

---

# ART

## Network in RAC

---

By

Riyaj Shamsudeen



©OraInternals Riyaj Shamsudeen



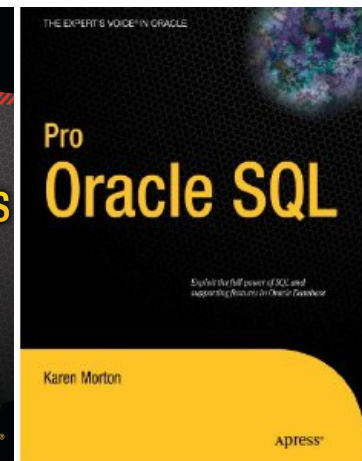
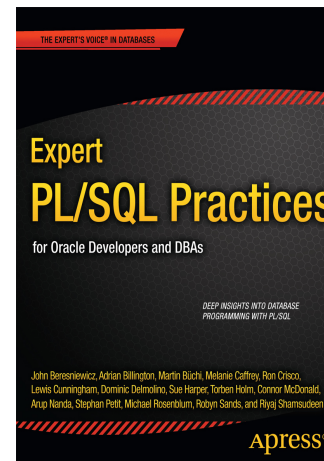
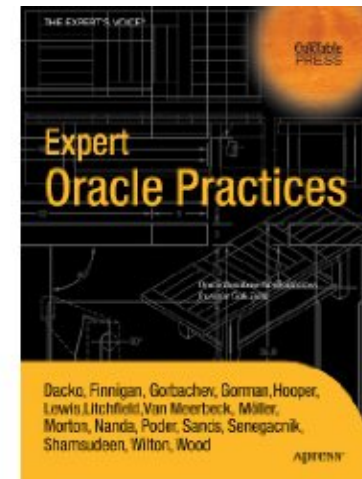
**Thank you to our Sponsor!**



# Who am I?



- 18 years using Oracle products/DBA
- OakTable member
- Oracle ACE
- Certified DBA versions 7.0,7.3,8,8i,9i &10g
- Specializes in RAC, performance tuning, Internals and E-business suite
- Chief DBA with OraInternals
- Email: [rshamsud@orainternals.com](mailto:rshamsud@orainternals.com)
- Blog : [orainternals.wordpress.com](http://orainternals.wordpress.com)
- URL: [www.orainternals.com](http://www.orainternals.com)



