

10/7 CS240 - LiveLock

Announcements

For next class (Thursday 10/9)

1. Read: [Memory Resource Management in VMware ESX Server](#)
2. Submit answers to reading questions (see course schedule) before class

Lab 1 available now on today's course schedule entry

Write-up: <http://cs240.stanford.edu/labs/lab1>

Due: Sunday, Nov 2, 2025 end of day (11:59:59 PM PDT)

Implement a user-level threads package on the x86_64 architecture

Paper background

- Digital Equipment Corporation Western Research Lab (DEC WRL)
- Usenix Technology Conference 1996
- Example of hardware technology driven OS research

I/O device interface into computer (CPU & memory)

- Interrupts
 - Interrupt priorities
 - Livelock
- Polling
 - Tradeoffs with interrupts
- Direct Memory Access (DMA)
- Programmed Input/Output (PIO)
- Interactions with the OS scheduler

First workstation NICs: 3COM 3C501 EtherLink

- Used on the early Sun workstation and early personal computers
- One receive and one transmit buffer on NIC
 - Programmed I/O - Needed to copy packet in or out of the NIC
 - Receive buffer corrupted if not copied out by the time the next packet arrives
- Unix network:
 - Assign high priority interrupt to copy packet in or out of NIC
 - Do the rest of packet processing at lower interrupt level (software interrupts)

AMD LANCE Am7990

Local Area Network Controller for Ethernet

Used on the many workstations (e.g Sun-3, DECstation as in paper)

- Ring buffers pointing to packets in memory
- DMA
- Ownership flag notation
- Some scatter/gather capability

