速さと効率の新時代：xFusion DFSがスーパーコンピューティングを支える理由

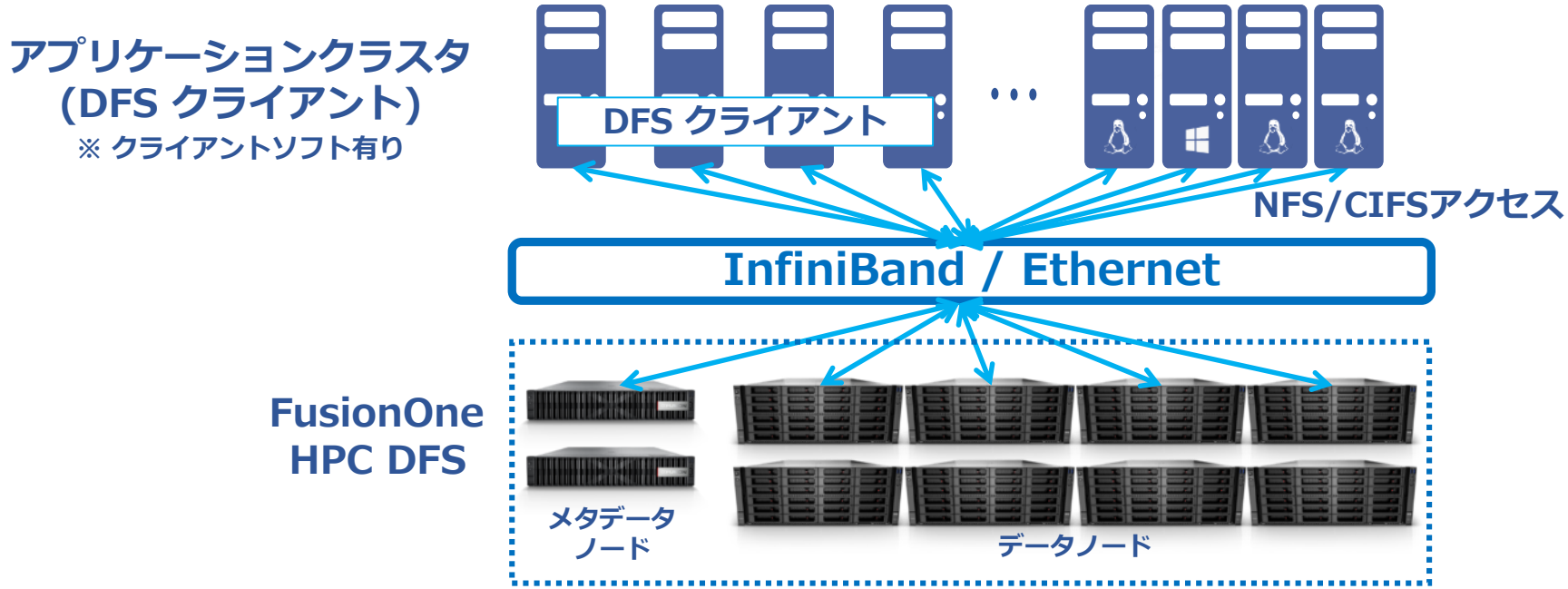
FusionOne HPC DFS 分散ストレージソリューション

xFusion技術日本株式会社　製造・公共営業本部
シニアシステムズエンジニア　伊東浩之
2025年2月４日

FusionOne HPC DFS とは

メタデータノード、データノード、ストレージソフトウェアで構成され
計算サーバにインストールするクライアントソフトによりRDMA接続も実現する

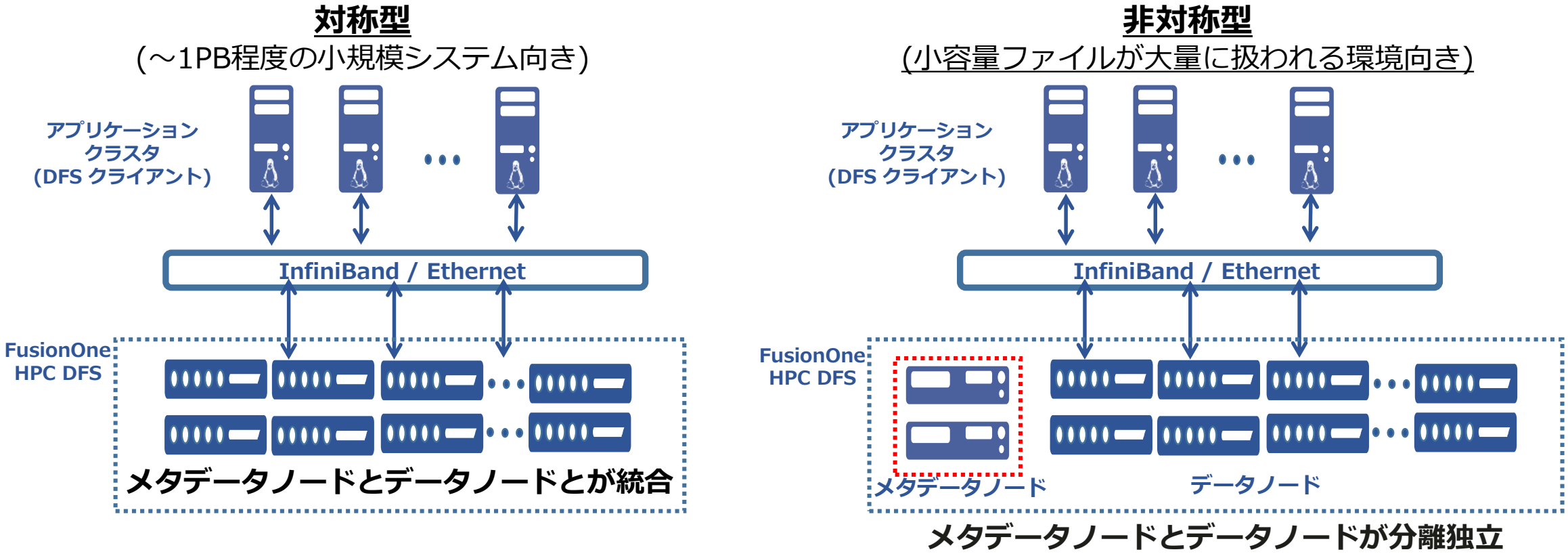
並列分散ストレージシステム



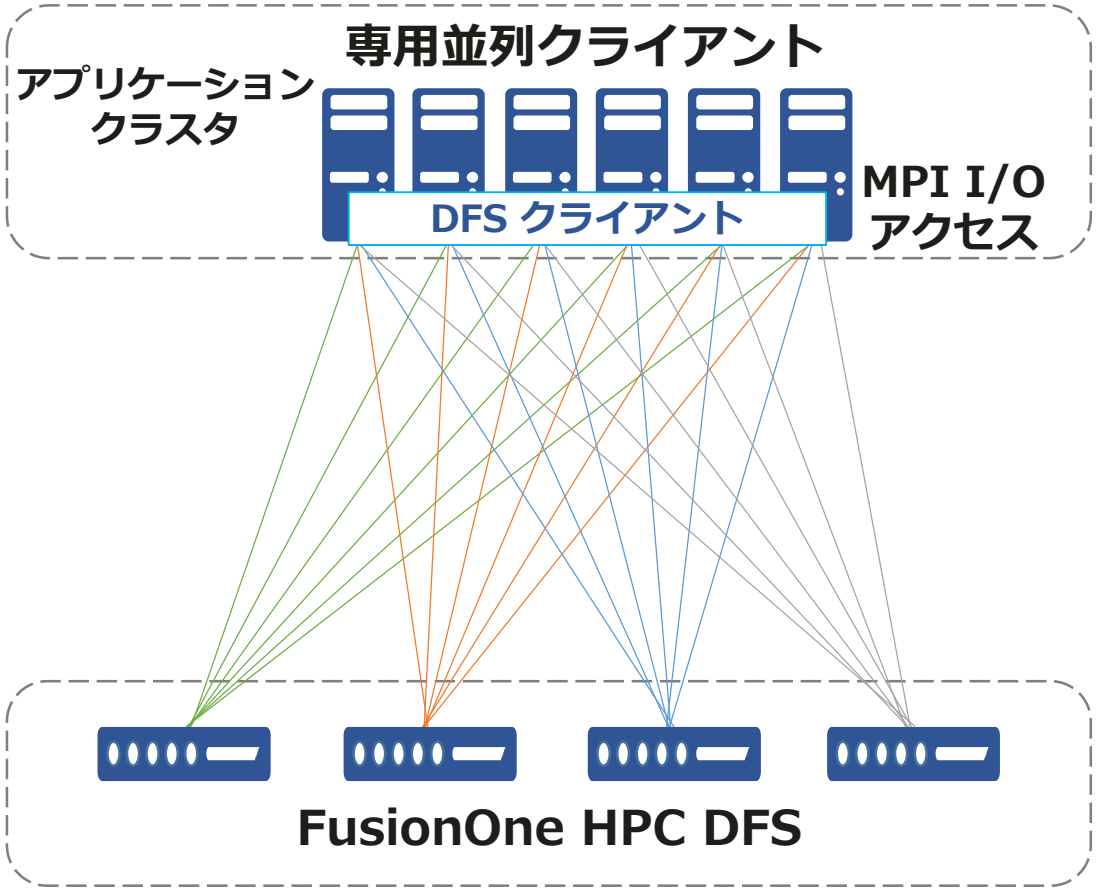
FusionOne HPC DFS とは

