分布式算法讲义

(V0.3)

黄 宇

2024 年 11 月 4 日

目录

前言

以前试图写一个重量级的授课讲义，推进非常缓慢。现在尝试写一个轻量级的讲义，以主要文献的推荐和简评为主，推进一段时间，试试效果。

对学习分布式算法比较深入的学生，我经常会问他一个问题，你是把分布式算法，当一个数学对象来看，还是当一个软件对象来看。换一个更“实用”的角度来说就是，你学了分布式算法，是去搞理论计算机科学、数学，还是去搞分布式系统软件。

这是一个基本的定位问题。这本讲义，主要还是将分布式算法当一个软件对象来看，主要面向的是分布式系统的建设者，特别是有一定的实践经验，理论基础比较薄弱的建设者。里面有一些理论性的结果与讨论，也被看作分布式系统构建的further reading。

根据我与系统软件开发人员的接触，以及看他们在网络发布的技术内容，我发现有一部分分布式系统软件开发人员，基本概念不够清晰，相对严谨的建模与论证能力尚有不足。从我的角度而言，目前分布式算法课本是缺位的，一些翻译的国外经典教科书，对于分布式系统开发人员也是不太合适的。希望（未来）本书的内容能对他们有所帮助。

此外，在校的同学，从本科生的算法课一路学下来，学了分布式算法的基本概念，再去实际分布式系统的构建中去深入体会，也是可以的。每年招新学生的时候，分布式算法知识的入门是一个反复出现的任务。一般而言，分布式算法知识和大家的本科课程距离较远，难以自然地衔接。分布式算法的经典课本、论文在本科毕业的时候直接阅读往往比较困难。另外一方面，网上相关技术内容不少，但是质量堪忧，有误导性的“负作用”。为此陆续整理这一本讲义，希望能将每年重复给新学生罗列的学习内容沉淀下来，复用起来。

这本讲义完全是“on demand”式地逐步添加的，而且论述非常简要，更像是向学生讲解所需阅读的入门文献时，所做的注解。未来有一天，如果这本讲义的内容充实到一定程度，或许它能成为一本合格的《分布式算法入门》课本。

分布式算法的系统学习，对本科生《算法设计与分析》课程的学习有一定的要求。具体知识点的多少，是次要的，重要的是学习和思考的范式，两门课程是一以贯之的。进一步内容可以参考我的本科生算法设计与分析课程：《战疫时期的算法课（2020年春季）》、《黑板上的算法课》（2023、2024年春季）等[1]。

[1]视频发布在我的B站空间：https://space.bilibili.com/474662253

第一部分

分布式系统建模

这一部分讲解计算模型的基本概念。

这一部分还讲解分布式算法的正确性规约，它脱胎于分布式系统设计的需求规约，以及分布式算法的设计者对需求规约的“强化”。

第一章　抽象计算模型的概念

1.1　基本模型

我们讨论抽象的分布式算法设计与分析，主要基于消息传递模型和共享内存模型。这里将主要讲解计算模型的基本组成部分。而模型的具体细节，将放到具体的章节（第三章、第十章）去讲。

我们首先讨论计算模型的两个基本的维度。一个维度是通信的载体，一个维度是时间模型。

通信的载体主要包括消息传递（Message-passing，MSG）和共享内存（Shared Memory，SHM）。我们后续的章节也是按照这个维度来组织。

时间模型主要包括异步模型（asynchronous model）和同步模型（synchronous model），也包括更深入的半同步模型（partially synchronous model）。

✍ 参考两本教材中系统模型相关章节[Asp19, AW04]。[Asp19, Sec 2.1.1]，作者对两本书（[Asp19]和 [AW04]）中，对于“消息传递系统”不同建模方式的“异”和“同”进行了比较，并回顾了相关的历史。理解不同建模背后共性的部分，是学习如何对分布式系统进行合理抽象，并进行后续的问题定义、算法设计、算法分析的基础。 □

1.2　建模进阶

建模中的一个重要概念是计算模型之间的模拟（simulation）。

✍ Simulation的内容可以参考[AW04, Part II]。作者Attiya的代表工作[ABND95]是在MSG模型上，模拟一个SHM的atomic register。因为其研究的原因，Attiya在写教材的时候，也比较偏重模拟技术的讲解与应用。例如对于异步环境中共识不可能性的结论，她就专门使用模拟技术来证明。而直接证明的方式其实更有名，也非常有学习意义，它就是著名的FLP的证明[FLP85]。 □

一个与“模型之间的模拟”有些类似，又有重要区别的概念是“问题之间的归约（reduction）”。在深入研究分布式计算的理论问题的时候，这两个技术是以既有结论为跳板，更“省劲”地证明新结论的利器。

其它还有一些高级的系统模型，可以作为主要模型的补充，对照着做一些初步的了解。

✍ 对于一些更高级模型的介绍，参考[Asp19, Part III]。 □

1.3　编程抽象

将系统建模推到一个更深层次的做法是，以“编程抽象”为核心载体，去解构一个分布式系统。对于分布式系统的使用者而言，分布式系统就是不同层级的各种编程抽象；对于分布式系统的构建者而言，其核心任务就是用设计层的算法、协议和系统层的代码去实现不同的编程抽象。在编程抽象的视角下，分布式系统就是层层编程抽象的组合，一个再复杂的系统，都像是乐高搭成的，都可以解构为更基本的单元。

