データベース・システム 31－3
（1982. 7. 15）

放送通信機能を用いた分散型データベースシステムの通信処理の実現について

竜沢 誠
（財）日本情報処理開発協会

1. 序

分散型データベースシステム（以降DDBSと記す）とは、意味的に関連するデータを有する複数のデータベースシステム（DBS）が、通信網によって相互結合され、利用者に1つの仮想DBS（i.e. 1つのスキーマと1つのアクセス言語）としてのDBSサービスを提供出来るシステムである[TAKIM 78, 79]。DDBSとしては、70年代後半に、SDD-1[ROTH J80], POLYPHEME[ADIBM80]等が開発されている。これらは、(i)関係DBSから成る同種システム、(ii)広域パケット交換網（e.g. ARPANET, CYCLADES）の利用、(iii)問合せの分散処理の実現等を特徴としている。これらを、第I期DDBSと呼ぶ。ここでの主要な問題の1つは、広域網の低速性から生ずる性能問題であった。これらに対して、80年頃開始されている新DDBSは、次の様な特徴を有している。

(i) 異種DBS（e.g. CODASYL型DBS）の統合（JDDBS[TAKI80], MULTIBASE[SMITH J81]）
(ii) ローカル網（LAN）（e.g. Ethernet）の利用（SIRIUS-DELTA[LEBI J81]）による通信問題の解決。
(iii) 分散更新処理の実現
(iv) 従来の大型DBSに加えて、個人用小型DBS（パソコン＋DBS＋DP）の組込み
(v) オフィス情報処理への適用

当協会で1977年より開発を進めているJDDBS（Jipdec DDBS）は、上記5つの目標達成を目指している。これらのシステムを第II期DDBSと呼ぶ。第II期DDBSの主要な特徴の1つは、高速（1～10Mbps）ローカル網によって第I期DDBSにおける通信問題の解決を試みていることである。

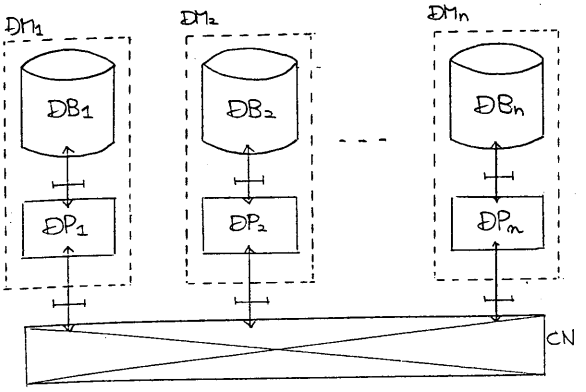
今後の情報システムは、各組織体（例えば会社の部、課）毎に、組織内の個人用DBS、既存大型DBSがローカルに結合されたローカル情報システムを基本単位として、これ等が広域網で結合された形態を取ると考えられる。ローカル情報システム（LIS）内では、各DBS間での大量、頻繁なデータ利用が、大容量LANを通して統合的にたされる。一方、LIS間では、大量かつ頻繁なデータ通信処理に対する要求はあまりないと考えられる。又、異なった組織体間でのデータ統合は、データの意味、定義域の差等が問題となり、実現にかなり困難とみられる。従って、DDBSは、まずLIS毎にLANを用いて実現される。

Ethernetの様なLAN，又無線網の様に通信媒体（e.g. 同軸ケーブル，無線）を共有する形態のものでは、従来の広域パケット交換網に対して、高速、高信頼であることに加えて、物理的に1対多（放送）通信機能をサポート出来る特徴を有していて、DDBSにおける通信処理は、冗長コピーの同時更新問題[BERN P81]に見られる様に本質的に1対多通信に基づいている。この為、LISとしてのDDBSを共有媒体型LANで実現することは有効である。

本論文では、開発中のJDDBS[TAKIM 78, 79, 80, 82a, b]で実現を試みている放送通信機能に基づいた通信処理方式について論じる。異種DBSのサポートについては、[TAKIM 79, 80, 82b]を参照されたい。

2. ＤＤＢＳモデル

分散型データベースシステム(DDBS)は、図2.1に示す様に通信網で結合された複数のデータモジュール($DM_1, \cdots, DM_n$)の集合としてモデル化する。各データモジュール(DM)は、データベース(DB)とDBプロセッサ(DP)とから成る。DBには、関係の集合が格納されている。DPは、自分のDB内の関係に対する関係演算と、組単位の更新が出来るプロセッサである。又、DPは、通信網CNを介して、他のDMへ／から関係を転送出来る。DPが受信した関係、DPが処理した結果関係は、自分のDB内に格納される。CNは、DM間での高信頼な基本通信機能を有している。DDBSに対する要求の無は、DM間での通信と、DM内での処理とによって得られる。



