

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7671008号
(P7671008)

(45)発行日 令和7年5月1日(2025.5.1)

(24)登録日 令和7年4月22日(2025.4.22)

(51)国際特許分類

F I

H O 4 L 47/30 (2022.01)

H O 4 L 47/30

請求項の数 10 (全27頁)

(21)出願番号	特願2023-552663(P2023-552663)	(73)特許権者	000004226
(86)(22)出願日	令和3年10月8日(2021.10.8)		日本電信電話株式会社
(86)国際出願番号	PCT/JP2021/037380		東京都千代田区大手町一丁目5番1号
(87)国際公開番号	WO2023/058232	(74)代理人	100083806
(87)国際公開日	令和5年4月13日(2023.4.13)		弁理士 三好 秀和
審査請求日	令和6年3月27日(2024.3.27)	(74)代理人	100129230
			弁理士 工藤 理恵
		(72)発明者	日比 智也
			東京都千代田区大手町一丁目5番1号
			日本電信電話株式会社内
		(72)発明者	市川 潤紀
			東京都千代田区大手町一丁目5番1号
			日本電信電話株式会社内
		(72)発明者	間野 暢
			東京都千代田区大手町一丁目5番1号
			最終頁に続く

(54)【発明の名称】 通信システム、中間装置、通信方法、および、プログラム

(57)【特許請求の範囲】

【請求項1】

第1の中間装置および第2の中間装置を備える通信システムであって、
前記第1の中間装置および前記第2の中間装置は、リモートダイレクトメモリアクセスを用いてデータを転送する第1の装置と第2の装置の間に配置され、
前記第1の中間装置は、
前記第2の中間装置との間の網の遅延に関する遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するバッファ管理部と、
前記第1の装置から前記第2の装置へ転送するデータが、前記第2の中間装置のクレジットより小さい場合、前記データを含むリクエストを前記第2の中間装置に転送する転送部と、
前記リクエストに対する疑似レスポンスを生成して前記第1の装置へ返却する生成部と、を備え、
前記第2の中間装置は、
前記遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するバッファ管理部と、
前記リクエストを前記第2の装置に転送し、前記リクエストを前記バッファに格納して自身のクレジットを更新する転送部と、
前記第2の装置から受信した前記リクエストに対するレスポンスを破棄し、前記バッファに格納したリクエストを削除し、前記クレジットを更新する破棄部と、を備える

通信システム。

【請求項 2】

前記第 1 の中間装置および前記第 2 の中間装置は、前記遅延情報を測定する網状態計測部を、それぞれ備える

請求項 1 記載の通信システム。

【請求項 3】

前記遅延情報を算出または測定し、前記第 1 の中間装置および前記第 2 の中間装置に通知するオーケストレータを備える

請求項 1 記載の通信システム。

【請求項 4】

前記第 1 の中間装置のバッファ管理部は、前記網に配置された第 1 の伝送装置から前記遅延情報を取得し、

前記第 2 の中間装置のバッファ管理部は、前記網に配置された第 2 の伝送装置から前記遅延情報を取得する

請求項 1 記載の通信システム。

【請求項 5】

前記第 1 の装置と前記第 2 の装置とを備え、

前記第 1 の装置および前記第 2 の装置のそれぞれは、

前記遅延情報に基づいて、前記第 1 の中間装置および第 2 の中間装置を経由して通信するか否かを判定する判定部と、

前記第 1 の中間装置および前記第 2 の中間装置を経由しない場合、前記遅延情報に基づいて、自装置のキューの深さを決定するキュー管理部と、を備える

請求項 1 から 4 のいずれか 1 項に記載の通信システム。

【請求項 6】

リモートダイレクトメモリアクセスを用いてデータを転送する第 1 の装置と第 2 の装置の間に配置される中間装置であって、

転送先の中間装置との間の網の遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するバッファ管理部と、

前記第 1 の装置から前記第 2 の装置へ転送するデータが、前記転送先の中間装置のクレジットより小さい場合、前記データを含むリクエストを前記転送先の中間装置に転送する転送部と、

前記リクエストに対する疑似レスポンスを生成して前記第 1 の装置へ返却する生成部と、を備える

中間装置。

【請求項 7】

第 1 の中間装置および第 2 の中間装置を備える通信システムが行う通信方法であって、

前記第 1 の中間装置および前記第 2 の中間装置は、リモートダイレクトメモリアクセスを用いてデータを転送する第 1 の装置と第 2 の装置の間に配置され、

前記第 1 の中間装置は、

前記第 2 の中間装置との間の網の遅延に関する遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するステップと、

前記第 1 の装置から前記第 2 の装置へ転送するデータが、前記第 2 の中間装置のクレジットより小さい場合、前記データを含むリクエストを前記第 2 の中間装置に転送するステップと、

前記リクエストに対する疑似レスポンスを生成して前記第 1 の装置へ返却するステップと、を行い、

前記第 2 の中間装置は、

前記遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するステップと、

前記リクエストを前記第 2 の装置に転送し、前記リクエストを前記バッファに格納し

10

20

30

40

50

て自身のクレジットを更新するステップと、

前記第 2 の装置から受信した前記リクエストに対するレスポンスを破棄し、前記バッファに格納したリクエストを削除し、前記クレジットを更新するステップと、を行う通信方法。

【請求項 8】

請求項 6 に記載の中間装置としてコンピュータを機能させるプログラム。

【請求項 9】

リモートダイレクトメモリアクセスを用いてデータを転送する第 1 の装置と第 2 の装置の間に配置される中間装置であって、

転送元の間装置との間の網の遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するバッファ管理部と、

前記転送元の間装置から転送された、前記第 1 の装置から前記第 2 の装置へ転送するデータを含むリクエストを、前記第 2 の装置に転送し、前記リクエストを前記バッファに格納して自身のクレジットを更新する転送部と、

前記第 2 の装置から受信した前記リクエストに対するレスポンスを破棄し、前記バッファに格納したリクエストを削除し、前記クレジットを更新する破棄部と、を備える中間装置。

【請求項 10】

請求項 9 に記載の中間装置としてコンピュータを機能させるプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、通信システム、中間装置、通信方法、および、プログラムに関する。

【背景技術】

【0002】

通信技術の進展により、広帯域かつ長距離を接続する新たなネットワークの提供が検討されている。光伝送システムに搭載されるトランスポンダ等の装置が、新たなネットワークにおいて、サーバ等のクライアントシステムに搭載されるシステムが検討されている。このようなシステムにおいて、電気信号は、伝送途中で電気-光変換がされることなく、通信相手のサーバ等に届く。

【0003】

光伝送システムにおいて、網状態（距離、信号品質など）から最適な通信モード（変調方式、ボーレート、キャリア数など）を選択することで、少ない網リソース（周波数など）で高速な伝送路を確立する技術が提案されている（特許文献 1、2 参照）。このような技術により、通信装置間の長距離・高速通信が、少ない網リソースで実現できる可能性がある。

【先行技術文献】

【特許文献】

【0004】

【文献】米国特許第 8 9 6 4 5 8 1 号明細書

【文献】国際公開第 2 0 2 0 / 0 3 1 5 1 4 号

【発明の概要】

【発明が解決しようとする課題】

【0005】

Infiniband のトランスポートレイヤに使用される R D M A （Remote Direct Memory Access）は、離れた距離にある通信端末同士で高速かつ高信頼なデータ転送を行う通信プロトコルである。R D M A は、送信装置のメモリ領域から受信装置のメモリ領域へ直接メモリアクセスするため、高速な通信が可能である。

【0006】

しかしながら、R D M A では、ネットワークが長距離になるほど転送性能は低下すると

10

20

30

40

50

いう問題がある。これは、長延化した回線ではより多くの転送時間が必要であり、コネクション型のプロトコルではデータ転送完了を知らせるパケットを受け取るまで次のパケットを送信できないからである。

【０００７】

本発明は、上記事情に鑑みてなされたものであり、本発明の目的は、ＲＤＭＡの転送先が長距離であっても、高速かつ高信頼なデータ転送を実現することにある。

【課題を解決するための手段】

【０００８】

上記目的を達成するため、本発明の一態様は、第１の中間装置および第２の中間装置を備える通信システムであって、前記第１の中間装置および前記第２の中間装置は、リモートダイレクトメモリアクセスを用いてデータを転送する第１の装置と第２の装置の間に配置され、前記第１の中間装置は、前記第２の中間装置との間の網の遅延に関する遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するバッファ管理部と、前記第１の装置から前記第２の装置へ転送するデータが、前記第２の中間装置のクレジットより小さい場合、前記データを含むリクエストを前記第２の中間装置に転送する転送部と、前記リクエストに対する疑似レスポンスを生成して前記第１の装置へ返却する生成部と、を備え、前記第２の中間装置は、前記遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するバッファ管理部と、前記リクエストを前記第２の装置に転送し、前記リクエストを前記バッファに格納して自身のクレジットを更新する転送部と、前記第２の装置から受信した前記リクエストに対するレスポンスを破棄し、前記バッファに格納したリクエストを削除し、前記クレジットを更新する破棄部と、を備える。

【０００９】

本発明の一態様は、リモートダイレクトメモリアクセスを用いてデータを転送する第１の装置と第２の装置の間に配置される中間装置であって、転送先の中間装置との間の網の遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するバッファ管理部と、前記第１の装置から前記第２の装置へ転送するデータが、前記転送先の中間装置のクレジットより小さい場合、前記データを含むリクエストを前記転送先の中間装置に転送する転送部と、前記リクエストに対する疑似レスポンスを生成して前記第１の装置へ返却する生成部と、を備える。

【００１０】

本発明の一態様は、第１の中間装置および第２の中間装置を備える通信システムが行う通信方法であって、前記第１の中間装置および前記第２の中間装置は、リモートダイレクトメモリアクセスを用いてデータを転送する第１の装置と第２の装置の間に配置され、前記第１の中間装置は、前記第２の中間装置との間の網の遅延に関する遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するステップと、前記第１の装置から前記第２の装置へ転送するデータが、前記第２の中間装置のクレジットより小さい場合、前記データを含むリクエストを前記第２の中間装置に転送するステップと、前記リクエストに対する疑似レスポンスを生成して前記第１の装置へ返却するステップと、を行い、前記第２の中間装置は、前記遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファを確保するステップと、前記リクエストを前記第２の装置に転送し、前記リクエストを前記バッファに格納して自身のクレジットを更新するステップと、前記第２の装置から受信した前記リクエストに対するレスポンスを破棄し、前記バッファに格納したリクエストを削除し、前記クレジットを更新するステップと、を行う。

