
# FastBit: An Efficient Indexing Technology For Data-Driven Science

FastBit implements a set of state-of-art bitmap indexing technologies. A key innovation is an efficient compression technique that is 10 times faster than commercially available ones. FastBit has been demonstrated to significantly speed up distributed data analysis, query-driven visualization, network traffic analysis, and drug discovery.

**John Wu, Kurt Stockinger, Ekow Otoo, Doron Rotem, Arie Shoshani**
Scientific Data Management, Berkeley Lab

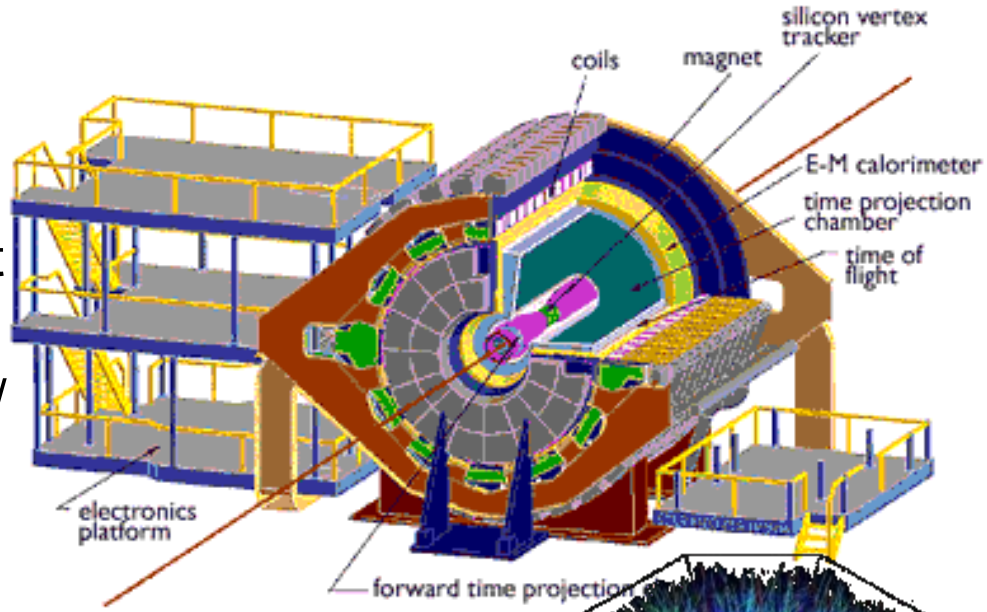
http://sdm.lbl.gov/fastbit

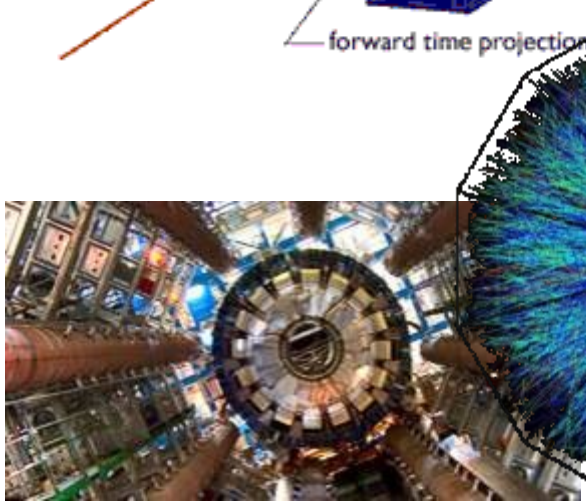
SDM
U.S. Department of Energy Contract No. DE-AC02-05CH11231
BERKELEY LAB

