# Caltech Mesh-Routing Chips

*Charles L. Seitz*
*Professor of Computer Science*
*California Institute of Technology*
*Pasadena CA 91125*
chuck@vlsi.caltech.edu

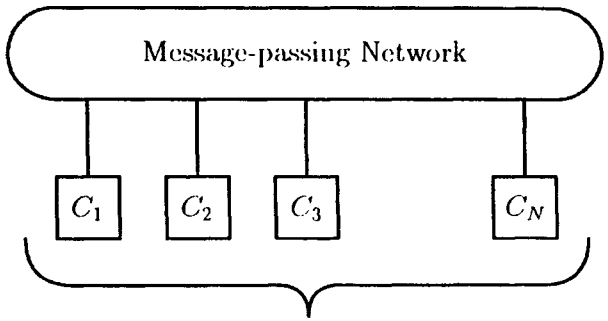
**Abstract:** The Caltech Mesh-Routing Chips (MRCs) perform cut-through packet routing on a two-dimensional mesh. An early version of MRC was used in the Symult S2010 multicomputer, and a current version is being used in the Intel Touchstone project and several other prototype MIMD systems. The five input and five output channels include channels to and from the node and to and from the four compass directions. Each channel is byte-wide and self-timed. In 1.2$\mu$m CMOS, the channels operate at approximately 100MB/s (800Mb/s), and the path-formation time is approximately 50ns per step. The internal design is self-timed, and is based on elementary cut-through routing circuits.

The research on which this talk is based was sponsored by the Defense Advanced Research Projects Agency. Caltech's proprietary interests in these chips are protected by maskwork registrations and patents.

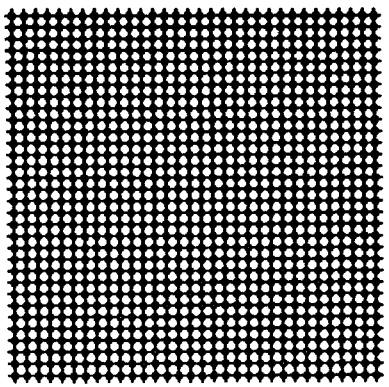


## Design Context

The Caltech Mesh-Routing Chips were designed to implement multicomputer message-passing networks:

