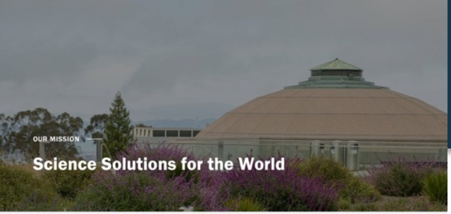
FasTensor: Efficient Tensor Computation for Large-Scale Data Analysis

Kesheng (John) Wu, Bin Dong, Suren Byna
Scientific Data Management Group
Lawrence Berkeley National Laboratory
http://sdm.lbl.gov/

Lawrence Berkeley National Lab



For more than 90 years...
EXPLORE OUR HISTORY

...we have been driven to excellence...
READ ABOUT OUR PEOPLE

...dedicated to a culture of belonging...
COMMITMENT TO DIVERSITY

...in partnership with the nation's leading public university...
OUR PARTNERSHIP WITH UC

...and building on the strengths of our communities.
COMMUNITY ENGAGEMENT

BERKELEY LAB

About | News | Careers | Search the Lab

Research | Capabilities | People | Engage

Our research focuses on discovery science and solutions for clean energy and a healthy planet.

- Research Overview

DISCOVERY SCIENCE

- AI, Math, and Data
- Accelerator Technologies
- Cosmic Frontiers
- Frontier Computing Sciences
- Materials and Chemical Sciences
- Microelectronics and Beyond
- Mysteries of Matter
- Quantum Science

CLEAN ENERGY

- Alternative Energy
- Bioenergy and Biofuels
- Biomanufacturing
- Carbon Management
- Energy Efficiency
- Energy Storage
- Energy Systems
- Hydrogen
- Manufacturing Decarbonization
- Sustainable Transportation

HEALTHY EARTH SYSTEMS

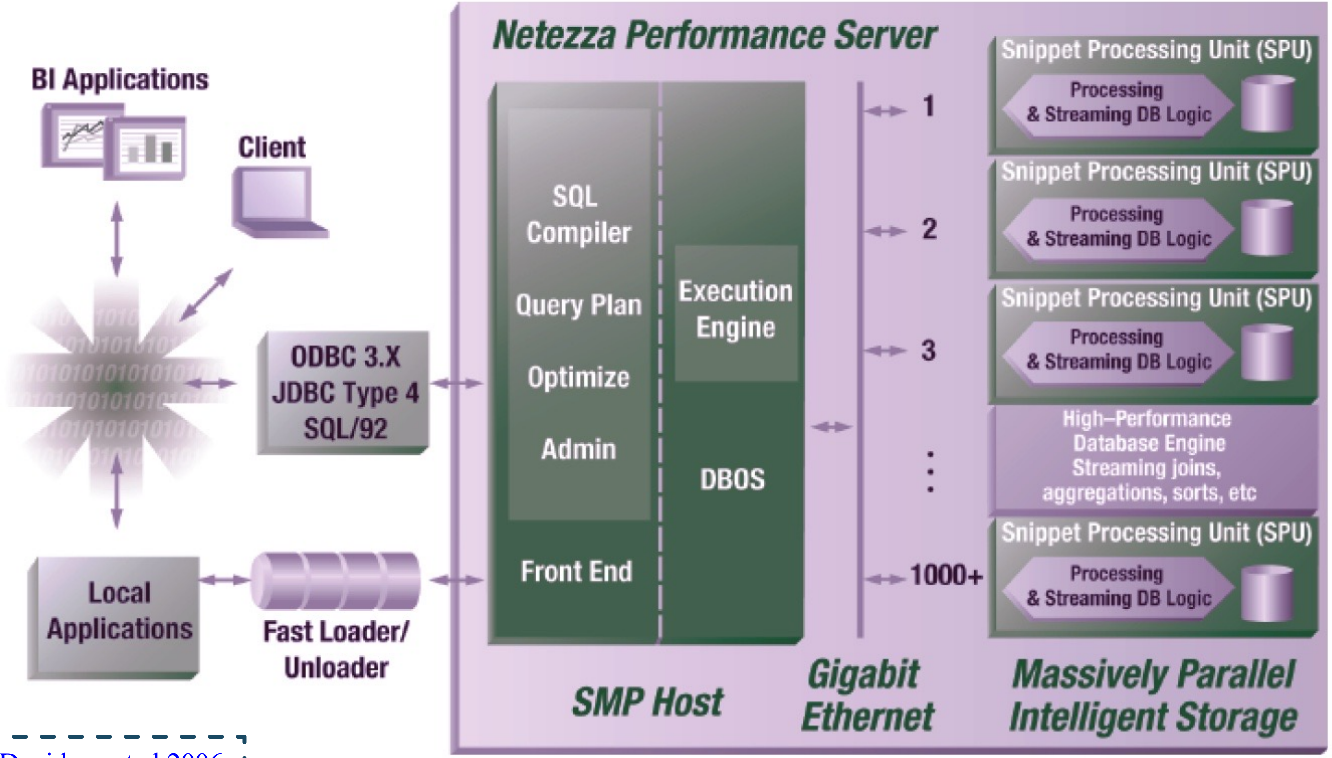
- Clean and Affordable Water
- Climate
- Earth and Ecological Systems
- Human Health
- Microbes to Ecosystems

