

(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

(43) 国際公開日
2023年3月30日(30.03.2023)



(10) 国際公開番号

WO 2023/047567 A1

(51) 国際特許分類:

G06F 15/173 (2006.01)

(21) 国際出願番号: PCT/JP2021/035305

(22) 国際出願日: 2021年9月27日(27.09.2021)

(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: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(72) 発明者: 市川 潤 紀 (ICHIKAWA Junki); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 西沢 秀樹 (NISHIZAWA Hideki); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 清水 健司 (SHIMIZU

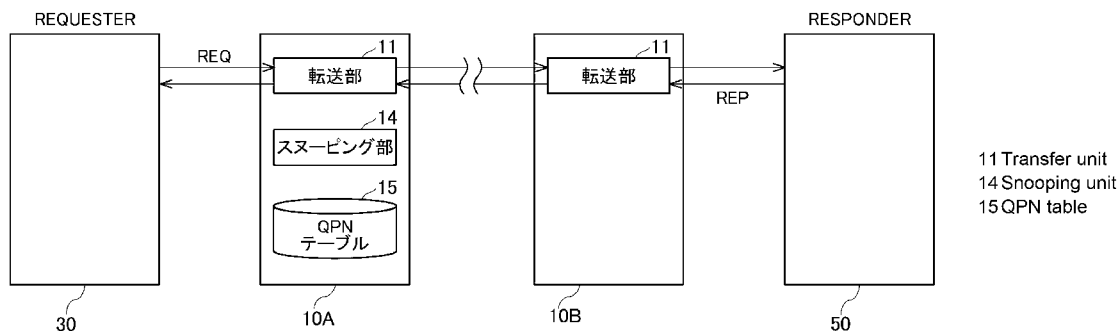
Kenji); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 築島 幸男 (TSUKISHIMA Yukio); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 日比 智也 (HIBI Tomoya); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 井上 綺泉 (INOUE Kiwami); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP).

(74) 代理人: 三好 秀和, 外 (MIYOSHI Hidekazu et al.); 〒1050001 東京都港区虎ノ門1丁目2番8号 虎ノ門琴平タワー 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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ,

(54) Title: INTERMEDIATE DEVICE, COMMUNICATION METHOD, AND PROGRAM

(54) 発明の名称: 中間装置、通信方法、およびプログラム



(57) Abstract: An intermediate device 10A is placed between a requester 30 and responder 50 for transferring data using RDMA. The intermediate device 10A extracts a combination of a QPN of the requester 30 and a QPN of the responder 50 from a packet transmitted/received at the time of establishing a connection, and registers the combination into a QPN table 15. The intermediate device 10A manages an MSN in a WQ table 17, switches the MSN number when receiving a request with a set Last or Only flag from the requester 30, and returns a pseudo-response including the switched MSN.

(57) 要約: 中間装置 10A は、RDMA を用いてデータを転送するリクエスタ 30 とレスポнда 50 の間に配置される中間装置 10A である。中間装置 10A は、コネクションを確立するとき送受信されるパケットからリクエスタ 30 の QPN とレスポнда 50 の QPN の組み合わせを抽出して QPN テーブル 15 に登録する。中間装置 10A は、MSN を WQ テーブル 17 で管理し、リクエスタ 30 から Last または Only のフラグが立ったリクエストを受信したときに MSN 番号を遷移して、遷移後の MSN を含む疑似レスポンスを返却する。

EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH,
KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,
MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,
NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT,
QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, WS, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

- 国際調査報告 (条約第21条(3))

明 細 書

発明の名称： 中間装置、通信方法、およびプログラム

技術分野

[0001] 本発明は、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] RDMA (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。RDMAは、送信端末のメモリ領域から受信端末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。RDMAは、クレジットベースのフロー制御機能を持つほか、タイマーアウトやパケットロス検知に基づく再送制御を行うため、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。

先行技術文献

非特許文献

[0003] 非特許文献1: InfiniBand Architecture Specification Volume 1, Release 1.4, April 7, 2020

発明の概要

発明が解決しようとする課題

[0004] 図1を参照し、RDMAの通信モデルについて説明する。RDMAは、Send Queue (SQ) とReceive Queue (RQ) からなるQueue Pair (QP) を用いてデータを転送する通信モデルである。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。SQ/RQのキューサイズ分、WQEをFIFO

O (F i r s t - I n - F i r s t - O u t) で S Q / R Q に積むことができる。Q P の間で W Q E の処理が正常に完了すると、S Q / R Q の W Q E は削除され、次の W R の受け入れが可能になる。

[0005] 従来技術では、ネットワークが長距離になると、R D M A の転送性能が低下するという問題があった。R D M A では、リクエストの W Q E を完了させて S Q に空きを作らないと次の W R に進むことができない。コネクション型のサービスタイプでは、リクエストの W Q E を完了させるためにはレスポンドから A C K を受信する必要がある。A C K を受信するまで未完了の W Q E が滞留すると、新しく W Q E をキューに積むことができなくなり転送性能が低下してしまう。

[0006] 本発明は、上記に鑑みてなされたものであり、R T T (R o u n d T r i p T i m e) が大きいネットワークサービス上でも高帯域なデータ転送を実現することを目的とする。

課題を解決するための手段

[0007] 本発明の一態様の中間装置は、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る中間装置であって、前記第 1 の装置と前記第 2 の装置との間でパケットを転送する転送部と、前記第 1 の装置と前記第 2 の装置との間でコネクションを確立するときに送受信されるパケットから前記第 1 の装置の宛先情報と前記第 2 の装置の宛先情報の組み合わせを抽出して宛先テーブルに登録する登録部を備える。

[0008] 本発明の一態様の中間装置は、前記第 2 の装置における前記リクエストの完了状態を表すメッセージシーケンス番号を管理テーブルで管理する管理部を備え、前記管理部は、コネクション確立時に、前記第 2 の装置の宛先情報をキーとして、前記第 1 の装置の宛先情報と初期化したメッセージシーケンス番号を前記管理テーブルに登録し、前記第 1 の装置から所定のリクエストを受信したときにメッセージシーケンス番号を遷移し、前記管理テーブルから前記第 1 の装置の宛先情報と遷移後のメッセージシーケンス番号を取得し、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

する生成部を備える。

[0009] 本発明の一態様の通信方法は、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による通信方法であって、前記中間装置が、前記第1の装置と前記第2の装置との間でパケットを転送し、前記第1の装置と前記第2の装置との間でコネクションを確立するときに送受信されるパケットから前記第1の装置の宛先情報と前記第2の装置の宛先情報の組み合わせを抽出して宛先テーブルに登録する。

[0010] 本発明の一態様の通信方法は、前記中間装置が、前記第2の装置における前記リクエストの完了状態を表すメッセージシーケンス番号を管理テーブルで管理し、コネクション確立時に、前記第2の装置の宛先情報をキーとして、前記第1の装置の宛先情報と初期化したメッセージシーケンス番号を管理テーブルに登録し、前記第1の装置から所定のリクエストを受信したときにメッセージシーケンス番号を遷移し、前記管理テーブルから前記第1の装置の宛先情報と遷移後のメッセージシーケンス番号を取得し、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。

発明の効果

[0011] 本発明によれば、RTT (Round Trip Time) が大きいネットワークサービス上でも高帯域なデータ転送を実現できる。

図面の簡単な説明

[0012] [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、本実施形態の中間装置を備えた通信システムの構成の一例を示す図である。

