

# スーパーコンピュータ「京」の 高性能・高信頼ファイルシステム

## High-Performance and Highly Reliable File System for the K computer

● 酒井憲一郎      ● 住元真司      ● 黒川原佳

---

### あらまし

理化学研究所と富士通は、10 PFLOPSの演算性能を持つスーパーコンピュータ「京」の開発を進めている。「京」は8万を超える膨大な計算ノードに加え、数十Pバイトのディスク容量とTバイト/秒のI/O帯域を持つ世界トップレベルのストレージシステムで構成される。このストレージシステムを構成する数千台のファイルサーバとストレージ装置を数万クラスの計算ノードから高速にアクセス可能とするため、高性能・高信頼なクラスタ型の分散ファイルシステム「FEFS(Fujitsu Exabyte File System)」を開発した。

本稿ではFEFSに関し、大規模システムで解決しなければならない課題とその施策について述べる。

### Abstract

RIKEN and Fujitsu have been developing the world's fastest supercomputer, the K computer. In addition to over 80 000 compute nodes, the K computer has several tens of petabytes storage capacity and over one terabyte per second of I/O bandwidth. It is the largest and fastest storage system in the world. In order to take advantage of this huge storage system and achieve high scalability and high stability, we developed the Fujitsu Exabyte File System (FEFS), a clustered distributed file system. This paper outlines the K computer's file system and introduces measures taken in FEFS to address key issues in a large-scale system.

---

## まえがき

近年、スーパーコンピュータの性能向上は目覚ましく、2011年には演算性能が10 PFLOPSに達した。計算ノードの総数・総コア数および搭載メモリ量の増加に伴い、ファイルシステムに求められる容量と総スループット性能も増大し、容量・性能ともに1年で10倍近いペースで飛躍的に伸びており、今後、容量は100 Pバイト級、性能は1 Tバイト/秒級に達すると予想される。このため、ファイルシステムの主流は単一サーバ型からクラスタ型へと変わりつつある。

スーパーコンピュータ「京」<sup>(注)</sup>の世界トップクラスの演算性能にふさわしい容量と性能のファイルシステムを実現するため、クラスタ型の分散ファイルシステム「FEFS (Fujitsu Exabyte File System)」を開発した。

本稿では、「京」のファイルシステムの概要について述べた後、大規模システム上での重要な課題についてFEFSの施策を紹介する。

(注) 理化学研究所が2010年7月に決定したスーパーコンピュータの愛称。

## 「京」のファイルシステムの概要

「京」のファイルシステムは世界トップクラスにふさわしい性能と安定したジョブ実行を実現するため、ジョブ専用の高速一時保存領域として使用する「ローカルファイルシステム」と、ユーザのファイルを保存する大容量の共用保存域として使用する「グローバルファイルシステム」の二階層のファイルシステムモデルを採用した(図-1)。この二つのファイルシステム間をシステム制御のファイルステージングによりデータ転送することによりジョブ実行を行う。以下にそれぞれの役割を述べる。

### (1) ローカルファイルシステム

バッチジョブとして実行されるアプリケーションのファイルI/O性能を最大限に引き出すための、ジョブ専用の高速なテンポラリ領域である。ファイルステージング機能により入出力ファイルをグローバルファイルシステムとの間で転送し、ローカルファイルシステム上には実行中または実行待ちのジョブのファイルが一時的に置かれる。

