ON LANs AND MANs: AN EVOLUTION FROM Mbit/s TO Gbit/s*)

Pitro Zafiropulo
IBM Research Division, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland

This paper focuses on the trend towards ever higher speeds in shared-medium based networks such as LANs and MANs. The token passing ring protocol is introduced with emphasis on its performance and simplicity advantages over the token passing bus approach. The issues of scaling up the token ring principle to higher speeds is considered with the help of the Early Token Release and FDDI protocols. Although networks based on token passing operate efficiently at speeds up to 100 Mbit/s it is shown that they exhibit inherent performance limitations in the Gbit/s speed range. This leads one to consider the advantages of protocols based on slot reservation at very high speeds. An example of this is the Distributed Queue Dual Bus (DQDB) MAN proposal. Unpredictable behavior and lack of fairness, however, make it inadequate at very high speeds so that other approaches must be considered.

1. Introduction

Local Area Networks (LANs) are being used today in myriad different applications spanning office and industrial environments. Many different LAN networks support these applications adequately at speeds ranging from 1 to 20 Mbit/s. As the established LAN markets have grown, the need for interconnecting LANs into larger networks has spawned new market potentials not only in the area of LAN bridges, routers and gateways, but also in backbone networks of interconnecting LANs for larger geographical coverages. Important applications of emerging Metropolitan Area Networks (MAN) are such backbone functions which include enterprise, corporate and campus wide configurations. MANs are high-speed networks covering geographical distances of 50 km and greater, and which may include public domain. They are similar to LANs in that they use a common medium access control (MAC) protocol but are required to support isochronous traffic. Today, backbone networks range from LAN speeds up to 100 Mbit/s and are envisioned for 155 Mbit/s. However, as multiple LAN networks grow and LAN speeds increase, fiber-based backbone networks well beyond 155 Mbit/s will be required.

The bulk of LANs in use today carry alphanumeric data, but graphic payload has the potential of greatly surpassing that of alphanumerics in certain application environments. A single medium to high-resolution graphics display can require anything from 0.3 to 6 Mbytes per picture frame. High-resolution graphics workstations operating on files provided by graphics servers in a client/server LAN environment with browsing capabilities may well require data rates in the many hundreds of Mbit/s to even the gigabit per second (Gbit/s) speed range. Attention is also being focused on motion graphics such as visualization and animation which require even greater bandwidths when operated interactively over supercomputer distribution networks.

More graphics-based applications have yet to be invented, but when one considers the potential of motion graphics as a new form of man/machine interface[1] then it is time to consider LANs and MANs for Gbit/s and near Gbit/s speeds. The seriousness of this trend is reflected by the increased activity in such networks.[2-6] This development may well be accelerated by the forthcoming 0.8 Gbit/s high-speed channel standard HSC (newly renamed to High-Performance Peripheral Interface, HPPI) proposed by ANSI X3T9.3.[7] This standard developed for transporting visualization payload from a supercomputer to a workstation has the potential of becoming an attractive attachment to Gbit/s networks.

This paper focuses on the trend towards ever higher speeds in networks based on shared media access caused by these applications, a situation which may be compounded in the future. The token passing ring protocol is introduced with emphasis on its advantages of better performance and greater simplicity over the token passing bus approach. The issue of scaling up the token ring principle to higher speeds is investigated with the help of the Early Token Release and FDDI protocols. Although networks based on token passing operate efficiently at speeds up to 100 Mbit/s it is shown that they exhibit inherent performance limitations in the gigabit/s speed range. This leads us to consider the advantages of slot reservation based protocols for very high speeds. An example of this is the Distributed Queue Dual Bus (DQDB) MAN proposal. However, lack of fairness and unpredictable behavior make it inadequate at very high speeds so that other approaches must be considered. The paper concludes with a promising MAC protocol for Gbit/s speeds.

2. Token Bus, Token Ring Evolution

The token passing bus originally owes its significance to its higher throughput under heavy load than can be achieved by the CSMA/CD bus protocol for small packet lengths.[8] Its principle[9] as outlined in Fig. 1 is that stations wait to receive the token before sending data. On ending transmission, stations pass the token to their 'next station neighbor' via a next station (NS) address. The token thereby circulates past all attached stations to form a

*) This paper appeared in the Proceedings of the EFOC/LAN Conf., Munich, 1990.

logical ring mapped onto the physical bus. The length of this logical ring can become large when next station neighbors do not map into physical bus neighbors, see Fig. 1. This may cause large access delays for larger network coverages.