FusionOne HPC DFSはアーキテクチャとして対称型と非対称型が選択可能で、ストレージの冗長性にはイレージャコーディングを採用。

ユーザのご利用形態に応じて柔軟なストレージシステムを構成可能



高速性：専用並列クライアントにより、高性能並列アクセスを実現



多重並行処理による高いパフォーマンス

- アプリケーションクラスタ専用の並列クライアントにより、すべてのデータノードへ接続し、負荷分散した同時データ読み出し/書き込みを実現
- パフォーマンスのボトルネックがなく、専用クライアントはアプリケーションクラスタに応じて無限に拡張でき、パフォーマンスもリニアに向上

MPI I/O対応

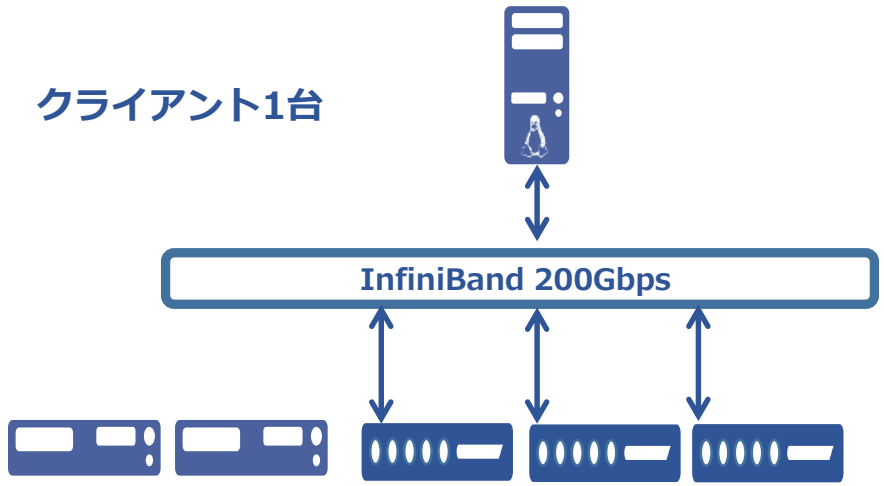
- 専用の並列クライアントは、POSIX標準 API に基づいて MPI I/O を実装し、さまざまな MPI I/O ライブラリへの基盤となるファイルアクセスを提供