データブロックにアクセスするファイル

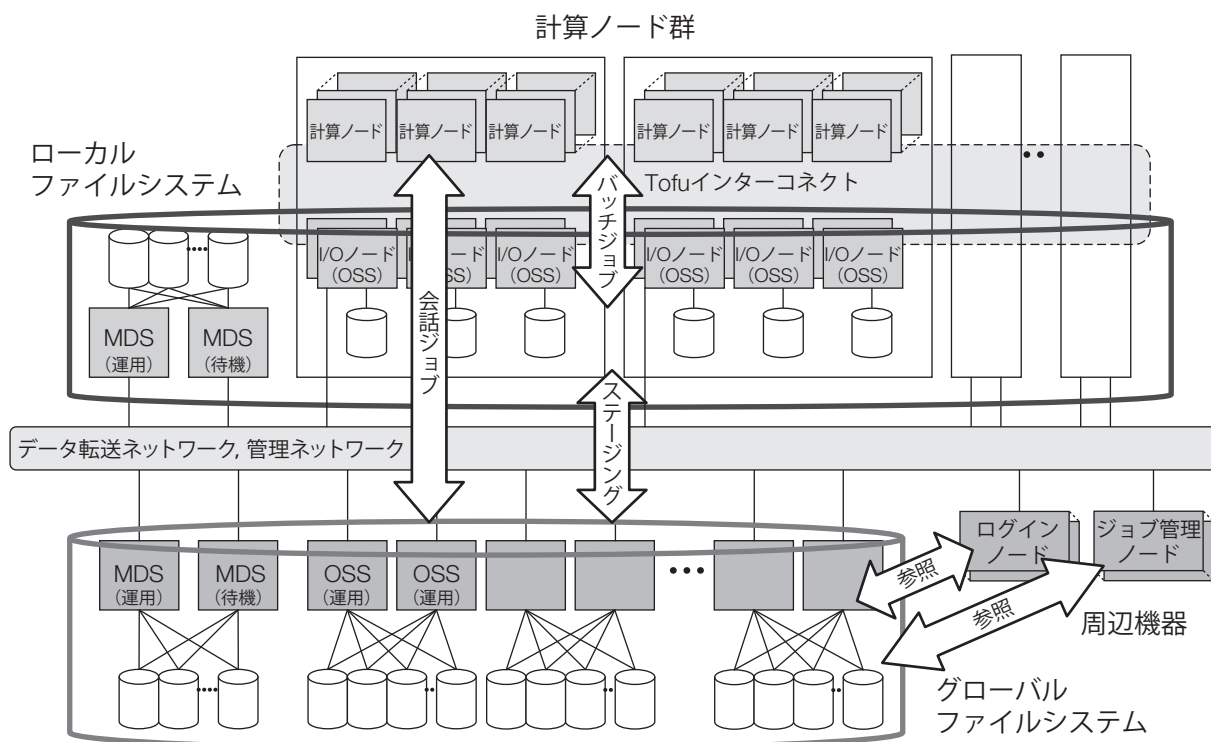


図-1 二階層のファイルシステムモデル

サーバには計算ラック内のI/O専用ノード（I/Oノード）を用い、計算ノードとTofuインターコネクトで接続し、RDMA（Remote Direct Memory Access）通信により低レイテンシ・高スループットのファイルデータ転送を実現している。

## (2) グローバルファイルシステム

計算ラックの外部に配置し、ジョブの入出力データなどユーザファイルを格納する大容量の共用域として利用する。ログインノードからのファイルアクセスに加え、会話型ジョブ利用などでは計算ノードから直接ファイルアクセスすることで、ジョブ出力をユーザが確認しながらのデバッグとチューニングが可能である。

計算ノードは、前述したI/Oノードに搭載するInfiniBandでグローバルファイルシステムのファイルサーバ群と接続する。I/OノードはTofuとInfiniBandとの間のファイルデータの転送を中継し、計算ノードはI/Oノードを中継してファイルサーバにアクセスする。

## (3) ファイルステージング

ローカルファイルシステムとグローバルファイルシステム間のファイルの受渡しは、ファイルステージング機能によりシステムが自動的に行う。

ファイルステージング機能はジョブ運用ソフトウェアと連携して動作する。ジョブ開始までに入力ファイルをグローバルファイルシステムからローカルファイルシステムに転送（ステージイン）し、ジョブ終了後に出力ファイルをローカルファイルシステムからグローバルファイルシステムに転送（ステージアウト）する。ユーザはステージイン・ステージアウトするファイルをジョブスクリプトで指定する。

次章でこれらの役割を実現するファイルシステムFEFSを紹介する。

## 超大規模クラスタファイルシステムFEFS

「京」のローカルファイルシステムとグローバルファイルシステムにおいては、世界トップクラスの計算性能にふさわしいファイルシステムを提供するためFEFSを開発した。FEFSの開発においては次の目標を設定した。

