

## スカラーパラレルコンピュータ AP3000の運用管理機能について

山根 恒美  
富士通  
HPC本部第二開発統括部

AP3000は複数の汎用ワークステーションをAP-Netと呼ぶ高速の通信ネットワークで接続した分散メモリ型のスカラー並列機である。高速なノード間通信性能により、高性能並列コンピュータとして高い実効性能を得ることができるとともに、ワークステーションクラスタとして利用することも可能である。各ノードではSolaris オペレーティングシステムが動作し、各種運用管理ソフトウェアとハードウェアのシステム制御機構により、AP-Netに接続されたノードコンピュータ全体を1システムイメージで運用することができる。また、ノードコンピュータを用途に応じてグループ分割し、様々な利用形態に柔軟に対応することも可能である。

## Scalar Parallel Computer AP3000 System Operation and Management Functions

Tsunemi Yamane  
Development Div. II  
High Performance Computing Group  
Fujitsu

The AP3000 is a distributed-memory scalar parallel server consisting of multiple general workstations connected via a high speed communication network which is called as an AP-Net. By its supreme performance for inter-node communication, the AP3000 can be used as a high performance parallel computer, or it also can be used as a workstation cluster. The Solaris operating system runs on each node, and by using several kinds of operation management software with the system controll feature of the AP3000, all node computers connected to the AP-Net can be managed by one system image operation. Furthermore, node computers can also be splitted into seviral node groups as required to meet to various flexible utilization.

## 1. AP3000概要

AP3000は64ビットマイクロプロセッサ UltraSPARCを採用した汎用のUNIXワークステーションをノードコンピュータとして高速通信ネットワークAP-Netで接続し、最小4ノードから最大1024ノードまでの拡張性を有する分散メモリ型のスカラー並列コンピュータである。個々のノードではSolarisオペレーティングシステムが動作し、既存のSolarisアプリケーションをそのまま実行することができる。各ノードを単体のUNIXサーバとして使用することも可能であり、また複数のノードを用いた並列実行により超高速処理を実行することも可能である。AP3000の適用分野として、現時点では以下を想定している。

- ・R&D分野向け共同利用サーバ

多数のハイエンド・プロセッサによるスループットサーバとしての利用形態であり、大学における情報処理センターや、企業内の部門R&Dサーバとしての利用

- ・アプリケーション専用サーバ

並列処理による高いスカラー性能と広いI/Oバンド幅および高速データ転送機能を利用したデータハンドリング能力をいかした実験・観測データ処理システム、CAD/CAE等のシミュレーション専用サーバ、大規模情報検索システム、およびデータマイニング処理等への利用

- ・並列処理研究用

高速のノード間通信機能により効率のよい並列処理が可能であり、また汎用ワークステーション上の各種開発・評価ツールなども利用可能である。

## 2. AP3000システム構成

AP3000のハードウェア構成を図-1に示す。サーバタイプの汎用UNIXワークステーションがノードとして、独自開発のAP-Netに接続される。AP-Netは二次元のトーラス状のトポロジーを持つ通信ネットワークであり、各通信路は同時双方向に200MB/Sのデータ転送が可能である。システム制御機構は電源制御や統合コンソールを実現するための機構であり、各ノードとはシリアルインタフェースで接続され、またAP3000の外部にある制御ワークステーションと直結される。制御ワークステーションはAP3000システムの運用に用いるものであり、各ノードとは制御ネットワークを介して接続される。

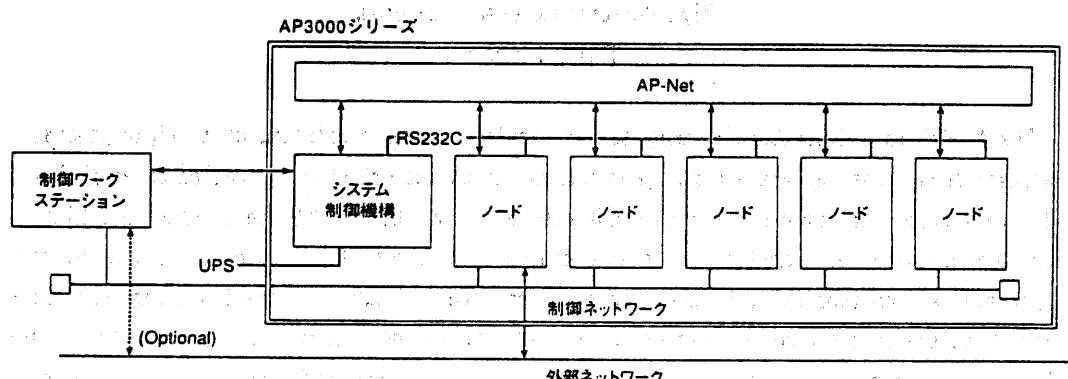


図-1 AP3000のシステム構成

### 3. ソフトウェアの構成

図-2はAP3000のソフトウェア構成を示している。ノードのOSとしてアプリケーションの豊富な汎用OSであるSolarisを採用し、オープン性の維持、すなわちSolarisアプリケーションの動作保証のため、また新バージョンへの追隨時間を最短にするため、OS層には可能な限り変更を加えないように設計した。具体的にはSolaris2.5.1にAP-Netドライバを組み込み、これ以外はすべてOSより上位のレイヤーで、運用管理を含む種々の機能を実現している。