FUTURE SCIENCE

- Emerging Ideas
- Scientists of Tomorrow
- Emerging Capabilities

http://www.lbl.gov

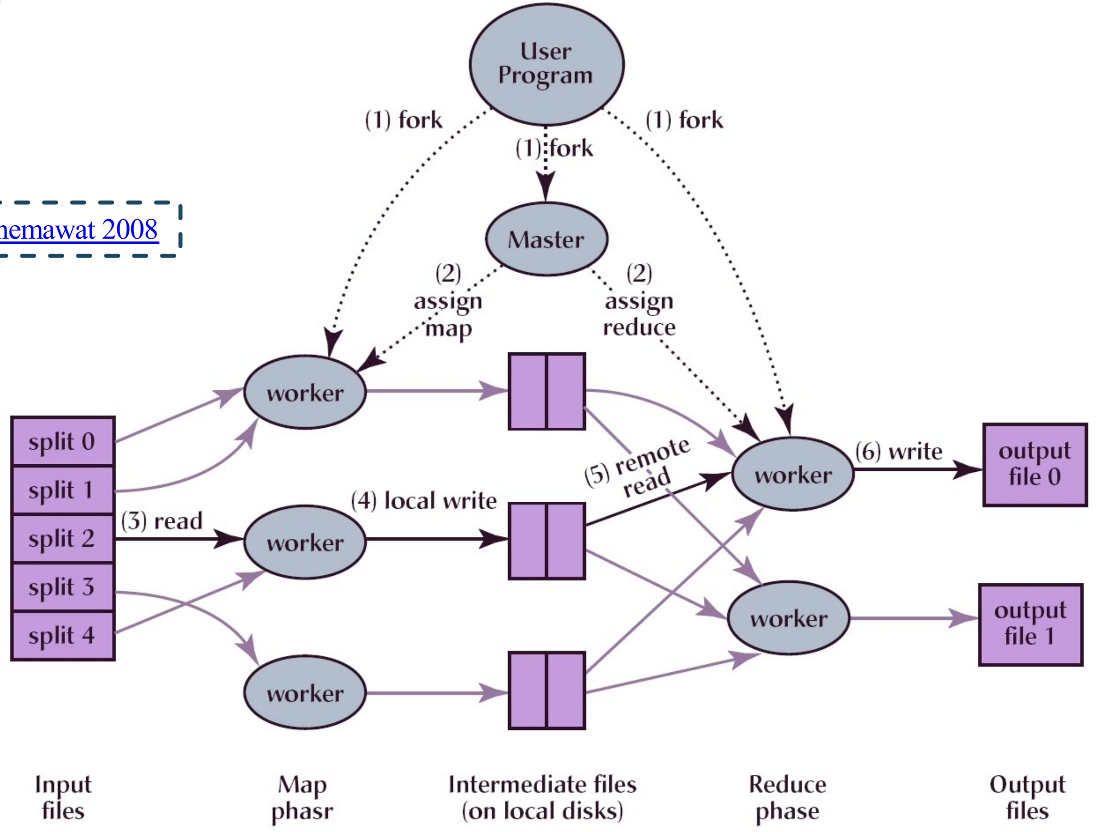
Classical Data Management -- Parallel Database Systems



[Davidson et al 2006]

Modern Data Management -- MapReduce

[Dean and Ghemawat 2008]



Big Data is Everywhere in Business

Is there anything special in scientific applications?

http://mattturck.com/bigdata2017/

BERKELEY LAB

Is Scientific Data Different?
Ex 1: Remote Sensing → Complex Computation

Subsurface Related Applications

Distributed Acoustic Sensing (DAS) turns fiber optic cable into high-precision seismic sensors. It is being studied for a variety of applications that monitors earthquake, soil properties, permafrost, oil production, oil transport pipelines, and so on. It promises to save lives by improving earthquake prediction and boost economy by improving production from hydraulic fracturing.

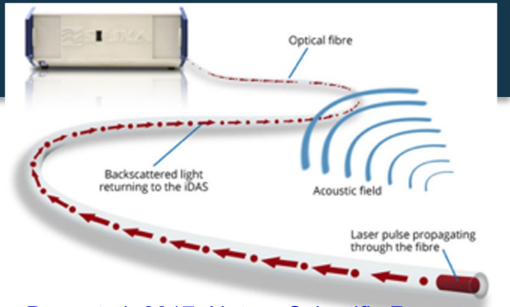
State of Art

DAS are deployed in several demonstrations and are producing hundreds of terabytes of data per installation. The data collected are typically returned to data centers by transporting the hard drives from field.

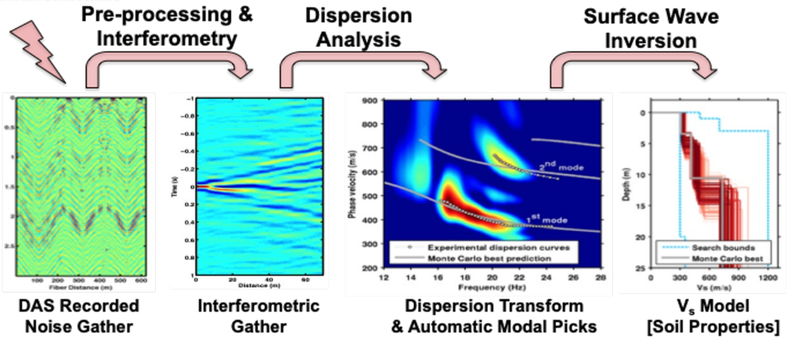
Major Challenges: processing the volume of data collected at the sensor

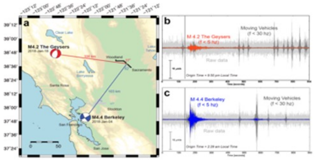
As illustrated in the data processing workflow on the right (middle), the pre-processing and interferometry step needs to ingesting petabytes of raw sensor data, reducing the raw data into interferometry data in the field will dramatically reduce the volume of data to be transported and make DAS more effective tool.

Schematics of Distributed Acoustic Sensing: using the back scattering to deduce the motion of the fiber optic cable. Sample applications are shown below.

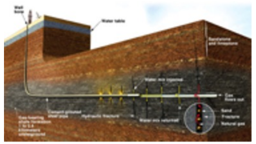


[Dou, et al. 2017. Nature Scientific Report.]





Earthquake detection



Oil well monitoring

Oil pipeline monitoring

From J Ajo-Franklin, V R Tribaldos, and J Wu

Is Scientific Data Different?
Ex 2: Precision Agriculture → Multiple Data Sources

Environment and Agriculture

Advances in monitoring technologies for environment, hydrology and agriculture will increase the volume of data produced and making real-time streaming analysis an essential to these scientific and engineering activities.

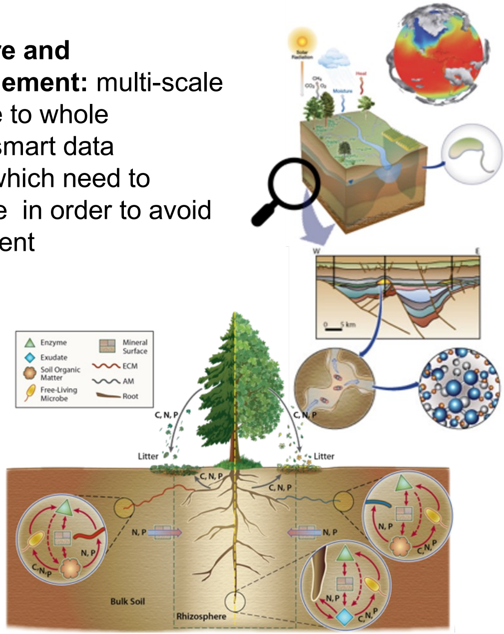
State of Art

Data collection, processing, integration and dissimilation are ad hoc and require considerable amount of manual processing involving many different data management systems

Major Challenges

- Environmental monitoring results such as water level (top right image) could change quickly, effective real-time analysis and reporting is critical to human life and safety
- New generation of data collection tools such as Nanopore (lower left image) is anticipated to produce many gigabytes of data per sample.
- How to effectively integrate data from multiple sources in the field is a unique challenge

Precision Agriculture and Environment Management: multi-scale data (from DNA-scale to whole watershed) requires smart data integration, some of which need to performed at the edge in order to avoid massive data movement



