HP-UX Network Performance Tuning and Application Troubleshooting Tools

Pat Kilfoyle
Hewlett-Packard Co.

pat.kilfoyle@hp.com

HP WORLD 2003
Solutions and Technology Conference & Expo

Purpose

- Describe some common problems seen at the link, transport, and application layers.
- Describe the tools used to isolate network problems at the link, IP/UDP/TCP transport, and application levels and detail their useful features.
- Describe methodologies that most quickly narrow down the scope of a network application problem…which tools first and why.
- Review real-world case studies illustrating the tools and methodologies described

Agenda

- **Link layer**
  - Common problems
  - Detailed description of tools
- **IP/UDP/TCP layer**
  - Common problems
  - Detailed description of tools
- **NFS client and server subsystem**
  - Got a few days?…a brief summary then
- **Socket/Application layer issues and tools**
  - Common problems
  - Detailed description of tools
- **Case studies**
  - Peeling the onions

Link Layer

- **Common problems**
  - Ethernet 10/100/1000 BaseT/SX switched topologies
    - Link level connectivity and physical level errors.
    - Speed/duplex mismatches
    - Interconnect links
    - Switch buffering for speed step-downs
    - Max throughput expectations.
    - Card/CPU interrupt distribution
    - Trunking configurations
- **Tools**
    - Lanscan, landiag, linkloop, nettl, HW/SW analyzers, topology maps, stats from switches/routers and other interconnect equipment.

Link Layer (cont)

- **Link level connectivity checks & packet errors**
  - ***linkloop* command**
    - Basic local LAN (local subnet) connectivity tool
  - ***landiag/lanadmin* utility can be used to display per interface link stats.**
    - In general link stats for full duplex connections should be squeaky clean….no FCS, collisions, carrier sense errors.
  - ***lanadmin –x stats internal (drv) ppa#***

Link Layer (cont)

- **Speed & duplex mismatches**

  Symptoms run from no link level connectivity, FCS errors, collisions for half duplex modes, and packet loss in general.

  - ***lanscan –v***
    - Provides interface specifics:

      HW path

      MAC address

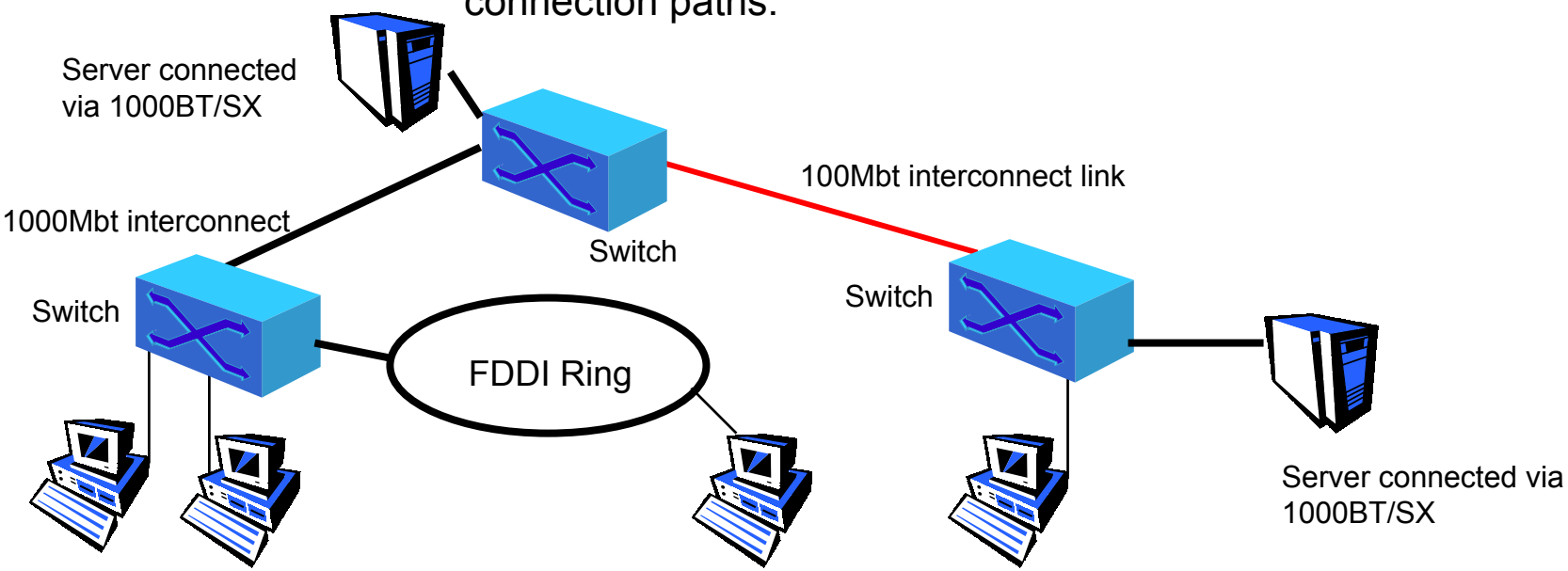
      Card instance number (ppa #)

      Driver name

      Interface name…i.e.. Lan2
  - ***lanadmin –x <ppa>***
    - Displays speed, duplex, and autonegotiation state
  - ***/etc/rc.config.d/<card config file>***
    - Config files for interface cards specifying speed/duplex to be set during bootup.

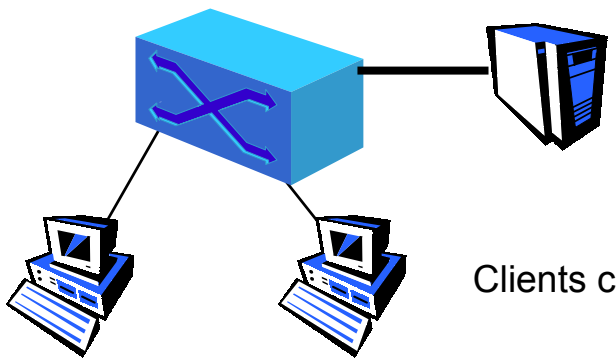
Link Layer (cont)

- **Interconnect links**
  - **Low bandwidth**
    Symptoms are low throughput and packet loss
  - **FDDI – ethernet bridging**
    IP fragmentation/MTU changes can cause performance issues
  - **Load balancing equipment**
    Application level load balancers and firewalls can complicate the connection paths.



Link Layer (cont)

- **Switch buffering for speed step downs**
  - Symptoms are packet loss and poor throughput at upper layers
  - In some high traffic scenarios large bursts of packet trains on the gig side will overrun the buffering of the switch. This is usually a sustained high packet rate with little or no upper level protocol flow control.
    - NFS PV3 over UDP with 32K read/write sizes…..this will put 22 packets in a burst on the wire for every read from the server. UDP being connectionless means there is no pacing or flow control, so frames are pumped out as fast as the card can drive the link.



Server connected via 1000BT/SX Gigabit

Clients connected via 100Mbt

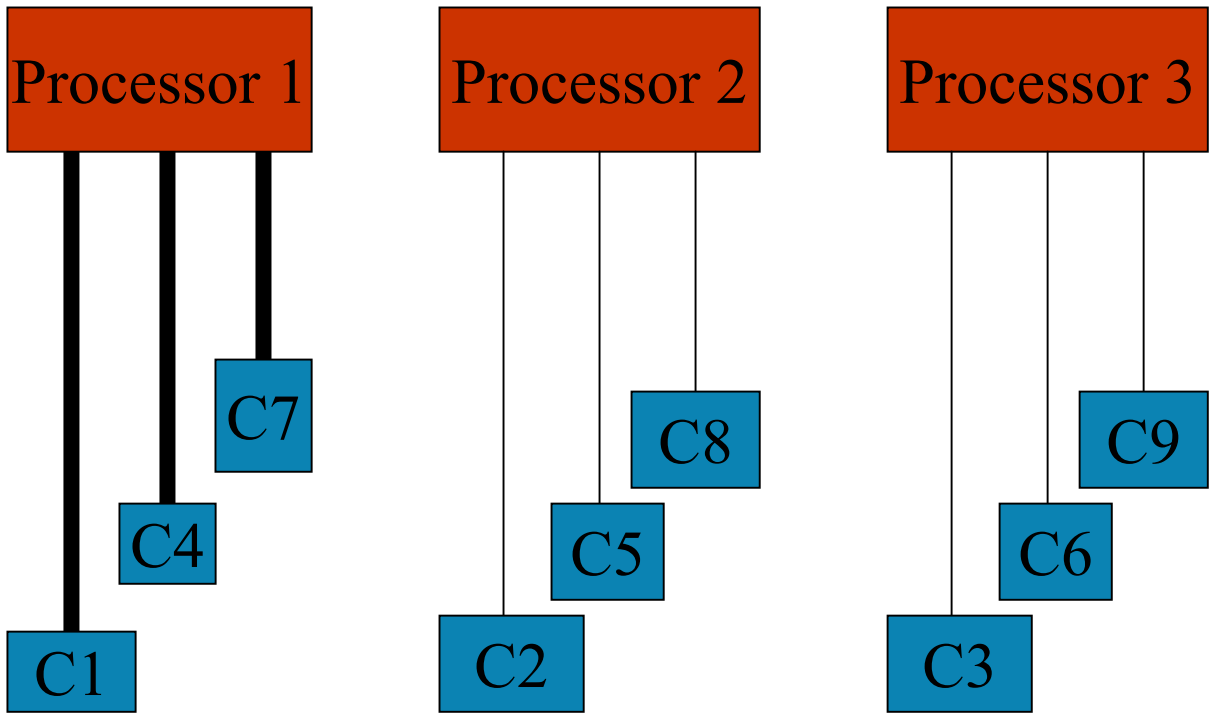
Link Layer (cont)

- **Max throughput expectations**
  - **#### BaseT is not a promise of #### Mbit/sec**
    System CPU speeds and card DMA rates both inbound and outbound are typically the limiting factor.
    Even if the card can do it, the application/transport driving the connection may be the limiting factor.
    Trunking (Auto Port Aggregation) has load balancing schemes that play a key role in trunk utilizations and throughput for any one TCP connection.
  - **The specific test used to measure throughput is also a critical factor. You need to understand exactly what the test tool is doing and what limitations the tool has.**

Link Layer (cont)

- **Card/CPU interrupt distribution -**
  - round robin interrupt allocation can result in heavy interrupt load on the same CPU
  - Starting with 11i, you can use "interrupt migration" functionality. This is delivered in a free package available on software.hp.com.

