# ネットワーク結合並列ディスクにおける耐故障制御の影響

味松康行† 横田治夫†

我々は，高信頼並列ディスクシステムを構成するために，ディスクを相互接続ネットワークで結合しパリティ計算により故障ディスクをマスクするDR-netを提案し，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の2種類のパリティ分散保持方式や，ディスク故障が発生した場合のデータ再構築処理の進め方に関する2種類の戦略等を提案してきた．本稿では，これらの各方式の効果および，ディスク故障が性能に与える影響について報告する．システムのスループットに関するモデルを提示し，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た実測値との比較を行う．その結果は，読み出し性能においてモデルにほぼ一致する性能向上が見られること，パリティの分散保持により固定保持した場合と比較して書き込み性能が改善されること，2つのデータ再構築戦略では逐次的な戦略が優れていることなどを示した．

# Effect on Fault Tolerant Control in a Parallel Disk Array with an Interconnection Network

YASUYUKI MIMATSU† and HARUO YOKOTA†

We have proposed DR-nets, Data-Reconstruction networks, to construct highly reliable parallel disk systems. We study two methods for parity distribution and two strategies for data reconstruction for DR-nets. In this paper, we present a model for system throughput and show results of experience using a DR-net prototype. The results indicate that the performance of read operation meets with the model and that performance of write operation improves with parity distribution.

## 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，動画データをサーバに格納してリアルタイムで提供するシステム等でも広いバンド幅を持つ信頼性の高い大容量二次記憶が望まれている．

しかし，プロセッサやメモリのような半導体製品とは異なり，ディスクシステムには機械的な動作がともなうため，単体ディスクにおける性能および信頼性の大幅な向上は望めない．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て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[2)]．信頼性に関しても，ディスクの数が増えた場合にはパリティグループ内での単一故障の前提は必ずしも適切ではない[3)]．また，1つのデータディスクに対して複数のチェックディスクを設けることにより，多重故障を扱えるRAIDも提案されているが[4),5)]，さらにバスの負荷が増大するためスケーラビリティが十分とはいえない．

我々は，RAIDで用いられ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[6)~8)]．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おり，通信が分散されるため高並列化が望める．また，ディスク故障が存在する際のデータ再構築計算は局所的なサブネット内で行われるため，ネットワークのサイズが大きくなった場合でもデータ再構築の速さはあまり劣化しない．また，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

† 北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science and Technology

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すべてのデータを再構築することが可能であり，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[8]，ほぼ同数のディスクを持つシステムで比較すると，同じ冗長情報の割合を持つRAIDレベル5あるいは6よりも高い信頼性を有することが分かっている[9]．

DR-netは，RAIDレベル5同様パリティの分散保持が可能で，これまでに2つの分散方式が提案されている[7]．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ている[10]．本稿では，提案されたこれらの方式・戦略の実際の性能面への影響について調査した．トランスピュータと小型ディスクからなる実験システム[11]を用いたスループットの測定結果から，パリティ分散による性能改善の効果を示し，システムのモデルから算出される値との比較を行う．また，ディスク故障が存在する場合の性能特性を明らかにし，2つのパリティ分散方式および2つの再構築戦略が与える影響を示す．

## 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[8]，またパリティグループの形状も様々に変更可能である[9]が，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の例を考える．

### 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ともなう．新しいパリティはRAIDレベル4または5と同様に次式で算出される．

新パリティ＝旧パリティ **xor** 旧データ **xor** 新データ　(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．十字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．

