4T－2

Smart Scatter: インターネット／イントラネットの負荷分散機構 - パケット振りわけ部 - *

今井祐二 †岸本光弘 ‡高場浩一 §矢作毅彦 ¶
††（株）富士通研究所‖ §¶（株）富士通

1 はじめに

我々は，大規模ネットワークサーバのためのフロントエンドミドルウェア「Smart Scatter」を開発している．Smart Scatterは，複数のサーバノードにTCP,UDPセッションを分配し，ノード故障時もリクエスト配送先を故障していないノードに切替えることで，大容量／高信頼でかつクライアントから１台に見えるネットワークサーバーをクラスタによって構築することを可能にする．Smart Scatterはカーネル内で実際にパケットの振りわけを行なう「パケット振り分け部」と，振り分け方法を指示する「制御部」からなる．本稿ではこのうちパケット振り分け部に関して述べる．

2 機能

Smart Scatterは，クラスタ全体を表す代表IPアドレス宛のTCP,UDP/IPパケットを，管理者が定義する割合でクラスタ内のノードに分配することができる．Smart Scatterでのパケット振り分け方法を，パターン＆ハッシュ方式と呼んでいる．(図1) 振りわけ部は代表アドレス宛のパケットを指定されたパターンを用いて分類する．パターンはパケットのUDP,TCP,IPヘッダに含まれるIPアドレスとポート番号で表現されている．パターンに合致したパケットは同じくヘッダの情報をパラメータとするハッシュ関数を用いてハッシュ値を計算し，クラスタノードを要素とする宛先ノード表から対応項目を参照し配送ノードを決定する．

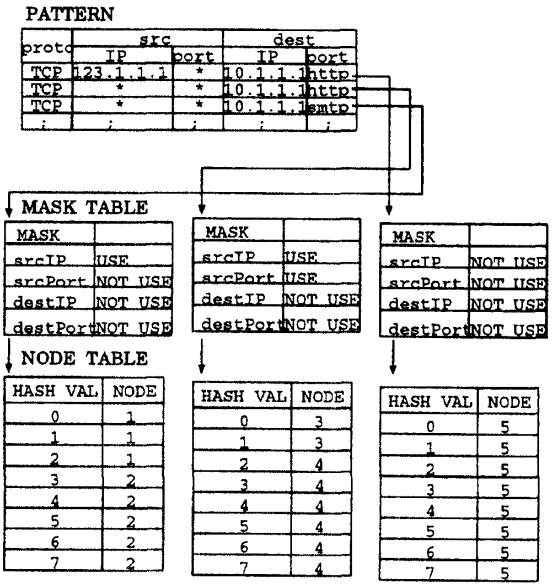

図 1: パターン＆ハッシュ

単一のTCPセッションのデータグラムは宛先IPアドレス／ポート，送り元IPアドレス／ポートが同一なので，ハッシュ値も同一となり，同一ノードへの配送が保証される．配送されたノードでは，ノード固有のIPアドレスの他に代表アドレスを，Logical interface(例：Solaris 2.5.1 における le0:1)等を用いて論理的に設定し，自ノード宛の通常パケットと同様に代表アドレス宛データグラムの処理を行なう．

ハッシュ関数へ渡すヘッダの情報はマスクによって選択が可能で，これにより配分手法を選択できる．例えばデータベース操作をCGIを経由して行なうWebアプリケーションの場合，同一クライアントからの連続したアクセスを常に同一ノードへ収容する事が望まれる．このような場合送り元IPアドレスのみをハッシュ関数引数として使用す

*Smart Scatter -scatter module-

†Yuji Imai
‡Mitsuhiro Kishimoto
§Kouichi Takaba
¶Takehiko Yahagi
‖Fujitsu Laboratories, 4-1-1 Kamiodanaka, Nakahara, Kawasaki, Japan

ることで，連続したhttpセッションが他のノードへ配送される事を防げる．

宛先ノード表はパターン毎に設定が可能で，サービス毎に異なるノードグループを配送先に選んだり，特定クライアントからのリクエストを特別に用意したノードへ配送したりすることが可能である．

クラスタ内に性能に差があるノードを混在させたい場合，宛先ノード表でのノードの出現割合を性能に比例して配分することで調節可能である．また，宛先ノード表の大きさを大きくすることで，割り振りの粒度を所望の程度に調整可能である．

制御部がノードの故障を検出した場合，もしくはノードの予防保守などのために一時的にノードの切り離しを行ないたい場合には，宛先ノード表の切り離しノードの項目を，運転中のノードに書き換える．

予防保守時や振り分け割合変更時に宛先ノード表を書き換えた場合には，接続中のTCPセッションが切断されるおそれがある．これを防止するために，接続中のTCPセッション状態の監視を行なうことができる．宛先ノード表切替時に接続されていたTCPセッションが存在する場合には，後続のパケットが既接続セッションのものかどうかを検査し，既接続セッションの場合には宛先ノード表が指すノードを無視し，切替え前のノードにパケットを配送する．同様に，前述のCGIによるデータベース連係時などの，連続TCPセッションを同一ノードへ配送する必要がある場合のために，クライアントIPアドレス毎の最近のTCPセッション配送先を記録し，割り振り変更後も，以前のノードにパケットを配送することもできる．

3 構成

Scat moduleはUNIX SVR4のSTREAM moduleとしてIP moduleとDLPI driverの中間に挿入される．(図2) 外部ネットワークからのデータグラムはイーサーネットなどによって受信

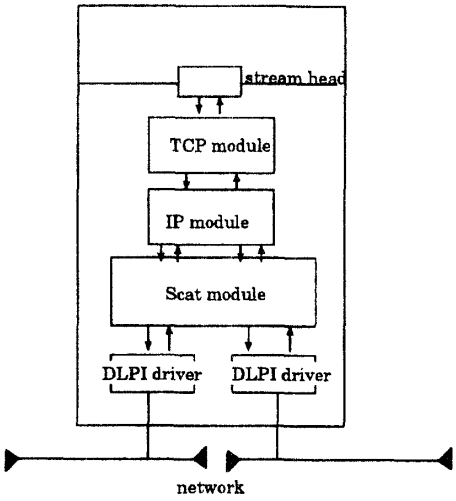

図2: Scat module

され，DLPIドライバから上位モジュールであるscatモジュールが呼び出される．scatモジュールは，データグラム中のIPヘッダを見て，代表アドレス宛のデータグラムであるかどうかを判断する．代表アドレス宛である場合には上記パターン&ハッシュ手法により配送先ノードを決定し，物理層アドレスの解決と書き換えを行なった後，クラスタ内ネットにパケットを送出するためDLPIドライバを起動する．サーバノードからクライアントへのリプライパケットなど，代表IPアドレス宛でない場合には，上位のIPモジュールにパケットを転送する．IPモジュールではルーティング処理が行なわれ，下位のScatモジュールに処理が依頼される．Scat moduleはこのパケットを，そのまま下位のDLPIドライバに渡すので，全体として通常のフォーワーディングが行なわれることになる．

参考文献

[1] 細井他,"Smart Scatter - 全体構成と制御部-", 第55回情報処理全国大会論文集 4T-01, 1997