7L－2　分散オペレーティングシステムDM-2におけるサイトの動的な追加及び切り離し方式

河原 功志† 篠原 拓嗣† 藤川 賢治† 大久保 英嗣†† 津田 孝夫†
† 京都大学工学部情報工学科 †† 立命館大学理工学部情報学科

1 はじめに

分散OS [1] は従来のネットワークOSとは異なり、ユーザにネットワークの物理的配置を意識させないネットワーク透過性を実現している。ネットワーク透過性には位置透過性など様々な概念が含まれるが、規模透過性の実現をも分散OSには求められている。規模透過性とは分散システムの稼働中においても計算機の台数を増減することが可能で、かつ、システム構成の変更をユーザに意識させない性質である。本稿では規模透過性の実現する手法としてサイトの追加・切り離し方式を提案し、我々が現在開発中の分散OS DM-2に実現した。これらの機能によってDM-2は実行中のタスクを消滅や中断させずに、任意のサイトを停止して切り離したり追加して起動したりできるようになる。本稿ではサイトの動的な追加・切り離し機能を実現するために必要なローカルスタート・ローカルシャットダウン機能を提案する。

2 分散OS DM-2

我々はメモリ資源とタスクの位置透過性を実現するために分散仮想記憶と呼ばれる方式に基づく分散OS DM-2を開発してきた。

分散仮想記憶とは、ネットワーク上に構築された仮想的な単一の記憶空間のことであり、その実体は分散システム上の各サイトの主記憶及び2次記憶である。すべてのメモリ資源は分散仮想記憶空間上に写像される。仮想記憶上の連続したアドレスに割り当てるメモリ資源をメモリオブジェクトと呼び、その実体はページ単位で管理される。

DM-2はネットワークを介して接続された同一アーキテクチャをもつ計算機の上に実装されており、単一の仮想記憶空間を共有している計算機の集合をDM-2クラスタと呼び、クラスタに属する各計算機をDM-2サイトと呼ぶ。

DM-2ではプログラムの実行をタスク・スレッドモデル [2] に基づいて管理しており、一つのタスクを複数のスレッドから構成できる。そしてスレッドは位置透過にシステム内で実行され、タスク内の各スレッドをネットワーク上の複数のサイトに分散させた並列処理が可能である。

また、DM-2ではOSの諸機能をマイクロカーネル部分とタスク群に分けて構成している。これらのタスクにはスケジューラ、メモリオブジェクトマネージャ、タスクマネージャ及びスレッド分配機構であるスレッドディストリビュータがある。これら以外のタスクをユーザタスクと呼ぶ。カーネルは分散カーネルとしてサイト間通信と分散仮想記憶の機能を提供する。また本研究ではサイトの追加機能を実現するためにサイト間接続機構をカーネル内部に実現し、クラスタの停止やサイトの切り離し機能を提供するタスクであるシャットダウンサーバを実現した。

3 DM-2におけるサイトの追加・切り離し機能

分散仮想記憶に基づいたDM-2は柔軟な運用形態をユーザに提供することができる。すなわち、ひとたび起動したDM-2システムは十分な（2次記憶も含めた）記憶資源と計算資源がある限り、サイトの数を増減しながらも処理を中断することなく継続的に動作することが可能である。

DM-2は規模透過性を有することによって、例えば分散システムを使用中にサイトの一部だけを保守のために電源を落したい、あるいはシステムの処理能力が不足しているのでさらにサイトを追加して処理能力の向上を図りたい、といった要求に応えることができる。

DM-2では、すでに起動されているクラスタに新たにサイトを追加する際にそのサイトで行なわれる処理をローカルスタートと呼ぶ。そしてクラスタからサイトを切り離す際に行われる処理をローカルシャットダウンと呼ぶ。なお本稿では、ローカルシャットダウンの時

Method of Run-time Cluster Reconfiguration in the Distributed Operating System DM-2
Kohji Kawahara†, Takuji Shinohara†, Kenji Fujikawa†, Eiji Okubo†† and Takao Tsuda†
† Dept. of Information Science, Kyoto University
†† Dept. of Computer Science, Ritsumeikan University

には十分な2次記憶がネットワーク上に存在し、ローカルスタートを行うサイトの2次記憶には分散仮想記憶上のメモリオブジェクトは存在しないと仮定する。

4 ローカルスタート

ローカルスタートはサイトの追加に関わる処理であり、次のような手順で行われる。

まず、追加サイトがDM-2クラスタに物理的に接続されてからカーネルが起動し、クラスタに加わりたい旨をすべてのサイトにブロードキャストする。クラスタ内のサイトは追加サイトに通信の接続を試み、追加サイトに選ばれたそのうちの一つのサイトが追加サイトとの通信を確立する。このサイトを接続サイトと呼ぶ。接続サイトは追加サイトに、分散仮想記憶にアクセスするために必要な情報と、クラスタに属するすべてのサイトの情報を伝える。そして接続サイトから追加サイトにシステムタスクのスレッドが移送される。次に追加サイトとすべてのサイトとの間に通信が確立し、追加サイトがDM-2サイトとして稼働する。あとはスレッドディストリビュータによってユーザタスクのスレッドが移送されてくる。また必要なメモリオブジェクトがローカルにない場合はページフォルトが発生することよって検出され、分散仮想記憶管理部によって必要なページが自動的にリモートサイトから取得される。

5 ローカルシャットダウン

切り離されるサイトではシャットダウンサーバがローカルシャットダウンモードに移行する。シャットダウンサーバは自分以外のユーザスレッドを他のサイトに移送するようにスレッドディストリビュータに依頼する。次に他サイトに複製が存在しないメモリオブジェクトを他サイトに移送するように、メモリオブジェクトマネージャに依頼する。必要なメモリオブジェクトとスレッドの追い出しが終るとカーネルに制御が移り、カーネルがすべてのシステムタスクを停止する。そしてクラスタとの通信を切断した後カーネルが終了する。

6 おわりに

本稿で提案したサイトの動的な追加及び切り離し方式は、システムの可用性を向上させることを目標に設計された。すなわち、計算機の稼働中にシステムの規模を自在に変更することができれば、その時々で利用できる計算機資源を最大限使いながら計算の続行が可能となる。分散OS DM-2における規模透過性は、ユーザに計算機が追加されたことを特に意識させることなく、より高い処理能力を提供することができる、という利点を有する。

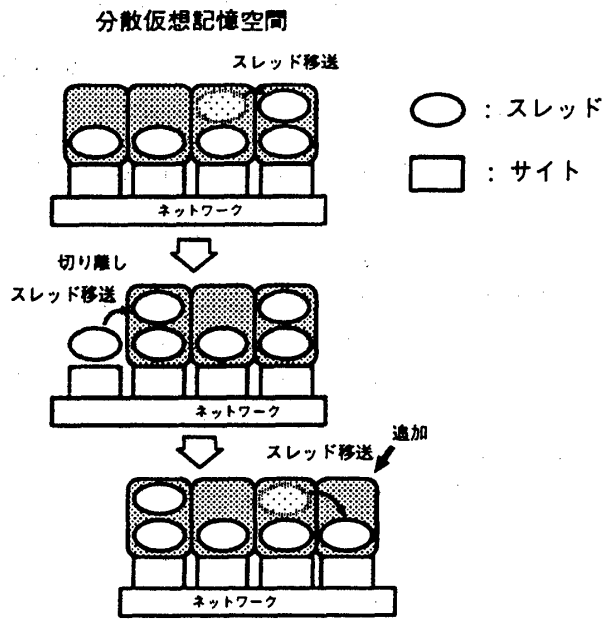

図 1: DM-2におけるサイトの追加及び切り離し

また、分散システムの障害対策機能においても、ある利点を有する。分散システムに発生することが考えられる主な障害には、ネットワークの異常や、サイトの2次記憶の故障などが挙げられる。DM-2では、特に予測可能な2次記憶の障害については、その発生が予測された時点で問題のあるサイトを動的に切り離すことで障害を回避することができる。また、そのサイトが異状から復帰した場合もそのサイトを動的に追加することによって切り離しによって低下した性能を回復することができる。すなわち、障害の影響を回避しつつ計算を継続することが可能である。したがって、本研究が提案するサイトの動的な追加及び切り離し方式は障害対策機能にも有効であると考えられる。

参考文献

[1] 前川 守, 所 真理雄, 清水 謙太郎 : 分散オペレーティングシステム UNIX の次に来るもの (共立出版, 1991).

[2] 藤川 賢治, 篠原 拓嗣, 大久保 英嗣, 津田 孝夫 : 分散仮想記憶に基づくオペレーティングシステム DM-1 におけるタスク・スレッドモデル, 第48回情報処理学会全国大会論文集, 3F-07 (1994).