



(51) International Patent Classification:

H04L 12/851 (2013.01) H04L 12/863 (2013.01)

H04L 12/721 (2013.01) H04L 12/931 (2013.01)

H04L 12/743 (2013.01) H04L 12/935 (2013.01)

(21) International Application Number:

PCT/US2020/024259

(22) International Filing Date:

23 March 2020 (23.03.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/852,273 23 May 2019 (23.05.2019) US

62/852,203 23 May 2019 (23.05.2019) US

62/852,289 23 May 2019 (23.05.2019) US

(71) Applicant: CRAY INC [US/US]; 901 5th Ave STE 1000, Seattle, Washington 98164-2008 (US).

(72) Inventor: BEECROFT, Jonathan P.; 901 5th Ave STE 1000, Seattle, Washington 98164 (US).

(74) Agent: FEBBO, Michael et al.; HEWLETT PACKARD ENTERPRISE, 3404 E. Harmony Road Mail Stop 79, Fort Collins, Colorado 80528 (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, JO, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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,

(54) Title: SYSTEM AND METHOD FOR FACILITATING DATA-DRIVEN INTELLIGENT NETWORK

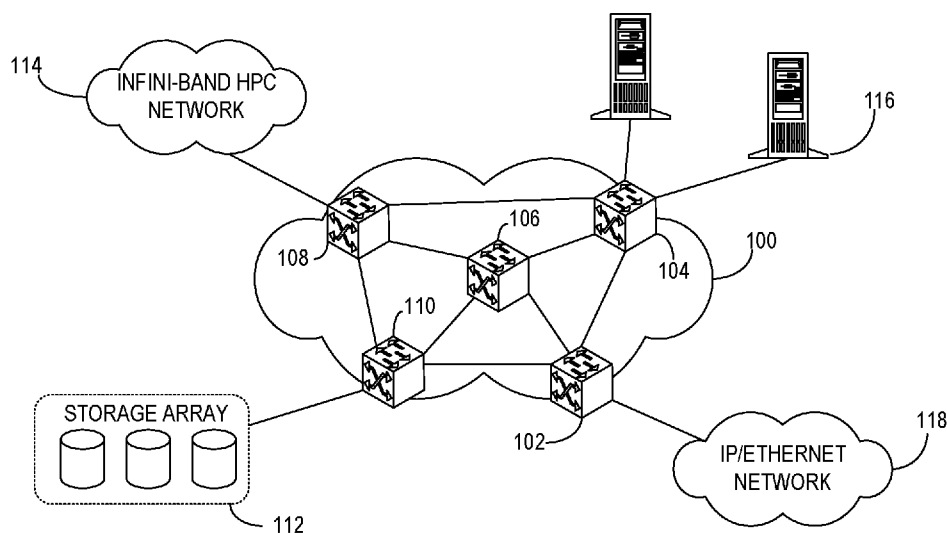


FIG. 1

(57) Abstract: Data-driven intelligent networking systems and methods are provided. The system can accommodate dynamic traffic with fast, effective congestion control. The system can maintain state information of individual packet flows, which can be set up or released dynamically based on injected data. Each flow can be provided with a flow-specific input queue upon arriving at a switch. Packets of a respective flow can be acknowledged after reaching the egress point of the network, and the acknowledgement packets can be sent back to the ingress point of the flow along the same data path. As a result, each switch can obtain state information of each flow and perform flow control on a per-flow basis.



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).

Declarations under Rule 4.17:

- *as to the identity of the inventor (Rule 4.17(i))*
- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*

SYSTEM AND METHOD FOR FACILITATING DATA-DRIVEN INTELLIGENT NETWORK

5

Inventor: Jonathan P. Beecroft

BACKGROUND

Field

10

[0001] This is generally related to the technical field of networking. More specifically, this disclosure is related to systems and methods for facilitating scalable, data-driven intelligent networks with congestion control.

Related Art

15

[0002] As network-enabled devices and applications become progressively more ubiquitous, various types of traffic as well as the ever-increasing network load continue to demand more performance from the underlying network architecture. For example, applications such as high-performance computing (HPC), media streaming, and Internet of Things (IOT) can generate different types of traffic with distinctive characteristics. As a result, in addition to

20 conventional network performance metrics such as bandwidth and delay, network architects continue to face challenges such as scalability, versatility, and efficiency.

SUMMARY

25

[0003] A data-driven intelligent networking system is provided. The system can accommodate dynamic traffic with fast, effective congestion control. The system can maintain state information of individual packet flows, which can be set up or released dynamically based on injected data. Each flow can be provided with a flow-specific input queue upon arriving at a switch. Packets of a respective flow are acknowledged after reaching the egress point of the network, and the acknowledgement packets are sent back to the ingress point of the flow along

30 the same data path. As a result, each switch can obtain state information of each flow and perform flow control on a per-flow basis.

BRIEF DESCRIPTION OF THE FIGURES

35

[0004] FIG. 1 shows an exemplary network that facilitates flow channels.

[0005] FIG. 2A shows an exemplary switch that facilitates flow channels.

[0006] FIG. 2B shows an example of how switches along a data path can maintain flow state information.

[0007] FIG. 3A shows an exemplary fabric header for a data packet.

[0008] FIG. 3B shows an exemplary acknowledgement (ACK) packet format.

5 [0009] FIG. 3C shows the relationship between different variables used to derive and maintain state information of a flow.

[0010] FIG. 4A shows an example of how flow channel tables can be used to deliver a flow.

[0011] FIG. 4B shows an example of an edge flow channel table (EFCT).

10 [0012] FIG. 4C shows an example of an input flow channel table (IFCT).

[0013] FIG. 4D shows an example of an output flow channel table (OFCT).

[0014] FIG. 5 shows an example where an unfair share of link bandwidth can occur in a network.

[0015] FIG. 6 shows an example of endpoint congestion.

15 [0016] FIG. 7A shows a flow chart of an exemplary process of generating an explicit endpoint-congestion-notification ACK.

[0017] FIG. 7B shows an exemplary endpoint congestion management logic block.

[0018] FIG. 8 shows a flow chart showing of exemplary process of generating an ACK in response to a packet being dequeued from an output buffer.

20 [0019] FIG. 9A shows a flow chart of an exemplary fine grain flow control (FGFC) process.

[0020] FIG. 9B shows an example of a FGFC-enabled network interface controller.

[0021] FIG. 10 shows an example of fabric link congestion.

25 [0022] FIG. 11 shows a flow chart of an example process of applying credit-based flow control on a congested fabric link.

[0023] FIG. 12 shows an exemplary edge switching system that facilitates flow channels.

[0024] FIG. 13 shows an exemplary intermediary switching system that facilitates flow channels.

[0025] In the figures, like reference numerals refer to the same figure elements.

30

DETAILED DESCRIPTION

[0026] Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present disclosure. Thus, the present invention is not limited to the embodiments shown.

35

Overview

[0027] The present disclosure describes systems and methods that accommodate dynamic data traffic with fast, effective congestion control by maintaining state information of individual packet streams. More specifically, packets injected into a network of switches can be categorized into streams, which can be mapped to their layer-2, layer-3, or other protocol-specific header information. Each stream can be marked by a distinctive identifier that is local to an input port of a switch, and provided with a stream-specific input buffer, so that each stream can be individually flow-controlled. In addition, packets of a respective stream can be acknowledged upon reaching the egress point of the network, and the acknowledgement packets can be sent back to the ingress point of the stream along the same data path in the reverse direction. As a result, each switch can obtain state information of active packet streams it is forwarding, and can perform highly responsive, stream-specific flow control. Such flow control can allow the network to operate at higher capacity while providing versatile traffic-engineering capabilities.

[0028] In this disclosure, packet streams can also be referred to as “packet flows,” or simply “flows.” The data path traversed by a flow, together with its configuration information maintained by switches, can be referred to as a “flow channel.” Furthermore, the terms “buffer” and “queue” are used interchangeably in this disclosure.

[0029] FIG. 1 shows an exemplary network that facilitates flow channels. In this example, a network 100 of switches, which can also be referred to as a “switch fabric,” can include switches 102, 104, 106, 108, and 110. Each switch can have a unique address or ID within switch fabric 100. Various types of devices and networks can be coupled to a switch fabric. For example, a storage array 112 can be coupled to switch fabric 100 via switch 110; an InfiniBand (IB) based HPC network 114 can be coupled to switch fabric 100 via switch 108; a number of end hosts, such as host 116, can be coupled to switch fabric 100 via switch 104; and an IP/Ethernet network 118 can be coupled to switch fabric 100 via switch 102. In general, a switch can have edge ports and fabric ports. An edge port can couple to a device that is external to the fabric. A fabric port can couple to another switch within the fabric via a fabric link.

[0030] Typically, traffic can be injected into switch fabric 100 via an ingress port of an edge switch, and leave switch fabric 100 via an egress port of another (or the same) edge switch. An ingress edge switch can group injected data packets into flows, which can be identified by flow ID's. The concept of a flow is not limited to a particular protocol or layer (such as layer-2 or layer-3 in the Open System Interface (OSI) reference model). For example, a flow can be mapped to traffic with a particular source Ethernet address, traffic between a source IP address and destination IP address, traffic corresponding to a TCP or UDP port/IP 5-tuple (source and

destination IP addresses, source and destination TCP or UDP port numbers, and IP protocol number), or traffic produced by a process or thread running on an end host. In other words, a flow can be configured to map to data between any physical or logic entities. The configuration of this mapping can be done remotely or locally at the ingress edge switch.

5 **[0031]** Upon receiving injected data packets, the ingress edge switch can assign a flow ID to the flow. This flow ID can be included in a special header, which the ingress edge switch can use to encapsulate the injected packets. Furthermore, the ingress edge switch can also inspect the original header fields of an injected packet to determine the appropriate egress edge switch's address, and include this address as a destination address in the encapsulation header. Note that
10 the flow ID can be a locally significant value specific to a link, and this value can be unique only to a particular input port on a switch. When the packet is forwarded to the next-hop switch, the packet enters another link, and the flow-ID can be updated accordingly. As the packets of a flow traverses multiple links and switches, the flow IDs corresponding to this flow can form a unique chain. That is, at every switch, before a packet leaves the switch, the packet's flow ID can be
15 updated to a flow ID used by the outgoing link. This up-stream-to-down-stream one-to-one mapping between flow ID's can begin at the ingress edge switch and end at the egress edge switch. Because the flow ID's only need to be unique within an incoming link, a switch can accommodate a large number of flows. For example, if a flow ID is 11 bits long, an input port can support up to 2048 flows. Furthermore, the match pattern (one or more header fields of a
20 packet) used to map to a flow can include a greater number of bits. For instance, a 32-bit long match pattern, which can include multiple fields in a packet header, can map up to 2^{32} different header field patterns. If a fabric has N ingress edge ports, a total number of $N \cdot 2^{32}$ identifiable flows can be supported.

[0032] A switch can assign every flow a separate, dedicated input queue. This
25 configuration allows the switch to monitor and manage the level of congestion of individual flows, and prevent head-of-queue blocking which could occur if shared buffer were used for multiple flows. When a packet is delivered to the destination egress switch, the egress switch can generate and send back an acknowledgement (ACK) in the upstream direction along the same data path to the ingress edge switch. As this ACK packet traverses the same data path, the
30 switches along the path can obtain the state information associated with the delivery of the corresponding flow by monitoring the amount of outstanding, unacknowledged data. This state information can then be used to perform flow-specific traffic management to ensure the health of the entire network and fair treatment of the flows. As explained in more detail below, this per-flow queuing, combined with flow-specific delivery acknowledgements, can allow the switch

fabric to implement effective, fast, and accurate congestion control. In turn, the switch fabric can deliver traffic with significantly improved network utilization without suffering from congestion.

[0033] Flows can be set up and released dynamically, or “on the fly,” based on demand. Specifically, a flow can be set up (e.g., the flow-ID to packet header mapping is established) by an ingress edge switch when a data packet arrives at the switch and no flow ID has been previously assigned to this packet. As this packet travels through the network, flow IDs can be assigned along every switch the packet traverses, and a chain of flow IDs can be established from ingress to egress. Subsequent packets belonging to the same flow can use the same flow IDs along the data path. When packets are delivered to the destination egress switch and ACK packets are received by the switches along the data path, each switch can update its state information with respect to the amount of outstanding, unacknowledged data for this flow. When a switch’s input queue for this flow is empty and there is no more unacknowledged data, the switch can release the flow ID (i.e., release this flow channel) and re-use the flow-ID for other flows. This data-driven dynamic flow setup and teardown mechanism can obviate the need for centralized flow management, and allows the network to respond quickly to traffic pattern changes.

[0034] Note that the network architecture described herein is different from software-defined networks (SDN’s), which typically uses the OpenFlow protocol. In SDN, switches are configured by a central network controller, and packets are forwarded based one or more fields in the layer-2 (data link layer, such as Ethernet), layer-3 (network layer, such as IP), or layer-4 (transport layer, such as TCP or UDP) headers. In SDN such header-field lookup is performed at every switch in the network, and there is no fast flow ID-based forwarding as is done in the networks described herein. Furthermore, because the OpenFlow header-field lookup is done using ternary content-addressable memory (TCAM), the cost of such lookups can be high. Also, because the header-field mapping configuration is done by the central controller, the setup and tear-down of each mapping relationship is relatively slow and could require a fair amount of control traffic. As a result, an SDN network’s response to various network situations, such as congestion, can be slow. In contrast, in the network described herein, the flows can be set up and torn down dynamically based on traffic demand; and packets can be forwarded by a fixed-length flow ID. In other words, flow channels can be data driven and managed (i.e., set up, monitored, and torn down) in a distributed manner, without the intervention of a central controller. Furthermore, the flow ID-based forwarding can reduce the amount of TCAM space used and as a result a much greater number of flows can be accommodated.

[0035] Referring to the example shown in FIG. 1, suppose that storage array 112 is to send data using TCP/IP to host 116. During operation, storage array 112 can send the first packet

with host 116's IP address as the destination address and a predetermined TCP port specified in the TCP header. When this packet reaches switch 110, the packet processor at the input port of switch 110 can identify a TCP/IP 5-tuple of this packet. The packet processor of switch 110 can also determine that this 5-tuple currently is not mapped to any flow ID, and can allocate a new flow ID to this 5-tuple. Furthermore, switch 110 can determine the egress switch, which is switch 104, for this packet based on the destination (i.e., host 116's) IP address (assuming switch 110 has knowledge that host 116 is coupled to switch 104). Subsequently, switch 110 can encapsulate the received packet with a fabric header that indicates the newly assigned flow ID and switch 104's fabric address. Switch 110 can then schedule the encapsulated packet to be forwarded toward switch 104 based on a fabric forwarding table, which can be computed by all the switches in fabric 100 using a routing algorithm such as link state or distance vector.

[0036] Note that the operations described above can be performed substantially at line speed with little buffering and delay when the first packet is received. After the first packet is processed and scheduled for transmission, subsequent packets from the same flow can be processed by switch 110 even faster because the same flow ID is used. In addition, the design of the flow channels can be such that the allocation, matching, and deallocation of flow channels can have substantially the same cost. For example, a conditional allocation of a flow channel based on a lookup match and a separate, independent deallocation of another flow channel can be performed concurrently in nearly every clock cycle. This means that generating and controlling the flow channels can add nearly no additional overhead to the regular forwarding of packets. The congestion control mechanism, on the other hand, can improve the performance of some applications by more than three orders of magnitude.

