B-032

# スーパーテクニカルサーバSR11000の高速分散ファイルシステムHSFS

## HSFS:Hitachi Striping File System for Super Technical Server SR11000

清水 正明† 鵜飼 敏之† 三瓶 英智‡ 飯田 恒雄‡ 藤田不二男‡
Masaaki Shimizu Toshiyuki Ukai Hideaki Sanpei Tsuneo Iida Fujio Fujita

## 1．はじめに

近年のスーパーコンピュータに対する性能要求では、演算性能ばかりではなく、演算データの入出力を行なうディスク装置およびファイルシステムの入出力性能が重要視されている。特に多数のノードで構成する並列計算機においては、複数ノードに分散したディスク装置を統合し、全ノードに対し単一ファイルシステムツリーを提供する分散ファイルシステムが必要である。

本稿ではスーパーテクニカルサーバSR11000において上記要求に応える為に開発した高速分散ファイルシステムHSFSについて報告する。

## 2．スーパーテクニカルサーバSR11000の概要

SR11000は、ベクトル・スカラ融合型サーバSR8000シリーズの後継機として開発を進めてきた科学技術計算サーバである。

SR11000モデルJ1では1.9GHz動作のPOWER5プロセッサを16個搭載するメモリ共有ノードを基本構成単位とし、多段クロスバネットワークにより複数ノードを接続する分散メモリ型並列計算機の構成を取っている。最大512ノードを接続することで理論ピーク性能62.2TFLOPSを実現し、大規模な科学技術計算に対応可能である。また、演算性能と入出力性能の要求に応じて、全ノードにディスクアレイ等の入出力装置を接続することも、または一部ノードにのみ入出力装置を接続することも可能である。

オペレーティングシステムはAIXをベースとして、SR8000用の日立独自OS HI-UX/MPPの特長技術である分散ファイルシステム、拡張記憶機能、予実算管理機能、専用ディスクアレイドライバを追加している。

## 3．分散ファイルシステムに求められる要件

SR11000の分散ファイルシステムに対する要件を示す。

(1) アクセスの透過性

SR11000は複数のノードで構成するクラスタシステムであり、分散ファイルシステムに対する第一の要件は、クラスタのいずれのノードからもファイルに透過的アクセス可能な環境を提供することである。つまり、複数のノードに分散したディスク装置とファイルを統合し、任意のノードからアクセスできる機能が必要である。

(2) ディスクとノードの性能スケーラビリティ

第二の要件は高い単体ディスク性能の達成とともに、複数ノードの同時入出力に対してもスケーラブルに総合入出力性能が向上することである。

(3) 大規模単一ファイル入出力性能と多数ファイル同時入出力性能の両立

第三の要件は、大規模単一ファイルのシーケンシャル入出力性能および、TSS利用時または/home領域等における多数ファイル同時入出力性能を両立させることである。

## 4．高速分散ファイルシステムHSFS

3章の要件を満たすために SR8000 で実績のある HI-UX/MPP の分散ファイルシステム HSFS ( Hitachi Striping File System ) をクラスタシステムSR11000上に実現した。

### 4.1 HSFSの概要

HSFSは複数のノードで単一ファイルシステムツリーを共有可能にする分散ファイルシステムであり、ディスクを接続した複数の入出力ノードにファイルを分割・分散（ストライピング）配置することで、入出力の並列化によるトータルスループットの向上を可能にしている。このとき、ファイルのストライピング方法として、ブロックストライプ方法だけでなくファイルストライプ方法を利用できることが特長である（図1）。

