Global Server Load Balancing

Dima Krioukov [dima@nortelnetworks.com]
Alex Kit [akit@winstar.com]
October 24, 2000
Purpose

• Existing methods
• New technique
• Analysis
• Applicability considerations
Plan

• Introduction
  – What are ASPs?
  – Requirements to IDCs

• LSLB
  – Load Sharing NAT (LSNAT)
  – Direct Server Return (DSR)
  – Tunneling

• GSLB
  – DNS Based
  – Host Route Injection (HRI)
  – Triangle Data Flow (TDF)
  – Latest Trends

• New Technique – Virtual Block Injection (VBI)
  – Description
  – Testing
  – Analysis

• Applicability Considerations

• Conclusions and References
Abbreviations

- LB = Load Balancing/Balancer
- SLB = Server LB
- LSLB = Local SLB
- GSLB = Global SLB
- HA = High Availability
- RS = Real Server/Service
- VS = Virtual Server/Service
- VIP = VS IP address
- LSNAT = Load Sharing NAT
- DSR = Direct Server Return

- PRP = Proximity Report Protocol
- LRP = Load Report Protocol
- LPRP = PRP + LRP
- HRI = Host Route Injection
- VBI = Virtual Block Injection
- TDF = Triangle Data Flow
- IDC = Internet Data Center
- CDN = Content Delivery Network
- ASP = Application Service Provider
- CASP = Content/Collocation and Application Service Provider
- AIP = Application Infrastructure Provider
- xyP = ?

GSLB - 3
1. Introduction

Logic: GSLB ← IDC ← ASP ← Hosting
ASP

- End Customer
- ASP
- Applications
- Operations
- IDC
- ISP/Backbone
- Access
- Infrastructure

Operations
IDC

IDC

**Core (Routing)**

**Distribution (L3 Switching)**

Tier  Tier  Tier

LB Tier

**Load Balancing (L4 Switching)**

**Port Density (L2 Switching)**

Servers

SAN
Requirements to IDCs

- **High Availability (HA)**
  - Local
  - Global
- **Load Balancing (LB)**
  - Local
  - Global
  - Proximity (“including” congestion)
  - Load
2. Generic SLB and LSLB

SLB = VS $\rightarrow$ RS

- **Health Checking**
  - Layer 2
  - Layer 3
  - Layer 4
  - Layer 7

- **SLB Algorithm**
  - Round Robin
  - Least Connections
  - Server Response Time
  - Server Load
  - Hashing

- **SLB Forwarding**
  - Session Tables
  - Timers
LSLB Forwarding

- LSNAT
- DSR
- Tunneling
### LSNAT

<table>
<thead>
<tr>
<th>Segment</th>
<th>Layer</th>
<th>src/ dst</th>
<th>Ingress</th>
<th>Egress</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>L2</td>
<td>src</td>
<td>Router_MAC</td>
<td>Virtual_MAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>Virtual_MAC</td>
<td>Router_MAC</td>
</tr>
<tr>
<td>X</td>
<td>L3</td>
<td>src</td>
<td>Client_IP</td>
<td>Virtual_IP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>Virtual_IP</td>
<td>Client_IP</td>
</tr>
<tr>
<td>X</td>
<td>L4</td>
<td>src</td>
<td>Client_Port</td>
<td>Virtual_Port</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>Virtual_Port</td>
<td>Client_Port</td>
</tr>
<tr>
<td>Y</td>
<td>L2</td>
<td>src</td>
<td>LB_MAC</td>
<td>S1_MAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>S1_MAC</td>
<td>LB_MAC</td>
</tr>
<tr>
<td>Y</td>
<td>L3</td>
<td>src</td>
<td>Client_IP</td>
<td>S1_IP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>S1_IP</td>
<td>Client_IP</td>
</tr>
<tr>
<td>Y</td>
<td>L4</td>
<td>src</td>
<td>Client_Port</td>
<td>S1_Port</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>S1_Port</td>
<td>Client_Port</td>
</tr>
</tbody>
</table>
## LSNAT + Source NAT

