

PREEMPTIVE RETRANSMISSION OF BUFFERED DATA IN A NETWORKField of the Invention

The present invention relates to the field of distributed, networked, fault tolerant systems, and more particularly to those systems in which the speed and reliability of data transmission are important.

Background of the Invention

Distributed systems are becoming an increasingly popular means of providing platforms for a variety of functions, including storage controller type functions. Their popularity arises from the flexibility and scalability such systems offer. Fault tolerance is implemented in a number of mutually supporting ways, such as providing redundant network infrastructures, or redundant storage attachments. Distributed applications depend on network connectivity and communications capability to perform their function. These fault tolerance features improve the availability of the system. High availability is also increasingly important for many applications.

Many systems implement retry algorithms for their network interfaces which operate as follows:

1. An error happens, such as a packet being dropped.
2. A timeout interval expires.
3. The network hardware or protocol stack detects the error.
4. The error is reported to the issuer of request.
5. The issuer attempts the request a second time, possibly using alternative hardware.

Such schemes are simple, and are adequate when it is acceptable to wait for the original request to fail. However, for some important environments, bare availability (lack of actual failures while attempting to access a service) is insufficient. In these environments, it is important to receive a response within a certain time. Failure to respond within that time will have a penalty comparable to that of a full-scale failure in accessing the service, for example, another application might time out and return an error condition, or an Internet user might click away to a competitor's web-site in frustration.

Thus, it would be advantageous to provide a method which would enable retries within a desired time limit, but existing systems do not naturally enable this. There are a number of problems to be addressed. First, the system might try to detect an error in a more timely fashion. It might be possible to reduce timeout intervals in interface adapters or hardware, but often there are architectural limits in how low these intervals can be. For example, in Fiber Channel, an exchange that fails to complete normally cannot be reused until the expiry of an error timeout interval defined by the switch, often 10 seconds. Further, many network implementations do not behave robustly if the network error timeout interval is reduced too much.

It might be possible to use some other timeout mechanism -- one not associated with the interface software or hardware that has failed -- to attempt to redrive the request, but this does not solve the problem, because the original request is still active. Any attempt to redrive the original request, using an alternative path offered by redundant hardware, will be blocked until the original request has completed, because there are resources associated with the original request that are still in use.

As a concrete example, where a Fiber Channel adapter is used to implement the transmission interface, and a multithreaded user process is attempting to retransmit a buffer, a second transmission attempt will be blocked by the virtual memory system because the memory is still in use by the original transmission.

Another possible solution might be to attempt to avoid this memory blocking problem by retaining a copy of the transmission data in a private buffer. This copy could then be used to create a second transmission should the first be deemed to have taken too long. The private copy could be accessed without being blocked by the virtual memory manager, but this scheme would increase the cost of every transmission, including the majority which do not encounter a problem, because the data would have to be copied for each transmission before the first transmission attempt is made. Such an additional processing cost is likely to prove unacceptable in most modern networks.

Summary of the Invention

The present invention accordingly provides, in a first aspect, an apparatus for data transmission over a network having multiple paths comprising a buffer for containing a data item to be transmitted to a

receiver; a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval; a first accessor of buffer to access data for a first transmission and to start a first timeout clock set to said interval in said timer mechanism; and a second accessor of said buffer responsive to said signaling by said timer mechanism to access said data for retransmission and to start at least a second timeout clock set to said interval in said timer mechanism, said second accessor attempting to select a path for transmission of said data avoiding use of path elements used by said first accessor.

The apparatus preferably further comprises a reference counter for maintaining a count of the references to said buffer by said first accessor and said second accessor, said reference counter being incremented on each of said references and decremented on completion of each data transmission by said first accessor and said second accessor, said reference counter signaling when said count reaches zero; and a memory manager adapted to permit read access to said buffer by said first accessor and said second accessor and to return said buffer to a free buffer pool responsive to said reference counter signaling when said count reaches zero.

The apparatus preferably further comprises an analysis mechanism for determining said optimum interval for retransmission of said data; and a tuning mechanism for tuning said timing mechanism to said optimum interval for retransmission of said data.

Preferably, said analysis mechanism is operable to use network monitoring data to determine said optimum interval for retransmission of said data.

Preferably, said network comprises a storage network.

Preferably, said network comprises the Internet.

