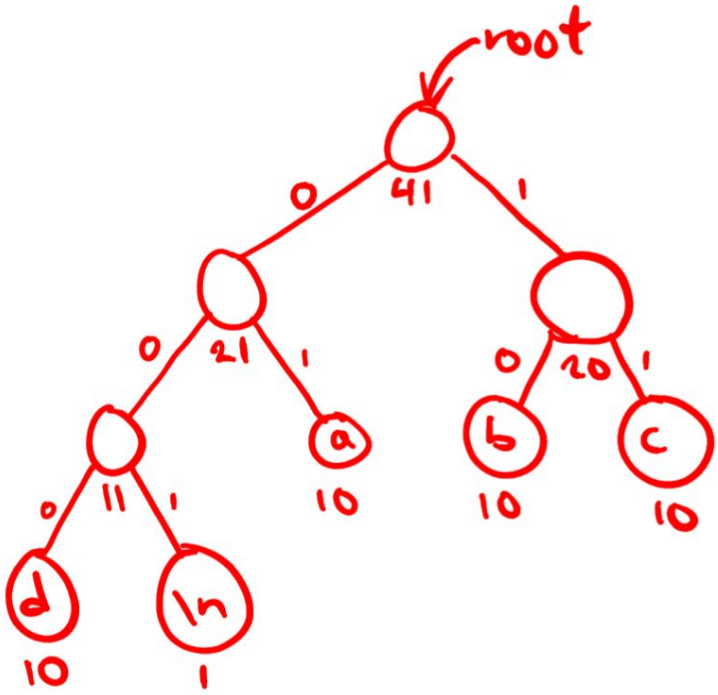
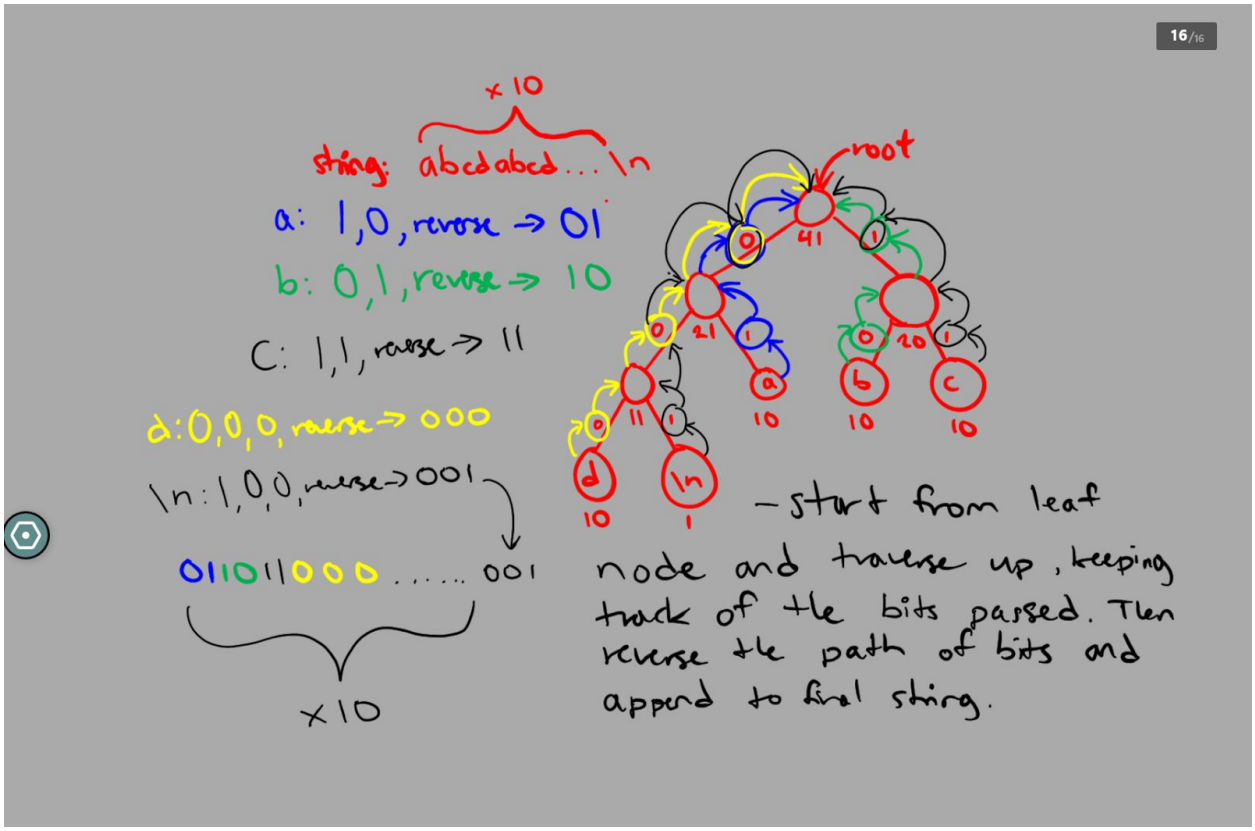
Kent Nguyen
Timothy Ferido

