステム 20 から受信したパケットをバッファ 120 に登録する。

【0108】

比較器 122 は、バッファ 118 とバッファ 119 から出力される CRC を比較し、一致した場合には、バッファ 121 にパケットを出力するように指示する。

【0109】

セクタ 123 は、CPU サブシステムがロックステップ動作している場合、また CPU サブシステム 10 のみが稼働している場合には、バッファ 121 を選択する。CPU サブシステム 20 のみが稼働している場合には、バッファ 120 を選択する。

【0110】

[本実施形態の動作]

CPU サブシステム 10 と CPU サブシステム 20 がロックステップ動作している場合と、ロックステップ動作していない場合とで動作が異なる。まず、CPU サブシステム 10 と CPU サブシステム 20 がロックステップ動作している場合の動作を説明する。

【0111】

CPU サブシステムがロックステップ動作している場合には、プロセッサ 100 とプロセッサ 200 は、同じ時刻に同じパケットを生成する。ここで、生成されたパケットの宛先が IO デバイス 130 であるとする。生成されたパケットは、パケット送信部 101、パケット送信部 201 を経由してパケット受信部 111 とパケット受信部 211 に受信される。パケット受信部 111 は、受信したパケットを出力し、バッファ 121 に登録する。また、CRC 抽出器 115 は、受信したパケットの CRC を取り出し、バッファ 118 に登録する。

【0112】

一方、パケット受信部 211 は、パケットの宛先が IO デバイス 130 宛てである場合には、バッファ 221 にパケットを登録しない。セクタ 213 にはパケット受信部 211 が受信したパケットと、CRC 抽出器 215 がパケットから取り出した CRC が入力されているが、CPU がロックステップ動作しているため、CRC を送信部 216 に送付する。送信部 216 は、CRC を信号伝送路 32 を介して受信部 117 に送信する。受信部 117 は、受信した CRC をバッファ 119 に登録する。比較器 122 は、バッファ 118 に登録された CRC と、バッファ 119 に登録された CRC とを比較し、一致している場合にはバッファ 121 に指示して、バッファ 121 に登録されているパケットを出力させる。セクタ 123 は、バッファ 121 から受け取ったパケットをパケット送信部 124 に出力する。パケット送信部 124 は、受け取ったパケットを IO デバイス 130 に送

10

20

30

40

50

信する。

【0113】

次に、I Oデバイス130がパケットを生成した場合の動作を説明する。C P Uサブシステムがロックステップ動作している場合には、このパケットはプロセッサ100とプロセッサ200の両方に送信する必要がある。パケット送信部132は、I Oデバイス130が生成したパケットをパケット受信部125に送信する。パケット受信部125は、受信したパケットをパケット送信部112と、送信部116に出力する。パケット送信部112は、受け取ったパケットをパケット送信部102に送信する。送信部116は、受け取ったパケットを信号伝送路31を介して受信部217に送信する。受信部217は、受信したパケットの宛先を確認し、パケット送信部212に出力する。パケット送信部212は、受け取ったパケットをパケット送信部202に送信する。

10

【0114】

このように、C P Uサブシステム10とC P Uサブシステム20がロックステップ動作している場合には、プロセッサからI Oデバイス宛てのパケットはC R Cのみがモジュール間リンク30上で送信され、I Oデバイスからプロセッサ宛てのパケットはパケット全体がモジュール間リンク30上で送信される。

【0115】

次に、C P Uサブシステム10とC P Uサブシステム20がロックステップ動作していない場合の動作を説明する。ここでは、C P Uサブシステム10が稼働しており、C P Uサブシステム20は稼働していない場合の説明を行う。

20

【0116】

C P Uサブシステム10のみが稼働しているため、I Oデバイス130宛てのパケット、I Oデバイス230宛てのパケットを発行するのはC P Uサブシステム10のみである。まず、C P Uサブシステム10がI Oデバイス130宛てのパケットを発行した場合の説明を行う。プロセッサが発行したI Oデバイス130宛てのパケットは、パケット送信部101、パケット受信部111を経由してバッファ121に入力される。ここで、C P Uサブシステム10とC P Uサブシステム20はロックステップ動作していない為、バッファ121に入力されたパケットは、セレクト123、パケット送信部124を経由して、パケット受信部131により受信される。

【0117】

30

次に、C P Uサブシステム10がI Oデバイス230宛てのパケットを発行した場合の説明を行う。同様に、プロセッサが発行したパケットは、パケット送信部101、パケット受信部111を経由するが、宛先がI Oデバイス230であるため、バッファ121には登録されない。その代わりに、I Oデバイス230宛てのパケットは、セレクト113、送信部116を経由して、信号伝送路31経由で受信部217により受信される。受信部217は、受信したパケットの宛先を確認し、I Oデバイス230宛なので、バッファ220に登録する。登録されたパケットは、セレクト223、パケット送信部224を経由して、パケット受信部231により受信される。このように、C P Uサブシステム10とC P Uサブシステム20がロックステップ動作していない場合、宛先が他系のI Oデバイスであるパケットは、その全体がモジュール間リンク30上で送信される。