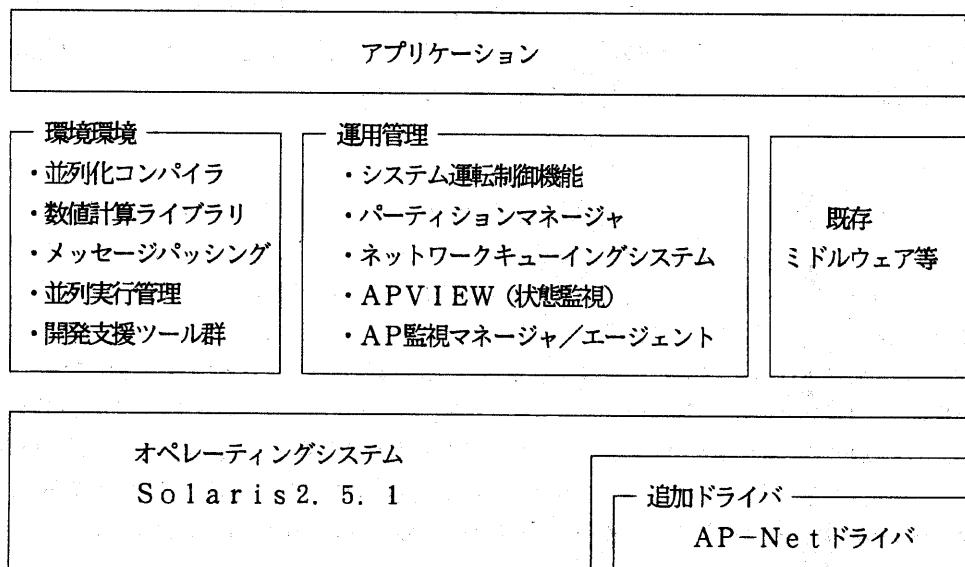


図-2 AP3000のソフトウェア構成

### 4. AP3000の運用管理

AP-Netは並列プログラム用のユーザレベル通信とともに、ワークステーション環境で使用されるIP処理のためのシステムレベル通信もサポートしており、TCP/IPやNFSなどのワークステーション環境に必須の通信の高速化が図られている。このためAP-Netを高速LANとして、AP3000システムをワークステーションクラスタの形態で利用することが可能である。

しかし、通常のワークステーションによる分散環境の欠点として、多数のワークステーションの管理が困難であることが挙げられる。地理的に分散していたり、管理ノウハウが分散または不在であるためにシステムとしての統制が取れないという問題・悩みは常に指摘されている。このためにUNIX環境用の運用管理のためのいわゆるミドルウェアが多くベンダーから提供されている。AP3000ではSolaris用の運用管理ミドルウェアは目的に応じてそのまま利用できるようにしている。

一方、AP3000では分散から集約へ、すなわち複数のワークステーションを筐体に一括収納し、利用イメージは分散環境のまま、システム運用の一元管理を実現し、運用の簡易化と管理コストの低減を可能とする専用ソフトウェアも開発した。また、各ノードを利用目的に応じてパーティションと呼ぶいくつかのグループに分割運用し、さらに各ノードの稼働状況を把握し、利用率の低いノードへのジョブの自動配分等、システムの運用性と利用効率の向上を目指している。以下にこれらの集約化を支える専用ソフトウェアの機能を説明する。

#### ・システム運転制御機能

本機能は、電源制御とコンソールの集中制御を行うシステム制御機構と呼ぶハードウェアの機構と相まって、AP3000システム全体の電源の一括投入と切断、各ノードの起動と停止、各ノードのコンソール操作と運用ログの集約を制御ワークステーションから行うためのソフトウェアである。オペレータは制御ワークステーションからAP3000を構成するすべてのノードを、1システムイメージで操作することができる。またシステム保守についても、一箇所に集約されたログ情報を解析することにより、全ノードの障害の情報を取得することができる。さらに各ノードへのソフトウェアのインストールやパッチの適用は、制御ワークステーションから制御ネットワークを介して行われる。システム制御機構はUPSや防災監視装置等の外部装置との接続も可能であり、電源や環境に異常を検出した場合の自動対処を可能としており、システムの自動運転を図ることができる。

