Developing SLA apps by using ORACLE Solaris 11.3 DSCP flows

Samsi Serghei

blog: http://sscdvp.blogspot.com

Real-world SLA application

Codul sursă disponibil online:

https://github.com/sscdvp/flow-mgmt

Implementat în termen de 30 de zile datorită tehnologiei anuntate de ORACLE in 03.2015

Solaris Flow

Din manual flowadm(1M):

"... a flow is defined as a set of attributes based on Layer 3 and Layer 4 headers, which can be used to identify a protocol, service, or a virtual machine"

"... can be used on any type of data link, including physical links, virtual NICs, and link aggregations"

Caracteristice de bază ale flow

- Gestionează QoS pentru stiva virtualizată de reţea
- Flow QoS este integrat în stiva de protocoale și nu este un layer separat
- Diferențierea serviciilor se bazează pe atributele L3/L4:
  - protocol (UDP/TCP/SCTP/ICMP) – se suportă IPv6
  - adresă IP (SRC/DST) – se acceptă masca de reţea
  - port (SRC/DST)
  - DS field – se acceptă valoarea şi masca
- Efectuarea controlului de bandă cu un efort minim
- Partajarea lățimei de bandă PNIC/VNIC între mai mulţi clienţi. Beneficienții pot fi VM-uri sau chiar socket-uri
- Clasificarea traficului
- Marcarea traficului prin DSCP
- Integrat în Solaris Zones (administratorul zonei poate gestiona flow-urile aferente)

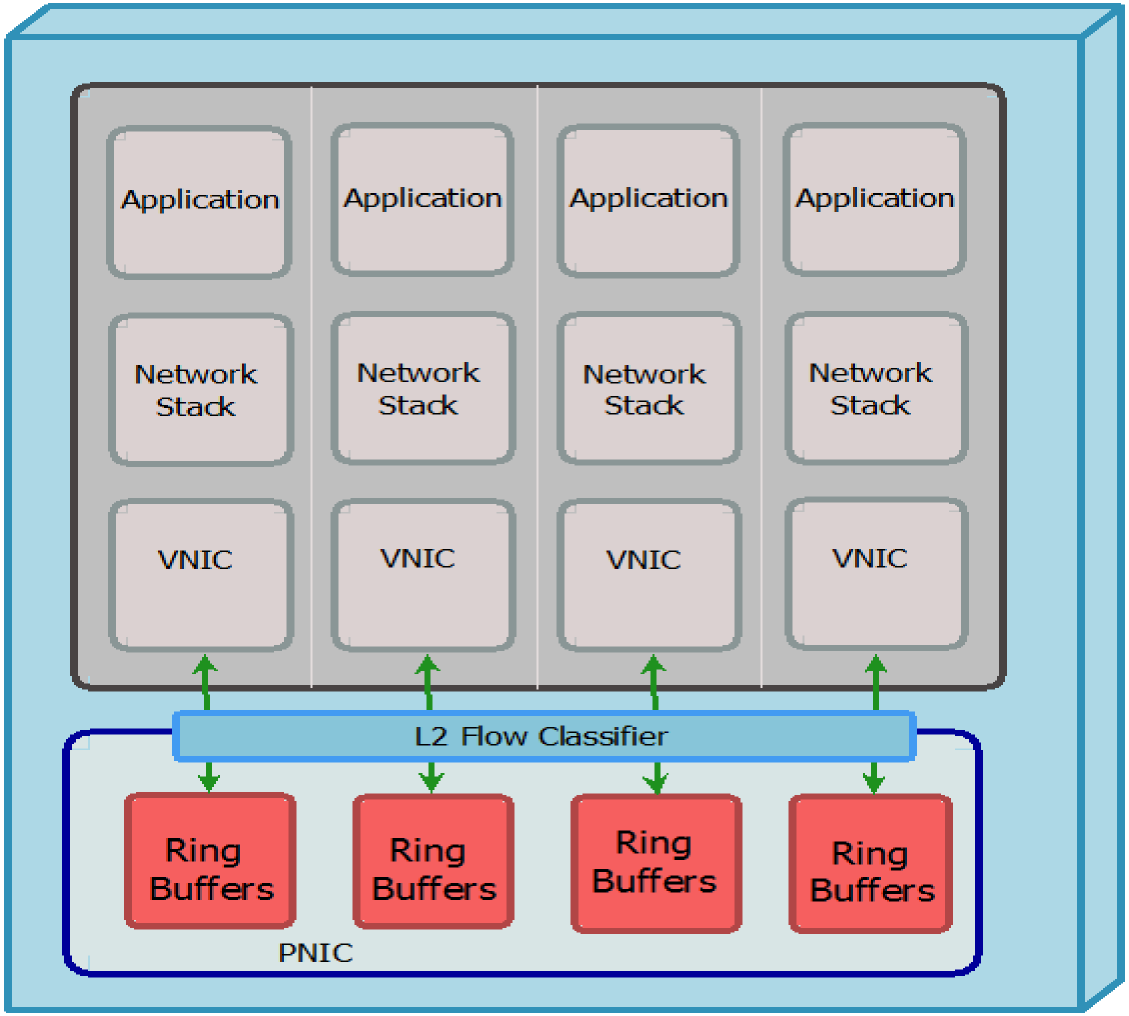
November 2015	Se adaugă: marcarea DSCP, flow-uri unidirecţionale, ridicarea constrângerilor privind combinaţia atributelor flow pe un datalink, flow ranking	**Oracle Solaris 11.3.0.30.0**
May 2015	Se adaugă: marcarea DSCP	**Oracle Solaris 11.2.8.4.0**
August 2014	Se adaugă: componentul SDN - application-driven flows (SO_FLOW_SLA), prioritizarea	**Oracle Solaris 11.2.0.0.42**
November 2011	Se adaugă suport pentru Solaris Zones	**ORACLE Solaris 11 11/11**
November 2008	Prima apariţie a elementului cheie în virtualizarea de reţea – Solaris flows: controlul lăţimii de bandă, lăţimea zero dacă e dorită sistarea traficului, moştenirea set-urilor CPU de la datalink atribuit, stocarea configuraţiei flow-urilor în fişiere pentru păstrare după restart	**OpenSolaris (Crossbow)**

Cronologia evoluţiei Solaris Flow

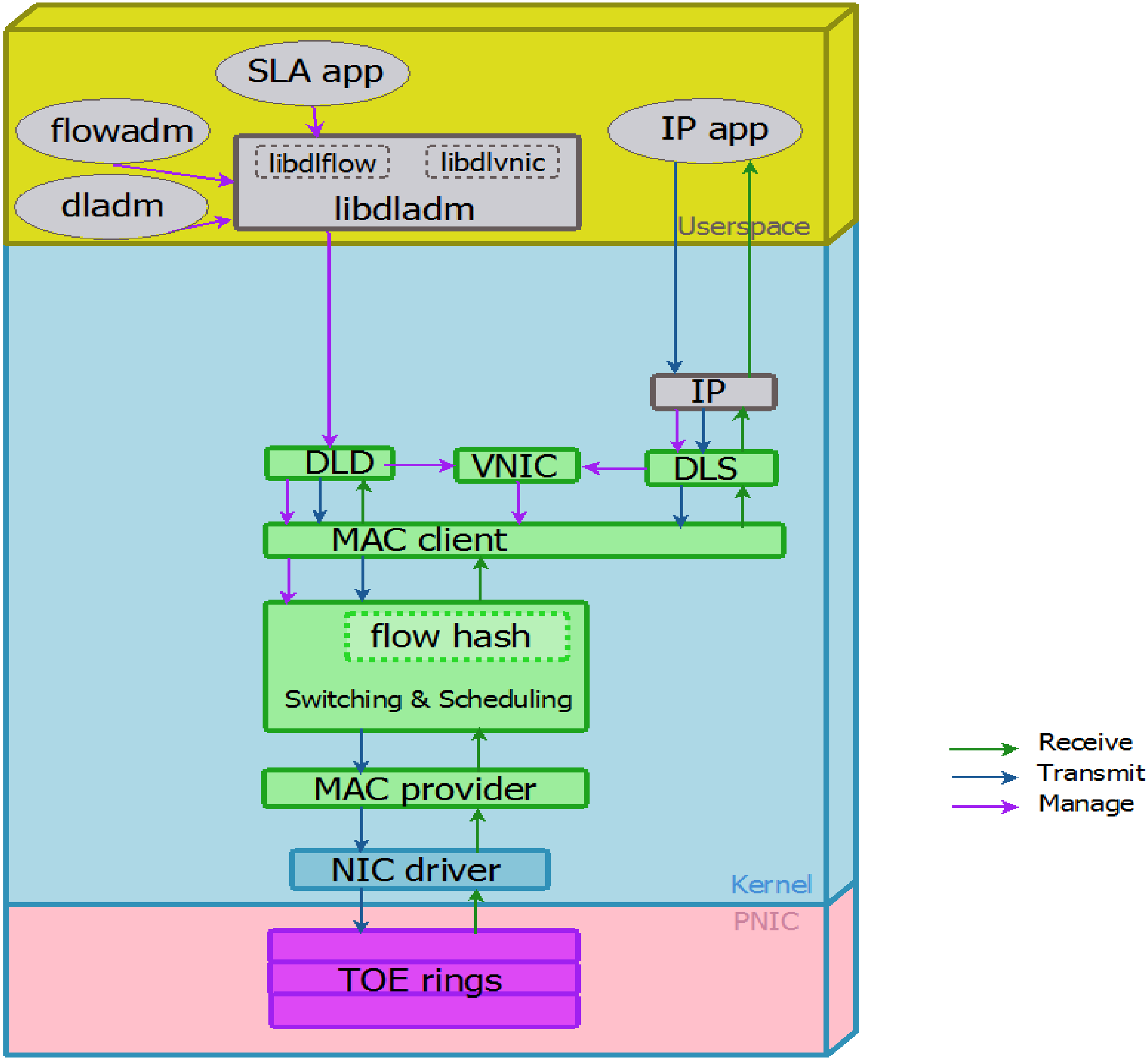
Arhitectura virtualizare de reţea ORACLE Solaris

