[**《Paxos Made Simple》翻译**](http://www.cnblogs.com/YaoDD/p/6150498.html)

## 1 Introduction

可能是因为之前的描述对大多数读者来说太过Greek了,Paxos作为一种用于实现*容错的分布式系统*的算法被认为是难以理解的。但事实上，它可能是最简单，最显而易见的分布式算法了。它的本质其实就是*共识算法*。

在下一节中我们将展示，该共识算法基本满足了所有我们想要它满足的特性。

最后一节则展示了完整的Paxos算法，通过直接应用*协商一致*的状态虚拟机来构建分布式系统——这种方法应该是广为人知的，因为这可能是分布式系统理论中被引用最多的领域。

## 2 The Consensus(共识) Algorithm

### 2.1 The Problem

假设有一些进程可以提出value。共识算法保证了在所有**提出**的value里只有一个会被**选中**。如果没有value被提出，那么也就没有value会被选中。如果一个value被选中了，那么其他的进程应该能够获取该value。协商一致的要求如下：

* 只能选择已经被提出的value
* 只能选择一个value
* 进程只能获取那些真正被选中的value

我们不会尝试指定精确的要求。但是我们的目标是要确保总有一些被提出的value会被选中，如果一个value最终被选中了，那么其他进程最终要能够获取该value。

我们用三类agent来代表*共识算法中的三类角色*：proposers, acceptors和learners。在具体的实现中，一个进程可能扮演不止一类agent，但是从agent到进程的映射我们在这里并不关心。

假设agent之间可以通过发送message互相通信。我们使用customary asychronous，non-Byzantine model，其中：

* Agents以任意速度执行，可能发生故障，可能重启。因为所有的agent都可能在一个value被选中之后故障并重启，因此一般的方法是不可行的，除非agent能记住一些信息，即使发生了故障或重启。
* 发送的message可以是任意长度的，可能重复，也可能丢失，但是它们不会被损坏

### 2.2 Choosing a Value

存在一个单一的acceptor agent是最简单的选择value的方式。proposer向acceptor发送proposal[提议]，后者从中接收最早收到的那个。虽然很简单，但是这种方法是不能满足要求的，因为acceptor的故障就将导致接下来的操作都无法进行。

因此我们需要尝试另一种选择值的方式。这次我们将有多个而不是一个acceptors。proposer将会向一个acceptor的集合发送value。acceptor可能会接受value。但是该value只有在足够多的acceptor都接受它的情况下才算被选择了。那么怎样才算足够大呢？为了确保只有一个value会被选中，我们可以认为一个足够大的agent集合由任意的agent majority组成。因为任意两个majority都至少有一个公共的agent，因此如果一个agent最多只能接收一个value，那么这种方法是可行的。

在没有故障和message丢失的情况下，我们想要有一个value能被选中，即使只有一个proposer提出了一个value。这就需要满足以下要求：

**P1. acceptor必须接收第一个它收到的proposal**

但是这个要求会引起这样一个问题。不同的proposer可能会在几乎同时提出好几个不同的value，这会导致这样一种情况：每个acceptor都接受了一个value，但是没有一个value是被一个majority接受的。即使只提出了两个value，而它们各自被一半的acceptor所接收，那么任意单个acceptor的故障都将让我们无法获取它选择了哪个value。

P1以及value只有被majority个acceptor接受才被算被选中的要求就导致了我们的acceptor必须能接受超过多于一个的proposal。我们通过给每个proposal赋予一个编号来追踪不同的proposal，所以一个proposal由一个proposal number和一个value组成。为了防止出现歧义，我们要求不同的proposal要有不同的number。这里我们仅仅只是做出这个假设，具体的实现可能有所不同。当一个proposal被一个acceptor的majority所接收时，我们就认为该value被选中了。这种情况下，我们说这个proposal(同时也包括它的value)被选中了。

我们可以允许多个proposal被选中，但是我们必须保证被选中的proposal有相同的value。通过归纳proposal number，足以保证：

**P2.如果一个value为v的proposal被选中，那么所有被选中的higher-numbered proposal都具有value v**

因为number都是有序的，条件P2保证了只有单一的value被选中这一重要特性。为了能够被选中，proposal必须至少被一个acceptor所接受。所以我们可以通过满足以下条件满足P2：

**P2a.如果一个value为v的proposal被选中，那么每一个被任何acceptor所接受的high-numbered proposal都有value v**

我们依然需要满足P1从而确保有proposal被选择。因为交互是异步的，一个proposal可能被一些特定的，没有接受过任何proposal的acceptor c选中。假设一个新的proposal "wake up"并且发送了一个带有不同value的high-numbered proposal。P1要求c接受这个proposal，但是却违背了P2a。为了同时满足P1和P2a,需要加强P2a：

**P2b.如果一个value为v的proposal被选中，那么之后每个proposer发出的high-numbered proposal都有value v**

因为一个proposal 在被acceptor接受之前都要首先由proposer发出。因此满足P2b就满足了P2a,也就满足了P2。

为了明白如何满足P2b,我们先对它进行证明。我们假设有一些number为m，value为v的proposal已经被选中了，接下来我们证明任何的有number n>m的proposal的value都为v。我们可以通过归纳到n来简化证明，首先假设每一个发出的number在m...(n-1)的proposal都有value为v，其中i...j代表范围为i到j的所有number。既然有number为m的proposal被选中了，那么必然有这样一个集合C，它由一个acceptor的majority组成，而C中的每个majority都接受它。结合归纳假设，从m被选中的假设可以推出：

C中的每一个acceptor都接受了number在m...(n-1)中的一个proposal，而每个被任意acceptor接受的number在m...(n-1)的proposal都有value为v。

因为任何集合由majority个acceptor组成的集合S必然和C存在公共的元素，我们可以通过满足以下条件来确保标号为n的proposal有value为v：

**P2c.对于任意的v和n，如果有一个value为v，number为n的proposal被发出了，那么就存在一个由majority个acceptor组成的集合S。要么(a)S中没有一个acceptor接受过number小于n的proposal，要么(b)v是S中的acceptor接受过的number小于n的最高number的proposal的value**

因此我们可以通过满足P2c来满足P2b。为了满足P2c，如果proposer想要发出一个number为n的proposal，那么它必须要获取已经或者将要被一个majority接受的最高number不大于n的proposal的value。获取那些已经被接受的proposal是非常简单的，但是对未来的接受情况进行预测就非常困难了。为了避免对未来进行预测，proposer通过获得不会有这样的接受情况的承诺来进行控制。换句话说，proposer请求acceptor不要接受任何number小于n的proposal。这就导出了以下发送proposal的算法：

1、一个proposer选择了一个新的proposal number n并且给一些acceptor集合的每个成员发送了一个请求，并希望它们回复：

（1）、承诺再也不接受number小于n的proposal

（2）、如果已经接受了最高number小于n的proposal，返回

我们将这样的请求称为number为n的prepare request

2、如果proposer接受了来自一个majority对于请求的回复，那么它就可以发送一个number为n，value为v的proposal。其中v要么是返回的最高number的proposal的value，如果没有proposal返回，那么proposer随意选择一个值。

proposer将proposal发送给一些acceptor的集合，请求它们接受。（这里的acceptor的集合不一定要和回复初始请求的acceptor的集合是同一个集合）我们将这样的请求称之为accept request。

这里描述的是proposer的算法。那么acceptor呢？它可以从proposer那里接受到两种请求：prepare request和accept request。acceptor可以忽略任何没有compromising safety的请求。因此我们只需要讨论它允许回复请求的情况。它总允许对prepare request进行回复。它可以对一个accept request进行回复，代表接受一个proposal，如果它没有承诺不那么做的话。换句话说：

**P1a.acceptor可以接受number为n的proposal，如果它之前没有回复任何number大于n的prepare request**

可以发现P1a包含P1

现在我们已经为满足安全性要求地选择value提供了一个完整的算法——假设proposal number唯一的情况。最终的算法只是在这基础之上做了一个小小的优化。

假设一个acceptor接受了一个number为n的prepare request，但是它已经回复了一个number大于n的prepare request。因此它承诺不会接受任何number为n的新的proposal。但是该acceptor是不会接受该proposer想要发出的number为n的proposal的，所以该acceptor没有理由回复这个prepare request。因此，我们让acceptor忽略这样的prepare request。同样，我们对于已经接受的proposal对应的prepare request也是忽略的。

经过这样的优化以后，acceptor只需要记住它接受过的最高number的proposal以及它回复过的最高number的prepare request。即使发生了故障也要满足P2c,因此即使发生了故障或者重启了，acceptor也要记住这些信息。需要注意的是proposer总是可以放弃一个proposal并且将它忘得一干二净——只要它不再发送另一个具有相同number的proposal。

将proposer和acceptor的行为结合在一起，我们可以看到算法分两阶段执行。

**Phase 1.**（a）proposer选择一个proposal number n，然后向一个majority发送number为n的prepare request。

（b）如果一个acceptor接受了一个number为n的prepare request，并且n大于任何它已经回复的prepare request的number，那么它将承诺不再接受任何number小于n的proposal，并且回复已经接受的最高number的proposal（如果有的话）。

**Phase 2.**（a）如果proposer接受了来自majority对它的prepare request的回复，那么接下来它将给这些acceptor发送一个number为n，value为v的proposal作为accept request。其中v是收到的回复中最高number的proposal的value，或者如果回复中没有proposal的话，就可以是它自己选的任意值。

（b）如果acceptor收到一个number为n的accept request，如果它没有对number大于n的prepare request进行过回复，那么就接受该accept request。

一个proposer可以生成多个proposal，只要它能满足算法的要求。它可以在协议的任意时刻放弃proposal。（正确性依然是得到满足的，即使请求或者回复在对应的proposal被放弃了很久之后才到达目的地）当其他的proposer已经开始发出更高number的proposal时，最好放弃当前的proposal。因此。如果acceptor因为它已经接受了更高number的prepare request而忽略了其他的prepare或者accept request，那么它应该通知对应的proposer放弃该proposal。这是对性能优化，并不会影响正确性。

### 2.3 Learning a Chosen Value

为了获取一个已经被选中的value，learner必须要确定已经有一个proposal被majority接受了。最显而易见的算法就是让每个acceptor在接受了一个proposal之后向所有的learner发送这个proposal。这能让learner尽快地找到被选中的value，但这需要acceptor对每个learner进行回复——回复的数量为acceptor和learner数量的乘积。

non-Byzantine failures的假设允许我们让一个learner可以从另一个learner处获取已经被接受的value。我们可以让acceptor把它们的接受情况都发送给一个distinguished learner，这个learner再转而通知其他的learner有个value被接受了。这种方法需要额外的一个round来让所有的learner发现被选中的value。同时这样也是非常不可靠的，因为这个distinguished learner可能故障。但它只需要acceptor和leaner数目的和的回复数。

更一般地，acceptor可以将它们的接受情况发送给一个distinguished learner的集合，而它们中的任意一个都能在value被选中的时候通知所有的learner。通过提供更大集合的distinguished learner在增加通信复杂度的同时提供了更高的可靠性。

因为message的丢失，可能没有learner知道已经有value被选中了。learner可以直接问acceptor它们接受了什么proposal，但是acceptor的故障可能让我们无法知道是否有majority接受了一个特定的proposal。这种情况下，learner只能在有新的proposal被接受的时候才能确定被选中的value是什么。如果一个learner想要知道一个value是否被选中，它可以让一个proposer发送一个proposal，使用上文描述的算法。

### 2.4 Progress

我们很容易构建这样一个场景，两个proposer持续发送比对方的number高的proposal，并且最终它们两者没有一个被选中。Proposer p通过proposal number n1完成了phase 1。另一个proposer q接着通过了proposal number n2 > n1完成了phase 1。proposer p在phase 2的以n1标记的accept request会被所有acceptor拒绝，因为它们已经承诺不接受任何number小于n2的proposal。因此proposer p开始用新的proposal number n3 > n2来开始并完成phase 1，而这又导致了proposer q在phase 2的accept被忽略。如此反复进行。

为了保证流程的执行，我们必须选出一个distinguished proposer，作为唯一的proposal发送者。如果distinguished proposer能和majority进行通信，并且使用了一个比所有已经使用的proposer number都大的number，那么它就能成功发送一个已经被接受的proposal。通过拒绝已有的proposal并且通过更高的proposal number重试，distinguished proposer最终会选择到一个足够大的proposer number。

如果系统足够多的部分都工作正常（proposer, acceptors以及交互网络），那么通过选出一个单一的distinguished proposer就能保持系统的活力。由Fischer, Lynch, and Patterson著名的结论可得，选举一个proposer的可靠算法必须要么使用randomness，要么使用real time——比如，使用超时。但是，无论选举成功还是失败，安全性总是可以保证的。

### 2.5 The Implementation

Paxos算法假设了一个进程网络。在这个共识算法中，每个进程扮演着proposer, acceptor和learner的角色。该算法需要选择一个leader，来扮演distinguished proposer和distinguished learner的角色。Paxos共识算法正如上文所描述的那样，请求和回复都以ordinary message的形式发送。（Response message会用相应的proposal number标记为了防止混淆）我们需要使用stable storage（会在故障时候保存）来维护那些acceptor必须保存的信息。acceptor会在真正发送response之前将它记录下来。

接下来所有的内容都将用于描述如何保证两个proposal不会有相同的number。不同的proposer会从不相交的数据集中选择number，所以不同的proposer不会发送具有相同number的proposal。每个proposer都会用stable storage记住它尝试发送的最高number的proposal并用一个比所有已经使用过的number都高的number开始phase 1。

## 3 Implementing a State Machine

实现一个分布式系统最简单的方式就是一个client的集合向一个central server发送命令。central server可以被看做是一个以一定顺序执行client命令的deterministic state machine；它通过将输入作为命令，并产生输出和一个新的状态。比如，分布式银行系统的client可以看做是teller，而所有用户的account balancer可以看做state-machine的状态。一个撤销操作可以通过执行当且仅当balance大于amount withdrawn的时候减小account's balance这一state machine command完成。

使用单一的central server的实现，如果central server发生故障，整个系统就会发生故障。因此，我们使用了一个server的集合，它们各自独立地实现一个state machine。因为state machine是确定性的，所以在执行完相同序列的命令之后，所有的server都会产生相同的状态序列和输出。client可以使用任意server的输出。

为了保证所有的server执行相同序列的state machine commands，我们实现了一个序列的Paxos共识算法的单独实例，第i个实例选择的value作为序列中第i个state machine command。每一个server都扮演了该算法中的所有角色（proposer，acceptor和learner）。现在，我假设server集合是固定的，因此该共识算法的所有实例都使用相同的agent的集合。

在通常的操作中，只有一个server能够被选为leader并作为distinguished proposer（唯一的proposer发送者）在共识算法的所有实例中。client将命令发送给leader，leader决定每个命令应该放在序列的哪个地方。如果leader决定一个特定的client命令应该作为第135个命令，那么它就会将该命令作为共识算法第135个实例。这通常都会成功。但也有可能因为发生故障或者有另一个server认为自己是leader并且对第135条命令是什么有它自己的想法。但是共识算法确保了第135条命令最多只有一个。

Paxos共识算法效率的关键在于直到phase 2之前都不对提出的value进行选择。回忆一下，是在完成了phase 1之后才知道要发送的value要么已经被决定了，要么proposer可以被任意选择。

我现在要讨论的是在正常执行时，Paxos state machine是怎么工作的。之后，还会描述什么情况下会出错。我考虑的是当前一个leader刚刚发送故障但是新的leader还没有选举出来的情况（系统刚刚启动时是一个特殊的情况，那时候还没有任何命令被提交）。

新的leader，也是共识算法所有实例的leader，应该了解已经选择的大多数命令。假设它知道命令1-134,138和139——即实例1-134，138和139选择的值（接下来我们会知道命令序列中的gap是怎么产生的）。之后，它将执行实例135-137的phase 1以及所有大于139的实例（下面我将描述这是如何完成的）。假设这些操作的执行结果确定了实例135和140的value，但是其他实例的value还是未确定的。之后，leader将会执行实例135和140的phase 2，从而选择了命令135和140。

leader以及那些获取了leader已知的所有command的server现在可以执行命令1-135。然而，它仍然不能执行命令138-140，即使它已经知道它的内容了，因为命令136和137并没有被选择。leader可以将接下来client请求的两个命令作为命令136和137。然而，我们通过发送特殊的让状态不发生改变的"noop"命令来马上填充gap（通过执行共识算法的实例136和137的phase 2来实现）。一旦这些no-op命令被选中，命令138-140就可以执行了。

命令1-140已经选择完毕。leader也完成了共识算法所有大于140的实例的phase 1。它将client发送的下一个请求赋值为141，并且将它作为共识算法实例141的value。之后再将用户的下一个请求作为命令142，如此往复。

leader可以在它知道已经发送的命令141被选择之前就发送命令142。可能发送命令141的所有数据都会丢失，命令142也可能在所有server都不知道leader发送的命令141的任何内容之前被选择。当leader没有收到它希望得到的关于实例141的phase 2信息的回复时，它会对这些信息进行重发。如果所有运行正常的话，发送的命令将会被选中。然而，在一开始可能会发生故障，从而在已选中的命令序列中留下一个gap。一般来说，假设一个leader可以提前获取α个命令——这意味着在命令1到i被选中的前提下，它可以发送命令i + 1到i + α之间的命令。因此，一个至多为α−1条命令大的gap可能会出现。

一个新的被选中的leader可以执行无数多个共识算法实例的phase 1——在上面的场景中，即为实例135-137以及所有大于139的实例。通过对所有实例使用同一个proposal number，它可以用给其他server发送一个single reasonably short message来实现。在phase 1，如果一个acceptor已经从一些proposer收到phase 2信息的时候，它就会不仅仅回复一个简单的Ok。（在例子中，就是对于实例135和140）因此，server（作为acceptor）可以用一个single reasonably short message来回复所有的instance。在无数个实例的phase 1这样执行不会产生任何问题。

因为leader的故障和新的leader的选举都是小概率事件，因此执行state machine command的花费——即实现command/value的共识——主要是共识算法phase 2的执行。可以看出Paxos共识算法的phase 2在所有会出现故障的情况能达到共识的所有算法里有着最小的可能花费。因此，Paxos算法基本上是最优的。

关于系统执行的正常操作假设除了当前leader发生故障，新的leader还未选出的短暂时间外，总是存在一个单一的leader。在一些意外的情况下，leader的选举可能发送故障。如果没有server作为leader执行，那么就不能有新的命令被发送。如果有多个server认为它们是leader，那么它们可以对共识算法的同一个实例发送value，而这会防止任何value被选择。然而安全性总是被保留的——两个不同的server用于不会不同意已经被选为第i个state machine command的value。单一leader的选举仅仅只是为了保证流程的执行。

如果server的集合是可以改变的，我们必须要有办法确定哪些server实现了共识算法的哪些实例。实现这个最简单的方法就是通过state machine它自己。当前的server的集合可以作为状态的一部分并且可以随着ordinary state-machine command而改变。在执行完第i个state machine command之后，我们可以让leader提前获取α个命令，通过让server的集合执行共识算法的第i + α个实例。