The Part-Time Parliament

LESLIE LAMPORT

Digital Equipment Corporation

最近在Paxos岛上的考古学发现揭示了尽管立法者们都很逍遥，但是议会仍然运作。立法者们维护着议会记录的一致的副本，尽管他们经常不在室内而且他们的传信者也总是健忘。Paxos的议会协议提供了一种新的状态机实现去设计一个分布式系统。

关键词：计算机网络交互，分布式系统，操作系统，行政数据处理

最近在TOCS社论办公室的文件柜后发现了这个提案。尽管因为它年龄的问题，但是主编认为这个是值得出版的。这位作者现在正在希腊岛做相关的工作，所以我被要求准备这次出版。

作者似乎是一位只对计算机科学感兴趣的考古学家…（无关紧要的内容）

正文：

1. The Problem

1.1 The Island of Paxos

在这个千年的初期，爱琴海上的Paxos岛是一个蓬勃发展的商业中心。财富导致政治复杂化，Paxos的人用政府的议会取代了民主制度。但是岛上的公民在贸易面前，没人愿意一直待在议会。所以Paxos的议会需要一个好的处理方法，这样即使议员不停地进出议会厅也能够对事务进行处理。

兼职的议会的问题是如何出色的解决对应的问题以面对如今的容错分布式系统，即使有议员处理对应任务但是部分议员不在议会导致失败。Paxos的解决方案可能因此成为计算机科学中的一部分。我在这里的Paxos议会协议的一小段描述，是为了引出分布式系统的讨论。

Paxos文明被外来侵略者所破坏，考古学家最近才开始挖掘这段历史。我们关于Paxos议会的了解因此而模糊。虽然我们知道基本协议，但是我们还缺少很多细节。这些细节很有趣，我会自由的想象议会是怎么做的。

1.2 Requirements

议会的主要任务是决定岛的法律，这个由通过的法律序列定义的。现代的议会会雇佣秘书来记录行为，但是在Paxos没人愿意停留在办公室始终作为一个秘书。每个Paxos议员维护了一份记录着通过的法律序号的账本。比如，议员λ的账本有这么一句：

155: The olive tax is 3 drachmas per ton

如果她相信155号法令被议会通过。账本会由不可消除的墨水书写，并且他们的不会被修改。

议会协议的第一个要求是账本的一致性，也就意味着两份账本不能有冲突的信息。如果议员Φ的账本中有

132: Lamps must use only olive oil

就不能有其他议员在132这条规则上存在不同。然而，其他的议员可能在他的账本上没有132这一条法令，如果他没有学习过已经通过的法令。

只有账本的一致性是不够的，因为可以简单的使账本为空白。还需要一些要求用以保证法令会被最终通过并记录到账本上。在现代议会中，通过法令时会被不同意的部分议员阻扰。Paxos的情况并非如此，议员之间相互信任。Paxos的议员愿意通过被提出的任意提案。然而，他们的周期倾向造成了一个问题。如果一个议员团体通过法令

37: Painting on temple walls is forbidden

并举行宴会，不管另一个议员团体是否在议员办公室，并对他们刚做的决定一无所知，又刚好通过了一个冲突的法令

37: Freedom of artistic expression is guaranteed

此时，一致性就会丢失。

进展不能被保证除非有足够的议员在一段足够长的时间内待在议员办公室。因为Paxos的议员不愿意减少他们外出的活动，这不能够完全确定任何法令一定能被通过。然而，议员们愿意保证，在议员办公室的议员，他们和他们的助手会及时的通知所有的议会相关事宜。这个保证允许Paxos人们设计一个议会协议满足下面进展条件。

如果大部分的议员在办公室并且在足够长的一段时间内没有人进入或离开，办公室中议员提出的任何法令都会被通过，并且每个通过的法令都会被记录到每个在办公室的议员的账本上。