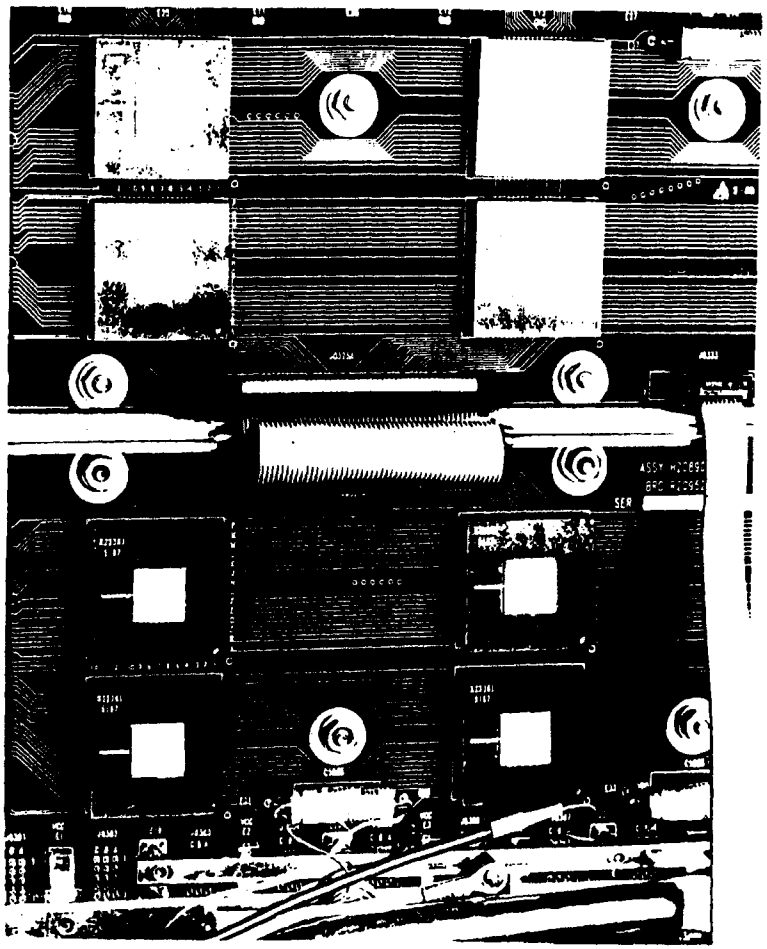



Such as:



## Design Context, continued

*Part of the routing-mesh backplane of the Symult S2010*



CLS

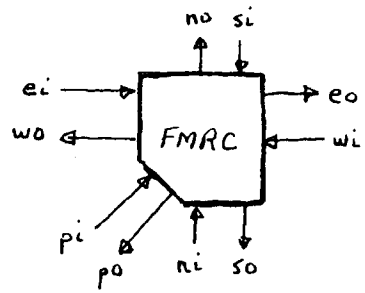


## "Frontier" Series of Mesh-Routing Chips (FMRC)

Designed for use in the DARPA-sponsored Intel Touchstone project. Also used in numerous other experimental MIMD systems.

*Signal Names*

The FMRC connects to 10 channels, 5 input and 5 output:




Each channel is 11 wires. The 110 signals have 3-character names formed as:

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | r | Request |
| | | | | | a | Acknowledge |
| | | | | | 0 | Data bit 0 |
| North | n | | | | 1 | Data bit 1 |
| South | s | | i | | 2 | Data bit 2 |
| East | e | + | o | + | 3 | Data bit 3 |
| West | w | | | | 4 | Data bit 4 |
| Node | p | | | | 5 | Data bit 5 |
| | | | | | 6 | Data bit 6 |
| | | | | | 7 | Data bit 7 |
| | | | | | t | Tail bit |

CLS

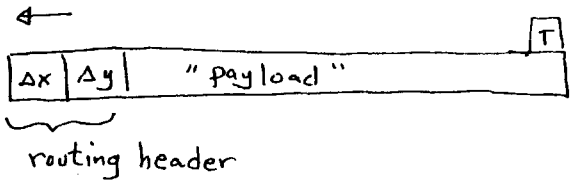
## FMRC, continued

*Request/Acknowledge timing*

The request, acknowledge, and data signals conform to the 2-cycle signaling discipline shown in figure 7.16 of Mead & Conway. (If the request signal were a clock, this is a 0-setup-time device, in which the hold time is determined by the corresponding acknowledge signal.)

*Packet Format*

A packet may be of any length so long as it includes the appropriate header.




*Routing*

Oblivious, dimension-order, cut-through routing on a 2D mesh. 19-stage internal queue on the typical internal path.

