INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/dac.872

Potential performance bottleneck in Linux TCP

Wenji Wu*,† and Matt Crawford

Fermilab, MS-368, P.O. Box 500, Batavia, IL 60510, U.S.A.

SUMMARY

Transmission control protocol (TCP) is the most widely used transport protocol on the Internet today. Over the years, especially recently, due to requirements of high bandwidth transmission, various approaches have been proposed to improve TCP performance. The Linux 2.6 kernel is now preemptible. It can be interrupted mid-task, making the system more responsive and interactive. However, we have noticed that Linux kernel preemption can interact badly with the performance of the networking subsystem. In this paper, we investigate the performance bottleneck in Linux TCP. We systematically describe the trip of a TCP packet from its ingress into a Linux network end system to its final delivery to the application; we study the performance bottleneck in Linux TCP through mathematical modelling and practical experiments; finally, we propose and test one possible solution to resolve this performance bottleneck in Linux TCP. Copyright © 2007 John Wiley & Sons, Ltd.

Received 28 April 2006; Revised 15 November 2006; Accepted 21 November 2006

KEY WORDS: Linux; TCP; networking; process scheduling; performance analysis; protocol stack

1. INTRODUCTION

The transmission control protocol (TCP) is the most widely used transport protocol on the Internet today. It carries the vast majority of all traffic over the Internet for various network applications, including end-user applications (such as web browsing, remote login, and e-mail), bandwidth-intensive applications (such as GridFTP for bulk data transmission [1]) and high-performance distributed computing [2, 3]. TCP has been and will continue to be an evolving protocol. Over the years, various TCP flavours have been implemented. Early TCP implementations used a go-back-N model and required the expiration of a retransmission timer to recover any loss. TCP Tahoe added the slow start, congestion avoidance, and fast retransmit algorithms to TCP [4]. Based on TCP Tahoe, TCP Reno added fast recovery

*Correspondence to: Wenji Wu, Fermilab, MS-368, P.O. Box 500, Batavia, IL 60510, U.S.A.
†E-mail: wenji@fnal.gov

Contract/grant sponsor: U.S. Department of Energy; contract/grant number: DE-AC02-76CH03000

algorithm, first implemented in 1990 [5]. TCP SACK allows receivers to selective ACK out of sequence data and it aimed at eliminating the timeouts that arise in TCP Reno due to multiple losses from the same window [6, 7]. TCP Vegas [8] is another implementation of TCP, which adjusts transmission rate according to anticipated congestion. It employs a new retransmission mechanism and slow start mechanism from Reno. TCP Westwood [9, 10] is proposed to handle random or sporadic losses. It continuously measures at the TCP source the rate of the connection by monitoring the rate of returning ACKs. The estimate is then used to compute congestion window and slow start threshold after a congestion episode. Recently, due to requirements for high bandwidth transmission, TCP variants, such as FAST TCP [11], BIC TCP [12], HTCP [13], and HSTCP [14], are also proposed and implemented.

To improve TCP performance, researchers have been primarily working in the fields of TCP flow control [15], TCP congestion control [4–14, 16–18], and TCP offloading [19, 20].

Linux-based network end systems have been widely deployed in the high-energy physics (HEP) community, at laboratories and universities. At Fermilab, thousands of network end systems are running Linux operating systems; these include computational farms, trigger processing farms, servers, and desktop workstations. From a network performance perspective, Linux represents an opportunity since it is amenable to optimization due to its open source support and projects such as web100 and net100 that enable tuning of network stack parameters [21, 22].

In all previous versions of Linux, the kernel itself cannot be interrupted while it is processing. Linux 2.6 is preemptible [23, 24]. The 2.6 kernel can be interrupted mid-task, so that the system is more responsive and interactive. However, we note that preemption in certain sections incurs some serious negative effects on networking performance. In this paper, we investigate these problems. Our analysis is based on Linux kernel 2.6.14. The contribution of the paper is as follows: (1) we systematically describe the trip of a TCP packet from its ingress into a Linux end system to its final delivery to the application; (2) we point out the performance bottleneck in Linux TCP from both mathematical analysis and practical experiments; and (3) we propose and test one possible solution to resolve the performance bottleneck in Linux TCP.

The remainder of the paper is organized as follows: in Section 2, the Linux packet receiving process is presented. Section 3 analyses the performance bottleneck in Linux TCP. In Section 4, we show the experiment results to further study the Linux TCP performance issues, verifying our conclusions in Section 3. In Section 5, we propose a potential solution to resolve the performance bottleneck in Linux TCP. And finally in Section 6, we conclude the paper.

2. TCP PACKET RECEIVING PROCESS

The Layer 2 technology is assumed Ethernet network media, since it is the most widespread and representative LAN technology. Also, it is assumed that the Ethernet device driver makes use of Linux's 'New API,' or NAPI [25, 26], which reduces the interrupt load on the central processing units (CPUs).

Figure 1 demonstrates generally the trip of a TCP packet from its ingress into a Linux end system to its final delivery to the application [25, 27, 28]. We will not generally observe the distinctions among datalink frames, IP packets, and TCP segments, as the data structures moved along protocol stack in the Linux kernel represent any of these things at different times, and the data remain at the same memory location. For simplicity, we use the single term 'packet'

DOI: 10.1002/dac

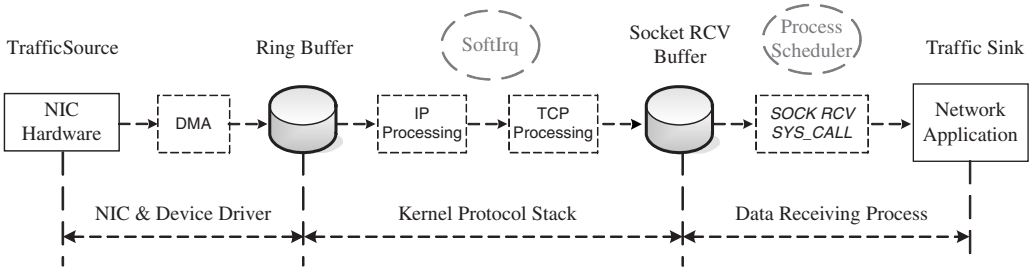


Figure 1. Linux networking subsystem: TCP packets receiving process.

wherever it will not cause confusion. In general, the packet's trip can be classified into three stages:

- Packet is transferred from network interface card (NIC) to ring buffer. The NIC and device driver manage and controls this process.
- Packet is transferred from ring buffer to a socket receive buffer, driven by a software interrupt request (softirq) [23, 27]. The kernel protocol stack handles this stage.
- Packet data are copied from the socket receive buffer to the application, which we will term the *data receiving process*.

The following subsections detail these three stages.

2.1. NIC and device driver processing

The NIC and its device driver perform the layer 1 and 2 functions of the OSI 7-layer network model: packets (datalink frames) are received and transformed from raw physical signals, and placed into system memory, ready for higher layer processing. The Linux kernel uses a structure *sk_buff* [25, 27] to hold any single packet up to the MTU (maximum transfer unit) of the network. The device driver maintains a 'ring' of these packet buffers, known as a 'ring buffer,' for packet reception (and a separate ring for transmission). A ring buffer consists of a device- and driver-dependent number of packet descriptors. To be able to receive a packet, a packet descriptor should be in 'ready' state, which means it has been initialized and preallocated with an empty *sk_buff* that has been memory mapped into address space accessible by the NIC over the system I/O bus. When a packet comes, one of the ready packet descriptors in the reception ring will be used, the packet will be transferred by direct memory access (DMA) [29] into the preallocated *sk_buff*, and the descriptor will be marked as used. A used packet descriptor should be reinitialized and refilled with an empty *sk_buff* as soon as possible for further incoming packets. If a packet arrives and there is no ready packet descriptor in the reception ring, it will be discarded. Once a packet is transferred into the main memory, during subsequent processing in the network stack, the packet remains at the same kernel memory location.