[図3]図3は、Queue Pair Numberテーブルの一例を示す図である。

[図4]図4は、コネクション確立時の処理の流れの一例を示すシーケンス図である。

[図5]図5は、コネクション解放時の処理の流れの一例を示すシーケンス図で

ある。

[図6]図6は、本実施形態の中間装置を備えた通信システムの構成の一例を示す図である。

[図7]図7は、Work Queueテーブルの一例を示す図である。

[図8]図8は、データ転送時の処理の流れの一例を示すシーケンス図である。

[図9]図9は、データ転送時の処理の流れの一例を示すシーケンス図である。

[図10]図10は、中間装置をNIC上に構成した一例を示す図である。

[図11]図11は、中間装置のハードウェア構成の一例を示す図である。

発明を実施するための形態

[0013] 以下、本発明の実施の形態について図面を用いて説明する。

[0014] [Queue Pair Number (QPN) の管理]

図2を参照し、リクエスタ30のQPNとレスポнда50のQPNのペアを管理する中間装置10Aを備えた通信システムの一例について説明する。QPNとは、QPのエンドポイントのそれぞれに付与された番号である。S/Q/RQは対向のQPNを認識しており、リクエストおよびレスポンスなどのRDMAパケットを生成する際はdestination QPNをヘッダに含める。

[0015] 図2は、本実施形態の中間装置10A、10Bを備えた通信システムの構成の一例を示す図である。中間装置10A、10Bは、RDMAを用いてデータを転送するリクエスタ30とレスポнда50の間に配置される。より具体的には、中間装置10Aは、リクエスタ30側の長距離ネットワークの前段に配置され、中間装置10Bは、レスポнда50側の長距離ネットワークの前段に配置される。なお、リクエスタ30はローカル、レスポнда50はリモートと呼ばれることもある。

[0016] 中間装置10Aは、転送部11、スヌーピング部14、およびQueue Pair Number (QPN) テーブル15を備える。

[0017] 転送部11は、リクエスタ30からのリクエストをレスポнда50へ転送し、レスポнда50からのレスポンスをリクエスタ30へ転送する。リクエ

スタ３０とレスポнда５０との間で送受信されるリクエストとレスポンスには、コネクション確立時のREQ、REP、およびRTU、データ転送時のリクエストとレスポンス、コネクション解放時のDREQとDREPがある。

[0018] スヌーピング部１４は、RDMAのCommunication Management (RDMA-CM) のコネクション確立フェーズにおいてリクエスト３０とレスポнда５０との間で送受信されるRDMA-CMのパケットを傍受し、リクエスト３０のQPNとレスポнда５０のQPNのペアを持つ双方向のQPNエントリをQPNテーブル１５に登録する。

[0019] リクエスト３０とレスポнда５０は、コネクションにおける通信を一意に識別するための識別子としてCommunication ID (CID) を設定する。このCIDは、コネクションが破棄されるまで変化しない。リクエスト３０とレスポнда５０それぞれのCIDに紐づいてリクエスト３０とレスポнда５０のそれぞれのQPNも一意に識別される。CIDおよびQPNはコネクション確立フェーズにおいてリクエスト３０とレスポнда５０の間で交換され、CIDペア、QPNペアが設定される。コネクション確立フェーズは、Connect Request (REQ)、Connect Reply (REP)、およびReady To Use (RTU) の３-way ハンドシェイクによって構成される。

[0020] また、スヌーピング部１４は、RDMA-CMのコネクション解放フェーズにおいてリクエスト３０とレスポнда５０との間で送受信されるRDMA-CMのパケットを傍受して、QPNテーブル１５から対応するQPNエントリを削除する。

[0021] コネクション確立フェーズにおいて設定されたCIDおよびQPNのペアは、RDMA-CMのコネクション解放フェーズで解放される。コネクション解放フェーズは、Disconnect Request (DREQ) およびDisconnect Reply (DREP) のハンドシェイクによって構成される。

[0022] QPNテーブル15は、CIDをキーとして、ローカル側（リクエスタ側）QPNとリモート側（レスポнда側）QPNのペアを関連付けて管理する。図3に、QPNテーブル15の一例を示す。リクエスタ30とレスポнда50との間にコネクションが確立されると、リクエスタ30の観点でコネクションを一意に識別するためのLocal CIDをキーとするQPNエントリと、レスポнда50の観点でコネクションを一意に識別するためのLocal CIDをキーとするQPNエントリがQPNテーブル15に作成される。具体的には、リクエスタ30観点のLocal CIDをキーとするQPNエントリには、Local QPNにリクエスタ30側のQPに割り当てられたQPNが登録され、Remote QPNにレスポнда50側のQPに割り当てられたQPNが登録される。レスポнда50観点のLocal CIDをキーとするQPNエントリには、Local QPNにQPのレスポнда50側に割り当てられたQPNが登録され、Remote QPNにQPのリクエスタ30側に割り当てられたQPNが登録される。

[0023] 中間装置10Bは、転送部11を備える。

[0024] 転送部11は、中間装置10Aの転送部11と同様に、リクエスタ30からのリクエストをレスポнда50へ転送し、レスポнда50からのレスポンスをリクエスタ30へ転送する。

[0025] 次に、図4のシーケンス図を参照し、コネクション確立時の処理の流れの一例について説明する。図4では、中間装置10A、10Bを中間装置10として図示している。

[0026] ステップS11にて、リクエスタ30は、レスポнда50へREQを送信する。REQは、Local CIDおよびLocal QPNを含む。REQの含むLocal CIDは、リクエスタ30がコネクションを識別するための識別子である。REQの含むLocal QPNは、リクエスタ30のQPに付与されたQPNである。

[0027] 中間装置10は、REQをレスポнда50へ転送する際、REQの含むLocal CIDをキーとしたQPNエントリをQPNテーブル15に作成

し、そのQPNエントリのLocal QPNにREQの含むLocal QPN（リクエスト30のQPN）を登録する。

[0028] レスポンダ50はREQを受信すると、ステップS12にて、リクエスト30へREPを送信する。REPは、Local CID、Remote CID、およびLocal QPNを含む。REPの含むLocal CIDは、レスポンダ50がコネクションを識別するための識別子である。REPの含むRemote CIDは、リクエスト30がコネクションを識別するための識別子であり、REQの含むLocal CIDと同じである。REPの含むLocal QPNは、レスポンダ50に付与されたQPNである。

[0029] 中間装置10は、REPをリクエスト30へ転送する際、REPの含むRemote CIDをキーとするQPNエントリをQPNテーブル15から検索し、そのQPNエントリのRemote QPNにREPの含むLocal QPN（レスポンダ50のQPN）を登録する。また、中間装置10は、REPの含むLocal CIDをキーとした逆方向のQPNエントリをQPNテーブル15に作成する。具体的には、中間装置10は、REPの含むLocal CIDをキーとし、Local QPNにレスポンダ50のQPNを登録し、Remote QPNにリクエスト30のQPNを登録したQPNエントリをQPNテーブル15に作成する。

[0030] リクエスト30はREPを受信すると、ステップS13にて、レスポンダ50へRTUを送信する。RTUは、Local CIDとRemote CIDを含む。

[0031] 以上の処理により、リクエスト30とレスポンダ50との間にコネクションが確立されるとともに、リクエスト30のQPNとレスポンダ50のQPNをペアとする双方向のQPNエントリがQPNテーブル15に作成される。コネクション確立後は、リクエスト30とレスポンダ50との間でデータ転送が行われる。中間装置10は、リクエスト30とレスポンダ50の間でパケットを転送するとともに、疑似レスポンスを生成して送出する。

[0032] 次に、図5のシーケンス図を参照し、コネクション解放時の処理の流れの一例について説明する。図5では、中間装置10A、10Bを中間装置10として図示している。

[0033] ステップS31にて、リクエスト30は、レスポнда50へDREQを送信する。DREQは、Local CID、Remote CID、およびRemote QPNを含む。DREQの含むLocal CIDは、リクエスト30がコネクションを識別するための識別子である。DREQの含むRemote CIDは、レスポнда50がコネクションを識別するための識別子である。DREQの含むRemote QPNは、レスポнда50に付与されたQPNである。

