

Performance Evaluation of Myrinet-based Network Router

Information and Communications University

2001. 1. 16

Chansu Yu, Younghee Lee, Ben Lee

SECU.T.COM



Contents

- Suez : Cluster-based Router Suez Implementation
- Implementation of Suez
- Inside the Router : Pipelined Data Move
- Limitations of Suez
- Suez Components
- Suez Technologies
 - Routing Table Lookup
 - FGFFQ Scheduling
- Current Status & Future Work



Suez : Cluster-based Router

■ Advantages

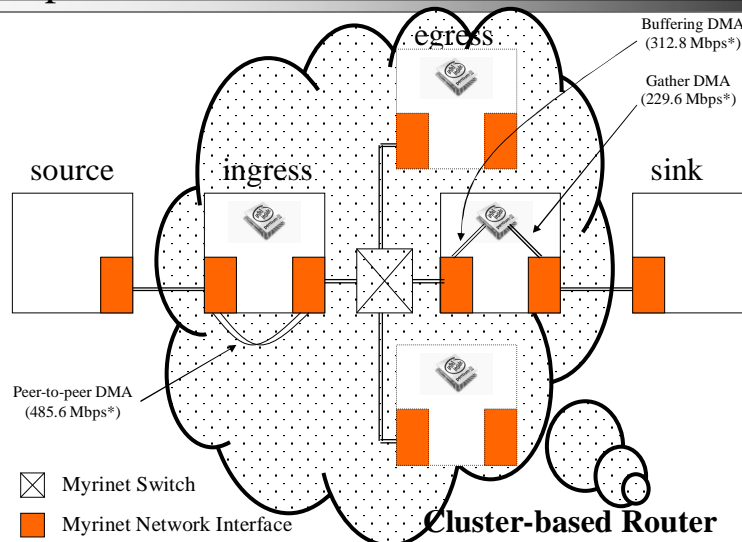
- scalable
- extensible
- cost-effective

■ Key technologies

- Cache-conscious Routing Table Lookup
 - exploit host CPU's cache
- FGFFQ (Fixed-Granularity Fluid Fair Queueing)
 - support a number of real-time connections
- Five-stage pipelined data movement
 - Peer-to-peer DMA
 - Buffering & Gather DMA's

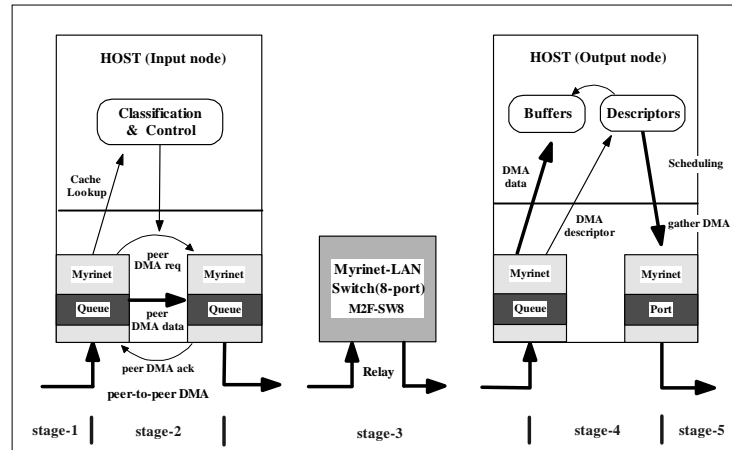


Implementation of Suez



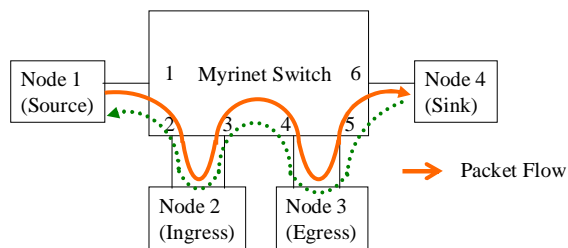
* Batch size :4096 bytes

Inside the Router : Pipelined Data Move



Limitations of Suez

- Difficult to use in practice
 - Myrinet & BIP-based external interface is assumed
 - Ethernet interface is not supported
 - Ingress and egress functions are not integrated (data move in reverse direction is not considered)

