

# **第 XXIV 部**

## **IX の運用技術**



## 第 24 部

### IX の運用技術

#### 第 1 章 はじめに

NSPIXP ワーキンググループでは、商用インターネットを相互に接続する場合の問題点について、実証的な手法で研究をおこなうことを目的として発足し、1994 年の NSPIXP-1 運用開始、1996 年の NSPIXP-2 運用開始、1997 年の NSPIXP-3 運用開始を経て、現在では東京に分散配置した DIX-IE、大阪に分散配置した NSPIXP-3、IPv6 に特化した NSPIXP-6 の運用を基盤とした実証実験をおこなっている。

特に、Internet eXchange Point を物理的に配置した環境での ISP 間のトラフィック交換において、高信頼性および高効率性を考慮した上での分散ネットワークアーキテクチャに着目し、メディアおよび制御技術の実証および展開の検討と議論をおこなっている。現在、今後のトラフィックボリュームを考えた新規ネットワークポロジの検討を終え、移行に向けた試験をおこなっている。

本報告書では、「商用インターネットの相互接続に関する研究」の 2007 年度の活動を

第 2 章. DIX-IE/NSPIXP-3/NSPIXP-6 の現状

第 3 章. 運用関連情報

第 4 章. 次期 DIX-IE アーキテクチャと移行準備としてまとめる。

#### 第 2 章 DIX-IE/NSPIXP-3/NSPIXP-6 の現状

##### 接続拠点一覧

2007 年 12 月時点での DIX-IE/NSPIXP-3/NSPIXP-6 の接続拠点を表 2.1 に示す。

表 2.1. DIX-IE/NSPIXP-3/NSPIXP-6 の接続拠点一覧

|          |                                    |
|----------|------------------------------------|
| DIX-IE   | KDDI 大手町 (WIDE)                    |
|          | NTT 大手町 (NTTCom)                   |
|          | NF 西大井 (MIND)                      |
|          | ComSpace1 (SBIDC)                  |
|          | @Tokyo (KDDI → @Tokyo)             |
| NSPIXP-3 | NTT 堂島 (WIDE)                      |
|          | SBIDC 福島 (WIDE)                    |
|          | K-Opt 湊町 (WIDE)                    |
| NSPIXP-6 | KDDI 大手町 (WIDE)                    |
|          | NTT 大手町 (WIDE)                     |
|          | NTT 堂島 (WIDE)                      |
|          | Bangkok Thai Tower, TH (Thai Sarn) |
|          | Global Switch, SG (SingAREN)       |

##### DIX-IE 接続組織数

2007 年 5 月時点での DIX-IE の接続組織数を表 2.2 に示す。

表 2.2. DIX-IE 接続組織数

|        |              |         |
|--------|--------------|---------|
| 接続組織数  |              | 68 (−1) |
| 接続ポート数 |              | 94      |
| メディア別  | 10GbE        | 9       |
|        | 1GbE         | 58 (−2) |
|        | FastEthernet | 24 (+1) |
|        | Ethernet     | 2       |
| 分散拠点別  | KDDI         | 61 (+2) |
|        | NTT Com      | 18 (−2) |
|        | MIND         | 5       |
|        | SBIDC        | 7 (−2)  |
|        | @Tokyo       | 2 (+1)  |