PA3 Report

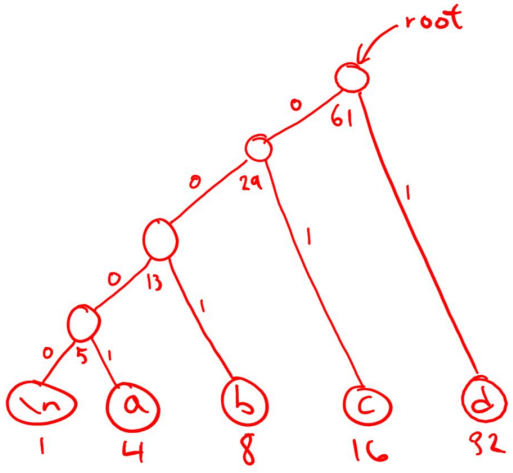
- The output for the command ./compress check1.txt checkoutput1.txt is:
- Line 11: 1
- Line 98: 10
- Line 99: 10
- Line 100: 10
- Line 101: 10
- Encoded string:
  01101100001101100001101100001101100001101100001101100001101100001101100001101100001101100001101100001



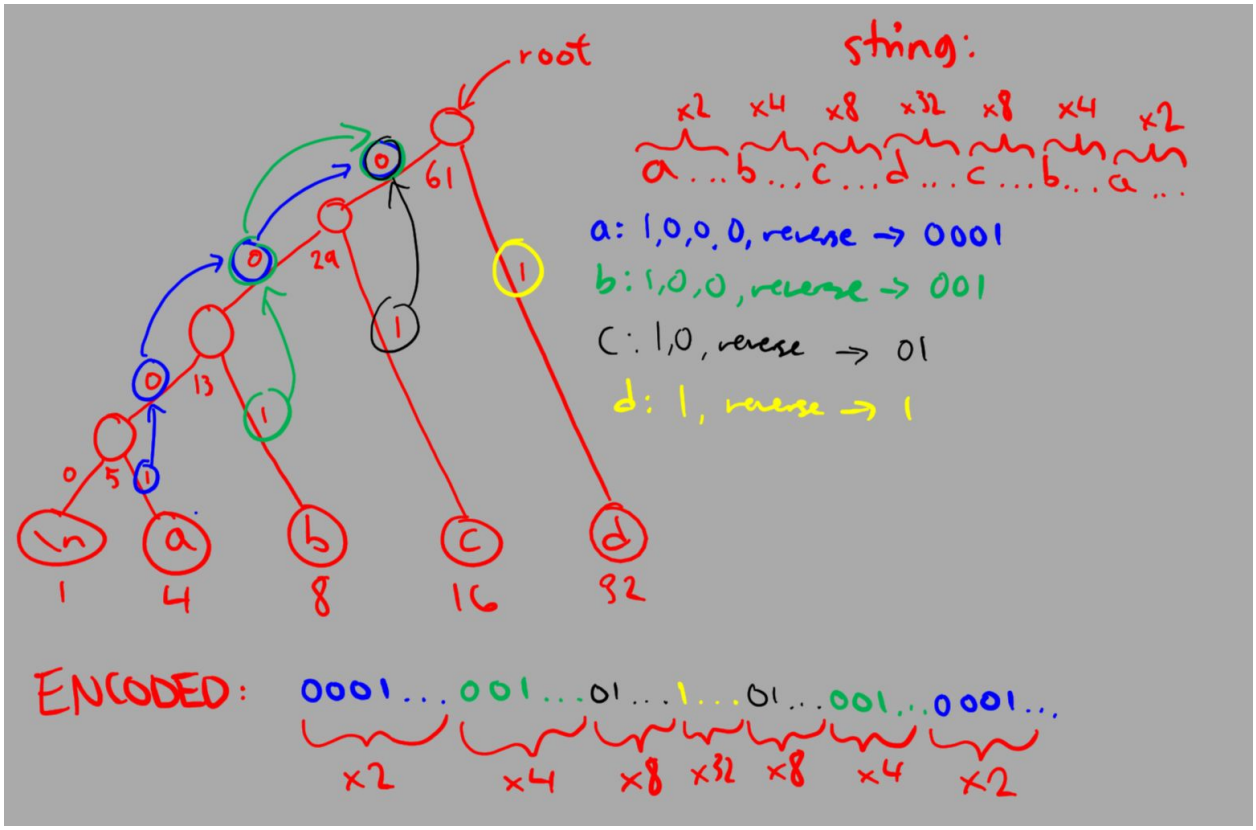
- We built the tree by taking the two nodes with the lowest frequencies and pushing them onto a priority queue. Then two nodes are popped off at a time and a parent node is created that connects them and has the combined frequency. This node is then pushed back onto the priority queue and the process repeats until there is only one node, this is the root.
- To find the code word for each byte, we traverse down from the root, going right if the bit read is ‘1’ and going left if the bit read is ‘0’. Once a leaf node is reached (no children) then the symbol is returned



- The output for the command ./compress check2.txt check2output.txt is:
- Line 11: 1
- Line 98: 4
- Line 99: 8
- Line 100: 16
- Line 101: 32
- Encoded string:
  11101110110110110110101010101010101000000000000000000000000000000010101010101010110110110110111011011101111
- The manual huffman tree we got looks like this:



- Finding the manual encoding:

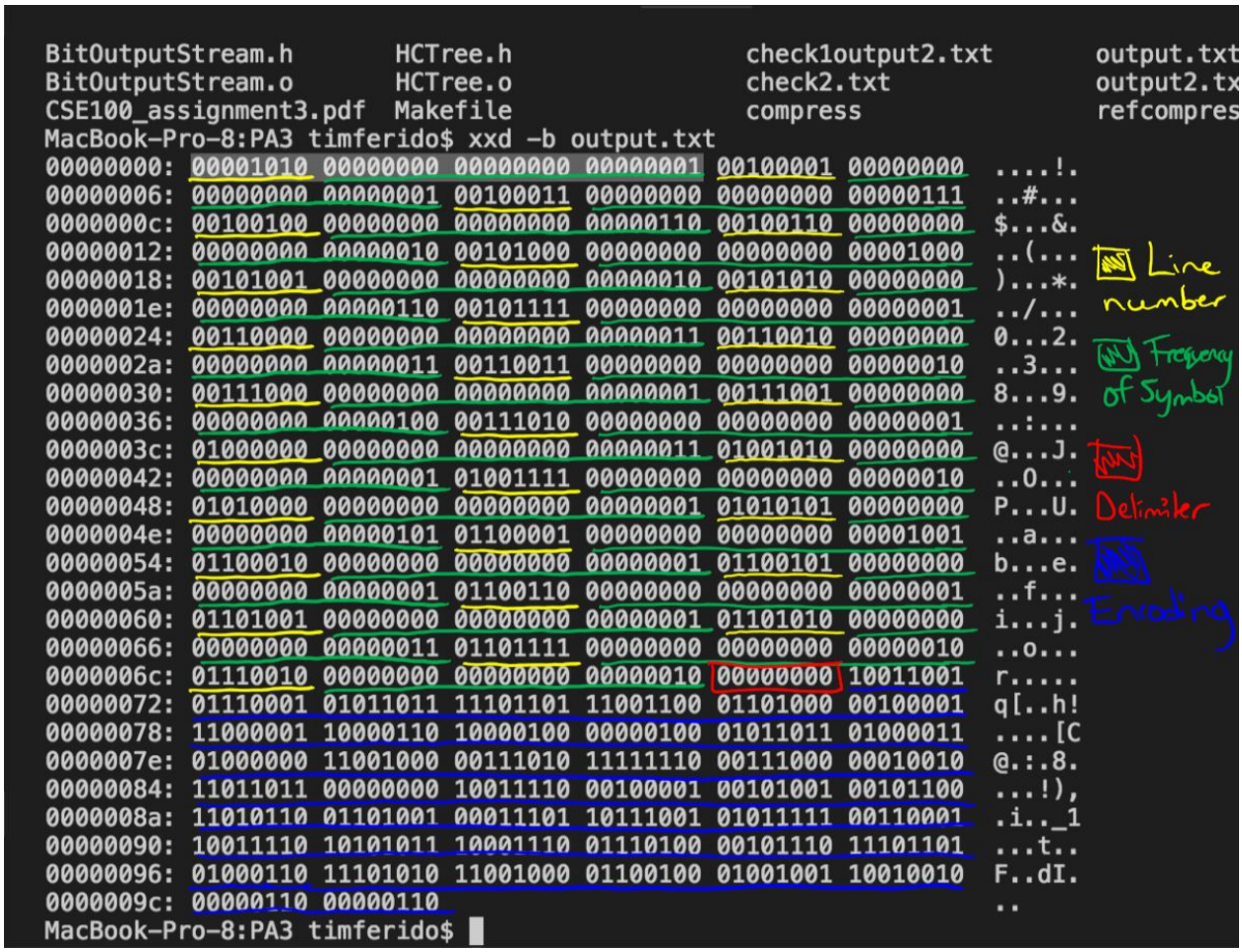


- As we can see our manual encoding produced a string much like the one our compress algorithm produced EXCEPT the 0's and 1's are flipped. This is because when we manually encoded the tree, the 0 child and the 1 child were assigned in reverse.

Efficient Header Re-Design

- In order to implement a more efficient header when compressed, we decided to represent the *non-zero* frequencies as a 4 byte combination of two parts. The first part is a 1 byte number representing the line number / index of the freqs vector to build the Huffman tree. The second part is a 3 byte number representing the frequency. Additionally, the newline characters were removed. Here is a hexdump of the redesigned

header:



- The only “entries” are frequencies that are non-zero. The first “Line Number Byte” that is 00000000, and is not the first byte, is read to be the delimiter and the remaining bytes create the encoding, using the tree.