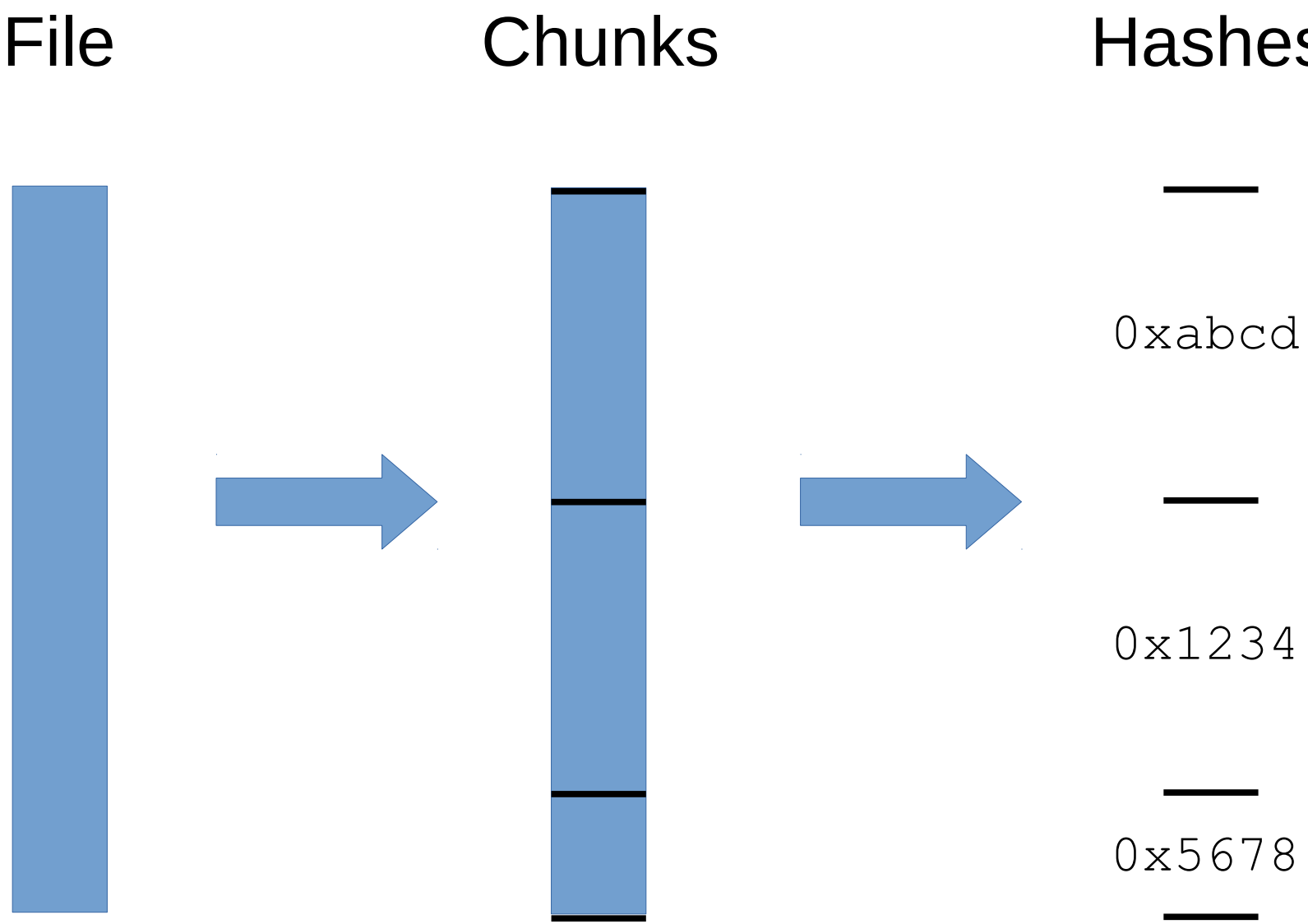
Accordion

Multi-Scale Recipes for Adaptive Detection of Duplication

Russell Lewis
John H. Hartman

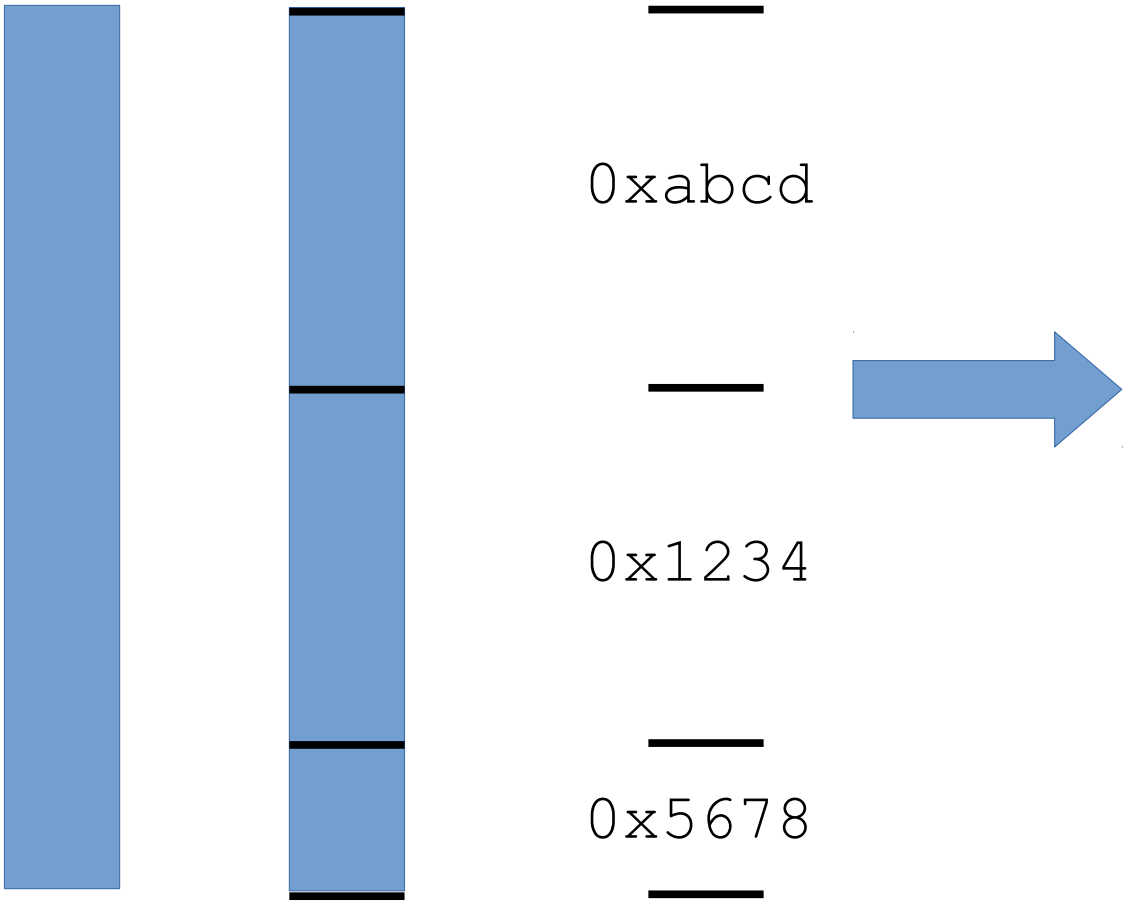
University of Arizona

File Chunking



Recipes

File Chunks Hashes Recipe



$(pos_1, len_1, 0xabcd)$
$(pos_2, len_2, 0x1234)$
$(pos_3, len_3, 0x5678)$

Finding Duplication

- Compare hashes
  - File-to-file
  - File-to-library

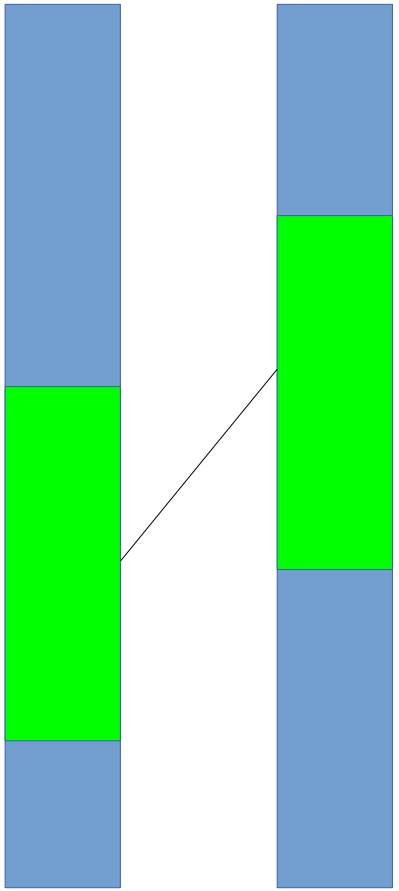
rsync

content-addressable storage

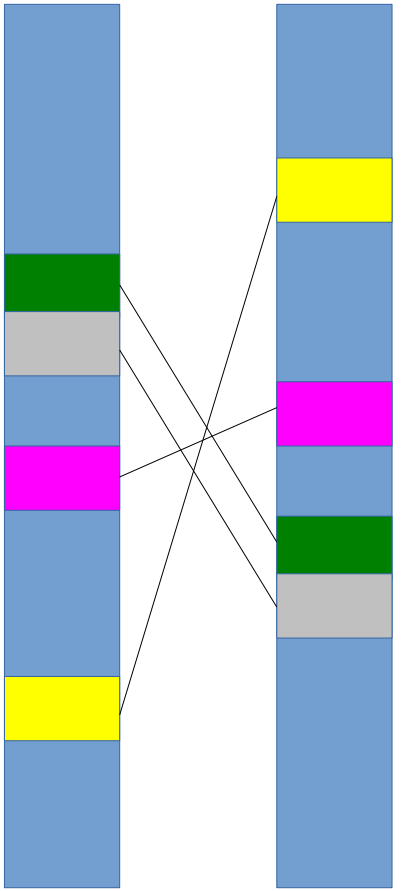
dedup appliances

Recipe Scale

Large scale

