平成 19 年 1 月 26 日

次期システムのご紹介

九州大学情報基盤センター　　計算機システム室長　　天野浩文[1]

本稿では，平成 19（2007）年 6 月 1 日稼働予定の次期スーパーコンピュータ・高性能アプリケーションサーバ両システムの概要を紹介します.

1. はじめに

現在運用しているスーパーコンピュータシステム VPP5000 およびスカラー並列サーバ GS320 は，平成 13 年 1 月に導入されたものです．これまで，本センター全国共同利用計算サービスの主戦力として多くの利用者ジョブを処理してきましたが，導入後 6 年を経過して，最先端の科学技術計算の需要を満たすのは難しくなっております．そこで，VPP5000 および GS320 は平成 19 年 2 月末をもって運用を終了し，システム更新作業に入ります．新たに導入される 2 つのシステム：

- **次期スーパーコンピュータシステム**
- **高性能アプリケーションサーバシステム**

は，どちらも現在の最先端の計算機技術を採用した，現有システムよりもはるかに大規模かつ高性能のシステムとなっております.

2. 次期スーパーコンピュータシステムの概要

次期スーパーコンピュータシステムは，計算サービスの主力となる 2 種類のバックエンドサーバ，フロントエンドサーバおよびファイルサーバ，ディスクアレイ装置とバックアップ装置から構成されます．本システムは，平成 19（2007）年 6 月から平成 23（2011）年 2 月まで，45 ヶ月間運用する予定です.

2.1 バックエンドサーバ A

バックエンドサーバ A は，富士通株式会社製の共有メモリ型並列計算機 PRIMEQUEST580（図 1）32 ノードです．各ノードは，Intel Itanium2 プロセッサ 1.6GHz（デュアルコア）を 32 プロセッサ（＝64 コア）搭載します．各ノードの主記憶容量は 128GB です.

PRIMEQUEST580 の特色は，各ノードが大規模 SMP（symmetric multi-processor）であり，各ノード内のどのプ

図 1：PRIMEQUEST580
（写真提供：富士通株式会社）

[1] amano@cc.kyushu-u.ac.jp

ロセッサから当該ノードの主記憶のどの部分にアクセスする場合でも，アクセス時間に差異がないことです．この性質は，プログラムを多数のスレッドによって並列化した場合の安定した性能向上に大きく寄与します．

バックエンドサーバ A の性能諸元を表 1 に示します．

表 1：バックエンドサーバ A の性能諸元

演算ノード	富士通株式会社 PRIMEQUEST580 Intel Itanium2 プロセッサ 1.6GHz（デュアルコア）×32 プロセッサ（＝64 コア） 主記憶容量 128 GB
総ノード数	32 ノード
総プロセッサ（コア）数	1,024 プロセッサ（2,048 コア）
理論演算性能の総和	13.1 TFLOPS
主記憶容量の総和	4 TB
相互結合網	ノードあたり InfiniBand 4x DDR（20 Gbps）×4 ポート（80 Gbps） ＝片方向 40Gbps（冗長符号分を除くと理論転送性能は片方向 4GB/s） スイッチ：SiverStorm9240

バックエンドサーバ A に導入される主なソフトウェアを表 2 に示します．

表 2：バックエンドサーバ A のソフトウェア

オペレーティングシステム	Red Hat Enterprise Linux AS (v.4 for Itanium)
ファイルシステム	Parallelnavi SRFS (Shared Rapid File System) for Linux
バッチジョブ管理システム	Parallelnavi for Linux Advanced Edition
言語処理系	Parallelnavi Language Package for Linux Fortran 処理系（OpenMP 対応，自動並列化機能有り）， C 処理系（OpenMP 対応，自動並列化機能有り）， C++
メッセージパッシングライブラリ	MPI（動的プロセス生成などを除き MPI 規格 2.0 に対応）
数値計算ライブラリ	Parallelnavi Language Package for Linux （BLAS，LAPACK，ScaLAPACK，PARDISO 等を含む）
科学技術計算アプリケーション	Gaussian，GAMESS，Molpro，AMBER
その他	デバッグ・チューニングツール等

2.2 バックエンドサーバ B

バックエンドサーバ B として，富士通株式会社製 PRIMERGY RX200S3 192 ノードからなる PC クラスタを，2 セット導入します．PRIMERGY RX200S3 の単体は，図 2 に示すような外観のラックマウント型サーバです．

図 2：PRIMERGY RX200S3
（写真提供：富士通株式会社）

今回，本センターとして初めて全国共同利用大規模計算サービスのために PC クラスタを導入することにした理由は，ますます増大する計算需要にお応えしていくためにどうしても大規模な PC クラスタの導入が不可欠であったこと，および，研究室等で小規模なクラスタをお持ちの利用者が気軽にセンターを利用できるようにし

たいと考えたことです.

バックエンドサーバ B の性能諸元を表 3 に示します.

