



- (51) **International Patent Classification:**
G06F 3/00 (2006.01)
- (21) **International Application Number:**
PCT/US2017/022208
- (22) **International Filing Date:**
13 March 2017 (13.03.2017)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
62/306,966 11 March 2016 (11.03.2016) US
- (71) **Applicant:** PURDUE RESEARCH FOUNDATION
[US/US]; 1281 Win Henschel Blvd., West Lafayette, IN 47906 (US).
- (72) **Inventors:** THOTTETHODI, Mithuna, Shamabhat;
1032 Onyx Street, West Lafayette, IN 47906 (US). VI-JAYKUMAR, Terani, N.; 1694 Fernleaf Drive, West Lafayette, IN 47906 (US). XUE, Jiachen; 465 Northwestern Ave, West Lafayette, IN 47907 (US).
- (74) **Agent:** JALAIE, Bobak, P.; Purdue Research Foundation,
1281 Win Henschel Blvd., West Lafayette, IN 47906 (US).
- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

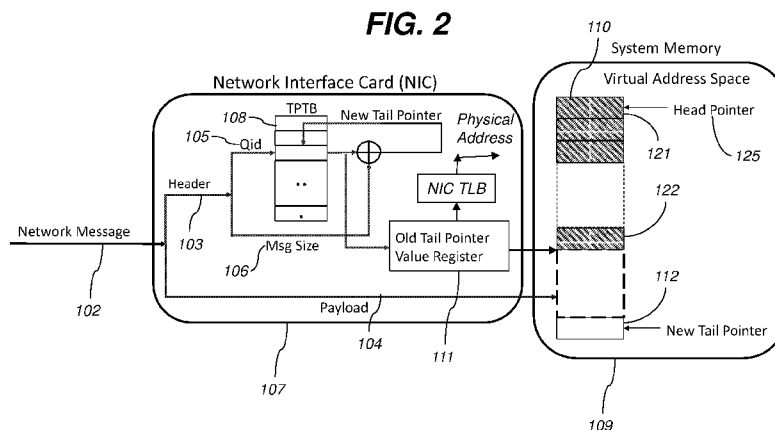
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

(54) **Title:** COMPUTER REMOTE INDIRECT MEMORY ACCESS SYSTEM



(57) **Abstract:** A remote indirect memory access system and method for networked computer servers. The system comprises a network interface card having a network interface memory and a system memory operatively connected to the network interface card. The system memory has a plurality of electronic memory queues, wherein each of the memory queues corresponds to one of a plurality of receive processes in the computer server, with each of the memory queues having a corresponding head pointer and tail pointer. Each of the memory queues is assigned to receive electronic messages from a plurality of sender computers. The NIC comprises a tail pointer table, with the tail pointer table comprising initial memory location data of the tail pointers for the memory queues. The memory location data referenced by corresponding queue identifiers.

COMPUTER REMOTE INDIRECT MEMORY ACCESS SYSTEM

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present patent application is related to and claims the priority benefit of U.S. Provisional Patent Application Serial No. 62/306966, filed March 11, 2016, the contents of which is hereby incorporated by reference in its entirety into the present disclosure.

TECHNICAL FIELD

[0002] The present application relates to computer communication buffer management, and more specifically, to a remote indirect memory access system.

BACKGROUND

[0003] Data center network performance is critical for most modern web/cloud services whose backends run within data centers. While the dominant TCP/IP over Ethernet model has many advantages including scalability and stability, it suffers from significant latency (e.g., milliseconds). In contrast, remote direct memory access (RDMA), such as RDMA over Infiniband and RDMA over converged Ethernet (RoCE) are attractive for their low latency (e.g., microseconds). Unfortunately, there remain several challenges in achieving high performance in an efficient manner.

[0004] RDMA's basic primitives, while well-suited for high-performance computing (HPC) where there are a limited number of communication nodes, are a mismatch for the communication typical in modern data centers. Specifically, the request-response model of data centers must accommodate the following traffic constraints. First, there are unexpected sources of data, which can be from any source. This is because of sharding of content, where objects may be widely distributed across many servers any of which may respond (e.g., data may be sharded across 5000 servers in a modern data center). Second, there are unexpected message sizes (i.e., the messages can be of any size). This

is because the key/values that are exchanged vary from small text updates (tens/hundreds of bytes) to large videos/images (MBs). Third, there is unexpected burstiness (any-number). Message arrival in modern data centers may be bursty with the peak rate limited only by the line rate (e.g., 56Gb/s for Infiniband FDR). In contrast, HPC workloads typically have fewer sources and less variation in message sizes. Even in HPC, when the number of potential sources increase, the challenges of RDMA limit its effectiveness.