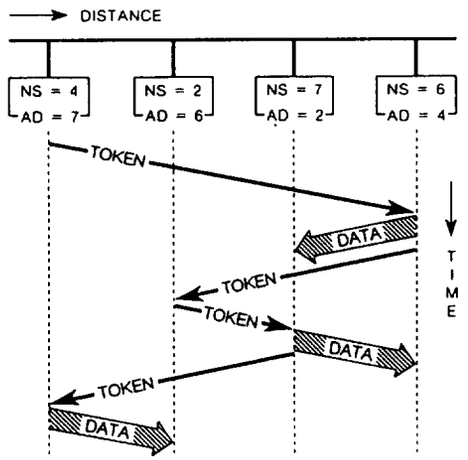

Fig. 1: Principle of the token passing bus protocol.

Station insertion is complicated because of this logical to physical mapping. Consider the simplest case shown in Fig. 2: if the token rotation time, i.e. bus loading, is sufficiently small, then station i (belonging to a predefined station subgroup) broadcasts a Solicit Successor message with i as its own address and j as that of its NS. If a station, the address k of which lies between i and j, wishes to enter, it responds by taking j as its NS and issuing a Set-Successor k message. Station i responds to this message by modifying its NS to k. The NS rearrangement ensures that the token will pass to the newly inserted station. The algorithm[9] must be extended to include cases where the inserting station address k is larger/smaller than that of all active stations; k then lies outside all (i, j) pairs. Further complications occur when a plurality of inserting nodes have addresses that lie within the same (i, j) bound.

Mapping the logical token ring onto a physical ring yields the Token Ring media access protocol.[10-12] The Token Ring principle is shown in Fig. 3. A node wishing to send modifies the passing free token status to busy, opens the ring, and issues the data frame preceded by a source, destination address pair. The receiver, recognizing the destination address as its own, copies the frame. The sending node issues a free token on detecting its source address, thereby enabling downstream nodes to transmit. Finally the ring closes when the frame completely removes itself from the ring.

The token ring minimizes the token rotation time compared with that of the token bus, thereby improving throughput.[13] It also dramatically simplifies node insertion because the token automatically passes all stations on the ring, including newly inserted ones. The simplicity of the token passing ring algorithm is the reason for its robustness.

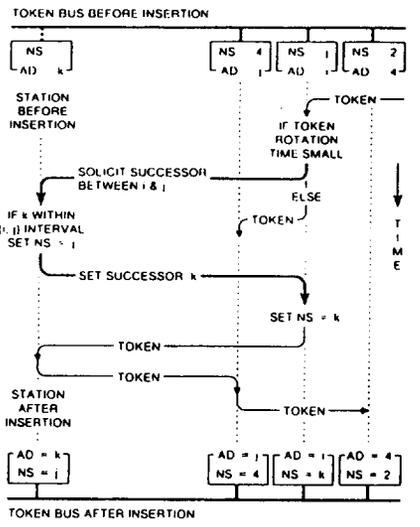

Fig. 2: Example of station insertion into a token bus network.

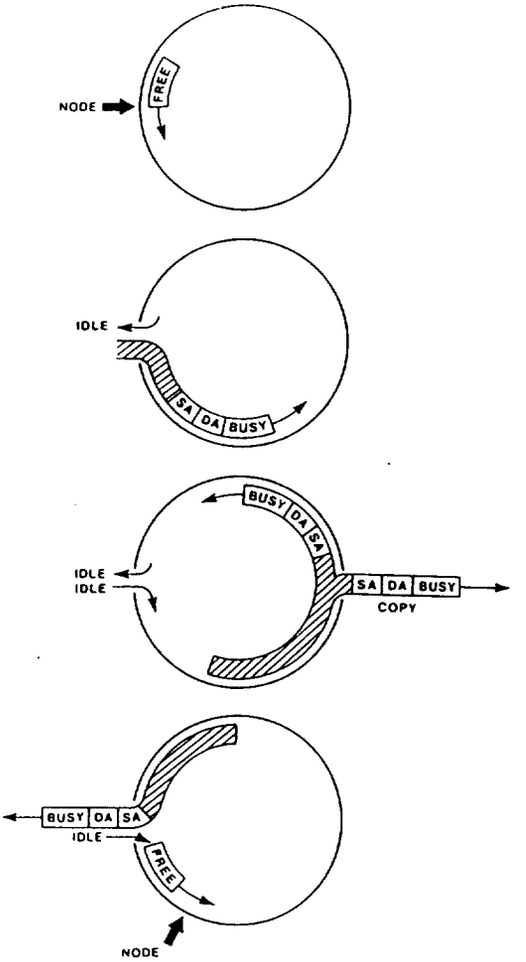

Fig. 3: Principle of the token passing ring protocol.