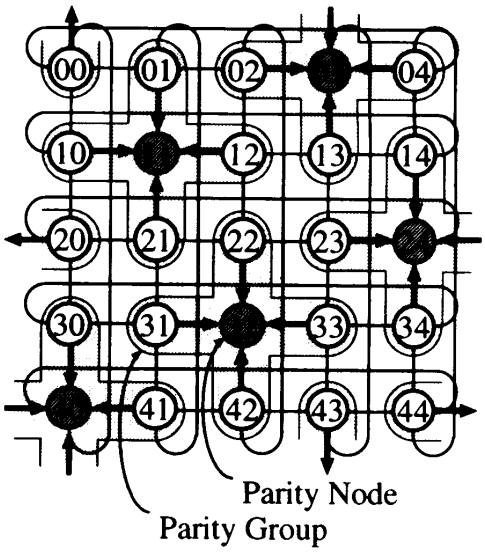


**図1** FPG（First Parity Groups）の構成
Fig. 1 Structure of FPGs (First Parity Groups).

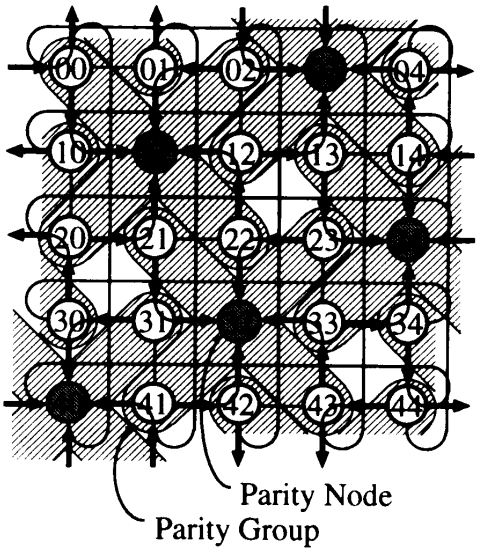


**図2** SPG（Second Parity Groups）の構成
Fig. 2 Structure of SPGs (Second Parity Groups).

### 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．

新パリティ＝新データ **xor** データA **xor** データB **xor** データC　(2)

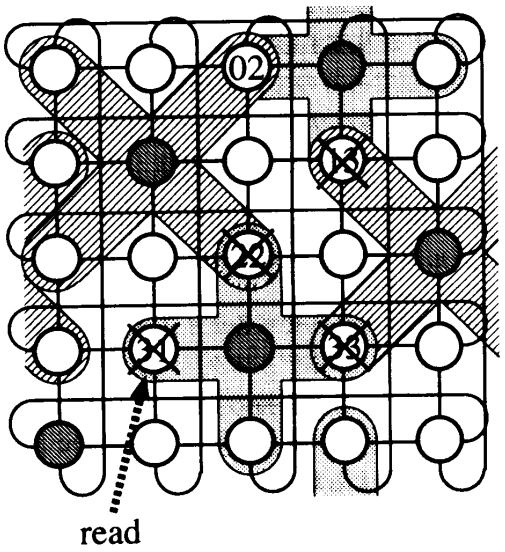


図3 FPG, SPG を併用したデータ再構築
Fig. 3 Data reconstruction with FPGs and SPGs.

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．

1 つのパリティグループ内で複数のディスク故障が発生した場合には，前述の 2 種類のパリティグループ（FPG, SPG）を併用することで対応できる．同じパリティグループ（FPG）内の 3 つのデータノードでディスク故障が発生している例（**図 3**）で，ノード 31 のデータを読み出すことを考える．この場合は，故障ディスクを 1 つしか含まない SPG を用いればただちにデータを再構築できるが，同じパリティグループ（ここでは FPG）内に複数の故障があっても対応できることを示すために，ここでは FPG による再構築を考える．

(1) ノード 31 が属する FPG を用いてデータの再構築を試みる．再構築の際には，同じ FPG に属する他の 3 つのデータノードのデータおよびパリティノードのパリティが必要である．

(2) 必要なデータを保持する 2 つのデータノード 22, 33 で，ディスクが故障していることが検出される．

(3) まず，ノード 22, 33 のデータをそれらが属する SPG を用いて再構築する（33 は，まず 13 をそれが属する FPG を用いて再構築した後に再構築できる）．

(4) その結果を用いてノード 31 の内容を再構築する．

このような流れで，1 つのパリティグループ内に複数のディスク故障が存在する場合のデータ再構築が可能である．再構築に必要なデータやパリティが 1 つでも得られない場合は，再構築は不可能である．上の例では，もしノード 22 のデータが（SPG のパリティノードが壊れているなどの理由で）再構築できなかった場合，ノード 31 の FPG を用いた再構築は失敗となる．FPG, SPG のいずれを用いた場合にも再構築が失敗するような故障が起きれば，そのディスクに格納されたデータは失われる．そのような故障の例としては，あるデータディスクとそれが属する FPG, SPG 双方のパリティディスクが故障した場合や，依存関係の推移閉包が構成される場合（図 3 で 02 ノードが故障した場合）などが考えられる．

### 2.3 データ再構築戦略

故障ディスク内にあったデータを読み出すにはデータの再構築が必要であるが，その場合，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，その逆もありうる．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[10]．

#### 2.3.1 LRS

LRS（Lazy Reconstruction Strategy）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．

#### 2.3.2 ERS

ERS（Eager Reconstruction Strategy）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，すべてのデータノードに書き込み要求があるとした場合には各パリティノードは合計8つのパリティ更新要求を受けることになる．したがって，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固定して保持するのではなく，RAIDのレベル5のようにすべ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[7),8)]．

#### 2.4.1 MPG

1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うことが考えられる．このパリティ分散方式をMPG（Moving Parity Groups）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[9)]．

#### 2.4.2 MPN

もう1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MPN（Moving Parity Nodes）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[9)]．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．

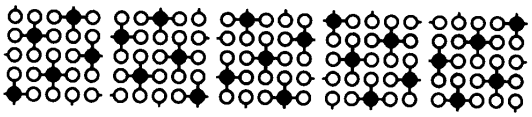

**図4** MPG: Moving Parity Groups（フェーズによるパリティグループの移動）

Fig. 4 Phase switches of the MPG (Moving Parity Groups).

