

大型計算機システムOS IVにおける性能管理ツールの統合

住田 宏己¹⁾ 村瀬 真一郎¹⁾ 野口 守²⁾ 金澤 正憲³⁾ 三科 淳⁴⁾ 飯田 記子⁵⁾

¹⁾ 富士通(株) ²⁾ 理化学研究所 ³⁾ 京都大学

⁴⁾ 高エネルギー物理学研究所 ⁵⁾ 学術情報センター

大型計算機システムを運用するためには、様々な性能管理ツールを使い、膨大な性能データを収集し分析しなければならない。これらの性能管理作業を軽減しつつ的確で効率的な管理を行うために、性能関連作業の全体を包括した作業モデルを提案する。この作業モデルに則って、システム監視、チューニング、キャパシティ・プランニング等の作業を支援する各種の性能管理ツールを統合する。特に、ホストマシンのオーバヘッド削減とツールの操作性向上を狙い、ホストマシンでは性能測定とデータ蓄積、パソコン側では性能分析を行うという、ホスト・パソコン連携による性能管理ツール体系を実現する。

Integration of performance management tools in large-scale computer system OS IV

Hiroki Sumida¹⁾ Shinichirou Murase¹⁾ Mamoru Noguchi²⁾

Masanori Kanazawa³⁾ Atsushi Mishina⁴⁾ Noriko Iida⁵⁾

¹⁾ FUJITSU, Ltd. 140 Miyamoto, Numazu-shi Shizuoka, 410-03 Japan

²⁾ The Institute of Physical and Chemical Research.

³⁾ Kyoto University.

⁴⁾ National Laboratory for High Energy Physics.

⁵⁾ National Center for Science Information System.

A performance control system of a large-scale computer system should be designed to reduce many works of a system manager, such as skillfull manipulation of various tools and analyzing huge data. Here, we propose a model of works related to performance analysis. Efficiency of the works may be improved by systemizing various tools in accordance with the model of works. Further, in order to reduce the overhead of a host and improve the human-computer interface of tools, we realize a new performance control system in cooperation with a host and a personal computer, in such a way that the host takes measurement and data accumulation and the personal computer takes real-time monitoring and analysis.

1. はじめに

大型汎用計算機システムの運用のために、様々な性能管理ツールが提供されている。システム稼働監視、チューニング、キャパシティ・プランニングといった性能関連作業を行う際には、多くのツールを巧みに操り、膨大な測定データを収集し分析しなければならない。

システム管理者に委ねられたこれらの作業を軽減しつつ確で効率的な管理を行えるようにすることが、これからの大汎用計算機システムに求められている課題である。そのためには多様な性能管理ツールの統合が必要であると考えた。

そこで、様々な性能関連作業を一連の手順にまとめた作業モデルを考案した。この作業モデルに則って、各種の性能管理ツールを統合することで、性能関連作業の効率化が図れる。

従来から、各種のツールを体系的に整備するための地道な努力が続けられている。今回提案する作業モデルの特徴は、測定する性能データ自体に注目し、各種の分析ツールでデータを再利用することを前提にデータ中心のアプローチを探っている点にある。

一方、作業モデルに則った実際のツール体系においてはホスト・パソコン連携による性能分析作業を実現する。ホスト・マシンでの処理は性能測定とデータ蓄積に限定し、モニタリングや性能分析のための集計作業はすべてパソコン（またはワープステーション）側で行うものである。これにより、ホスト・マシンのオーバヘッド削減、分析ツールのH C I (Human Computer Interface)の向上が図れる。

我々のツール体系のもう一つの狙いは、様々な性能管理ツールを開発し易くすることであり、そのためのA P I (Application Interface)の整備を進めている。

本稿では、我々の提案する性能関連作業モデルの詳細と、そのモデルに則ったツール体系について述べる。

本研究は大学や研究機関を中心とし

たPACOMユーザで構成される「Scientific System 研究会」の中で、筆者らを含む「性能測定WG」の活動として行っている〔1〕、〔2〕。上記の作業モデルにもとづき、ツールの評価を主目的としてリアルタイム・モニタを試作した〔3〕。作業モデルやリアルタイム・モニタの試作版については、「Scientific System 研究会」の会員による評価も行っており合わせて報告する。