[0005] Memory allocation to hold incoming packets is a key challenge in managing communication. Transmission Control Protocol (TCP) uses shared buffers to temporarily hold packets of multiple messages, which are then delivered by the operating system (OS) to the appropriate process. Because the buffers are shared across processes, the amount of memory allocated for the buffers can be small but OS involvement is necessary which is slow. In contrast, under the traditional RDMA read/write model, memory for private buffers is proactively pre-allocated and pinned via memory registration. The address of such pinned memory is exchanged out-of-band which then enables fast read/write from remote servers without OS involvement. This strategy effectively means that memory regions must be pre-allocated and pinned per sender; sharing of a region among multiple senders is not possible as one sender's data may be overwritten by another before being read (due to the any-source constraint). Further, because of any-size, and any-number constraints, the allocated memory must be large. Effectively, handling the above constraints via proactive pre-allocation must provision for the worst case of any possible communication which results in large, grossly-underutilized memory pre-allocation.

[0006] In contrast to RDMA's read/write model, the send/receive model requires the receiver to proactively post pre-allocated per-message memory buffers before the messages arrive. The semantics of send/receive is that newly-received messages are copied to the next unutilized buffer at the head of a receive queue. Because distinct messages are copied to distinct memory buffers, sharing of the receive queue is possible

(un- like in RDMA read/write where writes to a given address do overwrite one another) which alleviates the any-source constraint. The Infiniband Shared Receive Queue (SRQ) has been introduced specifically for this purpose. Further, as long as enough memory is pre-allocated and posted to the receive queues to handle the line rate, one can be sure of not being overrun by incoming messages (thus handling the any-number constraint). Unfortunately, the any-size constraint continues to pose a problem because the message must fit in the pre-allocated buffer. In fact, a single Shared Receive Queue can be used to handle only a single message size because posted memory buffers are consumed in a strictly first- in-first-out order.

[0007] A naive approach, which frees the programmer from reasoning about the any-size constraint, is to use message buffers of the maximum possible message size (e.g., 1MB). While easy for the programmer, this approach has serious disadvantages because of a huge memory footprint (which is typically underutilized) and the corresponding performance impact (e.g., NIC TLB misses, opportunity cost of reduced memory for the server). An alternative (and realistic) approach to avoid the huge overhead requires the programmer to handle communication via multiple SRQs, each of a different size. Given a wide spread of sizes, the number of SRQs must be tuned to minimize the memory overhead to a manageable level. While clearly better than the naive approach in terms of memory footprint, testing shows that as many as 42 distinct SRQs are needed to reduce the over-head (the number of slab sizes used in Facebook's memcached memory allocator). Therefore, improvements are needed in the field.

SUMMARY

[0008] The present disclosure provides remote indirect memory access (RIMA) in a computer server, which utilizes append semantics to shared queues. The computer server is typically located in a data center (DC), and receives messages from other sender computers in the data center. The RIMA's primary 'verb' is 'append' which uses a tail pointer to a shared region of system memory in the server. At setup, the senders receive from the computer server the ID of the shared queue to which they can append messages. At the receiver end, a network interface card (NIC) in the receiving computer server uses the ID to identify the queue's tail pointer, and atomically updates the tail pointer and appends the incoming message to the newly allocated space (only the update is atomic; the message append is not). As in traditional RDMA, the network interface card of the receiving computer server is responsible for pointer lookup, address translation and enforcing virtual memory protections. This indirection of specifying a queue ID (and not the actual tail-pointer virtual address, which remains hidden from senders) handles the typical DC traffic of any sender sending any number of messages of any size to a shared queue. At the same level of programmer effort, RIMA achieves significantly smaller memory footprint than Infiniband SRQ whose huge memory footprint can degrade performance and increase cost. On the other hand, while SRQ can be crafted to minimize memory footprint by expending significant programming effort, RIMA provides those benefits with no additional programmer effort.