Figure 2 shows a general packet receiving process at NIC and device driver level. When a packet is received, it is transferred into main memory and an interrupt is raised only after the packet is accessible to the kernel. When CPU responds to the interrupt, the driver's *interrupt handler* is called, within which the *softirq* is scheduled by calling *netif_rx_schedule( )*. It puts a

DOI: 10.1002/dac

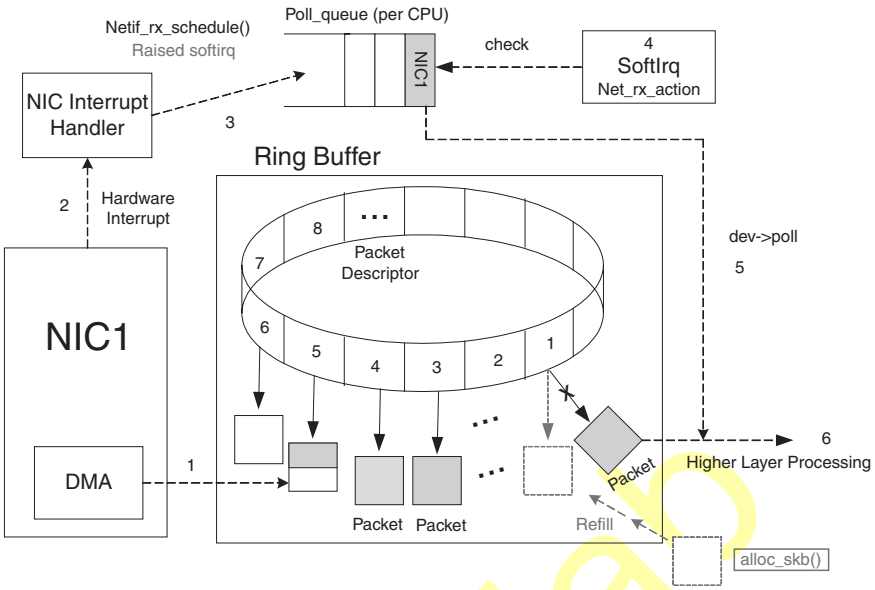


Figure 2. NIC and device driver packet receiving.

reference to the *device* into the *poll queue* of the interrupted CPU. The interrupt handler also disables the NIC's receive interrupt until all packets in its ring buffer are processed.

The softirq is serviced shortly afterward. The CPU polls each *device* in its *poll queue* to get the received packets from the ring buffer by calling the *poll* method *dev->poll( )* of the *device driver*. In *dev->poll( )*, each received packet is passed upwards for further processing by *net_receive_skb( )*. After a received packet is dequeued from its receiving ring buffer for further processing, its corresponding packet descriptor in the reception ring buffer needs to be reinitialized and refilled.

2.2. Kernel protocol stack

2.2.1. IP processing. The IP protocol receive function *ip_rcv( )* gets called from within *net_receive_skb( )* during the processing of a softirq, whenever an IP packet is dequeued from its receiving ring buffer. This function performs all the initial checks on the IP packet, which mainly involve verifying its integrity (IP checksum, IP header fields, and minimum packet length). If the packet looks correct and passes the *netfilter* hook, *ip_rcv_finish( )* is called. *ip_rcv_finish( )* deals with the routing functionality of IP. It checks whether the packet should be forwarded to another machine or if it is destined to the local host. In the latter case, the packet is given to *ip_local_deliver( )*. In case the packet is fragmented, IP fragment reassembly is performed here. Then, the packet passes another *netfilter* hook, and finally goes to the *ip_local_deliver_finish( )* function. There, an IP packet undergoes the last stage of IP-level processing: IP header data are trimmed and the higher layer protocol is determined so that the packet is ready for transport ('layer 4') processing. For each transport layer protocol, a corresponding entry handler function is defined: *tcp_v4_rcv( )* and *udp_rcv( )* are two examples. When a packet is passed upwards, the corresponding protocol entry handler function is called.

DOI: 10.1002/dac

2.2.2. TCP processing. When a packet (TCP segment) is handed upwards for TCP processing, the function *tcp_v4_rcv( )* first performs the TCP header processing. Then *__tcp_v4_lookup( )* is called to find the corresponding *socket* that is associated with the packet. A packet without a corresponding *socket* will be dropped. A socket has a lock structure to protect it from un-synchronized access. If the socket is locked, the packet waits on the backlog queue before being processed further. If the socket is not locked, and its *data receiving process* is sleeping for data, the packet is added to the socket's *prequeue* and will be processed in batch in the *process context*, instead of the *interrupt context* [23]. Placing the first packet in the prequeue will wake up the sleeping data receiving process. If the prequeue mechanism does not accept the packet, which means that the socket is not locked and no process is waiting for input on it, the packet must be processed immediately by a call to *tcp_v4_do_rcv( )*. The same function also is called to drain the backlog queue and prequeue. Except in the case of prequeue overflow, those queues are drained in the *process context*, not the *interrupt context* of the softirq. In the case of prequeue overflow, which means that packets within the prequeue reach or exceed the socket's receive buffer quota, those packets should be processed as soon as possible, even in the *interrupt context*.

tcp_v4_do_rcv( ) in turn calls other functions for actual TCP processing, depending on the TCP state of the connection. If the connection is in the *tcp_established* state, *tcp_rcv_established( )* is called; otherwise, *tcp_rcv_state_process( )* is called to handle state transitions and connection management, if there are no header or checksum errors. *tcp_rcv_established( )* performs key TCP actions such as sequence number checking, RTT estimation, acknowledging, and data packet processing. Here, we focus on the data packet processing.

When a data packet is handled on the *fast path*, *tcp_rcv_established( )* checks whether it can be delivered to the user space directly, instead of being added to the *receive queue*. The data's destination in user space is indicated by an *iovec* structure provided to the kernel by the *data receiving process* through a system call such as *recvmsg( )*. The conditions for checking whether to deliver the data packet to the user space are as follows:

- the socket belongs to the currently active process; AND
- the current packet is the next in sequence for the socket; AND
- the packet will entirely fit into the application-supplied memory location.

When a data packet is handled on the *slow path* it will be checked whether the data are in sequence (packet is the next one to be delivered to the user). Similar to the *fast path*, an in-sequence packet will be copied to user space if possible; otherwise, it is added to the *receive queue*. Out of sequence packets are added to the socket's *out-of-sequence queue* and an appropriate TCP response is scheduled. Unlike the *backlog queue*, *prequeue* and *out-of-sequence queue*, packets in the *receive queue* are guaranteed to be in order, already acknowledged, and contain no holes. Packets in out-of-sequence queue would be moved to *receive queue* when incoming packets fill the preceding holes in the data stream. Figure 3 shows the TCP processing flow chart within the *interrupt context*. In the figure, 'A' and 'B' stand for measurement points that will be referred to in later sections.

As previously mentioned, the *backlog* and *prequeue* are generally drained in the *process context*. The socket's *data receiving process* obtains data from the socket through socket-related receive system calls. For TCP, all such system calls eventually lead to *tcp_recvmsg( )*, which is the top end of the TCP transport receive mechanism. As shown in Figure 4, *tcp_recvmsg( )* first

DOI: 10.1002/dac

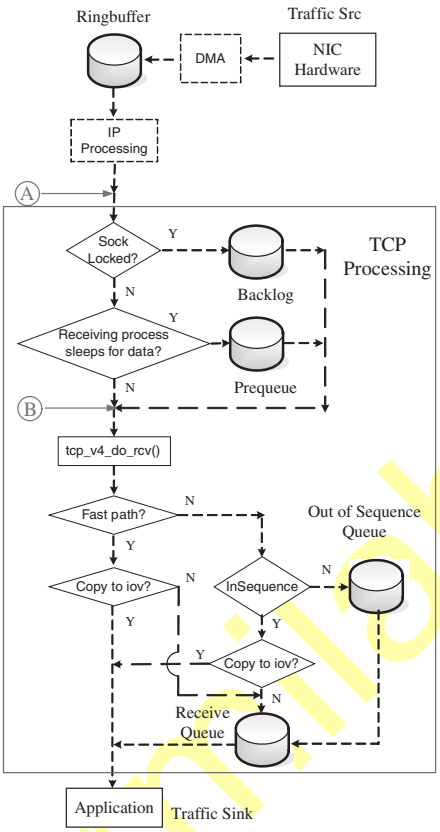


Figure 3. TCP processing—interrupt context.

locks the socket, then checks the *receive queue*. Since packets in the receive queue are guaranteed in order, acknowledged, and without holes, data in *receive queue* are copied to user space directly. After that, *tcp_recvmsg( )* will process the *prequeue* and *backlog queue*, respectively, if they are not empty. Both result in the calling of *tcp_v4_do_rcv( )*. Afterward, processing similar to that in the *interrupt context* is performed. *tcp_recvmsg( )* does not return to user space until the *prequeue* and *backlog queue* are drained. *tcp_recvmsg( )* may need to fetch a certain amount of data before it returns to user code; if the required amount is not present, *sk_wait_data( )* will be called to put the data receiving process to sleep, waiting for new data to come. The amount of data is set by the data receiving process. Before *tcp_recvmsg( )* returns to user space or the data receiving process is put to sleep, the lock on the socket will be released. As shown in Figure 4, when the data receiving process wakes up from the sleep state, it needs to relock the socket again.

