

Load balancerと俺

ブロードコムコミュニケーションズ
システムズ株式会社

甲野 謙一



私は誰



世界を船で旅していたころ
@シンガポール

甲野 謙一(36) as KonoKono

- 2000年から様々な形でLoad Balancerを担当
- BrocadeではADXの機能設計やPOST系を担当

ブロード

ー 事業

- L2/L3とL4-7スイッチ
- Fiber Channel SANスイッチ

今は、SANもIP/LANも両方やっています



今日お話しすること

- 商用LB業界を振り返る
- Load Balancer基礎
- L2DSRとL3DSR構成 ←メインTOPIC
- Brocade LAB環境の紹介

Load Balancer 業界の変遷

とメジャープレイヤー達



founded in 1996

2008



Gigabit Ethernet NIC
Jumbo Frame

Bought by Nortel Networks



Bought by Radware



founded in 1996



2000

2009

2010



1997 Local Director



1998

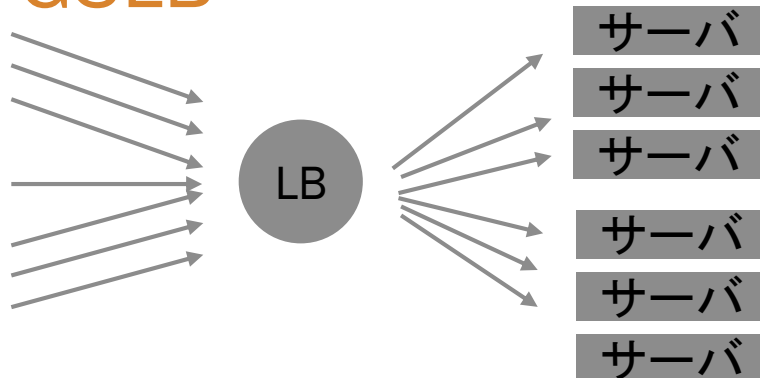
founded in 1996



Load Balancingとは

復習

SLBとGSLB



LB技術により、集中と分散をコントロールし、運用を最適化

SLB