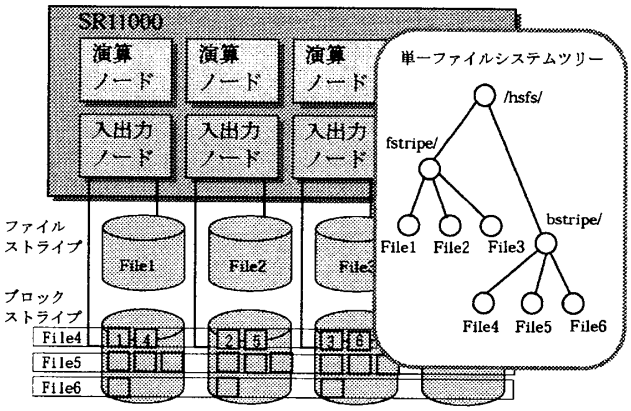


図1 HSFSの概要

ファイルを指定サイズ単位で複数ディスクにストライピング配置するブロックストライプ方法は大規模単一ファイルの入出力性能を高める際に有効であるが、多数ファイルを同時に入出力する場合に適用するとすべてのファイルがそれぞれ多数のディスクにストライプされるため、ディスクアクセスのランダム化やデータ転送のためのノード間通信が多発し、ディスク装置の性能や台数効果を発揮しにくくなる。多数のファイルに同時入出力する場合には、一つのファイルは一つのディスク内に固定して配置するファイルストライプ方法が有効である。

### 4.2 HSFSのSR11000における実装

(1) HI-UX/MPPのシングルOS機能の導入

第一の要件であるディスク装置とファイルを透過的に管理するために、HI-UX/MPPの特長技術であるシングルOS技術を導入した。SR8000においては、分散OS Machと

† (株) 日立製作所 中央研究所
Central Research Laboratory, Hitachi, Ltd.
‡ (株) 日立製作所 ソフトウェア事業部
Software Division, Hitachi, Ltd.

OSF/1 MK-AD UNIXサーバのシングルOS機能を使用して複数のノードのディスクやファイルシステムを透過的に扱うことで共有を実現していた。しかしSR11000は各ノードでAIXが稼動するクラスタシステムであり、各ノードでディスク装置とファイルシステムは独立しているため共有はできない。

そこで、AIX上でMachカーネルとOSF/1 UNIXサーバの実行を可能にする技術を開発し、シングルOS機能を利用してディスクとファイルシステムの共有を実現することにした☆。

SR11000における実装では、AIXのカーネル拡張機能を利用してMachカーネルとOSF/1 UNIXサーバの一部であるファイルサーバ（分散ファイルシステムHSFS）を実装した（図２）。Machカーネルがデバイス管理、メモリ管理、ノード間メッセージ通信機能を実現し、OSF/1サーバはMachの機能を利用してディスク装置とファイルに対する透過的なアクセスを実現している。ユーザはOSF/1サーバが提供する分散ファイルシステムHSFSを利用することですべてのノードから共通のファイルシステムツリーにアクセスできる。

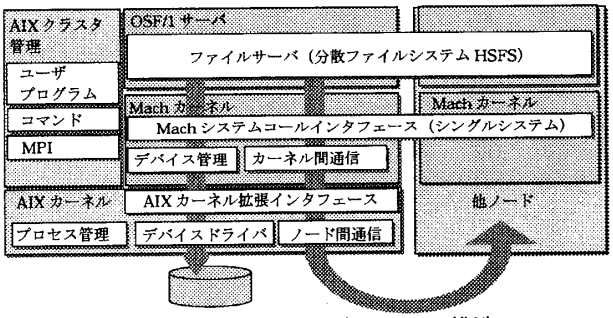


図２　AIXクラスタ上のHSFS構造

(2) サブファイルシステム構造

HSFSのストライピング機能は、ファイルをストライピングする方法としてサブファイルシステム構造を採っている。具体的には分散ファイルシステムの各ノードのディスクに分散配置した実ファイル群をサブファイルとして利用し、利用者にはサブファイルの集合を一つの論理ファイルとして見せている。ブロックストライプでは一つの論理ファイルを複数のノードのサブファイルにストライピングし、ファイルストライプでは一つの論理ファイルを特定ノードのサブファイルに対応させている。