DM : data module
DB : database　　DP : DB processor
CN : communication network

図2.1　DDBSモデル

3. 放送網のモデル

放送網では、あるデータモジュール(DM)から発せられたメッセージを、他の全てのDMは、同時に受信出来る。Ethernet、無線網の様に、同軸ケーブル、無線といった通信媒体を共有した形態の網では、あるDMから出された信号は、網に結合されている全てのDMによって受信出来る。従って、これ等の網では、放送通信機能が物理的にサポートされていると言える。

利用者から見た通信網は、届け先と届けたい情報量に対するコスト(全通信時間)によってモデル化出来る。ここで、$C_{i\{j_1,\cdots,j_n\}}(x)$ ($n \geq 1$)を、DM_iから、n個の$DM_{j_1}, \cdots, DM_{j_n}$に情報量$x$を転送する時の全通信コスト(時間)とする。広域パケット交換網では、一対の通信実体間の高信頼な通信機能が、通信距離に依存せず通信量のみに依存したコストのもとでサポートされている。従って、(1)式の様になる。

$$C_{i\{j_1,\cdots,j_n\}}(x) \triangleq n(a + b \cdot x) \quad \cdots (1)$$

ここで、aとbとは定数である。aは通信リンクの初期化、切断のコストである。一方、$R_{i\{j_1,\cdots,j_n\}}(x)$を、一つの宛先の応答時間とする。同時に、$n$本の通信リンクが設定出来る場合には、最良の応答時間$a + b \cdot x$をもたらす。よって

$$R_{i\{j_1,\cdots,j_n\}}(x) \geq a + bx \quad \cdots (2)$$

これに対して、放送網では網上の信号を全てのDMが受信出来る為に、全通信

コストは、転送先の数にも依存になり、(3)式の様になる。

$$C_{i\{j_1,\cdots,j_n\}}(x) \triangleq a + b\,x \qquad \cdots (3)$$

又、放送網では、通信媒体が全てのDMで共有されている為に、同時に1つのDMのみが信号を出すことが出来る。よって、応答時間は(3)と同一になる。

$$R_{i\{j_1,\cdots,j_n\}}(x) \triangleq a + b\,x \qquad \cdots (4)$$

実際の通信網では、経路選択、待行列、競合等による統計的要因が加わり(1)~(4)式はより複雑なものとなる。ここでは、網は軽負荷であると仮定する。この仮定のもとで、(3)及び(4)式が成り立つ網を放送網と定義する。

4. 分散問合せ処理

複数のデータモジュール(DM)内の関係を参照する問合せを、DM間での通信と、各DM内でのローカル処理(i.e. 通信処理[TABAK80])によって解くことを分散問合せ処理と呼ぶ。本章では、3.で定義した放送網を用いた分散問合せ処理アルゴリズムについて論じる。

4.1 仮定

分散問合せ処理問題を考えるうえでの基本仮定について述べる。第1に、コストについて考える。放送網の通信コストは、(3)及び(4)式が成り立つとする。各DMでの処理コストは、広域網と比較して、放送網の高速性とDMとして小型システムも考えることから、通信コストと比肩し得るものとする。R_1, R_2を関係、a, bを各々の属性とする。$j(R_1,R_2)$, $p(R_1)$, $r(R_1)$を各々、結合$R_1[a=b]R_2$, 制限$R_1[a=\alpha]$, 射影$R_1[a]$とする。又、$|R_1|$と$|R_2|$を各々R_1, R_2のサイズとし、演算ωに対して$P(\omega)$をその処理コストとする。又、$st(a)$を、属性aの選択度[HEDNA78]とする。この時、関係演算コストを次の様に仮定する。

$$P(j(R_1,R_2)) = |R_1| + |R_2|$$
$$P(r(R_1)) = |R_1| \cdot st(a)$$
$$P(p(R_1)) = |R_1|$$

各DMには、互いに独立な関係が格納されていて、冗長コピーの存在は考えないものとする。関係の所在が不可視な視野(ビュー)(全体概念スキーマ[TAKI78, 79])に対する問合せは、問合せ修正手法[STON75]を用いて、各DM内の関係を参照する問合せ(全体問合せと呼ぶ)に既に変換されているものとする。

問題を簡単にする為に、全体問合せは、sum, maxといった統計関数(aggregate)を含まず、結合として等結合のみから成るものとする。全体問合せは、関係計算言語QUEL[STON76]によって記述されるものとする。

最後に、分散問合せ処理の最適化の目標について考える。通信処理スケジュール(又は単にスケジュール)とは、DM間でのデータ通信と、DM内での処理(関係演算)との実行順序を表す順序集合とする。全体問合せに対して複数の可能なスケジュールが存在するが、この中から以下の目標を達成する様なスケジュールが選ばれる[HEDNA78]。