40

【0118】

次に、I Oデバイス130がパケットを生成した場合の動作を説明する。C P Uサブシステム10のみが稼働しているため、このパケットはプロセッサ100にのみ送信する必要がある。I Oデバイス130が生成したパケットは、パケット送信部132、パケット受信部125、パケット送信部112を経由して、パケット送信部102に受信される。

【0119】

次に、I Oデバイス230がパケットを生成した場合の動作を説明する。C P Uサブシステム10のみが稼働しているため、このパケットもプロセッサ100にのみ送信する必要がある。I Oデバイス230が生成したパケットは、パケット送信部232、パケット受信部225、送信部216、信号伝送路32、受信部117、パケット送信部112を

50

経由して、パケット送信部 102 に受信される。このように、CPU サブシステム 10 と CPU サブシステム 20 がロックステップ動作していない場合、IO デバイスが生成したパケットの宛先が他系のプロセッサである場合、パケット全体がモジュール間リンク 30 上で送信される。

【0120】

[本実施形態の効果]

本実施形態の効果として、モジュール間リンクに必要となる帯域を大幅に減少することが可能になる。その理由は以下の通りである。

【0121】

本発明に関連する第 2 のフォールトトレラントシステムでは、CPU サブシステムがロックステップ動作しているかどうかにかかわらず、モジュール間リンク上でパケット全体を送信していた。その為、自系の CPU サブシステムから他系の IO サブシステムへのモジュール間リンクと、自系の IO サブシステムから他系の CPU サブシステムへのモジュール間リンクを独立して用意する必要があった。一方で、本実施形態によるシステムでは、CPU サブシステムがロックステップ動作している場合には、自系の CPU サブシステムから他系の IO サブシステムへは CRC のみを送付すれば良い。パケット全体のデータ量に対して、CRC のみのデータ量は数十分の一に抑えることが出来るため、自系の CPU サブシステムから他系の IO サブシステムへのデータ転送量が大幅に削減できる。一方、CPU サブシステムがロックステップ動作していない場合には、稼働している CPU サブシステムから他系の IO サブシステムへはパケット全体を送信する必要があるが、稼働していない CPU サブシステムから他系の IO サブシステムへのデータ転送が不要である。その為、モジュール 1 からモジュール 2、またはモジュール 2 からモジュール 1 全体で見た場合、モジュール間リンク上で転送されるデータ量は、CPU サブシステムがロックステップ動作している場合よりも、CRC の送信にかかる分だけ少ない。

【0122】

以上より、本発明に関連する第 2 のフォールトトレラントシステムではパケット全体の送信に必要となる帯域を P とした場合、モジュール 1 からモジュール 2 へ 2P の帯域が、モジュール 2 からモジュール 1 へ 2P の帯域が必要となっていた。一方で、CRC のみの送信に必要となる帯域を C とした場合、モジュール 1 からモジュール 2 へ P+C の帯域が、モジュール 2 からモジュール 1 へ P+C の帯域のみがあれば良い。これにより、モジュール 1 からモジュール 2 への送信に必要となる帯域、モジュール 2 からモジュール 1 への送信に必要となる帯域が P-C 削減される。前述の通り C は P よりも十分小さいため、大幅にデータ帯域を減らすことが可能になる。

【0123】

[第 4 の実施形態]

【0124】

図 12 は本発明の第 4 の実施形態を示している。本発明の第 4 の実施形態では、本発明の第 3 の実施形態における CRC 抽出器 115、CRC 抽出器 215 が、チェックサム生成器 126、チェックサム生成器 226 で置き変わっている。チェックサム生成器 126 とチェックサム生成器 226 は同等であるため、チェックサム生成器 126 を例に動作の説明を行う。

【0125】

チェックサム生成器 126 は、パケット受信部 111 から受け取ったパケットと、パケットを受け取った時刻情報とから、チェックサム情報を生成する。そして CPU サブシステムがロックステップ動作している場合には、チェックサム情報を他系の IO サブシステムに送信する。CRC を送信する代わりにチェックサムを送信する利点は、チェックサムには時刻情報が付加されている点である。

【0126】

例えば、プロセッサ 100 の内部で訂正可能なエラーが発生した場合、プロセッサ 100 から f t チップセット 110 にパケットが送信される時刻と、プロセッサ 200 から f

tチップセット210にパケットが送信される時刻に、若干のズレが発生する場合が多い。しかし、例えばPCI-Expressの仕様では、ftチップセット110とftチップセット210が、各々プロセッサ100とプロセッサ200から同じパケットを異なる時刻に受信した場合でも、パケットに含まれるCRCは同じになる。そのため、第3の実施形態では比較器122は差分を検出できない。一方、チェックサム情報には時刻情報が含まれるため、ftチップセット110とftチップセット210が、各々プロセッサ100とプロセッサ200から同じパケットを異なる時刻に受信した場合、チェックサム情報が異なる為、比較器122が差分を検出することができる。これにより、プロセッサ100の異常を早期に検出することができる。

