北京大学
PEKING UNIVERSITY

# 当数学结构遇见数学猜想

杨仝
北京大学
yangtong@pku.edu.cn
https://yangtonghome.github.io/

# 报告摘要

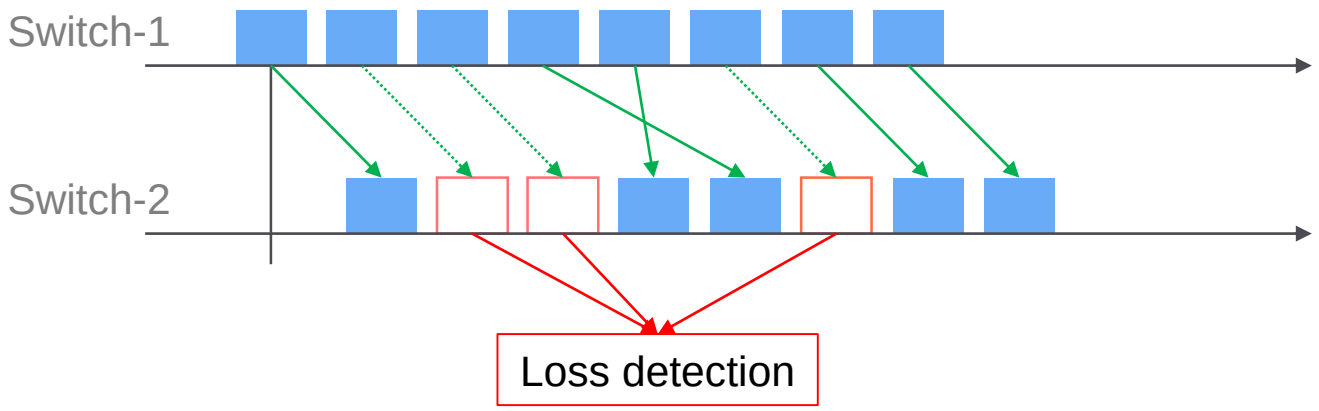
北京大学 PEKING UNIVERSITY

- 在数学的深邃海洋中，猜想犹如闪耀的灯塔，指引我们航行
- 通过融合这些猜想，我们设计出一些独特的概率数学结构——Sketch，以解决“集合—元素—集合”的关系问题
  - 丢包检测：费马小定理→费马Sketch
  - 多集合查询：四色定理→四色过滤器
  - 哥德巴赫猜想?

# 费马Sketch

# 问题定义

- 问题：单重/多重集合的差异
  - 丢包、多包
  - 组播、多播、黑洞
  - 安全设备：D清洗、网闸、NAT、病毒检测、防火墙、入侵检测/阻止



# 现有方案

北京大学 PEKING UNIVERSITY

- 两种解决方案
  - LossRadar
    - 内存开销与丢包数成正比
  - FlowRadar
    - 内存开销与总流数成正比

# 费马小定理

- 如果 $p$ 是质数，且 a 不是 $p$ 的倍数，我们有
  - $\mathrm{a}^{p-1}\ mod\ p = 1$
- 费马小定理历史
  - 1636年首先由费马发现
  - 1683年莱布尼茨在未发表的手稿中给出了证明
  - 1736年欧拉首次给出了正式的证明
  - 2500年前的中国猜想是费马小定理的一种特殊情况

# 费马Sketch / 定义

北京大学 PEKING UNIVERSITY

- 一种概率数据结构：丢包检测
- 特点：
  - 采用哈希映射
  - 几乎无错
  - 核心操作过程利用费马小定理还原丢包流ID以及丢包数
- 解码成功，无错
- 解码失败，出错

# 费马Sketch / 思想

北京大学 PEKING UNIVERSITY

- 将丢包问题转化为多重集合求差集问题
- 多重集合中的元素的结构为<流ID, 流频数>
- 多重集合的差集即为丢包流的集合
- 用模操作将流ID压缩到素域中
- 用费马小定理解码还原流ID

