[2007]九州大学情報統括本部年報 : 2007年度

https://doi.org/10.15017/1470731

出版情報：九州大学情報統括本部年報. 2007, 2008. 九州大学情報統括本部
バージョン：
権利関係：

第6章　プロジェクト紹介

6.1　ペタスケール・システムインターコネクト技術の開発

プロジェクトの概要

PSI プロジェクトとは, 文部科学省の委託事業「次世代 IT 基盤構築のための研究開発」, 研究開発領域「将来のスーパー コンピューティングのための要素技術の研究開発」（平成１７年度～１９年度）に採択された研究開発課題「ペタスケール・システムインターコネクト技術の開発」に関するプロジェクトである.

本プロジェクトでは, ペタフロップス超級スーパーコンピュータシステムの構成において数千～数十万規模の高速計算ノードを相互結合するシステムインターコネクト技術を対象に, 現状のシステムよりもコスト対性能比で 1 桁上を目指して, 高性能化, 高機能化, 低コスト化を同時に達成するための３つの要素技術, すなわち, 光パケットスイッチと超小型光リンク技術, 動的通信最適化による MPI 高速化, システムインターコネクトの総合性能評価技術を, 九州大学の他に財団法人福岡県産業・科学技術振興財団, 財団法人九州システム情報技術研究所, 富士通株式会社と共同で開発した.

光技術を用いた超高バンド幅スイッチング技術の開発

- 「光パケットスイッチング技術の開発」:

  ペタフロップス級の次世代計算機を構築する際には, そのインターコネクト部の物量, 消費電力がシステム構築する際の大きな問題のひとつである. スイッチ装置における OE/EO モジュール数を削減する手段として, 光信号を光のまま切り替える光スイッチ技術を導入することが, 幹線網の技術において導入が開始されているが, 切り換え速度が遅く, ペタスケール・システムインターコネクトに必要なナノ秒オーダの高速スイッチング特性と数十～100 ポート以上の多ポート化を両立できる, 光パケットデータ転送, スイッチ制御技術, スイッチハードウェア技術については確立されていない.

  本研究は, 上記の課題を解決するために, 異なる信号が変調されている複数の光波長を, 光の領域で多重, 分離する波長分割多重（WDM：Wavelength Division Multiplexing）技術によるテラビット/秒級の超高バンド幅インターコネクションと, 各計算機ノードで WDM 光パケット信号を生成し, スイッチ部分において, WDM 信号を OE/EO 変換することなく光信号のままでの一括切替が可能な光電気ハイブリッドスイッチ (以下, 光パケットスイッチと記す) の要素技術を開発することを目的とし, 広帯域波長多重光パケット信号を高効率で転送, 交換を可能とする新しい光スイッチング方式の要素技術確立を目指した.

  本年度において, 1.2 マイクロ秒長の光パケット信号を用いて, イーサネットフレーム—光パケット信号の変換機能および, 2 台の計算ノード間でのデータ転送動作を確認し, 高バンド幅光パケットスイッチシステムの要素技術を確立した.

- 「光電気変換部集積化技術の開発」:

ボード（プリント基板）間インターコネクトの光化のためにはノードのボード端部に光送受信部を多数配置する必要があるが，従来技術ではその大きさのため実現が困難である．本研究ではボード間インターコネクトを数十ギガビット／秒に高バンド幅化でき，かつ，現状最新の光送受信部（XFP：10 Gigabit Small Form Factor Pluggable）に比べ 1/10 以下の占有面積とコストを可能にする光電気変換部集積化の要素技術を開発した．

