ARCHITECTURE AND PROTOCOLS FOR HIGH-SPEED NETWORKS

EDITED BY

OTTO SPANIOL
ANDRÉ DANTHINE
WOLFGANG EFFELSBERG

KLUWER ACADEMIC PUBLISHERS

ARCHITECTURE AND PROTOCOLS FOR HIGH-SPEED NETWORKS

EDITED BY

Otto Spaniol
Technical University of Aachen
Aachen, Germany

Andre Danthine
University of Liege
Liege, Belgium

Wolfgang Effelsberg
University of Mannheim
Mannheim, Germany

KLUWER ACADEMIC PUBLISHERS
BOSTON / DORDRECHT / LONDON

Library of Congress Cataloging-in-Publication Data

Architecture and protocols for high-speed networks / edited by Otto Spaniol, André Danthine, Wolfgang Effelsberg.
p. cm.
ISBN 0-7923-9512-3 (alk. paper)
1. Computer network protocols. 2. Computer network architectures. 3. Asynchronous transfer mode. 4. Computer networks. I. Spaniol, Otto, 1945- . II. Danthine, A. III. Effelsberg, Wolfgang.
TK5105.55.A73 1994
004.6'5--dc20 94-31167
CIP

ISBN 0-7923-9512-3

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Kluwer Academic Publishers incorporates
the publishing programmes of
D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

Sold and distributed in the U.S.A. and Canada
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands

ARCHITECTURE AND PROTOCOLS FOR HIGH-SPEED NETWORKS

PREFACE

This book contains a selection of contributions (together with most recent additions and modifications made by the respective authors) which were presented at the First International Workshop on "Architecture and Protocols for High-Speed Networks", Schloss Dagstuhl, Germany (August 30 - September 3, 1993). The workshop was attended on an invitation basis by 35 international experts who discussed about actual problems and solutions in the rapidly expanding area of high-speed networking.

Major topics of the seminar were:
- switched networks, in particular ATM
- local and metropolitan area networks
- new network and transport layer protocols
- network applications, in particular multimedia applications
- protocol implementation on multiprocessors, and
- formal description techniques.

The general purpose of the workshop was to bring together telecommunications engineers and computer scientists, two groups of people who not very often have a chance to talk with each other. One of the hot topics was the status and future of ATM (Asynchronous Transfer Mode). Although ATM was initially designed to provide a wide-area high-speed telecommunications infrastructure, almost all of the installations and of the practical experience is concentrated on ATM switches in a local environment.

The new generation of applications in high-speed networks will contain multimedia data streams, i.e. digital audio and video. Continuous media streams, however, require transmission with guaranteed performance, in particular guaranteed bandwidth and bounds for delay and jitter. In addition to that, many multimedia applications will require peer-to-multipeer communication. Guaranteed performance can only be provided with resource reservation in the network, and efficient multipeer communication must be based on multicast support in the lower layers of the network.

Several manuscripts deal with internal structures for high-speed communication nodes. It is generally agreed that the performance bottleneck is currently in the end systems, upper layers and applications rather than in the

MAC adapters, on the links or in the switch fabrics. Parallel implementation of protocols on multiprocessors is considered as a promising solution.

All presentations were really excellent but due to the page number limitation the editors had to make a selection; less than 50 percent of the offered material could be included in this book. After a lot of discussions between the editors it was decided to concentrate on two areas which are in the center of interest for research and implementation of communication systems:

- protocol related aspects (switched networks, ATM, MAC layer, network and transport layer, traffic control, parallel processing,...)
- services and applications (multimedia systems, quality of service,...).

Even with such a concentration it turned out that a further selection was unavoidable. Since almost all authors delivered their manuscript in due time (the editors had never expected such an acceptance rate) the final versions had to be even more 'condensed' but this final procedure led to a significant increase in the quality of presentation (confer the following bon mot made by a famous person: "I'm writing a long letter since I couldn't afford the time to formulate a shorter one"; we apologize for writing a rather long preface!).

In most cases, it is difficult or impossible to exactly associate the manuscripts of the book to exactly one area since all relevant work must reflect aspects of different topics. The manuscript ordering, nevertheless, tries to follow basically the 'classical' rule: from lower layers to higher layers and finally to applications.

The manuscripts include a lot of new ideas resulting from very lively discussion rounds which were held in evening sessions during the workshop itself. The atmosphere of Dagstuhl castle was extremely positive for such intensive and fruitful discussions. Moreover, the fact that the participants were real experts in the field became a guarantee for critical but constructive comments; the editors are convinced that this interaction is visible in the manuscripts which have been updated several times, which have been thoroughly refereed and which present original unpublished material.

