A Modular Approach to Shared-Memory Consensus, with Applications to the Probabilistic-Write Model

James Aspnes *
Department of Computer Science
Yale University
New Haven, USA
aspnes@cs.yale.edu

ABSTRACT

We define two new classes of shared-memory objects: **ratifiers**, which detect agreement, and **conciliators**, which ensure agreement with some probability. We show that consensus can be solved by an alternating sequence of these objects, and observe that most known randomized consensus algorithms have this structure.

We give a deterministic m-valued ratifier for an unbounded number of processes that uses $\lg m + \Theta(\log \log m)$ space and individual work. We also give a randomized conciliator for any number of values in the probabilistic-write model with n processes that guarantees agreement with constant probability while using one multiwriter register, $O(\log n)$ expected individual work, and $\Theta(n)$ expected total work. Combining these objects gives a consensus protocol for the probabilistic-write model that uses $O(\log n)$ individual work and $O(n \log m)$ total work. No previous protocol in this model uses sublinear individual work or linear total work for constant m.

Categories and Subject Descriptors

F.1.2 [**Modes of Computation**]: Parallelism and concurrency

General Terms

Algorithms, Theory

Keywords

distributed computing, shared memory, consensus, randomization

1. INTRODUCTION

The consensus problem [26] is to devise a protocol so that n processes, each with a private input, can agree on a single common output that is equal to some process's input. For asynchronous deterministic processes, consensus is known to be impossible with a single crash failure in either a message-passing [21] or shared-memory [25] model. These impossibility results can be overcome using randomization [16], and can even tolerate up to $n-1$ crash failures in a shared-memory model [1]. There is a substantial literature on randomized consensus protocols that tolerate crash failures; see [6] for a survey.

In this paper, we concentrate on **wait-free** [22] consensus protocols in various shared-memory models with only atomic registers as their base objects (see Section 2). Our goal is to find efficient algorithms, that minimize both the total number of operations carried out by all processes (the **total work** T_{total}) and the maximum number of operations carried out by any single process (the **individual work** $T_{\text{individual}}$).

An important factor in determining the cost of consensus is the strength of the adversary scheduler that determines the interleaving of operations; this is an effect of what information the adversary is allowed to use when choosing each step. For an adaptive adversary, which can observe the entire state of the system when choosing which process moves next, optimal algorithms are known for both total expected work [12] and individual expected work [8]. But for weaker adversaries, finding tight bounds on expected work has remained an open problem.

Recently, Attiya and Censor [11] have shown that any randomized consensus protocol for 2-valued consensus tolerating f failures must fail to terminate after $k(n-f)$ total steps with probability at least $\frac{1}{c^k}$. This bound is very general: it applies even with a very weak adversary that cannot observe the process's execution, and works even in a model that allows global coins visible to all processes and $O(1)$-cost snapshot operations. They also show that it is tight for models with correlated local coins, using as a matching upper bound the BINARY-CONSENSUS protocol of [13].

Unfortunately, it gives little insight into the *expected* cost of consensus; the geometric series converges to $\Omega(n)$, a trivial lower bound on total work. It also leaves open the possibility that any algorithm with uncorrelated local coins or that allows more than two input values will necessarily be more expensive.

The best currently known upper bounds on expected work for weak adversaries in a asynchronous shared-memory model without correlated coins are $O(\log n)$ expected individual work (and $O(n \log n)$ expected total work) assuming a **value-oblivious adversary** that cannot observe internal states of processes or the contents of registers (but can see where

*Supported in part by NSF grant CCF-0916389.

PODC'10, July 25–28, 2010, Zurich, Switzerland.

processes choose to read and write) [13]; and $O(n \log\log n)$ expected total work (with the same bound on individual work) in the **probabilistic-write model**, where a process can flip a coin and choose whether or not to execute a write operation based on its outcome [19]. (We argue in Section 2 that this model is a special case of assuming a **location-oblivious adversary**, one that cannot distinguish between operations on different registers.)

The present work improves on previous upper bounds for weak-adversary randomized consensus by giving a new family of consensus protocols for the probabilistic-write model. For 2-valued consensus, we obtain $O(\log n)$ expected individual work and $O(n)$ expected total work with a single algorithm. This is the first weak-adversary consensus protocol with optimal total work, and demonstrates that the Attiya-Censor lower bound is asymptotically tight for the probabilistic-write model. For m-valued consensus, the individual work remains the same, but the total work rises to $O(n \log m)$; there may still be room for improvement in both this bound and the individual-work bound in both the binary and m-valued cases.

Our algorithms are structured as classic Las Vegas style randomized algorithms [15]: we explicitly separate the problems of producing agreement and detecting it into abstract **conciliator** objects, which produce agreement with some probability, and **ratifier** objects, which cause the processes to decide only once agreement has been reached. This modular approach allows these tasks to be optimized separately, with the expected cost of a consensus object built in this fashion proportional to the sum of the cost of a conciliator and ratifier. Furthermore, since a consensus object satisfies the specifications for both classes of objects, a lower bound on either conciliators or ratifiers will translate directly into a lower bound on consensus. This may help in determining the limits to weak-adversary consensus protocols.

