Delay-Optimal Quorum Consensus for Distibuted Systems

Ada Waichee Fu [1]
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Email: adafu@cs.cuhk.hk

October 31, 1997

Abstract

Given a set of nodes S, a *coterie* is a set of pairwise intersecting subsets of S. Each element in a coterie is called a *quorum*. Mutual exclusion in a distributed system can be achieved if each request is required to get consensus from a quorum of nodes. This technique of quorum consensus is also used for replicated distributed database systems, and *bicoteries* and *wr-coteries* have been defined to capture the requirements of read and write operations in user transactions. In this paper we are interested in finding coteries, bicoteries and wr-coteries with optimal communication delay. The protocols take into account the network topology. We design delay-optimal quorum consensus protocols for network topologies of trees, rings, and clustered networks.

Index Terms: mutual exclusion, quorum consensus, wide-area networks, distributed systems, replicated database systems, coteries, communication delay, network topology.

[1] This research was supported by the RGC (the Hong Kong Research Grants Council) grant UGC REF.CUHK 495/95E

1 Introduction

The use of coteries is one approach to mutual exclusion in distributed systems. In this approach, a coterie [11] is defined for a given network $G = \{V, E\}$, where V is the set of nodes and E is the set of edges which connects the nodes. A coterie is a set of subsets of V. Each set in a coterie is called a *quorum*. A process must obtain consensus from a quorum in the coterie before it can use a certain resource. Each quorum intersects with every other quorum in the coterie and so mutual exclusion can be enforced. For transaction management in replicated database systems, the idea of coterie has been extended to *bicoteries* and *wr-coteries*, which capture the requirements of read and write operations. Many protocols based on quorum consensus have been proposed ([1, 2, 4, 6, 7, 12, 15, 16, 20, 21, 23, 25]).

Availability is often used to evaluate a quorum consensus protocol. Some previous work has dealt with the problem of designing coteries with optimal availability ([22, 18, 19]). There has also been work to minimize the communication cost, and quorum size has been used as an estimation of the communication cost (e.g., [2]). However, not much work has been done on minimizing the communication delay, which can be an important factor for the system response time. In this paper we shall examine this problem.

In quorum consensus, since messages are sent to multiple nodes in the network in order to ensure consistency of the operations, the messaging can create a bottleneck in the response time. Minimizing communication delay is important in replicated distributed database systems, since locking is a common technique in concurrency control in transaction management and remote locks are typically held for the duration of communication. Minimizing the delays not only improves the response time for the transaction that issues the lock request, but also reduces the chance of blocking other transactions waiting for the same lock. Therefore, we are interested in finding quorum consensus protocols with minimal communication delay which take into account the network topology.

We want to minimize delay when no failure occurs in the network. This is justified if the network is highly available. We define the **virtual distance** between nodes a and b as the time required to send a message of a certain size from a to b or vice versa. We make

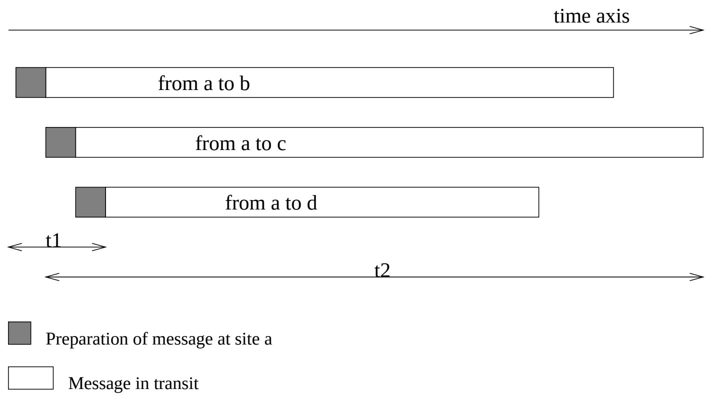


Figure 1: Message delivery from a to b, c, d

the assumption that the message propagation time from a to b is equal to the message propagation time from b to a. This value may be a linear function of the the physical distance in "surface" networks, while a satellite link can introduce a communication delay of about 300 milliseconds.

Suppose we are given a coterie for a connected network. Given an operation at a node s, any operational quorum in the coterie may be used. Node s can choose a quorum such that its virtual distance y from the furthest node in this quorum is minimized. We show here that this virtual distance y is important in measuring the communication delay in two important kinds of networks: long haul networks and local ring networks.

In the case of a ring, instead of sending one message to each receiver, it is possible to pass only one message in each of the clockwise and anti-clockwise directions around the ring to reach multiple receivers. We may consider that the time required to send a message to a set of nodes to be the virtual distance to the furthest node of the set.

For a long haul network, suppose a (logical) message is sent from a to a set of nodes b, c, and d (see Figure 1). Three physical messages may be delivered from a to b, c and d. If the virtual distances among the nodes are large enough, then the total preparation time, t_1, of the logical message at the sender node will be insignificant compared to the time the logical message spends in transit.

Let us consider some realistic timings for this example. Most of the figures in the following are from [13] and [14]. Suppose we have a wide-area network that connects a to b, c, d and that the maximum physical distance ac is around 3000 kilometers. With the speed of light (approximately 300,000 kilometers/second) the propagation delay for sending of a message for a distance of 3000 kilometer is about 10 milliseconds. However, if transmission is via satellite links, this delay may rise to 300 milliseconds. The available bandwidth contributes to the communications delay; in the 1990's it is around 100kb/s. For this, if the message consists of 100 bytes, the message transmission time is about 10 milliseconds. In the network, we may first send the message from a to a local gateway process, which then sends the message to a remote gateway process, which in turn passes the message on to the destination process. We estimate that at a there are 2,500 CPU instructions to be executed for the communication protocol, and among these 1,250 CPU instructions are executed for preparation of a message to c. There are 10,000 more instructions to be executed in the remaining path to c. Assuming CPU rate of about 10 mips, at a the CPU time needed for message sending to c will be 0.125 millisecond, and the CPU time needed down the road will be about 0.6 millisecond. The total communication delay for ac will be about 20.7 milliseconds. The round-trip delay for both sending a physical message and receiving a response will be about 41.4 milliseconds.

In the above, the local processing time at a is very small compared to the communication delay of the quorum consensus down the longest virtual path. The total local processing time may increase with the number of nodes in the quorum. However, we may set up the communication protocol to send the messages in the order of the virtual distances of the nodes, so that the message to the furthest node is sent first. In this way, the effect of the number of nodes in the quorum can be reduced. The bandwidth of advanced wide area networks is forecast to be 1Gb/s in the 2000's, and much faster CPU rate will become common. The dominating factor will then be the physical distance.

From the above example, it is seen that the dominating factor in the communication delay in rings and in long haul networks is the longest virtual distances t_2 among the paths of ab, ac and ad. Therefore, we use this distance in our definition of a delay-optimal coterie.

Another important type of network is a local area network which is not a ring. For such networks, the size of the quorum can sometimes be used to determine the communication delay, and much of the previously mentioned work has dealt with this metric. Other interesting metrics may include the sum of the virtual distances such as ab, ac and ad which in some cases gives the total bandwidth consumption, but it does not correspond to the message delay. Another possible metric for the communication cost for a node s to reach a quorum is the total weight of a weighted minimum spanning tree whose root is at node s and which spans each node in the quorum (see [26]).

Given a coterie, each node s chooses a quorum such that the virtual distance D_s to the furthest node in the quorum is minimal. The maximal value of D_s among all s will be an upper bound on the delay to reach a quorum for any node. The average value of D_s of all s will give us the average delay to reach a quorum for a node, provided that workload distribution is uniform. Given a network, we try to find a coterie that minimizes either the maximal value or the average value. Delay-optimal quorum consensus protocols for a network are protocols that minimize the delay under either metric.

We consider three types of networks: trees, rings and clustered networks. Tree structures are found in common telecommunication networks when the minimal length of cables is desired. Rings are common for optical networks. A clustered network is a model of the national or international networks, where local communication within a city involves much less delay than inter-city or international communication. We derive delay-optimal coteries for the above topologies.

In [11], the notion of non-domination (ND) of coteries is introduced. A coterie that is dominated can be improved in some sense. This notion has been extended to bicoteries and wr-coteries ([17]). We show that we can always search among ND coteries, bicoteries, wr-coteries for the delay-optimal coteries, bicoteries, or wr-coteries, respectively.