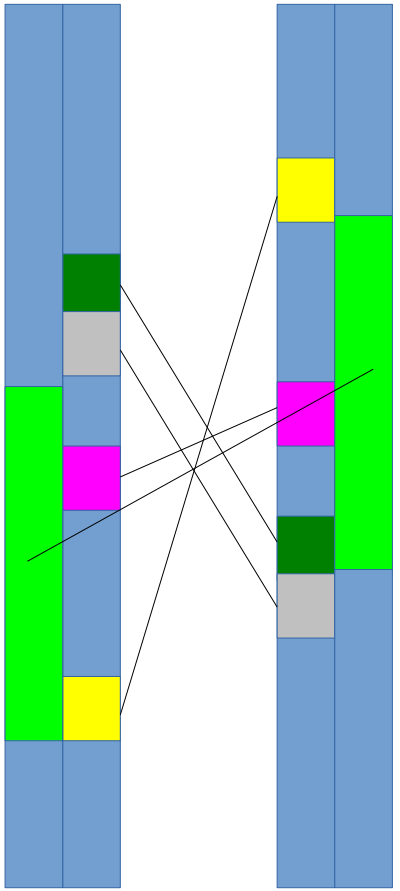


Small scale



Multi-Scale

Union of single-scale recipes



Managing Cost

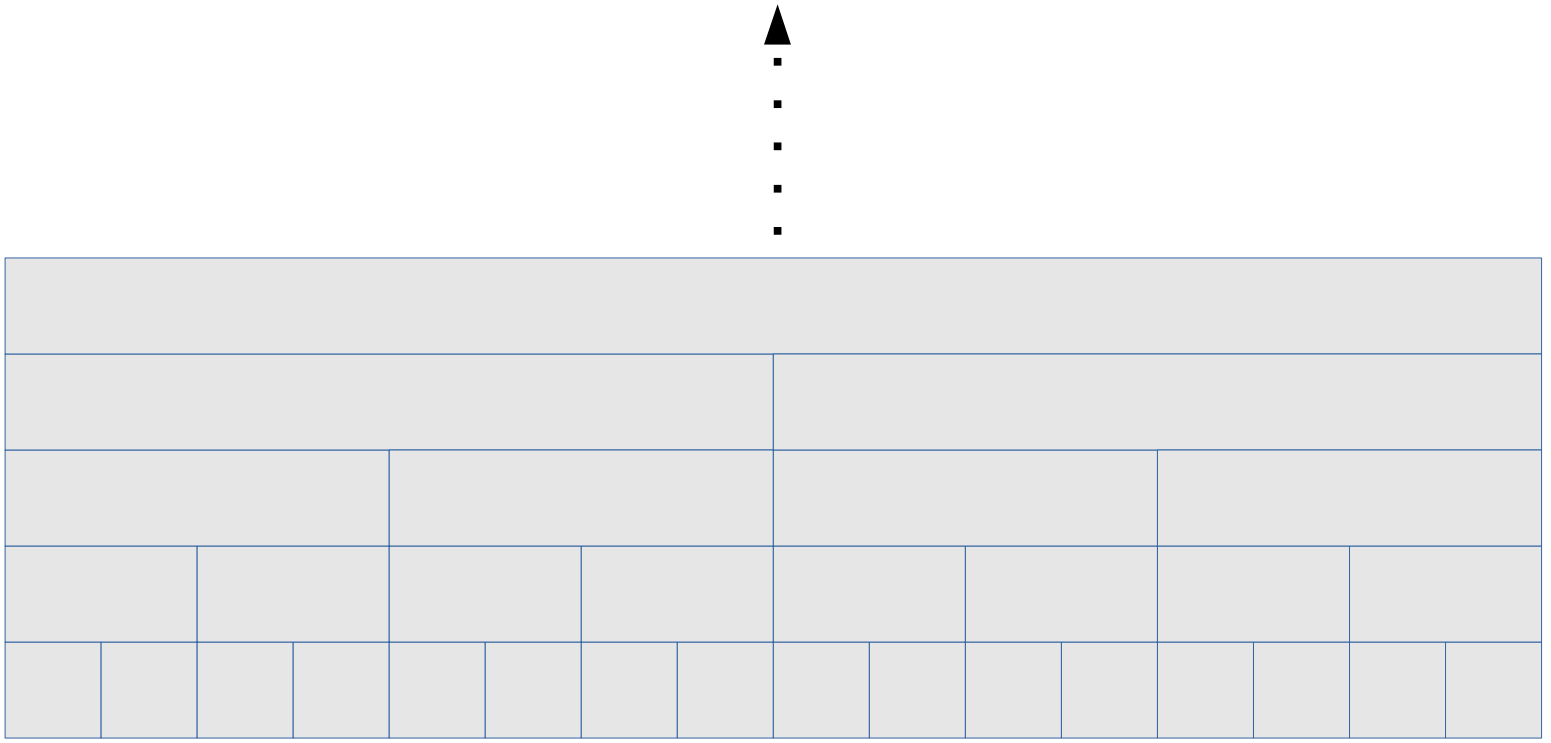
- Multi-scale recipes larger than single-scale
- Need to skip entries to manage cost

Algorithm 1 (top-down)

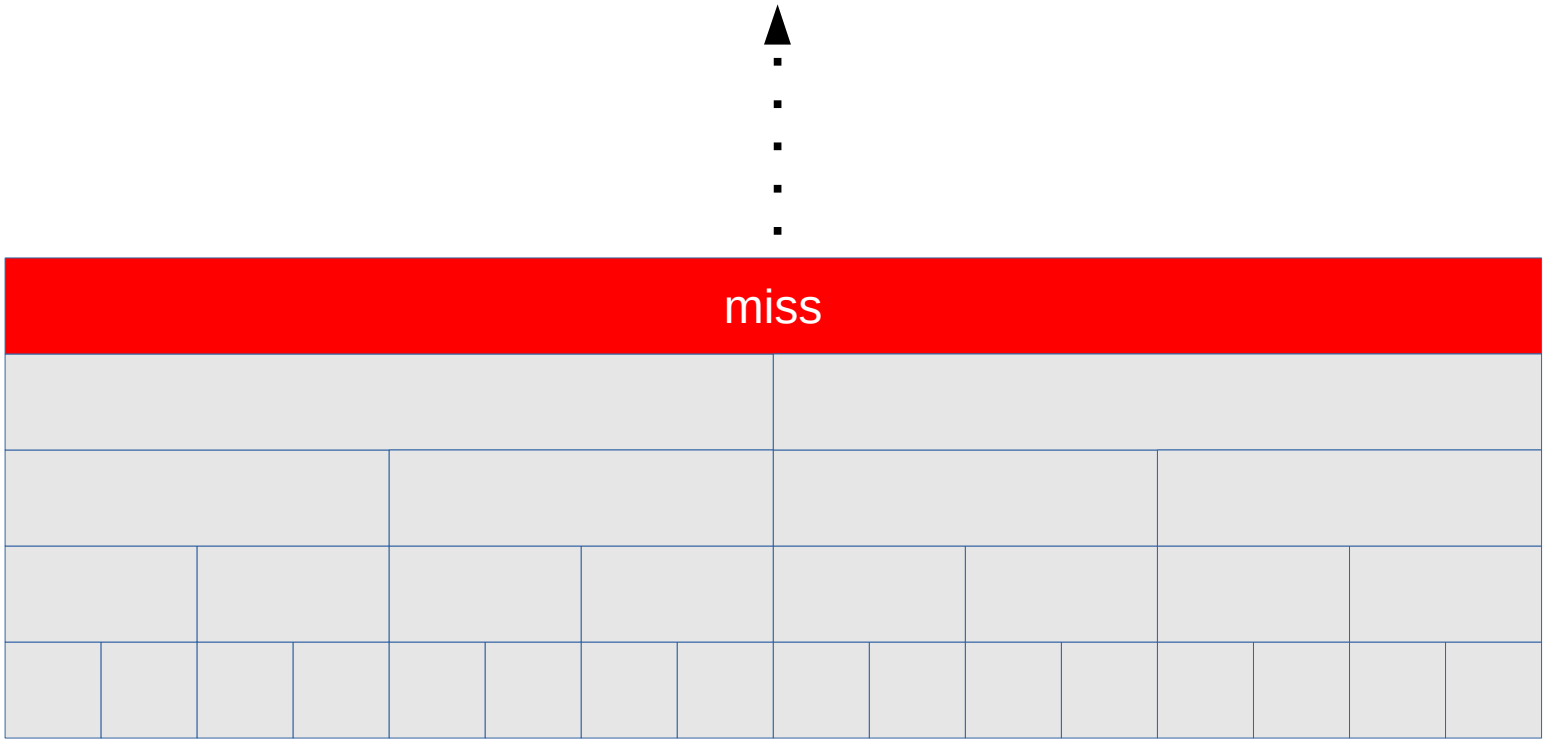
- Recipe Sort:
  - size (descending)
  - position (ascending)

- Best case: whole-file match
- Worst case: 2x single-scale

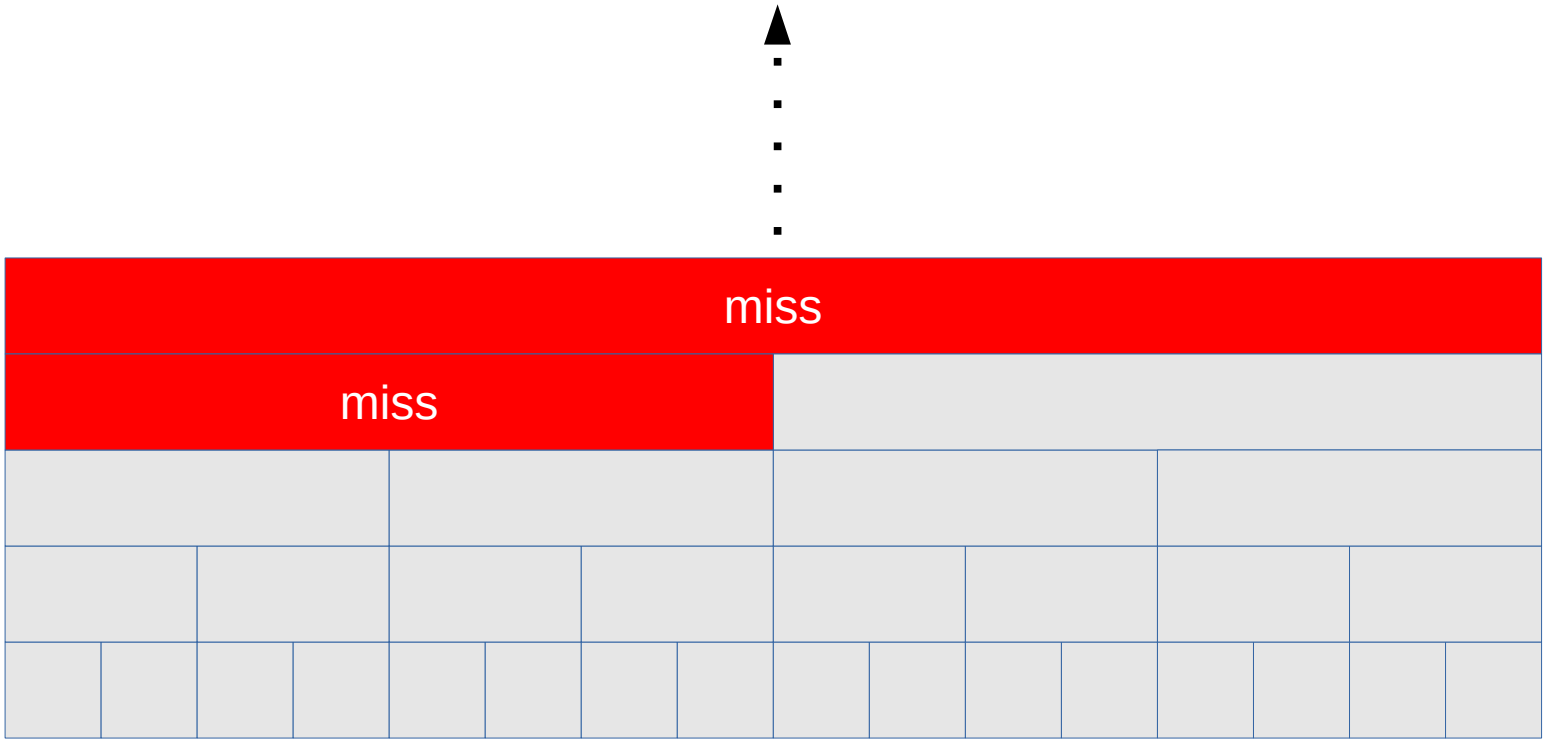
Algorithm 1 (top-down)



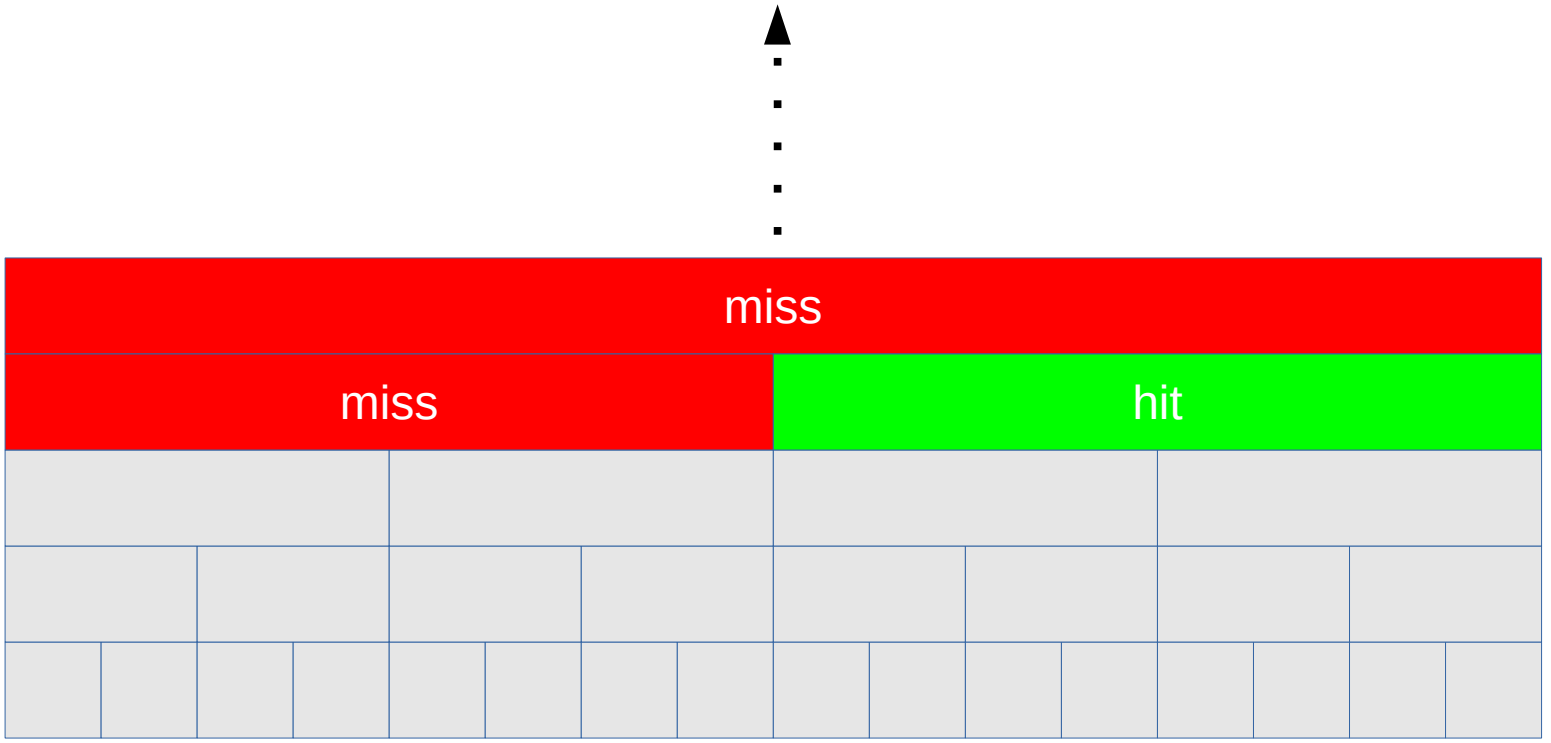
Algorithm 1 (top-down)



