2012　PCクラスタワークショップ in 北海道

日立のテクニカルコンピューティングへの取り組み

2012/3/9

株式会社　日立製作所
中央研究所
清水 正明

HITACHI
Inspire the Next

目　次

1 日立テクニカルサーバラインアップ

・SR16000
・HA8000

1-1 日立テクニカルサーバ：History & Future

S-810/S-820は国立科学博物館より
「重要科学技術史資料」第00042号
として認定されました。

SR16000
Power7ﾌﾟﾛｾｯｻ採用、
複数ﾓﾃﾞﾙをﾗｲﾝﾅｯﾌﾟ
(ﾓﾃﾞﾙ XM1,M1,VM1)

ﾍﾞｸﾄﾙ・ｽｶﾗ融合型

Next Generation HPC Servers

SR11000
ﾉｰﾄﾞ当り理論性能
147.2GFlops(ﾓﾃﾞﾙK2)

CP-PACA(SR2201)は、
1996年の
Top500で1位を獲得

ｽｶﾗ並列型

SR8000
世界初のﾍﾞｸﾄﾙ・ｽｶﾗ
融合型HPCﾏｼﾝ

CP-PACS(SR2201)
商用ﾍﾞｰｽ日本初の
分散ﾒﾓﾘ型並列機

HA8000-tc/HT225
(Opteron Base)

HA8000-tc/RS425
(Opteron Base)

ﾍﾞｸﾄﾙ型

S-3800
単体CPUの理論性能
8GFlops達成 (世界初)

クラスタシステム向け
高性能サーバ

HA8000/RS210
(Xeon Base)

S-820
単体CPUの理論性能
3GFlops達成

PCサーバ

S-810
日本国産初の
ﾍﾞｸﾄﾙ型ｽﾊﾟｺﾝ

HA8000Series(Xeon Base)

ﾊｲｴﾝﾄﾞのItanium(日立独自設計)ｻｰﾊﾞ
含め幅広いﾗｲﾝﾅｯﾌﾟ

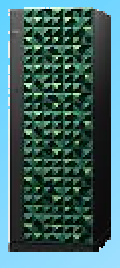
1980 1985 1990 1995 2000 2005 2010 2015

1-2 日立テクニカルサーバ　ラインアップ

スカラSMPからPCクラスタまでラインアップ拡充

スカラSMPクラスタ (SR16000シリーズ)

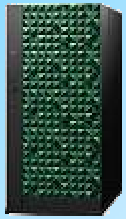
大規模SMP モデルVM1



POWER7 ～256way

最大ノード構成
- 8.2 TFLOPS
- 8 TBメモリ

高効率/高集積 モデルM1



POWER7 32way

最大ノード構成
- 980 GFLOPS
- 256 GBメモリ

システム構成
- 32～512ノード
- 最大500 TFLOPS

エントリ モデルXM1

POWER7 32way

最大ノード構成
- 844 GFLOPS
- 256 GBメモリ

システム構成
- 1～512ノード
- 最大432 TFLOPS

PCクラスタ (HA8000シリーズ)

HA8000-tc/HT225

AMD/Opteron

最大ノード構成
- 294 GFLOPS
- 64 GBメモリ

HA8000/RS210

Intel/Xeon

最大ノード構成
- 146 GFLOPS
- 192 GBメモリ

次世代Xeonプロセッサにも対応予定

InfiniBand QDRｻﾎﾟｰﾄ
- Fat-Tree
- 3D-Torus

1-3 ＳＲ１６０００の紹介

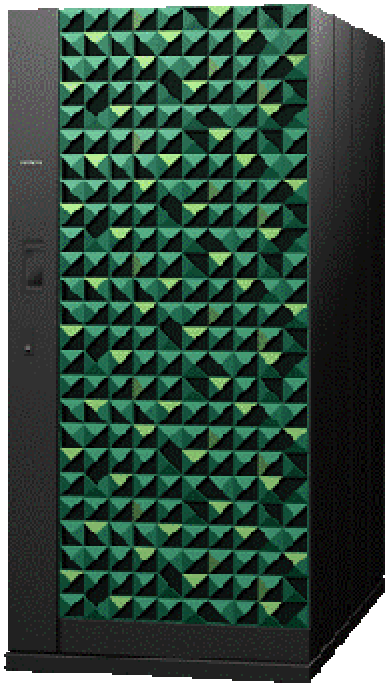
SRシリーズの特長を継承・強化させ、最先端H/W
テクノロジにより高性能・低消費電力を両立させる

◆ 実績と将来性を見据えたシステム・アーキテクチャ
- 高性能スカラプロセッサのSMP＆並列

◆ 最先端ハードウェアテクノロジの適用
- 最新プロセッサPOWER7（高性能・低消費電力）

◆ 各モデルの特長の明確化
- 設備条件の緩和（空冷モデルの継承、耐荷重の軽減）
- 世界最高クラスの高実装密度
- 大規模共有メモリを有する最高性能のSMPサーバ

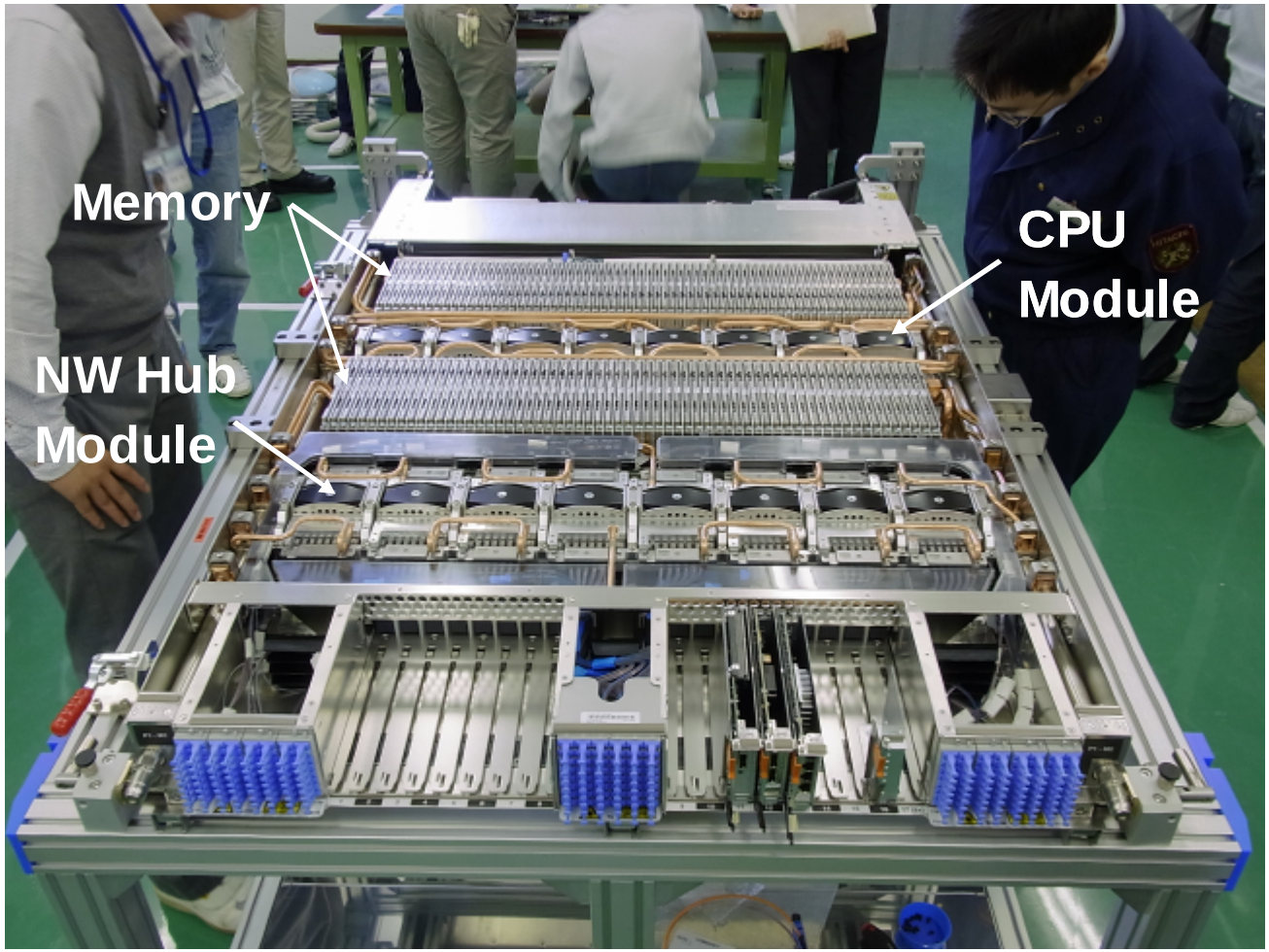
◆ HPC向け技術の継承
- チューニング技術（ｱﾌﾟﾘまでを見たﾄｰﾀﾙﾁｭｰﾆﾝｸﾞ）
- 運用技術：センター運用管理、単一ｼｽﾃﾑｲﾒｰｼﾞ等