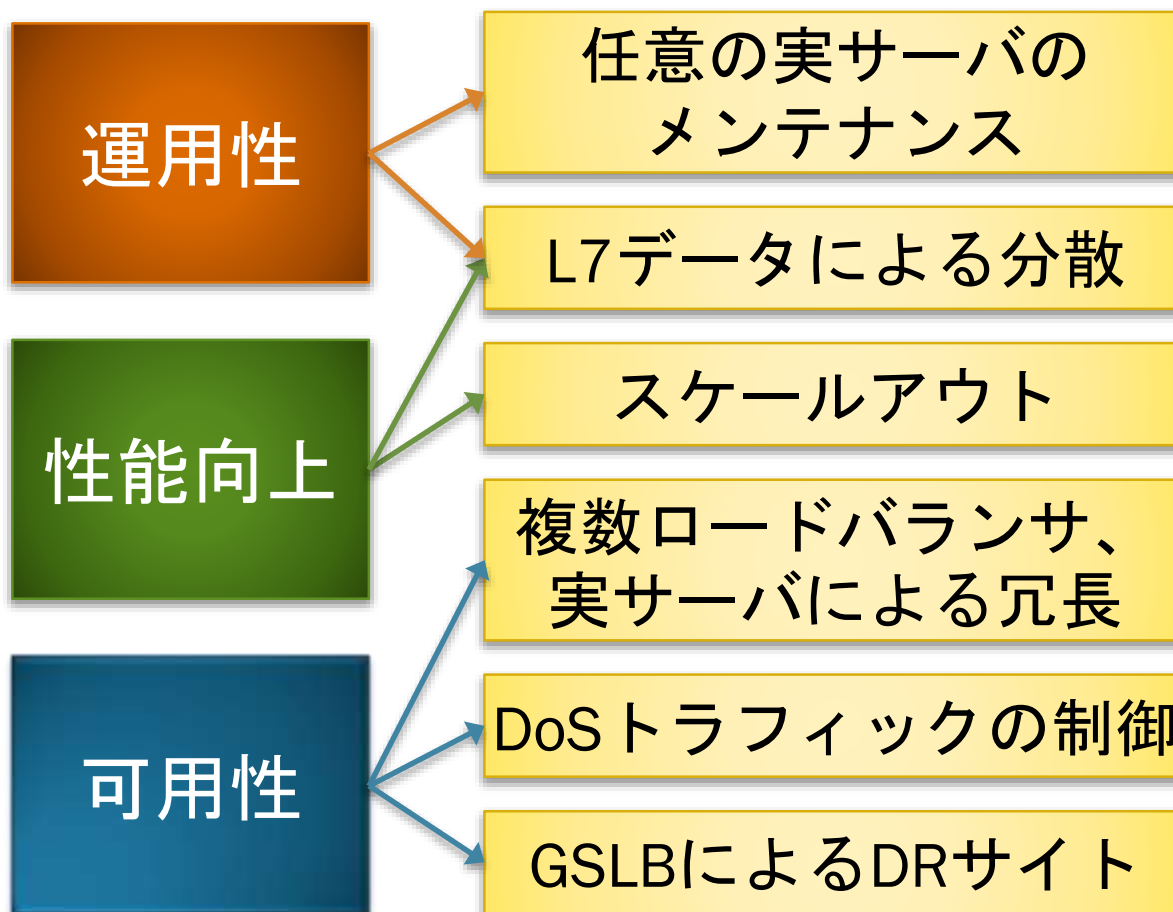
- 特定の宛先IPに集中したTrafficを分散させる技術。

GSLB

- Trafficが特定の宛先IPに集中する前に、Trafficを分散させる技術。LB業界では、DNSによる分散が一般的。

Load balancer

よくある導入の目的

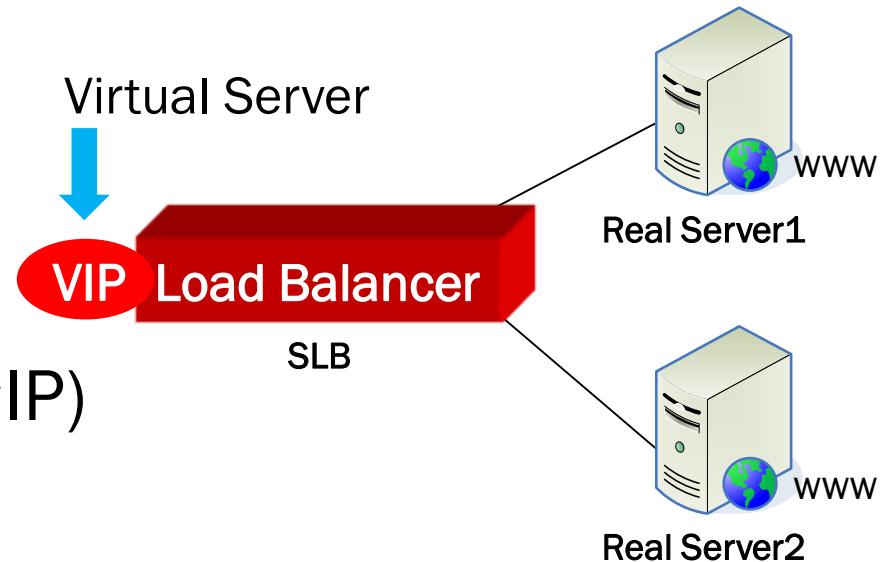


Load Balancerとは

復習

4つの基本機能

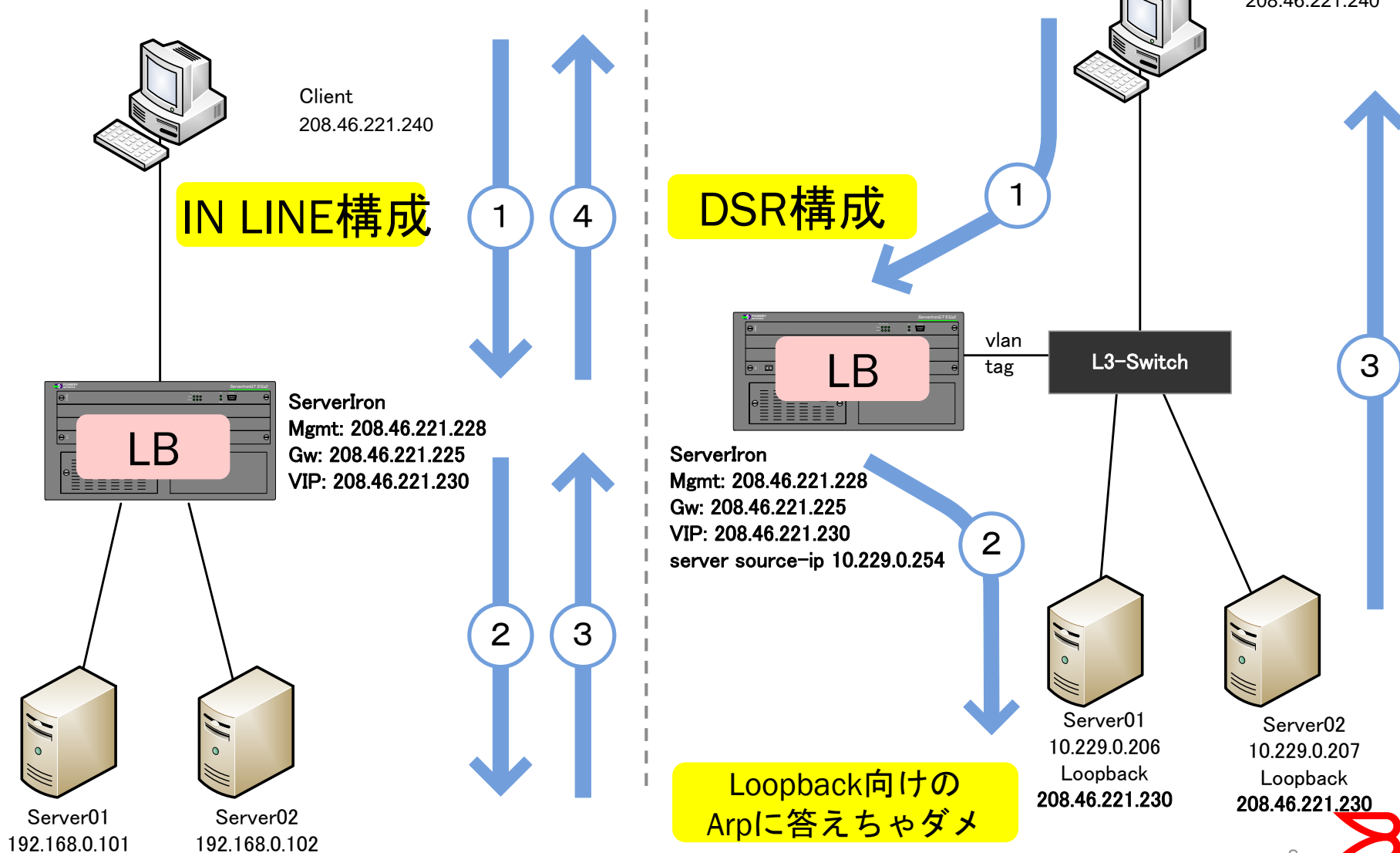
1. 負荷分散機能
2. ヘルスチェック機能
3. アドレス変換機能(MACとIP)
4. セッション維持機能



