SNS Control Systems



A new Tool to study Network Stack Exhaustion in

VxWorks

Epics Collaboration Meeting

Dec. 8, 2004

Sheng Peng Ernest L. Williams Jr. David Thompson

ICS _ Software Engineering Group













- When we were dealing with "IOC Disease" earlier this year we got pretty good at using vxWorks diagnostics tools, mBufShow, inetStatShow, and a few that WRS gave us like ifQValuesShow.
- We got pretty good at "tuning" by setting mbufs, driver queues, and the "if_Q length".
- We found and fixed several causes of depleted buffers.
- We still have errors! Diagnostics like ifShow indicates txErrors and we still get white screens.
- The end driver with debugging turned on also reports txErrors.









The first round of cures



inetstatShow

Active Internet connections (including servers) PCB Proto Recv-Q Send-Q Local Address Foreign Address (state)

1b4a990 TCP 0 8184 172.31.124.20.5064 172.31.124.107.51553 << Archive server

••••

mbufShow

CLUSTER POOL TABLE

size clusters free usage

64 800 772 9859 128 1600 1531 105147601 256 800 800 2138545 512 400 400 34635 1024 200 200 1913 2048 300 300 27947 4096 20 20 6197

Eventually the archive server would consume all of the buffers because daily restarts never closed the old sockets. It usually took several days to a week for the IOC to crash, especially with large buffer configurations.

Other clients, were problems as well. Edm with a stuck mouse would do the same thing!

The net result is that we understand this and have fixed the problems with clients for the most part.





mm









- Now what? We still have problems and the IOCs have plenty of free buffers in the network stack.
- Maybe it is time to look at traffic patterns.
- Bring in etherreal!













• IOC Under Study:

- » ScI-hprf-ioc05 (without Beckhoff driver)
 - Connected to CISCO 2950 layer2 switch
 - Iin-ics-netsw3b1 ---- port 1
 - Port 1 is mirrored for observation via a linux-based packet capture and analysis system.

Tools used:

- » Laptop with "Ethereal" packet capture software
 - NIC 1 (eth0) ---- used for remote access to the packet capture station
 - NIC 2 (eth1) --- connected to the CISCO port mirror.





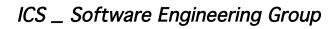




What we will study?



- What we will study:
 - » Network Memory Resources Model
 - What are mBufs?
 - What are mBlks?
 - What are cBlks?
 - What are clusters?
 - » Flow diagram of the vxWorks Network Stack
 - How are packets moved in out with respect to the OSI model?
 - The journey of a network packets as seen through the eyes of a network sniffer in an EPICS environment.
 - We will make a timeline of events from the time an IOC is booted to when it is "open for business"
 - CISCO port auto-negotiation and turn-on
 - Loading of vxWorks image from boot server
 - Re-setting of IOC's network hardware by vxWorks
 - Loading startup file common to all vxWorks IOCs (i.e. common.cmd)
 - Loading application specific startup file (i.e. st.cmd)
 - IocInit
 - » What protocols are showing up?
 - Needed in the context of EPICS
 - Nuisance Protocols/traffic