本年度は光電気変換部の小型化と広い温度範囲での動作を得るために，放熱とクロストークに関する設計改善，及びボード上の固定機構を省スペース化する実装技術開発を行った．さらにこの技術を適用し，10Gbps × 4ch の光送信モジュール及び光受信モジュールを試作して特性を評価した．その結果，光電気変換部を「長さ 8.5mm ×幅 8mm ×厚さ 4mm」に小型化し，環境温度 0〜70 ℃での良好なアイ開口波形と高感度の符号誤り率特性を得た．また，筐体カバーに押圧用バネを内蔵させてボード上での着脱を可能にした．以上より，光電気変換部の目標サイズ（長さ 10mm ×幅 10mm ×厚さ 5mm）を達成した．これは，業界標準品では最小となる SFP +（8.5 and 10 Gigabit/s Small Form Factor Pluggable）光モジュールのサイズ（長さ 56mm ×幅 13.4mm ×厚さ 12mm）に比べて 1/10 以下に相当する．本サブサブテーマで開発した技術は，テラビット級光データ伝送に必要な光電気変換部（例：10Gbps × 4ch 光モジュール 25 個）を標準的サイズ（例：約 40cm 角）のボード上に搭載可能とする．

高機能・高性能システムインターコネクト技術の開発

- 「コレクティブ通信をサポートする高機能スイッチの開発」：

  ノード間の通信経路上で通信データに対して種々の演算を施すことが可能なスイッチのハードウェアを設計し，これを搭載した高機能スイッチ装置の開発を行う．本高機能スイッチ装置を用いて，総和をはじめとする種々のコレクティブ通信機能をハードウェアにオフロードし，アプリケーション全体の高速化を狙う．本研究では，当該高機能スイッチを用いた数十ノード規模の評価システムを構築し，コレクティブ通信を従来比で 5 倍以上高速化することを目標に試作機での性能検証を行った．

  本年度は，高機能スイッチハードウェアおよび MPI ライブラリ（文献 (1)）を含むソフトウェアスタック（ドライバ，高機能スイッチライブラリ，MPI ライブラリの階層構造）の動作を確認した．実際にシステムとして設計し動作を確認したことにより，ソフトウェアのインターフェース設計についての基盤を固めた．また，高機能スイッチによるコレクティブ通信の性能測定と評価高機能スイッチを 17 ノードのシステムに接続した環境で性能測定を実施した．事前見積もり通りの性能となっており，高機能スイッチが，大規模システムへの適用に際して有効である事を確認した．

- 「動的最適化を用いた MPI 高速化技術の開発」：

  ペタフロップス級の次世代計算機のように大規模な計算機システムでは，割り当てられる計算資源の配置や他のジョブの存在等によって，プロセス間通信の性能が大きく変動すると予測される．本研究では，そのような状況でも並列処理性能を維持するための技術として，適応的に通信パラメータ等を調整する動的通信最適化技術の開発を行った．

  本年度の主な成果は以下の通りである．

  1. 実行時の性能に応じて動的に全体通信のアルゴリズムを選択する高速化技術遅いアルゴリズムを早い段階で選択の候補から外していくことによって効率良くアルゴリズムの選択を行う技術を実装した．実験の結果，図 6.1 に示す通り，提案手法がほとんどの場合で最適なアルゴリズムを自動的に選択することを確認した．

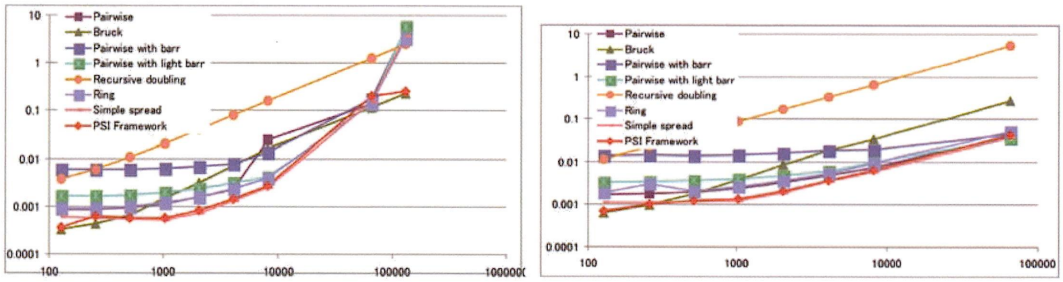


図 6.1: メッセージサイズに応じた MPI_Alltoall の所要時間 (左: 64 並列, 右: 128 並列)

