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

4.3 高信頼分散システム構築支援OS：知的分散™OS †

関　俊文††　長谷川　哲夫††

1. はじめに

技術革新にともなってユーザニーズが多様化し，拡張性・信頼性・保守性・適応性に優れたシステムが求められている．このような要求に応える計算機アーキテクチャとして宣言型の分散システムが提案されている[1]．そこでは，各々固有の能力が宣言された要素の単なる集合としてシステムが定義され，これら要素群が互いに自律的に協力・協調することによって状況に適応する．したがって集中管理機構が排され，各要素は集中管理機構による制約を受けることなく，状況に応じて自身の役割を決定し，その最大の能力を常に発揮することができる．このような宣言型システムが，上記要求を完全に実現するためには，個々の要素が関連要素の存在や位置，多重度に依存せず独立に動作する性質が重要である．本稿では，宣言型システムの性質を保存した高信頼分散システムの実現を支援するため，各要素を自律的に動作可能なオブジェクトとして表現し，それらを放送通信を用いて動的に結合できるようにした知的分散OS[2]について紹介する．

システムの高信頼化にあたっては，どのような障害にどのように対処するかによって様々な方法が存在するが，知的分散OSはハードウェア故障に対してソフトウェアの多重化によって高信頼化を支援するものである．

宣言型分散システムの性質を維持しながら高信頼化するために，次に示す性質が要求される．

1) アプリケーションプログラムは，多重化要素を並列多重で動作させるか待機冗長で動作させるか，といった高信頼化方式に対して透過でなければならない．

2) 要素は信頼度変更や高信頼化方式変更にオンラインで対応できるよう，位置・多重度・実行モード（稼動状態，待機状態）から透過でなければならない．

3) 多重化要素内の要素間の内部状態の一貫性が絶えず保証されていなければならない．

ソフトウェアモジュールの多重化による高信頼システム実現の関連研究としては，ISIS[5]やDelta-4[6]，Auragen[7]などが存在するが，これら研究との最大の違いはシステムレベルの管理機構存在の有無である．すなわち，これらの方法にはシステムレベルでソフトウェアモジュールの多重度や高信頼化方式を集中管理する機構が存在するが，知的分散OSによる方法には存在しない．本稿では，知的分散OSにおける上記性質実現のための機構を中心に述べる．知的分散OS全般については参考文献を参照願いたい．

2. 知的分散OSの基本構造

高信頼な宣言型分散システムの実現を支援するため，OS機能も集中管理機構を保持せず要素群の協力・協調によって実現されなければならない．従来のOSでも，OS機能を核とプロセスに分離し，核の機能をできるだけ小さく抑える試みがされている．しかし，そこではOS機能の各々を専用のプロセスに割り当てる機能分散の形がとられており，すべての計算機からの要求が特定のプロセスに集中してしまい，処理のボトルネックが生じる恐れがある．

さらに，動的に役割を決定するための要素の動的結合を支援する通信機構が必要となる．要素間の接続のみならず，要素の位置や多重度も動的に変更可能でなければならないため，これらの変更に柔軟に対応する通信機構が必要となる．

† IDPS Operating System for Constructing Fault-Tolerant Systems by Toshibumi SEKI and Tetsuo HASEGAWA (TOSHIBA Corporation).

†† (株) 東芝研究開発センター

このような要求を満たすため，知的分散OSは，オブジェクトモデルと放送通信を採用している．すなわち，OS機能を含むすべての各処理要素はオブジェクトモデルにおけるオブジェクトとして記述し，それら複数計算機に分散したオブジェクト群を放送通信で動的に結合する．