【００１１】

本発明の一態様は、上記中間装置としてコンピュータを機能させるプログラムである。

【発明の効果】

【００１２】

本発明によれば、ＲＤＭＡの転送先が長距離であっても、高速かつ高信頼なデータ転送を実現することができる。

【図面の簡単な説明】

【0013】

【図1】図1は、RDMAの通信モデルを説明するための図である。

【図2】図2は、RDMAのSENDを説明するための図である。

【図3】図3は、第1の実施形態の通信システムの構成の一例を示す図である。

【図4】図4は、第1の中間装置の構成例を示す図である。

【図5】図5は、第2の中間装置の構成例を示す図である。

【図6】図6は、ローカルの構成例を示す図である。

【図7】図7は、図3に示す通信システムの処理の流れの一例を示すシーケンス図である。

【図8】図8は、テーブルを作成してレスポンスの宛先QPNを解決する方法の一例を説明するための図である。

10

【図9】図9は、ソースQPNを通知してレスポンスの宛先QPNを解決する方法の一例を説明するための図である。

【図10】図10は、第2の実施形態の通信システムの構成の一例を示す図である。

【図11】図11は、オーケストレータの構成例を示す図である。

【図12】図12は、伝送装置の構成例を示す図である。

【図13】図13は、図10に示す通信システムの処理の流れの一例を示すシーケンス図である。

【図14】図14は、第2の実施形態の変形例1の通信システムの構成の一例を示す図である。

20

【図15】図15は、第2の実施形態の変形例2の通信システムの処理の流れの一例を示すシーケンス図である。

【図16】図16は、第2の実施形態の変形例3の通信システムの構成の一例を示す図である。

【図17】図17は、第3の実施形態の通信システムの構成の一例を示す図である。

【図18】図18は、第3の実施形態の中間装置を経由しない場合の通信システムの構成の一例を示す図である。

【図19】図19は、第3の実施形態の他の通信システムの構成の一例を示す図である。

【図20】ハードウェア構成例である。

【発明を実施するための形態】

30

【0014】

以下、本発明の実施の形態について、図面を参照して説明する。

【0015】

1. RDMAについて

RDMAは、送信装置のメモリ領域から受信装置のメモリ領域へ直接メモリアクセスする通信プロトコルである。RDMAは、クレジットベースのフロー制御機能を持つほか、データ転送の完了を確認して処理を進めるCompletion制御を行うため、高信頼な通信が可能である。RDMAは、SSD (Solid State Drive) とGPU (Graphics Processing Unit) との間での、Host-to-Device, Device-to-Deviceのデータ通信のトランスポート方式としても使われる。

40

【0016】

図1に示すように、RDMAは、ローカル装置とリモート装置の間でQP (Queue Pair) を構成し、QPを用いてデータを転送する通信モデルである。QPは、SQ (Send Queue) とRQ (Receive Queue) のセットである。RDMAの通信単位はWR (Work Request) と呼ばれる通信要求であり、WQE (Work Queue Element) という単位でSQ/RQに積まれる。WRには、送信要求であるSend WRと受信要求であるReceive WRがある。

【0017】

Send WRでは、送信したいデータのメモリ領域をWQEに指定してSQへ積む。

Receive WRでは、データを受信したいメモリ領域をWQEに指定してRQへ積

50

む。SQ/RQのキューサイズ分、WQEをFIFO (First-In-First-Out) でSQ/RQに積むことができる。QPの間でWQEの処理が正常に完了すると、SQ/RQのそれぞれに対応したCQ (Completion Queue) に、正常完了を示すCQE (Completion Queue Entry) が積まれる。QPの間でWQEの処理がエラーで終了すると、CQにはエラーを示すCQEが積まれる。WQEが正常完了すると、SQ/RQのWQEは削除され、次のWQEの受け入れが可能になる。

【0018】

RDMAのサービスタイプは、Reliable/Unreliable、Connection/Datagramの区分により、RC (Reliable Connection)、RD (Reliable Datagram)、UC (Unreliable Connection)、およびUD (Unreliable Datagram) の4つに大別される。RCとUDが一般に使用される。

10

【0019】

RCは、ACK/NAKによる通信の成功・異常の確認と再送の仕組みによって、メッセージの順序性と到達性を保証するものである。また、RCは、コネクション型でもあり、ローカルーリモートのQP間で1対1の通信を行う。

【0020】

UDには、確認応答および再送の仕組みはないものの、通信ごとに宛先を指定して複数のQPへの送信および複数のQPからの受信といった多対多の通信が可能である。

【0021】

RDMAにおけるオペレーションタイプは、SEND、RDMA WRITE (WITH Immediate)、RDMA READ、およびATOMIC Operationsの4つに大別される。RCではこれら全てが使用できる。UDではSENDのみ使用できる。

20

【0022】

RDMAにおける再送制御は、ACK/NAKが返ってこない場合、RNR (Receiver-Not-Ready) NAKが返ってきた場合、およびOut-of-sequence NAKが返ってきた場合の3つのパターンに分類される。リモート側からACKまたはNAKが一定時間以内に返ってこない場合、ローカル側はタイムアウトとして再送する。また、リモート側は、RQでWQEが用意できない場合にRNR NAKを返す。リモート側からRNR NAKが返ってきた場合、ローカル側は一定時間後に再送する。また、リモート側は、受信したパケットのPSN (Packet Sequence Number) が順序通りでなかった場合にOut-of-sequence NAKを返す。リモート側からOut-of-sequence NAKが返ってきた場合、ローカル側は待たずに再送する。

30

【0023】

次に、サービスタイプがRCのオペレーションについて説明する。ここでは、SENDを例として説明する。

【0024】

図2は、RDMAのSENDを説明する説明図である。SENDは、RDMAの基本的な送受信モデルであり、ローカルからリモートへデータを送信する。

【0025】

ローカルはSQを用意し、リモートはRQを用意して、ローカルとリモートはそれぞれSQとRQにWQEを積む。

40

【0026】

通信の準備ができると、ローカルはSENDでデータを送信する。リモートは、データの受信が成功すると、CQにCQEを積み、RQのWQEを解放するとともに、ローカルへACKを返却する。ローカルは、ACKを受信すると、CQにCQEを積み、SQのWQEを解放する。

【0027】

また、SENDには、特殊なオペレーションであるSEND w/Imm (SEND with Immediate) が用意されている。SEND w/Immでは、ローカルはSQのWQEに特殊フィールド (imm_data) を設定することができ、ローカルが

50

らリモートへのデータ送信時に `imm_data` を同時に送信できる。リモートは、データの受信が成功すると、CQに `imm_data` を含むCQEを積む。リモートでは、CQEを参照することで、`imm_data` の内容を知ることができる。

【0028】

2. 第1の実施形態

<通信システムの構成>

次に、図3を参照し、本実施形態の中間装置10A、10Bを備える通信システムの構成の一例について説明する。

【0029】

中間装置10A、10Bは、RDMAを用いてデータを転送するローカル30とリモート50の間に配置される。より具体的には、中間装置10Aは、ローカル30側の長距離の網9（ネットワーク）の前段に配置され、中間装置10Bは、リモート50側の網9の前段に配置される。

10

【0030】

データ転送機能として、中間装置10Aは、ローカル30からリクエスト（SEND等）を受信し、当該リクエストに対する疑似レスポンスをローカル30へ返却する。中間装置10Bは、中間装置10Aから送信されたリクエストをリモート50に転送し、リモート50からのレスポンス（ACK等）を破棄する。フロー制御機能として、中間装置10A、10Bは、クレジットベースのフロー制御を実施し、自装置のクレジット（空きバッファ容量）を相手装置に所定のタイミングで定期的に送信する。

20

【0031】

通信システムは、中間装置10A、10Bの他に、RDMAを用いてデータを転送するローカル30およびリモート50を備えてもよい。ローカル30（第1の装置）は、データ転送元の装置である。リモート50（第2の装置）は、データ転送先の装置である。

【0032】

図4に、中間装置10A（第1の中間装置）の構成例を示す。中間装置10Aは、転送部11と、生成部12Aと、網状態計測部13と、バッファ14（一時データ保存部）と、バッファ管理部15と、クレジット管理部16と、通信部17とを備える。

【0033】

図示する例では、転送部11と、生成部12Aと、網状態計測部13とを、ソフトウェア（CPU、メモリ、ストレージ等）上で実装しているが、これらをNIC上で実装してもよい。また、図示する例では、バッファ14と、バッファ管理部15と、クレジット管理部16と、通信部17とは、NICで実装しているが、これらの一部をソフトウェアで実装してもよい。また、中間装置10Aは、仮想マシンやコンテナとして実装してもよい。

30

【0034】

転送部11は、ローカル30からリクエストを受信してリモート50へ転送する。このリクエストは、例えば、前述のSEND、SEND w/Imm、RDMA WRITE、RDMA WRITE w/Imm、またはATOMIC Commandなどである。リクエストには、ローカル30からリモート50へ送信するデータまたはデータに対する操作が含まれる。転送部11は、通信部17を介してリクエストを送信する。

40

【0035】

本実施形態の転送部11は、ローカル30からリモート50へ転送するデータ（データサイズ）が、中間装置10Bのクレジット（バッファの空きサイズ）より小さい場合、前記データを含むリクエストを中間装置10Bに転送する。クレジット管理部16は、定期的に中間装置10Bから送信されるクレジットを受信する。

【0036】

なお、本実施形態では、転送部11は、中間装置10Bのクレジットより小さいデータを送信するが、例えば、中間装置10Bから送信されるクレジットの開放通知履歴から、中間装置10Bのバッファ14のクレジットを推定し、推定したクレジットの範囲内であれば、中間装置10Bから通知されたクレジットを超えるサイズのデータを送信しても良

50

い。また、再送コストが小さいと推測される場合、通信データが中間装置 10B に届く頃には中間装置 10B のバッファに空きが出来ていると推定される場合など、転送部 11 は、中間装置 10B から通知されたクレジットを超えるサイズのデータを送信しても良い。これにより、クレジットベースのフロー制御において、データ通信の効率を向上させることができる。

【0037】

一方、転送部 11 は、ローカル 30 からリモート 50 へ転送するデータが、中間装置 10B から送信されたクレジット以上の場合、すなわちクレジットが不足している場合、バッファ 14 にリクエストをキューイング（格納）し、中間装置 10B のクレジットが増えるまで待機する。この場合、転送部 11 は、自身のバッファ 14 のクレジットを更新する。

10

【0038】

生成部 12A は、リクエストを中間装置 10B に転送すると、当該リクエストに対する疑似レスポンス（pseudo-Response）を生成してローカル 30 に返却する。具体的には、生成部 12A は、ローカル 30 から送信され、かつ Only または Last のフラグが立ったリクエストをピックアップし、リクエストに含まれる PSN を用いて疑似レスポンスを生成する。生成部 12A は、生成した疑似レスポンスをローカル 30 へ返却する。なお、リクエストに対する疑似レスポンス（例えば疑似 ACK）には、Only または Last のリクエストと同じ PSN の値が用いられる。

【0039】

ローカル 30 は、疑似レスポンスを受信すると、リモート 50 からのレスポンスと認識して CQ に CQE を積み、正常完了する。これにより、ローカル 30 の SQ の WQE を強制的に解放できる。

