Proposal for an architecture for TUMULT based on a serial data link

J. Scholten
J. Hofstede
G.J.M. Smit
University of Twente
Department of Computer Science
P.O.Box 217
7500 AE Enschede
the Netherlands

Keywords: multiprocessors, interprocessor communication, serial networks, lan

A. Introduction

1. a short history of TUMULT

TUMULT is a project of the Department of Computer Science of the University of Twente and its aim is the design and implementation of a multiprocessor system. The project started in 1984 and now (at the end of 1986) two prototypes based on different communication networks are running. The software consists of a multitasking kernel, interprocess communication primitives, a distributed filesystem, a link manager, global and local task managers and utility software. The software is written in Modulair Pascal and Modula 2. A survey of the operating system and the communication primitives is given in [JANSEN].

2. the architectural model

TUMULT is a multiprocessor system in which up to 63 processing elements (PE) are interconnected [FIGURE 1]. The processors are loosely coupled, which means that there is no shared memory. Exchange of data is by means of message passing through an interprocessor communication network (IPC).

3. design considerations

There are some important design considerations for the IPC network:

the topology of the network must offer an any-to-any communication,
the network is the potential bottleneck for the overall system. If the network is not overloaded, performance is increased by adding processing elements. Hence the bandwidth of the network should be high compared with the communication speed of each processing element,
- the network must offer reliable communication links, such that a faulty communication path does not hold the system,
- for the programmer the logical complexity of the network is important.

4. the parallel IPC network

The second version of TUMULT is implemented with a parallel network. It uses a ring topology, which is relatively simple and has a good performance. A schematic representation of the system is shown in [FIGURE 2].

Figure 1: The architectural model of TUMULT

Figure 2: The parallel interprocessor communication network

Messages, containing 16 bits of data, are transferred in parallel over the network.
a) the network elements

The basic element of the network is the Message Switching Element (MSE), which switches incoming messages, according to the destination address, to the next MSE, or to the attached Processing Element (PE). The ring is synchronous, and works on a central clock. Each MSE has a buffer, which contains the message. At the clock pulse the message is transferred to the next MSE, or the PE. The clock rate is 10 MHz, so the data rate is 160 Mbit/sec, and the mean throughput is twice as much. This is very high in comparison to the communication speed of a PE, which is limited by the DMA controller (approx. 16 Mbit/sec). Transmission errors are handled by hardware and software: the hardware has a parity detection mechanism, the software maintains a checksum over the transmitted data and has a time-out mechanism. The main disadvantage of the ring topology is its reliability: if a MSE fails the whole system goes down. Therefore other topologies are considered now.

b) the network interface

For synchronization of the PE and the ring an interface is provided for. The interface is described in [SCHOLTEN], [SMT] and [Hoff].

B. Design objectives

The present TUMULT systems are implemented with 1: VME bus and 2: a parallel ring network as IPC. The PEs are housed in one 19" rack and are not geographically distributed. For several reasons it is desirable to have a serial IPC instead of a parallel one:

a- it has a less complicated interconnection backplane,
b- it makes geographical distribution possible,
c- one can construct hierarchical systems, in which a serial ring is used as backbone network,
d- redundant communication paths are implemented easily.

In the next chapter the alternatives for a serial network for TUMULT are given. In the remainder of this chapter the design or selection criteria are given.

1. performance

The performance of the network has to be optimal. The most important factors are:

throughput
the throughput of the network must be as high as possible. It is the product of the bandwidth and the network utilization. Throughput should stay very high under a load of very small messages. The aim is a bandwidth greater than 20 Mbit/sec,
distance
the distance between two adjacent nodes will be small.
access mechanism
the access mechanism must be suitable for realtime applications. The access delay time to the network must be sufficiently low, it must be fair, and it has to have a priority mechanism,
transmission delay
the transmission delay must be as short as possible.

transparency
the network should offer a transparent, non-complex service and interface to the software.