---

## Importance of network

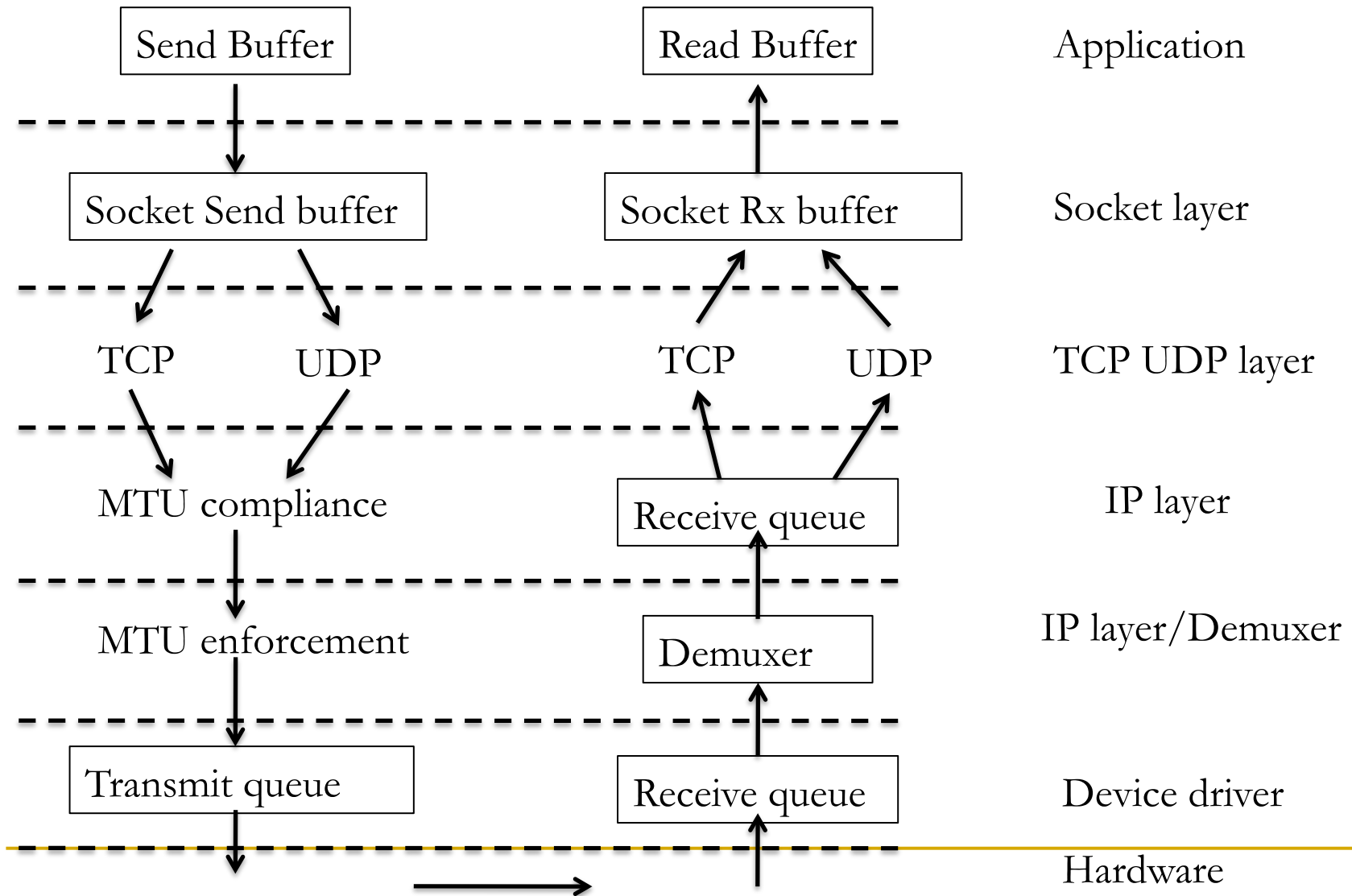
- In a typical organization, there are few types of networks employed:
  - Public Network – for public access to database and applications
  - Private Network – for cluster interconnect
  - Storage Network – Server to SAN access
  - Backup Network – For backup data traffic
- Generally, public network uses TCP/IP protocol. Private network uses one of UDP/LLT/RDS protocols in UNIX platform.. Storage networks generally uses TCP/IP protocol.
- Windows platform uses TCP/IP for private interconnect though.

---

## Network architecture

- Understand network architecture in your environment.
- A scalable network infrastructure is essential for application scalability.
- If you use parallelism extensively, use 10GB and Aggregated interfaces to keep interconnect traffic streaming.
- Latency (different from bandwidth) can cause performance issues to the application.

# Network layers



---

## What is UDP?

- UDP protocol can be employed for private interconnect traffic (Cache Fusion traffic.)
- UDP stands for Unreliable Datagram Protocol. No, that doesn't mean to scare you.
- UDP is in fact a layer over IP layer, so technically, it should be called as UDP/IP, similar to well known TCP/IP.
- Unreliable != data loss.