20

【0040】

網状態計測部 13 は、中間装置 10A、10B の間の網 9 の状態を測定する。具体的には、網状態計測部 13 は、パケットを送信するなどして、網 9 の遅延に関する遅延情報を測定する。遅延情報には、網 9（中間装置 10A、10B 間のパス）の伝送遅延、伝送容量などが含まれる。

【0041】

バッファ管理部 15 は、ローカル 30 とリモート 50 の間の網 9 の遅延情報を取得し、遅延情報に基づいてバッファ 14 のバッファサイズを決定し、決定したサイズのバッファ 14 を確保する。例えば、バッファ管理部 15 は、遅延情報の伝送遅延と伝送容量（帯域）の積の 2 倍以上の値をバッファサイズとして算出してもよい。バッファ 14 には、一時的にデータが格納される。

30

【0042】

クレジット管理部 16 は、クレジットベースのフロー制御を実施する。クレジット管理部 16 は、クレジットなどのフロー制御に関するデータを管理する。クレジット管理部 16 は、通信部 17 を介して、中間装置 10B から定期的にクレジット信号を受信し、当該クレジット信号と、所定のクレジット更新方式により、クレジット管理部 16 に格納されているクレジットを更新する。中間装置 10B から受信するクレジット信号は、中間装置 10B で受信可能な空きバッファサイズを示すクレジットを含む（Kung&Morris, Credit-based flow control for ATM networks, IEEE Network, 9(2), 40-48(1995)）。また、クレジット管理部 16 は、自身のバッファ 14 のクレジット（空きバッファサイズ）を、通信部 17 を介して中間装置 10B に定期的に送信してもよい。

40

【0043】

通信部 17 は、ローカル 30 および中間装置 10B 等の他の装置と通信するためのネットワークインタフェースである。

【0044】

図 5 に、中間装置 10B（第 2 の中間装置）の構成を示す。中間装置 10B は、転送部 11 と、破棄部 12B と、網状態計測部 13 と、バッファ 14 と、バッファ管理部 15 と、クレジット管理部 16 と、通信部 17 とを備える。

50

【0045】

転送部11は、中間装置10Aを介してローカル30が送信したリクエストをリモート50へ転送する。本実施形態の転送部11は、リクエストをリモート50に転送し、前記リクエストをバッファ14にキューイングして自身のクレジットを更新する。

【0046】

破棄部12Bは、リモート50から受信したリクエストに対するレスポンスを破棄する。これにより、ローカル30でのレスポンスの重複受信を防止できる。さらに、リモート50から送信されるRNRまたはout-of-sequenceによるNAKがローカル30へ到着すると誤作動を引き起こす可能性があるため、破棄部12Bは、これらのNAKも破棄する。本実施形態の破棄部12Bは、リモート50から受信した前記リクエストに対するレスポンスを破棄し、バッファ14にキューイングしたリクエストを削除し、クレジットを更新する。

10

【0047】

中間装置10Bの網状態計測部13、バッファ14、バッファ管理部15、クレジット管理部16および通信部17は、中間装置10Aの網状態計測部13、バッファ14、バッファ管理部15、クレジット管理部16および通信部17と同様である。

【0048】

図6に、ローカル30の構成を示す。ローカル30は、アプリケーション部31と、キュー管理部32と、網状態管理部33と、一時データ保存部34と、判定部35と、パケット振分部36と、通信部37とを備える。

20

【0049】

アプリケーション部31は、RDMA通信を用いて、転送データを含むリクエストをリモート50と送受信する。

【0050】

キュー管理部32は、一時データ保存部34の各キュー(CQ、SQ)を管理する。キュー管理部32は、網9の遅延情報に基づいて、必要なキューサイズを算出し、一時データ保存部34のキューの深さ(QueueDepth)を決定してもよい。ローカル30の一時データ保存部34は、キューとしてCQとSQとを備える。

【0051】

網状態管理部33は、網9の状態を管理する。具体的には、網状態管理部33は、他の装置から網9の遅延に関する遅延情報を取得してもよいし、パケットを送信するなどして、遅延情報を測定してもよい。遅延情報には、伝送遅延、伝送容量などが含まれる。

30

【0052】

振分部36は、通信部37を介して受信したデータ(パケット)を、対応する一時データ保存部34のQPに振り分ける。また、受信したデータが遅延情報の場合、網状態管理部33またはキュー管理部32に振り分ける。

【0053】

判定部35は、遅延情報に基づいて、中間装置10A、10Bを経由してリクエストを送信するか否かを判定してもよい。キュー管理部32は、中間装置10A、10Bを経由しないと判定した場合、遅延情報に基づいて一時データ保存部34のキューの深さを決定してもよい。本実施形態のローカル30は、網状態管理部33および判定部35を備えなくてもよい。

40

【0054】

通信部37は、中間装置10A等の他の装置と通信するためのネットワークインタフェースである。

【0055】

なお、キュー管理部32および網状態管理部33は、NICの機能として実装してもよい。アプリケーション部31は、ローカル30のCPUおよびメモリに実装される場合に限定されず、別の内部バスで接続されたGPU、FPGAまたはNICなどのハードウェアアクセラレータで実装されてもよい。

50

【0056】

リモート50の構成は、図6に示すローカル30と同様である。ただし、リモート50の一時データ保存部34は、CQおよびRQを備える。

【0057】

<通信システムの動作>

図7のシーケンス図を参照し、本実施形態の通信システムの処理の流れの一例について説明する。

【0058】

ステップS10において、各中間装置10A、10Bは、網9の遅延情報（伝送遅延、伝送容量）を測定し、測定結果に基づいてバッファ14を確保する。具体的には、中間装置10A、10Bは、パケットを送信するなどして網9の状態を測定する。中間装置10A、10Bは、遅延情報に基づいて必要なバッファサイズを算出し、当該バッファサイズのバッファ14を確保する。例えば、中間装置10A、10Bは、伝送遅延時間と伝送容量（帯域）の積の2倍以上の値をバッファサイズとして算出してもよい。

【0059】

なお、ステップS10において、中間装置10A、10Bは、ステップS10の計測結果から、自装置で利用可能な通信モードを相手装置と交換して、使用する通信モードを決定してもよい。

【0060】

次に、ステップS20において、ローカル30およびリモート50は、中間装置10A、10Bを介してデータを送受信する。具体的には、ローカル30は、SQにWQEを積み、リモート50へリクエストを送信する（ステップS21）。ここでは、SENDのリクエストが送信された場合を例として説明する。

【0061】

中間装置10Aは、ローカル30から受信したリクエストのデータサイズと中間装置10Bから受信したクレジットとを比較する。リクエストのデータサイズが中間装置10Bのクレジットより小さい場合、すなわち、クレジットが残っている場合、中間装置10Aは、リクエストを中間装置10Bに送信する（ステップS22）。このとき、中間装置10Aは、ローカル30から受信したリクエストをバッファにキューイングし、自身のクレジットを更新してもよい。

【0062】

中間装置10Aは、リクエストに含まれるPSNを用いて疑似レスポンス（疑似ACK）を生成し、疑似レスポンスをローカル30へ返却する（ステップS23）。ローカル30は、疑似レスポンスを受信すると、CQにCQEを積み、SQのWQEを解放する。

【0063】

中間装置10Aから送信されたリクエストにパケットロスが発生していない場合、中間装置10Bは、中間装置10Aから送信されたリクエストをリモート50に送信し（ステップS24）、当該リクエストをバッファ14にキューイングし、クレジットを更新する（ステップS25）。

【0064】

一方、パケットロスが発生している場合、中間装置10Bは、NACKを中間装置10Aに送信し（ステップS26）、中間装置10Aから再送されたリクエストを受信する（ステップS27）。再送されたリクエストにパケットロスが発生していない場合、中間装置10Bは、S24に進む。

【0065】

また、中間装置10Aは、ステップS21のリクエストのデータサイズが中間装置10Bのクレジット以上の場合、すなわち、中間装置10Bのクレジットが不足している場合、バッファ14にリクエストをキューイングして、中間装置10Bのクレジットが増えるまで待機する（ステップS28）。中間装置10Bのクレジットがリクエストのデータサイズより大きくなった場合、中間装置10Aは、S22に進む。

10

20

30

40

50

【0066】

リモート50は、ステップS24のリクエストの受信に成功した場合、ACKのレスポンスを中間装置10Bに送信する（ステップS29）。中間装置10Bは、リモート50から受信したレスポンスを破棄し、ステップS25でバッファ14にキューイングしたリクエストを開放し、クレジットを更新する（ステップS30）。

【0067】

一方、リモート50は、ステップS24のリクエストを正常に受信できない場合、NACKのレスポンスを中間装置10Bに送信し（ステップS31）、中間装置10Bから再送されたリクエストを受信する（ステップS32）。再送されたリクエストを正常に受信した場合、リモート50はステップS29に進み、中間装置10BはステップS30を行う。

10

【0068】

S20のデータ送受信とは非同期に、中間装置10Bは、図示しないタイマーなどを用いて、所定のタイミングで自身のクレジットを中間装置10Aに定期的に送信する（ステップS30）。これにより、中間装置10Aは、中間装置10Bのクレジットを取得し、クレジット管理部16に格納された中間装置10Bのクレジットを更新する。

【0069】

<宛先QPNの解決>

RDMAのインタフェースでは、QPは、エンドポイントごとに異なるQPNを持つ。SQ/RQは対向のQPNを認識しており、RDMAパケットを生成する際はdestination QPNをヘッダに含める。しかしながら、送信元のQPNはヘッダに含まれない。中間装置10Aが疑似レスポンスを生成する際、受信したリクエストに送信元のQPNを示す情報がいないため、疑似レスポンスの送り先が不明である。そこで、本実施形態では、以下に示す2通りの方法で、疑似レスポンスの送り先を特定する。

20

【0070】

第1の方法は、本来のRDMAのリクエストとレスポンスの交換を検査して、QPNの組み合わせをテーブルに記憶する方法である。RDMAパケットのOnlyまたはLastのリクエストとACKには同じPSNが使用される。そこで、中間装置10Aは、通過するリクエストとレスポンスを検査し、同じPSNを持つOnlyまたはLastのリクエストとACKのそれぞれのヘッダのdestination QPNを組み合わせとしてテーブルに追記する。図8の例では、同じPSNを持つリクエストとレスポンスのヘッダのdestination QPNは、それぞれ0x000020と0x000010であるので、0x000010と0x000020の組み合わせをテーブルに追記する。なお、図8では、ローカル30はリモート50Aとリモート50Bのそれぞれとの間でQPを構成している。

30

【0071】

中間装置10Aが疑似レスポンスを生成する際は、テーブルからリクエストのdestination QPNを含むQPNの組み合わせを取得し、組み合わせのもう片方のQPNを疑似レスポンスのdestination QPNに設定する。例えば、destination QPNが0x000020のリクエストを受信した場合、中間装置10Aは、テーブルから0x000020を含む0x000010と0x000020の組み合わせを取得し、疑似レスポンスのdestination QPNに0x000010を設定する。