这一编程抽象的视角，对于系统学习分布式系统理论而言，是非常有帮助的。对于实际开发者而言，它的主要意义在于将核心的概念，层层解构，辨析清楚。

✍ [CGR11]这本书就是基于编程抽象，来组织分布式系统编程相关内容的。 □

第二章　分布式算法的基本性质

2.1　Safety and Liveness

本科生算法课中讨论的正确性规约，是必要的基础。基于此，我们可以发现这一规约对于讨论分布式系统做的对不对，是不够的。

分布式系统中的正确性讨论，主要基于safety和liveness这两个核心概念，它们是串行程序中“partial correctness”和“complete correctness”概念的泛化。

✍ 一直以为这两个核心概念是Lamport首创，[Mal19]的Introduction章节中也有提及。但是后面又看到这两个概念是在[AS85]中定义的（这一工作获得了2018年分布式计算理论领域的Dijkstra奖），Lamport应该是对于异步分布式算法，进一步深化了这两个概念。 □

上述基本性质之外，还要讨论几个更深入的性质。

公平性（fairness）。

模拟（simulation）。从外在使用者的角度，它是计算模型之间的转换。从模拟算法设计者的角度，它是一种性质或者说规约。

下界与不可能性（impossibility）。这不是单个算法的性质，它是一个问题的所有可能算法的整体性质。

✍ 上述性质的讨论，主要参考了[Asp19, Sec 1.2] □

2.2　正确性规约的由来

分布式系统设计的正确性规约，显然来自于用户的需求。但是除此之外，分布式系统的正确性规约还有一类更“微妙”的来源，这和分布式系统设计的基本范型（paradigm）有关。

分布式系统实际的核心挑战是，如何让独立的，只有局部知识的节点，共同完成一个全局性的任务。应对这一挑战的主要手段，是设计一些“精巧”的全局不变式。不同节点间的协作，致力于维持这一全局不变式，而全局不变式的成立蕴含了用户需求的满足。

所以在很多时候，例如在分布式算法/系统验证的时候，我们所面对的正确性规约就是这样的全局不变式。具体的案例可以参考§6.2和§6.3。

第二部分

消息传递算法

这部分讲解MSG模型上的经典分布式算法。

第三章　MSG模型

第一章主要是介绍基本概念。MSG模型的技术细节，拓展性讨论将在此展开。

节点端主要使用一组变量in_buf和out_buf来建模消息的收发。

对于一个包含n个节点的系统，节点间通信主要通过信道$channel_{ij}(1 \leq i, j \leq n)$来建模。

✍ MSG模型的具体内容参考[AW04, Sec 2.1]、[Asp19, Chap 2]。 □

第四章　领导者选举问题

领导者选举问题（Leader Election，LE）是一个很重要的理论、技术问题。

4.1　LE问题的定义

在具体算法之前，有两个理论的概念需要辨析：何谓一个LE算法是anonymous的？何谓一个LE算法是uniform的？对于“anonymous”性质的辨析，使得我们从理论上认识到，LE算法中，假设每一个process有一个ID是合理的，且是必须的。

算法从n^2向nlogn的改进，让我们想起了熟悉的comparison-based sorting。

这些课本上的LE算法，特别是同步模型下的LE算法，它们对process的ID值的应用是让程序员感到“不适”的，感觉更像是在做智力游戏，而不是在真正解决分布式系统中的实际问题。这一感觉是对的，上述算法的讲解，主要目的就是帮助学习者厘清基本概念，掌握基本理论知识，不直接指导系统的构建，但是为系统构建打下理论基础。

下界相关的结果暂时略去，它们属于impossibility results的专题。

✍ 课程的讲解主要参考[AW04, Chap 3]中的算法。 □

4.2　一点引申的讨论

对于先入门了一点分布式系统开发，再学习分布式系统基本概念的人而言，有一个有意思的“鸡生蛋、蛋生鸡”的辨析：共识算法需要选举一个leader保证liveness（如[Lam01]）。同时共识算法（包括与共识等价的atomic broadcast算法）又可以支持LE机制的实现（如ZooKeeper官方的“Recipes and Solutions”）。那么，到底是使用共识实现LE，还是有了LE才能实现共识呢？

第五章　基本共识算法

5.1　问题的定义

首先需要对共识问题的严格定义，仔细地学习与体会。特别是validity的概念，初学的时候容易get不到它在说什么，为什么需要它。

✍ [AW04]比较简单直接的引入共识问题。[Lyn96]的讨论比较系统。 [Asp19]的讨论，看上去比较“凌乱”地分散在好多章节中，问题、模型都在变化，但其实大概这就是共识问题，及其解法，形成的过程。所以，入门可以看第一个，学院派、学习者先看第二个，有点融会贯通的，看第三个。 □

共识算法的正确性，包含其safety和liveness。这两类性质的保证，一般就是两套“割裂”的机制，所以在理解一个共识算法的设计时，要把这两方面的property分开来看。

✍ 基础的Paxos算法中，只谈safety，而把liveness“外包”出去 [Lam01]。Zab协议也是把leader election机制外包出去，论文中不谈 [HKJR10, JRS11]，后来又补充了一个FLE (Fast Leader Election)的机制[Med12]。第6.4.1小节提供了一个鲜明的例子，它使用randomness来保证liveness。 □