[0037] At each switch along the data path (which includes switches 110, 106, and 104), a dedicated input buffer can be provided for this flow, and the amount of transmitted but unacknowledged data can be tracked. When the first packet reaches switch 104, switch 104 can determine that the destination fabric address in the packet's fabric header matches its own address. In response, switch 104 can decapsulate the packet from the fabric header, and forward the decapsulated packet to host 116. Furthermore, switch 104 can generate an ACK packet and send this ACK packet back to switch 110. As this ACK packet traverses the same data path, switches 106 and 110 can each update their own state information for the unacknowledged data for this flow.

[0038] In general, congestion within a network can cause the network buffers to fill. When a network buffer is full, the traffic trying to pass through the buffer ideally should be slowed down or stopped. Otherwise, the buffer could overflow and packets could be dropped. In conventional networks, congestion control is typically done end-to-end at the edge. The core of

the network is assumed to function only as “dumb pipes,” the main purpose of which is to forward traffic. Such network design often suffers from slow responses to congestions, because congestion information often cannot be sent to the edge devices quickly, and the resulting action taken by the edge devices cannot always be effective in removing the congestion. This slow response in turn limits the utilization of the network, because to keep the network free of congestion the network operator often needs to limit the total amount of traffic injected into the network. Furthermore, end-to-end congestion control usually is only effective provided that the network is not already congested. Once the network is heavily congested, end-to-end congestion control would not work, because the congestion notification messages can be congested themselves (unless a separate control-plane network that is different from the data-plane network is used for sending congestion control messages).

[0039] In contrast, the flow channels can prevent such congestion from growing within the switch fabric. The flow channel mechanism can recognize when a flow is experiencing some degree of congestion, and in response can slow down or stop new packets of the same flow from entering the fabric. In turn, these new packets can be buffered in a flow channel queue on the edge port and are only allowed into the fabric when packets for the same flow leave the fabric at the destination edge port. This process can limit the total buffering requirements of this flow within the fabric to an amount that would not cause the fabric buffers to become too full.

[0040] With flow channels, the switches have a reasonably accurate state information on the amount of outstanding in-transit data within the fabric. This state information can be aggregated for all the flows on an ingress edge port. This means that the total amount of data injected by an ingress edge port can be known. Consequently, the flow channel mechanism can set a limit on the total amount of data in the fabric. When all edge ports apply this limit action, the total amount of packet data in the entire fabric can be well controlled, which in turn can prevent the entire fabric from being saturated. The flow channels can also slow the progress of an individual congested flow within the fabric without slowing down other flows. This feature can keep packets away from a congestion hot spot while preventing buffers from becoming full and ensuring free buffer space for unrelated traffic.

Operation of Flow Channel

[0041] In general, flow channels can define a path for each communication session across the switch fabric. The path and amount of data belonging to each flow can be described in a set of dynamically connecting flow tables associated with each link of the switch fabric. On every ingress port, edge and fabric, a set of flow channel queues can be defined. There can be one queue for each flow channel. As packets arrive, they either can be assigned to a flow channel on

an edge port, or have been assigned to a flow channel by the link partner's egress fabric port on a fabric ingress port. The flow channel information can be used to direct the packets into the appropriate flow channel queue.

[0042] FIG. 2A shows an exemplary switch that facilitates flow channels. In this example, the switch can include a crossbar switch 202. Crossbar switch 202 can have a number of input ports, such as input port 204, and a number of output ports, such as output 208. Crossbar switch 202 can forward packets from an input port to an output port. Each input port can be associated with a number of input queues, each assigned to a different incoming flow arriving on that input port. For example, data arriving on a given port of the switch can first be separated, based on their individual flows, and stored in flow-specific input queues, such as input queue 206. The packets stored in the input queues can be dequeued and sent to crossbar switch 202 based on scheduling algorithms designed to control congestions (described in more detail in later sections). On the output side, once a packet passes crossbar switch 202, it can be temporarily stored in an output transmission queue, such as output transmission queue 210, which can be shared by all the flows leaving on the same output port. Meanwhile, before a packet is dequeued from the output transmission queue and transmitted on the outgoing link, the packet's header can be updated with the flow ID for the outgoing link. Note that this hop-by-hop flow ID mapping can be done when the first packet in the flow travels across the network. When the packet reaches the next-hop switch, the packet can be stored again in a flow-specific input queue and the same process can be repeated. Note that a flow ID is used to distinguish between flows traveling on the same fabric link, and can be typically assigned by the transmitter end of this link, which is the output port of the switch that is transmitting onto this link.

[0043] By providing flow-specific input queues, the switch can allow each flow to move independently of all other flows. The switch can avoid the head-of-queue blocking problem, which is common with shared input buffers. The flow-specific input queue also allows the packets within a single flow to be kept in order. When a flow passes through the switches, a flow-specific input queue on each input port can be allocated for this flow and these input queues become linked, effectively forming one long queue that reaches across the entire fabric for this flow, and the packets of this flow can be kept in order.

[0044] The progress of successful delivery of packets belonging to a flow can be reported by a sequence of ACKs generated by the edge port of an egress switch. The ACK packets can travel in the reverse direction along the data path traversed by the data packets and can be forwarded by the switches according to the forwarding information maintained in flow tables. As ACK packets travel upstream, they can be processed by each switch's input queue manager, which can update the corresponding flow's state information based on information carried by the

ACK packets. The ACK packets can have a type field to provide advanced information about the downstream data path, such as congestions. A switch's input queue manager can use this information to make decisions, such as throttling the transmission rate or changing the forwarding path, about the pending data packets currently buffered in its input queues. In addition, the input queue manager can update the information carried in an ACK packet based on state information of a buffered flow, so that the upstream switches can make proper decisions. For example, if an input queue for a given flow is experiencing congestion (e.g., the amount of data in the queue is above a predetermined threshold), the input queue manager can update an ACK packet that is being forwarded to the next upstream switch to include this congestion information.

[0045] If an ACK corresponds to the last packet of a flow, a switch can determine that there is no more unacknowledged data for that flow. Correspondingly, the switch can free the flow channel by removing the corresponding entry in the flow table.

[0046] As mentioned above, the input queue manager at each switch can maintain information about transmitted but unacknowledged data of a given flow. FIG. 2B shows an example of how switches along a data path can maintain flow state information. In this example, the data path taken by a flow can include switches 222, 224, and 226. The amount of transmitted but unacknowledged flow data can be indicated by a variable "flow_extent," which can be measured in number of fixed-length data units, such as 256 bytes. Furthermore, flow_extent and other flow state information can be maintained by a switch's input queue manager, which can continuously monitor all the flow-specific queues.

[0047] In the example in FIG. 2B, the value of flow_extent at the input queue manager of switch is 1, because there is one unit of data that has been sent out of the input queue and forwarded through the crossbar switch. Note that a data packet sent by an input queue might be temporarily buffered in the output transmission buffer due to the scheduling of all the data packets to be transmitted via an output link. When such a packet is buffered in the output port's transmission buffer, the packet can still be considered by the input queue as transmitted for the purpose of updating the flow_extent value.

[0048] Correspondingly, because the input queue for the given flow at switch 226 has six queued data units, and two additional data units are in transit between switches 224 and 226, the flow_extent value at switch 224 is 9. Similarly, the flow_extent value at switch 222 is 13, because there are three data units stored in the input queue at switch 224 and one data unit in transit between switches 222 and 224.

[0049] In general, a flow channel can remain allocated to a single flow until all the ACKs for all the packets sent on the flow channel have been returned. This means that flow channel

table entries can remain active for longer near the fabric ingress edge port than near the egress edge port. If a single packet is injected into the network, a flow channel can be allocated for the ingress edge port and then another flow channel can be allocated for the next fabric link the packet traverses and so on, until the last flow channel is allocated when the packet reaches the last fabric link. Each allocation can generate a flow ID, denoted as variable “flow_id,” to identify the entries of the flow tables of the fabric link. (More details on flow channel tables are provided in the description below in conjunction with FIG. 4A.) This first packet may cause the allocation of a different flow_id, on each of the fabric links the packet traverses across the switch fabric.

[0050] At the input queue of each switch, the flow channel table entries can indicate each flow’s state information, including the flow_extent value, from this point downstream to the flow’s egress destination edge port. Packets received on the local input port can increase this flow_extent value by the amount of incoming data, and ACKs can reduce the flow_extent by the amount of acknowledged, delivered data.

[0051] When a packet reaches the final destination egress port, an ACK packet can be generated and returned for that packet. This ACK can be routed using the data path information stored in the corresponding entry of the flow channel tables at every switch along the data path. Optionally, the ACK packet itself does not need to carry path information and therefore can be small and light weight. If no other data packet is sent on the flow, the ACK can release each flow channel in the reverse order. Once released, the flow channel at each switch can be allocated to a different flow.

[0052] If another packet follows the first packet on the same flow, the ACK corresponding to the second packet would need to be received before the flow channel can be released at a given switch. In one embodiment, the flow channel can only be released when ACKs for all the transmitted packets of the same flow have been returned.

[0053] Typically, various protocols may require in-order packet delivery. The flow channels can be used to guarantee this delivery order, even when the fabric uses adaptive routing for load balancing across multiple data paths. If packets between an ingress edge port and an egress edge port, perhaps in a different switch on the far side of the fabric, are injected at a very low rate, then each packet injected could reach its destination and return an ACK back to the source before the next packet is injected. In this case, each packet can be a lead packet and free to take any path across the fabric, using the best available dynamic adaptive routing choice. This is possible because the first packet can define the flow’s path through the fabric.

[0054] Now assume that the packet injection rate is increased slightly to the point where the next packet of the same flow is injected before the current packet’s ACK has returned to the

source. The second packet can pass the ACK of the first packet somewhere along the flow's data path. Beyond this passing point, the ACK will have released the flow channels allocated to the first packet, because the flow_extent value associated with the first packet is returned to zero when the ACK is processed by the flow channel's logic. Meanwhile, the second packet can now
5 define a new flow, because it is again causing flow channels to be allocated on each of the subsequent fabric links. This second packet, while it is causing flow channels to be allocated beyond the passing point, can be forwarded to a different path based on dynamic adaptive routing. On the other hand, before the passing point, the second packet can extend the outstanding flow created by the first packet to include the second packet. This means the first
10 packet's ACK may not reduce the flow_extent value to zero and the flow channels may remain active before the passing point. It also means that the second packet may follow the exact path taken by the first packet up to the passing point. Note that while it is following the previous packet, the second packet cannot arrive at the egress edge port before the first packet does, and therefore correct packet order can be maintained.

15 **[0055]** If the injection rate for this flow is increased further, the second packet will pass the first packet's ACK at a location closer to the destination edge port. It is also possible that a third, fourth, fifth, or additional packet may enter the fabric before the first packet's ACK is returned to the source edge port, depending on the data packet injection rate of this flow and the data packet-ACK round trip delay. The maximum packet rate can depend on the size of the
20 packets and the bandwidth of the links. The round trip delay of the data packet and ACK can be an important parameter for a fabric implementation and can be used along with the maximum packet rate to calculate the maximum required number of flow channels for each link. Ideally, a design can provide a reasonable number of unallocated flow channels regardless of the traffic pattern. The demand for the number of flow channel can be high when a large number of packets
25 arriving at an ingress edge port have different destinations and these packets have small sizes and high injection rates. In the most extreme case, each packet could be allocated a different flow channel. These flow channels are freed when the packets' ACKs are returned. Correspondingly, the number of flow channels needed can be calculated as ((Packet rate) * (Average packet to ACK round trip latency)).

30 **[0056]** Note that packet rate on a single flow channel is not to be confused with packet rate on a link. If the traffic pattern is such that many small packets are being sent to different destinations, then successive packets sent onto the link can have different destinations. This means that each packet could belong to a different flow and could be the only packet to use the corresponding flow channel. In this example, the link can experience a high packet rate, but the
35 packet rate of individual flows can be low. Optionally, a number of ACKs (e.g., 48 ACKs) can

be aggregated together into a single ACK frame for transmission over a link and protected by a Frame Check Sequence (e.g., a 32-bit FCS). For example, the ACKs can occupy 25 bits each, and there can be a 9-byte overhead to the frame. That is, the overhead per ACK on a full size frame is approximately $9/(25/8 * 48) * 100\% = 6\%$. The logic can optimize the number of ACKs per frame so an ACK does not need to wait too long to be aggregated when the ACKs are arriving slowly. For example, the ACK aggregation logic block can use three timers to manage ACK transmission based on the activity of an outgoing link. These timers can be started when a new ACK arrives at the ACK aggregation logic block. If the outgoing link is idle, a first timer, which can for example be set at 30ns, can be used to hold the ACK while waiting for additional ACKs to arrive. When this timer expires, all the ACK received within the corresponding time window can be aggregated into one frame and transmitted onto the outgoing link. If the outgoing link is busy, a second timer, which can for example be set at 60ns, can be used to wait for additional ACKs. Using this second timer can allow more ACKs to be aggregated into a single frame, and this frame can be transmitted only if a predetermined number of ACKs are collected. Note that due to the Ethernet framing constrains, some numbers of ACKs in a single frame can use less wire bandwidth per ACKs than other numbers of ACKs. If no efficient number of ACKs are collected, and the outgoing link remains busy sending normal data packets, then a third timer, which can for example be set at 90ns, can be used. Once this third timer expires, all the ACKs that have been collected can be aggregated in a frame and transmitted onto the link. By using these three timers, the system can significantly reduce the overhead of sending ACKs on the outgoing link.

[0057] In some examples, the ingress edge port of a switch can encapsulate a received data packet with a fabric header, which allows the packet to be forwarded using flow channels. FIG. 3A shows an exemplary fabric header for a data packet. The fabric header can include a flow_id field, which can identify the flow channel, and a “data_flow” field, which can indicate the progression of the entire flow.

[0058] When a data packet is delivered to its destination, at least one ACK can be generated. FIG. 3B shows an exemplary ACK packet format. An ACK packet can include a “flow_id” field, an “ack_flow” field, an “ACK type” field, and a cyclic redundancy check (CRC) field. The flow_id field can indicate the flow this ACK packet belongs to. The ack_flow field can indicate the data packet to which this ACK packet acknowledges. Recall that each switch can maintain a flow_extent value which indicates the amount of transmitted but unacknowledged data. The value of flow_extent can be derived as $\text{flow_extent} = \text{data_flow} - \text{ack_flow}$, wherein data_flow value is taken from the last transmitted data packet.

[0059] The ACK type field can indicate different types of ACKs. As mentioned above, during normal operation, when a data packet is delivered to the destination edge port, a regular ACK packet can be generated and sent back to the source. Correspondingly, the ACK type field in the ACK packet can indicate a normal ACK. When congestion occurs in the fabric, the ACK type field can be used to indicate various types and severity of congestion, such as a new congestion, a persistent congestion, or a severe congestion at the egress edge port that calls for rerouting of the flow. In addition, under special circumstances such as the presence of a severely congested fabric link, dropped packets, or link error, an ACK can also be generated by an intermediate switch that is not the final destination, and the ACK type field can be used to notify upstream switches of different types of network condition. Other additional fields can also be included in an ACK packet.