カッコ内は前年比

### DIX-IE のトラフィックの推移（総トラフィック量）

DIX-IE の総トラフィックの推移を図 2.1 に示す。

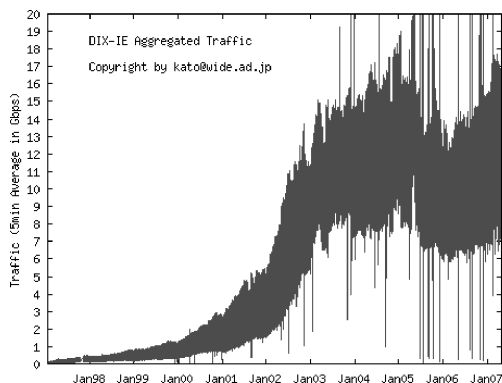


図 2.1. DIX-IE の総トラフィック量の推移

### DIX-IE のトラフィックの推移（総トラフィック量、ログスケール）

DIX-IE の総トラフィックの推移（ログスケール）を図 2.2 に示す。

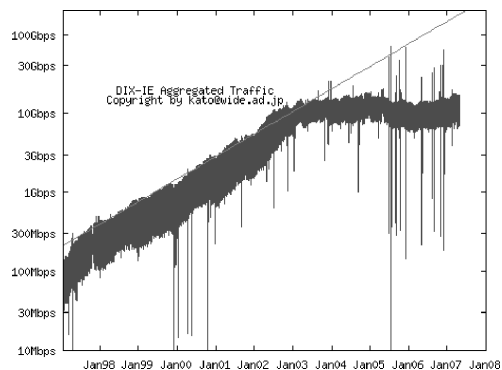


図 2.2. DIX-IE の総トラフィック量（ログスケール）の推移

### DIX-IE のトラフィックの推移（日単位）

DIX-IE の総トラフィックの推移（日単位）を図 2.3 に示す。

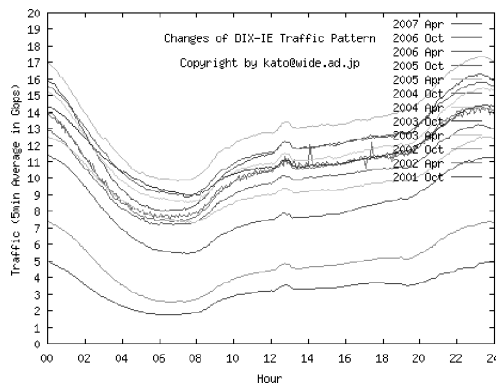


図 2.3. DIX-IE の総トラフィック量の推移（日単位、2001～2007 年）

### DIX-IE のトラフィックの推移（週単位）

DIX-IE の総トラフィックの推移（週単位）を図 2.4 および 2.5 に示す。

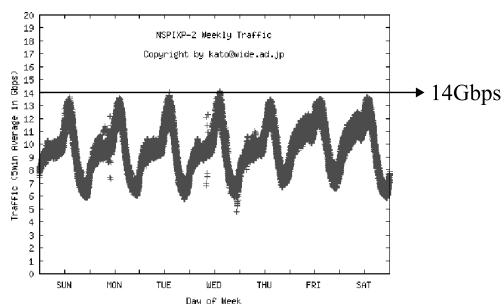


図 2.4. DIX-IE の総トラフィック量の推移（週単位、2006 年 5 月）

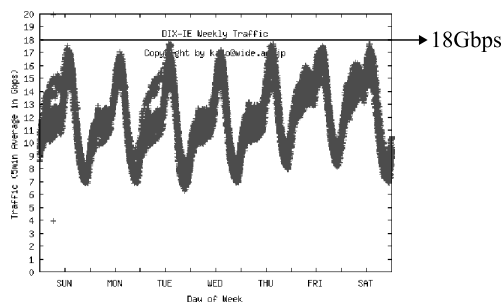


図 2.5. DIX-IE の総トラフィック量の推移（週単位、2007 年 5 月）

### NSPIX-3 接続組織数

2007 年 9 月時点での NSPIX-3 の接続組織数を表 2.3 に示す。

表 2.3. NSPIXP-3 接続組織数

|        |              |         |
|--------|--------------|---------|
| 接続組織数  |              | 19 (-2) |
| 接続ポート数 |              | 25 (-1) |
| メディア別  | 1GbE         | 23 (+1) |
|        | FastEthernet | 2 (-2)  |
| 分散拠点別  | NTT 堂島       | 22 (-1) |
|        | SBIDC 福島     | 0       |
|        | K-Opt 湊町     | 3       |

