



**Argonne**  
NATIONAL  
LABORATORY

*... for a brighter future*



U.S. Department  
of Energy

UChicago ►  
Argonne<sub>LLC</sub>



A U.S. Department of Energy laboratory  
managed by UChicago Argonne, LLC

# The Scientific Data Management Center

<http://sdmcenter.lbl.gov>

**Principal Investigator:** Arie Shoshani

## Co-Principal Investigators

### DOE Laboratories

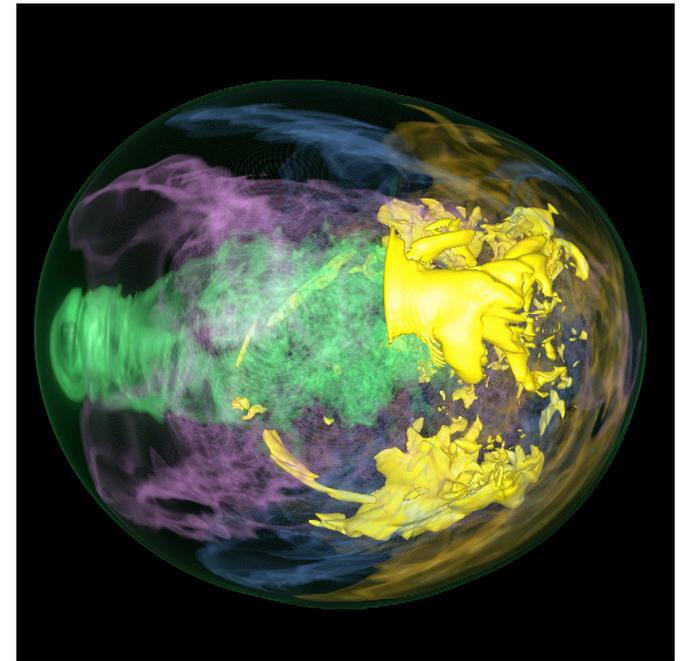
**ANL:** Rob Ross  
**LBNL:** Doron Rotem  
**LLNL:** Chandrika Kamath  
**ORNL:** Nagiza Samatova  
**PNNL:** Terence Critchlow

### Universities

**NCSU:** Mladen Vouk  
**NWU:** Alok Choudhary  
**UCD:** Bertram Ludaescher  
**SDSC:** Ilkay Altintas  
**UUtah:** Claudio Silva

# What is SciDAC?

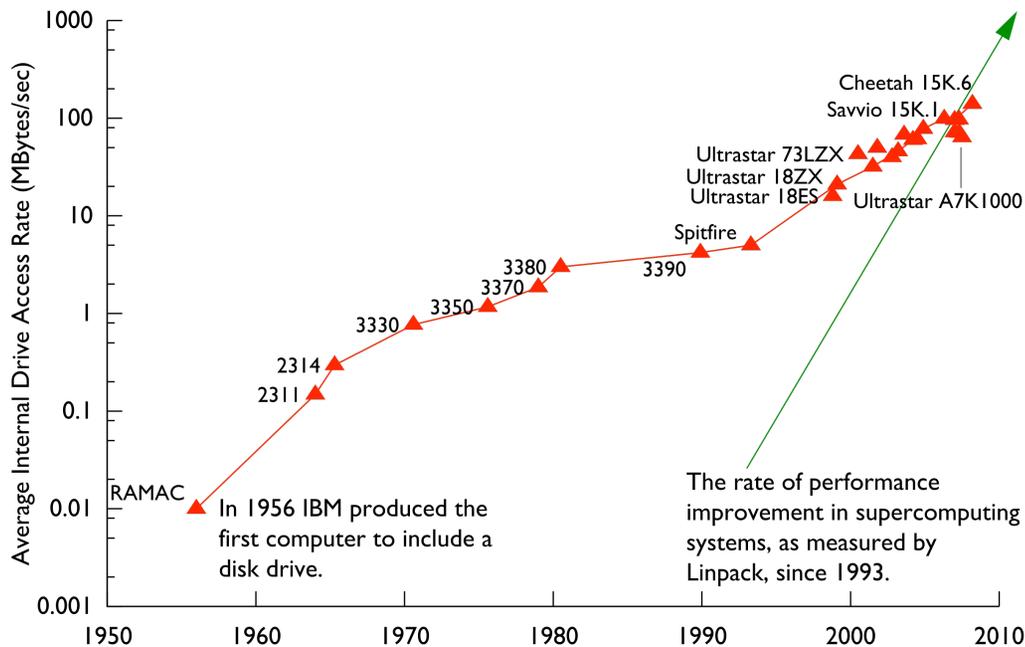
- Department of Energy program for Scientific Discovery through Advanced Computing
- Brings together physical scientists, mathematicians, computer scientists, and computational scientists
- Science projects in Nuclear Physics, Fusion Energy Sciences, Climate, Combustion, etc.
- Two computer science projects you're likely to be familiar with:
  - Scientific Data Management Center (A. Shoshani, PI)
  - Petascale Data Storage Institute (G. Gibson, PI)



