PCクラスタにおけるVLANイーサネットのトポロジの評価

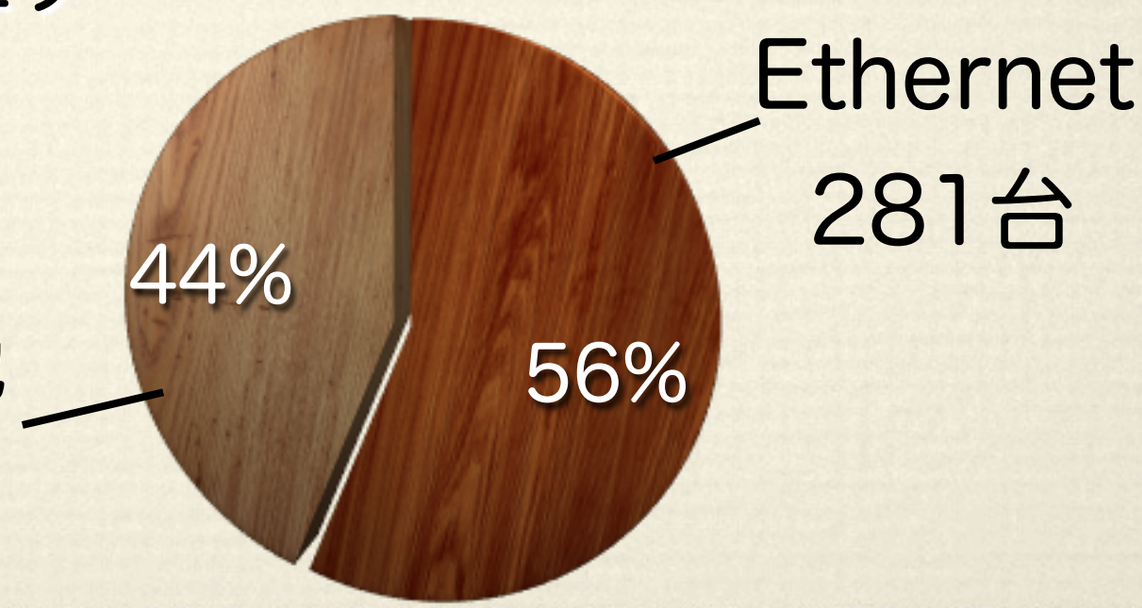
廣安知之[1]，渡辺崇文[2]，○中尾昌広[3]
大塚智宏[4]，鯉渕道紘[5]

1. 同志社大学 生命医科学部
2. 株式会社 インターネットイニシアティブ
3. 同志社大学 工学研究科 博士課程3年
4. 慶應義塾 インフォメーションテクノロジーセンタ
5. 国立情報学研究所 / 総合研究大学院大学

研究背景

- PCクラスタではEthernetが広く用いられている
  - 安価
  - コモディティ製品
  - 10GBASE-Tの標準化

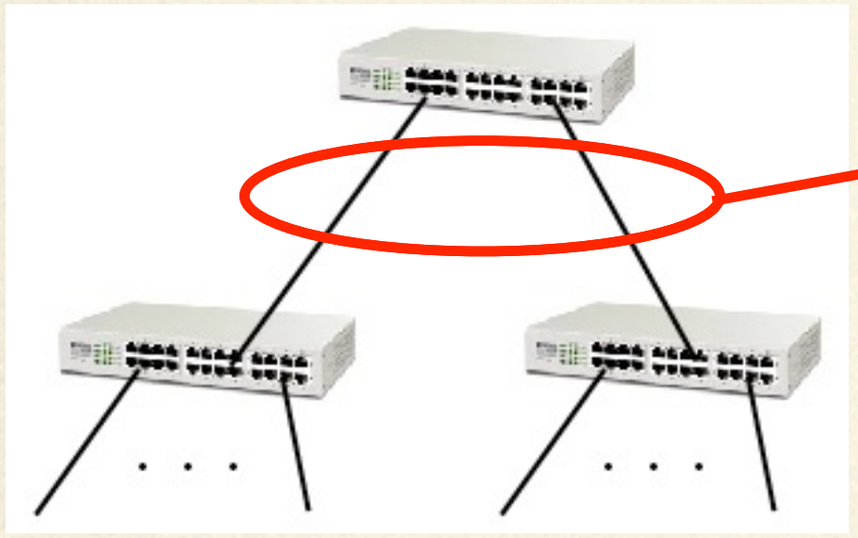
- Top500のEthernetのシェア（2008年11月）



Ethernetの問題点

- ネットワークトポロジに制限がある

  Ethernetはツリートポロジしか作成できない

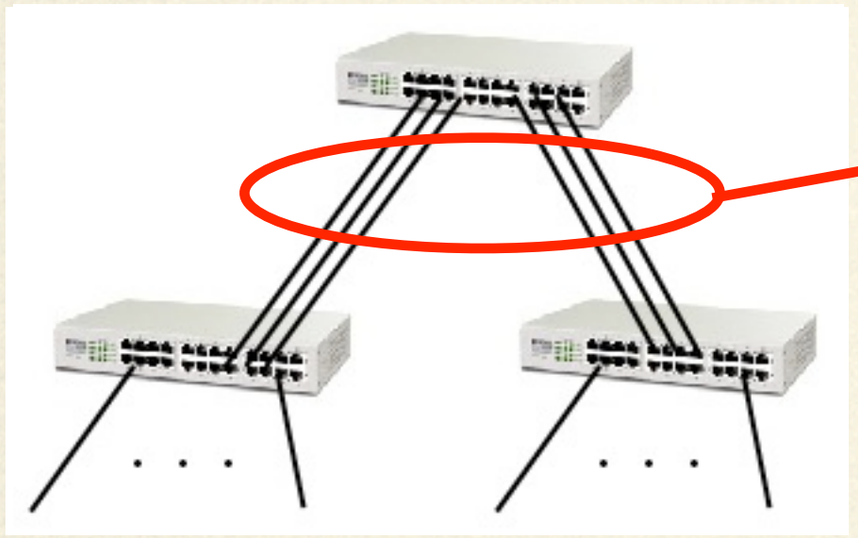


InfinibandやMyrinetなどの高速インターコネクトは
様々なトポロジを作成できるが，高価である
（Ethernetとのコスト比は数倍から10倍程度）

Ethernetの問題点

- ネットワークトポロジに制限がある

  Ethernetはツリートポロジしか作成できない



トラフィックが集中

リンクアグリゲーションで緩和できるが・・・

InfinibandやMyrinetなどの高速インターコネクトは様々なトポロジを作成できるが，高価である

（Ethernetとのコスト比は数倍から10倍程度）

研究目的

EthernetにおいてVLAN技術を応用することで,
PCクラスタ上で様々なネットワークトポロジを構成し,
その影響を明示する

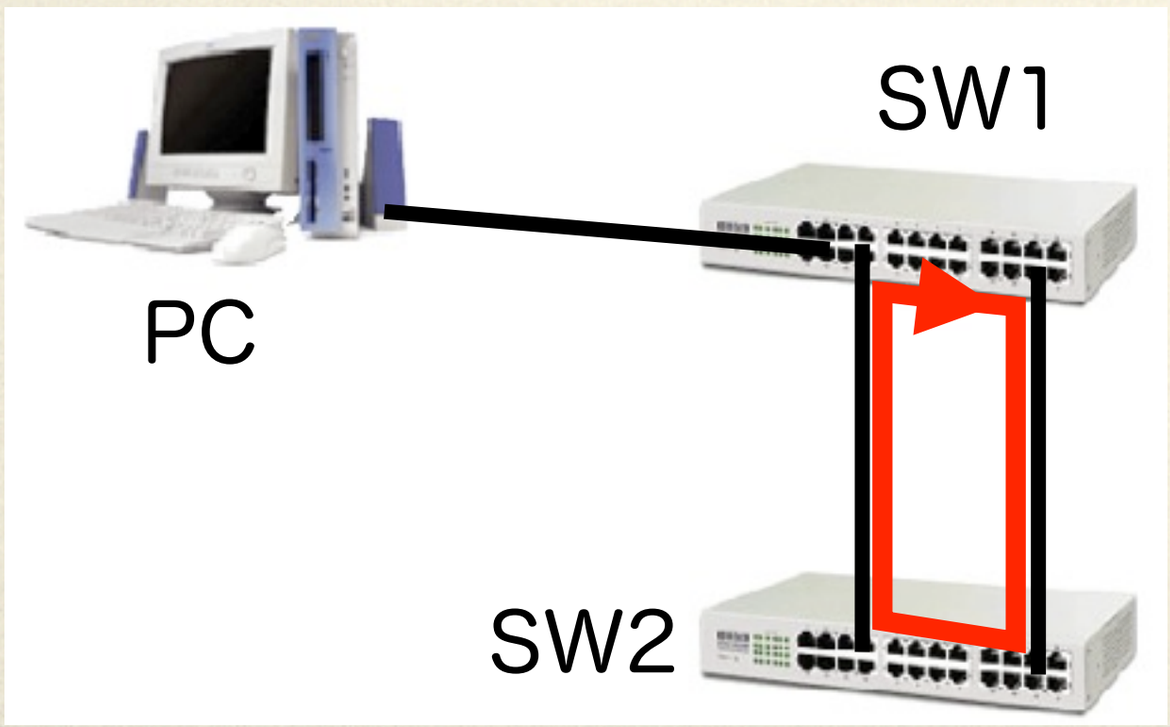
- 200ホスト以上の環境でも，簡易に管理できる
  ネットワークトポロジの実装方法
- ベンチマークを用いた性能評価

安価で高性能なPCクラスタ用のネットワークの構築

トポロジ制限の理由

Ethernetでループを含むトポロジを作成すると,
ブロードキャストストームが発生する

ブロードキャストが帯域幅を占有してしまい,
他の通信が行えないようになる現象のこと

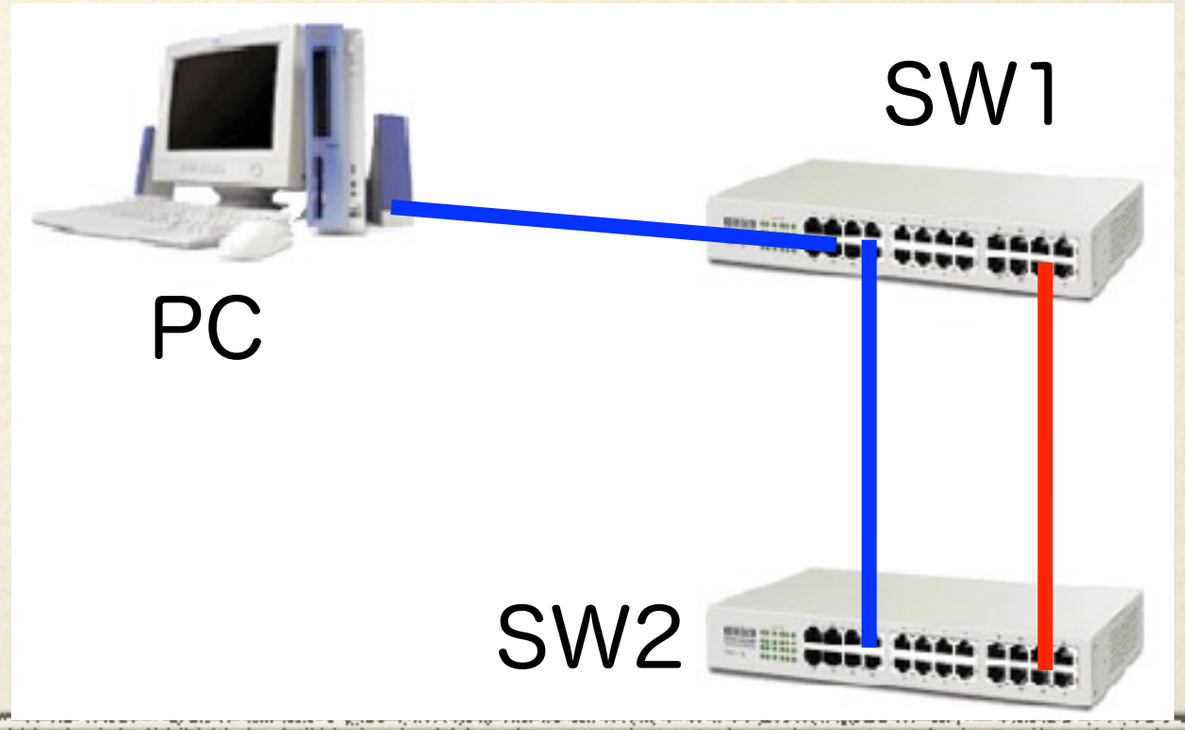


ループを含むような
ネットワークトポロジの場合,
ブロードキャスト
フレームが常に流れる

VLAN (Virtual LAN)

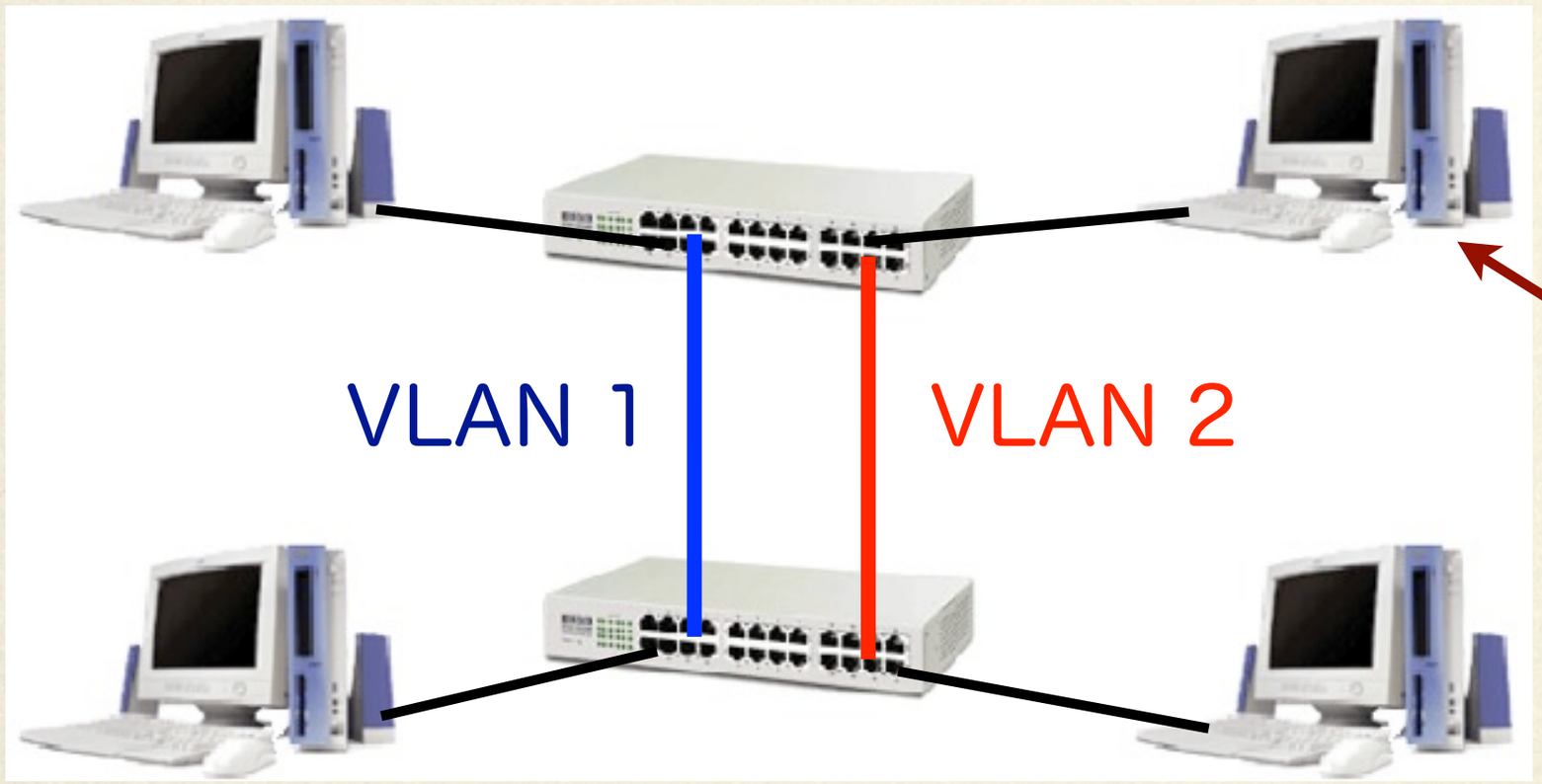
- ネットワーク機器により，物理的な接続とは別の仮想的なネットワークを構成する機能
- ギガビットイーサネットの普及に伴い，VLANを作成できるスイッチが安価で手に入るようになった

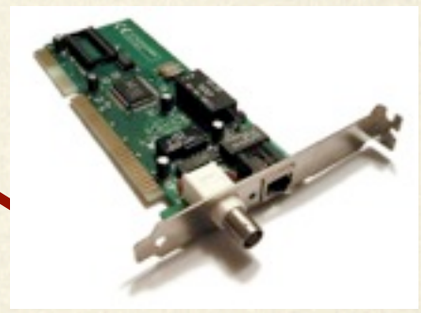
ポート毎にグループを設定可能



VLANルーティング法

“VLANを用いた複数パスを持つクラスタ向き
L2 Ethernet ネットワーク”（工藤ら，産総研，2004）





仮想インタフェースにより，各ホストは複数のIPアドレスを付与する

ホスト毎にネットワークパスを選択する

歴史と研究の位置付け

2004
- VLANルーティング法の提案（工藤ら，産総研）
  - ホスト側でIP経路テーブルの設定などが必要

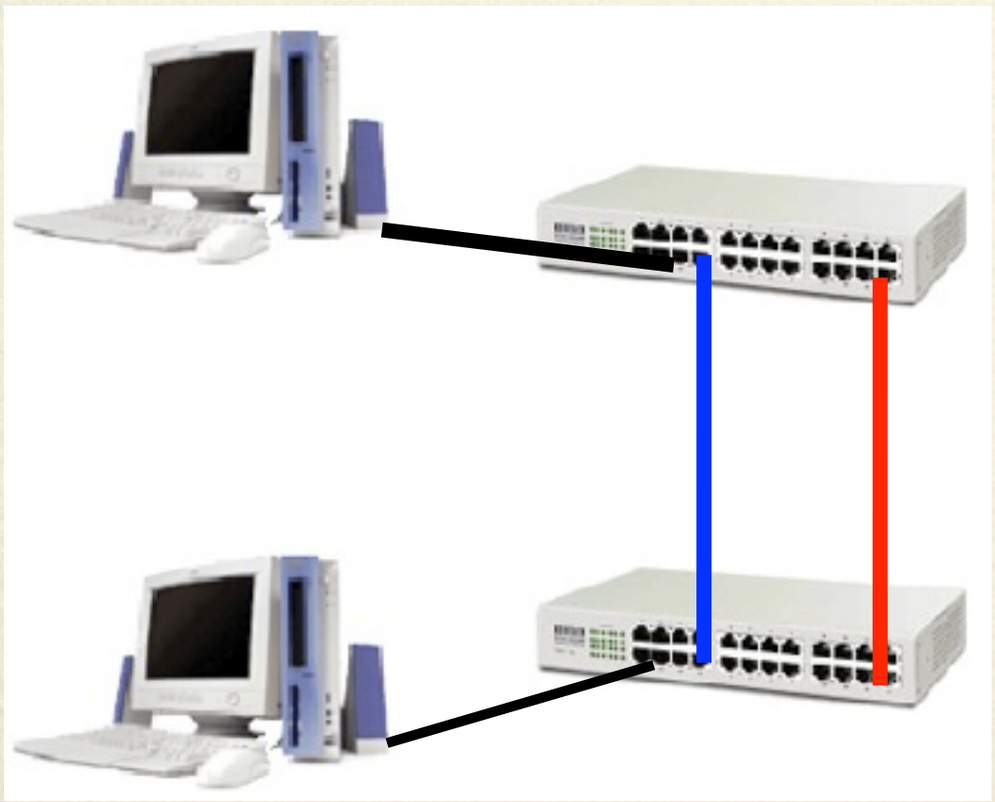
2005
- 動的ルーティング，耐故障性（三浦ら，筑波大）
  - ドライバの改良が必要

2006
- Switch-tagged VLANルーティング法（大塚ら，慶大）
  - ホスト側の設定は必要なし（スイッチで設定）
  - 32台のホストで構成したPCクラスタの評価

現在
- 200台超のホストで構成したPCクラスタを用いた評価
- スイッチにおけるMACアドレス管理の工夫

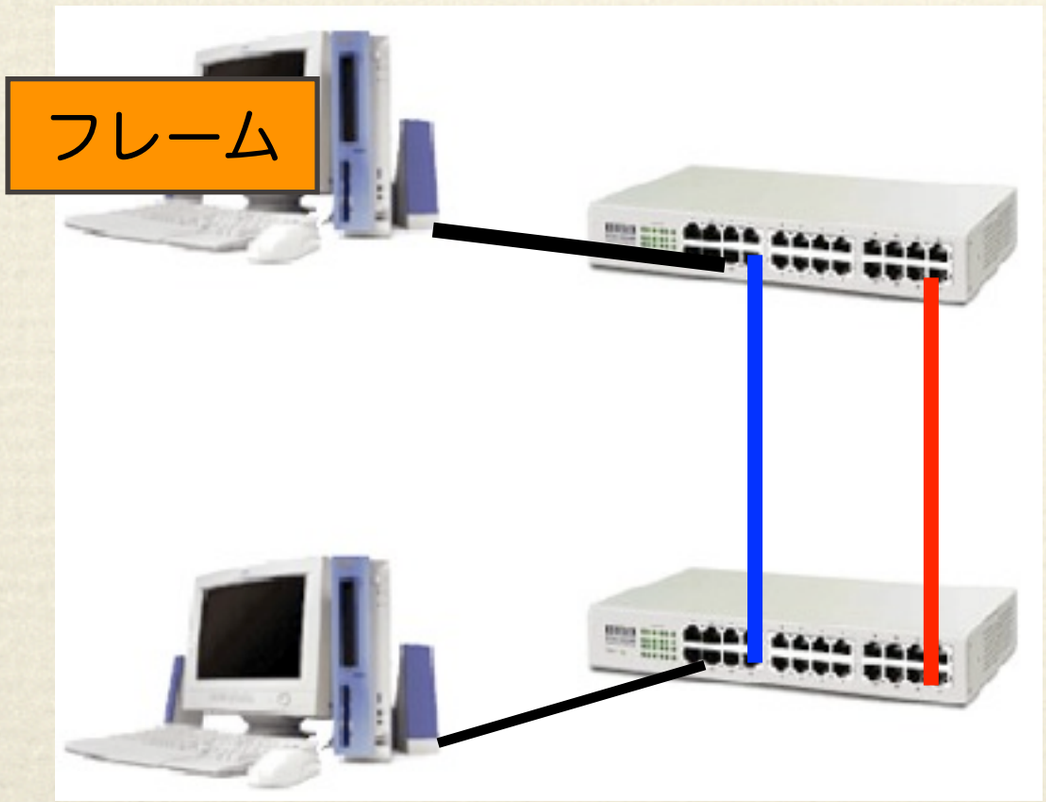
Switch-tagged VLANルーティング法

- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



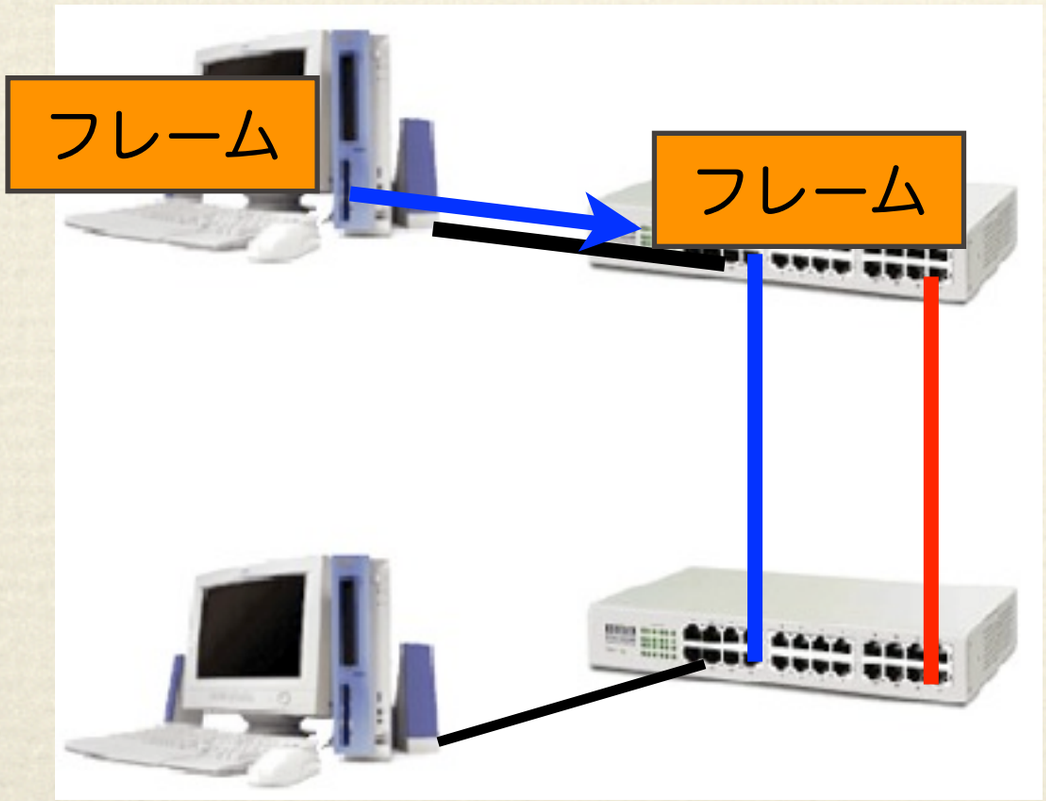
Switch-tagged VLANルーティング法

- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



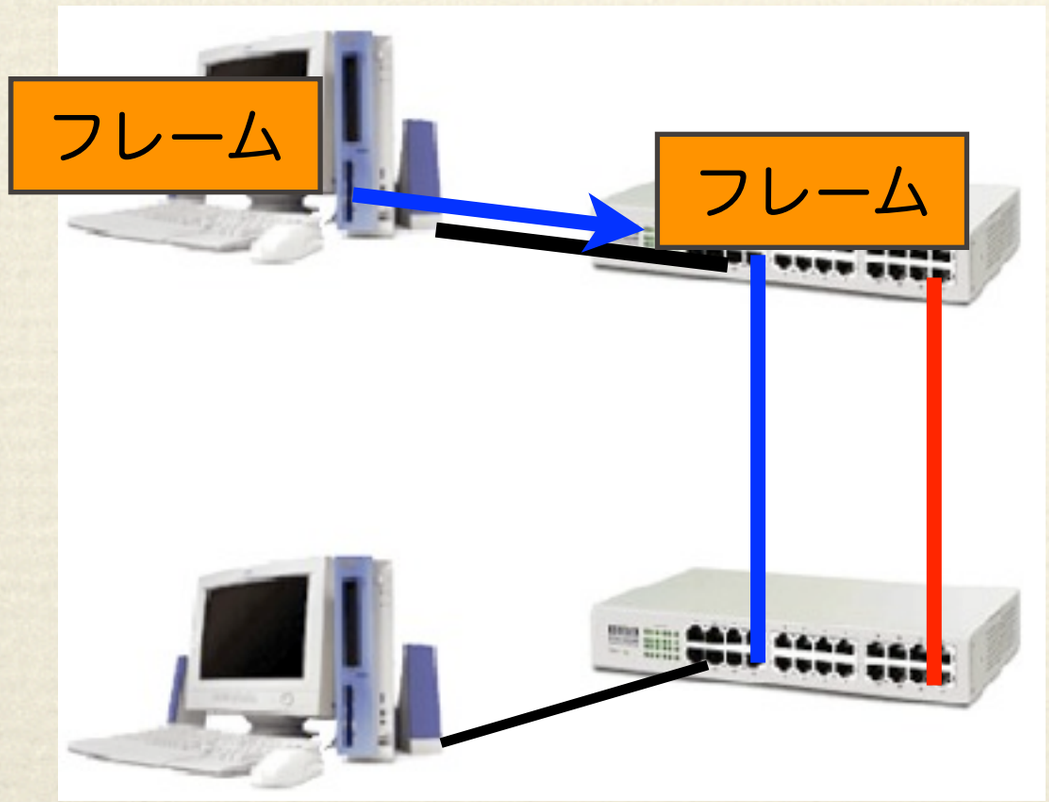
Switch-tagged VLANルーティング法

- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



Switch-tagged VLANルーティング法

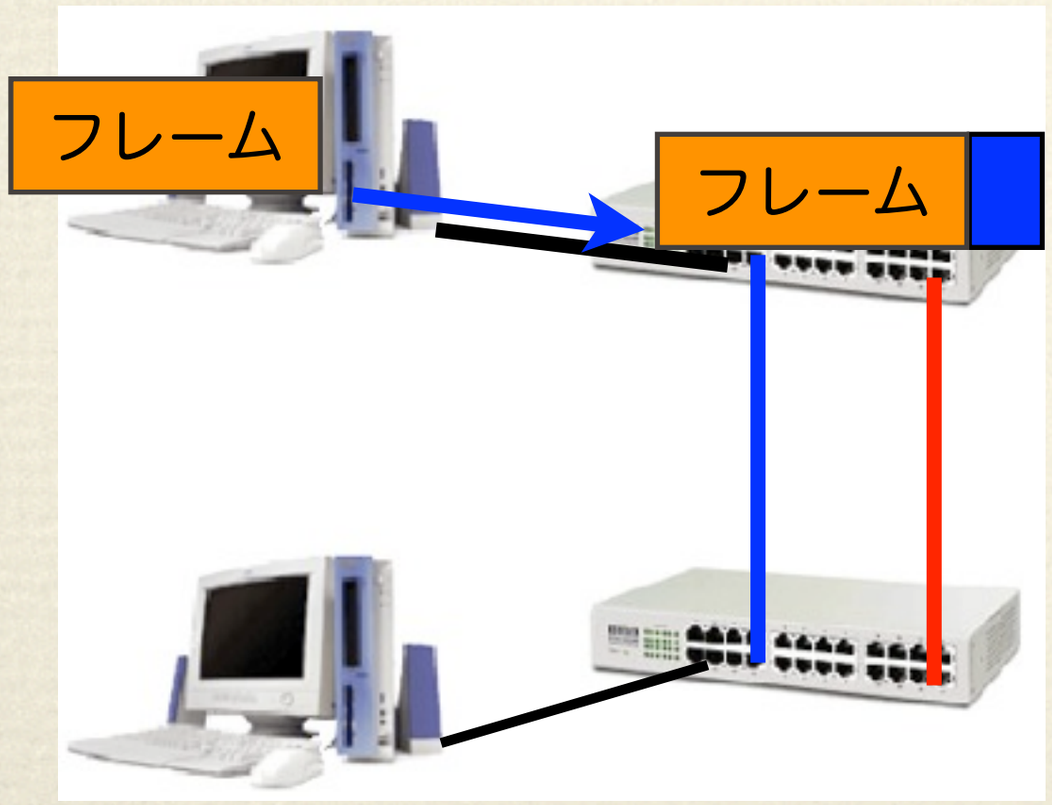
- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



