High-Performance Algorithms for Large-Scale Singular Value Problems and Big Data Applications

Lingfei Wu, Andreas Stathopoulos (Advisor), College of William and Mary, Williamsburg, Virginia, USA

Introduction

❖ The Era of Big Data
- Increasing influence on our daily life and research activities
- Poses significant challenges both on memory requirements and computational expense in various research areas
- Some applications demand fast solution of large-scale SVD problems
- Some others require extracting knowledge from large-scale data in real time

❖ The Objective of This Dissertation
- Develop efficient numerical methods and practical data mining techniques to cope with very large-scale problems on extremely large parallel machines

❖ Main Contributions of This Dissertation
- Propose a preconditioned two-stage SVD method that significantly advances the current state-of-the-art in singular value problem solving
- Develop a high quality SVD software supporting accurate computation of both largest and smallest singular triplets on a massively parallel machine
- Present a high-performance region outlier detection method for finding blob -filaments in real fusion experiments or numerical simulations

Motivation and Related Work

❖ Importance of Efficient SVD Solver
- Largest and smallest SVD problems: machine learning, image processing, control theory, least squares problem and low rank approximation

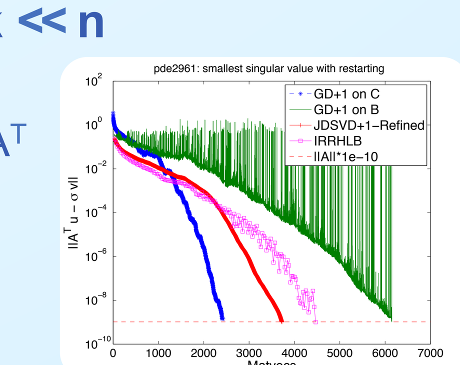
❖ The problem: find k extreme singular triplets of $A^{m \times n}$

$$Av_i = \sigma_i u_i, \quad i = 1, 2, \ldots, k, \quad k \ll n$$

❖ Iterative methods for computing SVD:
1. Hermitian eigenvalue problem on $C = A^TA$ or $C = AA^T$
2. Hermitian eigenvalue problem on $B = [0\ A^T; A\ 0]$
3. Lanczos bidiagonalization method (LBD)

$$A = PB_dQ^T \text{ and } B_d = X\Sigma Y^T$$

where $U = PX$ and $V = QY$.

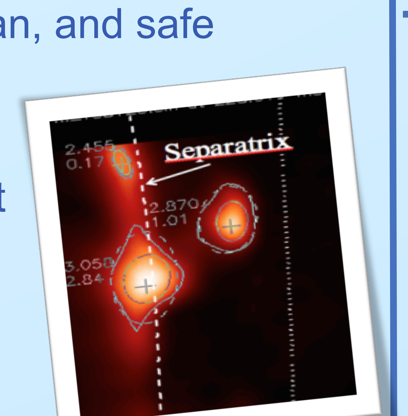


❖ Scarcity of SVD software for large-scale problems

Software	State-of-the-art Methods	Classic Methods	Lang	MPI/ SMP	Precond-itioning	Extreme SVs	Fast Full Accuracy
PRIMME	PRIMME_SVDS	Multimethod	C	Both	Y	Y	Y
N	IRRHLB	N/A	Matlab	N	N	Y	Y
N	IRLBA	N/A	Matlab	N	N	Y	Y
N	JDSVD	N/A	Matlab	N	Y	Y	Y
N	SVDIFP	N/A	Matlab	N	Y	Y	N
SLEPc	N/A	Many	C	MPI	Y	Y	N
PROPACK	N/A	LBD	F77	SMP	N	Y	N
SVDPACK	N/A	Lanczos	F77	N	N	Y	N

❖ Importance of Fusion Energy
- Fusion is a viable energy source in future: Inexhaustible, clean, and safe

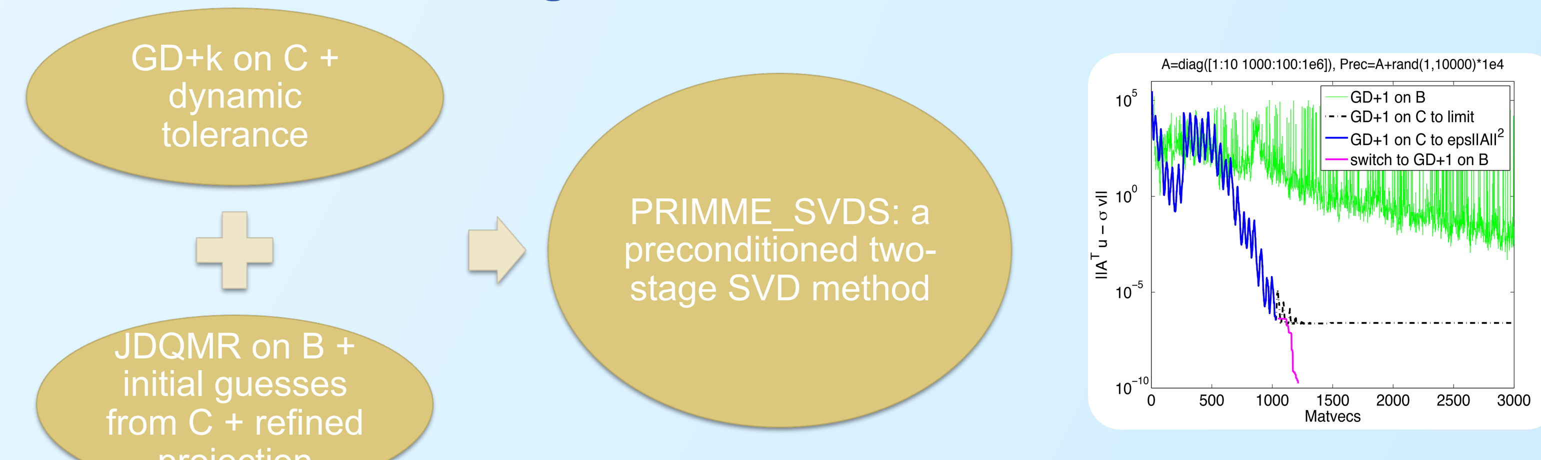
❖ Tokamak and Blobs
- Blobs carry high energy and plasma outside the magnetic confinement that causes loss of heat, degrading confinement