主要OSとのクライアント互換性

- 並列クライアントは、Windows や Liunx などの主要な OS と互換性があり、高いパフォーマンスのデータアクセスを実現

高速性：パフォーマンス例

サンプル構成(非対称型EC4+2)



メタデータノード2台

(1台あたり)
CPU：Xeon8368 * 2
MEM：256GB
SSD：OS用SATA * 2
SSD：メタデータ用 NVMe * 2
IB：200Gbps HCA

データノード3台

(1台あたり)
CPU：Xeon5318Y * 2
MEM：256GB
SSD：OS用SATA * 2
SSD：データ用1.6TB NVMe * 16
IB：200Gbps HCA

性能例(Blocksize:4K・ご参考)

非同期ランダムREAD(シングル)：

約230,000 IOPS

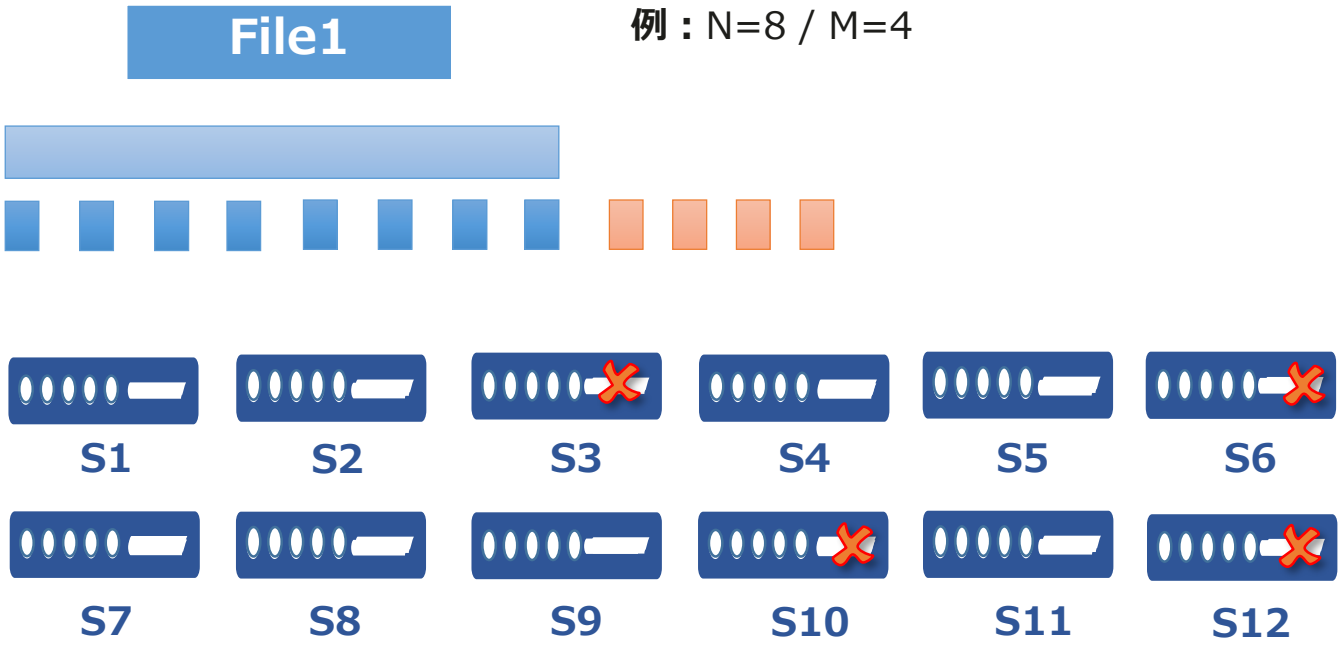
非同期ランダムREAD(マルチ)：

約700,000 IOPS

※上記性能例は、あくまでもサンプル構成による参考値であり、性能を保証するものではありません。

高効率：Erasure Coding(EC)による高効率と高信頼の実装

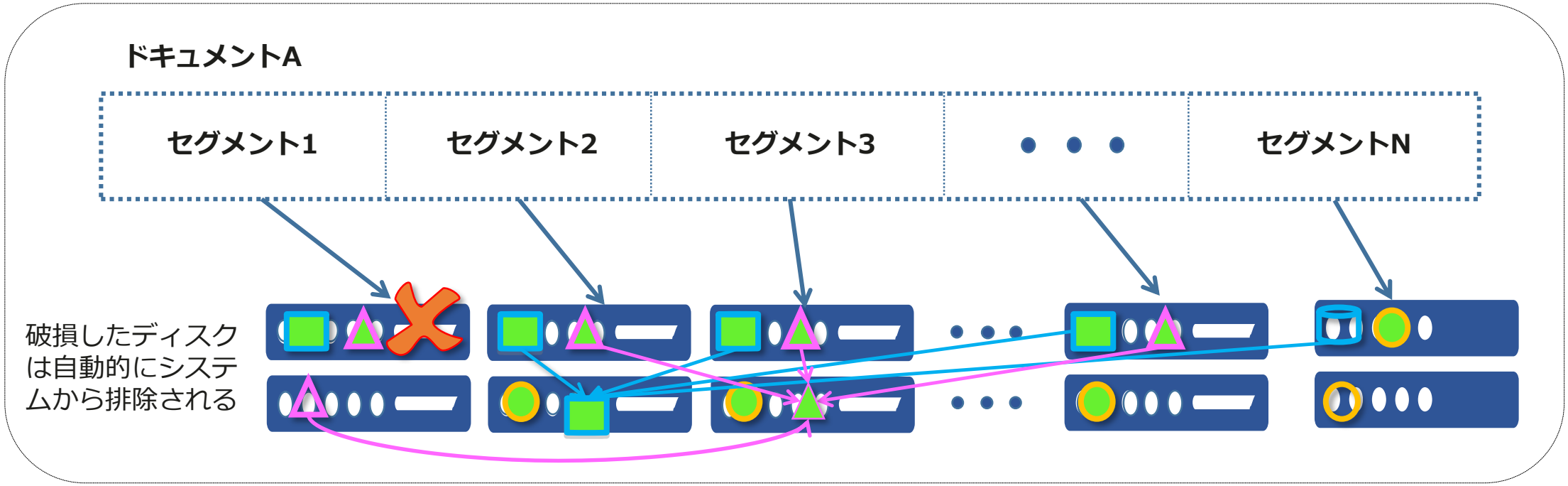
データブロックからパリティブロックを生成し、そのパリティを別々のディスクに保存することで複数台ディスクが壊れても復旧可能な状態を保持。近年RAIDに変わる技術として広く採用されてきている。



