RAIDコントローラ T380

T380 RAID Controller for MAGNIA$_{TM}$ Series of Dual-Processor Servers

大森　幹雄　　川村　和也　　藤本　真吾
■ OMORI Mikio　■ KAWAMURA Kazunari　■ FUJIMOTO Shingo

MAGNIA$_{TM}$シリーズに使われている東芝製RAID（Redundant Array of Independent（Inexpensive）Disks）技術のRAID Master$_{TM}$に基づき，RAIDコントローラ T380を開発した。RAID Masterは，高性能，高信頼性，高可用性（高耐障害性），及び保守性向上を狙ったもので，ソフトウェア（SW）RAIDのMAGNIA ATA RAIDを製品化し，ハードウェア（HW）RAIDであるT380にも適用した。

T380は，デュアルコアI/O（Input/Output）プロセッサなど最新のデバイスを採用し，通常動作とは独立した診断プログラムがRAIDコントローラの状況を監視することで，高性能と高信頼性を実現している。また，磁気ディスク装置（HDD）が2台故障してもデータを失わないRAID 6をサポートし，エラーの頻度が高くなった場合に，診断プログラムの結果を利用して自動的に待機系HDDにデータをコピーする予防保全機能があり，簡単にデータを失わないように可用性を高めている。また，障害発生時に解析情報を記録しており，コントローラを解析すれば障害情報が得られる工夫を行っている。更に，バッテリーは外部型と内蔵型の2種類を用意し，簡単に交換できるようにして保守性の向上を図った。

Toshiba has developed the T380 redundant array of inexpensive disks (RAID) controller based on RAID Master$_{TM}$, a collective name for our RAID technology that aims to achieve high performance, high reliability, high availability, and high maintainability and is implemented in the MAGNIA advanced technology attachment (ATA) RAID as well as in the T380.

The T380 RAID controller provides high performance and reliability with the introduction of a diagnostic program, which constantly inspects the status of the controller, and the latest dual-core input/output (I/O) processors. It also provides high availability by supporting both RAID 6, which can recover data from the loss of two hard disk drives (HDDs), and a preventive maintenance function to automatically copy data into a standby HDD in case the error occurs frequently. Information on abnormalities obtained by analysis of the controller is stored on flash memory as error information. Furthermore, the system is equipped with two types of battery—an external type and a built-in type—that can be easily exchanged, to improve maintainability.

1　まえがき

RAIDコントローラ T380は，東芝製RAID技術であるRAID Masterに基づいたHW-RAIDコントローラである。T380は，デュアルコアI/Oプロセッサなどの最新のデバイスを採用し，通常動作とは独立した診断プログラムがRAIDコントローラの状況を監視することで，高性能と高信頼性を実現している。また，待機系HDDにデータをコピーする予防保全機能があり，簡単にデータを失わないように可用性を高めている。

ここでは，T380の特長，仕様，及びRAS（Reliability（信頼性），Availability（可用性）and Serviceability（保守性））機能などについて述べる。

2　RAIDコントローラ T380の特長

T380は，次の四つの特長を持っている。

（1）　高性能　　デュアルコアIntel®(注1) IOP348を搭載し，主記憶，ディスクキャッシュ用にバッテリーによる電源バックアップ機能を備えたDDR2（Double Data Rate 2）メモリや8ポートのSAS（Serial Attached SCSI（Small Computer System Interface））1.1 HDDやPCI（Peripheral Component Interconnect）Express X8などの最新のコンポーネントを搭載している。

（2）　高信頼性　　バックグランドでRAIDカード上のメモリやHDDのパトロール診断を行って障害を未然に検出し，代替処理を行う。当社が開発したRAID技術であり，障害発生時の対応やフィールド情報の迅速なフィードバックなどができる。更に，外部機能評価を中心としたブラックボックス評価とともに，ホワイトボックス評価(注2)を行って品質向上を図っている。