一般的なLB構成

復習

IN LINEとDSRのPACKET Walkthrough



なぜいまL3DSRか

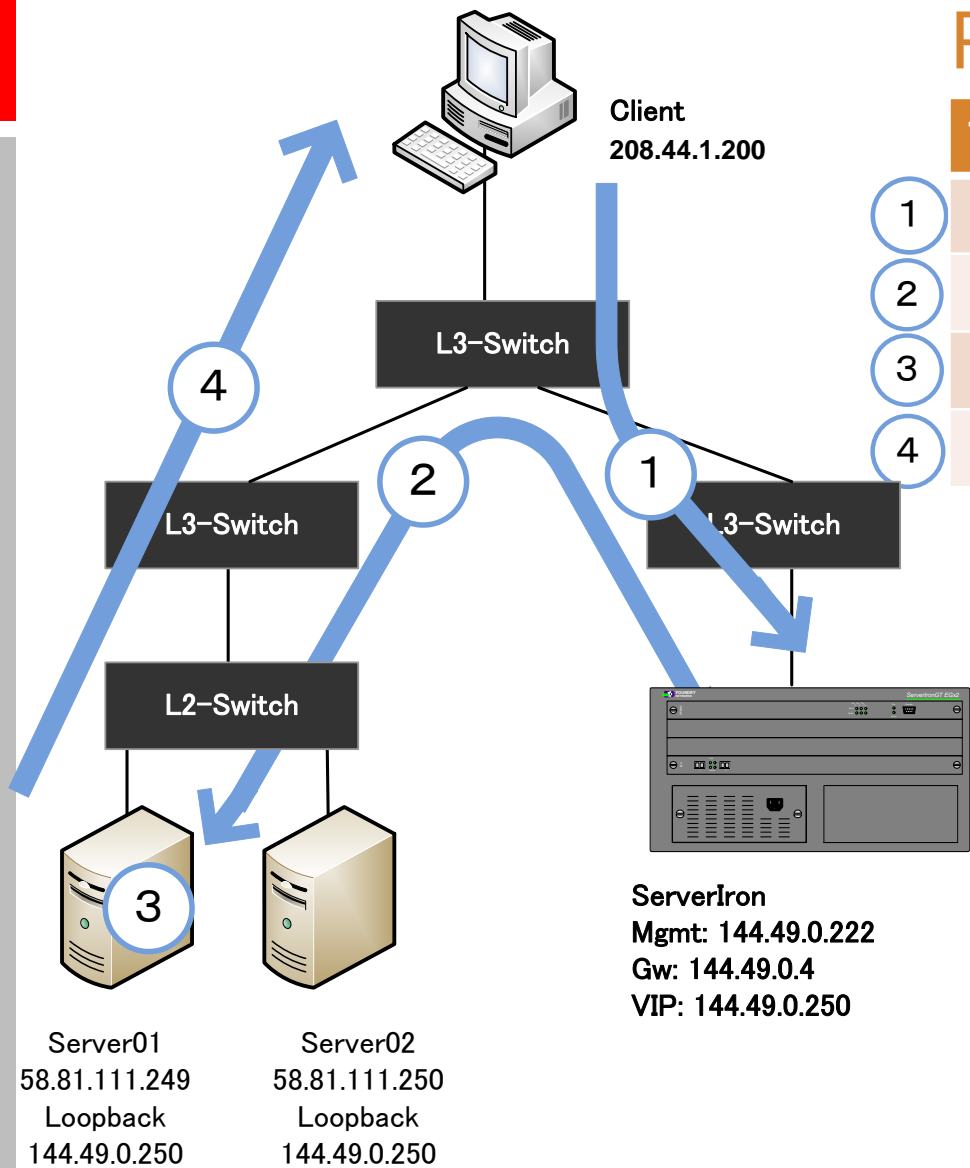
検討項目	L2DSR	L3DSR
パフォーマンス (スループット)	Outbound trafficは LBを経由しない。	←同じ
クライアントの Source IPを変更	SNATしないので、 SIPはそのまま。	←同じ
LBが全滅した時	サーバ自身が サービスを継続可能	←同じ
L3をまたいだReal Server設置	×	可能！

- 構成が柔軟なため、インターネット系のCloudサービスにうまくはまる。
- Anti Load Balancerの人(会社)であっても、DSRであれば、柔軟に構成できるため、In-LINEと比較して導入が容易。
- A10、Brocade、CitrixがL3DSR構成を正式サポート。F5でもできるはず。

L3DSR構成の話

ここ3年の話

PACKET Walkthrough



	TOS	Source IP	Destination IP
1	0x0	204.44.1.200	144.49.0.250
2	0x7	204.44.1.200	58.81.111.249
3	0x7	204.44.1.200	58.81.111.249
4	0x0	144.49.0.250	204.44.1.200

Serverが宛先IPを
144.49.0.250に変更する点が
ポイント

FreeBSD用のReference module
9.1/8.3/7.4-RELEASE

<https://github.com/yahoo/l3dsr>

L3DSR Packet Walkthrough

DSCP⇔VIP変換表に関して

DiffServは6bitしかないのに、IPアドレスはどう格納しますか。

Answer: LBとサーバそれぞれで、変換テーブルを持っています。

FreeBSDの
変換表

```
l3dsr01# sysctl -a | grep dscp↓  
net.inet.ip.dscp_rewrite.7: 144.49.0.250↓  
net.inet.ip.dscp_rewrite.enabled: 1↓
```

LBの
変換表

```
server virtual vip0 144.49.0.250  
tos-marking 7 hc-l3-dsr↓  
port http↓  
bind http rs5 http↓
```

VIP

DSCP

L3DSR Packet Walkthrough

RealServer上でPacketを見てみます。

No.	Time	Source	Destination	Protocol	Length	Info
25	3.089689	208.44.1.200	58.81.111.249	TCP	74	38655 > http [SYN] Seq=0 win=14600 Len=0 MSS=1460 SACK_PERM=1 TSval=46966173 TSecr=0 ws=16
27	3.089975	208.44.1.200	58.81.111.249	TCP	66	38655 > http [ACK] Seq=1 Ack=1 Win=14608 Len=0 TSval=46966173 TSecr=941726988
28	3.090335	208.44.1.200	58.81.111.249	HTTP	240	GET /index.html HTTP/1.1
30	3.090673	208.44.1.200	58.81.111.249	TCP	66	38655 > http [ACK] Seq=175 Ack=213 win=15680 Len=0 TSval=46966173 TSecr=941726988
31	3.091423	208.44.1.200	58.81.111.249	TCP	66	38655 > http [FIN, ACK] Seq=175 Ack=213 win=15680 Len=0 TSval=46966173 TSecr=941726988
34	3.091781	208.44.1.200	58.81.111.249	TCP	66	38655 > http [ACK] Seq=176 Ack=214 win=15680 Len=0 TSval=46966173 TSecr=941726988

