Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

# クラウド環境によるOpenNSIMインターコネクトシミュレーションサービス

柴村英智[†1]　薄田竜太郎[†1]　平尾智也[†1]
吉田真[†1]　神戸隆行[†2]　三輪英樹[†3]
三吉郁夫[†3]　井上弘士[†4]　村上和彰[†1,†4]

次世代の大規模インターコネクトの性能評価を目的としたインターコネクトシミュレータ OpenNSIM を開発した．OpenNSIM は TaaS と呼ぶクラウド環境で動作し，Web ブラウザから利用可能である．トポロジ，ノード数，評価アプリケーションなどの煩雑なシミュレーションパラメータの設定は，GUI によって指定・選択でき，シミュレーション後の評価データも Web ブラウザから確認できる．FatTree や 2 次元，3 次元のメッシュ／トーラス網をサポートしており，8 ノードから 128K ノード構成までのインターコネクトの幅広いシミュレーションが可能である．また，実機に近い設定や将来の大規模システムを想定した設定を施したシミュレーションを行うことで，システム規模の変化にともなうアプリケーション実行性能の変化を推定したり，詳細シミュレーションによってプログラムのデバッグや性能チューニングの支援にも利用できる．本稿では，OpenNSIM と TaaS の概要，ならびに評価事例を報告する．

# OpenNSIM Interconnect Simulation Service via a Cloud Environment

Hidetomo Shibamura ,[†1] Ryutaro Susukita ,[†1]
Tomoya Hirao ,[†1] Makoto Yoshida ,[†1]
Takayuki Kando ,[†2] Hideki Miwa ,[†3] Ikuo Miyoshi ,[†3]
Koji Inoue [†4] and Kazuaki Murakami [†1,†4]

We have developed an interconnect simulator OpenNSIM for performance evaluation of next-generation large scale interconnects. OpenNSIM works on a cloud environment TaaS and setting of simulation parameters such as topology, number of nodes, evaluation application, and access to evaluation reports after simulation can be done from a web browser. OpenNSIM supports fat-tree, 2D-/3D- meshes, and tori and can simulate varisized interconnect up-to 128K nodes. We also can estimate change of execution performance of application through some simulations assumed production-like or extreme-scale system. In addition, detailed simulation helps us program debugging and/or performance tuning. In this paper, the overview of OpenNSIM and TaaS are described and some evaluation results are reported.

## 1. はじめに

高い処理性能を発揮する大規模並列システムを実現するには，プロセッサ要素のみならず，それらを搭載した多数の計算ノードを高い通信効率で相互接続するインターコネクト（相互結合網）の設計開発が重要である．インターコネクトの基本的な通信性能は，トポロジ（網形態），ノード数，リンク通信性能によって特徴づけられ，これらのパラメータを，開発当時に利用可能な技術水準や，製造時のコストを踏まえて適切に設定しなければならない．ここで，多数のパラメータの組み合わせから成る広大な設計空間から，所望の性能を達成する最適なパラメータセットを導くためには，大規模なコンピュータシミュレーションが不可欠であり，これまでに様々なインターコネクトシミュレーションが行われてきた．

近年の大規模並列処理では，科学技術計算をはじめとする高度に複雑化されたプログラムを実行するため，計算ノード間で発生する通信のパターンはプログラム毎，さらにはプログラム内の通信ブロック毎に異なる．また，インターコネクトの通信性能が高くとも，通信混雑が頻繁に発生する通信パターンでは通信遅延時間が大きくなりアプリケーションの実行性能が十分に出ない．すなわち，インターコネクト設計時には，単に物理的な通信性能だけでなく，様々なプログラムに内在する通信パターンとの親和性も要件としなければならない．

今後，計算ノード数が数十万規模に増加するとインターコネクトの性能がアプリケーションの実行に与える影響はいっそう大きくなるため，インターコネクトの通信性能とアプリ

†1 (財) 九州先端科学技術研究所 次世代スーパーコンピュータ開発支援室
Institute of Systems, Information Technologies and Nanotechnologies
†2 クオリアークテクノロジーソリューションズ株式会社
Quoliarc Technology Solutions Ltd.
†3 富士通株式会社
Fujitsu Ltd.
†4 九州大学
Kyushu University

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

ケーションの通信パターンを併せた定量的な性能評価は，ペタ・エクサスケールコンピューティング時代に向けた重要な課題といえる．

そこで我々は，数十万ノードを接続する大規模インターコネクトの性能評価を目的とし，並列プログラムを駆動させながら通信パターンを生成する，インターコネクトシミュレータ NSIM を開発した[1)]．NSIM は，設計・開発現場での利用を念頭に，現実的な時間内でシミュレーションを完了することを目指しており，すでに各種の評価実験を行っている[2)]．

近年では，インターネットを利用したソフトウェアツールの提供が盛んに行われており，プロセッサやシステム LSI などの開発ツールをはじめ，各種のシミュレータ，解析ツール，ならびにベンチマークなどが各サイトから広く公開されている．また，集合的な公開サイトとしては 3) がある．

ツール利用者側の視点では，このようなツールを利用するためにはダウンロード，ビルド，インストールといった煩雑な作業を伴うことが多く，ツールの実行環境や仕様・制約についても留意しなければならない．また，これらの作業はツールの試用や評価を目的とした場合でも発生し，場合によっては実行環境への初期投資コストも必要となる．

一方，ツール提供側は，デモや期間を定めた試用によって，開発したツールの普及促進や新規利用者の開拓を図ることが多い．しかし，前述のような作業や負荷を利用者側に強いることになる．また，機能追加やバグ修正などによるソフトウェア更新が発生すると，利用者の手間暇も増え負担はさらに大きくなる．このような問題は，例えば，新しいハードウェアの設計開発と並行して，開発ツールを頻繁に更新するような場合には顕著となる．