(1) 全通信処理コストの最小化

(iii) 応答時間の最小化

4.2 放送網による特徴

分散問合せ処理の実現において、以下の点を検討する必要がある。

(i) 通信形態

(ii) スケジュール決定方法

(iii) スケジュール実行の管理手(controller)の所在

(i)の通信形態としては、既存広域網の1対1通信と、共有媒体型のローカルネットワーク等の1対多通信とがある。我々は後者を選択している。1対多通信を用いる方法としては、既に[TAKIM81][TOANN81][KAMB81]等がある。[TOANN81]では制御情報だけが放送される。

(ii)のスケジュールの決定方法としては、実行前に全てのスケジュールを決定してしまう静的方法と、実行中にその都度、次に何をするか決定する動的決定法とがある。[TOANN81]では両者を混合した方法が取られている。

(iii)では、各DM間/内の通信処理の実行の制御を、1つのDMが行う(集中制御)か、各DMが独自に判断する(分散制御)かの選択がある。[VINED81]はデータフロー方式の分散制御を提案している。

我々の方法は、以下の特徴を有している。

(i) データ及び制御情報の放送 (ii) 動的決定 (iii) 完全分散制御

各DMから出された情報は、放送網で結合された他の全てのDMによって受信出来る。これによって、1種の属性についてのm個($m \geq 2$)の異ったDM上の関係間の結合を行う時、1つの関係を放送すれば、他のm-1個の関係との結合をm-1個のDM上で行える[TAKIM81]。1回の通信で、m-1個の処理が出来る。又、各DMの状態情報(e.g. 関係のサイズ)を放送するならば、1回の通信によって、他の全てのDMはこれを知ることが出来る。即ち、DDBS内の全てのDM$_1$, ..., DM$_n$は、DDBS全体についての同一状態情報を、n回の通信によって得れる。よって、各DMは、自分の有するDDBS状態情報によって、他とは独立に、かつ同一の決定を行える(完全分散制御)。又、スケジュールの各ステップ毎に、各DMがその処理結果を放送することによって、その時点での最適なスケジュールを動的に決定出来る。これ等の特徴は、分散問合せ処理アルゴリズムを簡単なものとし、実現と運用とを容易なものとする。

4.3 全体問合せ

利用者から問合せが入力されたデータモジュール(DM)を、結果DM(RDM)と呼ぶ。結果DMでは、分散性が不可視な視野に基づいた問合せから、各DM内の関係を参照する全体問合せを生成する。全体問合せは、QUELによって、次の様に記述される。

$$\left.\begin{array}{l}\underline{\text{range}}\ (r_1, R_1:d_1)\ (r_2, R_2:d_2)\ \cdots\ (r_m, R_m:d_m);\\ \underline{\text{retriere into}}\ R\ (a_1 = exp_1, \cdots, a_\ell = exp_\ell)\ \underline{\text{where}}\ quals;\end{array}\right\} \cdots (5)$$

R_iは、DM_{d_i}に存在する関係であり、range文はR_iに対する組変数r_iを定義している($i=1, \cdots, m$, $m \geq 1$)。Rは結果関係名で、$a_1, \cdots, a_\ell$はその属性である。exp_kは、変数を$r_i.b_i$(r_iは関係R_iの組変数で、b_iはR_iの属性)とする算術式である

$(\ell=1,\cdots,k, k\geq 1)$. qual は 結合述語 $r_i.b_i = r_j.b_j$ $(r_i \neq r_j)$, 制限述語 $r_i.b_i \ \theta\ r_i.b_i'$ $(b_i \neq b_i')$ 又は $r_i.b_i \ \theta\ v$ $(\theta \in \{<, \leq, =, \geq, >, \neq\}$, v は定数) との2種の述語の積として表される論理式と仮定する.

例として, 図4.1に示す関係を考えよう. DM_1 には部と部員情報が格納され, DM_2 には従業員の会議予定(日時(date), 場所(place), テーマ(thema)) が, 又 DM_3 には社内のプロジェクト番号情報が格納されている. この時, 「"A"氏と"DDBS"の会議に同席する従業員の部と所属プロジェクト名」が知りたいとすると, 全体問合せは QUELで次の様に表せる.

DM_1 EMP (eno, ename, tel)
DE (dno, dname, eno)

DM_2 SCDL (eno, ename, date, place, thema)

DM_3 PROJ (pno, pname, eno, position)

図4.1 例

```
range (e, EMP:1) (de, DE:2) (s, SCDL:2)
      (p, PROJ:3) (s1, SCDL:2);
retrieve into R(d.dname, p.pname)
where s1.ename = "A"  and  de.eno = e.eno  and
      e.eno = s.eno  and  s1.thema = s.thema  and
      s1.thema = "DDBS"  and  s.eno = p.pno  and
      e.eno = p.pno ;        --- (6)
```

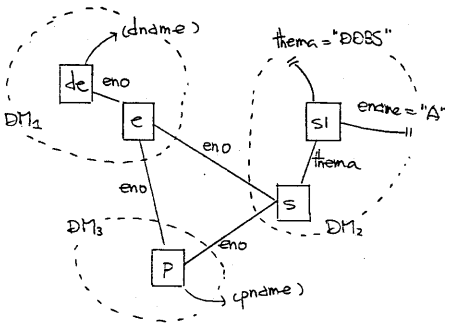


図4.2 (6)の問合せグラフ

(6)の問合せは, 図4.2の様にグラフ表現出来る. 節点は組変数を表し, 節点間の辺は結合式を表す. 節点に付加されている ⟶| と →は各々制限式と目標属性を表している.