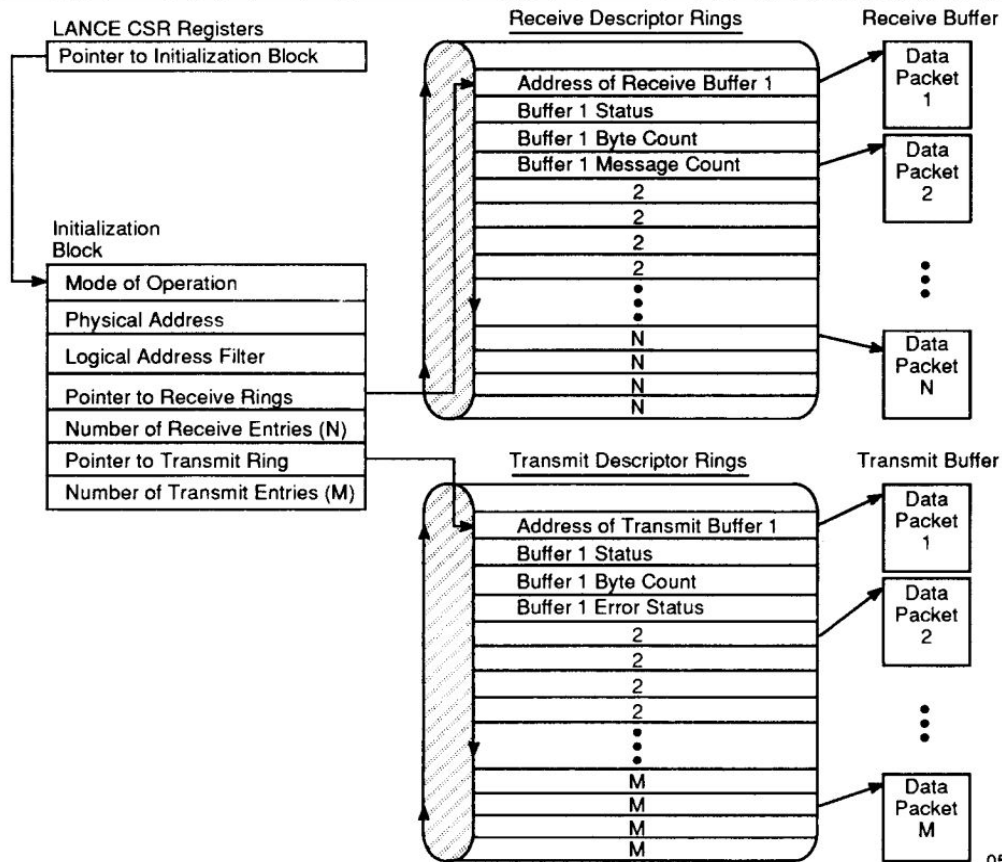


Figure 2-2. LANCE Memory Management

Driver programming error

11:00	BCNT	BUFFER BYTE COUNT is the length of the buffer pointed to by this descriptor, expressed as a two's complement number. This field is written by the host and is not changed by the LANCE. Minimum buffer size is 64 bytes for the first buffer of packet.
-------	------	--

What is Flow Control and why does it matter here?

- Transmission Control Protocol on Internet Protocol (TCP/IP)

vs

- Host-based routing
 - e.g. firewalls
- Network monitoring
 - Promiscuous mode
- Network File System (NFS)
 - User Datagram Protocol (UDP)

Networking Metrics

- Throughput
- Latency
- Jitter
- Fairness
- Stability

Maximum Lost Free Receive Rate (MLFRR)

Problems

- Receive livelock
- Increased latency
- Starvation of transmit
- Others?

Experiment

- Two machines
 - DECstation 3000/300 "router-under-test" machine
 - DECstation 3000/400 load generator
 - Modern PC about 500x-750x faster
- Workload
 - 10000 UDP packets with 4 byte payload (Section 6)
 - Minimum-sized packets (Section 7)
- Ethernet
 - $10 \text{ Mb/s} \div 672 \text{ bits per minimum packet} = 14,801 \text{ packets per second}$
Modern:
 - PC: $10 \text{ Gb/s} \div 672 \text{ bits per minimum packet} = 14,880,952 \text{ packets per second}$
 - Server: $800 \text{ Gb/s} \div 672 \text{ bits per minimum packet} = 1,190,476,190 \text{ packets per second}$

