誰でも使えるPC普及の一方で ハイパフォーマンス

国家間競争になって ハイパフォーマンス

2. 地球シミュレータの構成技術

今年3月末に，米国から日本のHPCを調査するための政府レベルの2つの調査団が来日し，地球シミュレータセンターやJAXA，核融合研究所，理化学研究所等，日本を代表するスーパーコンピュータを保有する研究機関や文部科学省および関連法人を訪問し，利用状況，日本の科学技術政策の中でのスーパーコンピュータ開発の経緯，今後の開発計画等について精力的なヒヤリングを行った。

この調査レポートは，今後，公表されると思うが，暫定版が米国ナショナルアカデミーから刊行されている。その中に，地球シミュレータに関する項目がある。引用すると「地球シミュレータには，目新しい技術は使われてはいないが，高メモリーバンド幅とレーテンシ隠蔽ハードウェアを持つ，専用化されたマイクロプロセッサ技術とノード間のネットワークにかなりの費用を振り向けて，トータルな性能を実現している」としている。

これは，まったく正鵠を得た指摘である。日本流に言い換えれば，「地球シミュレータは，当時としては，世界最新の0.15ミクロン幅の半導体技術でチップが製造されており，ベクトルプロセッサアーキテクチャが採用され，ノード間はワンレベルクロスバで構成されている」ということになる。

0.15ミクロン技術は，本誌9月号でもふれたように，米国半導体工業会（SIA）のロードマップ上でも，いわゆる「ムーアの法則」からも予想されていた線上にあるわけだし，ベクトルプロセッサ技術は，もともと米国のシーモア・クレイ氏によって1970年代に開発された技術である。

ワンレベルクロスバにいたっては，コンピュータの出現より古い，交換技術の初期の段階から存在していた技術である。なぜ，米国は「コンピュートニック」なる言葉まで作り，「TOP500」で地球シミュレータに連続5回もの独走を許す羽目になったのか，その分析をし，「次」には，前回の轍を踏まぬよう細心の注意で取り組むという米国の強い意思が感じられるレポートである。

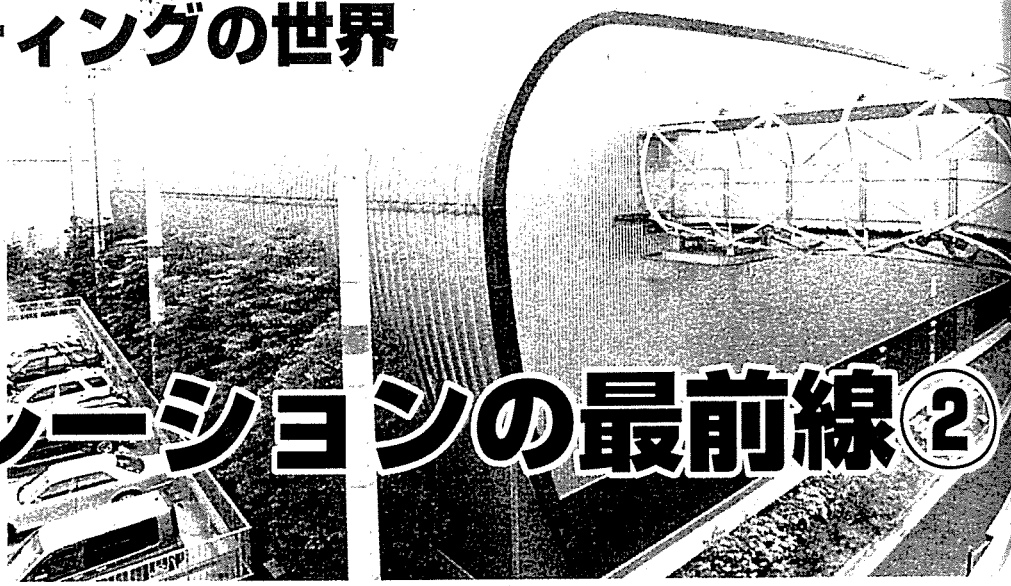
日本には「温故知新」という言葉がある。地球シミュレータ開発にあたって，古い技術がどのように磨かれ，極限近くまで引き出されたのか振り返ってみたい。

2.1 システム全体構成

地球シミュレータ全体の構成は，大変複雑では

着実に進展している コンピューティングの世界

いる シミュレーションの最前線②



避雷用のワイヤと支柱，地球シミュレータ棟への渡り廊下

独立行政法人　海洋研究開発機構
地球シミュレータセンター センター長補佐　**平野　哲**

あるが，640台のスーパーコンピュータ（プロセッサノード）が超高速ネットワーク（IN）で接続されており，これらの制御のために16ノードずつを束ねるクラスタ技術が開発されている。

各ノードは，クラスタコントローラ（CCS）を介してスーパークラスタコントローラ（SCCS）に接続されている。SCCSはホットスタンバイ構成となっており，バックアップ含めて二組ある。

通常，一組は，バージョンアップや開発作業に使われている。40クラスタのうち1クラスタは，Sクラスタといって，1ノード内のバッチジョブ用として14ノード，会話型処理用として2ノードに分けている。残りの39クラスタはLクラスタといって，大規模並列処理用に利用されている。