カッコ内は前年比

### NSPIXP-3 のトラフィックの推移 (総トラフィック量)

NSPIXP-3 の総トラフィックの推移を図 2.6 に示す。

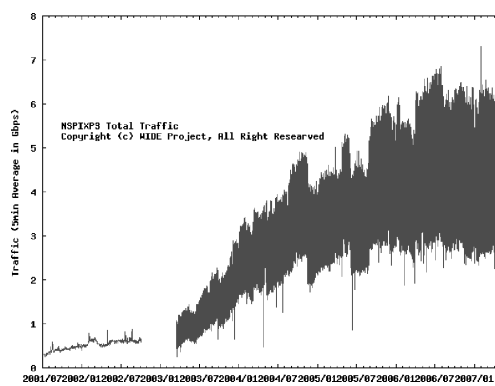


図 2.6. NSPIXP-3 の総トラフィック量の推移

### NSPIXP-3 のトラフィックの推移 (総トラフィック量、ログスケール)

NSPIXP-3 の総トラフィックの推移 (ログスケール) を図 2.7 に示す。

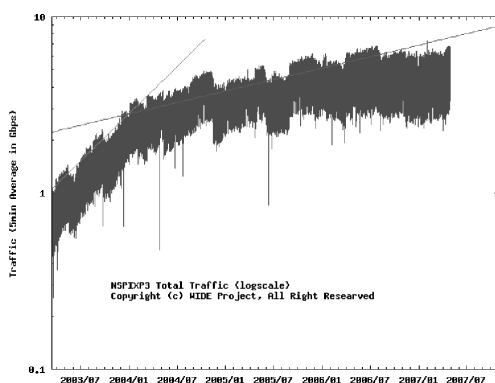


図 2.7. NSPIXP-3 の総トラフィック量 (ログスケール) の推移

### NSPIXP-3 のトラフィックの推移 (日単位)

NSPIXP-3 の総トラフィックの推移 (日単位) を図 2.8 および 2.9 に示す。

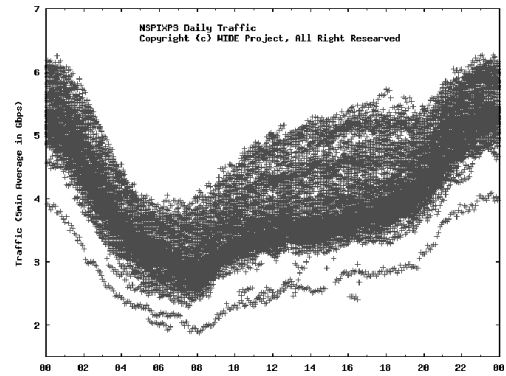


図 2.8. NSPIXP-3 の総トラフィック量の推移 (日単位、2006 年 5 月)

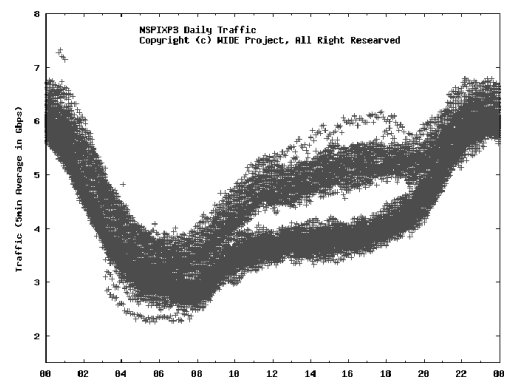


図 2.9. NSPIXP-3 の総トラフィック量の推移 (日単位、2007 年 5 月)

### NSPIXP-3 のトラフィックの推移 (週単位)

NSPIXP-3 の総トラフィックの推移 (週単位) を図 2.10 および 2.11 に示す。

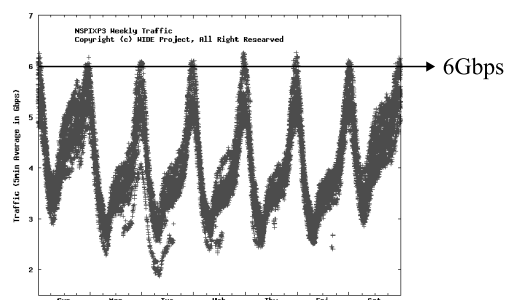


図 2.10. NSPIXP-3 の総トラフィック量の推移 (週単位、2006 年 5 月)