In a second aspect, the present invention provides a storage controller including An apparatus for data transmission over a network having multiple paths comprising a buffer for containing a data item to be transmitted to a receiver; a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver,

said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval; a first accessor of buffer to access data for a first transmission and to start a first timeout clock set to said interval in said timer mechanism; and a second accessor of said buffer responsive to said signaling by said timer mechanism to access said data for retransmission and to start at least a second timeout clock set to said interval in said timer mechanism, said second accessor attempting to select a path for transmission of said data avoiding use of path elements used by said first accessor.

In a third aspect, the present invention provides a network appliance including An apparatus for data transmission over a network having multiple paths comprising a buffer for containing a data item to be transmitted to a receiver; a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval; a first accessor of buffer to access data for a first transmission and to start a first timeout clock set to said interval in said timer mechanism; and a second accessor of said buffer responsive to said signaling by said timer mechanism to access said data for retransmission and to start at least a second timeout clock set to said interval in said timer mechanism, said second accessor attempting to select a path for transmission of said data avoiding use of path elements used by said first accessor.

Preferred features of the storage controller of the second aspect and the network appliance of the third aspect correspond to the respective preferred features of the apparatus of the first aspect.

In a fourth aspect, the present invention provides a method for data transmission over a network having multiple paths comprising the steps of providing a buffer for containing a data item to be transmitted to a receiver;

providing a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval; accessing data in a buffer by a first accessor for a first transmission; starting a first timeout clock set to

said interval in said timer mechanism; accessing data in a buffer by a second accessor responsive to said signaling by said timer mechanism; retransmitting said data by said second accessor; starting a second timeout clock set to said interval in said timer mechanism; and attempting by said second accessor to select a path for transmission of said data avoiding use of path elements used by said first accessor.

The method preferably further comprises the steps of maintaining by a reference counter a count of the references to said buffer by said first accessor and said second accessor, said reference counter being incremented on each of said references and decremented on completion of each data transmission by said first accessor and said second accessor, said reference counter signaling when said count reaches zero; and permitting by a memory manager read access to said buffer by said first accessor and said second accessor and returning said buffer to a free buffer pool responsive to said reference counter signaling when said count reaches zero.

The method preferably further comprises the steps of determining said optimum interval for retransmission of said data; and tuning said timing mechanism to said optimum interval for retransmission of said data.

Preferably, said step of determining uses network monitoring data to determine said optimum interval for retransmission of said data.

Preferably, said network comprises a storage network.

Preferably, said network comprises the Internet.

In a fifth aspect, the present invention provides a computer program comprising computer program code to, when loaded into a computer system and executed, perform the steps of the method of the fourth aspect.

The system is thus preferably structured so that multiple transmissions of data from the same application buffers can be outstanding simultaneously.

Some resources, such as hardware control blocks, are always consumed for the duration of a hung transmission. The system is thus preferably structured so that such resources are segregated into independent transmission paths. In consequence, a hung transmission on one path will not impede the ability to transmit data down a second path.

While a transmission is active (and not acknowledged), a timer is preferably operational to provide early detection of any unusual delay in receiving a response. When the timer threshold is reached, this indicates that there may be an advantage in trying an alternative path. The transmission is thus suitably reissued immediately down the alternative path, without having to wait or block for the first transmission to complete.

It will be clear to one skilled in the art that the interval set for the timer should be selected to optimize performance, by striking a balance, suited to the particular network and its attached devices, between failing to respond quickly enough to a possible transmission failure and, on the other hand, burdening the system with additional retransmissions based on "false positives", when the system responds too frequently and the retransmission overhead is increased too much. In a preferred embodiment of the present invention the interval used may be tuned to set it to the optimum value. In a most preferred embodiment, the timer interval is set based on monitoring of the network so as to ensure that the interval is tuned to the optimum value based on the network performance. In a further or alternative embodiment, quality-of-service requirements may be taken into account in determining the optimum interval.

The preferred embodiment of the present invention further alleviates the problem caused by the fact that conventional virtual memory systems of the prior art do not support multiple concurrent transmissions from application buffers because they are unable to track the progress of the multiple transmissions or police application access to the transmission buffers. In conventional systems of the prior art, once a buffer has been assigned to contain a particular data item, and has had its real address given to a hardware I/O device, it is in effect "frozen" and cannot be accessed by any entity other than the hardware I/O device, even including the operating system, which cannot, for example, be permitted to swap the page containing the buffer out to disk, as the hardware I/O device must be able to rely on the memory reference it has been given for the duration of the transmission. The preferred embodiment of the present invention alleviates this limitation by providing a mechanism for controlling the multiple accesses outside the scope of the hardware I/O device and operating system to permit timely retransmission to improve system performance in spite of network path disruptions.

