〈基本ソフトウェア〉

SX-8オペレーティングシステムSUPER-UXの概要

Summary of Operating System SUPER-UX for SX-8

梁川貴志* Takashi Yanagawa　宮崎恵美子* Emiko Miyazaki　大谷敦久* Atsuhisa Ohtani　長谷川晶一** Syouichi Hasegawa

要　旨

SX-8のオペレーティングシステムSUPER-UXは，SX-5/6/7シリーズで信頼と実績を築き上げてきたSUPER-UXをさらに強化したものです。

SUPER-UXでは，従来からの特長である，高速・大規模システム，高信頼性を引き継ぎながら，いっそうの大規模システムでの使い勝手の良さ，運用管理の充実を追求しています。

本稿では，SUPER-UXの特長と GFS，NQSII の強化内容について述べます。

The SUPER-UX of SX-8 provides the most advanced supercomputing environment which has been enhanced from the matured SUPER-UX of SX-5, SX-6 and SX-7 Series.

The SUPER-UX realizes user friendliness of UNIX for larger system and substantial system administration functions retaining high reliability, high-performance and large-scale system of the existing SUPER-UX.

This paper describes the features and characteristics of SUPER-UX that realizes the higher-speed and larger-scale SX-8 System, the features of GFS and NQSII.

1．はじめに

ハードウェア技術の進歩による，スーパーコンピュータの計算能力の向上・コストパフォーマンスの向上は，官公庁や大学の計算センターなどの大規模ユーザから，民間企業，さらには研究室単位の導入まで，その利用分野を大きく拡大させています。

近年，HPC分野では，ノード内の共有並列では並列数の上限があるため，ノード数を増やすことによって並列数を上げるマルチノードシステムが主流になり，そのノード数も増えつつあります。

このため，単なる大規模化，高速化への対応だけではなく，システムの導入から運用，プログラムの開発をより容易にかつ柔軟に行えるとともに，標準化・オープン化に対応していくことが，スーパーコンピュータにとってますます重要となっています。

本稿では，スーパーコンピュータSX-8のオペレーティングシステムSUPER-UXの特長である，大規模・高信頼性，高速性，使い勝手の良さ，充実した運用管理，新標準やオープンシステムへの対応と，最近の強化内容を紹介します。

2．SUPER-UXの特長

SUPER-UXは，オペレーティングシステムとして業界標準となっているUNIX System Vをベースに，BSDおよびSVR4.2MPからも機能を取り込み，さらにスーパーコンピュータに必要な機能を強化したオペレーティングシステムです（**図1**）。

SX-8のSUPER-UXは，SX-5/6/7シリーズでもそのまま動作することが可能です。これは，最新の機能をより多くのSXシリーズで利用できると同時に，SX-5/6/7シリーズからの互換性を維持しながら SX-8への移行を保証しています。

またSUPER-UXは，カーネル自身の高並列化によりシングルノードシステムモデルで最大32CPU（SX-7）までをサポートし，マルチノードシステムにおいては最大512ノードをクラスタ接続することにより，最大4,096CPU（SX-8）までをサポートしています。

さらに，大規模なメモリへの対応としてシングルノードシステムで最大128Gバイト（SX-8），512ノードのマルチノードシステムで最大メモリサイズ64Tバイト（SX-8）を実現しています。

このようにSUPER-UXは，シングルノード32CPUから大規模ノードまでの高スケーラビリティを保障するために柔軟な資源管理，カーネル・I/Oの高い並列処理性を持っています。

またノード数が多くなってもプログラムの実行や管理が煩雑にならないように各種運用ツールの強化を図っています。

＊ 第一コンピュータソフトウェア事業部
1st Computers Software Division

＊＊ NECシステムテクノロジー　サーバソフトウェア事業部
NEC System Technologies, Ltd.