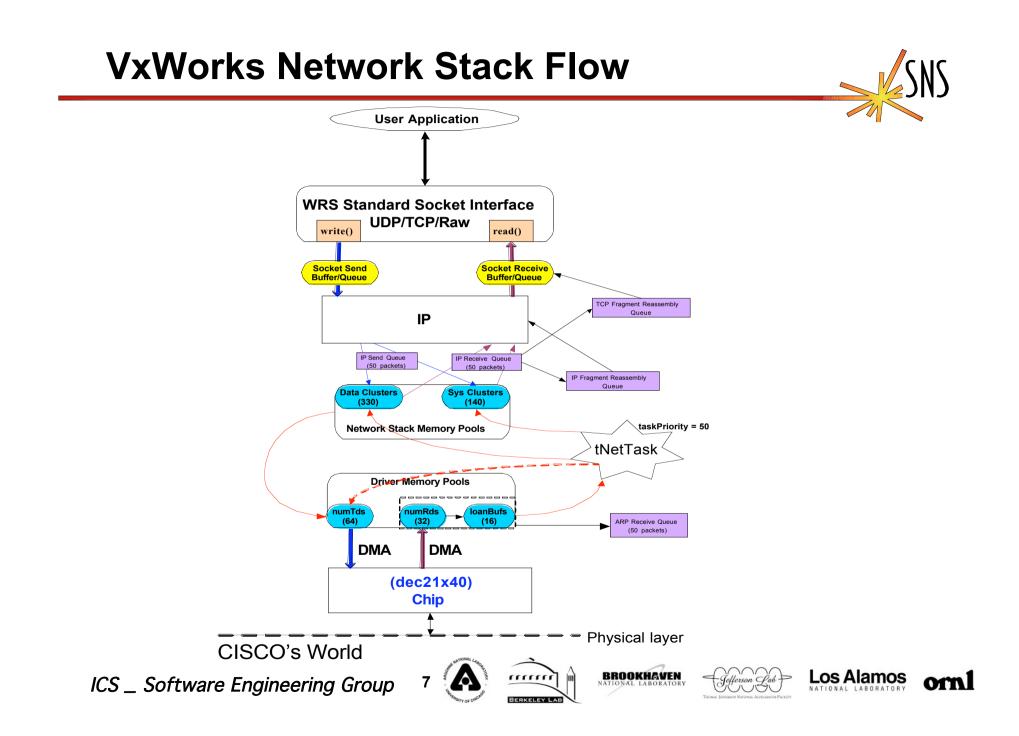












Real-Time OS Considerations



- » Pre-allocated buffers as opposed to dynamic from the global heap at run-time
- Timers
 - » Connection management
 - » Timeouts
 - » Retries
- Latency
 - » Fast and deterministic interrupt handling interfaces
 - » Small thread context switch times
- Concurrency
 - » Smart use of semaphores
- Minimized Data Copying
 - The TCP/IP implementation should minimize the amount of data copying. The data within each frame can be maintained in the same buffer so it doesn't need to be copied and re-copied by the CPU at each stage of the protocol. The networking chip's DMA places the packets directly in the managed buffer pool where the packet is passed up through the stack by manipulating pointers and not by copying data. The mbuf mechanism has been extended to allow the data to be shared between mbufs and mblocks where there are STREAMS protocols also present in the system.









Protocols we deal with in EPICS



- UDP port 5065
 - » CA beacons ("I am here" Heartbeat)
 - » Used to re-establish CA TCP virtual circuits
 - » CA beacons do not expect any replies
 - » The CA Beacon Daemon is listening on UDP port 5065.
 - A.K.A caRepeater
- UDP port 5064
 - » CA search message
 - » A response is expected within some timeout interval
- TCP port 5064
 - » CA server establishes a virtual circuit on port 5064
- NFS
 - » UDP port 111
 - Loading up IOC application
 - Running autosave/restore
 - Re-directing IOC files to boot server
- NTP
 - » UDP port 123
 - Keep IOCs time in synch
 - At SNS we should see this about every 10 seconds in our current configuration.