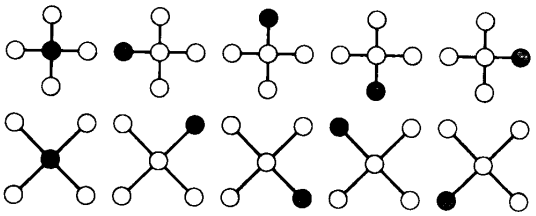

**図5** MPN: Moving Parity Nodes（フェーズによるパリティノードの移動）

Fig. 5 Phase switches of the MPN (Moving Parity Nodes).

## 3. スループットのモデル

ディスク故障が存在せず各ディスクのアクセス頻度が一様なときの，単体ディスク，パリティ固定保持のDR-net，パリティ分散保持のDR-netにおけるそれぞれの平均スループット算出のモデルを提示する．ディスク故障が存在する場合にはシステムに対しデータ再構築のための負荷が加わるが，同じ故障数でも故障ノードの位置により負荷の重さやノード間での負荷の分散の度合が異なるため，単純なモデル化は困難である．

### 3.1 単体ディスク

単体ディスクにおける，1ページの読み出し，書き込みの所要時間をそれぞれ $t_R$, $t_W$ (ms)，また，1回のアクセスデータ長を $L$ (Bytes) とすると単体ディスクの読み出し，書き込みのスループット $T_{SR}$, $T_{SW}$ (KB/s) はそれぞれ

$$T_{SR} = \frac{L}{t_R}, \quad T_{SW} = \frac{L}{t_W}$$

となる．

### 3.2 パリティ固定保持方式

読み出しはパリティディスクを除いたすべてのディスクで並列に処理される．データディスク数を $N_D$ とすると，読み出しスループット $T_{FR}$ は

$$T_{FR} = \frac{N_D L}{t_R + OH_{FR}}$$

となる．式中の $OH_{FR}$ はディスクアクセス以外のオーバヘッドである．

書き込みではパリティの更新がともなう．すべてのデータノードに1回の書き込み要求が出されると，式(1)から分かるように，データディスクでは旧データの読み出しおよび新データの書き込みのためにread, writeが1回ずつ発生する．一方，パリティディスクには前述のように8つのデータノードからパリティ更新要求が集中するため，旧パリティの読み出しおよび新パリティの書き込みのためのread, writeが8回ずつ発生する．したがって全体のスループットは，最も負荷の集中するパリティディスクにおける処理のスループットに依存する．パリティノードが8回のread, writeを行う間にシステム全体で $N_D$ 回の要求を処理

すると考えられるから，

$$T_{FW} = \frac{N_D L}{8(t_R + t_W) + OH_{FW}}$$

となる．

### 3.3 パリティ分散保持方式

読み出しはシステムの全ディスクで並列に処理される．全ディスク数を $N$ とすると，

$$T_{DR} = \frac{NL}{t_R + OH_{DR}}$$

となる．

書き込みの際のパリティ更新は分散して行われる．1回の書き込み操作につき，データの更新，FPGのパリティ更新，SPGのパリティ更新が必要であるから，データノードおよびパリティノードにおいて合計で3回のread，writeが発生する．すべてのディスクに対して1回の書き込み要求が出され，パリティ更新がすべてのディスクで均等に分散されるとすると，各ディスクでread，writeが3回ずつ発生する．3回のread，writeを行う間に $N$ 回の要求を処理すると考えられるから，

$$T_{DW} = \frac{NL}{3(t_R + t_W) + OH_{DW}}$$

となる．

## 4. 性能評価実験

各ノードにトランスピュータT805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[11)~13)]を用いて性能評価実験を行った．実験では，システムに対し1000回のディスクアクセスを要求し，それぞれのアクセスのスループットの平均値を測定した．アクセスはすべ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，アクセスするページはランダムで，ページサイズは0.5~32 KBまで変化させた．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った．また，現在の実験システムの構成はインタフェースノードが1つしかないので，外部とデータを入出力するとインタフェースノードがボトルネックとなってしまう．そこで本実験では各ノードが自ら自分宛のアクセス要求パケットを発行することで，ユーザとのコミュニケーションに代えている．

### 4.1 故障が存在しないとき

実験システムにおいて，故障ディスクが存在しないときの読み出しおよび書き込み操作のスループットを計測した．結果を**図6，7**に示す．比較のため，各構成のスループットモデルの式でオーバヘッドを0とした場合の値も示してある．