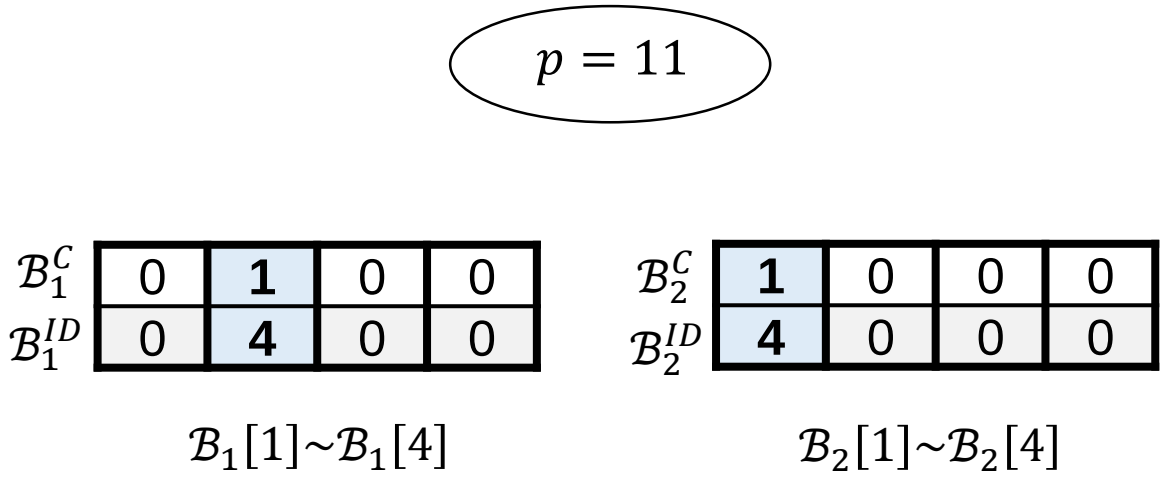
# 费马Sketch/ 数据结构与操作

北京大学 PEKING UNIVERSITY

- 数据结构
- 插入Insertion/encoding
- 删除deletion
- 解码decoding

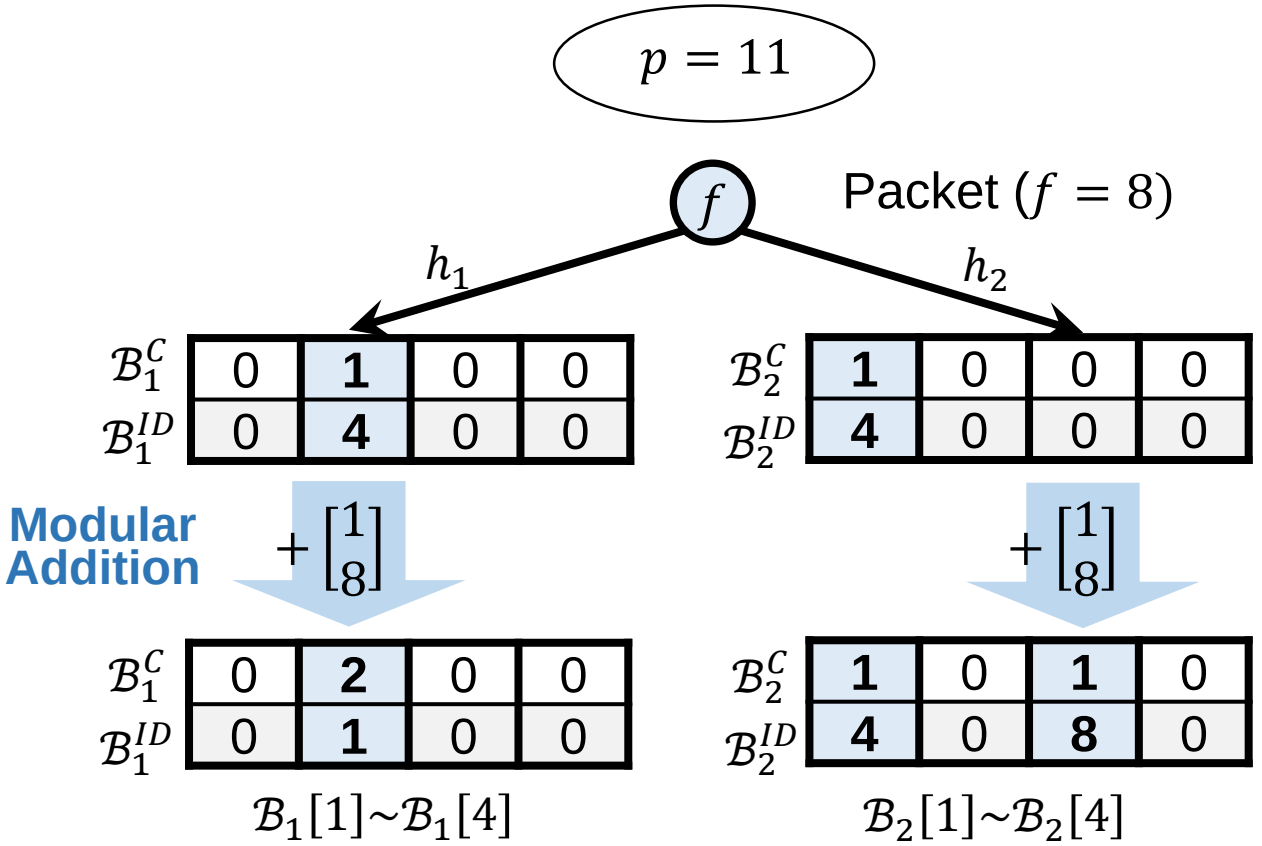
# 费马Sketch/ 数据结构

- $d$个桶数组
- Count 域
- IDsum 域
- 大质数$p$



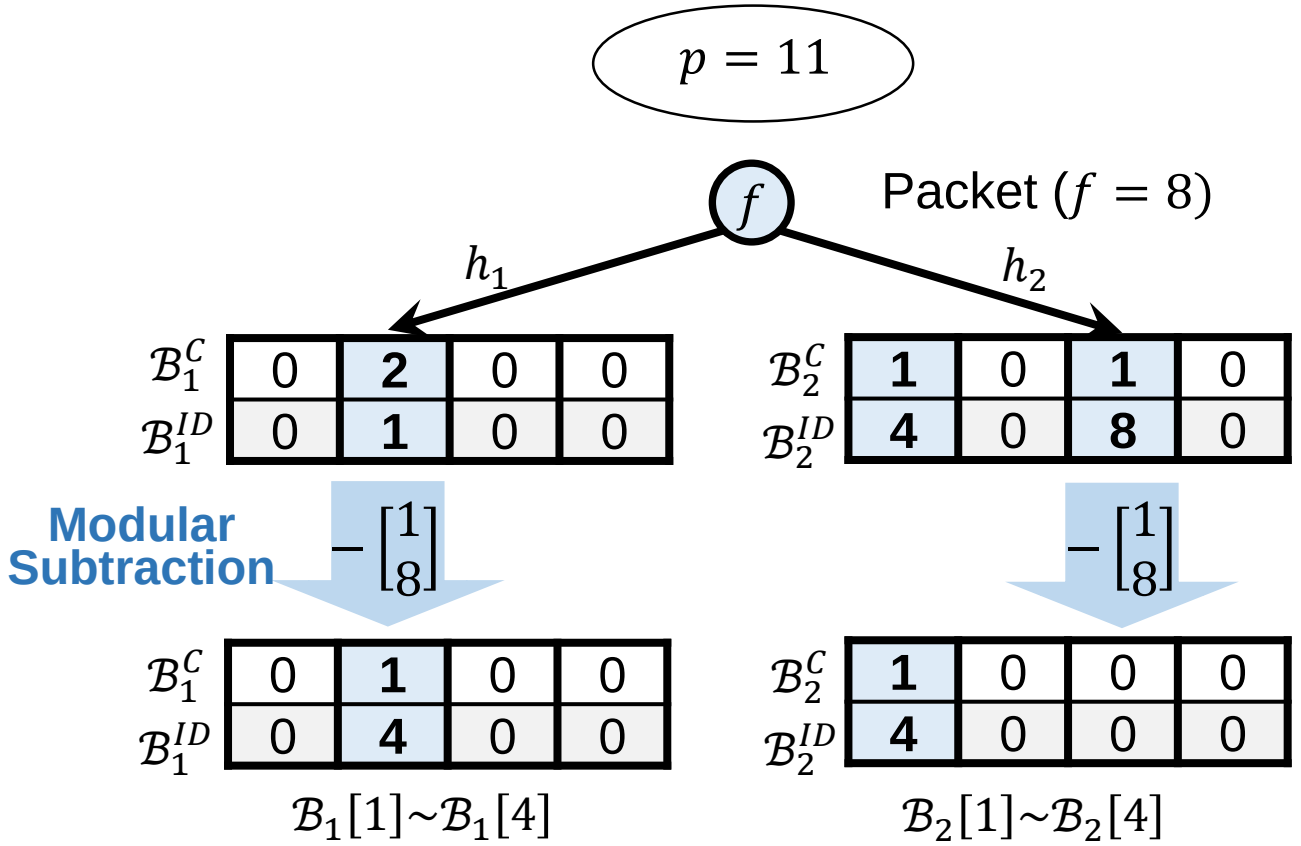
# 费马Sketch/ 插入

- 流ID为 f 的数据包
- 哈希定位桶
- Count 域加1
- IDsum域加$f$后模$p$



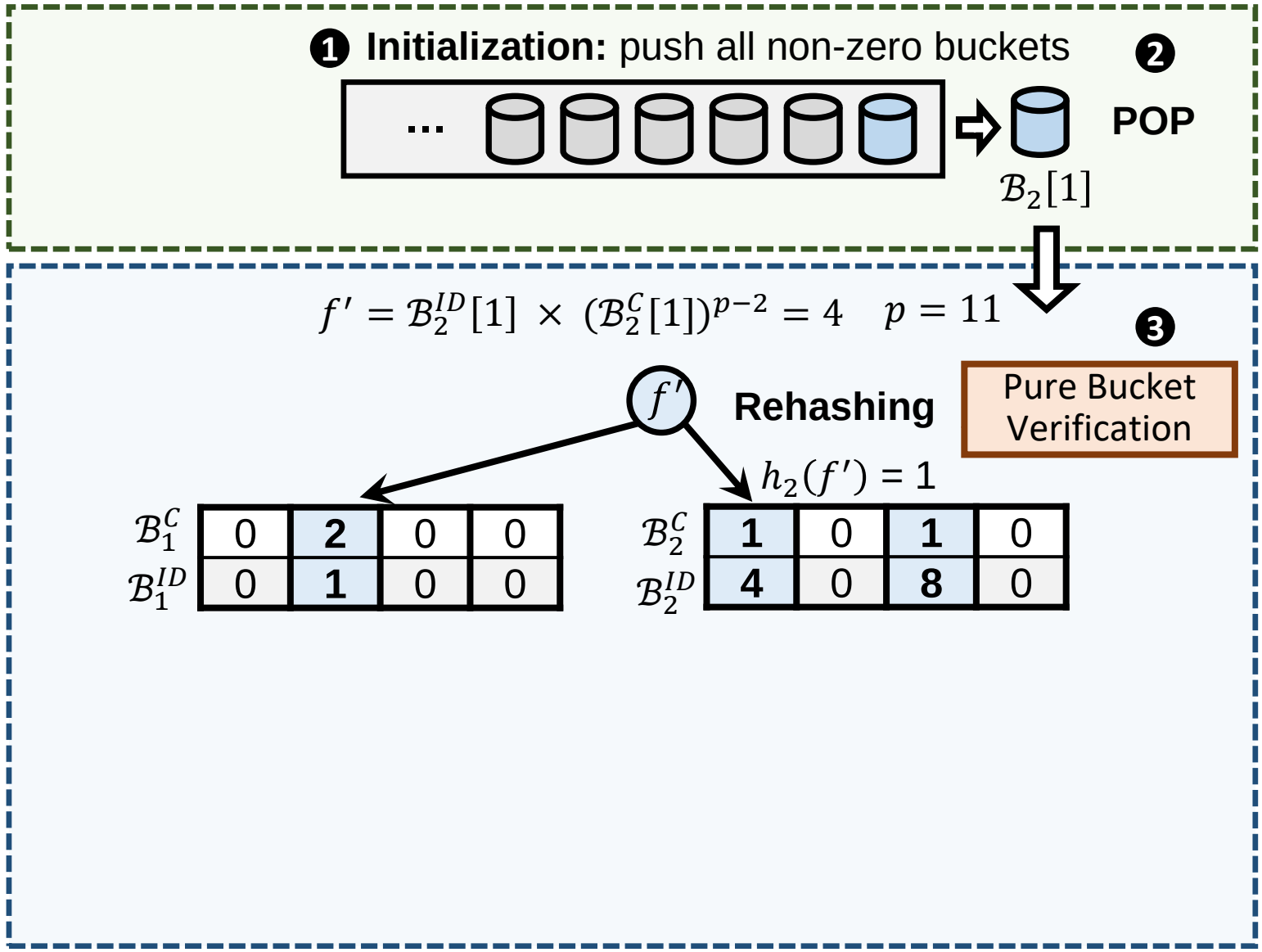
# 费马Sketch/ 删除

- 流ID为 f 的数据包
- 哈希定位桶
- Count 域减1
- IDsum域减$f$后模$p$



# 费马Sketch/ 解码

- 七步循环
- 纯净桶校验