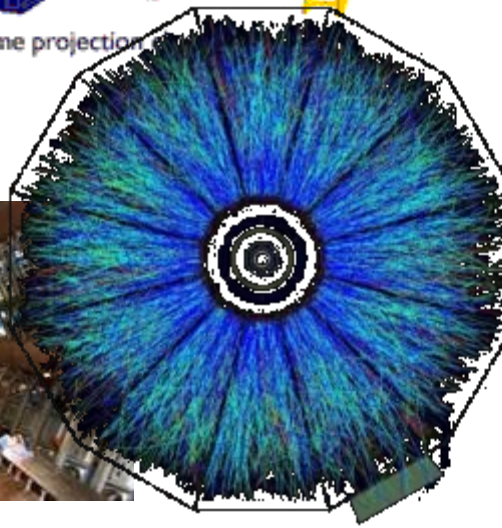
# FastBit Started In a Big Smash

Searching for clues of Quark-Gluon Plasma in a large set of high-energy collision data

- High-Energy Physics experiment STAR
- 600 participants / 50 institutions / 12 countries
- Data rate 200 MB/s
- Data collected 5 PB
- ~ 1 Billion collision events, 5 MB per event (equivalent to having millions of variables)
- Challenge: finding 100 or so events with the best evidence of QGP