Differentiated Services Field: 0x1c (DSCP 0x07: Unknown DSCP; ECN 0001 11.. = Differentiated Services Codepoint: Unknown (0x07) ..00 = Explicit Congestion Notification: Not-ECT (Not ECT))
Total Length: 60
Identification: 0x767b (30331)

Protocol: TCP (6)
Header checksum: 0xc0aa [correct]
Source: 208.44.1.200 (208.44.1.200)
Destination: 58.81.111.249 (58.81.111.249)
[Source GeoIP: unknown]
[Destination GeoIP: unknown]
Transmission Control Protocol, Src Port: 38655 (38655), Dst Port: http (80), Seq: 1, Ack: 1, Len: 174
Hypertext Transfer Protocol
GET /index.html HTTP/1.1\r\n
[Expert Info (Chat/Sequence): GET /index.html HTTP/1.1\r\n]
Request Method: GET
Request URI: /index.html
Request Version: HTTP/1.1
User-Agent: curl/7.22.0 (x86_64-pc-linux-gnu) libcurl/7.22.0 OpenSSL/1.0.1 zlib/1.2.3.4 libidn/1.23 librtmp/2.3\r\nHost: 144.49.0.250\r\nAccept: */*\r\n\r\n[Full request URI: http://144.49.0.250/index.html]

Differentiated Services Codepoint (ip.d... Packets: 42 Displayed: 6 Marked: 0 Load time: 0:00.015 Profile: Default

DSCP 0x07
0x07(HEX)
= 28(DEC)



L3DSR Packet Walkthrough

Real Server自身にPingしてみます。

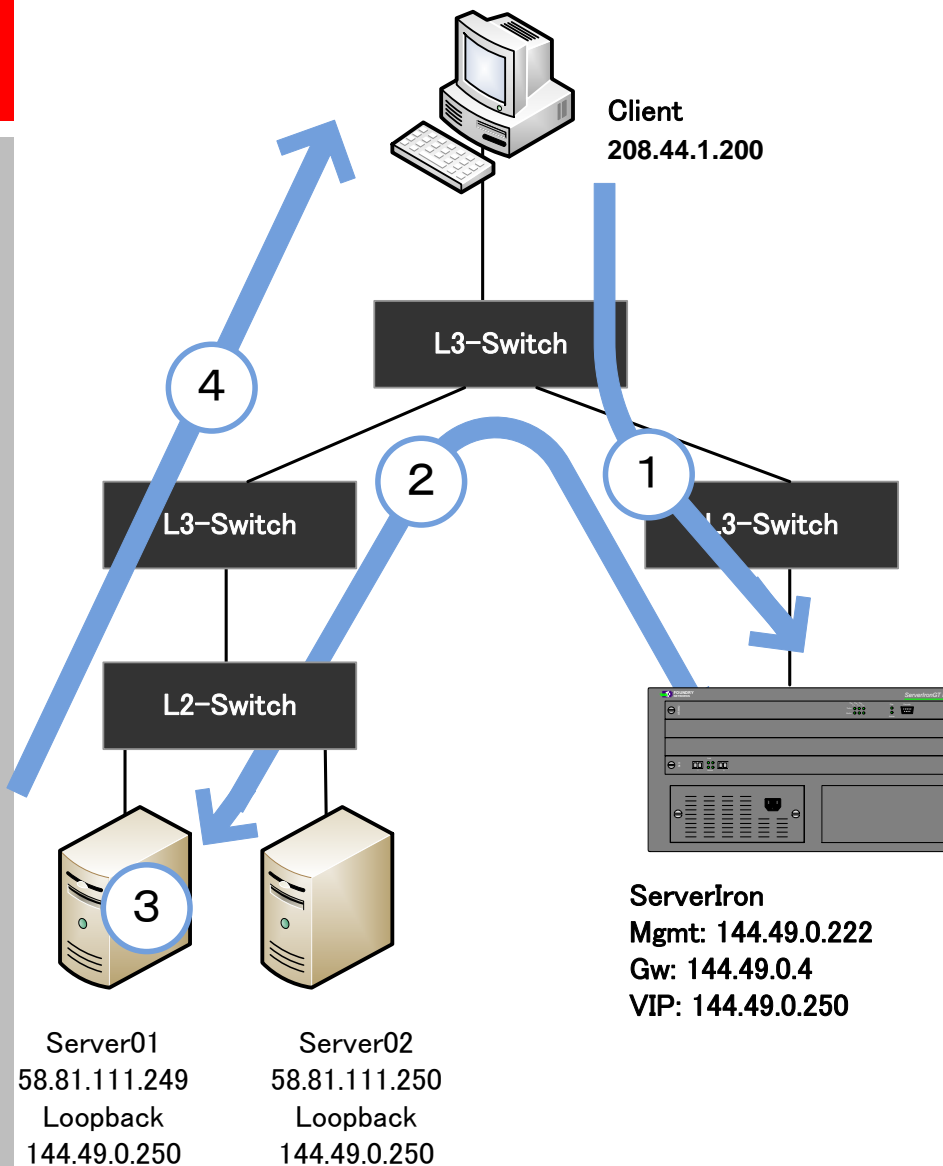
```
l3dsr01# ping -c 3 -z 28 58.81.111.249↓
PING 58.81.111.249 (58.81.111.249): 56 data bytes↓
64 bytes from 144.49.0.250: icmp_seq=0 ttl=64 time=0.027 ms↓
64 bytes from 144.49.0.250: icmp_seq=1 ttl=64 time=0.027 ms↓
64 bytes from 144.49.0.250: icmp_seq=2 ttl=64 time=0.029 ms↓
↓
--- 58.81.111.249 ping statistics ---↓
3 packets transmitted, 3 packets received, 0.0% packet loss↓
round-trip min/avg/max/stddev = 0.027/0.028/0.029/0.001 ms↓
l3dsr01# ↓
```

