Vol. 36 No. 8　　情　報　処　理　　Aug. 1995

解　説

日本におけるオペレーティングシステム研究の動向

1. 2　分散 OS Galaxy †

清　水　謙多郎††　　前　川　　守†††　　芦　原　　評††

1. はじめに

Galaxy は，真にネットワークの存在を意識させない情報・資源の共有，信頼性の向上，性能の向上を追求した分散オペレーティングシステムである[1)~3)]．これらは，分散システムの共通の設計目標であり，より良い分散処理環境を実現する上で不可欠のものである．もちろん，これらすべてを最初から完全に達成することは不可能に近い．Galaxy では既存のシステムの限界を克服し，上記の目標を可能な限り達成することを目指して開発を行っている．さらに，大規模分散システムに対応できる規模適応性（スケーラビリティ）の実現も Galaxy の重要な設計目標である．大域的な情報を分散管理し，ブロードキャスト・プロトコルや大域的ロック，大域的なタイムスタンプによる順序付けを使用しないなどの工夫を行っている．

2. オブジェクト・ネーミング

2.1 基本方式

上記の目的を達成していこうとする上で，とくに重要なのがオブジェクトのネーミングである．Galaxy では，**図-1** に示すような 3 階層のネーミング方式を用いている[4),5)]．

外部名 (external name) はユーザがオブジェクトに与える文字列の名前である．外部名はオブジェクトの存在位置とは独立であり，大域的に一意に定義される．オブジェクトを移動しても名前を変更する必要はない．外部名は，階層構造を持つ名前として定義され，システムで 1 つの大域的なディレクトリ階層に登録される．ディレクトリの各エントリには，オブジェクトまたは（下位の）ディレクトリの外部名と ID の対が記載されている．

Galaxy の各オブジェクトは，システムで一意に定まる**オブジェクト識別子 (ID)** を持つ．ID は，オブジェクト生成時にシステムによって割り当てられ，ユーザのプログラムやシステムがオブジェクトを識別するのに用いられる．ID は，**ID マネジャ**により，オブジェクトの存在位置（オブジェクトが存在するノードのアドレス）に変換される．この操作を位置付け (locating) と呼ぶ．外部名から ID への変換とは独立に，ID だけで大域的なオブジェクトの位置付けが可能である．

2.2 位置付け機構

ID は，(a)オブジェクトが生成されたノードのアドレスと，(b)そのノードにおけるオブジェクト生成時のタイムスタンプから構成され，システム全体での一意性を実現している．ただし，オブジェクトの移送に対応するため，(a)の情報をオブジェクトの位置付けには使用しない．位置付けには，ID Table と呼ぶ大域的なデータベースを使用する．各ノードには，少なくとも次の条件を満たす ID の ID Table エントリ (IDTE) が登録される．

（1） そのノードに存在するディレクトリに含まれる ID

（2） そのノードで実行中のプロセスが用いる ID

これらの IDTE は，必要に応じてレプリケートされ，ノード間で重複して保持することができる．このような機構により，Galaxy では，ID を

† The Galaxy Distributed Operating System by Kentaro SHIMIZU (Department of Computer Science, The University of Electro-Communications), Mamoru MAEKAWA (Department of Information Systems Science, Graduate School of Information Systems, The University of Electro-Communications) and Hyo ASHIHARA (Department of Computer Science, The University of Electro-Communications).

†† 電気通信大学情報工学科
††† 電気通信大学情報システム学研究科情報システム設計学専攻

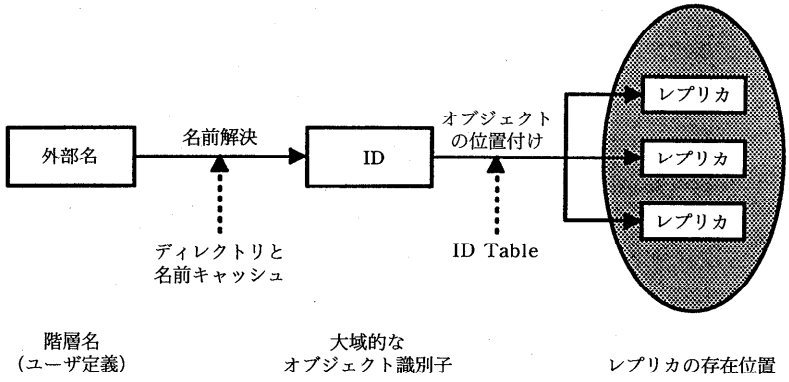


図-1 3階層ネーミング方式

もとに，アクセスするノードでオブジェクトの位置付けが可能となる．従来のシステムの位置付け機構としては，(a)特定のノードが集中して位置情報を管理する方式，(b)全ノードが全位置情報を管理する方式，(c)ブロードキャストで問い合わせる方式，(d)生成ノードを起点として移動先の位置情報をチェインで保持する方式などが用いられてきたが，どれも性能，信頼性，規模適応性のいずれかの面で支障があった．