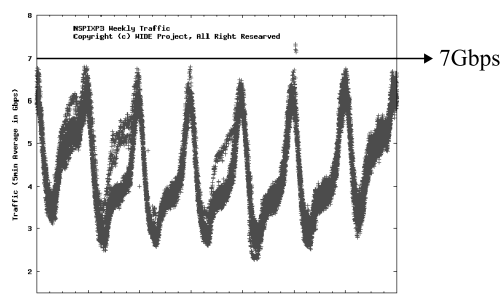


図 2.11. NSPIXP-3 の総トラフィック量の推移（週単位、2007 年 5 月）

### 第3章 運用関連情報

#### NSPIXP-3、NSPIXP-6

NSPIXP-3 の IDC 福島拠点の収用ユーザがいいため、廃止・移転を含めて検討をおこなっている。現状、特に廃止への反対や移転への要望がないため、廃止予定である。

また、NSPIXP-6 については、DIX-IE および NSPIXP-3 の IPv4/v6 デュアルスタック化は完了しており、その役割を終えたといえる。現在、2008 年 6 月 6 日をもって運用終了を予定している。

#### 運用に関する議論

スイッチ障害時の対応として、スイッチ間リンク不良時の運用として、どうあるべきかに関して、1) 帯域減：そのまま（あふれる）、2) 帯域減：リンクダウンさせる、3) 同じ帯域：迂回路への切替、のそれぞれについて議論をおこなっている。

4Octets-AS 番号については、Peer 経路として流れてきた場合に対応できないルータがあることをふまえて議論をおこなっている。

また、KDDI 研究所が、4Octets-AS の実験のため、以下の項目を実行することが報告され、問題があった場合の対応などの議論をおこなった。

- 2Octets AS とのインターオペラビリティのテスト
- 4Octets AS の内部に PlanetLab ノードを設置し、オーバレイネットワークのトラフィックの輻輳に対するトラフィック分散の制御方式について研究

- AS7667 をトランジットし、ftp.kddilabs.jp のトラフィック制御について研究

4Octets-AS 番号は AS2.6 であり、経路交換ポリシーとしては、AS-path ^2.6\$ および ^2.6\_7667\$ を IPv4 でアナウンスする。

BFD (Bidirectional forwarding detection) について、DIX-IE および NSPIXP-3 としての見解として、基本的に関与せず Peer 同士で自由におこなうものとする事が確認された。

Flow Collect について、公開に際するデータの取り扱いについては細心の注意を払うこととし、sFlow によるデータ収集をおこなう事が確認された。

### 第4章 次期 DIX-IE アーキテクチャと移行準備

#### 次期アーキテクチャの選定

前年度より、NSPIXP 内での議論や各候補メーカー開発部署との議論を続け、Foundry Networks 社の NetIron MLX series を採用した。構成は、コアとして MLX-16 を 2 台配置し、拠点に MLX-4 あるいは 8 を配置する方針とし、拠点側装置については、マネージメントモジュールの冗長をすれば 1 台でも良いこととした。選定理由を以下に示す。

- 転送メモリとして TCAM を使用している事による debug のしやすさ
- Packet プロセッサとして FPGA を使用している事による機能拡張性
- EtherOAM への対応
- IEEE802.1ad Provider Bridge (Q in Q) への対応
- IEEE802.1ah Provider Backbone Bridging (MAC in MAC) への対応

今後は、次期アーキテクチャへの移行を考え、事前試験をおこないながら運用切替について議論をおこなっていく予定である。

#### 次期アーキテクチャへの移行準備

新アーキテクチャへの移行準備として、選定された機器を用いた各種試験を 8 月 8 日におこなった。この試験では、主に障害発生時における冗長経路への切り替わり、ならびに冗長モジュールへの切り替

