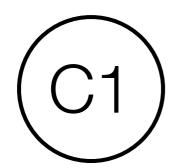
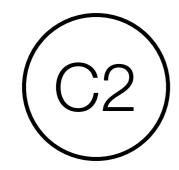
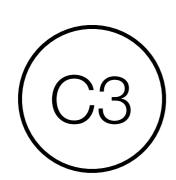
A High Throughput Paxos Variant That Mere Mortals Can Understand

Neil Giridharan, Joseph M. Hellerstein, Ion Stoica, Adriana Szekeres, **Michael Whittaker**

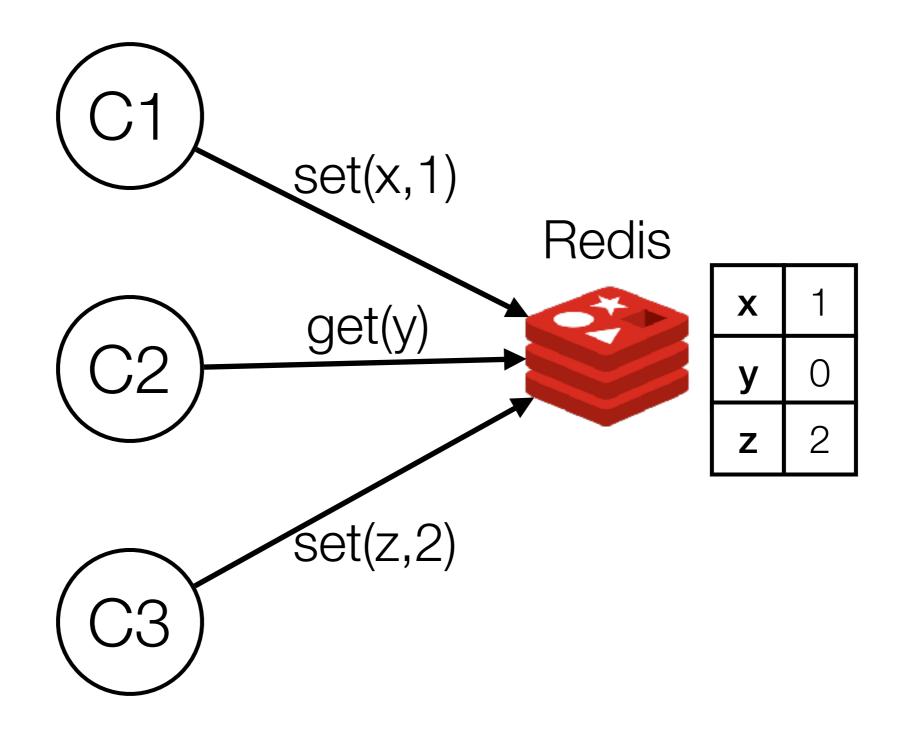


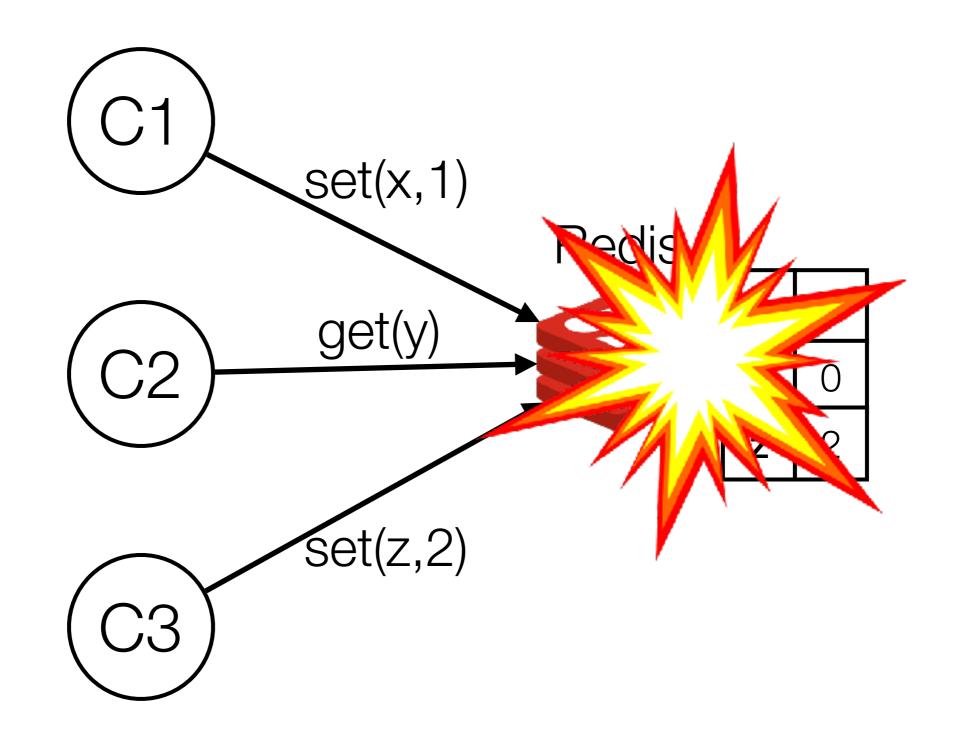


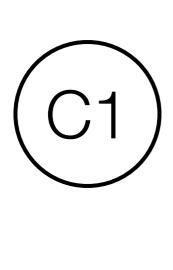


Redis

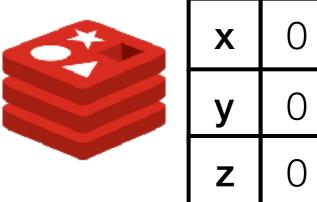


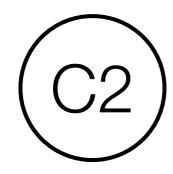






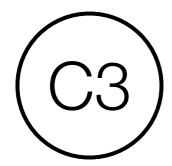






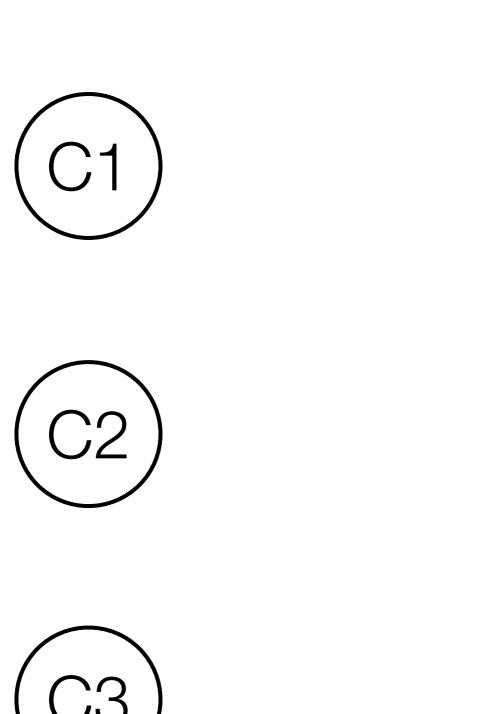


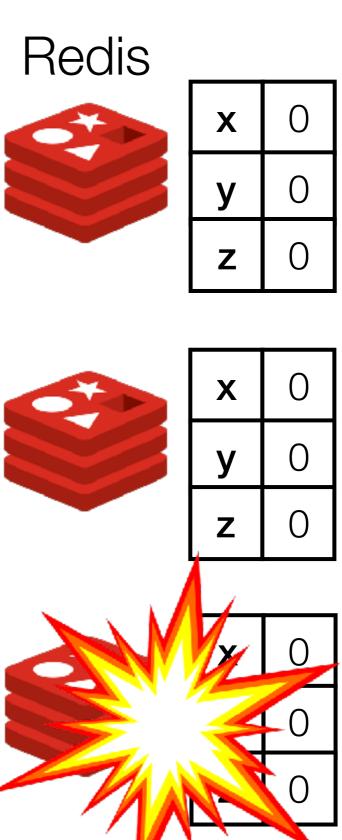
X	0
У	0
Z	0

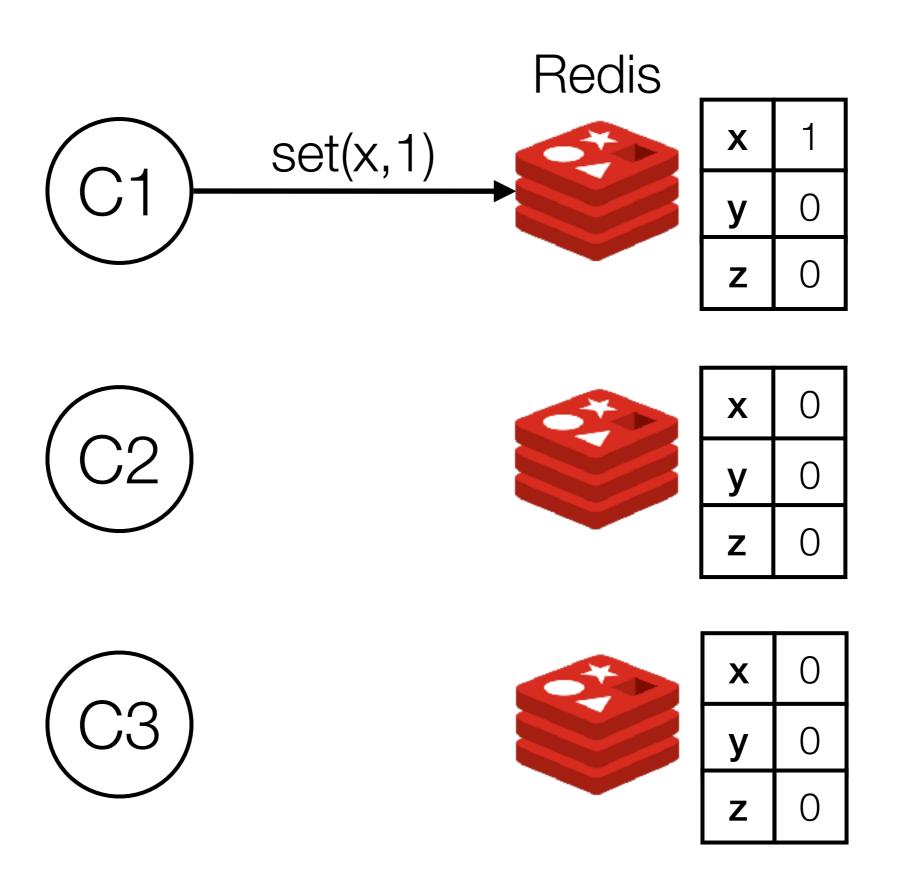


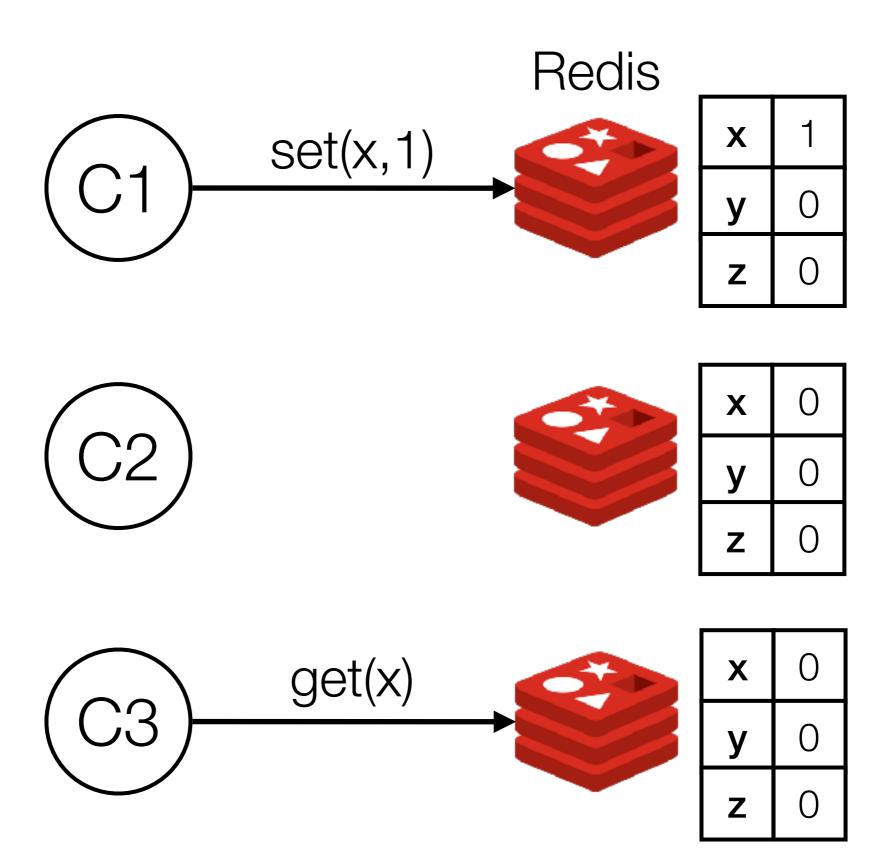


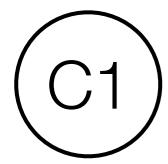
X	0
у	0
Z	0









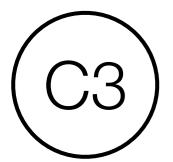




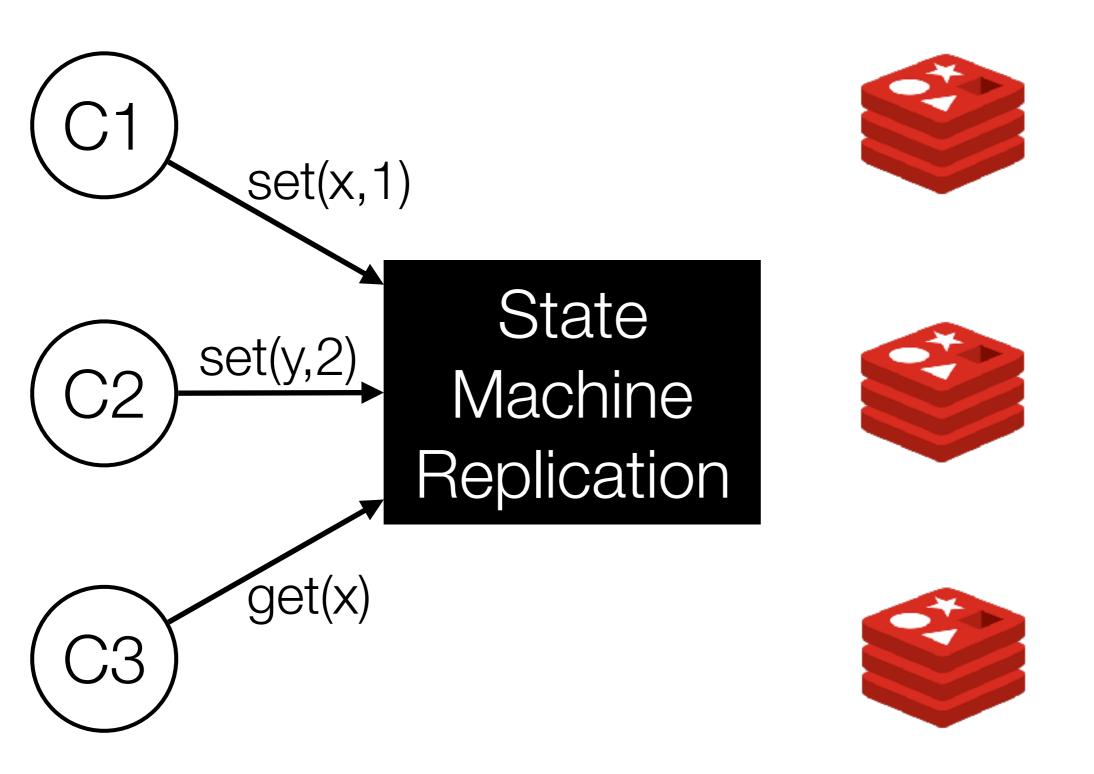


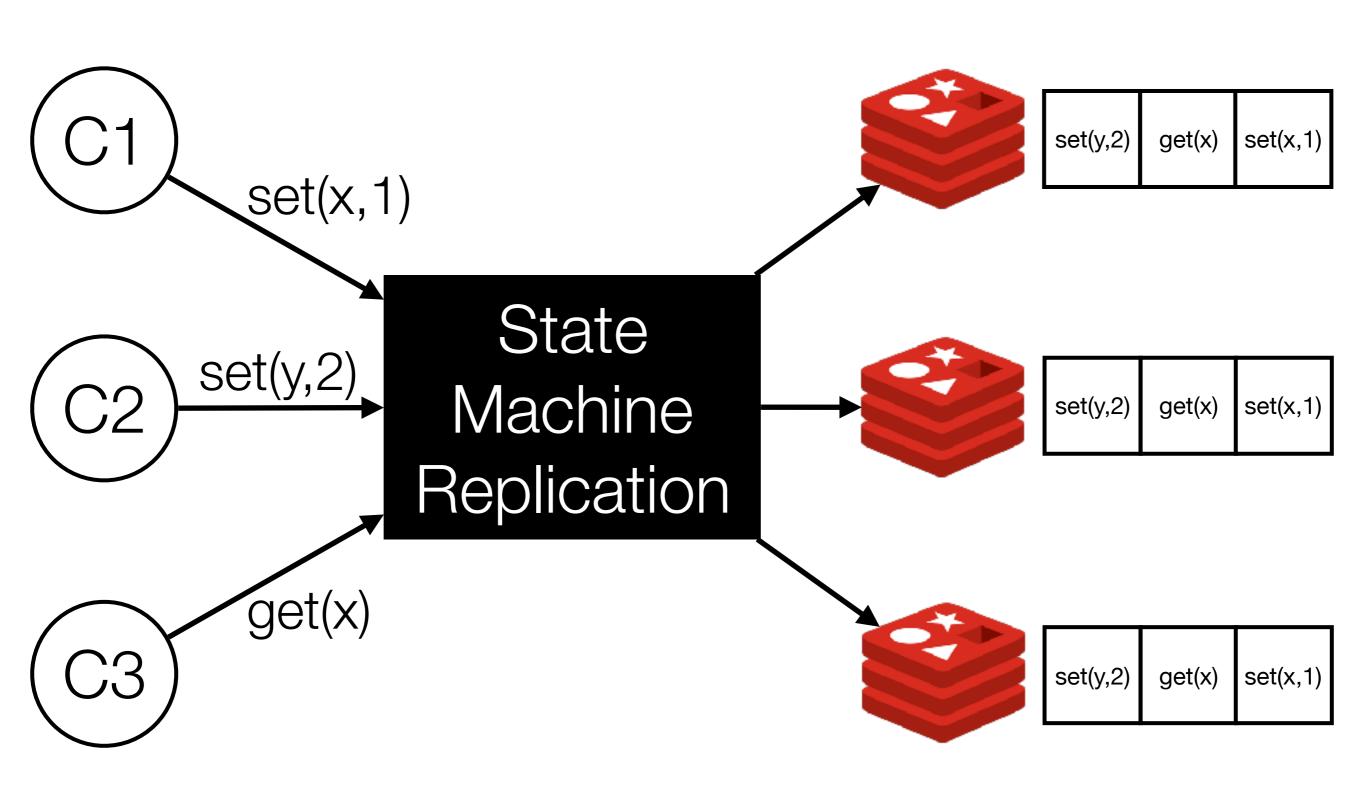
State
Machine
Replication











MultiPaxos

clients







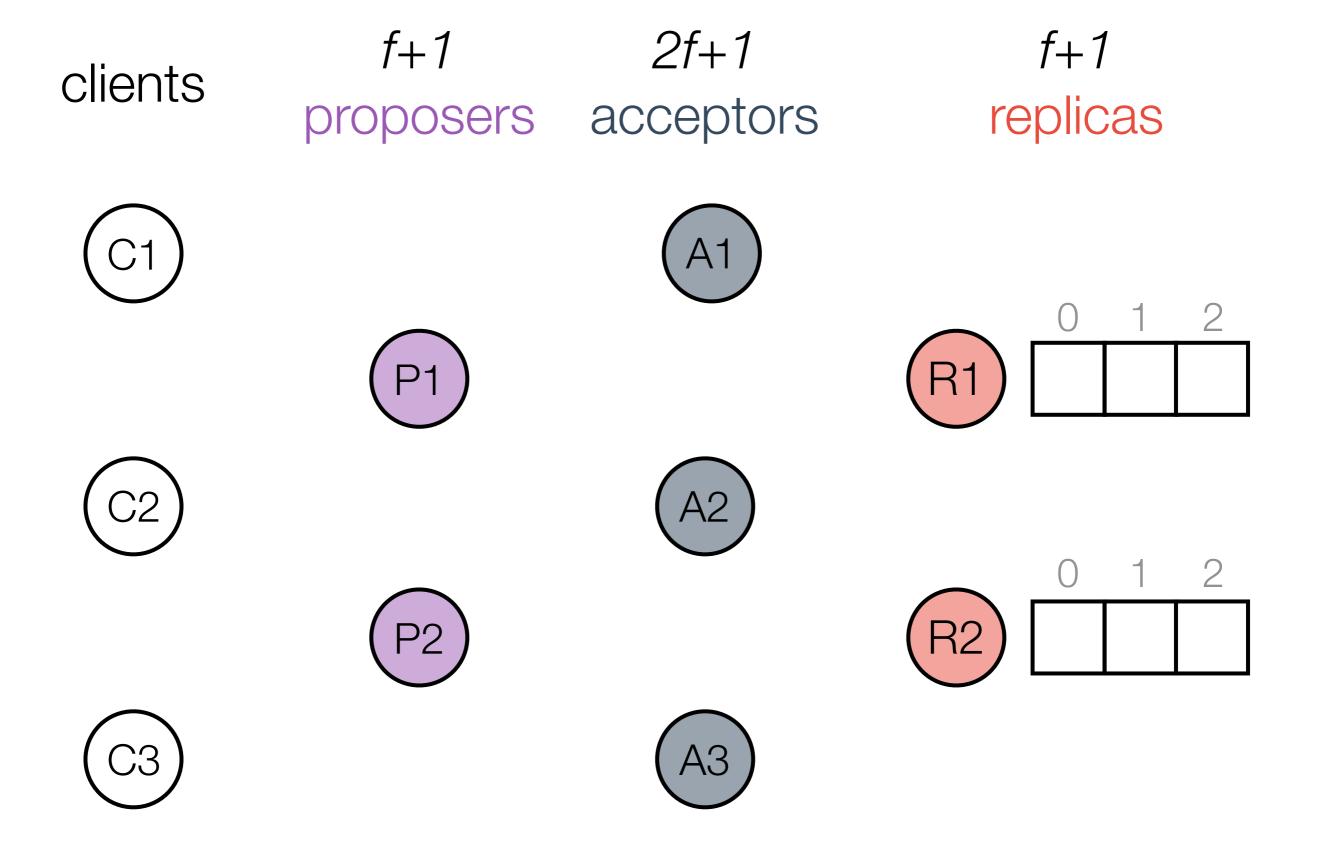
clients f+1 proposers

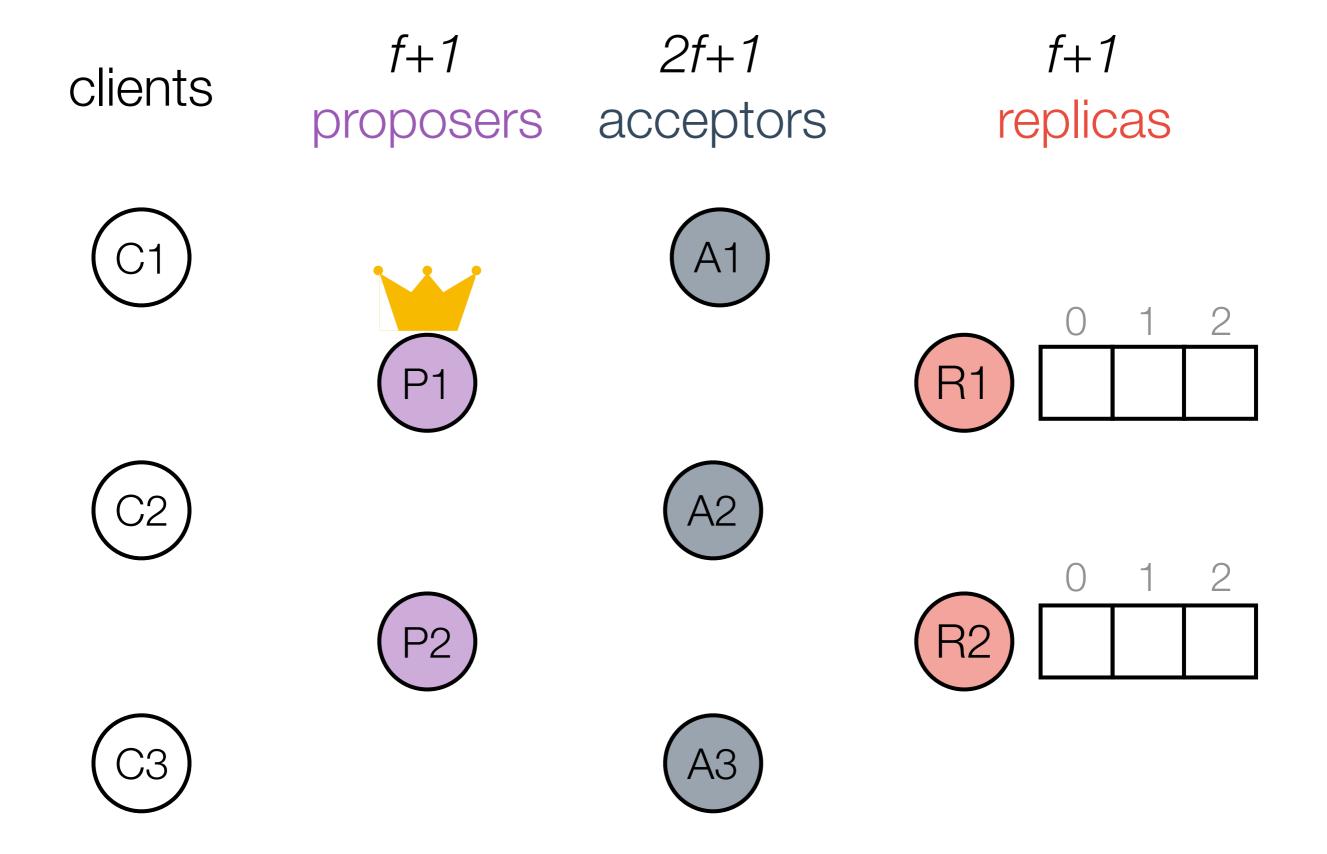
(C1)
(P1)