[0034] 中間装置10は、DREQをレスポнда50へ転送する。

[0035] レスポнда50はDREQを受信すると、ステップS32にて、リクエスト30へDREPを送信する。DREPは、Local CIDとRemote CIDを含む。DREPの含むLocal CIDは、レスポнда50がコネクションを識別するための識別子である。DREPの含むRemote CIDは、リクエスト30がコネクションを識別するための識別子である。

[0036] 中間装置10は、DREPをレスポнда50へ転送する際、DREPの含むLocal CIDをキーとするQPNエントリを検索してQPNテーブル15から削除するとともに、DREPの含むRemote CIDをキーとするQPNエントリを検索してQPNテーブル15から削除する。

[0037] 以上の処理により、リクエスト30とレスポнда50との間のコネクションが解放されるとともに、当該コネクションに対応するQPNエントリがQPNテーブル15から削除される。

[0038] [Message Sequence Number (MSN) の管理]

図6を参照し、MSNを管理する中間装置10Aを備えた通信システムの一例について説明する。MSNとは、サービスタイプがReliable Connection (RC) の通信において、リクエスト30によるリク

エストがレスポнда50によってどの程度完了したのかを通知する数値である。MSNは、ACKのACK Extender Transport Header (AETH) に記述されて、リクエスト30に通知される。リクエスト30のWQEには、このMSNと一対一で対応するSend Sequence Number (SSN) が設定されている。リクエスト30は、ACK受信時にMSNに記述された値までのSSNのWQEを解放する。SSNとMSNは、RDMA-CMのコネクション確立フェーズにおいてSSN=1、MSN=0にリセットされる。レスポнда50は、受信したリクエストのヘッダ情報からMSNの状態を管理する。なお、MSNはメッセージ単位のシーケンス番号であり、Packet Sequence Number (PSN) はパケット単位のシーケンス番号である。

[0039] 図6は、本実施形態の中間装置10A、10Bを備えた通信システムの構成の一例を示す図である。図2に示した通信システムと同様に、中間装置10A、10Bは、RDMAを用いてデータを転送するリクエスト30とレスポнда50の間に配置される。

[0040] 中間装置10Aは、転送部11、生成部12、トレーシング部16、およびWork Queue (WQ) テーブル17を備える。中間装置10Aは、図2に示したスヌーピング部14とQPNテーブル15を備えてもよい。

[0041] 転送部11は、リクエスト30からのリクエストをレスポнда50へ転送し、レスポнда50からのレスポンスをリクエスト30へ転送する。

[0042] 生成部12は、データ転送時に、リクエスト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、リクエストに対する疑似レスポンス (pseudo-Response) を生成し、生成した疑似レスポンスをリクエスト30へ返却する。疑似レスポンスを生成する際、生成部12は、OnlyまたはLastのリクエストと同じPSNの値を用いる。また、生成部12は、後述のWQテーブル17を参照し、疑似レスポンスに含めるdestination QPNおよびMSNを決定する。

- [0043] リクエスト30は、疑似レスポンスを受信すると、レスポнда50からのレスポンスと認識して、疑似レスポンスのMSNに記述された値までのSSNのWQEを解放する。
- [0044] トレーシング部16は、コネクション確立時に、レスポнда50のQPNをキーとして、リクエスト30のQPNとMSNを持つWQエントリを後述のWQテーブル17へ登録する。この時、トレーシング部16は、WQエントリのMSNを0にリセットしておく。
- [0045] トレーシング部16は、QPNテーブル15におけるQPNエントリの作成を契機として、WQエントリを作成してWQテーブル17へ登録してもよい。この時、QPNエントリのRemote QPNをキーとして、Local QPNとMSNを持つWQエントリを作成する。
- [0046] トレーシング部16は、リクエスト30から受信したリクエストのヘッダ情報に基づいてメッセージ単位を識別し、MSNの値を遷移させる。具体的には、トレーシング部16は、OnlyまたはLastのフラグが立ったリクエストを受信したときに、対応するWQエントリのMSNの値を遷移させて、レスポнда50のMSNの状態を模擬する。
- [0047] トレーシング部16は、WQテーブル17のWQエントリをコネクション解放後に削除してもよい。例えば、図2のスヌーピング部14がQPNエントリを削除するときに、トレーシング部16は、削除されるQPNエントリに対応するQPNをキーとするWQエントリを削除する。
- [0048] WQテーブル17は、レスポнда50のQPN(dst QPN)をキーとして、リクエスト30のQPN(src QPN)とMSNを管理する。図7に、WQテーブル17の一例を示す。WQテーブル17には、リクエストを受信するレスポнда50のQPNをキーとして、リクエストを送信するリクエスト30のQPNとMSNを持つWQエントリが登録される。
- [0049] 生成部12は、リクエストに対する疑似レスポンスを生成する際、リクエストの宛先のQPNに一致するキーを持つWQエントリを検索し、そのWQエントリのMSNの値を持つ疑似レスポンスを生成する。疑似レスポンスの

`destination QPN`にはWQエントリの`src QPN`を設定する。

[0050] 中間装置10Bは、転送部11と破棄部13を備える。

[0051] 転送部11は、中間装置10Aの転送部11と同様に、リクエスタ30からのリクエストをレスポнда50へ転送し、レスポнда50からのレスポンスをリクエスタ30へ転送する。

[0052] 破棄部13は、データ転送時のリクエスタ30からのリクエストに対する、レスポнда50からの真のレスポンスを破棄する。これにより、リクエスタ30でのレスポンスの重複受信を防止できる。破棄部13は、レスポнда50からリクエスタ30へ返却されるメッセージのうちリクエスタ30が誤動作を引き起こす可能性があるメッセージを破棄してもよい。

[0053] なお、中間装置10Aが破棄部13を備えて、中間装置10Bを配置しなくてもよい。

[0054] 次に、図8のシーケンス図を参照し、データ転送時の処理の流れの一例について説明する。

[0055] ステップS21にて、リクエスタ30は、SQにWQEを積み、レスポнда50へリクエストを送信する。リクエストは、転送するデータを含む。リクエストは、中間装置10A、10Bを介してレスポнда50へ転送される。

[0056] ステップS22にて、中間装置10Aは、WQテーブル17を参照して`destination QPN`を決定し、リクエストに対する疑似レスポンスを生成する。

[0057] ステップS23にて、中間装置10Aは、疑似レスポンスをリクエスタ30へ返却する。リクエスタ30は、疑似レスポンスを受信すると、SQのWQEを解放する。

[0058] 以降、リクエスタ30が、SQにWQEを積み、レスポнда50へリクエストを送信すると（ステップS26）、中間装置10Aは、リクエストを転送すると疑似レスポンスを生成してリクエスタ30に返却する（ステップS

27, S28)。

[0059] 一方、レスポнда50は、リクエストの受信が成功すると、ステップS24にて、レスポンスをリクエスタ30へ送信する。

[0060] ステップS25にて、中間装置10Bは、受信したレスポンスを破棄する。

[0061] 以降、レスポнда50はリクエストを受信するとレスポンスを返却し、中間装置10Bはレスポンスを破棄する。

[0062] 次に、図9のシーケンス図を参照し、データ転送時の処理の流れの一例について説明する。ここでは、MSNの管理の観点で処理を説明する。図9では、中間装置10A, 10Bを中間装置10として図示している。

[0063] ステップS41からステップS43にて、リクエスタ30は、レスポнда50へデータ転送のリクエストを送信する。いずれのリクエストもSSN=1が関連付けられたWQEに関するリクエストであるとする。図中のrはリクエストを表し、rに続く数字はPSNを表す。r17のリクエストは、Lastフラグが立ったリクエストである。中間装置10は、リクエストをレスポнда50へ転送し、レスポнда50はリクエストを受信する。

