January 2023

# High Precision Programmable Delay in Networking Switches

n/a

# High Precision Programmable Delay in Networking Switches

ABSTRACT

Fairness in computer networks used in application domains such as financial markets, prediction markets, etc., can be achieved by ensuring that all participants receive the same multicast information at the same time. However, in practice, due to factors such as differing processing delays at network nodes, differing optical fiber lengths, etc., some participants may receive data sooner than others. This disclosure describes techniques to reduce timing variance and unfairness in multicast networks. Since multicasting works by replicating information to destination nodes, the replication engine at a node of a given network layer is programmed to delay retransmission of information received by it until all nodes at that layer have received the information. All switches at a given network layer send their replications to the next layer simultaneously, such that timing variance throughout the network is reduced and network fairness is improved.

KEYWORDS

- Multicast
- Data replication
- Timing fairness
- Network fairness
- Colocation
- Colo
- High-frequency trading (HFT)

## BACKGROUND

Fairness in computer networks used in application domains such as financial markets, prediction markets, etc., can be achieved by ensuring that all participants receive the same multicast information at the same time. However, multicasting works by replicating the same information to destination nodes. Since replication requires a certain amount of time, multicasting as practiced today has structural unfairness in timing such that some participants may receive the data sooner than others. Other sources of timing unfairness exist, e.g., differing cable lengths.

Current techniques to reduce timing unfairness amount to physically measuring timing differences and changing cable lengths to adjust latency on certain network paths. This is not always possible, and even when possible, laborious, unscalable, and brittle. Industry agreements to limit timing variance across market participants can be difficult to enforce, especially as networks grow big, and especially with recent developments, where even infinitesimally small advantages in timing can be transformed to substantial profits. Reduction in timing variance and scaling of the network beyond a certain point have thus far proven almost mutually incompatible. Market participants often detect or suspect large variances in timing, leading to complaints, dissatisfaction, and reduction in trust.

## DESCRIPTION

This disclosure describes software-based techniques to address the structural unfairness in timing that can arise from multicast by replication. The techniques also address other sources of structural unfairness in timing, e.g., originating in cables of differing lengths. The replication engine at a node (e.g., switch or router) of a given network layer is programmed to delay retransmission of multicast information it receives until all nodes at that layer have received the

information. The variance in delay at each layer of the network is reduced to the point that all switches at the layer send their replications to the next layer at the same time.

