

システムサイドから見たこれまでの経験と将来像

地球シミュレータ開発

平野 哲

(元日本電気(株) 現(独)海洋研究開発機構)



講演者紹介

名前：平野 哲（ひらの ひろし）

現職：独立行政法人 海洋研究開発機構 情報技術審議役

略歴：1970年 早稲田大学大学院理工学研究科電気工学専攻修了

同年 日本電気㈱入社

大学, 研究機関, 中央官庁の大型コンピュータ, ネットワークシステムのSEとして従事

1987年 官庁システム事業部システム部長, 企画, 営業を経て

2001年 NECネットワークス主席技師長

2002年 4月より海洋科学技術センター(現 海洋研究開発機構)

地球シミュレータセンター長特別補佐,

地球シミュレータシステム管理・運用に従事、

2009年 現職

文部科学省、厚生労働省、総務省の各種専門委員を歴任、

情報処理学会, 日本物理学会



地球シミュレータ開発

2011. 9. 10

平野 哲

地球シミュレータ開発・史



2010年12月 開発史をWebに公開

<http://www.jamstec.go.jp/es/jp/publication/index.html#es-d>

既に 60,000ダウンロード



本日 展示会場にてCDを配布中

既に 本と合わせ 1,200部配布

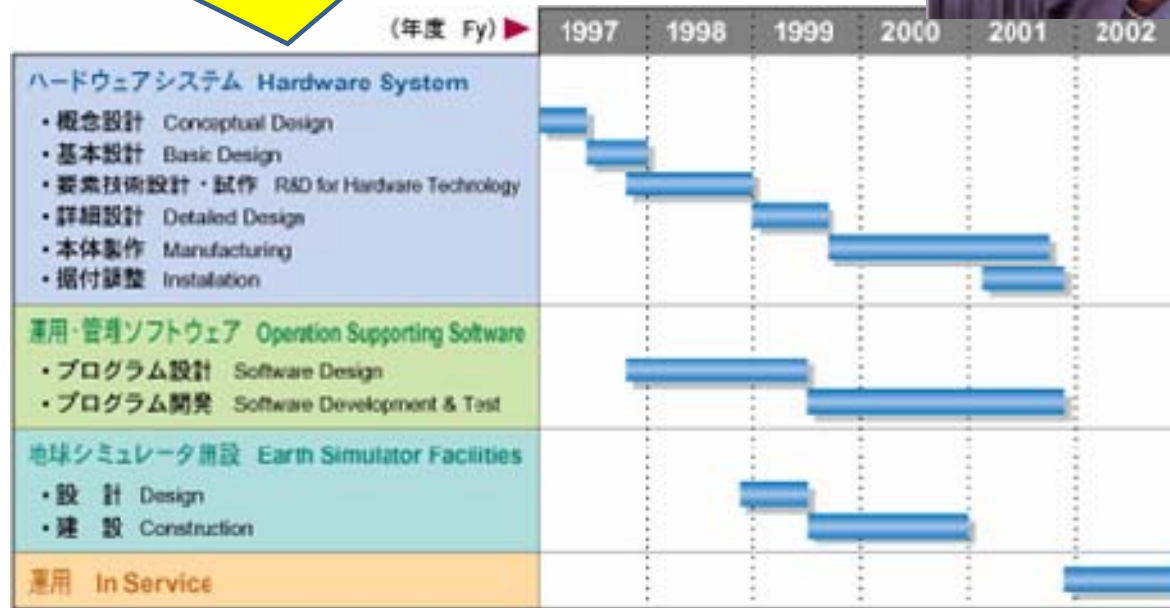
地球シミュレータ開発

- ◆ NECとの人脈創りが最初の仕事
- ◆ バランスを絶えず気にして 義理人情を大切に
- ◆ 時にはテクノロジーの実現性で掛けをして負けた事も
- ◆ 現場に行かれるのは何時も列車 無類の飛行機嫌い
- ◆ 2001.10 ESC来所、体力無く地球の見学叶わず
- ◆ 死者は生者を走らせるな、が持論 自身の葬儀も



三好 甫 先生

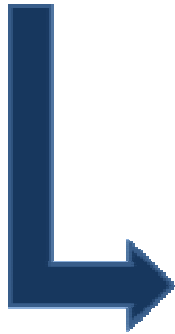
- ・国家的な視点
- ・広い見識・視野
- ・卓越した先見性
- ・現実を見極める
- ・旺盛な知識欲
- ・公平無私
- ・切磋琢磨を求め
- ・高い目標への挑戦
- ・同じ目線での議論



地球シミュレータ開発

◆ 三好先生のいだかれていたスパコンのイメージ

- ◆ 実用システム (実用に耐える) である事
 - 実行性能が目標以上である事 (ピーク狙いは不可)
- ◆ データ供給能力が重要 (主記憶重視・キャッシュ依存は不可)
- ◆ ネットワークはクロスバーで有るべき (隣接接続は好ましくない)
 - 研究的な危険なアーキテクチャ&ハードは採用しない
- ◆ ユーザがアプリ性能の予測が出来る事 (実行時の性能保証)

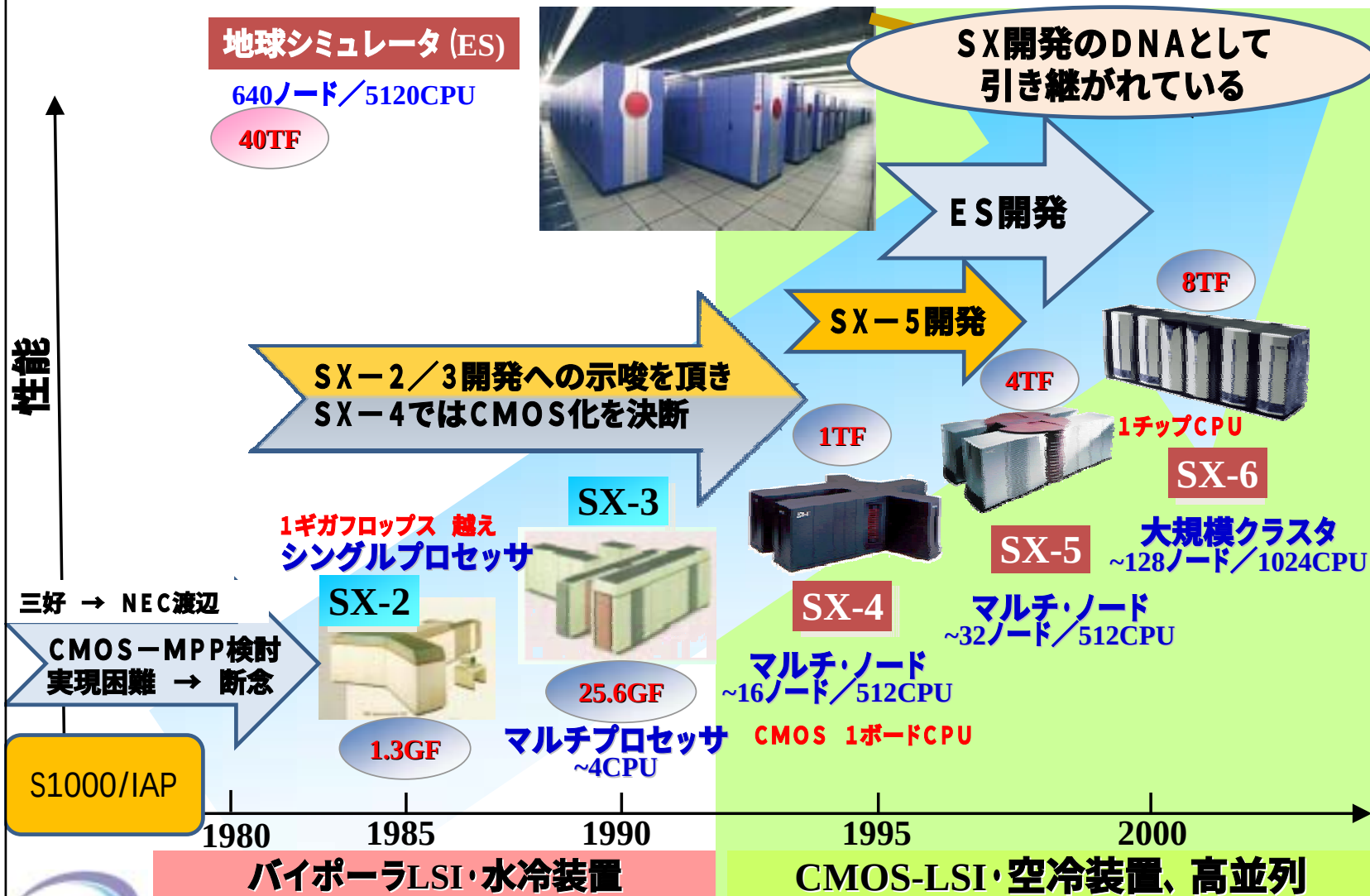


● NECの対応

- 1チップ CMOS ベクトルプロセッサ を実現:
 - メモリーインタフェース回路の工夫や分散で1チップを実現
 - コスト/サイズ/消費電力/信頼性 の大幅改善
- 新規な技術をむやみに採用しない:
 - NECで実績のある アーキテクチャ/OS/コンパイラ
 - 多重ベクトルパイプライン/ノード間は電気転送の採用
 - 代替えが無ければ革新的な技術を開発 (ビルドアップ基板)
- 高主記憶性能: 特注 FPLRAMの採用 (独自開発品の最後)
- 信頼性の確保: ノード間にバースト故障対応のコード付加 等

