広域疎結合分散システムのためのデータ配送機構の設計

井澤 志充　三輪信介　篠田陽一

北陸先端科学技術大学院大学
情報科学研究科

ネットワーク上に広域に分散したシステム間でデータ配送を行う場合には、TCP に基づいた独自のプロトコルを用いてデータ通信を行うのが一般的である。例えば、CORBA などの分散オブジェクトインフラストラクチャや NetNews などで用いられている NNTP などがこれにあたる。しかし、これらのデータ配送機構はネットワーク故障に対する頑健性やネットワークトポロジの変化への追従性等において問題がある。そこで本稿では、これら従来用いられてきたデータ配送機構の問題点を指摘し、大規模災害時に発生するネットワークの輻輳や断絶、ネットワーク特性が著しく異なるメディアを用いた場合にも適用できるデータ配送機構の提案を行う。また、提案する機構の適用範囲やその拡張性についても論じる。

A design issue of data transport mechanism for the loosely coupled widely distributed system

Yukimitsu Izawa　Shinsuke Miwa　Yoichi Shinoda

School of Information Science,
Japan Advanced Institute of Science and Technology, Hokuriku

Wide spread distributed transport system, for example CORBA and NetNews system with NNTP has been applied to network infrastructure. But these data transport system is fragile in the situation that network trouble however network partitioning and network congestion. Robustness is required for the data transport system that against for the network trouble, changing network link topology for these system. But these former data transport system do not have robustness. In this paper, we describe data transport mechanism design. This design aim to making robustness, scalable and extendable data transport system. We discuss about usability and robustness of this mechanism.

1　はじめに

従来、広域に分散したシステム間のデータ通信をサポートするデータ配送機構は、安定したネットワークを前提にしているものが多く、安定して稼働するサーバがネットワーク上を移動するシステムのサポートを行うなど、何らかの安定したシステムを前提としたモデルで設計されている。

このようなモデルで設計されたデータ配送機構を用いる問題点として、大規模災害時に発生するようなネットワークの激しい輻輳や分断が起こった場合などを想定していないため、そのような場合には利用することができない。

これは従来のデータ配送機構が、配送における頑健性を念頭に置いて設計されていないことに起因している。つまり、配送における頑健性を提供することができるデータ配送機構であれば、大規模災害時でも有効に機能することができるといえる。

従来のデータ配送機構を用いて上記のような問題に対応するために、データ配送機構を用いるアプリケーションでこれらの問題を解決する方法が多くみられる。

しかしながらアプリケーションのレベルでこのような問題を解決する方法は、

- アプリケーション毎にこれらの機能を実現する必要がある。
- アプリケーションが下位の状態を把握する必要があるため、ネットワークレイヤの観点からみるとレイヤバイオレーションをおこしている。
- すべての情報がアプリケーションで処理されることになるため、プロトコルオーバーヘッドが大きくなる。

等の問題点を持つ。

そこで、データ配送機構に広域負荷分散やレプリケーション、グループ化などをサポートする機能を持たせることが上記の問題を解決する一つの手法である。

データ配送機構自身が、このような機能を内包することで、より効率的な広域負荷分散の実現やより一般的で様々なデータ配送を必要とするアプリケーションをサポートすることができる可能性を持つと考えられる。

2 従来の技術

従来、広域に分散するシステム間のデータ通信をサポートする機構には以下のような機能を持つことが要求されてきた。

- アプリケーションがホストとリソースを特定し、そのリソースに対してのデータ配送を行う。
- 様々なアプリケーションから利用できるような一般的なインターフェースを提供する。

このような要求に対し、データ配送に関する機能は最小限の機能を提供し、アプリケーションがこの最小限の機能を利用して、付加的なデータ配送のしくみを実現する方法が一般的に採られている。[7, 8]

