{ 第1部 } 高い生産性を持つ高信頼ソフトウェア作成技術の開発 * 2

# 次世代高性能コンピュータシステム上の高信頼ソフトウェアシステムの開発支援技術

石川　裕 *1　佐藤　充 *2
中島　耕太 *2　住元　真司 *2
久門　耕一 *2　木村かず子 *3
鈴木　敦夫 *4

*1 東京大学
*2 (株)富士通研究所 IT システム研究所
*3 NEC ソフト(株) IT システム事業部
*4 NEC ソフト(株) PF システム事業部

平成 15 年度から 3 年間，東京大学，富士通研究所，NEC ソフトは，信頼性を実現するシステムソフトウェアを開発するための支援技術として以下の研究開発を行った．1) User Mode Linux 上にハードウェア故障を模擬する故障模擬ソフトウェアシステムを開発し，2) そのシステムを利用した保守・監視系ソフトウェアシステムを開発した．また，3) 通信機器故障を模擬するハードウェアおよび開発支援システムを開発し，4) そのシステムを利用したリモートメモリアクセス通信機構を開発した．さらに，5) これら開発したシステム群を組み合わせた単一並列コンピュータシステムイメージを提供するプロトタイプシステムを開発した．

## 背景と概要

コンピュータを高速 LAN で接続したクラスタ環境（**図 -1**）が，HPC（High Performance Computing）あるいは HA（High Availability）の分野で使われてきた．HPC の分野では，高性能通信ライブラリによる並列化，あるいはパラメータサーチのようなジョブの間で通信を必要としない大量ジョブの並列実行に使われている．HA の分野では，ファイルオーバやジョブ負荷分散の形態でクラスタ環境が使われている．

一方，基幹系アプリケーションでは，大量のジョブを実行するために，大容量のディスクや Storage Area Network につながった共有メモリ型並列コンピュータが使われてきた．

平成 15 年度時点での PC の技術では，CPU コア・メモリ間のデータ転送幅は数 GB/sec，メモリ・I/O 間の合計データ転送幅では CPU・メモリ間のデータ転送能力の約半分の 4GB/sec を備えるチップセットも登場していた．10Gbps を超える高性能ネットワークをトランク（複数束ねる）することにより，I/O 能力を最大限活かすことが可能となる．データ遅延の観点では CPU とメモリに比べてネットワークは 100 倍以上遅くなるが，I/O バスである PCI-Express などでは，PCI-Express

図 -1　PC サーバ群が高速 LAN で接続されたクラスタ環境

で PC 同士をつなげる規格も出現してきており，遅延に関してもその差が縮まっている．

このようなクラスタに関するハードウェア環境で，我々は，従来のクラスタで扱っていた疎結合的利用ではなく，比較的密結合な利用，すなわち従来の共有メモリ型並列コンピュータ上で実行されてきた基幹系アプリケーションが利用可能になると考えている．すなわち共有メモリ型並列コンピュータを置き換えるシステムとして，単一並列コンピュータイメージを提供するブレードサーバシステムが実現可能である．

しかし，このような環境を実現するためには，ネットワークプロトコル，オペレーティングシステム，運用ツール等，多岐にわたるシステムソフトウェアを開発しなければならない．このために，そのようなシステムソフトウェア開発に必要な開発環境に関する研究開発を行った．

多数のプロセッサやディスクが接続されるシステムにおいては，構成要素の一部であるプロセッサやディスクが動かなくなっても他の正常なハードウェアを使用して，単一並列コンピュータとして動作させることが可能となる高信頼システムソフトウェアが必要となる．このような高信頼システムソフトウェア開発においては，さまざまなハードウェア故障を系統的に模擬し，開発したソフトウェアが耐故障性を有する高信頼性を保っているかどうかを検証するための開発支援環境が必須である．