3. Extending the Token Ring Principle to Higher Speeds

Higher transmission speeds affect the token ring in the way shown in Fig. 4. Increasing the bit rate shrinks the length coverage of a given frame on the ring. Thus, if single frames just fit the ring for a transmission speed S_0, then the ring's throughput does not increase with a further increase in speed. This is so because the token issuing time is then determined by the ring latency and no longer by the frame transmission time. Adverse performance results when traffic is dominated by short frames. If for example a 40 km ring (neglecting node delays) is momentarily loaded by a flurry of 128 byte frames then the aggregate ring throughput drops below 5.12 Mbit/s, regardless of how fast the ring is operated above S_0 (= 5.12 Mbit/s).

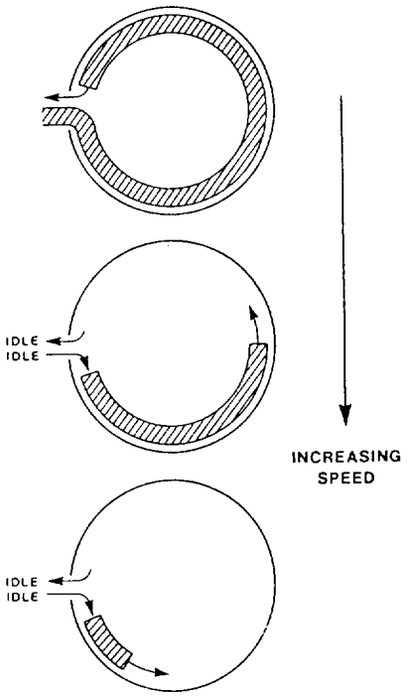

Fig. 4: Shrinkage of a frame's geographical coverage with increasing ring speed.

This performance bottleneck is eliminated by removing the 'one frame on the fly' limitation. Allowing multiple frames maintains a heavily loaded ring full, even in the presence of short frames and can therefore significantly increase aggregate throughput. It is achieved by sending stations issuing the free token immediately after the frame-end is transmitted. This way the next sender can place its frame directly behind the previous one, see Fig. 5. This extension, fittingly called the Early Token Release (ETR),[14,15] is successfully being used on 4/16 Mbit/s token rings.

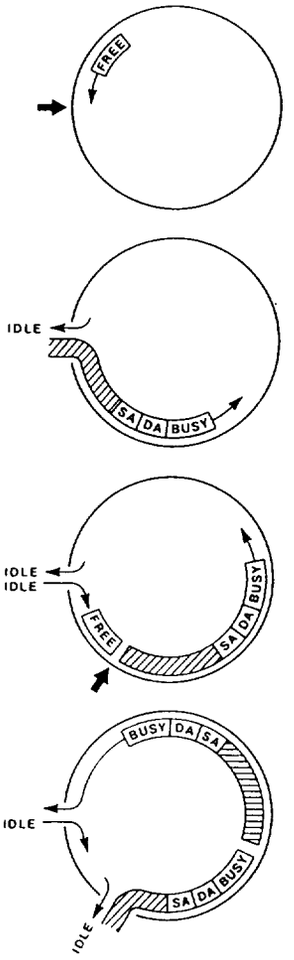

Fig. 5: Principle of the Early Token Release protocol.

The original token ring protocol supports priorities[11] by adding a priority field in front of and a reservation field behind the token. The underlying principle is as follows: a high-priority node wishing to send when the passing token is busy sets the reservation field. On seeing this, the low-priority sender responds by setting the priority field as it generates the free token. This modifies the token for exclusive use by high-priority nodes. When the priority traffic subsides, the low-priority node that initiated the change reverts the token to its original low-priority status.

This priority scheme does not work well with the ETR protocol. The reason is that sending nodes can issue the free token before the transmitted frame returns, thus the free token is already gone when the returning reservation arrives. This shortcoming is largely overcome by having a high-priority node repeat the reservation on successive passing frames until it obtains a free token, be it of high or low priority.

Loading an ETR ring slows down the free token rotation because no free token is issued while a frame is being placed on the ring. Two known priority schemes are based on this slowing effect. The first one[16] extends the token

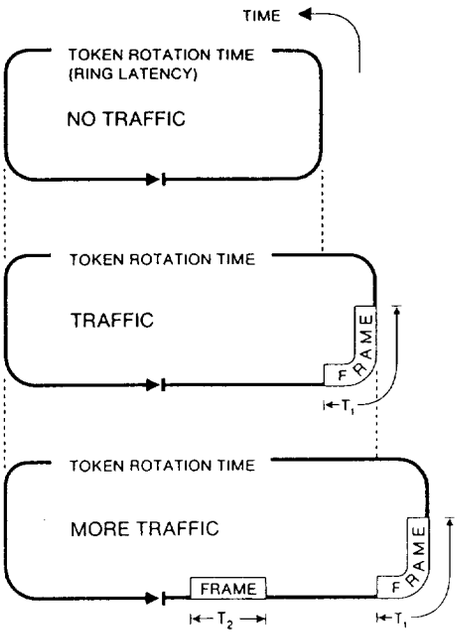

Fig. 6: Token rotation time increase with growing traffic load. Each frame issued during a token rotation interval increases rotation time by that frame's transmit time *T*.

ring priority mechanism as follows: a sender node responds to a priority setting in its frame reservation field by letting the header go by on the ring. The header acts as priority change signal that catches up with and modifies the slower travelling token. The second scheme uses a different approach based on measuring the token rotation time. The token rotation time approach first appeared in the token bus protocol[9] but is mainly known from its use in the FDDI token ring[17,18] the principle of which we will describe in the context of the speed issue.

