# BGP FOR SYSADMIN

**ExaBGP ou comment gérer ses IPs de service**

SYSADMIN #4
28th of Febuary 2013

Thomas Mangin
Exa Networks

exa networks

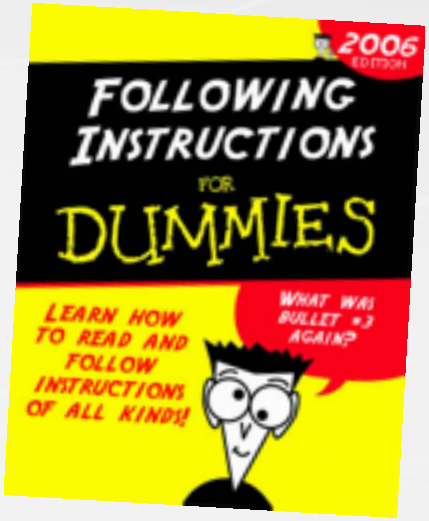
Whatever a speaker is missing in depth he will compensate for in length

*Montesquieu*

# NO NETWORKING 101

I ASSUME THAT ...

- You have basic networking knowledge
  (connected, static routes)
- Your organisation use some routers you can break
- You know what IPs, netmasks, gateways are




I WILL COVER

- Quickly Dynamic Routing
- What is BGP, the Border Gatway Protocol
- Why BGP is a great protocol for sysadmins

I WILL NOT COVER

- How to configure a BGP router for general purpose

exa networks

Truth is more valuable if it takes you a few years to find it.
*Renard*

# ASN

## AUTONOMOUS SYSTEM NUMBER

**Unique Network identifier**

initially 16bits
32 bits usage is a negotiated feature (RFC 4893)

**Like RFC 1918, its reserves some IPs**

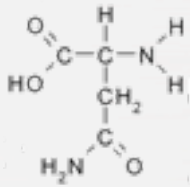
Some ASNs are reserved for documentation (like the 192.0.2.0/24 range)
The range 64496-64511
Some ASNs are reserved for private use
The range 64512-65532

**Given to** LIR (LOCAL INTERNET REGISTRY)

In France, this means RIPE members
does not mean ISP only

http://as30740.peeringdb.com/

**Asparagine** (abbreviated as **Asn** or **N**)



A little learning is a dangerous thing
*Alexander Pope*

exa networks

# BORDER GATEWAY PROTOCOL

**To share routing information between ASN**

Many RFCs (main one being 4271), many optional features
http://www.bgp4.as/

Open Source implementation in BIRD, Quagga, OpenBGPD

To use it, you do **NOT** need to :
- be connected to the internet
- have real world IPs
- be or ask an ISP anything (but it can be useful)

NOT


Use TCP with its own failure detection mechanism.
-> minimum 3s for failure detection

There are many true statements about complex topics that are too long to fit on a PowerPoint slide
*Edward Tufte*

exa networks

# BGP transmits ROUTES

**What makes a route**

- A PREFIX (a block of IP) - the “destination IP regex”
- A DESTINATION (called next-hop)
- with many optional information (called ATTRIBUTES)
  - use to select one route over another

The next-hop is a machine that should know how to contact any IP in the prefix, it does not have to be locally connected but just “known”.

Some of the attributes are

- LOCAL PREFERENCE, a value to distinguish two 'identical routes'
- AS PATH, the chain of ISP who have seen and transmitted the route

**BGP will make sure**

- that the data is always sent to a machine nearer to the end point than itself
- that the decision process between multiple routes does not cause loops



