Stochastic Fairness Queuing*

Paul E. McKenney
IBM Beaverton
pmckenne@us.ibm.com†

Abstract

Fairness queuing has recently been proposed as an effective way to insulate users of large computer communication datagram networks from congestion caused by the activities of other (possibly ill-behaved) users. Unfortunately, fair queuing as proposed by Shenker et al. [DKS89] requires that each conversation[3] be mapped into its own queue. While there are many known methods of implementing this type of mapping, they are relatively slow, requiring numerous memory references, and thus do not lend themselves to a software or firmware implementation capable of operating in high-speed networks.

This paper presents a class of algorithms collectively called "stochastic fairness queuing" that are probabilistic variants of fair queuing. These algorithms do not require an exact mapping, and thus are suitable for high-speed software or firmware implementation. Furthermore,

*This work was supported by the Rome Air Development Center and the Defense Advanced Research Projects Agency under contract number F30602-89-C-0015, and by SRI Internal Research and Development.

†This work was performed while the author was with SRI International, Menlo Park, CA

[3]A "conversation" consists of all packets with a given source-destination address pair.

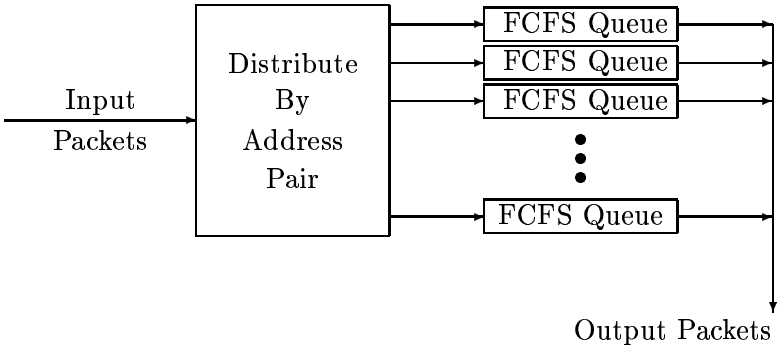


Figure 1: Fairness Queue

these algorithms span a broad range of CPU, memory, and fairness tradeoffs.

1 Introduction

Current datagram networks are vulnerable to congestive collapse when offered load approaches network capacity [BG87]. Although several end-to-end congestion-avoidance algorithms have been proposed [BG87, Hay81, JR87, RFS90, Jac88], none of them have been shown to perform optimally in today's high-speed, high bandwidth-delay-product networks [DKS89, BG87]. This has led some researchers to conclude that gateways must participate in congestion avoidance [Nag87]. To this end, it has been proposed that gateways use a fair queuing algorithm [Hah86, Nag87, DKS89, DH89].

This fair queuing algorithm operates by maintaining a separate first-come-first-served (FCFS) queue for each conversation, as shown in Figure 1. Since the queues are serviced in a manner that approximates bit-by-bit round-robin,[4] ill-behaved conversations that attempt to use more than their

[4]Bit-by-bit round-robin services queues in an order that allocates bandwidth equally

fair share of network resources will face longer delays and larger packet-loss rates than well-behaved conversations that remain within their fair share. Shenker et al. have presented results showing that this algorithm performs well with a variety of topologies and traffic patterns [DKS89].

Maintaining a separate queue for each conversation requires that the gateway be able to map from source-destination address pair to the corresponding queue on a per-packet basis. There are a number of methods of accomplishing this [Knu73]; however, they are relatively slow (requiring numerous memory references), and are therefore unsuitable for use in gateways operating in high-speed networks. See Section 2 for a discussion of possible implementations of fair queuing.

In summary, although fair queuing exhibits excellent behavior, its computational requirements render it infeasible for use in high-speed networks.[5] Since we cannot afford the perfect justice provided by fair queuing, we turn to stochastic fairness queuing, which will be shown to provide reasonable justice at a price we can afford.

Stochastic fairness queuing can be most easily understood by comparing it to strict fair queuing. The major differences are that the queues are serviced in strict round-robin order and that a simple hash function is used to map from source-destination address pair into a fixed set of queues, as in the (very small) six-queue example shown in Figure 2. If the number of queues is large compared to the number of conversations, each conversation is very likely to be assigned to its own queue. If two conversations

to the queues. If all packets are the same size, this degenerates to simple round-robin service.

[5]However, fair queuing is quite feasible in low-speed networks, many of which still exist [Lou89].

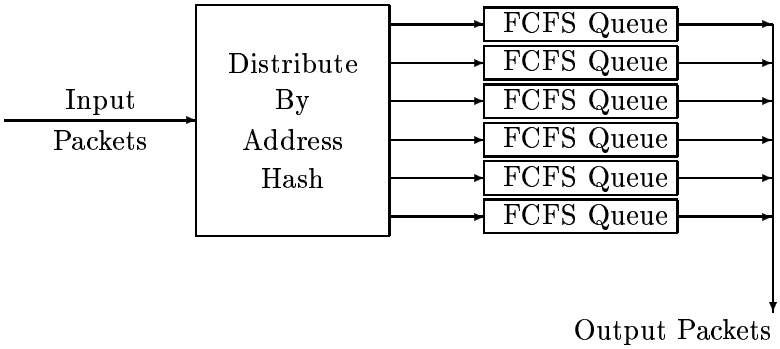


Figure 2: Stochastic Fairness Queue

collide, they will continue to collide, resulting in each conversation of the pair persistently receiving less than its share of bandwidth. This situation is alleviated by periodically perturbing the hash function (as described in McKenney [McK89]) so that conversations that collide during one time period are very unlikely to collide during the next.

Simulation results presented in Section 5 show that stochastic fairness queuing can achieve a performance that approaches that of fairness queuing and greatly exceeds that of simple FCFS queuing.

Stochastic fairness queuing can act in concert with many end-to-end congestion-avoidance algorithms, such as the DEC-bit scheme and the Jacobson/Karels TCP [DKS89].