When the ring speed is further increased, a point is reached at which it becomes simpler to remove the token from the ring than to modify its status on the fly to busy. In the early 80s, this change made sense for 100 Mbit/s, owing to the speed of the cost-effective technology then available. This approach was chosen for FDDI – a sending node removes the token and then re-issues it, after placing its last frame for transmission on the ring. Token removal eliminates the need to differentiate between busy and free tokens. As mentioned above, the priority principle is based on nodes that measure the token rotation time which, in turn, depends on the ring loading: see Fig. 6. A rotation time threshold value determines whether only high-priority frames or both high and low-priority frames can be transmitted. Today FDDI is standardized at 100 Mbit/s[18] where it exhibits good aggregate throughput.

4. Beyond 100 Mbit/s

How does the FDDI token passing principle stand up to speeds well beyond 100 Mbit/s? The key to this question is token propagation. When the token travels from one node to the next, no station can access the ring for transmission. One can therefore speak of a nonproductive token propagation delay. The faster the transmission the shorter the frames on the ring for some upper frame size limit. Thus the effect of nonproductive token propagation delay becomes more dominant with increasing ring speed and size. Consider the simplified case of N active stations, each transmitting frames during the same time interval t. Let T be the ring latency, then the ring utilization U is given by

$$U = N \cdot t/(N \cdot t + T)$$

$$U = N/(N + \mathbf{a}) \ ,$$

where **a** is the propagation delay across the entire network medium. Expressed in numbers of back-to-back frames that can coexist in the medium,[19] it is given by

$$\mathbf{a} = (S \cdot L)/(V \cdot P) \ ,$$

where S is the speed of the medium in bit/s, L the length of the medium in meters, V the signal propagation speed over the medium in m/s and P the frame size in bits. The parameter **a** is thus directly proportional to the product of the speed and distance of the network. To obtain some understanding of the numbers involved, consider a 10 km, 100 Mbit/s FDDI ring supporting 10 powerful file servers that generate the bulk of the data consisting of 2 Kbyte frames. The maximum utilization U for this configuration is 0.97. This excellent utilization drops to a mere 0.21 if the above configuration is accelerated to 1.2 Gbit/s and extended to 100 km. One can argue that utilization is improved by allowing larger frames or more frame transmissions per token acquisition, which is true but only at the cost of increased ring access delay.

This inherent limitation also holds for schemes based on implicit token passing.[20] It can be avoided by selecting a configuration which allows nodes to access the network simultaneously in a non-sequential manner. One such configuration is the slotted ring or bus format (see Fig. 7) where node A can transmit before, during or after node B. Buffer insertion is another such medium (see Fig. 8) but is not well suited to supporting isochronous traffic, which is a MAN requirement.[21] Isochronous channels require fixed transfer delays, a property that buffer insertion cannot easily fulfill because the higher the loading, the more buffers are traversed and hence the longer is the delay.

Bus configurations require no garbage collection in that bit or burst errors eventually drop off the bus end, whereas ring configurations require identification of the error followed by removal and token generation. Furthermore slotted busses have no delay constraints as far as bus ends are concerned. On the other hand slotted rings require automatic delay equalization to ensure an integer number of bits and slots on the media.

In an environment with a light to medium load, sharing a slotted bus requires only a busy/free slot indicator, for there are enough available slots for the transmission needs of each user. However, when heavy loading or overloading occurs then a slot reservation scheme is needed to ensure fair access among all active

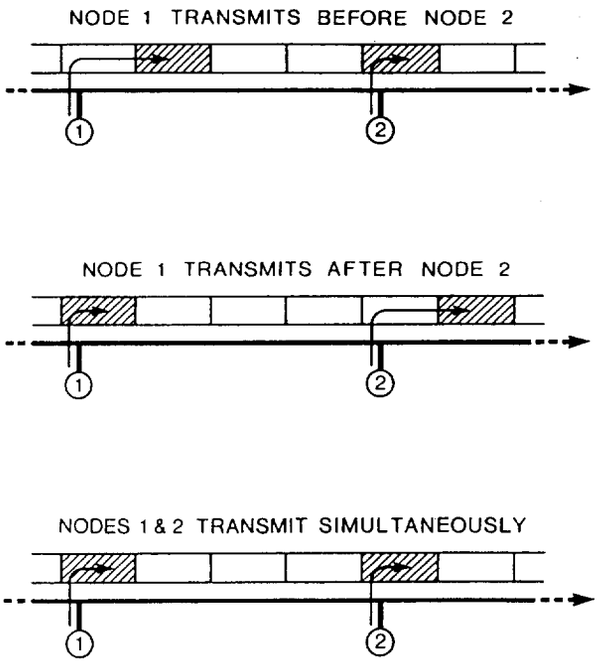

Fig. 7: Two nodes transmitting over a slotted format exemplify the nonsequential nature of slotted access.