【0127】

10

また、プロセッサとftチップセット間のリンクで転送されるパケットが、CRCを含まないリンク仕様を採用する場合、CRC抽出器を使用する第3の実施形態は動作しない。

【0128】

以上、本発明を幾つかの実施形態を挙げて説明したが、本発明は以上の実施形態にのみ限定されず、その他各種の付加変更が可能である。

【産業上の利用可能性】

【0129】

本発明は、フォールトトレラントシステム、特にハードウェアによって実現されるフォールトトレラントシステムに適用することができる。

20

【0130】

上記の実施形態の一部又は全部は、以下の付記のようにも記載され得るが、以下には限られない。

(付記1)

互いに同一のハードウェアで構成された複数のサブシステムを備えたフォールトトレラントシステムであって、

前記複数のサブシステムは、

前記複数のサブシステム間においてロックステップ同期の状態および非同期の状態で作可能なプロセッサ部と、

前記プロセッサ部に接続される入出力部と、

30

前記プロセッサ部および前記入出力部間に接続される制御部と、

前記制御部を介して前記複数のサブシステム間を互いに接続する信号伝送路とを有し、

前記制御部は、前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信を行う

フォールトトレラントシステム。

(付記2)

前記信号伝送路は複数の信号伝送路で構成され、

40

前記制御部は、前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、ロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信、自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のうちの何れか一方の送信のために使用する制御を行う

付記1に記載のフォールトトレラントシステム。

50

(付記 3)

前記制御部は、

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データの受信時刻と前記データとからチェックサムを生成し、該生成した前記チェックサムを前記誤り検出用データとして出力する誤り検出用データ生成器を有する

付記 2 に記載のフォールトトレラントシステム。

(付記 4)

前記制御部は、

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データが有する CRC を抽出し、該抽出した前記 CRC を前記誤り検出用データとして出力する誤り検出用データ生成器を有する

付記 2 に記載のフォールトトレラントシステム。

(付記 5)

前記制御部は、

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データと前記誤り検出用データ生成器から出力された前記誤り検出用データとのうちの何れか一方を選択する第 1 のセレクタを有する

付記 3 または 4 に記載のフォールトトレラントシステム。

(付記 6)

前記制御部は、

自制御部の前記第 1 のセレクタの出力と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータとを同じ一つの前記信号伝送路へ送信する第 1 の送信部と、

自制御部の前記誤り検出用データ生成器から出力された前記誤り検出用データを一時的に記憶する第 1 のバッファと、

前記信号伝送路から受信した前記誤り検出用データを一時的に記憶する第 2 のバッファと、

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データを一時的に記憶する第 3 のバッファと、

前記信号伝送路から受信した他の前記サブシステムの前記プロセッサ部から自サブシステムの前記入出力部へアクセスされる前記データを一時的に記憶する第 4 のバッファと、

前記第 1 のバッファに記憶された前記誤り検出用データと前記第 2 のバッファに記憶された前記誤り検出用データとを比較する比較器と、

前記プロセッサ部が前記ロックステップ同期の状態にあるときは、前記比較器で一致が検出されることを条件に前記第 3 のバッファに記憶された前記データを自サブシステムの前記入出力部へ出力し、前記プロセッサ部が前記ロックステップ非同期の状態にあるときは、前記第 3 のバッファに記憶された前記データまたは前記第 4 のバッファに記憶された前記データを自サブシステムの前記入出力部へ出力する第 2 のセレクタと

を有する

付記 5 に記載のフォールトトレラントシステム。

(付記 7)

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いる制御装置であって、

自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と前記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する信号伝送路に接続され、

前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送

10

20

30

40

50

信を行う

フォールトトレラントシステムで用いる制御装置。

(付記 8)

前記信号伝送路は複数の信号伝送路で構成され、

前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、ロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信、自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のうちの何れか一方の送信のために使用する制御を行う

10

付記 7 に記載のフォールトトレラントシステムで用いる制御装置。

(付記 9)

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データの受信時刻と前記データとからチェックサムを生成し、該生成した前記チェックサムを前記誤り検出用データとして出力する誤り検出用データ生成器を有する

付記 8 に記載のフォールトトレラントシステムで用いる制御装置。

(付記 10)

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データが有する CRC を抽出し、該抽出した前記 CRC を前記誤り検出用データとして出力する誤り検出用データ生成器を有する

20

付記 8 に記載のフォールトトレラントシステムで用いる制御装置。

(付記 11)

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データと前記誤り検出用データ生成器から出力された前記誤り検出用データとのうちの何れか一方を選択する第 1 のセレクタを有する