C2 P2

(C3)

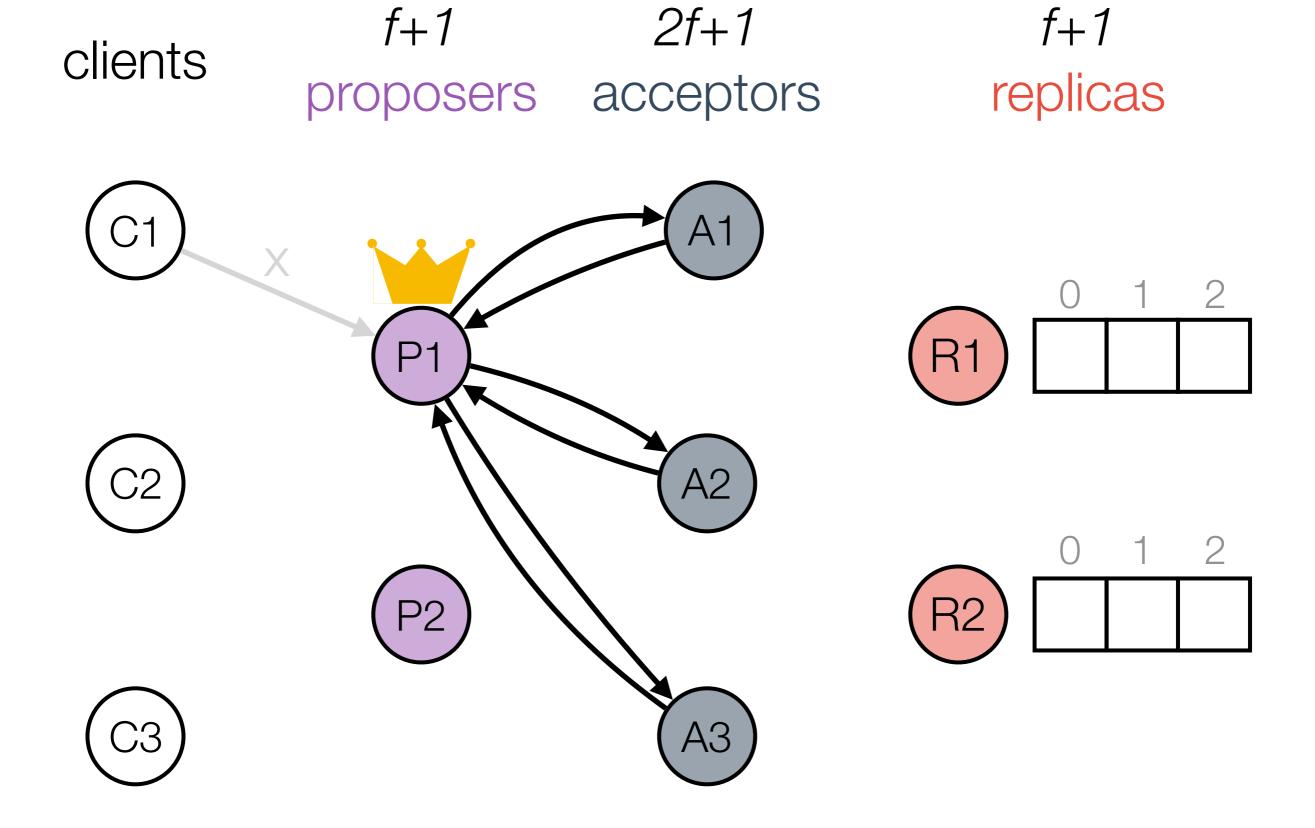
f+1 2f+1 clients proposers acceptors

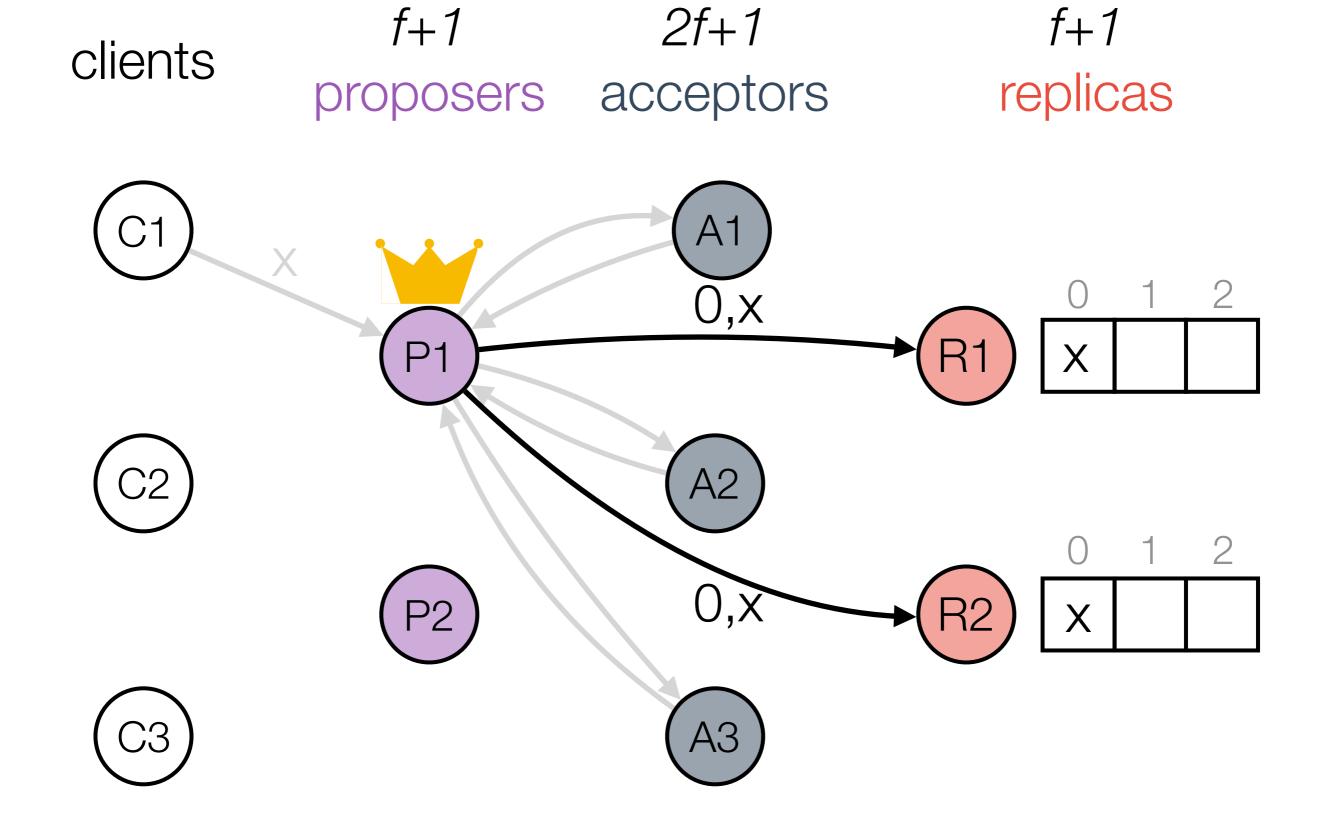


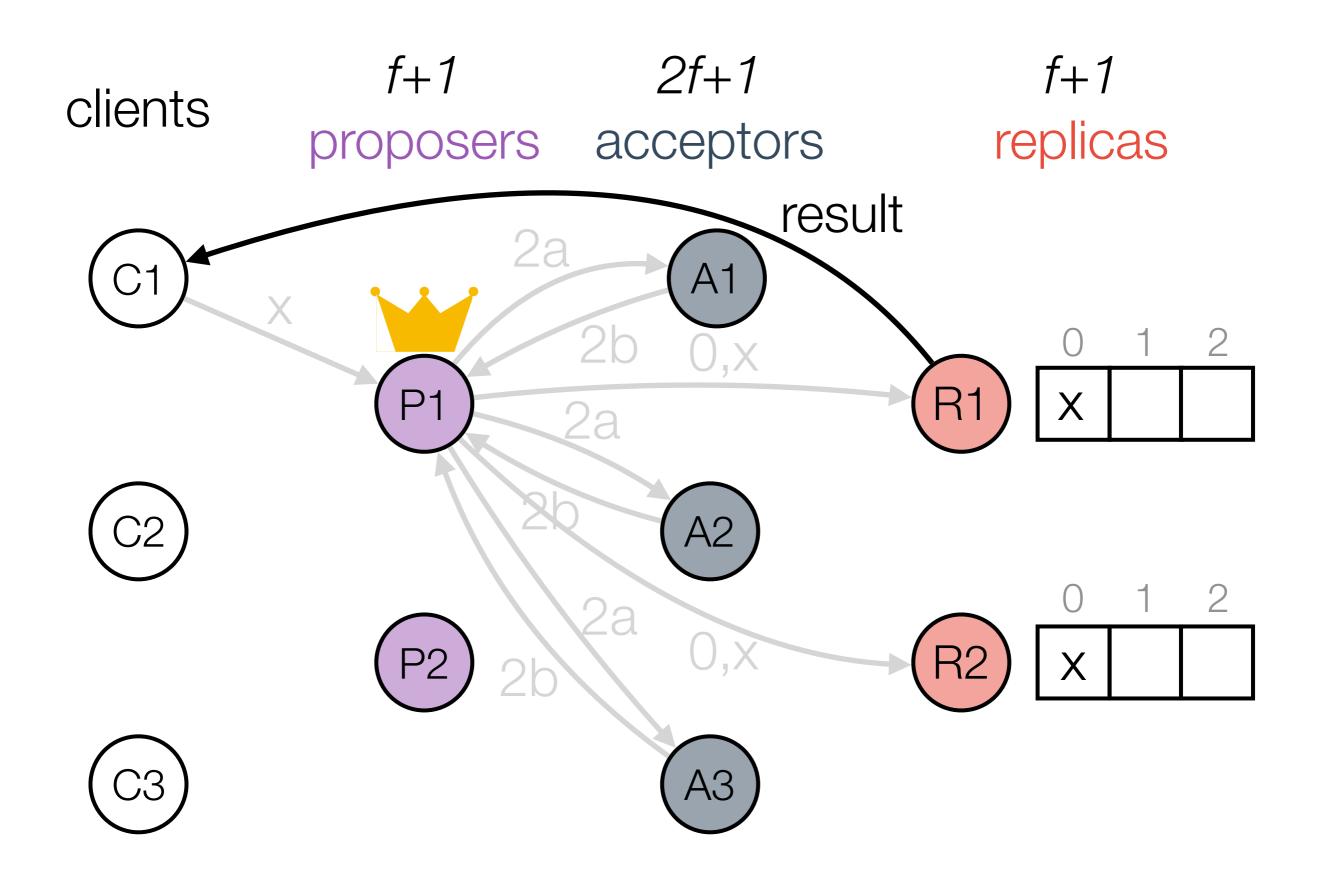


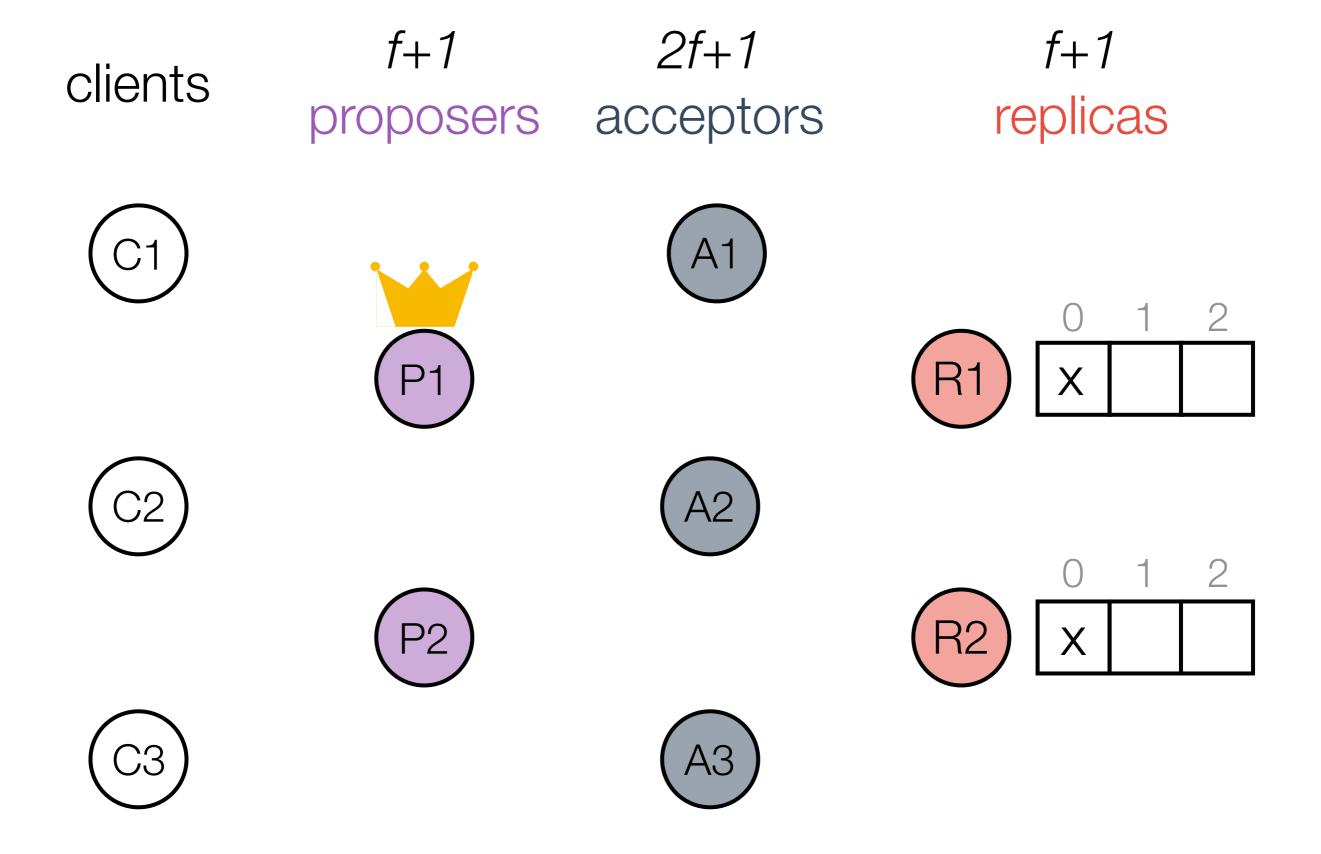
f+1 2f+1 f+1clients replicas proposers acceptors

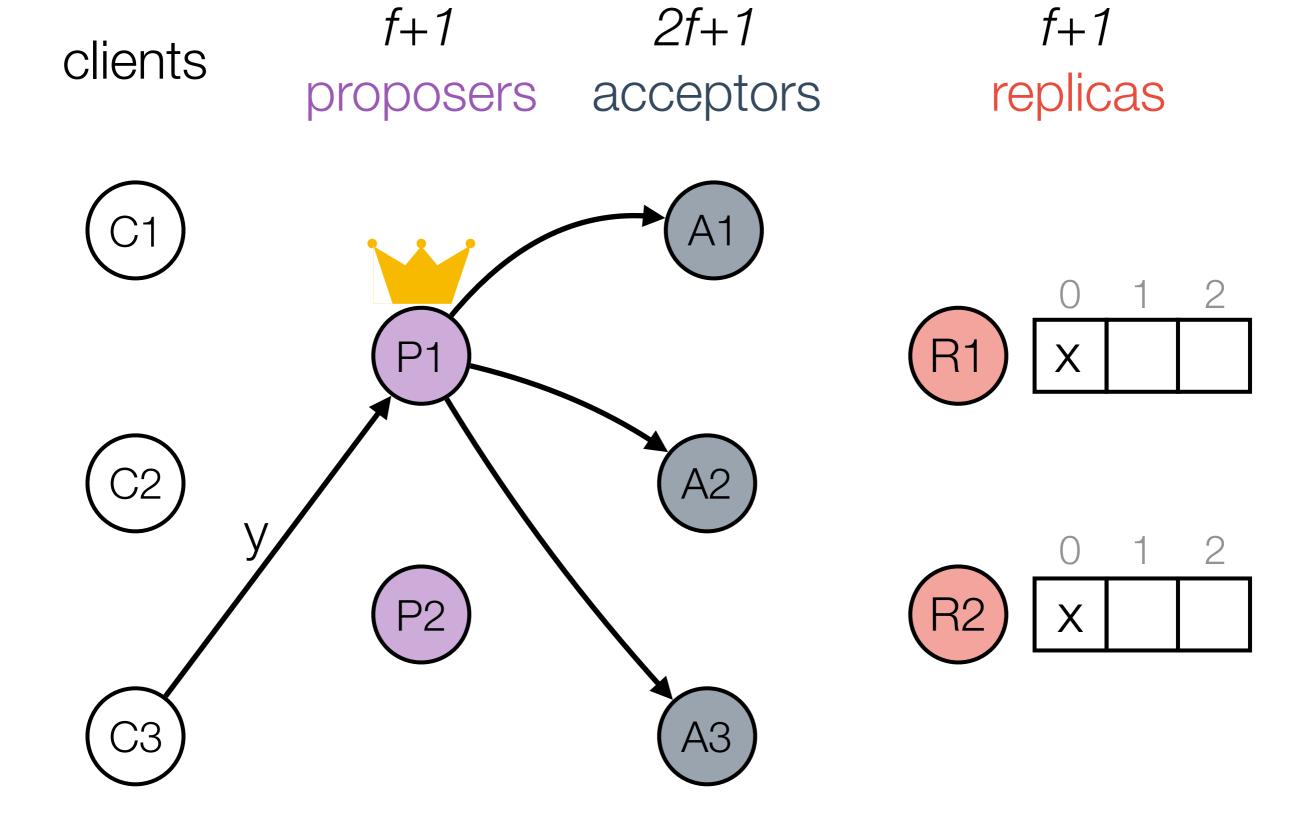
f+1 2f+1 f+1clients replicas proposers acceptors X

