Preview of Award 0916027 - Final Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	0916027
Project Title:	III: Small: RIOT: Statistical Computing with Efficient, Transparent I/O
PD/PI Name:	Jun Yang, Principal Investigator
Recipient Organization:	Duke University
Project/Grant Period:	09/01/2009 - 08/31/2014
Reporting Period:	09/01/2013 - 08/31/2014
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

Recent technological advances have enabled collection of massive amounts of data in science, commerce, and society. These large datasets have brought us closer than ever before to solving important problems such as decoding human genomes and coping with climate changes. Meanwhile, the exponential growth in the amount of data has created an urgent challenge. Today, much of advanced analysis is done with programs customdeveloped by statisticians. Unfortunately, progress has been hindered by the lack of easytouse statistical computing environments that scale to large datasets. Many existing tools assume that datasets fit in main memory; when applied to large datasets, they are unacceptably slow because of excessive disk input/output (I/O) operations. There have been many approaches toward I/Oefficiency, but none has gained traction with the statistical computing community. Diskbased storage engines and I/Oefficient function libraries provide only a partial solution, because many sources of I/Oinefficiency in a program remain at a higher, interoperation level: e.g., how large intermediate results are passed between operations, how much performance can be gained by deferring

and reordering operations, etc. Database systems seem to be a natural solution, with I/Oefficiency and a highlevel language (SQL) enabling many highlevel optimizations. However, work in integrating databases and statistical computing has mostly remained databasecentric, forcing statisticians to learn unfamiliar languages and deal with their impedance mismatch with the host language.

To make a practical impact on the statistical computing community, the RIOT project seeks to extend Ran opensource statistical computing environment widely used by statisticiansto transparently provide scalability over large datasets. Transparency means no SQL, or any new language to learn. Transparency means that existing code should run without modification, and automatically gain efficiency. RIOT is developing an endtoend solution that addresses issues on all fronts: I/Oefficient and parallel algorithms, deferred evaluation, pipelined execution, costdriven optimization, smart storage and materialization options, and seamless integration with database systems and the interpreted host language. Over the course of the project, we have also expanded the scope of RIOT to consider issues such as extending database systems to support matrix storage and computation, parallelization using GPU and cloud, and leveraging SSDs for better performance.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

A) RIOT-DB. We believe that many tried-and-true ideas from database research, such as external-memory algorithms, pipelined execution, and cost-based optimization, may be applied to RIOT. To test these ideas, we built RIOT-DB, which maps data and computation in R to an underlying database system. Operations such as multiplication are translated into database view definitions capturing the computation (without actually executing them). This method allows RIOT-DB to build up bigger expressions. When the result is needed, the query defining the view is optimized and executed by the database system, which provides I/O-efficiency. RIOT-DB shows it is indeed possible to achieve I/O-efficiency transparently. On the other hand, RIOT-DB also reveals inadequacies of generic database systems in statistical computation, which we set out to address later in the project. Despite these inadequacies, RIOT-DB significantly outperforms R for large datasets, demonstrating the promise of our approach.

B) RIOT. Based on the lessons from RIOT-DB, we built the next generation of RIOT from the ground up, without the overhead and inadequacies of existing database systems. We also optimize processing across the RIOT/database boundary to enable automatic push down of computation closer to data for efficiency. Specifically: 1) RIOT includes a custom storage manager for arrays (more in (C) below). 2) Relational operators turn out to be inappropriate for many statistical operations; we develop a new expression algebra of both relational and linear algebra operations, which is more amenable to optimization. 3) RIOT allows users to analyze data in databases without 'glue' code; RIOT automatically determines what queries should be pushed down. 4) For expression and database push-down optimization, we develop a combination of heuristics and cost-based techniques. 5) Pipelined

execution in database systems carries significant overhead; we apply compiler techniques to eliminate this overhead (more in (D) below). A prototype was demonstrated in ICDE 2010. Over the years, we have made a number of improvements as further detailed below.