2.3. Data receiving process

Packet data are finally copied from the socket's receive buffer to user space by *data receiving process* through socket-related receive system calls. The receiving process supplies a memory

DOI: 10.1002/dac

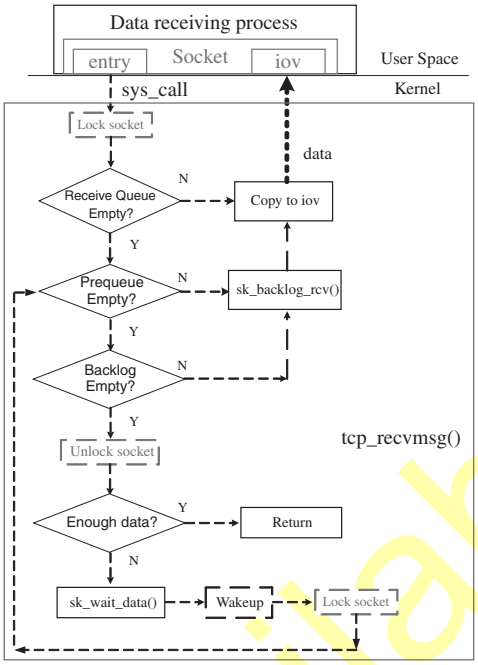


Figure 4. TCP processing—process context.

address and number of bytes to be transferred, either in a *struct iovec*, or as two parameters gathered into such a struct by the kernel. As mentioned in Section 2.2.2, all the TCP socket-related receive system calls result in a call to *tcp_recvmsg( )*, which will copy packet data from socket's buffers (*receive queue*, *prequeue*, *backlog queue*) through *iovec*.

3. POTENTIAL PERFORMANCE BOTTLENECK FOR TCP

As described above, TCP processing can be performed in interrupt or process context, depending on the status of the TCP receive socket and the data receiving process. To summarize, the different TCP packet processing scenarios are as shown in Figure 5. Since Linux is an interrupt-driven operating system, interrupt processing has higher priority than user-lever processes. TCP packets handled in the interrupt context are usually processed immediately by TCP protocol engine, independent of system load. However, when incoming TCP packets are on the prequeue or backlog queue, those packets will be handled in the process context. In that case, TCP processing is strongly related to system load and the Linux process-scheduling scheme. This leads to a potential performance bottleneck for TCP applications when the system is under load.

Linux 2.6 is a *preemptive multi-processing* operating system. Processes (tasks) are scheduled to run in a *prioritized round robin* manner [23, 24, 30], to achieve the objectives of fairness, interactivity, and efficiency. For the sake of scheduling, a Linux process has a dynamic priority and a static priority. A process' static priority is equivalent to its *nice* value, which is specified by

DOI: 10.1002/dac

Process \ Socket	locked	unlocked
sleep (by *sk_wait_data()*)	N/A	Wait on prequeue, (process context)
run	Wait on backlog, (process context)	Process immediately (interrupt context)

Figure 5. TCP packets processing scenarios.

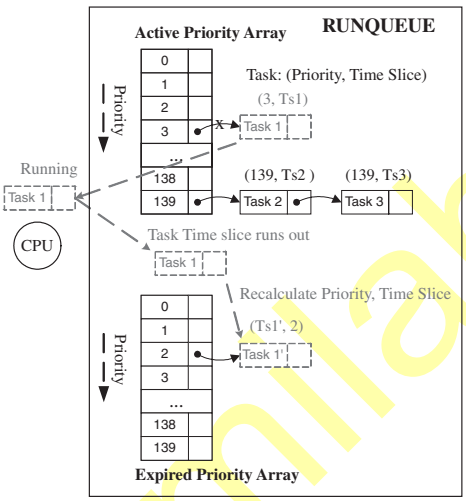


Figure 6. Linux process scheduling.

the user in the range −20 to +19 with a default of zero, and not changed by the kernel [23, 24]. Higher values correspond to lower priorities. The dynamic priority is used by the scheduler to rate the process with respect to the other processes in the system. An eligible process with a better (smaller-valued) dynamic priority is scheduled to run before a process with a worse (higher-valued) dynamic priority. The dynamic priority varies during a process' life. It depends on a dynamic priority bonus, from −5 to +5, and its static priority. The dynamic priority bonus depends on the process' interactivity status. Linux credits interactive processes and penalizes non-interactive processes by adjusting this bonus. There are 140 possible priority levels in Linux. The top 100 levels (0–99) are used only for real-time processes, which we do not address in this paper. The last 40 levels (100–139) are used for conventional processes.

As shown in Figure 6, process scheduling employs a data structure called *runqueue*. Essentially, a runqueue keeps track of all runnable tasks assigned to a particular CPU. One runqueue is created and maintained for each CPU in a system. Each runqueue contains two priority arrays: *active priority array* and *expired priority array*. Each priority array contains a queue of runnable processes per priority level. Higher (dynamic) priority processes are scheduled to run first. Within a given priority, processes are scheduled round robin. All tasks on a CPU begin in the active priority array. Each process' *time slice* is calculated based on its nice value. Table I shows the time slices for various nice values. When a process in the active priority array runs out of its time slice, it is considered *expired* and moved to the expired priority array if

DOI: 10.1002/dac

Table I. Nice value *vs* time slice.

Nice value	Time slice (ms)
+19	5
0	100
−10	600
−15	700
−20	800

it is not interactive, or reinserted back into the active array if it is interactive. During the move, a new time slice and dynamic priority are calculated. When there are no more runnable tasks in the active priority array, it is simply swapped with the expired priority array. A running process might be put into a wait queue to sleep, waiting for expected events (e.g. I/O). When a sleeping process wakes up, its time slice and priority are recalculated and it is moved to the active priority array. As for preemption, whenever a scheduler clock tick or interrupt occurs, if a higher-priority task has become runnable, it will preempt the running task if the latter holds no kernel locks.

Furthermore, when a process is termed interactive, its time slice is divided into smaller pieces. When it finishes a piece, the task will round robin with other tasks at the same priority level. This way execution will rotate more frequently among interactive tasks of the same priority, preventing interactive processes from blocking other interactive processes of the same priority.

Under the simplifying assumption that processes other than the data receiving process of interest do not sleep for long times, and hence use their entire time slices before the active priority array is empty, let us first consider the backlog scenario. The data receiving process is calling *tcp_recvmsg()* to fetch packet data from socket receive buffer to user space. The socket will be locked until the process returns to the user space. If the data receiving process' time slice ends before the lock is released, the data receiving process will be moved to the expired priority array with the socket locked. The socket will remain locked until the lock is released after the data receiving process resumes its execution in the next round of running. The time until the process resumes its execution is strongly dependent on the system load. Let us assume that when the expired data receiving process is moved to the expired priority array, there are, in all, N_1 running processes $(P_1, \ldots, P_{N_1})$ left in the active array, and N_2 expired processes $(P_{N_1+1}, \ldots, P_{N_1+N_2})$ in the expired array with priorities higher than that of the expired data receiving process. Considering that some of the N_1 processes, when expired, might move to the expired array with recalculated priorities higher than that of the expired data receiving process, the minimum time before the data receiving process could resume its execution would be

$$\sum_{j=1}^{N_1+N_2} \text{Timeslice}(P_j) \tag{1}$$

Here, Timeslice(P_j) denotes the time slice of process P_j.

As we have seen, during this period all the new incoming TCP packets for the data receiving process will wait on the socket's backlog queue without being TCP processed. No acknowledgements will be fed back to the TCP sender.

As for the prequeue scenario, the data receiving process might sleep within *tcp_recvmsg()* waiting for data. Before the process wakes up, all the incoming segments for the data receiving

DOI: 10.1002/dac

process will wait on the prequeue without being TCP-processed. Let us assume that when the woken-up data receiving process is moved to the active priority array, there are N_3 other processes in the active array whose priorities are higher. Assuming still that each process will use its full time slice, the minimum time before the data receiving process could resume its execution would be

$$\sum_{j=1}^{N_3} \text{Timeslice}(P_j) \tag{2}$$

Similarly, during this period no acknowledgements will be returned to the TCP sender.

Note that the data receiving process might be preempted within *tcp_recvmsg()* by other higher priority processes. In this case, packet might wait on the backlog queue. The analysis of how long packets would wait on the backlog queue is similar to the above.

Let us denote by T_{wait} the time a packet waits in the backlog queue or prequeue. In the worst case, we have

$$T_{\text{wait}} > \sum_{j=1}^{N_1+N_2} \text{Timeslice}(P_j) \quad \text{for backlog queue} \tag{3a}$$

$$T_{\text{wait}} > \sum_{j=1}^{N_3} \text{Timeslice}(P_j) \quad \text{for prequeue} \tag{3b}$$

The time that packets wait on the backlog queue or prequeue adds to the sender's estimate of the round-trip time (RTT), since ACKs have not been sent for segments on those queues.