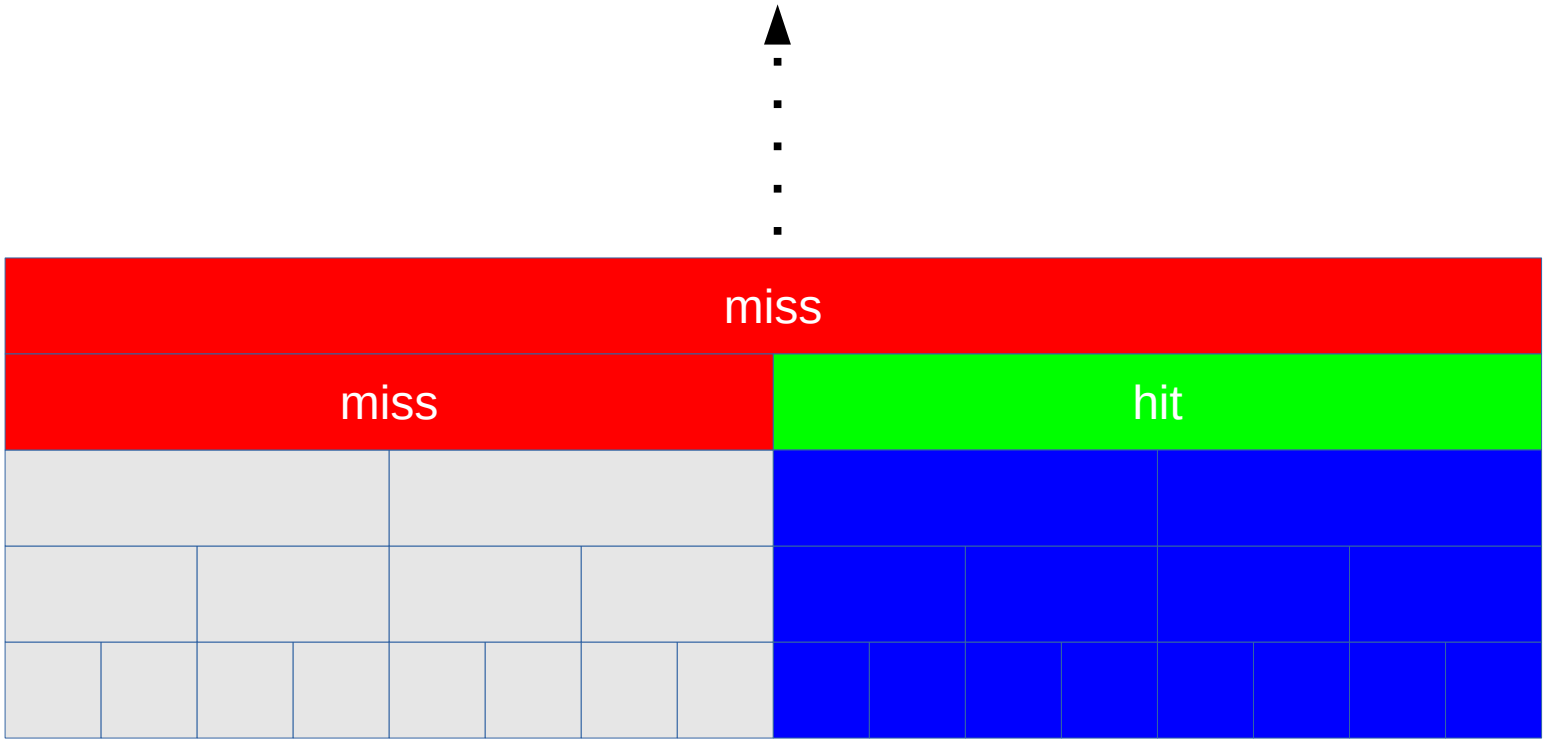
Algorithm 1 (top-down)



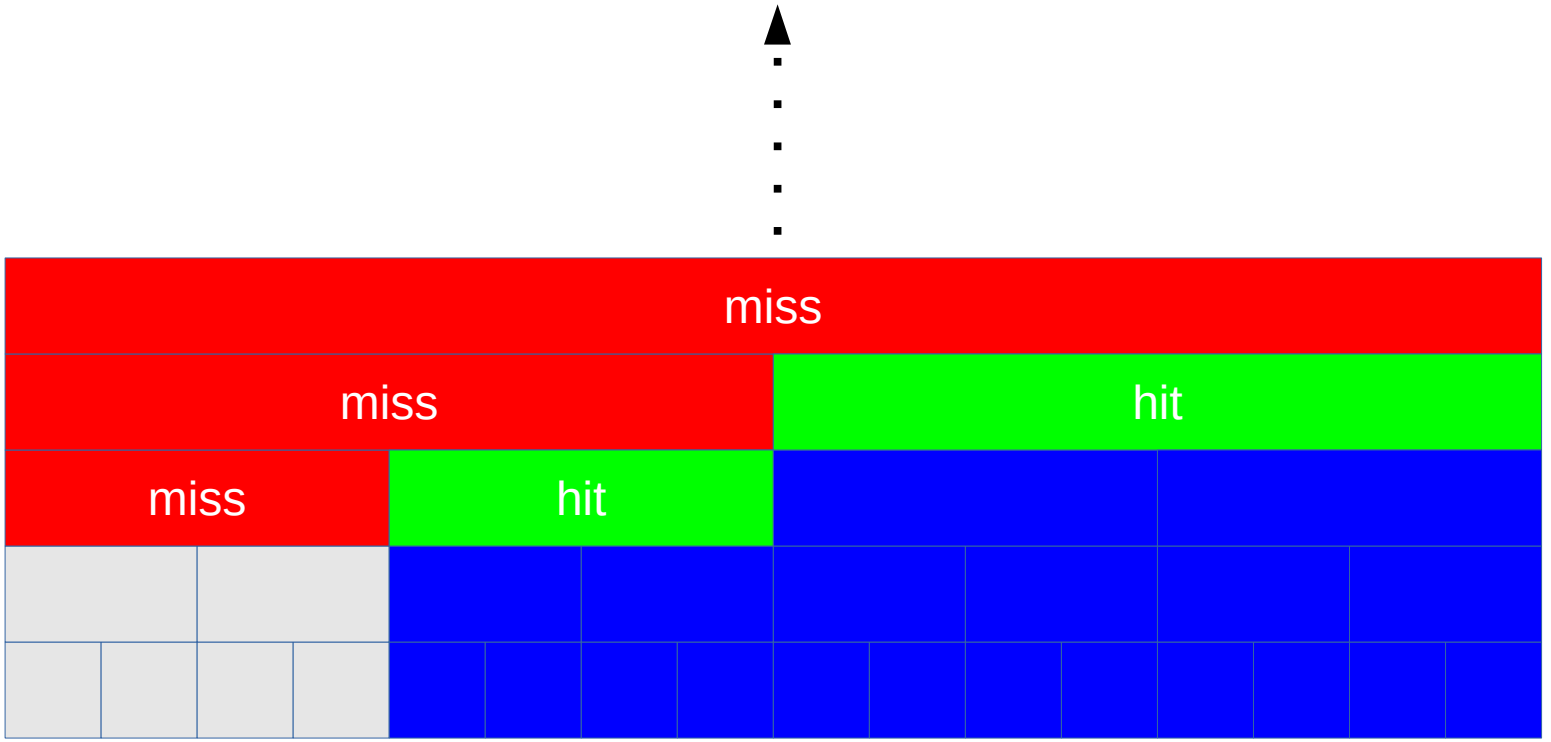
Algorithm 1 (top-down)



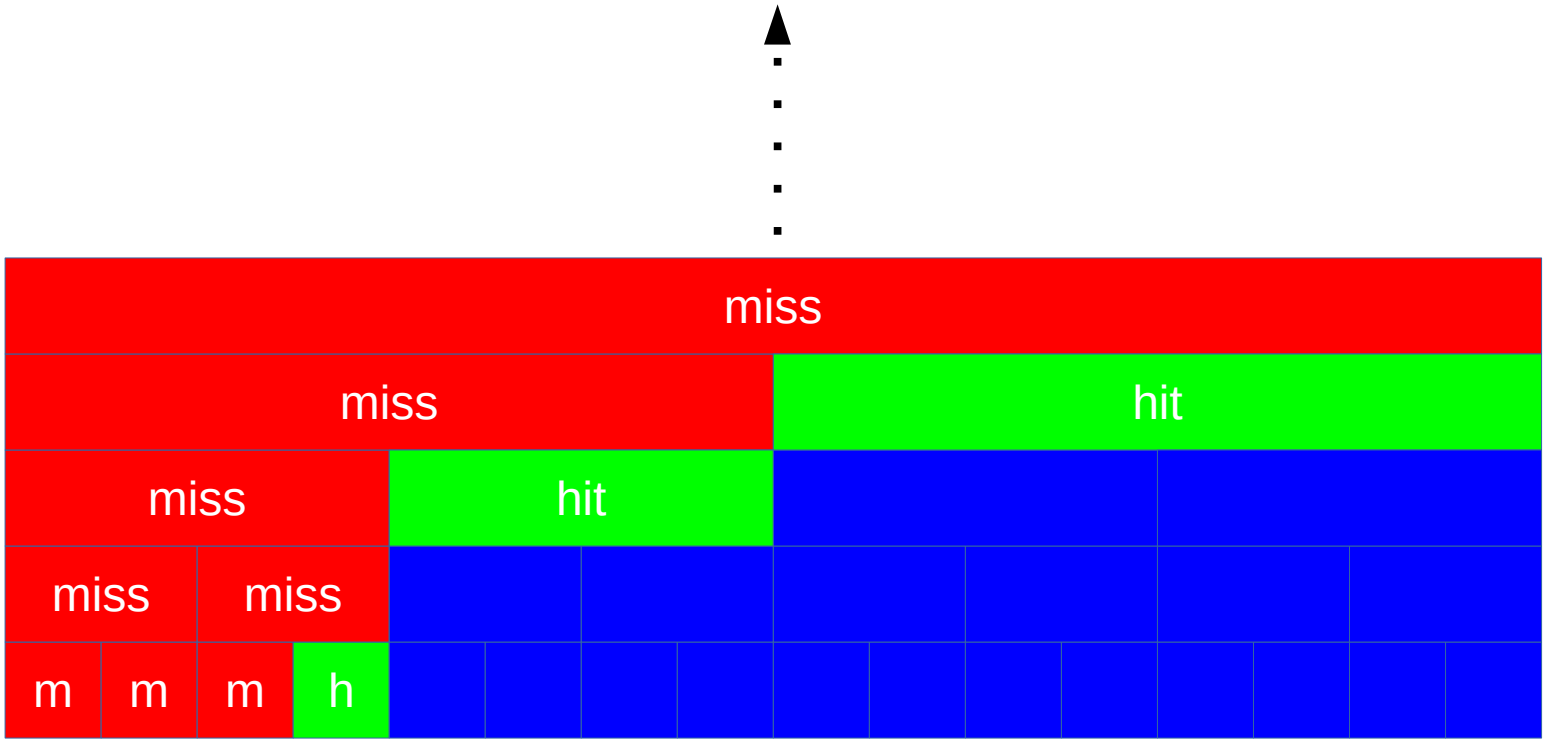
Algorithm 1 (top-down)