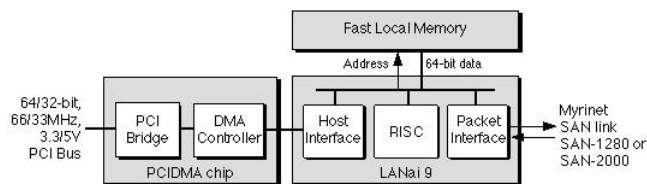


Suez Components

- Myrinet switch and NI's for ~1Gbps bandwidth
- GM : Myricom's message-passing system (messaging APIs + driver + Myrinet-interface control program)
- BIP (Basic Interface for Parallelism) for lightweight messaging software on top of Myrinet : Source & sink nodes
- Linux kernel modified : Ingress & egress nodes
 - Cache-conscious routing table lookup (ingress)
 - Peer-to-peer DMA (ingress)
 - Buffering DMA (egress)
 - Gather DMA (egress)

Suez Components : Myrinet

- Myrinet switch (inter-node connection)
 - M2F-SW8 : Myrinet-LAN Switch (8-Port)
- Myrinet network interface card (NIC)
 - M2F-PCI32C, Lanai 4.3 32-bit, 33MHz Myrinet-LAN/PCI interface, PCI short card, 1 MB SRAM



- MCP (Myrinet Control Program)
 - Firmware on the NIC
 - GM MCP, BIP MCP, Suez MCP

Suez Components : BIP (Basic Interface for Parallelism)



- Source & sink nodes
- BIP MCP + Linux Kernel + BIP Messaging
- Lightweight Messaging System
 - BIP Message
 - Send(), Receive() primitives with blocking/non-blocking
 - BIP features
 - User-level access to the communication buffer of network interface (NIC registers and BIP firmware) without any system calls
 - Avoid protected multiplexing of NIC
 - Zero-copy messaging
 - Fragmentation of messages and pipelining ==> 4~5 μ s Latency

Suez Components : GM for cluster nodes



- Message-passing system by Myricom
 - Includes
 - Messaging APIs
 - Driver
 - Myrinet-interface control program
 - Provides
 - Reliable and ordered delivery of messages
 - Protected as well as user-level access to NIC hardware
 - Latency : 8.5 μ sec
 - Extensible to allow simultaneous direct support of the GM API, IP (TCP/UDP), MPI, and other APIs



Suez Technologies: Routing Table Lookup

■ Conventional routing table lookup

- IP Routing Table Entry
 - Network Mask, Destination Address, Output Port Identifier
- Longest Prefix Matching
 - Index tree is used to avoid visiting unnecessary routing entry
- Issues
 - But, software-based routing table lookup can't be done with wire speed

■ Suez's solution

- HAC Cache (internal L1 cache)
 - HARC/IHARC
- NART Cache (external L2 cache)



Host Address Cache (HAC)

■ Network Packet Stream

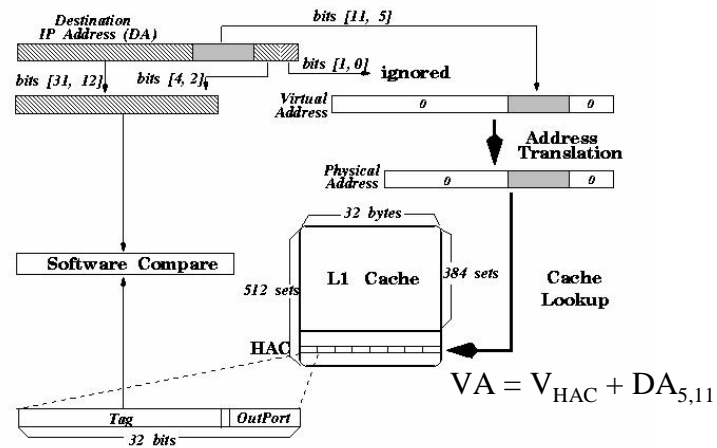
- Less spatial locality
- Temporal locality during network connection time
- Relaxed cache consistency

■ HAC

- Combined S/W & H/W approach
- Reserved L1 Cache
 - ensure HAC is always L1 cache resident
 - Linux kernel modification required

Cache Size	Block Size	Associativity	Miss Ratio
4K	32	1	57.09%
		2	53.25%
		4	50.92%
	8	1	36.51%
		2	31.29%
		4	29.00%
	1	1	12.71%
		2	8.42%
		4	6.86%
8K	32	1	43.70%
		2	40.78%
		4	38.05%
	8	1	26.35%
		2	21.72%
		4	19.33%
	1	1	7.57%
		2	4.59%
		4	3.29%
32K	32	1	18.65%
		2	16.52%
		4	15.58%
	8	1	9.59%
		2	6.66%
		4	5.49%
	1	1	2.39%
		2	1.07%
		4	0.75%

Host Address Cache (HAC)