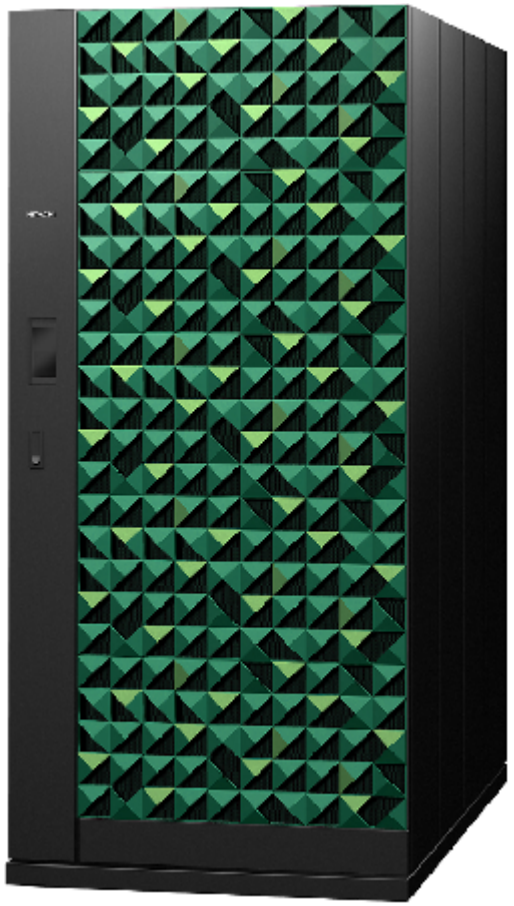


モデルM1

モデルXM1

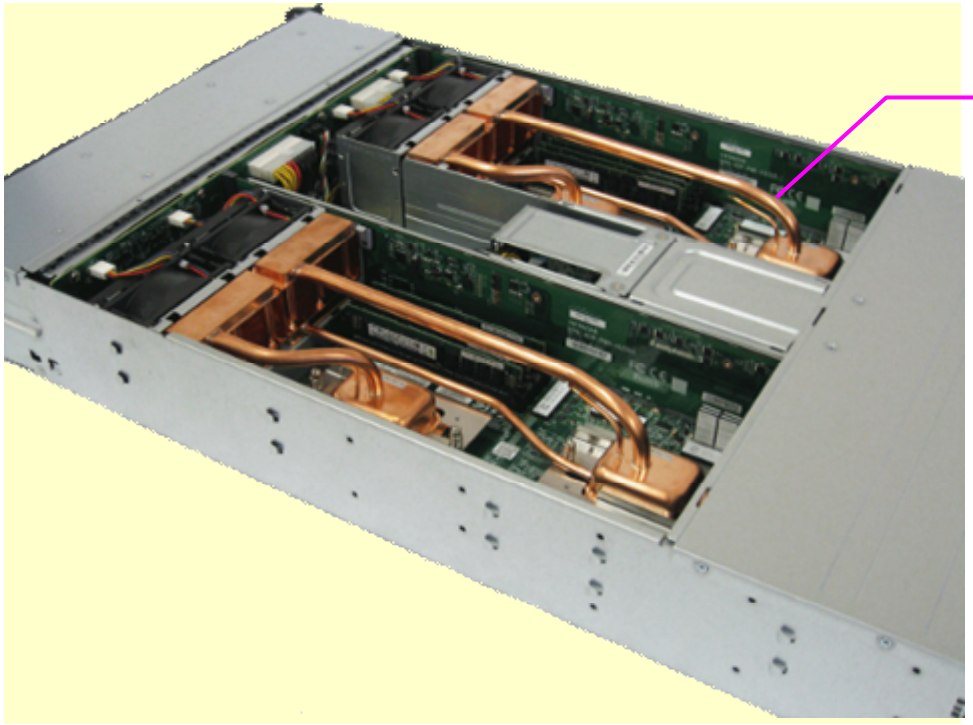
1-4 ＳＲ１６０００の紹介





8 node (CPU Module) / board

1-5 HA8000-tc/HT225の紹介



プロセッサ：AMD Opteron 6140/ 6276
(2.3GHz 16コア Interlagos)×2
294.4 GF

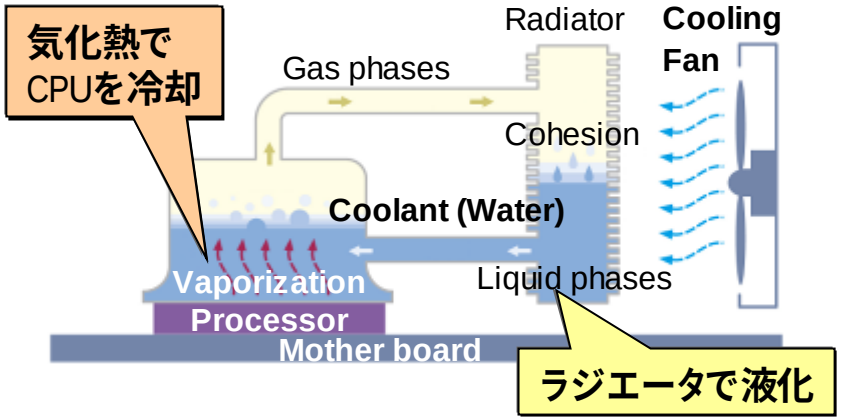
メモリ：最大 64GB (DDR3-1600)

HDD：2.5″SAS-2.0 HDD×4 (RAID 0,1,10)

