面向分布式系统的复制数据类型理论研究概述

(CCF 2018 第九届优博论坛)

魏恒峰

南京大学软件所

2018 年 08 月 08 日

面向分布式系统的复制数据类型理论研究概述

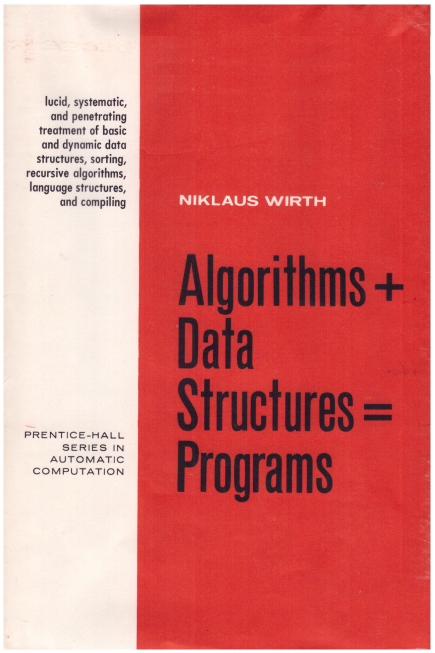
1. 研究背景
2. 两份工作

面向分布式系统的复制数据类型理论研究概述

1. 研究背景
2. 两份工作

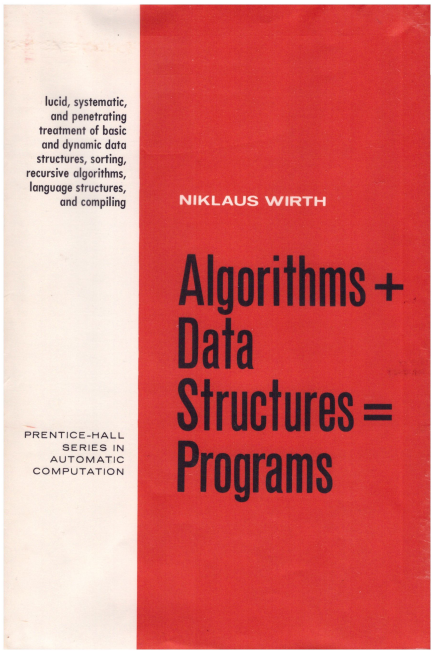
Abstract Data Types (ADT) [Liskov and Zilles, 1974]

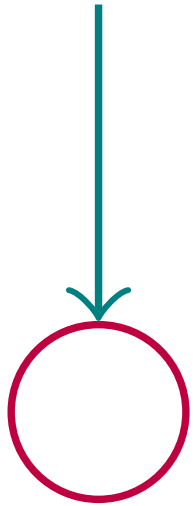
(单线程; 顺序语义)



Abstract Data Types (ADT) [Liskov and Zilles, 1974]

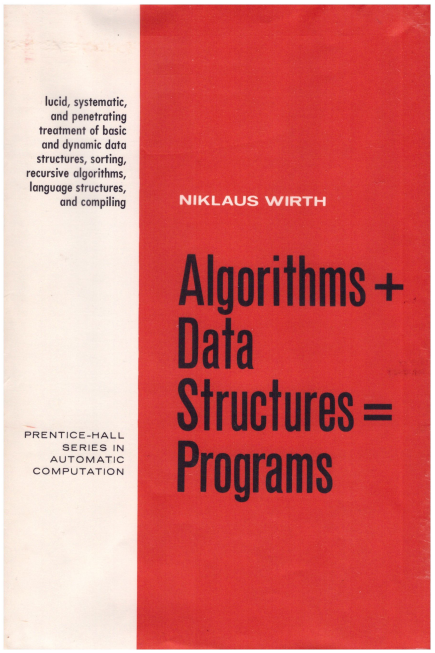
(单线程; 顺序语义)

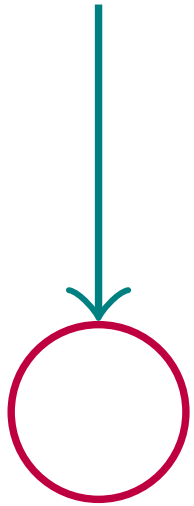




Abstract Data Types (ADT) [Liskov and Zilles, 1974]

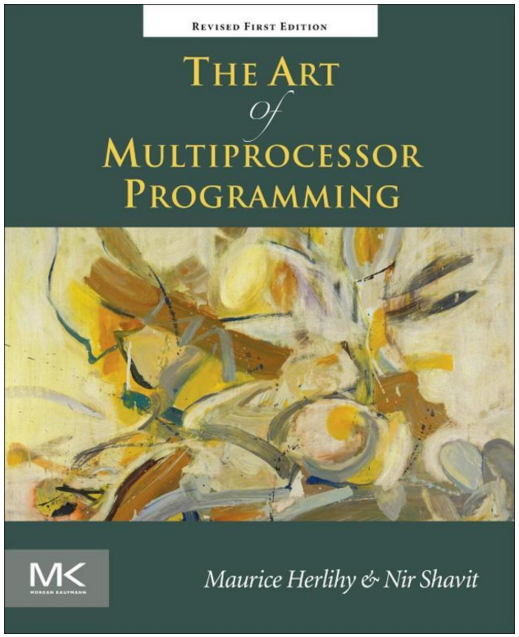
(单线程; 顺序语义)





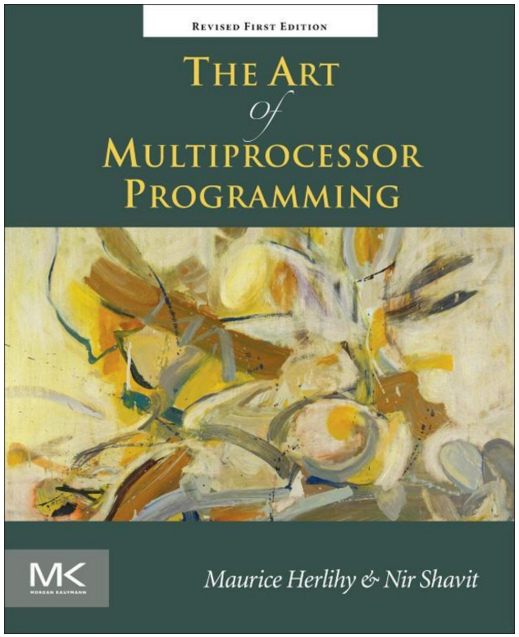
Concurrent Data Types [Herlihy and Wing, 1990]

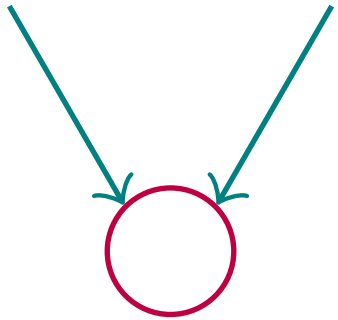
(多线程; 并发语义)



Concurrent Data Types [Herlihy and Wing, 1990]

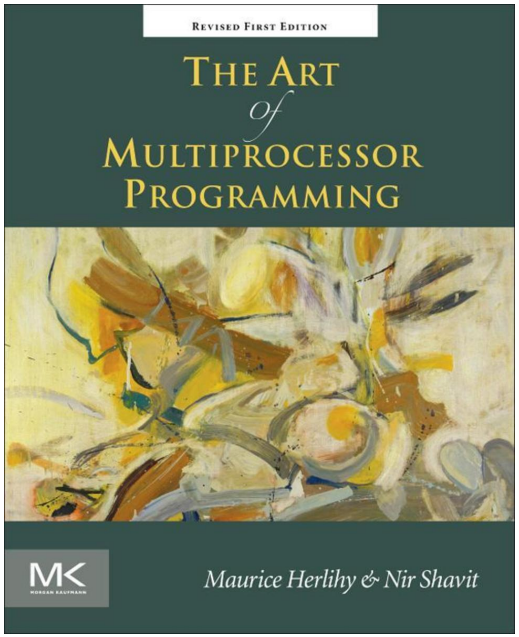
(多线程; 并发语义)

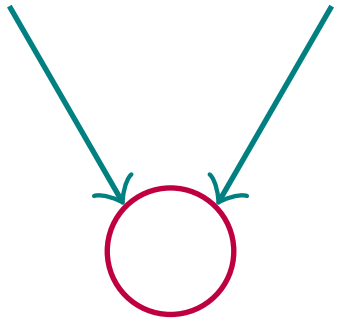




Concurrent Data Types [Herlihy and Wing, 1990]

(多线程; 并发语义)



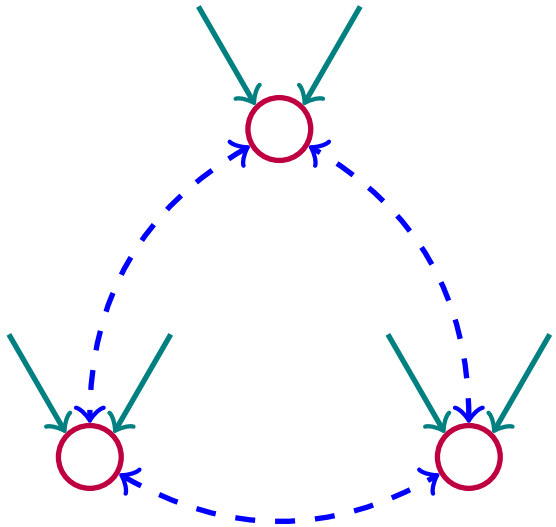


Replicated Data Types (RDT; $\approx$ 2010 年) [Burckhardt et al., 2014]

(多副本; 复制语义)

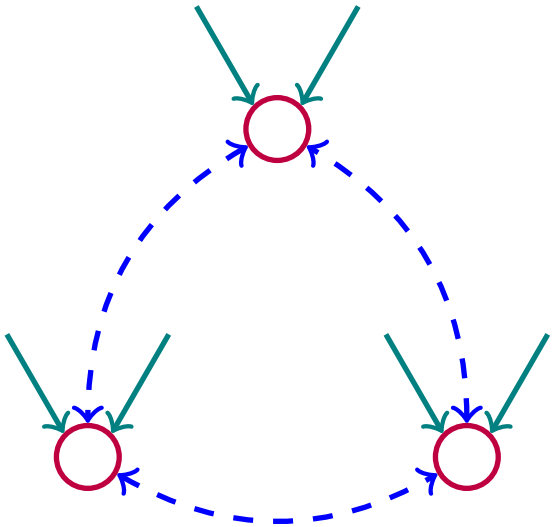
Replicated Data Types (RDT; $\approx$ 2010 年) [Burckhardt et al., 2014]

(多副本; 复制语义)



Replicated Data Types (RDT; $\approx$ 2010 年) [Burckhardt et al., 2014]

(多副本; 复制语义)




KEEP
CALM
AND
Why
Bother

新平台

大规模分布式系统

新浪微博社交应用 [1]:

- 日均用户近一亿名
- 日均消息近一亿条

[1] 2015 第三季度; 数据来自 China Internet Watch.