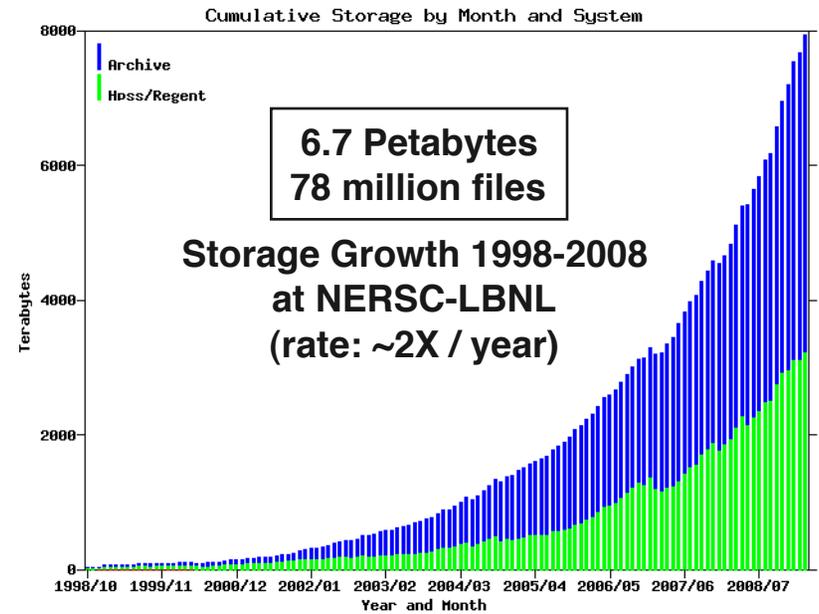
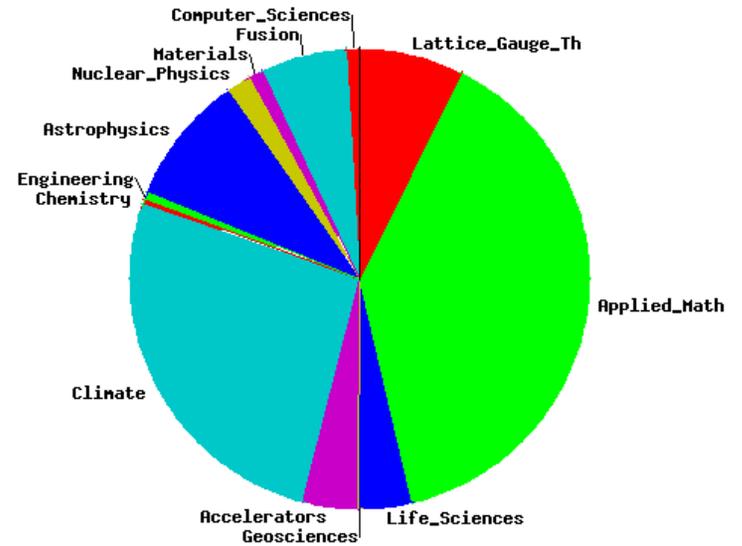
<http://www.scidac.gov>

# Scientific Data Management

Scientific data management is a collection of methods, algorithms and software that enables efficient capturing, storing, moving, and analysis of scientific data.



Storage Utilized by Discipline (2009/03)



# SDM Center Focus Areas

- High performance I/O
  - Parallel I/O, parallel file systems
  - Indexing, data movement
- Usability and effectiveness
  - Easy-to-use tools and APIs
  - Workflow and dashboards
- Establish dialog with scientists
  - Outreach and education
  - Partner with scientists
- Enabling data understanding
  - Parallelize analysis tools
  - Streamline use of analysis tools
  - Real-time data search tools
- Sustainability
  - Robustness
  - Productize software
  - Work with vendors and computing centers

# Two Examples

- Enabling efficient analysis of scientific data
- “Assisting” parallel file systems at scale



