Lightfleet®

Switched Network Latency Problems Solved

A Lightfleet Whitepaper by the Lightfleet Technical Staff

Overview

The biggest limiter to network performance is the control plane – the array of processors and interconnect that control the switches that steer the data from source to destination in the typical multi-server environment. Commercial network IT companies have attacked this source of overhead with increasingly sophisticated hardware or software. However, if the control plane could be eliminated altogether, a tough problem that has impacted network throughput in real-world applications can be significantly reduced or eliminated as well. This whitepaper discusses the issues with the typical control plane architecture and presents an innovative solution embodied in the new Data Distribution System (DDS) from Lightfleet. This new technology enables data packets to flow freely from source to destination nodes without requiring switching of the data's path, thereby eliminating switching delays that contribute to latencies and reduce throughput.

The typical computer network employs a centralized control structure to track the state of the individual switches, receive path requests, and operate specific openings and closings of individual switch elements. For top-of-rack installations where a single switch manages the attached host computers, this control may reside in the switch module itself. For larger systems, beyond the capacity of a single switching module and interconnecting more than a few dozen hosts, a controller integrated into the switching module is no longer possible, and a separate central controller is typically used, bringing with it an array of issues.

The burden of centralized control

A central control that communicates with each switching node independently of the network data paths greatly increases the number of required high-speed connections, depending on network topology. The controller unit itself is typically a high performance multi-processor with a communications capability of handling path requests from each switch in the network. Redundant controllers are often used to ensure reliability.

Even though the controller, also called the control plane to distinguish its function from the data plane (the network moving data between endpoints), makes use of high-speed communication links to the network, it introduces significant delays in network traffic. A message entering a switch must request a path from the source endpoint, often called the producer, to the destination endpoint or consumer. This request must travel from the switch to the controller, where a search is made for a sequence of switch closings that ensure proper delivery of the message. These switch closings or path segments are then communicated from the controller back to the various switches involved. At best, assuming that the search algorithms are much faster than the request and response transit times, a message in a typical three-layer network incurs a 67% overhead in the limit of zero traffic due to the request. In a large network, however, the time needed to identify and verify a complicated path of

switch closings can add significant latency to a message and, in the case of high traffic density, the likelihood that the request is denied approaches unity. Therefore, in high traffic situations, it's the controller that significantly increases the expected latency of a message.

Congestion compounds latency

A key issue in maintaining network efficiency at a high level in the presence of network congestion is the prevention of head-of-line (HOL) blocking, a situation where a blocked message due to a path being momentarily unavailable can prevent other messages in the same queue from advancing. On city streets, this is analogous to someone holding up a line of traffic as they wait to make a left turn.

Even though HOL blocking is a local issue involving a particular input queue in a particular switch, its solution is often relegated to the control plane in the course of the control plane's path-management function. An in-switch management of HOL blocking is only partially effective at maintaining a high level of efficiency, but this may be the preferred choice given that the requirement for a global solution to HOL blocking would increase transfer latency and affect network scalability.

Scalability breaks down

An assumption made in this paper is that the interconnect modules have a fixed radix, that is, each member in the network of interconnect modules has the same number of ports available for making connections to other modules and endpoints. For systems with a fixed radix, scalability is a matter of topology and not architecture.

There are two basic types of centralized controllers: (1) queueless controllers and (2) controllers that maintain large queues for re-routing messages and managing HOL blocking as discussed above. Local management of queues in each switch can maintain efficiency under certain types of congestion at a comfortably high level, however the path tables and management software required by the control plane makes scalability problematic even in this simple case.

