Parallel Feature Extraction from Large Science Data Sets

John Wu, Alex Sim, Suren Byna, Nagiza F. Samatova, Scott Klasky, Arie Shoshani, Xiaocheng (Chris) Zou, Lingfei Wu, Houjun Tang
CS Chang, Michael Churchill, Daniel Martin, Hans Johansen, Andreas Stathopoulos, Jong Choi

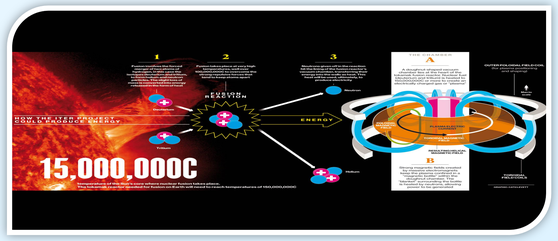
OVERVIEW

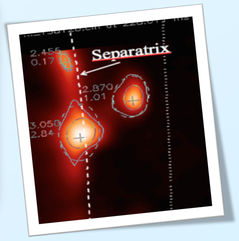
The SDM group at LBNL develops data management technology for large-scale analyses. This poster outlines two recent use cases where the spatial features are extracted from data records about individual mesh cells: the ice calving events in ice sheet modeling and blob detection in fusion data processing.

BLOBS IN MAGNETIC CONFINEMENT FUSION

❶ **Why fusion?**
- ✧ Fusion is a viable future energy
- ✧ Advantages of fusion: Inexhaustible, clean, and safe

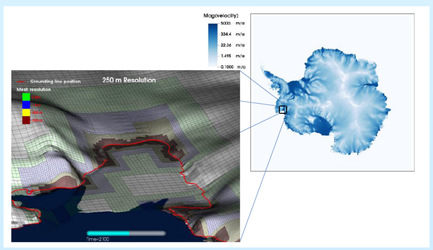
❷ **Blobs & Disruption Events**
- ✧ Continuous fusion requires steady plasma confinement
- ✧ Blob is associated with disruption of confinement
- ✧ Blobs carry hot plasma towards the interior wall of tokamak
- ✧ Blobs transport heat away from plasma core, degrade plasma confinement and damage the wall.





ICE CALVING

- ➢ A process of producing **free-floating icebergs** and **ice fracture**
- ➢ **Large-scale calving events** (e.g., iceberg twice the size of Atlanta breaks off of Antarctica) are highly important to global **sea level changes**
- ➢ Occasional ice calving events produce disconnected portions of floating ice shelves, lead to an **ill-posed system** in ice sheet modeling, and cause the **solvers to diverge**



Schematic showing computed ice velocity for Antarctic continent (right), and meshing for the Pine Island Glacier (left). The grounding line location is shown as red line.

BLOB DETECTION APPROACH

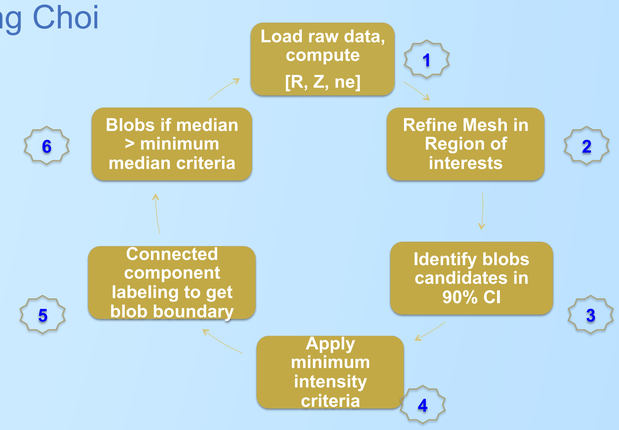
❖ **Basic approach: develop a real time outlier detection algorithm for finding blobs in numerical simulations and fusion experiments**