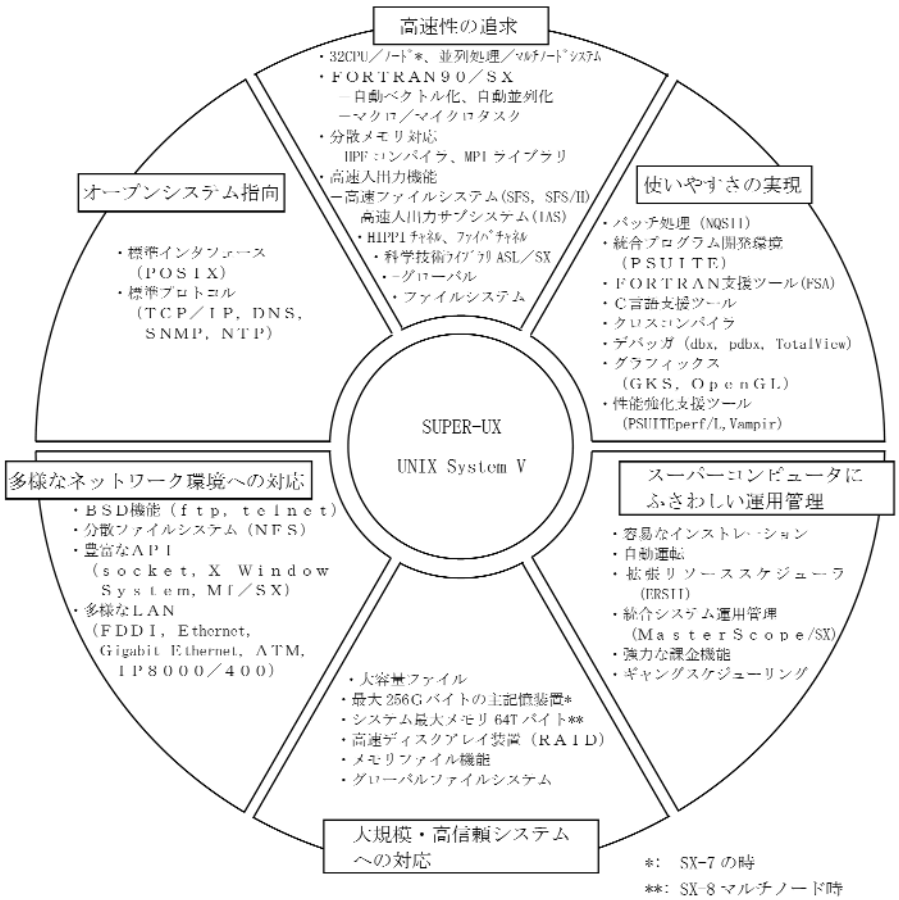

図1　SUPER-UXの特長
Fig.1　Characteristics of SUPER-UX.

2.1　大規模・高信頼性システムへの対応

（1）大規模メモリへの対応

1）大規模ページサポート

一般コマンド用の32Kバイト，コンパイラやシステムコマンドのための4Mバイトに加え，大規模なユーザプログラムのために16Mバイト（または64Mバイト）の3種類のページサイズをサポートしてします（16Mバイトと64Mバイトの切り替えは HW の SG で行います）。これにより，大配列を使用するプログラムの実行性能の向上を図るとともに，メモリ管理のオーバヘッドを削減しています。

2）プロセスサイズの拡大

単体のプロセスで主記憶を限界まで使用するようなプロセスを生成することが可能です。これにより，巨大なメモリを必要とする大規模なユーザプログラムも，容易に実行できるようになります。

3）仮想空間の拡大

ユーザ仮想空間は，約800Gバイトまで使用できます。これにより，巨大な主記憶を効率よく使用できるメモリレイアウトをサポートしています。

（2）メモリファイル機能

大容量のメモリ上に，通常のファイルシステムを構築して高速の入出力を可能にすると同時に，ディスクキャッシュとして利用することもできるSX-MFF（SX Memory File Facility）を提供しています。

（3）SX-Backstore

大規模階層記憶装置管理として，ファイルアーカイビングシステムSX-Backstoreを提供しています。

SX-Backstoreは，ファイルを他の大容量ストレージへマイグレーション（データをコピーし，元のデータを開放）することで，普段使用するファイルシステムの空き領域を増やすことができます。

マイグレーションされたファイルをユーザがアクセスした場合には，データを自動的にリコール（元のファイルシステムへのデータの復帰）を行います。SX-Backstoreを使用するためには，ファイルシステムを再構築する必要もなく，マイグレーションとリコールはシステムが自動的に行うため，ユーザが意識することもアプリケーションプログラムを変更する必要もありません。

マイグレーションされたデータを格納するストレージとしては，Storage Technology社のテープライブラリ装置，LTO2コンパクトオートローダ装置が使用できます。