Usually it is the case that

$$\text{RTT} = T_{\text{t}} + T_{\text{pd}} + T_{\text{q}} + T_{\text{pp}} \tag{4}$$

where T_{t} is the time the interface takes to send the packet. This will likely have a linear dependence on packet size. T_{pd} is the time taken for a signal to propagate through the network medium. In a single simple network link, this is equal to the distance between sender and receiver divided by propagation speed. T_{q} is the time spent in routing queues along the path. This will fluctuate over time depending on congestion in the path. And T_{pp} is the amount of time spent by sender and receiver and routers doing processing necessary for packet delivery. On current hardware, this is usually in the sub-microsecond range. For a given packet size, network path, sender, and receiver, it can be assumed that T_{t}, T_{pd}, and T_{pp} are constants. For packet switched data networks, T_{q} is usually a random variable, following some distribution. Hence, RTT is treated as a random variable.

Since TCP packets might wait on the backlog queue or prequeue in the receiver, we will have

$$\text{RTT} = T_{\text{t}} + T_{\text{pd}} + T_{\text{q}} + T_{\text{pp}} + T_{\text{wait}} \tag{5}$$

Clearly, if TCP packets are processed in interrupt context, $T_{\text{wait}} \approx 0$. In the receiver, since the system load is uncertain, whether, when, and how long TCP packets might wait on the backlog queue or prequeue is uncertain, and T_{wait} is also a random variable, following some distribution. We can see that T_{wait} and T_{q} are independent or effectively so.

DOI: 10.1002/dac

Hence, it will be the case that

$$E(T_t + T_{pd} + T_q + T_{pp} + T_{wait}) > E(T_t + T_{pd} + T_q + T_{pp}) \tag{6}$$

$$\mathrm{Var}(T_t + T_{pd} + T_q + T_{pp} + T_{wait}) > \mathrm{Var}(T_t + T_{pd} + T_q + T_{pp}) \tag{7}$$

Here, $E(*)$ and $\mathrm{Var}(*)$ are the expected value and variance of a random variable, respectively. From (6) and (7), it can be derived that when packets wait on backlog queue or prequeue, both RTT and its variance will increase.

In [31], Mathis *et al.*, derive

$$\mathrm{BW} = \frac{\mathrm{MSS}}{\mathrm{RTT}} \frac{C}{\sqrt{p}} \tag{8}$$

where BW is the achievable bandwidth from sender to receiver, MSS is the maximum segment size, C is a constant of order unity, and p is the packet drop probability along the path. Note: (8) is based on Reno TCP.

It follows from (6) and (8) that with increased RTT, the average achievable bandwidth from sender to receiver will decrease. Also, as we know, when competing TCP network traffic exists, increased RTT will put a TCP data stream in a disadvantaged position [32].

The TCP sender does not measure RTT precisely, but rather maintains 'smoothed' estimates of SRTT and RTTVAR of RTT and its variation, and uses the estimates in determining the RTO, or retransmit timeout period [33, 34], after which an unacknowledged packet is assumed lost and will be retransmitted. Estimates are updated as follows whenever a new measurement R of RTT is available from the acknowledgement of a timed data segment:

$$\mathrm{RTTVAR} \coloneqq \tfrac{3}{4}\mathrm{RTTVAR} + \tfrac{1}{4}|\mathrm{SRRT} - R| \tag{9}$$

$$\mathrm{SRTT} \coloneqq \tfrac{7}{8}\mathrm{SRTT} + \tfrac{1}{8}R \tag{10}$$

$$\mathrm{RTO} \coloneqq \max\{\mathrm{TCP\_RTO\_MIN}, \min\{\mathrm{SRTT} + 4 \times \mathrm{RTTVAR}, \mathrm{TCP\_RTO\_MAX}\}\} \tag{11}$$

The variation defined by (9) is not variance in the strict statistical sense, but is more easily calculated in the kernel and is commonly referred to as variance. Here, both TCP_RTO_MIN and TCP_RTO_MAX are constants, which are 200 ms and 120 s respectively. Experience has shown that clock granularity affects RTO calculation. Finer granularity ($\leqslant$100 ms) performs somewhat better than coarser granularities [34]. In Linux 2.6, the clock granularity is not a big concern, since it has reached the 1 ms level.

Also, from (6), (7), and (11), it can be seen that when the RTT's variance rises, the RTO in the sender will correspondingly increase. In that case, the TCP sender may be less responsive to packet losses, resulting in degraded performance.

When packets wait on the receiver's backlog queue or prequeue too long, it triggers the retransmission timeout in the sender. Assuming that when packets start to wait on the queue, the current RTO value in the sender is T_{RTO}. The sender will time out when

$$T_{wait} > T_{RTO} - T_t - T_{pd} - T_q - T_{pp} \tag{12}$$

If the system load is medium or high, condition (12) can be easily met. For example, if $N_1 + N_2 \geqslant 10$, and all the running processes have the default nice value of 0, from Equations (1)

DOI: 10.1002/dac

and (3a) and Table I, we can easily have $T_{wait} > 1$ s, large enough to outrigger the retransmission timer. Once RTO times out, the sender incorrectly assumes packet loss in the network. Such spurious timeouts affect TCP performance in two ways: (1) the sender unnecessarily reduces its offered load by setting its congestion window size to 1 segment and (2) the sender is forced into a go-back-N retransmission model. Spurious timeouts are usually due to sudden delay spike in the path, e.g. route changes. The Eifel algorithm [35] and F-RTO algorithm [36] have been proposed to solve the spurious timeout problem in the sender. However, since the spurious timeout in the case at hand is caused by Linux TCP implementation in the receiver, we seek a solution in the receiver.

4. EXPERIMENTS AND ANALYSIS

To verify our claims in Section 3, we ran data transmission experiments upon Fermilab's sub-networks. In the experiments, we run *iperf* [37] to send data in one direction between two computer systems. *iperf* in the receiver is the data receiving process. As shown in Figure 7, the sender and the receiver are connected to two Cisco 6509 switches connected to each other by an uncongested 10-Gbit/s link. During the experiments, the background traffic in the network is low, and there is no packet loss, or packet reordering in the network. The sender and receiver's features are as shown in Table II.

In order to study the detailed packet receiving process, we have added instrumentation within Linux packet receiving path. Specifically, to collect statistics and provide insights for T_{wait}, we have added measurement points A and B in Linux packet receiving path as shown in Figure 3. For each packet, the times that it passes over point A (t^A) and point B (t^B) are recorded; we collect the statistics for the time difference $t^{diff} = t^B - t^A$. It can be assumed that $T_{wait} \approx t^{diff}$, to within a few CPU cycles. Since it is difficult to keep track of every packet, we classify t^{diff} into six different groups: $0 \leqslant t^{diff} \leqslant 1$ ms, $1\text{ ms} \leqslant t^{diff} \leqslant 10$ ms, $10\text{ ms} \leqslant t^{diff} \leqslant 0.1$ s, $0.1\text{ s} \leqslant t^{diff} \leqslant 0.2$ s, $0.2\text{ s} \leqslant t^{diff} \leqslant 1$ s, and $t^{diff} > 1$ s. We collect the histogram for each category.

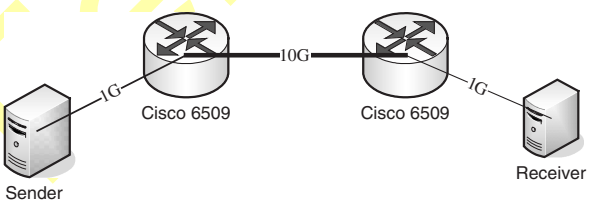


Figure 7. Experiment network and topology.

Table II. Sender's and receiver's features.

	Sender	Receiver*
CPU	Two Intel Xeon CPUs (3.0 GHz)	One Intel Pentium III CPU (1 GHz)
System memory (MB)	3829	512
NIC	Tigon, 64 bit-PCI bus slot at 66 MHz, 1 Gbps twisted pair	Syskonnect, 32 bit-PCI bus slot at 33 MHz, 1 Gbps twisted pair

*We ran experiments on different types of Linux receivers in Fermilab, and similar results were obtained.

DOI: 10.1002/dac

Table III. iperf output results.

System load in receiver	Time (s)	Data transmitted	End-to-end throughput (Mbits/s)
BL0	20	1.17 Gbytes	504
BL10	20	174 Mbytes	72.1

To create a variable system load in the receiver, we compiled the Linux Kernel with n parallel tasks by running `make` $-nj$ [23]. The different values of n corresponds to different levels of background system load, e.g. `make` $-10j$. For simplicity, they are termed as 'BLn'. The background system load implies load on both CPU and system memory. In the TCP experiments, sender transmits one TCP stream to receiver for 20 s. All the processes are running with a nice value of 0, and iperf's receive buffer is set to 40 MB. In the sender, we use *tcpdump* to record TCP streams, and later use *tcptrace* [38] to analyse the recorded packets. Consistent results were obtained across repeated runs. In the following sections, we present two groups of experiments, with background loads of BL0 and BL10, respectively. The experimental data are from both sender and receiver side. Table III shows the iperf output results in the sender.

