SR8000 システム導入について

大阪大学レーザー核融合研究センター　○大橋裕子、福田優子、斎藤昌樹
広瀬華子、長友英夫、西原功修

1. はじめに

平成 12 年 3 月 1 日に、大阪大学レーザー核融合研究センター（以下、当センターと略す）のコンピュータシステムが、日本電気製 SX4/2C（2GFLOPS×2CPU、主記憶 1GB） を中心とするシステム（以下､旧システムとする）から、日立製作所製 SR8000（12GFLOPS、主記憶 8GB）コンパクトモデル を中心とするシステム（以下、新システムとする）に更新され、運用を開始した。

本システムの利用目的は、当センターで行われる実験管理、データ管理・解析、プログラム開発、画像処理、およびサイバーメディアセンターのスーパーコンピュータのデータ処理に大別される。

本システムの導入にあたって特に重視したのは、新しい研究の展開に伴う計算ニーズの増大に対応できる性能・機能であった。システム更新に伴い、ホストコンピュータや各種サーバの記憶容量、および補助記憶装置容量が増加し、可視化装置も旧システムより増強され、今後の研究の進展が期待される。

2. システム構成

システムの基本的な構成は旧システムと同じ考え方で構成した（機能階層別構成　参考文献[2]）。新システムは、ホストコンピュータ SR8000（12GFLOPS,8GB）、ファイルサーバ、プログラミングサーバ、グラフィックサーバ・端末、そして 10 台の端末クライアントから構成される。(図 1 参照)

旧システムでは、ファイルサーバとプログラミングサーバ、グラフィックサーバの機能を 3 台のホスト系 WS と呼ばれる EWS に行わせていたが、高負荷による問題がたびたび発生していたため、新システムでは、それぞれの機能別にサーバを用意した。ホストコンピュータ、ファイルサーバへのユーザのログインは認めない運用形態をとっている。

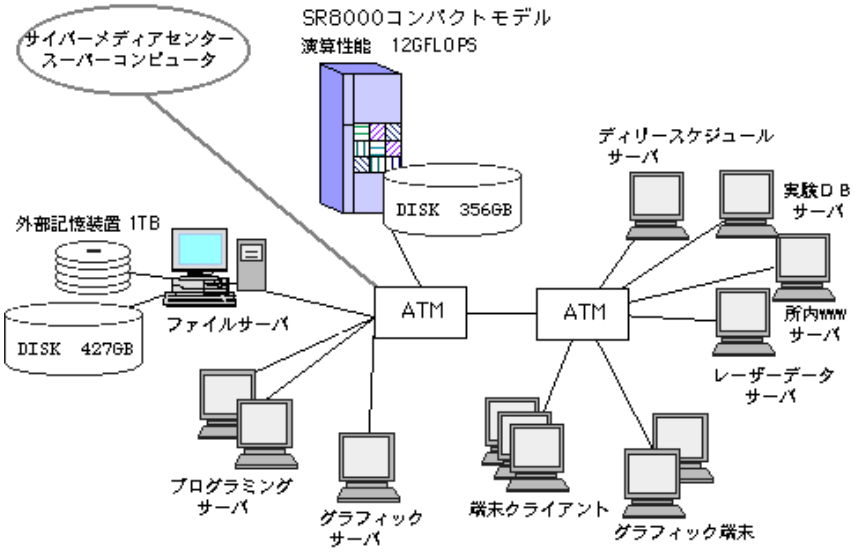

図 1. システム構成図

3. 導入に際して

日本電気（以下、NEC とする）から日立製作所（以下、日立とする）へのメーカー変更により、プログラム修正の必要性や運用ツールの動作確認など、利用者および管理者の負担が予想されたので、移行への準備は落札決定時より速やかに開始した。

3.1 WG（ワーキンググループ）の発足

日立に落札決定後、直ちに、全体、運用、ネットワーク、プログラム、可視化、DB-G、DB-M の 7 つの WG を立ち上げた。これらは当センターおよび日立の、それぞれに対応する担当者から構成されるものであり、実情に即した実のある話し合いができ、スムーズなシステム移行を行うことができた。

