MessageWay

High Performance Intra- and Inter-Computer Communication System

IETF-BOF, Danvers, MA April-4-1995

Danny Cohen & Chuck Seitz
Myricom, Inc.
325 N. Santa Anita Ave, Arcadia, CA 91006
818-821-5555 <Cohen@Myri.com> Fx:818-821-5316
http://www.myri.com

MsgWay

1. The Problem
2. The Approach

An Observation

Modern multi-processors are moving from intra-system, inter-processor buses to internal networks ("MPP-networks").

Buses are still popular for intra-board communications.

MsgWay's Goals

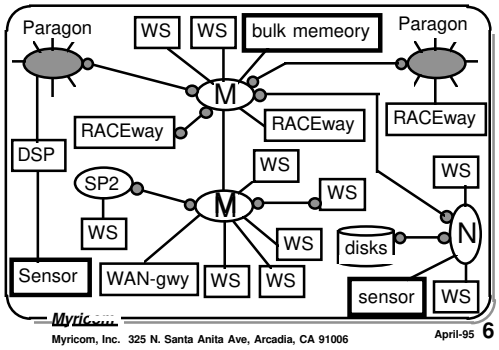
- **Interoperability ("internetting") of the heterogeneous networks of computers/MPs, using a variety of high-performance communication technologies.**
- **Clustering MPPs and distributed workstations to work together as a single MPP.**

Sounds like IP. Why not IP?

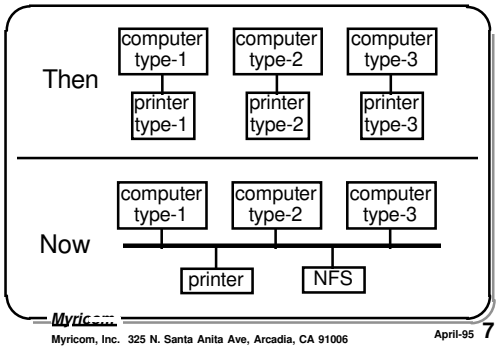
MsgWay is ...

a set of heterogeneous systems and high-performance nets, with uniform end-to-end protocols, allowing processors to locate each other and to communicate with each other at high-performance.

Example MsgWay

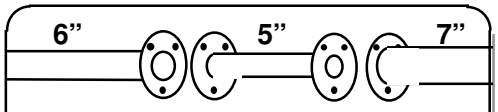


Interoperability



Myricom, Inc. 325 N. Santa Anita Ave, Arcadia, CA 91006 April-95 7

What to Standardize?



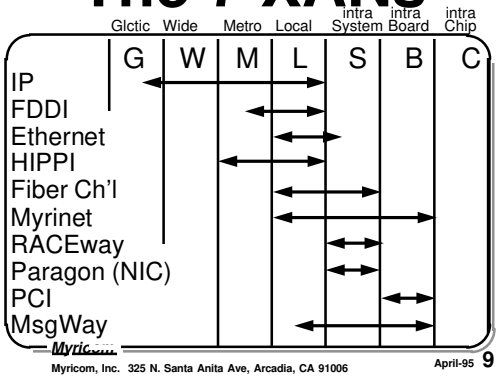
Standardize the flanges (Level-3), not the pipes (Level-2).

Level 2 brings performance.
Level 3 brings interoperability.

(Standardize the essentials, not the incidentals!)

Myricom, Inc. 325 N. Santa Anita Ave, Arcadia, CA 91006 April-95 8

The 7 XANs



Myricom, Inc. 325 N. Santa Anita Ave, Arcadia, CA 91006 April-95 9

Communication Levels

IP **WANs+MANs**

LANs ***MsgWay***

SANs+BANs

IP unifies Networks into a coherent end-to-end communication system.

MsgWay unifies “backplane-nets”.

Myricom, Inc. 325 N. Santa Anita Ave, Arcadia, CA 91006 April-95 10

WAN

- ☹ **low performance**
- ☺ **10^6meter distances**
- ☺ **heterogeneous**
- ☺ **general topology**
- ☹ **host-unit addressing**
- ☺ **packets /** ☹ **cells**
- ☺ **scalable (size, population)**

Examples: NSFnet and ESnet

Myricom, Inc. 325 N. Santa Anita Ave, Arcadia, CA 91006 April-95 11

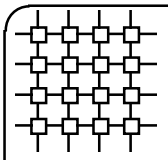
LAN

- 😐 **medium performance**
- ☺ **10^3meter distances**
- ☺ **heterogeneous**
- 😐 **restricted topology**
- ☹ **host-unit addressing**
- ☺ **packets**

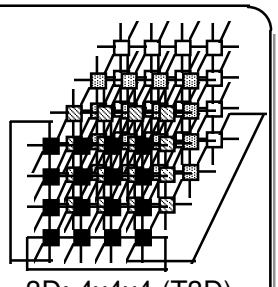
Examples: Ethernet, FDDI, ATM, Myrinet,...