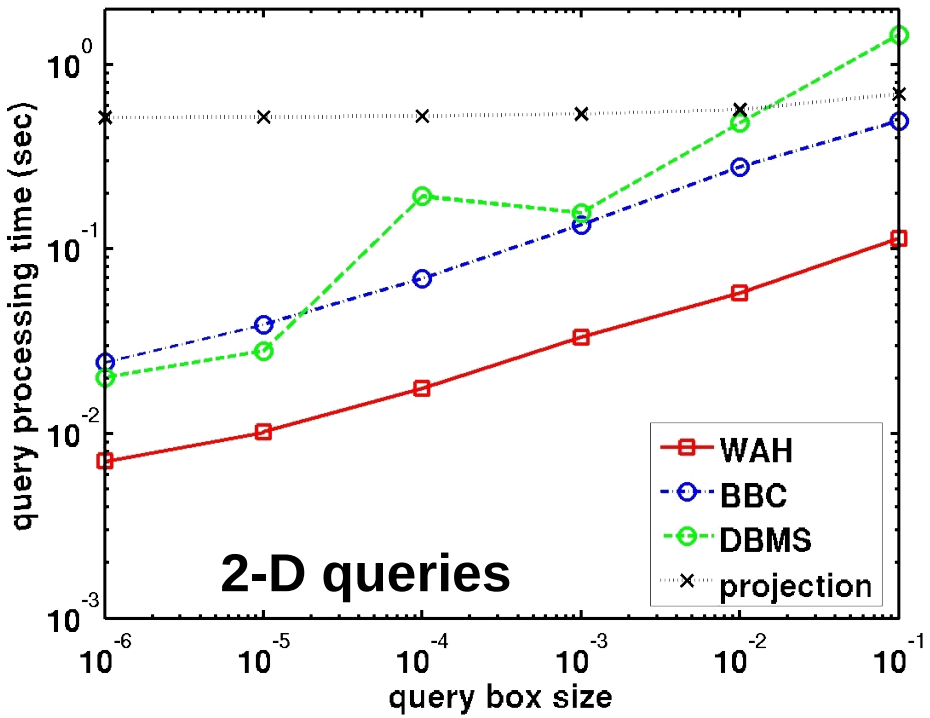



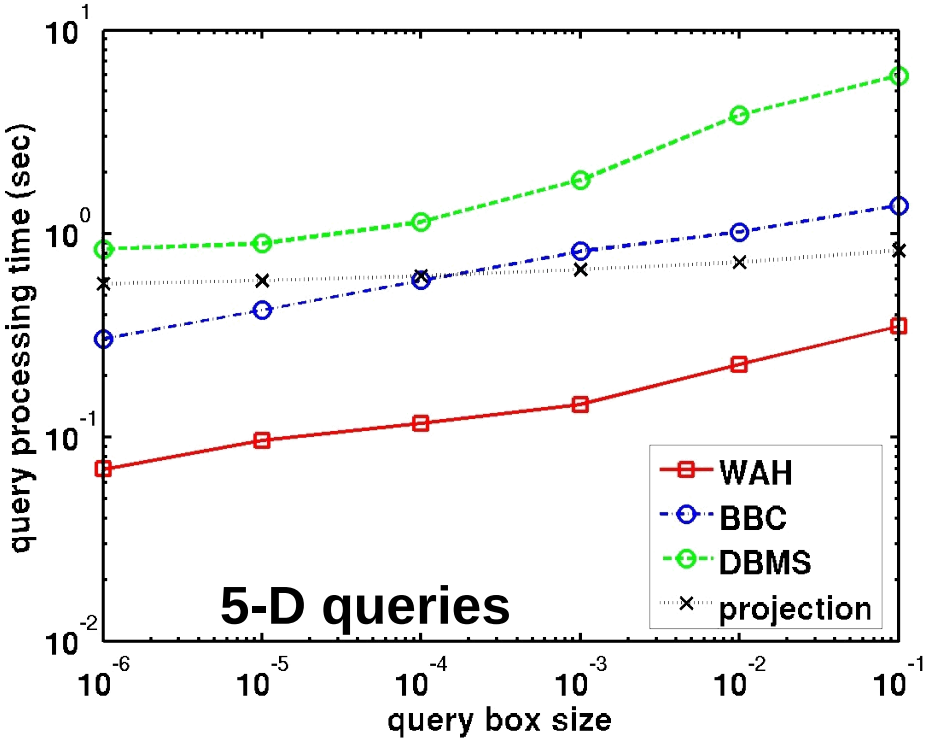


**Best Paper Award (ISC 05)**

**[Wu, Gu, Lauret, Poskanzer, Shoshani, Sim and Zhang 2005]**

# FastBit 10x Faster than DBMS



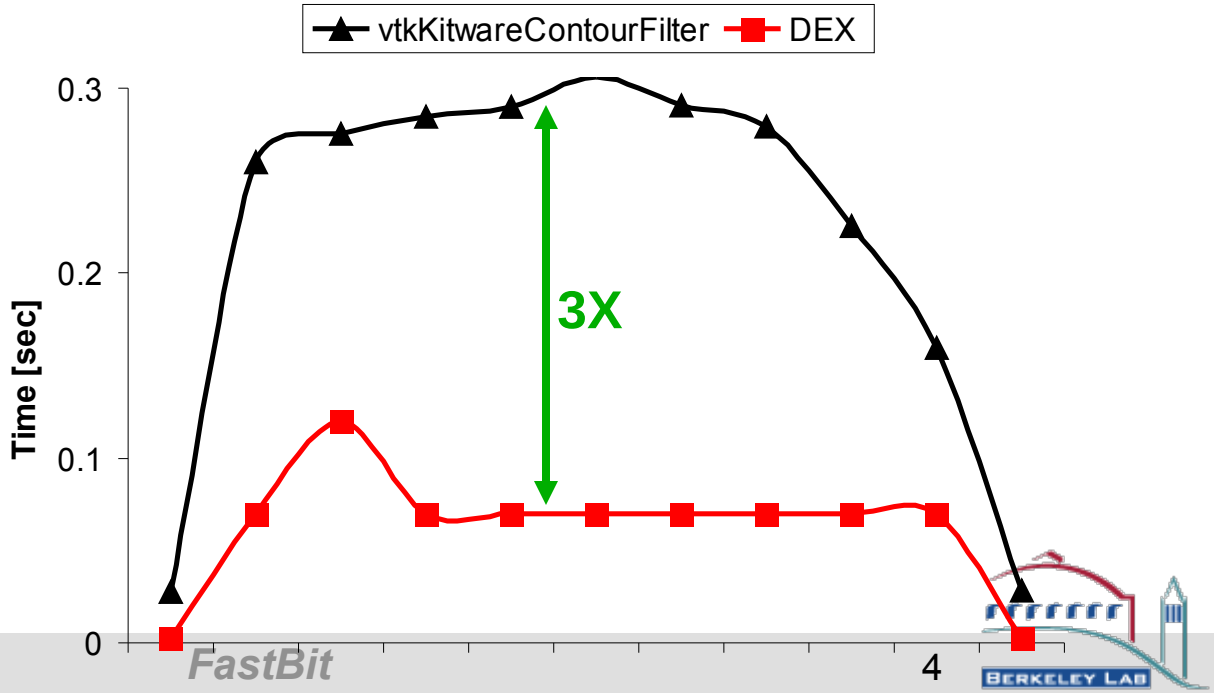




- Queries on 12 most queried attributes (2.2 million records) from STAR High-Energy Physics Experiment, average attribute cardinality 222,000
- Experiments confirm that:
  - WAH compressed indices are 10X faster than bitmap indices from a DBMS (using BBC), 5X faster than our own implementation of BBC
  - Size of WAH compressed indices is only 30% of raw data size (a popular DBMS system uses 3-4X for B+-tree indices)