CCSは40台あり，CCS，SCCSは，ギガビットイーサネット（GbE）接続で制御情報がやり取りされている。ジョブの投入は，通常，地球シミュレータ本体とLAN接続されているログインサーバーから行われる。

ジョブの実行は，SクラスタまたはLクラスタのノードで行われる。ファイルは，Sクラスタに接続されているユーザーディスク（Sディスク）として230TB，Lクラスタに接続されているワークディスクとして460TB，マスデータプロセシングシステム（MDPS）にあるユーザーディスク（Mディスク）として250TBの合計940TBの容量を持っている。

さらに，MDPSは，Mディスクとテープライブラリ1.5PBで階層型ファイルを実現している。

実行に先立っては，MDPS機能を使って必要となるファイルがワークディスクに転送される（ステージイン）。処理結果は，速やかにMDPS機能を使ってワークディスクからユーザーディスクにコピーされる（ステージアウト）という。

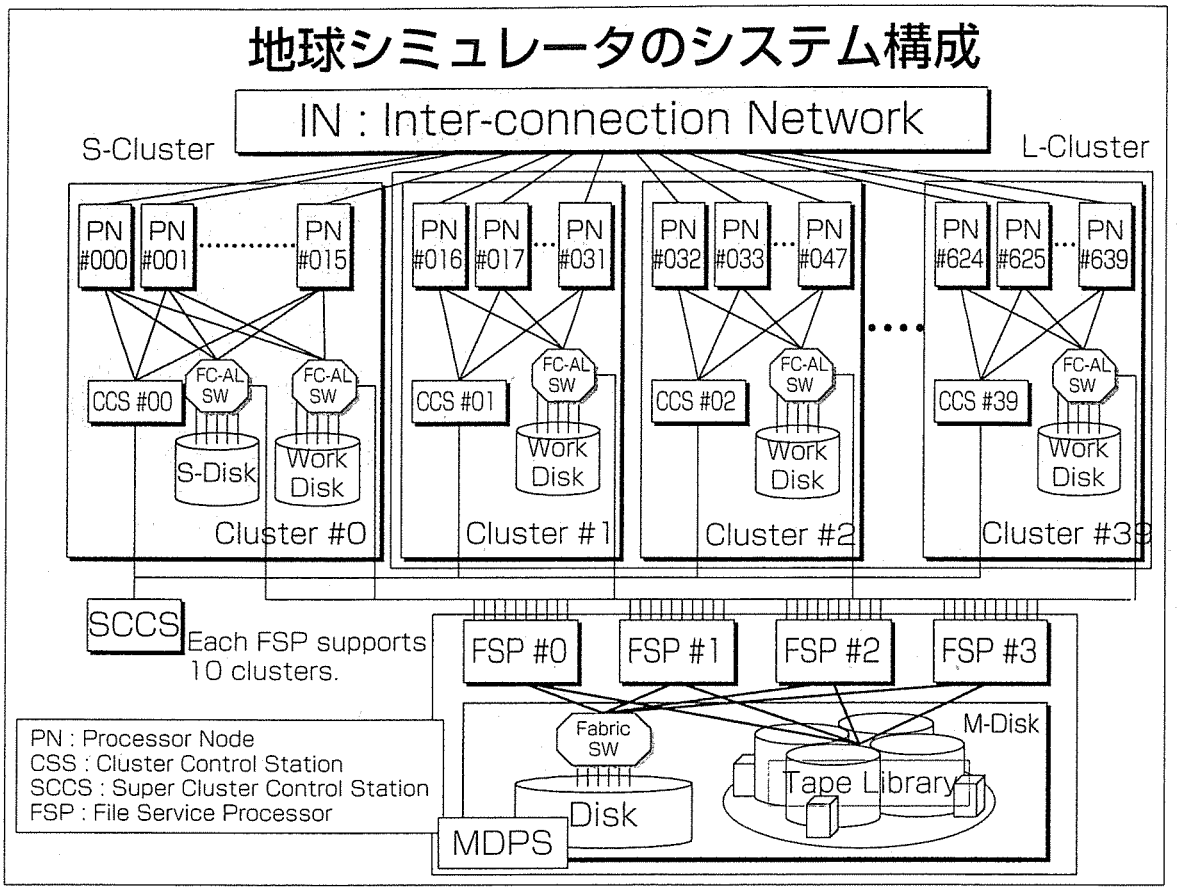
この機能により，遅滞なくLクラスタを計算処理に投入させることが可能となっている。

