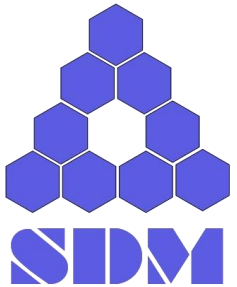
FastBit: an indexing technology for data-intensive science

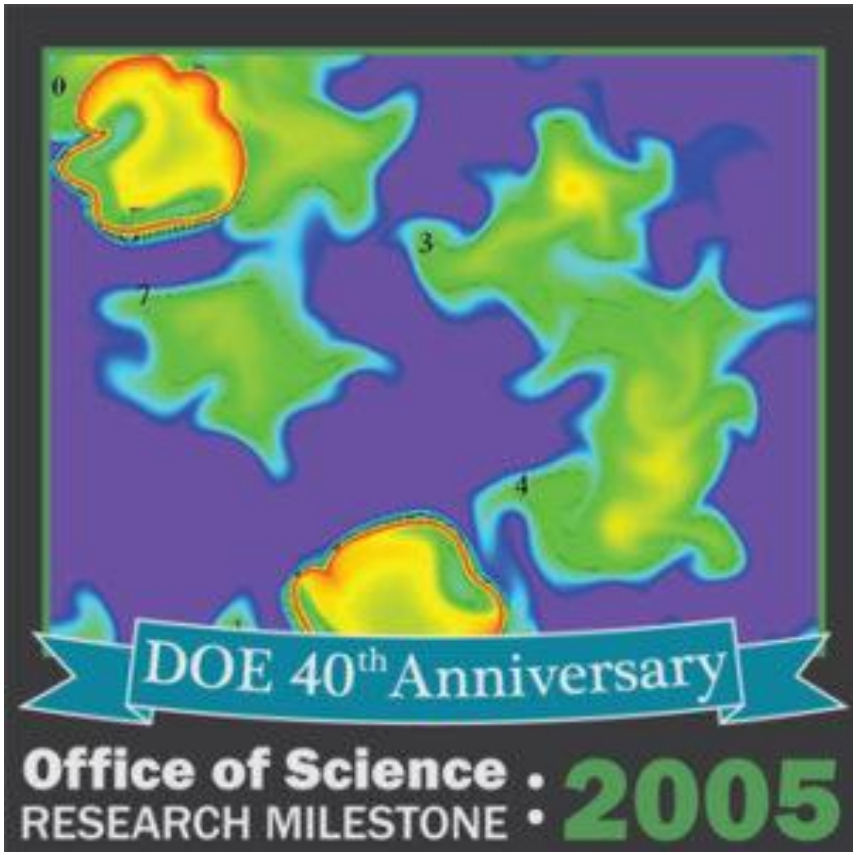
John Wu

U.S. DEPARTMENT OF ENERGY | Office of Science

Lawrence Berkeley National Laboratory



FastBit: an efficient indexing technology for accelerating data-intensive science



DOI: 10.1088/1742-6596/16/1/077

- **Compressed bitmap index:** Ekow Otoo and Arie Shoshani
- **Grid Collector:** Jerome Lauret, Wei-Ming Zhang, Alexander Sim, Junmin Gu, Arie Shoshani, Arthur Poskanzer, and Victor Perevoztchikov
- **DEX**: Kurt Stockinger, John Shalf, Wes Bethel, Wendy Koegler, Jacqueline Chen and Arie Shoshani

Big Science Data Example: Around 2000

Collaboration	# members /institutions	Date of first data	# events/year	total data volume/year-TB
STAR	350/35	2000	10^7-10^8	300
PHENIX	350/35	2000	10^9	600
BABAR	300/30	1999	10^9	80
CLAS	200/40	1997	10^{10}	300
ATLAS	1200/140	2004	10^9	2000

