高い耐故障性を実現した2重化計算機

7L−2

市川純一 (川崎製鉄㈱)　合田雅直 (川崎製鉄㈱)　吉澤功 (川崎製鉄㈱)　白岩正文 (川崎製鉄㈱)
山本浩文 (川崎製鉄㈱)　W.D. Shambroom (Charles River Data Systems, Inc.)

1．まえがき

デュープレックス方式による2重化により高い耐故障性を有する2重化計算機システムを開発した。本システムは電源部，CPU部，磁気ディスク部をモジュール構造化し活性保守を可能としている。　またソフトウェアによる2重系相互間の監視による故障検出，故障モジュールの切離し，回復処理についての高速化，高信頼化を図っている。以下本システムの概要と耐故障性機能について報告する。

2．システム構成

図1に本システムの基本的な構成を示す。I486 CPUとEISA(Extended Industry Standard Architecture)バスを使用した，2系統の標準的な計算機から構成される。磁気ディスクはミラー（2重）化されSCSIバスのマルチホスト機能を利用して，A，B両系からアクセス可能である。電源部は3台の電源を並列接続しておりどれか1台が故障してもシステムは正常に動作可能である。RS232C回線を使ったハートビートリンクによりA，B両系は相互にメッセージを交換して，故障監視を行う。システムコントローラは自系と相手系の電源ON／OFF制御および各種スイッチ等制御回路のコントロールと温度等の環境監視を行う。

また，図1の点線で囲った磁気ディスクと電源については，故障時には個別に活性保守が可能である。CPU，主記憶，バス，各種コントローラは各系毎に1体化してモジュール化されており，正常系動作中に故障モジュールの交換修理（活性保守）が可能である。

3．ハードウェアの信頼性評価

図2は本システムの基本ハードウェア構成の主要部をモデル化したものである。CPU−MはCPUと主記憶，コントローラ等を含む。信頼性の評価尺度としては，本システムのような冗長システムの評価には必ずしも最適ではないが，MTTF(Mean Time To Failure)を用いる。

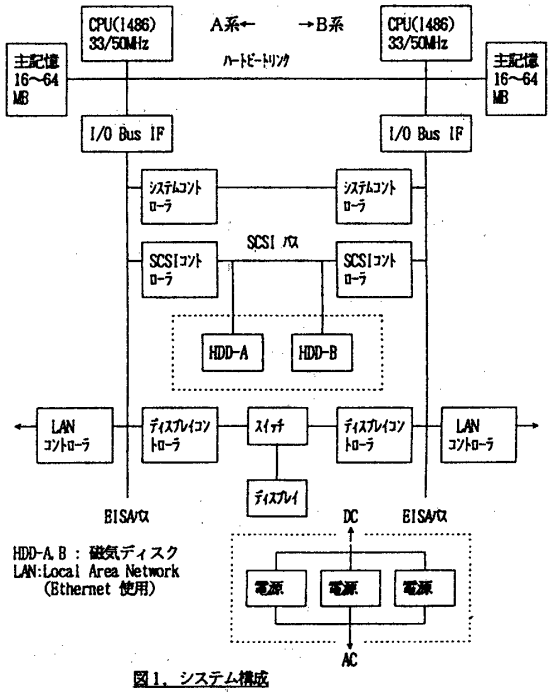

図1．システム構成

電源については3個の内2個が正常であればシステムとして正常であり，CPU−MとHDDについては2個の内1個が正常であればシステムとして正常であると仮定する。

システムのMTTFは以下のように求められる。[1]

$$\frac{1}{MTTF\text{-}system} = \frac{1}{\frac{1}{6}\frac{MTTF\text{-}ps^{**}2}{MTTR\text{-}ps} + \frac{5*MTTF\text{-}ps}{6}} + \frac{1}{\frac{1}{2}\frac{MTTF\text{-}cpu\text{-}m^{**}2}{MTTR\text{-}cpu\text{-}m} + \frac{3*MTTF\text{-}cpu\text{-}m}{2}} + \frac{1}{\frac{1}{2}\frac{MTTF\text{-}hdd^{**}2}{MTTR\text{-}hdd} + \frac{3*MTTF\text{-}hdd}{2}}$$

MTTR : Mean Time To Repair

MTTF-ps = MTTF-cpu-m = MTTF-hdd = 20,000 Hour,

MTTR-ps = MTTR-cpu-m = MTTR-hdd = 48 Hour

として計算すると，MTTF-system = 841,740 Hour

Duplex computer which realizes highly fault resiliency

Junichi Ichikawa, Masanao Gouda, Isao Yoshizawa, Masafumi Shiraiwa, Hirofumi Yamamoto :Kawasaki Steel Co.
W.D Shambroom :Charles River Data Systems, Inc.