- ・世界最速のI/O性能、高性能なMPI-IO
- ・外乱排除で安定したジョブ実行時間
- ・世界最大となるファイルシステム容量
- ・性能・容量がハードウェア増設でスケラブルに向上
- ・高信頼性（サービスの継続、データの保水性）
- ・使いやすさ（多数ユーザによる共用）
- ・フェアシェア（多数ユーザによる公平性）

これらの目標を達成するためのファイルシステムは、世界的に業界標準のオープンソースファイルシステムであるLustreをベースとして開発を進め、FEFSで必要な仕様と機能の拡張を図った。表-1に示すように、FEFSは8 E（エクサ）バイトクラスの規模までファイルシステムサイズ、ファイルサイズなどの仕様を既存のLustreから拡張しており、サーバとストレージの増強により運用中においてもファイルシステムのサイズの拡張が可能となっている。

次章以降では、FEFSの超大規模対応のために重要な次の4点について述べる。

- ・通信バス・ディスクの分離によるI/O競合排除
- ・ハードウェア二重化とFailoverによる運用継続
- ・階層的なノード監視と自動交替

表-1 LustreとFEFSの仕様比較

機能		Lustre	FEFS
システム仕様上の制限	最大ファイルシステムサイズ	64 Pバイト	8 Eバイト
	最大ファイルサイズ	320 Tバイト	8 Eバイト
	最大ファイル数	4 G個	8 E個
	最大OSTボリュームサイズ	16 Tバイト	1 Pバイト
	最大stripe数	160個	2万個
	最大ACLエントリ数	32エントリ	8191エントリ
スケラビリティ	最大OSS数	1020個	2万個
	最大OST数	8150個	2万個
	最大クライアント数	128台	100万台
ブロックサイズ (ldiskfs) (Backend File System)		4 Kバイト	～ 512 Kバイト

・運用に応じたQoSポリシー選択

通信パス・ディスクの分離によるI/O競合排除

クラスタ型ファイルシステムであるFEFSは、多数のファイルサーバを束ねることで並列スループット性能をファイルサーバの台数に比例して向上させることが可能である。しかし、特定のファイルサーバにアクセスが集中すると、通信の輻輳やディスクアクセスの競合が発生し、I/O性能が低下して計算ノードのジョブ実行時間がばらつくことになる。安定したジョブ実行のためには、ファイルI/Oの競合を徹底して排除することが重要である。

このため、FEFSはジョブ単位と計算ノード単位の2段階でファイルI/Oのアクセスを分離することで、ファイルI/Oの競合を排除している（図-2）。

ジョブレベルでは、ジョブごとにファイル格納先I/Oノードの範囲を分けることでサーバやネットワーク、ディスク上でのファイルI/Oの競合を排除する。

また、ジョブ内の計算ノードレベルでは、Tofu通信のホップ数が最小のI/Oノードを経由してファイルデータの送受信を行うことで、計算ノード間のファイルI/Oの競合を最小化している。

ハードウェア二重化とFailoverによる運用

性能と並んで大規模システムで重要なのは、高

い信頼性である。クラスタ型ファイルシステムは多数のファイルサーバ、ストレージ装置、およびネットワーク機器で構成されるため、それらの一部が故障またはダウンした場合や保守中であってもファイルシステムのサービスを持続し、システムの運用を継続させなければならない。

しかし、数百台を超えるファイルサーバとストレージ装置でFEFSを構成する場合、ネットワークアダプタ、サーバなどのハードウェアが故障し保守待ちの状態になる可能性がある。このような状況下でもシステム全体の運用を継続させるためには、故障を自動的に検出し、故障箇所を迂回してファイルシステムのサービスを持続させることが重要である。

このため、FEFSはハードウェアを二重化し、ハードウェア故障時にソフトウェア制御によって仕掛中のサービスを正常に回復させるFailover処理を行い、サーバおよびI/O通信の経路を切り替えることで、一つの障害でファイルシステムのサービスが停止することなく、システムの運用が継続できるようにしている（図-3）。