- Virtualization lane: conţine resurse hardware şi software destinate pentru procesarea traficului
- Resursele PNIC: ring-urile Tx şi Rx
- Resursele MAC: softring-urile
- Resursele de transport: cozile de serializare
- Resursele CPU

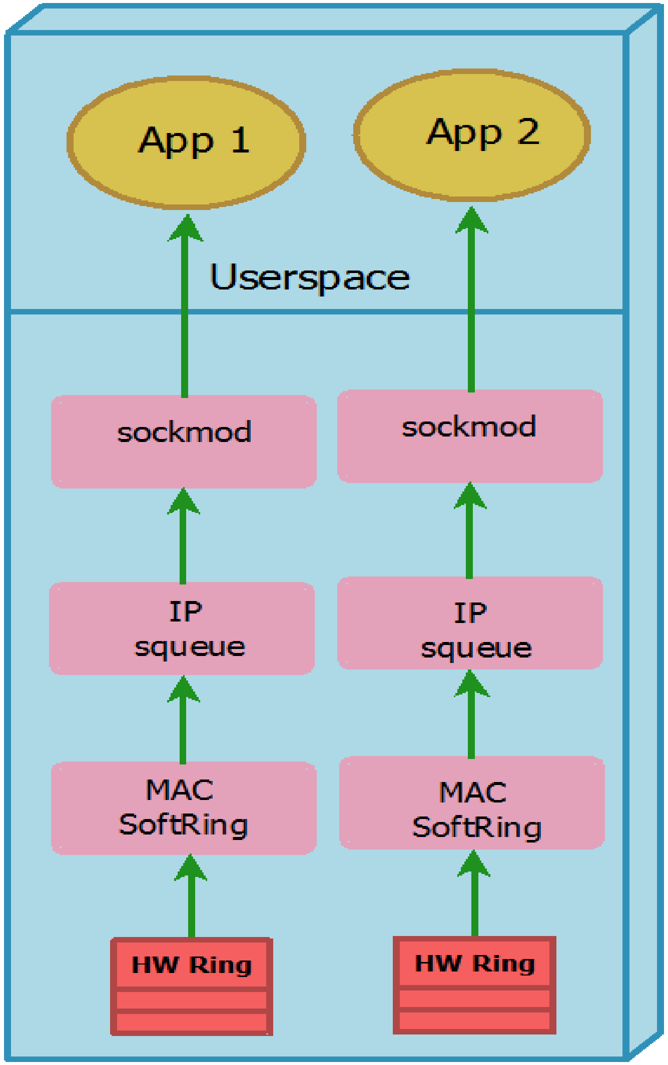
Izolarea și controlul resurselor în stiva virtualizată de rețea ORACLE Solaris