- top or vsar can be used to show % CPU interrupt load.



Link Layer (cont)

- **HP APA or Trunking**
  - Trunking or Auto-Port-aggregation configurations offer redundancy and higher bandwidth logical links.
  - *lanscan –v* can be used to see which individual links are in which aggregates. The the *lanadmin/landiag* interface can be used to review per interface stats.
  - *netstat –in* will show which IP's are assigned to which links…that includes APA trunks.
  - System startup config files */etc/rc.config/hpapa** control the trunk configurations.
  - Switch side configurations

Link Layer Tools - lanscan

- **Displays information about LAN interfaces installed that the system SW supports/recognizes**
- **Typical usage:**

  lanscan –v | *more* or simply *lanscan*
- **Key fields:**
  - HW IO path for card
  - HW MAC address
  - Instance # or ‘PPA #’
  - Netname…ie. lan1 etc.
  - Driver name for this card
  - APA port assignment

Link Layer Tools – landiag/lanadmin

- **Admin tool to manage LAN interfaces at the link layer**
  - Display and change the station address.
  - Display and change the 802.5 Source Routing options (RIF).
  - Display and change the maximum transmission unit (MTU).
  - Display and change the various card settings…eg. Speed/duplex
  - Clear the network statistics registers to zero.
  - Display the interface statistics.
  - Reset the interface card, thus executing its self-test.
- **Basic data gathering info utility**

Link Layer Tools - linkloop

- **The linkloop command uses IEEE 802.2 link-level test frames to check connectivity within a local area network (LAN).**
  - A good sanity check of basic link level connectivity when dealing with IP connectivity problems within the same subnet or vlan.
  - Uses DLPI to talk to the interface card…no Streams/transport stack involved.
  - Not a good throughput measurement tool since design is request-reply model.

Link Layer Tools - nettl

- **The nettl tracing and logging subsystem**
  - Kernel subsystems log at various levels of severity to /var/adm/nettl.LOGXXX *nettl –ss* to list them
  - At the link level it can be used to trace the packets at the driver level.
  - Significant driver events are also logged by default.
  - Traces to raw binary files which must be formatted using the *netfmt* command
  - Post filtering is done with *(-c filterfile)* option to netfmt
  - High speed links can overrun the trace buffer causing holes in the trace data.

Link Layer Tools – nettl (cont) – netfmt formatter

- **The nettl tracing and logging subsystem**
  - The netfmt formatter has a flexible filter file format
  - The nettl trace header record has useful info like Kernel threadID's of the process sending down the packet and timestamps of course, and this can be specified in the filter file. Inbound packets have a thread ID of –1 for ICS.
  - Multiple subsystems can be traced at the same time. The –e option specifies the entity/subsystem to trace and *'-e all'* is a valid option.
  - With light enough traffic and enough CPU speed, real-time formatting through a filter file is possible.

Link Layer Tools – nettl (cont) – sample output

- **The nettl tracing and logging subsystem**
  - netfmt formatted output in one line presentation helps searching a large trace file and formats quicker and the output is smaller.
  - Once the specific connection or time period is identified, the –N or nice formatting can be done using a filter file.
  - The ‘-n -1T’ option specifies one-line with Timesamps and name/address translation suppressed.
  - The ‘–n –lN’ option specifies ‘nice’ formatting with name/address translation suppressed.

Link Layer Tools – HW/SW analyzers

- **External analyzers come in two forms**
  - Hardware designed to listen/connect to the network
  - SW using std NIC's running in promiscuous mode
- **Objective data gathering.**
  - Helps resolve the "we sent it out" finger pointing
- **Can be difficult to insert in data path.**
  - Monitor ports on switches and Gig fiber links
- **$$$**
- **Varying trace file formats**
- **Expert modes**

Link Layer Tools – SW analyzers

Capture:2 (Summary)

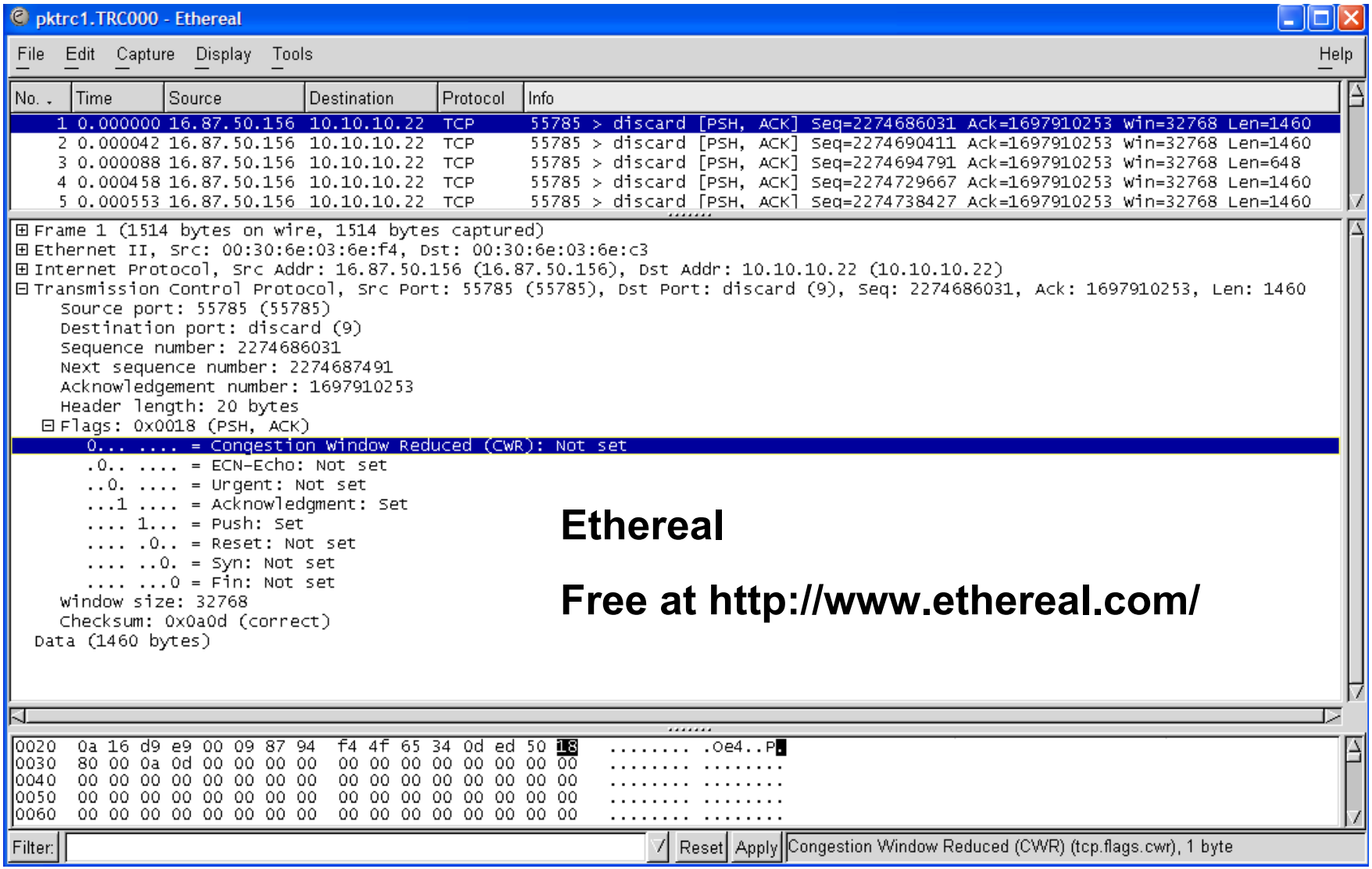
MAC Addr	Dst MAC Addr	Protocol	Description	Src Other Addr	Ds
NGE2410	*BROADCAST	NBIPX	Delete Name 1<00><00><00><00><00><00><00><00><00><00>É'3p+	M4LANGE2410	0.
41339E	*BROADCAST	ARP_RARP	ARP: Request, Target IP: 15.41.105.104		
8341901C	*BROADCAST	ARP_RARP	ARP: Request, Target IP: 15.24.50.121		
8341901C	*BROADCAST	ARP_RARP	ARP: Request, Target IP: 15.24.47.60		
0C31	006083419010	UDP	Src Port: Unknown, (1051); Dst Port: Unknown (4000); Length = 2	C2410C31	15
8341901C	C2410C31	UDP	Src Port: Unknown, (4000); Dst Port: Unknown (1051); Length = 2	15.40.184.55	C2
B014E0A2	*BROADCAST	SAP	General Svc Resp [0060B014E0A280... - HP Print Server]	1.0060B014E0A2	1.
DA22C270	*BROADCAST	ARP_RARP	ARP: Request, Target IP: 15.24.50.179		
NGE2410	*BROADCAST	SAP	General Svc Resp [M4LANGE2410!!!... - Unknown Service]	M4LANGE2410	1.
8341901C	C2410C31	TCP	.AP..., len: 178, seq:1904803299-1904803476, ack: 398672, w	15.56.32.48	C2
0C31	006083419010	TCP	.A...., len: 0, seq: 398672-398672, ack:1904803477, win:	C2410C31	15
0JXB	*BROADCAST	NETLOGON	SAM LOGON request from client	F2410JXB	15
0JXB	*BROADCAST	NETLOGON	SAM LOGON request from client	F2410JXB	15

```
+FRAME: Base frame properties
-ETHERNET: ETYPE = 0x0800 : Protocol = IP:  DOD Internet Protocol
  +ETHERNET: Destination address : 0050DA29F833
  +ETHERNET: Source address : 006083419010
    ETHERNET: Frame Length : 232 (0x00E8)
    ETHERNET: Ethernet Type : 0x0800 (IP:  DOD Internet Protocol)
    ETHERNET: Ethernet Data: Number of data bytes remaining = 218 (0x00DA)
+IP: ID = 0x7658; Proto = TCP; Len: 218
-TCP: .AP..., len:  178, seq:1904803299-1904803476, ack:    398672, win: 6432, src: 5222  dst: 1088
    TCP: Source Port = 0x1466
    TCP: Destination Port = 0x0440
    TCP: Sequence Number = 1904803299 (0x7188FDE3)
    TCP: Acknowledgement Number = 398672 (0x61550)
    TCP: Data Offset = 20 (0x14)
    TCP: Reserved = 0 (0x0000)
  +TCP: Flags = 0x18 : .AP...
    TCP: Window = 6432 (0x1920)
    TCP: Checksum = 0xBBF0
    TCP: Urgent Pointer = 0 (0x0)
```

