Extremely Large Databases

Jacek Becla
SLAC National Accelerator Laboratory

August 21, 2012

Hands-on Experience With Petabytes

- Production: 1999-2008
- ~5 PB raw data
- 1 PB database (in 2002)

CNN.com / SCI-TECH

SEARCH GO

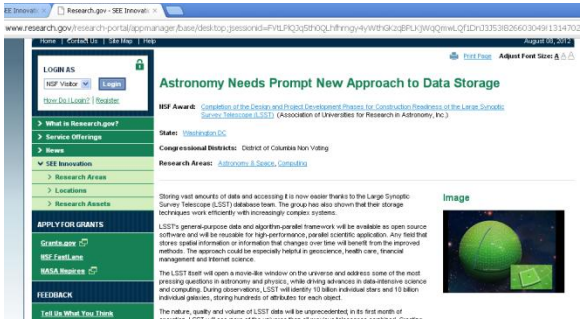
MAIN PAGE
WORLD
U.S.
WEATHER
BUSINESS

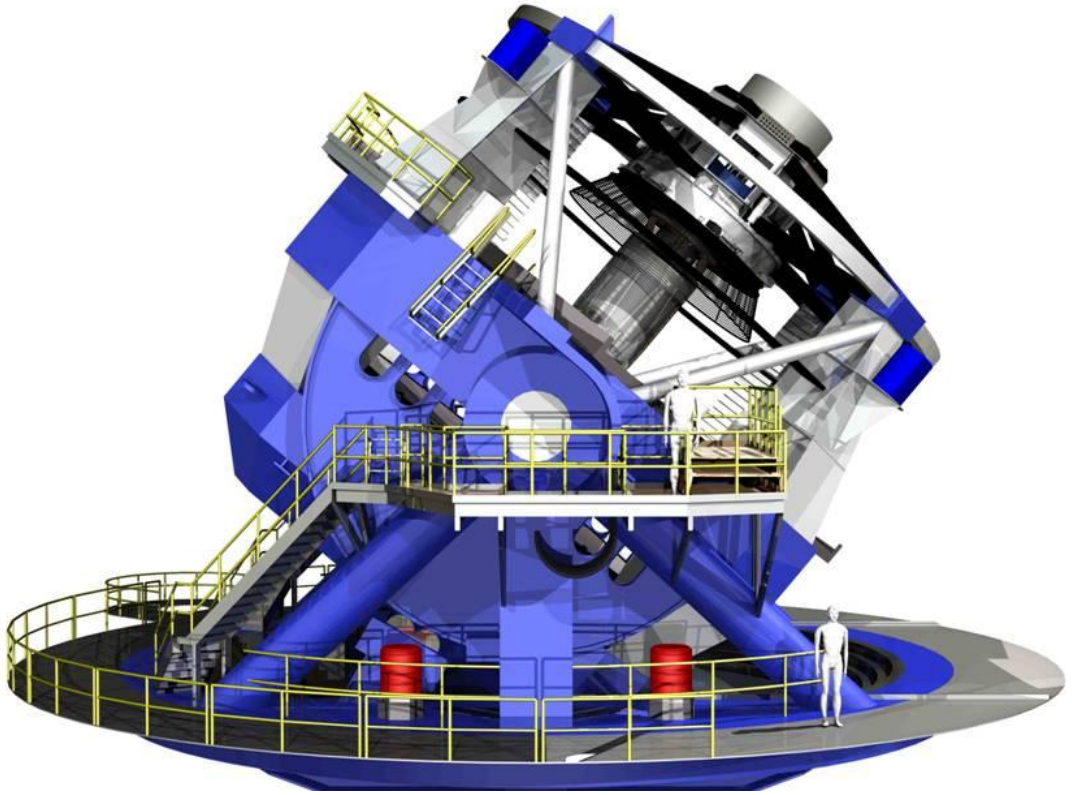
Stanford researchers may have largest database

April 18, 2002 Posted: 8:15 a.m. EDT (1215 GMT)