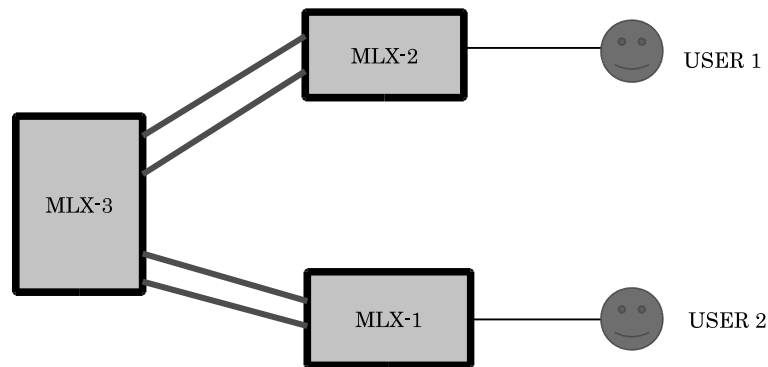


図 4.1. 次期アーキテクチャ試験トポロジ

わりに関する試験をおこなった。

試験には、3 台の Foundry NetIron MLX series を用いておこなった。試験トポロジを図 4.1 に示す。図中の太線は 10 Gigabit Ethernet (10Gbase-LR)、図中の細線は 1 Gigabit Ethernet (1000baseT) を示す。

図 4.1 中の MLX-3 は Foundry Networks 社製の MLX-16 であり、MLX-1、MLX-2 は Foundry Networks 社製の MLX-8 を用いた。どの MLX 筐体もマネージメントモジュールを 2 枚搭載し、管理制御システムは冗長化構成となっている。また、SFM (Switching Fabric Module) も 3 枚搭載することにより、スイッチ内部の packets 転送システムも冗長化された構成となっている。

MLX-1、MLX-2 に疑似ユーザを接続し、USER 1、USER 2 が間隔 0.01 秒の ping を打ち合うことにより、疑似障害発生時におけるパケットロスの傾向を測定した。

なお、図中にて MLX 同士を結ぶ 10 Gigabit Ethernet は 802.1ad を用いたリンクアグリゲーションによって構成されており、障害を検知し冗長経路に切り替えるためのプロトコルとして、802.1w を用いて構成されている。

試験は以下の項目についておこなった。

- a) 障害により装置自身が再起動した場合
- b) ファームウェア更新時における hitless-reload 機構の動作
- c) 障害により一部モジュールが再起動した場合
- d) 障害により SFM (Switching Fabric Module) が故障した場合
- e) 障害によりスタンバイマネージメントモジュールが故障した場合
- f) 障害によりアクティブマネージメントモジュールが故障した場合

ルが故障した場合

- g) 障害により両方のマネージメントモジュールが故障した場合

なお、すべてのテスト項目は、経路冗長化に用いている 802.1w プロトコルのパスコストを調整することにより、正常時は

USER1 → MLX-2 → MLX-3 → MLX-1 → USER2 という経路にて通信が行われるよう設定した。

各項目の試験結果を以下に示す。

#### a) 障害により装置自身が再起動した場合

MLX-3 に障害が発生したと仮定し、MLX-3 の電源ケーブルを引き抜き電源を落とし、電源を再投入することによって擬似的な障害を発生させた。その場合の USER1 から USER2 への ping 結果を図 4.2 に示す。

グラフの縦軸は RTT (Round Trip Time) を対数軸で示したものである。横軸は ping 開始時からの経過時間を示す。濃色の楕円と淡色の楕円に囲まれた部分に大きな RTT の乱れが観測された。濃色の楕円の場所が、MLX-3 の電源が落ちた部分であり、淡色の楕円の部分が、MLX-3 が再起動した部分である。それぞれ、802.1w プロトコルにより経路が切り替わったことを示す。濃色の楕円において、USER1 から USER2 への通信経路は