大规模分布式系统

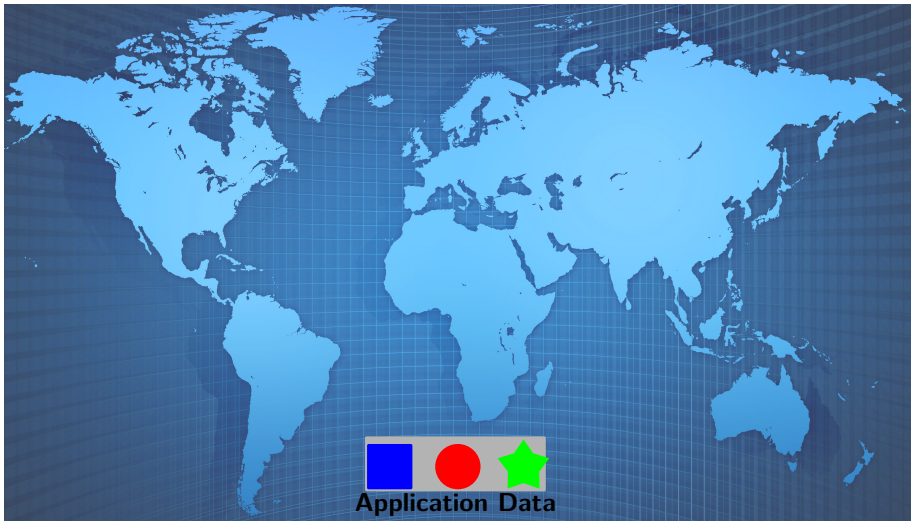
新浪微博社交应用 [1]:

- 日均用户近一亿名
- 日均消息近一亿条

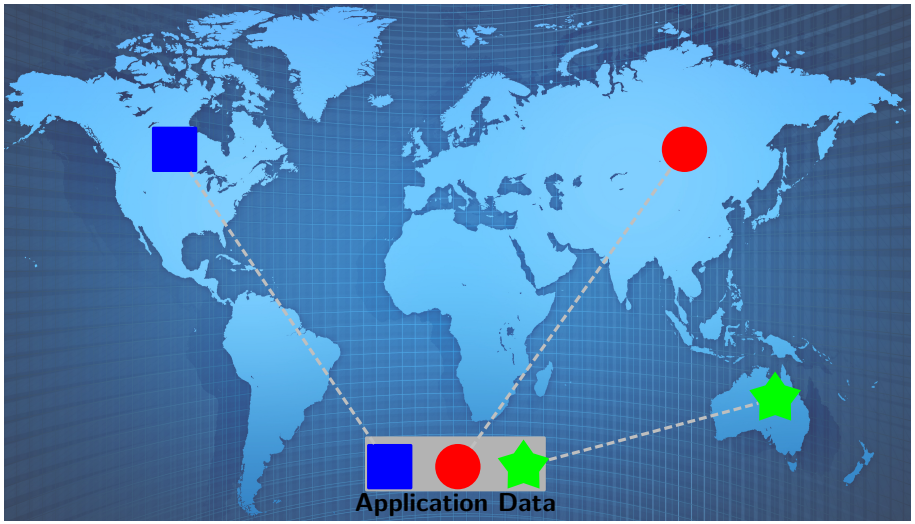
特性需求:

- 低延迟, 高可用性 (4 个 9 [2])
- 高容错性, 高可扩展性

[1] 2015 第三季度; 数据来自 China Internet Watch.
[2] 数据来自 InfoQ.

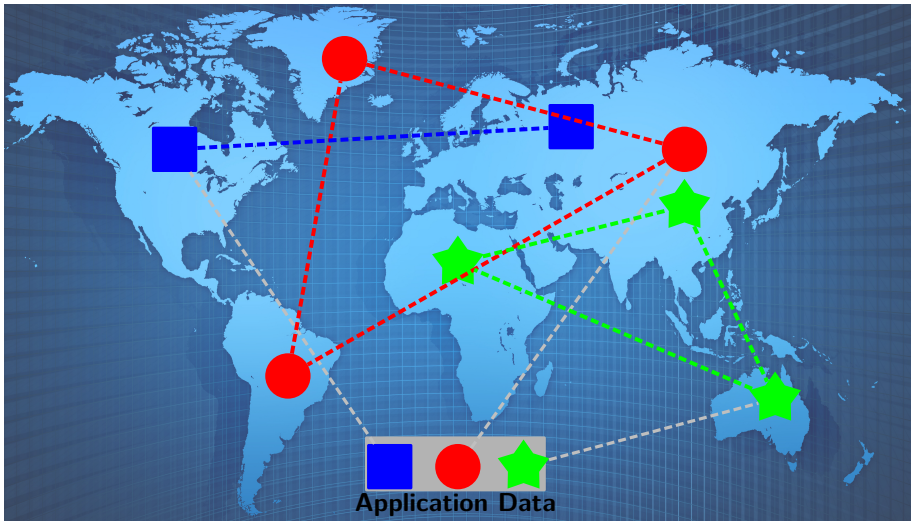


分布数据 (distributed data):



分布数据 (distributed data):

1. 分区 (partition): 水平扩展



分布数据 (distributed data):

1. 分区 (partition): 水平扩展
2. 副本 (replication): 就近访问, 容灾备份

复制数据类型 [Shapiro et al., 2011a]

- Read/Write Register
- Counter
- Set
- List
- HashMap
- Disjoint Set
- Graph
- ...

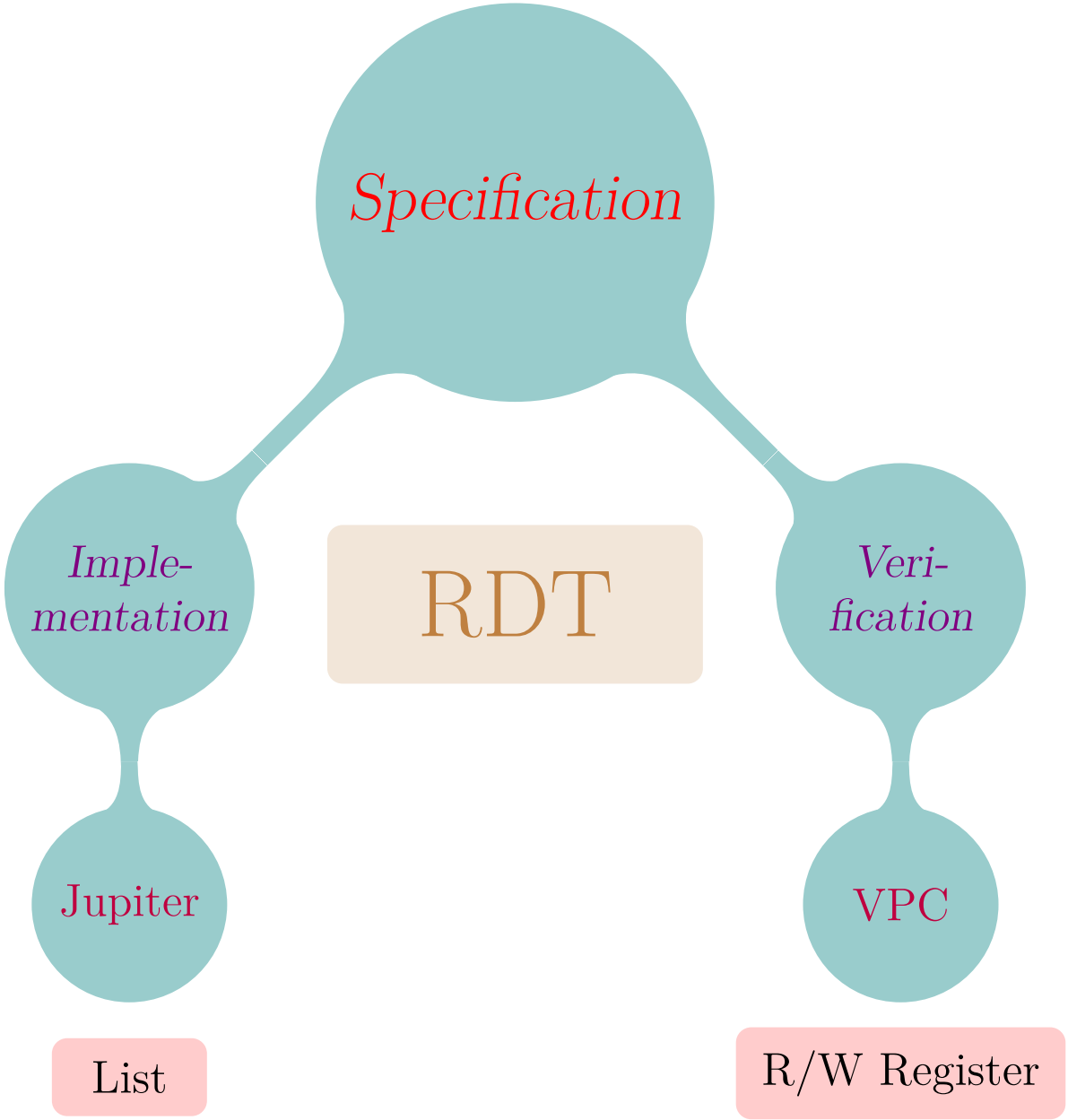
新问题，新挑战

新问题，新挑战

Replicated Data Types: Specification, Verification, Optimality

Sebastian Burckhardt　Alexey Gotsman　Hongseok Yang　Marek Zawirski

[Burckhardt et al., 2014]


Specification
Imple-
mentation
RDT
Veri-
fication
Jupiter
VPC
List
R/W Register

面向分布式系统的复制数据类型理论研究概述

面向分布式系统的复制数据类型理论研究概述

1. 研究背景
2. 两份工作
   - Jupiter

Brief Announcement @ PODC'2018 [3]

实现复制列表的 Jupiter 协议 [Nichols et al., 1995][a] 满足 weak list specification [Attiya et al., 2016][b].

[a] David A. Nichols et al. (1995). "High-latency, Low-bandwidth Windowing in the Jupiter Collaboration System". In: *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*. UIST '95. ACM, pp. 111–120.

[b] Hagit Attiya et al. (2016). "Specification and complexity of collaborative text editing". In: *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*. PODC '16. ACM, pp. 259–268.

[3] 猜想 @ PODC'2016 [Attiya et al., 2016]

Brief Announcement @ PODC'2018 [3]

实现复制列表的 Jupiter 协议 [Nichols et al., 1995][a] 满足 weak list specification [Attiya et al., 2016][b].