[0064] 中間装置10は、Lastフラグが立ったr17のリクエストを受信すると、ステップS44にて、対応するWQEエントリのMSNの値を遷移させ、遷移後のMSNの値を持つ疑似レスポンスを生成し、リクエスタ30へ送信する。図中のp-aは疑似レスポンスを表し、p-aに続く数字はPSNを表す。カッコ内の数字はMSNを表す。図9の例では、中間装置10は、WQEエントリのMSNの値を0から1に遷移させて、PSN=17、MSN=1とした疑似レスポンスを生成し、疑似レスポンスをリクエスタ30へ送信した。

[0065] リクエスタ30は、疑似レスポンスを受信すると、疑似レスポンスのMSNの値までのSSNを持つWQEを解放する。図9の例では、疑似レスポンスのMSNは1であるので、リクエスタ30は、1以下のSSNを持つWQEを解放した。

- [0066] 一方、レスポнда50は、リクエストの受信が成功すると、ステップS45からステップS47にて、レスポンスをリクエスタ30へ送信する。図中のaはレスポンスを表し、aに続く数字はPSNを表す。カッコ内の数字はMSNを表す。r15とr16のリクエストに対するレスポンスはa15とa16であり、いずれもMSN=0である。r17のリクエストに対するレスポンスはa17であり、MSN=1である。
- [0067] 中間装置10は、レスポнда50から受信したレスポンスをリクエスタ30へ転送せずに破棄する。
- [0068] 以上説明したように、本実施形態の中間装置10Aは、RDMAを用いてデータを転送するリクエスタ30とレスポнда50の間に配置される中間装置10Aである。中間装置10Aは、リクエスタ30とレスポнда50との間でコネクションを確立するときに送受信されるパケットからリクエスタ30のQPNとレスポнда50のQPNの組み合わせを抽出してQPNテーブル15に登録する。これにより、中間装置10Aは、コネクション確立時に疑似レスポンスの返却先を特定できる。
- [0069] 本実施形態の中間装置10Aは、レスポнда50におけるリクエストの完了状態を表すMSNをWQテーブル17で管理し、コネクション確立時に、レスポнда50のQPNをキーとして、リクエスタ30のQPNと0に初期化したMSNをWQテーブル17に登録し、リクエスタ30からLastまたはOnlyのフラグが立ったリクエストを受信したときにMSN番号を遷移して、遷移後のMSNと、リクエスタ30のQPNから、リクエストに対する疑似レスポンスを生成してリクエスタ30へ返却する。リクエスタ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、リクエスタ30とレスポнда50の間のRTTが大きい場合でも、リクエスタ30はレスポнда50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。リクエスタ30からレスポнда50へのPUSH型のデータ転送、特にRDMA Writeのデータ転送の長距離高速化を実現できる。

[0070] 以上では、リクエスタ30とレスポнда50の間に中間装置10A, 10Bを設置する構成で説明したが、図10に示すように、リクエスタ30の装置のNIC (Network Interface Card) 上に中間装置10Aを構成し、レスポнда50の装置のNIC上に中間装置10Bを構成してもよい。また、中間装置10A, 10Bは、物理サーバで構成してもよいし、仮想サーバで構成してもよい。スイッチまたはルータなどのネットワーク装置が中間装置10A, 10Bの機能を備えてもよい。

[0071] 上記説明した中間装置10A, 10Bには、例えば、図11に示すような、中央演算処理装置(CPU)901と、メモリ902と、ストレージ903と、通信装置904と、入力装置905と、出力装置906とを備える汎用的なコンピュータシステムを用いることができる。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、中間装置10A, 10Bが実現される。このプログラムは磁気ディスク、光ディスク、半導体メモリなどのコンピュータ読み取り可能な記録媒体に記録することも、ネットワークを介して配信することもできる。

符号の説明

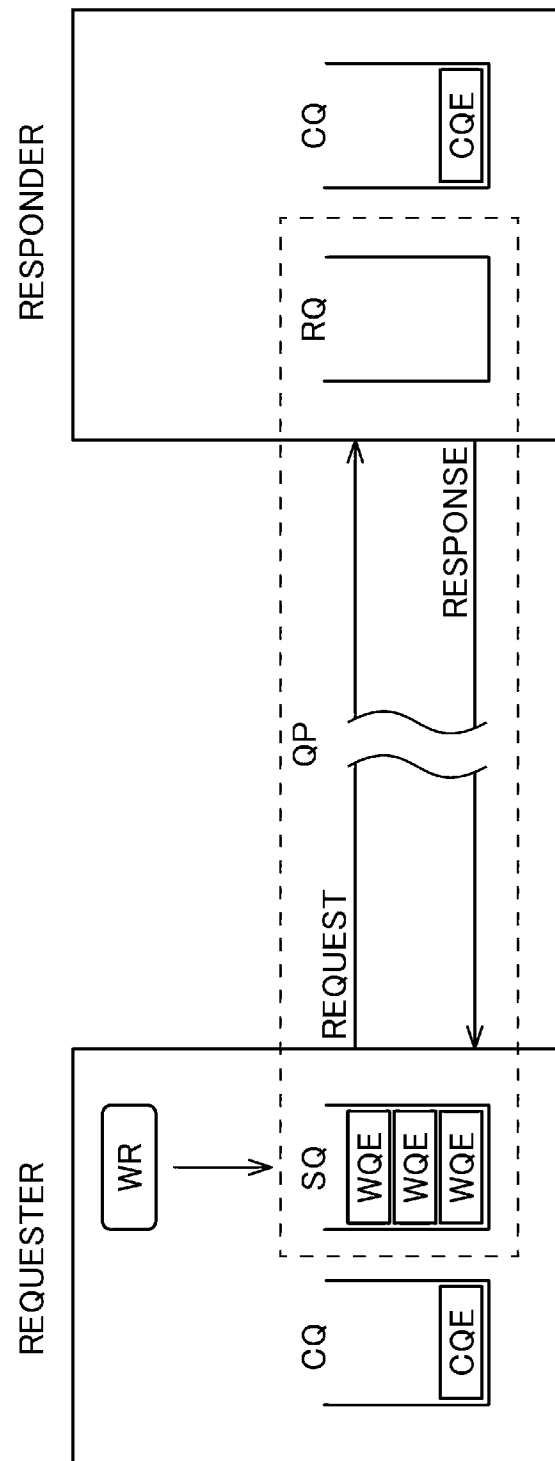
[0072] 10, 10A, 10B…中間装置
11…転送部
12…生成部
13…破棄部
14…スヌーピング部
15…QPNテーブル
16…トレーシング部
17…WQテーブル
30…リクエスタ
50…レスポнда