地球シミュレータとNECスーパーコンピュータ開発

5



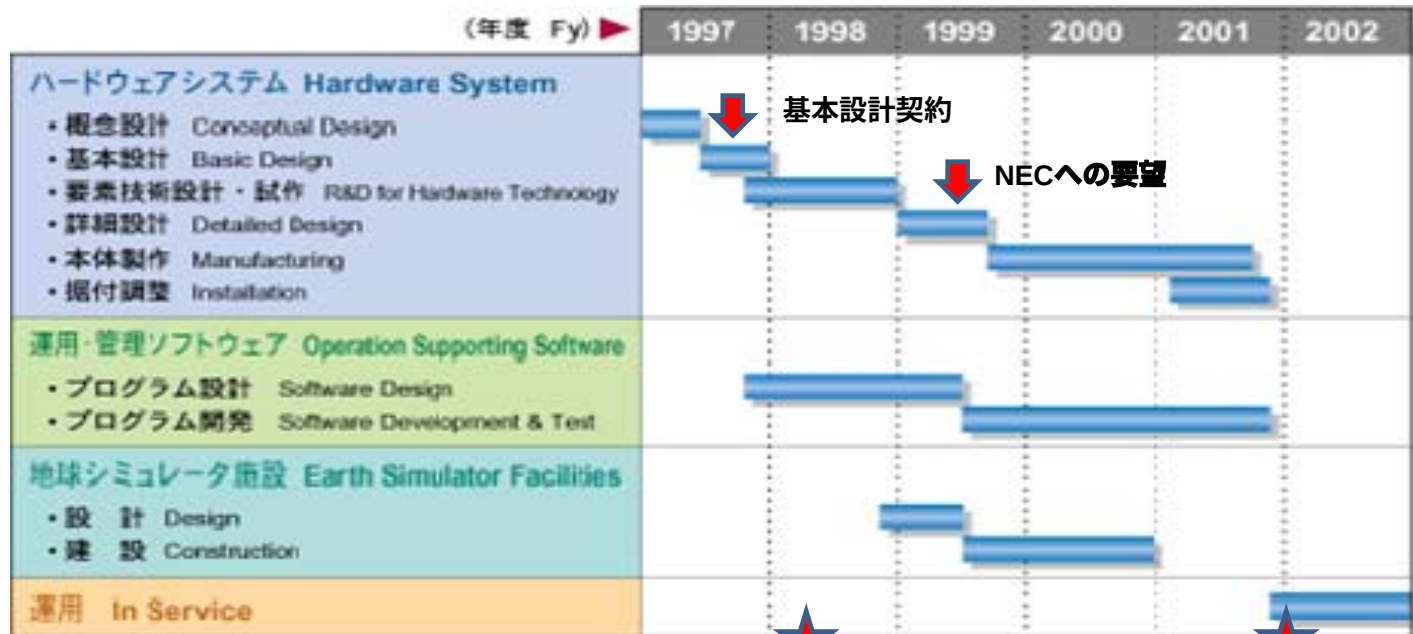
地球シミュレータ プロジェクト

当時のスパコンの1000倍の実行性能を実現

気象・気候・地球科学でブレークスルーを起こす

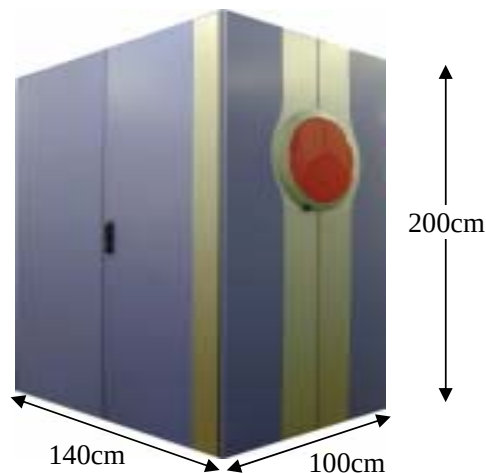
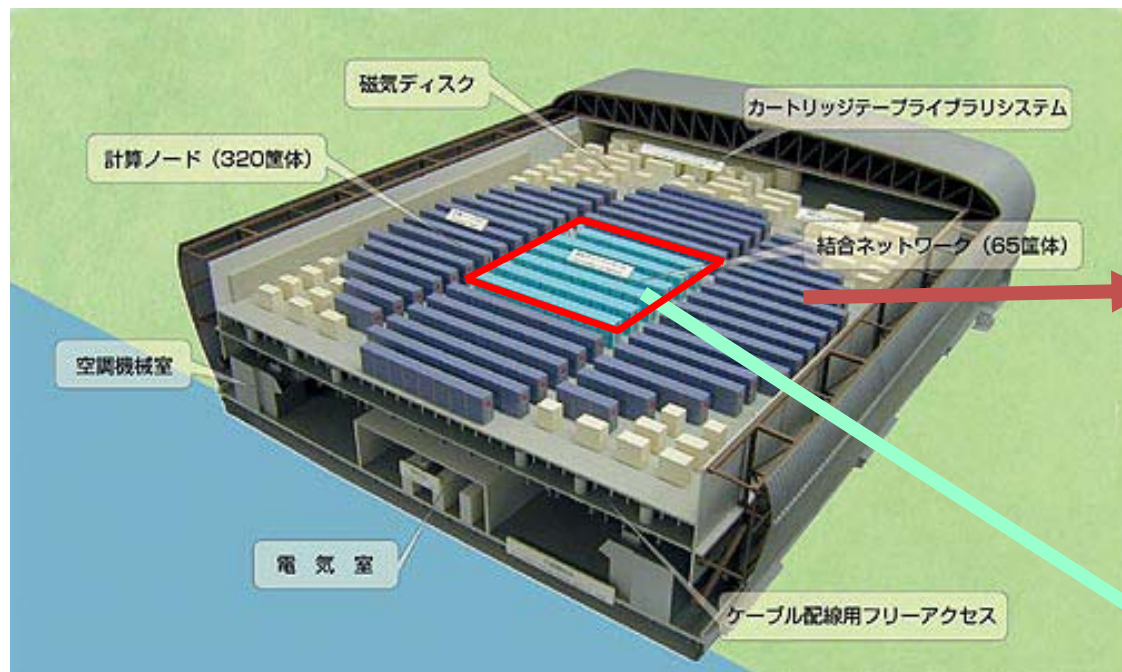
- 今までに経験した事のない規模を克服
- テクノロジーは1.5～2世代先を目指す
- 現行のSX開発との並行開発の実現

信頼性の確保・製造問題克服
要素技術も同時開発
人・物・金の調整

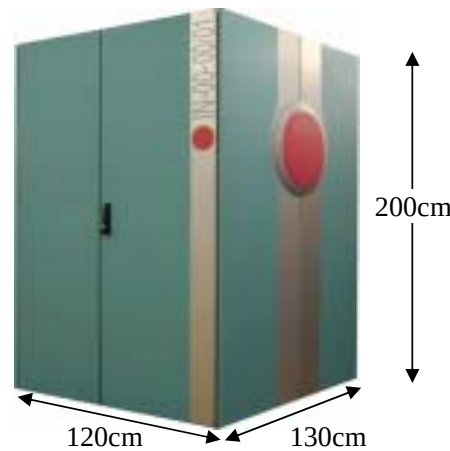


地球シミュレータシステム

7



PN筐体



IN筐体

PN筐体 320台 (2PN/PN筐体)

プロセッサ: 5120台、 ノード: 640台

総合最大ベクトル演算性能: 40TFLOPS、

総合最大主記憶容量: 10TB

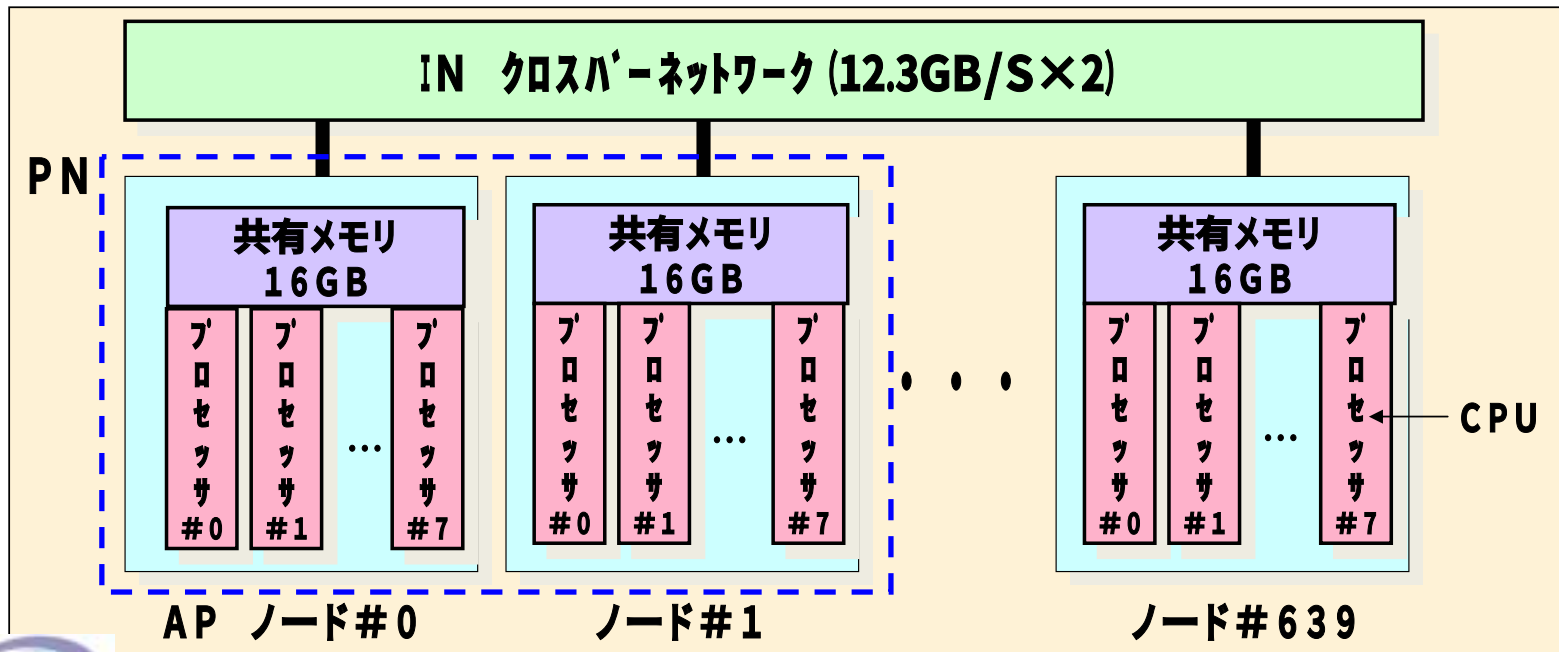
IN筐体 65台 (XCT筐体 1、XSW筐体 64)

データ転送性能 (ノードあたり): 12.3GB/秒*2、

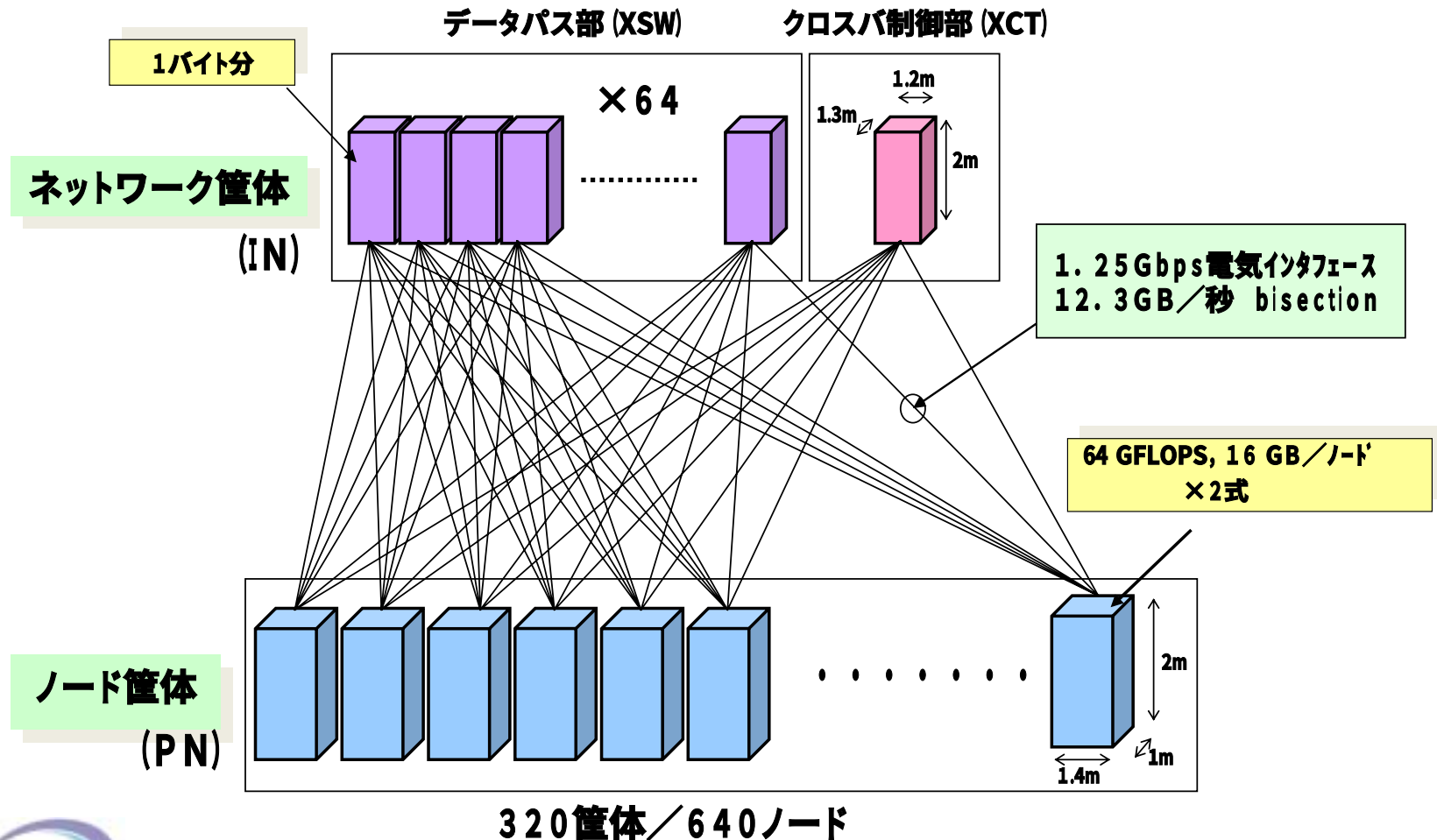
総合最大データ転送性能: 7.8TB/秒

本体系システム構成

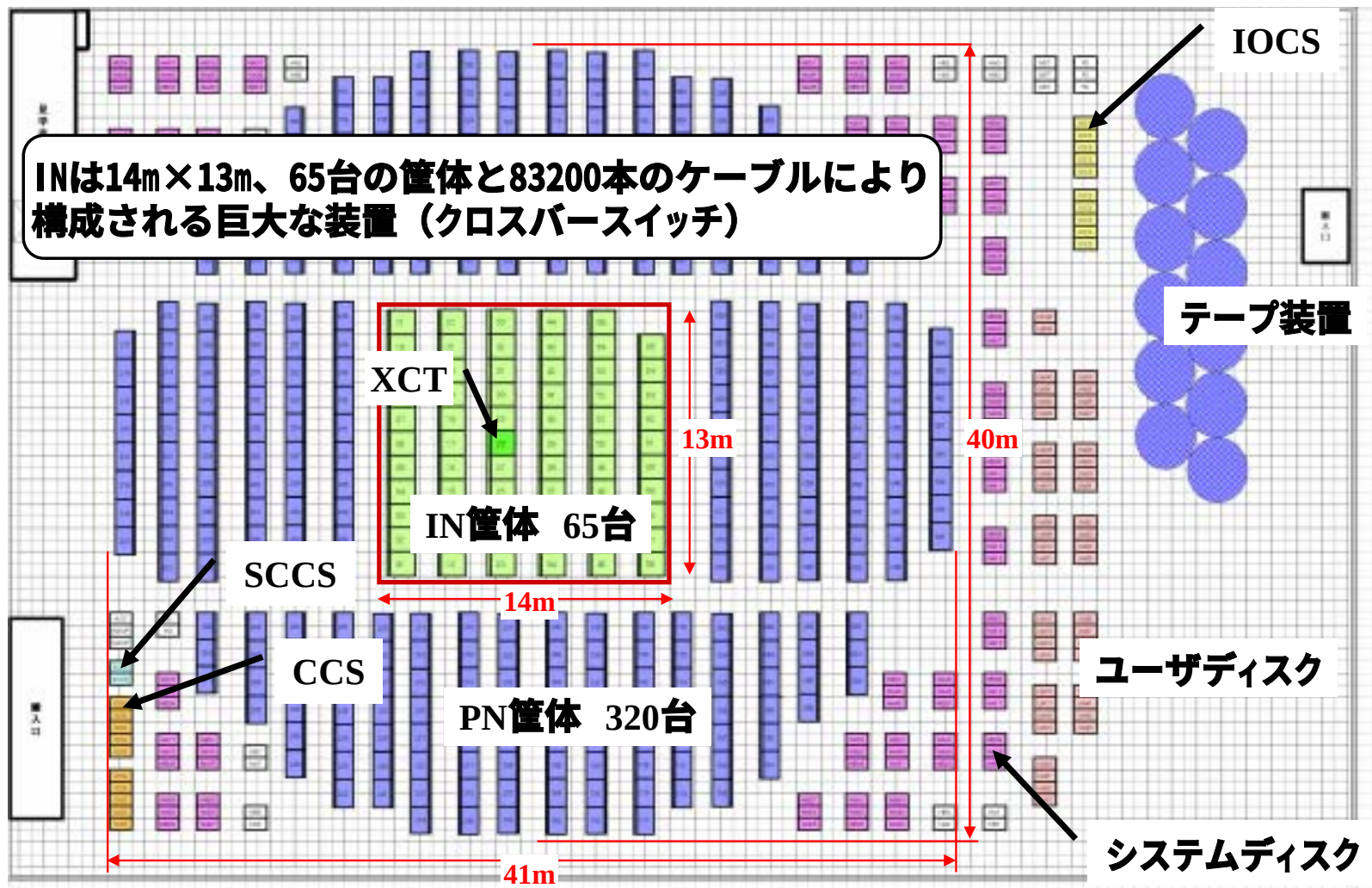
◆ ノード数	640	・ プロセッサ数	5,120
◆ 総ピーク性能	40TFLOPS	・ ピーク性能／プロセッサ	8GFLOPS
		・ ピーク性能／ノード	64GFLOPS
◆ 総メモリ容量	10TB	・ メモリ容量／ノード	16GB



