2008－MPS－68 (43)
2008／3／6

広域ネットワークにおけるデータの耐災害配置モデルの提案

城 啓輔[†1] 合田憲人[†2,†1]

近年，データグリッドに代表されるような広域ネットワーク上のサイトにデータの分散配置を行い，データ処理を行う技術が注目されている．しかし，データを配置するサイトごとの災害リスクを考慮したデータ配置手法については十分考慮されていない．また，耐災害性を重視して同一のデータを複数のサイトに配置する場合，サイトごとにデータに対する課金が行われることを考慮するのであれば，データの総配置コストはユーザの資産以内である必要がある．そこで，本稿では総配置コストを制約条件に持たせた広域ネットワークにおける耐災害性の高いデータの配置モデルの提案を行い，シミュレーションによる性能評価によりモデルの有効性を示した．

Disaster Tolerant Data Allocation Model on Wide Area Network

Keisuke Jo[†1] and Kento Aida[†2,†1]

Data processing technology using distributed data on WAN is recently much used on data grids. However, fault tolerant distribution schemes among distributed sites, where the sites have different disaster risks, have not been well discussed. Some schemes save replicated data on multiple sites to improve the tolerance. In this case, data need to be allocated so that the amount of allocation cost is within the user's budget. This paper proposes a disaster tolerant data allocation model, which allocates data considering both disaster tolerance and allocation cost, on WAN. The Simulation results show the effectiveness of the proposed model.

1. はじめに

近年，インターネットに代表されるような広域ネットワークに接続された個々のサイトにデータを分散配置して，仮想的に巨大なストレージを実現させ，大規模なデータ処理を行うデータグリッド技術が着目されている．データグリッド技術を生かした事例としては，欧州原子核研究機構 (CERN) で行われている高エネルギー物理学のアプリケーションなどが挙げられる[1]．また，本アプリケーションでは，膨大なデータを各地のサイトに分散配置させ，研究者はネットワークを経由し必要なときに必要なデータを処理することが可能となっている．また，データの信頼性やアクセス性能向上を目的としたデータの複製手法に関する研究も進められている[2]．

しかし，これらの研究では，主にアクセス性能の向上を目的としたデータ配置手法については議論が行われているが，データを配置する複数のサイトの災害リスクを考慮したデータの耐災害性の観点からの議論は十分になされていない．データの耐災害性を考慮し同一のデータを複数のサイトに配置する場合，データの損失リスクは減少するが，仮にサイトごとにデータ配置に対する課金が行われる場合，データの総配置コストは増大する．このようにデータの耐災害性と総配置コストの関係は互いにトレードオフの関係にある．

本稿では総配置コストを制約条件に持たせた広域ネットワークにおける耐災害性の高いデータの配置モデルの提案を行う．シミュレーションによる評価の結果，提案モデルの有効性が確認された．

2. 関連研究

耐災害性とコストのトレードオフを考慮した研究として，Benjaminによる Redundancy Allocation Program モデル (RAP モデル) が挙げられる[3]．RAP モデルでは災害性を考慮した LAN などの小規模なネットワークにおいて，ルータや Web サーバ等のシステムに必要な機能を担当する機器を総コスト以内で冗長化させ，システム全体の耐災害性を高めることを目的としている．しかし，RAP モデルでは，単一サイトの耐災害性を対象としており，機器が複数のサイトに配置され，かつ，サイト毎の災害リスクが異なる場合については議論されていない．

†1 東京工業大学
Tokyo Institute of Technology
†2 国立情報学研究所
National Institute of Infomatics

3. 提案モデル

本節では，本稿が提案するデータの耐災害配置モデル (ReRAP モデル) について述べる．本モデルは，n_s 個のサイトに F 種類のデータ (またはそのコピー) を配置する問題として表現できる．サイト (i) にはそのサイトで災害が発生する確率 $P_{di} \in (0,1)$，データ (f) にはその重要度 (w_f) がそれぞれ与えられる．また，データ f をサイト i に配置する際には，費用 (コスト) として C_{fi} が発生する．各サイトで発生する災害の総数を D，ユーザが利用可能なコストの総額 (資産) を B とすると，提案モデルは (1)〜(4) 式により表現される最適化問題として定義される．ここで，X_{fi} はサイト $i \in \{1,...,n\}$ へのデータ f の配置を表す関数で，データが配置される場合は $X_{fi}=1$，配置されない場合は $X_{fi}=0$ である．