知的分散OSは，**図-1**に示すように各計算機に存在する知的分散OS核と各オブジェクトに分散されたグローバルな管理知識からなる共通知識部の結合からなり，OS自体も集中管理部を持たない．各計算機上の知的分散OS核は，個々の計算機のCPUやメモリの管理，あるいはオブジェクト間で交換されるメッセージの送受信処理などローカルな資源管理を行う．各オブジェクトへの役割割り当てなどオブジェクトのグローバルな動作管理は，関連するオブジェクト群が持つ共通知識の動的結合によって実現される．本OSの最大の特長は，このようにグローバルな管理機構がアプリケーション・プログラムと同一レベルの個々のオブジェクトに分散していることである．

このため，計算機間での負荷分散やオブジェクトの名前管理，オブジェクト間で並列に進行する処理の同時実行制御，多重化されたオブジェクト間での状態の一貫性保証など，システム全体にわたる動作のすべての管理は，特定のオブジェクトによって実現されるのではなく，オブジェクト群の共通知識部の結合によって実現される．また各オブジェクトの固有知識は，個々のオブジェクトに固有の能力に関する知識である．共通知識と固有知識は，それぞれオブジェクト内の機能単位であるメソッドとして実現される．

また放送通信は，オブジェクトの位置と多重度からの独立性を実現する．すなわち放送通信を用いることにより，結合相手の論理名を指定するだけで，各オブジェクトはどの計算機に結合相手のオブジェクト群が存在するかを知ることなく動的に結合できる．また高信頼化のために多重処理が必要な場合は，同時に複数の同名オブジェクト群と結合する．このことによって，オブジェクトがシステム高信頼化のために多重化された場合でも，同時にすべての同名オブジェクト群と結合するので，要求の変化や故障によって多重度が変更されたり移動したとしても，そのような変化を意識することなく動作を継続できる．

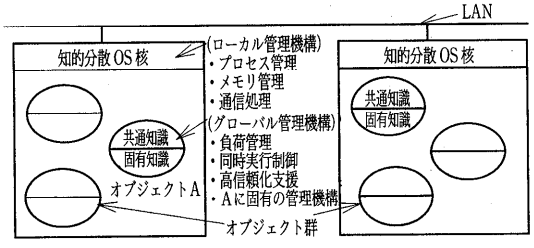


図-1 知的分散OSの構成

ここで，知的分散OSでいうオブジェクトは資源割付けの基本単位である．メソッドは外部からアクセス可能な手続きであり，関連オブジェクトのメソッドからメッセージ通信によって起動される単位である．メソッドの駆動形態としては，手続き呼出しに相当するメソッド駆動型，データ名を指定したデータ駆動型，起動ルールが満たされたメソッドを起動するルール駆動型を持つ．

さらに知的分散OSは基本機能として，オブジェクトの動的な生成，移動，複写機能を持つ．これらはシステムの処理性，拡張性，保守性，信頼性向上のために必要不可欠な機能となっている．

3. 知的分散OSのシステム高信頼化支援機構

知的分散システムでは，オブジェクトを多重化し，それらをすべて稼動系として並列に動作させたり，いくつかを待機状態にして待機冗長方式で動作させたりすることによりシステムの高信頼化を図る．そのため，知的分散OSがオブジェクトの位置や多重度のみならず，稼動/待機といったオブジェクトの実行モードからの独立性を実現することで，高信頼化による拡張性の低下を防止している．

オブジェクトの位置・多重度独立性を実現するために放送通信を用いることはすでに述べたが，**3.1**ではLAN（Local Area Network）に接続された全計算機が同一順序で同一メッセージを受信する全順序性を保証した高信頼放送通信機構について述べる．この通信機構により，多重化オブジェクト間の内部状態の一貫性を保証することが容易になる．**3.2**では多重化オブジェクトを並列多重方式で動作させたり，待機冗長方式で動作させるときの多重化オブジェクト間の内部状態の一貫性を保証するための機構について示す．