This paper is organized as follows: In Section 2, we give the definitions used in the paper and state the problem to be addressed. Sections 3, 4, 5 consider the tree, ring and clustered networks, respectively. Section 6 is a conclusion.

2 Preliminaries

We represent a network by an undirected graph and set the length of the edge that joins nodes a and b to be the virtual distance between a and b. Some definitions from [22] are adopted in this paper. Let $G = \{V, E\}$ be a network, where V is the set of nodes and E is the set of edges which connect the nodes. For $Q \subseteq V$, let $G|_Q = \{Q, E_Q\}$, where E_Q is the set of edges in G which connect nodes in Q. A set C of subsets of V is a **coterie** for G iff the following conditions hold:

1. *(Intersection)* $\forall Q_1, Q_2 \in C : Q_1 \cap Q_2 \neq \phi$.
2. *(Nonredundancy)* $\forall Q_1, Q_2 \in C$: $Q_1 \not\subseteq Q_2$.
3. *(Connectivity)* $\forall Q \in C$: $G|_Q$ is connected.

Each element in a coterie is called a **quorum**. The intersection property says that any two quorums in a coterie have at least one common node. The connectivity of the nodes in each quorum of a coterie has been considered in [3], [22] and [18]. In a network, if a node x is connected to node z only through node y, then we consider $\{x, y, z\}$ as a quorum rather than $\{x, z\}$. This is because a message between x and z must go through y. Hence we require that the nodes in each quorum be connected in the given graph.

Given a coterie C for a network $G = \{V, E\}$. Let $dist(a, b)$ be the shortest distance between two nodes a and b in G, the **delay** of node s in C, or $delay(s, C)$, is given by

$$delay(s, C) = \min_{Q \in C}\{\max_{v \in Q}\{dist(s, v)\}\}$$

We consider two different metrics of delays: the maximum of the delays among all nodes, and the arithmetic mean of delays of all nodes.

For a network $G = \{V, E\}$ the **max-delay** of a coterie C, or *max-delay*(C), is given by

$$max\text{-}delay(C) = \max_{s \in V}\{delay(s, C)\}$$

and the mean-delay of C is given by

$$mean\text{-}delay(C) = \frac{1}{|V|}\sum_{s\in V} delay(s, C)$$

Given the set CG of all coteries for G, a **max-delay optimal coterie** D is a coterie such that

$$max\text{-}delay(D) = \min_{C\in CG}\{max\text{-}delay(C)\}$$

and a **mean-delay optimal coterie** F is a coterie such that

$$mean\text{-}delay(F) = \min_{C\in CG}\{mean\text{-}delay(C)\}.$$

Let R and S be coteries for G. R **dominates** S iff $R \neq S$ and $\forall Q \in S\ \exists Q' \in R$: $Q' \subseteq Q$.

A coterie S for G is **dominated** iff there is a coterie for G which dominates S. If there is no such coterie, then S is **nondominated (ND)**.

Since an ND coterie cannot be enhanced in terms of communication cost or availability by simply replacing one of its elements, Q, by Q's subset, or by adding any other element, ND characterizes to some extent communication cost and availability optimality: some "optimal" coteries in terms of low communication cost or high availability can be found amongst the ND coteries. The complexity of the problem of deciding if a given coterie is ND is open.[2] We shall see that a max-delay or mean-delay optimal coterie or bicoterie can always be found among the ND coteries. Deciding that a given coterie is not ND rules out the possibility of the coterie being max-delay or mean-delay optimal. We believe that finding a max-delay or mean-delay optimal coterie, bicoterie or wr-coterie is not easy in the general case.

[2]This problem was related to the problem of self-dual positive functions in [17], and [8] developed a pseudo polynomial time algorithm for a problem equivalent to the self-duality problem.

2.1 Bicoteries and Wr-coteries

Here we consider the extension of the ideas of coteries to transaction management in distributed replicated database systems. We distinguish read quorums and write quorums for read operations and write operations, respectively. We make sure that a read quorum intersects all write quorums, and a write quorum intersects all other write quorums. For a discussion of consensus in replicated database see [5]. We have the following definitions [9, 17]:

Bicoterie: Given a network $G = \{V, E\}$, an ordered pair $B = (\alpha, \beta)$, where α, β are non-empty sets of subsets of V, is a bicoterie for G iff

1. $\forall Q \in \alpha \cup \beta$: $Q \neq \phi$.
2. (*Intersection Property*) $\forall Q_1 \in \alpha$, $Q_2 \in \beta$: $Q_1 \cap Q_2 \neq \phi$.
3. (*Nonredundancy*) $\forall Q_1, Q_2 \in \alpha$: $Q_1 \not\subseteq Q_2$, and $\forall Q_1$, $Q_2 \in \beta$: $Q_1 \not\subseteq Q_2$.
4. (*Connectivity*) $\forall Q \in \alpha \cup \beta$: $G|_Q$ is connected. $\Box$

Wr-coterie: A bicoterie $B = (W, R)$ for G is a wr-coterie for G iff
$\forall Q_1$, $Q_2 \in$ W: $Q_1 \cap Q_2 \neq \phi$ (i.e., W is a coterie). $\Box$

A bicoterie $A = (\alpha, \beta)$ is **dominated** by bicoterie $B = (R, S)$ (or B dominates A) iff the following conditions hold:
1. $\forall Q_1 \in \alpha\ \exists Q_2 \in R$: $Q_2 \subseteq Q_1$.
2. $\forall Q_1 \in \beta\ \exists Q_2 \in S$: $Q_2 \subseteq Q_1$.
3. $(\alpha, \beta) \neq (R, S)$. $\Box$

A bicoterie (wr-coterie) B is said to be **nondominated (ND)** if no bicoterie (wr-coterie) dominates B. $\Box$

A wr-coterie (W, R) can be used to form **read** and **write quorums** by considering each element of R and W as read and write quorum, respectively. A bicoterie (α, β) which is not a wr-coterie can also be used to form read and write quorums: let a read quorum be any element in α or any element in β, and a write quorum be the union of any element in α and any element in β. We thus form a wr-coterie from the bicoterie.

Recall that $dist(a,b)$ denotes the virtual distance between nodes a and b in a network $G = \{V, E\}$. Given a bicoterie $C = (WQ, RQ)$, the **read(write)-delay** for node s is given by

$$\begin{aligned} \textit{read-delay}(s,C) &= \min_{Q \in RQ}\{\max_{v \in Q}\{dist(s,v)\}\} \\ \textit{write-delay}(s,C) &= \min_{Q \in WQ}\{\max_{v \in Q}\{dist(s,v)\}\} \end{aligned}$$

The **delay** of node s in C and the **max-delay** of C are given by

$$\begin{aligned} \textit{delay}(s,C) &= \max\{\textit{read-delay(s,C)},\ \textit{write-delay}(\mathrm{s},\mathrm{C})\} \\ \textit{max-delay}(C) &= \max_{\mathrm{v}\in\mathrm{V}}\{\textit{delay}(\mathrm{v},\mathrm{C})\} \end{aligned}$$

If BC_G is the set of bicoteries for G, D is a **max-delay optimal bicoterie** for G iff

$$\textit{max-delay}(D) = \min_{C \in BC_G}\{\textit{max-delay}(C)\}$$

Suppose that the probability of read operations among read/write operations at s is known to be p. The mean delay of node s is then given by

$$\textit{mean-delay}(s,C) = p(\textit{read-delay}(s,C)) + (1-p)(\textit{write-delay}(s,C))$$

The arithmetic mean of the delays of all nodes is called the **mean-delay** of the bicoterie:

$$\textit{mean-delay}(C) = \frac{1}{|V|}\sum_{v \in V}\{\textit{mean-delay}(v,C)\}$$

If BC_G is the set of bicoteries for G, F is a **mean-delay optimal bicoterie** for G iff

$$\textit{mean-delay}(F) = \min_{C \in BC_G}\{\textit{mean-delay}(C)\}$$

The following lemma allows us to reduce the problem of finding a max-delay optimal wr-coterie to that of finding a max-delay optimal coterie.

Lemma 1 *For a given network, the max-delay of a max-delay optimal coterie is less than or equal to the max-delay of a max-delay optimal wr-coterie.*

Proof: If a max-delay optimal wr-coterie is $\{P, Q\}$, then we can form a coterie with P, and the max-delay of P is less than or equal to the max-delay of $\{P, Q\}$. ∎

From the above lemma, given a network, after finding a max-delay optimal coterie P, a max-delay optimal wr-coterie will be $\{P, P\}$.