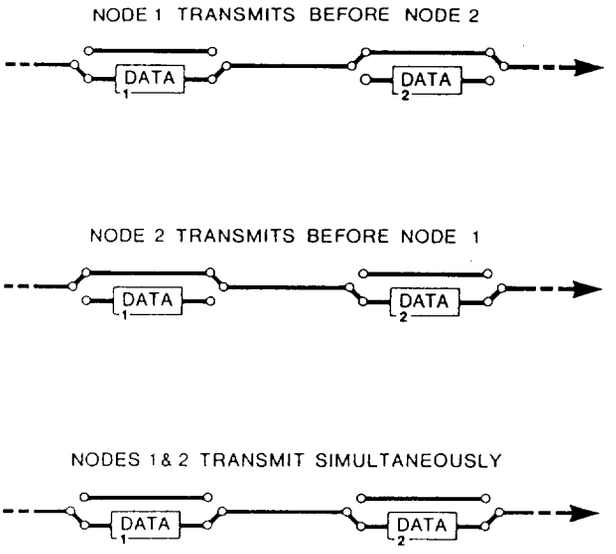

Fig. 8: Nonsequential access nature of buffer insertion.

nodes. Thus a viable approach to very high-speed LANs and MANs seems to be a slotted bus approach allowing good utilization over large geographical coverages coupled with a slot reservation scheme to achieve fair behavior under heavy loading. Figure 9 shows two slotted bus structures of interest. Firstly, there is a folded structure in which a headend located at one end of the bus generates a continuous slot stream. A node wishing to transmit invokes a reservation algorithm for slots on the outbound bus section whereby reception occurs on the inbound section. Secondly, there is a dual bus structure in which headends placed at alternate ends of each bus generate continuous slot streams. A node wishing to transmit to a right-hand (or left-hand) partner invokes a reservation algorithm for slots on bus A (or bus B).

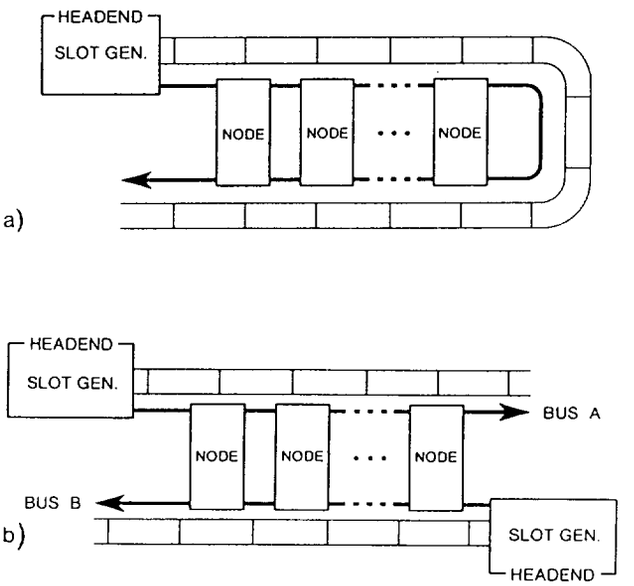

Fig. 9: Slotted bus configurations.

5. The DQDB Protocol

The IEEE 802.6 MAN standard proposal called DQDB (Distributed Queue Dual Bus)[22,23] exhibits precisely those features that are well suited to very high-speed operation, namely a slotted bus with reservation. The DQDB structure belongs to the class shown in Fig. 9b. Its basic principle is explained (see Fig. 10) in terms of FIFO queues.[24] Slot headers include a busy/free bit followed by a request bit for reservation purposes. The FIFO queue in Fig. 10a represents a node's mechanism for slot reservation on bus A. Nodes wishing to transmit place a REQ in the first available request field on bus B. Nodes duplicate all passing REQs into their FIFO queue. If a free slot passes on bus A when the FIFO head element is a REQ then the node refrains from using it and removes the head element. This ensures available slots on bus A for downstream nodes that made reservations. When a node wants to transmit, it places a tagged REQ (TREQ) duplicate in its FIFO queue (see Fig. 10b). Eventually the TREQ percolates to

