

Data Mining with Hadoop at TACC

Weijia Xu

Data Mining & Statistics

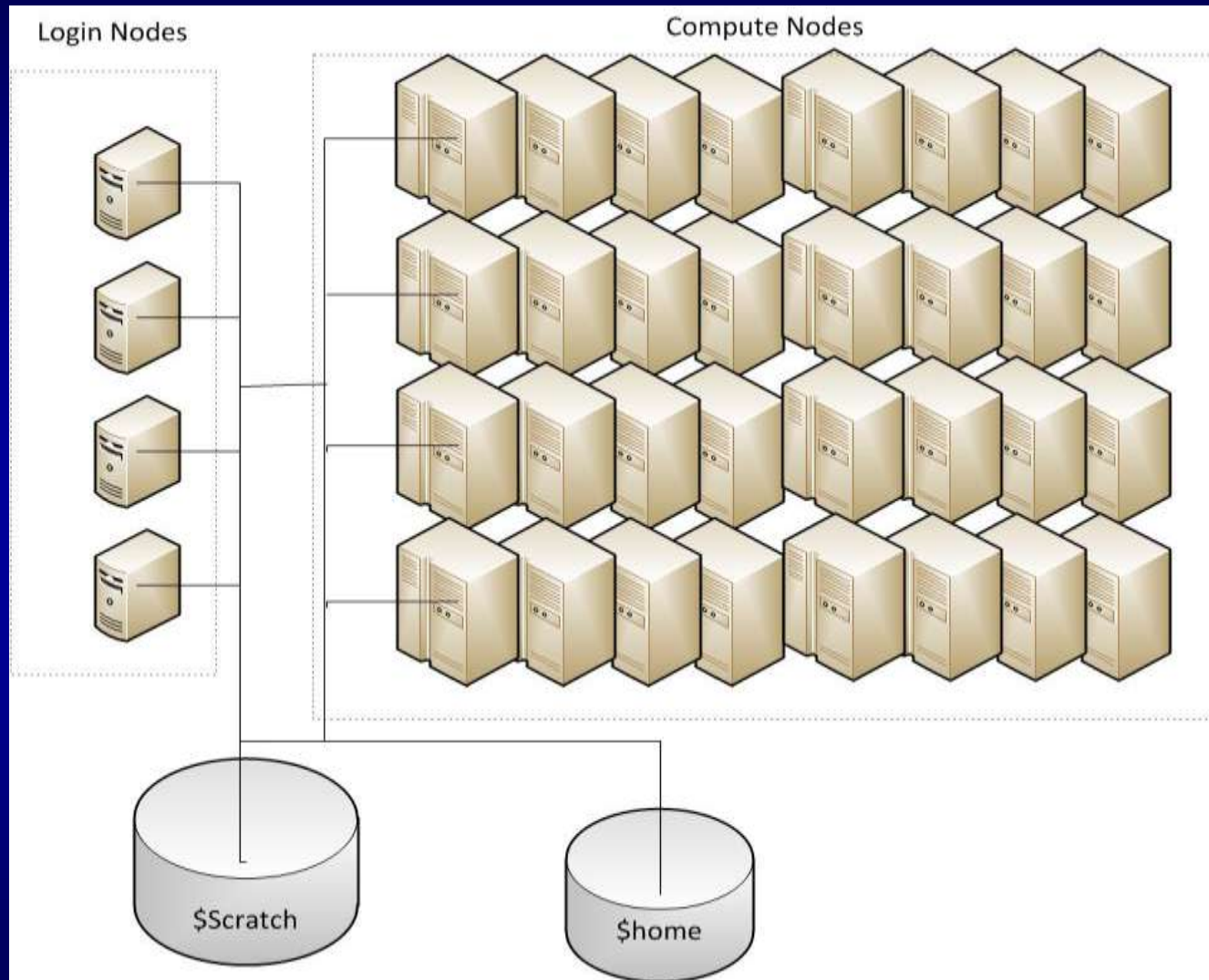
Data Mining & Statistics Group

- Main activities
 - Research and Development
 - Developing new data mining and analysis solutions for practical problems from various domains through collaborations.
 - Consulting
 - Providing advice on suitable analytic approaches for users.
 - Helping user to scale-up existing solutions.
- Active areas of collaborative projects
 - Data intensive computing
 - Statistical Computing with R
 - Visual Analytic