拡張I/O：PCI-Express(x16) 1スロット,
PCI-Express (x8) 2スロット

電源：シャーシ内 2ノードで共用冗長構成

サーモサイフォン冷却



**冷却用ファンの回転数低減により、
省電力、低騒音を実現**

2 日立サーバラインアップ

・ブレードサーバ
・ラックマウントサーバ／タワーサーバ

2-1 BladeSymphony ラインアップ

各製品、充実のラインナップで、用途に応じた製品を提供
仮想化環境やソリューションを含めたシステム提供も可能

ターゲット	ブレードサーバ部	ストレージ部	ネットワーク部
Web3階層統合システム 大規模データベース等	ハイエンドモデル BS2000	Hitachi Storage Solutions	CommuniMax
PCサーバ統合 データセンター 中小規模の部門サーバ	小型高集積モデル BS320	BR1600E BR1600S BR1200	ロードバランサ AX2500

Windows Server 2008

redhat

vmware

サーバ仮想化　日立サーバ仮想化機構 Virtage　VMware vSphere™　Windows Server® 2008 Hyper-V™

BladeSymphony システム管理ソフトウェア

ミドルウェア　JP1　HiRDB　Cosminexus　主要ISV ソフトウェア

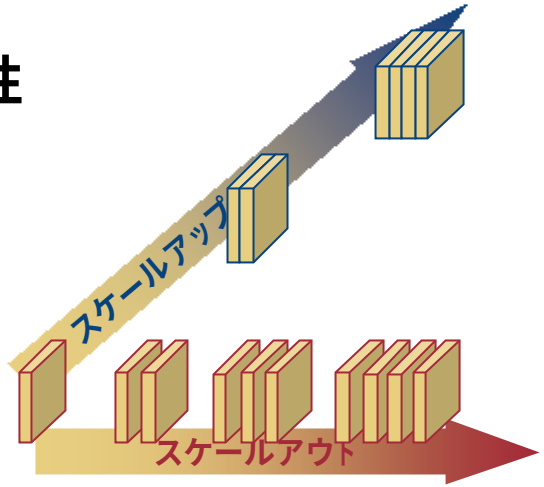
プラットフォームソリューション／サポートサービス

2-2 ハイエンドモデル　BS2000

ミッションクリティカル・システムにフォーカス BS2000

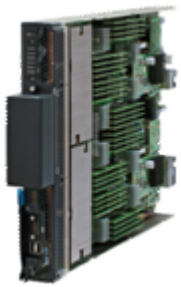
■ 仮想統合を実現する高信頼スケーラブル・ブレードサーバ

- 仮想化による集約、高速処理に適応した性能・拡張性
  (ブレード間SMP接続/ 大容量メモリー/ I/Oスロット拡張装置)
- 日立サーバ仮想化機構 *Virtage* 標準搭載(*1)
- メインフレームの高信頼・高可用化技術を継承
- 高効率電源の採用 (80 PLUS® GOLD認証取得(*2))
- ハードウェア長期保守対応 (ロングライフサポートサービス 7年/10年(*3))



*1：Essentialモデル
*2：電源負荷50%時の変換効率92%を実現
*3：BS2000 Eタイプにてサポート


標準サーバブレード


高性能サーバブレード


シャーシ：最大8ブレード/10U


I/Oスロット拡張装置

2-3 ハイエンドモデル　BS2000

＜2011年度の主な強化ポイント＞

■標準サーバブレード　性能強化
- ・最新Intel® Xeon® 5600番台プロセッサ
- ・16GB DIMMサポート

■高性能サーバブレード　性能強化
- ・最新Xeon E7ファミリー プロセッサ
- ・16GB DIMMサポート

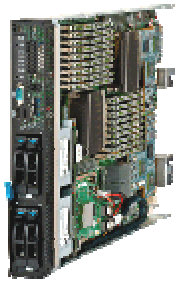
■日立サーバ仮想化機構Virtage 強化
- ・標準サーバブレード： 30LPAR
- ・高性能サーバブレード： 60LPAR

■ I/O系RAS機能強化

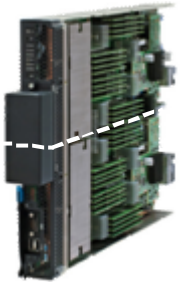
■キャパシティオンデマンド
- ・初期導入費用低減＆長期運用時の拡張性

BS2000

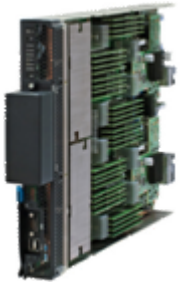
シャーシ：最大8ブレード/10U



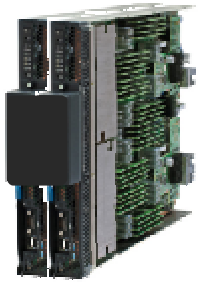
標準ｻｰﾊﾞﾌﾞﾚｰﾄﾞ



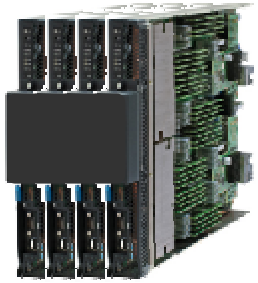
ｷｬﾊﾟｼﾃｨ
ｵﾝﾃﾞﾏﾝﾄﾞ



高性能
ｻｰﾊﾞﾌﾞﾚｰﾄﾞ



2ﾌﾞﾚｰﾄﾞSMP構成



4ﾌﾞﾚｰﾄﾞSMP構成

I/Oスロット拡張装置

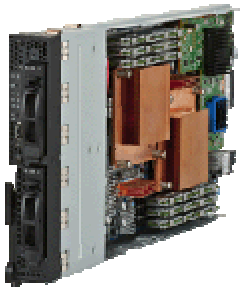
2-4 小型高集積モデル BS320

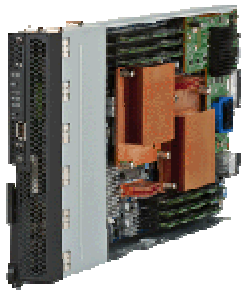
より軽く、より小さく 高密度実装を追求 BS320

■ 幅広い用途に対応する高集積･省電力ブレードサーバ

- 高さ6U(約27cm)に最大10ブレード搭載可能
- 最大重量約98kg/シャーシの軽量設計
- 用途に応じた多彩なサーバブレードをラインアップ
- 日立サーバ仮想化機構 *Virtage*に対応(*1)
- 高効率電源の採用 (CSCI Silver基準適合，80 PLUS® SILVER認証取得(*2))
- ハードウェア長期保守対応 (ロングライフサポートサービス：7年)

*1：PCI拡張ｻｰﾊﾞﾌﾞﾚｰﾄﾞ *Virtage*ﾓﾃﾞﾙで提供
*2：負荷50%時の変換効率89%以上を実現


標準サーバブレード


SAN専用サーバブレード


HDD拡張サーバブレード


PCI拡張サーバブレード

2-5 小型高集積モデル BS320

＜2011年度の主な強化ポイント＞

■最新Intel® Xeon® 5600番台プロセッサー

■大容量メモリ/次世代SSD

・32GB DIMMサポート

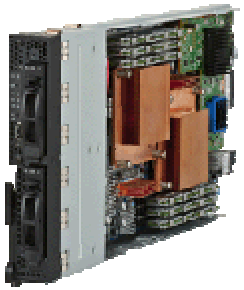
・SSD搭載サポート

■日立サーバ仮想化機構Virtage 強化

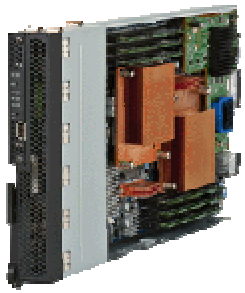
・LPAR数増強 (16LPAR)