付記 9 または 10 に記載のフォールトトレラントシステムで用いる制御装置。

(付記 12)

前記第 1 のセレクタの出力と自サブシステムの前記入出力部から他サブシステムの前記プロセッサ部へアクセスされるデータとを同じ一つの前記信号伝送路へ送信する第 1 の送信部と、

30

前記誤り検出用データ生成器から出力された前記誤り検出用データを一時的に記憶する第 1 のバッファと、

前記第 2 の信号伝送路から受信した前記誤り検出用データを一時的に記憶する第 2 のバッファと、

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データを一時的に記憶する第 3 のバッファと、

前記信号伝送路から受信した他サブシステムの前記プロセッサ部から自サブシステムの前記入出力部へアクセスされる前記データを一時的に記憶する第 4 のバッファと、

40

前記第 1 のバッファに記憶された前記誤り検出用データと前記第 2 のバッファに記憶された前記誤り検出用データとを比較する比較器と、

前記プロセッサ部が前記ロックステップ同期の状態にあるときは、前記比較器で一致が検出されることを条件に前記第 3 のバッファに記憶された前記データを自サブシステムの前記入出力部へ出力し、前記プロセッサ部が前記ロックステップ非同期の状態にあるときは、前記第 3 のバッファに記憶された前記データまたは前記第 4 のバッファに記憶された前記データを自サブシステムの前記入出力部へ出力する第 2 のセレクタと

を有する

付記 11 に記載のフォールトトレラントシステムで用いる制御装置。

(付記 13)

50

互いに同一のハードウェアで構成された複数のサブシステムを備え、前記複数のサブシステムは、前記複数のサブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と、前記プロセッサ部に接続される入出力部と、前記プロセッサ部および前記入出力部間に接続される制御部と、前記制御部を介して前記複数のサブシステム間を互いに接続する信号伝送路とを有するフォールトトレラントシステムの動作方法であって、

前記制御部が、

前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信を行う

10

フォールトトレラントシステムの動作方法。

(付記 1 4)

前記信号伝送路は複数の信号伝送路で構成され、

前記制御部が、前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、ロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信、自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のうちの何れか一方の送信のために使用する制御を行う

20

付記 1 3 に記載のフォールトトレラントシステムの動作方法。

(付記 1 5)

前記制御部は、

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データの受信時刻と前記データとからチェックサムを生成し、該生成した前記チェックサムを前記誤り検出用データとして使用する

30

付記 1 4 に記載のフォールトトレラントシステムの動作方法。

(付記 1 6)

前記制御部は、

自サブシステムの前記プロセッサ部から前記入出力部へアクセスされる前記データが有する CRC を抽出し、該抽出した前記 CRC を前記誤り検出用データとして使用する

付記 1 4 に記載のフォールトトレラントシステムの動作方法。

(付記 1 7)

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いられ、自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と前記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する信号伝送路に接続される制御装置の動作方法であって、

40

前記制御装置が、前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行い、ロックステップ非同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信を行う

フォールトトレラントシステムで用いる制御装置の動作方法。

(付記 1 8)

前記信号伝送路は複数の信号伝送路で構成され、

50

前記制御装置が、前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行い、ロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信、自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のうちの何れか一方の送信のために使用する制御を行う

付記 17 に記載のフォールトトレラントシステムで用いる制御装置の動作方法。

10

(付記 19)

同一のハードウェアで構成された複数のサブシステムで構成されるフォールトトレラントシステムの各サブシステムで用いられ、自他サブシステム間においてロックステップ同期の状態および非同期の状態で作動可能なプロセッサ部と前記プロセッサ部に接続される入出力部との間に接続され、また、自他サブシステム間を互いに接続する信号伝送路に接続された制御装置を構成するコンピュータに、

前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信を行わせ、

ロックステップ非同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信を行わせるためのプログラム。

20

(付記 20)

前記信号伝送路は複数の信号伝送路で構成され、

前記コンピュータに、

前記プロセッサ部がロックステップ同期の状態にあるときは、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータから生成した誤り検出用データの他サブシステムへの送信と自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信とで同じ一つの前記信号伝送路を共用する制御を行わせ、ロックステップ非同期の状態にあるときは、前記同じ一つの前記信号伝送路を、自サブシステムの前記プロセッサ部から前記入出力部へアクセスされるデータの他サブシステムへの送信、自サブシステムの前記入出力部から前記プロセッサ部へアクセスされるデータの他サブシステムへの送信のうちの何れか一方の送信のために使用する制御を行わせる

