課題番号 jh150042-MD03

科学技術計算における効率の良い複数拠点利用とそれを実現するユーザ駆動型・拠点協調フレームワークの開発と検証

實本英之（東京工業大学）

概要　広域に分散した計算機資源を有効に活用するためのフレームワークのプロトタイプについて、実アプリを利用した大規模検証を実施する。また、これを元にフレームワークの機能追加／安定化および、適切な用法を構築し、基盤センターもしくは HPCI 環境でのサービスとして提供可能な品質条件を明らかにすることを目的とする。フレームワークが対象とするアプリケーションは、引き続きシミュレーションと可視化を伴うもの、あるいはマルチスケールな構造をもった連成アプリケーションを対象とする。

1. 共同研究に関する情報

(1) 共同研究を実施した拠点名

東京大学、東京工業大学、九州大学、北海道大学

(2) 共同研究分野

□　超大規模数値計算系応用分野

□　超大規模情報システム関連研究分野

(3) 参加研究者の役割分担

<u>東京工業大学（システム設計・環境整備）</u>

實本英之（代表）：研究統括

システム設計・構築

三浦信一：RENKEI-VPE 環境調整

佐藤仁：ストレージ情報提供

<u>東京大学（システム設計及びアプリ検証）</u>

松本正晴：アプリ検証

中島研吾：アプリ提供

塙敏博：システム設計情報提供

片桐孝洋：アプリ情報提供

奥田洋司：アプリ情報提供

<u>九州大学（アプリ検証）</u>

小林泰三（副代表）：アプリ検証と提供

システム設計

<u>北海道大学（研究環境構築・整備）</u>

棟朝雅晴：検証環境調整

<u>理化学研究所（システム設計補助）</u>

滝澤真一朗：システム設計情報提供

2. 研究の目的と意義

2014 年度度行った、広域分散計算や連成計算の必要条件と要素技術の検討・整備を元に一部構築した「広域に分散した計算機資源を有効に活用するためのフレームワーク」のプロトタイプについて、実アプリを利用した大規模検証を実施する。また、これを元にフレームワークの機能追加／安定化および適切な用法を構築し、基盤センターもしくは HPCI 環境でのサービスとして提供可能な品質条件を明らかにすることを目的とする。フレームワークが対象とするアプリケーションは、引き続きシミュレーションと可視化を伴うもの、あるいはマルチスケールな構造をもった連成アプリケーションを対象とする。

現在のフレームワークはログインノードを Point of Presence(PoP)とし、ここでメッセージのリダイレクションを行うプロセスを立ち上げることで、各拠点のプライベートな計算リソース上で動いている連携アプリケーション間でのメッセージ送受信を可能としている。しかしながら、この PoP 上サービスプロセス（PoP サーバ）は、拠点ポリシに則り規定時間で強制終了される可能性がある。これはログインノードのリソースを使い尽くさないための手法であり、ほとんどの大規模計算拠点において採用されているものである。このため、2014 度の設計において、PoP サーバを必要に応じて再起動し、連成計算環境の自動的な再構築を行う手法を検討

した。しかし、この手法は、ログインノードのリソースを各拠点の定めるポリシを越えて利用することになるため、実利用において、どの程度の環境負荷を与えるのか詳細に測定し、導入コストを明らかにする必要がある。さらに、環境負荷をなるべく与えない手法を検討する必要があり、そのためには連成アプリケーション自体の構造においても最適化を行う必要がある。これらの検討・最適化を行うため、および、大規模利用による高度なアプリケーション実行結果をめざし、実アプリケーションを用いた大規模連成実行を行う。

