OSPF と RIP

WIDE

加藤 朗

東京大学大型計算機センター

kato@wide.ad.jp

© 1997, Akira Kato. All rights reserved.

Agenda

- 経路制御概論
- **RIP, RIP2**
- **OSPF**
- **OSPF** の運用
- まとめ

概論

IP の経路制御 (1)

- 宛先アドレスによる経路選択
  - パケット毎に経路表の探索
  - **hop-by-hop** な経路選択
  - 行きだけの経路制御
    - 帰りは一般には異なる経路
- 古典的には
  - 宛先ホスト単位
  - 宛先ネットワークアドレス単位
    - 古典的経路集約

IP の経路制御 (2)

- サブネットの導入
  - 共通なネットマスクの使用
  - 二段階の経路集約
    - サブネット、(natural) ネットワーク
- **CIDR** 時代
  - 経路には全てマスク長を付加
  - 可変長サブネット
    - アドレスの効率的な使用
  - 経路情報の集約
    - 経路制御の効率化

CIDR 時代 (1)

- **CIDR** の理由
  - インターネットの普及
  - **Class B** 空間の枯渇
  - 複数の **Class C** アドレスの割当
  - 経路数が爆発的に増大
- 暫定的な解決
  - アドレス空間は一定
  - 効率化で延命
- 長期的解決
  - **IPng**

CIDR 時代 (2)

- プリフィックス長で経路を表現
  - **203.178.136.0/21**
  - 連続したネットマスク
    - 255.255.0.255 は不可
- **Class A, B, C** は忘れよう
  - 「歴史的」表現
- **Netmask** の表現には / は使わない
  - **203.178.140.192/28**
  - **203.178.140.193 mask 255.255.255.240**

経路制御方式 (1)

- **Distance Vector (Bellman-Ford) 型**
  - 宛先に対する距離 ：**metric**
    - metric の小さい経路を選択
  - ルータが互いの経路表を交換
    - 繰り返しで収束
    - count-to-infinity 問題
  - 経路情報とトラフィックは逆方向
  - 選択的な経路送信も可能
    - ポリシの実現

経路制御方式 (2)

- **Link State 型**
  - トポロジデータベースを作成
    - 仮想的には一つ
    - 実際には各ルータにコピー
    - コピーの同期が問題
  - 経路の計算
    - 自ルータを根とする Spanning Tree
    - Dijkstra のアルゴリズム (SPF)
  - 収束とループの解消が高速
  - フィルタの実装は困難

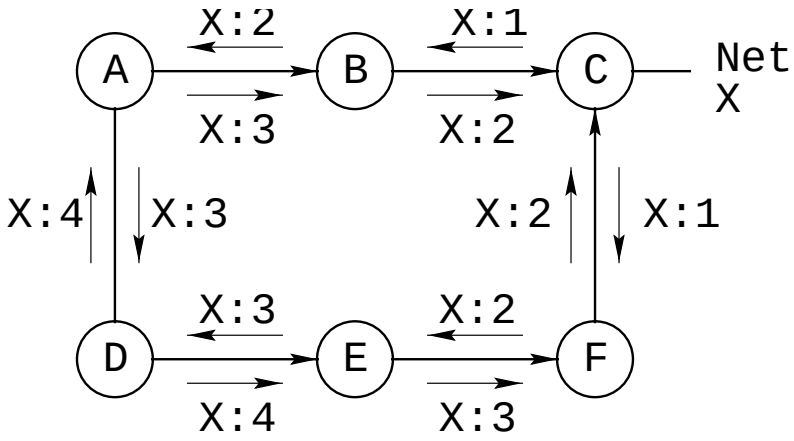
RIP, RIP2

RIP の概要

- **RFC1058 で定義**
  - 「Historic」プロトコル
  - **UDP port 520 を使用**
- **Distance Vector 型の経路制御プロトコル**
  - ルータ相互間での経路制御の交換
  - **Metric** で経路の質を表現
    - 1 ~ 16 の範囲
    - 16 は無限大
  - 宛先 **0.0.0.0**
    - default 経路

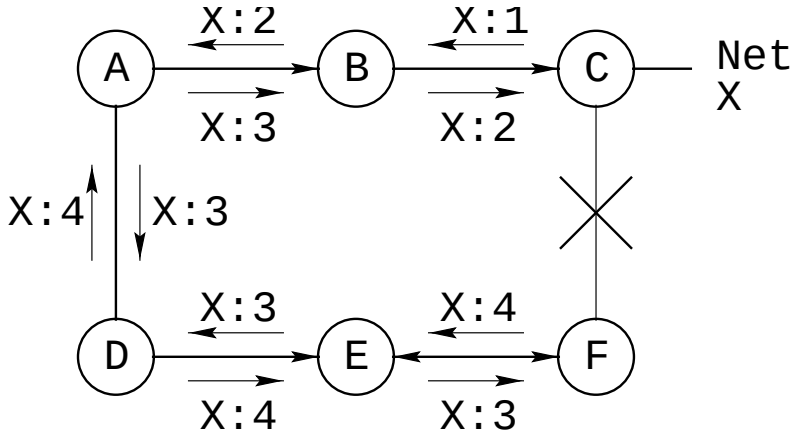
RIP の概念

- ルータが自分の経路表を順次教える
  - **30** 秒に 1 回 **flooding**
  - **180** 秒来ない経路は **down**
    - metric を 16 にする
  - さらに **120** 秒経つと消去



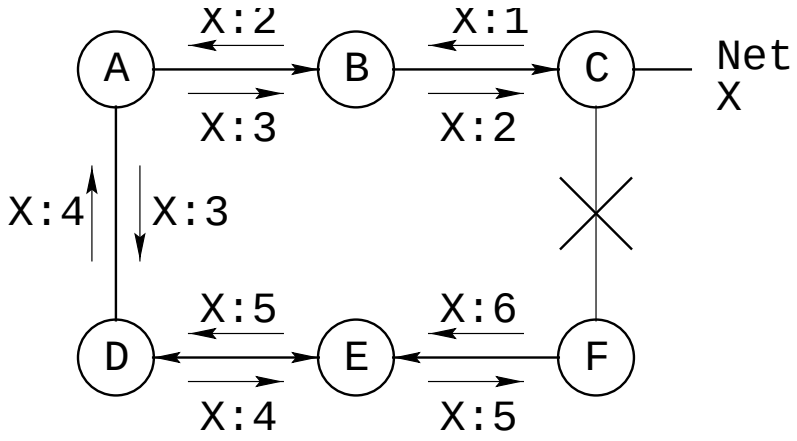
リンクが切れたら (1)