30

付記 19 に記載のプログラム。

【符号の説明】

【0131】

1000…フォールトトレラントシステム

1100、1200…サブシステム

1110、1210…プロセッサ部

1120、1220…入出力部

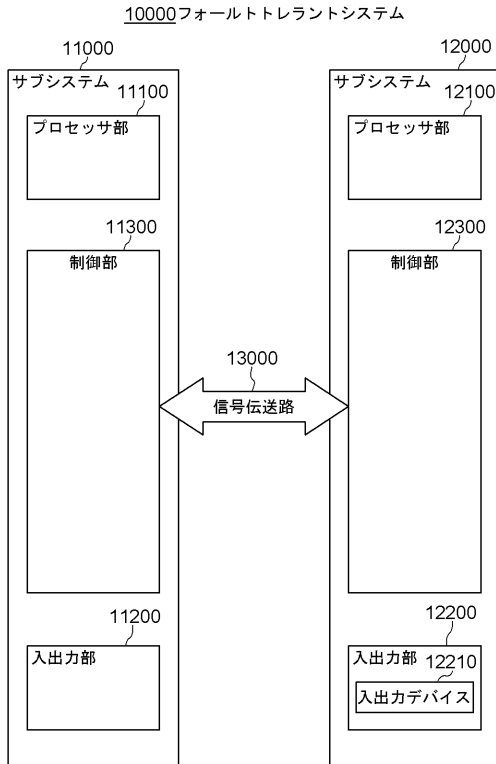
1121、1221…入出力デバイス

1130、1230…制御部

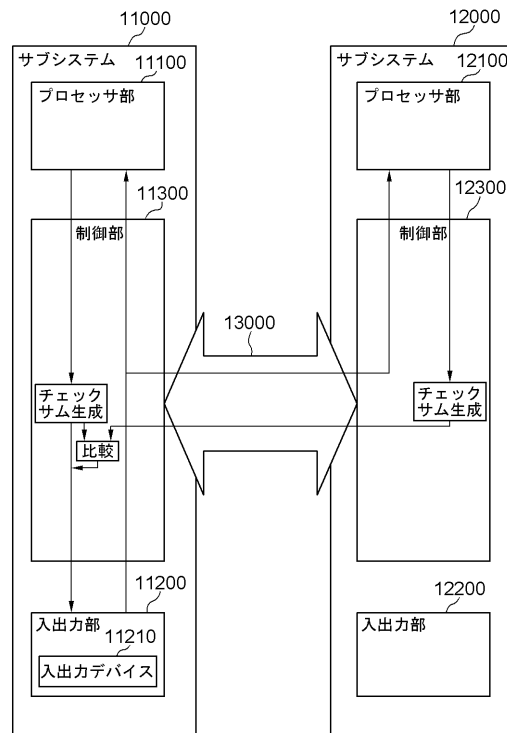
1300、1400…信号伝送路

40