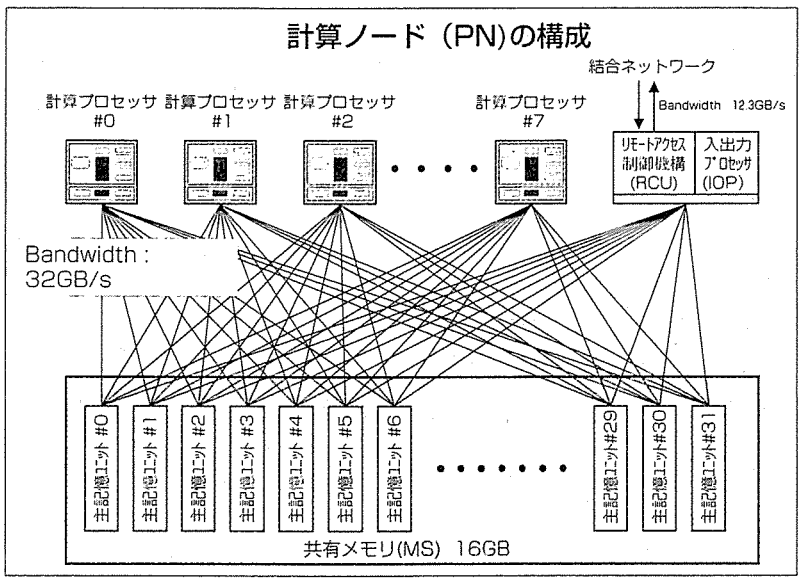
2.2　プロセッサノード(PN)

地球シミュレータは，640台のPNから構成されている。PNには8個のベクトルプロセッサがある。

1個のベクトルプロセッサの性能は，8GFlopsなので，PNとしては，64GFlopsの性能を持っている。メモリは，8個のベクトルプロセッサで共用される16GBの容量を持つ，いわゆるSMP（共有メモリ並列プロセッサ）方式のアーキテクチャとなっている。

PNを構成するベクトルプロセッサは，言うま





でもなく，地球シミュレータの要であることは間違いないことではあるが，このベクトルプロセッサが高い実効性能を出すためには，メモリおよびメモリとのインターフェイスが重要な役割を担っていることを説明したい。

スーパーコンピュータでは，多くのアプリケーションがメモリの広いアドレス空間に対してアクセスを行うが，特に多いのは，配列データのように規則的に並んだデータである。ベクトルプロセッサの8 GFlopsという高い性能を如何なく発揮させるためには，メモリとのデータのやりとりが途切れないようにスムースに行われなければならない。

PNでは，9個のポート，すなわち8個のベクトルプロセッサ，入出力／結合ネットワーク（IN）と32個のメモリ装置との間をクロスバ接続することにより，1プロセッサあたり32GB／sのバンド幅を持つ転送ができる設計となっている。

したがって，PNにおけるメモリアクセスのトータルバンド幅は256GB／sという巨大なものと

なる。スカラパラレル方式のスーパーコンピュータでは，メモリアクセスに対して，キャッシュメモリを多用しているが，大規模アプリケーションでは，キャッシュ上でヒットするケースが少ないことから演算性能が発揮されないことが多い。

もちろん，特定のアプリケーションで，キャッシュを効率よく使えるものもあることも事実である。メモリチップからの読み出し／書き込み速度は，30ナノ秒程度であるから，PN全体では，2048ウェイのインターリーブをかけて，しかも，配列データをメモリに配置するときにアクセスが衝突しないようコンパイラが処理を行って，上記バンド幅を実現しているのである。

メモリチップとしては，地球シミュレータ用に開発された128MBのフルパイプラインメモリが採用されている。システム全体では，100万個以上のメモリチップが使われている。

地球シミュレータの開発においては，ベクトルプロセッサの１チップ化の他にも，メモリ制御のためのチップ，リモート制御アクセス機構（RCU）のチップ，他にINの制御のために２種類の合計５種類のLSIが開発された。

ベクトルプロセッサは，何故速いのかという質問をされることがある。以上述べた説明からもある程度は理解されようが，やはり，速く処理できるように物量が投入されているという答えが正解と思うのだが，質問者は，もう少し，画期的な何かがあると期待しているらしく，この答えに「ピン」とは来ないようである。

さてベクトルプロセッサ（AP）チップの説明に移ろう。このチップは，NECの一世代前のスーパーコンピュータであるSX５のアーキテクチャを踏襲している。このAPチップの概要を図に示す。

８セットのベクトル乗算パイプラインと加算／シフトパイプラインが同時動作可能なようにデータがベクトルレジスタから連続的に供給されることで，８GFlops性能が発揮されている。

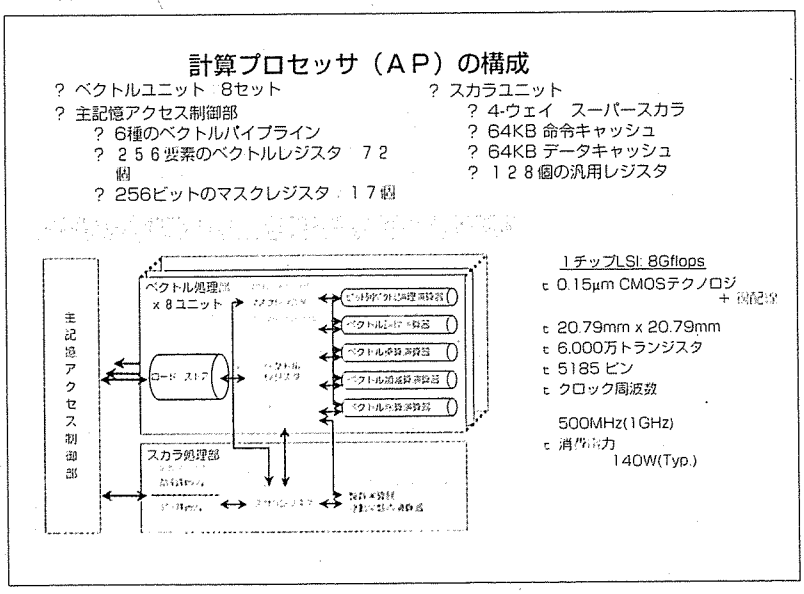
このチップは，スカラユニットを除き，キャッシュを持たず，もっぱらメインメモリとの高いバンド幅を持つメモリアクセスに委ねることにより，高い実効性能が維持されている。