[0060] FIG. 3C shows the relationship between different variables used to derive and maintain state information of a flow. In this example, a switch can use the variable “total_extent” to track the total amount of unacknowledged transmitted data and data currently queued at the switch. The value of total_extent can equal the sum of flow_extent, which is the amount of transmitted and unacknowledged data, and queue_extent, which is the amount of data stored in the input queue for the corresponding flow. The variable “ack_flow” can indicate the data position that corresponds to the latest ACK for this flow. The variable “data_flow” can indicate the position of the next data packet to be transmitted, which also corresponds to the data packet stored at the head of the input queue. The variable “next_data_flow” can indicate the position of the next data packet that the switch can expect to receive from the upstream switch. Note that $\text{queue_extent} = \text{next_data_flow} - \text{data_flow}$, and $\text{flow_extent} = \text{data_flow} - \text{ack_flow}$.

[0061] In some examples, flow channel tables can be used to facilitate flow channels throughout a fabric. Flow channel tables are data structures that store the forwarding and state information for a given flow at the port of a switch. FIG. 4A shows an example of how flow channel tables can be used to store state information associated with multiple flows. This state information can be specific to each flow and efficiently stored in a table. Assume that a source host 402 is sending data packets to a destination host 404 via a fabric. The data path traversed by the data packets can include an ingress edge switch 406, intermediate switches 408 and 430, and egress edge switch 432.

[0062] When a packet arrives on an ingress edge link 403 of switch 406, the packet's header can be analyzed by an address translate logic block 410. Address translate logic block 410 can determine the destination fabric address of the egress switch (which in this case is switch 432) based on the packet's Ethernet, IP, or HPC header information. Note that header

information associated with other protocols or a combination of different protocols can also be used by address translate logic block 410. The fabric destination address determined by address translate logic block 410 can then be used to perform a lookup in an edge flow channel table (EFCT) 412. EFCT 412 can perform a lookup operation for the packet using the packet's fabric destination address and optionally additional values extracted from the packet's header, which can be referred to as a match pattern. EFCT 412 can compare the packet's match pattern against stored match patterns of all existing allocated flows. If a match is found, then this packet is part of an existing flow and the previously allocated flow ID can be returned for this packet. If no match is found, a new flow ID can be allocated for this packet, and a match pattern can be added to EFCT 412. In other words, EFCT 412 can be used to determine whether a flow channel already exists for the incoming packet, or whether a new flow channel needs to be allocated. In addition to the destination fabric address, other packet header information such as traffic class, TCP or UDP port number, and process or thread ID can be used to map or allocate flow IDs.

[0063] The flow ID obtained by EFCT 412 can then be used as an index to map to an entry in an input flow channel table (IFCT) 414. Each entry in IFCT 414 can be indexed by a flow ID and store state information for the corresponding flow. An entry in IFCT 414 can store the values of next_data_flow, data_flow, and ack_flow (see FIG. 3C) associated with a flow. In addition, an IFCT entry can store other parameters for congestion control and dynamic routing for a flow.

[0064] The flow ID can also be used to identify or allocate a flow-specific input queue in which the incoming packet can be temporarily stored. The state information for a particular queue, as well as parameters for monitoring and controlling the queue (such as threshold for detecting congestion) can be stored in the corresponding entry in IFCT 414. An input queue management logic block can determine when a packet can be dequeued from the input queue and sent to a data crossbar switch 413 based on flow-control parameters stored in the entry of IFCT 414.

[0065] When a packet is dequeued from the input queue and sent through crossbar switch 413 to an output port, the packet is sent with the input port number on which it has arrived at switch 406. When the packet reaches an output port's transmission buffer, the packet's header can be updated, based on the packet's flow ID and input port number, with a new flow ID to be used by the next-hop switch (i.e., switch 408) for the same flow. This is because each link, in each direction, can have its own set of flow channels identified by their respective flow IDs. The mapping from the incoming flow ID to the outgoing flow ID used on the next link can be done by looking up an output flow channel table (OFCT) 416. OFCT 416 can perform a lookup using a match pattern that is a combination of the local input port number corresponding to link 403

and the packet's flow ID which is produced by EFCT 412. If a match is found, then the flow has already been defined, and the packet's flow ID is updated with the value corresponding to the match pattern (this new outgoing flow ID is to be used by the downstream next-hop switch 408). If a match is not found, then a new flow channel can be allocated with a new, outgoing flow ID, which can be mapped to the input port number and the previous, incoming flow ID. An entry including the outgoing flow ID, input port number, and incoming flow ID can be stored in OFCT 416.

[0066] In the case where the packet is the first packet in the flow, a lookup in OFCT 416 would not produce any mapping. In turn, OFCT 416 can allocate for the packet a flow channel with a flow ID to be used by the input port and IFCT 418 on switch 408. This new flow channel, identified by its flow ID, can be added to the packet header for transmission onto link 417, and can be used by the link partner's (which is switch 408) IFCT 418 to access the flow channel's congestion information. As before, OFCT 424 can further generate a new flow channel if no match is found, using the match pattern of its immediate upstream input port number and flow ID associated with link 417. OFCT 424 can then allocate a new flow channel identified by a new flow ID. Note that OFCT 416 can also function as a forwarding table for ACKs of this flow in the upstream direction. After being forwarded upstream from switch 408 to switch 406, the ACK packet can be updated with the flow ID associated with edge link 403 and forwarded to the appropriate input port on switch 406 as indicated by the corresponding entry in OFCT 416. The ACK packets can be forwarded to the input port by an ACK crossbar switch 415 in the upstream direction.

[0067] Subsequently, when the packet arrives at switch 408, its flow ID can be used to identify an input queue to use and to determine an entry in IFCT 418. If the packet's flow ID has not been previously allocated by switch 408, a new input queue can be provided and a new entry in IFCT 418 can be created. From this point onward, a similar process can be performed to forward the packet across switches 408 and 430 until the packet reaches egress switch 432.

[0068] When the packet reaches switch 432, after the packet is forwarded by a data crossbar switch 423, an ACK generator logic block 420 can generate an ACK packet based on the packet's flow ID and input port number. This ACK packet can then be forwarded in the upstream direction by an ACK crossbar switch 422. At the same time, based on the ACK packet, an IFCT 421 can update the state information for the flow in the corresponding table entry. When the ACK packet reaches switch 430, an OFCT 419 can be looked up to determine the upstream flow ID and upstream input port to which the ACK packet is to be forwarded. The ACK packet can then have its flow ID updated and be forwarded to the appropriate input port in

the upstream direction. As the ACK packet traverses the data path upstream in a similar way, the IFCT at each switch can update its table entry for the flow based on the ACK.

[0069] Note that the flow_extent variable can be an important parameter, because it represents the total amount of downstream packet data for a flow. A flow channel is considered free to be reallocated to another flow when the flow_extent of an entry is zero. In general, on receipt of a new packet, the input logic can make a request to send data to an output port. The selected output port can be a function of the flow_extent stored in the IFCT. If flow_extent is zero, there are no packets downstream in the flow to the destination egress edge port. As a result, the switch can use a load based adaptive route selection to choose any valid path that leads to the destination. In a multi-path network, dynamic adaptive routing can be done without the packet being reordered. If flow_extent is not zero, and if in-order delivery is required, the packet can use the same route taken by previous packets. The IFCT can have a field that stores a previous output port number, which is loaded when a packet request is made to an output port and can be used to ensure a connection to the previously used output port.

[0070] As mentioned before, the flow channels can use a match function to recognize packets belonging to an existing flow. Received Ethernet frames or other types of packets can be parsed in real time when the frame or packet is received on an ingress edge port and some fields of the packet header can be used for a lookup in a CAM or Ternary Content Addressable Memory (TCAM). If there is a match, the match address can become the flow ID used to select a flow channel. When no match occurs, the switch hardware can load the pattern that fails to match directly onto a free line of the CAM, which can be done without additional delay. As a result, any following packet can be matched to this new entry without significant amount of buffering. The free entry chosen becomes the new flow ID for the new flow channel entry. Note that no external software intervention is required for the loading of the new entry. The process can be completed autonomously by the switch hardware.

[0071] The de-allocation of flow IDs and corresponding CAM match lines can also be automatically performed by the hardware when the last ACK is returned for the flow. The de-allocation can occur in hardware with respect to potentially matching new packets, without external software intervention.

[0072] In some examples, ingress edge switch 406 can include a fine-grain flow control logic block 434, which can communicate with a network interface controller (NIC) 401 on host 402 to apply flow control on a per-flow basis. More details on fine-grain flow control are provided below in conjunction with the description on congestion management.

[0073] FIG. 4B shows an example of an EFCT. In this example, an EFCT can include a data_flow field 454, an ACK_flow field 456, and optionally additional fields. The EFCT can be associated with an input port, and entries in the EFCT can be indexed by flow_ID values, such as flow_ID 452. In one embodiment, the match pattern field can reside in the match function logic block, which can include a CAM or TCAM. The match function logic block can use the match pattern to generate the flow_ID value, which in turn can be used as an index to the corresponding EFCT entry. From this EFCT's perspective, the flow_extent (i.e., data_flow - ack_flow) can include all the unacknowledged data downstream of this table, which can include the local flow_queue plus the corresponding IFCT's flow_extent value.

[0074] FIG. 4C shows an example of an IFCT. In this example, an IFCT can be associated with an input port, and can include a follow_port field 466, a next_data_flow field 468, a data_flow field 470, an ACK_flow field 472, an ep_congestion field 474, an upstream metering (UM) flag field 477, a downstream metering (DM) flag field 478, and optionally additional fields. An incoming packet's flow_ID value, such as flow_ID 464, can be used as an index to look up the output port number, which is indicated by follow_port field 466, and the state information associated with the corresponding flow. Congestion-control information associated with endpoint congestion (such as ep_congestion field 474) and (hop-by-hop credit-based flow control (such as UM flag field 477 and DM flag field 478), which is described in more detail later in this document, can also be stored in the IFCT. The IFCT can further store information related to dynamic routing associated with different flows.

[0075] FIG. 4D shows an example of an OFCT. In this example, an OFCT can be associated with an output port, and can include an input_port field 482, an input_port_flow_ID field 484 (which corresponds to a packet's existing flow_ID upon its arrival at an input port), a data_flow field 486, an ACK_flow field 488, and optionally additional fields. Data_flow field 486 and ACK_flow field 488 can be used to determine the value of flow_extent from this OFCT onward. The combination of input_port field 482 and input_port_flow_ID field 484 (which can also be referred to as "incoming flow_ID") can be used to determine or allocate the outgoing flow_ID of a packet that is ready for transmission onto the outgoing link corresponding to this OFCT. In one embodiment, the outgoing flow_ID values, such as flow_ID 486, can be used as an index to look up entries in the OFCT.

Congestion Management

[0076] As described above, each flow at a given switch can have its own private queue of packets. This configuration facilitates separate flow control for each flow. As a result, the network can remain mostly lossless, and one flow using a link can be blocked without blocking

any of the other flows using the same link. Unlike a traditional packet switched network, congestion in one part of the network can only affect the flows that are contributing to the congestion. For example, in a conventional network, the buffers before a congested link can quickly fill up with the packets causing the congestion. This in turn can force the switch to issue a pause command or use some other flow control method to prevent neighboring switches from sending packets toward the congested link. Consequently, the packets causing congestion can be stopped or slowed down, and all other packets, which may not be heading to the congested link, can also be stopped or slowed down. As a result, the congestion could spread sideways and increase the size of the saturation tree from a topological perspective.

[0077] In contrast, with flow channels, the load corresponding to flows contributing to congestion can be reduced on the links leading up to the congestion. This reduction of load can allow other flows that are sharing these links to use more link bandwidth and deliver their payload more quickly, while only the packets contributing to the congested link are slowed down.

[0078] Typically, conventional networks can operate normally provided the network load is not at or near full capacity. This can be the case for small or medium sized networks most of the time. With large or very large networks operating with multiple bandwidth-hungry applications, however, at any point in time part of the network can be saturated with traffic load. Under these circumstances, unfair packet delivery could occur even if individual switches implement locally fair policies.

[0079] FIG. 5 shows an example where an unfair share of link bandwidth can occur in a network. In this example, each of the sources A to K is trying to send a stream of packets to destination L, forming an incast scenario where multiple sources are sending packets to a single destination. Source nodes A, B, and C are coupled to switch 502; source nodes D, E, and F are coupled to switch 504; source nodes G, H, and I are coupled to switch 506; and source nodes J and K, and destination node L are coupled to switch 508. Assume that each switch has a fair arbitration policy of selecting an equal number of packets from each of its input ports to any particular output port. However, as can be seen in FIG. 5, sources closer to the destination can receive a much higher proportion of the final link bandwidth than sources the traffic of which needs to pass through more stages of switching. Switch 508 has three sources of incoming data from nodes J, K and switch 506, and can divide the bandwidth on the outgoing link to node L equally among each source. Hence, nodes J, K can each take 33.3% of the bandwidth on the outgoing link toward destination node L.

[0080] The next nearest switch, which is switch 506, can do the same and so on. In this example, with only four stages of switches and only three or four inputs on each stage, and only

with a total of 11 inputs trying to send to the destination node L, three input sources (nodes A, B, and C) only take 1/48 the bandwidth taken by two other input sources (nodes J and K) on the outgoing link toward destination node L. Hence, even with locally fair arbitration policies, nodes that are far away from the destination can suffer from very unfair treatment. A more realistic network topology can involve more switching stages, greater numbers of switch inputs, and more sources trying to send to a single destination. A moderate-sized incast could result in six orders of magnitude difference between the delivered bandwidths among different sources.

[0081] The unfairness problem described above is often caused by the fact that the arbitration policies implemented by a switch are based on the input ports. That is, the bandwidth throttling is done with a per-port granularity. In contrast, by facilitating flow channels and implementing flow-specific throttling, a network can significantly reduce the amount of unfairness among different flows. For example, in the scenario shown in FIG. 5, when the switches implement a fair per-flow bandwidth allocation policy, all the eight source nodes can take substantially equal share of the bandwidth of the edge link between switch 508 and destination node L. By providing a much fairer flow based arbitration policy, extreme tail latencies of individual packets can also be substantially reduced. For large system installations, controlling the maximum latencies through a network is often a major concern for architects. Often, this can only be achieved by restricting the input bandwidth into a network to a small percentage of the peak bandwidth. For example, a input bandwidth limit of 20% of the peak bandwidth can be typical for large datacenters. With flow channels and proper control mechanisms, in contrast, it is now possible to build a network that does not impose such restrictions.

[0082] In addition to fairness, another challenge faced by network architects is congestion. In general, two types of congestions can occur in a network. The first type is endpoint congestion, where an egress edge link coupled to a destination device is congested. The second type is fabric link congestion, where an intermediate fabric link is congested.

[0083] FIG. 6 shows an example of endpoint congestion. In this example, two source hosts 602 and 604 are sending data to a destination host 606. Traffic from source hosts 602 and 604 converges at edge switch 610, and an egress edge link 608 between switch 610 and host 606 can become congested. This congestion scenario can typically occur with incast, where multiple sources are sending traffic to a single destination. Congestion can occur when egress edge link reaches its full data rate capacity, or when destination host 606 cannot process all the incoming packets at a sufficiently fast rate. In any case, the output transmission buffer on switch 610 that is coupled to link 608 can experience an increase in its stored data amount when endpoint congestion occurs.