近年，新規にハードウェアを設計開発する際は，専用のアプリケーション開発ツールや評価・解析ツールといった支援ツールも並行して開発することが多い．優れたハードウェアを開発していたとしても，適切な支援ツールが提供されなければアプリケーションの開発やハードウェアの評価ができず，機能拡張や性能改善に向けた知見を得ることもできない．したがって，ハードウェアと支援ツールの足並みを揃えた開発のためにハードウェア開発期間のできるだけ早い時期から支援ツールを提供することが課題となる．

このような課題に対して，様々なツールを迅速に提供するためにはツールの Web サービス化が有効と考え，ツールの試用やデモを共通のフレームワークで提供する TaaS（Tools as a Service）と呼ぶクラウド環境を開発した．現在，九州先端科学技術研究所では，TaaS を利用して，インターコネクト・シミュレーションのみならず，動的再構成可能な ASIP 向けの開発ツールやキャッシュ・ミス率予測ツールの実行サービスを一般に公開している．

本研究では，開発ツールのデモや試用を迅速かつ容易に行え，利用者やツール提供者に対して，ツールを利用する際の初期コストやツール提供のための開発・運用コストを低減できるクラウドコンピューティング環境の構築を目的とする．

本稿では，TaaS と OpenNSIM によるインターコネクトシミュレーションサービスについて述べる．以下，第 2 章では，インターコネクトシミュレータ OpenNSIM について概説し，第 3 章では，TaaS の内部処理について説明する．第 4 章では，TaaS を用いた OpenNSIM の評価事例について述べる．そして，第 5 章でまとめと今後の課題について述べる．

## 2. OpenNSIM：インターコネクトシミュレータ

OpenNSIM は，我々がこれまでに開発してきた NSIM をベースに，安定したシミュレーション実行を TaaS で提供できるよう機能調整を施したものである．これは，NSIM において機能拡張や高速化などの先行的な開発を行い，その後，OpenNSIM にフィードバックするという開発方針を採っているためである．

従来のシミュレータは，実機から取得した通信ログや人工的に生成した通信パターンを利用するものが多く，メッセージの到着順序によって通信パターンが変化する場合は正確なシミュレーションを行うことが難しい．また，数万ノード規模のインターコネクトを対象とした場合，必要とする通信ログは非常に大きくなるため，取得や生成することは困難である．

OpenNSIM の入出力ファイルを**図 1** に示す，NSIM では MGEN プログラム（**図 2**）と呼ぶ MPI 相当の C プログラムを駆動させ，シミュレーション中にオンデマンドで通信イベントを生成している．したがって，実際の MPI プログラムの通信パターンに則したインターコネクトのシミュレーションができる．ただし，シミュレーションではメッセージデータの授受は行わないため，受信メッセージの内容に応じて通信パターンが変化するようなプログラムのシミュレーションについては対象外としている．現在，基本的な 1 対 1 の同期・非同期通信をはじめ，ランデブー通信，ゼロコピー通信をサポートしており，基本的な集団通信は MGEN プログラムとして別途 Web サイトから提供している．また，インターコネクトの詳細仕様をインターコネクト・コンフィグレーションファイル（**図 3**）と呼ぶ設定ファイルで与えるため，実機のみならず新規のインターコネクトの性能予測ツールとしても利用できる．そして，MPI ランクと物理ノードの対応を任意に与えられるよう，ランク・ノード変換テーブルファイルを入力とする．

OpenNSIM は，シミュレーション終了後，性能情報ファイル，統計情報ファイル，通信履歴ファイルを出力する．性能情報ファイルは，シミュレーションによって得られた評価対象インターコネクトの性能情報（MGEN プログラムの予測実行時間，総転送パケット数，

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

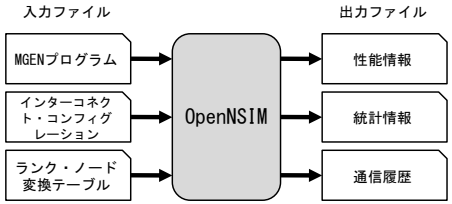


**図 1** NSIM の入出力ファイル

```
int MGEN_Main( int argc, char *argv[] )
{
    :
  MGEN_Comm_rank(MGEN_COMM_WORLD, &myrank);
  MGEN_Comm_size(MGEN_COMM_WORLD, &mysize);
  // Pairwise exchange
  int i,src,dst;
  MGEN_Request req[2];
  MGEN_Status stat[2];

  for ( i=1; i<mysize; i++ ) {
   dst = src = myrank ^ i;
     MGEN_Irecv(NULL,msg_sz,MGEN_BYTE,src,tag,comm,&req[0]);
     MGEN_Isend(NULL,msg_sz,MGEN_BYTE,dst,tag,comm,&req[1]);
     MGEN_Wait(&req[1], &stat[1]);
     MGEN_Wait(&req[0], &stat[0]);
  }
}
```

**図 2** MGEN プログラム（一部）

```
(network
 name="3dt-32x16x16"
 rank-node-filename="3dt-32x16x16.rnc"
 number-of-NIC=1
 number-of-virtual-channels=2
 packetsize=2048 ; byte
 flitsize=16 ; byte
 mtusize=2048 ; byte
 packetheadersize=32 ; byte
 virtual-channel-buffersize=8192 ; byte
 routing-computation-time=4000 ; ps
 virtual-channel-allocation-time=4000 ; ps
 switch-allocation-time=4000 ; ps
 flit-traversal-time=4000 ; ps
 switch-latency=78000 ; ps
 cable-latency=10000 ; ps
 (topology
   name="3D Torus"
   X=32 Y=16 Z=16
 )
```

**図 3** インターコネクト・コンフィグレーションファイル（一部）

総転送データ量，実効バンド幅，リンクスループット，リンクビジー率など）を含む．また，統計情報ファイルは，リンクスループット，仮想チャネルバッファの利用率，通信レイテンシ，ネットワークレイテンシなどの詳細な統計情報である．そして，シミュレーションされた通信イベントの詳しい履歴を通信履歴ファイルに出力する．OpenNSIM の出力に関する詳細は，TaaS+NSIM ユーザーズマニュアル[4]，ならびに OpenNSIM マニュアル[5] を参照されたい．なお，NSIM は並列離散事象シミュレーション（PDES）モデルに基づき，MPI で実装しており，多くの並列処理プラットフォームで実行可能である．