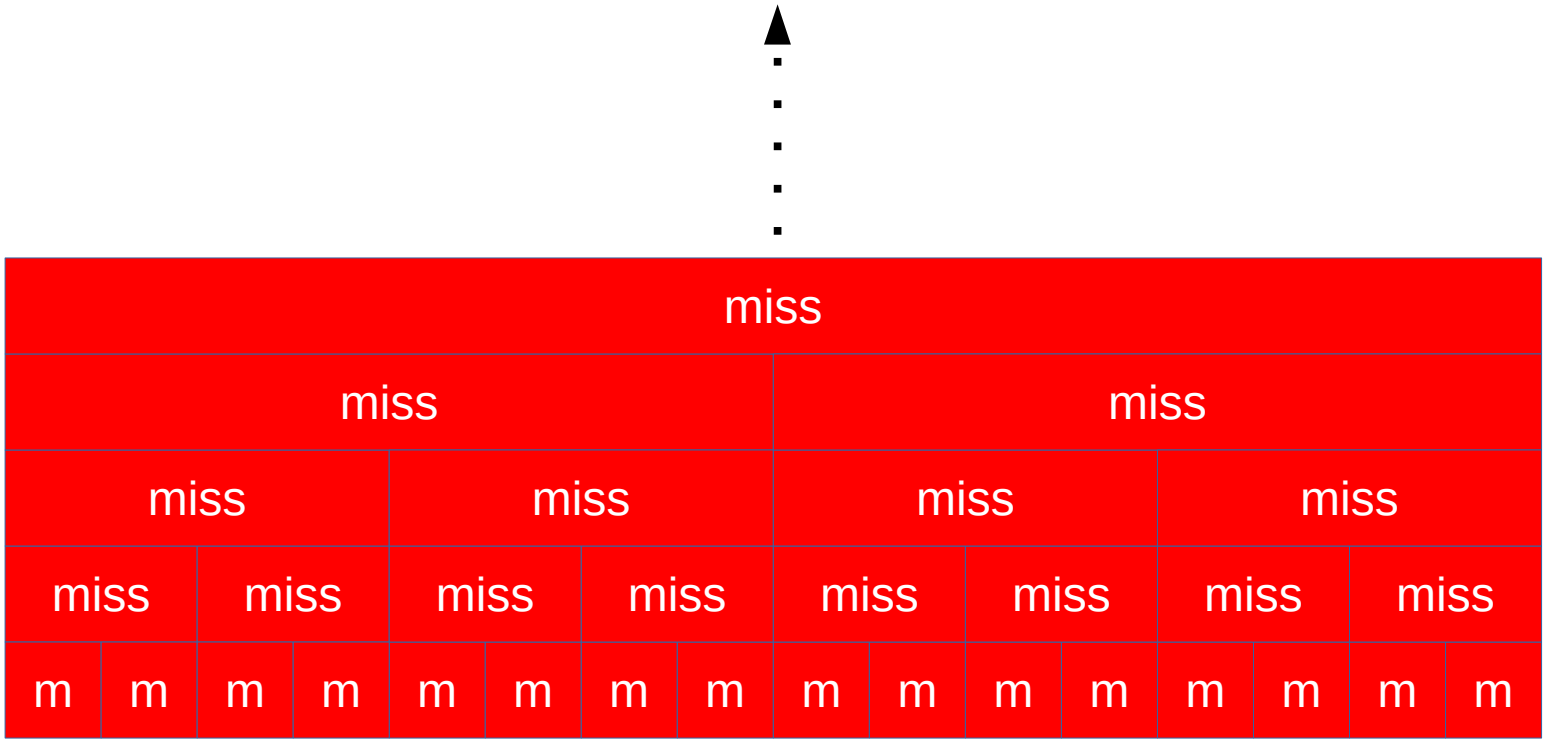
Algorithm 1 (top-down)



Algorithm 1 (top-down)



Algorithm 1 Worst Case

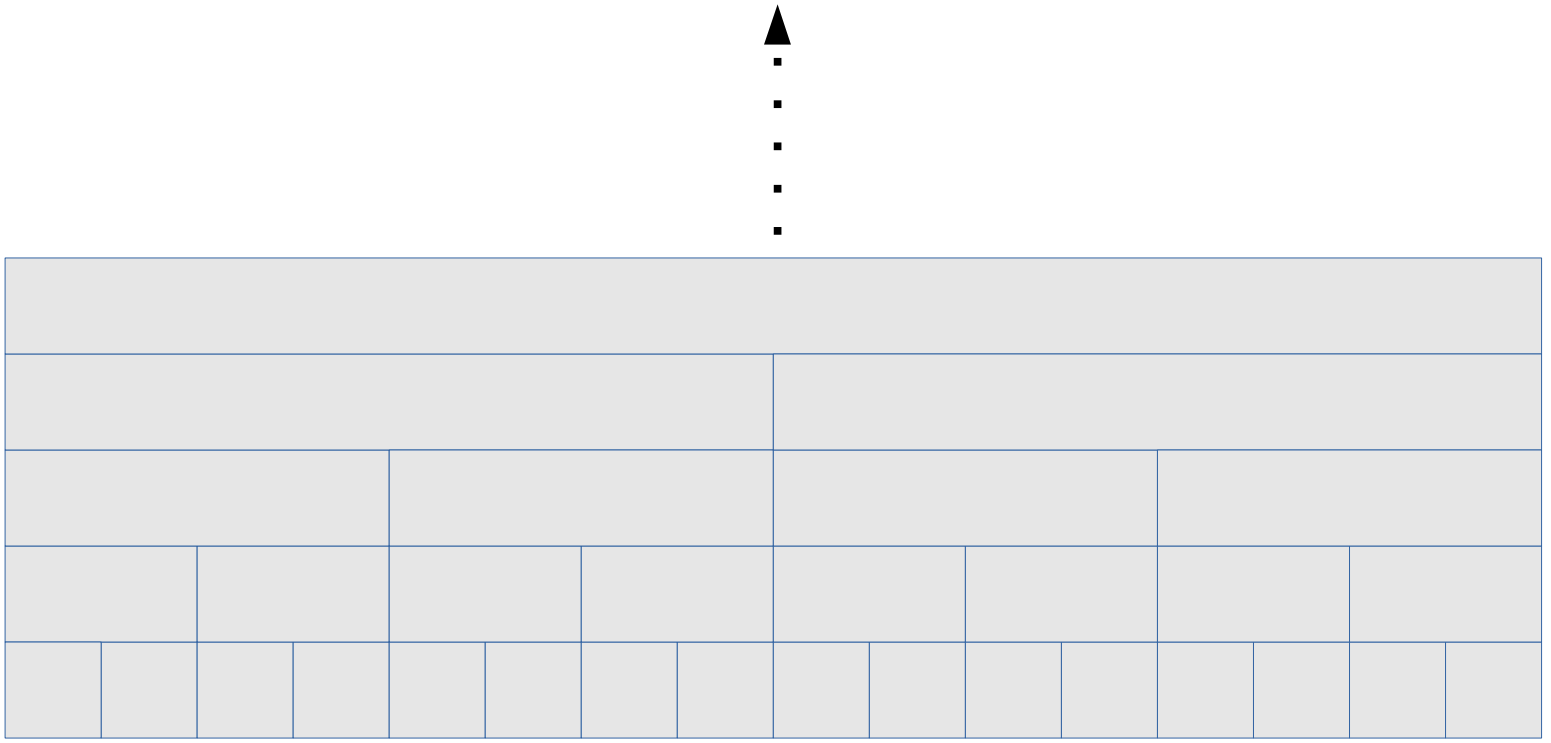


Algorithm 2 (bottom-up)

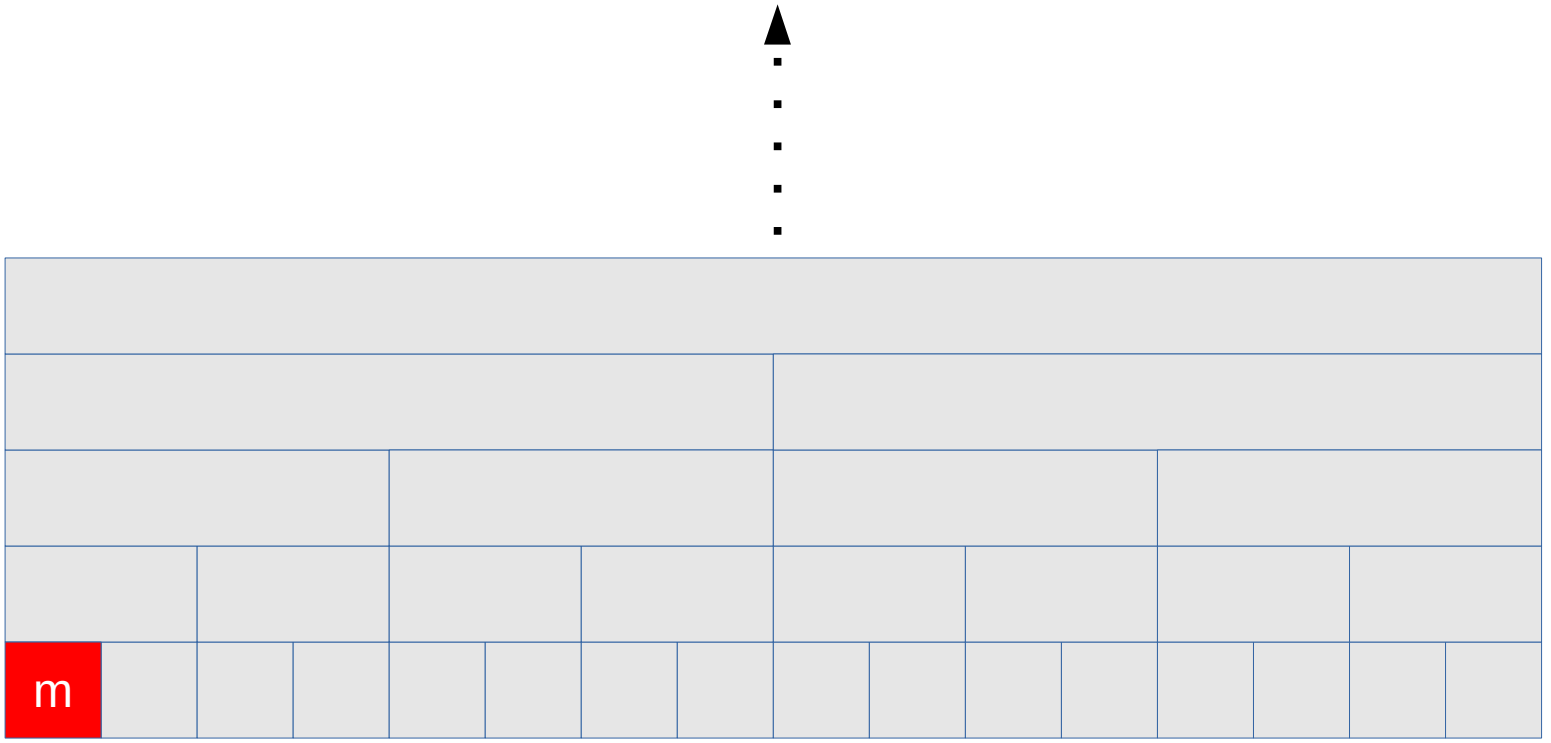
- Recipe Sort:
  - position (ascending)
  - size (ascending)
- On miss, skip all @ the location

- Best case: close to Algo 1
- Worst case*: match single-scale

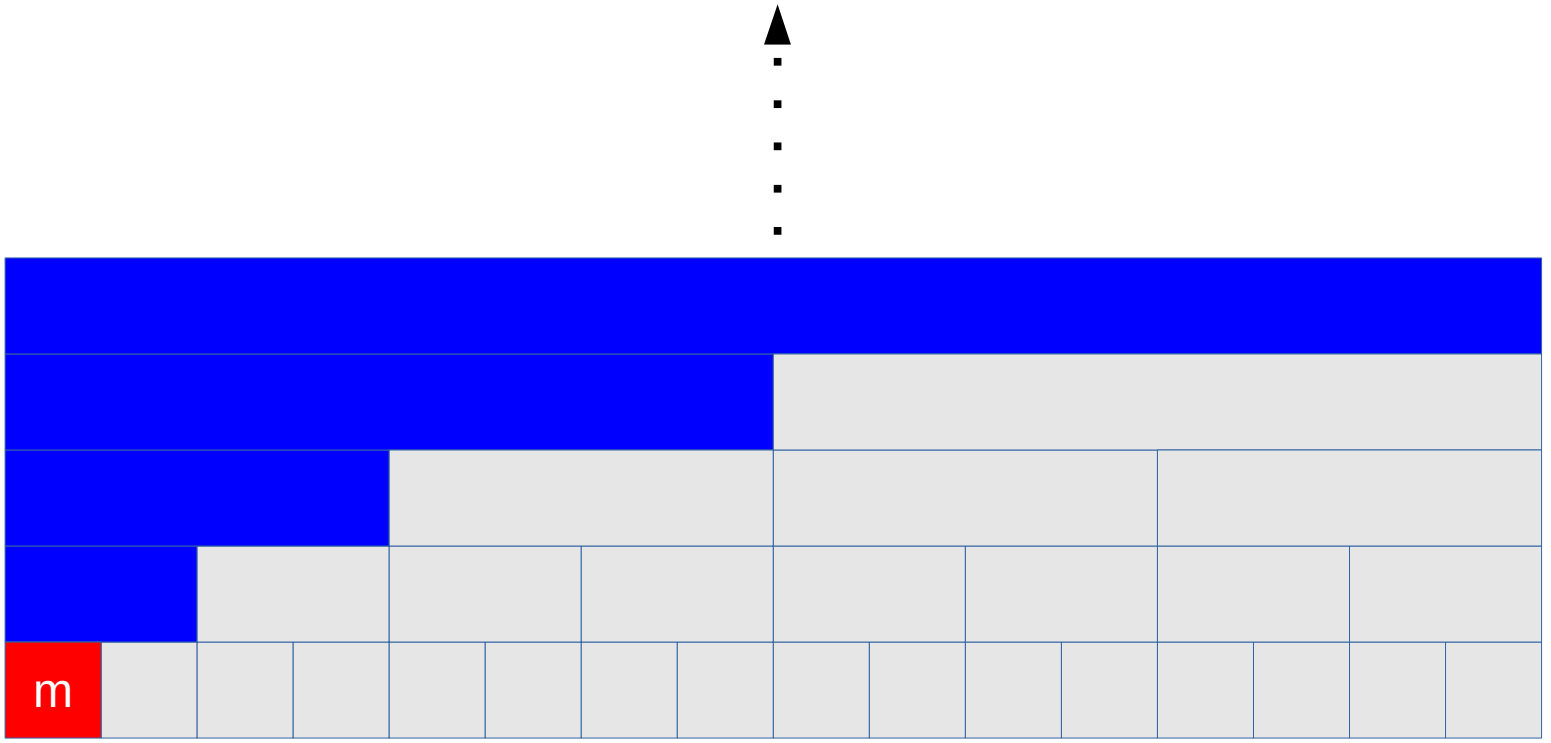
Algorithm 2 (bottom-up)