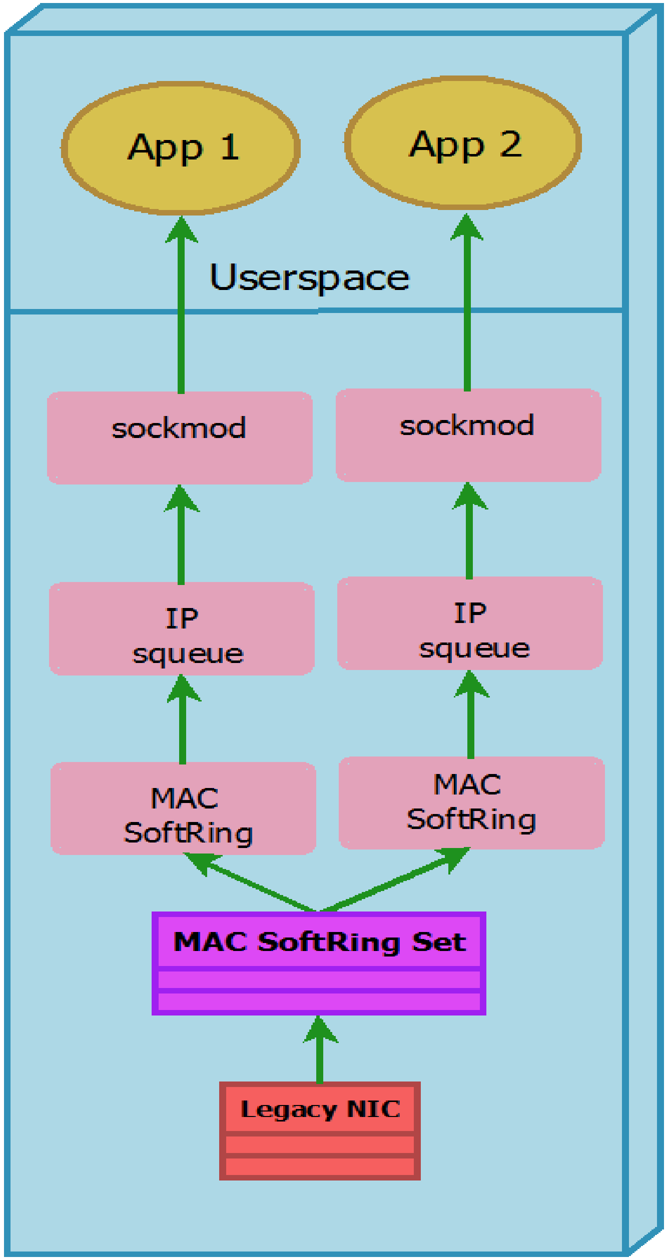


Componente stivei virtualizate



Rx ring-urile HW și SW





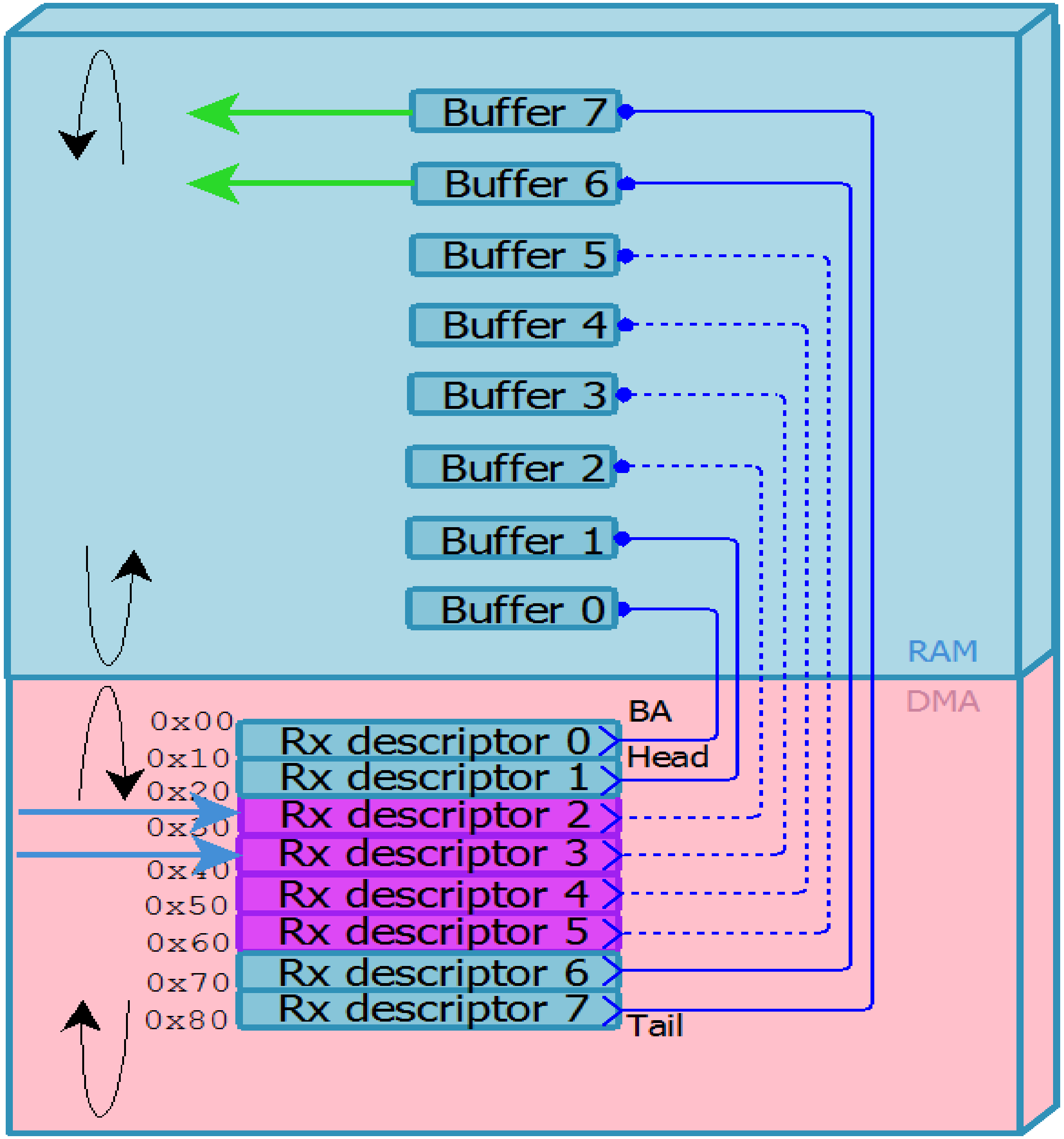
Interacțiunea între TOE (*TCP/IP Offload Engine*) şi sistem

- Afinitatea intreruperilor MSI (Receiver Side Scaling)
- Afinitatea pachetelor ce constituie un flow (Receiver Packet Steering)
- Afinitatea la nivel de virtualization lane (Receive Flow Steering)
- Pachetele sa fie procesate in batch
- Filtrarea direct pe PNIC

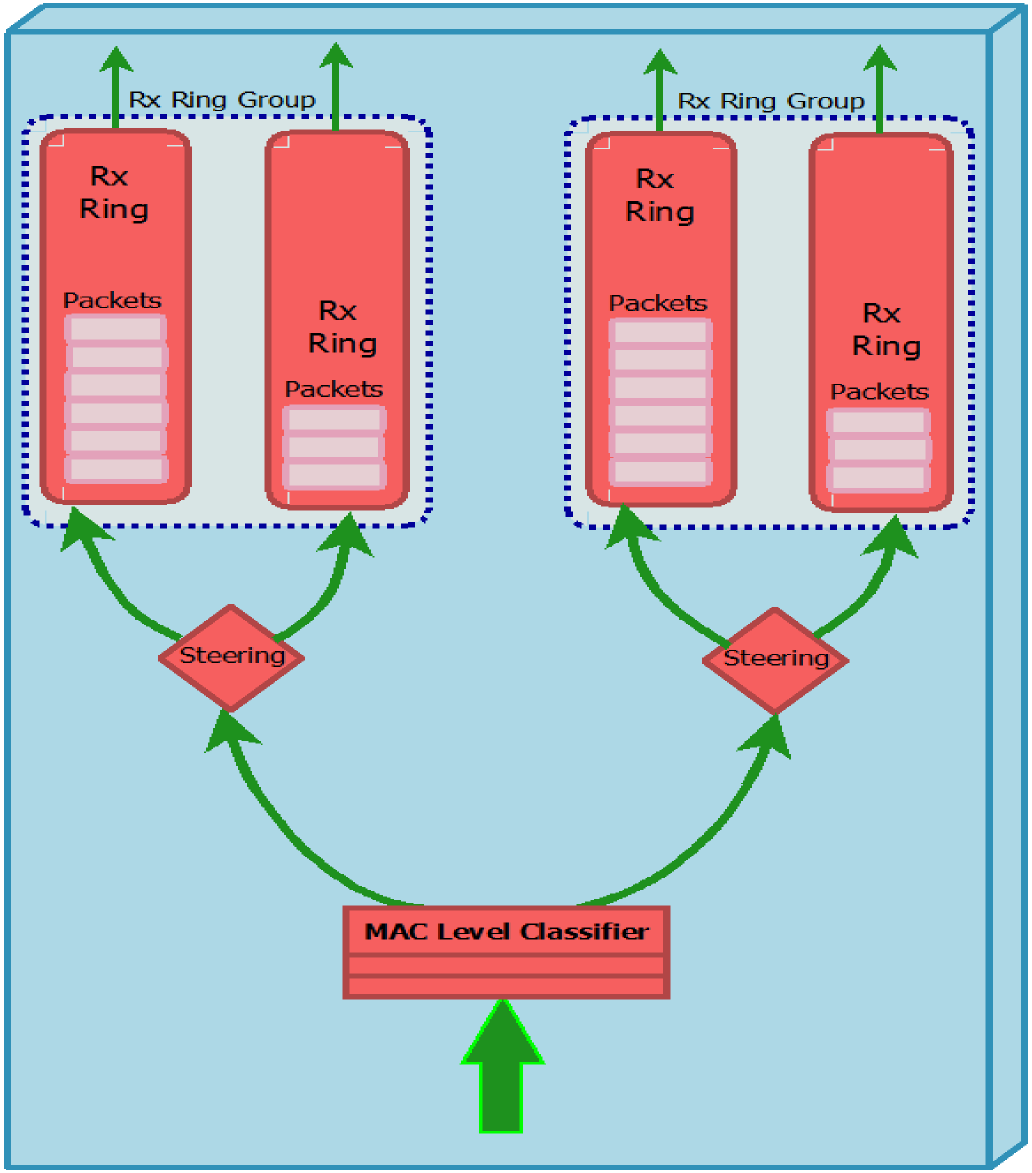
Avantaje RFS

- Independenta de NIC hardware
- Orice protocol nou poate fi adoptat in filtrele software
- Utilizeaza IPI si nu afecteaza IRQ