The preferred embodiment of the invention has the further advantage that it does not require extra copying of the data. In cases in which an alternative path is working, the timely availability of data in the system may be maintained in spite of a disruption or other delay on one of the transmission paths.

Brief Description of the Drawings

A preferred embodiment of the present invention will now be described by way of example only, with reference to the accompanying drawings, in which:

Figure 1 is a block diagram representing an apparatus according to a preferred embodiment of the present invention;

Figure 2a represents an exemplary communications flow according to a preferred embodiment of the present invention;

Figure 2b shows a logic flow according to a preferred embodiment of the present invention; and

Figure 3 represents the logic flow of a further preferred refinement of one embodiment of the present invention.

Detailed Description of the Preferred Embodiment

The presently most preferred embodiment of the present invention is implemented in a distributed storage controller, but it will be clear to one skilled in the art that the present invention can equally advantageously be implemented in other networked systems, including, but not limited, to networks of client and server computer systems. Such systems may be wired or wireless, and may comprise devices having local processing capabilities as well as devices lacking such capabilities, such as simple I/O devices.

In the preferred environment of a distributed storage controller, the memory management component maintains I/O Buffer descriptions for pieces of data up to 32kB in size. The I/O Buffer contains data that was received from some external source, such as data from a host system received as part of a SCSI write operation. The I/O Buffer can be translated to a scatter-gather list of PCI addresses, and these can be used as part of the data transfer instructions to a Fiber Channel adapter.

The memory manager does not block a client to construct the scatter-gather list, or when a transmission is active. The buffer must remain "pinned" (that is, the virtual memory used for the buffer is forced to remain real) as long it has been used by any I/O hardware device and that I/O hardware device's transmission has not yet completed with either definite success or definite failure.

Figure 1 shows an apparatus (102) according to a preferred embodiment of the present invention, having a buffer memory (104) accessible by memory manager (114) responsive to access requests by accessors (106, 110). Memory manager (114) is in communication with counter (108). Counter (108) maintains a count of the number of references to the buffer memory (104) by the accessors (106, 110), the count being incremented on each reference and decremented on completion of each accessor's data transmission. The counter (108) is further adapted to signal when the count reaches zero. Accessors (106, 110) are associated with respective timer clock instances (116, 120). The timer clock instances (116, 120) are provided within a timer mechanism (122), which is adapted for timing a predetermined optimum interval, which is less than the "round trip" time that would be required to return a real error condition from the network. The timer mechanism (122) is further adapted to signal the end of the predetermined optimum interval as set by each timer clock instance (116, 120).

Memory manager (114) is adapted to lock buffer memory during write activity, to permit read access to the buffer memory (104) by accessors (106, 110) and to return the buffer memory to a free buffer pool when counter (108) signals that the count has reached zero.

To track the multiple concurrent processes, the memory manager thus maintains a reference count of concurrent accessors. Concurrent accessors of a memory buffer are allowed to read the data, but not write (change) the data. Consequently, multiple processes can be active concurrently reading the source of the data, provided none of the processes attempts to change the data. As each process completes, it 'frees' its reference to the buffer; the reference count is decremented. When the last process completes, the reference count reaches zero, and the buffer is put into the pool of free buffers.

Thus this memory manager allows multiple processes to transmit the data from the buffer, since transmission does not involve changing the content of data.

The memory manager allows a second or further process to start to transmit the same buffer of data, constructing its own PCI scatter-gather list to access the buffer, even though an original request, or even many prior requests, are still active.

In a preferred embodiment of the present invention, each alternative path is chosen so as to offer reasonable confidence that it is likely to be more successful than the first or other earlier paths. This is ensured by selecting the alternative paths from a completely separate set of hardware elements, for example, a separate interface adapter, or a different physical network. Paths may be further selected using additional criteria, such as available bandwidth, response times assessed from ping responses during periods when the controller or appliance was idle, and the like.

In a further preferred embodiment, an interval tuner (126) is used to tune the interval used by clocks (116, 120), based on analysis of data by a network monitoring data analyzer (128).

The memory manager of the preferred embodiment, combined with a timer mechanism adapted to optimize performance by balancing the need for rapid retransmission in the case of a probable failed transmission against the need to avoid excessive retransmission overhead, thus allows timely retransmission of the data from a buffer. At the same time, the counter of the preferred embodiment allows the buffer memory to be allocated, "pinned", and freed without preventing read accesses by further accessors.

The mechanism of the preferred embodiment requires a protocol be implemented at the receiver of the transmission to enable it to cope with duplicate receptions of data, in case a timed-out transmission is delayed and not permanently lost. Such schemes are well understood in the art, and are already implemented in protocols such as TCP/IP, and so they will not be discussed further here.

