University of California, Berkeley
College of Engineering
Computer Science Division — EECS

Fall 2007 John Kubiatowicz

Midterm II

December 3rd, 2007

CS162: Operating Systems and Systems Programming

Your Name:	
SID Number:	
Circle the letters of CS162 Login	First: a b c d e f g h I j k l m n o p q r s t u v w x y z Second: a b c d e f g h I j k l m n o p q r s t u v w x y z
Discussion Section:	

General Information:
This is a **closed book** exam. You are allowed 1 page of **hand-written** notes (both sides). You have 3 hours to complete as much of the exam as possible. Make sure to read all of the questions first, as some of the questions are substantially more time consuming.

Write all of your answers directly on this paper. *Make your answers as concise as possible.* On programming questions, we will be looking for performance as well as correctness, so think through your answers carefully. If there is something about the questions that you believe is open to interpretation, please ask us about it!

Problem	**Possible**	**Score**
1	20	
2	25	
3	20	
4	35	
Total		

[This page left for π]

3.14159265358979323846264338327950288419716939937510582097494 4

Problem 1: True/False [20 pts]

In the following, it is important that you *EXPLAIN* your answer in TWO SENTENCES OR LESS (Answers longer than this may not get credit!). Also, answers without an explanation *GET NO CREDIT*.

Problem 1a[2pts]: The Clock Algorithm requires hardware support for a "use" bit in the PTE.

True / (False)

Explain: *When a hardware use bit is not available, the PTEs can be set to invalid when they are unused, so that when the page is accessed, a trap will occur, and the operating system can set the use bit in software.*

Problem 1b[2pts]: The Aloha algorithm for broadcast networking permitted a new transmitter to interrupt a message that was already partially transmitted.

(True) / False

Explain: *Yes, because Aloha sends blindly without checking to see if there are any ongoing communications (it relies on retransmissions after detecting garbled packets).*

Problem 1c[2pts]: Memory mapped I/O devices cannot be accessed by user-level threads.

True / (False)

Explain: *Memory-mapped I/O is accomplished using load/store instructions to a special region of memory; a user-level thread can access this region if the user's page table has a mapping for it.*

Problem 1d[2pts]: It is possible to make Remote Procedure Calls (RPCs) with integer arguments between clients that use big-endian integers and servers that use little-endian integers.

(True) / False

Explain: *The marshalling/unmarshalling code of RPC will automatically convert between network and host byte orders, thus accommodating any combination of host byte orders.*

Problem 1e[2pts]: A dictionary attack could be used against a publicly readable password file that contains encrypted passwords.

(True) / False

Explain: *An attacker could encrypt every word in the dictionary and compare against the encrypted passwords stored in the publicly readable file. Note that the "encryption" used the UNIX password file acts as a one-way function.*

Problem 1f[2pts]: The rate of page faults in a virtual memory system can always be reduced by adding more memory.

True / (False)

Explain: *There are circumstances in which adding memory can actually increase the number of faults. Specifically: Belady's anomaly can come into play for a FIFO replacement policy and certain access patterns.*

Problem 1g[2pts]: *Compulsory* misses in a cache can be reduced with prefetching.

(True) / False

Explain: *If the data is prefetched, it will already be in memory, so when that data is accessed the first time, there will not be a compulsory miss.*

Problem 1h[2pts]: A "memoryless" probability distribution provides a poor model for any real sources of events.

True / (False)

Explain: *When events from many independent sources are combined together the result can often be well approximated by a memoryless distribution—even if the individual sources are not memoryless.*

Problem 1i[2pts]: Nonvolatile Ram (NVRAM) can improve the durability of a file system that uses delayed writes.

(True) / False

Explain: *Data that has been buffered for delayed writes can be stored in NVRAM. If a power failure occurs before the data is written out to disk, the data can be recovered from the NVRAM.*

Problem 1j[2pts]: Because of the 32-bit IP-V4 address space, it is impossible for more than 2^{32} computers to communicate over the internet.

True / (False)

Explain: *Network Address Translation allows a single IP address to be used by several computers.*

Problem 2: Virtual Memory and Paging [25pts]

Consider a two-level memory management scheme on 22-bit virtual addresses using the following format for virtual addresses:

Virtual Page # (7 bits)	Virtual Page # (7 bits)	Offset (8 bits)

Virtual addresses are translated into 16-bit physical addresses of the following form:

Physical Page # (8 bits)	Offset (8 bits)

Page table entries are 16 bits in the following format, *stored in big-endian form* in memory (i.e. the MSB is first byte in memory).