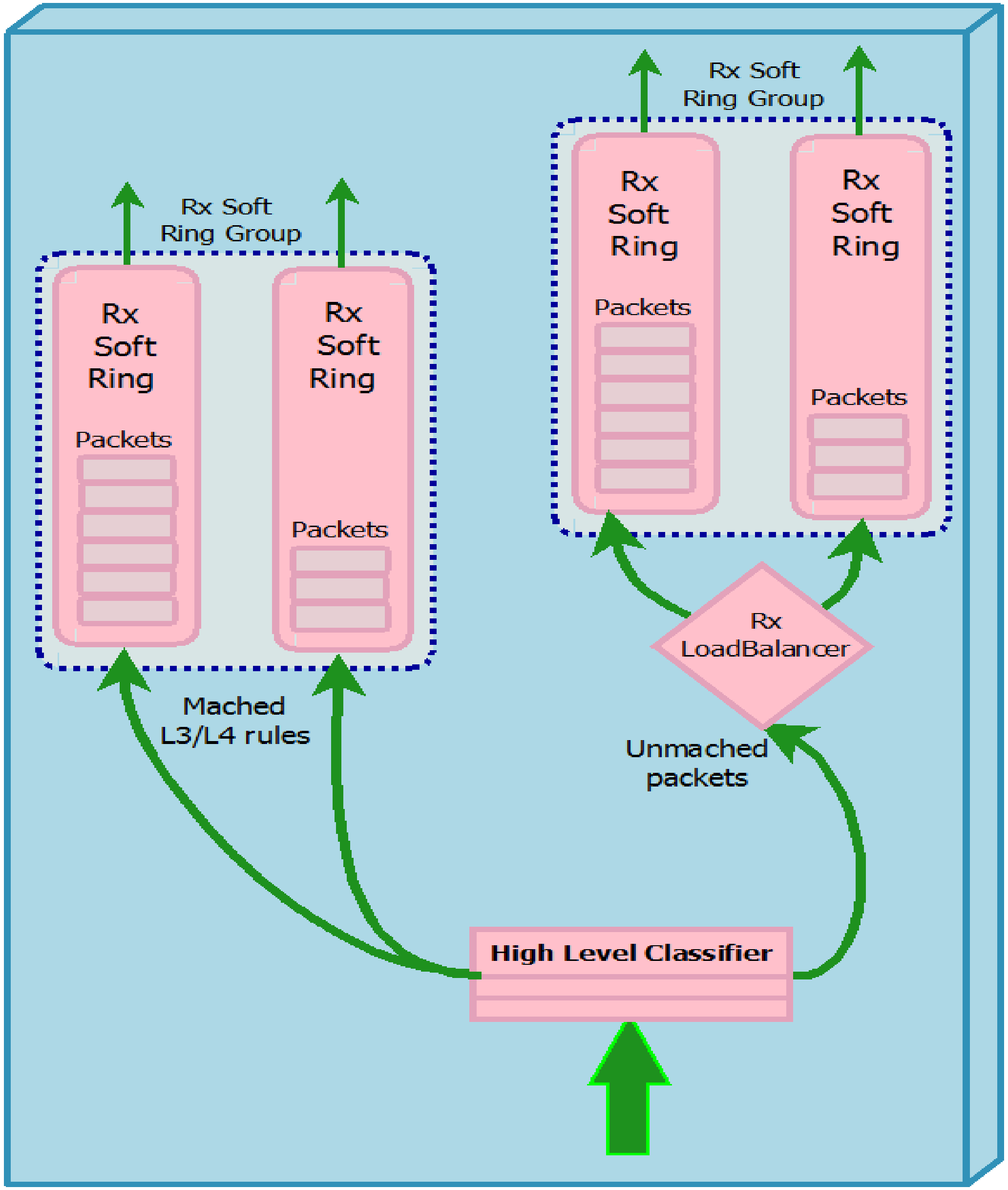
Rx-descriptor



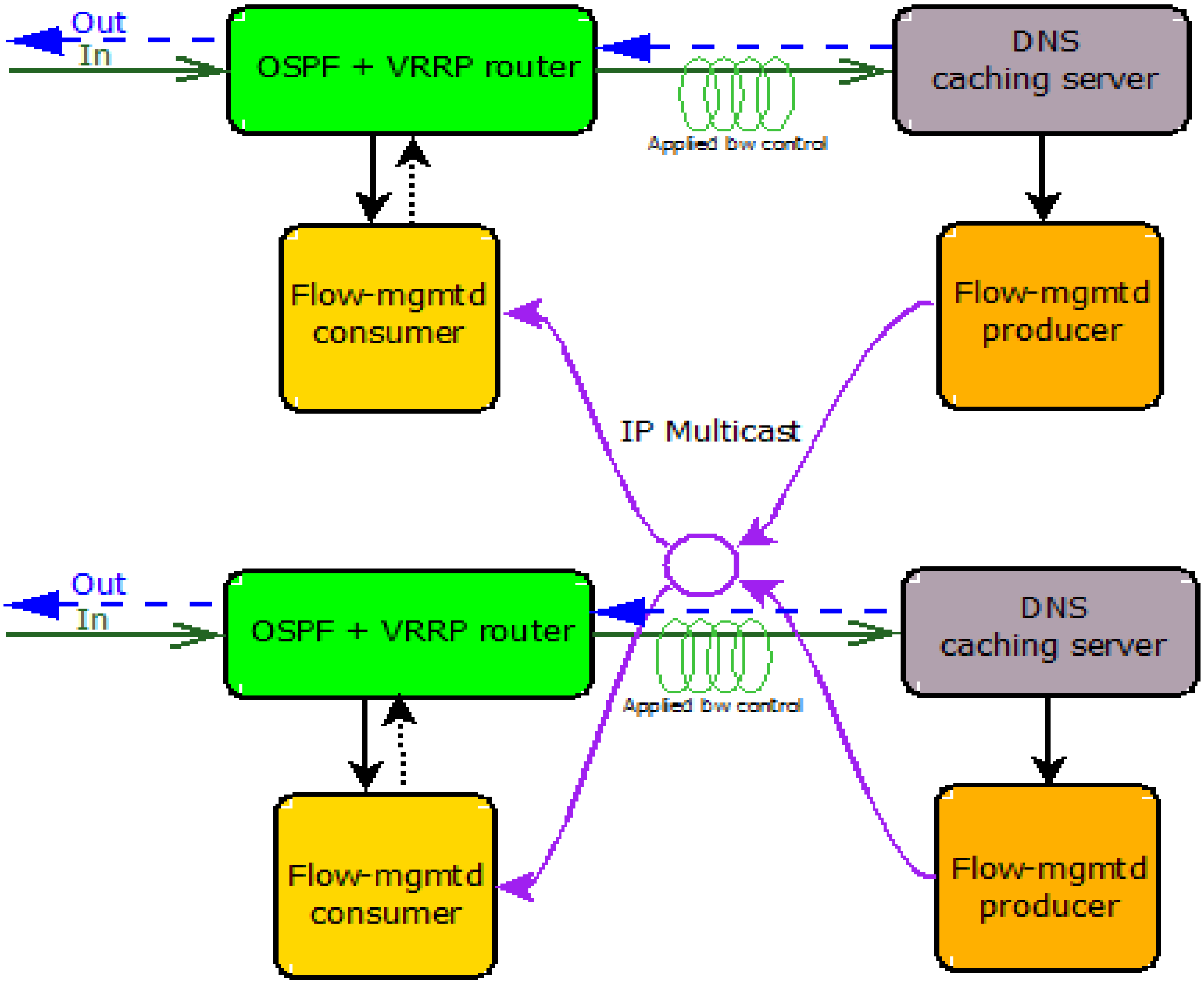
Clasificator MAC Level



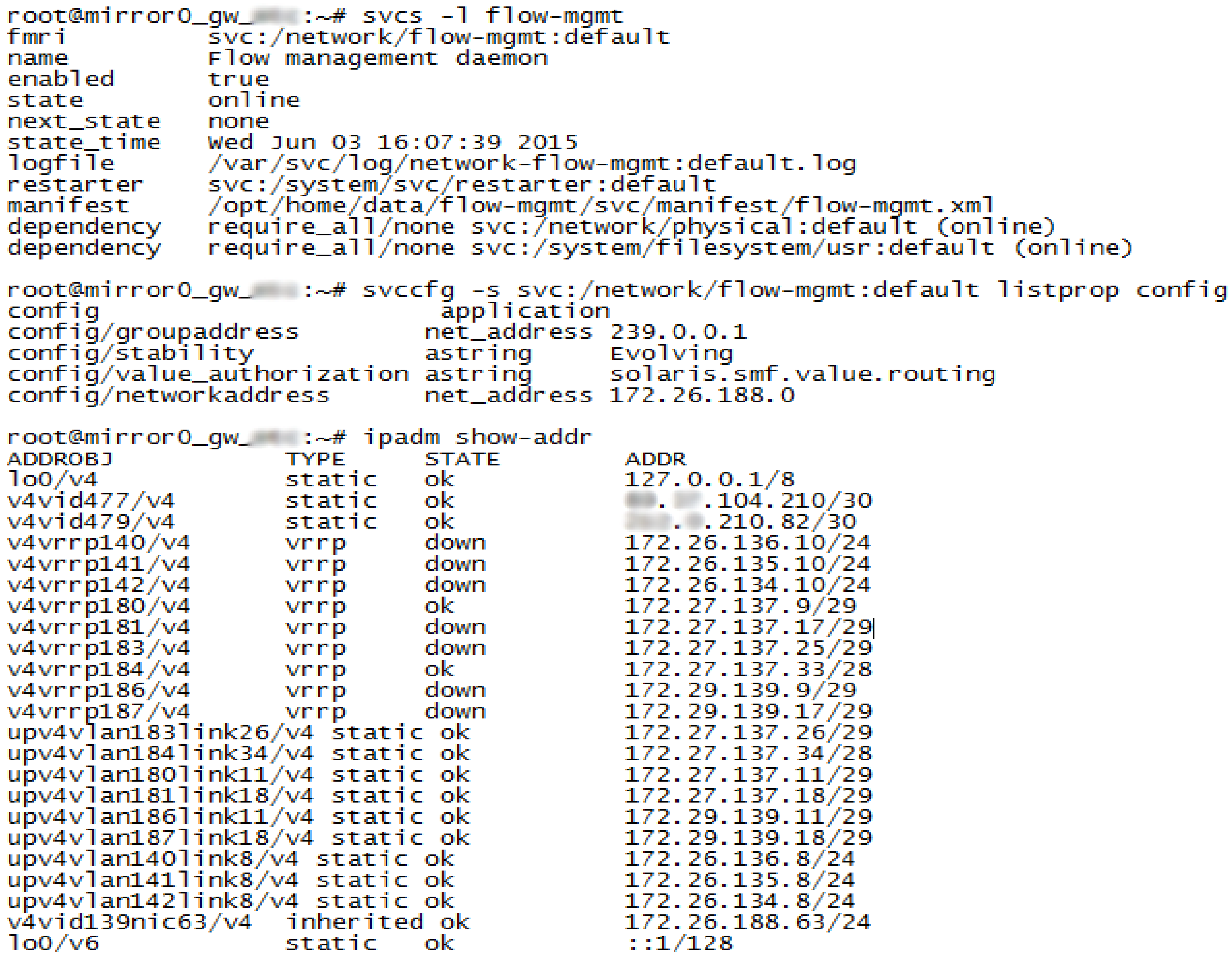
Clasificator High Level



Workflow aplicației SLA

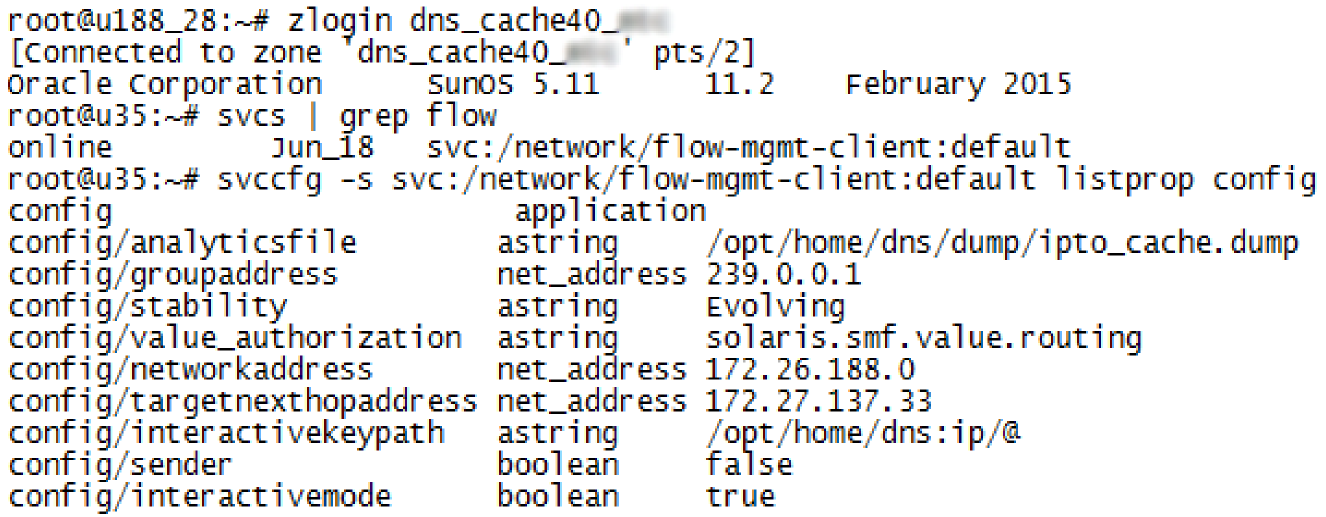


SMF configuratia aplicatiei din partea zonei Gateway



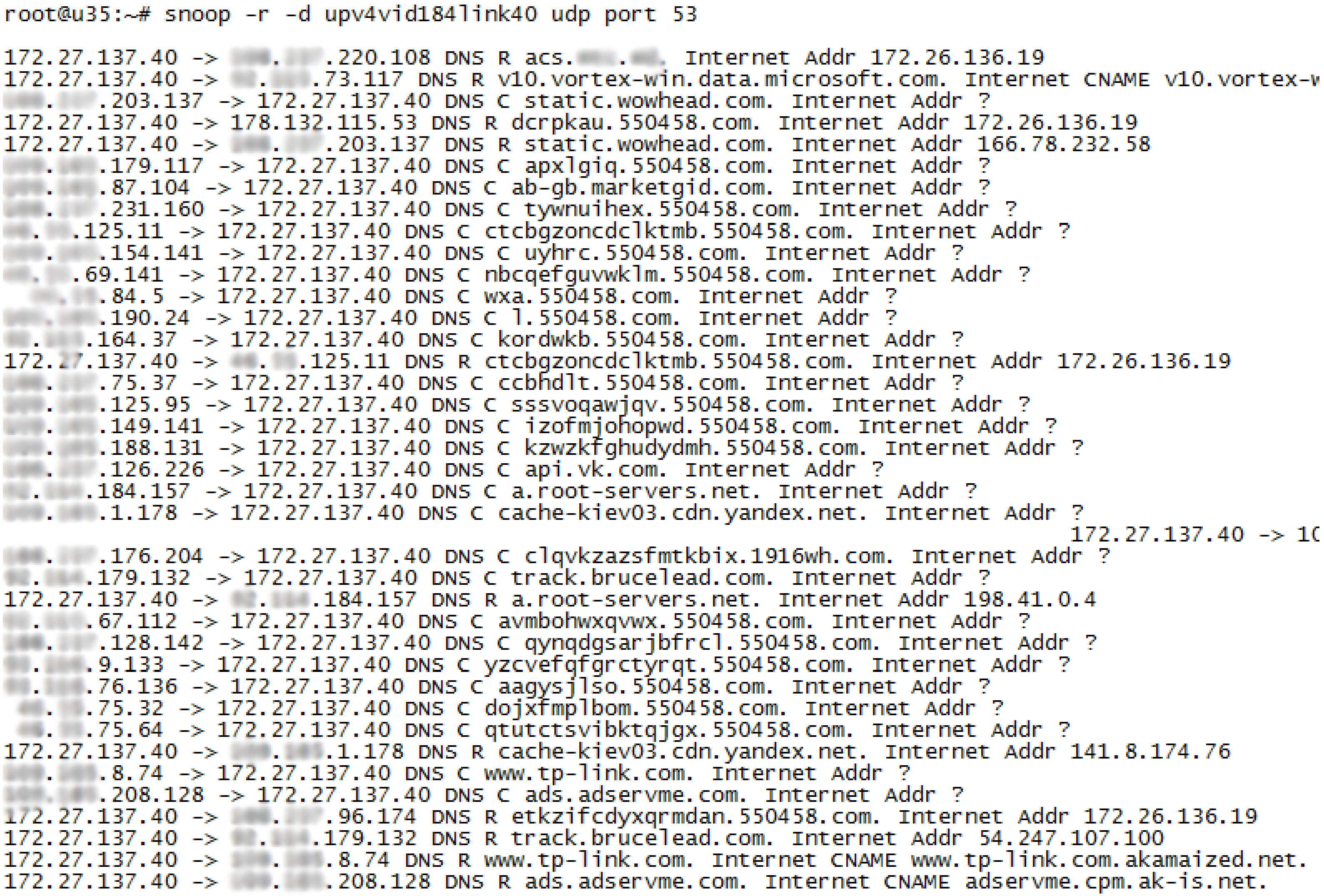
```
root@mirror0_gw_[redacted]:~# svcs -l flow-mgmt
fmri         svc:/network/flow-mgmt:default
name         Flow management daemon
enabled      true
state        online
next_state   none
state_time   Wed Jun 03 16:07:39 2015
logfile      /var/svc/log/network-flow-mgmt:default.log
restarter    svc:/system/svc/restarter:default
manifest     /opt/home/data/flow-mgmt/svc/manifest/flow-mgmt.xml
dependency   require_all/none svc:/network/physical:default (online)
dependency   require_all/none svc:/system/filesystem/usr:default (online)

root@mirror0_gw_[redacted]:~# svccfg -s svc:/network/flow-mgmt:default listprop config
config                      application
config/groupaddress        net_address 239.0.0.1
config/stability           astring     Evolving
config/value_authorization astring     solaris.smf.value.routing
config/networkaddress      net_address 172.26.188.0

root@mirror0_gw_[redacted]:~# ipadm show-addr
ADDROBJ           TYPE     STATE        ADDR