40

【0072】

第2の方法は、Source QPNをRDMAパケットにのせる方法である。WQEには32bitのimmDt (immediate Date) フィールドが存在し、SEND with immediateまたはRDMA WRITE with immediateに限り、immDtフィールドに任意の32bitの情報を記載できる。図9では、ローカル30が挿入部38を備え、挿入部38がローカル30側SQのWQEのimmDtフィールドにローカル30側SQのQPNを書き込む。

50

【0073】

中間装置10Aが疑似レスポンスを生成する際は、受信したリクエストのimmDtフィールドに書き込まれたQPNを疑似レスポンスのdestination QPNに設定する。

【0074】

以上説明した本実施形態は、中間装置10Aおよび中間装置10Bを備える通信システムであって、中間装置10Aおよび中間装置10Bは、RDMAを用いてデータを転送するローカル30とリモート50の間に配置され、中間装置10Aは、中間装置10Bとの間の網9の遅延に関する遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファ14を確保するバッファ管理部15と、ローカル30からリモート50へ転送するデータが、中間装置10Bのクレジットより小さい場合、前記データを含むリクエストを中間装置10Bに転送する転送部11と、前記リクエストに対する疑似レスポンスを生成してローカル30へ返却する生成部12Aと、を備える。中間装置10Bは、遅延情報に基づいてバッファサイズを決定し、前記バッファサイズのバッファ14を確保するバッファ管理部15と、前記リクエストをリモート50に転送し、前記リクエストをバッファ14に格納して自身のクレジットを更新する転送部11と、リモート50から受信した前記リクエストに対するレスポンスを破棄し、バッファ14に格納したリクエストを削除し、前記クレジットを更新する破棄部12Bと、を備える。

10

【0075】

本実施形態では、ローカル30は中間装置10Aからの疑似レスポンスに応じてSQのWQEを解放するので、ローカル30とリモート50の間のRTT(Round Trip Time)が大きい場合でも、リモート50からのレスポンスを待つことなく、高帯域なデータ転送を実現できる。

20

【0076】

また、本実施形態では、中間装置10A、10Bにクレジットベースのフロー制御を導入することで、リモート50側の中間装置10Bの性能以上のデータが送信されてしまうことを抑制することができる。すなわち、中間装置10Aは、中間装置10Bが受け取ることができるデータ量を考慮してデータを送信する。このため、中間装置10Bが受け取ることができる性能以上のデータの送信を抑止し、リモート側でのデータのロスの発生を回避することができる。したがって、本実施形態では、RDMAの転送先が長距離であっても、高速かつ高信頼なデータ転送を実現し、中間装置10A、10B間の通信を保証することができる。

30

【0077】

また、本実施形態では、中間装置10A、10BにNACKに対する再送機能を付与することで、データロスに対応することができる。

【0078】

3. 第2の実施形態

<通信システムの構成>

第1の実施形態の通信システムでは、データ転送を開始する際に、中間装置10A、10Bが網9の状態を測定するため、データ転送開始までに時間を要する。本実施形態では、オーケストレータ70が、網9の光パス(伝送路)を設定する際に、遅延を計算または測定し、中間装置10A、10Bに通知する。これにより、本実施形態では、第1の実施形態の中間装置10A、10Bによる網9の状態の測定処理を省略し、データ転送開始までの時間を短縮することができる。

40

【0079】

図10は、本実施形態の通信システムの構成の一例を示す図である。本実施形態の通信システムは、中間装置10A、10Bと、オーケストレータ70と、を備える。

【0080】

本実施形態の中間装置10A、10Bは、第1の実施形態の中間装置10A、10B(図4、図5)と同様である。ただし、本実施形態の中間装置10A、10Bの網状態計測

50

部 13 は、網 9 状態を計測することなく、オーケストレータ 70 から遅延に関する遅延情報（伝送遅延、伝送容量、光パス情報など）を取得する。

【0081】

オーケストレータ 70 は、光トランスポートネットワーク等の網 9 全体を監視・管理して集中的に制御する。オーケストレータ 70 は、上位レイヤの計算機（ローカル 30、リモート 50）と下位レイヤの光トランスポートネットワークとの間で動作し、計算機からの通信要件と光トランスポートネットワークの光伝送路の状態とに基づき光ネットワーク装置を設定・制御することで、光トランスポートネットワーク内に光パスを構築する。これにより、地理的に分散した計算機からの要求に応じて光トランスポートネットワークを人手によらず自動的に最適制御可能となる。

10

【0082】

図 9 は、オーケストレータ 70 の構成の一例を示す図である。本実施形態のオーケストレータ 70 は、ローカル 30 からの通信要件に応じ、網 9（光トランスポートネットワーク）に光パスを設定し、当該光パスの遅延情報を算出または測定する。

【0083】

図示するオーケストレータ 70 は、通信要求受信部 71 と、ACK 送信部 72 と、ACK 受信部 73 と、結果出力・送信部 74 と、スケジューラ部 75 と、設計部 76 と、トポロジ情報格納部 77 と、ノード情報格納部 78 と、状態監視・管理部 79 と、を備える。

【0084】

通信要求受信部 71 は、ローカル 30（またはリモート 50）から通信に関する通信要件を含む通信要求を受信する。通信要件とは、通信を行うアプリケーションの種類、必要帯域、データ総量、許容遅延時間、タスク完了時間、符号誤り率（BER：Bit Error Rate）、電力等である。通信要求とは、10msec 以内に本タスクを完了したい、データをバックアップする、1Tbite のデータをローカル 30 からリモート 50 へ転送する等である。

20

【0085】

ACK 送信部 72 は、網 9 内の各伝送装置 20A、20B（光 NW 装置）から設定情報の設定完了応答を受信した後に、光パスの設定完了通知をローカル 30 およびリモート 50 へ送信する。つまり、ACK 送信部 72 は、網 9 で光パスの構築が完了した後に、要求元のローカル 30 と要求先のリモート 50 が通信できる状態になり次第、光パスの構築完了を示す ACK を要求元のローカル 30 に通知する。

30

【0086】

ACK 受信部 73 は、各伝送装置 20A、20B から、当該装置が備える光 NW 装置への設定情報の設定完了応答をそれぞれ受信する。

【0087】

結果出力・送信部 74 は、各伝送装置 20A、20B の光 NW 装置への設定情報を各伝送装置 20A、20B へそれぞれ送信する。つまり、結果出力・送信部 74 は、網 9 を最適制御するために設計・選択された以下のような設定情報を、網 9 を構成する光 NW 装置群へ送信する。

【0088】

- ・ 計算機間の光パス
- ・ OEO (optical-electro-optical) 変換点 (=DSP の挿入位置)
- ・ 各伝送装置 20A、20B のトランシーバに設定する、(i) 変調方式、(ii) ボーレート、(iii) 送信電力、(iv) FEC (Forward Error Correction) の overhead

より具体的には、以下のような設定情報が想定される。

40

【0089】

- ・ 通信要件を満たすように設計・選択された、計算機間の光パス、OEO 変換点、各伝送装置 20A、20B に存在する光 NW 装置群への (i) 変調方式、(ii) ボーレート、(iii) 送信電力、(iv) FEC (overhead) の設定

- ・ 光トランスポートネットワーク全体のリソース使用量が最小となるように設計・選択された、計算機間の光パス、OEO 変換点、各伝送装置 20A、20B に存在する光 NW 装

50

置群への (i) 変調方式、(ii) ボーレート、(iii) 送信電力、(iv) FEC (overhead) の設定

・光トランスポートネットワーク全体の電力使用量が最小となるように設計・選択された、計算機間の光パス、OEO変換点、各伝送装置 20A、20B に存在する光NW装置群への (i) 変調方式、(ii) ボーレート、(iii) 送信電力、(iv) FEC (overhead) の設定

結果出力・送信部 74 は、上記設定情報を各伝送装置 20A、20B へそれぞれ送信し、各伝送装置 20A、20B の光NW装置に当該設定情報を設定させる。なお、光NW装置とは、トランシーバ (トランスポンダ)、光クロスコネクタ (OXC: optical cross connect)、光分岐挿入装置 (ROADM: Re-configurable Optical Add/Drop Multiplexer)、増幅器等である。

10

【0090】

スケジューラ部 75 は、ローカル 30 から受信した通信要求を管理・スケジューリングする。光パスの設計・開通が完了するまでには一定の時間を要するので、複数の通信要求が混在する場合、スケジューラ部 75 は、優先度の高い通信要求からスケジューリングする。

【0091】

設計部 76 は、ローカル 30 から受信した通信要件と網 9 の光伝送路の状態とに基づき、ローカル 30 の通信に用いる光パスを決定し、当該光パスを構築するために各伝送装置 20A、20B の光NW装置に設定される設定情報をそれぞれ計算する。

【0092】

設計部 76 は、光パス設計・選択部 761 と、必要帯域計算部 762 と、伝送モード選択部 763 と、電力計算部 764 と、OSNR 計算部 765 と、BER 計算部 766 と、マルチフロー判定部 767 と、遅延計算部 768 と、タスク完了時間計算部 769 と、を備える。

20

【0093】

光パス設計・選択部 761 は、候補となる光パスの集合を列挙し、ローカル 30 から受信した通信要件に基づき最適な光パスを選択する。最適とは、例えば、遅延、タスク完了時間、必要帯域、網 9 全体のリソース使用量、網 9 全体の消費電力について、最適である場合をいう。

【0094】

最適な光パスを選択するため、光パス設計・選択部 761 は、必要帯域計算部 762 ～遅延測定部 770 と連携し、トポロジ情報格納部 77 ～状態監視・管理部 79 に格納された網 9 に関する情報を参照する。なお、伝送容量と距離にはトレードオフの関係があるが、その関係を考慮して距離に応じた最適な光伝送モードの選択する手法については、後述の参考文献 1 ～ 3 を利用できる。

30

【0095】

必要帯域計算部 762 は、ローカル 30 とリモート 50 (要求元、要求先) の通信において、ローカル 30 から受信した通信要件に基づき必要な帯域を算出する。

【0096】

伝送モード選択部 763 は、ローカル 30 とリモート 50 の通信で必要とされる帯域やアプリケーションの種類を基に、候補となる伝送モードを列挙する。伝送モード選択部 763 は、ローカル 30 とリモート 50 間のある候補となる光パス上で、候補となる伝送モードに対し、各トランシーバに設定すべき変調方式、ボーレート、FEC等を算出する。なお、これらのパラメータを基に最適な伝送モードを選択する方法については、参考文献 4 (W02020/031514 A1) を利用できる。

