サイエンティフィック・システム研究会　　1999 年 8 月 20 日
科学技術計算分科会　第１回会合資料

大規模システムの運用

理化学研究所
情報基盤研究部
情報環境室
堀木　豪
horiki@postman.riken.go.jp

１．はじめに

理化学研究所では平成１１年２月ベクトル並列型スーパーコンピュータ VPP700E/128 を導入、運用を開始した。運用には情報基盤研究部情報環境室（情報環境室と略す）が当たっており、ユーザの利用動向の把握・解析、要望・問題解決への対応、ユーザ登録管理を遂行している。運用開始より 6 ヶ月が経過し、ここに現在までの実績と問題点をまとめ報告することにする。これにより大規模システム運用の今後の課題を明確にしていきたい。

２．システム概要

本研究所の大規模システムは、計算サーバとして大規模な科学技術計算（数値計算、シミュレーション、大容量データ処理）を行うことを目的とする VPP700E/128 を中心として、以下に示すハードウェアにより構成されている。図 1 にシステムの概要図を示す。

２．１　計算サーバ

サーバ計算機はベクトル並列型計算機 VPP700E/128 であり、CPU は１１８個の S-PE と１０個の IOPE（P-PE、IMPE×９）で計１２８個 PE（processor element）の構成である。各 PE 演算性能は 2.4GFLOPS で 2GB のメモリを有している。全体としての性能は 2.4×128=307.2GFLOPS、メモリ容量は 2×128=256GB であり、ディスクの容量は 4.4TB（GEN5 XLE）である。

２．２　フロントエンド計算機

フロントエンド計算機として S－7／7000Umodel450 を持ち CPU は UltraSPARC2（336MHz）×4、メモリ 4GB、ディスク容量 1.2TB（GEN5 XLE）を有し、一般利用者と管理者の端末からのアクセスを受け付けている。利用者のホームディレクトリ領域もここに確保しており、NFS（Network File System）マウントにより計算サーバからも同一パス名でアクセス可能である。

２．３　大容量アーカイブシステム

このシステムは大容量テープライブラリ装置 Powderhone/SD-3 Redwood（総容量 200TB）と S-7/7000Umodel450（UltraSPARC2（250MHz）×4CPU、メモリ 2GB、ディスク容量 1.6ＴB（GEN5 XLE））から構成され、利用者によるユーザデータの長期保存、ユーザ領域のバックアップに使用している。

２．４　リアルタイム可視化装置

計算サーバにより流体等のシミュレーション計算をしながら結果をリアルタイムで可視化するシステムで、SGI Onyx2 InfiniteReality（R10000（195MHz）×6CPU、メモリ 10GB、ディスク容量 100GB）を使用している。また、このサブシステムとして三次元可視化装置 SGI Onyx2 InfiniteReality（R10000（195MHz）×2CPU、メモリ 512MB、ディスク容量 18GB）を設置し、プロジェクターによりスクリーン上で三次元動画のデモ等に使用している。

