6T－3

セションリダイレクト機構を用いたエニーキャストゲートウェイの提案と実装

木下真吾　山下博之
NTT 情報通信研究所

1. はじめに

WWW･FTP サーバのフォールトトレランス，負荷分散を目的として，ミラーサーバを設置するサイトが増加している．従来，ユーザは複数のミラーサーバのアドレスを意識し，また稼動状況（故障･負荷状況）などを予測してアクセスする必要があった．

エニーキャストは，同一機能を有するホスト集合のうち"nearest"ホストへのアクセスを保証する通信機構である．"nearest"とは，単に物理的な意味でなく，すばやい応答が得られるなど時間的な意味も含んでいる．"nearest"ホストの決定は，ネットワークによって自動的に行われ，ユーザは，どのホストが"nearest"かを意識する必要がなくなる．

このエニーキャストを上記ミラーサーバ群へのアクセスに適用した場合，ユーザは代表アドレスにアクセスするだけで，"nearest"サーバとのアクセスが可能になる．

ラウンドロビン DNS(以下 rrDNS)は，1 つの名前に対して登録された複数のミラーサーバの IP アドレスをリクエスト毎に順番に回答することで，エニーキャストを実現する．しかし，rrDNS は，常に"nearest"サーバへのアクセスを保証しているわけではない．まず，rrDNS には，故障サーバの検出機能がない．そのため，故障サーバの IP アドレスをユーザに対して回答してしまう可能性がある．また，DNS 情報は，ユーザ側の DNS サーバにキャッシュされる．これを利用するユーザのアクセスが，同じアドレスに対して行われるため，rrDNS の意図した負荷分散が実現できなくなる．

本稿では，フォールトトレランスと効率的な負荷分散機構をもつエニーキャストゲートウェイの提案とその実装方法，性能について報告する．

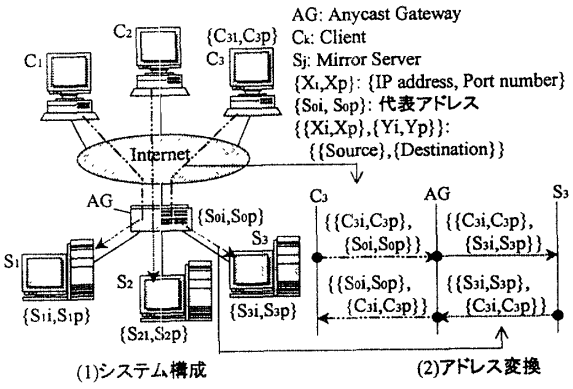

図1 エニーキャストゲートウェイ適用システム

2. エニーキャストゲートウェイ

提案するエニーキャストゲートウェイの適用システム例を図1(1)にを示す．エニーキャストゲートウェイ（以下 AG）は，ミラーサーバ群のフロントエンドに位置し，クライアントからの代表アドレス S_0i に対するアクセスを，S_1, S_2, S_3 のうちの"nearest"サーバへリダイレクトする．

AG は，クライアントからのアクセス毎にミラーサーバ群の稼動状況に応じて"nearest"サーバを決定するため，rrDNS に比べて効果的なフォールトトレランスや負荷分散を実現できる．

3. セションリダイレクト制御

AG におけるセションリダイレクトは，"nearest"サーバ決定処理とパケットのアドレス変換処理によって実現される．

3.1. "nearest"サーバ決定処理

AG は，ミラーサーバの故障状況と負荷状況に応じて"nearest"サーバを決定する．

故障検出には，全ミラーサーバに対して定期的に行うヘルスチェックを用いる．故障が検出されたミラーサーバは"nearest"サーバ候補から除外される．また，故障検出後もヘルスチェックは続けられ，復旧が確認されたミラーサーバは再び"nearest"サーバ候補に加えられる．

負荷の推定には，ミラーサーバ性能と処理中のセション数を考慮する．すなわち，あらかじめ測定したミラーサーバの性能に対するセション数の比率が最小となるものを"nearest"サーバとして選択する．ミラーサーバにおけるセション処理負荷やセション寿命を推定することで，より効率のいい負荷分散を実現できる．しかし，インターネットの不安定な転送レート，処理負荷や寿命予測に要する計算オーバヘッド，人の操作が介在する Telnet や FTP などの寿命予測の困難さなどの問題がある．