Logic will get you from A to B. Imagination will take you everywhere

*Albert Einstein*

# SHOW (me a) ROUTE

BGP only has one active route for a prefix at a time (the one indicated with *)
BUT can use multiple links to get to the next-hop (depending on the IGP)

```
> show route 192.175.48.0

192.175.48.0/24     *[BGP/170] 6w1d 00:57:41, localpref 175
                       AS path: 112 I
                     > to 82.219.2.177 via ge-0/3/0.17

                     [BGP/170] 6w1d 00:57:40, localpref 175, from 82.219.0.69
                       AS path: 112 I
                       to 82.219.1.85 via ge-0/3/0.9
                     > to 82.219.2.202 via ge-1/3/0.28
                       to 82.219.2.155 via ge-1/3/0.30
                       to 82.219.2.194 via ge-0/3/0.32

                     [BGP/170] 4d 03:09:59, localpref 75
                       AS path: 286 8312 35627 112 I
                     > to 134.222.89.0 via ge-1/3/0.142
```



Work delivers us from three great evils: boredom, vice and want.

Voltaire.

# BGP CONVERSATION

**Two routers establish one TCP connection** (port 179)

**exchange some information about what they can do** (OPEN messages)
- what extra address family they support (IPv6, IPvpn, ...)
- what advanced features (GRACEFUL RESTART, 32 bits ASN, ...)

**Send each other what they know about the network** (UPDATE messages)
- this is where the routes exchanges occurs
- each UPDATE can be to
  - announce a new route(s) or
  - withdraw a previously known route(s)

**BGP does not rely on TCP for link failure for peer failure detection**
- instead send heartbeat data every few seconds (KEEPALIVE messages)
- failing to send 3 messages in a row kills the connection
- smaller delay between message 1s -> minimum 3s for failure detection



Nothing is more humiliating than to see idiots succeed in enterprises we have failed at

*Flaubert*

# BGP CONVERSATION

**IDLE > ACTIVE > CONNECT > OPEN SENT > OPEN CONFIRM > ESTABLISHED**

| | |
|---|---|
| **IDLE** | Configured but not ready |
| **ACTIVE** | Configured and ready |
| **CONNECT** | TCP connection established |
| **OPEN SENT** | The router sent its OPEN packet |
| **OPEN CONFIRM** | The peer replied with its OPEN then KEEPALIVE |
| **ESTABLISHED** | The router sent its KEEPALIVE packet |

**Once ESTABLISHED**

| | |
|---|---|
| UPDATE | A packet with routing information (both way) |
| KEEPALIVE | The heartbeat packet |



Nothing is more humiliating than to see idiots succeed in enterprises we have failed at

*Flaubert*

# EBGP vs IBGP

**Same protocol - totally different usage**

**EBGP**

- used by different services providers to interconnect
- both routers are in different Autonomous systems
- Most often the next-hop of received route will be rewritten to “self”

**IBGP**

- Can be used as an IGP replacement
- Each router is fully meshed with all the others (many TCP session)
- **configured as route-reflector** a router can become a “repeater” for other BGP peers

**BOTH**

- Can be used to inject any route in a network



Nothing is more humiliating than to see idiots succeed in enterprises we have failed at

*Flaubert*

# BGP ROUTE SELECTION

**The more specific the route, the better**

/32 better than /31, better than /30, ...

**Warning: protocols have preferences**

CONNECTED > STATIC > IGP > BGP (last)

**Must be a valid routes**

must be SYNCHRONISED with the IGP
(let's turn that off on the router)
The NEXT_HOP must be reachable.

**Route selection (in order)**

Highest WEIGHT (cisco proprietary)
Highest LOCAL PREF (used within an AS)
Prefer LOCALLY ORIGINATED route
Shortest AS-PATH
...