2. costs

The costs must be as low as possible. Factors are:

simple design
the complexity of the hardware influences both development costs and realization costs,
standardization
in general it is cheaper to buy (inexpensive) standard boards than to develop them oneself.

3. reliability

simple and careful design
the quality of design is one of the most important factors in the reliability of a network [CLARKE], robustness
this is the inherent insensitivity to negative effects, such as transmission errors and failure of components,
redundancy
when the robustness of a given topology is poor, the reliability may be increased by adding redundant links to the basic structure.

4. modular structure

The structure of the network must be such, that a stepwise extension is possible.

5. homogeneous structure

Costs, reliability and modular extendibility are influenced by application of identical nodes in the network.

6. flexible structure

The system must be flexible toward future developments and be able to incorporate them.

C. Alternatives

In this chapter alternatives for the most important aspects of the network are discussed. The first section explains that mechanisms, terminology and technology of local area networks are applicable to TUMULT. Then topology, access control mechanism, error recovery and physical aspects of the network are discussed and a selection is made. In the last section design decisions for the TUMULT network are given.

1. local area network characteristics

'A Computer Network is a set of autonomous, intelligent attributes ('computer-systems'), interconnected by communication means ('data-links'), via which interconnections are performed in order to complete one or more tasks.' [VAN SINDEREN]

Obviously, following this definition, TUMULT satisfies all conditions to be called a computer network. One way to classify these systems is by geographical spread of the computers. The following classes can be distinguished:
2. topology

The topology is an apparent aspect of a computer network. It has great impact on many other aspects, like capacity, access control mechanisms and reliability. The most important and well known topologies are:

- **meshed topology:** a topology with an irregular structure. It is used in WANs, where links are arranged according to load requirements in order to minimize the costs. Complexity routing decisions have to be made because messages between source and destination may be forwarded via different links.
- **ring topology:** in a ring or a loop, data flows always in the same direction and follows the same route. Therefore complexity of protocols for routing and data flow control can be low. Messages are repeated at every node between source and destination.
- **bus topology:** nodes are connected to a shared transmission medium. Messages travel from source to destination without interfering intermediate nodes. A bus is less vulnerable to node failures than a ring.
- **star topology:** a central computer is linked to all other systems by separate links. Protocols are simple. If each node has to communicate with any other node the central system will become the bottleneck. A failure of the central computer blocks the whole system.

### System types

**local area network (LAN):** interconnections are used for communications between systems in one room, or perhaps in one frame,

**wide area network (WAN):** communications performed over greater distances, like between cities or countries.

Whilst TUMULT with the parallel implementation of the datapath is classified as multiprocessor, TUMULT with the serial implementation of the datapath must be classified as local area network. A LAN covers a large section in this classification and it has certain properties in common with multiprocessors and WANs:

- a LAN has in common with a WAN that it has to bridge a noteworthy distance (<100 km) and the data has to transmitted serially,
- a LAN has in common with a multiprocessor bus that there is a shared communication medium and that some sort of distributed arbitration has to be done.

Important aspects LANs of different sizes share, are:

- data is transmitted serially,
- error recovery, because the links offer unreliable transmission facilities,
- a significant propagation delay,
- a distributed access control mechanism.

The following sides of LANs will be discussed in the next sections: transmission, topology, protocols and service.

3. access mechanism

Both ring and bus are a shared communication medium. Therefore the system must provide some sort of access control mechanism.

a) access control selection criteria

In [VAN SINDEREN] some criteria are given for evaluating access control mechanisms: (1) performance (resource utilization, access delay, throughput), (2) fairness, (3) priority and (4) robustness. These criteria will be used when discussing some general access mechanisms.

In a (packet switched) network the shared communication medium is allocated to a node for the duration of the transmission of a message. The access mechanism defines the time a node may access the transmission medium, and which node may access next the medium. The access mechanism can be centralized or distributed. The distributed version will be in general more robust.