Images from B Brown, C Varadharajan, and W Collins

Is Scientific Data Different?
Ex 3: Smart Infrastructures → Real-time Control

Self-Managing and Self-Healing Future Esnet

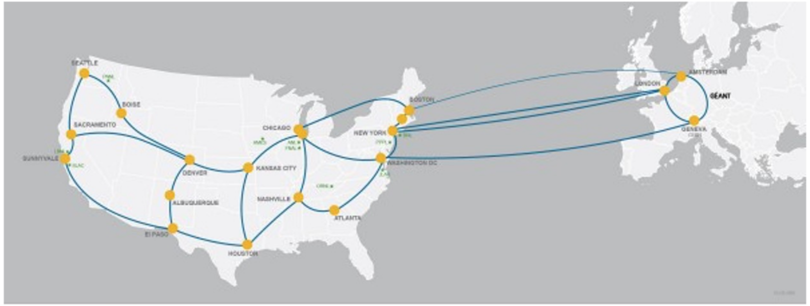
ESnet is the backbone network connecting DOE Office of Science facilities. It is responsible for moving petabytes of data per week. The data rate has been doubling every 18 months.

State of Art

Much of network management is still labor-intensive, lots of potential to automate mundane tasks

Major Challenges

- More sites and more variety of devices are being connected to ESnet. The networking components are becoming more "software-defined," which requires more software control and offers more opportunities for automation.
- Distributed real-time control would be important to achieve the self-managing and self-healing vision.



ESnet
ENERGY SCIENCES NETWORK

Schematics of ESnet: ESnet is the backbone network connecting large DOE science facilities and national laboratories. As many scientific activities have become more collaborative, ESnet has grown to support DOE's participation in international collaborations such as LHC and ITER. AI at the Edge would play a critical role to make the future ESnet self-managing and self-healing.

From Mariam Kiran, Chin Guok, and John Wu

Is Scientific Data Different?
Ex 4: Systems of Systems → Decentralized Control

Connected and Autonomous Mobility System

DOE is conducting research and development that investigates how disruptive connected and autonomous vehicles will impact energy consumption in transportation. These efforts will result in large amounts of data being generated by many disparate types of sensors located both on vehicles and infrastructure.

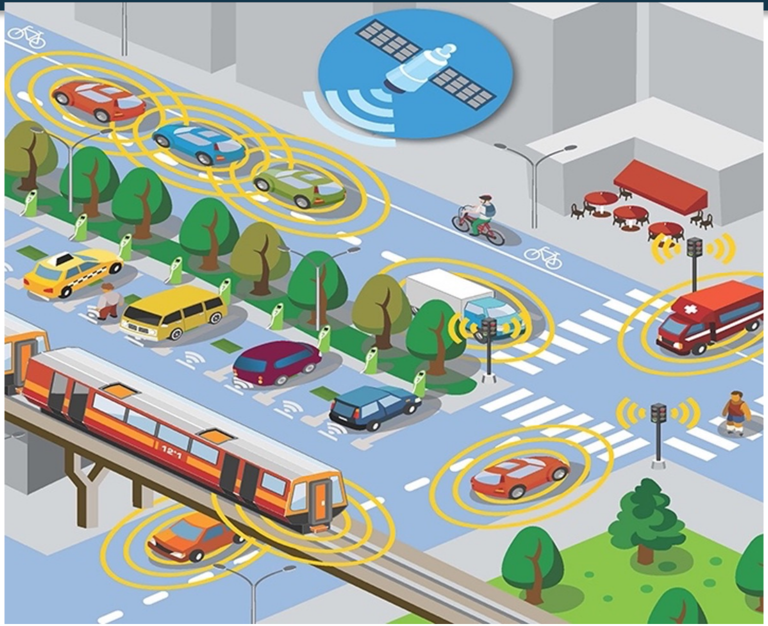
State of Art

Offloading limited amounts of data to data-centers often performing shallow analysis.

Major Challenges

As connected, autonomous vehicles begin to be deployed the amount of data generated will grow enormously. It will no longer be possible to perform all analysis at a central location. Edge computing will be a necessary component to leveraging all data created and making real time decisions.

The collection sensors deployed will be disparate and thus will require different types of edge AI resources. This will necessitate an edge infrastructure capable of supporting and integrating heterogeneous resources involving many autonomous vehicles interacting with traffic signaling system and electric power grid among others, and thus requiring an effective way to integrate systems of systems.