もちろんコンパイラは，各種の演算パイプラインを遊ばせることなく，データをスムースに供給できるよう，最適なオブジェクトコードを生成している。このチップは，6,000万個のトランジスタを詰めこんだものだけに21mm角という大きなものになった。

通常半導体チップというものは，一枚のウエハーから何万個という機能素子を作り出すことにより，大幅なコストダウンが計られるものであるが，このチップはまったく違う発想で「これが小さくできないとシステム全体が構築できない」ということからきているわけで，本誌９月号で述べたように，大いに気を揉んだところであった。

当然のことながら１枚のウエハーからまともに動くチップが取り出されるようになるには，かなり長い時間を要したのである。図は，このチップと基板，冷却のヒートパイプである。

APチップのみならず，チップを搭載する基板も新規開発する必要があった。APチップは，5,185本のピン（シグナル，電源，アースなど）を持ち，21mm角内に集積されているので，基板としても，この集積度でピンを受けねばならず，6層の多層基板をベースとして，上面に４層，下面に４層の配線を半導体を作るように積層したのであ



る。

基板上のライン／スペース幅は，25／25ミクロンmである。APチップは，140Wの消費電力があり，発生する熱も多く，冷却も大変重要な課題となった。最終的には，ヒートパイプ方式により解決したが，中の気圧を下げて冷媒を気化しやすくしてある。さらに冷媒が，気相，液相変化する過程で混ざりあって冷却効率が下がるのを防ぐために，冷風が当たる面に，角度をつけたラジエータ構造となっている。

PNにはこの装置が8個実装されている。

一つの筐体には，2ノードが格納されており，電源などを共通化することにより，省スペース化が図られている。PNの筐体は，全部で320ある。

以上のような方法でコンパクト化を計ったが，過去のスーパーコンピュータと床占有面積を比較すると図のようになる。SX5と較べると，実に1／18の大きさとなった。さらに，ネットワークの筐体も小型化を図り，あわせて50m×65mの建屋の中に収めることが可能となった。この18倍のスペースというのは，考えるのも恐ろしい広さである。電力も一世代前では50KVAであったが，PNは8KVAであり，1／6となった。現在は空調関係を含んで，6,000KW／hを使用しているので，一世代前では36,000KW／hと膨大な電力量となり，電力の安定確保が大きな問題となったろう。

2.3 ノード間ネットワーク(IN)

大規模なアプリケーションプログラムは，多数のノードに処理を分散化することにより，処理時間の短縮を図る。当然なこととして，多数のノード間において共有メモリ上のデータを交換する必要に迫られる。ネットワークアーキテクチャは交換されるデータの特性に依存するのは明白である。

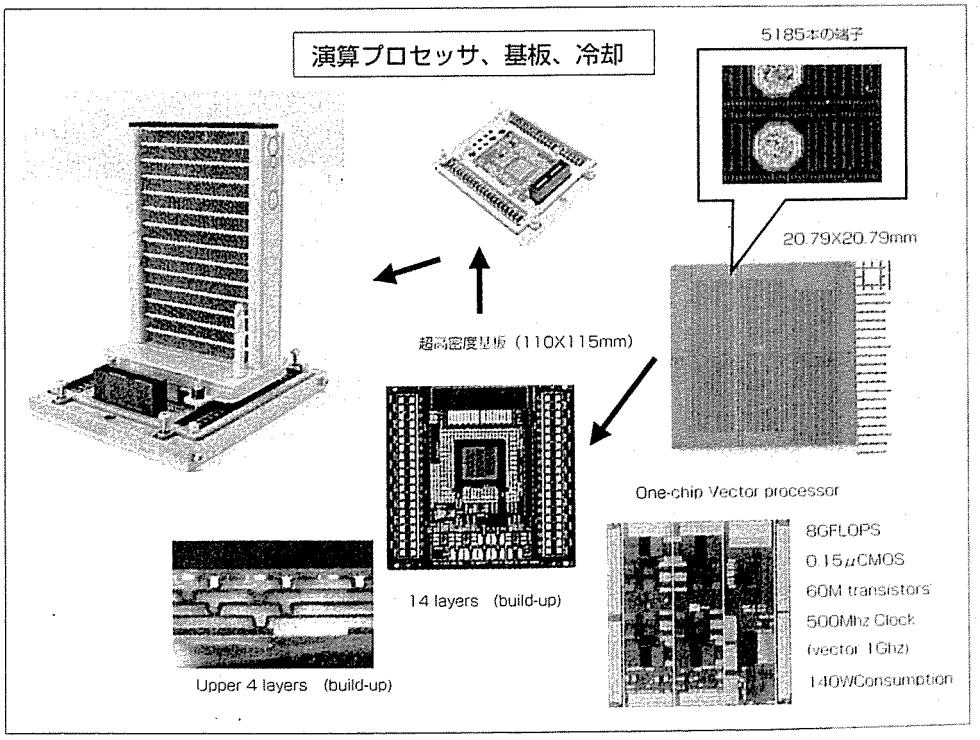
ネットワーク内で経過する時間は，それ自身，演算性能を高める方向には寄与しないので，様々なアプリケーションに対し，