[0009] In certain embodiments, the disclosed RIMA method decouples the memory allocation into two parts. The heavyweight allocation, as visible to the heap-manager and the operating system (OS), is done *a priori* to accommodate enough space given the line rate. Thus, the online allocation work is limited to atomically incrementing the tail pointer. To make sure that there is adequate allocated memory, the disclosed method uses a water-mark based approach which increases the memory allocation whenever the available space is running low, and decreases the allocation when message queues are

drained. As such, queue sizes grow and shrink as needed without exposing the latency of memory allocation to the queue management.

[0010] According to one embodiment, a computer server is disclosed, comprising a network interface card having a network interface memory located on the network interface card, and a system memory operatively connected to the network interface card, the system memory having a plurality of electronic memory queues, each of the memory queues corresponding to one of a plurality of receive processes in the computer server, each of the memory queues having a corresponding head pointer and tail pointer, wherein each of the memory queues is assigned to receive electronic messages from a plurality of sender computers. The NIC comprises a tail pointer table, the tail pointer table comprising initial memory location data of the tail pointers for the memory queues, the memory location data referenced by corresponding queue identifiers.

[0011] According to a further embodiment, a method of managing incoming network messages in a computer server is disclosed, comprising establishing a plurality of electronic memory queues in a system memory of the computer server, each of the memory queues corresponding to one of a plurality of receive processes in the computer server, each of the memory queues having a corresponding head pointer and tail pointer, wherein each of the memory queues is assigned to receive electronic messages from a plurality of sender computers, and establishing a tail pointer table in a network interface card (NIC) of the computer server, the tail pointer table comprising initial memory location data of the tail pointers for the memory queues, the memory location data referenced by corresponding queue identifiers.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] In the following description and drawings, identical reference numerals have been used, where possible, to designate identical features that are common to the drawings.

[0013] FIG. 1 is a diagram illustrating a plurality of computer servers connected by a network in a data center according to various aspects.

[0014] FIG. 2 is a diagram illustrating a network interface card and a system memory connected within a computer server according to various aspects.

[0015] The attached drawings are for purposes of illustration and are not necessarily to scale.

DETAILED DESCRIPTION

[0016] In the following description, some aspects will be described in terms that would ordinarily be implemented as software programs. Those skilled in the art will readily recognize that the equivalent of such software can also be constructed in hardware, firmware, or micro-code. Because data-manipulation algorithms and systems are well known, the present description will be directed in particular to algorithms and systems forming part of, or cooperating more directly with, systems and methods described herein. Other aspects of such algorithms and systems, and hardware or software for producing and otherwise processing the signals involved therewith, not specifically shown or described herein, are selected from such systems, algorithms, components, and elements known in the art. Given the systems and methods as described herein, software not specifically shown, suggested, or described herein that is useful for implementation of any aspect is conventional and within the ordinary skill in such arts.

[0017] According to one embodiment, the disclosed RIMA system (1) implements a layer of indirection in specifying the destination memory address for messages sent to a computer server over a computer network, and (2) re-actively allocates enough memory at the tail of the queue to accommodate incoming messages.

[0018] To implement the RIMA functionality, an interface is provided which allows the user customize the system to create and destroy RIMA queues at the receiver computer server. The queue identifier (or handle) is communicated out- of-band (possible over TCP/IP) to potential sender computers. Each queue offers a single remote operation: sending a message of arbitrary size to a remote queue on the target receiver computer server. Receiver computers monitor a completion queue to monitor message arrival. Upon seeing a completion notice, the application at the receiver computer end may read the message from the receive-buffer.

[0019] Using the disclosed RIMA system (unlike prior art SRQ systems), senders are free to intermingle messages of any size. Further, applications do not need any

synchronization to coordinate with other senders (unlike in RDMA read/write) who may be sending to the same queue.

[0020] Although RIMA's independence from message size means that it does not need multiple queues, programmers are free to use multiple queues for logically distinct communication channels. For example, in certain embodiments, the system may be configured or programmed to store control messages (e.g., heartbeat) and data messages (e.g., objects from the database) on separate queues.

[0021] Table 3 below describes an example application programming interface (API) for managing the disclosed RIMA system "append queue" at the local node according to one embodiment. The RIMA append queue (AQ) is a substitute of the SRQ. When the sender posts send using the RIMA interface, it uses the newly introduced IBV_WR_APPEND opcode instead of the old IBV_WR_SEND opcode. Pre-posting receives is no longer required at the receiver's side. This table shows one example API. Other APIs to achieve the disclosed RIMA system's append queue may also be used.