From the following lemma, we know that to look for a delay optimal coterie (bicoterie), we can always search among the ND coteries (bicoteries).

Lemma 2 *For a given network, if a coterie/bicoterie/wr-coterie A is dominated by a coterie/bicoterie/wr-coterie B, then the max-delay (mean-delay) of B is less than or equal to that of A.*

Proof: Let A be dominated by B. Then, for any quorum Q_1 in A, there exists a corresponding quorum Q_2 in B, such that $Q_2 \subseteq Q_1$. For any node s, therefore, either $delay(s, A) \geq delay(s, B)$, if A and B are coteries; or *read-delay*$(s, A) \geq$ *read-delay*(s, B) and *write-delay*$(s, A) \geq$ *write-delay*(s, B), which implies the lemma. ∎

3 Trees

In this section, we consider networks that are of the form of a tree. In [3], it is shown in Theorem 4.7 that for an ND-coterie, each quorum contains only nodes from a single biconnected component. As a corollary,

Corollary 1 *For a network that is a tree, any ND-coterie consists of only one quorum which is a single node in the tree.*

To prove Corollary 1, one can extract part of the proof for Theorem 1 in [18], skipping over the arguments about availability.

Max-Delay Optimal Coteries

From Lemma 2, we can search among the ND coteries for max-delay optimal coteries. Since any ND-coterie for a tree consists of only one single-node quorum (from Corollary 1), one max-delay optimal coterie is the center of the tree (graph). The center of a graph $G = (V, E)$ is a node of minimum eccentricity, where the *eccentricity* of a node v is $\max_{w \in V}$ { minimum length of a path from w to v }.

Lemma 3 *A max-delay optimal coterie for a tree can be computed in $O(n)$ time.*

Proof: Efficient algorithms can be found in the references of [24], a survey paper on the p-center and p-median problems. For a tree network, the unweighted 1-center problem is equivalent to the problem of finding a max-delay optimal coterie. From results cited in [24], an algorithm is known that requires $O(n)$ time. ∎

From Lemma 1, the max-delay optimal coterie can be used to generate a max-delay optimal wr-coterie for the tree.

Mean-Delay Optimal Coteries

For a mean-delay optimal ND coterie, the single node in Corollary 1 must be one such that the arithmetic mean of the minimum distances from every other node is minimum.

Lemma 4 *A mean-delay optimal coterie for a tree can be computed in $O(n)$ time.*

Proof: The unweighted 1-median problem is equivalent to the problem of finding a mean-delay optimal coterie. From results cited in [24], an algorithm is known that requires $O(n)$ time. ∎

Mean-delay optimal wr-coterie depends on the ratio of read and write operations. In one extreme, the read-one-write-all approach can be used. An ND wr-coterie for a tree must contain only one write quorum; if there are two or more write quorums, the wr-coterie will be dominated by one which has the intersection of all these write quorums as the only write quorum.

In [22], a definition of optimal performance in the sense of availability is given as the maximal weighted expected number of working node, where a node is working if it can communicate with a quorum. It is found that for a tree where all edges have equal length and all nodes have equal weights, if the nodes are reliable (probability of failure for each node is small), then a coterie with optimal availability coincides with the mean-delay optimal coterie given above.

4 Rings

In this section, we consider the case where the network forms a ring. Since long haul networks are built on telecommunication networks which are seldom in the form of rings, we restrict our scope to local area ring networks (e.g. IBM LAN, FDDI rings). For such networks, if we assume that machines of comparable speeds are being used at each node, then the virtual distance between two adjacent nodes may be approximated by a constant. This is because the dominating factor in communication delay in a local area network is the CPU processing time [14]. Therefore, we consider rings where the nodes are evenly spaced, i.e., the virtual distance between any two adjacent nodes is a constant.

From [22] we have the following definition. Consider a ring with n nodes, let k be a positive integer where $1 < 2k+1 \leq n$. Choose $2k+1$ nodes $u_1, ..., u_{2k+1}$ on the ring in a clockwise order. For example in the ring in Figure 2(a), suppose $k = 2$, and let v_1, v_2, v_3, v_4, v_5 be five chosen nodes on the ring. There are $2k+1$ arcs on the ring which contain exactly $k+1$ of these nodes with two chosen nodes at the endpoints. The chosen nodes at the endpoints are called **end-nodes**. In the example, arcs $\langle v_1 v_2 v_3 \rangle$, $\langle v_2 v_3 v_4 \rangle$, $\langle v_3 v_4 v_5 \rangle$, $\langle v_4 v_5 v_1 \rangle$, $\langle v_5 v_1 v_2 \rangle$ [3], are 5 distinct arcs, each containing exactly 3 of the chosen

[3]Unless otherwise specified, if $x_1, x_2, ...x_n$ are nodes on the ring, an arc $\langle x_1 x_2 ... x_n \rangle$ refers to an arc with endpoints at x_1 and x_n and which passes through x_1, x_2, ... x_n in a clockwise order.

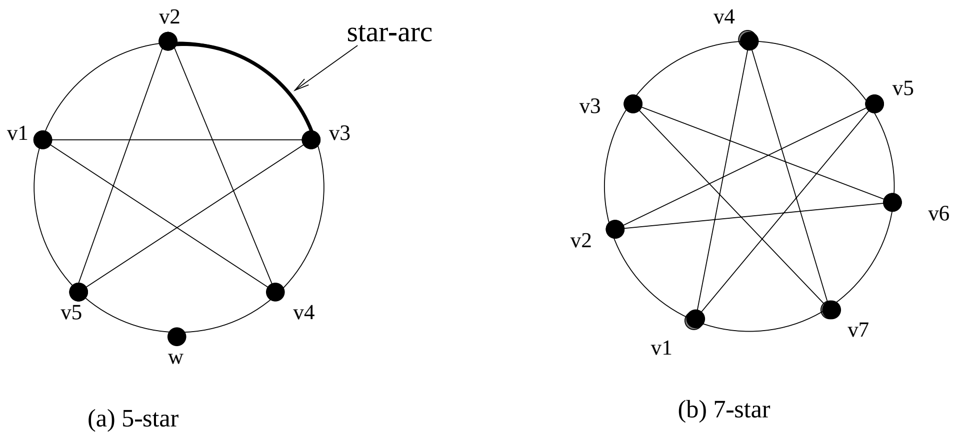


Figure 2: Star polygons

nodes and having two of these nodes as end-nodes. (v_1 and v_3 are end-nodes of arc $\langle v_1 v_2 v_3 \rangle$.) The coterie defined by the $2k+1$ sets of nodes which correspond to the nodes in the above $2k+1$ arcs is called a $(2k+1)$**-oligarchy**. Such a $(2k+1)$-oligarchy is also called an **odd oligarchy**. For the example in Figure 2(a), if v_1, v_2, v_3, v_4, v_5 are the only nodes on the ring, then there is only one 5-oligarchy: $\{\{v_1, v_2, v_3\}, \{v_2, v_3, v_4\}, \{v_3, v_4, v_5\}, \{v_4, v_5, v_1\}, \{v_5, v_1, v_2\}\}$.

Note that all the $2k+1$ chosen nodes are end-nodes of some of the $2k+1$ arcs in the above. We refer to the remaining nodes as **non-endpoint nodes**. The following theorem tells us that one can find optimal coteries from among the odd oligarchies.

Theorem 1 *(from the proof of Theorem 3 in [22]): An ND coterie for a ring network is always an odd oligarchy.*

As described in [22], the structure of an odd oligarchy can be represented by a *canonical odd star polygon*. A canonical odd star polygon has $2k+1$ nodes on a circle, and edges connecting every k-th node. Every edge of such a polygon intersects all other edges. Figure 2(a) shows a ring with 5 end-nodes, and the corresponding canonical odd star polygon. We can think of an edge xy (x and y are the two end-nodes of the edge) in the star as a chord on the circle (ring) which divides the circle into two arcs with overlapping end-nodes. The nodes on the arc that contains $k+1$ end-nodes form the quorum that

corresponds to the chord (edge). For example, in Figure 2(a), if v_1, v_2, v_3, v_4, v_5 are the only nodes on the ring, then the quorum corresponding to chords v_1v_3 and v_2v4 are $\{v_1, v_2, v_3\}$ and $\{v_2, v_3, v_4\}$, respectively.

