

第2章 計算機システム

2.1 新数値解析予報システムについて¹

2.1.1 スーパーコンピュータ

新数値解析予報システム(以下「NAPS-8」という。)は、図2.2.1のような構成となっている。数値予報ルーチンはこのNAPS-8で実行される。主な数値予報ルーチンは構成図一番上にあるスーパーコンピュータにて実行される。以下に個々の計算機の役割について概要を説明する。

スーパーコンピュータはノードという単位で構成され、1ノードには16個のCPUと64GBのメモリが搭載されている。サブシステム2及び3は、80ノードで構成され、CPUはPOWER5+の2.1GHzである。ノードはクラスタ結合²され、ノード間はMPIで通信を行う。ノード内の処理分散はSMPで処理され、計算効率を高めている。各サブシステムには4個のI/Oノードがあり、ネットワークとの通信、フロントディスクへの入出力を行っている。一方、隣接している高速磁気ディスク装置へは全てのノードから入出力が可能となっている。また、拡張機能としてメモ

リの一部をディスクとして利用している。

サブシステム3では、全球、メソを初めとする数値予報ルーチンが実行される。サブシステム2では、サブシステム3で実行される数値予報ルーチンの確認のための準ルーチン及び数値予報課が担当するモデルの開発・検証が実施される。なお、サブシステム3の保守時及び重障害時にはサブシステム2が代替えとなり数値予報ルーチンを実行する。これにより1日8回のメソ数値予報の実行が安定的に可能となる。サブシステム3の高速磁気ディスク装置は、切り替えによりサブシステム2からも直接アクセス可能となる。

2.1.2 サーバ類

(1) 降水短時間予報用サーバ

降水短時間予報ルーチンとして解析雨量作成及び降水短時間予報等が30分毎に実行される。本サーバは、UNIXサーバ3台で冗長構成³され、スーパーコンピュータと同様に稼働系サーバの保守時及び障害時には待機系サーバが降水短時間予報ルーチンを実行する。