Stochastic fairness queuing can also be combined with resource reservation algorithms such as flows [Zha89], MCHIP [Par90], and ST [Gro90]. These algorithms require per-connection information (such as maximum and average allowed throughputs), which is obtained from the user, possibly through negotiation with a network resource scheduler.

A network that provided both best-effort and reservation services would

use a reservation algorithm to control the bandwidth used by predictable traffic, and stochastic fairness queuing to allocate the remaining bandwidth fairly. The reservation algorithm would leave some fraction of the total bandwidth for best-effort services.

Note that this hybrid datagram/connection approach allows so-called "hot pairs"[6] of hosts to be handled in a natural way; these hosts could use reservation services to obtain the needed bandwidth.

2 Alternatives for Fair Queuing Implementations

The performance benefits offered by stochastic fairness queuing do not come for free. The price is loss of determinism and of absolute guarantees of fairness. Readers with experience in the areas of caches and hash tables may be very familiar with this tradeoff; these readers may wish to skip to the next section.

Much of the motivation for stochastic fairness queuing stems from two aspects of fair queuing that do not lend themselves well to high-speed implementations.

The first aspect is the packet scheduling technique that fair queuing uses to provide bit-by-bit round-robin service. This technique requires addition to and deletion from a priority queue of length equal to the number of conversations flowing through the queue. The best known priority queue algorithms provide time complexity of $O(\log(n))$, where n is the length of the priority queue. The number of conversations in a gateway has been

[6] A "hot pair" of hosts needs to exchange an unusually high volume of traffic. An example of a hot pair of hosts might be a mail gateway.

measured to be as large as 180 [Fel89] and is expected to grow larger (see Section 3.2). The average computational cost of just the packet scheduling part of fair queuing, when applied to this number of conversations, can exceed the worst-case total computational cost of an efficient version of stochastic fairness queuing.

The second aspect is the technique of using a one-to-one mapping from source-destination address pair into the corresponding queue. The remainder of this section examines alternative implementations of this mapping and shows how each is deficient for high-speed networks. In some cases it is necessary to look at machine-language implementations; in these cases, the Motorola MC68020 processor will be used.

At first glance, it would appear that whatever strategy was used to look up routes would suffice for address-pair mapping. However, while routing updates (which can modify the routing data structure) occur relatively infrequently, modifications to the fair queue structure can occur on a per-packet basis. Therefore, unlike routing, fair queueing cannot amortize the cost of data structure modifications over a long time period.

Another approach is to rely on hardware assists such as content-addressable memories. Such assists can work quite well for gateways supporting a single protocol that is not subject to growth and revision. However, the growing demand for gateways that support multiple protocols, most of which are still evolving, render this approach impractical for internetworking gateways for the foreseeable future.

The simplest and fastest way of mapping from source-destination address pair into queue is to use a simple array, indexed directly by the binary number formed by concatenating the bits representing the source and des-

tination addresses. For example, IP has relatively small 32-bit addresses, so that the corresponding index would be a 64-bit quantity. Unfortunately, this results in an infeasible 2^{64} element array, eliminating this approach from further consideration.

Various types of search trees are heavily used in database applications [Knu73]. These methods are relatively slow, requiring numerous memory references for access and complex algorithms for updates, making them unsuitable for use in switches operating in high-speed networks.

A particularly seductive alternative is the trie [Knu73]; the following paragraphs examine it in detail. For concreteness, imagine maintaining a 256-way trie indexed by successive bytes of the IP address pair. The first byte of the IP address pair would be used as an index into a table of 256 possibly NULL pointers. Each non-NULL pointer would point to its own table of 256 pointers, again possibly NULL, indexed by the second byte of the IP address. These tables of pointers would form an eight-level tree; non-NULL pointers at the eighth level would point to a queue header. NULL pointers at all levels are placeholders for address pairs that do not correspond to any currently active conversation.

We will assume that conversations average at least three packets in length; thus an implementation of a fair queuing trie should be optimized for the second and subsequent packets in a conversation. Since there are only eight levels in an IP trie, it makes sense to fully unroll the loop that traverses the trie. Assuming that a pointer to the concatenated IP address pair and a pointer to the root of the trie are preloaded into registers, each segment of the unrolled loop will contain three instructions. The first instruction will load the next byte of the address pair into a register while incrementing the

pointer to the address, the second instruction will use this byte to index into the current level of the trie, loading a pointer to the next level, and the third instruction will branch to a special handler if this pointer is NULL. NULL pointers would be encountered upon receipt of the first packet of a new conversation; the special handler would allocate and initialize memory needed to add the new address pair to the trie.

Use of a trie thus requires 24 instructions to map from IP address pair to the corresponding queue in the best case; this ignores the added instructions needed to allocate structures for new conversations and needed to scan the trie periodically to dispose of structures corresponding to conversations that have ended. In contrast, an efficient implementation of a stochastic fairness queue requires fewer than 10 MC68020 instructions in the *worst* case to map from IP address pair to the corresponding queue.[7] As noted earlier, IP has relatively short addresses; stochastic fairness queuing's advantage is greater for longer addresses.

A final alternative is the use of hash tables with chaining. The best case instruction count for a hashed fair queue in an IP network is almost as good as the worst case for stochastic fairness queuing. The difference is due to the fact that the hashed fair queue must compare the address in the packet to that of the first queue header in the chain; fair queuing must reference address fields three times as often as stochastic fairness queuing.[8]

[7]This assumes that the hashing function is implemented in software; the instruction count might decrease somewhat given a hash function implemented in hardware.

[8]Stochastic fairness queuing must scan the packet's address pair once in order to compute the hash function. Fair queuing must in addition scan the packet's address pair and the queue header's address pair in order to compare the two. Protocols with short address fields such as IP may allow fair queuing implementations that cache the packet's address

In addition, a hashed fair queue must periodically scan its queues in order to dispose of those corresponding to conversations that have ended, and must allocate and initialize new queues upon arrival of a packet that is part of a new conversation. The overhead due to these activities will depend on traffic statistics.

