OPENFABRICS ALLIANCE

14th ANNUAL WORKSHOP 2018

A NEW APPROACH TO SWITCHING NETWORK IMPLEMENTATION

Harold E. Cook
Director of Software Engineering
Lightfleet Corporation

Lightfleet

April 9, 2018

OBJECTIVES

- **Discuss efficiency and reliability issues in routable networks due to packet structures and software required to move them through the fabric**
- **Present a new approach which overcomes issues in typical software based routing and delivers new levels of performance and flexibility**

PACKET STRUCTURES

- **In general, routable network packets consist of:**
  - Control information
    - 1 or more packet headers depending on protocols
    - Verification information (checksums)
  - Data or payload

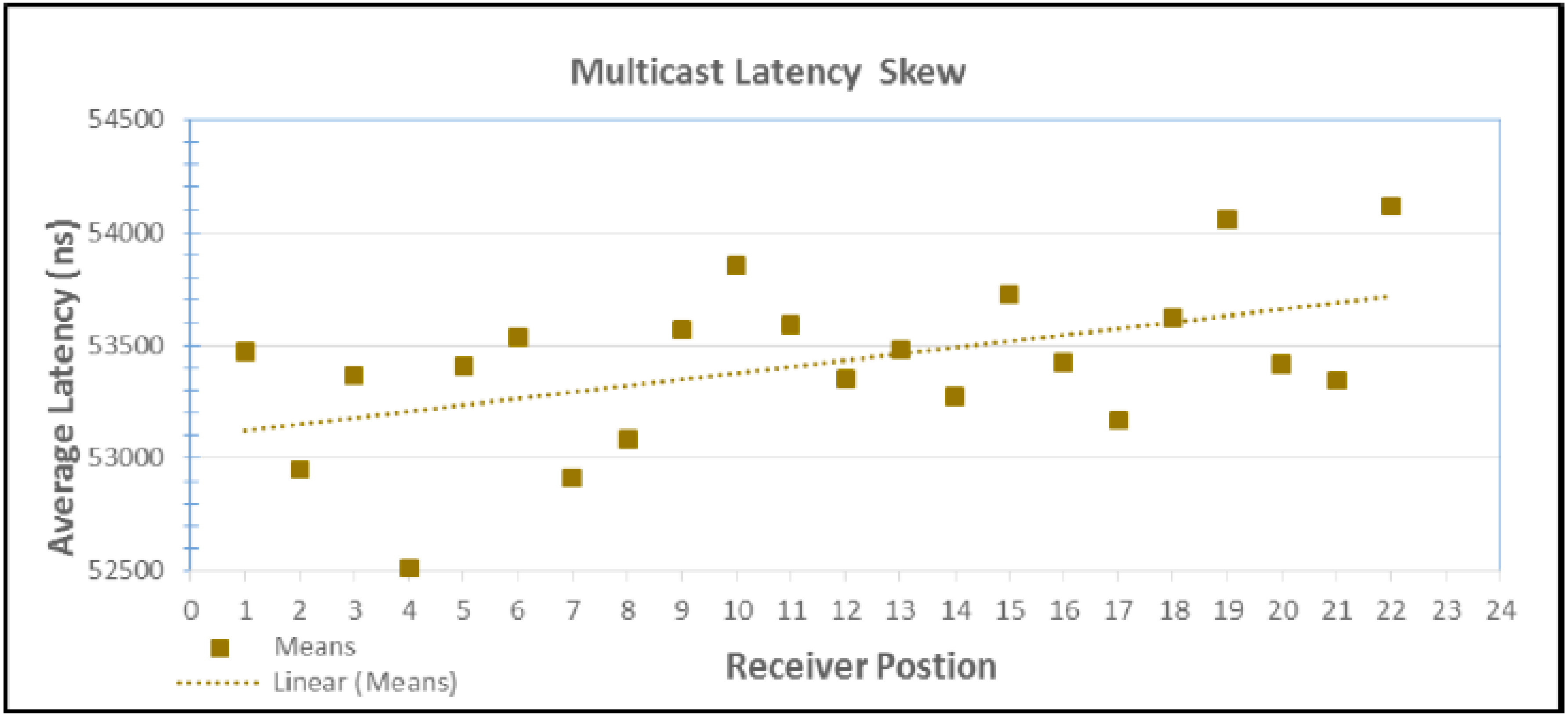
PACKET SWITCHING

- **Switches use software stacks to examine control information to determine packet routing**
  - In many cases, Software based table lookups
  - Overhead varies depending upon
    - Packet type
      - Unicast, Multicast or Broadcast
    - Pass-through mode of the switch
      - Store-and-forward or Cut-through