各IDTEには，ID，アクセス制御情報，オブジェクトのレプリカの存在するノードのリスト(レプリカリスト)，IDTE自身のレプリカの存在するノードのリスト（コピーリスト）を保持している．レプリカリストには，そのオブジェクトの全レプリカの位置が記載されているので，変更時には，そこに記載されている全ノードに直接変更要求を出すことができ，参照時には，その中から適当なレプリカを選択することができる（選択方針はサーバによって与えられる)．また，ノードやネットワークの障害によりアクセス不可能なレプリカが生じても，柔軟に対応することができる．コピーリストは，IDTEに対する変更をそのレプリカが存在するすべてのノードに直接伝えるためのものである．コピーリストもレプリカリストと同様，可用性・信頼性の向上に役立つ．

IDTEの変更で最も頻繁に行われるのが，レプリカリストとコピーリストの変更である．これらの情報は，独立した項目から構成されるので，その変更は項目の追加と削除に限られる．この場合，同一項目に対する変更でない限り並行して実行することが可能である[6]．これらの操作は，発行後ただちに全ノードで実行される必要がなく，IDTEレプリカ間で一時的な不一致を許す．すなわち，IDTEの変更は，局所ノードのIDTEを変更するだけで操作を終了し，他ノードのIDTEの変更はバックグラウンドで行うようにしている．IDTEの大域的なロックや大域的なタイムスタンプによる順序付けは必要としない．また，IDTEレプリカの追加がシステム内で同時に行われたとき，これらの新しいレプリカに，他方のレプリカの追加が最終的に反映されるよう，レプリカを生成したノードが責任を持って，その後のIDTE変更要求を生成したレプリカに伝播させる．

変更要求は，全ノードに伝わるまでバッファリングされるため，同一項目の重複した追加および削除はそれぞれ1つの操作として実行し，同一項目の追加と削除の対は（その順序によらず）相殺することができる．

2.3 ディレクトリの管理

ディレクトリは，他のオブジェクトと同様，IDにより位置付けが行われる．外部名からIDへの変換において，ルートディレクトリまたはカレントディレクトリを起点として，パス名中に現れるディレクトリが順次探索される．これらのディレクトリは，システム内に分散していてかまわない．また，ディレクトリは，他のオブジェクトと同様，レプリカを作成することができる．一般に，ルート近くのディレクトリは，葉の近くのディレクトリに比べ，多くのノードから頻繁に読み出される一方，変更はむしろ少ない．そのため，レプリカを多数生成することが効率向上に役立つ．

大域的なディレクトリ階層では，名前空間全体

にわたってネーミング機構の信頼性を高めることは非常に困難である．一般に，ユーザが参照するオブジェクトは全体のごく一部であり，ユーザが作業を行うノードも限られている．そこで，限られた範囲のノード，オブジェクトに対して信頼性に関するパラメータをユーザレベルで設定する機構を提供している[7]．

（1） subpath reliability：他ノードに問い合わせることなく局所ノードで解決できるパス名の長さ s

（2） m-stage reliability：パス名解決の経路の中で必要となるノード探索のステップ数 m

（3） k-path reliability：パス名解決の経路の中で任意の2ノード間でとり得る経路の数 k

パス名の長さ（パス名に含まれる要素数の総数）を n とすると，$0 \leq s \leq n, 0 \leq m \leq n, 1 \leq k$ で定義される．

3. プロセス管理

3.1 エグゼキュータ/ドメイン・モデル

Galaxy では，計算モデルとしてエグゼキュータ/ドメイン・モデルを採用している．**エグゼキュータ**とは能動的な実行主体であり，**ドメイン**とはエグゼキュータによってアクセスされる資源の集合である．エグゼキュータ自身を記述する記述子（オブジェクト参照，事象の発生を伝える範囲，その他の実行状態を含む）もドメインの1つである．従来のプロセスは，エグゼキュータとそれがアクセスするドメインを合わせたものとして定義できるが，ある程度完結したエグゼキュータとドメインの組合せをプロセスとしてユーザが明示的に指定することができる．プロセスは，移送，レプリケーションなどの単位としても利用される．エグゼキュータは，互いにドメインの任意の部分を共有することができるため，任意の重さを持つことのできる，いわゆる**可変ウェイトプロセス**が実現される．