In summary, the worst-case execution speed of stochastic fairness queuing is faster than the best-case execution speed of all of these implementations of fair queuing, and this advantage is larger for protocols with longer addresses, e.g., the ISO protocol suite.

3 Analysis

The following sections analyze expected queue occupancy and a bound on the number of conversations passing through a gateway.

3.1 Expected Queue Occupancy

The analysis of stochastic fairness queuing closely parallels that of hash tables with chaining. The only difference is that a collision in a hash table causes a search of only half of the linked list (on the average), while a collision in a stochastic fairness queue causes all of the colliding conversations to share the queue. An analysis of hash table performance may be found in Graham et al. [GKP89]. Adapting this analysis to stochastic fairness queues and assuming a large number of queues gives the number of conversations that

pair in machine registers, but this is not likely to be practical for protocols with longer address fields. Note that it is not possible to overlap computation of the hash function with comparison of the address pairs, since the hash function must be computed before the queue can be located.

a given conversation can expect to share its queue with (counting itself), represented by

$$EC = \alpha + 1 \tag{1}$$

and

$$VC = \frac{\alpha^2}{6} + \alpha \quad , \tag{2}$$

where EC is the expected number of conversations, VC is the variance in the expected number of conversations, and α is the ratio of the number of conversations to the number of queues.

Consider a stochastic fairness queue that is empty (its occupancy is zero). Then a new conversation is guaranteed to be given a queue with exactly one occupant (the conversation itself). On the other hand, a stochastic fairness queue with as many conversations as queues (occupancy of one) has a value of 2 for EC and a value of about 1.17 for VC. This indicates that a new conversation will share its queue with one other conversation on the average, but that the actual number of conversations in the queue may vary considerably from that value.

Note that the expected number and variance of conversations sharing a given queue will be low when the occupancy is low. This indicates that the stochastic fairness queue will behave in a very consistent, predictable manner when given a sufficient number of queues (for instance, about five or ten times the number of conversations). Feldmeier has collected data showing that the number of concurrent conversations passing through MIT's ARPANET gateway in early 1988 was almost always less than 180 [Fel89]. This would indicate that about 1000 to 2000 queues would suffice; this number can be easily accommodated by today's large random-access memories.

The number of queues needed scales with the speed of the network, as shown in the next section, and thus more memory is required for higher speed networks.

The fact that both EC and VC are continuous with respect to the occupancy indicates that the algorithm is not prone to sudden failure, but instead gracefully degrades under overload.

3.2 Bound on Number of Conversations

A fair queueing algorithm such as stochastic fairness queueing can guarantee fair service only if it maintains separate state for each and every conversation. If state (such as queues or finish-time counters) is shared between two conversations, then the algorithm will be unable to distinguish the conversations and will thus be unable to prevent one of the conversations from stealing resources from the other. The maximum number of conversations that may through a gateway at a given time is vitally important, as this number defines the amount of state information needed to allow the fair queueing algorithm to guarantee fair service.

This number may be calculated given a maximum per-conversation packet interarrival time. For example, if a three-hour-long conversation contains a one-hour gap, it is reasonable to model it as two smaller conversations. In the following, we derive a formula for the maximum number of conversations as a function of the link bandwidths, average packet size, and maximum inter-packet gap allowed in a conversation.

The average number of packets per second arriving at the gateway, as-

suming that each link is fully utilized, is given by

$$P = \frac{\sum_{i=1}^{N} C_i}{S} \quad , \tag{3}$$

where S is the average packet size in bits, N is the number of interfaces, and C_i is the capacity, in bits per second, of the i^{th} interface. If G is the largest gap allowed between consecutive packets belonging to a given conversation, then the maximum number of simultaneous conversations is simply GP, which expands to

$$M = \frac{G \sum_{i=1}^{N} C_i}{S} \quad . \tag{4}$$

This maximum occurs when each conversation transmits exactly one packet during each time interval of length G, in other words, when the gateway is giving the conversations perfectly fair service.

Table 1 shows the maximum number of conversations for a four-interface gateway under typical (maximum inter-packet gap of 10 seconds, 10% of conversations accounting for 90% of throughput, and average packet size of 1000 bytes) and worst-case (maximum inter-packet gap of 10 seconds, each conversation providing equal throughput, average packet size of 50 bytes) conditions. The number of conversations under typical conditions is given by

$$M_t = \frac{B_L}{B_H} M \quad , \tag{5}$$

where B_L is the fraction of bandwidth alloted to the high-bandwidth conversations, B_H is the fraction of bandwidth alloted to the low-bandwidth conversation, and M is computed as shown in the previous equation. These figures show that congestion-avoidance algorithms must scale well with increasing numbers of simultaneous conversations if they are to be usable in high-speed networks.

Table 1: Bounds on Number of Conversations

Medium	Typical	Worst-Case
ARPANET (56kbps)	31	5,600
T1 (1.5 Mbps)	853	153,600
T3 (45 Mbps)	25,600	4,608,000
Fiber (1 Gbps)	555,556	100,000,000

4 Example Implementation

The following subsections describe the requirements for the hash function, exhibit a particular function that meets those requirements, and demonstrate the data structures and algorithms used by a specific implementation of stochastic fairness queuing.

4.1 Hash Function

The example implementation uses a hash function to map from source-destination address pair to queue index. This hash function must give a high information content, as defined by Jain in [Jai89], but must also allow perturbation such that address pairs that collide for one perturbation value are very unlikely to collide for a different perturbation value.

Two hash functions were used in simulations. The first is the high-level data-link control (HDLC) procedure (ISO-3309-1979) cyclic redundancy check (CRC) [Int79]. Hardware implementations of the HDLC CRC are commercially available. This hash function is perturbed by multiplying each byte by a sequence number in the range from 1 to 255 before applying the CRC.