C) RIOT native store. Previous work found databases inadequate in storing arrays, because the relational model does not exploit the ordered nature of arrays. Although B-trees are good at handling sparse arrays, it is difficult to support different array layouts. Furthermore, since dense and sparse arrays require different storage strategies, we need to handle matrices of arbitrary sparsity, which may vary over time and over different regions. We have developed a structure called LAB-tree, which supports flexible matrix layouts and automatically adapts to varying sparsity. We reexamine the leaf splitting strategies and batched update flushing policies. We obtained theoretical and empirical results that contribute to the fundamental understanding of these problems.

D) RIOT Execution and Optimization Issues. Database queries exhibit a limited number of data access patterns. However, linear algebra gives rise to a larger space of access patterns typically expressed as loops. We have developed a framework that allows these patterns to be capturedeither through code analysis or user specificationin a form amenable to automatic optimization. With this framework, we considered how to exploit I/O sharing opportunities. Experiment results show that our optimizer is capable of finding execution plans that exploit nontrivial I/O sharing opportunities with significant savings. We also considered joint optimization of I/O and data layout choices.

E) Parallelization with Cloud. The rise of cloud in recent years offers a promising possibility for bigdata analytics. However, it remains difficult to use the cloud for nontrivial statistical analysis of big data. First, development requires a great deal of expertise and effort. Second, deployment in the cloud is hard, with a maddening array of choices ranging from resource provisioning, software configuration, to execution parameters and implementation alternatives. We have developed a system called Cumulon, as a natural extension of RIOT to the cloud, aimed at helping users rapidly develop and intelligently deploy matrixbased bigdata analysis programs in the cloud. Cumulon is elastic---it can take advantages of the socalled "spot instances," which are cheap but can be taken away when their market prices exceed the bidding prices set by the user. We apply a suite of benchmarking, simulation, modeling, and search techniques to support effective costbased, uncertainty-aware optimization.

F) Extending Database Systems and Parallelization with GPU. In collaboration with HP Labs, we studied how to overcome disadvantages of DBMS in handling matrices. Specifically, we reduce the overhead of element-oriented storage and iterator-based execution by handling matrices in 'chunks' (submatrices). We work around inefficient DBMS execution using optimized numerical libraries on a per-chunk basis. We avoid the awkwardness of SQL with user-defined functions. We developed

and evaluated alternative matrix linearization and chunking schemes, and showed superior flexibility and performance. We also considered different strategies for implementing matrix operations within PostgreSQL. Overall, we have demonstrated that highly efficient matrix computation and GPU acceleration can be possible within a DBMS. Our findings were summarized in a paper titled 'MaSSA (Massive-Scale Statistical Analysis) DBMS,' presented at HP Tech Con 2011, and a provisional patent has been filed.

G) Working with SolidState Drives (SSDs). Solidstate drives are becoming a viable alternative to magnetic disks for many workloads. We have considered them for various database and linear algebra workloads. We designed an index structure called FD+tree and an associated concurrency control scheme called FD+FC. We then studied the problem of permuting (and resorting) on SSDs (one application being matrix layout conversion). While external merge sort is often used for permutation, it is an overkill that fails to exploit the property of permutation and carries unnecessary overhead in storing and comparing keys. We proposed faster algorithms with lower memory requirements for a large, useful class of permutations. We also tackled practical challenges, such as the cost asymmetry between reads and writes. Finally, we investigated the problem of optimizing the merge operation fundamental to many multi-level index structures such as LSM trees and FD+trees. We considered intelligent strategies that selectively merges portions of levels in a way that avoids writing new blocks as much as possible.

H) Educational Activities. In Fall 2011, 2012, and 2013, the PI taught the undergraduate database course at Duke University. He won the David and Janet Vaughan Brooks Teaching Award in 2013. In Spring 2010, he introduced a new graduate course titled 'Database and Programming Languages: Crossing the Chasm,' which covered research related to RIOT. In Spring 2012, he introduced a new graduate course, 'Projects in Computational Journalism,' devoted to the nascent research area of computational journalism. In Spring 2014, he introduced a new undergraduate course, 'Everything Data,' aimed at exposing students to various aspects of working with data---acquisition, integration, querying, analysis, and visualization---and data of different types---from unstructured text to structured databases.