# Enabling Efficient Analysis of Scientific Data



# FastBit: Accelerating Analysis of Very Large Datasets

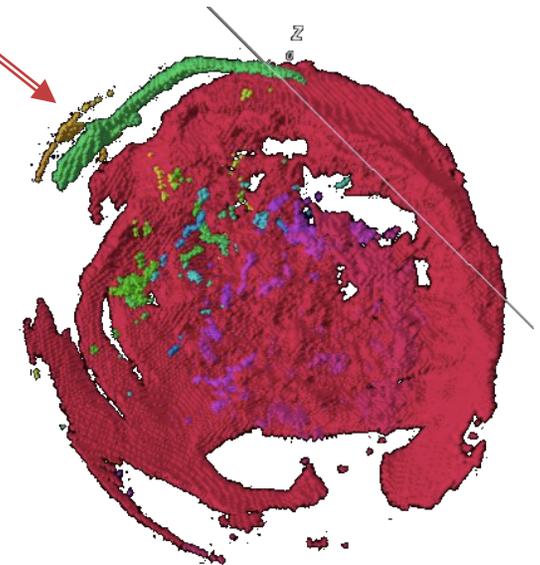
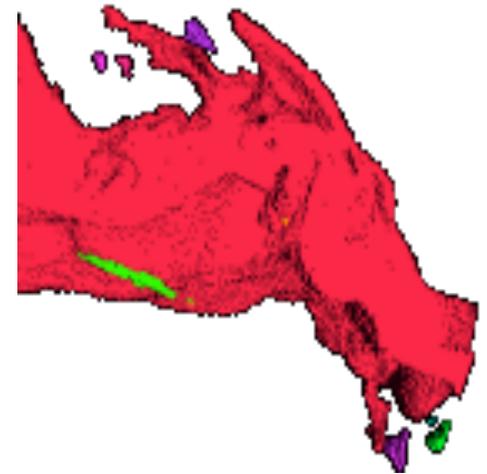
- Most data analysis algorithm cannot handle a whole dataset
  - Therefore, most data analysis is performed on a subset of the data
  - Need very fast indexing for real-time analysis
- FastBit is an extremely efficient compressed bitmap indexing technology
  - Excels for high-dimensional data
  - Can search a billion data values in seconds
  - FastBit improves the search speed by 10x – 100x of times over best known indexing methods
- FastBit indexes are modest in size compared to well-known database indexes
  - On average about 1/3 of data volume compared to 3-4 times in common indexes (e.g. B-trees)

Kesheng Wu, “FastBit: An Efficient Indexing Technology For Accelerating Data-Intensive Science”, Journal of Physics: Conference Series, 2005.



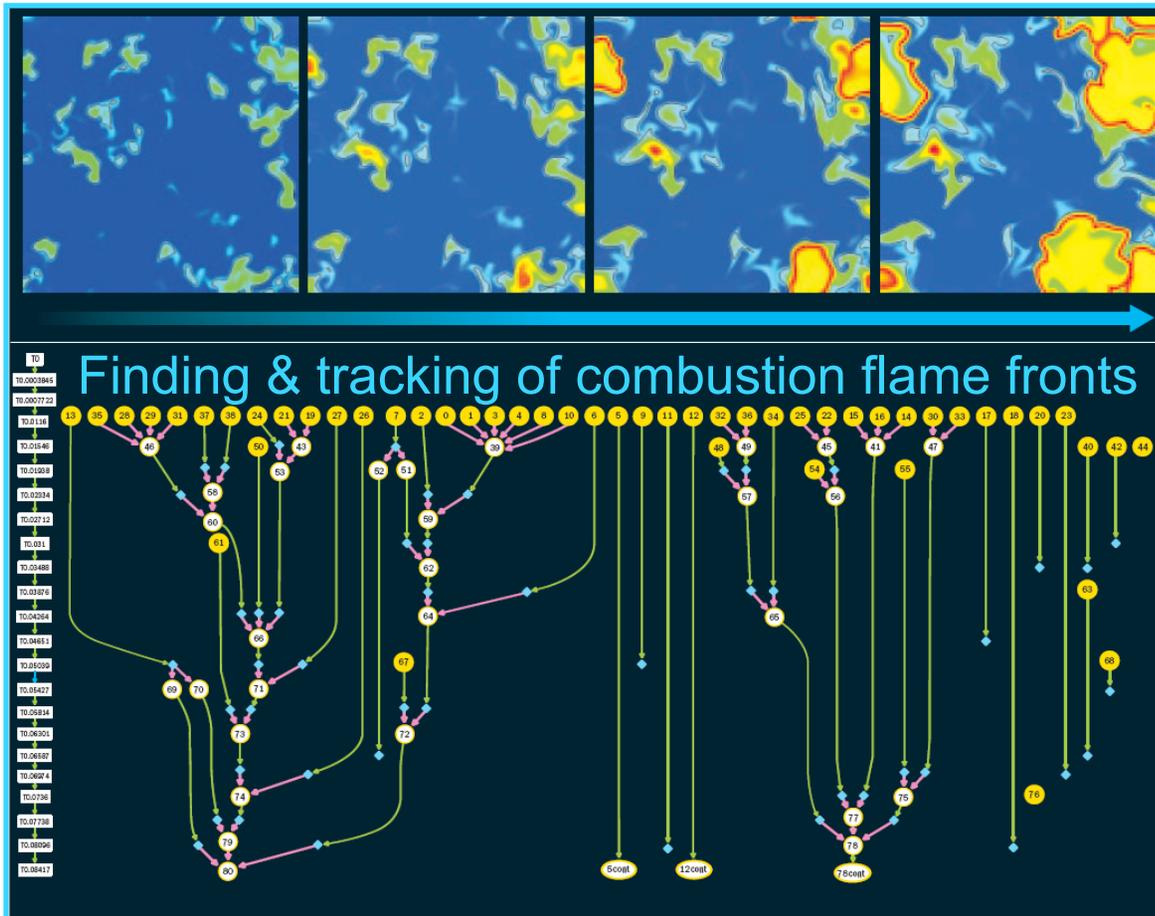
# Searching Problems in Data Intensive Sciences