Explain what is happening here

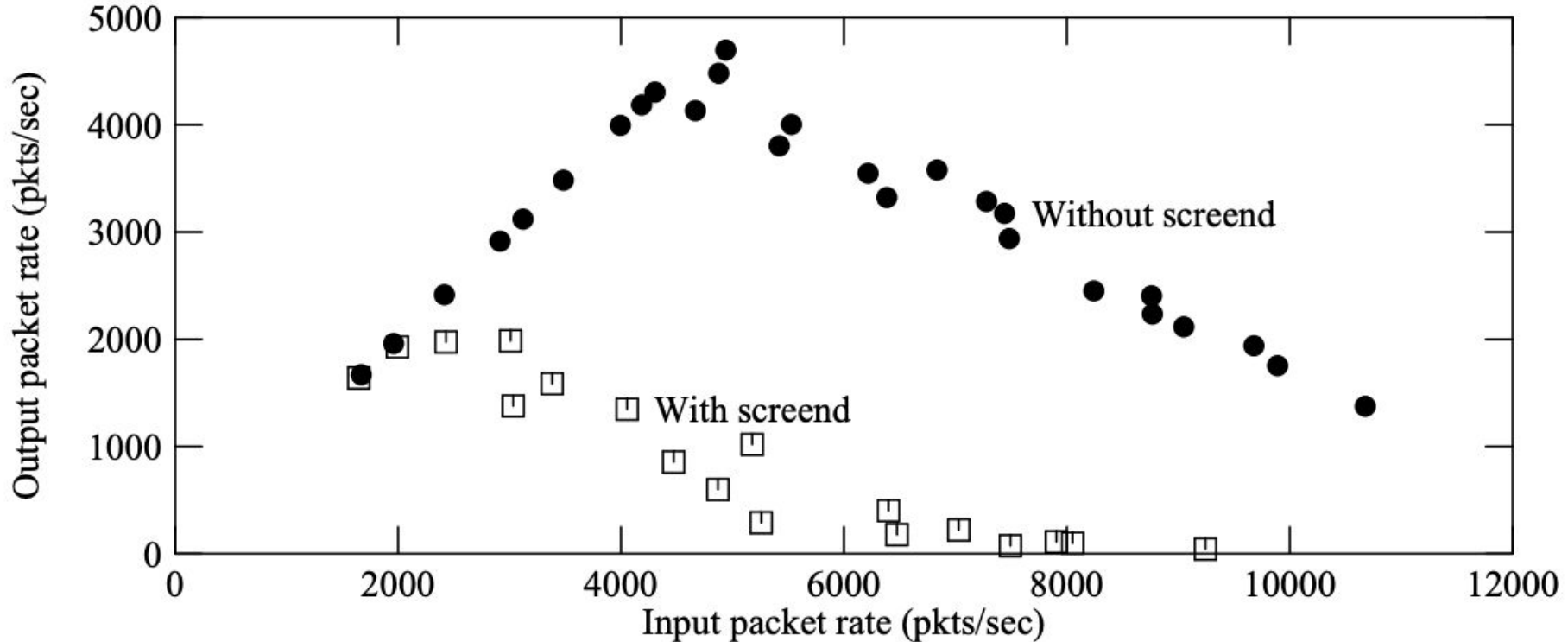


Figure 6-1: Forwarding performance of unmodified kernel

Paper solution to the livelock problem

- What was the kinda obvious approach for enqueued packets?
- How did they decide to poll or enable interrupts?
- What is the tradeoff between everything at high IPL versus nothing at IPL?
 - How is this related to Naked Notify in Mesa?