Table 3: RIMA's API for managing append queue resource.

ibv_create_aq(...);	Creating an append queue with initial attributes
ibv_destroy_aq(...);	Destroy an append queue
ibv_query_aq(...);	Query attributes of queue
ibv_modify_aq(...);	Modify attributes of queue

[0022] FIG. 1 shows a data center 100 having a plurality of computer servers 120 connected by a network 122 according to one embodiment. Although only two servers 120 are shown in FIG. 1, the data center 100 may have many more servers 120. The

network 122 may be configured to support Infiniband or RoCE (RDMA over converged Ethernet) for sending and receiving messages between the servers 120. As shown, the computer servers may include a network interface card (NIC) 107, a system central processing unit (CPU) 113, and a system memory 109. The components 107, 112, and 109 may be communicatively connected by a system bus (e.g., a PCI or PCIe bus). The NIC 107 may also comprise a network interface memory 115 and a processing unit 117 operatively connected to the network interface memory 115.

[0023] FIG. 2 illustrates one embodiment of a remote indirect memory access system. Append queues 110 are created in the system memory 109. Each append queue 110 has a head pointer 125 which references memory location 121 as shown (where the oldest message is stored) and a tail pointer 111 which references memory location 122 (where the next incoming message will be stored). The append queues 110 may be optionally created in virtual memory addresses which map a continuous address space to physical system memory locations. The append queues 110 may also be implemented as circular queues, which map a fixed-size memory space as if it were connected end to end. The network interface card (NIC) 107 maintains a tail pointer table (TPTB) 108 for all the append queues 110. In certain embodiments, the TPTB 108 may be implemented as a hardware table in the NIC 107 separate from the network interface memory 115, however in other embodiments, the TPTB 108 may be located within the network interface memory 115. In other embodiments, the TPTB 108 may be located within the network interface memory 115 of the NIC 107.

[0024] In operation, as shown in FIG. 2, an incoming RIMA message 102 is sent from a sender computer and received by the NIC 107 of the receiver computer server. The message 102 contains a header 103 and a payload 104. The header 103 includes a queue identifier (Qid) 105 and message size value (MsgSize) 106. The payload contains the underlying message information. Upon receipt of the header 103, the TPTB 108 is looked up (using the processor 117) to translate the queue-id 105 to the tail pointer 111. In certain embodiments, the tail pointer value is atomically increased by the size of the

message (to new memory location 112). In other embodiments (where the memory address convention is reversed), the tail pointer is atomically decreased by the size of the message. In other embodiments, the tail pointer is atomically adjusted (either increased or decreased) by slightly more than the size of the message (e.g., up to 10% more, up to 25% more or up to 50% more). The tail pointer adjustment may be done modulo arithmetic in the case of a circular queue, or regular arithmetic for other types of queue address space schemes. In the disclosed RIMA system, the tail-pointer 111 is part of the append queue's attributes and can be maintained in the TPTB by the NIC 107. As such, the RIMA's atomic-fetch-and-add of the tail pointer 111 is very fast. In contrast, the prior art RDMA's native atomic fetch-and-op operation operates on arbitrary locations in host memory (e.g., system memory 109) which requires traversing the PCI system bus. This is part of the reason why emulating the append operation using existing verbs incurs a huge performance penalty. As used herein, the term "atomic" shall be understood to mean an operation acting on shared memory which completes in a single step relative to other threads. When an atomic store is performed on a shared variable, no other thread can observe the modification half-complete.

[0025] After the tail-pointer 111 is atomically increased (or decreased), the message payload 104 is copied (using the system bus) to the region of system memory 109 starting at the old tail pointer value. Upon completion of the message copy to the queue 110, a completion message is posted to a completion queue, which is also located in the system memory 109.

[0026] The system polls the completion queue using the CPU 113 and processes messages after delivery. Upon completion, the head pointer of the completion queue is increased to remove the processed messages from the completion queue and to free its memory region.