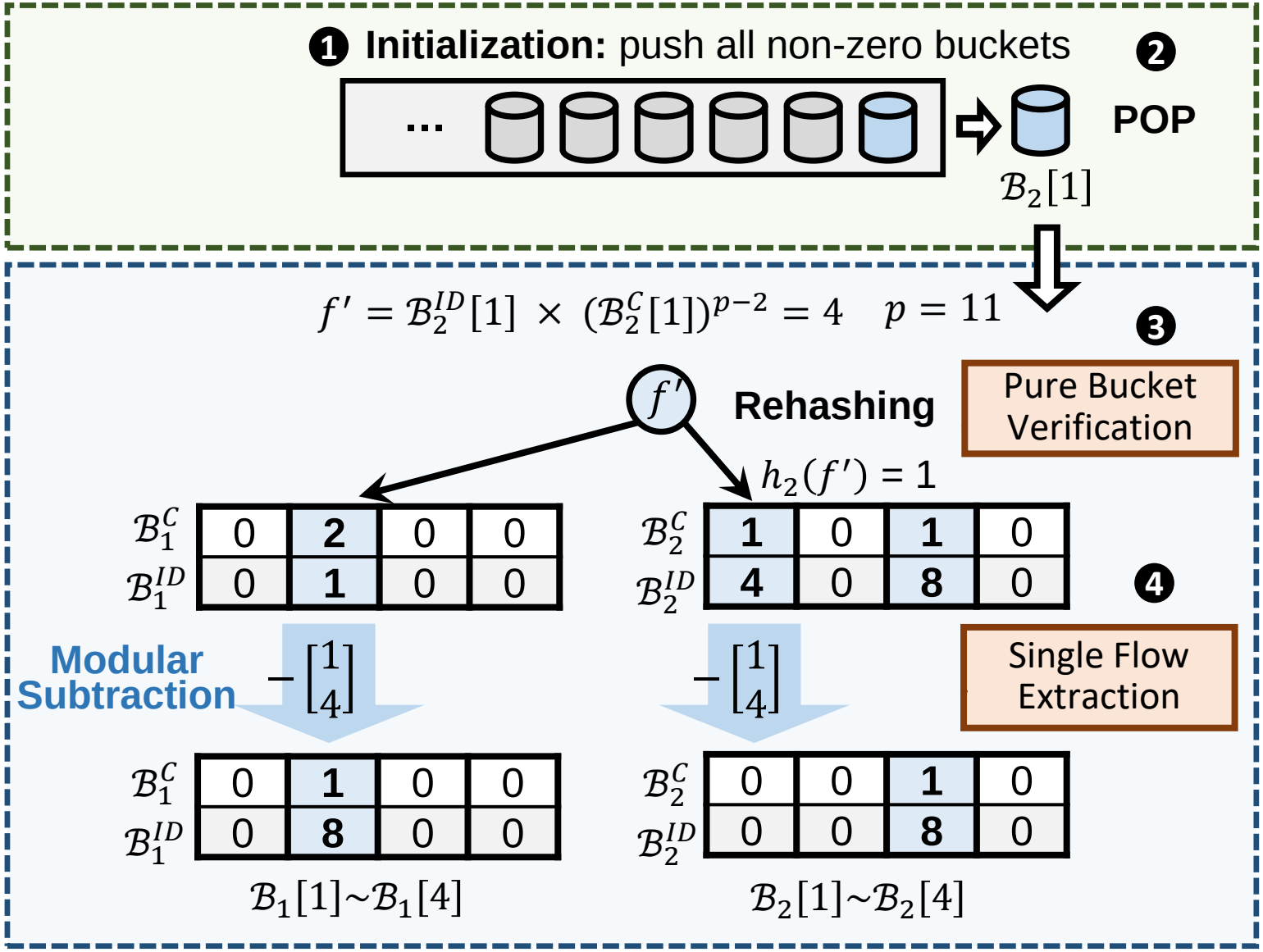


# 费马Sketch/ 纯净桶校验

- 纯净桶：只被一条流哈希映射到的桶
- 重哈希校验（Rehash verification）
  - 假设桶$\mathcal{B}_i[j]$中只有一条流，标识为ID
  - 应满足$\mathcal{B}_{\mathrm{i}}^{ID}[j]$= (ID $*$ $\mathcal{B}_{\mathrm{i}}^{c}[j]$) mod $p$
  - 由费马小定理，因为$\mathcal{B}_{\mathrm{i}}^{c}[j]$与$p$互质，($\mathcal{B}_{\mathrm{i}}^{c}[j]$ ^(p-1)) mod p =1
  - ($\mathcal{B}_{\mathrm{i}}^{ID}[j]$ * $\mathcal{B}_{\mathrm{i}}^{c}[j]$ ^(p-2)) mod p= (ID * $\mathcal{B}_{\mathrm{i}}^{c}[j]$ ^(p-1)) mod p = ID
  - 通过$h_{\mathrm{i}}$(ID)是否等于j 来判断是否这个桶是纯净桶

