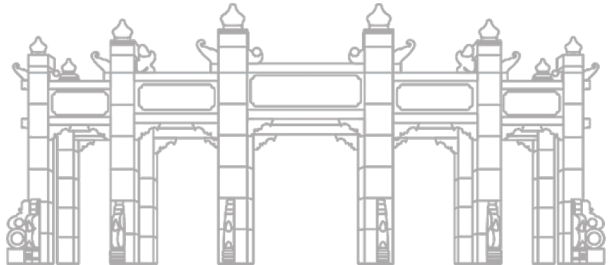
SYSU 2024
PRESENTATION SLIDE

中山大學
SUN YAT-SEN UNIVERSITY

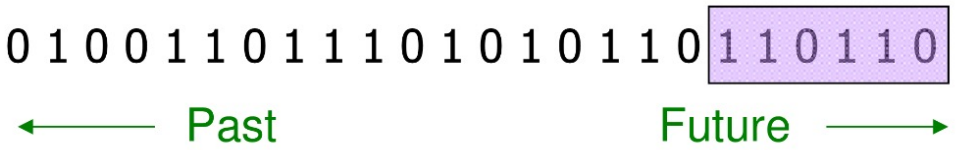
Improvement of DGIM method

林宇浩

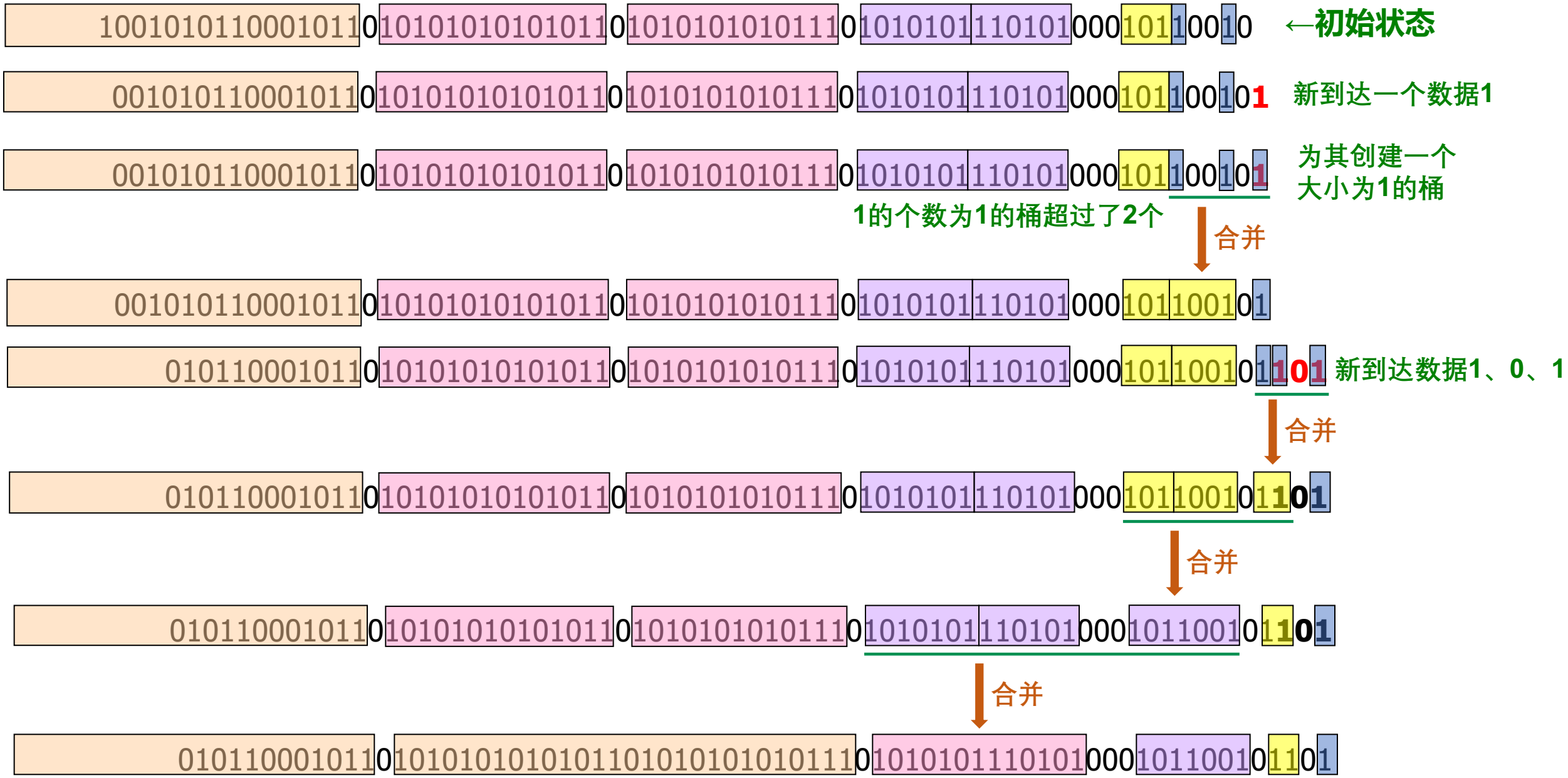


上节课介绍的