Hands-on Experience With Petabytes

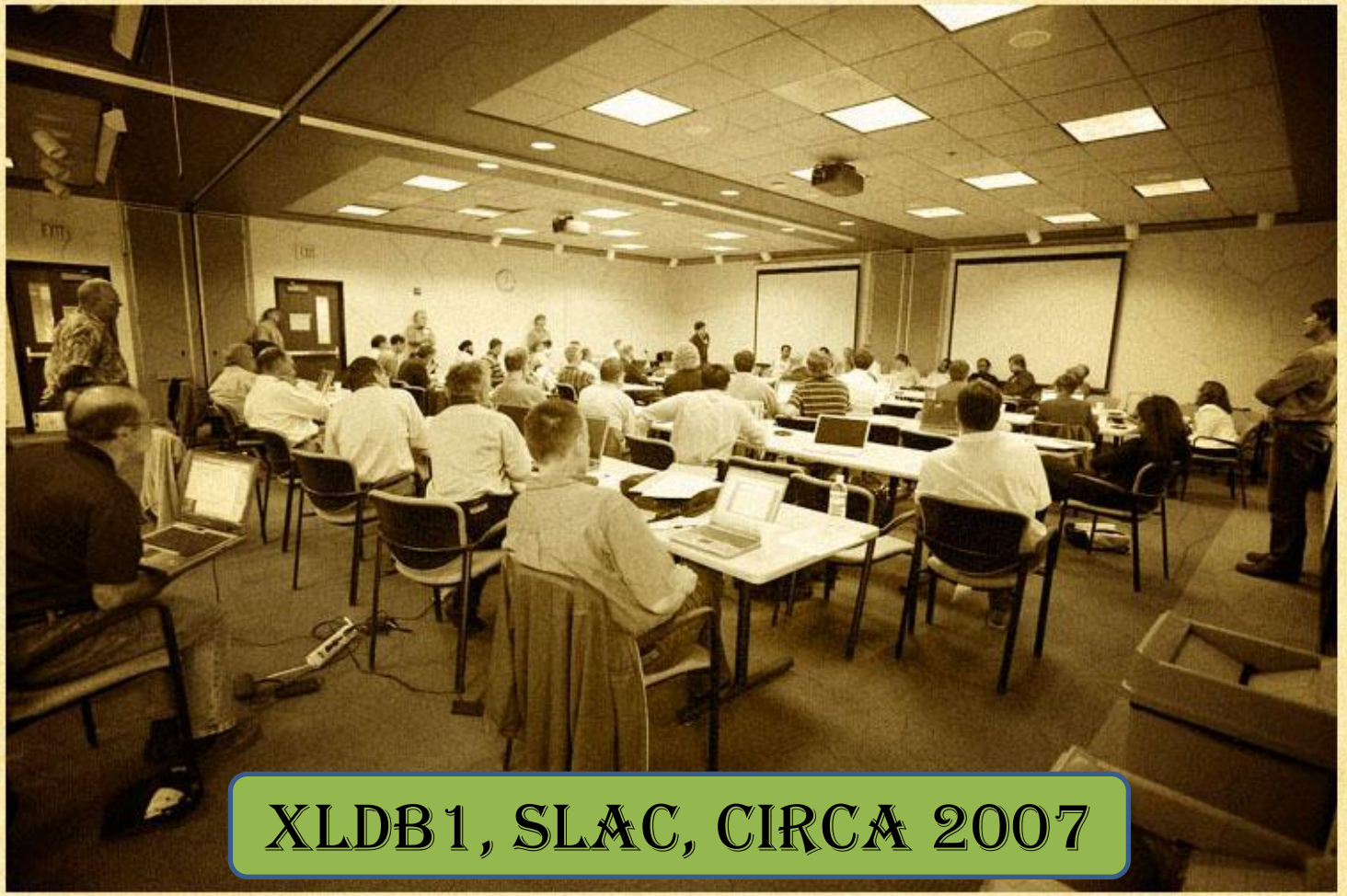
- Operations: 2022-2031
- 50+ PB images
- 45 PB database






