様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １１ 日現在

機関番号：１４４０１
研究種目：基盤研究(B)（一般）
研究期間：2016～2018
課題番号：１６Ｈ０２８０２

研究課題名（和文）計算機資源の動的再構成機能を有するベアメタルクラウド構築手法の確立

研究課題名（英文）A method for building baremetal cloud with dynamic reconfigucation funtion of computing resources

研究代表者
下條　真司（Shimojo, Shinji）

大阪大学・サイバーメディアセンター・教授

研究者番号：００１８７４７８

交付決定額（研究期間全体）：（直接経費）　　13,000,000円

研究成果の概要（和文）：本研究では、ユーザが記載する計算機資源要求を受け付け、それら資源要求間で有限の計算機資源の調整・スケジューリングを行い、アクセラレータ等の計算機資源を動的に計算ノードに割り付けることで、それぞれの計算機資源要求および性能要求に合致したベアメタルクラウドをオンデマンドに構築・提供する技術の実現を目的とした。その結果、利用者の計算ニーズに応じてGPU演算加速器を計算ノードに動的に割り当て計算可能にする技術を実現し、利用者の多様な計算ニーズに対応可能なベアメタルクラウド構築技術の設計・実装が可能となった。また、民間クラウド資源の動的な統合を可能にする技術のプロトタイプ実装に成功した。

研究成果の学術的意義や社会的意義
本研究では、多様化する計算ニーズを収容可能とする計算インフラ技術の実現を目的として研究開発を推進した。結果、計算機にあわせてユーザが計算を行うのではなく、ユーザの計算ニーズにあわせ計算インフラが計算環境を提供することを可能とする基盤技術を実現できた。これにより、大規模計算を必要とする学術研究に携わる研究者の研究効率向上に貢献できる。

研究成果の概要（英文）：This research aimed to realize a technology that build and construct a bare-metal cloud environment in an on-demand manner based on users' request. Specifically, it receives a resource request from users and then attach such devices as accelerators to compute nodes. As the result of this reserach, we have succeeded to design and implement the bare-metal cloud building technology that can accommodate a variety of users' computing needs. It attaches GPU resources to compute nodes dynamically in response to user computing needs. Also, the technology is extended to integrate cloud resources.

研究分野：情報ネットワーク

キーワード：ベアメタル　資源管理　クラウド構成技術

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

プロセッサ、アクセラレータ技術の急速な発展は、高性能計算技法・手法をますます複雑化・多様化する傾向にある。例えば、今日では、マルチコア化が急速に進展する CPU に加え、NVIDIA 社製 Tesla 等の GPU などのアクセラレータ、Intel 社製 Xeon Phi に代表されるメニーコア型プロセッサを選択・併用することにより、数百ギガフロップス (Giga Flops) 級の理論性能値を有する計算機システムを容易に構成できる。さらに、低遅延・広帯域なネットワーク(インターコネクト)でそれら複数台 (ノード)を接続することで、数百テラフロップス級 (Tera Flops) 級の高性能クラスタシステムをも実現可能である。しかし、そのような高性能な計算機システム上でハードウェア性能を引き出すためには、MPI(Message Passing Interface) を用いたノード間での分散並列処理、OpenMP を用いたノード内分散並列処理、GPU や Xeon Phi などのアクセラレータ、SSD などの高速 2 次記憶装置などによる分散処理などの高性能計算技法を駆使する必要性がある。そのような背景から、今日では、ユーザの計算手法は様々に異なり、計算機資源に対する要求もまた著しく多様化する傾向にある。

ユーザの計算機資源要求にあわせたシステム構成を提供可能とする IaaS (Infrastructure as a Service) 型クラウド上で、科学計算をはじめとする高性能計算を試みる研究開発が今日までに数多く報告されている。しかし、それら先行研究の多くは、中核技術として仮想計算機技術 (Virtual Machine Technology)を用いた IaaS (Infrastructure as a Service) 型クラウドであり、プロセッサやアクセラレータ等が密結合された高性能計算機のハードウェア性能を最大限にひきだそうとするユーザの性能要求を満たすことは未だ実現できていない。また、今日では、仮想計算機技術を用いないベアメタル(Bare Metal)型クラウドに関する応用事例も報告されつつあるが、例えば、ノードに GPU を 3 枚ずつ割り付け、10 ノードから構成するクラスタを利用したい、といったユーザの資源要求に合致したベアメタルクラウドをオンデマンドに構成する技術は未だ実現されていない。