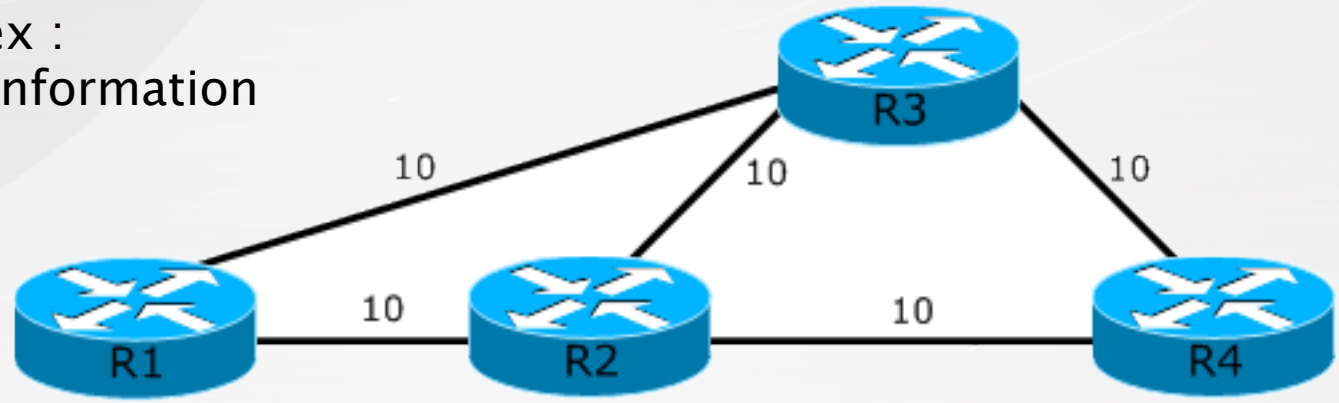
exa networks

# WHAT IS AN ... I.G.P ?

A routing protocol used by routers

RIP : obsolete, use OSPF
OSPF : use TCP, adaptive routing (available everywhere)
IS-IS : use an ISO L2 protocol, adaptive routing (higher end kit)
EIGRP : use multicast, distance-vector routing (cisco only)

They are all complex :
- share topology information
- election process
- ...




Patience is bitter, but its fruit is sweet
*Rousseau*

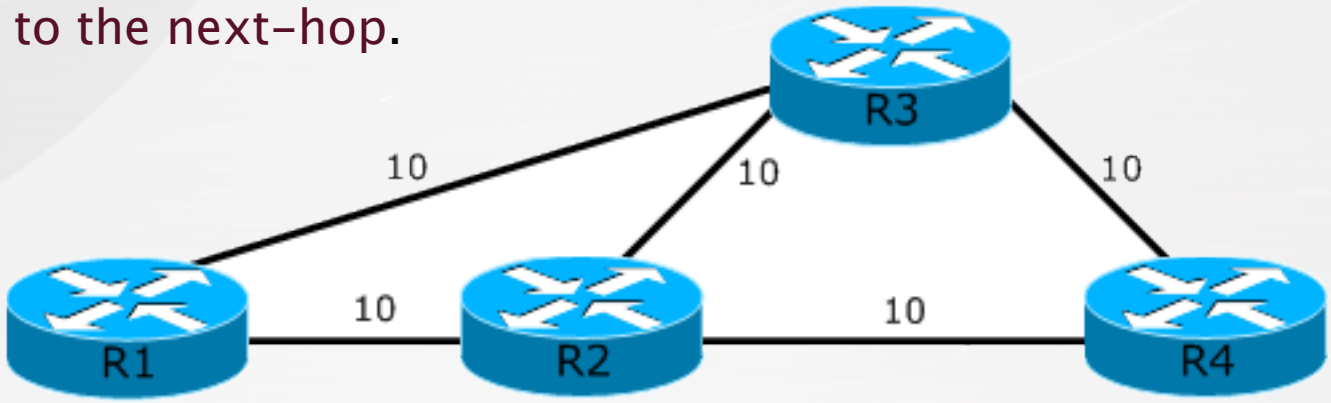
exa networks

# WHAT IS AN ... I.G.P ?

Should contain as few routes as possible (P2P and connected networks)
    Converge quickly (find a alternative path) in case of link failure

Multiple routes per prefix is possible
    Traffic load balancing between links

BGP only has one active route for a prefix at a time but the IGP may use multiple paths to get to the next-hop.