XLOB

History



Goals

Exchange Information

Identify trends, commonalities and major roadblocks related to building extremely large data stores

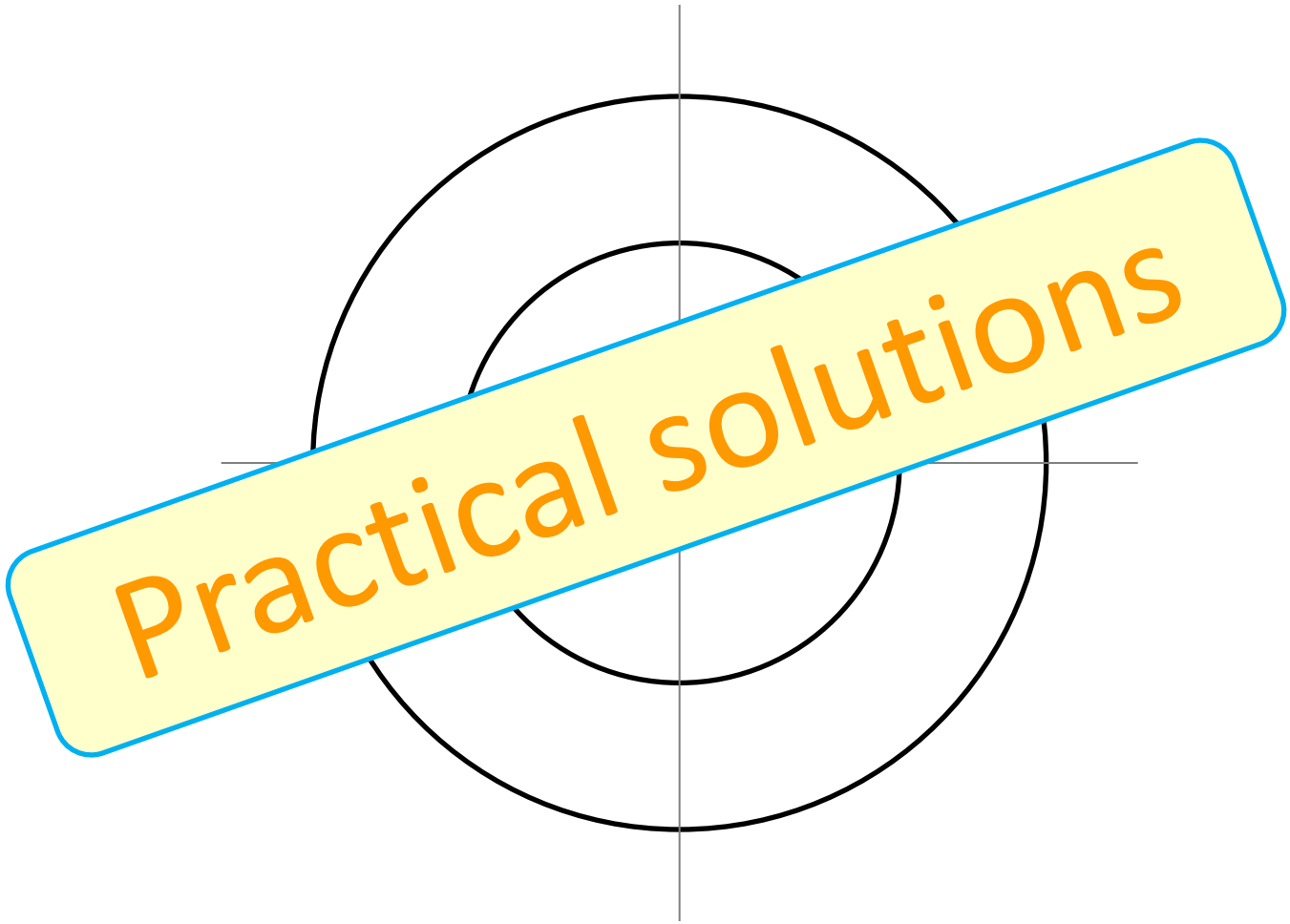
Bridges

Bridge the gap between users trying to build extremely large data stores and solution providers worldwide