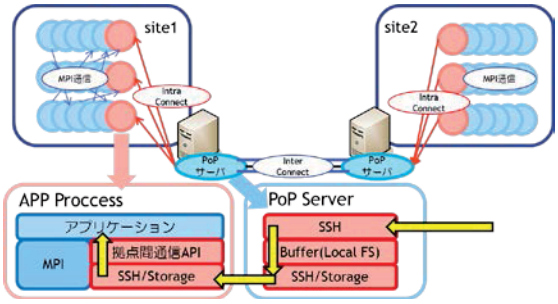

図 1. フレームワークにおけるメッセージリダイレクション

3. 当拠点公募型共同研究として実施した意義

拠点間連携システムの構築、および拠点間アプリケーションの特性を検証するにあたり、実際に多拠点が SINET で結びついた環境を利用できることは大きな利点である。

また、各拠点の計算資源のネットワークへの接続方法、利用法などの知識の提供、さらに構築途上のシステムテストのためにある程度検証に向いた環境設定を各拠点で行える研究者との共同研究は必須であり、運用システムと研究者が密に連携可能である本公募型共同研究によりこれが達成された。

実際に、本共同研究の資源である RENKEI-VPE 環境は、小規模な共同研究で構築することは難しいが、本研究課題の核心となる評価環境として大きな支援を得ている。

4. 前年度までに得られた研究成果の概要

フレームワークの実装・構築に先立ち、拠点間連携を行う既存手法である「HPCI 先端ソフトウェア運用基盤 分散環境ホスティングサービス」を利用した連成計算アプリケーションの評価を行った。対象とする連成計算アプリケーションとして、3 次元熱伝導方程式の有限差分法（FDM）解析コードに可視化処理用のルーチンを導入したもの用いた。

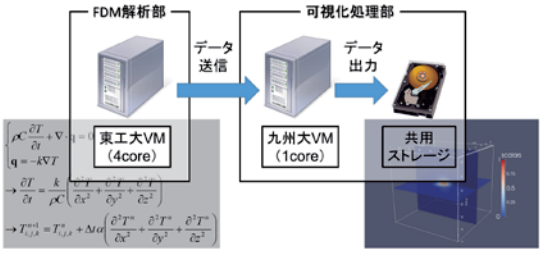

図 2. 連成計算アプリ評価概要

このコードは FDM 解析部と可視化処理部がそれぞれ特定の MPI プロセスで分割される。このため、各 MPI プロセスと各拠点の VM を対応させることによって、FDM 解析部と可視化部をそれぞれ別拠点で実行することができるものである。可視化部へのデータ出力間隔と FDM 解析時間を調整し、通信の隠蔽が十分行われる状況において、1) 東工大 4core のみと 2) 東工大 4core+九州大 1core を利用し解析部 4 プロセス、可視化部 1 プロセスのジョブを実行したところ、1) では 8 sec. 2) では 9 sec. の実行時間となった。

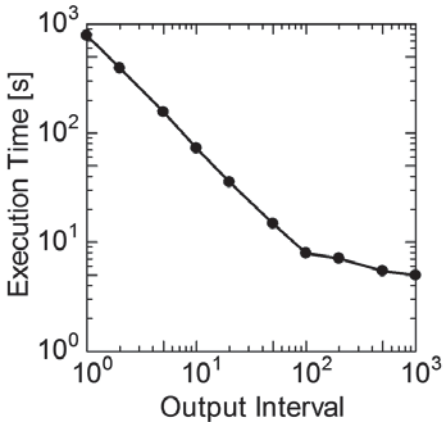

図 3：データ出力間隔に対する実行時間

この結果は適切なパラメータ設定が行われた場合、多拠点連携処理に効果がある可能性が示唆される。この結果は、システム設計の検討材料とした。

検証結果と合わせ、システム設計も行った。本研究で提案するユーザ駆動型・拠点協調フレームワークは、連成アプリケーションを多拠点実行する際に問題となるものとして、1)アプリケーション間メッセージをどのように最適に送受信するか、2)各拠点の基盤システムの差異を埋めながらアプリケーションを多拠点に同時投入しているように見せかけるにはどうすれば良いかに注目してフレームワークを設計した。