Patience is bitter, but its fruit is sweet
*Rousseau*

exa networks

# Options for service resilience

**HSRP, VRRP**

resilience for the gateway, not the host

**Linux-HA solutions** (Heartbeat, Pacemaker, Wackamole,..)

Need both machine in the same Layer 2

Lack of IPv6 support !

ARP (relation MAC/IP) expiry 4 to 6 hours ..

MAC (relation ARP/Port) expiry 5 minutes

some kit only allow configuration per interface, not VLAN

enabling gratuitous ARP is a security risk on shared networks (cloud)

**Yahoo! L3DSR load balancing solution**

Layer 3 Load Balancing, encoding the destination IP in the DSCP field

http://www.nanog.org/meetings/nanog51/presentations/Monday/NANOG51.Talk45.nanog51-Schaumann.pdf

# WHERE DOES BGP FITS ?

**External BGP** : connecting to other networks
- protection from ISP outages

**EBGP or IBGP**
- Anycast : announce the same IP at different location (CDN, DNS, ...)
- DDOS "mitigation" : prevent bad traffic to reach servers
- Flow Routes (firewall rules deployment using BGP)

**Internal BGP** : fully controlled BGP
- block/redirect some traffic (customers, countries, organisations, ...)
- Servers announcing some Service IPs

exa networks

I love fools' experiments. I am always making them.
*Charles Darwin*

# BE YOUR OWN ISP

IPV6 ONLY

**RIPE Membership**

- Become your own ISP
- IPV4 – ran out !

**Provider Aggregate versus Provider Independant**

- PA: a block of IP owned by the LIR (often the ISP)
  - changing ISP forces you to renumber
- PI : a block of IP owned by the end users
  - changing ISP is a routing change

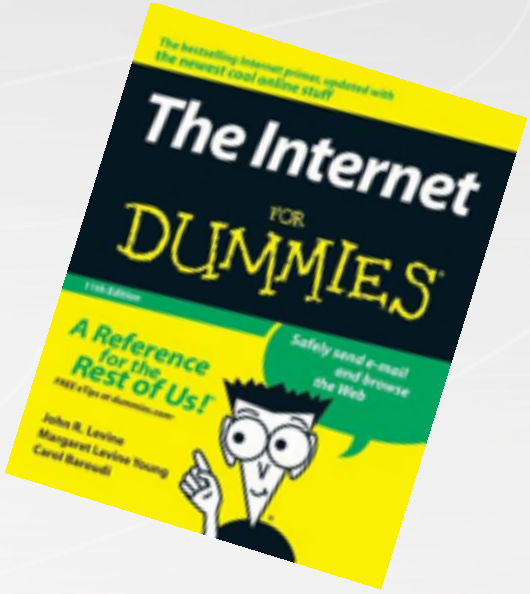
**Announce your network to the world via BGP**

- Not as hard as it sounds
- Ask you ISP

# ANYCAST

**Split personality ..**

- Announcing the same IP with BGP in different location
- Another RFC (4786)
- The network finds the nearest server
- Not best suited for long lived TCP connections
  - routing can change




**On the internet used by**

- Root servers (UDP mainly)

**Within a networks**

- caching DNS (UDP)
- CDN local DNS (UDP)
- Proxies (TDP, near DSL exit points, requires very stable routing)
- ...

# RTBH

**Tell your provider to stop sending you traffic for some IPs**

Announce some more specific routes (/32, ...) part of your network
and TAG the route with communities
so it can be filtered (dropped by the router)

Most useful when you have a public ASN and buy transit
Traffic is dropped before it is billed

Many Talks (NANOG, APRICOT, ...) on the topic and an RFC (5635)
> google RTBH or REMOTELY TRIGGERED BLACKHOLE

The goal is to skip the transit provider NOC and NOC response time in time of emergency.