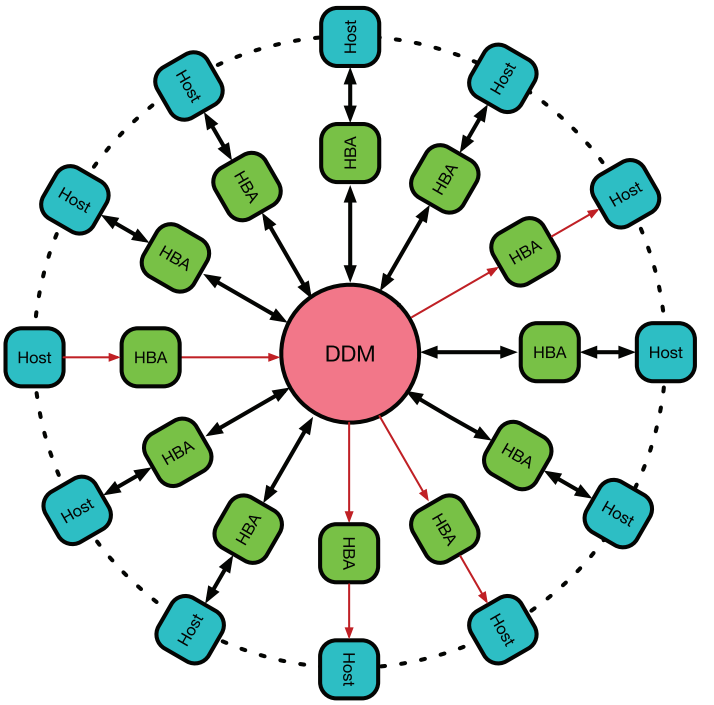
For network-wide HOL management, all semblance of scalability is lost as the extensive virtual output queues grow along with the path tables containing the spanning trees and the current switching patterns. On top of this, the software needed to control a large number of switches is complex and often difficult to maintain, adding yet another layer of expense to the switched network. For example, the algorithms needed to compute and prune large spanning trees, a subject of much study over the past three decades, are known to be unstable, introducing an element of indeterminable risk to system operation.

A preferred architecture: Data-directed distribution

Instead of a top-down control plane, so necessary for operating interconnects based on switches, the delays that are inherent to switched architectures can be avoided by incorporating the logic necessary for deciding on the most efficient delivery path of the message to its destination into the server data interfaces themselves. Extra data links needed in a switched network for global system knowledge would not be required, since each interface only requires knowledge about its nearest neighbors' location and status. Similar to how traffic circles enable vehicular throughput to be accelerated by eliminating stoplights, the distributed control philosophy embodied in Lightfleet's DDS fabric enables data packets to flow smoothly from source to destination relying solely

on destination information contained in the packets themselves.

At the core of Lightfleet's solution is the DataRotor™ Data Distribution Module (DDM) which handles the data flow, interacting with Host Bus Adapters (HBAs) associated with computers or servers connected to the fabric (see illustration -- The dashed circle indicates a plurality of inputs, connections and hosts, while the red arrows illustrate a particular multicast transmission originating in the far-left host).



Circumventing scalability risks

Since each Lightfleet interconnect module contains its own control logic, scalability is not an issue as in the switched network's control plane. A network is expandable by merely adding new modules, provided each module recognizes its position in the network. In the case of a switched network, where modules also require location information, the control plane must be replaced with a new, upgraded version that accommodates the increased number of communications ports to and from the network along with additional internal queues to ensure proper HOL control and handle any potential re-routing. Software upgrades are generally required to handle the geometrically increasing number of routing trees. Avoiding these drawbacks, unlimited scalability is a fundamental feature of a network built using Lightfleet technology.

Adaptive routing efficiency and network resilience

In a network with distributed control, a message is forwarded from one network node to the next using only local knowledge of the network topology and addressing information contained in the message itself. As automobile traffic can make its own way around a traffic circle with reference only to destination information posted at the exits, a message can enter the network and effectively find its own way to its destination(s) along optimal paths with minimal distance between the producer and consumer(s). The path is dynamically chosen based only on local information about the location of the node where the message currently resides and the state of that node's nearest neighbors.

The status of each network node or channel is passed to nearest neighbors, making the entire network fail-safe so that messages may self-route around failed nodes and channels. This form of network resiliency allows a network with failed components to keep functioning at a reduced capacity, and, depending on architectural particulars, modules and other components can be replaced without interrupting the operation of

the remainder of the network.

Handling message bursts

Simple background traffic without multicast or addressing a particular receiver with more than a single sender can cause severe stress if some of the senders are transmitting at a high rate. The input queues in a switched system can fill faster than new paths can be located by the control plane, causing throughput and hence efficiency to drop. By contrast, in the Lightfleet system, the internal queues will start to fill, but since their function is to evenly distribute their queued messages to the next available output queue, network output will not suffer and the efficiency will remain close to 100%.

High Performance Applications Benefit

For example, even at a 100% load factor (i.e., 100% of the hosts are transmitting) at maximum rate (i.e., no delay between sequential messages from a given host), the efficiency of the Lightfleet DDS is effectively 100%. For the same conditions, the efficiency of the switched network in simulation remains well below 100%, and, for a switched network to reach high efficiency for a 100% load factor, the mean host transmission rate must be reduced below the maximum.