lo0/v4            static   ok           127.0.0.1/8
v4vid477/v4       static   ok           [redacted].[redacted].104.210/30
v4vid479/v4       static   ok           [redacted].[redacted].210.82/30
v4vrrp140/v4      vrrp     down         172.26.136.10/24
v4vrrp141/v4      vrrp     down         172.26.135.10/24
v4vrrp142/v4      vrrp     down         172.26.134.10/24
v4vrrp180/v4      vrrp     ok           172.27.137.9/29
v4vrrp181/v4      vrrp     down         172.27.137.17/29
v4vrrp183/v4      vrrp     down         172.27.137.25/29
v4vrrp184/v4      vrrp     ok           172.27.137.33/28
v4vrrp186/v4      vrrp     down         172.29.139.9/29
v4vrrp187/v4      vrrp     down         172.29.139.17/29
upv4vlan183link26/v4 static ok          172.27.137.26/29
upv4vlan184link34/v4 static ok          172.27.137.34/28
upv4vlan180link11/v4 static ok          172.27.137.11/29
upv4vlan181link18/v4 static ok          172.27.137.18/29
upv4vlan186link11/v4 static ok          172.29.139.11/29
upv4vlan187link18/v4 static ok          172.29.139.18/29
upv4vlan140link8/v4 static ok           172.26.136.8/24
upv4vlan141link8/v4 static ok           172.26.135.8/24
upv4vlan142link8/v4 static ok           172.26.134.8/24
v4vid139nic63/v4  inherited ok          172.26.188.63/24
lo0/v6            static   ok           ::1/128
```