The second is the simple software algorithm given by

$$\text{hash} = \text{ROL}(\text{src}, \text{seq}) + \text{dst} \quad , \tag{6}$$

where "ROL" is the circular rotate-left function implemented as a single instruction on many computers, "src" is the Internet Prococol (IP) source address, "seq" is a sequence number in the range from zero to 31 that is used to perturb the hash function, and "dst" is the IP destination address. Although this hash function does not give as high an information content as the HDLC CRC, it can be implemented very efficiently on many computers and performs very well in simulations provided that the number of queues is not a small integral multiple of a power of two.

Note that this software hashing function is specific to IP; more work is needed to find a software hash function that can efficiently handle variable-length addresses such as those found in the ISO protocols.

4.2 Data Structures and Algorithm

A particular algorithm from the class of stochastic fairness-queuing algorithms was simulated in order to provide proof of concept. This section describes the data structures used by this particular algorithm (see Figure 3); pseudo-code is given in Reference [McK90]. See Section 6 for discussion of other instances of stochastic fairness queuing.

The type of stochastic fairness queue simulated consists of an array of finite-length queues (the lettered boxes in Figure 3); a doubly-linked "active list" that includes only those queues that are non-empty (in this case, the queues labeled "A", "B", and "D"); a round-robin pointer that points to the queue that is to be serviced next (currently "D"); an array of doubly-linked

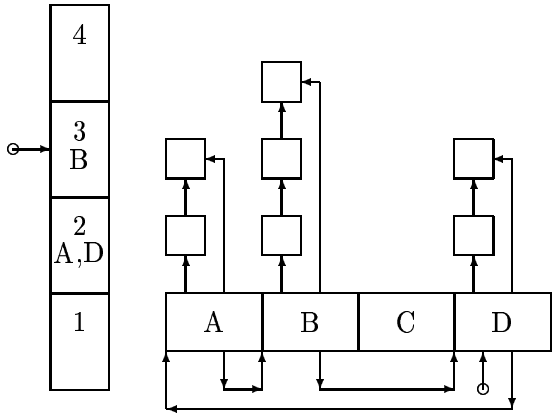


Figure 3: Stochastic Fairness Queuing Data Structures

"number-of-elements lists" (the numbered boxes in Figure 3, one for each possible non-zero queue length); and a maximum-size pointer to the element of this array corresponding to the longest queue (currently three). The links for the number-of-elements lists have been omitted from the figure in the interest of readability; list one is empty, list two contains queues "A" and "D", list three contains queue "B", and list four is empty. Queue "C" is empty, and therefore does not appear on either the active list or any of the number-of-elements lists.

The purpose of the active list and the round-robin pointer is to allow the next departing packet to be located without wasting time in scanning over empty queues. The purpose of the number-of-elements list and the maximum size pointer is to allow the longest queue to be located without wasting time searching.[9] The hash function perturbation is done in-place,

[9]This occurs when the buffer pool is exhausted, in which case buffers will be stolen from the longest queue in order to accommodate arriving packets. This buffer theft is the only purpose of the number-of-elements list; eliminating this data structure would allow

so no special data structure is required to control this operation.

Note that all of the algorithm's operations are time complexity $O(1)$,[10] suitable for implementation in high-speed software or firmware. In particular, adding a packet to a queue requires two doubly-linked-list operations and one singly-linked-list operation (in addition to the hashing operation), unless the buffer pool was exhausted, in which case it requires an additional two doubly-linked-list operations and one singly-linked-list operation (for a total of four doubly-linked-list operations, two singly-linked-list operations, and one hashing operation). Deleting a packet from a queue always requires two doubly-linked-list operations and one singly-linked-list operation.

Because of the fact that it is never necessary to do any scanning of the data structures comprising a stochastic fairness queue or any source-destination-address-pair comparisons, the operation count is quite small, compared to that of fair queuing.

5 Simulation

The behavior of stochastic fairness queuing was studied using three different simulations. The first examines the behavior of the perturbable hash functions in isolation. The second consists of a single overloaded node with no transport protocol action; this was used to do parametric studies. The third is a more realistic simulation of multinode networks with several transport

significant speedup. This and other variants of stochastic fairness queuing are discussed in Section 6.

[10]The "big-O" notation describes the asymptotic performance of an operation or algorithm within a constant factor [Knu73]. Thus, an $O(1)$ operation is guaranteed to complete in a fixed amount of time, regardless of the size of the problem.

protocols [Kes89].

5.1 Hash Functions

The hash functions are tested against 10,000 randomly selected pairs of IP address-pairs from the HOSTS.TXT file (available from NIC.DDN.MIL). The address pairs are hashed by the function under test with each possible perturbation value in turn; the number of perturbation values for which the address pairs' hash collides is counted. This process is repeated for each possible queue header array size in the range from 2 to 512, inclusive. The value from the hash function is taken module to queue header array size; collisions are of course checked *after* the modulo operation.

The output of an ideal hash function would be indistiguishable from an uniformly-distributed random variable. This would result in M/N collisions on the average, where M is the number of perturbation values and N is the queue header array size.

The results of this simulation along with the 95% confidence interval for the CRC hash are shown in Figure 4. The 95% confidence intervals bracket the M/N curve; thus the behavior of the CRC hash is very close to ideal.