# 费马Sketch/ 解码

- 七步循环
- 纯净桶校验
- 单流萃取



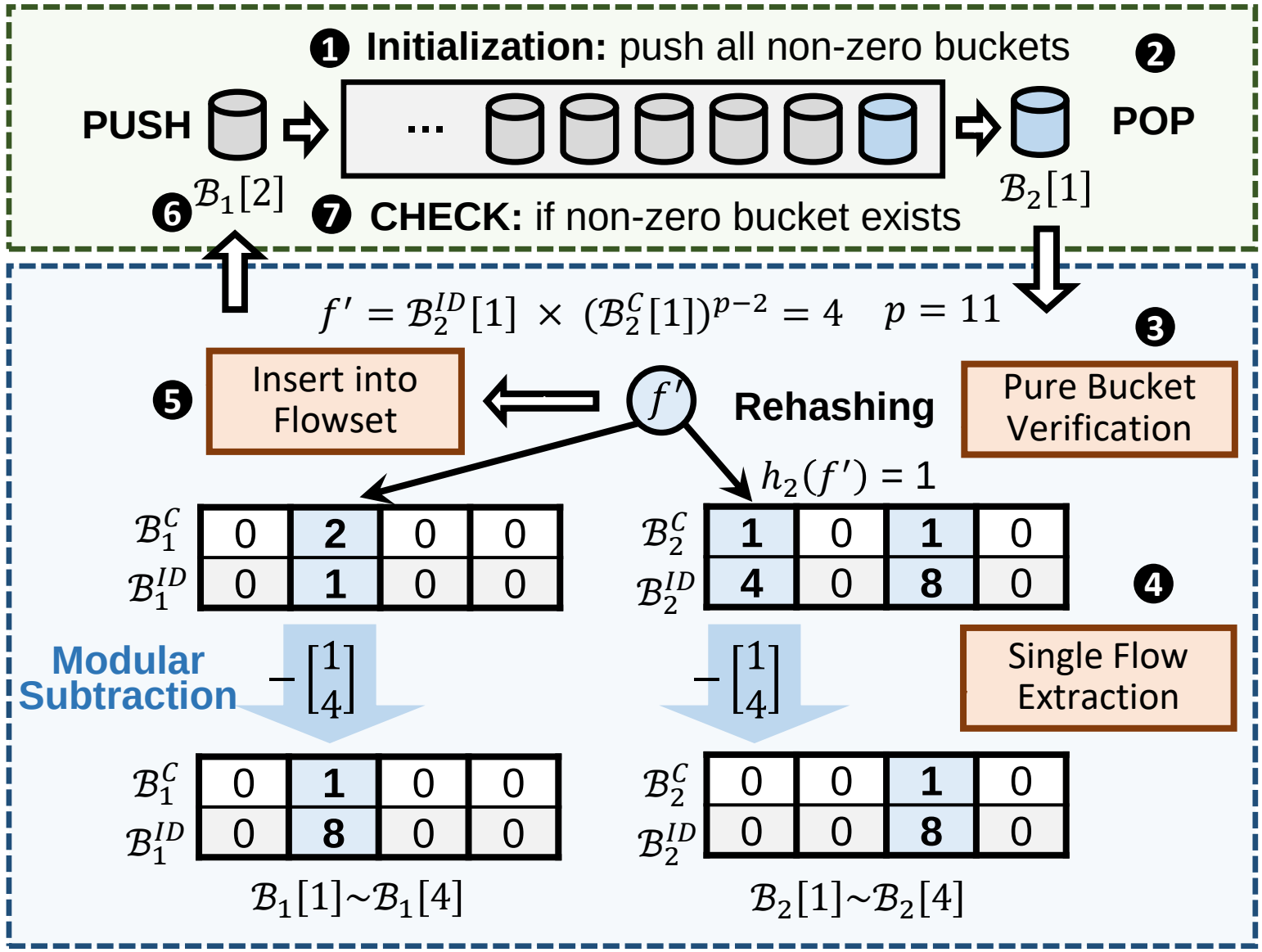
# 费马Sketch/ 单流萃取

北京大学 PEKING UNIVERSITY

- 若桶$\mathcal{B}_i[j]$是纯净桶，提取它记录的唯一流ID以及频数
- ID = $(\mathcal{B}_{\mathrm{i}}^{ID}[j]$ * $\mathcal{B}_{\mathrm{i}}^{c}[j]$ ^(p-2)) mod p
- Freq = $\mathcal{B}_{\mathrm{i}}^{c}[j]$
- 提取完后将该流从费马Sketch中删除

# 费马Sketch/ 解码

- 七步循环
- 纯净桶校验
- 单流萃取
- 无非空桶→解码成功
- 有非空桶→解码失败



# 费马Sketch / 分析

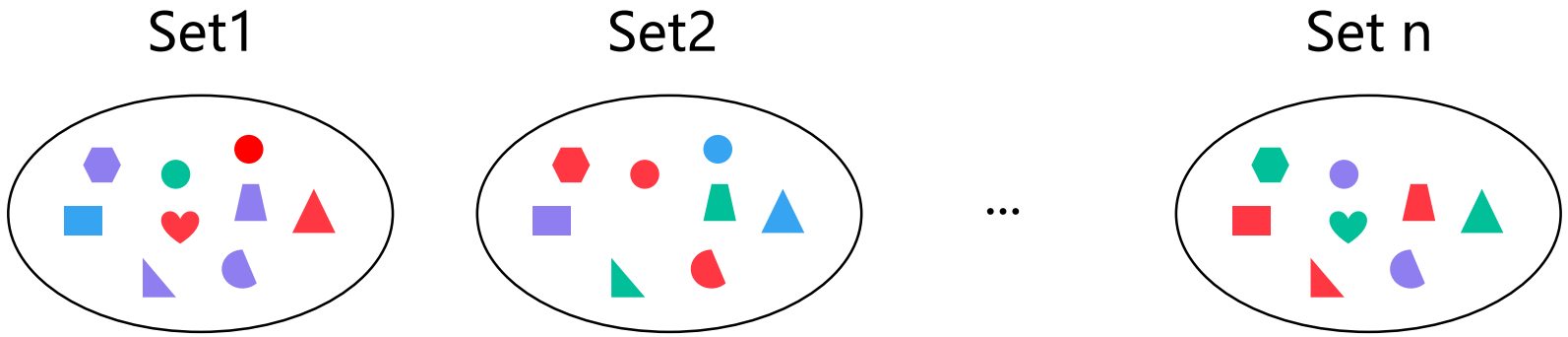
- 优点：
  - 为什么内存低？从流数/丢包数到丢包流数
- 优化方向
  - 更快
  - 更准
- 应用前景
  - PISA架构芯片
  - 华为海思芯片