[a] David A. Nichols et al. (1995). "High-latency, Low-bandwidth Windowing in the Jupiter Collaboration System". In: *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*. UIST '95. ACM, pp. 111–120.

[b] Hagit Attiya et al. (2016). "Specification and complexity of collaborative text editing". In: *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*. PODC '16. ACM, pp. 259–268.

[3] 藏在脚注里的猜想 @ PODC'2016 [Attiya et al., 2016]

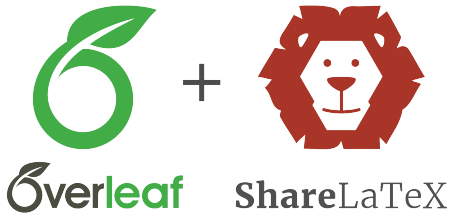
Weak List Specification

基于副本的协同文本编辑系统

(a) Google Docs

(b) Apache Wave

(c) Wikipedia



(d) LaTeX Editor

复制列表对象：建模编辑系统的核心功能

$\textsc{Ins}(a, p)$: 在 p 位置插入元素 a

$\textsc{Del}(p)$: 删除 p 位置上的元素

$\textsc{Read}$: 返回该列表

定义 (最终收敛性 (Eventual Convergence) [Ellis and Gibbs, 1989])

当用户不再提交更新操作时，每个 *replica* 上的列表是相同的。

定义 (最终收敛性 (Eventual Convergence) [Ellis and Gibbs, 1989])

当用户不再提交更新操作时，每个 *replica* 上的列表是相同的。

定义 (强最终一致性 (Strong Eventual Consistency) [Shapiro et al., 2011b])

如果两个 *replica* 处理了同一组用户操作，那么这两个 *replica* 上对列表是相同的。

定义 (最终收敛性 (Eventual Convergence) [Ellis and Gibbs, 1989])

当用户不再提交更新操作时，每个 *replica* 上的列表是相同的。

定义 (强最终一致性 (Strong Eventual Consistency) [Shapiro et al., 2011b])

如果两个 *replica* 处理了同一组用户操作，那么这两个 *replica* 上对列表是相同的。

对系统的中间状态缺少足够的约束

Specification and Complexity of Collaborative Text Editing

Hagit Attiya
Technion

Sebastian Burckhardt
Microsoft Research

Alexey Gotsman
IMDEA Software Institute

Adam Morrison
Technion

Hongseok Yang
University of Oxford

Marek Zawirski*
Inria & Sorbonne Universités,
UPMC Univ Paris 06, LIP6

定义 (Weak List Specification $\mathcal{A}_{weak}$ [Attiya et al., 2016])

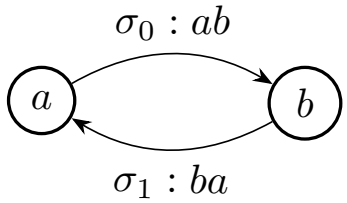
Informally, $\mathcal{A}_{weak}$ requires the ordering between elements that are not deleted to be consistent across the system.

定义在系统所有列表状态上的全局性质

定义 (状态对兼容性 (Pairwise State Compatibility Property))

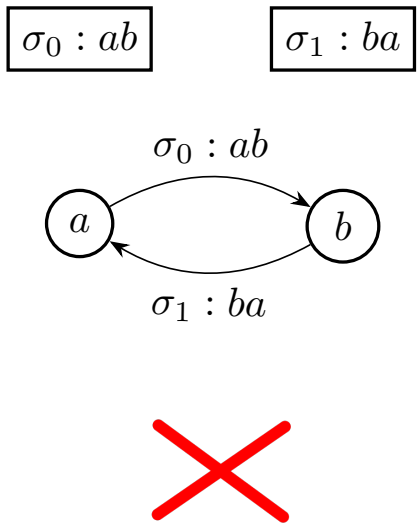
任给两个列表状态 σ_0、σ_1, 若它们含有两个共同元素 a、b, 则 a、b 在 σ_0 与 σ_1 中的相对顺序保持一致。

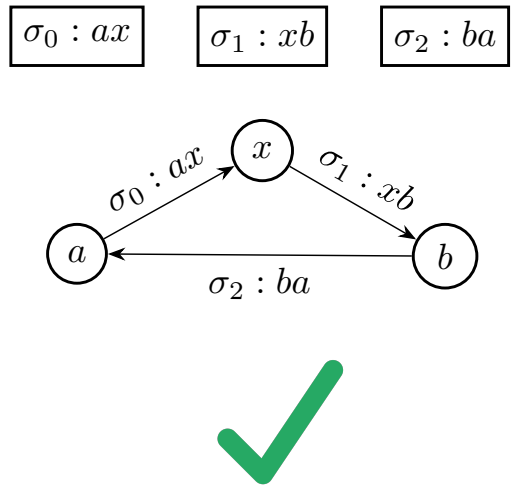
$\boxed{\sigma_0 : ab}$ $\boxed{\sigma_1 : ba}$



定义 (状态对兼容性 (Pairwise State Compatibility Property))

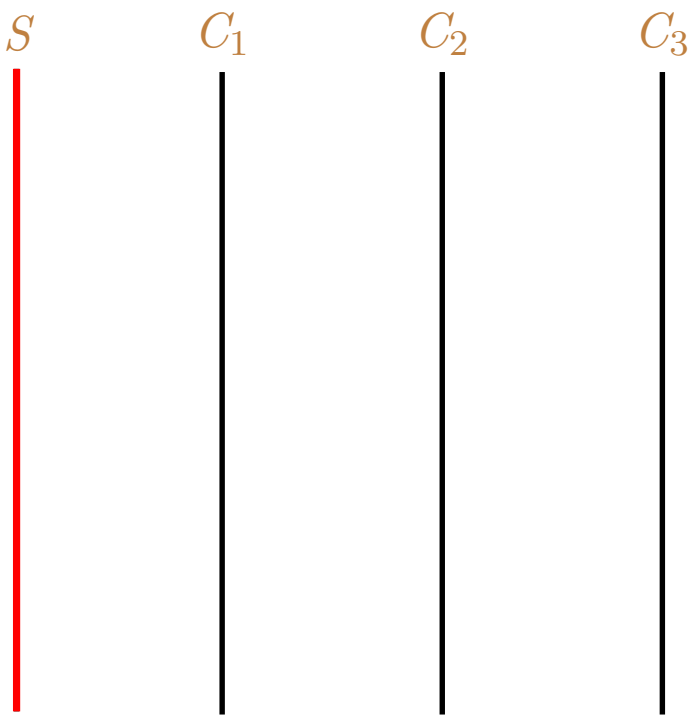
任给两个列表状态 σ_0、σ_1, 若它们含有两个共同元素 a、b, 则 a、b 在 σ_0 与 σ_1 中的相对顺序保持一致。



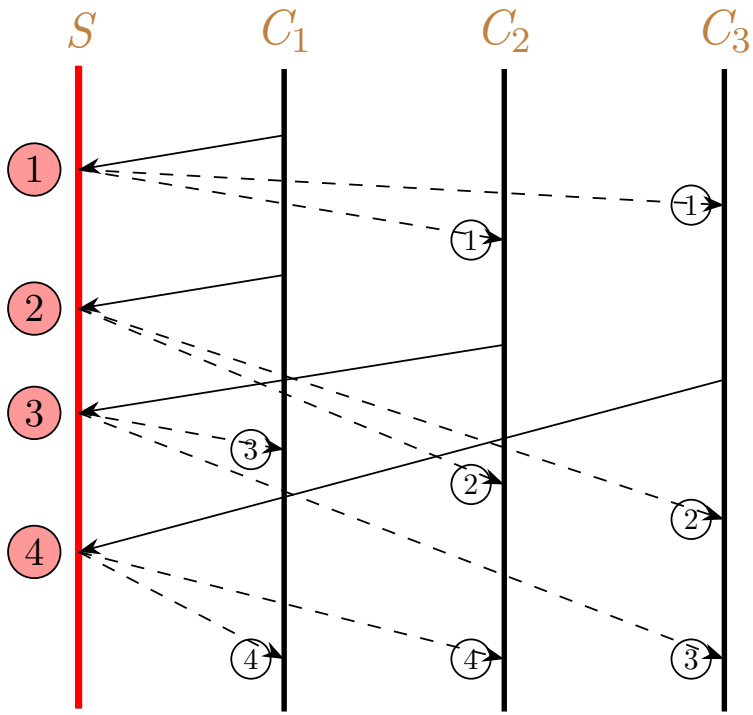


Jupiter

$(n+1)$ replica $\triangleq$ (n) Client $+$ (1) Server [Nichols et al., 1995]



$(n+1)$ replica $\triangleq$ (n) Client $+$ (1) Server [Nichols et al., 1995]