	EC	レプリケーション
信頼性	高い	非常に高い
容量使用率	高い (66%-94%)	低い (25%-50%)
パフォーマンス	高い	低い
シナリオ	ビッグデータ、 HPCなど	データベース、 信頼性重視

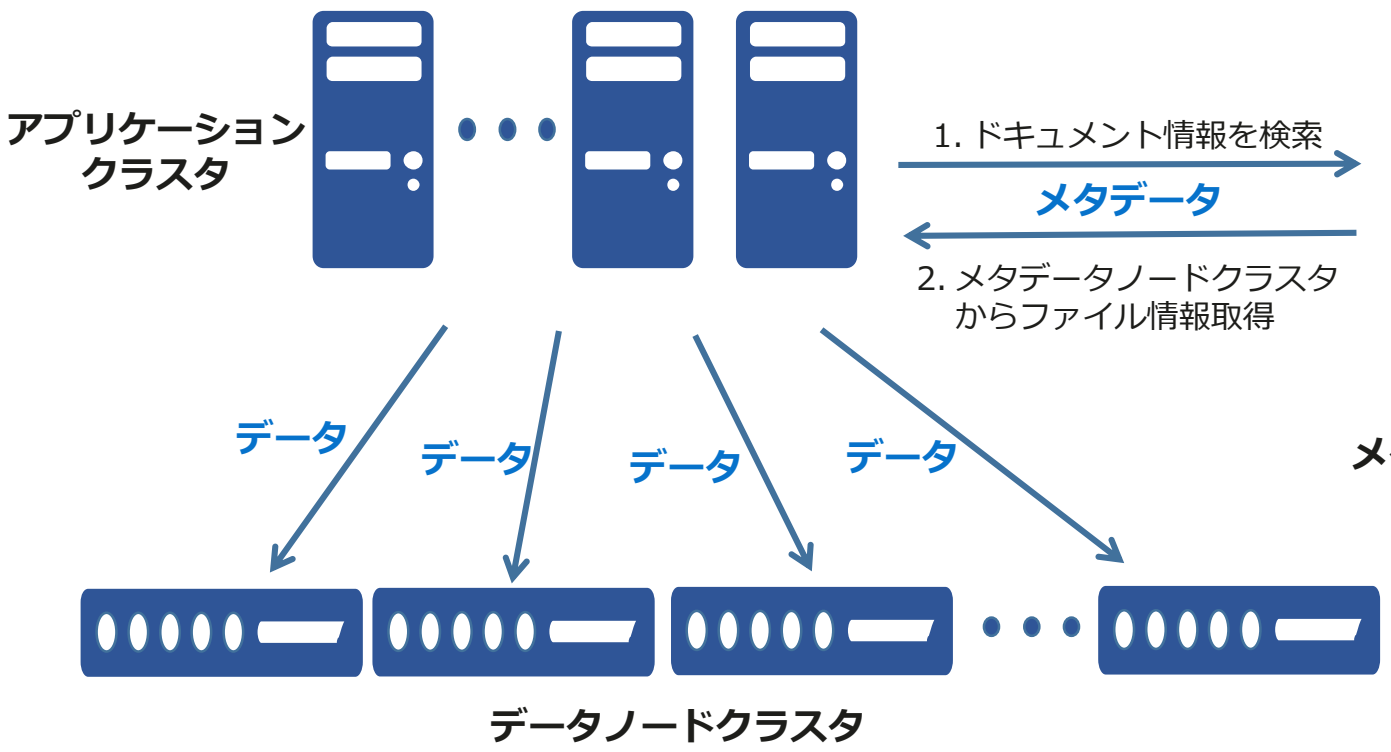
最大4ノード障害でも
サービス中断なし

高効率：高いデータ復元効率と高い復元性能　1TBデータ復元＜30分



1. クラスタ化されたストレージデバイスにノードレベルの障害回復プロセスを適用し、ノードがダウンしても、業務の中断は無し
2. ストレージプールの全ノードがデータ再構築プロセスに参加し、複数ディスクから複数ディスクへ再構築し、30分以内に1TB のデータ再構築が可能