Microsoft Netmon

```
00000000   00 50 DA 29 F8 33 00 60 83 41 90 1C 08 00 45 00   .P+)°3.`âAÉ...E.
00000010   00 DA 76 58 40 00 3A 06 5D 25 0F 38 20 30 0F 18   .+vX@.:.]%¤8 0¤.
00000020   2E 21 14 66 04 40 71 88 FD E3 00 06 15 50 50 18   .!¶f.@qê²p..§PP.
00000030   19 20 BB F0 00 00 3C 70 72 65 73 65 6E 63 65 20   . +=..<presence
00000040   66 72 6F 6D 3D 27 6B 61 79 5F 64 6F 77 64 61 6C   from='kay_dowdal
```

Link Layer Tools – SW analyzers



Ethereal

Free at http://www.ethereal.com/

IP/UDP/TCP Layer

- **Common IP layer problems**
  - Path MTU
  - Route through network
  - Multiple Interface, multiple IP addr confusion
  - Arp cache entries in switches and routers
  - ICMP redirect, unreachable, sourcequench
- **Tools**
  - *ifconfig*
  - *route*
  - *arp* – displays/modifies arp cache entries
  - *traceroute* - maps IP routes through network
  - *netstat –rn* – displays transport routing tables
  - *nettl* - Network tracing and logging subsystem
  - ndd - utility to get/set transport tunables
  - *ping* - Utility to send ICMP echo packets

IP/UDP/TCP Layer (cont)

- Path MTU issues
-
  - netstat –rn can be used to list routing table entries and MTU's
  - landiag/lanadmin can be used to display/set an interfaces current MTU setting.
  - Ping can be used to send ICMP packets of varying sizes
  - ndd tunables for path MTU management

IP/UDP/TCP Layer (cont)

- IP routing
  - Default gateway definitions
    - /etc/rc.config.d/netconf config file for IP
  - Dynamic routes
    - *netstat –rnv* command to see routing table
    - *ndd* tunables for managing IP routing

      ip_ire_hash - Displays all routing table entries, in the order searched when resolving an address

      ip_ire_status - Displays all routing table entries

      ip_ire_cleanup_interval - Timeout interval for purging routing entries

      ip_ire_flush_interval - Routing entries deleted after this interval

      ip_ire_gw_probe_interval - Probe interval for Dead Gateway Detection

      ip_ire_pathmtu_interval - Controls the probe interval for PMTU

      ip_ire_redirect_interval - Controls 'Redirect' routing table entries
  - *traceroute* tool to map out the IP

IP/UDP/TCP Layer (cont)

- **Multiple interfaces/IP addresses**
  - Application binding to proper interfaces
  - Multihomed hosts still only have one default GW.
    - Traffic might go out one interface and back another
  - Moving secondary IP's between interfaces and systems
    - Switches and routers need to remap the MAC/IP addresses.
  - Duplicate IP addresses…..a bad idea.
  - Cabled to wrong subnet
    - nettl tracing can help see subnet bcasts and ID what subnet you're really on.

IP/UDP/TCP Layer (cont)

- **ARP cache tables in switches/routers**
  - Relocatable IP addresses can confuse them
    - Only one ARP is sent when ifconfig is issued
  - Clearing them may be necessary
  - Temporary use of a dumb hub/repeater, or point to point cable to verify link level connectivity.

IP/UDP/TCP Layer (cont)

- **ICMP redirect, unreachable, sourcequench**
  - Network and host redirects
    - System routing tables will be updated and a 5 minute time started on the 'learned' or 'dynamic' route entries that result.
  - Network and host unreachables
    - Stderr messages will usually result.
      ENETUNREACH errno 229
      EHOSTUNREACH errno 242
  - Sourcequench
    - HPUX generates them by default when UDP or raw IP socket buffers overflow
    - Ndd can disable them
    - Typically nothing to worry about and can be useful in troubleshooting udp socket overflows. The ICMP message will contain the IP/UDP header of the packet that caused the overflow.

IP/UDP/TCP Layer (cont)

- **Common UDP related problems**
  - IP fragmentation of UDP datagram
  - No congestion/flow control at the UDP layer other than ICMP sourcequenches
  - Reserved and anonymous port range usage
- **Tools**
  - Nettl
  - Netstat
  - Ndd
  - lsof
  - Glance – ‘Thread List’ screen

IP/UDP/TCP Layer (cont)

- **IP fragmentation of UDP datagram larger than the interface MTU size.**
  - Performance concerns
    - IP fragmentation reassembly memory is fixed but can be tuned.
    - Timeout of IP fragments waiting for reassembly
  - Intermediate network equipment needs to support IP fragmentation if FDDI/Token Ring/Ethernet topologies are mixed.

IP/UDP/TCP Layer (cont)

- **No congestion/flow control at UDP other than ICMP sourcequench messages**
  - Most frequent UDP abuser is NFS PV3 with 32k read/write sizes. The 32k read/write bursts are comprised of 21+ 1500 byte packets in a burst and can contribute to network congestion if the server is at Gigabit speeds and clients at 100bt.
  - ICMP source quench messages are not typically acted upon.
    - Can be useful in finding which UDP port is overflowing and which process owns the port.
  - ndd tunables for default UDP socket buffer size on 11.11+

IP/UDP/TCP Layer (cont)

- **Reserved UDP port range usage**
  - Ports < 1024 are considered reserved for superuser
  - A weak implication of security.
  - There are only 1023 of them…..
  - What process owns them?
  - What are they being used for?
  - NFS/NIS/ONC heavy users
- **Anonymous UDP port range**
  - *ndd –h* | *grep anon* for tunable limits
  - 49152-65535 is default range
  - Low end may need to be dropped

IP/UDP/TCP Layer (cont)

- **Common TCP related problems**
  - TCP Connection setup
  - TCP Data transfer …Established state
  - TCP Connection teardown
- **Tools**
  - Nettl, netstat, ndd
  - Ttcp, netperf
  - ftp, rcp, telnet
  - Sample TCP socket code
  - Tusc, lsof

IP/UDP/TCP Layer (cont)

- **TCP Connection setup**
  - 3-way handshake

    Kernel will put conn in EST before listener accept()s. The listen backlog queue size determines how many can be waiting. System default is 20 .

    Take note of TCP options in SYN packets..MSS and window scaling options.
  - Connection timeouts

    ndd –h tcp_ip_abort_cinterval 75 sec on HPUX 11.X

    netstat –an | *grep SYN_SENT* is a clue

    ndd –h tcp_conn_grace_period
  - Connection rejected

    resets of failed connection attempts may tack on added text info to the reset packet

    netstat –sp tcp to see drops due to queue full or no listener

IP/UDP/TCP Layer (cont)

- **TCP Data transfer …Established state**
  - Retransmission timeout

    netstat –sp tcp | more
  - Fast Retransmission

    netstat –sp tcp | grep retrans

    netstat –sp tcp | grep rexmit
  - Slow Start Algorithm

    A kinder gentler transport
  - Keepalives

IP/UDP/TCP Layer (cont)

- **TCP connection teardown**
  - *netstat –an* | *grep –E 'CLOSED|FIN'* to see connections which are in the process of shutting down.
  - Connections that linger in the *CLOSED_WAIT* state indicate a local process owning the port has not issued a close() call against the socket
  - Likewise connections that linger in *FIN_WAIT2* indicate that the node and the _other_ end of the connection is not closing his socket.
  - Some connections can be terminated non- gracefully with a TCP RESET packet. Typically the setsockopt() socket option enables so_linger, but sets the linger time value to zero. This will result in a RESET on the connection as soon as the socket is closed.
  - Some network load balancers and firewalls will forcibly reset idle or problematic connections with RESET packets.