（3）　高可用性　　HDD2台の故障に対応するRAID 6（リリース予定）のサポートを行う。また，2台の故障に対応

（注1）　Intelは，米国及びその他の国における米国Intel Corporation又は子会社の登録商標又は商標。
（注2）　システム内部の構造を理解したうえで，それら一つ一つが意図したとおりに動作しているかを確認する評価。

できないRAIDレベルにおいては，リビルド処理(注3)時に2点障害，すなわち最初の障害が発生後，コピー元に更に障害が発生した場合でも，バットスポット管理(注4)によってリビルド処理を継続し，コピー元の障害のないデータをエラーとしないことでデータをすべて失わない対策を行っている。また，予防保全機能によって，SMART (Self-Monitoring, Analysis and Reporting Technology) だけではなく，HDDの代替回数の頻度など障害情報を統計的に検査し，ホットスペア (待機HDD) がある場合は自動的にデータをコピーして，未然に障害を防ぐ機能も実装している。

(4) 保守性の向上　　障害が発生したときに障害情報や操作の手順などを記録するディスクログ機能と，ファームウェア (FW) 停止時にRAIDカード内のメモリの内容をRAIDカード上のフラッシュメモリにダンプするメモリダンプ機能がある。これらの情報を解析することで障害の状況を容易に把握できる。障害情報がカード内にあることで，カードが返却されればこれらの情報を得られるようになった。また，メモリ電源バックアップ用バッテリーをカード外部に実装できるようにしたことで，定期的に必要となるバッテリー交換の作業性を向上させた。

3　RAIDコントローラ T380の仕様

T380は，MAGNIAシリーズのDP (Dual Processor) サーバに搭載される。T380は，当社独自開発のHWとSWを用いており，他社に勝る高い信頼性を確保するとともに，更に，可用性と保守性を向上させたユーザーメリットのある機能を盛り込んだ。また，チューニングを十分に行い，高いディスクアクセス性能を実現した。T380の外観を**図1**に，主な仕様を**表1**に示す。また，ブロック図を**図2**に示し，その詳細を以下に述べる。

3.1　I/Oプロセッサ

T380では，I/OプロセッサにデュアルコアIntel® IOP348を採用した。このI/Oプロセッサはデュアルコア構成で，片方のコアにRAID FWを，もう一方のコアにはI/Oインタフェースを実行するFWを実装し，負荷の分散を図っている。

従来のRAIDコントローラは，I/OプロセッサとI/Oインタフェースの2種類のチップ構成で作られることが多かったが，ワンチップ構成で組むことができるため部品点数が減少し，信頼性が向上した。

3.2　I/Oインタフェース

T380はSAS 1.1をサポートしている。SFF-8087(注5)準拠のコネクタとして2チャネル搭載し，8台のHDDを制御することができる。

HDDのアクセス状況や障害の状態を表示する発光ダイオード (LED) を制御するために，SGPIO (Serial General Purpose Input/Output) もサポートしている。

3.3　電源バックアップ機能付きメモリ

T380は，256 MバイトのDDR2メモリを搭載している。このメモリは，I/Oプロセッサのワーク領域やディスクキャッシュとして利用されている。ディスクキャッシュとしてはライトバックキャッシュもサポートしている。ライトバックキャッシュでは，メモリ内にHDDに書き込むべきデータが残っており，不意の停電などによってそのデータが消えてHDDに書き戻せない場合の対策として，バッテリーによる電源バックアップが行われ

(注3)　HDDが故障した場合に，HDD交換後にコピー元HDDからデータを復元すること。
(注4)　リビルド中にデータを復元できなかった場所を記録する方法。
(注5)　4チャネルのSAS HDDを接続できるコネクタ規格。