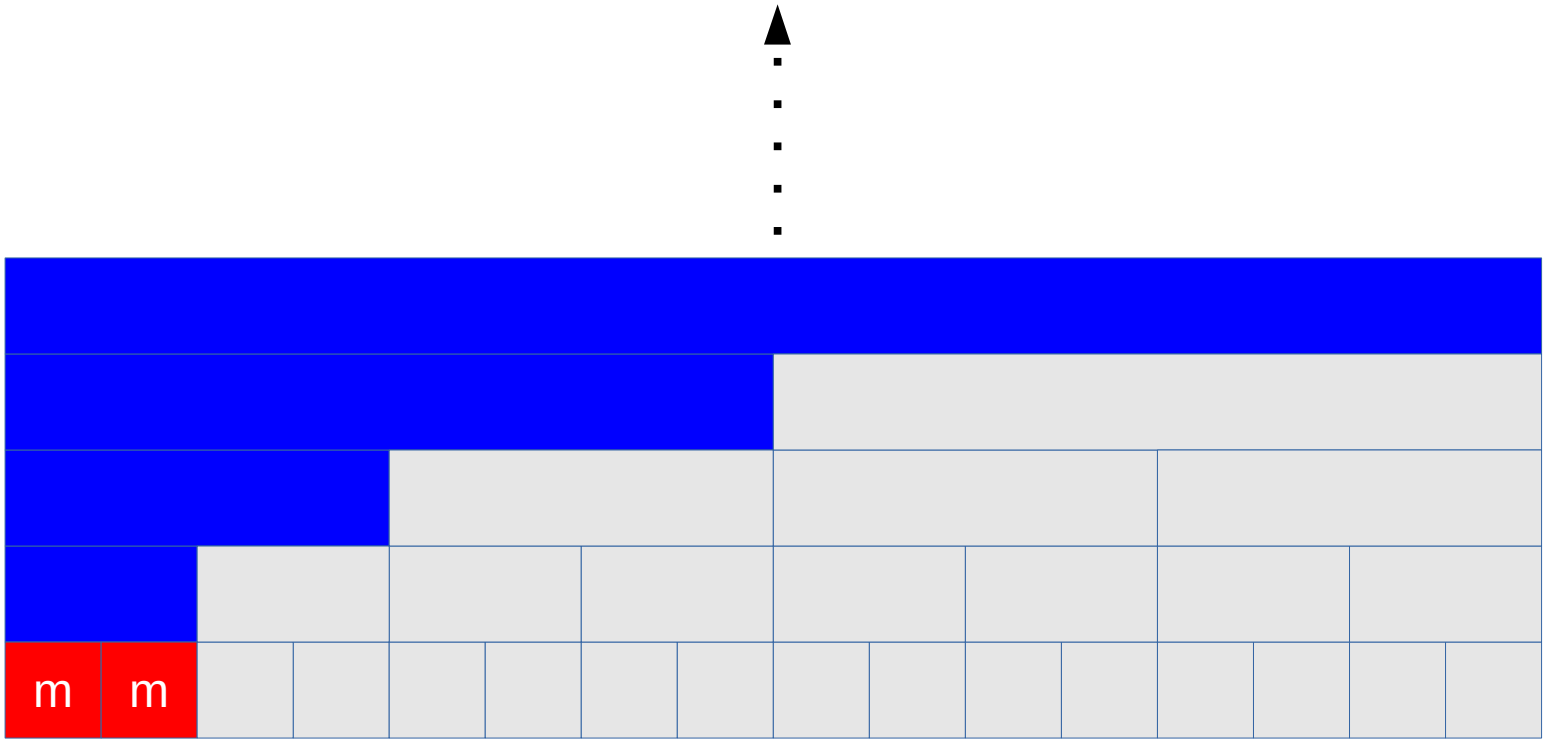
Algorithm 2 (bottom-up)



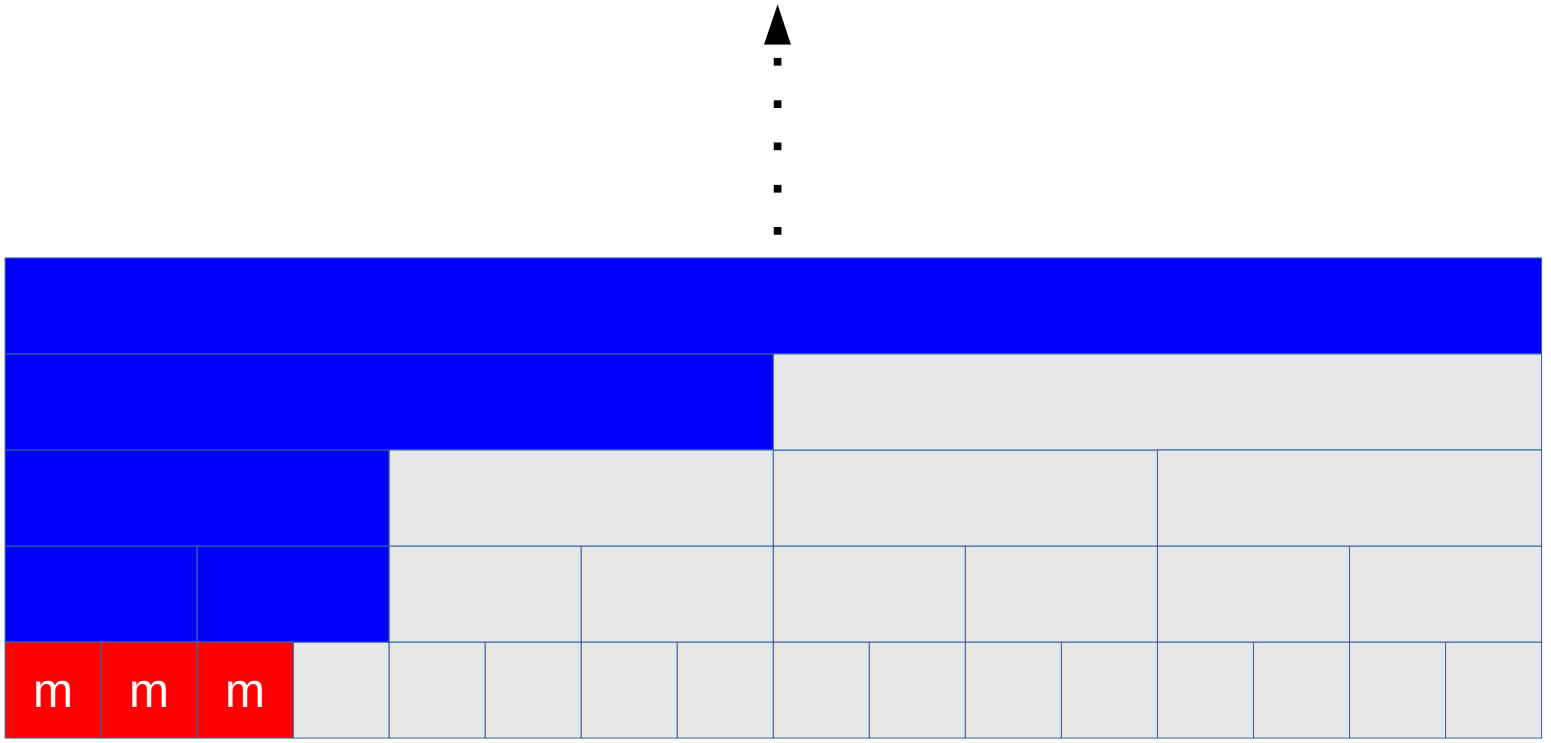
Algorithm 2 (bottom-up)



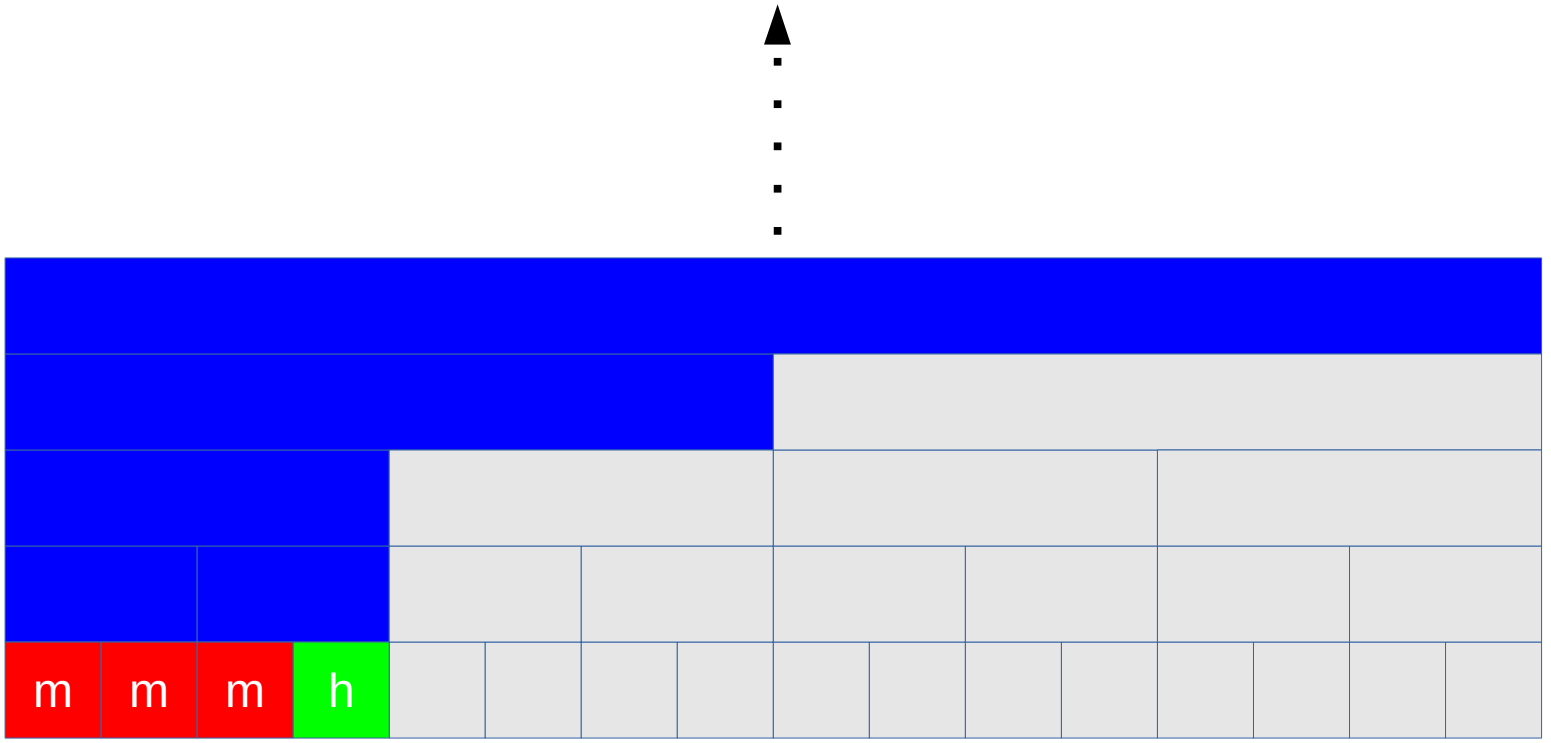
Algorithm 2 (bottom-up)