IP/UDP/TCP Tools - netstat

- **netstat – show network status**
  - netstat –in
    - IP interfaces configured
  - netstat –rnv
    - IP routing table
  - netstat –s
    - IP/UDP/TCP/ICMP/IGMP/ IPv6/ICMPv6 stats
  - netstat –an
    - transport AF_UNIX and AF_INET connection list
  - netstat –g
    - IP multicast address group membership

IP/UDP/TCP Tools - arp

- **arp – address resolution display and control**
  - arp –an
    - display the arp cache using IP addresses instead of names.
  - arp –d
    - delete an arp cache entry, forcing ARP resolution on next reference
  - arp –s
    - add a static arp cache entry manually
- **ndd –h | grep arp**

IP/UDP/TCP Tools - nettl

- **network tracing and logging utility.**
  - *nettl –ss | more* to see configured subsystems
  - *ns_ls_ip ns_ls_udp ns_ls_tcp*
    - *'-e all'* option traces all subsystems
  - packets can be traced at every layer
    - did it make it to the next layer?
    - what was the latency between layers?
  - Outbound packets are stamped with the kernel thread ID of the sending thread
  - 99Mbyte max raw trace file for 11.0 more for 11.11+

IP/UDP/TCP Tools - ndd

- **The ndd command allows the examination and modification of several tunable parameters that affect networking operation and behavior.**
  - *ndd –h* | *more* for general list of tunables
  - *ndd –h* <*specific tunable*> for detailed description
  - */etc/rc.config.d/nddconf* for tunables to set at startup time
  - tunables vary from 11.0 to 11.11+
  - tunables for IP, TCP, UDP, RAWIP, ARP, IPSEC, SOCKET

IP/UDP/TCP Tools - ifconfig

- **Configure or display network interface parameters**
  - *ifconfig lan0 inet 15.24.46.28 netmask 255.255.248.0 up*
  - *ifconfig lan0*

    lan0: flags=843<UP,BROADCAST,RUNNING,MULTICAST>
    inet 15.24.46.28 netmask fffff800 broadcast 15.24.47.255
  - *ifconfig lan0:1 inet 15.24.46.29 netmask 255.255.248.0 up*

```
# netstat -in | grep lan0
lan0:1        1500 15.24.40.0         15.24.46.29             34
  0
lan0          1500 15.24.40.0         15.24.46.28       1430849
  344305
```

IP/UDP/TCP Tools - route

- **manually manipulate the routing tables**

  /usr/sbin/route [-f] [-n] [-p pmtu] add [net|host] destination [netmask mask] gateway [count]

  /usr/sbin/route [-f] [-n] delete [net|host] destination [netmask mask] gateway [count]

  - adding network or host routes with altered path MTU
  - override default gateway assignment for networks/hosts
  - delete routes manually
  - refer to netstat –rn to see current routing table entries
  - *ndd –h* | *grep ip_ire* to see ndd tunables referring to routes.

IP/UDP/TCP Tools - traceroute

- /usr/contrib/bin/traceroute
- Maps out the path through a network between two nodes.
- traceroute [-dFlnrvx] [-f first_ttl] [-g gateway] [-i iface]
  [-m max_ttl] [-p port] [-q nqueries]
  [-s src_addr] [-t tos] [-w waittime]
  host [packetsize]

IP/UDP/TCP Tools - ping

- **Send ICMP Echo Request packets to network host**
  - *ping [-oprv] [-i address] [-t ttl] host [-n count]*
  - *ping [-oprv] [-i address] [-t ttl] host packet-size [[-n] count]*
- **Used to check IP connectivity**
- **Used to probe response to differing packet size**
  - *IP fragmentation check*
- **The –v and –p options**
  - useful for decoding the ICMP error messages routers may send in reply to you ping packet…
    - path MTU updates, sourcequenches

IP/UDP/TCP Tools - lsof

- **List Open Special Files utility.**
  - Lists open files for every running process and where possible tries to map it to a meaningful file name/path or type
  - Lists local and remote IP and UDP/TCP port numbers
  - provides socket address pointers for AF_INET and AF_UNIX sockets
  - shows what shared libs are mmap'ed in.
  - cwd for the process
  - Does not show kernel owned sockets…i.e. NFS

IP/UDP/TCP Tools – lsof (cont)

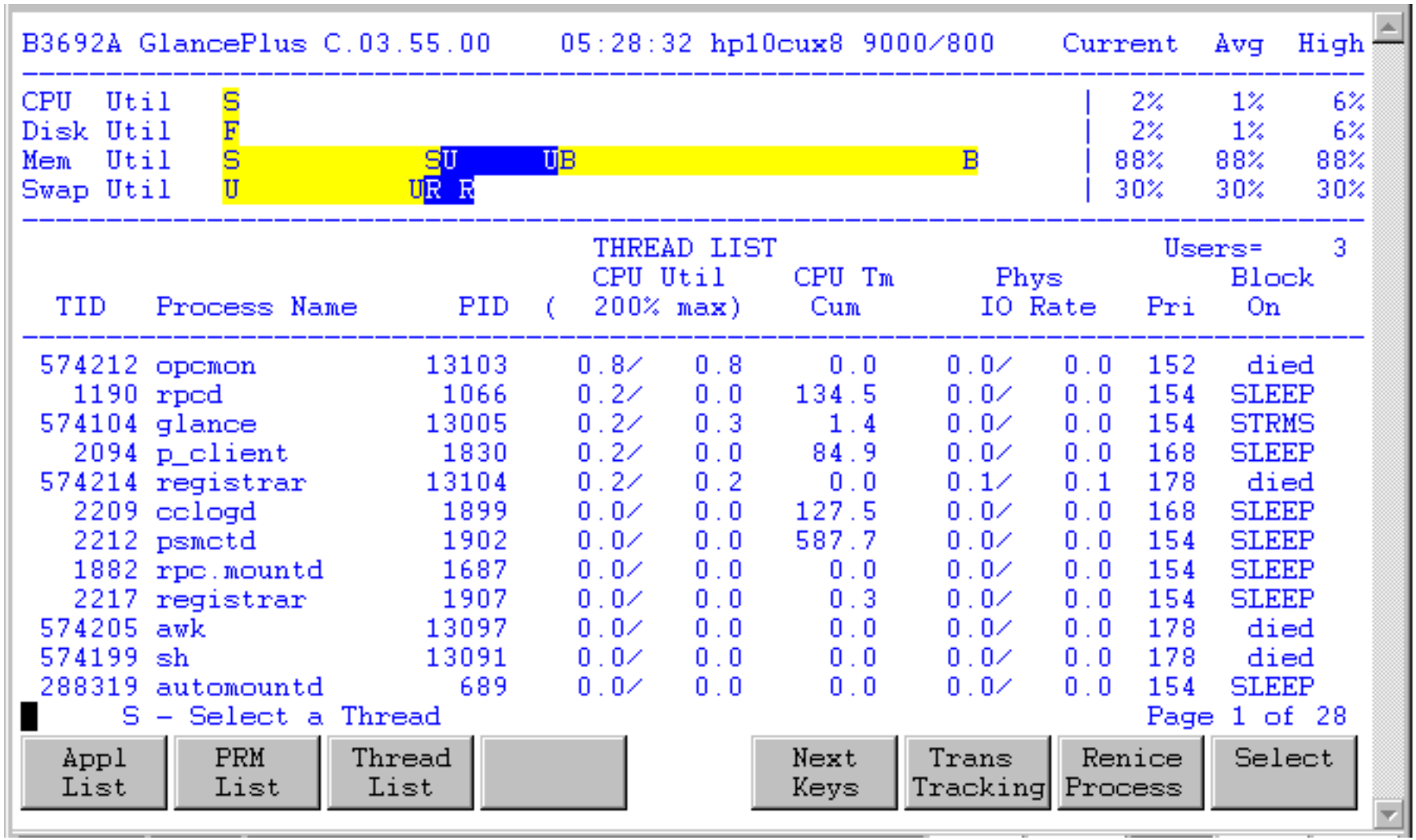

```
COMMAND   PID USER   FD   TYPE     DEVICE SIZE/OFF  NODE NAME
telnetd 16432 root  cwd    DIR     64,0x3     1024     2 /
telnetd 16432 root  txt    REG     64,0x6    90112 10827 /usr/lbin/telnetd
telnetd 16432 root  mem    REG     64,0x6    24576 13966 /usr/lib/libnss_dns.1
telnetd 16432 root  mem    REG     64,0x6    45056 13967 /usr/lib/libnss_files.1
telnetd 16432 root  mem    REG     64,0x6   135168   118 /usr/lib/libxti.2
telnetd 16432 root  mem    REG     64,0x6   724992   120 /usr/lib/libnsl.1
telnetd 16432 root  mem    REG     64,0x6    45056   165 /usr/lib/libnss_nis.1
telnetd 16432 root  mem    REG     64,0x6  1044480 14006 /usr/lib/libsis.sl
telnetd 16432 root  mem    REG     64,0x6    24576 13903 /usr/lib/libdld.2
telnetd 16432 root  mem    REG     64,0x6  1843200 13855 /usr/lib/libc.2
telnetd 16432 root  mem    REG     64,0x6   155648 13586 /usr/lib/dld.sl
telnetd 16432 root  mem    REG     64,0x7      532 10909 /var/spool/pwgr/status
telnetd 16432 root    0u  inet 0x4284c0c0      0t0   TCP 15.24.46.28:telnet->15.
24.46.33:1272 (ESTABLISHED)
telnetd 16432 root    1u  inet 0x4284c0c0      0t0   TCP 15.24.46.28:telnet->15.
24.46.33:1272 (ESTABLISHED)
telnetd 16432 root    2u  inet 0x4284c0c0      0t0   TCP 15.24.46.28:telnet->15.
24.46.33:1272 (ESTABLISHED)
telnetd 16432 root    3u   STR     32,0x1    0t101   499 /dev/telnetm->pckt->telm
telnetd 16432 root    4u  unix     64,0x7      0t0 11326 /var/spool/sockets/pwgr
                                                    /client16432 (0x43710700)
```

IP/UDP/TCP Tools - Glance

- **GlancePlus system performance monitor for HP-UX**
- **Useful screens…..**
  - Thread List
  - Process syscalls
  - Open files
    - shows offset within files and file types/names
  - Network by interface
  - NFS global and by system
    - stats plus read/write rates for client and server
  - Memory report
    - buffercache size and pagein/out rates