40

【0097】

電力計算部 764 は、ローカル 30 とリモート 50 間のある候補となる光パスにおいて、各トランシーバに対して設定すべき適切な送信電力を計算する。

【0098】

OSNR 計算部 765 は、ローカル 30 とリモート 50 間のある候補となる光パス上に

50

におけるOSNRを計算する。OSNR計算部765は、ローカル30とリモート50間のある光パスのネットワーク状態を入力値とし、入力光パス上のOSNRを出力値とする。OSNRの計算方法については、参考文献1を利用できる。

【0099】

参考文献1は、「A. Ferrari、外8名、“GNPy: an open source application for physical layer aware open optical networks”、in Journal of Optical Communications and Networking、vol.12、no.6、2020年、p.C31-p.C40」である。

【0100】

BER計算部766は、ある光パス上におけるOSNRやマージンを基に、BERを計算する。例えば、参考文献2を参考にすると、波長 λ で光パス p を確立したときの $BER_{p,\lambda}$ は、式(1)で計算可能である。

【0101】

$$BER_{p,\lambda} = \psi(OSNR_{p,\lambda} - M_T(\tau) - M_d(\tau)) \quad \dots (1)$$

なお、 $\psi(\cdot)$ は、特定の光パスの特定の変調フォーマットに応じたBERの計算式である。 $M_T(\tau)$ は、経年劣化を踏まえたシステムマージンである。 $M_d(\tau)$ は、デザインマージンである。

【0102】

参考文献2は、「P. Soumplis、外4名、“Multi-period planning with actual physical and traffic conditions”、in IEEE/OSA Journal of Optical Communications and Networking、vol.10、no.1、2018年、p.A144-p.A153」である。

【0103】

マルチフロー判定部767は、ローカル30とリモート50間のある候補となる光パス上において、ローカル30から受信した通信要件に基づき当該光パスを構成する区間上でマルチフロー化が必要か否かを判定する。通信のマルチフロー化に関する技術については、参考文献3を利用できる。

【0104】

参考文献3と、「M. Jinno、外4名、“Multiflow optical transponder for efficient multilayer optical networking”、in IEEE Communications Magazine、vol.50、no.5、2012年、p.56-p.65」である。

【0105】

遅延計算部768は、ローカル30とリモート50間のある光パス上における伝送遅延時間を計算する。伝送遅延時間の計算において、遅延計算部768は、ローカル30とリモート50間のある光パス及び網9のトポロジ情報を入力値とし、入力光パス上の伝送遅延を出力値とする。なお、伝送遅延時間は、基本的にはノード間の距離で決まる。

【0106】

タスク完了時間計算部769は、選択された候補の光パスを基に通信のタスク完了に要するタスク完了時間を計算する。

【0107】

遅延計測部770は、ローカル30とリモート50間のある光パス上における伝送遅延時間を、各伝送装置20A、20Bを用いて実際に計測してもよい。伝送遅延時間は、遅延計算部768が計算してもよく、遅延計測部770が測定してもよい。

【0108】

トポロジ情報格納部77は、伝送装置20A、20B間の接続及び距離に関するトポロジ情報を保持する。当該トポロジ情報は、光パス設計、マルチフロー判定等に必要となる。

【0109】

ノード情報格納部78は、伝送装置20A、20Bに存在するトランシーバの種類・数に関するノード情報を保持する。当該ノード情報は、トランシーバの種類によって、受信感度(必要なOSNR)が異なるため、光パス設計に必要となる。ノード情報格納部78は、各伝送装置20A、20Bに存在する増幅器の種類・数に関するノード情報も保持する。当該ノード情報は、ラマン増幅やEDFA(Erbium Doped Fiber Amplifier)など、増

10

20

30

40

50

幅器の種類によって雑音指数（NF：Noise Figure）が異なるため必要となる。ノード情報格納部 78 は、各伝送装置 20A、20B に存在する光ノード（OXC、ROADM）の種類・形態（ポート数等）に関するノード情報も保持する。

【0110】

状態監視・管理部 79 は、各伝送装置 20A、20B に存在する増幅器、トランシーバ、光ノードの使用状況及び状態に関する監視・管理情報を保持する。状態監視・管理部 79 は、各リンクにおける波長の使用状況、経年劣化、各リンクにおけるスプライス、コネクタ箇所等、光信号のロスに関する監視・管理情報も保持する。当該監視・管理情報は、非線形効果（特に XPM（Cross Phase Modulation）、FWM（four-wave mixing））が波長使用状況によってロスや BER に影響するため、光パス（波長パス）の選択時に必要な情報となる。状態監視・管理部 79 は、光パスの追加・削除に伴い、格納情報を更新する。状態監視・管理部 79 は、各伝送装置 20A、20B の光伝送路計測部に接続されており、定期的に伝送路の状態・マージンに関する伝送路情報を受け取り、当該伝送路情報を逐次更新可能に格納する。

10

【0111】

図 12 は、各伝送装置 20A、20B の構成の一例を示す図である。伝送装置 20A、20B は、網 9 を構成するノードである。各伝送装置 20A、20B は、結果受信部 21 と、制御部 22 と、光 NW 装置 23 と、ACK 送信部 24 と、光伝送路計測部 25 と、を備える。

【0112】

結果受信部 21 は、オーケストレータ 70 から自ノードの光 NW 装置 23 への設定情報を受信する。つまり、結果受信部 21 は、オーケストレータ 70 から光 NW 装置 23 へ設定すべき設定情報を受信し、当該設定情報を制御部 22 へ渡す。

20

【0113】

制御部 22 は、オーケストレータ 70 から受信した設定情報を自ノードの光 NW 装置 23 に設定・制御する。つまり、制御部 22 は、オーケストレータ 70 から受信した設定情報を基に光 NW 装置 23 を以下のように設定・制御する。

【0114】

・各トランシーバ（トランスポンダ）へ伝送モード（変調方式、ボーレート、FEC、等）や送信電力を設定

30

・add-drop/throughする波長をROADMやOXCへ設定

・増幅すべき波長の増幅器への割当

制御部 22 は、設定・制御が完了し次第、その旨を ACK 送信部 24 へ通知する。

【0115】

光 NW 装置 23 は、トランシーバ、OXC、ROADM、増幅器等、網 9 を構成する装置である。

【0116】

ACK 送信部 24 は、光 NW 装置 23 への設定情報の設定を完了した後に、設定完了応答を当該オーケストレータ 70 へ送信する。つまり、ACK 送信部 24 は、光 NW 装置 23 への設定・制御が完了し次第、当該完了を示す ACK をオーケストレータ 70 へ通知する。

40

【0117】

光伝送路計測部 25 は、網 9 の光伝送路（光バス）の状態・マージンを計測してオーケストレータ 70 へ送信する。つまり、光伝送路計測部 25 は、実際の光伝送路の状態・マージンを計測する計測装置である。例えば、光伝送路計測部 25 は、参考文献 5 より、コヒーレント DSP で受信した信号をもとに、光伝送路の状態を定期的又は要求があった際に推定計算し、当該光伝送路の状態情報をオーケストレータ 70 へ送信する。伝送路計測部 25 は、光伝送路の伝送遅延時間、伝送容量などの遅延に関する遅延情報を計測してもよい。

【0118】

50

参考文献 5 は、「T. Sasai、外 8 名、“Simultaneous Detection of Anomaly Points and Fiber types in Multi-Span Transmission Links Only by Receiver-side Digital Signal Processing”、in Optical Fiber Communication Conference (pp. Th1F-1)、Optical Society of America、2020 年」である。

【0119】

本実施形態のローカル 30 のアプリケーション部 31 は、リモート 50 と通信する際に、通信に関する通信要件を含む通信要求をオーケストレータ 70 へ送信する。また、網状態管理部 33 は、オーケストレータ 70 から網 9 で構築された光パスの設定完了通知 (ACK)、および、設定された光パスの遅延情報を受信する。遅延情報には、伝送遅延時間、伝送容量、光パスに関する情報などが含まれる。ローカル 30 は、波長を動的に切替可能な波長可変トランシーバを備えてもよい。リモート 50 は、ローカル 30 と同様である。

10

【0120】

＜通信システムの動作＞

図 13 は、本実施形態の通信システムの動作を示すシーケンス図である。

【0121】

ローカル 30 は、リモート 50 と通信するため、通信要件を指定した通信要求 (パス設定要求) をオーケストレータ 70 へ送信する (ステップ S51)。通信要求とは、データやファイルの転送、映像の伝送や配信等である。通信要件には、当該通信を行うアプリケーションの種類、必要帯域、データ総量、許容遅延時間、タスク完了時間、BER、電力等が含まれる。ローカル 30 は、通信要求を中間装置 10A または伝送装置 20A に送信してもよい。その場合、中間装置 10A または伝送装置 20A は、オーケストレータ 70 に通信要求を送信する。

20

【0122】

オーケストレータ 70 は、ローカル 30 からの通信要件と、網 9 に配置された各伝送装置から予め受信していた網 9 の状態とに基づき、ローカル 30 の通信に用いる光パスを決定し、当該光パスを構築するために各伝送装置 20 の光 NW 装置 23 に設定される設定情報をそれぞれ計算する (ステップ S52)。

【0123】

例えば、オーケストレータ 70 は、ローカル 30 とリモート 50 間の光パス、OEO 変換点、各伝送装置 20 のトランシーバに設定する、(i) 変調方式、(ii) シンボルレート、(iii) 送信電力、(iv) FEC (overhead)、に関する設定情報を計算する。

30

【0124】

その際、オーケストレータ 70 は、遅延、タスク完了時間、必要帯域、網 9 全体のリソース使用量、網 9 全体の消費電力が最適になるように計算する。例えば、オーケストレータ 70 は、上記通信要件を満たすように設定情報を計算するとともに、網 9 全体のリソース使用量が最小となるように設定情報を計算し、かつ／又は、網 9 全体の電力使用量が最小となるように設定情報を計算する。

【0125】

オーケストレータ 70 は、上記設定情報を伝送装置 20A および伝送装置 20B へそれぞれ送信し、光パスの設定を指示する (ステップ S53、S54)。

40

【0126】

伝送装置 20A および伝送装置 20B は、それぞれ、上記設定情報を自ノードの光 NW 装置 23 に設定・制御する。例えば、伝送装置 20A および伝送装置 20B は、上記設定情報をもとに、各トランシーバへ伝送モード (変調方式、シンボルレート、FEC、等) や送信電力を設定し、add-drop/through する波長を ROADM や OXC へ設定し、増幅すべき波長の増幅器への割当を実行する。

【0127】

伝送装置 20A および伝送装置 20B は、それぞれ、光 NW 装置 23 への設定・制御が完了し次第、当該完了を示す ACK をオーケストレータ 70 へ通知する (ステップ S55、S56)。