図 4 に示すフレームワークの全体像について、機能 2, 3, 5 に関して詳細な設計を行った。

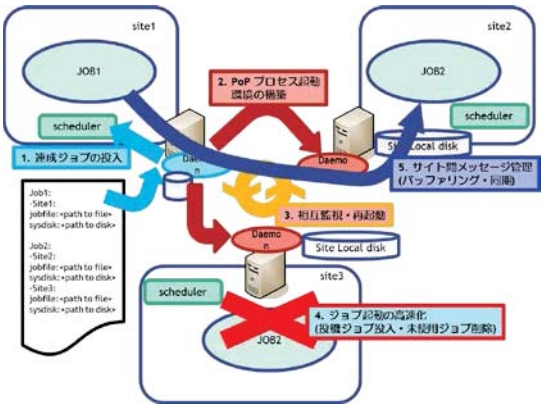

図 4. フレームワーク全体像

主要な機能は機能 5 についてであり、アプリケーション間メッセージは、ゲートウェイノードを通して、他サイトに転送する。ゲートウェイノードには PoP サーバと呼ばれる、メッセージ通信、ジョブ投入を管理するプロセスを実行する。このシナリオについては前章、図 1 に記した。サイト間ネットワーク(Inter connection),サイト内ネットワーク(Intra connection) については、スケーラビリティの必要性、各拠点に適応・有効利用するための拡張正の程度に着目し設計を行った。結果、Inter connection については SSH トンネリングを用いた拠点間 1 対 1 通信を、Intra Conncection については、SSH を併用した TCP/IP コネクションもしくは共有ストレージによる通信を対象とし、さらに共有ストレージの並列入出力機能を最大限に利用するなど、他のシステム研究の成果を考慮した拡張性の高いものにすることとした。そのほか、機能 3 に関しては、ACK 応答を利用することによりサイト間メッセージの受信側 PoP プロセス、送信側 PoP プロセスの生存パターンおよび ACK 応答の有無に応じた 5 種類の対応を洗い出し、対応策を検討した。

さらに、連成アプリケーションを達成する手法の一つとして OpenFOAM に対し、ポスト処理を連成可能になるような拡張を行った。これに関して試験実装を行い、今後の大域化に向けて用いる連成アプリケーションについて単拠点大規模実行を行い、成果を得た。このアプリケーションは昨年度 JHPCN 課題 14-NA17 曲管を有する管楽器を対象とした大規模並列流体音シミュレーション（代表 小林泰三）で行ったフルートの歌口部分の流体シミュレーションを対象として OpenFOAM を本課題で拡張することによりポスト処理連成を含めたアプリケーションとしたものである。

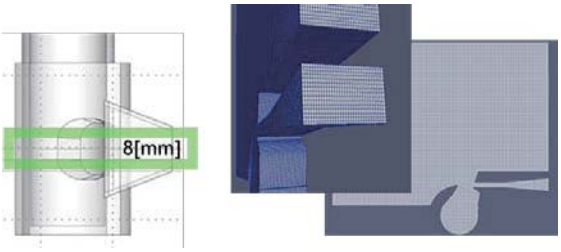

図 5. フルートの歌口補助具の解析

5. 今年度の研究成果の詳細

昨年度積み残していたフレームワークを実現するシステムソフトウェア群の実装を行った。サイト間ネットワークについては SSH トンネリングを行った上での TCP/IP 通信を利用し、サイト内ネットワークでは送信側、受信側双方とも共有ファイルシステム

を利用した手法を優先的に実装した。各メッセージに関する送受信は多種の入出力用ディスクリプタを一元的に扱うことができるイベントドリブンフレームワーク提供ライブラリである libevent を利用した。