Otto Spaniol — **Technical University of Aachen, Germany**
André Danthine — **University of Liège, Belgium**
Wolfgang Effelsberg — **University of Mannheim, Germany**

CONTENTS

1

SIZE AND SPEED INSENSITIVE DISTRIBUTED QUEUE NETWORK

Z.L. Budrikis, A. Cantoni, J.L. Hullett

Australian Telecommunications Research Institute, Curtin University of Technology, Perth, Western Australia

ABSTRACT

A shared medium ATM switch in the form of a distributed queue dual bus (DQDB) network is described. Access to the network is in multiple stages, a separate stage for each branch of the network. Performance of the queue protocol is affected by distance within a stage, but not by distances between stages. In consequence a DQDT network can be of arbitrary extent without thereby limiting the rate, and DQDT networks can be designed with a speed-distance product that is orders of magnitude larger than is possible for DQDB, FDDI and other high speed LANs.

A DQDT network is asymmetrical in operation: Information is gathered from leaves towards the root of the concentrator tree, and is dispersed from the root towards the leaves of the distributor tree. The asymmetry makes DQDT a suitable component of a star-based network. It can serve as the first or concentrator stage in a wider ATM network.

1 INTRODUCTION

A shared medium network may be viewed as a time-shared communications server. As is well known, first-in-first-out, or FIFO queue service, outranks all other forms of service in efficiency and quality of performance. Little wonder then that, despite comparatively recent appearance, the distributed queue dual bus (DQDB) network has achieved prominence and attracted much attention and acceptance as a network standard [3,13,12,7].

As was pointed out by Mischa Schwarz [14], the performance of an M/G/1 queue and, for the case of constant length packets, of the M/D/1 queue, give an upper bound to the performance of *all possible* multiaccess strategies, where performance is judged by shortness of waiting time and smallness in the variance of the average waiting time. Strictly, the bound is established only for Poisson arrivals. But it appears true for all arrival patterns. Given a single server of fixed rate, the

average waiting time will be minimum if the server is never idle while there is a customer waiting for service. This is true for all queued service, irrespective of service discipline. A further significant minimum can be attributed to the FIFO discipline, namely that a server, serving in FIFO order, achieves not just a minimum in the average waiting time, but also a minimum variance in the average waiting time. This far-reaching observation was made by Kingman [8].

The variance in the waiting time is an important attribute of a communications service, particularly when the communications include real-time signals, because in reconstituting the time fabric at the receiver, the *largest* delay that occurs during the course of a connection contributes directly to the end-to-end delay of the signal [10]. Generally, a smaller variance in the delay signifies also a smaller largest delay that is possible.

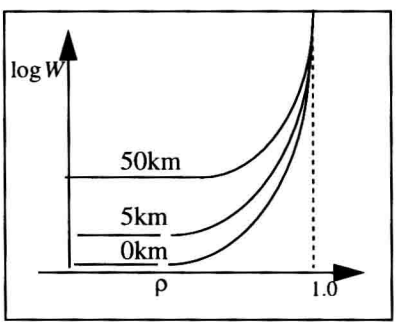

Figure 1 Waiting time with central scheduler

While a queue service gives the shortest possible waiting time, a *central* scheduler in a distributed network will not achieve this. The central scheduler would approach the theoretical bound only if both its processing time and the propagation time between it and the terminals on the network were negligible. However these delays are increasingly significant as the speed and size of the network grow. In fact the processing time and the two-way propagation time become an added constant to the waiting delay. This is illustrated in Figure 1, bringing out the effect of distance between terminal and scheduler.

Even with the additional delay due to signal propagation, the central scheduler will outperform most other practical multiaccess schemes. For instance, it will give a better performance in terms of maximum throughput, waiting time, and variance in waiting time than can be obtained with token passing, or with cyclic service, or with round-robin multiplexing. All these schemes have reduced performances due to larger protocol produced server idle times.

DQDB is exceptional among known practical schemes in performing better than the central scheduler. DQDB shares with the central scheduler FIFO queueing, but gives service that is almost totally free of protocol induced idleness and that does not suffer in the same way as that of the central scheduler from signal propagation delay. In DQDB access delay, averaged over all access points, is very little larger than the average waiting time in an ideal FIFO queue, and the

maximum throughput remains essentially 100 per cent even for very large networks.