❖ Difficulties in large-scale data analysis
- Generating massive data: a few terabytes per second!
- Single-threaded, only for post-run analysis and slow
- Real fusion experiments demand real-time data analysis

PRIMME_SVDS: A High-Performance SVD Solver

- ❖ **Our goal:** solve large-scale, sparse SVD problems with unprecedented efficiency, robustness and accuracy
- ❖ **Our approach:** a preconditioned hybrid, two-stage SVD method on top of the state-of-the-art eigensolver PRIMME



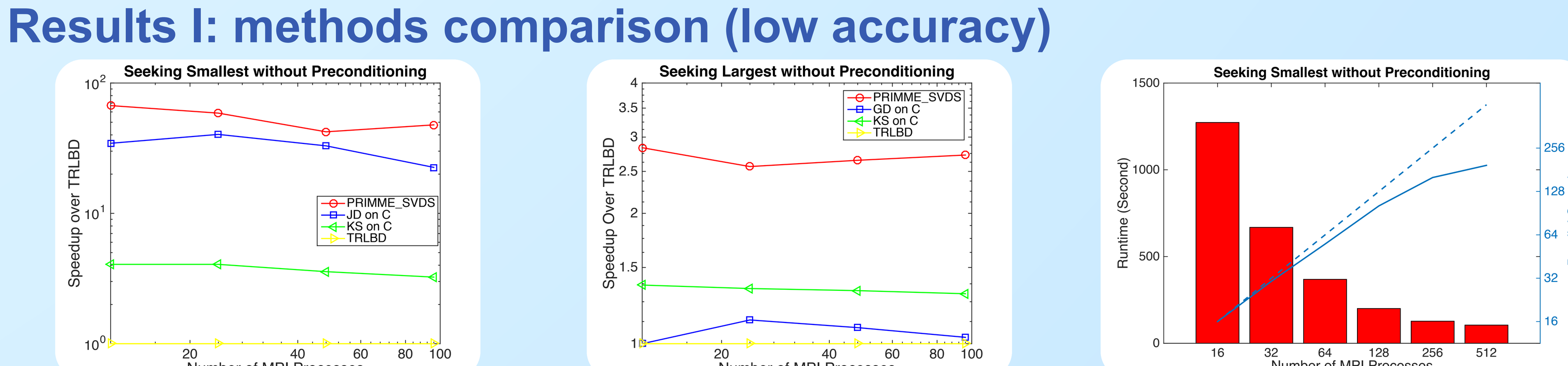
Hybrid, two-stage SVD method and an example

❖ Parallel Implementation of PRIMME_SVDS
- User defined data distribution among processes
- User defined parallel matrix-vector and preconditioning functions
- User provided global summation for dot products

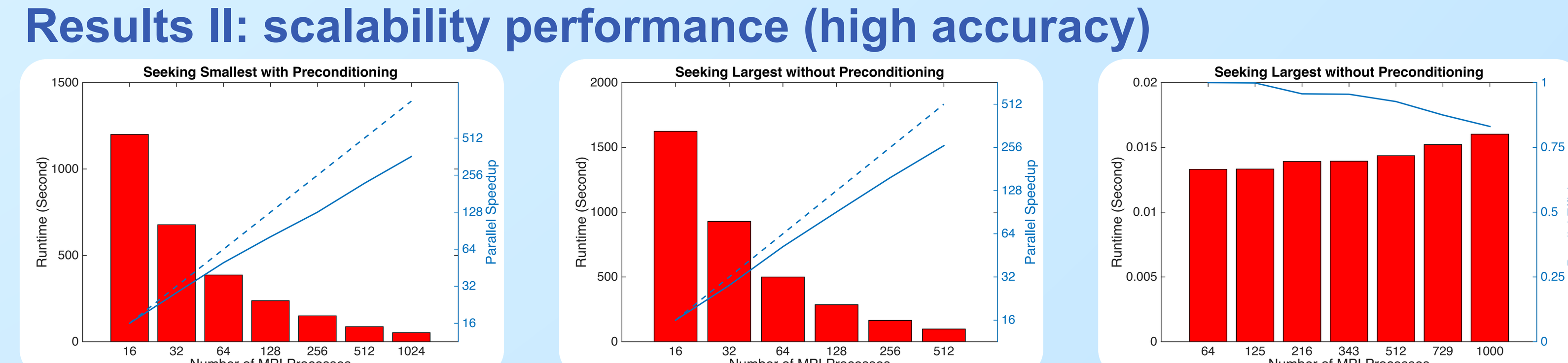
Operations	Kernels or Libs	Cost per iteration	Scalability
Dense algebra: MV, MM, Inner-Prods, Scale	BLAS (eg, MKL, ESSL, OpenBlas, ACML)	O((m+n)*numSVs)	Good
Sparse algebra: SpMV, SpMM, Preconditioner	User defined (eg, PETSc, Trilinos, HYPRE, librsb)	O(1) calls	Application dependent
Reduction	User defined (eg, MPI_AllReduce)	O(1) calls of size O(numSVs)	Machine dependent