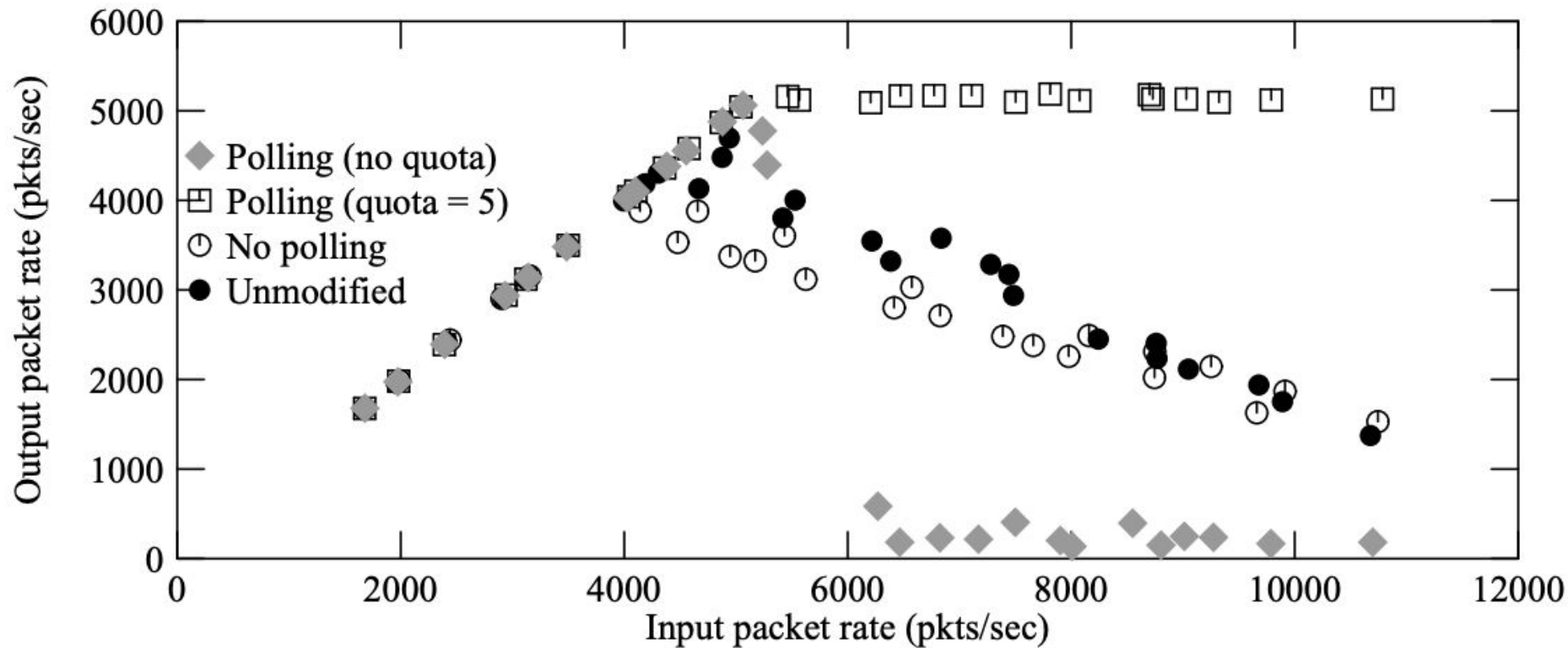


Figure 6-3: Forwarding performance of modified kernel, without using *screend*

What is feedback? What is the problem with it?

