様式 C-19

科学研究費補助金研究成果報告書

平成　23　年　6　月　26　日現在

機関番号：12608
研究種目：特定領域研究
研究期間：2006 ～ 2010
課題番号：18049028
研究課題名（和文）情報爆発時代に対応する高度にスケーラブルな高性能自律構成実行基盤

研究課題名（英文） Highly Scalable, High Performance and Autonomous Distributed Execution for Information Explosion Environments

研究代表者
松岡　聡（MATSUOKA　SATOSHI）
東京工業大学・学術国際情報センター・教授
研究者番号：20221583

研究成果の概要（和文）：
100 万のオーダーのノードからなる超分散環境上で多種なアプリケーションを安全安心に実行するための真の高度にスケーラブルな自律的実行基盤「リジリエント・グリッド(Resilient Grid)」の構築のための研究開発を推進し、高性能基盤技術、実行基盤の自律的構成、次世代ネットワークと実行基盤の融合、性能モデリングとシミュレーションの点において要素技術を確立した。

研究成果の概要（英文）：
We have conducted several fundamental research activities for constructing highly scalable, high performance and autonomous distributed execution environments, called “resilient grids”, for the information explosion era. We have built the constituent techniques, including modeling and simulation, for the resilient grids in terms of autonomous construction of high performance application execution environments and federation of future-networks and the environments.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６	17,100,000	0	17,100,000
２００７	16,100,000	0	16,100,000
２００８	18,200,000	0	18,200,000
２００９	16,800,000	0	16,800,000
２０１０	18,900,000	0	18,900,000
総　計	87,100,000	0	87,100,000

研究分野：情報爆発時代に向けた新しい IT 基盤技術の研究
科研費の分科・細目：情報学／計算機システム・ネットワーク
キーワード:ディペンダブル・コンピューティング, ハイパフォーマンスコンピューティング, グリッド, P2P, 分散処理

１．研究開始当初の背景

情報爆発はその膨大な情報量を扱う分散システムの規模を百万ノード以上のオーダーに肥大化・複雑化させ、革新的な利用法が期待される反面、全体の高度なセキュリティ・百万オーダーへのスケーラビリティ・大規模な故障に対する自律的なリジリエンシー(堅固さ)が欠如し、全体の安全安心は確保できていない。例えば、Blaster ウィルスによる大規模 DDoS(分散 Denial of Service)アタックでは、数千から数万規模のノードが乗っ取られ、大企業のサーバが攻撃を受け、北米の大停電等に繋がった可能性があるとIEEE のセキュリティ会議にて報告されてい

る。また、最近の交通管制システムのダウン、銀行システムの停止、携帯メイルの不調などの社会の根幹の IT 基盤システムの不安定性は、システムエンジニアの不断の努力のみが頼りの基盤 IT システムが、今後の情報爆発による管理の限界を超えることの初期の具現化といえる。さらには、グリッド技術の台頭により大規模並列計算科学アプリケーションを分散環境で実行し、それにより科学技術を進歩させ安全安心に貢献することが期待されるが、環境の本質的な不安定さ・安全安心な動作環境の欠如等により、現状のグリッドは情報爆発に対応できない。

２．研究の目的

情報爆発時代に対応できる計算基盤として、100 万のオーダのノードからなる超分散環境上で多種なアプリケーションを安全安心に実行するための真の高度にスケーラブルな自律的実行基盤「リジリエント・グリッド(Resilient Grid)」を構築することを目的とし、それらの実現のための要素技術の研究開発を目指す。

３．研究の方法

「リジリエント・グリッド(Resilient Grid)」の実現に必要な要素技術を以下の 4 点に分類し、研究を推進する。

1. 高性能実行基盤技術
2. 実行基盤の自律構成
3. 次世代ネットワークと実行基盤の融合
4. 性能モデリングとシミュレーション

４．研究成果

3 章で述べたとおり、「リジリエント・グリッド(Resilient Grid)」の実現に必要な要素技術の研究開発を推進した。ここではその一部について述べる。

4.1 高性能実行基盤技術

4.1.1. スケーラブルかつ高速な仮想クラスタの構築

カスタマイズ性と高速性を兼ね備えた構築システムの予備的な設計と実装による実環境での検証を行った。仮想マシンイメージのキャッシュ、再利用により構築時間を削減する方式を提案し、キャッシュ対象の仮想マシン構成を自動的に選択するアルゴリズムを開発した。同キャッシュ方式の有効性を評価するために、 クラスタ構築ツール Lucie をベースにした仮想クラスタ構築システムのプロトタイプを実装した。同プロトタイプを用いた評価により、200 台程度の仮想クラスタを 40 秒程度で実際に構築できることを確認した。