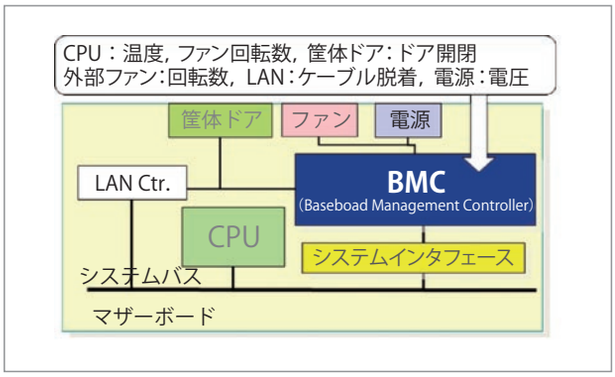


図 -2　IPMI 規格

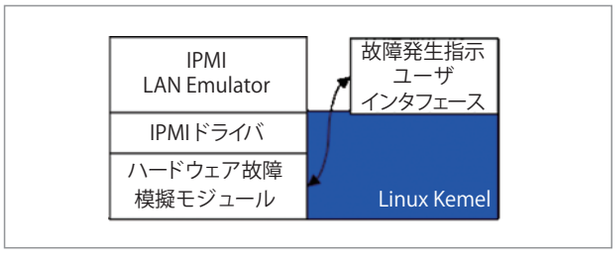


図 -3　故障シミュレータ

また，高信頼システムソフトウェアを早期に実現するための，基盤ソフトウェアの整備が必須である．我々は，開発支援環境および基盤ソフトウェアの整備として，以下のような研究開発を行った．

1) PC 単体故障模擬ソフトウェアシステムと保守・監視系システムの開発

ネットワークにつながった PC 環境上でのシステム開発は非効率的であり，また，ハードウェア故障を模擬するためのツールが必要となる．我々は，User Mode Linux 上に PC 単体のハードウェア故障を模擬するソフトウェアシステムを開発した．また，本システムを利用して保守・監視系ソフトウェアシステムを開発した．

2) 基盤ソフトウェアモジュール

通信機器の故障に対応可能な高信頼ネットワークプロトコル処理ソフトウェアを実現するために必要となる通信機器故障を模擬するハードウェアおよびそのハードウェアを使用した開発支援システムを開発した．また，本開発支援システムを用いて，リモートメモリアクセス通信機構を開発した．

さらに，開発したシステム群を組み合わせた単一並列コンピュータイメージを提供するプロトタイプシステムを開発した．

## 故障模擬ソフトウェアシステム

近年の PC サーバでは，IPMI[1] と呼ばれる規格に基づいたハードウェア故障を監視するプロセッサ（BMC）がボード上に搭載されている．**図 -2** に示す通り，BMC は，IPMI が規定するボード上の電圧，CPU のファン回転数，CPU の温度，メモリエラー，筐体ドアの開閉等のセンサデータ情報を監視している．電圧，回転数，温度などスレッシホールドを持つようなデータに対しては，上限値，下限値を設定できる．スレッシホールドを越えるような状況が発生すると故障を通知する機能がある．また，Watchdog timer によりオペレーティングシステムが正常動作しているかどうかを監視する機能もある．

IPMI 機能を利用することにより，ハードウェア故障状況に応じた耐故障機能を組み込んだ高信頼システムソフトウェアが実現可能となる．このような高信頼システムソフトウェアが設計通りに実現されているかをテストするために，IMPI 規格に基づくハードウェアの故障模擬ソフトウェアシステムを開発した（**図 -3**）．

本故障シミュレータは，オープンソースプログラムである OpenIPMI[2] で提供されている IPMI ドライバを利用している．図 -3 に示す通り，IPMI ドライバの下位層にハードウェア故障模擬モジュールを実装した[3]．ハードウェア故障模擬モジュールは，ソフトウェア的に BMC の挙動をエミュレーションするものであり，IPMI で定められたプロトコルに従って，管理ソフトウェアに BMC が保持している情報を返す．故障発生指示ユーザインタフェースは，ハードウェア故障模擬モジュールにおいてエミュレーションされた BMC の内部パラメータを変更するユーザインタフェースを提供する．実機上のハードウェアセンサ情報を取得し，故障シミュレータに設定する機能も有する．

## 保守・監視系ソフトウェアシステム

**図 -4** に示す高信頼システムを支援する保守・監視系ソフトウェア Tenjin を開発した[4]．

Tenjin では，次のような機能が実現されている．

- Web サーバ

機器の表示系，システム運用ポリシーの設定などを行う．新たに，ATML（A Temporal Markup Language）と呼ばれるタグ体系を設計した．Web サーバは apache を使用し，ATML を解釈する apache モジュールを開発した．

- データベースエンジン

DMTF[5] が規格化している CIM（Common Information Model）に基づいたデータベースエンジンであり，機器情報および機器の状態が格納される．デー