RSH

- » UDP port 514
 - Remote login support
 - Cat in vxWorks image



rrrrr





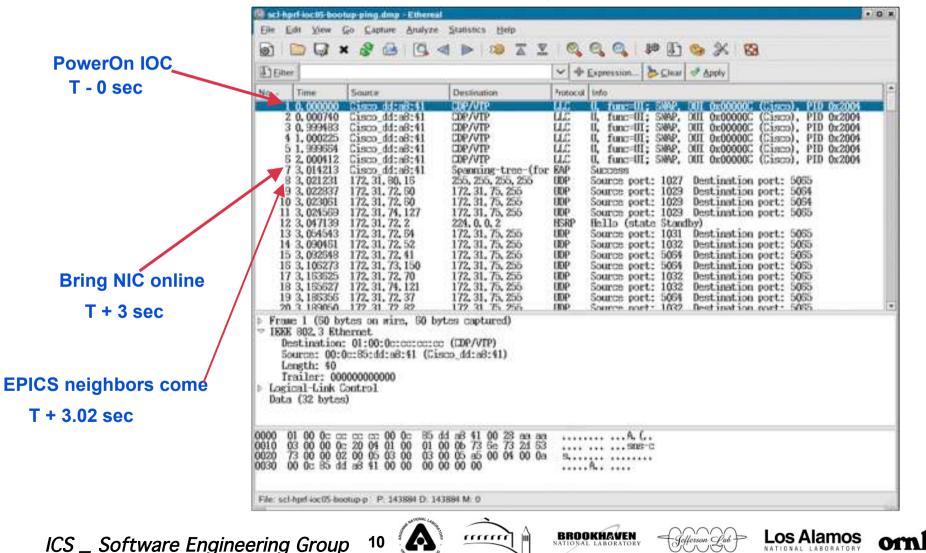




Network Traffic Analysis



Ethereal Packet Analysis Timeline



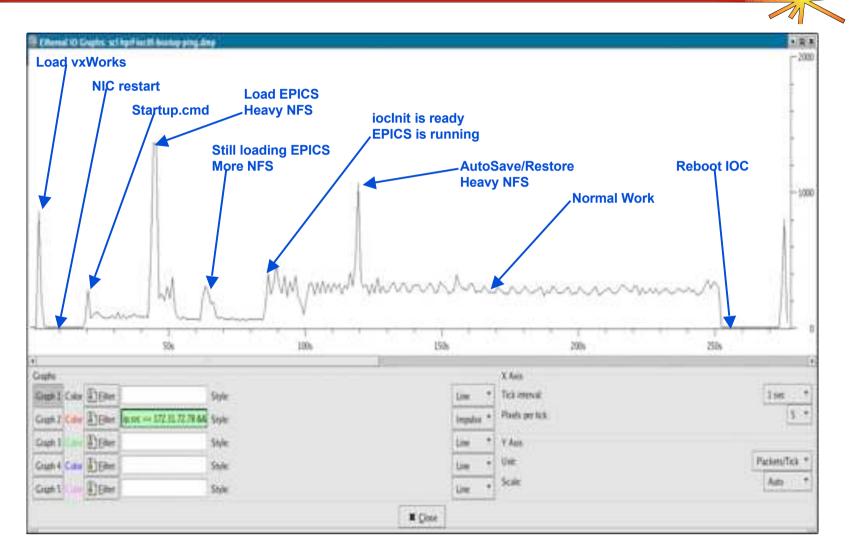
BERKELEY LA



Ethereal Packet Analysis Timeline

	Nu., Time Source Destruction Protocol Mo
	52948 313, 30428 172, 31, 72, 78 172, 31, 72, 101 975 1972 centramemistation of f3008 92 LONDP Call (Reply In 3309), DECMENDIQUES-1-7 52948 313, 30428 172, 31, 72, 78 172, 31, 72, 101 975 1972 centramemistation of f3012 92 LONDP Call (Reply In 3309), DECMENDIQUES-1-7 52952 313, 30646 172, 31, 72, 78 172, 31, 72, 101 975 1972 centramemistation of f3012 92 LONDP Call (Reply In 3313), DECMENDIQUES-1-7 52953 313, 30646 172, 31, 72, 78 172, 31, 72, 101 975 1972 centramemistation of f3014 92 LONDP Call (Reply In 3315), DECMENDIQUES-12000 (PEEC) DECMENDIQUES-12000 (PEEC) DECMENDIQUES-12000 (PEEC) DECMENDIQUES-12000 (PEEC) DECMENDIQUES-12000 (PEEC) DECMENDIQUES-110000 (PEEC) DECMENDIQUES-10000 (PEEC) DECMENDIQUE
/arning!! S is heavy	 Frame 2187 (152 bytes on mirs, 152 bytes captured) Elternet II, Sec: 00:001:af:00:25:03, Det: 00:05:5b:3e:7b:0a Internet Protocol, Sec Addr: 172, 31, 72, 79 (172, 31, 72, 70), Det Addr: 172, 31, 72, 101 (172, 31, 72, 101) Oner Detagma Protocol, Sec Port: iron (396), Det Port: ofn (2040) Remote Procedum Call, Type:Call KID:00:09:098357 Network File System, LONOF Call DE:0m#3320bo/bootLine Why is a retransmit necessary, hmmm?
	0000 00 05 34 35 34 00 01 of 02 25 13 06 00 45 00 D.L