1.3 Assumptions

议会协议的要求只有当议员被提供足够的资源时才能实现。每一个议员收到一本包含法令记录的账本，一支笔和一盒不会消失的墨水。议员可能在离开办公室后忘了他们做了什么，所以他们会写笔记在账本后面以提醒自己一些重要的议会任务。一个列表中的法令实体永远不会被改变，但是笔记可以被划掉。达到这个进展的条件要求议员可以测量时间的流逝，所以有一个简单的沙漏。

议员随身携带着他们的账本，可以一直阅读法令清单和未被划掉的笔记。账本由最好的羊皮纸做成，只写重要的笔记。一个议员可以写其他笔记在纸上，这些纸在他离开办公室后可能丢失。

办公室内的声音很轻，不能做演讲。议员可以通过信差交流，并且有足够的金钱雇佣很多的信差。信差不能修改内容，但是他可能会忘了他已经传过消息，然后再传一遍。就像议员一样，信差也是兼职的。信差可能离开办公室去度假。也可能永远离开，也就是让他传的消息永远不会被送到。

尽管议员和信差可以在任意时间进入和离开，当他们在办公室内，他们会只专注于工作。当他们留在房间内时，信差迅速分发消息给其他议员，并且议员也会最快时间内处理并回复他们收到的信息。

Paxos官方记录显示要求议员和信差完全的诚信和遵守议会协议。大部分学者宣传这些内容，为了显示Paxos的人在道德上更好。不诚实的说，虽然罕见，但确实存在。然而，因为不是在官方文档中提到，所以我们对议会如何应对不诚实的议员和信差的手段知道的不多。这些被揭露的证据在3.3.5节会被提到。

2. The Single-Decree Synod

Paxos议会是从每19年召开的早期祭祀仪式演变而来，以选择一个单独的象征性的法令。数百年来，会议根据常规程序选择了一个个法令，这些程序要求所有的牧师都在场。但是由于商业的蓬勃发展，当会议在仪式中时，牧师开始不停的徘徊于室内外。最后，旧协议失败了，会议在结束时没有选择任何法令。为了避免重复神学的困难，Paxos的宗教领导要求数学家制定一个协议可以选择会议的法令。这个协议的要求和假设实质上是和后来的议会期望一致的，也就是包含一个法令的序列号，一个账本对应的最多有一个法令。会议协议的描述在本节中，议会的协议将在第3节中描述。

数学家们通过多步派生出会议协议。首先，他们证明结果表明协议满足一些约束能够保证一致性并允许进度。一个初步的协议在这些限制下衍生了出来。受限版的初步协议提供了基础协议保证一致性，但不包括进度。完整的会议协议，满足一致性和进度要求，是通过对基础协议的再限制得到的。

数学结果在2.1节描述，协议非正式的描述在2.2-2.4节。对基础协议的正式描述和正确证明在附录中。

2.1 Mathematical Results

会议的法令通过一系列编号的选票，每张选票只能投给一个法令。每个牧师只能选择投或不投这张选票。和选票相关的是叫法定人数的牧师集合。一张选票当且仅当每个法定人数中的牧师都投了这个法令才叫投票成功。正式的来说，选票B由以下四部分组成（除非另有说明，集合就代表了有限集合）

一个法令（被投票的那个）

牧师的非空集（选票法定人数）

牧师集合（投给了这个法令的人）

选票数量

一张选票B被认为成功当且仅当，所以一张成功的选票是法定人数中的每一个都投的。

选票数量是从无限的有序集合中选出的。如果，那么选票被认为落后于选票B。然而，这不代表选票被进行的顺序；后面的选票可以替代更早的。

Paxos的数学家们定义了三个条件在选票集合B上，并且证明了如果选票的集合可以符合这些条件，一致性将被保证，同时程序也是可行的。开始的两个条件很简单，可以被如下的表示简单说明：