表 3：バックエンドサーバ B の性能諸元

演算ノード	富士通株式会社 PRIMERGY RX200S3 Intel Xeon プロセッサ 3.0 GHz（デュアルコア）×2 プロセッサ（＝4 コア） 主記憶容量 8 GB
総ノード数	192 ノード×2 セット
総プロセッサ（コア）数	384 プロセッサ（768 コア）×2 セット
理論演算性能の総和	18.4 TFLOPS
主記憶容量の総和	3 TB
相互結合網	ノードあたり InfiniBand 4x DDR（20 Gbps）×1 ポート ＝冗長符号分を除くと理論転送性能は 2GB/s スイッチ：SiverStorm9240

バックエンドサーバ B に導入される主なソフトウェアを表 4 に示します.

表 4：バックエンドサーバ B のソフトウェア

オペレーティングシステム	Red Hat Enterprise Linux WS (v.4 for x86, AMD64 and EMT64)
ファイルシステム	Parallelnavi SRFS (Shared Rapid File System) for Linux
バッチジョブ管理システム	Parallelnavi NQS for Linux V2.0, PBS Professional（グリッド用）
言語処理系	Parallelnavi Language Package for Linux Fortran 処理系（OpenMP 対応，自動並列化機能有り）， C 処理系（OpenMP 対応，自動並列化機能有り）， C++
メッセージパッシングライブラリ	MPI
数値計算ライブラリ	Parallelnavi Language Package for Linux（BLAS, LAPACK, ScaLAPACK, PARDISO 等を含む）
科学技術計算アプリケーション	GAMESS，CHARM，AMBER
その他	デバッグ・チューニングツール等，NAREGI ミドルウェア

2.3 フロントエンドサーバおよびファイルサーバ

利用者用のフロントエンドサーバおよびファイルサーバとして，バックエンドサーバ A のノードと同型で主記憶容量を 512GB に増強した富士通株式会社製 PRIMEQUEST580 を 1 台導入します.

本センターでは，この大規模共有メモリ型並列計算機を 2 つのパーティションに分割し，それぞれを 32 コア・主記憶容量 256GB の並列計算機として運用します．平常時には，一方をフロントエンドサーバ，他方をファイルサーバとして使用しますが，いずれかに障害が発生した場合には残りのサーバが機能を代行することができます.

各パーティションが 4 Gbps ファイバーチャネルインタフェースを 16 本ずつ搭載します.

フロントエンドサーバに導入される主なソフトウェアを表 5 に示します.

表 5：フロントエンドサーバのソフトウェア

オペレーティングシステム	Red Hat Enterprise Linux AS (v.4 for Itanium)
ファイルシステム	Parallelnavi SRFS (Shared Rapid File System) for Linux
バッチジョブ管理システム	Parallelnavi for Linux Advanced Edition
言語処理系	Parallelnavi Language Package for Linux Fortran 処理系（OpenMP 対応，自動並列化機能有り）， C 処理系（OpenMP 対応，自動並列化機能有り）， C++
メッセージパッシングライブラリ	MPI（動的プロセス生成などを除き MPI 規格 2.0 に対応）
数値計算ライブラリ	Parallelnavi Language Package for Linux （BLAS，LAPACK，ScaLAPACK，PARDISO 等を含む）
科学技術計算アプリケーション	Gaussian，α-FLOW, LS-DYNA, AutoDock, GAMESS, Molpro, AMBER, CHARMM，ANSYS CFX，FIELDVIEW，AVS，Materials Exploler
その他	デバッグ・チューニングツール等

2.4　ディスクアレイ装置およびバックアップ装置

利用者のデータを格納するディスクアレイ装置およびそのバックアップ装置として，富士通株式会社製 ETERNUS8000 モデル 2100 を 1 台導入します.

近年のディスクアレイ装置の大容量化に伴い，磁気テープメディアを使用したバックアップ装置ではバックアップ作業に多大な時間を要するという問題が顕在化しつつありました．しかし，利用者のデータの安全な保管は本センターの重要な任務であり，バックアップ作業を完全に省略することには大きな問題があります．このため，今回ついに，磁気ディスクのバックアップに磁気ディスクを採用することといたしました.

RAID レベル 5 構成時の総実効容量約 500 TB のうち，半分の 250 TB をオンラインディスクとして，残りの 250 TB をバックアップディスとして使用します．バックアップディスクは，バックアップ作業時間以外はスピンドルの回転を止めることができます.

ディスクアレイ装置全体で 4 Gbps ファイバーチャネルインタフェースが 32 本搭載されます.

3.　高性能アプリケーションサーバシステムの概要

前述の次期スーパーコンピュータとは別に，新たに高性能アプリケーションサーバシステムを導入します．これは，平成 19（2007）年 6 月から平成 21（2011）年 2 月まで 21 ヶ月間運用したのち，現行の高性能演算サーバ（IBM eServer p5 モデル 595）と合わせて次々期システムへ更新する予定です．これにより，平成 21 年以降は，ほぼ同規模の計算機システムが 2 年に 1 度，交互に更新されるようになります.

3.1　バックエンドサーバおよびフロントエンドサーバ兼ファイルサーバ

株式会社日立製作所製の共有メモリ型並列計算機 SR11000（図 3）のモデル J1 を 16 ノード，および，同モデル K2 を 8 ノード導入します.