[Wu, Otoo, Shoshani 2001]

# FastBit Finds Volumes Faster Than Best Isocontour Finder

- FastBit finds volume of interest efficiently with compressed representation of the volume
- FastBit identifies volumes of interest as efficient as the best algorithm that identify the surface only (isocontouring), in theory
- FastBit is three times faster than the best isocontouring algorithm in VTK

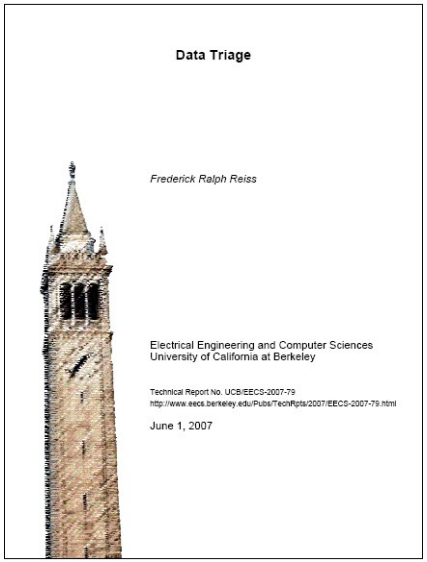



[Wu, Koegler, Chen, Shoshani 2003]

[Stockinger, Shalf, Bethel, Wu 2005]

# FastBit Milestones

- *2007/08*: FastBit speed up drug discovery tool (first publication not involving any FastBit developers)
- *2007/08*: First public release, version a0.7
- *2007/06*: Physical design reviewed
- *2007/06*: First PhD thesis involving FastBit completed
- *2006/03*: Prove formal optimality
- *2006/02*: Work on Enron data made headline at PRIMEUR
- *2005/05*: Appeared in ACM TechNews
- *2005/05*: Grid Collector wins ISC Award
- *2005/01*: CRD news report on FastBit
- *2004/12*: WAH patent issued


Data Triage

Frederick Ralph Reiss

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-79
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-79.html