２．研究の目的

本申請研究は、上述の背景と申請者らの実システム運用経験から着想を得、ユーザの多様なシステム構成要求に応じ、フレキシブルな計算機資源の動的再構成機能を有するベアメタルクラウドの実現を目的とした実践的研究開発を推進する。より具体的には、計算機の内部バス PCIExpress をイーサネット上で拡張することで計算機資源を高い自由度で拡張可能とする ExpEther 技術を応用し、ユーザのシステム構成ニーズに応じて GPU、SSD、HDD などの計算機資源を任意の組み合わせで構築・提供する再構成可能なベアメタルクラウド構築手法および技術を確立することを目的とした。

３．研究の方法

申請時点までに、申請者らは、プロセッサ資源、メモリ資源といった計算機資源情報のみに基づき資源割当を行う従来のジョブ管理システムの問題点に着眼し、高度先進ネットワーキング技術 Software-Defined Networking によってもたらされるネットワークプログラミング性を統合することで、ユーザの資源要求に最適な計算機資源とネットワーク資源を動的に割り当てるジョブ管理フレームワーク SDN-JMS を提案・実装してきた。本申請研究では、SDN-JMS の資源割当機構を中核技術とし、さらに、クラウド管理ツール OpenStack および ExpEther 技術を加えることでベアメタルクラウド構築を実現する。

本申請研究で提案したベアメタルクラウドは、ユーザが記載する計算機資源要求スクリプトを受け付け、それら資源要求間で有限の計算機資源の調整・スケジューリングを行い、GPU、SSD、HDD 等の計算機資源を計算ノードに割り付け、ユーザ占有型ベアメタルクラウドを提供する。本構想の実現のために、本申請課題では、以下の 3 点のマイルストーン課題を設定して研究開発を推進した。

[課題 1] 計算機資源要求に基づく動的再構成フレームワークの設計と実装

本課題では、(a)ユーザからの資源要求を受信・解析する資源要求インタフェースおよび資源要求解析パーサ、(b)ユーザ資源要求間の調整・スケジューリング機能、および(c)再構成可能計算機資源の状態を管理し、システム仮想化技術 ExpEther との連携により計算機資源の割当制御を行う機能から構成される、ベアメタルクラウドの中核技術となる動的再構成フレームワークの設計と実装を行う。

[課題 2] ユーザ資源要求に基づく計算機資源の多次元スケジューリング手法の考案

本課題では、[課題 1]で実現する(b)ユーザ資源要求間の調整・スケジューリング機能のためのアルゴリズムを設計・実装する。具体的には、プロセッサ資源、GPU 資源、SSD 資源およびネットワーク資源をユーザ資源要求間で調整・スケジューリングし、クラウドを構成するこれらの計算機資源の利用率、ユーザジョブの計算性能およびスループットに着眼し、多次元スケジュー

リング手法の設計・実装を行う。

[課題 3] 実環境での性能評価と有用性・運用性の検証
本課題では、図 1 に示すベアメタルクラウドを実際に構築し、当該環境上で計算機資源の動的再構成に伴うオーバヘッド等の計測を通じて有用性・運用性を検証する。また、実際高性能計算を行う科学者や研究者に当該環境を公開することにより、実際の計算機資源要求に対する[課題 2]で考案する多次元スケジューリング手法についての有用性・実用性について検証する。これらにより、本研究の成果をより実効性の高いものへと高めていく。

4．研究成果

本研究の主要 2 点の成果を以下にまとめる。

(1) ジョブスケジューラ連動型計算機環境配備システム
本研究では、計算機資源の利用率向上およびジョブ実行待ち時間の短縮のため、従来は静的である各計算ノードが備える計算資源量やミドルウェアといった計算環境を、ジョブの要求に応じて動的に再構成して割り当てるジョブスケジューラ連動型計算機環境配備システムを実現した。また、検証実験を行うために計算環境の再構成アルゴリズム、および本システムの機能を備えたジョブスケジューリングシミュレータを設計・実装し、計算資源の利用率およびジョブの平均待ち時間について従来の資源割当と比較することで本システムの有用性を示した。