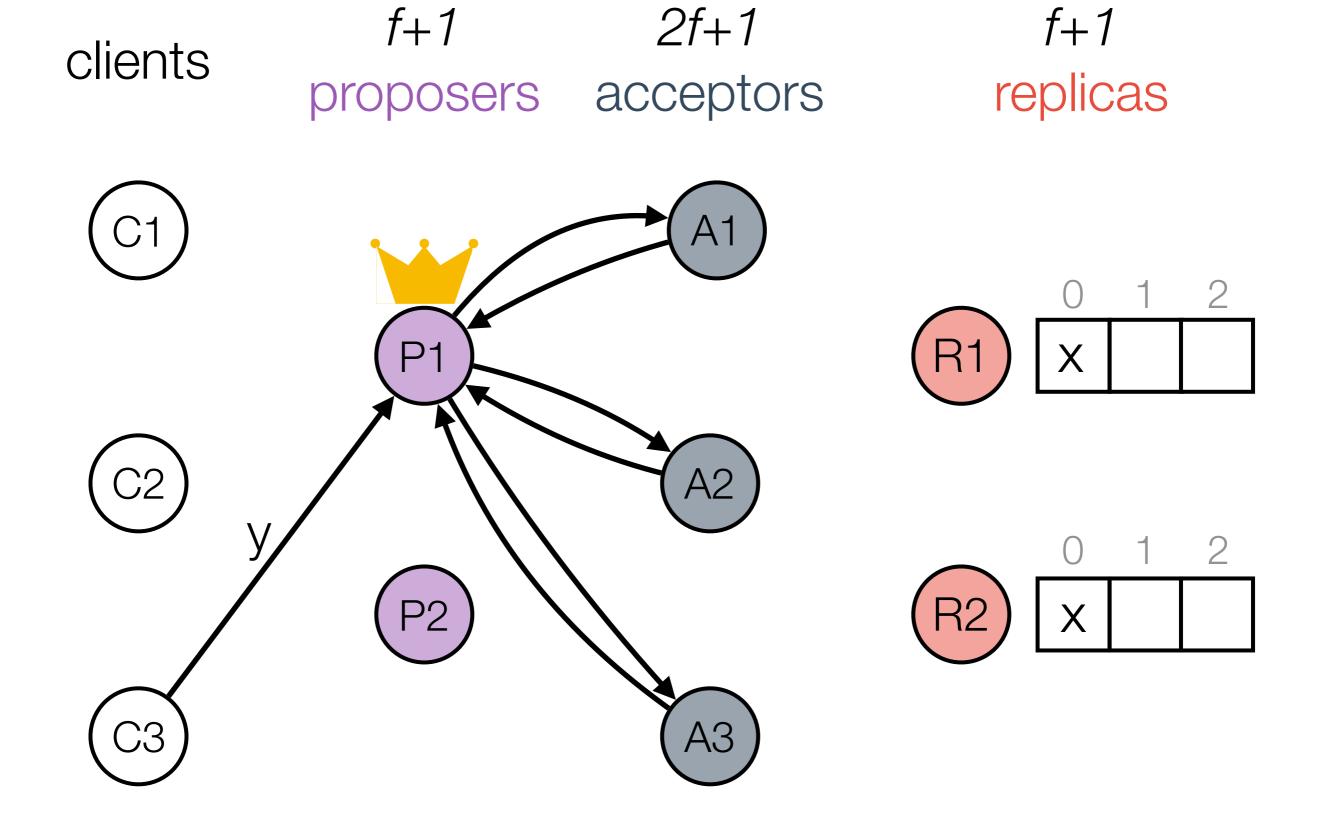


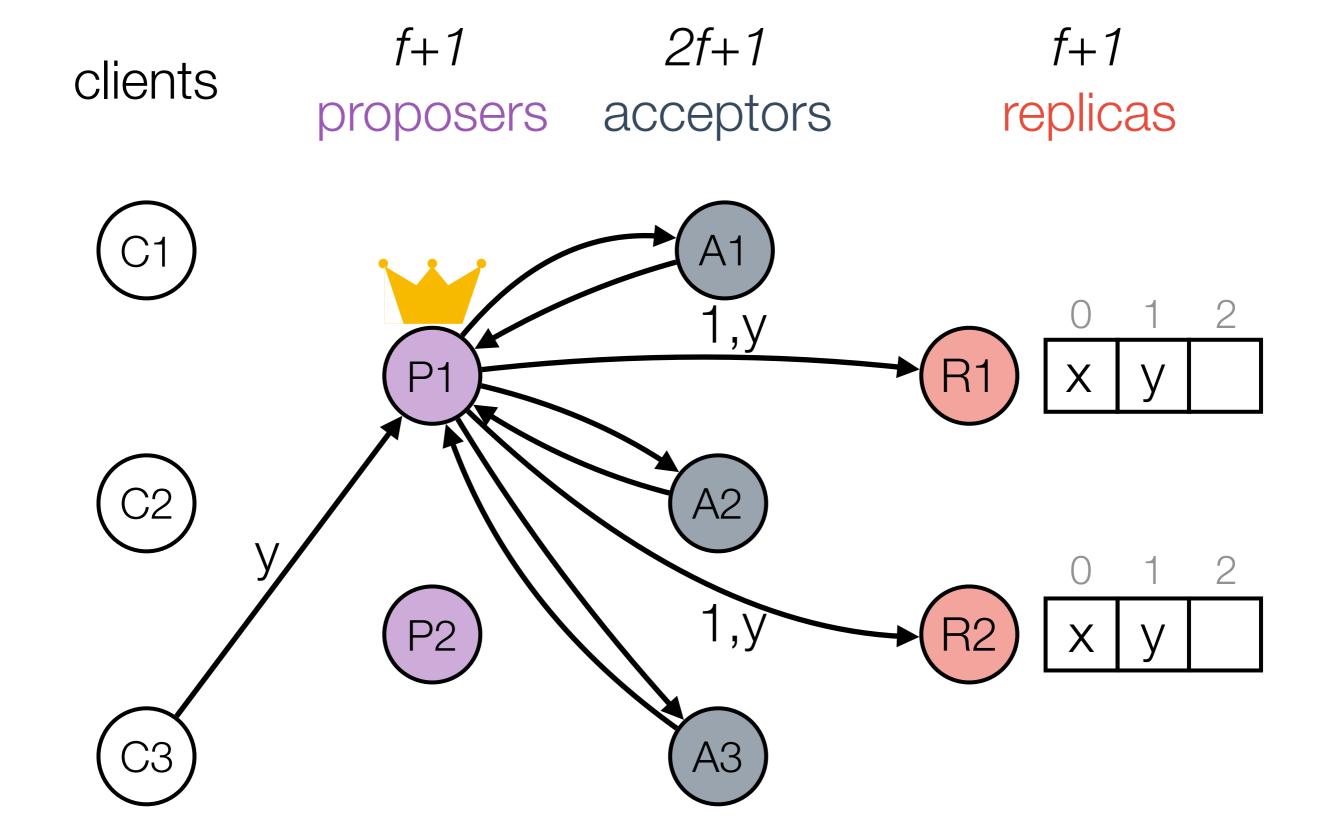


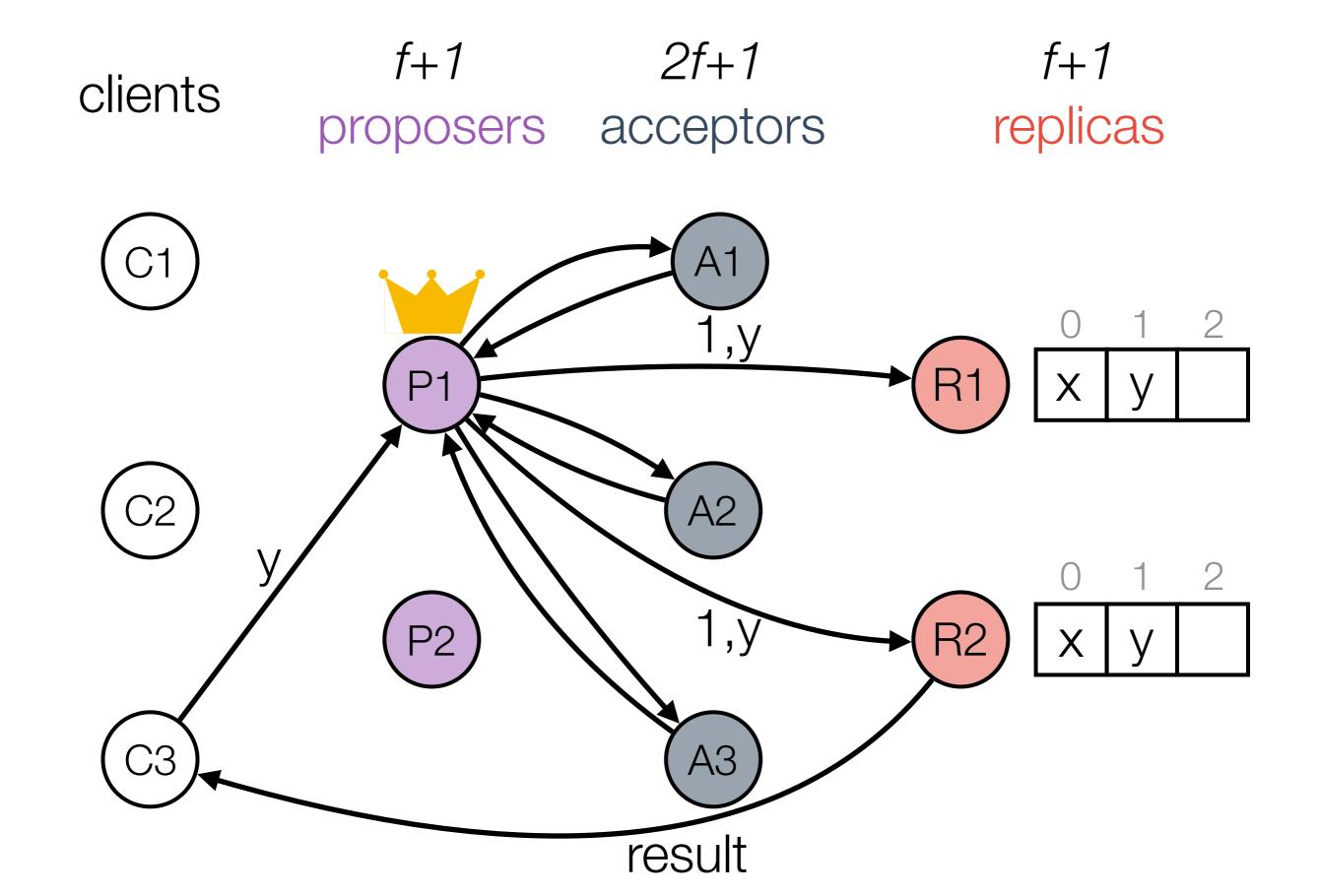










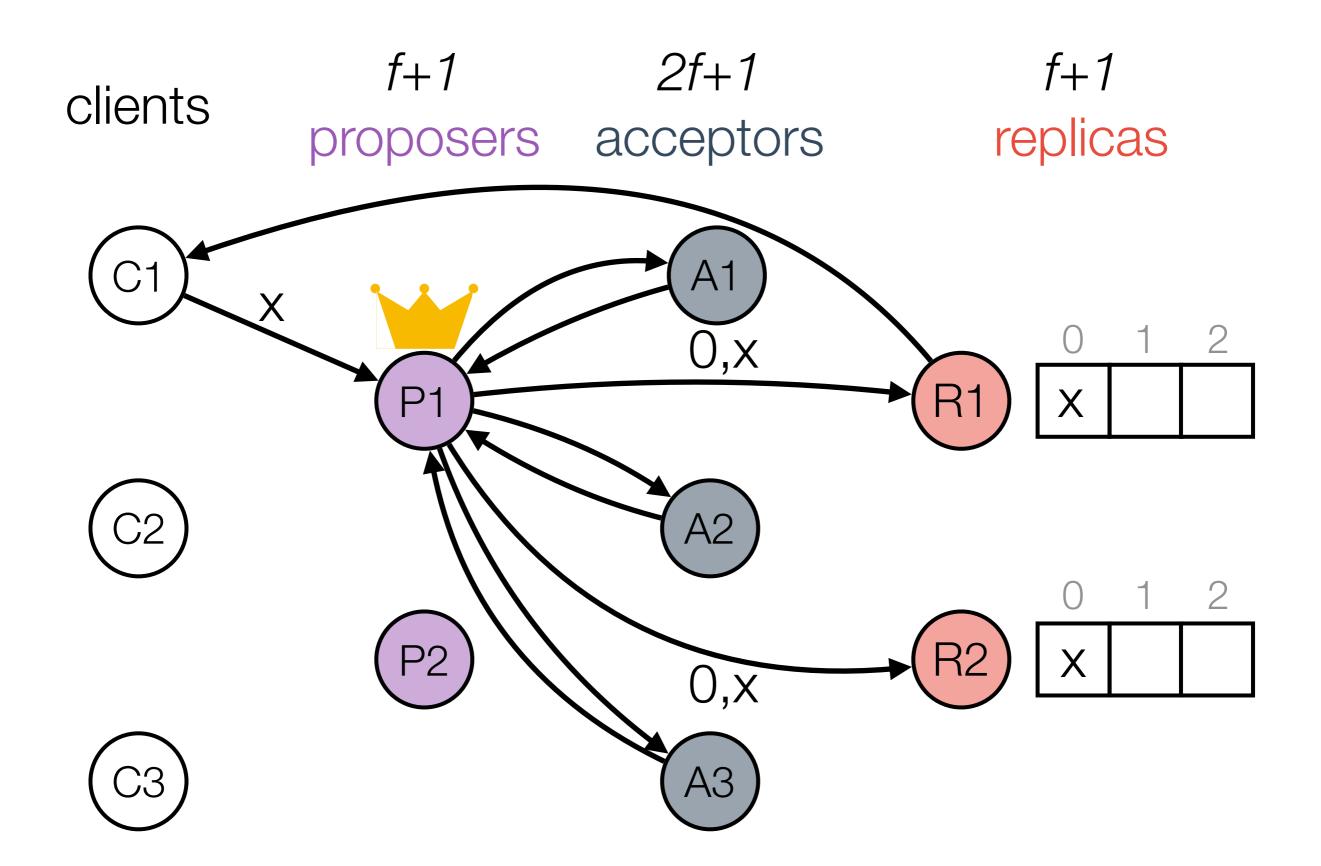


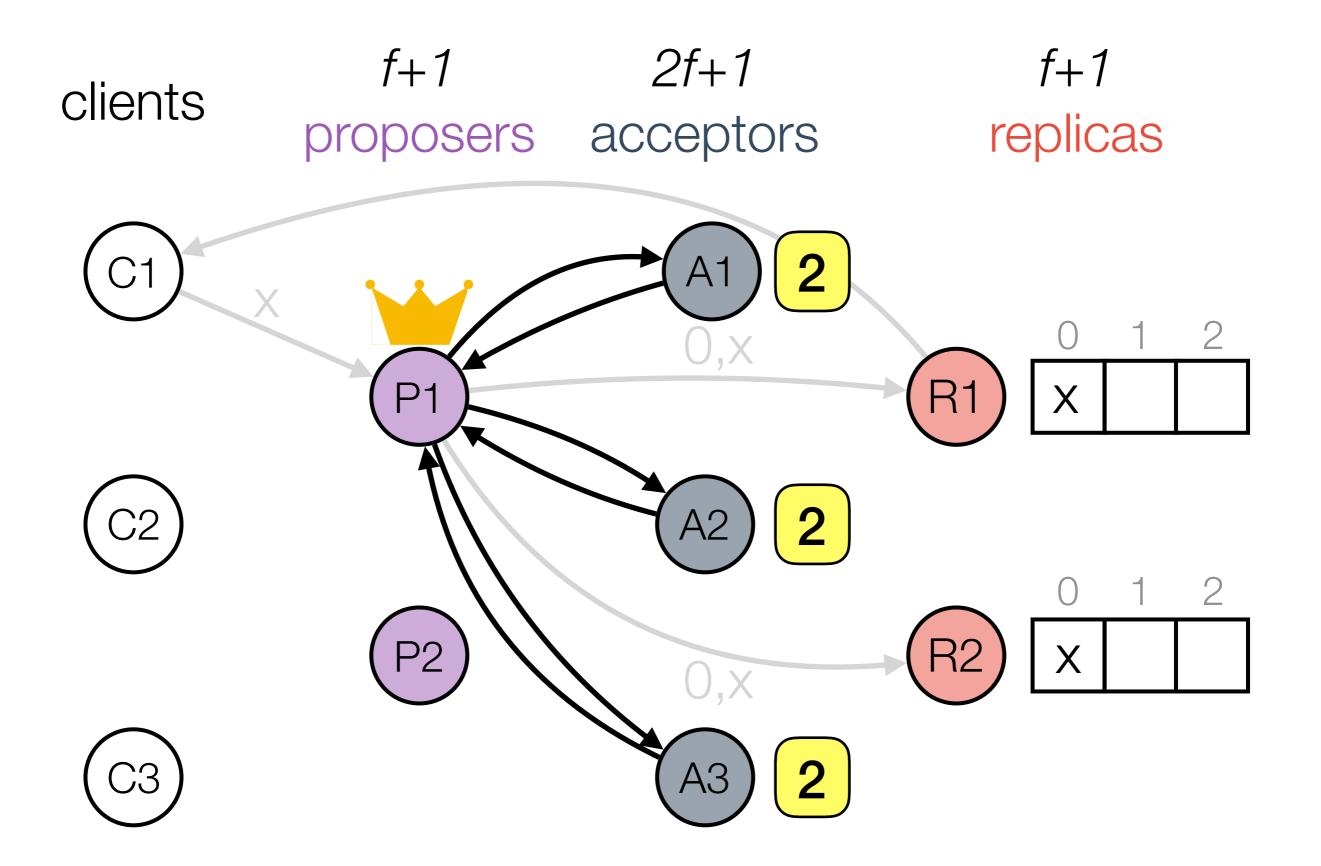
The MultiPaxos leader is a throughput bottleneck

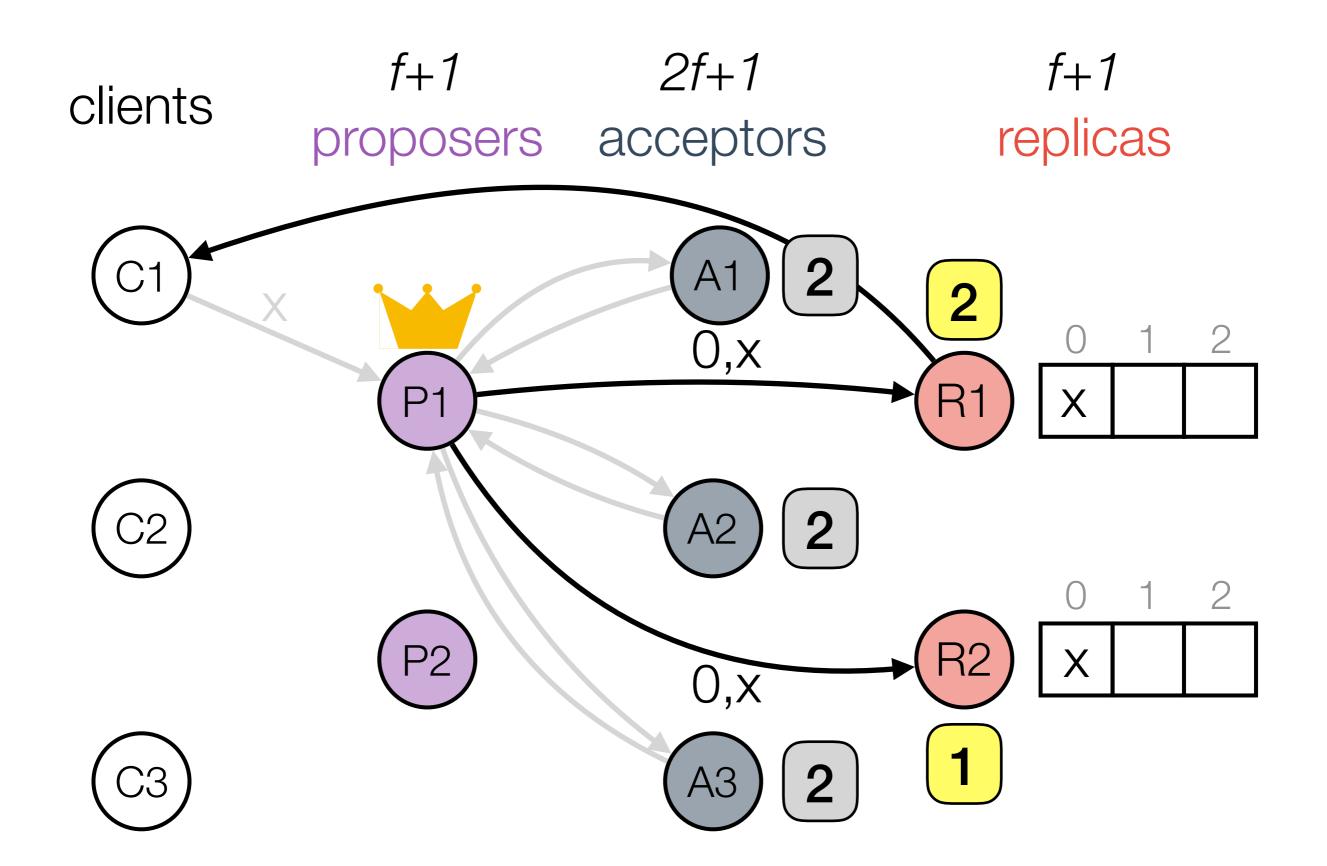
"The leader in Paxos is a bottleneck that limits throughput" - Mencius

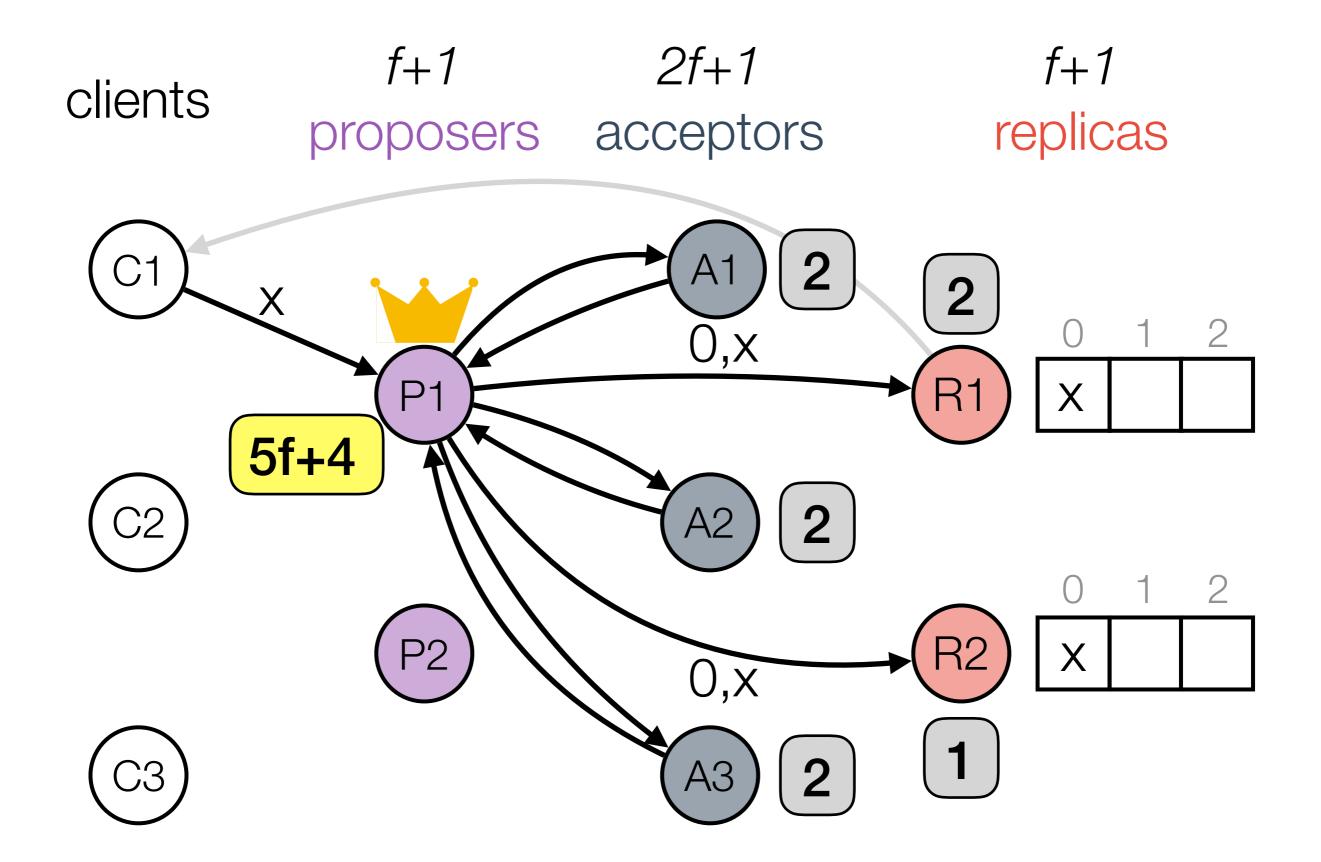
"It impairs scalability by placing a disproportionately high load on the master, which must process more messages than the other replicas" -EPaxos

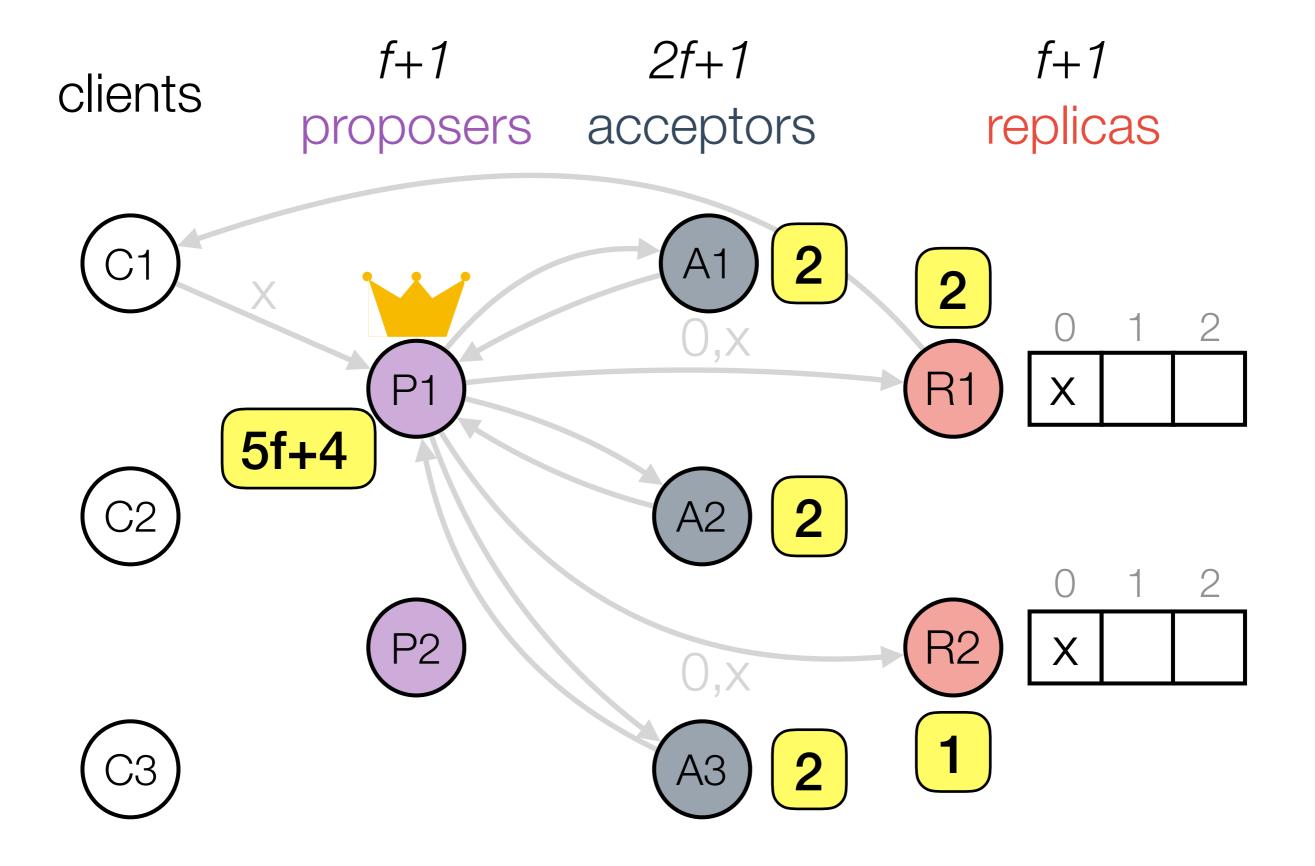
"In practice, [MultiPaxos'] performance is tied to the performance of the leader" -Caesar











Goal: to avoid the leader bottleneck and increase MultiPaxos' throughput.