本体系システム物理構成

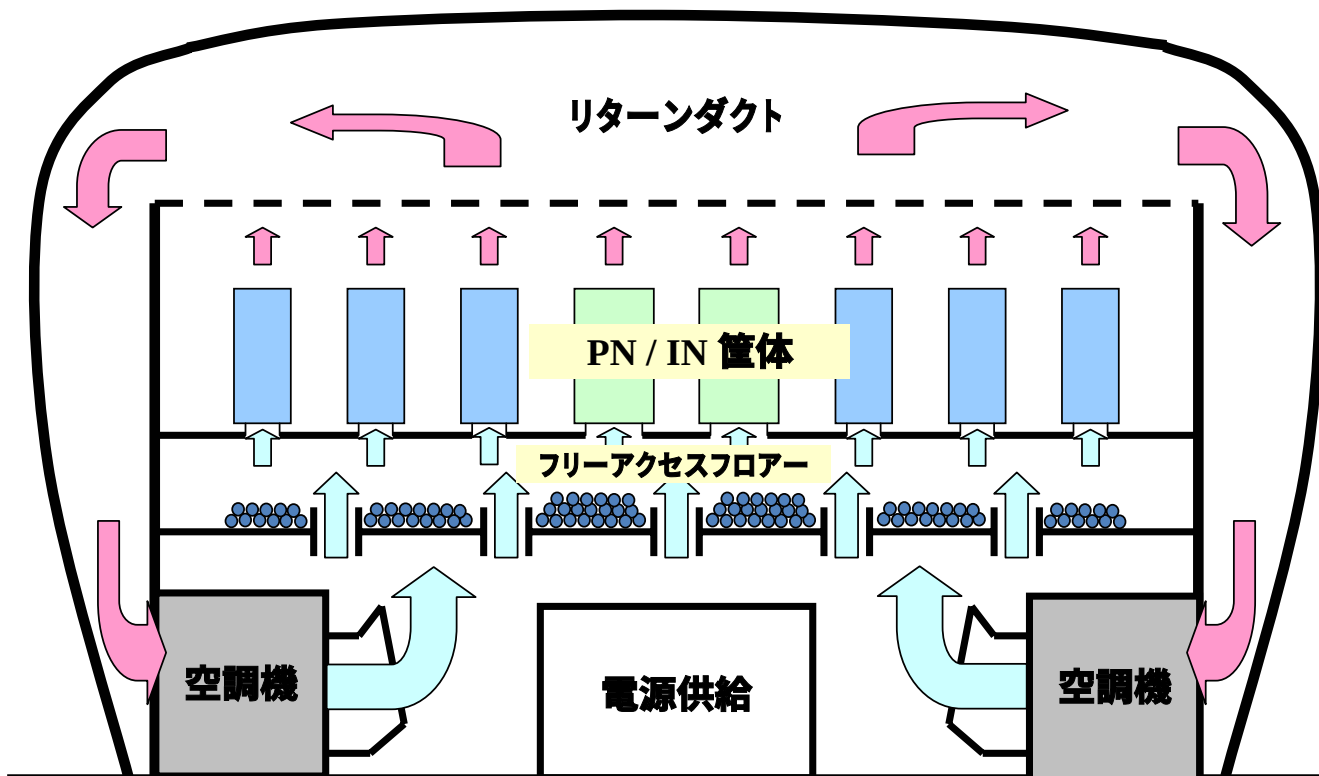


装置の配置



空調システム（階下式空調方式）

11



プログラム実行環境

12

● オペレーティングシステム

- UNIX (NEC SXシリーズ SUPER-UX)
- 科学技術計算用高速ファイルシステム
- 並列ファイルシステム (MPI_IO, HPF)

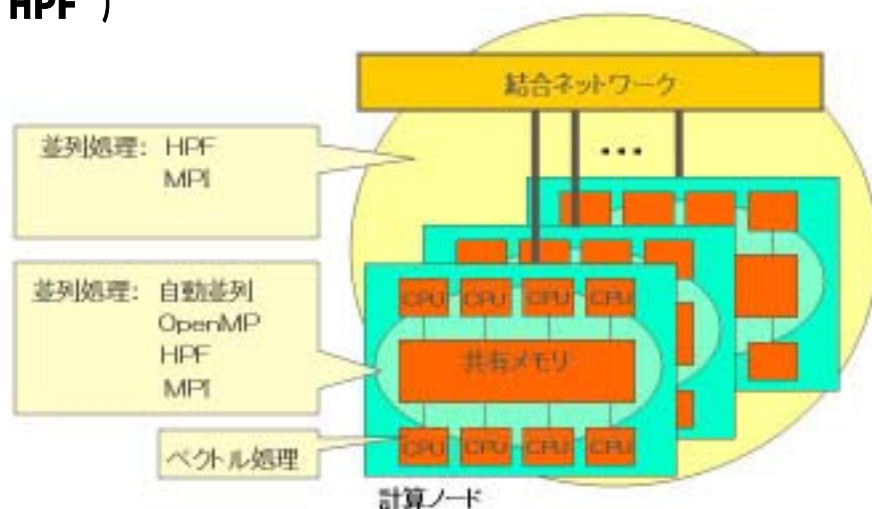
ベクトル処理と階層型並列処理

● プログラミング環境

- Fortran90, C, C++
- 自動ベクトル化, 自動並列化
- MPI2, OpenMP, HPF2
- Tuning tools
- Parallel debugger

● ジョブスケジューラ

- 独自のジョブ制御・リソースアロケーターを開発
- ジョブスケジューラ用 ジョブスクリプトの開発

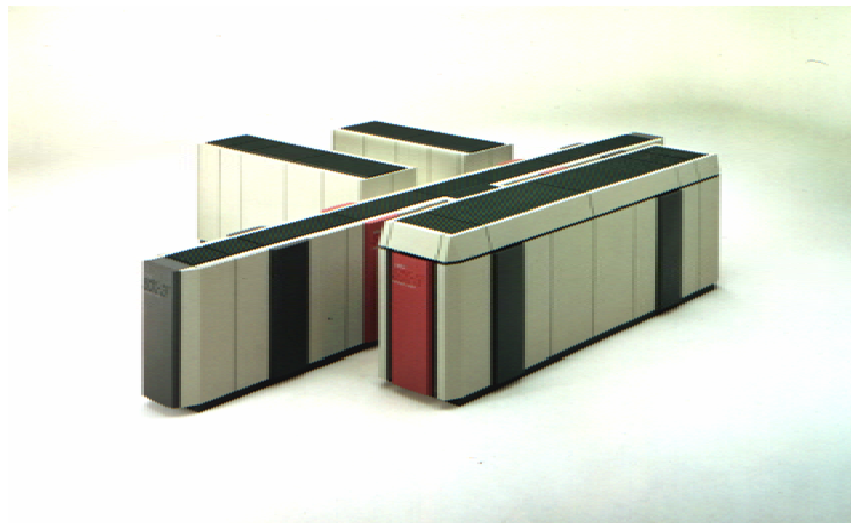


最近 似たような構造のシステムが...

10年間の技術発展

13

SX-3



22GFlops/4Cpu

1990

地球シミュレータ

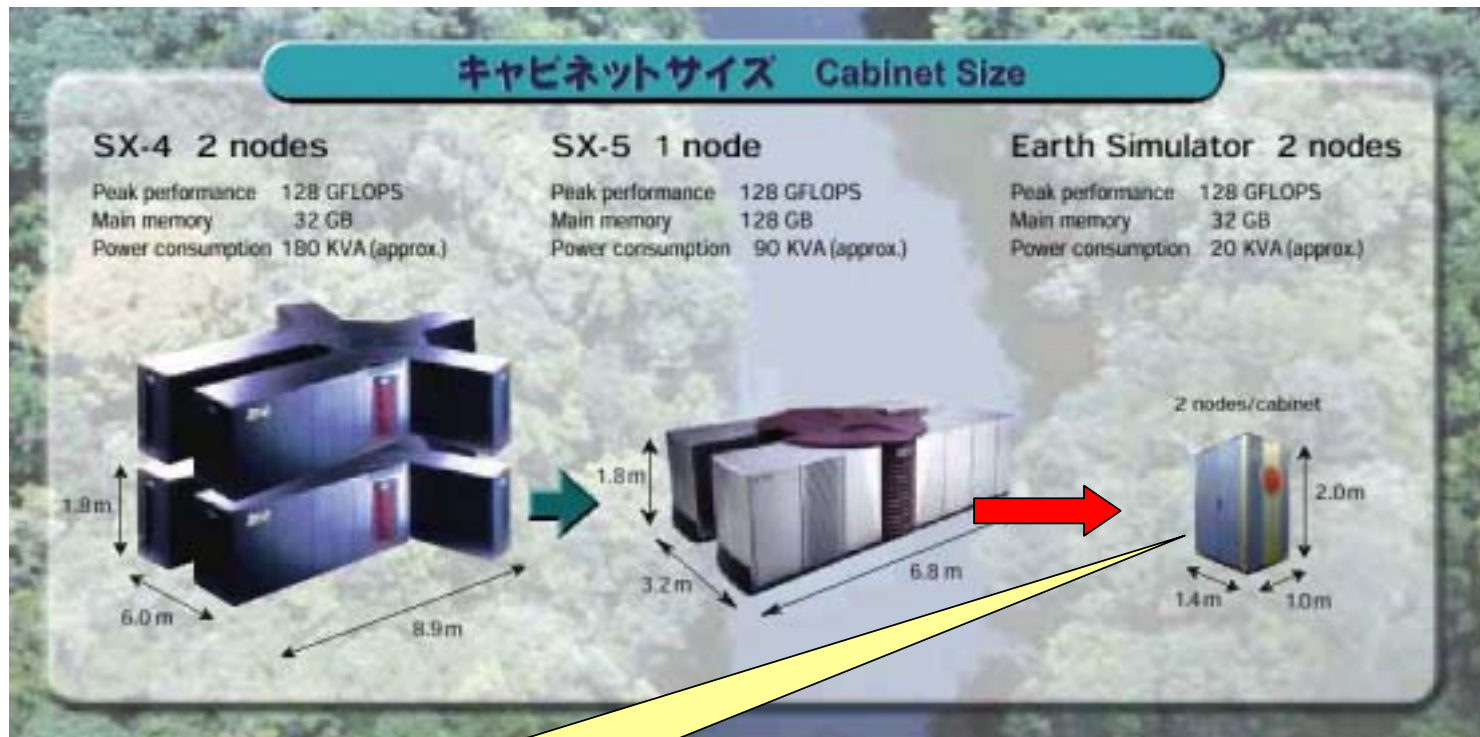


40TFlops/5120Cpu

2002

10年で1000倍！

PN筐体の比較 (128GFの筐体)