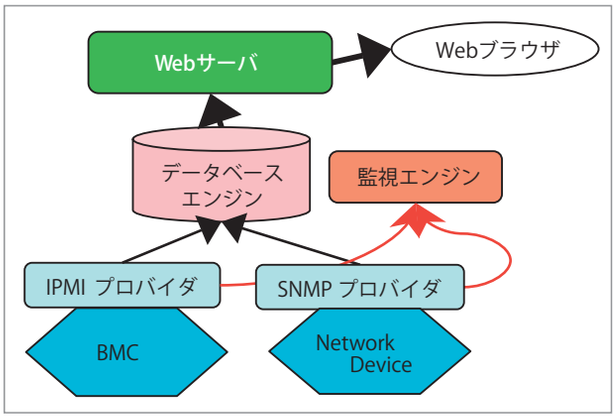


図 -4　Tenjin ソフトウェアアーキテクチャ

```
<H3>MyCluster情報</H3>
<TABLE>
<TR><TH>管理者</TH><TH>連絡先</TH></TR>
<TR><TD>石川 裕</TD>
    <TD>ishikawa@is.s.u-tokyo.ac.jp</TD>
<TR><TD>本郷太郎</TD>
    <TD>hongo@virtualsupport.co.jp</TD>
```

図 -6　実行例

タベースエンジンには，SQL データベースを使用している．CIM はオブジェクト指向データベースであるために直接 SQL を使用することはできない．CIM で定義されたデータベースを SQL で処理できるリレーショナルデータベースに変換する処理系を開発した．

- 監視エンジン

  Tenjin 監視スクリプト言語によって記述された監視プログラムを実行する．Tenjin 監視スクリプト言語と ATML は 1 つの言語として設計されている．すなわち，ATML のタグ体系に対応して Tenjin 監視スクリプト言語構文が定義されている．ATML は，Zope[13] が提供する DTML（Dynamic Template Markup Language）のようなタグとして設計実装した．ATML 使用例を**図 -5** に示す．データベース検索のために query コマンドが用意されている．本コマンドの戻り値は検索結果のリストである．図 -5 では，CIM_Cluster テーブル中，Name カラムが“MyCluster”であるエントリを検索している．atml-foreach ブロック内では，atml-eval コマンドを用いて，選択されたエントリのカラムを参照することができる．**図 -6** に実行例を示す．

- プロバイダ

  プロバイダは，機器情報を取得し，データベースエンジンに情報を登録，あるいは，機器からの Alert メッセージを受け取って，監視エンジンに故障の事象を発生させる機能を提供する．

```
<H3>MyCluster情報</H3>
<TABLE>
<TR><TH>管理者</TH> <TH>連絡先</TH></TR>
<atml-foreach query="select * from CIM_Cluster
                where Name='MyCluster'">
<TR>
   <TD><atml-eval
         name="PrimaryOwnerName"></TD>
   <TD><amtl-eval
        name="PrimaryOwnerContact"></TD>
</TR>
</atml-foreach>
</TABLE>
```

図 -5　ATML 使用例

## 通信機能故障模擬・開発支援システム

高速・高信頼通信ネットワークの開発には，通信状態の正確な把握が必要不可欠である．ハードウェア障害によるパケット喪失やデータ化けが生じてもプロトコルが正しく実装されているかどうかを検証するとともに期待通りの性能が達成されているかを確認する必要がある．10Gigabit Ethernet 環境で被測定環境に影響を与えることなく通信状態を把握することができ，高性能リモートメモリアクセス等の新しいネットワークプロトコルを開発するときに使用されるネットワーク装置とそれを利用した開発環境を構築した．ネットワーク装置は，10Gigabit Ethernet 用に XGE-ProtoDevel と UZURA の 2 つを開発した．XGE-ProtoDevel は，UZURA を開発する前段階に開発された装置である．以下，XGE-ProtoDevel とそれを使用した PacketPlot ツールについて解説したのち，UZURA 装置と，それを使用したネットワークプロトコル実装例を紹介する．

### XGE-ProtoDevel [6]

XGE-ProtoDevel は，富士通 InfiniBand HCA と 10Gigabit Ethernet ネットワークとの間に入り，InfiniBand と 10Gigabit Ethernet のプロトコル変換，擬似的なエラー挿入，送受信パケットのロギングなどを行う FPGA で構成された装置である（**図 -7**）．XGE-ProtoDevel では，すべての処理は FPGA で構成されたハードウェアで実現され，被測定環境には影響を与えることがない．