これを図で示すと図1のようなレイヤ図であらわすことができる。

このようなデータ配送技術を用いてデータ通信を効率的に行うため、図2のようなレプリケーションや図3のようなオーソリティサーバを用いる方法が一般的である。

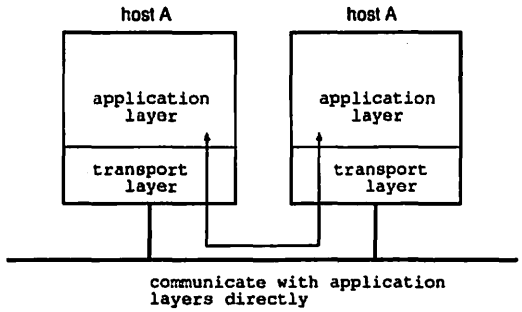

図 1: 従来型データ配送機構

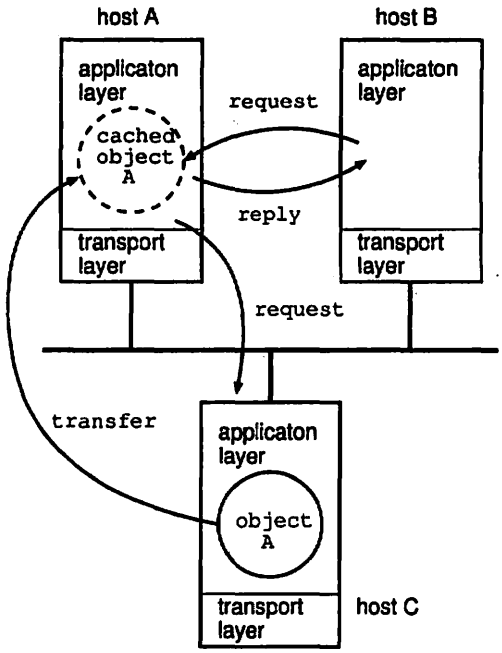

図 2: レプリケーション

いずれの機構も、データ配送における頑健性を考慮したモデルではないため、このような場合には有効に利用できない。

3 設計方針

本稿で提案する手法のねらいは、以下のような要求をもつデータ通信における信頼性を向上させることである。

- 広域に分散したシステム間での同報データ通信
- 即時性は問わない
- 耐規模性に優れる

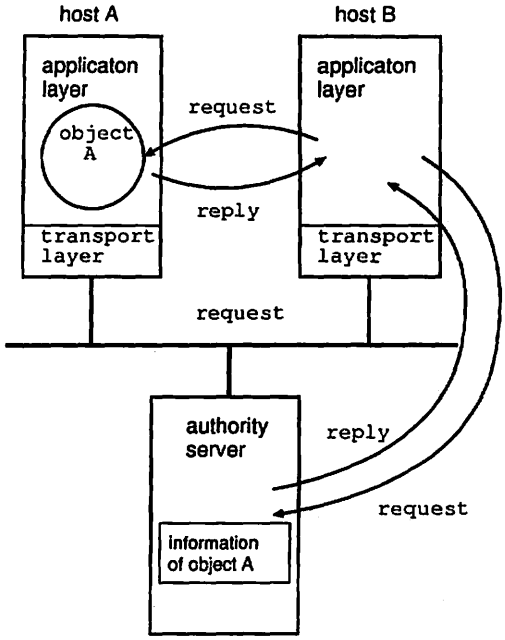

図 3: オーソリティサーバ

1番目の項目における同報データ通信とは、データ配送機構によって接続されている分散したシステム全てに同様の情報を伝搬する機能を有する必要があることをさしている。これは、1対多での通信をサポートすることを意味する。1対多通信をサポートすることで、例えばこのデータ配送機能によって結ばれる分散データベースの全てのクラスタで同様のレコードを容易に保持することができるようになり、分散データベース全体の系に冗長性を持たせることができる。

2番目の項目としてデータ配送の即時性を重要視しないことを挙げた。これは、即時性と頑健性がトレードオフの関係に陥り易いことに起因しており、我々は即時性よりも頑健性を重要視しているためである。

3番目の項目として挙げた耐規模性は、このデータ通信機構が様々なアプリケーションが利用できるようにするためには重要な要素である。

またその他に、アプリケーションに対して分散環境を隠蔽することができることも必要である。あらかじめ分散環境で稼働することを考慮したアプリケーションから利用できるのは当然のことながら、アプリケーション側からは分散環境を利用していることが分からないようなインターフェースを提供することが必要である。

これは、アプリケーションに分散環境を意識したコードを含む必要性を無くすことで、本手法のデータ配送機構の適用範囲を広げることにもなる。

現在、このような要求をデータ配送機構のみで満たすものはみられず、データ配送機構を用いるアプリケーション側で必要に応じた処理を行う必要がある。

そこで本稿では、データ配送機構に広域負荷分散や同報性といった機能を拡張できるような枠組みを持たせ、データ配送において要求される機能をデータ配送機構に組み込むことが出来るシステム設計を提案する。