Turning to Figure 2a there is shown an example of a simple communications flow illustrative of the method according to a preferred embodiment of the present invention.

The communications flow involves a timer, a memory manager, accessors 1 and 2 (further accessors 3,4...n are also shown to indicate that a plurality of such accessors may exist), and a network. The flows will be defined in terms of a set of times T1, T2 and so on. The

communications flow illustrates an exemplary case in which a first accessor times out and a second accessor is permitted to access the same buffer in order to retransmit the data. Both the flows complete (either with successful receipt of the data or with a definite error return).

At T1, a first accessor accesses a buffer, and the counter is incremented to record its access. At T2, the clock is started to count down for the predetermined optimum timeout interval. The buffer is pinned at T3, and at T4 accessor 1 transmits the data over the network. At T5, accessor 1's timeout interval expires, and accessor 2 is triggered to access the buffer. This occurs at T6, and the count is incremented to record the access. At T7, the clock is started to count down to timeout for accessor 2. At T8, the buffer is pinned, and at T9, accessor 2 transmits the data. At T10, accessor 1's transmission is acknowledged by the network and accessor 1 releases its access to the buffer. The count is decremented to record the end of accessor 1's access. At T12, accessor 2's transmission is acknowledged by the network. At T13 accessor 1 releases its access to the buffer and the count is decremented. At T14, the count being zero, the buffer is returned to the free buffer pool. It will be clear to one skilled in the art that this is a much simplified example created solely to set forth the basic set of flows that can occur within the preferred embodiment.

The logic flow of Figure 2b shows a method for transmission of data according to a preferred embodiment where there is provided a buffer memory (104), a memory manager (114) adapted to lock buffer memory (104) during write activity, to permit read access to the buffer memory (104) and to return the buffer memory (104) to a pool of free buffer memory, and a timer mechanism for timing a predetermined optimum interval for retransmission of said data, said interval being less than that required for receipt of any error signal from said network, and for signaling the end of said predetermined optimum interval.

At step (202), the runtime logic process according to the presently preferred embodiment is entered. At step (204), a first accessor accesses data in a buffer memory for transmission. At step (206), the count is incremented to record the accessing of the buffer memory (104) by the first accessor, and at step (208) the timeout clock is set to the predetermined optimum interval for retransmission of the data, that interval being less than the round-trip time that would be required for receipt of any error signal from the network.

At step (212), the system waits for either notification from the network that a transmission has reached completion (whether a successful receipt of the transmission or a definite failure signal from the network) or the expiry of the timeout interval. At step 214, a test for timeout is performed, and if there has been no completion received before expiry of the timeout interval, the logic initiates step (216), in which the clock is stopped. At step (218) a second or further accessor accesses the same data in the same buffer memory for retransmission over the network. At step (220) the second or further accessor selects an alternative path for transmission of the data, the step of selecting being adapted to select a complete set of different path elements if possible, and if that is not possible, to select as many alternative path elements as possible. Each non-first accessor attempts to avoid use of path elements used by any prior accessor.

The alternative path might be selected further by means of an optimizing logic element to select, for example, the least transmission cost path or the highest performance path, or perhaps the path offering the best quality-of-service guarantees. Such quality-of-service guarantees are well known in the network communications art and need not be described here. In a presently preferred embodiment, the alternative path is selected on the basis of an analysis of network performance monitoring statistics.

The logic flow proceeds to step (206) where the count is incremented to record the step of accessing the buffer memory by the second or further accessor and continues as before to set the clock, etc.

On any iteration at step (214) if there is a positive response to the test indicating that there has been a completion (whether a successful receipt of the transmission or a definite failure signal from the network) before expiry of its respective transmission's timeout, the clock is stopped at step (222) and the count is decremented at step (224). At step (226), a further test is performed to determine whether the count has reached zero or not. If it has not reached zero, this part of the logic process returns to step (228) to be triggered by the next completion. If on any iteration, the count is determined to have reached zero, the memory manager (114) releases the buffer memory (104).

Turning now to Figure 3, there is shown a further preferred refinement of the exemplary embodiment. The logic process of Figure 3 starts at step (302) and at step (304) receives network monitoring data.

Such data may be provided by any of numerous network monitoring devices, systems or computer programs well known in the art. From this, an assessment is made at step (306) of the round-trip response time that can be expected, and at step (308) any quality-of-service parameters that have been set are examined. Based on these inputs, at step (310) the optimum retransmission timeout interval is set. At step (312) this part of the process ends. As will be clear to one skilled in the art, this process may be iterative, and may be performed at intervals to reset the optimum interval according to the most recent data received from the network and from any setter of quality-of-service parameters.