3.2 先行導入・テスト運用

平成 11 年 11 月には、プログラミングサーバとなる機種を先行導入し、ユーザ環境の確認等、運用に必要な環境構築を行った。

また、一般利用者への公開を行い、テスト的に利用してもらうことで、利用者からの意見も取り入れることができたことは有益であった。

4. システム構築・ファイル移行

システム構築に際し、我々の希望は、「現在のままの状態で利用できること」、特に、利用者が旧システムとほとんど違和感なく利用できることを目標とした。前回のシステム更新は、OS の変更ということで利用者はかなり苦労したが、幸い今回のシステム更新は、基本的な OS が UNIX 同士である。4 年間の UNIX システム利用によって利用者が十分 UNIX に慣れていたことも幸いし、利用者にとっては使用形態がほぼ同じであり、そういう意味ではメーカーの違いはさほど苦にならなかった。

SX4/2C から SR8000 への変更に加え、旧システムでは端末クライアントとして EWS を利用していたが、新システムでは HP マシンを始め、可視化用に SGI マシン、端末クライアントには日立製 PC を利用するという、異なる機種での運用を決めた。OS は旧システムと同じ UNIX、端末クライアントは Linux とした。

4.1 ディスク構成

当センターでは、ファイルサーバ上に利用者のホームディレクトリや様々な運用ツールを置き、NFS(Network File System)を用いることにより、どのマシンからログインしても同じ環境で利用できる。旧システムからのディレクトリ構成の変更がないように、構成を考えた。

テスト運用開始後、マシンが異なることにより、ログイン時の環境が若干異なることがわかり、source.cshrc を修正するなどの作業が生じた。

4.2 ファイル移行

ユーザファイル、および運用ツールの移行作業に当たっては、新システムとのファイルが混在することを避けるため、新ホームディレクトリとは別にディレクトリを作成し、そこへ移行することで、新ディレクトリを空の状態で使用できるようにした。本更新の機会に、利用者にディレクトリを整理してもらうことを期待しての試みであったが、この試みは十分成功したと言える。

5. プログラム移行

SX4/2C で実行していたプログラムを、SR8000 に則したプログラムへ移行する作業が発生することから、先行導入したプログラミングサーバにクロスコンパイラを早期に使える環境を用意し、利用者自身がコンパイルしてみることができるようにした。また、当センターの実験の為の基幹となるプログラムや、ライブラリを多種多様に使用しているプログラムなど、数本のプログラムを日立に提供し移行作業を依頼した。

その結果、標準的な FORTRAN の記述ならば問題ないことがわかった。しかし当然ながら、修正が必要なプログラムも存在し、SX4 から SR8000 に移行する際に特に修正が必要なプログラムの問題点が判明した。

・NEC 独自の拡張仕様を使用している
・NEC 固有のライブラリを使用している

移行のための変換ライブラリも用意し、プログラム修正なく、SX4/2C でも SR8000 でも実行できる環境を用意した。

その他、移行時にわかった事項は速やかに利用者へ情報提供を行い、ユーザ用テキストにも反映することができた。

6. ジョブ運用

6.1 検討段階で問題になった事柄

ジョブ運用について日立と検討を行った際に、特に、我々と日立との間で生じた初期のくい違いは、用語の意味の違い、あるいは我々の希望する機能が SR8000 にないことから生じるものが多かった。同じ言葉でも、その意味する事が若干違ったりすることにより、我々の希望がうまく伝わらない、といったようなことがおきた。

例えば、NEC では CPU と呼ぶものは、日立では IP であり、特に NQS（Network Queing System）関係の用語の定義が NEC と日立では随分異なっていた。疑問が生じるたびに問い合わせながら作業を進めた。

特に困ったのは、今まで行ってきたサスペンド機能を用いてのジョブ運用が、できなくなったことであった。現在は、サスペンド機能を用いない運用を行っているが、利用しにくいことは否めない。また、今まで利用していた、いくつかの便利なコマンドがないこともわかった。これについては、今までのものと同等の機能をもつコマンドを、日立に提供してもらうことで解決した。

6.2 NQS 利用情報について

qstat コマンドでは、現在実行されているジョブの、現時点でのメモリや CPU の使用状況がわからないため、新規にコマンドを提供してもらい、ジョブの現状を 5 分おきに採取するようにした。

この情報を定期的に採取することで、夜間にジョブトラブルが発生した場合でも、発生時間の絞込みや、

その時点で実行されていたジョブについての情報を得ることができる。
また、項番 7.2 にも示すが、NQS 利用情報を月ごとに集計し、利用状況を確認している。
採取している情報は、1) ユーザ毎のジョブ数、利用時間、最大使用メモリ、2) キュー毎のジョブ数、最大使用メモリ、総 CPU 利用時間、などである。
これらの情報を参考に、キュー構成の見直しや、ジョブ運用の検討を行っている。