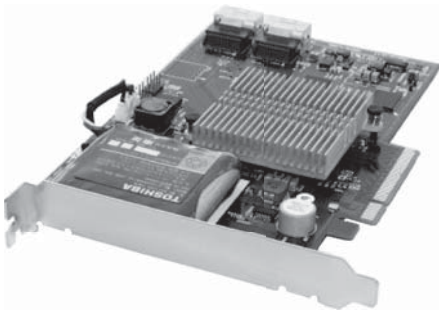
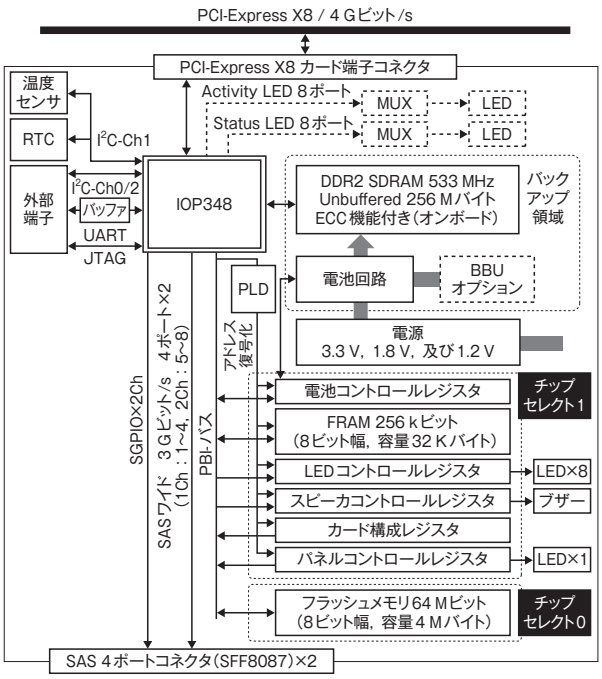


図1. RAIDコントローラ T380 — 従来のコントローラは，I/OプロセッサとSCSIコントローラの2チップ構成であったが，T380はI/Oプロセッサの中にSCSIにあたる部分を集約して，1チャネル構成になっている。
T380 RAID controller

表1. RAIDコントローラ T380の主な仕様
Main specifications of T380 RAID controller

項　目	仕　様
I/Oプロセッサ／ホストインタフェース	Intel® IOP348 (デュアルコア) PCI-Express Rev1.0a×8
キャッシュメモリ	PC2-4200 (DDR2-533) 256 Mバイト ECC機能付き (オンボード(*1))
HDDコントローラ	SAS1.1/SATA Ⅱ 8ポート
HDDインタフェース	SAS1.1
HDDコネクタ	SFF-8087準拠×2チャネル
サーバ当たりの搭載可能枚数	1枚
HDD接続台数	8台
サポートRAIDレベル	RAID 0, 1, 5, (6), 10, 50
RAS	SGPIO
動作温度	10～35 ℃(*2)
動作湿度	20～80 %（ただし，結露ないこと）
バッテリーバックアップ時間	72 h (256 Mバイト：初期購入時)
外形寸法	167.65×111.15 mm (スタンダードハイト ハーフレングス)

＊1：LSIチップが，マザーボード基板上に直接搭載されていること
＊2：組み込んだ状態での外気温度



RTC : Real Time Clock
I²C : Inter-Integrated Circuit
Ch : チャネル
UART : Universal Asynchronous Receiver and Transmitter
JTAG : Joint Test Action Group
Activity : HDDアクセス信号
Status : HDDの正常，縮退などを表示
MUX : Multiplexer
SDRAM : Synchronous DRAM
ECC : Error Correcting Code
BBU : Battery Backup Unit
PLD : Programmable Logic Device
PBI : Peripheral Bus Interface
FRAM : FeRAM

図2. RAIDコントローラ T380の構成 — プロセッサにデュアルコアIntel®IOP348，主記憶及びディスクキャッシュ用にDDR2メモリ，その他8ポートのSAS 1.1 HDDやPCI Express X8など最新コンポーネントで構成されている。

Configuration of T380 RAID controller

ている。バックアップ時間は72時間可能であり，時間内に復帰すれば次の起動時にキャッシュのデータは書き戻される。

3.4 バッテリー

メモリの電源バックアップには，耐温度性能，保管性能，輸送条件などにおいてリチウムイオン電池に優るニッケル水素電池を選択し，実装している。バッテリーの実装は，RAIDカード内実装型と外部実装型の2種類の方法を提供し，使用するシステムの形態に合わせた最適な方法が取れるようになっている。特に外部実装型バッテリーは，交換に際し，内部実装型に比べて作業性の向上が図られている。