3.1 高信頼放送通信機構

知的分散OSでは全順序性を保証する放送通信プロトコルとしてフェイル・ストップ放送通信機構[3]を持つ．従来の提案方式は，メッセージの順序管理を行う集中管理機構が存在したり[8]，宛先数を意識した方法[9]となっており，宣言型システムの性質に合い入れない．提案プロトコルは，各計算機の通信機構が集中管理部を介さず分散的に故障計算機を検出するため，負荷が特定の計算機に集中することはなく，しかも同期機構が不要なため受信確認を待ち合わせるなどのオーバヘッドがない．以下に本プロトコルの概要を述べる．

各計算機の通信機構は，**図-2**に示すようにLANを流れるメッセージを傍受し，それまでに送受されたメッセージの総量に対応する値として通信累積量(Accumulated Message Number，以下AMNと記す)を計算・記憶する．そしてメッセージ送信時には，自通信機構が記憶しているAMNをメッセージに付加する．このとき正常に動作している計算機群は同一のメッセージを受信するので，送受信が正しく行われていれば，受信メッセージ中のAMNと受信計算機内のAMNは一致する．そこで2つのAMNが異なる場合は，受信計算機は送信元にて通信フォールトが発生したと判断して受信メッセージを無視すると共に，送信元に通信エラー検出通知(Site-Fail)を送る．さらに各通信機構は，一定時間以内にあらかじめ定められた故障確信数以上の計算機からSite-Failを受信すると自身が通信フォールトを発生したと判断して，その動作を止めるかあるいは再送要求を出す．故障確信数はシステムに要求される信頼度に応じて決めることができる．また，通信エラー検出通知は，通信フォールトを起こした計算機が送信をしなくてもいずれかの計算機の送信動作によって即検出される．このため通信障害による処理中断時間が短い．

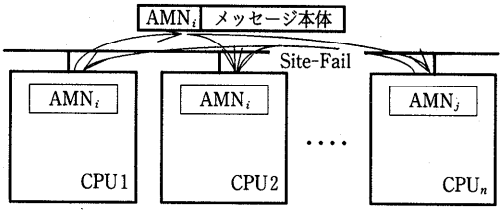


図-2 フェイル・ストップ放送通信

システム中の計算機群が同一順序でメッセージを受け取る機構は，通信機構がメッセージをLANに正しく送出したのを確認してから，自計算機上のオブジェクトに転送することにより実現できる．

本機構により，多重化オブジェクトは同一メッセージを同一順序で受信可能となる．

3.2 高信頼要素制御機構

知的分散OSに基づくシステムでは，オブジェクトの多重度や高信頼化方式(並列多重方式か待機冗長方式かの選択)は，オブジェクトごとに要求される信頼度や，故障時に許容される切替え時間に応じて選択できる．すなわち，機能の無停止性が要求されるオブジェクトには並列多重方式を，故障時にある程度の処理の中断が許されハードウェア資源量に強い制限がある場合には待機冗長方式を用いる．

従来はシステム全体として1つの方式に固定されてしまい，高信頼化方式のそれぞれの利点を十分に生かすことができなかったが，知的分散OSではオブジェクト個々の処理形態を考慮してそれらを使い分けることができる．

すなわち知的分散OSは，アプリケーションプログラムに並列多重方式か待機冗長方式かを意識させず，どこに配置しても何重に多重化しても，それら多重化オブジェクト間の一貫性保証を支援するものである．このような支援機構の実現によって，個々のオブジェクトは処理形態，信頼性，実行環境などに応じた最適な高信頼化方式を選択することが可能となる．さらに各オブジェクトの処理内容に応じて，両方式を**図-3**に示すように簡単に混在させることもできる．