- Find the HEP collision events with the most distinct signature of Quark Gluon Plasma
- Find the ignition kernels in a combustion simulation
- Track a layer of exploding supernova
- These are not typical database searches:
  - Large high-dimensional data sets (1000 time steps X 1000<sup>3</sup> cells X 100 variables)
  - No modification of individual records during queries, i.e., append-only data
  - Complex questions:  $500 < \text{Temp} < 1000 \ \&\& \ \text{CH}_3 > 10^{-4} \ \&\& \dots$
  - Large answers (hit thousands or millions of records)
  - Seek collective features such as regions of interest, histograms, etc.



# Flame Front Tracking with FastBit

Flame front identification can be specified as a query, efficiently executed for multiple timesteps with FastBit.



## Cell identification

Identify all cells that satisfy user specified conditions:  
“600 < Temperature < 700  
AND HO<sub>2</sub> concentr. > 10<sup>-7</sup>”

## Region growing

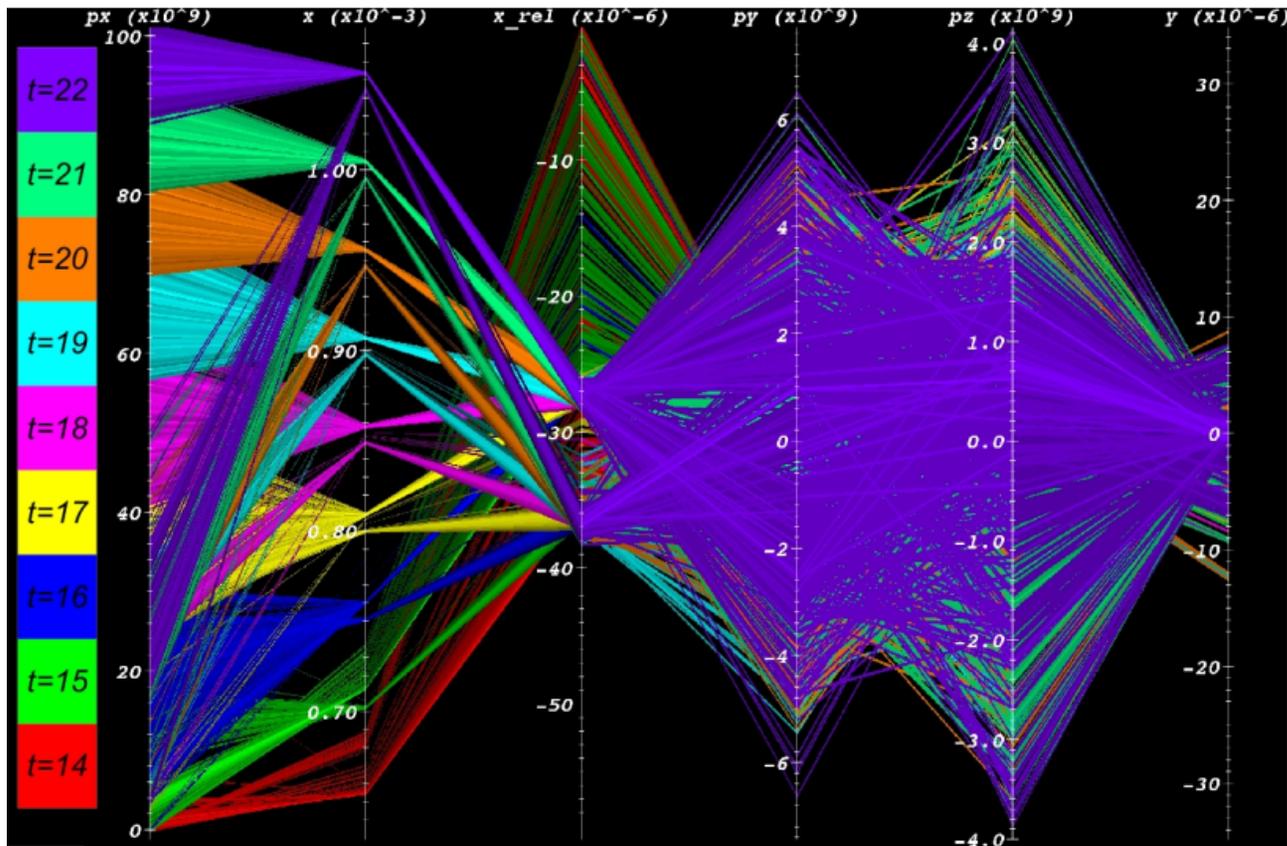
Connect neighboring cells into regions