Let us call each edge of an odd star polygon a **star-chord**. If we have a $(2k+1)$-star polygon (an odd star polygon with $2k+1$ nodes), where k is a positive integer, then the star polygon divides the ring into $2k+1$ arcs. In the 5-star in Figure 2(a), the 5 arcs are $\langle v_1v_2\rangle$, $\langle v_2v_3\rangle$, $\langle v_3v_4\rangle$, $\langle v_4v_5\rangle$, and $\langle v_5v_1\rangle$. Let us call these the **star-arcs**. If a set of star-arcs forms a single arc on the ring, we say that the star-arcs are **consecutive**. Let us call an arc formed by x consecutive star-arcs an x**-arc**. For example, $\langle v_1v_2v_3\rangle$ in Figure 2(a) is a 2-arc, formed by arcs $\langle v_1v_2\rangle$ and $\langle v_2v_3\rangle$. For a $(2k+1)$-oligarchy, some $(k+1)$-arc will have the greatest length, we call such an arc a **max-arc**.

4.1 Max-delay Optimal Coteries

In this section, we assume a ring with n nodes and with a circumference of θ, the distance between two adjacent nodes in the ring is $\frac{\theta}{n}$.

Theorem 2 *Given a ring with circumference θ, on which n nodes are evenly spaced, if $2k+1 \leq n$, a $(2k+1)$-oligarchy with minimal max-delay among all $(2k+1)$-oligarchies can be constructed in $O(k)$ time.*

Proof: Each quorum Q in an odd oligarchy corresponds to a star-chord z_1z_2. Suppose that for star-chord z_1z_2, the nodes in $\langle z_1z_2\rangle$ form a quorum Q. If a node exists at the midpoint of arc $\langle z_1z_2\rangle$, then we call this node x. Otherwise, there must exist two nodes closest to the midpoint of arc $\langle z_1z_2\rangle$, and without loss of generality, let x be the node closer to z_1. Node x can choose Q for minimum delay. It is because if x chooses any other quorum Q', there must be a node in the quorum Q' whose distance from x is greater than or equal to the greater of the clockwise distances z_1x or xz_2. The delay of x is thus the greater of the lengths of arcs $\langle z_1x\rangle$ and $\langle xz_2\rangle$. In fact, if z_3 is the next end-node outside $\langle z_1z_2\rangle$ and clockwisely next to z_2, then all of the nodes on the small arc clockwise from x to the midpoint y of the arc $\langle z_1z_2z_3\rangle$ will be able to choose Q as the optimal quorum. For example, in Figure 2(a), all the nodes from v_4 to w (where w is the midpoint of $\langle v_3v_4v_5v_1\rangle$) can choose the quorum of star-chord v_3v_5, and w has the maximum delay among them. Moreover, any node in the ring lies within an arc such as $\langle xy\rangle$ in the above.

From this we shall show that the max-delay of the coterie is determined by the maximum of the lengths of arcs such as $\langle z_1z_2z_3\rangle$ above. Any $(k+1)$-arc is a possible choice. One may distinguish the case where a node exists at the midpoint of some max-arc from the case where there is no node at the midpoint of any max-arc.

1. If a midpoint node exists for some max-arc $\langle gh\rangle$, the delay of the node is exactly half the length of $\langle gh\rangle$.
2. Otherwise, let $\langle gh\rangle$ be a max-arc. The delay at a node nearest to the midpoint of $\langle gh\rangle$ is the length of $\langle gh\rangle$ minus half the distance $\frac{\theta}{n}$, in which case, we need to show that it still gives us the max-delay. We need only consider the $(k+1)$-arc $\langle ef\rangle$ of the second longest length. Its length must be less than the length of gh by at least $\frac{\theta}{n}$, hence even if a midpoint node exists for $\langle ef\rangle$, its delay is the same as the delay for the near midpoint node for $\langle gh\rangle$.

Hence, the maximum length among all $(k+1)$-arcs, i.e., the length of a max-arc, will determine the max-delay. We would like to minimize this value. In the following the length of an arc $\langle x_1x_2\rangle$ on the ring is indicated by $|x_1x_2|$. The following lemma will be useful.

Lemma 5 *The sum of the lengths of all $(k+1)$-arcs for a $(2k+1)$-oligarchy is $(k+1)\theta$.*

Proof: In the summation, each star-arc appears in k+1 distinct $(k+1)$-arcs and is therefore counted $k+1$ times, we also know that the sum of the lengths of all star-arcs is θ. ■

For example for the 5-star in Figure 2(a) we have $(|v_1v_2|+|v_2v_3|+|v_3v_4|)+(|v_2v_3|+|v_3v_4|+|v_4v_5|)+(|v_3v_4|+|v_4v_5|+|v_5v_1|)+(|v_4v_5|+|v_5v_1|+|v_1v_2|)+(|v_5v_1|+|v_1v_2|+|v_2v_3|)=3\theta$.

Lemma 6 *If* $\frac{n}{2k+1} = m$ *and* $m \geq 1$*, then a max-arc of an* $(2k+1)$*-oligarchy has length at least* $\lceil (k+1)m \rceil \frac{\theta}{n}$*.*

Proof: Suppose on the contrary that the max-arc length is less than $\lceil (k+1)m \rceil \frac{\theta}{n}$, then all the $(k+1)$-arcs have length less than $\lceil (k+1)m \rceil \frac{\theta}{n}$. Summing up the lengths of the $2k+1$ $(k+1)$-arcs will give a value less than $(k+1)\theta$. From Lemma 5, we have a contradiction. ∎

Lemma 7 *If* $2k+1$ *divides* n *exactly, then a* $(2k+1)$*-oligarchy in which each star-arc has length* $\frac{n}{2k+1}$ *is a max-delay optimal* $(2k+1)$*-oligarchy.*

Proof: Let $\frac{n}{2k+1} = m$, where m is a positive integer. From Lemma 6, the length of a max-arc will be at least $(k+1)m(\frac{\theta}{n})$. From the sum of $(k+1)\theta$ in Lemma 5, a max-arc length of $(k+1)m(\frac{\theta}{n})$ is achieved when all $(k+1)$-arcs have length equal to $(k+1)(m)\frac{\theta}{n}$. This is the case when each star-arc has length m, since the difference in length between two $(k+1)$-arcs such as $(|v_1v_2| + |v_2v_3| + |v_3v_4|)$ and $(|v_2v_3| + |v_3v_4| + |v_4v_5|)$ in Figure 2(a) is zero and it means that $\langle v_1v_2 \rangle$ and $\langle v_4v_5 \rangle$ are equal in length. ∎

In the following let $m = \frac{n}{2k+1}$. If m is not an integer, then the following shows that the max-arc length of $\lceil (k+1)m \rceil \frac{\theta}{n}$ can also be achieved. We conjecture that some max-delay optimal coterie has at most two different lengths of $(k+1)$-arcs: $\lfloor (k+1)m \rfloor \frac{\theta}{n}$, and $\lceil (k+1)m \rceil \frac{\theta}{n}$.

Lemma 8 *If there exists an odd oligarchy where the difference between any two* $(k+1)$*-arcs is at most* $\frac{\theta}{n}$*, then the max-arc has a length of* $\lceil (k+1)m \rceil \frac{\theta}{n}$*.*

Proof: From Lemma 6, the max-arc has length at least $\lceil (k+1)m \rceil \frac{\theta}{n}$. If the length is greater than $\lceil (k+1)m \rceil \frac{\theta}{n}$, then the sum of all $(k+1)$-arcs will be greater than $(k+1)\theta$, in contradiction to Lemma 5. ∎

Lemma 9 *One can construct in* $O(k)$ *time a* $(2k+1)$*-oligarchy where the difference in* *length between any two j-arcs is at most $\frac{\theta}{n}$, for any integer j, $1 \leq j < n$, for any integer $k \geq 1$.*

Proof: Let us name the nodes on the ring clockwisely starting from any node by w_0, w_1, w_2, ... w_{n-1}. We create star-arcs clockwisely. The star-arcs created will be named S_0, S_1, S_2, ... S_{2k}. S_i contains the nodes that are clockwise from node $w_{\lfloor im \rfloor}$ to and including node $w_{\lfloor (i+1)m \rfloor}$.

First we show that in the resulting $(2k+1)$-oligarchy, the length of any j-arc is given by $(\lfloor (i+j)m \rfloor - \lfloor im \rfloor)\frac{\theta}{n}$ for some integer i, j where $0 \leq i, j \leq 2k+1$, and $j > 0$.