4.4 初期ローカル問合せ処理

全体問合せの中で, 各DMで独立に処理出来る部分と, 各DMで処理し, DM間の通信処理に必要な部分のみから成る関係を生成することを初期ローカル問合せ処理(ILQP)と呼ぶ. 図4.2の例では, 点線で囲まれた部分が, 各DMで処理され, 図4.3の様な結果となる. 即ち, DM内の結合, 制限式が処理され, GCS問合せの目標属性と共に, DM間の結合属性だけから成る結果関係が生成されて。

ILQPの目的は, 後述するDM間での通信処理の為の通信量の縮小と共に, 異種DBSから条件を満足するデータを関係形式に導き出す目的を持っている. CODASYLモデルの様に構造を有したデータモデルでは, 関係モデルの様な閉包性を有していない. 通信処理では中間結果の動的な生成, 消滅操作が必要となり, 閉包性が求められる. 従って, まず異種DBSから, 関係インタフェースLDP[TAKI80, 826]によって結果を関係として導出する必要がある. ILQPで生成された関係は, 各DMのDBに格納される.

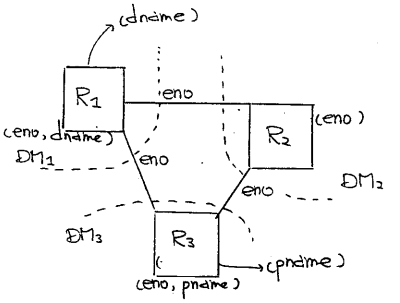


図4.3 ILQP後の問合せグラフ

4.5 単純問合せのDM間通信処理

ILQP後の全体問合せは、DM間の結合だけから成る。きず、図4.4に示す様に、各々異ったDM上のn個の関係$R_1,\cdots,R_n$間に、単一結合属性aがある全体問合せを単純問合せと呼び、これについて考える。[TAKIM81]では、動的決定と分散制御に基づく次の様なアルゴリズムを提案している。

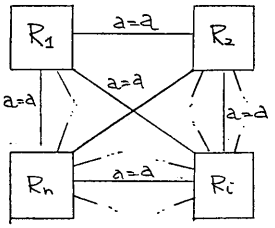


図4.4 単純問合せ

(1) $\Omega = \{R_1,\cdots,R_n\}$ とする。 $\Omega' \leftarrow \Omega$;

(2) $R \in \Omega$ で $|R[a]|$ が最小のものを見つけ、これを放送し、ΩからRを除く。

(3) $\forall R' \in \Omega$ $\quad R' \leftarrow R'[a=a](R[a])$;

(4) (2),(3)を $\Omega = \phi$ となるまで繰り返す。最後の関係の放送時に、(3)では$\forall R' \in (\Omega' - \Omega)$に対して特別な行為を行う。

(3)の終了した時点で、各DMは正のacknowledgement(ACK)に得られた結果のパフォーマンス情報(サイズやカーディナリティ)を乗せて放送する。全てのDMは、他の全てのDM内の関係のパフォーマンス情報を知ることが出来るので、(2)について同一の決定を行うことが出来る。[TAKIM81]では、更に[HEVNA79]の1対1通信に基づいたアルゴリズムとの比較により、大幅な通信コストの減少をもたらすことを示している。これを、アルゴリズムBA(broadcast algorithm)と呼ぶ。

ここでは、更に、[KAMBY81]によるデータ圧縮技法を加えることにより、より通信量を減少出来るアルゴリズムを示す。我々の方式と同様な方法が、上林[KAMBY82]によって独立にたされている。アルゴリズムBAでは、1つの関係の射影($R[a]$ $R \in \Omega$)の放送に対して、各DMのACKは、単に結合結果のカーディナリティのみを含んでいる。この為、n個の関係に対して、n-1回分の関係の放送と処理、Ack転送が必要になる。これに対して、関係の放送のACKに、カーディナリティに加えて、生成された関係$R'[a]$の$R[a]$に対するビットマップを乗せてしまう方法である。各DMは他のDMからのACKを受信しビットマップの積を取っていく事により最終結果を得ることが出来る。この方法にさえば、1回の関係の放送と、各DMでの処理とAck放送だけで処理されてしまう。これを、アルゴリズム BBA(bit-map BBA)と呼ぶ。