直径0.6mm 高速細径同軸ケーブル
小型低電圧大電流電源モジュール

640台のPNを設置するため、
PNの小型化が必須

PN筐体
PN
演算性能
記憶容量
総台数

2台
64GFLOPS×2
16GB×2
320台

CPU テクノロジーの比較



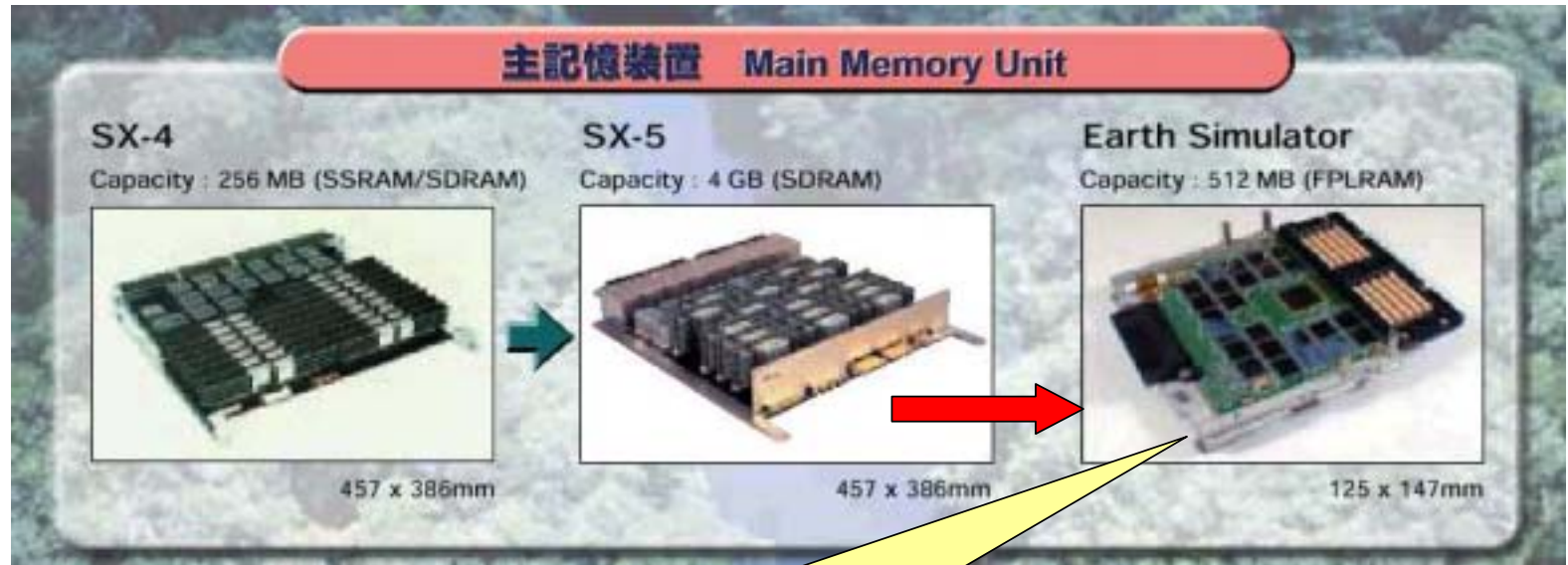
0.15 μ m CMOS-LSI (5700万Tr、銅配線)
 微細ビルドアップ基板 (最短配線幅 25 μ m)
 直径0.6mm 高速細径同軸ケーブル用コネクタ
 (最短ピン間隔 0.5mm)
 高効率ヒートシンク (ヒートパイプ使用)
 AP用D-Dコンバータ

プロセッサパッケージ

演算性能	8GFLOPS
動作周波数	500MHz
消費電力	140W
総使用数	5120枚 (8枚/PN)

640台のPNを設置するため、
PNの小型化が必須

メモリ テクノロジーの比較

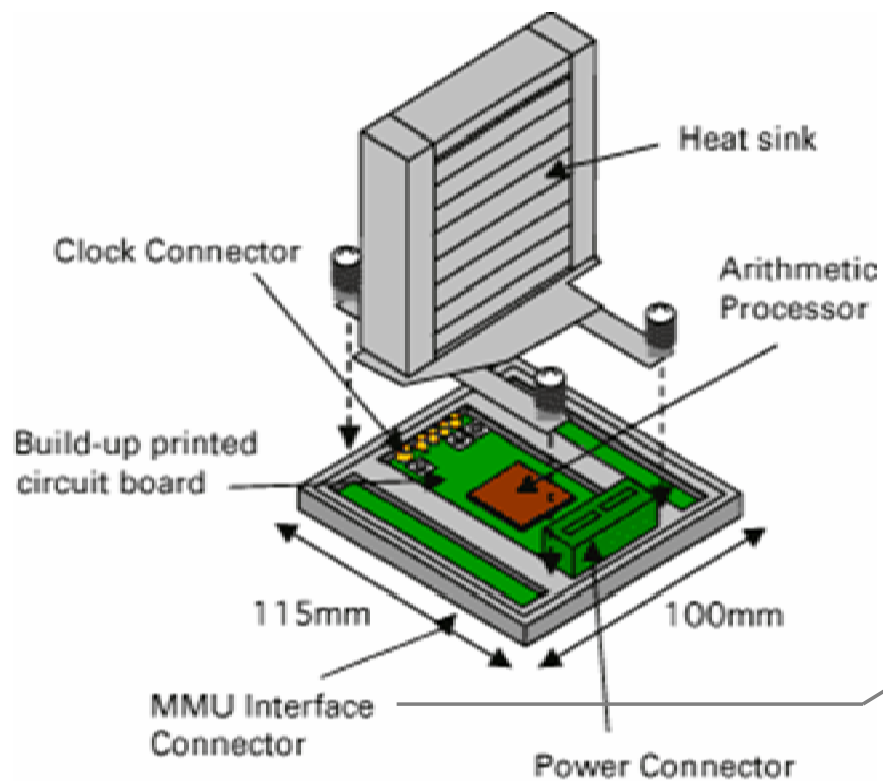


0.15um CMOS-LSI (銅配線)
 128Mb FPLRAM (8bank, 2Mword×8bit構成)
 微細ビルドアップ基板 (最短配線幅 25um)
 直径0.6mm 高速細径同軸ケーブル用コネクタ
 (最短ピン間隔 0.5mm)
 MM用D-Dコンバータ

主記憶パッケージ
 総使用数 20480枚
 (32枚/PN)
 記憶容量 512MB
 (16GB/PN)

640台のPNを設置するため、
PNの小型化が必須

CPUパッケージ



地球シミュレータのプロセッサLSI (AP)

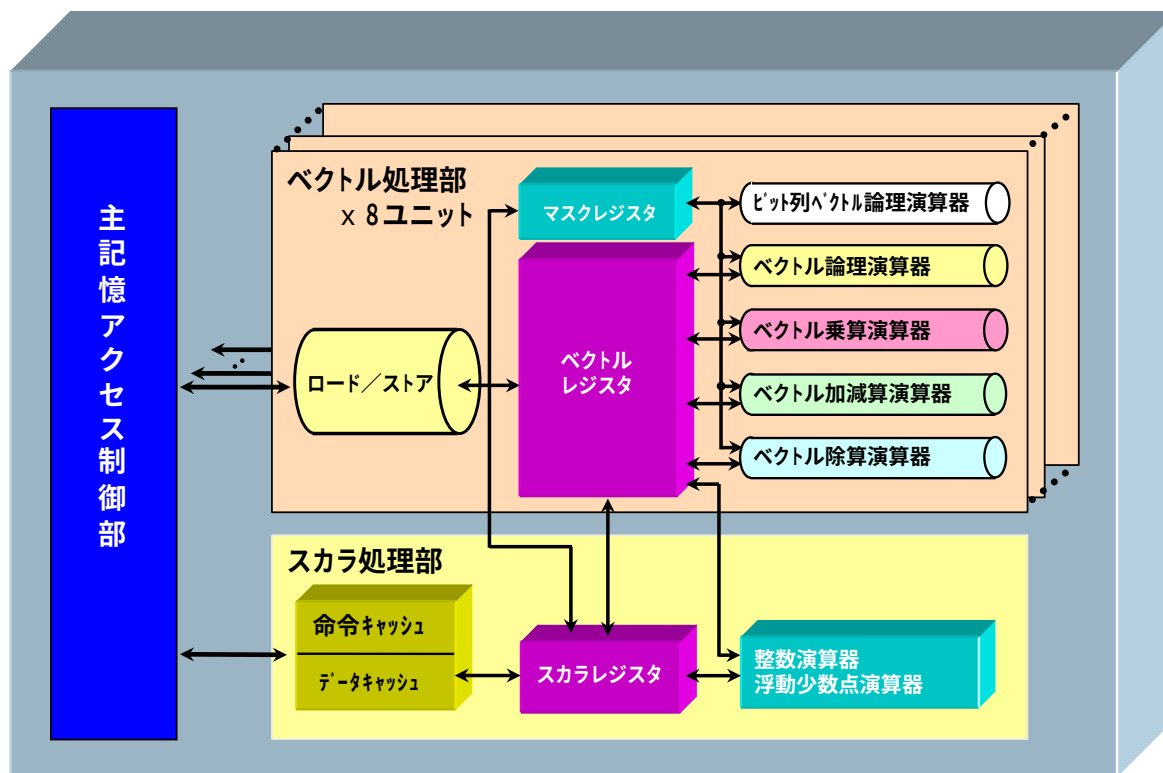
○ ベクトルユニット: 8セット

- ◆ 6種のベクトルパイプライン
- ◆ 256要素のベクトルレジスタ: 72個
- ◆ 256ビットのマスキレジスタ: 17個

○ 主記憶アクセス制御部

○ スカラユニット

- ◆ 4-ウェイ スーパースカラ
- ◆ 64KB 命令キャッシュ
- ◆ 64KB データキャッシュ
- ◆ 128個の汎用レジスタ

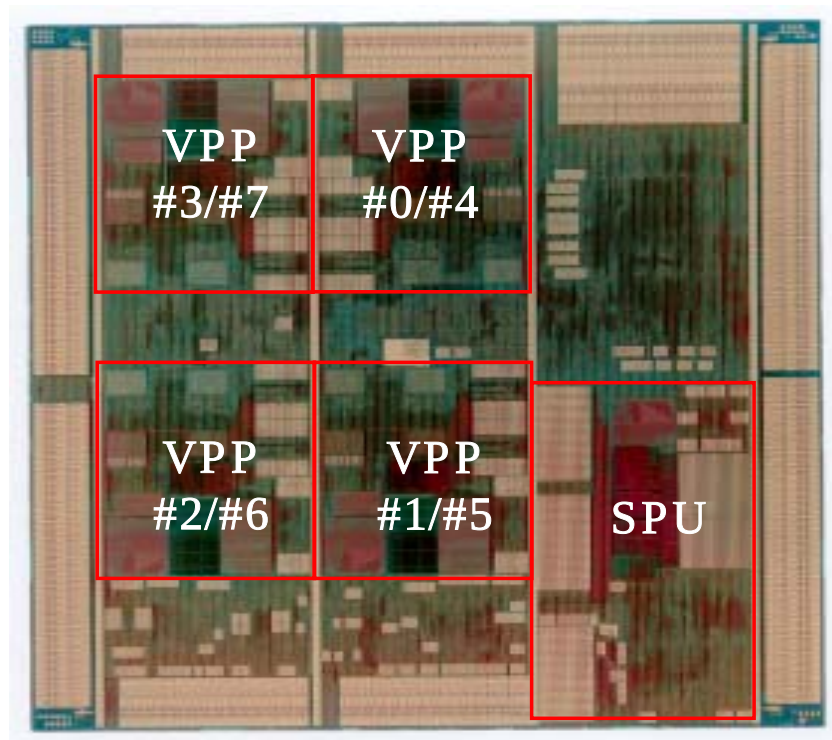


1チップLSI: 8Gflops

- ◆ 0.15 μ m CMOSテクノロジー + 銅配線
- ◆ 20.79mm x 20.79mm
- ◆ 5,700万トランジスタ
- ◆ 5185 ピン
- ◆ クロック周波数 500MHz
- ◆ 消費電力 140W

地球シミュレータのプロセッサLSI (AP)

19

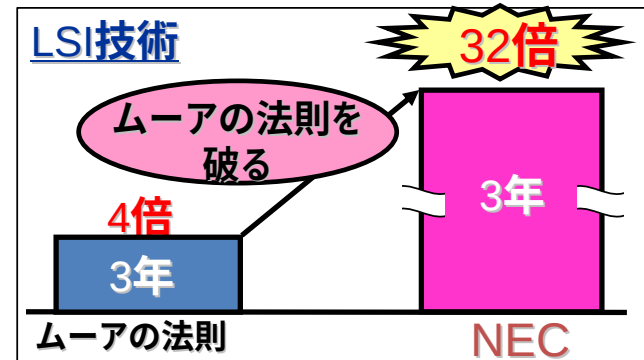


- 0.15 μ CMOS
- 8層銅配線
- 20.79mm * 20.79mm
- 60million トランジスタ
- 5185ピン (信号 & 電源)
- Clock Frequency : 500MHz (1GHz)
- Power Consumption: 140W (typ.)