Formal definitions of these objects, together with methods for composing them, are given in Section 3. Generic algorithms for consensus (that do not specify the details of their component conciliators and ratifiers) are given in Section 4.

In Section 5, we give an implementation of a conciliator for arbitrarily many values that produces agreement with constant probability at the cost of $O(\log n)$ individual work in the worst case and $O(n)$ expected total work. (This dominates the cost of the ratifier for binary consensus, hence our binary-consensus bound.) The implementation is very simple, and uses only a single multi-writer register large enough to hold one input value or a null value $\perp$. We also show that **weak shared coins** as defined in [9] can be turned into conciliators. While this does not give any new results, it shows how the conciliator+ratifier framework can be used to model previous shared-memory consensus protocols.

In Section 6, we give a deterministic implementation of an m-valued ratifier that uses $\lceil \lg m \rceil + O(\log\log m)$ registers and has the same individual work cost. For m-valued consensus, the total work cost of the ratifier becomes dominant. Though we have not fully resolved the cost of consensus in weak-adversary models, we believe that our decomposition of consensus into separate conciliation and ratification steps is natural, and that the limitations of our current implementations hint at where further improvements would be needed to obtain better bounds.

2. MODEL

We use the standard model for asynchronous shared memory: there are n processes that communicate by reading and writing atomic registers, with the value returned by each read equal to the last value written. Asynchrony is modeled by interleaving. Each process that has not halted has exactly one **pending operation** in each state; an **execution**, which consists of a sequence of operations and their return values, is constructed by repeatedly applying pending operations. A **partial execution** is one in which not all processes have halted. The choice of which pending operation occurs in each state is determined by an **adversary scheduler**, a function from partial executions to process ids. We consider only **wait-free** protocols, where there are no fairness conditions on executions.

Processes are randomized: they have access to **local coins** that are not predictable by the adversary but also not visible to other processes. Formally, we can think of the local coins as probabilistic read-only objects private to each process, so that executing a coin-flip is an operation.

The **total work** or **total step complexity** of an execution is the total number of operations it contains. The **individual work** or **individual step complexity** is the maximum number of operations carried out by any single process. Local computation (including local coin-flip operations) is excluded from both measures.

2.1 Strong and weak adversaries

The strength of the adversary has a large effect on the cost of randomized consensus. An **adaptive** or **strong** adversary has no restrictions on its choice of which process carries out the next operation. A **weak adversary** is any adversary that is not strong. Typically this means that the adversary's knowledge is limited in some way; this can be modeled by an equivalence relation on partial executions, and requiring that the adversary choose the same process in equivalent conditions.

Some examples of weak adversaries:

- **Oblivious adversary**. An oblivious adversary has no knowledge of the execution and schedules processes in a fixed order. Here two executions are equivalent if they have the same length.
- **Value-oblivious adversary**. A value-oblivious adversary cannot observe internal states of processes, the contents of registers, or the values in pending write operations, but it can distinguish between pending operations of different types (e.g., read vs. write) or to different locations. Value-oblivious adversaries of various kinds are used by [13, 14, 18]. The useful property of value-oblivious adversaries is that they allow the outcome of a local coin-flip to be stockpiled in a register as a pre-flipped global coin that is visible to all processes but still not predictable by the adversary until some process uses its value.
- **Location-oblivious adversary**. A location-oblivious adversary cannot observe internal states of processes, *can* observe the contents of memory and the values of pending write operations, but cannot distinguish between pending write operations to different locations. This allows for **probabilistic writes** as defined in the "strong model" of [1] to represent the assumptions of

the Chor-Israeli-Li protocol [20] and as used more recently by Cheung [19]. These are write operations that take effect only with some probability, where the adversary cannot choose whether to allow the write operation based on the outcome of the coin-flip. In a location-oblivious model, they can be implemented by having a process randomly choose between writing to the desired location or to some dummy location (corresponding to an omitted write).

For some of our results, we assume probabilistic writes, which can be implemented using a location-oblivious adversary as described above.

The practical justification for weak adversaries is that realistic implementations of shared memory are not likely to distinguish between operations with similar properties. This is a very natural assumption for value-oblivious adversaries. For location-oblivious adversaries, it may be less natural, but it is still plausible if we assume that the relevant memory locations are all stored on the same page, and that the main source of timing uncertainty lies in a page-based memory management system that treats all locations on the same page as equivalent.

Other restrictions on the adversary have been used in the consensus literature. These include requiring that the adversary implement a quantum or priority-based scheduling algorithm [2, 3, 27] or that the adversary include some random jitter in its scheduling decisions [5]. We discuss implications of some of these assumptions for our framework in Section 4.2.

3. DECOMPOSING CONSENSUS

Here we show how a consensus protocol can be decomposed into a collection of conciliators and ratifiers. This provides an alternative to previous general frameworks for implementing consensus, such as the rounds-based weak-shared-coin framework of Aspnes and Herlihy [9].

Recall that a **consensus protocol** must satisfy three properties:

- **Validity**. Every process's output equals some process's input.
- **Termination**. Every process terminates with probability 1.
- **Agreement**. Every process outputs the same value.

