CMPS340 File Processing
An Example of Calculating the Running Time of Merging Sorted Runs

We are given that the time to do a seek (including rotational delay) is, on the average, $20ms$, the time to transfer the contents of the file to/from RAM from/to disk (ignoring seeking) is 200 seconds, and the file is 240 times as large as available RAM. Translating this into the notation of the sibling document, *Estimating the running time for the merging phase of external (disk) sorting*, we have $s + r = 20$ *ms*, $t_f = 200$ *sec*, and $F = 240 \cdot R$. In particular, $F/R = 240$. Using the formula found about two-thirds of the way down the second page of the sibling document, we derive the estimated time for doing a single pass using P-way merging:

$$\begin{aligned} t &= P \cdot \frac{F}{R} \cdot (s + r) + t_f \\ &= P \cdot 240 \cdot (0.02sec) \ + \ 200sec \\ &= P \cdot 4.8sec \ + \ 200sec \end{aligned}$$

(a) Using 240-way merging (thereby finishing in one pass).
Solution: Plugging 240 in for P in the above, we get

$$\begin{aligned} t &= 240 \cdot 4.8sec + 200sec \\ &= 1152sec + 200sec \\ &= 1352sec \end{aligned}$$

(b) Using 16-way merging on the 1st pass and 15-way merging on the 2nd.
Solution: Plugging 16 in for P gives us an estimated time t_1 for the first pass; plugging 15 in for P gives us an estimated time t_2 for the second pass.

$$\begin{aligned} t_1 &= 16 \cdot 4.8sec + 200sec \\ &= 76.8sec + 200sec \\ &= 276.8sec \\ t_2 &= 15 \cdot 4.8sec + 200sec \\ &= 72sec + 200sec \\ &= 272sec \end{aligned}$$

In total, the two passes require about 548.8 seconds.

(c) Using 12-way merging on the 1st pass and 20-way merging on the 2nd.
Solution: Plugging 12 in for P gives us an estimated time t_1 for the first pass; plugging 20 in for P gives us an estimated time t_2 for the second pass.

$$\begin{aligned} t_1 &= 12 \cdot 4.8sec + 200sec \\ &= 57.6sec + 200sec \\ &= 257.6sec \\ t_2 &= 20 \cdot 4.8sec + 200sec \\ &= 96sec + 200sec \\ &= 296sec \end{aligned}$$

In total, the two passes require about 553.6 seconds. Notice that this result is not quite as good as the result of (b), the reason being that, in (b), $|P_1 - P_2|$ is smaller. That is, our choices of P_1 and P_2 in (b) were closer together. (Recall from the accompanying document that it is best to choose the P_i's to be as close to one another in value as possible.)

(d) Using 6-way merging on the 1st pass, 5-way on the 2nd, and 8-way on the 3rd.
Solution: Plug in 6 for P in order to calculate t_1, plug in 5 for P to compute t_2, and plug in 8 for P to compute t_3. (From the fact that three passes are performed, we know that our estimate of total time will exceed $3 \cdot t_f = 600sec$ (which is larger than our estimates in (b) and (c)), and thus we would be better off using only two passes.)

$$\begin{aligned} t_1 &= 6 \cdot 4.8sec + 200sec \\ &= 28.8sec + 200sec \\ &= 228.8sec \\ t_1 &= 5 \cdot 4.8sec + 200sec \\ &= 24sec + 200sec \\ &= 224sec \\ t_3 &= 8 \cdot 4.8sec + 200sec \\ &= 38.4sec + 200sec \\ &= 238.4sec \end{aligned}$$

In total, this requires about 691.2 seconds.

We could have used the last formula in the sibling document to arrive at the same running time estimates. However, that formula does not yield a separate running time for each pass, and we judged it to be useful to obtain that information.

The reader may have found the solutions given above to be unsatisfying, because they do not give any indication of how many sorted runs existed on each pass or of how many seeks had to be made to each of them during a pass. In order to make these values explicit in our formula, refer to equation (2) in the sibling document:

$$t_i = k_i \cdot m_i \cdot (s + r) + t_f$$

Here, k_i denotes the number of seeks performed on each sorted run during pass i and m_i denotes the number of sorted runs existing at the beginning of pass i.

