Lightfleet®

Introduction to Lightfleet's Multiflo™ Data Distribution System

Overview

Lightfleet's Multiflo™ Data Distribution System (DDS) introduces a new approach to interconnecting multicomputer systems that is based on the new concept of Self-Directed Data Flow. The DDS accomplishes the same basic functions as a traditional switched fabric, but does so more efficiently and with lower latency. Unlike those fabrics, the native mode for data movement in the Multiflo architecture is multicast. Also, those fabrics are top-down architectures, Multiflo is a bottom-up architecture. The supervisory function that was developed by switch-makers as a solution to the need to control data movement has actually become the major impediment to reliable high-speed and high-volume data transfers. Lightfleet's fabric does not need a separate data path controller to ensure efficient routing of information. Flow-control is in effect "woven" into the fabric, enabling the Lightfleet system to avoid the problems of high latency jitter and dropped packets that legacy fabrics can suffer under extreme conditions, including those found in cloud-based applications such as multimedia conferencing and video playback or online gaming, or in securities trading. Applications such as airline scheduling, scientific modeling, and distributed interactive simulations can also benefit.

Components

The Lightfleet Multiflo DDS consists of the company's innovative Data Distribution Module (DDM) serving multiple Lightfleet Host Bus Adapters (HBAs) schematically illustrated in Figure 1. (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.) Each HBA resides in or is closely coupled to an endpoint host computer or server, and is connected to the DDM by means of optical fibers which carry data packet streams.

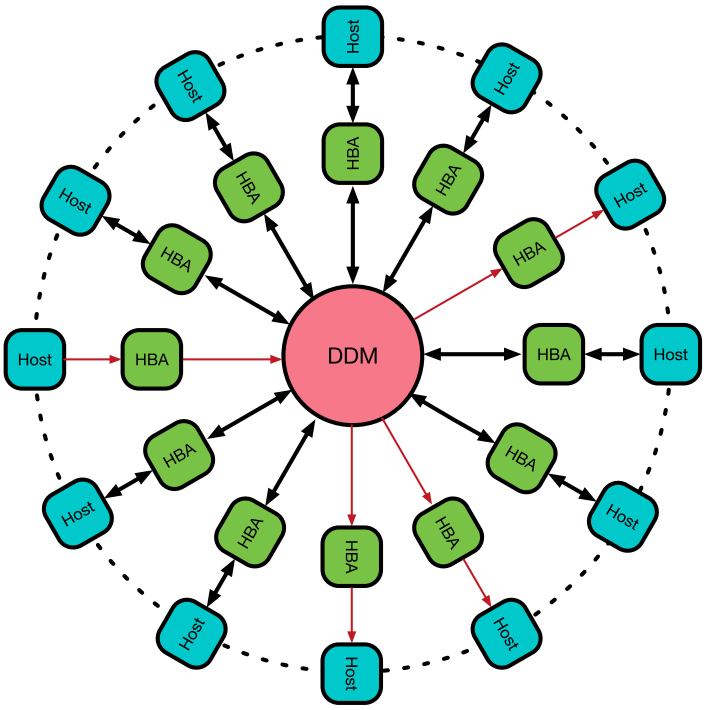

Figure 1. DDS Components. Illustration of a DDS with the DDM (pink) in the center connected to multiple HBAs (green) which are connected to their Hosts (blue).

Packet Protocol

Lightfleet's internal packet protocol has lower overhead than either Ethernet or InfiniBand, resulting in a more efficient means of data transfer, while also supporting the messaging function of Ethernet systems. A Multiflo packet consists of data frames representing a payload with a digital wrapper on each end. The first bytes of each packet contain the destination (or subscriber information), which travels along with and introduces each packet to the various components of the DDS. The last bytes of a packet contain cyclic redundancy check (CRC) information to verify the integrity of the payload. All information pertaining to the packet is therefore contained in the wrapping, and all packet control is provided by the interaction of the wrapper with the elements of the DDS. Each packet "finds its own way" through the DDS to the desired destination(s) in an adaptive manner without any outside or top-down supervision being needed to manage the data flow.

The self-directing properties embedded in a packet allow a Multiflo DDS fabric to operate without the spanning-tree machinery required by switched systems; such spanning trees both enable and frustrate data traffic in traditional switched fabrics. Without this complexity, the DDS fabric is not prone to the congestion and dropping of packets that occurs with switched systems when traffic gets "bursty" or when many sources overwhelm systems by trying to send to many recipients simultaneously.

Self-directed Data Flow

When received by the DDM, the destination field encoded in each packet is interpreted as an index into a "subscription" table located in the DDM and maintained by software residing in the source and destination computers/servers. A table entry for a specified group of data recipients is a bit map of the exit points leading to all members of the indicated group. The subscription table entry furnishes all the necessary information to the DDM's internal routing mechanism, allowing the packet to flow through to the exit port(s) corresponding to the group member(s) specified by the bit map.

The contents of an offset field, included in the initial bytes of a packet, are interpreted in the receiving HBA relative to a base address that is maintained by the receiving host. The offset plus the base address are used by the receiving HBA(s)' direct-memory-access (DMA) hardware to write the data portion of the packet directly into the correct location in the receiving host's memory -- without any copying step or operating system intervention. The distinguishing feature of the offset field used in DDS transfers is that it accomplishes its purpose without modifying the data portion of the packet (since it is a part of the packet header). The data packet location information has no interaction with the protocol stack on either end of the transmission, being used immediately and then discarded in the HBA.