高効率：メタデータノードのマルチノード化によるボトルネック解消



メタデータノードのクラスタリング技術により、オンデマンドの拡張に対応

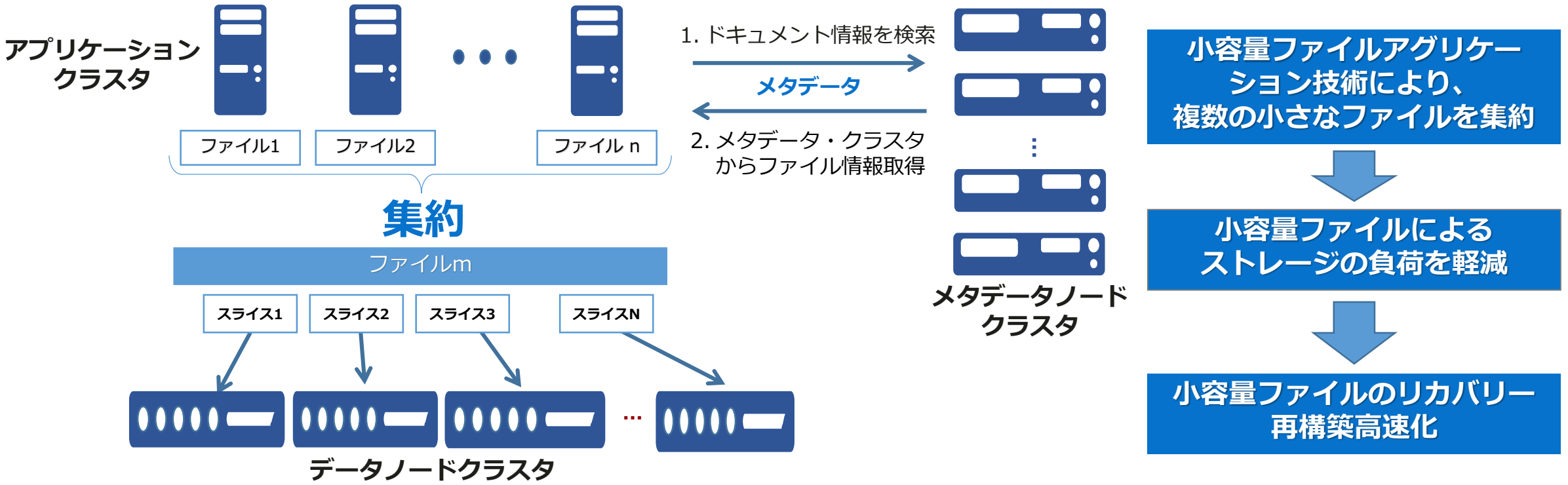
マルチノード・マルチライブモードにより、シングルノードのボトルネックを解消

ディレクトリやファイル数の制限をなくし、ファイルシステム全体で最大5,000億ファイルまで対応

1. アプリケーションクラスタは、まずメタデータクラスタを通じてファイルの位置情報などを取得し、そのファイル情報に基づいてデータクラスタ内のデータに直接アクセス
2. アプリケーションクラスタは、ファイルクライアントを介してファイルをスライスし、ECポリシーに従って複数のフラグメントに格納し、各データノードが同時に読み書きする。これにより、ファイルがデータノードに書き込まれた後に、スライスされて他のデータノードに分散される必要がないため、ストレージ効率と性能が改善される