50

【0128】

オーケストレータ70は、設定した光パスの伝送遅延、伝送容量、光パスの情報などを含む遅延情報を、中間装置10A、10Bに送信する（ステップS57、S58）。

【0129】

伝送装置20Aおよび伝送装置20Bは、オーケストレータ70から送信された遅延情報に基づいてバッファサイズを算出し、当該サイズのバッファ14を確保する（ステップS59、S60）。例えば、伝送装置20Aおよび伝送装置20Bは、伝送遅延と伝送容量の積の2倍以上の値をバッファサイズとして算出する。

【0130】

オーケストレータ70は、全ての伝送装置20からACKを受信した後、光パスの設定完了通知を示すACKを要求元のローカル30へ通知する（ステップS61）。ローカル30は、オーケストレータ70からACKを受信した後、網9に構築された光パスを介してリモート50との間で通信を開始する。

10

【0131】

以上説明した本実施形態では、遅延情報を算出または測定し、中間装置10A、10Bに通知するオーケストレータ70を備える。第1の実施形態では中間装置10A、10Bが実際にパケットを送受信して遅延情報を測定するが、本実施形態ではオーケストレータ70が光パスを設定する際、遅延情報を計算または測定し、光パスの確立時に遅延情報を中間装置10A、10Bに通知する。これにより、本実施形態では、中間装置10A、10Bが、遅延情報を測定することなく、通信を開始までに要する時間を短縮することができる。

20

【0132】

なお、オーケストレータ70は、事前に計算または測定した遅延情報を読み出して、中間装置10A、10Bに通知してもよい。また、オーケストレータ70は、網9の状態を定期的に、または、網9が変更された場合に、中間装置10A、10Bに遅延情報を通知してもよい。

【0133】

<第2の実施形態の変形例1>

図14は、第2の実施形態の変形例1の構成を示す図である。第2の実施形態では、中間装置10A、10Bは、オーケストレータ70から光パスの遅延情報を取得したが、変形例1では、中間装置10A、10Bは、伝送装置20A、20Bから光パスの遅延情報を取得する。この場合、オーケストレータ70は、図13のステップS53およびS54で、算出した遅延情報を伝送装置20A、20Bに通知してもよい。

30

【0134】

<第2の実施形態の変形例2>

変形例1では、伝送装置20A、20Bは、オーケストレータ70から取得した光パスの遅延情報を、中間装置10A、10Bに通知した。変形例2では、伝送装置20A、20Bは、オーケストレータ70により設定された光パスの遅延情報を測定し、中間装置10A、10Bに通知する。本変形例の構成は、図14に示す変形例1の構成と同様である。

【0135】

図15は、変形例2の通信システムの動作を示すシーケンス図である。伝送装置20Aは、ローカル30（または、中間装置10A）からの要求を受信すると、伝送装置20Bと連携してローカル30とリモート50間に設定された光パスの伝送モードを決定し、遅延情報を測定する（ステップS71）。伝送装置20A、20Bは、測定した遅延情報を、中間装置10A、10Bに送信する（ステップS72、S73）。

40

【0136】

中間装置10Aは、伝送装置20Aまでの遅延情報を測定し、中間装置10Bは、伝送装置20Bまでの遅延情報を測定する（ステップS74、S75）。

【0137】

中間装置10Aは、ステップS71の遅延情報と、ステップS74の遅延情報とに基づ

50

いて、バッファサイズを算出し、当該サイズのバッファ 14 を確保する（ステップ S 7 6）。同様に、中間装置 10 B は、ステップ S 7 1 の遅延情報と、ステップ S 7 5 の遅延情報とに基づいて、バッファサイズを算出し、当該サイズのバッファ 14 を確保する（ステップ S 7 7）。

【0138】

なお、一般的に中間装置 10 A、10 B から伝送装置 20 A、20 B までの距離は比較的近く、伝送装置 20 A、20 B 間の距離は比較的遠い。そのため、伝送遅延の大部分は伝送装置 20 A、20 B 間の伝送遅延である。そのため、中間装置 10 A、10 B は、S 7 4、S 7 5 を省略し、伝送装置 20 A、20 B 間の遅延情報に基づいて、バッファサイズを算出し、当該サイズのバッファ 14 を確保してもよい。

【0139】

＜第 2 の実施形態の変形例 3＞

図 16 は、第 2 の実施形態の変形例 3 の構成を示す図である。変形例 3 では、中間装置が伝送装置の機能を備える。具体的には、中間装置 10 C は、第 2 の実施形態の中間装置 10 A が、伝送装置 20 A の機能を備えた装置であって、中間装置 10 D は、第 2 の実施形態の中間装置 10 B が伝送装置 20 B の機能を備えた装置である。

【0140】

本変形例では、オーケストレータ 70 が光パスを設定した後、中間装置 10 C、10 D に実装された伝送装置 20 A、20 B の機能が、トランシーバ（トランスポンダ）で波長を選択する際に遅延情報も測定する。中間装置 10 C、10 D は、測定された遅延情報に基づいてバッファサイズを決定する。

【0141】

4. 第 3 の実施形態

遅延と伝送容量が大きい場合、または、要求される通信要件（SLA）が厳しくない場合、ローカル 30 およびリモート 50 の一時データ保存部 34 のキューの深さ（QueueDepth）の値を大きくすることで、中間装置 10 A、10 B がなくても性能を維持できる場合がある。そのため、ローカル 30 は、通信要件に即した光パスをオーケストレータ 70 に要求し、オーケストレータ 70 から取得した光パスの遅延情報からローカル 30 のキューの深さを調整してもよい。これにより、中間装置 10 A、10 B を経由することなく、キューの深さの調整によって伝送遅延または伝送容量の増加による性能劣化を回避することができる。

【0142】

ただし、ローカル 30 およびリモート 50 は、ハードウェア、プロトコルなどの制約によってキューサイズを際限なく大きくすることはできない。そこで、本実施形態のローカル 30 は、アプリケーションの通信要件と、オーケストレータ 70 から取得した遅延情報とを用いて、中間装置 10 A、10 B を経由して通信するが、あるいは、キューの深さを調整するか否かを判定する。

【0143】

図 17 は、第 3 の実施形態の通信システムの一例を示す図である。本実施形態のシステムは、ローカル 30 と、リモート 50 と、中間装置 10 A、10 B と、オーケストレータ 70 と、ネットワーク装置 40 A、40 B と、を備える。本実施形態の中間装置 10 A、10 B およびオーケストレータ 70 は、第 2 の実施形態の中間装置 10 A、10 B およびオーケストレータ 70 と同様である。

【0144】

ローカル 30 およびリモート 50 は、第 1 の実施形態のローカル 30 およびリモート 50 と同様である。ただし、本実施形態のローカル 30 の判定部 35 は、遅延情報に基づいて中間装置 10 A および中間装置 10 B を経由して通信するか否かを判定する。例えば、判定部 35 は、遅延情報の伝送遅延と伝送容量の積の 2 倍以上の値を、一時データ保存部 34 のキューサイズとして確保できる場合は、中間装置 10 A および中間装置 10 B を経由しないと判定する。この場合、ローカル 30 のキュー管理部 32 は、遅延情報に基づい

10

20

30

40

50

て一時データ保存部 34 のキューサイズおよびキューの深さを決定する。キューの深さは、キューサイズおよびパケットサイズに基づいて決定される。

【0145】

同様に、リモート 50 の判定部 35 も、遅延情報に基づいて中間装置 10A および中間装置 10B を経由するか否かを判定し、中間装置 10A および中間装置 10B を経由しない場合、キュー管理部 32 は、遅延情報に基づいて一時データ保存部 34 のキューサイズおよびキューの深さを決定する。

【0146】

ネットワーク装置 40A は、ローカル 30 の指示に従って、ローカル 30 が送信するリクエストをルーティングする。具体的には、ネットワーク装置 40A は、リクエストの宛先情報からルーティングテーブルなどを用いて、リクエストの転送先を中間装置 10A とするか、またはネットワーク装置 40B とするかを決定する。

【0147】

同様に、ネットワーク装置 40B は、リモート 50 の指示に従って、リクエストに対するレスポンスの転送先を中間装置 10B とするか、またはネットワーク装置 40A とするかを決定する。

【0148】

図 18 は、中間装置 10A、10B を経由しない場合の通信システムの構成を示す図である。ローカル 30 およびリモート 50 は、オーケストレータ 70 から取得した遅延情報に基づいて中間装置 10A、10B を経由しないと判定し、遅延情報に基づいてキューの深さを調整する。そして、ローカル 30 は、図示しないネットワーク装置 40A を介してリクエストをリモート 50 に送信する。リモート 50 は、図示しないネットワーク装置 40B を介してレスポンスをローカル 30 に送信する。

【0149】

なお、図 19 に示すように、本実施形態の通信システムは、ネットワーク装置 40A、40B を備えることなく、ローカル 30 およびリモート 50 がリクエスト又はレスポンスの宛先を書き換えてもよい。具体的には、ローカル 30 およびリモート 50 の判定部 35 は、ヘッダの宛先情報を書き換えることで、リクエスト又はレスポンスの転送先を制御してもよい。

【0150】

本実施形態では、ローカル 30 およびリモート 50 が、遅延情報に基づいて、中間装置 10A、10B を経由して通信するか否かを判定する。このように、ローカル 30 およびリモート 50 が、遅延情報に応じて中間装置 10A、10B を経由するか、ローカル 30 およびリモート 50 のキューの深さを調整するかを自律的に選択することで、高速な通信を行うことができる。したがって、光パスや通信要件が動的に変更される場合であっても、変更後の光パスや通信要件の遅延情報に応じた最適な通信システムを自動で選択することができる。

【0151】

なお、ローカル 30 およびリモート 50 は、例えば、遅延情報の伝送遅延と伝送容量の積が所定の値以上の場合、伝送装置 20A、20B の高機能部に中間装置 10A、10B を自動で構築し、構築した中間装置を経由した通信を行い、遅延情報の伝送遅延と伝送容量の積が所定の値より小さい場合、キューの深さを調整してもよい。伝送装置 20A、20B に構築する中間装置 10A、10B の設定、もしくは中間装置 10A、10B を構築する伝送装置 20A、20B は、ローカル 30 およびリモート 50 の端末側で指定してもよいし、オーケストレータ 70 が端末からの距離、光パスの経路、回線の輻輳などの状況に基づいて判定し、決定しても良い。

【0152】

また、ローカル 30 およびリモート 50 は、通信要求や光パスの変更をオーケストレータから取得してもよく、あるいは、アプリケーション部 31 で監視してもよく、別アプリケーションで ping などによって遅延を監視したりして、中間装置 10A、10B を経由す

10

20

30

40

50

るか否かを判別してもよい。

【0153】

5. ハードウェア構成

上記説明したローカル30、リモート50、中間装置10A、10B、伝送装置20A、20B、オーケストレータ70のそれぞれは、例えば、図20に示すような汎用的なコンピュータシステムを用いることができる。図示するコンピュータシステムは、CPU（Central Processing Unit、プロセッサ）901と、メモリ902と、ストレージ903（HDD：Hard Disk Drive、SSD：Solid State Drive）と、通信装置904と、入力装置905と、出力装置906とを備える。メモリ902およびストレージ903は、記憶装置である。このコンピュータシステムにおいて、CPU901がメモリ902上にロードされた所定のプログラムを実行することにより、ローカル30、リモート50、中間装置10A、10B、伝送装置20A、20B、オーケストレータ70のそれぞれの機能が実現される。