<table>
<thead>
<tr>
<th>Segment</th>
<th>Layer</th>
<th>src/dst</th>
<th>Ingress</th>
<th>Egress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>L2</td>
<td>src</td>
<td>Router_MAC</td>
<td>Virtual_MAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>Virtual_MAC</td>
<td>Router_MAC</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>src</td>
<td>Client_IP</td>
<td>Virtual_IP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>Virtual_IP</td>
<td>Client_IP</td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>src</td>
<td>Client_Port</td>
<td>Virtual_Port</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>Virtual_Port</td>
<td>Client_Port</td>
</tr>
<tr>
<td>Y</td>
<td>L2</td>
<td>src</td>
<td>LB_V_MAC</td>
<td>S1_MAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>S1_MAC</td>
<td>LB_V_MAC</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>src</td>
<td>LB_V_IP</td>
<td>S1_IP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>S1_IP</td>
<td>LB_V_IP</td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>src</td>
<td>LB_V_Port</td>
<td>S1_Port</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dst</td>
<td>S1_Port</td>
<td>LB_V_Port</td>
</tr>
</tbody>
</table>
## Tunneling

### Diagram

- **Router**
- **LB**

### Table

<table>
<thead>
<tr>
<th>Layer</th>
<th>src/dst</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>src</td>
<td>R_MAC</td>
<td>LB_MAC</td>
<td>S1_MAC</td>
</tr>
<tr>
<td></td>
<td>dst</td>
<td>V_MAC</td>
<td>S1_MAC</td>
<td>R_MAC</td>
</tr>
<tr>
<td>L3</td>
<td>src</td>
<td>C_IP</td>
<td>Ext: LB_IP</td>
<td>Int: C_IP</td>
</tr>
<tr>
<td></td>
<td>dst</td>
<td>V_IP</td>
<td>Ext: S1_IP</td>
<td>Int: V_IP</td>
</tr>
<tr>
<td>L4</td>
<td>src</td>
<td>C_Port</td>
<td>C_Port</td>
<td>V_Port</td>
</tr>
<tr>
<td></td>
<td>dst</td>
<td>V_Port</td>
<td>V_Port</td>
<td>C_Port</td>
</tr>
</tbody>
</table>
3. GSLB

- DNS Based
- HRI
- TDF
- Latest Trends
3.1 DNS Based

GSLB = Name $\rightarrow$ VS (DNS+)

- **Smart DNS**
  - Load and availability awareness $\rightarrow$ Load Report Protocol (LRP)
  - Proximity and congestion awareness $\rightarrow$ Proximity Report Protocol (PRP)

- **LB DNS Functionality**
  - DNS Server
  - DNS Proxy
    - Caching
  - DNS Traffic Intercept
LPRP

- **Transport**
  - UDP
  - TCP
  - HTTP

- **Operation**
  - Periodic Updates
  - Periodic Requests
  - Triggered Updates
PRP

- RTT
- Effective bandwidth
- Number of hops
- Number of AS hops
- IGP metric

Proximity to the client LDNS, not to the client
**LRP**

- **VS Health**
  - Up
  - Down
  - Backup only

- **VS Load**
  - Number of sessions
  - Response Time

- **LB Load**
  - Number of sessions
  - Capacity threshold
  - CPU

- **RS/Content Load**

- **Network Load**
  - bps
  - pps

- **QoS**

- **Security**
How it works

1. Client queries the ADNS for the LDNS IP address.
2. The ADNS resolves the LDNS domain name to an IP address.
3. The Client contacts the LDNS using the resolved IP address.
4. The LDNS forwards the request to one of the IDCs.
5. The Client is directed to the appropriate IDC by the IDC's Load Balancer (LB).
6. The Client communicates directly with the server(s) within the IDC.
How it works
Analysis

Pros
• Accurate load info
• Accurate proximity info
• Perfect solution… in some cases and if certain conditions are met

Cons
• DNS – wrong target
• Proximity between client and its LDNS
• Caching
  – LB
  – LDNS
  – Application
• Complexity
• Hard to find optimal values for various timers (TTL, cache timeouts, etc.) and prefix lengths
3.2 HRI

GSLB = Routing+

• To what?
  – BGP
  – IGP

• By what?
  – RS
  – Router
  – LB
To what

- IGP?
- BGP
  - Route filtering (both ways)
  - No ECMP
By what

RS

IDC1
Router
BGP
RS

IDC2
Router
BGP
RS
By what

Router

IDC1
- Router
- RS

IDC2
- Router
- LB
- RS
- RS
By what
Analysis

Pros

• Simplicity
• No new protocols are needed
• Proximity is handled by routing
• Load handling?