Outline

- Background
 - Common HPC system
 - Hadoop
- Enabling dynamic Hadoop cluster on Longhorn
- A project example

Background: Typical HPC Architecture



Background: HPC Architecture

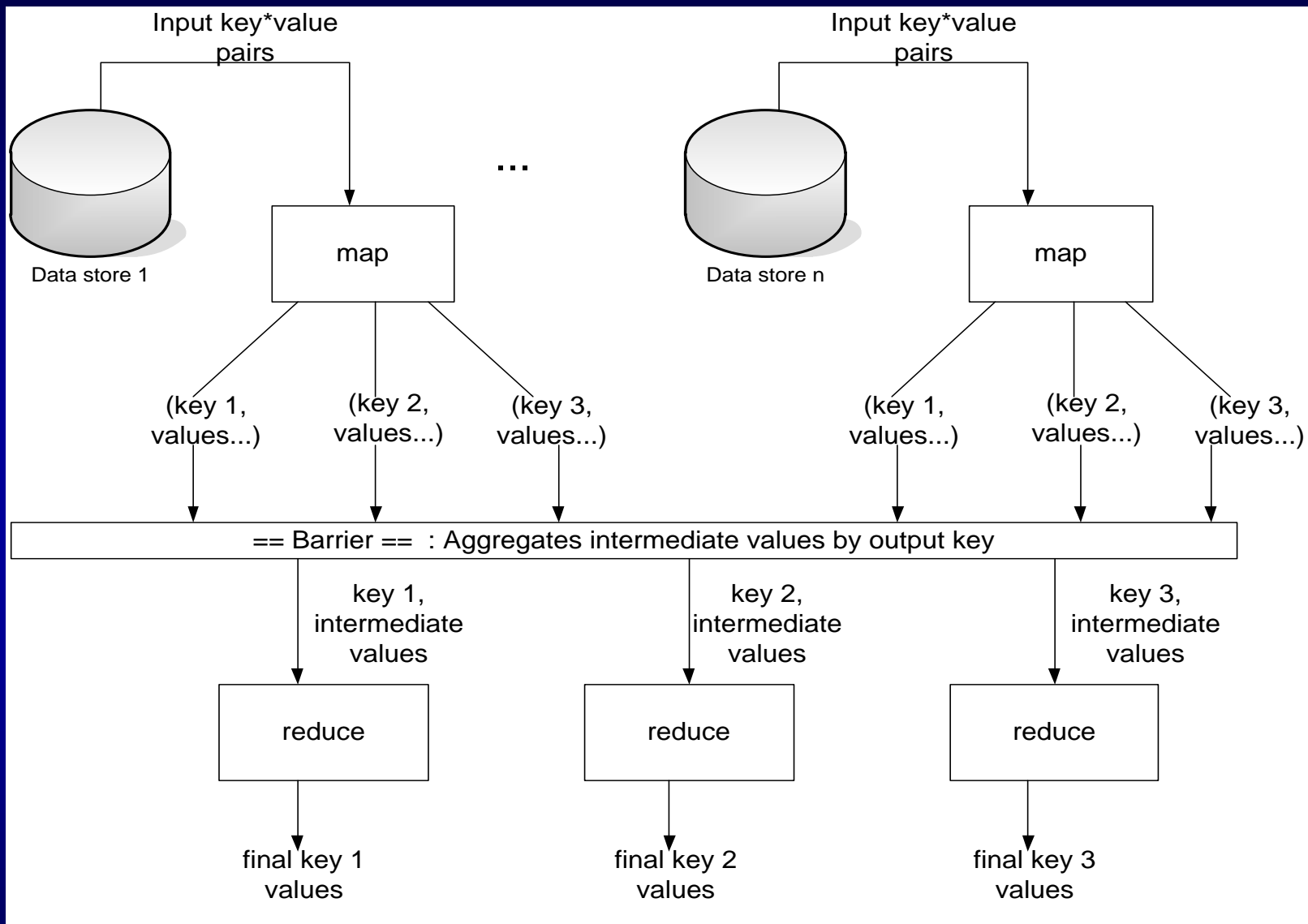
- All data (raw input, intermediate data, output) are stored in a shared high performance parallel file system.
- Each compute node can access all the data.
- No/minimum local storage at each compute node
- Data are transferred from storage to compute node memory via high performance network connections.