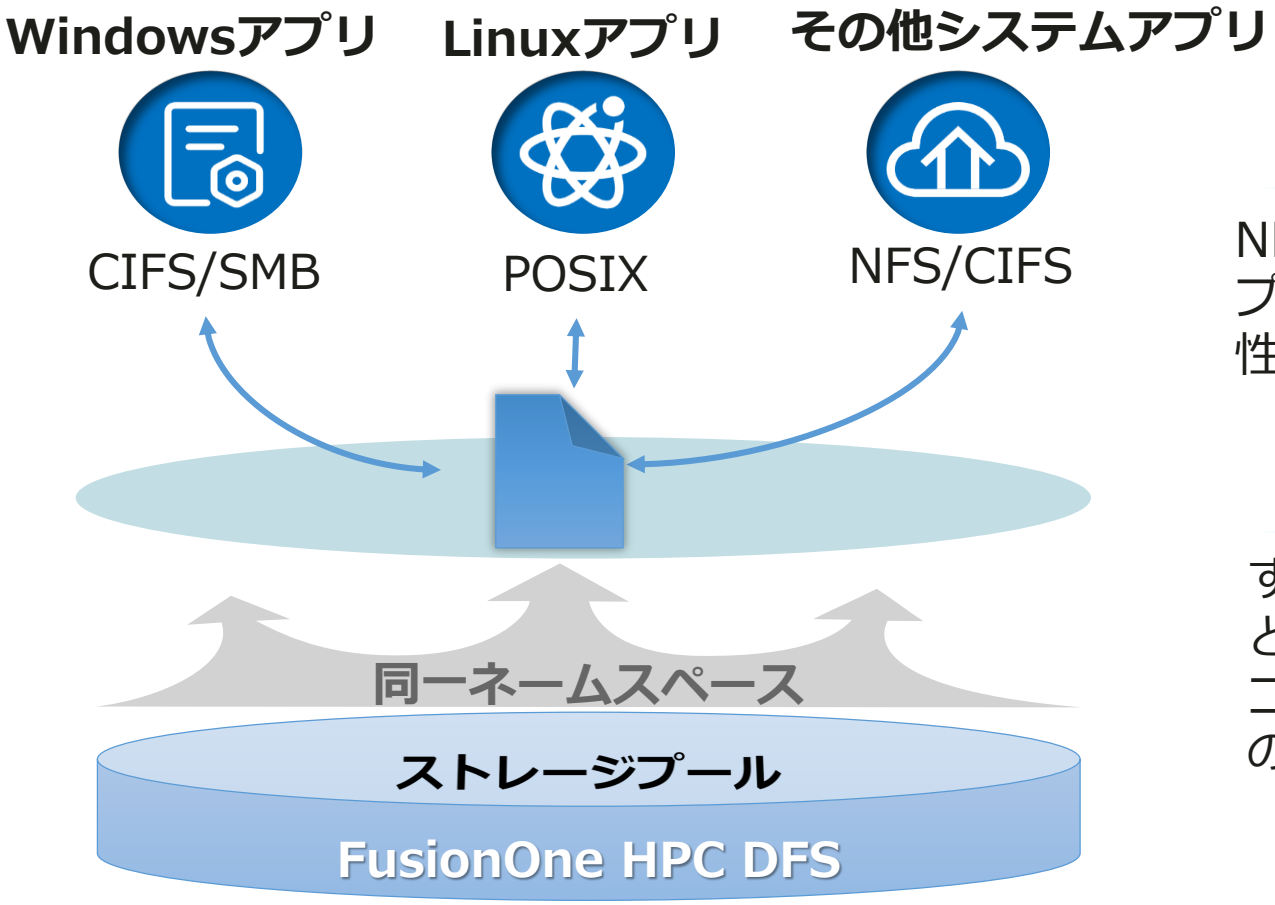
高効率：クライアント側の小さなファイルを集約

クライアント側の小さなファイルを集約し、IO ドロップとデータ読み書き回数を低減



1. 小容量ファイルのアクセス効率を高めるために、小容量ファイルのアグリゲーション転送・保存機構を提供
2. 複数の小さなファイルを1つのデータブロックにまとめて転送することで、データの読み書きの負担を軽減し、帯域幅の消費を回避

高効率：マルチプロトコル互換、データ移行なくアクセス効率の向上



マルチプロトコル互換

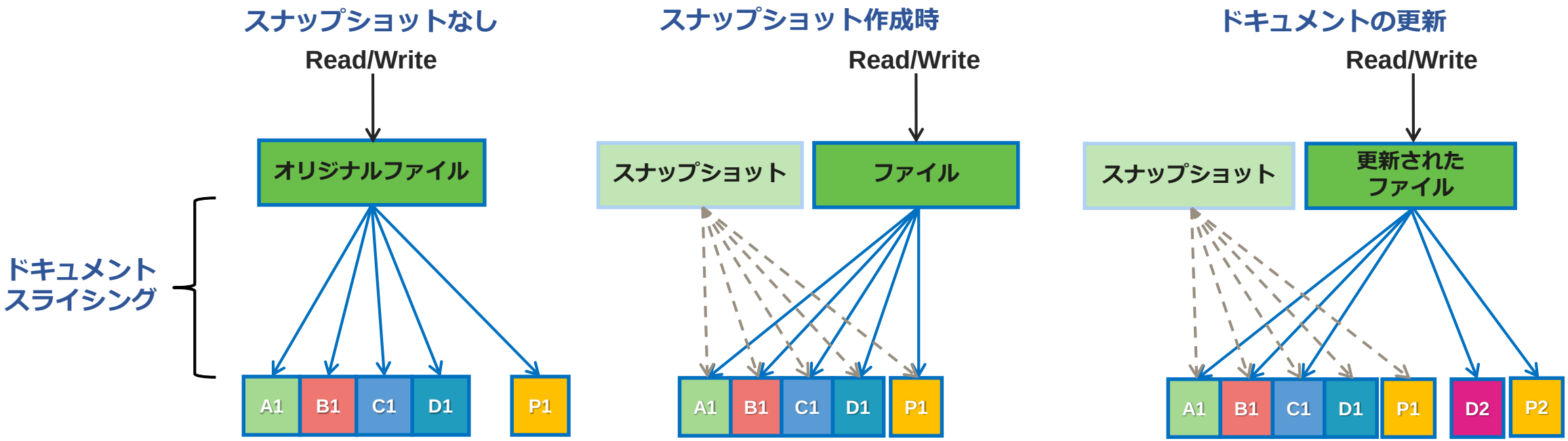
NFS、CIFS/SMB、プライベートクライアントプロトコルなどさまざまなプロトコルとの互換性を持つ

データ共有の効率化

すべてのファイルデータはさまざまなプロトコルと互換性を持ち、データを共有し、マルチプロトコルアクセスにより、データ移行不要で、コピーの重複保存を避け、アクセス効率を向上

高効率：ROW方式スナップショット

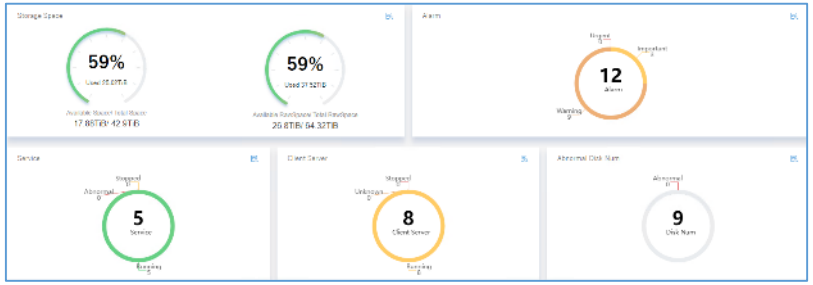
パフォーマンスに影響を与えず省スペースを実現



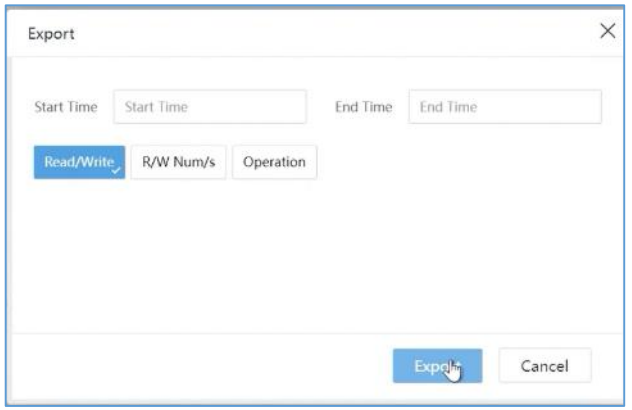
1. ROW方式スナップショット、メタデータ管理によりスナップショットの断片化問題を解消。COW方式と比較してデータコピーと書込み増加がなく、データ書き込み性能への影響無し
2. 多数のスナップショットの共存をサポートし、スライスベースのスナップショットはスナップショットの粒度が小さく、省スペース
3. 瞬時の作成と実行