USER1 → MLX-2 → MLX-1 → USER2

という経路に切り替わり、淡色の楕円の部分にて元の正常時の経路に戻ったことが確認された。

#### b) ファームウェア更新時における hitless-reload 機構の動作

マネージメントモジュールを冗長化した際に利用できる、hitless-reload 機構の動作テストを行った。

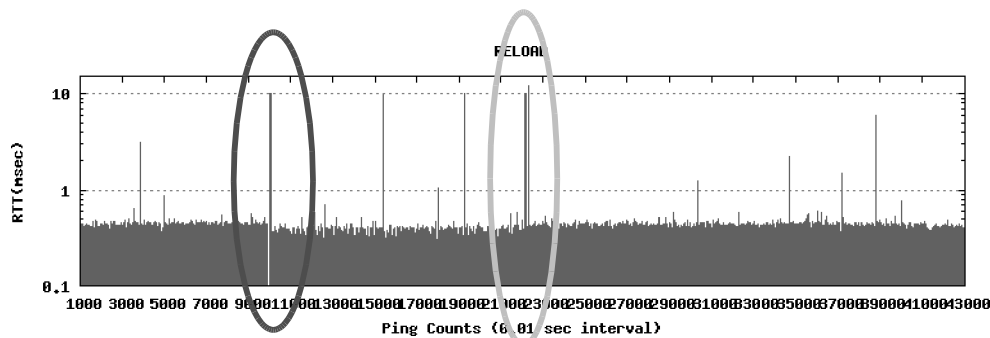


図 4.2. 試験項目 a) の ping 結果

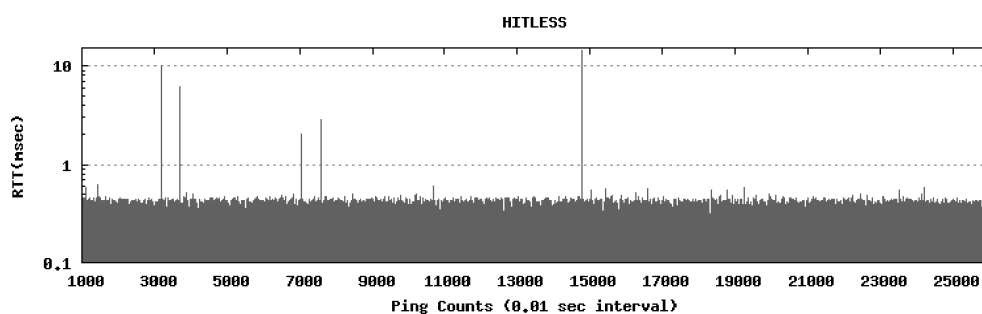


図 4.3. 試験項目 b) の ping 結果

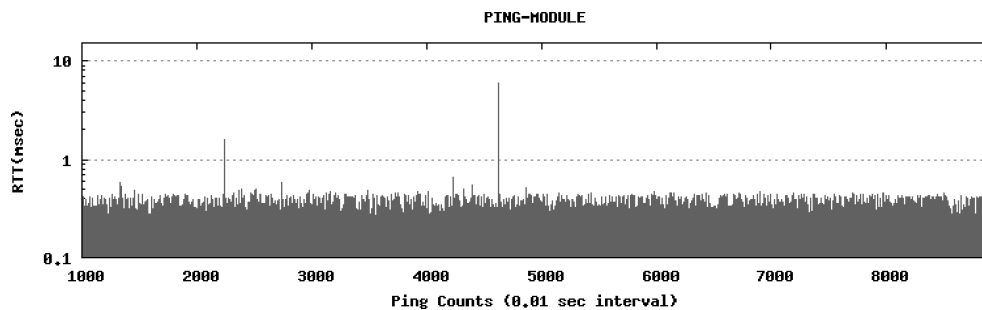


図 4.4. 試験項目 c) の ping 結果

これは、スイッチのファームウェア更新時に、スイッチ自体を再起動することなくファームウェア更新が行える機構である。MLX-3 のアクティブマネジメントモジュールにて hitless-reload コマンドを投入することによりテストを実施した。その場合の USER1 から USER2 への ping 結果を図 4.3 に示す。