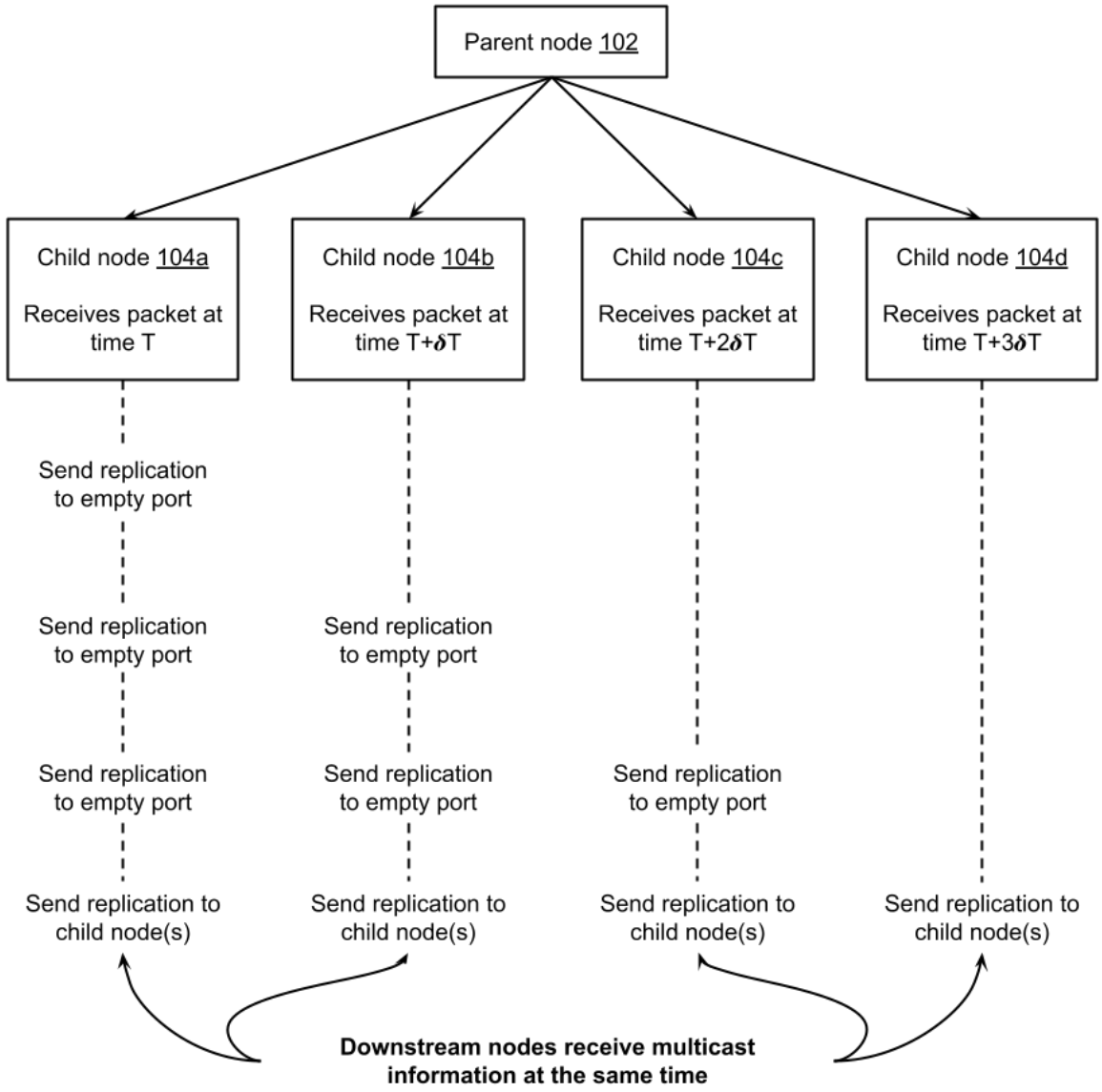


**Fig. 1: Reducing timing unfairness arising from multicast by replication**

Fig. 1 illustrates reducing timing unfairness arising from multicast by replication. A parent node (102) multicasts information to child nodes (104a-d) by replicating information at the parent node. In this example, there are four child nodes, such that the parent node replicates information four times, once for each child node. Replication takes a time $\delta T$, e.g., one

nanosecond, such that if the first child node 104a to receive the information receives it at a time $T$, the second child 104b node to receive the information receives it at time $T+\boldsymbol{\delta}T$, the third child node 104c to receive the information receives it at a time $T+2\boldsymbol{\delta}T$, and the fourth child node 104d to receive the information receives it at a time $T+3\boldsymbol{\delta}T$.

Each child node in turn multicasts its received information to its child nodes, such that the information eventually propagates throughout the entire network. To account for the relative delay in receiving information at individual child nodes, a child node in a given network layer that receives information from a parent performs as many null replications (replications to an empty port) as necessary to enable the first and last receiving child nodes in that layer to simultaneously multicast information to their child nodes in the following layer.

For example, node 104a is the first child node that receives information from its parent. Node 104a sends three replications to an empty port before replicating to its child nodes. The second child node 104b sends two replications to an empty port before replicating to its child nodes. The third child node 104c sends one replication to an empty port before replicating to its child nodes. The fourth child node 104d sends no replication to an empty port before replicating to its child nodes. By thus coordinating their multicast replication to the next layer, nodes of the next layer receive multicast information simultaneously (or with smaller delays), thereby reducing timing variance and unfairness throughout the network.

Here, the assumption is that the time-to-replicate is the same for the parent and the child nodes. If there are differences in the time-to-replicate between the parent and the child nodes, the number of replications to empty port done by the child nodes can be adjusted such that nodes in the layer that follow the child nodes still get the information simultaneously. For example, if the parent node takes two nanoseconds per replication while the child node takes one nanosecond

per replication, then in the topology of Fig. 1, child node 104a performs six replications to empty port before replicating to its child node; child node 104b performs four replications to empty port before replicating to its child node; etc.

The techniques described herein can also be used to balance structural variances in timing introduced by other sources, e.g., unequal cable lengths between switches. For example, if a certain node at a given layer is found to receive packets later than other nodes in the layer due to a longer leading cable, other nodes in that layer can delay their multicast retransmissions to their child nodes such that all nodes in the following layer receive the information substantially simultaneously. As explained earlier, a delay can be introduced to a multicast retransmission by replicating to an empty port.

The described techniques can be used in any context where synchronization in data replication is important. For example, cloud computing service providers can use the described techniques for clients that utilize the service for financial or other applications that benefit from such replication.

## CONCLUSION

This disclosure describes techniques to reduce timing variance and unfairness in multicast networks. Since multicasting works by replicating information to destination nodes, the replication engine at a node of a given network layer is programmed to delay retransmission of information received by it until all nodes at that layer have received the information. All switches at a given network layer send their replications to the next layer simultaneously, such that timing variance throughout the network is reduced and network fairness is improved.