仮想マシンイメージのキャッシュ(特定のソフトウェアパッケージをあらかじめインストールした仮想マシンのイメージファイル)は、パッケージインストール時間が仮想クラスタ構築時間全体において支配的であり、その時間を削減することが高速化に有効だからである。キャッシュイメージにインストールするパッケージの選択においては、キャッシュ化による高速化への有効性を見積り、有効な組合せのイメージを優先的に選択する。パッケージ組合せの有効性は、今後その組合せをユーザからどの程度の頻度で要求されるか、また実際にその組合せを用いた際にインストール時間がどの程度削減されるかの二点に基づく。我々はパッケージの組合せとその将来のインストール要求頻度を、過去のインストール履歴に階層的クラスタ分析を応用することで決定する。キャッシュ対象パッケージの選択は、利用可能ディスクスペースに収まる範囲内で、有効性の高い組合せから順に選択する。選択された組合せについて、そのパッケージをインストールした仮想マシンイメージをファイルとして生成し、キャッシュが作成された後のインストール要求では、キャッシュに対する差分パッケージのみをインストールする。

また、これらの研究をベースとして、性能モデルに基づいた資源選択による高速な仮想クラスタの構築について取り組んだ。大規模グリッド環境上で仮想クラスタを用いるためには、スケーラブルかつ高速な構築が必要であるが、ハードウェア性能が不均一となりうるため、ノード選択によっては性能が極端に低いノードによって全体の構築時間が大幅に増加しうる。そのため、ノード毎の性能が非均一な大規模グリッド環境においても安定して高速な構築を行うために、構築時間のモデル化に基づいた資源選択手法の提案、開発を行った。本手法では、構築プロセスを論理的なステップに分割し、ステップごとに実行時間を予測する。性能モデルは各計算ノードの性能（CPU 周波数、ディスク読み書き速度、インストールパッケージ容量など）をパラメータとし、それらの線形結合で各ステップの実行時間を表現する。モデル係数の決定は、仮想クラスタを実際に構築し、その性能データを基に重回帰分析を行う。仮想クラスタインストーラ VPC について、このモデルに基づく計算ノード選択機能を拡張し、評価実験を行った。実験では、エミュレートされた 2 サイト上で仮想クラスタを構築し、各サイトではインストールパッケージ容量を変化させた。その結果、モデルに基づく選択法は、各サイトでインストールパッケージ容量が異なる場合に特に有効であり、ランダムに計算ノードを選択する最も単純な手法(FIFO)に比べて最大 68%、各計算ノードの CPU 周波数だけを考慮した選択法に比べて最大 60%、ディスク性能だけを考慮した選択法に比べて最大 58%の構築時間短縮を実現

できることが分かった。

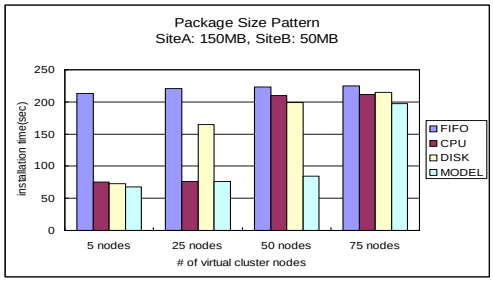


図 1 各資源選択における
仮想クラスタ構築時間の比較

4.1.2. グリッドファイルシステムにおけるアクセスパターンと性能を考慮した複製配置

ファイルシステムを用いたグリッド環境での大規模なデータ共有は，シングルシステムイメージを実現し，ユーザの利便性を向上させる有効な手法である．しかし，ファイルへのアクセス集中や遠方へのファイルアクセスなど，不均質な環境での，煩雑なデータ管理が発生することが問題となる．我々は，ファイルのアクセス頻度や管理ポリシーに応じて自動的にファイルの複製配置を決定するアルゴリズムを提案した．提案アルゴリズムでは，この複製配置問題をアクセス時間，ストレージ容量，及び，転送時間の最小化を関数とする 0-1 整数計画問題に帰着し，ファイルアクセスのモニタリングにより得られた情報を利用することにより解く．シミュレーションでの評価では，複製作成を行わない手法，アクセス時に複製をキャッシュする手法，　サイト毎に複製を持つ手法などの単純な複製管理手法と比較して，ストレージ使用量を低く保ちつつ，かつ，高いスループット性能を達成する複製配置を自動的に実現することを
確認した．また，現在，この提案アルゴリズムを既存のグリッドファイルシステム(Gfarm)に適用し，InTrigger テストベッドに配備して，実アプリケーション(Blast)を用いた有効性を示した。