エラーモードとしてパケットのデータ化けやパケット損失を実現することができ，擬似的に信頼性の低いネットワークにおける通信処理の挙動を調査することができる．また，富士通の InfiniBand HCA は内部にマイクロコントローラを備えた NIC であり，プログラミング

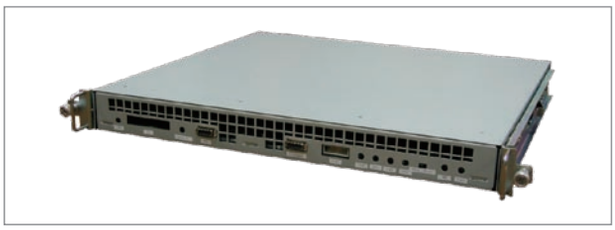

図 -7　XGE-ProtoDevel

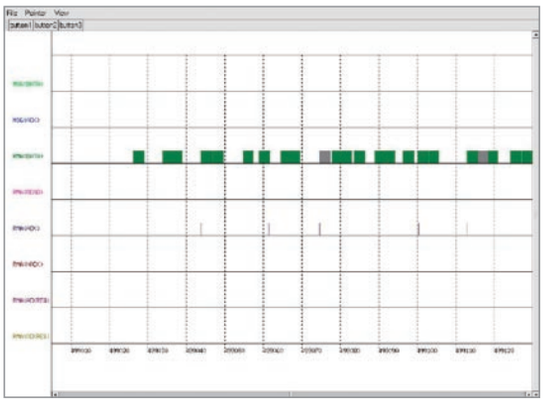

図 -8　PacketPlot

によって独自のプロトコル開発を行うことができる．

### PacketPlot

PacketPlot は XGE-ProtoDevel でロギングした送受信パケットを時系列に表示するツールとして開発された．PacketPlot は，**図 -8** のように，通信プロトコル情報を元にロギングした送受信パケットを解釈し，時系列に視覚的に表示できる．また，各パケットを GUI 上で選択し，シーケンス番号や ID 等の詳細情報を表示できる．また，パケットギャップ，再送率，パケット種別ごとの割合等の統計解析機能を持つ．

### UZURA

XGE-ProtoDevel および PacketPlot により，高速・高信頼ネットワークプロトコルの構築が可能となった．これらの成果を用いて，10Gigabit Ethernet のホスト・アダプタ UZURA の作成を行った．UZURA は **図 -9** のような PCI-X バスに接続する 10Gigabit Ethernet アダプタであり，従来 InfiniBand HCA と XGE-ProtoDevel の 2 つで実現していた 10Gigabit Ethernet 機能を単独で実現している．

UZURA に搭載されている FPGA により，これまで InfiniBand HCA のマイクロコントローラ上ファームウェアとして実現されていた RDMA プロトコルをハードウェア実装している．さらに XGE-ProtoDevel で実現されていた故障模擬機能，パケットキャプチャ機能も搭載している．UZURA を用いることにより，従来はア

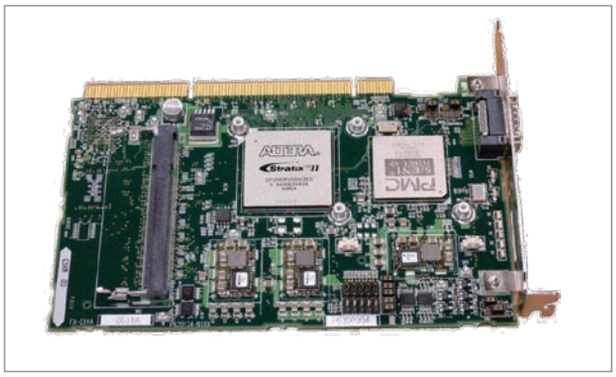

図 -9　10Gigabit Ethernet ホストアダプタ：UZURA

ダプタカードと外付け BOX の組合せで実現していた機能が，アダプタカードのみで実現できるようになり，より広範な機器への接続が可能になった．

UZURA では PCI-X インタフェースを含め多くの部分を FPGA で実現しているため，柔軟なハードウェア構成が実現できる．特に PCI-X のホストバスインタフェースに関しては，細かなパラメータ調整を行い，高効率データ転送を実現している．