STAR: Solenoidal Tracker At RHIC

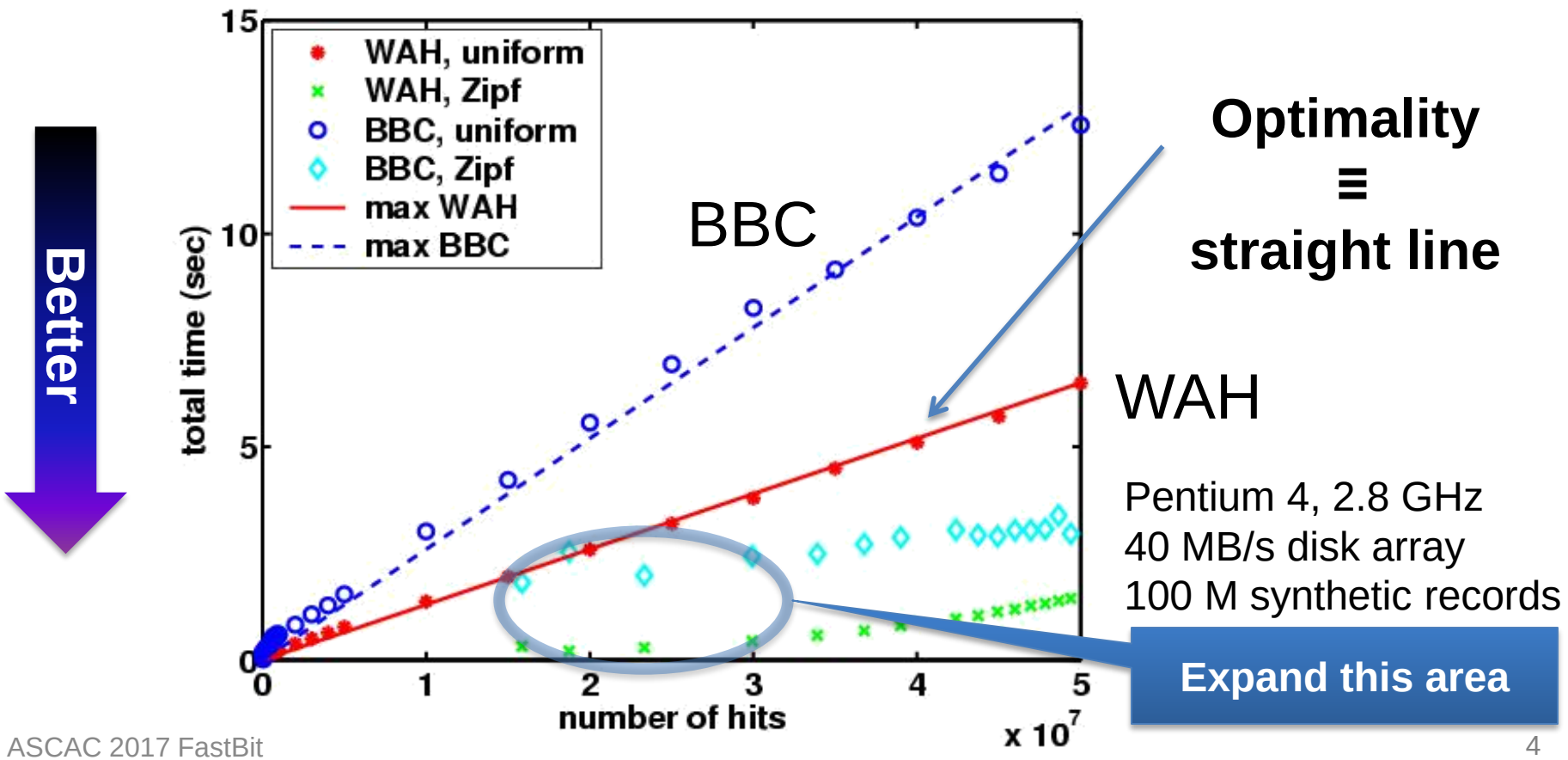
RHIC: Relativistic Heavy Ion Collider

Key science goal: find evidence of quark-gluon plasma

- **Observation:** the evidence might be in a few thousand collision events
- **Needle in the haystack** type of problem