このうち，モデル J1 の 1 ノードがフロントエ

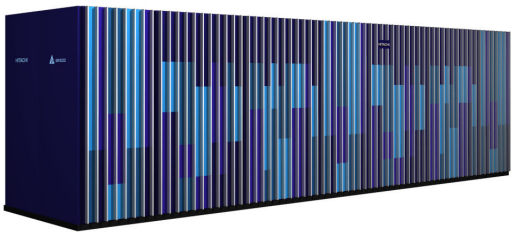

図 3：SR11000
（写真提供：株式会社日立製作所）

ンドサーバ兼ファイルサーバとなり，モデル J1 の残り 15 ノードと K2 の全 8 ノードを合わせたものがバックエンドサーバとなります.

各ノードは POWER5／POWER5+プロセッサ 8 個（16 コア）を搭載した SMP であり，各ノードの主記憶容量は 128GB です．ノード間は 4GB/s（片方向）×2 のクロスバーネットワークで接続されます.

バックエンドサーバの性能諸元を表 6 に示します.

表 6：高性能アプリケーションサーバの性能諸元

演算ノード	株式会社日立製作所 SR11000 モデル J1（16 ノード，うち 1 ノードがフロントエンド兼ファイルサーバ） IBM POWER5 プロセッサ 1.9 GHz（デュアルコア）×8 プロセッサ（＝16 コア） 主記憶容量 128 GB モデル K2（8 ノード） IBM POWER5+プロセッサ 2.3 GHz（デュアルコア）×8 プロセッサ（＝16 コア） 主記憶容量 128 GB
総ノード数	（バックエンドサーバ合計）15＋8
総プロセッサ（コア）数	（バックエンドサーバ合計）184 プロセッサ（368 コア）
理論演算性能の総和	（バックエンドサーバ合計）3 TFLOPS
主記憶容量の総和	（バックエンドサーバ合計）2.9 TB
相互結合網	専用クロスバーネットワーク ノードあたり 4 GB/s（片方向）×2

高性能アプリケーションサーバシステムに導入される主なソフトウェアを表 7 に示します.

表 7：高性能アプリケーションサーバのソフトウェア

オペレーティングシステム	AIX 5L 5.3
ファイルシステム	GPFS (General Parallel　File System)
バッチジョブ管理システム	LoadLeveler
言語処理系	最適化 Fortran90（OpenMP 対応，自動並列化機能有り）， XL C/C++（OpenMP 対応，自動並列化機能有り）
メッセージパッシングライブラリ	MPI（動的プロセス生成などを除き MPI 規格 2.0 に対応）
数値計算ライブラリ	BLAS，LAPACK，ScaLAPACK，PARDISO MATRIX/MPP，MATRIX/MPP/SSS，MSL2
科学技術計算アプリケーション	Gaussian，GAMESS，Molpro，AMBER，CHARMM，TINKER，VASP，PHASE，MSC.Marc/Mentat，MSC.Nastran，MSC.Patran，CONFLEX，CFX，IDL
その他	デバッグ・チューニングツール等

3.2　ディスクアレイ装置およびバックアップ装置

ディスクアレイ装置には，株式会社日立製作所製の SANRISE AMS500 を 2 台導入します．実効容量の総和は，20.7TB となり，SR11000 とは，4 Gbps ファイバーチャネルにより接続されます.

また，このディスクアレイのバックアップ装置として，ソニー株式会社製のテープライブラリ PetaSite S60 を導入します．実効容量の総和は，非圧縮で約 10 TB です.

4. おわりに

本稿では，平成 19（2007）年 6 月 1 日稼働予定の次期スーパーコンピュータ・高性能アプリケーションサーバ両システムの概要を紹介しました.

また，今回の更新に合わせ，計算機システムの月額レンタル料金の予算を見直し，運転に必要な電力料金をできるだけ本センターに交付される運営費交付金の中から負担できるようにいたしました．これにより，電力料金の不足分にあてるため利用者の皆様にこれまでお支払いいただいていた「利用負担金」の金額を大きく引き下げることができる見込みです．大規模かつ高性能のシステムの登場と合わせて，どうぞご期待ください.

これらのシステムと新たな利用負担金制度が皆様の研究活動を大きく前進させる力となりますよう祈念いたしております.

なお，今回の更新から，従来のように後継機を導入したのちに旧機器を撤去する空間的な余裕がなくなってしまいました．また，電源設備や空調設備も現有のままでは次期システムを安全に運用できなくなってしまいました．このため，旧システムの解体・撤去を行った後に，計算機室拡張工事・電源設備改修工事・空調設備増強工事などの一連の工事を行ってから，新システムを導入することとなりました，これに伴い，平成 19 年 3 月～5 月の 3 ヶ月間のサービス停止期間[2]が必要になり，利用者の皆様に大変ご迷惑をおかけいたしますが，どうか事情をご理解の上，今しばらくご辛抱くださるようお願い申し上げます.

[2] 高性能演算サーバ IBM eServer p5 モデル 595 は，この間もサービスを継続します.