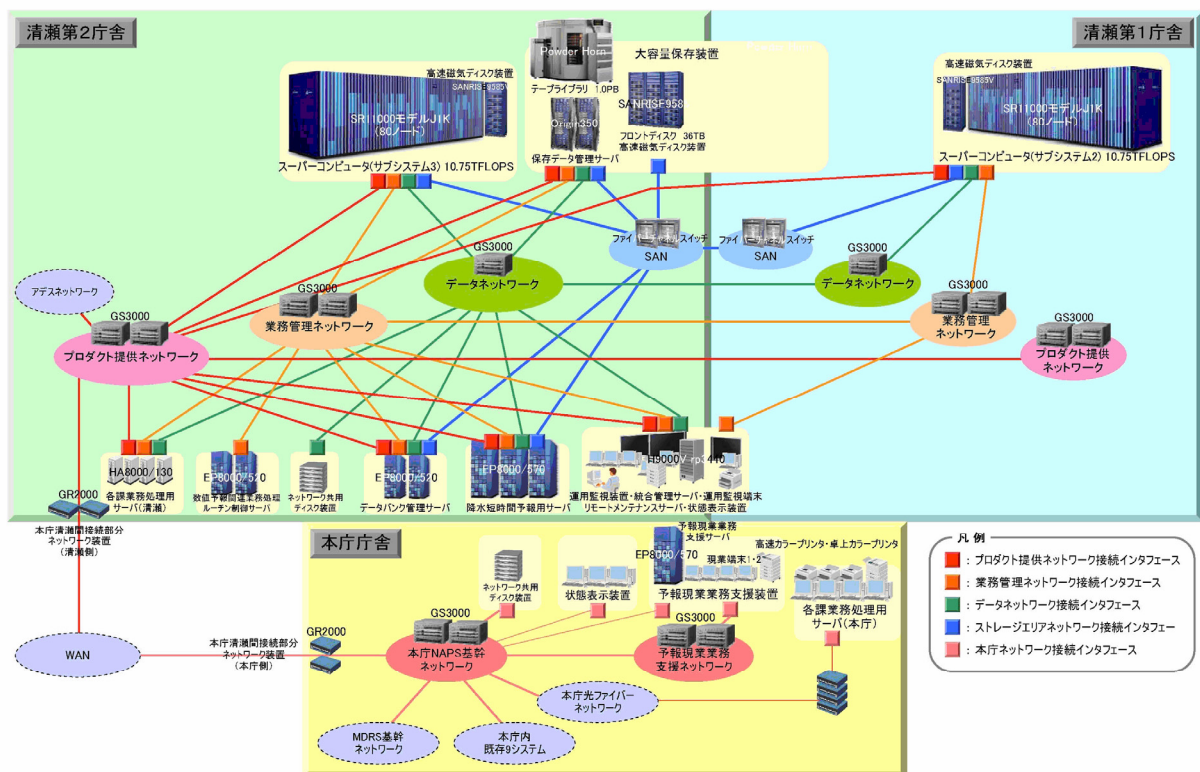


図 2.1.1 新数値解析予報システムの構成

¹ 中山 博義

² 複数のコンピュータを相互に接続し、ユーザやほかのコンピュータに対して全体で1台のコンピュータであるかのように振舞わせること

³ 一台が故障しても、残りのサーバが補填することでシステム全体の停止を防ぐ構成のこと

(2) 保存データ管理サーバ

UNIXサーバで冗長構成され、図2.2.1で同サーバに隣接した高速磁気ディスク装置、フロントディスク及び大容量保存装置のデータ管理を行っている。SANに接続された計算機間ではデータの共有が可能となっている。また、大容量保存装置もSANで共有され、ユーザからはフロントディスクにより大きな磁気ディスクのような感覚で利用可能である。

(3) データバンク管理サーバ

UNIXサーバで冗長構成され、NAPS-8で作成された共用データの管理を行っている。基本的に利用者はPandoraによりデータをアクセスする。

(4) 数値予報関連業務処理ルーチン制御サーバ

UNIXサーバで冗長構成され、数値予報ルーチン及び降水短時間予報ルーチンの制御及び監視を実施している。数値予報ルーチン及び降水短時間予報ルーチンは基本的に予め設定された時刻になると本サーバにより自動的に起動され、ジョブ間の複雑な依存関係を保ちながら実行される。またジョブの異常終了時には、そのことが本サーバにより検知され、統合管理サーバに伝える。統合管理サーバはシステム運用室の現業者に対して異常発生を鳴動等により知らせる。

(5) 各課業務処理用サーバ(清瀬)

Linuxサーバ18台で構成され、2台1組となっている。数値予報課では本サーバにて課ルーチン⁴を実行するほかに、清瀬に設置された計算機資源を使用した開発のために利用している。

(6) ネットワーク共用ディスク装置

清瀬及び本庁に設置され、それぞれ清瀬及び本庁に設置されたサブシステム及びサーバにNFSでマウントされている。本装置は、一般ユーザのホームディレクトリがあり、主に開発利用のデータを保存するために整備されている。書き込み及び読み込みはネットワークを通じて行うため速度は高速ではないが、通常近傍の全てのサーバ等からアクセス可能なため利便性が高い。

2.1.3 ネットワーク

これらの機器を接続する目的で、清瀬には3種類のネットワークが整備される。3種類のネットワークの利用目的は以下のとおりである。プロダクト提供ネットワークでは、通信システムと接続し、プロダクト、気象電文及び観測データの転送及び取得を行う。業務管理ネットワークでは、数値予報ルーチン及び降水短時間予報ルーチン制御、並びに機器監視のためのデータを扱う。データネットワークでは、主にネットワーク共用ディスク装置上の開発用データを扱う。