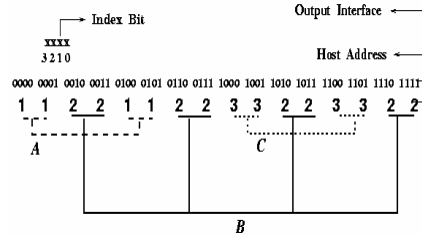


Host Address Range Cache (HARC)

- To cache host address ranges instead of individual address
- Two additional processing steps for HARC
 - Culling step
 - every address range in the IP address space is covered by *exactly one* routing table entry
 - Merging + Align step
 - adjacent address ranges that share the same output interface should be merged into larger ranges
 - *minimum_range_granularity* is calculated
 - $range_size = \log(\text{minimum_range_granularity})$

Intelligent Host Address Range Cache (IHARC)

- To merge non-contiguous address ranges into a single cacheable unit



- Index bit selection algorithm

```

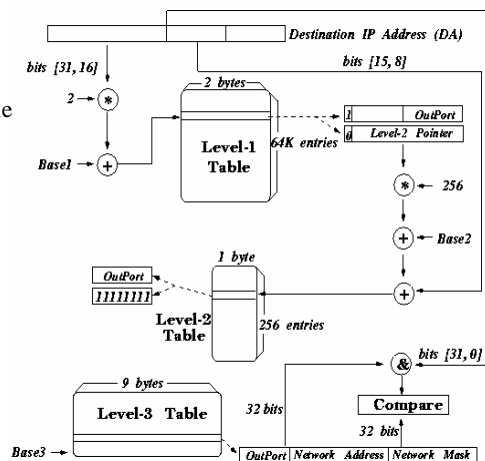
S = 0;
for (i=1; i ≤ K; i++) {
    score = ∞;
    candidate = 0;
    for (j=range.size+1; j ≤ N; j++) {
        if (!(j ∈ S)) {
            currentscore = Score(S,j);
            if (currentscore < score) {
                score = currentscore;
                candidate = j;
            }
        }
    }
    S = S ∪ {candidate};
}
    
```

Network Address Routing Table (NART)

- NART Construction

- o Three levels of tables
 - o one Level 1 Table
 - o variable Level 2 Table
 - o one Level 3 Table

- NART Lookup



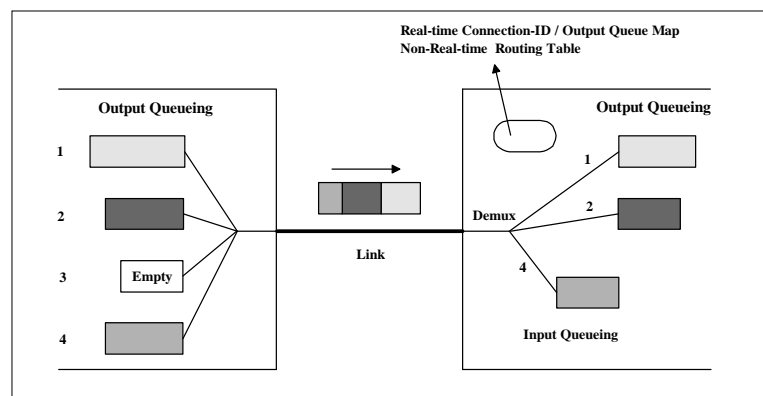


Suez Technologies: FGFFQ Scheduling

- Fixed-Granularity Fluid Fair Queueing
 - Solves problems with WFQ (weighted fair queueing)
 - No Virtual-to-Physical time Mapping
 - No Packet Sorting
 - Multiplexing / De-multiplexing mechanism on top of link layer protocol
 - Emulate fluid fair queueing model by transmitting data quantum instead of packet in round-robin fashion
 - Support Work-conserving & Non-work-conserving scheduling



Stream Mux/Demux in FGFFQ Scheduling





Future Works

- Questions
 - Fragmentation and Reassembly : where and how
 - at Myrinet switch
 - at sink
 - BIP or GM
 - FGFFQ queueing model
 - concept of fluid queueing model
 - how does it handle congested case
 - Cached routing table
 - Host's L1/L2 cache : how does it configured (Linux kernel)
 - Myrinet NIC's cache (?) utilized (?)
 - Software structure at source/sink, ingress and egress
 - BIP, GM, MCP, driver, Linux
 - Linux (which part is modified),
- Future works : Measurement of throughput at each of five pipeline stages
 - Ingress Node Receive
 - Peer-to-Peer DMA at Ingress
 - Myrinet Switching (wormhole)
 - Buffering & Gather DMAs
 - Egress Node Send