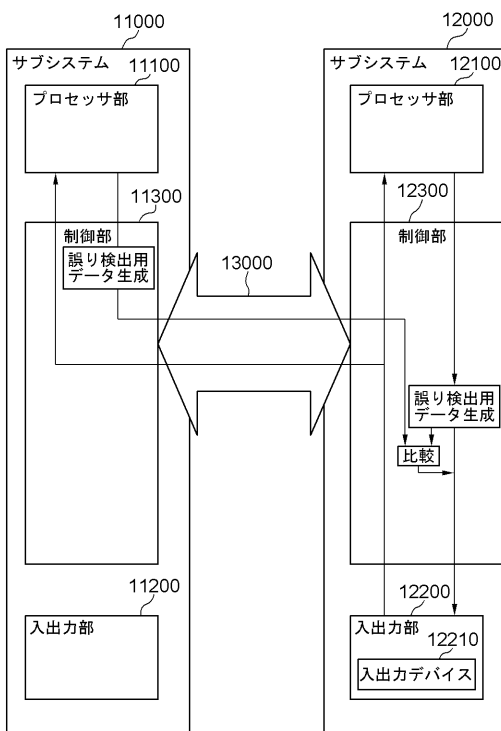
【図 1】



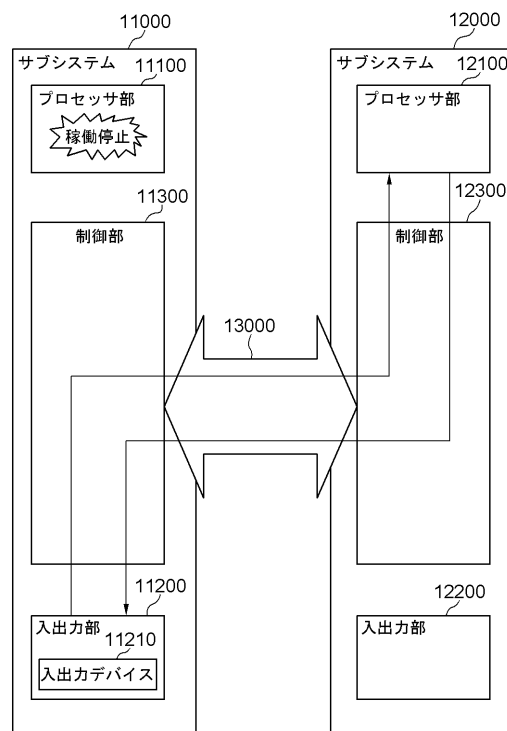
【図 2】



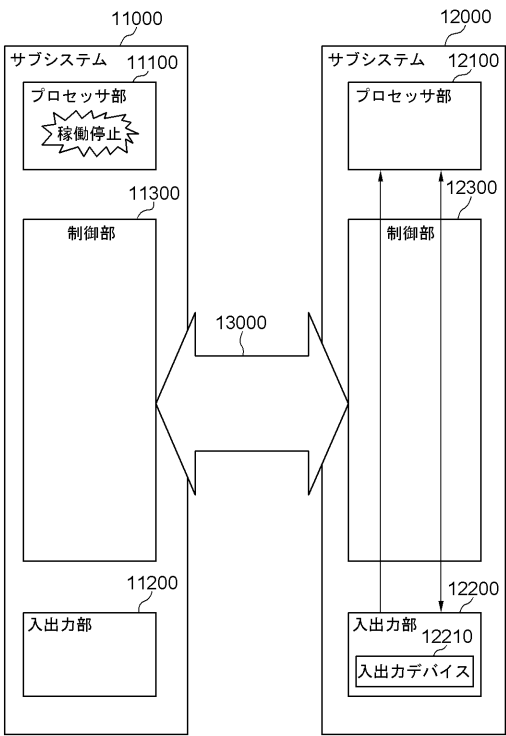
【図 3】



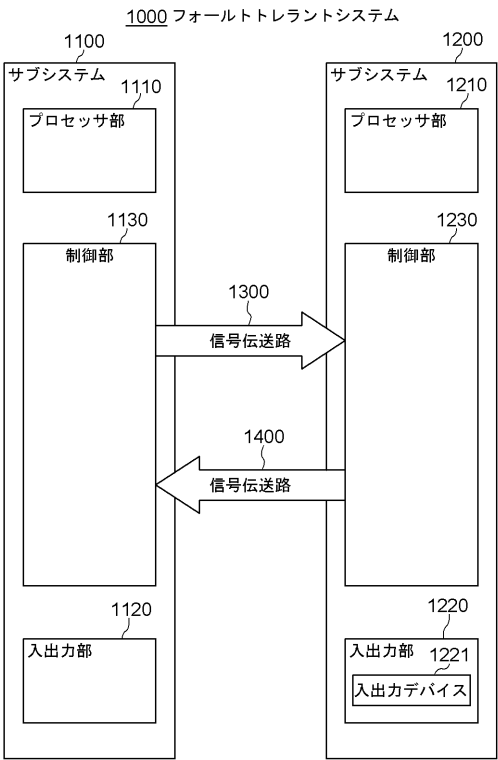
【図 4】



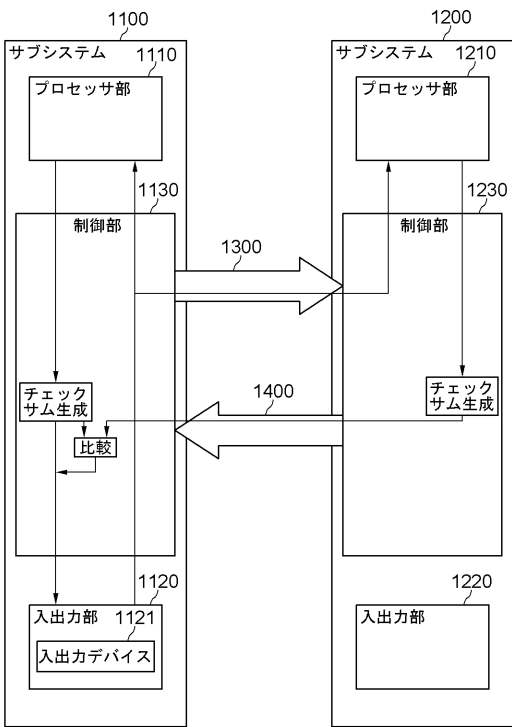
【図 5】



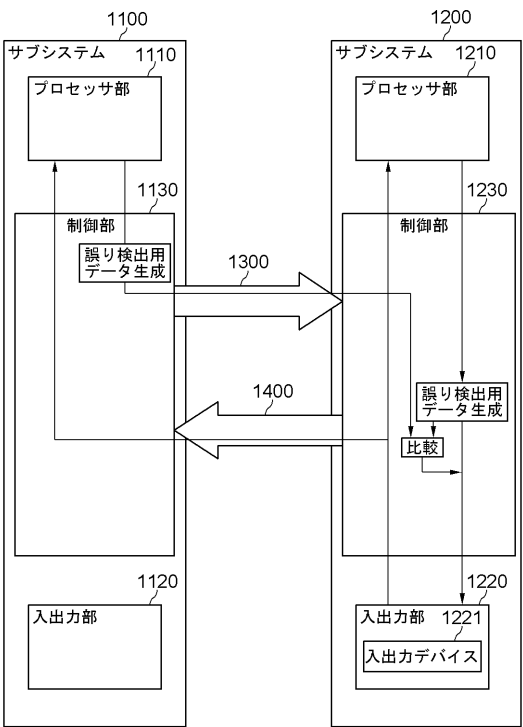