[0084] A switch can detect and mitigate endpoint congestion by monitoring the output buffer on an egress edge link and by sending ACKs with congestion information to upstream switches and source nodes. More specifically, the output buffer coupled to an egress edge link can monitor the state of the buffer and detect congestion when certain criteria are met. When a packet arrives at or leaves an output buffer, the output buffer can compute three congestion-detection parameters, such as: (1) the amount of data stored in the buffer, (2) the number of packets stored in the buffer, and (3) the rate of change of buffer depth (amount of data stored in the buffer). Three threshold values can be set respectively for these three monitored parameters, although more or less can be set. Congestion is considered to be present when at least one of these parameters exceeds the corresponding threshold.

[0085] When congestion is detected, the switch can generate and transmit an endpoint-congestion-notification ACK corresponding to the packet that has just entered the output buffer. The ACK can include a value indicating the severity of the congestion. Note that this endpoint-congestion-notification ACK is not intended to notify upstream switches of the successful delivery of the packet, but to inform them of the presence and degree of congestion at the egress edge link. (In fact when this endpoint-congestion-notification ACK is sent, the packet may still be stored in the output buffer waiting to be transmitted onto the egress edge link.) This fast, explicit congestion notification mechanism allows the switches to act quickly on a specific flow contributing to the congestion.

[0086] In addition, the output buffer can update the congestion-detection parameters when a packet is dequeued and transmitted onto the egress edge link. If no congestion is present, a regular ACK is generated and sent, which can clear any previous congestion notifications received by the upstream switches operating on the corresponding flow. If congestion is present, the ACK can be marked with a flag, which allows the ACK to notify the switches of persistent congestion at the egress edge link as well as the successful delivery of the packet.

[0087] FIG. 7A shows a flow chart of an exemplary process of generating an explicit endpoint-congestion-notification ACK. During operation, the system can continuously monitor an egress edge link's output buffer. The system can then receive a packet at the output buffer (operation 702). Upon receipt of the packet, the system can compute the three congestion parameters (total amount of data, total number of packets, and rate of change of buffer depth) for the output buffer (operation 704). The system can further determine whether any of the parameters exceeds a corresponding threshold (operation 706). If at least one parameter exceeds the threshold, congestion is considered to be present. Accordingly, the system can generate and send an explicit endpoint-congestion-notification ACK packet corresponding to the packet's flow

to the upstream switches (operation 708). If no congestion is detected, the system can return to normal operation.

[0088] FIG. 7B shows an exemplary endpoint congestion management logic block. In this example, an endpoint congestion management logic block 730 can include an output buffer monitor 732, a congestion parameter computation logic block 734, and an endpoint-congestion-notification ACK generation logic block 736. During operation, output buffer monitor 732 can monitor the state of an output buffer associated with an egress edge link. Based on the state of the monitored output buffer, congestion parameter computation logic block 734 can compute the three congestion parameters (see operation 704 in the flow chart in FIG. 7A). When one of these parameters exceeds the corresponding threshold, endpoint-congestion-notification ACK generation logic block 736 can generate an endpoint-congestion-notification ACK and transmit the ACK to the upstream switch.

[0089] FIG. 8 shows a flow chart showing of exemplary process of generating an ACK in response to a packet being dequeued from an output buffer. In this example, the system first dequeues a packet from the output buffer (operation 802). The system can then compute the three congestion parameters (total amount of data, total number of packets, and rate of change of buffer depth) for the output buffer (operation 804). The system can determine whether any of the parameters exceeds a corresponding threshold (operation 806). If at least one parameter exceeds the threshold, congestion is considered to be present. Accordingly, the system can generate an ACK packet with a marked flag indicating persisting congestion (operation 808). If no congestion is detected, the system can generate a regular ACK packet (operation 809). The system can subsequently send the ACK packet to the upstream switches (operation 810), and transmit the dequeued data packet onto the egress edge link (operation 812).

[0090] Note that the endpoint congestion management logic block shown in FIG. 7B can also perform the operations described by the flow chart shown in FIG. 8. In other words, endpoint congestion management logic block 730 can potentially generate endpoint-congestion-notification ACKs upon the arrival of a packet at the output buffer as well as the departure of the packet from the output buffer.

[0091] As an endpoint-congestion-notification ACK traverses the fabric, the IFCT's of the switches along the path can apply bandwidth limitations to the flow corresponding to the ACK. Effectively, the fabric can slow down the delivery of that flow in a distributed way at each switch along the data path. When an endpoint-congestion-notification ACK passes an IFCT its value can be stored in the flow's table entry as an ep_congestion value, which can be used to select a desired maximum bandwidth for the flow. Each value of ep_congestion can have a corresponding set of high, target, and drop watermark values. For high levels of congestion,

when ep_congestion has a high value, the watermark values can have lower values, so that the congestion can be mitigated more aggressively. For low levels of congestion, when ep_congestion has a low value, a different set of greater high, target, and drop watermark values can be used for higher flow bandwidth. For example, a table indexed by the ep_congestion value can be used. For each ep_congestion value, the table can indicate a corresponding set of high, target, and drop watermark values. The entries of this table can be predetermined, so that when an endpoint-congestion-notification ACK is received, the switch can use the ep_congestion value to perform a lookup in this table, and apply the three corresponding watermark values to the identified flow.

[0092] In some cases, if the source is injecting data in a greedy manner, only slowing down the forwarding inside the network might not be sufficient to fully remove the congestion. To address this problem, an ingress edge switch can be configured to instruct the source device (which typically resides outside the fabric) to limit data injection on a fine-grain, per-flow basis. This switch-to-host flow control mechanism can be referred to as Fine Gran Flow Control (FGFC).

[0093] In particular, especially in an HPC environment, an end host or computing node could have a large number of cores running numerous threads, processes, or virtual machines, each of which could be injecting their own stream of data into the network through a common physical network interface controller (NIC). When congestion is present, a per-port based flow control can only throttle the overall data rate over a single port on the NIC, which can be 40 Gb/s or more. Pushing back on the total data rate on the entire port can cause unfairness to flows that are not contributing to congestion. FGFC can extend the concept of the individual flows or group of associated flows to their ultimate source, which can be a single thread executing on one of the cores.

[0094] To slow down data injection from the source, an FGFC logic block on an ingress edge switch (for example, FGFC logic block 434 in edge switch 406 in FIG. 4A) can use a pause-credit hybrid method to throttle incoming data associated a particular flow or group of flows. A pause-based method typically involves a receiving end issuing a pause command to the transmitter end, which in response can stop transmission until further notice. With a credit-based method, the receiving end can issue transmission credits to the transmitting end, which allows the transmitter to send more data but only up to the amount specified by the credit value. This mechanism allows the receiving end to control more precisely its input buffer depth to avoid overflow while allowing transmission to continue. FGFC can use a hybrid method, in which upon detection of congestion the ingress edge switch can issue a FGFC frame for one or more flows with a set timer value to the end host NIC (such as NIC 401 on end host 402 in FIG. 4A).

After the FGFC frame is received, the ingress edge switch may turn on a credit-based flow control mode. In response, the NIC can throttle the transmission data rate for the corresponding flow(s) based on the received credit, while allowing other flows to be transmitted at normal data rate. After the predetermined timer expires, the end host NIC can revert to normal transmission for the throttled flow(s), unless another pause command is received. Note that a throttled flow can be identified by any field derived from a packet. A throttled flow can be specific to a single process or thread executed on the end host.

[0095] FGFC can implement the control communication between an edge switch and an end host NIC using an Ethernet frame with an Organizationally Unique Identifier (OUI) extended Ether_Type field. These frames can indicate one or more of the following: (1) the protocol used by the flow being controlled; (2) an identifier to indicate the source (e.g., application, process, or thread) generating the packets that need to be throttled; (3) a pause time value for which the flow control is to last (which can prevent a lockup if subsequent FGFC frames are lost due to errors), and (4) a credit value, which can be zero, to indicate the number of frames or amount of data that can be sent during the pause period.

[0096] Note that the identifier for indicating the source flow subject to flow control can be different based on the protocol associated with the flow. For layer-2 Ethernet virtual local area network (VLAN) traffic, the identifier can include the VLAN number. For IPv4 traffic, the identifier can include a source/destination IP address pair, a UDP or TCP/IP 5-tuple that includes UDP or TCP port numbers, or an optional flow label. For IPv6 traffic, the identifier can include one or more IPv6 addresses or an IPv6 flow label. For proprietary HPC protocol traffic, the identifier can include a process or thread ID. In general, this identifier is also stored in the EFCT of the edge switch, since it is used to map the corresponding traffic to a flow ID.

[0097] To trigger FGFC, the IFCT of an ingress edge switch can monitor its flow-specific input queues. For each queue, the corresponding IFCT entry can indicate three watermark values: high, target, and drop, which can be used to measure the queue depth. In some examples, these watermark values can be included as additional fields in the IFCT as shown in FIG. 4C, or can be stored in a separate table and linked by a field in the IFCT. When the queue depth is less than the target value, no FGFC is necessary. When the queue depth reaches the target watermark value, the IFCT can communicate with an FGFC logic block to initiate FGFC with an end host's NIC. When the queue depth reduces to below the drop watermark value, FGFC can be stopped and normal transmission for the flow can be resumed.

[0098] FIG. 9A shows a flow chart of an exemplary FGFC process. During operation, at an ingress edge switch, the system can monitor the flow-specific input queues (operation 902). The system can further determine, for a respective flow, whether FGFC is currently turned on

(operation 904). If FGFC is currently turned on for this flow, the system can then determine whether the queue depth is below the drop watermark (operation 906). If the queue depth has not reduced to below the drop watermark, the system can continue the credit based transmission in the FGFC mode (operation 912). If the queue depth has reduced to below the drop watermark, the system can revert to normal transmission for the flow (operation 914). Referring back to operation 904, if FGFC is currently not turned on, the system can determine whether the queue depth is greater than the target watermark (operation 908). If so, the system can initiate FGFC for the flow (operation 910). The FGFC logic block in the edge switch can obtain flow identifying information (e.g., VLAN tag, TCP/IP 5-tuple, thread ID, etc.) from the EFCT entry corresponding to the flow and send an FGFC Ethernet frame to the NIC on the end host. Subsequently, the system can continue to monitor the input queues (operation 902). If the queue depth is not greater than the target watermark, the system can continue regular data transmission (operation 914)

[0099] To facilitate FGFC, a NIC can be configured to process the FGFC Ethernet frame, so that the NIC can communicate to the application or process on an end host that is generating the data. Parsing of the FGFC Ethernet frame and communication to the application or process can be done in software, hardware, or a combination of both. FIG. 9B shows an example of a FGFC-enabled NIC. In this example, a NIC 930 can include a processor 932, a memory 934, a transmitter 936, a receiver 938, a FGFC logic block 940, and a communication logic block 942. During operation, transmitter 936 and receiver 938 can perform communication to and from an edge switch via an edge link. Communication logic block 942 can perform communication via a data bus (such as a Peripheral Component Interconnect Express (PCIe) bus) with the central processing unit of the end host in which NIC 930 resides. Processor 932 and memory 934, which are internal to NIC 930, can perform local processing of the data. During operation, FGFC logic block 940 can work with an edge switch to apply FGFC on a per-flow basis. In addition, FGFC logic block 940 can communicate via communication logic block 942 with the end host's central processing unit to throttle and data injection of an individual application or process corresponding to the specific flow subject to FGFC, thereby controlling the amount of data injected into the fabric.

[00100] As mentioned above, two types of congestions can occur in a network. A first type is endpoint congestion, and a second type is fabric link congestion. FIG. 10 shows an example of fabric link congestion. In this example, two intermediate switches 1002 and 1006 are in communication via a fabric link 1004. Multiple source/destination pairs can be sending traffic via fabric link 1004. As a result, fabric link 1004 can experience congestion, although the links

leading up to and away from fabric link 1004 might not be congested. Fabric link 1004 can appear to be a “hot spot” when such congestion occurs.

[00101] To mitigate fabric link congestion, a switch can apply dynamic per-flow credit-based flow control. At a switch, if an input queue starts to fill up, and the queue_extent value for this flow reaches a predetermined threshold, the switch can generate a special ACK to notify the upstream switch’s IFCT of the congestion. This special per-hop ACK can be referred to as “HeadroomACK.” Upon receiving the HeadroomACK, the upstream switch’s IFCT can start a credit based flow control with the downstream switch. In the downstream IFCT entry, a flag Upstream Metering (UM) can be set to indicate that the data transmission from the upstream switch is now metered based on the credits. The HeadroomACK packet can also include a credit value.

[00102] When the upstream switch receives a HeadroomACK, a flag called Downstream Metered (DM) can be set in the corresponding entry of the IFCT. The IFCT can also store a signed headroom field in the IFCT entry with the credit value carried by the HeadroomACK (i.e., the headroom value indicates the number of credits). This headroom field can represent the amount of data that can be forwarded to the downstream switch. This establishes a credit based flow control for the corresponding flow. If the upstream IFCT receives a HeadroomACK while the DM flag in the flow’s entry is already set, the credit value carried by the HeadroomACK can be added to the existing headroom value.

[00103] New packets received by the upstream IFCT can be blocked if the headroom value is not greater than zero (i.e., there is no credit available). These packets can fill this flow’s input queue and may in turn cause the IFCT to initiate per-flow credit based flow control with its upstream IFCT, and so on. If the headroom value is greater than zero, a packet stored in the input queue can be dequeued and forwarded to the downstream switch, and the headroom value can be decremented by the size of the forwarded packet, which may cause the headroom value to become zero or negative.

[00104] With the flow restricted from sending new packets to the downstream IFCT, the downstream IFCT’s input queue can start to drain at some rate depending on its downstream congestion. As described above, each flow’s input queue can have three queue-depth watermark values, namely high, target, and drop, which can be used to manage credit-based flow control. The target watermark can be approximately the ideal queue depth for the desired flow bandwidth. It indicates sufficient buffering is available for transmitting data downstream. When there is congestion, the credit-based flow control mechanism can attempt to keep the flow’s queue_extent value approximately at this target watermark.

[00105] If the queue_extent value is between the high watermark and drop watermark, and is greater than the target watermark, when a packet is forwarded, slightly less than this packet's size of credit can be returned with a HeadroomACK to the upstream switch. If the queue_extent value does not exceed the target watermark, when a packet is forwarded, slightly more than this packet's size of credit can be returned with the HeadroomACK to the upstream switch.

[00106] If the queue_extent depth is greater than the high watermark, no credit is returned when packets are forwarded. This mechanism can bring the queue_extent value down more quickly and is usually used when congestion is detected for the first time. If the congestion clears, the flow's input queue can start to empty more quickly. When the queue depth is less than the drop watermark, the credit-based flow control can be switched off. This can be done by clearing the UM flag in the IFCT entry and returning a HeadroomACK with the maximum credit value to the upstream switch. When received by the upstream IFCT the HeadroomACK clears the entry's DM flag and flow control against the headroom value is turned off.

[00107] Note that in a typical network topology there can be a number of switches and between two endpoints there can be multiple data paths. In a multi-path network, it is possible to use various methods to control fabric link congestion. For example, the injection limits, described later in this document, can control the maximum total amount of data in the entire fabric. This means that if a particular fabric link is overloaded, a flow can use a different data path that does not go through the congested link. It is possible to detect an overloaded link and generate "reroute" ACKs for a set of flows. The reroute ACKs can temporarily block the flow in an upstream switch, and when all the ACKs for that flow have been returned, the flow can be unblocked and become free to use a different path across the fabric. A dynamic load-based adaptive routing mechanism can then direct the lead packet to use a different uncongested fabric link. In turn the load across the entire fabric can become more balanced.