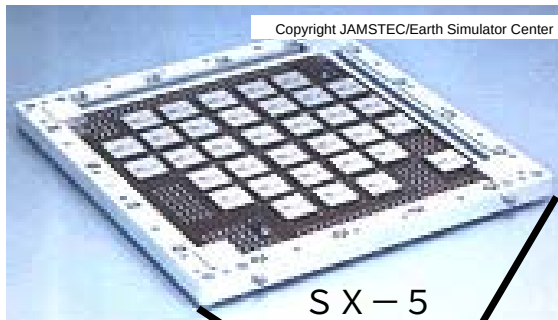
超高性能、超高密度実装実現のため、1Chip化に挑戦

1Chipベクトルプロセッサ

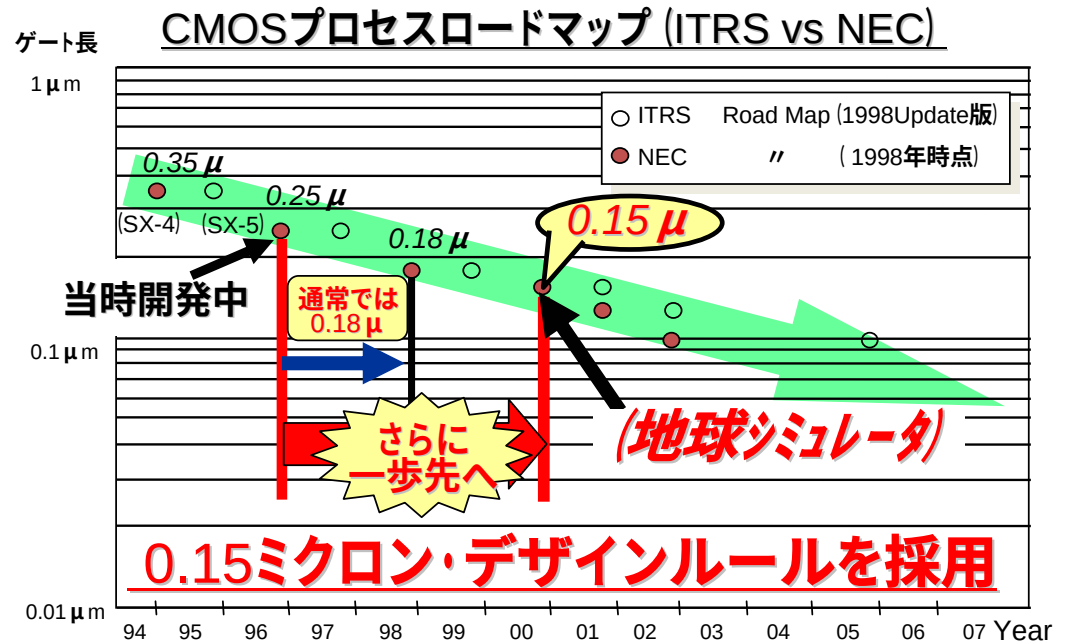
0.15 μ プロセスの先取りと、
Cu配線プロセスで高速化



従来32チップを1Chip化



地球シミュレータ

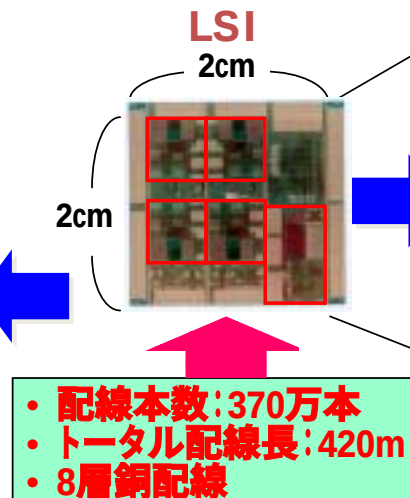


テクノロジーへの挑戦

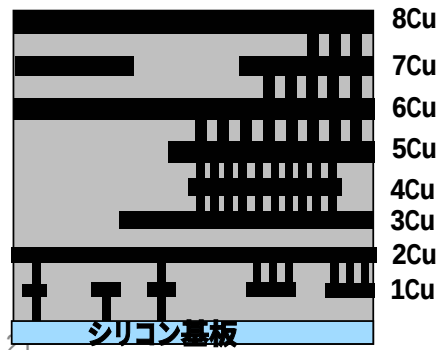
21

膨大な回路を小さなチップに詰め込む
ためのテクノロジー

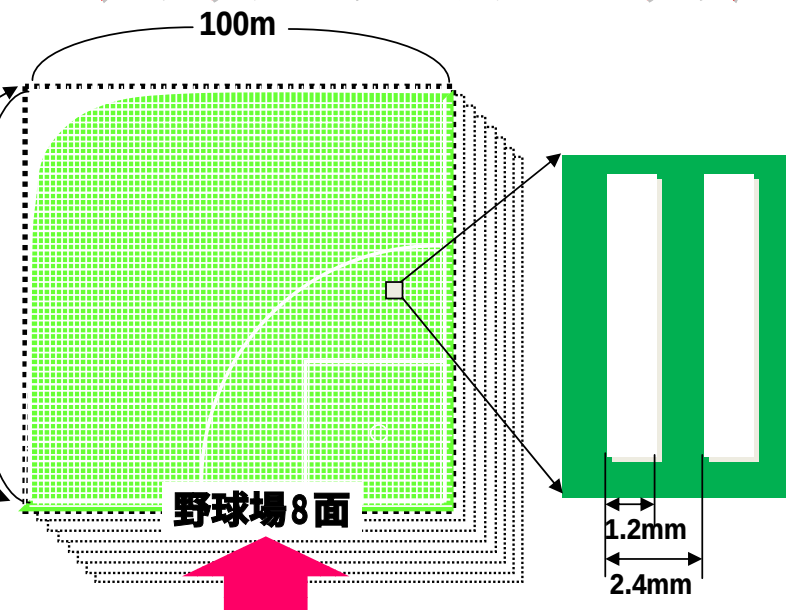
平均的日本人の
髪の毛の太さ



8層銅配線断面図



(1チップベクトルプロセッサ)



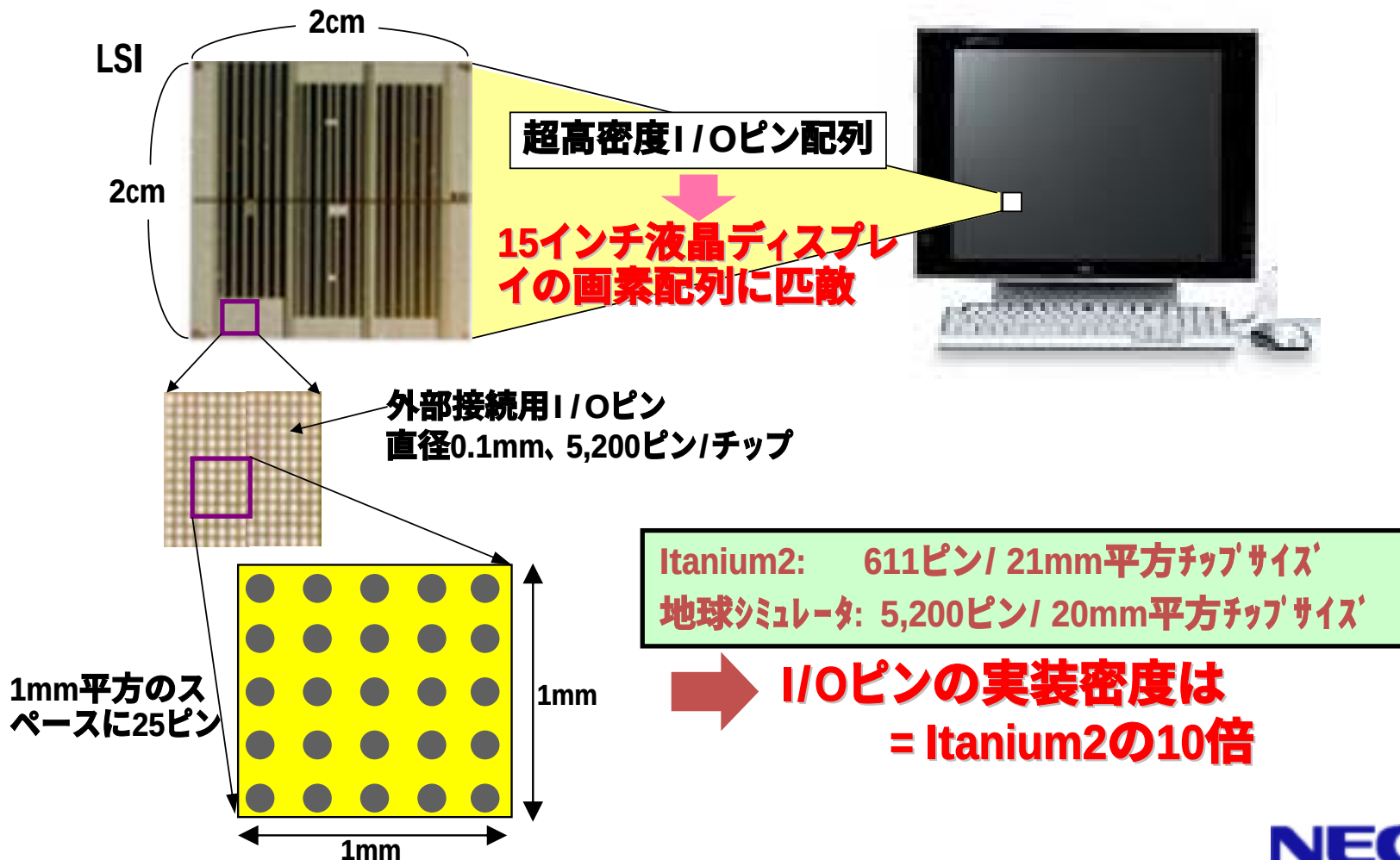
1.2mm幅、2.4mmピッチの配線を
野球場8面に敷き詰めるのと等価
配線本数370万本→総配線長2,100km!

- CADによる配線では数千本の引き残しが発生
- 引き残された配線は人手で接続
(天文学的な配線ルート組み合わせの中から、
熟練技術者が最適なルートを選択)

テクノロジーへの挑戦 (LSI I/Oピン)

22

高いメモリバンド幅を実現するために、1,000ビット以上のデータを並列に転送
→多数のLSI I/Oピンを小さなチップ上に実装するためのテクノロジー



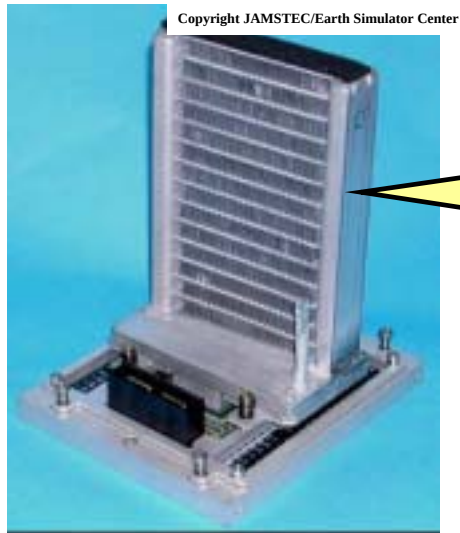
地球シミュレータの冷却技術 (空冷)

消費電力が140WのLSIを冷却する技術

熱発生密度は調理器具のホットプレートの20倍

→ 従来技術による冷却は実現不可と思われた。

→ 色々な分野の冷却技術を調査した結果、
自動車業界の技術を採用した。(異業種連携)



Copyright JAMSTEC/Earth Simulator Center

沸騰冷却器を採用することにより体積を1/5にできた。

→ 高密度実装が可能となった。

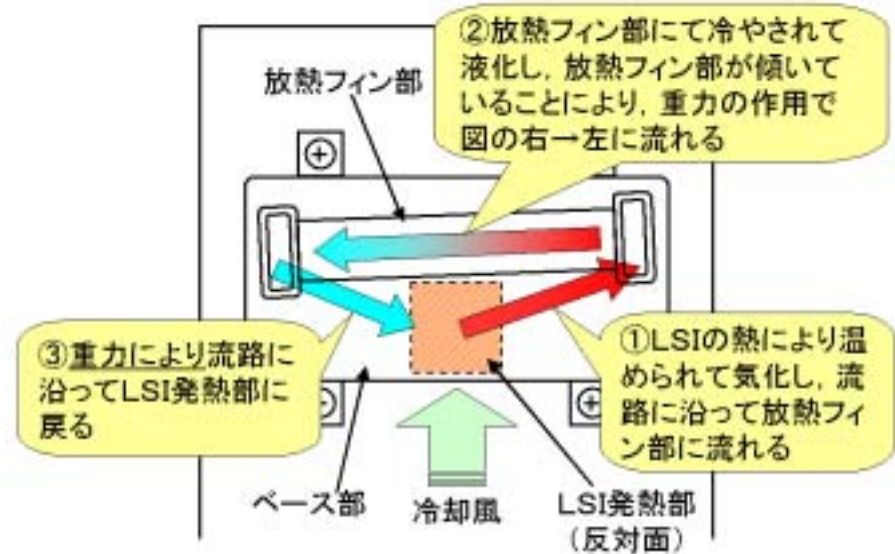
最大170WのLSIまで冷却が可能

沸騰冷却器を搭載したCPUパッケージ

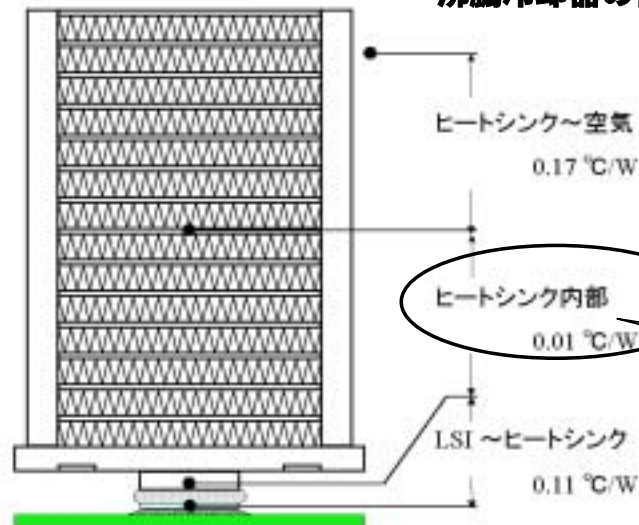
冷却技術



沸騰冷却器を取付けた
CPUパッケージ



沸騰冷却器の作動原理 (上から見た図)