【図 6】



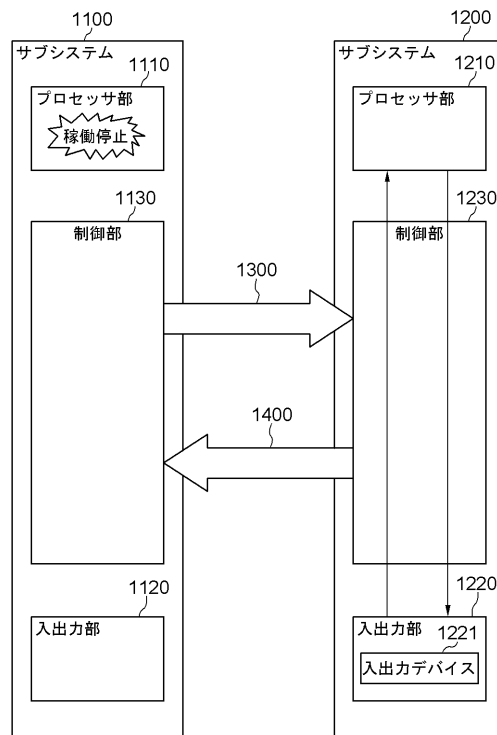
【図 7】



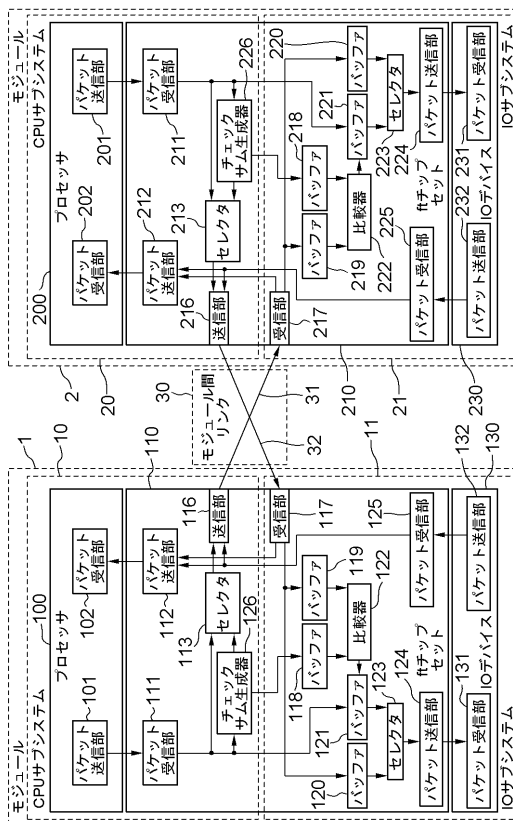
【図 8】



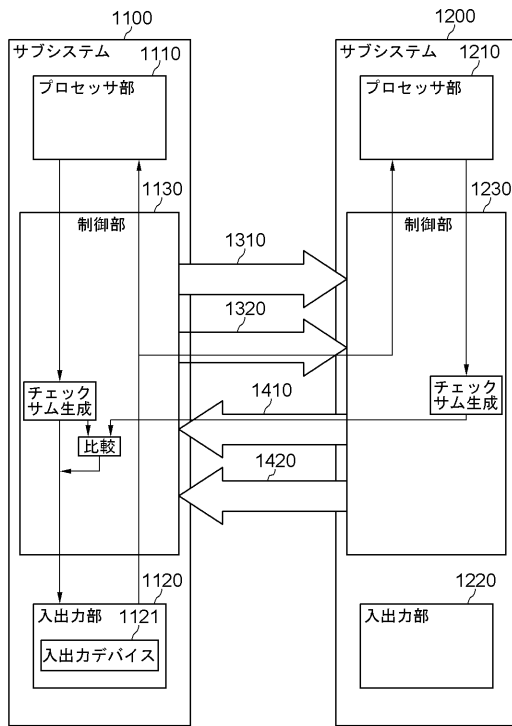
【図 10】



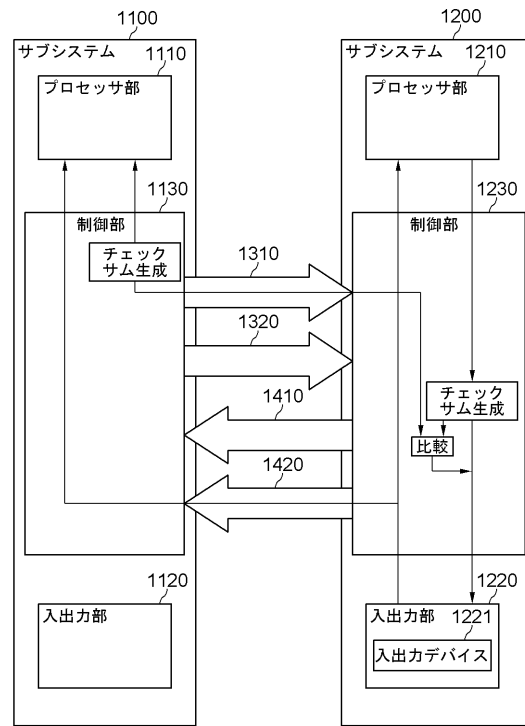
【図 1 2】



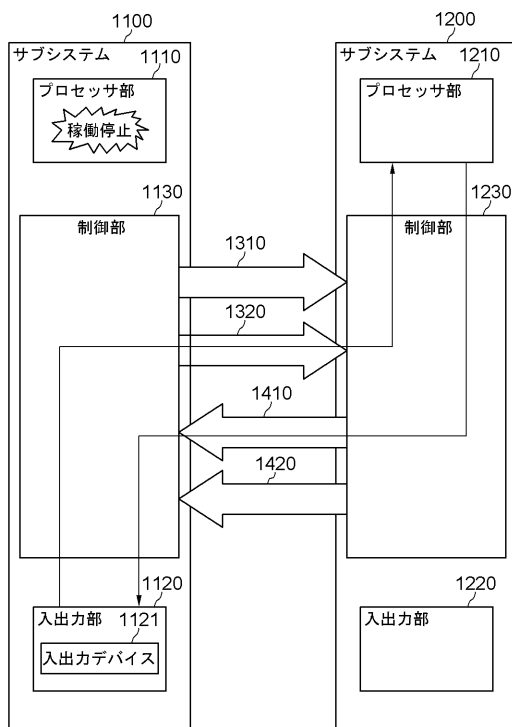