2. 通信タイミングを考慮し, 通信衝突を回避するランク配置最適化技術昨年度ツリートポロジ向けに開発した最適化技術を IBM BlueGene/L の 3D メッシュトポロジに適用し, 実装して評価した. 実験の結果, 最適化を行わない場合に対して通信の所要時間を大幅に削減できることを示した. また, 最適なランク配置を探索する発見的アルゴリズムについて提案し, 特に通信の頻度が多い場合に従来手法に対して, より短時間により衝突の少ないランク配置を見つけることが出来ることを示した.
3. 昨年度開発した, 負荷の状況に応じて全体通信内部の一対一通信の順序を調整する技術について, MPI プログラムから呼び出し可能なライブラリとして整備し, 性能評価を行った. また, 圧縮格納された疎行列同士の乗算を行うプログラムに適用して効果を確認した.

ペタスケール・システムインターコネクトの性能評価環境の構築

- 「ペタスケール・アーキテクチャの開発」:

  本研究では, 昨年度までに, 高演算性能, 高エネルギー効率を目指す SIMD 拡張部を持つプロセッサアーキテクチャの開発, および, FORTRAN ソースコードのコンパイルから, SIMD 計算ノードの動作をシミュレートし実行時間を評価することが出来る計算ノードシミュレータ PSI-PSIM までの一連のツールの開発を行った.

  今年度は, 前年度までに実施した SIMD 計算ノードアーキテクチャの定義, 性能評価ツールに加えて, 電力モデルと消費電力の評価機能を開発した. この機能を用いて演算あたりの消費電力を求め, SIMD 拡張アーキテクチャにより, 演算あたりの消費電力を 1/2〜1/4 に削減できることが確認された. これにより, 提案のアーキテクチャは, 本研究の目標である高性能で高効率のペタスケールシステム向きのプロセッサアーキテクチャであることが定量的に確認できた. また, ペタスケール級のアプリケーションである HPL と PHASE について, 小規模なシステムで高速に実行できるスケルトンコードを開発した. これらのスケルトンコードをサブサブテーマ 3- 開発の BSIM を用いて実行することにより, これらのアプリケーションを, SIMD 拡張アーキテクチャの計算ノードを用いるペタフロップス級のシステムで実行する場合の性能を予測した. これにより, サブテーマ３の性能評価メソドロジにより, ペタスケールシステムにおけるアプリケーション実行性能の予測が可能であることを示した. 本研究で開発した環境をベースとして拡充することにより, ペタスケールシステムの開発に際して計算ノードアーキテクチャ設計やインターコネクト設計のトレードオフの検討が可能になると考えられる.

- 「ペタスケール・システムの性能予測技術の開発」:

  サブサブサブテーマ３－ －（１）では, これまでに開発した各種ツールの完成度を高めると共に, ペタスケール・アプリケーションを対象とした性能予測実験を実施した. 具体的には, 次

世代のスーパーコンピュータを念頭におき, 性能評価用ベンチマーク・プログラムの作成, ならびに, ペタスケール・システム性能予測を行った.

その結果, 以下の研究成果を達成した.

1. テラスケール小規模ホストマシンを用いたペタスケール大規模ターゲットマシンの性能予測可能性の実証

   ペタスケール・ターゲットの性能予測を可能にすべく, 仮想超並列実行環境を中心とした新しい性能予測フローを考案した. また, サブテーマ３－　で考案した仮想ペタスケール・スーパーコンピュータを対象に性能予測実験を行い, ピーク性能 2.1PFlops のターゲット性能を 1.6TFlops もしくは 200GFlops のホストマシンで予測した結果, HPL においては実効性能が 1.01PFlops（3D-Torus インターコネクションを想定した場合）であり, 性能予測に要した時間は約 6 時間程度であった. また, PHASE プログラムに関するペタスケール性能予測も行い, 約 4 時間を費やすことでその性能が 0.65PFlops であることを示した. これらの実験により, テラスケール小規模ホストマシンを用いたペタスケール大規模ターゲットマシンの性能予測が可能である事を実証した.