Background: Hadoop

- Hadoop is a library to enable efficient distributed data processing easily.
- Hadoop is an open source implementation of MapReduce programming model in JAVA with interface to other programming language such as C/C++, python.
- Hadoop includes several subprojects
 - HDFS, a distributed file system based on google file system (GFS), as its shared file system.
 - Mahout, scalable machine learning and data mining library
 - Pig, a high-level data-flow language and execution framework for parallel computation.
 - ...

Background: MapReduce Programming Model

- Data storage is distributed among multiple nodes.
- Bring the computation to the data.
- Automatic parallelization & distribution
- Users implement interface of two functions:
 - `map (in_key, in_value) -> (out_key, intermediate_value) list`
 - `reduce (out_key, intermediate_value list) -> out_value list`



Hadoop vs. HPC

- Data storage
 - HPC: a shared global high performance parallel file systems.
 - Hadoop: duplicated across all the data node.
- Parallelism
 - HPC: computing parallel
 - Hadoop: Data parallel
- Which one is better?
 - Is there enough memory?
 - How many times data need to be read/write?

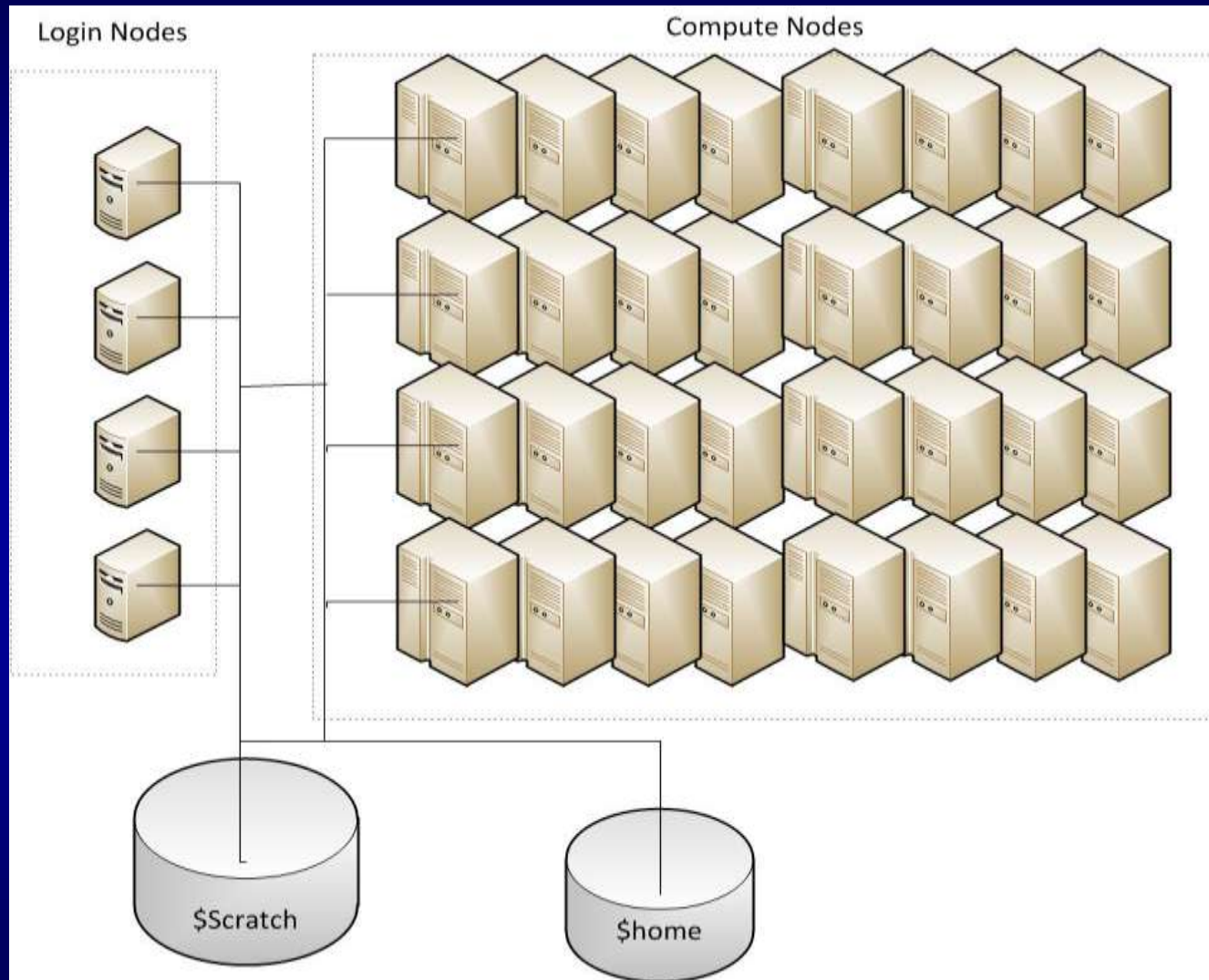
Enabling Dynamic Hadoop Cluster on Longhorn

