特别放送 | 那些你不能错过的分布式系统论文

2019-11-15 刘梦馨 来自北京

《分布式技术原理与算法解析》

> 你好，我是聂鹏程。

> 古人云“以史为鉴，可以知兴替。”说的就是追本溯源的力量。通过学习和思考技术的发展和演进，我们方能更好地把握未来。而对分布式技术追本溯源的方式，无疑就是精读相关经典轮文了。

> 为此，今天我特地邀请了我的朋友刘梦馨，来与你系统分享下分布式系统领域的经典论文。你有时间和耐力的话，可以逐一阅读、学习下这些论文。

> 刘梦馨是灵雀云容器平台高级研发工程师，负责容器平台的架构、容器网络方案的设计和实现，也是开源 Kubernetes 网络插件 Kube-OVN 作者。他平时非常喜欢阅读论文，也总结了很多高效阅读论文的方法。

> 话不多说，我们来看刘梦馨的分享吧。

你好，我是刘梦馨。

分布式系统领域有着最令人费解的理论，全链路的不确定性堪比物理中的量子力学。同时，分布式系统领域又有着当代最宏伟的计算机系统，Google、Facebook、亚马逊遍布全球的系统支撑着我们的信息生活。

显然，能够征服分布式系统的，都是理论和实践两手抓两手都要硬的强者。然而，分布式系统领域还有着最高的上手门槛，没有大规模的基础设施、没有潮水般的流量，分布式领域幽灵般的问题并不会浮出水面。

那么，我们应该**如何开启征服分布式系统的征程**呢？

好在这条路上我们并不孤独。学术大牛们在五十年前就开始探索各方面理论上的问题，全球规模的互联网公司也有着丰富的实践和经验。而这些分布式领域人类的智慧，最终都沉淀为了一篇篇的经典论文。

和普通的技术文章相比，论文的发表有着极为严格的要求，随之而来的也是极高的质量。通过阅读分布式领域的经典问题，我们可以快速吸收前人的智慧，领略大型系统的风采，并收获最为宝贵的实战经验。

现在，就让我们从一篇篇经典论文开始，踏上征战分布式系统的征程吧！

我**按照从理论到实践的顺序**，将经典的分布式系统论文分成了分布式理论基础、分布式一致性算法、分布式数据结构和分布式系统实战四类，帮助你快速找到自己需要的论文。

这些论文我都给到了标题，你可以直接去 Google 学术里搜索。

分布式理论基础

分布式理论基础部分的论文，主要从宏观的角度介绍分布式系统中最为基本的问题，从理论上证明分布式系统的不确定、不完美，以及相互间的制约条件。研读这部分论文，你可以了解经典的 CAP 定理、BASE 理论、拜占庭将军问题的由来及其底层原理。

有了这些理论基础，你就可以明白分布式系统复杂的根源。当再碰到一些疑难杂症，其他人不得其解时，你可以从理论高度上指明方向。

以下就是分布式理论基础部分的论文：

- Time, Clocks, and the Ordering of Events in a Distributed System
- The Byzantine Generals Problem
- Brewer’s Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services
- CAP Twelve Years Later: How the “Rules” Have Changed
- BASE: An Acid Alternative
- A Simple Totally Ordered Broadcast Protocol
- Virtual Time and Global States of Distributed Systems

分布式一致性算法

只要脱离了单机系统，就会存在多机之间不一致的问题。因此，分布式一致性算法，就成了分布式系统的基石。

在分布式一致性算法这一部分，我将与你推荐 2PC、Paxos、Raft 和 ZAB 等最知名的一致性算法。分布式算法的复杂度比普通算法要高出几个数量级，所以这部分论文是最为烧脑的一部分。

搞明白这部分论文，你的空间想象力和统筹规划能力都会得到质的提升。

- A Brief History of Consensus, 2PC and Transaction Commit
- Paxos Made Simple
- Paxos Made Practical
- Paxos Made Live: An Engineering Perspective

Raft: In Search of an Understandable Consensus Algorithm

ZooKeeper: Wait-Free Coordination for Internet-Scale Systems

Using Paxos to Build a Scalable, Consistent, and Highly Available Datastore

Impossibility of Distributed Consensus With One Faulty Process

Consensus in the Presence of Partial Synchrony

分布式数据结构

分布式数据结构部分的论文，将与你介绍管理分布式存储问题的知名数据结构原理。通过它们，你可以构建自己的分布式系统应用。

这部分论文的涵盖范围大致包括两部分：一是，分布式哈希的四个著名算法 Chord、Pastry、CAN 和 Kademlia；二是，Ceph 中使用的 CRUSH、LSM-Tree 和 Tango 算法。

和分布式一致性算法类似，分布式数据结构也极其考验空间想象力和统筹规划能力。不过，在经过分布式一致性算法的锻炼后，相信这些对你来说已经不再是问题了。

Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications

Pastry: Scalable, Distributed Object Location, and Routing for Large-Scale Peer-to-Peer Systems