2. 自動スケルトンコード生成可能性の実証

   昨年度までに開発したフロントエンド部ならびにバックエンド部, そしてコンパイラ・フレームワークである COINS を統合して PARSER を完成させた. また, 本ツールを用いて NAS Parallel Benchmark FT の連接構造部分を対象とし, 自動でスケルトンコードを生成した. 人手で作成した FT スケルトンコードと比較した結果, 精度低下は見られるものの, 性能予測誤差が 36 あった. これにより, 限定的ではあるがスケルトンコードの自動生成が可能であることを実証した.

3. プログラム実行を俯瞰可能な大規模システム向け表示/解析ツールの開発

   インターコネクション・ネットワークにおいて通信混雑が発生するプログラム実行期間を高速に特定するための通信検索エンジンを開発した. また, これをグループワークビューアに統合し, プログラム実行全体を俯瞰/解析可能な大規模表示/解析ツール ANA を完成させた. これにより, ペタスケール級の大規模アプリケーション開発, ならびに, システム開発において利用可能な新しい表示/解析法を示した.

4. 高速かつ正確な仮想超並列実行環境 BSIM の開発小規模ホストマシンでの高速なスケルトンコード実行を可能にするため, 実際の通信を行うことなしにその実行結果（通信時間）を通信プロファイルに反映する機能を追加した. これにより, より高速なスケルトンコード実行を可能にした. 特に, 本機能は通信を多く含むアプリケーション・プログラムの実行において有効である. また, 与えられた通信遅延情報に基づき BSIM 仮想タイマを更新する機能を実装した. これにより, BSIM においても通信を考慮した通信プロファイルの生成ならびに実行時間予測を可能にした.

5. PSI-SIM による高精度な性能予測実現可能性の実証

   200GFlops（または 50GFlops）の小規模ホストマシンを利用して, 6.5TFlops 大規模ターゲットマシンの性能予測実験を実施した. その結果, 多くのベンチマーク（FMO-ERI, HPL, PHASE）に関して, 性能予測誤差が 4～10 %程度と極めて低いことを示した. また, 性能予測に要する時間に関しては, ターゲットマシンでのオリジナルコード実行時間の半分程度の場合もあることを示した（FMO-ERI, HPL）. これらの結果より, PSI-SIM を用いることで, 2～3 桁大規模なターゲット・システムの性能予測を高速かつ正確に行えることを実証した.

6. ペタスケール・インターコネクト・シミュレータの開発

実用時間内に精度の高いシミュレーションを実施できるよう，実行性能の改善を図り，最大で 68.9 倍の性能向上を達成した．また，性能予測精度を高めるフリットレベル・シミュレーション機能の実装を行い，実機レベルのインターコネクトをモデリングできるようにした．次いで，通信プロファイルを分割し，部分読込み実行方式を実装することでシミュレーションに必要なメモリ容量を大幅に削減し，数十万ノード規模のシミュレーションにも対応するよう機能拡張を行った．そして，20G バイトの巨大通信プロファイルを用いたシミュレーションや 4,096 ノードを持つ大規模インターコネクトのシミュレーションを実施し，実用的な時間内で良好な性能を得られることを実証した．

- 「ナノアプリケーションのコード分析と超並列化」

本研究では，次世代スパコンのためのターゲットアプリケーションの 1 つである FMO プログラムの超並列化を目標にして，超並列化を行うための並列化手法を決定して，その方針に従って，超並列化 FMO プログラムの実コードの作成を行った．その結果，本年度は，以下のような成果を得た．

1. 超並列化を行う際の並列化手法の検討

   FMO 計算におけるホットスポットの 1 つである 2 電子積分ルーチンに対して，MPI を用いたプロセス並列化，OpenMP を用いたスレッド並列化，ならびに，これら 2 つを組み合わせたハイブリッド並列化，を行い，その並列性能の比較を行った．その結果，プロセス並列化のみを用いた場合よりも，スレッド並列化を用いた（併用した）方が，並列性能が高いことがわかった．また，超並列 FMO 計算では，グループサイズを大きくする必要があり，その場合，複数の計算ノードを利用することが一般的であるため，プロセス並列化とスレッド並列化を組み合わせたハイブリッド並列化を行うことで，超並列 FMO プログラムを作成することに決定した．