Page Table Entry (PTE)

Physical Page # (8 bits)	Kernel Only	Uncacheable	0	0	Dirty	Use	Write	Valid

Note that a virtual-physical translation can fail at any point if an incompatible PTE is encountered. Two types of errors can occur during translation: "invalid page" (page is not mapped at all) or "access violation" (page exists, but access was illegal).

Problem 2a[2pts]: Can you give a logical reason why the designer might have made the virtual page # fields 7 bits each?

*With a 7 bit virtual page #, the size of each page table is equal to the size of one page of memory. 2^7 entries/table * 2 bytes/entry = 2^8 bytes/table; each page is 2^8 bytes.*
Grading: 2 points for recognizing that a page table fits exactly in one page; no other answer accepted.

Problem 2b[2pts]: What is the maximum amount of physical memory addressable by this system? Can you think of a way to increase the amount of available physical memory without altering the widths of the virtual addresses or PTEs?

16-bit byte-addressed physical addresses => 2^{16} bytes of physical memory is addressable.
You can use the two zero bits in the page table entry to address a larger number of physical pages (2^{10} physical pages). You can also increase the size of pages (by increasing offset width), or add registers for segmentation.
Grading: 1 pt. for each part.

Problem 3c[2pts]: How many total bits of storage will each entry of the TLB consume (including the tag and/or other fields)? Explain.

Simplest answer: TLB valid bit, 14-bit VPN, 16-bit PTE = 29 bits
Alternate: TLB valid, 14-bit VPN, 8-bit PPN (1/2 pt. each) , dirty, used, write (1/2 pt. for all 3) = 26 bits. We ignored kernel only and uncacheable bits.

Grading: -1 if you include 8-bit offset. -1 if you stated just "extra bits." -1/2 if you included zero bits in the PTE. If you said 16 bits for PTE, we also required you to explicitly state TLB valid bit.

Problem 2d[4pts]: The contents of physical memory are given on the next page. *Assume that the page-table base pointer = 0x2000, and that the CPU is in user-mode*. Translate the following virtual addresses. If there is an error during translation, make sure to say what the error is. Errors can be either "invalid page" or "access violation". Two answers are give. *Hint: be careful to look at the virtual address format!*

Virtual Addr	Physical Addr
0x10123	**0x0023**
0x60423	**Invalid Page**
0x28156	*0x5156*

Virtual Addr	Physical Addr
0x38256	*0x0056*
0x10F00	*Invalid page*
0x00278	*0x3278*

Grading: 1pt each, no partial credit

Problem 2e[6pts]: Consider the same multi-level memory management scheme. Once again, *assume that the page-table base pointer = 0x2000 and that the CPU is in user-mode.* Please return the results from the following load/store instructions. Addresses are virtual. The return value for load is an 8-bit data value or an error, while the return value for a store is either "**ok**" or an error. For errors, please specify which type of error (either "invalid page" or "access violation"). When the result is not an error, indicate the new value for the PTE after the access is complete. The first two results are given:

Instruction	Return Value	New PTE
Load [0x380FE]	**0xEE**	**0x3005**
Store [0x380FE]	**Access violation**	**------**
Load [0x41015]	*Access violation*	------
Load [0x00115]	*0x57*	*0x3105*
Store [0x08310]	*OK*	*0x000F*
Load [0x10102]	*0x10*	*0x0007*
Store [0x10731]	*Access violation*	------
Load [0x62345]	*Invalid page*	------

*Note: keep in mind that we are accessing things at user-level. This fact results in the "access violation" for **Load[0x41015]**, for instance. Also, for **Store [0x08310]**, we accepted "access violation," which is the result if you look at the write bit in the top level page table.*
Grading: ½ pt. each box for non-errors, 1 pt. each row if error.

Problem 2f[2pts]: Suppose that there are two processes loaded on the ready queue. Process A has page-table base pointer = 0x2000. Process B has page-table base pointer = 0x2F00. Give the *physical address* of a data page that is shared read-write (RW) between processes A and B? Is this data page mapped at the same place in the *virtual* address spaces of the two processes? Explain.