## 3. TaaS：ツール提供を目的としたクラウド環境

### 3.1 設 計 方 針

TaaS では，各種ツールのデモや試用を迅速かつ容易に行えることが要件である．そこで，本研究では，利用者とツール提供者に対して，以下の利便性を提供することを念頭とし，TaaS の設計方針を立てた．

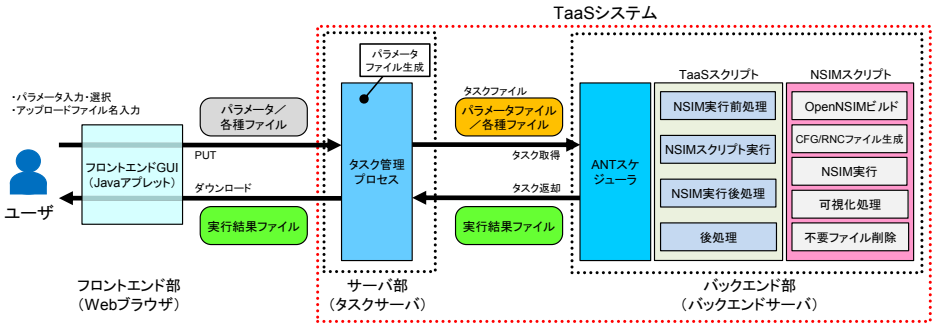


**図 4** TaaS 実行環境の構成と処理フロー（OpenNSIM の場合）

利用者に提供する利便性は，ツールを利用する際の初期コストを低減できることである．そこで，まず，利用者によるダウンロード，ビルド，インストールなどの作業を一切不要とするために，ツールの実行を Web サービス化とし，一般的な Web ブラウザから利用できるものとする．また，他のツールを利用する場合でも，新たに利用方法を習得する期間を削減するために，Web ブラウザからの操作方法の共通化を図る．そして，ツールの実行が手元の実行環境で行われた場合と同等にするために，実行結果ファイルを ZIP 化し，TaaS から返却するものとする．

一方，ツール提供者に提供する利便性は，ツール提供のための開発・運用コストを低減できることである．そこで，まず，HTTPS や Basic 認証の利用など最低限のセキュリティを提供するために Web サービスの基本部分を共通化する．また，ツール利用時のパラメータ設定をテンプレート化し Web ブラウザで表示する GUI を自動生成することで，ツール固有のオプション設定や利用手順の簡単化を図る．そして，これらのサービスを提供するサーバと，実際にツールを実行するサーバを分離し，柔軟に負荷分散する構造を採るものとする．

### 3.2 実行環境の構成と処理フロー

**図 4** に TaaS 実行環境の構成と OpenNSIM における処理フローを示す．TaaS 実行環境は，フロントエンド部，サーバ部，バックエンド部の 3 つから構成される．また，サーバ部とバックエンド部を合わせて TaaS システムと呼ぶ．以下に全体の処理の流れを示す．

**フロントエンド部**

(1) ユーザは Web ブラウザによって，サーバ部のポータルサイトにアクセスし，ツール利用のためのメールアドレスやワンタイムパスワードを登録する（仮申込み）．

(2) ツール利用のための専用の URL とユーザ ID が，サーバ部からユーザにメールで送

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

られる．Basic 認証によってその URL にアクセスするとツール毎に各種パラメータを入力・選択するための Java アプレットがブラウザに送信される．このアプレットによってフロントエンド GUI が形成され，必要なパラメータをブラウザから設定し，必要に応じてローカルにあるアップロードファイルとともにサーバ部（タスクサーバ）に送信される（本申込み）．

**サーバ部**（単一のタスクサーバから構成）

（3） フロントエンド部から受信したパラメータについて，所定のチェック（データ有効範囲やサイズチェックなど）を行い，後述するバックエンド部で実行する実行スクリプト用のパラメータファイルを生成する．そして，アップロードされたファイルと共に ZIP 化し，タスクファイル（#id.ZIP）としてツール毎のキューに登録・管理する．

**バックエンド部**（ツール毎に複数のバックエンドサーバから構成）

（4） バックエンドサーバで常時起動している ANT スケジューラ（Ant で記述）によって，定期的にタスクサーバをアクセスし，当該ツールのキューを検索する．新しいタスクが登録されている場合，そのタスクファイルを取得し，バックエンドサーバ内のワークディレクトリに展開する．

（5） そして，ANT スケジューラは TaaS シェルスクリプトを実行し，NSIM 実行の前処理を行った後に，NSIM シェルスクリプトを実行する．

（6） NSIM シェルスクリプトでは，OpenNSIM の実行ファイルをビルドし，パラメータファイルに従ってシミュレーションに必要なファイルを生成した後に OpenNSIM を実行する．その間，ANT スケジューラによって所定のタイムアウト時間の超過を監視しており，タイムアウト時間を超えると強制的に実行を停止させる．

（7） 出力された統計情報ファイルを可視化，ならびに不要ファイルを削除し，NSIM シェルスクリプトを終了する．なお，OpenNSIM の実行結果は，この時点でワークディレクトリにすべて格納されている．

（8） TaaS シェルスクリプトは，ファイルサイズの超過チェックやレポートファイルといった後処理を行い，ANT スケジューラに処理を戻す．

（9） 実行が終了したワークディレクトリ内の実行結果ファイルを ZIP 化し，タスクサーバに送信する．

**サーバ部**

（10） バックエンド部から実行結果ファイルを受信すると，当該ツール専用のプールディレクトリに格納し，メールによってツールの実行が終了したことをユーザに通知する．

