

NewSQL(Basis)

- NoSql([Bigtable](#) [LSM-Tree](#) [Dynamo](#))
- Paxos(Synchronous Replication across Data Centers)
- RDBMS

Transaction
ACID
2PL
MVCC
2PC

- Secondary Index
- Consistency

current read
snapshot read
inconsistent read
linearizability



Turing award winners and Berkeley alumni (from left to right): Ken Thompson, Butler Lampson, Jim Gray, Nicklaus Wirth



[Unix&C](#) UTF-8 ->Go [Spanner](#)
Alto->Mac 2PC Paxos的流行
ACID Transaction
Pascal 算法+数据结构=程序

Oracle 1979

1973 System R -> DB2 1983
1973 Ingres->Sybase->SQLServer

1960s
GE(IDS)
IMS(IBM)

1970s

网状数据库(Charles Bachman)

关系数据库(Edgar Codd/Jim Gray/[Michael Stonebraker](#) 2014 Turing)

2000s

NoSql([Bigtable](#)/[Dynamo](#))

2010s

NewSql([MegaStore](#)/[Spanner](#)/[F1](#))

数据库阵营

分布式阵营

新硬件
新场景
[BigData](#)



Cockroach DB



数据库先锋([Michael Stonebraker](#))
2007:主要的RDBMS都是过时的技术
2008:MapReduce: 一个巨大的倒退
2011:FaceBook在MySQL上生不如死

[Readings in Database Systems](#)

两手抓：理论+实践 研究+产业
Postgres->Greenplum
StreamBase->Aurora
CStore->Vertica
HStore->VoltDB

NewSql(What's Really New with NewSQL?)

中间件:分库分表/
查询路由/分布式事务

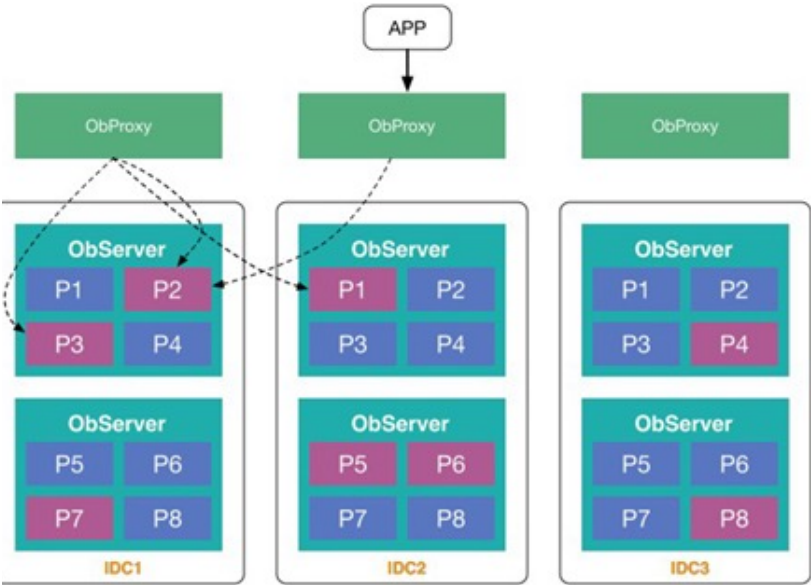
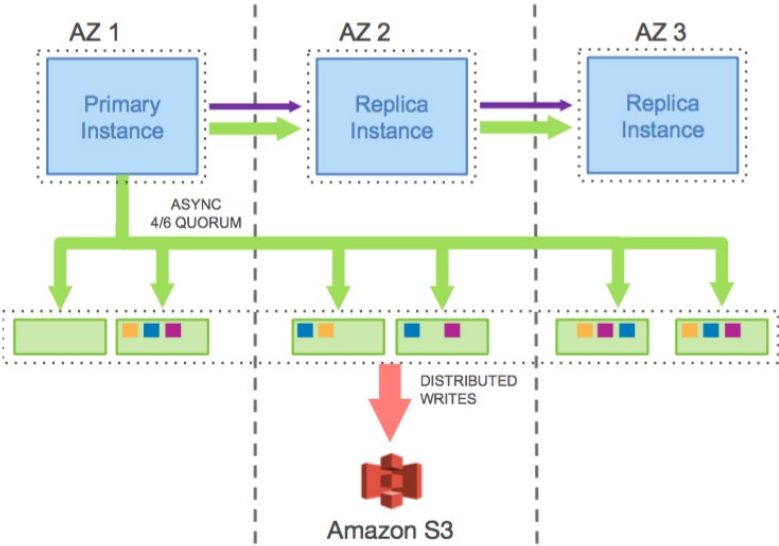
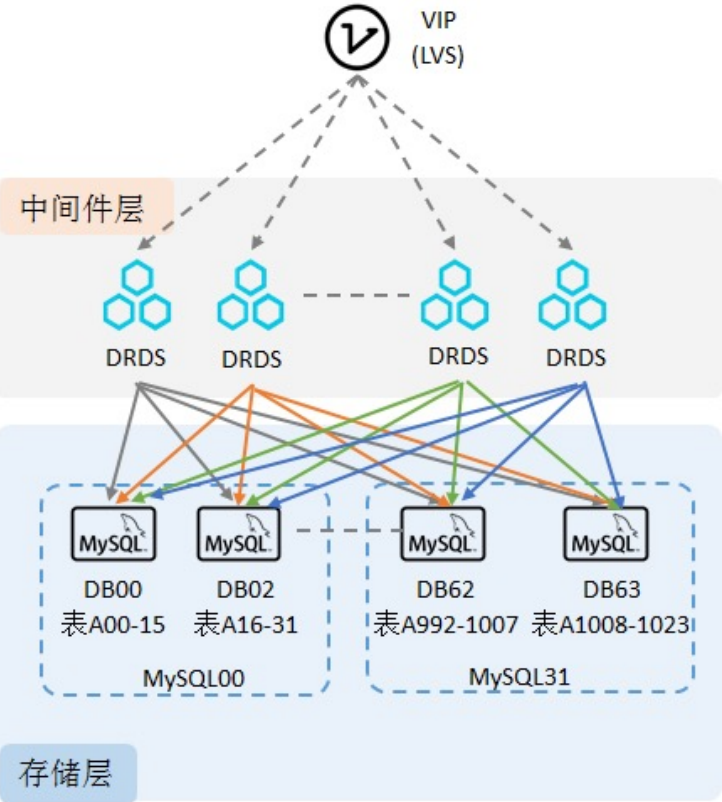
Drds