#### ・パーティションマネージャ

AP3000では、多くのノードを効率よく運用管理するために、二階層のグループングの考え方を導入している。一つは大学における学部や学科単位にノードをグループ化するケースであり、各ノードをファイルサーバやユーザ登録などを含めた利用環境毎に分割するものであり、PMクラスタと呼ぶ。他の一つは同一の利用環境でありながら、利用形態毎にグループ化するものであり、パーティションと呼んでいる。たとえばバッチ処理用パーティション、会話処理用パーティションという分類である。一つのPMクラスタには複数のパーティションが存在する。パーティションマネージャはこの二階層に分けたノードの運用を支援するソフトウェアであり、グループ毎に環境設定や異常発生時の措置を決定することが可能である。またパーティションマネージャは会話処理における負荷分散を実現しており、低負荷ノードへのリモートログイン/リモートシェル機能、および低負荷ノードへのスイッチ(再ログイン)機能をサポートしている。

#### ・ネットワークキューリングシステム

パーティションマネージャによりバッチ処理用に分割された複数ノード上で、バッチ処理を実現するソフトウェアである。UNIXの標準バッチ処理システムに対し以下の機能追加を行い、特に並列ジョブ実行のための運用性の向上を図っている。

##### [実行ノードの選択とジョブの保全]

ジョブキュー毎に複数のノードを実行ノードとして定義しておき、ジョブ実行時にその中から必要なノードをシステムが自動的に割り当てる。ジョブ実行中にノードが故障した場合、特に指定がなければジョブは別ノードを使用して再実行が計画される。

##### [ノードの排他的利用]

ジョブキューにSIMPLEX属性を与えることにより、そのジョブを実行している間、ノードを他のジョブと共にせず、排他的に利用することができる。そのため、他ジョブの影響による実行時間の遅延やシステム資源の競合の心配がない。AP-Netのユーザレベル通信によるノード間高速通信を使用する並列プログラムを、ネットワークキューリングシステム経由でバッチジョブとして実行する場合は、そのジョブに割り当てられたノードはSIMPLEXモードで実行される。なお、一つのノード上で同時に複数のジョブを実行する場合は、ジョブキューにSHARE属性を設定する。

##### [ノード固定ジョブキュー]

特定のジョブキューに特定のノードを固定的に割りつけることができる。この機能を利用すること

により、特定のノードを特定のアプリケーションサーバとして運用することが可能である。

#### [負荷分散]

ジョブ実行に際し、ジョブキューに定義されている複数のノードの中から、現在未使用的ノードを選択、あるいは未使用的ノードがない場合には、使用中のノードの中から最もシステム負荷（CPU 使用率）の低いノードを選択するため、システム全体の使用効率の向上が実現される。

#### ・APVIEW

AP3000を構成するすべてのノードの稼働状況を制御ワークステーションに一括表示するMotifベースのGUIツールである。これにより各ノードの電源の状況、負荷状況（CPU使用率の10段階表示）、実装メモリ量、ログインユーザ数、実装位置、およびAP-Netの電源状態、ネットワークトポロジーなどを一括監視することができる。パーティションマネージャとも連携しており、パーティションの状況を表示することもできる。

#### ・AP監視マネージャ／エージェント

AP監視マネージャ／エージェントは運用管理のためのミドルウェアである「集中監視マネージャ」と「Open-Eyes」をAP3000向けに統合したシステム管理者用の統合コンソール機能である。AP3000向けに最適化しており、インストール後直ちに利用することができる。また、他のミドルウェア製品との連携も可能であり、ネットワーク管理のための「NetWalker」や業務スケジュール管理のための「ジョブスケジューラ」と組み合わせて運用管理システムを構築できる。

### 5. 今後の課題

AP3000の運用管理機能を中心に説明してきたが、今後さらなる運用性の向上のために、以下の課題に取り組む必要があると考えている。

#### [ノードの負荷状況のより正確な把握]

各ノードの負荷の平準化のために、現在はCPU使用率の低いノードを選択してジョブの割り付けを行っている。各ノードのCPU性能やCPU数に大きな差異がなければ、この方式で大きな支障はない。しかし、今後はノードのCPUがさらに高速化され、またノードあたり最大30CPUのSMP(Symmetrical Multi Processor)をAP-Netに接続可能とする計画である。このように各ノードのCPUパワー（単体性能×数）に大きな差が生じるようになると、負荷分散を図るうえで単なるCPU使用率のみならず、CPUパワーの各ノード毎の重み付けを考慮する必要がある。

#### [ネットワークアドレスの削減]

並列アプリケーションおよび分散アプリケーションの両方を動作可能とするために、現在はすべてのノードにIPアドレスを割り振っている。IP資源の節約と運用性の向上のために、このIPアドレス削減のための取り組みが必要である。

### 6. AP3000の今後の展開

AP3000の今後の方針として、AP-Netの通信性能のより一層の高速化、並列ファイルシステムによるファイルアクセスの高速化、AP3000を構成するコンポネントの二重化やテークオーバ方式による信頼性と可用性の向上、VPP(ベクトルプロセッサ)や他のUNIX機を含む異機種分散コンピューティングの実現を目指しており、運用管理機能もこれらと相まって発展させていきたい。