メッセージは以下の手順により送出される。1) アプリケーションから呼び出される API により、送出データをパケットサイズに分割、ヘッダを付与して共有ファイルシステムに保存、2)PoP サーバは通信用ファイルからパケット単位でデータを非ブロッキング読み出し。読み出しが終了次第、次のデータの読み込みを非ブロッキングで開始しつつ、他拠点の PoP サーバに送出する。3)受信元の PoP サーバはパケットヘッダを読みながら、(送信元 ID, 送信先 ID) のセット毎に共有ファイルシステム上に作られた通信用ファイルにパケットを書き込む。4) アプリケーションから呼び出される API がヘッダを削りつつパケットをまとめてメッセージを復元する。現時点ではメッセージのバイナリ形式に対して拠点間のアーキテクチャ差異をうめるような実装はなされていない。これに関しては今後、形式をコンバートする API を提供する予定である。これは、MPI のデータ型に似た実装となる予定である。

中間報告以降、本システムの動作チェックとデバグおよび、基礎的な評価を行った。評価は東工大 TSUBAME2.5 のログインノードおよび、TSUBAME-KFC のログインノードを用いて行い、システムを用いた転送バンド幅を測定した。

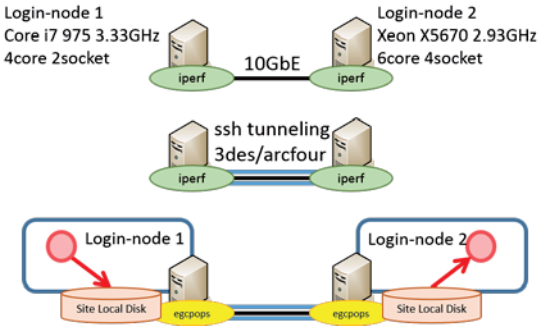

図 6. 実験環境とセッティング

提案システムは、ログインノード上にアプリケーションプロセスとしてメッセージ送受信アプリを配置し、ログインノードにマウントされた共有ファイルシステムを用いた転送を行った。ただし、本試験では同じノード上（ログインノード上）でファイルの作成、読み書きを行っており、転送サイズがノードのメモリ量より大きく少ないため、ファイルによる通信性能はメモリ転送である。

比較対象として、双方のログインノード上でネットワークスループット測定ツール iperf を用いたもの、さらに ssh トンネリングを行った上で、iperf を用いたものを利用した ssh トンネリングは、well-known サービスで本環境を構成する為に必須であるが、経路の暗号化を伴うため、転送性能に影響がある。今回は ssh のデフォルト設定である 3des および、比較的コストの低い arcfour 暗号化アルゴリズムについて比較を行った。iperf/egcpops(提案システムのサーバプログラム) 双方とも転送ブロックサイズは 8192 バイトとし、全体で 4G バイトの転送を行った。egcpops においては、1 ブロックにつき 20 バイトのヘッダが付与されるため、実際の転送量は 4.01GB 程度となる。実際に iperf については 10 秒毎の測定について最も性能の良いデータ、egcpops では、最後の 1G バイトの転送についてのデータとする。全ての評価は 5 回の平均となっている。結果は表 1 に記載した。

表 1: バンド幅性能比較 (Gbps)

cipher	iperf	egcpops
none	6.58	
3des	0.822	0.630
arcfour	1.65	1.20

次に、TSUBAME-KFC 側について、読み込み側アプリケーションを計算ノードに配置した上で、arcfour 暗号化アルゴリズムを用い

た際のバンド幅性能を測定した。その他のパラメータは先の実験と同じものである。利用される共有ファイルシステムは NFS であり、bonnie++により測定された性能はシーケンシャル I/O について書き込み 896Mbps、読み込み 904Mbpss である。

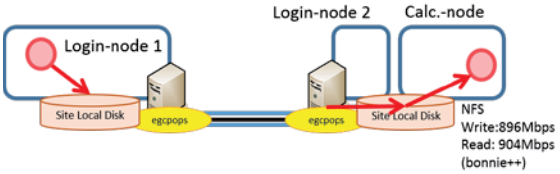

図 7. 共有ストレージを介した通信性能測定

先の実験と同様に 5 回測定を行った上での平均は 878Mbps となった。以上より、アプリケーション・egcpops 間通信における性能低下はおおよそ 3%程度、また、egcpops 間における性能低下はおおよそ 25%前後と考える。本システムは、github からアプリケーションのサンプルコード、設定ファイル例とともに一般に公開している（その他 1）。

