設計自動化 72−17
(1994. 10. 28)

DR-net におけるパリティ配置/データ再構築戦略の影響

味松 康行　　横田 治夫
mimatsu@jaist.ac.jp　　yokota@jaist.ac.jp

北陸先端科学技術大学院大学 情報科学研究科
〒923-12 石川県能美郡辰口町旭台 15

あらまし 我々は高信頼並列ディスクシステムを構成するために DR-net を提案してきた。DR-net に関するこれまでの研究で、データ書き込みの際のパリティ更新処理が特定ノードに集中することを避けるために、2つのパリティ分散保持方式が提案されている。また、故障ディスク内のデータへの読み出し要求に対しては冗長情報を利用したデータの再構築が必要であるが、再構築処理の進め方に関して2つのデータ再構築戦略が提案されている。本稿では、トランスピュータと小型ディスクによる実験システムを用い、2つのパリティ分散方式および2つのデータ再構築戦略の性能への効果·影響に関する評価実験を行なった結果について報告する。実験結果は、パリティの分散保持により書き込み性能がほぼ理論値に一致して改善されること、2つのデータ再構築戦略のうち当初不利と思われた逐次的な戦略が優れていることを示している。

和文キーワード　DR-net, 冗長ディスクアレイ, RAID, パリティ分散配置, データ再構築

Parity Distribution and Data Reconstruction Strategy of DR-net

Yasuyuki MIMATSU　　Haruo YOKOTA
mimatsu@jaist.ac.jp　　yokota@jaist.ac.jp

School of Information Science, Japan Advanced Institute of Science and Technology
15 Asahidai, Tatsunokuchi Nomi-gun Ishikawa 923-12, Japan

Abstract　We have proposed DR-nets, Data-Reconstruction networks, to construct highly reliable parallel disk systems. In the previous studies on DR-nets, two methods for parity distribution and two strategies for data reconstruction were proposed. The two parity distribution methods avoid load imbalance during parity update in write processes. The usage of parity groups differs in two reconstruction strategies. In this paper, we compare these two methods and strategies using the DR-net prototype. The results indicate that the both two methods improve the performance almost same as the estimation, and that one strategy unexpectedly provides better performance despite its sequentiality.

英文 key words　DR-net, Redundant disk arrays, RAID, Parity distribution, Data reconstruction

1 はじめに

近年、プロセッサの処理速度向上にともない、ディスクシステムに対する性能向上の要求が高まっている。また、大容量二次記憶の利用には、ディスクシステムの信頼性向上が不可欠である。RAID として知られる冗長ディスクアレイは、多数のディスクを並列に動作させることにより二次記憶装置の性能を向上させ、さらに冗長情報を用いた信頼性の向上を図っている [1][2]。

しかし、非常に多数のディスクを用いた場合に RAID は必ずしも十分な性能、信頼性を提供するとは限らないように思われる。多数のディスク装置をつなぐ1本のバスにがボトルネックとなり、性能を低下させることが考えられる。信頼性に関しても、1つのパリティグループ内では単一のディスク故障を仮定しているが、ディスクの数が増えた場合には、単一故障の前提は必ずしも適当ではなく、また、ディスク装置単体の MTTF が製造ロットによって偏っていることも考えられる。

我々は、RAID で用いられているパリティ計算の手法を相互接続ネットワークに適用し、上記の問題を解決する方法を提案してきた [3][4][5]。データ再構築ネット (Data-Reconstruction Networks:DR-net) では、ディスクは相互接続ネットワークの各ノードに接続され、その相互接続ネットワークのサブネットでパリティグループを形成する。各ノードは1本のバスではなく、相互にネットワークで接続されているためバスボトルネックは存在しない。

DR-net では、2種類のパリティグループをネットワーク上に重ね合わせて配置することにより、高い信頼性を実現することができる。5 × 5 の 2 次元トーラスネットワークを用いた構成では、いかなる2つのディスク故障に対しても全てのデータを再構築することが可能であり、4 つの故障でも 95%以上が再構築可能である。また、故障ディスクの位置関係によっては最大 9 つのディスク故障に対しても対応できることがわかっている [5]。

基本的な DR-net の構成では、パリティデータはパリティノードと呼ばれる特定のノードで集中して保持されるため、データ書き込みの際にパリティ更新処理がパリティノードに集中し、性能が低下する原因となる。これは RAID のレベル 4 と同様の問題である。また、2種類のパリティグループについてパリティ情報を保持しなければならないため、パリティノードのディスクだけ2倍の容量を必要とする。これらの不均衡に対処するために、パリティを全てのディスクで分散保持する2つの方式が提案されている [4]。また、故障ディスクが保持するデータを再構築する場合には、2種類のパリティグループの利用法により、2通りの再構築戦略が考えられる [6]。