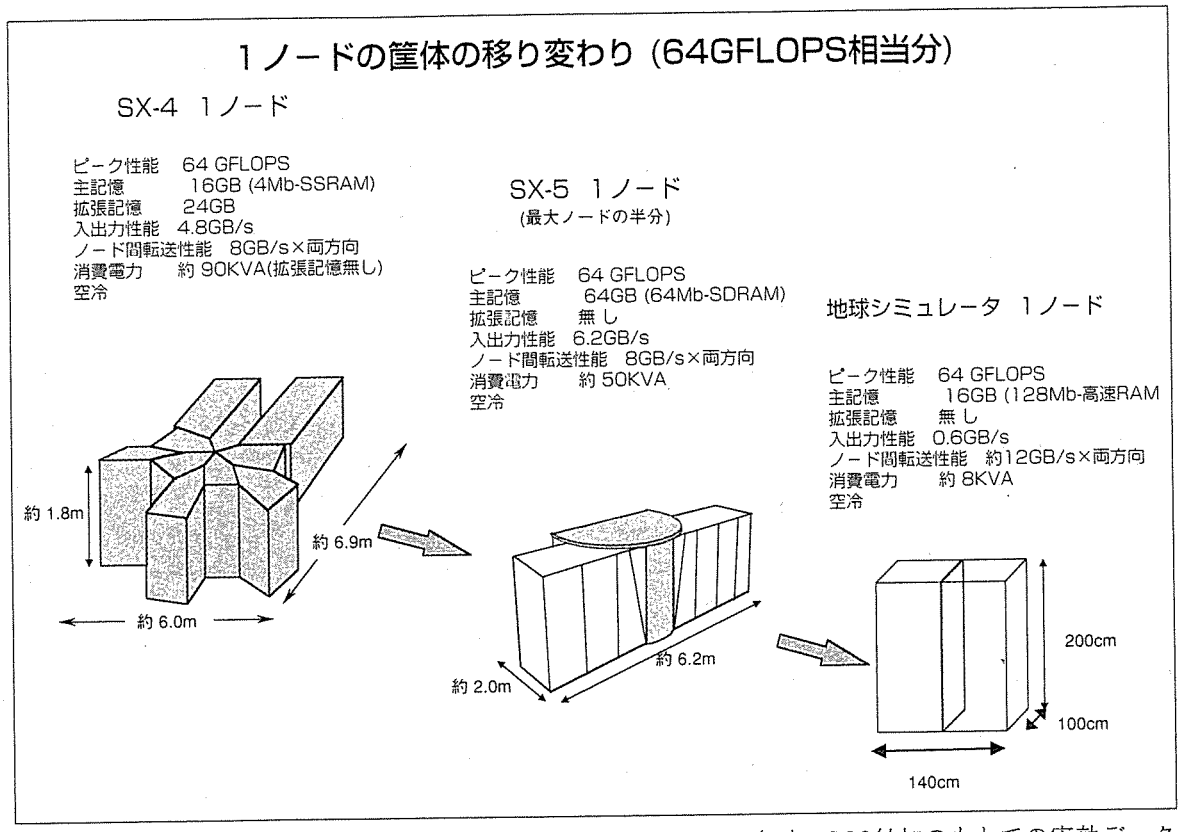
①ネットワーク遅延（レーテンシ）をできるだけ小さくする，

②できるだけ大量のデータを迅速に交換できるようにする，

という二つの用件を満たす必要がある。光の速度（＝電気の伝わる速度）は有限であるので，ケーブルの長さも短いに越したことはない。真空中では，1mあたり3ナノ秒かかる。同軸ケーブルでは遅くなるので，倍ぐらいの6ナノ秒とすると，30mでは，200ナノ秒くらいかかることになる。全体で40TFlopsで動作している演算性能からみると決して無視できる時間ではない。

光ケーブルではさらに，電気・光変換，光・電気変換の時間が純増となるので，採用は得策でなかった。また，ノードを大きく作ると，設置スペースが広くなり，ケーブルが長くなるのでデータ交換に





時間がかかるようになり，実効性能を下げる要因になる。

特に，レーテンシについてはアプリケーションが特定できないので，もっとも単純なN:N単段（1レベル，N＝640）クロスバで実現することした。

開発目標としては，PNのメモリバンド幅である32GB／sの１／４以上である８GB／s以上とした。ノード間を１本が1.25GB／sのメタリックケーブル128本を束ねて，同時にクロスバで切り替えることにより12.3GB／sのデータ転送能力を実現した。