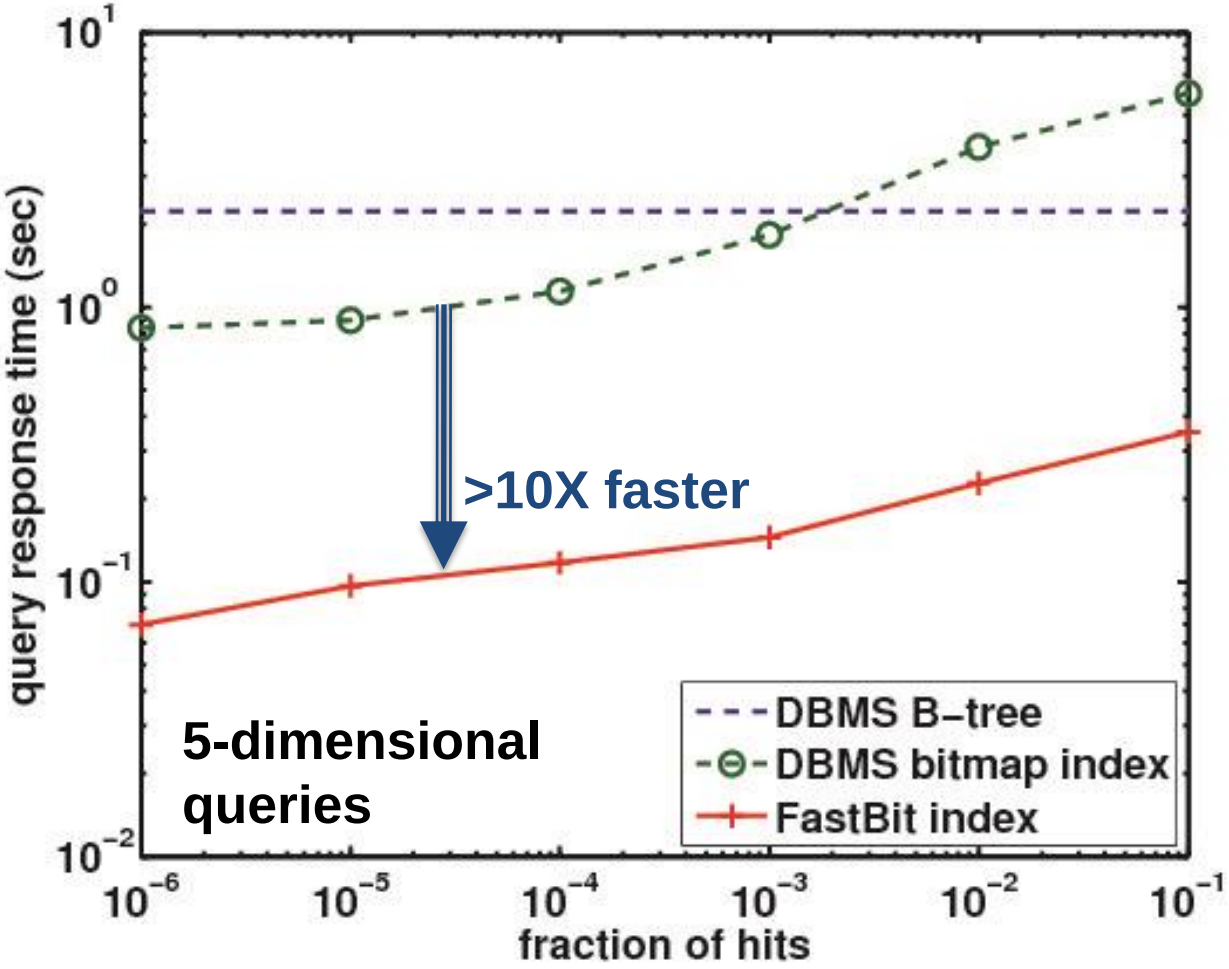
WAH Compressed Index Is Optimal

- In the worst case, query response time is a linear function of the number of hits, H
- WAH Compressed indexes are optimal for one-dimensional range queries, search time O(H)



Multi-Dimensional Query Performance

- Queries 5 out of 12 most popular variables from STAR (2.2 million records)
- Average attribute cardinality (distinct values): 222,000
- FastBit uses WAH compression
- DBMS uses BBC compression
- FastBit >10X faster than DBMS
- FastBit indexes are 30% of raw data sizes



