# A Survey of
# High-speed Networks

By Ning Zhang
University of Waterloo

# Questions to Answer

- What is the role of network subsystem in OS?
- What is new in high-speed network?
- What is the usage of high-speed networks in current I/O devices?
- What can be done to the DDBMS based on these changes?

# Network Subsystem

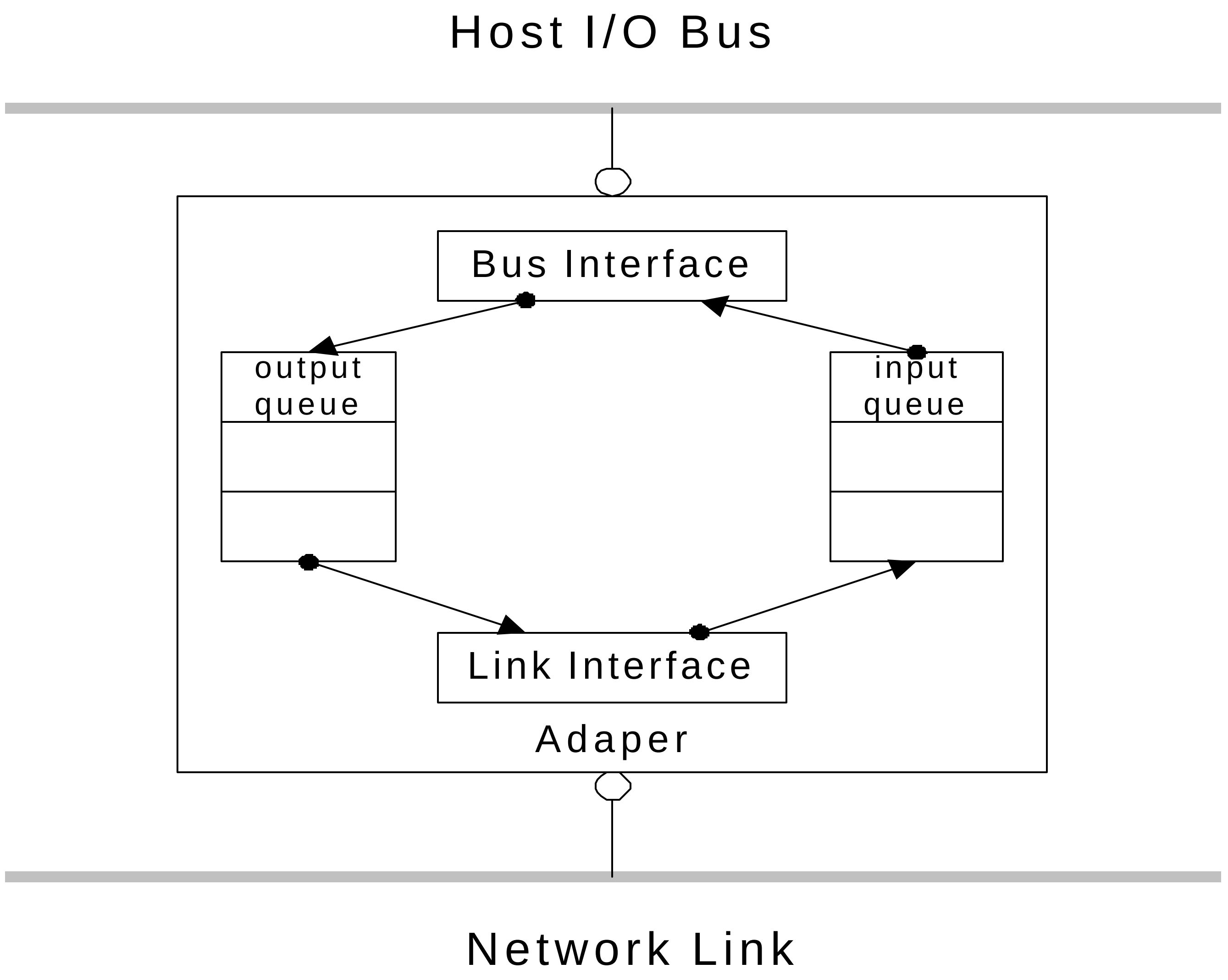
- In OS, all the software components (NIC driver, protocol implementations …) related to network communications are called **network subsystem**.
- Three components in Network subsystem: hardware architecture of the host, the host software system, and the network interface.
- The common services provided to the application determines their functionality and efficiency.
- It is crucial to understand what services provided to the upper layer applications.

# Broadband and High-speed Network

- Network performance is determined by 2 parameters: bandwidth and latency.
- Bandwidth: the number of bits can be transferred in a unit of time. Broadband means the number is large.
- Latency: the time it takes to transfer a certain size of message from one end to the other end. High-speed means the latency should be short.