两种任务类型

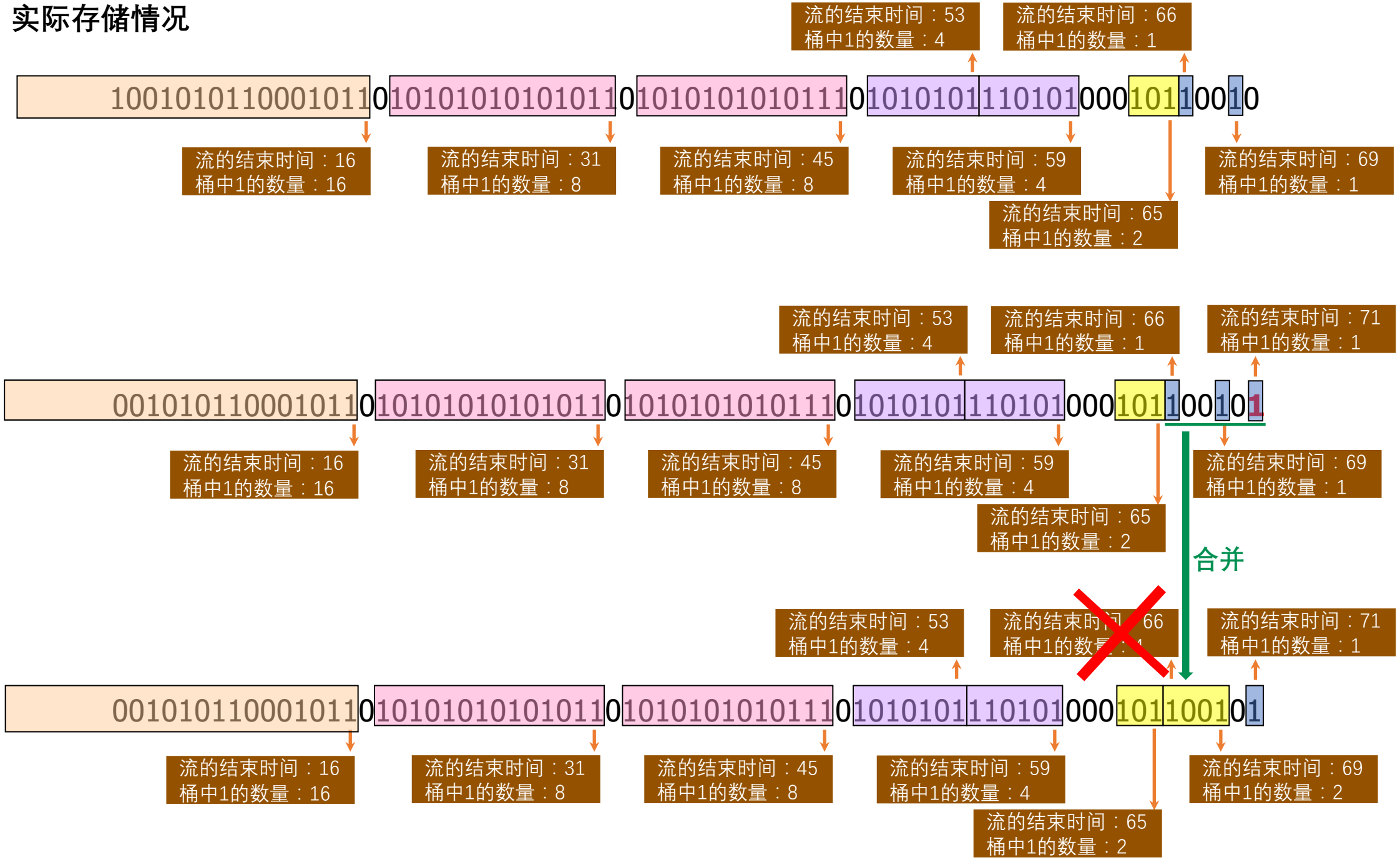
- Sampling data from a stream（从流中采样数据）
- Queries over sliding windows（对滑动窗口查询）
  问题：给定一个0和1的流，最后k位中有多少个1？
  - 暴力算法：
    空间复杂度O(N)
  - 一致性假设：
    无法适应流分布变化的情况
  - DGIM方法