Page 0x51 (addresses 0x5100-0x51FF) are shared RW between the two processes.

The page is mapped at different places in the two address spaces, as they appear at different offsets into the top level page tables.

Grading: 1 pt. for the correct page, ½ pt. for "No", ½ pt. for explanation. Partial credit was given for 0x2800 (which is not a data page, but the second-level page table shared by the two processes.)

Virtual Address Format

Virtual Page # (7 bits)	Virtual Page # (7 bits)	Offset (8 bits)

Page Table Entry (PTE)

Physical Page # (8 bits)	Kernel	Not Cacheable	0	0	Dirty	Use	Write	Valid

Physical Memory

Address	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
0x0000	0E	0F	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D
0x0010	1E	1F	20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D
….																
0x1010	40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
0x1020	40	07	41	06	30	06	31	07	00	07	00	00	00	00	00	00
….																
0x2000	21	01	22	01	25	01	22	01	2F	03	28	03	30	03	22	03
0x2010	40	81	41	81	42	81	43	83	00	00	00	00	00	00	00	00
….																
0x2100	30	05	31	01	32	03	33	07	34	00	35	00	36	00	37	00
0x2110	38	00	39	00	3A	00	3B	00	3C	00	3D	00	3E	00	3F	00
….																
0x2200	30	01	31	83	00	01	00	0F	04	00	05	00	06	00	07	00
0x2210	08	00	09	00	0A	00	0B	00	0C	00	0D	00	0E	00	0F	00
….																
0x2500	10	01	00	03	12	85	13	05	14	05	15	05	16	05	17	05
0x2510	18	85	19	85	1A	85	1B	85	1C	85	1D	85	1E	85	00	00
….																
0x2800	50	01	51	03	00	00	00	00	00	00	00	00	00	00	00	00
….																
0x2F00	60	03	28	03	62	00	63	00	64	03	65	00	66	00	67	00
0x2F10	68	00	69	00	00	00	00	00	00	00	00	00	00	00	00	00
0x2F20	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
….																
0x30F0	00	11	22	33	44	55	66	77	88	99	AA	BB	CC	DD	EE	FF
0x3100	01	12	23	34	45	56	67	78	89	9A	AB	BC	CD	DE	EF	00
0x3110	02	13	24	35	46	57	68	79	8A	9B	AC	BD	CE	DF	F0	01
….																
0x4000	30	00	31	06	32	07	33	07	34	06	35	00	43	38	32	79
0x4010	50	28	84	19	71	69	39	93	75	10	58	20	97	49	44	59
0x4020	23	87	20	07	00	06	62	08	99	86	28	03	48	25	34	21

Problem 2g[4pts]: For the following problem, assume a hypothetical machine with 4 pages of physical memory and 7 pages of virtual memory. Given the access pattern:

A B C D E F C A A F F G A B G D F F

Indicate in the following table which pages are mapped to which physical pages for each of the following policies. Assume that a blank box matches the element to the left. We have given the FIFO policy as an example.

Access→		**A**	B	C	D	E	F	C	A	A	F	F	G	A	B	G	D	F	F
FIFO	1	A				E									B				
	2		B				F										D		
	3			C					A									F	
	4				D								G						
MIN	1	*A*															*D*		
	2		*B*														*D*		
	3			*C*									*G*				*D*		
	4				*D*	*E*	*F*												
LRU	1	*A*				*E*							*G*						
	2		*B*				*F*										*D*		
	3			*C*											*B*				
	4				*D*			*A*										*F*	

For MIN, we accepted the final D in any of the first three pages.
Grading: 2 pts. each for MIN and LRU; -1 for each error.

Problem 2i[3pts]: What is a precise exception and why would we want a software TLB fault to generate a precise exception?

A precise exception is one where the state of the machine is preserved as if the program executed up to the offending instruction. All previous instruction have completed, and the offending instruction and all following instructions act as if they had not even started.

We would like the hardware to generate a precise exception on a software TLB fault because it makes implementation of the fault handler easier. The fault handler simply restarts execution at the offending instruction (easy). It does not need to figure out the state of the system and potentially rollback or complete execution (in software) of partially completed instructions.

Grading: 1.5 pt. for each part.

Problem 3: File Systems [20pts]

Please keep your answers short (one or two sentences per question-mark). *We may not give credit for long answers.*

Problem 3a[2pts]: What is spatial locality and which type of file access pattern exploits spatial locality?