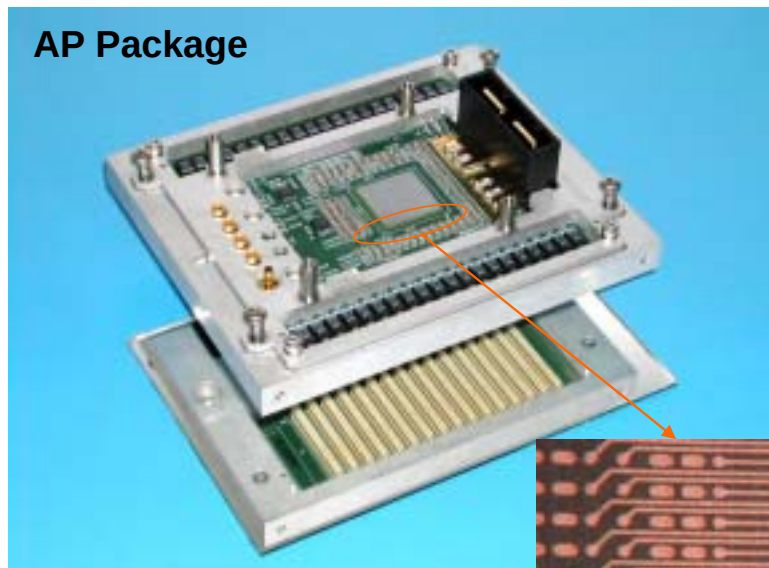
熱抵抗の内訳

沸騰冷却器の採用によりこ
の部分が大きく改善：
従来ヒートシンクの1/5～1/10

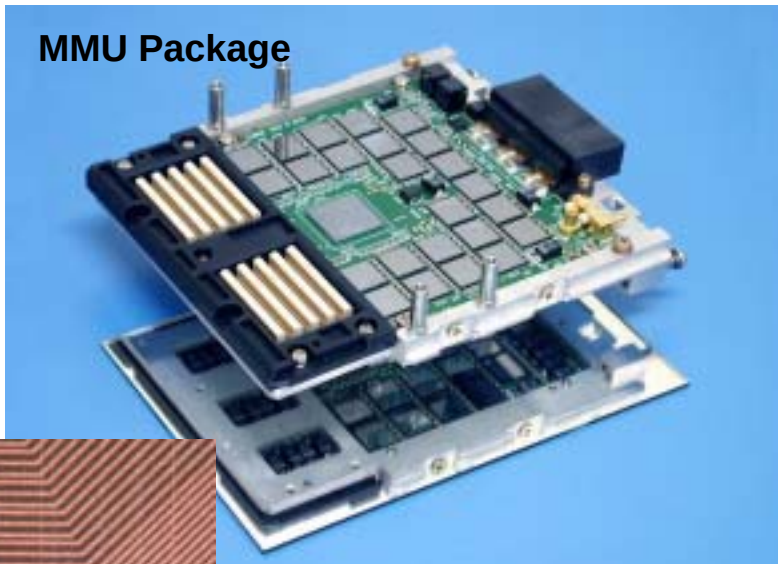
ビルドアップ基板

25

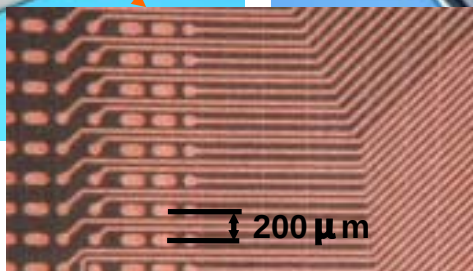
AP Package



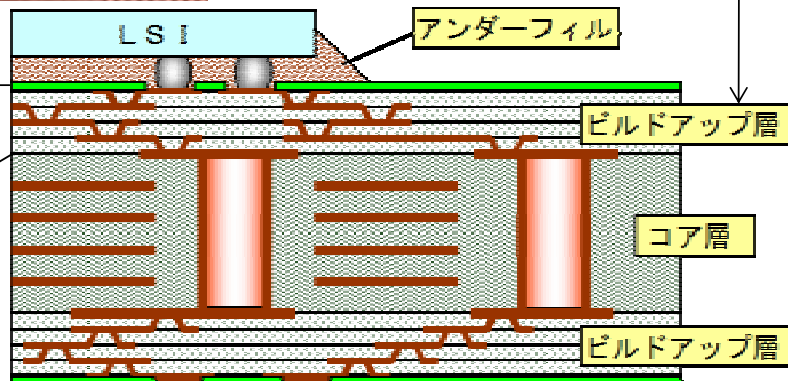
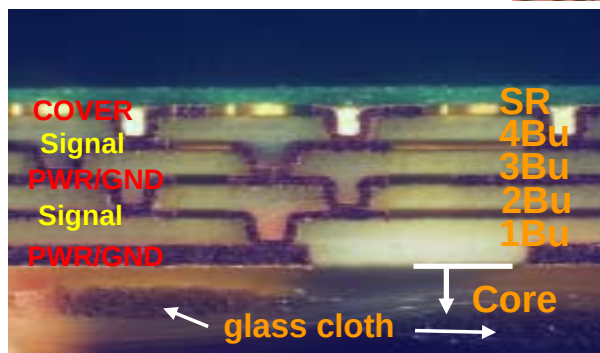
MMU Package



LSI近傍配線
(内層)



プリント板にLSI製造並みの技術を導入



ベアチップ実装 (LSI実装技術)



APモジュールの外観

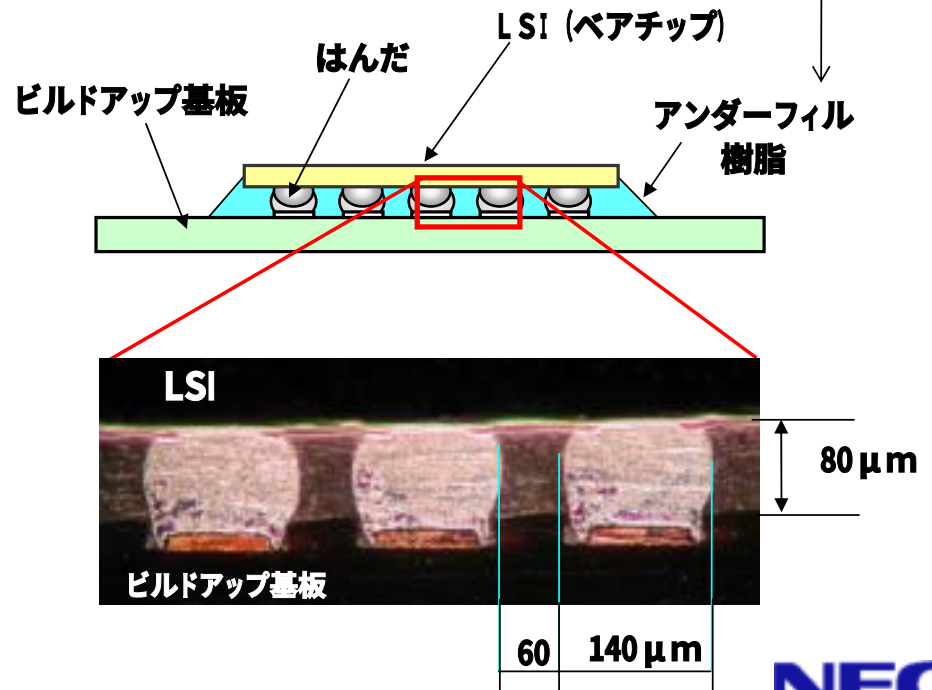
ボイド



ボイド写真

基板寸法	100×115mm
LSI端子数	5,185
LSI端子ピッチ	200 μ m
LSI寸法	20.79mm□

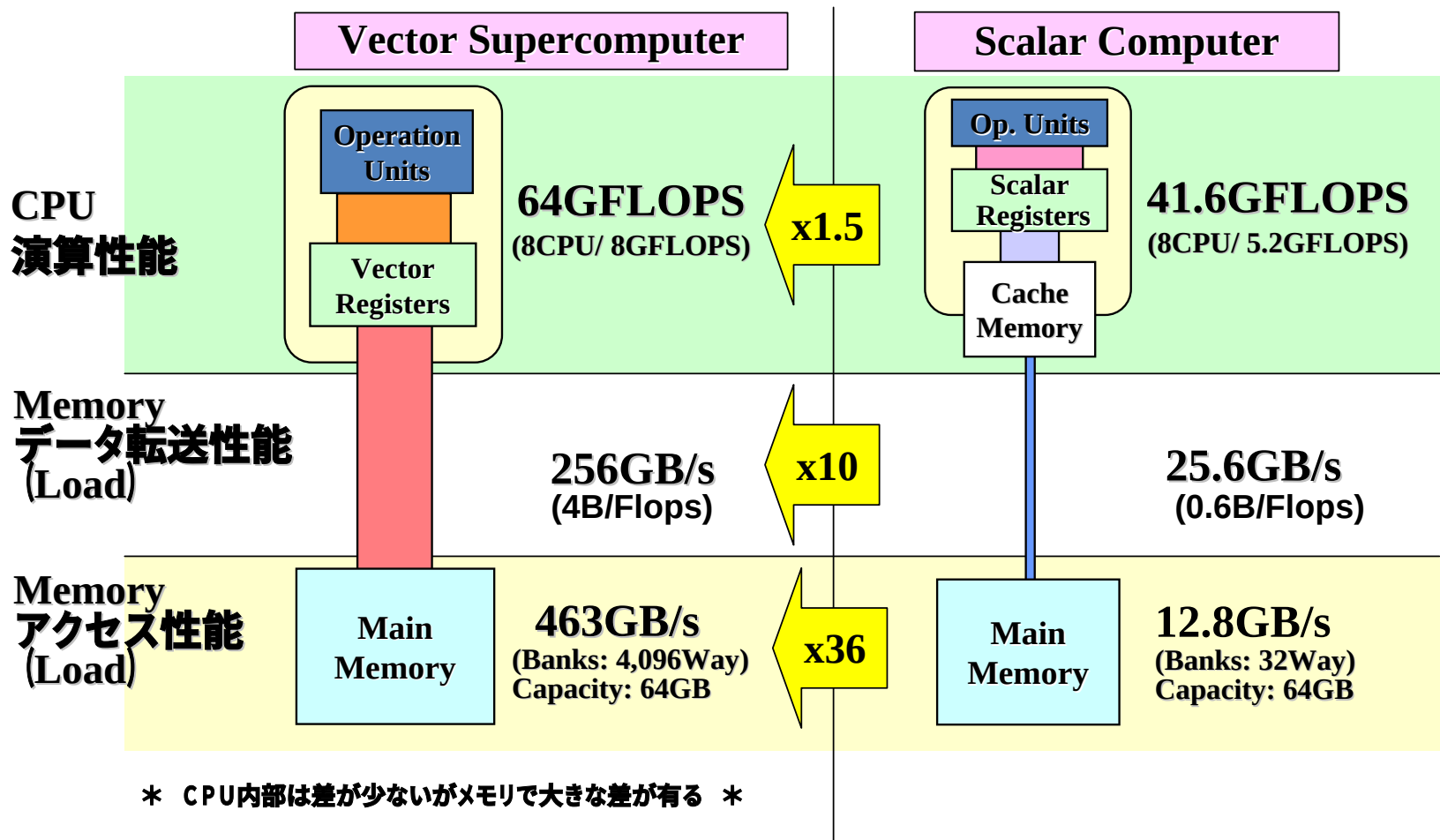
膨張率の異なる部品の接続の為に



地球シミュレータの伝送技術

27

(CPU-Memory間データ転送)



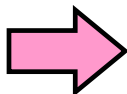
CPU・MMU間ケーブル

28

CPU=MMU間接続ケーブル

(1本の太さはボールペンの芯並み)

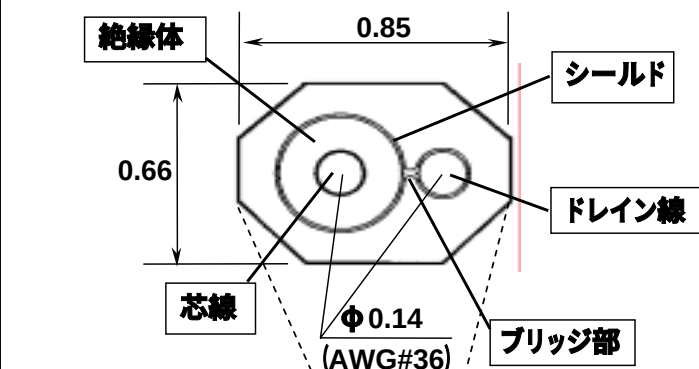
超高速接続には、
超極細同軸ケーブルが必須



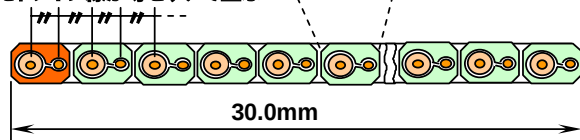
医療機器業界の製品を応用
(異業種連携)

- ・ 布線BOX内には 70cm のケーブルが 約20,000本 (14km/ノード)
- ・ システム全体で 9000km のケーブル
(東京ーフランクフルト間に相当)

ケーブル設計形状 (断面)



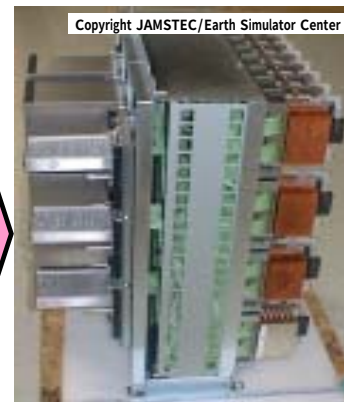
芯線とドレイン線が等ピッチで並ぶ



30芯フラットイメージ



配線作業工程
(同軸単芯約2万本)



配線完了
(カードを取付けた状態)

10回以上、試作・評価を
繰り返し、完成

地球シミュレータの伝送技術 (ノード間ケーブル)

**83,200 本のノード間ケーブルを使用。
床下に積み上げられたケーブルの最高部は床から40cm**

- ・余長が発生しないように40種類の長さのケーブルを使用 (10m~40m/cable)
- ・ケーブルの積み上げ高と冷気の移動空間確保、および消防法の制約から床上げ高は1.7m