Each ISP implements it differently ..
level3 > whois -h whois.ripe.net AS3356 | grep -B1 -A15 Blackhole

It is dangerous to be right in matters on which the established authorities are wrong
*Voltaire*

# FLOW ROUTES

**Use BGP to transmit firewall like rules**

- RFC 5575, Juniper routers (Alcatel / Perhaps IOS XR)
- Can be used to transproxy in the core things like ... spammers

**Match possible components making the flow**

- Prefix (source and destination)
- IP Protocol (list of <action, value>)
- Port (source, destination, either)
- ICMP (type, code)
- TCP flag
- Packet Len
- DSCP value
- Fragment (don't, is, first, last)

**Then take action**

- Drop, Rate-limit, Redirect

exabpg is the only OSS application to support Flow Routes

The secret of business is to know something that nobody else knows

*Aristotle Onassis*

exa networks

# REDIRECT / BLOCK TRAFFIC

**Intercept some traffic injecting BGP routes**

- the route must be more specific or have an higher LOCAL PREF

**Your own IPs**

- Move a machine to another geographical location
  - connected traffic always preferred to a gateway
- Intercept traffic
  - web server (using another server with destination NAT)

**Another network IPs**

- Block bad sources of traffic : spammers, proxies, TCP scanners, ...
  - You are affecting the return packets
    - it will not stop a UDP, SYN flood attack
    - will prevent TCP 3 way handshake (block the SYN-ACK)
- Force outgoing traffic to use one upstream over another
  - even if default routes and do not use BGP today

Success is a result, not a goal

*Flaubert*

# SERVICE IP ANNOUNCEMENT

**Use BGP to announce service IP**

An extra IP added to a server for the purpose of providing a public service (ie: pop, imap, web, reverse proxy, vpn IP, ...)

provide IP stability, not physically bound to a location/machine

people SHOULD use DNS entries ... but don't
firewall configuration, etc ...

Have servers announcing their own service IP
Server outage means the IP stops to be routed


Or provision service IPs from a centralised location

I have always believed that to succeed in life, it is necessary to appear to be mad and to act wisely
*Montesquieu*

# SERVICE IP ANNOUNCEMENT

**Single server**

Use GRACEFUL RESTART so the router does not forget the route for a programmed number of seconds when BGP goes down unexpectedly

**Active / Passive**

Use LOCAL PREFERENCE (BGP route preference)
Use ipvsadm on the active to still balance traffic

**Active/Active**

For machine within the same Layer 2, look at using OSPF
Otherwise ANYCAST (if suitable)

In revolution there are only two sorts of men, those who cause them and those who profit by them