Server 负责将所有操作序列化

用户操作 —提交→ Client —FIFO→ Server —FIFO→ 其它 Clients

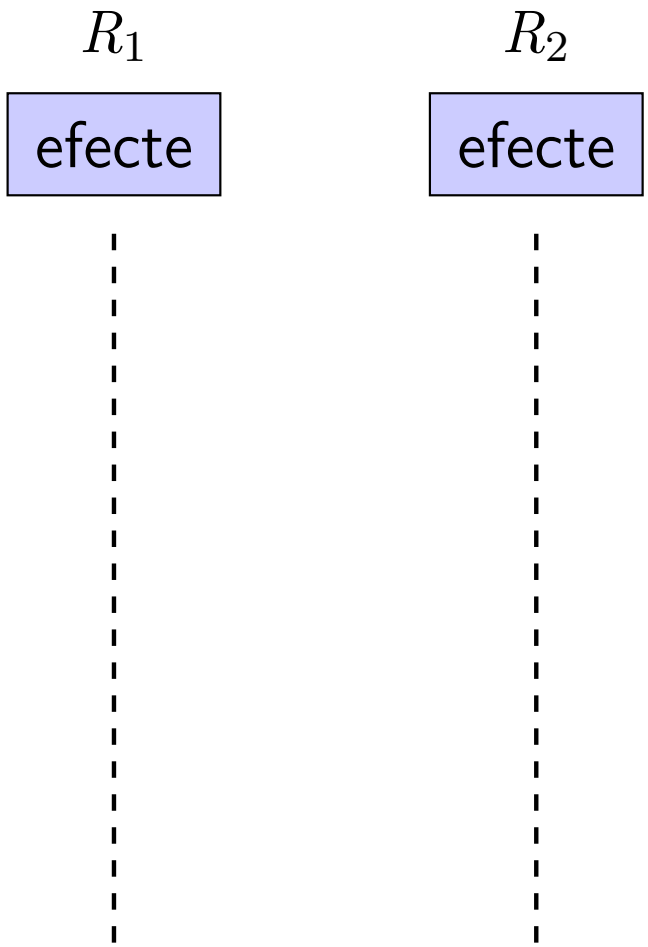
立即返回

更新操作

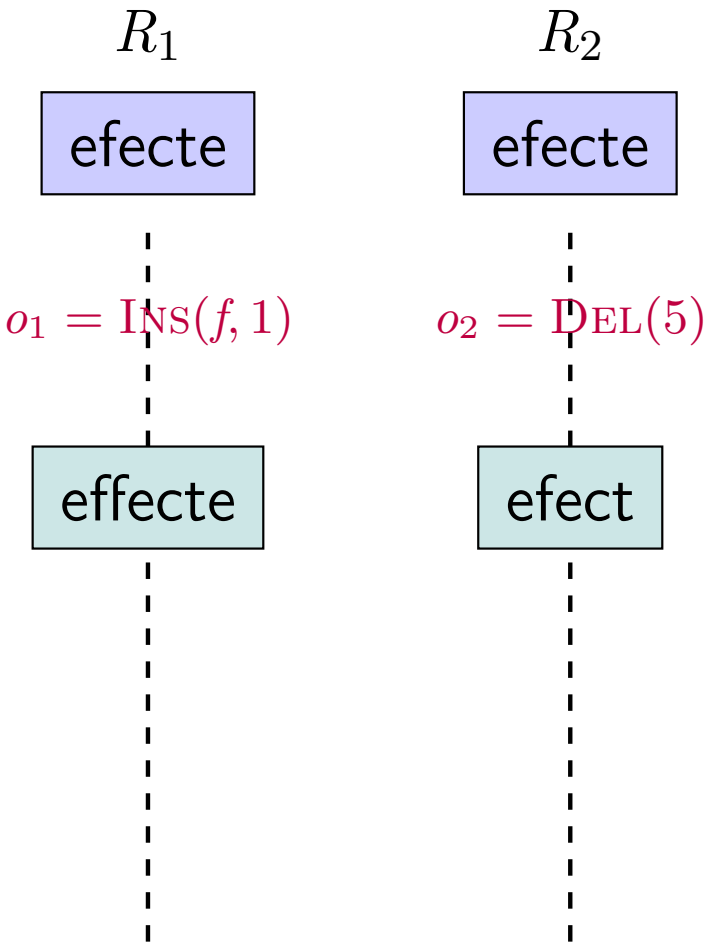
更新操作

操作转换 (Operational Transformation; OT) [Ellis and Gibbs, 1989] 技术

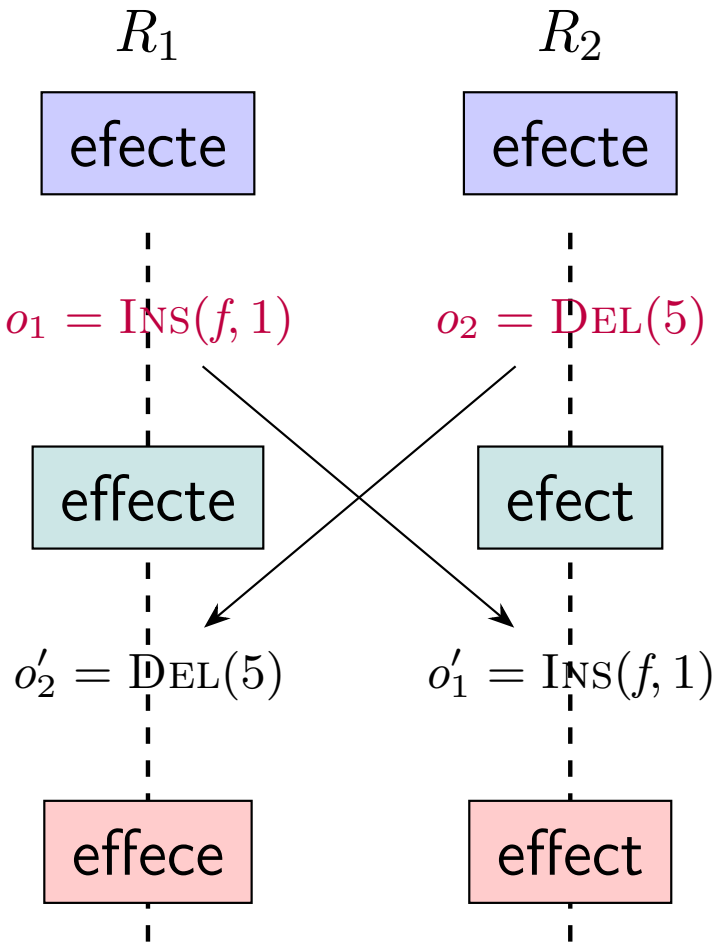
操作转换 (Operational Transformation; OT) [Ellis and Gibbs, 1989] 技术



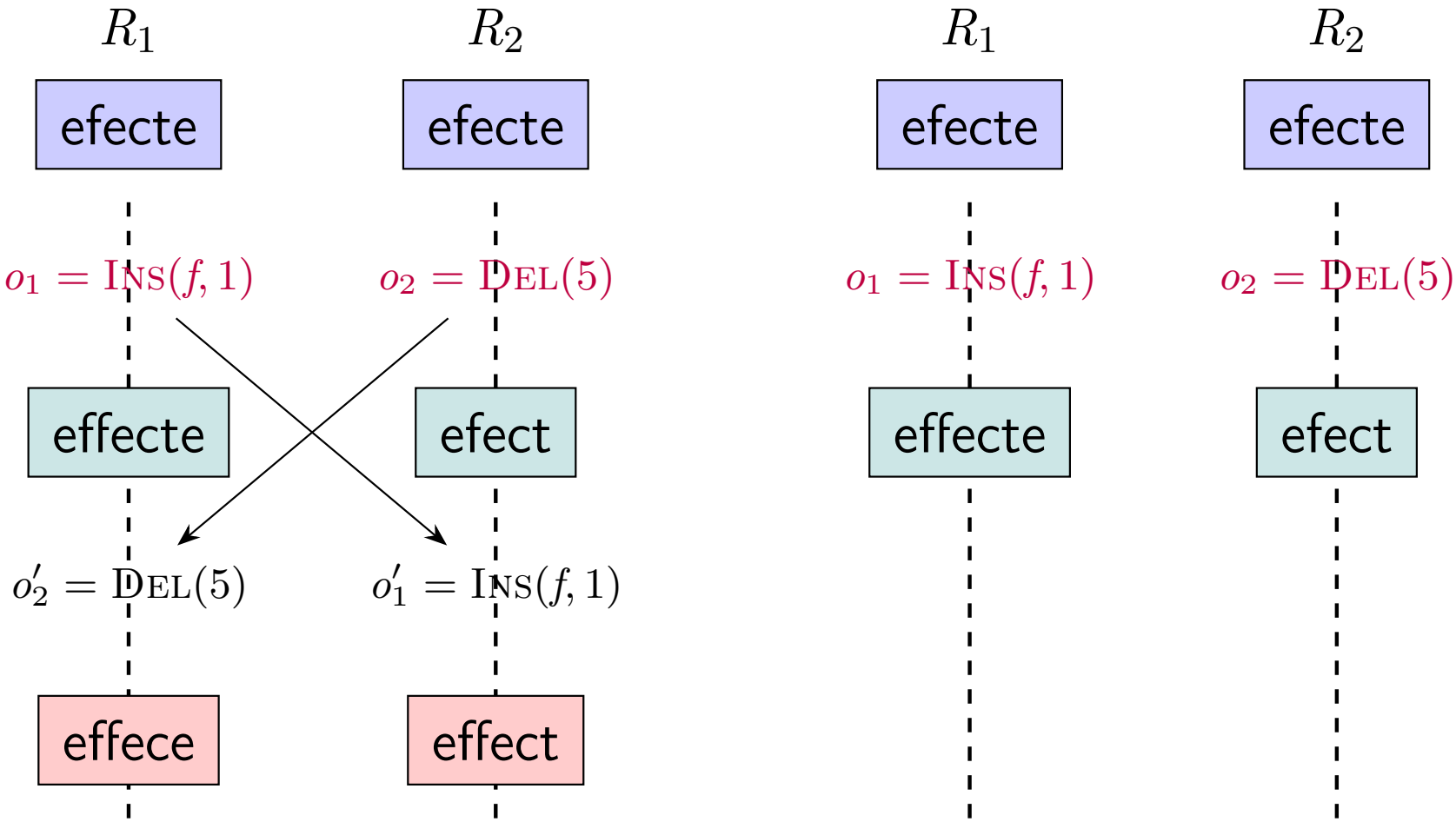
操作转换 (Operational Transformation; OT) [Ellis and Gibbs, 1989] 技术



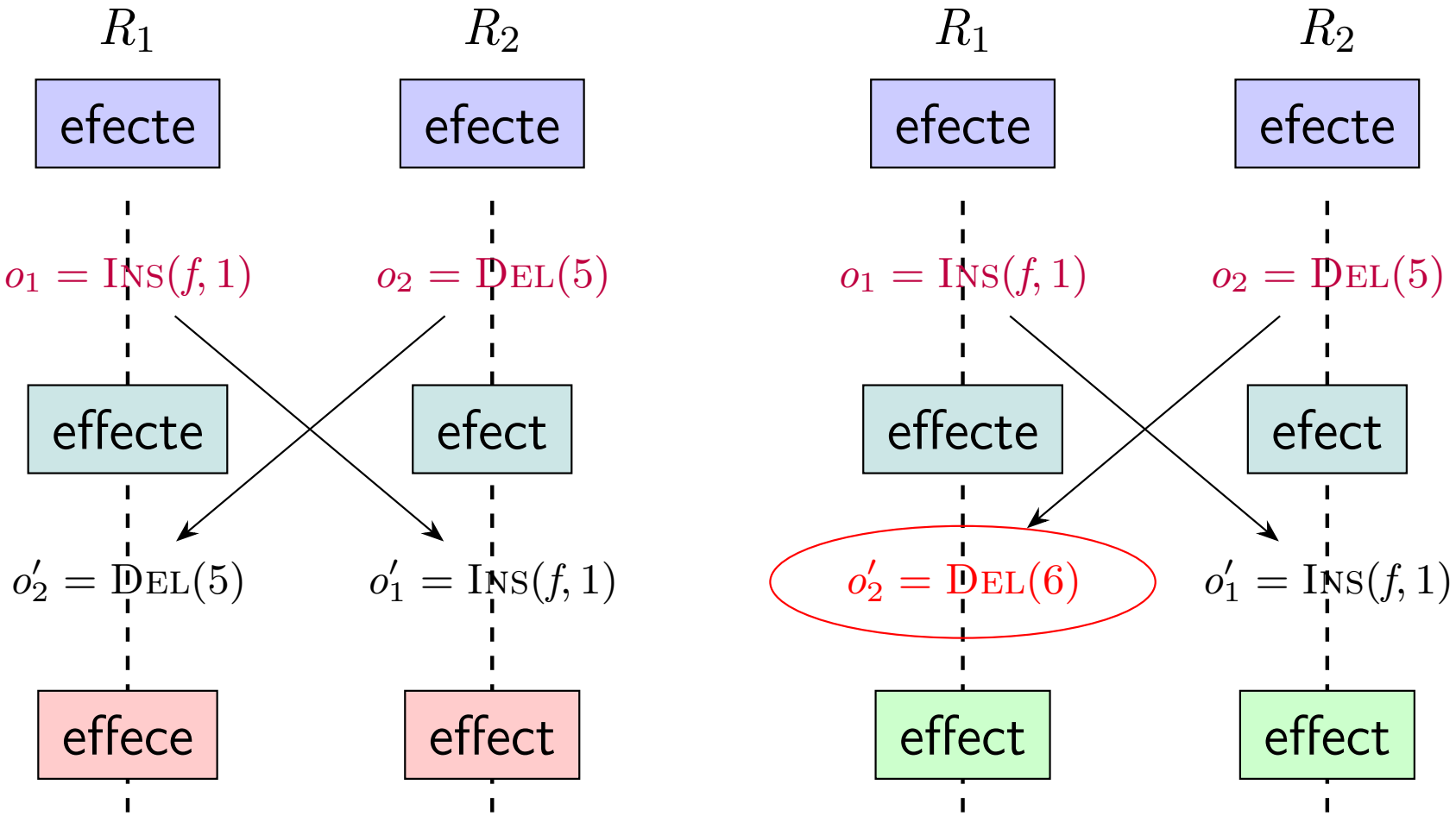
操作转换 (Operational Transformation; OT) [Ellis and Gibbs, 1989] 技术