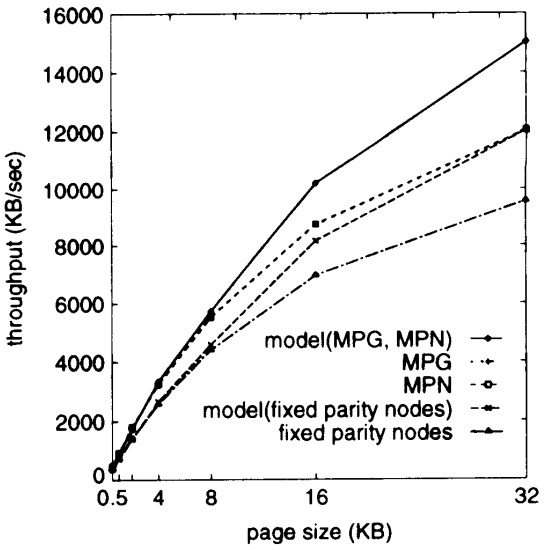


**図6** 各構成の読み出しスループット（ディスク故障なし）
Fig. 6 Throughput of read operations (without disk failures).

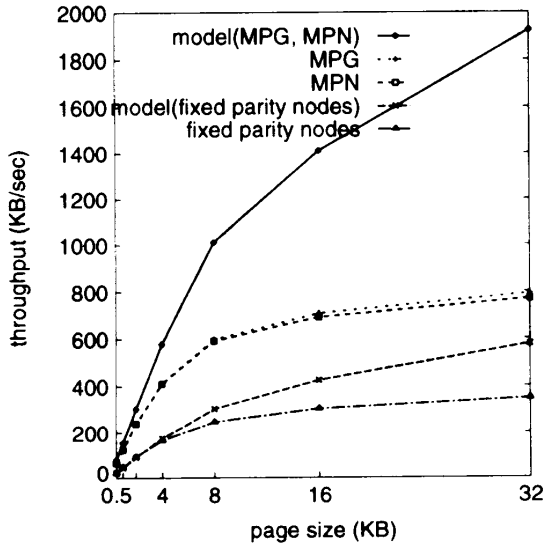


**図7** 各構成の書き込みスループット（ディスク故障なし）
Fig. 7 Throughput of write operations (without disk failiures).

#### 4.1.1 モデルとの比較

図6から，読み出しに関しては実測値とモデル値がかなり一致することが分かる．また，単体ディスクのスループット（ページサイズが0.5，2，8，16，32 KBのときそれぞれ17.8，69.7，230，408，602 KB/sec）と比較すると，ほぼデータディスクの台数倍の性能が得られた．読み出しでは，ページサイズの増加とディスク台数に見合う性能向上が達成されていることが分かる．

一方，書き込みスループットはモデル値よりもかなり低い値となっている（図7）．モデルと比較してオー

バヘッドを算出すると，オーバヘッドの大きさがページサイズに比例していることが分かった（図8）．このことから，オーバヘッドの原因としてはパリティ計算やノード間通信の時間がディスクアクセスで隠蔽されていないことやデータのコピーおよび転送など，データ長に関係するソフトウェアオーバヘッドが考えられる．原因を確かめるため，パリティ計算を省略あるいは通信時に送るデータ量を削減した実験を行った（図9）．この結果，固定パリティノードではパリティ計算を省略することによりモデルに近い値が得られた．

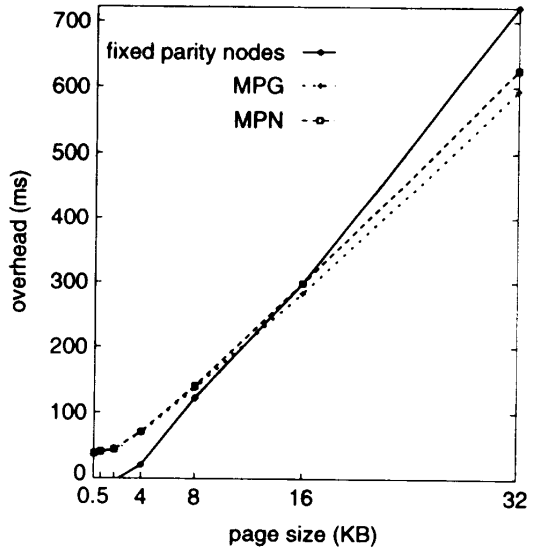


図8 各構成の書き込みオーバヘッド（ディスク故障なし）
Fig. 8 Overhead of write operations (without disk failures).

