サイエンティフィック・システム研究会　　2011 年 10 月 19 日
2011 年度科学技術計算分科会 会合

世界トップクラスシステムに相応しい 超大規模ストレージを目指す 「京」＊のストレージシステム

住元 真司

富士通株式会社 次世代TC開発本部 ソフトウェア開発統括部

[アブストラクト]

「京」のストレージシステムは世界トップクラスのシステムに相応しく、かつ、次世代のエクサバイトクラスでも利用に耐えるものとなることを目標に、ハードウェアとソフトウェアの開発を進めている。本講演では、大規模ストレージシステムでは何が問題になり、「京」のストレージシステムでは、どのように解決しようとしているのかを米国での大規模センターでのストレージ事例を織り交ぜながら議論したい。

＊2010 年 7 月に理化学研究所様が発表した「次世代スーパーコンピュータ」の愛称です

[キーワード]

HPC、 大規模ストレージ、クラスタファイルシステム、Lustre

1. はじめに

理化学研究所様と共同開発を進めている、スーパーコンピュータ「京」は、最先端の 100 万コアクラスの超並列スーパーコンピュータである。このような大規模なシステムを支えるストレージシステムは、IO ノードと RAID システムがそれぞれ数千台規模と、従来と一線を画した大規模なストレージになる。「京」の 2 階層のストレージを設計するにあたり、京の演算性能を最大限に引き出す、様々なユーザが大規模データをストレスなく扱える、高度なセンター運用に耐える、という３つの要件を満たし、エクサの時代にも使いものになるストレージ設計を目標とした。

2. 海外の大規模ストレージ事例調査

当時の日本においては、ペタスケールの計算システム上でのストレージシステムに対する事例がなかったため、海外の大規模ストレージ事例や調査事例を調べた。この結果、欧米では、単なるストレージだけではなく、シミュレーションを実行する全体のプロセス処理の効率化を目指した、科学データ管理(Scientific Data Management)とう考え方を元に、データ処理を階層化して、ツール、データライブラリが研究開発されていることがわかった。また、大規模データを扱うためのクラスタファイルシステムと並列データ処理を行う MPI-IO と HDF5, netCDF のデータライブラリから構成されており、これらがこれからの大規模ストレージの重要なコンポーネントとなることがわかった。

3. 「京」のストレージシステム設計

海外の大規模ストレージの事例調査から判明した実現課題として、TB/s クラスのファイル IO 性能と運用機能、ジョブへのファイル IO の影響を徹底的に排除、単一障害に耐える信頼性、計算機センターとしての運用性確保、の４つを設定して、「京」のストレージ設計にあたった。

「京」のファイル IO アーキテクチャとしては、ジョブ利用と共用利用を独立したストレージとして配置することにより相互のファイル IO の影響を排除したほか、Z 軸直下にストレージを均等に配置し、ファイル IO も可能な限り直下に実施されるようファイルのステージングを実施することにより、ジョブ間でのファイル IO の影響を抑えている。

また、ファイルシステムは様々なシステムから利用可能なように業界で標準的に利用されているファイルシステムが望ましいと考え、Lustre ベースにエクサバイトクラスまで対応可能なように仕様を拡張している。この拡張が施されたファイルシステムは FEFS と名付けられ、「京」の２階層ストレージの両方の階層に用いられる。FEFS に施された Lustre の拡張には、大規模化、スケーラビリティ、センター運用向け、性能向上、高可用性、使いやすさの６つの観点からの拡張がある。これらの拡張については将来の Lustre に反映されるよう Lustre コミュニティに contribution していく予定である。

4. まとめ

「京」のストレージシステム設計について、その課題と設計について述べた。現在、「京」のシステムソフトウェアは開発中であるが、ストレージシステムとしても、現在世界で最も大規模なものとなっている。きちんとシステムを安定稼働させ、これからの大規模ストレージのレファレンスとなれるよう開発に取り組みたい。

[参考文献]

(1) 次世代スパコン「京(けい)」のコアテクノロジ ， 追永勇次，応用物理、第 80 巻、第 7 号、p.0590-0593 (2011)