June 1, 2007


# FastBit: An Efficient Indexing Technology For Data-Driven Science

Recent advances

- Two-level encoding
- Feature identification on toroidal mesh

http://sdm.lbl.gov/fastbit

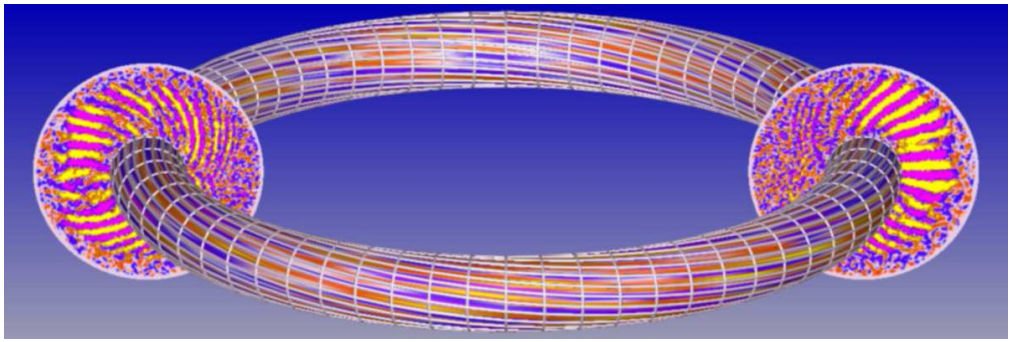
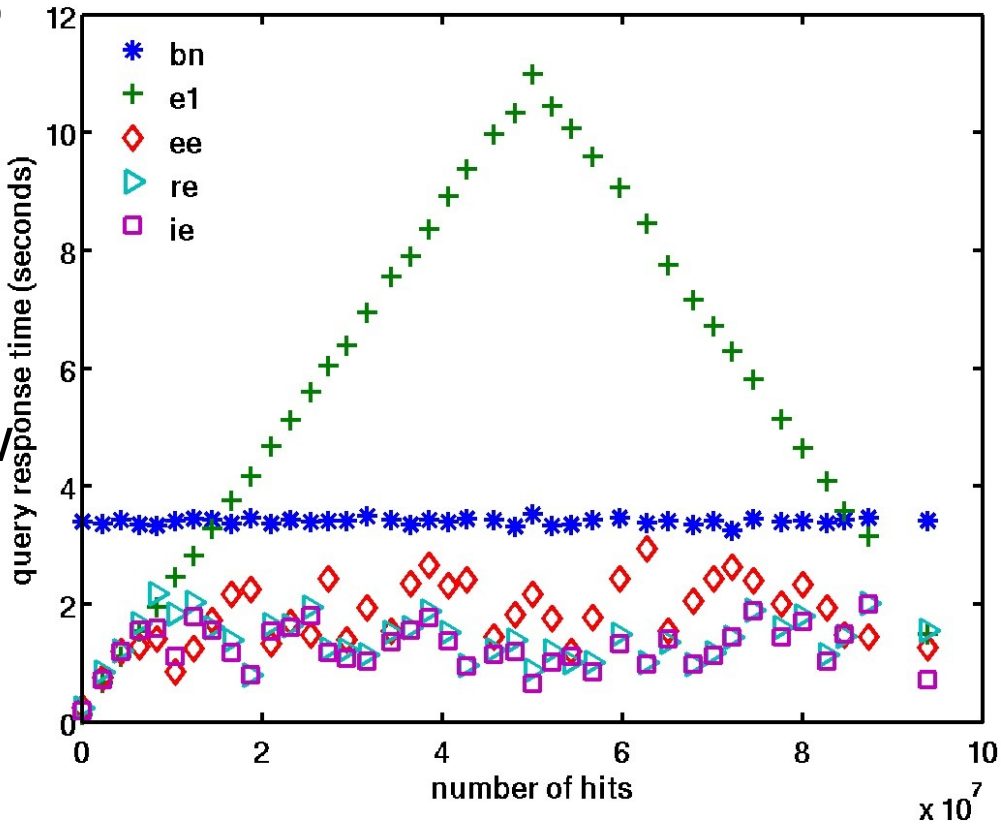