存储计算分离:
共享存储 一写多读

Amazon Aurora
PolarDB

新架构:ShareNothing/
多节点并发/分布式查询

Spanner
OceanBase



Basis(Bigtable)

- Distributed multi-dimensional sparse map
(row, column, timestamp) → cell contents

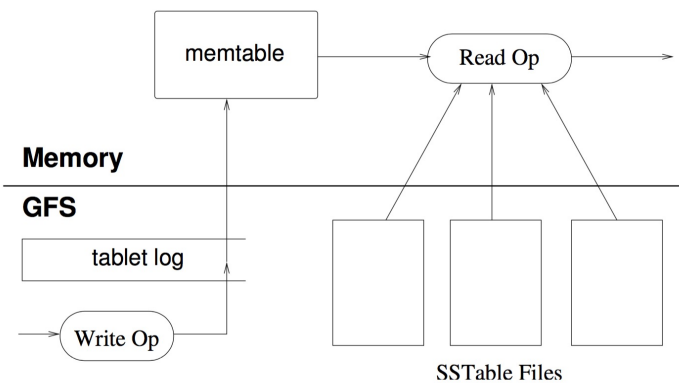
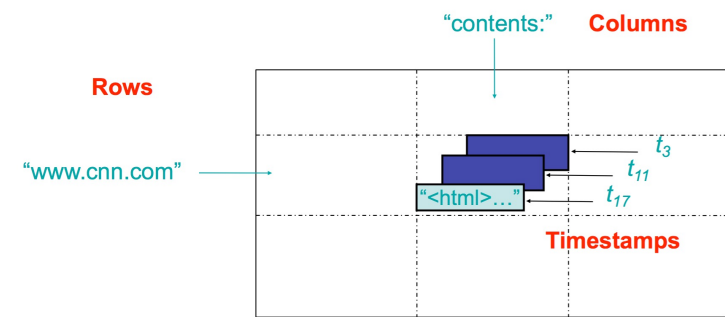
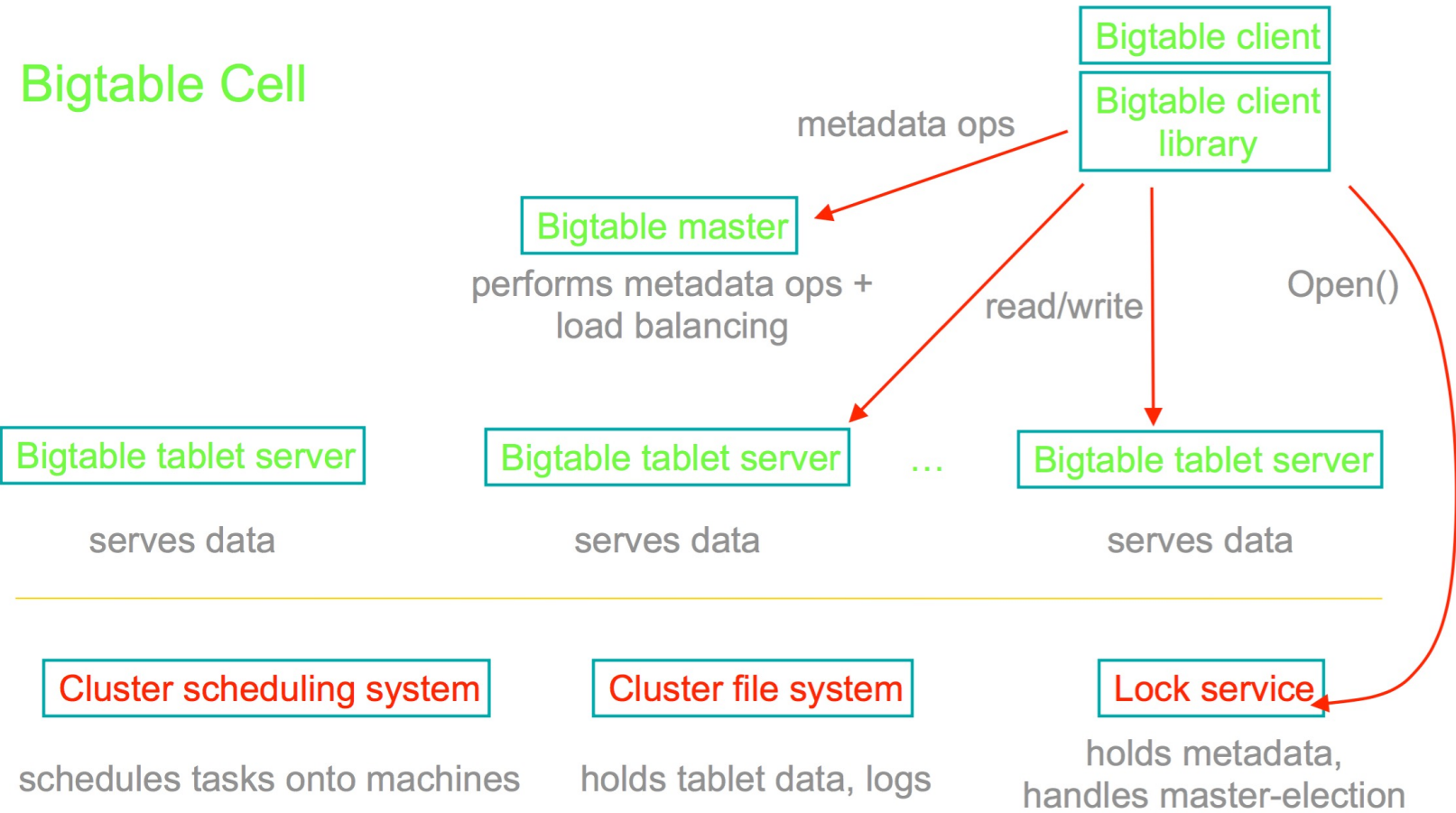