### 【開発事例】

#### ◉リモートメモリアクセス機能の実現

開発した通信機能故障模擬・開発支援システムを用いて，通信プロトコル開発支援システムの効果を試すため，高速・高信頼な Remote DMA（RDMA）を用いたネットワークプロトコルを開発した [7]．

Ethernet はハードウェアレベルでパケット到着を保証していないため，Ethernet 向け RDMA プロトコルは，パケット損失を検出し，再送処理を行うよう実装する必要がある．NIC は，転送ディスクリプタと呼ばれる資源により転送データを管理している．転送ディスクリプタは，転送開始から終了まで使用中となり，完了後に再使用が可能になる．

NIC は複数の転送ディスクリプタを持てるが，転送ディスクリプタ数を増加させると NIC 上の他の資源に使用できるメモリ量が減少するため，必要最小限の使用が重要である．また，パケット到着を確認するために受信側は，ACK を送ることにより送信側にパケット到着を伝える．ACK が一定時間内に送信されないことをタイムアウトにより検出し，再送を行う．

ディスクリプタの数，タイムアウト時間を最適に設定するために，開発支援システムが使われた．当初，パケット損失率 10% の環境下で，8KB のデータ転送を行った場合，性能劣化率は 88.9% であった．これを本ツールにより通信プロトコルの挙動を可視化するとともにパケットギャップ時間や物理ネットワーク上でのパ

ケット占有率を解析し，ネットワーク実装上の問題点を突きとめた．問題点を解析しプロトコルを改善することにより，性能劣化率は 32.4% までに改善された．詳細は，論文 7）を参照のこと．**図 -10** は PacketPlot による改善前と改善後の通信プロトコルの挙動の比較結果である．このように XGE-ProtoDevel と PaketPlot を用いることによって，プロトコルの挙動を視覚的に把握することが可能になった．

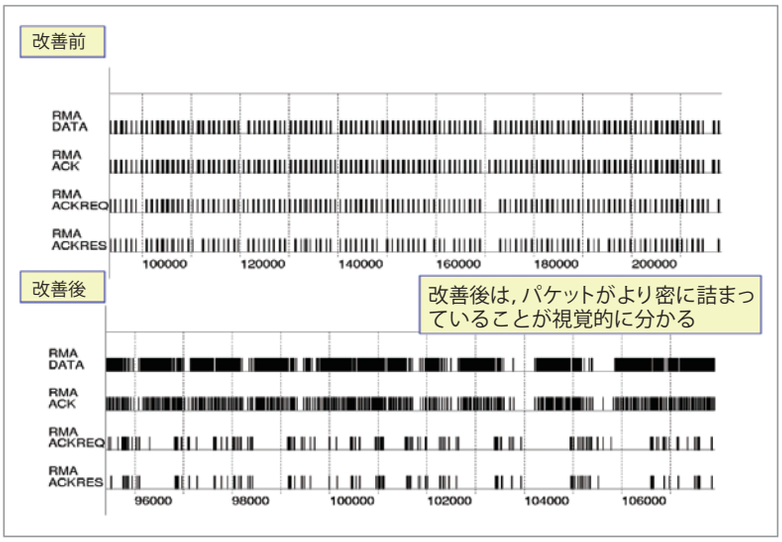

図 -10　PacketPlot による改善前と改善後の比較

### ◉ストレージシステムへの応用 [8)]

富士通で開発されている WSS（Wire Speed Storage）は，ネットワーク接続されたキャッシュノードから高速にデータを転送することができるストレージシステムである．本研究で開発された高速･高信頼なネットワークを，この WSS に応用した．

ファイルシステムにおいてはカーネルレベルでの通信が必要となる．ユーザレベル通信とは異なり，カーネルレベルでの通信は小サイズのデータ転送が頻発し，ポーリングによる同期が使えないため，従来のインタフェースを拡張したカーネルレベル・インタフェースを提供する必要があった．カーネルレベル・インタフェースを新たに作成し，ストレージシステムと組み合わせることにより，高速・高信頼通信システムのストレージへの応用可能性が確認できた．