分散ファイルシステムを管理する各ノードから見た場合、各サブファイルはノード内の一つの実ファイルという点がサブファイルシステム構造を持つHSFSの性能上の利点となっている。ノード内のファイル入出力の性能はディスクブロック割り当て方法や先読み、まとめ書き等により最適化を図ることが容易である。またノード毎に独立して入出力を行なうためノード間の性能独立性が高くなり、高いスケーラビリティを実現できる。

このようなサブファイルシステム構造とノード毎のファイル入出力性能の最適化によって、３章の二つ目の要件である高い単体ディスク性能およびディスクとノードの高い性能スケーラビリティを達成した。また、ブロックストライプ機能およびファイルストライプ機能により、三つ目の要件である大規模単一ファイル入出力性能と多数ファイル同時入出力性能の両立を実現した。

☆SR11000においてはAIXをベースに本シングルOS機能およびHSFSを実装したが、これらの機能はLinux等のコモディティOS上においても容易に実現可能であるように移植性を高く設計した。

## 5. HSFSの性能評価

8ノード構成のSR11000において、各ノードに1台ずつディスクアレイ装置を接続し、実装したHSFSの性能評価を行なった。各ディスクアレイ装置では10,000rpmのFC SCSIディスクを用いて4D+1P構成で論理ディスクを構成し、SR11000と2Gbps Fibre Channelで接続した。

性能評価は、まず1ノードでディスクアレイ装置1台に対する基本入出力性能を測定し、次に8ノード構成でブロックストライプとファイルストライプの性能を測定してスケーラビリティを確認した。基本性能測定プログラムでは2MB単位で合計8GBのファイルのシーケンシャル読み書き性能を測定し、ブロックストライプ性能測定プログラムでは64GBの論理ファイルに対して8ノードから8GBずつの割り当て部分に対して同時に読み書きを行なった性能を測定した。またファイルストライプ性能測定では、8ノードがそれぞれ1個の8GBファイルの論理ファイルに対して読み書きを行なった性能を測定した（図3）。

性能測定の結果、基本性能では2Gbps Fibre Channelの80%以上の性能を達成し、大規模単一ファイルのシーケンシャル入出力に対応するブロックストライプおよび多数ファイル同時入出力に対応するファイルストライプでは基本性能に対して94%以上のスケーラビリティを確認できた。

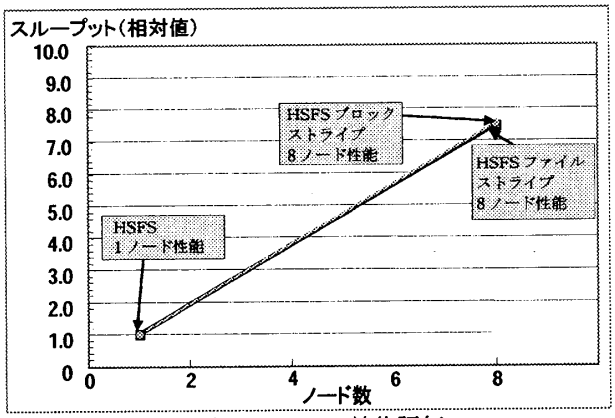


図３　HSFSの性能評価

## 6．おわりに

スーパーテクニカルサーバSR11000において、高速分散ファイルシステムHSFSを開発した。HI-UX/MPPのシングルOS技術を用いた分散ファイルシステム機能、サブファイルシステム構造のストライピング機構を利用することで、ファイルシステムに対する要件である、透過性、性能スケーラビリティ、大規模単一ファイル入出力性能と多数ファイル同時入出力性能の両立を実現した。8ノード構成における性能評価の結果、高い基本性能とブロックストライプおよびファイルストライプにおける高いスケーラビリティを確認した。

**謝辞**　本ファイルシステムの開発にあたり、多大なるご指導をいただいた東京大学情報基盤センターの金田康正教授に深く感謝いたします。