Determinism in Audio Computer Control Network Applications

January 1995

Introduction

A control network is a system of sensors, actuators, displays, and logging devices (referred to as "nodes") that are linked together to monitor and control electrical devices. Supervisory functions are typically handled automatically and require no manual intervention except to respond to faults that the system cannot itself correct. In audio applications, a control network may monitor equipment fault conditions, adjust gain and crossover settings, monitor MIDI inputs, transport SMPTE time codes, sequentially turn equipment on or off, control special effects equipment, and display network status on a computer or custom status display.

The list of criteria that are typically considered essential in an audio control network include:

- Predictable response times
- Inexpensive node cost
- Compatibility with existing systems
- Control architecture
- Connectionless data transfer
- Pre-setable environments
- Fault tolerance
- Network-wide synchronization
- Predictable network latency
- Multiple media in the network
- Peer-to-peer architecture
- Inexpensive development cost
- Open architecture
- Application Programming Interface (API)
- Connection-based topology
- Object orientation
- Efficiency
- Multicast data channels
- Consistency with the OSI model
- Open availability and support

The first item on the list, predictable response times, stands out because it is often confused with deterministic behavior. An audio control network needs predictable response time, especially during periods of network overload. Deterministic behavior actually interferes with the ability of an audio computer control network to operate with predictable response times, though the reasoning behind this statement may not be obvious.

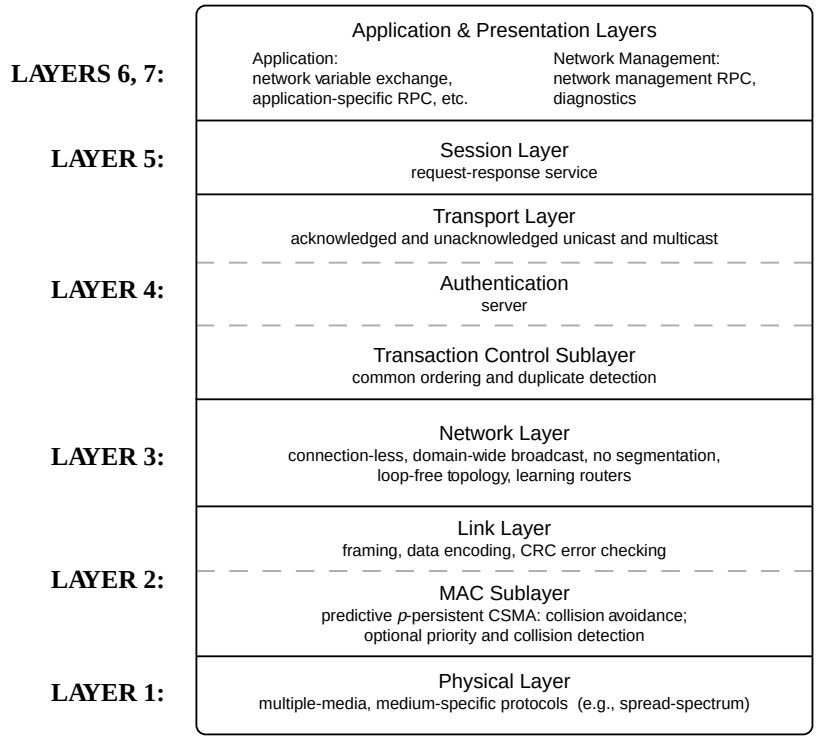
This paper will deal with the importance of response time, and likewise the unimportance of determinism, as an element in an audio control network.

We will begin with a review of determinism, and then review different protocols to assess their level of determinism. We will then review what factors are critical to the operation of a control network under conditions of overload, and examine why many factors other than determinism are necessary to ensure reliable performance under these conditions.

Determinism

By determinism, it is usually meant that access to the control network by a node may be delayed by at most some time *t*, where *t* is known. Every node thus has fair and equal access to the network since no one node will be delayed from gaining access for longer than time *t*. Using the OSI reference model shown in table 1 as a framework, a node's ability to access the network is a function of the Media Access Protocol (MAC), which is a sublayer of the Link layer known as OSI layer 2.

Table 1 OSI Reference Model



Proponents of determinism claim that knowing the value of *t* makes it much easier to design control networks, and that networked audio control systems

should only use deterministic protocols. The proponents also claim that the carrier sense multiple access (CSMA) family of protocols is non-deterministic and therefore not suitable for audio computer control applications. In reality, what happens above the MAC sublayer and at the application determines whether a control network will function or not. As an investigation of protocols will reveal, deterministic protocols actually undermine the performance of an audio control network under overload conditions, while properly implemented CSMA protocols are in fact ideal for audio control applications.