# 四色过滤器

# 问题定义

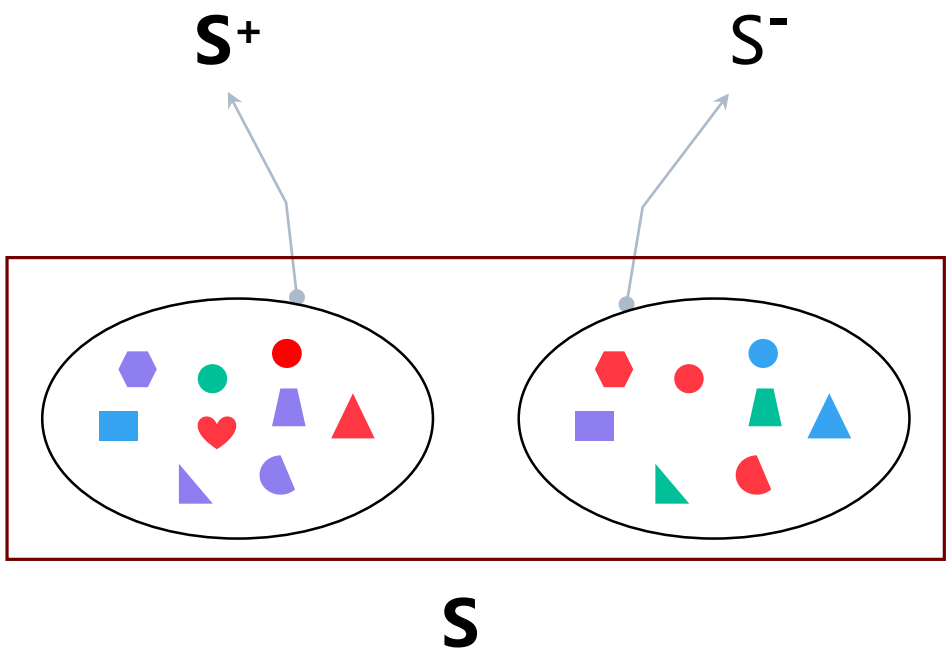
北京大学 PEKING UNIVERSITY

- 问题：多集合查询
  - 多集合查询定义



# 问题简化

- 先考虑最简单的情况：两个集合$S^+$ 和 $S^-$
  - 交集为空
  - 并集为全集 $S^+ \cup S^- = S$

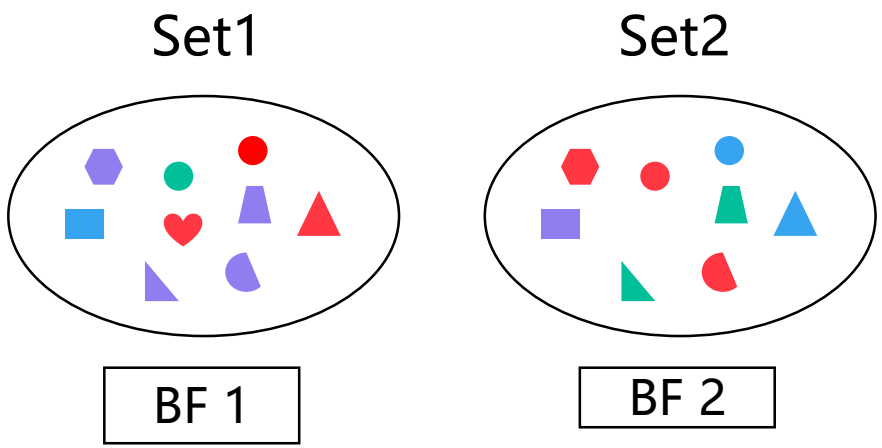



给定全集**S**中的任意一个元素，
回答属于哪个集合？

答案：$S^-$

# 现有方案

- 两种解决方案
  - 哈希表
  - Bloom filter方案




**BF的优点：内存消耗降低了100倍以上**

**目标：再降低10到20倍?**

# 四色猜想与四色定理

北京大学 PEKING UNIVERSITY

- 平面图三染色?
- 四染色历史
  - 1852年首先由一位英国大学生**F.古色利**提出。
  - 1920年弗兰克林证明了**25**个国家四色一定可染。
  - 1926年雷诺兹将国家的数目提高到**27**个。
  - 1936年弗兰克林将国家的数目提高到**31**个。
  - 1968年挪威数学家奥雷证明了提高到**40**个。

规则：

**相邻异色**

# 四色猜想与四色定理

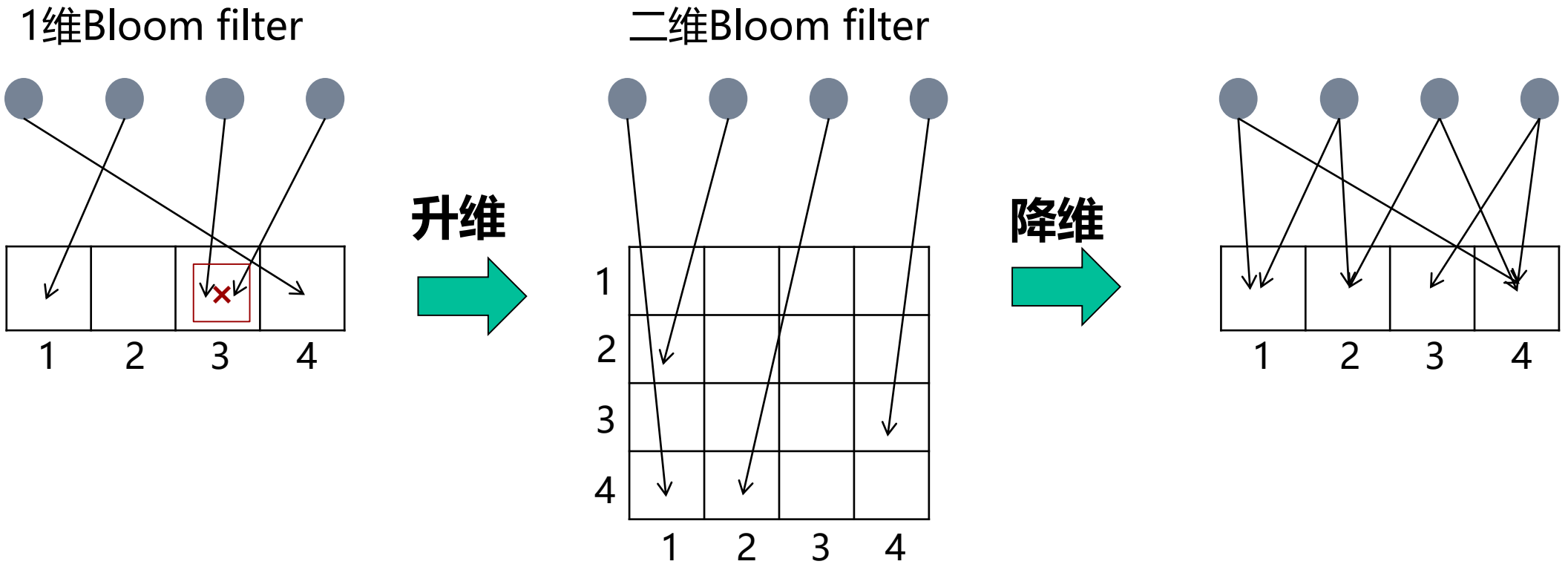
北京大学 PEKING UNIVERSITY

- 四色猜想更名
  - 1976年6月，**哈肯**和**阿佩尔**，3台计算机，证明了四色猜想
    - 基于前人成果
    - 枚举暴力染色
  - 1977年发表时，邮局纪念邮戳（FOUR COLORS SUFFICE）
  - 至今，一些数学家依旧追求纯数学证明方法