2. 超並列 FMO 計算を行うための実プログラム作成

   ハイブリッド並列化による並列 FMO プログラムを作成するために，FMO 計算で用いる，スレッドセーフな各種積分ルーチンを作成した．これらの積分ルーチンを，昨年度までに作成済みの OpenFMO のスケルトンコードに導入することによって，FMO 計算を行うための実コードを作成した．

3. スケルトンコードの評価

   大規模システム性能評価を実現すべくベンチマーク・プログラムを開発した．具体的には，FMO-ERI ならびに Open-FMO の 2 つのスケルトンコードを開発し，6.5TFlops ターゲットマシンを対象とした性能予測実験を実施した．その結果，FMO-ERI に関しては性能誤差が 4% 程度と極めて高精度に性能予測が可能である事を示した．一方，Open-FMO に関しては性能予測誤差が大きかったものの，モノマーとダイマーに関する抽象化箇所の実行時間見積りを個別に行うことで精度を向上できる見通しを立てた．

6.2　次世代検索エンジン・プロジェクト

プロジェクト概要

本プロジェクトは科学技術振興機構による大学発ベンチャー創出推進事業として平成１８年度に採択され、課題番号：1812、研究開発課題名：「データマップ法と概念グラフによる次世代検索エンジンの研究開発」として、３年間の期間で実施しているプロジェクトです。

インターネットの普及により、誰もが大量の文書情報に直面しなければならなりません。利用者は、知りたい情報があれば検索エンジンを用いて調べますが、ほとんどの場合、検索結果の先頭に現れるいくつかの文書しか閲覧しません。また、情報提供側は、伝えたい情報があっても、検索結果の上位に表示させるために連動広告や検索エンジン最適化が必要になっています。しかしながら、ユーザーが意図した情報と違う情報が現れるケースも増えてくると同時に、検索結果の先頭に現れない情報の中にも新たな気づきにつながる有益な情報が隠れていることが多くあります。

プロジェクトの目標は、言葉のつながりによる検索結果全体の把握と絞込みを支援する概念グラフに基づく規模耐性のある高速検索エンジンの開発です。概念グラフは、単語間に共起頻度による順序関係を与える単純な原理ですが、把握と絞込みの効率についての精度の保証を与え、理論的にも競合優位性を立証することも目標としています。具体的にはマトリックスと概念グラフというテキストマイニング技術を用いる検索エンジンを開発しています。

研究の特徴

大量の情報に対する次世代検索エンジンの中核技術として期待されています。本研究の特徴は、全体像の効率的な把握と効率的な絞込みの支援にあります。本プロジェクトでは、言葉のつながりとして大量文書群を可視化するシステムを開発しています。言葉のつながりによって大量の文書を俯瞰的に表示することで、利用者に従来の情報の中から新たな気づきを与え、より多くの情報が有効活用されることが期待できます。

2.1 マトリックス検索エンジン

従来の検索エンジンでは、ユーザーのキーワードに対する検索結果を順位付きリストのような一次元表示しかできません。我々のマトリックス検索システ ムでは、複数の観点からの分析を実現します。ユーザーは２つの観点を選択すると、それぞれの観点に基づいた分類が行なわれ検索結果は２次元マップとして表 示されます。表示された各セルには、観点ごとに特徴語が自動的に抽出されているので、セルの解釈だけでなく、検索拡張や絞り込みのためのキーワードとして 利用できます。さらに、各セルをクリックするだけでズーミング検索ができます。

2.2 概念グラフ

概念グラフは文書群に現れる特徴語の概念的な上下関係を表す可視化システムです。ユーザーが入力するキーワードを含む文書群を検索結果 として求めるところまでは通常の検索エンジンと同じですが、コンセプト・グラフ・エンジンは検索結果のこれらの文書に含まれる特徴語を抽出します。さら に、それら特徴語の間の概念的な上下関係を出現頻度にもとづき抽出し、グラフとして可視化します。検索結果が数百件あったとしても、それらの文書でどのよ うなことがポイントになっているか大局的につかむことができます。具体的な文書を読む前に、あるいは、検索結果をじっくり読んだ後で、概観をまとめたり分 析するための強力なツールとなります。

検索結果可視化の従来手法の多くは、文書あるいは単語を一つの点として表示します。二つの対象の関係は、類似度を距離として力学的バネモデルにより平面上に配置するため高速ですが、つながりの方向には意味がない可視化でした。自己組織化モデルによるクラスタリング表示するものもあり、見やすい分類としては有効な可視化ですが、絶対座標には論理的な意味付けがありません。確率的因果関係によるベイズ・ネットなどでは方向性はありますが、表示する単語群を人手で事前に固定しなければなりません。本プロジェクトで使う概念グラフでは、単語の頻度情報という単純な統計指標にもとづき、特徴語の自動抽出と、特徴語間の上位下位関係化も抽出も自動的に行い、可視化します。これにより、検索結果は単語群の有向グラフとして可視化され、効率的な全体像把握と、効率的な絞り込みのための支援が実現できます。

本年度の研究成果

研究の成果として以下のような成果を得ることができました。**マトリックス・エンジンについては、**AND、OR、NOT、検索範囲指定、複数検索条件設定などの機能を追加することができました。クラスタリングについては、数値項目、地名、分類記号などの固定属性値のクラスタリングを実装しました。頻度にもとづき抽出した特徴語だけを使う文書分類法を考案し特許出願して理論的評価実験を進めています。

概念グラフの理論的評価として、シソーラス自動構築の他手法と階層構造の被覆率(図1)、階層の粒度という定量的比較(図2)を行い、本手法が優位であることを検証し、その結果を国際会議で発表しました。

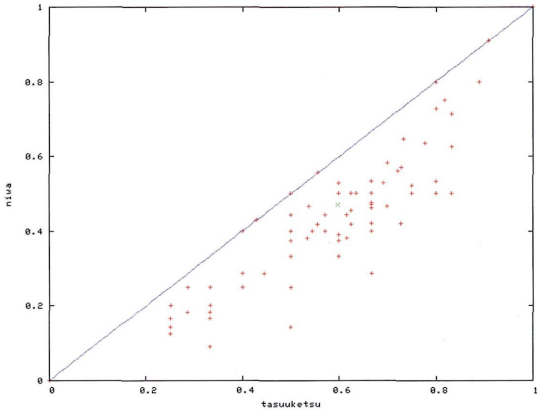


図1丹羽の手法との比較(被覆率)

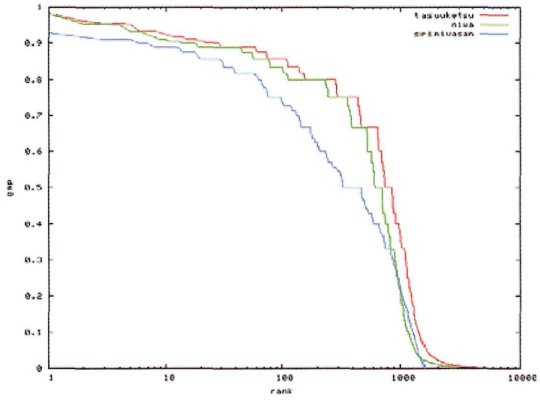


図2丹羽、Srinivasan の手法との比較(粒度)

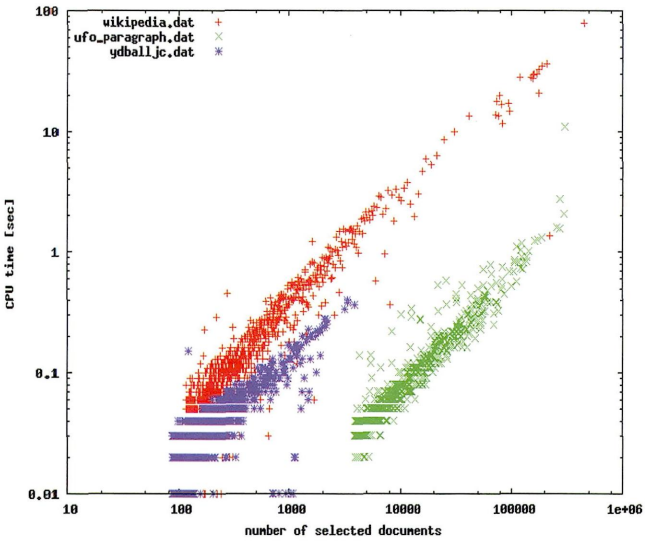
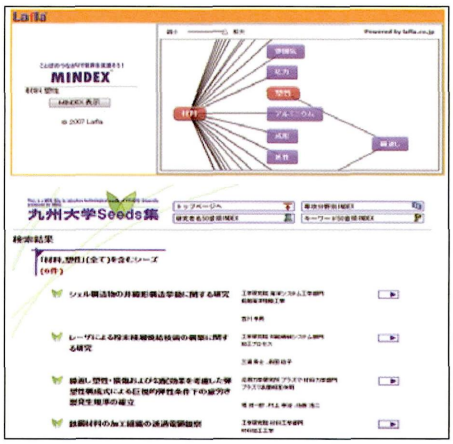


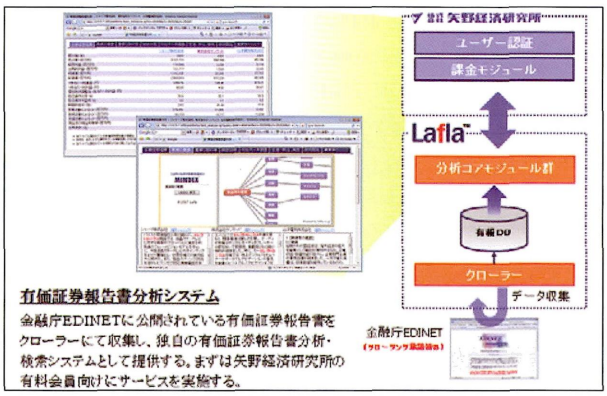
図3　グラフ生成時間

高速化・コンパクト化を目的として従来の Perl 版を全面的に C 言語にて書き換えました。1 万件の文書群に対する概念グラフを 100 ミリ秒で求めることができる規模耐性のある高速エンジンが完成しました。これは、当初の目標であったシンクタンクの白書に対する検索エンジンとしての実用レベルの性能を満たしています。図3は、Wikipedia(赤)、レポートデータ(青)、有価証券報告書(緑)などの実データでの処理速度を表したものです。

これにより、実用化の目処が見えてきました。例えば、平成19年8月から矢野経済研究所の市場調査レポートの検索システムとして公開しています。また、平成20年2月には、有価証券報告書検索システム UFO Lenz(図5)も公開しました。また、九州大学研究 seeds 集(図4)を2月に公開しました。詳しくはプロジェクトホームページ http://www.lafla.co.jp をご覧ください。