What Is A Deterministic Protocol?

There are four basic families of media access protocols: time division multiplexing (TDM), token bus, token ring, and CSMA. Of these four, all but CSMA are commonly considered deterministic.

TDM Media Access

TDM protocols assign a unique time for each node relative to a clock pulse (commonly called "start of frame") on the medium. Each node counts down to its time and transmits typically a single byte per frame. As long as all nodes have a unique time to transmit, and all nodes receive the same start of frame signal at the same time, the delay to access the network is bounded by the number of time slots assigned. For example, if there are 32 nodes and 32 slots each 1 byte wide, then each node may transmit a single byte for each 32 byte times on the medium, less the overhead for the start of frame signal.

The principle weaknesses of TDM protocols are that they waste bandwidth if all the nodes do not always have something to send, and a loss of synchronization to the start of frame signal results in the loss of all communications - a single point of failure condition. Furthermore, high levels of input/output activity at a node may cause more message traffic to be generated than can be handled by the assigned time slot. Under conditions of high traffic, system responses to input and output can vary considerably in time.

Token Passing Protocols

Token bus and token ring protocols, typified by ARCnet™ and IEEE 802.5, both pass a special message called the "token" which, upon receipt, grants the right

to transmit on the medium. Each node may only hold the token for a limited time before passing it on to the next node in the network. In this way, access to the network is bounded by the maximum latency of the token as it is passed to each of the nodes. In a token ring network, only the next station on the wire receives the token. In a token bus network, all stations receive the token and then must decide if it is intended for them.

Relying on topology to route the token message is an important advantage for a token ring because it eliminates the need for addressing information in the token message. Even more important, eliminating the need for addressing information also eliminates the need for each station on the network to know the address of the next station on the ring. This means that stations may enter and leave the ring without any configuration of the existing stations on the ring. On the other hand, reliance on topology to route the token precludes the use of communications media for which there can be no topological control, e.g., radio frequency and power line.

A weaknesses common to both of these token-based protocols is that the token may be lost due to error and then may be difficult to recover quickly. Most protocols use a random (non-deterministic) method to recover tokens; ARCnet is a typical example.

Errors may also result in the generation of multiple tokens. The presence of multiple tokens results in the loss of all tokens, and cause the network to initiate a token recovery process.

Knowing which node is the "next" node to pass the token is also a problem in token bus systems. If two nodes have the same address - each believes that it is the "next" node - then they will both acknowledge the token and cause a temporary loss of communications. This communication outage will repeat indefinitely and at regular intervals until the error is corrected.

One final problem occurs when nodes enter or leave the token bus. When this occurs the entire bus must automatically reconfigure itself in order to determine the sequence of node addresses on the bus and allow orderly token passing to resume. The time required for reconfiguration is proportional to the number of nodes on the bus, so if the failure mode is one where a node comes and goes due to, say, a watchdog timer reset, then the bus will effectively go down due to continuous reconfigurations.

CSMA-based Protocols

CSMA is a listen-before-transmit scheme in which a node with a message to transmit first listens to the network. If no message traffic is detected, evidenced by the absence of a carrier signal, then the node will transmit. Unlike the other media access protocols, the CSMA protocol family has many variants. The best known form of the CSMA protocol is Ethernet, also known IEEE 802.3. Ethernet was not designed for control systems and exhibits very poor characteristics near network overload; for this reason it is not often used for control systems. The limitations of Ethernet have created the impression that CSMA protocols are not suitable for computer control applications, though at least one variation of CSMA is ideally suited for such an application.

CSMA is fully deterministic when used in master-slave operation, however, it cannot be deterministic in peer-to-peer operation because nodes are not provided with equal access to the network. Nodes transmit on the basis of their ability to resolve packet collisions, and sometimes on the basis of priority messaging as well, and it is precisely these features which make a properly implemented, non-deterministic CSMA protocol so well suited for control applications.