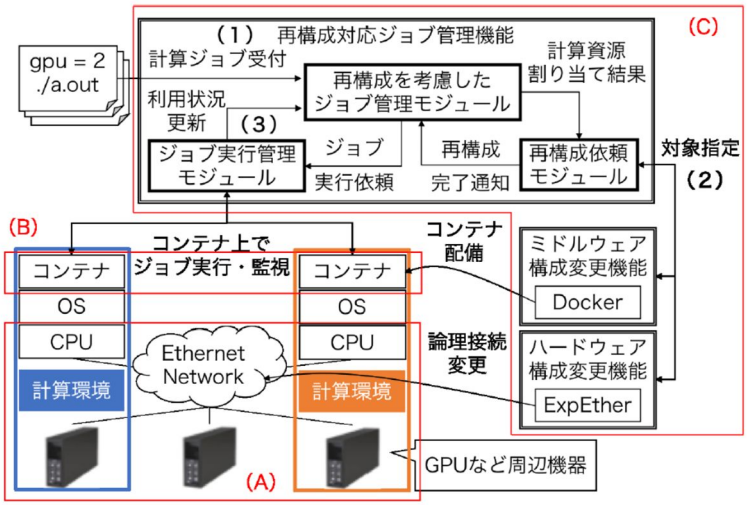
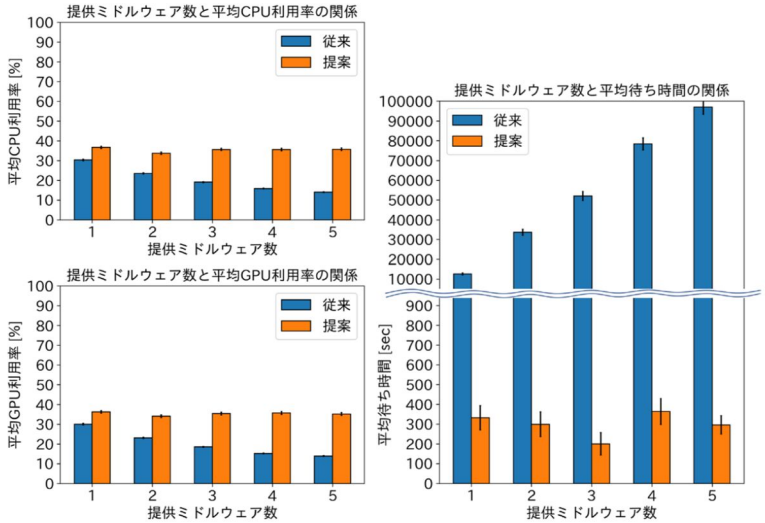


図 1：ジョブスケジューラ連動型計算機環境配備システムの構成

図 1 にジョブスケジューラ連動型計算機環境配備システムの構成を示す。本システムは、図 1 に示す (A) 従来のサーバ単位より細かい、GPU デバイス単位で再構成可能なシステム・ハードウェア、(B) ハードウェア構成との依存性を解消し、目的とするアプリケーションを効率よく実行するための計算実行環境（ミドルウェア）を動的に配備するコンテナ、および (C) ハードウェア、ミドルウェアの動的再構成を考慮しつつジョブの実行管理を可能とするジョブスケジューラで構成される。本システムの動作は以下になる。(1) 投入されたジョブのジョブスクリプトに記述された計算環境要求、資源割当時点での割当済み計算機資源の計算環境の構成、および利用可能な計算機資源の 3 つの情報に基づき、後述する再構成アルゴリズムが投入されているジョブの中から次に実行するジョブを決定する。(2) (1)の決定に基づき、再構成依頼モジュールがハードウェアとコンテナの再構成を行い、その完了をジョブ実行管理モジュールに通知する。(3) ジョブ実行管理モジュールがジョブに記述された処理をコンテナ上で実行し、その実行状態を監視する。上記 (1) の再構成アルゴリズムについて、本研究では各ジョブ実行後に毎回計算資源を配備・解放する All-Reconf と、無作為に選ぶ First-Fit、並びに再構成回数が最小になるように選ぶ Best-Fit の 3 種類の再構成アルゴリズムを考案して実装した。
図 2 に実装したジョブスケジューラ連動型計算機環境配備システムの機能を備えたジョブスケジューリングシミュレータによる平均 CPU 利用率，平均 GPU 利用率，平均待ち時間の比較結果を示す。本実験における計算機の構成は、100 台の計算ノードに 200 個の GPU と 5 種類のミドルウェアを配備した GPU クラスタシステムとした。実験で投入するジョブセットは、計算資源の組み合わせを変化させ、GPU とミドルウェアの要求資源がランダムなジョブで構成される。また、再構成のアルゴリズムは今回作成した Best-Fit を適用し、再構成を行うための性能劣化を 2%、再構成に必要な時間を 30 秒とした。結果より、提供ミドルウェア数が大きい場合、提案システムによって、従来の再構成を行わないサーバ単位の管理を行う場合と比較して CPU およ