【図 1 3】



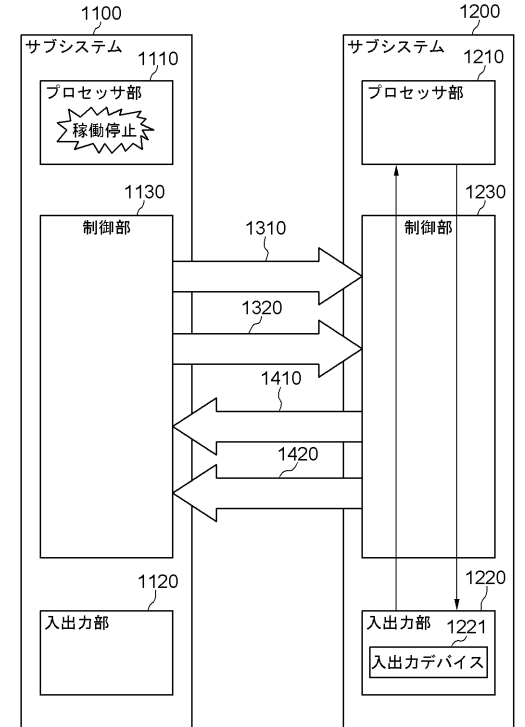
【図 1 4】



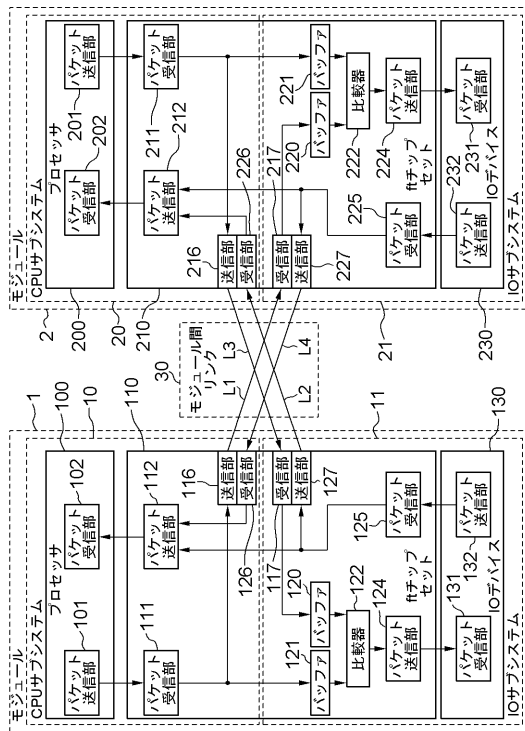
【図 1 5】



【図 1 6】



【図 17】



フロントページの続き

(56)参考文献 特開2006-178615 (JP, A)
特開2010-218370 (JP, A)

(58)調査した分野(Int.Cl., DB名)
G06F 11/18
G06F 11/16