## Region tracking

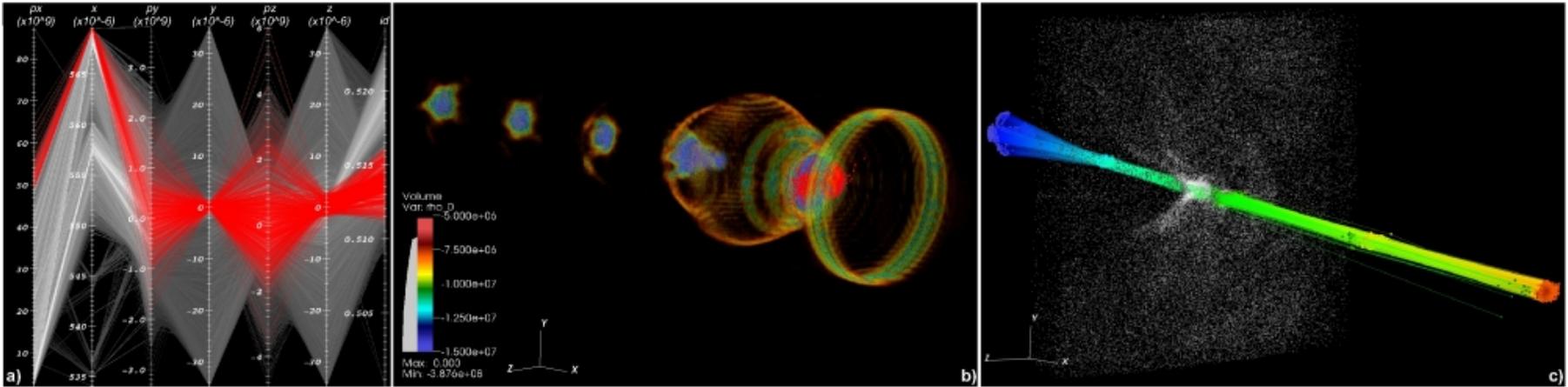
Track the evolution of the features through time

# Parallel Coordinates

Parallel coordinates is a method for viewing multivariate data. A point in simulated space is represented by a line. Intersections with vertical axes provide the value for that point for some variable (e.g., position in **X** dimension, velocity in **Z** direction).



# Query-Driven Visualization and FastBit



Parallel coordinates

Particle density

Fast particles over time

## ■ Collaboration between SDM and VACET centers

- Use FastBit indexes to efficiently select data for visual analysis in VisIt

## ■ Above example: laser wakefield accelerator simulation (VORPAL)

- Finding and tracking particles with large momentum is key to accelerator design
- Brute-force algorithm is super-linear over all particles (5 min for 0.5M particles)
- FastBit time is linear in the number of results (takes 0.3 s) = 1000x speedup

Rubel et al, “High Performance Multivariate Visual Data Exploration for Extremely Large Data,” SC08, November, 2008.