アルゴリズム BBA

(1) $\Sigma = \{R_1,\cdots,R_n\}$ とする。

(2) $R \in \Sigma$ で $|R[a]|$ が最小のものを見つけ、これを放送する。$\Sigma \leftarrow \Sigma - \{R\}$;

(3) $\forall R' \in \Sigma$ で、 $R' \leftarrow R'[a=a](R[a])$;
$R'[a]$の$R[a]$に対するbit-map BM'_aを生成する。

(4) BM'_aの放送を行い、他の全てのDMからのbit-map BM_aを待つ。

(5) BM_aを受信したたらば $BM'_a \leftarrow BM'_a \wedge BM_a$;

(6) 全てのDMからBM_aを受信したたらば、BM'_aを基にして、最終結果$R'[a]$を生成する。

図4.5には、アルゴリズムBAの例を示してある。

次に通信処理コストについて考える。BAにおいて、関係は $R_1, R_2,\cdots,R_{n-1}$の順に放送されるとする。又、属性aの選択度をmとし、各関係のaの値は均一分散しているとする。この時 通信コストC_{BA}と処理コストP_{BA}は (7),(8)式の様

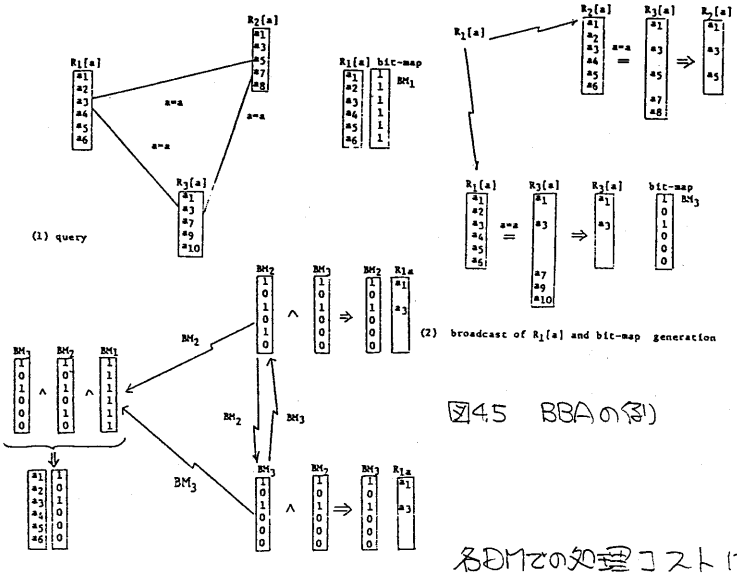


図4.5 BBAの例

の様になる。

$$C_{BA} = |R_1[a]| + \alpha|R_2[a]| + \cdots + \alpha^{m-1}|R_m[a]| = \sum_{i=1}^{m} \alpha^{i-1}|R_i[a]| \quad \cdots (7)$$

$$P_{BA} = (m-1)|R_1[a]| + |R_2[a]| + \cdots + |R_m[a]|) + \alpha((m-2)|R_2[a]| + |R_3[a]| + \cdots + |R_m[a]|) + \cdots + \alpha^{m-2}(|R_{m-1}[a]| + |R_m[a]|) + (|R_1[a]| + \alpha|R_2[a]| + \cdots + \alpha^{m-2}|R_{m-1}[a]| - (m-1)\alpha^{m-1}|R_m[a]|) \quad \cdots (8)$$

各BMでの処理コストは、4.1の仮定に基づいている。1つの関係$R_i[a]$の放送は、$m-1$の結合処理$J(R_i[a], R_2), \cdots, J(R_i[a], R_m)$をもたらすので、処理コストは、$(m-1)\cdot|R_i[a]| + \sum_{j=2}^{m}|R_j[a]|$ となる。この結合処理で$R_i[a]$は、$\alpha\cdot|R_i[a]|$ のサイズになる。ヌ、(7)では ACKメッセージの通信コストを無視してある。

BBA の通信コストC_{BBA}，処理コストP_{BBA}は、$R_i[a]$が放送された関係とすると(9)、(10)式の様になる。

$$C_{BBA} = |R_i[a]| + (m-1)\cdot card(R_i[a]) \quad \cdots (9)$$

$$P_{BBA} = \sum_{i=2}^{m}(|R_i[a]| + |R_i[a]|) = (m-1)|R_i[a]| + \sum_{i=2}^{m}|R_i[a]| \quad \cdots (10)$$

(9)式の右2項は、ビットマップの転送コストである。処理コストでは、ビットマップ演算が主記憶内で出来るとして、無視してある。$P_{BBA} \leq P_{BA}$は明らかである。$|R_i[a]|$は最小であるので

$C_{BA} \geq |R_i[a]|(1+\alpha+\cdots+\alpha^{n-1})$　　$C_{BBA} = |R_i[a]|\left(1 + \frac{m-1}{|a|}\right)$。よって、$|a| \geq (m-1)\cdot\frac{1-\alpha}{\alpha-\alpha^m}$ であれば、$C_{BBA} \leq C_{BA}$ である。$|a|$を2バイト(16 bits)、$m=2$の時 $\alpha \geq 0.06$ であれば、BBAの方が有利になる。