During this project, the PI supervised a number of undergraduate researchers: Weiping Zhang helped with (C); Jiaqi Yan studied RIOT applications and helped with (F); Andrew Shim, Emre Sonmez, and Seokhyun Song worked on applications of data analysis to fact-checking. The PI supervised three PhD researchers: Yi Zhang is the lead student on the project and graduated with a PhD in spring 2012; Risi Thonangi and Botong Huang, who work primarily on (G) and (E) respectively, continue to make solid progress in their PhD study.

Specific Objectives: See the section of this report on "Major Activities."

Significant Results: A) RIOT-DB. We completed the proof-of-concept implementation of RIOT on top of a database system, and published the RIOT 'vision' paper in CIDR 2009. The code for RIOT-DB is publicly available on the project Web site.

B) RIOT. We presented a demonstration of a prototype for the next generation of RIOT in ICDE 2010.

C) RIOT native store. We have developed a B-tree variant called the Linearized Array B-tree, or LAB-tree, which supports flexible matrix layouts and automatically adapts to varying sparsity across parts of a matrix and over time. We developed new leaf splitting strategies and batched update flushing policies. A paper on the RIOT storage engine was published in PVLDB 2011.

D) RIOT Execution and Optimization Issues. We successfully applied compiler techniques to the problem of optimization I/O for linear algebra workloads. Our new framework on exploiting I/O sharing opportunities was published in PVLDB 2012. Extension of this work is published as part of Yi Zhang's PhD dissertation (2012).

E) Parallelization with Cloud. We have completed a two iterations of implementation of Cumulon on top of Amazon EC2, which significantly improves the ease and efficiency of development and deployment of matrixbased statistical computing workloads in the cloud. Our first paper on Cumulon was published in SIGMOD 2013. We also presented Cumulon at the IBM Workshop on Big Data Analytics in June 2013. An overview/vision paper was published in a special issue of the IEEE Data Engineering Bulletin (2014). Results on supporting elasticity with spot instances are now under submission.

F) Extending Database Systems and Parallelization with GPU. We have presented our results at HP Tech Con 2011, and filed a provisional patent (with HP Labs) on GPUbased matrix computation inside a database system.

G) Working with SolidState Drives (SSDs). We developed a new index structure called FD+tree for the SSDs and the associated concurrency control scheme a paper on efficient; the work was published in CIKM 2012. Results on permuting and re-sorting data on SSDs (with application in matrix lay conversion and beyond) was published in PVLDB 2013; a journal version is currently under preparation. Our result on merge optimization for multi-level index structures is also under preparation for submission.

Key outcomes or Other achievements: See the section of this report on "Significant Results."

*** What opportunities for training and professional development has the project provided?**

Ph.D. students: Yi Zhang, Risi Thonangi, Botong Huang, Herodotos Herodotou. Undergraduate students: Weiping Zhang, Jiaqi Yan, Andrew Shim, Emre Sonmez, Seokhyun Song.

These students have gained research experience in algorithms, compilers, databases, high-performance computing, and programming languages.

*** How have the results been disseminated to communities of interest?**

Most of our research results have been published and presented at

reputable international venues for database research. See the section of this report on "Significant Results" for more details. It is worth noting that in addition to publications at academic venues, we have also used several other methods for dissemination in this project: 1) The RIOT-DB code is publicly available. 2) We presented a software demonstration of RIOT at ICDE 2010. 3) We filed a provisional patent with HP Labs. 4) We disseminated our results at industry venues, including the HP Tech Con 2011 and IBM Workshop on Big Data Analytics in 2013. 5) In terms of educational materials, all PI's course materials are posted online and open to the general public.