この能力は，PN内でプロセッサが共有メモリを介してメモリコピーを行う場合（ロード，ストア）とすると16GB／sとなるので，コンパラな能力となる。すなわち，PN内も外も，データ交換に要する時間は大した違いがないということになる。

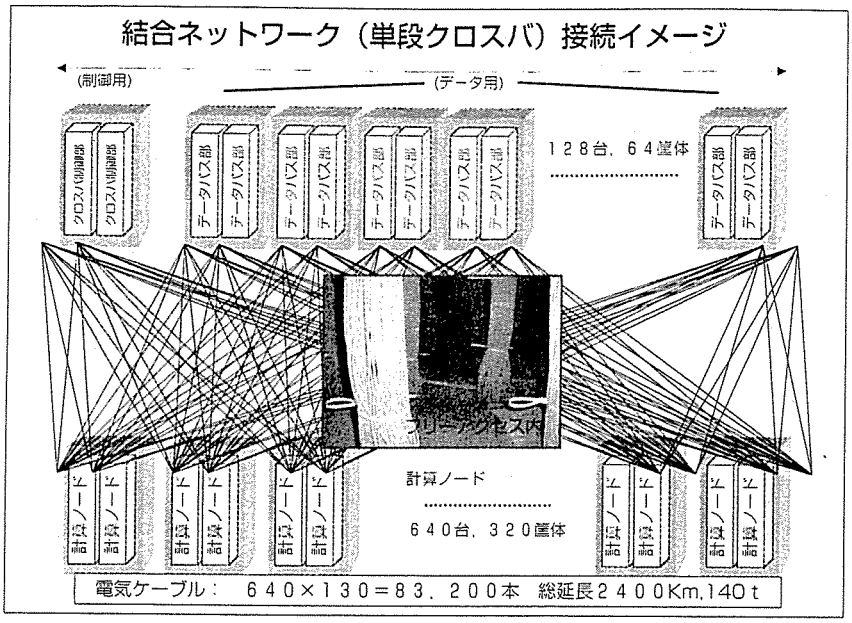
オフィスでワークステーション間がギガビットイーサで接続されている環境と較べてみると，この違いは際立ったものである。ケーブルにメタリックケーブルを使って，差動信号により信号の授受を行うが，ノイズなどによる誤作動を防止するためにECCを付加して冗長性を高めてある。

12.3GB／sは，ECC付加のもとでの実効データ転送性能である。物理転送性能は16GB／s。制御ケーブルとして２本を必要とするので合計130本のケーブルが１ノードあたり使用されている。システム全体では，83,200本のケーブル本数となり，トータルな敷設距離は2,400Km，重量140トンである。

このケーブルを引くために，50人×3.5ヶ月を要した。MPI（OS込み）によるノード間のレーテンシは，5～７マイクロ秒，MPI_SENDのスループットは11.8GB／sが測定されている。

INの重要な機能にグローバルバリヤ同期機構がある。多数のノードにまたがって処理が行われるので，それぞれのノード内で動作するプログラムが同期をとるようにしないと効率のダウンが起こる。これをいちいち，ノード間で連絡をとりあうと時間はかかるし，時間もずれることとなる。

グローバルバリヤ同期機構は，INの制御部の中にレジスタがあり，このレジスタに同期が必要となるノード数をセットしておけば，それぞれのノードが同期点に達すると，このレジスタが減じられ，0になるとすべてのノードに通知されるハー



ドウェア機能であり，640ノード全部が同期をとる場合でも3.3マイクロ秒で実行できる。

2.4　オペレーティングシステム

地球シミュレータのオペレーティングシステムは，NECのスーパーコンピュータであるSXシリーズのオペレーティングシステム「SUPER-UX」をスケーラビリティの面から拡張したものに，巨大システムの運用管理上必要となるスケジューラ機能や複数ノードにまたがる並列ファイルアクセスを可能にする機能などが強化されている。

それぞれのノードが持っているクロックの時刻合わせひとつにしても，数が多いと困難が伴うものである。640ノードという多数のノードを制御するために，16ノードを1クラスタとし，クラスタコントロールステーション（CCS）が制御し，さらに40クラスタを統括するためのスーパークラスターコントロールステーション（SCCS）を設け，SCCSからCCS経由で，ノードに向かってコマンドを発して，CCSによって集められた診断情報，システム情報からシステム全体の効率的な運用を司っている。

診断情報は，PNごとにある診断プロセッサにより常に情報が収集されている。なお，定期保守は，クラスタ単位で1週間で1日だけ停めて行っている。それでもINを除き，640ノード全部を保守するのに40週間かかることになる。

Lクラスタにおいては，利用者は，ジョブの投入にあたり，必要なノード数と時間の指定を行う。この時間は，そのノードの8個のプロセッサ（64GFlops）とメモリ（16GB），ワークディスク（約300GB程度）は占有利用となる。

占有利用となるので，ノード内にある8個のプロセッサは，最大限使い切って欲しいということをお願いしている。すなわち，ベクトル化率，並列化の良し悪しがノードの効率を決めるので，利用者にはそれらの最適化をくどいほどお願いしているのは，この理由があるからである。

