1K-04

マルチラック環境における HDFSの効率的なレプリカ再配置手法の提案

日開 朝美† 竹房 あつ子‡ 中田 秀基‡ 小口 正人†
†お茶の水女子大学 ‡産業技術総合研究所

1. はじめに

Hadoop Distributed File System(HDFS) では，ノードが故障するとデータノード間でレプリカ再配置処理を行いデータの可用性を維持する．しかしながら，HDFS のレプリカ再配置では，生成元・生成先をランダムに選択するため，データ移動に偏りが生じ非効率な再配置処理が行われている．この問題を解消するために，我々は指向性リング構造に基づいたデータ転送により各ノードの負荷を均衡化する生成元・生成先選出アルゴリズムと，可用性とネットワーク負荷を考慮してスケジューリングを行う手法を提案し，2 ラックからなる HDFS クラスタにおいて，その特性及び性能が向上することをシミュレーションを用いて示してきた [1]．本稿では既提案手法を応用して，3 つ以上のラックからなる HDFS クラスタにおける効率的なレプリカ再配置手法を提案し，評価する．

2. マルチラックの HDFS のレプリカ再配置

2.1 レプリカ配置ポリシーとレプリカ再配置

複数のレプリカはレプリカ配置ポリシーに基づいて，第 1 レプリカはローカルノードに，第 2 レプリカは第 1 レプリカと異なるラックに，第 3 レプリカは第 2 レプリカと同一ラックの異なるノードに配置される．データ転送に関しては，ラック間転送よりもラック内転送が優先される．ノード故障などによりレプリカが不足した場合には，レプリカ配置ポリシーに基づき残りのノード間で不足レプリカを補う．残りのレプリカが異なるラックに存在する場合には，ラック内に再配置を行うラック内転送が行われ，同一ラックに存在する場合には、異なるラックに再配置しなければならないため，ラック間転送が行われる．この時，各データノードが同時に送信できるストリーム数は2である．

2.2 レプリカ再配置の問題点

レプリカ生成元・生成先をほぼランダムに選出するデフォルトのレプリカ再配置を，1 ラックあたりのノード数を 8 台とし，4 章に述べる 3 ラック構成のシミュレーション環境にて評価を行った際の，各ラックの送受信のブロック数と，1 ブロックあたりの平均転送時間を表 1，2 に示す．削除ノードを含むラックを failure rack，削除ノードを含まないラックを normal rack と呼ぶ．一見すると各ラックの送受信ブロック数の合計がほぼ等しく，処理が分散し，効率良く再配置処理が行われているようであるが，failure rackC のノードは，ラック間転送の生成元になり得ないが，生成先には選出されるため，受信ブロック数が圧倒的に多くなり，転送が集中して，rackC に関する転送時間が長引く事態が発生し，非効率なレプリカ再配置が行われている．

表 1: 各ラックの送受信のブロック数 (デフォルト手法)

		normal rack		failure rack
		rackA	rackB	rackC
ラック間転送	送信	311	310	0
	受信	166	157	298
ラック内転送	送信	380	388	451
	受信	380	388	451
送信合計		691	698	451
受信合計		546	545	749
送受信合計		1237	1243	1200

表 2: 1 ブロックあたりの平均転送時間

		転送時間 [sec]
ラック間転送	A to B	4.364
	A to C	5.867
	B to A	4.328
	B to C	5.710
ラック内転送	A	4.260
	B	4.323
	C	6.225

3. マルチラックのレプリカ再配置手法の提案

前節の問題を解消するために，本稿では既提案手法の指向性リング構造に基づくデータ転送を応用し，送受信のブロック数をそれぞれ均衡化させることにより，各ラックひいては各ノードの負荷を均衡化させ，効率良くレプリカ再配置を行う制御手法を提案する．