図-3は，オブジェクトA，B，C，Dがそれぞれ3重化，2重化，3重化，2重化され，放送通信を用いて情報交換している．このうちオブジェ

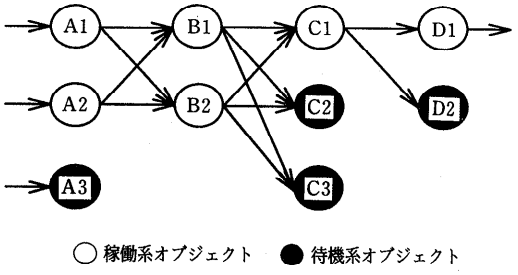


図-3 多重化オブジェクトによる動作例

クトAは，2つの稼動オブジェクトと1つの待機オブジェクトからなり，オブジェクトA1とA2は並列多重方式で制御され，A3は待機オブジェクトであるという混在型である．また，オブジェクトBは並列多重方式で制御され，オブジェクトCとDは待機冗長方式で制御されている．ここで，オブジェクトAのような混在型多重方式は，処理のノンストップ性と資源の有効利用の両方の性質を合わせ持つ．

つまり，各オブジェクトは自身の状態(たとえば，稼動/待機といった実行モード)のみに依存して振る舞い，多重化オブジェクトのすべてが稼動状態に宣言されれば，自動的に並列多重方式で制御され，1オブジェクトだけが稼動系で他のオブジェクトが待機系であるときは待機冗長方式で制御される．もちろん，2つ以上のオブジェクトが稼動状態で他が待機状態に宣言されれば，混在型多重方式で制御される．

稼動状態にあるオブジェクトは，それぞれ独自に処理結果を関連オブジェクトに放送する．このとき，放送通信を用いてメッセージを送信するので，関連オブジェクトの位置・多重度・実行モードを意識しない．たとえば，図-3におけるオブジェクトA1，A2はそれぞれの処理結果をオブジェクトBに放送する．受信側オブジェクトは，送信元オブジェクトが多重化されている場合は，その多重化オブジェクトから送られるメッセージ中から正しいメッセージを選択受信する．これは，知的分散OSの持つ，オブジェクトが定められた所定数の内容が一致するメッセージを受信した段階でそれを有効化する多重化メッセージ選択機構によって実現する[2),3)]．たとえば，オブジェクト$B_i(i=1,2)$は，$A_i(i=1,2)$から送られるメッセージ内容を比較して正しいメッセージを選択する．

多重化メッセージ選択機構が正しく動作するためには，比較対象となるメッセージを識別する必要がある．したがって知的分散OSでは，多重化されたオブジェクト群が，同一の処理過程で送信するメッセージに対して同一のメッセージIDを付加する機構を持つ．このメッセージIDは，個々のオブジェクトの持つ情報だけで発番するので，オブジェクトの独立性が保たれている．

よって並列多重方式で制御される多重化オブジェクトは，同一初期状態が与えられ，決定的動作が仮定された状況下では，多重化オブジェクトは同一メッセージを選択し，同一状態に至り一貫性は保証される．

待機冗長方式で制御される多重化オブジェクトは，待機系オブジェクトと稼働系オブジェクトの状態を一致させ，稼動系故障時にその処理を待機系オブジェクトが引き継ぐため，チェックポイントにて稼働系オブジェクトの状態を待機系オブジェクトにコピーする．知的分散OSではこのチェックポイント機構を，オブジェクトの位置と稼働/待機状態にあるオブジェクト数および，アプリケーションプログラムから独立に実現している[4)]．このために混在型多重方式の場合でも，稼動/待機オブジェクトの数や位置に依存せずに待機オブジェクトが処理を引き継ぐことができる．

チェックポイントのタイミングは，待機オブジェクトで受信するメッセージ量に応じて設定するので，単に周期的にチェックポイントをとる方法に比べ無駄なチェックポインティングを回避し，かつユーザ透過性を実現している．さらに稼動オブジェクトのチェックポイントタイミング設定処理による本来の処理の妨げを排除している．たとえば図-3におけるオブジェクト$C_i(i=2,3)$は，受信メッセージ量に応じてチェックポイントタイミングと判定すると，それぞれC1に対して状態コピーを要求する．C1は多重化メッセージ選択機構を用いてただ1回要求を受け付けると，その内部状態を$C_i(i=2,3)$に対して放送で返信する．よって，多重化オブジェクト間の状態一貫性が保証される．