A **consensus object** is a shared-memory object with a single **consensus** operation, such that if each process executes this operation exactly with its input, the resulting outputs satisfy the requirements for randomized consensus.

We are going to replace consensus objects by sequences of objects satisfying weaker conditions. Like consensus objects, these weaker objects will be **one-shot** objects in which each process executes at most one operation on the object, and we will expect inputs to be from the set of possible decision values Σ. But the outputs will be annotated by a **decision bit** that indicates whether the protocol should terminate immediately or continue to the next object in the sequence: an output of $(1, v)$ means to decide v immediately, while an output of $(0, v)$ means to continue to the next object, using v as input. We call an object annotated with a decision bit in this way a **deciding object**.

These objects will generally satisfy only **Las Vegas** requirements [15], where an object is not required to guarantee agreement unless it returns a decision bit of 1. This requires replacing agreement with a weaker notion of **coherence**, which says only that non-deciders stick to any value chosen by a decider.

- **Coherence**. If any process outputs $(1, v)$, then no process outputs (d, v') for $v' \neq v$.

A **weak consensus object** is an object that satisfies validity, termination, and coherence. Note that weak consensus objects may be very weak indeed: an object that simply copies its input to its output with decision bit 0 satisfies all three properties. But we will use weak consensus objects as a basis for building stronger objects.

3.1 Conciliators and ratifiers

3.1.1 Conciliators

A **conciliator** is a weak consensus object whose outputs agree with constant probability, but that does not detect when or if agreement occurs. Formally, it satisfies the additional requirement of **probabilistic agreement**:

- **Probabilistic agreement.** There is a fixed **agreement probability** $\delta > 0$ such that, for any adversary strategy, the probability that all return values are equal is at least δ.

A conciliator object may correspond to a weak shared coin (with machinery added to ensure validity) or to the first-mover-wins technique of Chor-Israel-Li-style protocols [19, 20]. We give examples of both types of constructions in Section 5. Because conciliators are not expected to detect agreement, our constructions will satisfy coherence vacuously, by always returning a decision bit of 0.

3.1.2 Ratifiers

Actual decisions are produced by **ratifiers**, objects that detect agreement. A ratifier is a weak consensus object that satisfies the additional requirement of **acceptance**:

- **Acceptance**. If all inputs are equal to v, all outputs are $(1, v)$.

Because ratifiers are not required to produce agreement, they can be implemented deterministically with low space and work complexity. In Section 6, we give an implementation of an m-valued ratifier that uses $O(\log m)$ multiwriter registers and $O(\log m)$ individual work. For binary consensus, this reduces to 3 registers and at most 4 operations per process.

3.2 Composing objects

Our consensus protocols will consist of an alternating sequence of ratifiers and conciliators. To define this formally, it helps to have a notion of composing objects.

If $X : A \to B$ is a deciding object with inputs in A and outputs in B and $Y : B \to B$ is a deciding object with inputs and outputs in B, then their composition $(X; Y) : A \to B$ is a deciding object whose operation $\mathrm{op}_{(X;Y)}(x)$ is given by the code in Procedure `Composition`.

Informally, we perform op_X first, and feed the result to op_Y only if op_X does not decide on its own. We can also

```
1 (d, v) ← op_X(x)
2 if d = 1 then
3     return (1, v)
4 else
5     return op_Y(v)
6 end
```

Procedure `Composition(X,Y,x)`

think of this rule as implementing an exception mechanism where a decision by X immediately terminates the composite object without executing Y.

Note that in $(X;Y)$, X comes first. This is the reverse of the usual rule for function composition. Note also that because the output of $(X;Y)$ could be generated by either X or Y, both must have the same output type, which must also be equal to Y's input type.

It is easy to see that composition is associative: $((X;Y);Z)$ has exactly the same behavior as $(X;(Y;Z))$ for all objects X, Y, Z with appropriate types. We will generally omit the parentheses and write $(X;Y;Z)$ in this case. Similarly, we can define compositions $(X_1;X_2;\ldots;X_k)$ for arbitrarily many objects, and even define infinite compositions $(X_1;X_2;\ldots)$ in the obvious way.

If P is some predicate on objects, we say that composition **preserves** P if $P(X)$ and $P(Y)$ implies $P(X;Y)$. This naturally extends by induction to longer finite compositions.

3.2.1 Composing weak consensus objects

The property of being a weak consensus object is preserved by composition.

LEMMA 1. *Validity is preserved by composition.*

PROOF. Suppose that X and Y satisfy validity. Then any return value of $(X;Y)$ is either (a) a return value of X, and thus equal to an input to $(X;Y)$; or (b) a return value of Y, and thus equal to an input to Y, which is in turn an output of X and thus equal to an input of $(X;Y)$. □

The converse does not hold. It may be that the first object always scrambles its inputs (but does not decide) while the second unscrambles them.

LEMMA 2. *Termination is preserved by composition.*

PROOF. Immediate. □

A partial converse holds for termination: if $(X;Y)$ terminates with probability 1, so does X.

