Pairs Approach

Map output records=243241

Map output bytes=2917812

Reduce input records=243241

GC time elapsed (ms)=87972

Physical memory (bytes) snapshot=18800640

Strip Approach

Map output records=540

Map output bytes=2431870

Reduce input records=540

GC time elapsed (ms)=3461

Physical memory (bytes) snapshot=19443712

As we can see from the comparison above, the pairs approach has used more resources in term of cpu (time), but less memory since it its not using hashmaps as in stripe approach.

While we can figure out the huge difference between them in map output records, in stripe approach we save lots of network traffic while in pairs approach we sent 2 emits for each (k,val in w{k}).

The stripe approach maybe more memory consuming, but in our point of view, stripe approach is better in performance and network traffic saving than pairs approach.