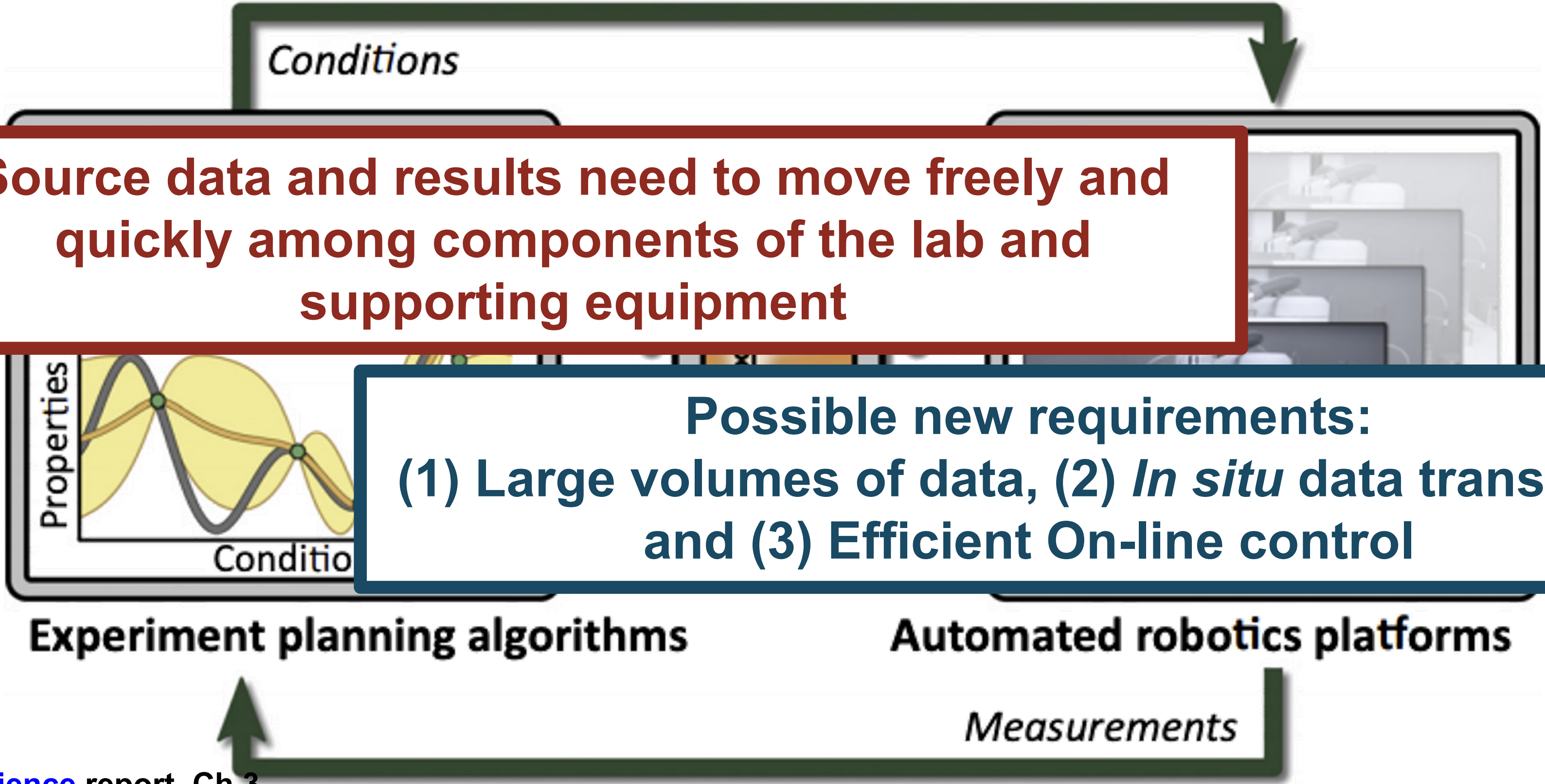


Connected Vehicles and Infrastructure: DOE is conducting research and development that investigates how disruptive forces such as automated, connected, electric and/or shared (ACES) vehicles will impact energy consumption in transportation. It also helps communities determine how they can plan for and encourage energy efficiency increases in mobility.

Image courtesy DOE Vehicles Technology Office (VTO)

https://www.energy.gov/eere/vehicles/energy-efficient-mobility-systems

Ex 5: Self-driving Biology Lab



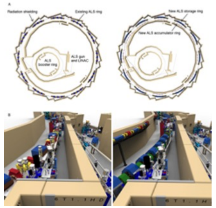
Source data and results need to move freely and quickly among components of the lab and supporting equipment

Possible new requirements:
(1) Large volumes of data, (2) *In situ* data transport, and (3) Efficient On-line control

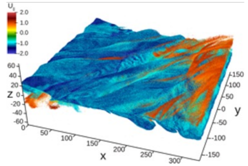
AI for Science report, Ch 3

FasTensor: Designed to Address Two Fundamental Challenges

(1) Size → parallelization

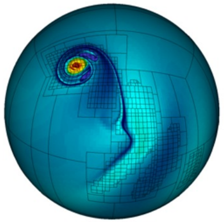


Light Source
180 PB/year
(ALS-U at Berkeley Lab)

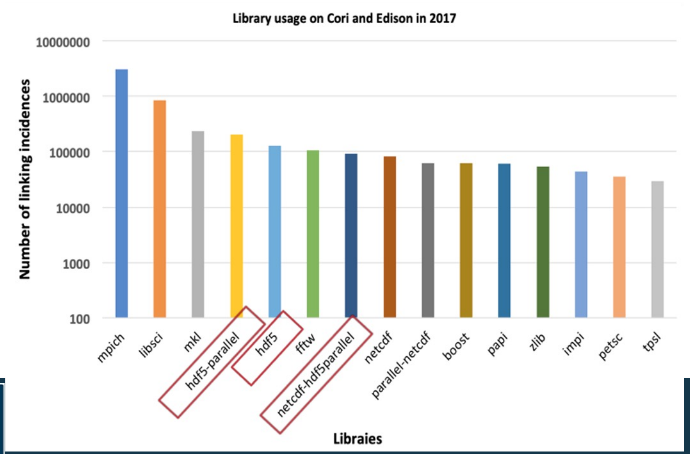


High Energy Physics
200 PB/year

Genomics
10 PB/year



Climate
100 EB/year



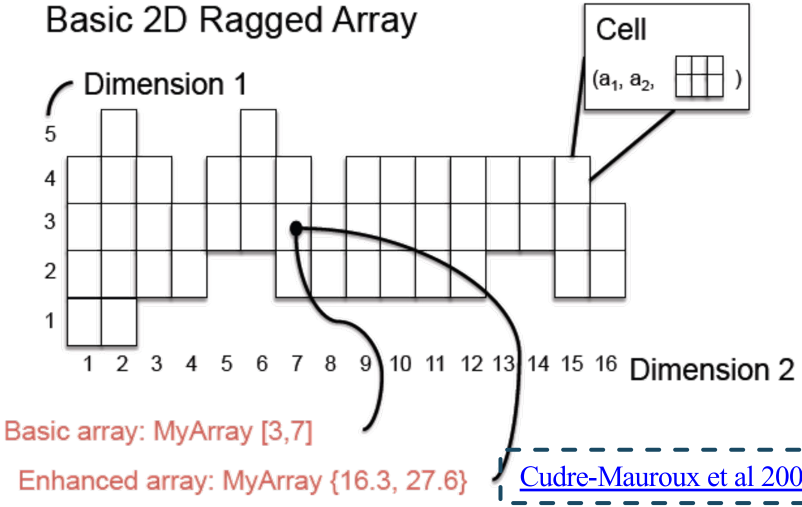
(2) Organization → files

Most are multidimensional arrays, stored in file formats like HDF5, PNetCDF, ADIOS, etc

Sources: L. Nowell, D., Ushizima, S. Byna, JGI and ALS at LBNL, etc.

Why FasTensor: Scientific Data in Arrays

- **Approach 1: a database system for scientific applications, e.g., SciDB**
- **SciDB features:**
  - Array-oriented data model
  - Append-only storage
  - First-class support for user-defined functions
  - Massively parallel computations

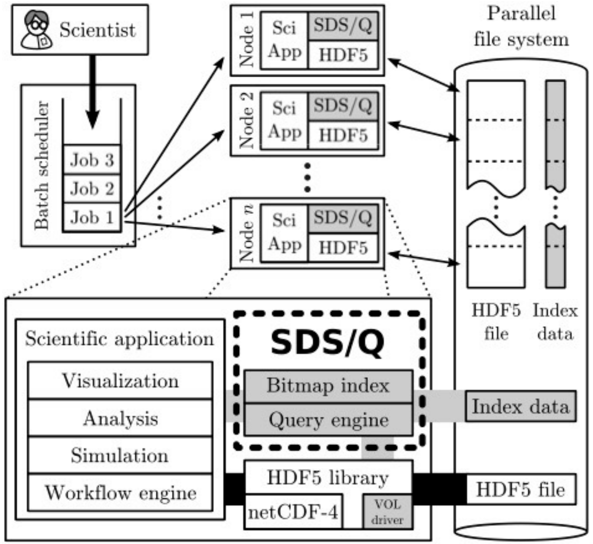


[Cudre-Mauroux et al 2009]