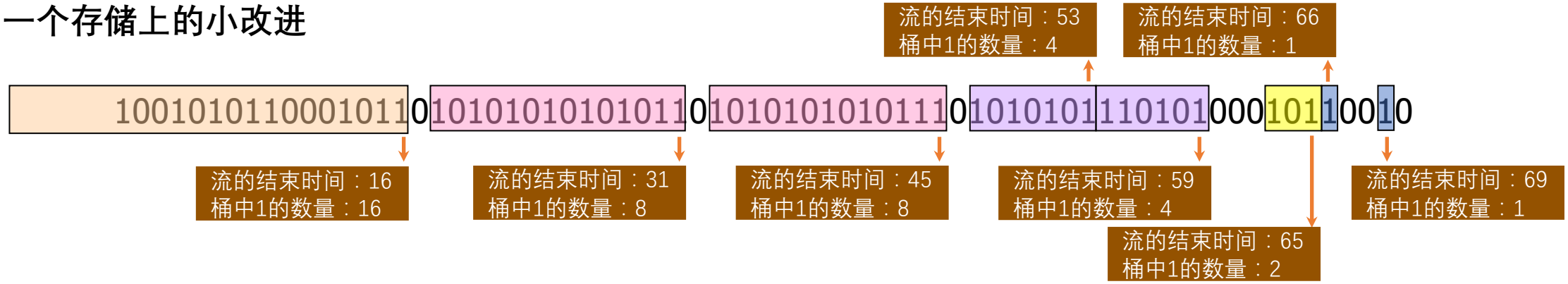
DGIM算法内容回顾

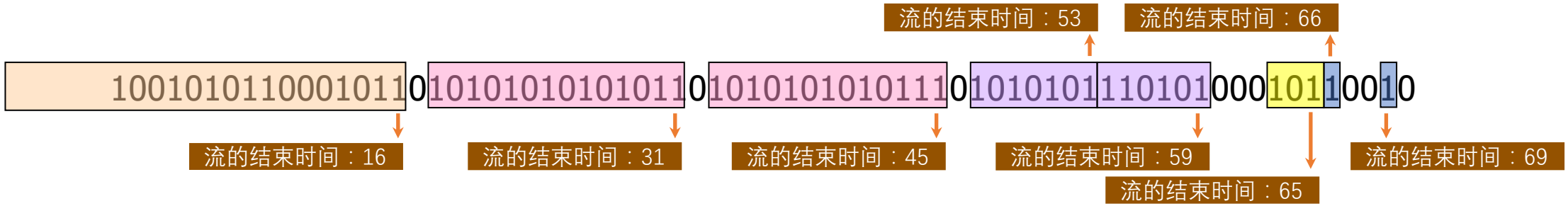

←初始状态
新到达一个数据1
为其创建一个
大小为1的桶
1的个数为1的桶超过了2个
合并
新到达数据1、0、1
合并
合并
合并

实际存储情况


流的结束时间：53
桶中1的数量：4
流的结束时间：66
桶中1的数量：1
1001010110001011 0 1010101010101011 0 1010101010101111 0 1010101 110101 000 101 1 00 1 0
流的结束时间：16
桶中1的数量：16
流的结束时间：31
桶中1的数量：8
流的结束时间：45
桶中1的数量：8
流的结束时间：59
桶中1的数量：4
流的结束时间：65
桶中1的数量：2
流的结束时间：69
桶中1的数量：1
流的结束时间：53
桶中1的数量：4
流的结束时间：66
桶中1的数量：1
流的结束时间：71
桶中1的数量：1
0010101100010110 1010101010101011 0 1010101010101111 0 1010101 110101 000 101 1 00 1 0 1
流的结束时间：16
桶中1的数量：16
流的结束时间：31
桶中1的数量：8
流的结束时间：45
桶中1的数量：8
流的结束时间：59
桶中1的数量：4
流的结束时间：65
桶中1的数量：2
流的结束时间：69
桶中1的数量：1
合并
流的结束时间：53
桶中1的数量：4
流的结束时间：71
桶中1的数量：1
0010101100010110 1010101010101011 0 1010101010101111 0 1010101 110101 000 101 1001 0 1
流的结束时间：16
桶中1的数量：16
流的结束时间：31
桶中1的数量：8
流的结束时间：45
桶中1的数量：8
流的结束时间：59
桶中1的数量：4
流的结束时间：65
桶中1的数量：2
流的结束时间：69
桶中1的数量：2

