Cs 5050

04 08 2014

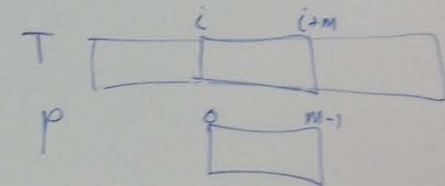
String matching

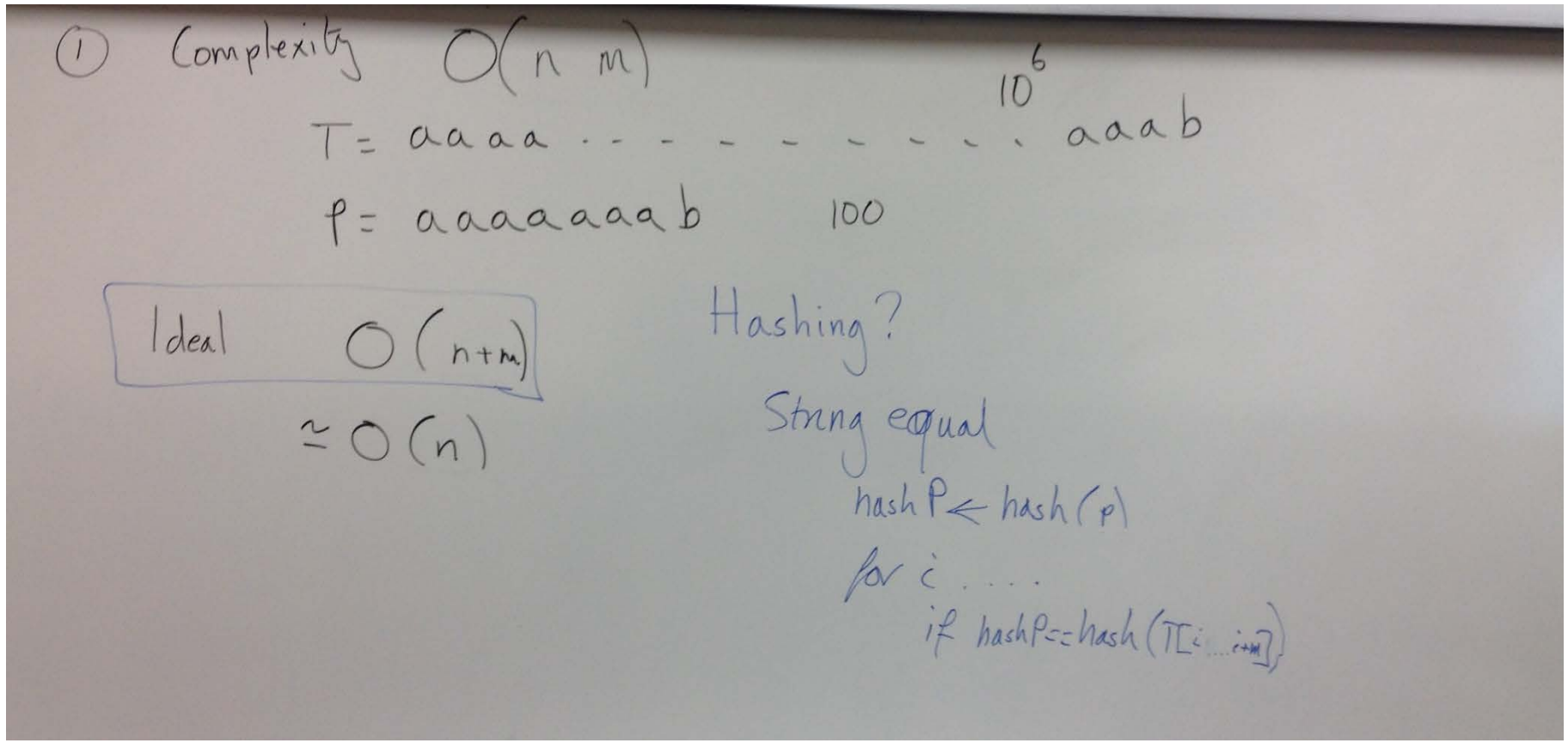
Given P small size m P[]

T text target big size n T[]

Find if P[0] == T[i] && P[1] == T[i+1] && ....

```
find =
for i=0 to n-1-m
    found = True
    for j=0 to m-1
        found = found && P[j] == T[i+j]
    end
end
```




① Complexity O(n m)
10^6
T = aaaa . . . . . aaab
P = aaaaaaaab
100
Ideal O(n+m)
≃ O(n)
Hashing?
String equal
hashP ← hash(P)
for i . . . .
if hashP == hash(T[i . . . i+m])