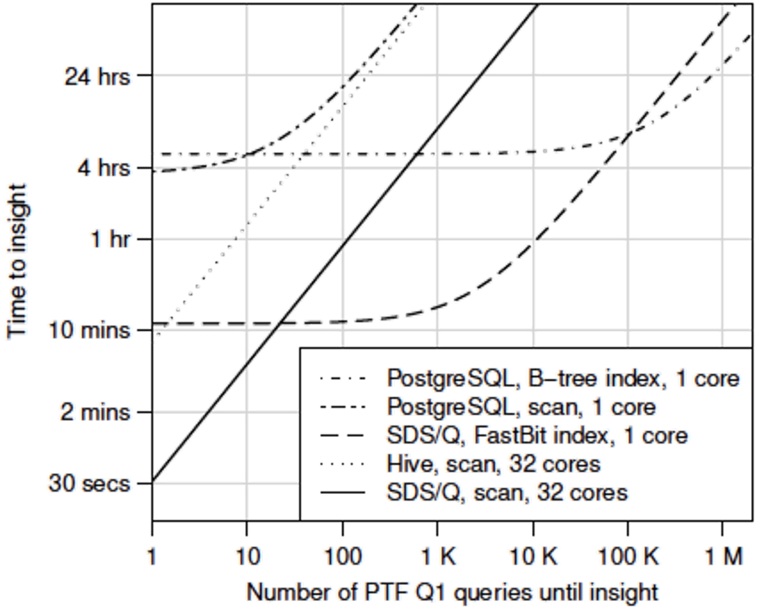
Why FasTensor: Scientific Data in Files

Approach 2

- **Relational parallel query processing directly on scientific file formats**
- **Using database technology requires costly loading of data and converting results**



Overview of SDS/Q, the querying component of the Scientific Data Services framework.



Time to insight for a PTF query: 150X faster than PostgreSQL and 10X faster than Hive

S. Blanas, K. Wu, S. Byna, D. Bin, A. Shoshani, SIGMOD 2014

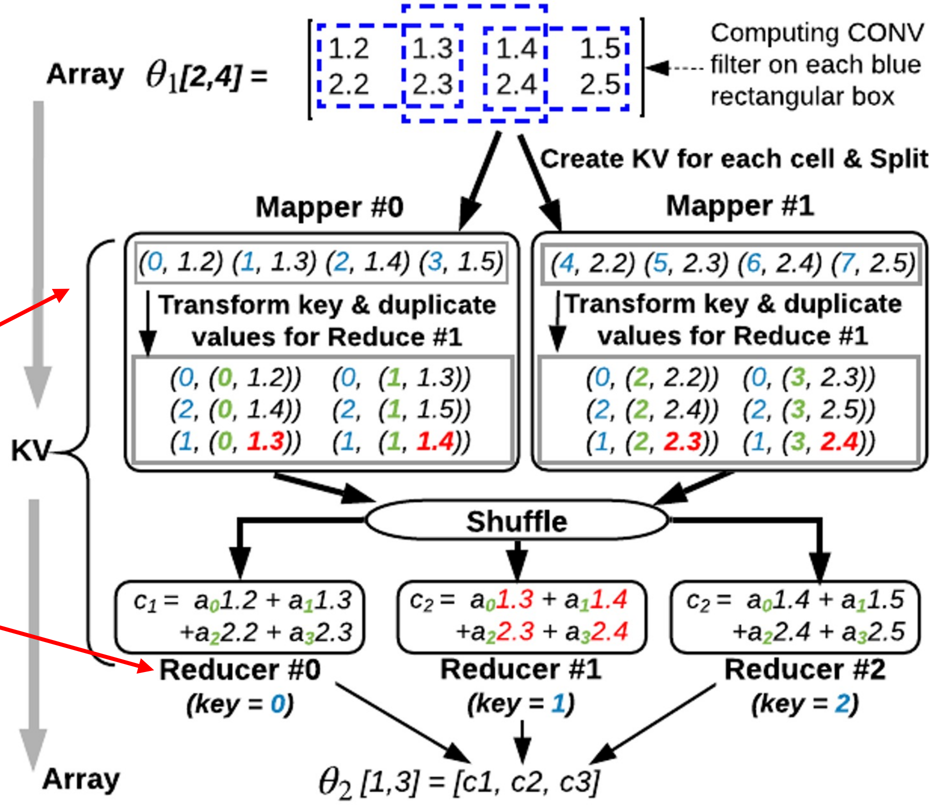
Why FasTensor: Slow Operations in MapReduce

Convolution on a 2 by 4 2D Tensor (Array)

Kernel is 2 by 2

1. Mismatched Data Model
-- Convert Tensor to KV list at Map stage

2. Expensive reduce
-- Duplicate KV for Reduce stage



*More details at **"SLOPE: Structural Locality-aware Programming Model for Composing Array Data Analysis"**, ISC 2019*

FasTensor: New Data-Parallel Programing Model on Arrays

Inspiring by: a Tensor is a multidimensional array with transformations

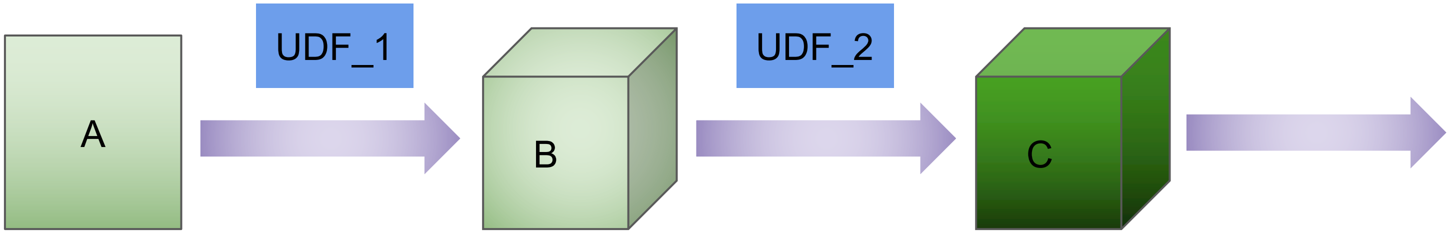
FasTensor is a generic parallel data programming model

Tensor = Multidimensional **Array** + **Transform** Rules

Data Model ← Array

Transform → User Defined Function (UDF)

Array



B = A→Transform(UDF_1)

C = B→Transform(UDF_2)

