"If you know a mathematical proposition, that's not to say you yet know anything. If there is confusion in our operations, if everyone calculates differently, and each one differently at different times, then there isn't any calculating yet; if we agree, then we have only set our watches, but not yet measured any time.

If you know a mathematical proposition, that's not to say you yet know anything.

I.e., the mathematical proposition is only supposed to supply a framework for a description."

(Ludwig Wittgenstein, "Remarks on the Foundation of Mathematics").

CHAPTER 1.

INTRODUCTION

1.1. Overview

Presented in this thesis a design (description), reasoning, and analysis of a large area network, which is based on the hypergraph topology. One of the main challenges in computer engineering today is the question of how to connect a large number of computing nodes, which are distributed over a large area, into one integrated system. The functions performed by the system may vary from electronic mail and distributed file servers to parallel processing and real–time data communication (e.g., voice and video). The proposed hypergraph network is a realistic system architecture for a very large network, and can perform the above functions efficiently and in a distributed manner.

The work in this research is part of an AT&T project for developing a metropolitan area network. This term denotes a network having thousands of nodes within an area of about 1000 square kilometers and integrates the transmission of data, voice, and video. It will be based on high bandwidth, single mode, and optical fiber links.

The network topology is a hypergraph, which is a set of nets; each net is a set of two or more nodes [Ullm84]. The nets form the edges of a hypergraph, and the nodes are its vertices. The vertices of a particular edge have complete connectivity, so that a message can be transferred between any pair of its vertices. The nets or edges can be viewed also as buses, and each node has ports to several buses (at least one). This type of system is often called a bus–based topology or bus–connected architecture [RePa85].

One of the motivations for this research is to exploit recent technological advances, which make it possible to construct a high–performance, large–area network. The design will make use of the advances in the following areas:

(i) Single mode optical fibers, with a very high bandwidth, low attenuation (less than 0.5 db/km), and low spectral–group delay difference [Kapr85]; and low–loss, passive, optical star couplers.

(ii) High–speed logic devices from GaAs and ECL technologies, which make it possible to construct a digital electronic interface to a serial link with a transmission rate of more than one gigabit/second.

(iii) Very high–density custom and semicustom integrated circuits, for executing complex network control algorithms in real time.

An optical fiber is a dielectric waveguide which can transfer electromagnetic energy over long distances with very low losses. Electrical transmission lines have relatively low

efficiencies at frequencies higher than a few gigahertz. Optical waveguides are feasible well above these frequencies.

Optical fibers (see [Kapr85] and [Pers83]) have the following properties:

(1) Electrical Isolation – Optical communication enables data transfers among subsystems that are electrically isolated from one another, eliminating the problems of ground loops and ground noise.

(2) Noise Immunity – optical links are immune to most electric and electromagnetic noise sources; e.g., RFI, EMI.

(3) Low Loss – less than 0.5dB/km for single–mode fibers.

(4) Security – A fiber–optic link is difficult to tap without detection, and it is practically impossible to sense the transmitted information (due to its low losses).

(5) Passive optical coupling – This enables electromagnetic energy to be split between two fibers. The star topology is based on optical couplers. It has very good fault tolerant properties, and can continue to operate in the event of multiple failures.

The major motivation for high–speed fiber optic links (over 1 gigabit/second) is to decrease the dimension of the hypergraph, and to increase the width (number of ports from different nodes) on every hypergraph edge. As a result, the overall system's communication and computation control is simpler. It will be shown that it is possible to design high efficiency algorithms for a two– and three–dimensional hypergraph. These algorithms are executed in real time, independent of the distributed operating system software.

Each net of the network is realized as a centralized, passive, optical star. The messages from the net's ports are merged into one small area in space (a few inches), and are then broadcast back to all the net's ports. The centralized net's topology simplifies the synchronization of all the nodes on the net. Time is slotted, such that one time slot is about one message or packet interval. The centralized star topology can tolerate multiple failures and can detect collisions of packets of any size. The simple net synchronization together with the low hypergraph dimension make it possible to achieve global event synchronization, which is very important for implementing distributed functions and integrating real–time data communication.