3.1 ReRAP モデル式

maximize

$$S = \frac{1}{D}\sum_{d=1}^{D}\Big[\sum_{f=1}^{F} w_f\Big(1-\prod_{i=1}^{n_s} {P_{di}}^{X_{fi}}\Big)\Big] \tag{1}$$

subject to

$$\sum_{i=1}^{n_s} X_{fi} \geq 1, \sum_{f=1}^{F}\sum_{i=1}^{n_s} C_{fi}X_{fi} \leq B \tag{2}$$

$$C_{1i} = C_{2i} = \cdots = C_{Fi} \tag{3}$$

$$X_{fi} = \{0,1\} \tag{4}$$

上式の提案モデルの制約条件 (2) は，各データが少なくとも 1 つ以上のサイトに配置され，全てのデータを割り当てるために必要なコストは資産 B 以下であることを意味している．この制約条件内で X_{fi} の組合せ探索を行い，S を最大化する．X_{fi} の組合せが，ReRAP モデルにおける最も耐災害性の高いデータ配置となる．

4. 実装方法

提案した ReRAP モデルは非線形の目的関数を持つ 2 値の整数プログラミングモデルである．小規模な ReRAP では X_{fi} の列挙を行うことで ReRAP モデルを解くことが可能であるが，大規模な ReRAP モデルの場合は多大な計算時間を要し実践的ではない．そこで，計算機に ReRAP モデルの実装を行うために以下のような作業を行う．

4.1 目的関数の展開

まず，目的関数 S を展開することにより，ReRAP モデルを最小化問題に変換する．

$$\begin{array}{cccc}
Z_1(1_1) & Z_1(2_1) & \cdots & Z_1(B-1_1) \quad Z_1(B_1) \\
Z_2(1_2) & Z_2(2_2) & & Z_2(B-1_2) \quad Z_2(B_2) \\
\vdots & & \ddots & \vdots \\
Z_{F-1}(1_{F-1}) & Z_{F-1}(2_{F-1}) & & Z_{F-1}(B-1_{F-1}) \quad Z_{F-1}(B_{F-1}) \\
Z_F(1_F) & Z_F(2_F) & \cdots & Z_F(B-1_{F-1}) \quad Z_F(B_F)
\end{array}$$

図 1 $Z_f(t_f)$ 行列

$$S = 1-\Big(\frac{w_1}{D}\sum_{d=1}^{D}\prod_{i=1}^{n_s} {P_{di}}^{X_{1i}} + \cdots + \frac{w_F}{D}\sum_{d=1}^{D}\prod_{i=1}^{n_s} {P_{di}}^{X_{Fi}}\Big) \tag{5}$$

各データ f において，f を割り当てる際に用いる資産を $t_f \in B$ とし,(6)(7) 式を満足する X_{fi} の組合せ $(X_{f1}, X_{f2}, \cdots, X_{fns-1}, X_{fn_s})$ 及び $Z_f(t_f)$ を求める．

minimize

$$Z_f(t_f) = \frac{1}{D}\Big(w_f\sum_{d=1}^{D}\prod_{i=1}^{n_s} {P_{di}}^{X_{fi}}\Big) \tag{6}$$

subject to

$$\sum_{i=1}^{n_s} X_{fi} \geq 1, \sum_{i=1}^{n_s} C_{fi}X_{fi} \leq t_f, X_{fi} = \{0,1\} \tag{7}$$

以上の操作をすべての $f \in F$ において行うと図 1 に示した行列が生成される．

$Z_f(t_f)$ 行列において，制約条件 (2) を満たすためには，(8)(9) 式を満たす必要がある．

minimize

$$Z = \sum_{f=1}^{F} Z_f(t_f) \tag{8}$$

subject to

$$1 \leq t_f \leq B, \sum_{f=1}^{F} t_f \leq B \tag{9}$$

制約条件 (9) を満たすように，総当り方式で (8) 式の最小値を求めるには，計算量は $O(B^F)$ となり，割り当てデータ数 F が大きくなるにつれ計算量は爆発的に増加する．よって，計算量を削減するために動的計画法を導入し，制約条件 (9) を持った，(8) 式を再帰方程式へ変換を行うと (10)(11) 式に変換される．

$$\varphi_f(T_f) = min[Z_f(t_f) + \varphi_{f-1}(T_f - t_f)]$$
$$f = 2, \cdots, F,\ 1 \leq T_f \leq B \tag{10}$$

$$\varphi_1(T_1) = min[Z_1(t_1)], 1 \leq T_1 \leq B \tag{11}$$

よって，(10)(11) 式の再帰方程式により，ReRAP モデルの計算量を占める部分である (8) 式の計算量が $O(FB^2)$ になることで，ReRAP モデルは実用的な時間で解く事が可能となる．これにより，$1-\varphi_F(B)$ が ReRAP モデルにおける利用可能資産 B を用いた際のデータ全体の生存確率となり，各 $f \in F$ に対応した X_{fi} の組合せ $(X_{f1}, X_{f2} \ldots X_{fns-1}, X_{fn_s})$ がデー