一个存储上的小改进





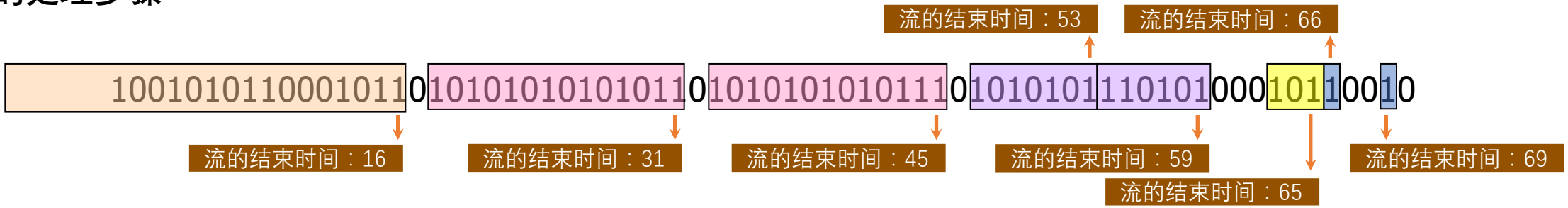
桶中1的数量	…	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	0	1	1	0	1

桶中1数量为1的桶有2个

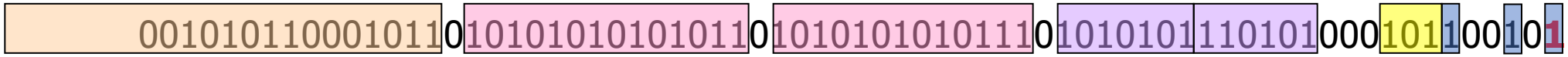
桶中1数量为2的桶有1个

桶中1数量为4的桶有2个

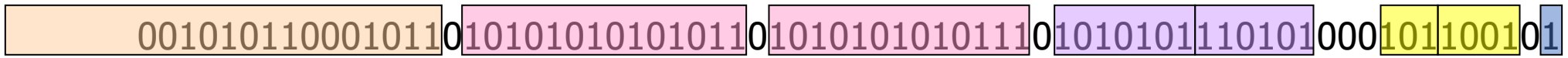
改进后的处理步骤



桶中1的数量	…	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	0	1	1	0	1



桶中1的数量	…	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	0	1	1	0	1
						+1
	…	0	1	1	1	0



改进后可以利用二进制加法代替原先的多次合并

00101011000101101010101010110101010101011101010101110101000101100101

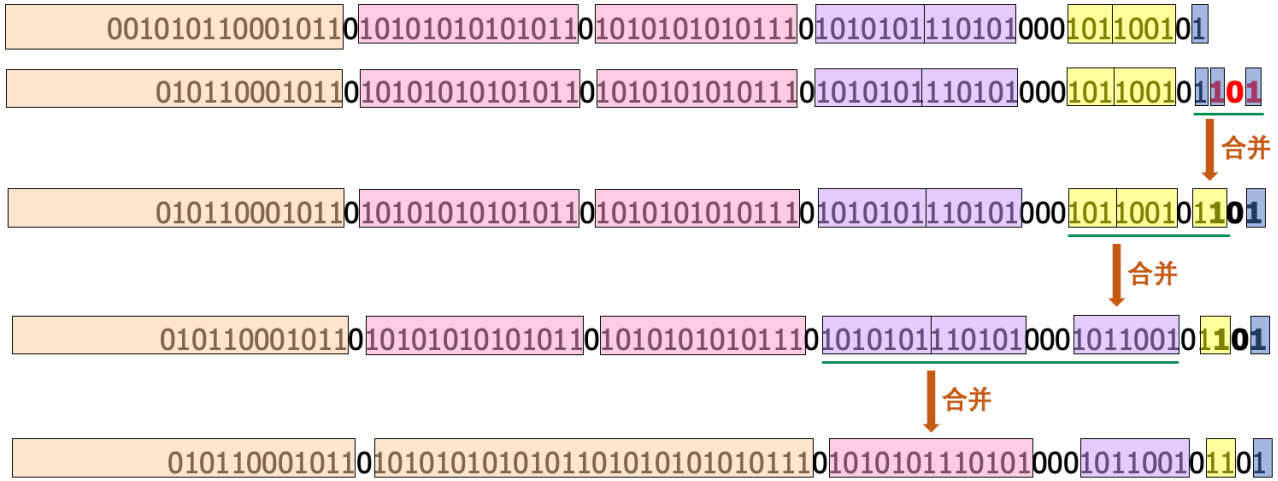
桶中1的数量	…	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	0	1	1	1	0