Plate 1 shows the time-sequence diagrams for the recorded TCP traces from the sender side. Plate 1(a) shows that with a background load of BL0, the sender sends packets smoothly and continuously. Packets are acknowledged in time and no RTO happens. However, the time-sequence diagram in Plate 1(b) shows another story. With BL10 in receiver, sender sends packets intermittently. In the diagram, the small red 'R' represents retransmission. There are quite a few retransmissions. Since there are no packet losses or reordering in the network, those unnecessary retransmissions are due to spurious timeouts in the sender. As we have analysed in Section 3, when packets wait on backlog queue and prequeue too long, no ACKs are returned to the TCP sender in time, leading to RTOs in the sender.

Now let us study the issues from the receiver side, to further verify the conclusions of Section 3.

Figures 8 and 9 show various TCP queues at BL0 and BL10, respectively.

- Normally, prequeue and out-of-sequence queue are empty. The backlog queue is usually not empty, even during the periods that the data receiving process is not running. Packets are not dropped or reordered in the test network. However, packets might be dropped by the NIC in the reception ring buffer [39], causing subsequent packets to go to the out-of-sequence queue.
- With BL0, the receive queue is approaching full. In our experiment, since the sender is more powerful than the receiver, this scenario is as expected. The experiment results have confirmed this point. With BL10, since RTOs in the sender seriously degrade the TCP performance, the receive queue is not approaching full, even with a background load of BL10.
- In contrast to Figure 8, the backlog and receive queues in Figure 9 show some kind of periodicity. The periodicity matches the data receiving process' running cycle [39]. In Figure 8, with BL0, the data receiving process runs almost continuously, but at BL10, it runs only intermittently.

Though Figures 8 and 9 provide information about status of various TCP queues, they provide no clue about how long packet will wait on backlog queue or prequeue. Table IV gives the statistics about t^{diff} with BL0 and BL10. As we have known that there is $T_{\text{wait}} \approx t^{\text{diff}}$, the

DOI: 10.1002/dac

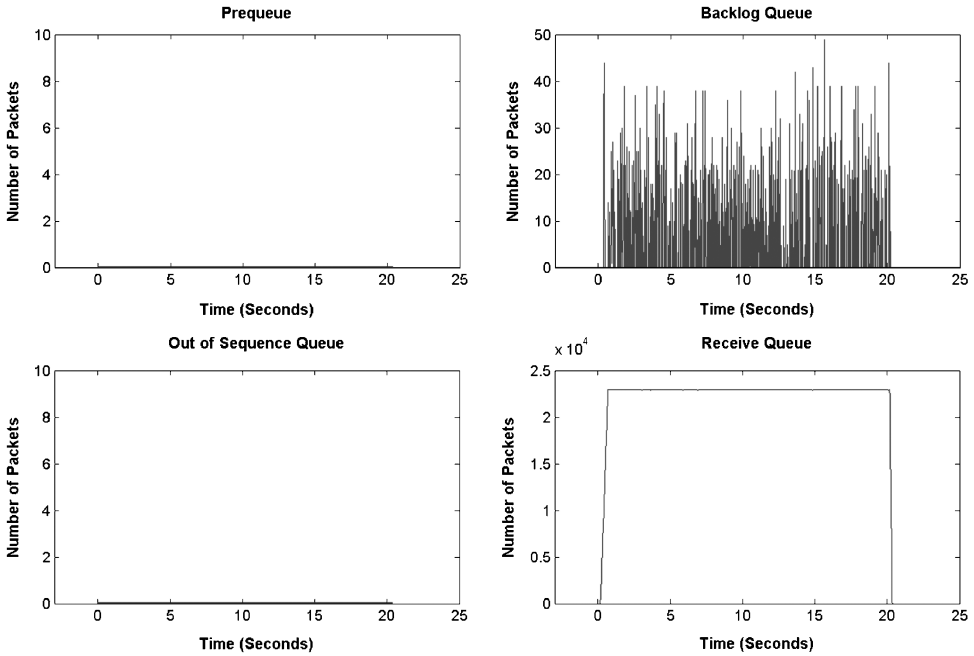


Figure 8. Various TCP receive buffer queues—BL0 (receiver side).

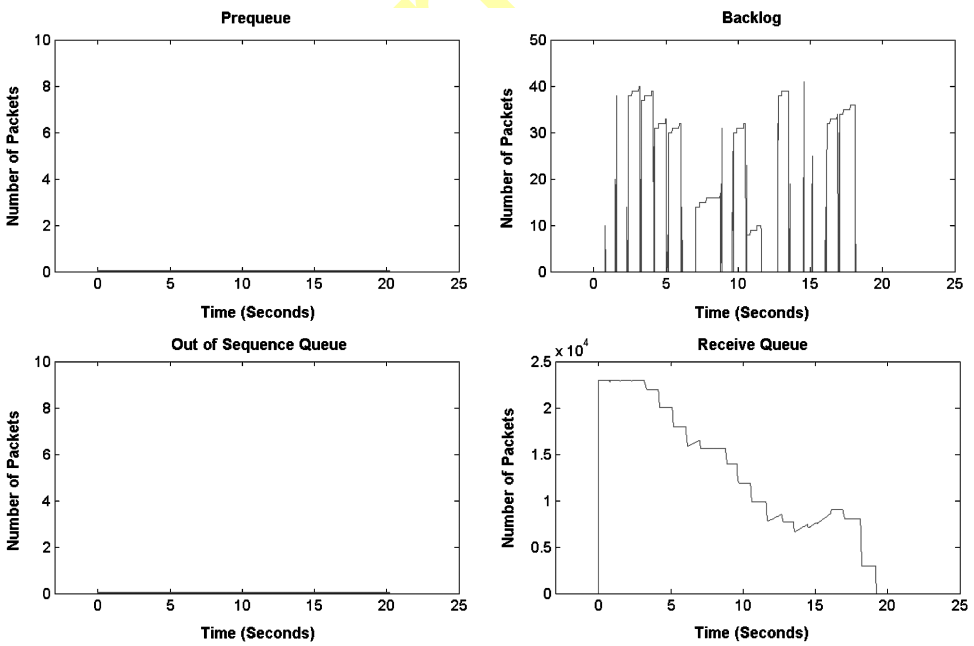


Figure 9. Various TCP receive buffer queues—BL10 (receiver side).

DOI: 10.1002/dac

Table IV. t^{diff} statistics (receiver side).

System load	<1 ms	1–10 ms	10–100 ms	100–200 ms	200 ms–1 s	>1 s
BL0	862 429	636	0	0	0	0
BL10	122 657	744	896	40	300	75

statistics about t^{diff} will provide us further information about T_{wait}. When the load is light in the receiver, packets would go to prequeue or backlog queue (as shown in backlog queue of Figure 8). Since the data receiving process runs continuously, packets within backlog queue or prequeue are processed soon, T_{wait} will not be large. However, when the system load increases, as it has been analysed in formula (3a) and (3b), T_{wait} might be large, which has been confirmed by Table IV. As shown in Table IV, with BL0, there are no packets with $t^{diff} \geqslant 10$ ms. However, with BL10, some packets even have $t^{diff} \geqslant 1$ s, waiting on the backlog queue or prequeue for quite a long period of time, without being TCP-processed. This is why we have seen in Plate 1(b) that RTO occurs.

5. A POSSIBLE SOLUTION

As described above, the TCP performance bottleneck is due to the fact that TCP packets might wait on the backlog queue or prequeue in the receiver without being TCP-processed. To resolve the performance bottleneck issue, there might be two basic approaches. Naturally, the first approach is to always do TCP processing in the interrupt context, not in the process context at all. However, this would require the overhaul of the whole Linux TCP protocol engine, which might be complex and time consuming. The second approach is to reduce T_{wait} for packets waiting on prequeue or backlog queue. As implied by formulas (1)–(3), the underlying idea here is that when there are packets waiting on the prequeue or backlog queue, do not allow the data receiving process to release the CPU for long. Relatively, the second approach is easier to implement. We have modified the Linux process scheduling policy and *tcp_recvmsg()* to implement a solution of the second sort. The pseudo-code for scheduling is as shown in Listing 1. The code changes for *tcp_recvmsg()* is as shown in Listing 2, highlighted with a lighter shade. To summarize, the solution works as follows: an expired data receiving process with packets waiting on backlog queue or prequeue is moved to the active array, instead of expired array as usual. More often than not, the expired data receiving process will continue to run. Even if it does not run, the wait time before it resumes its execution will be greatly reduced. However, this gives the process extra runs compared to other processes in the runqueue. For the sake of fairness, the process would be labelled with the *extra_run_flag*. Also considering the facts that: (1) the resumed process will continue its execution within *tcp_recvmsg()* and (2) *tcp_recvmsg()* does not return to user space until the *prequeue* and *backlog queue* are drained. For the sake of fairness, we modified *tcp_recvmsg()* as such: after *prequeue* and *backlog queue* are drained and before *tcp_recvmsg()* returns to user space, any process labelled with the *extra_run_flag* will call *yield()* to explicitly yield the CPU to other processes in the runqueue. *yield()* works by removing the process from the active array (where it currently is, because it is running), and inserting it into the expired array [23]. Also, to

DOI: 10.1002/dac

prevent processes in the expired array from starving, a special rule has been provided for Linux process scheduling (the same rule used for interactive processes [24]): an expired process is moved to the expired array without respect to its status if processes in the expired array are starved.