稼動オブジェクト故障検出後の復旧処理は，待機オブジェクトが稼動オブジェクトの故障を検出した後，即稼動系となって処理を引き継ぐ．複数の待機オブジェクトが存在した場合は，稼動系となって処理を引き継いだ後に調停処理を行って，不要な新稼動オブジェクトを待機状態に戻す．このことによって，稼動系オブジェクト故障時の処理引継ぎのための新稼動オブジェクト選択処理においても，稼動系や待機系オブジェクトの位置や多重度に依存せず，かつできる限り短時間で処理を引き継ぐ．

このような高信頼化支援機構により，システム稼動中にオブジェクトの多重度や実行モードを動

的に変更することができる．たとえば，オブジェクトの信頼度を向上させるために稼動オブジェクトをコピーして多重度を増やしたり，処理の無停止性を実現するため待機オブジェクトを稼動系に変えることが，他のオブジェクトの動作に影響を与えることなく容易に実現できる．

4. おわりに

ソフトウェアモジュールの多重化による高信頼分散システム構築を支援する知的分散OSについて紹介した．ソフトウェアモジュール単位の多重化による高信頼化方式は，ハードウェア単位の多重化方式に比べ，潜在的に次に示す優れた性質を保持している．

1)要求の変化に対応して多重度や高信頼化方式を柔軟に変更可能．

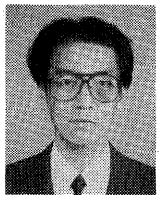
2)重要なモジュールのみを多重化することが可能となり，効率的資源利用が可能．

3)ソフトウェアモジュールの計算機間移動による負荷分散が可能．

参 考 文 献

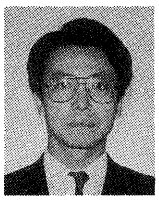
1) 田村他：知的分散システムのアーキテクチャ，電気論文誌C, Vol. 108, No. 6 (1988).
2) Seki, T. et al.: An Operating System for the Intellectual Distributed Processing System—An Object Oriented Approach Based on Broadcast Communication—, J. of Infomation Processing, Vol. 14, No. 4 (1991).
3) 関他：知的分散システムにおける高信頼放送通信機構，信学会論文誌，D-I, Vol. J 73-D-I, No. 2 (1990).
4) 関他：オブジェクト指向分散システムにおける放送待機冗長処理方式，電気論文誌D, Vol. 114, No. 3 (1994).
5) Birman, K. P.: The Process Group Approach to Reliable Distributed Computing, CACM, Vol. 36, No. 12, pp. 37-53 (1993).
6) Chereque, M. et al.: Activc Replication in Delta-4, Proc. of 22nd FTCS, pp. 28-37 (1992).
7) Borg, A. et al.: A Message System Supporting Fault Tolerance, Proc. of 9th ACM Symp. on OS Principles, pp. 90-99 (1983).
8) Kaashoek, M. F. et al.: Group Communication in the AMOEBA Distributed Operating System, Proc. of 11th ICDCS, pp. 222-230 (1991).
9) Birman, K. P. et al.: Lightweight Causal and Atomic Group Multicast, ACM Trans. on Computer Systems, Vol. 9, No. 3, pp. 272-314 (1991).

(平成6年9月14日受付)



関 俊文（正会員）

1960年生．1983年早稲田大学理工学部電気工学科卒業，1985年同大学院修士課程修了．工学博士．1985年(株)東芝入社．現在，同社研究開発センター，システム・ソフトウェア生産技術研究所勤務．分散システム全般に興味があり，分散OSや通信機構を含めた高信頼分散システムに関する研究に従事．電子情報通信学会，電気学会各会員．



長谷川哲夫（正会員）

1961年生．1985年早稲田大学理工学部電気工学科卒業．1987年同大学院修士課程修了．1987年（株）東芝入社．現在，同社研究開発センター，システム・ソフトウェア生産技術研究所勤務．分散システムに関する研究に従事．