び GPU 利用率は最大 22%向上し、平均待ち時間は最大 99.8%低下した。

図 2：平均 CPU 利用率，平均 GPU 利用率，平均待ち時間の比較結果



本研究で設計実装したジョブスケジューラ連動型計算機環境配備システムにより、従来の資源割当で生じていた使用されない GPU のジョブへの割当や、ジョブの計算環境要求に適さないため使用されない状態となっていた計算ノードを大幅に削減することができ、計算資源の利用効率の向上およびジョブの待ち時間の削減を達成することができた。また、本システムでは構成の異なる計算ノードの追加や新たなミドルウェアが追加されても柔軟に対応が可能なことから、システムのスケーラビリティを高めることができる。

(2) クラウドバースティング機能の実現

上述の(1)をさらに発展させ、本研究では、民間クラウドベンダの提供する IaaS (Infrastructure as a Service)型クラウドを利用し、スケジューラと連動させることで、オンプレミス環境にある計算資源の負荷に応じて、クラウド環境上の計算資源を動的に統合した環境を実現するクラウドバースティング機能を実現した。さらに、申請代表者の所属する大阪大学サイバーメディアセンター設置のスーパーコンピュータ OCTOPUS に、当該機能を実戦配備し、実際の利用者ジョブを扱った実証実験・検証を行った。

図 3: OCTOPUS のクラウドバースティング機能拡張.

図 3 に OCTOPUS に実装されたクラウドバースティング機能の概要を示す。OCTOPUS ではスーパーコンピュータシステムに投入される利用者からのジョブ要求を負荷分散・スケジュールするために NEC 製スケジューラシステム NQS が配備されている。本研究では、当該スケジューラを拡張し、OCTOPUS の高負荷状態時に、管理者の判断に応じて、マイクロソフト社の提供する IaaS 型クラウドサービス Azure 上に配備した仮想計算機を動的に起動し、OCTOPUS の計算ノードとして収容する技術を実現した。

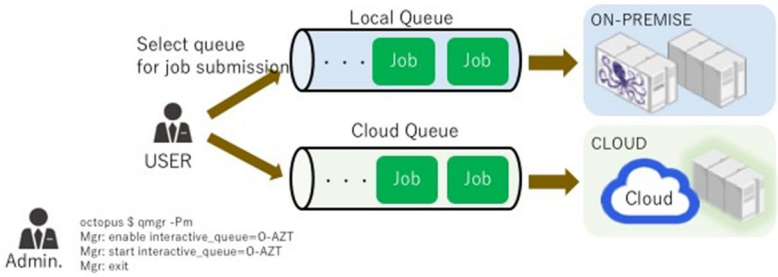


図 4: クラウドバースティング機能のプロトタイプ実装.

図 4 に本報告書執筆時点におけるクラウドバースティング 機能のプロトタイプ実装の詳細を示す。現時点でプロトタイプ化されたクラウドバースティング機能では、管理者がその判断によってクラウドバースティング機能の利用可否を破断する。例えば、オンプレミス環境側のスーパーコンピュータシステムでの高負荷状態が発生し、利用者の待ち時間が長くなる状況が継続する場合に、管理者はクラウド資源へのジョブ投入用キューを利用可能にする。その後、クラウドキューに利用者からのジョブが要求されると、クラウドバースティング機能はクラウド上の仮想計算機を起動し、ジョブを実行できる状態にする。一方、クラウドキューへのジョブ投入が一定時間失われると、当該機能はクラウド上の仮想計算機を停止状態にする。これにより、クラウド上での計算資源の利用を最小化する設計・実装としている。

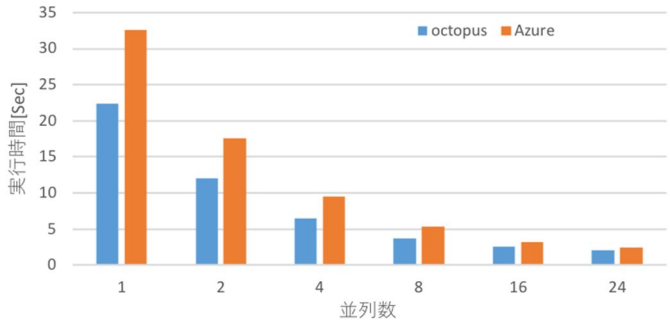


図 3: 性能評価.