**ケーブルの総延長
2,400km**



【ケーブル布設】

- ・プリント基板のパターン設計ツールを改造し、ケーブル布設ルートを決定
- ・NEC府中事業場の体育館を使用し、ケーブル布設作業の事前評価を実施
- ・一日に1,200本のケーブルを布設
23秒ごとに1本布設

(ケーブル布設作業に3.5 カ月掛け、
ピーク時70 名が作業)

ノード間ケーブル敷設 = 事前評価 =

30

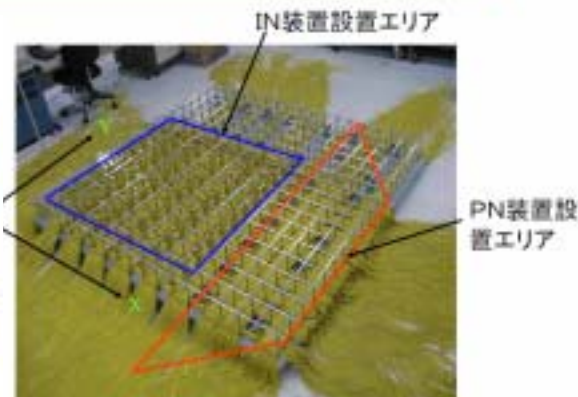
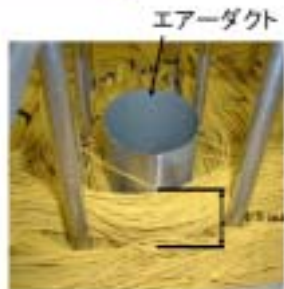
ケーブル自重圧力による評価

縮尺モデル(1/10スケール)敷設評価全景

縮尺モデルの内部詳細

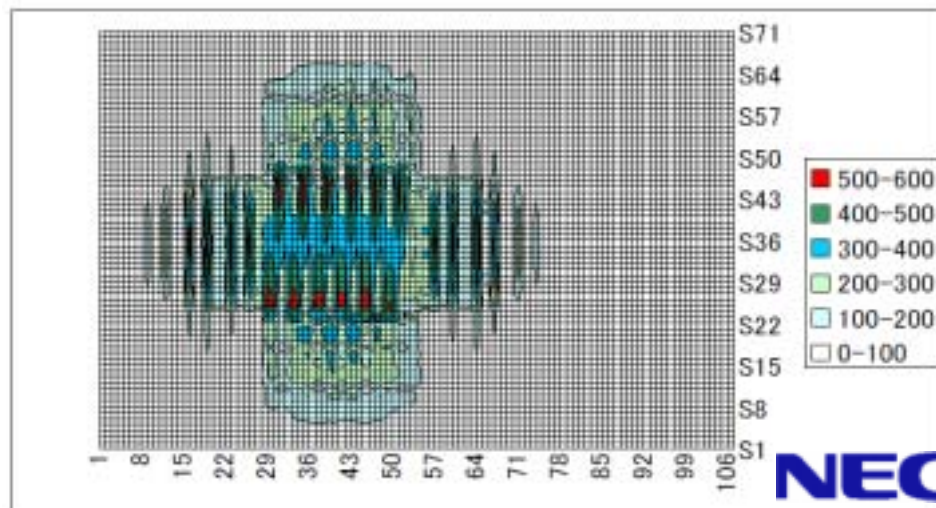


PN-Aエリア床下イメージ



床下ケーブル敷設高さシミュレーション結果

府中体育館でのケーブル敷設訓練



ケーブル敷設の事前評価 (1)

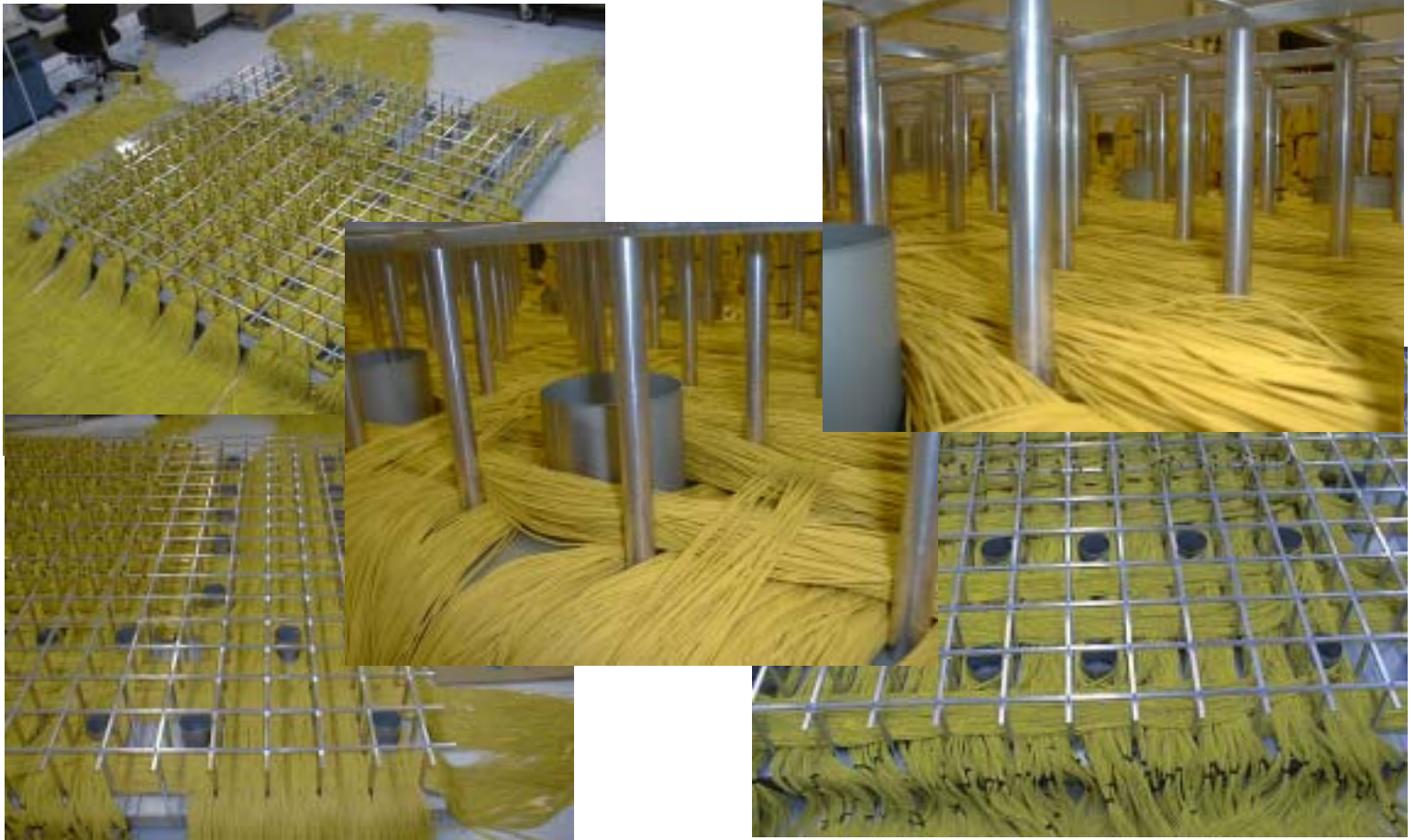
31

同じ太さのケーブルを使用したケーブル積上げ高の評価



ケーブル敷設の事前評価 (2)

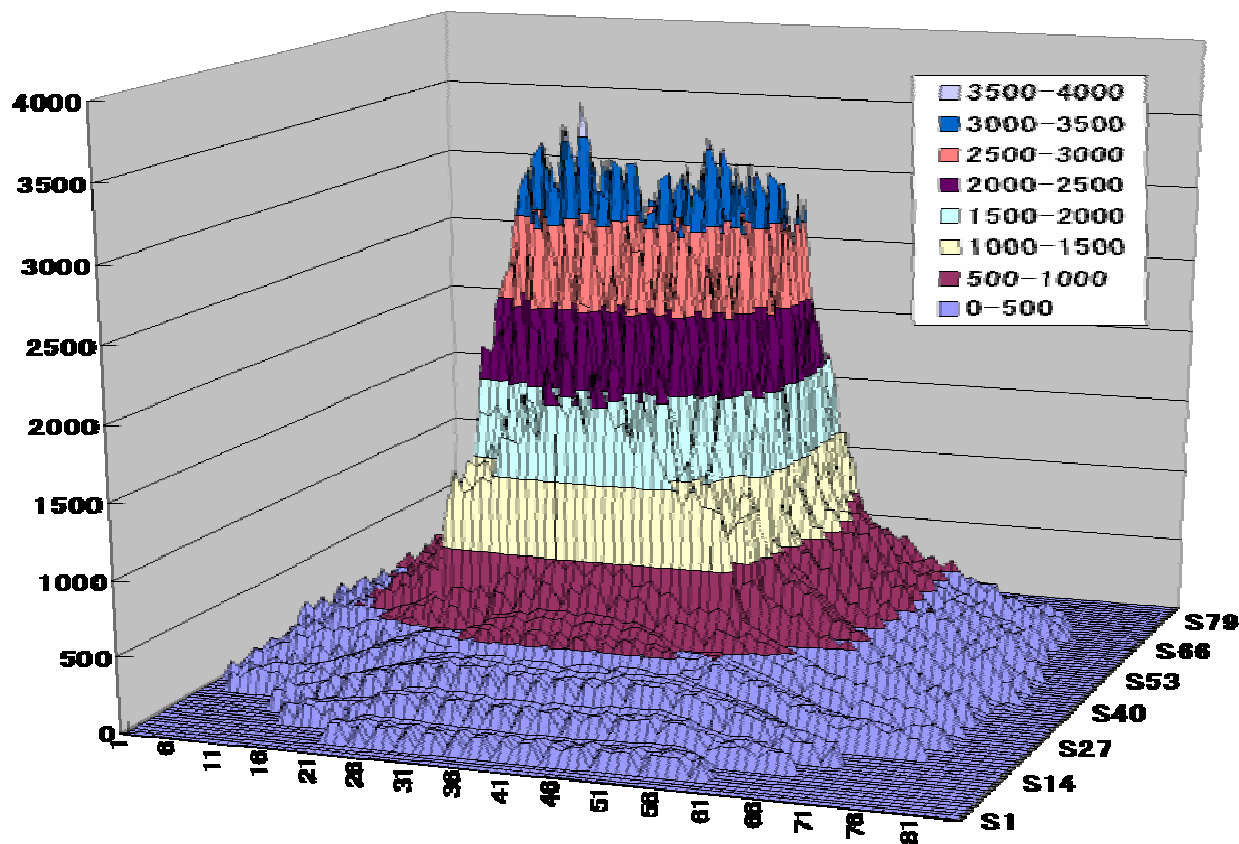
1/10スケールモデルによるケーブル敷設順序とケーブル積上高の評価



ケーブル敷設の事前評価 (3)

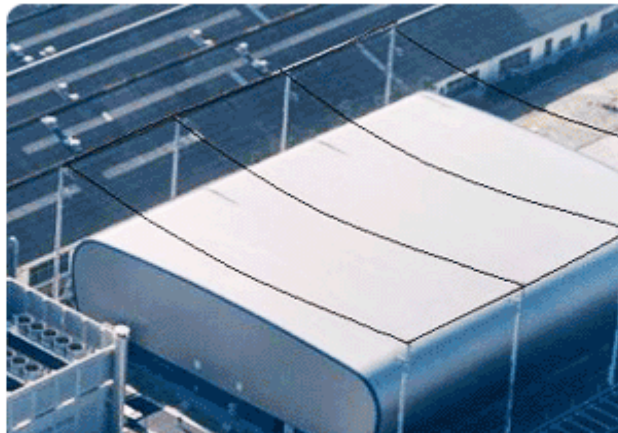
33

基板のパターン生成プログラムを改造したケーブル敷設シミュレータを使用した
ケーブル積上高の評価



シミュレータ棟の避雷、照明設備、免震

24m pole x 8 pcs



避雷システム (架空地線方式)

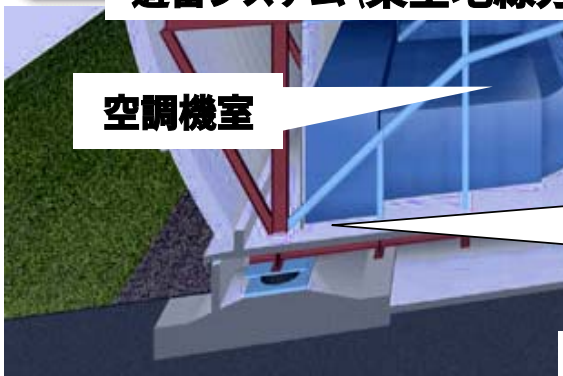


ライトチューブ

照明システム (ライトガイド方式)



フリーアクセス
(高さ 1.5m)