Network Traffic Analysis (Annotated)





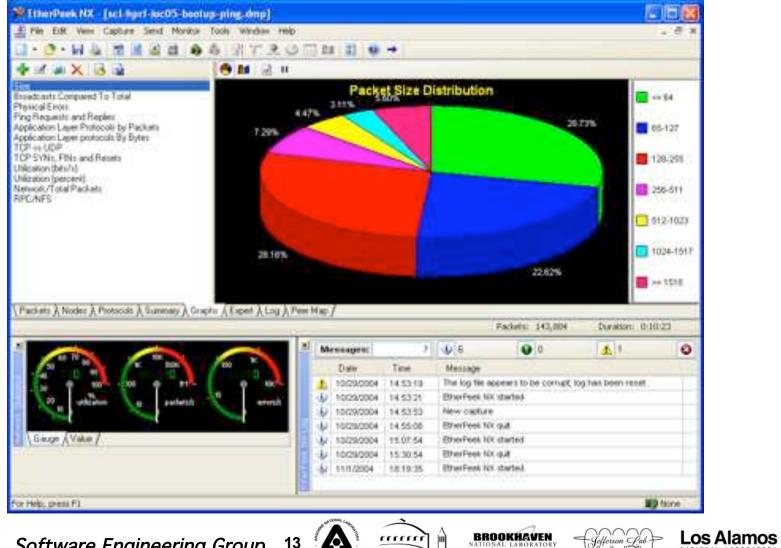






Network Analysis (Packet Size Distribution)

Scl-hprf-ioc05



ICS _ *Software Engineering Group*

A 13

BERKELEY LAB

NAL LABORATORY

UUGU

orn

Network Analysis (NFS/RPC statistics)

Scl-hprf-ioc05

k Elle Edit Yere Capitur	a Send Monitos Iools Window	19/0:					2.50
	134 06 RT3	and the second second second	21 21 4	-			
Details for RPC NFS	the probability of the local data and the second						
and the second s	6,353 Largerst packet: 05,402 Smalleret packet: 351072 Average packet size:	1,508 74 445	0-94	234	81		
Node			Percentage *				Packets /
18-172.31.72.101		39,9639 +				2,257,392	3,176 7,111
#11:36:78:94		39,9039				2,267,392	3,178 3,178
F-172.31.72.78	1000 B	10.0179				568,070	3,177
Protocol		Percentage				Eytes	Packets =
Sithernet Type 1		0.000%				0	01
(E) 17P	0,000% 0,000%				0	0	
- 100/873	110,000					2,035,402	¥,202]
		1 10	Messages:	7	QU 6	0 0	1
50 10 10	12 T 102 100 1	Date Time Mestage				- Million - Contractor	
	a a a a					rers to be comust, to	g has been re
the second	a salent ann		10/29/0804	14:53:21	EtherPeek tox st	leftel.	
	ALL DOT OF THE OWNER		10/29/2004	14 53 53	New capture		
and the second second		10/29/2004 14:55:08 EtherPeek NV 0					
Gauge (Value /			4 10/29/2904	15.07:54	EtherPeak Nil st	A 4 5 1 0	
			10(29/2004	15:30:54	EtherPeek KK guit		
			11/1/2004	10.10.25	EtherPeek HK d	w161	
		13					

ICS _ Software Engineering Group

14

BERKELEY







Network Analysis:

Data Collection on Network Queues



PROTOCOL RECEIVE QUEUES

• Healthy:

» dtl-IIrf-ioc1a> protocolQValuesShow
 IP receive queue max size = 50
 IP receive queue drops = 0

ARP receive queue max size = 50 ARP receive queue drops = 0 value = 28 = 0x1c

• Unhealthy:

» scl-hprf-ioc05> protocolQValuesShow
 IP receive queue max size = 50
 IP receive queue drops = 107

ARP receive queue max size = 50 ARP receive queue drops = 0 value = 28 = 0x1c











Network Analysis: Data Collection on Network Queues

IP SEND QUEUES

• Healthy:

» dtl-IIrf-ioc1a> ifQValuesShow("dc0") dc0 drops = 0 queue length = 0 max_len = 100 value = 46 = 0x2e = '.'