B1(*B*) 每个*B*中的选票有一个独特的选票序号

B2(*B*) *B*中任意两个选票的法定人数至少有一个共同的牧师。

第三个条件略微复杂。一份Paxos的手稿包含了如下描述，相当混乱：

B3(*B*) 每个*B*中的选票B，如果任意选票B法定人数中的牧师早就投给了*B*中的更早的选票，那这个B中的法令等于之前的选票中最新的法令。

这段密文的解释在图1所示的手稿中，说明了条件B3(*B*)，五位牧师A, B, Г, Δ, E组成的会议中有5个选票。*B*集合有五张选票，对于每张选票，投票者的集合是法定人数牧师的子集。举个例子，14号选票有法令α，一个法定人数包含三个牧师和包含两个投票者的集合。条件B3(*B*)有形式“对每一个*B*中的选票B:….”，其中“...”是选票B上的条件。图1中这五个选票B的条件如下。

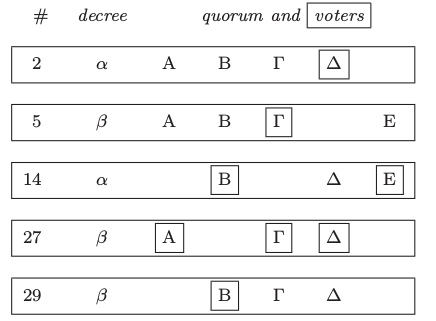


图1. Paxos手稿显示了一个集合*B*，包含了五个选票，满足条件B1(*B*)-B3(*B*)。

2. 2号选票是最早的选票，所以这张选票显然满足条件。

5. 5号选票的四个法定人数成员没人投票给更早的选票，所以这样选票也满足条件。

14. 14号选票的法定人数中只有一个 Δ投票给了更早的选票，也就是2号选票，所以根据条件要求14号选票的法令必须等于2号选票的法令。

27. （这是一个成功的选票）27号选票的法定人数是A, Г和 Δ。牧师A没有投票给更早的选票，Г的唯一一张更早的选票是投给5号选票的，同时 Δ的更早一张选票是给2号的。最近的这些两张选票是5号选票，所以根据条件要求，27号选票的法令必须等于5号选票的法令。

29. 29号选票的法定人数是B, Г和Δ。B投的更早一张选票是14号选票，牧师Г投了5号和27号，Δ投了2号和27号。最近的这四张中更相近的是27号选票，所以根据条件要求，29号选票的法令必须等于27号选票的法令。

为了更好的声明B1(*B*)-B3(*B*)，需要更多的符号。一张投票*v*定义为三部分：一个牧师，一个选票序号，一个法令。它代表了由牧师投给了选票的法令。Paxos居民们也定义了*null*投票，并且，其中并且BLANK不是法令。对于任意的牧师*p*，定义为唯一的null投票*v*当。

Paxos岛的数学家们对于所有投票定义一个总的排序，但是包含定义的手稿的部分丢失了。剩下的部分表明如下，对于任意投票v和v’，如果，那么。并不清楚当时v和v’的相对排序如何定义。

对于任意的选票集合*B*，*B*中的Votes(*B*)定义为包括所有的投票v即对于部分 B ∈ *B*,

。如果*p*是牧师并且*b*是选票序号或者，那么是被定义为由*p*投出的Votes(*B*)中最大的投票v，并且或者如果没有这个投票。因为比任何*p*的真实投票都小，也就意味着是集合中最大的投票。

对于任意的牧师非空集合*Q*，被定义为等于当*p*在*Q*中时的。

条件B1(*B*)-B3(*B*)正式描述被如下：

虽然的定义依赖于投票的顺序，B1(*B*)暗示了是独立于如何选择被排序的选票号相同的投票的。

为了展示这些条件如何实现一致性，Paxos岛民们首先展示了B1(*B*)-B3(*B*)，如果一个*B*中的选票B成功了，那么任意之后的*B*中的选票是和选票B的法令一样的。

引理： 如果满足B1(*B*), B2(*B*), B3(*B*)，则对于任意*B*中的B, B’

引理证明：

对于任意的*B*中的选票B，让作为*B*中晚于选票B的且法令不同的选票集合。

为了证明lemma，如果则是空的足以展示这一点。Paxos的人给出了矛盾的证明。他们假设存在*B*且和，并且获得了一个如下的矛盾：

1. 选择，则

证明：C存在因为非空且有限

2.

证明：按1和的定义。

3.