2. 性能関連作業モデル

大型汎用計算機システムの運用における性能作業のモデルを図1に示す。性能関連作業は、短期作業と長期作業に大別される。日常的で、短い周期の作業を短期作業と呼ぶ。性能データの測定や蓄積、リアルタイム・モニタリング、システム運転状況の監視、トラブルの原因究明と対処、チューニング等が短期作業ととらえられる。一方、年に1度、半年に1度、1年に1度程度の頻度で行う作業を長期作業と呼ぶ。月間のシステム運用状況の把握、需要変動の把握、長期的な計算機設備の計画等、いわゆるキャパシティ・プランニングに関する作業が長期作業ととらえられる。

作業モデルが示すように、測定データは様々な分析作業で再利用される。したがってデータをどのように利用するか整理した上で測定方式を実現すべきである。また性能作業の連続性を考慮した上で各種の分析ツールを構築すべきである。我々は、この作業モデルにもとづいてどのような性能

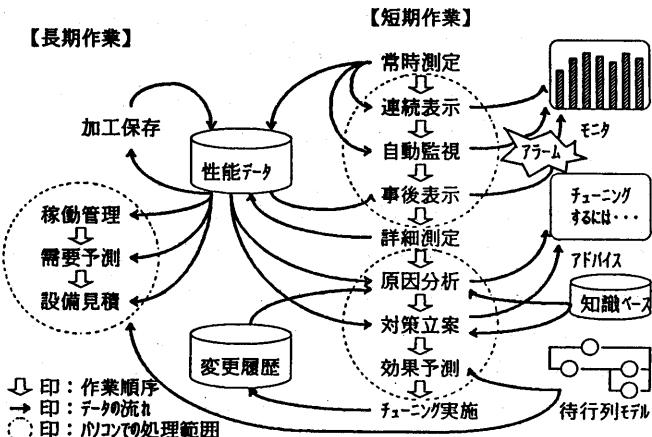


図1. 性能関連作業モデル

データがどの段階で必要になるか整理した上で、測定ツール、分析ツールを再構築している。

以下に、性能関連作業モデルの個々の作業内容について述べる。これらの作業内容のうち、詳細な検討とツールの試作まで行ったのは、常時測定、詳細測定、連続表示、事後表示までである。その他の項目、特に自動監視や、長期的な需要予測等については、今後の検討課題である。

2.1 常時測定

性能関連作業の流れから見て、性能データは、システム資源の使用状況やサービス・レベルを監視するためのデータと、トラブルが発生した場合に原因を分析し対策を立案するためのデータと、需要予測の基礎値とするためのデータの3種類に分けられる。このうち監視や予測のためにデータを常に測定しておくことを常時測定と呼ぶ。常時測定の項目はオーバヘッドとの兼ね合いから必要最小限のものに限定する。

常時測定のデータを元に、後述の連続表示、自動監視、事後表示、加工保存を行う。

2.2 詳細測定

トラブルの原因分析や対策立案のために、一時的に測定することを詳細測定と呼ぶ。詳細な測定が必要であるためオーバヘッドが大きくなる。ただし、トラブル検出時には常時測定データを用いてトラブル発生箇所を絞り込むことができる。したがって測定範囲を限定することが可能であり、詳細測定のオーバヘッドをある程度抑えられる。

1日のうちで時々、応答時間が悪化するようなケースではトラブル検出後に測定したのでは間に合わない。その場合でもトラブルが再現するまでの一定期間だけ、項目を絞って詳細なデータを測定することで対応できる。

詳細測定のデータを元に、後述の原因分析、対策立案、効果予測を行う。

2.3 連続表示

システム導入時やチューニング実施後に、予定どおりの処理能力が発揮されているか否か確認しながら運転するためのものである。いわゆるリアルタイム・モニタリングを行うものであり、数秒程度の間隔で測定データを画面に表示する。

2.4 自動監視

定常運転の際に、いつも画面を眺めていなくても済むように、トラブル発生を自動的に検出するものである。性能項目ごとに評価基準値と比較したり、データの時系列変化やデータの相対値を評価することで可能になると考える。