清瀬と本庁間は広域LANで冗長接続されている。本庁側では本庁NAPS基幹及び予報現業業務支援ネット

ワークと接続している。予報現業業務支援のため、支援装置及び端末が設置されている。また本庁光ファイバーネットワークを通じて、本庁内各課から業務及び開発のためNAPS-8の利用ができる。数値予報課からも当該課整備の課内LANを通じて各自の机上のパソコン端末からNAPS-8の各機器へアクセスが可能である。

2.1.4 数値予報ルーチンとその制御方法

数値予報ルーチンでは、気象電文及び観測データの主なものは通信システムから取得する。また、気象衛星センター(MSC)で処理している衛星データは、本システムが衛星データ処理システムと結合しているため、当該データをファイルとして取得する。その他には、インターネットに公開されているデータがあるが、こういったデータの取得は本庁の MDCDS を通じて課ルーチンとして運用されているジョブにより取得され、数値予報ルーチン実行時にデータを参照する。

取得されたデータはデコード処理され、その後全球解析され、全球予報が実行される。この結果を境界値としてより領域が狭く格子間隔の小さい解析・予報が実行される。これらの処理結果は定期的にサブシステムに接続された高速磁気ディスク装置に保存される。前述の処理が正常終了の後、プロダクト作成のための処理が適時実施され、気象電文、FAX またはファイルに加工される。

作成されたファイルは、データバンクとして公開され、同時に通信システムを初めとする他システムへ転送される。

このように NAPS-8 で実行される数値予報ルーチンは、数千にも及ぶジョブが相互に関連し、また、一連のジョブが実行される計算機が複数に渡ることもあり、制御のためのソフトウェアが不可欠である。このような運用形態を行う目的で本システム受注メーカから数値予報ルーチン業務運用支援ソフトウェア(JNOS3)が納入・提供されている。本ソフトウェアを用いて数値予報ルーチンは、数値予報課で維持・管理され、システム運用室で運用・実施されている。

⁴業務のために実行されているルーチンで、数値予報ルーチンよりも必須性が小さい。

2.2 Pandora プロジェクト¹

2.2.1 はじめに

数値予報課では、第 2.1 節で述べた数値解析予報システム (NAPS) など、近年の大規模な計算機システムを利用する上でのデータの取り扱いと、可視化等のデータの利用手段に要するコストについての問題意識から、平成 13 年度に Pandora プロジェクトと称してデータの所在やファイル形式の違いを吸収する転送方式の実用化に向けた開発を始めた(豊田 2001)²。プロジェクト名の Pandora とは「すべてのデータ」という言葉に相当するギリシャ語であり、ここでは多様なデータにその所在やファイル形式を意識することなくアクセスできるようにしようという目標を表している。この仕組みを利用することにより、NAPS ではファイル形式の違いを意識することなく、リモートホストにあるさまざまなデータをローカルディスクにあるかのように読み出すことが可能になる。また、平成 17 年度に更新した気象情報伝送処理システム(アデス)では、従来の電報配信とは違った新しいデータの交換方式と、データの所在や形式によらずにアクセスできるデータベース機能の導入が検討され、数値予報データの交換には Pandora による転送方式が採用されている。

本節では分散システムにおける多様なデータの取り扱いに関する問題を Pandora がどのような仕組みで解決するのか、簡単に紹介する。略語は巻末の付録にまとめてある。

2.2.2 分散システムにおけるデータの扱いに関する問題

近年の大規模な計算機システムは、スーパーコンピュータを中核として、多数の支援コンピュータ群をネットワークで接続した分散システムの形態をとるようになってきた。たとえば、NAPS においては、数値予報モデルなどの大規模計算をスーパーコンピュータで実行し、データの可視化や検証、プログラムの開発などをおもに各種のサーバで実行するため、各マシンはネットワークで接続されている。また、アデスは、データサーバおよび東西の中核システムを中核に、府県等の端末を結ぶ TCP/IP 基盤通信網として構成されている。

最近の高性能計算に関する話題に目を向けると、並列計算機による大規模計算だけでなく、ネットワークを利用して分散した計算機資源を有効活用するグリッドコンピューティングに関する技術の発展が注目されるよう