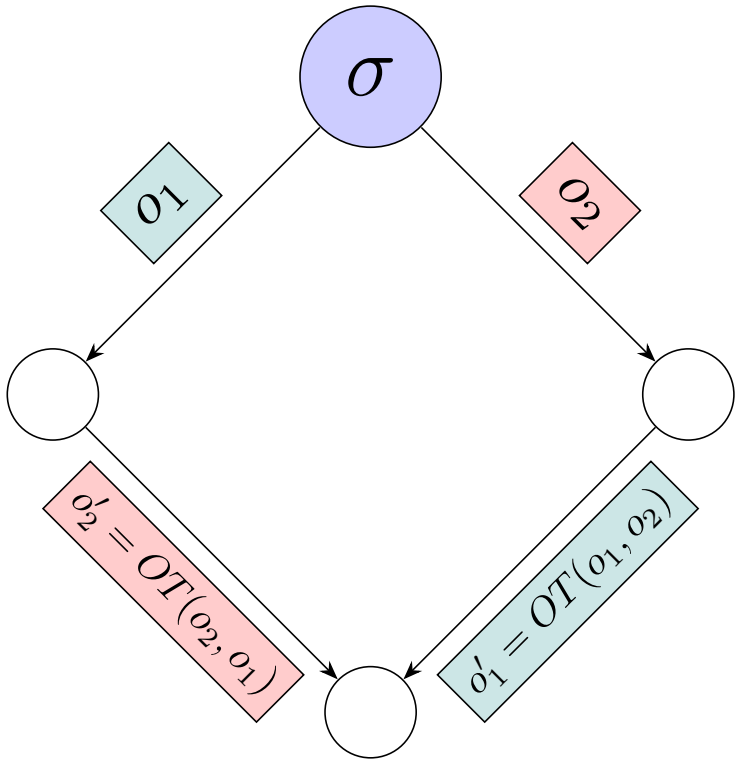


操作转换 (Operational Transformation; OT) [Ellis and Gibbs, 1989] 技术



操作转换 (Operational Transformation; OT) [Ellis and Gibbs, 1989] 技术





交换律 $\sigma; o_1; o_2' \equiv \sigma; o_2; o_1'$

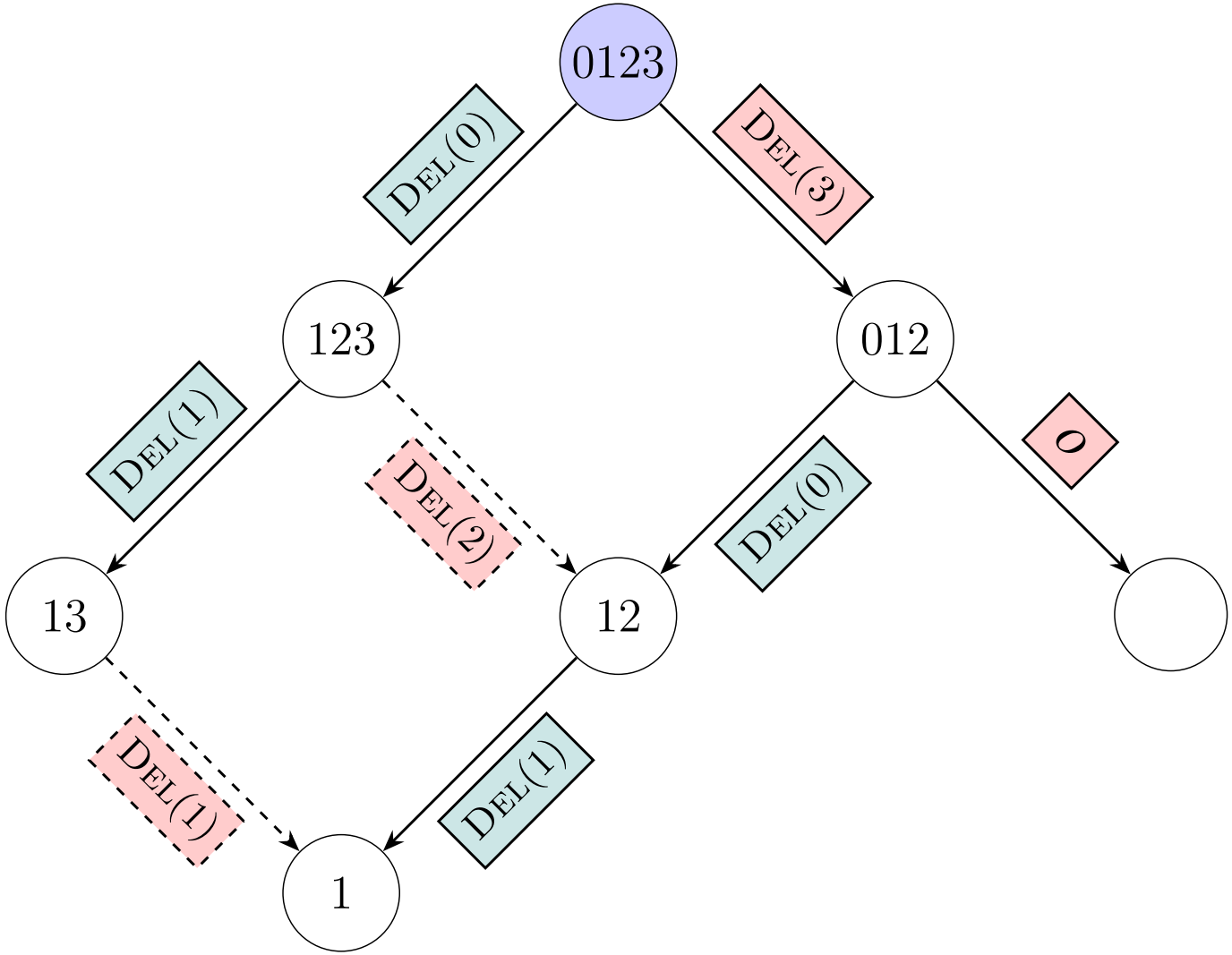
针对列表的操作转换函数 [Ellis and Gibbs, 1989]

$$OT\Big(\textsc{Ins}(a_1, p_1, pr_1), \textsc{Ins}(a_2, p_2, pr_2)\Big) = \begin{cases} \textsc{Ins}(a_1, p_1, pr_1) & p_1 < p_2 \\ \textsc{Ins}(a_1, p_1 + 1, pr_1) & p_1 > p_2 \\ \text{NOP} & p_1 = p_2 \wedge a_1 = a_2 \\ \textsc{Ins}(a_1, p_1 + 1, pr_1) & p_1 = p_2 \wedge a_1 \neq a_2 \wedge pr_1 > pr_2 \\ \textsc{Ins}(a_1, p_1, pr_1) & p_1 = p_2 \wedge a_1 \neq a_2 \wedge pr_1 \leq pr_2 \end{cases}$$

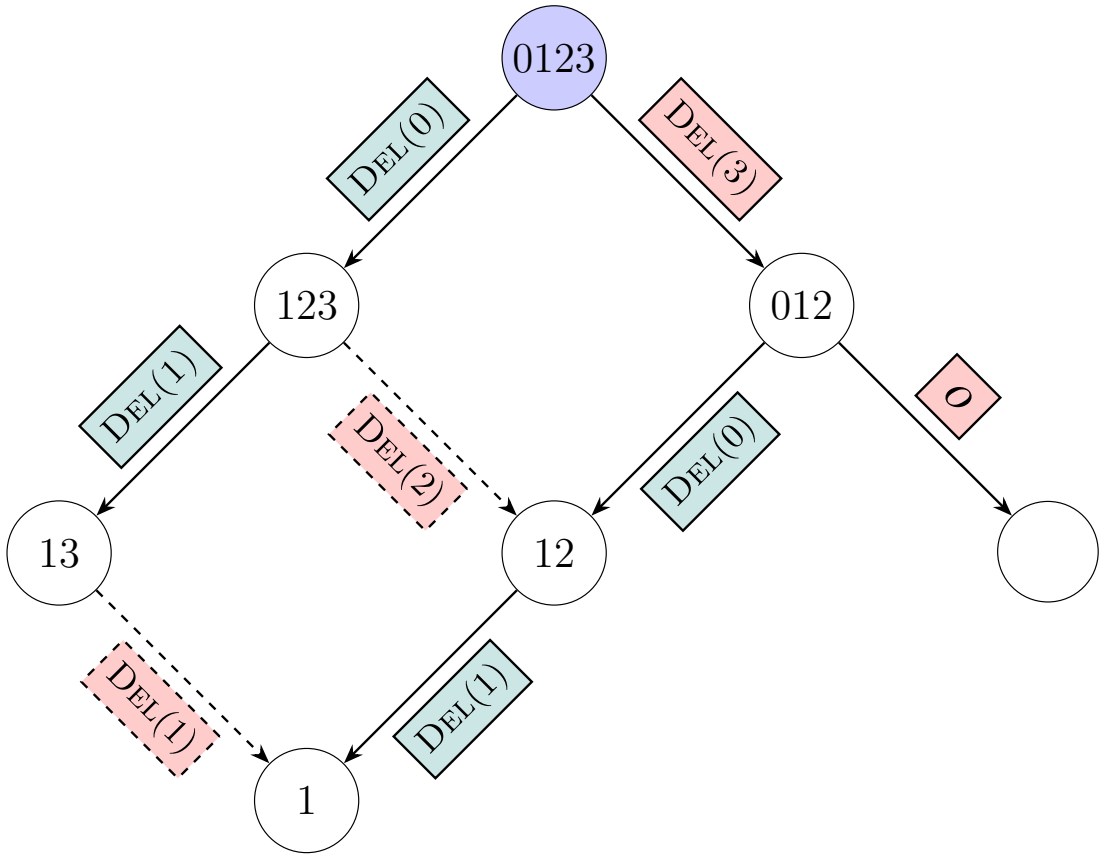
$$OT\Big(\textsc{Ins}(a_1, p_1, pr_1), \textsc{Del}(\_, p_2, pr_2)\Big) = \begin{cases} \textsc{Ins}(a_1, p_1, pr_1) & p_1 \leq p_2 \\ \textsc{Ins}(a_1, p_1 - 1, pr_1) & p_1 > p_2 \end{cases}$$