Spatial locality is accessing a location that is close to or next to recently accessed location. Sequential access of a file exploits spatial locality.

Grading: 1 point for definition and 1 point for access pattern.

Problem 3b[2pts]: Which component of disk access time is the *disk scheduling algorithm* trying to minimize?

The disk scheduling algorithm is attempting to minimize overall time wasted to moving the disk arm (i.e. it optimizes seek time).

Grading: 2 point for seek time or 1 point for some reasonable explanation of another component.

Problem 3c[3pts]: How does a *Journaled* file system improve the durability of data on disk? Give two reasons why a journaled file system would have higher performance than a file system that forced every block to the disk immediately, while at the same time giving the same or better durability.

Journaled file systems keep a log of changes to the file system; this log is forced to disk so that if a crash occurs the file system can recover all the transactions that haven't been committed properly to the disk.

1) *Since only modifications are forced to the log, the total traffic forced to disk might be a lot smaller than if every modified block must be sent to disk.*
2) *The log has great spatial locality (written on a set of consecutive sectors. Consequently, it can be forced to disk with little or no head movement.*
3) *When the file system is actually updated, the scheduler can order requests for higher efficiency – since the correctness depends on the log, not the order of updates to the rest of the disk.*

Grading: 1 point for explanation of Journaled and 1 point each for reasons.

Problem 3d[2pts]: Name at least two ways in which the *buffer cache* is used to improve performance for file systems.

1) *Reads can be from cache instead of disk*
2) *Allow delayed writes to disk, thereby permitting better disk scheduling an/or temporary files to be created and destroyed without even being written to disk.*
3) *Used to cache kernel resources such as disk blocks and name translations*

Problem 3e[2pts]: The Fast File System (FFS) of Berkeley 4.2 Unix utilized "Skip Sector Positioning" to improve performance. Explain what "Skip Sector Positioning" is and why this optimization may no longer be important.

Skip sector positioning placed successive sectors of a file on every other sector (or with multiple sectors between). This helps to avoid the situation in which the processor must do so much work after reading each sector that it misses the next sector and has to wait for a complete revolution. Modern disk controllers have track buffers that keep data from a complete track in memory, thereby eliminating the need for sector interleaving

Problem 3f[3pts]: The Network File System (NFS) uses a "stateless" protocol between clients and servers. What does this mean? Name one advantage and one disadvantage of stateless filesystem protocols.

All the requests to the NFS server are self-contained: they include all the arguments required for execution and do not assume that information is maintained between successive server requests.

Advantage: Server can crash and restart transparently to the client.
Disadvantage: Cannot track which clients are caching data, thus making it difficult to get clean caching semantics.

Grading: 1 point for explanation, 1 for advantage and 1 for disadvantage.

Problem 3j[6pts]: Suppose that a new disk technology provided access times that are of the same order of magnitude as memory access times. What, if anything, must be changed in the following three OS components to take advantage of the quicker access time? If something doesn't change, be very specific why it doesn't change. If it will change, contrast these changes with current implementation and be as specific as possible in your answers (i.e. identify what would change and why).

1. **Process Scheduler**
   Scheduler deals with processes at task level with arrival and run times so while run time might change the scheduler does not need change.

