Michele Zoncheddu – 545227

Files compressor assignment report

After checked that the mmap don't really map immediately the file in memory, but works with on-demand paging, I left that fast work to the emitter.

The first version is a farm without a collector, whose workers compress the whole file and writes it on the disk.

The second version uses the user-defined parameter BIGFILE_LOW_THRESHOLD (default: 5MB) to split the files bigger than the threshold in chunks at most big as the threshold. The chunks are assigned by the emitter, simply changing the size to read and the starting point of the mapped memory by an offset for each chunk.

The Task structure has now these additional parameters:

- The number of the current chunk (used to write the name of the part)
- The total size of the file (used by the collector to count how many bytes are left)

The workers now compress and write every file or file chunk, unmapping the memory only if the file was not a splitted one. In that case, the task is sent to the collector.

The collector uses a hash map with the name of the file as key, and the pair <remaining_bytes, starting_pointer> as a value, to keep track of all the splitted files and their remaining size to be compressed. When all the parts of a file have arrived, the whole file is unmapped (the starting pointer is saved when the first chunk arrives); then, the tar command to pack all the parts in a single tarball.

The division in chunks has checked to be correct by decompressing the file and comparing the md5 checksum with the original file.

The files to test the application are taken from the Silesia corpus, an open-source compression benchmark counting 12 different files (medical images, executables, databases, text, files, …), with sizes between 6 and 51MB.

I obtained a maximum speedup of 21 with 62 workers, using the blocking mode and the default mapping. With the non-default mapping, the performance were a little worse. The sequential reference times were taken from a sequential version, not from the pipe one. Here are the results.