*Napoleon Bonaparte*

# RESILIENCE

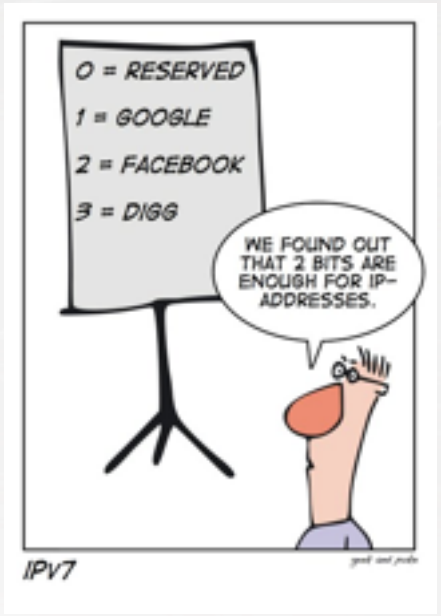
for IPv4
or IPv6

**Resilience with IPv6**

2x Router Advertisement
-> two default routes

BGP (over an IPv4 or IPv6 TCP connection)
-> announce the IPv6 service IP

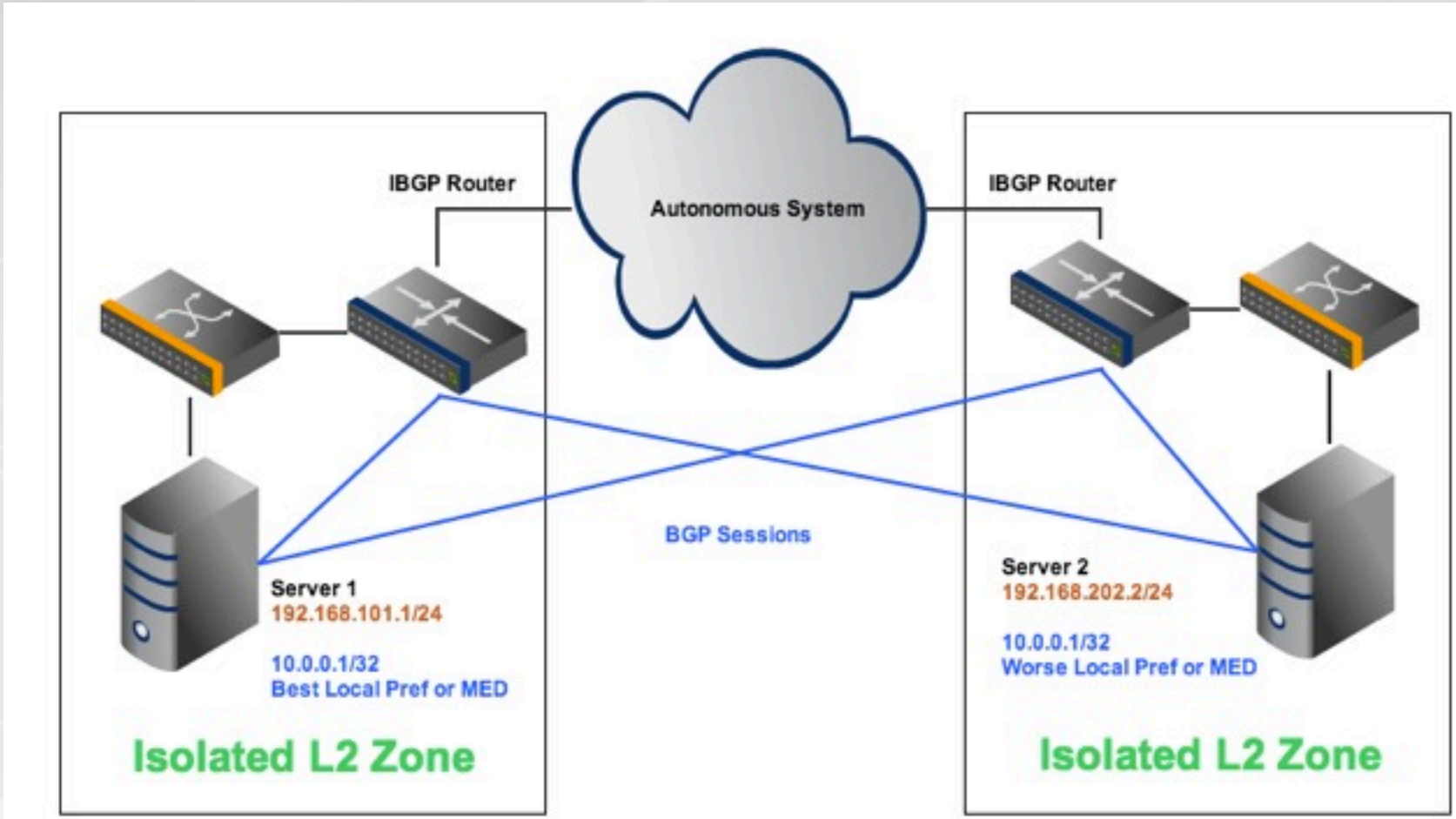
AVAILABLE TODAY




It is easier to ask for forgiveness than permission
*Stewart's law of retraction*

exa networks

# GEOGRAPHICALLY RESILIENT SERVICE




exa networks

# ACTIVE / PASSIVE SCENARIO

Configure IP /32 on the loopback interface
linux (debian/Ubuntu)