The results of this simulation for the rotation hash are shown in Figure 5. The 95% confidence intervals bracket the M/N curve except for those values of the queue header array size N that are powers of two or small integral multiples of powers of two. These values of N are where the upward spikes occur in Figure 5, for example at 128, 160, 192, and 256. The reason for this poor behavior is that the rotation hash mixes the bit patterns of its input less thoroughly than does the CRC hash, and thus relies on the modulo operation to do additional mixing. When the modulus is a small multiple

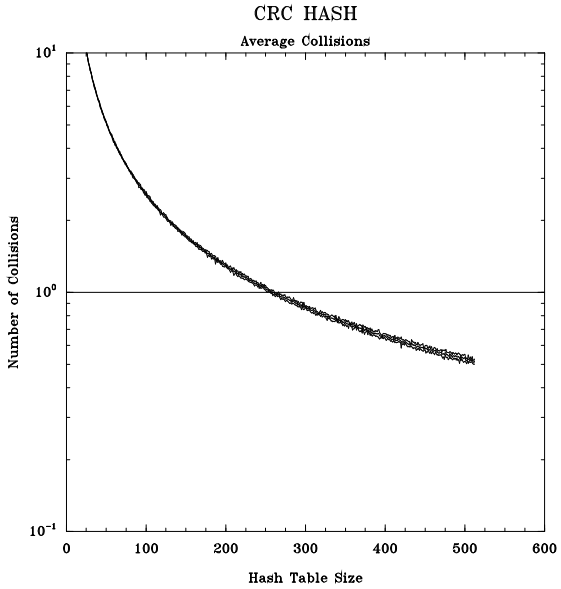


Figure 4: Collisions For CRC Hash Function

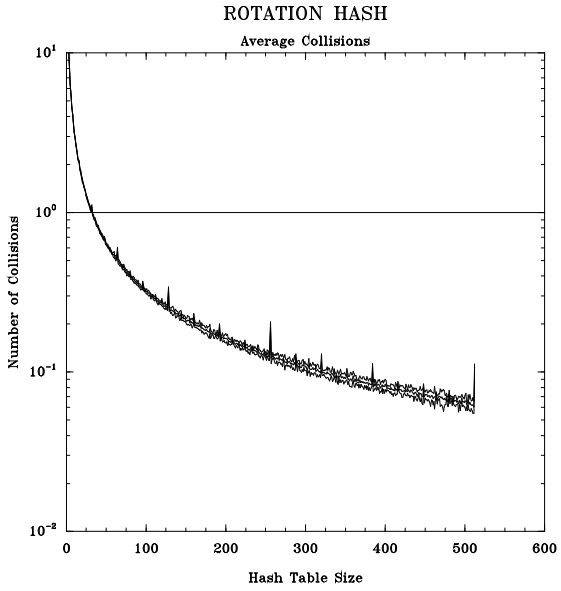


Figure 5: Collisions For Rotation Hash Function

of a power of two, the modulo operation acts more to discard bits than to mix them. In fact, when the modulus is an exact power of two, the modulo operation simply discards the high-order bits.

This non-ideal behavior of the rotation hash can be avoided simply by avoiding queue header arrays sizes that are small integral multiples of powers of two.

5.2 Parametric Studies

The object of the simulation is to determine whether the behavior of stochastic fairness queuing is a good approximation of that of fair queuing. A very simple simulation suffices for this purpose. The simulation consists of a single node with four saturated input lines and one output line. All packets are pure datagrams of equal length; no transport-layer protocol was simulated. There are 20 conversations, one of which is ill-behaved, generating as much input traffic on the average as the other 19 combined. During each time interval, one packet departs from the node and four are offered to the node. The conversation to which a given input packet belongs is randomly chosen.

Each queue making up the stochastic fairness queue is a finite FCFS queue, and a perturbable variant of the HDLC CRC is used as the hash function. Hash function switching is done in such a way as to avoid packet reordering; newly occupied queues are appended to the end of the active list, and buffers are stolen from the beginning of the longest queue to accommodate packets that arrive when the buffer pool is exhausted. A per-conversation fairness policy is used, and the fairness granularity is irrelevant, since all packets are of equal size.

The baseline stochastic fairness queuing run used 160 queues (eight times

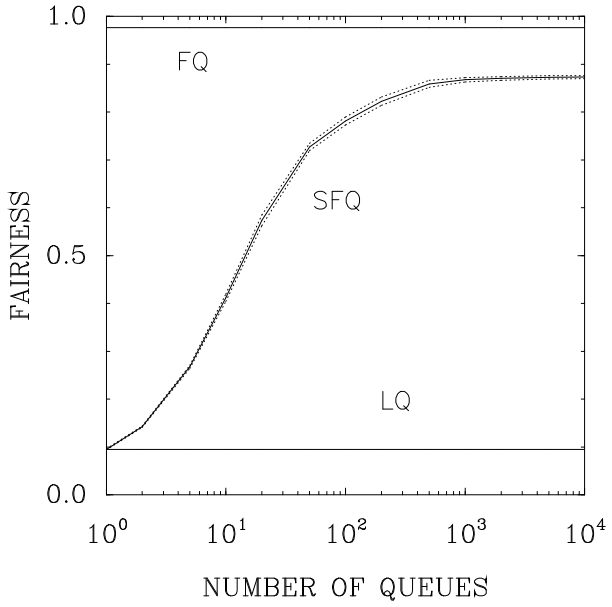
the number of conversations), each with a maximum length of five packets, a buffer pool containing space for up to 160 packets, and a hash function switch interval such that the hash function was perturbed for every 1000 input packets. The simulation runs until 10,000 packets have been offered to the node,[11] at which point the number of packets output per conversation is printed.

The figure of merit used to analyze the results is the ratio of the bandwidth granted to the least-fortunate conversation to that granted to the most-fortunate conversation. A perfectly fair algorithm will have a fairness of one (since it will treat all conversations exactly equally). As points of reference, the fairness of fair queuing, baseline stochastic fairness queuing, and of a length-five FCFS queue[12] are 0.98, 0.81, and 0.095 packets per conversation, respectively.

Increasing the number of queues in the stochastic fairness queue increased its performance, as shown in Figure 6. The lines labeled "SFQ" show the mean value of the fairness for stochastic fairness queuing taken over five runs and the 95% confidence interval. The lines labelled "FQ" and "LQ" show the performance of fair queuing and length-five FCFS queuing, respectively. The 95% confidence bounds for FQ and LQ are barely wider than the line itself, and are not shown. The performance of SFQ should converge to that of LQ as the number of queues approaches one, and should converge to that of FQ as the number of queues grows without bound. The

[11] Or, equivalently, until about 2500 packets (125 per conversation, on the average) have been output from the stochastic fairness queue (since the node is running under four-times overload).