1. スイッチはホストからのフレームにVLANタグを挿入

Switch-tagged VLANルーティング法

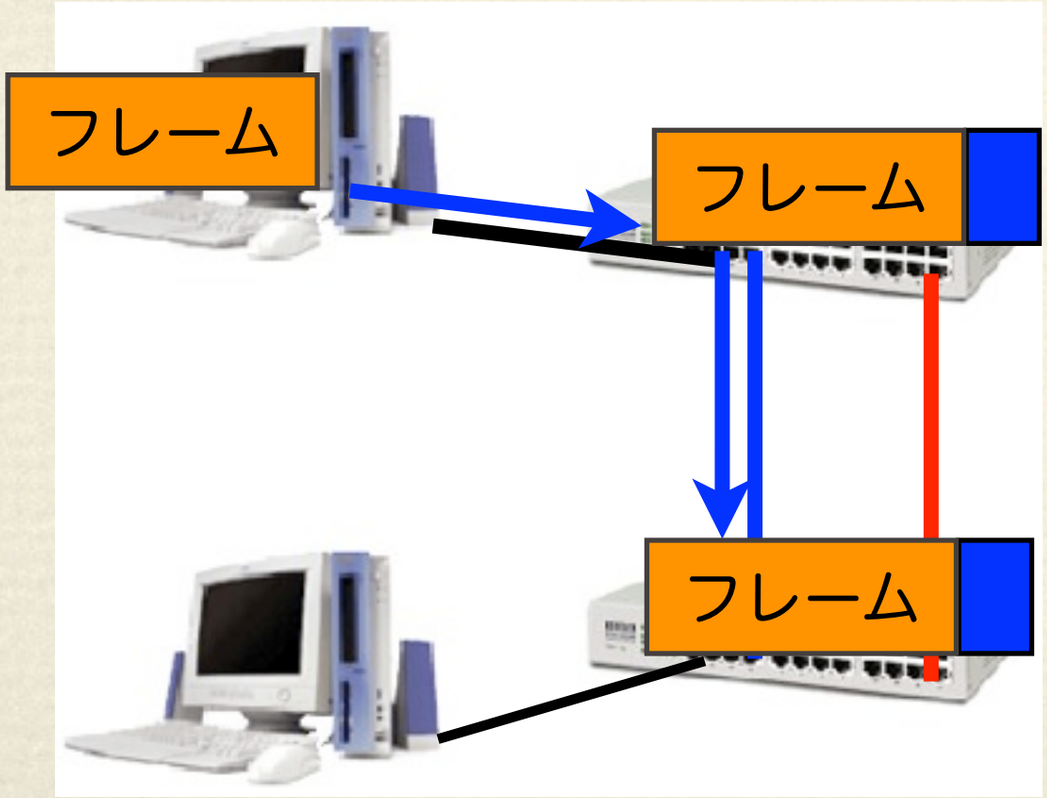
- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



1. スイッチはホストからのフレームにVLANタグを挿入

Switch-tagged VLANルーティング法

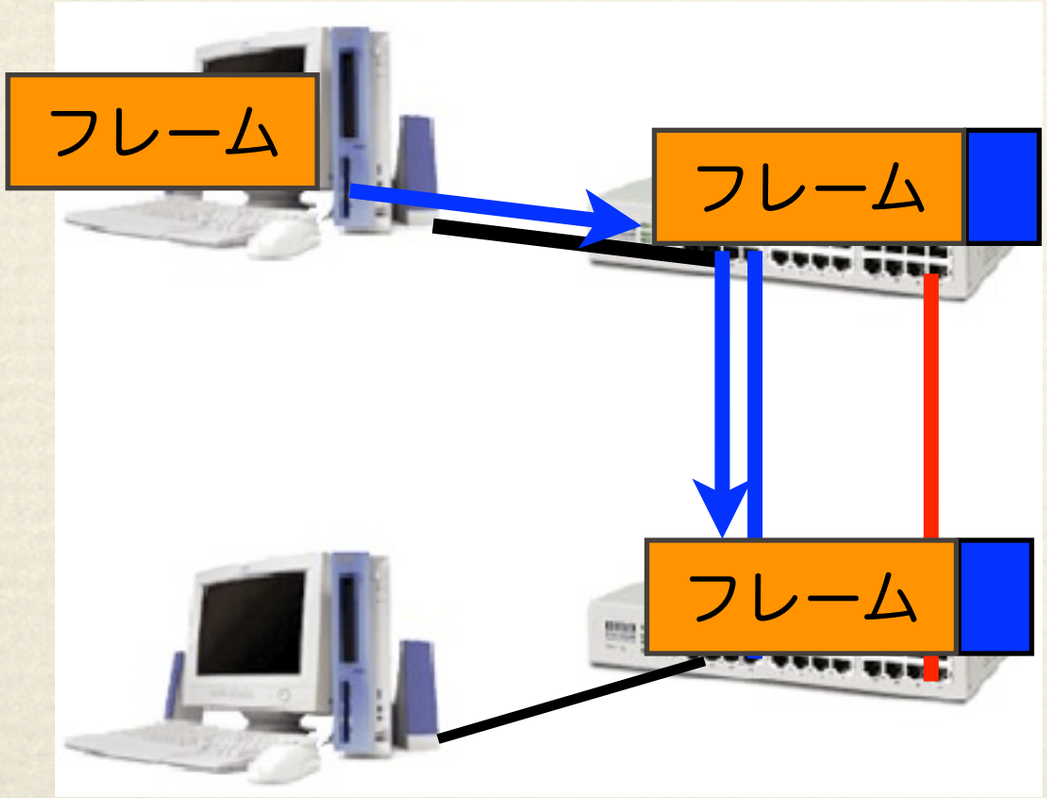
- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



1. スイッチはホストからのフレームにVLANタグを挿入

Switch-tagged VLANルーティング法

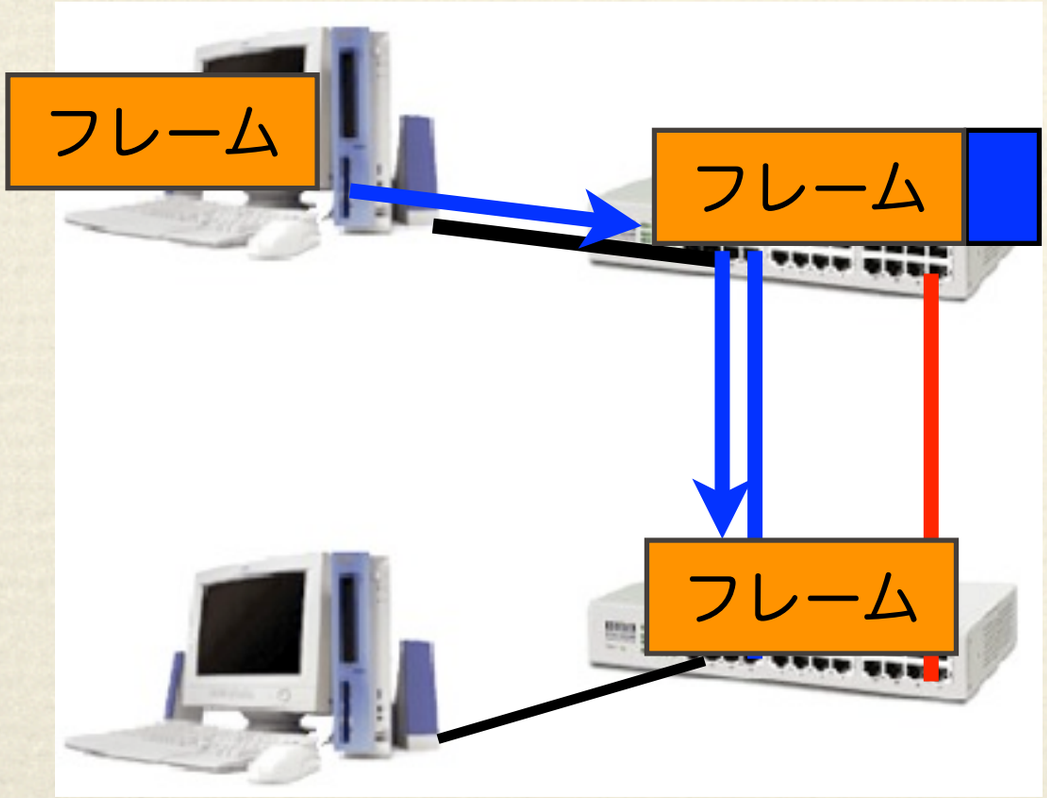
- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



1. スイッチはホストからのフレームにVLANタグを挿入
2. ホストへ出力するフレームはVLANタグを除去

Switch-tagged VLANルーティング法

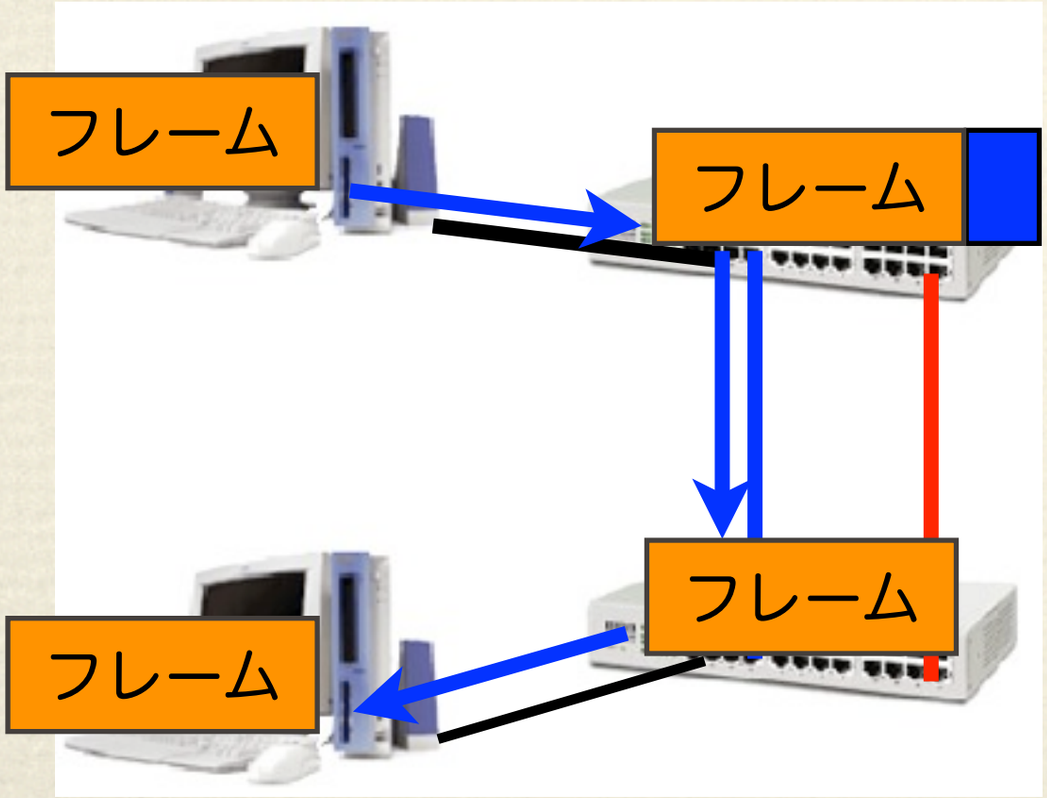
- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする



1. スイッチはホストからのフレームにVLANタグを挿入
2. ホストへ出力するフレームはVLANタグを除去

Switch-tagged VLANルーティング法

- スイッチにVLANの設定を行う
  - 全ポートにVLAN IDを設定
  - 全ポートを全VLANのタグなしメンバとする

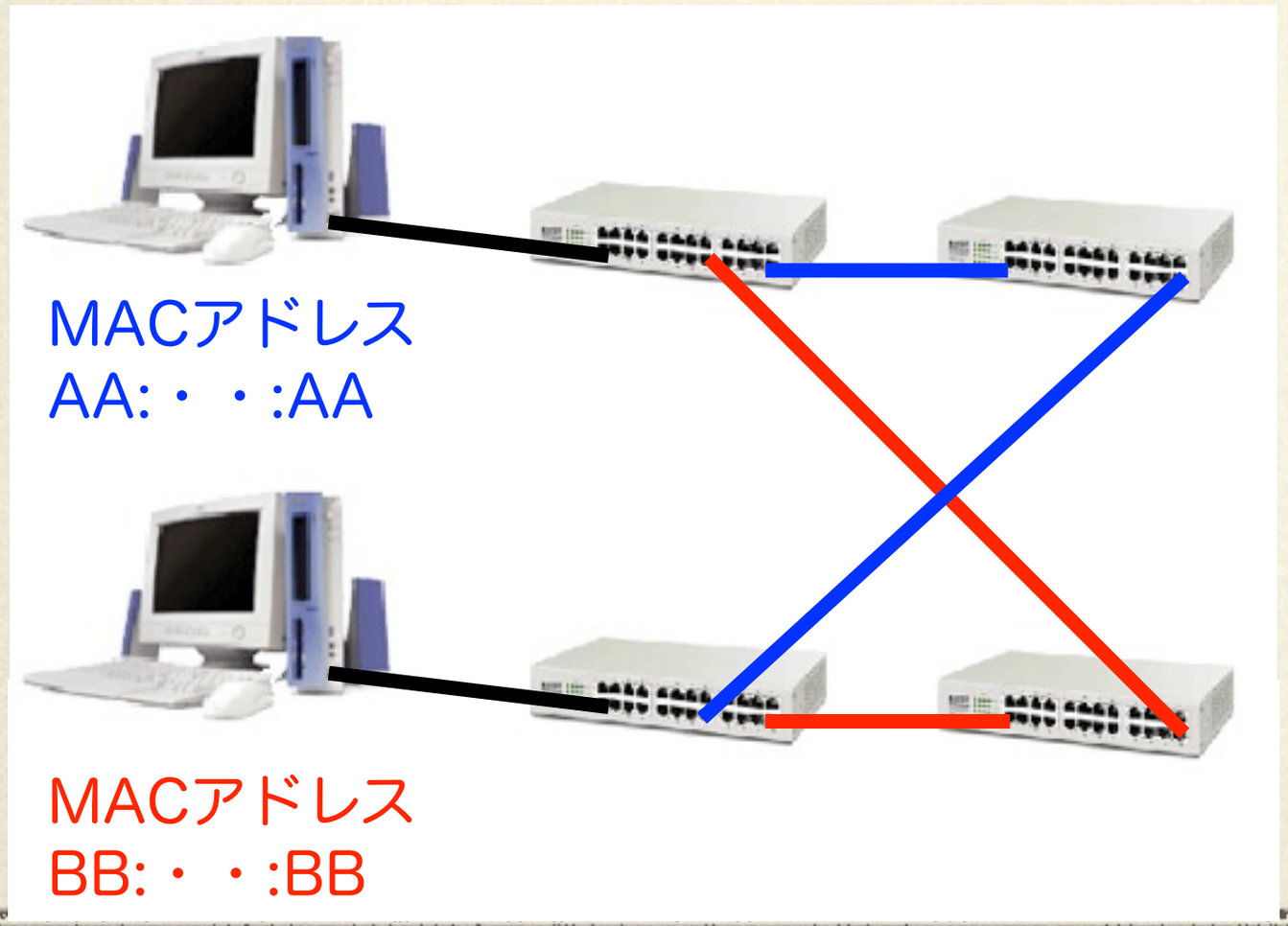


1. スイッチはホストからのフレームにVLANタグを挿入
2. ホストへ出力するフレームはVLANタグを除去

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

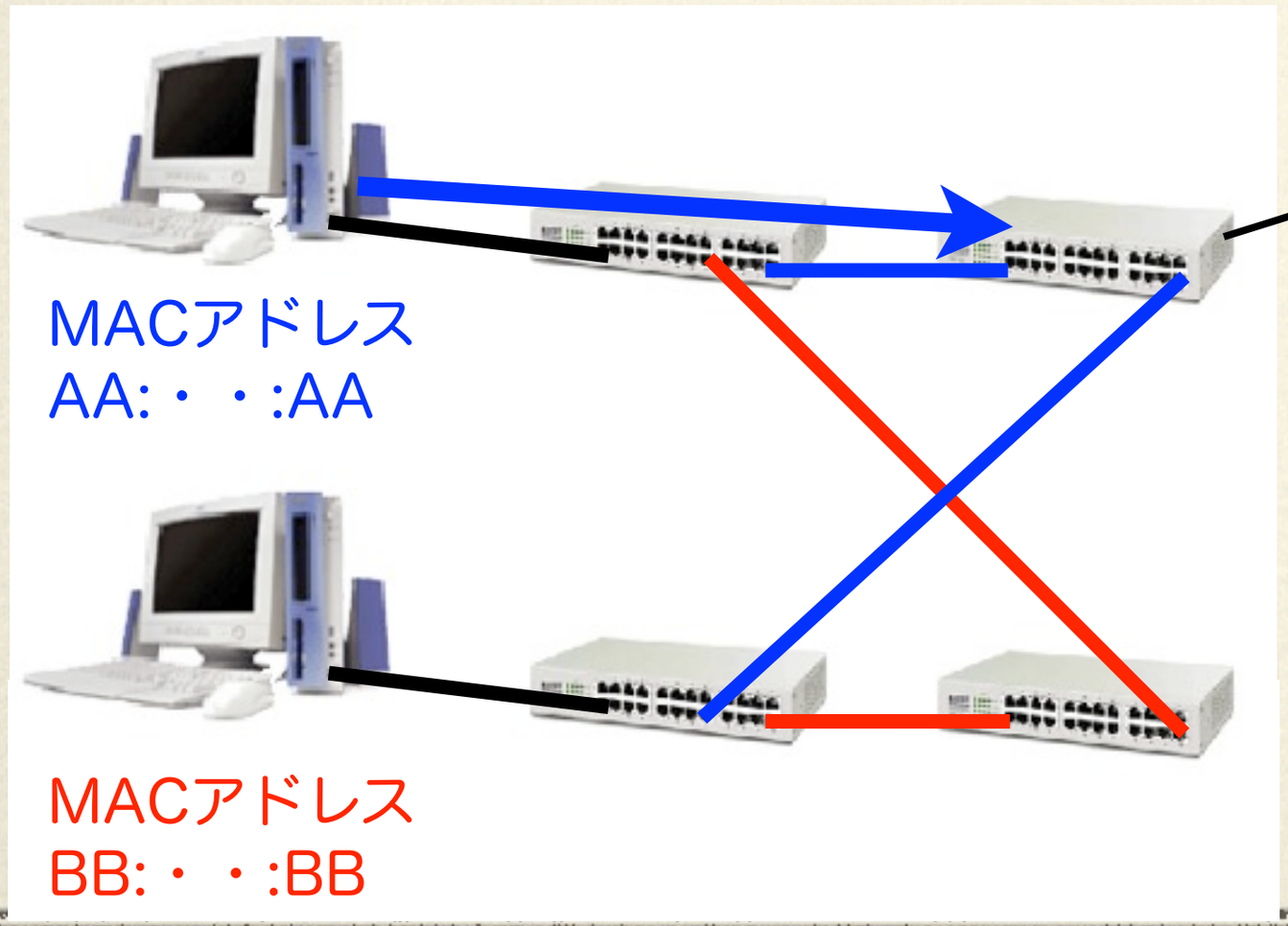
➡ スイッチに宛先MACアドレスが登録されない



MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない

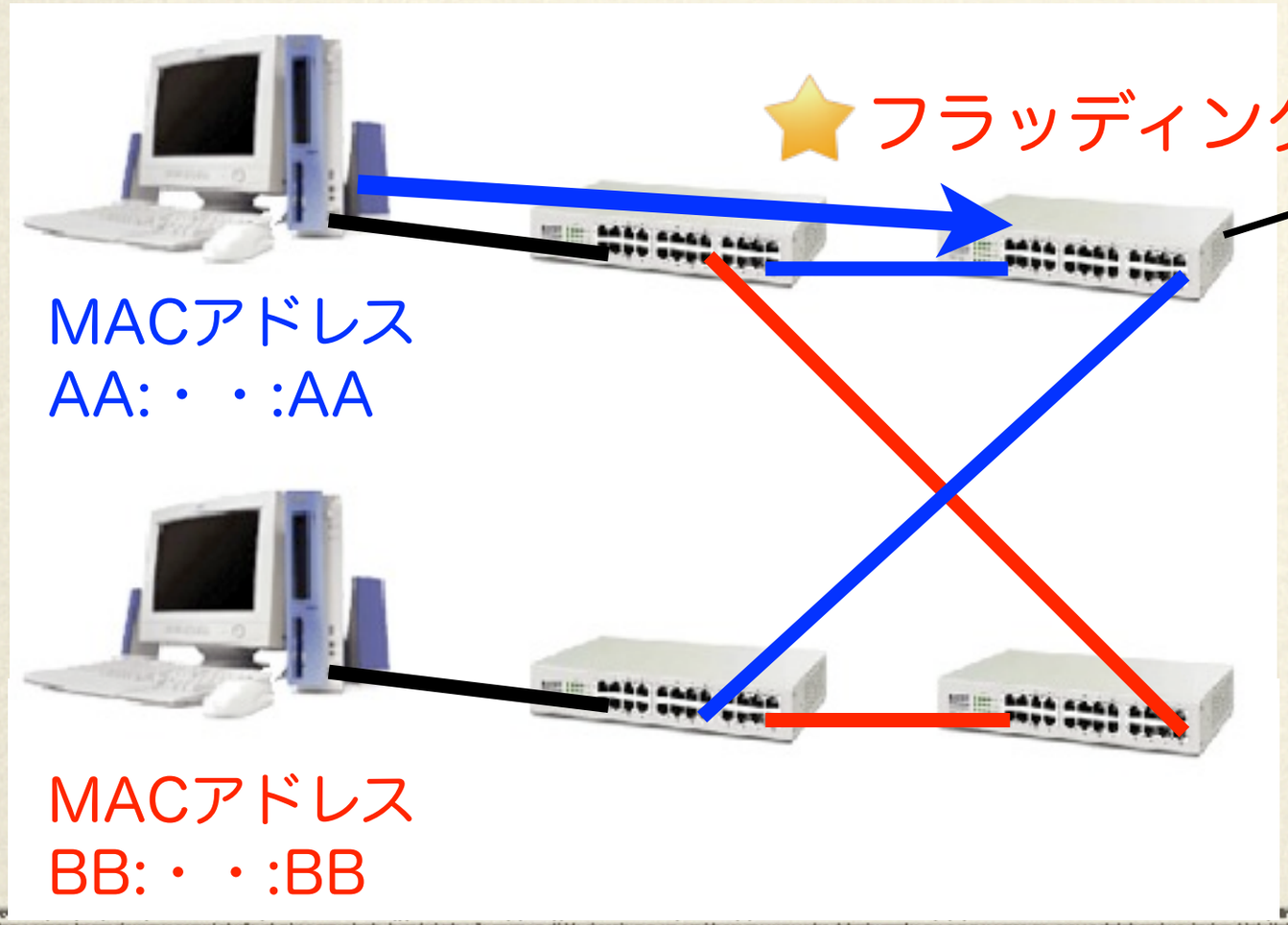


VLAN	port	MACアドレス
1	1	AA:・・:AA

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない

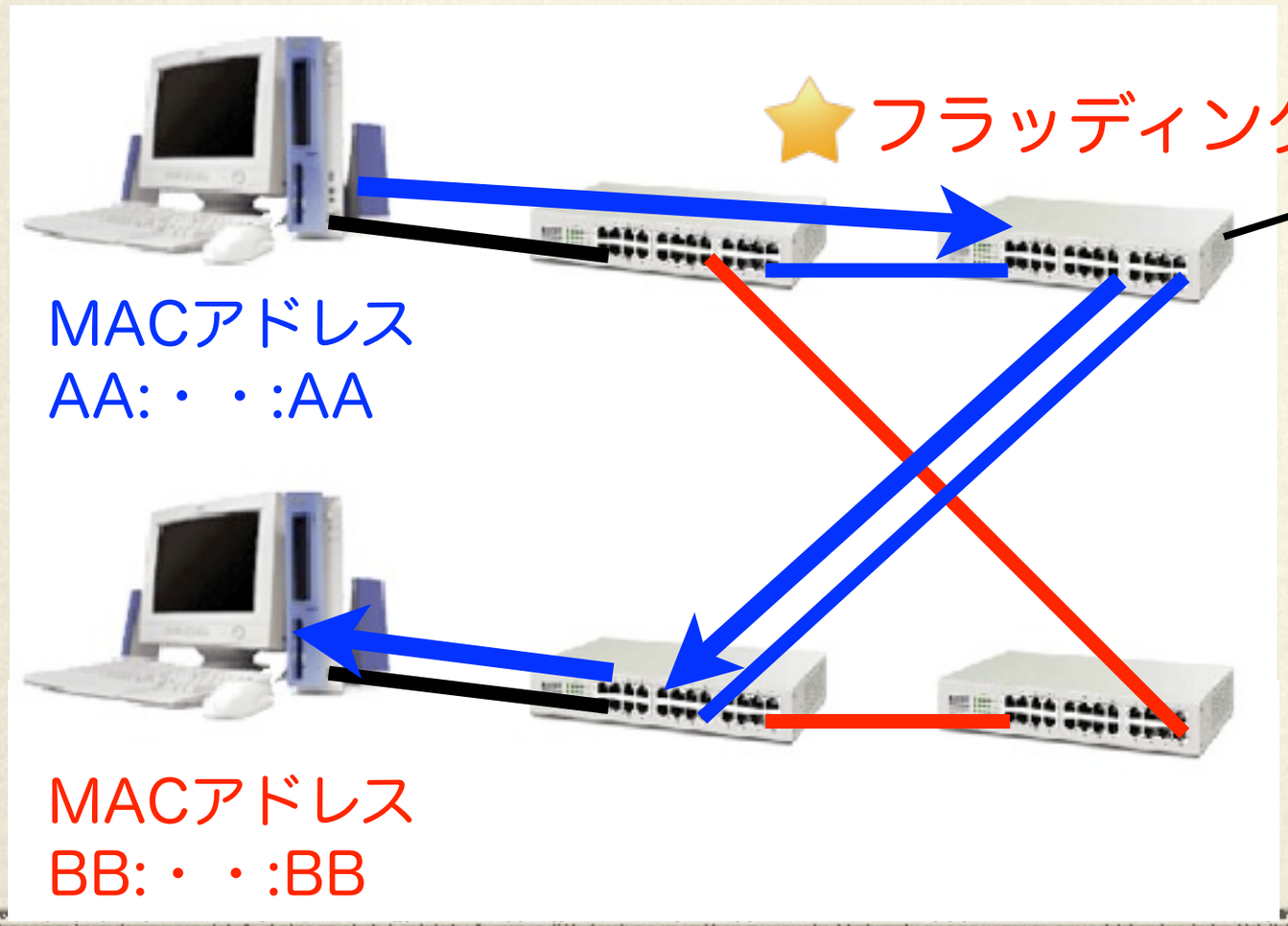


VLAN	port	MACアドレス
1	1	AA:・・:AA

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない

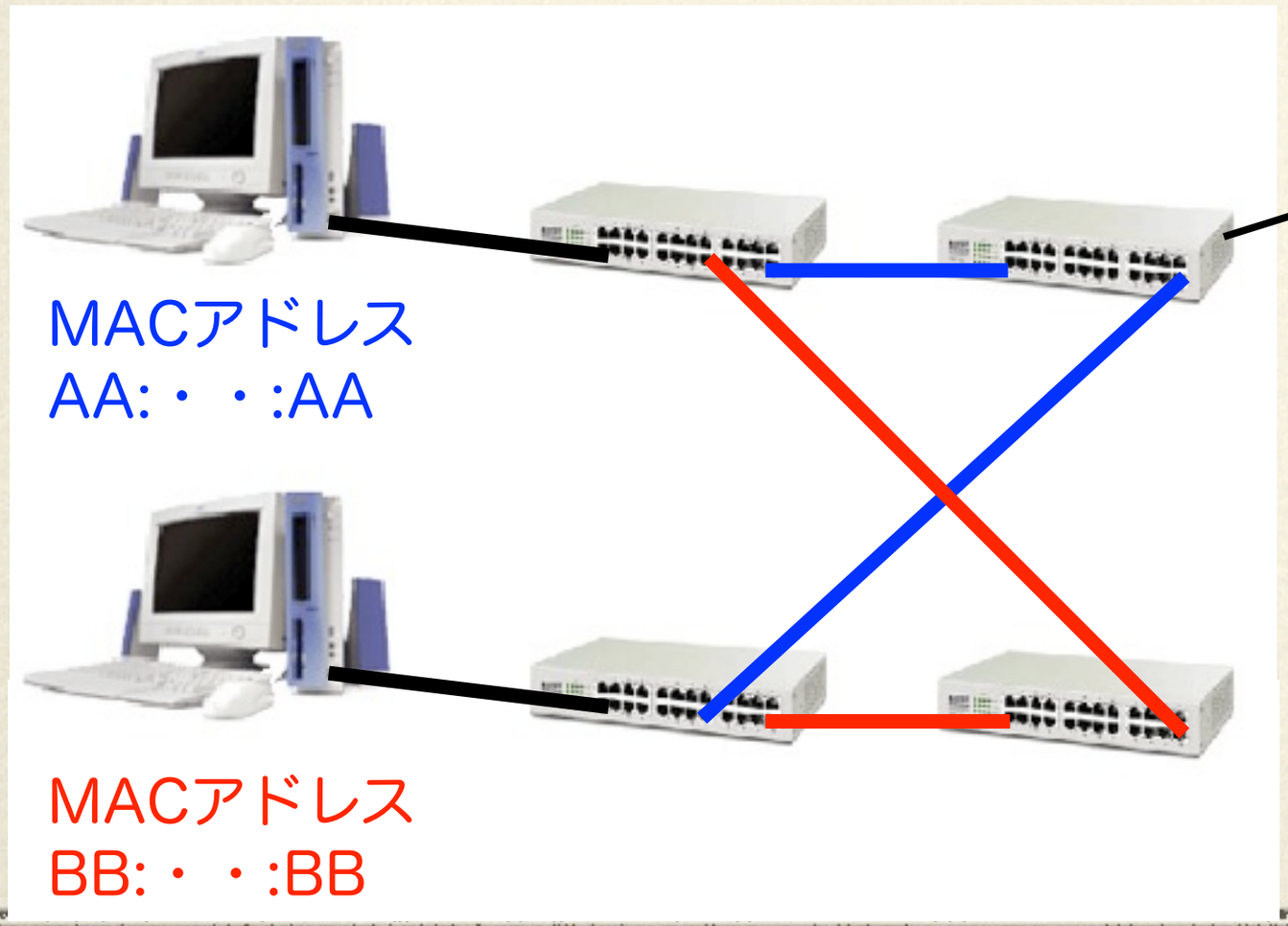


VLAN	port	MACアドレス
1	1	AA:・・:AA

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない

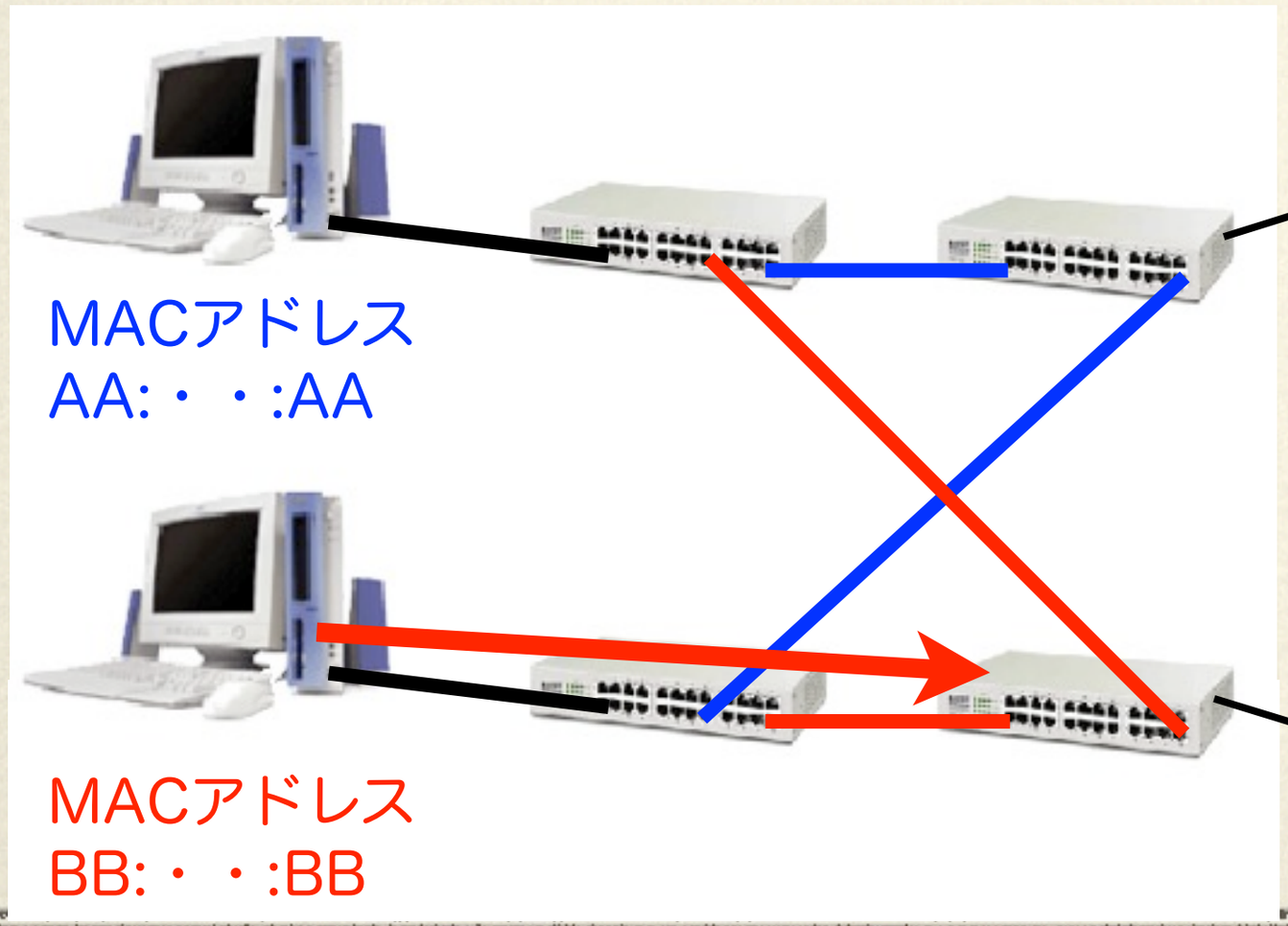


VLAN	port	MACアドレス
1	1	AA:・・:AA

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない



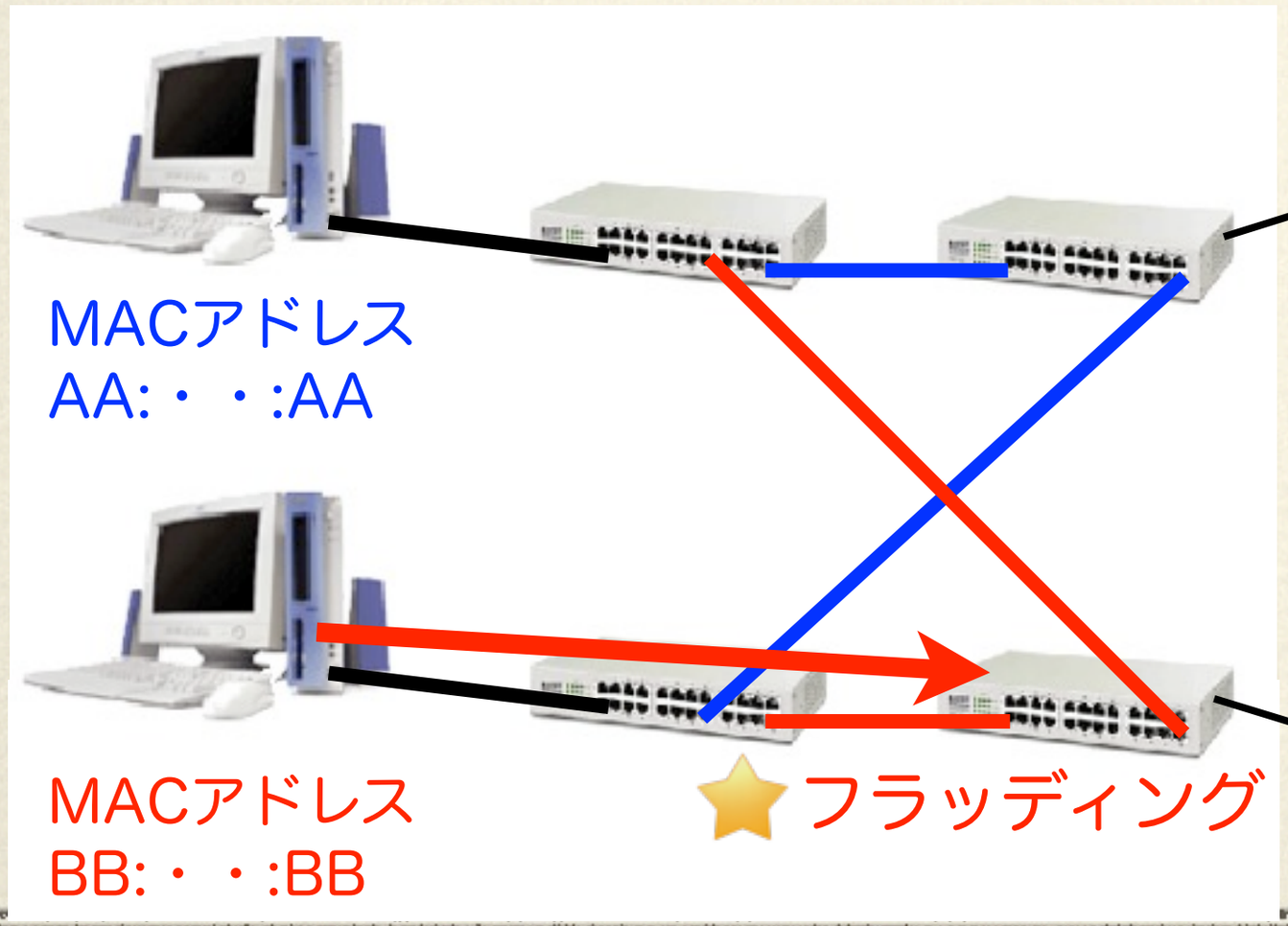
VLAN	port	MACアドレス
1	1	AA:・・:AA

VLAN	port	MACアドレス
2	2	BB:・・:BB

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない



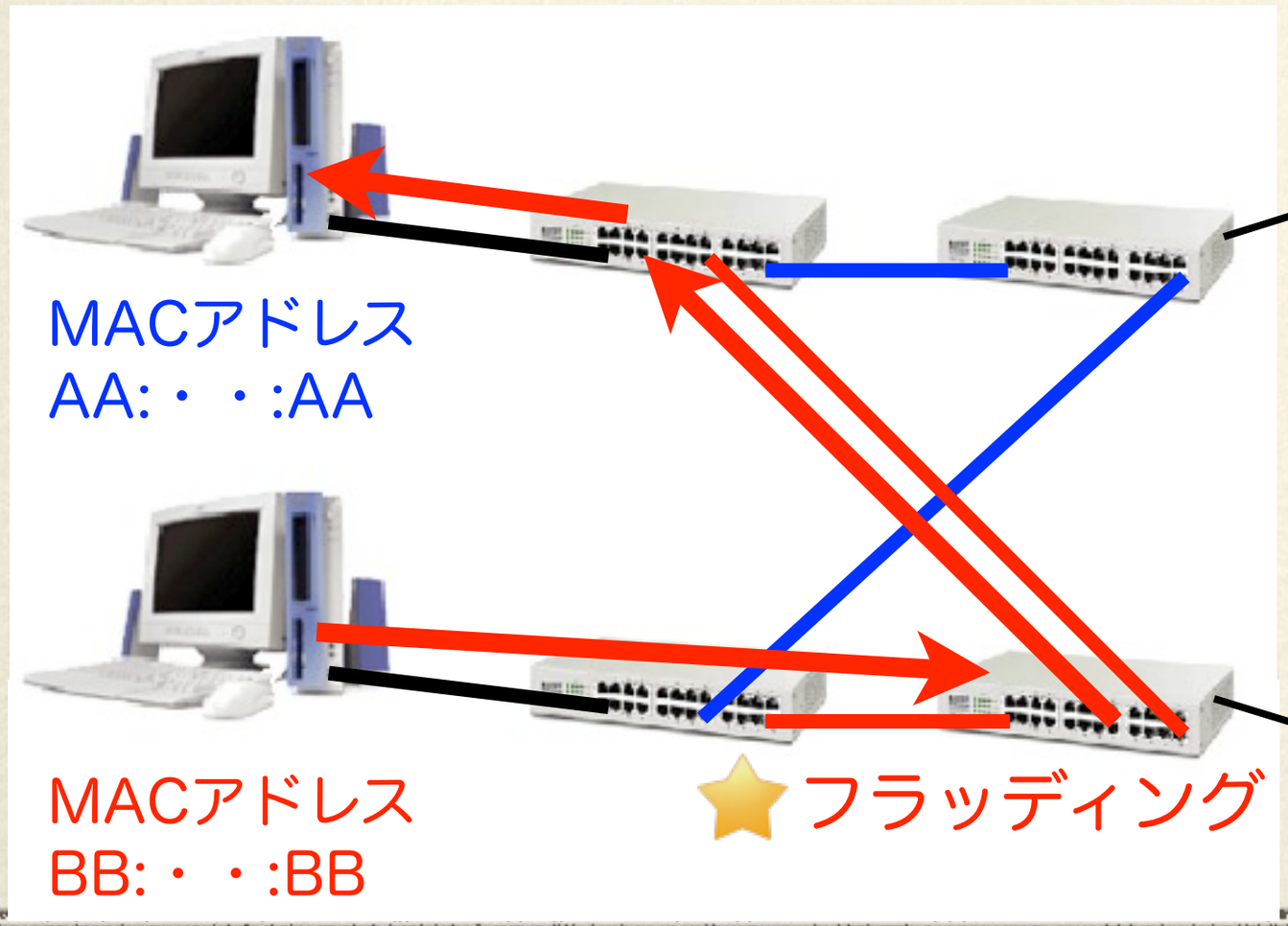
VLAN	port	MACアドレス
1	1	AA:・・:AA

VLAN	port	MACアドレス
2	2	BB:・・:BB

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない



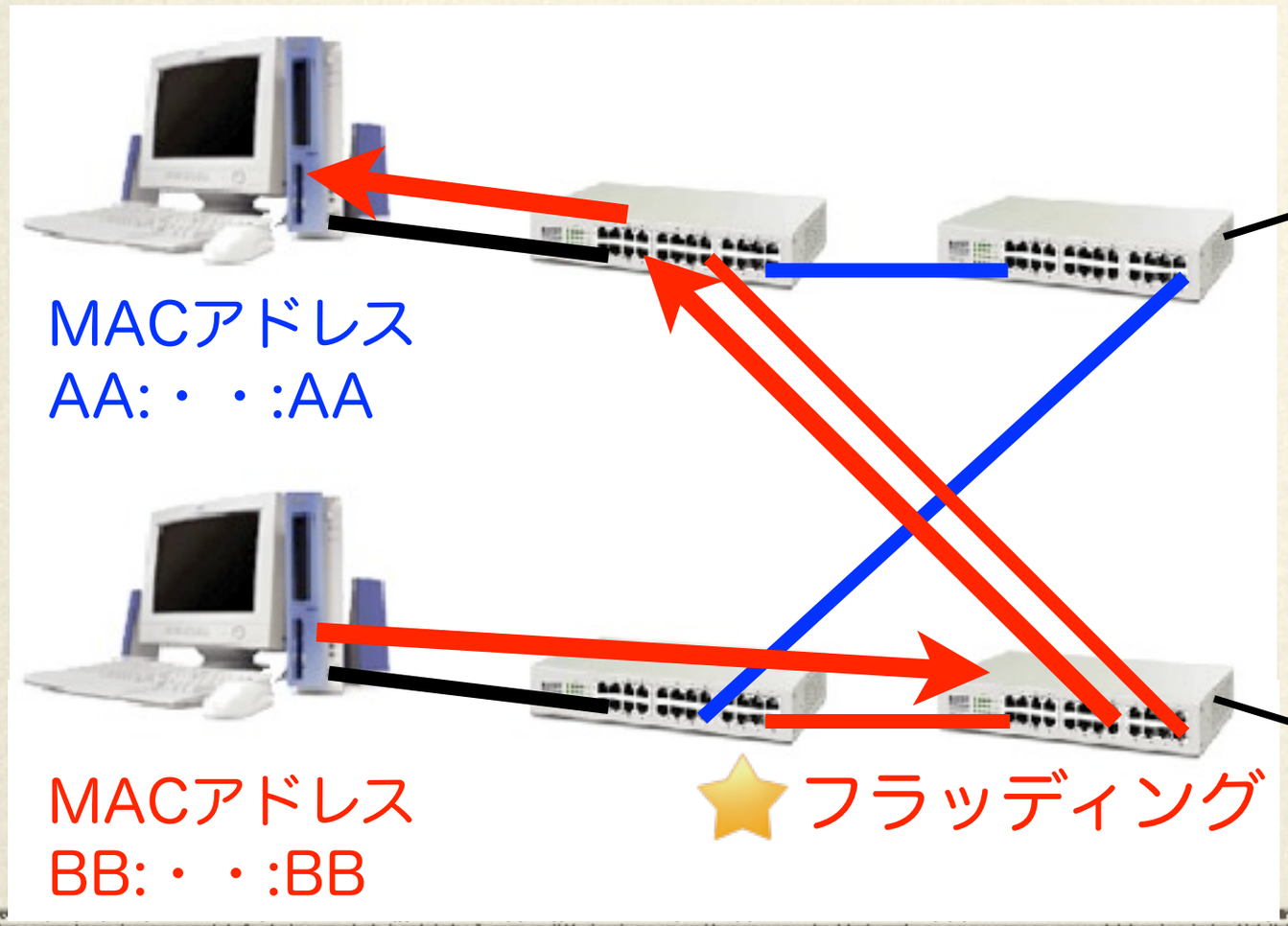
VLAN	port	MACアドレス
1	1	AA:・・:AA

VLAN	port	MACアドレス
2	2	BB:・・:BB

MACアドレス管理の必要性

VLANルーティング法では，ホスト間で通信の送受信を行う経路が異なる

➡ スイッチに宛先MACアドレスが登録されない



VLAN	port	MACアドレス
1	1	AA:・・:AA

フラッディングが
毎回発生してしまう

VLAN	port	MACアドレス
2	2	BB:・・:BB

スイッチに行う設定

- Switch-tagged法では，MACアドレスを静的に各スイッチに設定する
- 各ホストがどの経路で通信を行うかを決定する必要がある
- 各スイッチが持つMACアドレステーブルで，全ての経路情報を管理してルーティングを行う

#VLANid	#MAC アドレス	#ポート	#登録方式
101	001E.C944.C88B	ch1	Dynamic
101	001E.C94C.7E43	ch1	Dynamic
101	001E.C94C.7E44	ch1	Dynamic
101	001E.C94C.845F	1/g2	Dynamic
101	001E.C94C.8460	1/g13	Dynamic
101	001E.C94C.8462	ch1	Dynamic
101	001E.C94C.8463	ch1	Dynamic
101	001E.C94C.90AF	ch1	Dynamic
101	001E.C94C.90B0	ch1	Dynamic
101	001E.C94C.90B2	ch1	Dynamic
101	001E.C94C.90B3	ch1	Dynamic

MACアドレス管理の問題点

- 静的にMACアドレスを登録できる数は,
  スイッチによって異なる

Dell PowerConnect 6248

今回用いたスイッチの
静的に登録可能な数は100個
大規模な環境では実装が困難

必要な登録数： ホスト数 x 使用VLAN数 x NIC数/ホスト
実験用クラスタ ： 225 x 8 x 1 = 1800 個必要

スイッチの自動学習機能では8000個登録可能

MACアドレスの管理手順

- 提案する実装方法の工夫点

  仮想インタフェースとスイッチの自動学習機能によって
  MACアドレスを登録し，その状態を最大時間保持する

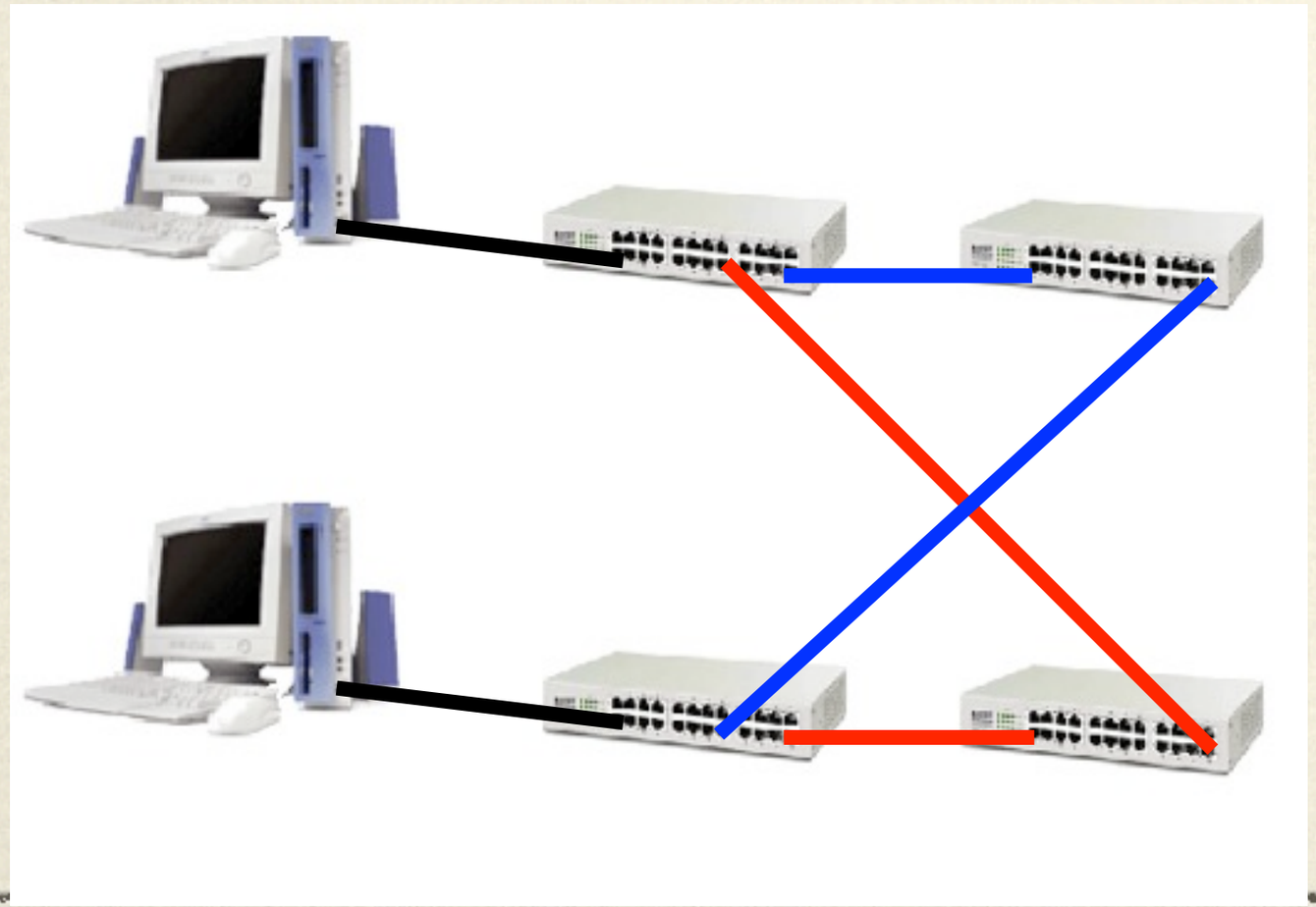
- 具体的手順

  1. スイッチの情報保持期間（Aging Time）を最大に設定
  2. 各ホストで仮想インタフェースを作成
  3. 各ホストから全IPアドレスにブロードキャスト

MACアドレスの管理手順

1. スイッチのMACアドレスの情報保持期間を最大に設定

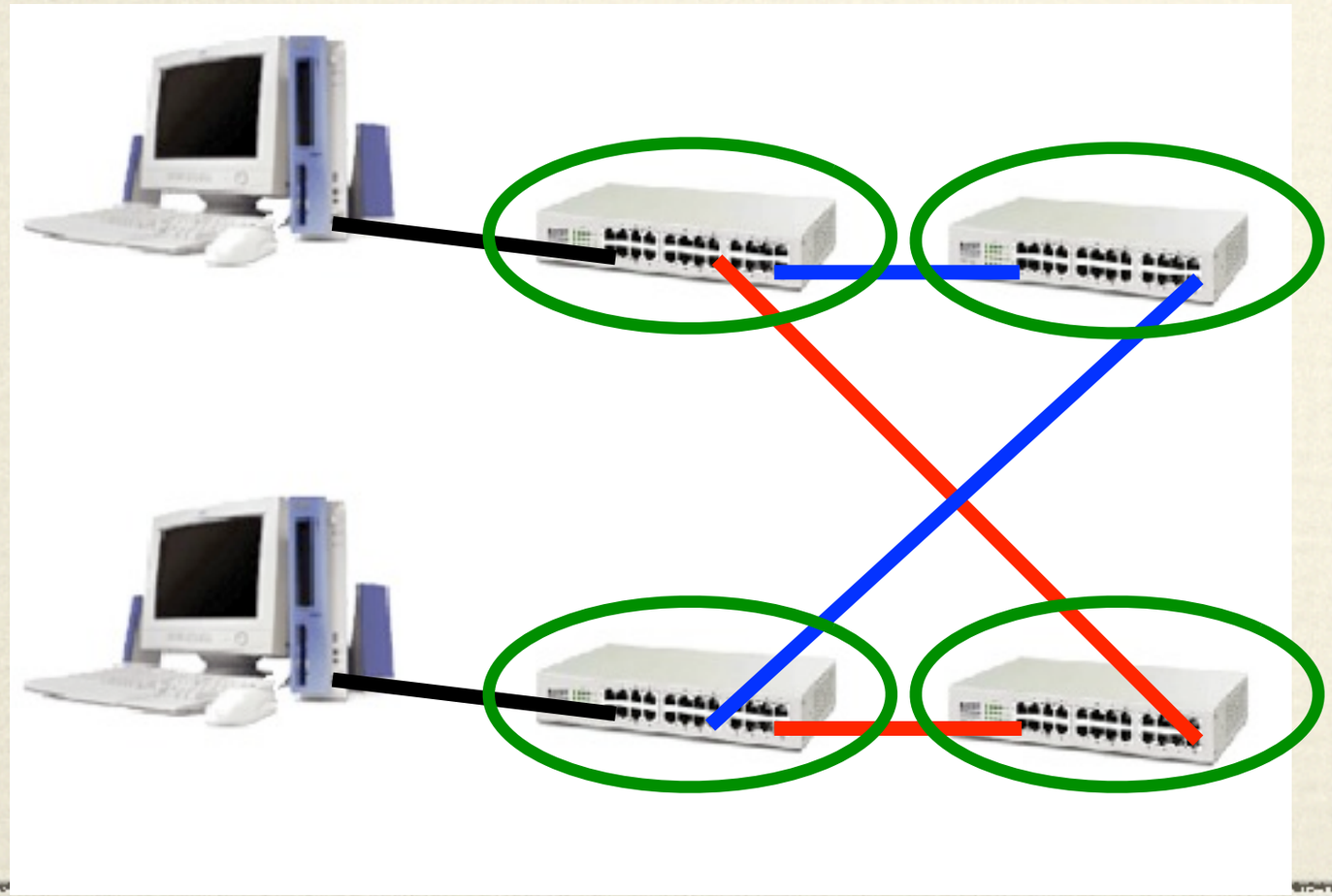
   Dell PowerConnect 6248の場合：1,000,000秒 = 約12日