となる。ここで使用した各モジュールのMTTFの値は実際より低めに見積っており，この結果から本システムの信頼性はアーキテクチャ的には問題ないことがわかる。
しかしながら現実には，ソフトウェアの障害，操作ミス，両系間の共通部品の故障等の問題で，障害の発生頻度は，これよりかなり高くなるものと思われる。

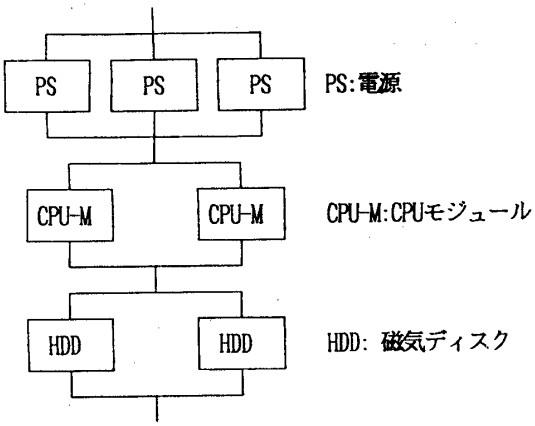

図2．本システムの信頼性評価モデル

4．故障監視，回復機構

本システムでは，図3のように主系，従系の各系でハートビートモジュールと呼ばれるプログラムが動作し，一定時間間隔でメッセージを交換することにより，相互の異常監視を行い，タイムアウト検出または相手系からの通知により故障検出を行う。

各系のハートビートモジュールは次の１０種類の状態の内の１つをとり，一定周期で自分の状態を相手系に送る。相手系は受け取った状態の内容によって状態遷移を引き起こす。このようにしてシステム全体として有限状態マシンとして動作する。

状態０　ＮＵＬＬ（初期状態）
状態１　ＰＩ（主系として起動途中の状態）
状態２　ＳＩ（従系として起動途中の状態）
状態３　ＰＲ（主系として運用中の状態）
状態４　ＳＲ（従系として運用中の状態）
状態５　ＲＰ（相手系の再起動を試みている状態）
状態６　ＳＰ（相手系を停止させている状態）
状態７　Ｆ（自ら故障を検出した状態）
状態８　Ｈ（ＯＳがハングした状態）
状態９　Ｄ（ハードウェアの致命的な故障状態）

上記状態の内，状態１～７ではハートビートモジュールが動作しており，定期的に自己の状態をハートビートリンクを通して送信している。状態遷移は，相手系から送信される状態メッセージおよびタイムアウト事象によって行われる。表１に故障の種類毎に故障検出と処理を示す。
表１において，ハートビート回線故障の場合には必ず主系のハートビートモジュールでタイムアウトが発生するように，タイムアウト値を設定している。これによってハートビート回線が故障した時にシステムが動作不定となったり無用の系切り換えが発生することを防いでいる。

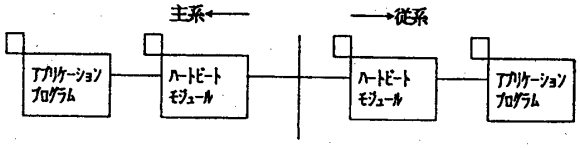

図3．相互監視機構

表１

故障の種類	故障の検出	処理
回復不能なハードウェアの故障	ハートビートタイムアウト	系切り換え，ＡＰＬ再起動
回復不能なソフトウェアの障害	ハートビートタイムアウト	系切り換え，ＡＰＬ再起動
回復可能なハードウェアの故障	エラー検出プログラム／ＡＰＬ	系切り換え，ＡＰＬ再起動
回復可能なソフトウェアの障害	ＡＰＬタイムアウト検出	ＡＰＬ再起動または系切り換え
ハートビート回線故障	ハートビートタイムアウト	従系電源ＯＦＦ／主系継続

APL:APpLication program

5．結び

比較的単純な構成ながら，高い耐故障性を持つ，２重化計算機システムを開発した。特に電源の３重化と，主要モジュールの活性保守機能の実現によって，理論的には非常に高いＭＴＴＦ値を実現した。またハートビートモジュールと呼ばれる機構により，相互監視による確実かつ迅速な故障検出と系の切り換えを実現した。

本システムはいわゆるホットスタンバイ方式を採用しているため，故障時の系の切り換えにシステムで約１分程度の時間がかかる。また系が切り換わった時のデータの一貫性と処理の一貫性の保証は，基本的にアプリケーションプログラムの責任である。今後は，系が切り換わった時のアプリケーションプログラムに対する透過性をより高めることが課題である。

参考文献

[1] Daniel P. Siewiorek, Robert S. Swary "Reliable Computer Systems" Design and Evaluation Second Edition. Digital Press. P840.