九大システム・アプリケーションチームにおいては、OpenFOAM のポスト処理連携に関して、本研究課題により作成された egcpops を利用すべく並行して準備・開発してきたが、現段階での egcpops を試験適用するにあたり、これまで開発してきたクラスへの組み込みではなく、OpenFOAM に標準として用意されている FunctionObjects を利用する方法に変更して試験を行った。これは新潟大学の大嶋拓也先生(H24 年度 JHPCN 採択課題 12-NA18「実在地域における建築・都市環境の総合数値予測」）からご教示を戴いた手法を用いた実装である。テスト環境と条件は以下である。研究室設置の PC クラスタの OpenFOAM で流体音計算を実行し、音圧データを手元のサーバへ転送する設定で試験を行った。転送する対象ファイルを一つにまとめることで数ステップの短時間小規模な疎通テストは確認できたが、基盤センターでのサービスとして提供するまでには検証が進んでいない。

一方で、研究開発成果を用いて行った数値計算結果からは、学術論文 1 編、国際会議 5 編と学会発表等 1 3 編を発表することができた。国際会議発表 2,3,5 と国内会議発表 1-5,8-15 は、本研究課題で開発した OpenFOAM のポスト処理連携を試験的に用いて流体音数値解析を行った結果の学術発表である。発表した学術分野は物理学・流体力学・音響学である。発表内容については、当該分野でこれまでにない精密な計算結果に基づく議論が可能になり、エアリード楽器における流体音の発音メカニズムにおける新たな解釈を提案している。特に、流体から音へのエネルギー遷移が起こる領域を数値的に確認できたことが最大の成果である。

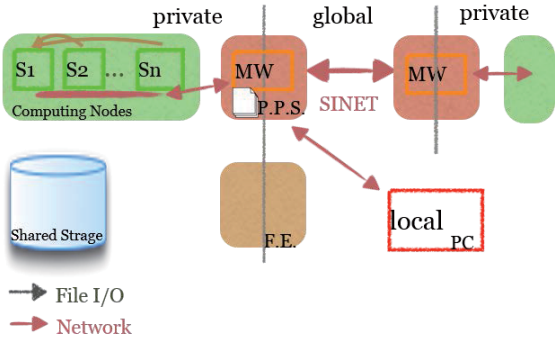

図 8. 拡張 OpenFOAM へのフレームワーク適用

国際会議発表 4 は、本研究課題の成果をまとめた発表であり、現在の過渡現象数値解析が抱える問題を視点に据えて、目指すべき連成・連携計算の形（図 8）を提案したものである。図 8 は多拠点間連成計算の概念を示している。この発表では、*in situ* visualization の可能性や、数値計算からポスト処理へのデータ転送にデータ圧縮などのデータ転送量の削減の必要性について活発な議論が交わされ、本研究課題の成果が広く待たれていることを確認できた。

国際会議発表 1, と国内会議発表 16 は、本研究課題のような複数拠点を協調して利用する条件に関する考察に関するものである。複数拠点を利用するには、各拠点やユーザが抱える種々の条件を満たす必要があるが、それ

らの多くはジョブを実行する、あるいは実行している最中でなければ決定できない不確定要素を多く含むものである。そのような不確定な状況をマネジメントするシステムには、アルゴリズムを実行プログラムとしてコーディングする従来のソフトウェエア開発手法では対応困難であり、本研究課題で提案しているような、制御構造をプログラムの外に置く方法をとらざるを得ないことを指摘した。

また、フレームワークを実現するシステムソフトウェアと現在開発中の連成機能を組み込んだ OpenFOAM の協調手法について詳細検討を行った。

6. 今年度の進捗状況と今後の展望

システムソフトウェアの実装が昨年度から続いており、公開を行った(その他 1)。また、基本性能測定を行い、その過程での改良点も抽出している。しかしながら大規模アプリケーションへの対応・検証がほとんど進んでおらず、システム側の目標達成率としては 30%ほどである。