SMF configuratia aplicatiei din partea zonei DNS



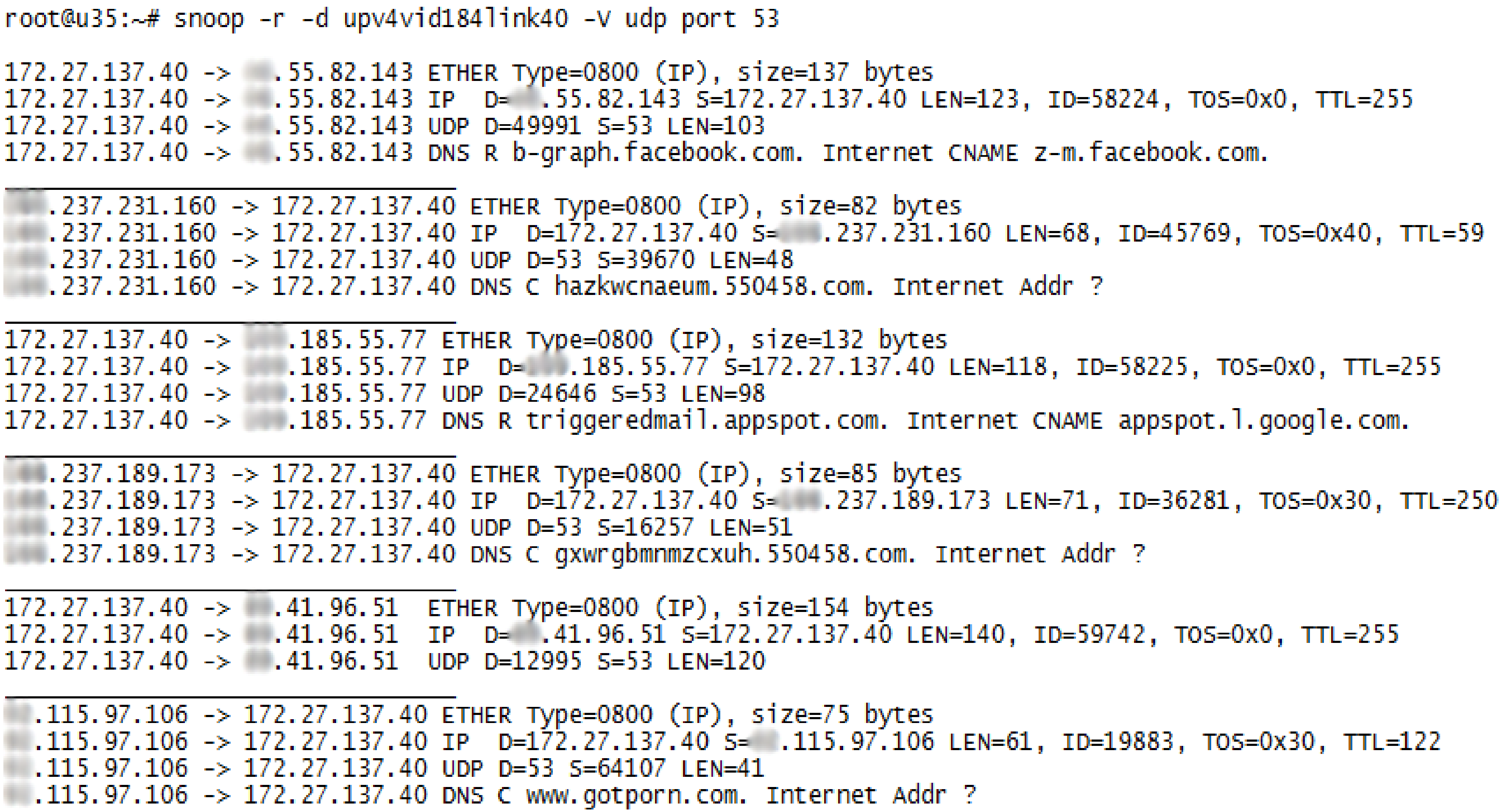
```
root@u188_28:~# zlogin dns_cache40_[redacted]
[Connected to zone 'dns_cache40_[redacted]' pts/2]
Oracle Corporation      SunOS 5.11      11.2    February 2015
root@u35:~# svcs | grep flow
online         Jun_18   svc:/network/flow-mgmt-client:default
root@u35:~# svccfg -s svc:/network/flow-mgmt-client:default listprop config
config                       application
config/analyticsfile        astring     /opt/home/dns/dump/ipto_cache.dump
config/groupaddress         net_address 239.0.0.1
config/stability            astring     Evolving
config/value_authorization  astring     solaris.smf.value.routing
config/networkaddress       net_address 172.26.188.0
config/targetnexthopaddress net_address 172.27.137.33
config/interactivekeypath   astring     /opt/home/dns:ip/@
config/sender               boolean     false
config/interactivemode      boolean     true
```