Let the first star-arc clockwisely in the j-arc be S_i. The above is obviously true if the j-arc does not contain w_0 as a non-endpoint node so that we have $i + j \leq 2k+1$ in the above expression. Otherwise, the j-arc will be divided into two subarcs by node w_0: $\langle w' w_0 \rangle$ and $\langle w_0 w'' \rangle$. The subarc $\langle w' w_0 \rangle$ contains $2k+1-i$ star-arcs. The length of $\langle w' w_0 \rangle$ is given by $(n - \lfloor im \rfloor)\frac{\theta}{n}$. The other subarc $\langle w_0 w'' \rangle$ will contain $j - (2k+1-i)$ star-arcs, and its length is given by $\lfloor (j - (2k+1-i))m \rfloor \frac{\theta}{n}$. Since $(2k+1)m = n$, summing the lengths of $\langle w' w_0 \rangle$ and $\langle w_0 w'' \rangle$ gives $(\lfloor (i+j)m \rfloor - \lfloor im \rfloor)\frac{\theta}{n}$

Hence the difference in length of any two j-arcs is given by

$$
\begin{aligned}
& (\lfloor (i+j)m \rfloor - \lfloor im \rfloor)\frac{\theta}{n} - (\lfloor (i'+j)m \rfloor - \lfloor i'm \rfloor)\frac{\theta}{n} \\
= \; & \{[(i+j)m - \delta_1] - [(i'+j)m - \delta_2] - ([im - \delta_3] - [i'm - \delta_4])\}\frac{\theta}{n} \\
= \; & [(\delta_2 - \delta_1) - (\delta_4 - \delta_3)]\frac{\theta}{n}
\end{aligned}
$$

for some integers i, i' where $i \neq i'$ and $0 \leq \delta_t < 1$ for $t \in 1, 2, 3, 4$.

Since $|\delta_2 - \delta_1|$ and $|\delta_4 - \delta_3|$ are both less than 1, the magnitude of the difference is less than $2\frac{\theta}{n}$, and since it is a multiple of $\frac{\theta}{n}$, it is at most $\frac{\theta}{n}$. The construction of $S_0, S_1, ..., S_k$ takes $O(k)$ time. ∎

From Lemma 8, the max-arc of the oligarchy constructed in the proof of Lemma 9 has a length of $\lceil (k+1)m \rceil \frac{\theta}{n}$, which is optimal from Lemma 6. Therefore, from Lemma 7 and

Lemma 9 we conclude that we can construct a $(2k+1)$-oligarchy with minimal max-delay among all $(2k+1)$-oligarchies in $O(k)$ time. ∎

Corollary 2 *If a ring has n nodes evenly spaced, then the max-delay for a $(2k+1)$-oligarchy which is max-delay optimal among all $(2k+1)$-oligarchies is determined as follows:*

If the midpoint of any max-arc contains a node, then the max-delay is given by

$$c_1(k) = \frac{1}{2}\left\{\left\lceil (k+1)\frac{n}{2k+1}\right\rceil \frac{\theta}{n}\right\}$$

If none of the midpoints of the max-arcs contains a node, then the max-delay is given by

$$c_2(k) = c_1(k) - \frac{1}{2}\left(\frac{\theta}{n}\right)$$

Proof: From the proof of Theorem 2, the length of a max-arc in a max-delay optimal coterie is $\left\lceil (k+1)\frac{n}{2k+1}\right\rceil \frac{\theta}{n}$. A node in the middle of a max-arc (if any) has a delay of half the length of a max-arc, which is the max-delay of the coterie. Hence we have the above expression for $c_1(k)$.

If none of the max-arcs has a node at the midpoint, then the node closest to the midpoint of a max-arc, which is at a distance of $\frac{\theta}{2n}$ from the midpoint, will have the longest delay of $c_1(k) - \frac{1}{2}\left(\frac{\theta}{n}\right)$. ∎

For example, in Figure 2(a), suppose that arc $\langle v_3 v_4 v_5 v_1\rangle$ is a 3-arc of maximal length, and the midpoint w contains a node. Then the max-delay is $c_1(k)$. If the midpoint w is not a node, then w must be of distance $\frac{\theta}{2n}$ to a node y, y can reach quorum $\{v_4, v_5, v_1\}$ or $\{v_3, v_4, v_5\}$ with a smaller delay than that for a node at w.

Lemma 10 *For a ring with n evenly spaced nodes, the midpoint of a max-arc contains a node if the number of nodes in the max-arc is odd. If a max-arc has an even number of* *nodes, then the midpoint does not contain a node. The number of nodes in the max-arc is given by* $v(k) = \left\lceil (k+1)\frac{n}{2k+1} \right\rceil + 1$.

From Lemma 10, for a ring with 5 nodes, the 5-oligarchy has no midpoint node in the max-arc; for a ring with 7 nodes, the 7-oligarchy has a midpoint node in the max-arc.

Theorem 3 *Given a ring with n evenly spaced nodes, a max-delay optimal coterie can be computed in $O(n)$ time.*

Proof: From Lemma 2, one can search for delay-optimal coteries from among the ND coteries. From Theorem 1, an ND coterie is always an odd oligarchy. Corollary 2 and Lemma 10 give us a way to determine the max-delay of a max-delay optimal $(2k+1)$-oligarchy given a ring with evenly spaced nodes. The problem then is to find the optimal value of k. We can deduce the delays of different values of k by first checking if $v(k)$ in Lemma 10 is even or odd, then apply $c_1(k)$ or $c_2(k)$ in Corollary 2 correspondingly. Hence the optimal value of k can be computed in $O(n)$ time. Finally, the procedure in the proof of Lemma 9 constructs an max-delay optimal $(2k+1)$-oligarchy in $O(k)$ time. ∎

It is found in [22] that an optimal coterie (in terms of availability defined therein) is always a *canonical* $(2k+1)$-oligarchy. The max-delay optimal coterie obtained above for the evenly spaced ring is also a canonical $(2k+1)$-oligarchy. We may compute both the delay and availability factors for different values of k to select an desirable k.

4.2 Mean-delay Optimal Coteries

In this section, we study the problem of finding mean-delay optimal coteries in a ring. To simplify our discussion, we assume that the distance between two adjacent nodes in the ring is normalized to 1. Since we consider rings with n evenly spaced nodes, the circumference of the ring is n.

For example, if $k = 3$, we have 7 end-nodes in the $(2k+1)$-oligarchy. Figure 3 shows a 7-oligarchy with star arcs $\hat{d}_1$, $\hat{d}_2$, $\hat{d}_3$, $\hat{d}_4$, $\hat{d}_5$, $\hat{d}_6$, $\hat{d}_7$. The length of $\hat{d}_i$ is given by d_i. In

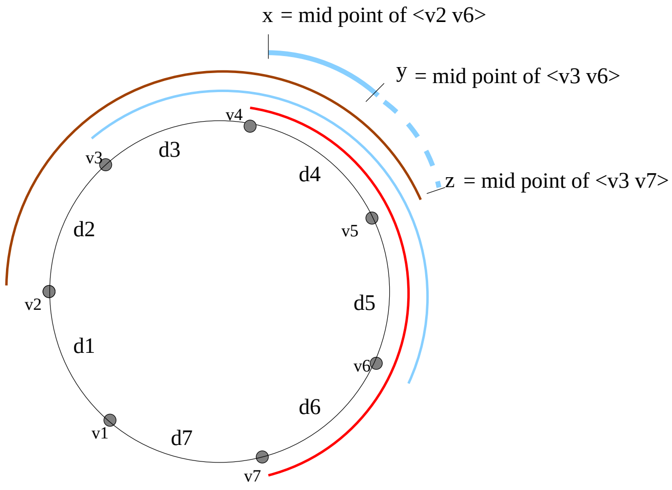


Figure 3: A ring network

the figure, let x be the midpoint of arc $\langle v_2v_6\rangle$, y be the midpoint of arc $\langle v_3v_6\rangle$, z be the midpoint of arc $\langle v_3v_7\rangle$. The nodes on the arc $\langle xy\rangle$ (arc clockwise from x to y) can choose the quorum of $\{v_3, v_4, v_5, v_6\}$ and the length of this arc is given by $\frac{1}{2}d_2$. The delay of a node at y, if it exists, is $\frac{1}{2}(d_3 + d_4 + d_5)$. As we move anti-clockwise from y towards x, delays of nodes increase until at x, the delay is $\frac{1}{2}(d_3 + d_4 + d_5 + d_2)$.