- Motivations
 - To run existing hadoop code
 - To test/learn if the Hadoop is the right solutions.
 - To experiment what is the best Hadoop cluster setups.
 -
- Funded through Longhorn Innovation Fund for Technology (LIFT) through UT.

Goal

- Enable dynamically starting Hadoop cluster session by user through SGE.
 - Expanding local storage of Longhorn nodes
 - Data is distributed to local storage for further processing
 - Support data analysis jobs written with Hadoop library.
 - The Hadoop cluster is fully customizable by users.

Background: Typical HPC Architecture



Compute Nodes on Longhorn

- 240 Dell R610
 - 2 Intel Nehalem quad-core processors (8 cores)
@ 2.53 GHz
 - 48GB RAM, 73GB local disk
- 16 Dell R710
 - 2 Intel Nehalem quad-core processors (8 cores)
@ 2.53 GHz
 - 144GB RAM, 73GB local disk

Hadoop on Longhorn

- Local Storage Expansion
 - 192 500GB 7.2k drives are installed on 48 R610 nodes on Longhorn.
 - 112 146GB 15k drives are installed on 16 R710 nodes on Longhorn.
- /hadoop file system
 - Expanded Storage are available as /hadoop at each node.
 - R610 nodes and R710 nodes are accessible via different queue
 - 48 R610 nodes: ~2TB available per node, 96TB total available through Hadoop queue.
 - 16 R710 nodes: ~1TB available per node, 16TB total available through largemem queue

Hadoop job queue on Longhorn

- Queue priority on Longhorn
 - Normal queue has the highest priority,
 - Hadoop queue is same as the long, largemem queue.
- For a job request, computes nodes will be assigned in the following order to minimize the usage on “hadoop nodes”
 1. 192 R601 nodes without /hadoop
 2. 48 R610 nodes with /hadoop
 3. 16 R710 nodes with /hadoop

Running Hadoop on Longhorn

- Default job submission scripts are provided for user.
- User can start Hadoop session using any number of nodes up to 48 (384 cores).
- The name node and data node in Hadoop cluster are configured dynamically for each job.
- The configuration file is stored in the user home directories for user to customize settings.