階層的なノード監視と自動交替

大規模システムでは、人手によらない故障の検出と、その故障が影響する範囲のノードへの交替通知を自動化することが必須である。計算ノードとファイルサーバ間で監視パケットを飛ばして

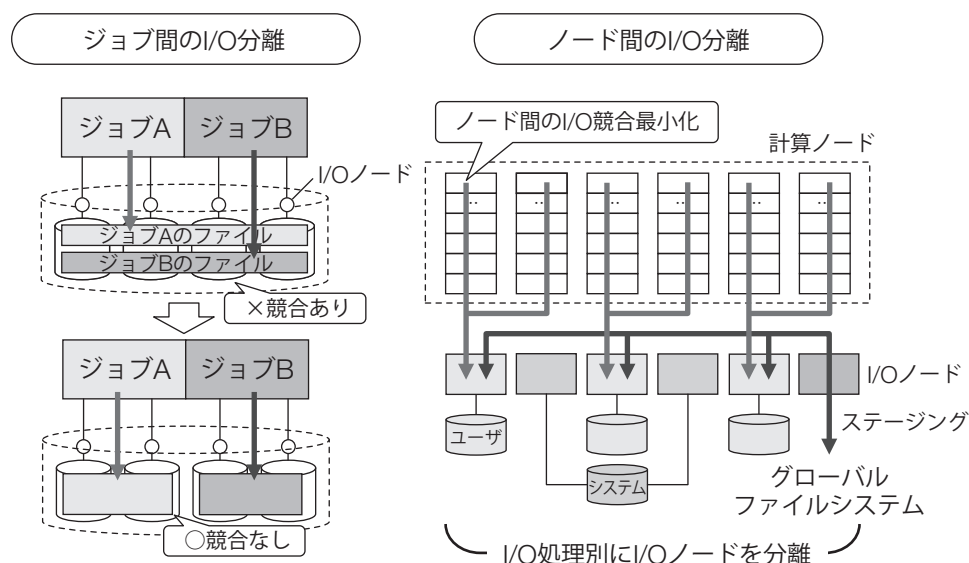


図-2 I/O分離による競合排除

ノード状態を監視する方式では、規模のべき乗に比例して監視パケットが大量に発生し、計算ノード間のMPI通信および計算ノードとファイルサーバとの間のデータ通信を阻害する。

FEFSでは、この対策としてシステム管理ソフトと連携して階層的なノード監視と交替の制御を行い、通信負荷を最小化している。計算ラック内の監視、複数ラックをグループ化したノードグループ単位での監視、その上位のノードグループの監視とツリー上に監視を行うことで効率的な自動交替を実現している（図-4）。

### 運用に応じたQoSポリシー選択

センター運用のように多数のユーザが利用する大規模システムでは、特定のユーザが大量にファイルI/Oを行った場合でも他ユーザに悪影響を与えないことが求められる。また、計算ノードのジョブからファイルアクセスが行われている場合でも、

