第 19 部

IX の運用技術

第 1 章

NSPIXP(Network Service Provider Internet eXchange Project)

1.1 NSPIXP-1 の実験修了

商用ネットワーク相互接続実験プロジェクト（NSPIXP）は、商用ネットワークを相互に接続した IX を構築し、この IX を実験環境とた実証実験を行っている。このたび 1999 年 3 月末日をもって、実証実験環境の 1 つである NSPIXP-1 の実験を修了した。NSPIXP-1 での研究成果を用い、引き続き NSPIXP-2 および NSPIXP-3 での実証実験は行っていく。

NSPIXP-1 は、NSPIXP における最初の実証実験環境であり、1995 年に WIDE インターネットの東京オペレーションセンター内（神保町）に構築された。当時は、IX における技術的な問題点や運用上の問題点などがはっきりしない時代であったが、共同研究各社との議論を元に、日本ではじめとの IX として実験を開始することができた。当初 US における先行的な例を参考にしながら、第 3 層（IP 層）による IX として実験をスタートし、その後 IX における経路制御交換ポリシに関する研究の結果、現在主流となる第 2 層（データリンク層）への移行を行った。

NSPIXP-1 は、実験基盤技術としてスイッチ技術を用いた 10Mbps の Ethernet を利用し、約 40 の ISP を相互に接続していた。各 ISP から NSPIXP-1 への回線には、T.1(1.5Mbps) という制限を設けていた。このような制限を設けることにより IX の実験をスロースタートな形ではあるが、実験を立ち上げることができた。しかし、実験を進めるにつれて、より現実的な IX での実験への移行を進めることになり、主な実験基盤を NSPIXP-2 へと移行してきた。主な実験基盤を NSPIXP-2 および NSPIXP-3 へと移す中、NSPIXP-1 では、新たに IX における各種サーバの実験（News サーバの構築、ルートサーバの構築）を行ってきたが、これらの実験も NSPIXP-2 への移行が完了した。従って 1999 年 3 月末日をもって、NSPIXP の最初の実証実験の場である NSPIXP-1 での実験を修了した。

1.2 NSPIXP-2 でのトラフィック

NSPIXP では、IX における主なデータリンク技術として、NSPIXP-2 では、FullDuplex FDDI を、NSPIXP-3 では、Full Duplex FastEthernet を用いているが、これらのデータリンク技術は単方向 100Mbps（双方向 200Mbps）であり、ISP を相互に接続する IX としては、充分な帯域を確保することができなくなってきている。NSPIXP-2 で交換されているトラフィックの変動を図 1.1に示す。このデータが示すように、NSPIXP-2 で交換されているトラフィックは、増加の一途をだどり、現在は、ピークで 850Mbps のトラフィックが交換されている。NSPIXP-2 では、現在 50 の ISP が F ー FDDI(Full duplex FDDI) で相互にトラフィックを交換している。また、すでに 1 ISP が複数の FDDI を用いるなどして帯域を確保しているが、複数のリンクを用いる方法は、ISP 側での運用を複雑にするなどの問題点があると同時に、IX の規模性に対する問題も発生してきている。具体的には、現在 NSPIXP で利用している GigaSwitch と呼ばれる FDDI スイッチで提供可能なポート数の上限を迎えようとしている。