*Performance*

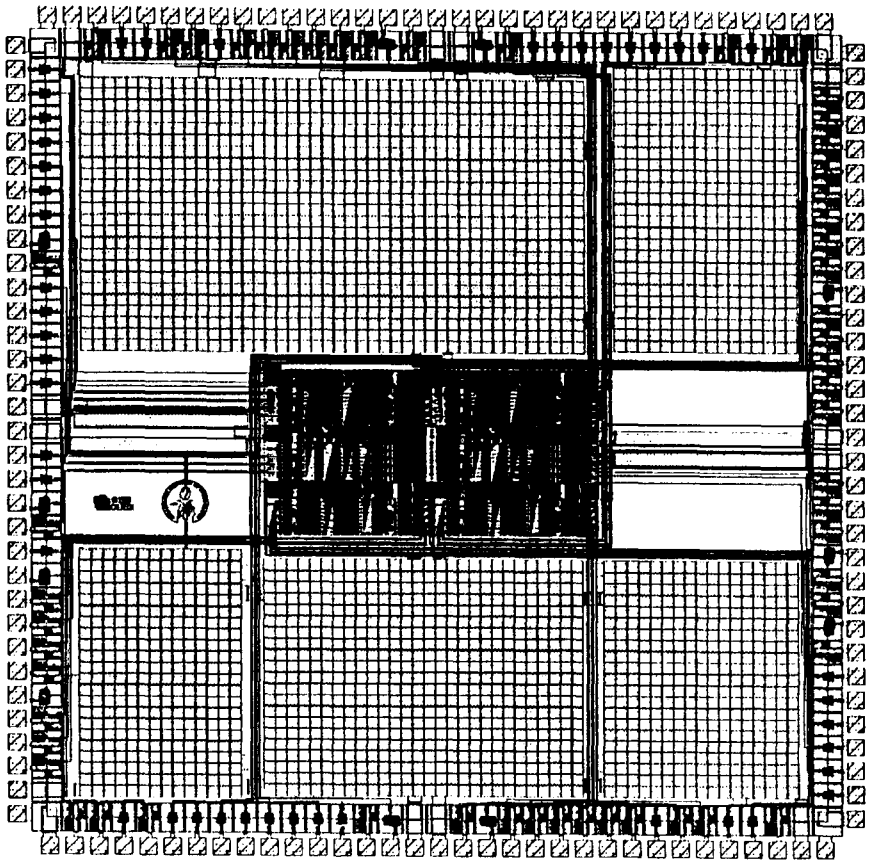
In 1.2$\mu$m MOSIS SCMOS (HP CMOS34) technology, $T_p = 50$ns, $B = 100$MB/s $= 800$Mb/s. ($B_{\text{internal}} \approx 200$MB/s.)

*Area*

Core area of the router is $A \approx 2\text{mm}^2$.



## Chip plot



## Small-area packet-routing networks

*Caution:* These multicomputer message-passing networks have little in common with local-area or wide-area networks.

- In a small, physically protected environment, the channels can operate at very high bandwidths, very small delays, and immeasurably low error rates.
- The networks are *regular* to permit simple, fast, algorithmic routing (no routing tables).
- The networks are *direct* and *bidirectional* to allow locality in the communication patterns to be exploited.

The design of these small-area packet-routing networks involves many mutually interdependent design choices and goals.

*Design choices:*

- Network topology.
- Flow-control methods.
- Routing methods.
- Deadlock-avoidance.
- Fairness.

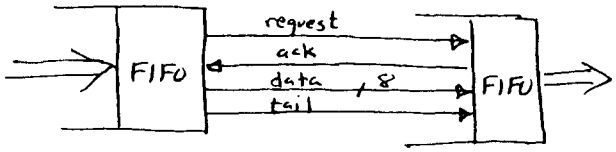
*Design goals:*

- High throughput.
- Low latency.
- High reliability.
- Low cost.
- Scalability.



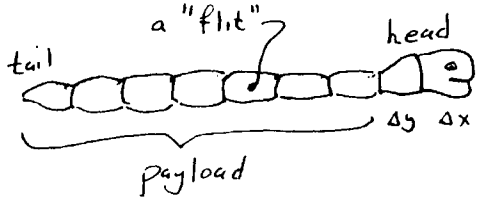
## The flow-control hierarchy:

The small delay on the physical channels allows the regulation of information flow in small units, referred to as flow-control units, or *flits*. For example, the flit in the FMRC is an 8-bit-parallel data item:




The channel can be considered to be a queue that conveys a sequence of flits. Each flit is acknowledged. If contention blocks a channel, the flow is blocked by the queue discipline.

Flits do not individually carry routing information. A sequence of flits forms a *packet* whose initial flits are a *header* that carries routing information, and whose last flit is tagged as the *tail*.