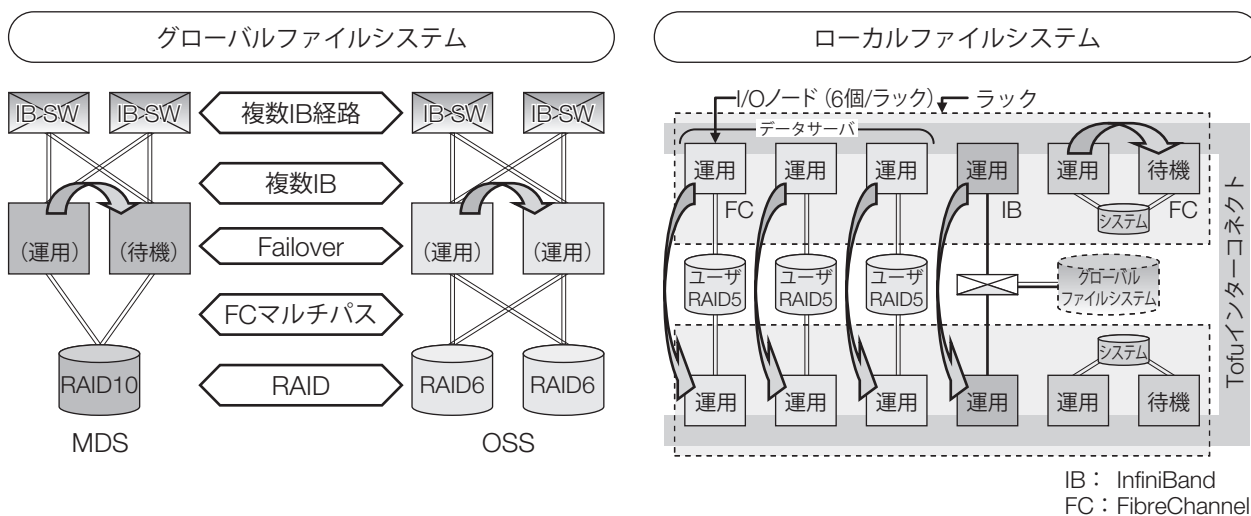


図-3 ハードウェア二重化による耐故障性向上

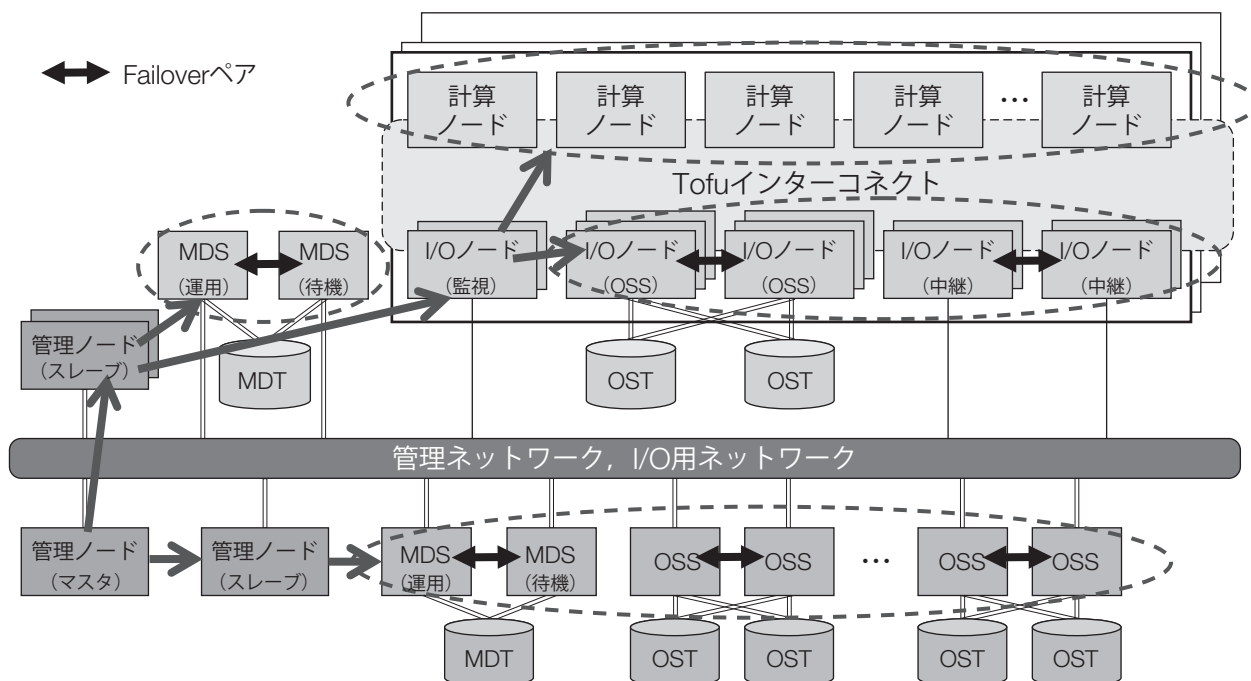


図-4 階層的なノード監視と自動交替



ログインノード上のユーザのレスポンスに影響しないことが求められる。

FEFSは、ユーザ間のファイルI/Oのフェアシェア機能と、TSS（Time Sharing System）環境におけるレスポンス保証の機能によりこれらの課題を解決している。