請求の範囲

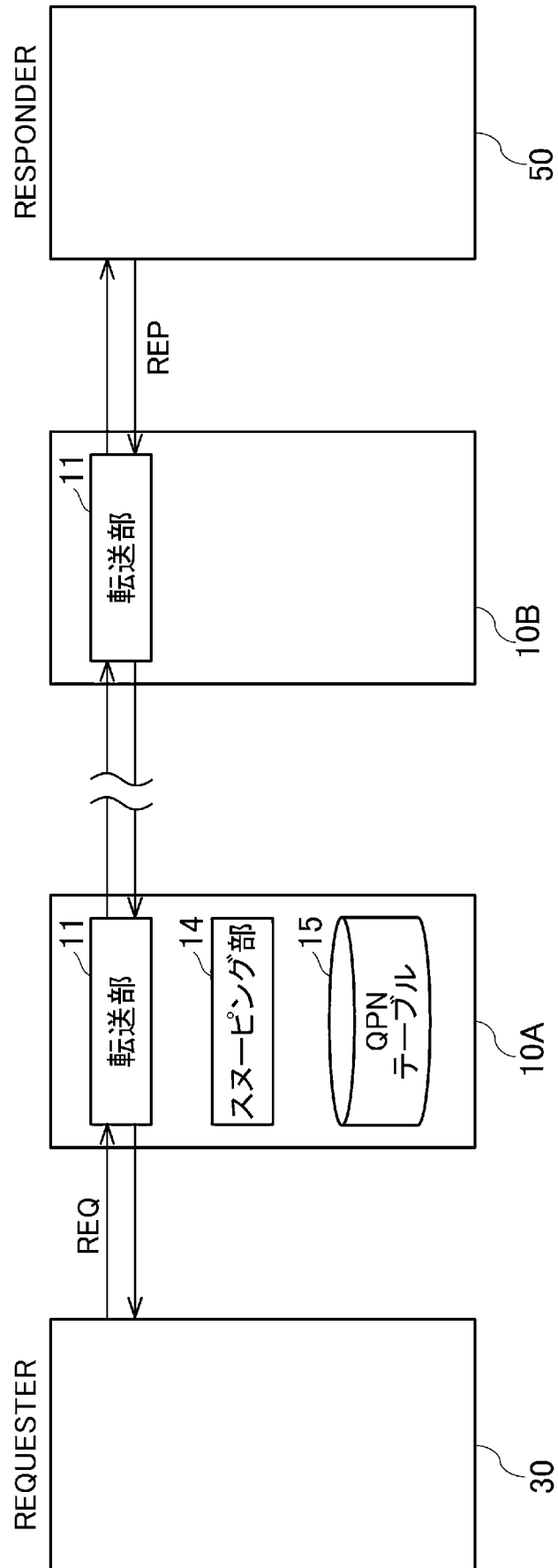
- [請求項1] リモートダイレクトメモリアクセスを用いてデータを転送する第1の装置と第2の装置の間に配置される中間装置であって、
- 前記第1の装置と前記第2の装置との間でパケットを転送する転送部と、
- 前記第1の装置と前記第2の装置との間でコネクションを確立するときに送受信されるパケットから前記第1の装置の宛先情報と前記第2の装置の宛先情報の組み合わせを抽出して宛先テーブルに登録する登録部と、を備える
- 中間装置。
- [請求項2] 請求項1に記載の中間装置であって、
- 前記登録部は、前記第1の装置および前記第2の装置のそれぞれが前記コネクションを識別するためのコネクション識別子に基づいて前記第1の装置の宛先情報と前記第2の装置の宛先情報の組み合わせを特定し、前記コネクション識別子をキーとして前記組み合わせを前記宛先テーブルに登録する
- 中間装置。
- [請求項3] 請求項1または2に記載の中間装置であって、
- 前記第2の装置におけるリクエストの完了状態を表すメッセージシーケンス番号を管理テーブルで管理する管理部を備え、
- 前記管理部は、コネクション確立時に、前記第2の装置の宛先情報をキーとして、前記第1の装置の宛先情報と初期化したメッセージシーケンス番号を前記管理テーブルに登録し、前記第1の装置から所定のリクエストを受信したときにメッセージシーケンス番号を遷移し、
- 前記管理テーブルから前記第1の装置の宛先情報と遷移後のメッセージシーケンス番号を取得し、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を備える
- 中間装置。

- [請求項4] 請求項3に記載の中間装置であって、
前記管理部は、L a s t または O n l y のフラグが立ったリクエストを受信したときに前記管理テーブルで管理されるメッセージシーケンス番号を遷移する
中間装置。
- [請求項5] 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による通信方法であって、
前記中間装置が、
前記第1の装置と前記第2の装置との間でパケットを転送し、
前記第1の装置と前記第2の装置との間でコネクションを確立するときに送受信されるパケットから前記第1の装置の宛先情報と前記第2の装置の宛先情報の組み合わせを抽出して宛先テーブルに登録する通信方法。
- [請求項6] 請求項5に記載の通信方法であって、
前記中間装置が、
前記第2の装置におけるリクエストの完了状態を表すメッセージシーケンス番号を管理テーブルで管理し、
コネクション確立時に、前記第2の装置の宛先情報をキーとして、前記第1の装置の宛先情報と初期化したメッセージシーケンス番号を管理テーブルに登録し、前記第1の装置から所定のリクエストを受信したときにメッセージシーケンス番号を遷移し、
前記管理テーブルから前記第1の装置の宛先情報と遷移後のメッセージシーケンス番号を取得し、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
通信方法。
- [請求項7] 請求項1ないし4のいずれかに記載の中間装置の各部としてコンピュータを動作させるプログラム。

[図1]



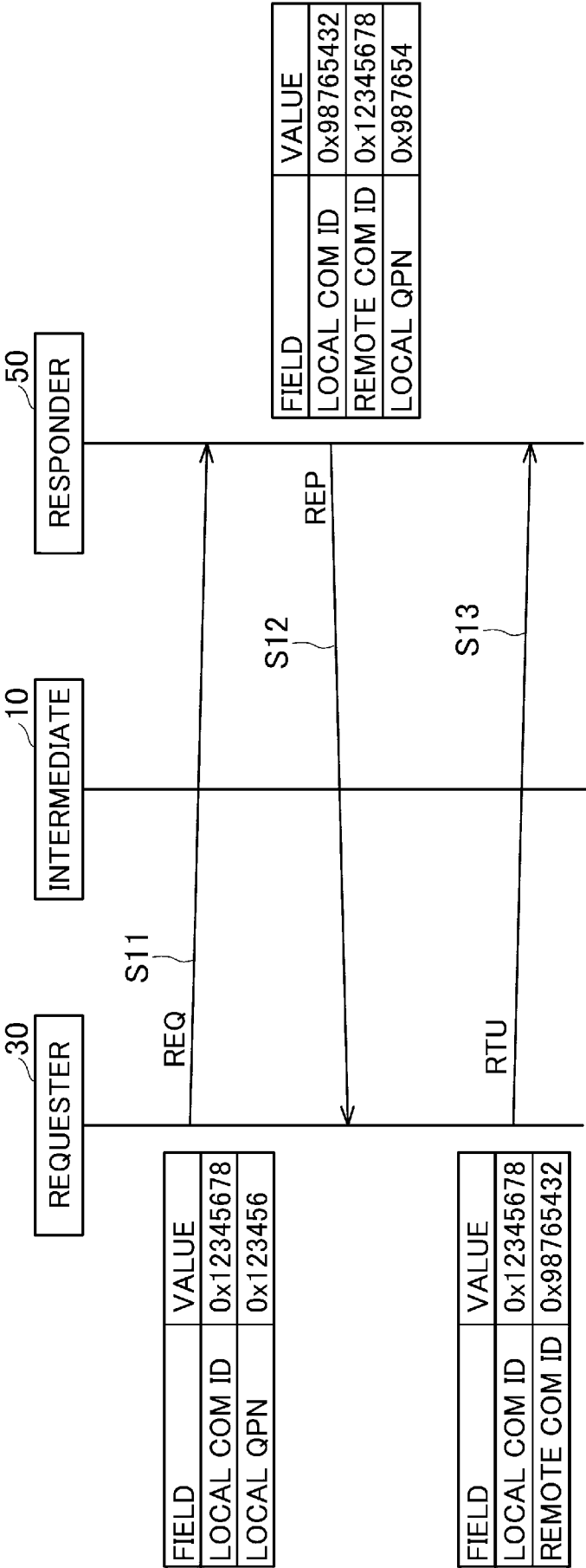
[図2]