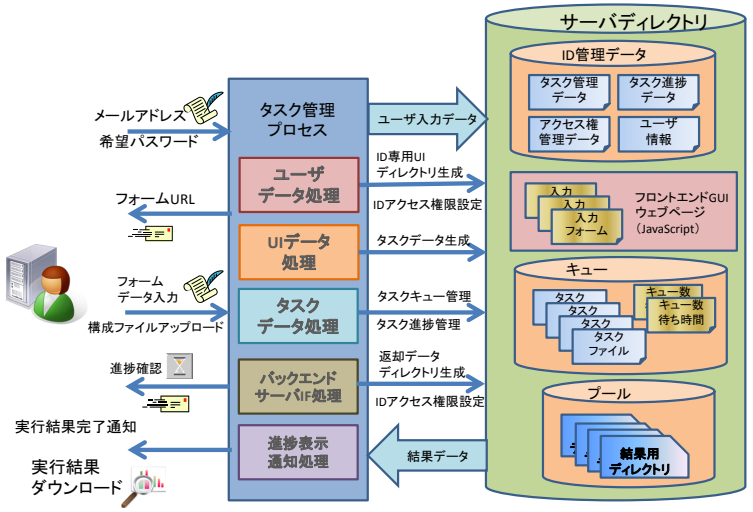


**図 5** タスクサーバの機能ブロック図

**フロントエンド部**

（11） タスクサーバから受信したツール実行の終了メール内には，タスクサーバが管理している実行結果ファイルのダウンロード専用ページへの URL が記載されており，そのページをアクセスすることで，ツールの実行結果を得る．

以上の処理フローに従って，OpenNSIM が TaaS システム上で実行される．なお，ブラウザからの操作手順は共通化しているため，他のツールについても同様の手順で処理される．

### 3.3 タスクサーバの構成

タスクサーバは，主にユーザに対する Web サービスインタフェースを提供し，ユーザから要求されたツールの実行をタスクとして管理する．**図 5** にタスクサーバの機能ブロック図を示す．また，各機能の処理内容を以下に示す．

**ユーザデータ処理**：ユーザが入力するメールアドレスと希望パスワードを用いて，仮登録メール送信および本登録申込み用の Web ページ URL を送信する．また，仮申込みで発行された ID 番号とそのタスクの進捗を管理する ID 管理データファイルへの登録，更新処理も行う．バックエンドサーバから実行結果ファイルが返却されると，完了通知メールを送信した後，メールアドレスとパスワードはサーバの ID 管理データから削除される．ただし，実行結果ファイル確認用の Web ページのアクセスパスワードは，結果ファイルページ

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

が保存される 3 日間だけ残される．なお，OpenNSIM ではユーザとタスクサーバ間の認証は PHP の crypt 関数を用いた DES アルゴリズムを用いている．

**UI データ処理**：OpenNSIM のユーザインタフェースは JavaScript を用いている．Ajax（ウェブブラウザ内で非同期通信とインターフェイスの構築などを行う技術）通信を用いて，Web ブラウザとサーバ間でのデータの双方向通信を実現し，データのチェック応答，受信状態通知などを実装している．なお，Web サーバから共通の Javascript でサーバ側に接続した場合，サーバ側にはどのタスク ID かを認識する手段がないため，各タスク ID で専用のディレクトリを生成し，各 ID 専用の URL で POST や GET を行う JavsScript コード生成し，各 ID のディレクトリに出力する．したがって，各 ID のディレクトリにはユーザ入力のパスワードでしかアクセスできない PHP コードを置いている．また，ID 番号は固定でパスワード入力のみとしているため，他のユーザの ID とパスワードでは試行もアクセスも不可能でありセキュリティが高くなるという利点がある．

**タスクデータ処理**：本申込み完了後のキュー管理，進捗データ表示を担当する．キューにタスクファイルが投入されると，各タスク ID の進捗管理データをキュー待ち状態にして，最大待ち時間表示のための，各タスクのタイムアウト設定時間をタイムアウト管理ファイルに記録する．ユーザから進捗状況の表示要求が来ると，キューに存在するファイルリストからキューの個数とタイムアウト時刻の積算を行い，進捗ページに表示する．

**バックエンドサーバ IF 処理**：バックエンドサーバから定期的にキュー状態の取得要求がある．バックエンドサーバからタスクサーバへのアクセスが認証されると，キューの有無，キューがある場合はそのファイル名が通知され，バックエンドサーバは#id.zip のファイルをダウンロードする．ダウンロード後，タスクサーバでは，キューからダウンロードされた ID のファイルおよび，ファイル名をファイルリストから削除する．

バックエンドサーバで，取得したタスクファイルの展開に成功すると，所定のタスクサーバの PHP ファイルへタスク ID（#id）の HTTP アクセスを行う．これを受けてタスクサーバでは，進捗状態を実行待ちから実行中に更新する．

バックエンドサーバから結果ファイルが POST によりアップロードされると，ID 専用の実行結果ファイルのダウンロードページ用ディレクトリを生成し，ID 専用のパスワード PHP 認証のページおよび#id.zip が置かれ，その URL をユーザへメールで通知する．OpenNSIM ではグラフデータを含み，HTML レポートも生成するため，結果確認ページからその HTML レポートファイルを閲覧できるよう，#id.zip を展開している．なお，PHP 認証済みの結果確認ページのリンクを用いず，#id.zip や HTML レポートを直接アクセス