#### (1) 1ユーザが発行できるI/Oの上限管理

FEFSでは、クライアントとサーバでそれぞれ発行・処理するI/Oリクエスト数を制御し、特定ユーザによるI/O資源（I/Oネットワーク、サーバ、ディスク装置）の占有を防止する（図-5）。

クライアント側では、1ユーザが同時に発行できるI/Oリクエスト数の上限を制限し、1ユーザから大量にI/Oリクエストが発行されI/O資源を占有してしまうことを抑止する。

計算ノードのように、複数クライアントから同じユーザのアプリケーションが同時にI/O要求を発行すると、I/O資源が占有される可能性がある。このため、ファイルサーバ側で1ユーザが利用できるサーバ処理能力を制御し、1ユーザからのI/O要求

でI/O資源が占有されることを防止する。

#### (2) ログインノードのレスポンス保証

ユーザが体感するアクセスレスポンスは、システムの使いやすさに直結するため、ジョブに対するアクセスレスポンスよりも重要である。

FEFSは、ログインノードでTSS利用するユーザのアクセスレスポンスを保証するため、ログインノードからのI/O要求を処理するサーバ資源を割り当てる機能を持つ。これにより、計算ノードのジョブがファイルI/Oを行っていても、ログインノードでファイル操作するユーザのレスポンスを確保できる。

#### (3) I/O帯域のベストエフォート活用

サーバ資源を余さず有効活用できるように、ベストエフォート型の運用を可能とした（図-6）。ログインノードと計算ノードからのI/O要求がある場合には2者で全ノード資源を共有する（図-6左）。また、計算ノードからのI/O要求がない状態ではログインノードが全サーバ資源を利用する（図-6右）。