3.1 レプリカ生成元・生成先の選出

送受信のブロック数を均衡化するために，failire rack はラック間転送に関与させず，ラック内転送のみを行うものとし，その分ラック内転送の負荷を normal rack よりも増加させる．ここで，ラック数を R とし，1 台のノードが削除された際，複製の必要なブロック数を B とすると，確率的にラック間転送が行われるブロック数 $B_{inter} = \frac{1}{3}B$，ラック内転送が行われるブロック数 $B_{inner} = \frac{2}{3}B$ である．各ラックの送信ブロック数が均衡化するのは $\frac{B}{R}$ であるため，表 3 のように再配置処理を割り当てる．

生成元の選出に関して，ラック間転送に関しては，ラック間生成元選出回数が最小のノードを選出することで，表 3 を満たすことが出来る．ラック内転送に関しては，生成元候補ノードの中から，表 3 の処理の割り当てを考慮し，normal rack の場合は $3(R-1)$ 倍，failure rack の場合は $(2R-3)$ 倍して，ラック内生成元選出回数の比較を行い，選出回数が少ないノードを選出する．生成先の選出は，既提案手法の指向性リング構造に基づいて 1 つ先のノードを選出する手法を応用する．ラック内転送のために，ラック毎に論理的なリング構造を構成し，ラック間転送のために，normal rack に含まれる全てのノードを繋ぐ，1 つの論理的なリング構造を作成する．この時，前後のノードは異なるラックに属するノードとなるようにする．このようなリング構造により，ラック間及びラック内転送それぞれにおいて，一意のノードが生成先として選出される．そして各ノードの送信ブロック数の均衡化に伴い，受信ブロック数

A Proposal of Effective Replica Reconstruction Schemes for Multi-rack HDFS Environment
† Asami Higai, Masato Oguchi
‡ Atsuko Takefusa, Hidemoto Nakada
Ochanomizu University (†)
National Institute of Advanced Industrial Science and Technology (AIST)(‡)

表 3: 各ラックへの再配置処理の割り当て

	normal rack	failure rack
ラック間転送	$B_{inter} \times \frac{1}{R-1}$	0
ラック内転送	$\left(B_{inner} - \frac{B}{R}\right) \times \frac{1}{R-1}$ $= B_{inner} \times \frac{2R-3}{2R(R-1)}$	$B \times \frac{1}{R}$ $= B_{inner} \times \frac{3}{2R}$

表 4: シミュレーション環境

シミュレータ	SimGrid-3.10
ラック数	3
1 ラックあたりの DataNode 数	8, 16, 32
ブロックサイズ	67MB(default)
レプリカ数	3(default)
不足ブロック数 (削除ノードが保持するブロック数)	80*正常な DataNode 数
ラック内ネットワーク帯域幅，遅延	125 MB/sec，0.1msec
ラック間ネットワーク帯域幅，遅延	1.25 GB/sec，0.1msec
ディスク性能	67 MB/sec

も付随して均衡化される．

3.2 スケジューリング制御

発生することは稀であるが，万が一レプリカ再配置中にラック全体に渡る障害が発生した場合，同一ラックに残りのレプリカが存在するブロックは復元不可能になってしまう．そのため，ラック間転送が必要なブロックを先に再配置することは可用性の向上に繋がる．そこで，ラック間転送を行うブロックに高い優先度をつけて先にスケジューリングした後に，ラック内転送を行うブロックをスケジューリングする制御手法を優先度付手法とする．一方，これら 2 つの状態のブロックを区別することなく，任意の順にスケジューリングする制御手法を優先度無手法とする．

4. 評価実験

デフォルト手法と提案手法を用いて，ある 1 つのラックのうちの 1 台のノードを削除した際のレプリカ再配置を表 4 に示す環境において，シミュレーションにより評価する．