证明：按B2(*B*)和假设

4.

证明：按2,3和的定义

5.

证明：按4（表明非空）和的定义

6.

证明：按5和B3(*B*)

7.

证明：按1,6 和 定义可得

8.

证明：按4，因为7和B1(*B*)表明

9.

证明：按7,8 和定义

10.

证明：按定义

11. 矛盾

证明：按1,9,10得

引理证明结束

根据Lemma，很容易得出，如果B1-B3满足，那么任意两个成功的选票会是同一个法令。

定理：如果满足B1(B), B2(B) 和B3(B)，则对于任意*B*中的B, B’

定理证明：

如果，则B1(*B*)表明。如果，则定理满足引理。

定理证明结束：

2.2 The Preliminary Protocol

Paxos人随后证明了这个定理断言，如果会议中有足够的牧师，则有可能产生一个成功的选票当它维持B1-B3。虽然这不能保证进度，但它至少展示了基于选票协议不会造成死锁。

定理2：使*b*作为选票序号，*Q*作为牧师集合，则对于所有的，都有和。如果B1(*B*)，B2(*B*)，B3(*B*)保持，则存在一个选票B’使得和，则保持。

定理证明：

条件基于B1(*B*)，的选择，和关于*b*的假设。条件基于B2(*B*)，的选择和关于*Q*的假设。如果，则使为任意法令，否则让其等于。条件基于B3(*B*)。

定理证明结束：

Paxos岛民们从要求B1(*B*)-B3(*B*)为真，当*B*作为所有进行过的选票集合的情况下派生出了初步协议。协议的定义指明*B*如何更改，却从未被明确计算过。Paxos的人们将B的数量视为只有上帝可以知道，因为它从不被任何凡人所知。

每个选票有牧师定义，他选择序号，法令和法定人数。每一个法定人数中的牧师再决定是否投票。这个为了维护B1(*B*)-B3(*B*)而派生出的规则决定了初始人如何选择选票号，法令和法定人数，以及牧师决定是否投票。

为了保持B1，每一个选票必须收到一个唯一序号。通过记住他以前初始化的选票，一个牧师可以轻松的避免初始化两张同样序号但内容不同的选票（通过账本上的笔记）。为了保持不同牧师用同样的序号初始化选票，可能的选票序号集合在牧师间被划分了。当不清楚这是否被使用时，一个明显的方法将会让选票序号作为一对由序号和牧师组成的数，使用词典顺序，如

因为Γ比Λ早。在任何情况下，每个牧师都有一个无界的选票序号集合给他自己使用。

为了保持B2，一个选票的法定人数被选择包含牧师。原来，只意味着一个简单的多数。后来它被观察到胖牧师更少移动并且花费更多时间在办公室中，所以被用来代表总体重比一般牧师总体重更重的牧师的集合，不仅仅是简单的牧师中的多数。当瘦牧师的团体抱怨说不公平时，实际体重被基于牧师的出席记录代表的象征体重取代。主要的要求试任意两个保持的牧师集合中最少有一个公共的牧师。为了保持B2，牧师初始化选票B选择作为主体。

条件B3要求如果不为空，序号为b的选票和法定人数Q必须有法令。如果为空，则选票可以有任何的法令。为了保持B3(*B*)，初始化一个序号为b法定人数为Q的新选票前，牧师p必须找到。为了做到这点，p必须找到每一个Q中的牧师p的。

回顾是小于b的最大选票序号，由q投起或的投票。牧师p获得q的通过交换信息。因此，协议的最初两步是：

(1) 牧师p选择一个新的选票序号b并发送一个的信息给一些牧师集合。

(2) 一个牧师q收到了消息并回复给p，v是q投起的小于序号b的选票序号，或者当q没有任何选票序号小于b时为。

牧师q必须使用笔记在他的账本上记录他曾经发起的选票序号。

当q发送消息，v等于。但是*B*改变了。因此牧师p选择法令时将使用v作为的值，为了保持B3(*B*)成立，必须在q发送后不改变。为了保持不断变化，q必须不用到b之间的序号投票。通过发送信息，q承诺不投票（为了保证这个承诺，q必须在账本上记录必要的信息）。