5.2　课本上的共识算法

共识算法是自媒体上的热门topic，似乎分布式系统方面的程序员没有不懂共识算法的，说起paxos，raft，zk，拜占庭容错，区块链起来，熟得不行。

相比之下，课本上的共识算法就显得有些落寞。但是要真正理解工业级的共识协议及其系统实现，基本的概念，平淡无奇的“课本共识算法”是必要的铺垫。

[AW04, Sec 5.1.3]讲了一个非常简单，非常蛮力的共识算法。但是它对大家理解共识，理解如何利用冗余性来容错，有很重要的帮助。再厉害的共识算法，其实并没有跳出这个蛮力算法的路数。

同样对于拜占庭容错，课堂上我们也只讲蛮力算法[AW04, Sec 5.2.4, Sec 5.2.5]。大概的原理是，充分多的好人，充分交换意见。这样的话，真话占明显多数，少量的假话，存在也不影响，它们会因为跟大多数真话不一样，而被识别出来。

拜占庭共识算法总体上相对来说是难的，但是我们在课堂教学中，能够接受某些维度的不计成本。在这一前提下，它的难度是不大的。因而，这类算法更多是教学讲解的意义，它们并不是瞄着系统中的实际应用去的。它们更像是在帮助大家认识这个问题，而不是帮助大家得到一个实用

的solution。

学习共识算法，很有必要了解共识算法相关的一些不可能性结果。这一内容放在不可能性证明部分，单独讨论。

第六章　类Paxos共识算法族

包括Paxos [Lam01]，Raft [OO14]和Zab [HKJR10, JRS11]。

严格地讲，Zookeeper的Zab协议是一个atomic broadcast协议，由于atomic broadcast 和 consensus 两个问题可以互相规约，是等价的[CT96]，所以我们也经常不仔细区分，都称之为共识算法。

6.1　Paxos

6.1.1　基本Paxos算法

6.1.2　Multi-Paxos

6.2　Zab

6.3　Raft

6.4　Paxos算法族中的其他变体

6.4.1　复制状态机框架Rabia

除了一些关系比较紧密的Paxos变体外，还有一些改动相对大一点的共识算法。它们仍然算是脱胎于Paxos算法，它们的设计也主要与Paxos的各种变体作充分对比。所以这类算法仍然放在这一章。

专门提到Rabia的原因是，它的改进在方法学层面有比较重要的意义。

我们知道，共识算法的正确性要把safety和liveness分开来谈。[Lam01]中的基础Paxos算法，只能保证saftety，它把liveness分离开来单独谈。

由于FLP等一系列不可能性结果，异步环境的共识要保证liveness，必须要一些“额外的信息”。常见的额外信息包括时间信息、随机性等。

此处的Rabia就是利用randomness，保证了livess。并且Rabia还很高效地保证了liveness。一方面，用randomness保证liveness，省去了传统共识算法中leader选举，以及leader失效时选举新leader等复杂的机制。另一方面，面对单数据中心场景中网络条件比较好的有利条件，Rabia还保证了良好的性能。

✍ The Rabia SMR framework [PTZ+21]. □

第七章　VR共识算法族

7.1　背景和历史

背景知识和历史发展过程介绍[CBPS10, Chap 7]。

7.2　VR算法

✍ Liskov的这一系列算法，受2PC的影响比较大。为了更好地理解这一族算法，大家需要先了解一下基本的2PC算法，参见第九章。 □

7.3　PBFT算法

7.4　高山仰止：Paxos和VR

Lamport和他提出的Paxos不仅厉害，而且传奇。开始Lamport提出了Paxos算法，由于他的表述方式过于geek，导致读懂的人不多。对分布式方面的程序员而言，Paxos的出名是由于Google的推动。转述6.824（Spring 2021）的讲述，共识算法开始只是象牙塔里的一个结果而已，在众多结果中，也未见有特别过人、出奇之处。当多数据中心的平台逐步出现，构建大规模分布式系统的需求出现的时候，它的理论基础基本是就绪的，缺的就是一个团队真的把象牙塔里的协议给实实在在地实现出来。最早出现的这个团队是Google，它们实现了Paxos [CGR07]。后续大家都知道了，几乎每个相关的公司、程序员、自媒体，都来讲讲Paxos。

不仅Paxos有很多直接的变体，在回顾Lamport一身工作的这本书中[Mal19]，把几乎所有的共识算法，都看成是Paxos的某种变体。这种提法，是有一定道理的。理解不同共识算法之间的本质联系、成长演化的历程，是全面深入学习共识算法的必要条件。

有一个跟Paxos同一时期提出的共识算法叫Viewstamped Replication（VR），它的提出者是2008年图灵奖得主Barbara Liskov。在Paxos如此热门的大背景下，很多人不知道这个算法，或者知道但是大大低估了它的价值。

说VR多么多么厉害，可能有点偏颇，准确的说是要把VR放到Liskov一系列的研究中去看，Liskov这一系列的工作是“高山仰止”级的。Liskov在1980年左右，从data abstraction，CLU语言设计的工作，逐步转到（广义的）分布式数据库系统方面的工作。在这一过程中，为了提高系统的可靠性，她受2PC协议的启发，设计了VR算法。

VR本身不是太有名，或者说风头完全被Paxos盖住。但是对于拜占庭容错，经典的PBFT算法就是VR的派生。再考虑到后来Liskov在分布式事务方面的奠基性工作[Ady99]等，Liskov的这一系列工作是不输于Lamport的Paxos系列工作的。只不过在这一系列的工作中，VR是略显普通的一份子。

这一段历史，Liskov在[CBPS10, Chp. 7]中自己有简要论述。

第八章　分布式系统执行的观察与理解

8.1　分布式系统的全局切片

理解、构建一个异步MSG系统，一个重要的观念的跃迁，是从“绝对的、全序的时钟观”变成“相对的、偏序的时钟观”。形成这种异步分布时钟观的核心是消息传递的因果关系蕴含的时序happen-before关系和异步系统中时刻的概念：分布异步系统的consistent global snapshot。

异步分布式消息传递系统的执行轨迹（称之以为一个distributed computation，也有很多其它名称），天然自带happen-before关系。这一happen-before关系的定义参见Lamport奠基性的论文[Lam78]。这一happen-before关系可以用各种不同的逻辑时钟来表示。各种逻辑时钟机制，可以看成是表示偏序关系的代价和表示偏序关系的能力之间的权衡。虽然Lamport那篇奠基性的论文名气很大，但是学习逻辑时钟，建议从最标准、表达能力最强、代价最大的vector clock入手[Mat89]。

✍ 作为准备知识，需要读者对离散数学中学的偏序关系、格结构，有较为深入的了解。我为本科生通识课，做过一次相关的报告，可供参考：偏序集与格理论-及其在分布式系统中的应用[1]。 □

如何得到一个分布式系统的snapshot，可以参考[CL85]这一先驱性的经典工作。随着分布式系统的不断进化，各种snapshot技术也随着发展。近期snapshot技术，可以参考[YZZ+24]等。

8.2　观察：切片集合的格结构

了解happen-before关系及其vector clock编码之后，就可以正式理解distributed computation中的snapshot和snapshot之间的格结构，主要参见[Mat89]和[BM93]。

8.3　理解：全局谓词的规约和检测

把distributed computation的格结构刻画清楚之后，就可以从中提取观察者所感兴趣的信息。观察者所关注的，distributed computation所具有的性质，可以形式化规约为一个全局谓词（global predicate），每个谓词有相应的检测算法。

基于上面的准备，观察和理解distributed computation的基本过程是：

- 观察系统执行，得到trace。
- 解析用户规约的全局谓词。

[1] https://www.bilibili.com/video/BV1iU4y1M7SU/

- 在trace上检测谓词的成立情况。

谓词规约和检测的一些初步知识参见[CL85, CM91, BR95]等。一篇关于global temporal谓词的更深刻的理论分析见[CBDGF95]。

第九章　分布事务提交2PC算法

9.1　分布事务系统模型

分布式云原生数据库的基本结构。

分布事务系统建模。

数据的划分。数据的复制。

原子性提交（Atomic Commit，AC）。并发控制（Concurrency Control，CC）。基于复制的容错技术（Fault-tolerance based on Replication，RE）。

✍ 上述系统模型，包括下面的算法，主要参考[SKS20, Chap 23]。 □

9.2　2PC算法

原始的2PC算法。

9.3　基于共识的容错2PC算法

用共识解决原始2PC coordinator单点失败的问题。

第三部分

共享内存算法

这部分讲解SHM模型上的经典分布式算法。

第十章　SHM模型

第一章主要是介绍基本概念。

SHM模型的技术细节，拓展性讨论将在此展开。

主要讨论异步SHM模型。同步SHM模型在PRAM中有讨论，主要属于并行算法的范畴，这里不讨论。

主要的载体是共享寄存器（shared register）。对于共享寄存器，首先讨论它的访问方式（access pattern），然后讨论它的一致性模型。

我们这里将以原子寄存器（atomic register）为例，讲解SHM模型的基本概念。对于SHM一致性模型的更全面的讨论将在第十一章展开。

✍ SHM模型具体内容参考[AW04, Sec 4.1]和[Asp19, Chap 15]。 □

第十一章　SHM一致性模型

[AW04, Chap 4]所讲解的SHM，其实是一个退化版。因为对shared register的更新是原子的，所以它其实讲的是SHM上的atomic register抽象。Atomic register是一个非常重要的抽象，它是深入理解SHM模型的支点。它的原始定义参考[Lam86a, Lam86b]。

在MSG模型上，模拟出SHM模型中的atomic register抽象，是分布共享内存DSM中的经典问题。所以在一些DSM的经典工作中，也可以读到atomic register的定义，而且读起来可能比SHM相关文献中的定义还好懂一些，包括Attiya的DSM奠基性工作[ABND95]，Welch关于multi-writer情况的讨论[SWPL11]，atomic register fast access下界证明的经典工作[DGLV10]，也包括我们自己下界证明的工作[HHW20]和ASC(Almost Strong Consistency)的工作[WHL17, OHWL21]。

有了atomic register抽象的铺垫之后，可以更深入地了解SHM的一致性模型，包括带时间的atomic / regular / safe registers [Lam86a, Lam86b, SWPL11]。还包括不带时间的各种模型，及其背后的统一框架[SN04]。这个统一框架很深刻，把所有一致性模型都打碎，碎成乐高一样的原子块，进而所有模型都可以挑若干原子块搭建起来，并且通过上述解构，比较容易看出所有模型在强弱关键下，具有lattice结构。

第十二章　分布共享内存

12.1　基本概念

分布共享内存(Distributed Shared Memory, DSM)的概念可以从理论和系统两种不同的视角来辨析。

从理论的视角，它就是在MSG模型上，模拟出SHM的假象。我们的讨论主要基于这一视角，并且主要讨论atomic register的MSG实现。其它更复杂的DSM概念都可以基于此来逐步深入学习。

从系统的视角，DSM最早在[LH89]中提出。随着多数据中心平台的出现，随着分布式云储存、NoSQL、NewSQL、云原生数据库等新型计算平台上新型数据密集型基础设施软件的出现，DSM可以看成是上述实际软件系统的理论抽象。虽然上述实际系统比DSM概念所描绘的系统要复杂很多，但是DSM的概念是上述系统构建的基础铺垫与支撑。

12.2　Atomic register的MSG实现

Atomic register可以进一步分为single-writer和multi-writer的情况。Single-writer情况的MSG实现是Attiya的代表性工作[ABND95]。[LS97]中把它推广到了multi-writer的情况。

不过早期的论文都比较难懂，相关领域深入研究的读者顶多可以看看 [ABND95]，[LS97] 基本不用直接去看了。[Asp19, Chap 16]有提炼后的重新表述，从学习的角度而言，看课本就可以了，很多不必要的细节都被合理抛弃了。

12.3　DSM的高效实现

在实际系统中，DSM受到效率问题的制约，其应用并不广泛。为了提升DSM实际实现的效率，一种可行的办法是利用DSM中数据访问的特定模式，做针对性的优化。

在近期的一份工作中，研究人员利用了Rust语言中“ownership model”的概念，针对性地实现了一个支持SWMR regisgter的高效地DSM [MQL+24]。

这一ownership model的概念，以及SWMR shared regisgter，我们在经典的Bakery算法中（参见§13.3）会再次看到它的“simple but powerful”的应用。

✍ 在[MQL+24]中包含对于DSM系统的一个mini survey，关注DSM系统的读者可以用它来“快速入门”。 □

第十三章　互斥问题

有了SHM模型的基本概念之后，就可以比较系统地了解互斥问题（Mutual Exclusion，MUTEX）。

13.1　问题的定义

需要从safety和liveness两个侧面来理解MUTEX问题的定义。MUTEX其实不是一个确定的问题，而是一个问题族。一般它的safety部分没什么变化，而liveness部分有种类繁多的变体。

要真正掌握MUTEX问题的定义，你得能够看懂别人精确定义的问题，并且能根据自己实际面对的问题，给出最合适的系统建模和问题定义。

从宏观的视角来看，互斥算法可以看成是一个频谱：SHM提供的存储抽象越强，那MUTEX就越容易解决，反之就越难。

✍ 问题的定义及下面的算法，主要参考[AW04, Chap 4]。 □

13.2　基于强原语的MUTEX算法

[AW04, Chap 4]中先讲了几个用强大原语 - 包括test&set，read-modify-write等 - 的MUTEX算法。

作为一个实际系统中的例子，6.824（Spring 2021）的L4 - Primary-backup Replication里有一个用powerful原语的例子。当primary和backup断连时，它们都可以访问一个storage sverver，它们用test&set原语来更新storage server上的一个标志位来实现容错场景下最核心的协同。

13.3　Bakery算法：从“伪MUTEX算法”到“真MUTEX算法”

MUTEX算法中，比较经典的是Lamport提出的Bakery算法。可以分几个步骤来逐步理解这个算法。

- 协同的载体：首先理解进程p_i拥有（own）一个register的概念，或者说single-writer multi-reader register的妙用。每个proc拥有一个register，即只有我能写它，其他人都可以读。这样每个proc拥有一个register，大家就可以互相通过这些register比较自如地完成交流、协同。

- 取号机制的SHM实现：基于上述概念，就让每个proc类似银行排队取号一样，等待进入临界区。只不过这里的取号是通过每个进程 p_i 都拥有的register来实现的。每个proc看看别人的号，就知道自己取号要取所有号里最大号的下一个号。

- 承诺我主意不改了：上述取号机制的基本原理是对的，但是有比较直觉的漏洞。根据adversary argument的视角（adversary argument的基本概念可以参考我本科生算法课的L7），别人一读到我own的register，我就恶意地去改它，这是可以让上述机制出错的。为此我们课上直接构造了一个这样的例子。所以需要进一步引入一个类似承诺的概念，就是“买定离手”，我不改了。这个机制是通过让每个proc拥有一个flag register $Choosing_i$ 来实现的。

Bakery算法的基本设计讲完了，但是故事才刚刚开始。上面所有的MUTEX算在Lamport看来，都是伪互斥算法，因为底层SHM某种意义上已经提供互斥了，所以你的互斥都是伪互斥，是“剽窃”别人的。要真的自己实现互斥，你连atomic register都不能用，因为atomicity显然已经提供了某种互斥。而他所提出的Bakery算法是真互斥算法，也就是说Bakery算法是不依赖底层register的能力的，它基于最弱的Safe register就能实现。（注意，课题上为了入门学习的方便，我们是简化了模型，讲了Bakery算法的退化版本，即假设SHM提供了atomic register，而基于atomic reigster我们可以实现Bakery算法。）

大家比较熟悉各种register的语义之后，就可以深入辨析Bakery算法的正确性证明，来体会上面的结论。从这一角度来讲，Bakery算法是最牛的MUTEX算法。

13.4　Bakery算法的后续改进

Bakery算法在一个维度有明显的问题，即它要求register能存任意大的值。利用adversary argument的观点很好理解，大家疯狂取号，那这个号就会一直增下去，没有机会归零，总有一个时刻，register存不下应该表示的号了。

为此[Attiya04, Chap 4]讲了一套改进的办法，它是一系列算法，跳板式逐步改进。这一跳板式的改进思路，比算法的细节更重要：

- 2个人，不对等：只考虑两个proc，而且有一个人优先级高，同等条件下，总能抢占另一个人。基于这一条件，比较容易实现MUTEX。

- 2个人，对等：有了上面的跳板，让优先级高的人不固定，也就是每个时刻都是不对等的，但是两个人轮流做那个优先级高的人。

- n个人，对等：有了上面的跳板，就让大家锦标赛，两两比赛（比谁能顺利进入临界区），比出的冠军，直接进入临界区。

第四部分

分布式算法进阶选题

这部分讲解重要的分布式算法分析技术，以及分布式算法的正确性验证。

第十四章　不可能性结果

14.1　概述

当你熟悉分布式算法、分布式系统的基本概念，对它们有一点宏观性的思考的时候，一个很重要的概念就是分布式算法、分布式系统中的基本的不可能性结果。

✍ 早在89年的PODC上，Lynch就survey了“A hundred impossibility proofs for distributed computing”[Lyn89]，到了03年，Faith Ellen（与人合作）又写了“Hundreds of impossibility results for distributed computing”[FR03]。最终到了14年，Hagit Attiya和Faith Ellen汇集成书“Impossibility Results for Distributed Computing”[AE14] [1]。 □

14.2　FLP

不可能性结果的系统讨论超出introduction级别的教科书的范畴，有需要直接去看[AE14]即可。但是有两组结果有一些特别的原因，专门提一下。

一个是异步容错共识的不可能性结果FLP。由于共识算法的大热，它也老被人提起。它的原始证明见[FLP85]，直接看课本的表述更容易一些[AE14]。

看这类证明，必须对分布式系统的执行有一个“超然世外的上帝视角”。分布式系统的执行就像是电影拍摄的片段、素材，你就像剪辑导演，随便复制、改写、拼接，直到得到满足你要求的执行。

14.3　拜占庭共识：“叛徒”比例不能达到$\frac{1}{3}$

同步模型下解决拜占庭共识问题的一个基础结论是：当$3f > n$时，是不可能解决拜占庭共识问题的。其证明的关键在于，采用adversary argument的视角，可以让拜占庭错误的proc做你希望的任意动作，这样你就可以构造两个不可区分的执行。这两个执行的不可区分性与它们分别达成不同的共识，又是矛盾的。

具体的构造是巧妙的，需要结合[AW04, Sec 5.2.3]的图示来理解。证明是先讨论n=3的特殊情况，然后再据此推出任意n情况的证明。

[1] 上述理论结果背后一个有趣的现象是，上面所提到的作者都是女性，分布式算法领域两本教科书的作者也是女性，她们的女性学生后续也在分布式理论领域做出了出色的结果。

14.4 Chain Argument

这是不可能性结果证明中的一个简答而有效的技术，主要用于构造分布式系统执行之间的不可区分性（indistinguishability）。因为我们下面的理论证明用到了它，所以专门提一下。

这个技术可以让你很形象地理解什么叫做，以一个剪辑导演的视角，拼接分布式系统的执行片段。我们的PODC2020工作的会议报告中，有一个形象的简单例子 [2]。

具体内容可以看[AE14, Chap 2]。我自己由于下面研究的关系，实际看的是一份具体的研究工作[HNS16]（这是一篇PODC16的短文，具体细节可以看它的arxiv长文版本）。

14.5 Fast Access to Atomic Registers

实现atomic register经常需要要两轮（roundtrip）的通信，由此一轮通信的算法就被称为是fast的。

直觉上一轮通信是不可能实现atomic register抽象的，但是严格证明它却不那么容易。还是分single-writer和multi-writer的情况来逐个击破。

[DGLV10]证明了single-writer情况的fast实现不可能。

Multi-writer情况的不可能性结果，期间有一些受限情况下的证明。我们自己的一份工作最终解决了这一问题[HHW20]。

此外，既然fast是不可能严格实现atomic register的，那我们的另一系列的工作就讨论：假设fast是必须的，atomicity会被牺牲到什么程度。针对这一问题，我们提出了ASC（Almost Strong Consistency）的概念，并做了一些初步的理论分析与实验研究[WHL17, OHWL21]。

[2] https://www.bilibili.com/video/BV1K54y1D766/

第十五章　Failure Detector的概念与应用

Failure Detector是分布式系统建模与分布式算法设计的重要概念 [CT96]。

Weakest failure detector的这一份工作名气挺大，内容很深刻 [CHT96]。

✍ 我们在自己的工作[HHW20]中的一部分证明，其实就是这份工作证明的一个大幅简化版。我们做完[HHW20]我才真正看懂这份工作[CHT96]，并且收获很大，知道自己的证明其实可以仿照它，写得更严谨，更易懂。 □

第十六章　分布式算法正确性验证

16.1　SandTable

SandTable。

✍ 这是我们自己的一份验证方面的工作 [TSH+24]。 □

16.2　Remix

Remix。

✍ 这是我们对Zab协议和ZooKeeper系统实现进行验证的一份工作 [OST+25]。 □

16.3　TLA+

✍ TLA+是Lamport提出的一种形式化规约语言，主要针对并发与分布式系统设计。可以直接在Lamport维护的TLA+主页上了解它的基本知识。

或许是因为TLA+和Paxos都是Lamport发明的，TLA+在分布共识算法的规约与验证方面，取得了广泛的应用。通过开源社区大家可以获得丰富的TLA+规约与验证的实际案例。 □

附录 A　讲义的修订历史

- ✍ 2022/08/05：回到latex讲义环境，写了一小部分素材，发现写讲义和写知乎版介绍的概念还是有质的不同，甚至是冲突。回归到传统的讲义写作上下文，继续推进写作。
- ✍ 2022/05/29：知乎版讲义的初步效果还不错，有效推进了讲义的写作。随着基本框架的成型，后续的正式写作，转回到latex+github的写作、发布模式。这算是这本讲义的V0.2版。
- ✍ 2022/05/03：换一个方式，在知乎专栏上写了一版轻量级的讲义。这是当时的说明：“以前试图写一个重量级的授课讲义，推进非常缓慢。现在尝试写一个轻量级的讲义，以主要文献的推荐和简评为主，推进一段时间，试试效果。”。
- ✍ 2021/11/24：2021年陆续写了一个讲义的草稿，基本就是文献的列举，没有实质性的评述与讨论。这算是这本讲义的V0.1版。

参考文献

[ABND95] Hagit Attiya, Amotz Bar-Noy, and Danny Dolev. Sharing memory robustly in message-passing systems. *J. ACM*, 42(1):124–142, January 1995.

[Ady99] Atul Adya. *Weak Consistency: A Generalized Theory and Optimistic Implementations for Distributed Transactions*. PhD thesis, Cambridge, MA, USA, 1999. AAI0800775.

[AE14] Hagit Attiya and Faith Ellen. *Impossibility Results for Distributed Computing*. Morgan & Claypool, 2014.

[AS85] Bowen Alpern and Fred B. Schneider. Defining liveness. *Information Processing Letters*, 21(4):181–185, 1985.

[Asp19] James Aspnes. *Notes on Theory of Distributed Systems*. Yale University, CPSC 465/565, 2019.

[AW04] Hagit Attiya and Jennifer Welch. *Distributed Computing: Fundamentals, Simulations and Advanced Topics*. John Wiley & Sons, 2004.

[BM93] Özalp Babaoğlu and Keith Marzullo. *Consistent Global States of Distributed Systems: Fundamental Concepts and Mechanisms*, pages 55–96. ACM Press/Addison-Wesley Publishing Co., USA, 1993.

[BR95] Özalp Babaoğlu and Michel Raynal. Specification and verification of dynamic properties in distributed computations. *J. Parallel Distrib. Comput.*, 28(2):173–185, aug 1995.

[CBDGF95] Bernadette Charron-Bost, Carole Delporte-Gallet, and Hugues Fauconnier. Local and temporal predicates in distributed systems. *ACM Trans. Program. Lang. Syst.*, 17(1):157–179, jan 1995.

[CBPS10] Bernadette Charron-Bost, Fernando Pedone, and André Schiper, editors. *Replication: Theory and Practice*. Springer-Verlag, Berlin, Heidelberg, 2010.

[CGR07] Tushar D. Chandra, Robert Griesemer, and Joshua Redstone. Paxos made live: An engineering perspective. In *Proceedings of the Twenty-sixth Annual ACM Symposium on Principles of Distributed Computing*, PODC '07, pages 398–407. ACM, 2007.

[CGR11] Christian Cachin, Rachid Guerraoui, and Lus Rodrigues. *Introduction to Reliable and Secure Distributed Programming.* Springer Publishing Company, Incorporated, 2nd edition, 2011.

[CHT96] Tushar Deepak Chandra, Vassos Hadzilacos, and Sam Toueg. The weakest failure detector for solving consensus. *J. ACM*, 43(4):685–722, July 1996.

[CL85] K. Mani Chandy and Leslie Lamport. Distributed snapshots: Determining global states of distributed systems. *ACM Trans. Comput. Syst.*, 3(1):63–75, feb 1985.

[CM91] Robert Cooper and Keith Marzullo. Consistent detection of global predicates. In *Proceedings of the 1991 ACM/ONR Workshop on Parallel and Distributed Debugging*, PADD '91, pages 167–174, New York, NY, USA, 1991. Association for Computing Machinery.

[CT96] Tushar Deepak Chandra and Sam Toueg. Unreliable failure detectors for reliable distributed systems. *J. ACM*, 43(2):225–267, March 1996.

[DGLV10] Partha Dutta, Rachid Guerraoui, Ron R. Levy, and Marko Vukolić. Fast access to distributed atomic memory. *SIAM J. Comput.*, 39(8):3752–3783, December 2010.

[FLP85] Michael J. Fischer, Nancy A. Lynch, and Michael S. Paterson. Impossibility of distributed consensus with one faulty process. *J. ACM*, 32(2):374–382, April 1985.

[FR03] Faith Fich and Eric Ruppert. Hundreds of impossibility results for distributed computing. *Distributed Computing*, 16(2):121–163, Sep 2003.

[HHW20] Kaile Huang, Yu Huang, and Hengfeng Wei. Fine-grained analysis on fast implementations of distributed multi-writer atomic registers. In *Proceedings of the 39th Symposium on Principles of Distributed Computing*, PODC'20, page 200–209, New York, NY, USA, 2020. Association for Computing Machinery.

[HKJR10] Patrick Hunt, Mahadev Konar, Flavio P. Junqueira, and Benjamin Reed. ZooKeeper: wait-free coordination for internet-scale systems. In *Proc. ATC'10, USENIX Annual Technical Conference*, pages 145–158. USENIX, 2010.

[HNS16] Theophanis Hadjistasi, Nicolas Nicolaou, and Alexander A. Schwarzmann. Brief announcement: Oh-ram! one and a half round read/write atomic memory. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*, PODC '16, pages 353–355, New York, NY, USA, 2016. Association for Computing Machinery.