Running Hadoop on Longhorn

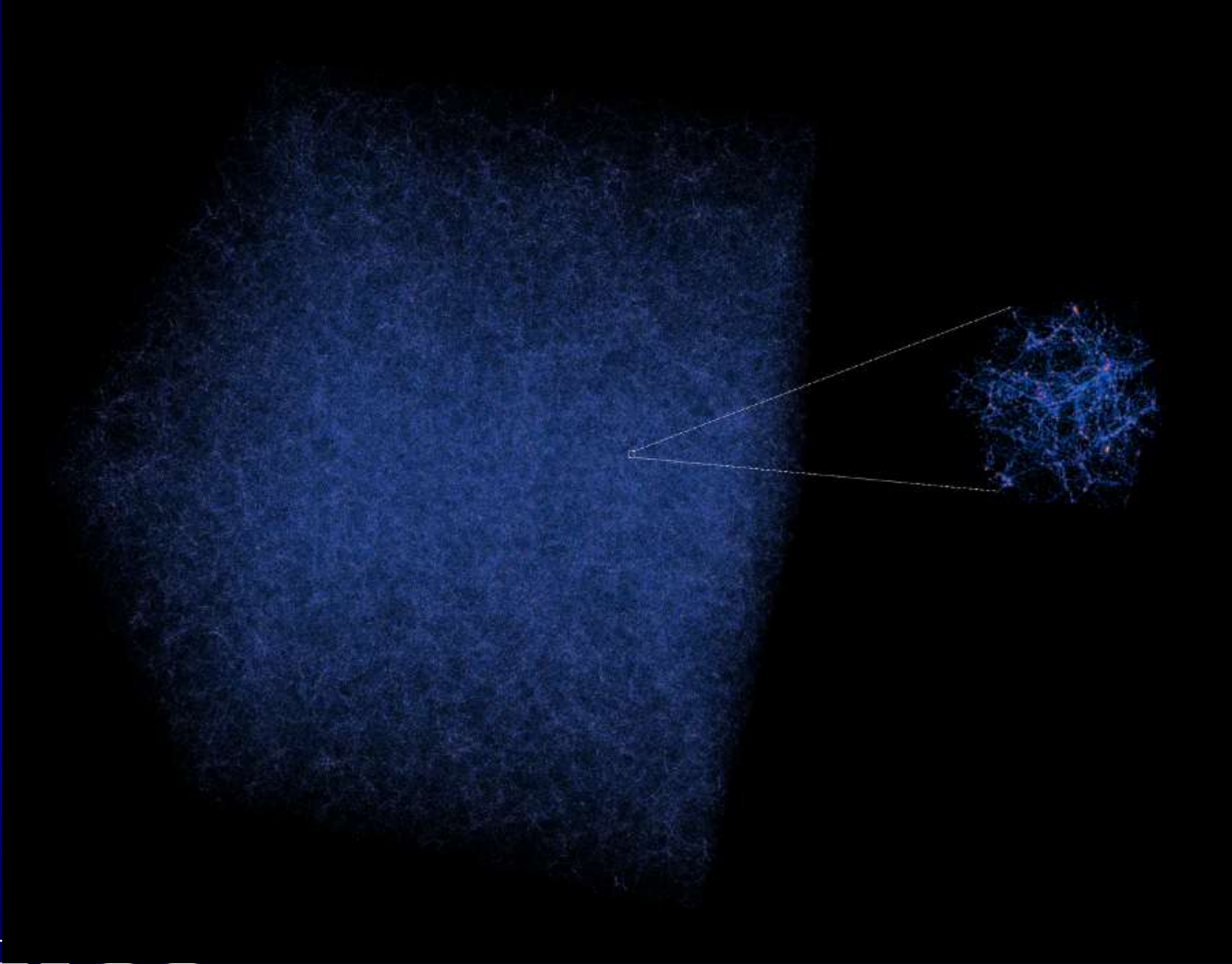
- Details at <https://sites.google.com/site/tacchadoop/>
- In a nutshell
 - Get and install tacc version Hadoop locally
 - Submit Hadoop scripts
 - Starts VNC sessions (optional)
 - Starts hadoop clusters
 - User interactions within VNC session / running data processing with existing Hadoop codes and scripts
 - User close vns session/ Job termination
 - Shutdown of hadoop clusters
 - Shutdown of vnc server.

Distributed, Scalable Clustering for Detecting Halos in Terascale Astronomy

- Problem:
 - ~200 Terabyte data has been generated through simulation to study evolution of the universe by astronomic researcher.
 - A 1GB data file can contain about ~15 million data objects
 - Difficult to analyze or visualize.
- Goal: exploring computational solution for
 - Automatically identify regions of interests as halo candidates for detailed rendering
 - Automatically identify substructure within halo candidates
- Algorithmic approach:
 - A density clustering based approach.
 - Identify dense area iteratively,
 - “shaves off” points in sparse area during each iteration.