Multicast gridlock

The concept of a multicast is aptly described in a Wikipedia article as "… the delivery of a message or information to a group of destination computers simultaneously in a single transmission from the source. Copies are automatically created in other network elements, such as routers, but only when the topology of the network requires it." The control plane in switched networks searches pre-allocated multicast trees, attempting to reserve path segments. Once all path segments (describing branches in the tree) are reserved, the control plane sends commands to the various switches so that paths along the tree, which is rooted at the source, may start receiving the multicast message from the source and distributing it along various precomputed branches for delivery to each specified endpoint. This method has the advantage of minimizing the spread in arrival times so that all recipients of a multicast transmission receive the message at approximately the same time.

While a multicast message is flowing along the various branches of a multicast tree, the control plane prevents other messages from using those same branches; that is, a multicast blocks any other traffic that would otherwise flow through any of the branches in an active multicast tree. If the tree is large enough or multiple multicast requests are made about the same time, any network traffic requiring any of the same path segments that are in use must wait for the completion of the multicast transmission. Thus multicast in a switched network is the source of spreading congestion, ultimately leading to a cessation of virtually all network traffic, driving the network output and efficiency towards zero.

There are two reasons why a multicast request in a control plane must remain in the input queue until all branches become free. The first reason is that the consumers of the multicast transmission typically need to receive the message with minimal temporal variance. The second reason is that the likelihood of a reserved path segment needing to be swapped with another to ensure timely release of the multicast message is

quite high in topologies where multiple paths between source and destination are possible. In practice, rearranging paths is computationally difficult and involves either storing messages in the control plane or canceling and resending messages along the rearranged segments. Both of these choices are difficult to carry out and result in additional delays for the messages being re-routed as well as for the multicast messages themselves. Switch vendors typically ignore the adverse effects of multicasting when they quote maximum latency times.

Multicast floods swamp efficiency

In a Lightfleet data fabric, where the control functions are distributed throughout the network rather than being centralized, a multicast message is released from the root or source as soon as it is generated in the host computer. The message sent into the multicast tree follows predetermined branchings that are managed by a software in one of the host computers. There is no need to wait for the availability of every branch since a multicast message has priority over regular traffic and simply flows through or around other messages underway (as emergency vehicles do in a multi-lane traffic circle). Thus multicast traffic is additive to any existing network traffic; it does not need to wait for other traffic nor does it block traffic from other sources. Multicasting in a Lightfleet fabric does not cause spreading congestion and the network efficiency remains unaffected. The variance or spread in arrival times of a multicast message is also minimal since multicast messages have priority over other messages.

Multicast messages can also generate flash floods if there are multiple simultaneous multicast requests. In a switched network such floods often cause the efficiency to decrease rapidly towards zero if the requests are repeated frequently enough. A network with large background traffic is particularly susceptible to multicast flooding. The efficiency of a switched network can drop precipitously during repeated multicast requests, while that of the Lightfleet DDS fabric remains high even during the stress period.

Big data — an example

Updates to a distributed database are typically handled by multicast requests. Likewise queries to such a database require a multicast operation unless the database is highly organized and ideally indexed. An example of such a distributed database is to be found in installations handling “big data,” a term that refers to very large and complex data sets that are difficult or impossible to manage using the traditional database tools. Such issues involving big data are easily managed in a Lightfleet fabric since each DataRotor DDM is particularly efficient at performing multicasting tasks. Suppose 100 million (or more) database entries identified by headings such as license-plate numbers, or social-security numbers, or telephone numbers are distributed among a large number of host computers (perhaps thousands of millions). Each host computer may have one or more connections to an external network for distributing updated information about a particular item or group of items and handling queries about the items stored.

Summary

An incoming request from outside the system is received by one of the hosts where it is directed to a particular item by hashing the item into a 100 million-entry look-up table (table size is a few gigabits, small by today’s standards). The output of this process is a multicast message containing the essentials of the request for the particular item. The message then enters the network, and quickly arrives at the destination computer or computers responsible for maintaining the particular item. The

This white paper discussed the factors that can limit the performance and reliability of switched network topologies, making them suboptimal for mission-critical applications such as securities trading and processing of big data. A solution to these problems was presented, Lightfleet's data flow fabric technology, which offers better all-around performance and a solution to problems causing network congestion and increased latency.

Interested in evaluating Lightfleet technology?
Lightfleet will provide all the details to qualified organizations. Please call Ivy Yap at 360.816.2840 or email iyap@lightfleet.com to order or discuss specific requirements.

For more information, visit
www.lighfleet.com

Lightfleet Corporation
4800 NW Camas Meadows Dr.
Camas, WA 98607-7671
USA