We repeated the TCP experiments as described in Section 4 on Linux updated with the new scheduling policy described as above. We compare the new experiment data with those obtained in Section 4. The old experiment data will be prefixed with 'O-'; whereas, the new data are prefixed with 'N-'.

Listing 1. Pseudo-code for scheduling policy.

```
If (process->timeslice - - == 0) {
        recalculate timeslice and priority;
        if (packets are waiting on backlog queue or prequeue) {
                if (processes in expired array are starved)
                        move the process to expired array;
                else {
                        move process to active array;
                        if (process is non-interactive)
                        set process->extra_run_flag:=TRUE;
                }
        }
        else {
                ... as usual ...
        }
}
else {
        ... as usual ...
}
```

Listing 2. Code changes for *tcp_recvmsg()*.

```
tcp_recvmsg{
        ... as usual ...

        TCP_CHECK_TIMER(sk);
        release_sock(sk);

        if (process->extra_run_flag == TRUE){
                set process->extra_run_flag:=FALSE;
                yield();
        }

        return copied;

        ... as usual ...
}
```


DOI: 10.1002/dac

Table V. iperf output results.

System load in receiver	Time (s)	Data transmitted	End-to-end throughput (Mbits/s)
O-BL0	20	1.17 Gbytes	504
O-BL10	20	174 Mbytes	72.1
N-BL10	20	220 Mbytes	88.8

Table V shows the iperf output results in the sender. It can be seen that the TCP performance of N-BL10 is much better than that of O-BL10. The TCP end-to-end throughput of N-BL10 reaches as high as 88.8 Mbits/s; however, with O-BL10, the corresponding TCP end-to-end throughput is only 72.1 Mbits/s. This implies that our proposed solution is effective in resolving the TCP performance bottleneck issue. This point is verified by experiment data from both sender and receiver sides. Plate 2 shows the time-sequence diagrams for the recorded TCP traces from the sender side. For comparison, Plate 2(a) shows old kernel's time-sequence diagram with a background load of BL10. And Plate 2(b) shows the time-sequence diagram with N-BL10. In Plate 2(b), there are no retransmissions due to packets waiting on backlog queue or prequeue too long; packets are acknowledged in time and no RTO happens. Nevertheless, the sender still transmits intermittently. This is caused by zero window advertisements from receiver. In our experiments, sender is a more powerful machine than the receiver; in addition, the receiver runs with a high background load. When packets cannot be consumed by the data receiving process in time, the data receive buffer in receiver is approaching full. Then receiver will advertise zero windows to stop sender transmitting. The small 'Z' in Plate 2(b) represents a window advertisement of 0 bytes received from the receiver. Later, from Figure 10 we can see that the receive buffer is approaching full.

Figure 10 shows various TCP queues with N-BL10. It can be seen that the receive queue is approaching full. Compared with Figure 9, it can be concluded that TCP performance is really enhanced. Table VI gives observations of t^{diff} for N-BL10. There are now no packets with $t^{\text{diff}} \geqslant 200$ ms, which implies that no packets waited long on the prequeue or backlog queue. It further verifies that our proposed solution is effective in resolving the TCP performance bottleneck issue.

Now, let us study the fairness issues of our proposed solution. Readers might suspect that our proposed solution might cause fairness issues. The following experiments and analysis will show that it does not significantly do so.

As described above, Linux scheduler moves an expired process to the expired priority array if the process is not interactive, or reinserts it back into the active array if the process is interactive. To better evaluate the fairness performance of our proposed solution, we try to eliminate the influence of interactive processes in the following experiments: non-interactive processes run as background loads. To create non-interactive processes in the receiver, we develop a CPU-intensive application that executes a number of operations in a loop. In all the following experiments, the sender transmits one TCP stream to the receiver for 20 s. In the receiver, iperf is run as '*time iperf -s -w 20M*'. All the processes are running with a nice value of 0. Further, since the transmission lasts 20 s, in the receiver we calculate iperf's experiment CPU share as: (stime + utime)/20 s. We compare the iperf's experiment CPU shares with its fair CPU share. If there are M background processes, iperf's fair CPU share is: $1/(M+1)$. Consistent results were obtained across repeated runs. In the following sections, we present a group of experiment results in Table VII.

DOI: 10.1002/dac

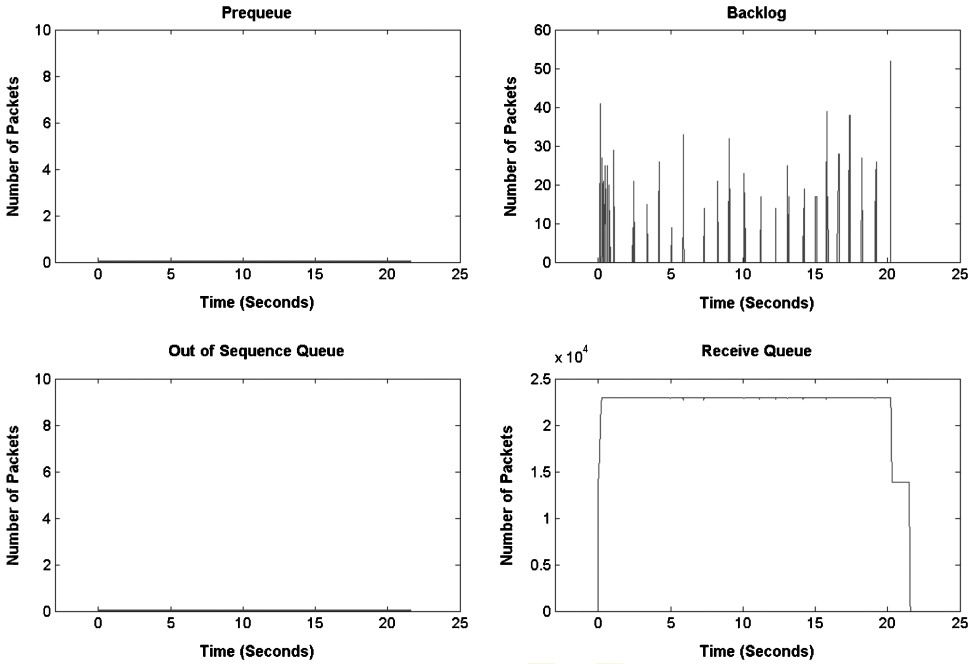


Figure 10. Various TCP receive buffer queues—BL10 (new kernel, receiver).

Table VI. t^{diff} statistics (receiver side).

System load	< 1 ms	1–10 ms	10–100 ms	100–200 ms	200 ms–1 s	> 1 s
O-BL0	862 429	636	0	0	0	0
O-BL10	122 657	744	896	40	300	75
N-BL10	156 300	851	618	29	0	0

Table VII. Fairness experiments.

Background processes in receiver (No. of processes)	Iperf experiment results				
	End-to-end throughput (Mbps)	Utime (s)	Stime (s)	Iperf experiment CPU share (%)	Iperf fair CPU share (%)
1	265	0.048	10.47	52.58	50
3	143	0.028	5.644	28.38	25
4	117	0.032	4.728	23.8	20
9	70	0.028	2.744	13.86	10

As shown in Table VII, in the worst case, iperf gains the extra CPU shares of 3.86%. Considering the possibilities that iperf itself might sometimes be termed interactive and gain extra runs, the experiment results show that our proposed solution will not cause fairness issues. The proposed solution tradeoffs a small amount of fairness performance to resolve the TCP

DOI: 10.1002/dac

performance bottleneck. The reason that our proposed solution will not cause serious fairness issues is due to the facts that:

(1) Each time when an expired data receiving process with packets waiting on backlog queue or prequeue is moved to the active array, it gains at most the *tcp_recvmsg()* amount of extra time compared to other processes in the runqueue.
(2) Each calling of *tcp_recvmsg()* will not take long. When Linux kernel processes packets within backlog queue or prequeue, the processed data will be fed to the socket out of sequence queue or receive queue, then TCP flow control will take effect to slow down or throttle sender.
(3) The possibility that a data receiving process runs out of its timeslice and is moved to the expired array with packets waiting on backlog queue or prequeue does not occur often, compared to the Linux scheduling time scale. This has been shown in Plate 1(b). As iperf does pure data transmission, the received data will not be further processed in the user space. Therefore, for a real network application, this possibility is even lower.