Development & Growth

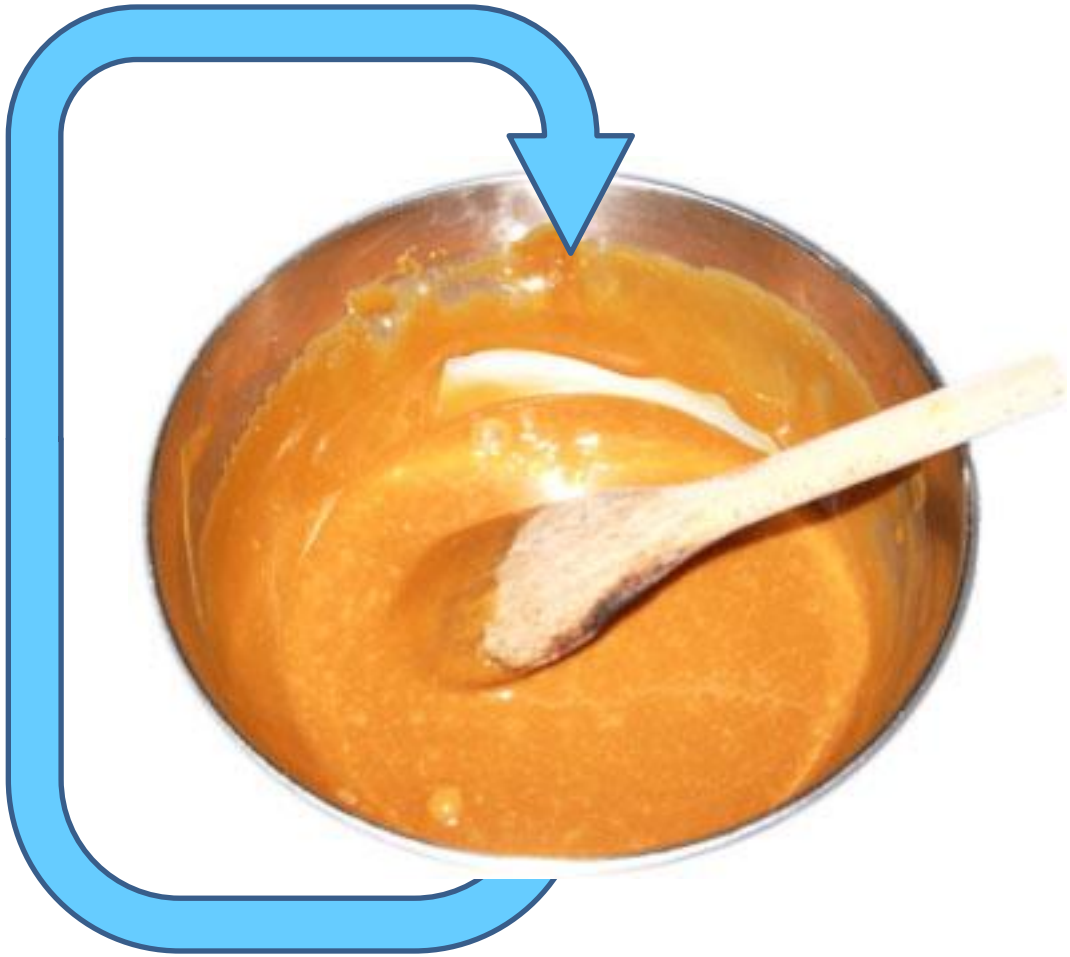
Facilitate development and
growth of practical technologies
for extremely large data stores