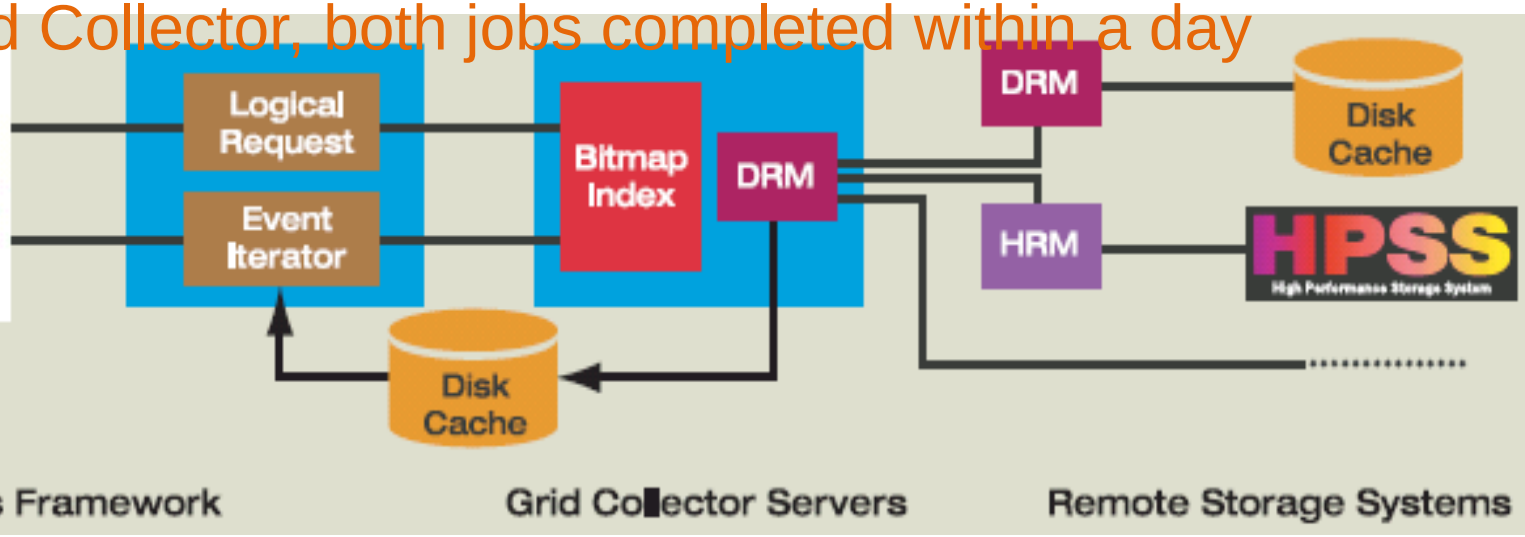
[Wu, Otoo and Shoshani 2002]

FastBit in STAR

- Searching for anti-^{3}He
- Lee Barnby, Birmingham, UK
- Previous studies identified collision events that possibly contain anti-^{3}He, need further analysis

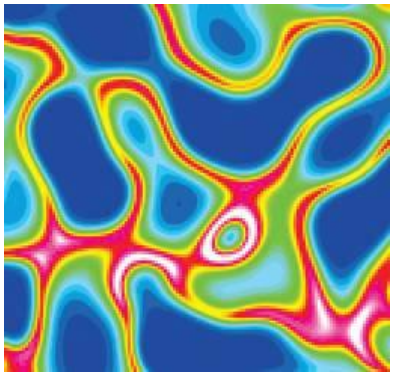
- Searching for strangelet
- Aihong Tang, BNL
- Previous studies identified events that behave close to strangelets, need further investigation

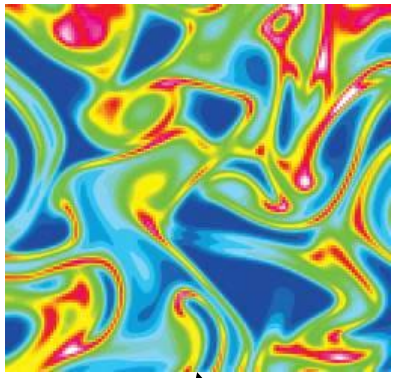
- Without Grid Collector, one has to retrieve many files from mass storage systems and scan them for the wanted events – may take weeks or months, no one wants to actually do it
- With Grid Collector, both jobs completed within a day