4 RAS機能

この章では，T380に搭載されている代表的なRAS機能について詳細を述べる。

4.1 パトロール機能

システムの安定運用のためには，定期的な診断を行うことが欠かせない。T380では，通常動作とは別にバックグランドで，HDD全体を2，4，又は6週間の選択式で1周する，メディア検査を行っている。

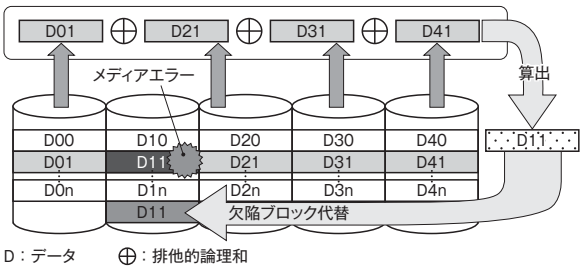


図3. 冗長性を利用したデータ修復 — データが壊れたときに，残りのディスクから冗長性を利用して壊れたデータを修復し，書き込む。

Data recovery using redundant HDDs

このメディア検査をパトロール機能と言い，その目的は二つある。一つは，HDDの高密度化で特定の部分にアクセスが集中すると，隣接した部分が読込みエラーになってしまうITI(Inter-Track Interference)が発生するが，この発生確率を軽減することである。ITIが起こっても，HDDは読込みすれば弱っている部分を再書込みする修復機能を持っているので，パトロール機能によって定期的にまんべんなくHDD全体を読込みすることは有効である。

もう一つは，HDDには一部のセクタ(1セクタは512バイト)が読み込めなくなるメディアエラーが発生することがあり，この問題を対策することである。T380は，**図3**に示すように冗長構成(注6)に組んであり，パトロール機能でメディアエラーを検出した場合には，HDD上にある代替ブロックを使って代替処理を行い，メディアエラーの起こった場所にアクセスしないようにする。更に，残ったHDDの内容を使ってデータを復元し，新たに代替した場所に書き戻す処理を行う。

パトロール機能によって，HDDの障害を事前に検出し，HDDが代替処理できるかぎりはエラーにすることがないため，可用性は高くなる。この機能はHDD全域を対象として行われるため，メディアエラーによるリビルド中断も軽減する。更に，メディアエラーを潜在的に持っている場所に突然アクセスすることが少なくなるため，容量が増えていっても，事前確認によってエラーとの遭遇を防ぐことができる。また，I/Oが低負荷なときに実施されるため性能への影響は少なく，ユーザーが意識することなく，設定されたスケジュールに従い実施される。

4.2 予防保全機能

HDDの故障予測としては，SMARTが一般的である。SMART情報も判断材料となるが，T380では，4.1節で述べたパトロール機能と組み合わせてメディアの代替情報に一定のしきい値を設け，HDDの故障が予測される場合には，RAID

(注6) 構成を一重から二重にすることで，片方に障害などが起きても通常の動作が継続できるようにした構成。

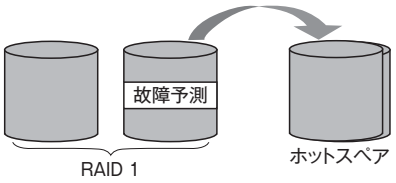


図4. 予防保全の概念 — パトロール機能によって代替セクターが急増したと診断された場合に，該当するHDDの内容をバックアップ系のHDDにコピーし，HDD障害を軽減する。

Concept of preventive maintenance

の冗長性を保ったままでそのHDDのデータをホットスペアへコピーし，安全に交換を行う機能を持っている（**図4**）。

4.3 RAID 6機能

従来のRAIDコントローラには，冗長構成としてRAID 1（ミラーリング）やRAID 5（分散データとガーディング（保護））が用いられていた。これらのRAIDレベルでは，1台のHDDが故障した場合だけデータが復旧できる。HDDが壊れたとき復旧作業をするリビルド時において，読込み側と復旧側のHDDが故障した場合は，バットスポット管理などによってデータの全体を失わないで，読み込めるところは復旧するなどいろいろ工夫しているが，完全な復旧ができなかった。