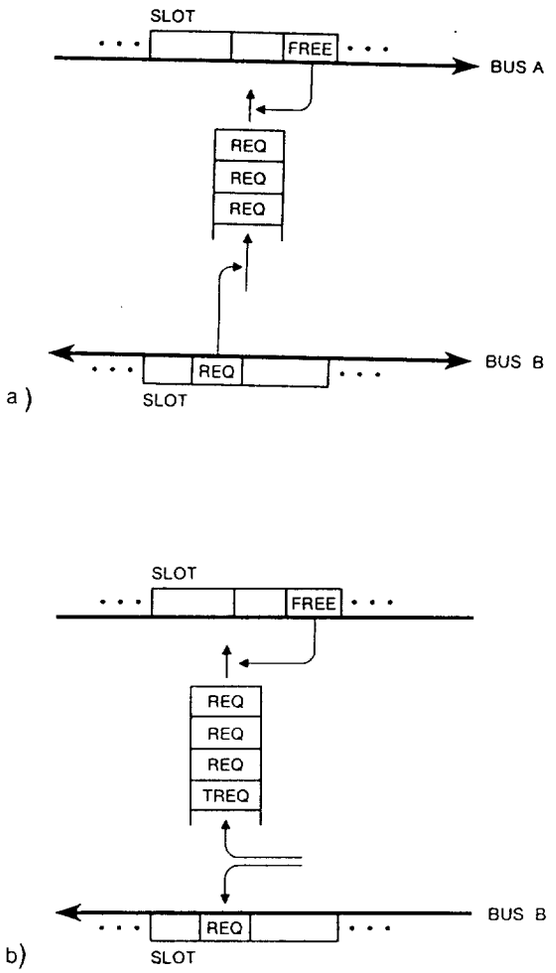

Fig. 10: Illustration of the DQDB distributed queue principle with the help of FIFO queues.

the queue head, thereby permitting the node to use the next free slot to pass on bus A. An identical FIFO mechanism controls data transmission on bus B via REQ reservations on bus A.

This represents a smart mechanism for distributing a first come, first served reservation discipline over a network of nodes. It works well as long as propagation delays between nodes are sufficiently small so that individual queues can be considered updated simultaneously. For larger network configurations, however, bus sections between consecutive nodes may be so long as to carry entire slots on the fly. If these slots carry REQs then the individual queues, belonging to consecutive nodes, are no longer consistent, which results in complex behavior. DQDB then exhibits unfair node throughput, unpredictable behavior and makes ineffective priority assignments during heavy loading and overloading.[25,26] Generally speaking, the unfairness becomes more pronounced the larger the number of slots on the fly in the network medium, i.e. the larger the product of the speed and distance of the bus. This proved to be so severe[27] that DQDB was extended to include a non-mandatory bandwidth balancing mechanism,[28] whose performance still exhibits fairness and priority inadequacies.[29,30] Although bandwidth balancing is currently the favored correction in the IEEE 802.6 MAN standards committee, other extensions have been proposed as well.[24,31]

6. Beyond DQDB: A Conclusion

Although DQDB has serious deficiencies (see Section 5) it maintains high aggregate throughput performance even at high bus speed and over large distances. This is because every slot carries payload (be it unfairly distributed) when the network saturates. One can conclude that the slotted bus nature of DQDB should be maintained when going to Gbit/s, but that a new reservation approach is required which operates efficiently when most needed, namely during heavy loading and overloading.

What requirements and properties, then, does a slot reservation scheme need to operate effectively in the gigabit speed range? Obviously, it must have predictable behavior, exhibit throughput fairness independent of node location and be expandable to support a traffic priority hierarchy. A reservation scheme should also work well whether the bus latency corresponds to 100, 1000 or 10000 slots. Thus a viable reservation scheme can be expected to be effective over large ranges of bus lengths and hence be applicable to both LAN and MAN networks. Another requirement is a high network utilization capability for all active user combinations while maintaining throughput and delay fairness. The extreme case of a single heavy user in an otherwise lightly loaded network must also be included for it represents the scenario of a powerful host or supercomputer attached to a LAN/MAN via an HPPI-like interface.

CRMA (Cyclic-Reservation Multiple-Access) is a slotted bus-based shared medium access protocol[24,32,33] with a promising slot reservation scheme that fulfills the above requirements. The principles of this protocol are illustrated in Fig. 11 for a folded bus structure but it operates equally well for dual bus configurations.[33] Numbered reservation commands periodically issued by the headend collect requests for slots as they pass nodes wanting to transmit. The nodes increment a cycle-length field by the number of slots they need. On returning, a command's cycle-length field plus number (cycle number) is entered in a headend FIFO queue. Entries percolate through the FIFO with the top element being removed to generate as many slots as its length number indicates. This group of slots form a cycle, preceded by a cycle start command carrying the appropriate cycle number. On matching a passing cycle number with one associated with a previous reservation, a node begins transmitting in the first free slot it observes in that cycle.