Generalized Paxos

Generalized Consensus and Paxos

Leslie Lomport

3 March 2004 revised 15 March 2005 corrected 28 April 2005

Microsoft Research Technical Report MSR-TR-2005-33

EPaxos

There Is More Consensus in Egalitarian Parliaments

Iulian Morara, David G. Andersen, Michael Kaminsky Carregic Meilon University and Intel Laba

Abstract

This paper describes the design and implementation of Egalisation Phases (EPacos), a new discripted consensus algorithm based on Phases, EPacos achieves three goals: (I) optimal coment leaves the wide-tase, when infer-sting one and two failures, under realistic conditions; (2) authors lead behaving across all replices (they achieving high throughparts and (3) greatful performance degradation without performance degradation and consideration of the performance of th

ton when replices are slow or used.

Explication benears to one knowledge the first proteased to artiface the previously stated goals efficiently—that is, requiring only a simple majority of replices to be semifacily, using a market of moreogen linear in the number of replices to choose a communant, and committing commands the just one communant are made (one remains) in the common case or after at most two rounds in any case. We prove Egolitamian based a proportionally for a proposition of the properties of the province of the province

1 Introduction

Distributed computing places two main demands on regionation protocols: [1] high demands as regionation protocols: [1] high demands are regionation included accounts the centers. Today's clusters use fault-information charges such as Chathry [4], Revised [21], or Zoo-Kenper [12] for activities including operation sequencing, oscillation, leader election, and resecute discovery. Many flatabases are recreased structured (2, 8).

An important limitation on these systems is that during efficient, failure five approxime, all clients communirate with a single master (or leader) server at all times.

Permission to make slightly or hard argins of past or all off this work for general or alterances out is protect without for protect due that could be not continued or of starboard for protect or control of starboard could be register bear this testice and the full starboard for fire gauge. Copyrights for this purply compresses of this work must be the cond. For all other way, contact the December 1991.

Caparight is held by the Owner-Parliam of .

SEEF 10, Nov. 1-6, 2013, Farmingson, Ferninghamb, USA.

ACCH 2018-1-4003-2048-491311.

May Produce of the 1994-1994-2591338

This optimization—termed "Multi-Facus" for systems based on the Bases personal [15]—in important to addering high throughput in practical systems [7]. Changing the leader sugment involving additional consensus machinisms that substantially reduce performance.

This algorithmic limitation has several important consequences. First, it impairs scalability by placing a disproportionately high load on the master, which most per cess more messages than the other replicas (23). Second, when performing 500 replication, oftents inour additional latency for communicating with a senant master. Third, as we show in this paper, traditional Parces variants are sensitive to both lone term and transient load sollers and notwork delays that increase latency at the master. Fionly. his single-master optimization can have socilability: if the master fails, the system cannot service requests until a new master is elected. Previously proposed tolatters, such as partitioning or using precy servers and underirable because they restrict the type of operations the cluster can perform. For example, a partitioned cluster carnet perform atomic operation; across metitions without using additional techniques.

Egyllistian Pages (Efferon) has no designated leader process. Instead, clients can choose at every step which replica to submit a command to, and in most cases the commend is committed without interfering with other concurrent commands. This allows the system to evenly distribute the lead to all replieus, climinating the first betthereok identified above the, having one server that must be on the critical path for all communication). EPence's flexible load distribution better handles permanently or transimily slaw usels, as well as the binney beterageneity crossed by geographical distribution of replicas, this substantially reduces both the median and tail commit laimes. Finally, the system can provide higher systability and higher performance under fullures because there is no transient interruption caused by leader efact one these is no leader, and hence, no need for leader election, as long as more than half of the replicas are available.

We begin by reviewing the core Fasos algorishm and the intration behind Egolization Percecin Section 2. We then describe several Perces variants that reduce carcheal or commit latency in Section 3. Throughout the paper we compare extendedly against Modi-Percecinal towards

Generalized Paxos



Leslie Lomport

J March 2004 revised 15 March 2005 corrected 28 April 2005

Microsoft Research Technical Report MSR-TR-2005-35

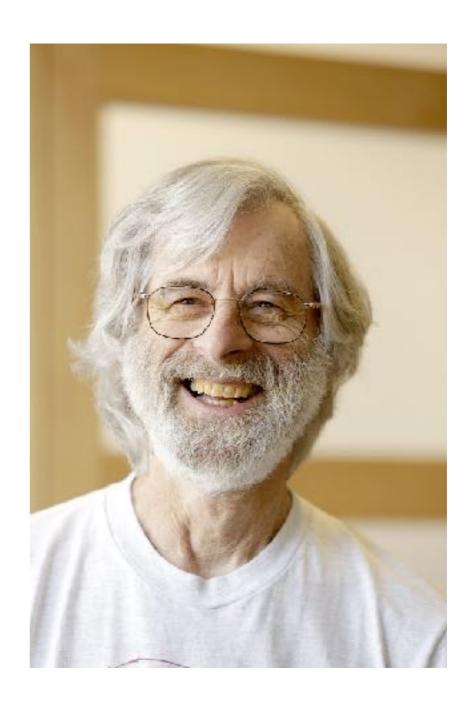
Generalized Paxos

Generalized Consensus and Paxos

Leslie Lomport

J March 2004 revised 15 March 2005 corrected 28 April 2005

Microsoft Research Technical Report MSR-TR-2005-35



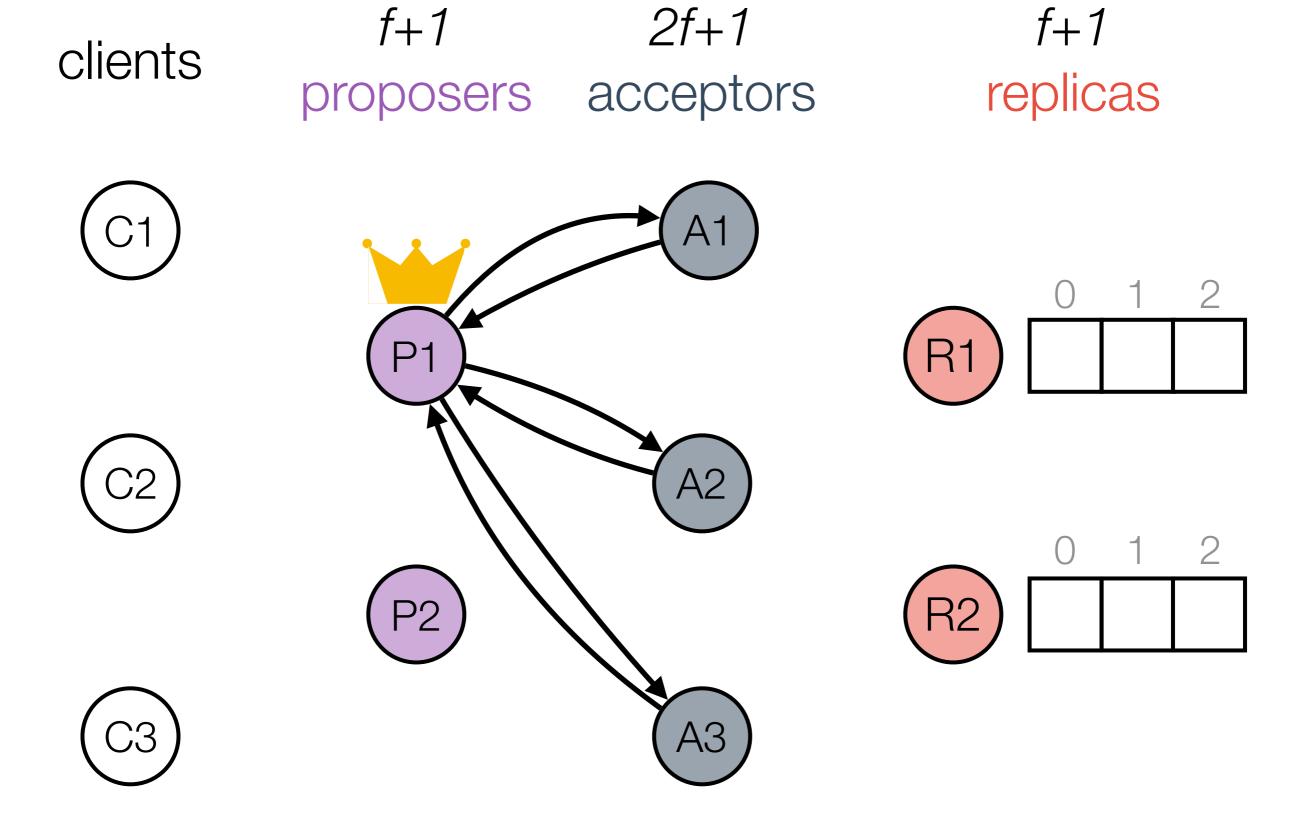
To increase MultiPaxos' throughput, we must trade off complexity?

Key Insight: The MultiPaxos leader has two independent responsibilities.

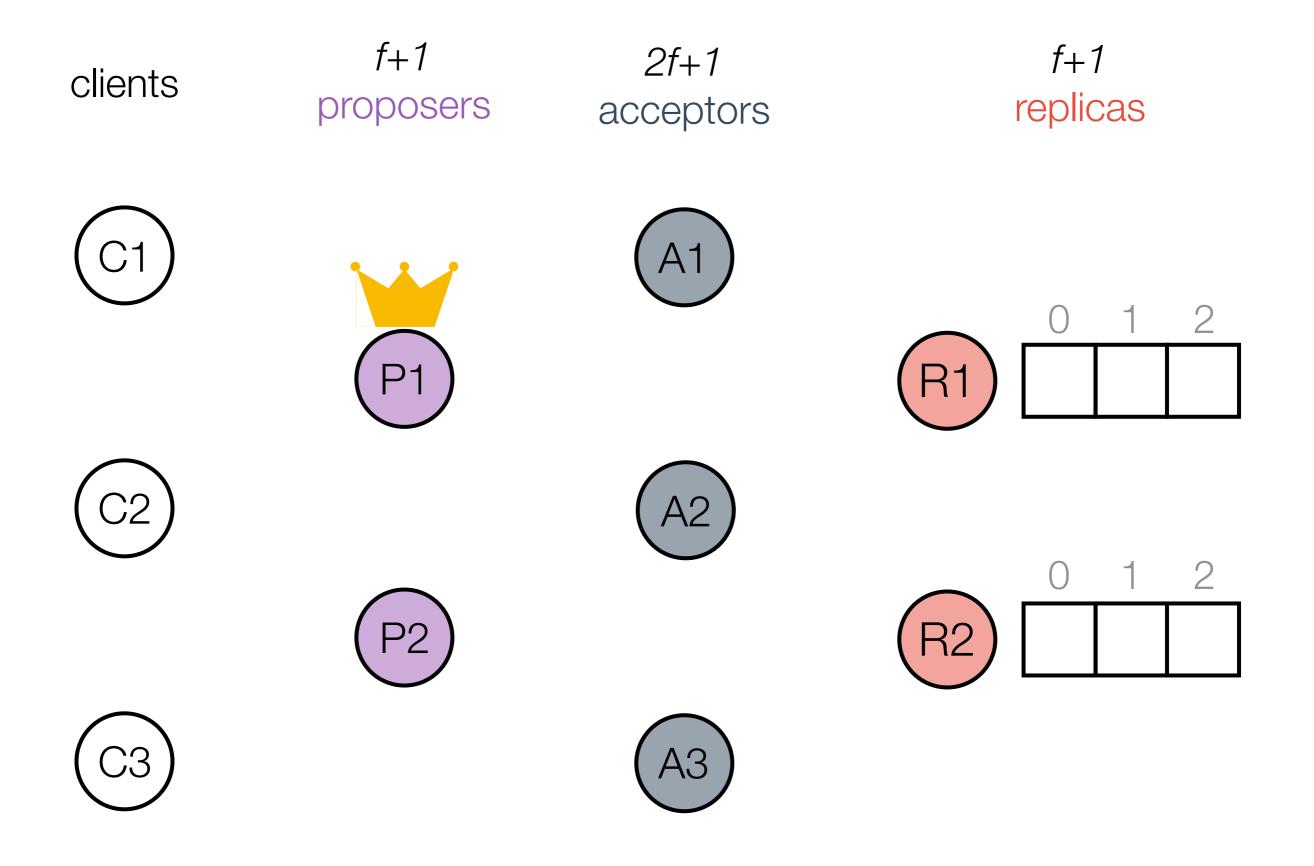
f+1 2f+1 f+1clients replicas proposers acceptors X

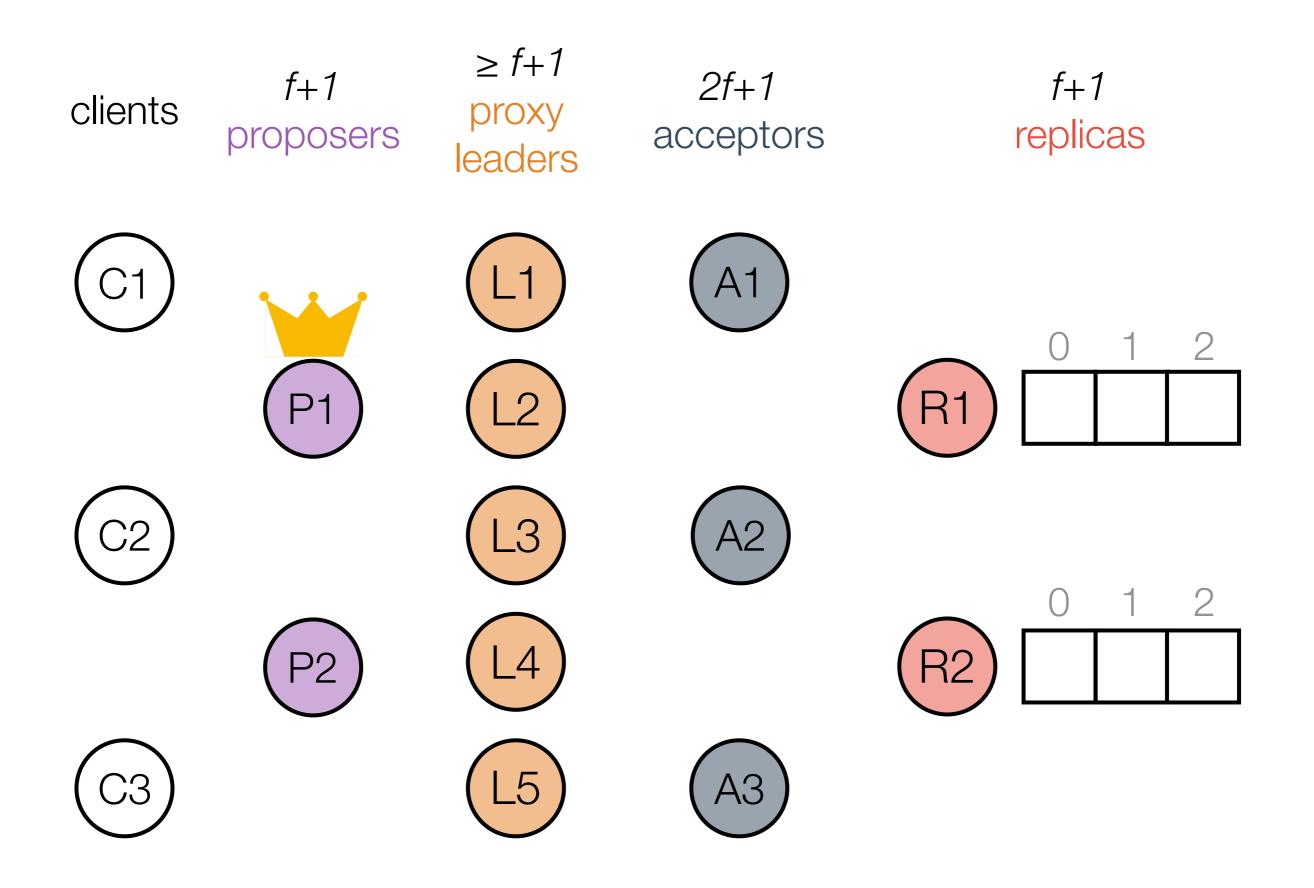
f+1 2f+1 f+1clients replicas proposers acceptors

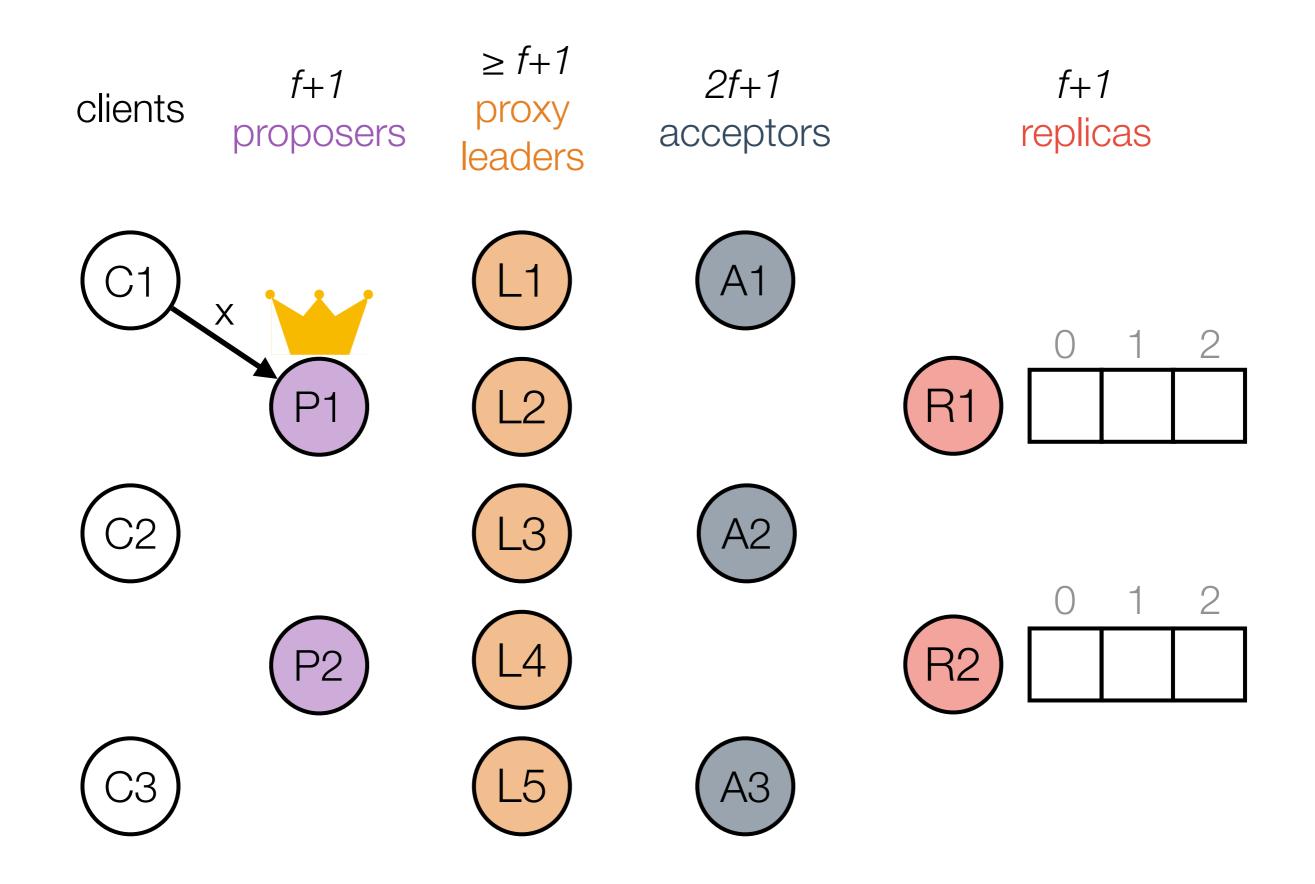
f+1 2f+1f+1clients replicas proposers acceptors