２．５　その他

上記システム、装置は HIPPI スイッチおよび FDDI によるネットワークでお互いに接続され、高速データ転送を可能としている。

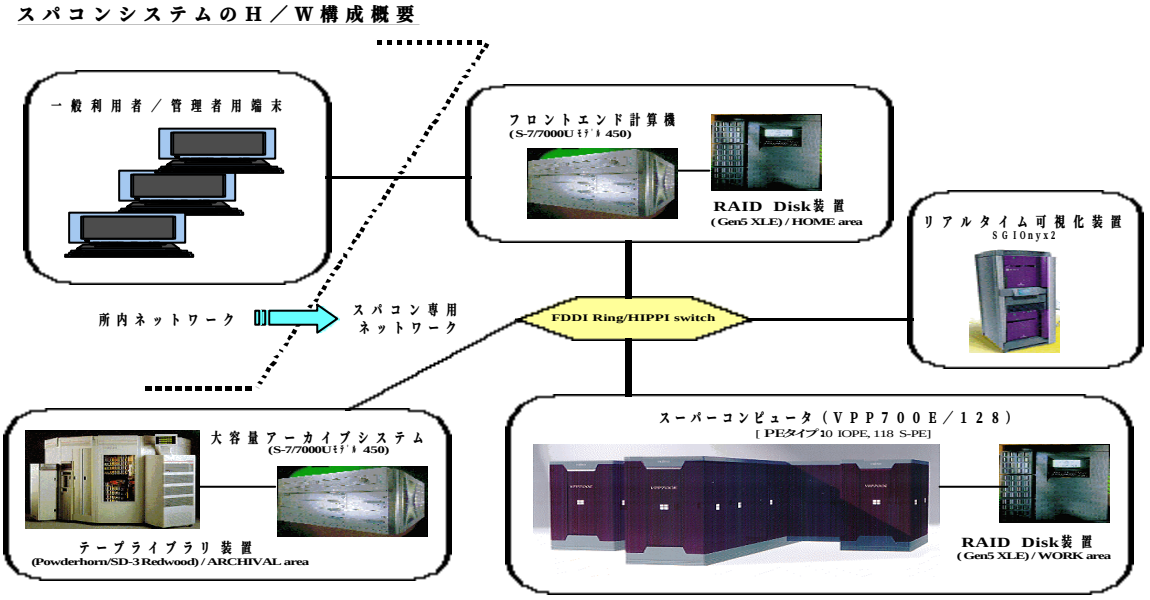

図 1

３．運用の方針と概要

情報環境室では大規模システムの運用にあたり、主要な方針としてユーザフレンドリィを謳っている。つまり利用者に快適な計算機利用環境を提供すること、質問・要望に迅速に対応することを第一にしている。さらに巨大ジョブ、長時間ジョブをより実行し易くする為に CPU 課金は行っていない。ただし、磁気ディスクリソースについては有限であるため利用容量に応じて段階的な課金を行っている。（表１参照）

以上を基に具体的には会話型のジョブグループを用意しコマンドレベルによるジョブの実行を可能にしている。もちろん複数ＰＥでの並列処理も可能である。また、多種のバッチジョブキューを用意しプログラムの規模に応じたキューの選択が可能である。少ないＰＥ数での長

時間ジョブや分子構造解析ソフトウェア GAUSSIAN 等のアプリケーションに対する特別なキューも用意し、利用者の要望に応えている。

現在の利用登録者数は約 300 名で、60 %の利用者が本研究所の外部からのアクセスにより利用している。本来セキュリティと利用の簡便さは常に背反するものであるが、このような状況の中でセキュリティは重要な問題である。本研究所ではフロントエンド計算機にアクセスする際にファイヤウォールを通過しなければならない。これは従来、計算機への接続、ファイル転送はユーザが自由に利用できてきたが、セキュリティ強化のため暗号化通信を行うツールによる接続、ファイル転送が必要となってきた。しかしユーザにとってこれは自分の端末パソコンやワークステーションにツールを設定し、今まで利用していたコマンドとは違うコマンドを使用するという負担がかかってくる。従って、セキュリティに関してはファイヤウォールを設け、認証サーバと連携した利用者認証の強化を図り、計算機接続、ファイル転送に関して、利便性を損なわない環境（今まで利用していたコマンドを利用できる）を構築した。

ユーザ領域はフロントエンド計算機に確保しており、各ユーザ当たり 1GB が利用可能である。ワーク領域はスーパーコンピュータに確保しており各ユーザ当たり 10GB 利用可能で、大容量の出力データの保持に使用され、必要な場合には各ユーザによりアーカイバに退避される。ユーザ領域のバックアップは、大容量アーカイブシステムにより、毎週末定期的に行っている。

■ディスク課金（１ヶ月当たり）

ホームディレクトリ	
0円	0KB≦使用量＜２５６KB
1円	２５６KB≦使用量＜1MB
2円	1MB≦使用量＜2MB
：	：

アーカイブディレクトリ	
0円	0KB≦使用量＜２５６KB
5円	２５６KB≦使用量＜１００MB
10円	１００MB≦使用量＜２００MB
：	：

表１

４．主要なサービス

情報環境室で行っているサービスとして、一般利用者向けに定期的な講習会を行っている。各種ドキュメント、利用手引は所内 Web サーバのホームページに掲載し、英文化も進めている。登録手続きなどの一部はメール、ホームページからの申請も受け付けている。

また一方では、運用管理を行う上で管理者側も運用支援という形で SI（System Integrator）の富士通から必要なサービスを受けている。このうちで利用者にとって有用なツールであれば管理用のツールも積極的に利用者に開放することにしている。

４．１ 利用者運用支援サービス

以下に利用者に対する支援サービスを列挙する。このうちプログラム相談、Ｑ＆Ａのメールでの対応は利用者にも浸透してきている。

1	一般利用者向けツール群	［1］現状表示ツール群 ■会話／バッチ型ジョブの待ち・実行状況 ■会話型ジョブの実行状況 ■会話／バッチ型ＰＥの振分け・使用状況
2	管理者向けツール群	［1］現状表示ツール群 ■会話／バッチ型ジョブの全体状況一覧 ■会話／バッチ型ジョブの待ち・実行状況 ■会話型ジョブの実行状況 ■会話／バッチ型ＰＥの振分け・使用状況 ■ログインＰＥ別のログイン利用者状況 ■ＰＥ／ジョブの割付け状況 ■バッチ型ジョブの待ち理由 ［2］統計表示ツール群 ■会話／バッチ型ジョブの日別ジョブ実行状況 ■会話／バッチ型ＰＥの日・時間別ＰＥ振分け・使用状況 ■ＣＰＵの日別稼働率 ■会話／バッチ型ジョブの日・時間別待ち・実行状況 ■長時間／大規模ＰＥ並列ジョブの日・キュー別実行状況
3	利用者向け講習会	■プログラミング初心者講習会 ＦＯＲＴＲＡＮ／ＶＰ、ＦＯＲＴＲＡＮ／ＶＰＰ、 Ｃ／ＶＰ、ＭＰＩ ■スパコン初心者利用講習会
4	プログラム相談	■質問窓口及びメールにて受け付け対応
5	その他Ｑ＆Ａ対応	■質問窓口及びメールにて受け付け対応

４．２ 連携支援サービス

利用者認証の強化については前項でも述べたが、もう少し説明すると、スーパーコンピュータ利用登録者に一人一人にセキュアＩＤカードを配布し、スーパーコンピュータに接続するとき毎にワンタイムパスワードを発行して利用することにしている。認証が確認されて始めてフロントエンド計算機へのログインが可能となる。FTP によるファイル転送を行う場合もこのワンタイムパスワードが求められる。

1	ネットワーク・ セキュリティ環境	■Ｆｉｒｅｗａｌｌ及びユーザ認証サーバを設置し、 ワンタイムパスワード利用によるセキュリティ強化 ■本研究所内外を問わずＦｉｒｅｗａｌｌを経由し、 以下のサービスが利用可能 ｔｅｌｎｅｔ ｆｔｐ ｈｔｔｐ
2	ジョブ操作環境	■フロントエンド計算機上からＮＱＳパイプキューを 利用した以下のバッチジョブ操作が可能 ジョブ実行 ジョブ状態表示 ジョブ削除
3	ファイル操作環境	■フロントエンド計算機上にホームディレクトリ領域を 配置し、ＮＦＳ共有によりスパコン本体と同一のパス 名でファイル操作が可能 ■ファイルサーバ上にテープアーカイバ用のディスク キャッシュ領域１ＴＢを配置し、ＮＦＳ共有によりスパ

		コン本体からのファイル操作で、テープアーカイバへのマイグレーション／リコールが可能
4	利用者向け最適化環境 (1) ログインホストのＰＥ自動選択 (2) スパコン本体から各ＷＳへの通信経路負荷分散	■９台のＩＭＰＥのＣＰＵ負荷により、ログインするＰＥが自動選択され、最適なＰＥへログイン可能 ■スパコン本体から各ＷＳ資源をネットワーク経由でジョブがアクセスする際、ネットワーク接続されている複数のＩＯＰＥのＣＰＵ負荷により、経路が自動選択され最適な経路でアクセス可能

４．３ ジョブ運用サービス

ジョブの効率的実行は運用に際しての重要ポイントである。従って、各ジョブキューでの実行待ち行列に対しても実行順序の公平性を図る種々の方策がなされている。たとえば同一利用者ジョブの連続実行抑止であるとか、最大待ち時間超過ジョブの優先実行である。しかし、これらの方策も飽和的利用状況では実行待ちジョブの増加を抑えることはできていない。

1	会話型ジョブとバッチ型ジョブの混在	■利用者は、ログインしたＰＥ上から会話型ジョブ又はバッチ型ジョブを実行可能
2	ジョブのＰＥ割付け最適化環境 (1) 会話型／バッチ型リクエスト選択の最適化 (2) 会話型／バッチ型ＰＥ振り分けの最適化	■バッチ型ジョブの待ちリストから実行可能なジョブを選択し、運用要件にあったリクエストを実行する ［運用要件］ バッチ型ジョブの連続実行抑止 ＮＱＳキュー内同時実行数の緩和 利用者毎のＮＱＳキュー内最大同時実行数抑止 バッチ型ジョブの最大待ち時間の監視／制御　など ■会話型ジョブの要求が増えてきた場合、自動的に会話型のＰＥ群を増加（最適化）する
3	限定利用者向けの特殊対応 (1) 長時間／大規模ＰＥ並列バッチ型ジョブの運用 (2) 大容量ファイル対応の運用	■利用者の要望に合わせ、以下のようなバッチ型ジョブ運用が可能 逐次ジョブを長時間実行 ６４／１１８ＰＥジョブ実行（申請制） ■利用者の要望に合わせ、最大約６００ＧＢの大容量ファイルを利用可能（申請制）

４．４ 遠隔監視サービス

スパコンの運用時間は２４時間連続運転を基本としており、定期保守以外での運用停止は極力避けたいのであるが、障害によるシステム利用が不可能になる場合は早急に復旧を目指さなければならない。このため、休祭日、夜間での遠隔監視および緊急連絡体制を敷いている。障害対応は基本的には翌日就業時間内対応となるが、緊急の場合は時間外にもＳＥに対応してもらう事態もある。

５．ジョブ運用

従来のスーパーコンピュータＶＰＰ５００システムではバッチジョブが主体であったが本システムより会話型ジョブを導入し、コマンド入力でのプログラム実行を可能とし簡便な利用環境を実現している。表２にスーパーコンピュータシステムのジョブクラスの設定を示す。

■ バッチ型ジョブ用のジョブクラス

バッチキュー名	b001	g001	s001	b004	g004	b008	b016	b032	b064	b118
最大使用可能PE数	1			4		8	16	32	64	118
キュー毎最大同時実行多重度	15	1	1	5	1	5	4	2	1	1
ユーザ毎最大同時実行の多重度制限	3	1	—		1	—				
CPU時間制限	5H	50H		5H	50H	5H				60H
ユーザ毎同時実行の緩和制御	ON							OFF	ON	
キュー毎の同一ユーザ連続実行制御	ON									
キュー毎アクセス制限	—	申請		—	申請	—			申請	

■ 会話型ジョブ用のジョブクラス

ジョブグループ名	t001	t004	t008	t016
最大使用可能PE数	1	4	8	16
ジョブグループ毎の最大同時実行多重度数	10	4	3	2
ユーザ毎最大同時実行の多重度制限	1			
ＣＰＵ時間制限	3H			
ジョブグループ毎アクセス制限	—			

表２

５．１　会話型ジョブの運用

会話型ジョブの実行にはジョブグループと呼ばれるＰＥ数が１，４，８，１６の４グループを設定している。

５．２　バッチジョブの運用

バッチ型のジョブに対しては平日利用できるキューとして最大１，４，８，１６，３２ＰＥが利用できるキューと週末に実行できる６４ＰＥジョブ、最大５０時間という長時間利用可能な１ＰＥ，４ＰＥ用のキューおよびＧＡＵＳＳＩＡＮ用のキューを用意している。ただし、６４ＰＥジョブは１ヶ月のうち２週末に限り実行することにしている。長時間ジョブは１ヶ月という期限を設けている。更に、この時間でも研究課題により、計算時間の不足が有るようであれば、運用に支障がきたさない範囲で再度１ヶ月延長も可能である。これらのジョブ利用に関しては、ベクトル化率、並列度その他の条件による許可制で運用している。

６．稼動状況

４月から７月初旬までの運用状況を示すものとして、図２に CPU 稼働率、図３に各ジョブグループの会話型ジョブの実行数、実行時間の内訳、図４に各キューのバッチジョブの実行数を示す。

6．1　ＣＰＵ稼働率

グラフから分かるように CPU 稼働率は平均８５％で変動は少ない。９０％を超えている日もかなり存在する。全体的にみても CPU 稼働率の変動は少なく密度の高い利用であることが分かる。

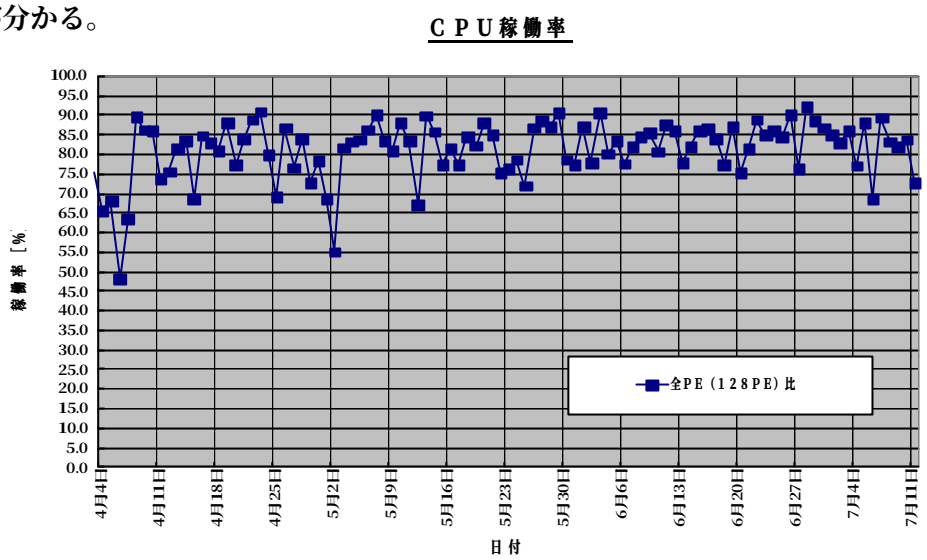

図２

CPU 稼働率(%)＝会話ジョブとバッチジョブ CPU 時間の総計／（稼動時間×128PE）

6．2　会話型ジョブ

会話型ジョブの傾向としては並列度が上がるとジョブ数が少なくなっている。これは同時に確保できる PE 数により確保できる確率が少なくなることも有るが、ジョブグループ毎の平均実行時間をみると、１，４ＰＥのジョブグループでは最大実行可能時間まで利用していることは少なく、バッチジョブのためのデバッグに使用しているものと思われる。８，１６ＰＥのジョブグループでは少し実行時間の長いジョブが有るが、８０％は 10 分以内でジョブが終了している。（図３-２参照）情報環境室では会話型ジョブ利用を促進したいと考えるところである。

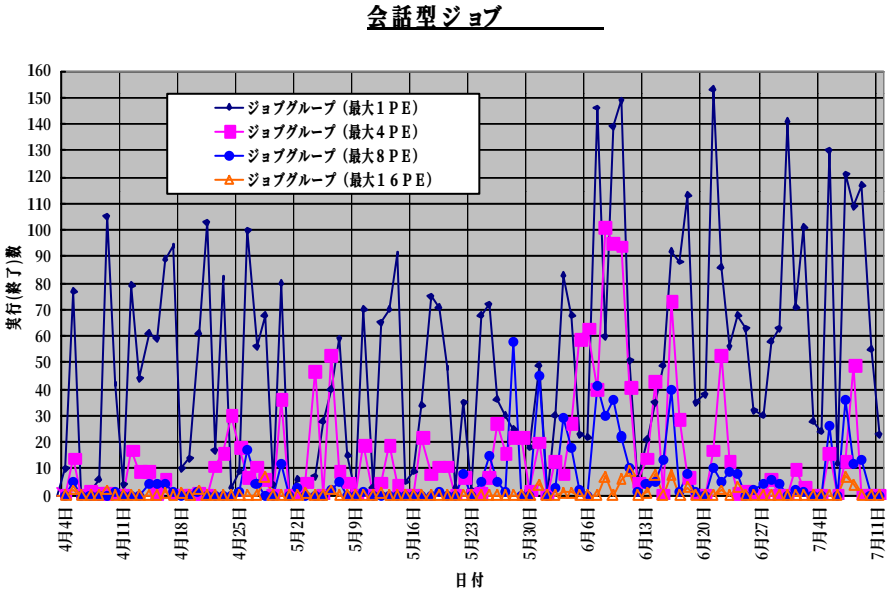

図３－１

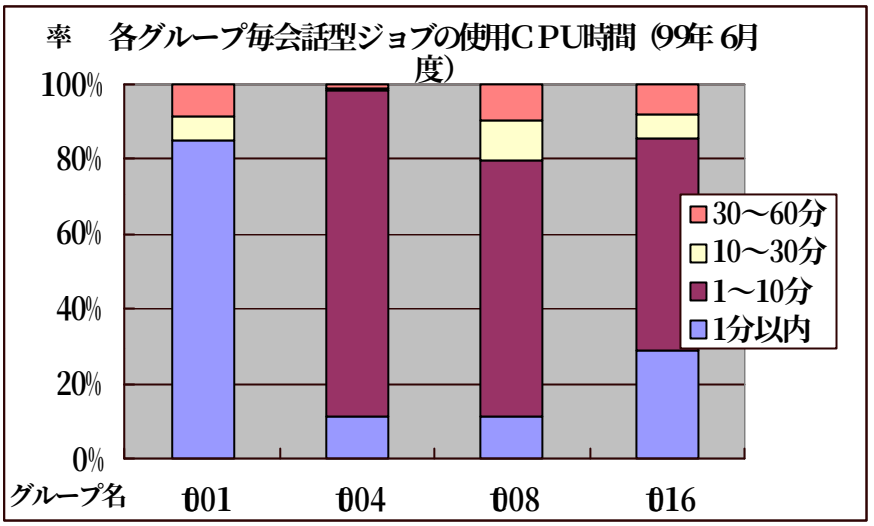

図３－２

６．３　バッチ型ジョブ

バッチ型ジョブにおいては、各キューとも処理件数は違ってはいるが、一様に処理されている。週末における特別利用の状況もグラフ上に明確に現れている。

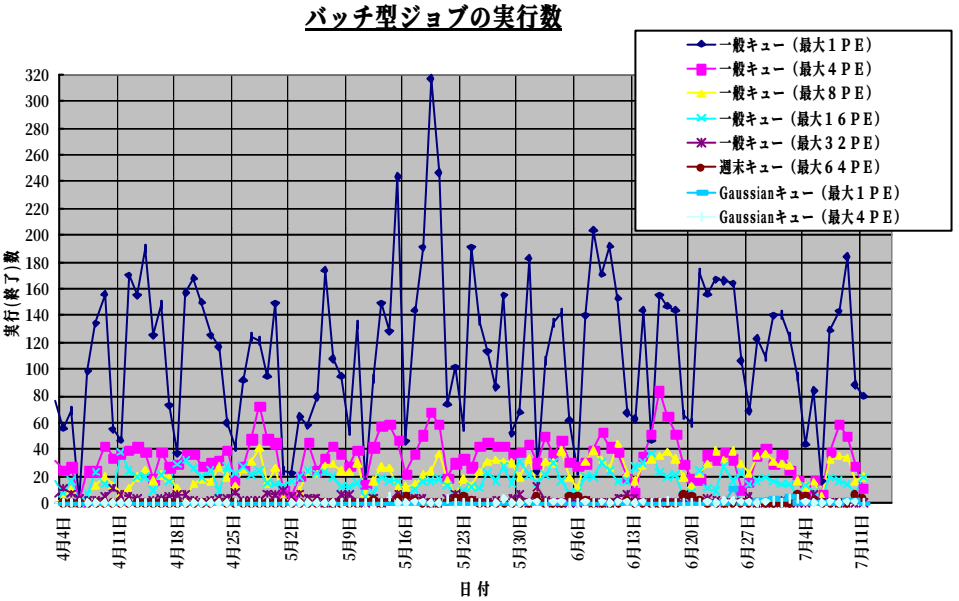

図４

７．おわりに

ＶＰＰ700Ｅ本体の稼働率をみると、導入後６ヶ月にしてフル稼動状態にある。しかも緊急保守を含め、障害等で利用者が使用できない時間が有り、この間もバッチジョブは継続して実行されているにもかかわらずジョブ実行待ちが解消されない状態である。CPU 稼働率は、各種ジョブ実行のための方策、SE のジョブ監視努力により高率を示している。しかし、ジョブ待ち時間、待ち数の解消には至っていない。このため、現在全ＰＥを利用するジョブの運用は見合わせているところである。来年には現在の１２８ＰＥから１６０ＰＥに増設したシステムで運用することになっており、この状態から脱することを期待する

最後に、現在抱えている問題をあげる。高速データ通信を行うために各システム間を接続している HIPPI スイッチ関係における障害が連続して発生し、利用者に対する快適な利用環境の提供が阻害される状況にあり、苦慮するところである。関係者の適切な対応で早期に解決して頂きたい点である。