MACアドレスの管理手順

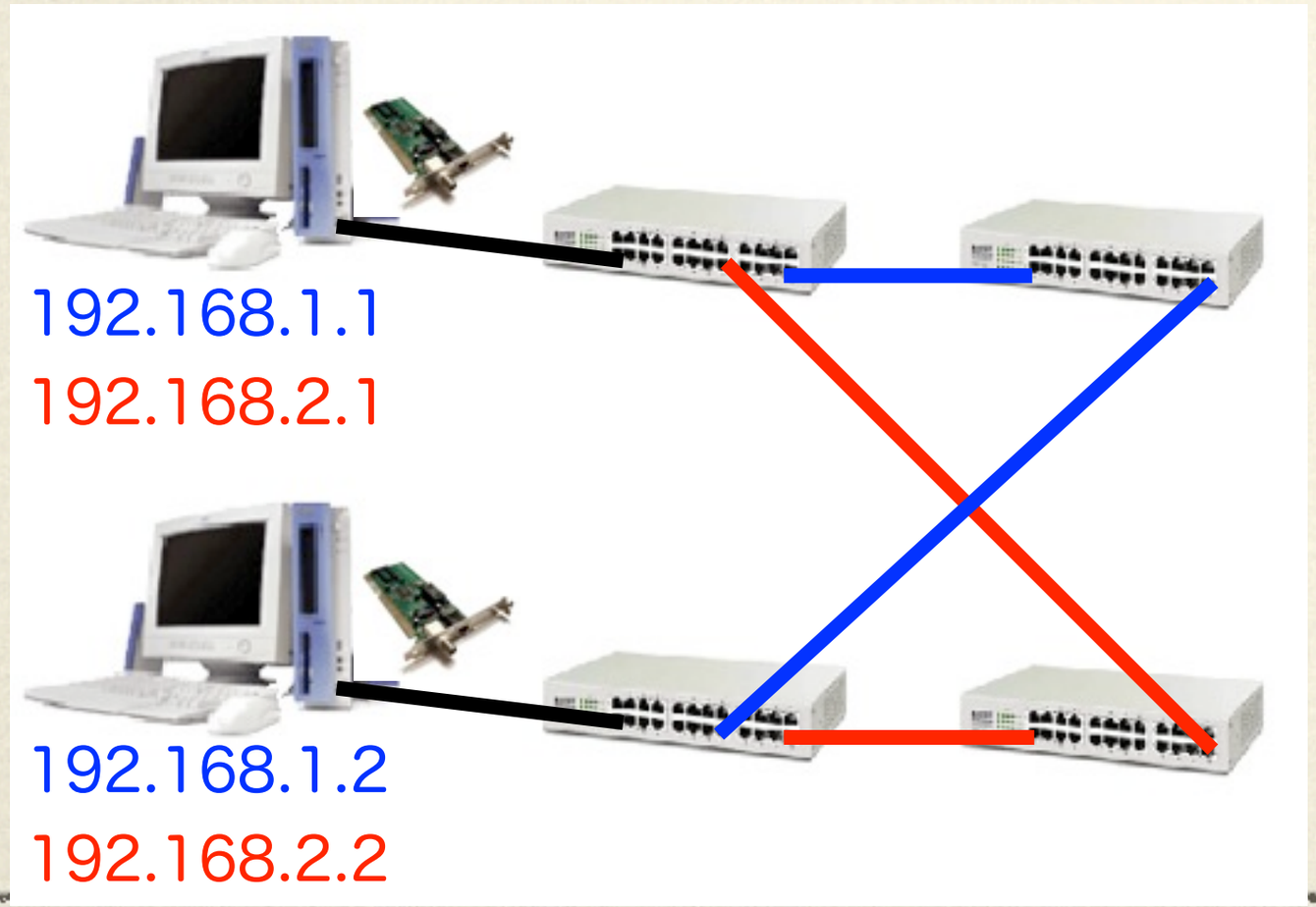
1. スイッチのMACアドレスの情報保持期間を最大に設定

   Dell PowerConnect 6248の場合：1,000,000秒 = 約12日



MACアドレスの管理手順

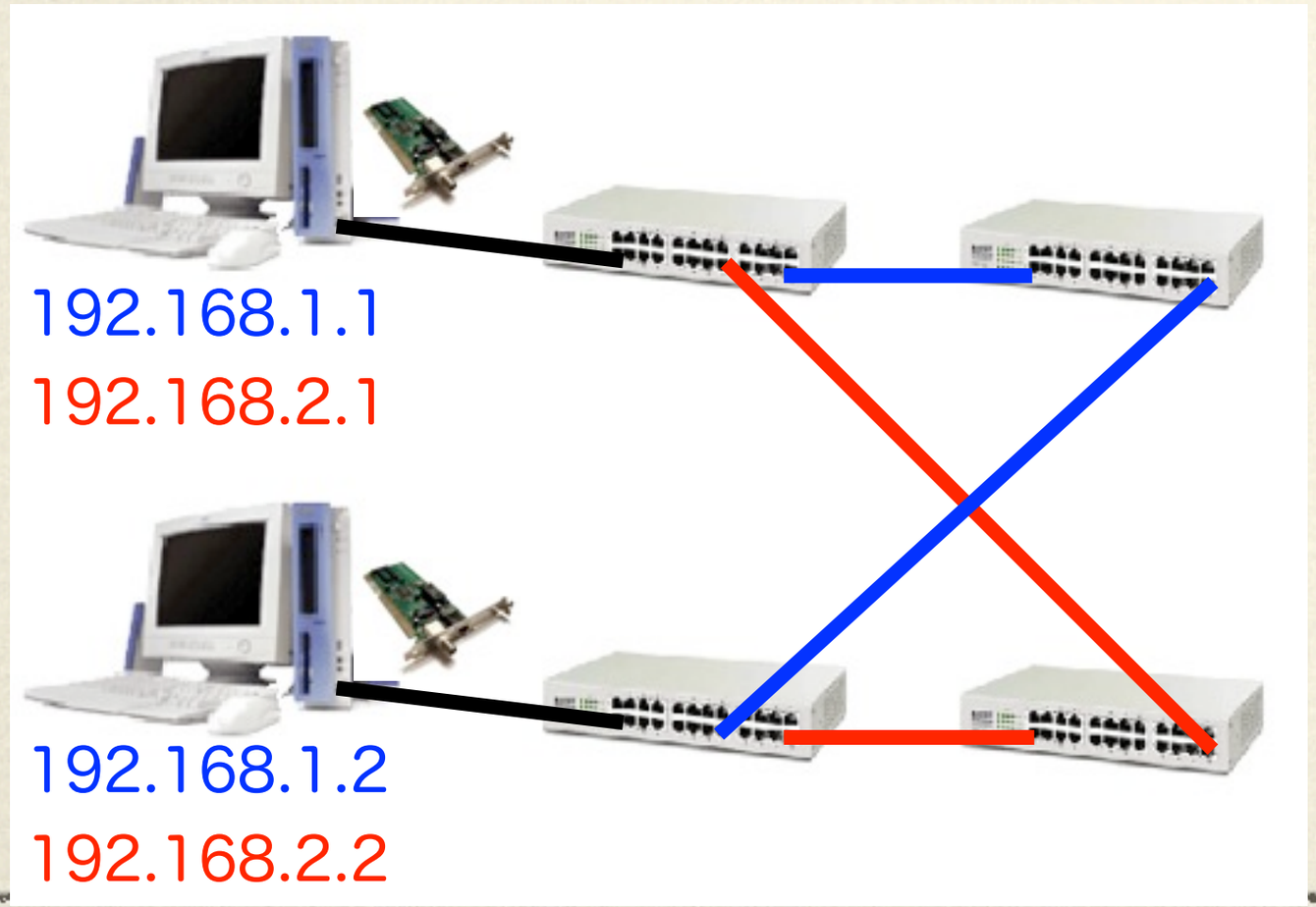
1. スイッチのMACアドレスの情報保持期間を最大に設定
   Dell PowerConnect 6248の場合：1,000,000秒 = 約12日
2. 各ホストに仮想インタフェースを作成し，IPアドレスを
   VLANの数だけ付与する



MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

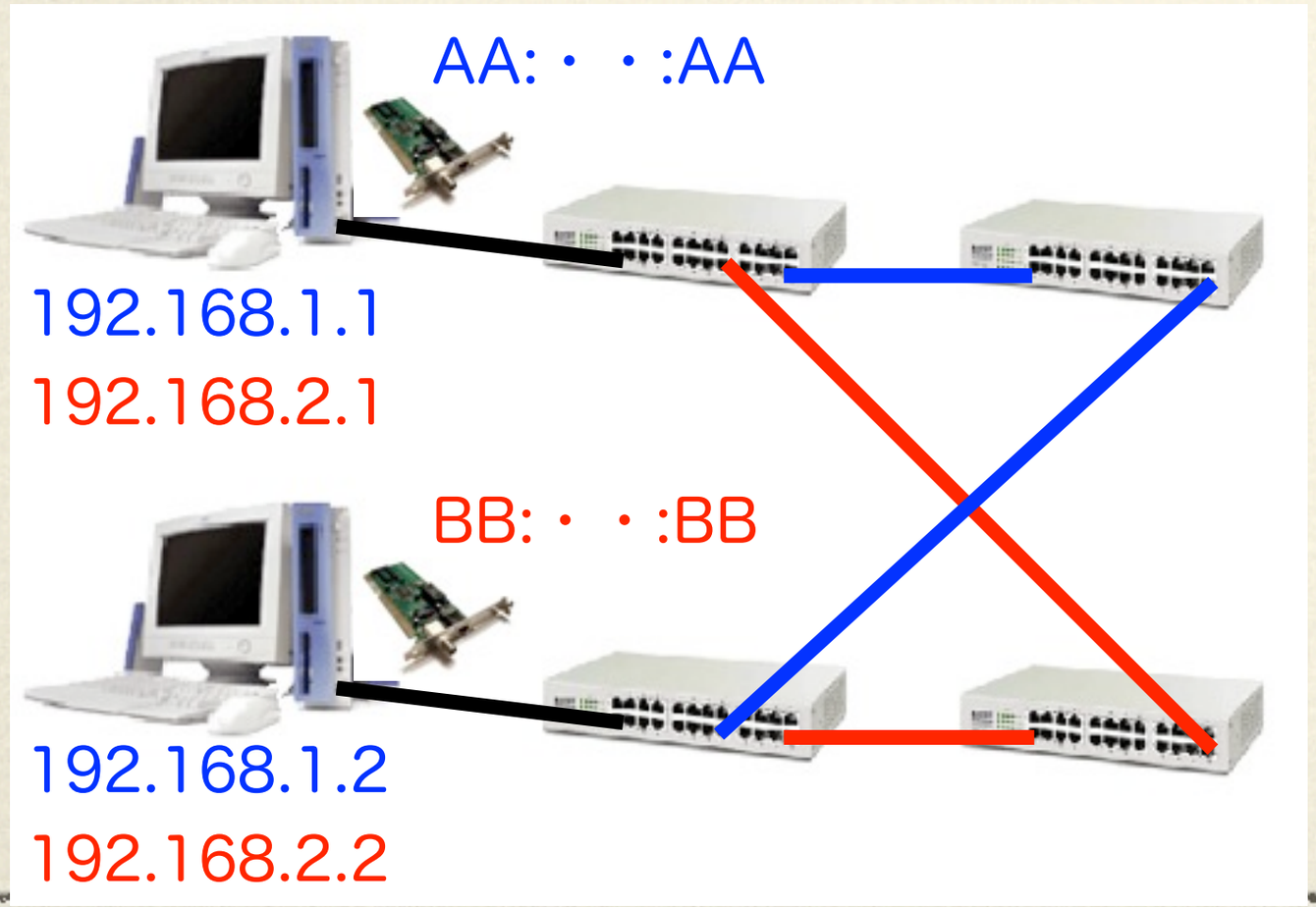
```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```



MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

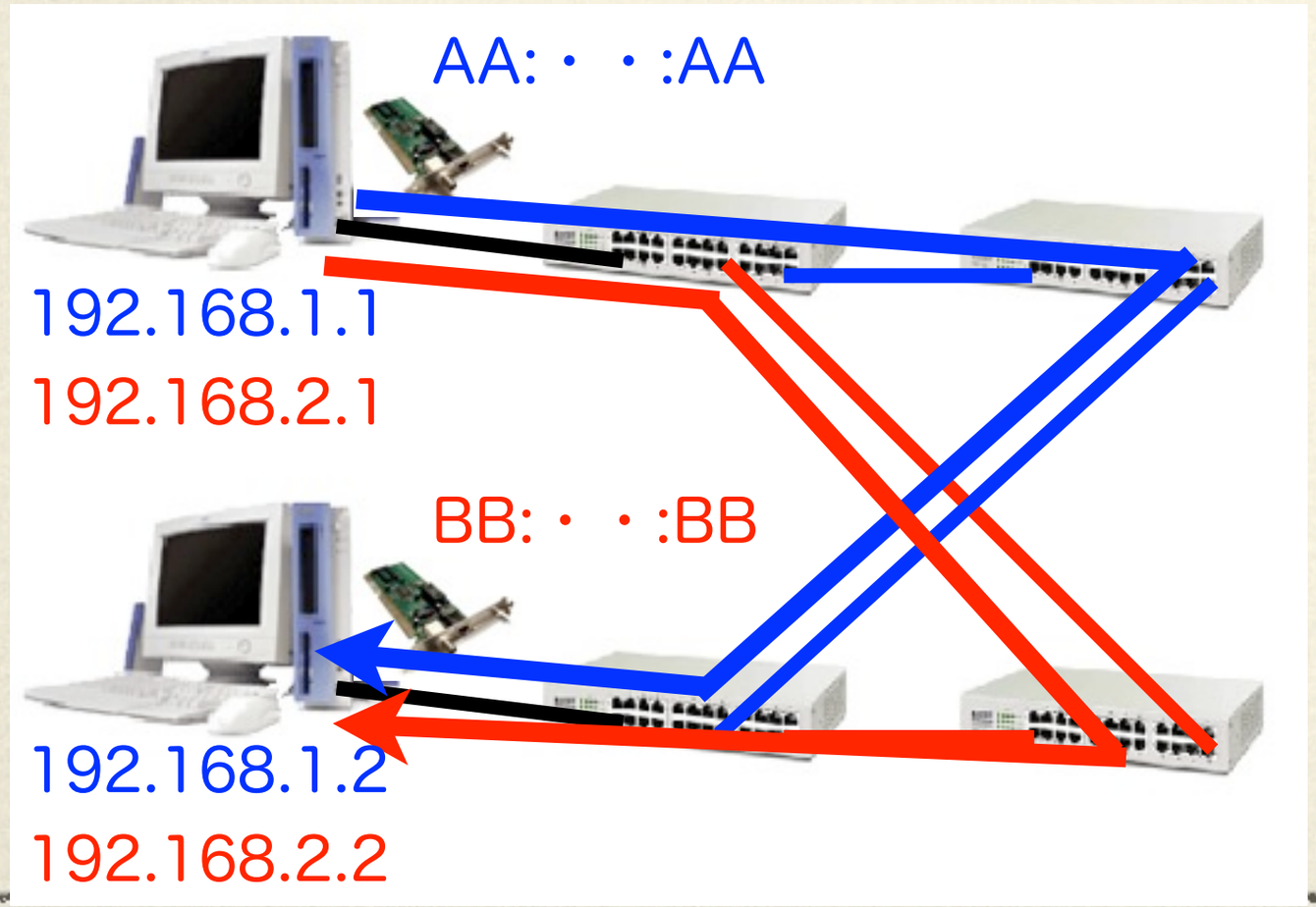
```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```



MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

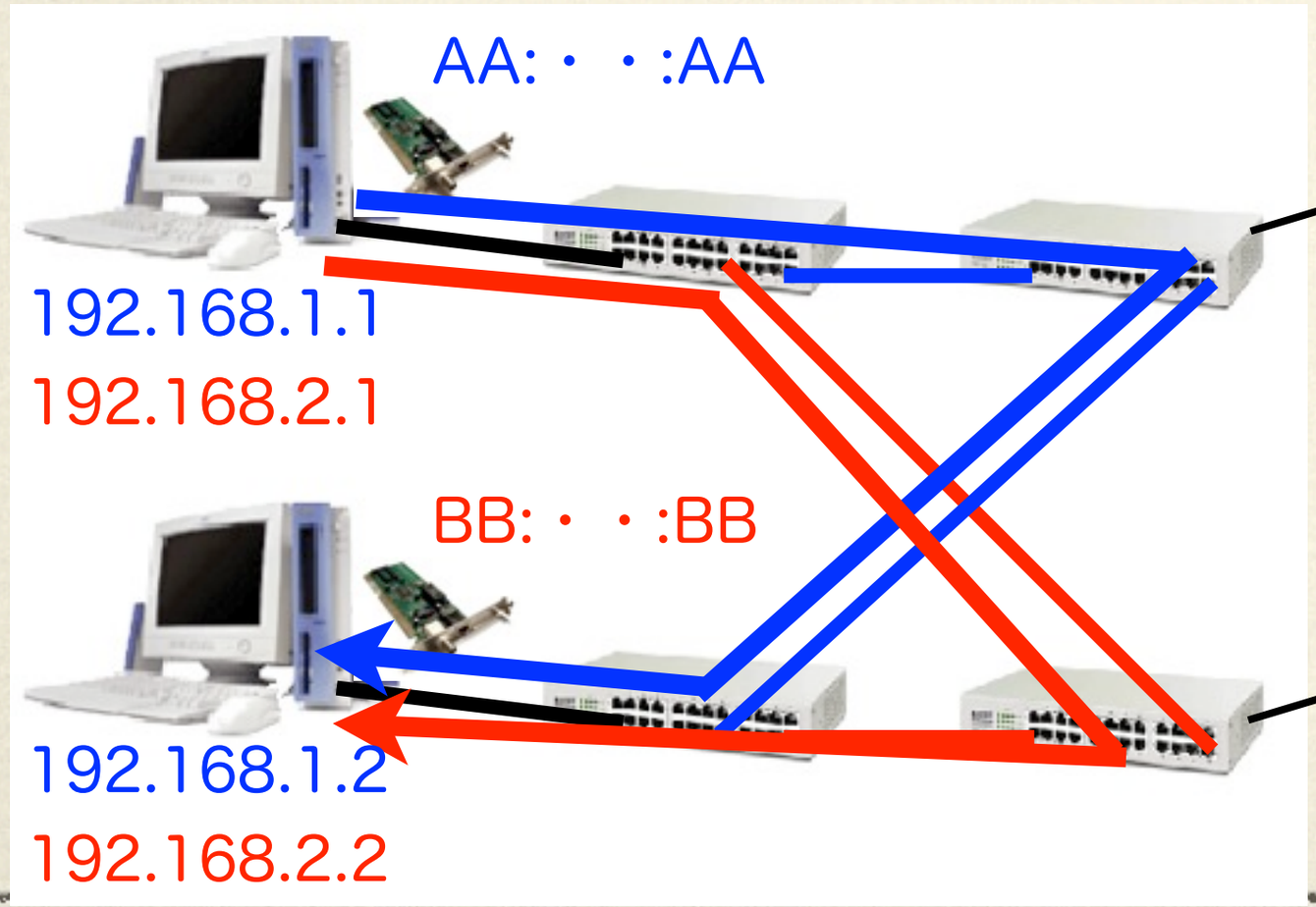
```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```



MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```



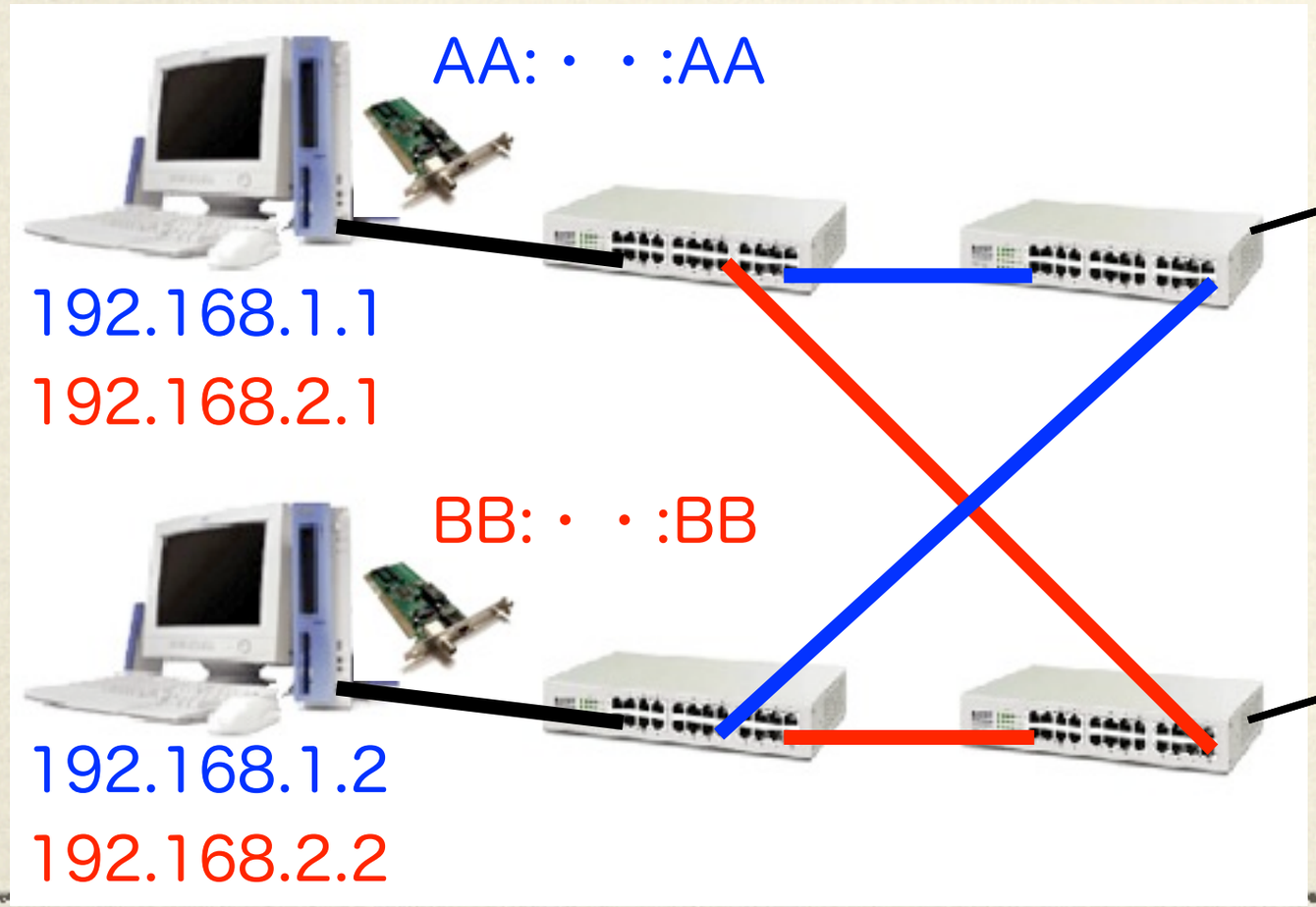
VLAN	port	MACアドレス
1	1	AA:・・:AA

VLAN	port	MACアドレス
2	23	AA:・・:AA

MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```



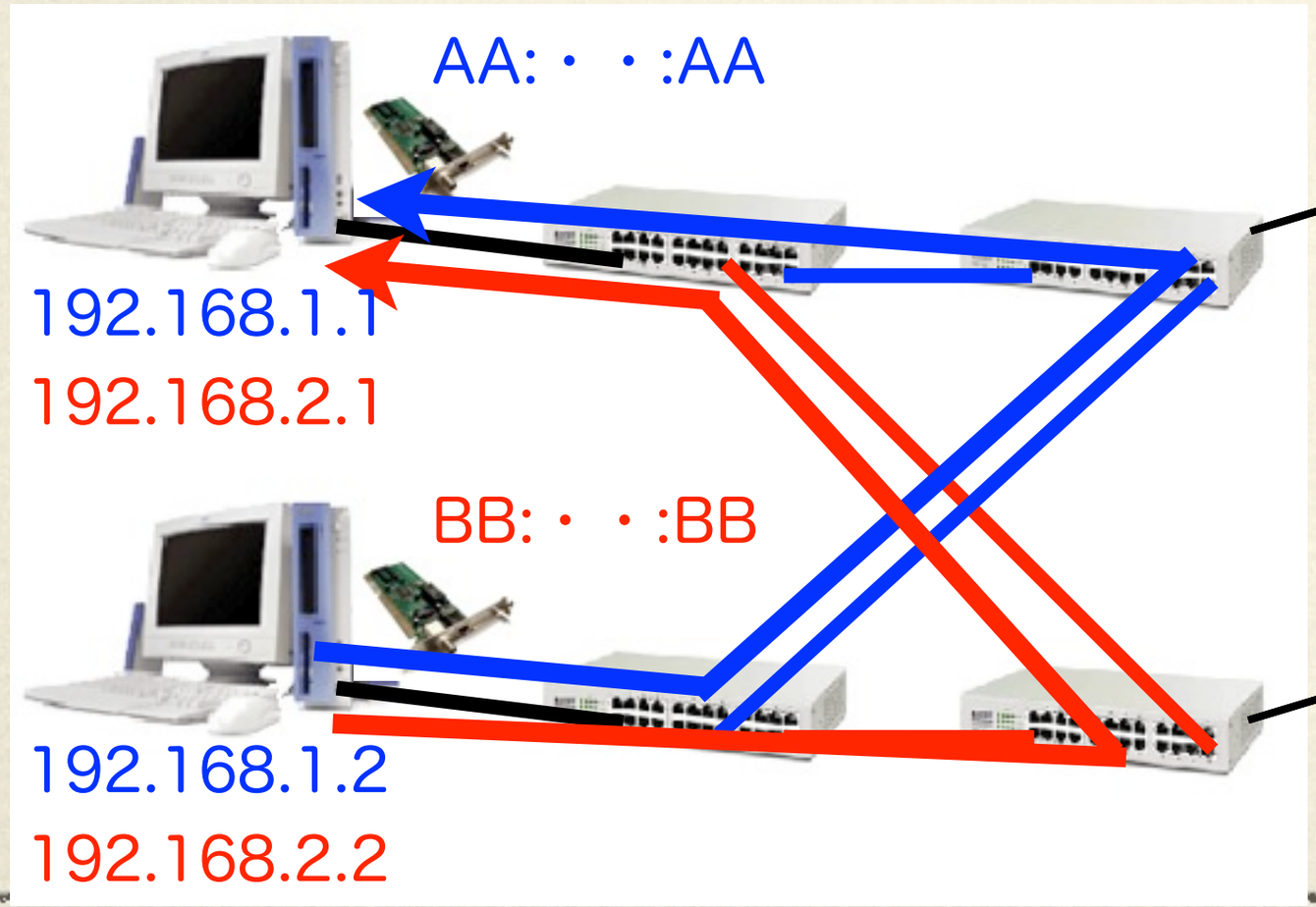
VLAN	port	MACアドレス
1	1	AA:・・:AA

VLAN	port	MACアドレス
2	23	AA:・・:AA

MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```



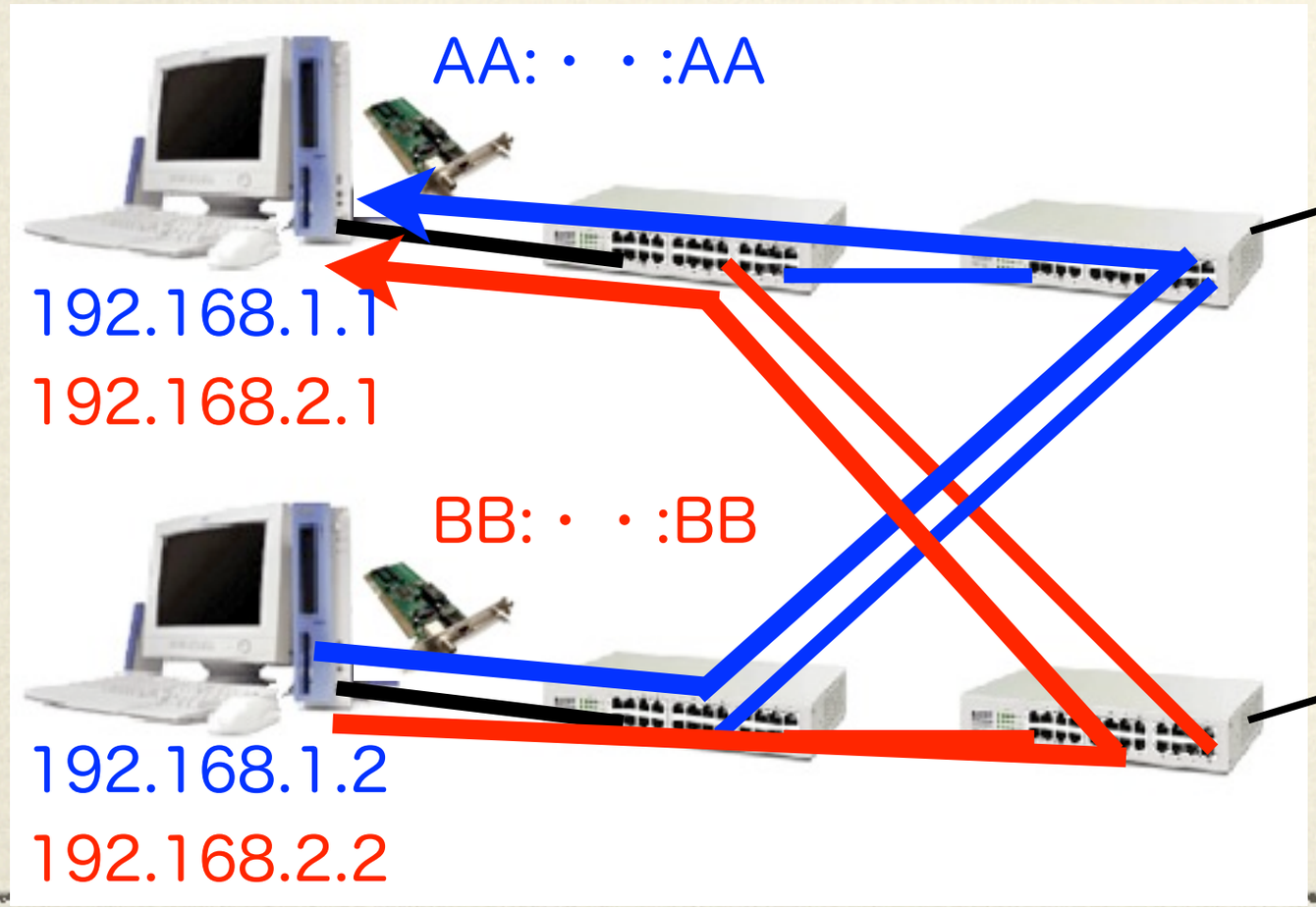
VLAN	port	MACアドレス
1	1	AA:・・:AA

VLAN	port	MACアドレス
2	23	AA:・・:AA

MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```



VLAN	port	MACアドレス
1	1	AA:・・:AA
1	24	BB:・・:BB

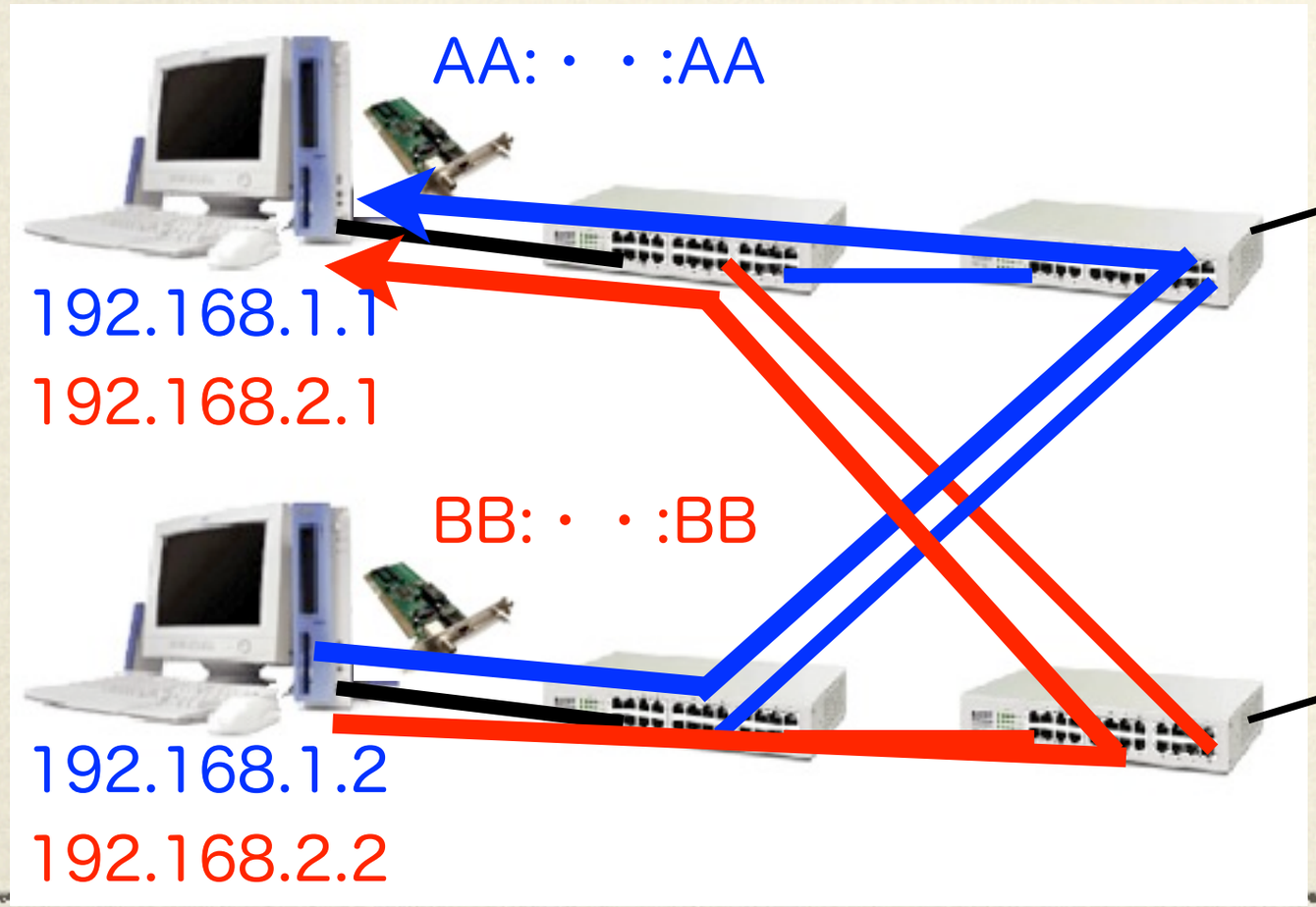
VLAN	port	MACアドレス
2	23	AA:・・:AA
2	2	BB:・・:BB

MACアドレスの管理手順

3. 全ホストの全IPアドレスにブロードキャストを送信