PACKET SWITCHING SPECIAL CASE: MULTICAST

- Multicast packets are especially onerous
- Generally need to be replicated on a subset of the available ports – serial retransmission to each port
- Skew and jitter in transit times from first port to last port
- Creates opportunity for congestion in the network that will result in dropped packets in switches under load

MULTICAST SKEW AND JITTER



CONGESTION AND PACKET LOSS

- **Congestion and packet loss is reality in software oriented switches**
- **Lost/dropped packets must be detected in protocol software stacks**
  - Recovery incurs additional overhead

NETWORK SECURITY

- **Software Stack based switches are vulnerable to cyber attacks, including:**
  - Denial of Service
  - Malicious code attacks targeted at the processors in switches, e.g., Spectre and Meltdown
  - Spoofed protocol packets or “man in the middle”
  - Others…

ROUTABLE NETWORK SUMMARY

- **Packets carry everything necessary to be routed to their destination**
- **Packets examined by every switch along the way to determine where the packet is going**
  - Software table look-up latency
  - Multicast poorly handled in switch software

A NEW APPROACH

A NEW APPROACH

SWITCHLESS NETWORKING

- Use a protocol to enable hardware routing and eliminate the software overhead from the switch

SHIFT IN NETWORKING PARADIGM

- **Move from:**
  - Packets carry everything necessary for the network to “figure out” where the packet goes
    - Requires significant software overhead
      - Network switching and routing software
      - OS based network stack
- **To:**
  - The application defines its needs (i.e. groups) and the network adapts to fulfill these needs

FOR YOUR CONSIDERATION

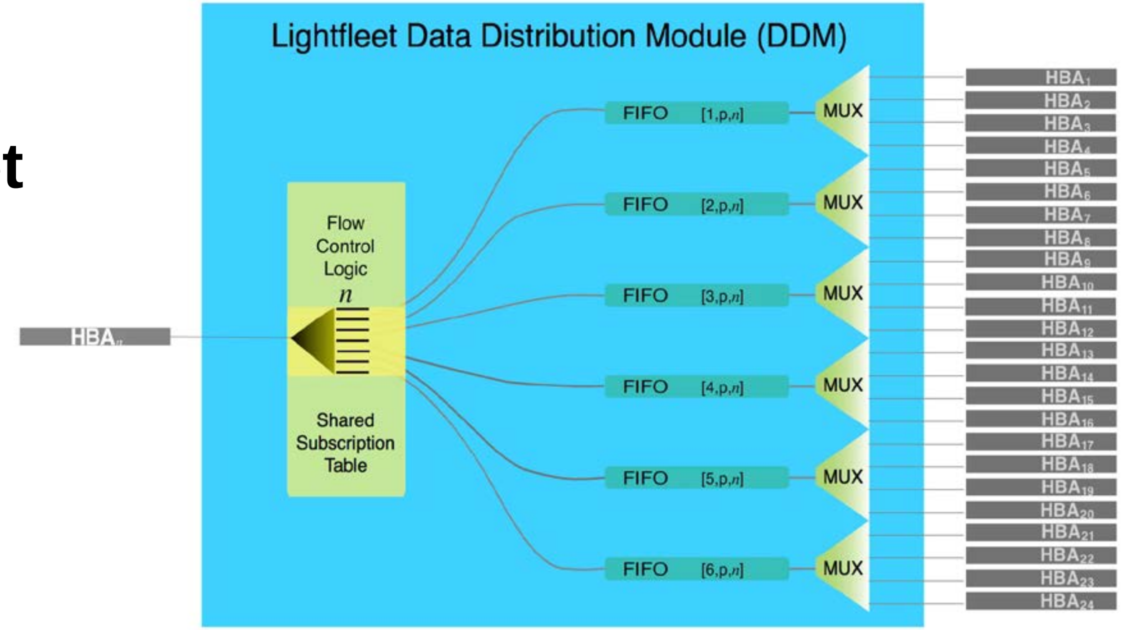
- **Many applications involve a set of hosts, working together in a bounded environment to provide services**
- **In this type of environment why tolerate:**
  - Throughput penalty of generalized network protocol(s), and
  - The software overhead that is required to support them