• Unhealthy:

» scl-hprf-ioc05> ifQValuesShow("dc0")
dc0 drops = 200 queue length = 0 max_len = 100
value = 48 = 0x30 = '0'



mm









What can go wrong with the Network Stack?

- Disruption of tNetTask via deadlock causing sockets not to be read.
- User tasks in general should have a priority lower than tNetTask. (i.e. greater than 50)
- Do not create and then take SEM_INVERSION_SAFE semaphores before making a socket call or your task could be promoted to run at tNetTask level
 - » tNetTask netTask 1cee480 0+I PEND











What can go wrong with the Network Stack?

- Application may have deadlock conditions which prevent them from reading sockets.
- If inetstatShow (or equivalent in other systems) displays data backed up on the send side and on the receive side of the peer, most likely there is a deadlock situation within the client/server application code.
- Running both server and client in the target by sending to 127.0.0.1 or to the target's own IP address is a good way to detect this kind of problem.
- Heavy NFS traffic may require an increase in driver memory pool.











Results/Conclusions



- The Network Analysis allows tuning of the network stack from apriori information as well as empirical data collected from the real environment.
- We have discovered some devices on our network that have improper configurations and hence cause unnecessary traffic.
- We have discovered that NFS is really a heavy hitter and that autosave/restore request files should be stored in one location.
- We have discovered that IGMP snooping must be supported on the CISCO edge switches to contain Allen Bradley Control Logix PLC multicast traffic. Multicast traffic should be contained in general.
 - » We moved from the CISCO 3500 series to the CISCO 2950 series
 - CISCO 3500 series only supported CGMP snooping
- We learned that sometimes IOC application errors are the main cause of Network Stack Exhaustion and/or failure.
- We have added an "open-source" network sniffer (Ethereal) to our EPICS Network trouble-shooting ToolKit.
- We have built in the Network diagnostics show routines from WRS in to our IOC's common support library.









Outline



- Introduction
- Implementing a network stack in the context of a Real-Time OS (RTOS)
- Basic Definitions and Memory Pools
- Network Stack Flow Diagram
- Network Traffic Analysis (w/ethereal)
- What can go wrong with the Network Stack?
- Results/Conclusion











Basic Definitions



Fundamental Data Structures

- Mbufs (deprecated):
 - » stores small stack data structures such as socket addresses, and packet data. Mbufs were designed to facilitate passing data between network drivers and the network stack, and contain pointers that can be adjusted as protocol headers are added or stripped. Mbufs contain space within them to store small amounts of data. Larger amounts of data were stored in fixed-sized clusters (typically 2048 bytes), which could be referenced and shared by more than one mbuf.
- Clusters:
 - » Network Data containers of various sizes in bytes
 - » Data containers must be a power of two
- cBlks:
 - The cBlk is a structure that contains a pointer to the cluster data, the cluster size, and an external reference count. the "cluster block" was added, supporting the zbuf sockets

interface and multiple network pools in addition to cluster sharing.

- » One cluster block is required for each cluster
- mBlks:
 - The mBlk is a structure that contains a pointer to a cBlk or another mBlk. mBlks are basically a modified version of the BSD style mbufs. The difference is that they now reference external clusters rather than carrying data directly. They are now called "mblocks."











- Network Stack "Data" Pool:
 - Data pools are used for packet send data with extra space for protocol headers. Clusters from the pool are allocated in the socket layer. The function which offers info about it is netStackDataPoolShow(). You can configure it with the definition of NUM_64, ..., NUM_2048.
 - » application layer \rightarrow network stack layer \rightarrow network driver
- Network Stack "System" Pool:
 - » System pools are used for network structures (sockets, routes, etc). The function to offer info about it is netStackSysPoolShow(). You can configure it with the definition of NUM_SYS_64, ..., NUM_SYS_512
- Network "Driver" Interface Pool:
 - » Buffer pool for each network interface. Data from the wire is received in clusters from a network device pool. These buffers are then passed up to the network stack. This pool is also used for staging packets to be transmitted by the target. The driver pool can be shown with the following utility routine: endPoolShow("dc",0) for our MVME2101 boards. Call muxShow() to show network driver info.