[00108] FIG. 11 shows a flow chart of an example process of applying credit-based flow control on a congested fabric link. During operation, a switch system can monitor its flow-specific input queues (operation 1102). The system can determine whether an entry in its IFCT has a UM flag set (operation 1104). If the UM flag is set, which means that credit-based flow control is on, the system can further determine whether the queue_extent value is less than the drop watermark value (operation 1106). If the queue_extent value is less than the drop watermark value, the system can clear the UM flag, turn off the credit-based flow control, and resume normal data transmission (operation 1014). If the queue_extent value is greater than the drop watermark value, the system can continue the credit-based flow control (operation 1106). Referring back to operation 1104, if the UM flag is not set, which means the system is in regular

transmission mode, the system can determine whether the queue_extent value is greater than the target watermark value (operation 1108). If so, the system can initiate credit-based flow control and send a HeadroomACK to the upstream switch (operation 1110). If the queue_extent value is not greater than the target watermark value, the system can continue with regular data

5 transmission (operation 1112).

[00109] In general, a flow channel switch can use a combination of several congestion detection and control mechanisms. For example, different degrees of endpoint congestion can be reported using the endpoint-congestion-notification ACK that can be returned from the final fabric egress edge port. This ACK type can be used to manage the bandwidth of
10 flows into a significantly congested egress edge port. The system can also use a per-hop credit-based flow control to manage fabric link congestion. This per-hop congestion management mechanism can be effective against low to moderate levels of congestion, because the response time can be much shorter than the network-wise round trip delay.

[00110] If the congestion is severe, perhaps caused by a wide incast, the system
15 can also apply a per-flow injection limit. A flow's injection limit can be determined based on the ep_congestion value. The injection limit can be compared with the flow_extent value in all IFCTs the flow passes through. If the flow_extent is greater than this limit the IFCT can block the forwarding of packets from the input queue for this flow. This mechanism can reduce the rate of forwarding of packets over an entire flow to as little as a single packet.

[00111] The system can also protect unrelated traffic from extreme congestion
20 caused by incasts with a large number of contributors. In this case, the ep_congestion value can be set to a high value and the average amount of data of a flow can be reduced to a small fraction of a packet. This can be achieved by only releasing the next packet of an individual flow into the fabric from the ingress edge port's IFCT after a programmable delay has elapsed since when the
25 ACK of the previous packet has been received.

[00112] In addition to per-flow injection limits, the system can measure the amount
of data that has been injected into the fabric on a per-ingress-port basis, and set injection limits to impose a cap on the total amount of data a port can inject into the fabric. Since every ingress
30 port can apply this injection limit, the system can control the maximum amount of data allowed inside the fabric. Limiting the total amount of data into the fabric can ensure that buffer exhaustion does not occur where bandwidth is scarce. As a result, traffic which is not using the
paths with reduced bandwidth are not affected.

[00113] To facilitate per-port injection limit, an IFCT can maintain a total traffic
count. Each time a packet is injected into the fabric from the edge port the total count can be
35 incremented. When a flow's ACK is returned, the total traffic count can be decremented. Once

all the ACKs of all the flows of an ingress port have been returned (i.e., when the sum of the flow_extent values for all the flows becomes zero), the total traffic count can be set to zero.

[00114] FIG. 12 shows an exemplary edge switching system that facilitates flow channels (which, for example, can correspond to switch 406 in FIG. 4A). In this example, a switch 1202 can include a number of communication ports, such as port 1220. Each port can include a transmitter and a receiver. Switch 1202 can also include a processor 1204, a storage device 1206, and a flow channel switching logic block 1208. Flow channel switching module 1208 can be coupled to all the communication ports and can further include a crossbar switch 1210, an EFCT logic block 1212, an IFCT logic block 1214, and an OFCT logic block 1216.

[00115] Crossbar switch 1210 can include one or more crossbar switch chips, which can be configured to forward data packets and control packets (such as ACK packets) among the communication ports. EFCT logic block 1212 can process packets received from an edge link and map the received packets to respective flows based on one or more header fields in the packets. In addition, EFCT logic block 1212 can assemble FGFC Ethernet frames, which can be communicated to an end host to control the amount of data injected by individual processes or threads. IFCT logic block 1214 can include the IFCT, and perform various flow control methods in response to control packets, such as endpoint-congestion-notification ACKs and fabric-link credit-based flow control ACKs. OFCT logic block 1216 can include a memory unit that stores the OFCT and communicate with another switch's IFCT logic block to update a packet's flow ID when the packet is forwarded to a next-hop switch.

[00116] FIG. 13 shows an exemplary intermediary switching system that facilitates flow channels (which, for example, can correspond to switches 408 and 430 in FIG. 4A). In this example, a switch 1302 can include a number of communication ports, such as port 1320. Each port can include a transmitter and a receiver. Switch 1302 can also include a processor 1304, a storage device 1306, and a flow channel switching logic block 1308. Flow channel switching module 1308 can be coupled to all the communication ports and can further include a crossbar switch 1310, an EFCT logic block 1312, an IFCT logic block 1314, and an OFCT logic block 1316.

[00117] Crossbar switch 1310 can include one or more crossbar switch chips, which can be configured to forward data packets and control packets (such as ACK packets) among the communication ports. EFCT logic block 1312 can process packets received from an edge link and map the received packets to respective flows based on one or more header fields in the packets. In addition, EFCT logic block 1312 can assemble FGFC Ethernet frames, which can be communicated to an end host to control the amount of data injected by individual processes or threads. IFCT logic block 1314 can include the IFCT, and perform various flow control methods

in response to control packets, such as endpoint-congestion-notification ACKs and fabric-link credit-based flow control ACKs. OFCT logic block 1316 can include a memory unit that stores the OFCT and communicate with another switch's IFCT logic block to update a packet's flow ID when the packet is forwarded to a next-hop switch.

5 **[00118]** In summary, the present disclosure describes a data-driven intelligent networking system that can accommodate dynamic traffic with fast, effective congestion control. The system can maintain state information of individual packet flows, which can be set up or released dynamically based on injected data. A packet flow can be mapped to their layer-2, layer-3, or other protocol-specific header information. Each flow can be provided with a flow-specific input queue upon arriving at a switch. Packets of a respective flow are acknowledged
10 after reaching the egress point of the network, and the acknowledgement packets are sent back to the ingress point of the flow along the same data path in the reverse direction. As a result, each switch can obtain state information of each flow and perform flow control of a per-flow basis. Such flow control allows the network to be better utilized while providing versatile traffic-engineering and congestion control capabilities.
15

[00119] The methods and processes described above can be performed by hardware logic blocks, modules, or apparatus. The hardware logic blocks, modules, or apparatus can include, but are not limited to, application-specific integrated circuit (ASIC) chips, field-programmable gate arrays (FPGAs), dedicated or shared processors that execute a piece of code
20 at a particular time, and other programmable-logic devices now known or later developed. When the hardware logic blocks, modules, or apparatus are activated, they perform the methods and processes included within them.

[00120] The methods and processes described herein can also be embodied as code or data, which can be stored in a storage device or computer-readable storage medium. When a
25 processor reads and executes the stored code or data, the processor can perform these methods and processes.

[00121] The foregoing descriptions of embodiments of the present invention have been presented for purposes of illustration and description only. They are not intended to be exhaustive or to limit the present invention to the forms disclosed. Accordingly, many
30 modifications and variations will be apparent to practitioners skilled in the art. Additionally, the above disclosure is not intended to limit the present invention. The scope of the present invention is defined by the appended claims.

What Is Claimed Is:

1. A switch, comprising:

a storage device to buffer an incoming packet flow, wherein the packet flow corresponds to packets with one or more common header fields and is identified by an incoming flow

5 identifier that is unique within an input port via which the packet flow is received; and

an input flow logic block to maintain state information of the packet flow, and to free the incoming flow identifier in response to determining that all forwarded data of the packet flow have been acknowledged, thereby allowing the freed flow identifier to be used by another packet flow;

10 wherein the state information indicates an amount of forwarded but unacknowledged data belonging to the packet flow; and

wherein the input flow logic block is coupled to the storage device.

2. The switch of claim 1, wherein the storage device comprises a flow-specific queue that is dedicated to the packet flow, thereby allowing forwarding of the packet flow to be

15 controlled independently from other flows.

3. The switch of claim 1, wherein the input flow logic block is further to update the amount of forwarded but unacknowledged data belong to the packet flow based on received acknowledgement messages, wherein the acknowledgement messages indicate delivery of packets belonging to the packet flow to a destination.

20 4. The switch of claim 1, further comprising:

an output flow logic block to update a packet belonging to the packet flow with an outgoing flow identifier, wherein the outgoing flow identifier replaces the incoming flow identifier.

5. The switch of claim 4, wherein the output flow logic block is further to perform a
25 lookup for the outgoing flow identifier based on a match pattern that indicates the incoming flow identifier and the input port.

6. The switch of claim 5, wherein, in response to the lookup not resulting in a match, the output flow logic block is further to assign a value that is not used for any flow to the outgoing flow identifier.

7. The switch of claim 4, wherein, in response to determining that the switch is a destination for the packet flow, the output flow logic block is further to remove a fabric header from a packet belonging to the packet flow, wherein the fabric header includes the incoming flow identifier.

8. A switch, comprising:
a storage device to store packet header information and flow identifiers, wherein a respective flow identifier is unique within an input port and identifies a packet flow which corresponds to packets with one or more common header fields; and
an edge flow logic block to perform a hardware-based lookup in the storage device for a flow identifier using a match pattern that includes header information of a received packet, and to create an entry in the storage device that maps the match pattern to a new flow identifier in response to the lookup not resulting in a match.

9. The switch of claim 8, wherein the edge flow logic block is further to encapsulate the packet with a fabric header that includes the new flow identifier.

10. The switch of claim 8, wherein the storage device comprises a ternary content addressable memory (TCAM) or content addressable memory (CAM).

11. A method, comprising:
buffering an incoming packet flow in a storage device, wherein the packet flow corresponds to packets with one or more common header fields and is identified by an incoming flow identifier that is unique within an input port via which the packet flow is received;
maintaining state information of the packet flow, wherein the state information indicates an amount of forwarded but unacknowledged data belonging to the packet flow; and
freeing the incoming flow identifier in response to determining that all forwarded data of the packet flow have been acknowledged, thereby allowing the freed flow identifier to be used by another packet flow.

12. The method of claim 11, further comprising providing in the storage device a flow-specific queue that is dedicated to the packet flow, thereby allowing forwarding of the packet flow to be controlled independently from other flows.

13. The method of claim 11, further comprising updating the amount of forwarded but
5 unacknowledged data belong to the packet flow based on received acknowledgement messages, wherein the acknowledgement messages indicate delivery of packets belonging to the packet flow to a destination.

14. The method of claim 14, further comprising updating a packet belonging to the packet flow with an outgoing flow identifier, wherein the outgoing flow identifier replaces the
10 incoming flow identifier.

15. The method of claim 14, further comprising performing a lookup for the outgoing flow identifier based on a match pattern that indicates the incoming flow identifier and the input port.

16. The method of claim 15, further comprising assigning a value that is not used for
15 any flow to the outgoing flow identifier in response to the lookup not resulting in a match.

17. The method of claim 14, further comprising removing a fabric header from a packet belonging to the packet flow in response to determining that the switch is a destination for the packet flow, wherein the fabric header includes the incoming flow identifier.

18. A method, comprising:
20 storing in a storage device packet header information and flow identifiers, wherein a respective flow identifier is unique within an input port and identifies a packet flow which corresponds to packets with one or more common header fields;
performing a hardware-based lookup in the storage device for a flow identifier using a match pattern that includes header information of a received packet; and
25 creating an entry in the storage device that maps the match pattern to a new flow identifier in response to the lookup not resulting in a match.

19. The method of claim 18, further comprising encapsulating the packet with a fabric header that includes the new flow identifier.

20. The method of claim 18, wherein creating the entry comprises storing the entry in a ternary content addressable memory (TCAM) or content addressable memory (CAM).