---

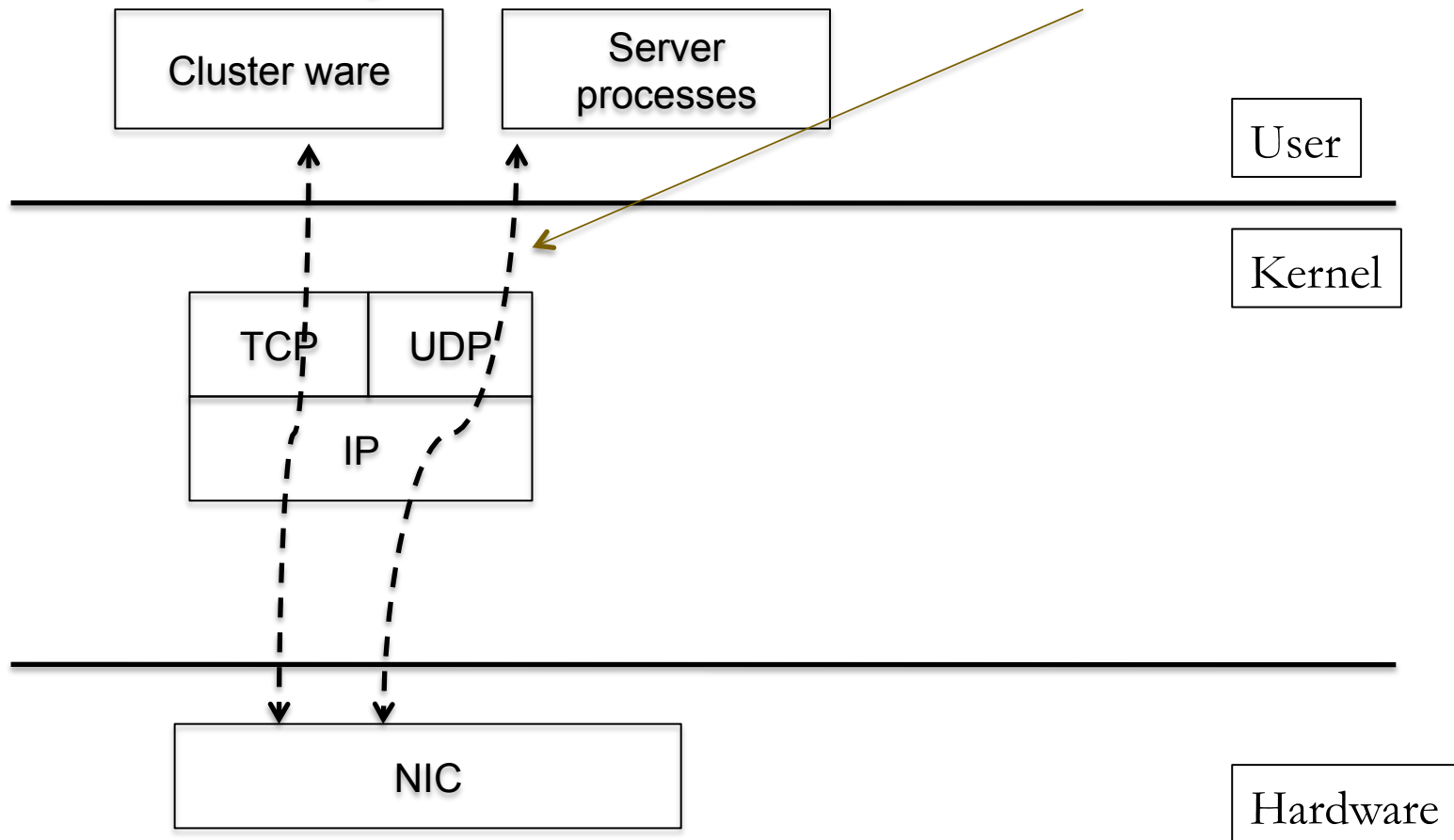
## UDP is different from TCP

- In TCP/IP, for every packet sent, an acknowledgement is received from the TCP layer in the receiving side.
- Packet must be acknowledged, within a timeout window. If not, packet is retransmitted.
- On the contrary, UDP is a send-and-forget protocol.
- Once the packet is sent, the packet send is considered complete. It is up to the application to handle error conditions.
- In RAC world though, RAC background processes sends back an acknowledgement packet, either as a grant/a buffer/a message.

# UDP vs TCP

Clusterware uses TCP for network heartbeat traffic between the nodes.