Another justification for our proposed solution is that 'Fairness is often a hard attribute to justify maintaining because it is often a tradeoff between over global performance and localized performance. For example, in an effort to provide maximum disk throughput, the Linux 2.4 block I/O scheduler may starve older requests, in order to continue processing newer requests at the current disk head position. This minimizes seeks and thus provides maximum overall disk throughput—at the expense of fairness to all requests' [40].

6. CONCLUSIONS

Suspension of TCP processing for incoming packets induces both an increase and a greater variability of the round-trip time measured by the sender. A moderate or high system load on the receiver can delay TCP processing so long as to cause a timeout in the sender, which will then resume sending at the minimum possible rate. Current storage implementations in the high-energy physics community and elsewhere exploit the disk space of compute-farm worker nodes [41], making this a very topical concern.

So far, we have been discussing how the proposed solution behaves in improving TCP performance, and resolving the bottleneck. Also, we evaluate the fairness performance of our proposed solution. The proposed solution trades a small amount of fairness performance to resolve the TCP performance bottleneck. Our experiments and analysis have shown that our proposed solution will not cause serious fairness issues. The criteria for a good scheduling algorithm also include efficiency, response time, turnaround, and throughput [42]. How our first cut at a throughput-enhancing process scheduling policy affects other processes and overall system performance needs further study. We will cover this topic in other papers.

REFERENCES

1. Allcock W *et al*. The globus striped GridFTP framework and server. *Proceedings of the ACM/IEEE Supercomputing 2005 Conference*, Seattle, U.S.A., 2005.
2. Foster I *et al*. *The Grid*: *Blueprint for a New Computing Infrastructure* (2nd edn). Wiley: New York, 2004.

DOI: 10.1002/dac

3. Berman F *et al*. *Grid Computing*: *Making the Global Infrastructure a Reality*. Wiley: New York, 2003.
4. Jacobson V. Congestion avoidance and control. *Proceedings of the ACM SIGCOMM*, Stanford, CA, August 1988; 314–329.
5. Fall K *et al*. Simulation-based comparison of Tahoe, Reno and SACK TCP. *ACM Computer Communications Review* 1996; **5**(3):5–21.
6. Braden R *et al*. *TCP Extensions for Long Delay Paths*. RFC 1072, 1988.
7. Braden R *et al*. *TCP Extensions for High-Speed Paths*. RFC 1185, October 1990.
8. Brakmo LS *et al*. TCP Vegas: end to end congestion avoidance on a global Internet. *IEEE Journal on Selected Areas in Communications* 1995; **13**(8):1465–1480.
9. Casetti C *et al*. TCP Westwood: bandwidth estimation for enhanced transport over wireless links. *Proceedings of ACM Mobicom 2001*, Rome, Italy, 2001; 287–297.
10. Gerla M *et al*. TCP Westwood: congestion window control using bandwidth estimation. *Proceedings of IEEE Globecom 2001*, vol. 3, San Antonio, U.S.A., 2001; 1698–1702.
11. Jin C *et al*. FAST TCP: from theory to experiments. *IEEE Network* 2005; **19**(1):4–11.
12. Xu L *et al*. Binary increase congestion control for fast long-distance networks. *Proceedings of IEEE INFOCOM 2004*, Hong Kong, 2004.
13. Leith DJ *et al*. H-TCP: a framework for congestion control in high-speed and long-distance networks. *Hamilton Institute Technical Report*, August 2005.
14. Floyd S. *HighSpeed TCP for Large Congestion Windows*. RFC 3649, December 2003.
15. Gardner MK *et al*. User-space auto-tuning for TCP flow control in computational grids. *Computer Communications* 2004; **27**(14):1364–1374.
16. Widmer J *et al*. A survey on TCP-friendly congestion control. *IEEE Network Magazine*, *Special Issue on Control of Best Effort Traffic* 2001; **15**(3):28–37.
17. Martin J *et al*. Delay-based congestion avoidance for TCP. *IEEE/ACM Transactions on Networking* 2003; **11**(3).
18. Mathis M *et al*. Forward acknowledgment: refining TCP congestion control. *Proceedings of SIGCOMM'96*, Stanford, CA, August 1996.
19. Freimuth D *et al*. Server network scalability and TCP offload. *Proceedings of the 2005 USENIX Annual Technical Conference*, Anaheim, CA, April 2005; 209–222.
20. Clark DD *et al*. An analysis of TCP processing overheads. *IEEE Communication Magazine* 1989; **27**(2):23–29.
21. Mathis M *et al*. Web100: extended TCP instrumentation for research, education and diagnosis. *ACM Computer Communications Review* 2003; **33**(3).
22. Dunigan T *et al*. A TCP tuning Daemon. *Proceedings of the ACM/IEEE Supercomputing 2002 Conference*, Baltimore, U.S.A., 2002.
23. Love R. *Linux Kernel Development* (2nd edn). Novell Press: Indianapolis, IN, U.S.A., 2005. ISBN: 0672327201.
24. Bovet DP *et al*. *Understanding the Linux Kernel* (3rd edn). O'Reilly Press: Sebastopol, CA, 2005. ISBN: 0-596-00565-2.
25. Rio M *et al*. *A Map of the Networking Code in Linux Kernel 2.4.20*. March 2004.
26. Mogul JC *et al*. Eliminating receive livelock in an interrupt-driven kernel. *ACM Transactions on Computer Systems* 1997; **15**(3):217–252.
27. Wehrle K *et al*. *The Linux Networking Architecture—Design and Implementation of Network Protocols in the Linux Kernel*. Prentice-Hall: Englewood Cliffs, NJ, 2005. ISBN 0-13-177720-3.
28. www.kernel.org.
29. Corbet J *et al*. *Linux Device Drivers* (3rd edn). O'Reilly Press: Sebastopol, CA, 2005. ISBN: 0-596-00590-3.
30. Rodriguez CS *et al*. *The Linux(R) Kernel Primer*: *A Top-Down Approach for x86 and PowerPC Architectures*. Prentice-Hall, PTR: Englewood Cliffs, NJ, 2005. ISBN: 0131181637.
31. Mathis M *et al*. The macroscopic behavior of the congestion avoidance algorithm. *Computer Communications Review* 1997; **27**(3).
32. Henderson TH *et al*. On improving the fairness of TCP congestion avoidance. *IEEE Globecomm Conference*, Sydney, 1998; 539–544.
33. Sarolahti P *et al*. Congestion control in Linux TCP. *Proceedings of 2002 USENIX Annual Technical Conference*, Freenix Track, Monterey, CA, June 2002; 49–62.
34. Paxson V *et al*. *Computing TCP's Retransmission Timer*. Internet RFC 2988, November 2000.
35. Ludwig R *et al*. The Eifel algorithm: making TCP robust against spurious retransmissions. *ACM Computer Communications Review* 2000; **30**(1).
36. Sarolahti P *et al*. F-RTO: an enhanced recovery algorithm for TCP retransmission timeouts. *Computer Communication Review* 2003; **33**(2):51–63.
37. http://dast.nlanr.net/Projects/Iperf/
38. http://www.tcptrace.org/
39. Wu W *et al*. The performance analysis of Linux networking–packet receiving. *Computer Communications*, in press. DOI: 10.1016/j.comcom.2006.11.001.
40. Love R. Interactive kernel performance—kernel performance in desktop and real-time applications. *Proceedings of the Linux Symposium*, Ottawa, Ont., Canada, July 2003.