EXAMPLES

- **Supercomputers, HPC clusters, Big Data Analytics clusters, etc.**
- **Multi-host applications which run long periods**
  - e.g. market analysis/trading, billing, inventory systems, microservice environments, etc.
- **Storage networks**
  - Front side or back end of large storage arrays
  - NVMe fabrics

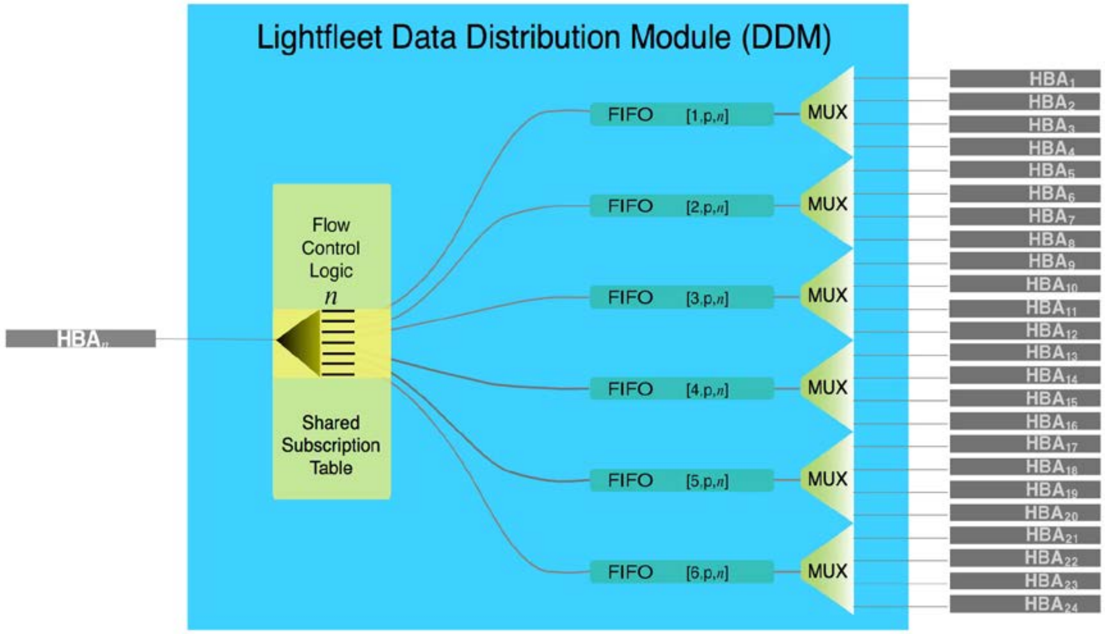
THE ALTERNATIVE

- **A connection-oriented protocol**
- **Deterministic packet routing at hardware speeds**
- **Reliable data transmission**
  - Zero Lost Packets
- **Hardware flow control**