3.2. アドレス変換処理

図 1(2)は，1 つの要求パケットとそれに対する 1 つの応答パケットで完了する単純なセションのアドレス変換例である．このような特徴をもつセションの例として NTP，DNS などがあげられる．図の例では，AG は，パケットの宛先を代表アドレス$\{S_0i, S_0p\}$から"nearest"サーバとなった S_3 のアドレス$\{S_0i, S_0p\}$へと変換している．

このアドレス変換によって IP アドレスが変更されるため，AG は，IP ヘッダのチェックサムフィールド，TCP の場合は TCP のチェックサムフィールドを再計算し変更する．

AG はクライアントからの要求を受けた時点（TCP を利用している場合は，SYN フラグで検出する）で，クライアント C_k と AG 自身 S_0 およびリダイレクト先 S_j それぞれの IP アドレス(C_ki, S_0i, S_ji)とポート番号(C_kp, S_0p, S_jp)（これら 6 つの情報の組をセション管理情報：SMI と定義する）とをセション管理テーブル（以下 SMT）に登録し，クライアントへの応答パケット生成時に参照する．

HTTP セションのアドレス変換

HTTP などの TCP を利用するほとんどのアプリケーションプロトコルでは，1 セションで複数の要求･応答パケットがやりとりされる．このようなセションでは，一度リダイレクト先が決定された後，セションが終了する（TCP コネクションが切断される）までの間，後続パケットを同じサーバに到着させる必要がある．AG は，すべてのパケットを監視し，登録済み SMI の$\{\{C_ki, C_kp\}, \{S_0i, S_0p\}\}$とパケットの宛先が一致するかどうかを調べる．一致した場合は，パケットの宛先を SMI の$\{S_ji, S_jp\}$に変更して送出する．

この手法は，1 セションが 1 つの TCP コネクションからなる Telnet などのプロトコルにも適用できる．

"A Proposal and Implementation of the Anycast Gateway using Session-Redirection"
Shingo KINOSHITA, Hiroyuki YAMASHITA
NTT Information and Communication Systems Laboratories

FTP セションのアドレス変換

FTP は，1つのセションが制御用･データ転送用の 2 つのコネクションからなる点で，HTTP などのセションとその制御方法が大きく異なる．2 つのコネクションは，同じサーバとの間で確立される必要がある．

データ転送コネクションに用いるデータ転送用ポート番号の情報は，確立後の制御コネクション上でやりとりされる．その方法は2通りある．1つは，PORT コマンドによるもの（Active モード）で，クライアントが自分の IP アドレスとデータ転送用に用意したポートの番号 $C_k$dp をサーバに通知する．もう 1 つは，PASV コマンドによるもの（Passive モード）で，クライアントからの PASV コマンドに対して，サーバが"227"ステータスコードとあわせて，自分の IP アドレスとデータ転送用に用意したポートの番号 $S_j$dp をクライアントに通知する．これらの情報は，すべて ACSII データで記述されている．

･Active モードのアドレス変換

AG は，制御コネクション上を流れるデータを常に監視する．PORT コマンドを検出すると，制御コネクション用に登録した SMI をコピーし，それを SMT にデータ転送用 SMI として追加する．データ転送用 SMI のクライアントポート番号に PORT コマンドの引数 $C_k$dp を設定（C_kp を上書き）する．その後，サーバからの{C_ki, $C_k$dp}宛てのコネクション確立要求パケットを検出すると，その TCP ヘッダのソースポート番号をデータ転送用 SMI の AG ポート番号，リダイレクト先ポート番号に設定（S_0p, S_jp を上書き）する．この結果，データ転送用 SMI が完成し，データ転送コネクションが制御コネクションと同じリダイレクト先に確立されることになる．

･Passive モードのアドレス変換

"PORT"コマンドの変わりに"227"ステータスコードを検出すること以外は基本的に Active モードと同様の処理を行う．ただし，"227"ステータスコードの引数にはリダイレクト先のサーバ IP アドレス，データ転送用ポート番号{S_ji, $S_j$dp}が設定されているが，それを AG の{S_0I, S_0p}に変更する必要がある．この情報は，ASCII で記述されているため，ユーザデータ長が変化する可能性がある．変化した場合，AG は，その差分を計算し，IP ヘッダの TL フィールドを変更する．また，後続パケットにおいて，TCP ヘッダの ACK，SEQ フィールドが変化するため，これらのフィールドも変更する．