3.2 マイクロプロセス

エグゼキュータのスケジューリングは，ユーザの指定により，カーネルとユーザレベルの協調によって行うことができる．この場合，ユーザから見える実行主体を**マイクロプロセス**という．ユーザ・アドレス空間内には，マイクロプロセスの基本的な実行コンテキスト（レジスタ退避域など）およびスケジューラ（ユーザ・スケジューラと呼ぶ）が置かれる．ユーザ・スケジューラは，タイムスライス，封鎖型システムコールの発行，ページフォールトの発生をきっかけとして，カーネルからスケジュール要求を受け，次に実行すべきマイクロプロセスを選択する．マイクロプロセスの機構により，ユーザのアプリケーションに合ったスケジューリング方針を適用し，マイクロプロセス間の効率的な切替え，同期・通信を実現することができる．

3.3 プロセス移送

プロセス移送の機構上の問題点として最も重要なものは，移送にともなうオーバヘッドである．中でも特にメモリ領域の転送の問題は重要であり，この点については，すでにページ単位で転送する方式が知られている．従来，提案されてきたオーバヘッド軽減のための方策は，(1)制御を移動する前にページを前もって転送する先行コピーと，(2)制御を移動した後ページフォールト発生時に転送する要求コピーの2つに大別される．(1)は，複数のページを一括して転送することによるページ転送のオーバヘッドの軽減が見込めるが，移送後にアクセスする可能性のないページも転送しなければならず，またすでに転送したページに書込みが行われたときは，ページの再転送が必要であるという問題がある．一方，(2)の方式は，無駄なページ転送は不要であるが，ページ転送のオーバヘッドが増大する，負荷の移行が緩慢であるという問題がある．Galaxy では，それらの中間の方式として，ワーキングセット内のページは制御移動時に一括して転送し，さらにその後でアクセスされる可能性の高いページは先行コピー（ただし，1回だけメモリに存在しているページを走査する），上記以外のページは要求コピーを適用するという方式を用いている[8]．

プロセス移送の目的は，平均実行時間の短縮，特定のプロセスの実行時間の短縮，スループットの向上，可用性の向上など様々であり，それぞれの目的に応じた方針が用いられる．とくに平均実行時間の短縮には，負荷の均衡化が必要である．具体的な方針については，シミュレーションおよび確率過程モデルの数値解析により[9)~11)]，次の方針が有効であるとの結果を得た．

（1） 情報収集方針：ノード間で定期的に情報

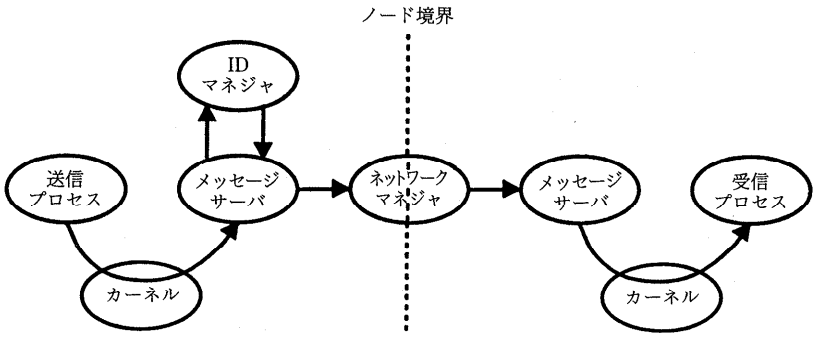


図-2 ネットワーク透過なプロセス間通信機構

を交換して重負荷/軽負荷ノードの候補を絞り，それらのノードに対して入札を適用する．

（2） 移送起動方針：送り手主導，受け手主導の双方を適用した対称主導を用いる．システムの平均負荷が重い場合，軽い場合ともに有効である．

（3） ノード選択方針：移送先は負荷最小のノード，移送元は負荷最大のノードを選択する．

（4） プロセス選択方針：優先度が低下したプロセスで，移送先で高い優先度で実行されるもの．ただし，この方針は性能にそれほど大きな影響を与えない．移送にともなうデータ転送量も考慮に入れるべきである．