Since CSMA was invented, there has been a great deal of research focused on modifying the initial protocol to make it perform better near network saturation. These efforts have been successful, as exemplified by non-persistent CSMA protocols (a node waits for a random period of time before checking if a busy channel is free to transmit) and p-persistent CSMA protocols (a node uses probability calculations to determine when and when not to transmit on slotted channels). For example, Echelon's LonTalk® media access protocol uses a predictive p-persistent CSMA protocol - a variant of the p-persistent CSMA protocol - to dynamically adjust the number of packet time slots based on predicted network traffic. By dynamically allocating network bandwidth, the predictive p-persistent CSMA protocol permits the network to continue operating in the presence of very high levels of network traffic without slowing the network during periods of light traffic. The benefits of this technology are its high efficiency, low overhead, low cost hardware, elimination of the need for network wide synchronization, and lack of loss-prone tokens.

CSMA protocols have been criticized because it is *theoretically* possible that a node could be prohibited from successful media access for an unbounded

time due to unresolved collisions. This theoretical result fails the test of determinism because colliding nodes may experience considerable delays accessing the network. If collisions could be resolved with CSMA protocols, so the argument goes, then 100% network utilization could be achieved. Transceivers have in fact been designed for the predictive p-persistent CSMA protocol which resolve collisions; Echelon's power line carrier transceivers are examples of such designs, and since 1992 have been widely used in thousands of industrial control, building management, and home automation applications that require collision resolution. When the predictive p-persistent CSMA protocol is coupled with a transceiver that implements collision resolution, the protocol operates with no packet loses due to collisions. Such a node will achieve successful media access since it will resolve collisions. The predictive p-persistent CSMA protocol has proven its robustness in field applications, and since 1992, Motorola and Toshiba have sold several hundred thousand Neuron® Chips embedded with Echelon's predictive p-persistent CSMA protocol. These firms are currently designing third generations of this family of chips using 0.6 micron geometry.

The predictive p-persistent CSMA protocol also overcomes the issue of unsuccessful media access by employing transceiver designs that limit the number of stations on a single network segment. In addition, each node is limited to a single outgoing transaction at a time; transmitters stop and wait for an acknowledgment prior to accessing the communications medium again. These two implementation details overcome a key limitation of other CSMA protocols by making it impossible for a working station to be denied access to the communications medium indefinitely.

The predictive p-persistent CSMA protocol is not the only non-deterministic CSMA protocol to offer robust performance. Various automotive protocols, such as Chrysler's Carlink™, use a special encoding scheme to eliminate collisions by resolving them in favor of a single message getting through. As a packet is transmitted, arbitration for access to the media is computed for each bit transmitted. The encoding scheme permits stations to monitor the line during the transmissions of bits corresponding to logic level '1' and not during transmission of logic level '0,' or vise versa. In this way, access to the network is arbitrated a bit at a time with the zero bits dominating the one bits. Thus stations transmitting a 1 bit when another is transmitting a 0 see the other's transmission and stop their own transmission. It should be noted that this protocol does not include fair access. The MAC delay for a packet which has a bit pattern which always loses in the arbitration is unbounded, and this problem must be handled at higher levels in the system.

Network Overload

The proponents of deterministic protocols claim that having determinism makes it much easier to design a networked control system. This assertion requires closer scrutiny.

When a network is lightly loaded (and this should be its condition in normal operation), response times will be good with both deterministic and non-deterministic systems. The reason that light loading should be the condition for normal operation is because the traffic is usually not uniformly distributed over time. Instead it arrives in bursts, and these bursts should not exceed the capacity of a network for very long.

When a network is heavily loaded, as is often the case during periods of emergency or error, response times will be poor with both deterministic and non-deterministic systems. This is because the offered traffic exceeds the bandwidth of the network and messages queue within nodes while awaiting access to the medium. If the overload persists, the nodes may run out of buffer memory to queue additional messages. This will either stop or reset the application in the node, causing further delays. The property of determinism only exacerbates delays during overloads since determinism includes "fair and equal access" to the network as its central feature. Allowing equal access to both critical and non-critical packets is simply not appropriate during emergencies. When a microphone is dropped and results in feedback howl, or an amplifier starts to smoke, or dynamic equalization must be performed to avoid phasing effects, equal access interferes with the real task at hand - namely to send high priority packets to correct the problem ASAP. At these times, offering unbounded delays for non-priority messages actually conserves network bandwidth for critical functions. Thus a network which supports prioritized access and is therefore not deterministic is a better choice for control applications, provided that the non-deterministic protocol also offers high network utilization.

All token passing schemes offer linearly increasing network delays up to a saturation point, but most CSMA protocols do not. The predictive p-persistent CSMA protocol does offer this feature, and represents a distinct improvement over, for example, Ethernet's 1-persistent CSMA approach.