そこで，RAID 5を拡張し，2台のHDDが同時に故障してもデータが復旧可能なRAID 6のサポートを行う。RAID 6では，パリティ（データの誤り検出）用に使用するHDD台数が2台必要になるが，2台の故障に対応できるので，その分信頼性は大幅に向上することになる。

T380のRAID 6の機能では，**図5**に示すように“P+Q”は，算出方法の異なる2種類のパリティ（PとQ）を，RAID 5と同様に複数のHDDをまたぐように格納する方式を取っている。HDDが故障した場合は，この2種類のパリティを利用してRAIDの再構築ができる。2種類のパリティを計算することから書込み時の計算負荷は高いが，ハードウェアによるP+Qの計算サポート機能をIOP348に利用することでこの問題を克服した。2種類のパリティをすべてのHDDでローテーションしながら書き込むため，HDDに対するアクセス負荷が均一になり，

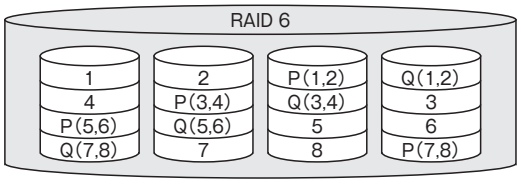


図5. RAID 6のP+Qパリティ配置 — 従来は，一般的にRAID 5やRAID 1などの1台の冗長性しか対応していなかったが，T380ではRAID 6によりパリティが2段になるので，2台の冗長性を拡張している。

P and Q parity locations of RAID 6

いっそう高速化することができた。

4.4 カード内フラッシュメモリへのメモリダンプ書込み機能

T380は，HWが致命的なエラーを検知したりFWが矛盾を検出すると，データ破損などの二次被害を防ぐため，FWを停止する。今回，FW停止時に，障害データの概要をRAIDカード上のNVRAM（Non Volatile RAM）に保存するとともに，**図6**に示すように，RAID制御領域のメモリダンプをRAIDカード上のフラッシュメモリに保存する機能を実装した（ただし，ユーザーデータが含まれるキャッシュ領域は保存しない）。これにより，障害解析でRAIDカードだけ返却された場合にも，解析を進めることができるようになった。

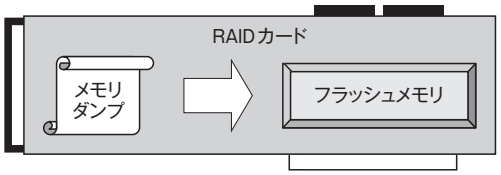


図6. メモリダンプをカード内のフラッシュメモリへ書き込む機能 — メモリダンプをカード内のフラッシュメモリへ書き込んで，後日カードからデータを取得する。

Function to move memory dump files to flash memory in card

5 あとがき

RAIDコントローラT380は，RAID Masterに基づいて，高性能，高信頼性，高可用性，及び保守性向上の四つの特長を兼ね備えている。現行のMAGNIAシリーズと組み合わせ，これらの特長を生かせる業種又は業界で様々なユーザーに広く使われる製品として，そのラインアップの拡充を図っていく。今後は，更に市場ニーズに応える製品の開発に注力していく。

大森　幹雄　OMORI Mikio

PC&ネットワーク社 PC開発センター サーバ・ネットワーク設計部主務。サーバ及びRAID関連のファームウェアとハードウェア，ソフトウェアRAIDのBIOS，ハードウェアRAIDの開発に従事。

PC Development Center

川村　和也　KAWAMURA Kazunari

PC&ネットワーク社 PC開発センター サーバ・ネットワーク設計部主務。IAサーバMAGNIA向けRAIDコントローラの設計・開発に従事。

PC Development Center

藤本　真吾　FUJIMOTO Shingo

PC&ネットワーク社 PC開発センター サーバ・ネットワーク設計部主務。IAサーバMAGNIA，RAID関連のハードウェア設計・開発に従事。

PC Development Center