- **A, F** 間のリンクがダウンすると :
  - **A** からの **RIP** が途絶える
    - 180 秒でタイムアウト
  - **F** の **X** 行き経路が **E** を向く
  - **E, F** 間で経路ループが発生



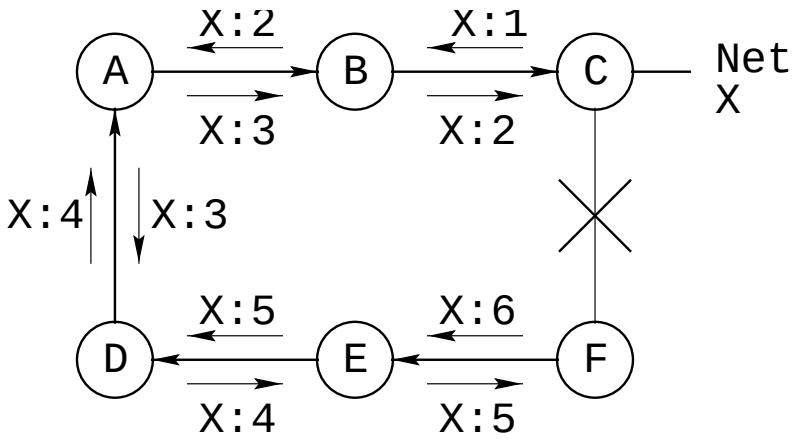
リンクが切れたら (2)

- **F** から **E** に **X:6** が送られると :
  - **E** は **X** 行き経路として **D** を選択
  - **D, E** 間で経路ループが発生

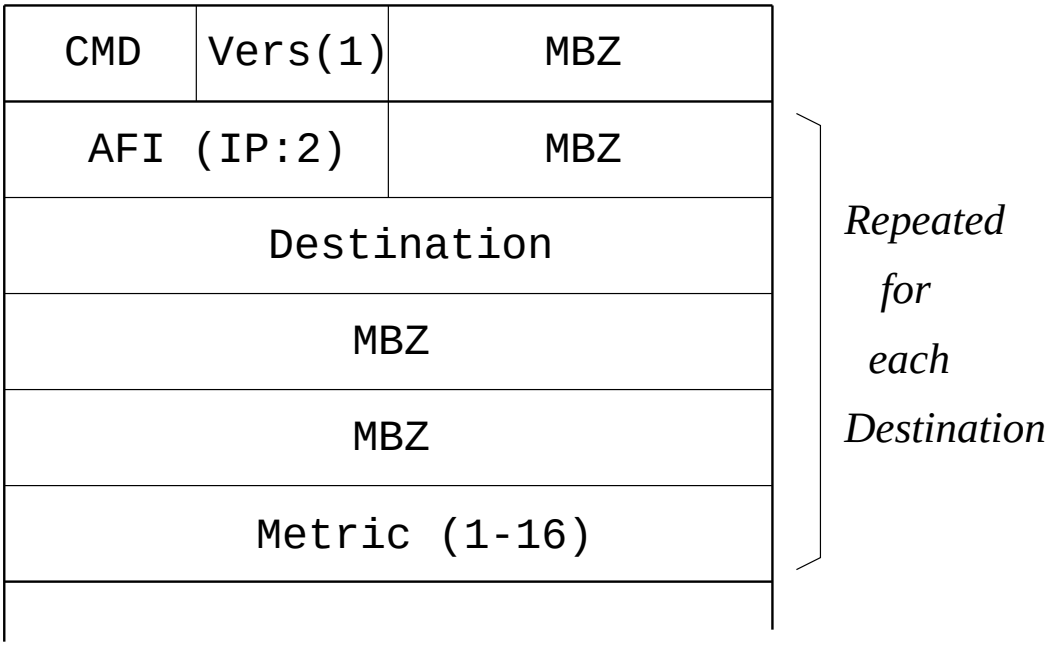


リンクが切れたら (3)

- **E** から **D** に **X:5** が送られると :
  - **D** は **X** 行き経路として **A** を選択
  - 経路ループの無い状態に辿り着く
  - ループ解消に時間が掛かる



RIP パケット



MBZ: Must Be Zero

RIP コマンド

- **1: request**
  - 隣接ルータに対する経路要求
  - 特定の宛先の問い合わせ
  - 経路表全体の要求
    - AFI: 0, Metric: 16
- **2: response**
  - 問い合わせに対する返事
  - **30** 秒毎の更新メッセージ
  - **triggered update**

RIP の改良 (1)

- 現在の次段ルータからの **metric 16**
  - **metric** を **16** に設定
  - **120** 秒の **GC** タイマを起動
- **Triggered Update**
  - 経路変動は **30** 秒を待たずにアナウンス
  - 変化した経路を素早く伝搬

RIP の改良 (2)

- **Split Horizon**
  - 経路が来た方向にはアナウンスしない
  - 誤った経路選択を防止
- **Poisoned Reverse**
  - **Metric 16** でアナウンスする
  - ループ解消が速い場合がある
  - オーバヘッドは増加

RIP の問題点

- **30 秒に一回の Flooding**
  - 一パケットで **25** 経路
  - 経路数の増加でルータの負荷が増大
  - ルータの入力キューの溢れ
- **CIDR** に適合しない
  - **Netmask** の伝達ができない

RIP の適用範囲

- 小さなネットワーク
  - 数十サブネット、直径で数ホップ
- ルータの存在の通知
  - **default** 経路のみ通知
  - ホストは受信だけ
    - routed -q
    - IPv6 では RA として実現