More on the Driver's Pool



- Cluster size for ethernet is 1520
 - » Cluster size has to be big enough to receive or transmit the maximum packet size allowed by the link layer. In this case that is 1518 bytes
 - » Two extra bytes are required to align the IP header on a 4 byte boundary for incoming data.
 - » Default number of clusters is 80.
 - » END network drivers lend all their clusters.
 - » Clusters = numRds + numTds + NUM_LOAN
 - Where numRds (32) is the number of receive descriptors
 - Where numTds (64) is the number of transmit descriptors
 - NUM_LOAN (16) is the number of loan buffers
 - mBlks = 4(numRds + NUM_LOAN)
 - Currently in the field for SNS06a and SNS06c we have:
 - numRds = 32, numTds = 64, and NUM_LOAN = 16
 - mBlks = 192, Clusters = 112
 - Should this be increased for some Apps? If yes, we need a configuration parameter in the "other" field.
 - Driver Pool for END drivers are configured in:
 - \$(WIND_BASE)/target/config/<bsp>/configNet.h









Basic Definitions (Cont'd)



Network Stack Queues

- Queues are used to hold data waiting to be processed
 - » Queues are implemented as a linked-list.
 - » Clusters are chained the to the queue's linked list

• Types of Queues:

- » Receive Queues:
 - IP PROTOCOL RECEIVE QUEUE
 - FRAGMENT REASSEMBLY QUEUE
 - ARP RECEIVE QUEUE
 - TCP REASSEMBLY QUEUE
 - SOCKET RECEIVE QUEUES
- » Send Queues:
 - SOCKET SEND QUEUES
 - IP NETWORK INTERFACE SEND QUEUES