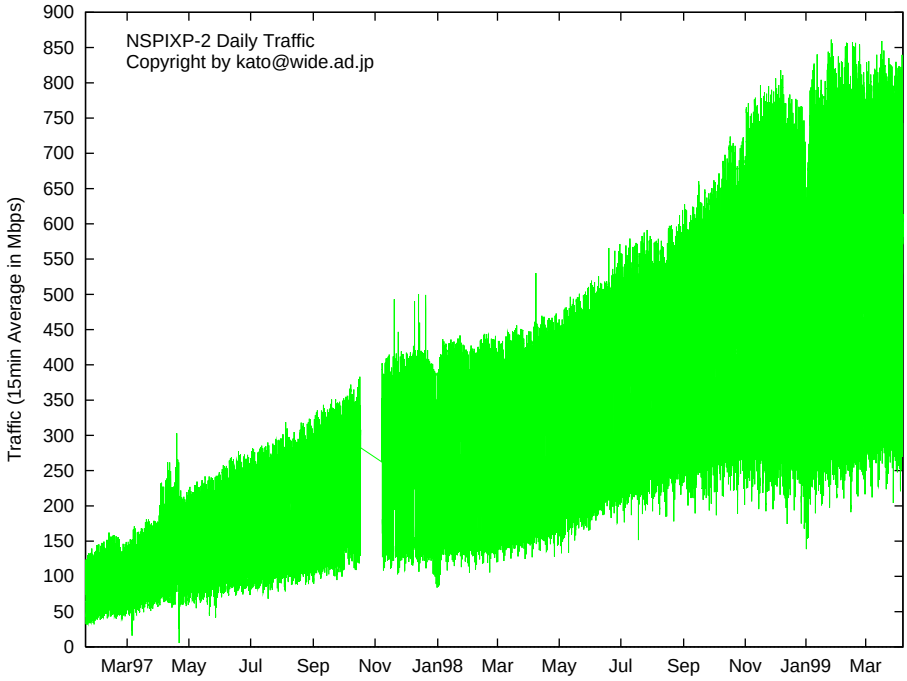

図 1.1: NSPIXP-2 トラフィックの遷移

NSPIXP では、より広帯域で拡張性のある IX を構築するための実験として、このたび Gigabit Ethernet をデータリンク技術として用いた実験を開始することにした。Gigabit Ethernet を用いることにより、ISP と IX 間の接続帯域を 1 Gbps にすることが可能となる。また、Gigabit Ethernet は、現在多くのプロダクトが開発されており、GigaSwitch 以

上のポート数を提供可能なプロダクトも出始めている。本実験は、NSPIXP-2 の機能拡張という形で実験を開始する。本実験は、2000 年の 3 月末日までを実験開始のための準備期間とする。この実験開始のための準備期間における主な研究項目として以下のような課題を解決する。

1. Gigabit Ethernet の相互接続性の検証
2. Gigabit Ethernet の安定性の検証
3. 既存の IX との相互関係を明確にする
4. Gigabit Ethernet の拡張性に関する検討
5. 耐故障性に関する検討

1.3　NSPIXP でのニュースサーバーの運用

NSPIXP1 での IX 運用の研究が始まり、ISP 間での BGP による経路情報の交換が行われるようになったため、IX でのインターネットニュースサーバーの運用を行ってみた。インターネットニュースは、ニュースサーバー間で NNTP による接続を行い、ネットニュースの交換を行う。通常、IX に接続された ISP のサーバーが相互にニュース交換を行うためには、各 ISP のサーバーが自 ISP 以外の ISP とメッシュ状の NNTP 接続を行う必要があり、そのための各サーバーの設定は非常に煩わしいものとなる。しかし IX 近傍にニュースサーバーがあれば、各 ISP ではそのニュースサーバーとスター状の NNTP 接続を行うだけで済むため、各 ISP のサーバーの設定はかなり単純化される。またこのモデルになることで、IX での NNTP によるトラフィックを削減することが可能となる。　通常であれば IX に接続されるのはルータであり、ルーター間が BGP 等によるルーティング交換を行い、ネットニュースなどのアプリケーションサーバーはルータを経由して接続される。しかし本ニュースサーバーは、あえてルータ等を介さずに NSPIXP のセグメントに直接接続を行い、そこでアプリケーションサーバーの運用を行った。　NSPIXP1 で、実際にニュースサーバーの運用を始めたのは、1997 年 5 月の初旬である。このサーバーは、ホスト名を『news.nspixp.wide.ad.jp』といい、この時点でのスペックは次のようなものである。