Built-in Multicast

From the above description and the depiction of the potential data paths in Figure 2, it is seen that the traffic-flow mechanism built into the DDM at the architectural level is that of multicast. Unicast is merely a case where the data distribution "group" consists of a single endpoint.

With the multicast capabilities inherent in the Multiflo architecture, memory coherence can easily be initiated and maintained across an entire set of hosts distributed across a network. Physical memory in a given host may contain mirrored segments belonging to

multiple groups, and coherence is maintained independently for each group supported by a given host. Such multiple, independent, group-coherent memories are often quite difficult to establish and costly to maintain in switched network environments.

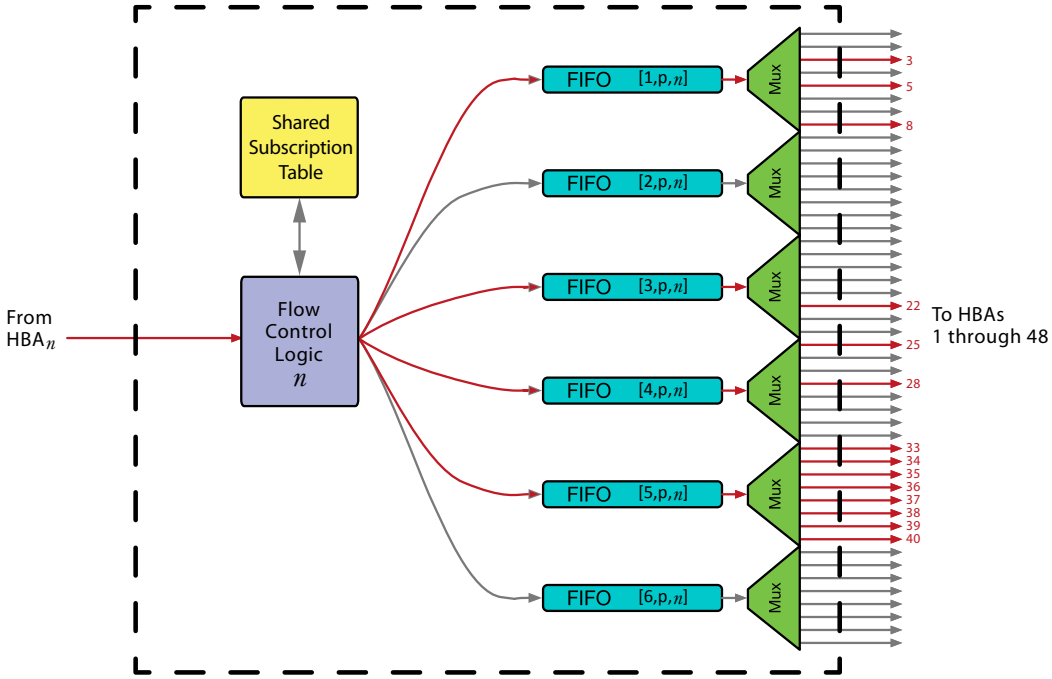

Figure 2. DDM Internal Architecture
Inside the DDM's data-flow architecture. Shown are the possible data paths from one HBA input to up to 48 HBAs connected to the DDM's output. The lines in red show the transmission path to a particular group of outputs whose members consist of hosts 3, 5, 8, 22, 28 and 33 through 40.

Distributed Flow Control

The control of data flow within and to the DDS is distributed throughout the fabric and is accomplished by issuing high-priority, single-frame (64 bit) control messages when an input queue in either an HBA or the DDM is nearing capacity. A flow-control frame sends a simple "off" or "stop-sending" command to the source feeding the complaining input queue, halting the transmission until such time as the queue is able to accept more data. In sharp contrast to a switch with centralized traffic control, DDS flow control is more responsive to internal conditions both in halting traffic to a module when conditions require it and in allowing traffic to resume. This accounts for why any congestion dissipates much faster through the DDS than with switched networks.

Handshaking in the form of acknowledgements (and negative acknowledgments) also makes use of these high-priority control frames and provides a point-to-point guarantee of correct message arrival, ensuring that there are no lost packets due to transmission errors.

In summary, the distributed flow control based on the fast priority messages enables the DDS to keep the outputs to the HBAs maximally busy and contributes to enhanced system efficiency.

Extending the Fabric

The structure of a Multiflo DDS fabric has the advantage of being modular, with each node in the fabric being identical to all others. To extend a DDS from a single DDM to a fabric or network of multiple DDMs is accomplished by computing the group subscription tables for the set of DDMs according to their topological position in the fabric, to allow any message to reach its destination by the most efficient route.

High Performance Applications Benefit

As described above, Lightfleet's Multiflo fabric provides high-efficiency and low-latency solutions to a variety of challenging data distribution problems. The architecture of the system ensures a lower-latency traffic flow and higher resource utilization compared to that of an Ethernet or InfiniBand system. Applications that benefit include high-frequency trading and high-performance computing where low latency and high efficiency enhance performance and resource utilization. Cloud computing benefits from the coherent-memory capability of the DDS, making maintenance of database applications and load balancing of virtual operating systems faster and more secure. Lightfleet's technology has the additional advantage of consuming less power than the alternatives and functions with high reliability (no dropped messages during multicast due to the flow control method used).

Interested in evaluating Lightfleet technology?
Lightfleet will provide all the details to qualified organizations. Please call Jay Brandon at 360.816.5700 or email jbrandon@lightfleet.com to order or discuss specific requirements.

Lightfleet®

For more information, visit www.lightfleet.com

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