[0027] The pointers in the TPTB 108 are virtual addresses (i.e., the queues are in virtual address space) of the system memory 109. The NIC 107 TLB (translation lookaside buffer) provides translations from the virtual-address pointers to physical

memory addresses. The NIC107 TLB is already present in current RDMA NICs. When a virtual-to-physical translation is not present in the NIC TLB, the TLB invokes the system operating system (in software) or a page-table walk (in hardware) to look up the translation in the page table. If the translation does not exist in the page table (i.e., the physical memory corresponding to a virtual address has not been allocated) then the operating system (OS) is invoked to allocate physical memory and update the page table which then provides the translation to the NIC 107 TLB.

[0028] The disclosed RIMA system relies on light-weight allocation of physical memory in the common case in order to efficiently create room for messages dynamically. On one hand, there is a strong incentive to use a pre-allocated pool of physical memory to facilitate easy allocation with simple tail-pointer manipulation. However, under this approach, all of the pre-allocated memory, including that beyond the tail pointer, contributes towards the memory footprint of RIMA. On the other hand, the memory footprint minimization goal pushes the design in the opposite direction, since it is better to grow and shrink the memory dynamically to minimize wasted memory allocations.

[0029] The disclosed RIMA system overcomes this challenge by decoupling heavyweight physical memory allocation (and possibly OS-visible page mapping) from lightweight tail-pointer manipulation. In one embodiment, the disclosed RIMA system uses a reserve pool of physical memory beyond the tail pointer 111 to ensure that incoming messages can be accommodated with minimal delay. The reserve pool does add to the physical memory footprint. However, the size of the reserve pool is likely not large in absolute terms, especially in modern computer servers, and is much less than prior art SRQ which typically has many more queues (and corresponding reserve pools).

[0030] Further, to ensure that the heavyweight memory allocation latency is hidden, the allocation may be triggered at an appropriately set low watermark to ensure that the reserve pool is replenished before it is exhausted. For this replenishment, the sender computer does not reserve receiver computer memory and is not involved. Upon hitting

the low watermark, the receiver CPU 113 is interrupted to allocate more physical memory to the queue while the NIC 107 continues to process appends by using the reserve pool. The low watermark may be computed using Little's law. Assuming that the latency of allocating new pages is L (e.g., 1 millisecond) and some line rate (e.g., 10 Gbps), a reserve pool that is larger than $(L \times \text{linerate})$ (= 1.2 MB in our example) can accommodate any burst at linerate indefinitely if the newly allocated memory also equals (or exceeds) $L \times \text{linerate}$. By allocating a larger amount (say 12 MB), the receiver CPU 113 interrupts can be made less frequent. For de-allocation, as the tail pointer 112 moves forward, the older pages are effectively unused. As such, unused pages will be swapped out naturally by the OS (via normal virtual memory operation) to avoid unnecessary memory pressure.

[0031] The invention is inclusive of combinations of the aspects described herein. References to “a particular aspect” and the like refer to features that are present in at least one aspect of the invention. Separate references to “an aspect” (or “embodiment”) or “particular aspects” or the like do not necessarily refer to the same aspect or aspects; however, such aspects are not mutually exclusive, unless so indicated or as are readily apparent to one of skill in the art. The use of singular or plural in referring to “method” or “methods” and the like is not limiting. The word “or” is used in this disclosure in a non-exclusive sense, unless otherwise explicitly noted.

[0032] The invention has been described in detail with particular reference to certain preferred aspects thereof, but it will be understood that variations, combinations, and modifications can be effected by a person of ordinary skill in the art within the spirit and scope of the invention.

CLAIMS:

1. A computer server, comprising:

a network interface card (NIC) having a network interface memory; and

a system memory operatively connected to the network interface card, the system memory having a plurality of electronic memory queues, each of the memory queues corresponding to one of a plurality of receive processes in the computer server, each of the memory queues having a corresponding head pointer and tail pointer, wherein each of the memory queues is assigned to receive electronic messages from a plurality of sender computers;

wherein the NIC comprises a tail pointer table, the tail pointer table comprising initial memory location data of the tail pointers for the memory queues, the memory location data referenced by corresponding queue identifiers.
2. The computer server of claim 1, wherein the NIC is configured to receive an incoming network message from a sender computer external to the computer server, the network message comprising a header and a payload, the header comprising a queue identifier and a message size value.
3. The computer server of claim 2, wherein the NIC is configured to:

determine an assigned memory queue and corresponding tail pointer for the incoming message by referencing the queue identifier of the incoming message in the tail pointer table; and

atomically adjust the tail pointer of the assigned memory queue by at least the message size value.