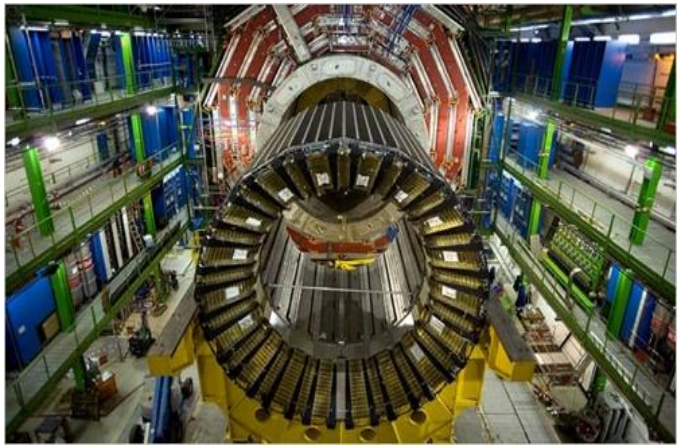
Focus



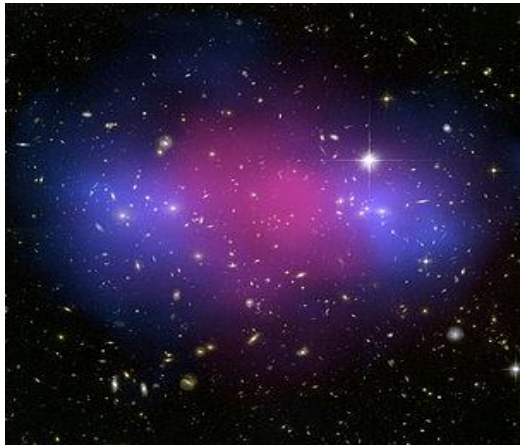
Our Recipe for XLDB Events

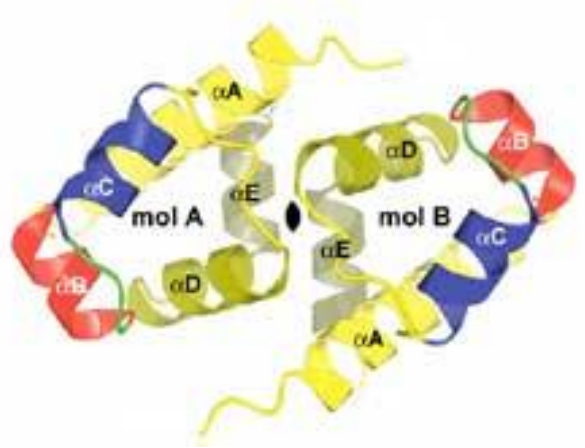
- Ingredients
  - Big science
  - Data-intensive industries
  - Vendors
  - Academia
- Directions
  1. Mix together
  2. Repeat yearly