LEMMA 3. *If X satisfies coherence, and Y satisfies both validity and coherence, then $(X;Y)$ satisfies coherence.*

PROOF. We consider two cases, depending on whether any process skips Y:

1. X outputs $(1,v)$ for some process. Then coherence for X implies all processes obtain $(-,v)$ from X, and any other process either obtains the same return value directly from X or obtains the same return value from Y because of the validity of Y.

2. X does not output $(1,-)$ for any process. Then all processes execute op_Y and coherence follows from coherence of Y. □

COROLLARY 4. *The property of being a weak consensus object is preserved by composition.*

PROOF. Apply Lemmas 1, 2, and 3. □

4. RECOMPOSING CONSENSUS

We give three constructions of consensus protocols. The first uses an infinite sequence of ratifiers and conciliators (but terminates after using only a constant number on average); the second truncates the infinite sequence by falling back to some fixed-space consensus protocol with low probability; and the third omits the conciliators and relies on scheduling restrictions to terminate.

4.1 Consensus with conciliators and ratifiers

4.1.1 Unbounded construction

Let R_i be a separate ratifier object for each $i \geq -1$ and C_i a conciliator object with agreement probability δ for each $i \geq 1$. Construct the composite object

$$U = R_{-1}; R_0; C_1; R_1; C_2; R_2; \ldots.$$

Observe that this object satisfies termination (and thus yields well-defined return values), because with probability 1 some C_i eventually produces agreement, causing the following R_i to force every process to decide.

Suppose now that every process decides in some prefix $R_{-1}; R_0; C_1; R_1; \ldots; C_k; R_k$. Then this prefix is a weak consensus object by Corollary 4, so the output values are consistent with validity and coherence (and thus agreement because the processes decide). It follows that object U is a randomized consensus object.

The initial prefix $R_{-1}; R_0$ implements a fast path for the case where some process p finishes R_{-1} before any process with a different input arrives.[1] If this case holds, then the acceptance condition implies that p must decide, because it cannot distinguish this execution from one in which all processes have the same input. But then coherence implies that all processes have the same output from R_{-1} and thus decide in R_0. This avoids the overhead of running a conciliator when conciliators are more expensive than ratifiers.

In a general execution, the cost of object U depends on the cost of the R_i and C_i. The expected waiting time until some C_i successfully produces agreement is at most $1/\delta$, so given some time measure T we have $\mathrm{E}[T(U)] \leq 2T(R) + (1/\delta)(T(C)+T(R)) = O(T(C)+T(R))$ when δ is constant. So the cost of consensus will be asymptotically equal to the worse of the costs of conciliation and ratification.

4.1.2 Bounded construction

The preceding construction requires unbounded space. We can avoid this by leveraging any bounded-space construction to truncate the sequence of objects (e.g. the polynomial-time bounded-space construction of [4]). Let R_i and C_i be as above and let $B = (R_{-1}; R_0; C_1; R_1; C_2; R_2; \ldots; C_k; R_k; K)$, where K is a bounded-memory consensus protocol. That B is a consensus object follows from Corollary 4 and the fact that K decides if no earlier object does. The expected complexity of B for any time measure T is bounded by $O((1/\delta)(T(R)+T(C)) + (1-\delta)^k T(K))$. If δ is constant and $T(K)$ is polynomial in n, then for some $k = O(\log n)$ this reduces to $O(T(R)+T(C))$ as in the previous case.

[1] I am indebted to Azza Abouzeid for suggesting this idea.

We state this result as a theorem:

THEOREM 5. *Given any bounded-space implementations of conciliators $\{C_i\}$ with constant agreement probability and ratifiers $\{R_i\}$, there exists an implementation of consensus that uses bounded space, with expected individual work $O(\max(T_{\text{individual}}(C_i), T_{\text{individual}}(R_i)))$ and expected total work $O(\max(T_{\text{total}}(C_i), T_{\text{total}}(R_i))$.*

4.2 Consensus with ratifiers only

Under severe restrictions on the adversary, it is possible to solve consensus without using conciliators at all. Let $R^* = R_1; R_2; \ldots$ consist of an unbounded sequence of ratifiers. If during an execution of R^*, some process p completes R_i before any process with a conflicting value enters R_i, then p decides by the same analysis as for the fast-path prefix above. So if we can guarantee that this eventually happens, R^* implements a consensus protocol.

For binary consensus using a constant-individual-work ratifier, R^* is essentially equivalent to the LEAN-CONSENSUS protocol of [5], so the analysis there shows that R^* will terminate in $O(\log n)$ individual work with a noisy scheduler. This is a scheduler that chooses in advance the timing of all steps of the algorithm, but has its choices perturbed by random errors that accumulate over time. The proof in [5] shows that eventually this cumulative error will push some process ahead of all the others. We expect that comparable results can be obtained for m-valued consensus, but as our current m-valued ratifier requires $\Theta(\log m)$ work, additional analysis would be needed.

