広域分散ファイルシステム Blobseer-wan/HGMDS の設計と初期評価

鷹津冬将，平賀弘平，建部修見（筑波大学），Gabriel Antoniu (INRIA)

1 背景

近年データインテンシブコンピューティングなどの分野では，複数の拠点間で効率的に扱う広域データ解析の要求が高まっており，広域に分散した拠点間において大量のデータを共有するために，広域分散ファイルシステムが注目されている．しかし既存のファイルシステムでは各サーバ間の通信がネットワークにおける遅延が非常に大きいことがボトルネックとなる．

2 設計

Blobseer-wan は，分散データストレージである Blobseer [1] を広域向けに現在も開発されている広域分散ストレージである．

HGMDS [2] は，広域分散ファイルシステムのために実装された Multi-master のメタデータサーバであり，拠点間の遅延が大きい環境においても高いメタデータ操作性能を示す．

先述の Blobseer-wan と HGMDS を用いて Blobseer-wan/HGMDS を設計した．その構成を図 1 に示す．図に示されるようにクライアントやメタデータサーバを複数台設置することができ，各クライアントは使用するメタデータサーバをひとつ指定し，さまざまなオペレーションに対して処理を行う．

3 評価

評価には mdtest [3] を使用し，シングルノードにおいて 1 秒間に実行できるオペレーション数を 3 種類のオプションをつけて評価した．一つ目（引数なし）はファイルを作成し stat 構造体を取得した後削除する．二つ目（-w 4096）は作成した後，ファイルの場合にはデータを書き込む．三つ目（-w 4096 -y）はデータを書き込んだ場合に sync を行う．

ストレージノードのファイルシステムとして ext4 を使う Gfarm を同じオプションで評価を行い比較した．

評価した結果を図 2 に示す．図 2 のグラフでは縦軸は単位時間あたりのオペレーション数を示している．

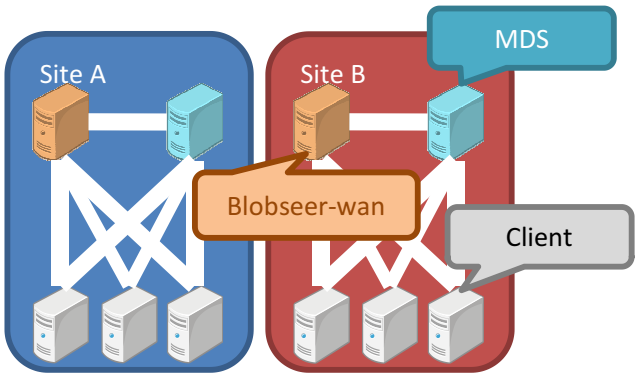

図 1 Blobseer-wan/HGMDS の構成

4 まとめと今後の課題

本稿では，広域分散ファイルシステム Blobseer-wan/HGMDS の設計・実装し，初期評価としてシングルノードにおける性能評価について述べた．シングルノードにおける性能評価では，期待していた性能が発揮されなかった．

本稿ではシングルノードにおける IO 性能のバンド幅が調査できていない．また，一つのノード上ですべてのプロセスを動かしているが，実際には高遅延環境において各プロセスを個別のノード上で動かすことになり，その場合に高い性能が出る設計となっている．

そこで，今後の課題としてはシングルノードにおける IO におけるバンド幅の調査や，複数台における各性能の評価，高遅延環境における評価などが挙げられる．

謝辞

本研究の一部は JST-ANR FP3C による．

参考文献

[1] Bogdan Nicolae, Gabriel Antoniu, Luc Bougé, Diana Moise, and Alexandra Carpen-Amarie. BlobSeer: Next Generation Data Management for Large Scale Infrastructures. *Journal of Parallel and Distributed Computing*, Vol. 71, No. 2, pp. 168–184, February 2011.

[2] 平賀弘平, 建部修見. 広域ファイルシステム HGFS のための分散メタデータサーバの実装と性能評価. 情報処理学会研究報告. [ハイパフォーマンスコンピューティング], Vol. 2010, No. 29, pp. 1–9, jul 2010.

[3] mdtest HPC benchmark. `http://mdtest.sourceforge.net/`.

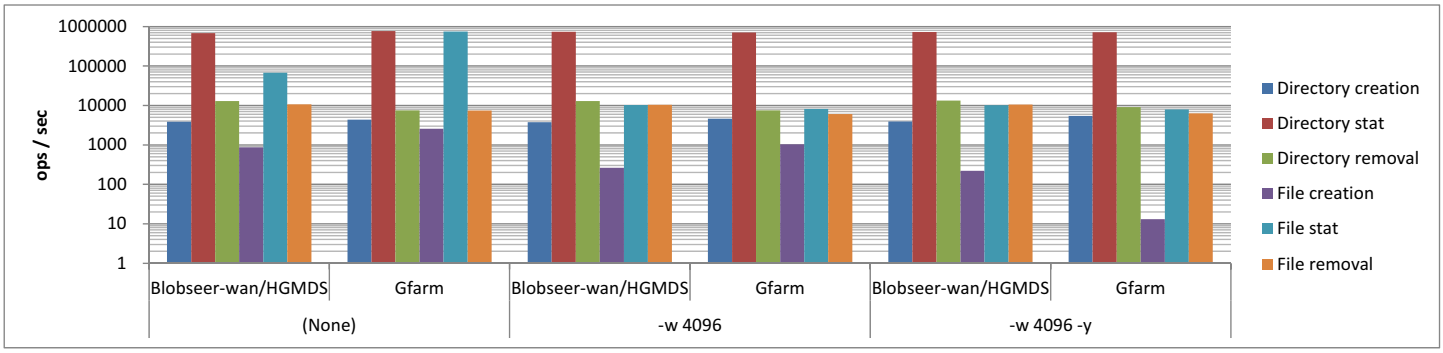

図 2 mdtest の結果