これまで、上記のパリティ分散保持方式によるシステム性能向上の定量的な評価や、2つのデータ再構築戦略の性能面における比較は行なわれていなかった。本稿ではトランスピュータと小型ディスクを用いた構成で実験を行ない、2つのパリティ分散方式の効果、および2つの再構築戦略の比較に関する評価結果を報告する。

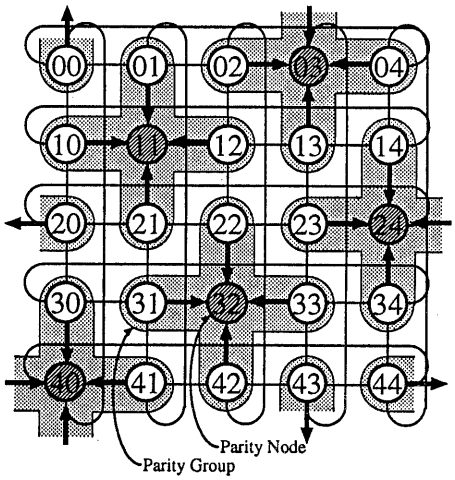

図 1: FPG(First Parity Groups) の構成

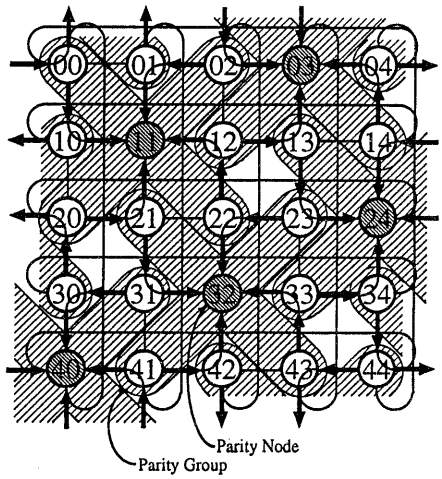

図 2: SPG(Second Parity Groups) の構成

2 DR-net の概要

DR-net を構成するネットワークトポロジは各種考えられるが [5]、以下では単純な例として 5 × 5 の 2 次元トーラスネットワークを考える。その他のネットワークトポロジの場合には、パリティグループの形状等も変更することになる。

2.1 パリティグループの構成

パリティグループはデータを保持する複数のデータノードと、それらのデータのパリティを保持するパリティノードから構成される。データノードにデータを書き込む場

合にはパリティの更新が伴う。新しいパリティは次式で算出される。

$$\text{新パリティ} = \text{旧パリティ xor 旧データ xor 新データ} \quad (1)$$

パリティ計算の際に、パリティグループ内での通信を局所的に行なうことが重要であるから、各データノードはなるべくパリティノードの近傍に配置されていることが望ましい。我々は、2種類のパリティグループとして、パリティノードを中心とする十字型と斜め十字型を選んだ。これらを5×5のトーラス上に配置したのが図1, 2である。十時型のパリティグループをFPG(first parity groups)、斜め十字型のパリティグループをSPG(second parity groups)と呼ぶ。

2.2 故障ディスクへのアクセス要求

故障ディスクが保持するデータに対して読み出し要求があった場合、そのノードが属するパリティグループ内の他の4つのノードで保持されているデータおよびパリティの排他的論理和を計算し、データを再構築する。故障ディスクに対する書き込みは、そのノードが属するパリティグループのパリティを更新することで実現される。その際、式(1)のパリティ計算に必要な書き換え前の旧データは故障ディスク内にあるため、パリティは次式で計算される。

$$\begin{aligned}\text{新パリティ} \;=\; &\text{新データ xor データ A} \\ &\text{xor データ B xor データ C} \quad (2)\end{aligned}$$

ここで、データA, B, Cは新データを書き込むデータノードと同じパリティグループに属する3つのデータノードにそれぞれ保持されているデータである。

2.3 データ再構築戦略

データを再構築する場合、2つのパリティグループ(FPG, SPG)のどちらも利用できる。FPGを用いた再構築に失敗してもSPGを用いた再構築に成功する場合がある。また、その逆もあり得る。再構築のために利用できるパリティグループが2種類あることから、次の2つの再構築戦略が考えられる[6]。