The R^* protocol also works with priority-based scheduling as described by [27]. Here each process is assigned a unique priority that does not change over the duration of its execution of the consensus protocol, and each step is taken by the highest-priority process that currently has a pending operation. It is easy to see that in this model, the highest-priority process to execute the protocol will eventually overtake all other processes and reach some ratifier alone, unless a decision occurs earlier. In either case, we achieve consensus. The R^* protocol is less efficient than the protocol from [27], which uses only two registers and terminates in at most six operations per process. Part of the reason for this inefficiency is that we are assuming R is an arbitrary ratifier object. Particular implementations of ratifier objects might offer better performance.

5. IMPLEMENTING A CONCILIATOR

Here we show how to implement a conciliator. We first show that the classic weak shared coin approach of [9] fits in our framework, and then give a new conciliator for the probabilistic-write model with very strong performance bounds.

5.1 Conciliators from weak shared coins

A **weak shared coin** [9] is a protocol in which each process decides on a bit, and for some agreement probability $\delta > 0$ it holds that the probability that all processes decide 0 and the probability that all processes decide 1 are both at least δ, regardless of the adversary's choices. This is both stronger than our definition of a conciliator—in that a conciliator need not be unpredictable—and weaker—in that a weak shared coin need not respect validity. Fortunately, validity is easily enforced.

```
   shared data:
      binary registers r_0 and r_1, initially 0;
      weak shared coin SharedCoin
1  r_v ← 1
2  if r_¬v = 1 then
3      return (0, SharedCoin())
4  else
5      return (0, v)
6  end
```

Procedure `CoinConciliator(`v`)`

Code for a shared-coin-based binary conciliator is given as Procedure `CoinConciliator`.

THEOREM 6. *Given a weak shared coin* `SharedCoin` *with agreement probability δ, Procedure* `CoinConciliator` *satisfies termination, validity, coherence, and probabilistic agreement with agreement probability at least δ.*

PROOF. Termination follows from termination of `SharedCoin` and the lack of loops. Validity follows from the fact that if all inputs are v, then no process writes $r_{\neg v}$ and all processes skip the shared coin. Coherence is satisfied vacuously.

For probabilistic agreement, essentially the same argument as used in [9] applies: if some process p skips the shared coin and returns v, then it must have written r_v before reading 0 from $r_{\neg v}$. It follows that any process with input $\neg v$ reads r_v after writing $r_{\neg v}$, sees 1, and executes the shared coin. So only v can be returned by processes skipping the shared coin, and with probability at least δ, all processes executing the shared coin also return v. □

The cost of this implementation is likely to be dominated by the cost of `SharedCoin`; the other parts add 2 registers and 2 register operations. Note that this implementation only provides a 2-valued conciliator. How to extend a shared coin to more values is not obvious; for one choice, see [23]. Our new algorithm below avoids this restriction to a bounded set of values, at the cost of requiring a weaker adversary.

5.2 Conciliators using probabilistic writes

In the probabilistic-write model, we can build a conciliator for arbitrarily many values, following the approach of the classic algorithm of Chor, Israeli, and Li [20] and its more recent optimization by Cheung [19]. The basic idea is to have a single multi-writer register, which in an ideal execution is written only once by some winning process. This is enforced by having each process attempt to write the register with small probability only if it has not yet observed a value in the register. Assuming the probabilities are set correctly, there will be a constant chance that some process writes the register but the next $n-1$ writes do not, meaning that the corresponding processes read the winning value and return it.

Previous protocols in this model have used a constant $\Theta(1/n)$ probability for each write. This gives a bound on both total and individual work of $O(n)$. We generalize this approach to allow processes to become impatient over time and increase their probabilities (analogously to the increasing weighted votes of [7, 8, 10]). This new approach will

```
   shared data: register r, initially ⊥
1  k ← 0
2  while r = ⊥ do
3      write v to r with probability 2^k/2n
4      k ← k + 1
5  end
6  return (0, r)
```

Procedure `ImpatientFirstMoverConciliator(`v`)`

allow us to get optimal $O(n)$ total work while reducing the individual work bound to $O(\log n)$.

Because the process writes with probability 1 once 2^k reaches $2n$, we immediately get a bound of $2\lceil \lg 2n \rceil + 2 = 2\lceil \lg n \rceil + 4$ on the number of operations.[2] At the same time, since each write succeeds with probability at least $\frac{1}{2n}$, we retain the total work bound of $6n$ expected operations. But it may be that the cost of impatient processes is that we lose the constant agreement probability. That this is not the case is shown in the following theorem:

THEOREM 7. *Procedure* `ImpatientFirstMoverConciliator` *satisfies termination in expected* $6n$ *total work and at most* $2\lg n + O(1)$ *individual work; validity; coherence; and agreement with probability at least* $(1-e^{-1/4})(1/4) \approx 0.0553$.

PROOF. The bounds on individual work and total work have already been established. Validity and coherence are immediate from inspection of the algorithm.

For agreement, observe that the adversary's choices are effectively limited to choosing what order the processes will attempt their probabilistic writes, and that it succeeds only if it can get one of the remaining $n-1$ processes to carry out a successful write after some initial process successfully writes to r. Fix some adversary strategy, and let p_i be the probability that the i-th write succeeds if no previous write succeeds. Let t be the minimum value for which $\sum_{i=1}^{t} p_i \geq 1/4$. We will argue that with at least constant probability, some write succeeds in the first t attempts, and thereafter no process writes r on its next attempt. In this case exactly one value is written to r and this value is returned by all processes.