これらの要求の程度はシステム運用者ごとに異

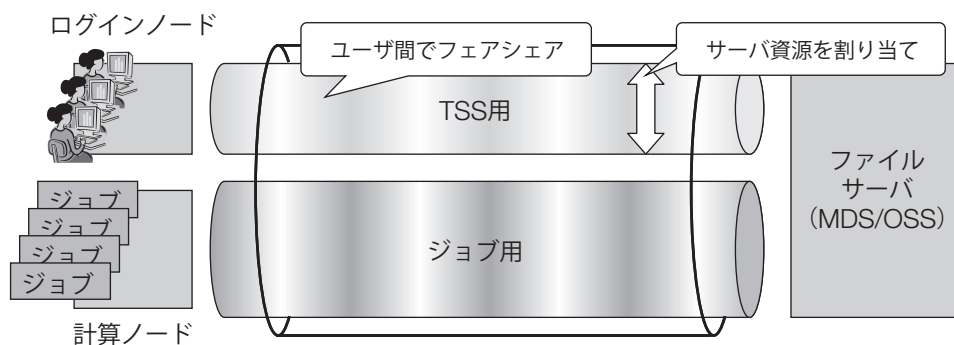


図-5 ログインノードのレスポンス保証

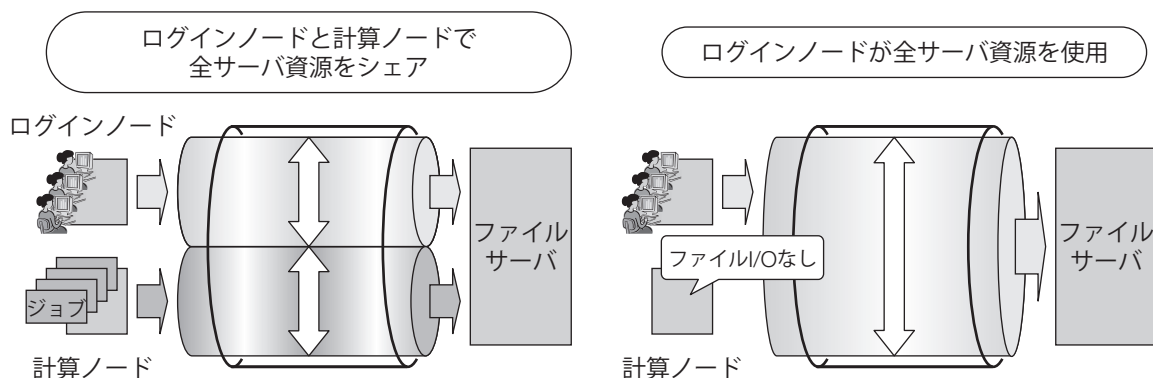


図-6 I/O帯域のベストエフォート型利用

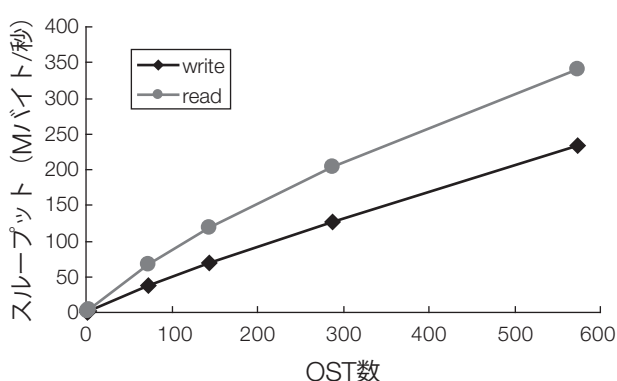


図-7 スループット性能(IOR)

なるため、FEFSではそれぞれの保証とその程度(QoSポリシー)を選択可能としている。

### FEFSのデータ転送性能

現在、「京」は開発中であり、FEFSについても徐々に規模を拡大しながら運用試験を続けている。システム全体9列中2列分規模のファイルI/O性能を図-7に示す。本性能をIORと呼ばれるI/Oの標準ベンチマークを用いて測定した。測定の結果、580サーバを並列I/Oすることで340 Gバイト/秒のファイル参照性能を達成している。これは、現時点での世界トップクラスの性能である。

### む す び

本稿では、スーパーコンピュータ「京」向けに開発したクラスタ型分散ファイルシステム「FEFS」について紹介した。FEFSは数千台規模のファイルサーバを束ね、100 Pバイト級の大容量とTバイト/秒級のI/O性能を持つ高性能なファイルシステムをも実現する。また、「京」の運用管理ソフトと連携することで、人手に頼らない効率的なノード監視・自動交替の仕組みを提供する。さらに、大規模システムを共用する利用者間やジョブ間のファイルアクセスによる相互影響を低減するためのQoS機能により、ファイルアクセスの使いやすさも向上させる。

今後も、FEFSの性能と信頼性向上のための開発を継続し、「京」の安定運用に努める一方、富士通は「京」のノウハウを商用スーパーコンピュータ「PRIMEHPC FX10」やPCクラスタシステムにも適用し、ミドルエンドからハイエンドのスーパーコンピュータのファイルシステムを提供していく。また、「京」での開発成果はLustreの開発コミュニティであるOpenSFSやLustre専門開発企業であるWhamcloud社と協力し、将来のLustreに反映させ標準とするよう作業を進めていく。

### 著者紹介



#### 酒井憲一郎 (さかい けんいちろう)

次世代テクニカルコンピューティング開発本部ソフトウェア開発統括部 所属  
現在、スーパーコンピュータ向け大規模ファイルシステムのソフトウェア開発に従事。



#### 黒川原佳 (くろかわ もとよし)

独立行政法人理化学研究所次世代スーパーコンピュータ開発実施本部 所属  
現在、理研においてシステム全般の開発に従事。



#### 住元真司 (すみもと しんじ)

次世代テクニカルコンピューティング開発本部ソフトウェア開発統括部 所属  
現在、MPI通信ライブラリ、クラスタファイルシステムなど高性能通信に関わるHPCシステムソフトウェア全般の技術開発に従事。