```
$ ping -c1 -w1 -b 192.168.1.255
$ ping -c1 -w1 -b 192.168.2.255
```

このコマンドは1秒で終了する.
cronに登録しておく



VLAN	port	MACアドレス
1	1	AA:・・:AA
1	24	BB:・・:BB

VLAN	port	MACアドレス
2	23	AA:・・:AA
2	2	BB:・・:BB

実験

- 評価環境：2台のPCクラスタを用いる

	Supernova
Year	2003
Nodes	225
CPU	Opteron 1.8GHz × 2
Memory	DDR-SDRAM 2GByte
NIC	Broadcom BCM95704A7 1000BaseT
ピーク性能	1620 GFlops

Dell PowerConnect 6248

特徴：ノード数が多い

実験

- 評価環境：2台のPCクラスタを用いる

	Misc
Year	2008
Nodes	66
CPU	Quad Opteron 2.3GHz × 2
Memory	DDR2-SDRAM 8GByte
NIC	Broadcom BCM95721 1000BaseT × 2
ピーク性能	4868 GFlops

特徴：8コア/Node,　NICが2枚

Dell PowerConnect 6248

実験

- 評価環境：2台のPCクラスタを用いる

	Misc
Year	2008
Nodes	66
CPU	Quad Opteron 2.3GHz × 2
Memory	DDR2-SDRAM 8GByte
NIC	Broadcom BCM95721 1000BaseT × 2
ピーク性能	4868 GFlops

192.168.1.1

異なるVLANのIP addrを設定

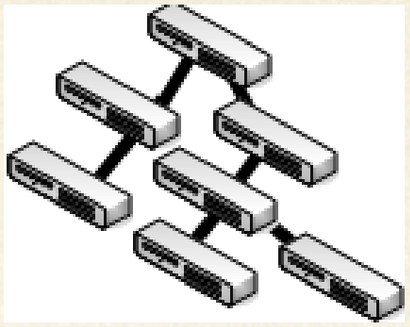
192.168.2.1

特徴：8コア/Node,　NICが2枚

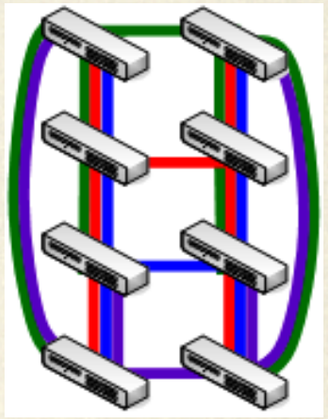
Dell PowerConnect 6248

ネットワークトポロジ

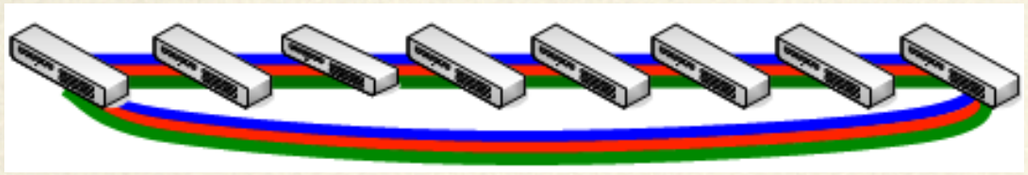
- Tree　（VLAN 使用せず）
- 4×2 Torus　（VLAN 4個使用）
- 完全結合　（VLAN 8個使用）
- 4×2 Mesh　（VLAN 4個使用）
- 8×1 Ring　（VLAN 8個使用）



Tree

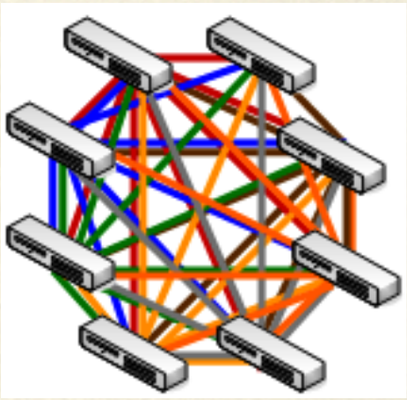


Torus



Ring

Mesh

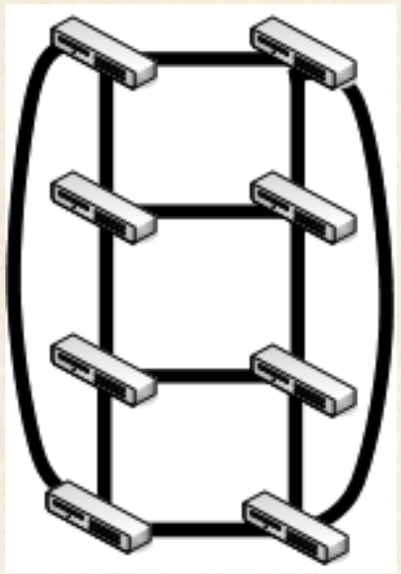


完全結合

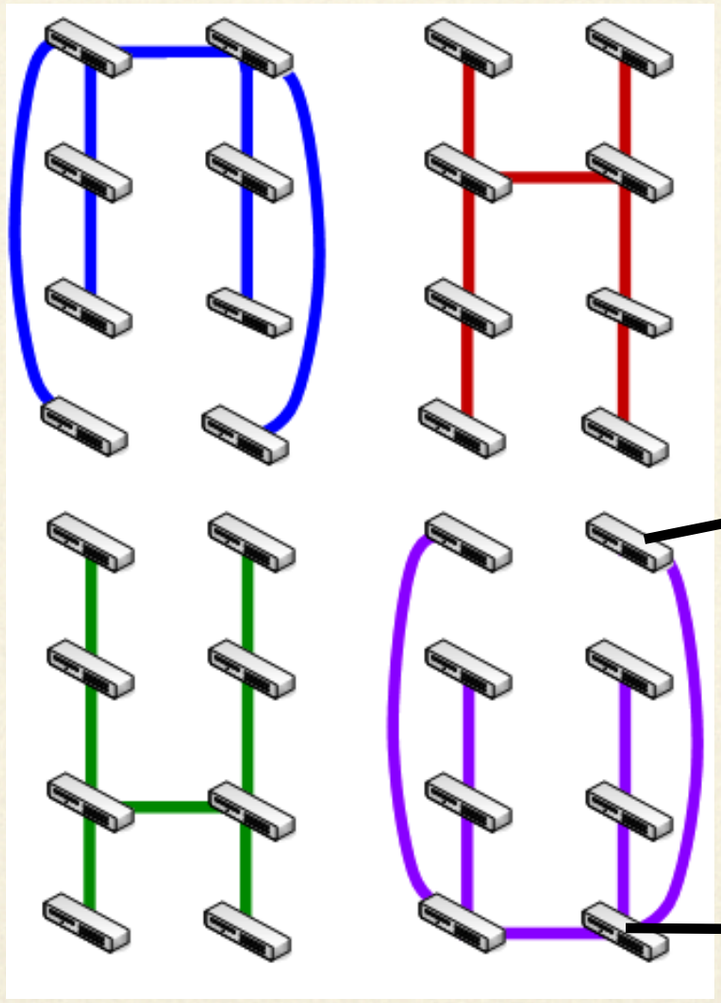
スイッチのリンクアグリゲーション機能を使用して,
各トポロジのスイッチ間リンクを 1～ 8本にして評価を行った

Torusトポロジ詳細

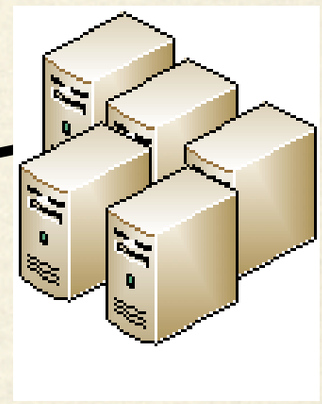
- ４ × ２ Torusトポロジ

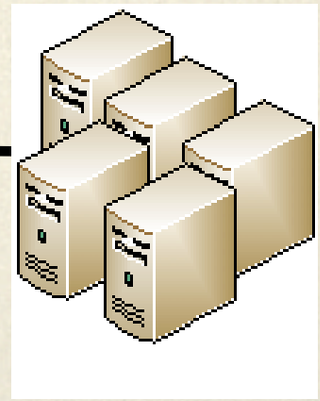


物理配線



各スイッチのホスト数 =
全ホスト数 / スイッチ数

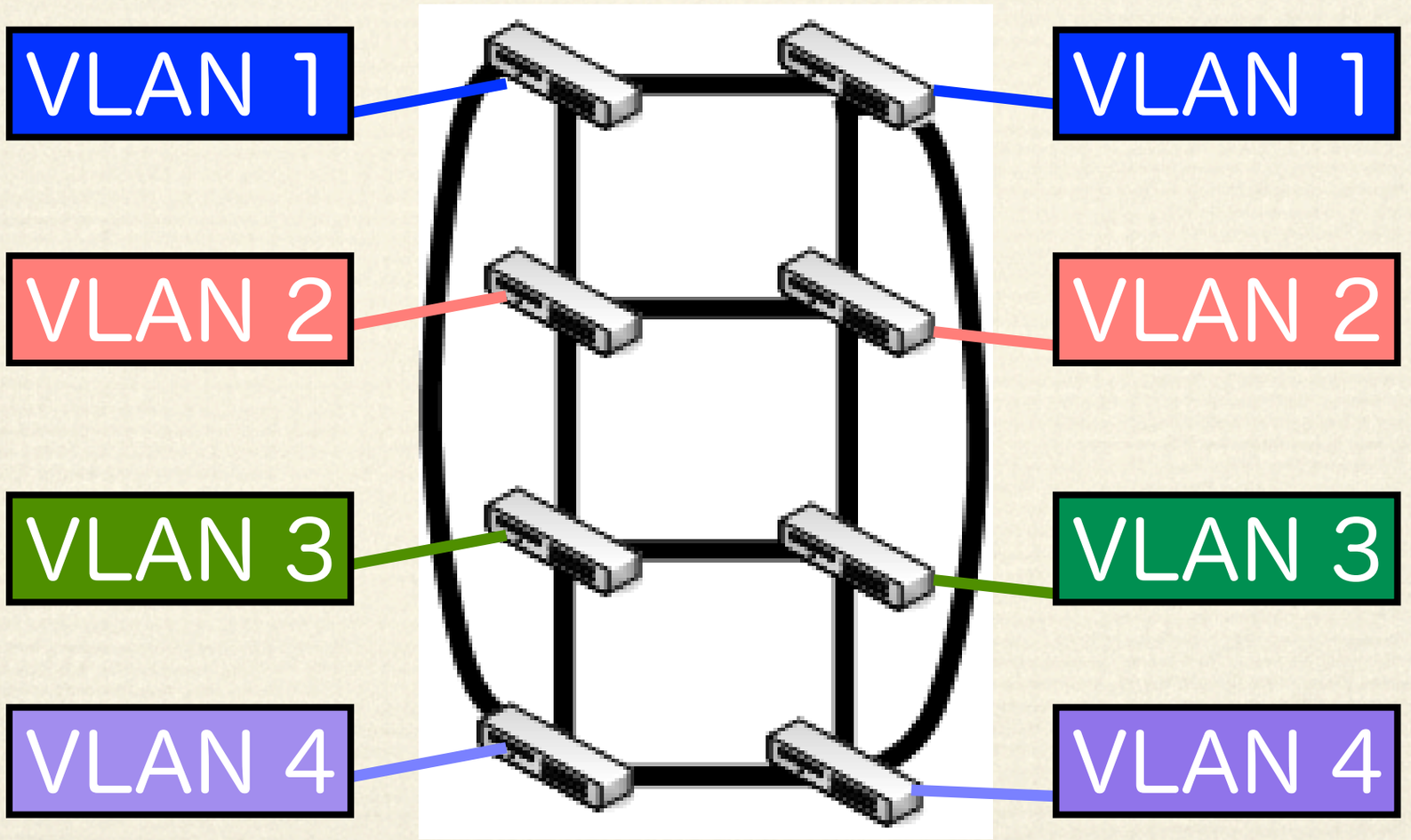




4つの論理的なトポロジが存在する

Torusトポロジ詳細

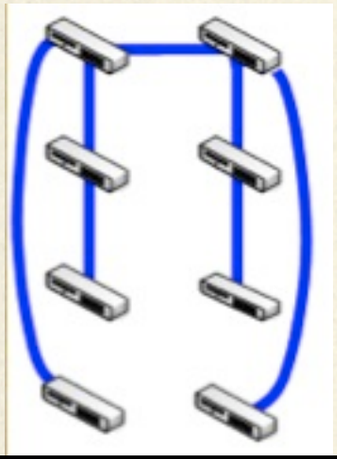
- 4 × 2 Torusトポロジ



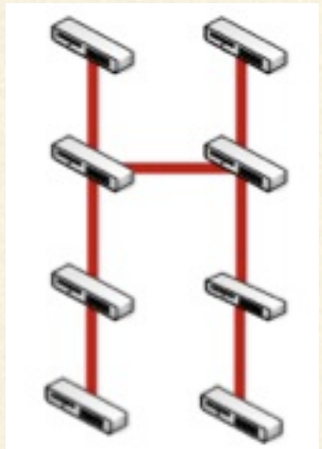
各スイッチに接続したホストからの入力フレームに,
それぞれのスイッチで設定したVLANタグを挿入する

Torusトポロジ詳細

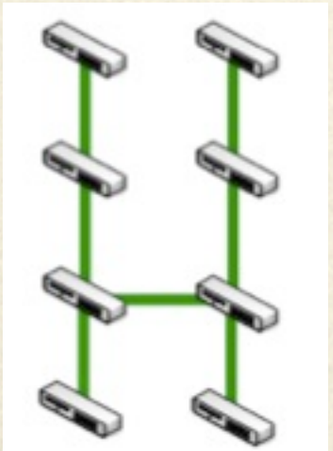
- ４ × ２ Torusトポロジ



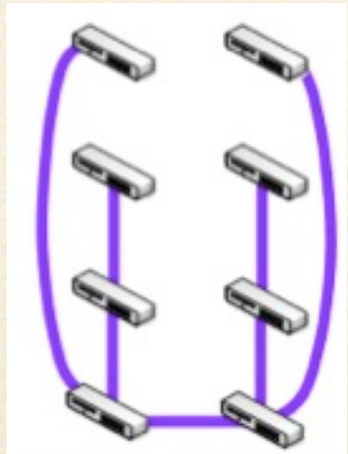
VLAN 1



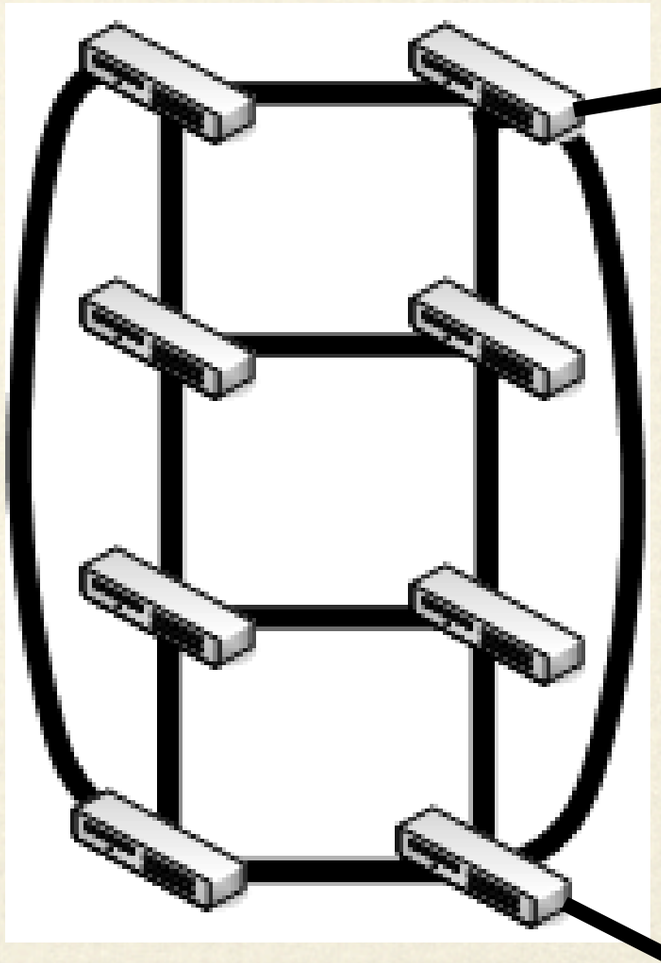
VLAN 2

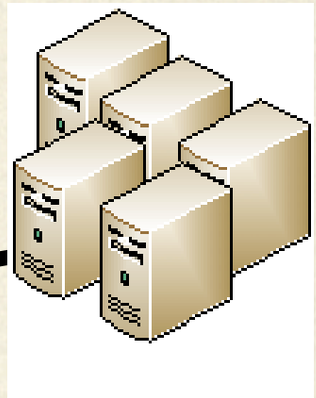


VLAN 3

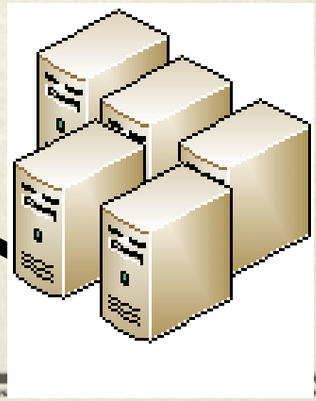


VLAN 4



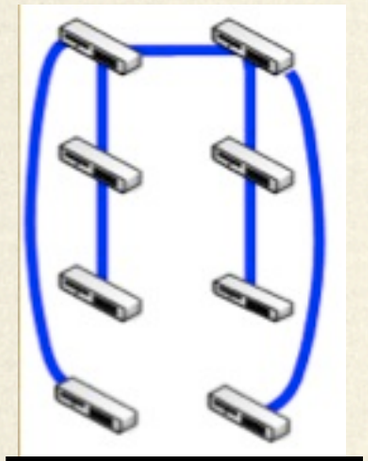


VLAN 1

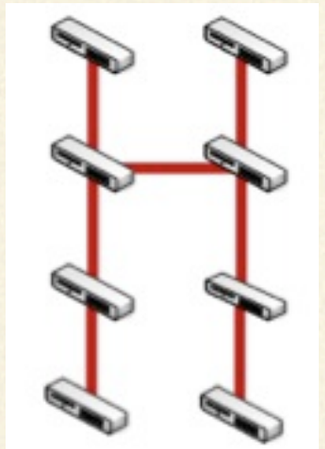


Torusトポロジ詳細

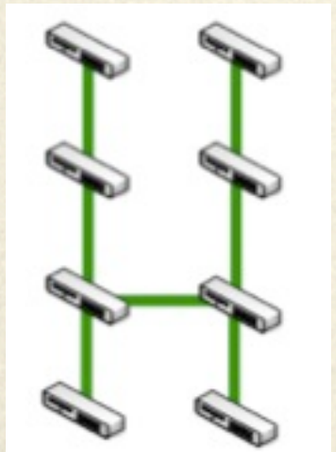
- ４ × ２ Torusトポロジ



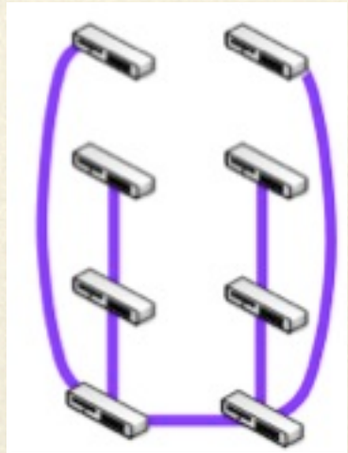
VLAN 1