In the case of overload conditions, there are three features which are essential if a protocol is to be used for a control system:

1. Graceful degradation: some number of messages must get through regardless of the offered traffic load. During an overload condition, neither determinism or CSMA will make the system work because the network will not be fast enough to carry all of the messages, however, the system must respond in a controlled manner to such a condition without failing catastrophically;

2. Priority: not all messages will get through in time during overload so it is essential that emergency messages have priority over other messages. Priority messaging works better than token passing in situations of overload because, unlike token passing protocols, high priority messages lock-out low priority messages. This has the effect of dedicating the network bandwidth to emergency traffic and holding off messages that the customer has decided can be deferred. Token passing schemes do not have this important characteristic and allow equal network access to all messages;

3. End-to-end acknowledgment: it is vital for an application either to know that a message got to its destination or that it did not get there within the real time requirements of the system.

Note that determinism is not a necessary requirement. During periods of overload (and graceful degradation), all of the system's real time requirements may not be met. The lack of acknowledgment is a way to detect this failure and to initiate the priority messages to save the process under control.

TDM and token protocols lack these features, even though some of these protocols claim deterministic behavior. The predictive p-persistent CSMA protocol has all three of these features even without a collision resolving communications transceiver. The reason for this situation is simple: the predictive p-persistent CSMA protocol was designed specifically for control network applications, while the other protocols emerged from data transmission network applications. Since control applications must continue operating reliably and/or degrade gracefully during overload conditions, the predictive p-persistent CSMA protocol was tailored to such a mode of operation. Data networks, on the other hand, are optimized for hauling large data files, and do not have the same response time or overload requirements.

One might question whether a predictive p-persistent CSMA protocol can provide well defined network timing given that the number of timing slots vary dynamically. Using the end-to-end acknowledgment service within a predictive p-persistent CSMA protocol (as typified by the LonTalk protocol) allows the application to know whether an operation has succeeded or not within a bounded amount of time. The protocol tracks elapsed time from the point that an application requests that a packet be sent, rather than when the packet is actually sent on the communications medium. For example, suppose that a packet must be sent and acted upon within 50 milliseconds, and if it is not acted upon in that time, the sender of the packet must take immediate action. Such a packet could be sent using acknowledged service with a retry count of 2 and a retry interval of 16 milliseconds. In this way, the application will either know that the transaction completed successfully by receiving the acknowledgment, or the application will know that the transaction failed in 48 milliseconds.

If the sample transaction described above fails, the application might then send an emergency message using the priority feature of the protocol. The priority feature uses separate buffers within each node to allow outgoing priority packets to precede non-priority packets which have already been queued for transmission. Additionally, the priority feature uses dedicated bandwidth (also referred to as "priority slots") at the end of each packet to eliminate contention for the communications medium after the transmission of a packet. Collision resolving transceivers can also be used when the channel bandwidth is limited and/or there is a need to run the network at its maximum capacity for a sustained time.

Conclusion

A control network must monitor and control electrical devices during both normal and emergency conditions. The ability of a control network to function during emergencies, including periods of high network traffic, is dependent on the network's ability to allocate bandwidth to important messages. Should a network become saturated beyond its capability or experience a fault condition, a graceful failure mechanism should allow for fast recovery without catastrophic failure. The predictive p-persistent CSMA protocol performs well under both normal and overload/fault conditions, and provides the predictable response times needed by an audio control network. The predictive p-persistent CSMA protocol offers significant

performance advantages over TDM, token ring, and token bus protocols under similar circumstances.

While the predictive p-persistent CSMA protocol can be used in a deterministic manner, determinism is neither necessary or sufficient to ensure predictable response times under overload or fault conditions. Yet, these are precisely the circumstances in which an audio computer control network needs to operate predictably; an amplifier fault or CPU failure during a live concert - where revenue and reputations are at stake - is when a user relies most heavily on the robustness of the control network. The predictive p-persistent CSMA protocol works robustly at the edge of the network's capability, a critical point at which deterministic behavior matters not at all.

Echelon Corporation
4015 Miranda Avenue
Palo Alto, CA 94304
Telephone (415) 855-7400
Fax (415) 856-6153

Echelon Europe Ltd
Elsinore House
77 Fulham Palace Road
London W6 8JA
England
Telephone +44-81-563-7077
Fax +44-81-563-7055

Part Number 005-0052-01A

Echelon Japan K.K.
Kamino Shoji Bldg. 8F
25-13 Higashi-Gotanda 1-chome
Shinagawa-ku, Tokyo 141
Telephone (03) 3440-7781
Fax (03) 3440-7782