ping commandのoptionにおける `-z 28`は、
`0x07(16進数)`を10進数に変換した値です。

```
kkono@l3dsr03:/home/kkono % man ping↓
PING(8)                                FreeBSD System Manager's Manual
↓
~省略~↓
↓
-z tos  Use the specified type of service.↓
```



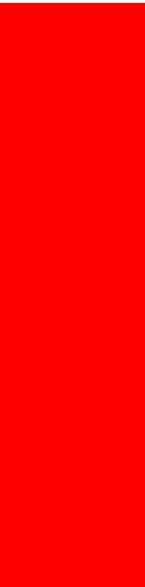
L3 DSR構成の話(おさらい)



ポイントのおさらい

- LBとServerは同じ変換表を共有
- LBは、宛先IPを変換してServerに送信。
- ServerはDSCP bitを確認。
- Serverは変換表に基づいて宛先をloopbackアドレスに書き換える。
- 後は、L2DSRと同様、ServerはVIPをSourceにしてClientに送信

その他 Kono Kono LABの紹介

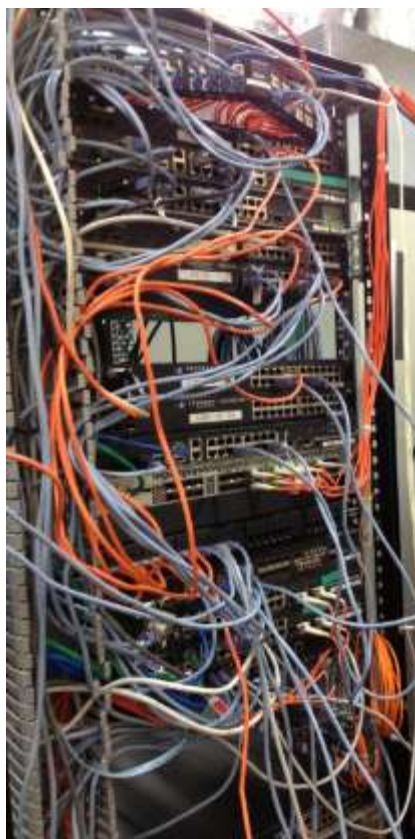


Brocade LAB環境の紹介

私が占有しているのだけです。



LB 4台
Server 7台
L3 2台



LB 10台
L3 8台
FC Switch 1台



Server 12台
esxi
10g x 2 port NIC
1g x 4 port NIC



7面モニター！



Thank you!