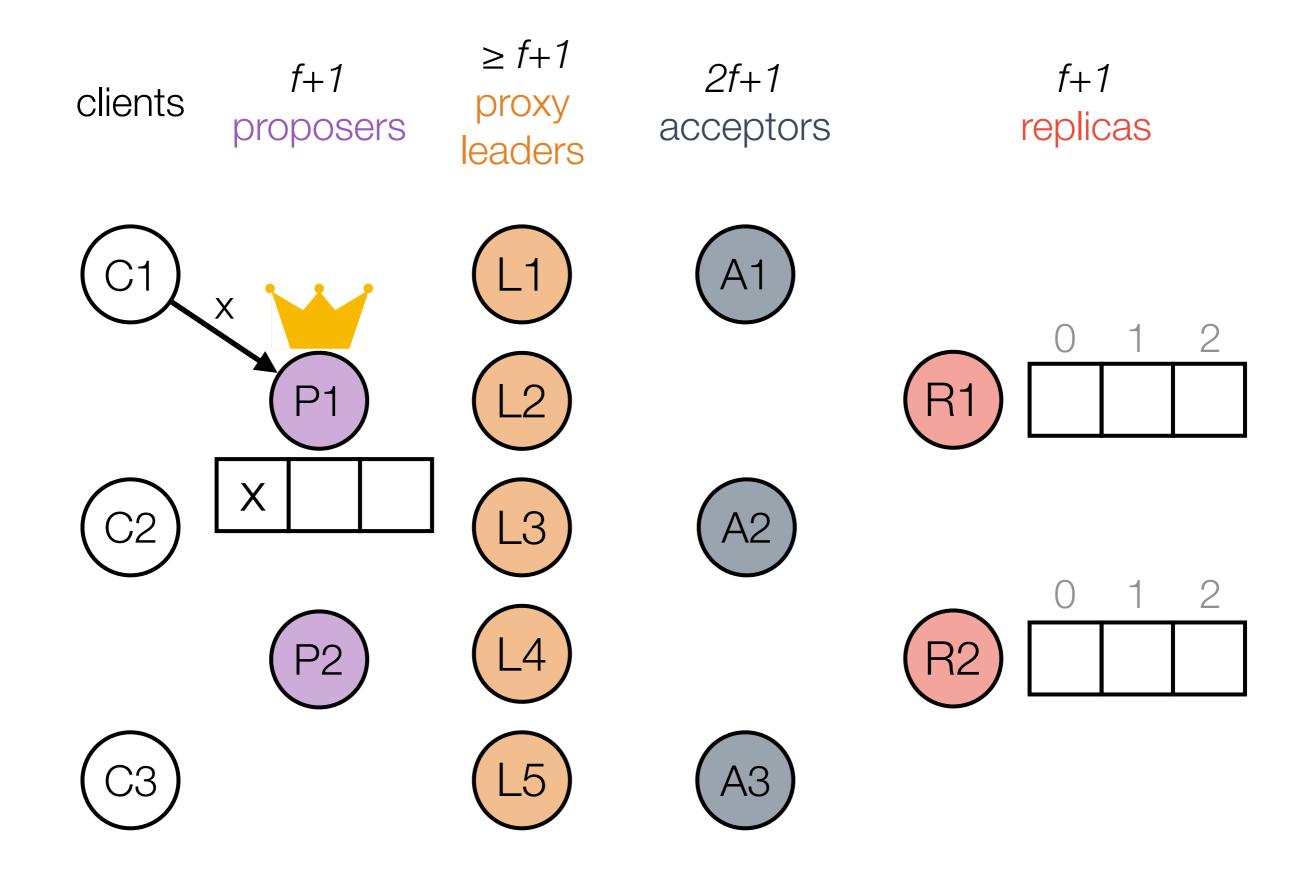


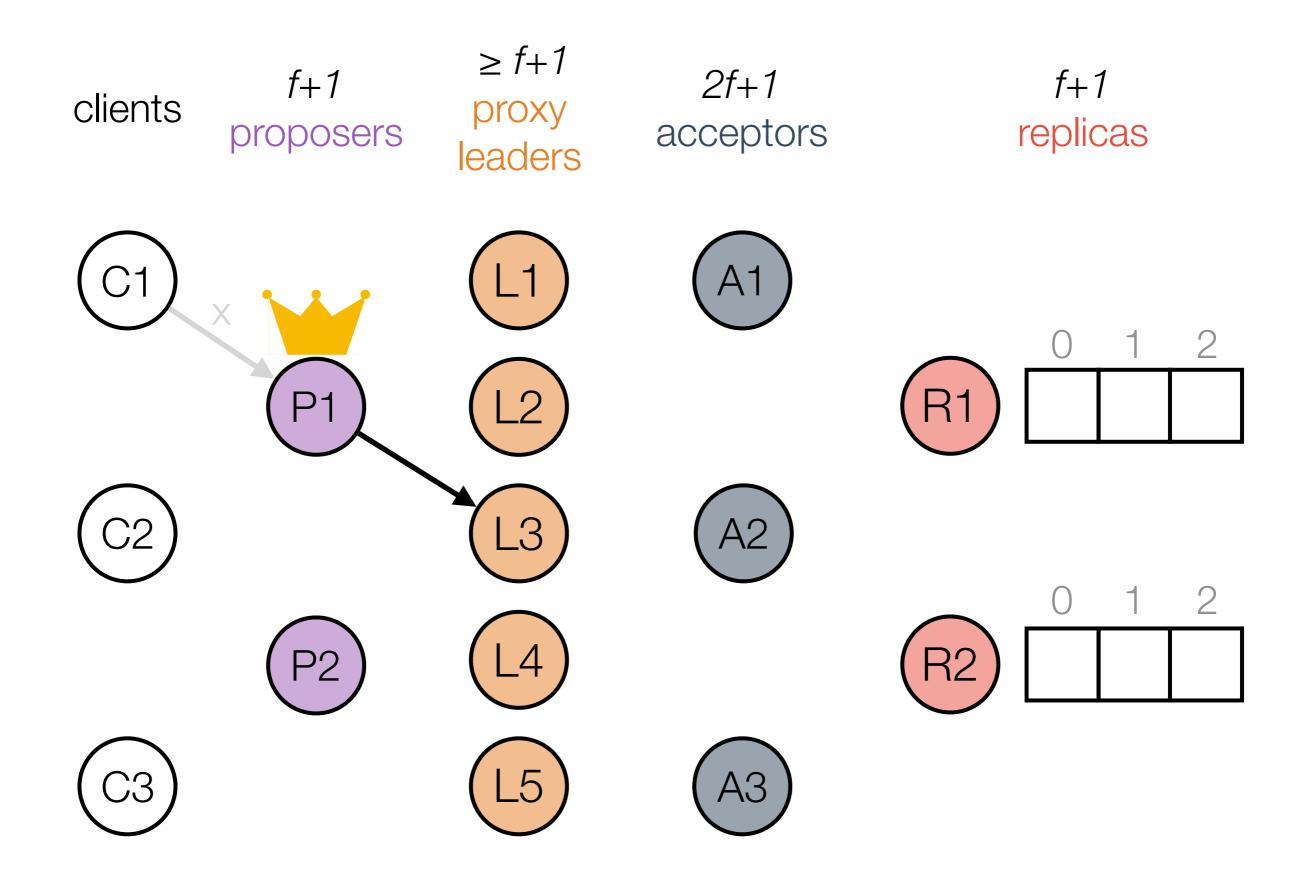
Key Insight: Decouple the MultiPaxos leader's two responsibilities