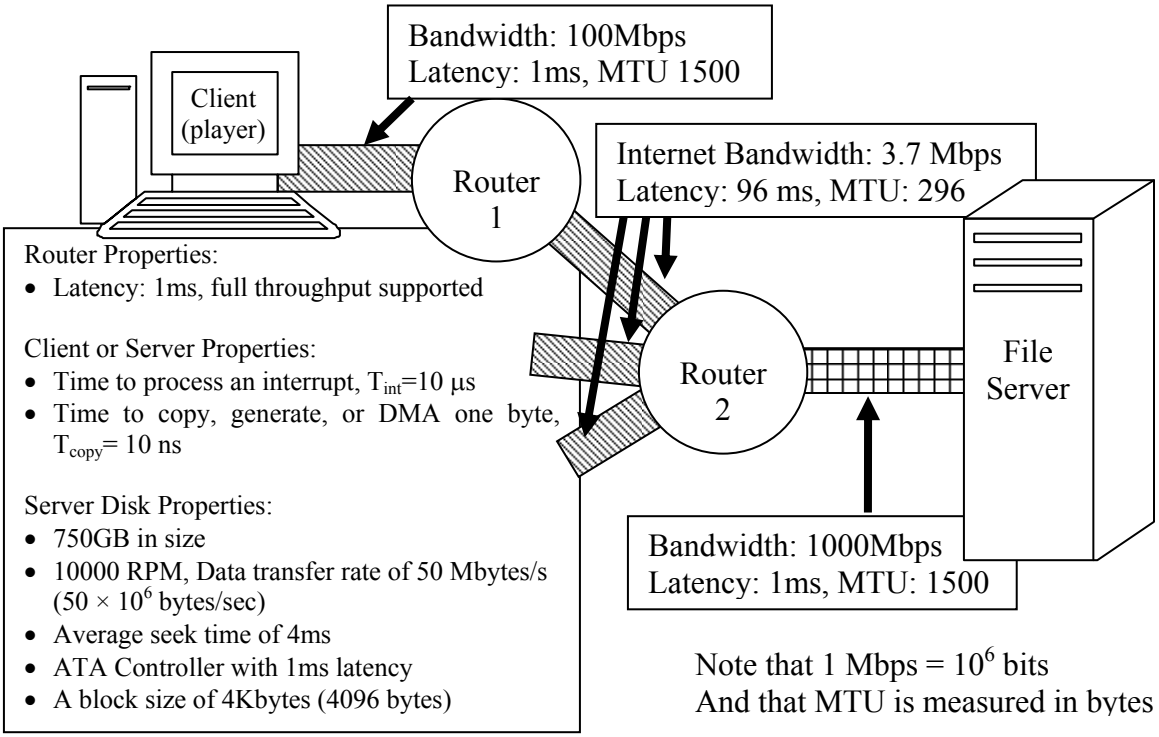
2. **Memory Management**
   Write through instead of write back since it's cheaper to go to disk. Less need to prefetch and instead save on the cost of getting unused memory location. Do more on-demand paging instead of having to rely on predicting access pattern and missing. Mainly, allows memory management to balance hit cost and miss cost instead of focusing on increasing hit rates only.

3. **Device Driver for the new disk**
   Use polling instead of interrupt for new disk device driver since it will be in heavy use due to lower access. Or no change to the structure of the device driver from OS point of view.

Grading: 2 points for good reasoning for change or no change with some modification if there is a change. 1.5 points for good explanation of the choice to modify or not. 1 point for incomplete reasoning but in the right direction.

[This page intentionally left blank]

Problem 4: Network File Server [35 pts]



The above figure illustrates a network in which one client interacts with a server to display video files. Each link is characterized by its Bandwidth, one-way Latency (for the first bit to arrive), and Maximum Transfer Unit (MTU) in *bytes*. All links are full-duplex (can handle traffic in both directions at full bandwidth). Each router is highly pipelined and will start forwarding a packet to an output port 1ms after it receives the first bit of the packet. Assume that no packets will be lost.

Both the video player (on the client) and file server run at user level. To display a video, the client opens a TCP connection to the server, sends multiple requests, then waits for the response. Assume that the time to setup the connection is negligible relative to the overall viewing time.

To send data over the network, a user-level process issues a write system call on a socket. The OS will copy the data to a kernel buffer (inside the socket). Then, the TCP routines generate headers (by copying them from templates) and use DMA to copy the completed packet to the network controller. Immediately after receiving all of the data for a packet, the network controller sends it. Afterwards, it generates an interrupt (one per packet).

At the destination, the network controller DMAs the packet to memory, then interrupts the CPU. Subsequently, the TCP routines transmit an acknowledgement (ACK) back to the sender. For simplicity, you may treat ACKs as if they are zero bytes long. After sending an ACK, the OS copies the received data into user-level buffer of the receiving user-level process.

A user-level process reads data from the disk by executing a *read()* system call with a user-level buffer for the results. The OS splits read requests into chunks that consist of sequential blocks on the same track and that are 16 blocks or less in size. It queues requests in the *device driver request queue (DDRQ).* Whenever the disk controller finishes reading a chunk of data into kernel memory, the controller generates an interrupt, causing the OS to dequeue the next request from the DDQ. The OS then copies data from kernel memory into the user's buffer and returns from *read().*

Problem 4a[3pts]: What is the maximum amount of *data* that the client or server can send in each packet while avoiding fragmentation along the way? Remember that the TCP+IP header is 40 bytes. Explain.