グラフが示すとおり、大きな乱れは発生しておらず、ping のタイムアウトを示す空白もグラフ上に存在しない。hitless-reload 機構により、通信に乱れを発生させることなくファームウェア更新がおこなえたことを示している。

#### c) 障害により一部モジュールが再起動した場合

障害により MLX-3 のモジュールカードが再起動

した場合を想定した試験をおこなった。MLX-3 において、通信に利用していない 10 Gigabit Ethernet モジュールカードを、動作中に物理的に引き抜くことにより実施した。その場合の USER1 から USER2 への ping 結果を図 4.4 に示す。

グラフが示すとおり、ping 結果に大きな乱れやタイムアウトは発生していない。これはモジュール障害が発生した場合も、問題なく通信が行われていることを示している。

#### d) 障害により SFM (Switching Fabric Module) が故障した場合

障害により MLX-3 の SFM が故障した場合を想定した試験をおこなった。MLX-3 の SFM モジュー



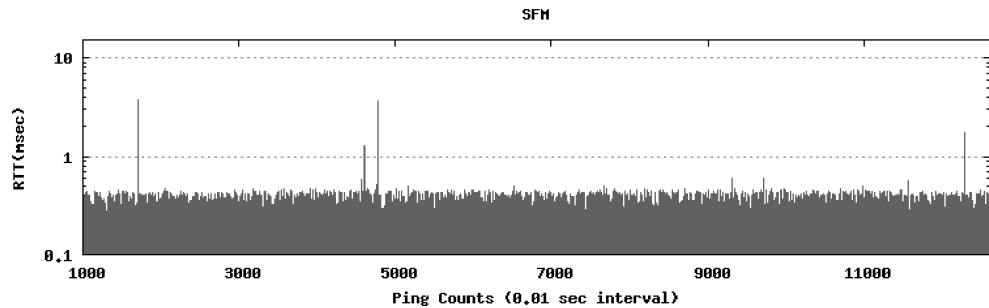


図 4.5. 試験項目 d) の ping 結果

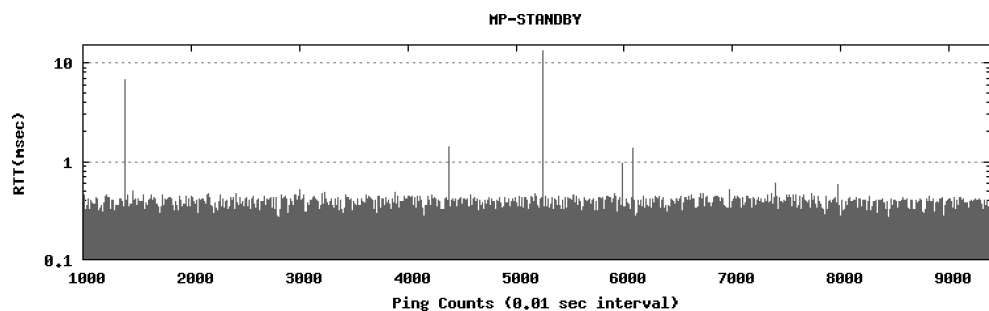


図 4.6. 試験項目 e) の ping 結果

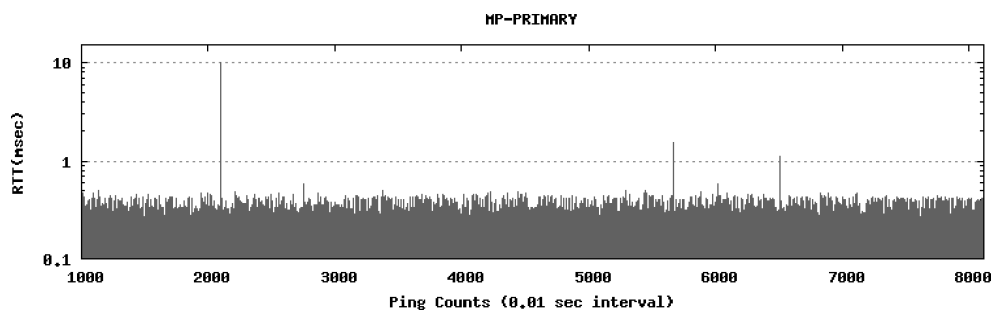


図 4.7. 試験項目 f) の ping 結果

ルを動作中に物理的に引き抜くことにより実施した。その場合の USER1 から USER2 への ping 結果を図 4.5 に示す。

グラフが示す通り、ping 結果に大きな乱れは発生していない。これは、SFM が故障した場合にも、通信に影響することなく代替 SFM に切り替わることを示している。