4 設計

前節で述べたように、従来のデータ配送機構と比較して高い拡張性や耐規模性、頑健性を提供するため本稿で提案するデータ配送機構には、大別して4層からなるレイヤモデルを採用している。

本節では設計のうち特に本方式のレイヤ構成について述べる。

4.1 レイヤ構造

本稿で提案するデータ配送機構は、端点に位置するシステム同士の通信 (point-to-point) から、グループ化されたデータベースクラスタへのデータ送信まで、様々なデータ通信の要求を満たすことができるように設計されている。

またアプリケーションには仮想的なホストを提供し、これらとのデータ通信を行わせる方式を採用することで、アプリケーションに対して実際のネットワークトポロジーを隠蔽することができるようにした。

これは、配送ネットワークの再構成やグループ化、あるいは激しい輻輳や分断といったネットワーク故障が起こった場合の自動対応などをアプリケーションに意識させる事なく行えることを意味している。あるいはアプリケーションがこれらの機能を明示的に指定することも可能である。

このような粒度の違う要求をみたすため、本稿で提案するデータ配送機構に以下の4つのレイヤを定義した。この4つのレイヤは図4のように構成されている。

- リンクアソシエーション層
- リンクコンフィギュレーション層
- レコード層
- レコードコントロール層

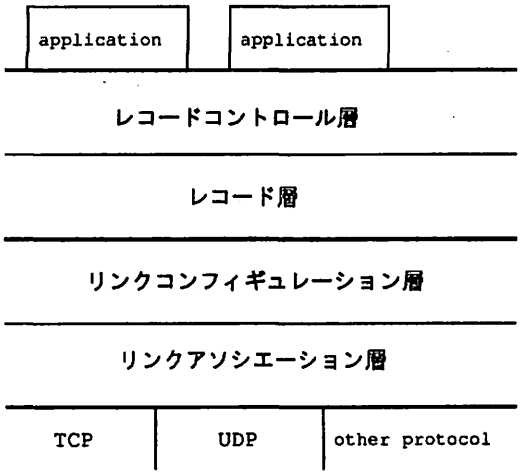

図 4: 4 レイヤ構成

これら 4 層からなるデータ配送機構の上位層にはアプリケーションレイヤがあり、これにサービスを行う。また下位層には TCP といったセッション指向のプロトコルの他にも UDP のようなデータグラム指向のプロトコルも使用することができる。

下位層にどのようなプロトコルを使用するかは、上位レイヤに対してどのようなサービスを提供するかに依存しており、また本機構内での実現方法によっても変わってくるため、実際にはさまざまなプロトコルを混在して使用することになることが考えられる。

次にこれら 4 つのレイヤそれぞれの役割について説明する。

4.1.1 リンクアソシエーション層

リンクアソシエーション層では、端点同士の接続に関する部分を担うレイヤである。この層は実際にあるホストとあるホストが通信を行うことを上位レイヤに保証する必要がある。

ここでいう保証とは、端点から端点へ実際にデータが誤りなく配送されたあるいは配送することが出来ないことを明らかにすることを意味している。

例えばリンクアソシエーション層の下位レイヤとして UDP などの確認応答がないプロトコルを用いる場合この層で確認応答を実現する必要がある。

また、ネットワークとして使用するメディア特性の違いはこの層で隠蔽されることになる。例えば、通信衛星網を用いた場合の単方向リンクを用いた場合などがこれにあたる。

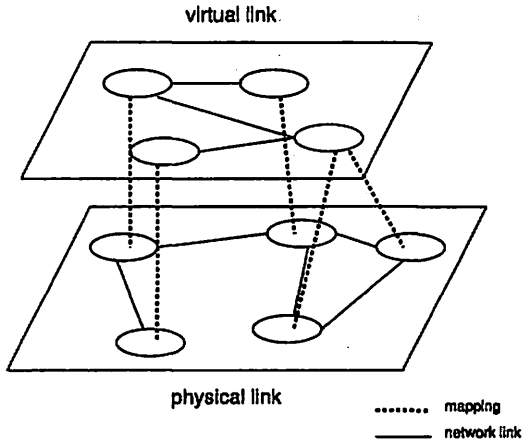

図 5: リンクコンフィギュレーション層の内部構成

4.1.2 リンクコンフィギュレーション層

リンクコンフィギュレーション層は以下の 2 つのサービスを行うレイヤである。

- データ通信を行うシステム間のネットワークを構築する。
- 上位レイヤには仮想的な配送ネットワーク構成を提供する。

図5に示すように、リンクコンフィギュレーション層は内部に実際のシステム間のネットワーク構成をもちながら、上位レイヤにはそのネットワーク構成を隠蔽した仮想的なネットワークを提供する。