各手法のレプリカ再配置の実行時間を図 1 に，1 ラックあたりの DataNode 数が 8 台の場合の 1 ブロックあたりの平均転送時間を表 5 に示す．図 1 より，提案手法によりレプリカ再配置の実行時間が減少し，最大で 18%削減できた．1 ラックあたりの DataNode 数が 8，16 台の場合，優先度無手法より優先度付手法の方が実行時間削減に有効である．優先度付手法の性能が高い理由は，優先度付手法は，normal rack のデータ転送に関して，処理の冒頭はラック間転送のみが行われ，それらが終了した後にラック内転送が実行されるため，時系列的にみると生成元と生成先が一対一に対応した転送が行われ，最大受信ブロック数が 2 に制限される．一方で優先度無手法では，ラック間転送とラック内転送が混在し，リング構造に基づいたデータ転送であっても、ある DataNode に関して最大で 4 つのブロックを受信してしまう事態が発生するからである．このことから，優先度付手法は可用性の向上だけでなく，比較的小規模な環境においてはより効率の良い手法である．しかしながら，ラック内の DataNode 数が 32 台と多い場合，優先度付手法ではラック間転送の集中により，ラック間のネットワーク帯域が飽和して，性能が低下してしまうため，適切なストリーム制御が必要であることが分かる．また表 5 からデフォルト時には，failure rackC に関連する

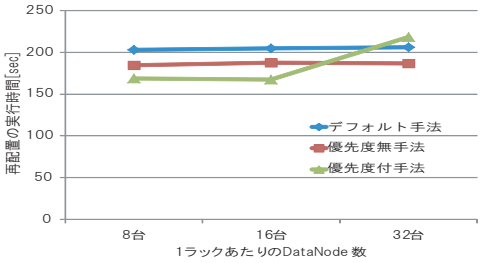

図 1: 各手法におけるレプリカ再配置の実行時間

表 5: 1 ブロックあたりの平均転送時間 [sec]

		デフォルト	優先度無	優先度付
ラック間転送	A to B	4.364	4.666	4.036
	A to C	5.867	0	0
	B to A	4.328	4.565	4.015
	B to C	5.710	0	0
ラック内転送	A	4.260	4.307	3.971
	B	4.323	4.500	3.980
	C	6.225	3.990	3.990

データ転送時間が長くなってしまっていたが，提案手法により，全ての転送形態についてほぼ等しく，効率良く転送が行われていることが分かる．

5. 関連研究

鈴木ら [2] は，クラスタ間で効率良くファイルを複製するには，単一ディスクへのアクセス集中による性能低下とネットワークの通信性能低下の回避が重要であると述べ，適切なノードにファイル複製を割り当てる「ファイル複製選択アルゴリズム」とそれらを適切なコネクションに割り当てる「転送順序スケジューリングアルゴリズム」を提案している．前者のアルゴリズムを線形計画法及び貪欲法を用いて解き，後者をリストスケジューリング法を用いて解く手法をシミュレーションにより評価し，それらの提案手法が有効であることを示している．適切な複製元選択とスケジューリングは本研究に通ずるが，複製先のノードも選択しなければならない点と，複製先が同一ラック内の場合もある点が本研究と異なる．

6. まとめ

3 つ以上のラックからなる HDFS クラスタのレプリカ再配置において、削除ノードを含むラックはラック間転送を行わないようにした上で，各ノードの送信ブロック数を均衡化し，指向性リング構造に基づいて，一対一のデータ転送を行うことで，受信ブロック数も均衡化し効率良く処理を行う手法を提案し評価した．評価実験から，提案手法により，各ラックの送受信ブロック数が均衡化され，最大で再配置の実行時間を 18%削減することができた．

参考文献

[1] Asami Higai, Atsuko Takefusa, Hidemoto Nakada, Masato Oguchi, ”A Study of Effective Replica Reconstruction Schemes for the Hadoop Distributed File System,” IEICE Trans. Inf. & Syst., Vol.E98-D,No.4,Apr. 2015 (To be appeared)

[2] 鈴木克典，建部修見，”PC クラスタ間ファイル複製スケジューリング ”情報処理学会論文誌コンピューティングシステム Vol.3 No.3 pp.113-125