IP/UDP/TCP Tools - Glance

- Thread List screen



```
B3692A GlancePlus C.03.55.00    05:28:32 hp10cux8 9000/800     Current   Avg   High
-------------------------------------------------------------------------------
CPU  Util   S                                                  |   2%     1%    6%
Disk Util   F                                                  |   2%     1%    6%
Mem  Util   S          SU      UB                          B   |  88%    88%   88%
Swap Util   U          UR R                                   |  30%    30%   30%
-------------------------------------------------------------------------------
                                THREAD LIST                          Users=     3
                                CPU Util      CPU Tm      Phys                Block
  TID   Process Name      PID  ( 200% max)    Cum       IO Rate    Pri    On
-------------------------------------------------------------------------------
 574212 opcmon          13103    0.8/  0.8     0.0      0.0/  0.0  152   died
   1190 rpcd             1066    0.2/  0.0   134.5      0.0/  0.0  154   SLEEP
 574104 glance          13005    0.2/  0.3     1.4      0.0/  0.0  154   STRMS
   2094 p_client         1830    0.2/  0.0    84.9      0.0/  0.0  168   SLEEP
 574214 registrar       13104    0.2/  0.2     0.0      0.1/  0.1  178   died
   2209 cclogd           1899    0.0/  0.0   127.5      0.0/  0.0  168   SLEEP
   2212 psmctd           1902    0.0/  0.0   587.7      0.0/  0.0  154   SLEEP
   1882 rpc.mountd       1687    0.0/  0.0     0.0      0.0/  0.0  154   SLEEP
   2217 registrar        1907    0.0/  0.0     0.3      0.0/  0.0  154   SLEEP
 574205 awk             13097    0.0/  0.0     0.0      0.0/  0.0  178   died
 574199 sh              13091    0.0/  0.0     0.0      0.0/  0.0  178   died
 288319 automountd        689    0.0/  0.0     0.0      0.0/  0.0  154   SLEEP
     S - Select a Thread                                          Page 1 of 28
 Appl List | PRM List | Thread List |  | Next Keys | Trans Tracking | Renice Process | Select
```

IP/UDP/TCP Tools - Glance

- **Process syscalls**

```
B3692A GlancePlus C.03.55.00    05:39:00 hp10cux8 9000/800    Current   Avg   High
------------------------------------------------------------------------------
CPU  Util   S                                                  |  1%     2%    13%
Disk Util   F          F                                       | 21%     8%    85%
Mem  Util   S             SU      UB                        B  | 91%    88%    91%
Swap Util   U           UR R                                  | 30%    30%    30%
------------------------------------------------------------------------------
System Calls PID: 13364, sh              PPID: 13363 euid:     0 User: root
                                                 Elapsed                 Elapsed
System Call Name          ID     Count   Rate      Time   Cum Ct  CumRate CumTime
------------------------------------------------------------------------------
fork                       2         0    0.0   0.00000       2      0.0  0.00062
read                       3         0    0.0   0.00000      37      0.7 21.90270
write                      4         0    0.0   0.00000      39      0.7  0.00111
time                      13         0    0.0   0.00000       2      0.0  0.00000
stat                      38         0    0.0   0.00000       4      0.0  0.00018
ioctl                     54         0    0.0   0.00000      12      0.2  0.00030
setpgrp2                  82         0    0.0   0.00000       2      0.0  0.00001
sigvector                108         0    0.0   0.00000       8      0.1  0.00004
sigprocmask              185         0    0.0   0.00000     172      3.4  0.00039
waitpid                  200         0    0.0   0.00000       4      0.0 19.32314

        Cumulative Interval:      51 secs
                                                                  Page 1 of 1
Process Resource | Wait States | Memory Regions | Open Files | Next Keys | Process Syscalls
```

IP/UDP/TCP Tools - Glance

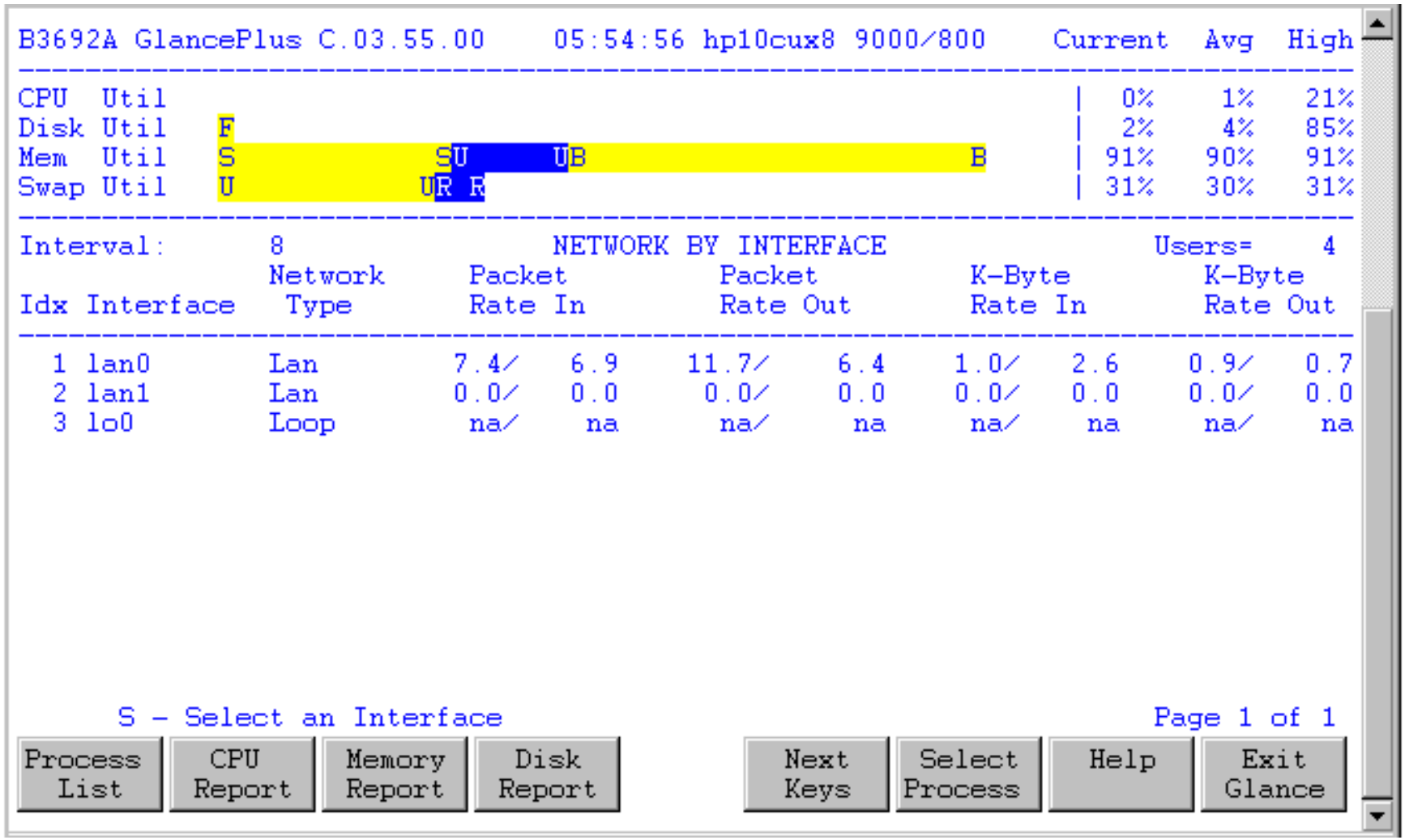
- Open files

```
B3692A GlancePlus C.03.55.00    05:45:50 hp10cux8 9000/800     Current   Avg   High
-------------------------------------------------------------------------------
CPU  Util                                                    |   1%     1%    21%
Disk Util   F   F                                            |  10%     7%    85%
Mem  Util   S            SU     UB                       B   |  91%    90%    91%
Swap Util   U            UR R                                |  31%    30%    31%
-------------------------------------------------------------------------------
Open Files PID: 13363, telnetd        PPID:   710 euid:     0 User: root
                                                      Open   Open
FD  File Name                                         Type   Mode   Count      Offset
-------------------------------------------------------------------------------
  0 inet,tcp                                          socket rd/wr    3            0
  1 inet,tcp                                          socket rd/wr    3            0
  2 inet,tcp                                          socket rd/wr    3            0
  3 /dev/telnetm                                      stream rd/wr    1          101
  4 unix /var/spool/sockets/pwgr/client13363          socket rd/wr    1            0

                                                                        Page 1 of 1
Process Resource | Wait States | Memory Regions | Open Files | Next Keys | Process Syscalls |  |
```