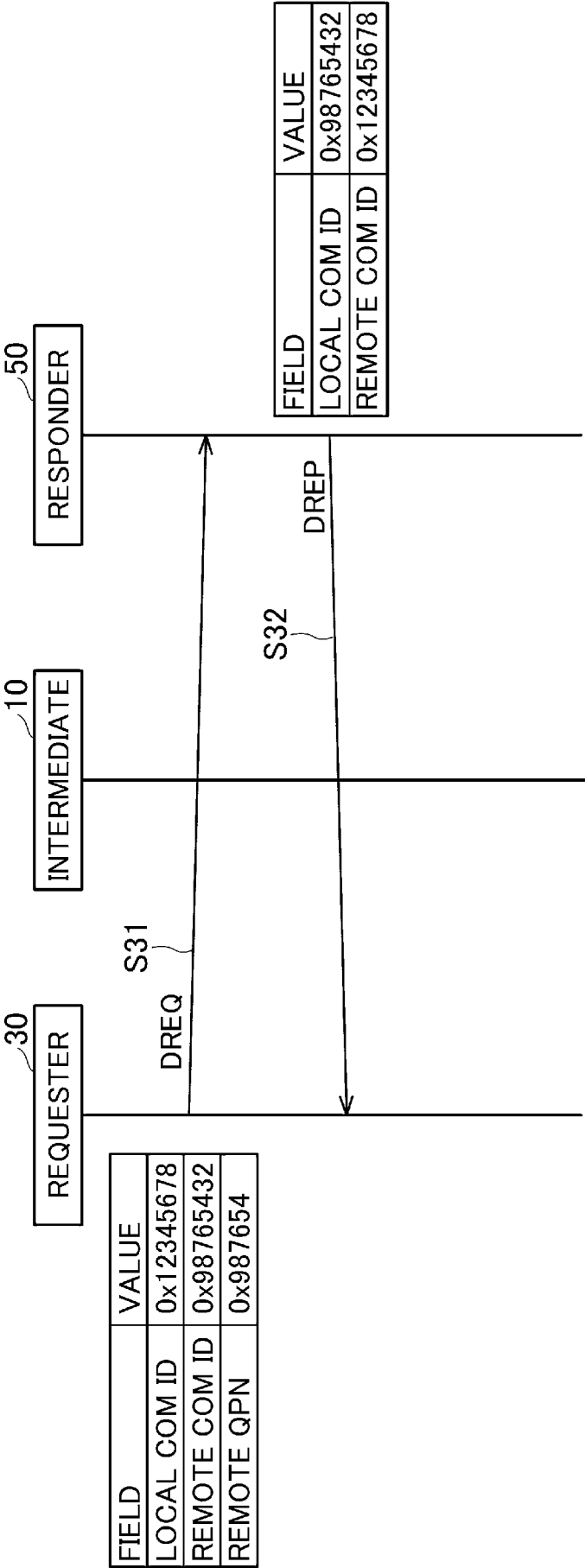
[図3]

key (Com ID)	Local QPN	Remote QPN
0x12345678	0x123456	0x987654
0x98765432	0x987654	0x123456
...	...	...

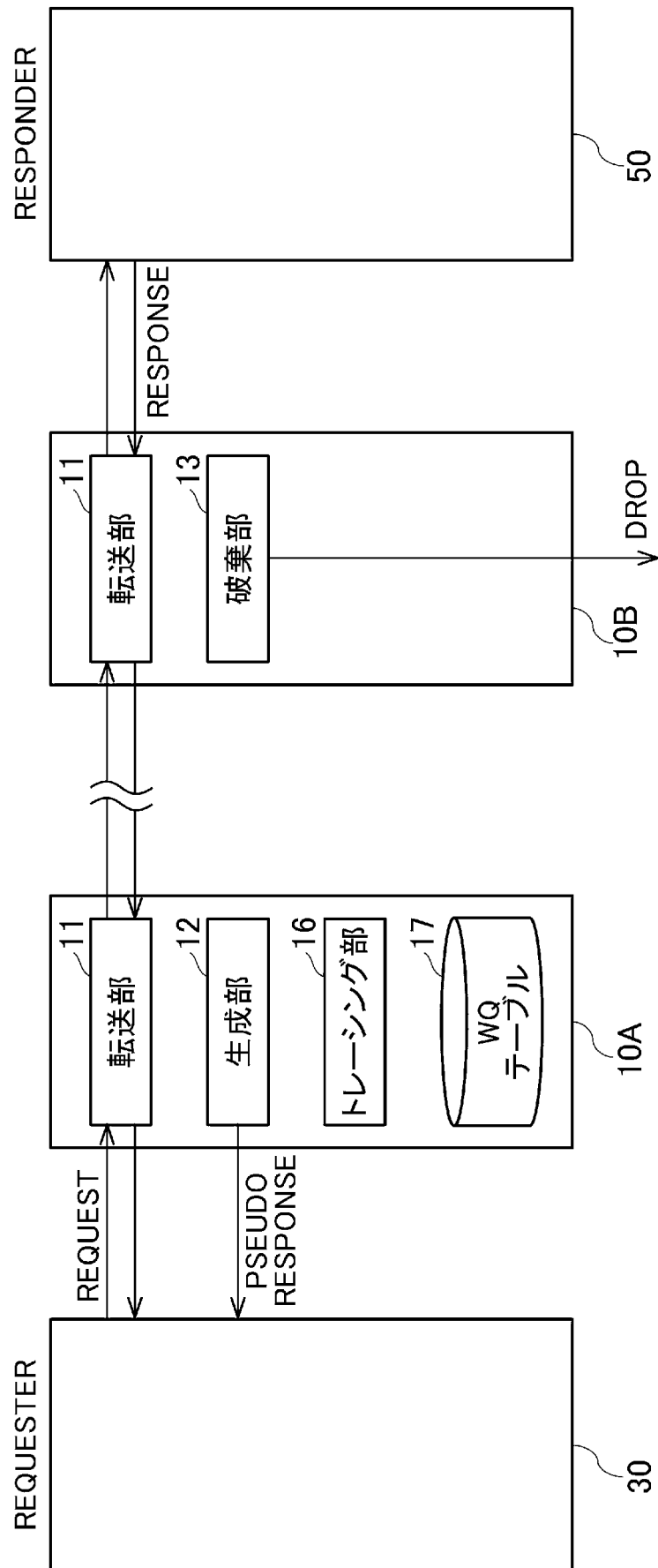
[図4]



[図5]



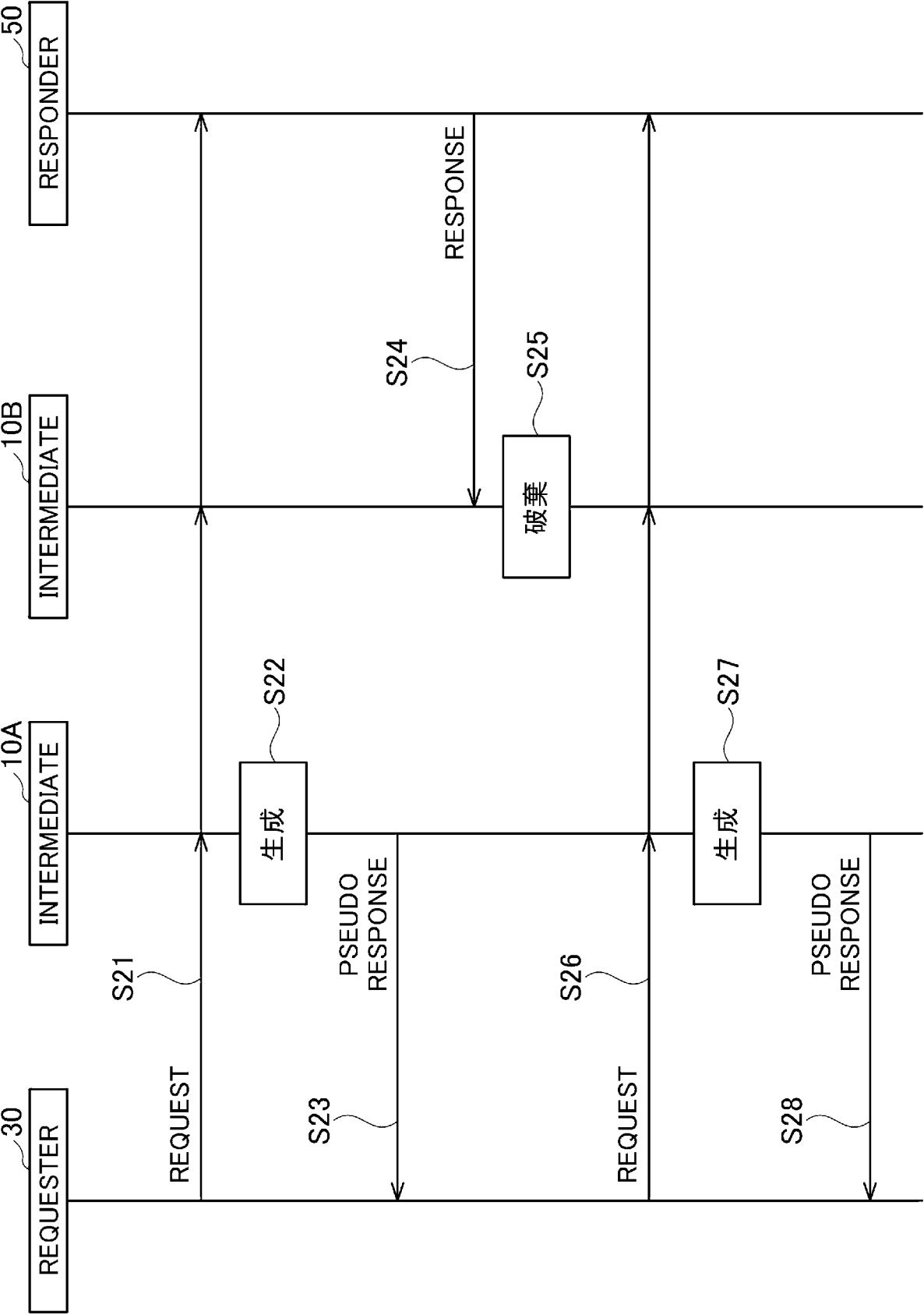
[図6]