2.5 事後表示

トラブルの状態を確認したり原因を切り分けるために、任意の時点で、最新の稼働状況や過去の数時間の稼働状況を表示するものである。最新の稼働状況の表示では、トラブル発生時だけ必要となる情報も扱う。例えば、応答時間が悪化した時点で、特定の資源を占有しているタスク名や、その資源を待っているタスク名等を確認すれば、トラブルの原因を絞り込める場合がある。

また、過去の稼働状況の表示では、ある程度の時間範囲で把握したい情報も扱う。例えば、単位時間当たりのTSSコマンド処理件数は、数秒単位の連続表示には馴染まないが、TSS応答時間が悪化した時、過去の数時間について負荷の増減の影響を確認するのには有効である。あるいはバッチジョブの終了時刻が従来より遅くなっている場合、ある程度の時間範囲でとらえると、どのジョブが競合しているか調査し易くなる。

このように事後表示によって、トラブルの一次切り分けが可能になる。どの装置がボトルネックになっているか、およその見当が付いたならば、該当装置に絞って詳細測定を行う。

2.6 原因分析

原因分析とは、トラブル発生時に、常時測定データと詳細測定データを元に原因を分析する作業である。知識ベース・システムとして、熟練者のノウハウや過去の事例を蓄積利用することで、システム管理者の作業を支援できる〔4〕。

2.7 対策立案と効果予測

対策立案とは、トラブル発生時にトラブルの原因を除去するための対策を立案する作業である。対策の候補が幾つおりか存在する場合は、実施効果、危険度、容易度、影響範囲、費用等の要件を考慮して選択することになる。対策の立案と選択についても、原因分析作業と同様に知識ベース・

システムにより支援できる。特に、実施効果に関しては、過去の経験だけでなく、待ち行列モデルや、見積り式等によって定量的な予測も可能である。

2.8 チューニング実施

性能上のトラブルに対する作業の最終段階はシステム・チューニングである。システムが自動的に行うのが理想であるが、システム管理者が判断し人手によって行なうことが実際的である。各種パラメタの動的変更、データセットの再配置、データベースの再構成等、様々なチューニング作業を支援するツール群が必要になる。

2.9 加工保存

常時測定データを加工し、データ量を削減した上で保存するものである。キャパシティ・プランニングのためには、1~2年程度の長期間の性能データが必要である。常時測定データにはキャパシティ・プランニングに必要なデータが収集されているが、そのまま保存すると龐大な量になるのでデータ量を削減する必要がある。

例えば、常時測定データとして、ジョブごとのCPU使用時間や、データセットごとのI/O発行回数がある。これらに対して、業務ごとの合計や1日分の合計などによりデータ量を削減する。

1時間分を合計するか、1日分を合計するか等、加工する範囲は一意には決められない。データを再利用する際の便利さと保存容量との兼ね合いで加工する範囲を設定する。

2.10 稼働管理

加工保存されたデータを元に、週報・月報等のシステム稼働情報レポートを作成するなどして稼働状況を把握する作業である。週や月のデータを集計することで、負荷の変動傾向を把握し、トラブルを未然に防止することができる。

2.11 需要予測と設備見積り

半期や1年ごとの作業として、システム資源が飽和する時期を予測し必要な設備計画を立案する作業である。既存業務については、過去の稼働実績を元に需要の伸びを予測する手法を用いることが多い。そのための基礎データとして、加工保存されたデータが役立つ。

3. 性能関連作業モデルの有効性

我々は製品に反映することを最終目標として検討を行っている。そこでアンケート調査によって「Scientific System 研究会」の会員に、利用者の立場から我々の提案する作業モデルを評価していただいた。

3.1 性能測定に対する条件

アンケート調査による評価結果を表1に示す。

表1. アンケート結果

回答内容	会員数
基本的に賛成	13会員
条件付き賛成	9会員
回答なし	37会員
合計	59会員

回答率は36%であり、決して高くはない。しかし、実際の利用者から見た評価として賛同するという意見をいただいている。我々の提案する作業モデルの有効性を確認できた。その中で、具体的な条件を示した上で賛成するという貴重なコメントがある。共通点は、「性能測定にはオーバヘッドがつきものであり無条件に多くのデータを測定すべきではない」というものである。示された条件の詳細を以下に示す。