2.2　高速性の追求

（1）IXS制御機能

マルチノードシステムでは，IXS（Internode Crossbar

Switch）という超高速のノード間接続装置をサポートしています。

通常，MPI（Message Passing Interface）やHPF（High Performance Fortran）でのノード間のメッセージ転送が発生すると，データ転送のためにはOSを呼び出す必要があります。しかし，SUPER-UXがサポートしているIXS制御機能を用いることにより，ユーザプログラムが他のノードに，OSを介することなく直接データを転送することができるため，MPIやHPFなどの分散並列プログラムを効率よく動作させて，高性能を達成しています。

また，IXS上でTCP/IP（Transmission Control Protocol/Internet protocol）を実現することにより，ftp（file transfer protocol）によるファイル転送やNFSによるファイル共有も，高速に実現しています。

（2）大規模SANにおける高速ファイル共有

ファイバチャネルによる大規模SAN（Storage Area Network）環境において，SXシリーズのマルチノードシステムだけでなく，Linux，Solaris，HP-UXなど，マルチプラットフォームでの高速ファイル共有を可能にしました。

（3）ネットワーク

ネットワークでは，10 BASE/100 BASE-TX Ethernetおよび，より高速なネットワークドライバとして，ギガビットEthernetをサポートしています。

2.3 使いやすさの追求

（1）MasterScope/SX

SUPER-UXでは，ユーザサイトで容易にシステムチューニングが行えるよう，各種システム統計情報や性能解析ツールを用意しています。

たとえば，MasterScope/SX（統合ネットワークシステム運用管理機能）は，分散ネットワーク管理の業界標準であるHP OpenViewをプラットフォームに採用し，SX-8からPC，LAN接続までの分散システム管理を支援するミドルウェアです。MasterScope/SXを用いることにより，GUIを用いたグラフィカルな表示・入力で，性能管理・障害管理・ハードウェア構成管理・運転管理を行うことができます。

（2）IOXソフトウェア

IOX（Integrated Operation Station for SX）ソフトウェアのインストールでは，シングルノードはもちろんマルチノードにおいてもWebによるインストールが可能です。これは，Internet ExplorerやNetscape Navigatorをブラウザとして使用することで，グラフィカルなユーザインタフェースによるインストールを実現し，導入から立ち上げまでを短時間で行うことができます。

さらに修正物件配布ツールがサポートされ，マルチノードにおいて修正物件を一度に配布・適用することが可能となり，システムの保守性も向上しています。

IOXとしては，HP-UXを搭載したスケーラブルサーバTX7によるTX7版IOXおよびOSとしてLinuxを採用したPC版IOXをサポートしています。

特にPC版IOXでは，OSとしてRed Hat Linuxを採用し，大幅な価格低減・小型化を実現しました。

2.4 スーパーコンピュータ/高性能計算サーバにふさわしい運用管理

ギャングスケジューリング機能により，並列処理やMPIプログラム内のタスクやプロセスを同時に実行/停止するように制御することで，複数の並列プログラムが同時に実行されても，CPUを効率よく割り当てて，スループットを向上させることができます。

ギャングスケジューリングはマルチノードで実行するMPIプログラムに対しても有効です。

3．GFS

近年，HPCの分野では，多数のノードからファイルを共有し，かつ各ノードからこれらファイルに高速にアクセスする機能を求められることが一般化しています。このような機能を実現するために，ファイバチャネルを用いたSANに基づく高速ファイル共有の機能としてGFS（Global File System）を開発してきました[1]。さらに，SXシリーズのノード間でのファイル共有に留まらず，Linuxを搭載したIPFマシンのTX7など，GFSのマルチプラットフォーム化にも取り組んでいます。

本章では，GFSのマルチプラットフォーム化，およびSX-6とSX-8のGFS性能の比較について言及します。

3.1 マルチプラットフォーム化への試み

主要な計算はSX-8で行い，それ以外の前処理や後処理において他のマシンを使用することが実ユーザでは少なくありません。システム全体としてみた場合，SX-8間でのファイル共有だけではユーザの使い勝手の面で必ずしも良いとは言えません。このため，**図2**に示したようにSX-8と他のマシン間においてもGFSを用いてファイルを共有できることが望ましいと考えられます。