Figure 5: Tablet Representation

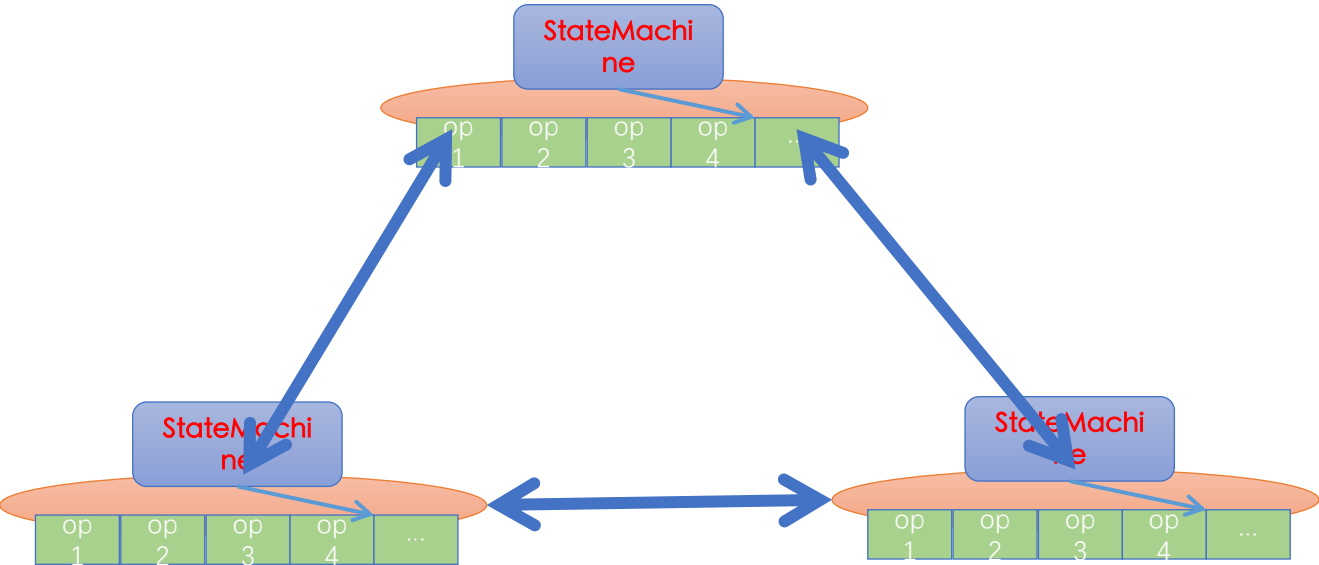
BigTable System Structure

Bigtable Cell



Basis(Paxos)

Replicated State Machines



Basic Paxos(Lampson Lampport)

- Majority
- Multi-Round



1988

Viewstamped Replication: A New Primary Copy Method to Support Highly-Available Distributed Systems (Liskov 2008 Turing)

1989

The Part-Time Parliament (Lampport 2013 Turing)



1996

How to Build a Highly Availability System using Consensus (Butler W.Lampson 1992 Turing)

1998

1999

Revisiting the PAXOS algorithm (Lampson/Lynch)



2001

Paxos Made Simple(Lampport)

2004

Cheap Paxos(Lampport)

2005

Fast Paxos(Lampport)

2006

Chubby

2007

Paxos Made Live

2011

MegaStore

2012

Spanner

2014

Mesa RAFT

2017

PaxosStore

Basis(Transaction—Jim Gray 1976)

- ACID

Atomicity(AllOrNothing): WAL(redo/undo), 2PC

Consistency: 完整性约束(唯一性、外键、Check)

Isolation: 并发事务相互不可见

Durability: log

- Concurrent Control

2PL(1976)

MVCC(1978)

- Distributed Transaction Commit

2PC(1976)

3PC(1981)

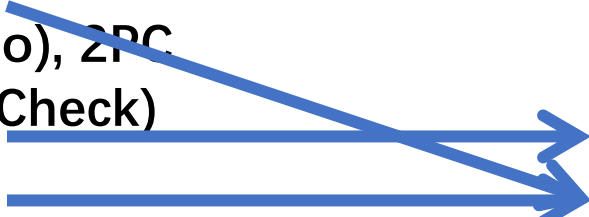
Paxos Commit(2004)

- Time&Clock&HappenBefore

Lamport Clock(1978)