■iSCSI対応N+1コールドスタンバイ

■省電力機能＆低電圧プロセッサ/メモリ



標準サーバブレード

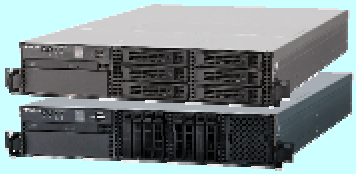


SAN専用サーバブレード

HDD拡張サーバブレード

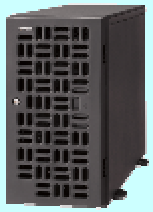
PCI拡張サーバブレード

2-6 HA8000ラインアップ/ラックサーバ

	2010/下	2011/上	2011/下～2012/上
4プロセッサー	HA8000/RS440(K1) 	HA8000/RS440(L)  ・CPUエンハンス ・低電圧メモリサポート ・10G LANサポート ・管理ツール添付拡充 ・ﾊﾟｯｸ型の保守商品を追加	HA8000/RS440  ・継続強化
2プロセッサー	HA8000/RS220(K1) 	HA8000/RS220(L)  ・CPUエンハンス ・RAIDｷｬｯｼｭ SC*ﾊﾞｯｸｱｯﾌﾟ ・2.5型SAS HDD容量追加 ・管理ツール添付拡充 ・ﾊﾟｯｸ型の保守商品を追加	HA8000/RS220  ・次世代プロセッサ ・ラインアップ拡充 等
	HA8000/RS210(K1) 	HA8000/RS210(L)  ・CPUエンハンス ・2.5型SAS HDD容量追加 ・管理ツール添付拡充 ・ﾊﾟｯｸ型の保守商品を追加	HA8000/RS210  ・次世代プロセッサ ・ラインアップ拡充 等
1プロセッサー	HA8000/RS110(K1) 	HA8000/RS110(L)  ・CPUエンハンス ・2.5型SAS HDD/SSDｻﾎﾟｰﾄ ・3.5型HDD搭載数アップ ・RAIDｷｬｯｼｭ SC*ﾊﾞｯｸｱｯﾌﾟ ・管理ツール添付拡充 ・簡単取り付けラックレール ・ﾊﾟｯｸ型の保守商品を追加	HA8000/RS110  ・次世代プロセッサ ・ラインアップ拡充 等

SC*：スーパーキャパシタ

2-7 HA8000ラインアップ/タワーサーバ

	2010/下	2011/上	2011/下～2012/上
2プロセッサー	HA8000/TS20(K1) 	HA8000/TS20(L)  ・CPUエンハンス ・2.5型SAS HDD容量追加 ・管理ツール添付拡充 ・ﾊﾟｯｸ型の保守商品を追加	HA8000/TS20  ・次世代プロセッサ ・ラインアップ拡充 等
1プロセッサー	HA8000/TS10(K1) 	HA8000/TS10(L)  ・CPUエンハンス ・2.5型SAS HDD容量拡大 ・SSDサポート ・RAIDｷｬｯｼｭ　SC*ﾊﾞｯｸｱｯﾌﾟ ・管理ツール添付拡充 ・ﾊﾟｯｸ型の保守商品を追加	HA8000/TS10  ・次世代プロセッサ ・ラインアップ拡充 等
	HA8000/SS10(K1) 	HA8000/SS10(L)  ・CPUエンハンス ・2.5型SAS HDD容量/搭載数拡大 ・SSDサポート ・管理ツール添付拡充 ・ﾊﾟｯｸ型の保守商品を追加	HA8000/SS10  ・継続強化
HA8000-bd	HA8000-bd/BD10(x1)  ・Windowsサポート ・SSDサポート ・iSCSIサポート ＋ HDDレスモデル	・大容量SSDサポート ・コンテンツ配信事業者や DC事業者向けサービスの充実	HA8000-bd/BD10(x2)  ・次世代プロセッサ ・ラインアップ拡充 等