**Photo credit:** Crawford, Ma, Huang, Klasky, Ethier, 2004

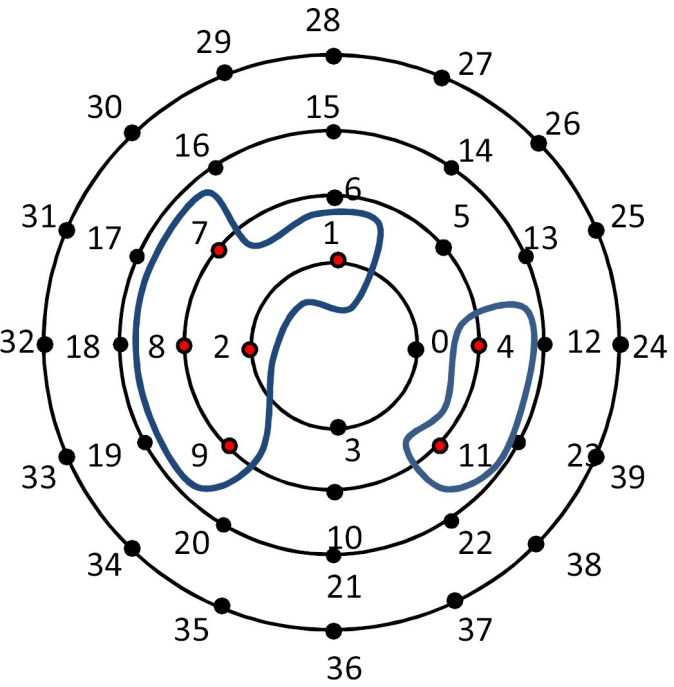
# Two Levels Are Better Than One

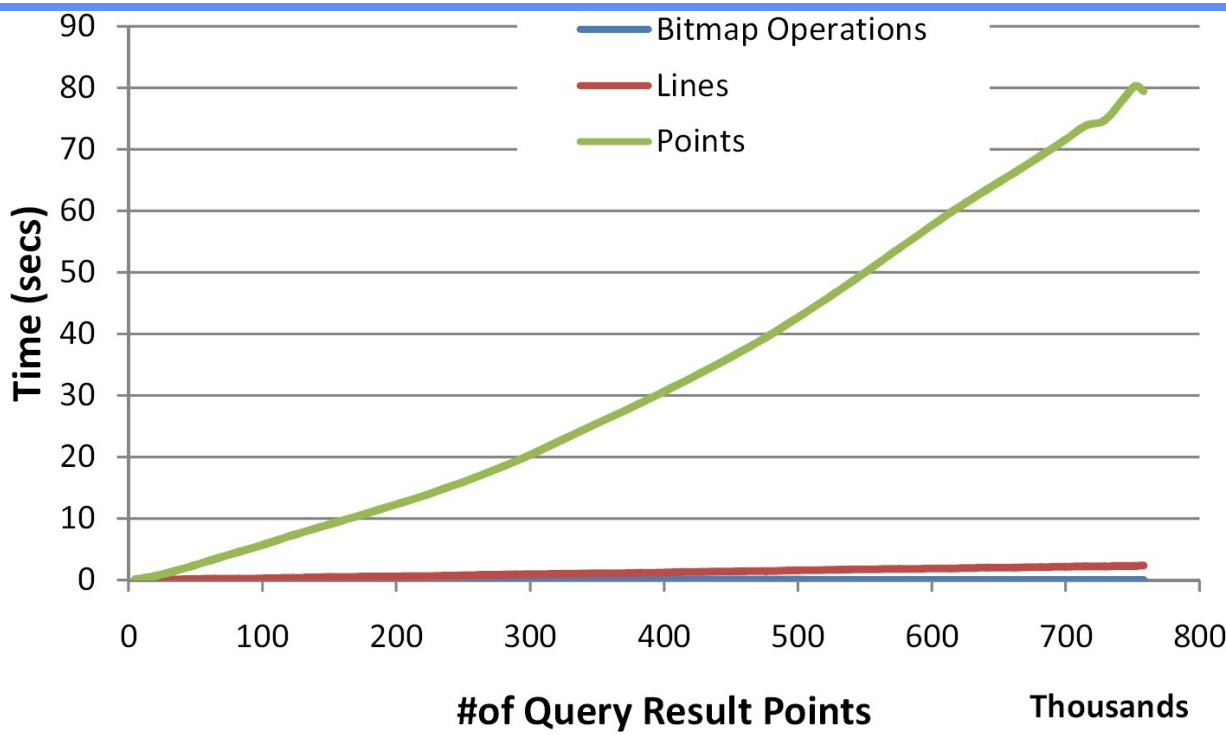
- Most commonly used bitmap index is one-level equality encoded (e1)
- Multi-level encoding was postulated to improve performance, but no satisfactory guidance on how to do it [Wu, Otoo, Shoshani, 2000] [Sinha, Winslett, 2007]
- We found the optimal parameters and three efficient two-level encodings: equality-equality encoding, range-equality-encoding, and interval-equality encoding [Wu, Stockinger, Shoshani]




bn = binary encoding
e1 = one-level equality
ee = equality-equality
re = range-equality
ie = interval-equality

# Feature Identification on Toroidal Mesh




- Two ways to speed up the feature identification
  - work with lines instead of points
  - use an efficient connected component labeling algorithm
- 10 – 100 times faster than working with points [Sinha, Winslett, Wu]