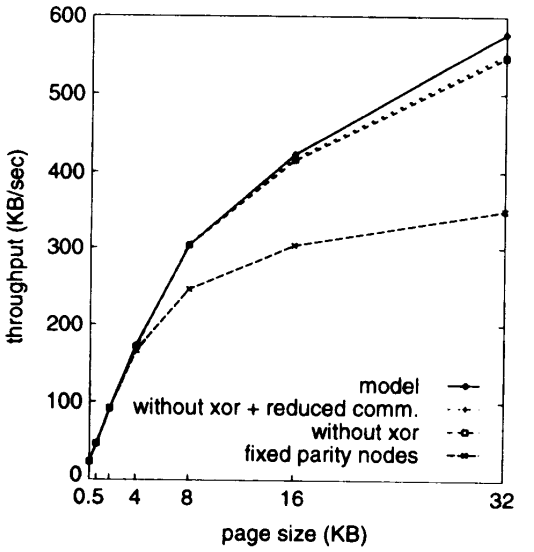


図9 パリティ計算，通信量を削減した固定パリティノードの書き込みスループット（ディスク故障なし）
Fig. 9 Throughput of write operation using fixed parity nodes reducing parity computation and/or communication (without disk failures).

また，パリティ分散を行った場合でも，パリティ計算の省略により大幅なスループット向上が見られた．一方，通信量を削減した場合には，あまり性能は向上しなかった．このことから，トランスピュータを用いた今回の実験システムにおけるモデルと実験結果の不一致の原因は，パリティ計算がディスクアクセスにオーバラップされていないことが原因と考えられる．

### 4.1.2 パリティ分散保持の効果

MPG，MPNを用いたパリティ分散保持によるスループットの向上について考察する．読み出しスループットの改善は，データを保持するディスク台数の増加により，パリティを分散しない場合に比べ，より高並列な動作が可能となったことによる．書き込みでは，それに加えてパリティ更新負荷の分散がスループットを向上させている．各ページ長でのパリティ分散をしない場合に対する向上率は**表1**のようになり，パリティ分散保持の効果が確認できる．一方，モデルから書き込みスループットの向上率 $T_{DW}/T_{FW}$ を求めると

$$\frac{T_{DW}}{T_{FW}} = \frac{NL}{3(t_R+t_W)+OH_{DW}} \times \frac{8(t_R+t_W)+OH_{FW}}{N_D L}$$

$OH_{FW}$，$OH_{DW}$ を0とすると，$N=25$，$N_D=20$ であるから，

$$\frac{T_{DW}}{T_{FW}} = \frac{25}{3} \times \frac{8}{20} = \frac{10}{3}$$

したがって，スループットは10/3倍になると期待できる．この値と実測値との差は書き込みにおけるオーバヘッドが0でないことによると思われる．

また，図7でMPGとMPNを比較すると，ページサイズが大きくなるとMPGの方が若干高いスループットを示している．これは，MPNの方がMPGに比べて各データノードからパリティノードまでの平均リンクホップ数が大きいため，ノード間通信やメモリ内でのデータ転送の時間が大きくなることが原因であると考えられる．

表1 パリティ固定保持方式に対する，パリティ分散方式の書き込みスループットの向上率
Table 1 Throughput ratio of distributed parity to fixed parity for write operations.

| page size (KB) | MPG/fixed | MPN/fixed |
|---|---|---|
| 0.5 | 2.67 | 2.68 |
| 2 | 2.58 | 2.60 |
| 8 | 2.42 | 2.40 |
| 16 | 2.33 | 2.27 |
| 32 | 2.26 | 2.20 |

### 4.2 ディスク故障が存在するとき

実験における故障ディスク数は0〜7とした．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データ再構築の際の負荷が異なり，また，再構築が不可能な場合もある．ここでは，再構築不可能な故障パタンでの性能については考慮せず，各パリティ保持方式で再構築が可能なパタンについて実験を行った．その際，すべ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10の再構築可能な位置パタンを選び，それらの結果の重み付き平均値を評価結果とした．パタンの選択は，各パタンについてすべてのノードをアクセスしたときにシステムで発生するディスクアクセス数を調べ，その値が近いものは1つにまとめるようにし，各故障数ごとに最大10のパタンで代表させた．その代表パタンについて実験を行い，それぞれまとめたパタンの数だけ重みをつけて全体の平均をとった．また，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，故障はただちに検出されシステムはデータ再構築等の故障時動作を行う．

なお，今回スペアドライブは前提としていないが，ディスク故障が発生した場合，故障ディスクをシステムにあらかじめ用意してあるスペアドライブに切り替え，データの再構築を行うことも考えられる[14)〜16)]．DR-netにおけるスペアドライブの取扱いは今後の課題である．しかし，非常に大規模なシステムでは高い頻度でディスク故障が発生することが予想でき，故障ディスクをそのつど交換せずある期間ごとに保守を行うような運用の場合，スペアドライブを使い果たしディスク故障が存在する状態での性能が重要となる．ここで示す結果はそのような状況における性能としてみることができる．

