HPCA Submission #XXX– Confidential Draft – Do Not Distribute!!

Enhanced Base-Delta Compression with Data Splitting and Memory Pooling

Aditya Bhandaru (akbhanda@andrew.cmu.edu) Gennady Pekhimenko (akbhanda@andrew.cmu.edu)
Onur Mutlu (akbhanda@andrew.cmu.edu)

Abstract

Recent literature on cache compression has shown great potential for increasing the effective cache capacity on chip. Specifically, a technique called Base-Delta (B+Δ) compression has presented excellent compression (about 1.4X) and improvements in overall performance. However, B+Δ suffers from poor compressibility when adjacent data in memory have a high range in value.

We show here, as proof of concept, that existing techniques such as ***Data Splitting and Memory Pooling can enhance the B+Δ compressibility*** *of data in memory. Our simulations over various micro-benchmarks show that B+Δ with pooling results in an 8% reduction in MPKI, and a compression ratio of 2.6X over the baseline.*

1. Introduction

The memory bottleneck is a well known problem in computer architecture. Caching has become a standard for alleviating contention for data, the bus, and memory. As we trend to more cores, more applications, and larger computing problems, there is a much greater demand for data. Simply scaling cache size to compensate is too expensive, both in power and chip area.

Data compression in the cache is a promising alternative to increasing effective on chip cache capacity. For the same physical cache space, we can store more blocks per set. The ideal cache compression implementation would be fast, simple, and offer a high compression. Many ideas from older literature on cache compression suffer from either poor compression or incur high hardware complexity or long decompression latencies.

Why is fast decompression more important than fast compression? Decompression is on the critical path for a read. In order to supply the requested word, we must decompress the cache line. During a cache fill, compression can occur in the background while we bypass the requested word.

2. Preparation Instructions

2.1. Paper Formatting

All submissions should contain a maximum of 11 pages of single-spaced two-column text and figures excluding references. You can use an unlimited number of extra pages for references. If you are using LaTeX [?] to typeset your paper, then we strongly suggest that you use the template available at http://www.eecg.toronto.edu/~enright/hpca20template.tar.gz – this document was prepared with that template. This document can also be downloaded at http://www.eecg.toronto.edu/~enright/hpca20sample.pdf. If you are using a different software package to typeset your paper, then please adhere to the guidelines given in Table 1.

Field	**Value**
Page limit	**11** pages, **not including references**
Paper size	US Letter 8.5in × 11in
Top margin	1in
Bottom margin	1in
Left margin	0.75in
Right margin	0.75in
Separation between columns	0.25in
Body font	10pt
Abstract font	10pt, italicized
Section heading font	12pt, bold
Subsection heading font	10pt, bold
Caption font	9pt, bold
References	8pt, no page limit list all authors' names

Table 1: Formatting guidelines for submission.

Please ensure that you include page numbers with your submission. This makes it easier for the reviewers to refer to different parts of your paper when they provide comments.

Also, please ensure that your submission has a banner at the top of the title page, similar to this one, which contains the submission number and the notice of confidentiality. If using the template, just replace XXX in the template with the submission number you receive from the submission website.

2.2. Content

~~Author List.~~ All submissions are double blind. Therefore, please do not include any author names in the submission. You must also ensure that the metadata included in the PDF does not give away the authors. If you are improving upon your prior work, refer to your prior work in the third person.

Figures and Tables. Ensure that the figures and tables are legible. Please also ensure that you refer to your figures in the main text. The proceedings will be printed in gray-scale, and many reviewers print the papers in gray-scale. Therefore, if you must use colors for your figures, ensure that the different colors are highly distinguishable in gray-scale. If a figure is not easily understandable in gray-scale, then assume it will not be understood by the reviewers. In many cases, it is better to just prepare your documents without color.

Main Body. Avoid bad page or column breaks in your main text, i.e., last line of a paragraph at the top of a column or first line of a paragraph at the end of a column. If you begin a new section or sub-section near the end of a column, ensure that you have at least 2 lines of body text on the same column.

References. There is no length limit for references. **Each reference must explicitly list all authors of the paper** (no *et al.*). Author of NSF proposals should be familiar with this requirement. Knowing all authors of related work will help find the best reviewers.

3. Submission Instructions

3.1. Paper Authors

Declare all the authors of the paper upfront. Addition/removal of authors once the paper is accepted will have to be approved by the program chair. The paper selection process is carefully run in a way that maximizes fairness by seeking to eliminate all conflicts of interest. Late changes to author lists can invalidate that process.

3.2. Declaring Conflicts of Interest

The authors must register all their conflicts into the paper submission site. Conflicts are needed to resolve assignment of reviewers. Please get the conflicts right. You have several days between the registration of the paper and final submission – there is no need to do the conflicts in a rush at the last second. If a paper is found to have an undeclared conflict that causes a problem, the paper may be rejected.

Please declare a conflict of interest with the following for any author of a paper:

1. PhD advisor
2. Other past or current advisors
3. Current or past students
4. People whom you have collaborated in the last 5 years. This collaboration can consist of a joint research or development project, a joint paper, or when there is direct funding from the potential reviewer (as opposed to company funding) to an author of the paper. Co-participation in professional activities, such as tutorials or studies, is not cause for conflict. When in doubt, the author should check with the PC chair.
5. People with the same current affiliation or who were in the same institution in the last 5 years.
6. Between people whose relationship prevents the reviewer from being objective in his/her assessment. Please be reasonable. For example, just because a reviewer works on similar topics as the paper you are submitting is on, you cannot declare a conflict of interest with them.

All conflicts must be justified. You will have to declare all conflicts with PC members as well as non-PC members with whom you have a conflict of interest. When in doubt, contact the program chair.

3.3. Concurrent Submissions and Resubmissions of Already Published Papers

By submitting a manuscript to HPCA-20, the authors guarantee that the manuscript has not been previously published or accepted for publication in a substantially similar form in any conference or journal. The authors also guarantee that no paper which contains significant overlap with the contributions of the submitted paper is under review to any other conference or journal or workshop, or will be submitted to one of them during the HPCA-20 review period. Violation of any of these conditions will lead to rejection.

Extended versions of papers accepted to IEEE Computer Architecture Letters can be submitted to HPCA-20. If you are in doubt, contact the program chair.

3.4. Submission Site

The submission site is located at `http://hpca20.eecg.toronto.edu/conf`.