$$OT\Big(\textsc{Del}(\_, p_1, pr_1), \textsc{Ins}(a_2, p_2, pr_2)\Big) = \begin{cases} \textsc{Del}(\_, p_1, pr_1) & p_1 < p_2 \\ \textsc{Del}(\_, p_1 + 1, pr_1) & p_1 \geq p_2 \end{cases}$$

$$OT\Big(\textsc{Del}(\_, p_1, pr_1), \textsc{Del}(\_, p_2, pr_2)\Big) = \begin{cases} \textsc{Del}(\_, p_1, pr_1) & p_1 < p_2 \\ \textsc{Del}(\_, p_1 - 1, pr_1) & p_1 > p_2 \\ \text{NOP} & p_1 = p_2 \end{cases}$$

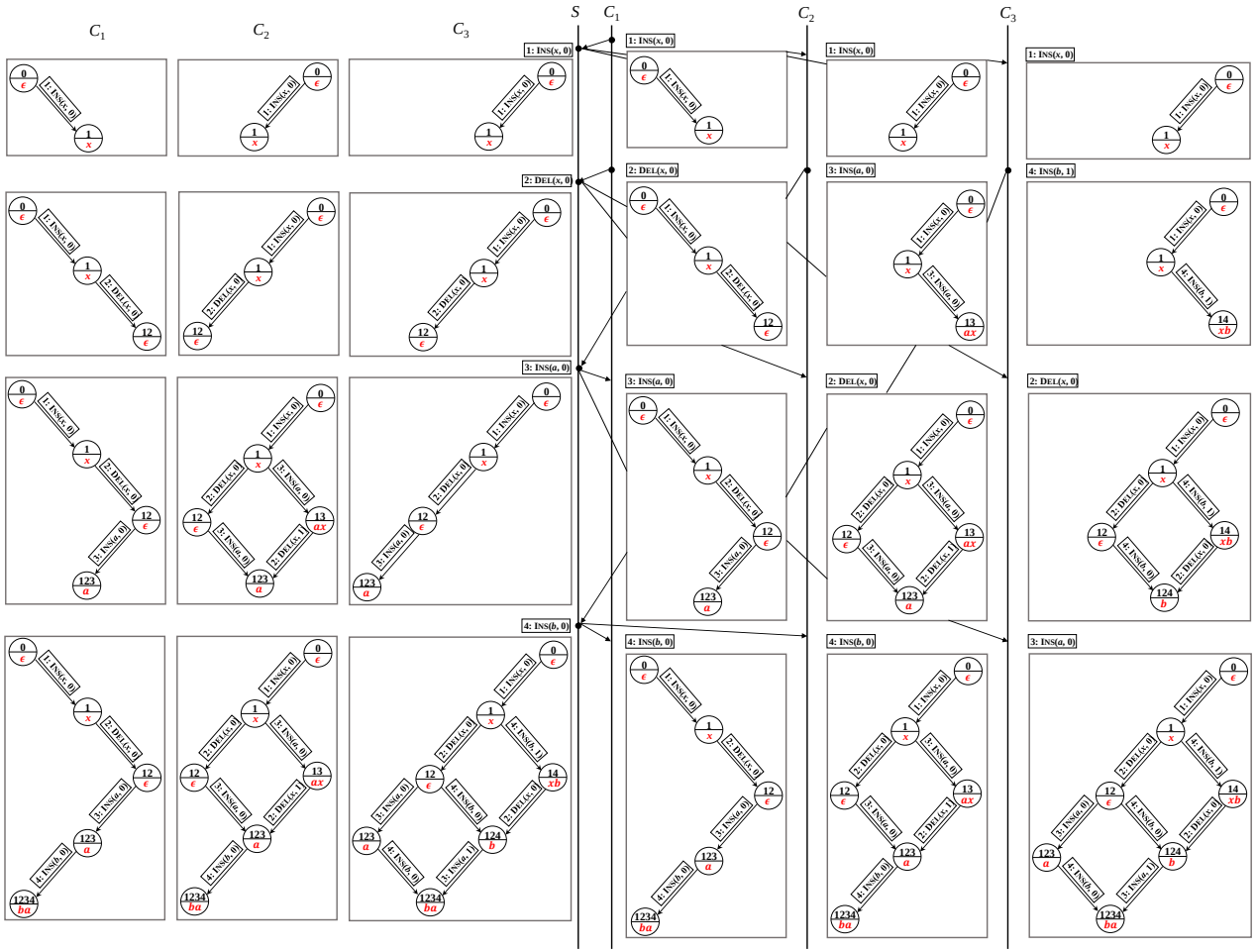

0123
Del(0)
Del(3)
123
012
Del(1)
Del(2)
Del(0)
0
13
12
Del(1)
Del(1)
1

利用数据结构 $2D$ 状态空间 [Xu, Sun, and Li, 2014] 控制何时以及如何执行 “操作转换”



$2D$: Local *vs.* Global

每个 Client 维护一个 $2D$ 状态空间



Server 维护 n 个 $2D$ 状态空间, 与 n 个 Clients 对应

$\mathcal{A}_{\text{weak}}$ 所规定的全局性质

Jupiter 协议中，每个 replica 所维护的局部视图

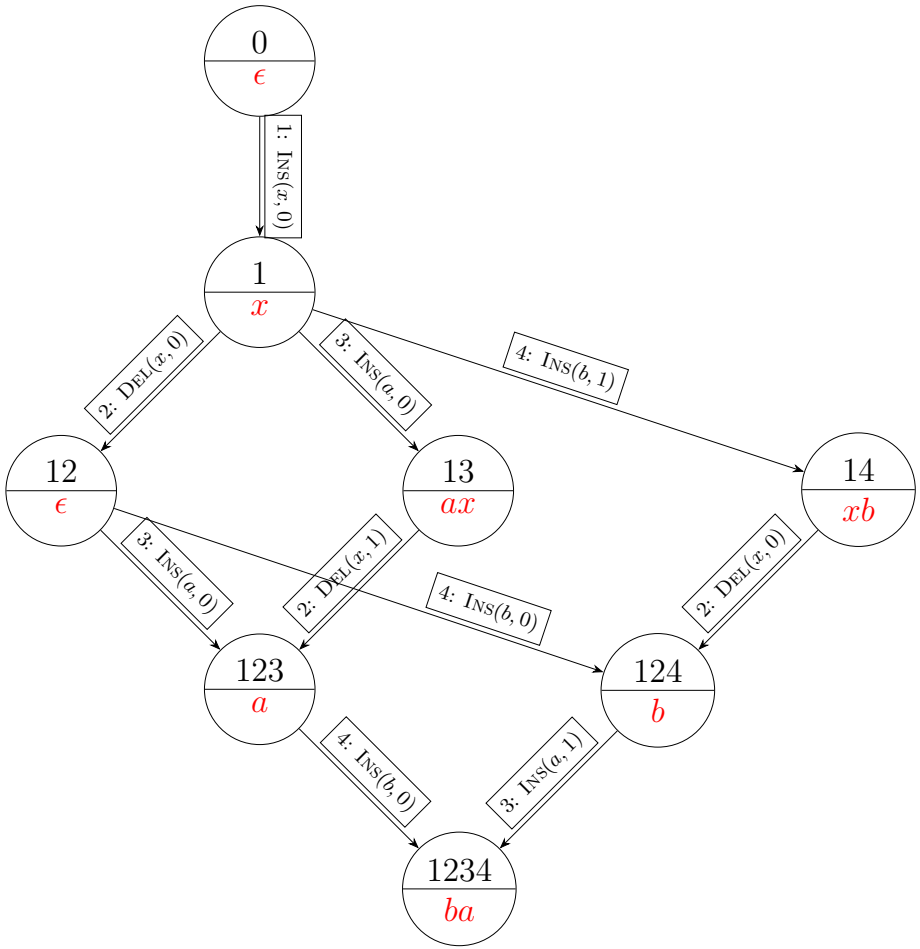
CJupiter (Compact Jupiter)

CJupiter (Compact Jupiter)

Theorem (等价性)

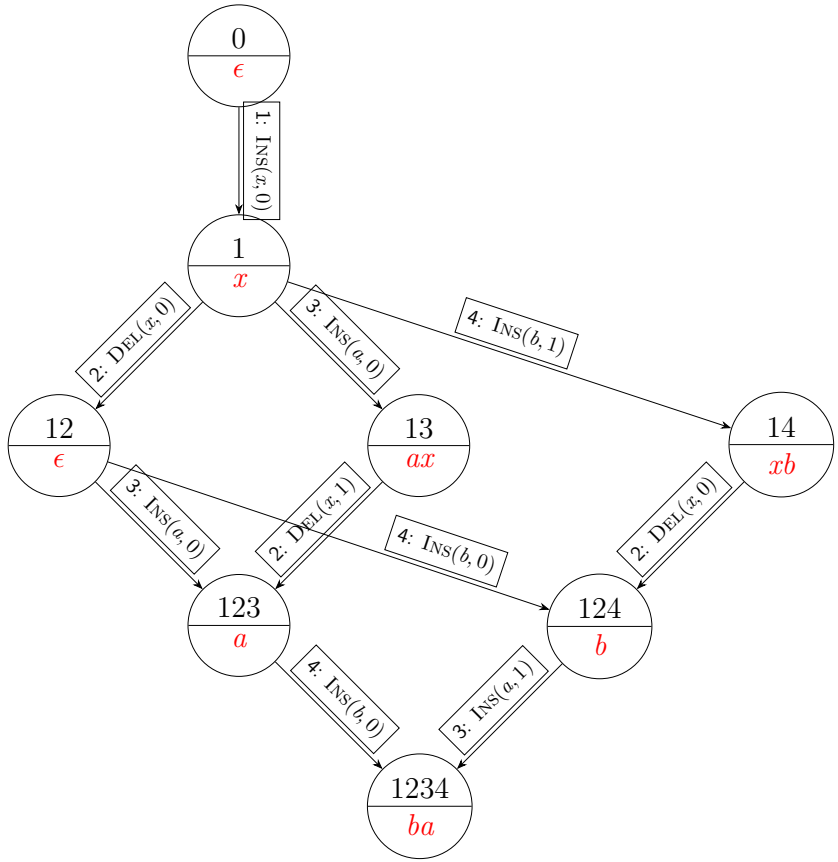
在相同的操作调度下，*CJupiter* 与 *Jupiter* 中的对应 *replica* 的行为 *(*状态序列*)* 是相同的。