2.3.1 Lazy Reconstruction Strategy(LRS)

LRSは、まず1つのパリティグループ(FPGあるいはSPG)で再構築を試み、成功すればその結果を返す。失敗した場合にのみ、もう一方のパリティグループを用いた再構築を試みる。この戦略では最初の再構築に成功したときには、1つのパリティグループだけを使うため、次のERSに比べて無駄なディスクアクセスが減る。しかし、最初の再構築が不可能な場合には、再構築失敗が確定してからもう一方のパリティグループを用いた再構築を始めるため、再構築完了までにより長い時間を要すると考えられる。

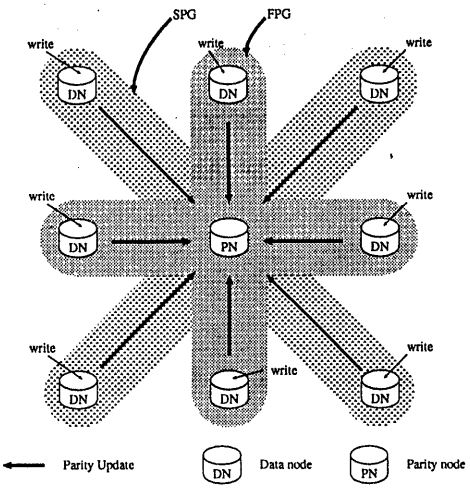

図3: パリティノードへの負荷の集中

2.3.2 Eager Reconstruction Strategy(ERS)

ERSは、故障ノードが属する2つのパリティグループを同時に利用し、双方で並列に再構築処理を行なう戦略である。2つの再構築のうち、最初に返ってきた結果を利用するため、より速いレスポンスを期待できる。反面、最初に返ってきた再構築結果が成功であった場合、後から返される結果は利用されず捨てられてしまう。つまり、無駄な再構築を行なっていることになり不必要なディスクアクセスが増加してしまう。

2.4 パリティ分散保持方式

あるデータノードにデータを書き込む場合、そのノードが属する2つのパリティグループのパリティノードにおいてパリティを更新する必要がある。各パリティグループは1つのパリティノードと4つのデータノードから構成され、各パリティノードは2つのパリティグループに属するため、各パリティノードは合計8つのデータノードからパリティ更新要求を受けることになる(図3)。従って、書き込みアクセスが多くなった場合、パリティノードに負荷が集中し性能の低下を招く。これはRAIDのレベル4と同様の問題である。そこで、パリティを特定のノードで集中して保持するのではなく、RAIDのレベル5のように全てのノードに分散して保持することが考えられる。書き込み性能改善のためのパリティ分散保持の方式として次の2つが提案されている[4]。

2.4.1 Moving Parity Groups(MPG)

一つは、ネットワークの対称性を利用しネットワーク上のパリティグループの配置をフェーズによりずらす方法である(図4)。5×5の2次元トーラスネットワークの場合では、図4のようにどのノードも5つのフェーズのいずれかでパリティノードとなる。フェーズの切替えはディスクのアクセスするページ、セクタ、トラック等の単位で行なうことが考えられる。このパリティ分散方式を

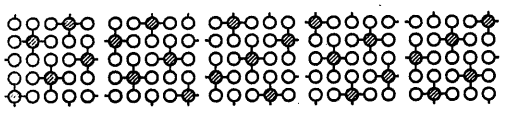

図 4: フェーズによるパリティグループの移動 Moving Parity Groups (MPG)

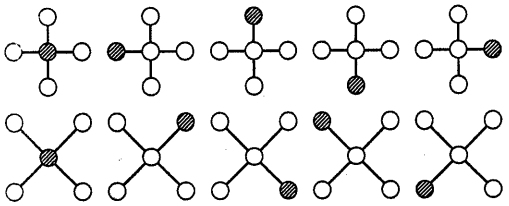

図 5: フェーズによるパリティノードの移動 Moving Parity Nodes (MPN)

Moving Parity Groups(MPG) と呼ぶ。MPG では、パリティグループの移動により 1 つの故障ノードが及ぼす影響の範囲が広がるので、パリティノードを固定した場合よりも信頼性が低下する。

2.4.2 Moving Parity Nodes(MPN)

もう一つの方法は、パリティグループの配置は固定し各パリティグループ内でパリティノードの位置を移動させるものである (図5)。この方式を MovingParity Nodes(MPN) と呼ぶ。MPN は非対称なネットワークにも適用することができる。また、1 つの故障ノードが及ぼす影響の範囲は固定パリティノード方式と同じであるので、信頼性は MPG よりも高い。しかし、各パリティグループ内でデータノードからパリティノードまでの通信距離が若干増大する欠点を持つ。