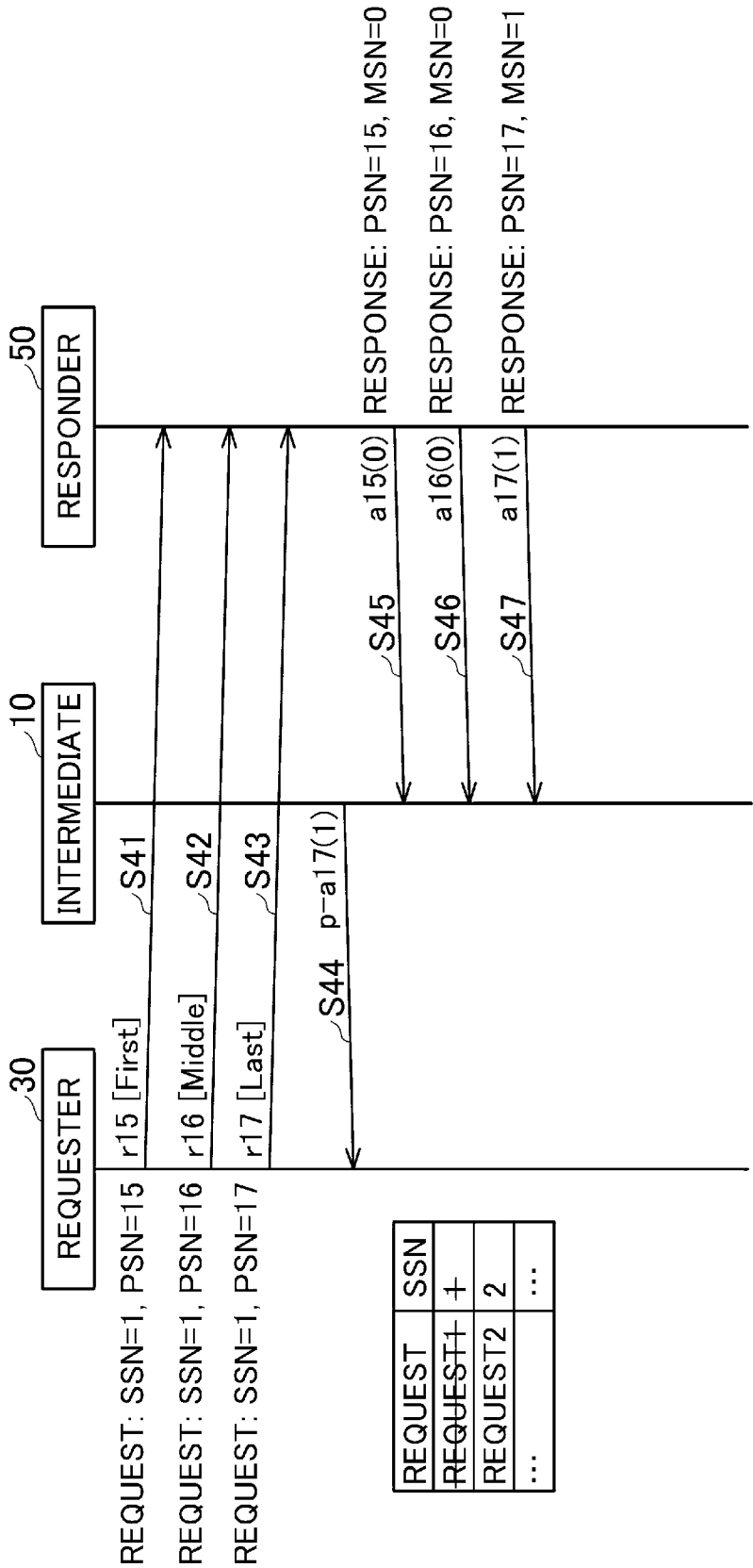
[図7]

key (dst QPN)	src QPN	MSN
0x987654	0x123456	0
...	...	...

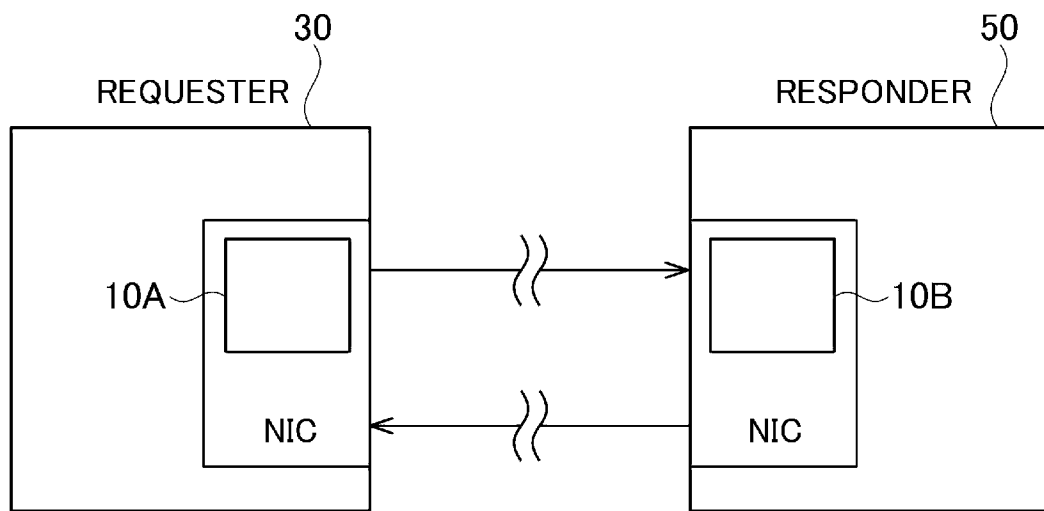
[図8]