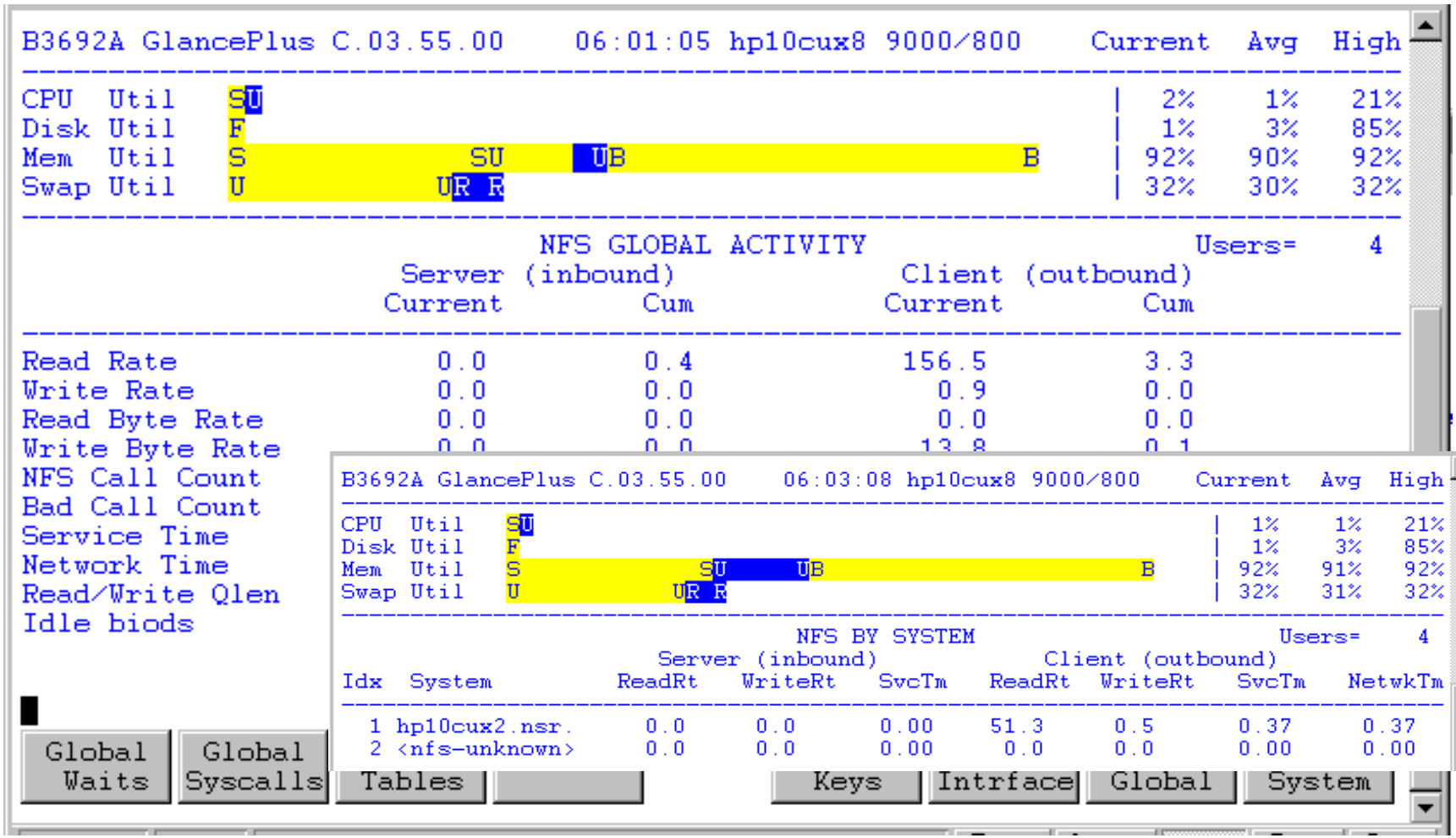
IP/UDP/TCP Tools - Glance

- Network by interface



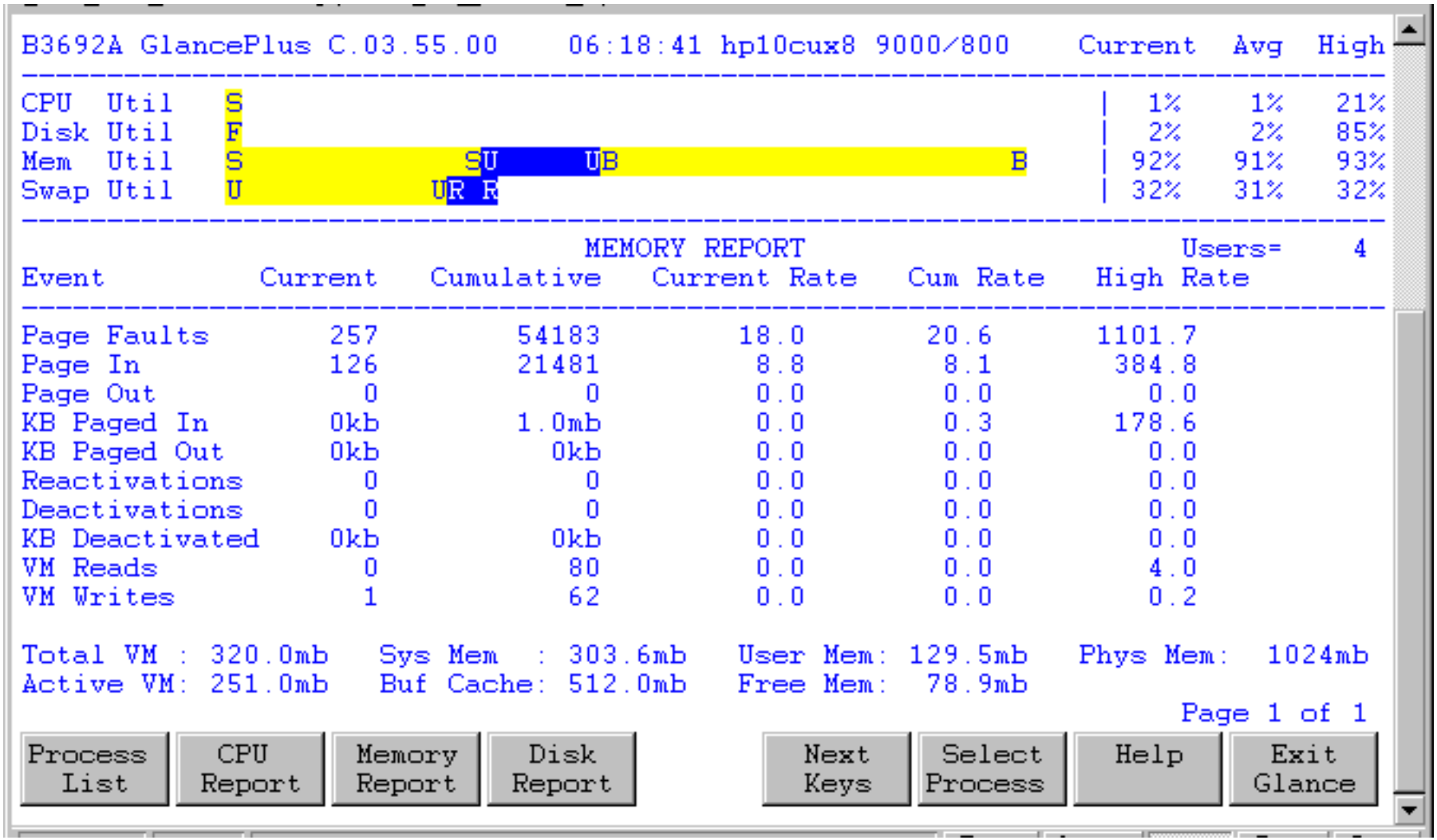
IP/UDP/TCP Tools - Glance

- NFS global and by system



IP/UDP/TCP Tools - Glance

- Memory report



```
B3692A GlancePlus C.03.55.00    06:18:41 hp10cux8 9000/800     Current   Avg   High
-------------------------------------------------------------------------------
CPU  Util   S                                                  |   1%     1%   21%
Disk Util   F                                                  |   2%     2%   85%
Mem  Util   S                SU     UB                     B  |  92%    91%   93%
Swap Util   U              UR R                               |  32%    31%   32%
-------------------------------------------------------------------------------
                                MEMORY REPORT                      Users=     4
Event            Current    Cumulative   Current Rate   Cum Rate   High Rate
-------------------------------------------------------------------------------
Page Faults          257         54183          18.0       20.6      1101.7
Page In              126         21481           8.8        8.1       384.8
Page Out               0             0           0.0        0.0         0.0
KB Paged In          0kb        1.0mb           0.0        0.3       178.6
KB Paged Out         0kb          0kb           0.0        0.0         0.0
Reactivations          0             0           0.0        0.0         0.0
Deactivations          0             0           0.0        0.0         0.0
KB Deactivated       0kb          0kb           0.0        0.0         0.0
VM Reads               0            80           0.0        0.0         4.0
VM Writes              1            62           0.0        0.0         0.2

Total VM : 320.0mb   Sys Mem   : 303.6mb   User Mem: 129.5mb   Phys Mem:  1024mb
Active VM: 251.0mb   Buf Cache: 512.0mb   Free Mem:  78.9mb
                                                          Page 1 of 1
Process List | CPU Report | Memory Report | Disk Report | Next Keys | Select Process | Help | Exit Glance
```

IP/UDP/TCP Tools - ttcp

- **Test TCP (TTCP) is a command-line sockets-based benchmarking tool for measuring TCP and UDP performance between two systems.**
- **simpler than netperf, UDP/TCP only**
- **public domain**

  http://hpux.cs.utah.edu/hppd/hpux/Networking/Admin/nttcp-1.47
- **typical usage:**

  ttcp –stp9 <host>
  - sends 2048 8k buffers to TCP port 9 (inetd's discard port) on target host.
  - can be run in server mode if no discard port available
- **No use of file system buffercache to skew results**

IP/UDP/TCP Tools - netperf

- **Benchmark tool for unidirectional and end-to-end latency testing. Test environments :**
  - TCP and UDP via BSD Sockets
  - DLPI
  - Unix Domain Sockets
  - Fore ATM API, HP HiPPI Link Level Access
- **Client/server model – netperf & netserver**
  - netserver started via command lien or inetd
  - two connections, control and test
- http://www.netperf.org/netperf/NetperfPage.html

IP/UDP/TCP Tools - iperf

- **A newer ttcp-like tool**
- **IPv6 Support**
- **IP multicast support**
- **Interface binding options**
- **Where to get it:**
  - http://dast.nlanr.net/Projects/Iperf

IP/UDP/TCP Tools - ftp

- **Quick and dirty.**
  - target file should be /dev/null to avoid disc/buffercache influence
  - run multiple time so source file is (hopefully) in buffercache.
  - Uses sendfile() system call.
  - 64k socket buffer max, default 56k.
  - ensure test file fits in buffercache

IP/UDP/TCP Tools – rcp/remsh

- **quick and dirty**
- **target should be /dev/null to avoid buffercache File system usage**
- **There are –S –R socket buffer size options**
- **remshd launched by inetd at remote**
- **Data read from pipes is in 1024 byte chunks**
- **ndd tcp tunables for default socket size**
  - tcp_recv_hiwater_def specifies the recv TCP window size used by default on the system…..don't forget to restart inetd after changes