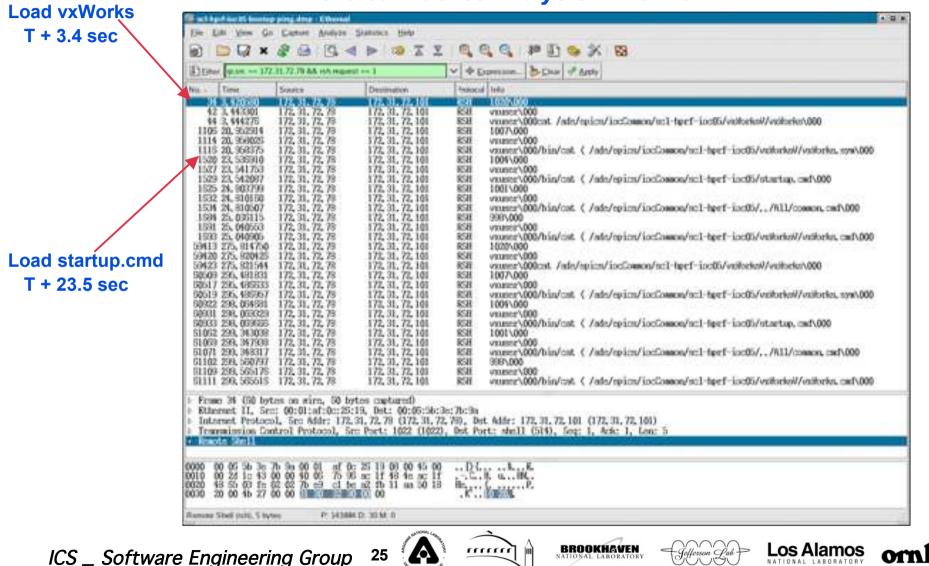






UUGU

Ethereal Packet Analysis Timeline

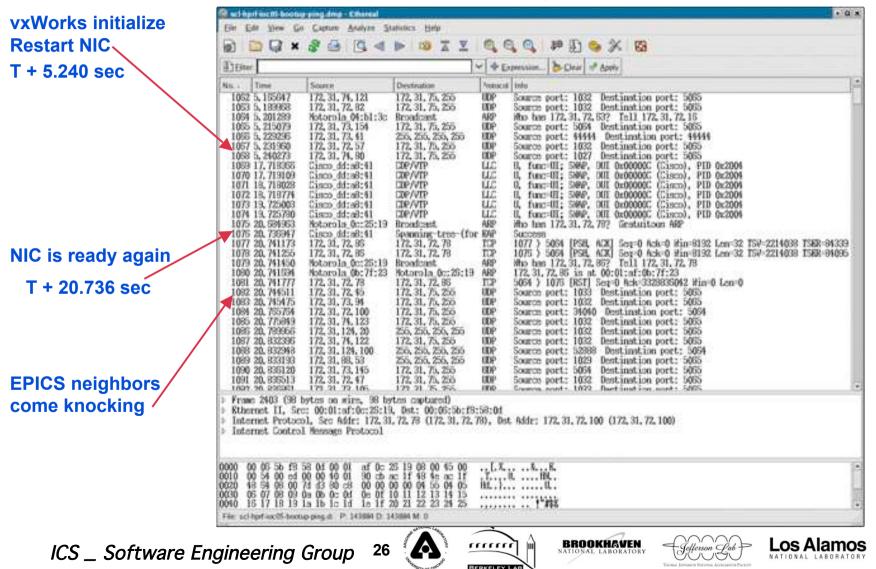


BERKELEY LAB



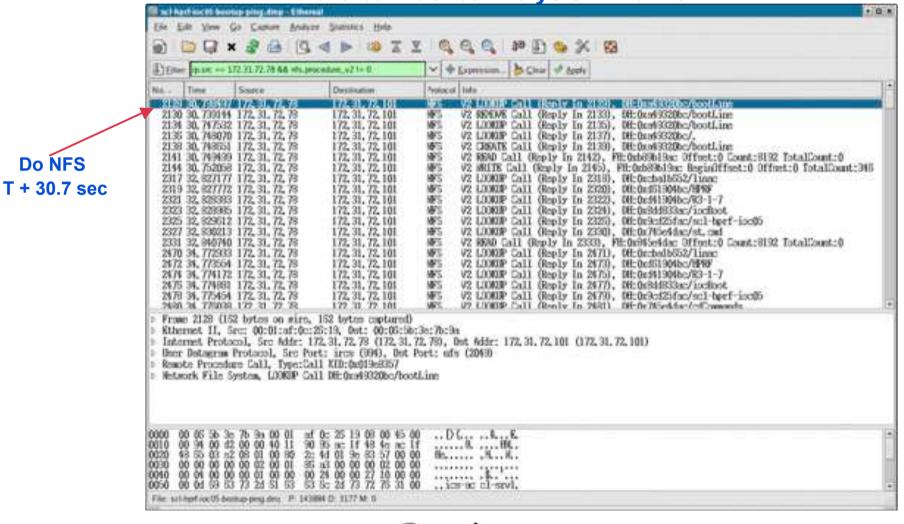
om

Ethereal Packet Analysis Timeline





Ethereal Packet Analysis Timeline



ICS _ *Software Engineering Group*