<http://vis.lbl.gov/Vignettes/Incite7/>

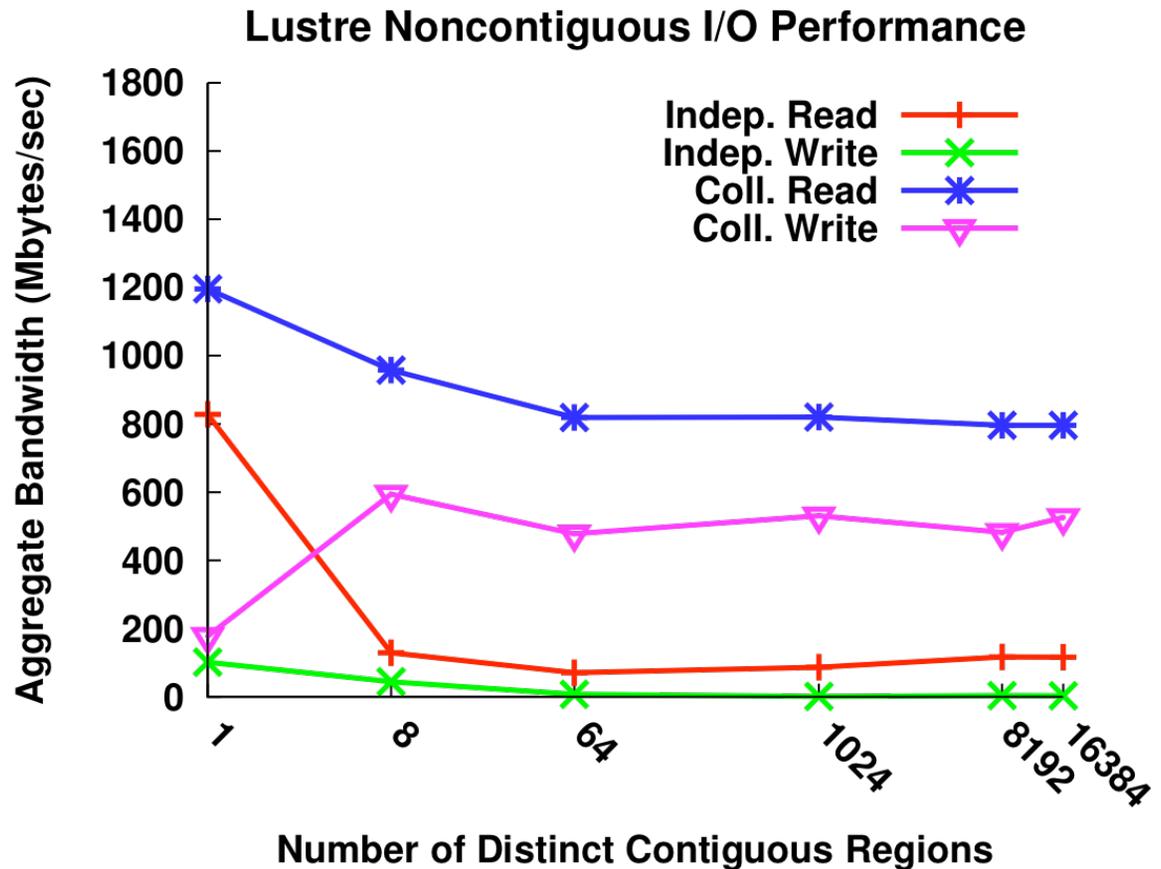
# Assisting Parallel File Systems at Scale



# The Problem of Noncontiguous Access

Tests from Thunder at LLNL, 1024-node cluster. Peak performance on this platform was approx. 2Gbytes/sec (11/2005).

Without collective I/O (MPI-IO), file system is useless for shared file access if processes access more than one region of file.



Subsequent studies show that **Lustre's locking mechanism is the culprit**, and research begins into how to work around this deficiency.

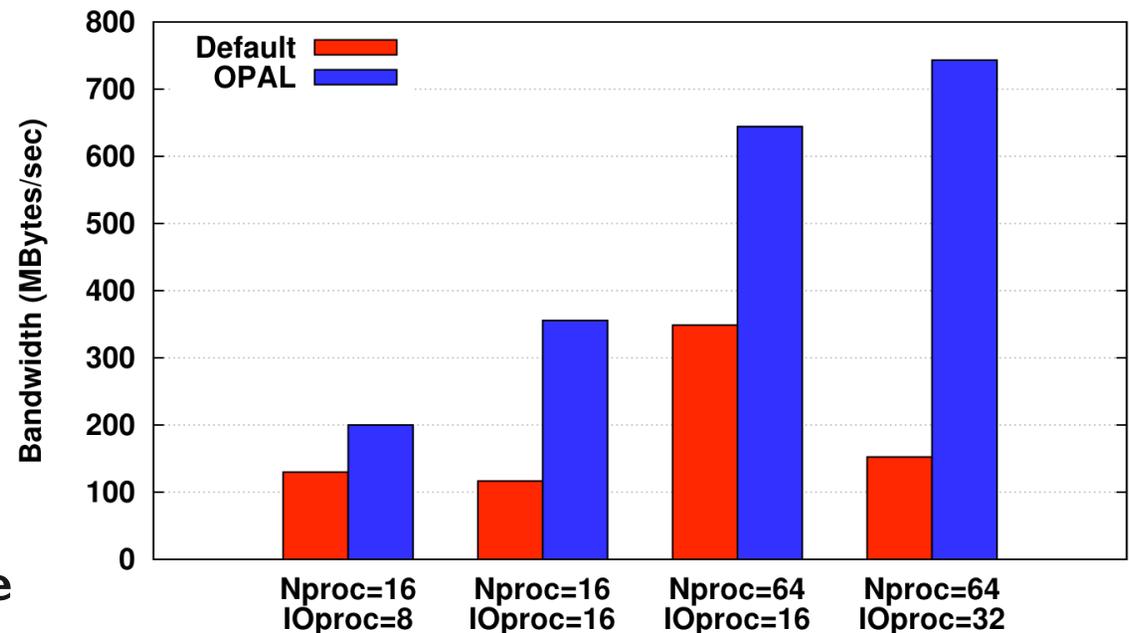
# Algorithmic Workarounds

**Optimized MPI-IO (OPAL),** developed specifically for Lustre, achieves much higher bandwidth (4x).

**BTIO benchmark exhibits typical noncontiguous I/O behavior, good test case.**

**Optimizations passed on the the Lustre team at Sun, who productized the code in conjunction with the ROMIO MPI-IO developers.**