タ f が割り当てられるサイトの組合せとなる．

5. 実験方法

本節では，提案モデルの有効性を検証するために行ったシミュレーション実験について述べる．

5.1 実験モデル

本実験モデルは以下のように設定を行った．

- アプリケーションモデル
  本実験が対象とするアプリケーションは，複数のジョブから構成され，各ジョブは広域ネットワーク上に分散配置された複数のデータを参照するものとする．
- 広域ネットワークモデル
  本実験が対象とする広域ネットワーク環境では，各サイトは1つのルータに接続され，ルータを介して別のサイトと通信を行う．各ジョブは，データが配置されるサイトとは別の1サイト上で実行され，実行に必要なデータを保持するサイトからデータを転送して参照する．また，本実験では，サイト内での災害を対象とし，サイト間ネットワークでは災害は発生しないと仮定している．
- 災害の発生方法
  災害の発生はトラフィックの時系列変動を利用する．各サイト間の時系列において，災害が発生した時点でのバンド幅使用率を 10^{-6} にし，災害による影響がある期間だけバンド幅使用率を減らすことにより，災害を再現した．
- シミュレータ
  実験を可能にするため，データグリッドシミュレータ OptorSim[5] を改変し，シミュレータを作成した．具体的な改変作業は，明示的なサイト上のデータ配置，長期間のトラフィック時系列変動の設定である．
- 評価指標
  災害が起こった場合，サイトにリンクされているバンド幅は極端に小さくなり，サイトからのデータ転送に時間がかかる．よって，本実験ではアプリケーションの実行時間を性能指標とした．
- データの重要度
  ReRAP モデルアルゴリズムでは，データの重要度 w_f を考慮する．そこで，データの重要度生成式をジョブ j(総数 J) を基に以下の式で定めた．

$$w_f = \sum_{j=1}^{J} \text{ジョブ } j \text{ の発生確率} \times \frac{\text{ジョブ } j \text{ でのデータ } f \text{ の出現数}}{\text{ジョブ } j \text{ で使用するデータの個数}} \tag{12}$$

- 配置可能資産率
  ネットワーク上のサイト i において，データ f の配置コストが最小のサイトの配置コストを $C_{fi_{min}}$ とする．全てのデータは必ず1つ以上配置されなければならないので，データ数が F の場合，最小配置資産は $F \times C_{fi_{min}}$ となる．また，配置資産の最大値は全てのサイトに全てのデータを配置する場合であり，最大配置資産は $F \times \sum_{i=1}^{n_s} C_{fi}$ となる．配置資産を B とするならば，これを基に配置可能資産率は以下の式で表される．

$$\text{配置可能資産率} = \frac{B - F \times C_{fi_{min}}}{F \times \sum_{i=1}^{n_s} C_{fi} - F \times C_{fi_{min}}} \times 100 \tag{13}$$

5.2 比較対象アルゴリズム

ReRAP モデルアルゴリズムの性能比較のために以下のデータ配置アルゴリズムを用意した．

- バンド幅優先配置アルゴリズム (BAND)
  バンド幅優先配置アルゴリズムでは，データの重要度が高い順に，バンド幅が大きいリンクに繋がっているサイト順にデータ配置を行う．また，制約条件としてデータは必ず一つ以上割り当てられなければならないので，最大バンド幅のリンクに繋がっているサイトに全てのデータの種類を割り当てることができなければ，データ割り当て途中で配置を行うサイトを変えることにより，必ず全ての種類のデータが配置されることを保証する．
- 災害確率優先配置アルゴリズム (DIS)
  災害確率優先配置アルゴリズムはバンド幅同様の流れになっており，バンド幅でデータを割り当てるサイトの優先度を決めるのではなく，災害確率で決定する．

6. 実験結果

実際はネットワークモデルを2種類，アプリケーションモデルを2種類用意し，それぞれの組合せで実験を行ったが，本稿では紙面の都合上ネットワークモデルを EUDATA モデル[6]，アプリケーションモデルを IBCP アプリケーション[4] の結果を掲載する．