4. 通　　信

Galaxyでは，(1)局所ノードに最適化された効率的なプロセス間通信と，(2)大域的なプロセス間通信の両方を提供している[1),3)]．(1)については，例外処理のためのメッセージ付きソフトウェア割込み，プロセス間の最も基本的な通信手段である固定長メッセージ受渡し，アドレスマッピングの切替えによる仮想ページ転送がある．(2)については，次のような機構が実現されている．

（a） メッセージ受渡しをノード間通信に拡張したもの

局所プロセスに対するメッセージ受渡しはカーネル内で処理され，遠隔プロセスに対するメッセージ受渡しはユーザレベルの**メッセージサーバ**によって処理される．メッセージサーバは，IDマネジャに相手プロセスの位置を問い合わせ，得られたノードとの間でメッセージ転送を行う．**図-2**に，ネットワークを介した送信操作の動作例を示す．

（b） グループ通信

Galaxyでは，オブジェクトをまとめて操作・管理するための機構として**オブジェクトグループ**[12)]の機構を提供している．1つまたは複数のプロジェクトをオブジェクトグループとしてまとめることにより，これらのプロセスに対して多様なセマンティクスのマルチキャスト通信（k-reliable通信，バッファリング，タイムアウトなど）を利用することができる．

（c） 分散共有メモリによるデータ共有

Galaxyの分散共有メモリの特徴は，アドレス空間の統一した操作と柔軟な共有セマンティクスにある．操作については，IDとID内アドレスの対によって形成されるネットワーク透過な大域アドレス空間上で，ページ単位でデータのコピーを行うcopy，ページ単位で領域を同一視するunifyを提供する．共有セマンティクスについては，読み出すデータが反映している変更の時刻，一貫性を持たせるノード群などを明示させる弱い一貫性を実現している[13),14)]．

5. データレプリケーション

Galaxyは，ファイルオブジェクトの部分的なレプリカを自動的に作成する**ダイナミックレプリカ**[15)]と，ファイル全体のレプリカを（多くの場合ユーザの指定により）作成する従来型の**スタティックレプリカ**の2つをサポートする．ダイナミックレプリカは，最初は空でアクセスがなされるたびにページ単位で実体が生成される．ダイナミックレプリカは，遠隔ファイルに対するアクセス要求に応じて作成されるが，作成するかどうかの決定および作成するノードの選択は，ページごとの

読出し/書込み比，クライアントとスタティックレプリカとの距離，同一ファイルの共有の度合いによる．レプリカ間の一貫性プロトコルについては，スタティックレプリカでは更新プロトコル，ダイナミックレプリカでは無効化プロトコルが用いられている．スタティックレプリカは，そこからコピーが作成される複数のダイナミックレプリカに対して一次サイトとして働く．

ダイナミックレプリカは，データの配置に関して大域的な最適解を与えるものではない．生成にあたっては，限られたノード間で局所的に決定がなされる．しかし，現実の分散システムにおいて，多くのファイルは，比較的限られたノードからアクセスされないし，あるノードでアクセスするファイルも限られている．このような観点から，上に述べたようなダイナミックレプリカの設計を行った．

6. 現状と今後

Galaxyの主要部分はすでにIBM RTワークステーション上に実装し，実際に動作している．システムのほとんどは数百行のアセンブリ言語の部分を除いて，C言語で記述している．現在，システムの完成を目指してさらに開発を進めているという状況である．今後の課題として，異種分散システムへの対応，セキュリティの強化，リアルタイム応用の支援などがあげられる．