#### 4.2.1 2つの再構築戦略の比較

**図10**はパリティをMPGによって分散したときの読み出しスループットに関して，ERS，LRSの2つの再構築戦略を比較した結果である．従来，ERSは2つのパリティグループで並列に再構築を行い，2つの結果の早く得られた方を利用するため，LRSよりも処理時間が短く，レスポンス時間は短いがスループットは低いと予測していた．しかし，文献17)でレスポンスはLRSの方が良いことが判明した．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うことによる他ノードの負荷増大が予想以上に処理時間を増加させ，その結果LRSの方が良い性能を示したと考えられる．今回の実験でスループットについても，パリティを固定保持した場合，およびMPG，

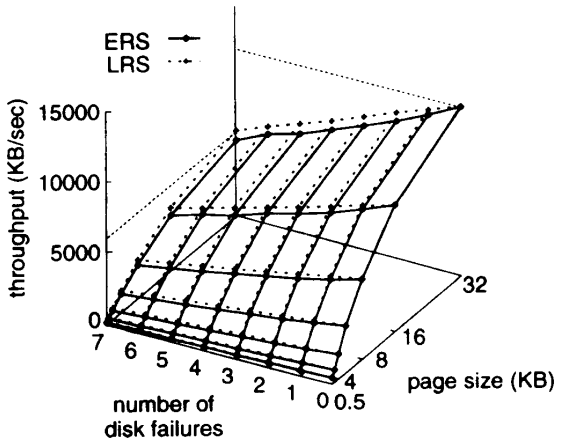


**図10** LRS，ERSの比較（MPGでの読み出し）
Fig. 10 Comparison between LRS and ERS (for read operations with MPG).

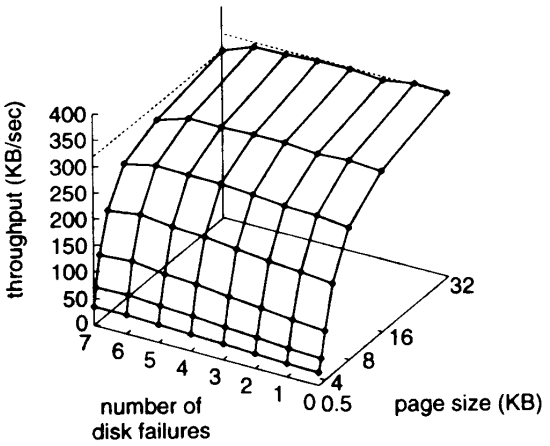


**図11** パリティ固定保持の書き込みスループット（ディスク故障あり）
Fig. 11 Throughput of write operations with fixed parity nodes (with disk failures).

MPNによって分散保持した場合のいずれにおいてもLRSの方が良いことが示された（図10ではMPGでパリティを分散する場合のみを示したが，パリティを固定保持した場合，MPNで分散保持した場合にも同様の結果が得られた）．

#### 4.2.2 ディスク故障が存在する場合のパリティ固定保持方式の書き込み性能

**図11**に故障が存在する場合のパリティ固定保持方式での書き込みスループットの変化を示す．パリティを分散させた場合には，故障ディスク数の増加にともなって書き込みのスループットも緩やかに低下していくが（図13参照），パリティ固定保持方式では，スループットが故障数に関して単調に低下せず，故障数が6以下ではほとんど性能が劣化していないことが分かる．正常なディスクへの書き込みにともなうパリティ更新は式(1)で計算されるが，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ともな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．したがって，書き込みを要求されたディスクが故障していると，

- 負荷の重いパリティノードにおける負荷の軽減
- 負荷の軽い他のデータノードにおける負荷の増大

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，パリティ固定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，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スループットが減少すると考えられる．

### 4.2.3 MPGとMPNの比較

故障ディスクがあるときのMPGとMPNを比較すると（**図12, 13**），故障数が増加するにつれてわずかながらMPNの方が良い性能を示すようになる（読み出しではLRSを用いた場合を示したが，ERSを用いた場合も同様の結果が得られた）．この差はわずかではあるが，ページサイズなどの条件を変化させたり，同一条件で複数回実験を行った場合に，いずれの結果も同じような傾向を示す．原因の1つとして，故障ディスクのデータを再構築するために必要なディスクアクセス数が考えられる．ある故障パタンで各フェーズごとにすべてのディスクに1回ずつアクセス要求を出したときにシステム全体で合計何回のディスクアクセスが起きたか数え，各故障数ごとに平均すると**図14**のようになる．これらの図から分かるように，MPGはつねにMPNよりも多くのディスクアクセスが必要となっている．このことが原因でMPGよりもMPNの方が良い結果を出したと考えられる．同様に，書き込みに関しても**図15**に示すようにMPGの方が多くのディスクアクセスが必要である．しかし，故障ディスクがないときの書き込みではMPGの方がMPNよりも良い性能を示しているため（図7），故障数が少ない場合の多少のアクセス数増加による影響は打ち消されてしまいMPGが優位を保つ．しかし，故障数が多くなるとアクセス数増加の影響が顕著になりMPNの方が良い結果を出すと考えられる．

