CSCI-GA.3033-004 - Graphics Processing Units (GPUs): Architecture and Programming

Antonio Mallia

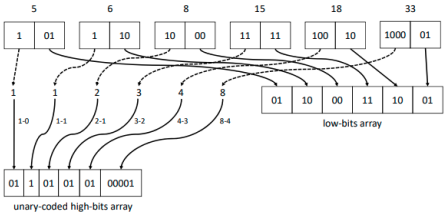
New York University, New York, USA
me@antoniomallia.it

Project proposal

In many important applications - such as search engines - data is stored in the form of arrays of integers. Integer encoding is a fundamental component of any inverted index representation.

The goal of the project is to implement state-of-the-art encoding algorithms to work on GPUs, with a particular focus for decoding speed. We aim to explore the literature and select the most suitable techniques for a parallel implementation. Promising candidates are Elias-Fano, Simple and Binary Packing.

Elias-Fano: To compress a sequence of integers, EF encoding divides each integer into high bits and low bits, and encodes them into the low-bits array and the high-bits array. For the list with n integer elements and U as the maximum possible value, the low-bits array stores the (fixed) $b = \lfloor \log \frac{U}{n} \rfloor$ bits of each element contiguously. . The high-bits array then stores the remaining upper bits (with variable lengths) of each element as a sequence of unary-coded d-gaps of these elements. To decompress these integers, we just need to recover the high bits from the unary-coded d-gaps array, find its corresponding low-bits, and concatenate them.



Simple: Several algorithms including *Simple-9*, *Simple-16*, and *Simple-8b* try to pack as many numbers as possible into one machine word to achieve fast decoding. For example, Simple8b is 64bit word-sized encoder that packs multiple integers into a single word using a 4 bit selector values and up to 60 bits for the remaining values. Integers are encoded using the following table:

Selector	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Bits/element	0	0	1	2	3	4	5	6	7	8	10	12	15	20	30	60
N	240	120	60	30	20	15	12	10	8	7	6	5	4	3	2	1
Unsused bits	60	60	0	0	0	0	12	0	4	4	0	0	0	0	0	0

Binary Packing: Also referred in literature as *PackedBinary*, it uses groups of integers with fixed number of values, where for each group the least value of b that can be used to code the largest element of the group is determined. b is typically store at the beginning of the block as an header, which is then used for decoding.

Experiments

Experiments will be conducted on synthetic data sets first and then results will be confirmed using large realistic data sets based on TREC collections ClueWeb09 and GOV2. The main focus of the experiments will be sequential decoding. A direct comparison with CPU and SIMD-enabled versions of the proposed econdings will be performed.

References

1. Vo Ngoc Anh and Alistair Moffat. 2010. Index compression using 64-bit words. Softw. Pract. Exper. 40, 2 (February 2010), 131-147. DOI=http://dx.doi.org/10.1002/spe.v40:2
2. Sebastiano Vigna. 2013. Quasi-succinct indices. In Proceedings of the sixth ACM international conference on Web search and data mining (WSDM '13). ACM, New York, NY, USA, 83-92. DOI: https://doi.org/10.1145/2433396.2433409