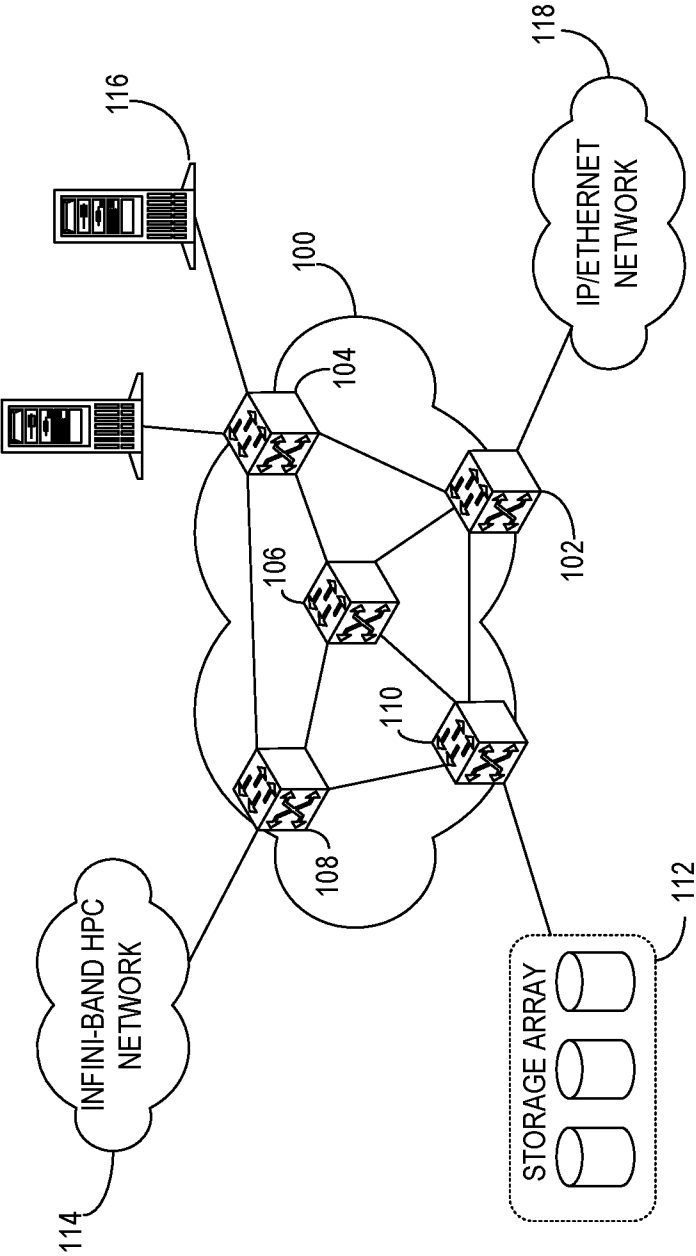


FIG. 1

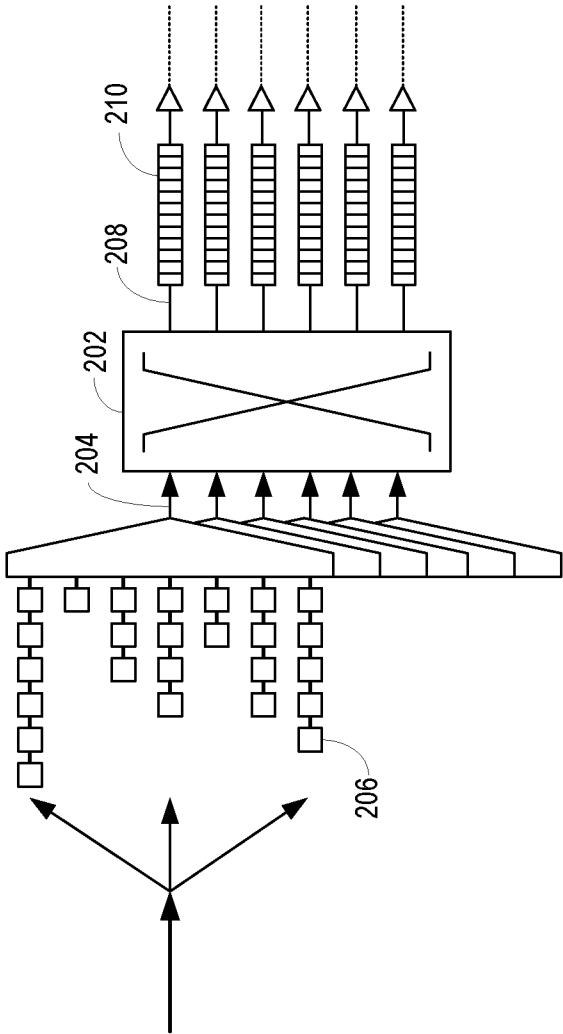


FIG. 2A

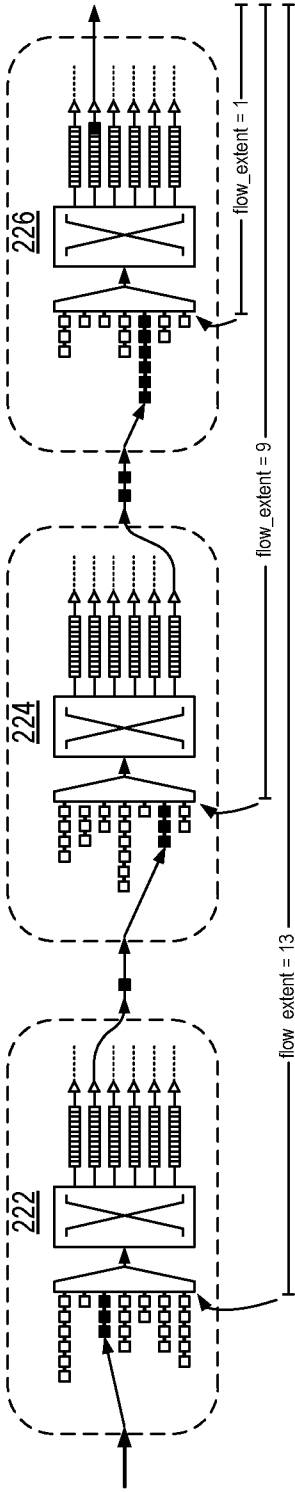


FIG. 2B

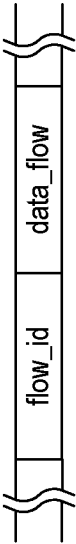


FIG. 3A

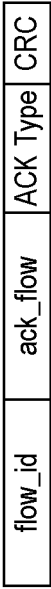


FIG. 3B

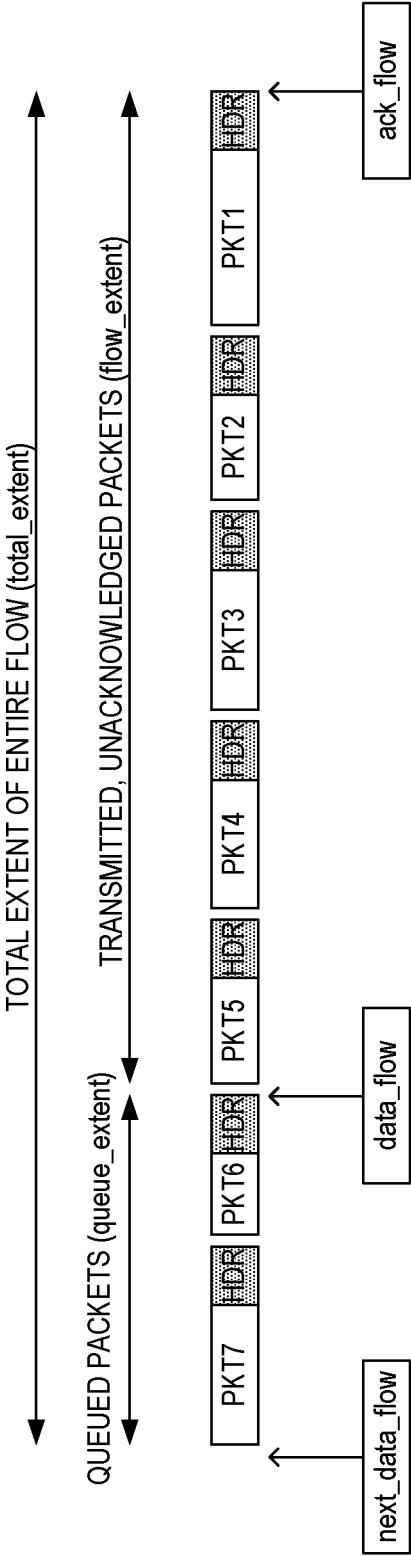


FIG. 3C

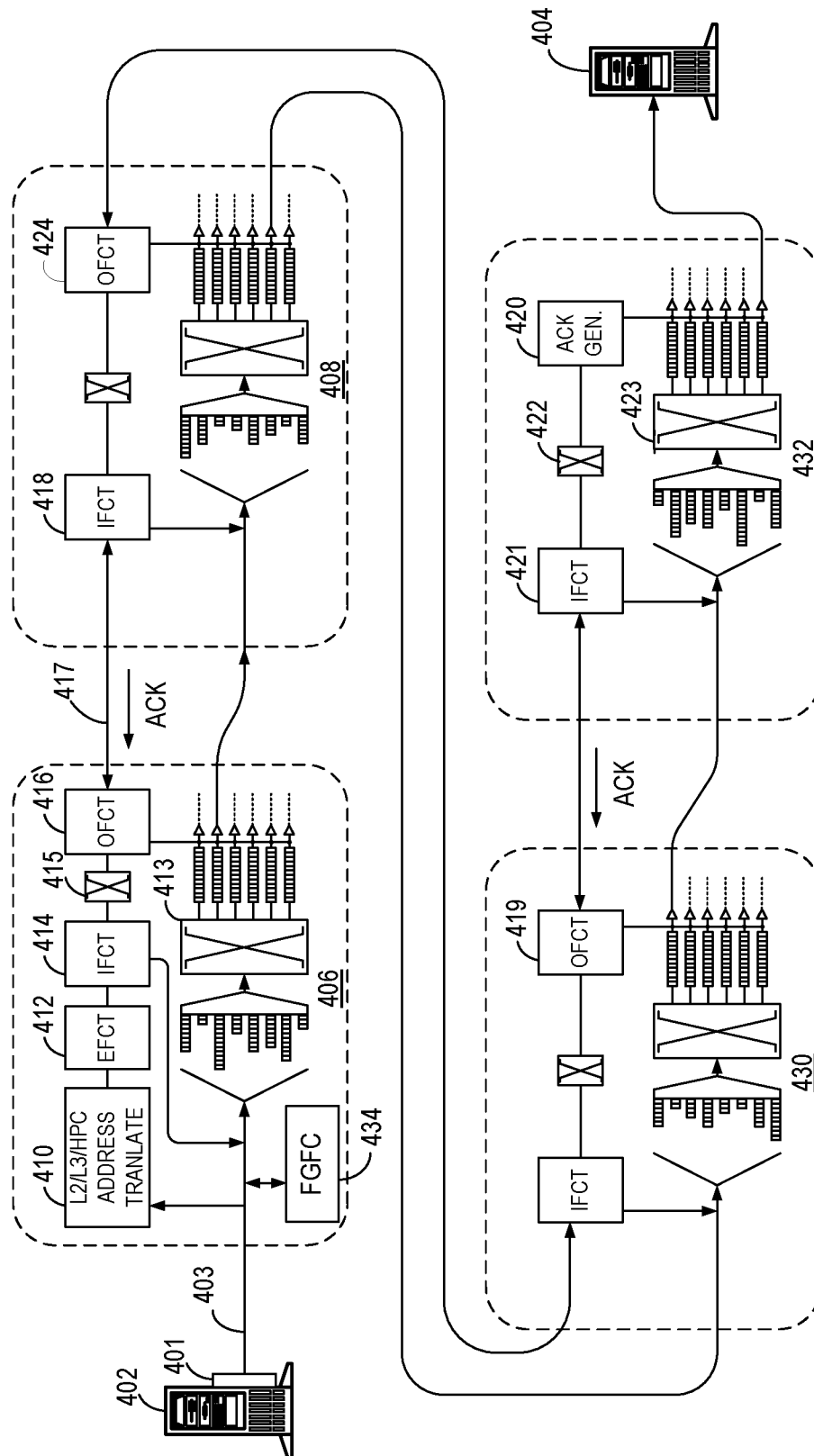
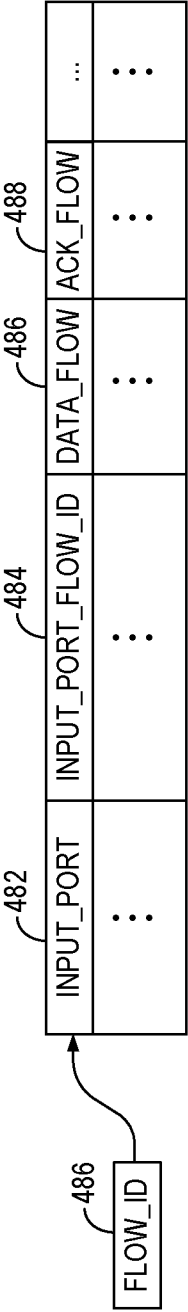
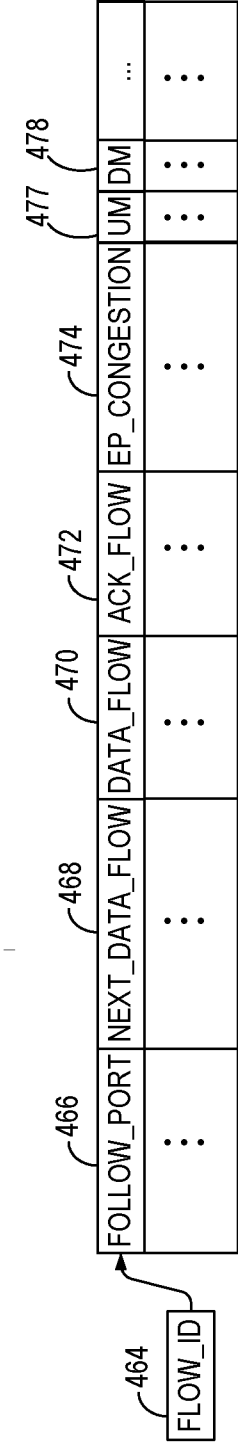
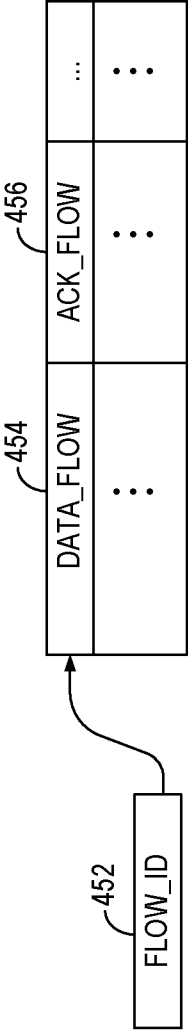