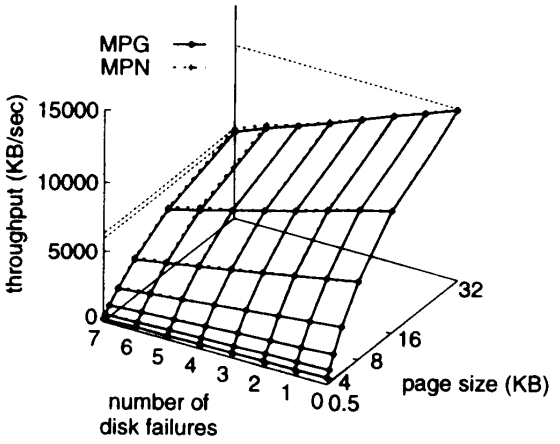


**図12** LRSを用いた読み出しスループットでのMPGとMPNの比較

Fig. 12 Throughput comparison between MPG and MPN for read operations with LRS.

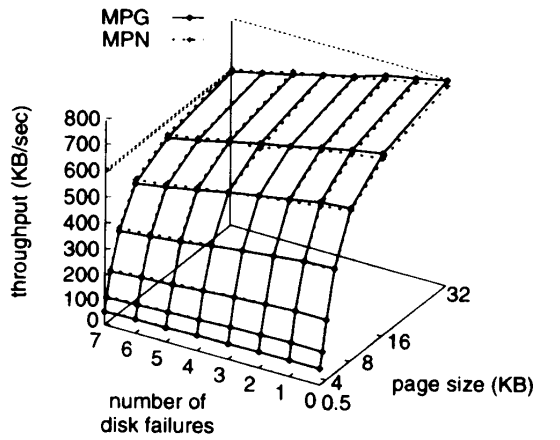


**図13** 書き込みスループットでのMPGとMPNの比較

Fig. 13 Throughput comparison between MPG and MPN for write operations.

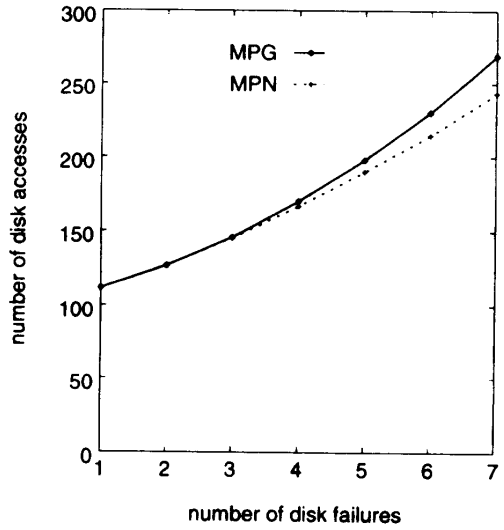


**図14** LRSを用いた読み出しに必要なディスクアクセス数の比較

Fig. 14 Comparison the numbers of disk accesses in read operations with LRS.

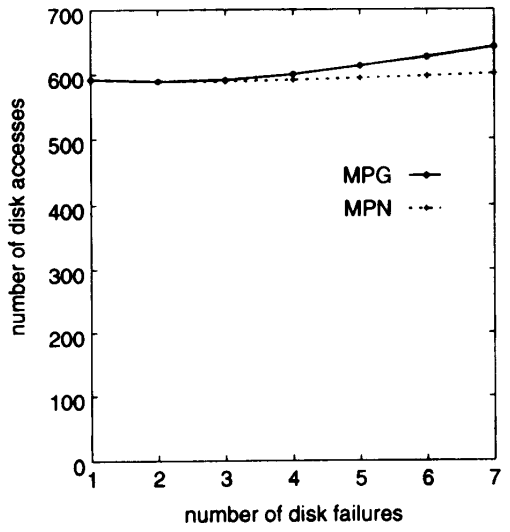


図15 書き込みに必要なディスクアクセス数の比較
Fig. 15 Comparison the numbers of disk accesses in write operations.

## 5. おわりに

システムスループットのモデルを示し，DR-netにおけるパリティ分散保持の効果とその2つの方式の違い，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実験について報告した．

実験により，読み出しに関しては，モデルと比較した場合，パリティ保持方式にかかわらずページサイズの増加とディスク台数に見合うだけの性能が得られることが確認された．

一方，書き込みに関してはどの方式でもモデルと実験結果の比較からオーバヘッドが存在することが分かった．そのオーバヘッドは主にパリティ計算がディスクアクセスにオーバラップされていないことが原因であることが判明した．また，MPG，MPNを用いたパリティ分散保持方式では，いずれも書き込み性能は改善されたが，特にページサイズが大きいときは書き込み処理で生じるオーバヘッドによりモデルから導かれる性能には及ばなかった．