Captura traficului DNS în timpul desfăşurării atacului DoS



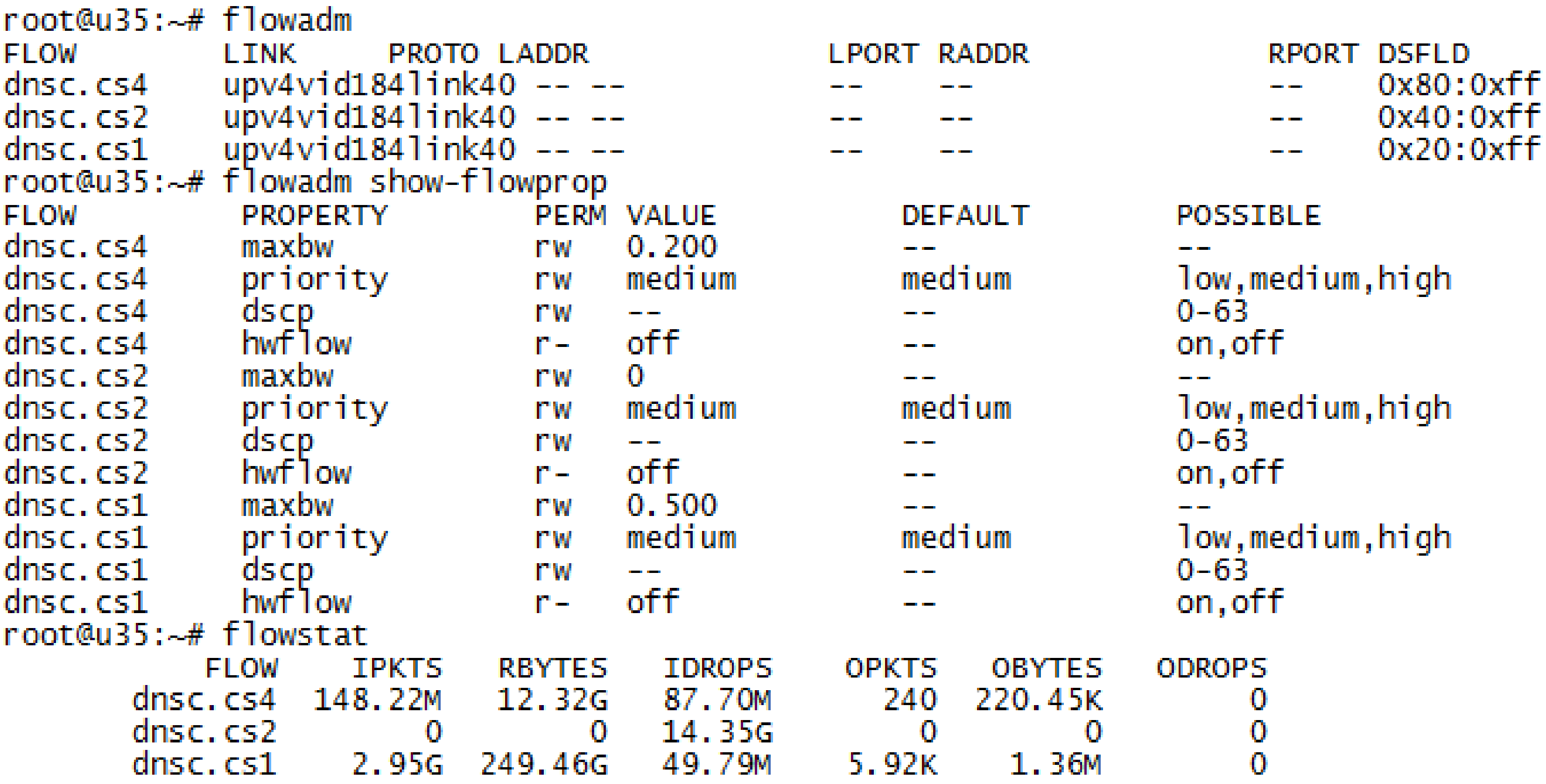
```
root@u35:~# snoop -r -d upv4vid184link40 udp port 53

172.27.137.40 -> [redacted].[redacted].220.108 DNS R acs.[redacted].[redacted]. Internet Addr 172.26.136.19
172.27.137.40 -> [redacted].[redacted].73.117 DNS R v10.vortex-win.data.microsoft.com. Internet CNAME v10.vortex-w
[redacted].[redacted].203.137 -> 172.27.137.40 DNS C static.wowhead.com. Internet Addr ?
172.27.137.40 -> 178.132.115.53 DNS R dcrpkau.550458.com. Internet Addr 172.26.136.19
172.27.137.40 -> [redacted].[redacted].203.137 DNS R static.wowhead.com. Internet Addr 166.78.232.58
[redacted].[redacted].179.117 -> 172.27.137.40 DNS C apxlgiq.550458.com. Internet Addr ?
[redacted].[redacted].87.104 -> 172.27.137.40 DNS C ab-gb.marketgid.com. Internet Addr ?
[redacted].[redacted].231.160 -> 172.27.137.40 DNS C tywnuihex.550458.com. Internet Addr ?
[redacted].[redacted].125.11 -> 172.27.137.40 DNS C ctcbgzoncdclktmb.550458.com. Internet Addr ?
[redacted].[redacted].154.141 -> 172.27.137.40 DNS C uyhrc.550458.com. Internet Addr ?
[redacted].[redacted].69.141 -> 172.27.137.40 DNS C nbcqefguvwklm.550458.com. Internet Addr ?
  [redacted].[redacted].84.5 -> 172.27.137.40 DNS C wxa.550458.com. Internet Addr ?
[redacted].[redacted].190.24 -> 172.27.137.40 DNS C l.550458.com. Internet Addr ?
[redacted].[redacted].164.37 -> 172.27.137.40 DNS C kordwkb.550458.com. Internet Addr ?
172.27.137.40 -> [redacted].[redacted].125.11 DNS R ctcbgzoncdclktmb.550458.com. Internet Addr 172.26.136.19
[redacted].[redacted].75.37 -> 172.27.137.40 DNS C ccbhdlt.550458.com. Internet Addr ?
[redacted].[redacted].125.95 -> 172.27.137.40 DNS C sssvoqawjqv.550458.com. Internet Addr ?
[redacted].[redacted].149.141 -> 172.27.137.40 DNS C izofmjohopwd.550458.com. Internet Addr ?
[redacted].[redacted].188.131 -> 172.27.137.40 DNS C kzwzkfghudydmh.550458.com. Internet Addr ?
[redacted].[redacted].126.226 -> 172.27.137.40 DNS C api.vk.com. Internet Addr ?
[redacted].[redacted].184.157 -> 172.27.137.40 DNS C a.root-servers.net. Internet Addr ?
[redacted].[redacted].1.178 -> 172.27.137.40 DNS C cache-kiev03.cdn.yandex.net. Internet Addr ?
                                                                                  172.27.137.40 -> 1(
[redacted].[redacted].176.204 -> 172.27.137.40 DNS C clqvkzazsfmtkbix.1916wh.com. Internet Addr ?
[redacted].[redacted].179.132 -> 172.27.137.40 DNS C track.brucelead.com. Internet Addr ?
172.27.137.40 -> [redacted].[redacted].184.157 DNS R a.root-servers.net. Internet Addr 198.41.0.4
[redacted].[redacted].67.112 -> 172.27.137.40 DNS C avmbohwxqvwx.550458.com. Internet Addr ?
[redacted].[redacted].128.142 -> 172.27.137.40 DNS C qynqdgsarjbfrcl.550458.com. Internet Addr ?
[redacted].[redacted].9.133 -> 172.27.137.40 DNS C yzcvefgfgrctyrqt.550458.com. Internet Addr ?
[redacted].[redacted].76.136 -> 172.27.137.40 DNS C aagysjlso.550458.com. Internet Addr ?
 [redacted].[redacted].75.32 -> 172.27.137.40 DNS C dojxfmplbom.550458.com. Internet Addr ?
 [redacted].[redacted].75.64 -> 172.27.137.40 DNS C qtutctsvibktqjgx.550458.com. Internet Addr ?
172.27.137.40 -> [redacted].[redacted].1.178 DNS R cache-kiev03.cdn.yandex.net. Internet Addr 141.8.174.76
[redacted].[redacted].8.74 -> 172.27.137.40 DNS C www.tp-link.com. Internet Addr ?
[redacted].[redacted].208.128 -> 172.27.137.40 DNS C ads.adservme.com. Internet Addr ?
172.27.137.40 -> [redacted].[redacted].96.174 DNS R etkzifcdyxqrmdan.550458.com. Internet Addr 172.26.136.19
172.27.137.40 -> [redacted].[redacted].179.132 DNS R track.brucelead.com. Internet Addr 54.247.107.100
172.27.137.40 -> [redacted].[redacted].8.74 DNS R www.tp-link.com. Internet CNAME www.tp-link.com.akamaized.net.
172.27.137.40 -> [redacted].[redacted].208.128 DNS R ads.adservme.com. Internet CNAME adservme.cpm.ak-is.net.
```

