LUMINOUS NETWORKS

Assessment of Scalable Coherent Interface (SCI)

Jason C. Fan

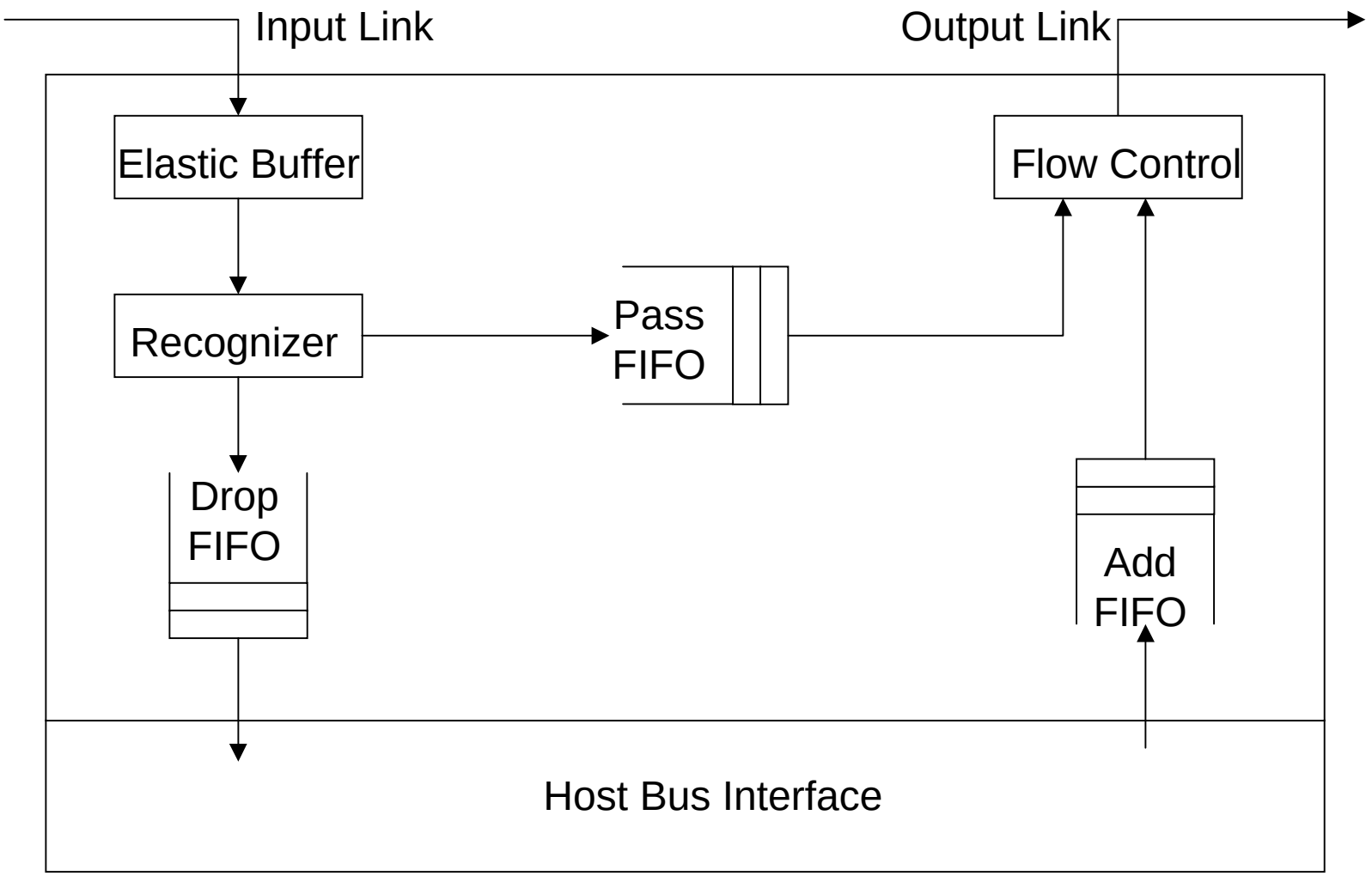
jfan@luminousnetworks.com

Purpose of SCI

LUMINOUS NETWORKS

- SCI serves as an “open” distributed bus
  - Provides low-latency supercomputer interconnect
  - Solves scalability problems of centralized buses
  - Provides bus-like behavior and services
    - Transactions: read, write, lock memory locations
    - Optional cache coherence
  - Enables connection of up to 64K nodes in topologies including single-ring or multi-ring

Block Diagram of SCI



SCI: General Information

LUMINOUS NETWORKS

- Standardized in IEEE Std 1596-1992
- Plesiochronous
  - 16-bit “Idles” between packets provide elasticity control
  - Separate clock and SOP/EOP flag signals
- Packetized
  - Variable-length packets, 16 bytes header, up to 256 bytes payload
  - Header includes
    - Bandwidth allocation and flow control fields
    - 16-bit addresses
    - Transaction number
    - Time of death
  - Packet types: request send, request echo, response send, response echo (only 2 virtual circuits per node)

SCI: Fully Reliable

LUMINOUS NETWORKS

- SCI must ensure fully reliable communications
  - Uses node-to-node request/response/acknowledge protocol
  - Uses split transactions to prevent deadlocks
    - Requests and responses are independent

- This protocol is not required in RPR
  - Would result in unnecessary increased complexity
  - Each node supports only requests and responses without external interfaces - an isolated network

SCI: Extremely Low-Latency

LUMINOUS NETWORKS

- SCI must ensure extremely low latency (a few microseconds or less)
  - Interconnect length usually limited to 200 meters for fiber, 15 meters for copper
  - Pass traffic always prioritized over add traffic
  - Node-to-node flow control and packet sequencing ensure that responses occur in phased order (overloaded server case)
    - Phase A responses occur before Phase B requests are accepted
    - This approach used as a transport layer backoff mechanism
- This is a different application from RPR
  - RPR will be over significantly larger distances
  - No need for phased requests/responses

SCI: Bandwidth Allocation

LUMINOUS NETWORKS

- SCI does not fully enforce fairness
  - Pending requests overflow at node => node enters state A, during which new requests are rejected and moved to state B
  - No delay measurements, bandwidth estimation, or random packet discard
- BW allocation messaging is via modifiable idle characters
  - Node can prevent starvation by preventing exit from state A (or B) while it still has A requests to send
  - Node can force insertion of higher priority packet
    - Other nodes forced to stop transmitting based on notification

Issues with SCI BW Allocation

- SCI uses its own physical layer encoding - non-Ethernet based and non-SONET based
  - BW allocation tied to this physical layer mechanism
- SCI handling of priority is inefficient
  - Other nodes forced to stop transmitting for a node to be able to transmit a high priority packet

Summary of SCI Assessment

- SCI is built for an application with fundamentally different requirements from RPR
  - Unnecessarily complex for RPR
  - Division of functions among networking layers is different from normal data networking approach