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分かった．また，ディスク故障が存在する場合のパリティ固定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負荷が軽減されることから性能がほとんど劣化しないが，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，性能が悪化することが判明した．MPGとMPNの比較では，故障がない場合の書き込みではMPGの方が優れていることが分かった．

今後，ディスク故障が発生した場合に利用するスペアドライブをシステムに組み込む必要がある．また，スペアドライブにデータを再構築する際の戦略や，そのときのシステムの性能を評価しなければならない．また，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を実装し評価する予定である．システム構成の面では，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が，DR-netの概念は2次元トーラスに依存するものではなく，他のネットワークトポロジにも適用できる．したがって，別の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．また，故障ディスクを取り替えた後の再構築過程や複数のインタフェースノードを用いたときのファイル管理法などについても検討したい．

## 参考文献

1) Patterson, D.A., Gibson, G. and Katz, R.H.: A Case for Redundant Arrays of Inexpensive Disks (RAID), *Proc. ACM SIGMOD Conference*, pp.109–116 (1988).
2) Chen, P.M., et al.: RAID: High-Performance, Reliable Secondary Storage, *ACM Computing Surveys*, Vol.26, No.2, pp.145–185 (1994).
3) Burkhard, W.A. and Menon, J.: Disk Array Storage System Reliability, *Digest of Paper FTCS 23*, pp.432–441 (1993).
4) Gibson, G., Hellerstein, L., Karp, R.M., Katz, R.H. and Patterson, D.A.: Coding Techniques for Handling Failures in Large Disk Arrays, *ASPLOS-III*, pp.123–132 (1989).
5) Blaum, M., Brady, J., Bruck, J. and Menon, J.: EVENODD: An Optimal Scheme for Tolerating Double Disk Failures in RAID Architectures, *Proc. 21st ISCA*, pp.245–254 (1994).
6) 横田治夫：RAIDのネットワーク上への展開と信頼性向上，信学技報，CPSY 93–11, pp.79–86 (1993).
7) 横田治夫：データ再構築ネット（DR-net）における不均衡対策，信学技報，FTS 93–20, pp.9–16 (1993).
8) Yokota, H.: DR-nets: Data-Reconstruction Networks for Highly Reliable Parallel-Disk Systems, *Proc. 2nd Workshop on I/O in Parallel Computer Systems*, pp.105–116 (1994). (also in *ACM Computer Architecture News*, Vol.22, No.4).

9) 味松，横田：並列ディスクシステムのパリティグループの構成の変化と信頼性の比較，信学技報，FTS 95-34, pp.25-32 (1995).
10) 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).
11) Tomonaga, S. and Yokota, H.: An Implementation of a Highly Reliable Parallel-Disk System using Transputers, *Proc. 6th Transputer/Occam Intn'l Conf.*, IOS Press, pp.241-254 (1994).
12) Yokota, H. and Tomonaga, S.: The Performance of a Highly Reliable Parallel Disk System, *Proc. World Transputer Congress '94*, Gloria, A.D., Jane, M.R. and Maini, D. (Eds.), pp.147-160, IOS Press (1994).
13) 横田，友永：高信頼並列ディスクプロトタイプのアクセス性能，信学技報，FTS 94-37, pp.17-24 (1994).
14) Munts, R.R. and Lui, J.C.S.: Performance Analysis of Disk Arrays Under Failure, *Proc. 16th VLDB*, pp.162-173 (1990).
15) Menon, J. and Mattson, D.: Comparison of Sparing Alternatives for Disk Arrays, *Proc. 19th ISCA*, pp.318-329 (1992).
16) Holland, M., Gibson, G.A. and Siewiorek, D.P.: Fast, On-Line Failure Recovery in Redundant Disk Arrays, *Digest of Paper FTCS 23*, pp.422-431 (1993).
17) 味松，横田：DR-net におけるパリティ配置/データ再構築戦略の影響，信学技報，FTS 94-60, pp.25-30 (1994).

(平成 7 年 9 月 1 日受付)
(平成 8 年 3 月 12 日採録)

**味松 康行**

平成 5 年東京工業大学工学部電気電子工学科卒業．平成 7 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，同大博士後期課程在学中．ディスクアレイシステム等の大規模二次記憶システムの高性能化，フォールトトレランスおよび並列データベースシステムに興味を持つ．

**横田 治夫**（正会員）

昭和 55 年東京工業大学工学部電子物理工学科卒業．昭和 57 年同大大学院理工学研究科情報工学専攻修士課程修了．同年富士通（株）入社．同年 6 月（財）新世代コンピュータ技術開発機構研究所．昭和 61 年（株）富士通研究所勤務．平成 4 年北陸先端科学技術大学院大学情報科学研究科助教授．工学博士．主としてデータベース，データ工学向けの並列アーキテクチャ等に関する研究に従事．電子情報通信学会，人工知能学会，IEEE，ACM 各会員．