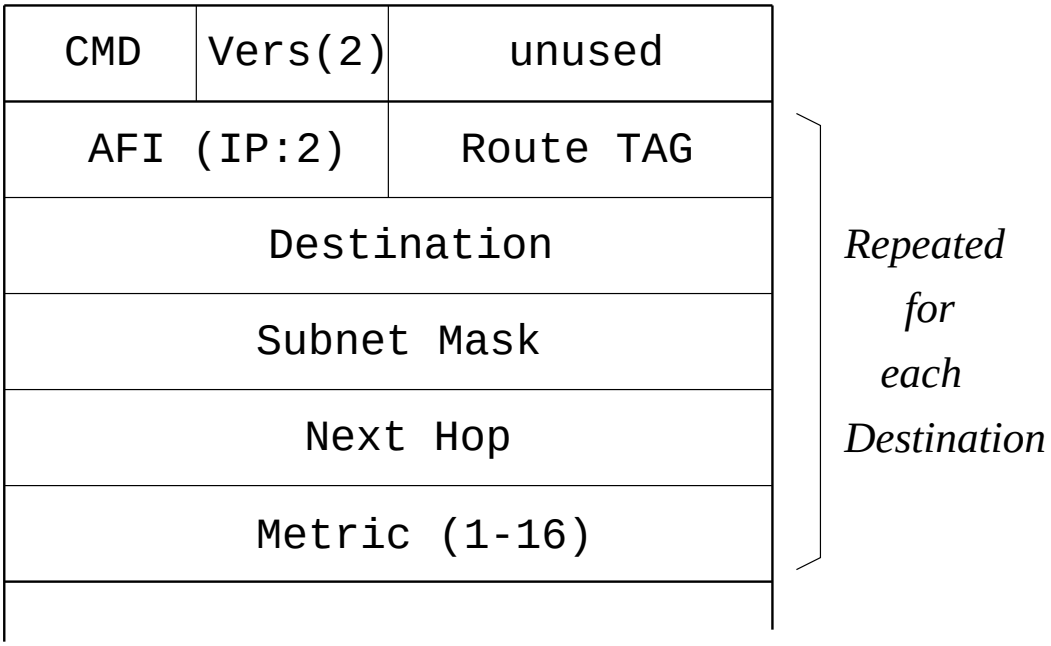
RIP version 2

- **RIP の改良版**
  - **RFC1723**
  - **CIDR に対応**
  - **RIP に上位互換**
  - マルチキャストの使用
    - 関係の無いホストの負荷低減
  - **Metric は 1 ~ 16**
    - 変更できず...

RIP-2 の特徴

- **Netmask** の搬送
  - **CIDR** に対応
- **Nexthop** アドレスの通知
  - **Redirect** を避ける
- **Tag** の添付
  - **BGP** との連携
- 認証機能
  - パスワード認証
  - 認証ヘッダ ：1 経路分の場所 (**AFI: 0xffff**)

RIP-2 のパケット


CMD
Vers(2)
unused
AFI (IP:2)
Route TAG
Destination
Subnet Mask
Next Hop
Metric (1-16)
Repeated for each Destination

RIP-2 の適用

- 実装 ：
  - **gated, cisco, bay** 等
- 実際にはあまり嬉しくない
  - **OSPF** が普及、安定に稼働
  - ホストでの実装が少ない
    - Default だけなら RIP-1 で充分
    - DHCP による default 経路の設定

OSPF

OSPF の特徴 (1)

- **Open Shortest Path First**
  - **RFC2178 (RFC1583)** で定義
    - 210 page を超える厚さ ！
  - リンク状態型経路制御
  - **LSA** データベースの同期機構
  - エリアによる階層構造
  - エリア境界での経路集約
  - 外部経路の取扱い

OSPF の特徴 (2)

- プロトコル
  - **IP プロトコル番号 89**
  - 独自の再送プロトコル
  - ケーブルマルチキャスト
    - 再送はユニキャスト
  - 認証によるセキュリティ
    - 単純パスワード
    - MD5

リンク状態型経路制御

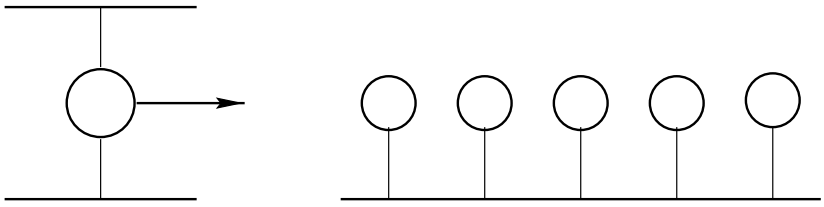
- トポロジデータベース
  - 沢山の **LSA** から構成
- 各ルータで同一コピーを保持
  - 信頼性のある通信が必要
  - 変動を実時間に更新
- 共通なデータベースから経路計算
  - 共通なアルゴリズムを使用
    - そうしないとループが発生
  - 経路ループは短時間で解消
  - ポリシの実現は困難

データベースの構成 (1)

- トポロジの記述
  - **Link State Advertisement** の集合
  - ５種類の **LSA**
    - Type-1 ~ Type-5
- ルータの名前
  - **Router ID**
- ネットワークの名前
  - 代表となるルータ (**DR**) のアドレス

データベースの構成 (2)

- **Type-1 LSA** : ルータの記述
  - 各ルータが生成
  - 接続ネットワークのリスト
- **Type-2 LSA** : ネットワークの記述
  - 各ネットワークで生成
  - 接続ルータのリスト



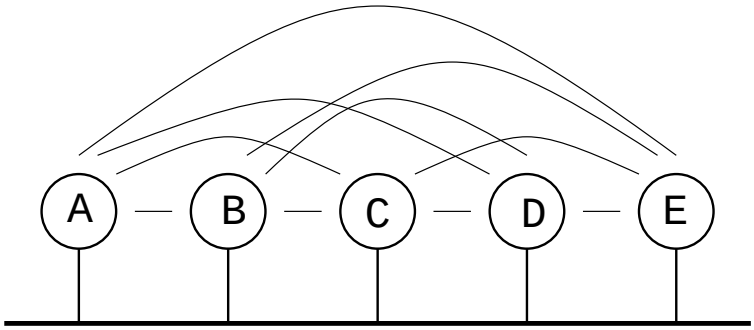
Type-1 LSA *Type-2 LSA*

データベースの構成 (3)

- **Type-3 LSA**
  - サマリ情報の記述
- **Type-4 LSA**
  - **AS** 境界ルータの記述
- **Type-5 LSA**
  - **AS** 外部経路の記述