The probability that none of the first t writes succeed is given by $\prod_{i=1}^{t}(1-p_i) \leq \prod_{i=1}^{t} \exp(-p_i) = \exp\left(-\sum_{i=1}^{t} p_i\right) \leq e^{-1/4}$. So with probability at least $1 - e^{-1/4}$, some process writes r at a time t' when $\sum_{i=1}^{t'-1} p_i$ is still small.

The reason we care about this is that we can use the bound on $\sum p_i$ to get a bound on the probability that some other process then overwrites r. Consider any single process p, and suppose that it has failed to carry out its first k_p writes. The probability that its next write succeeds is $\frac{1}{2n}2^{k_p} = \frac{1}{2n}\left(1 + \sum_{j=0}^{k_p-1} 2^j\right) = \frac{1}{2n} + \sum_{j=0}^{k_p} \frac{2^j}{2n}$.

If process q carries out the first successful write as one of the first t writes, we can bound the right-hand side of the above inequality when summed over all $p \neq q$ by letting each k_p count the number of unsuccessful writes carried out by p among the at most $t-1$ writes preceding q's write. The probability that q's value is overwritten is then bounded by

$$\begin{aligned}\sum_{p\neq q} \frac{2^{k_p}}{2n} &\leq \sum_{p\neq q}\left(\frac{1}{2n} + \sum_{j=0}^{k_p-1}\frac{2^j}{2n}\right)\\ &\leq \frac{n-1}{2n} + \sum_{i=1}^{t-1} p_i\\ &< \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.\end{aligned}$$

Note that this bound does not depend on which write succeeds. We thus have

$$\begin{aligned}\Pr[\text{only one write occurs}] &\geq \left(1 - e^{-1/4}\right)\left(1 - \frac{3}{4}\right)\\ &= \left(1 - e^{-1/4}\right)(1/4)\\ &\approx 0.0553.\end{aligned}$$

□

6. IMPLEMENTING A RATIFIER

Here we show how to build a deterministic m-valued ratifier with $\Theta(\log m)$ individual work.

6.1 Write and read quorums

The operation of our ratifier is similar to the follow-the-leader mechanisms in the classic Chor-Israeli-Li [20] and Aspnes-Herlihy [9] protocols. A process first announces that it has a particular value v by writing to all registers in a **write quorum** W_v that depends on v. It then examines a special *proposal* register proposal. If this register is empty, the process may propose its value by writing to the proposal register; otherwise, it adopts the previously proposed value as its new preferred value in place of v. Whatever value preference it obtains, it returns $(1, \text{preference})$ only if no other value has been announced; otherwise, it returns $(0, \text{preference})$. Conflicting values are detected by reading all registers in a read quorum $R_{\text{preference}}$. This works as long as for every two distinct values v and v', $R_{v'}$ includes some register written in W_v that is not written in $W_{v'}$.

Code for this implementation is given as Procedure `Ratifier`.

THEOREM 8. *Let* $W_v \cap R_{v'} = \emptyset$ *if and only if* $v = v'$. *Then Procedure* `Ratifier` *satisfies termination, validity, coherence, and acceptance.*

PROOF. Termination is immediate from the lack of unbounded loops. Validity holds because any return value is either an initial input v or an input read indirectly from proposal. Acceptance holds because if every process has input v, then no process announces any value $v' \neq v$, or writes any $v' \neq v$ to proposal, so preference $= v$ for all processes and the test in Line 10 always fails.

For coherence, suppose that some process p returns $(1, v)$ (it is not hard to see that a process can only return its own input in this case). Then p observed no conflicting v' when it executed Line 10. It follows that no process p' with a conflicting value v' had yet completed Line 3 before p finished executing all of the code up to Line 10. Since p either reads v from proposal in Line 4 or sets proposal to v in Line 8, we have that no process p' with input $v' \neq v$

[2] We assume here that each probabilistic write costs 1 unit whether or not the write succeeds. We do not require that a process detect if it performs a successful write, as it will leave the loop following its next read in any case. If we can detect success, the individual work bound can be reduced by 2 if we return immediately after a successful write.

```
   shared data:
      register proposal, initially ⊥;
      binary registers r_i, initially 0
   local data: preference, u
 1 foreach r_i ∈ W_v do
 2     r_i ← 1
 3 end
 4 u ← proposal; if u ≠ ⊥ then
 5     preference ← u
 6 else
 7     preference ← v
 8     proposal ← preference
 9 end
10 if r_i ≠ 0 for some r_i ∈ R_preference then
11     return (0, preference)
12 else
13     return (1, preference)
14 end
```

Procedure `Ratifier(`v`)`

finishes Line 3 before proposal is set to v. Thus no such p' writes any $v' \neq v$ to proposal, and every process adopts v as its preference—and hence its ultimate return value—before reaching Line 10. □

6.2 Choice of quorums

It remains only to specify a choice of quorums satisfying the condition in Theorem 8. Because the cost of Procedure `Ratifier` is dominated by writing W_v and checking $R_{\text{preference}}$, our goal is to keep both write quorums and read quorums as small as possible.