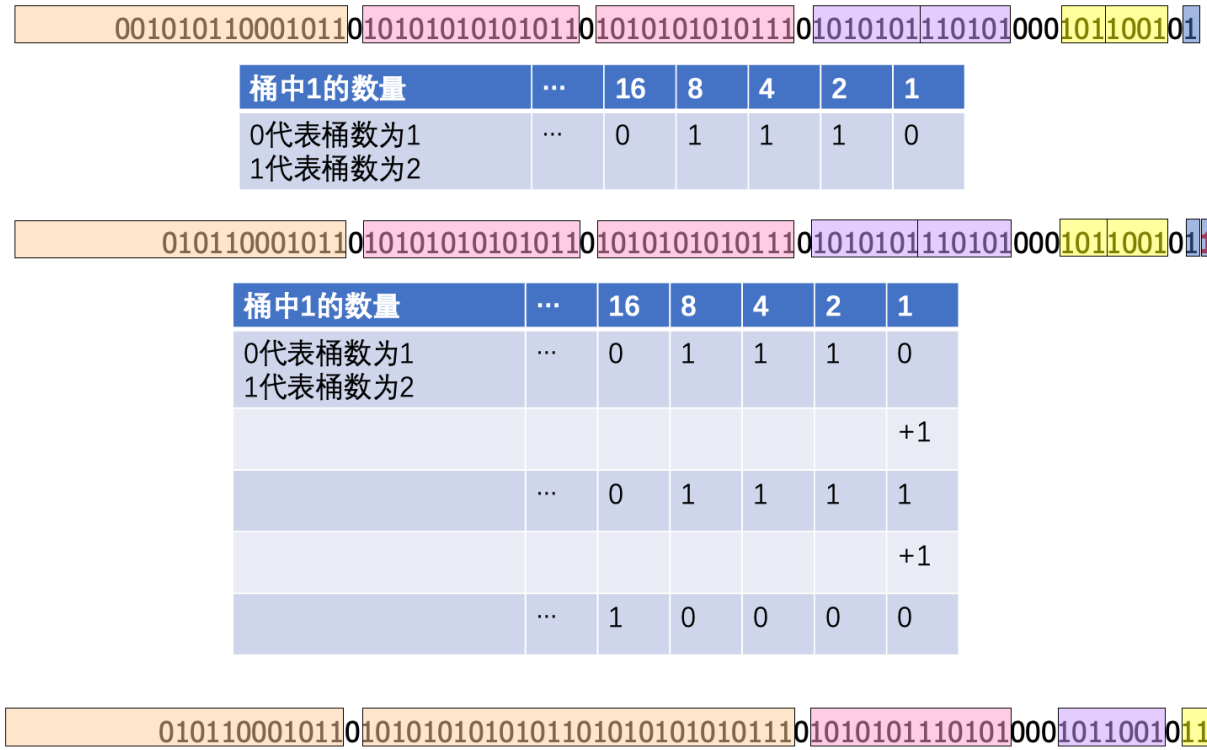
01011000101101010101010110101010101011101010101110101000101100101101