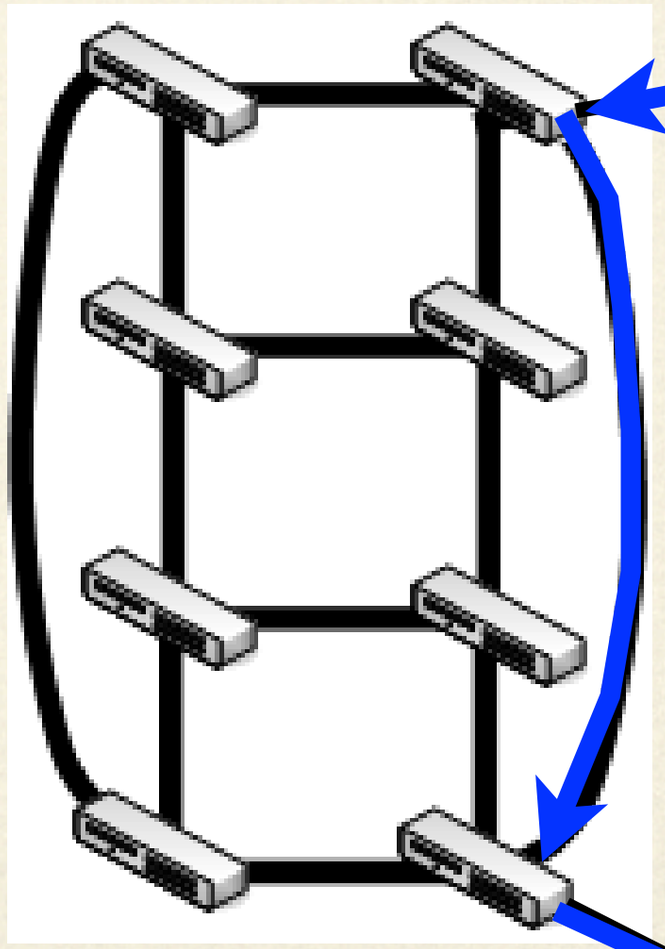
VLAN 2

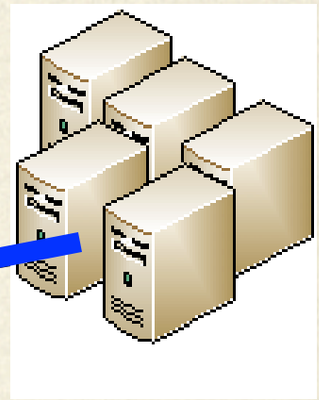


VLAN 3

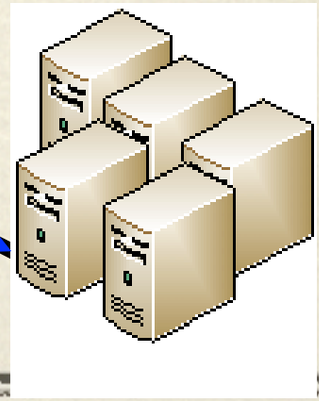


VLAN 4



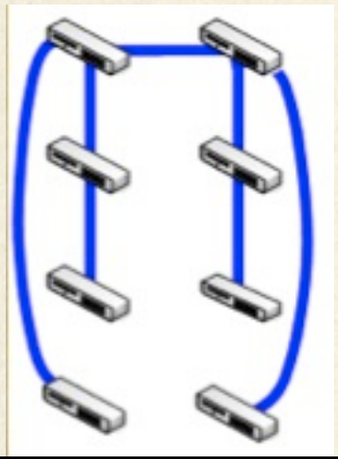


VLAN 1

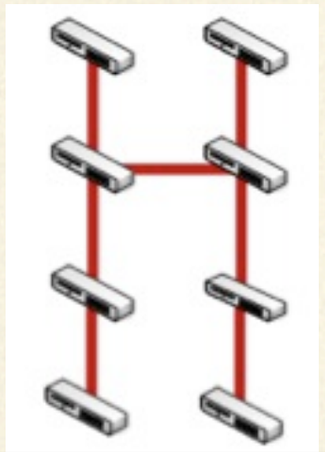


Torusトポロジ詳細

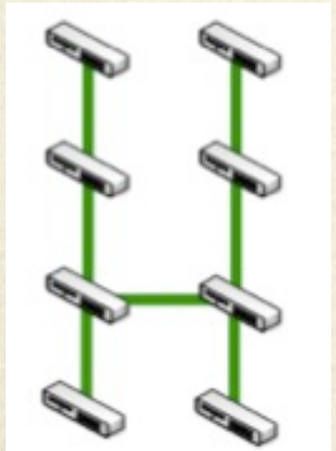
- ４ × ２ Torusトポロジ



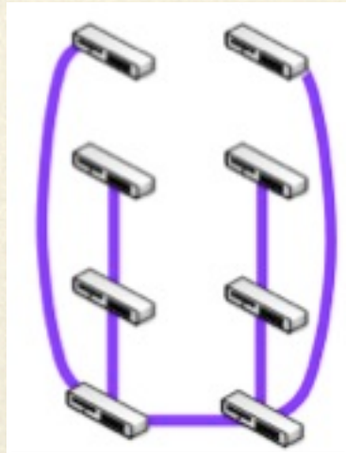
VLAN 1



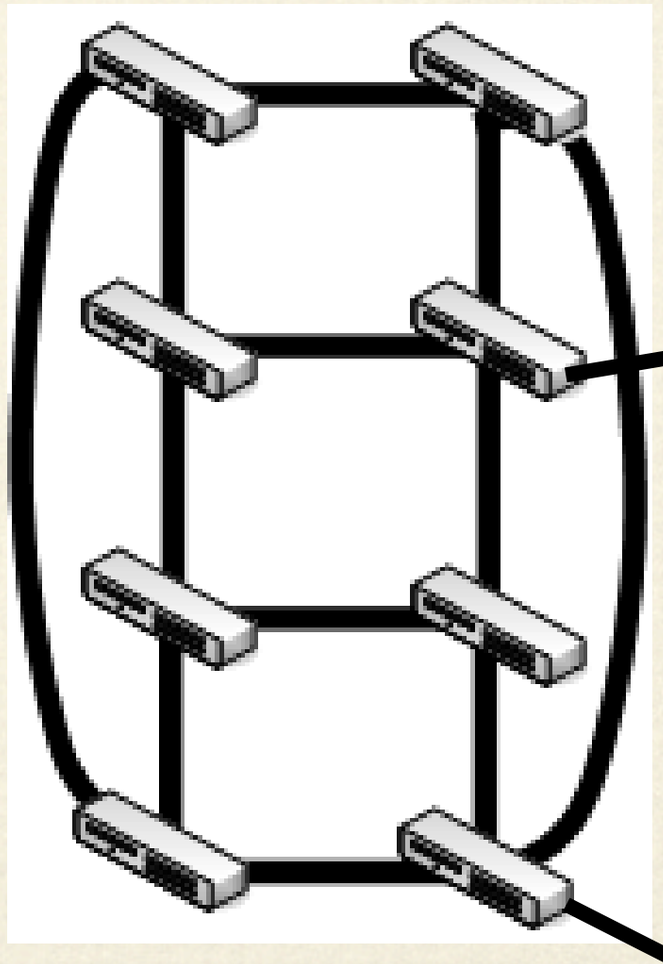
VLAN 2

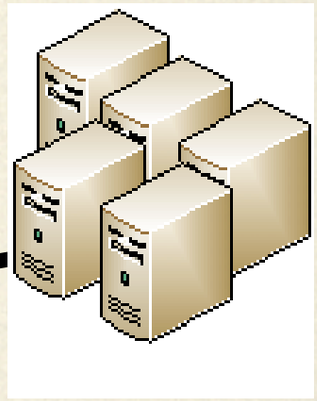


VLAN 3

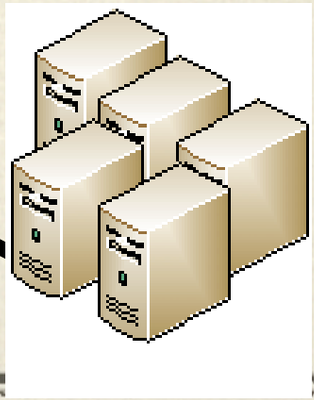


VLAN 4



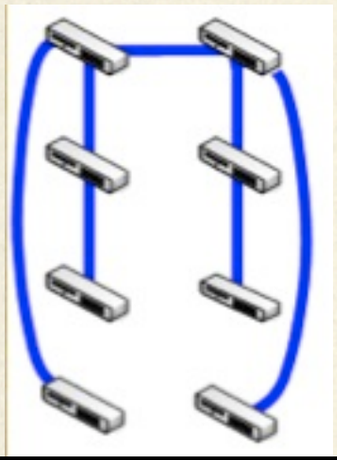


VLAN 2

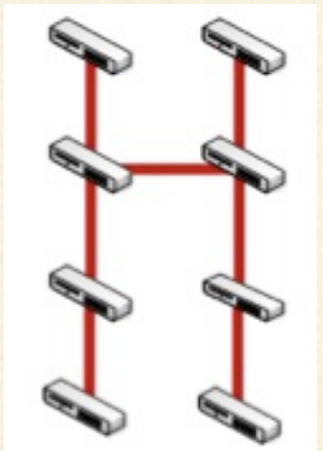


Torusトポロジ詳細

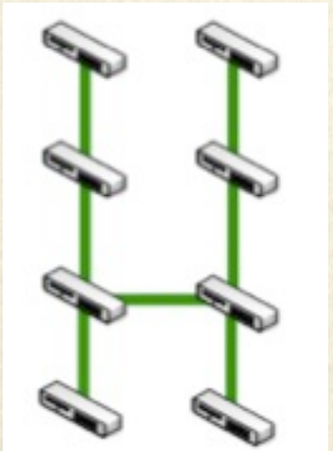
- ４ × ２ Torusトポロジ



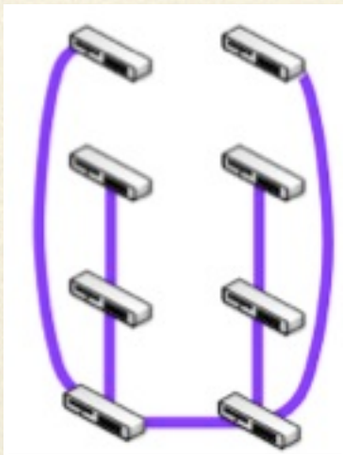
VLAN 1



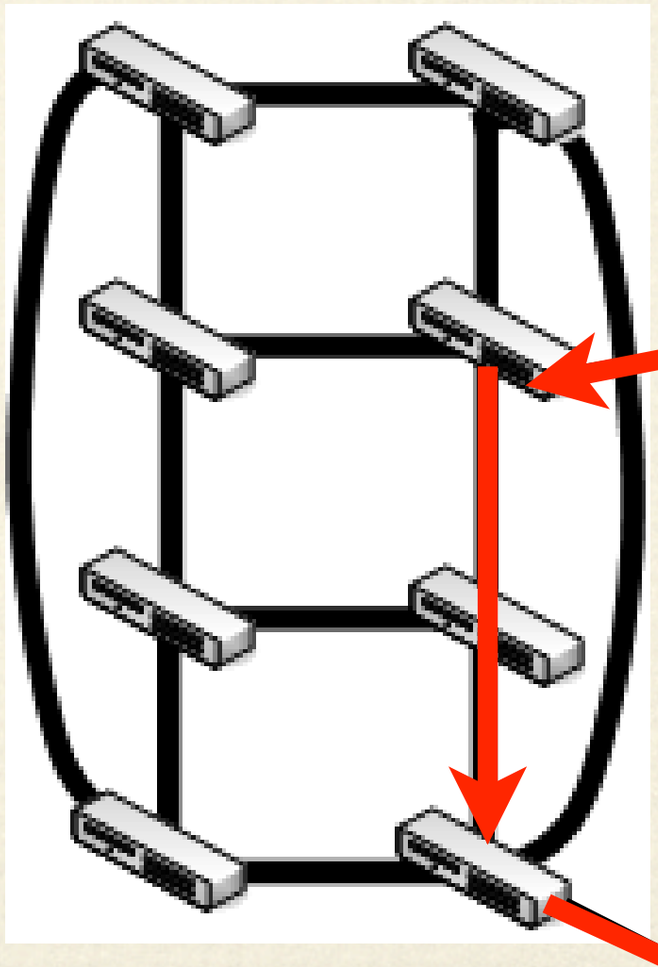
VLAN 2

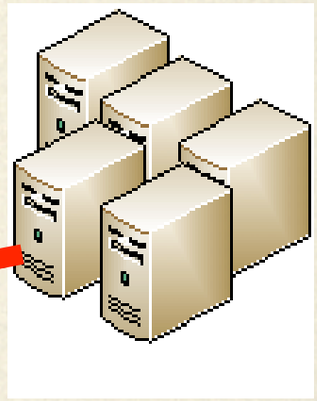


VLAN 3

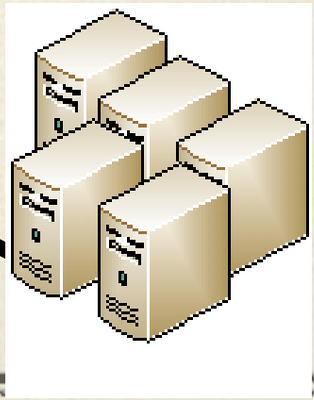


VLAN 4





VLAN 2



実験環境構築風景

評価に用いるベンチマーク

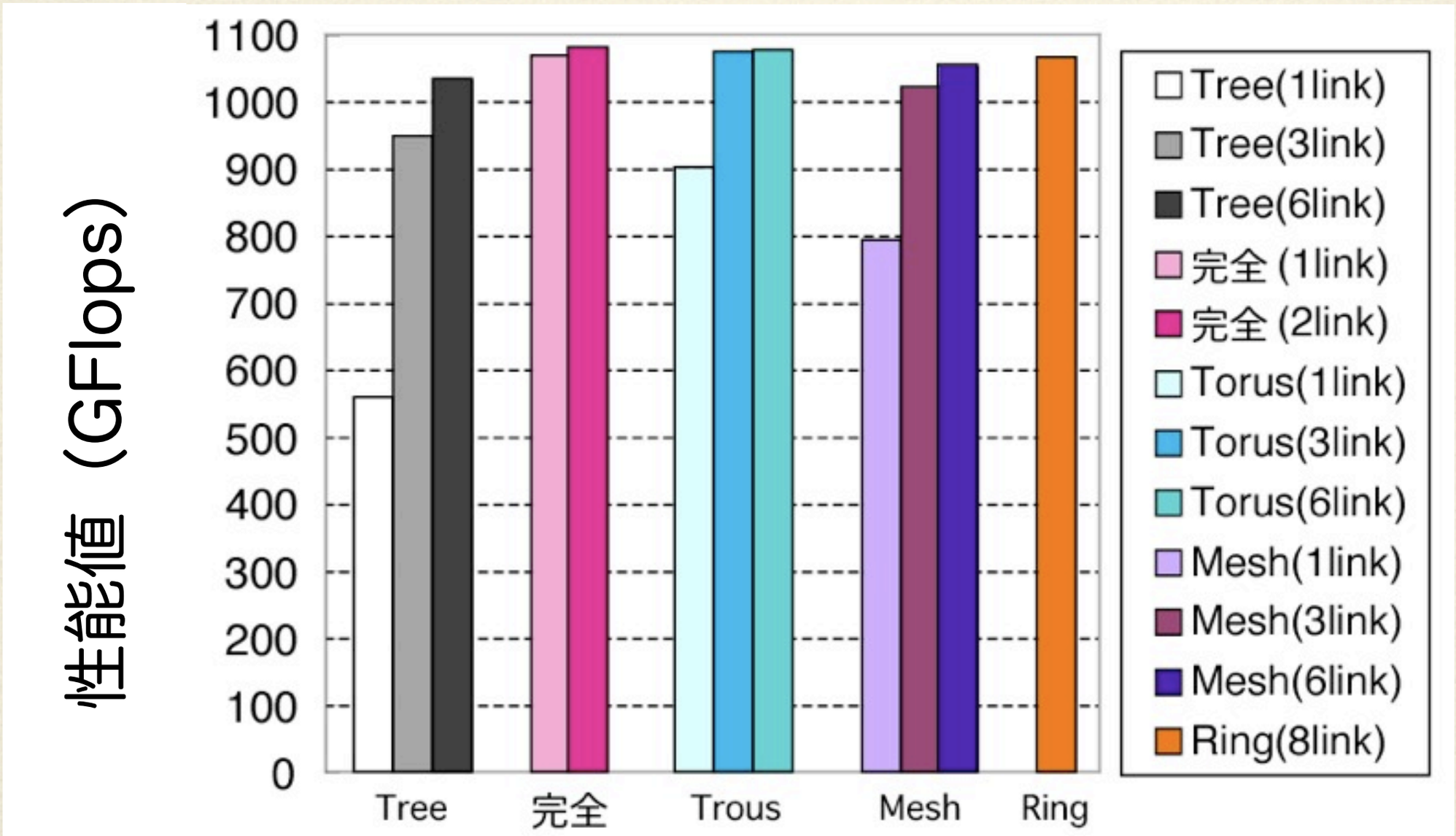
- High Performance LINPACK Benchmark（HPL）
  - Top500公式ベンチマーク
  - 密行列連立一次方程式をガウス消去法で解く速度を測定

- NAS Parallel Benchmarks（NPB）
  - NASAが開発した並列計算機用ベンチマーク
  - 航空関連の流体シミュレーションの実行性能評価
  - 8つのベンチマークから構成

結果（HPL）

- Supernova(225ホスト)の結果

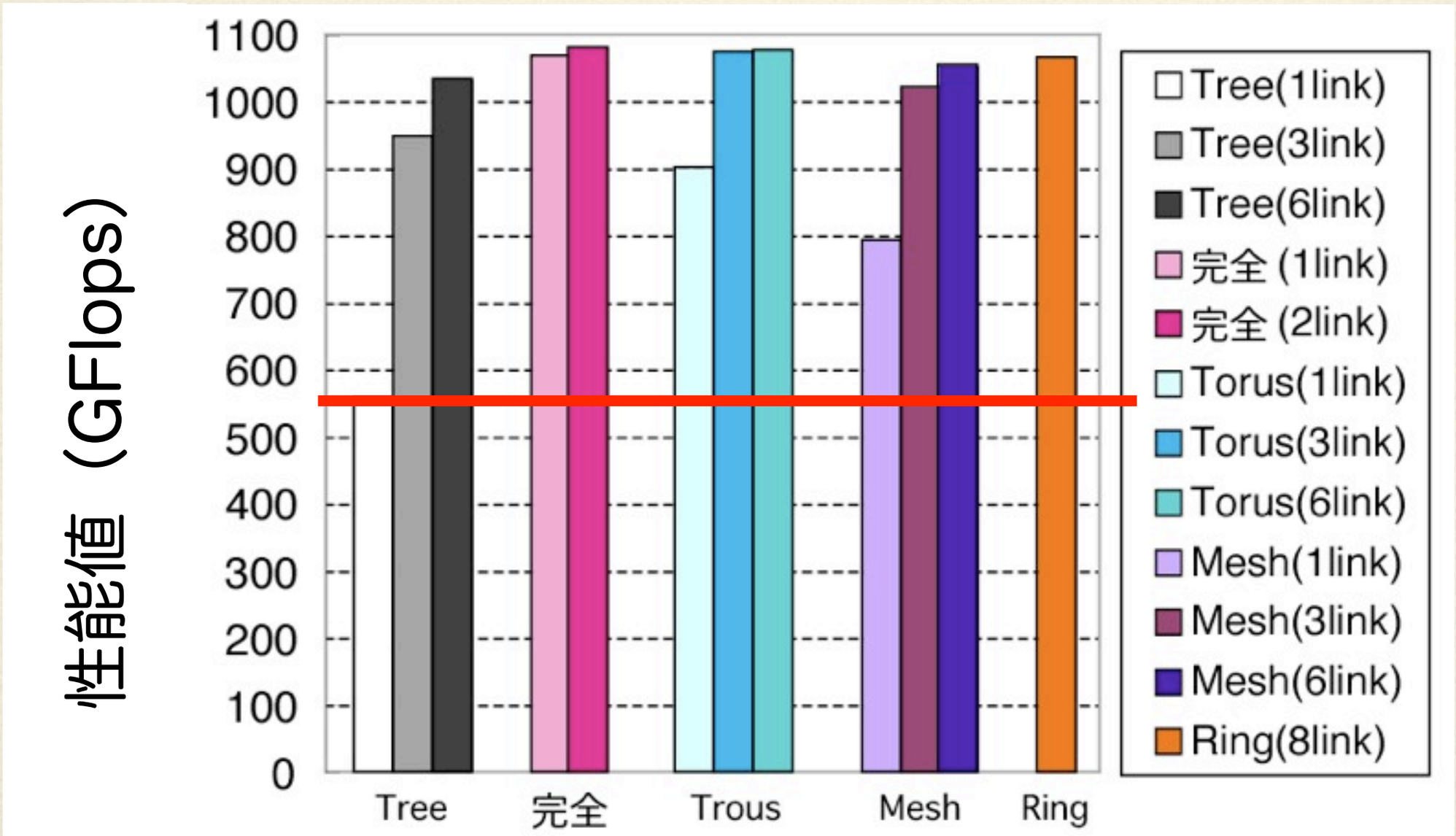
Tree(1link)と比較して，すべて性能が向上した



結果（HPL）

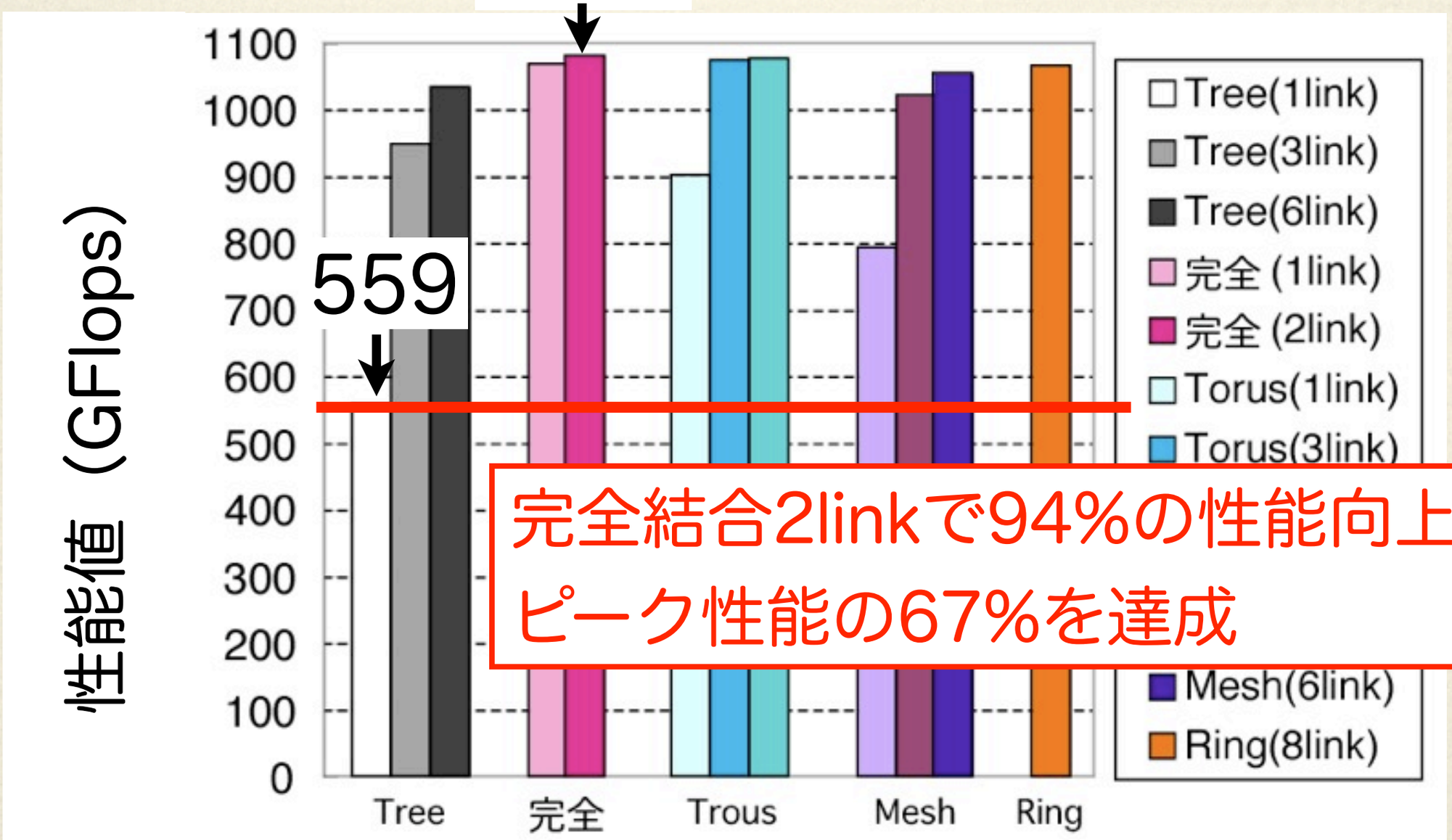
- Supernova(225ホスト)の結果

Tree(1link)と比較して，すべて性能が向上した



結果（HPL）

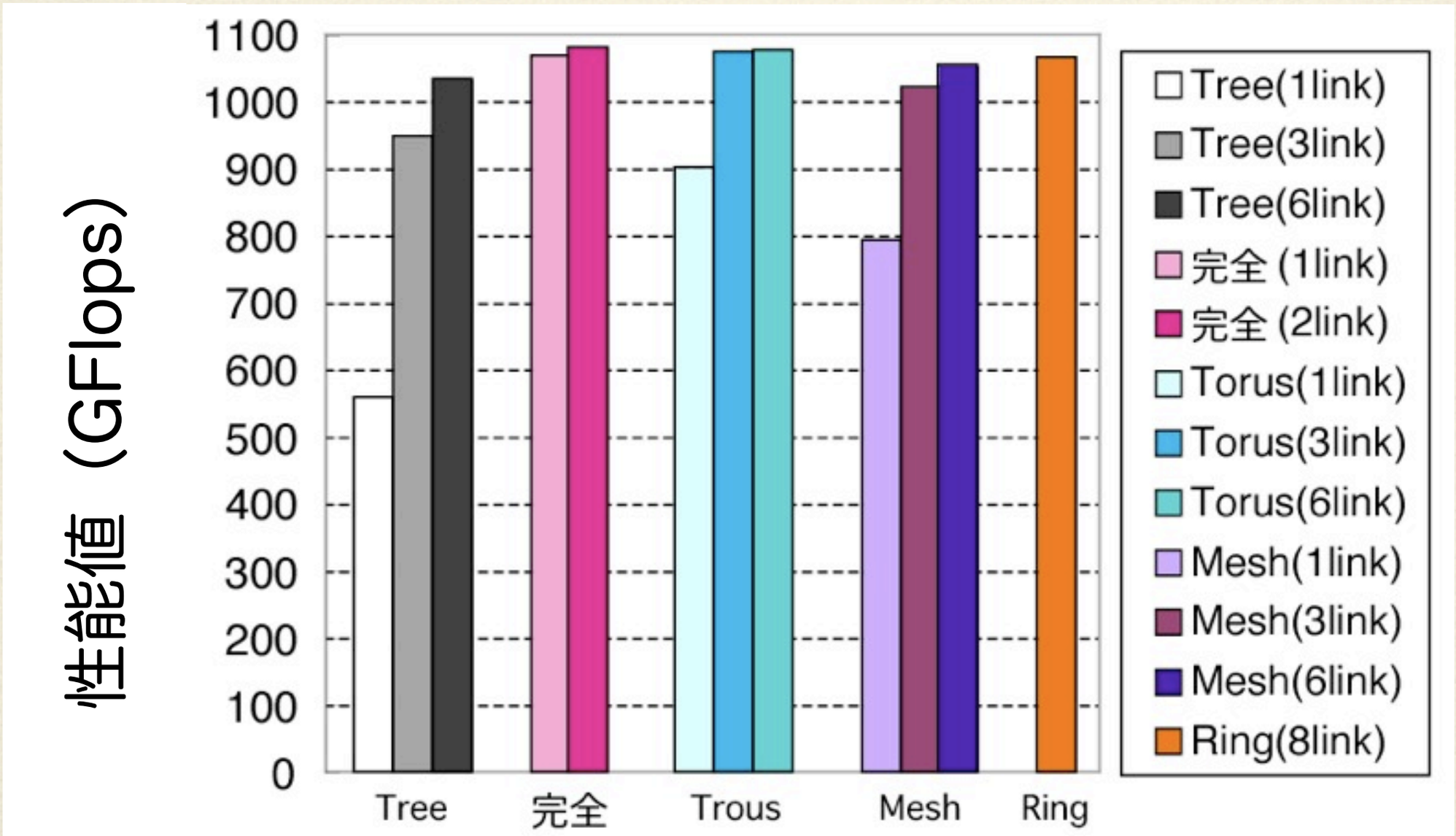
- Supernova(225ホスト)の結果



結果（HPL）

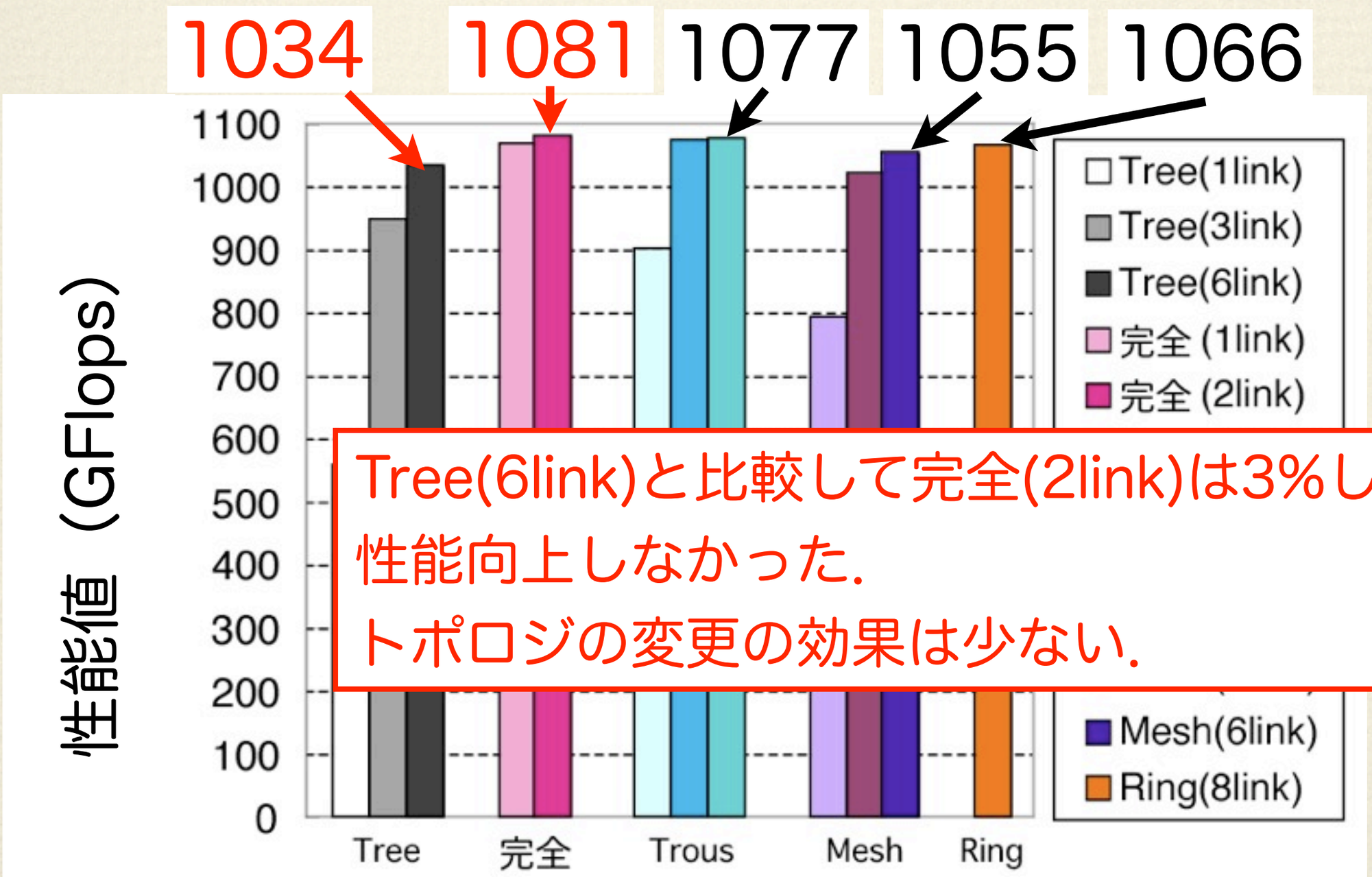
- Supernova(225ホスト)の結果

スイッチ間のリンク集約化が効果的



結果（HPL）

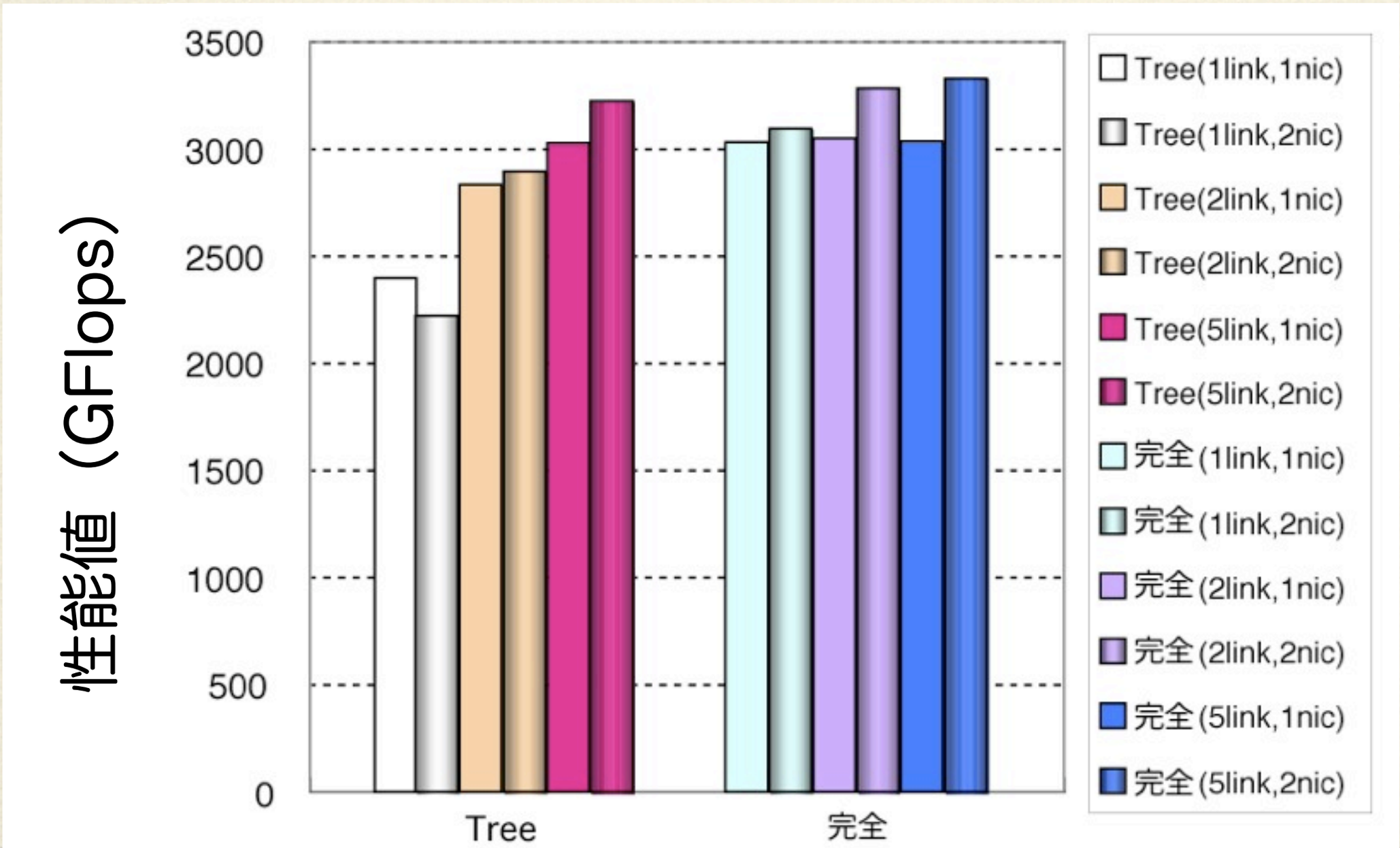
- Supernova(225ホスト)の結果



Tree(6link)と比較して完全(2link)は3%しか性能向上しなかった.
トポロジの変更の効果は少ない.

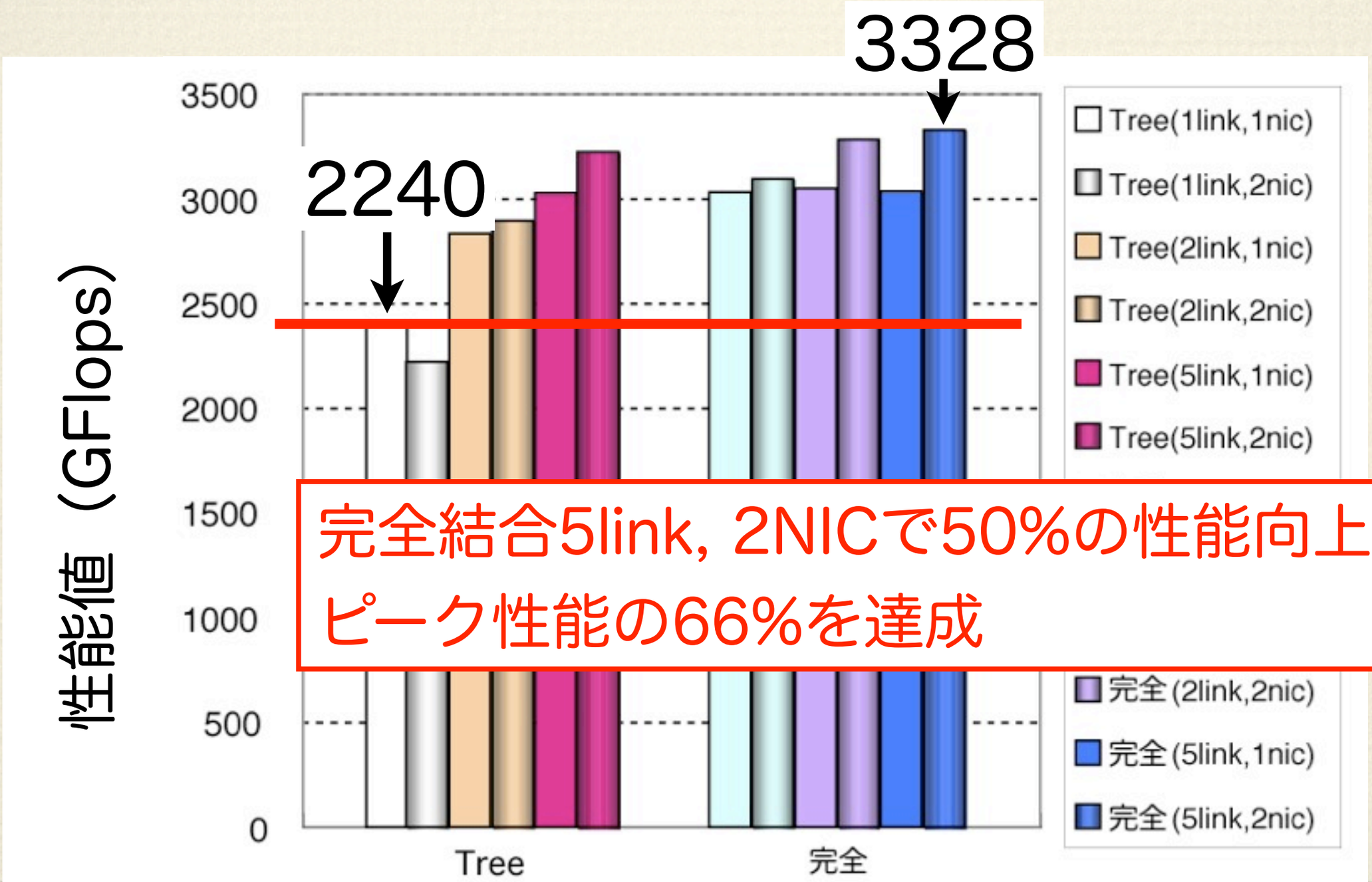
結果（HPL）

- Misc (66ホスト) の結果（Treeと完全結合のみ）



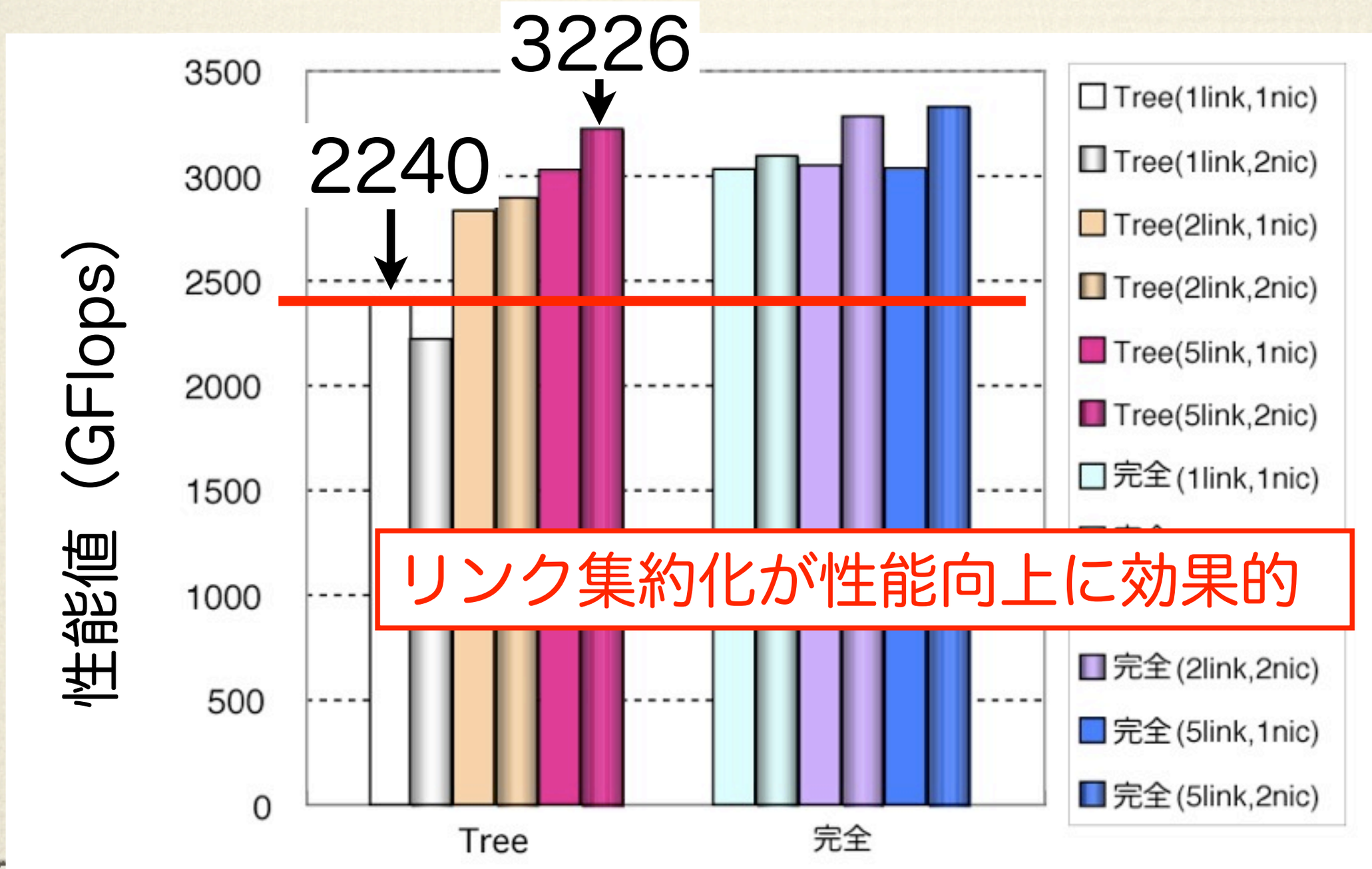
結果（HPL）

- Misc (66ホスト) の結果（Treeと完全結合のみ）



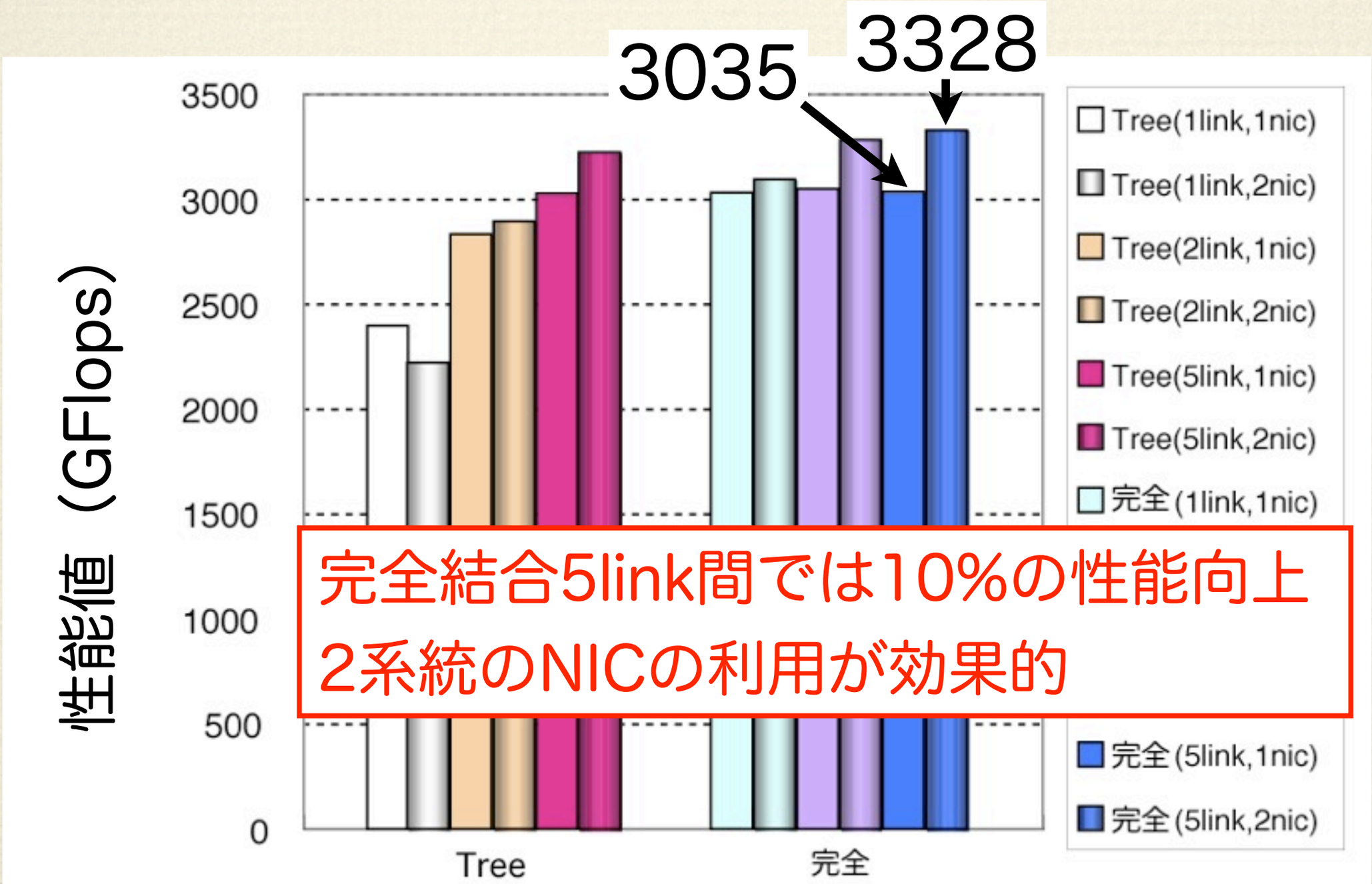
結果（HPL）

- Misc (66ホスト) の結果（Treeと完全結合のみ）



結果（HPL）

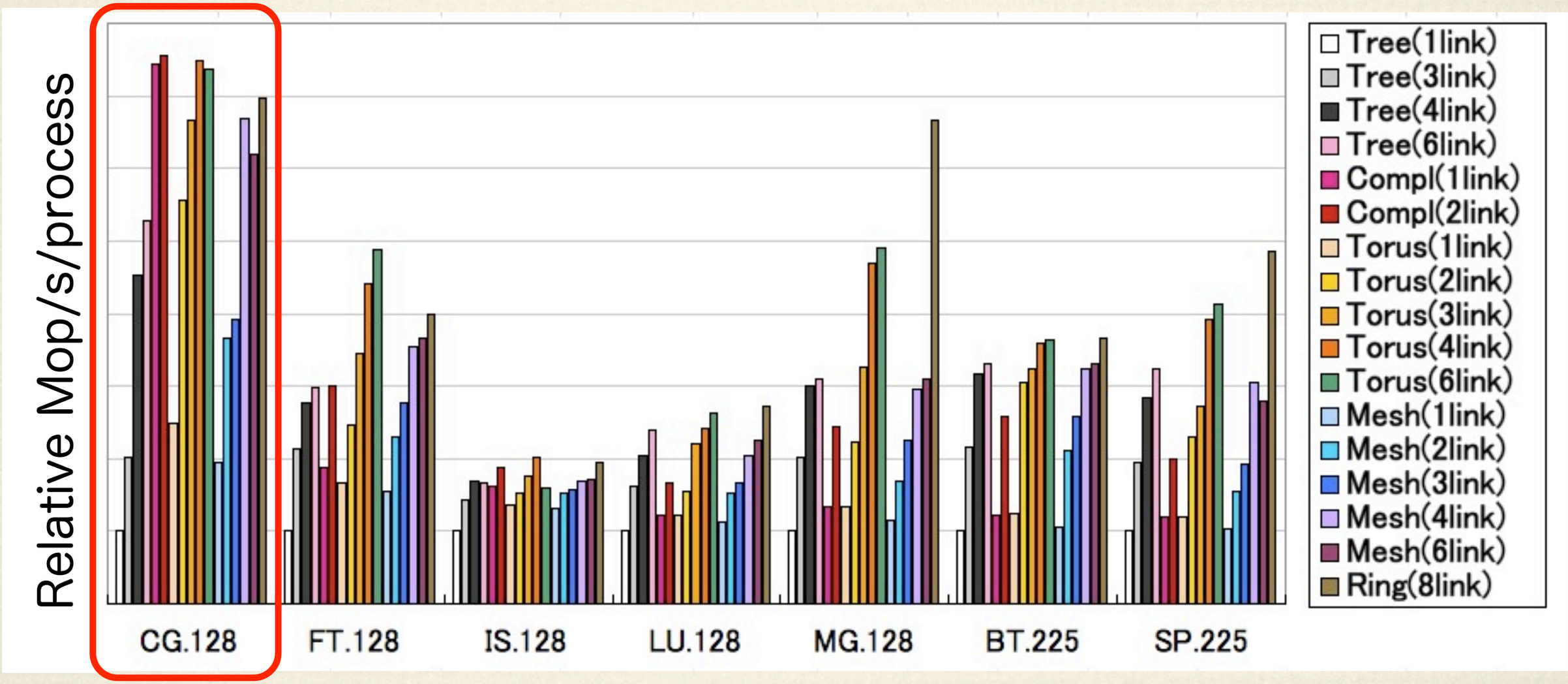
- Misc (66ホスト) の結果（Treeと完全結合のみ）



完全結合5link間では10%の性能向上
2系統のNICの利用が効果的

結果（NPB）

- Supernova (225ホスト) の結果

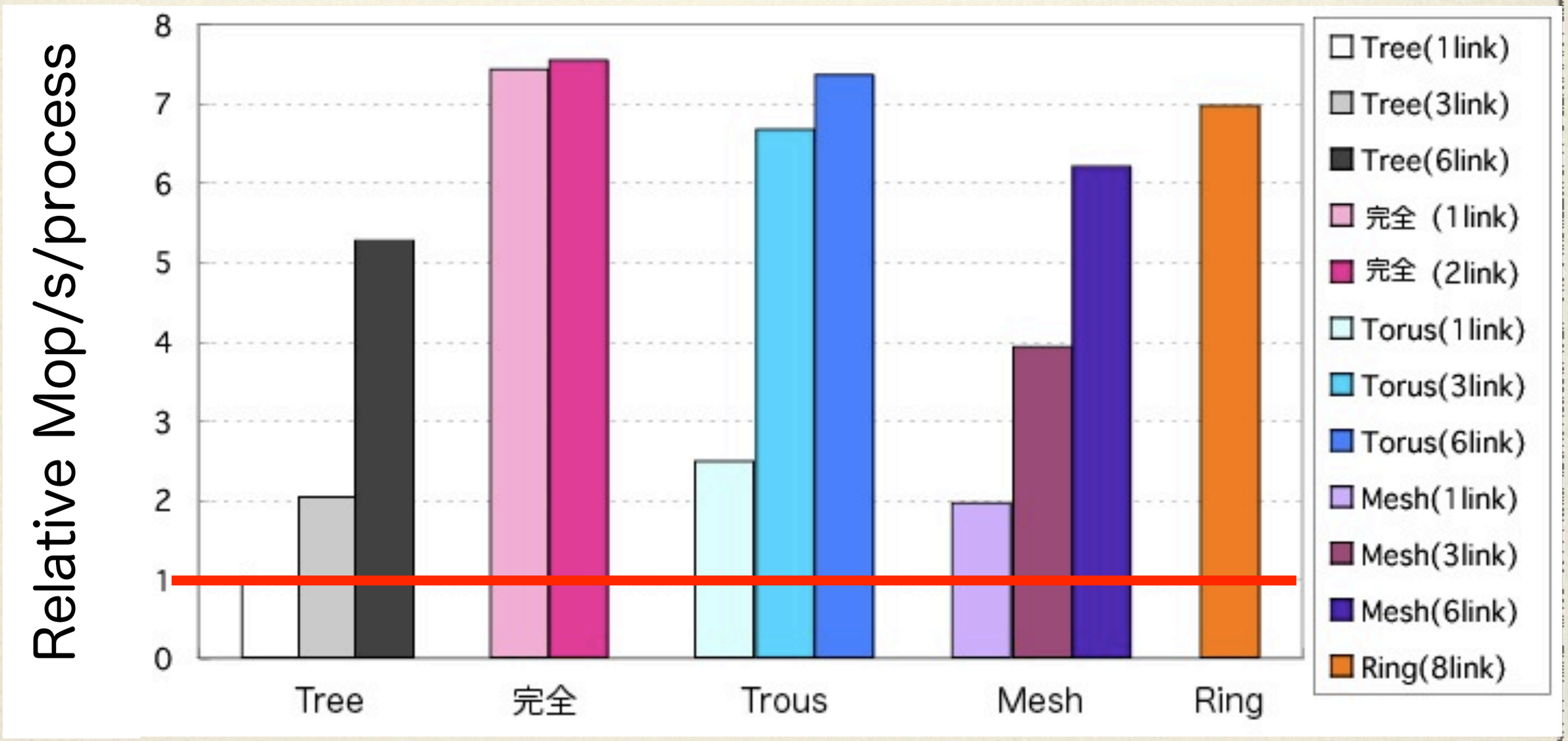


他のベンチマークについては論文集を参考にして下さい

結果（NPB：CG法）

- Supernova (225ホスト) の結果