CRMA's cyclic nature features great flexibility in bandwidth allocation. Fairness can be achieved by imposing a bound or window on the number of slots reservable in one command. Its behavior under both light and heavy loads is predictable in a straightforward, intuitive way, and it can support prioritized traffic by superimposing multiple instances of the reservation scheme onto the same bus. It achieves maximum network utilization for all active node

configurations by offering more reservation than payload capacity, i.e. the rate of payload reservation is higher than the payload transport rate of the network. Yet the scheme minimizes access delay for light users in the presence of heavy loading in a simple way by temporarily stopping the reservation process when the backlog suffices to ensure full aggregate network utilization.[32,33]

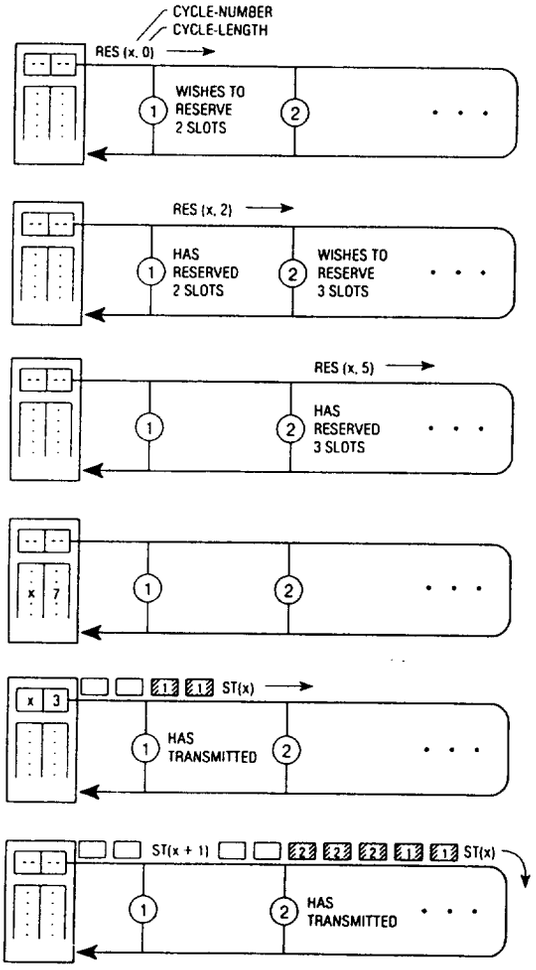

Fig. 11: Principle of CRMA as it applies to folded bus configurations.

Acknowledgment

I would like to acknowledge the valuable help of my co-workers, P. Heinzmann, H.R. Mueller, M.M. Nassehi, H.R. van As, K.T. Wilson, J.W. Wong, E.A. Zurfluh, W. Bux and F. Closs, without whom this paper could not have been written.

REFERENCES

1. "Special Report: Visualization in Scientific Computing – a Synopsis," IEEE Computer Graphics and Applications, Vol. 7, No. 7 (1987) pp. 61-70.
2. A. Albanese, M.W. Garret, A. Ippoliti, H. Izadpanah, M.A. Karr, M. Maszczak, D. Shia, "Bellcore Metrocore Network – A Test-Bed For Metropolitan Area Network Research," Globecom '88, Hollywood, FL, Nov. 28, 1988, pp. 1229-1234.
3. G. Watson, S. Ooi, "A Prototype Gbit/s Network," Third IEEE Workshop on Metropolitan Area Networks, San Diego, CA, March 28-30, 1989.
4. D.J. Greeves, D. Lioupis, A. Hopper, "The Cambridge Backbone Ring," Preprint, INFOCOM Conference Proceedings, San Francisco, CA, June 5-7, 1990.
5. "The Ultranetwork System," Ultranetwork Technologies Documentation, Ultranetwork Technologies, San Jose, CA.
6. Esprit Project N.2100: "MAX – Metropolitan Area Communications System Draft," Esprit Work Program, July 1987.
7. American National Standard Draft Proposal, "High Speed Channel, Mechanical, Electrical and Signalling Protocol Requirements," ANSI X3T9.3, June 1, 1989.
8. W.R. Hawe, M.F. Kempf, A.J. Kirry, "The Extended Local Area Network Architecture and LAN Bridge 100," Digital Technical Journal, No. 3, Sep. 1986, pp. 54-72.
9. ANSI/IEEE Std 802.4-1985, ISO 8802/4, "Token Passing Bus Access Method and Physical Layer Specifications," IEEE 1985.
10. W. Bux, F.H. Closs, K. Kuemmerle, H.J. Keller, H.R. Mueller, "Architecture and Design of a Reliable Token Ring Network," IEEE J. Select. Areas Commun., Vol. SAC-1, No. 5 (1983) pp. 756-765.
11. R.D. Dixon, N.C. Strole, J.D. Markov, "A Token-Ring Network for Local Data Communications," IBM Systems Journal, Vol. 22, Nos. 1/2 (1983) pp. 47-62.
12. ANSI/IEEE Std 802.5-1985, ISO 8802/5, "Token Ring Access Method and Physical Layer Specifications," IEEE 1985.
13. W. Bux, "Performance Issues in Local Area Networks," IBM Systems Journal, Vol. 23, No. 4 (1984) pp. 351-374.
14. N.C. Strole, "Enhancement to the Token-Ring Proposal: Early Token Release," Contribution 802.5-86-22 to IEEE Token Ring Standards meeting, San Diego, CA, Nov. 17, 1986.
15. N.C. Strole, "Inside Token Ring Version 2," Data Communications, January 1989, pp. 117-125.
16. M.D. Dias, A. Goyal, "A High Speed Token Ring Network," EFOC/LAN Conference Proceedings, Amsterdam, The Netherlands, June 29 - July 1, 1988, pp. 403-407.
17. R.M. Grow, "A Timed Token Protocol for Local Area Networks," Electro/82, Token Access Protocols (17/3), May 1982.