The total delay of nodes in $\langle xy\rangle$ can be approximated by

$$\sum_{s\in V(xy)} delay(s, C) \approx \int_{\frac{1}{2}(d_3+d_4+d_5)}^{\frac{1}{2}(d_3+d_4+d_5+d_2)} x dx$$

where $V(xy)$ is the set of nodes on arc $\langle xy\rangle$, including x and not including y. Similarly, the total delay of nodes in arc $\langle yz\rangle$ is approximated by

$$\sum_{s\in V(yz)} delay(s, C) \approx \int_{\frac{1}{2}(d_3+d_4+d_5)}^{\frac{1}{2}(d_3+d_4+d_5+d_6)} x dx$$

Let us call an arc pair such as $\langle xy\rangle$ and $\langle yz\rangle$ a pair of **twin-arc**. In general, let the end-nodes of a $(2k+1)$-oligarchy be labeled clockwisely as v_1, v_2, v_3, ... v_{2k+1}, starting with any node as v_1. Let the distance between v_i and v_{i+1} be d_i. We represent the star-arc that corresponds to d_i by $\hat{d}_i$. We refer to a j-arc formed clockwisely by j star arcs $\hat{d}_1, \hat{d}_2, ...\hat{d}_j$ by $\hat{d}_1\hat{d}_2...\hat{d}_j$. The total delay of all nodes in the network is approximated by

$$\sum_{s \in V} delay(s, C) \approx \sum_{i=1}^{2k+1} \left[\int_{a_i}^{b_i} x dx \right] \tag{1}$$

where

$$a_i = \frac{1}{2} \sum_{j=i}^{i+k} e_j, \quad b_i = \frac{1}{2} \sum_{j=i}^{i+k+1} e_j \quad \text{and} \quad e_j = d_{\{j \bmod (2k+1)\}}$$

For example for a ring of 7 end-nodes as in Figure 3, this is given by

$$\begin{aligned}
& \tfrac{1}{4}\Big[\quad \frac{1}{2}(d_1+d_2+d_3+d_4+d_5)^2 - \frac{1}{2}(d_1+d_2+d_3+d_4)^2 + \frac{1}{2}(d_1+d_2+d_3+d_4+d_7)^2 - \frac{1}{2}(d_1+d_2+d_3+d_4)^2 + \\
& \quad \frac{1}{2}(d_2+d_3+d_4+d_5+d_6)^2 - \frac{1}{2}(d_2+d_3+d_4+d_5)^2 + \frac{1}{2}(d_2+d_3+d_4+d_5+d_1)^2 - \frac{1}{2}(d_2+d_3+d_4+d_5)^2 + \\
& \quad \dots + \\
& \quad \frac{1}{2}(d_7+d_1+d_2+d_3+d_4)^2 - \frac{1}{2}(d_7+d_1+d_2+d_3)^2 + \frac{1}{2}(d_7+d_1+d_2+d_3+d_6)^2 - \frac{1}{2}(d_7+d_1+d_2+d_3)^2 \Big] \\
= \; & 0.25\Big[\quad d_7(d_1+d_2+d_3+d_4) + \frac{1}{2}d_5^2 + d_5(d_1+d_2+d_3+d_4) + \frac{1}{2}{d_7}^2 + \\
& \quad d_6(d_2+d_3+d_4+d_5) + \frac{1}{2}{d_6}^2 + d_1(d_2+d_3+d_4+d_5) + \frac{1}{2}{d_1}^2 + \\
& \quad \dots + \\
& \quad d_4(d_7+d_1+d_2+d_3) + \frac{1}{2}{d_4}^2 + d_6(d_7+d_1+d_2+d_3) + + \frac{1}{2}{d_6}^2 \Big]
\end{aligned}$$

Lemma 11 *In the above expression,* d_j*, for* $1 \leq j \leq 7$*, is multiplied by each* d_i *where* $1 \leq i \leq 7$ *and* $i \neq j$ *exactly twice.*

Proof: If $\hat{d}_i$ is adjacent to a k-arc of $\hat{d}_{j_1}\hat{d}_{j_2}...\hat{d}_{j_k}$, then a product term appears for each of $d_i d_{j_l}$, for $1 \leq l \leq k$. Consider star-arc $\hat{d}_1$ in Figure 3. It is adjacent to two arcs $\hat{d}_2\hat{d}_3\hat{d}_4$ and $\hat{d}_5\hat{d}_6\hat{d}_7$. In general a star-arc meets with two k-arcs which contain all other star-arcs exactly once. From this consideration, we have the first set of product terms for d_j.

Again consider star-arc $\hat{d}_1$ in Figure 3. It lies on 3 k-arcs : $\hat{d}_1\hat{d}_2\hat{d}_3$, $\hat{d}_7\hat{d}_1\hat{d}_2$, and $\hat{d}_6\hat{d}_7\hat{d}_1$. $\hat{d}_1\hat{d}_2\hat{d}_3$ is adjacent to $\hat{d}_7, \hat{d}_4$. $\hat{d}_7\hat{d}_1\hat{d}_2$ is adjacent to $\hat{d}_6, \hat{d}_3$. $\hat{d}_6\hat{d}_7\hat{d}_1$ is adjacent $\hat{d}_5, \hat{d}_2$. In general, a star-arc lies within each element of a set of k-arcs which are adjacent to each other star-arcs exactly once. From such adjacencies, we have the second set of product terms for d_j. ∎

Lemma 12 *The right hand side of equation (1) can be simplified to* $0.25n^2$.

Proof: Since

$$\left(\sum_{i=1}^{i=2k+1} d_i\right)^2 = \sum_{i=1}^{i=2k+1} d_i^2 + 2 \sum_{1 \le i,j \le 2k+1, i \ne j} d_i d_j \, , \qquad \text{and}$$

$$\sum_{i=1}^{i=2k+1} d_i = n$$

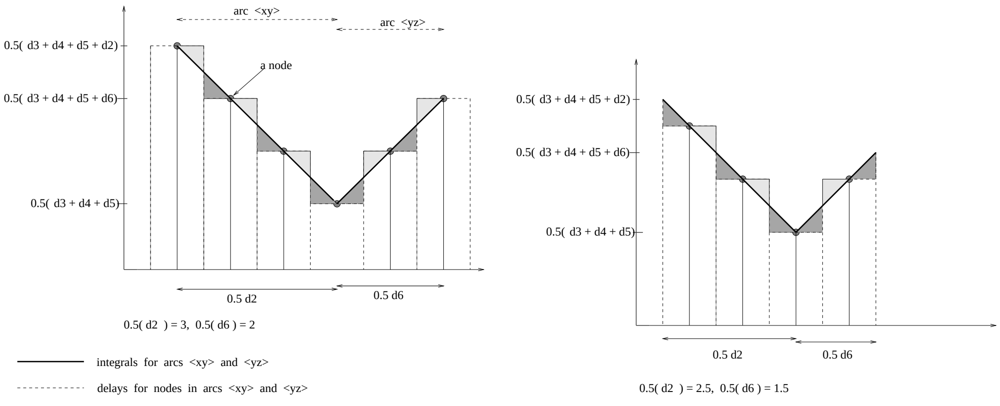
Therefore, from the proof of Lemma 11, the right hand side in equation (1) can be simplified to $0.25n^2$. ∎

To see how close the above approximation is, we look at six cases which are the only possible cases without loss of generality of the points x, y and z in Figure 3. These cases identify different combinations of a node existing or not existing at each of x, y and z. The cases are:

(a) a node exists at each of x, y, and z.
(b) a node exists at y only.
(c) a node exists at each of x and z only.
(d) no node exists at any of x, y, or z.
(e) a node exists at x only.
(f) a node exists at each of x and y only.

Two of the cases, (a) and (b), are shown in Figure 4. In these diagrams, we show the integration for the twin-arcs $\langle xy \rangle$ and $\langle yz \rangle$ in Figure 3, although the values of d_i's for different values of i's are altered in each case. In each case, the total darker grey area minus the total lighter grey area is how much the integration in the right hand side of (1) is greater than the actual sum of delays for the nodes on the twin-arcs $\langle xy \rangle$ and $\langle yz \rangle$. One can readily see that setting different possible values for d_i's has no impact on this difference in area.

In cases (a) and (d) there is no difference between the two kinds of grey area. In case (c), the integration is less than the actual delays by 0.25. In case (b), the integration is more than the actual delays by 0.25. In case (e), the integration is less than the actual delays by 0.125. In case (f), the integration is more than the actual delays by 0.125. Therefore, the delay is less if case (b) appears more frequently.