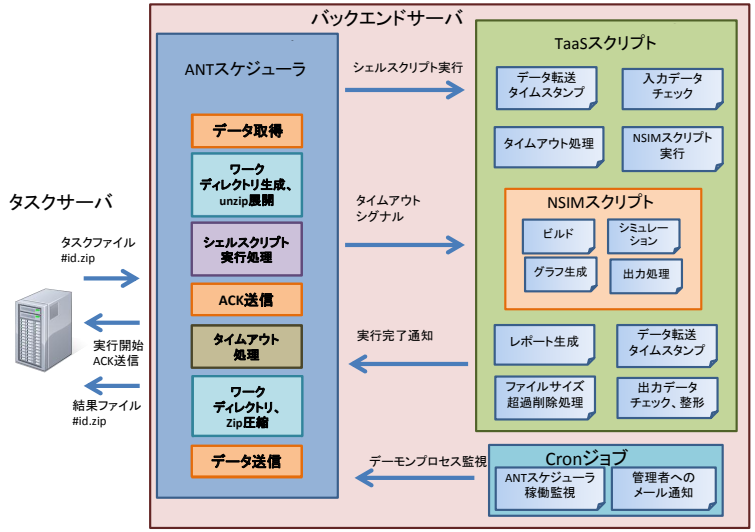


**図 6**　バックエンドサーバの機能ブロック図（OpenNSIM 利用時）

できないよう，.htaccess により直接アクセスは禁止する設定にしてあるため，ID 番号もしくは URL を第三者が知ったとしても，パスワード入力なしでは結果を一切知ることができないようアクセスを制限している．

**進捗表示・通知処理**：進捗状況ページおよびメール通知機能を実現する PHP プログラムである．進捗状況は，ID のタスク進捗状況を管理しているファイルに適宜アクセスして進捗を表示する．なお，OpenNSIM ではキューに存在する場合は，待ちキュー数，最大待ち時間，本登録してからの時間表示，といった機能拡張を行っており，キューに存在する場合の実行取り消しや，実行タスクの中断申込みの拡張を予定している．

### 3.4　バックエンドサーバの構成

バックエンドサーバは，主にタスクサーバから受け取ったタスクファイルを基に必要な前処理を行い，実際にツールを実行した後に，結果を再度タスクサーバに返却する．**図 6** に，バックエンドサーバの機能ブロック図（OpenNSIM 利用時）を示す．また，各機能の処理内容を以下に示す．

#### 3.4.1　ANT スケジューラ

**データ取得**：タスクサーバから未処理タスクの ZIP ファイルを取得する．

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

**ワークディレクトリ生成**：取得したタスクファイル（#id.zip）から ID 番号を抽出し，ID 用のディレクトリを ANT のワークディレクトリ内に生成する．

**UNZIP 展開**：取得したファイルを UNZIP 展開し，ワークディレクトリに格納する．

**シェルスクリプト実行処理**：展開したファイルに含まれている，ANT 用設定ファイルに基づき，TaaS スクリプトを実行する．

**ACK 送信**：ZIP ファイルを展開後，タスクサーバが指定する PHP ファイルにアクセスし，ツールの実行が開始したことを通知する．

**タイムアウト処理**：TaaS スクリプトの実行中に設定したタイムアウトを超過すると，実行中のスクリプトに対して SIGTERM シグナルを送信し，ツール実行を中断させる．

**ZIP 圧縮**：TaaS スクリプトの実行が完了，もしくは中断されると，当該タスクのワークディレクトリ内に存在する実行結果ファイルを圧縮し，#id.zip のファイルを生成する．その後，ワークディレクトリを削除する．

**データ送信**：タスクサーバに生成した実行結果ファイル#id.zip を POST 送信し，一連の処理を完了する．

#### 3.4.2 TaaS スクリプト

**データ転送**：NSIM スクリプトは指定のディレクトリで実行するように設計しており，ANT のワークディレクトリから，必要なファイルを NSIM のワークディレクトリにコピーする．

**タイムスタンプ**：レポート作成用に，処理を開始した時刻をファイルに記録する．

**入力データチェック**：NSIM スクリプトで最初の環境変数設定に必須の Param.tcsh ファイルが入力に存在するか確認する．存在しない場合は NSIM スクリプトは動作しないため，処理を中断して終了する．/home/taas/log には，この Pram.tcsh をログとして実行時間ファイル名として保存しており，タイムアウトエラーなどの分析に利用できる．

**タイムアウト処理**：TaaS スクリプトや NSIM スクリプトは，サブ・シェルスクリプトやツール実行のために複数のプロセスを起動するため，タイムアウト時に TaaS スクリプトを停止するだけではいくつかのプロセスが残る．そこで，TaaS スクリプト内に，ANT スケジューラからの SIGTERM の受信で動作するタイムアウト処理関数を定義し，NSIM スクリプトが起動したプロセスを親，子，孫と継承先までプロセス ID を追跡し，該当するプロセスを全て停止させている．NSIM スクリプトが停止すると，NSIM スクリプトのワークディレクトリからすべてのファイルを ANT のワークディレクトリに移動し，タイムアウト中断の発生時刻をファイルに記録し TaaS シェルスクリプトの実行を終了する．

**レポート生成**：ツールの実行が正常に終了すると，NSIM スクリプトのグラフ生成ツールの実行結果，実行開始時刻，終了時刻，ユーザが指定したパラメータ一覧などがレポートファイルに出力される．一方，タイムアウトエラーやファイルサイズ超過エラーが発生した場合は，タイムアウト時刻や超過によって削除したファイルの情報を出力する．

**ファイルサイズ超過削除処理**：実行結果ファイルのサイズが所定のサイズを超えた場合，ファイルサイズが大きいファイルから削除し，所定サイズ以内に収まるよう実行結果ファイルを調整する．

**出力データチェック**：ホワイトリストを参照し，指定以外のファイルが生成されていないかチェックし，存在する場合は削除する．これは，MGEN プログラムによるいたずらや，ハッキングを目的としたファイル生成，あるいはファイルの持ち出しを防止するためである．

#### 3.4.3 NSIM スクリプト

NSIM スクリプトは，主に OpenNSIM を実行しその出力結果を操作するためのスクリプトである．具体的には，OpenNSIM のビルド，実行，統計情報ファイルからのグラフ生成，不要ファイル削除の順に実行し，終了すると TaaS スクリプトの実行に戻る．