今後は、性能が出づらい実装手法をとった部分の改良を行う。本システムの実運用に当たっては、ジョブが複数の拠点で同時に動くことが仮定できないため、ジョブ実行開始時刻のずれが全ジョブの終了時刻に大きく寄与すると考えられる。このため、送信側・受信側のプロセス間で同期が発生しないようなシステム設計を行っている。よって本システムによるコストは本質的ではないが、受信側のジョブ終了時刻は送信側のデータ転送が終了しているかどうかに関わっており、なるべく帯域の性能を使い切れるようチューニングを進めていく。また、より高度な目標として設定していた 3 拠点連携についての設計も引き続き進める。

先にも述べたが、東大・九大とも本システムを用いた大規模アプリケーション検証に関しては、システムソフトウェアの実装遅延により従来アプリケーションの改造中であり、大規模な性能測定に至らなかった。本件に関しては今後も進めていく予定である。

OpenFOAM に対するポスト処理連携の研究開発面においては、実装方法を FunctionObjects を利用するものに変更して egcpops の試験を行った。短時間小規模な疎通テストは確認できたが、基盤センターでのサービスとして提供するまでには検証が進んでいない。

今後は、egcpops の特徴・先進面である多拠点連携機構と、これまで OpenFOAM に独自に実装してきた NStream クラスを組み合わせて、介在させるファイルの置き方を再検討し、可能であればファイルを介在させない方法の検討を進める。

7. 研究成果リスト

(1) 学術論文

1. K. Takahashi, S. Iwagami, T. Kobayashi, T. Takami, “Theoretical Estimation of the Acoustic Energy Generation and Absorption Caused by Jet Oscillation”, J. Phys. Soc. Jpn., Vol.85, No.4, Article ID: 044402

(2) 国際会議プロシーディングス

(3) 国際会議発表

1. T. Kobayashi, “Uncertainty and Dynamical Process on Computation”, International Workshop on Advanced Future Studies, Mar. 14-16, Kyoto, Japan, 招待講演
2. T. Kobayashi, S. Iwagami, T. Takami, K. Takahashi, “Vortex Sound from a wavy jet and Howe’s energy corollary”, XXXV Dynamics Days Europe, P.3, 6-10 September 2015, University of Exeter, UK, 査読付
3. T. Takami, M. Shimokawa, T. Kobayashi, “Multiscale descriptions for nonlinear dynamics”, XXXV Dynamics Days Europe, P.1, 6-10 September 2015, University of Exeter,

UK, 査読付

4. T. Kobayashi, Y. Morie, H. Jitsumoto, T. Takami, M. Aoyagi, “A New Bottleneck in Large-Scale Numerical Simulations of Transient Phenomena, and Cooperation Between Simulations and the Post-Processes”, PANACM 2015, 1st. Pan-American Congress on Computational Mechanics, [3.17 MS: High Performance Computing and Related Topics I], April 27-29, 2015, Buenos Aires, Argentina, 査読付
5. S. Iwagami*, G. Tsutsumi, K. Nakano, T. Kobayashi, T. Takami, K. Takahashi, “Numerical Analysis on the Lighthill Sound Sources of Oscillating Jet”, PANACM 2015, 1st. Pan-American Congress on Computational Mechanics, [Contributed Session on Advanced Methods in Computational Fluid Dynamics II], April 27-29, 2015, Buenos Aires, Argentina, 査読付
6. Hideyuki Jitsumoto, “A Framework for Jointing Computation Center with User-Level Management System”, 2nd Annual Meeting on Advanced Computing System and Infrastructure (ACSI2016), Jan 2016, 査読無し