Products

Books

Book Chapters

Conference Papers and Presentations

Risi Thonangi, Shivnath Babu, and Jun Yang (2012). *A practical concurrent index for solid-state drives*. 2012 International Conference on Information and Knowledge Management. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Botong Huang, Shivnath Babu, and Jun Yang (2013). *Cumulon: optimizing statistical data analysis in the cloud*. 2013 ACM SIGMOD International Conference on Management of Data. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Yi Zhang, Weiping Zhang, and Jun Yang (2010). *I/O-efficient statistical computing with RIOT*. 2010 International Conference on Data Engineering. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Yi Zhang, Herodotos Herodotou, and Jun Yang (2009). *RIOT: I/O-efficient numerical computing without SQL*. 2009 Conference on Innovative Data Systems Research. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Inventions

Journals

Botong Huang, Nicholas W. D. Jarrett, Shivnath Babu, Sayan Mukherjee, and Jun Yang. (2014). Cumulon: cloud-based statistical analysis from users perspective. *IEEE Data Engineering Bulletin*. 37 (3), 77-89. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Risi Thonangi and Jun Yang (2013). Permuting data on random-access block storage. *Proceedings of the VLDB Endowment*. 6 (9), 721-732. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Yi Zhang and Jun Yang (2012). Optimizing I/O for big array analytics. *Proceedings of the VLDB Endowment*. 5 (8), 764-775. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Yi Zhang, Kamesh Munagala, and Jun Yang (2011). Storing matrices on disk: theory and practice revisited. *Proceedings of the VLDB Endowment*. 4 (11), 1075-1086. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Licenses

Other Products

Other Publications

Patents
Two-Level Chunking for Data Analytics. UNITED STATES. Application Date = 03/16/2012. Status = Submitted

Technologies or Techniques

Thesis/Dissertations
Yi Zhang. *Transparent and efficient I/O for statistical computing*. (2012). Duke University. Acknowledgement of Federal Support = Yes

Websites

http://db.cs.duke.edu/projects/riot

Participants/Organizations

Research Experience for Undergraduates (REU) funding

Form of REU funding support:	REU supplement
How many REU applications were received during this reporting period?	2
How many REU applicants were selected and agreed to participate during this reporting period?	2
REU Comments:	

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Yang, Jun	PD/PI	6
Huang, Botong	Graduate Student (research assistant)	9
Thonangi, Risi	Graduate Student (research assistant)	11
Shim, Andrew	Research Experience for Undergraduates (REU) Participant	1
Song, Seokhyun	Research Experience for Undergraduates (REU) Participant	3
Sonmez, Emre	Research Experience for Undergraduates (REU) Participant	2

Full details of individuals who have worked on the project:

Jun Yang
Email: junyang@cs.duke.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 6

Contribution to the Project: Jun Yang has been serving as the faculty lead on this project.

Funding Support: NSF IIS-0713498, NSF IIS-1320357, and 2010 HP Labs Innovation Research Award

International Collaboration: Yes, China
International Travel: Yes, Italy - 0 years, 0 months, 7 days; Mexico - 0 years, 0 months, 7 days

---

Botong Huang
Email: bhuang@cs.duke.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 9

Contribution to the Project: Botong Huang is investigating how to run RIOT on data-parallel cloud computing platforms such as Hadoop MapReduce and Amazon EC2. He has developed Cumulon, which aims at making R-style, matrix-based statistical analysis easier to develop and deploy in the cloud.

Funding Support: NSF IIS 0917062, 0964560, and 1320357.

International Collaboration: No
International Travel: Yes, China - 0 years, 1 months, 0 days

Risi Thonangi
Email: rvt@cs.duke.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 11

Contribution to the Project: Risi Thonangi has been researching how to leverage the recently emerging SSDs (solid state drives) to improve performance.

Funding Support: No other funding sources.

International Collaboration: No
International Travel: Yes, Italy - 0 years, 0 months, 7 days; India - 0 years, 1 months, 0 days

---

Andrew Shim
Email: andrew.shim@duke.edu
Most Senior Project Role: Research Experience for Undergraduates (REU) Participant
Nearest Person Month Worked: 1

Contribution to the Project: Andrew Shim is an undergraduate student who started working with the PI in Spring 2013. He has been looking at how quantitative analysis of data can be used in the context of journalism and fact-checking.