FIG. 4A



6/14

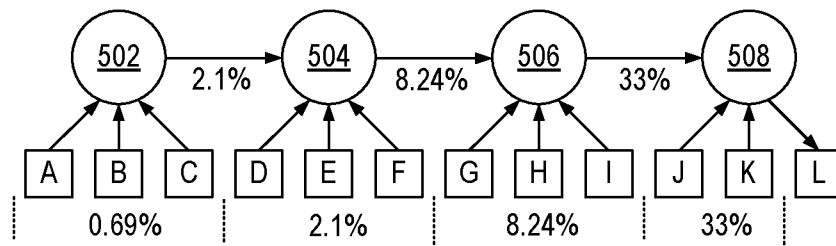


FIG. 5

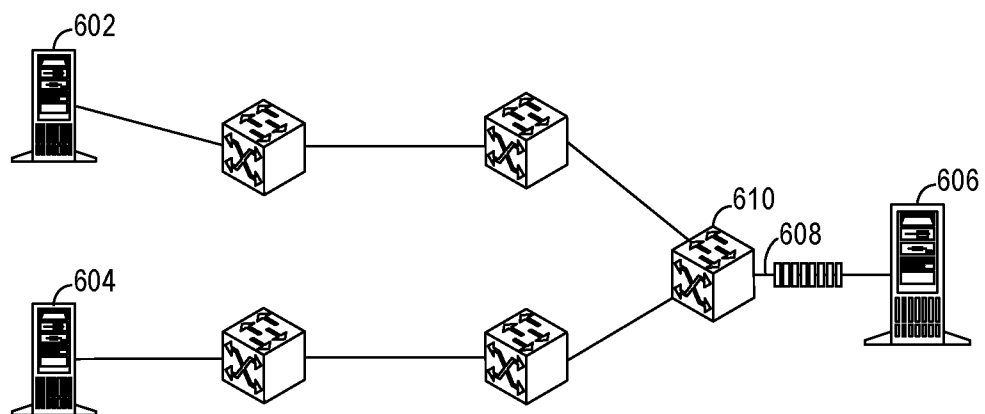


FIG. 6

7/14

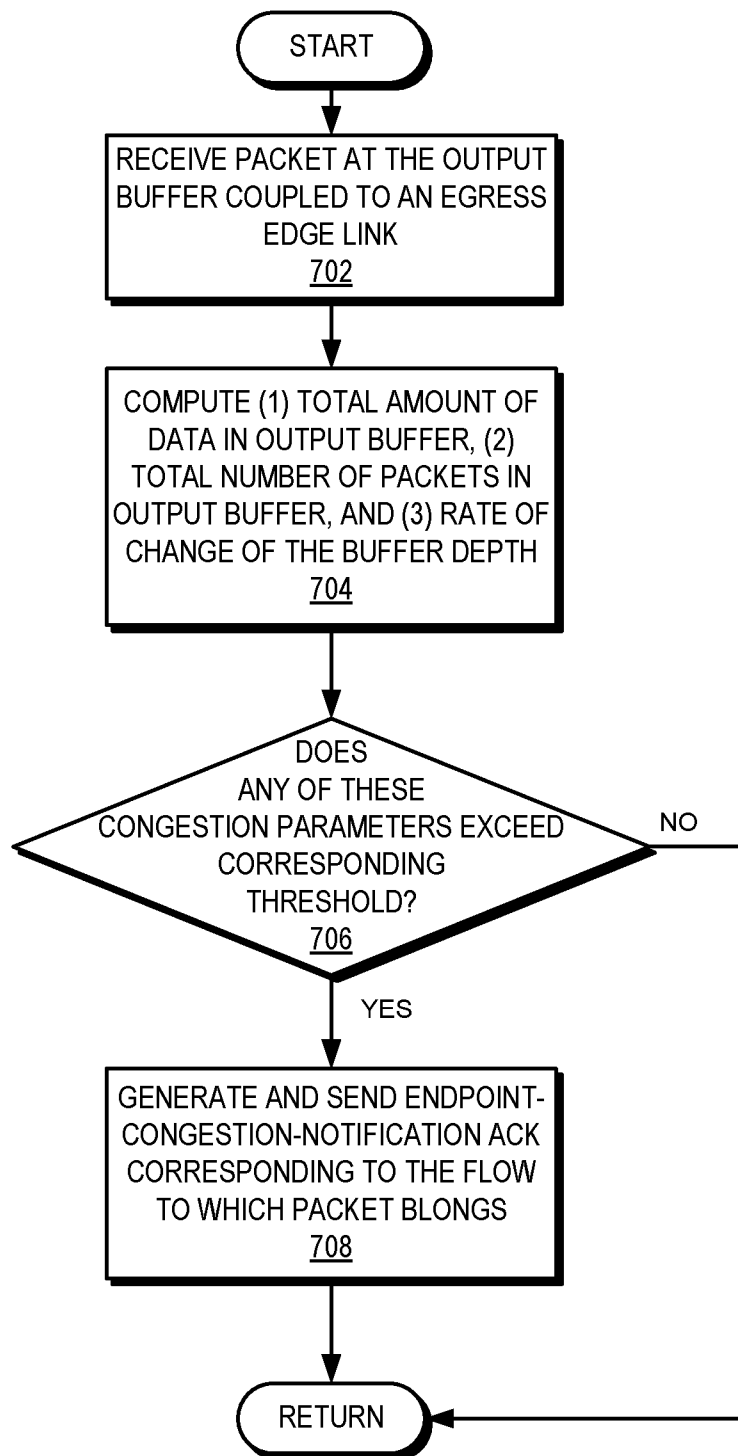


FIG. 7A

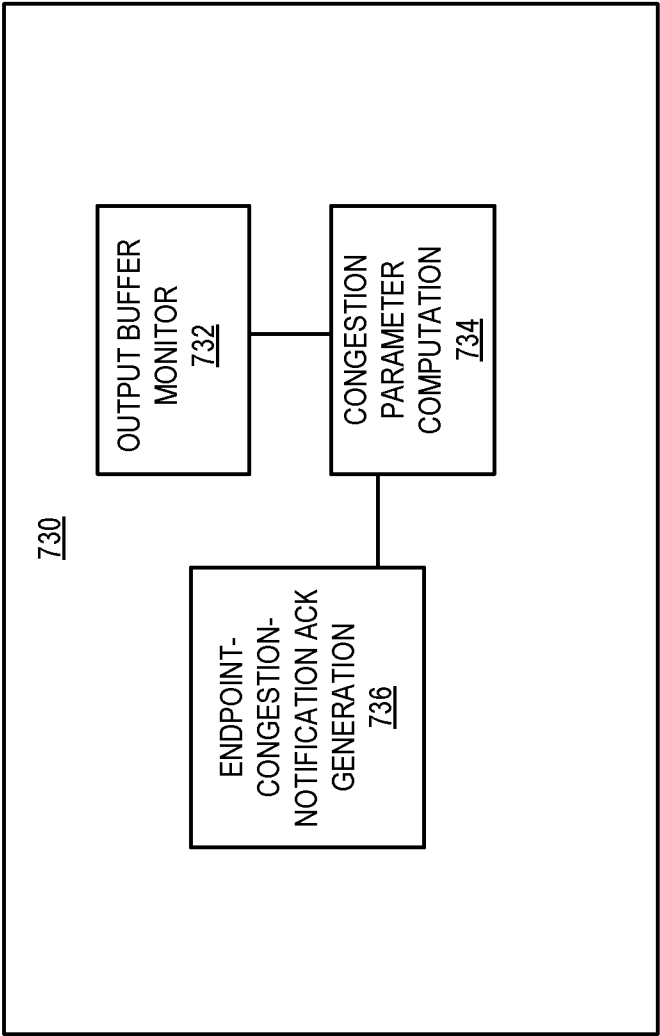


FIG. 7B

9/14

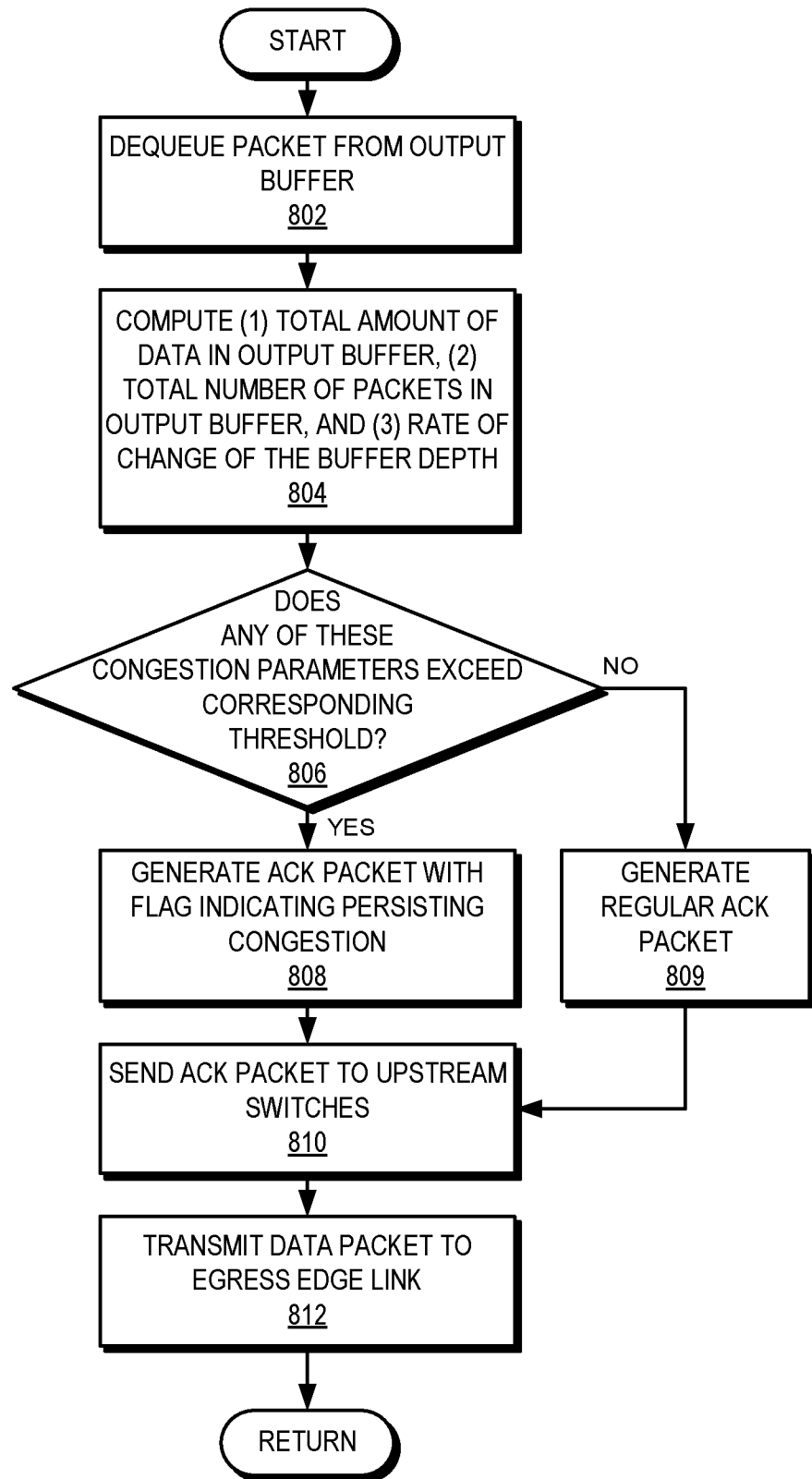
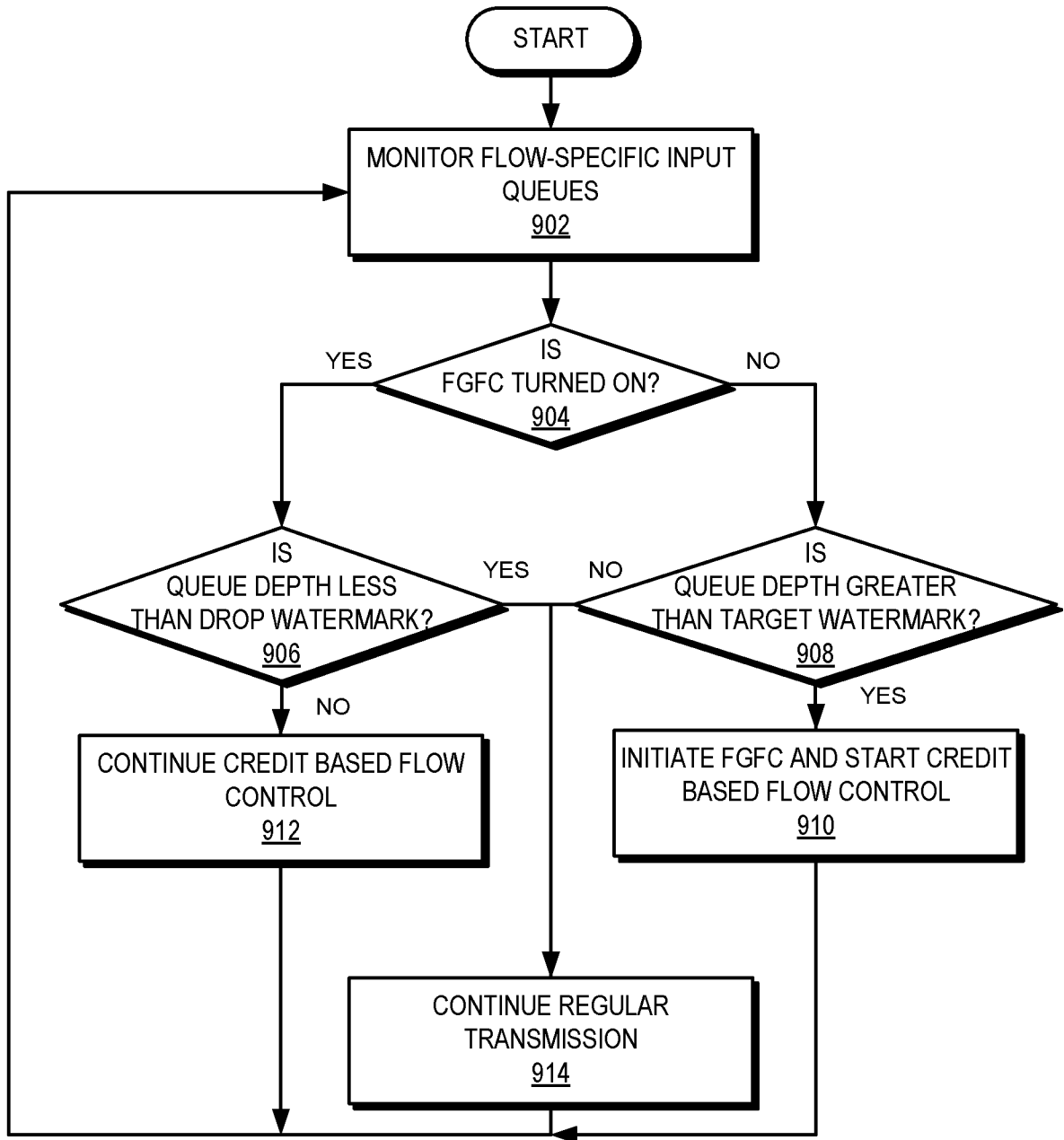