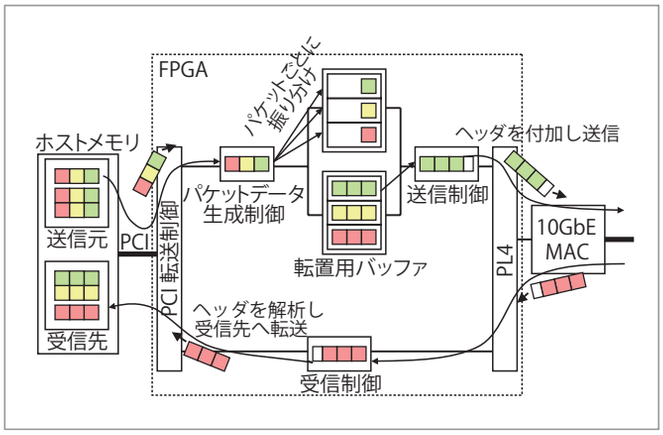

図 -11　配列転置データ転送機構

### ◉並列 FFT への応用 [9)]

並列化された FFT（Fast Fourier Transform）アプリケーションでは，配列の転置を伴うデータ転送が多用される．配列の転置では，連続するデータが不連続なアドレス上に分散配置されるため，RDMA による高速化は困難である．

そこで，**図 -11** のような配列転置データ転送機構を UZURA 上の FPGA を用いて実装した．この機構は，DMA 転送性能を高めるために，アドレスが連続する方向にホスト上のメモリを走査し DMA 転送を行う．そして，複数の FIFO からなる転置用バッファを用いて転置処理を行う．UZURA 上に実装できるメモリ量には物理的制約から限りがあるため FIFO の本数や深さには制約がある．FIFO の本数と深さは，それぞれ転送元と転送先の PCI-X バス上のバースト転送長に比例するため，DMA 転送性能に大きな影響を与える．そこで，転送元と転送先の PCI-X バスを解析し，最適な FIFO の本数と深さを設計した．これにより，限られたメモリ量において最大の転送性能を実現できた．

この配列転置データ転送機能を実際の FFT ライブラリに適応させる並列 FFT の高速化が実現できることを確認した．

## 単一並列コンピュータイメージプロトタイプ

外部から単一の IP アドレスでアクセスできるクラスタ環境のプロトタイプシステムおよびクラスタ内のノード間でプロセス移送を実現し [12)]，本研究課題で開発した基盤ソフトウェアモジュールと統合した．2005 年 11 月に開催された国際会議 SC05 の研究展示においてシステムのデモンストレーションを行った．本デモンストレーションの概要を**図 -12** に示す．従来のシステムで実現できていなかった TCP/IP コネクションを保持したままプロセス移送する機能を示すために以下の手順でデモンストレーションを行った．

1. 計算ノードからクライアントに TCP/IP を用いて画像データを送信．


図 -12　デモンストレーション概要

2. 計算ノード上のファンを故障させる．
3. 管理ノードが異常を検知し，計算ノード上のプロセスを待機ノードに移送し，計算ノードをシャットダウンする．

## その後の展開

本稿では，平成 15 年度から 3 年間，東京大学，富士通研究所，NEC ソフトとともに行った，信頼性を実現するシステムソフトウェアを開発するための支援技術の研究開発の成果について紹介した．本プロジェクト終了後，石川研究室では，基盤ソフトウェアモジュールで開発したシステムを発展させるべく，科学技術振興機構の CREST 事業の研究領域「ディペンダブル・オペレーティングシステム」において，「並列・分散型組込みシステムのためのディペンダブルシングルシステムイメージ OS」として研究開発を継続している．また，富士通研究所では，基盤ソフトウェアモジュールで開発したシステムの発展研究として，仮想マシン移動への応用 [10]，遠隔スワップメモリへの応用 [11] などが行われている．