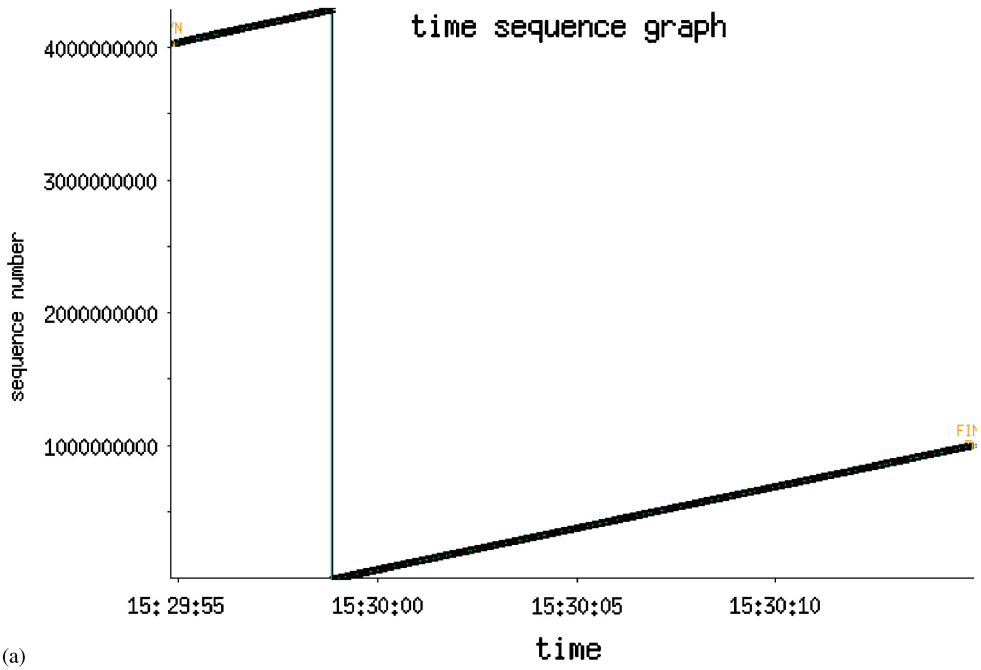
41. Kulyavtsev A *et al*. Resilient dCache: replicating files for integrity and availability. *Proceedings of Computing in High Energy Physics* (*CHEP*), Mumbai, India, 2006.
42. Silberschatc A *et al*. *Operating System Concepts* (7th edn). Wiley: New York, 2004. ISBN: 0471694665.

AUTHORS' BIOGRAPHIES

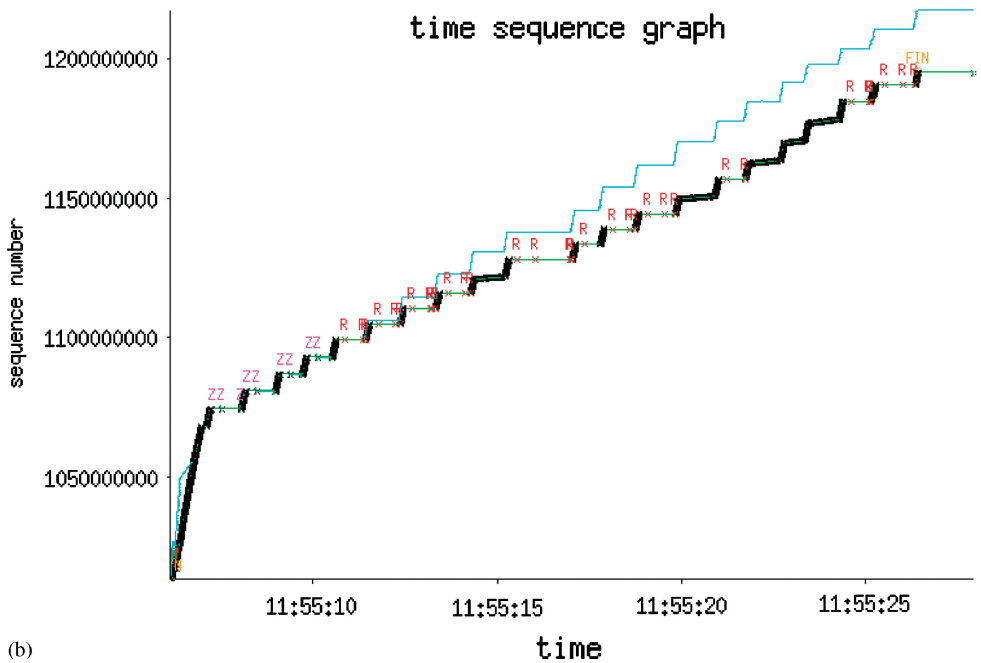
Wenji Wu holds a BA degree in Electrical Engineering (1994) from Zhejiang University (Hangzhou, China), and doctorate in computer engineering (2003) from the University of Arizona, Tucson, U.S.A. Dr Wu is currently a Network Researcher in Fermi National Accelerator Laboratory. His research interests include high performance networking, optical networking, and network modeling & simulation.

Matt Crawford is head of the Data Movement and Storage department in Fermilab's Computing Division. He holds a bachelor's degree in Applied Mathematics and Physics from Caltech and a doctorate in Physics from the University of Chicago. He currently manages the Lambda Station project, and his professional interests lie in the areas of scalable data movement and access.

DOI: 10.1002/dac



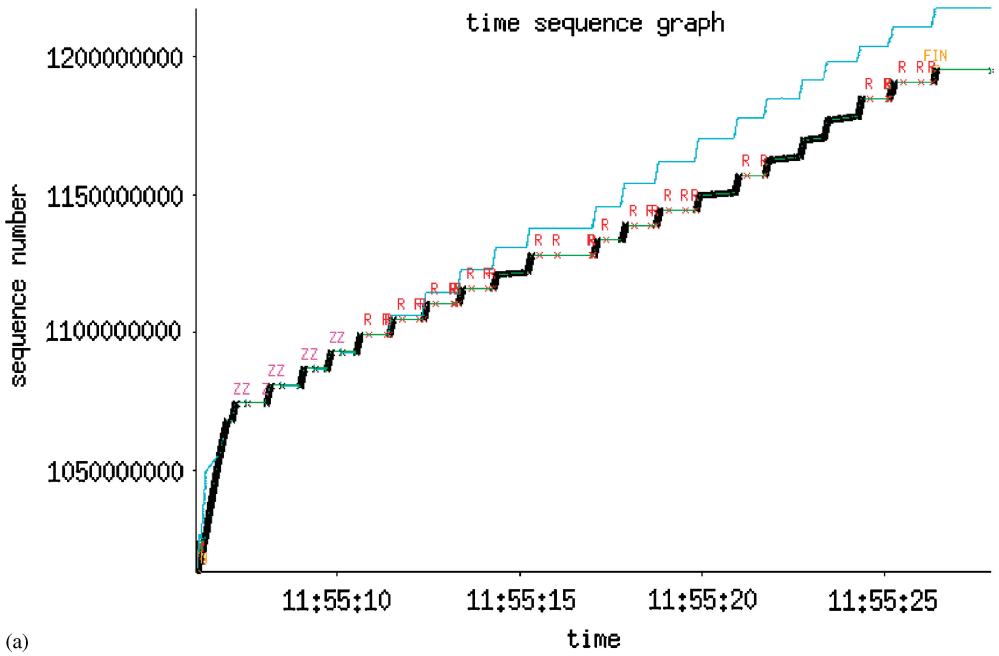
(a)



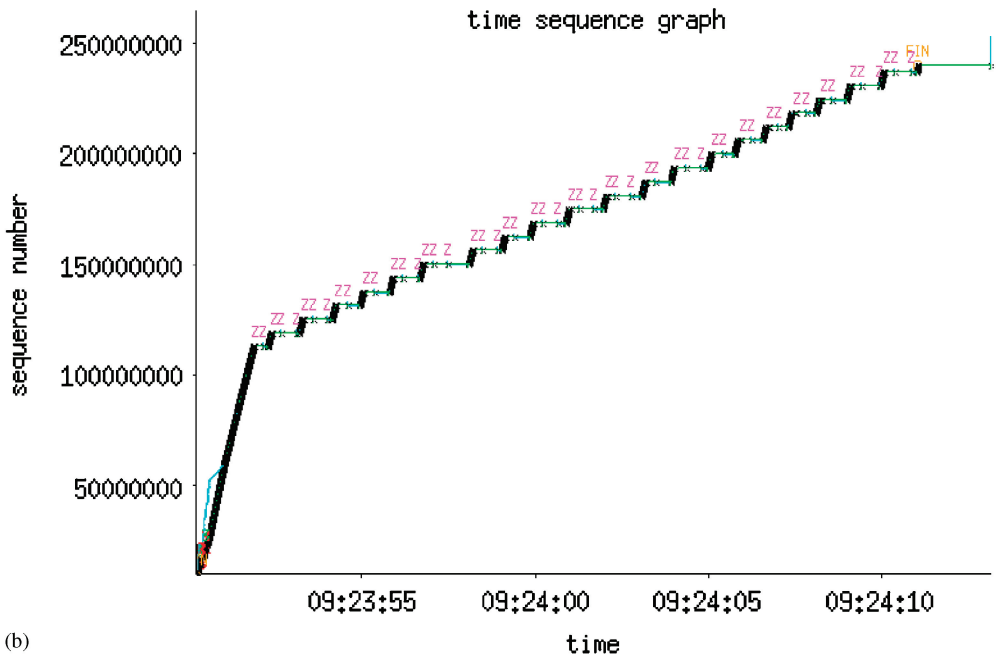
(b)

Plate 1. Sender time-sequence diagram (sender side): (a) with BL0 in receiver and (b) with BL10 in receiver.

DOI: 10.1002/dac



(a)



(b)

Plate 2. Sender time-sequence diagram (sender side): (a) with BL10 in receiver (old kernel) and (b) with BL10 in receiver (new kernel).

DOI: 10.1002/dac