3 性能評価実験

上記の 2 つの再構築戦略および 2 つのパリティ分散保持方式の性能への影響を調べるため、各ノードにトランスピュータと 2.5″小型ハードディスクを持つ 5×5 トーラスネットワーク構成の試作機 [7][8][9] を用いた性能評価実験を行なった。実験では、システムに対し 1000 回のディスクアクセスを与えられた間隔で要求し、それぞれのアクセスのレスポンスタイムの平均値を測定した。アクセスは全てのノードに均等に分散され、1 回のディスクアクセスは 1 ページ (512 バイト)、アクセスするページはランダムである。実験はディスク故障が存在しない場合と存在する場合に分けて行なった。

3.1 ディスク故障が存在しないとき

アクセスの読み出し/書き込みの比率およびアクセス要求の発行頻度を評価の際のパラメータとした。読み出し/書き込みの比率は 1000/0~0/1000、アクセスの頻度は 50 ~ 600 I/Os per second の範囲で変化させ、それぞれのパリティ保持方式 (集中保持, MPG, MPN) について評価を行なった。実験結果を図 6, 7に示す。

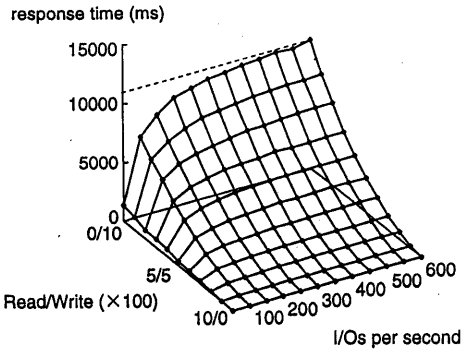

図 6: パリティ集中保持 (ディスク故障なし)

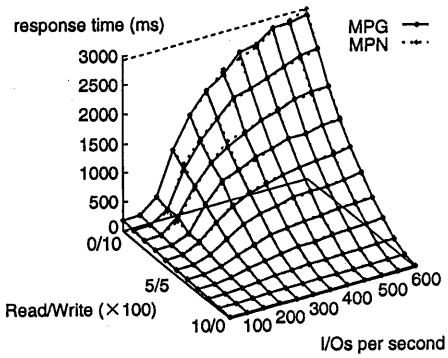

図 7: パリティ分散保持 (ディスク故障なし)

MPG, MPN を用いたパリティ分散保持による書き込み性能の改善について、各データノードに 1 回の書き込み要求が出された場合を考えると、パリティノードには、同じ FPG に属する 4 つのデータノードおよび同じ SPG に属する 4 つのデータノードから、合計 8 回のパリティ更新が要求される。つまり、パリティノードにはデータノードの 8 倍の負荷がかかる。MPG, MPN によりパリティを分散保持した場合、負荷が均等に分散されたとすると各ノードでは 1 回のデータ書き込みと 2 回のパリティ更新が行なわれ、パリティノードの負荷は 3/8 になる。パリティを特定のパリティノードで集中保持した場合、パリティノードにアクセスが集中し、パリティ更新処理にかかる時間がレスポンスタイムを支配するため、負荷が 3/8 になるとレスポンスタイムもおよそ 3/8 に短縮されると考えられる。

また、パリティを集中保持した場合には 5 つのノードはパリティノードとして使用されるため、ユーザからのアクセスは受け付けない。しかし、パリティを分散保持した場合には、全てのディスクがユーザからのアクセスを受け付けるため、ユーザからのアクセス要求がより多くのディスクで並列に処理されることになる。この実験では、パリティを集中保持したときには 20 ノード、分散保持した

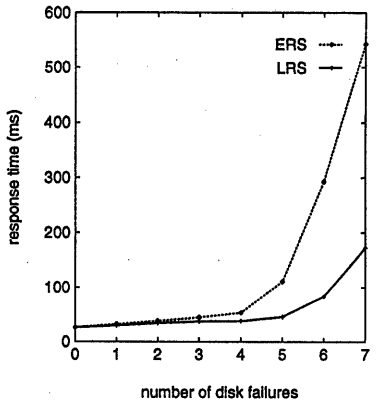

図 8: パリティ集中保持の読み出し (ディスク故障あり)

ときには 25 ノードでアクセスを受け付ける。処理の並列度に比例してレスポンスタイムが短縮されると考えると、パリティ分散によってレスポンスタイムは 20/25 = 4/5 に短縮される。