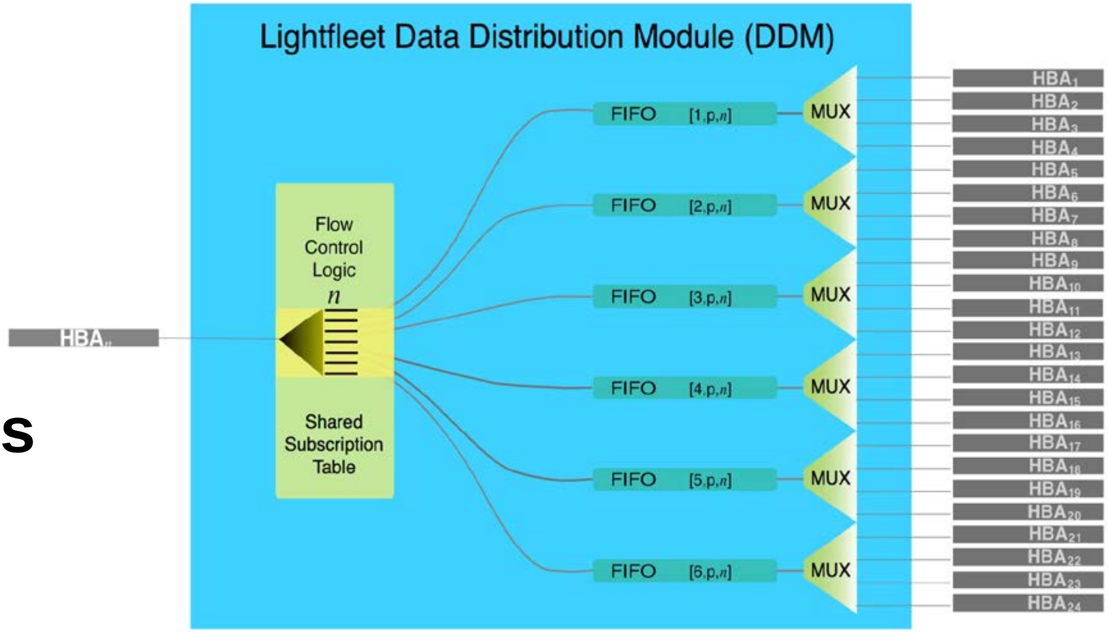
THE ALTERNATIVE

- **Everything is inherently multicast**
  - No skew in end point arrival time
  - Unicast is simplified multicast case



THE ALTERNATIVE

- **Kernel bypass architecture**
- **User space memory transfers**
- **Supports standard APIs and frameworks**



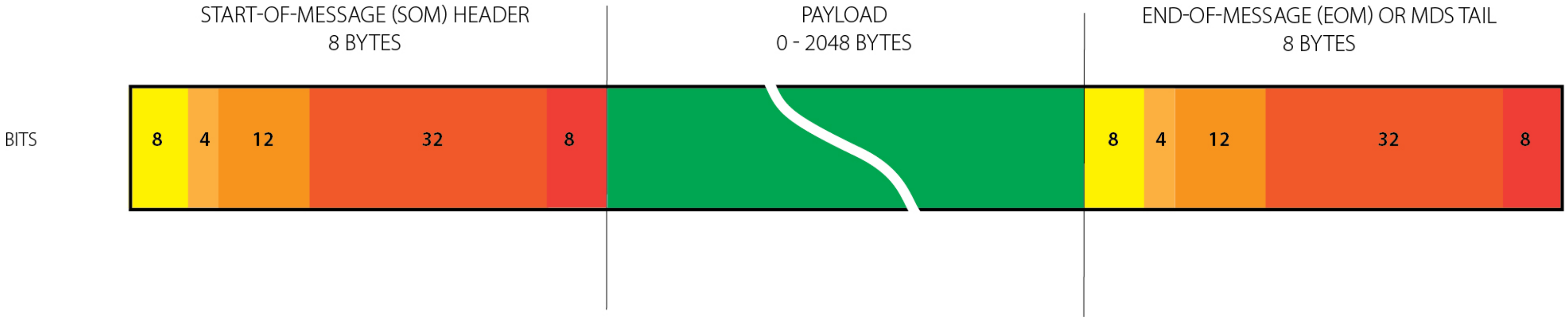
CONNECTION-ORIENTED PROTOCOL

- **Application defines “groups” of one or more servers that receive data**
  - Data written to the group is transferred to all members of the group
  - Groups are dynamic
    - Nodes enter & leave as needed

PACKET ROUTING AT HARDWARE SPEEDS

- **Packet routing determined by group identifier**
- **Lookup is done in hardware not software**
  - Latency greatly reduced!