Cache fusion traffic can use UDP



---

## System calls

- Socket system calls copy network buffers from user space to kernel space.
- These system calls consume CPU in Kernel mode, call downstream system calls, and transfer packets to the interface.
- So, high cache fusion traffic can lead to higher kernel mode CPU usage.
- Buffers are copied from User space to Kernel space resulting in a buffer copy operation. In the receiving side, buffers are copied from Kernel to User space: Resulting in a double-copy, double buffer operations.

---

## Cluster\_interconnects and OCR

- Both DB and ASM queries OCR to get IP address for the cluster\_interconnect.
- This IP address will be used by both ASM and Database.
- Prior to 11gR2, clusterware used private node name for Heart beat traffic.
- From 11gR2 onwards, cluster\_interconnects parameter in OCR is queried by clusterware for node heartbeat.
- It is very important to specify correct IP address so that clusterware can detect the failures quickly.

---

## oifcfg

- Oifcfg command can be used to query the cluster\_interconnect IP address.

```
oifcfg getif  
agg1 10.188.244.0 global public  
agg3 172.29.1.0 global cluster_interconnect  
agg4 172.29.1.0 global cluster_interconnect
```

- In the above example, clusterware will choose an IP in the 172.29.1.X range and use that for heart beat.
- If you don't specify cluster\_interconnect parameter explicitly, then database/ASM also will choose an IP address from OCR.

## Cache Fusion IP

- `Gv$cluster_interconnects` view shows current interconnect details.

INST_ID	NAME	IP_ADDRESS	IS_	SOURCE
1	e1000g1	1.3.1.170	NO	cluster_interconnects parameter
2	e1000g1	1.3.1.180	NO	cluster_interconnects parameter

- You can also get the same from `cluster_interconnects` parameter value:

```
Show parameter cluster_interconnects
```

NAME	TYPE	VALUE
cluster_interconnects	string	1.3.1.170

## Another way..

- Another way to verify the use of IP address and protocol used for cache fusion traffic is, to use oradebug ipc command.

```
oradebug setmypid
```

```
oradebug ipc
```

```
Information written to trace file.
```

- View the trace file generated and search for SSKXPT

```
SSKGXPT fffffd7ffccb44f8 flags 0x0 sockno 11 IP 1.3.1.170 UDP 33689 lerr 0
```

- If pfiles is supported in your platform, then pfiles of the dedicated server process will show the IP address too:

```
11: S_IFSOCK mode:0666 dev:298,0 ino:28470 uid:0 gid:0 size:0
```

```
O_RDWR|O_NONBLOCK FD_CLOEXEC
```

