

Title: Tight Lower Bounds for Query Processing on Streaming and External Memory Data External Memory

Authors: Martin Grohe, Christoph Koch, and Nicole Schweikardt

Abstract: We consider a scenario where we want to query a large dataset that is stored in external memory and does not fit into main memory. The most constrained resources in such a situation are the size of the main memory and the number of random accesses to external memory. We note that sequentially streaming data from external memory through main memory is much less prohibitive. We propose an abstract model of this scenario in which we restrict the size of the main memory and the number of random accesses to external memory, but do not restrict sequential reads. A distinguishing feature of our model is that it admits the usage of unlimited external memory for storing intermediate results, such as several hard disks that can be accessed in parallel. In practice, such auxiliary external memory can be crucial. For example, in a first sequential pass the data can be annotated, and in a second pass this annotation can be used to answer the query. Koch's (VLDB'03) ARB system for answering XPath queries is based on such a strategy. In this model, we prove lower bounds for sorting the input data. As opposed to related results for models without auxiliary external memory for intermediate results, we cannot rely on communication complexity to establish these lower bounds. Instead, we simulate our model by a non-uniform computation model for which we can establish the lower bounds by combinatorial means.

*Last modified: Sun May 1 12:24:52 CEST 2005
Martin Grohe*