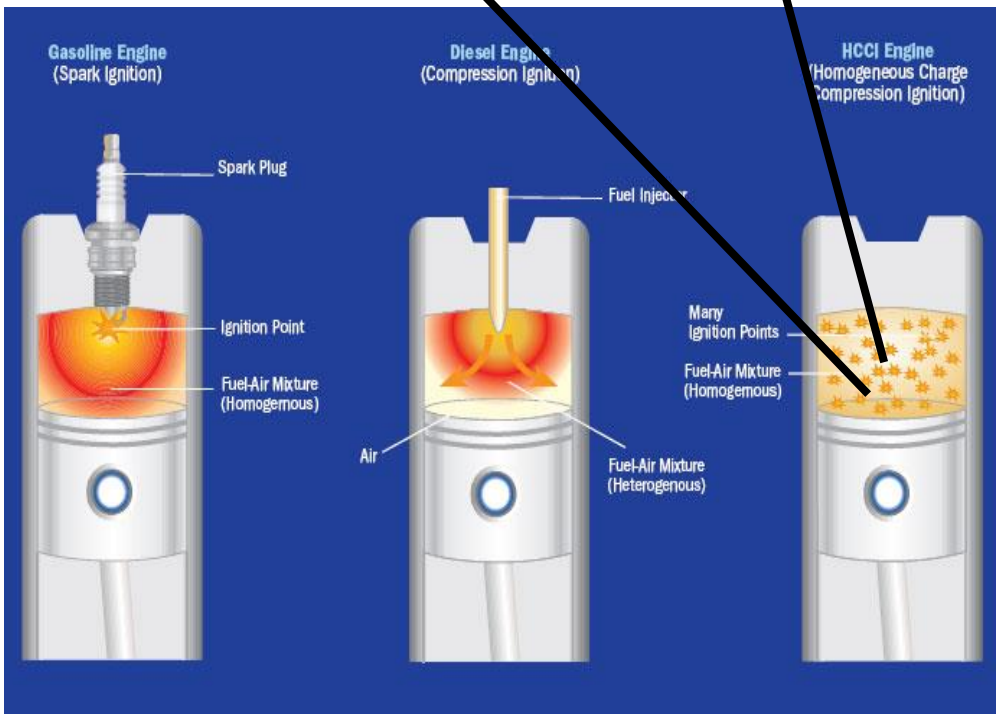


FastBit in Combustion

- Searching for a more fuel efficient combustion engine (Homogeneous-Charge Compression Ignition engine)
  - Require detailed numerical simulation with hundreds of variables
  - Simulation mesh: 1000 x 1000 x 1000
  - 1000s time steps per simulation
  - Challenge: finding and tracking ignition kernels