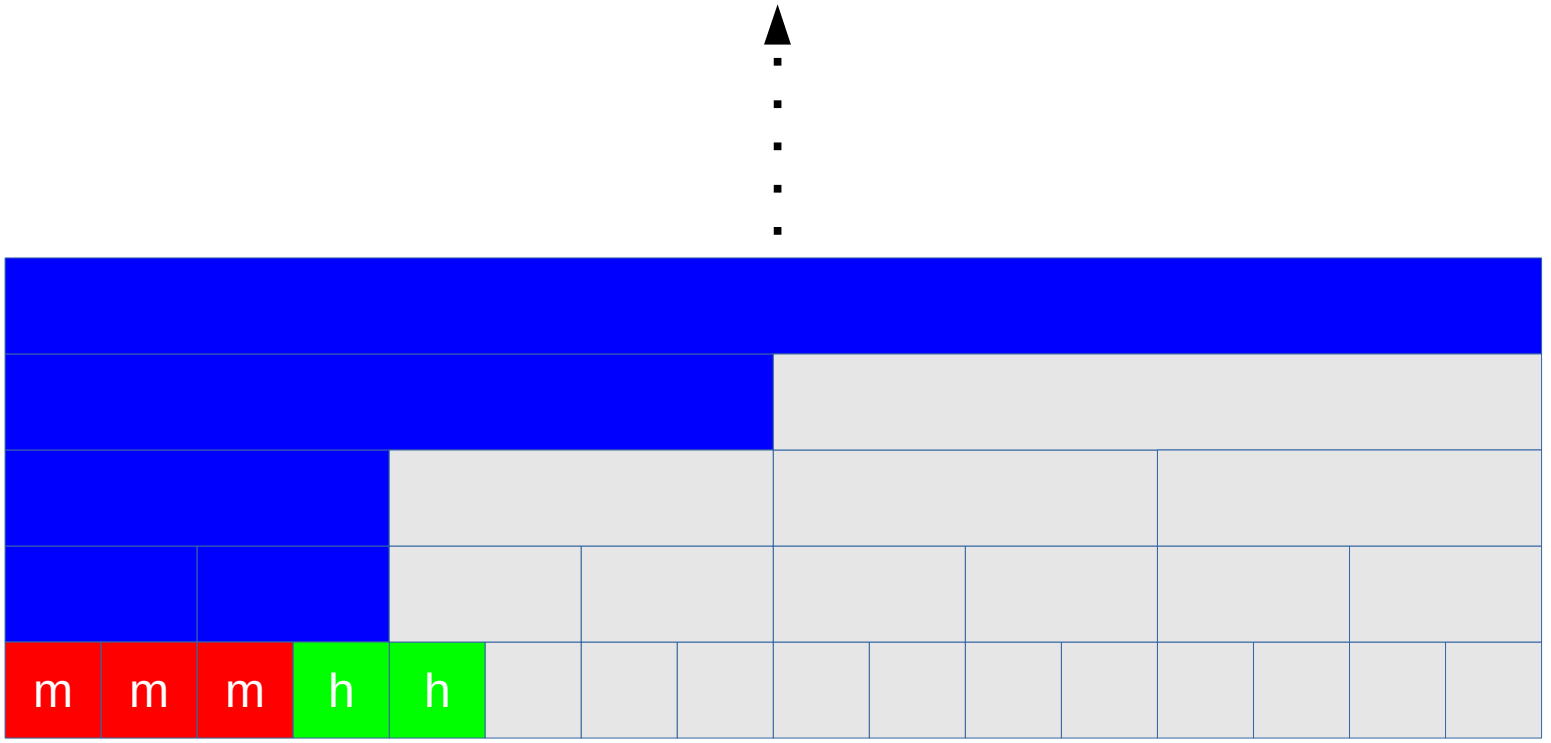
Algorithm 2 (bottom-up)



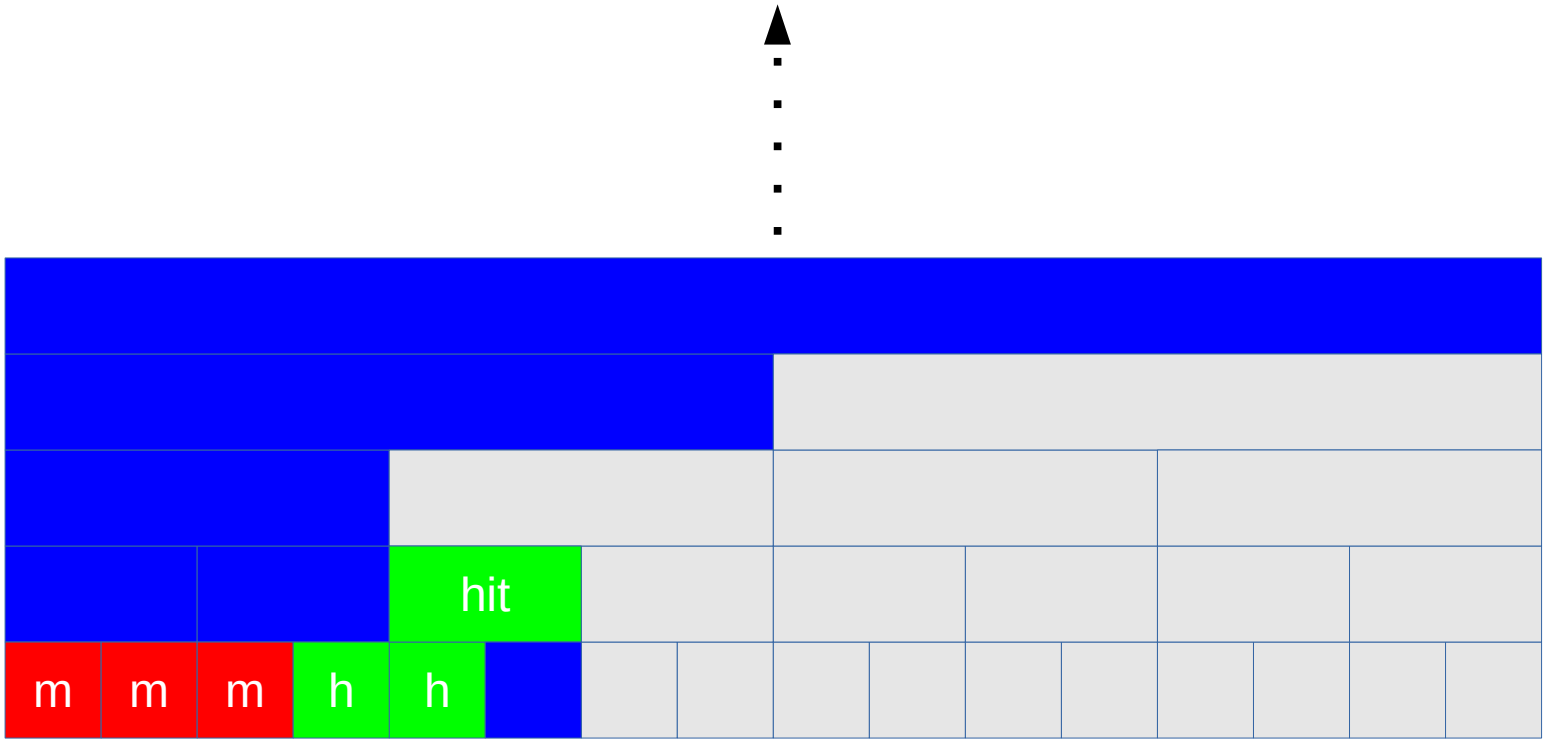
Algorithm 2 (bottom-up)



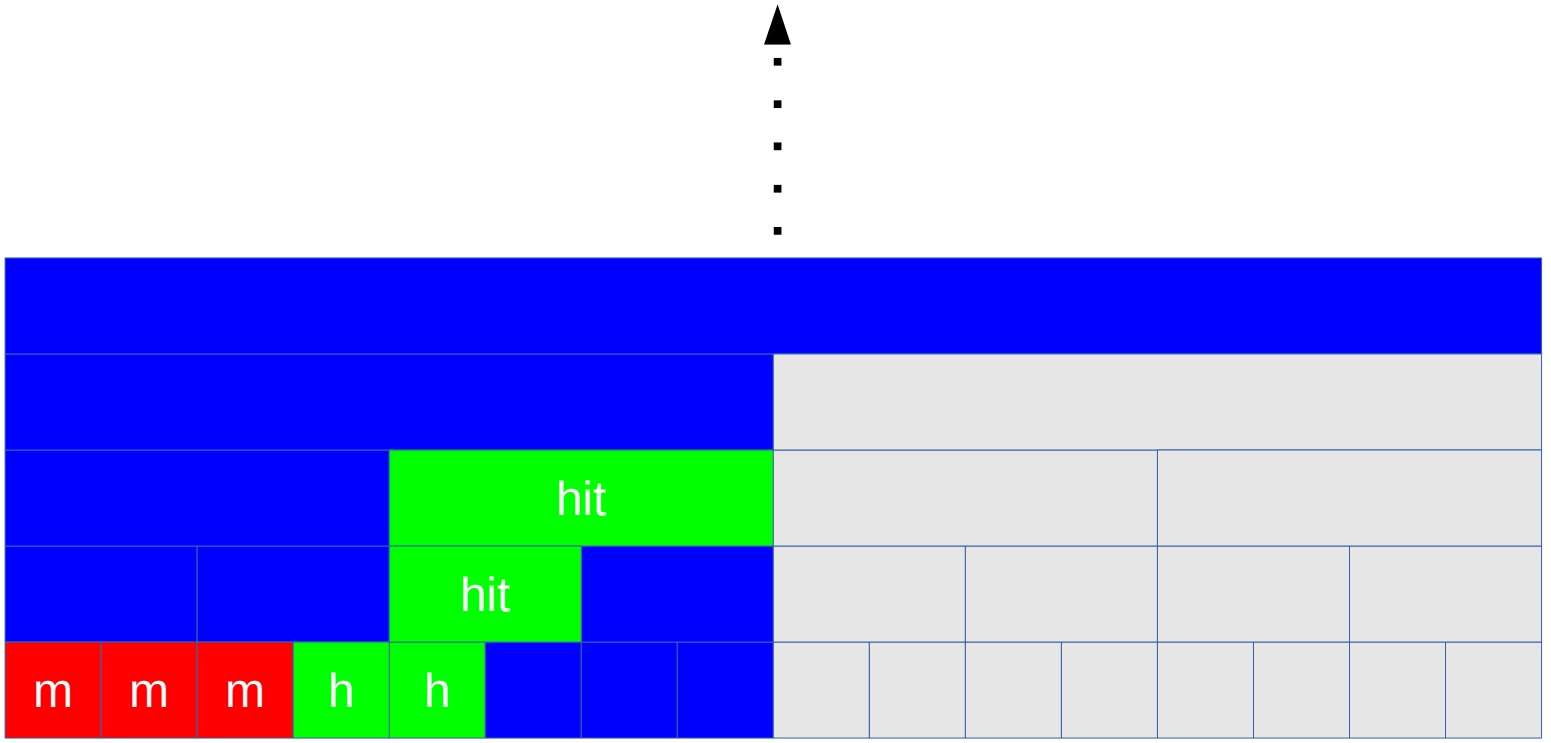
Algorithm 2 (bottom-up)