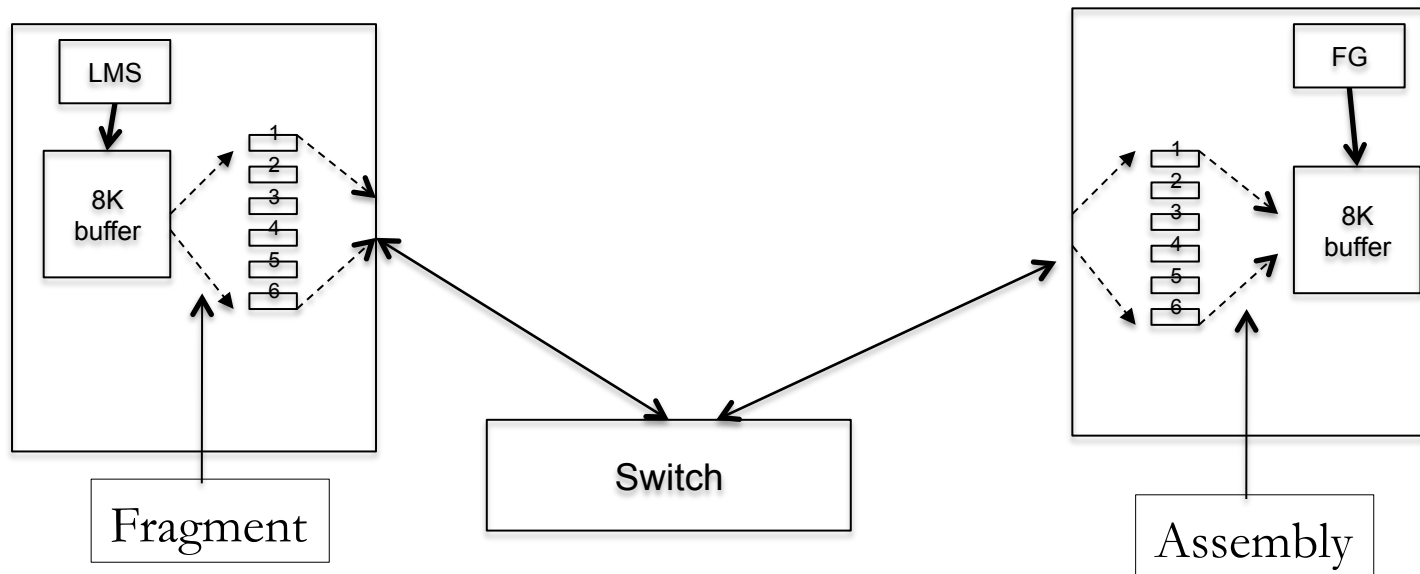
```
SOCK_DGRAM
```

```
SO_SNDBUF(57344), SO_RCVBUF(57344), IP_NEXTHOP(0.224.0.0)
```

```
sockname: AF_INET 1.3.1.170 port: 33689
```

## MTU

- MTU defines Maximum Transmission Unit of a packet. Essentially, limits the size of a packet, default is ~1500 bytes.
- For example, to transfer a buffer of 8K size, 6 packets must be transmitted. These packets can leave and arrive any order.



---

## Jumbo frames – Why?

- Assembly and Fragmentation of network packets are CPU intensive operations.
- Since this operation is performed in system calls, CPU is used in Kernel mode.
- Jumbo frames can be helpful if there is CPU starvation already.
- With Jumbo frame usage, MTU is increased beyond 8K, typically 9000 bytes.
- Just one packet is needed to transmit 8K buffer eliminating the need for fragment and assembly operations.

---

## Checking MTU

- ifconfig command, in UNIX platform will show the MTU size.

```
/sbin/ifconfig -a|more
```

```
...
```

```
e1000g1: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 3  
    inet 1.3.1.170 netmask ffffffff broadcast 1.3.1.255
```

```
...
```

- If jumbo frame is setup, then you would see the mtu adjusted.