Captura traficul DNS parțial marcat DSCP



```
root@u35:~# snoop -r -d upv4vid184link40 -V udp port 53

172.27.137.40 -> [redacted].55.82.143 ETHER Type=0800 (IP), size=137 bytes
172.27.137.40 -> [redacted].55.82.143 IP  D=[redacted].55.82.143 S=172.27.137.40 LEN=123, ID=58224, TOS=0x0, TTL=255
172.27.137.40 -> [redacted].55.82.143 UDP D=49991 S=53 LEN=103
172.27.137.40 -> [redacted].55.82.143 DNS R b-graph.facebook.com. Internet CNAME z-m.facebook.com.
________________________________
[redacted].237.231.160 -> 172.27.137.40 ETHER Type=0800 (IP), size=82 bytes
[redacted].237.231.160 -> 172.27.137.40 IP  D=172.27.137.40 S=[redacted].237.231.160 LEN=68, ID=45769, TOS=0x40, TTL=59
[redacted].237.231.160 -> 172.27.137.40 UDP D=53 S=39670 LEN=48
[redacted].237.231.160 -> 172.27.137.40 DNS C hazkwcnaeum.550458.com. Internet Addr ?
________________________________
172.27.137.40 -> [redacted].185.55.77 ETHER Type=0800 (IP), size=132 bytes
172.27.137.40 -> [redacted].185.55.77 IP  D=[redacted].185.55.77 S=172.27.137.40 LEN=118, ID=58225, TOS=0x0, TTL=255
172.27.137.40 -> [redacted].185.55.77 UDP D=24646 S=53 LEN=98
172.27.137.40 -> [redacted].185.55.77 DNS R triggeredmail.appspot.com. Internet CNAME appspot.l.google.com.
________________________________
[redacted].237.189.173 -> 172.27.137.40 ETHER Type=0800 (IP), size=85 bytes
[redacted].237.189.173 -> 172.27.137.40 IP  D=172.27.137.40 S=[redacted].237.189.173 LEN=71, ID=36281, TOS=0x30, TTL=250
[redacted].237.189.173 -> 172.27.137.40 UDP D=53 S=16257 LEN=51
[redacted].237.189.173 -> 172.27.137.40 DNS C gxwrgbmnmzcxuh.550458.com. Internet Addr ?
________________________________
172.27.137.40 -> [redacted].41.96.51  ETHER Type=0800 (IP), size=154 bytes
172.27.137.40 -> [redacted].41.96.51  IP  D=[redacted].41.96.51 S=172.27.137.40 LEN=140, ID=59742, TOS=0x0, TTL=255
172.27.137.40 -> [redacted].41.96.51  UDP D=12995 S=53 LEN=120
________________________________
[redacted].115.97.106 -> 172.27.137.40 ETHER Type=0800 (IP), size=75 bytes
[redacted].115.97.106 -> 172.27.137.40 IP  D=172.27.137.40 S=[redacted].115.97.106 LEN=61, ID=19883, TOS=0x30, TTL=122
[redacted].115.97.106 -> 172.27.137.40 UDP D=53 S=64107 LEN=41
[redacted].115.97.106 -> 172.27.137.40 DNS C www.gotporn.com. Internet Addr ?
```

Statistica flow-urilor din partea serverului DNS (coloana IDROPS)



```
root@u35:~# flowadm
FLOW        LINK        PROTO LADDR             LPORT RADDR             RPORT DSFLD
dnsc.cs4    upv4vid184link40 -- --              --    --                --    0x80:0xff
dnsc.cs2    upv4vid184link40 -- --              --    --                --    0x40:0xff
dnsc.cs1    upv4vid184link40 -- --              --    --                --    0x20:0xff
root@u35:~# flowadm show-flowprop
FLOW         PROPERTY        PERM VALUE          DEFAULT        POSSIBLE
dnsc.cs4     maxbw           rw   0.200          --             --
dnsc.cs4     priority        rw   medium         medium         low,medium,high
dnsc.cs4     dscp            rw   --             --             0-63
dnsc.cs4     hwflow          r-   off            --             on,off
dnsc.cs2     maxbw           rw   0              --             --
dnsc.cs2     priority        rw   medium         medium         low,medium,high
dnsc.cs2     dscp            rw   --             --             0-63
dnsc.cs2     hwflow          r-   off            --             on,off
dnsc.cs1     maxbw           rw   0.500          --             --
dnsc.cs1     priority        rw   medium         medium         low,medium,high
dnsc.cs1     dscp            rw   --             --             0-63
dnsc.cs1     hwflow          r-   off            --             on,off
root@u35:~# flowstat
           FLOW    IPKTS   RBYTES   IDROPS    OPKTS   OBYTES   ODROPS
       dnsc.cs4  148.22M   12.32G   87.70M      240  220.45K        0
       dnsc.cs2        0        0   14.35G        0        0        0
       dnsc.cs1    2.95G  249.46G   49.79M    5.92K    1.36M        0
```

Rata de rejectare pentru trei flow-uri de agregare:

59%, 100%, 1%

Dificultăţi utilizând libdladm API

- Lipsa documentaţiei API
- Memory leak-uri (Soluționat: Oracle Solaris 11.3.3.6.0)
- Constrângere în ioctl DLDIOC_WALKFLOW (există workaround)

Output MDB pentru aplicație:

```
>::findleaks -dvf
findleaks: elapsed CPU time => 0.0 seconds
findleaks: elapsed wall time => 0.0 seconds
findleaks:
CACHE LEAKED BUFCTL CALLER
086f5010 1 0879dc40 libdladm.so.1`do_check_dscp+0x3c
086f5010 1 0879dbc8 libdladm.so.1`do_check_maxbw+0x34
------------------------------------------------------------------------

Total 2 buffers, 32 bytes

umem_alloc_16 leak: 1 buffer, 16 bytes
ADDR BUFADDR TIMESTAMP THREAD
CACHE LASTLOG CONTENTS
879dc40 8798fa0 1e2f61f9b5d0b6 1
86f5010 0 0
libumem.so.1`umem_cache_alloc_debug+0x157
libumem.so.1`umem_cache_alloc+0x19d
libumem.so.1`umem_alloc+0x76
libumem.so.1`malloc+0x2d
libdladm.so.1`do_check_dscp+0x3c
libdladm.so.1`i_dladm_flow_proplist_extract_one+0x198
libdladm.so.1`dladm_flow_proplist_extract+0x37
libdladm.so.1`dladm_flow_add+0x83
do_add_flow+0x33a
main+0x118
_start+0x7d

umem_alloc_16 leak: 1 buffer, 16 bytes
ADDR BUFADDR TIMESTAMP THREAD
CACHE LASTLOG CONTENTS
879dbc8 8798fc0 1e2f61f9b50f33 1
86f5010 0 0
libumem.so.1`umem_cache_alloc_debug+0x157
libumem.so.1`umem_cache_alloc+0x19d
libumem.so.1`umem_alloc+0x76
libumem.so.1`malloc+0x2d
libdladm.so.1`do_check_maxbw+0x34
libdladm.so.1`i_dladm_flow_proplist_extract_one+0x198
libdladm.so.1`dladm_flow_proplist_extract+0x37
libdladm.so.1`dladm_flow_add+0x83
do_add_flow+0x33a
```

Output MDB pentru tool-ul *flowadm*:

```
#env LD_PRELOAD=/usr/lib/libumem.so.1 UMEM_DEBUG=default
/usr/bin/i86/mdb /usr/sbin/flowadm
>::load libumem
>::sysbp _exit
>:r add-flow -l vlink35 -a local_ip=10.10.7.7 -p dscp=38 test2
mdb: stop on entry to _exit
mdb: target stopped at:
0xec88cd88: nop
mdb: You've got symbols!
Loading modules: [ ld.so.1 libumem.so.1 libc.so.1 libuutil.so.1 ]
> ::findleaks
CACHE LEAKED BUFCTL CALLER
0852d290 1 0856ec40 libdladm.so.1`do_check_dscp+0x3c
0852d290 1 0856ebc8 libdladm.so.1`do_check_maxbw+0x34
---------------------------------------------------------------------
---
Total 2 buffers, 32 bytes
```

CR #	Description	Fixed in version	SR date	Resolution date
15606330	restriction on flow creation can be relaxed in some cases	Oracle Solaris 11.3.0.30.0	**16.01.15**	**23.02.15**
15806736	some flow hash tables scale poorly with a large number of flows	Oracle Solaris 11.3.0.30.0	**16.01.15**	**23.02.15**
17649247	inbound/outbound traffic only flows	Oracle Solaris 11.3.0.30.0	**16.01.15**	**23.02.15**
20981017	libdladm leaks memory while adding flows	Oracle Solaris 11.3.3.6.0	**23.04.15**	**02.06.15**

Lista CR-urilor deschise sau escaladate în MOS

Link-uri utile

https://docs.oracle.com/cd/E53394_01/html/E54847/ntwkg.html#SOLWNgpqhs

https://docs.oracle.com/cd/E53394_01/html/E54764/flowadm-1m.html

https://blogs.oracle.com/yenduri/entry/new_flowadm_features_in_s11

https://tools.ietf.org/html/rfc2474

ORACLE - Writing Device Drivers

http://docs.oracle.com/cd/E23824_01/html/819-3196/gkbnv.html#gld3-datapaths

Interrupt handlers in ORACLE Solaris

http://www.oracle.com/technetwork/server-storage/solaris10/interrupt-handlers-141289.html

Mulţumesc pentru atenţie!