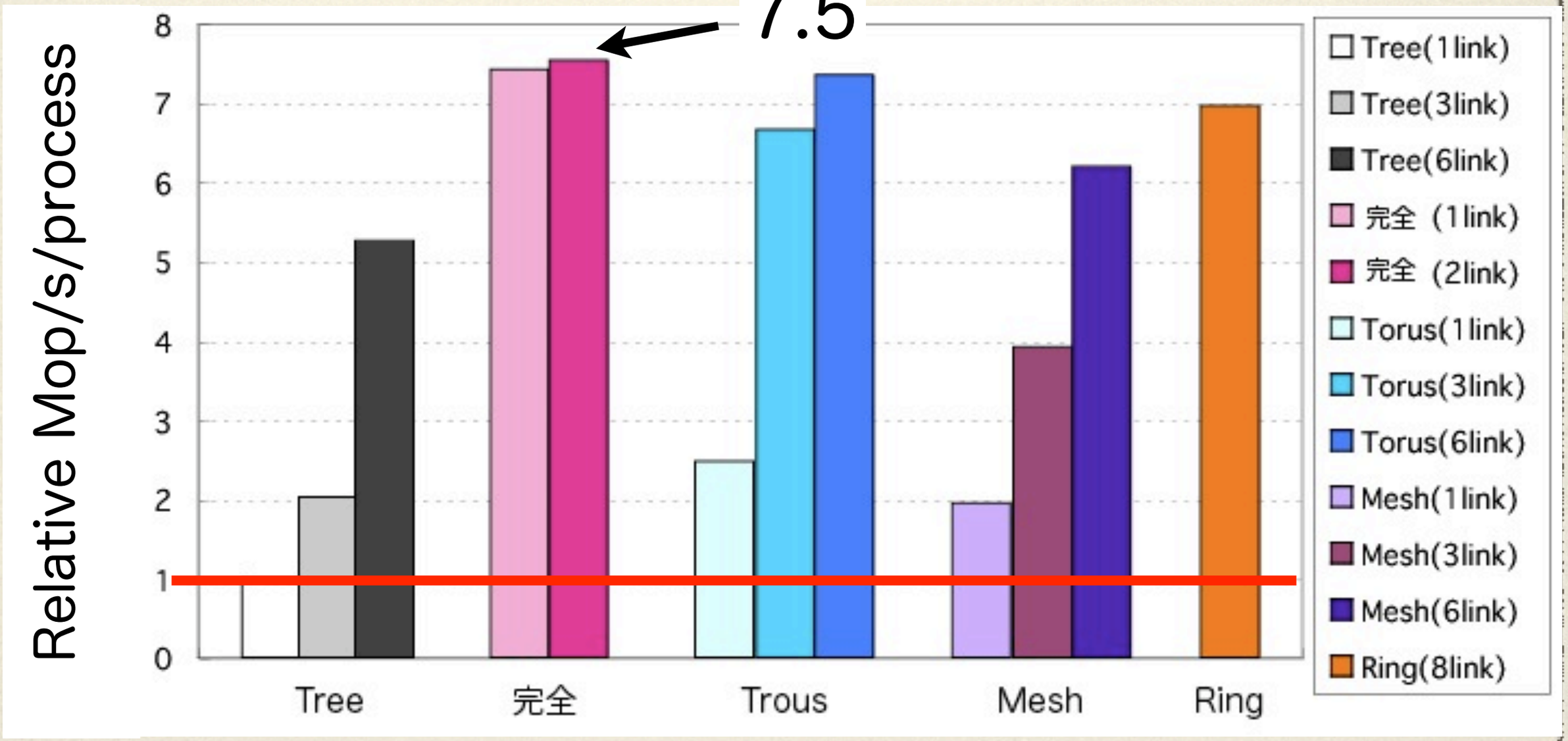
Tree(1link)と比較して，すべて性能が向上した



結果（NPB：CG法）

- Supernova (225ホスト) の結果

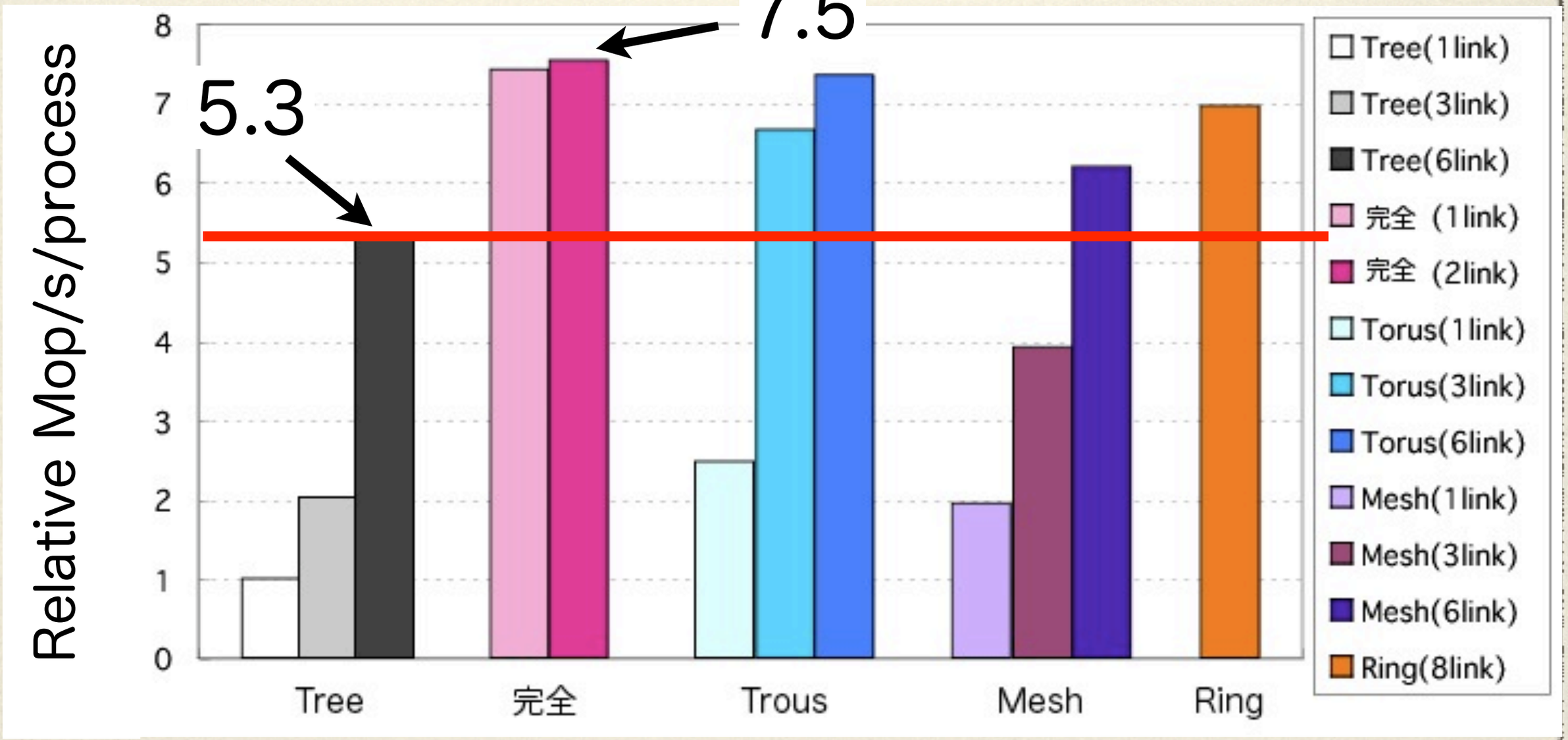
Tree(1link)に比べて完全(2link)では650%の性能向上を達成



結果（NPB：CG法）

- Supernova (225ホスト) の結果

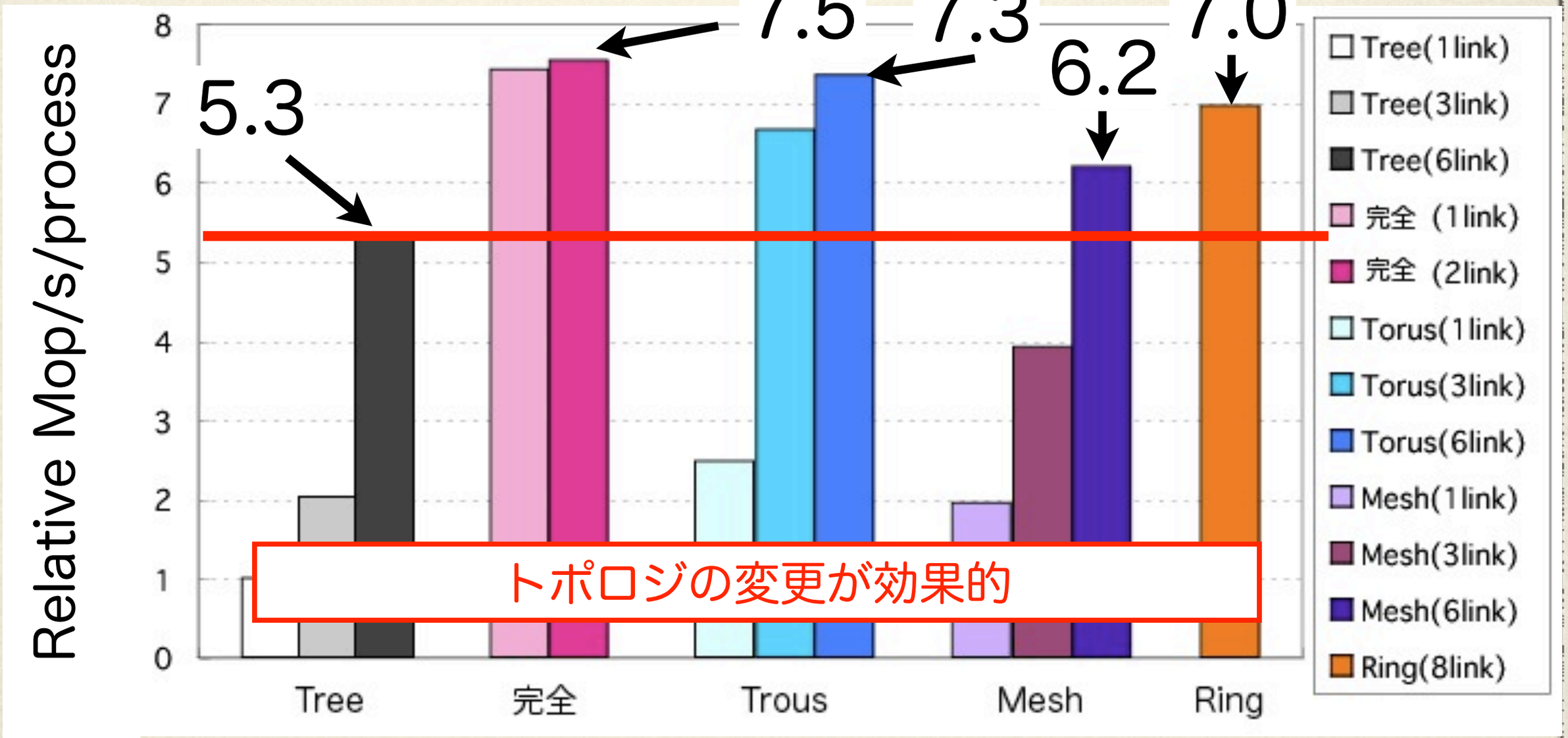
Tree(6link)に比べて完全(2link)では43%の性能向上を達成



結果（NPB：CG法）

- Supernova (225ホスト) の結果

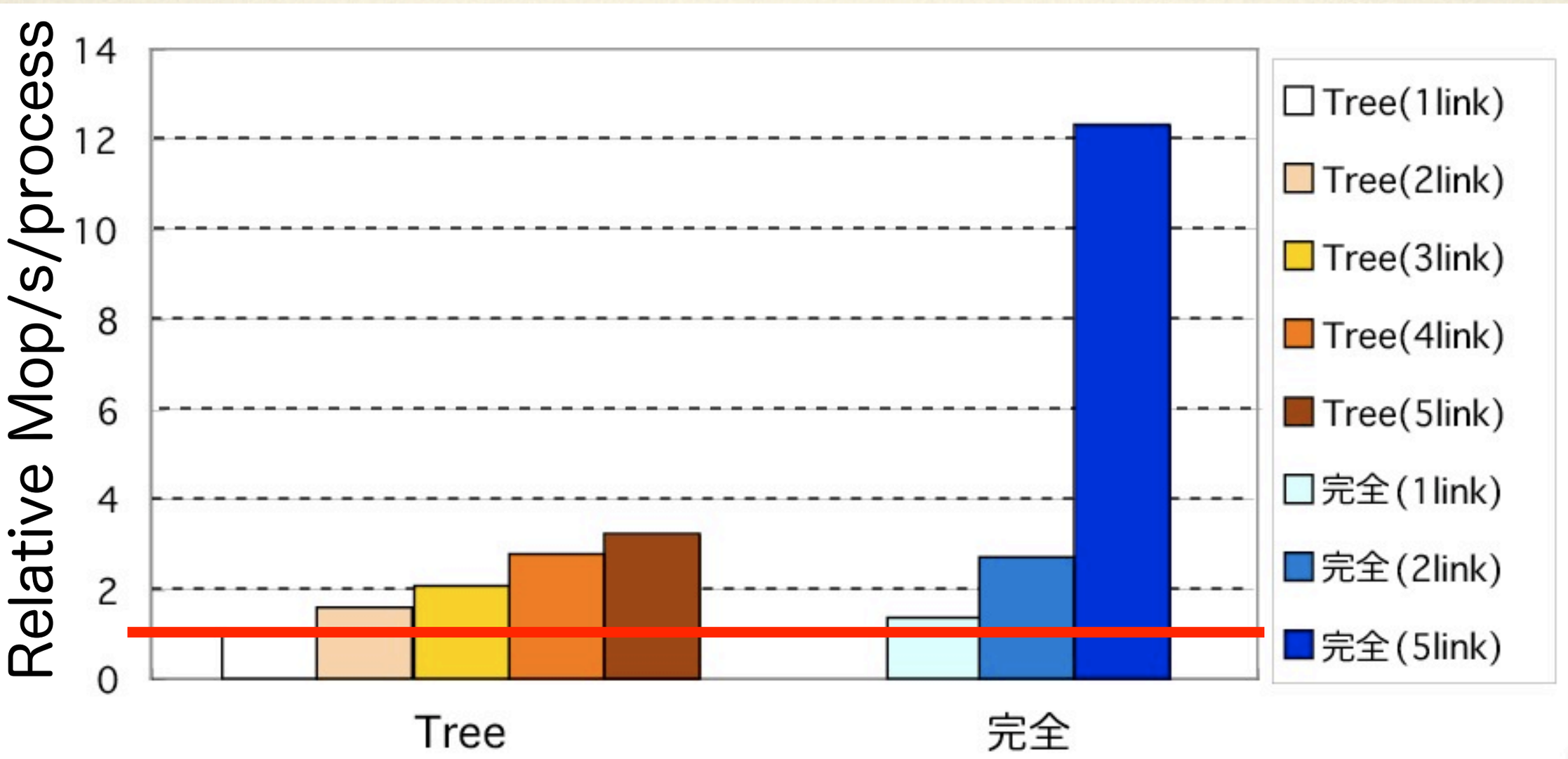
Tree(6link)に比べて完全(2link)では43%の性能向上を達成



結果（NPB：CG法）

- Misc (66ホスト) の結果（Treeと完全結合のみ）

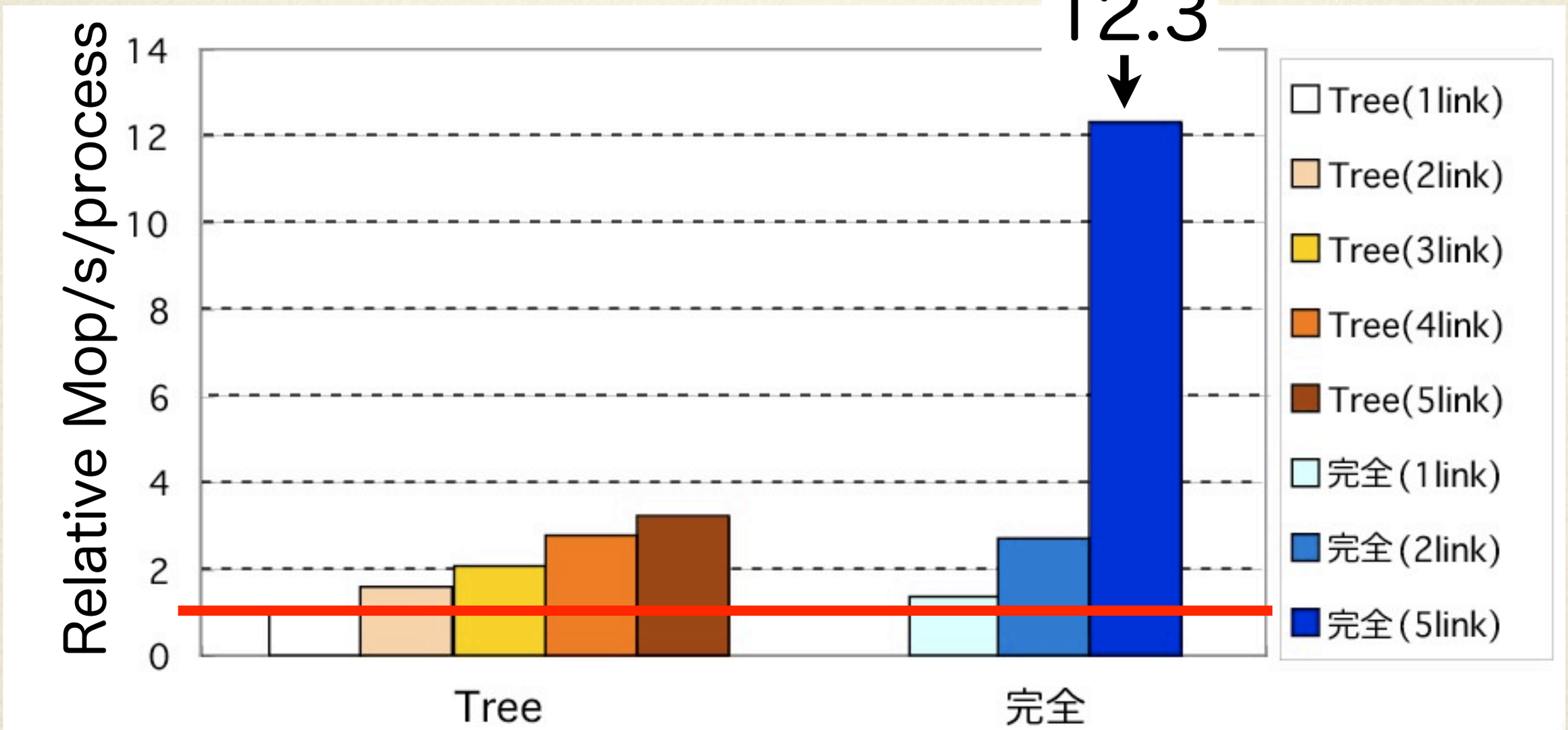
Tree(1link)と比較して，すべて性能が向上した



結果（NPB：CG法）

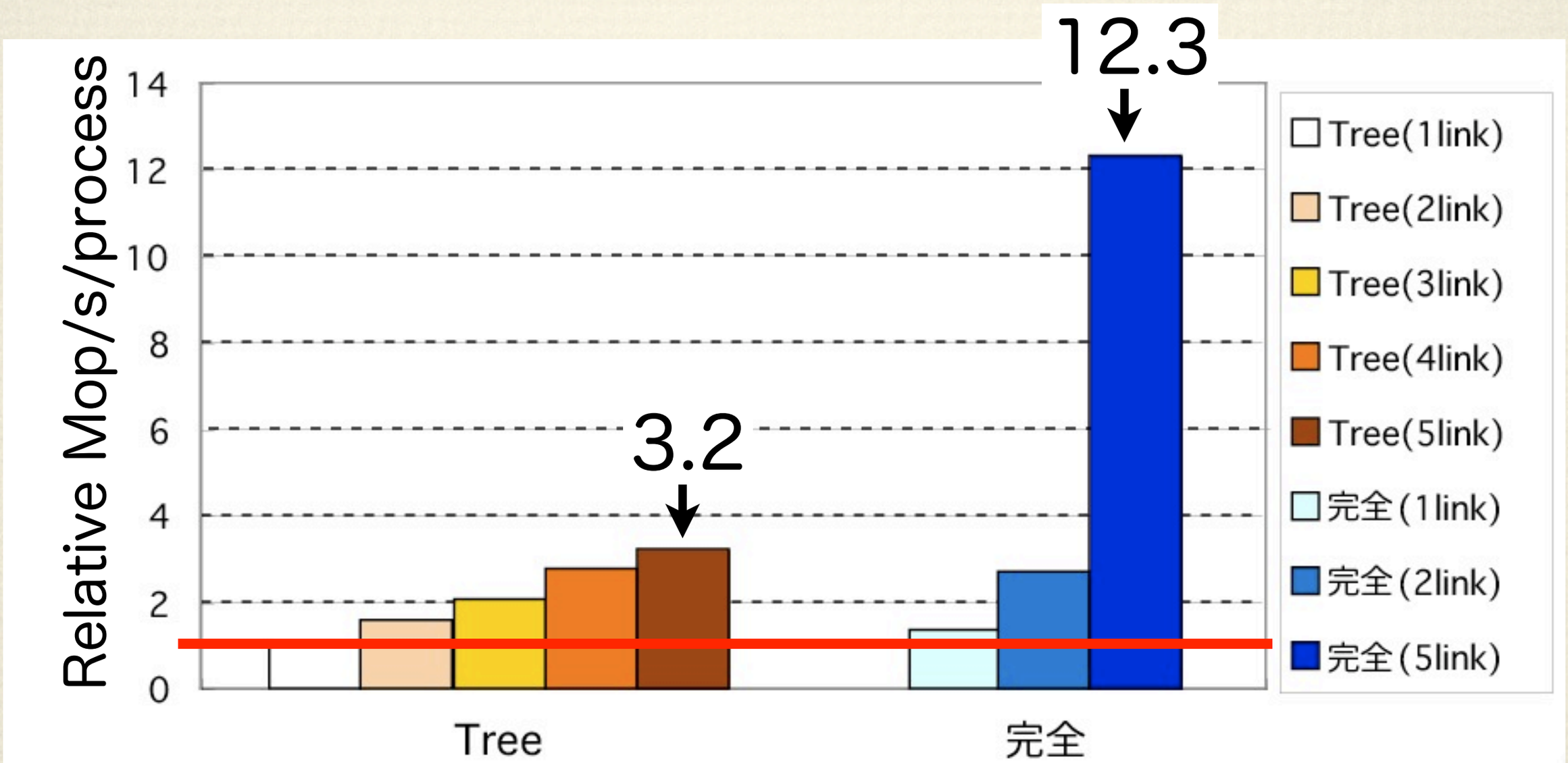
- Misc (66ホスト) の結果（Treeと完全結合のみ）

Tree(1link)と比較して，すべて性能が向上した



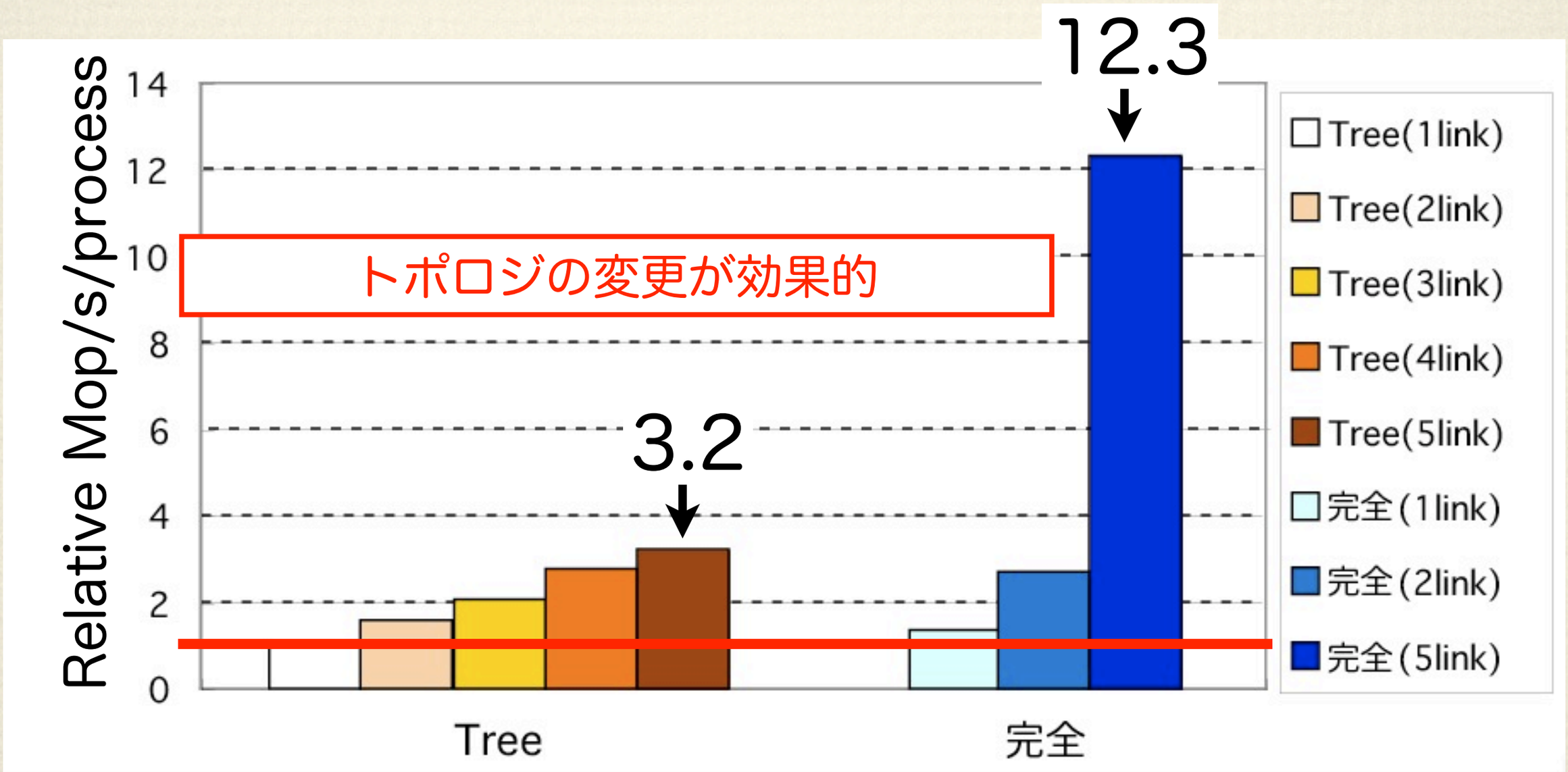
結果（NPB：CG法）

- Misc (66ホスト) の結果（Treeと完全結合のみ）



結果（NPB：CG法）

- Misc (66ホスト) の結果（Treeと完全結合のみ）



評価のまとめと考察

- High Performance LINPACK Benchmark
  - スイッチ間のリンク集約化が効果的
  - 225ホストのクラスタでピーク性能の67%を達成.
    なお，Top500のEthernetを用いたシステムの内,
    最大性能割合は63%，約9割のシステムが55%以下
  - クラスタの拡張を考慮する場合，完全結合1linkが良い

	Tree(6link)	完全(1link)	完全(2link)
スイッチ間リンク数	42	28	56
Tree(1link)との比較	185%	191%	193%

評価のまとめと考察

- NAS Parallel Benchmarks
- トポロジの変更が効果的
- 完全結合，トーラス，リングにおいて性能が高い.
  単純ツリー構造と比較して分散された経路を用いることができるため，大きな性能差になったと考えられる

以上より，アプリケーションにより適したトポロジは異なるが，ツリー以外のトポロジを選択することで，性能が向上する場合が多いことがわかった.

まとめ

- EthernetとVLAN技術を用いることで,
  高速なPCクラスタ用インターコネクトを実装
  - MACアドレス管理を工夫することで，管理の簡易化
- 2台のPCクラスタを用いた性能評価
  - HPL, NPBともに高い性能を示した

ホスト数の比較的多いPCクラスタ上で,
安価で性能の高いネットワークを構築できる

補足資料

Infinibandの価格

2009年5月末の価格

- ケーブル（4X - Quad 1X InfiniBand, 1m）：3万円
- NIC（PCI-Express x8）：10万円
- スイッチ（48ポート、シャーシ型）：80万円

通信基礎評価（1/2）

- Bit reversal
  - node0からnode63について数字のビット列を反転させた番号のノードとそれぞれ通信を行う

    送信元：node1　：$(000001)_2$

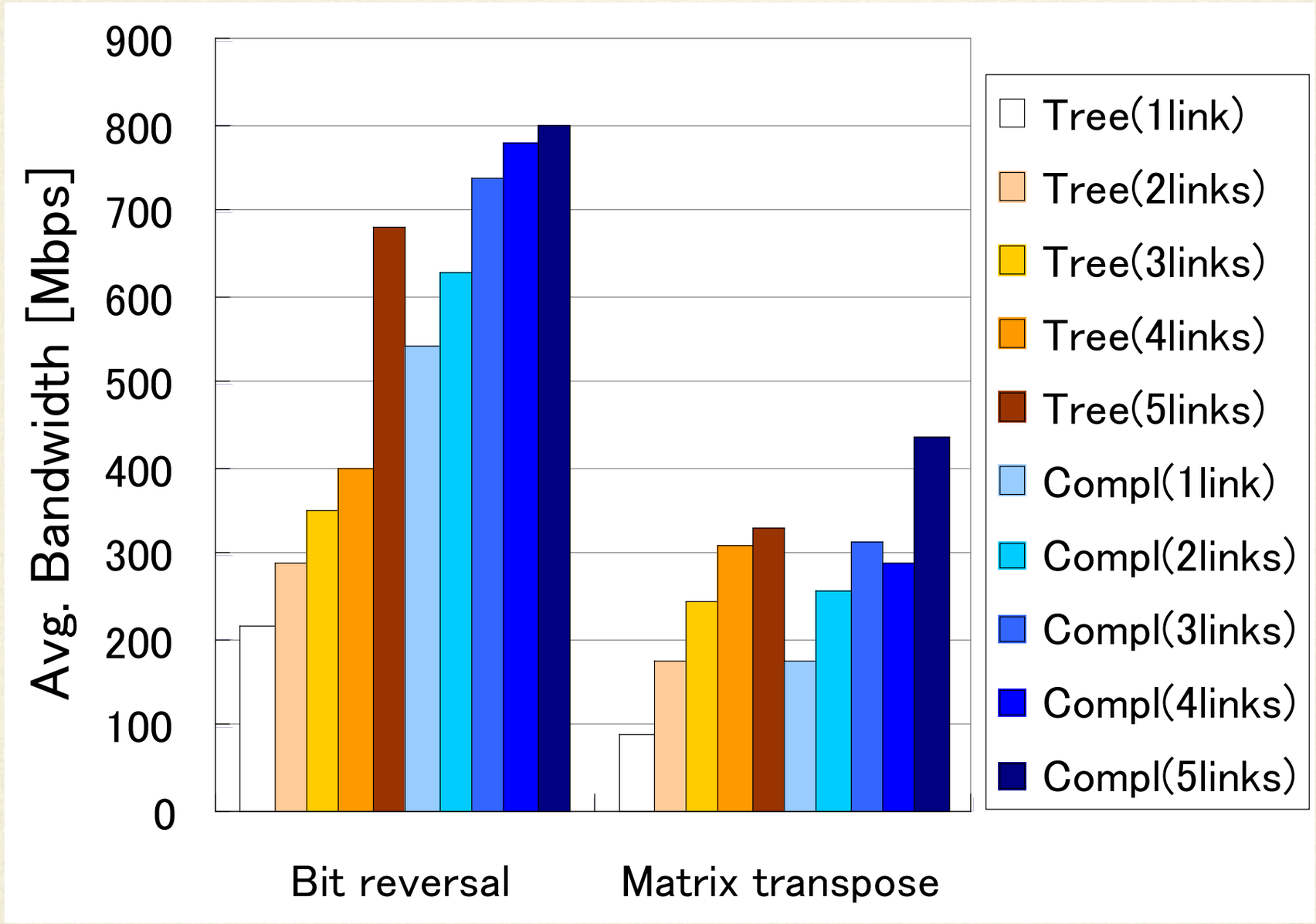
    宛先　：node62：$(111110)_2$

- Matrix transpose
  - node0からnode63について以下のような組合せで通信を行う

    送信元：node63，node62，・・，node0

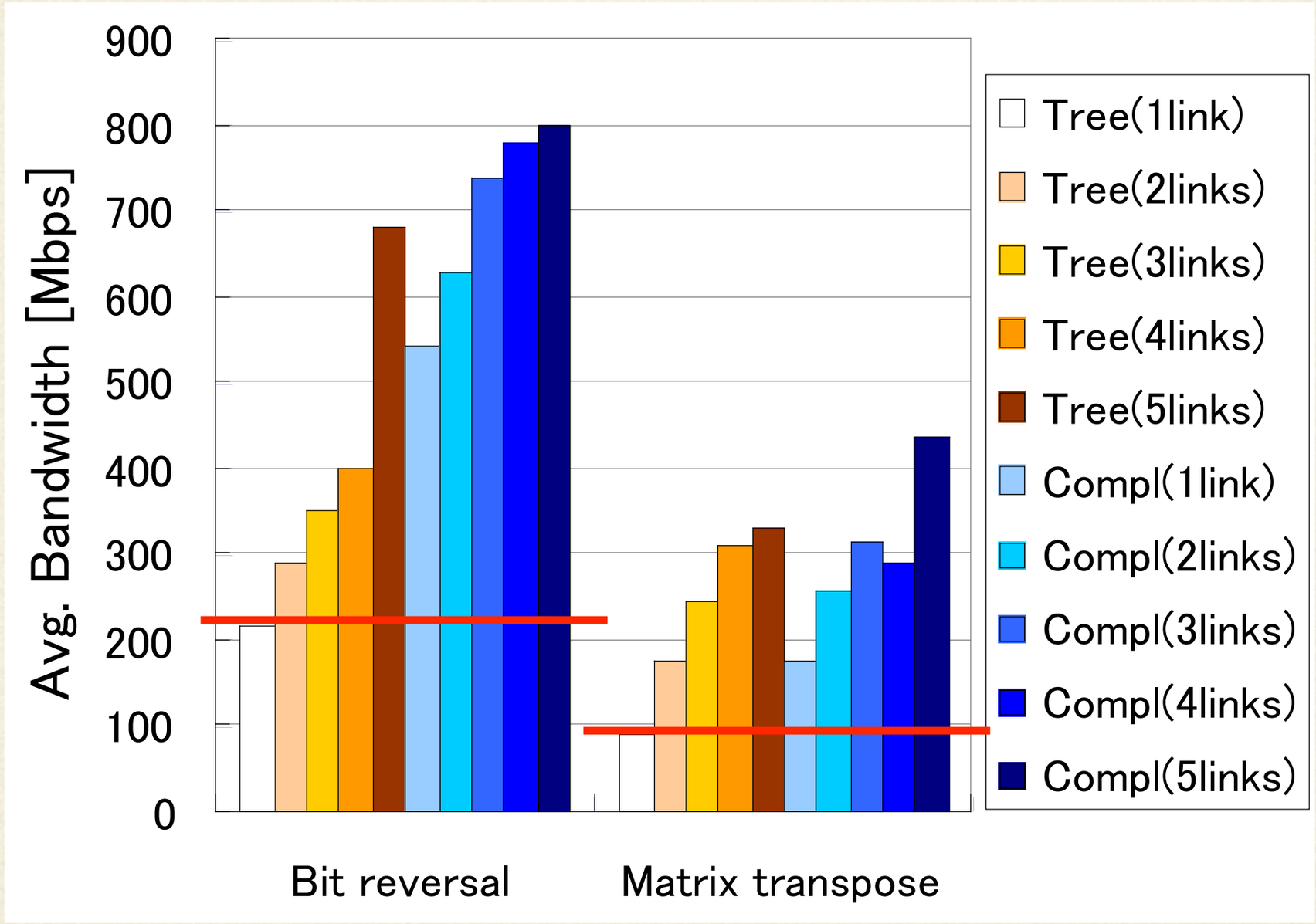
    宛先　：node31，・・node0，node1，node32

通信基礎評価（2/2）



Tree(1link)と比較して約4倍の性能

通信基礎評価（2/2）



Tree(1link)と比較して約4倍の性能

費用対効果

Supernova（225ホスト）における比較で
高価なスイッチ1台を用いた時と同等の性能を達成

スイッチ	Force 10 E1200 ×1	Dell PowerConnect　6248 ×8
ポート数	336	48
LINPACK性能（実行効率 ）	1.169 (63.4%)	1.081 (66.7%)
コスト（万円)	4,000	160（ 20 x 8)

1/25

スイッチの設定例