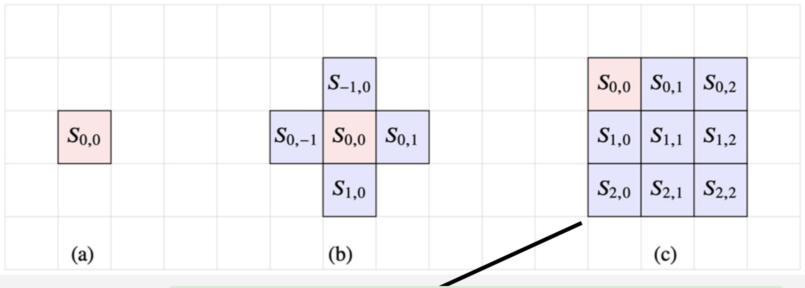
How FasTensor works

- ➔ **Multidimensional Array Model**
  - ◆ Disk (e.g. HDF5, ADIOS, netCDF)
  - ◆ Memory (e.g., DASH)
- ➔ **Flexible Stencil Data Structure**
  - ◆ Flexible UDF functions
- ➔ **Execution Engine**
  - ◆ Auto-parallel: MPI/OpenMP/hybrid
  - ◆ Optimized Chunking Size
  - ◆ Optimized Ghost Zone
  - ◆ In-place Modification Semantic
  - ◆ Fault-tolerance Support

Stencil

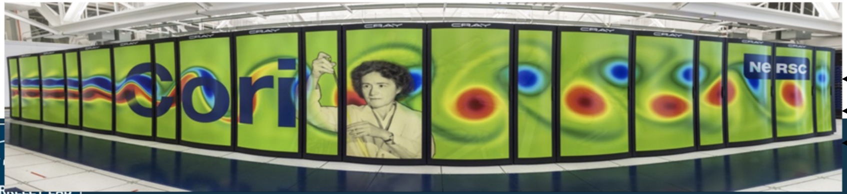
Base Cell

Neighbor Cells -- relative offset(s)



Example: define a sequential **Sum** as udf_Window_Aggregates()

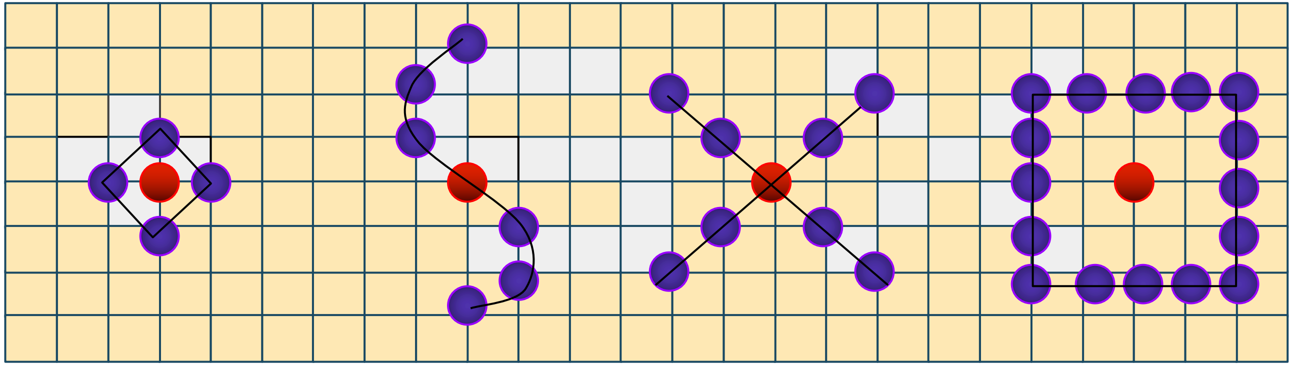
```
inline Stencil<float> udf_Window_Aggregates(const Stencil<float> &iStencil)
{
    Stencil<float> oStencil;
    oStencil =   iStencil(0, 0) + iStencil(0, 1) + iStencil(0, 2)
               + iStencil(1, 0) + iStencil(1, 1) + iStencil(1, 2)
               + iStencil(2, 0) + iStencil(2, 1) + iStencil(2, 2);
    return oStencil;
}

int main(int argc, char *argv[]){
   B = A->Transform(udf_Window_Aggregates)
}
```



Stencil: Abstract Data Type

Stencil

- An abstract data structure to represent a neighborhood of an Array
- Definition: *S*(Base Cell, Neighbor Cells -- relative offsets)



Structural Locality

Flexible geometric shapes to represent patterns of computation

An Example of 3-point Moving Average $(v_{t-1}+v_t+v_{t+1})/3$

```
int main(int argc, char *argv[])
{
    //Init the MPICH, etc.
    FT_Init(argc, argv);
    // set up the chunk size and the overlap size
    std::vector<int> chunk_size = {4, 4};
    std::vector<int> overlap_size = {0, 1};
    //Input data
    Array<float> A("EP_HDF5:tutorial.h5:/data", chunk_size, overlap_size);
    //Result data
    Array<float> B("EP_HDF5:tutorial_ma.h5:/data");
    //Run
    A.Transform(udf_ma, B);
    FT_Finalize();
    return 0;
}
inline Stencil<float> udf_ma(const Stencil<float> &iStencil)
{
    Stencil<float> oStencil;
    oStencil = (iStencil(0, -1) + iStencil(0, 0) + iStencil(0, 1)) / 3.0;
    return oStencil;
}
```

Input Array A, a 2D 16 x 16 dataset in HDF5 file, where each row is a time series from a sensor

Output Array B, a 2D dataset in HDF5 file

Execute the Transform, either sequentially or in parallel

Rules of the Transform from A to B

$$\frac{V_{t-1} + V_t + V_{t+1}}{3}$$

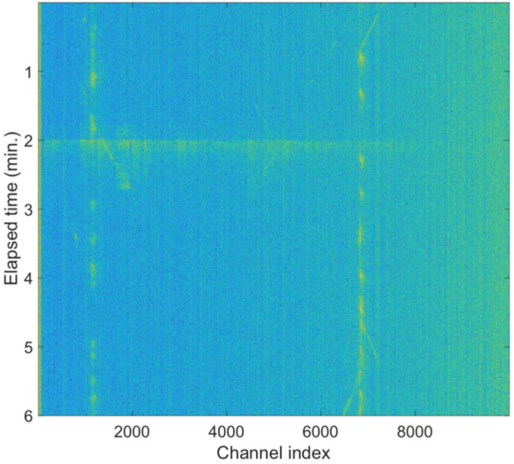
Relative offsets

Distributed Acoustic Sensing and its Data Analysis

DAS: Distributed Acoustic

- Record strain or strain-rate along fiber-optic cables in subsurface
- Provides high spatial and temporal resolutions for geoscience, e.g., earthquake detection, seismic imaging

