分散システムにおけるクラスタ型機能分割方式の開発

6T－2

真矢　譲[1]、源馬　英明[2]、木下　俊之[1]
[1](株)日立製作所システム開発研究所、[2]同オフィスシステム事業部

1．はじめに

近年、コンピュータシステムは大規模化、広域化が進み、ネットワークを用いた分散システム上で、オンライントランザクション処理(OLTP)を行うシステムが増えつつある。このOLTPの処理量は毎年増加しており、処理能力の向上と24時間365日連続処理が要求されている。

そこで、複数のサーバを高速なシステムバスにより接続し、各サーバ上の電文処理をいくつかのプロセスに分割し、プロセス単位にホットスタンバイ制御を行うクラスタ型機能分割方式を開発した。本発表では、提案方式の概要と、その可用性と処理能力を示す計算アベイラビリティの評価結果について報告する。

2．クラスタ型機能分割方式

（1）システム構成

提案方式のシステム構成を図1に示す。各サーバはプロセッサ、メモリ、バス制御装置、IOP、回線制御装置およびディスク制御装置からなり、これらを高速なシステムバスにより接続する。

電文処理は、通信プロセス、トランザクションプロセス(以下、Trプロセス)およびファイルプロセスに分割され、このプロセス単位にホットスタンバイ制御を行う。その際、通信プロセス1の現用プロセスはサーバ1に、待機プロセスはサーバ2に搭載し、一方、通信プロセス2の現用プロセスはサーバ2に、待機プロセスはサーバ1に搭載する。このように、各プロセスとも相互にバックアップをとる構成とする。

（2）処理内容

(a)チェックポイントデータ取得手順：現用プロセスは、待機プロセスと同時にサーバあるいは端末からの電文を受信する。そして現用プロセスはI/O要求直前にチェックポイントデータ(CD)を待機プロセスに転送し、一方待機プロセスはこれをメモリに格納する。

(b)引継ぎ処理：現用サーバは、一定周期毎に制御用パスを介してaliveメッセージを待機サーバに送信する。待機サーバはこれの途絶により障害を検出する。引き継ぎ後、新現用サーバは他のサーバに障害発生を通知した後、CDから障害現用サーバのレジスタ群を回復し、再開命令を実行する。これにより、新現用サーバは最新のI/O要求発行時点から処理を実行する。

(c)再2重化処理：旧現用サーバの修復が完了し、その通知がされると、新現用サーバは他のすべてのサーバに修復完了を通知し、2重系に復帰する。

3．評価

提案方式について、計算アベイラビリティ(状態確率とその状態の処理能力の積)を算出する。

（1）処理能力

各サーバの処理能力は、障害検出のためのaliveメッセージ処理、障害サーバの引継ぎ処理のためのチェックポイントデータの取得、およびプロセス間通信のオーバヘッドを除いたものである。各プロセスの処理能力はサーバの処理能力の和であり、一方、システム全体の処理能力は3プロセスのうち最低のものになる。

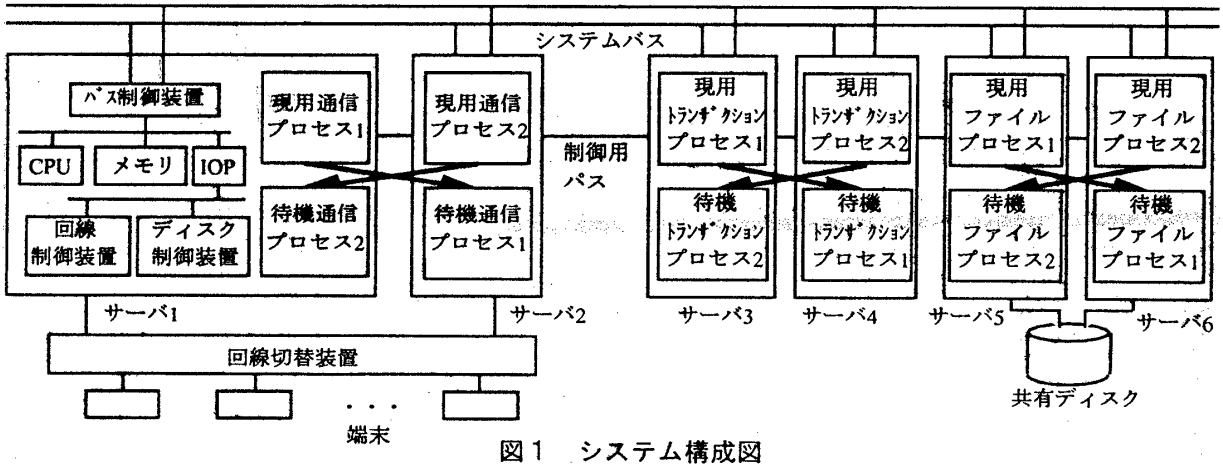

図1　システム構成図

A Proposal of Clustered Functional-Partitioning Method in Distributed Computer Systems
Yuzuru MAYA, Hideaki GEMMA, and Toshiyuki KINOSHITA
[1] Systems Development Laboratory,Hitachi,Ltd., [2]Office Systems Division,Hitachi,Ltd.