LIGHTFLEET PACKET

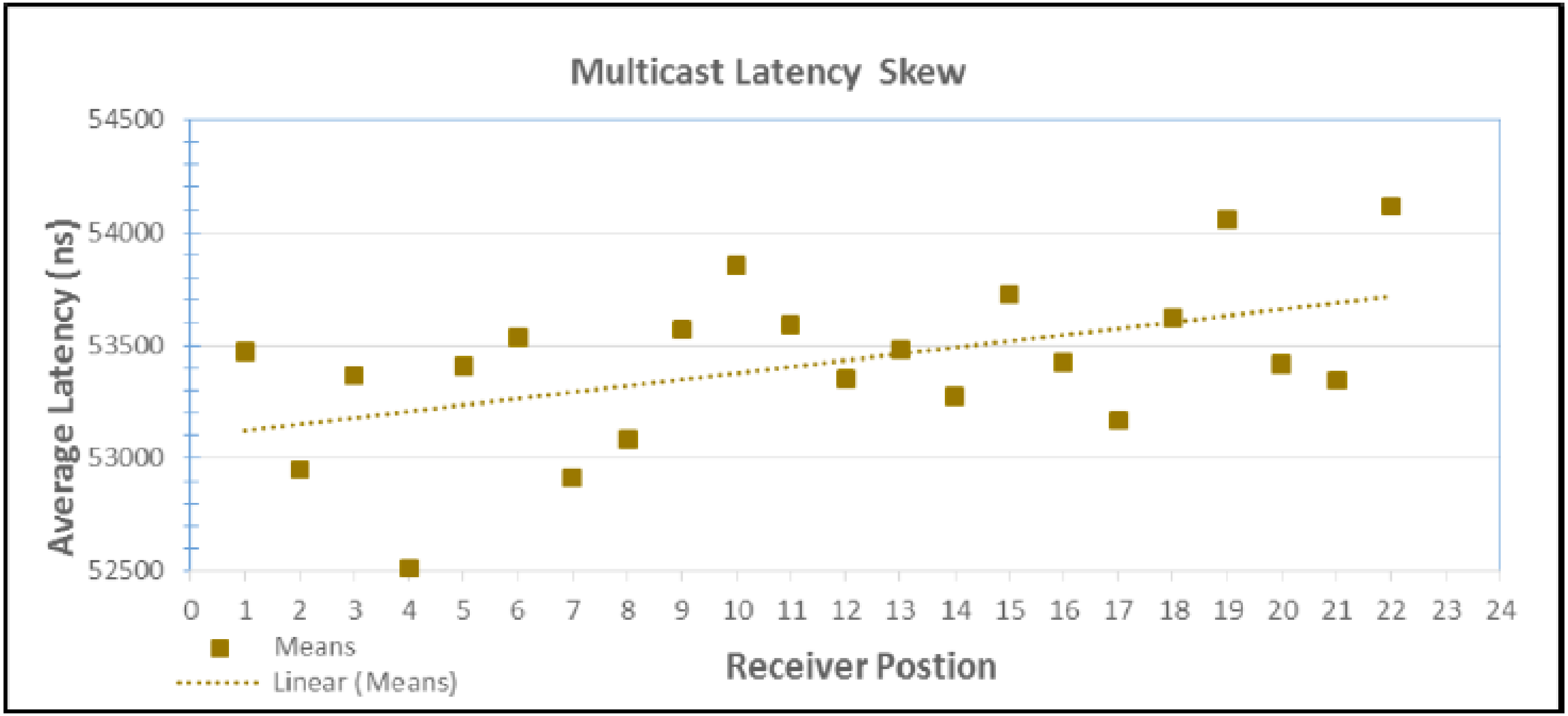


START-OF-MESSAGE (SOM) HEADER
8 BYTES
PAYLOAD
0 - 2048 BYTES
END-OF-MESSAGE (EOM) OR MDS TAIL
8 BYTES
BITS
8
4
12
32
8
8
4
12
32
8