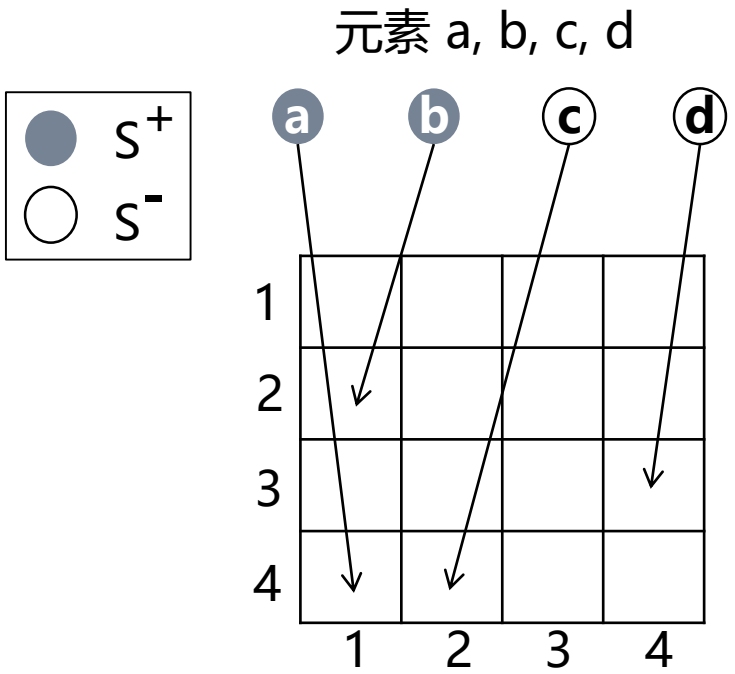
# 四色过滤器 / 定义

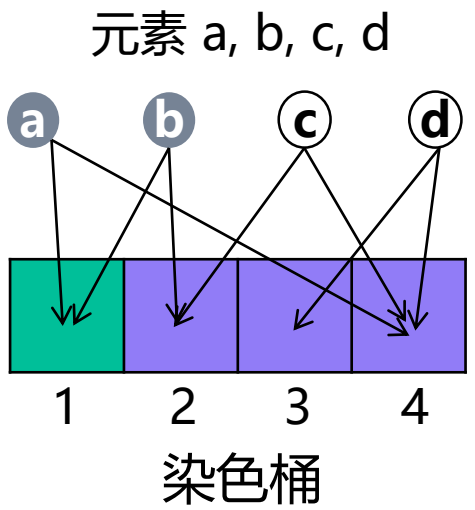
- 一种概率数据结构：多集合查询
- 特点：
  - 采用哈希映射
  - 允许小误差
  - 核心操作过程类似地图四染色
- 染色成功，无错
- 染色失败，出错

# 四色过滤器 / 思想



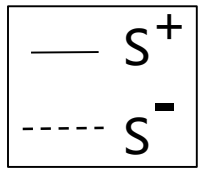
# 四色过滤器 / 如何降维?

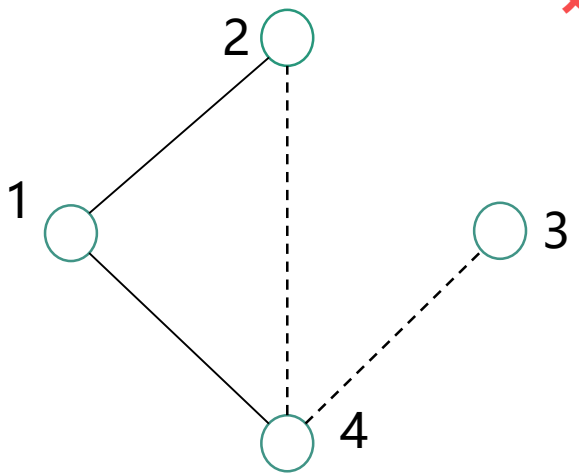


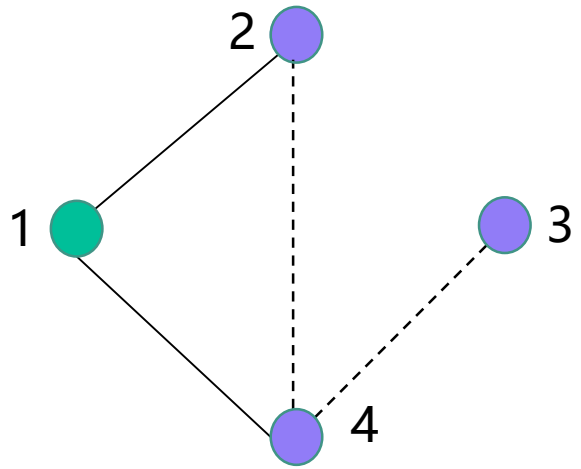

规则:

**正边异色**
**负边同色**




如何染色?