Two main types of access control mechanism exist; the first type is based on collision detection, the second on collision prevention. Under a low load of the network the access delay is low with collision detection. However, if the load is high, throughput and access delay are better with collision prevention.

Examples of collision detection based systems are Ethernet and Twentenet (NIEMEGERS].

Three important access control mechanisms based on collision prevention are [VAN SINDEREN]:

- **token passing:** high throughput at high load is guaranteed. The mean access delay time grows with the load. Special care has to be taken to be sure the token does not disappear by a transmission error,
- **time slots:** two ways to control the access to the slots exist: the reservation ring and the demand assignment ring. The latter is not fair, because the maximal access delay can not be specified. It is used in the TUMULT parallel network and in the Cambridge ring.
- **register insertion:** the transfer delay is variable, depending on the number of output buffers switched into the ring.

Comparing the three access mechanism, the most important conclusions are:

- it is difficult to implement a slotted ring with a fair or priority based access to the ring, while keeping throughput and network utilization high,
- token passing and register insertion need special measures to increase robustness. In a token ring, a token may be corrupted or duplicated by transmission errors; in a register insertion ring all
Token passing performs well in terms of network utilization and fairness, but the access delay grows as the network load grows. To be suitable in a realtime environment a priority mechanism must be implemented.

b) token passing with priority

The basic principle for a priority mechanism is that when the network load increases, the low priority nodes are discriminated in accessing the network compared with high priority nodes. Arbitration can be done by (1) comparing the priority of all nodes and then allocating the token to the node with the highest priority, or by (2) measuring the network load at each node, whereupon the decision is made by the node to take the token or not. An example of (1) is the IEEE 802.5 Token Passing Standard ("IBM Token Ring"), which has 8 priority levels. An example of (2) is the ANSI X3T9.5-b FDDI proposal. This proposal uses a timed token access scheme. While in concept simple, it offers a powerful and elegant mechanism for priorities. It is meant for high performance LANs (100 Mbits/sec.) where the IEEE 802.5 mechanism is too complex and slow.

4. reliability

One of the motivations for building a multiprocessor is to increase the reliability of the system: if one PE goes down its tasks must be reallocated to other PEs. The network is the central part of such a system. Failure of the network might be fatal for the system. Therefore the network must offer a reliable service to the PEs, even when a failure occurs. Three kinds of failures can be distinguished: transient, intermittent and permanent failures.

Transient failures are caused by the unreliability of the medium. They may be detected and corrected by parity checks, checksums etc.

Permanent failures are caused by failure of a hardware component, for instance a cable break. The failing component must be isolated and its function must be taken over by redundant components.

Intermittent failures are also caused by malfunctioning hardware, but take place irregularly.

Transient or transmission errors can corrupt data and token messages in such a way that no PE at all can access the medium. The inherent tolerance of a system to such failures is called robustness. There is robustness of topology, of protocols and of access mechanism. Resilience is taken to mean the ability of the network to continue to handle traffic without loss of data: the face of faults and breakages [RICHARDSON]. The user should not be involved in the actions to accomplish continuation of service, although he might be aware of an (temporary) increase in the network access time ('graceful degradation'). Without special measures taken to achieve resilience, some inherent resilience is associated with the network topology, which is called robustness. Comparing bus, ring and star topologies, the bus has the best and the ring the least robustness in case of a node failure. If a link fails, the star has the best and the ring the least robustness. In any case the ring is the least robust one. However, there are techniques known, (reconfiguration and replication) to achieve a more resilient LAN [RICHARDSON].

For reconfiguration the network has some redundant paths to change the route of the messages. For replication all or part of the network is duplicated; in case of an error communications are switched to the redundant structure. One can distinguish three classes of faults in the LAN: the node failure, the tranceiver failure and the cable failure. Since the node is isolated from the network by the tranceiver, a node failure generally affects only one node. The tranceiver functions as a controller that checks the proper workings of the node. Tranceiver failures are serious, because they can bring down the whole network. If this type of failure occurs, the erroneous tranceiver must be bypassed. Problem is, how this bypass is activated when adjacent nodes detect the failure.