LSA の伝搬

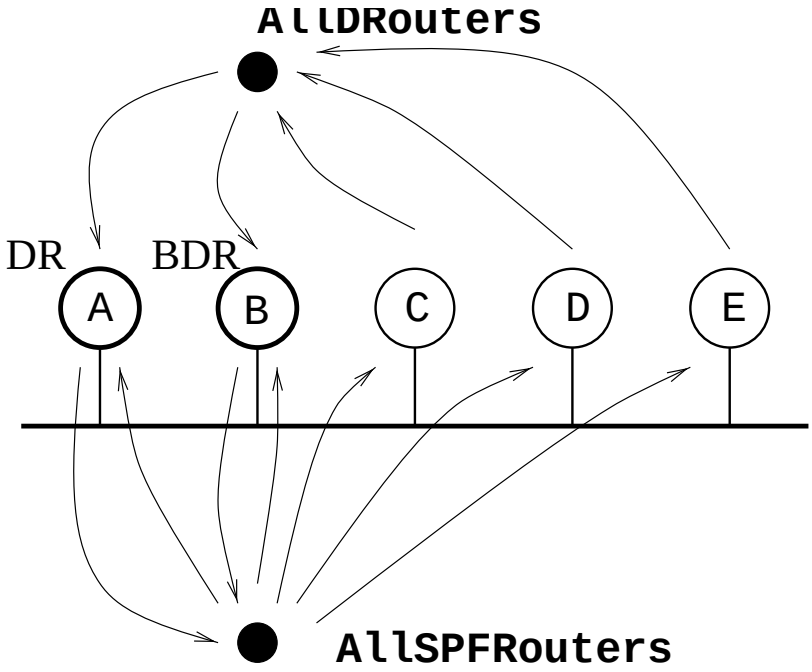
- 信頼性のある伝搬
- 効率的な伝搬



マルチキャストの利用

- **DR/BDR**
  - **Designated Router：指定ルータ**
  - **Backup BDR：バックアップ指定ルータ**
- ルータの関係
  - **DR (BDR) とその他のルータ：Adjacent**
  - それ以外：**Neighbor**
- 2 つのマルチキャストアドレス
  - **AllSPFRouters：全 OSPF ルータ**
  - **AllDRouters：DR/BDR**

DR/BDR との関係



DR の役割

- **Adjacency** の一端
- 信頼性のあるデータ転送
- 各ルータから **DR/BDR** への通信
  - **AllDRouters :224.0.0.6**
- **DR** から各ルータへの通信
  - **AllSPFRouters: 224.0.0.5**
- 再送
  - 常に **Unicast** を使用

隣接関係

- 隣接関係の状態
  - **Init : Hello を受信**
  - **2-Way : 相手の Hello に自分を発見**
    - DR/BDR でない場合には完了
  - **ExStart : Adjacency の初期段階**
    - Master/Slave 選定中
  - **Exchange : Database 記述を交換中**
  - **Loading : LSA を交換中**
  - **Full : Adjacency が完了**
    - Database の同期も確立

Router ID

- **32bit** の値
- 各ルータの対する **ID**
- ドメイン内で一意でなければならない
  - 複数のルータが同一 **ID** を用いると大混乱
- **IP** アドレスのいずれかを利用
- なるべく変動しないもの
  - 安定なインターフェース
  - **Software** インターフェース

Hello プロトコル (1)

- **Hello Interval** 毎に送出
- 受信した **Router ID** のリストを添付
  - 双方向の通信を確認
- **DR/BDR** を決定
  - 既存の **DR/BDR** を優先
    - 選択の安定性
  - **Priority** の大きなルータ
    - Router ID の大きなルータ
  - **Priority 0: DR** にならない

Hello プロトコル (2)

- **RouterDeadInterval** 内に **Hello** が来ないと
  - ダウンしたものと認定
  - 必要なら **DR/BDR** の選出
    - BDR が DR になる
    - 新 BDR を選出
- **RouterDeadInterval**
  - **HelloInterval** の数倍
- 推奨値：**Ethernet** では
  - **HelloInterval : 10 sec**
  - **RouterDeadInterval : 40 sec**

データベースの同期

- 指定ルータとの間や **p2p** リンクの両端
- **DD Description** の交換
  - **LSA** ヘッダだけを交換
  - 保持していない **LSA** をチェック
  - 古い **LSA** をチェック
  - **Window Size : 1**
- **LSA** の交換
  - **LSA update/ack**
  - 複数の **LSA** を一度に交換

Serial リンクや NBMA

- **Serial リンク**
  - 両端同士でデータベースの同期
  - **Priority** は無意味
- **NBMA**
  - 各ルータに **DR/BDR** 候補を設定
  - **HelloInterval : 30 sec**
  - **RouterDeadInterval : 120sec**
  - **Down** している **neighbor** にも **Hello**
    - PollInterval : 120 sec

共通パラメータ

- ネットワークで共通
  - **HelloInterval**
  - **RouterDeadInterval**
  - **Network Mask**
  - 認証データ
- エリアで共通
  - 認証方式
  - **エリア ID**
  - **E bit**
    - stub エリアかどうか

LSA

- **Link State Advertisement**
  - **LSA Type、LS Age**
  - **Link State ID**
  - **Sequence Number**
    - 0x80000000 ～ 0x7fffffff
    - 一周したら特別な処理
  - **LS checksum, LS length**
  - **Expire Time**
    - 1800 sec で再生成
    - 3600 sec でタイムアウト

LSA の生成 (1)

- **Router LSA**
  - **Type 1 LSA**
  - **各 OSPF ルータが生成**
  - **LSID : 生成したルータの Router ID**
  - **各接続ネットワークについて**
    - Link ID : **隣接リンクの** ID **等**
    - Link Data : Link **の** IP **アドレス等**
    - **外向きのコスト** (TOS **毎に**)

LSA の生成 (2)

- ネットワーク **LSA**
  - **Type 2 LSA**
  - マルチアクセスネットワークに対応
    - ２台以上のルータが接続したもの
    - 一台だけの場合には、stub network
  - **DR** によって生成
  - **LSID : DR の IP** アドレス
  - ネットワークマスク
  - 接続ルータの **Router ID** のリスト

Metric (1)