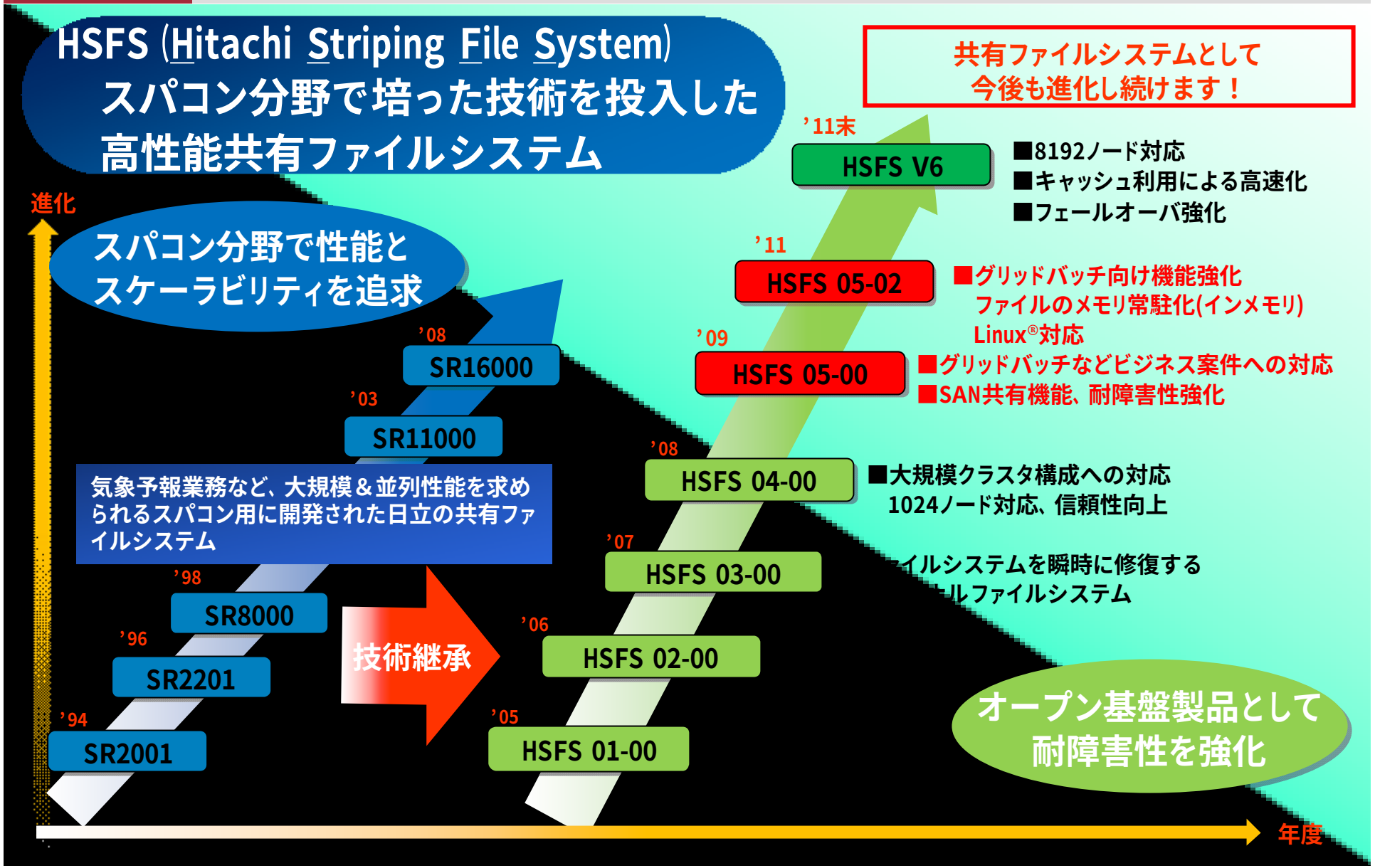
SC*：スーパーキャパシタ

3 事例紹介

・北海道大学殿「北海道大学アカデミッククラウド」
・東京大学医科学研究所殿

4 分散並列ファイルシステム HSFS Version6

4-1 日立の分散並列ファイルシステムHSFS



4-2 大規模システムにも柔軟に対応する並列ＦＳ

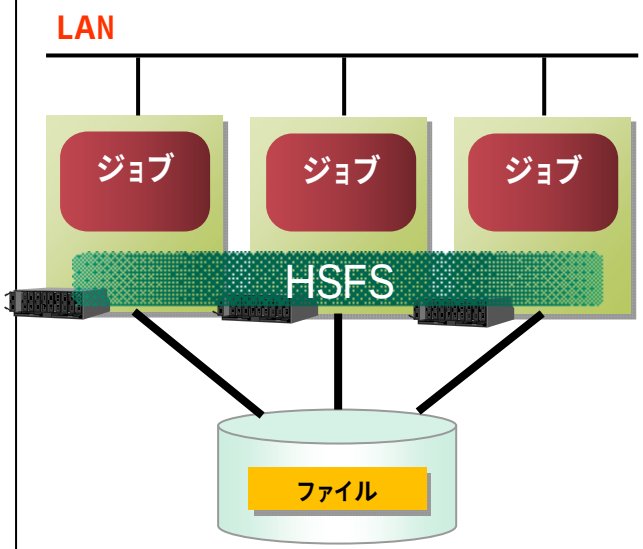
小規模から大規模まで台数に応じた構成を組むことが可能です

特長

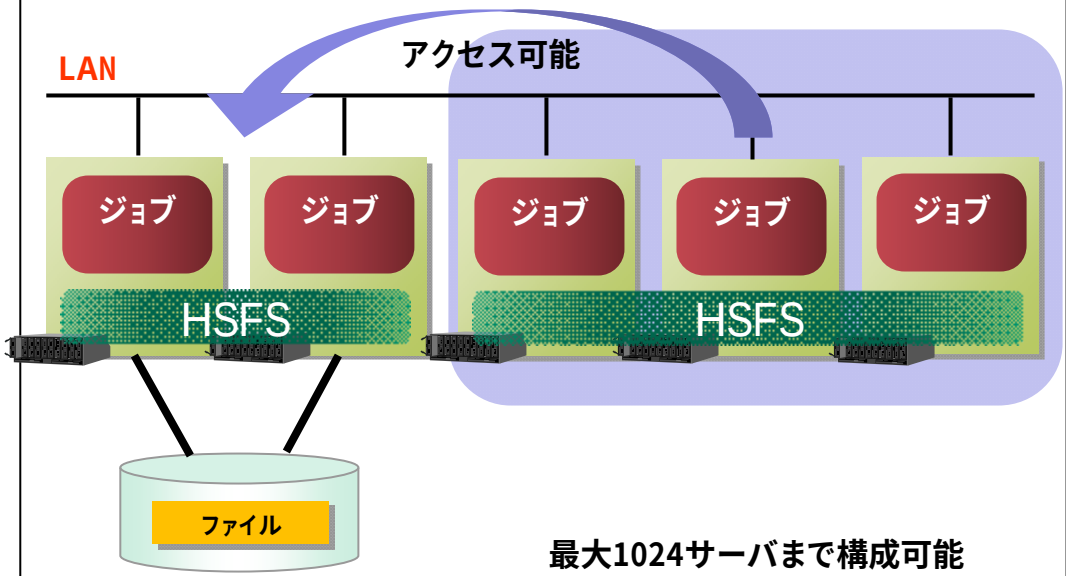
バッチ処理業務の増加に伴い、多数のサーバ台数が必要となる場合があります。
各サーバにディスク装置を接続できない大規模システムでも、ネットワークによるファイル共有を構築することができます。SAN共有機能とネットワーク共有の混在型も可能です。

SAN共有機能型

◆ サーバ台数が比較的少ない場合



ネットワーク共有型

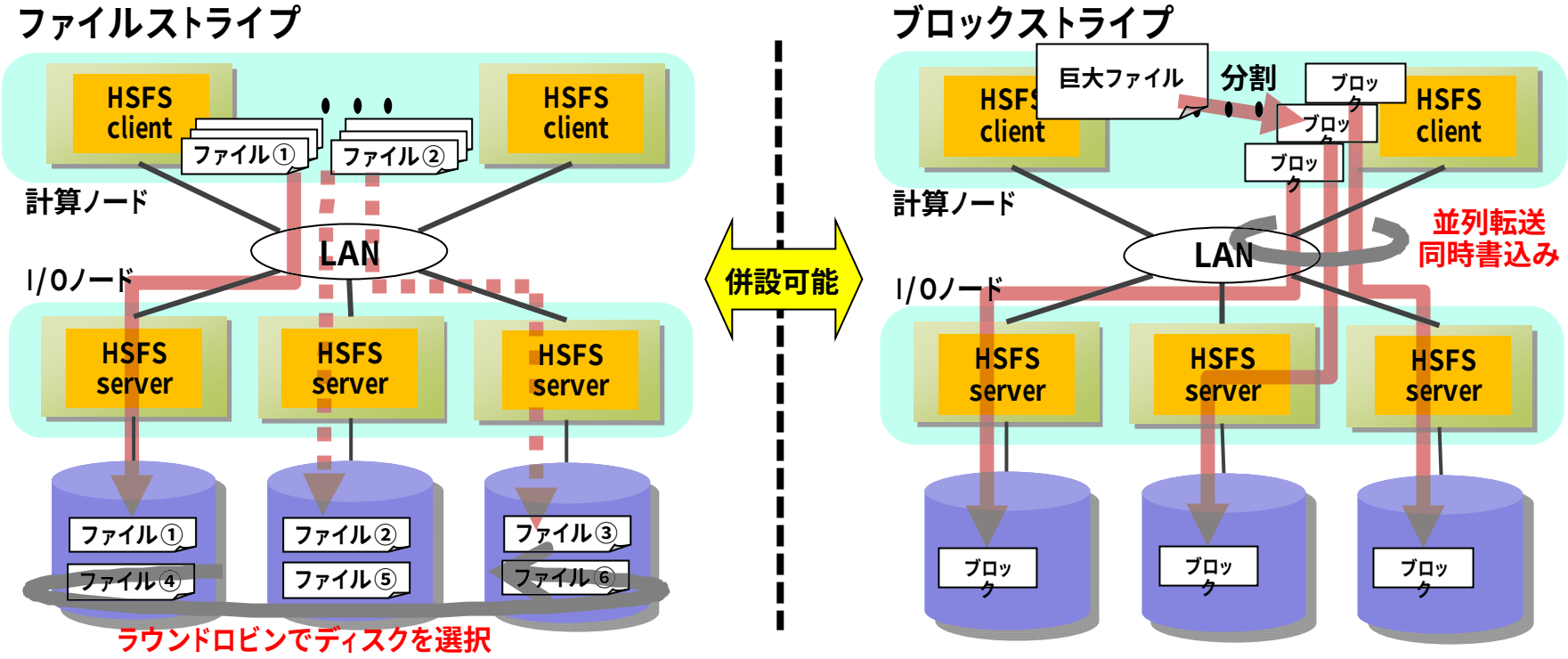


最大1024サーバまで構成可能

4-3 ストライピング機能 (2つのストライプ方式)