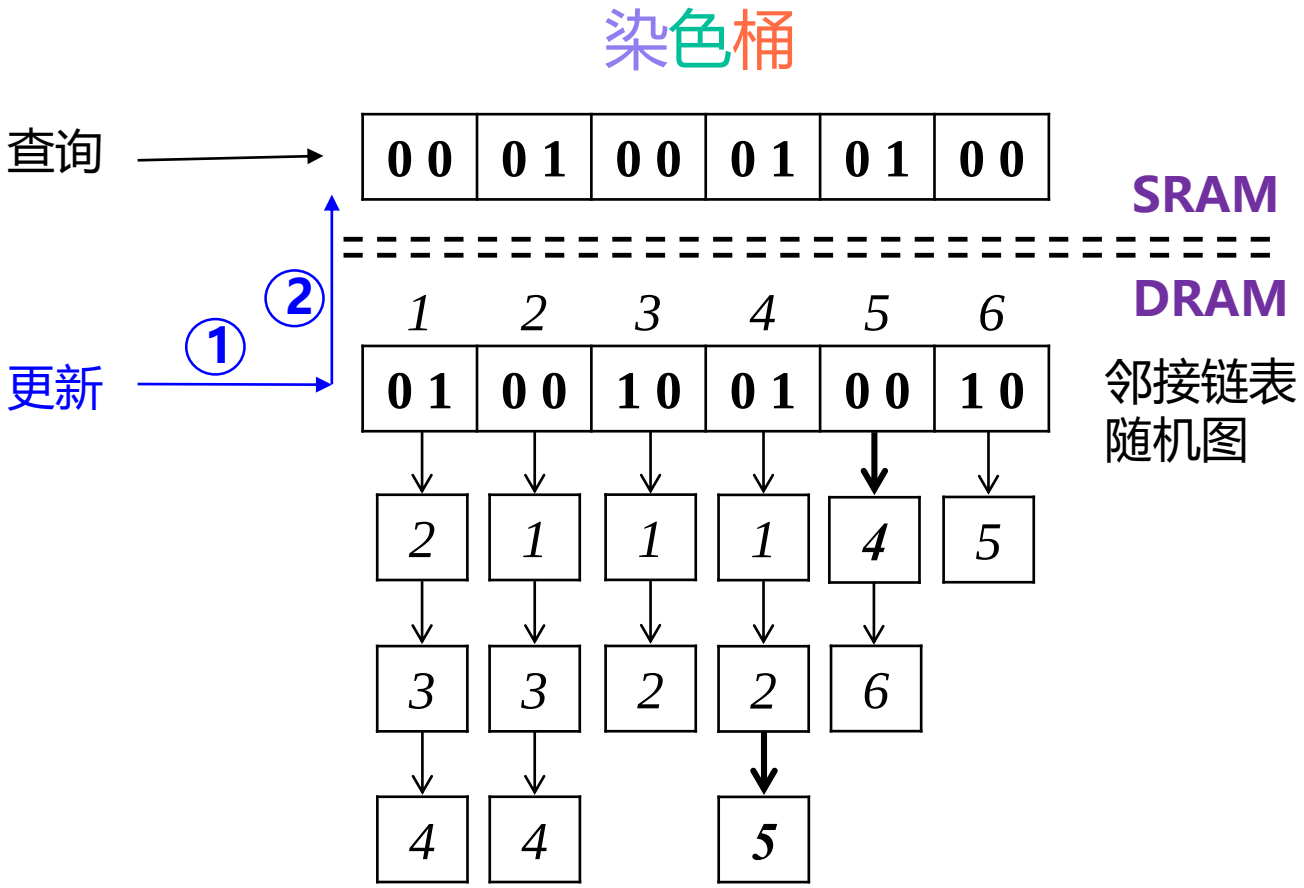


# 四色过滤器 / 数据结构与操作

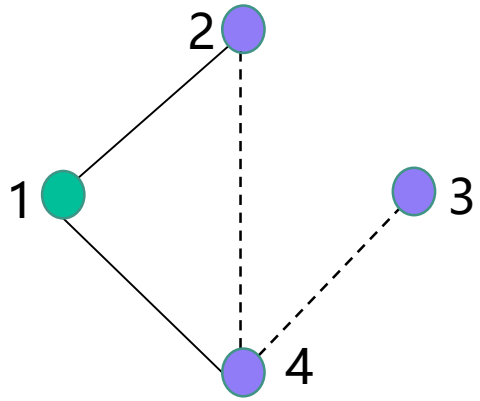
北京大学 PEKING UNIVERSITY

- 数据结构
- 构建Construction
- 插入Insertion

# 四色过滤器 / 数据结构




染色桶数量：m (6)
元素总个数：n
哈希函数　：2

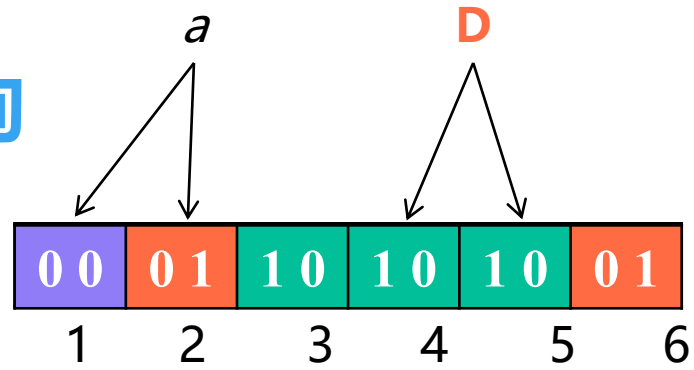


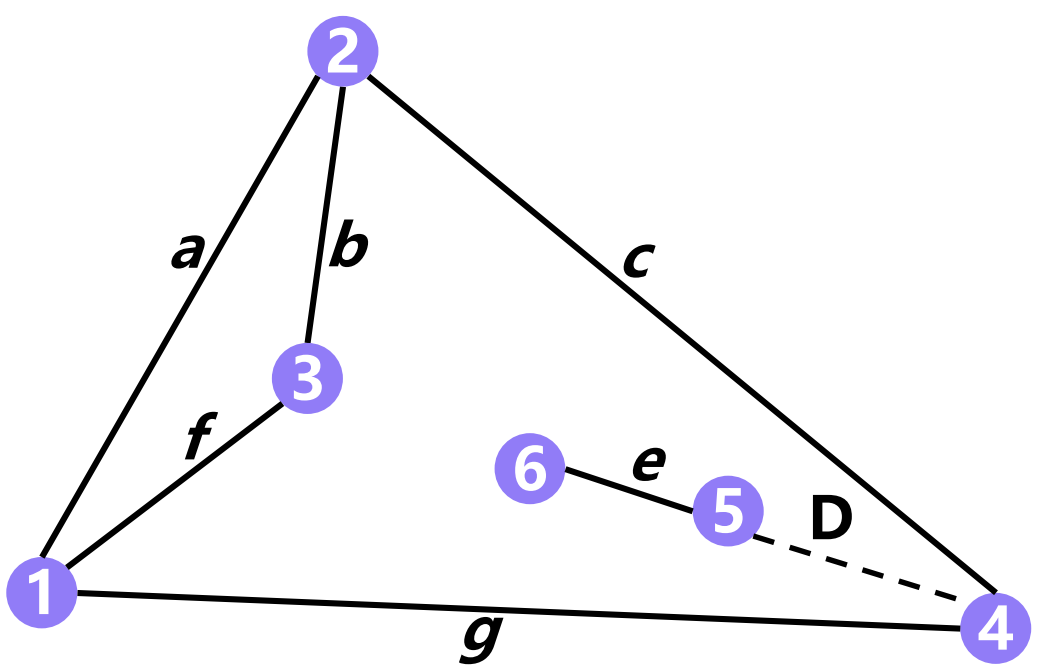
## 构建

正集合：*a,b,c,e,f,g*
负集合：D

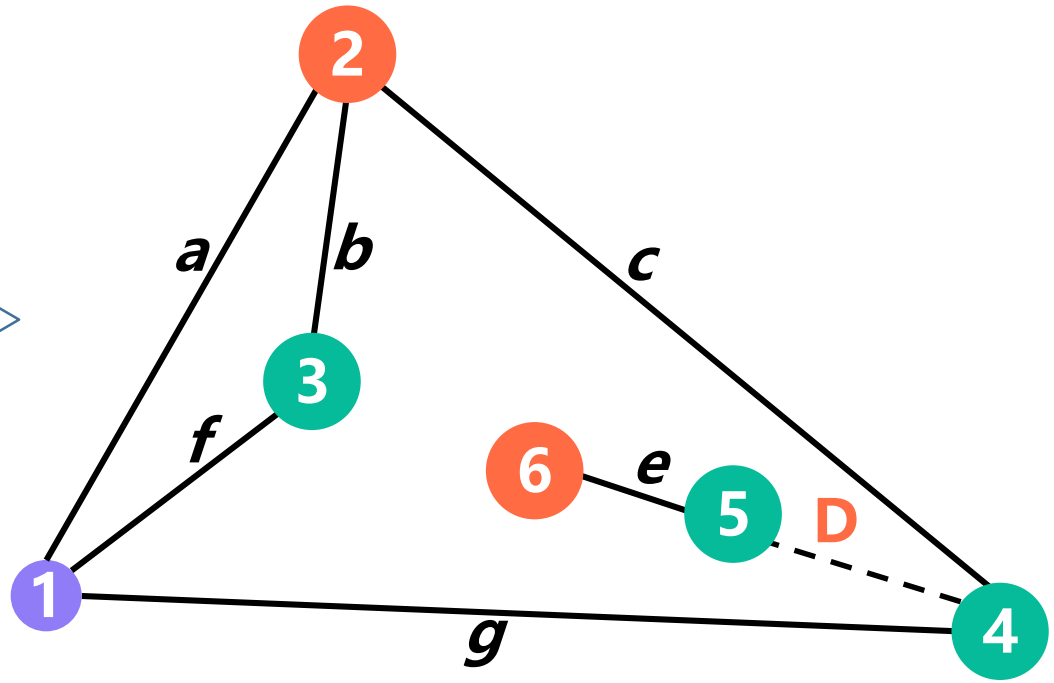
| 0 0 | 0 0 | 0 0 | 0 0 | 0 0 | 0 0 |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |

## 查询



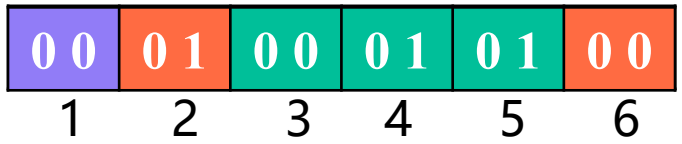




四染色

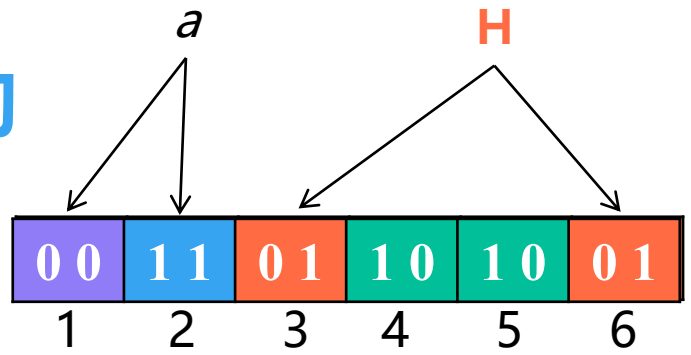


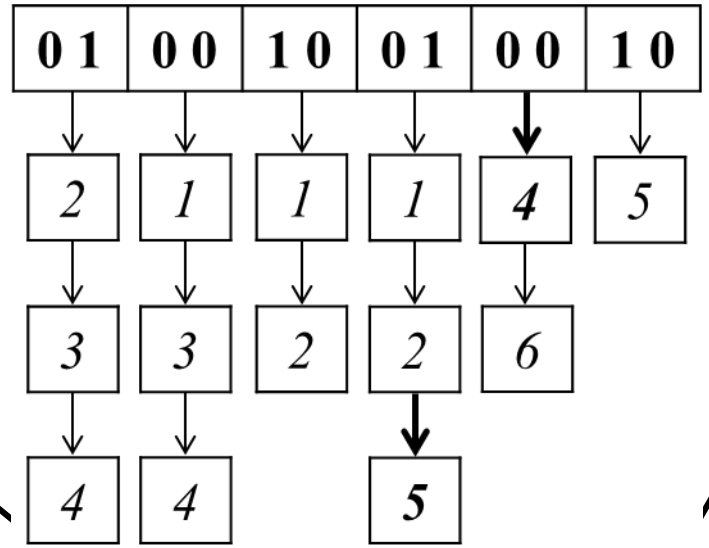
## 插入

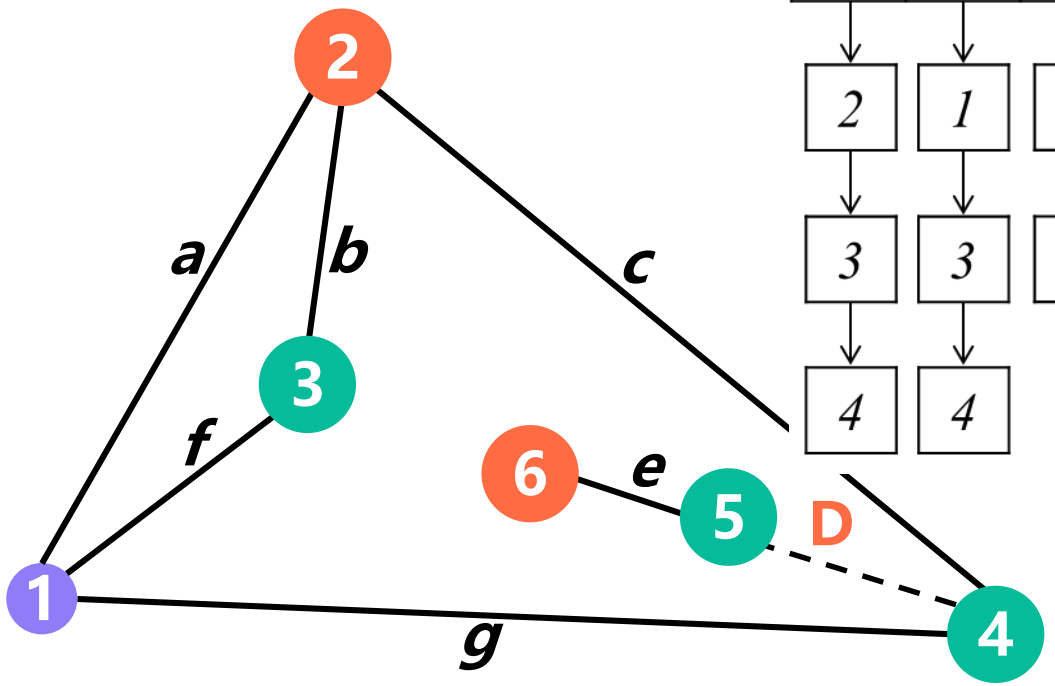
正集合：*a,b,c,e,f,g*
负集合：D，H

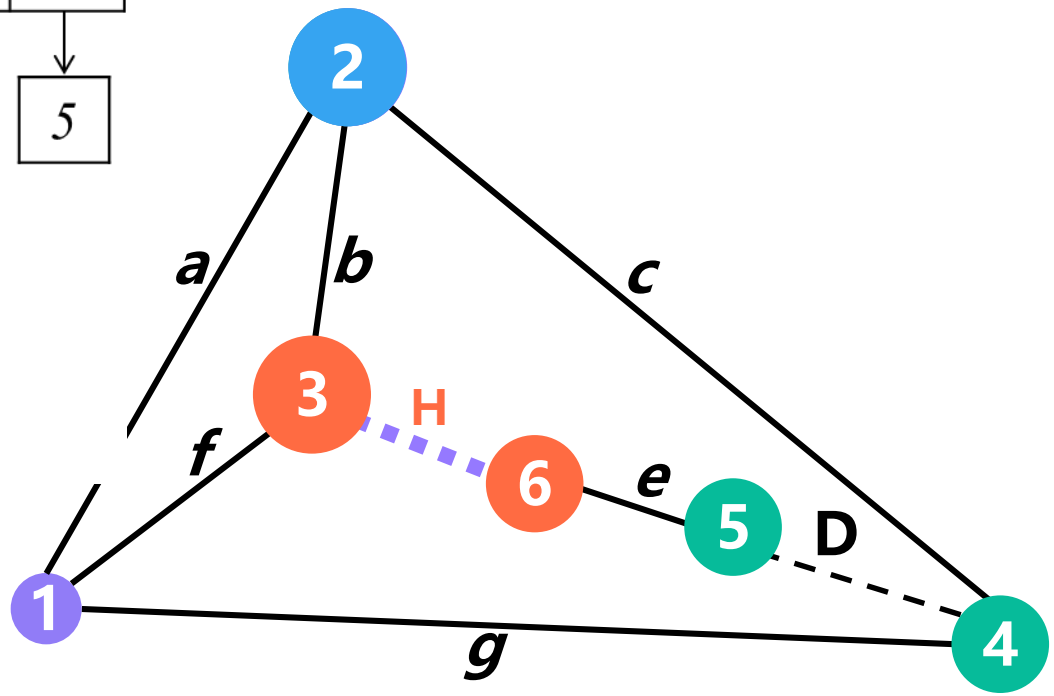



## 查询



# 3 四色过滤器 / 分析

- 优点：
  - 为什么内存低？20比特/元素降低到2.2比特
  - 为什么误差小，降低3个数量级
- 优化方向
  - 更快
  - 更准
- 应用前景
  - CDN分布式缓存
  - 华为海思芯片


杨仝
北京大学
yangtong@pku.edu.cn
https://yangtonghome.github.io/