Cable failure causes the greatest problem. Some redundant path must be available to continue communications. Problems of reconfigurating arise.

A bus structure can be made redundant by duplicating bus and tranceivers. If care is taken to make extra routing decisions, the second bus may be applied to increase the capacity of the network. The costs of a redundant bus structure are high.

A ring offers many possibilities to add redundant paths in a regular fashion. The simplest one is a bypass relay over a tranceiver, or a tranceiver together with adjacent links. This is done in the first (Star Shaped Ring or SSR) and second (Mesh Loop or ML) example represented in [Figure 3]. The third example shows counter rotating rings (CRR), where failing links and nodes are isolated readily.

Figure 3: Redundant paths for rings
These three concepts must be compared, considering effectiveness, efficiency, costs and employability.

**Effectiveness**: the ability to handle faults; in this case restricted to single faults,

**efficiency**: the ability to isolate the failing part of the network without isolating correct functioning parts,

**costs**: the extra costs to establish redundancy into the network,

**employability**: the complexity of the procedures for automatically locating and isolating failing parts.

The effectiveness of all three concepts are sufficient, but a problem arises when the switch in the centre of a SSR fails. Attention must be paid to its design.

The efficiency of the SSR depends on the number of nodes in a meander. The ML and CRR are adaptable to less efficient versions.

Costs depend on the hardware used. The ML and CCR have complicated hardware in the nodes, which have multiple inputs and outputs. For the SSR the extra hardware is located in the wiring centre.

Distributed procedures can be defined for ML and CRR topologies to locate failing parts. For a SSR these procedures will be more centralized.

5. physical data transfer

Traditionally a characterization of a LAN follows from the medium used, because in large this defines the performance. In TUMULT, where only short distances are to be traversed, the question is more which medium suits the demands at the lowest cost. The demands are: high throughput, low bit error rate and reasonable cost.

Although both coaxial cable and fiber optic fulfill these demands the latter has more prospects to the future. The bandwidth is enormous, the bit error rate is very low and the prices will be competitive with those of coaxial cable [Beale]. Both kinds may be used and will depend on the application.

6. a new design

It would be wise to take a commercial product if it suits the application. In terms of costs it is preferable to select available, standard products.

a) commercial products

**Ethernet and Twentenet**: two CSMA/CD bus networks. Although the physical signalling is of interest for TUMULT, the basic protocols are not suitable for TUMULT’s realtime applications.

**PHILAN**: a fiber optic based, elegant and powerful LAN [Brandsma]. It does not fit well in the communication concept of TUMULT, but it could be used as a backbone network for a hierarchy of TUMULT systems.

**FDDI**: still a proposal and will be very expensive.

**IEEE 802.5**: Texas Instruments has a chipset available that implements the functions of the physical layer, the medium access control and the logical link control [TMS380]. The raw speed is limited to 4 Mbit/sec. which is to low for TUMULT. Above all, it will take considerable effort to realize an application.

b) ring versus bus

Before giving final objectives for a new design a decision has to be made which topology to use. In the previous sections pros and cons are listed for ring and bus. Here are some considerations:

- since a timed token access mechanism is chosen, a ring is preferable because it implies efficient and fair token passing; in a bus structure a logical ring has to be simulated,
- in a ring structure the stations are actively coupled to the network; this causes a few bit-times delay every time a token passes a station,
- a ring structure is best suited when fiber optic is to be used,
- a ring can be as reliable as a bus, provided the necessary measures are taken.

Taken this all in account a ring topology is best suited for TUMULT.

c) objectives for a new design

Main objective is to realize a network with high performance, while being simple and not too costly.