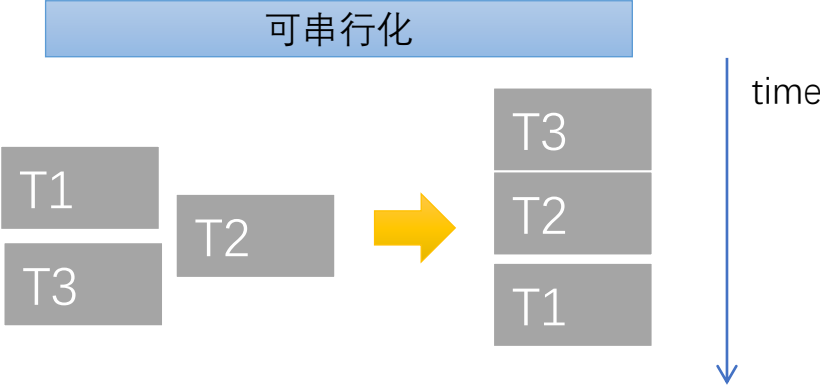
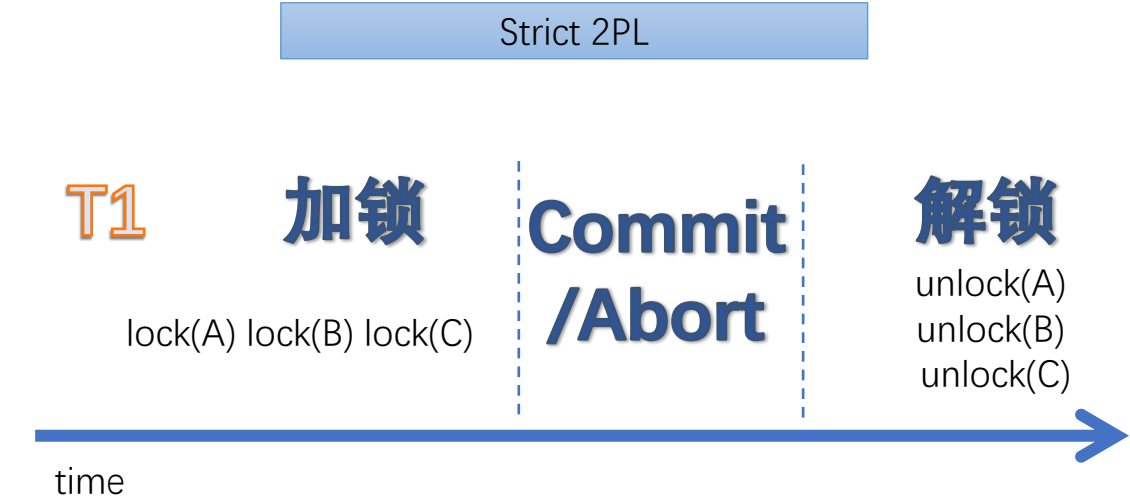
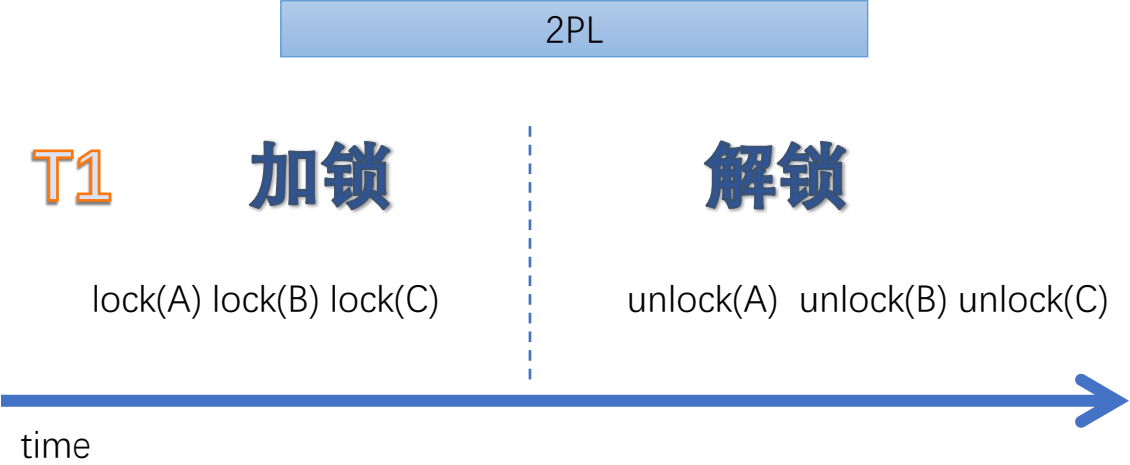
Lease(1989)

Linearizability(1990)