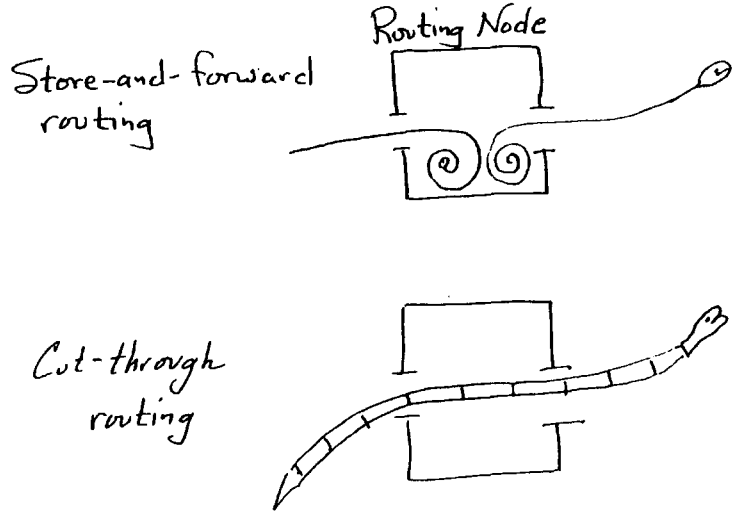



Fair interleaving of packets that require the same channel makes the packet the elementary unit not only of routing, but also of network fairness. Accordingly, packets have a maximal length. *Messages* can be composed of a sequence of packets.

## Cut-through routing:

The low error rate of the channels and the flit-level flow control allow aggressive routing methods. It isn't necessary to compute a check sum of a packet before advancing it to the outgoing channel. In *cut-through routing*, an incoming packet may advance directly to the required outgoing channel as soon as sufficient header is read.

For a packet of length $L$ and channels of bandwidth $B$, the time required to send a packet through a channel is $L/B$. *Store-and-forward routing* of this packet across $D$ channels then exhibits a network latency of $T_{\text{S\&F}} = D(L/B)$. However, if the packet can be advanced in cut-through routing in time $T_p$, the network latency is $T_{\text{CT}} = T_p D + L/B$.






## Contention and deadlock-freedom in cut-through routing:

- ○ *Virtual cut-through routing:* Revert to store-and-forward routing. [Kermani and Kleinrock, 1979]

  Deadlock-freedom is assured by the unbounded resources of the node.

- • *Wormhole routing:* Block the packet in place. [Seitz 1984, Dally & Seitz 1986 (virtual channels)].

  Deadlock-freedom can be assured by eliminating cyclic dependencies in the routing relations.

  For meshes (including hypercubes), cyclic dependencies can be eliminated by dimension-order (e-cube) routing. [Lang & Seitz 1981] This approach restricts the route to a unique path *(oblivious routing)*, but preserves packet order and performs well with little buffering in the network.

- ⋆ *Adaptive cut-through routing:* Be willing to misroute the packet. If possible, advance the packet into a channel that brings it closer to its destination, but if all such channels are blocked, divert the packet into an unprofitable channel. [Ngai & Seitz 1989]

  Deadlock-freedom is assured by the equal in- and out-degree of the network; however, a progress argument is required.

## Topology — why a 2D mesh?

A 2D mesh is a perfect fit to planar packaging technologies. It happens also to be nearly optimal with respect to packet latency.

Consider the family of $N$-node, $d$-dimensional (bidirectional) meshes, $N^{\frac{1}{d}} \times N^{\frac{1}{d}} \times N^{\frac{1}{d}} \times \ldots$ ($d$ times), $1 \leq d \leq \log_2 N$. Fix the *wire bisection* at $N/2$ wires in each direction; the wire bisection governs the throughput when packets are sent to randomly selected destinations, and also influences the cost.

The bisection is $N^{\frac{d-1}{d}}$ channels. If $b$ is the bandwidth of a single wire, the bandwidth allowed on each channel is $B = (b/2)N^{\frac{1}{d}}$. The average distance in the $d$-dimensional mesh is $\frac{1}{3}d(N^{\frac{1}{d}} - N^{-\frac{1}{d}})$.

What value of $d$ provides the lowest *worst-case-average-distance* latency with cut-through routing? In practice, $T_p \approx 2/b$, yielding $T_{CT} \approx (1/b)(\frac{2}{3}dN^{\frac{1}{d}} + 2LN^{-\frac{1}{d}})$. For $L = 256$ bits: