1995 Paper 3 Question 6

Data Structures and Algorithms

A reasonable approximation for the average number of probes to insert a new entry into an open hash table where a fraction β of the table is already in use is $1/(1-\beta)$.

Suppose that data is stored by initially creating a really tiny hash table (say a vector of size just 8). Entries are added to the table from time to time. Whenever the table becomes 3/4 full a new hash table, twice the size, is created: all existing data is taken out of the old table, and inserted instead into the new one. Thus in general the table that is in use will be between 3/8 and 3/4 full, and looking things up in it will be efficient.

Although the above method has good predicted costs for *retrieving* information stored in the table, there remains some worry that the repeated cost of copying data from smaller to larger tables may be excessive. Suppose that at some stage N items have been inserted and that the very last insertion provoked the copying step. Estimate the ratio between the *total* number of hash probes performed while building the table and N. How does it compare with the number that would have been used if the table had been built full-sized to start with rather than having to grow stage by stage on the way? [20 marks]

[If you really need their values, you may assume that $\ln(2) = 0.7$, $\ln(3) = 1$ and $\ln(5) = 1.6$, but note that you are not expected to perform any arithmetic tedious enough to call for a calculator: a reasonable estimate (for example to within a factor of 1.5) and a justification of how that value was arrived at is what is required.]