6.3 現在のジョブ運用

旧システムで運用していたジョブ管理形態（参考文献[2]）を引き継ぎ、各ジョブの特性（CPU 時間、メモリサイズなど）に合わせたキュー設定とスケジューリング、ジョブコントロールなど、効率的な運用を行えるようにしている。現在は、主にメモリ使用量で分けられた 4 つのキューで運用している。SR8000 を 1 ノードで利用しているため、大規模ジョブとそれ以外の通常ジョブとの兼ね合いを考え、大規模ジョブ用キュー bl については申請制とし、他のジョブをチェックポイントリスタートさせることにより、夜間に 1 ノードを占有してジョブ実行させることにした。
旧システムで開発、運用していたモードという考え方を取り入れ、通常モード（ss、ms、ml のジョブ実行。昼間）と大規模モード（bl のジョブのみ。夜間）での運用を行っている。切り替えツールを日立に提供してもらい、そのツールを実行させることによって行っている。

7. システム運用

7.1 利用者への情報提供

プログラム移行を始め、移行のためのサービス停止、新システムの概要についてなど、利用者へのアナウンス事項は非常に多かった。重要なアナウンスも数多くあり、メールのみでは徹底できない場合は、当センターで行われている全体会議や所内 Web を活用し、できるだけ迅速に情報提供を行うよう努力した。
また、新システムの概要説明のための講習会の他に、日立による SR8000 についての講習会を実施してもらった。当センターで、実際にプログラム開発や解析を行っている利用者が出席し、SR8000 でのプログラミング方法について講習を受けた。その際に利用者から出された質問などは、その後のテキストに反映され、より利用者にわかり易いテキストを作り上げることができた。
新システム運用開始までに、計 4 回の講習会を行い、運用開始後も日立による講習会を実施した。
さらに、プログラミングに関しての質問に対しては、現在でも行われている日立との定例会終了後に、担当者に来てもらい回答してもらうという形をとり、それ以外でも随時、利用者の質問に応じられるような体勢をとっている。

7.2 稼動情報

旧システム時に、運用の自動化を試みた（参考文献[1]）。新システムでは、その仕組みを引き継ぎ、改良を加えた上で、より管理効率が上がるように、運用の自動化を図っている。新システムで、定型処理として自動的に採取している情報としては、1) ディスク使用量、2) ログイン情報、3) CPU/SWAP 情報、4) NQS 利用情報、5) SR8000 稼動時間　などである。それぞれ日次的、あるいは月次的に情報採取し、利用状況の確認やトラブル時の確認に役立てている。
それぞれの情報採取用ツール､および採取したログは､ファイルサーバ上で一元管理している｡NFS（Network File System）を利用し、ファイルサーバから rsh、remsh を使用してツール実行を行うことで、マシン毎に設定を行う手間を省いた。

8. 終わりに

新システムの運用開始後、早数ヶ月が経った。やっと落ち着いてきたかな、というのが正直な感想である。
OS の変更がないので簡単に移行できるだろうと安易に考えていたが、なかなか思うようにいかずに困ったことも何度かあった。その原因には、HP、SGI、PC など、数種の異なるメーカーの機種を使用したこと、当センター固有のシステム設定に慣れすぎていたこと、当初メーカーとの意思疎通がうまくいかなかったことなどが挙げられる。
しかし、旧システムでの運用形態をベースに、日立のサポートや利用者の協力により、より良いシステム構築を行えたと思う。今後もより効率的な運用、利用者にとって利用しやすい環境作りを目指し、努力していきたい。

参考文献

[1] 田村篤和、岡本匡代、福田優子、斉藤昌樹 ”分散処理システムにおける運用の自動化の試み”
1998 年度 高エネルギー加速研究機構技術研究会報告集、279 頁～282 頁
S.Tamura, M.Okamoto, Y.O.Fukuda and M.Saito "Automatic system management on workstations"
Proceedings of the Meeting on Engireering and Technology in Basic Research
pp279-282(1999.3) KEK, Tsukuba, Japan March4-5,1999

[2] 岡本匡代、福田優子、島田京子、直江正美、和田幸裕、田村篤和、西原功修
”SX4-2C システム導入について” 1996 年度 技術研究会東京分科会 1996.9.19-20

[3] 福田優子、澤井和美、藤井丈暢、安井秀一 ” 計算機システム運用の自動化(ILE オペレータ)”
技術研究会報告集 113 頁～114 頁(1990.12)岡崎国立共同研究機構 分子科学研究所