Since the minimum MTU along the path is 296, and the header is 40 bytes, the maximum amount of data that the client or server can send in each packet is 296-40 = 256 bytes.
Grading:1pt for minimum MTU, 1 pt for subtracting header, 1 for some explanation

Problem 4b[2pts]: Explain how the client or server could automatically discover the answer to 4a:

The sending end starts with large packets with the "no fragment" flag set in the header and slowly reduces the size until packets start making it through without being dropped.
Grading: 1 pt for plausible endpoint-generated query, 1 pt for mentioning no frag bit

Problem 4c[2pts]: Under ideal circumstances (and ignoring interrupt and copying overheads and window sizes), what is the maximum *data* bandwidth that could be sent from the server to the client without dropping packets, assuming that the server utilizes the mechanism in 4b? Explain.

Assuming that we use the maximum packet size from 4a (i.e. 296) which contains 256 bytes of data in it, we can compute the fraction of the 3.7 Mbps that we have in the middle link as: 3.7 Mbps $\times$ (256/296) = 3.2 Mbps
Grading: 1pt identifying bottleneck bandwidth, 1 pt for scaling by correct fraction

Problem 4d[3pts]: Assume that a maximal sized packet (as in 4a) is sent from the TCP send buffer of the server to the TCP receive buffer of the client. Compute the total roundtrip latency from the point at which the server invokes the device driver to send the packet until it receive the ACK via an interrupt. *Hint: don't forget the interrupt at the receiving side and the DMA mechanism into and out of the network.*

Note that the fact that the router is pipelined (see description above) implies that the router begins routing toward the output port even before it has received the complete packet. This simplifies things so that we only have to track latencies of packets through the router as 1ms. At the receiving side, we technically need to wait for the packet to be completely received before we start DMAing it into kernel memory.

Also, we said that the ACK is zero length. If this is truly zero, then there is no header generation/transmission latency. Otherwise, we need to generate headers. There are two different options here: Ack latency (zero) or Ack latency (header) here.

We translate to ms here. 10ns=10^{-5}ms, 10μ = 10^{-2}ms. 1Gps=10^{12}bits/ms 100Mbps=10^{11}bits/ms

$$\text{Message latency} = [Gen_{header}+DMA_{kernel\rightarrow network}]+[Link_2+Route_2+Link_{middle}+Route_1+Link_1]+$$
$$[Receive_{server}+DMA_{network\rightarrow kernel}+Int_{message}]$$
$$=(40\times10^{-5}ms+296\times10^{-5}ms] + [1ms+1ms+96ms+1ms+1ms]+$$
$$[(296\times8\ bits)/(10^{11}bits/ms)+296\times10^{-5}ms+10^{-2}ms] \cong 100.02ms$$

$$\text{Ack latency (zero)} = [Link_2+Route_2+Link_{middle}+Route_1+Link_1]+Int_{ack} = 100.01\ ms$$

$$\text{Ack latency(header)} = [Gen_{header}+DMA_{kernel\rightarrow network}]+\text{Ack latency(zero)}+$$
$$[Receive_{client}+DMA_{network\rightarrow kernel}]$$
$$=[40\times10^{-5}ms+40\times10^{-5}ms]+100.01ms+$$
$$[(40\times8)/(10^{12}bits/ms)+40\times10^{-5}ms] \cong 100.01ms$$

Total Answer $\cong$ 200.03ms
Grading: 1 point on right track, -2 for missing all network components, -1 point for missing components of basic 200ms roundtrip, -½ for missing DMAs, -½ for missing ints, -½ for interrupt on send (not in critical path), -½ for copy into kernel (already in TCP buffer), -½ math error

Problem 4e[3pts]: What TCP window size is necessary to achieve the bandwidth of 4c without dropping packets in the network? Use simple constants and values that you computed for 4a-4d. Explain. Make sure that you correct for units.