ノード内の1個のプロセッサが故障しても残りの7個で動作させることはできるが，即座に，そのノードを止めて修理を行うこととしているのは，常に最高の環境を提供していくという考えに基づくものである。

利用者が投入してくるジョブ（リクエストという）の要求リソースは，10ノードから512ノードまで，時間も最大で12時間までと，ばらつくので，スケジューラはキューに登録されているリクエストの要求リソースを組み合わせて，最適な使用効率を得るように設計されている。

場合によっては，後のリクエストが前のリクエストを追い越すこともあり得る。

この部分は，当初は地球シミュレータ研究開発センターで開発したが，マスデータプロセシングシステム（MDPS）の導入に伴い，NECの標準サポートにロジックを取り込んでいる。リクエストが途中でキャンセルされたり，アボートしたりすると，当てはめ戦略を再計算する。

利用者のリクエストは，毎回同じノードで処理されるとは限らない。

最近，CAEの商用パッケージを地球シミュレータで走らせようとしたが，ほとんどのパッケージ

は使うプロセッサを固定して登録するので，地球シミュレータのような，どこで実行されるのかわからない方式では契約できないことがわかり，結局5,120個のCPU全部を登録するという，思ってもいない事態が発生して目を白黒したものである。

2.5 MDPS

マスデータプロセシングシステム（MDPS）は，昨年6月から試験運用が始まり，10月から本運用に入ったシステムである。

L系ノードで処理が終わったワークディスクのデータは，従来は，INを使ってS系のユーザーディスクに戻すか，L系に接続されている1.5PBの容量のあるテープライブラリにマイグレーションする方法のどちらかであった。

しかし前者は，PNの能力を単なる入出力のような，ほとんど計算しない処理に使うのは勿体ないし，後者は，テープライブラリからリコールによりワークディスクに戻し再利用するには都合がいいが，処理結果をそのまま使うには，外部に取り出すパスが細いという問題点があった。

その上，25,000巻ものテープと80台以上のアクチェータの組み合わせで障害に悩まされていたこともあり，MDPSの導入が決定されたのである。

MDPSは，4台のファイルサービスプロセッサ（FSP）と250TBのディスク，1.5PBのテープライブラリ，ワークディスクとのファイバーチャネル116本などから構成されている。

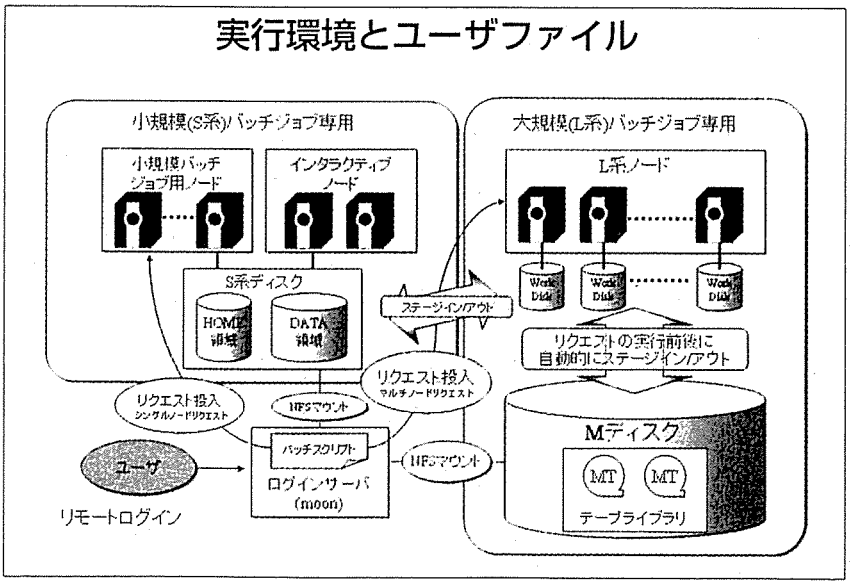
FSPは，ワークディスク⇒Mディスクのコピー処理を行う（ステジーアウト）と同時に，SCCSからの指示によりPNでの処理が始まる前に，必要となるファイルを予めユーザーディスクまたはMディスクからワークディスクへコピーする（ステジーイン）ことも行う。また，外部ネットワークへ大量のデータを移送する機能も持っている。計算能力が如何に高くても，足腰が弱ければ結果的に意味がないわけで，地球シミュレータのスループット／TAT向上に役立っている。