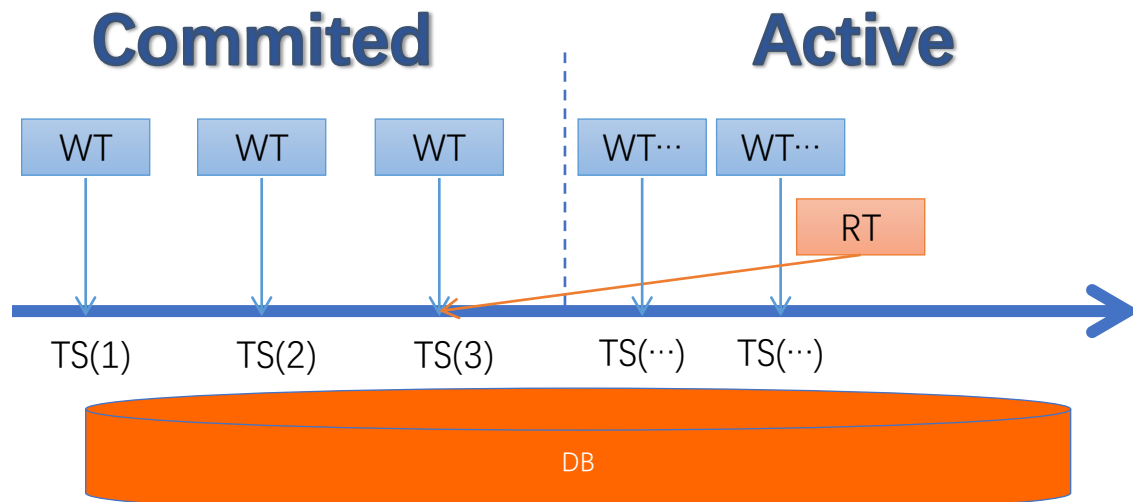


concurrent control
recovery(ARIES)

Basis(2PL-Eswaran 1976)

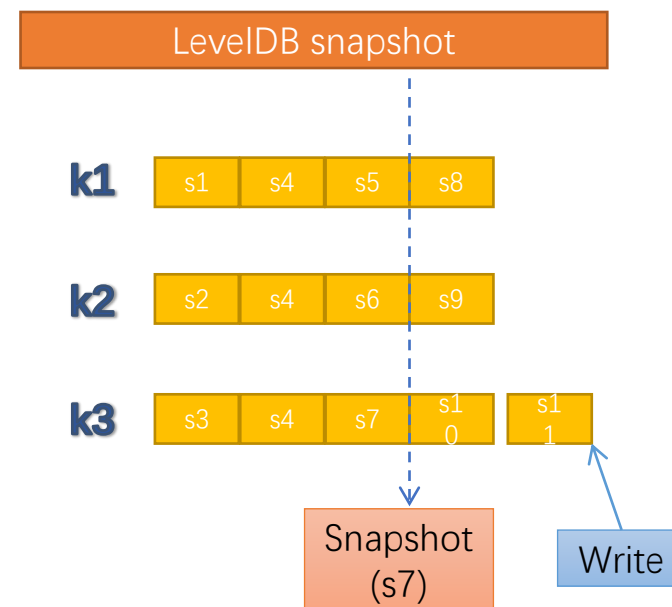


Basis(MVCC - David Patterson 1978-(2017 Turing-RISC/RAID)-Google TPU)



MVCC

- 只读/读写事务处理不冲突，只读事务不会被block
- 非覆盖写 保存多版本数据
- Version单调递增
- 写写冲突，结合2PL解决



Jeff Dean · @JeffDean · 45分
Congrats to John Hennessy (Stanford faculty member and also Alphabet board chair) & Dave Patterson (my Google Brain team colleague and long-time Berkeley faculty member) on receiving this year's A.M. Turing Award!