応答時間R_{BA}，R_{BBA}は、次の様になる。

$$R_{BA} = C_{BA} + \sum_{i=1}^{m} \max_{j\in\{1,\cdots,m\}}(\alpha^{i-1}[|R_j[a]| + |R_i[a]|]) \quad \cdots (11)$$

$$R_{BBA} = C_{BBA} + \max_{i\in\{2,\cdots,m\}}(|R_i[a]| + |R_i[a]|) \quad \cdots (12)$$

図4.6には、放送網における、1対1通信を用いた[HEVNA78]，及びのBA，BBAの各々における通信コストをまとめてある。(1)，(2)式で$a=0$，$b=1$としてある。データは、[HEVNA78]を用いた。図から、放送通信機能を用いたBA，BBAの優位性が解かるとともに、データ圧縮技法(ビットマップ)利用効果も明らかである。

4.6 一般問合せのDM間通信処理

次に、異なったDMに存在するm個の関係$R_1,\cdots,R_m$ ($m\geq 2$)が、n (≥ 1)個の結合属性$a_1,\cdots,a_n$によって等結合されている一般問合せについて考える。本来問合せ[BERNP対]でない問合せに対しては、複数の属性の組の値が転送される必要がある。以下に、一般問合せに対するアルゴリズムGAを示す。ここで、Ω_{R_i}を関係R_iの結合属性の集合とする。Ωをすべての結合属性の集合とする。各結合属性$a\in\Omega$に対して、$\overline{R}_a$をaを有する関係集合とする。Σを$R_1,\cdots,R_m$の集合とする。

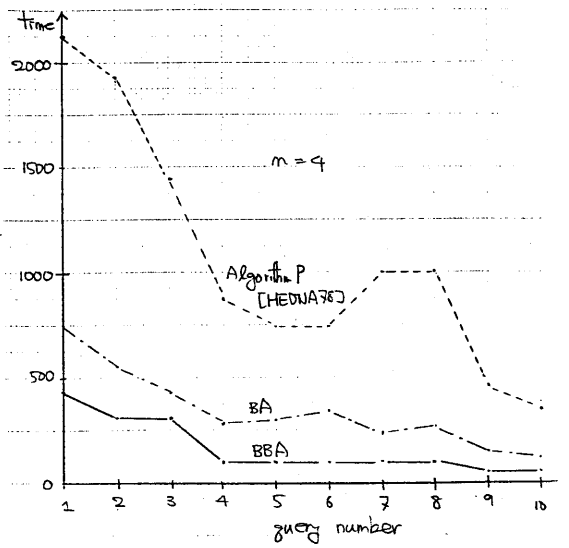


図4.6 全通信実問

アルゴリズムGA

(1) $STA\leftarrow\phi$; $\forall R_i\in\Sigma$ $STA_i\leftarrow\phi$; $\ell\leftarrow 1$;

(2) $\forall a\in\Omega$ $\forall R\in\overline{R}_a$ $R[a]$が最少のものを見つける。これを$R_\ell[a_i]$とする。a_iをℓ番目のステージ属性と呼び、R_ℓをソース関係と呼ぶ。$R_\ell[a_i]$の放送を行う。 $STA\leftarrow STA\cup\{a\}$; $\Omega\leftarrow\Omega-\{a_i\}$;

(3) $a_i\in\Omega_{R_j}$なる全てのDMで、 $\forall a\in STA_j$; $R_{ja}\leftarrow R_j[a]$; $STA_j\leftarrow STA_j\cup\{a_i\}$;
結合処理 $R_j\leftarrow R_j[a_i=a](R_\ell[a_i])$; $R_{ja_i}\leftarrow R_\ell[a_i]$;
$\forall a\in STA_j$に対して、$R_{ja}\supset R_j[a]$であれば、$R_j[a]$のR_{ja}に対するビットマップBM_{ja}をつくる。同時に $\forall b\in(\Omega_{R_j}-STA_j)$に対して、$R_j[b]$のサイズ$s_{jb}$をもとめる。

(4) これらのBM_{ja}とs_{jb}をACKメッセージとして放送する。他のDMからのACKを待つ。

(5) 他のDM$_k$からのBT_{ka} ($a\in STA_j$)を受信したならば、$BM_{ja}\leftarrow BM_{ja}\wedge BM_{ka}$;