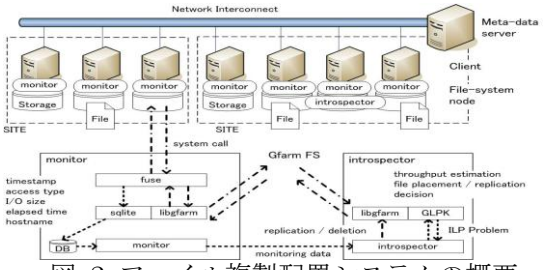
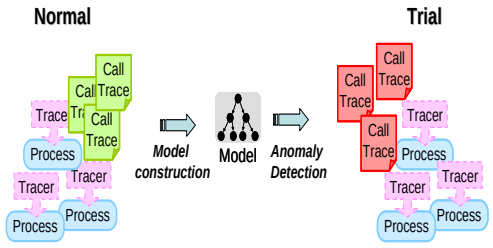


図 2 ファイル複製配置システムの概要

4.2. 実行基盤の自律構成

4.2.1. 自律的な障害解析に関する研究

数百万ノードに及ぶ情報爆発環境ではシステムがソフトウェア、ハードウェア両面において複雑、大規模化し、従来の人手による解析手法では対応が困難になっている。我々は分散システムの実行状態を常に監視し、異常原因の解析を行う方式を開発した。図２にあるように、同解析手法はシステムの各構成プロセスの関数呼び出しを常に記録し、正常に実行された場合と故障が発生した場合を比較することで障害の原因を解析する。まず、正常実行時のトレースよりシステムの正常な振る舞いを表すモデルを構築する。モデルでは各関数についてその呼び出し確率を推定する。システムに故障が発生した場合は、構築済みモデルと故障発生時のトレースを比較し、異常な関数呼び出しを検出する。本手法では、通常は呼ばれる関数が呼ばれなかった場合とその逆を異常と定義し、呼び出し確率を正常時トレースより推定することで異常な振る舞いを検出する。検出された異常な呼び出しをシステム管理者、開発者に通知することで、さらなる人手による障害解析を支援する。本手法をグリッド環境において発生する障害の解析に適用し、有効性を検証した。InTrigger 環境において複数拠点からなるグリッド環境において､MPICH が適切に MPI ジョブを起動できない障害が確認されている。我々はまず正常にジョブが実行
された場合のトレースを取得し、正常実行モデルを構築した。さらに障害原因を解析
するために 3 拠点､78 ノードを用いた構成において障害トレースを取得し、モデルを用いた異常検知を行った。その結果、同障害の原因バグに該当する箇所を容易に特定できることがわかり、本障害解析手法の有効性を確認できた。



4.3. 次世代ネットワークと実行基盤の融合

4.3.1. 光ネットワークを活用するスケーラブルな計算基盤

将来のペタフロップスを実現するスーパーコンピュータは数万～数十万機の数多くのプロセッサを搭載し、それらを接続するための大規模ネットワークを必要とするが、既存の電気ネットワークでは高消費電力であることが問題となる。本研究では、低消費電力で高バンド幅を実現できる光ネットワークに着目し、従来利用されている電気パケットネットワークと光サーキットネットワー

クを組み合わせたハイブリッドネットワークを提案した。各計算ノードは両ネットワークに対してそれぞれ1つのリンクで接続される。既存のパケットネットワークを使用したシステムに対し、光サーキットネットワークを追加するだけで容易に安価に拡張できる。また、提案ネットワーク上でMPIアプリケーションを実行する場合、各ノードは1つの光リンクしか持たないため、通信発生に対しon demandにサーキットの確立・解放を行うと実行時間の大幅な増加やサーキットのstarvationが起こりえる。そこでアプリケーションの通信パターンと局所性に着目した以下のMPI通信手法を提案した。