❖ **Outline of the blob detection approach:**

❶ The process i loads raw simulation or experimental data in each time frame and computes normalized intensity
❷ Refine the triangular mesh obtained in the region of interests
❸ Apply outlier detection to Identify blob candidates with 90% confidence level in the region of interests
❹ Compare the intensity of blob candidate with minimum intensity criteria
❺ Use connected component labeling to compute the size of different blobs
❻ A blob is found if its median satisfies minimum median intensity criteria

❖ **Hybrid MPI/OpenMP parallelization:**

❶ In high-level, use MPI to allocate n processes to process each time frame
❷ In low-level, use OpenMP to accelerate the computations with m threads

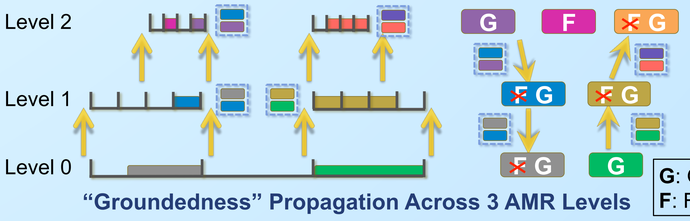


ICE CALVING ALGORITHM ON AMR DATA

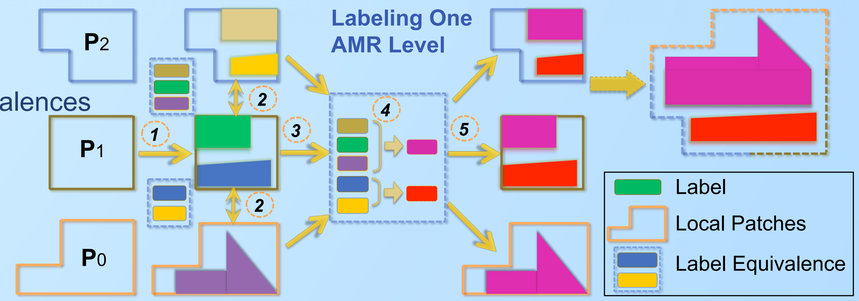
Part A: AMR-aware Connected Component Labeling

➢ **A.1 Multiple stages of labeling one AMR level**

1. Discover local connectivity: assign local labels, record label equivalences
2. Exchange labels on boundary and collect expanded equivalence
3. Aggregate label equivalences to the selected processor
4. Determine final labels using aggregated label equivalences
5. Disseminate final labels to each processor



"Groundedness" Propagation Across 3 AMR Levels

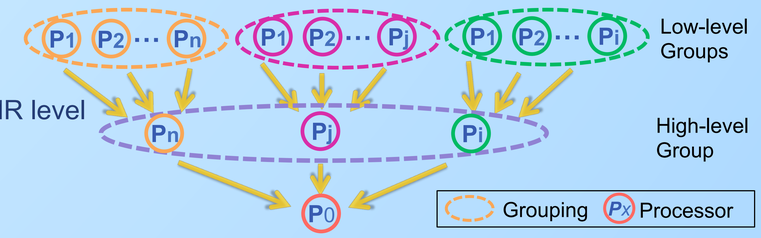


➢ **A.2 Propagate component's "Groundedness" across AMR levels**

1. Collect label equivalence between AMR levels using Chombo
2. Propagate "Groundedness" along the label equivalence chain to determine the "Groundedness" of the whole component

Part B: Hierarchical Data Aggregation

➢ Divide entire processors into two-level computing groups:
Low-Level group: a group of processors with data from the same AMR level
High-level group: selected processors from low-level groups
➢ Collect label equivalences and "Groundedness" info in parallel
➢ Avoid expensive global data communication overhead



SDAV Scalable Data Management, Analysis, and Visualization
NC STATE UNIVERSITY
BERKELEY LAB Lawrence Berkeley National Laboratory
U.S. DEPARTMENT OF ENERGY | Office of Science