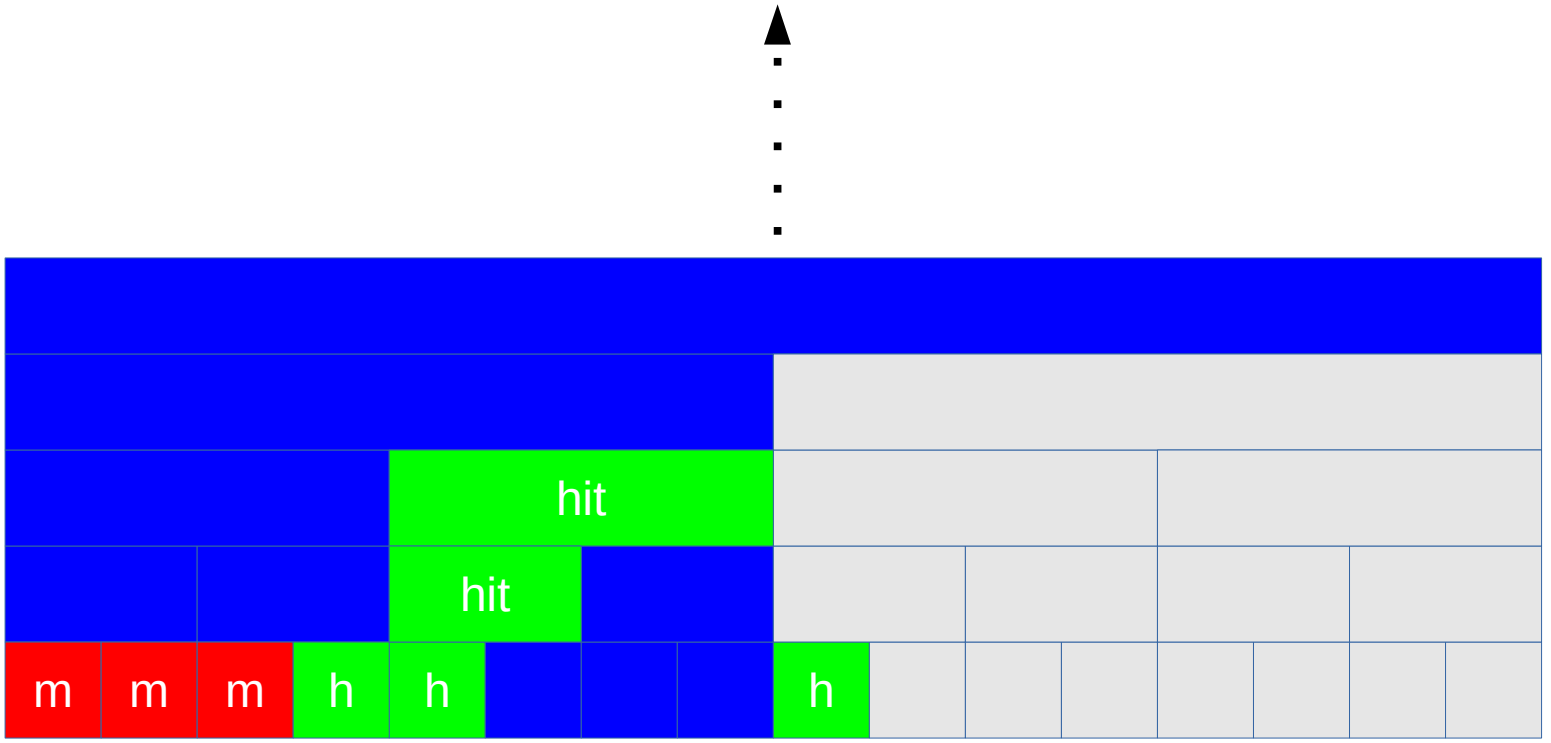
Algorithm 2 (bottom-up)



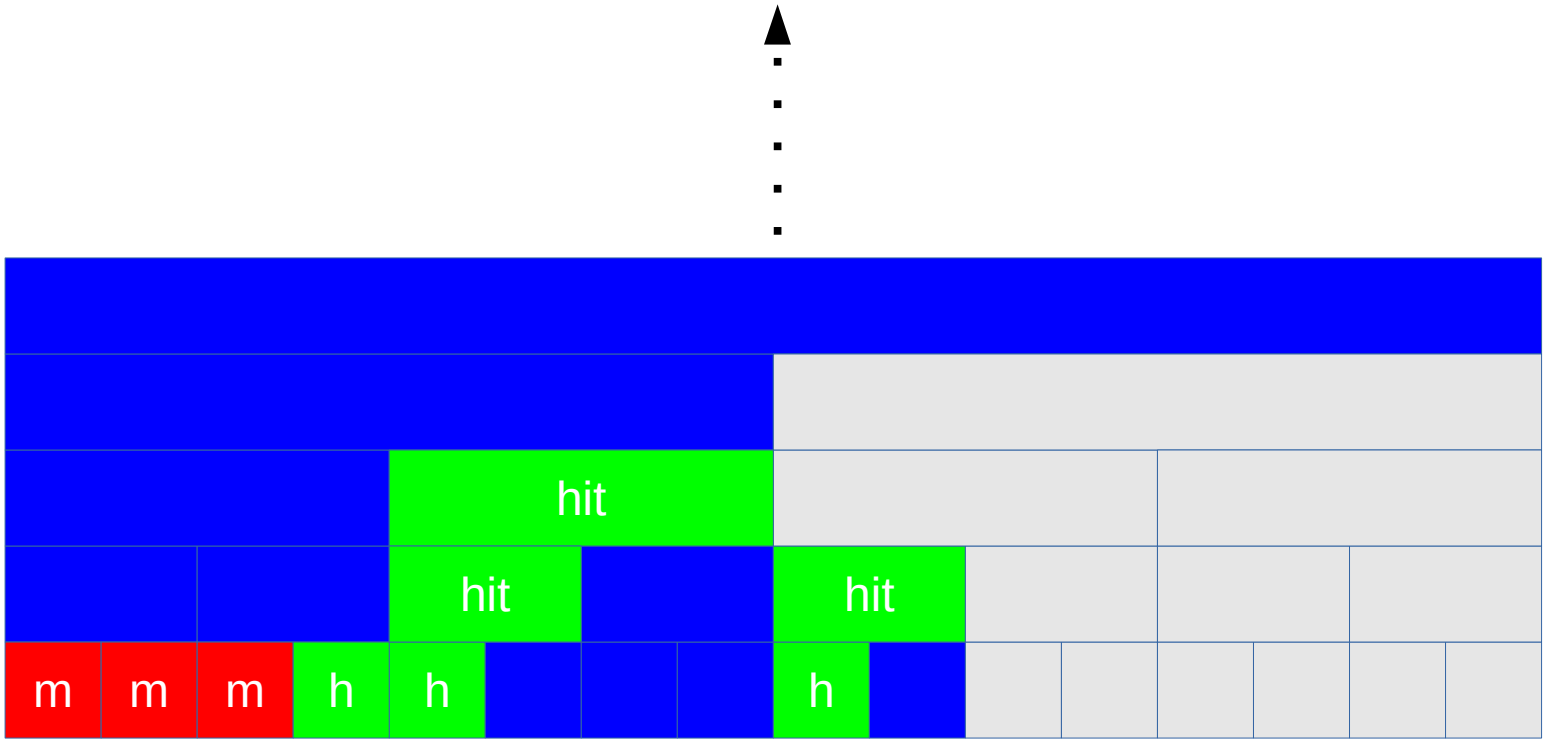
Algorithm 2 (bottom-up)



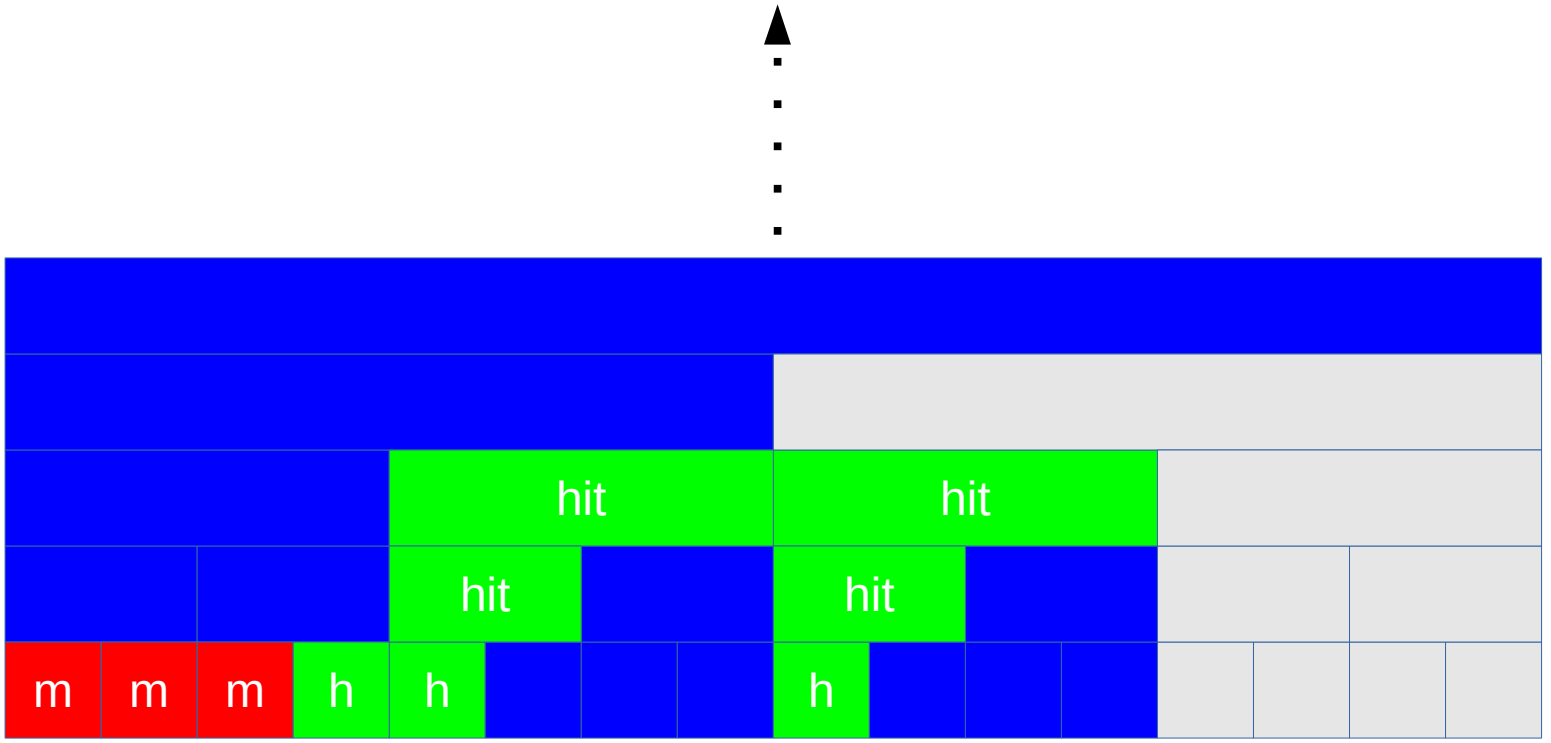
Algorithm 2 (bottom-up)



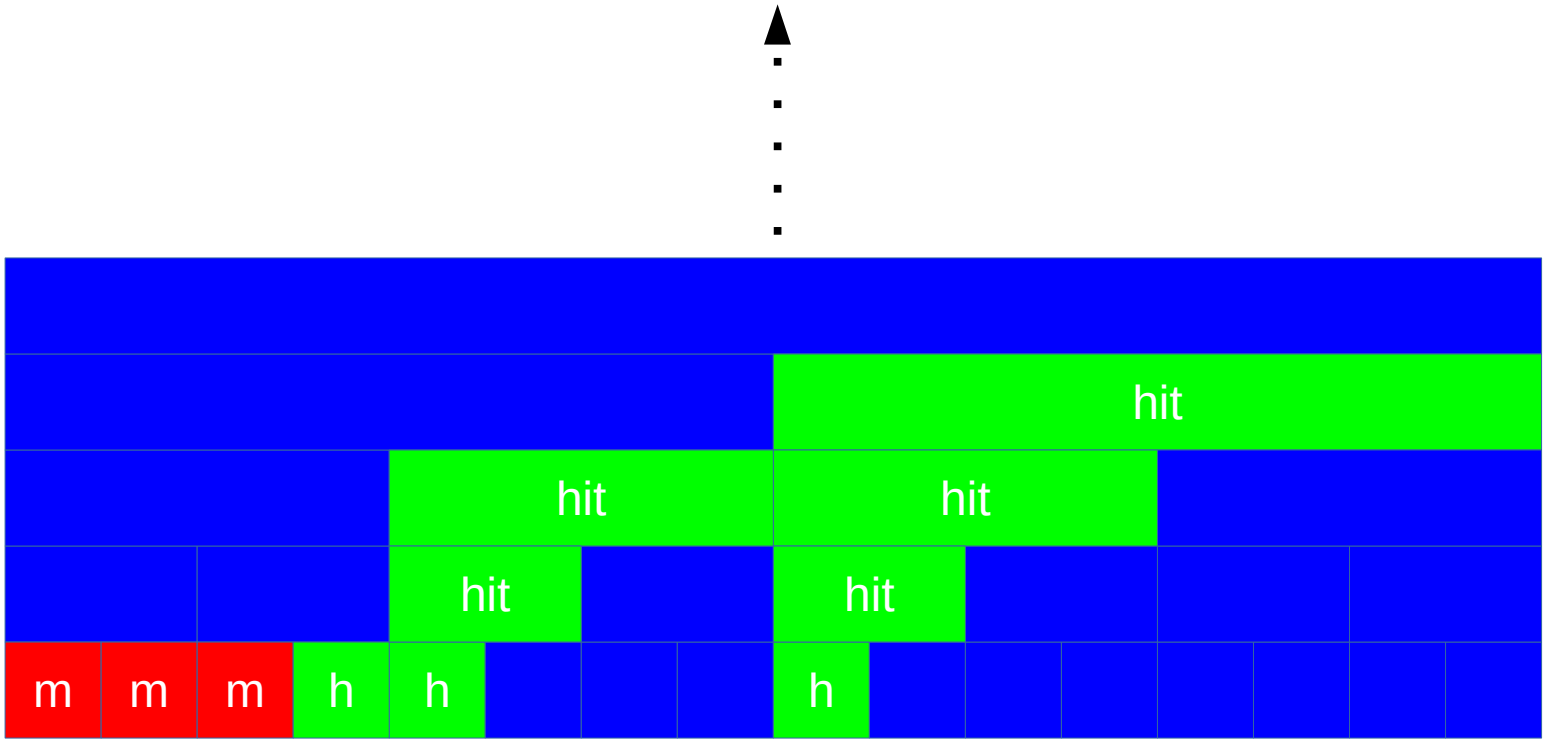
Algorithm 2 (bottom-up)