1.2. Thesis

A low–dimension, a high width (the number of ports on a net), globally synchronized optical hypergraph, can be efficiently (with low overhead) managed and controlled in a distributed manner, such that:

(1) The proposed distributed algorithms (e.g., access control, synchronization, voice integration, overflow prevention) improve their performance as the state/time information propagates faster through the system. Increasingly higher bandwidth makes the slot duration shorter, which makes the state update more often. The high bandwidth enables the construction of a low–dimension hypergraph; therefore, the state information should propagate via, at most, one, two or three nets. The net and network global synchronization will enable a well–defined state transition at the end of each time slot, which simplifies the implementation of distributed algorithms.

(2) Algorithms for access control, global synchronization, routing, buffer management, and voice integration are independent of the distributed operating system. Furthermore, these algorithms are based on one another. Their construction is done bottom–up; a higher–level algorithm is based on lower–level algorithms. Thus, the routing algorithm is based on the properties of the access control and synchronization algorithms, the access control is derived from properties of the encoding/decoding scheme, and the encoding/decoding is based on the properties of the fiber optic communication medium.

(3) The net interface can tolerate permanent and intermittent physical failures by dynamic reconfiguration and adaptation of the network control algorithms. It will be shown that the centralized, passive optical star is very reliable and inherently redundant. The objectives of the fault tolerant mechanism are to minimize the probability of packet loss and to avoid the transmission of packets which cannot reach their destinations. From the reconfigurability of the system, it is possible to derive many applicable variants of the basic system, e.g., from the regular hypergraph it is possible to derive the partial hypergraph, such that both have the same distributed operating principles.

1.3. The Basic Construction Steps

The basic methodology for the design proceeds bottom–up, as shown in Figure 1.1. Using this design methodology it is possible to exploit the unique properties of the optical medium and to guarantee the physical realizability (or feasibility) of the design.