/ETC/NETWORK/INTERFACES

```
AUTO LO:SERVICE
IFACE LO:SERVICE INET STATIC
    ADDRESS 192.0.2.1
    NETMASK 255.255.255.255
    NETWORK 192.0.2.1
    BROADCAST 192.0.2.1
```

Control ARP broadcast (as more than one machine has one IP on its loopback) and RPF check

/ETC/SYSCTL.CONF

```
NET.IPV4.CONF.ALL.ARP_FILTER = 1
NET.IPV4.CONF.ALL.ARP_IGNORE = 1
NET.IPV4.CONF.ETH0.ARP_IGNORE = 1
NET.IPV4.CONF.ALL.ARP_ANNOUNCE = 2
NET.IPV4.CONF.ETH0.ARP_ANNOUNCE = 2
```

# ACTIVE / PASSIVE SCENARIO

Active Server : an exabgp configuration (version 1.2.0 +)

```
GROUP ANNOUNCE-MY-SERVICE-IP-OF-192.0.2.1 {
    # ETH0 10.0.0.1/24 GATEWAY 10.0.0.254 (HSRP/VRRP)
    LOCAL-ADDRESS 10.0.0.1;

    # WE SETUP AN IBGP CONNECTION
    LOCAL-AS 64520;
    PEER-AS 64520;

    STATIC {
        # 150 IS A BETTER LOCAL-PREFERENCE VALUE THAN 100 (DEFAULT VALUE)
        ROUTE 192.0.2.1/32 NEXT-HOP 10.0.0.1 LOCAL-PREFERENCE 150;
    }
    NEIGHBOR 172.16.0.1 {
        DESCRIPTION "BGP ROUTER 1 RUNNING HSRP/VRRP";
    }
    NEIGHBOR 172.16.0.2 {
        DESCRIPTION "BGP ROUTER 2 RUNNING HSRP/VRRP";
    }
}
```

# ACTIVE / PASSIVE SCENARIO

Passive Server : an exabgp configuration (version 1.2.0 +)

```
GROUP ANNOUNCE-MY-SERVICE-IP-OF-192.0.2.1 {
    # ETH0 10.0.0.2/24 GATEWAY 10.0.0.254 (HSRP/VRRP)
    LOCAL-ADDRESS 10.0.0.2;

    # WE SETUP AN IBGP CONNECTION
    LOCAL-AS 64520;
    PEER-AS 64520;

    STATIC {
        # 100 (DEFAULT VALUE) IS A WORSE LOCAL-PREFERENCE VALUE THAN 150
        ROUTE 192.0.2.1/32 NEXT-HOP 10.0.0.1 LOCAL-PREFERENCE 100;
    }
    NEIGHBOR 172.16.0.1 {
        DESCRIPTION "BGP ROUTER 1 RUNNING HSRP/VRRP";
    }
    NEIGHBOR 172.16.0.2 {
        DESCRIPTION "BGP ROUTER 2 RUNNING HSRP/VRRP";
    }
}
```

# ACTIVE / PASSIVE SCENARIO

Router : Router 1 (cisco) BGP configuration example

```
BGP 64520
    NO SYNCHRONIZATION
    BGP ROUTER-ID 172.16.0.1

    NEIGHBOR SERVICE-IP PEER-GROUP
    NEIGHBOR SERVICE-IP REMOTE-AS 64520
    NEIGHBOR SERVICE-IP DESCRIPTION SERVICE IPS
    NEIGHBOR SERVICE-IP EBGP-MULTIHOP 5
    NEIGHBOR SERVICE-IP UPDATE-SOURCE LOOPBACK1
    NEIGHBOR SERVICE-IP DEFAULT-ORIGINATE
    NEIGHBOR SERVICE-IP ROUTE-MAP BGP-SERVICE-IP IN
    NEIGHBOR SERVICE-IP ROUTE-MAP DENY-ANY OUT

    NEIGHBOR 10.0.0.1 PEER-GROUP SERVICE-IP
    NEIGHBOR 10.0.0.2 PEER-GROUP SERVICE-IP

    NO AUTO-SUMMARY
!
```