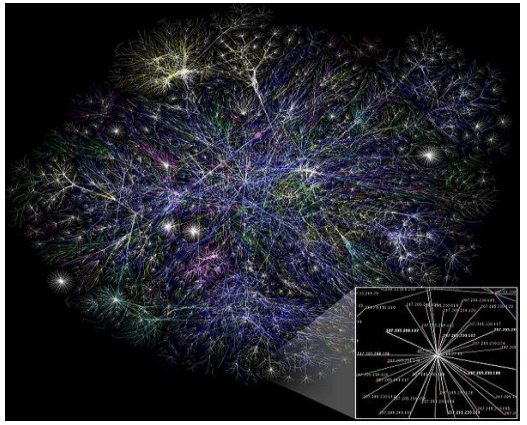


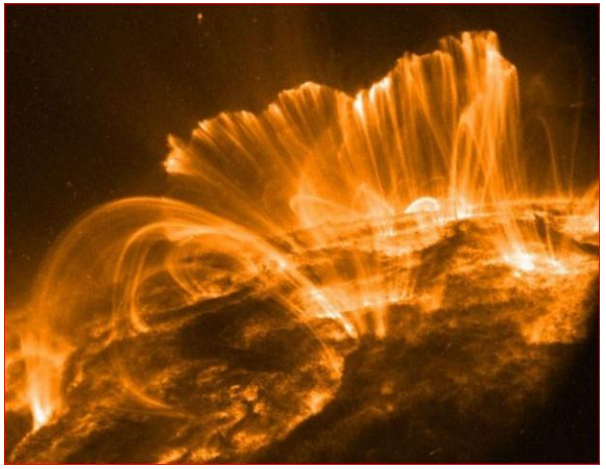


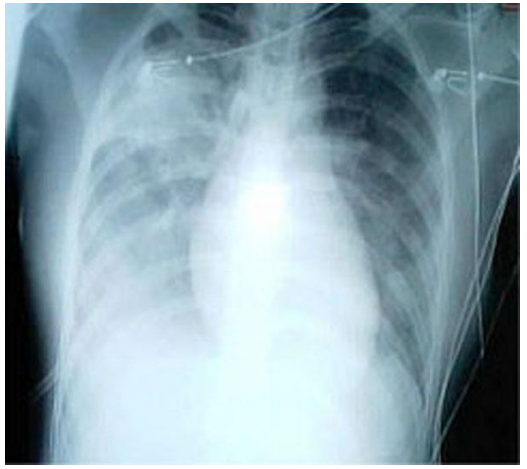

NYSE
The world puts its stock in us.



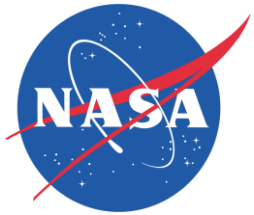

αA
αB
αC
αD
αE
mol A
mol B
αD
αE
αC
αB
αA







Laboratories and Agencies



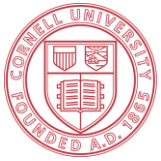
and more…

Data-Intensive Industries

at&t

and more…

Researchers



Cornell University

and more…

Vendors

and more…

XLDB Events

- 2007 & 2008
  - First two events. @ SLAC
- 2009
  - First international event (Lyon, France)
- 2010
  - First open conference
- 2011
  - First satellite workshop (Edinburgh, June 8-10)
  - Conference (SLAC, Oct 18-20)
- 2012
  - First event in Asia (Beijing, June 22-23)
  - Conference (Stanford, Sep 10-13). Tutorials, demos

Astronomy, HEP, telecom, web

Web, banking

Oil & gas

Bio, drug discovery, social networks

Healthcare, bio, retail

Attendance

XLDB Attendance

2007	2008	2009	2010		2011		
workshop	workshop	workshop	workshop	conference	workshop	conference	
43%	41%	45%	37%	24%	30%	21%	**science**
21%	19%	15%	26%	42%	41%	46%	**industry**
30%	19%	23%	18%	9%	13%	18%	**vendors**
6%	21%	15%	15%	24%	13%	14%	**academia**
0%	0%	2%	4%	1%	3%	1%	**DOE/NSF**
53	**62**	**52**	**57**	**150**	**57**	**293**	← **count**

XLDB – Tangible Results

- Detailed reports[+]
- SciDB[*]
- Use cases[+]
- Science benchmark [+]
- 1000+ user community[+]
- XLDB blog [+]
- XLDB mailing lists[+]
- ArrayQL standard [+]
  [*official announcement at XLDB-2012*]

[+] SLAC coordinates
[*] SLAC collaborates

SciDB

- Open source DBMS for scientific research
- Driven by needs of data-intensive users with array data model
  - Optical and radio astronomy, geoscience, biology
  - Web companies, drug discovery, Wall Street, oil & gas
- Designed for complex analyses on large data sets
  - Time series, spatial correlations, matrix operations

SLAC helped jump-start SciDB, including co-founding, chairing science advisory board

http://scidb.org

Summary

- Decade+ hands-on experience with peta-scale computing
- XLDB
  - Internationally recognized event
  - Started at / organized by SLAC
  - Benefits to science, industries, DOE and SLAC