John Hennessy and David Patterson will receive the 2017 ACM A.M. Turing...
acm.org

1

28

97

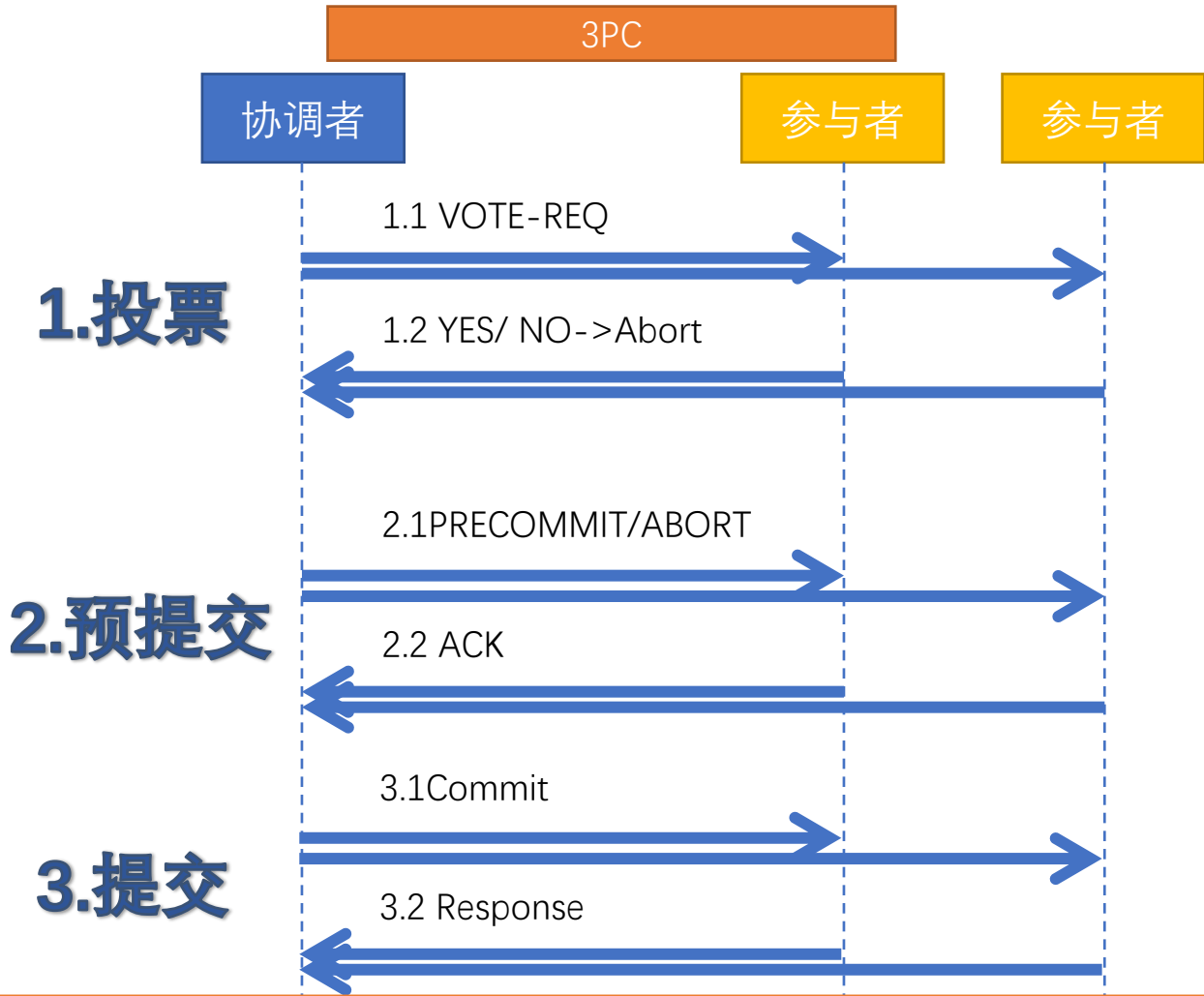


Basis(2PC-Lampson 1976)



Uncertainty period: After vote YES, Before receive info to know decision

Basis(3PC -1981 Dale Skeen VS Paxos Commit-2004 Jim Gray&Lamport)



不变性: 只要活着的进程中有人是uncertainty, 任意进程(包括失败的进程)都不会commit
方法: 增加一个Precommit阶段, 确认所有人走出uncertainty阶段, 然后再commit
问题: 只保证site-fail-stop下的正确性; uncertainty节点与commitable节点网络分区 -> 脑裂->Majority

Paxos vs 2PC

- 事务提交: 1.参与者不对等(拥有不同数据) -> **一票否决权** (完整性/冲突检查) 2.不同事务成员不固定
- Replicated State Mache: 1.多副本对等 高可用 2.成员固定
- Uniform consensus is harder than consensus