1. アプリケーションの通信パターン（プロセスごとの通信相手と通信量）を取得
2. 通信パターンとプロセス配置に応じてプロセスをグルーピング
3. グループ間通信に光サーキットを割り当て
4. 光・電気ハイブリッドネットワーク上でのフォワーディングテーブルを作成
5. フォワーディングテーブルを用いて通信を行う

提案ネットワークと通信手法によるMPIアプリケーションの実行時間と、電気パケットネットワーク上での実行時間をシミュレーションにより測定し、比較を行った。その結果、局所性の高い通信パターンをもつアプリケーションの場合は、少ない光サーキット数で、フルバイセクションバンド幅を実現するパケットネットワークよりも高速に実行できることを確認した。

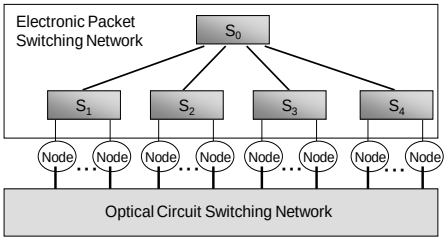


図3 光・電気ハイブリッドネットワーク

4.3.2. スケーラブルで自律的に実行可能なクラウドでのマルチキャスト手法

クラウドにおいて大規模データを用いた並列アプリケーションを実行する場合，処理データを各ノードに効率よく配布する必要があるが、従来の一般的な並列計算実行環境におけるマルチキャスト最適化手法では，動的にネットワーク性能が変化するクラウドにおいては十分な性能を常に発揮し続けることは難しい．そこで我々は，トポロジやバンド幅マップなどのモニタリング情報を用いずに，各ノードが自律的にマルチキャストスループットを最適化可能なアルゴリズムを提案した．提案アルゴリズムでは，(i)クラウドストレージから各ノードへのデータ転送部分と，(ii)各ノードから各ノードへのデータ転送部分のそれぞれに対して動的に最適化を行う．(i)においては，クラウドストレージからのダウンロードスループットが各ノードで変動するため，それぞれのノードが自律的に協調してダウンロードワークスティーリングを行い，ボトルネックリンクでの性能低下を補う．また，(ii)においては，各ノードがP2Pで主に用いられているBitTorrentプロトコルに似た手法を用いて，動的にノード間のボトルネックリンクを迂回して転送を行う．以上の手法により，各ノードのマルチキャストスループットを動的ロードバランスして最適化をすることが可能となる．このアルゴリズムを用いて，Amazon EC2/S3クラウドにおいてマルチキャストの実効性能を評価した．その結果，単純な手法に比べてノード数とデータサイズの増加に対してスケーラブルでかつ高い性能が得られることを確認した．また，提案手法は，常に全ノードが安定してマルチキャストを行えることも確認した．

4.4.性能モデリングとシミュレーション

4.4.1. 仮想マシンマイグレーションを考慮した大規模データ処理の最適化

仮想マシンマイグレーションを考慮した大規模データ処理の最適化を行った。既存研究でファイルキャッシュや複製を行うことにより向上をはかっているが、消費ストレージ容量やファイル転送時間が大きいという問題がある.そこで我々は、仮想マシンをアクセス対象のファイル、仮想マシンメモリサイズ、拠点間のネットワークスループットを考慮することでファイルが所在する拠点へ移動させることにより、この問題を解決する。

(1)仮想マシンマイグレーションの時間とファイルアクセスの時間のモデルを構築し、(2)アプリケーションのファイルアクセス履歴からファイルの依存関係を記述するマルコフモデルを構築し、そしてこれら2つのモデルから仮想マシンの移動パターンを有効非循環グラフ(DAG)として表現し、頂点の重みをファイルアクセス時間の期待値、辺の重みを仮想マシンマイグレーション時間として最短経路問題に帰着する最適化を行うことによって、データアクセスに最適な仮想マシンの移動先を決定する.提案手法をシミュレーションにより評価した結果、仮想マシンを移動させない手法(No Migration I/O)に比べ最大で38%、ファイルの存在する場所へ毎回移動しローカルアクセスを行う手法(Migration I/O)に比べ最大で47%のファイルアクセスのスループット向上を確認し、

性能モデル及びマルコフモデルに基づき仮想マシンを適切に再配置することにより、ファイルアクセスのスループット向上されることを示した．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）
〔雑誌論文〕（計 14 件）