CLAIMS

1. An apparatus for data transmission over a network having multiple paths comprising:

a buffer for containing a data item to be transmitted to a receiver;

a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval;

a first accessor of buffer to access data for a first transmission and to start a first timeout clock set to said interval in said timer mechanism; and

a second accessor of said buffer responsive to said signaling by said timer mechanism to access said data for retransmission and to start at least a second timeout clock set to said interval in said timer mechanism, said second accessor attempting to select a path for transmission of said data avoiding use of path elements used by said first accessor.

2. The apparatus as claimed in claim 1, further comprising:

a reference counter for maintaining a count of the references to said buffer by said first accessor and said second accessor, said reference counter being incremented on each of said references and decremented on completion of each data transmission by said first accessor and said second accessor, said reference counter signaling when said count reaches zero; and

a memory manager adapted to permit read access to said buffer by said first accessor and said second accessor and to return said buffer to a free buffer pool responsive to said reference counter signaling when said count reaches zero.

3. The apparatus as claimed in claim 1, further comprising:

an analysis mechanism for determining said optimum interval for retransmission of said data; and

a tuning mechanism for tuning said timing mechanism to said optimum interval for retransmission of said data.

4. The apparatus as claimed in claim 3, wherein said analysis mechanism is operable to use network monitoring data to determine said optimum interval for retransmission of said data.

5. A storage controller including an apparatus for data transmission over a network having multiple paths comprising:

a buffer for containing a data item to be transmitted to a receiver;

a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval;

a first accessor of buffer to access data for a first transmission and to start a first timeout clock set to said interval in said timer mechanism; and

a second accessor of said buffer responsive to said signaling by said timer mechanism to access said data for retransmission and to start at least a second timeout clock set to said interval in said timer mechanism, said second accessor attempting to select a path for transmission of said data avoiding use of path elements used by said first accessor.

6. The storage controller as claimed in claim 5, further comprising:

a reference counter for maintaining a count of the references to said buffer by said first accessor and said second accessor, said reference counter being incremented on each of said references and decremented on completion of each data transmission by said first accessor and said second accessor, said reference counter signaling when said count reaches zero;

a memory manager adapted to permit read access to said buffer by said first accessor and said second accessor and to return said buffer to a free buffer pool responsive to said reference counter signaling when said count reaches zero.

7. The storage controller as claimed in claim 5, further comprising:

an analysis mechanism for determining said optimum interval for retransmission of said data;

a tuning mechanism for tuning said timing mechanism to said optimum interval for retransmission of said data.

8. A network appliance including an apparatus for data transmission over a network having multiple paths comprising:

a buffer for containing a data item to be transmitted to a receiver;

a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval;

a first accessor of buffer to access data for a first transmission and to start a first timeout clock set to said interval in said timer mechanism; and

a second accessor of said buffer responsive to said signaling by said timer mechanism to access said data for retransmission and to start at least a second timeout clock set to said interval in said timer mechanism, said second accessor attempting to select a path for transmission of said data avoiding use of path elements used by said first accessor.

9. A method for data transmission over a network having multiple paths comprising:

providing a buffer for containing a data item to be transmitted to a receiver;

providing a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval;

accessing data in a buffer by a first accessor for a first transmission;

starting a first timeout clock set to said interval in said timer mechanism;

accessing data in a buffer by a second accessor responsive to said signaling by said timer mechanism;

retransmitting said data by said second accessor;

starting a second timeout clock set to said interval in said timer mechanism; and

attempting by said second accessor to select a path for transmission of said data avoiding use of path elements used by said first accessor.

10. A computer program product tangibly embodied in a computer readable medium and comprising computer program code means to, when loaded into a computer system and executed, perform the steps of:

providing a buffer for containing a data item to be transmitted to a receiver;

providing a timer mechanism for timing a predetermined optimum interval for retransmission of said data item to said receiver, said interval being less than that required for receipt of any error signal from said network, said timer mechanism further signaling the end of said predetermined optimum interval;

accessing data in a buffer by a first accessor for a first transmission;

starting a first timeout clock set to said interval in said timer mechanism;

accessing data in a buffer by a second accessor responsive to said signaling by said timer mechanism;

retransmitting said data by said second accessor;

starting a second timeout clock set to said interval in said timer mechanism; and

attempting by said second accessor to select a path for transmission of said data avoiding use of path elements used by said first accessor.