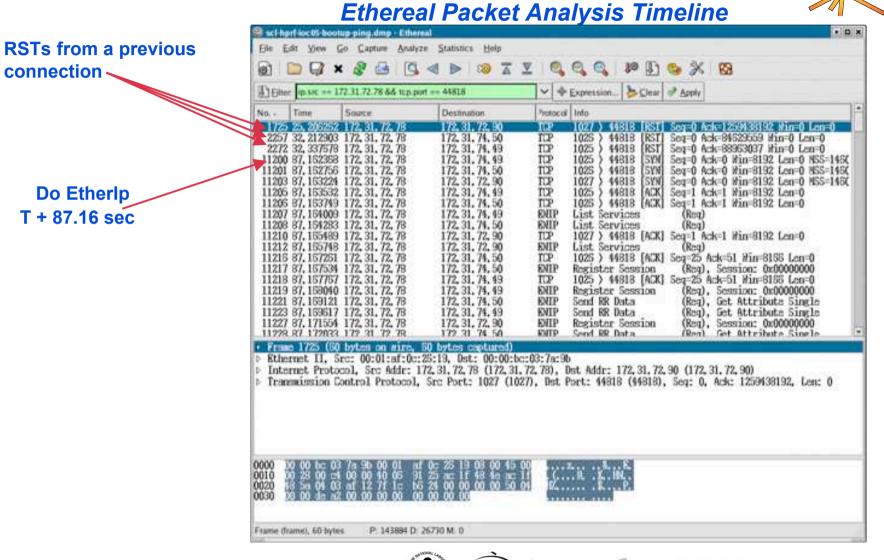






Los Alamos

om



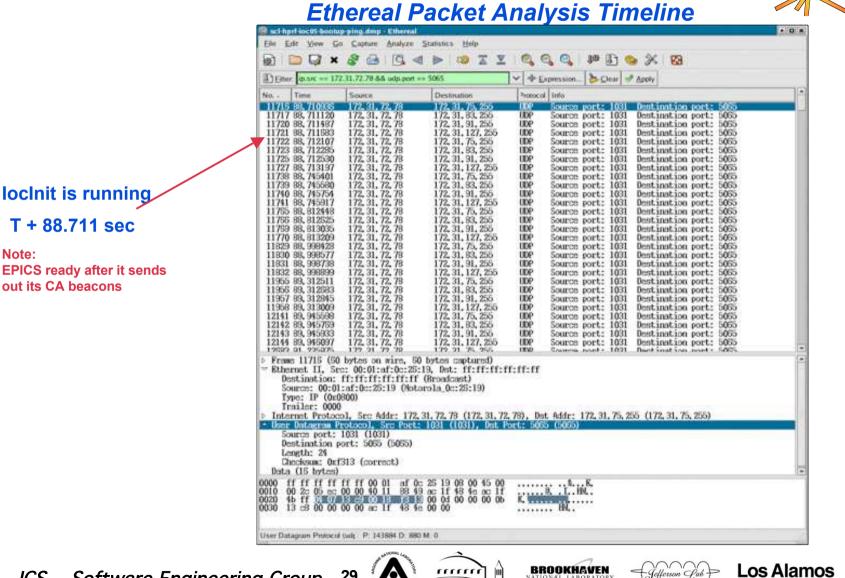
ICS _ *Software Engineering Group*











ICS _ Software Engineering Group

rrrrrr BERKELEY LAB

29

BROOKHAVEN

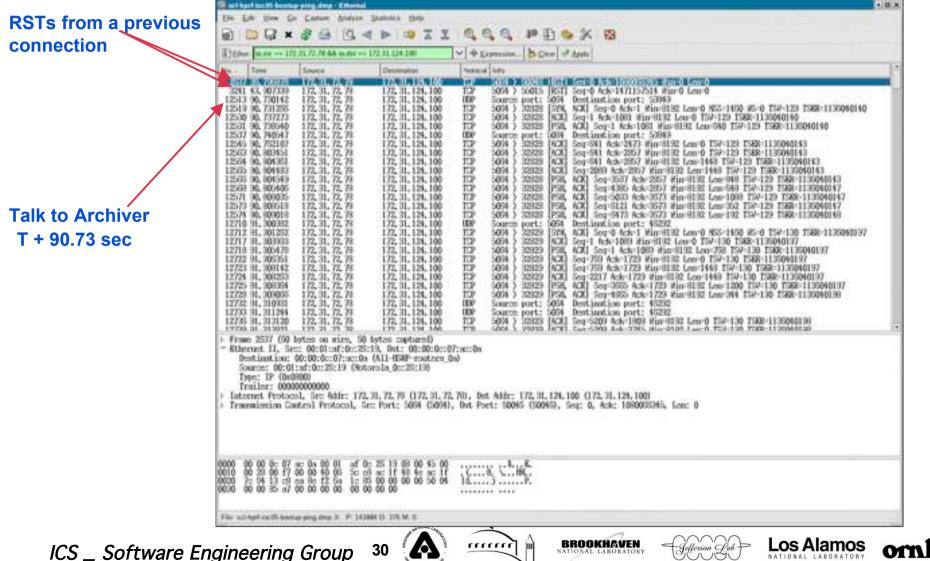
Jefferson Pab

UUGL

om



Ethereal Packet Analysis Timeline



BERKELEY LAB