図 5 に当該機能が用いたクラウド仮想計算機および OCTOPUS 計算ノード上でのアプリケーション実行による性能評価結果を示す。なお、Azure 上では、Intel Platinmu 8168 processor から構成される 48vCPU、96GiB メモリを有する Standard F48s_v2 仮想計算機を採用している。この結果から、現状では OCTOPUS よりも性能のよいプロセッサを用いても、OCTOPUS のほうが性能がよいことがわかる。

本研究で設計実装したクラウドバースティング機能は、研究者の研究ニーズに応じて動的に計算資源を統合させることで、利用者ニーズを満たしたクラウド連動型スーパーコンピュータの実現を可能にする。本研究推進段階では、民間クラウドの提供するベアメタル型クラウド資源を利用できなかったが、本研究で実装したクラウドバースティング機能をベアメタル型クラウド資源で展開することにより、ベアメタル型クラウドバースティング機能が完成する。当該研究成果は上述した(1)の成果と合わせることで、利用者の要望に基づいたアクセラレータを動的に計算資源に割り当てることで利用者の多様な計算ニーズを収容可能な計算環境をオンプレミス環境だけでなく、オンプレミスおよびクラウドのハイブリッドクラウド環境で実現することが可能となる。本研究の成果は、このようなベアメタルクラウド構成技術の可能性、すなわち、利用者が計算機にあわせた計算を行うのではなく、利用者の計算ニーズにあわせて計算環境を動的に構成可能とすることを示したことである。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Akihiro Misawa, Susumu Date, Keichi Takahashi, Takashi Yoshikawa, Masahiko Takahashi, Masaki Kan, Yasuhiro Watashiba, Yoshiyuki Kido, Chonho Lee, Shinji Shimojo	４．巻 864
２．論文標題 Dynamic Reconfiguration of Computer Platforms at the Hardware Device Level for High Performance Computing Infrastructure as a Service	５．発行年 2018年
３．雑誌名 Communications in Computer and Information Science	６．最初と最後の頁 177-199
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-319-94959-8_10	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Susumu Date, Hirotake Abe, Dashdavaa Khureltulga, Keichi Takahashi, Yoshiyuki Kido, Yasuhiro Watashiba, Pongsakorn U-chupala, Kohei Ichikawa, Hiroaki Yamanaka, Eiji Kawai, Shinji Shimojo	４．巻 22
２．論文標題 SDN-accelerated HPC Infrastructure for Scientific Research	５．発行年 2016年
３．雑誌名 International Journal of Information Technology	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計17件（うち招待講演　1件／うち国際学会　12件）

１．発表者名 Yuki Matsui, Yasuhiro Watashiba, Susumu Date, Takashi Yoshikawa, Shinji Shimojo
２．発表標題 Job Scheduling Simulator for Assisting the Mapping Configuration between Queue and Computing Nodes
３．学会等名 The 33rd International Conference on Advanced Information Networking and Applications (AINA-2019)，(国際学会)
４．発表年 2018年

１．発表者名 Susumu Date, Yuki Matsui, Yasuhiro Watashiba, Takashi Yoshikawa, Shinji Shimojo
２．発表標題 Job Scheduler Simulator Extension for Evaluating Queue Mapping to Computing Node
３．学会等名 28th Workshop on Sustained Simulation Performance (WSSP)（国際学会）
４．発表年 2018年

１．発表者名
Yuki Matsui, Yasuhiro Watashiba, Susumu Date, Takashi Yoshikawa and Shinji Shimojo

２．発表標題
Architecture of Job Scheduling Simulator for Evaluating Mapping Between Queue and Computing Node

３．学会等名
PRAGMA 34 workshop（国際学会）

４．発表年
2018年

１．発表者名
三澤明寛，高橋慧智，渡場康弘，伊達進，吉川隆士，阿部洋丈，野崎一徳，木戸善之，Lee CHONHO，下條真司

２．発表標題
医療応用を考慮した動的構成変更可能計算機クラスタの検討

３．学会等名
日本ソフトウェア科学会 第16回ディペンダブルシステムワークショップ

４．発表年
2018年

１．発表者名
松井祐希，渡場康弘，伊達進，吉川隆士，下條真司

２．発表標題
細粒度マッピング設定に対応したジョブスケジューリングシミュレータの構築

３．学会等名
第143回 システムソフトウェアとオペレーティング・システム研究会

４．発表年
2018年

１．発表者名
Susumu Date

２．発表標題
Secure Staging Mechanism towards AI-related research Treating Security-sensitive Scientific Data