```
no spanning-tree // スパニングツリープロトコルを無効にする
vlan database     // VLAN IDの登録
        vlan 101,102  // VLAN ID 101,102を登録
exit

// 隣接上位スイッチへのポートのVLAN設定
interface ethernet g1 // ポート1を設定
        switchport mode general // ポートのVLANモードの設定(802.1Qモード)
        // VLAN 101のタグ付メンバとして登録
        switchport general allowed vlan add 101 tagged
exit
//ローカルホストへのポートのVLAN設定
interface ethernet g5 // ポート5 の設定
        switchport mode general
        // VLAN 101,102のタグなしメンバとして登録
        switchport general allowed vlan add 101,102 untagged
        switchport general pvid 101
exit
```

トポロジの特徴

	バイセクションバンド幅	VLAN数
Tree(n link)	n	1
完全(n link)	16n	8
Mesh(n link)	2n	4
Trous(n link)	4n	4

ホスト数とスイッチ数の関係(1/2)

- 完全結合の場合

スイッチ数：n
1つのスイッチのポート数：k
・・全てのポート数：nk

完全結合網に用いるポート数(a link)：an (n-1)

ホストに用いる事ができるポート数 = nk -an(n-1)

ホスト数とスイッチ数の関係(2/2)

- 48ポートスイッチ，１ linkの場合