(6) 全てのDMからACKを受信したならば、$\forall a\in STA_j$ BM_{ja}に基づいてR_jを再書式する。

(7) $\ell\leftarrow\ell+1$; $\Omega\neq\phi$ ならば (2)～(7)を繰り返す。

最悪的に、ステージ属性は$a_1,\cdots,a_m$の順に取られたとする。i番目のソース関係$R_\ell[a_i]$のサイズを$|R_\ell[a_i]|$とする。この時、i番目のa_iを処理する為の通信コストC_iは、(13)式の様になる。

$$C_i = |R_\ell[a_i]| + \sum_{\substack{j=1\\(j\neq\ell)}}^{m}\delta_{ij}\cdot\left[card(R_\ell[a_i]) + \sum_{\ell=1}^{i-1}\tau_{j\ell}\cdot card(R_j[a_\ell])\right]$$

$$\text{ここで、}\ \delta_{ij}=\begin{cases}1 & \text{if } a_i\in\Omega_{R_\ell}\\ 0 & \text{otherwise}\end{cases}\qquad \tau_{j\ell}=\begin{cases}1 & \text{if } a_\ell\in STA_j\\ 0 & \text{otherwise}\end{cases}\qquad\cdots(13)$$

(13)式の第1項は$R_\ell[a_i]$の放送コスト、第2項は$BM_{\ell a_i}$、第3項はBM_{ja_ℓ} ($a_\ell\in STA_j$)の転送コストである。4.5と同様に、ビットマップ処理コストを無視すると、i番目のa_iの処理コストP_iは、(14)式の様になる。

$$P_i = \sum_{\substack{j=1\\(j\neq\ell)}}^{m}\delta_{ij}\cdot\left[|R_\ell[a_i]| + |R_j[a_i]|\right]\qquad\cdots(14)$$

応答時間R_iは、$R_i = C_i + \max_j\left(\delta_{ij}[|R_\ell[a_i]| + |R_j[a_i]|]\right)\cdots(15)$となる。

全通信コスト、処理コスト、応答時間は、これらの和($i=1,\cdots,m$)となる。

一般問合せも、放送網を用いることによって、容易にかつ有効に処理出来る。処理コストと通信コストの重みづけの割合は、今後の課題である。

5. 分散更新処理

利用者からみた原子的な実行単位は、トランザクションと呼ばれる。更新処理では、この原子性を保つ必要がある。実行が原子的であるとは、正常終了した時のみ、この実行によるデータ変化を、他のトランザクションが参照し得、もし正常終了しなければ何らの影響もデータベース(DB)に与えないことである。トランザクション実行の原子性を保障する為の制御は、コミットメント制御と呼ばれる。即ち、トランザクション中が、n個のデータオブジェクト(e.g. 組) $x_1, \cdots, x_n$ $(n \geq 2)$ の更新を行おうとする時、$x_1, \cdots, x_n$ の全ての更新が行われたか、全くなされなかったかのどちらかになる様に更新実行は制御される。この為の手法としては、二相コミット方式[Gra78]がある。これは、次の様である。

(i) $x_1, \cdots, x_n$ に対する更新データを結果DM(RDM)に生成する。(ii) $x_1, \cdots, x_n$ の各々をデータモジュール(即ち $DM_1, \cdots, DM_n$ とする)に、更新データを含めた precommit (x_i) $(i=1, \cdots, n)$ 命令を送る。(iii) 各 DM_i は、precommit (x_i) を受信したならば、更新データを当分の安全な格納領域(e.g. ログ)に待避する。格納出来たならば precommitted を RDM に送る。失敗すれば abort を送る。(iv) RDMは、全ての DM_i から precommitted を受信すれば、各 DM_i に commit 命令を出す。全ての DM から precommit を受信出来なければ、各 DM_i に abort を送る。(v) 各 DM_i は RDM から commit を受信すれば、安全領域での更新データによって DB_i 内の x_i の物理的な更新を行い、acknowledgement (Ack) を RDM に送る。この事に、合計で 4n 回の通信が必要になる。

放送網を用いて、分散制御を行わせると、二相コミット法を次の様に簡単化出来る。

(i) 放送通信による分散問合せ処理において、$x_1, \cdots, x_n$ に対する更新データが生成される。生成されたDMが、更新データを放送する(precommit に相当)。

(ii) 各 DM_i は、受信又は生成された更新データを安全領域に待避する。待避出来たならば、precommitted を放送する。

(iii) 他の全ての DM からの precommit を待つ。全てから precommit を受信したならば、DB_i の x_i に対する更新を、安全領域のデータを用いて行う。

(iv) 更新が出来たならば、Ack を放送する。

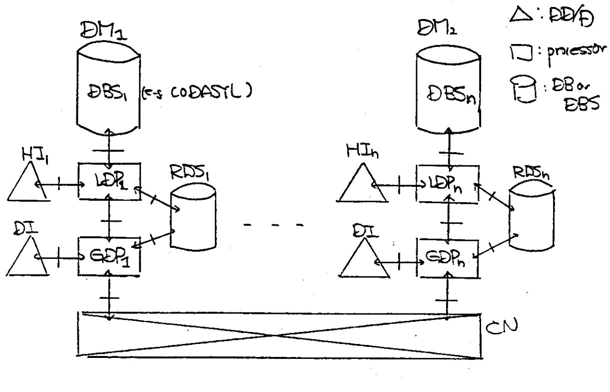
必要な通信時はいずれも1回、(ii)で n 回、(iv)で n 回、都合 2n + 1 回となり、前者に対して通信量を半減できる。冗長コピーを保有する場合も、冗長性を意識することなく分散更新を行うことができる。