# ACTIVE / PASSIVE SCENARIO

Router : Router 1 (cisco) BGP configuration example

```
!
INTERFACE LOOPBACK1
 DESCRIPTION BGP
 IP ADDRESS 172.16.0.1 255.255.255.255
!
IP PREFIX-LIST SERVICE-IP SEQ 10 PERMIT 192.0.2.1/32
IP PREFIX-LIST SERVICE-IP SEQ 99999 DENY 0.0.0.0/0 LE 32
!
IP ACCESS-LIST STANDARD MATCH-ANY
 PERMIT ANY
!
ROUTE-MAP BGP-SERVICE-IP PERMIT 10
 MATCH IP ADDRESS PREFIX-LIST SERVICE-IP
 SET COMMUNITY NO-EXPORT ADDITIVE
!
ROUTE-MAP DENY-ANY DENY 10
 MATCH IP ADDRESS MATCH-ANY
!
```

# DYNAMIC SERVICE MIGRATION

**Permanent configuration generation**

1 – Regenerating BIRD/Quagga/OpenBGPD configuration on change
2 – Getting the daemon to reload its configuration
3 – Go back to 1

## There must be a better way ...

OpenBGPD bgpctl
BIRD birdc
Quagga / Zebra telnet ..

## There must be a better way .....

Logic will get you from A to B. Imagination will take you everywhere
*Albert Einstein*

exa networks

# HOW ?

## flap.sh

1 - take your favourite language : perl, python, lua, C, shell, french ! ...
2 - create a forever loop
3 - print what you want to do ...
4 - ... profit ?

```
#!/bin/sh

# ignore Control C
trap '' SIGINT

while `true`;
do
    echo "announce route 192.0.2.1 next-hop 10.0.0.1"
    sleep 10
    echo "withdraw route 192.0.2.1 next-hop 10.0.0.1"
    sleep 10
done
```

# CONTROL BGP

**BGP configuration**

```
neighbor 192.168.127.128 {
        description "will flap a route until told otherwise";
        router-id 198.111.227.39;
        local-address 192.168.127.1;
        local-as 65533;
        peer-as 65533;

        # add and remove routes when flap.sh prints
        process loving-flaps {
                run etc/processes/flap.sh;
        }
}
```



Success is a result, not a goal

*Flaubert*

# WANT SIMPLER ?

**BGP configuration**

```
neighbor 192.168.127.128 {
        router-id 198.111.227.39;
        local-address 192.168.127.1;
        local-as 65533;
        peer-as 65533;

        process default-name-for-watchdog {
             run etc/processes/monitor.sh;
        }

        static {
              route 172.10.0.0/16 next-hop 192.0.2.1 watchdog service-one;
        }
}
```

# CONDITIONAL ANNOUNCEMENT ?

**only announce what works !**

The watchdog ...

```
#!/bin/sh
trap '' SIGINT
while `true`;
do
    state=`check-if-all-ok`
        if [ “$state” = “up” ]; then
            echo "announce watchdog service-one"
        fi
        if [ “$state” = “down” ]; then
            echo "withdraw watchdog service-one”
        fi
        # pick its name from the process section name
        echo “announce watchdog”
        sleep 5
done
```

# GET IT !

http://code.google.com/p/exabpg/

# Questions ?

Yes, API control works with flow routes too

Judge a man by his questions rather than by his answers

*Voltaire*

exa networks

# Other Questions ?

**Thank you for coming and listening.**

thomas.mangin@exa-networks.co.uk

http://code.google.com/p/exabgp/

Judge a man by his questions rather than by his answers

*Voltaire*

exa networks