#### 3.4.4 Cron ジョブ

ANT スケジューラのプロセスが存在しているか定期的にチェックする．異常終了あるいは，バックエンドサーバの再起動などでプロセスが存在しないときは，スケジューラを起動し，メールによってスケジューラがダウンしていたことを通知する．

### 3.5 TaaS による OpenNSIM のパラメータ設定

TaaS による OpenNSIM のシミュレーションでは，1）MGEN プログラム，2）インターコネクト，3）シミュレーション出力，4）タイムアウトの 4 項目についてパラメータ設定を行う．OpenNSIM のシミュレーションでは，ReadyMade モードと UserMGEN モードの二つのモードを用意している．前者の ReadyMade モードは，TaaS 側で用意している既成の MGEN プログラム，インターコネクト・コンフィグレーションファイル，ランク・ノード変換テーブルファイルを利用するものである．また，後者の UserMGEN モードは，ユーザが作成した MGEN プログラムやインターコネクト・コンフィグレーションファイルを TaaS へアップロードし，OpenNSIM への入力として利用するものである．なお，MGEN プログラム設定では ReadyMade モードとし，インターコネクト設定では UserMGEN モード，あるいはその逆のような指定も可能である．

#### 3.5.1 MGEN プログラム設定

ReadyMade モードで MGEN プログラムを選択する場合は，ブラウザに表示されている ReadyMade MGEN プログラムの利用を選択し，引き続き MGEN プログラムを選択する

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

**表 1** ReadyMade モードで提供する MGEN プログラム

| MGEN プログラム | プログラムの内容 |
|---|---|
| Point to Point | ランク 0 からランク 1 へのメッセージ通信（1 対 1 通信） |
| Random Ring | 全ランクを用いたランダムリング通信 |
| Alltoall(pairwise exchange) | pairwise exchange アルゴリズムによる全対全通信 |
| Alltoall(bruck) | Bruck アルゴリズムによる全対全通信 |
| Alltoall(ring) | Ring アルゴリズムによる全対全通信 |
| Allreduce(recursive doubling) | Recursive doubling アルゴリズムによる Allreduce 通信 |

（**図 7**）．そして，これらの通信を行う際のデータサイズ（メッセージのサイズ）を入力する．現在，MGEN プログラムは**表 1** に示す 6 つを利用できる．

一方，UserMGEN モードでユーザが記述した MGEN プログラムを利用したい場合は，ユーザファイルの利用を選択し，所定の欄に MGEN プログラムのファイル名を入力する．

### 3.5.2 インターコネクト設定

ReadyMade モードでコンフィグレーションファイルを選択する場合は，readymade コンフィグレーションの利用を選択し，トポロジとノード数の二つについて選択する（**図 8**）．現在，OpenNSIM では，2 次元メッシュ網，2 次元トーラス網，3 次元メッシュ網，3 次元トーラス網，FatTree 網の 5 つのトポロジをサポートしている．なお，詳細は省くが，ReadyMade モードでのインターコネクト設定は，リンクバンド幅：4GB/s，パケット長：2KiB，仮想チャネルバッファ：8KiB，1 ホップあたりの遅延は 100ns に設定している．

ユーザが記述したコンフィグレーションファイルを利用したい場合は，ユーザファイルの利用を選択し，入力欄にインターコネクト・コンフィグレーションファイル名とランク・ノード変換テーブルファイル名を与える．

### 3.5.3 シミュレーション出力設定

ユーザがダウンロードする実行結果ファイルに含める項目を選択する．選択項目には，大きく通信履歴の出力要求と各種の統計情報のグラフ出力がある．まず，MGEN プログラム実行時の通信履歴を出力するか選択し，次にそれぞれの統計情報についてグラフ化するかを選択する．また，統計情報を採取する観測間隔をマイクロ秒単位で入力する．統計情報のグラフ化は，**表 2** の 7 項目を選択することができる．それぞれにチェックを入れると当該項目について統計情報のグラフ化を行う（**図 9**）．

### 3.5.4 タイムアウト設定

OpenNSIM の実行についてタイムアウト時間を設定する．なお，現在設定できる最大実行時間は 60 分である．（**図 10**）

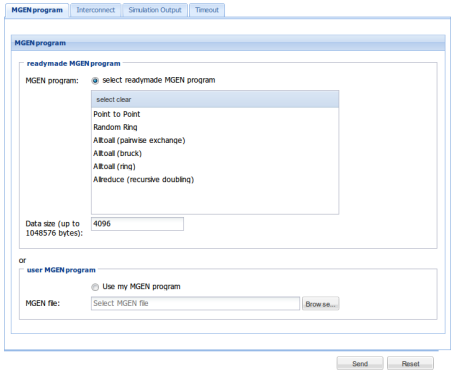


**図 7** MGEN プログラムの設定

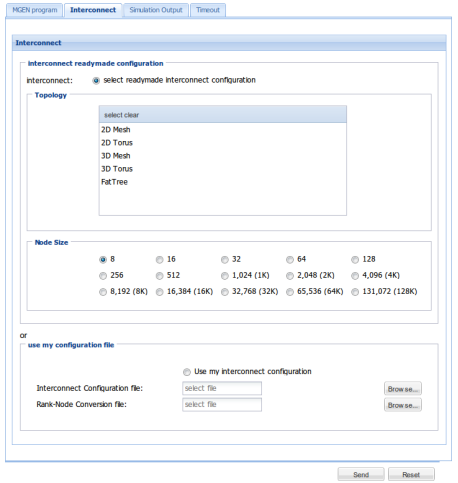


**図 8** インターコネクトの設定

**表 2** 選択可能なシミュレーション出力項目

| Generate Timeline Graph（時系列グラフ） | |
|---|---|
| Network Latency | ネットワークレイテンシ |
| Link Throghtput | リンクスループット |
| Buffer Utilization | バッファ利用率 |
| Communication Latency | 通信レイテンシ |
| Link throughput Distribution | リンクスループットの割合の変動 |
| Congenstion Point Ratio | 輻輳の発生割合 |
| Generate Congestion Point Graph | |
| Router & Port | 輻輳箇所をノード・ポート毎に上位順で表す |