Cons

• Single backbone*
  – Its own
  – Single ISP
• Too many routes
• Less accurate load and proximity info
  – Only local load
  – Optimal routing?
• Route flapping*
3.3 TDF

GSLB = X + TDF

- NAT Based
- Tunneling
Why “wrong” IDC?

• Failure of, disabled or non-implemented LPRP
• Cached DNS records
• Other retardation effects (LPRP, BGP)
## NAT Based

<table>
<thead>
<tr>
<th>L3</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>src</td>
<td>C</td>
<td>C</td>
<td>V1.1</td>
</tr>
<tr>
<td>dst</td>
<td>V1.1</td>
<td>V2.2</td>
<td>C</td>
</tr>
</tbody>
</table>

1. Client
2. IDC1, “wrong”
3. IDC2, “right”
"Remote Servers"

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>src</td>
<td>C</td>
<td>V1.1</td>
<td>V2.1</td>
<td>V1.1</td>
</tr>
<tr>
<td>dst</td>
<td>V1.1</td>
<td>V2.1</td>
<td>V1.1</td>
<td>C</td>
</tr>
</tbody>
</table>

Client

IDC1, "wrong"

IDC2, "right"
Tunneling

Next section
Analysis

Pros
• Fixes errors optimally

Cons
• `ip verify reverse-path`

Diagram:
- **Client** connecting to **Router**.
- **Router** connects to **IDC2, “right”**.
- **Router** connects to **IDC1, “wrong”**.
Analysis

**Pros**
- Fixes errors optimally

**Cons**
- `ip verify reverse-path`

Diagram:
- Client (connected to Router)
- Router (connected to IDC1, labeled "wrong")
- Router (connected to IDC2, labeled "right")
3.4 Latest Trends, Radicalism

- Internet infiltration
- Going to the client edge
- Going to the client
- Modifying the client

- LB presence in strategic locations (HydraGPS, Speedera)
- LDNS modifications (Speedera)
- Application modifications (SRV RRs)
Internet Infiltrations
Internet Infiltrations

Diagram showing the relationship between Customer, Client, LB, IDC1, and IDC2.
LDNS modifications in CDNs
4. Virtual Block Injection (VBI)

• Inject not VS host routes, but blocks of GSLB’ed VSs → IDC (LB) failures are handled by the routing protocol

• Use tunneling TDF in case of individual VS failure
How it works
How it works
How it works
Testing

Needed
• LB
• BGP
• Tunnels

Linux
• Linux Virtual Server (LVS, Wensong Zhang, Julian Anastasov)
• Zebra
• Tunnels
Analysis

Pros

• All of HRI, plus
• No host route injection
• Working TDF
• Perfect VS health handling
• VS load → LRP
• Obvious simplifications in more “ideal” cases

Cons

• LB load → stop advertisement?
• BGP – proximity tool?
• Discontinuous AS?
• Route flapping!
Route Flapping

ISP1
IDC1, R1/20
IDC2, R2/20

AS1
AS2

Client

TCP UDP

UDPTCP

Route Flapping
Solution for UDP

Session table entry exchange for long sessions
Solution for UDP

Session table entry exchange for long sessions
Solution for TCP

If LB receives packet
- Destined to a VS
- No SYN
- No session table entry
- Not via the tunnels
  Forward via all the tunnels
5. Applicability Considerations

GSLB of

• Small number of VSs (or RSs)
  – by an ISP*
  – by its customer

• Big number of VSs (between IDCs)
  – CASP = ISP
  – CASP ≠ ISP
    • CASP has its own backbone
      - CASP does not have control over customer access
      - CASP has control over customer access**
    • CASP does not have its own backbone
      - CASP is multihomed to the same ISP
      - CASP is multihomed to different ISPs*
6. Conclusions

• No ideal GSLB method

• For some “ideal” network scenarios, there are some “ideal” solutions

• For realistic network scenarios, there are rapidly improving realistic solutions

• Good competition

• Lack of comparative testing in the production-like environment
References

- On ASPs: Nortel, ASP Industry Consortium, Network Magazine, IRG
- Vendors: Alteon, ArrowPoint, Foundry, F5, Cisco, Nortel, Radware, HydraWEB, Speedera, Resonate
- RFCs: LSNAT, SRV, DNS for LB, SLB draft (work in progress)
- VBI Testing: http://www.krioukov.net/~dima/VBI/