Myricom, Inc. 325 N. Santa Anita Ave, Arcadia, CA 91006 April-95 12

Intra-MPP Comm



2D: 4x4 (Mosaic, RACEway, Delta, Paragon)



3D: 4x4x4 (T3D)

Intra-MPP Comm

Networks (fabric) replace buses
High performance (latency, rate, BER)
Point-to-point links
Cut-through switches
Packets
Flow Control
Short links
Homogeneous nodes
Regular topology

MPP Networks

- ☺ **high performance**
- ☹ **1meter distances**
- ☹ **homogenous**
- ☹ **fixed topology**
- ☺ **fine grain addressing**
- ☺ **packets**

Examples: Paragon, RACEway, SP2, ...

Observation

MPP networks are optimized for high rate, low latency, variable length packets, point-to-point FDX communication links, and cut-thru switches. Examples: Paragon, CM5, T3D, RACEway, and Myrinet.

Each being an island unto itself.

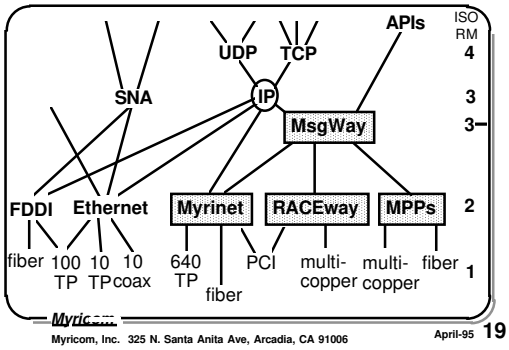
MsgWay

- ☺ **high performance**
- ☺ **10^3meter distances**
- ☺ **heterogeneous**
- ☺ **general topology**
- ☺ **node-unit addressing**
- ☺ **packets**
- ☺ **scalable**

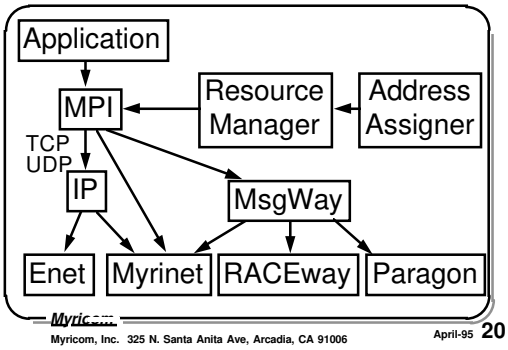
The Price

- ☹ **10^3meter distances only**
- ☹ **limited scaleability (10^3s)**
- ☹ **limited community**
- ☹ **for IP traffic: additional SW**

Where is MsgWay?



MPI's View



Unification

MsgWay could provide uniform services over SANs such as:

- √ **RACEway**
- √ **NIC (Intel's Paragon)**
- √ **Myrinet**
- **SP2**
- **Fiber Channel, ...**

Philosophy

- **The *Internet* philosophy for high-performance and fine-grain, for connecting computers+MPPs.**
- **Designed for local scale (not for a WAN).**
- **High reliance on native operation.**
- **Permissive framework (encourage deferred bindings)**

IP Deficiencies

(for high performance operation)

- ☹ **Very long addresses (32/128 bits)**
- ☹ **Addresses only, no S/R**
- ☹ **Need for routing by switches**
- ☹ **Host-addressing only (not finer)**
- ☹ **No flow control, fault recovery**
- ☹ **No error detection**
- ☹ **No support of DMA (alignment)**
- ☹ **Hierarchical de-muxing**

IP Efficiencies

(for general global operation)

- ☺ **Very long addresses (32/128 bits)**
- ☺ **Addresses only, no S/R**
- ☺ **Routing by switches, not hosts**
- ☺ **Host addressing, not finer**
- ☺ **Flow control by ends**
- ☺ **Fault recovery by ends**
- ☺ **Error handling by ends**
- ☺ **No support of host-idiosyncracies**
- ☺ **Hierarchical de-muxing**

MsgWay and IP

- IP was defined 20+ years ago, was designed for generality within the WAN environment, and has survived technology changes. MsgWay should learn from that example.
- **MsgWay intends to supplement IP for faster communication and for fine-grain addressing.**