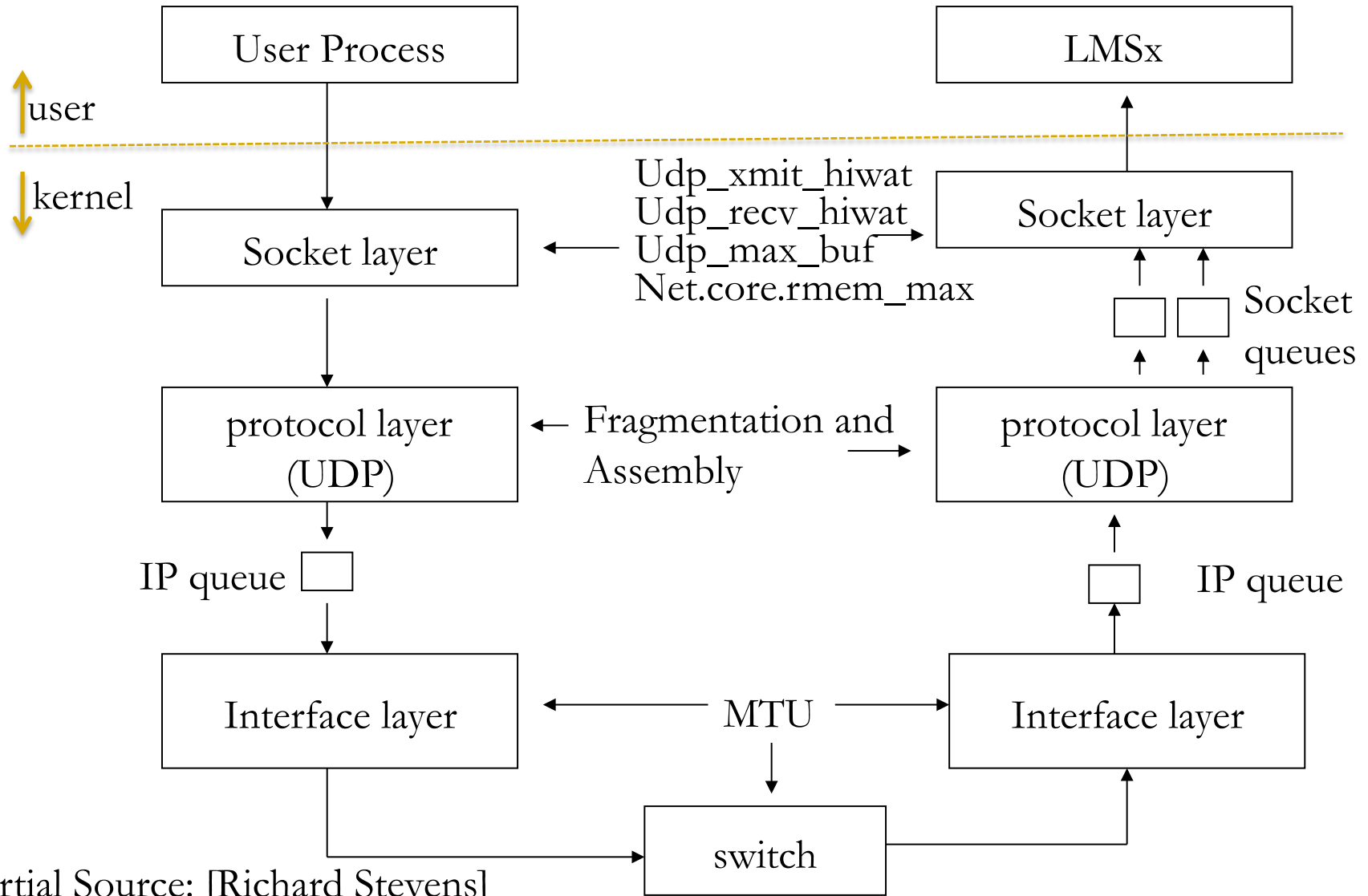
```
/sbin/ifconfig -a|more
```

```
...
```

```
e1000g1: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 9000 index 3  
    inet 1.3.1.170 netmask ffffffff broadcast 1.3.1.255
```

```
...
```

# Network layers



Partial Source: [Richard Stevens]

- There are varieties of tools available for network performance and measurement from the network side.
- As a DBA, you will have access to database server side tools such as netstat, ping, traceroute, ifconfig etc..
- Netstat utility provides performance counters at the interface and protocol level, in UNIX platforms.
- Ping provides ability to ping the packets to other nodes and test the performance.
- Traceroute prints details about the performance counter in the route.

## Ping

- Ping command is effective in testing the network. In fact, you might want to add the command to capture ping in OSWatcher.
- UDP packets can be sent and ping receives a message from other nodes.

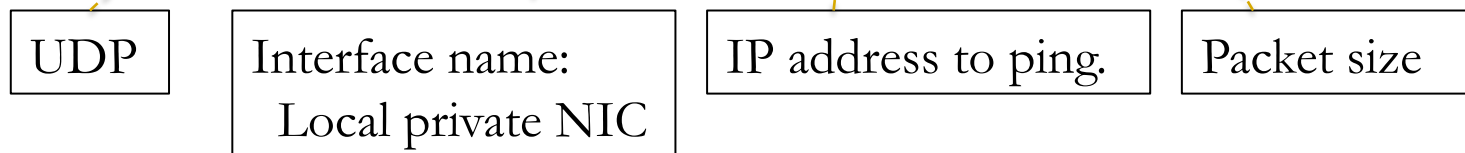
```
/usr/sbin/ping -s -U -i e1000g1 1.3.1.180 1492 6
```

```
PING 1.3.1.180: 1492 data bytes
```

```
92 bytes from solaris2_priv.solrac.net (1.3.1.180): udp_port=33434. time=0.559 ms
```

```
92 bytes from solaris2_priv.solrac.net (1.3.1.180): udp_port=33437. time=0.440 ms
```

```
...
```