Aspects of performance on which DQDB does deteriorate significantly with increasing network size and speed are adherence to strict FIFO service and to equality of service to all terminals under heavy load. Even though non-adherence to FIFO only affects the variance in waiting delay, and unfairness under overload can be effectively be eliminated at small cost in total throughput by bandwidth balancing [6], there is an effective limit on the practically possible size of a DQDB network. At 155 Mbps and cell size of 53 octets, the maximum end-to-end length of network may be put at 40 km [2].

We report the invention of another distributed queue network, based on a dual tree rather than dual bus topology, or DQDT [1], which in effect can circumvent the distance dependent problems of DQDB. It does so by employing staged queueing, thereby introducing the ability to restrict the distance over which the distributed queue protocol is implemented, without restricting the overall distance spanned by the network.

2 DESCRIPTION OF THE DQDT NETWORK

Just like DQDB, the DQDT is a network with oppositely directed unidirectional information flows, with the information in fixed size, contiguous time slots or cells. However, unlike DQDB which is operated symmetrically, DQDT is asymmetrical. Information is written only on the one flow stream, the *concentrator tree*, and read from the other, the *distributor tree*. The buses of DQDT are referred to as trees because, as illustrated in the schematic of Figure 2, they may be in the form of trees, with branchesup to arbitrary order. The network is a *dual tree* because the concentrator and distributor are each a tree, one with information flow towards its root, the other with flow away from root. In the case of the network functioning as a stand-alone switch, the information that arrives at the concentrator root would be transferred directly to the distributor root, as indicated in Figure 2. The network may also function as a first stage concentrator/distributor element in a larger switched network when the roots would be connected to input and output of a switch port.

The disk-shaped elements in Figure 2 are media adaptors (MAs), and the square elements represent terminal equipments (TEs). As a general rule, MAs are three-port devices with two in-line ports and one lateral port. Tree limbs (stem or branches) are formed by in-line interconnection of MAs. The lateral port of an MA can take the attachment of either a TE or of a next higher order branch of the tree.

In ordinary LAN practice the media adaption function would be contained in the terminal, without any exposed interface between MA and TE. In DQDT an exposed interface is mandatory because, for one and as already noted, a branch is created by plugging into a media adaptor a section of network in place of a terminal, and for another and more importantly, to exploit the distance capabilities of DQDT to the full, media adaptors on the same branch should be as close to each other as possible, suggesting that they be clustered into a hub, and hence that terminals and the media adaptors to which they attach, be physically separated from each other.

The function of the whole DQDT network becomes clear from the function of a single media adaptor. Figure 3 shows the block schematic of one media adaptor.U and V are shown as plugs and are at the crown end of the adaptor, V on the concentrator and U on the distributor. X and Y are sockets at the root end, and A