CJupiter 为每个 replica 维护一个 n-ary 有序状态空间



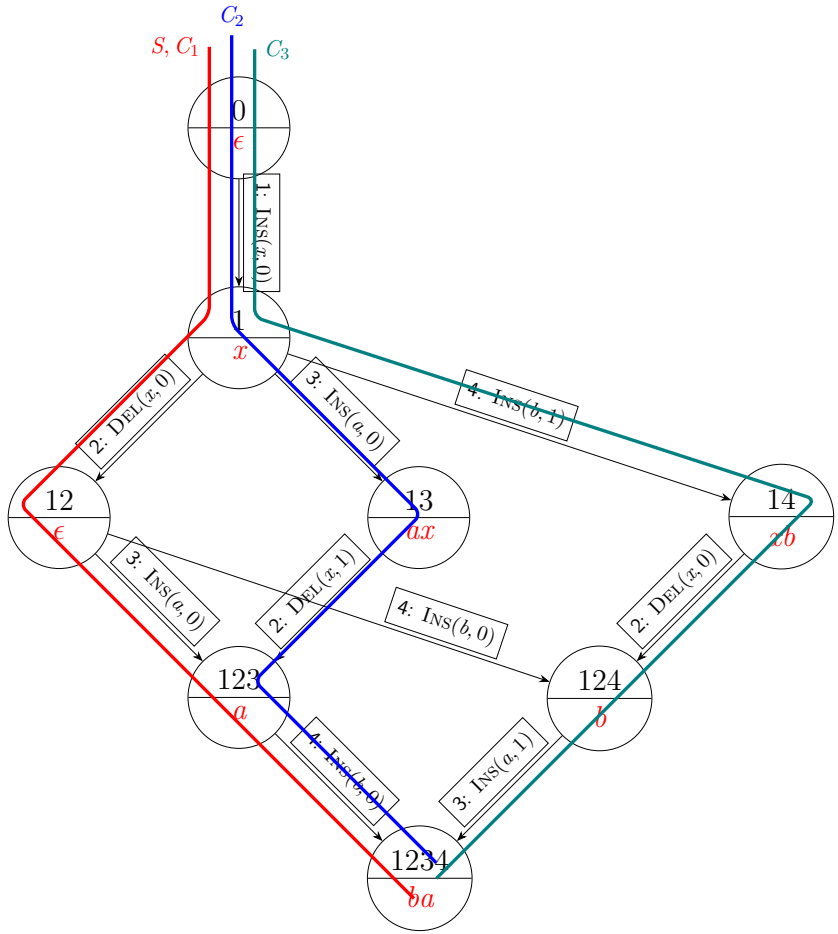
命题 (Compactness of CJupiter)

CJupiter 所维护的 $(n+1)$ 个 *n-ary* 有序状态空间是相同的。



命题 (Compactness of CJupiter)

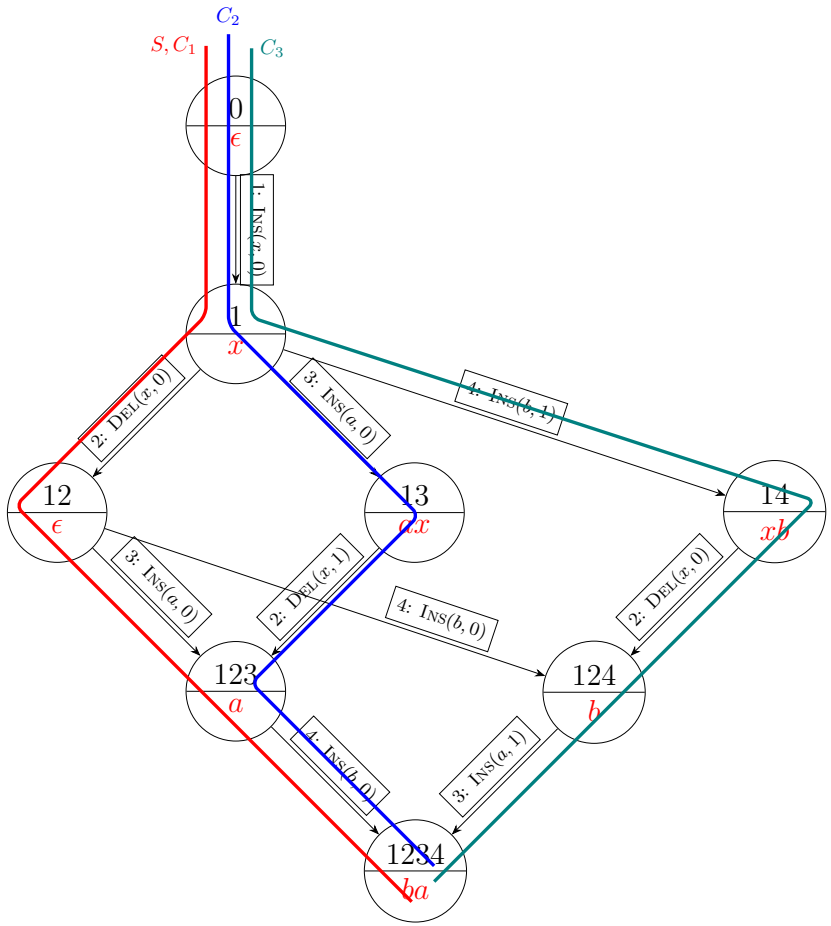
CJupiter 所维护的 $(n+1)$ 个 *n-ary* 有序状态空间是相同的。



每个 replica 的行为对应于该状态空间中的一条路径

CJupiter 满足 Weak List Specification

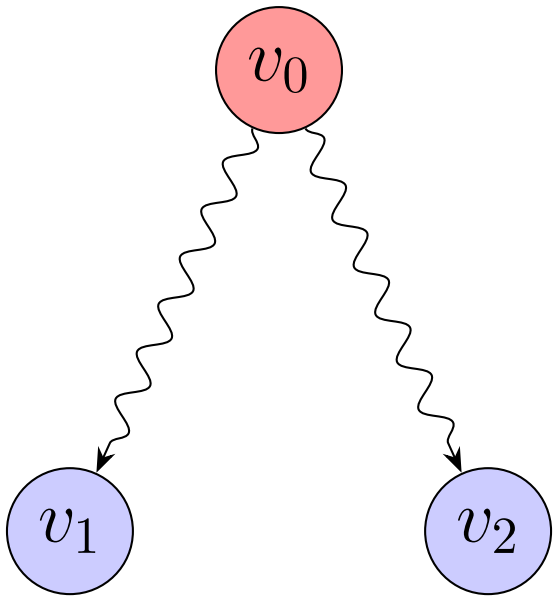
关注某个 n-ary 有序状态空间，三步骤证明状态对兼容性



1 任取两个状态节点 v_1 和 v_2

引理 (LCA (Lowest Common Ancestor))

n-ary 有序状态空间中的任意一对状态节点都有唯一的最近公共祖先。

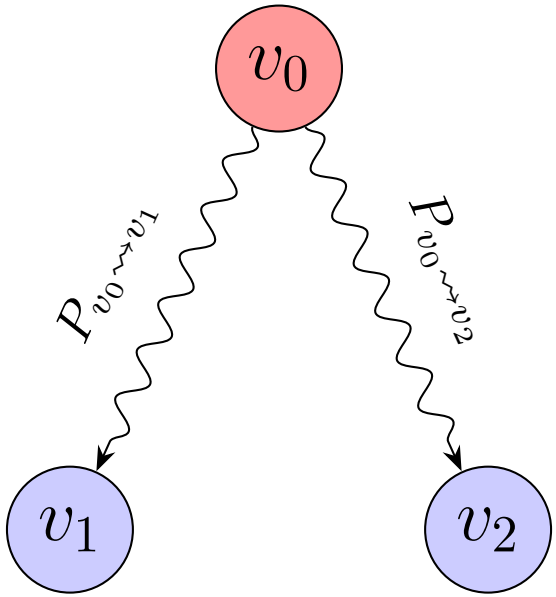


$$v_0 = \mathrm{LCA}(v_1, v_2)$$

2 考虑从 $v_0 = \mathrm{LCA}(v_1, v_2)$ 到 v_1 和 v_2 的两条路径

引理 (Disjoint Paths)

路径 $P_{v_0 \rightsquigarrow v_1}$ 上包含的操作集 $O_{v_0 \rightsquigarrow v_1}$ 与路径 $P_{v_0 \rightsquigarrow v_2}$ 上包含的操作集 $O_{v_0 \rightsquigarrow v_2}$ 不相交。

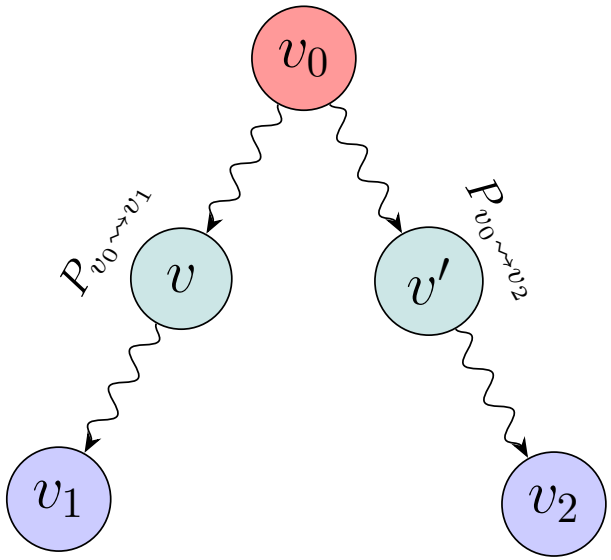


$$v_0 = \mathrm{LCA}(v_1, v_2)$$

3 考虑两条路径上的状态

引理 (Compatible Paths)