Demo: demo\_ping\_udp.ksh

## Traceroute

- Traceroute can help to determine if the network path is configured properly.

```
/usr/sbin/traceroute -i e1000g1 -s 1.3.1.170 1.3.1.180
traceroute to 1.3.1.180 (1.3.1.180) from 1.3.1.170, 30 hops max, 40 byte packets
1 solaris2_priv.solrac.net (1.3.1.180) 0.636 ms * *
```



- Make sure that there are no hops between the source and target IP addresses.

Demo: demo\_traceroute.ksh

## Netstat – UDP

- Netstat –s provides statistics applicable to UDP/IP traffic.

```
...  
UDP      udpInDatagrams      =45934963002      udpInErrors      =      0  
         udpOutDatagrams =46871333207      udpOutErrors      =      0  
...
```

Total number of UDP datagrams transmitted/received from the server start.

UDP errors. This number should be very small. 0 is ideal.

## Netstat - IP

- IP RX/TX indicates an idea about workload and errors in IP stack.

ipInReceives	=2741104633	ipInHdrErrors	=	0
ipInAddrErrors	= 0	ipInCksumErrs	=	0
...				
ipInUnknownProtos	=216332	ipInDiscards	=	1108067397
ipInDelivers	=489353125	ipOutRequests	=	2679008014
ipOutDiscards	= 18535	ipOutNoRoutes	=	3

Total number of IP packet receives.

IP stack errors. These errors indicate hardware or network path issues. Zero is ideal.

Checksum errors can happen due to bugs in checksum offloading feature of network interface.

## Netstat - reassembly

IP reassembly without any errors.  
Should be less than Reassembly  
required parameter.

IP Reassembly required mostly  
due to lower MTU size.

- Reassembly statistics shows any issues with Reassembly and failures in packet reassembly. Lower MTU size means more reassembly.

...

ipReasmTimeout	=	60	ipReasmReqds	=	1569208584
ipReasmOKs	=	1569208453	ipReasmFails	=	131
ipReasmDuplicates	=	19	ipReasmPartDups	=	0

...

Duplicates should be smaller. Higher  
number is usually a bug or hardware  
issue.

IP reassembly failures. Should be a  
very small number since failure  
indicates that reassembly was not  
successful. High CPU usage can cause  
failures to.

## Netstat - interfaces

Netstat -i provides number of input/output packets and packet errors.  
Focus on the interfaces with errors if netstat -s showing any failures.

MTU size of the interface.

netstat -i

Name	Mtu	Net/Dest	Address	Ipkts	Ierrs	Opkts	Oerrs	Collis	Queue
lo0	8232	loopback	localhost	178586469	0	178586469	0	0	0
ce0	1500	ipmp2	ipmp2 3866864208	0	3263283863	0	0	0	
ce2	1500	priv1	priv1 1844093895	7	1447518648	0	0	0	
ce3	1500	priv2	priv2 877998656	1	3630601630	0	0	0	
ce9	1500	nas1	nas1 25389640	0	5581696	0	0	0	

Priv1 is busier than priv2. Load balancing issues?

Errors are minimal for interconnect network hardware.

# Thank you for attending!

If you like this presentation, you will love  
My upcoming seminar in Aug 2011 & Sep 2011.

<http://blog.tanelpoder.com/seminar/>

Contact info:

Email: rshamsud@gmail.com

Blog : orainternals.wordpress.com

URL : www.orainternals.com