(4) 国内会議発表

1. 岩上翔, 堤元気, 小林泰三, 高見利也, 高橋公也,「エッジトーンの発音機構における数値解析」, 19aBU-7, 日本物理学会 第 71 回年次大会, 2016 年 3 月, 東北学院大学
2. 馬場玲於, 三木晃, 鬼束博文, 宮川矩昌, 岩上翔, 堤元気, 松清一樹, 小林泰三, 高見利也, 高橋公也,「フルート歌口近傍の流体音響解析 II」, 20pPSA-60, 日本物理学会 第 71 回年次大会, 2016 年 3 月, 東北学院大学
3. 松清一樹, 野見山亮, 馬場玲於, 岩上翔, 小林泰三, 高見利也, 高橋公也,「音孔のついた閉管の共鳴状態の流体音響解析 II」, 20pPSA-61, 日本物理学会 第 71 回年次大会, 2016 年 3 月, 東北学院大学
4. 高橋公也, 小林泰三,「多重遅延系のモード選択則」, 22pBU-8, 日本物理学会 第 71 回年次大会, 2016 年 3 月, 東北学院大学
5. 岩上翔, 堤元気, 小林泰三, 高見利也, 高橋公也,「エッジトーンにおける流体音源の数値的評価」, 第 29 回数値流体力学シンポジウム, A04-2, 九州大学, 2015 年 12 月
6. 實本 英之, ユーザ駆動型・大域アプリ連成フレームワーク, アカデミッククラウドシンポジウム, Sep. 2015.
7. 實本 英之, 小林 泰三, 松本 正晴, 滝澤 真一朗, 三浦 信一, 多拠点連成アプリケーションを実現するユーザ駆動型・拠点連携システム, 研究報告ハイパフォーマンスコンピューティング (HPC), 2015-HPC-151, Sep 2015
8. 小林泰三, 「フルートの発音 ～流体音入門～」}, 研究会「非線形現象の捉え方」, 2015 年 10 月 9~11 日, FIT セミナーハウス (湯布院)
9. 小林泰三, 岩上翔, 高見利也, 高橋公也,「エアリード楽器の流体音源と Howe's energy corollary」, 17aCW-1, 日本物理学会 2015 年秋季大会, 2015 年 9 月, 関西大学
10. 岩上翔, 堤元気, 小林泰三, 高見利也, 高橋公也, 「エッジトーンにおける Lighthill 音源の数値的評価」, 17aCW-2, 日本物理学会 2015 年秋季大会, 2015 年 9 月, 関西大学
11. 高橋公也, 岩上翔, 小林泰三, 高見利也,「トイモデルを用いた振動するジェットからの音響エネルギー発生メカニズムの解析」, 17aCW-3, 日本物理学会 2015 年秋季大会, 2015 年 9 月, 関西大学
12. 松清一樹, 野見山亮, 馬場玲於, 岩上翔, 小林泰三, 高見利也, 高橋公也, 「音孔のついた閉管の共鳴状態の流体音響解析」, 17pPSA-83, 日本物理学会 2015 年秋季大会,

2015 年 9 月，関西大学

13. 北崎祥一，馬場礼於，松清一樹，岩上翔，堤元気，小林泰三，高見利也，高橋公也，「バスレフポートスピーカーのポートノイズの流体音解析」，17pPSA-84，日本物理学会 2015 年秋季大会，2015 年 9 月，関西大学
14. 馬場玲於，三木晃，鬼束博文，宮川矩昌，岩上翔，堤元気，松清一樹，小林泰三，高見利也，高橋公也，「フルート歌口近傍の流体音響解析」，17pPSA-85，日本物理学会 2015 年秋季大会，2015 年 9 月，関西大学
15. 高橋公也，小林泰三,「3 重遅延系のモード選択則」，18pCQ-11，日本物理学会 2015 年秋季大会，2015 年 9 月，関西大学
16. 小林泰三，「動的過程の計算論」，京都大学基礎物理学研究所研究会「複雑システムにおける創造的破壊現象の原理に迫る」，2015 年 8 月，京都大学

(5) その他（特許，プレス発表，著書等）

1. https://github.com/RyzeVia/exgcoup
   開発ライブラリを GitHub にて公開