[JRS11] Flavio P. Junqueira, Benjamin C. Reed, and Marco Serafini. Zab: high-performance broadcast for primary-backup systems. In *Proc. DSN'11, IEEE/IFIP Conference on Dependable Systems and Networks*, pages 245–256. IEEE, 2011.

[Lam78] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7):558–565, 1978.

[Lam86a] Leslie Lamport. On interprocess communication. part i: Basic formalism. *Distributed Computing*, 1(2):77–85, 1986.

[Lam86b] Leslie Lamport. On interprocess communication. part ii: Algorithms. *Distributed Computing*, 1(2):86–101, 1986.

[Lam01] Leslie Lamport. Paxos made simple. *ACM SIGACT News (Distributed Computing Column) 32, 4 (Whole Number 121, December 2001)*, pages 51–58, December 2001.

[LH89] Kai Li and Paul Hudak. Memory coherence in shared virtual memory systems. *ACM Trans. Comput. Syst.*, 7(4):321–359, November 1989.

[LS97] N. A. Lynch and A. A. Shvartsman. Robust emulation of shared memory using dynamic quorum-acknowledged broadcasts. In *Proceedings of IEEE 27th International Symposium on Fault Tolerant Computing*, pages 272–281, June 1997.

[Lyn89] N. Lynch. A hundred impossibility proofs for distributed computing. In *Proceedings of the Eighth Annual ACM Symposium on Principles of Distributed Computing*, PODC '89, pages 1–28, New York, NY, USA, 1989. Association for Computing Machinery.

[Lyn96] Nancy A. Lynch. *Distributed Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996.

[Mal19] Dahlia Malkhi, editor. *Concurrency: The Works of Leslie Lamport*. Association for Computing Machinery, New York, NY, USA, 2019.

[Mat89] Friedemann Mattern. Virtual time and global states of distributed systems. In *Proc. International Workshop on Parallel and Distributed Algorithms*, pages 215–226, Holland, 1989.

[Med12] Andre Medeiros. Zookeeper' s atomic broadcast protocol: Theory and practice. http://www.tcs.hut.fi/Studies/T-79.5001/reports/2012-deSouzaMedeiros.pdf, 2012. Accessed: 08-19-2022.

[MQL+24] Haoran Ma, Yifan Qiao, Shi Liu, Shan Yu, Yuanjiang Ni, Qingda Lu, Jiesheng Wu, Yiying Zhang, Miryung Kim, and Harry Xu. DRust: Language-Guided distributed shared memory with fine granularity, full transparency, and ultra efficiency. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 97–115, Santa Clara, CA, July 2024. USENIX Association.

[OHWL21] Lingzhi Ouyang, Yu Huang, Hengfeng Wei, and Jian Lu. Achieving probabilistic atomicity with well-bounded staleness and low read latency in distributed datastores. *IEEE Trans. Parallel Distrib. Syst.*, 32(4):815–829, apr 2021.

[OO14] Diego Ongaro and John Ousterhout. In search of an understandable consensus algorithm. In *Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference*, USENIX ATC'14, pages 305–320, Berkeley, CA, USA, 2014. USENIX Association.

[OST+25] Lingzhi Ouyang, Xudong Sun, Ruize Tang, Yu Huang, Madhav Jivrajani, Xiaoxing Ma, and Tianyin Xu. Multi-grained specifications for distributed system model checking and verification. In *Proceedings of the Tenth European Conference on Computer Systems*, EuroSys'25, New York, NY, USA, 2025. Association for Computing Machinery.

[PTZ+21] Haochen Pan, Jesse Tuglu, Neo Zhou, Tianshu Wang, Yicheng Shen, Xiong Zheng, Joseph Tassarotti, Lewis Tseng, and Roberto Palmieri. Rabia: Simplifying state-machine replication through randomization. SOSP'21, pages 472–487, New York, NY, USA, 2021. Association for Computing Machinery.

[SKS20] Avi Silberschatz, Henry F. Korth, and S. Sudarshan. *Database System Concepts, Seventh Edition*. McGraw-Hill Book Company, 2020.

[SN04] Robert C Steinke and Gary J Nutt. A unified theory of shared memory consistency. *Journal of the ACM (JACM)*, 51(5):800–849, 2004.

[SWPL11] Cheng Shao, Jennifer L. Welch, Evelyn Pierce, and Hyunyoung Lee. Multiwriter consistency conditions for shared memory registers. *SIAM J. Comput.*, 40(1):28–62, January 2011.

[TSH+24] Ruize Tang, Xudong Sun, Yu Huang, Yuyang Wei, Lingzhi Ouyang, and Xiaoxing Ma. Sandtable: Scalable distributed system model checking with specification-level state exploration. In *Proceedings of the Nineteenth European Conference on Computer Systems*, EuroSys'24, pages 736–753, New York, NY, USA, 2024. Association for Computing Machinery.

[WHL17] Hengfeng Wei, Yu Huang, and Jian Lu. Probabilistically-atomic 2-atomicity: Enabling almost strong consistency in distributed storage systems. *IEEE Trans. Comput.*, 66(3):502–514, March 2017.

[YZZ+24] Liangcheng Yu, Xiao Zhang, Haoran Zhang, John Sonchack, Dan Ports, and Vincent Liu. Beaver: Practical partial snapshots for distributed cloud services. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 233–249, Santa Clara, CA, July 2024. USENIX Association.