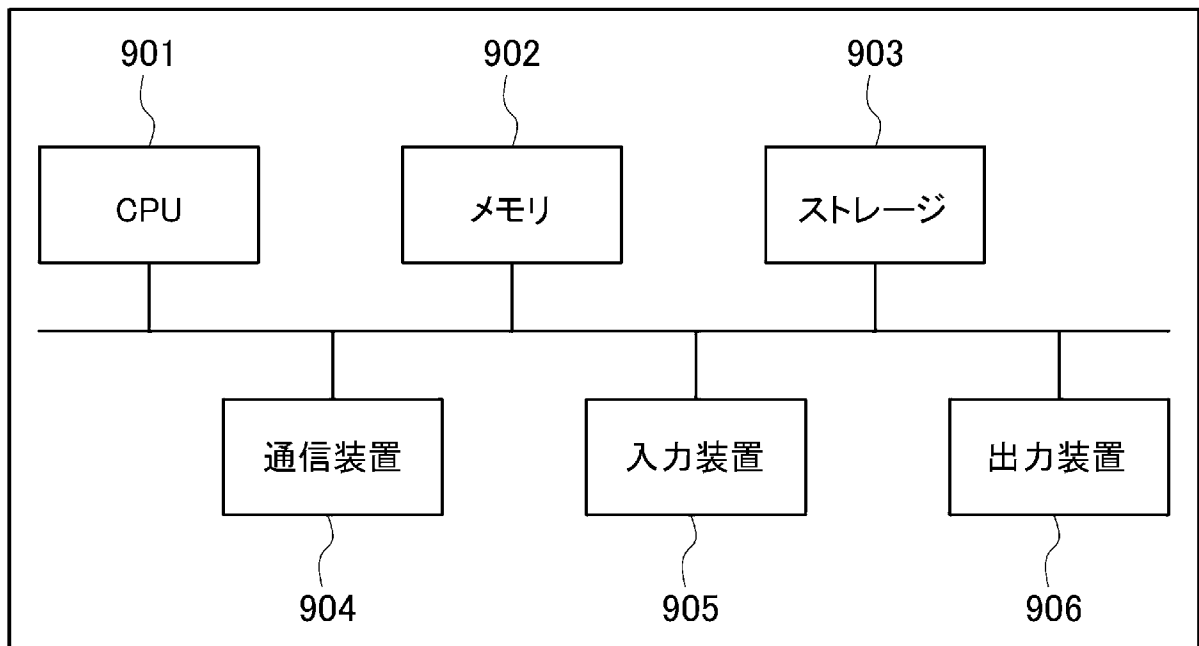
[図9]



[図10]



[図11]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2021/035305

A. CLASSIFICATION OF SUBJECT MATTER G06F 15/173 (2006.01)i FI: G06F15/173 665 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 Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Published examined utility model applications of Japan 1922-1996 Published unexamined utility model applications of Japan 1971-2021 Registered utility model specifications of Japan 1996-2021 Published registered utility model applications of Japan 1994-2021 Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)																	
C. DOCUMENTS CONSIDERED TO BE RELEVANT <table border="1"> <thead> <tr> <th>Category*</th> <th>Citation of document, with indication, where appropriate, of the relevant passages</th> <th>Relevant to claim No.</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>JP 2007-235674 A (NEC CORP.) 13 September 2007 (2007-09-13) paragraphs [0004]-[0008], [0049], fig. 8</td> <td>1-2, 5, 7</td> </tr> <tr> <td>Y</td> <td></td> <td>3-4, 6</td> </tr> <tr> <td>Y</td> <td>空閑洋平 他, 物理マシンと協調動作可能なソフトウェアによるPCIeデバイスエミュレーション手法, 情発処理学会 研究報告 システムソフトウェアとオペレーティング・システム(OS) 17 July 2019, 2019-OS-147, pages 1-6, ISSN: 2188-8795, (IPSI SIG Technical Report, System software and operating system(OS).), non-official translation (KUGA, Yohei et al. PCIe device emulation method by software that can cooperate with physical machine.) chapter 3, fig. 2(c)</td> <td>3-4, 6</td> </tr> <tr> <td>Y</td> <td>JP 2010-61194 A (HITACHI, LTD.) 18 March 2010 (2010-03-18) paragraphs [0098]-[0100]</td> <td>4</td> </tr> </tbody> </table>			Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.	X	JP 2007-235674 A (NEC CORP.) 13 September 2007 (2007-09-13) paragraphs [0004]-[0008], [0049], fig. 8	1-2, 5, 7	Y		3-4, 6	Y	空閑洋平 他, 物理マシンと協調動作可能なソフトウェアによるPCIeデバイスエミュレーション手法, 情発処理学会 研究報告 システムソフトウェアとオペレーティング・システム(OS) 17 July 2019, 2019-OS-147, pages 1-6, ISSN: 2188-8795, (IPSI SIG Technical Report, System software and operating system(OS).), non-official translation (KUGA, Yohei et al. PCIe device emulation method by software that can cooperate with physical machine.) chapter 3, fig. 2(c)	3-4, 6	Y	JP 2010-61194 A (HITACHI, LTD.) 18 March 2010 (2010-03-18) paragraphs [0098]-[0100]	4
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.															
X	JP 2007-235674 A (NEC CORP.) 13 September 2007 (2007-09-13) paragraphs [0004]-[0008], [0049], fig. 8	1-2, 5, 7															
Y		3-4, 6															
Y	空閑洋平 他, 物理マシンと協調動作可能なソフトウェアによるPCIeデバイスエミュレーション手法, 情発処理学会 研究報告 システムソフトウェアとオペレーティング・システム(OS) 17 July 2019, 2019-OS-147, pages 1-6, ISSN: 2188-8795, (IPSI SIG Technical Report, System software and operating system(OS).), non-official translation (KUGA, Yohei et al. PCIe device emulation method by software that can cooperate with physical machine.) chapter 3, fig. 2(c)	3-4, 6															
Y	JP 2010-61194 A (HITACHI, LTD.) 18 March 2010 (2010-03-18) paragraphs [0098]-[0100]	4															
☐ 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 02 December 2021		Date of mailing of the international search report 14 December 2021															
Name and mailing address of the ISA/JP Japan Patent Office (ISA/JP) 3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915 Japan		Authorized officer Telephone No.															

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2021/035305

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP	2007-235674	A	13 September 2007	US 2007/0206598 A1 paragraphs [0006]-[0010], fig. 1	
JP	2010-61194	A	18 March 2010	US 2010/0064070 A1 paragraphs [0121]-[0125], fig. 9	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06F 15/173(2006.01)i FI: G06F15/173 665		
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（IPC）） G06F15/173		
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 1922-1996年 日本国公開実用新案公報 1971-2021年 日本国実用新案登録公報 1996-2021年 日本国登録実用新案公報 1994-2021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 請求項の番号
X	JP 2007-235674 A（日本電気株式会社）13.09.2007（2007-09-13） 段落0004-0008, 0049, 図8	1-2, 5, 7
Y		3-4, 6
Y	空閑 洋平 他, 物理マシンと協調動作可能なソフトウェアによるP C I e デバイス エミュレーション手法, 情報処理学会 研究報告 システムソフトウェアとオペレー ティング・システム（OS）, 2019.07.17, 2019-OS-147, 第1頁-第6 頁, ISSN: 2188-8795 第3章, 図2(c)	3-4, 6
Y	JP 2010-61194 A（日立製作所）18.03.2010（2010-03-18） 段落0098-0100	4
☐ 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抵 触するものではなく、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 02.12.2021	国際調査報告の発送日 14.12.2021	
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 三坂 敏夫 5B 4178 電話番号 03-3581-1101 内線 3545	

国際調査報告
パテントファミリーに関する情報

国際出願番号

PCT/JP2021/035305

引用文献			公表日	パテントファミリー文献	公表日
JP	2007-235674	A	13.09.2007	US 2007/0206598 A1 Paragraphs0006-0010, Figure1	
JP	2010-61194	A	18.03.2010	US 2010/0064070 A1 Paragraphs0121-0125, Figure9	