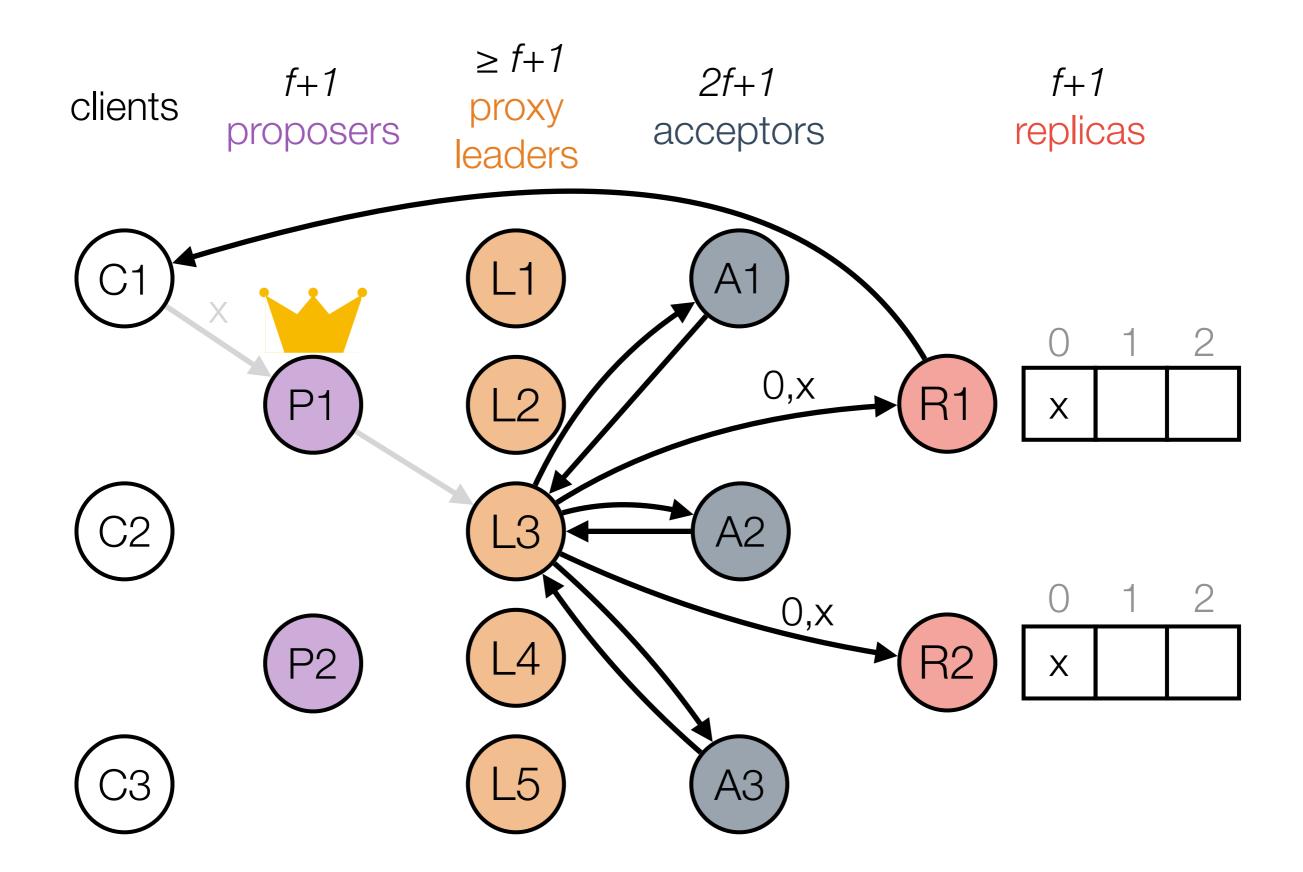




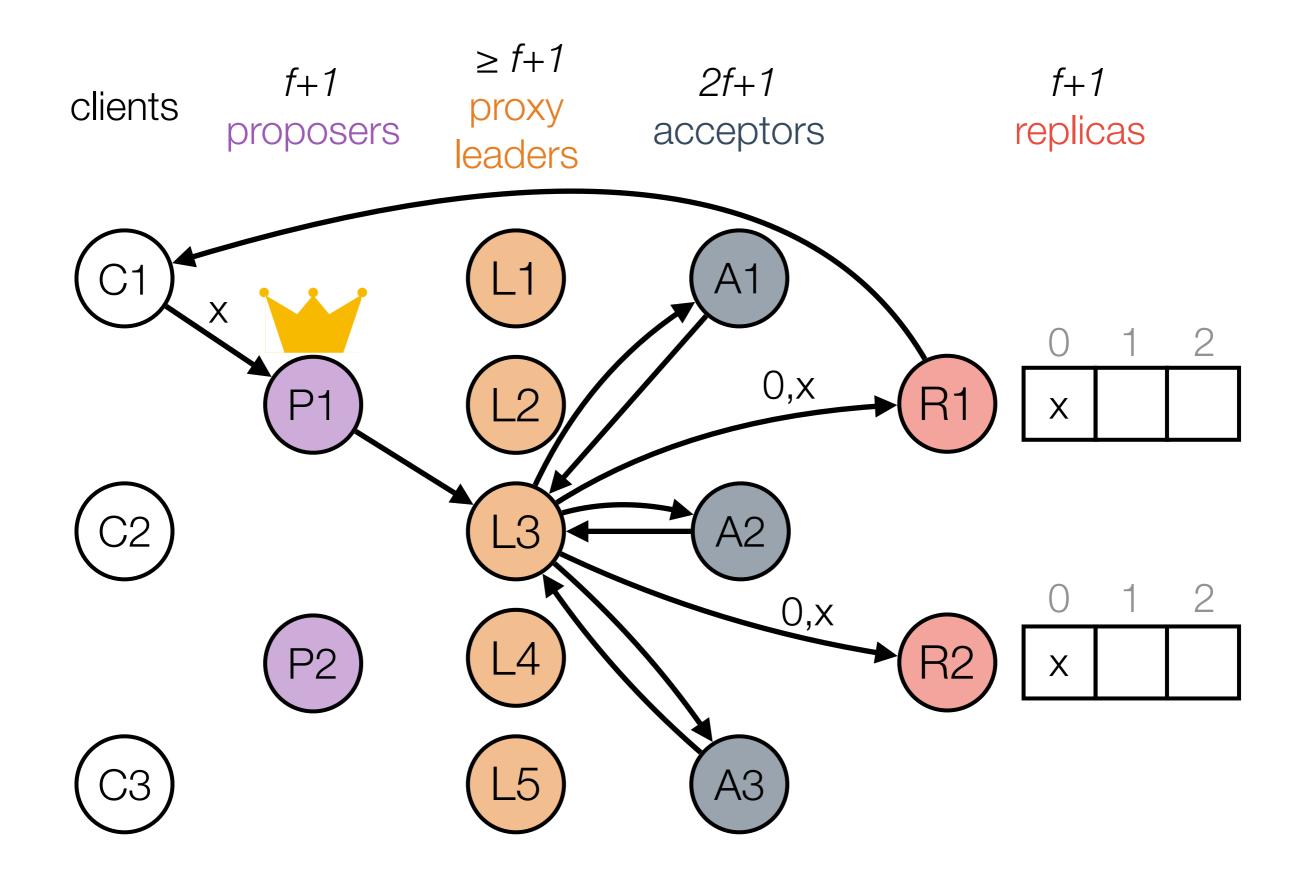


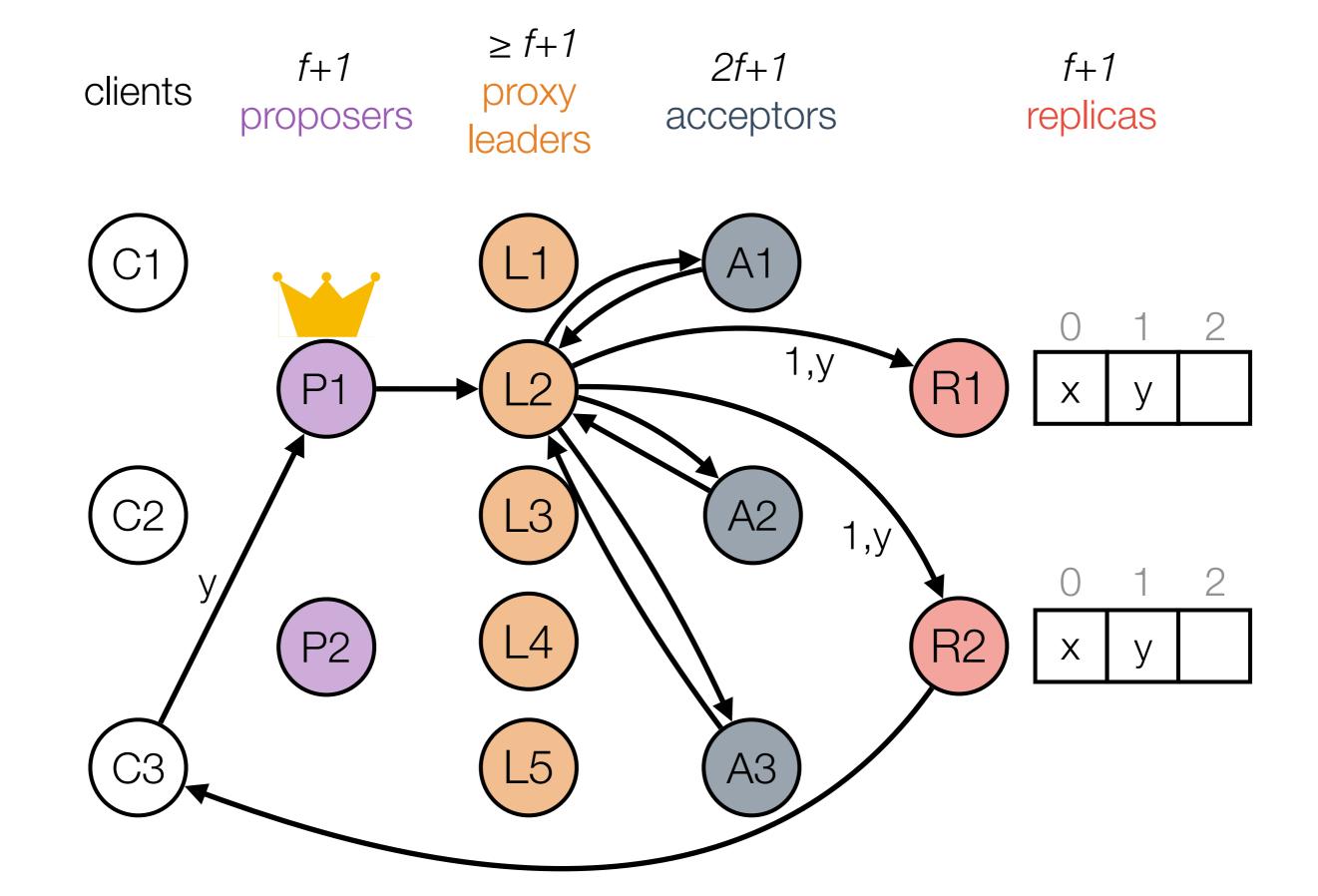


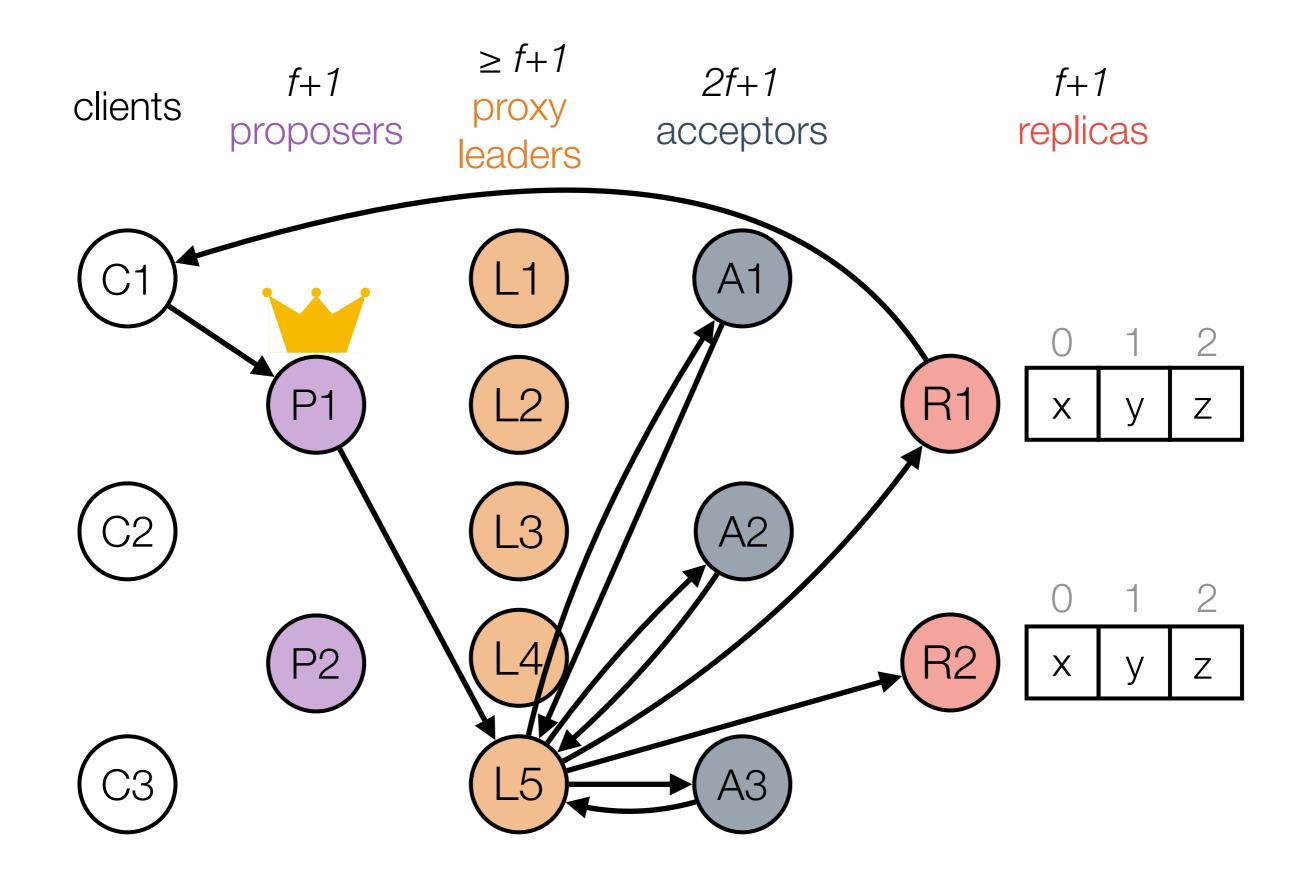


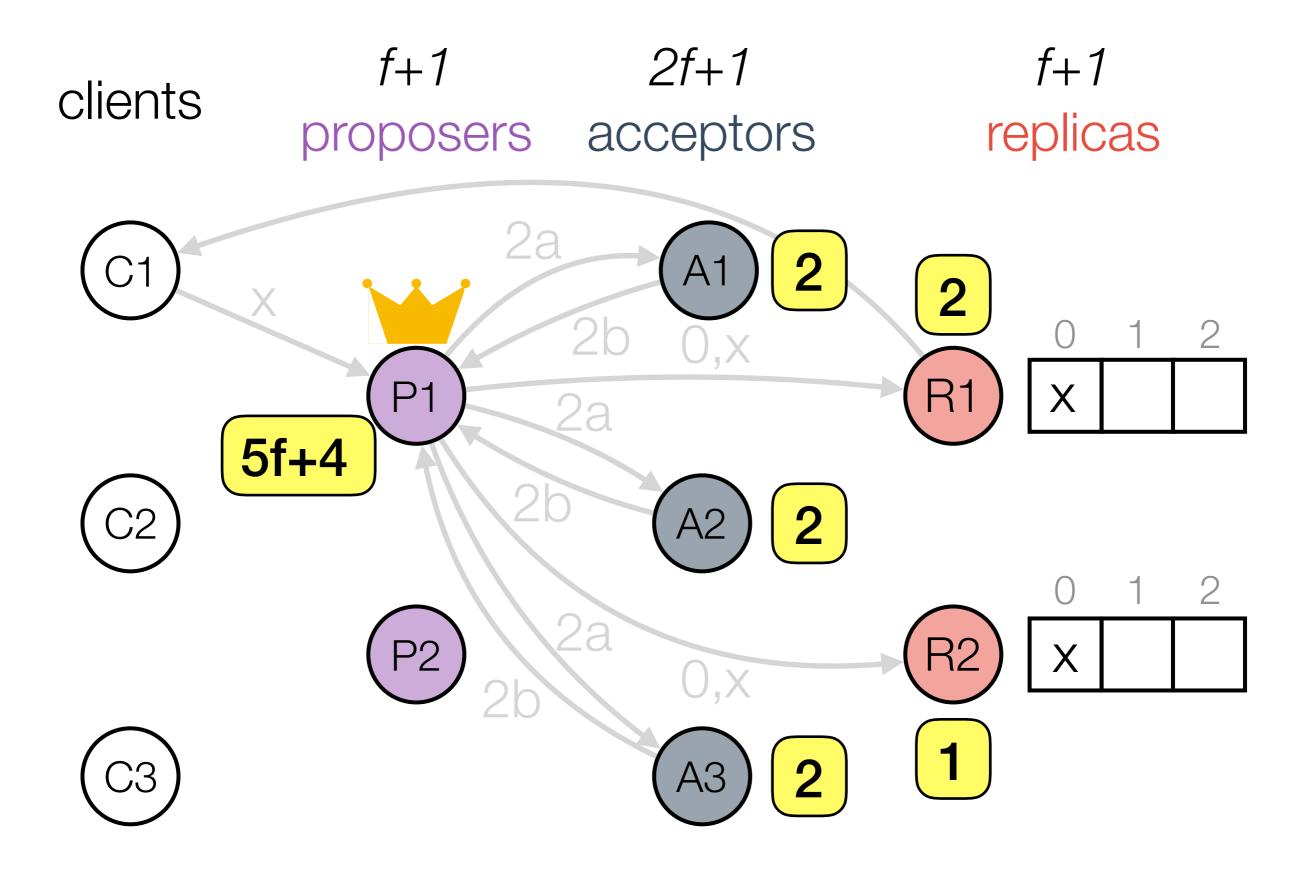


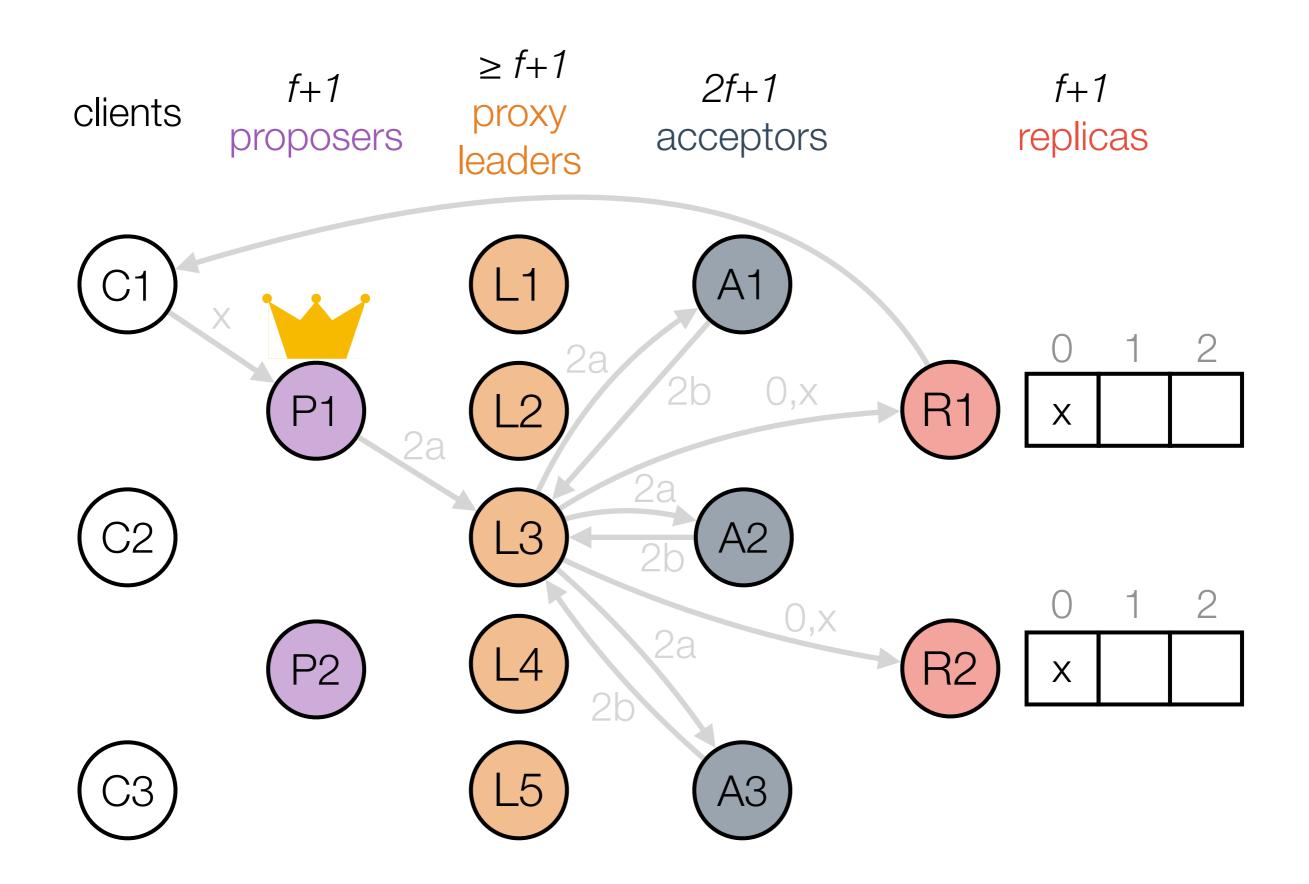
Key Insight: Scale up the proxy leaders

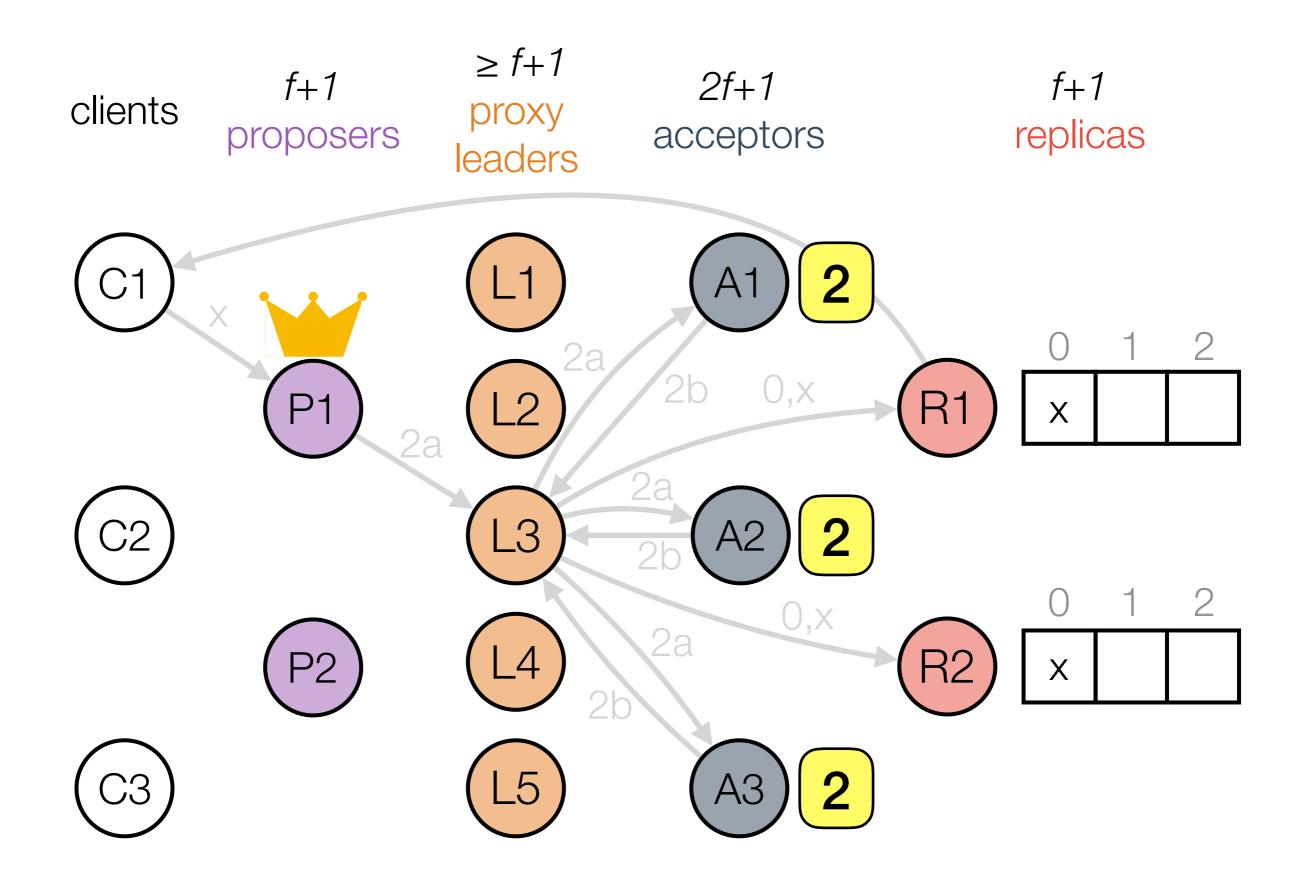


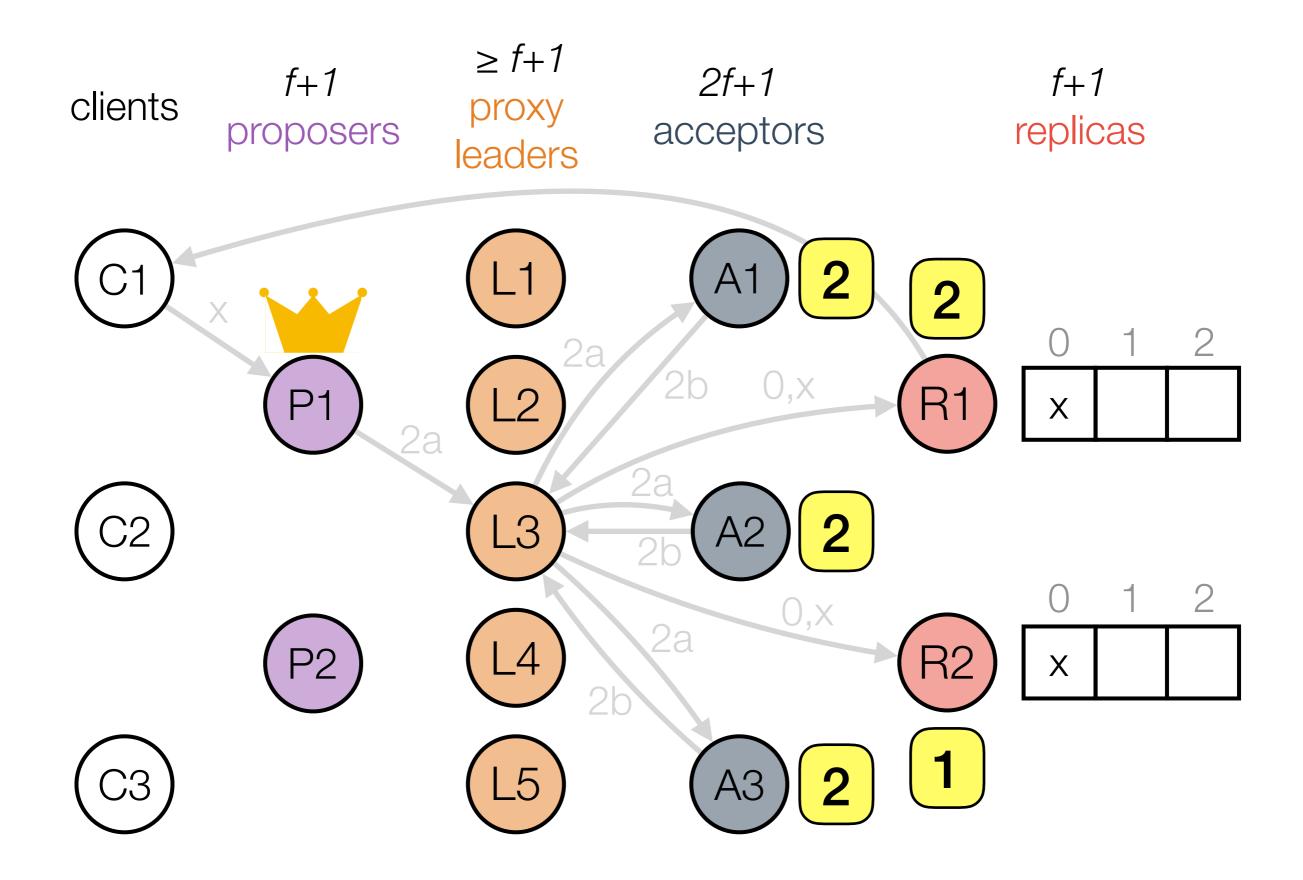


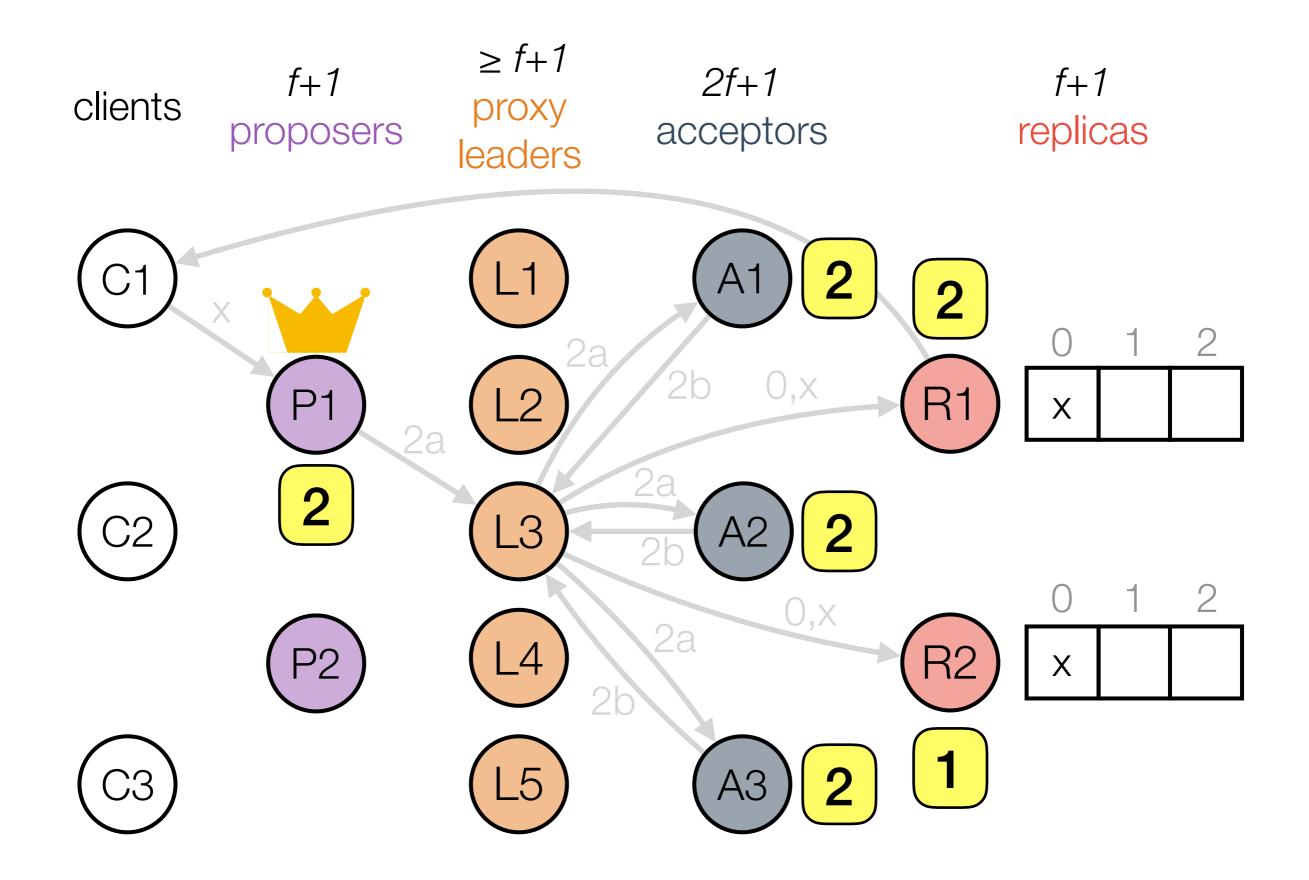


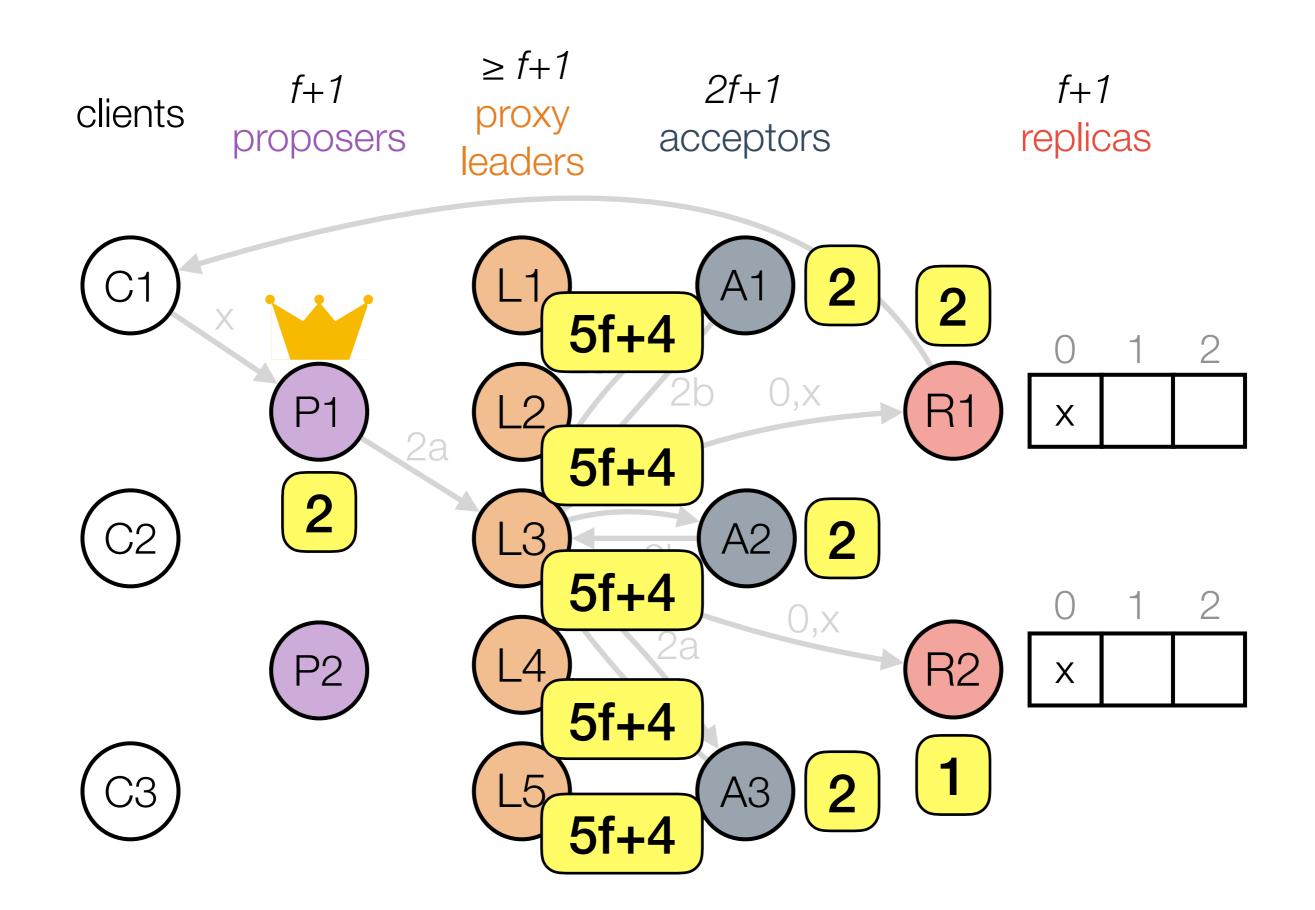




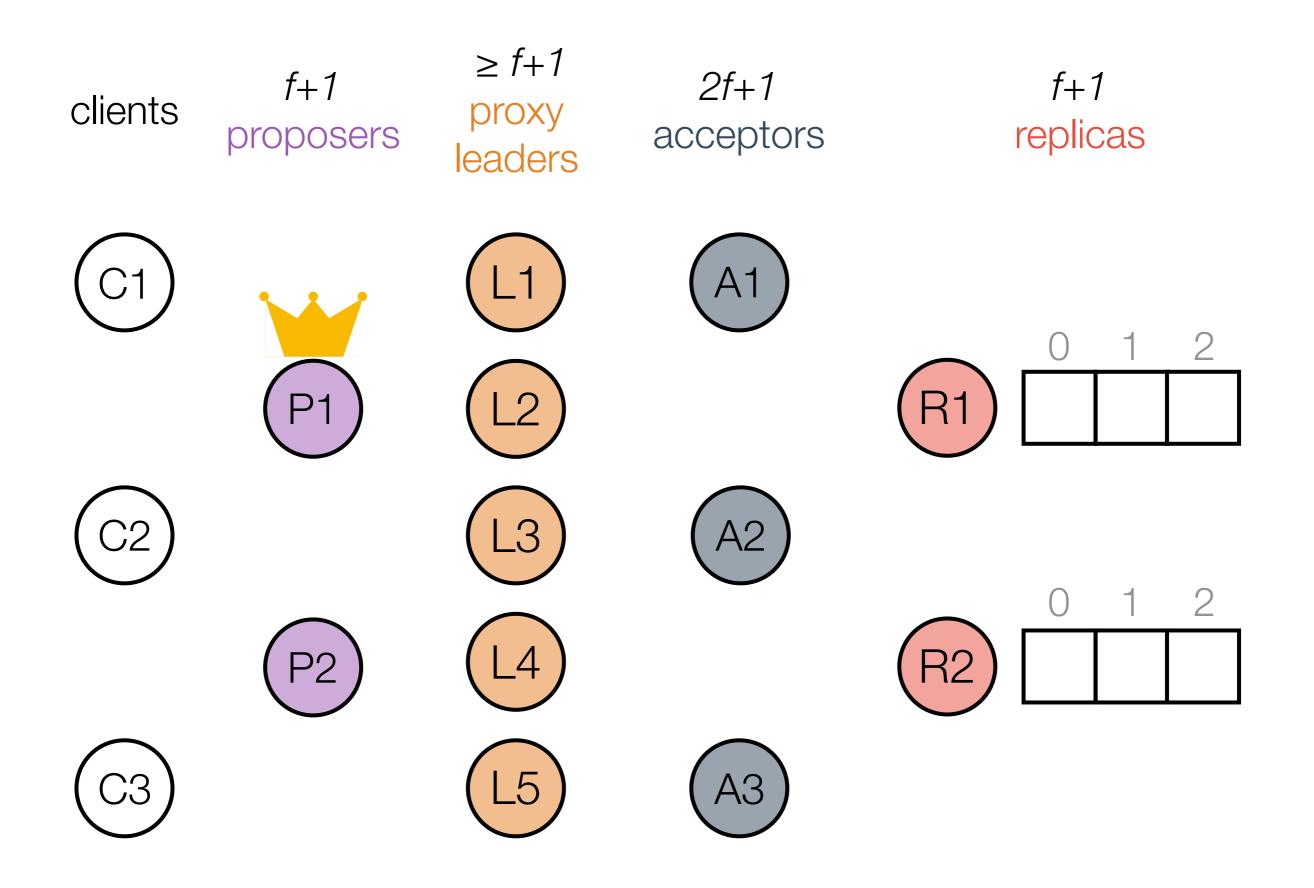


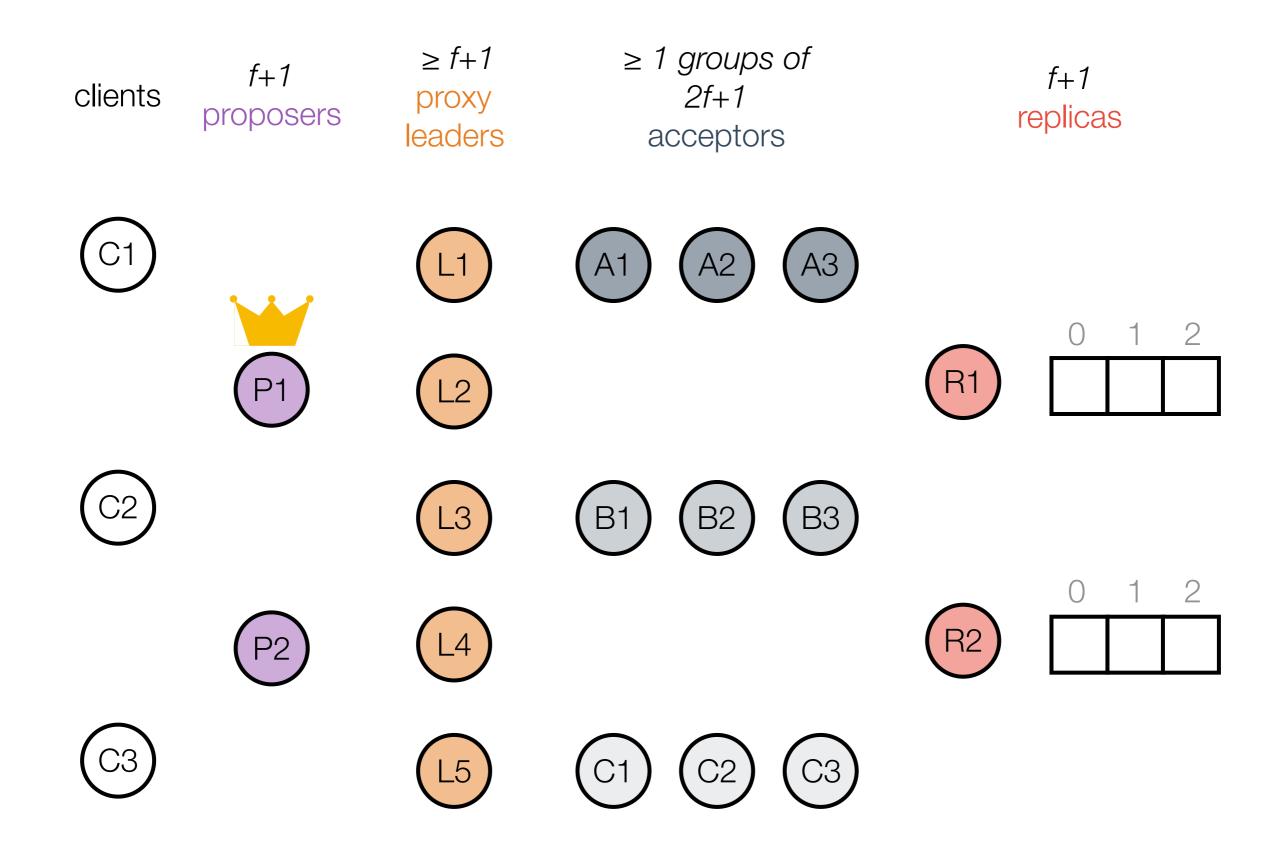


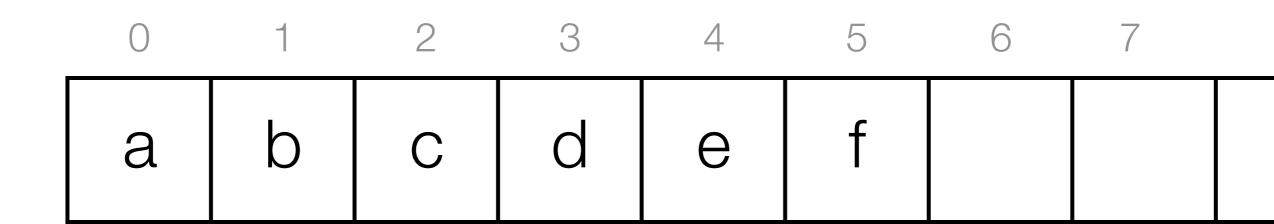




Compartmentalization: Decouple and scale







0 1 2 3 4 5 6 7

a b c d e f

A1

(A2)