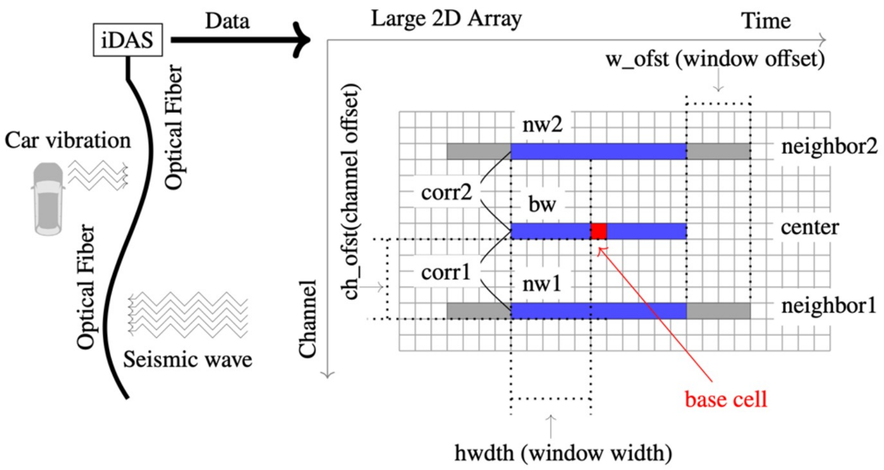
Data Analysis Challenges

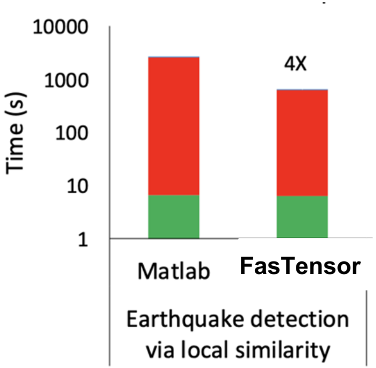


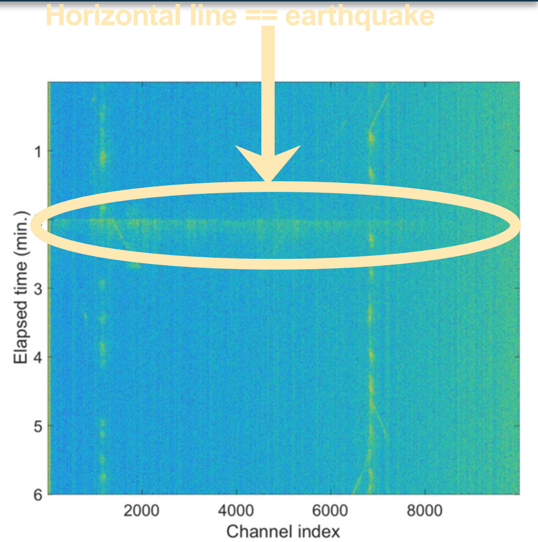
- DAS data size is large (TB/day), but scattered among many files
- Different analysis operations are required in different DAS data investigations.

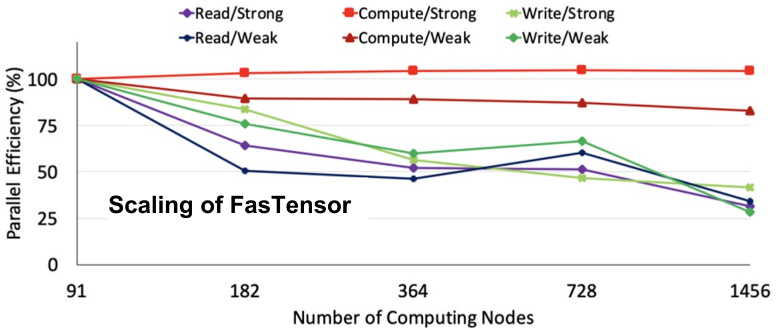
Task 1: Extract Earthquake Signals through Similarity

- Use local self-similarity calculation to identify earthquakes
- Could detect small earthquakes frequently missed









What FasTensor brings:
- Efficiently handling deep ghost zone (~500 layers)
- Easy parallelization

Dong, Bin, V. R. Tribaldos, etc. IPDPS 2020

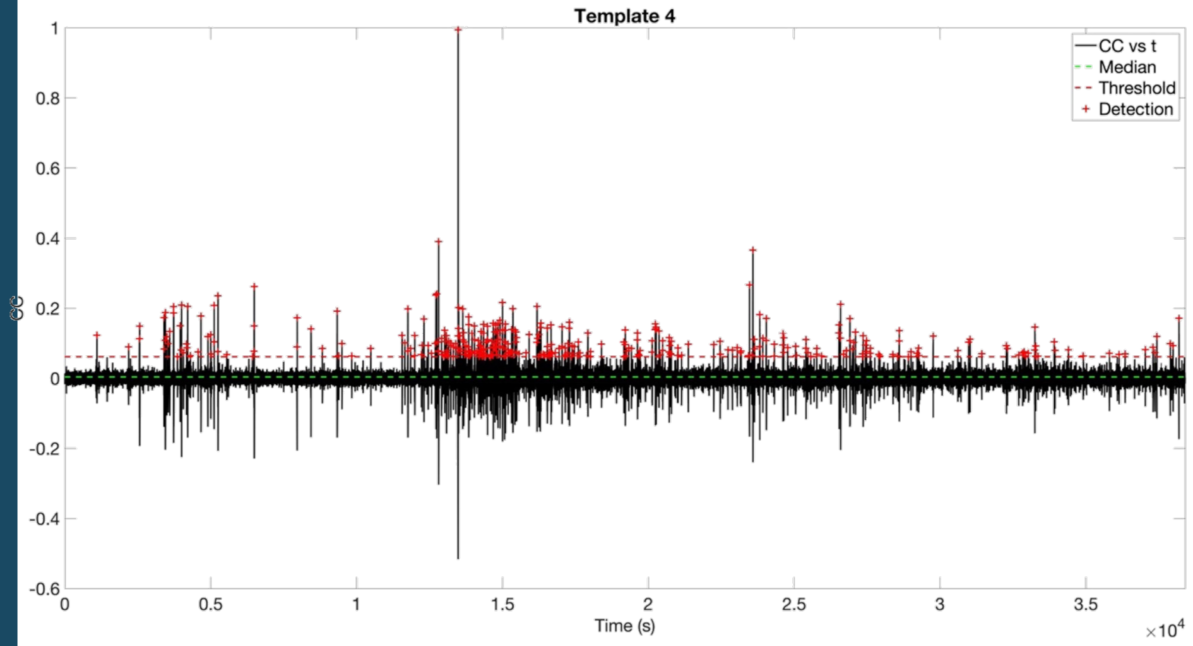
Task 2: Detect Even Smaller Earthquakes with Template

- For earthquakes less than magnitude 3, the above self-similarity approach is not sensitive enough
- Since earthquakes emanating from the same fault locations are likely to have the same wave form, could use a stronger quake as the template to find a weaker one by matching their shapes
- The matching is computed as similarity between a template and moving windows of observation data → <u>VERY MANY similarity calculations</u>
- Use FasTensor to parallelize the calculations (on-going work)