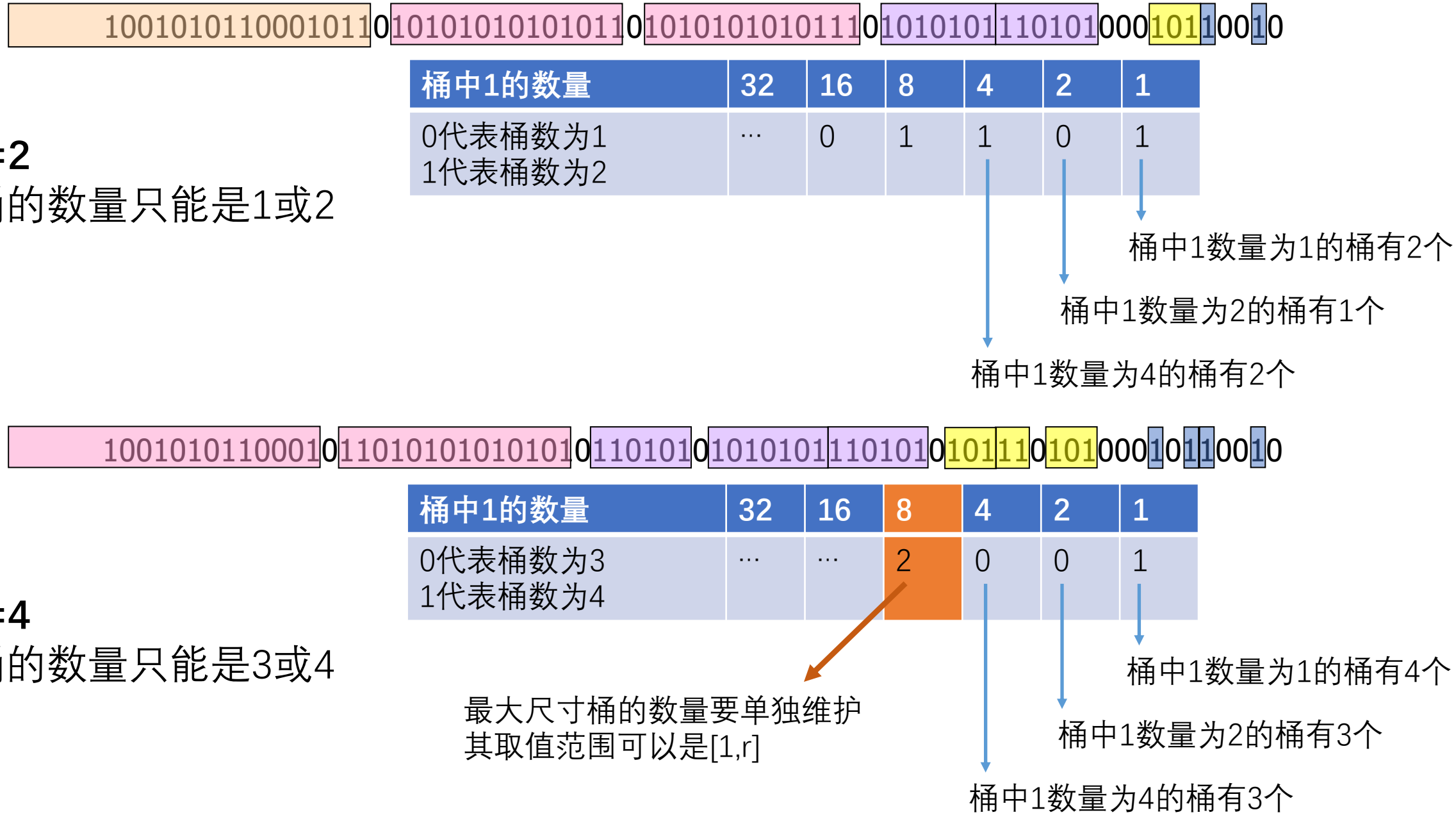
桶中1的数量	…	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	0	1	1	1	0
						+1
	…	0	1	1	1	1
						+1
	…	1	0	0	0	0