One must be aware that in the TUMULT system probably a significant part of all messages is small to medium sized. These messages must be transferred in an efficient way. Network speed must not be restricted by the speed of the PEs. Therefore a reasonable amount of buffer space must be available.

This implies that this buffer must be able to hold a small or medium sized message entirely, but not a large message. The maximum frame length will be small, and overhead for frame and token manipulation must be low. The mean token rotation time will also be low, which is an advantage for realtime applications.

D. Architecture

The new architecture for a serial data link for TUMULT is presented. Based on a thorough investigation of the alternatives and the design objectives, token passing is chosen as the basic concept for the network and a token rotation time scheme is used as priority mechanism.

The hardware is partitioned according to its functions.

1. hardware partitioning

In [Figure 4] a functional partitioning of the hardware of the network is given. There are three main components: (1) the processor interface (PI), (2) the ring access mechanism (RAC), and (3) the ring interface (RI).

(1) The PI’s basic function is synchronization of PE and network, and synchronization of sending and receiving PEs. A certain amount of buffering is therefore needed. The PI has six channels for sending and receiving data, control and exception messages. For data messages The PI has a hardware flow control mechanism. Its architecture is the same as for the parallel network, except for the output buffering. The basic information unit is the message, which is 16 bits wide.
(2) In the RAC the timed token access mechanism is implemented, which distributes available bandwidth among the channels according to relative priority. The other functions are relaying messages to other nodes (RAC is in REPEAT mode), and error detection. The basic information unit is a variable length frame, with a maximum size. The RAC contains a elasticity buffer, by which the incoming clock and the local clock are synchronized. It is combined with the hardware token grabbing functions, because both imply a delay for the messages.

(3) The function of the RI is the physical transfer of the bits of the frame over a certain distance.

2. processor interface

The PI forms the interface for the PE to the network. From the PE seen, the PI consists of a number of addresses. Through these addresses data, control and exception messages can be written and read. Besides that the PI has addressable status registers. As the PI is basically the same as the PI of the parallel version the reader is referred to [SCHOLTEN] an [SMIR]. Different is the buffering of the output data channel. Because the parallel version of the network is based on a slotted ring access mechanism, no output buffering is needed to be efficient. But in the new, serial version, several data messages are assembled into one frame, which is transmitted at a speed, that may be higher than the PE can handle. Therefore the output data channel is provided with a fifo.

3. ring access controller

The messages that flow through the PI have to be handled by the network. For a control and exception message an output channel at the sender is associated with an input channel at the receiver for the time it takes to transfer a single 16 bits message. For reasons of efficiency the association for a data channel is as long it takes to transfer a complete record, which may consist of several 16 bits messages. The RAC has control over the dynamic association of output and input channels at different nodes. Since messages are transferred over a shared medium, his medium must be allocated to a pair of input and output channels for a certain time. This is done through a token passing scheme, which is implemented in the RAC. To fulfill the medium access control functions, the RAC can be in one of two modes:

- **repeat mode:** all incoming data is repeated to the next node. If data is addressed to this node, the data is also copied to the PI.
- **transmit mode:** the ring is interrupted and incoming data is not repeated to the next node. Instead data from an output channel of the PI is transferred to the next node.

It is obvious that only the station that has the free token is allowed to be in the transmit mode. This station transmits one frame of variable length, which may contain messages to different nodes. Under normal conditions the working mode is controlled by RAC hardware. To recover from an error switching of modes is possible through software when executing a recovering procedure. A frame can contain four types of fields:

- **preamble (PA):** this field indicates the start of a frame,
- **frame control field (FCF):** it is always the first field of a frame, and contains the address of the sending node. Additionally it has an indication whether the token is free or busy. The format of the field is: token indication (TI) - 1 bit source address (SA) - 11 bits parity (Pr) - 2 bits unused - 2 bits
- **message control field (MCF):** this field determines the destination of the following information fields. Its format is:
  - type (Ty) - 2 bits
  - destination address (DA) - 11 bits
  - broadcast indication (BI) - 1 bit
  - parity (Pr) - 2 bits
- **information field (IF):** this field actually holds the information, which is 16 bits.