図　4　　有　報　　Lenz





http://www.ydb.jp/ufolenz

図5九州大学 Seeds 集

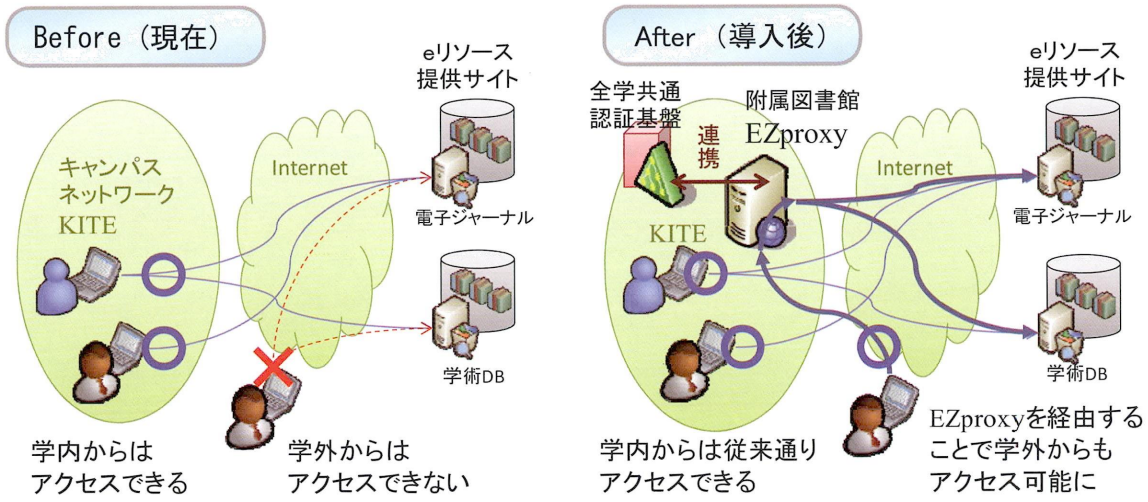
http://kyudaiseeds.lafla.imaq.kyushu-u.ac.jp

6.3　図書館との連携による学外からのｅリソース閲覧

大学活動の主な部分をしめる研究活動では，関連研究の調査やサーベイは必須であり，そのために既に発表された学術論文を調査することが日常的に行われてきた。近年，論文の電子化が進み電子ジャーナルとして利用されるようになってきた。電子ジャーナルの他にも電子化した電子ブックや学術資料データベースなど，様々な学術情報資源がオンラインで利用できる。これらの電子化された学術情報資源をｅリソースと呼んでいる。九州大学附属図書館では，全国の大学に先駆けてｅリソースの充実化を進めており，平成１９年度末には約４万タイトルのｅリソースが閲覧可能になっており，このタイトル数は世界有数の整備状況であると言える。

電子ジャーナルなどのｅリソースにより，論文や学術資料の閲覧場所は図書館以外にも拡張したものの，ネットワーク的な理由から閲覧場所の制約が残っている。現在，ｅリソースを提供する企業では，利用契約をした大学の構成員へのみ閲覧を許可しており，そのための閲覧制限にはインターネットの IP アドレスに基づく制御を用いる場合が多い。これは，大学はインターネットの初期から接続されていたためグローバルアドレスを保有する場合が多いことと，構成員のアカウント情報を一元管理する大学が少ないためであろう。実際，九州大学附属図書館が契約しているｅリソース提供企業でも，九州大学の保有する IP アドレスブロック（133.5.*.*/16）からのアクセスには情報を提供するように制御するものが多い。

このようなアクセス制御がなされているため，出張や留学などで九州大学のキャンパス外にいる場合は，九州大学の構成員であってもｅリソースが閲覧できなかった。そこで，九州大学が契約しているｅリソースを，学外からでも閲覧可能にする環境を整える活動を，附属図書館と情報統括本部が連携して平成 19 年度から開始した。具体的には，Web プロクシ機能と，利用者認証機能とを組み合わせた仕組みを構築している。



利用者認証では，情報統括本部が提供する全学共通認証サービスを用いる。情報統括本部では，全学構成員のアカウント情報を一元的に管理し，かつ利用者認証機能を学内の情報サービスへ提供する全学共通認証サービスを平成１９年９月から開始した。学内の構成員であれば，職員は SSO-KID と呼ばれる全学共通 ID が配布されている。学生，大学院生および研究生は，学生番号に基づく ID を保有している。利用者の ID には対応するパスワードが設定されており，それを用いて利用者認証を行うことができる。

Web プロクシは，EZproxy と呼ばれるソフトウェアシステムを購入し，附属図書館側で，全

学共通認証基盤との連携や，各電子ジャーナル提供サイトとの連携についての設定を行っている。また，学内の構成員に分かりやすいWebサイトの構築も行っている。

残念ながら，平成１９年度内は準備と試験だけで終わり，学内へのサービス開始にいたっていない。次年度の平成２０年度には試行的にサービスを開始できるよう準備を進める予定である。