これら 2 つの要因から、パリティ分散保持によりレスポンスタイムは $3/8 \times 4/5 = 3/10$ に短縮されると期待できる。ディスク故障が存在しないときの書き込み性能の実験結果を比較すると、パリティを集中保持したときは最大で約 10000 ms (図 6)、分散保持したときは最大で約 3000 ms (図 7) であるから、上記の理論値とほぼ一致することが確認できる。

3.2 ディスク故障が存在するとき

故障ディスク数を評価の際のパラメータとし、各パリティ保持方式について読み出し、書き込みをそれぞれ 1000 回行なった。読み出し性能の評価では、2 つの再構築戦略についてそれぞれ評価を行なった。

アクセス頻度は 100 I/Os per second で、アクセス要求が出されたノードのディスクが故障していた場合、故障は直ちに検出されデータ再構築等の故障時動作を行なう。故障ディスク数は 0~7 であるが、同じ故障数でもネットワーク内の故障ノードの位置パタンにより得られる結果が違い、また、故障数が多くなった場合にはデータが再構築できないようなパタンも存在する。全ての位置パタンについて実験することは困難であるため、ここでは、各故障ディスク数について 4 つの再構築可能な位置パタンを選び、それらの結果の重みつき平均値を評価結果とした。パリティを集中保持したときの読み出しおよび書き込みの実験結果を図 8, 9に示す。図 10, 11はそれぞれ MPG, MPN の読み出しの結果、図 12は MPG, MPN の書き込みの結果である。

ディスク故障が存在するときの読み出し性能の実験結果から 2 つの再構築戦略を比較すると、従来、ERS は 2 つのパリティグループで並列に再構築を行ない、2 つの結果の早く得られた方を利用するため、レスポンスタイムが短縮されると考えられてきた。しかし今回の実験では、パリティを集中保持した場合、および MPG, MPN によって分散保持した場合のいずれに関しても LRS の方が

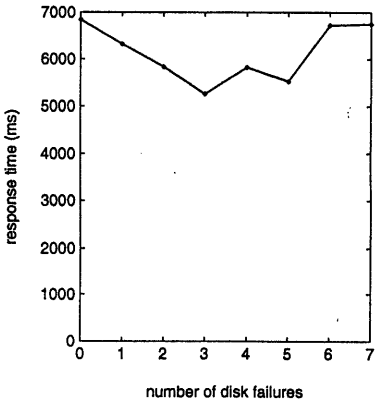

図 9: パリティ集中保持の書き込み (ディスク故障あり)

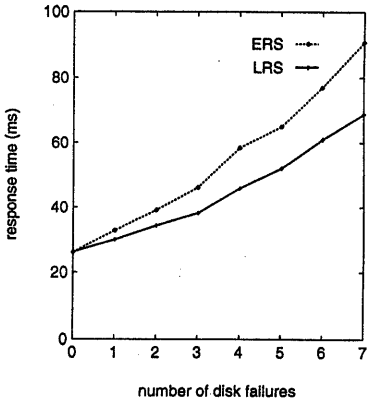

図 10: MPG の読み出し (ディスク故障あり)

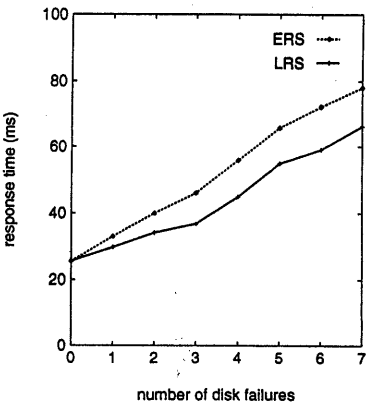

図 11: MPN の読み出し (ディスク故障あり)

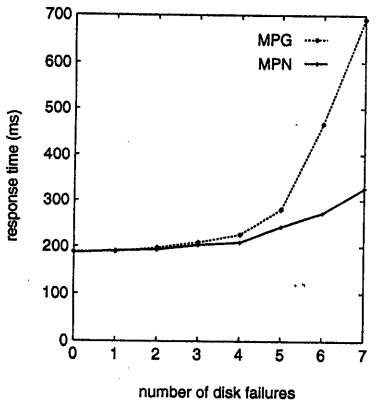

図 12: パリティ分散保持の書き込み (ディスク故障あり)