6.1 EUDATA モデル

EUDATA モデルとは EU DATA Grid Project にてアプリケーションに実際に使用されているトポロジ

モデルである．本実験ではデータ配置サイト数としてEU DATAモデル中のサイトの中から10サイトを選択した．災害数Dは1とし，災害発生確率P_{di}は0.001から0.3の区間でランダムに5パターンとした．災害によりバンド幅が影響を受ける期間は1時間，1日，1週間の3パターンとした．各サイトのデータ配置コストはバンド幅比例，一定，ランダムの場合を想定した．

6.2 IBCPアプリケーション

IBCPアプリケーションモデルとは大規模な蛋白質配列の処理により蛋白質の特定の部位の解析を行うアプリケーションである．このアプリケーションのデータはLyonにあるIBCP(Research Institute for the biology and chemistry of proteins)の実際のログから生成した[4]．ジョブが使用するデータは10個，各ジョブの発生個数は8873回とし，異なる配置可能資産率ごとにIBCPアプリケーションを実行し，各サイトの災害確率5パターンの平均実行時間を計測した．

6.3 考　　察

本実験結果より，データ配置コストがバンド幅比例，一定の場合は，ReRAPとDISの実行時間がほぼ同一となり，BANDに比べて優位であることが確認された．また，図2は災害によりバンド幅が影響を受ける期間が1週間，配置コストパターンがランダムの場合の結果を示しており，ReRAPがDISに比べて優位性を示すことが確認できる．これは災害確率優先配置アルゴリズムでは，災害確率を重視するあまり少数のサイトにデータが集中することで，性能が低下することが原因である．性能の違いを示す例として表1を挙げる．表1は各サイトにおけるランダムな場合の配置コスト及び災害確率を示し，さらに，配置可能資産率が27%での各アルゴリズムの各サイトでのデータの配置個数を示している．災害確率優先アルゴリズムでは，コストが高くても災害確率が小さいサイトに優先的データ配置するため，Torinoのように，災害確率は小さいがコストは高いサイトにデータを配置してしまうケースがある．しかし，ReRAPモデルアルゴリズムでは，冗長化指向が強いため，災害確率を優先しすぎて配置コストを考慮しないことは避けることができ，結果的に配置コストに見合わないサイトを避けることができることが確認できる．以上により，本研究にて提案を行ったReRAPモデルはネットワーク上のサイトのデータ配置コストがバンド幅に相関なくランダムに決定されている場合に有効であることが示された．

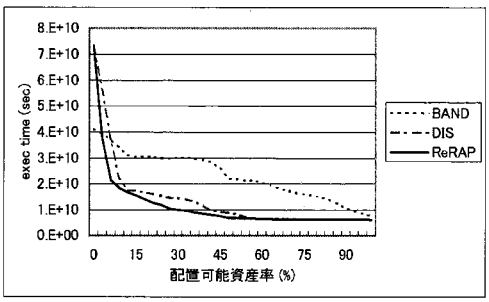

図2 災害によりバンド幅が影響を受ける期間:1週間 配置コストパターン:ランダム

表1 各サイト情報及びデータ配置個数

サイト	配置コスト	災害確率	DIS	ReRAP
NIKNEF	7	0.144	0	2
Lyon	8	0.277	0	0
CERN	6	0.038	23	23
Padova	3	0.254	0	23
Bologa	10	0.181	0	0
Canan	9	0.140	0	0
Torino	9	0.138	23	0
Mirano	2	0.132	23	23
RAL	4	0.250	0	21
Norbu	2	0.196	3	23

参 考 文 献

1) BT Human, R McNulty, T Shears, RS Denis, D Waters: "The CDF/D0 UK GridPP Project",CDF Internal Note, 2002
2) W. Hoschek, J. Jean-Martinez, A. Samar, H. Stockinger, and K. Stockinger: "Data Management in an International Data Grid Project", IEEE/ACM Int. Workshop on Grid Computing(Grid ' 2000), Bangalore, India, December 2000.
3) Benjamin B.M Shao: "Optimal redundancy allocation for disaster recovery planning in the network economy", IEEE Transactions on Dependable and Secure Computing,Vol.2, No.3,JULY-SEPTEMBER , 2005
4) Desprez F., Vernois A. : "Simultaneous scheduling of replication and computation for bioinformatics applications on the grid", Challenges of Large Applications in Distributed Environments, 2005. CLADE 2005. Proceedings Volume , Issue , 24 July 2005 Page(s): 66 - 74
5) OptorSim, http://edg-wp2.web.cern.ch/edg-wp2/optimization/optorsim.html
6) The DataGrid Project, http://eu-datagrid.web.cern.ch/