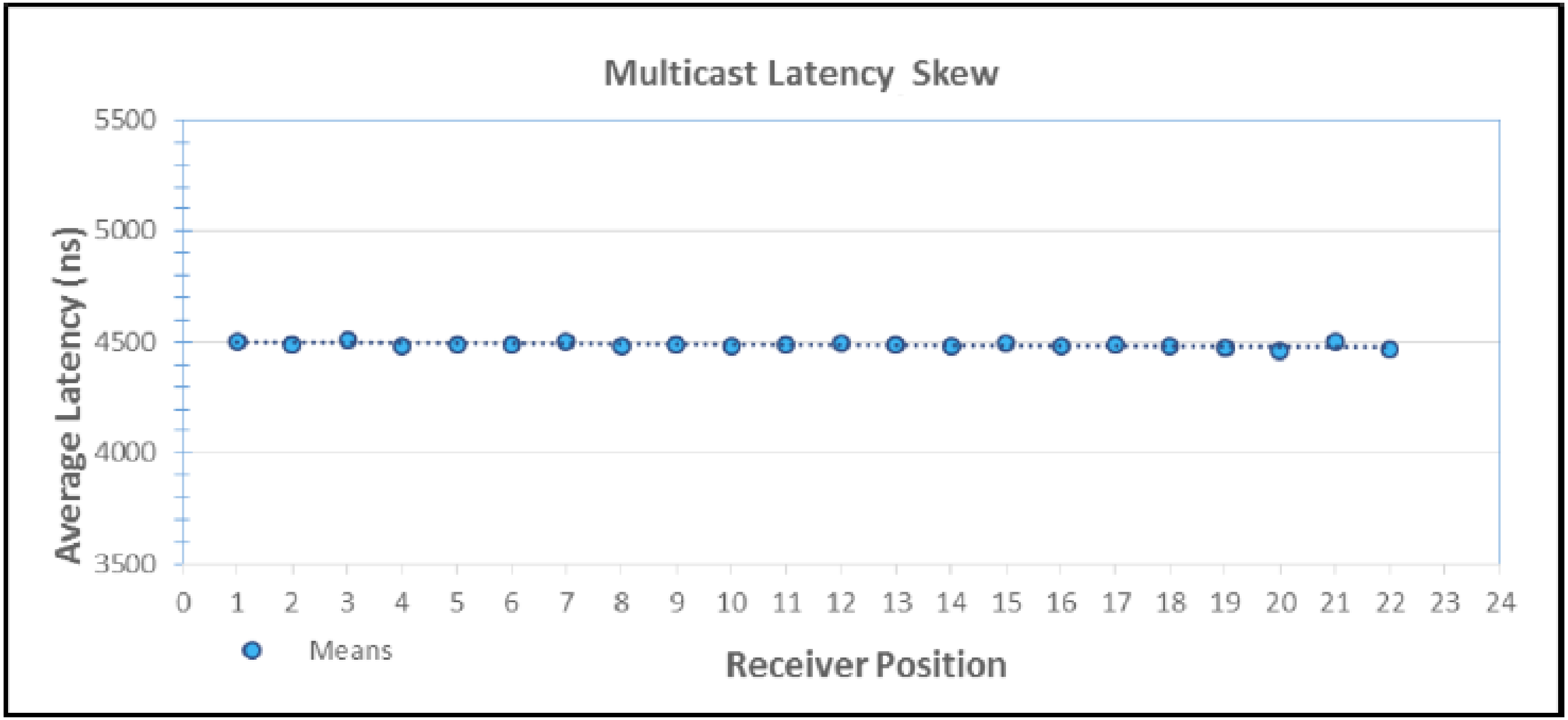
EVERYTHING IS INHERENTLY MULTICAST

- **Data moved to all exit ports simultaneously**
  - No skew and no jitter!
    - Critically important in time sensitive applications
- **True multicast was lost in the transition from bus based networks to star topologies**
  - Ongoing research & investigation into applications and benefits of Multicast.
    - Examples:
      - “High Performance Multicast”, AFRL, 2012, Birman, et al
      - “Building Smart memories and Cloud Services with Derecho”, Sagar Jha, et al, Cornell University

RECALL THIS SLIDE FROM EARLIER



SKEW-LESS AND JITTER-LESS MULTICAST



Multicast with no skew, no jitter and 12x faster*

*SOURCE: Tolly Report #216157, Nov. 2016

KERNEL BYPASS ARCHITECTURE

- **Improved latency and throughput**
  - No kernel or network stack overhead
- **User space to user space transfers**
  - Zero copy
- **Kernel drivers are used to initialize hardware and manage group subscription tables**

API AND FRAMEWORK SUPPORT

- **OFED, LibFabric, Verbs**
  - MPI and other Clustering
- **Netty**
  - Big Data Analytics & JAVA environments
- **Aeron, 29West, Informatica, Derecho**
  - Messaging based applications
- **Network emulation (i.e. Ethernet)**
  - Access standard networking interfaces

PERSISTENT MEMORY

- **Highest and best use cases for persistent memory are:**
  - Expanded front side memory bus architecture for local access such as Gen-Z, etc.
  - Lowest Latency, highest throughput reliable network for NVMeoF

NETWORK SECURITY

- **Hardware implementation means that there are no processors to attack**
- **All data is encapsulated by hardware, there is no point at which a protocol packets or headers can be spoofed**
  - No “man in the middle” opportunities
- **Denial of service not possible due to flow control implementation.**

CONCLUSION

- **By enabling hardware routing with a new protocol and eliminating software overhead, networking becomes:**
  - ✓Faster
  - ✓Simpler
  - ✓More reliable and secure

OPENFABRICS ALLIANCE

14th ANNUAL WORKSHOP 2018

THANK YOU

Harold E. Cook

hcook@Lightfleet.com

http://www.Lightfleet.com

Lightfleet®