01011000101101010101010110101010101011101010101110101000101100101101

改进后可以利用二进制加法代替原先的多次合并



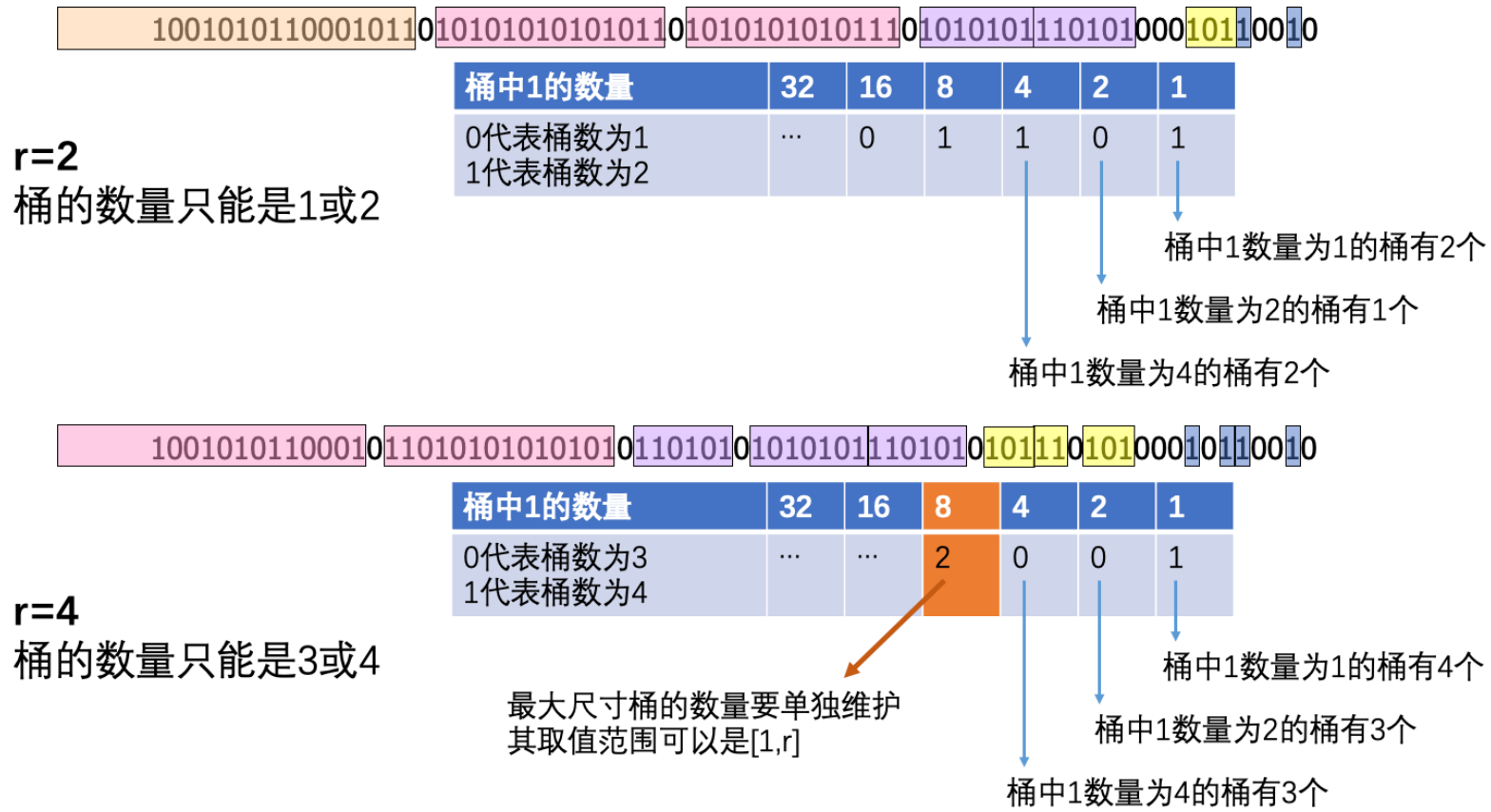



1001010110001011010101010101011010101010101110101010110101000101100010
桶中1的数量
32
16
8
4
2
1
0代表桶数为1
1代表桶数为2
…
0
1
1
0
1
桶中1数量为1的桶有2个
桶中1数量为2的桶有1个
桶中1数量为4的桶有2个
r=2
桶的数量只能是1或2
1001010110001011010101010101011010101010101110101010110101000101100010
桶中1的数量
32
16
8
4
2
1
0代表桶数为3
1代表桶数为4
…
…
2
0
0
1
最大尺寸桶的数量要单独维护
其取值范围可以是[1,r]
桶中1数量为1的桶有4个
桶中1数量为2的桶有3个
桶中1数量为4的桶有3个
r=4
桶的数量只能是3或4

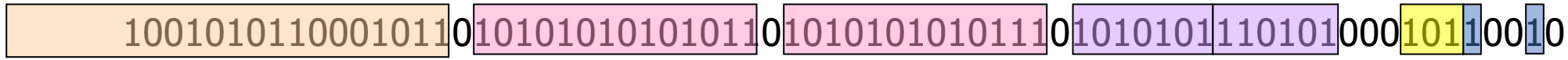
Further Reducing the Error

- Instead of maintaining **1** or **2** of each size bucket, we allow either ***r*-1** or ***r*** buckets (***r* > 2**)
  - Except for the largest size buckets; we can have any number between **1** and ***r*** of those
- **Error is at most *O*(1/*r*)**
- By picking ***r*** appropriately, we can tradeoff between number of bits we store and the error

	r值	错误率	桶的数量	存储桶中1的数量要用的空间	存储时间戳要用的空间	更新消耗
DGIM method	↑	↓	↑	↑	↑	↑
改进后	↑	↓	↑	↓	↑	≈