## 4. 評価事例

本節では，TaaS を用いて OpenNSIM を実行した評価事例について述べる．

本評価では，2,048 ノードの 3 次元トーラス網（16×16×8）を評価対象とした．また，評価アプリケーション（MGEN プログラム）には，pairwise exchange による Alltoall を用いた．OpenNSIM でシミュレーションする際のインターコネクト・コンフィグレーションの基本設定値を**表 3** にまとめる．これらの仕様は，現在の主流となる技術・性能を反映し，かつ，評価対象アプリケーションの評価を容易にできる値としている．

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

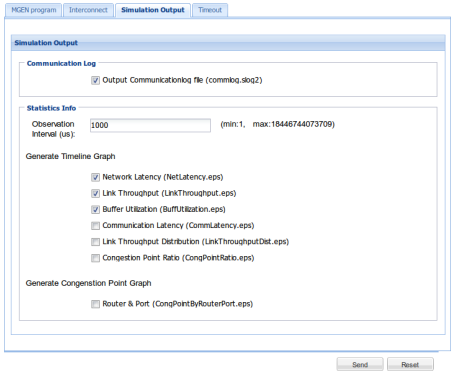


**図 9** シミュレーション出力の設定

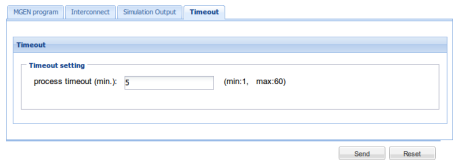


**図 10** インターコネクトの設定

**表 3** インターコネクト・コンフィグレーションの基本設定

| パラメータ | 設定値 | パラメータ | 設定値 |
|---|---|---|---|
| ルーティング方式 | 次元順+dateline | MTU | 2KiB |
| パケット転送方式 | VCT | パケット長 | 32B〜2KiB(MTU) |
| ノード間リンクバンド幅 | 4GB/s（単方向） | パケットヘッダ長 | 32B |
| ルーティング計算時間 (RC) | 4ns | 仮想チャネル数 | 2 |
| 仮想チャネル設定時間 (VA) | 4ns | 仮想チャネルバッファ | 8KiB (MTU×4) |
| スイッチ設定時間 (SA) | 4ns | フリット長 | 16B |
| フリット転送時間 (ST) | 4ns | DMA 転送レート | 16GB/s |
| スイッチ遅延時間 | 78ns | メモリバンド幅 | 16GB/s |
| ケーブル遅延時間 | 10ns | MPI 関数処理オーバヘッド * | 200ns |

* スタートアップレイテンシ／後処理遅延

TaaS システムのバックエンドサーバには，Intel Xeon E5440 2.83GHz (4cores×2) を搭載した Linux マシンを用いた．OpenNSIM 実行の結果，シミュレーションに要した時間は約 24 分であった．また，Alltoall の予想実行時間は 166.7ms であった．各種統計情報を可視化し，その中から各次元方向毎のリンクスループット，平均ネットワークレイテンシ，ならびに輻輳発生率の時系列変化を，それぞれ，**図 11**，**図 12**，**図 13** に示す．

リンクスループットに着目すると，時間の経過と共に各次元方向のスループットが大きく脈動していることがわかる．また，次元毎にその変化の具合（周期，振幅）が異なっている．

pairwise exchange アルゴリズムは，最初，近隣ランクとペアを組んでデータ交換を行うが，データ交換を繰り返す毎にペアを組むランクを離していく．すなわち，直接網である 3 次元トーラス網では，データ交換を行う度に通信先の物理ノードが遠くなる（ホップ数が増

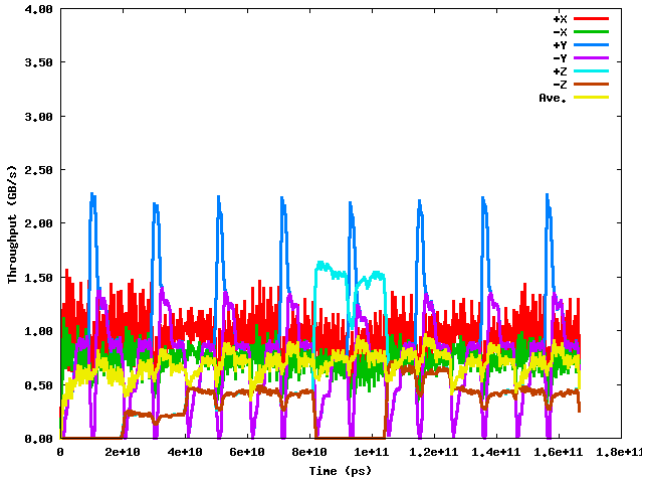


**図 11** リンクスループットの時系列変化

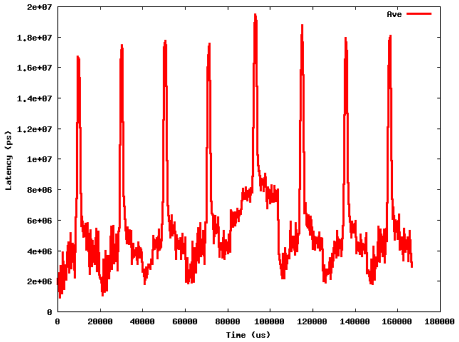


**図 12** 平均ネットワークレイテンシの時系列変化

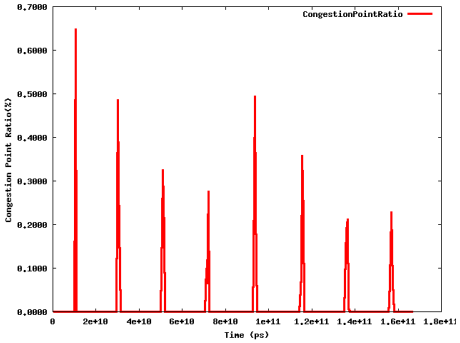


**図 13** 輻輳発生率の時系列変化

加する）．すると，一本のリンクを通過しようとするメッセージが多くなりリンクスループットは向上するが，パケットの衝突が増えるためネットワークレイテンシは増加してしまう．