Kademlia: A Peer-to-Peer Information System Based on the XOR Metric

A Scalable Content-Addressable Network

Ceph: A Scalable, High-Performance Distributed File System

The Log-Structured-Merge-Tree

HBase: A NoSQL Database

Tango: Distributed Data Structure over a Shared Log

分布式系统实战

分布式系统实战部分的论文，将介绍大量互联网公司在分布式领域的实践、系统的架构，以及经验教训。

Google 的新老三驾马车，Facebook、Twitter、LinkedIn、微软、亚马逊等大公司的知名系统都会在这一部分登场。你将会领会到这些全球最大规模的分布式系统是如何设计、如何实现的，以及它们在工程上又碰到了哪些挑战。

The Google File System

BigTable: A Distributed Storage System for Structured Data

The Chubby Lock Service for Loosely-Coupled Distributed Systems

Finding a Needle in Haystack: Facebook’s Photo Storage

Windows Azure Storage: A Highly Available Cloud Storage Service with Strong Consistency

Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing

Scaling Distributed Machine Learning with the Parameter Server

Dremel: Interactive Analysis of Web-Scale Datasets

Pregel: A System for Large-Scale Graph Processing

Spanner: Google’s Globally-Distributed Database

Dynamo: Amazon’s Highly Available Key-value Store

S4: Distributed Stream Computing Platform

Storm @Twitter

Large-scale Cluster Management at Google with Borg

F1 - The Fault-Tolerant Distributed RDBMS Supporting Google’s Ad Business

Cassandra: A Decentralized Structured Storage System

MegaStore: Providing Scalable, Highly Available Storage for Interactive Services

Dapper, a Large-Scale Distributed Systems Tracing Infrastructure

Kafka: A distributed Messaging System for Log Processing

Amazon Aurora: Design Considerations for High Throughput Cloud-Native Relational Databases

以上就是我为你准备的分布式系统经典论文清单了。这个清单里的每一篇论文，都是经典中的经典。很多论文对之后的工业界及学术界产生了翻天覆地的影响，开创了一个又一个火热的产业。

希望你没有被这个清单吓到，当你翻开这些论文后，就会发现它们的内容并不是高高在上，包含了很多很实际、很具体的问题。认真读下去，你甚至会有掌握了屠龙之技的快感，一发而不可收拾。

为了帮助你高效阅读这些论文，并汲取其中的精华，我再和你说说我阅读论文的一些心法吧。

如何高效地阅读论文？

一般来说，单篇论文大概会有 15 到 20 页的内容，**如果你是第一次读论文可以把重点放在前面的背景介绍、相关工作和概要设计上**。好的论文通常会很仔细地介绍背景知识，帮助你从宏观上先对整个问题有一个初步认识，了解当前现状。

接下来，你可以再**根据自己的兴趣，选择是否仔细阅读论文涉及的详细原理和设计**。这一部分，通常是论文中最精华的部分，包含了最具创新的理念和做法，内容通常也会比较长，需要花费较多的时间和精力去研究。这时，你可以根据自己的情况，选择一批论文重点突破。

论文最后通常是评测和数据展示部分。这部分内容对我们最大的参考价值在于，**学习作者的评测方法、用到的测试工具和测试样例**，以便将其运用到工作中。

阅读完一篇论文后，如果你觉得内容还不错的话，可以通过 Google 学术去搜索相关的文章，找到所有引用这篇论文的新作品。这样一来，你就可以通过一篇经典论文不断深入，全面掌握一个领域。

最后，我希望你可以通过经典论文的助力，迅速建立起自己的知识武器库，来攻克日常工作中的难题。

精选留言 (9)

Jackey
2019-11-15

论文down下来了，上不了Google学术的同学可自取

链接: https://pan.baidu.com/s/1LN9ZaIuSMCRKN_3LAlq4Iw 提取码: 1hp5

编辑回复: 优秀

共 11 条评论 >　　69

sunsun314
2019-11-17

公司组织了分布式理论的业余兴趣小组，之前一直苦于没有体系整理，跟着MIT的课程瞎看，这篇提供了整个体系的入口和方法。
说实话这课程光这篇文章就值了

共 1 条评论 >　　4

Jackey
2019-11-15

哇，确实有点被吓到了。决定还是先从基础的看起来，按每周一篇的进度吧...感谢老师的分享

共 2 条评论 >　　2

极客雷
2020-03-28

教材才会这样：好的论文通常会很仔细地介绍背景知识，帮助你从宏观上先对整个问题有一个初步认识，了解当前现状。
论文一般不会。

 　　1

钱
2020-02-20

优秀，感谢分享！
之前智慧老师说，学习技术最快的三步曲，第一步就是读论文，然后看官方文档，然后看源码。
😂尴尬，英文水平也是拉开距离的关键。

1

Eternal
2019-11-30

超级干货，带上华为云的实践案例，理解更深

PatHoo
2019-11-15

赞！曾经也很喜欢研读论文，对arXiv.org深深着迷。

starnavy
2019-11-15

感谢老师分享。省了好多search cost

随心而至
2019-11-15

好赞，不知道自己能读懂多少，读就完事了。