#### e) 障害によりスタンバイマネージメントモジュールが故障した場合

障害によって、MLX-3 に搭載されているスタンバイマネージメントモジュールが故障した場合を想定した試験をおこなった。MLX-3 のスタンバイマネージメントモジュールを、動作中に引き抜くことによって実施した。その場合の USER1 から USER2 への

ping 結果を図 4.6 に示す。

グラフが示す通り、ping 結果に大きな乱れは発生していない。これは、スタンバイマネージメントモジュールの故障が通信に影響を与えないことを示している。

#### f) 障害によりアクティブマネージメントモジュールが故障した場合

障害により、MLX-3 に搭載されたアクティブマネージメントモジュールが故障した場合を想定した試験をおこなった。MLX-3 に搭載され、アクティブマネージメントとして機能しているマネージメントモジュールを動作中に引き抜くことによって実施した。その場合の USER1 から USER2 への ping 結果を図 4.7 に示す。

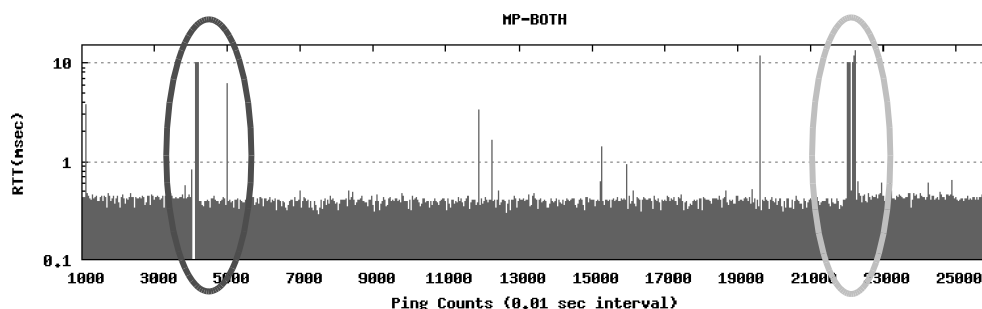


図 4.8. 試験項目 g) の ping 結果

グラフが示す通り、ping 結果に大きな乱れは発生していない。これは、通信中にアクティブマネージメントモジュールに障害が発生しても、通信に影響を与えないことを示す。

#### g) 障害により両方のマネージメントモジュールが故障した場合

障害により、MLX-3 に搭載された両方のマネージメントモジュールが故障した場合を想定した試験をおこなった。MLX-3 に搭載されたマネージメントモジュールを、2 枚とも動作中に引き抜くことにより実施した。その場合の USER1 から USER2 への ping 結果を図 4.8 に示す。

グラフにて 2 箇所、通信の乱れた箇所が観測された。濃色の楕円の部分は、マネージメントモジュールを 2 枚とも引き抜いた部分であり、淡色の楕円の部分は、マネージメントモジュールを再度挿入することによって、MLX-3 が再起動してきた部分である。それぞれ、802.1w プロトコルにより経路が切り替わったことを示す。濃色の楕円において、USER1 から USER2 への通信経路は

USER1 → MLX-2 → MLX-1 → USER2

という経路に切り替わり、淡色の楕円の部分にて元の正常時の経路に戻ったことが確認された。

#### まとめ

以上の試験により、Foundry Networks 社製の MLX シリーズを用いた DIX-IE 新アーキテクチャは、障害時の冗長化に関して期待している機能を実現可能であることが確認された。