Scale of the problem

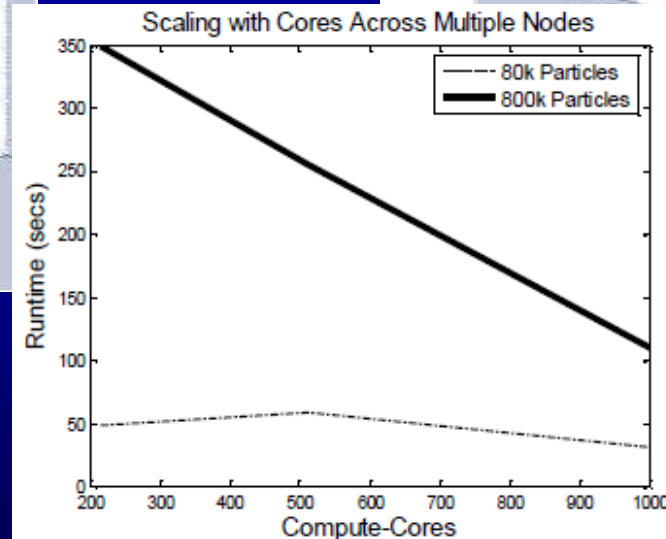
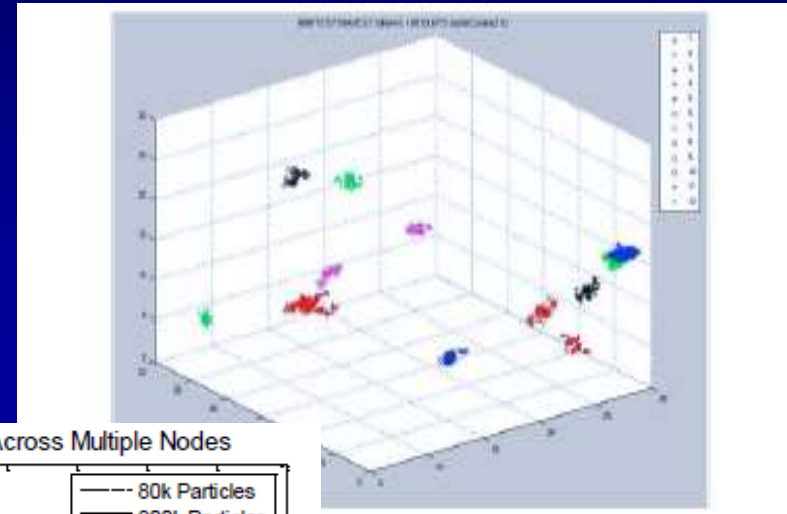
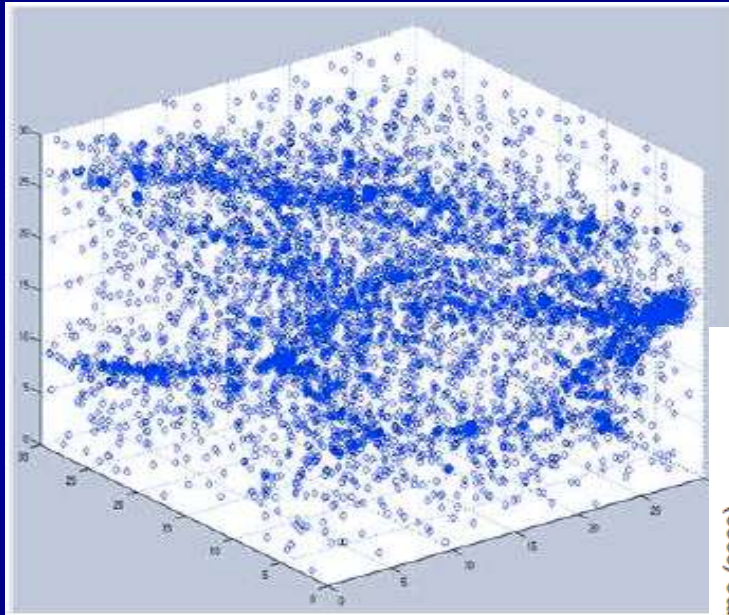
- An image demonstrates the scale of the problem.
 - http://www.utexas.edu/cio/itgovernance/lift/articles/images/cloud_large_1.png
 - An image demonstrate the scale of the problem.
 - Image shows a potential interesting halo area out of 6k-cubed simulation result.
 - Image curtsey of Paul Navratil.