分散方式	特徴	適性
ファイル ストライプ	ファイル単位で分散配置(ラウンドロビン) 複数のファイルを別々のディスクに格納する 各ディスクのファイル数を平準化	小サイズファイルI/Oで高性能を発揮 ・MBオーダー未満のI/Oプログラム ・TSS環境、コンパイル環境など
ブロック ストライプ	ファイルを複数ブロックに分割してから配置 1つのファイルをブロック分割(*1)してから、 ブロックを別々のディスクに並列転送	巨大なファイルのI/O時間短縮 (小サイズファイルの場合、ファイル分割 損が生じるため非効率となることがある)

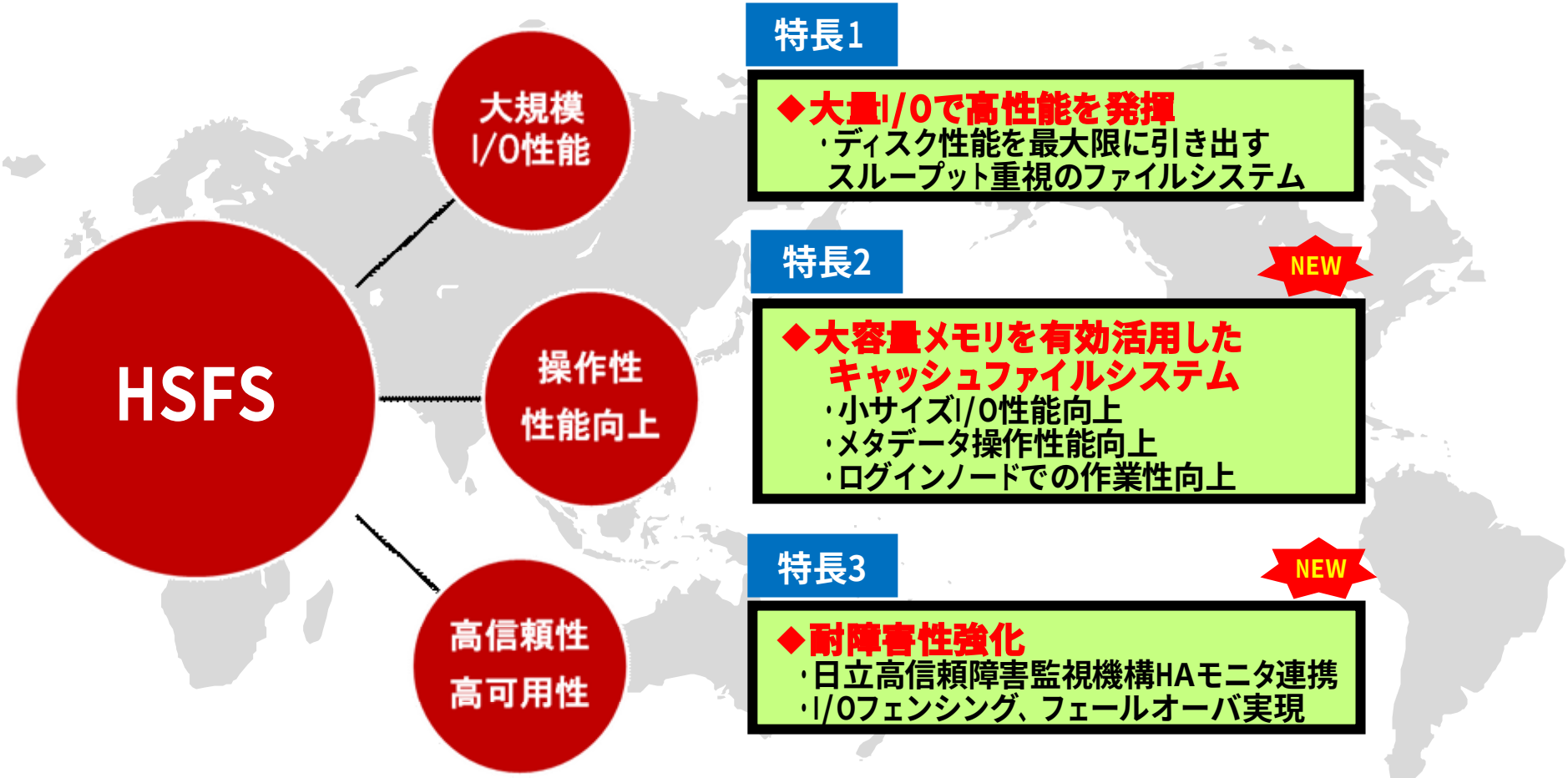
(*1) ブロックストライプのブロック分割数はシスパラで変更可



4-4 HSFSのコンセプトとVersion 6の特徴

大規模I/O重視型ファイルシステムをベースに、V6で更に進化します！

- 小サイズI/O性能も大きく向上させ、インタラクティブ操作の快適性の追求
- 障害発生時の確実なI/Oフェンシングとフェールオーバによるユーザ資産保護



特長1

◆大量I/Oで高性能を発揮
- ディスク性能を最大限に引き出す
  スループット重視のファイルシステム

特長2 NEW

◆大容量メモリを有効活用した
キャッシュファイルシステム
- 小サイズI/O性能向上
- メタデータ操作性能向上
- ログインノードでの作業性向上

特長3 NEW

◆耐障害性強化
- 日立高信頼障害監視機構HAモニタ連携
- I/Oフェンシング、フェールオーバ実現

4-5 特長1 大規模I/Oで性能を発揮

ディスク性能を最大限に引き出し、大規模･大量I/Oに強いファイルシステム

過去の実績

◆SR8000ではOS(HI-UX/MPP)の一機能として提供[～2002年]
　－大学や研究所等の日立スパコンユーザで多数の稼働実績あり
　－ハードピーク性能の90%を超える性能を発揮(1GbpsのFCで90MB/s)