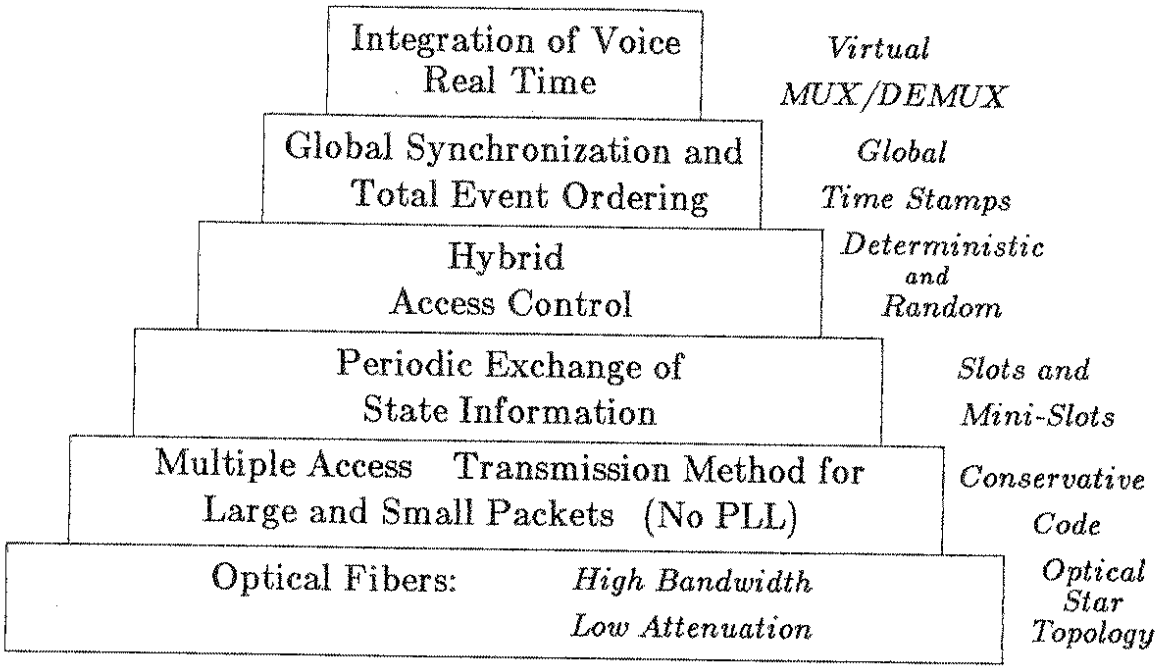

Figure 1.1: Design Methodology

Described in Chapter 2 the basic principles of the system's operation and provides an overview of its structure. The connectivity among the network nodes is defined, and the basic mechanism in which they interact is discussed. In Chapter 3 a new type of encoding scheme is proposed and analyzed. The code is designed to preserve the time information for self–clocking in serial communication over multiple–access channels. This new code is called conservative, since each codeword has a fixed number of transitions and a known delimiting transition at the end. It can be decoded without explicitly recovering the receiving clock with a phase–locked loop (PLL). Thus, using this new coding scheme, it is possible to broadcast short control messages, one after the other, from different and asynchronous nodes. These short control packets are essential for uniformly integrating various system functions. The code is analyzed under various constraints and is shown to be efficient. In Chapter 4 the design of a digital interface is

presented, making use of the conservative code. The critical timing path of the interface is optimized for maximum digital bandwidth.

In Chapter 5 a detailed design and analysis of a single net are presented. The emphasis in this chapter is on the mechanism for the periodic exchanging of state information and the way this information is used for the hybrid access control. Global synchronization and time stamping are incorporated into the system and presented in Chapter 6. The global event synchronization mechanism, which is imposed on the system, also uses the implicit and explicit time information, which is exchanged periodically by the control messages. All events are time stamped by a synchronous system clock, thereby preserving total event ordering.

The next step is combining the nets into an optical hypergraph network, which is described in Chapter 7, with emphasis on two basic configurations: two–dimensional regular hypergraphs and two–dimensional partial hypergraphs. The fault tolerance enhancement for the system is described in Chapter 8. In particular, it is shown how the loss of packets can be prevented, based again on the periodic exchange of state information. Chapter 9 shows how real–time (voice) communication is integrated into the system. The voice integration is an immediate outcome of the global synchronization. Final discussion, conclusions and further considerations are presented in Chapter 10.

1.4. Optical Architecture for Large Area Network

A metropolitan area network is supposed to be significantly larger than today's local area networks (e.g., Ethernet) with respect to the following parameters: (i) 10–100 times the number of nodes, (ii) 10–100 times the area, (iii) 10–100 times the medium

bandwidth, and (iv) capability to integrate several types of communication.

In general, an optical architecture is comprised of three parts: (i) the connectivity among the network nodes (i.e., the topology), (ii) the explicit physical properties of the network (e.g., transmission method, time partitioning, baud rate), and (iii) basic principles of operation, such as access control, routing, and buffer management. Many of the network examples which are found in the literature concentrate on only one of these three aspects, and often it is not clear how and how well the other architectural aspects are solved.

When examining the scope of optical architecture, the following classes can be identified:

(1) Centralized point–to–point architecture – similar to the telephone network. In this architecture high bandwidth fibers replace wires. This architecture is based on a centralized switch, which can be a major bottleneck. The major advantage of the centralized switch architecture is that it can be controlled more in a simple manner than a distributed architecture. A major disadvantage is that it is harder to extend this architecture. Two recent examples for this approach can be found in [LMHo86] and [VlLe84]. Clearly, the switch architecture is a possible solution, which is relatively well understood and simple to implement.

(2) Optical local area network (LAN) – this architecture is limited in its size and is constructed of a single shared medium (a ring or a bus), with possible gateways to other networks. A typical example is Fibernet II: A Fiber Optic Ethernet [SRNJB83], or Hubnet ([LeBo83] and [LeBoIk84]), D–NET [TCJ83], or active star LAN [Kama87].

(3) Backbone optical networks – which have been proposed for metropolitan area networks. These networks are characterized by having dual rings operating in opposite directions. Typical examples are FDDI [Joly84] and MAGNET [Lazar85]. In some configurations several of these dual rings are connected together [Sze85].

(4) Distributed networks – a network consisting of several nets, where each net is a shared medium or a bus. The switching of messages is done in a distributed manner. Most of these designs are top–down without explicitly showing how they are realized and operated. Typical examples are found in [Witt81], [Ullm84], and [RePa85].

The proposed optical hypergraph is a distributed network, as in the last example. The major differences are: (i) the design methodology is bottom–up, trying to show how all major aspects of the system can be realized, (ii) the design is based on some specific optical properties (single mode fibers, passive optical couplers), and (iii) the design tries to make an efficient use of the high bandwidth (about 1 gigabit/sec), utilizing different design principles (e.g., periodic exchange of state information).