高効率：標準で組み込まれた管理・監視ツールやAPI連携

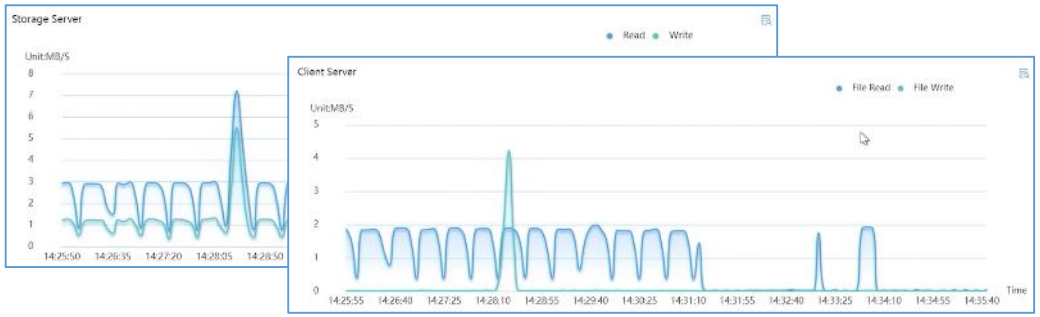
DFS管理ツール画面でストレージ状況をモニタリングすることが可能で、監視情報は時間範囲を指定してエクスポートすることが可能
また、APIも用意されており、Prometheus、Grafanaなどの外部監視ツールとの連携も可能



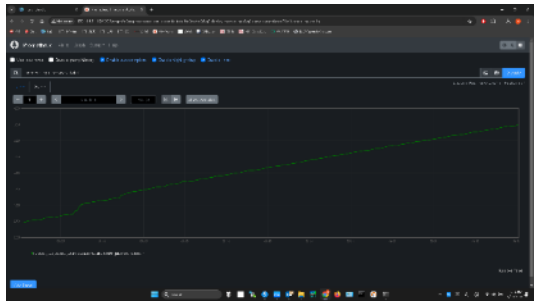
DFS管理ツール画面例。ストレージ空き容量、アラーム状況、サービス状況、クライアント状況、異常ディスクがモニタリングされる



監視情報は時間範囲を指定してエクスポート可能。また、APIが用意されていますので、外部監視ツールとの連携も可能

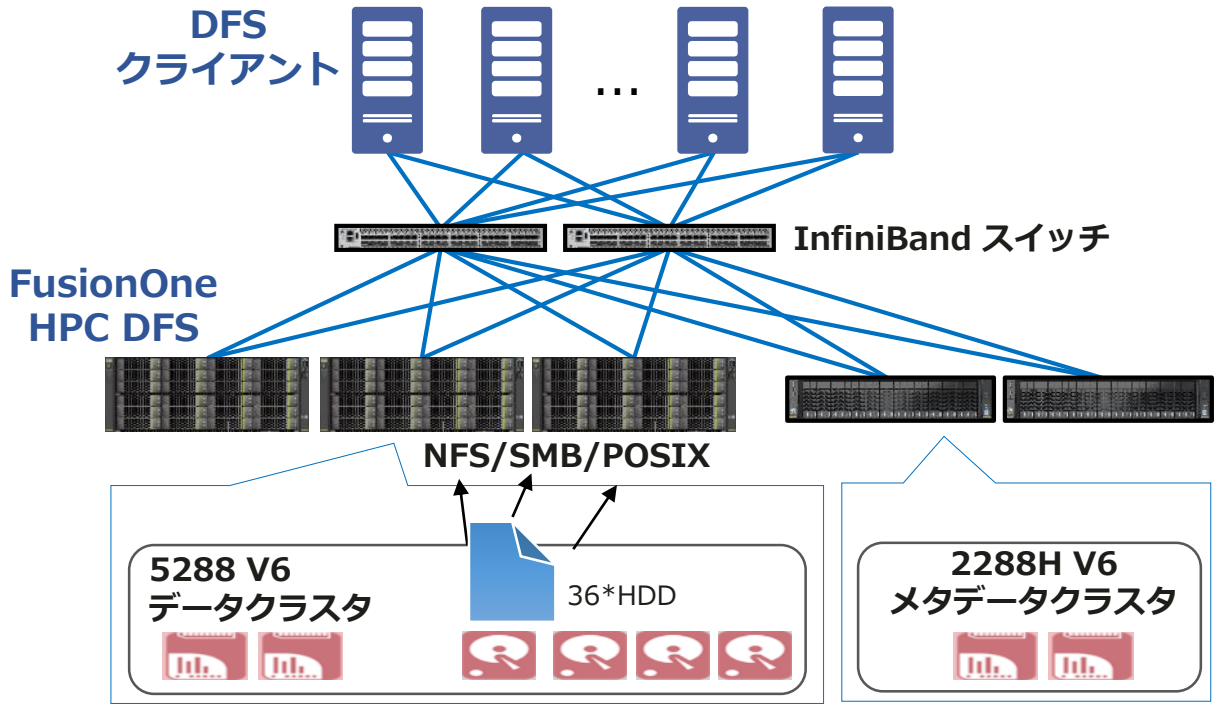


ストレージサーバおよびクライアントサーバのIO帯域幅のリアルタイム監視上協をモニタリング可能



Prometheusなどの外部連携も可能

FusionOne HPC DFS 構成例(非対称型)



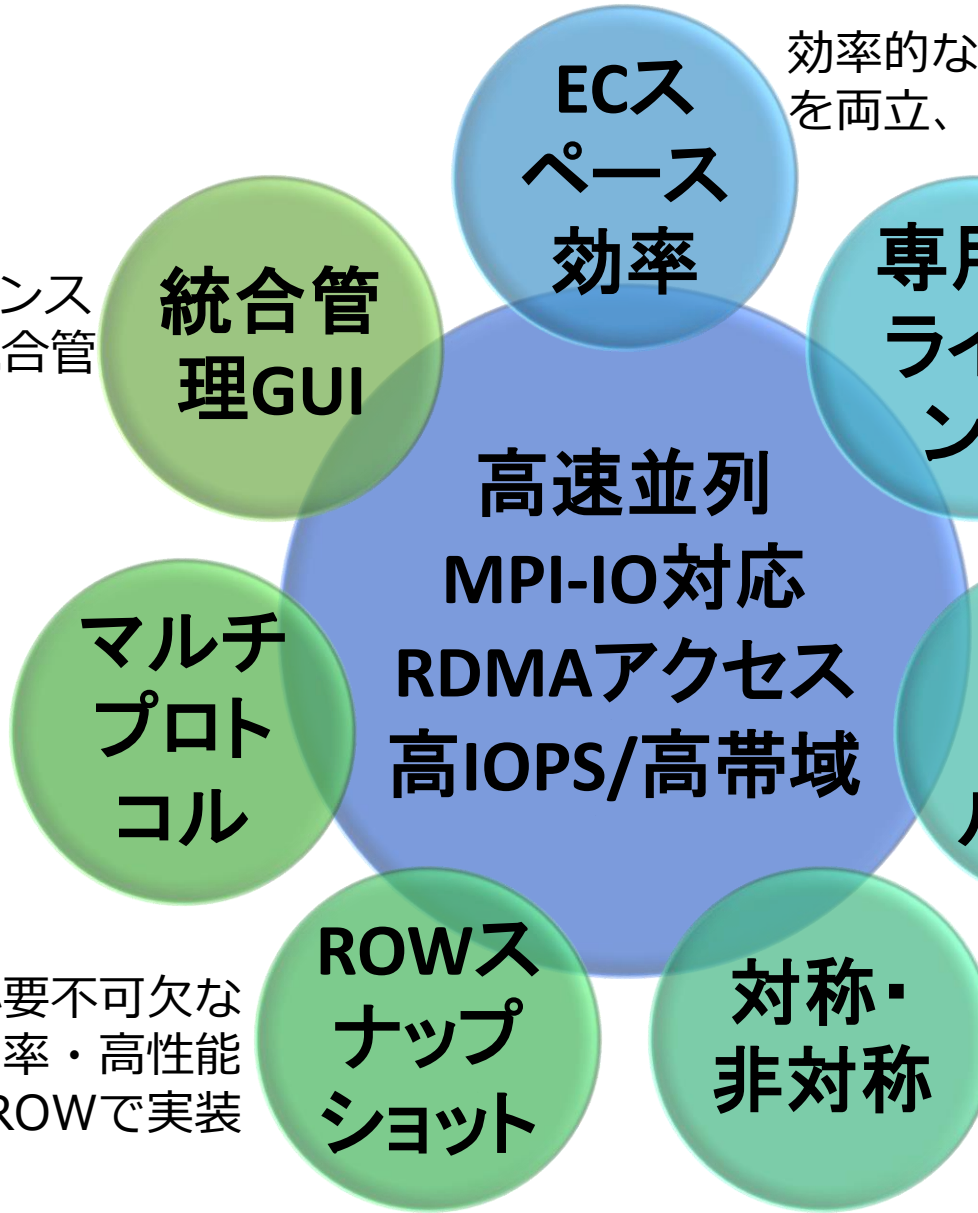
ネットワーク構成：3つのネットワークを構成

- ✓ ストレージネットワーク：InfiniBand で構成し、ファイルサービスを提供
- ✓ 帯域外管理ネットワーク：1*IPMI ポート (リモート電源 ON/OFF)
- ✓ サービス管理ネットワーク： 2*1GE ポート

FusionOne HPC DFS 構成

コンポーネント	説明
5288 V6 (データノード ≧3)	➢ 5288 V6 (36*3.5インチ HDD、シングルRAID シャーシ) ➢ 2*Intel 4314、256GB RAM ➢ システムディスク：2*240GB SSD(オプション) ➢ データディスク：36*HDD(SATA)、6T/8TB/10TB/12TB/14TB/16TB が利用可能 ➢ EDR(100Gb/s) シングル/デュアルポート IBカード ➢ 1*9460-8i-PCIe RAID カード-2GB
2288H V6 (メタデータノード≧2)	➢ 2288H V6(8*2.5インチ HDD エンクロージャー) ➢ 2*Intel 4314、256GB RAM ➢ システムディスク：2*240GB SSD(オプション) ➢ メタデータディスク：2*960GB SSD ➢ EDR(100Gb/s) シングル/デュアルポート IBカード ➢ 1*9460-8i-PCIe RAID カード-2GB
DFS ソフトウェア	➢ DFS ストレージソフトウェアライセンス(物理容量) ➢ DFS ソフトウェア SNS：1/2/3/5年(物理容量)
IB モジュール	InfiniBand 構成では、構成数は IB カードのインターフェース数と同じになります。
OS	Rocky Linux9.3

FusionOne HPC DFS のメリット



効率的なスペース利用と高信頼性を両立、高速リカバリ

高速性を実現しながら全ノードに効率よくデータ&パリティを分散配置

大容量ファイルと小容量ファイルのIO特性を同時に対応可能

柔軟な展開構成により必要に応じて効率の良い拡張が可能
SSD/HDDの混在構成も可能

ファイルサーバとして必要不可欠なスナップショットを高効率・高性能なROWで実装

データの無駄なコピーを排除し効率よく複数のプロトコルでアクセス可能

xFusion独自のアプライアンスで効率の良いHW/SWの統合管理・監視が可能

Thank you.

https://www.xfusion.com

xFusion

Let Computing Serve You Better