Here are some possible choices:

1. For a binary ratifier, use two 1-bit registers r_0 and r_1, and let $W_v = \{r_v\}$ and $R_v = \{r_{\neg v}\}$. Using this implementation, `Ratifier` requires at most 4 register operations and uses only 3 registers: r_0, r_1, and proposal.

2. A generalization of the preceding mechanism to m values is to use a pool of k registers $r_1 \ldots r_k$, where $k = \lg m + O(\log\log m)$ satisfies $\binom{k}{\lfloor k/2 \rfloor} \geq m$. For each value v, assign a distinct write quorum W_v of size $\lfloor k/2 \rfloor$ and a complementary read quorum $R_v = \overline{W}_v$. Then $R_v \cap W_v = \emptyset$ and $R_v \cap W_{v'} \neq \emptyset$ when $v \neq v'$. An m-valued instance of `Ratifier` constructed using this technique requires $\lg m + O(\log\log m)$ registers and $\lg m + O(\log\log m)$ individual work.

   This choice of quorums is optimal, in the sense of maximizing the number of distinct values m given a fixed bound k on $|W_v| + |R_v|$. This follows from a classic result in extremal set theory, known as Bollabás's Theorem [17]; the version we give here is taken from [24].

   THEOREM 9 ([24], THEOREM 9.8). *Let $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$ be two sequences of sets such that $A_i \cap B_j = \emptyset$ if and only if $i = j$. Then*

$$\sum_{i=1}^{m} \binom{a_i + b_i}{a_i}^{-1} \leq 1, \tag{1}$$

   where $a_i = |A_i|$ and $b_i = |B_i|$.

   Letting each A_i correspond to some W_i and each B_i to some R_i, each term in the left-hand side of (1) is minimized by setting $|W_i| + |R_i| = k$ $|W_i| = \lfloor k/2 \rfloor$; this gives $m = \binom{k}{\lfloor k/2 \rfloor}$ as above.

3. An alternate encoding of values into write quorums gives a simpler implementation with almost as good performance. Here we use a two-dimensional array of registers r_{ij} where $i \in \{1 \ldots \lceil \lg m \rceil\}$ and $j \in \{0, 1\}$. Writing each v as a $\lceil \lg m \rceil$-bit vector, let W_v equal $\{r_{iv_i}\}$ and R_v equal its complement. The rest of the analysis is the same as in the previous method; the space complexity is exactly $2\lceil \lg m \rceil + 1$ registers and the individual work is at most $2\lceil \lg m \rceil + 2$ operations.

4. Finally, in a cheap-snapshot or cheap-collect model (where reading an array of n single-writer registers takes $O(1)$ time), we can simulate write quorums of size 1 (and corresponding read quorums of size $m - 1$) by having each process announce its value by writing its own register and test if any process has announced a conflicting value using a single collect operation. The individual work bound in this case is 4 operations, as in the binary case. (While this model is not realistic, a cheap-collect ratifier helps put a limit on what lower bounds can be achieved in a cheap-collect model.)

Using either construction of $O(\log m)$-sized quorums with Theorem 8 gives the following result:

THEOREM 10. *For any m, an m-valued ratifier can be implemented deterministically for any number of processes using $O(\log m)$ atomic registers and $O(\log m)$ individual work.*

7. DISCUSSION

We have shown that separating consensus into explicit objects responsible for generating and detecting agreement allows for a simpler description of the overall protocol and for optimizing the resulting objects independently. For the probabilistic-write model, this gives the first known algorithm with sublinear individual work and the first algorithm with linear total work when only a constant number of input values are permitted.

Considering conciliators and ratifiers separately also offers guidance for seeking possible non-trivial lower bounds on expected work in weak-adversary models with uncorrelated local coins. For m-valued consensus, we currently have two obstacles to improved individual work: we would need to reduce both the $O(\log n)$ upper bound on conciliators and the $O(\log m)$ upper bound on ratifiers. While it is not necessarily the case that any consensus protocol separates cleanly into these components (for example, the protocol of [4] clearly does not), because a consensus object satisfies the specifications of both a conciliator and a ratifier, any lower bound on either object also gives a lower bound on consensus. So it may be that concentrating on proving lower bounds for conciliators or ratifiers separately could yield better lower bounds for consensus in general.

Acknowledgments

The decomposition of consensus into conciliator and ratifier objects was inspired in part by a desire to give a better unified presentation of known consensus protocols. I would

like to thank the students in Yale's Spring 2010 Theory of Distributed Systems class for serving as test subjects for the resulting pedagogical experiment and for offering several helpful comments. I would also like to thank Azza Abouzeid for suggesting that the consensus protocols of Section 4.1 should include a fast path that avoids conciliators when the fastest processes agree. Finally, I would like to thank Keren Censor Hillel for many useful comments on an early draft of this paper.