计算窗口内1的数量的估计值

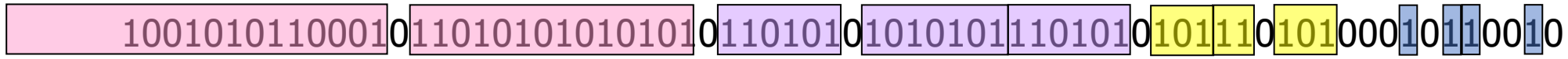


r=2

桶的数量只能是1或2

桶中1的数量	32	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	0	1	1	0	1
桶中1的数量		1×2^4	2×2^3	2×2^2	1×2^1	2×2^0

$$1\text{的总数} = [\frac{1}{2}\times1\times2^4 + (1-1)\times2^4\,] + 2\times2^3 + 2\times2^2 + 1\times2^1 + 2\times2^0$$



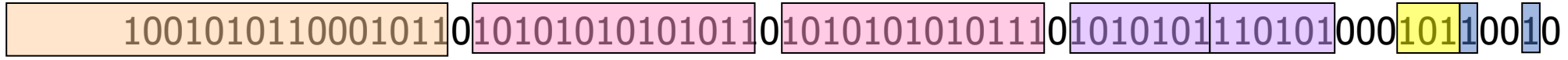
r=4

桶的数量只能是3或4

桶中1的数量	32	16	8	4	2	1
0代表桶数为3 1代表桶数为4	…	…	2	0	0	1
			2×2^3	3×2^2	3×2^1	4×2^0

$$1\text{的总数} = [\frac{1}{2}\times2^3 + (2-1)\times2^3\,] + 3\times2^2 + 3\times2^1 + 4\times2^0$$

对于r=2的情况还能在计算上优化



r=2
桶的数量只能是1或2

桶中1的数量	32	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	0	1	1	0	1
桶中1的数量		1×2^4	2×2^3	2×2^2	1×2^1	2×2^0

数值为0都会乘以1
数值为1都会乘以2
最高位会有一个桶乘以0.5

$$1\text{的总数} = [\frac{1}{2}\times1\times2^4 + (1-1)\times2^4\,] + 2\times2^3 + 2\times2^2 + 1\times2^1 + 2\times2^0 = 36$$

处理所有数值为0的

32	16	8	4	2	1
…	0	1	1	0	1

逻辑非

32	16	8	4	2	1
0	1	0	0	1	0

处理所有数值为1的

32	16	8	4	2	1
…	0	1	1	0	1

左移1位

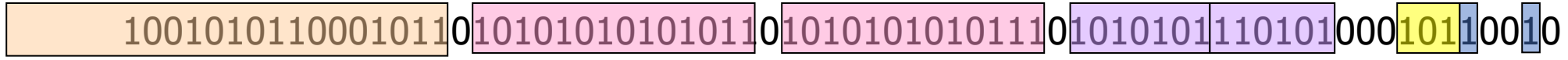
32	16	8	4	2	1
0	1	1	0	1	0

	32	16	8	4	2	1
	0	1	0	0	1	0
+	0	1	1	0	1	0
	1	0	1	1	0	0

	32	16	8	4	2	1
	1	0	1	1	0	0
-	0	0	1	0	0	0
	1	0	0	1	0	0

最左侧的桶乘以0.5

对于r=2的情况还能在计算上优化



r=2
桶的数量只能是1或2

桶中1的数量	32	16	8	4	2	1
0代表桶数为1 1代表桶数为2	…	1	1	1	0	1
桶中1的数量		2×2^4	2×2^3	2×2^2	1×2^1	2×2^0

数值为0都会乘以1
数值为1都会乘以2
最高位会有一个桶乘以0.5

$$1\text{的总数} = [\frac{1}{2}\times1\times2^4 + (2-1)\times2^4\,] + 2\times2^3 + 2\times2^2 + 1\times2^1 + 2\times2^0 = 52$$

处理所有数值为0的

32	16	8	4	2	1
…	1	1	1	0	1

逻辑非

32	16	8	4	2	1
0	0	0	0	1	0

处理所有数值为1的

32	16	8	4	2	1
…	1	1	1	0	1

左移1位

32	16	8	4	2	1
1	1	1	0	1	0

	32	16	8	4	2	1
	0	0	0	0	1	0
+	1	1	1	0	1	0
	1	1	1	1	0	0

	32	16	8	4	2	1
	1	1	1	1	0	0
-	0	0	1	0	0	0
	1	1	0	1	0	0

最左侧的桶乘以0.5

中山大學
SUN YAT-SEN UNIVERSITY

谢谢观看