Funding Support: REU supplement associated with this grant.

International Collaboration: No
International Travel: No
Year of schooling completed: Junior
Home Institution: Duke University
Government fiscal year(s) was this REU participant supported: 2014

Seokhyun Song
Email: seokhyun.song@duke.edu
Most Senior Project Role: Research Experience for Undergraduates (REU) Participant
Nearest Person Month Worked: 3

Contribution to the Project: Seohyun (Alex) Song is an undergraduate student who started working with the PI in Spring 2014. He has been looking at how quantitative analysis of data can be used in the context of journalism and fact-checking.

Funding Support: REU supplement associated with this grant.

International Collaboration: No
International Travel: No
Year of schooling completed: Junior
Home Institution: Duke University
Government fiscal year(s) was this REU participant supported: 2014

Emre Sonmez
Email: emre.sonmez@duke.edu
Most Senior Project Role: Research Experience for Undergraduates (REU) Participant
Nearest Person Month Worked: 2

Contribution to the Project: Emre Sonmez is an undergraduate student who started working with the PI in Spring 2014. He has been looking at how quantitative analysis of data can be used in the context of journalism and fact-checking.

Funding Support: REU supplement associated with this grant.

International Collaboration: No
International Travel: No
Year of schooling completed: Freshman
Home Institution: Duke University
Government fiscal year(s) was this REU participant supported: 2014

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
HP	Industrial or Commercial Firms	HP Labs in Beijing, China

Full details of organizations that have been involved as partners:

HP

Organization Type: Industrial or Commercial Firms
Organization Location: HP Labs in Beijing, China

Partner's Contribution to the Project:
Financial support
Collaborative Research

More Detail on Partner and Contribution: HP Labs provided additional funding to RIOT in Year 2 through their

Innovation Research Program. The PI visited HP Labs in Beijing in July 2010 to help jump-start the collaboration. The collaboration focused on the issues of database extensibility and parallelization, at both GPU and cloud level. See the section of this report on research and education activities for details.

Have other collaborators or contacts been involved? No

Impacts

What is the impact on the development of the principal discipline(s) of the project?

We have made a series of solid contributions toward enabling efficient statistical analysis over massive datasets. We have built three functional prototypes, RIOTDB, RIOT, and Cumulon, and published in CIDR 2009, ICDE 2010, PVLDB 2011, PVLDB 2012, CIKM 2012, SIGMOD 2013, PVLDB 2013, and IEEE Data Engineering Bulletin 2014. For detailed descriptions of these contributions, please refer to the sections of this report on "Major Activities" and "Significant Results."

What is the impact on other disciplines?

The PI has been part of other interdisciplinary projects funded by NSFone studied how to collect and analyze ecological data from a sensor network; one is investigating how to simplify the development and deployment of statistical data analysis in a cloud; a third is considering perturbation analysis of data queries with applications such as public policy. Much of the work in this project is motivated by the ecological data analysis problems faced in the first project on sensors, while many of the results from this project are now being applied in the second and third projects to problems in statistics, political science, and public policy.

What is the impact on the development of human resources?

Ph.D. students: Yi Zhang, Risi Thonangi, Botong Huang, Herodotos Herodotou. Undergraduate students: Weiping Zhang, Jiaqi Yan, Andrew Shim, Emre Sonmez, Seokhyun Song.

What is the impact on physical resources that form infrastructure?
Nothing to report.

What is the impact on institutional resources that form infrastructure?
Nothing to report.

What is the impact on information resources that form infrastructure?
Nothing to report.

What is the impact on technology transfer?

The PI received an HP Labs Innovation Research Award for work on RIOT. The work on integrating RIOT within GPUenabled DBMS was presented at HP Tech Con 2011, and a provisional patent has been filed. The work on statistical data analysis in the cloud was presented at the IBM Workshop on Big Data Analytics in 2013.

What is the impact on society beyond science and technology?
Nothing to report.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.