8. REFERENCES

[1] Karl Abrahamson. On achieving consensus using a shared memory. In *Proceedings of the 7th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 291–302, 1988.

[2] J. H. Anderson, R. Jain, and D. Ott. Wait-free synchronization in quantum-based multiprogrammed systems. In *Distributed Computing; 12th International Symposium; Proceedings*, volume 1499, pages 34–45, Andros, Greece, September 1998. Springer-Verlag.

[3] James H. Anderson and Mark Moir. Wait-free syncronization in multiprogrammed systems: Integrating priority-based and quantum-based scheduling. In *Proceedings of the Eighteenth Annual ACM Symposium on Principles of Distributed Computing*, pages 123–132, Atlanta, Georgia, USA, 3–6 May 1999.

[4] James Aspnes. Time- and space-efficient randomized consensus. *Journal of Algorithms*, 14(3):414–431, May 1993.

[5] James Aspnes. Fast deterministic consensus in a noisy environment. *Journal of Algorithms*, 45(1):16–39, October 2002.

[6] James Aspnes. Randomized protocols for aynchronous consensus. *Distributed Computing*, 16(2-3):165–176, September 2003.

[7] James Aspnes, Hagit Attiya, and Keren Censor. Randomized consensus in expected $O(n \log n)$ individual work. In *PODC '08: Proceedings of the Twenty-Seventh ACM Symposium on Principles of Distributed Computing*, pages 325–334, August 2008.

[8] James Aspnes and Keren Censor. Approximate shared-memory counting despite a strong adversary. In *SODA '09: Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pages 441–450, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics.

[9] James Aspnes and Maurice Herlihy. Fast randomized consensus using shared memory. *Journal of Algorithms*, 11(3):441–461, September 1990.

[10] James Aspnes and Orli Waarts. Randomized consensus in expected $O(N \log^2 N)$ operations per processor. *SIAM Journal on Computing*, 25(5):1024–1044, October 1996.

[11] Hagit Attiya and Keren Censor. Lower bounds for randomized consensus under a weak adversary. In *PODC '08: Proceedings of the Twenty-Seventh ACM Symposium on Principles of Distributed Computing*, pages 315–324, New York, NY, USA, 2008. ACM.

[12] Hagit Attiya and Keren Censor. Tight bounds for asynchronous randomized consensus. *Journal of the ACM*, 55(5):20, October 2008.

[13] Yonatan Aumann. Efficient asynchronous consensus with the weak adversary scheduler. In *PODC '97: Proceedings of the Sixteenth Annual ACM Symposium on Principles of Distributed Computing*, pages 209–218, New York, NY, USA, 1997. ACM.

[14] Yonatan Aumann and Michael A. Bender. Efficient low-contention asynchronous consensus with the value-oblivious adversary scheduler. *Distributed Computing*, 17(3):191–207, 2005.

[15] László Babai. Monte-carlo algorithms in graph isomorphism testing. Technical Report D.M.S. 79-10, Université de Montréal, 1979.

[16] Michael Ben-Or. Another advantage of free choice: Completely asynchronous agreement protocols (extended abstract). In *Proceedings of the Second Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, pages 27–30, Montreal, Quebec, Canada, August 1983.

[17] B. Bollobás. On generalized graphs. *Acta Mathematica Hungarica*, 16(3):447–452, September 1965.

[18] Tushar Deepak Chandra. Polylog randomized wait-free consensus. In *Proceedings of the Fifteenth Annual ACM Symposium on Principles of Distributed Computing*, pages 166–175, Philadelphia, Pennsylvania, USA, 23–26 May 1996.

[19] Ling Cheung. Randomized wait-free consensus using an atomicity assumption. In James H. Anderson, Giuseppe Prencipe, and Roger Wattenhofer, editors, *OPODIS*, volume 3974 of *Lecture Notes in Computer Science*, pages 47–60. Springer, 2005.

[20] Benny Chor, Amos Israeli, and Ming Li. Wait-free consensus using asynchronous hardware. *SIAM J. Comput.*, 23(4):701–712, 1994.

[21] Michael J. Fischer, Nancy A. Lynch, and Michael S. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(2):374–382, April 1985.

[22] Maurice Herlihy. Wait-free synchronization. *ACM Transactions on Programming Languages and Systems*, 13(1):124–149, January 1991.

[23] Keren Censor Hillel. Multi-sided shared coins and randomized set agreement. To appear, SPAA 2010, January 2010.

[24] Stasys Jukna. *Extremal combinatorics: with applications in computer science*. Springer-Verlag, 2001.

[25] Michael C. Loui and Hosame H. Abu-Amara. Memory requirements for agreement among unreliable asynchronous processes. *Advances in Computing Research*, pages 163–183, 1987.

[26] M. Pease, R. Shostak, and L. Lamport. Reaching agreement in the presence of faults. *Journal of the ACM*, 27(2):228–234, 1980.

[27] Srikanth Ramamurthy, Mark Moir, and James H. Anderson. Real-time object sharing with minimal system support. In *Proceedings of the Fifteenth Annual ACM Symposium on Principles of Distributed Computing*, pages 233–242, New York, NY, USA, 1996. ACM.