さらに，ホップ数が次元方向の半分になった時，それまでのペア同士の通信は次元軸に対して + 方向と − 方向のように逆向きに送信してものが，次元順ルーティングの性質（同ホップ数の経路が 2 つある場合は，+ 方向へ優先的にルーティングする）により，+ 方向と + 方向に送信する結果となる．すなわち，プログラムの実行中，局所的に同一方向への

Vol.2010-ARC-192 No.15
Vol.2010-HPC-128 No.15
2010/12/17

送信が行われるため，リンクスループットは一気に向上する．しかし，リンクを経由するメッセージ数が増えるため，ネットワークレイテンシが増加してしまう．一方，この同一方向への送信が終わると，送信方向はペア同士で逆転するため，パケットの衝突は減りネットワークレイテンシが減少するとともに，スループットも減っていくことになる．これは，ネットワークレイテンシや輻輳発生率の時系列変化のグラフにも顕著に表れている．

各次元軸の大きさを x, y, z とすると，このような局所的な同一方向への送信による脈動は，X 軸では y×z 回，Y 軸では z 回，Z 軸では 1 回となり，シミュレーションから得られたリンクスループットのグラフとも一致する．

このように，OpenNSIM を用いることで，所望の並列プログラムを実行した場合のインターコネクト内部の挙動を観測することができるため，実行性能の評価のみならず，プログラムのデバッグや性能チューニングの支援にも利用できる．

## 5. 関 連 研 究

最近の TaaS に類似したクラウドサービスとして，以下が挙げられる．

(1) 組込み開発ツール提供サービス（日立ソフト）[6]
(2) 開発環境提供サービス（日本ユニシス）[7]
(3) CollabNet TeamForge（CollabNet）[8]

（1）は，コンパイラやシミュレータの提供に焦点を置くという観点から，本研究との類似性がある．(2) と (3) は，チーム開発に焦点を置いてバージョン管理や継続的インテグレーションといった周辺サービスを提供している．将来的には開発ツールの提供サービスと，バージョン管理や継続的インテグレーションといったチーム開発支援的な周辺ツール提供サービスは併用や統合が望ましいと考えられる．

ツール提供サービスの実現方法に関しては上記のサービスと TaaS では異なる．我々が遠隔ビルド方式でビルドとツールの実行だけをサービス側で行い，開発環境そのものはユーザ側に置くのに対し，上記のサービスは開発環境を仮想マシンに搭載して IaaS 型のクラウドに置き，仮想デスクトップ技術により UI を提供している．これらのサービスのような仮想マシン化と TaaS を比較した場合，仮想デスクトップの GUI は応答が遅いという短所があるが，開発環境をそのままサービス化することはユーザ側の管理コストを低くすることに役立つという長所もある．また，現在の TaaS ではまだ計画中だが，遠隔ビルドを拡張して分散ビルド化することは容易であるため，ツール実行の負荷が大きい場合にはスケーラビリティを改善できる見込みが高い．

## 6. ま と め

本稿では，次世代の大規模インターコネクトの性能評価を目的としたシミュレータ OpenNSIM の概要，ならびにツールのデモや試用を迅速かつ容易に行え，利用者やツール提供者に対して各種のコストを低減するクラウド環境 TaaS について述べた．TaaS は，利用者に対してツールを利用する際の初期コストの低減という利便性を提供し，ツール提供者に対してはツール提供のための開発・運用コストの低減といった利便性を提供する．共通化されたフレームワークによって，ツール利用時の煩雑なパラメータ設定を簡単化し，Web ブラウザからの容易な利用を可能にすることができた．また，TaaS 環境を用いた評価事例を通じて，OpenNSIM がインターコネクトやプログラムの解析に利用できることを示した．

今後の課題としては，OpenNSIM で動作する MGEN プログラムの拡充，ならびに，TaaS のユーザビリティやセキュリティの向上がある．

**謝辞** 本研究を進めるにあたりご協力頂いた九州大学情報基盤研究開発センター稲富雄一氏，同大学村上研究室の学生諸氏，(財) 九州先端科学技術研究所 吉松則文氏，曽我武史氏，に感謝する．

## 参 考 文 献

1) 三輪英樹，薄田竜太郎，柴村英智，平尾智也，眞木淳，稲富雄一，井上弘士，安島雄一郎，三吉郁夫，清水俊幸，安藤壽茂：NSIM:将来の大規模相互結合網を対象とした通信シミュレータの開発，情処研報，Vol.2010-HPC-125, No.5, pp.1–9 (2010).
2) 柴村英智，三輪英樹，薄田竜太郎，平尾智也，安島雄一郎，三吉郁夫，清水俊幸，石畑宏明，井上弘士：パケットペーシングを用いた最適全対全通信アルゴリズムのシミュレーション評価，情処研報， Vol.2010-HPC-126, No.14, pp.1–9 (2010).
3) WWW Computer Architecture Page: http://arch-www.cs.wisc.edu/wwwarch/public/home.
4) 九州先端科学技術研究所：TaaS+NSIM ユーザーズマニュアル，https://ngarch.isit.or.jp/taas/opennsim/taas+nsim-users-man.pdf.
5) 九州先端科学技術研究所：OpenNSIM マニュアル，https://ngarch.isit.or.jp/taas/opennsim/nsim-man.pdf.
6) 日立ソフトが組込みソフトウェア開発向けの SaaS 型サービスを提供開始：http://cloud.watch.impress.co.jp/docs/release/20100730_384448.html.
7) クラウドコンピューティング概況と日本ユニシスの取り組み：http://www.unisys.co.jp/tec_info/tr100/10007.pdf.
8) CollabNet TeamForge: http://www.open.collab.net/jp/products/ctf/.