① Satoshi Matsuoka, Takayuki Aoki, Toshio Endo, Akira Nukada, Toshihiro Kato, Atsushi Hasegawa, GPU accelerated computing-from hype to mainstream, the rebirth of vector computing, Journal of Physics: Conference Series, 180 巻, 1 号, 12-43 項, 2009, 査読有
② Satoshi Matsuoka, Kazushige Saga, Mutsumi Aoyagi, Coupled-Simulation e-Science Support in the NAREGI Grid, IEEE Computer, 41 巻, 11 号, 42-49 項, 2008 年, 査読有
③ Laurent Baduel, Satoshi Matsuoka, Outil autonome de surveillance de grilles(in French), Revue de l'Ingenierie des Systemes d'Information, 12 巻, 3 号, 2007, 査読有
④ Masao Sakauchi, Shigeki Yamada, Noboru Sonehara, Shigeo Urushidani, Jun Adachi, Kazunobu Konishi, Satoshi Matuoka, Cyber Science Infrastructure Initiative for Boosting Japanese Scientific Research, CTWatch Quarterly Journal, 2 巻, 1 号, 20-26 項, 2006, 査読無

〔学会発表〕（計 93 件）

① Ali Cevahir, Akira Nukada, Satoshi Matsuoka, High Performance Conjugate Gradient Solver on Multi-GPU Clusters Using Hypergraph Partitioning, International Supercomputing Conference (ISC2010), May 30, 2010, Hamburg, Germany
② Naoya Maruyama, Akira Nukada, Satoshi Matsuoka, A High-Performance Fault-Tolerant Software Framework for Memory on Commodity GPUs, IEEE International Parallel and Distributed Processing Symposium (IPDPS2010), Apr 19, 2010, Alaska, USA
③ Toshio Endo, Akira Nukada, Satoshi Matsuoka, Naoya Maruyama, Linpack Evaluation on a Supercomputer with Heterogeneous Accelerators, IEEE International Parallel & Distributed Processing Symposium (IPDPS2010), Apr 19 , 2010, Alaska, USA
④ Akira Nukada, Satoshi Matsuoka, Auto-Tuning 3-D FFT Library for CUDA GPUs, 2009 ACM/IEEE conference on Super Computing (SC09), Nov 14, 2009, Oregon, USA
⑤ Hitoshi Sato, Satoshi Matsuoka, Toshio Endo, File Clustering Based Replication Algorithm in a Grid Environment, 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2009), pp204-pp211, May 18, 2009, Shanghai, China
⑥ Sumeth Lerthirunwong, Naoya Maruyama, Satoshi Matsuoka, Adaptive Resource Indexing Technique for Unstructured Peer-to-Peer Networks, 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2009), pp172-pp179, May 18, 2009, Shanghai, China
⑦ Akira Nukada, Yasuhiko Ogata, Toshio Endo, Satoshi Matsuoka, Bandwidth Intensive 3-D FFT kernel for GPUs using CUDA, the ACM/IEEE conference on Supercomputing (SC08), Nov 15, 2008,Texas, USA
⑧ Hitoshi Sato, Satoshi Matsuoka, Toshio Endo, Naoya Maruyama, Access -Pattern and Bandwidth Aware File Replication Algorithm in aGrid Environment, The 9th IEEE/ACM International Conference on Grid Computing (Grid 2008), pp250-pp257, Sep 29, 2008, Tsukuba, Japan
⑨ Naoya Maruyama, Satoshi Matsuoka, Model-Based Fault Localization in Large-Scale Computing Systems, the 22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS 2008), Apr 14, 2008, Florida, USA
⑩ Toshio Endo, Satoshi Matsuoka, Massive Supercomputing Coping with Heterogeneity of Modern Accelerators, IEEE International Parallel & Distributed Processing Symposium (IPDPS 2008), Apr 14, 2008, Florida, USA
⑪ Hideo Nishimura, Naoya Maruyama, Satoshi Matsuoka, Virtual Clusters on the Fly --- Fast, Scalable, and Flexible Installation, the Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007), May 14, 2007, Rio de Janeiro, Brazil
⑫ Tatsuhiro Chiba, Toshio Endo, Satoshi Matsuoka, High-Performance MPI Broadcast Algorithm for Grid Environments Utilizing Multi-lane NICs, the Se

venth IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007), May 14, 2007, Rio de Janeiro, Brazil

⑬ Alexander V. Mirgorodskiy, Naoya Maruyama, Barton P. Miller, Problem Diagnosis in Large-Scale Computing Environments, the 2006 ACM/IEEE conference on Supercomputing (SC06), Nov 11, 2006, Florida, USA