参 考 文 献

1) Sinha, P. K., Maekawa, M., Shimizu, K., Jia, X., Ashihara, H., Utsunomiya, N., Park, K. S. and Nakano, H.: The Galaxy Distributed Operating System, IEEE Computer, Vol. 24, No. 8, pp. 34-41 (1991).
2) Maekawa, M., Shimizu, K., Jia, X., Sinha, P. K., Ashihara, H., Utsunomiya, N., Nakano, H. and Yamaguchi, S.: The Galaxy Distributed Operating System, in Distributed Environments, Y. Ohno (ed.), Springer-Verlag, pp. 259-289 (1991).
3) Sinha, P. K., Mackawa, M., Shimizu, K., Jia, X., Ashihara, H., Utsunomiya, N., Park, K. S. and Nakano, H.: The Architectural Overview of the Galaxy Distributed Operating System, in Distributed Computing Systems, T. L. Casavant and M. Singhal (ed.), IEEE Computer Society Press, pp. 327-345 (1994).
4) 清水謙多郎, 前川 守, 芦原 評: 分散オペレーティング・システムにおける名前管理, コンピュータソフトウェア, Vol. 6, No. 3, pp. 19-34 (1989).
5) Sinha, P. K., Shimizu, K., Utsunomiya, N., Nakano, H. and Maekawa, M.: Network-Transparent Object Naming and Locating, J. Info. Process., Vol. 14, No. 3, pp. 310-324 (1991).
6) Jia, X., Nakano, H., Shimizu, K. and Maekawa, M.: Highly Concurrent Directory Management in the Galaxy Distributed System, Proc. 10th IEEE Int. Conf. Distributed Computing Systems, pp. 416-423 (1990).
7) Sinha, P. K., Maekawa, M. and Shimizu, K.: Improving the Reliability of Name Resolution Me chanism in Distributed Operating Systems, Proc. 12th IEEE Int. Conf. Distributed Computing Systems, pp. 589-596 (1992).
8) Sinha, P. K., Park, K. S., Jia, X., Shimizu, K. and Maekawa, M.: Process Migration in the Galaxy Distributed Operating System, Proc. 5th IEEE Int. Parallel Processing Symp., pp. 611-618 (1991).
9) 朴 圭成, 芦原 評, 清水謙多郎, 前川 守: 分散オペレーティング・システムにおけるプロセス移送の方式, 情報処理学会論文誌, Vol. 31, No. 7, pp. 1080-1090 (1990).
10) Ashihara, H., Mizuguchi, N. and Maekawa, M.: Reduction of Information Exchange for Adaptive Load Sharing in Distributed Systems, Proc. 4th ISMM Int. Conf. Parallel and Distributed Computing and Systems, pp. 92-96 (1991).
11) Ashihara, H., Mizuguchi, N. and Maekawa, M.: All-range Policies for Adaptive Load Sharing in Distributed Systems, Proc. Int. Workshop Parallel Computing, pp. 34-41 (1991).
12) Shimizu, K., Maekawa, M., and Hamano, J.: Hierarchical Object Groups in Distributed Operating Systems, Proc. 8th IEEE Int. Conf. Distributed Computing Systems, pp. 18-24 (1988).
13) Sinha, P. K., Ashihara, H., Shimizu, K. and Mackawa, M.: Flexible Address Space Sharing Mechanisms in the Galaxy Distributed Operating System, Proc. 10th IEEE Int. Conf. Computers and Communications, pp. 212-218 (1991).
14) Sinha, P. K., Ashihara, H., Shimizu, K. and Maekawa, M.: Flexible User-Definable Memory Coherence Scheme in Distributed Shared Memory of Galaxy, Lecture Notes in Computer Science, Vol. 487, Springer-Verlag, pp. 52-61 (1991).
15) Ashihara, H., Shimizu, K., Maekawa, M. and Hamano, J.: File Access Improvements in Distributed Systems By On-Demand Replication of Files, Technical Report 87-26, Depart-

ment of Information Science, Univ. of Tokyo (1987).

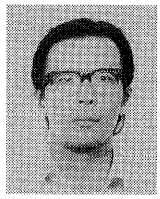
(平成7年5月24日受付)

清水謙多郎（正会員）

1957年生．1980年東京大学理学部情報科学科卒業．1985年同大学院理学系研究科博士課程修了．理学博士．東京大学理学部助手などを経て，1991年より電気通信大学助教授．オペレーティングシステム，並列/分散処理の研究に従事．著書「オペレーティングシステム」（岩波書店），「分散オペレーティングシステム」（共立出版，共著）など．本会論文誌編集委員．ACM，IEEE，電子情報通信学会，ソフトウェア科学会各会員．

前川　守（正会員）

昭和17年生．昭和40年京都大学工学部卒業．東芝，東京大学理学部助教授等を経て，現在電気通信大学大学院情報システム学研究科教授．Ph.D. 分散システム，マルチメディア，情報システム等の研究に従事．ACM，IEEE各会員．



芦原　評（正会員）

1964年生．1987年東京大学理学部情報科学科卒業．1992年同大学院理学系研究科博士課程修了．理学博士．同年より電気通信大学助手．オペレーティングシステム，並列/分散処理の研究に従事．ACM会員．