# Measure of Latency

- Latency $T_l = T_p + T_t + T_o$
  - – $T_p = d/c$, represents propagation time,
  - – $T_t = s/w$, represents transmission time,
  - – $T_o$ represents protocol overhead.
- High-speed is two-fold:
  - – $T_t$ is small → broadband network.
  - – $T_o$ is small → low overhead protocols and hardware support.


Host I/O Bus
Bus Interface
output queue
input queue
Link Interface
Adaper
Network Link

# Network Interface Card

- Who do the dirty work?
  - Direct Memory Access (DMA)
  - Programmable I/O (PIO)
- Where to put the data?
  - System’s space
  - User’s space.
- How close to the CPU?
  - I/O bus
  - System bus

# Proposed Changes to NIC

- Mapping NIC memory to virtual memory
- Connecting NIC to memory bus rather than I/O bus
- Allowing caching network interface registers, out-of-order and speculative access to the registers.
- Removing side-effects from the API in OS.

# Typical Levels in Memory Hierarchy

| **Level** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| Called | Registers | Cache | Main Memory | Disk Storage | Network |
| Bandwidth (MB/s) | 4,000-32,000 | 800-5,000 | 400-2,000 | 4-32 | 1-200 |
| Backed by | Cache | Main Memory | Disk | Network | Tape |

# JOIN Network and Disk

- Integrate the two distinct devices into one:
  - Pushing up NIC to system bus → make faster devices closer to CPU
  - Externalizing disk to high-speed networks → make slower devices farer from CPU
- System/Storage Area Network (SAN): creating a method of attaching storage to network.
  - Fibre Channel (FC): ANSI X3T9.3
  - InfiniBand Architecture (IBA): HP, IBM, Intel, Microsoft, Sun…
  - Others: ESCON, SCCI, HIPPI…

# Fibre Channel Features

- Allowing many well-known existing channels and network protocols to run under the same physical interface and media
- High bandwidth (>=100MB/sec)
- Flexible topologies
- Connectivity over several kilometers.
- Support for multiple data rates, media types, and connectors.

# Fibre Channel Layers

- FC-0: signaling, media specification, receiver/transmitter specification.
- FC-1: data encoding, link maintenance.
- FC-2: frame format, sequence management, flow control, classes of service, topologies.
- FC-3: undefined set of services.
- FC-4: mappings for Upper Level Protocols (ULPs).

# Fibre Channel Services

- Class 1: dedicated connection service. Frame orders are preserved. e.g. audio/video on-demand.
- Class 2: connectionless, but guarantees notification of delivery or failure  to deliver. e.g. client/server distributed computing.
- Class 3: connectionless, unacknowledged delivery. e.g. IP/UDP packets.

# InfiniBand Architecture vs. FC

- Higher bandwidth (250M~3GB/sec) and more sophisticated architecture than FC.
- Higher scalability: thousands of nodes per subnet. (FC -- 127 nodes in arbitrated loop).
- Higher flexibility: more complex topologies
- More layers and classes of services.
- Support IP v.6 and multicast.

# IBA Layers

- Physical layer: signaling, framing, etc.
- Link layer: packet format, flow control, subnet routing, etc.
- Network layer: routing between subnets.
- Transport layer: message segmentation and reordering.
- Upper layer protocols: network management protocols, etc.

# IBA Service Types

| **Service Type** | **Connection Oriented** | **Acknowledged** | **Transport** |
|---|---|---|---|
| Reliable Connection | Yes | Yes | IBA |
| Unreliable Connection | Yes | Yes | IBA |
| Reliable Datagram | No | Yes | IBA |
| Unreliable Datagram | No | No | IBA |
| RAW datagram | No | No | Raw |

# Lightweight Network Protocols (Trapeze as an example)

- Trapeze is built on Myrinet
- Lightweight Remote Procedure Call (RPC)
  - Programmable firmware on NIC
  - Zero-copy
  - Message pipelining on DMA
- Non-blocking RPC
  - Natural extension to select() system call.
  - Set up a hook procedure then return immediately.

# Impact on DDBMS?

- How to place data?
  - – Range partitioning
  - – Round robin partitioning
  - – Hashing partitioning
- How to make relational operators more parallel?
  - – Pipelined parallelism
  - – Partitioned parallelism

# Conclusion and Future Work

- As networks getting faster and faster, protocol overhead becomes the biggest time consumer.
  - Hardware solution: Storage Area Networks (FC, IBA)
  - Software solution: lightweight protocols (Trapeze)
- DDBMS needs to revise its assumption, algorithms, and strategies.