本方式の要求事項である耐規模性を鑑みた場合に、ある端点から全てのデータ配送先に向けて直接データを送るのは現実的ではないため、データ配送のためのネットワークを構成する必要がある。この配送ネットワークの構成と管理を行うのがこの層の役割の一つである。

また、上位レイヤに対しては仮想的な配送ネットワークを提供する。上位レイヤはこの仮想的な配送ネットワーク上に存在するシステムと通信を行う。

このように 2 層の配送ネットワークを用いる理由は、上位レイヤに対して実際の配送ネットワークをそのまま提供する方法を比較した場合以下のような利点があるためである。

- 仮想的な 1 つの宛先になっているが実際には複数の宛先になっているような、1 対多通信を上位レイヤに隠蔽して行える。

- 上記の技術を用いて、配送ネットワーク上のあるシステムを切り放したりあるいは新たに投入するといった、配送ネットワークを構成するシステム数の変化を、上位レイヤに隠蔽したまま行える。
- 配送ネットワークのリンク構成を上位レイヤに隠蔽したまま変更することができる。

ネットワーク状態に応じて動的に配送リンクを再構成することで、データ配送網としての頑健性を提供することができる。

このような利点を活用し、大規模災害時のようなネットワークの激しい輻輳や分断に対して対応する機能をこの層で実現することで、これより上位レイヤは配送リンクの変化に対応することなく、あるいは気付くことなくこのデータ配送ネットワークを利用し続けることができる。

また、配送ネットワーク上のあるシステムのリソースへアクセスした場合に実体が別のリソースにありそこへのポインタが返されるような、リダイレクション動作を上位レイヤ対し隠蔽することができる。

このレイヤより上位のレイヤは、仮想的な配送ネットワークの構成のみを意識することになる。

4.1.3 レコード層

レコード層は、下位層から提供された仮想的な配送ネットワーク上に存在する宛先に対してデータの入出力を行うレイヤである。上位レイヤに対しては、仮想的な宛先に対するデータレコード単位の入出力を提供する。このレイヤの役割は効率的な仮想的宛先へのデータ入出力である。

1レコード単位での通信を実際の配送ネットワーク上で行っていては効率が悪く、結果としてプロトコルオーバーヘッドが大きくなる。そこでこのレイヤでは図6に示すように、上位レイヤから渡されたレコードをある単位にまとめて配送し、また下位レイヤからまとまって渡されたレコード群を1レコード単位に分割して上位レイヤに渡す必要がある。

したがってこのレイヤが上位レイヤに提供するサービスは、仮想的な宛先とレコードの組から構成されるデータの入出力である。

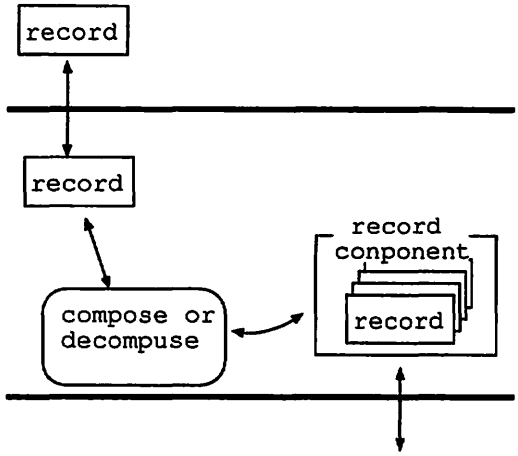

図 6: レコード層の内部構成

4.1.4 レコードコントロール層

レコードコントロール層は、下位レイヤから渡されたレコードをメッセージという形式にして上位レイヤに渡すのがその役割である。

メッセージとはレコードの内容に応じて生成される情報である。メッセージはレコードそのものである場合もあり、例えば上位レイヤがデータベースである場合には、レコードはそのままメッセージとして伝達される。

レコードそのものでない場合のメッセージとしては、上位レイヤに対する何らかのコントロールメッセージが考えられる。例えば、あるデータの削除を要求するレコードがこのレイヤで解釈されて上位レイヤに削除要求という形のメッセージとして伝達される。