4. The computer server of claim 3, wherein said adjustment comprises atomically increasing the tail pointer of the assigned memory queue by at least the message size value.

5. The computer server of claim 3, wherein said adjustment comprises atomically decreasing the tail pointer of the assigned memory queue by at least the message size value.

6. The computer server of claim 3, wherein the NIC is configured to:
store the payload at the initial memory location of the tail pointer.

7. The computer server of claim 1, wherein the computer server is configured to communicate a plurality of queue identifiers to the plurality of external sender computers.

8. The computer server of claim 1, wherein the computer server is located in a data center comprising a plurality of computer servers.

9. The computer server of claim 1, wherein the tail pointer table comprises a hardware table separate from the network interface memory.

10. The computer server of claim 1, wherein the tail pointer table is located within the network interface memory.

11. A method of managing incoming network messages in a computer server, comprising:

establishing a plurality of electronic memory queues in a system memory of the computer server, each of the memory queues corresponding to one of a plurality of receive processes in the computer server, each of the memory queues having a corresponding head pointer and tail pointer, wherein each of the memory queues is assigned to receive electronic messages from a plurality of sender computers;

establishing a tail pointer table in a network interface card (NIC) of the computer server, the tail pointer table comprising initial memory location data of the tail pointers for the memory queues, the memory location data referenced by corresponding queue identifiers.

12. The method of claim 11, further comprising:

using the NIC, receiving an incoming network message from a sender computer external to the computer server, the network message comprising a header and a payload, the header comprising a queue identifier and a message size value.

13. The method of claim 12, further comprising:

using the NIC, determining an assigned memory queue and corresponding tail pointer for the incoming message by referencing the queue identifier of the incoming message in the tail pointer table; and

atomically adjusting the tail pointer of the assigned memory queue by at least the message size value.

14. The method of claim 13, wherein said adjustment comprises atomically increasing the tail pointer of the assigned memory queue by at least the message size value.

15. The method of claim 13, wherein said adjustment comprises atomically decreasing the tail pointer of the assigned memory queue by at least the message size value.

16. The method of claim 13, further comprising:
storing the payload at the initial memory location of the tail pointer.

17. The method of claim 11, further comprising:
using the NIC, communicating a plurality of queue identifiers to the plurality of external sender computers.

18. The method of claim 11, wherein the computer server is located in a data center comprising a plurality of computer servers.

19. The method of claim 11, wherein the tail pointer table comprises a hardware table separate from the network interface memory.

20. The method of claim 11, wherein the tail pointer table is located within the network interface memory.

21. A network interface card (NIC), comprising:

a computer processor;

a network interface memory communicatively connected to the computer processor; and