３．学会等名
Southeast Asia International Joint Research and Collaboration Program in High-Performance Computing Applications and Networking Technology(SEAIP) 2018（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Susumu Date

２．発表標題
Secure Staging Structure for Treating Security-sensitive Scientific Data

３．学会等名
NUG 2018

４．発表年
2018年

１．発表者名
Yoshiyuki Kido

２．発表標題
ダイナミックセキュアステージングを用いた医療データ解析環境

３．学会等名
Small-workshop on Communications between Academia and Industry for Security (SCAIS2019)

４．発表年
2018年

１．発表者名
Masaharu Shimizu, Yasuhiro Watashiba, Susumu Date, Shinji Shimojo

２．発表標題
Adaptive Network Resource Reallocation for Hot-spot Avoidance on SDN-based Cluster System”

３．学会等名
NetCloud 2016 workshop, 8th IEEE International Conference on Cloud Computing Technology and Science (CloudCom2016) (国際学会)

４．発表年
2016年

１．発表者名
Yasuhiro Watashiba, Susumu Date, Hirotake Abe, Kohei Ichikawa, Yoshiyuki Kido, Hiroaki Yamanaka, Eiji Kawai

２．発表標題
Architecture of Virtualized Computational Resource Allocation on SDN-enhanced Job Management System Framework

３．学会等名
2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (国際学会)

４．発表年
2016年

１．発表者名
Susumu Date, Takashi Yoshikawa, Yasuhiro Watashiba, Yoshiyuki Kido, Shinji Shimojo, Masahiko Takahashi, Masaki Kan, Masaki Muraki

２．発表標題
A Proposal of On-demand Staging leveraging Job Management System and Software Defined Networking

３．学会等名
Workshop on Sustained Simulation Performance (WSSP)（国際学会）

４．発表年
2016年

１．発表者名
Seiya Murata, Chonho Lee, Chihiro Tanikawa, Susumu Date

２．発表標題
Towards a Fully Automated Diagnostic System for Orthodontic Treatment in Dentistry

３．学会等名
The thirteenth IEEE eScience Conference (e-science2017)（国際学会）

４．発表年
2017年

１．発表者名
Seiya Murata, Kobo Ishigaki, Chonho Lee, Chihiro Tanikawa, Susumu Date, Takashi Yoshikawa

２．発表標題
Towards a smart dental healthcare: an automated assessment of orthodontic treatment need

３．学会等名
The Second International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing (HEALTHINFO 2017)（国際学会）

４．発表年
2017年

１．発表者名
Susumu Date, Takashi Yoshikawa, Kazunori Nozaki, Yasuhiro Watashiba, Yoshiyuki Kido, Masahiko Takahashi, Masaya Muraki, Shinji Shimojo

２．発表標題
Towards a Software Defined Secure Data Staging Mechanism

３．学会等名
Sustained Simulation Performance 2017 (WSSP 2017)（国際学会）

４．発表年
2017年

１．発表者名
Arata Endo, Ryoichi Jingai, Susumu Date, Yoshiyuki Kido, Shinji Shimojo

２．発表標題
Evaluation of SDN-based Conflict Avoidance between Data Staging and Inter-Process Communication

３．学会等名
The 2017 International Conference on High Performance Computing & Simulation (HPCS 2017)（国際学会）

４．発表年
2017年

１．発表者名
Akihito Misawa, Susumu Date, Keichi Takahashi, Takashi Yoshikawa, Masahiko Takahashi, Masaki Kan, Yasuhiro Watashiba, Yoshiyuki Kido, Chonho Lee, Shinji Shimojo

２．発表標題
Highly Reconfigurable Computing Platform for High Performance Computing Infrastructure as a Service: Hi-IaaS

３．学会等名
The 7th International Conference on Cloud Computing and Services Science (CLOSER 2017)（国際学会）

４．発表年
2017年

１．発表者名
伊達進，吉川隆士，野崎一徳，渡場康弘，Lee Chonho，木戸善之，下條真司

２．発表標題
医療データを高性能計算機システムで利用するためのダイナミックセキュアなステージングシステム

３．学会等名
第37回日本医療情報学連合大会

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	伊達　進 (Date Susumu) (20346175)	大阪大学・サイバーメディアセンター・准教授 (14401)	

6．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	木戸　善之 (Kido Yoshiyuki) (70506310)	大阪大学・サイバーメディアセンター・講師 (14401)	
連携研究者	渡場　康弘 (Watashiba Yasuhiro) (60758275)	大阪大学・サイバーメディアセンター・特任講師 (14401)	