(a) (b)

Figure 4: Integration for the delays for arcs $\langle xy \rangle$ and $\langle yz \rangle$

Lemma 13 *The mean-delay for a ring with n evenly spaced nodes is at least* $0.25(n-1)$.

Proof: There can be at most n pairs of twin-arcs in a ring with n nodes, and case (b) as in Figure 4(b) is the best possible case for each pair. For case (b), the difference between the integration and the actual delays at each arc is given by 0.25. Hence, the optimal mean-delay will be given by $\frac{1}{n}(0.25n^2 - 0.25n)$, which is $0.25(n-1)$. ∎

Lemma 14 *If $n = 4a+1$, for some positive integer a, then a $(4a+1)$-oligarchy for a ring with n evenly spaced nodes is mean-delay optimal. The mean-delay is* $0.25(n-1)$.

Proof: In this case, there will be totally n pairs of twin-arcs and for each pair, case (b) as in Figure 4(b) holds. ∎

Lemma 15 *For a ring with* $n = 2a + 1$ *evenly spaced nodes, for some positive integer* a*, some* $(2a - 1)$*-oligarchy has a mean-delay of* $0.25(n - 1) + \frac{0.5}{n}$*, which is at most* $0.375/n$ *more than the optimal mean-delay.*

Proof: If $n = 2a + 1$, for some odd integer a, then we form a $(2a - 1)$-oligarchy in which one star-arc has length 3 while the remaining $2a - 2$ star-arcs have length of 1 each. Case (b) of Figure 4 will be the case for each pair of twin-arcs and there are $n - 2$ pairs of twin-arcs. The mean delay will be given by $\frac{1}{n}\left[0.25n^2 - 0.25(n - 2)\right]$, which is $0.25(n - 1) + \frac{0.5}{n}$. We know that the mean-delay of $0.25(n - 1)$ is not achievable since a $(2a + 1)$-oligarchy does not produce case (b) at all twin-arcs. The smallest possible mean-delay can only be $0.25(n - 1) + 0.125/n$. ∎

The following lemma can be deduced by examining the properties of the corresponding oligarchies.

Lemma 16 *For a ring with* 2^a *nodes, for some integer* a *greater than 2, a* $(2^a - 3)$*-oligarchy can be formed by creating one star-arc of length 4, and the remaining star-arcs of length 1. Such an oligarchy has a mean-delay of* $0.25n$*.*

Lemma 17 *There is a coterie for a ring with* n *evenly spaced nodes which has a mean-delay at most* $0.25n$*.*

Proof: For any ring which contains n nodes, where n is not a power of 2, n has a greatest odd factor, let it be f. For such a ring, an f-oligarchy can be formed by creating uniform length star-arcs. From Lemma 14 and Lemma 15 the mean-delay is less than $0.25n$. The case of $n = 2^2$ is trivial and Lemma 16 covers the remaining cases. ∎

Lemma 18 *By building coteries with the above methods, the mean-delay is at most 0.25 greater than the optimal value.*

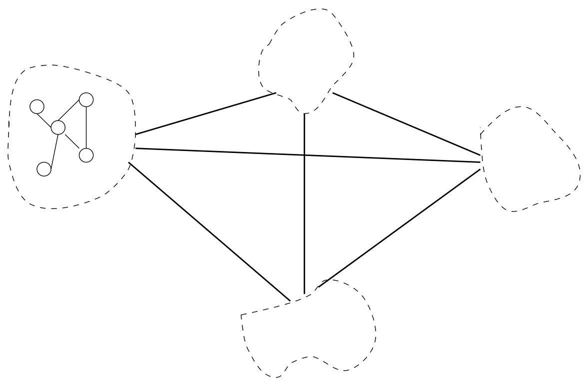


Figure 5: Sketch of a clustered network

Proof: From Lemma 13, the minimum possible mean-delay is $0.25(n-1)$. The Lemma follows from the proof of Lemma 17. ∎

The problem of finding a mean-delay optimal coterie for a ring network is thus partially solved. For a positive integer a, it is solved for $n = 4a + 1$ (Lemma 14), for $n = 2a + 1$, the coterie in Lemma 15 yields a mean-delay at most $0.375/n$ greater than that of an optimal solution. Otherwise, a coterie with mean-delay at most 0.25 greater than an optimal value can be found. The complexities in the above constructions are bounded by $O(n)$.

5 Clustered Networks

Suppose we have a wide area network connecting a number of big cities. Each city is represented by a set of nodes which we call a **cluster**. We assume a cluster is a complete graph. We assume that the message propagation delay within a cluster is small compared to the propagation delay between two different clusters. We call a network of the above form a **clustered network**. Figure 5 sketches a clustered network with 4 clusters, (not all nodes and links are shown). We assume even distribution of the operations among the nodes. We propose two bicoteries for the clustered network, one of which is a wr-coterie.

Let the network be a set of clusters $C = \{C_1, C_2, ...C_m\}$. Let each cluster be a set of nodes such that $C_i = \{s_1, s_2, ...s_{n_i}\}$. Let

$\alpha = C,$
$\beta = \{\{x_1, x_2, \cdots, x_m\} \mid x_i \in C_i\},$
$R = C \cup \beta,$
$W = \{Q_1 \cup Q_2 \mid Q_1 \in C, Q_2 \in \beta\}.$

Theorem 4 *$M = (\alpha, \beta)$ is a bicoterie. For a clustered network with at least 2 clusters, where each cluster has at least 2 nodes, $N = (W, R)$ is a wr-coterie.*

Proof: We can readily see that M satisfies both the intersection property and the nonredundancy property of bicoteries. For a network where each cluster has at least 2 nodes, none of the sets in C is a subset of any set in β. For a network where there are at least 2 clusters, none of the sets in β is a subset of any set in C. Therefore N satisfies the nonredundancy property. It also satisfies the intersection property between W and R and also within W. ∎

If the given network contains only one cluster, then each element in β is a subset of the only element in C, hence R is not nonredundant. If the given network contains some cluster C_i which has only one node, then the set R is not nonredundant since C_i is a subset of all elements in β.

Let us call $M = (\alpha, \beta)$ a **cluster bicoterie** and $N = (W, R)$ a **cluster wr-coterie**.

5.1 Delay of the cluster wr-coterie

Let the virtual distance between any two nodes within a cluster be d_c, the virtual distance between any two nodes from any two clusters be d_l. Probability of read is p, and probability of write is $q = (1 - p)$. For the cluster wr-coterie, each read operation issued from node s will read from the local cluster that contains s, and each write operation from s will write to all other clusters and also the local cluster. Therefore, the mean-delay of an operation is $pd_c + qd_l$.

Theorem 5 *Under our assumptions, for a clustered network with at least 2 clusters, where each cluster has at least 2 nodes, the cluster wr-coterie is mean-delay optimal if the probability of read operations is greater than* 2/3.

Proof: For any wr-coterie for the clustered network under our assumption, if a write quorum does not contain at least one node from each cluster, then some read quorum must contain a remote node. Since we assume uniform distribution of read and write operations among all nodes, there are two possibilities.

(1) Suppose that in such a wr-coterie, a write quorum contains remote nodes but does not contain at least one node from each cluster. Then there will be at least one read quorum which contains remote nodes in order to intersect with this write quorum. Let us call such wr-coteries **type A wr-coteries**. If f is the fraction of the reads that can read from local clusters, then the mean-delay of a type A wr-coteries is

$$\begin{aligned} & fpd_c + (1-f)pd_l + qd_l \\ = \; & pd_c + qd_l + (1-f)p(d_l - d_c) \end{aligned}$$

Since $f < 1$, and we assume $d_l > d_c$, this mean-delay is always greater than that of the cluster wr-coterie.