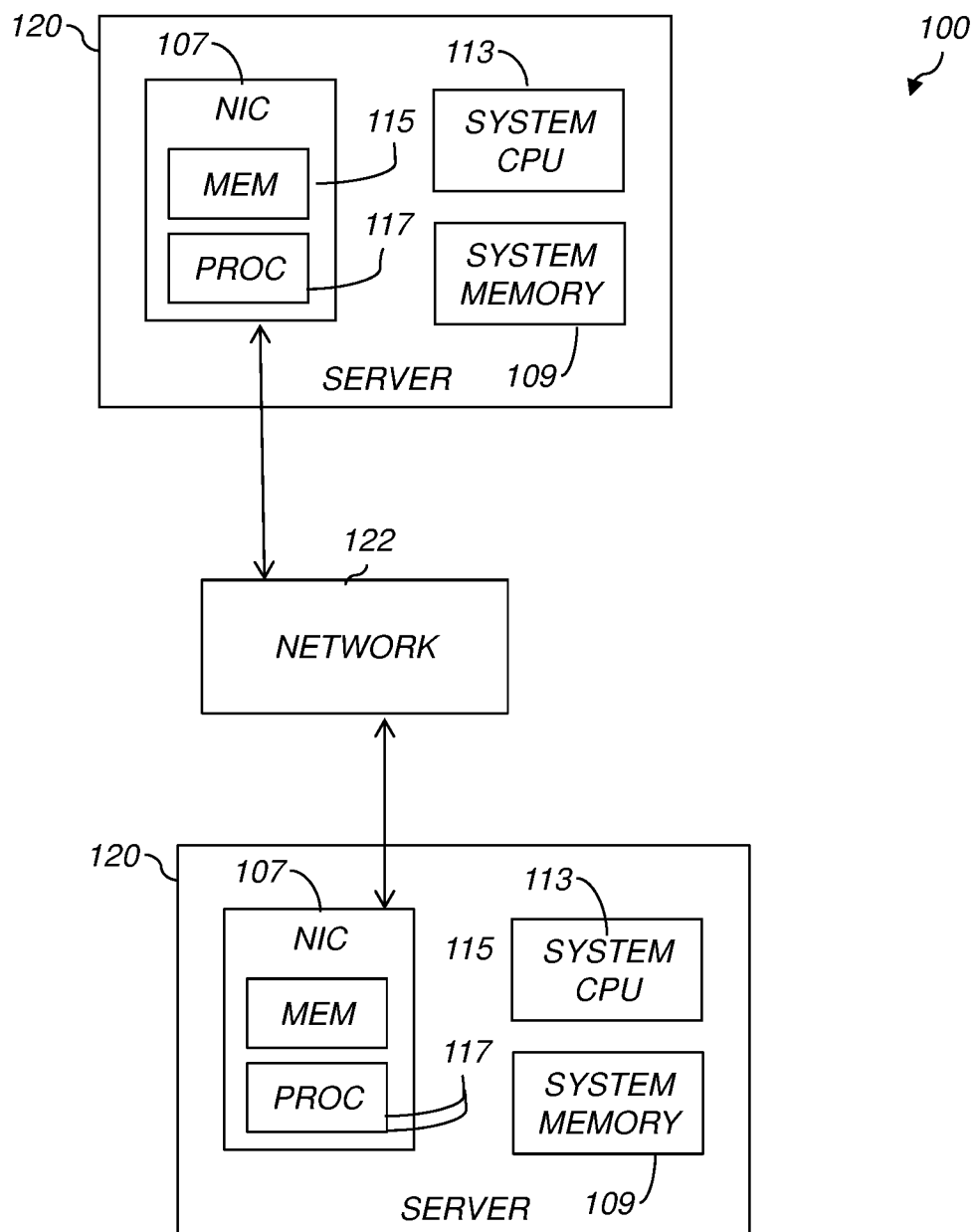
a tail pointer table, the tail pointer table comprising initial memory location data of a plurality of tail pointers for a corresponding plurality of memory queues, each of the memory queues corresponding to one of a plurality of receive processes received by the network interface card, the memory location data referenced by corresponding queue identifiers.

22. The network interface card of claim 22, wherein the NIC is configured to receive an incoming network message from a sender computer external and communicatively connected to the NIC, the network message comprising a header and a payload, the header comprising a queue identifier and a message size value.

23. The network interface card of claim 22, wherein the NIC is configured to:
- determine an assigned memory queue and corresponding tail pointer for the incoming message by referencing the queue identifier of the incoming message in the tail pointer table; and
- atomically adjust the tail pointer of the assigned memory queue by at least the message size value.
24. The network interface card of claim 23, wherein said adjustment comprises atomically increasing the tail pointer of the assigned memory queue by at least the message size value.
25. The network interface card of claim 23, wherein said adjustment comprises atomically decreasing the tail pointer of the assigned memory queue by at least the message size value.
26. The network interface card of claim 23, wherein the NIC is configured to:
- store the payload at the initial memory location of the tail pointer.
27. The network interface card of claim 21, wherein the tail pointer table comprises a hardware table separate from the network interface memory.

28. The network interface card of claim 21, wherein the tail pointer table is located within the network interface memory.