10

【0154】

また、これらの装置は、1つのコンピュータで実装されてもよく、あるいは複数のコンピュータで実装されても良い。また、これらの装置は、コンピュータに実装される仮想マシンであっても良い。これらの装置のプログラムは、HDD、SSD、USB（Universal Serial Bus）メモリ、CD（Compact Disc）、DVD（Digital Versatile Disc）などのコンピュータ読取り可能な記録媒体に記憶することも、ネットワークを介して配信することもできる。

20

【0155】

なお、本発明は上記実施形態に限定されるものではなく、その要旨の範囲内で数々の変形が可能である。例えば、上記実施形態の中間装置10A、10Bおよび／または伝送装置20A、20Bは、ローカル30またはリモート50のNICまたはアプリケーションとして実装されてもよい。

【0156】

また、ローカル30およびリモート50は、中間装置10A、10Bまでの距離を測定し、当該距離に応じて一時データ保存部のキューの深さを設定してもよい。

【0157】

また、本発明は、上記第1から第3の実施形態の少なくとも2つを組み合わせてもよい。

30

【符号の説明】

【0158】

10A、10B、10C、10D：中間装置

11：転送部

12A：生成部

12B：破棄部

13：網状態計測部

14：バッファ

15：バッファ管理部

16：クレジット管理部

17：通信部

20A、20B：伝送装置

30：ローカル30

31：アプリケーション部

32：キュー管理部

33：網状態管理部

34：一時データ保存部

35：判定部

36：パケット振分部36

37：通信部

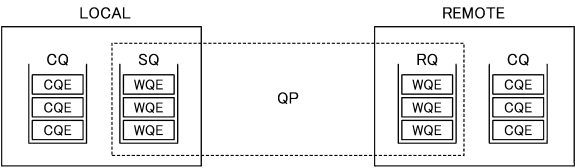
40

50

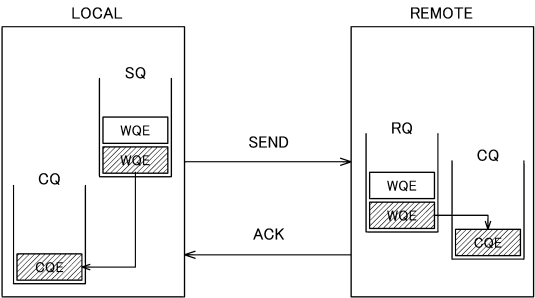
40A、40B：ネットワーク装置
50：リモート
70：オーケストレータ

【図面】

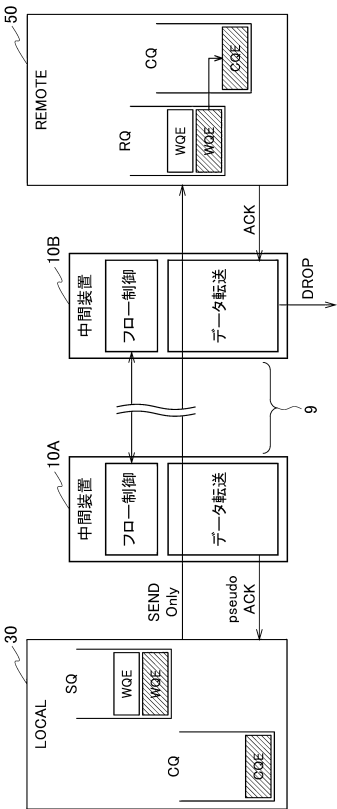
【図1】



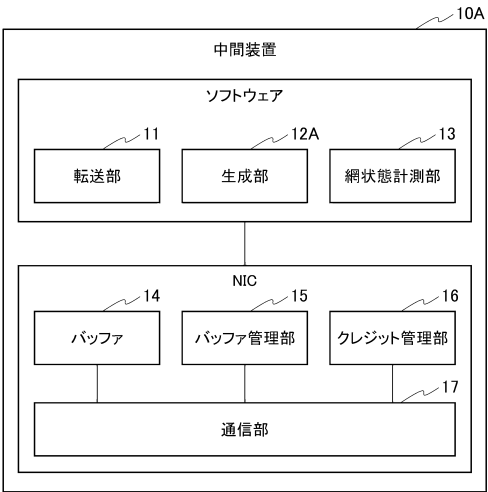
【図2】



【図3】



【図4】



10

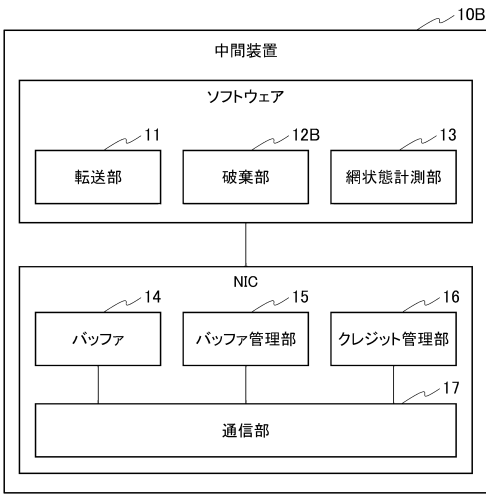
20

30

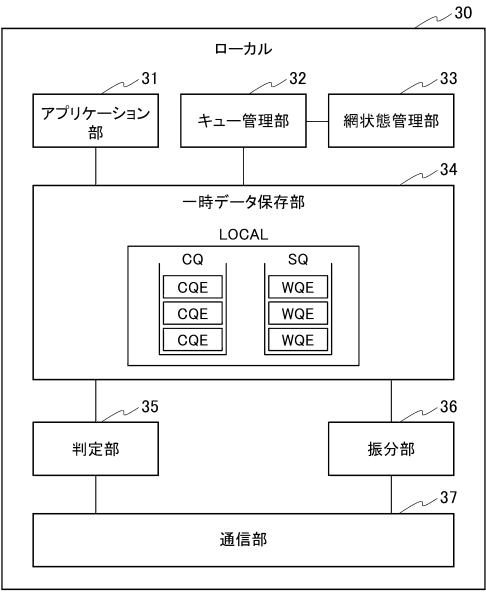
40

50

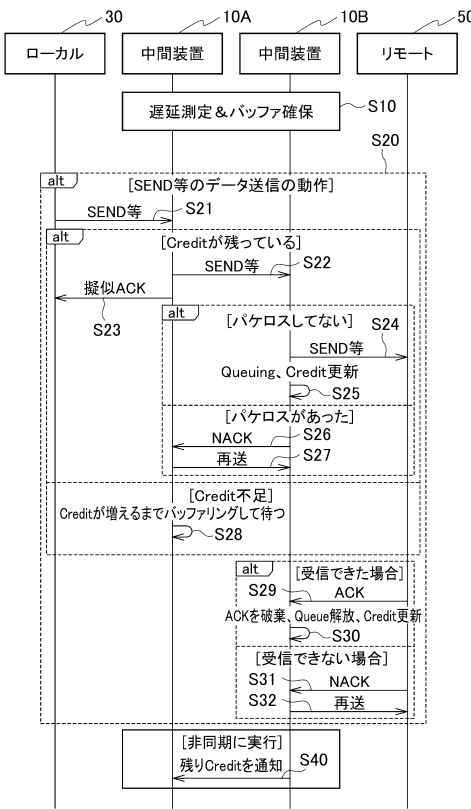
【図 5】



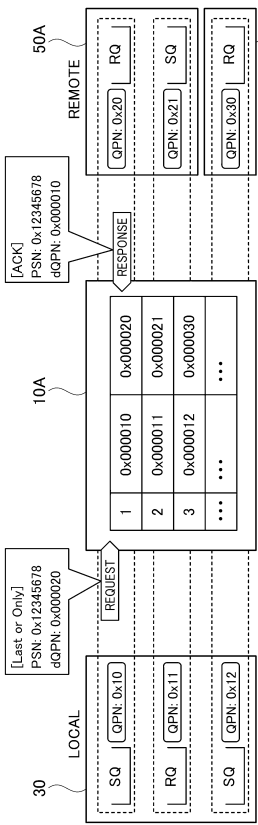
【図 6】



【図 7】



【図 8】