6. まとめと実現

本論文では、当研究室で現在開発中の分散型データベースシステム SDDBS において実現を試みている分散問合せ処理と分散更新処理アルゴリズムについて論じた。Ethernet, 無線網の様に、通信媒体を共有した形態の通信網によって物理的に提供される放送(1対多)通信機能を用いたアルゴリズムを述べた。放送網を用いることによって、各データモジュール(DM)での分散制御と、自由なスケジュール決定とが可能となり、DDBSの分散問合せ、更新処理の実現と運用を簡単化することができ

きる。今年度は、複数DMを、同一ホスト(M-170F)に実現し、JIPNETのエンド・エンドプロトコルスはAIM DB/DCによってDM間通信を行う予定である。既に、CODASYL DBS上の関係インタフェースシステム LDP-V2[TAKIM80,82b,c]はAIM(M-170F)及びADBS(Acos-700)上で実動している。

GDPは、本論文のDPに相当し、RDSは、DBに相当する。



LDP: local DB processor　HI: heterogeneity information
GDP global DB processor　DI: distribution information
RDS: relational working storage　CN: communication network

図6.1 JDDBSアーキテクチャ

謝辞

JDDBSの実現に協力頂いているシステラワ社の鈴木 信、当協会の菅家 実、塚本 元子の各氏に感謝します。日頃御指導頂いている当協会開発部の山本欣子、小関 眞美の両氏に感謝します。

参照文献

[ADIBM80] Adiba, M., et al., "An Overview of the Polypheme Distributed Database Management System," Proc. of the IFIP'80, Oct. 1980, pp.475-479

[BERNP81] Bernstein, P.A., et al., "Concurrency Control in DDBS," ACM Comp. Survey, Vol.13, No.2, June 1981.

[BERNP79] Bernstein, P.A., et al., "Full Reducers of Relational Queries Using Multi-Attribute Semi-Join," Proc. of IEEE Comp. Net., Dec. 1979

[GRAYJ79] Gray, J.N., "Notes on DB OS," in Operating Systems, Springer-Verlag, 1979

[HEVNA78] Hevner, A., et al., "Query Processing on a Distributed Databases," Proc. of the 3rd Berkeley Workshop, Aug 1978

[HEVNA78b] Hevner, A., et al., "Optimization of Data Access in Distributed Systems," TR281, Purdue Univ. July 1978.

[KAMBY81] Kambayashi, Y., "正規型準結合による分散型DBの質問処理," 研究会 AL81-54, Sep. 1981

[KAMBY82] Kambayashi, Y., "Theory of Databases," Comp. Scie. & Technology (Kitagawa, T. ed.) North-Holland and OHM., 1982

[LEBIJ80] Lebihan, J. et al., "SIRIUS: A French Nationwide Project on DDBS," Proc. of the 6th VLDB, Oct. 1978

[STONM75] Stonebraker, M. et al., "Design and Implementation of INGRES," ACM TODS, 1975

[SMITJ81] Smith, J., et al., "Multibase - Integrating Heterogeneous Distributed Database Systems," AFIPS Conf. Proc., 1981

[ROTHJ80] Rothnie, J.B. et al., "Introduction to a System for DDB (SDD-1)", ACM TODS, Vol.5, No.1, Mar. 1980

[TAKIM78] Takizawa, M. et al., "Resource Integration and Data Sharing on Heterogeneous RSS," Proc ICCC'78, Sep. 1978

[TAKIM79] Takizawa, M. et al., "The Four-Schema Concept as the Gross-Architecture of Distributed Database and Heterogeneity Problems," JIP, Vol.2, No.3, Nov. 1979

[TAKIM80] Takizawa, M. et al., "Query Translation in Distributed Databases," Proc. of the IFIP'80, Oct. 1980.

[TAKIM81] Takizawa, M., "分散型データベースシステムJDDBS-II と質問処理," 研究会 AL81-67, June 1981.

[TAKIM82a] Takizawa, M., "Distribution Problems in Distributed Databases - Integration and Query Decomposition," to appear in JIP, 1982

[TAKIM82b] Takizawa, M. et al., "CODASYLデータベース・システムに対する関係インタフェース・システム - LDP-V1.5 - の設計と実現," to appear in IPSJ 論文誌 1982

[TOANN81]

[VINED81] Vine, D.H., "A Data Flow Solution for Implementing Distributed Queries," Proc. of the 5th Berkeley, Feb. 1981