（2）状態確率

通信サーバ、Trサーバおよびファイルサーバはそれぞれ2台用意し、相互にバックアップする。通信サーバとファイルサーバはI/O処理を実行するため、処理能力が同一とし、次の4種類の状態を設ける。状態S_0は全サーバが正常な状態、状態S_1はTrサーバが1台障害の状態、状態S_2は通信サーバあるいはファイルサーバが1台障害の状態、状態S_3は同一のサーバが2台以上障害の状態であり、システム停止状態とみなす。ここで、故障率をλ、修復率をμとし、状態遷移を図2に示す。そして状態方程式(式(1))から、Si (i=0,1,2,3)の時刻(t)における状態確率p(i,t)を求める。

（3）評価結果

状態Siでの処理能力をTPS(i)とすると、計算アベイラビリティ(Ac(t))は式(2)となる。評価の前提条件を表Iに、提案方式と従来のN:1ホットスタンバイ方式の計算アベイラビリティの評価結果を図3と表2に示す。

従来のN:1方式では、バックアップ専用の予備サーバが必須であり、処理能力はその分低下する。また、1台のサーバでの障害発生時には引き継ぎ可能であるが、障害が2台以上で発生すると引き継ぎが不可能となりシステムは停止する。このように従来方式では、障害がシステムに与える影響が大きく、1,000日経過後の計算アベイラビリティは 30TPSに低下する。

一方提案方式では、バックアップ専用の予備サーバは不要であり、従来方式より処理能力が高い。また、1台あるいは異なるサーバで障害になった場合、処理能力は低下するが縮退運転が可能なためシステムに与える影響は小さい。同一プロセスのサーバが複数個障害になった場合にシステム停止となるが、この確率は従来方式よりかなり低い。この結果、提案方式の1,000日経過後の計算アベイラビリティは 75TPSとなり、従来方式の2.5倍に改善される。

4．おわりに

可用性および処理能力の向上を目標とし複数のサーバを高速なシステムバスにより接続し、電文処理をプロセスに分割するクラスタ型機能分割方式を提案し、従来方式より2.5倍程度向上できることを示した。

[参考文献]

1.当麻喜弘：コンピュータシステムの高信頼化技術入門；日本規格協会

$$\begin{bmatrix} \frac{d}{dt}p(0,t) \\ \frac{d}{dt}p(1,t) \\ \frac{d}{dt}p(2,t) \\ \frac{d}{dt}p(3,t) \end{bmatrix} = \begin{bmatrix} -6\lambda & \mu & \mu & 0 \\ 2\lambda & -(5\lambda+\mu) & 0 & 0 \\ 4\lambda & 4\lambda & -(3\lambda+\mu) & 0 \\ 0 & \lambda & 3\lambda & 0 \end{bmatrix} \begin{bmatrix} p(0,t) \\ p(1,t) \\ p(2,t) \\ p(3,t) \end{bmatrix} \quad (1)$$

$$Ac(t) = \sum_{i=0}^{3} \{TPS(i) \times p(i,t)\} \quad (2)$$

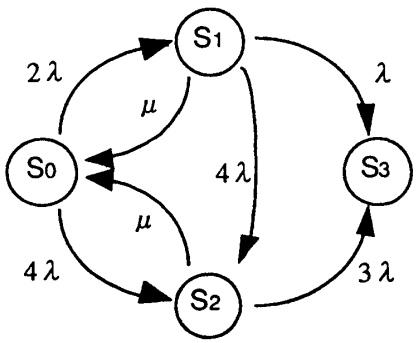

図2 状態遷移図

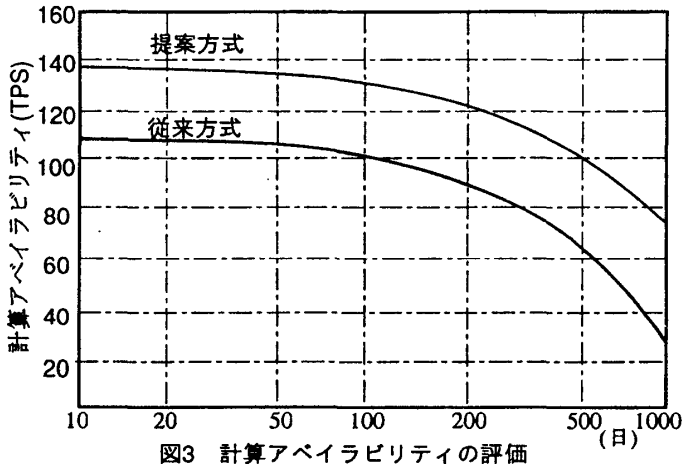

図3 計算アベイラビリティの評価

表1 前提条件

内容		数値(単位)
サーバの処理能力		10MIPS
プロセス間通信のステップ数		5kstep
各プロセスのダイナミックステップ数	通信サーバ	130kstep
	Trサーバ	135kstep
	ファイルサーバ	135kstep
1電文当たりのプロセス間通信の回数	通信サーバ	2回
	Trサーバ	3回
	ファイルサーバ	2回

表2 評価結果

方式	処理能力		可用性(障害の影響)		
		専用予備サーバ	単一障害	二重障害	
				異なるプロセスの場合	同一プロセスの場合
提案方式	高	無	縮退運転	縮退運転	システム停止
従来方式	低	有	通常運転	システム停止	