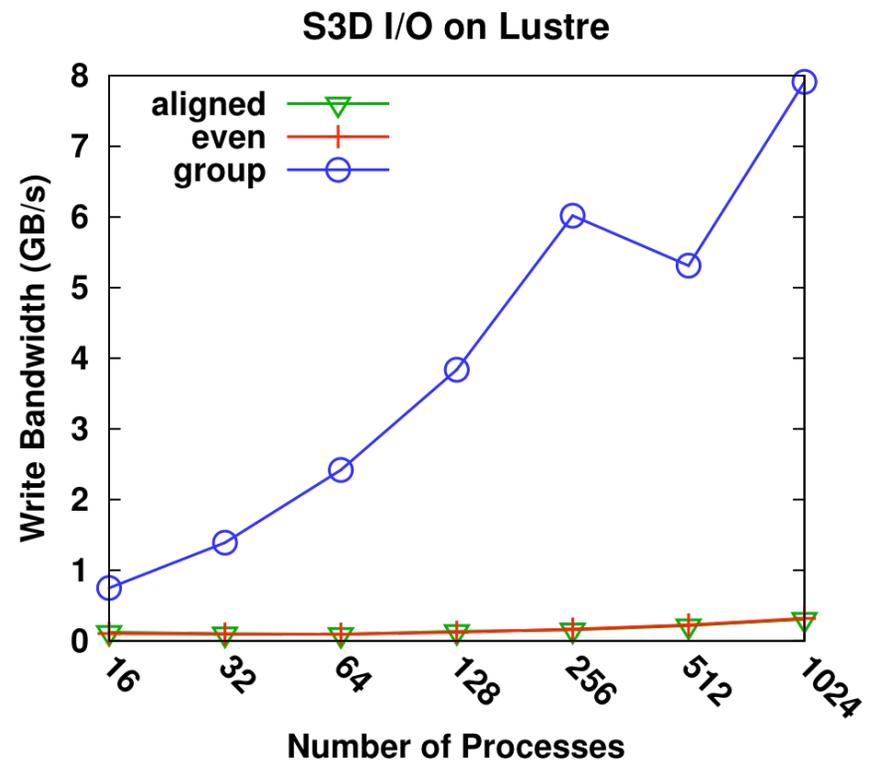
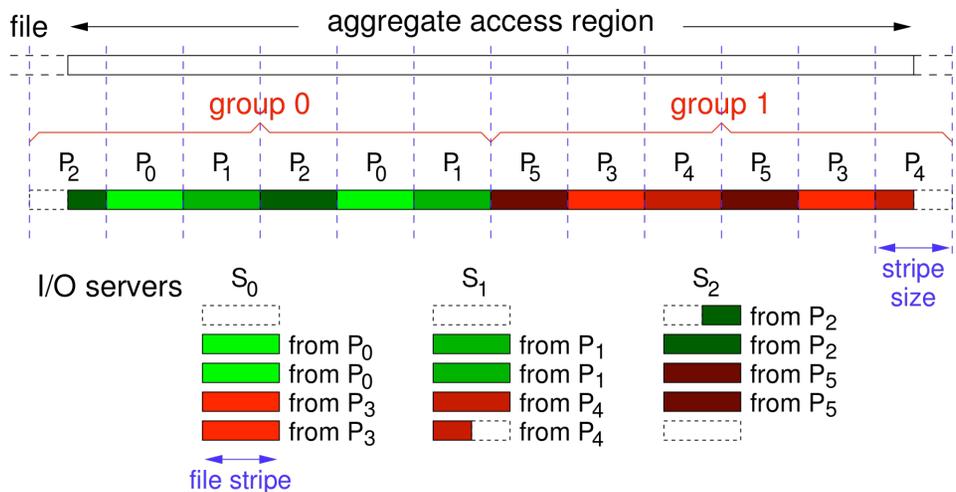
NAS BTIO Class B