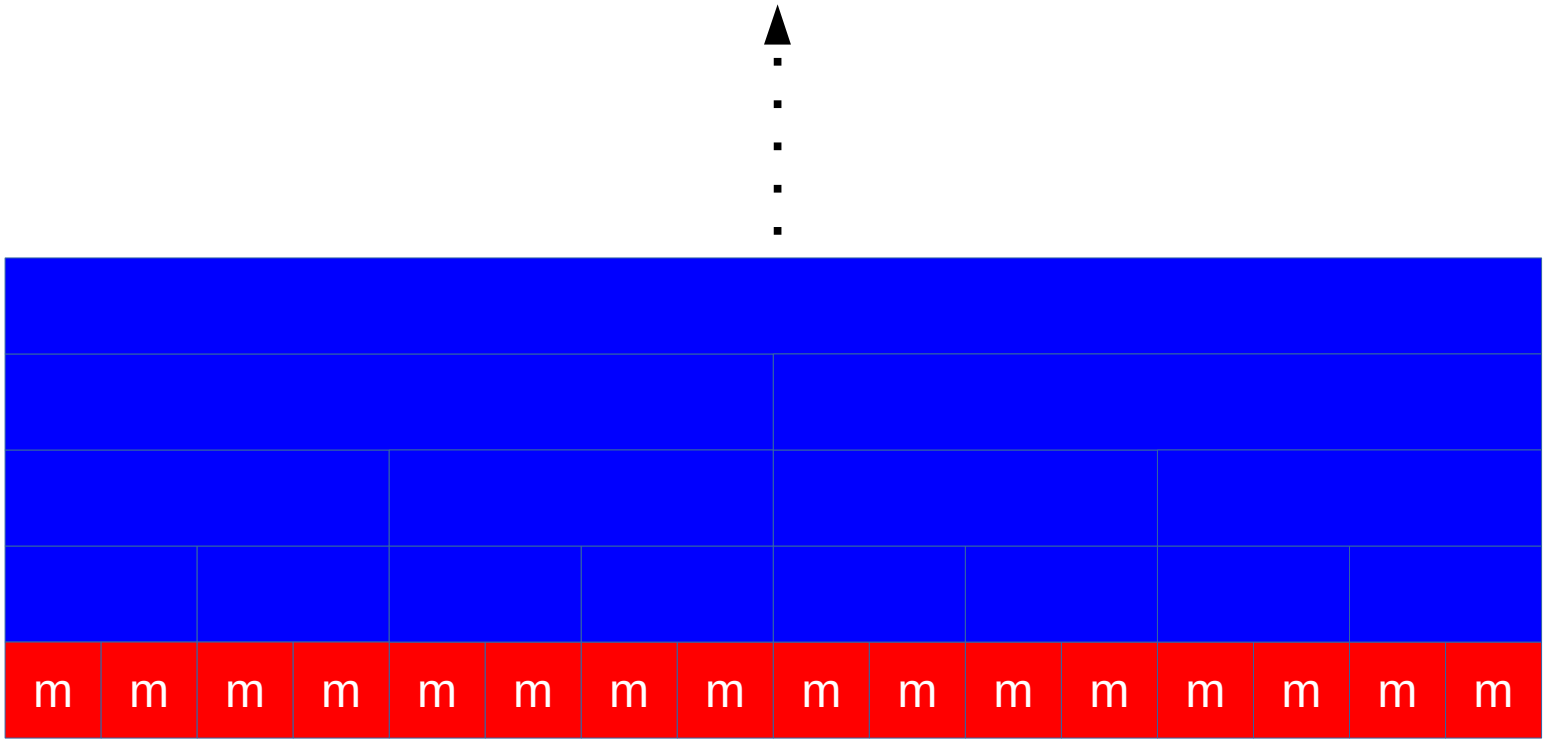
Algorithm 2 (bottom-up)



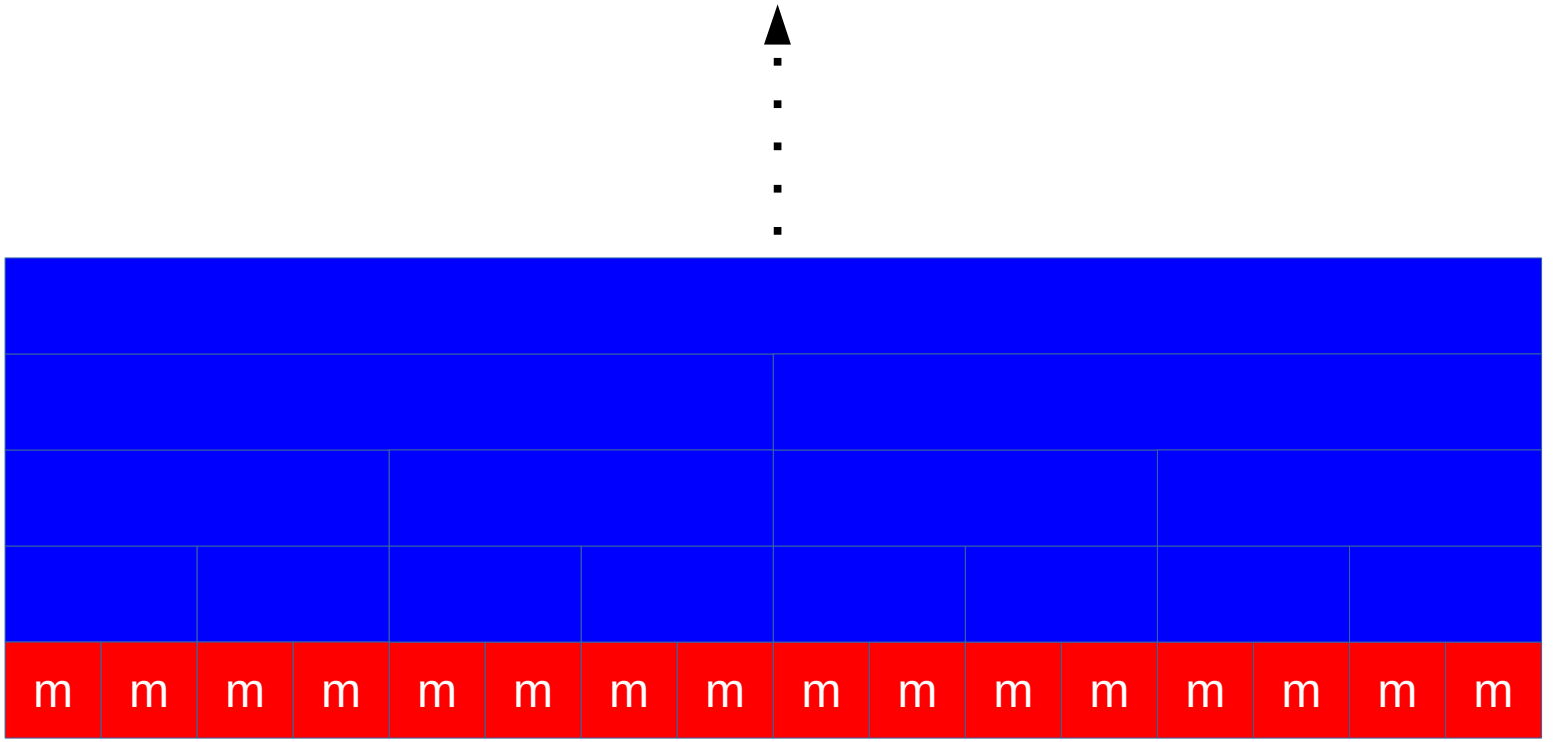
Algorithm 2 (bottom-up)



Algorithm 2 Worst Case

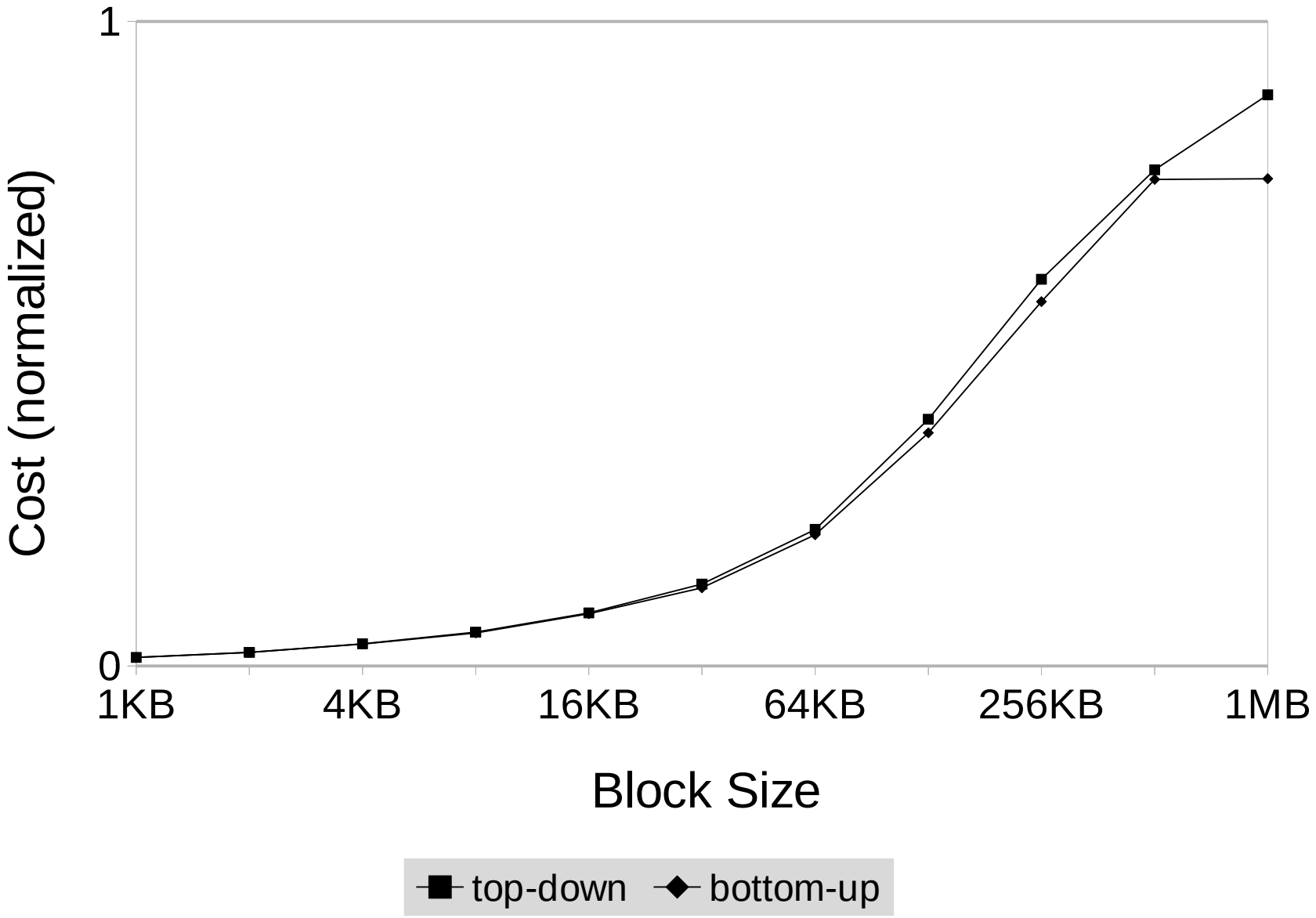


Algorithm 2 Worst Case



Recipe
miss
skip
skip
skip
skip
miss
miss
skip
miss
skip
skip
miss
miss
skip
miss
miss
skip
skip
...

Results: High Duplication



Results: Low Duplication