MsgWay

2. The Approach

Assumption

Each participating SAN is assumed to have "MPP-like" communication network:

scalable high performance
- high data rate (> 1Gb/s)
- low (path-formation) latency (< 1μs)
- low error rate (BER ?< 10^{-15})

moderate range (~10m Cu, ~1Km fiber)
addressing individual processors

MsgWay Goals

"MPP-like" communication network:
scalable high performance
- high data rate (> 1Gb/s)
- low (path-formation) latency (< 1μs)
- low error rate (BER ?< 10^{-15})

moderate range (~10m Cu, ~1Km fiber)
addressing individual processors

heterogeneous computers
general, robust, scalable topology
continuous mapping+monitoring [ATOMIC]

Desired Properties /1

FDX point-to-point links
Source-Route (and/or addresses)
Consumable header OK
- any number of bytes, padding OK

Very large (Indefinite?) body
- multiple of 32/64/128/... OK

"Type" for next level muxing
- in body

Desired Properties /2

Link flow control
End-to-end error detection
Fault detection and recovery
No ever-looping packets
Automatic mapping
Island interconnection / routers
Island interconnection / bridges

Desired Properties /3

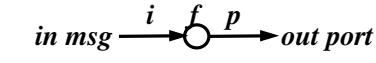
Fine grain addressing
Implementation independent
OS-environment independent
Vendor independent
Unified dictionary of Types
Interoperating routing
Unified mapping (L2 and L3)
Unified naming(s) + capabilities?

Who Routes?

IP uses Addresses:
Switches Route; Not Hosts.

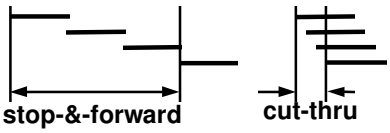
MsgWay may use Source Routes:
Hosts Route; Not Switches.

Each is right, for its environment.



Addresses: $p=f(i,S)$ SrcRt: $p=f(i)$

Format for Hi-Perf



Start with info needed for cut-thru
Format determined by each receiver (address or source-route)
"Type" and DMA info in header
Checksums in trailer

Layers Shortcuts

Lower levels should help higher levels (such as CRCs and DMAs to final destinations), even if it is against the strict interpretation of the ISORM.

MsgWay will support such assists by the hardware.

Mapping+Routing

MsgWay will be self-mapping.

It will use Level-3 routers for aggregate mapping among islands of different technologies, and use Level-2 bridges for fast direct data forwarding among them.

MsgWay Addresses

Addresses will be unique throughout the entire MsgWay.
Long addresses waste transmission time and must be hashed for TLU.
Short addresses are ideal for TLU, but require setup-computing for their unique assignment.
Recommendation:16 bits, initially.

Address Assignment

The short addresses could be assigned by the mapping process, in a distributed manner.

When interconnecting separate SANs some short-addresses conflicts may occur, to be treated (and fixed) just as faults (major problems, at low frequency).

Short Addresses

could be implemented:

either
* Statically (e.g., manual like IP-addr's)
or
* Dynamically (automatic)

and also

either
+ Hierarchically ((N)(16-N))
or
+ Packed

MsgWay Multicast

MsgWay multicast will be on a per-SAN basis.

It's up to each SAN to perform an internal multicast, by hardware or software, as it prefers.

MsgWay Priority

MsgWay priority will depend on the priority schemes (if any) of the participating SANs.

MsgWay

Summary

MsgWay does define

... how to achieve high performance interoperability among heterogen's islands, using a variety of comm'n implementations and media.

MsgWay is a seamless intra- and inter-host communication, ***but not a WAN***.

MsgWay doesn't define

Connectors,
Hardware,
Communication media,
Software (APIs, etc.),
Performance, error-rate,
Routing ,
How to flow control on links,
.....

Neither does it define

How to map SANs,
How resources are discovered,
How resources are assigned,
How addresses are assigned,
How routes are computed,
Preferred address format(s),
.....

Summary

Standardization at level 3- (slightly below IP) for interoperability, with emphasis on high performance.

MsgWay allows a variety of Level-2 and Level-1 implementations, including existing SANs (nearly) and workstations clusters.

The End

MessageWay

Mailing-List: <Msgway@myri.com>

POC: Danny Cohen
Myricom, Inc.
325 N. Santa Anita Ave,
Arcadia, CA 91006

<Cohen@Myri.com>
Tel: 818-821-5555
Fax: 818-821-5316