(2) We may allow a logical write operations of at most one cluster C_1 to write only at the local cluster, since otherwise the intersection property of the write quorums in a wr-coterie is violated. Let us call such wr-coteries **type B wr-coteries**. We assume that the logical operations are evenly distributed among the clusters. Therefore, if there are m clusters, we have one out of m writes that has a delay of d_c, and the rest of the writes have a delay of d_l. For the values written by operations at C_1 (these operations do not write to other clusters), a read operation from another cluster must read from the remote cluster C_1. The mean-delay of a type B wr-coterie is at least

$$\begin{aligned} & \left(\left(p - \frac{p}{m}\frac{m-1}{m}\right)d_c + \left(\frac{p}{m}\frac{m-1}{m}\right)d_l\right) + \left(\left(q - \frac{q}{m}\right)d_l + \frac{q}{m}d_c\right) \qquad (2) \\ = \; & \left(p - \frac{p}{m}\frac{m-1}{m} + \frac{q}{m}\right)d_c + \left(\frac{p(m-1)}{m^2} + q\left(1 - \frac{1}{m}\right)\right)d_l \\ = \; & pd_c + qd_l - \frac{p(m-1)-(1-p)m}{m^2}d_c + \frac{p(m-1)-(1-p)m}{m^2}d_l \end{aligned}$$

$$= \quad pd_c + qd_l + \frac{m - 2pm + p}{m^2} d_c - \frac{m - 2pm + p}{m^2} d_l$$

Since the mean-delay of the cluster wr-coterie is $pd_c + qd_l$, the above expression is greater than or equal to this mean-delay if

$$(m - 2pm + p)(d_c - d_l) \geq 0$$

If $d_c < d_l$, which is our assumption, then the above inequality becomes

$$\begin{aligned} & m - 2pm + p \leq 0 \\ \textit{hence,} \quad & p \geq \frac{m}{2m-1} \\ \Leftrightarrow \quad & p \geq \frac{1}{2} + \frac{1}{2(2m-1)} \end{aligned}$$

Since $\frac{1}{2} + \frac{1}{2(2m-1)} \leq \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$ for $m \geq 2$, we deduce that if d_c is smaller than d_l and p is above $2/3$, then the cluster wr-coterie will have a smaller mean-delay. ∎

Corollary 3 *Under our assumptions, if $p \geq \frac{1}{2} + \frac{1}{2(2m-1)}$, then the cluster wr-coterie is mean-delay optimal. If $p \leq \frac{1}{2} + \frac{1}{2(2m-1)}$, then one mean-delay optimal wr-coterie is a type B wr-coterie which contains a write-quorum C_1 that contains nodes only from a single cluster, and each remaining write-quorum contains at least one node from each cluster.*

If the logical operations are not evenly distributed among the clusters, then the lower bound on the mean-delay of type B wr-coteries may be different from the above. If the distribution is close to even, then the appearances of m in expression (2) above may be replaced by values close to m, and it would be possible to apply similar analysis as in the proof of Theorem 5.

5.2 Max-delay optimal wr-coteries

The only possible values for read-delay(s, C) and write-delay(s, C) for any s, C are d_c and d_l. Hence the only possible values for max-delay(C) for any C are also d_c and

d_l. However, it is not possible for the value of max-delay(C) to be d_c since it means that all the logical read and write operations read or write only the nodes at the local cluster, which violates the intersection property of wr-coteries. Hence any wr-coterie for the clustered network is a max-delay optimal wr-coterie.

6 Conclusion

Quorum consensus is a technique of providing mutual exclusion for distributed systems, and has also been found useful in replicated database systems. Such distributed systems are designed to meet the demands of high availability and short response time. However, with quorum consensus, complex transactions may require much communication overhead, which can be an important factor in the response time. We investigate coteries that help mitigate the effect of message propagation delay. We derived a number of delay-optimal quorum consensus protocols for common network topologies of trees, rings, and clustered networks. Delay-optimal quorum consensus for hypercubes and grids is investigated in [10].

Delay-optimal quorum consensus for the general graphs is an open problem. We have left open a number of other problems: finding delay-optimal quorum consensus of a ring with nodes not evenly spaced, finding delay-optimal quorum consensus for other types of network, etc. We also leave open the problem of delay-optimal coteries or bicoteries when the network reliability is not so high, so that the probability of getting a quorum should be taken into consideration. A delay-optimal coterie would be one that can achieve minimal expected delay considering failure probabilities.

Acknowledgements

The author thanks the anonymous referees for their thorough review of the initial drafts and very helpful comments and recommendations.

References

[1] A. El Abbadi and S. Toueg. Availability in partitioned replicated databases. In *Proc. of the ACM 5th Symposium on Principles of Database Systems*, pages 240–251, March 1986.

[2] D. Agrawal and A. El Abbadi. An efficient and fault-tolerant solution for distributed mutual exclusion. *ACM Transactions on Computer Systems*, 9(1):1–20, Feb 1991.

[3] D. Barbara and H. Garcia-Molina. The vulnerability of vote assignments. *ACM Transactions on Computer Systems*, 4(3):187–213, August 1986.

[4] P.A. Bernstein and N. Goodman. An algorithm for concurrency control and recovery in replicated distributed databases. *ACM Transactions on Database Systems*, 9(4):596–615, Dec 1984.

[5] P.A. Bernstein, V. Hadzilacos, and N. Goodman. *Concurrecncy Control and Recovery in Database Systems.* Addison-Wesley, Reading Mass., 1987.

[6] Agrawal D. and El Abbadi A. The generalized tree quorum protocol: An efficient approach for managing replicated data. *ACM Transactions on Database Systems*, 17(4):689–717, Dec 1992.

[7] A. El-Abbadi and S. Toueg. Maintaining availability in partitioned replicated databases. *ACM Transactions on Database Systems*, 14(2):264–290, June 1989.

[8] M. Fredman and L. Khachiyan. *On The Complexity of dualization of monotone disjunctive normal forms, Technical Report LCSR-TR-225.* Dept. of Computer Science, Rutgers University, 1994.

[9] A. Fu. *Enhancing Concurrency and Availability for Database Systems.* Ph.D. Thesis, Simon Fraser University, April 1990.

[10] A. Fu, M.H. Wong, T.W. Lau, and G.F.Ng. *Cost Optimal Coteries, Technical Report CS-TR-94-07.* Chinese University of Hong Kong, 1994.

[11] H. Garcia-Molina and D. Barbara. How to assign votes in a distributed system. *Journal of the ACM*, 32(4):841–860, Oct 1985.

[12] D.K. Gifford. Weighted voting for replicated data. In *Proc. of the 7th ACM SIGOPS Symposium on Operating Systems Principles*, pages 150–159, Dec 1979.

[13] J. Gray. The cost of messages. In *Proc. of the ACM 7th Symposium on Principles of Distributed Computing*, pages 1–7, 1988.

[14] J. Gray. *Paralleslism: The New Imperative in Computer Architecture.* Tutorial notes, International Conference of Very Large Databases, VLDB, 1994.

[15] M. Herlihy. A quorum-consensus replication method for abstract data types. *ACM Transactions on Computer Systems*, 4(1):32–53, Feb 1986.

[16] M. Herlihy. Dynamic quorum adjustment for partitioned data. *ACM Transactions on Database Systems*, 12(2):170–194, June 1987.

[17] T. Ibaraki and T. Kameda. A theory of coteries: mutual exclusion in distributed systems. *IEEE Transaction Parallel and Distributed Systems*, 4(7):779–794, July 1993.

[18] T. Ibaraki, H. Nagamochi, and T. Kameda. Optimal coteries for rings and related networks. In *12th International Conference on Distributed Computing Systems*, pages 650–656, June 1992.

[19] T. Ibaraki, H. Nagamochi, and T. Kameda. Optimal coteries for rings and related networks. *Distributed Computing*, (8):191–201, 1995.

[20] A. Kumar. Hierarchical quorum consensus: A new algorithm for managing replicated data. *IEEE Transactions on Computers*, 40(9):996–1004, 1991.

[21] M. Maekawa. A $\sqrt{N}$ algorithm for mutual exclusion in decentralized systems. *ACM Transactions on Computer Systems*, 3(2):145–159, 1985.

[22] C.H. Papadimitriou and M. Sideri. Optimal coteries. In *Proc. of ACM 10th Symposium on Principles of Distributed Computing*, pages 75–80, 1991.

[23] D. Peleg and A. Wool. Crumbling walls: A class of practical and efficient quorum systems. In *Proc. of the ACM 14th Symposium on Principles of Distributed Systems*, pages 120–129, August 1995.

[24] B. C. Tansel, R. L. Francis, and T. J. Lowe. Location on networks: A survey. Part I: The p-center and p-median problems. *Management Science*, 29(4):482–497, Apr 1983.

[25] R.H. Thomas. A majority consensus approach to concurrency control for multiple copy databases. *ACM Transactions on Database Systems*, 4(2):180–209, 1979.

[26] O. Wolfson and A. Milo. The multicast policy and its relationship to replicated data placement. *ACM Transactions on Database Systems*, 16(1):181–205, 1991.