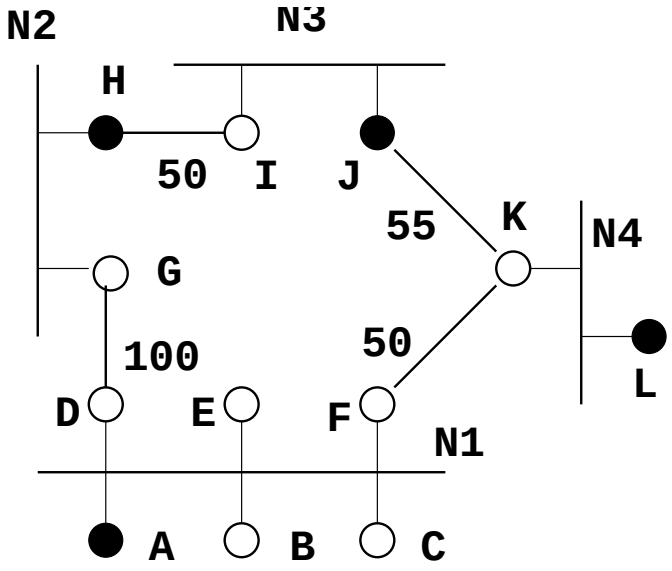
- リンクの出力側に **metric**
  - **Ethernet** : ルータ毎に設定
  - **Serial** : 方向別に設定
  - 一方向リンクは対称外
    - 近接関係を確立できない
    - UDLR working group で作業中
  - 入力側のコストは **0** と考える

Metric (2)

- **1 ∼ 65535 の範囲**
- **推奨値 : 10E8 / speed**
  - **FDDI : 1**
  - **Ethernet : 10**
  - **1.5Mbps : 66**
  - **64kbps : 1562**

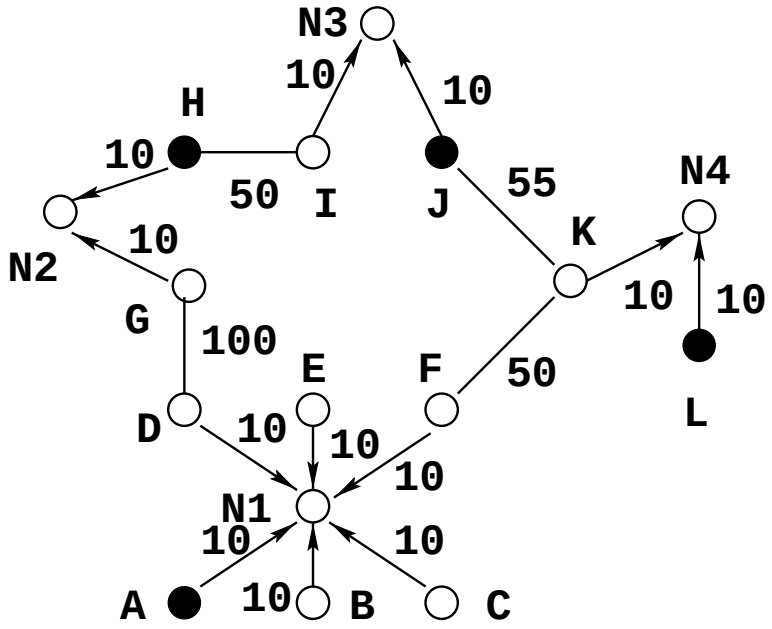
経路の計算 (1)

- Dijkstra のアルゴリズム



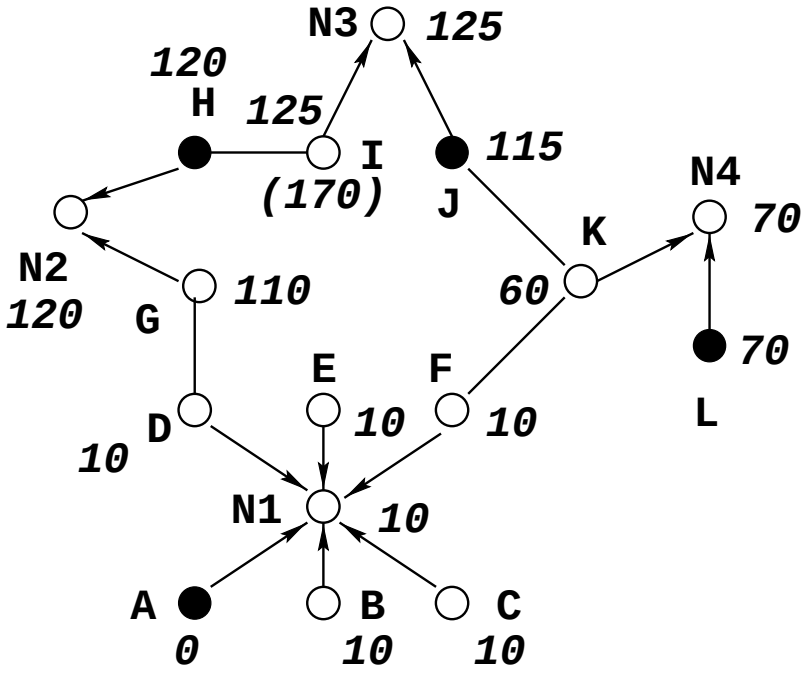
経路の計算 (2)

- マルチアクセスネットワークもノード
- 出力側に metric　(入力側は metric 0)



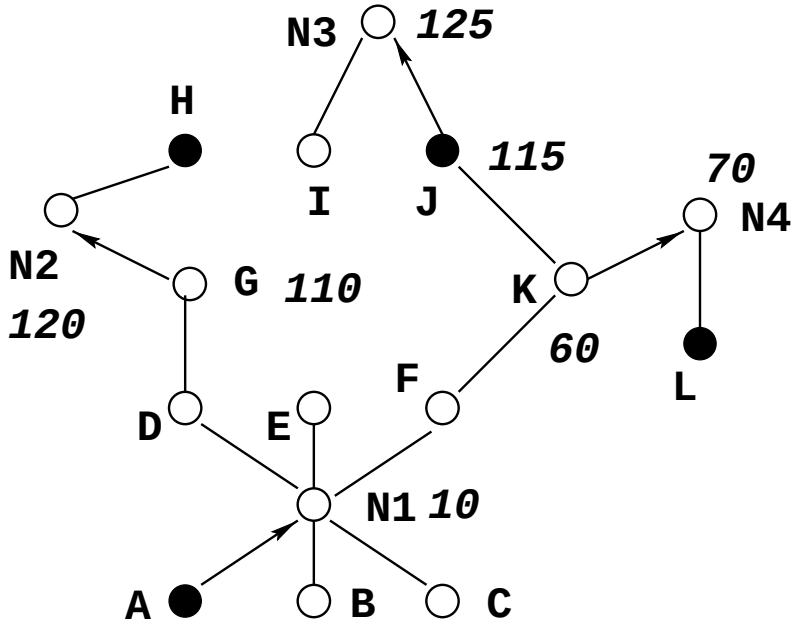
経路の計算 (3)