IP/UDP/TCP Tools – sample socket programs

- **/usr/lib/demos/networking/socket**
  - sample UDP/TCP client server source files
    - sync and async models
- **/usr/lib/demos/networking/af_unix**
  - local AF_UNIX socket equivalent of same client server programs
- **/usr/lib/demos/networking/dlpi**

IP/UDP/TCP Tools – q4

- **A dump analysis tool in /usr/contrib/bin which can also be used to look at kernel data structures live.**
  - ndd commands give pointers to some key data structures
  - used when command line tools do not provide enough detail
  - Many perl scripts included to help dump various kernel data structures
    - /usr/contrib/lib/Q4
  - Typically used by the RC/WTEC/Labs

NFS client/server

- **typical performance related issues**
  - Biod tuning
  - number of nfsds
  - PV3 vs. PV2 TCP vs UDP
  - Automount (legacy vs. autofs)
    - autofs and LOFS
  - Buffercache
  - HPUX 11.0 vs. 11.11 and beyond
- **Tools**
  - nfsstat, nettl, rpcinfo, deamon logging AND….
  - Optimizing NFS Performance: Tuning and Troubleshooting NFS on HP-UX Systems by David Olker

Socket/Application Layer

- **Common problems**
  - Server/Listener performance
    - External influences
  - multiprocess vs. multithreaded
  - Connection management
- **What you really want to know…..**
  - Where is this process spending its time?
  - if it's stuck, where, and who has the resource I need and why are they being a pig about it…i.e. where is *that* process spending its time?
- **Tools - Application debugging Swiss Army Knife**
  - Netstat, nettl
  - Tusc, lsof, ps, glance
  - Application logging
  - Sample socket code

Socket/Appl Layer (cont)

- **Server/Listener Performance**
  - socket() bind() listen() accept()* select()*
  - any work the listener does prior to going back to the listen socket for another accept/select attempt can impede performance….that includes fork() and the handoff of the new connection to a child process
    - authentication (gethostbyaddr(), getpwnam(), DNS, NIS)
    - IPC mechanisms to another slow-as-mud process
    - other socket connections made
  - Listen backlog queue size
    - maximum accept() rate?
    - *ndd –get /dev/tcp tcp_conn_request_max*
    - *netstat –sp tcp | grep 'full queue'*
    - on client *netstat –an | grep SYN_SENT*
  - *Glance* process syscalls screen for listen process
  - nettl trace of traffic to/from listen port

Socket/Appl Layer (cont)

- Multiprocess vs. Multithreaded
  - Attention to operations which are fork-safe and which are thread-safe.
  - XTI library routines are thread safe but not fork safe while the equivalent operation in a sockets based application is fork-safe.
    - t_accept() vs. accept() for example

Socket/Appl Layer (cont)

- **Connection management**
  - after the accept() who/what handles the client transactions and what do they spend their time doing?
    - ask the developers first
    - then use tusc, glance, and application logging to see if it looks even remotely like what they described.
  - socket options for send/recv buffer sizes tuned to link topology and data profile
    - bulk one-way data xfer or small bidirectional exchanges. Link bandwidth and latency.
  - connection close
    - shutdown() can be for read, write, or both
    - SO_LINGER socket option used?
    - nettl tracing can show latencies not obvious at the application level.
    - close() on a socket does not always mean the TCP connection has terminated gracefully.

Socket/Appl tools - netstat

- **Connection states via *netstat -an***
  - **SYN_SENT**
    - trying to contact remote host….either host is not up or the listen queue is full….otherwise you would have seen a RESET packet
  - **ESTABLISHED**
    - normal state for active TCP connection
  - **CLOSE_WAIT**
    - FIN received from remote end and waiting for local process owning the port to issue a close.
  - **FIN_WAIT2**
    - FIN sent to remote, and ACK'ed, but remote process has not closed his end of connection…the partner state to CLOSE_WAIT on the remote host.
  - **TIME_WAIT**
    - 60 second state after both ends close gracefully

Socket/Appl tools - nettl

- **Trace and filter by IP address or UDP/TCP port number or sending kernel thread ID.**
  - used to see the transport view of a client/server transaction/traffic.
  - will show TCP options being set that reflect socket calls such as specifying the recv socket buffer size.
  - will show flow control at TCP layer to indicate application level timeliness of reading from recv socket buffer.
  - Does not trace AF_UNIX sockets or pipes.

Socket/Appl tools - tusc

- **Trace unix System Call**
  - traces process syscalls
  - follows forks and threads
  - wall clock time and kernel syscall times provided
  - verbose mode decodes most syscall arguments in detail
  - select masks, socket IP/port info, semop args
  - shows all shared libs being mmap'ed in
  - can trace multiple PID's
  - Does affect performance…a bunch
- **The single most powerful tool for application debugging next to application logging.**
- **Latest version is 7.5 (7.3 also ok and is widely available)**

Socket/Appl tools - tusc


```
( Attached to process 22862 ("inetd") [32-bit] )
1027349371.048254 {650884} <0.000000> select(35, 0x7f7f08d0, NULL, NULL, NULL) [
sleeping]
                           readfds: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
                          writefds: NULL
                          errorfds: NULL
1027349385.649931 {650884} <0.000109> select(35, 0x7f7f08d0, NULL, NULL, NULL) =
 1
                           readfds: 6
                          writefds: NULL
                          errorfds: NULL
1027349385.650604 {650884} <0.000163> accept(6, NULL, NULL) = 35
1027349385.651985 {650884} <0.000031> getpeername(35, 0x7f7f09d0, 0x7f7f09ec) =
0
                           *fromlen: 16
                        sin_family: AF_INET
                          sin_port: 3561
                    sin_addr.s_addr: 15.24.46.33
1027349385.652323 {650884} <0.000045> stat("/var/adm/inetd.sec", 0x40004738) = 0
    - skipping forward a bit -
1027349385.655274 {650884} <0.000864> fork() ............. = 27728 {656254}
1027349385.655422 {656254} <-0.000000> fork() (returning as child ...) = 22862 {
650884}
    - skippping forward  a bit -
1027349385.692340 {656254} <0.001165> execve("/usr/lbin/telnetd", 0x4000aff0, 0x
7f7f055c) = 0 [32-bit]
```

Color code –

Wall clock time

Kernel TID

System CPU time

Syscall with args

Return value

verbose call detail

Socket/Appl tools - lsof

- **List Open Files**
  - finds every running process, maps out the open file descriptors and displays details.
  - shows cwd, mmap'ed file, regular files, sockets
    - resolves questions about which shared libs were really used
  - displays IP/UDP/TCP port info for AF_INET sockets
  - displays struct socket ptr for every socket…AF_INET and AF_UNIX
  - displays partner AF_UNIX stream socket addr
    - you can see which two processes own opposite ends of the AF_UNIX stream socket connection.
  - used to help map tusc data to process files
    - tusc doesn't always trace the open of a file and hence cannot provide the name

Socket/Appl tools – ps -elf

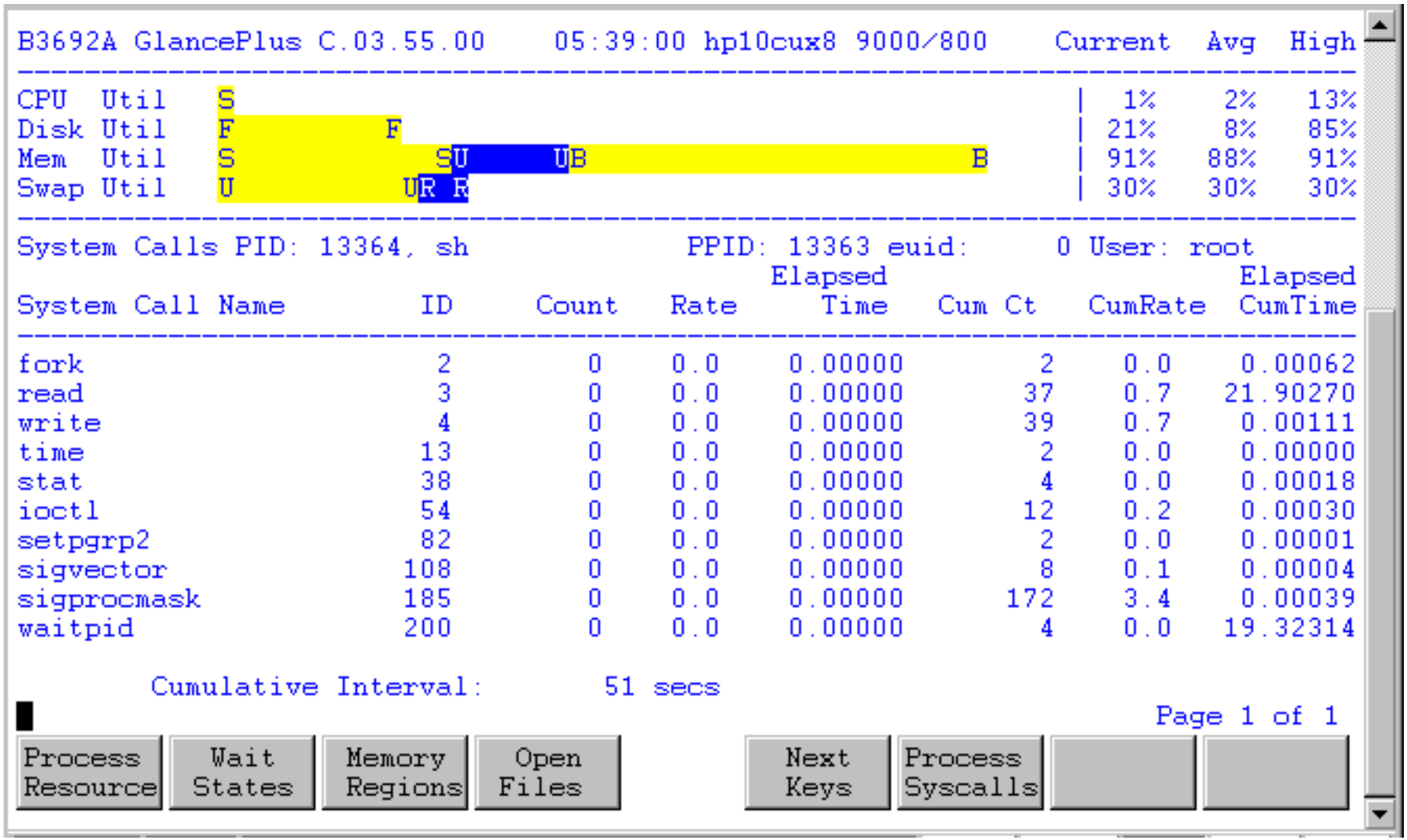
- **Maps process names to pids for cross reference with other tools**
- **The wait channel**
  - address/token passed to sleep()/sleep_one()
- **The proc structure address**
- **incore memory image size**
  - data, text, stack
  -
- **total CPU execution time**