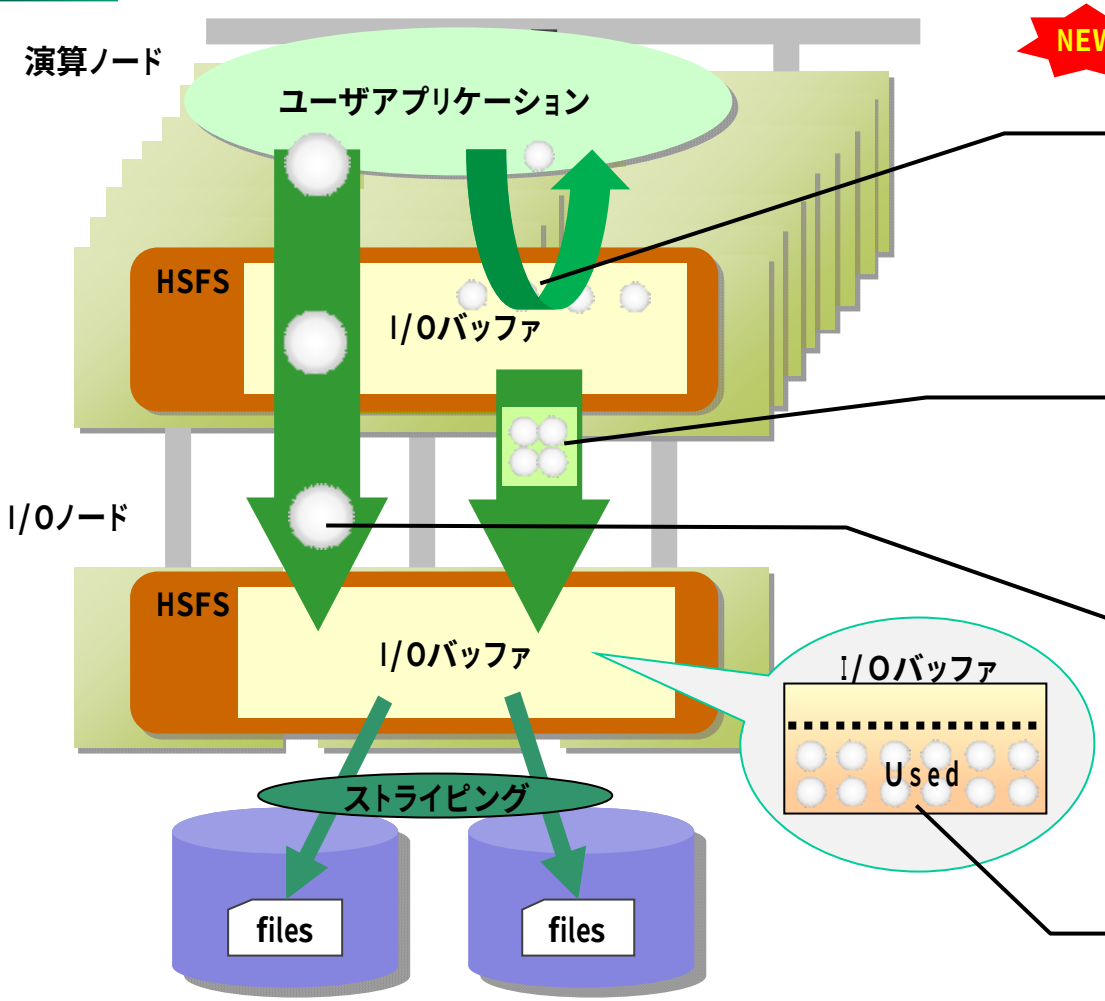
◆SR11000で高性能ファイルシステムとして製品化[2005年]
　－納入時のBMTで20GB/s達成(128ノードでの総スループット)

◆1台の大規模SMPサーバでの高い性能要求にも対応[2009年]
　－納入時のBMTで1台のサーバで6GB/s達成(SR16000,FC48本直結構成)