Mディスクが満杯になってくるとテープライブラリに掃き出される。テープの障害を考慮して，二重書きしているが，現在はどちらも満杯状態である。

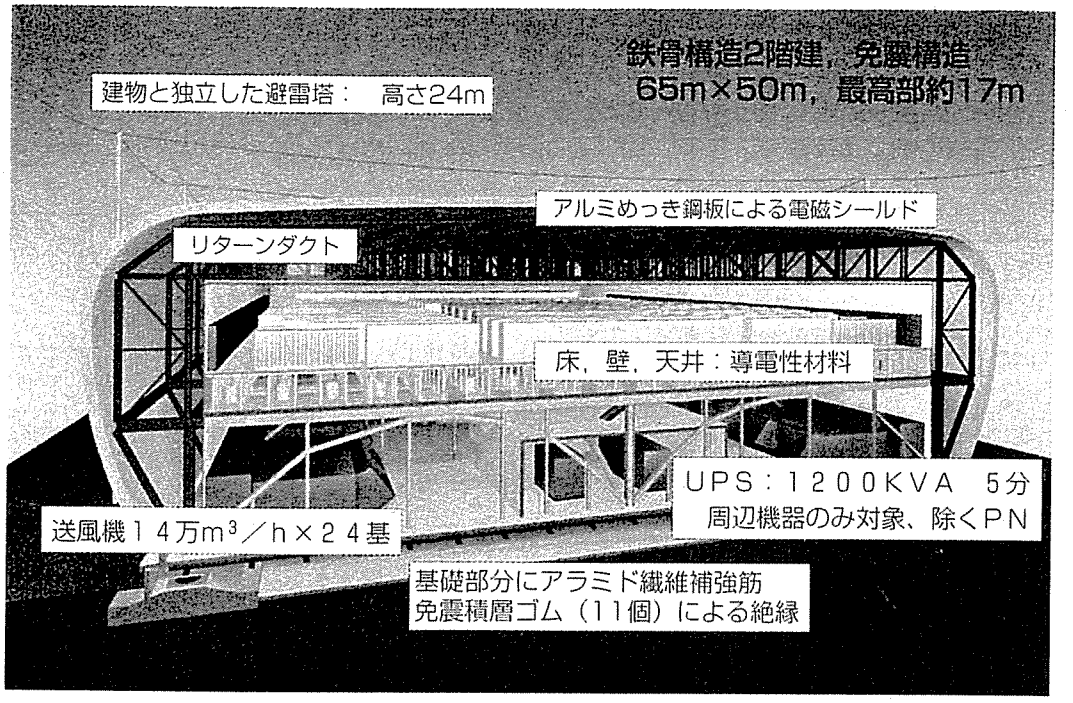
2.6 建屋および設備

地球シミュレータのような巨大なシステムは，耐震対策はもとより，落雷や電磁環境にも気を配らなければならない。また膨大な電力を必要とするし，その結果，発生する熱も大変大きなものとなる。そのために，コンピュータの設計技術者と建物の設計担当は，綿密な打ち合わせを行いつつ設計をすすめた。建物の大きさは横50m，奥行き65m，高さ17mである。

まず建物全体は，免震構造とするために，ゴムと鋼板を積層した直径1.1mのアイソレータ11個で支えられる構造とした。したがって，建屋であるシミュレータ棟と研究者の研究室がある研究棟4Fの渡り廊下は，それぞれの固有振動が異なるので，フレキシブルな構造としている。現在の地に建設が決まったときに，電磁環境の測定を行うと，近くを走る高速道路から27MHzの違法無線で，120dBμV以上の強電界が検出されたので，電磁シールドには相当の気を使っている。



シミュレータ棟の特徴的構造



建物全体は，電磁シールド仕様の鋼板でできているが，さらにコンピュータが設置される二階のフロア全体，INのケーブルが敷かれているフリーアクセス床内と三重のシールド構造となっている。トータルなシールドの性能は50dBである。

泣きどころは，分電盤ひとつ増やすにも工事費用が大変割高になるということである。アースについても，コンピュータ信号用，コンピュータのフレームグランド用，フリーアクセスフロア用，電力用など使い分けて，相互干渉を避けるため独立に距離を離して接地している。

また，コンピュータルームの中においても，本体から発する電磁波が建物の部材と共振現象を起こさないよう注意深く材料が選ばれている。避雷は，建屋から距離をおいて24mのポールを片側4本，合計8本を建てて，ワイヤを張って屋根に直撃がこないようにしている。

地下に流れた迷走電流による影響を避けるために，建屋の基礎には鉄筋を使っておらず，アラミドFRP筋樹脂で強度を保っている。空調設備は，建屋の外部にある冷却施設棟から冷水を引き込み，一階に設置してある24台の送風機（140,000m／h：2台は予備）で冷風を循環させている。一階と二階の間には，80cmのダクト230個があり，冷風（17度C）を吹き上げている。

この冷風は，二階のフリーアクセス，PNやINの筐体を通り抜けて冷却し，23度Cで二階の天井に抜けて側面の壁に沿って循環してくる仕掛けになっている。この送風量は，後楽園ドームの1.8倍に匹敵する。

地球シミュレータは，全体で約5,000KW／hの電力を消費する。この電力は，66KV特高変電設備から受電しているので，地球シミュレータセンターの立地については，候補地をいくつか挙げて検討され，現在の場所に落ち着いた経緯がある。

が，1,200KVAのUPSでは，ディスク装置，テープライブラリ，LAN，その他コントーラなどのバックアップを行って，5分程度持たせるようになっている。

PNは，研究用ということもあって商用直結となっている。この2年半の運転で初めて8月末に落雷による瞬断で停止したが，初めての経験だったので，ファイルの整合性チェックを念入りに行った後，すべての処理は，リスタート可能で復旧できた。(つづく)

(Hiroshi Hirano)