$P_{v_0 \rightsquigarrow v_1}$ 上的任一状态 v 与 $P_{v_0 \rightsquigarrow v_2}$ 上的任一状态 v' 是兼容的。

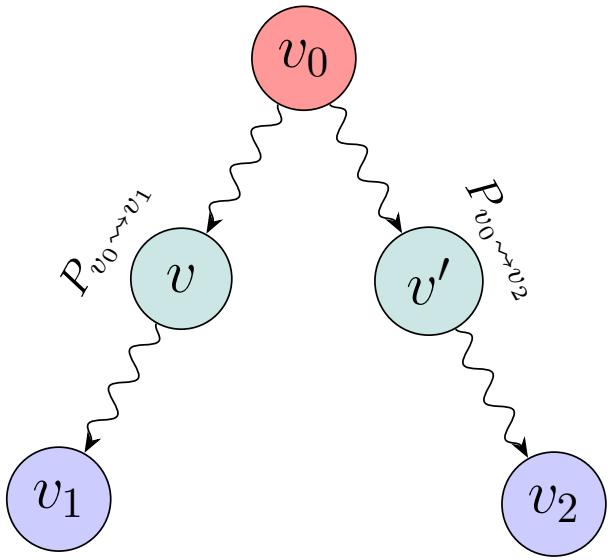


$$v_0 = \mathsf{LCA}(v_1, v_2)$$

3 考虑两条路径上的状态

引理 (Compatible Paths)

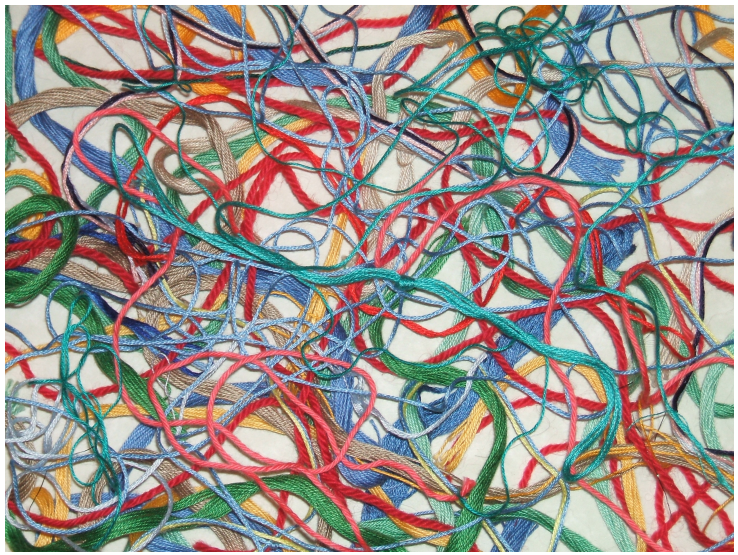
$P_{v_0 \rightsquigarrow v_1}$ 上的任一状态 v 与 $P_{v_0 \rightsquigarrow v_2}$ 上的任一状态 v' 是兼容的。



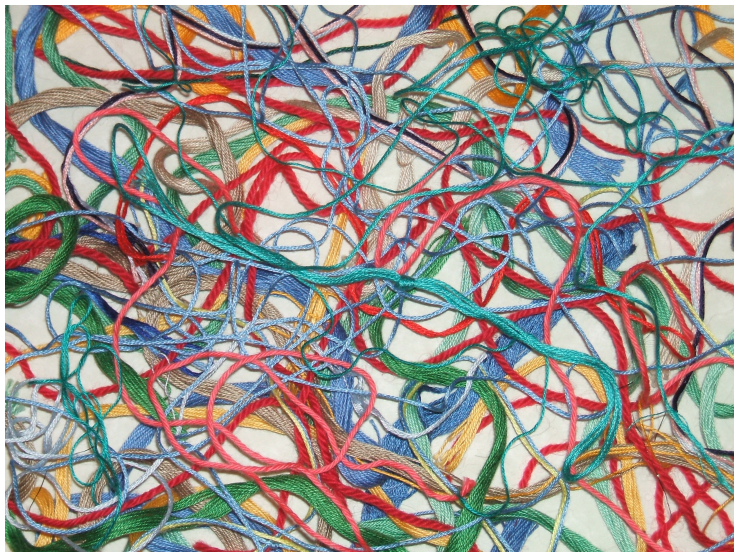
$$v_0 = \mathsf{LCA}(v_1, v_2)$$

$\therefore$ v_1 和 v_2 是兼容的

个人感觉: 基于 OT 思想的协议晦涩难懂



个人感觉: 基于 OT 思想的协议晦涩难懂

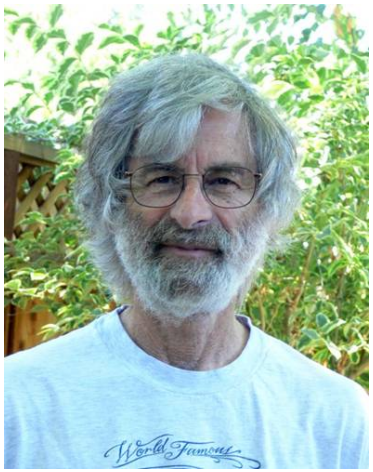


- 协议多种多样
- 经常不加证明
- 证明是错误的

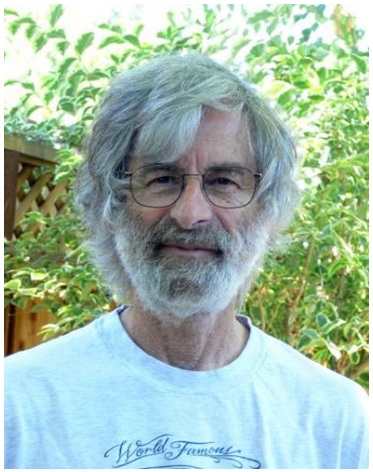
个人感觉: 基于 OT 思想的协议晦涩难懂

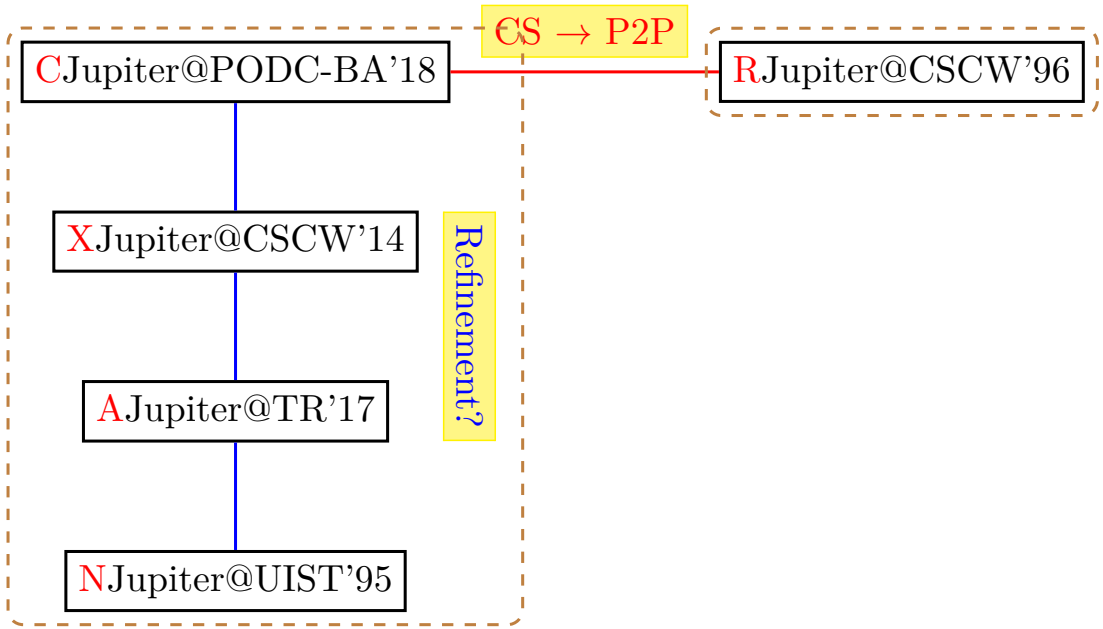
- 协议多种多样
- 经常不加证明
- 证明是错误的
- 勘误也是错的

Model Checking: 使用 TLAPLUS



Model Checking: 使用 TLAPLUS






Attiya, Hagit et al. (2016). "Specification and complexity of collaborative text editing". In: *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*. PODC '16. ACM, pp. 259–268.


Burckhardt, Sebastian et al. (2014). "Replicated Data Types: Specification, Verification, Optimality". In: *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL '14. San Diego, California, USA: ACM, pp. 271–284. ISBN: 978-1-4503-2544-8. DOI: 10.1145/2535838.2535848. URL: http://doi.acm.org/10.1145/2535838.2535848.


Ellis, C. A. and S. J. Gibbs (1989). "Concurrency Control in Groupware Systems". In: *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data*. SIGMOD '89. ACM, pp. 399–407.


Herlihy, Maurice P. and Jeannette M. Wing (1990). "Linearizability: A Correctness Condition for Concurrent Objects". In: *ACM Trans. Program. Lang. Syst.* 12.3, pp. 463–492. ISSN: 0164-0925. DOI: 10.1145/78969.78972. URL: http://doi.acm.org/10.1145/78969.78972.


Liskov, Barbara and Stephen Zilles (1974). "Programming with Abstract Data Types". In: *Proceedings of the ACM SIGPLAN Symposium on Very High Level Languages*. Santa Monica, California, USA: ACM, pp. 50–59. DOI: 10.1145/800233.807045. URL: http://doi.acm.org/10.1145/800233.807045.


Nichols, David A. et al. (1995). "High-latency, Low-bandwidth Windowing in the Jupiter Collaboration System". In: *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*. UIST '95. ACM, pp. 111–120.


Shapiro, Marc et al. (2011a). *A comprehensive study of Convergent and Commutative Replicated Data Types*. Research Report RR-7506. Inria – Centre Paris-Rocquencourt ; INRIA, p. 50. URL: https://hal.inria.fr/inria-00555588.


Shapiro, Marc et al. (2011b). "Conflict-free Replicated Data Types". In: *Proceedings of the 13th International Conference on Stabilization, Safety, and Security of Distributed Systems*. SSS'11. Springer-Verlag, pp. 386–400.


Xu, Yi, Chengzheng Sun, and Mo Li (2014). "Achieving Convergence in Operational Transformation: Conditions, Mechanisms and Systems". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work*. CSCW '14. ACM, pp. 505–518.

Thank You!