What is needed here is the bandwidth-latency product, where the latency is the roundtrip latency. The result will be the total amount of data that needs to be "in the network" before reception of the first ACK. Here, the window size is in data bytes (not including header) so, we use the data bandwidth, not total bandwidth. Also, we must correct for the fact that (4d) was computed in ms. Notice how all of the units cancel out properly! (we want bytes)

$$\begin{aligned} \textit{Window size} &\cong (4c)\times(4d)\times(10^{-3}\,s/ms) \,/\, (8bits/byte) \\ &\cong (3.2\times10^{6}\,bit/s)\times(200.03ms)\times(10^{-3}\,s/ms)\times(0.125\ bytes/bit) \\ &\cong 80012\ bytes \end{aligned}$$

Grading: -1 use incorrect bandwidth (unless consistent with 4c), -1 missed converting bits to bytes (unless give answer with units of "b," which means bits, -1 for not following directions (using only values from 4a through 4d or simple constants).

Problem 4f[6pts]: Suppose that video files are laid out in 64K (65536 bytes) chunks on the disk (i.e. 64K in successive sectors on a track). Compute the overhead for reading such a 64K chunk from a random place on the disk. Assume that the disk controller automatically DMAs the data to kernel memory in a fashion that is overlapped with reading it from the disk (so that you do not have to worry about DMA for this operation). After finishing, the controller generates an interrupt; the interrupt routine may submit another request to the controller (if one is queued on the DDRQ). Assume the disk parameters given above (repeated here):

- 750GB in size
- 10000 RPM, Data transfer rate of 50 Mbytes/s (50×10^6 bytes/sec)
- Average seek time of 4ms
- ATA Controller with 1ms controller initiation time
- A block size of 4Kbytes (4096 bytes)

What is the total time to read 64K chunk from a random place on the disk into memory including the interrupt? *Hint: there are 5 terms here including the interrupt!*

$$\begin{aligned} Time_{server} &= Time_{controller}+Time_{seek}+Time_{rotate}+Time_{transfer}+Time_{Int} \\ &= 1ms + 4ms + \\ &\quad \tfrac{1}{2}\,[\,60\ s/min \times 1000\ ms/s \,/\, 10000\ rev/min\,] + \\ &\quad 65536\ bytes \,/\, [(50\times10^{6}\ bytes/sec)\times 0.001\ sec/ms] + 0.01\ ms \\ &= 8.01ms + 1.31ms = 9.32ms \end{aligned}$$

Grading: 1 pt for each of 5 terms, 1 pt for calculation (can lose if have a unit conversion problem, like forgetting to convert from sec to ms or something similar, other reasons).

Problem 4g[3pts]: Now, assume that the video player works by sending requests for 64KB (=65536 bytes) at a time to the video server. Assuming that these requests are pipelined for maximum bandwidth, at what rate must it send these requests to achieve the bandwidth of 4c?

$Num_{req/client}$ = (3.2×10^6 bits / sec $\times$ 0.125 bytes/bit) / 65536 bytes/req = 6.1 req/sec

Grading: 1 pt if mostly on track, -1 pt for forgetting bits/byte conversion, -1 pt neglecting to use value from 4c.

Problem 4h[4pts]: Compute the total *processor* overhead for satisfying a 64K request at the server. Processor overhead includes (1) interrupts, (2) header generation overhead and (3) copying overhead. *Note that generating a byte of header is as expensive as copying.* Processor overhead does not include DMA or disk controller actions since these are overlapped with the CPU. The data portion of a request message is 16 bytes.

We translate to ms here. 10ns=10^{-5}ms, 10μ = 10^{-2}ms. 1Gps=10^{12}bits/ms 100Mbps=10^{11}bits/ms

$$\begin{aligned} Overhead &= Request[\, Int_{request} + Copy_{request}] + Disk[Int_{disk} + Copy_{filesystem \rightarrow user}] + \\ &\quad Send[Copy_{user \rightarrow socket} + 256 \times Gen_{header} + 256 \times Int_{network}] + Acks[256 \times Int_{network}] \\ &= Request[10^{-2} + 16 \times 10^{-5}] + Disk[\, 10^{-2} + 65536 \times 10^{-5}] + \\ &\quad Send[\, 65536 \times 10^{-5} + 256 \times (40 \times 10^{-5}) + 256 \times 10^{-2}] + Acks\,[256 \times 10^{-2}] \\ &= 0.01016ms + 0.75536ms + 3.31776ms + 2.56ms = 6.6438ms \cong 6.64ms \end{aligned}$$

Since TCP might generate less than one ack/packet, we can say that the result is at least > 4.08ms (the above number without the ACK interrupt term of 2.56ms).

