Voting Protocols

Instructor: Randal Burns
Lecture for March 11, 2000
Computer Science 600.416
Johns Hopkins University

Scheduling and Deadlock
Voting

• Voting is pessimistic replica control
  – Preserve mutual consistency among replicas at the expense of availability

• As opposed to optimistic protocols
  – Allow potentially conflicting updates to occur in separate partitions
  – Optimistic about transactions conflicting

• Accept updates only in the “majority” partition
  – We will see that majority has different meanings in different protocols
  – A better name is the “primary component”, indicating that only one group of computers are active at any one time.
Problem Formulation

• File $f$
• Replicated at $n$ sites
• Read and write quorums $R$ and $W$
  – The ROWA protocol has $R=l$ and $W=n$
  – The WORA protocol has $R=n$ and $W=l$
  – Any protocol must have the property that
    • $R+W \geq n$ for linearly ordered sites
    • $R+W > n$ for unordered sites
Static Voting

• Define read and write quorums
  \[ R = \left\lfloor \frac{n}{2} \right\rfloor + 1 \quad W = \left\lceil \frac{n}{2} \right\rceil \quad \text{or} \quad R = \left\lfloor \frac{n}{2} \right\rfloor \quad W = \left\lceil \frac{n}{2} \right\rceil + 1 \]

• Example – protocol uses a version number
  – Note that since every possible set \( R \) overlaps every possible set \( W \), that the most up to date copy will be inevitably be read.

• Majority in this case is always the same (hence static) and is a majority of all nodes
  – In partitions, one partition will have a majority of nodes and that is the majority partition.
Problems with Static Voting

• Low availability
  – How many site failures (or partitioned nodes) can the system withstand before replicas are unavailable?
    • \( \left\lfloor \frac{n}{2} \right\rfloor \)
Dynamic Linear Voting

• Dynamic
  – Concept: allow $n$ to evolve over time
  – A majority is defined in reference to the sites that wrote data on the previous instance, *i.e.* a majority of the previous majority

• Linear
  – Concept: order the nodes in order to break ties between groups that are the same size
  – If there are 2 groups of two, the group with the lower numbered node is the primary group
Dynamic Voting

- Define read and write quorums
  - \( R = \left\lceil \frac{n}{2} \right\rceil \) \( W = \left\lfloor \frac{n}{2} \right\rfloor \) subject to linear ordering

- Require extra state at each site
  - Version number (VN), as with static voting
  - Sites cardinality (SC), how many sites participated in the last write
  - Distinguished site (DS), lowest site (linearly) in the last majority

- Example

Scheduling and Deadlock
Advantages and Drawback

• DLV can tolerate n-1 failures

• DLV can be unavailable when static voting would be available
  – Example
Extensions and Other Concepts

• Weighted voting
  – Assign a larger number of votes to more important or more available sites
  – Simple extension, with a little more metadata

• Currency-based voting
  – Continuous version of weighted-voting
  – Allows for currency to be transferred site to site
  – No need to identify sites by name or know apriori who has currency, collect it at run-time

• Witnesses
  – Sites that do not replicate the data can vote
  – Can increase availability without increasing consistency and replication costs

• Voting without version numbers
  – Keep track of the set of sites that voted, rather than a version number and use this set of sites to determine majorities
  – Overhead scales with the number of sites rather than log(number of writes)