4-6 特長2　快適なインタラクティブ性能を提供

バッファメモリを利用した非同期I/Oをフル実装したキャッシュファイルシステム

特長　HSFSのI/Oバッファを利用したI/O完全非同期化により、新規ファイル生成/削除のコストを1msec未満に短縮し、快適なインタラクティブ性能を提供



NEW

(1) インタラクティブ操作性向上

I/Oは、I/Oバッファにデータを格納した時点で完了し、その後、非同期転送する。

※ I/Oバッファ上に蓄積されるデータは、fsync(),close()時に整合性を保証します

(2) I/Oデータ蓄積＆一括転送

演算ノードのI/Oバッファに小サイズデータを蓄積してから一括転送することで、ノード間ネットワークとDISKのスループットを上げます。

(3) 大規模データは即時転送

巨大データは、I/Oバッファに長時間置かず、即時I/Oノードに転送することで性能を確保します。

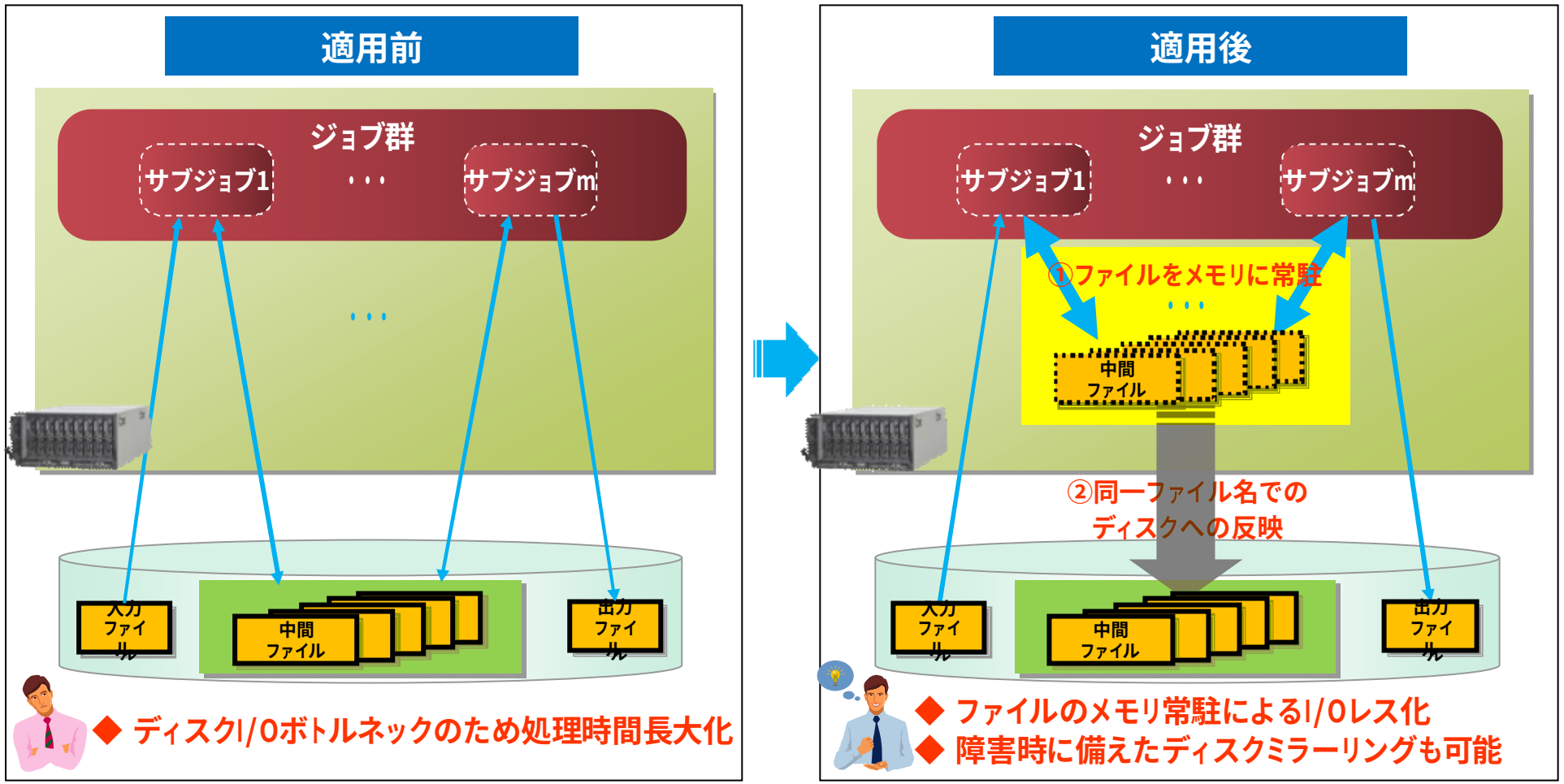
(4) I/Oバッファ有効活用

I/Oバッファ枯渇前に、I/OデータをI/Oノードに押出すことで、キャッシュの利用効率を促進します

4-7 特長2-1 バッチジョブ高速化のためのインメモリ機能

ファイルのメモリ常駐化機能(インメモリ)により、I/Oレスを実現

特長 ジョブ間で引き継ぐ一時的な中間ファイルなどを、各サーバのメモリ上に常駐することができます。
(本機能はサーバ間でメモリを共有するものではありません)

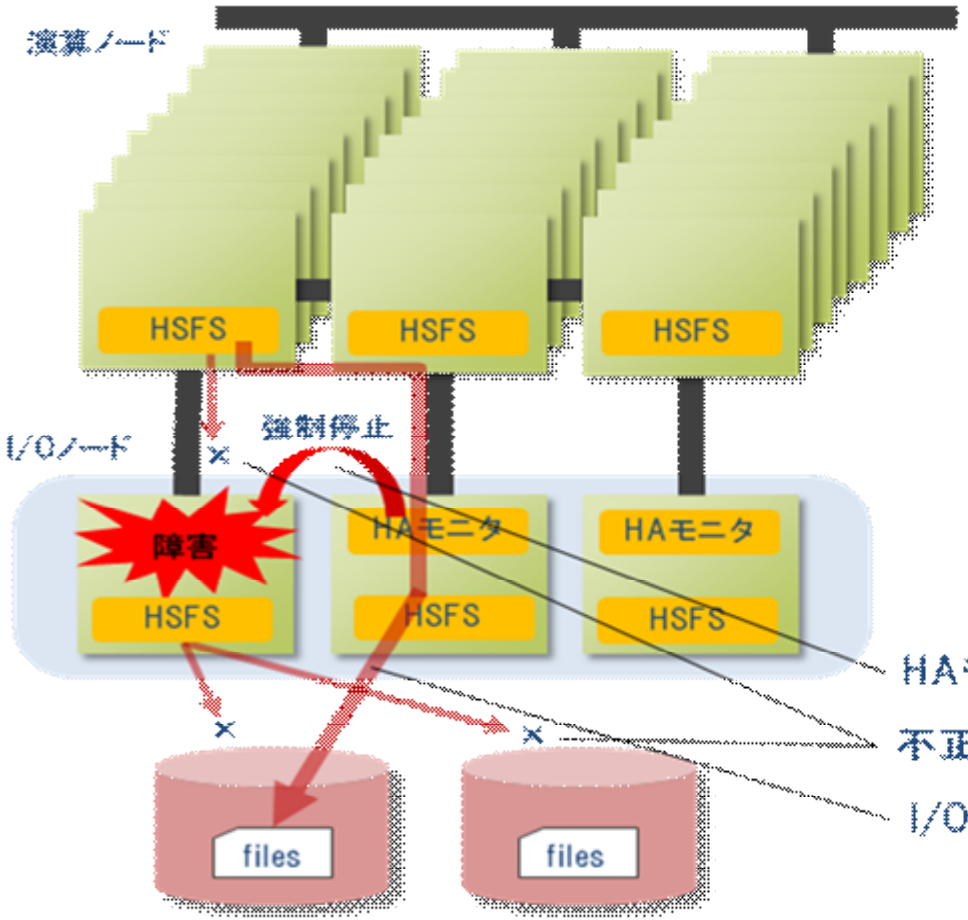


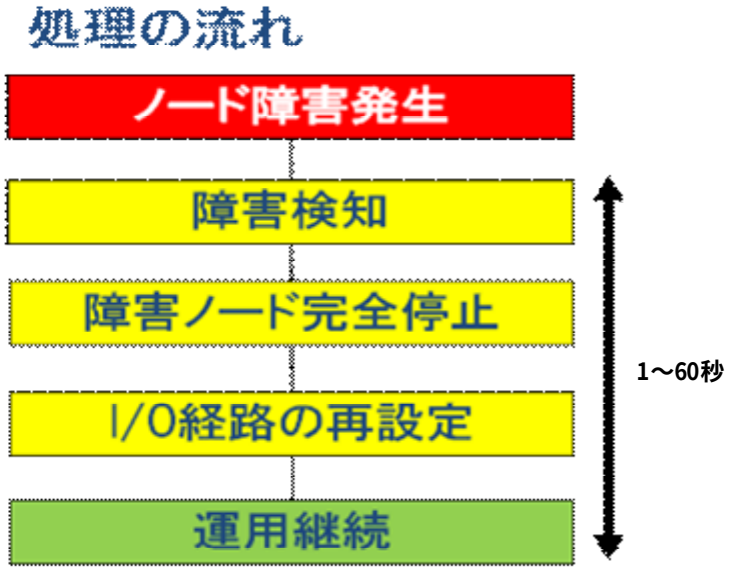
4-8 特長3　ファイルを保護する耐障害性機能

高信頼切換え機能「HAモニタ」と連携したI/Oフェンシングとフェールオーバ

特長

I/OノードのOSスローダウン時やハード障害時、
ファイルシステムを守るため、HAモニタ(高信頼障害監視機構)と連携し、
不安定なI/Oによるデータ破壊を確実に遮断し、安全なフェールオーバを実現します。





HAモニタで障害ノードを完全停止

不正アクセス遮断によるファイル破壊防止

I/O経路の再設定

HITACHI

Inspire the Next

他社商品名、商標等の引用に関する表示

■製品の内容･仕様は、改良のために予告なしに変更する場合があります。
■製品写真は出荷時のものと異なる場合があります。
■インテル、Intel、Xeon、Itaniumは、アメリカ合衆国およびその他の国におけるIntel Corporationの商標です。
■ AMD、Opteronは、Advanced Micro Devices, Inc.の商標または登録商標です。
■ Microsoft、Windows、Windows Server、Windowsロゴは、米国Microsoft Corpotationの米国およびその他の国における商標または登録商標です。
■ Linuxは、Linus Torvalds氏の日本およびその他の国における登録商標あるいは商標です。
■ Red HatならびにShadow Manロゴは、米国およびその他の国でRed Hat,Inc.の登録商標もしくは商標です。
■ VMwareは、VMware,Inc.の米国およびその他の国における登録商標または商標です。
■ AIXは、米国およびその他の国におけるInternational Business Machines Corporationの商標です。
■ AIX 5Lは、米国およびその他の国におけるInternational Business Machines Corporationの商標です。
■ Javaは、 Oracle Corporation 及びその子会社、関連会社の米国 及びその他の国における登録商標または商標です。