- ノード A に対する SPF 木の計算



経路の計算 (4)

- ノード A に対する SPF 木



経路の計算 (4)

- ノード A に対する経路表
  - **G : 110 (NH: D)**
  - **K : 60 (NH: F)**
  - **J : 115 (NH: F)**
  - **N1 : 10 (NH: A)**
  - **N2 : 120 (NH: D)**
  - **N3 : 125 (NH: F)**
  - **N4 : 70 (NH: F)**

エリア (1)

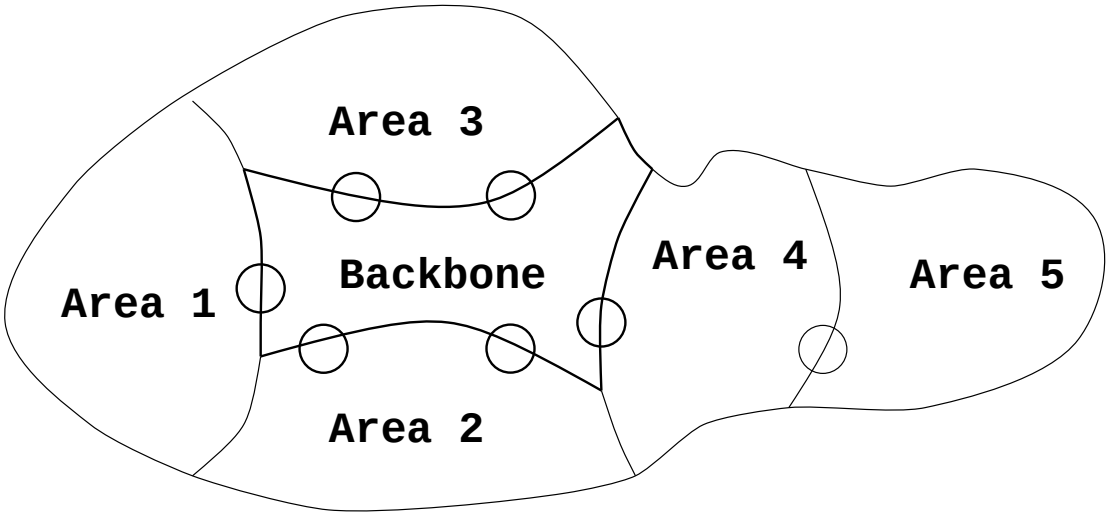
- **AS** (経路ドメイン) をエリアに分割
  - **32bit** エリア **ID**
    - 0.0.0.0 はバックボーン
  - 各ネットワークは単一エリアに所属
  - エリア境界はルータ
  - エリア外部はサマリ情報
  - エリア内だけ **Dijkstra** で経路計算
    - オーバヘッドの低減
    - サマリによる経路の集約

エリア (2)

- エリア間トラフィック
  - 発信元エリアから
  - バックボーンエリアを経由
  - 宛先エリアに
- バックボーンエリア
  - 全てのエリアに隣接
  - 必要な場合には **Virtual Link** で接続

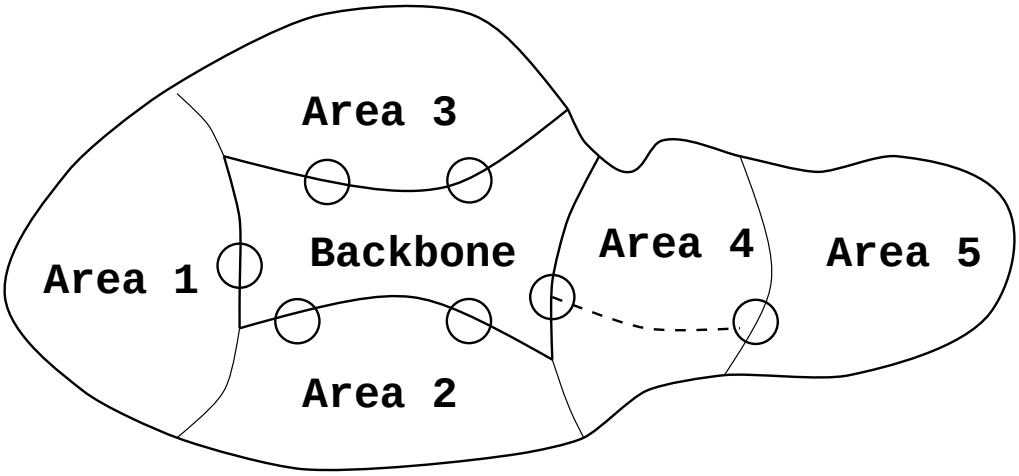
エリア (3)

- 各エリアは連結
  - バックボーンに隣接
- 正しくないエリア分割の例



エリア (4)

- **Virtual Link** で接続
  - バックボーンを拡張
  - 単一エリアを横断



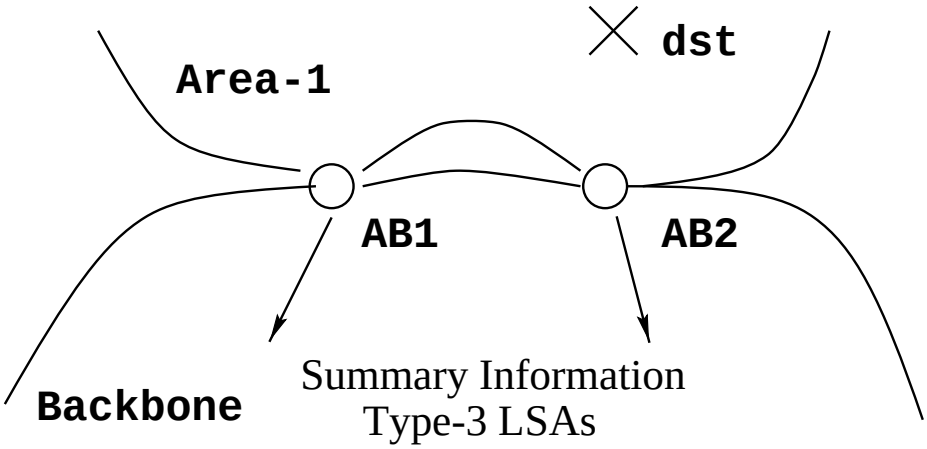
Virtual Link over Area 4

ルータの種類

- 内部ルータ
  - 単一エリアのみに接続
- バックボーンルータ
  - バックボーンに属しているもの
  - 内部ルータまたはエリア境界ルータ
- エリア境界ルータ
  - バックボーンと他のエリアを接続
- **AS** 境界ルータ
  - **AS** と他の **AS** を接続
  - **Type-5 LSA** を生成

エリア境界ルータ

- **Type-3 LSA** : サマリ情報の伝搬
  - エリア外部かつ **AS** 内の各宛先
  - **LSID:** 宛先アドレス
  - 各エリア境界ルータで生成

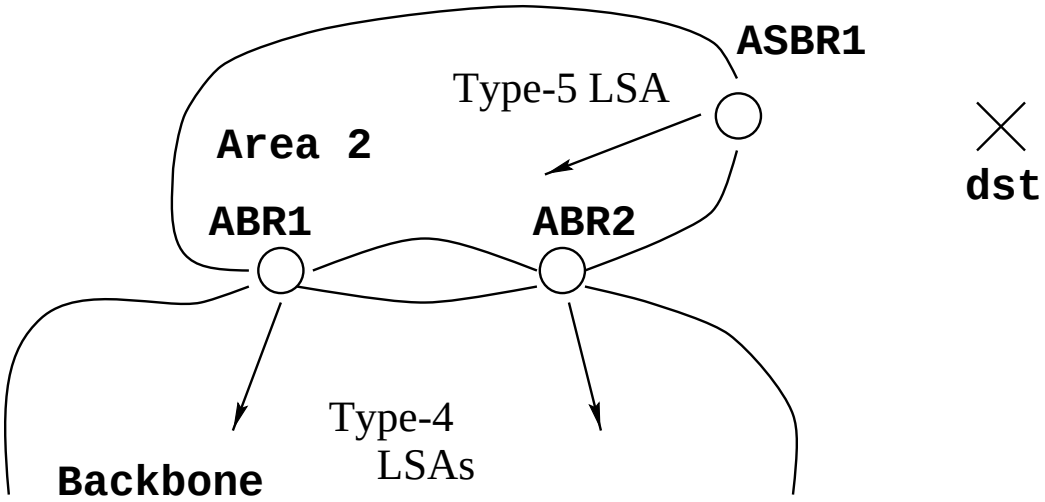


外部の経路 (1)

- エリア境界ルータ
  - **Type-4 LSA** も生成
  - **AS** 境界ルータのサマリ情報
  - **AS** 境界ルータへのコストを表現
  - **LSID: AS** 境界ルータの **Router ID**
- **AS** 境界ルータ
  - **Type-5 LSA** の生成
  - **AS** 外部の各宛先について一つ
  - **LSID:** 宛先アドレス

外部の経路 (2)

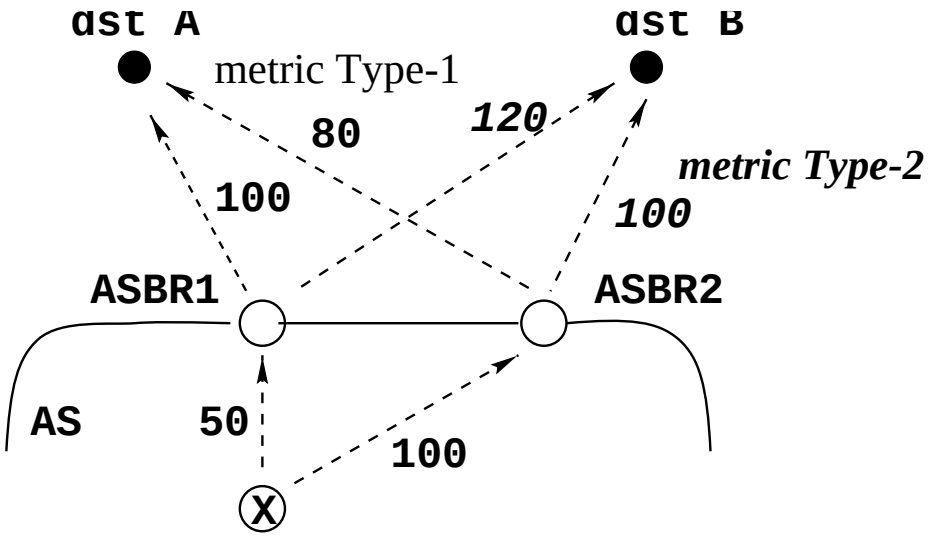
- **Type 4 LSA**
  - 各 **ASBR** について **ABR** が生成
- **Type 5 LSA**
  - 各外部宛先について **ASBR** が生成



Type 5 LSA (1)

- **Metric Type 1**
  - 外部 **metric** と内部 **metric** は比較可能
  - **metric** の合計が小さい経路を選択
- **Metric Type 2**
  - 外部 **metric** が支配的と考える
  - 外部 **metric** の小さな経路を選択
    - 内部 metric は考えず

Type 5 LSA (2)



- dst A へは ASBR1 経由 (150 : 180)
- dst B へは ASBR2 経由 (120 : 100)

Stub エリア

- **Type-5 LSA** を転送しない
- 全てのエリア内 **OSPF** ルータの合意が必要
- **Default** 経路を利用
  - エリア境界ルータで生成
  - **Type-3 LSA** で伝搬
- 外部経路の取り扱いをしない
  - **Type-5 LSA** を送らない
  - 例外 :**NSSA** オプション

OSPF と CIDR

- **LSID 規則の緩和**
  - **RFC1583 Appendix F**
  - **Type-3, Type-5 LSA**
  - マスクの異なる同一宛先を表現
    - ホスト部ビットを使用
    - 異なった LSID を命名

OSPF の運用

OSPF の設定例 (1)

- インターフェースの設定

```
interface ethernet 0/0
 ip address 192.168.1.1 255.255.255.0
 ip ospf authentication-key himitsu
 ip ospf cost 10
 ip ospf priority 1
 ip ospf hello-interval 10
 ip ospf dead-interval 40
```

OSPF の設定例 (2)

- **OSPF プロトコルの設定**

```
router ospf 100
 network 192.168.1.1 0.0.0.0 area 0.0.0.0
 area 0.0.0.0 authentication
```

Interface の確認


```
% show ospf interface
Ethernet0/0/0 is up, line protocol is up
   OSPF not enabled on this interface
Fddi5/0/0 is up, line protocol is up
  Internet Address 203.178.137.34/29, Area 0.0.0.0
  Process ID 2500, Router ID 203.178.136.4, Networ
  Transmit Delay is 1 sec, State BDR, Priority 1
  Designated Router (ID) 203.178.136.6, Interface
  Backup Designated router (ID) 203.178.136.4, Int
  Timer intervals configured, Hello 10, Dead 40, W
    Hello due in 00:00:07
  Neighbor Count is 1, Adjacent neighbor count is
    Adjacent with neighbor 203.178.136.6  (Designa
```

Neighbor の確認 (1)


```
% show ospf neighbor
Neighbor ID  Pri State   Dead Time Address
203.178.136.6  1 FULL/DR 00:00:35 203.178.137.33
150.65.0.1     1 FULL/ - 00:00:35 203.178.136.126
203.178.141.74 1 FULL/ - 00:00:38 203.178.141.74
203.178.136.7  1 FULL/ - 00:00:32 203.178.141.66
203.178.141.94 1 FULL/ - 00:00:36 203.178.141.94
203.178.141.85 1 FULL/ - 00:00:30 203.178.141.82
```

Neighbor の確認 (2)


```
% show ospf neighbor
Neighbor ID   Pri State        Dead Time Address
203.178.136.195 1 2WAY/DROTHER 00:00:55 203.178.136.167
203.178.137.50  1 FULL/BDR     00:00:56 203.178.136.164
203.178.136.196 1 2WAY/DROTHER 00:00:59 203.178.136.165
203.178.136.169 1 2WAY/DROTHER 00:00:55 203.178.136.169
203.178.136.241 1 FULL/DR      00:00:59 203.178.136.170
203.178.136.168 1 2WAY/DROTHER 00:00:55 203.178.136.168
```

Neighbor の確認 (3)


```
% show ip ospf neighbor detail
Neighbor 203.178.136.196, interface address 203.178.136.101
  In the area 0.0.0.0 via interface Serial1/0
  Neighbor priority is 1, State is FULL
  Options 2
  Dead timer due in 00:00:59
Neighbor 203.178.142.254, interface address 203.178.142.254
  In the area 203.178.136.100 via interface Serial1/3
  Neighbor priority is 1, State is FULL
  Options 2
  Dead timer due in 00:00:57
  Link State retransmission due in 00:00:04
      LSA in retransmission queue 2
```

Neighbor にならない場合

- パラメータの不一致
  - エリア ID
  - Netmask
  - タイミングパラメータ
  - 認証方式
  - 認証
- 状況の確認

```
# debug ip ospf adj
```

Database の確認 (1)


```
% show ip ospf database database-summary
Area ID         Rtr Net   SumNet SumASBR Subtotal
0.0.0.0          61  17     46      48      172
203.178.136.100 2     0     97      70      169
AS External                                4452
Total            63  17    143     118     4793
```

Database の確認 (2)


```
% show ip ospf database

                Router Link States (Area 0.0.0.0)
Link ID         ADV Router      Age Seq#       Checksum Link
133.27.2.3      133.27.2.3      340 0x80000A44 0x11F9  1
150.65.0.1      150.65.0.1      999 0x8000613E 0xEEF4  5
202.249.3.84    202.249.3.84    622 0x800015B4 0x53C0  1
202.249.3.86    202.249.3.86    666 0x80001638 0x2760  1
203.178.136.1   203.178.136.1   961 0x80004965 0x65D6  3
203.178.136.2   203.178.136.2  1201 0x8000144A 0x35EB  2
203.178.136.3   203.178.136.3   690 0x800054DA 0x7682  3
203.178.136.4   203.178.136.4   949 0x800018D2 0x1CCB 11
```

Database の確認 (3)


```
% show ip ospf database network 203.178.137.1
  Routing Bit Set on this LSA
  LS age: 854
  Options: (No TOS-capability)
  LS Type: Network Links
  Link State ID: 203.178.137.1 (address of DR)
  Advertising Router: 203.178.136.150
  LS Seq Number: 8000001E
  Checksum: 0x36B2
  Length: 32
  Network Mask: /29
        Attached Router: 203.178.136.150
        Attached Router: 203.178.136.16
```

Database の確認 (4)


```
% show ip ospf database database-summary
% show ip ospf database
% show ip ospf database router <LSID>
% show ip ospf database network <LSID>
% show ip ospf database summary <LSID>
% show ip ospf database asbr-summary <LSID>
% show ip ospf database external <LSID>
```

OSPF の設定ミスの例 (1)

- エリア分割
  - バックボーンに隣接しないエリアの作成
    - エリア分割の見直し
    - Virtual Link の使用
  - 単一エリアのルータ数
    - 数十程度以下が推奨
- **Full Route**
  - 多くのルータは **50,000** 経路の **OSPF** は無理
    - メモリや CPU の問題
    - トラフィック

OSPF の設定ミスの例 (2)

- **Router ID**
  - 同じ **Router ID** を複数のルータ
    - 誤った経路計算
    - ループの発生
    - 一時間以上経ってから

OSPF tips (1)

- 単一エリアに属すルータ
  - 全部のインターフェースを一つのエリアに

```
router ospf 100
  network 0.0.0.0 255.255.255.255 \
    area <AREA>
```

- 他にルータの無いネットワーク
  - **Hello** を止める

```
router ospf 100
  passive-interface ethernet 0/0
```

OSPF tips (1)

- **OSPF 以外の経路の導入**

```
router ospf 100
 redistribute rip route-map rip-ospf subnet
access-list 1 permit any
route-map rip-ospf permit 10
 match ip address 1
 set metric 1000
 set metric-type type-1
```

- 外部経路の再導入

```
# clear ip ospf redistribute
```

まとめ

RIP と OSPF

- **RIP**
  - 限られた場面では依然として有効
  - 大きいネットワークでは使用禁止
- **OSPF**
  - プロトコルは複雑
  - 実装は概ね安定している
  - 運用はそんなに難しくない
  - デバッグはちょっと大変 ？