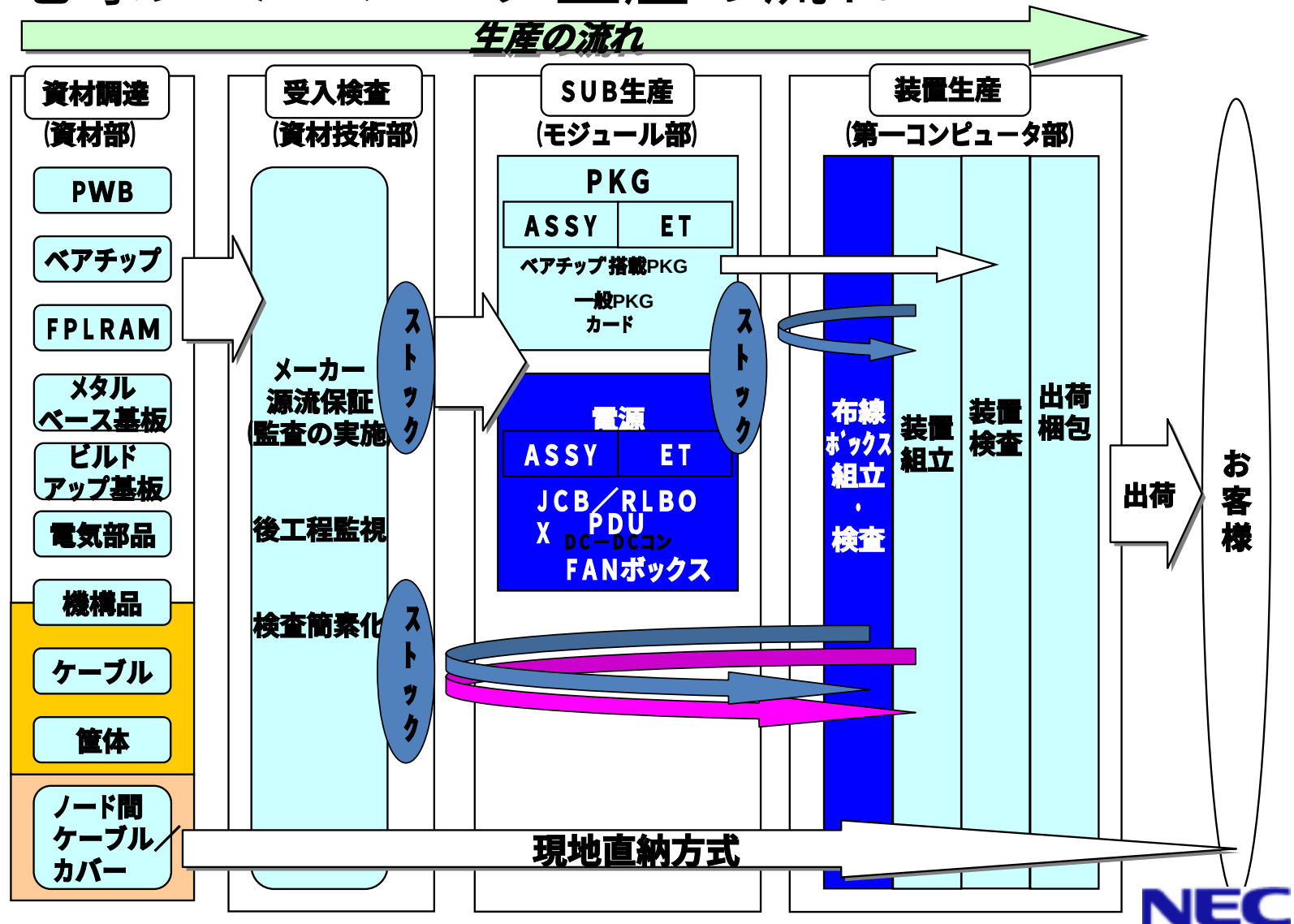
空調機室



免震システム
(11個の積層ゴムアイソレータ使用)

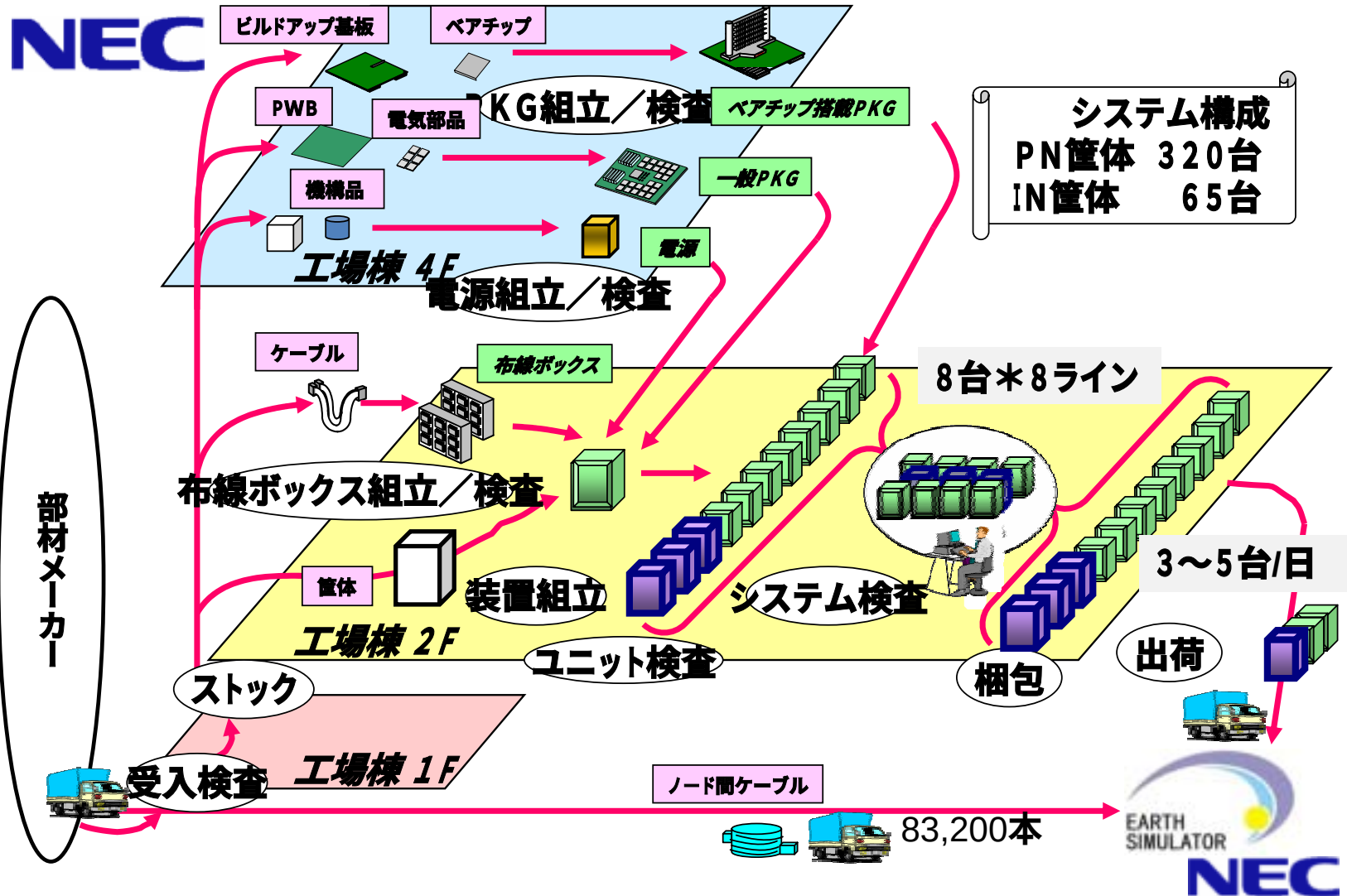


地球シミュレータ生産の流れ



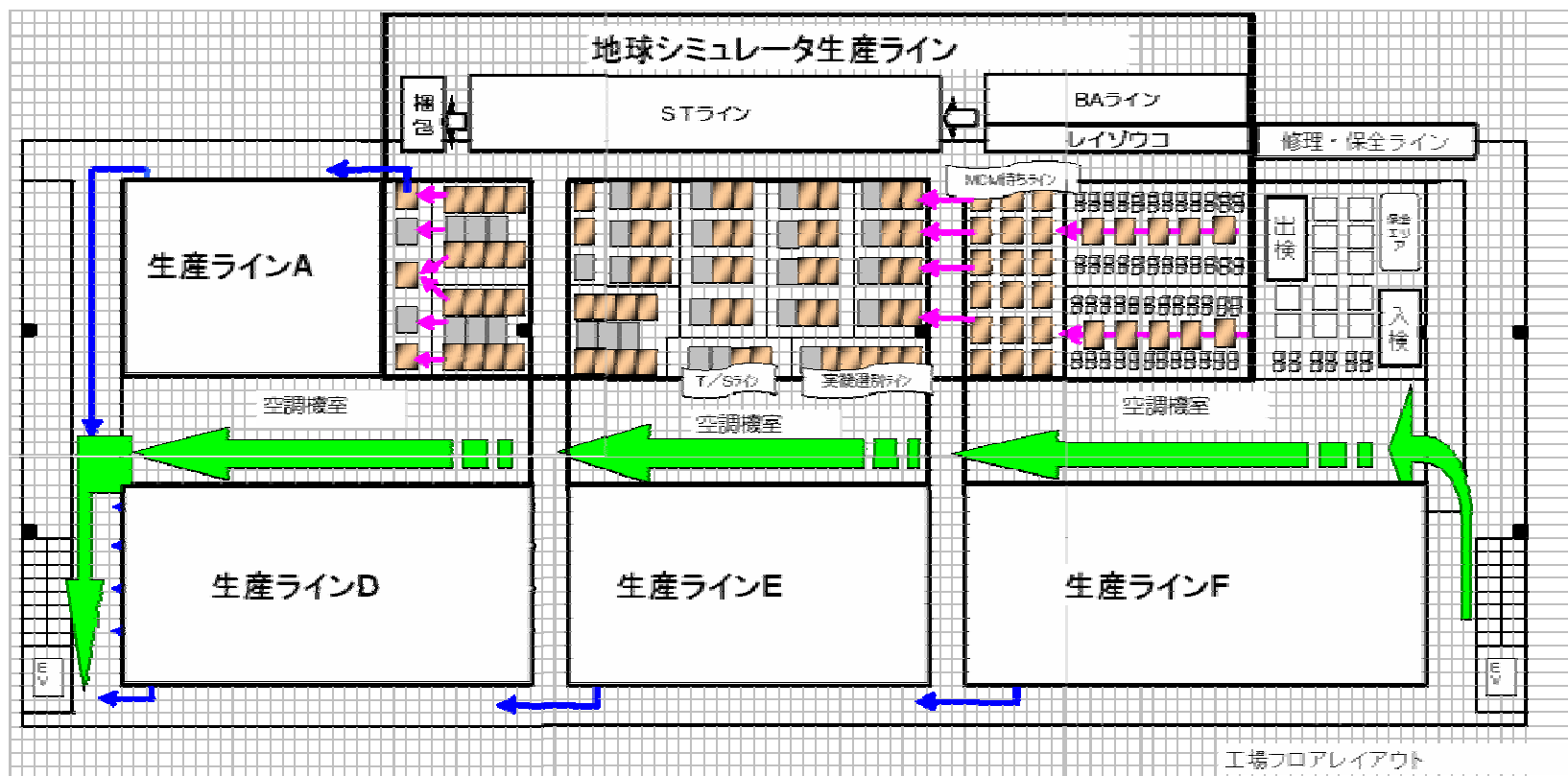
地球シミュレータ生産の流れ

36



NEC甲府生産ラインイメージ (工場フロアレイアウト)

37



ノード間ケーブルの敷設

ノード間ケーブル敷設作業
(2001. 2～2001. 5)



ハーネス治具 (PN時)



ノード間ケーブルの敷設



ケーブル本数： 83,200本
一本の長さ： 10～40m
1m単位で、40種使用
総ケーブル長： 約2400km
ケーブルの総重量： 140t
敷設に要した月数： 約3.3ヶ月
1日当たりの敷設本数： 1280本
ケーブル積上高： 最高部 39cm



ハーネス治具 (IN時)



ノード間ケーブルの敷設



計算ノード・結合ネットワーク間ケーブル敷設完了 (2001. 5)



搬入設置

41

搬入用仮設やぐら



搬入設置



高実効性能プログラムによるフリッカ現象の発生

- E S稼働開始後2002年5月～7月にかけてE S研究棟の照明がチラつき、人によっては気分が悪くなる現象が発生
- 乱流計算、AFES, LINPACKなど実効性能の高いジョブを流すと発生するらしい。→調査開始

E S計算中の瞬間的な電力変動は、1370KWあり、200Vラインでは、2.33Vの変動になっている。

フリッカ電圧 ΔV_{10} は、0.23となり、備え付けの安定器と蛍光管の組み合わせで発生することが確認

結局、余裕をみて、安定器、蛍光管の両方をHF式照明器具に全数交換

→2003年3月全面解決

まとめ

- 三好先生は、国産ベンダの技術力維持・向上のために、国策としてのスーパーコンピュータ開発を如何にすべきか考えていた。

各社技術者と個別に、打ち合わせを繰り返し、自分なりの確固たるイメージを作り上げていった。この過程でベンダ技術者の受けた影響は大きい。

また、ベンダの内部事情に詳しく、スーパーコンピュータの受注・失注がその先どのような影響を与えるか大変気にしていた。

技術者特に最適化を行うSEの行く末となる受け皿をどうすれば出来るかを真剣に考えていた。

- 地球シミュレータは、新規開発部分が多く、リスク回避のため、結果的には、競合ベンダからの資材調達もあり、オールジャパン製品といっても過言ではない。

- ◎地球シミュレータ性能についての三好先生の予想(2000年頃)

・ピーク性能: 1, 2年で破られるだろう	→ 2年半後 BlueGene/L
・LINPACK : 2, 3年かな	→ 2年半後 BlueGene/L
・アプリ : 5年は持つ、いやもっとかも。 (大気大循環)	→ 6年半後 Jaguar (XT5)

ご静聴ありがとうございました

本資料の作成にあたり、多大な協力を頂いた
古井利幸氏に、感謝申し上げます。