② hash(S)

$hash = base^{0} * atoi(S[0]) + middle\ part + base^{(m-1)} * atoi(S[m-1])$

$$hash = \left(\sum_{j=0}^{j=m-1} \underbrace{base^{j}}_{\text{prime \#}} \times \underbrace{atoi(S[j])}_{\text{byte } 0 \ldots 255} \right) \% \underbrace{\text{Array size}}_{\text{max int}}$$

S string

101

Principles

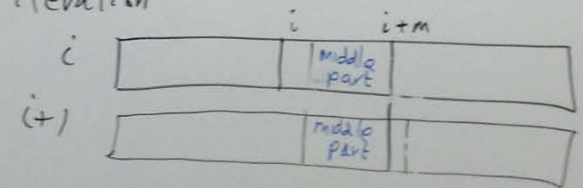
idea

"Compute once use two times"

incrementally update hash value

use for string matching

iteration



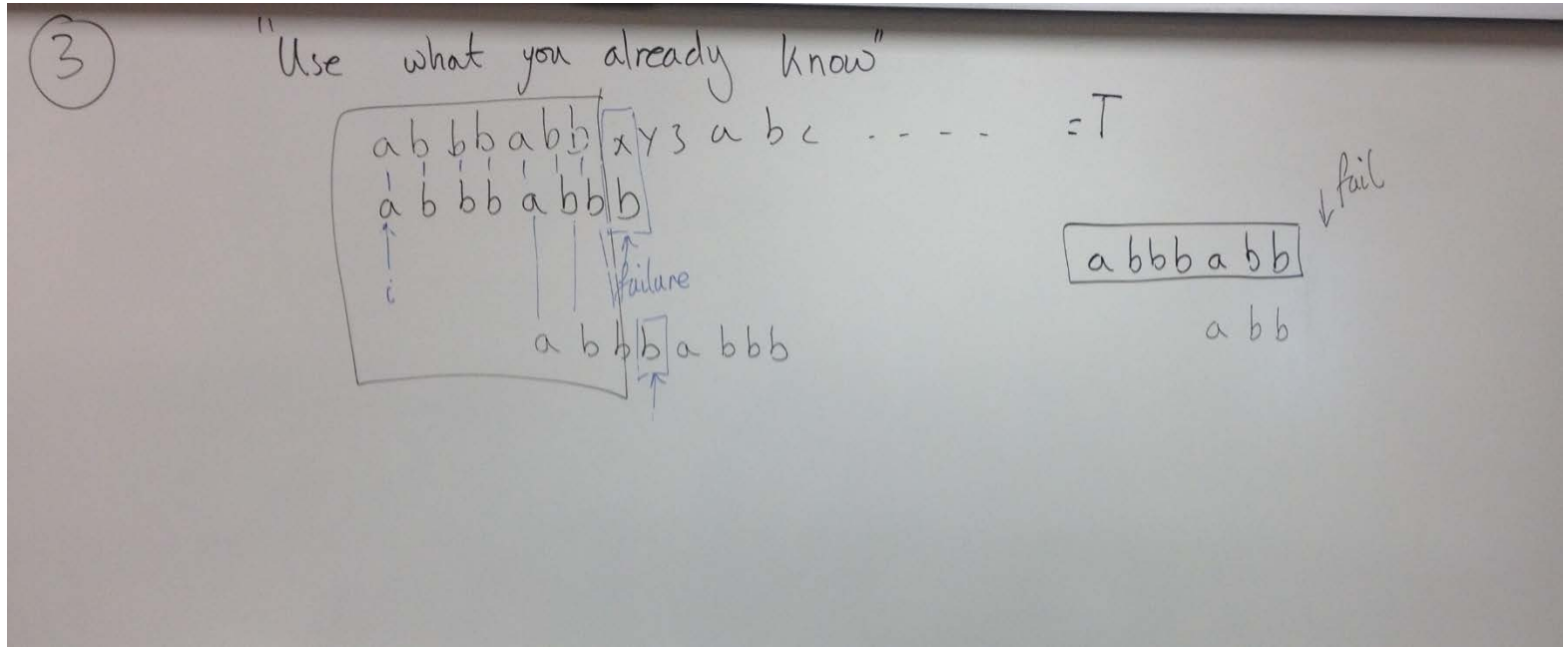
T

subtract $atoi(T[i]) * base^{(m-1)}$

multiply by base

add $atoi(T[i+m])$

roling hash


3
"Use what you already know"
a b bb a bb x y 3 a b c ---- =T
a b bb a bb b
i
failure
a b b b a b b b
a bbb a bb
fail
a bb