[12] Keep in mind that there is no end-to-end protocol action in this simulation, so the length of the FCFS queue does not affect the results.



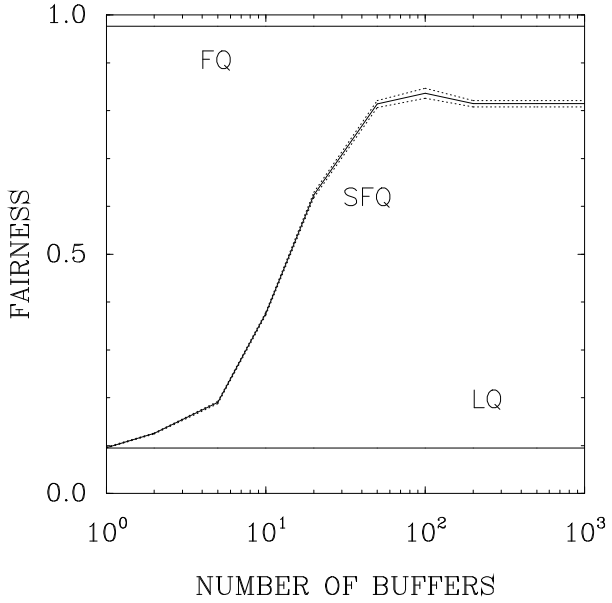
FQ –Fairness Queue

LQ –Limited FCFS Queue

SFQ–Stochastic Fairness Queue

Figure 6: Effect of Number of Queues

fairness for stochastic fairness queuing given 1000 queues (50 times the number of conversations) is 0.86, or 88% of that of pure fair queuing and over nine times that of FCFS queuing.



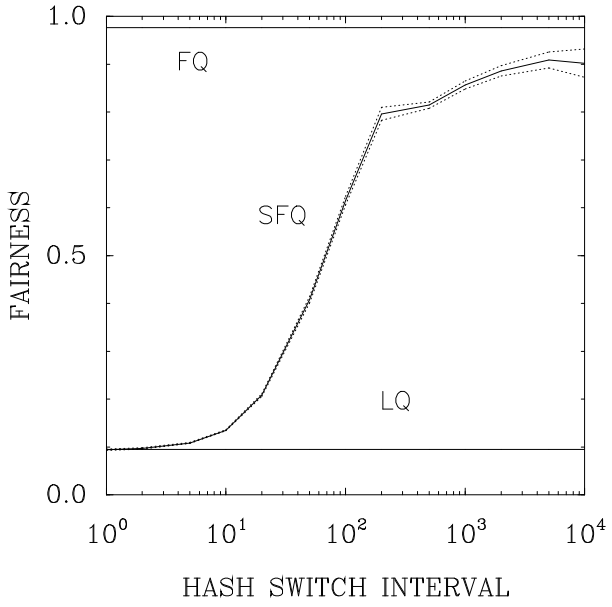
FQ –Fairness Queue

LQ –Limited FCFS Queue

SFQ–Stochastic Fairness Queue

Figure 7: Effect of Number of Buffers

Increasing the number of buffers in the stochastic fairness queue increased its performance, as shown in Figure 7. The lines labelled "SFQ" show the mean value of the fairness for stochastic fairness queuing taken over five runs and the 95% confidence interval. Since each queue can contain five packets at most, and there can be 20 queues at most (one per conversation), the algorithm can make use of 100 buffers at most. This can be seen in Figure 7; the performance levels off above 100 buffers.



FQ –Fairness Queue

LQ –Limited FCFS Queue

SFQ–Stochastic Fairness Queue

Figure 8: Effect of Hash Function Switch Interval

Increasing the length of the hash function switching interval increases performance up to a point, due to the fact that the transient unfairness associated with the switchover occurs less frequently.[13] However, as the duration of the switching interval approaches the average length of the conversation, performance starts decreasing, as can be seen in Figure 8. This is due to the fact that conversations that collide do so for most of their lifetime. Good values for the switch interval appear to lie between twice the queue-flush time of the stochastic fairness queue and one-tenth of the average conversation duration.

Additional simulations were run with buffer stealing disallowed and with a hash-function switching method that allows some packet reordering. Both of these modifications allow more CPU-efficient implementations. Disallowing buffer stealing has very little effect, given enough memory for each conversation to have a full queue. The faster hash-function switching method has almost no effect on fairness, but is more compatible with existing transport protocols, as will be seen in the next section.

These results demonstrate that the performance of stochastic fairness queuing can approach that of pure fair queuing and greatly exceed that of FCFS queuing.

5.3 Transport Protocol Studies

Transport protocol studies were performed using the REAL network simulation package [Kes89]. The scenarios described in Shenker et al. [DKS89]

[13] The transient unfairness is caused by the fact that an ill-behaved conversation can have two queues at its disposal during the switchover, while the well-behaved conversations will limit themselves to a throughput appropriate for a single queue.

were run using a version of stochastic fairness queuing that used buffer stealing, a per-conversation fairness policy, and a per-byte fairness policy. An efficient hash-function switching method was used that allowed packets to be reordered, as the method that avoids reordering can cause occasional packet loss from low-throughput conversations. This packet loss severely decreases the throughput when those conversions are using TCP with Van Jacobson's modifications. Runs were made using HDLC CRC and using the software algorithm described earlier for the hash function: as expected, essentially identical results were obtained. Except as noted below, the FTP throughput results for stochastic fairness queuing were within ten percent of those for fair queuing.