Note: We don't include the interrupt/header generation for the Ack to the request because Acks will likely ride the data packets going back to the client. We didn't take off points if you included these.

Grading: we didn't require the ack interrupts, although they are technically overhead. Roughly, we gave 1 point for Request overhead, 1 pt for Disk overhead (especially copy from kerne→user), 2 points for Send, divided into 1 point for user→ kernel copy for send and 1 point for packet generation of headers and packet interrupts.

Problem 4i[4pts]: Suppose that we wanted to handle 300 clients each at a request rate as given in (4g). Assume that multiple clients connect through network resources that are independent up to router #2. Each client would have its own user-level process working on its behalf. How many disks would we need to handle this rate? What else would need to be upgraded to handle this rate?

Number of requests / sec that disk can handle: (1000 ms / s) / (9.32 ms / req) = 107.3 req/s
Number of clients/disk = (107.3 req/sec) / (6.1 req/s/client) = 17.6 clients
300 clients / 17.6 clients/disk = 17.05 disks ⇒ we need 18 disks to meet the full demand.

Would need to upgrade network:
300 clients $\times$ 3.7 Mbps/client = 1.11×10^9 bps > 1Gbps!
Would need to upgrade processor:
Since overhead/req = 6.64ms, 300clients $\times$ 6.1req/client = 18300 req/s
⇒ request rate would require .00664 $\times$ 18300 = 121sec of processor time / sec!

Grading : 3 points for computing number of disks, 1 point for pointing out and justifying one upgrade. -1 pt for wrong disk BW (like simply using 50MBps) ; -1pt for computation error ; -½ point for naming an upgrade without giving justification ; Also, a couple of people said that memory would need to be upgraded, however, the memory requirements of this application are minimal – and you would have no way to know how much memory was there to start with.

Problem 4j[5pts]: Returning to our single-disk server, assume that 15 clients connect through network resources that are independent up to router #2. Each client has its own user-level process working on its behalf. Assume that all 15 clients send requests at a rate of (4g). If the disk response distribution can be described with C=1.5 and the aggregate network request rate can be described as exponential, what is the average length of the disk device driver request queue (DDRQ)? *Hint: the service time for the disk is the time between requests to the disk controller when the DDRQ is full. Possibly useful formulas include:*

Mean Service: $T_{server} = \frac{1}{n}\sum_{i=1}^{n} T_i$ *Variance:* $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(T_i - T_{server})^2$

M/G/1 queue: $T_q = T_{server} \times \left(\frac{1+C}{2}\right) \times \left(\frac{u}{1-u}\right)$ *Little's law:* $L_q = \lambda \times T_q$

Solution: *The important thing about solving this problem is figuring out what equations to use and which numbers to plug in. Since we are looking at the DDRQ queue, the service time is the time for one chunk (64KB) to be grabbed from the disk and to generate an interrupt. This service time is exactly what we computed in (4f). The processor overhead from (4h) is overlapped with these hardware-generated numbers, so is not included. Note that we can get average service time directly from (4f), so don't need to use the formula for "Mean Service" given above.*

The arrival rate (λ) includes 15 clients worth of requests, so is the result of multiplying (4g) by 15.

$\lambda = Num_{clients} \times Num_{req/client}$ = 15 × (4g) = 15 x 6.1req/sec = 91.5req/sec
$Time_{server}$ = (4f) = 9.32ms/req (overhead from 4h is overlapped and not counted)
$u = \lambda \times Time_{server}$
= 91.5req/sec × 9.32ms/req × 0.001 s/ms ≅ 0.853 (No UNITs!)
T_q = (9.32ms/req × 0.001s/ms) × ½(1+1.5)×[0.853/(1-0.853)]
= 0.0676
L_q = (15×6.1)× 0.0676 = 6.1854

⇒ *Average queue length = 6.1854*

Grading: 1 point for identifying formulas that get you all the way to the answer; 1pt for λ; 1pt for u; 1pt for $Time_{server}$; 1 point for calculation.

Note that we gave you credit for using numbers that you computed previously, even if they were incorrect. However, we gave -1pt if you computed a value for u that was > 1 but didn't say anything about it (i.e. didn't point out that you must have made a mistake).

Also, you could use -1pt for failing to correct for units (i.e. ms/s etc).

[Scratch Page (feel free to remove)]