になってきた³。グリッドコンピューティングは、計算機の CPU 資源を遠隔利用する計算グリッド、遠隔システムにあるデータへのアクセスを容易にするデータグリッドのほかに、WWW と計算グリッド、データグリッドなどの技術を背景にしたサービスグリッドに分類することができる。データグリッドにおいては、さまざまなシステムに分散した多用な形態のデータへの効率的なアクセスをはじめ、データハンドリングが重要な課題となる。分散システムにおいてデータを共有するためには、ファイルがディスクにあるのか、テープに格納されているのかといった格納場所の多様性や、ファイル形式の多様性を考慮する必要がある。たとえば、NAPS やアデスで扱う格子点値 (GPV) を格納するファイル形式には数値予報標準データセットシステム (NuSDaS) や国内二進形式をはじめ、GRIB などのさまざまな種類がある。したがって、数値予報の結果である GPV を配信先に応じたプロダクトとして加工したり、GrADS などのさまざまな可視化ツールに入力したりするためには、ファイル形式の変換が必要である。さらに、データを作成した計算機における CPU のバイトオーダーの違い、つまりビッグエンディアンであるかリトルエンディアンであるかの違いを考慮する必要がある⁴。現 NAPS や次期 NAPS のスーパーコンピュータはビッグエンディアンマシンであるのに対して、次期 NAPS の各課業務処理用サーバはリトルエンディアンマシンであるため、互いのマシンで作成した多バイトのバイナリデータをそのまま処理することはできず、処理プログラムにおいてバイトオーダーを変換しなければならない。

システムが分散して多様になるにしたがって、各種のデータ形式に対応してプログラムを書き換える作業量は膨大になる。したがって、多様なデータ形式の変換処理に何らかの枠組を設けて整理し、統合を図る必要がある。

2.2.3 Pandora による解決策

データハンドリングにまつわるさまざまな問題を、Pandora ではつぎの方法により解決する。

(1) ネットワークを通じたデータの通信に HTTP1.1

³ たとえば、2004 年 10 月に ECMWF で開催された「気象学における高性能計算技術の利用に関するワークショップ」では、DKRZ, NERC, NOAA が取り組んでいるグリッドコンピューティングや分散データ環境に関する講演があった。

http://www.ecmwf.int/newsevents/meetings/workshops/2004/high_performance_computing-11th/

⁴ 多バイトのデータをメモリに配置するときのバイトの並び順は CPU のアーキテクチャによって異なる。最上位のバイトから順に配置するビッグエンディアンと、最下位のバイトから順に配置するリトルエンディアンがあって、バイトオーダーが異なるマシンで多バイトのデータをやり取りするためには変換が必要である。

¹ 成田正巳

² 2001 年気象学会秋季大会の講演資料:

<http://www.gfd-dennou.org/arch/zz2001/msj-aut-toyoda/poster/>
第 3 回気象庁モデルフォーラム(2002 年)の講演資料:
http://pfi.kishou.go.jp/open/mdlfrm/forum_03/hasegawa/

を利用⁵。

インターネットにおける中心的なデータ転送仕様である HTTP1.1 を採用することにより、分散ホスト間のデータ通信における信頼性の確保や低水準処理を隠蔽するためのツールやライブラリの利用、通信経路において複数の拠点を經由する際の問題への対応、応答速度を確保するための手法など、これまでに十分な実績のある技術をそのまま利用できるようになる。また、HTTP は FTP と比べてキャッシュや負荷分散の処理が優れているため、大規模なサーバを構築することに適している。

(2) データ形式の記述に MIME メディアタイプの書式を利用⁶。

電子メールや WWW におけるマルチメディアデータの交換において基盤的な役割を果たしている MIME メディアタイプを採用することにより、文字データ以外の各種のデータの取り扱いが容易になる。