选票协议（牧师p的第1步开始）接下来的两步是：

(3)在从Q中大部分牧师收到消息后，牧师p初始化一个新的选票序号为b，法定人数Q，法令d，其中d满足B3。他之后在他的账本上记录这个选票，并发送一个消息给Q中的每一个牧师。

(4)收到消息后，牧师q决定是否给这个选票投票（他可能不会投票如果这会违背他给其他选票发送消息后做的承诺）。如果q决定投票给b，他就会发送消息给p，并且在他的账本上记录。

第3步的执行考虑到在*B*中添加选票B，其中和。第4步中，如果牧师q决定投票，则执行该步被认为是通过添加q至来改变*B*，当B∈*B*。

牧师可以在第4步选择不投票，即使投票也不会违背之前的承诺。实际上，协议中的所有的步骤都是可选的。举个例子，牧师q可以忽略消息而不是执行第2步。不采取行动可以防止进展，但是可能造成不一致因为违反了B1(*B*)-B3(*B*)。由于没有收到消息的唯一结果是不发生动作，消息丢失也不会造成不一致。因此，协议保证了一致性即使牧师离开办公室或者信息遗失。

收到多份消息的拷贝可能造成行动重复。除了步骤3，第二次执行动作不会有任何效果。举个例子，在第4步中发送多个消息和只发送一次是一样的。第3步的重复试通过账本中已执行的记录避免的。因此，一致性的条件被保证即使一个同样的消息被多次传递。

步骤1-4描述了初始化选票和投票的完整协议。剩下的是决定选票结果和当法令被选择后的宣布流程。回顾一个选票是成功的当且仅当法定人数中的每一个牧师都投给它。成功的选票的法令是被会议选择的。剩下的协议如下：

(5)如果p从每一个Q（选票号b的法定人数）中的牧师q收到消息，然后他在他的账本上写d（该选票的法令）并发送一个消息给每一个牧师。

(6)收到消息后，牧师在他的账本上记录法令d。

步骤1-6描述了一个独立的选票如何产生。这个初步的协议允许任何牧师在任意时间初始化新的选票。每一个都保持了B1(*B*)-B3(*B*)，所以整个协议也保持了这三个条件。因为一个牧师在他的账本上记录法令仅当法令被成功通过，定理1表示牧师们的账本是一致的。协议没有解决进程的问题。

在第3步中，如果法令d由B3决定，则可能法令早就在一些牧师的账本上被记录了。牧师不需要作为Q中的一员，他可以离开办公室。因此，一致性可能不被保证，如果第3步允许任意更大的自由空间去选择d。

2.3 The Basic Protocol

在初步协议中，一个牧师必须记录(1)他初始化的每一张选票(2)他投的每一个选票和(3)他发送的所有消息。对于忙碌的牧师而言，保持追踪所有的这些信息是很困难的。Paxos初步协议因此限制了基本协议中要求每个牧师p必须获得的内容为仅限于他账本后的以下信息：

*p*尝试初始化的最后一张选票序号，如果没有则为

*p*之前投的选票中序号最高的选票，如果没有则为

*p*发送消息后*b*的最大值，如果没有则为

初步协议的第1-6步描述了一个单独的选票如何由它的创造者*p*进行的。初步协议允许*p*并发生成Ballot的任意序号。在基本协议中，他一次只能生成一个序号即。在*p*初始化该选票后，他会忽略他之前初始化的其他选票相关的消息。牧师*p*保持所有的关于的信息在一张纸上。如果他丢失了这张纸，他就会停止该选票的进行。

初步协议中，每一个*q*回复的消息表示他承诺不投票给选票序号在和*b*之间的选票。在基本协议中，它表示强保证不新投票给任何小于*b*的选票序号。这个不投给其他选票的强保证在基本协议的第4步不可避免，但在初步协议中是可以避免的。然而，因为初步协议总是给q不投票的选择，基本协议不要求他做任何在初步协议中不允许的事。