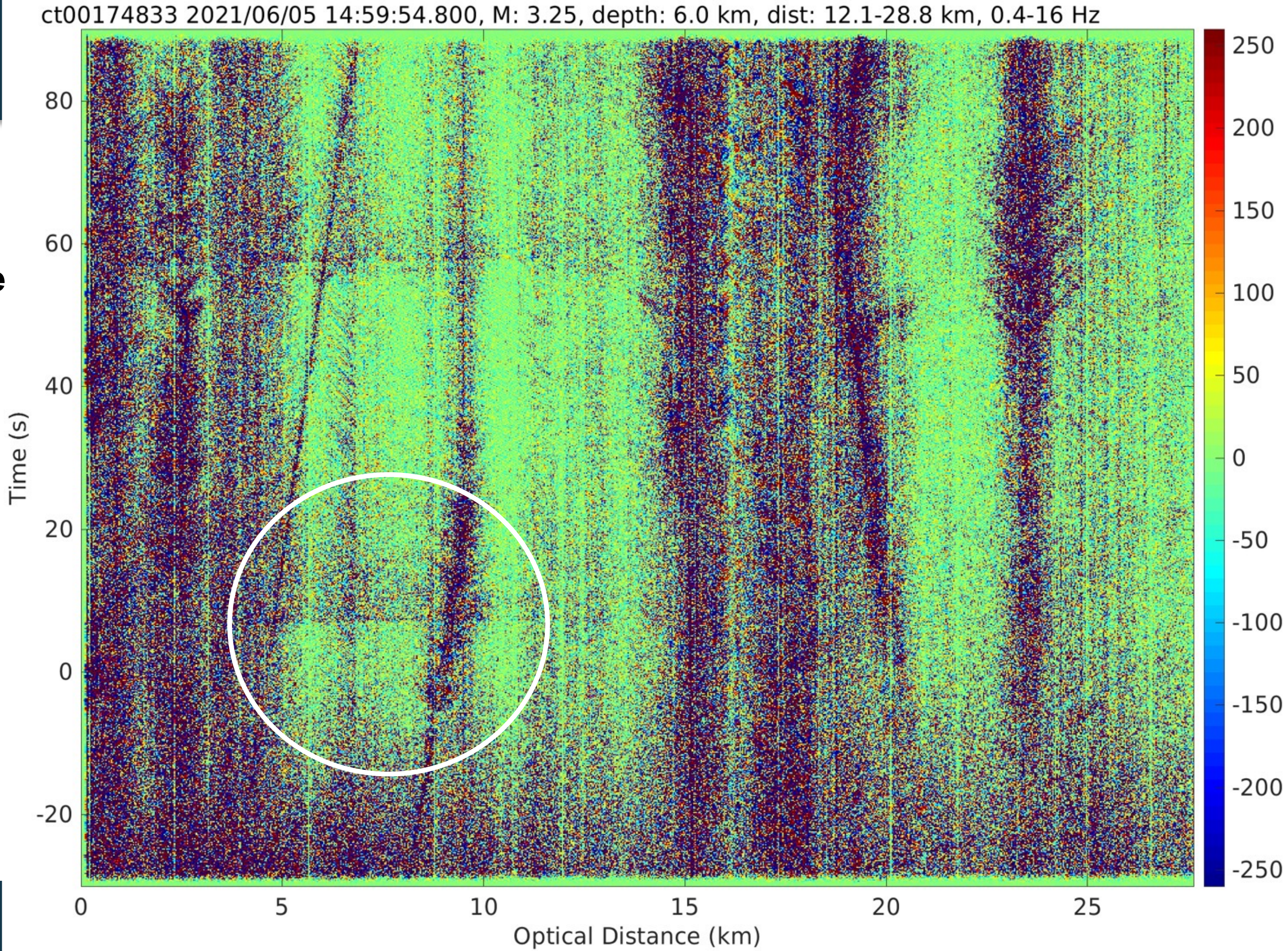
BERKELEY LAB

Template Matching Example

- Threshold = **12*std** + median
- 354 detections above threshold in unique 4 s windows
- 423 total earthquakes in catalog in this 10-hour period
- CC == cross correlation



Previously unknown earthquake example



BERKELEY LAB

FasTensor Summary

FasTensor website:
https://sdm.lbl.gov/fastensor/

Scientific Achievement

FasTensor, a <u>data parallel execution engine</u> for user-defined analysis, significantly reduces programming effort for various scientific analysis operations. It outperforms popular Big Data platforms such as Spark by ~50X to ~90X in executing machine learning computations.

Significance and Impact

FasTensor has been evaluated using:

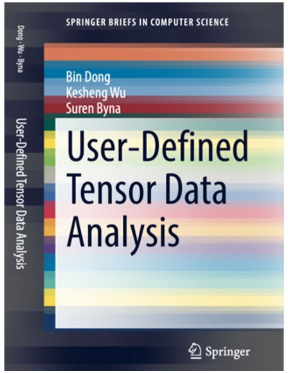
- **Earth science for detecting earthquakes and other subsurface events**
- **Fusion science for tracking field evolution**
- **Climate data analysis with Convolutional Neural Network (CNN) to predict extreme weather events**

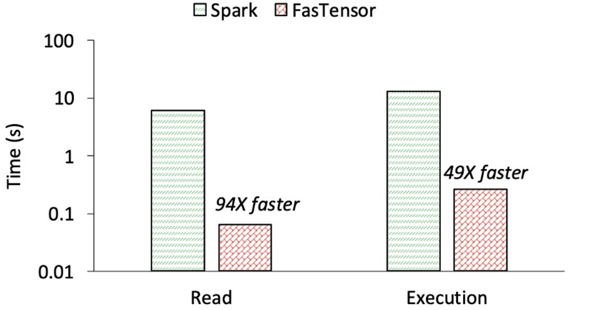
Research Details

FasTensor programming model consists of:

- **Simple data model (i.e, Stencil) abstraction well known in numerical computing**
- **Single operator (i.e., Transform) to execute user-defined analysis**
- **An execution engine for automatic parallelization**

Book (Springer 2021)





Performance comparison of FasTensor with Spark for completing CNN (CONV, Pooling and ReLU) on a 2D climate (CAM5) data

Acknowledgments

- Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, support for the SDS project and a DOE Career award under contract number DE-AC02-05CH11231, under ASCR research program (Margaret Lenz, Hal Finkle, Laura Biven, Lucy Nowell), SciDAC program (Lali Chatterjee), and Exascale Computing Project

U.S. DEPARTMENT OF ENERGY | Office of Science

- National Energy Research Scientific Computing Center

NERSC

My Team -- Scientific Data Management at Berkeley Lab

Jean Luca Bez

Bin Dong

Junmin Gu

Mariam Kiran

Imtiaz Mahmud

Alex Sim

Houjun Tang

Wei Zhang

Suren Byna (Faculty)

Horst Simon (Affiliate)

Florin Rusu (Faculty)

Arie Shoshani (retired)

More info --

- **http://sdm.lbl.gov/**
- **http://crd.lbl.gov/sdm/**
- **https://www.facebook.com/sdmberkeley/**