1/2

**FIG. 1**

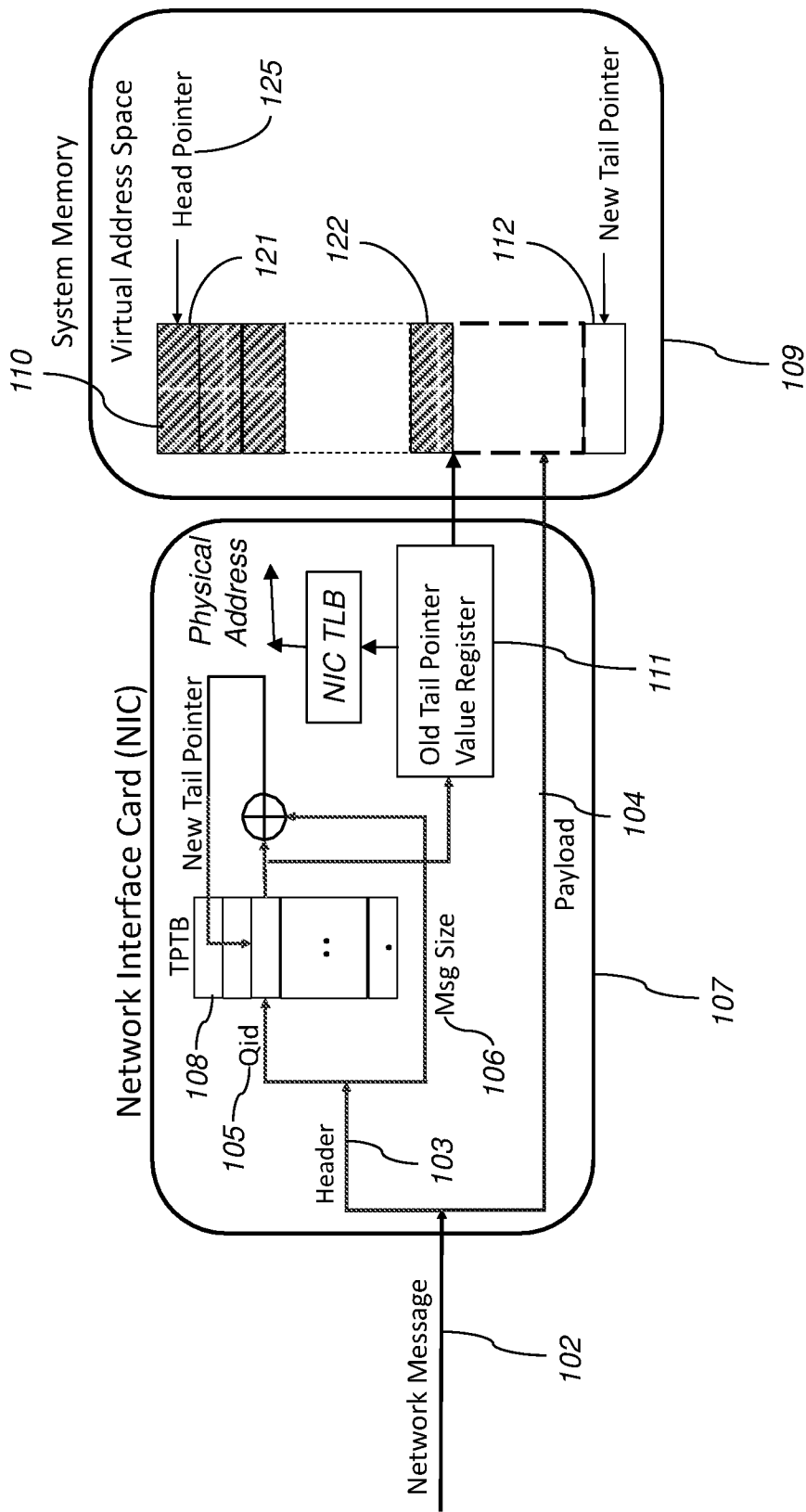


FIG. 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 17/22208

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 3/00 (2017.01)

CPC - G06F 3/0659

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2003/0200315 A1 (Goldenberg et al.) 23 October 2003 (23.10.2003), para. [0005], [0016], [0067], [0082], [0084]-[0085], and [0090], and Fig. 1 and 5-6, and claim 1.	1-2, 7-12, 17-22, 27-28 ----- 3-6, 13-16, 23-26
Y	US 2011/0173399 A1 (Chen et al.) 14 July 2011 (14.07.2011) (para [0072], [0076], [0112], [0126], [0136])	3-6, 13-16, 23-26
Y	US 2011/0145515 A1 (Serebrin) 16 June 2011 (16.06.2011) (para [0018], [0033]-[0034])	5, 15, 25

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

21 June 2017 (21.06.2017)

Date of mailing of the international search report

17 JUL 2017

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774