ERS よりもレスポンスタイムが短くなった (図 8, 10, 11)。これは、ERS が LRS に比べて多くのノードでディスクアクセスを行なうことによる他ノードの負荷増大が予想以上にレスポンスタイムを悪化させ、その結果 LRS の方が良い性能を示したと考えられる。

また、パリティ集中保持方式では、ディスク故障が存在するときの書き込み性能が故障数に関して単調に低下していない (図 9)。正常なディスクへの書き込みに伴うパリティ更新は式 (1) で計算される。式 (1) の計算には旧パリティが必要なことから、パリティノードにおいて 1 回の読み出し (旧パリティの読み出し) と 1 回の書き込み (新パリティの書き込み) が必要である。一方、故障ディスクへの書き込み要求に伴うパリティ更新は式 (2) で計算されるため、パリティノードにおいて旧パリティの読み出し処理は行なわれず、パリティノードの負荷は軽減される。その反面、式 (2) では書き込むデータノードと同じパリティグループに属する他の 3 つのデータノードのデータを必要とするため、他のデータノードの負荷が増大する。従って、書き込みを要求されたディスクが故障していると、

- パリティノードの負荷軽減
- 他のデータノードの負荷増大

という相反する 2 つの要因が考えられる。これらの要因により、パリティ集中保持方式での書き込みは、故障ディスク数が少ないときには性能のネックとなっていたパリティノードの負荷軽減によりレスポンスタイムが短縮されるが、故障ディスク数が多くなると他のデータノードの負荷増大により再びレスポンスタイムが悪化すると考えられる。

4 おわりに

DR-net におけるパリティ分散保持の効果、およびディスク故障が存在する場合の 2 つのデータ再構築戦略の性能への影響に関する評価実験について報告した。実験により MPG,MPN を用いたパリティ分散保持方式では、いずれもほぼ理論値通りの性能改善が示された。また、2 つのデータ再構築戦略に関しては、再構築の際に比較的少ないノードを必要とする LRS が有利であることがわかった。ERS は、従来は性能面で有利と考えられていたが、より多くのノードを並列に使用するため使用するノードの負荷の増大により結果的に平均レスポンスタイムが増大する。また、ディスク故障が存在する場合のパリティ集中保持方式の書き込み性能に関しては、故障ディスクの数が少ないときにはパリティノードの荷が軽減されることから性能が改善されるが、故障ディスク数が多くなると他のデータノードへの負荷が増大し、再び性能が悪化することが示された。

今後、DR-net を動画サーバ等として利用することを考慮した場合、単一の外部とのインタフェースではバンド幅が不足することが予想されるため、複数のインタフェースを用いた構成について研究を進め、現在の実験システムに実装し評価を行なう予定である。また、これまでの研究は主に 5×5 の 2 次元トーラスネットワークを用いた構成に限られていたため、その他の規模やネットワークトポロジを用いた構成についても研究を行いたいと考えている。

参考文献

[1] D. A. Patterson, G. Gibson and R. H. Katz: "A Case for Redundat Arrays of Inexpensive Disks (RAID)", Proc. of ACM SIGMOD Conference, pp. 109–116 (1988).

[2] G. A. Gibson: "Performance and Reliability in Redundat Array of Inexpensive Disks", Proc. of 1989 CMG Annual Conference (1989).

[3] 横田治夫: "RAID のネットワーク上への展開と信頼性向上", 信学技法 CPSY 93-11, 電子情報通信学会 (1993).

[4] 横田治夫: "データ再構築ネット (DR-net) における不均衡対策", 信学技法 FTS 93-20, 電子情報通信学会 (1993).

[5] H. Yokota: "DR-nets: Data-Reconstruction Networks for Highly Reliable Parallel-Disk Systems", Proc. of 2nd Workshop on I/O in Parallel Computer Systems, pp. 105 – 116 (1994).

[6] 友永誠史: "並列処理環境における二次記憶システムの信頼性に関する研究", 修士論文 (1994).

[7] S. Tomonaga and H. Yokota: "An Implementation of a Highly Reliable Parallel-Disk System using Transputers", Proc. of the 6th Transputer/Occam Intn'l Conf., IOS Press, pp. 241 – 254 (1994).

[8] H. Yokota and S. Tomonaga: "The Performance of a Highly Reliable Parallel Disk System", Proc. of the World Transputer Congress '94 (Eds. by A. D. Gloria, M. R. Jane and D. Maini), IOS Press, pp. 147–160 (1994).

[9] 横田治夫、友永誠史: "高信頼並列ディスクプロトタイプのアクセス性能", 信学技法 FTS 94-37, 電子情報通信学会 (1994).