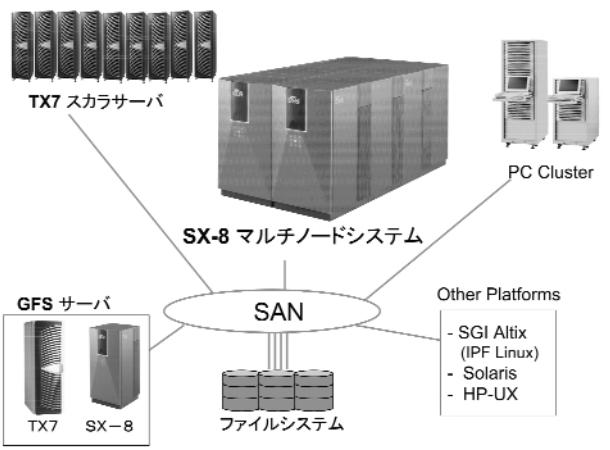

図2　マルチプラットフォーム化の概念図
Fig.2　Conceptual diagram of GFS on multi-platform environment.

GFSでは，サーバ，クライアント間で制御情報を送受信することによりGFSクライアントとディスク装置間での直接データ転送を実現[1]していますが，この制御情報は，特定のオペレーティングシステムやプラットフォームに依存しない形式になるように設計しているため，原理的には他のプラットフォームとの間でGFSを構成することも可能です。

そこで，すでに実現しているTX7上のNEC IA-64 Linuxに加えて，IA-32版Linux，SGI Altix上のIPF Linux，HP-UX，SolarisにおいてGFSクライアントの機能を実現しました。このうち，IA-32版LinuxとIPF Linuxについてはカーネルモジュール，HP-UX，Solarisについてはライブラリ機能としてインプリメントし，単体ディスクへのI/O性能としては，SX-8やTX7とほぼ同等な性能が得られていることを確認しました。

3.2 SX-6との性能比較

SX-6のI/OチャネルはPCIでしたが，SX-8ではPCI-Xが採用されています。また，これに伴いチャネルドライバの処理の効率化もなされています。そこで，SX-6とSX-8でGFSの基本性能にどの程度の違いがあるかを検証するため両者の性能の比較を行いました。基本性能を見ることが目的であるので，測定環境としては，1 LUN（Logical Unit Number），1ファイルシステム，ループバックマウントです。また，ディスク装置は，2Gビットファイバチャネルのディスク装置であるiStorage 2200（RAID 5, 4+P）を使用し，同ディスク上にSFS/Hファイルシステムを作成しました。

図3にSX-6，SX-8それぞれのリード性能を示します。I/Oサイズ128Kバイトで30%，1Mバイトで15%，8Mバイトで5%の性能向上が確認できました。今回は2Gb-FCでの測定ですが，今後4Gb-FCのサポートを予定しているため，I/Oチャネルの帯域が広がったことによる一定の効果が確認できたことは，今後システムの大規模化，高速化に貢献できると考えられます。

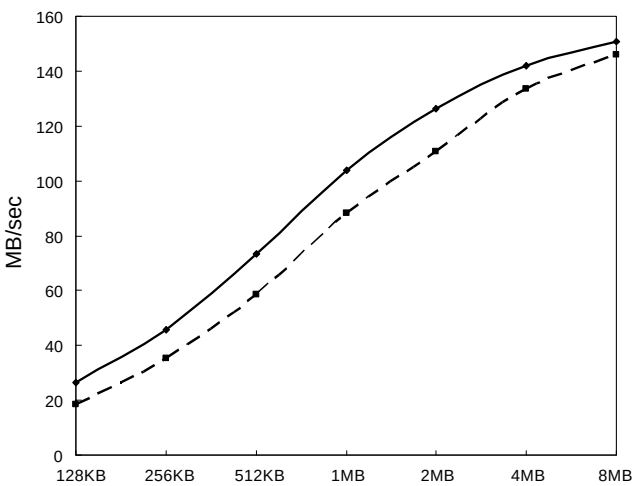

図3　SX-6とSX-8のGFS性能の比較（実線，点線はそれぞれSX-8，SX-6のリード性能を示す）

Fig.3　GFS performance comparison of SX-6 and SX-8 (Solid line and dotted line denote SX-8 and SX-6 read performance, respectively.)