Distributed, Scalable Clustering for Detecting Halos in Terascale Astronomy

- Methods and status
 - Parallel data flow approach with Longhorn:
 - Distribute data into each compute node for independent processing
 - Leveraging the multicores for computations at each node
 - About ~57k data points can be processed per compute node per hour
 - Distributed computation using Hadoop with Longhorn:
 - Data is stored in a dynamically constructed hierarchical distributed file system using local hard drives at each compute node.
 - Using Map –Reduce computing model implemented with Hadoop
 - Up to ~600k data points can be processed per compute node per hour

Distributed, Scalable Clustering for Detecting Halos in Terascale Astronomy



Dhandapani et. al. LNCS 2011

Result demonstration

- A movie demonstrate the clustering results of a 20k particle data set.
 - http://tacc-web.austin.utexas.edu/staff/home/xwj/public_html/20k.mpeg
 - First you see the original dataset, with each point plotted as a dot.
 - Points from one cluster (cluster 11) are shown with isosurface rendering at density of 10.0, using a 32^3 radial gaussian kernel on a 512^3 grid for the density calculation.
 - Then points of the cluster are added to the isosurface.
 - Followed by a volume rendering of the density function with opacity ramping up from density=10 to the data set max.
 - Movie is curtsey of Greg Abram.

Other Ongoing Projects Utilizing Hadoop on Longhorn

- TACC involved projects
 - Large scale document collection summarization and analysis with iterative density based clustering
 - Bae et al., ICDAR 2011
 - Classification of web page archives with supervised learning.
- User projects
 - Improve Blog Ranking by Mining Memes,
 - PI: Matt Lease, iSchool
 - TextGrounder, geo-spatio tagging of texts,
 - PI: Jason Baldrige, UTCL

Acknowledgment

--Enabling Cloud Computing at UT Austin

- People:
 - TACC staff:
 - Weijia Xu, Data Mining & Statistics
 - External collaborators:
 - Dr. Matthew Lease, Department of Computer Science / iSchool
 - Dr. Jason Baldridge, Computational Linguistics.
- Related References and links
 - Hadoop @ TACC wiki site:
 - <https://sites.google.com/site/tacchadoop/>
 - Feature story on recent project progress
 - <http://www.utexas.edu/cio/itgovernance/lift/articles/cloud-computing-update.php>
- Funding is provided through Longhorn Invoational Fund for Technology

Acknowledgement

-- Halo Detection Project

- People:

- TACC staff:

- Weijia Xu, Data Mining & Statistics
 - Paul Navratil, Visualization & Data Analysis

- External collaborators:

- Dr. Joydeep Ghosh, ECE, UT Austin
 - Sankari Dhandapani, Graduate student
 - Sri Vatsava, Graduate student
 - Dr. Nena Marin, Pervasive Inc., Austin, TX
 - Dr. Ilian Iliev, Astronomy, University of Sussex, UK

- Related Publication:

- Presented at KDCloud '10 in during ICDM '10

- Daruru, S., Dhandapani, S., Gupta, G., Iliev, I., Xu, W., Navratil, P., Marin, P. and Ghosh, J. (2010) Distributed, Scalable Clustering for Detecting Halos in Terascale Astronomy” in workshop proceedings of *IEEE International Conference on Data Mining: Knowledge Discovery Using Cloud and Distributed Computing Platforms (KDCloud'10)* Dec. 2010, Sydney, Australia, IEEE Computer Society, Washington, DC, USA, p.p. 138-147,

- Extended version to appear on LNCS

- Dhandapani, S., Daruru, S., Gupta, G., Iliev, I., Xu, W., Navratil, P., Marin, P. and Ghosh, J. (2011) Distributed, Scalable Clustering for Detecting Halos in Terascale Astronomy Datasets, *Lecture Notes in Computer Science (LNCS)*, in press

Question and Discussion

Additional References

- Hadoop @ TACC wiki site:
 - <https://sites.google.com/site/tacchadoop/>
- Feature story on recent project progress
 - <http://www.utexas.edu/cio/itgovernance/lift/articles/cloud-computing-update.php>