初步协议的第1-6步转变成基本协议中的如下6步。（所有的包括上面描述的3种信息的内容，都写在一张纸上）

(1) 牧师*p*选择一个新的选票序号*b*，同时*b*大于，设置为*b*，并且发送给其他牧师。

(2) 从*p*收到一个 消息并且 ，牧师*q*设置为*b*并且发送消息给*p*，其中*v*等于 。（当 时被忽略）

(3) 从牧师集合Q中收到大部分牧师的信息后，其中要求，牧师*p*初始化一个序号为*b*新的选票，法定人数Q，法令为*d*，其中*d*满足条件B3。之后他发送消息给Q中的每一个牧师。

(4) 收到满足的消息后，牧师*q*投票给选票*b*，设置为该选票，并且发送消息给p。（如果则被忽略）

(5) 如果*p*从Q中每一个牧师*q*处收到消息，其中，然后他在他他账本上写下该法令*d*，并且发送消息给每一个牧师。

(6) 收到消息后，牧师在他账本上记录法令*d*。

基本协议是初步协议的限制版本，意味着在基本协议中被允许的每一步都在初步协议中被允许。因为初步协议满足一致性条件，基本协议也满足条件。就像初步协议一样，基本协议不要求每一步都被执行，所以没有解决进度的问题。

B1-B3推导基本协议的的进度使满足一致性条件这一点被观察到。然而，一些类似的“观察”古老的智慧结果是错误的，并且怀疑的公民们要求更严格的证明。Paxos的数学家们证明协议满足一致性条件的进度在附录中。

2.4 The Complete Synod Protocol

基本协议保持了一致性，但是不能确保进度因为它只表明牧师可以做什么，不要求牧师做任何事。完整的协议由基本协议中同样的6步组成。为了帮助实现进度，它包括了明显的额外要求，牧师在2-6步要尽可能的快。然而，为了实现进度条件，一些牧师执行步骤1是必要的。完整协议的关键在于决定一个牧师何时应该初始化选票。

绝不初始化选票肯定会影响进度。然而，初始化太多选票同样会影响进度。如果*b*比任意其他的选票序号都大，然后在第2步中牧师*q*收到信息可能引起一个保证即在第4步中对之前任意初始化过的投票都不投。从而，一个新选票的初始化会影响任何之前成功初始化的选票。如果一个新选票是在之前的选票成功之前使用递增的选票序号初始化，则不会有任何进度产生。

实现进度条件需要初始化新的投票，直到成功，但不会太频繁地初始化。为了发展完整的协议，Paxos的人们首先需要直到信使需要多久传递消息以及牧师需要多久回复。他们决定在办公室中的信使总是需要最多4分钟传递信息，并且留在办公室的牧师在该事件中总是需要最多7分钟采取动作。从而，如果*p*和*q*都在办公室中，一些事件造成*p*发给*q*消息，然后*q*回复给*p*，如果双方的信使都没离开办公室，则*p*将会在22分钟内收到该回复。（牧师*p*将会用7分钟以内的时间发送该事件的信息，*q*收到需要4分钟，回复需要7分钟，*p*收到需要4分钟，总22分钟）。

假设只有一个牧师*p*正在初始化选票，并且他在协议的第1步给每个牧师发送信息。如果大部分牧师都在办公室中时，*p*初始化了一个选票，则他可以用22分钟执行步骤3，再用22分钟执行步骤5。如果他不能在这些时间内执行这些步骤，则有其他牧师或信使在*p*初始化选票后离开了办公室，或者有其他牧师之前初始化的更大序号的选票（在*p*成为唯一的初始化选票的牧师之前）。为了处理后一种可能，*p*必须学习关于其他牧师初始化的比更大的选票序号。这个可以由扩展协议完成，要求如果一个牧师*q*从*p*处收到了或者信息，其中，则他回复给*p*的消息中包含。牧师*p*接着会使用更大的数字进行初始化。

仍旧假设p是唯一的一个初始化选票的牧师，假设他被要求初始化一个新的选票当且仅当(i)