4．バッチジョブ管理システム

SUPER-UX NQSII（Network Queuing System II）は，ハイ・パフォーマンス・クラスタシステム計算リソースを最大限活用するためのバッチ処理システムです。NQSIIは，クラスタシステムを構成する計算ノードのワークロードをモニタリングし，クラスタシステム全体のロードシェアリングを実現します。また，SSI（Single System Image）を強化し，SSE（Single System Environment）を提供することによりシステムの運用性を向上しています。

4.1 NQSII

NQSIIの全体構成は**図4**のようになります。NQSIIシステムは，バッチリクエストを管理するバッチサーバホスト，ジョブの実行を管理するジョブサーバホスト（実行ホストまたは計算ノード）および，バッチリクエストを投入するクライアントホストから構成されます。クライアントホストは，NQSII/CUIをインストールすることにより，管理ホスト，実行ホスト以外に設定することができます。これにより，利用者は管理ホストへ直接アクセスすることなく，利用者のワーキングホストからバッチリクエストをキューにリモート投入し，実行状況参照や必要な操作ができます。

またNQSIIでは，グラフィカルユーザインタフェースNQSII/GUIをサポートしています。GUI機能により，利用ユーザに対してバッチリクエストの実行状況を監視・操作するためのユーザコンソール機能，システム管理者に対してバッチリクエスト監視・操作および，システム負荷状況のモニタリング，各種設定機能の使いやすさを向上させることができます。

その他のNQSIIの主要機能を紹介します。

(1) 動的なリソース管理機能

SUPER-UX NQSII は，ジョブサーバホスト上で実行されているジョブが使用中のリソース（CPU数，メモリ量）を一定時間間隔でモニタリングしリソースの使用状況をリアルタイムに管理します。これにより，精度の高いロードバランス機能をサポートし実行ホストの持つ高価な計算リソースを最大限利用できるようにしています。

(2) ファイルステージング機能

SUPER-UX NQSII は，ジョブの実行に関係するファイル群をクライアントホストと実行ホスト間で転送するファイルステージング機能をサポートしています。これにより，ジョブへの入力となるデータファイルをジョブ実行ホスト上にあらかじめ展開することができます（ステージイン機能と呼びます）。また，ジョブ実行後には任意のジョブ出力ファイルだけをユーザが指定した場所へ返却することもできます（ステージアウト機能と呼びます）。

(3) 動的ジョブマイグレーション機能

SUPER-UX NQSIIでは，あるノード上で実行中のジョブを他のノード上へ動的に移動させるジョブマイグレーショ

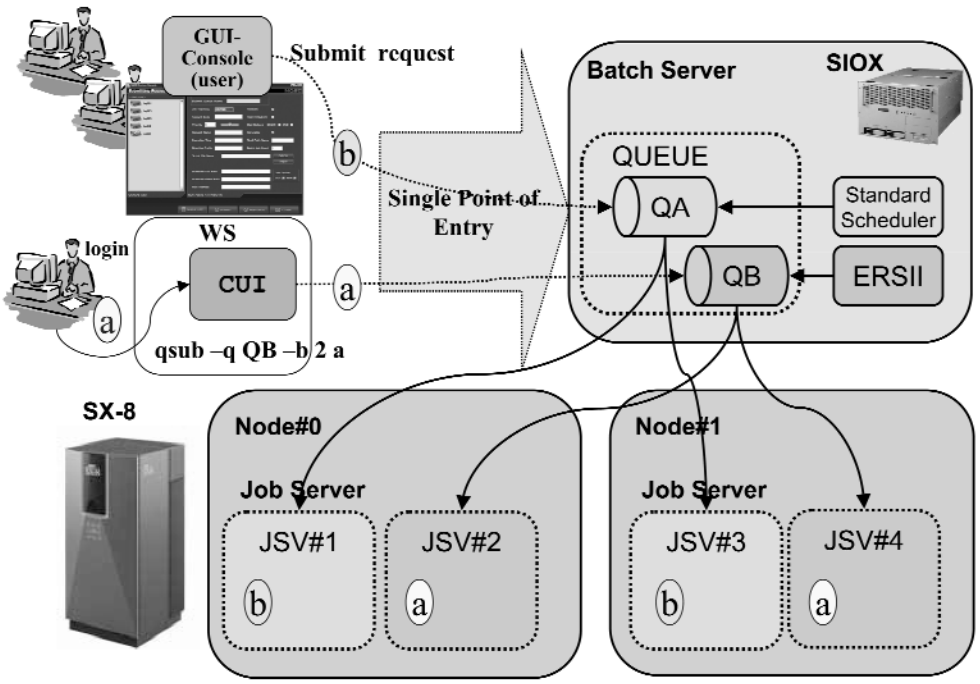

図4　NQSIIシステムイメージ
Fig.4　NQSII system image.