Socket/Appl tools - glance

- **Screens of interest for application level troubleshooting**
  - **process syscalls**
    - looking for excessive/unusual syscall counts/rates
    - looking for unusual CPU time associated with a particular syscall
  - **process resources**
    - context switches –forced vs. voluntary
  - **Wait states**
    - pipe, socket, stream, rpc
  - **memory regions**
    - RSS VSS and mmap'ed regions
  - **Open files**
    - names, types, open modes, offsets
  - **Thread list**
    - cross referencing for nettl

Socket/Appl Tools - Glance

- Process syscalls



```
B3692A GlancePlus C.03.55.00     05:39:00 hp10cux8 9000/800     Current   Avg   High
-------------------------------------------------------------------------------
CPU  Util    S                                                  |   1%     2%   13%
Disk Util    F          F                                       |  21%     8%   85%
Mem  Util    S            SU      UB                        B   |  91%    88%   91%
Swap Util    U          UR R                                    |  30%    30%   30%
-------------------------------------------------------------------------------
System Calls PID: 13364, sh              PPID: 13363 euid:     0 User: root
                                                 Elapsed                 Elapsed
System Call Name          ID     Count    Rate     Time    Cum Ct  CumRate CumTime
-------------------------------------------------------------------------------
fork                       2         0     0.0  0.00000         2     0.0  0.00062
read                       3         0     0.0  0.00000        37     0.7 21.90270
write                      4         0     0.0  0.00000        39     0.7  0.00111
time                      13         0     0.0  0.00000         2     0.0  0.00000
stat                      38         0     0.0  0.00000         4     0.0  0.00018
ioctl                     54         0     0.0  0.00000        12     0.2  0.00030
setpgrp2                  82         0     0.0  0.00000         2     0.0  0.00001
sigvector                108         0     0.0  0.00000         8     0.1  0.00004
sigprocmask              185         0     0.0  0.00000       172     3.4  0.00039
waitpid                  200         0     0.0  0.00000         4     0.0 19.32314

        Cumulative Interval:       51 secs
                                                                 Page 1 of 1
Process Resource | Wait States | Memory Regions | Open Files | Next Keys | Process Syscalls
```

Socket/Appl Tools - Glance

- **Process Resource**

```
B3692A GlancePlus C.03.55.00    05:45:12 hp10cux8 9000/800    Current  Avg  High
-------------------------------------------------------------------------------
CPU  Util                                                   |   0%    1%    2%
Disk Util   F                                               |   2%    3%    5%
Mem  Util   S             SU      UB     B                  |  52%   52%   52%
Swap Util   U  UR  R                                        |  15%   15%   15%
-------------------------------------------------------------------------------
Resources PID: 23559, glance         PPID: 23076 euid:     0 User: root
-------------------------------------------------------------------------------
CPU Usage (util):      0.2 Log Reads :       1 Wait Reason    :    STRMS
User/Nice/RT CPU:      0.0 Log Writes:       0 Total RSS/VSS  :  2.9mb/  6.3mb
System CPU       :     0.0 Phy Reads :       0 Traps / Vfaults:      1/      3
Interrupt CPU    :     0.0 Phy Writes:       0 Faults Mem/Disk:      0/      0
Cont Switch CPU  :     0.0 FS Reads  :       0 Deactivations  :      0
Scheduler        :    HPUX FS Writes :       0 Forks & Vforks :      0
Priority         :     154 VM Reads  :       0 Signals Recd   :      1
Nice Value       :      10 VM Writes :       0 Mesg Sent/Recd :      0/      0
Dispatches       :       2 Sys Reads :       0 Other Log Rd/Wt:     33/     12
Forced CSwitch   :       0 Sys Writes:       0 Other Phy Rd/Wt:      0/      0
VoluntaryCSwitch:       1 Raw Reads :       0 Proc Start Time
Running CPU      :       1 Raw Writes:       0    Fri Jul 26 05:44:43 2002
CPU Switches     :       0 Bytes Xfer:     0kb                :

     C - cum/interval toggle    % - pct/absolute toggle          Page 1 of 1
Process Resource | Wait States | Memory Regions | Open Files | Next Keys | Process Syscalls
```

Socket/Appl Tools – Glance

- **wait states**

```
B3692A GlancePlus C.03.55.00    14:45:31 hp10cux8 9000/800    Current  Avg  High
-------------------------------------------------------------------------------
CPU  Util                                                    |   1%    1%    3%
Disk Util                                                    |   0%    1%    4%
Mem  Util    S             SU    UB     B                   |  52%   52%   52%
Swap Util    U   UR R                                        |  15%   15%   15%
-------------------------------------------------------------------------------
Wait States PID:   740, automountd     PPID:     1 euid:     0 User: root

Event               %        Blocked On          %
-------------------------------------------------------------------------------
IPC          :      0.0       Cache       :      0.0       CPU Util   :     0.0
Job Control:        0.0       CDROM IO    :      0.0       Wait Reason:   SYSTM
Message      :      0.0       Disk IO     :      0.0
Pipe         :      0.0       Graphics    :      0.0
RPC          :      0.0       Inode       :      0.0
Semaphore    :      0.0       IO          :      0.0
Sleep        :     50.0       LAN         :      0.0
Socket       :      0.0       NFS         :      0.0
Stream       :      0.0       Priority    :      0.0
Terminal     :      0.0       System      :     50.0
Other        :      0.0       Virtual Mem:      0.0

      C - cum/interval toggle   % - pct/absolute toggle              Page 1 of 1
Process List | CPU Report | Memory Report | Disk Report | Next Keys | Select Process | Help | Exit Glance
```

Socket/Appl Tools - Glance

- Memory Regions

```
B3692A GlancePlus C.03.55.00    06:09:14 hp10cux8 9000/800    Current   Avg   High
-------------------------------------------------------------------------------
CPU  Util                                                    |   0%     1%     6%
Disk Util                                                    |   0%     1%     6%
Mem  Util    S             SU    UB    B                     |  50%    50%    52%
Swap Util    U  UR R                                         |  14%    14%    15%
-------------------------------------------------------------------------------
Memory Regions PID:   740, automountd     PPID:     1 euid:     0 User: root

Type          RefCt     RSS      VSS    Locked  File Name
-------------------------------------------------------------------------------
NULLDR/Shared    86      4kb      4kb       0kb <nulldref>
TEXT  /Shared     2     68kb     68kb       0kb <reg,vxfs,/...6,inode:13>
DATA  /Priv       1    1.4mb    1.5mb       0kb <reg,vxfs,/...6,inode:13>
MEMMAP/Priv       1     36kb    2.0mb       0kb <mmap>
MEMMAP/Priv       1      4kb      4kb       0kb /usr/lib/libnss_dns.1
MEMMAP/Priv       1      8kb      8kb       0kb <mmap>
MEMMAP/Priv       1      4kb      4kb       0kb <reg,vxfs,/...node:14017>
MEMMAP/Priv       1      8kb      8kb       0kb /usr/lib/libpthread.1
MEMMAP/Priv       1      4kb      4kb       0kb <mmap>

Text  RSS/VSS: 68kb/ 68kb  Data  RSS/VSS:1.4mb/1.5mb  Stack RSS/VSS: 48kb/ 48kb
Shmem RSS/VSS:  0kb/  0kb  Other RSS/VSS:2.2mb/4.5mb
                                                              Page 1 of 4
Process Resource | Wait States | Memory Regions | Open Files | Next Keys | Process Syscalls
```

Socket/Appl Tools - Glance

- Open files

```
B3692A GlancePlus C.03.55.00    06:12:45 hp10cux8 9000/800    Current  Avg  High
-------------------------------------------------------------------------------
CPU  Util                                                    |   0%    1%    6%
Disk Util    F                                               |   1%    1%    6%
Mem  Util    S              SU    UB    B                    |  50%   50%   52%
Swap Util    U  UR R                                         |  14%   14%   15%
-------------------------------------------------------------------------------
Open Files PID:   740, automountd     PPID:     1 euid:     0 User: root
                                                        Open   Open
FD  File Name                                           Type   Mode  Count      Offset
-------------------------------------------------------------------------------
  0 /dev/null                                           chr    read    18           0
  1 /etc/rc.log                                         reg    write   35       14142
  2 <reg,vxfs,/var,/dev/vg00/lvol7,inode:18>            reg    write    1      402274
  3 /dev/tlclts                                         stream rd/wr    1           0
  4 /dev/tlcotsod                                       stream rd/wr    1           0
  5 /dev/tlcots                                         stream rd/wr    1           0
  6 <fifo,hfs,inode:368>                                fifo   read    1           0
  7 <fifo,hfs,inode:368>                                fifo   write   1           0
  8 /dev/udp                                            stream rd/wr    1           0

                                                                        Page 1 of 1
Process Resource | Wait States | Memory Regions | Open Files | Next Keys | Process Syscalls |  |
```