External partition element finding

Lars Arge and Michail G. Lagoudakis

November 15, 1999

Algorithm for selecting $\sqrt{m}$ partitioning elements from a set S, where $|S| = N$.

1. Choose a subset G (green) of S of size $\frac{4N}{\sqrt{m}}$ as follows:
   - Load and sort N/M memory loads individually.
   - Pick every $\sqrt{m}/4$'th element from each sorted memory load.
2. Choose a subset R (red) of G of size $\sqrt{m}$ as follows:
   - Use the linear I/O selection algorithm $\sqrt{m}$ times to find every $\frac{4N}{\sqrt{m}}/\sqrt{m} = 4N/m$'th element of G.
3. Return R.

Lemma 1 *The algorithm performs $O(n)$ I/Os.*

Proof: The first step uses $O(|S|/B) = O(N/B) = O(n)$ I/Os. The second step uses

$$\sqrt{m} \cdot O(|G|/B) = \sqrt{m} \cdot O\left(\left(\frac{4N}{\sqrt{m}}\right)/B\right) = O(4N/B) = O(n)$$

I/Os. Overall, the algorithm performs $O(n)$ I/Os. □

Lemma 2 *The number of elements of S between two consecutive elements in R is less than $\frac{3}{2}\frac{N}{\sqrt{m}}$.*

Proof: There are $N/M = n/m$ sorted memory loads. The number of elements of S between two consecutive red elements r_1 and r_2 (r_1, r_2 might come from different memory loads) is bounded by the sum of the following (see Figure 1):

- The number of green elements between r_1 and r_2 which is at most $4N/m$ (because of the way reds were chosen from greens).
- The number of elements of S between two green elements between r_1 and r_2, which is at most

$$\frac{4N}{m}\left(\frac{\sqrt{m}}{4} - 1\right) = \frac{N}{\sqrt{m}} - \frac{4N}{m}$$

  To see this notice that there are $\sqrt{m}/4 - 1$ elements between a pair of consecutive greens in the same memory load. Since there are $4N/m$ greens between r_1 and r_2, there are at most $4N/m$ such pairs.

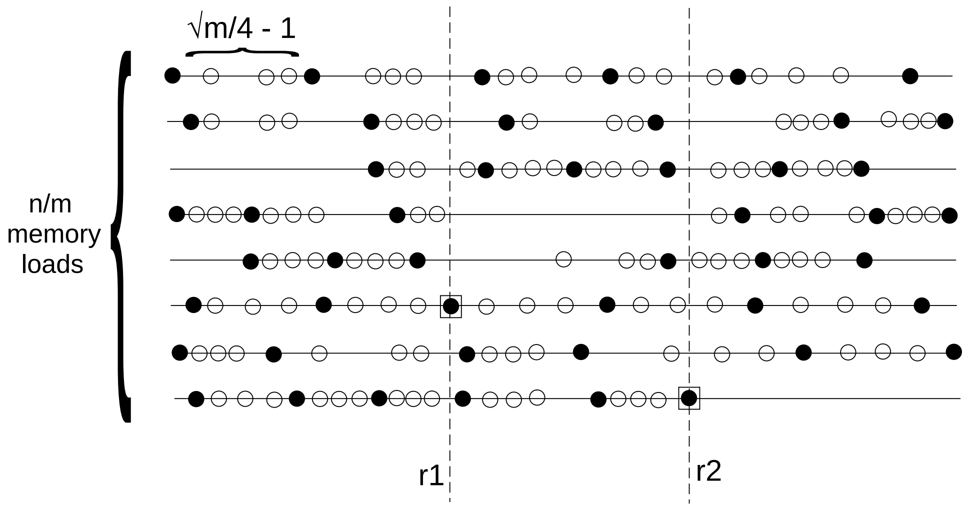

Figure 1: The sorted memory loads are depicted one below the other. Elements of S are shown as circles and their position reflects their rank in the total order. Green elements are shown as solid circles and red elements are enclosed in a square (only two reds (r_1 and r_2) are shown).

- The number of elements of S between r_1 and r_2 but not between two greens (i.e. they are between one green and r_1 or r_2), which is at most

$$2\frac{n}{m}\left(\frac{\sqrt{m}}{4}-1\right)=\frac{n}{2\sqrt{m}}-\frac{2n}{m}$$

  To see this notice that there are two "boundaries" (one defined by $r1$ and one by $r2$) and n/m memory loads. The number of elements of S between one of the boundaries and the closest green is at most $\sqrt{m}/4-1$ (otherwise there would be another green in between) in each memory load.

Summing up the above, we have:

$$\frac{4N}{m}+\frac{N}{\sqrt{m}}-\frac{4N}{m}+\frac{n}{2\sqrt{m}}-\frac{2n}{m} \leq \frac{N}{\sqrt{m}}+\frac{n}{2\sqrt{m}} \leq \frac{N}{\sqrt{m}}+\frac{N}{2\sqrt{m}} = \frac{3}{2}\frac{N}{\sqrt{m}}$$

□