We are given by equation (1) of the handout that, for all i, $m_i = F/S_i$ (or, equivalently, $S_i = F/m_i$). This is not enough information to determine the values m_2, m_3, etc. (or S_2, S_3. etc.). But consider that during pass i, the m_i sorted runs —each of size S_i— are merged into m_i/P_i sorted runs —each of size $S_i \cdot P_i$. That is, for all $i > 0$, $m_{i+1} = m_i/P_i$ and $S_{i+1} = S_i \cdot P_i$. It follows that, for $i \geq 1$,

$$\begin{aligned} m_i &= m_1/\hat{P}_i \\ S_i &= S_1 \cdot \hat{P}_i \end{aligned}$$

where $\hat{P}_k = P_1 \cdot P_2 \cdots P_{k-1}$. (Note that, by this definition, $\hat{P}_1 = 1$.) Now, in this homework, we are given $m_1 = 240$ and $S_1 = R$. Plugging into the above, we get $m_i = 240/\hat{P}_i$ and $S_i = R \cdot \hat{P}_i$.

Substituting for S_i in Equation (3), we calculate the number of seeks made to each sorted run during pass i:

$$\begin{array}{lll} & k_i & \\ = & (S_i \cdot P_i)/R & \text{(by Equation (3))} \\ = & (R \cdot \hat{P}_i \cdot P_i)/R & (S_i = R \cdot \hat{P}_i) \\ = & (R \cdot \hat{P}_{i+1})/R & (\hat{P}_i \cdot P_i = \hat{P}_{i+1}) \\ = & \hat{P}_{i+1} & (R/R = 1) \end{array}$$

Substituting for k_i, m_i, $s + r$, and t_f in Equation (2), we estimate the running time of pass i:

$$\begin{array}{rcl} t_i & = & k_i \cdot m_i \cdot (s + r) + t_f \\ & = & \hat{P}_{i+1} \cdot (240/\hat{P}_i) \cdot (0.02sec) \ + \ 200sec \end{array}$$

(The astute reader will notice that, by using the fact that $\hat{P}_{i+1} = \hat{P}_i \cdot P_i$ and then applying standard algebraic manipulations, the above can be simplified to $t_i = P_i \cdot 4.8sec + 200sec$, which is precisely the formula used in originally calculating the solutions above!)

To conclude, we answer part (d) according to this "new" formula.

(d) Using 6-way merging on the 1st pass, 5-way on the 2nd, and 8-way on the 3rd.
Solution: We have $P_1 = 6$, $P_2 = 5$, $P_3 = 8$, $\hat{P}_1 = 1$, $\hat{P}_2 = 6$, $\hat{P}_3 = 6 \cdot 5 = 30$, and $\hat{P}_4 = 6 \cdot 5 \cdot 8 = 240$.

$$\begin{array}{rcl} t_1 & = & \hat{P}_2 \cdot 240/\hat{P}_1 \cdot .02sec + 200sec \\ & = & 6 \cdot 240/1 \cdot .02sec + 200sec \\ & = & 6 \cdot 240 \cdot .02sec + 200sec \\ & = & 1440 \cdot .02sec + 200sec \\ & = & 28.8sec + 200sec \\ & = & 228.8sec \\ t_2 & = & \hat{P}_3 \cdot 240/\hat{P}_2 \cdot .02sec + 200sec \\ & = & 30 \cdot 240/6 \cdot .02sec + 200sec \\ & = & 30 \cdot 40 \cdot .02sec + 200sec \\ & = & 1200 \cdot .02sec + 200sec \\ & = & 24sec + 200sec \\ & = & 224sec \\ t_3 & = & \hat{P}_4 \cdot 240/\hat{P}_3 \cdot .02sec + 200sec \\ & = & 240 \cdot 240/30 \cdot .02sec + 200sec \\ & = & 240 \cdot 8 \cdot .02sec + 200sec \\ & = & 1920 \cdot .02sec + 200sec \\ & = & 38.4sec + 200sec \\ & = & 238.4sec \end{array}$$

The third line of the calculation of t_1 indicates that six seeks are made to each of the 240 sorted runs during pass 1. Similarly, the third line of the calculation of t_2 indicates that 30 seeks are made to each of 40 sorted runs during pass 2. Finally, the third line of the calculation of t_3 indicates that 240 seeks are made on each of the eight sorted runs during the third and final pass.