18. American National Standard, "FDDI Token Ring Media Access Control," ANSI, X3T9.5, Feb. 1986.
19. W. Stallings, "Local Area Networks: An Introduction," New York: Macmillan Publishing Company, 1984.
20. J.O. Limb, C. Flores, "Description of Fastnet — A Unidirectional Local Area Communications Network," Advances in Local Area Networks, IEEE Press 1987, pp. 190 - 205.
21. G.H. Clapp, "Broadband ISDN and Metropolitan Area Networks," Globecom Conference Record, Tokyo, Japan, Nov. 15-18, 1987, pp. 51.7.1-51.7.6.
22. R.M. Newmann, J.L. Hullett, "Distributed Queueing: A Fast and Efficient Packet Access Protocol for QPSX," Proc. Int'l Conf. on Computer Communications, Munich, Fed. Rep. Germany, Oct. 1986, pp. 294-299.
23. Proposed Standard: "DQDB Metropolitan Area Networks," IEEE unapproved drafts P802.6/D7 May 1989 to P802.6/D9 Aug 1989.
24. H.R. Mueller, M.M. Nassehi, J.W. Wong, E. Zurfluh, W. Bux, P. Zafiropulo, "DQMA and CRMA: New Access Schemes for Gbit/s LANs and MANs," to appear in INFOCOM Conf. Proc., San Francisco, CA, June 5-7, 1990.
25. J.W. Wong, "Throughput of DQDB Networks under Heavy Load," EFOC/LAN Conference Proceedings, Amsterdam, The Netherlands, June 14-16, 1989, pp. 146-151.
26. H.R. van As, J.W. Wong, P. Zafiropulo, "Fairness, Priority and Predictability of the DQDB MAC Protocol under Heavy Load," Int'l Zurich Seminar, Conference Proceedings, Zurich, Switzerland, March 5-8, 1990.
27. H.R. van As, J.W. Wong, P. Zafiropulo, "QA DQDB Analysis: Fairness, Predictability and Priority," IEEE 802.6-89/49 contribution, Montreal, Canada, Sep. 25-29, 1989.
28. E.L. Hahne, N.F. Maxemchuk, A.K. Choudhury, "Improving DQDB Fairness," IEEE 802.6-89/52 contribution, Montreal, Canada, Sep. 25-29, 1989.
29. M. Spratt, "A Problem in the Multi-Priority Implementation of the Bandwidth Balancing Mechanism," IEEE 802.6-89/61 contribution, Fort Lauderdale, FL, Nov. 6-10, 1989.
30. H.R. van As, "Performance Evaluation of Bandwidth Balancing in the DQDB MAC Protocol," to appear in EFOC/LAN Conference Proceedings, Munich, Fed. Rep. Germany, June 27-29, 1990.
31. J. Filipiak, "Access Protection for Fairness in a Distributed Queue Dual Bus MAN," ICC '89, Boston, MA, June 1989, pp. 635-639.
32. M.M. Nassehi, "CRMA: an Access Scheme for High-Speed LANs and MANs," SUPERCOM/ICC Conference Proceedings, Atlanta, GA, April 16-19, 1990.
33. M.M. Nassehi, "Cyclic-Reservation Multiple-Access Scheme for Gbit/s LANs and MANs Based on Dual-Bus Configuration," to appear in EFOC/LAN Conference Proceedings Munich, Fed. Rep. Germany, June 27-29, 1990.