10

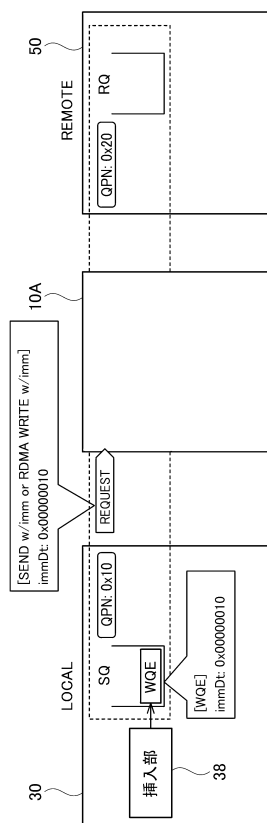
20

30

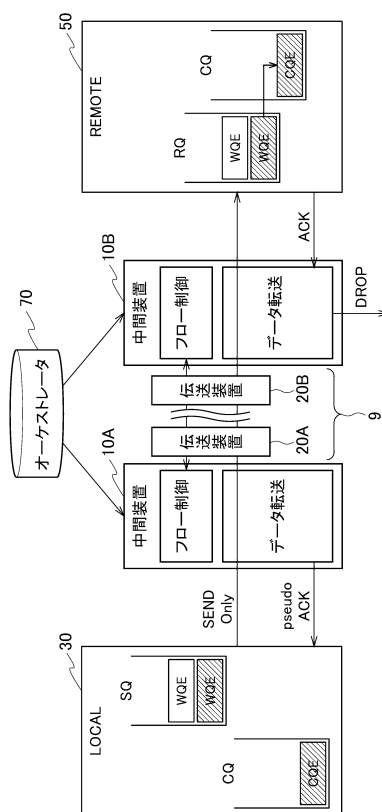
40

50

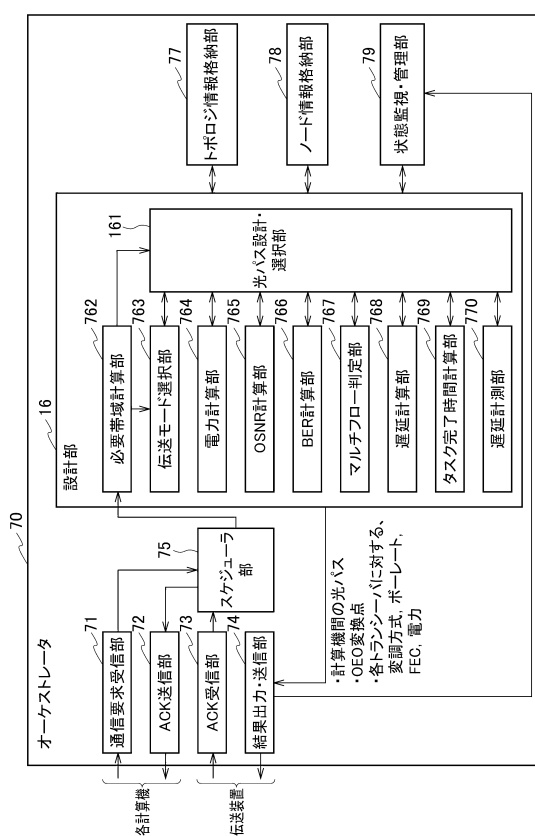
【圖 9】



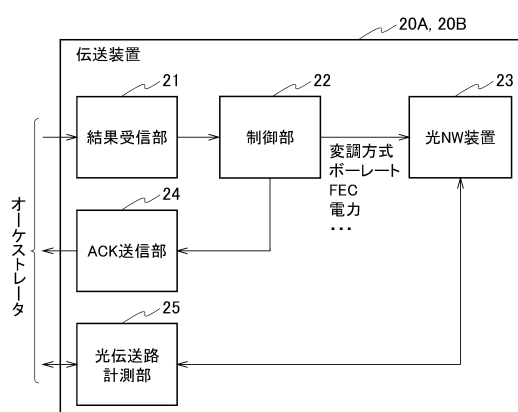
【図 10】



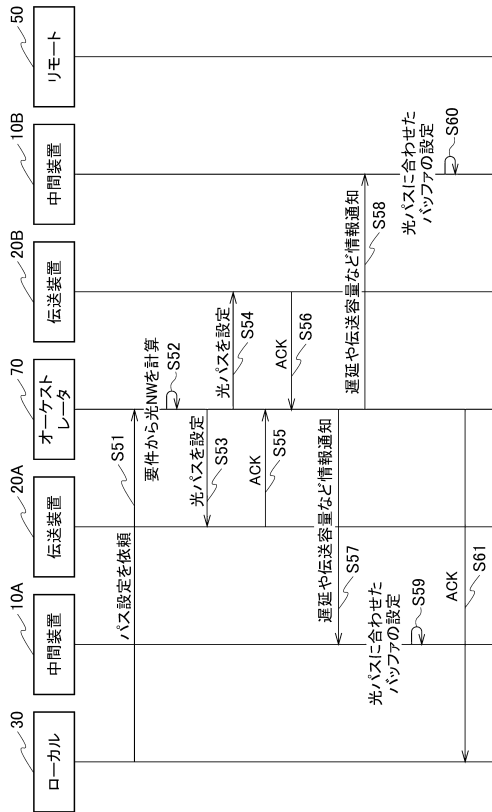
【図 1 1】



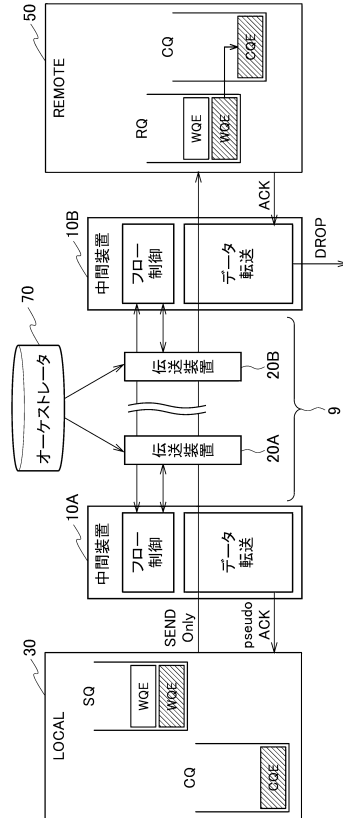
【図 12】



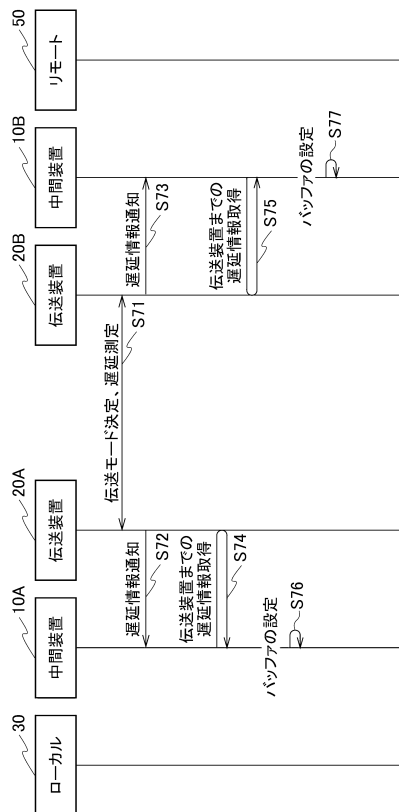
【図 13】



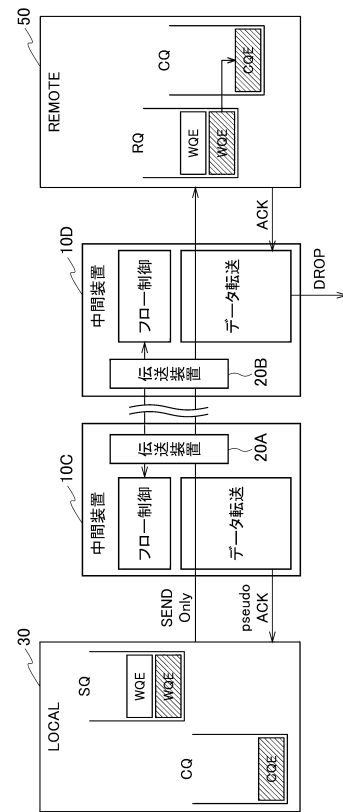
【図 14】



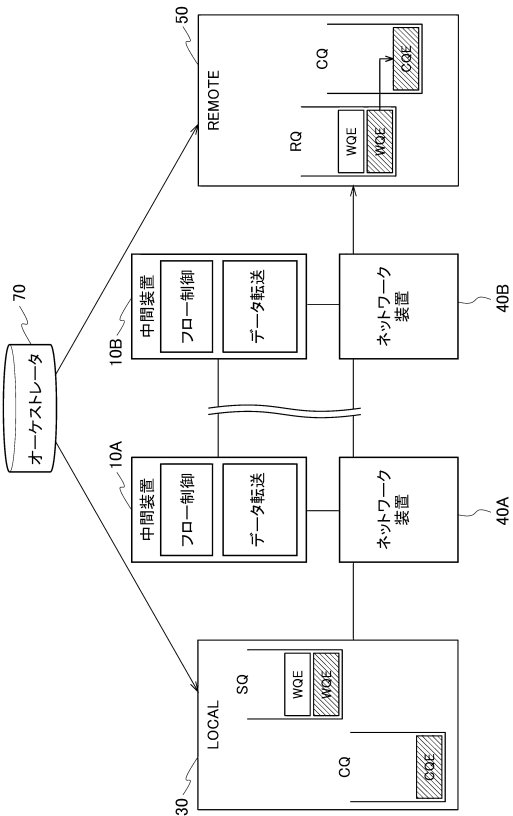
【図 15】



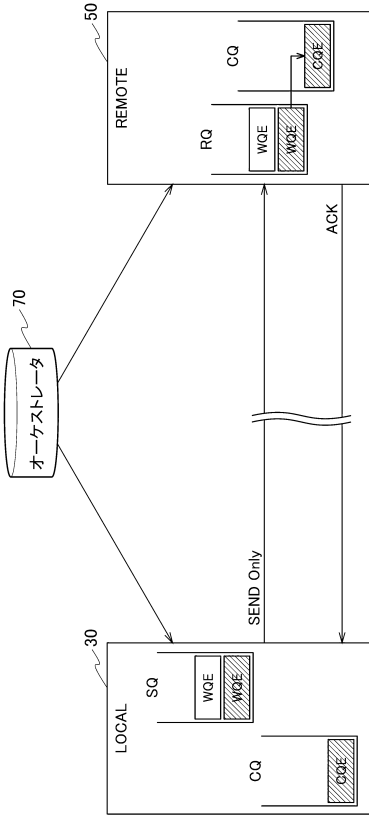
【図 16】



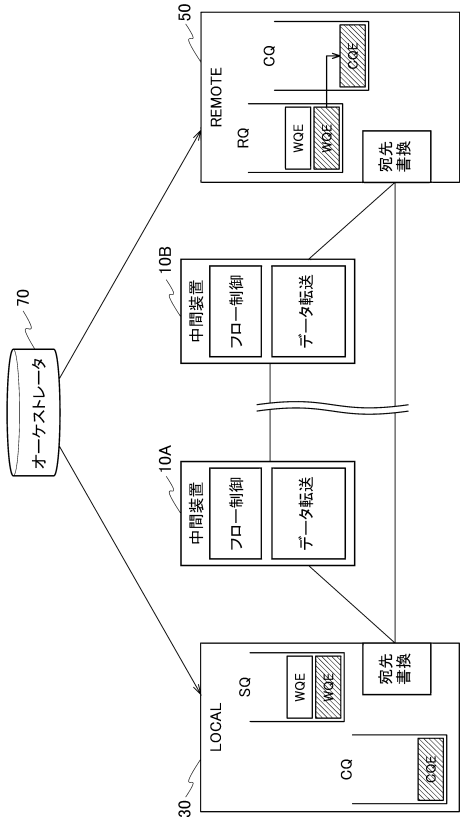
【図 17】



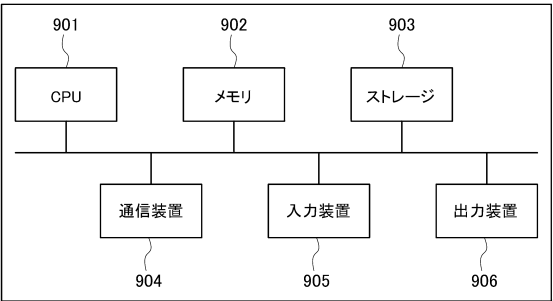
【図 18】



【図 19】



【図 20】



10

20

30

40

50

フロントページの続き

- 日本電信電話株式会社内
- (72)発明者 穴澤 和也
東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
- (72)発明者 築島 幸男
東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
- (72)発明者 清水 健司
東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
- (72)発明者 西沢 秀樹
東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
- (72)発明者 高杉 耕一
東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
- 審査官 田中 秀樹
- (56)参考文献 米国特許出願公開第2020/0089649 (US, A1)
特開2006-287293 (JP, A)
特開2014-057170 (JP, A)
特開2019-92218 (JP, A)
- (58)調査した分野 (Int.Cl., DB名)
H04L 12/00-12/66
H04L 41/00-101/695