The labels appearing in the "Policy" column of the results give the transport protocol and the queuing discipline, separated by a slash. "G" indicates a generic TCP and "VJ" indicates TCP with Van Jacobson's modifications. "FCFS", "FQ", and "SFQ" indicate first-come-first-served queuing, fair queuing, and stochastic fairness queuing, respectively. Results for the G/FCFS and G/FQ queuing disciplines are taken from Reference [DKS89].

Fairness queuing can provide lower delay to small packets than can stochastic fairness queuing, as a consequence of the greater amount of state maintained by the former. For example, in scenario 1 (an underloaded gateway passing two FTP and two Telnet conversations) fair queuing provides the Telnet conversations about 17 times lower delay than it provides to the FTP conversations, while stochastic fairness queuing provides roughly a factor of four improvement in delay; see Table 2.

None of the scenarios included Telnet and FTP conversations sharing a large bandwidth-delay-product link; it seems likely that this would greatly

Quantity	Policy	FTP		Telnet	
		1	2	3	4
Throughput (packets)	G/FCFS	1746	1746	99	96
	G/FQ	1746	1746	102	94
	G/SFQ	1735	1757	98	97
Average Roundtrip Time	G/FCFS	1.43	1.43	1.36	1.35
	G/FQ	1.43	1.43	0.08	0.09
	G/SFQ	1.28	1.26	0.36	0.36

Table 2: Scenario 1: Underloaded Gateway

reduce the importance of queuing delay.

Fairness queuing includes a mechanism that actively punishes conversations perceived to be malicious. Since stochastic fairness queuing does not attempt to judge users' intents, the results of scenario 3 (an overloaded gateway passing one well-behaved FTP, one well-behaved Telnet, and one ill-behaved FTP) differ significantly; see Table 3. Fairness queuing almost completely shuts down the ill-behaved FTP. However, stochastic fairness queuing grants both FTPs roughly equal bandwidth; the ill-behaved FTP gets about 20% more bandwidth in exchange for a packet-loss rate of over 95%.

These results demonstrate that stochastic fairness queuing works well in the presence of real-world transport protocols.

Quantity	Queue	FTP	Telnet	Ill-Behaved
		1	2	3
Throughput (packets)	G/FCFS	3	11	3497
	G/FQ	3491	95	5
	G/SFQ	1613	93	1883
Average Roundtrip Time	G/FCFS	1362.00	2.87	2.97
	G/FQ	0.72	0.08	903.00
	G/SFQ	1.28	0.44	2.08

Table 3: Scenario 3: Ill-Behaved Source

6 Alternative Implementations

The stochastic fairness queuing algorithms span a broad range of CPU, memory, and fairness tradeoffs. This allows an algorithm to be configured for a specific environment or range of environments. Configuration consists of selecting values for tuning parameters and structural parameters.

The tuning parameters are (1) number of queues; (2) number of buffers; (3) maximum queue length; and (4) hash function switching interval.

Each of these parameters was a command-line argument to the simulation program; the observed effects are described in Section 5. Note that some protocols (in particular, older versions of the Network File System protocol) place restrictions on the maximum queue length; if the queue length is too short, very poor performance will result [Hed89].

The structural parameters include (1) queuing discipline; (2) hash function; (3) hash function switching method; (4) active-list insertion policy; (5) buffer-theft policy; (6) fairness granularity; and (7) fairness policy.

The queuing discipline used in the simulation is finite FCFS; a packet that arrives while its queue is full is discarded. Alternative queuing disciplines include the various forms of random-drop queues.

As mentioned earlier, the simulation uses a variant of HDLC CRC and a simple rotate-and-add function as the hash functions. There are many possible alternative hash functions, including the Fletcher checksum used in OSI, other types of CRC, and various ad hoc functions consisting of sequences of simple functions such as shifts, adds, and exclusive-ORs.

The hash-function switching method simply perturbs the hash function without modifying the packet queues. This can result in packet reordering. Experiments with more complex hash-function switching mechanisms that carefully avoided packet reordering performed poorly when used with buffer-stealing and Van Jacobson's TCP. This is due to the fact that any method that avoids packet reordering while still maintaining O(1) performance must segregate the packets into one group of all packets arriving before the most recent switch, and those arriving after. While there are packets in both groups, the feedback loop from queue length (among the old packets being output) to the sender is broken. This results in significant numbers of packets being dropped from low-throughput conversations, which in turn results in TCP unnecessarily decreasing throughput. Ill-behaved conversations will of course ignore any drops and continue transmission at full speed.

The simulated algorithm always appends newly occupied queues to the end of the active list. An alternative method would be to probabilistically insert newly occupied queues containing small packets onto the head of the active list.[14] This would grant smaller average delay to small packets (which

[14]Consistently inserting newly occupied queues onto the head of the active list could

tend to be Telnet and Transmission Control Protocol [TCP] acknowledgement packets), but would increase the complexity of the algorithm.

The buffer-theft policy implemented in the simulation is to always remove the next packet that would have been output from the longest queue. Alternative policies include dropping a randomly selected packet from the longest queue and simply refusing to do buffer theft. The latter alternative is particularly attractive, as it allows the number-of-elements lists to be dispensed with, thereby greatly reducing the complexity of the algorithm.[15] Adding a packet to a stochastic fairness queue that does not do buffer theft requires one singly-linked-list operation, with an additional doubly-linked-list operation if the queue was initially empty. Deleting a packet from a queue also requires one singly-linked-list operation, with an additional doubly-linked-list operation if this was the last packet in the queue. Queues tend to be long under heavy load, so this configuration of stochastic fairness queuing actually consumes *fewer* CPU resources when heavily loaded. On the other hand, this method is likely to require more buffers than the method simulated to achieve a comparable level of fairness.

The simulation used a packet fairness granularity: that is, a single packet is output from each queue, regardless of packet length. An alternative would be an approximate bit or byte fairness granularity. This can be implemented efficiently using an approach similar to that used in Shenker's fair queuing algorithm [DKS89]. This alternative has the advantage of allocating bandwidth more fairly in the presence of differing packet sizes, but increases the

result in indefinite postponement of the other queues.