ン機能をサポートしています。これによりさらに精度の高いロードバランスを実現できるようになります。

また，SX-8ではジョブマイグレーションがより確実に成功するための機能強化をしています。ひとつには，マイグレーション時に任意のデータファイルもマイグレーション先ノードへ転送し，ジョブ実行に必要不可欠なデータファイルがマイグレーション先でも確実に参照できるようにしています。さらに，複数ノードで構成されるクラスタ環境下で，ジョブに割り当てられるプロセスID，ジョブIDをユニークに保つマルチノードIDリザーブ機能を実現しており，異なるノードへのジョブマイグレーションがID重複により失敗することを防いでいます。

(4) スケジューリング機能

SUPER-UX NQSIIでは図4のようにサイトの運用形態やポリシーに合わせて異なる複数のスケジューラをキュー単位に運用できます。またサイトスケジューラの構築，評価および，複数スケジューラの運用により柔軟なバッチ処理システムをサイト独自に構築することができます。

NQSII標準スケジューラでサポートするスケジューリング機能は，バッチリクエストを実行するジョブサーバをリソースベースに動的に割り当てるデマンドスケジューリングをサポートします。

4.2　ERSII

NQSII専用に高度なスケジューリングを実現するSUPER-UX ERSII（Enhanced Resource Scheduler II）を提供しています。ERSIIは，フェアシェアスケジューリング機能とジョブ実行ホストのロードバランス機能を実現しています。

フェアシェアスケジューリング機能では，ユーザ，グループ，およびアカウントコードを単位としたリソース使用量を常に監視し，それぞれに対して公平にリソースが分配されるようジョブの実行優先度（ジョブランク）を算出し，ジョブランクの高いジョブから実行するスケジューリングができます。

ロードバランス機能では，実行ホストのCPU，メモリ，およびMFFの使用状況を監視し，リソース使用量の少ない実行ホストからジョブを割り当て，実行します。また，各実行ホストのCPU，メモリなどの使用状況に応じて動的にジョブのホールド，リリース，ジョブマイグレーションなどを実施し，実行ホストのリソース使用量が設定された閾値を超えないように制御するとともに，各実行ホストの負荷を均等化することができます。

5．むすび

以上，スーパーコンピュータSX-8のオペレーティングシステムSUPER-UXについて紹介しました。ハードウェアの持つ性能を最大限に引き出すためには，スーパーコンピュータのハードウェアの技術の進歩に合わせて，オペレーティングシステムなどソフトウェアの技術の進歩が必要です。

今後も，高性能なスーパーコンピュータの適用分野がますます広がり，オペレーティングシステムへの要求もさらに高くなるものと思われます。ユーザ要求を先取りし，市場動向，技術動向を見極めつつ，一歩先んじたSUPER-UXの開発に努力する所存です。

参考文献

1） A. Ohtani, et al, “A File Sharing Method for Storage Area Network and Its Performance Verification”, NEC Res. & Develop., Vol.44, No. 1, pp.85-90, Jan. 2003.
2） M. A. Baker, G. C. Fox and H. W. Yau, “Cluster Computing Review,” Northeast Parallel Architectures Center http://www.npac.syr.edu/techreports/hypertext/sccs-748/cluster-review.html
3） COSMIC発行「The Network Queuing System」.

筆者紹介

Takashi Yanagawa
梁川　貴志　1990年，NEC入社。現在，コンピュータソフトウェア事業本部第一コンピュータソフトウェア事業部エキスパート。電子情報通信学会会員。

Emiko Miyazaki
宮崎恵美子　1987年，NEC入社。現在，コンピュータソフトウェア事業本部第一コンピュータソフトウェア事業部主任。

Atsuhisa Ohtani
大谷　敦久　1991年，NEC入社。現在，コンピュータソフトウェア事業本部第一コンピュータソフトウェア事業部主任。電子情報通信学会，日本リモートセンシング学会各会員。

Syouichi Hasegawa
長谷川晶一　1991年，NECソフトウェア神戸入社。現在，NECシステムテクノロジー サーバソフトウェア事業部テクニカル基盤ソフトウェアグループ主任。