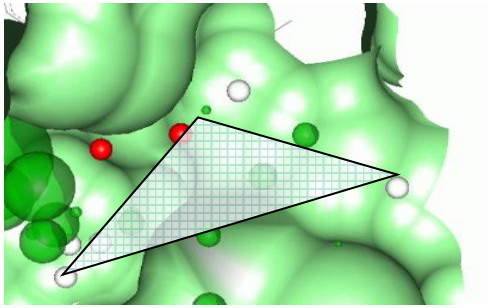


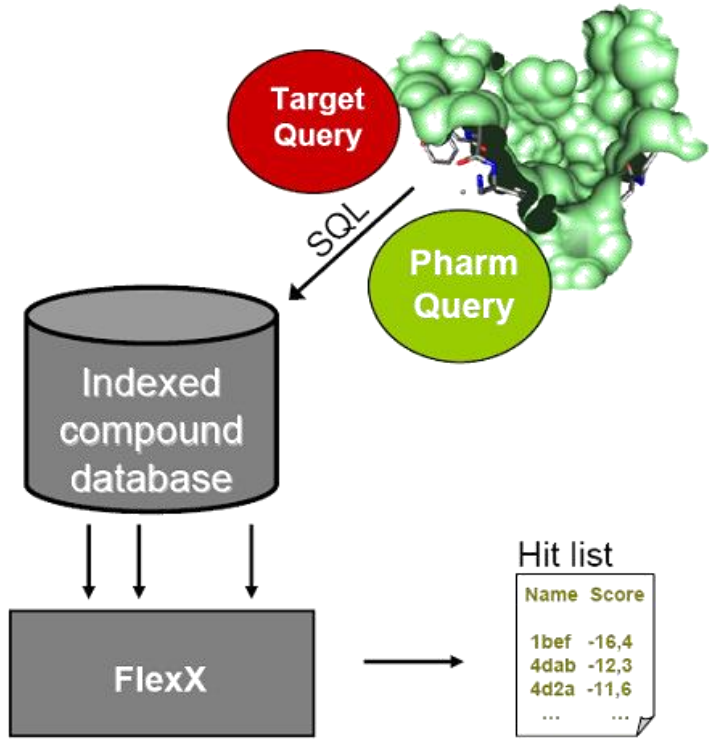


Use of FastBit for Molecular Docking

Method

- **Specification of the descriptor as triangle geometry**
  - Types of interaction centers
  - Triangle side lengths
  - Interaction directions
  - 80 bulk dimensions
- **Receptors**
  - Receptor descriptors are generated similarly
  - Using complementary information where necessary
- **Use of pharmacophore constraints on receptor triangles**
  - Reduces number of queries
  - Improved query selectivity because the pharmacophore tends to be inside the protein cavity

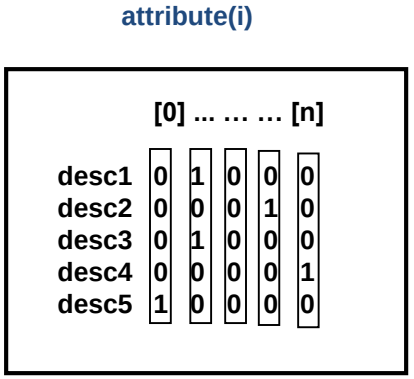




Use of FastBit for Molecular Docking

Method

- Indexing system
  - Properties of the problem:
  - Billions of descriptors (~ 1,000 for each ligand)
  - High dimensional query
- Properties of bitmap indexes
  - Well suited for those kind of queries
  - Can be run stand alone
  - Further compression possible
  - FastBit uses compression



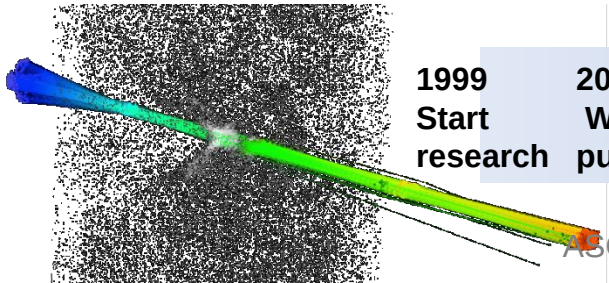
Bitmap index

Results

- TrixX-BMI is an efficient tool for virtual screening with average runtime in sub-second range
- screen libraries of ligands 12 times faster than FlexX without pharmacophore constraints
- **With pharmacophore constraints, speedup 140 – 250**

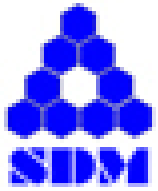
Long History of DOE Support

- **Problem**: given a large data collection, quickly find records satisfying user-specified conditions
  - Example: in billions of high-energy collision events, find a few thousand based on energy level, number of particles and so on
- **Solutions**
  - **Algorithmic research**: developed new indexing techniques, achieved 10-100 fold speedup compared with existing methods
  - **Efficient software:** available open source, received a R&D 100 Award
- **Enabled science**
  - **Laser Wakefield Particle Accelerator**: FastBit acts as an efficient back-end for identifying and tracking particles (lower left figure)
  - **Combustion**: FastBit identifies ignition kernels based on user specified conditions and tracks evolution of the regions
- **Testimonial** "FastBit is at least 10x, in many situations 100x, faster than current commercial database technologies" – *Senior Software Engineer, Yahoo!*



1999	2001	2004	2006	2007	2008
Start research	WAH published	WAH patented	- Query Driven Vis - Published theory	- FastBit released - BioSolveIT begin use	- R&D100 Award - Yahoo! begin use

[Wu, et al. 2009]



THANKS!

Program managers who have funded LBNL Scientific Data Management group:

Jim Pool, Dan Hitchcock, Fred Johnson, Lucy Nowell, Lali Chatterjee, Yukiko Sekine, Walt Polanski, Ceren Susut-Bennett, Steven Lee, Laura Biven

John's email John.Wu@NERSC.gov

FastBit software http://sdm.lbl.gov/fastbit/

Scientific Data Management group http://sdm.lbl.gov/

U.S. DEPARTMENT OF ENERGY | Office of Science