レコードコントロール層の役割は、下位レイヤから渡されたレコードの内容を解釈し適切なメッセージを生成することである。

5 議論

本方式の最も単純な構成はリンクコンフィギュレーション層における配送ネットワークと仮想ネットワークが同一で、さらにネットワークを構成するシステムが2つしかない場合である。

このような場合、本方式は RPC[9] のように振る舞うことができる。アプリケーションレイヤは、特定のルールに従ったメッセージを書き込むことで、ネットワークを通じてもう片方にメッセージを伝達することができる。

したがって、アプリケーションレイヤにファイ

ルシステム提供するアプリケーションを用いることで、本機構を NFS[10] のようなファイルの入出力として抽象化することができる。

また、リンクアソシエーション層は下位レイヤを選ばないため、同報通信の要求に対して、利用する下位レイヤとして RM(Reliable Multicast) 網を利用することでより効率の良い同報通信を行う枠組みを提供することができる。

リンクコンフィギュレーション層では、データ配送ネットワーク構成と上層に提供する仮想ネットワーク構成を変えることで、配送ネットワーク上のシステムのクラスタリングによる負荷分散や、グループ化によるデータ配送の効率化などを上位レイヤに意識させないで行うことができる。

この場合に、リンクコンフィギュレーション層やリンクアソシエーション層においてどのようなアルゴリズムを用いてリンクの構成を行うか、あるいはどの程度自動化するかは今後の課題として残されている。

6 展望

今後本稿で提案したデータ配送機構は、本稿で提案したレイヤデザインに基づいてインプリメンテーションデザインをし、それに基づいてインプリメンテーションを作成する。インプリメンテーションは WIDE Project Lifeline 分科会において現在稼働中の IAA システム [3, 4] のトランスポートレイヤとして置き換えて、性能評価等の実験を行い、その際に IAA クラスタの動的な配置やバルクデータの配送を試みる。

7 おわりに

本稿では、はじめに既存の分散システム間のデータ配送機構の特徴を述べた。次に疎結合分散システムに要求される機能を明らかにし、従来のデータ配送機構を適用することが困難であることを指摘した。そこで、従来方式の問題点を解決するための本稿でのアプローチを述べ、4 層からなるデータ配送機構の設計案とその可用性について論じた。

参考文献

[1] Yukimitsu Izawa, Shuji Ishii, Nobuhiko Tada, and Masaya Nakayama. Implementation and evaluation of widely distributed database system using satellite based multicast and netnews system for the transport mechanism. In *Proceedings of Internet Workshop '98(IWS'98)*, pp. 75–83. IEICE and ETL, 1998 年 3 月.

[2] 井澤志充. NetNews を使った信頼性のあるデータ通信の技法. 情報処理学会 分散システム運用技術 研究報告 No.9, pp. 49–54, May 1998.

[3] 多田信彦, 馬場始三, 井澤志充, 丸山太郎, 田中友英, 中山雅哉. インターネットを用いた安否情報システムの構築とその課題. インターネットコンファレンス'98 論文集 (pp.41–50), December 1998.

[4] 多田信彦. IAA システム全体のアーキテクチャについて. 情報処理学会 分散システム運用技術研究会, May 1998.

[5] S. Vinoski. CORBA: Integrating diverse applications within distributed heterogeneous environments. IEEE Communications, vol.14, no. 2, Feb. 1997

[6] The Common Object Request Broker: Architecture and Specification. Revision 2.2 February 1998 Available electronically as http://www.omg.org/library/c2indx.html

[7] Zakaria MAAMAR. Samdw-software agents meet data warehouses,new generation data warehouse technologies. In *IEICE TRANS INF.&SYST., Vol.E82-D, No.2*, pp. 189–198. IEICE, January 1999.

[8] BUDIARTO, Kaname HARUMOTO, Masahiko TSUKAMOTO, and Shojiro NISHIO. On relocation decision policies of mobile database. In *IEICE TRANS INF.&SYST., Vol.E82-D, No.2*, pp. 412–421. IEICE, February 1999.

[9] Sun Microsystems, Inc. RPC: Remote Procedure Call Protocol Specification Version 2. RFC 1057, June 1988.

[10] Sun Microsystems, Inc. NFS: Network File System Protocol Specification. RFC 1094, March 1989.

[11] 植原啓介, 西村篤, 村井純. LWPA:インターネット環境における広域無線通信メディア利用のためのアーキテクチャインターネットコンファレンス'97 論文集, December 1997.