(A3)

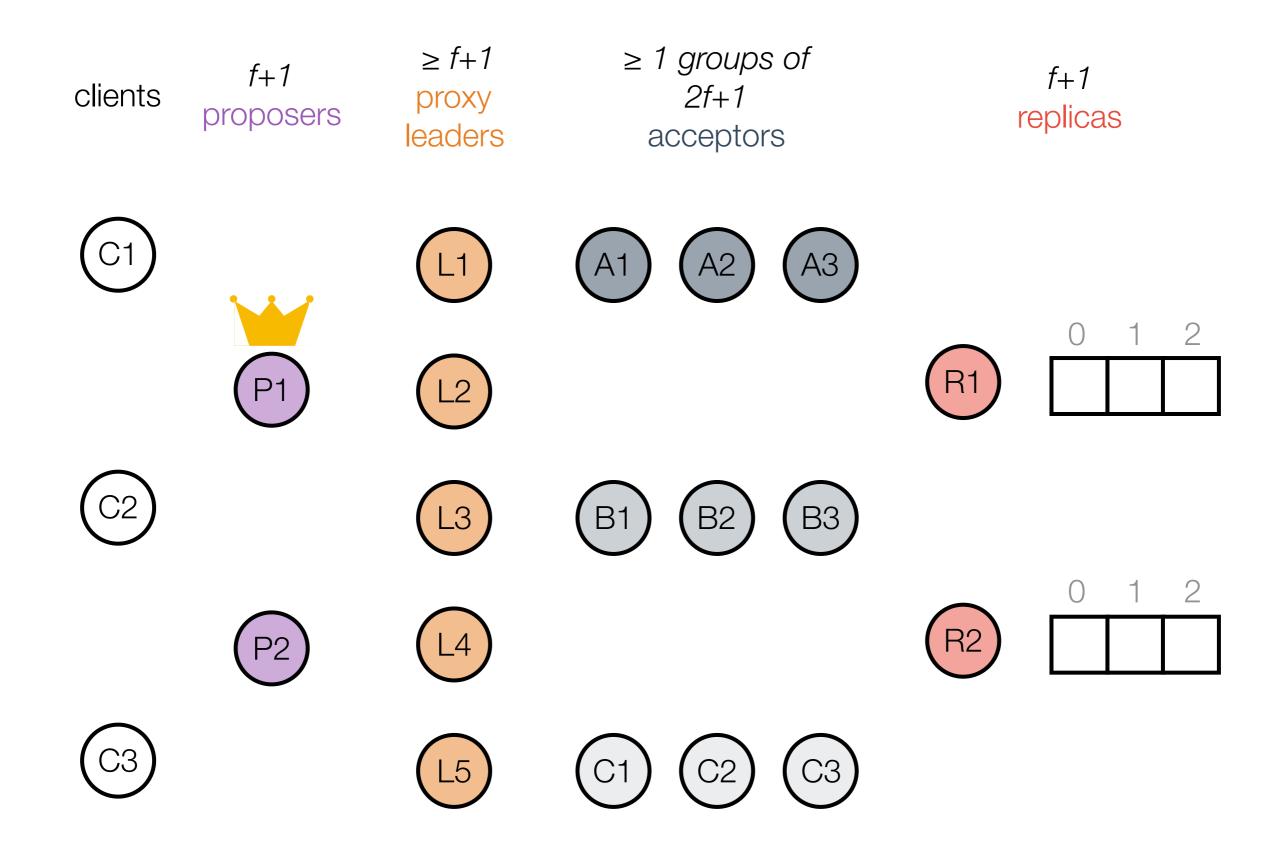
0 1 2 3 4 5 6 7

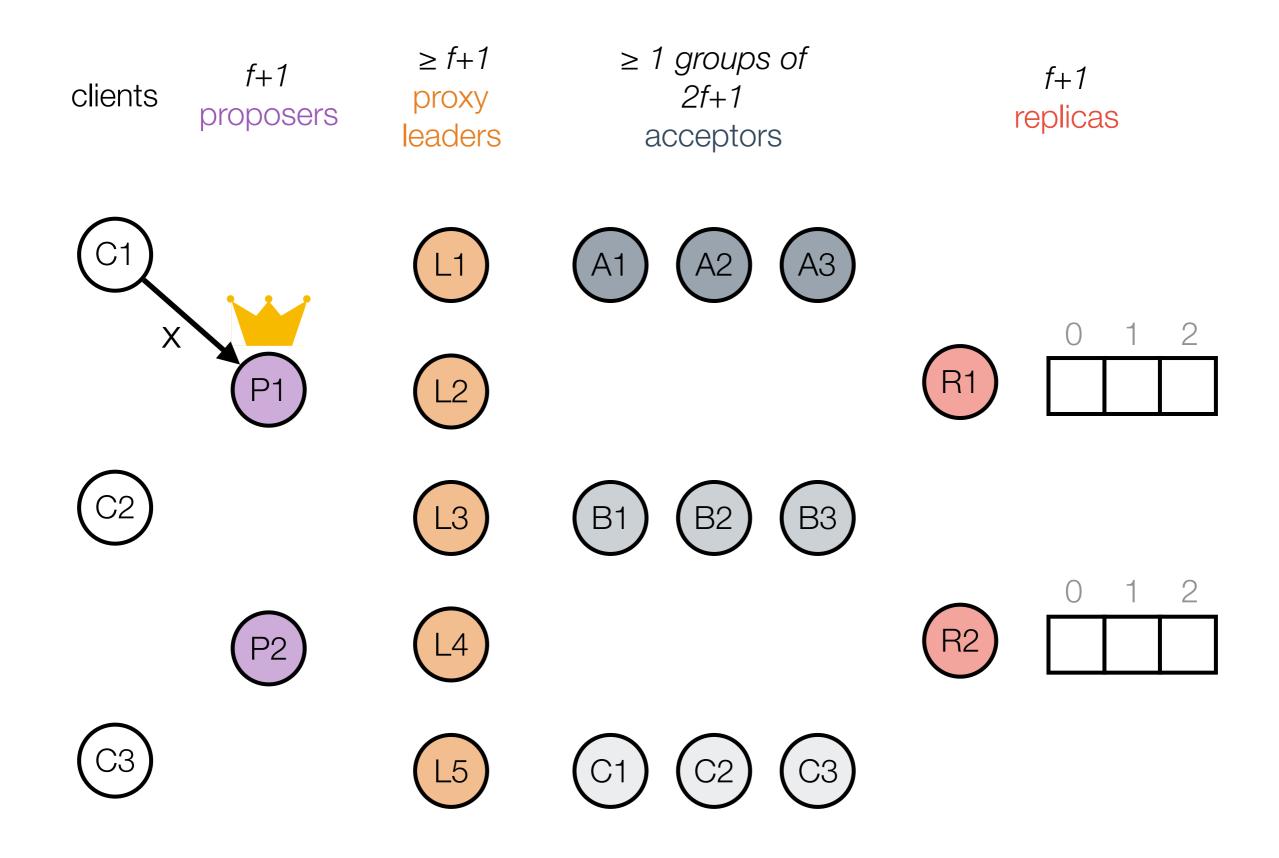
a b c d e f

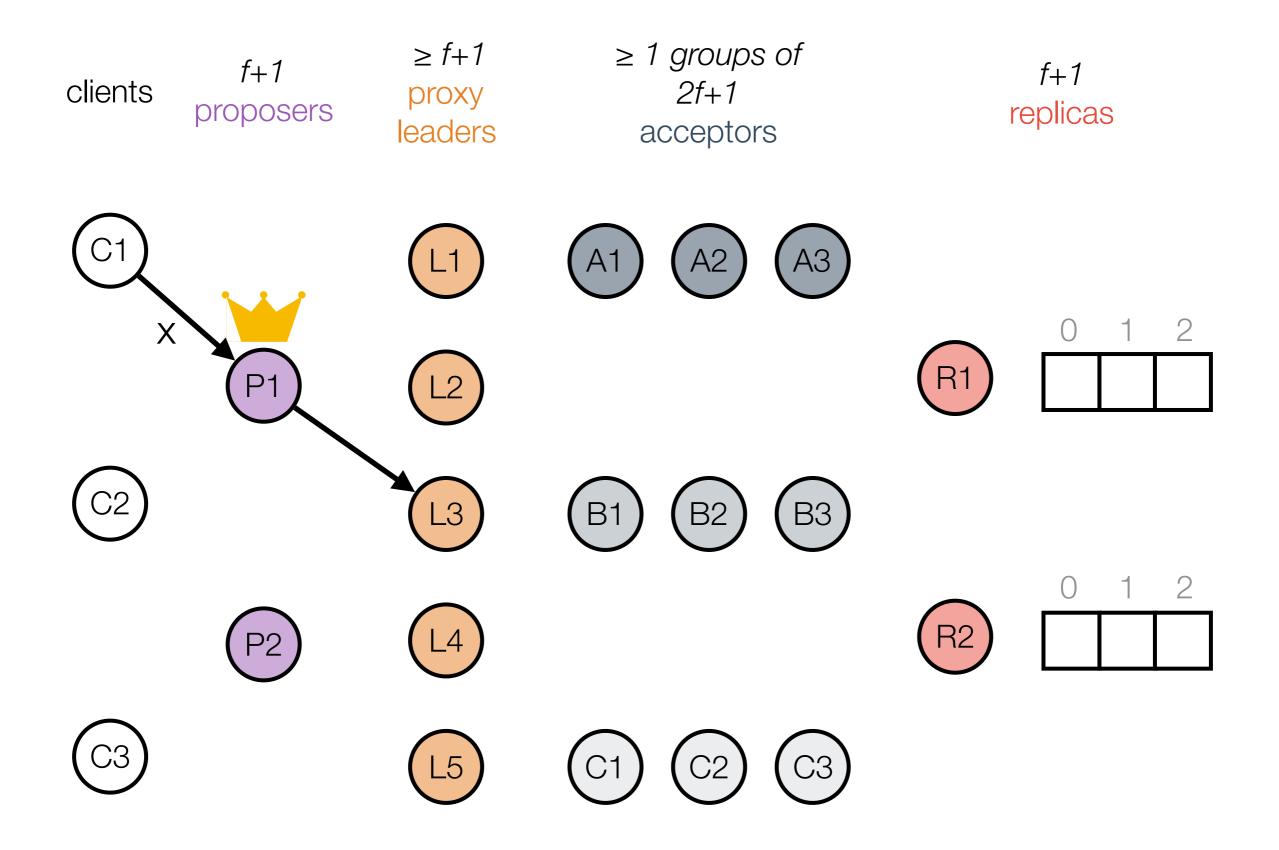
B1

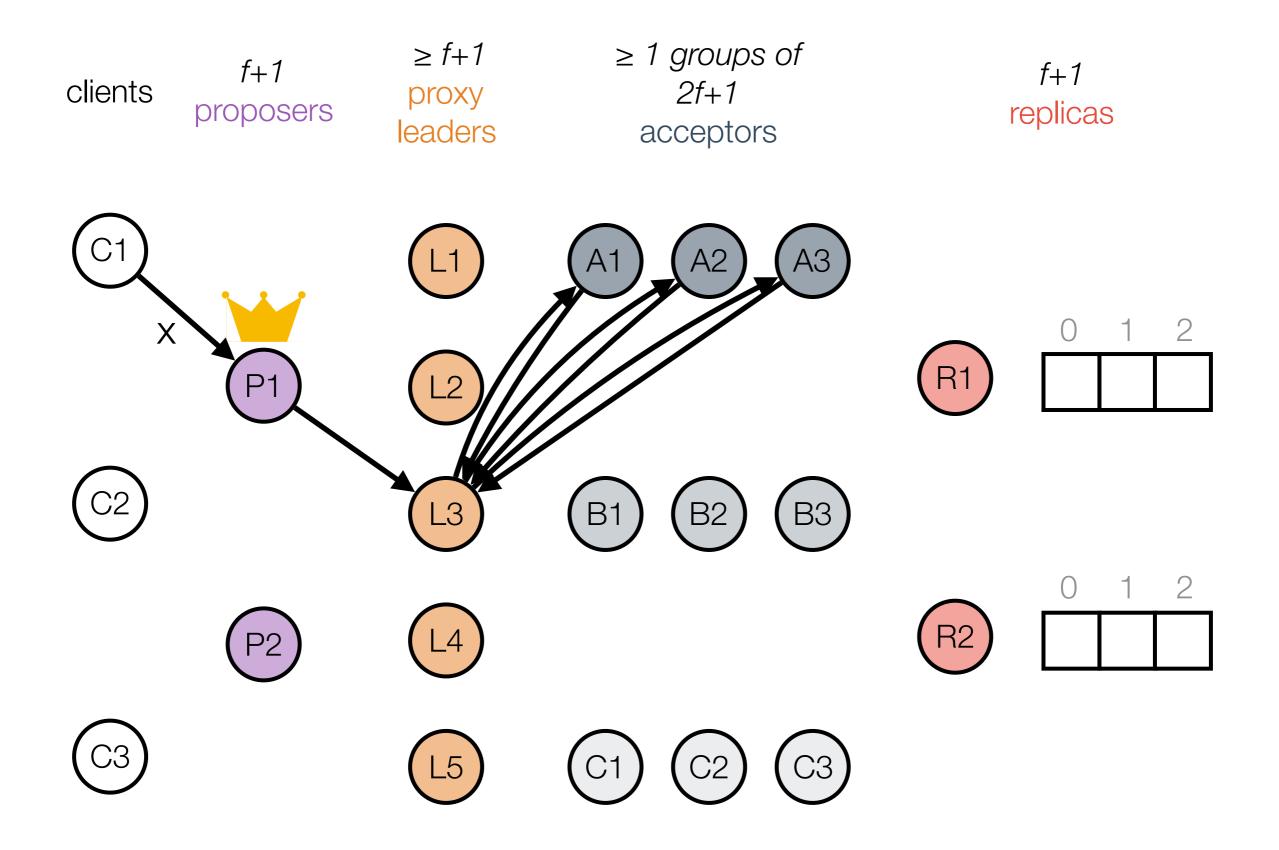
_	0	1	2	3	4	5	6	7	
	a	b	С	d	Ф	f			
	A1	B1	C1						
	A2	B2	C2						
	A3	B 3	C 3						

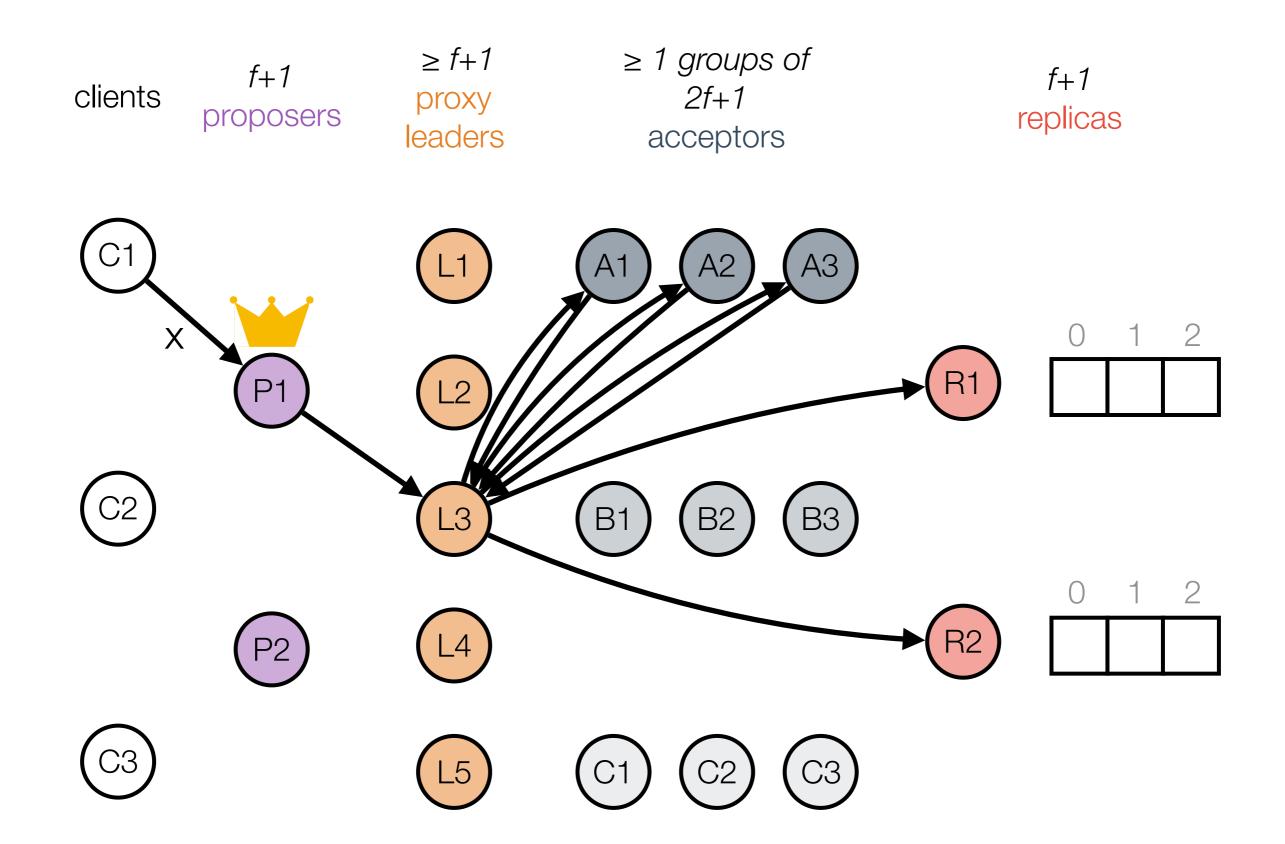
_	0	1	2	3	4	5	6	7	
	a	b	С	d	Ф	f			
	A1	B1	C1	A1	B1	C1	A1	B1	
	A2	B2	C2	A2	B2	C2	A2	B2	
	A3	B 3	C 3	(A3)	B 3	C 3	(A3)	B 3	