[15] Of course, the individual queues making up the stochastic fairness queue must be finite queues in order for buffer theft to be safely dispensed with.

complexity of the algorithm, especially since care must be taken to avoid indefinitely postponing packets already queued.

The fairness policy used by the simulation was equal allocation per host conversation. Many other fairness policies can be envisioned, many of which can be implemented by including different fields from the packet header in the hash function. For example, a policy of equal allocation per *network* conversation can be implemented by hashing only the network portions of the source-destination address pair. This fairness policy might be used in transit networks.[16] Another example would be a policy of equal allocation per TCP connection, which could be implemented by including the TCP port numbers in the hash function.[17]

7 Future Work

Additional work needs to be done to evaluate the different possible instances of the stochastic fairness queuing algorithms presented in Section 6. These results would provide the information needed to select the algorithm that best fits a given processor and network architecture.

The high-speed software hash function presented in Section 4.1 relies on a circular rotate instruction that is not present in some machines. This instruction can be easily simulated, but only with a substantial performance

[16] At first glance, this policy seems to have the disadvantage of encouraging institutions to register many different networks to increase their share of bandwidth. The transit networks can prevent this form of abuse by simply refusing to pass routing information for the excess networks.

[17] This policy might have the disadvantage of requiring very large numbers of queues in addition to the blatant (but not unprecedented) layering violation.

penalty. More work is needed to identify a hash function suitable for machines lacking a circular rotate.

The simulations performed to date used either a single-hop network with no protocol action (that is, pure datagram switching) and uniform packets, or a small network with some representative transport protocols. More work needs to be done to determine the effectiveness of SFQ in large networks and in real (as opposed to simulated) gateways in real networks.

8 Conclusions

This paper has presented and analyzed a class of probabilistic variants of Shenker's, et al., fair queuing algorithm (called "stochastic fairness queuing" algorithms) that are suitable for use in high-speed computer communications networks and that span a broad range of CPU, memory, and fairness tradeoffs. A particular instance of this algorithm has been shown to have behavior approaching that of fair queuing (when given sufficient resources), and to exhibit graceful degradation under overload, without sudden failure.

9 Acknowledgements

I owe many thanks to Craig Partridge, Mike Frankel, Phil Karn, Mark Lewis, John Nagle, Richard Ogier, K. K. Ramakrishnan, Vlad Rutenburg, Nachum Shacham, Scott Shenker, Greg Skinner, Lixia Zhang, and to the anonymous referees for discussions, comments, and much constructive criticism. I am indebted to Diane Lee and Mark Lewis for their support of this effort and I am especially grateful to Srinivasan Keshav for integrating a version of stochastic fairness queuing into his REAL network simulation system.

References

[BG87] D. Bertsekas and R. Gallager. *Data Networks.* Prentice-Hall, Inc., 1987.

[DH89] James R. Davin and Andrew T. Heybey. Router algorithms for resource allocation. Technical Report White Paper, MIT Laboratory for Computer Science, July 1989.

[DKS89] Alan Demers, Srinivasan Keshav, and Scott Shenker. Analysis and simulation of a fair queuing algorithm. *SIGCOMM '89*, pages 1–12, 1989.

[Fel89] David C. Feldmeier. Estimated performance of a gateway routing-table cache. Technical Report MIT/LCS/TM-352, MIT Laboratory for Computer Science, March 1989.

[GKP89] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics.* Addison-Wesley, 1989.

[Gro90] COIP Working Group. ST-II specification. Working document for IETF COIP Working Group, April 1990.

[Hah86] E. Hahne. Round robin scheduling for fair flow control in data communications networks. Technical Report LIDS-TH-1631, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, December 1986.

[Hay81] H. Hayden. Voice flow control in integrated packet networks. Technical Report LIDS-TH-1152, MIT Laboratory for Information and Decision Systems, 1981.

[Hed89] Chuck Hedrick. Protocols restrict queue lengths. Informal comments to the IAB Performance and Congestion Control Working Group, July 1989.

[Int79] International Organization for Standardization. *Data Communication – High-Level Data Link Control Procedures – Frame Structure*, 1979.

[Jac88] Van Jacobson. Congestion avoidance and control. In *SIGCOMM '88*, pages 314–329, August 1988.

[Jai89] Raj Jain. A comparison of hashing schemes for address lookup in computer networks. Technical Report DEC-TR-593, Digital Equipment Corporation, February 1989.

[JR87] Raj Jain and K.K. Ramakrishnan. Congestion avoidance in computer networks with a connectionless network layer. Technical Report DEC-TR-506, Digital Equipment Corporation, Maynard, Massachusetts, August 1987.

[Kes89] Srinivasan Keshav. REAL manuals. Technical Report UCB/TR/89/530, University of California at Berkeley, 1989.

[Knu73] Donald Knuth. *The Art of Computer Programming.* Addison-Wesley, 1973.

[Lou89] Kirk Lougheed. Low-speed lines still heavily used. Informal comments to the IAB Performance and Congestion Control Working Group, July 1989.

[McK89] Paul E. McKenney. High-speed event counting and classification using a dictionary-hash technique. *ICPP '89*, 1989.

[McK90] Paul E. McKenney. Stochastic fairness queuing. In *IEEE INFOCOM'90 Proceedings*, San Francisco, June 1990.

[Nag87] John Nagle. On packet switches with infinite storage. *IEEE Transactions on Communications*, pages 435–438, March 1987.

[Par90] Gurudatta M. Parulkar. The next generation of internetworking. *Computer Communications Review*, 20(1):18–43, January 1990.

[RFS90] John Robinson, Dan Friedman, and Martha Steenstrup. Congestion control in BBN packet-switched networks. *Computer Communications Review*, 20(1):76–90, January 1990.

[Zha89] Lixia Zhang. *A New Architecture for Packet Switching Network Protocols*. PhD thesis, Massachusetts Institute of Technology, July 1989.