〔図書〕(計 1 件)

① Satoshi Matsuoka,The Road to TSUBAME and Beyond,David Bader,Petascale Computing: Algorithms and Applications, Chapman & Hall/CRC Computational Science, pp289-310 (総ページ数530), 2007

〔産業財産権〕

○出願状況 (計 0 件)

○取得状況 (計 2 件)

①

名称: 支援プログラム、支援プログラム生成プログラム、支援プログラム生成方法、ファイル生成・配布方法、及びインストールサーバ

発明者:松岡聡, 高宮安仁, 竹内 義晴, 松岡浩司, 廣澤 治人

種類:特許

番号: 特願 2006-13022

出願年月日: 平成 18 年 1 月 20 日

取得年月日:平成 19 年 8 月 2 日

国内外の別:国内

②

名称:ファイル検査のための設定データ生成プログラム及びシステム

発明者:松岡聡, 脇田建, 高宮安仁

種類:特許

番号:特願 2004-257651

出願年月日: 平成 16 年 9 月 3 日

取得年月日:平成 22 年 10 月 22 日

国内外の別:国内

〔その他〕

① 松岡 聡, 2020 年に情報量 35Z バイト無尽蔵の需要に新技術で挑む, 日経エレクトロニクス, 2010/2/7

② 松岡 聡, スパコンも省エネ, 読売新聞(朝刊), 2010/1/30

③ 松岡 聡, マイクロソフト『産学連携』の凄み, 週刊東洋経済, 2010/12/4

④ 松岡 聡, スパコン『GPU』主流に, 日経産業新聞, 2011/1/17

⑤ 松岡 聡, 省エネ性能世界ランク 東工大スパコン 2 位, 読売新聞, 2010/11/24

⑥ 松岡 聡, アトムとウランの時事わーど百科『スパコン』, 読売新聞, 2010/12/3

⑦ 松岡 聡, 雪で冷却 節電スパコン, 読売新聞 (夕刊), 2011/1/11

⑧ 松岡 聡, GPU で最速スパコン, 朝日新聞(朝刊), 2011/11/19

⑨ 松岡 聡, 中国スパコン 軍の影, 朝日新聞 (朝刊), 2011/11/18

⑩ 松岡 聡, スパコン性能、中国が初の首位, 日経新聞 (夕刊), 2010/11/24

⑪ 松岡 聡, 東工大、世界初の GPU 採用スパコンに進化した「TSUBAME 1.2」を解説, Impress PC watch, 2008/12/3

⑫ Satoshi Matsuoka, 170 Tesla S1070 1u Systems Makes Tokyo Tech Tsubame Supercomputer, TechLime, 2008/11/18

⑬ Satoshi Matsuoka, Tokyo Tech upgrades TSUBAME supercomputer, Scientific Computing, 2008/11/19

⑭ 松岡 聡, 東工大の TSUBAME が汎用 GPU 計算アクセラレータで性能強化, IT media, 2008/12/2

⑮ 松岡 聡, 東工大、スパコン「TSUBAME」の性能を強化, 日経BP ITpro, 2008/12/2

６．研究組織

(1)研究代表者

松岡 聡 (MATSUOAKA SATOSHI)
東京工業大学・学術国際情報センター・教授
研究者番号：20221583

(2)研究分担者

合田 憲人 (AIDA KENTO)
国立情報学研究所・
リサーチグリッド研究開発センター・教授
研究者番号：80247212

中田 秀基 (NAKADA HIDEMOTO)
(独) 産業技術総合研究所・
情報技術研究部門・主任研究員
研究者番号：80357631

竹房 あつ子 (TAKEFUSA ATSUKO)
(独) 産業技術総合研究所・
情報技術研究部門・研究員
研究者番号：70345411

(3)連携研究者

丸山 直也 (MARUYAMA NAOYA)
東京工業大学・学術国際情報センター・助教
研究者番号：60532801

實本 英之 (JITSUMOTO HIDEYUKI)
東京大学・情報基盤センター・助教
研究者番号：00545311

佐藤 仁 (SATO HITOSHI)
東京工業大学・学術国際情報センター・
特任助教
研究者番号：00550633

滝澤 真一朗 (TAKIZAWA SHINICHIRO)
東京工業大学・学術国際情報センター・
特任助教
研究者番号：80550483