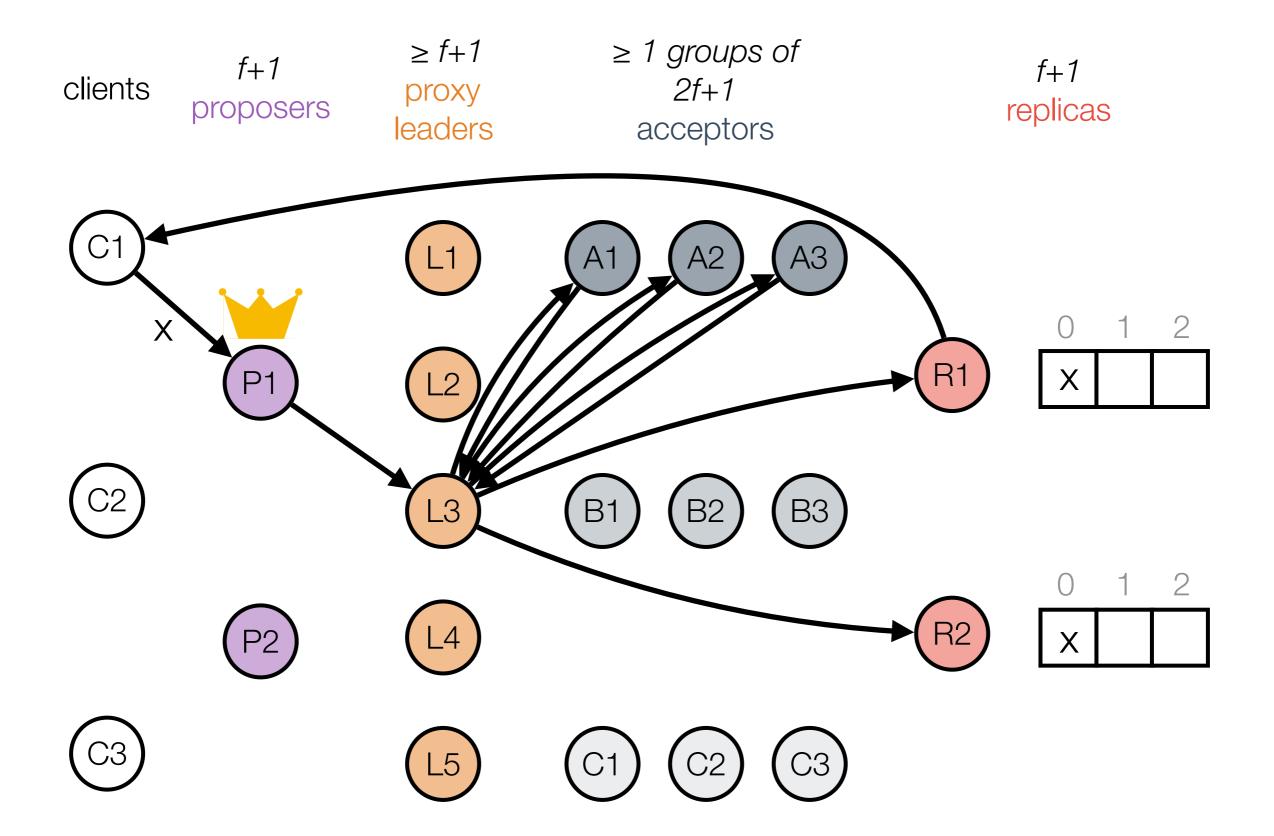


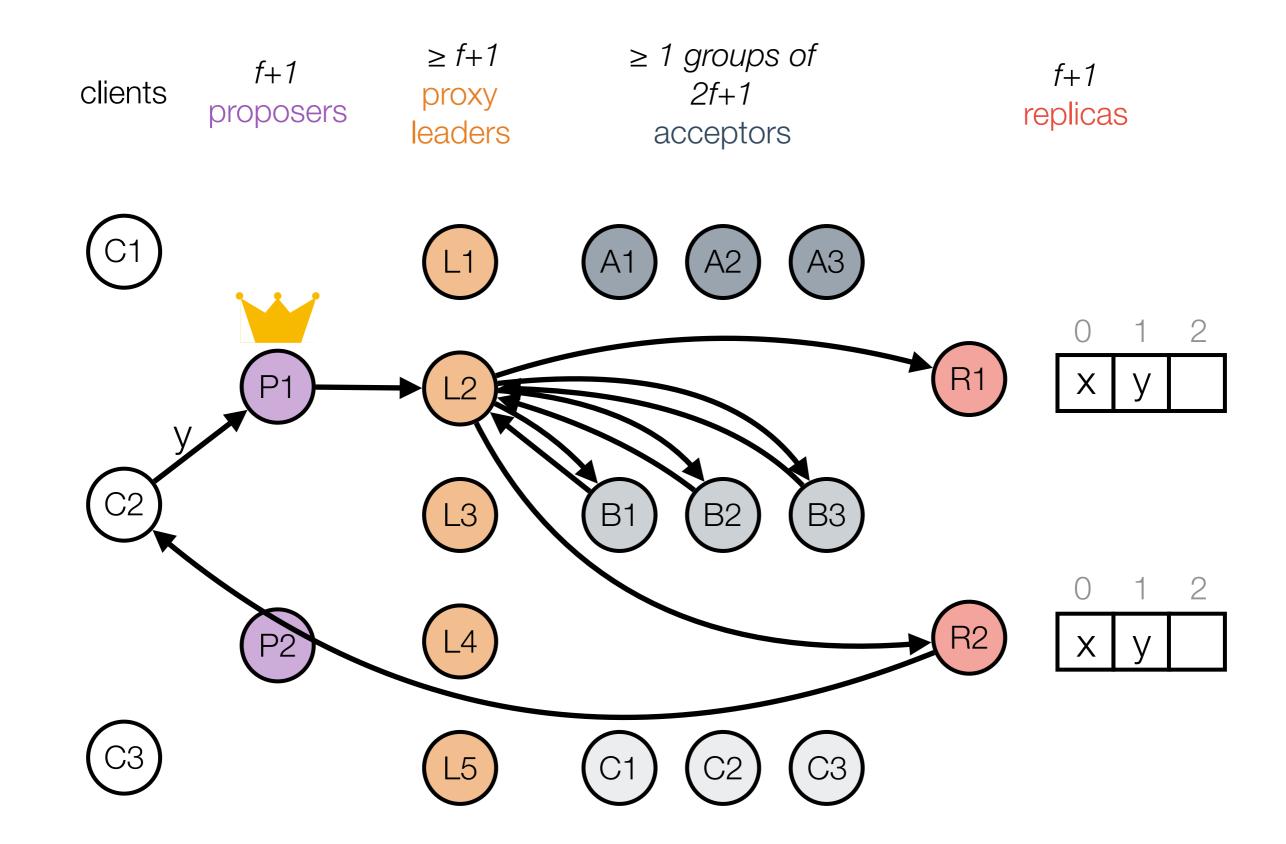


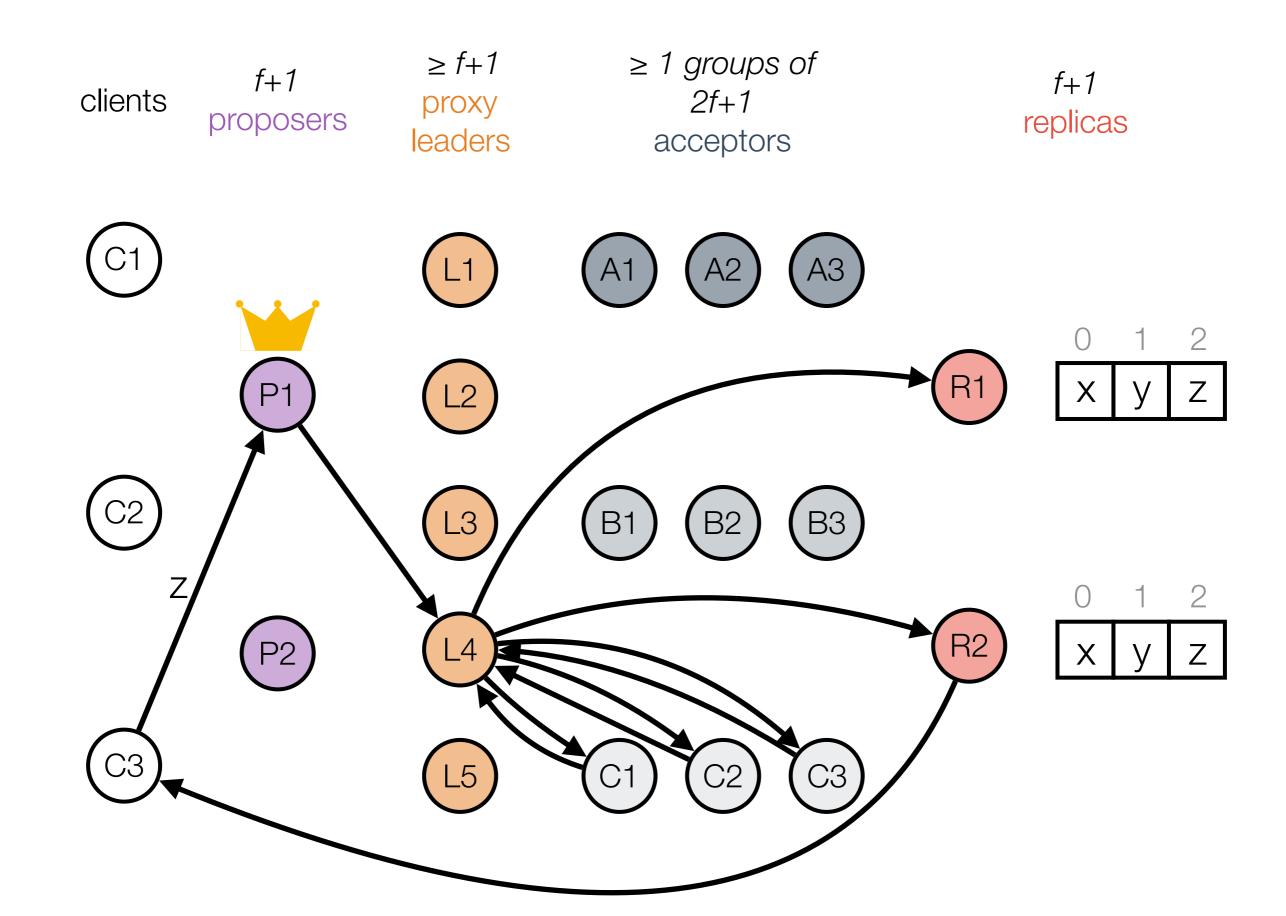






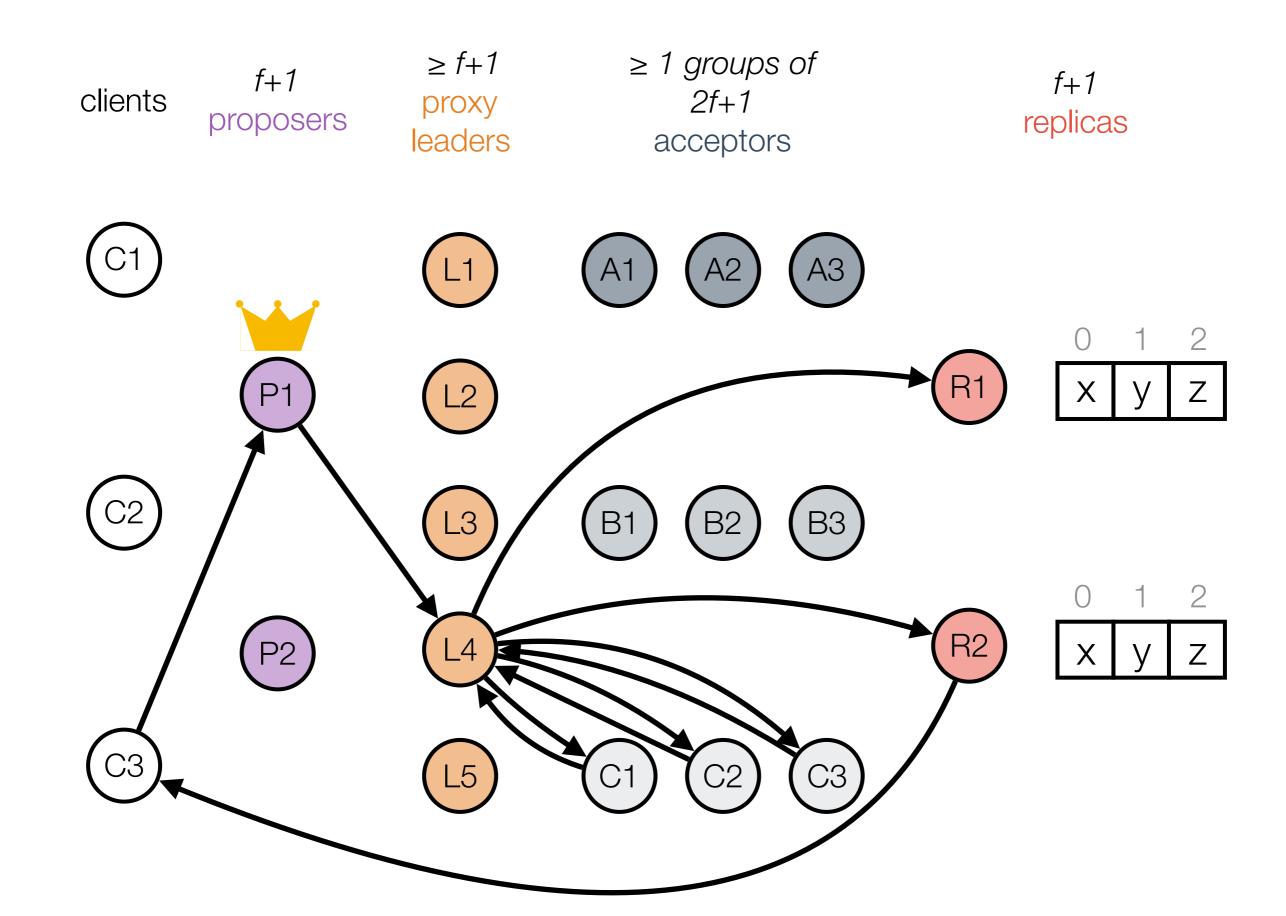


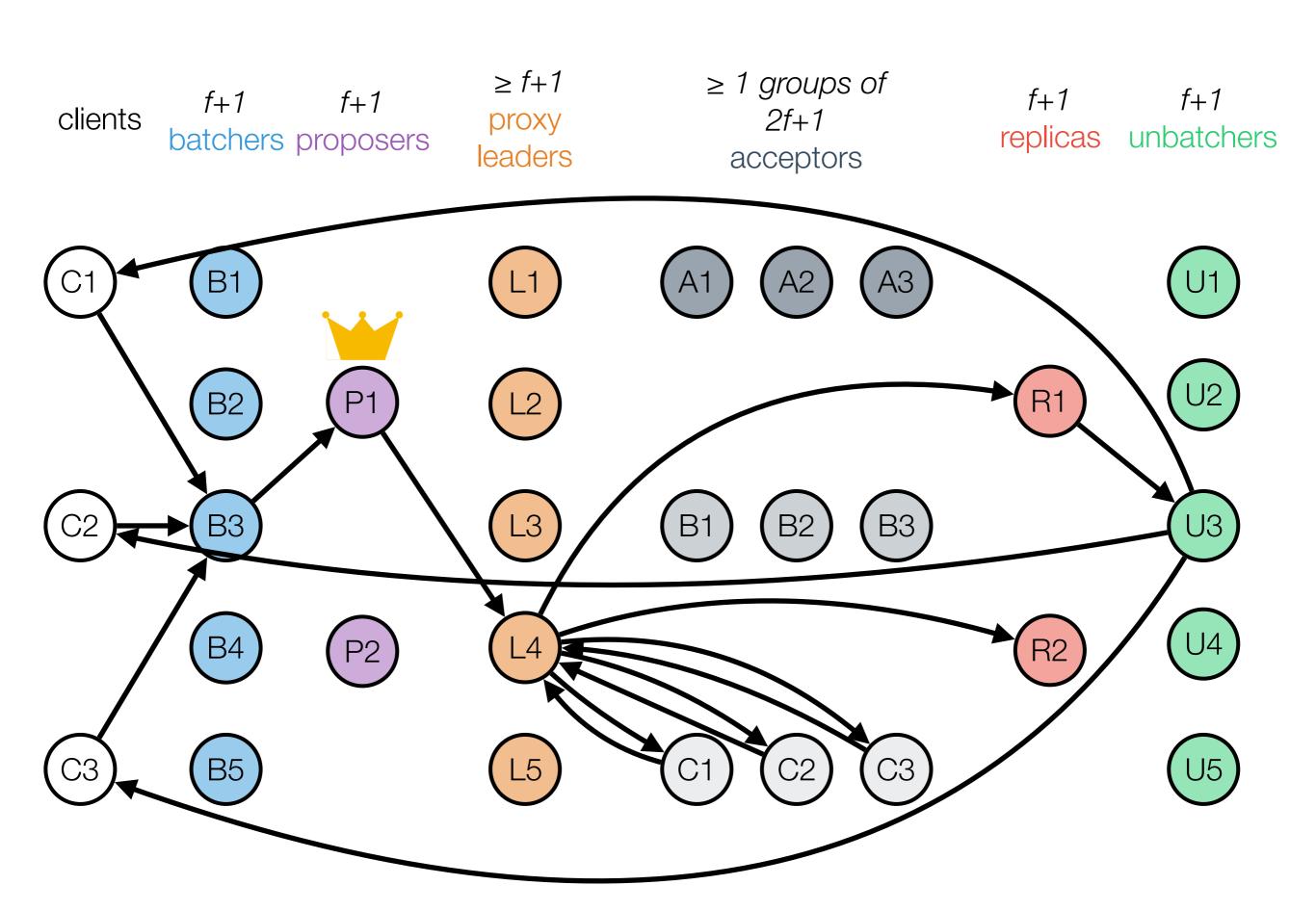




"Using more than 2f+1 [acceptors] for f failures is possible but illogical because it requires a larger quorum size with no additional benefit"

-TAPIR (SOSP 2015)





Mencius: round robin partition the log between multiple leaders

Mencius: Building Kllicient Replicated State Machines for WANs

Valina Mee CSE, UC Say Diego Say Jingo, CA - 78M meey uitue ∓os cos deda Florio P. Jungseim Baloo! Removeh Barcelone Ban slove, Catalona - Spain Bi@vdoo income Kriti Maradio CSE, UC See Disgo San Diego, CA - 1884, caccratio Oct. cost.edu

Abstract

We present a protocol for general state reachine septimization in method that proceds strong and retains, their backing perfectively, their backing perfectively. Their backing protocol Niteria has light broughput under high client head and transfer through a state of an about the distribution of a state of a state of a state of a state of their back. We develop our protocol is a development and be of their protocol protocol in the development can be changed or further refined to take about age of specific networks a application regularization.

1 Introduction

The condigment approach for providing religible acceleration in the set is explicated our machine acceleration [79]. Assuming a discussion of the course, it may not discuss to exclusive the color of the course and function is not bound acceleration of security and an autocarded sequence of consensus matrices in used to agree upon the commands they occurs. This approach provides strong-construction parameters and so to bounds applicable. Admission in effective command provides the consistent of system and instance, from the original application and instances. See an approximate resolution of consistent of the congression of the construction include Charles [8, 3], An Kanger [27] and Formwood [28].

With the rapid growth of wide-arms cardinal such as with services, gain bereiter, and to reads standard and occurrent a bank research specified in how to provide off-ficient state machine regulations in the wide area. One could always a supplication in the wide area. One could always a supplication for a large of one beautiful and ratios first the application (for a large of one beautiful participation gipper). Instead, we could get our department and only a former to reverse and the always are the always and the always are department of the high performance that high the regular made magnitude from the surrounder law of an identified and the participated from the standard rate for the surrounder law of an identified the department of the first of the first

connections have common food of earliest. And, we seed a solution that common with a derivative, the the popular connection protect forces as [22], we have been of first application [15].

Exhibit protects such as Theor. That Proces [25], and Celler F [13] har not, in general, the best concerns proceed of the side area applied from For example. Person ratios on a single leader to choose the request expected. Best for a short in the same of the support of a decrease of the same of the support of the process for a single leader in the same of the fact in the process of the same of the s

This paper process Revelor's small-deader suit machine replantion proceeds that derives them. Places indeapard to achieve high throughput under high citeriland and loss become and loss others lead, and markey to changing betweek and class to streams and.

The hale appears of Merckas is to partition the sequence of consents protect instance among the servers. For example, it is system with three acrees, one could make it is exercised to server 0 for exercises distance 0. 5, 6 ac, server 1 the consents instances 1, 4. 7, or and server 2 feet emissions instances 2, 3. 6 or. Definition of the basis of the consents instances 1, 5 acree 1 three partitions of the feet of the first partition of the first partition of the feet of

The blue of partitioning sequence numbers among

Mencius: round robin partition the log between multiple leaders

Mencius: Building Efficient Replicated State Machines for WANs

Venture Moo. CSE, UC Say Diego Sun Phogo, CA - USA meoyanitue@os.nesd.edn

Photo P. Januaran. Yakoo! Removeh Barcelone Banalone, Catalonia - Spain fri@wahoo iso.com.

Keith Marrello CSE, UC See Diego San Dingo, C4 - USA marrallo@es.nesd.eds.

Abstract

We present a protected for general state machine septicoron, a method that provides strong and istemy. The has high performance in a wide-area network. To particular, our protect! Meet his high finoughput under high-client 'each and less histories under high-three Client lend et al. on exder changing wide area perweck preferences and offers lead. We develop our protocol as a derivation from the well-known promost Parco. Such a development can be Changed or further refined to take advantage of specific network or application on change is.

The court general expression for practicing or highly surfable service is to use a suplicated state reaching to (Charles (9)). Asseming a determinable covers, the man and function is not could across a set of survey, and an embounded sequence of consessus mattaces in used to agree upon the commands they occurs. This approach provides strong consistency or armites, and so is boundly applicable. Advances in efficient communic process, have made this approach practical as well for a without of system architectures, from two principal again. curion of embedded systems (34) to asynchronous systerm. Recent examples of previous that we replicated use munitus inchele Chatty (4,3), Auflager (22)

With the rapid growth of wide-area cervices such as web stances, good services, and tervice countries and secures, a basic research question is how to provide of-Select state machine replication in the wide area. Onecould choose an application - for example, storale conmit to a very to-consist authorized, and powish and fider addition for the application (for a large class best and high throughput. Instead, we work a green beloken that only between the servers and the align save questiserver a wide area network. Wy sock high performance both high firenes and under righ short materal law laanyundariay charakadin in bear lehinging othe-

reconstruct insurance (and disertions And, we see a seleke the oracly in adamate, his thepople. conversely protected. Percent and (22), we it care the most first to apply it to a specific application [15].

Existing protectly such as Phases, Part Passes (25), and Cellet? [13] are not, in nonemi, the best converses prorecolls for with area applies flow. For example, Percos exites draw single leader to choose the request sequence. Deator's singuisty it has high than gight violant pasts generated by efforts in the same title to the frederingly invitating her diseas in other stacks we higher later ey. in addition, the leader in Paper is a beginner other limits. throughput. Finning a single leader also leads to an anbelianced communication present that limbs the will be-tion of bandwidth controlle to all of the network links connecting the servers. Fun Passe and CoRoff, on the other hand do not sells an axingle leader. They have low incarry ander law look, but have leverather gipper under high lead case to their higher through complexity.

This paper process blowers of a malli-leader water ma-

chine regilization protocol that derives from Phose. It is designed to achieve high throughput under high eiters and and loss impacy under loss oftens land, and roundary to the minimum twork and characteristic more and a

The hade approach of Merckin is to partition the sequence of ourseasus protocol instances among the servery. For example, in a system with three servery, one could make to server 0 the conserves incomes 0. 3, 6 pe, server I the consensus insurees 1, 4, 7, our and server 2 tip consenses inchanges 2, 3, 8 cm. (Bottog this priention the local of behalf behalf, which transport network is the borderack is carbifored leader scheme. note fully united the exclude burdwall to interest throughput. It also reduces betarroy, because of bury care use a local server is the leader for their requests; because of the design of Mercias, a diserved typically not have to wait for its server to get by men.

The blue of pattering sequence number sering

26-1089-11X Average in term Operating Systems Testion and Implementation 1995.

S-Paxos: decouple data flow from control flow

S-Paxos: Offloading the Leader for High Throughput State Machine Replication

Mertin Biely, Zurko Militsevic, Nano Santos, André Schiper Benle Polytechnique Pétérale de Lussanae (LPVL) Bruil: firement honomel/curich

Always Implementation of these as what leader-hand produced with the positions of these as what leader-hand produced. These produced we also benieve the leader produced with the positions are what leader-would be the produced with the leader of the leader

Reguents Passe, High Honglips Salarity, Star Machine Replication, Performance

1. Ізткоростюз

becomes a certail feature. One general approach for pro-(SMS) (II, (2), By replanting a service on unitigar servers. Tokestor estably took to higher throughput. disert are guaranteed that every if some rapides fills, the unvice is still available. The state machine replication approach. Popilistical SCM machine performs the following tasks: has been widely considered by both the theoretical $\{X_i,\{4\},$ [5] and systems research community [5], [7], and is also used in several real-world systems (8), (9), (10).

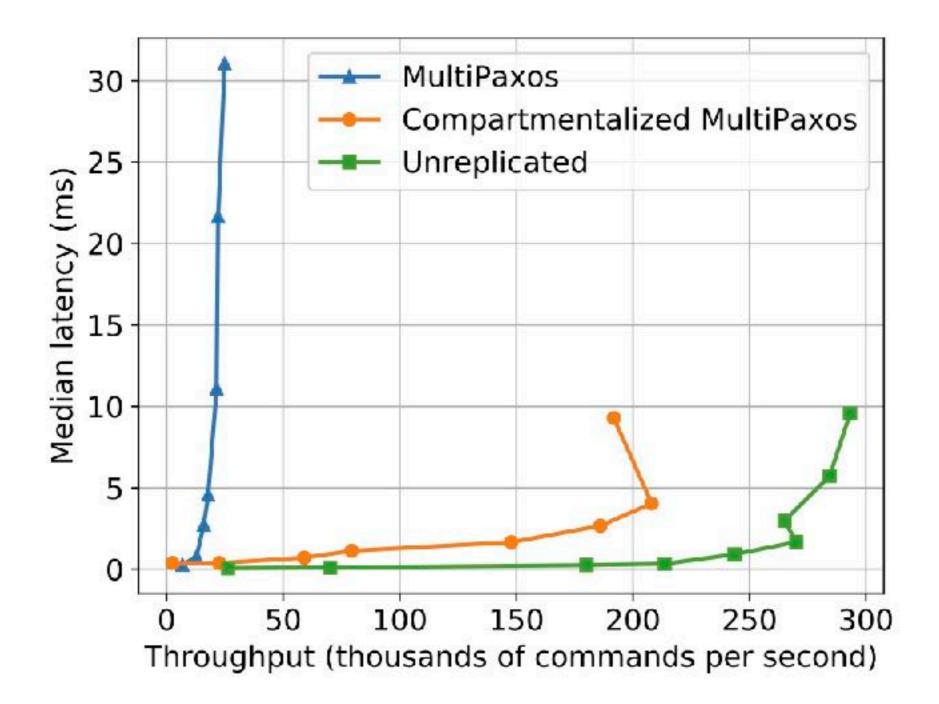
If otherwise requests is more much that existing them. has he system performance in finites by the performance of the service being rephrased. However, politic services are becausely lightweight. Her the Chabby lock service (8) or the

protocols, we propose S-Passe (S stands for scalable), a novel SMR protect) for clustered networks derived from Pages. So Paper achieves high throughput by balancing the load among region to are otherwise idle encages, functly creiting the bottlenerk at the leader. Purtnermore, its throughout keeps pressing with the condensof replicts (up to a resourch As politic services from the importance, their as distriby smarker). This way, S-Parce (womens the implicitual tradeof tesses full tolerage and thoughput potent in cost strike a highly available service it came matching particulars, care matching projection projected in Science Nober Both

In order to derive 5-Pases, we sunt by observing that a

- a Sequent dissociation; receiving sequent from classes and disaboring them to all regions
- · Entitlesing order reaching agreement to the order of Programming.
- Negacia companies concerting requires in the determined order and sanding replies to clients.

In leader-centric proposis such as Payes, most responsi-Zuderen austin den werder E. and van is de in van high billigt for time dans take som with the leader while the throughput provided that the ordering protect, one body up. Followers are left-only with adhesyledging the order proposed. the atthosy of Souther, at report is (11) that make Asymptotic by the feater and executing requests. The Softwar bey design et deutsteger is handed by the forengiant of the underlying gradeline bere devident the drive mais everly decreasesplace. nonle to de so, se ton the first two todas into separate We observe that implementations of state such its region. Levels disconnection and country. The disconnection form in tom map made all priving admits of Perce [9] for other leader. They indown among all replicat, with a find from invention. based promote. Two of to these proposals are also leaders and disconnected of the promote Since proposal contribution corne, in the sense that most of the wore is done by the in handles by the distermination layer the rule of the o-claring leader. Therefore the bottlesseld is found at the basics and layer is only to determine the order in which remains will the mechanic throughput in limited by the leader's resources—the encount. We use mornal flows for this layer with a



Some protocols become the bottlenecks they try to avoid!

mwhittaker.github.io

mwhittaker.github.io

mwhittaker.github.io/frankenpaxos

mwhittaker.github.io

mwhittaker.github.io/frankenpaxos

github.com/mwhittaker/frankenpaxos