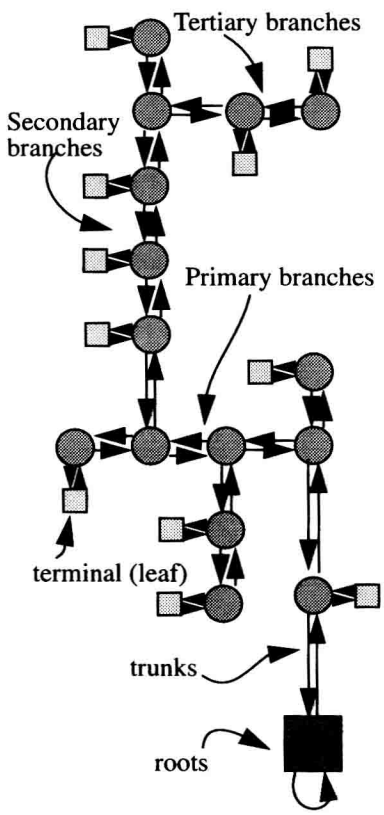

Figure 2 The DQDT Network

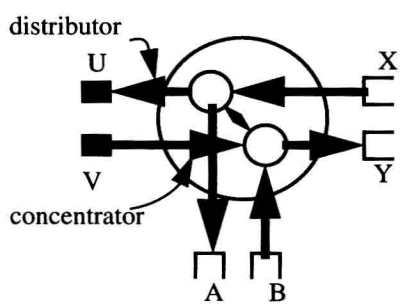

Figure 3 Schematic of a media adaptor

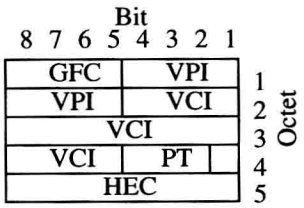

Figure 4 Header format

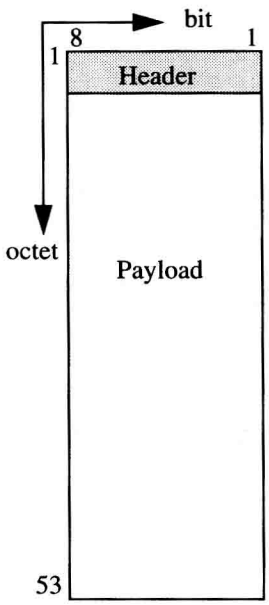

Figure 5 Cell format

and B are lateral sockets. Information enters the MA from a TE, or equivalently from a higher order branch, through B and has to be inserted into the concentrator stream which enters at V and exits at Y. Similarly, information enters in a distributor stream through X and is passed on to U as well as copied to terminal, or branch, through A. To insert the information into the concentrator stream, the MA uses the distributed queueing protocol.

Distributed queueing requires twin information streams, "upstream" and "downstream", carried in slots and accessed by all participating nodes. The information format must include an access control field. There are no further restrictions. Many different slot formats are possible. For the sake of description, and with an eye on future application for our network, we take the format defined for ATM at the user-network interface (UNI) in B_ISDN [4]. The slot, or cell as known in B_ISDN, is of 53 octets of which five octets are the header and 48 octets payload, as shown in Figure 4. The header format, as defined at the UNI, is shown in Figure 5. The very first four bits make up the Generic Flow Control Field (GFC) which can serve as access control information field for the distributed queueing in DQDT.

GFC protocol and procedures are still in the process of definition in ITU, with finality expected in 1995. The GFC that is being defined in ITU, is for control across the S and T reference points in the ITU defined customer premises network. and therefore will apply to the lateral, or TE-to-MA, flow in DQDT. The definition will provide for admission control on ATM connections that have no guaranteed bandwidths (the class of 'controlled' connections), and will give an effective back-pressure control through a credit reset/no-reset scheme.

The queue access control along branches and trunk must provide control for all access, including on connections that have guaranteed bandwidths.To achieve this it can be similar to the Access Control Field (ACF) of IEEE 802.6 in providing for queues of multiple priorities. In queueing, a node sends requests for empty slots to nodes upstream of itself, which for DQDT would mean that

requests are sent on the distributor tree. There is no essential role for GFC on the concentrator where it could for instance be defined to indicate whether the particular cell is on a bandwidth resourced or unresourced connection.

We propose that the GFC on the concentrator be (S, R_0, R_1, R_2) where S is a STOP bit, R_0, R_1, and R_2 are REQuest bits, respectively at priority levels 0, 1, and 2. Priority level 0 is the lowest and is used for unresourced connections corresponding to the 'controlled' connections of the ITU definition, while Priority levels 1 and 2 are for resourced connections. The highest level would be intended for connections that go over the T reference point to the public network, and the middle level for intra-premises or local connections. The STOP bit, when set, would stop access for one cell period to priority level 0 cells, i.e. to cells on 'controlled' connections.

Distributed queueing is implemented separately on each limb (trunk or branch) of DQDT. A higher order branch offers traffic to a lower order branch (or trunk) that it terminates onto, no differently from the traffic offered by a terminal. Thus at each priority level, the concentrator of DQDT a cascade of queue stages, as illustrated by Figure 6. Assuming that the distributed queueing performs ideally, each stage is equivalent to a FIFO buffer with parallel inputs.

The servers are in all cases slotted and have different effective rates for the three levels of priority: At priority level 2 the server is at bus rate, at level 1 it is at bus rate less the actual service at level 2, and at level 0 it is at bus rate less the actual service at levels 2 and 1. Queues at priority levels 2 and 1 are stable because inputs are regulated by contracts at Call Admission, and at priority level 0 are made stable by back-pressure control exerted through the GFC.

A distributed queue will approximate very closely the ideal single-server queue and will adhere strictly to priorities in the case of priority queueing, as long as REQuest propagation delays are shorter than one cell period [11]. With the queueing independent on each branch, the relevant delay is confined to a branch and, more precisely, to the node-to-node delay along the distributor member of the branch. Significantly, any delay between branch point and the nodes has no relevance to the queue protocol and hence a branch may have an arbitrary length from branch-point to first node, without affecting the functioning of the distributed queue protocol. To assure then ideal single-server queue behaviour and strict adherence to priority in service, the number of nodes on a branch and their spanned distance must be so limited that the maximum protocol delay does not exceed the set ceiling of one cell period.

The limit on spanned distance can be met easily by placing all nodes of a branch into a single cluster or hub. Figure 7 shows (with artistic license) how a DQDT network, with only one cluster per branch, might look. A cluster may have the extent of just a backplane and have a physical propagation span of less than one