(3) データの変換に CGI の機構を利用。

データの変換に用いるプログラムには一般的な CGI の機構を用いており、バイトオーダーの変換など、必要な処理を柔軟に追加できる。

2.2.4 Pandora のシステム構成

図 2.2.1 に Pandora の概念的なシステム構成を示す⁷。

(1) Pandora ライブラリ

可視化ツール等のアプリケーションプログラムがデータを入出力する低水準な階層として Pandora ライブラリを利用する。アプリケーションは、入出力においてデータの実際の所在を抽象化した指定でデータを要求する⁸。Pandora ライブラリはこれを HTTP に置き換えて Pandora サーバに要求を発行し、サーバの応答を解釈してアプリケーションに渡す。

(2) Pandora サーバ

クライアントである Pandora ライブラリからの要求を受け取り、応答を返す。この通信は HTTP で行う。要求を自身で処理できる場合は自身で管理するドライバを起動し、自身で処理できない場合はデータ所在データベースに登録された情報に基づいてほかの適切な Pandora サーバへ要求を転送し、クライアントからの応答を中継する。

(3) ファイル形式ドライバ群

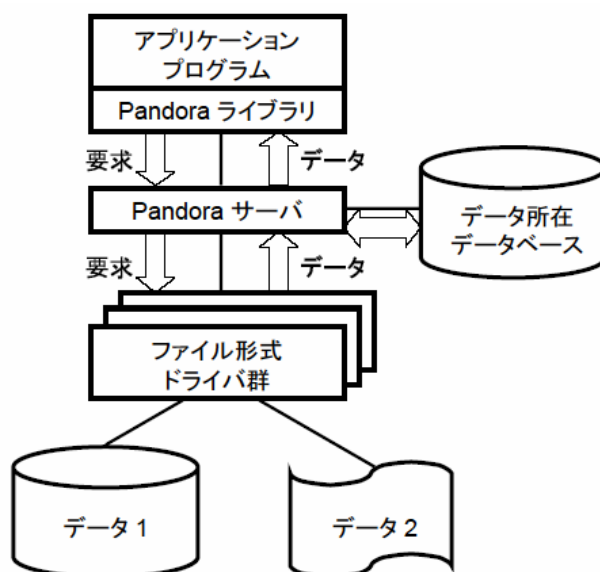


図 2.2.1 Pandora の概念的なシステム構成

クライアントに供給されるデータは、いずれかのサーバにおいて何らかの媒体に格納されたファイルとして管理されるものである。このファイルにアクセスするためのプログラムを、データを読み書きするドライバとして管理する。

以上の構成により、アプリケーションプログラムの利用者は、データがローカルディスクのファイルに存在する場合と同じように、ネットワークを通じて他ホストのデータにアクセスすることが可能になる。

2.2.5 おわりに

平成 15 および 16 年度に実施された予報課程特別研修では、Pandora の仕組みにより、本庁に設置されたデータサーバに保存されている数値予報 GPV をネットワーク経由で利用した。また、本庁に設置された防災情報モデル開発システムで実行したモデルの結果である GPV を、Pandora の仕組みにより各官署から利用することができるようになっている。

次期 NAPS においては、Pandora の仕組みを使って NAPS の内部からだけでなく、NAPS の外部にあるマシンから過去の数値予報の結果を保存しているデータバンクにネットワーク経由でアクセスできる仕組みを構築する。この仕組みの利用対象については、現在、検討を進めているところである。

参考文献

豊田英司, 2001: 数値データの入出力インターフェースの内部実装による Web 分散コンピューティングの構想. 気象学会秋季大会講演予稿集, 80, P333.

⁵ RFC 2616 - Hypertext Transfer Protocol -- HTTP/1.1: <http://www.faqs.org/rfcs/rfc2616.html>

⁶ RFC 2046 - Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types: <http://www.faqs.org/rfcs/rfc2046.html>

⁷ ここに示した構成は、Pandora プロジェクトの初期に想定されていたものである。

⁸ Web ページを指定する「http://ホスト名.ドメイン名/パス名/ファイル名」の形式と同じで、Pandora ではパス名の部分にデータの実際の所在ではなく要素名などの種別を指定する。