W. Yu, J. Vetter, and R. S. Canon, "OPAL: An Open-Source MPI-IO Library over Cray XT," SNAPI'07, September 2007.

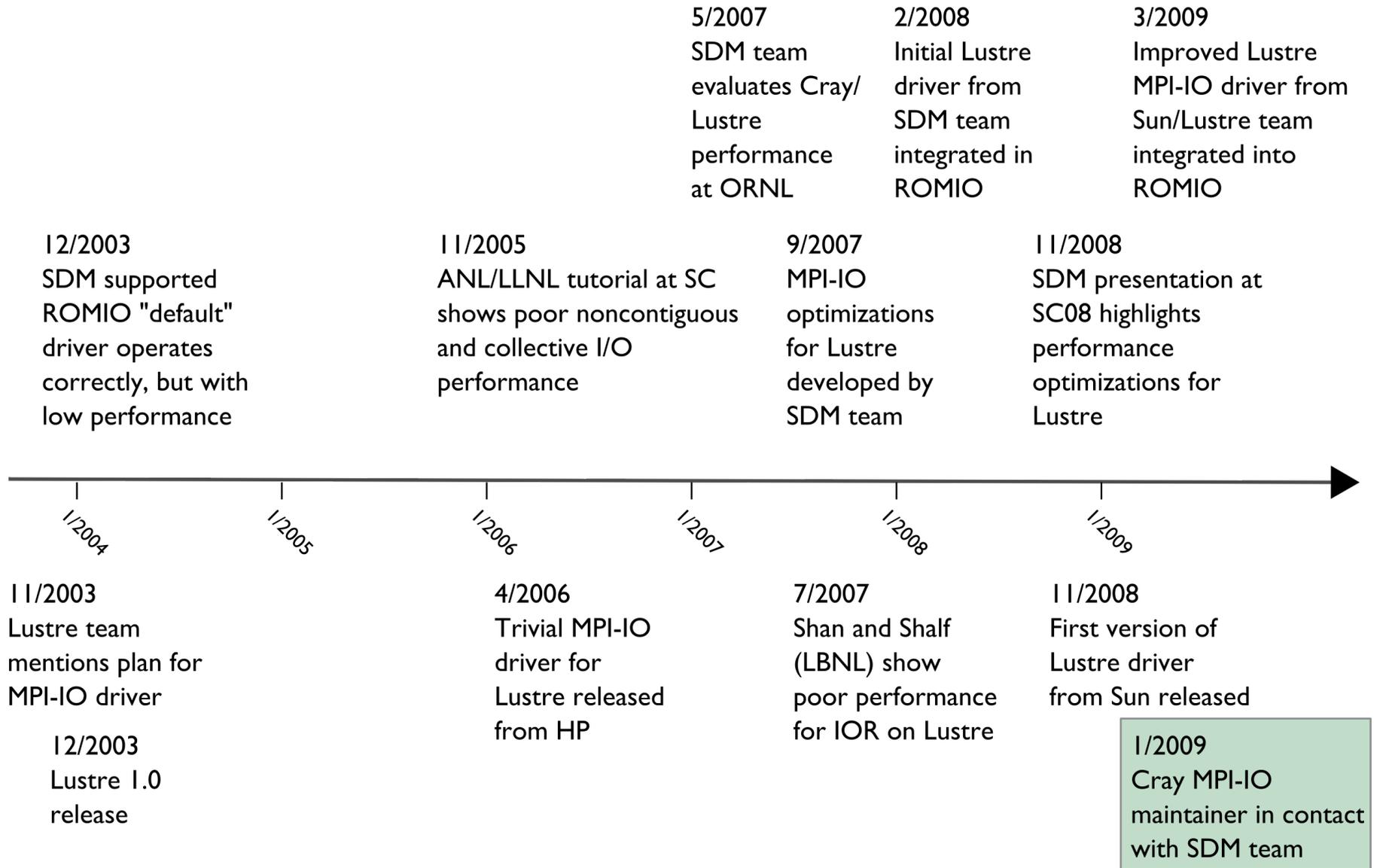
# Collective I/O and Distributed Locks

**Group-cyclic partitioning** assigns regions of the file to aggregators statically, in a round-robin fashion. Aggregators are placed in groups of  $N$ , where  $N$  is the number of servers, minimizing number of extent locks requested.



W.K. Liao and A. Choudhary, "Dynamically Adapting File Domain Partitioning Methods for Collective I/O Based on Underlying Parallel File System Locking Protocols," SC2008, November, 2008.

# Working with Vendors



# Concluding Remarks

- SciDAC has been very successful in bringing together scientists from many domains
- SDM Center balances research with development and support to solve real data management problems for application scientists
  - Workflow
  - Analysis
  - Parallel I/O
- Grounds our work, helps ensure the relevance of our research
- For more information:
  - Arie Shoshani (PI), [shoshani@lbl.gov](mailto:shoshani@lbl.gov)
  - <http://sdmcenter.lbl.gov>