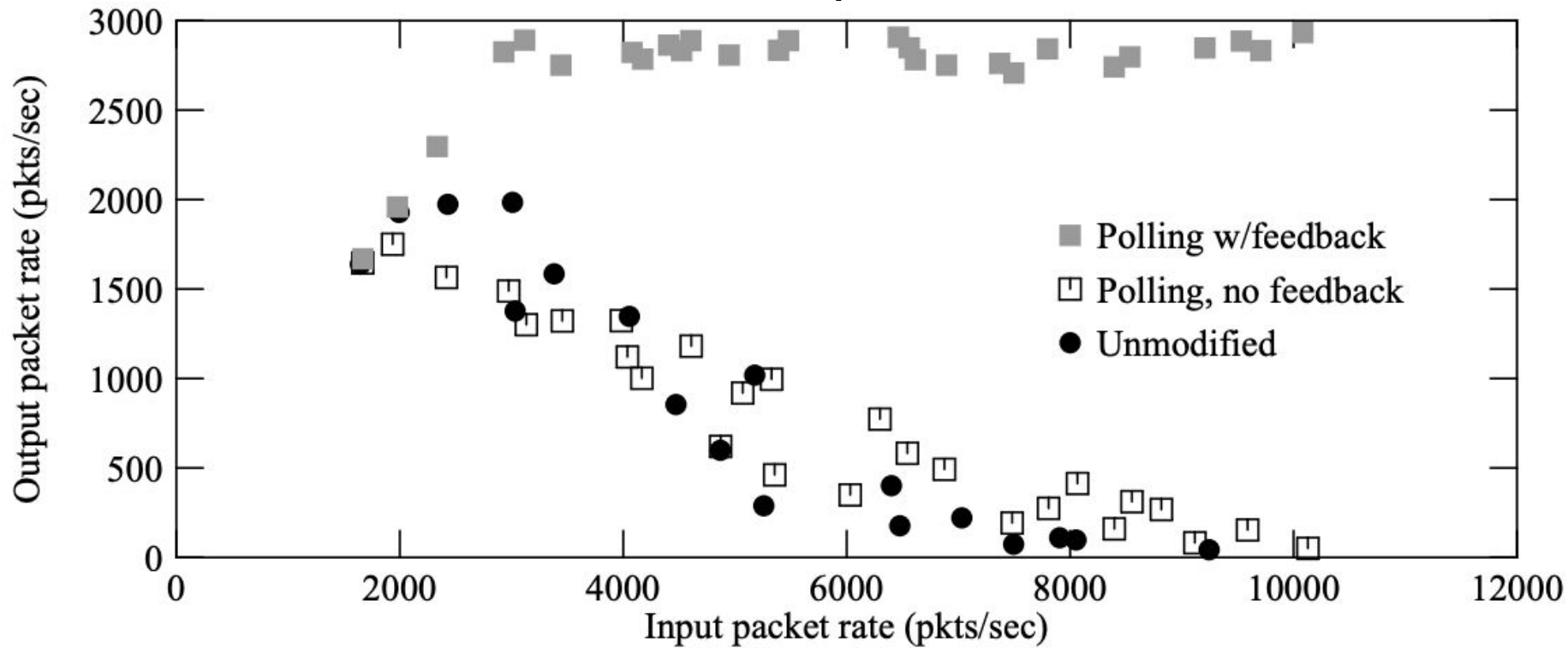


Figure 6-4: Forwarding performance of modified kernel, with *screend*

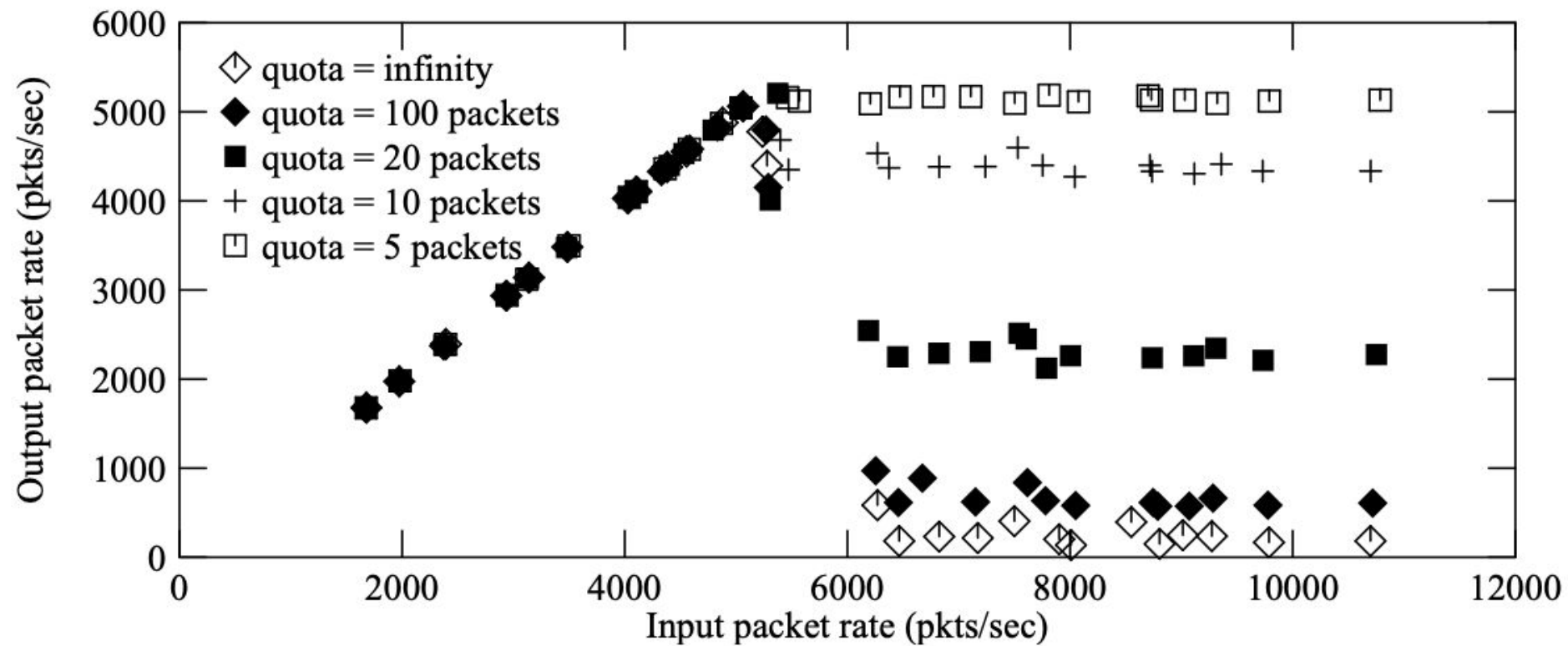


Figure 6-5: Effect of packet-count quota on performance, no *screen*

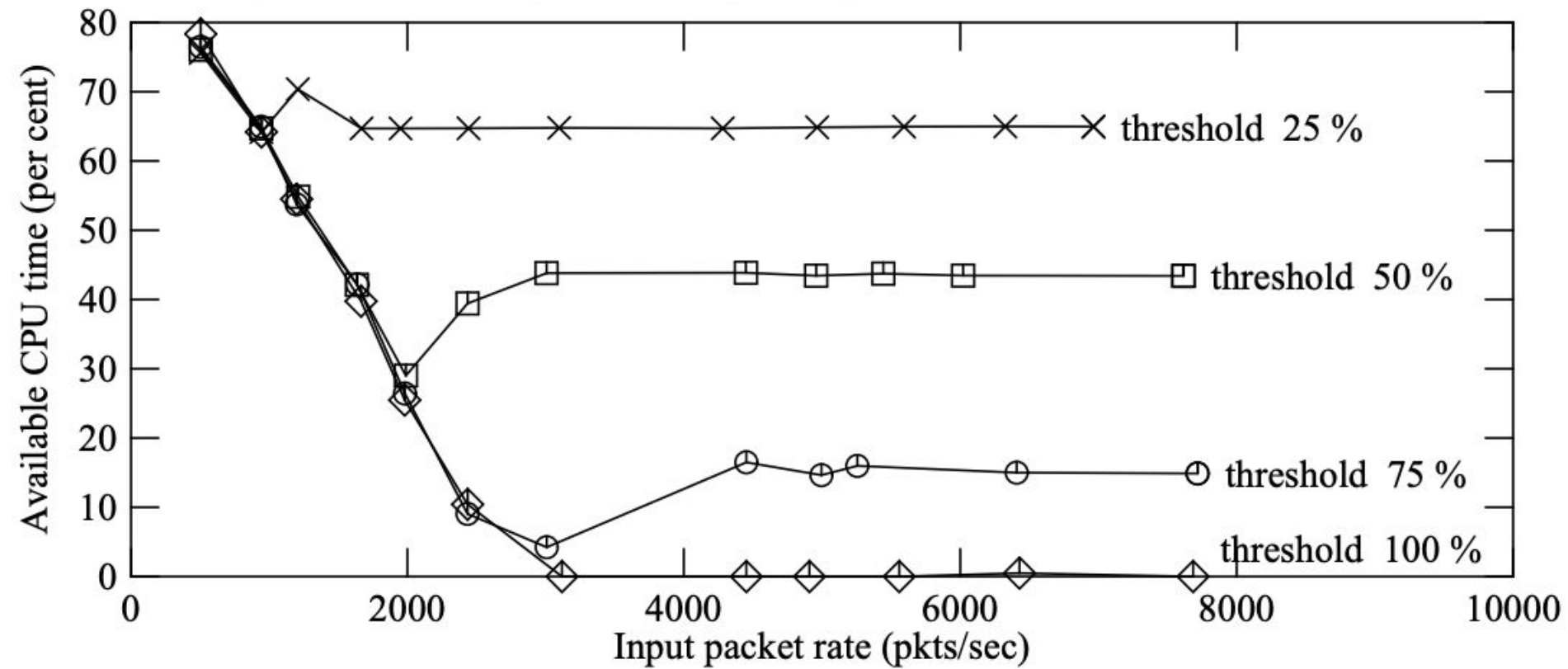


Figure 7-1: User-mode CPU time available using cycle-limit mechanism

Modern machines

- NIC interrupt processing
 - Interrupt after so many packets (batching)
 - Limit number of interrupts per second (e.g. limit to one interrupt per 80 μ s.)
- Multicore processors
 - NICs can have ring buffers per core and route traffic to particular rings
- Offload network functionality
 - Checksum calculation
 - Segmentation, etc.
- Enough to handle 10-100GbE for the paper's workload?

Reading questions

1. The [Livelock paper](#) has a hack to handle the problem of dropping packets on the "screend" queue. What limit does this solution have for multiple applications getting network data?
2. In Livelock: Figure 6-3: Why does "Polling (no quota)" drop almost immediately to zero rather than gradually decreasing similar to "unmodified?" Be very concrete.