CPU　Pentium 133MHz を使った AT 互換機

メモリ　128MB

ハードディスク　2.1G を 2 台, 4.3GB を 1 台

OS　FreeBSD 2.2.2

ハードディスクを 3 台用意したのは、2.1G を OS 用とニュースのヒストリー用、4.3GB のものをニュースのスプール用として、システムのスループットを上げる目的であった。 また、ネットニュースのニュースソフトウェアとしては、diablo(http://www.backplane.com/diablo/) のバージョン 1.08 を使用した。当時国内で diablo を利用しているサイトは殆どなく、まったくの未知数での運用開始であった。当時主に使われていた INN のバージョン 1 系のものに比べ、diablo は 5 〜 10 倍程度のパフォーマンスを発揮したようである。

IX でルーターを介さずにアプリケーションサーバーを運用する場合、もっとも問題となるのは、ルーティングの解決である。ルータがあればルーティング設定はルータに任せることができるため、サーバーは本来のアプリケーションの処理だけに専念することができる。またサーバー自体に BGP 等のルーティング処理が可能であるアプリケーション (例えば gated) をインストールして設定することで、直接ルーティング処理を行うこともできる。しかしネットニュースサーバーの場合、実際に通信を行う必要があるのは、ISP のニュースサーバーであり、いわゆる不特定多数のホストとの通信経路についてはあまり重要ではない。そこで gated 等を利用せず、ルーティングについては次の方針で設定を行った。

1. デフォルトの経路は WIDE のルーターへ設定する。
2. NNTP 接続を行う各 ISP のニュースサーバーへのホストルーティングを static に設定する。

つまり基本的に各 ISP のニュースサーバーへ static な経路設定を行い、主なトラフィックである NNTP については最短経路で通信を行い、あまり重要でない他のトラフィック (例えば DNS や SMTP) については、WIDE のルーターを経由するという形となる。ISP のニュースサーバーの IP アドレスが変わった場合、static な設定を行っているとその変化に対処できないという問題があるが、実際にはそのような IP アドレスの変更はほとんどなかったため、充分実用的な運用が可能であった。運用開始後は充分なパフォーマンスを示していたが、1997 年 8 月ぐらいになると、ISP 側のサーバーが diablo であっても、nspixp 側からのニュースの配送遅延が目立つようになった。vmstat などで観測すると CPU 的には常時 30-40 るように見うけられたが、9 月上旬に、思い切って CPU を PentiumPro の 200MHz へ交換した。この変更によって、そのような配送遅延はなくなった。この頃は 18 のサーバーと NNTP の交換を行い、1 日当たりの NNTP トラフィックは、incoming が 50-70 万通でボリュームにして 6-8G バイト程度、outgoing は 400-600 万通で、40G バイト程度であった。

この後メモリサイズを 256M バイトに増設するなどの増強を行い、diablo は適宜バージョンアップを行っていたが、incoming のトラフィックが増えると逆に outgoing のトラフィックが減るといった現象が見られるようになった。これは主にスプールに使っていたハードディスクの I/O 性能が足りなくなってきたためであった。またこのころになると incoming のトラフィックは 1 日あたり 10GB を越えるようになり、NSPIXP1 と ISP 間の回線の限界であった T1 回線程度では、容量不足という問題も起こってきた。

これらを解決するため、1998 年 3 月に、すでに運用が行われていた NSPIXP2 へニュースサーバーを用意することになった。当初 news.nspixp.wide.ad.jp をそのまま移設することも考えたが、ハードディスクの構成がそのままだと能力不足であるのは明らかなので、改めて、新規のマシンを用意することになった。このサーバーのホスト名は『news.nspixp2.wide.ad.jp』であり、その構成は次のようになっている。