4. 実装

AG は，セションリダイレクト制御を高速化するために，専用ハードウェアとして実装した．WS などを用いて，カーネルにその機能を実装した場合には，パケットのコピーなどのオーバヘッドが大きく，高速化が難しくなる．AG は，ネットワークコントローラチップ，パケットのバッファ用メモリ，SMT 高速検索用 LSI，セションリダイレクト処理ファームウェア，ファームウェア実行用汎用プロセッサで構成される．

さらに，AG は，Cut-Through 方式を採用し，さらなる高速化を図っている．Store-and-Forward 方式のように，パケット全体を受信した後，ヘッダ解析，アドレス変換を行う場合，ユーザデータ部の受信時間が無駄になる．一方，Cut-Through 方式では，パケットのヘッダ部の受信が完了した段階から，解析，アドレス変換を開始できるため，高速化が可能となる．

5. スループット解析と考察

AG のネットワークインタフェースを 10M, 100M Ethernet とした場合の AG のスループットを解析評価した．AG のスループットがネットワークインタフェースのスループットより大きい場合には，セションリダイレクトによるクライアントーサーバ間スループットの低下は起こらないと考えられる．

表 1 に，評価に用いた各セション毎の，Cut-Through 時に必要となる最小データサイズ D_s，ファームウェアのプロセッシング時間 T_fを示す．AG のスループットは，$D_t/(T+D_s/S)$で表される．ここで，D_tはパケットサイズ，Tは AG 内部処理時間，Sはネットワークスピードである．本評価では，T=T_f とした．これは，ネットワークコントローラチップからバッファメモリへのデータ転送時間，SMT 高速検索用 LSI による検索時間などの T_f以外の AG 内部処理時間が，T_fに比べて無視できるほど小さいためである．

表1 セション毎の D_s, T_f値

Data / Session	D_s[Byte] MAC	IP	UDP/TCP	Session	Total	T_f[μs]
NTP･DNS	14	20	8	0	42	78
HTTP	14	20	20	0	54	100
FTP(CTRL)	14	20	20	4～52	55～106	115
FTP(DATA)	14	20	0	0	54	105

図2にパケットサイズとスループットの関係を示す．全体的に，パケットサイズが大きくなるに従い，スループットが向上していることがわかる．10M Ethernet をネットワークインタフェースとした場合，Cut-Through 方式を採用した AG のスループットは，パケットサイズが 200Byte を超える範囲で，全セションともネットワークインタフェースのスループットを超えている．また 100M Ethernet の場合，Cut-Through 方式を採用しても，ほとんどがネットワークインタフェースのスループットを下回っていることがわかる．

特に，HTTP セションでは，頻繁に TCP コネクションの確立，切断が行われることから，小さなサイズのパケット比率が高いと考えられる．これにより AG におけるスループットが低下する可能性がある．高速なネットワークインタフェースを用いた場合は，さらに，その影響は大きくなる．そのため，ファームウェアによる処理時間の削減が必要となる．

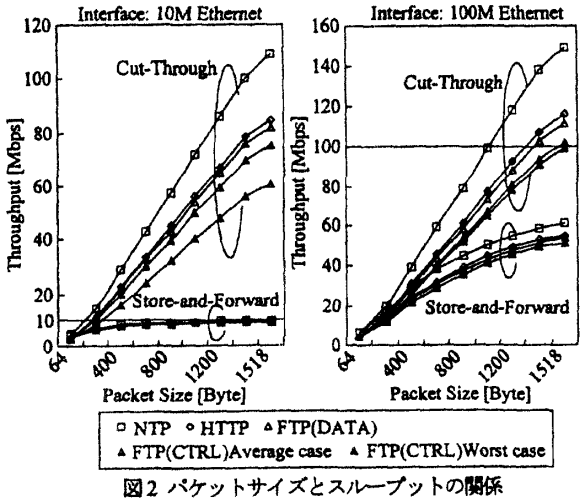

図2 パケットサイズとスループットの関係

6. おわりに

nDNS の問題点を解決するエニーキャストゲートウェイの提案と実装について報告した．各種セションのリダイレクト制御方法の提案と高速化に関する評価および考察を行った．現在 AG は，評価中であり，今後，実システムに適用し，そのスループットや負荷分散効果を測定する予定である．