参考文献
1）IPMI 規格：Intel, Hewlett-Packard, NEC and Dell, IPMI - Intelligent Platform Management Interface Specification, V1.5 (2002).
2）OpenIPMI：http://openipmi.sourceforge.net
3）岡家，木村，石川：IMPI 規格に基づく管理保守系システムソフトウェア，情報処理学会研究報告，03-OS-94 (SWOPP03)，情報処理学会，pp.53-60 (2004).
4）岡家，木村，石川：自立的システム管理ソフトウェアの設計，情報処理学会研究報告，2004-OS-97 (SWOPP04)，情報処理学会，pp.105-112 (2004).
5）DMTF：http://www.dmtf.org
6）Sumimoto, S., Sato, M., Nakashima, K., Kumon, K. and Ishikawa, Y. : XGE-ProtoDevel : A Communication Protocol Development Tool for 10Gbps Class Network, Dependable Software -Tools and Methods Workshop, IEEE International Conference on Dependable Systems and Networks (DSN2005) (June 2005).
7）中島耕太他：性能問題を解決する 10Gbps 級ネットワーク向け通信プロトコル開発支援ツールの提案，情報処理学会 SACSIS 2005, pp.321-328 (2005).
8）大江和一他：RDMA Storage I/O 向けカーネル通信 API の設計，情報処理学会論文誌コンピューティングシステム，Vol.46, No.SIG4 (ACS 9), pp.13-21 (2005).
9）中島耕太他：配列転置データ転送を高速化する 10Gb Ethernet インタフェースカードの設計，情報処理学会論文誌コンピューティングシステム，Vol.47, No.SIG12 (ACS 15), pp.74-85 (2006).
10）中島耕太他：10Gb Ethernet 上の RDMA 転送機能による仮想マシン移動の設計と評価，情報処理学会論文誌コンピューティングシステム，Vol.48, No.SIG18 (ACS 20), pp.69-82 (2007).
11）後藤正徳他：10Gb Ethernet 上の RDMA を用いた遠隔スワップメモリの実装，電子情報通信学会技術研究報告（CPSY），Vol.106, No.287, pp.7-12 (2006).
12）Matsuba, H. and Ishikawa, Y. : Single IP Address Cluster for Internet Servers, IEEE International Parallel and Distributed Processing Symposium (IPDPS 2007) (2007).
13）Zope, http://www.zope.org

（平成 20 年 8 月 8 日受付）

**石川　裕（正会員）** ishikawa@is.s.u-tokyo.ac.jp
1987 年慶應義塾大学大学院工学研究科電気工学専攻博士課程修了．工学博士．同年電子技術総合研究所入所．1993 年技術研究組合新情報処理開発機構出向．2002 年より東京大学大学院情報理工学系研究科コンピュータ科学専攻．教授．

**佐藤　充（正会員）** msato@labs.fujitsu.com
1969 年生．1992 年東京大学工学部電気工学科卒業．1997 年同大学院工学系研究科情報工学専攻博士課程修了．博士（工学）．同年富士通（株）入社．現在，（株）富士通研究所勤務．並列システムアーキテクチャの研究に従事．IEEE，ACM 各会員．

**中島　耕太（正会員）** kota@labs.fujitsu.com
2000 年九州大学工学部電気情報工学科卒業．2002 年同大学院システム情報科学府情報工学専攻修士課程修了．同年富士通(株)入社．現在，（株）富士通研究所勤務．高速通信機構に関する研究に従事．博士(工学)．

**住元　真司（正会員）** s-sumi@labs.fujitsu.com
1986 年同志社大学工学部電子工学科卒業．同年富士通（株）入社．1997 年より新情報処理開発機構に出向．2002 年より（株）富士通研究所にて高速通信機構の研究開発，大規模 PC クラスタ，UHPC システムの開発等に従事，並列分散システムのアーキテクチャ，システムソフトウェア等に興味を持つ．博士(工学)．

**久門　耕一（正会員）** kumon@labs.fujitsu.com
1979 年東京大学工学部電気工学科卒業．1981 年同大学院電子工学専門課程修士課程修了．1984 年同大学院博士課程中退．同年 (株) 富士通研究所入社．現在，同社 IT システム研究所に所属．CPU，メモリ，並列計算機アーキテクチャに関する研究に従事．GCC，Linux カーネル等の改良にも興味を持つ．

**木村かず子（正会員）** kimura@mxv.nes.nec.co.jp
1983 年日本電気ソフトウェア（株）（現：NEC ソフト（株））入社．NEC の汎用機，スーパーコンピュータの OS 開発に携わり，2004 年から Linux カーネルのダンプ機能の開発にも参画．

**鈴木　敦夫** suzuki-a@mxb.nes.nec.co.jp
1980 年日本電気ソフトウェア（株）（現：NEC ソフト（株））入社．NEC の汎用機の携わり，1999 年から NEC グループの Linux 関連ビジネスの立ち上げに参画．特定非営利活動法人 LPI-Japan 理事．