FIG. 8

10/14

**FIG. 9A**

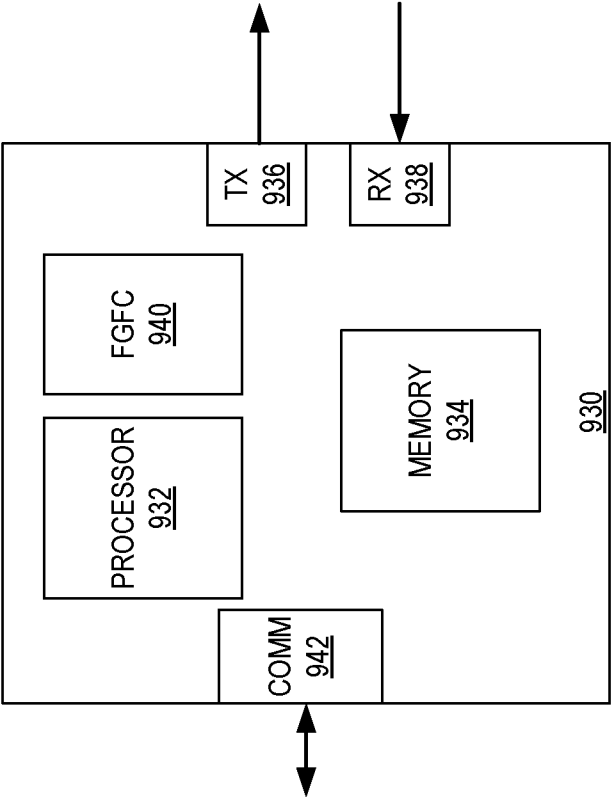


FIG. 9B

12/14

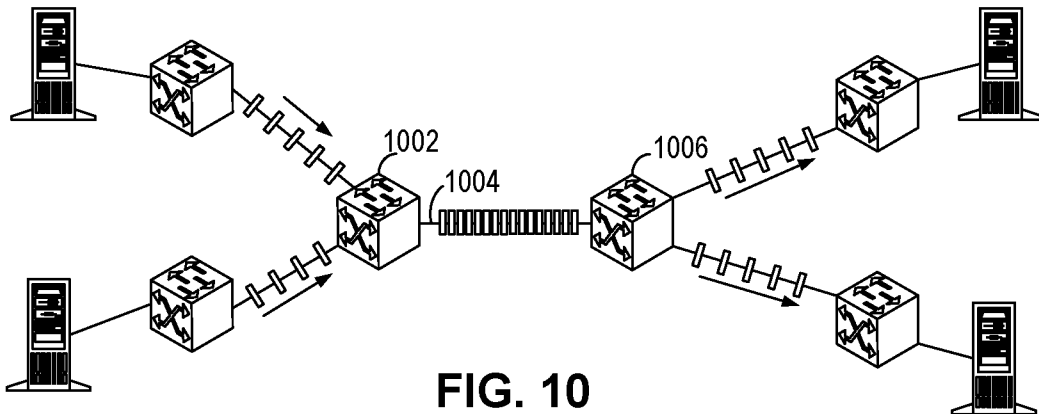


FIG. 10

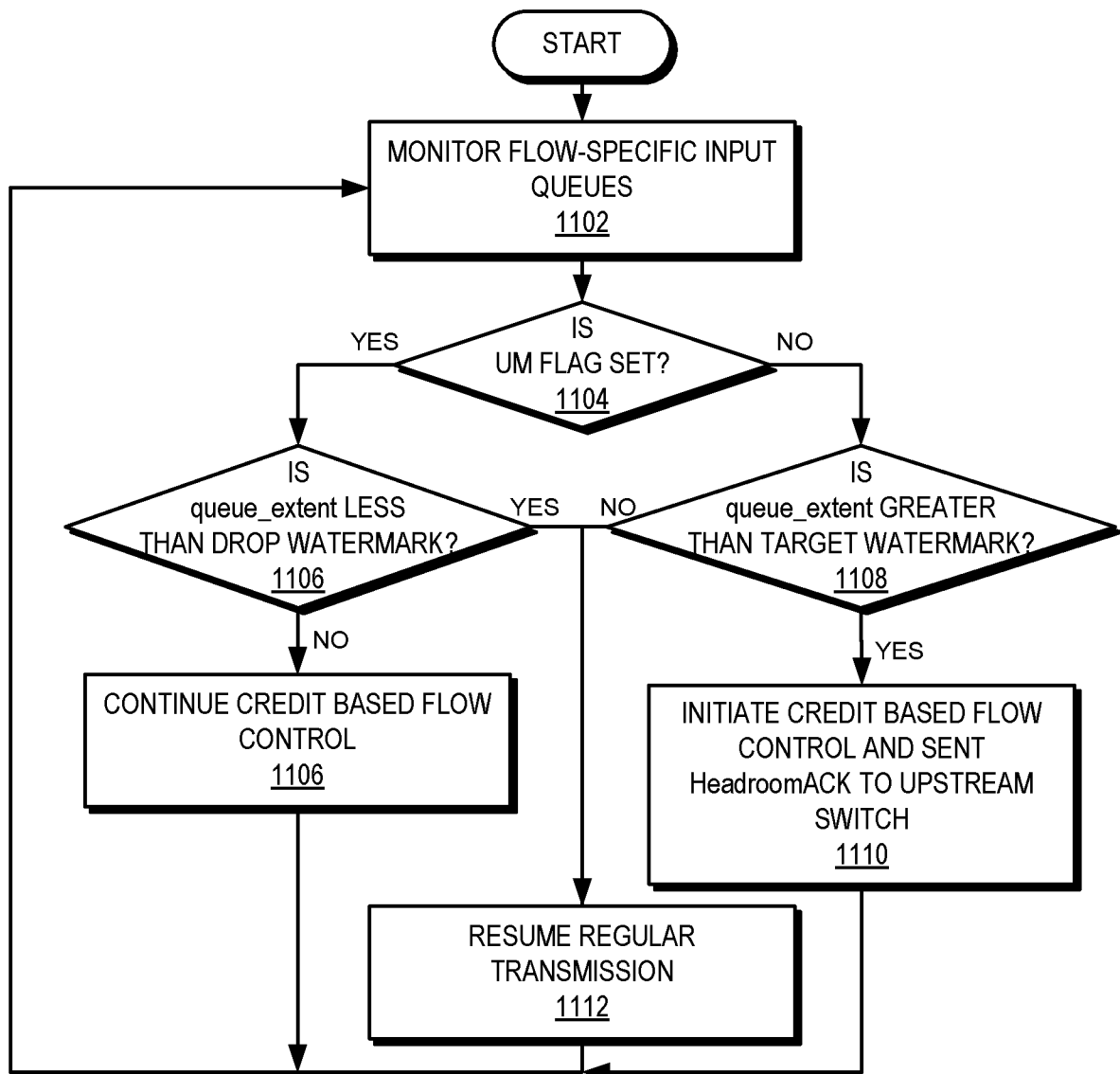


FIG. 11

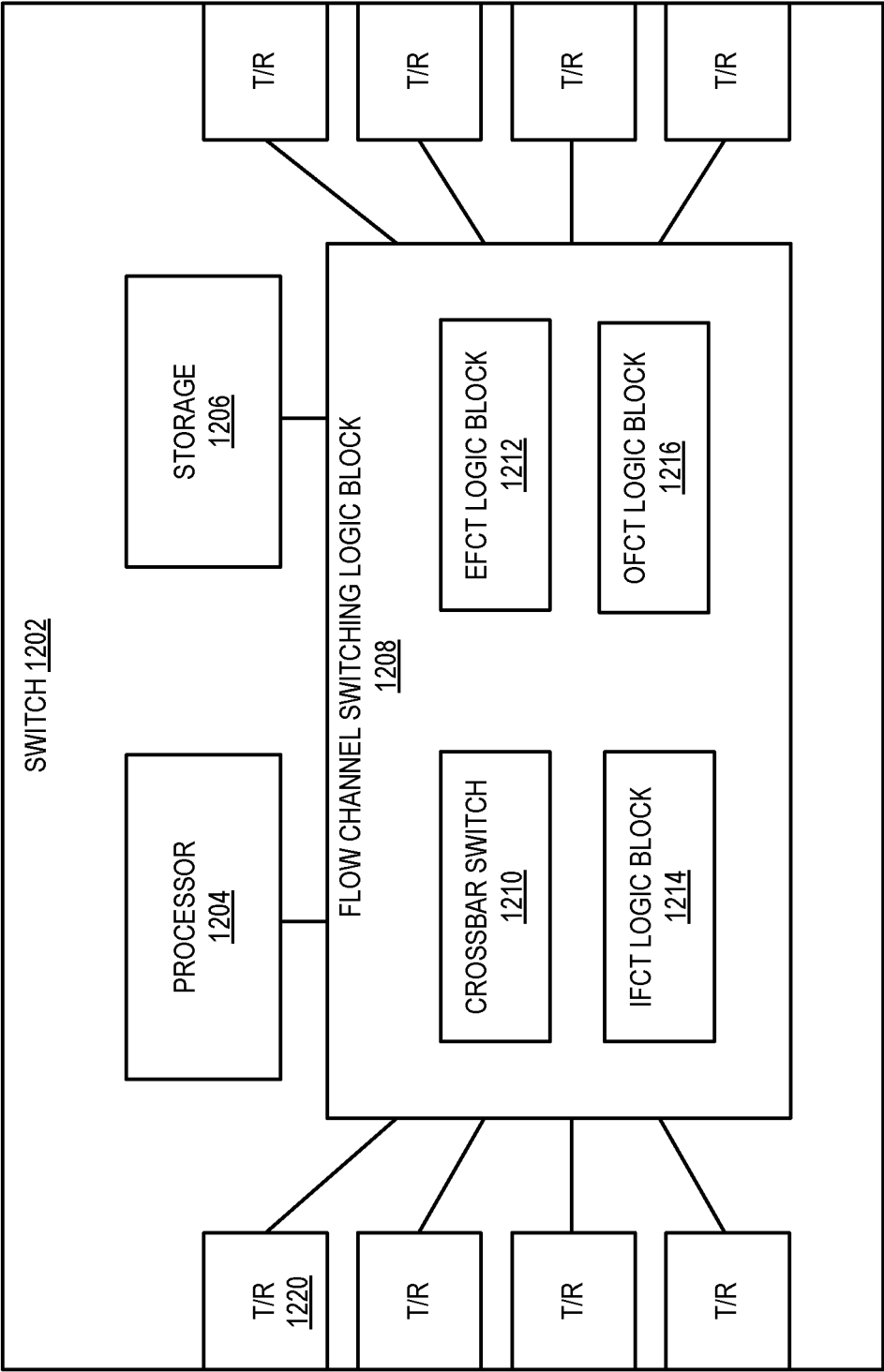


FIG. 12

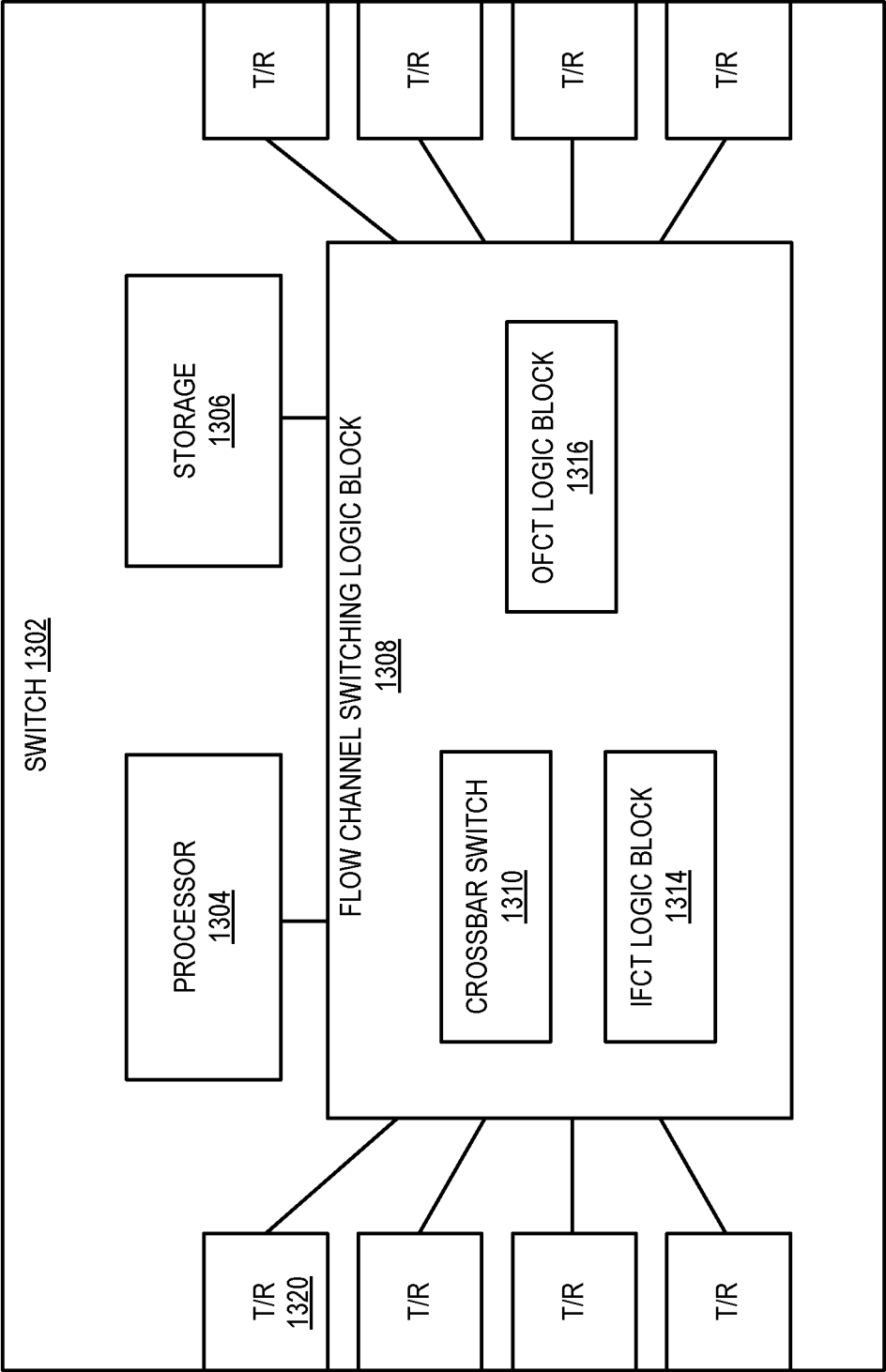


FIG. 13

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/851(2013.01)i, H04L 12/721(2013.01)i, H04L 12/743(2013.01)i, H04L 12/863(2013.01)i, H04L 12/931(2013.01)i, H04L 12/935(2013.01)i

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)

H04L 12/851; H04L 12/46; H04L 12/721; H04L 12/741; H04L 12/801; H04L 12/803; H04L 12/835; H04L 12/863; H04L 12/931; H04L 12/935; H04L 12/743

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

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

eKOMPASS(KIPO internal) & keywords: packet flow, flow identifier, free, unacknowledge, state information

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2018-0278540 A1 (AMAZON TECHNOLOGIES, INC.) 27 September 2018 paragraphs [0085], [0108], [0148], [0161]-[0163], [0179], [0209]; and figures 8-9B	1-20
Y	US 2018-0131602 A1 (ARGELA YAZILIM VE BILISIM TEKNOLOJILERI SAN. VE TIC. A.S.) 10 May 2018 paragraphs [0042]-[0043], [0052]; and figure 1	1-20
A	US 2014-0211804 A1 (SRIHARI MAKIKENI) 31 July 2014 paragraphs [0010]-[0033]; and figures 1A-5	1-20
A	US 2019-0058663 A1 (FUTUREWEI TECHNOLOGIES, INC.) 21 February 2019 paragraphs [0028]-[0066]; and figures 1-9	1-20
A	US 2014-0119367 A1 (FUTUREWEI TECHNOLOGIES, INC.) 01 May 2014 paragraphs [0029]-[0050]; and figures 1-14	1-20



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

"D" document cited by the applicant in the international application

"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

09 July 2020 (09.07.2020)

Date of mailing of the international search report

09 July 2020 (09.07.2020)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea



Facsimile No. +82-42-481-8578

Authorized officer

YANG JEONG ROK

Telephone No. +82-42-481-5709



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2020/024259

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2018-0278540 A1	27/09/2018	AU 2016-382952 A1 AU 2016-382952 B2 AU 2018-250412 A1 AU 2018-250412 B2 BR 112018013438 A2 CA 3010186 A1 CL 2018001771 A1 CN 108476209 A CN 108476209 B EP 3398315 A1 JP 2019-092217 A JP 2019-092218 A JP 2019-504557 A JP 6490310 B2 JP 6564960 B2 JP 6569020 B2 KR 10-1941416 B1 KR 10-2018-0089548 A KR 10-2019-0009419 A KR 10-2019-0108188 A KR 10-2023122 B1 SG 11201805409 A US 10148570 B2 US 2017-0187496 A1 US 2017-0187621 A1 US 2017-0187846 A1 US 2018-0278539 A1 US 2019-0173789 A1 US 9985903 B2 US 9985904 B2 WO 2017-117259 A1	12/07/2018 02/08/2018 08/11/2018 15/08/2019 24/04/2019 06/07/2017 12/10/2018 31/08/2018 05/11/2019 07/11/2018 13/06/2019 13/06/2019 14/02/2019 27/03/2019 21/08/2019 28/08/2019 12/04/2019 08/08/2018 28/01/2019 23/09/2019 20/09/2019 30/07/2018 04/12/2018 29/06/2017 29/06/2017 29/06/2017 27/09/2018 06/06/2019 29/05/2018 29/05/2018 06/07/2017
US 2018-0131602 A1	10/05/2018	US 10425327 B2	24/09/2019
US 2014-0211804 A1	31/07/2014	AT 388574 T CN 101044737 A CN 101044737 B CN 102427446 A CN 102427446 B DE 602005005219 D1 DE 602005005219 T2 EP 1813084 A1 EP 1813084 B1 TW 200642390 A TW 201208324 A TW I354473 B TW I411279 B US 2006-0104303 A1 US 2010-0020819 A1	15/03/2008 26/09/2007 16/11/2011 25/04/2012 14/01/2015 17/04/2008 12/03/2009 01/08/2007 05/03/2008 01/12/2006 16/02/2012 11/12/2011 01/10/2013 18/05/2006 28/01/2010

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2020/024259

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
		US 2011-0090920 A1	21/04/2011
		US 2017-0048142 A1	16/02/2017
		US 2018-0198709 A1	12/07/2018
		US 7620071 B2	17/11/2009
		US 8036246 B2	11/10/2011
		US 8718096 B2	06/05/2014
		US 9485178 B2	01/11/2016
		WO 2006-055494 A1	26/05/2006
US 2019-0058663 A1	21/02/2019	CN 110945845 A	31/03/2020
		WO 2019-034099 A1	21/02/2019
US 2014-0119367 A1	01/05/2014	US 9215093 B2	15/12/2015