CPU Pentium-II 333MHz を利用した AT 互換機

メモリ 384MHz

ハードディスク 4.3GB を 1 台, 9GB を 2 台

OS FreeBSD 2.2.5

CPU は当時最新のものであるが、速いものを望んで選んだというよりは、発熱が少ないという点を重視して選んだものである。メモリは、使用したマザーボードの限界まで搭載し、OS のディスクのバッファリング効果を最大に活用できるように構成した。またハードディスクは、4.3G のものを OS とニュースのヒストリー用にし、スプール用には、9GB の HD を 2 台使ってストライプを組むことで、I/O 性能を上げるように構成した。またネットワークインターフェースには、NSPIXP2 の GigaSwitch に直接接続できるよう、FDDI インターフェースを用意した。

NSPIXP2 のサーバーでは NSPIXP1 のサーバーで設定していたデフォルトルーティングの設定を止め、必要なサーバーへのホストルーティングを static に設定するだけにした。デフォルトを設定しないことで、ISP 側のサーバの IP アドレスが変わったときなどの場合に、不用意な経路 (特に WIDE の国際回線) にネットニュースの膨大なトラフィックがかかるのを防ぐことができるからである。もちろん、任意のホストと通信できないという問題があるが、ルーティング可能なホストを経由するなどして、現実はあまり問題にならなかった。

実際に運用してみたところ、news.nspixp2.wide.ad.jp は非常に良好なパフォーマンスを示し、現在では 30 サイト以上と NNTP 交換を行っている。この間ネットニュースのトラフィックは徐々に増え、news.nspixp2.wide.ad.jp での 1 日あたりのトラフィックは、次のように変化してきている。

	incoming		outgoing	
1998 年　7 月	55-85 万通	13-16GB	400- 600 万通	100-120GB
1998 年　9 月	50-90 万通	14-17GB	600-1000 万通	180-200GB
1998 年　12 月	55-80 万通	18-23GB	800-1000 万通	270-330GB
1999 年　4 月	55-60 万通	18-25GB	100-1300 万通	360-500GB

incoming の記事数があまり変わらないのに、ボリュームが増えているのは、巨大な記事の割合が増えていることに起因しているようであるが、これについてはあまり調べていな

い。outgoing が増えているのは、記事のボリュームの増加と、NNTP 接続のサイトが増えたことによる。

この数字からわかるように、現在 news.nspixp2.wide.ad.jp は平均で FDDI の 50 帯域を消費している。もちろんピーク時は FDDI がほぼ埋まるほどの膨大なトラフィックを生成しているのである。

多くの ISP では、news.nspixp2.wide.ad.jp からほとんどのニューストラフィックを受けるような形になってきている。news.nspixp2.wide.ad.jp の場合、定常時は特にニュースの配送が滞ることは無い。しかし、ISP で diablo を採用するところが増えたため、同じニュースシステムを使うことによる問題が発生することがわかっている。

diablo のデフォルトの設定では、月曜の早朝と木曜の早朝に、ニュースのヒストリファイルのトリミング処理を行い、ほんの 10 分程度であるが diablo はニュースの受信を停止する。このため news.nspixp2.wide.ad.jp から見ると、多くのサイトへネットニュースの送信ができない時間ができ、この間にネットニュースの incoming と outgoing のバランスが崩れ、メモリのバッファリングが破綻を起こす。この状態になると、ディスク I/O は通常運用時の数倍の負荷がかかるようになり、ニュースサーバーとして本来の性能が発揮できなくなって、そのままでは通常の運用状態に戻れなくなる。このとき outgoing のトラフィックは 300GB 程度まで減るが、配るより先に、expire によってニュースの記事本体が削除されてしまうからである。つまりスプールのディスク I/O の性能が、500GB 近い outgoing のトラフィックを支えられないことを示している。現在はヒストリのトリミングが終わるぐらいの時間に、強制的に配送キューを捨てるという方法でこの問題が起きないようにしている。少量の記事を捨ててでも、大多数の記事を配るべき配慮である。

定常状態であればマシンの負荷を示す load average は 2-5 程度であり CPU の空きは、10-60 非常に安定したニュースサーバーであるといえる。しかしながら、前述の問題があるため、今後はさらなるディスク I/O の性能を稼げるようなマシンに置き換えることを検討中である。