Two types of frame exist, distinguished by the TI:

- **token frame:** has a TI bit which indicates the free token. This frame only contains a FCF. A stations detecting a token frame knows that no data follows and can switch to transmit mode if it wants to send. Then it sends a data frame, followed by a token frame.
- **data frame:** has a TI bit which indicates a busy token, and the FCF is followed by one or more MCF-IF combinations. First control and exception messages are specified, then the packet acknowledgements (Pack), needed for the flow control, and finally the data messages. The end of the frame is indicated by the PA of the next frame.

4. ring interface

The function of the RI is to transfer the logical signals from node to node, by transforming them in a form which permits transmission over the physical medium. Furthermore a clock that indicates when data is valid, must be combined with the data to one signal. Data originating in the RAC are grouped in frames. For correct interpretation these frames must be preceded by a PA. It would be possible to generate the PA in the RC, but
its more efficient to do so in the RI:

1. PA is also used to synchronize the receiver clock with the incoming clock signal,
2. it is easy to generate a preamble by forcing errors into the applied coding, which depends on physical aspects.

Because the network has a ring structure, data travelling the ring is repeated at every station in the ring, which is an active act of the station. Therefore the station have to be powered. To overcome this demand the RI has a bypass relay that shortens the ring when the station is not powered.

F. Conclusion

Aim of the project is to design and realize a serial ring for TUMULT to replace the parallel version. In due course research developed in the direction of LAN technology. It was very laborious to get the right design objectives and to take the right decisions. Nevertheless we did come to a proposal for a token passing serial ring with a timed token priority mechanism. The next fase will be the implementation, which will take place in 1987-88.

Acknowledgements

Tumult is a joint research effort of several members (staff and students) of the department of computer science at Twente University of Technology. We would like to thank those who contributed to the project and this paper, especially Pierre Jansen and Harry Kaptein.

Glossary

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI</td>
<td>Broadcast Indication</td>
</tr>
<tr>
<td>CRR</td>
<td>Counter Rotating Ring</td>
</tr>
<tr>
<td>DA</td>
<td>Destination Address</td>
</tr>
<tr>
<td>FCF</td>
<td>Frame Control Field</td>
</tr>
<tr>
<td>IF</td>
<td>Information Field</td>
</tr>
<tr>
<td>IPC</td>
<td>Interprocessor Communication</td>
</tr>
<tr>
<td>LAN</td>
<td>Local Area Network</td>
</tr>
<tr>
<td>MCF</td>
<td>Message Control Field</td>
</tr>
<tr>
<td>ML</td>
<td>Meshed Loop or Chordal Ring</td>
</tr>
<tr>
<td>MSE</td>
<td>Message Switching Element</td>
</tr>
<tr>
<td>PA</td>
<td>Preamble</td>
</tr>
<tr>
<td>Pack</td>
<td>Packet Acknowledgement</td>
</tr>
<tr>
<td>PE</td>
<td>Processing Element</td>
</tr>
<tr>
<td>PI</td>
<td>Processor Interface</td>
</tr>
<tr>
<td>Pr</td>
<td>Parity</td>
</tr>
<tr>
<td>RAC</td>
<td>Ring Access Controller</td>
</tr>
<tr>
<td>RI</td>
<td>Ring Interface</td>
</tr>
<tr>
<td>SA</td>
<td>Source Address</td>
</tr>
<tr>
<td>SSR</td>
<td>Star Shaped Ring</td>
</tr>
<tr>
<td>TI</td>
<td>Token Indication</td>
</tr>
<tr>
<td>Tp</td>
<td>Type</td>
</tr>
<tr>
<td>TUMULT</td>
<td>Twente University Multiprocessor</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide Area Network</td>
</tr>
</tbody>
</table>

Literature