① オーバヘッド評価の必要性

- 測定項目ごとのオーバヘッドを明記する。
- 測定による性能低下分を評価する仕組を組込む。その結果から常時測定項目をさらに限定する。

② 測定項目の選択

- システム負荷とデータ量が問題となる。情報が多いほど良いというのは安易である。すべての測定項目に抑止機能を付加する。

③ 常時測定機能の必要性

- 課金情報は常時収集しているが、性能測定は週1回3時間だけ行っている。問題が発生した時点で詳しくデータを収集すればよい。
- 小規模システムでは常時測定を行うとオーバヘッドが問題になる。TSS応答時間が悪化した時のように、トラブルが生じた時だけ測定すればよい。

④ 性能データベースの公開

- 性能データベースは単純な構造としデータ形式を公開する。

⑤ 詳細測定の自動起動

- トラブルの再現待ちを避けるため自動監視の結果を元に自動的に詳細測定を起動する。

⑥ 事後表示方式の工夫

- 幾つかの性能項目を組合せて表示する。

⑦ コンソールとの一体化

- 性能分析用のパソコンとシステムコンソールは共通化する。

上記のいずれについても、我々の作業モデルで対応できる。③は常時測定に否定的な考え方であるが、この作業モデルは、常時測定を行わないという運用形態を不可とするものではない。常時測定を抑止し、トラブル発生後に詳細測定を行い、トラブルの再現を待った後、原因を分析するという運用も可能である。一方で別の会員からは、問題が発生した時、よく事象再現待ちになってしまふと指摘されている。道に迷った時だけ地図を広げても正しい道には辿り着けないというのが常時測定を設けた理由である。いずれにしろ、どの項目を測定するかは、オーバヘッドと原因分析の容易さとの兼ね合いであり、システム管理者の判断によって決定されるものである。

このように、我々の提案する作業モデルは、あらゆる運用形態に十分対応できる体系であると考えられる。

3.2 オーバヘッド

性能管理ツールを利用する上での最大の関心事はオーバヘッドである。そこで、既存の性能測定ツールのオーバヘッドを「Scientific System 研究会」の会員ごとに見積った。

測定オーバヘッドは、システムの運用状況、測定対象資源、測定の詳細度（精度）によって異なる。そこで「Scientific System 研究会」会員の代表的な運用パターンから3種類のシステム規模のモデルを作成し、それらに対して測定プログラムの実行命令数を数えることにより見積った。

その結果を表2に示す。大規模システムの場合CPU処理能力が高いので相対的なオーバヘッド

は小さくなっている。小規模システムの場合で詳細なデータを採取した場合のオーバヘッドは、無視できない大きさである。この見積り値は、見積もりモデルにほぼ対応する実システムの実測値と良い一致を示した。

少なくとも測定項目を絞り込むことによって、システムのオーバヘッドを運用に差支えない程度に抑えられると見えよう。

表2. 測定オーバヘッドの見積り

シス テ �ム 規 �模	オーバヘッド	
	ケース1	ケース2
大規模：M780/20 使用会員	0.2 %	0.3 %
中規模：M780/10S使用会員	0.5 %	0.6 %
小規模：M760/6 使用会員	1.7 %	2.3 %

※ケース1：CPU, チャネル, デバイスなどを測定

ケース2：ケース1+メモリ, ページング, 磁気ディスク・ヘッド移動状況, SVC(スーパーバイザ・コール), ENQ(排他)資源, SDM(System Decision Manager)管理情報 を測定

4. 性能管理ツールの一部試作

2章に示した性能関連作業モデルの中の連続表示を支援するツールとして、リアルタイム・モニタを試作した。図2に示すように、ホスト・マシンでは性能データの測定と蓄積を行い、表示、監視、分析等の作業はパソコン側で行うというホスト・パソコン連携により実現した。この方式には、従来のホスト・マシン上のツール体系に比べて以下の長所がある。

① 表示や解析のためのホスト・マシンの負荷を軽減できる。

② 扱い易いHCIを実現できる。

③ 利用者独自の解析ツールを作り易い。

このリアルタイム・モニタは「性能測定WG」のメンバの属する計算機センターに導入し、試験運用を行っている。試作段階であり、以下の項目に限って表示している。

- CPU使用率
- チャネル使用率
- 磁気ディスク装置ビギー率
(DASD:Direct Access Device)
- 秒当たりのページング・ページ数
- 仮想記憶領域の使用率 (SQA:System Queue Area, CSA:Common Service Area)

表示画面の一部を図3に示す。これらの項目の

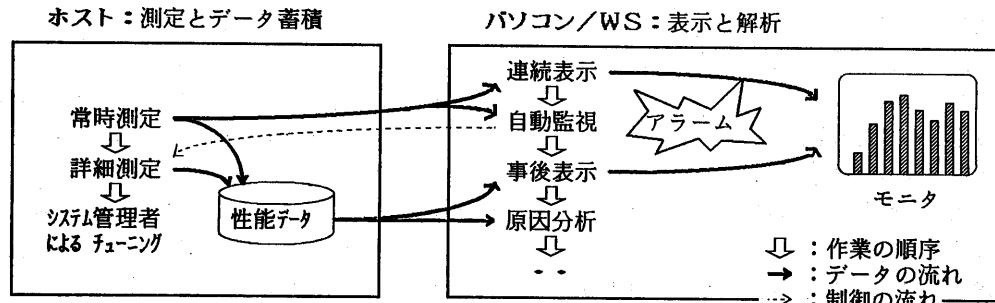


図2. ホスト・パソコン連携方式

うち、チャネル、磁気ディスク装置は多数存在するので、どれを表示するかを選択できるようにした。先頭画面には各種の性能項目をまとめて表示している。画面をめくれば、すべてのチャネルや磁気ディスク装置を一覧表示することもできる。評価基準と比較して使用率の高い項目については黄色や赤色などの注意を引く色に変えて表示する。

ホスト・パソコン間のデータ転送には、TSS配下でパソコン(FMR)と通信するためのプロトコルであるDUE T会話方式を利用した。試験運用では、数百台の磁気ディスク装置の性能データも含め、1回につき4KB程度のデータを3秒間隔で送信し、パソコン側でリアルタイムにグラフ表示させた。また、公衆電話回線を利用して遠隔地のホスト・マシンのデータをリアルタイム表示するデモンストレーションも行ったが、1200 bpsの通信速度で5秒間隔の表示が可能であった。普

普通での通信手順については、転送効率や接続性等の考慮が必要であり、今後の検討課題である。

なお、この試作版は「Scientific System 研究会」の会員に広め評価を重ねる予定である。

5. 性能関連作業モデルによる統合ツールの具体化

2章で述べた性能関連作業モデルに則って各作業を支援する、統合ツール体系について述べる。全体像を図4に示す。

5.1 API (Application Interface)

ハードウェア資源使用状況、ジョブ別資源使用状況、各種サブシステム資源の使用状況等のホスト・マシン内部の性能データは、すべて新PDL(仮称)と呼ぶ性能測定ツールにより測定される。測定データのうち、リアルタイムな分析に必要な項目については、リアルタイムAPIと呼ぶアプリケーション・インターフェースにより実メモリ経由で受け渡される。このインターフェースは基本的にOSIVシステムで共通であり、この上でリアルタイム・モニタや自動監視ツールなど様々なツールを開発できる。

一方、新PDLにより測定されるデータは、常時測定、詳細測定のいずれについても、すべて性能データベースに蓄積される。トラブル検出のように必要となった時点で、この性能データベースの中から性能データを引用し、原因分析を開始することができる。この性能データベースは、長期間のデータ保存の機能も備え、保存量を削減するためにデータを加工処理する。性能データベースは蓄積型APIによりアクセスできる。リアルタイムAPIと同様に、このインターフェースを利用してチューニング支援やキャパシティ・プラン

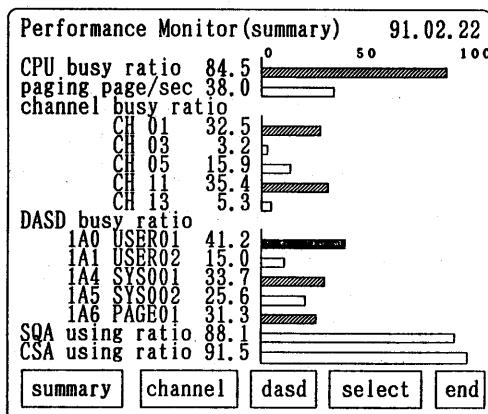


図3. リアルタイム・モニタ試作版の画面例

ニングのためのツールを開発できる。

これらのAPIは様々な性能管理ツールを効率良く開発できるように整備して公開される。

5.2 HCI (Human Computer Interface)

各種の分析ツールをパソコンで実現することにより、操作性の良いHCIを提供できる。また、パソコンで実現していることにより、コンソール機能を追加して総合的な運用管理端末に発展させることができる。以下では本ツール体系の中で特にリアルタイム・モニタに絞り、図5に示す画面例を用いてHCIの概要を述べる。

リアルタイム・モニタは、オペレータによるトラブル対応作業を支援するものである。エンド・ユーザへのサービス・レベルを確認したり、実際にサービス・レベルが悪化した際には即座にトラブルの原因を究明することを狙っている。図5(a)で資源バランスを把握することにより、性能トラブルを早期に検出できる。また、図5(d)で業務別

応答時間を把握することにより、応答時間の悪化を直ちに検出できる。図5(a)または(d)の情報から、応答時間の悪化を引起している部分が、回線の伝送時間の遅延か、ホスト内でのCPU資源やページングの待ちか、磁気ディスク装置へのI/O処理の遅延かあるいは特定タスク処理の待ちが発生しているのか切り分けることができる。

例えば、応答時間が悪化した場合、通常に比べてCPU待ち時間が長くなっているのであれば、図5(b)により、その日の稼働状況を確認する。応答時間が何時頃から悪化しているのか、どのような傾向か、CPU使用率と応答時間の相関関係等を確認する。CPU使用率が高くなったのが影響してCPU待ち時間が伸びているのであれば、他の業務の影響が推測されるので、実際にどの業務がCPUを多用しているのか確認する。図5(c)の例では業務Bの立ち上がりに伴ってCPU使用率が高くなり、業務Aの応答時間が悪化していると

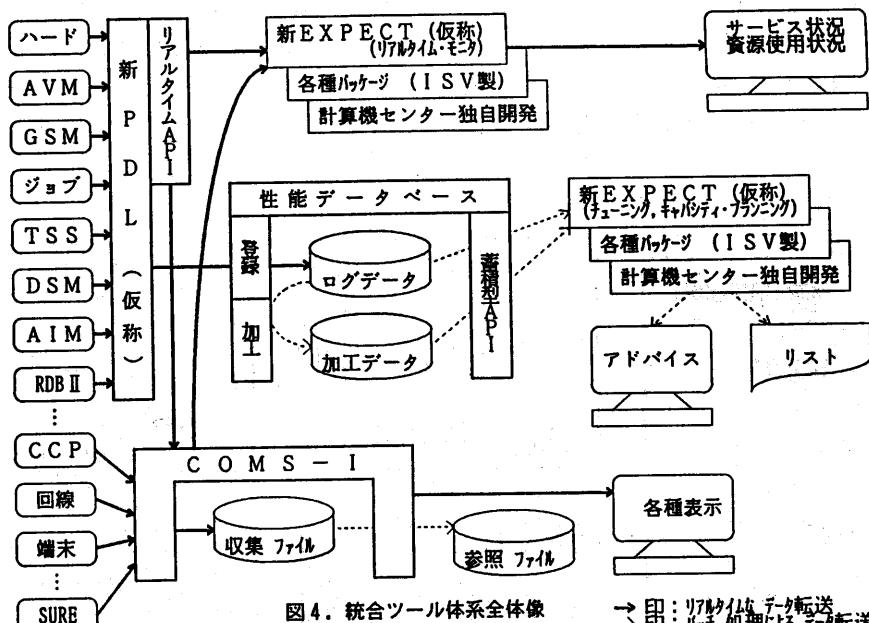


図4. 統合ツール体系全体像

→印: リアルタイムデータ転送
→印: バッチ処理によるデータ転送

AVM : Advanced Virtual Machine
GSM : Global Storage Management System
TSS : Time Sharing System
DSM : Distributed System Manager
AIM : Advanced Information Manager
RDB II : Relational Database II
CCP : Communication Control Processor
SURE : 次世代通信処理プロセッサ

COMS-I : Corporate Network Management System of Information processing
EXPECT : Expert System of Performance Consulting and Tuning
PDL : Performance Data Logger
API : Application Interface
ISV : Independent Software Vender

判断できる。このようにして性能上のトラブルの原因を究明できる。どちらの業務を優先するかは業務の運用責任者の判断を仰げばよい。

一方、応答時間の内訳で I/O 处理が遅延している場合には、図 5(e)により該当業務が使用しているボリュームの負荷状況を確認する。その中で該当業務の応答時間に大きな影響を及ぼしているボリュームを見つける。さらに、図 5(f)により該当ボリューム内のデータセット毎の負荷を調査し、どの業務が該当ボリュームの負荷を上げているのかを確認する。図 5(f)から業務 A の中の APL1 と APL2 の 2 つのジョブが同一ボリュームに大量の I/O 处理を行ったため I/O 处理が遅延したのだと判断できる。

6. おわりに

本稿では、性能関連作業モデルを提案し、そのモデルにしたがった統合ツール体系について述べた。現在、リアルタイム・モニタについては試作版の試行評価を終え、普及版の開発を行っているところである。その他のツールについても、順次、開発する予定である。

一方、本稿で述べたリアルタイム API と、蓄積型 API についても、具体的な仕様の設計を進めているところである。これらの API により、

従来よりも格段に使いやすいツール群を効率良く開発できると考えている。

本稿で述べた体系は、OS IV システムで共通なものとした。今後は、他の OS への展開と、分散処理システムでの性能管理体系との融合を行うことが課題である。

本研究に参画していただいた「性能測定 WG」の他のメンバー、貴重な御意見を寄せていただいた「Scientific System 研究会」会員の方々並びに事務局の方々に感謝します。

参考文献

- [1] 住田 他, "性能測定ツールについて", Scientific System 研究会 Newsletter №52, 富士通㈱, 1991年。
- [2] 田島 他, "システム性能評価ツールについて", Scientific System 研究会 Newsletter №48, 富士通㈱, 1989年。
- [3] 村井 他, "ホスト・パソコン連携による性能測定システム及びプロトタイプ版の概要", 情報処理学会第39回全国大会, 1989年。
- [4] 住田 他, "性能監視エキスパートシステム : EXPECT", 情報処理学会オペレーティング・システム研究会, 第38回, 1988年。

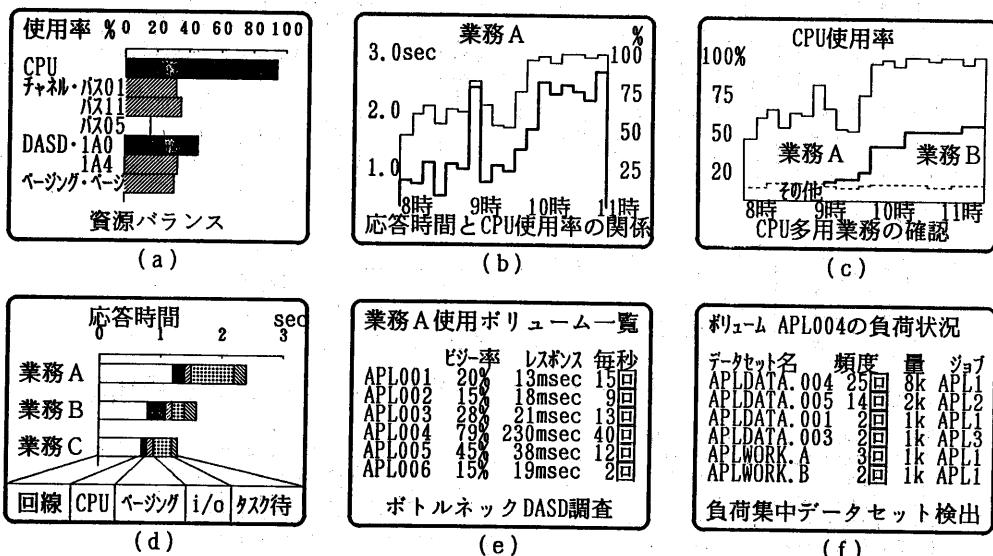


図 5. リアルタイム・モニタ概要