❖ Experimental setup and results
- Applications: DNA, Least squares, Graph clustering, and QCD (with problem size ten million)
- Matrices and vectors distributed on total 1024 processes on Edison or 96 on SciClone (W&M)

❖ Results I: methods comparison (low accuracy)

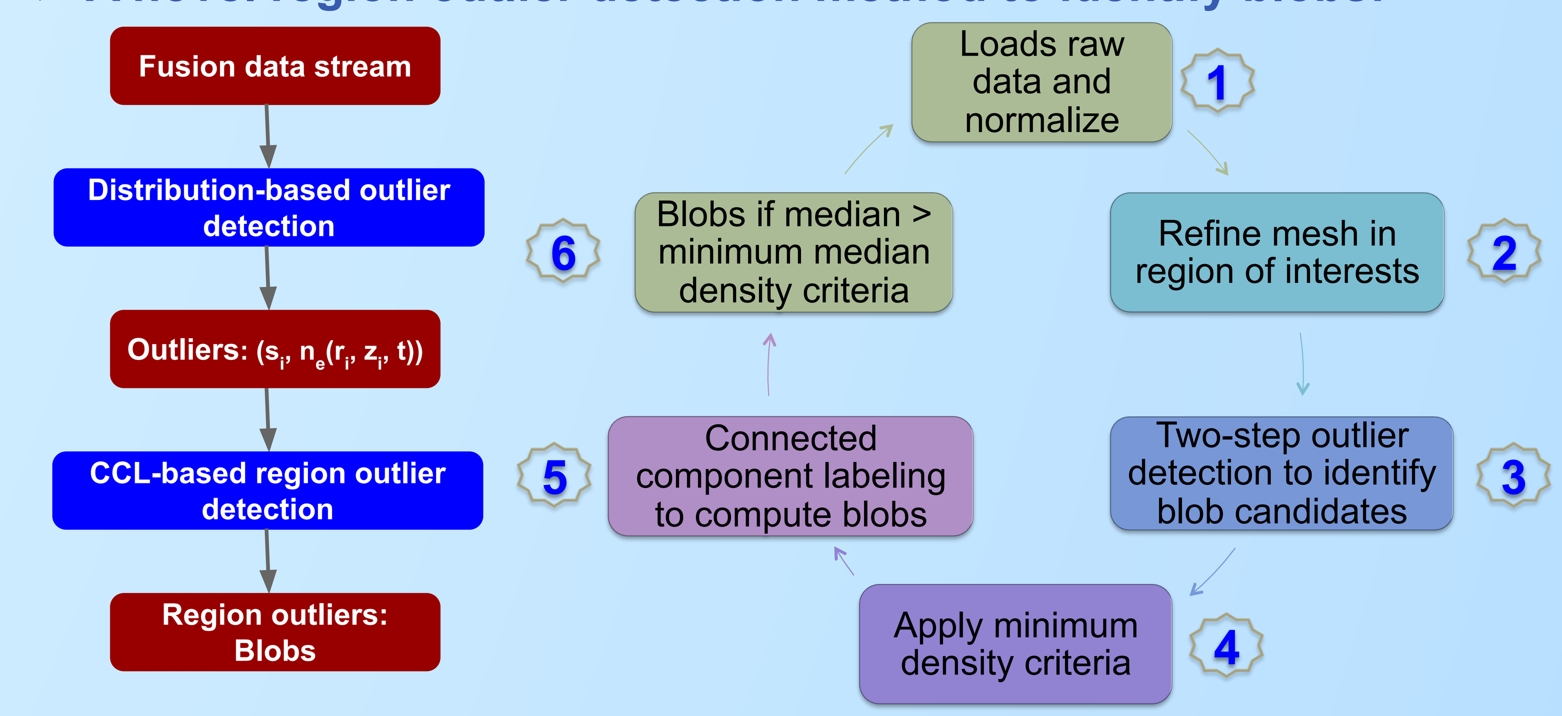


❖ Results II: scalability performance (high accuracy)



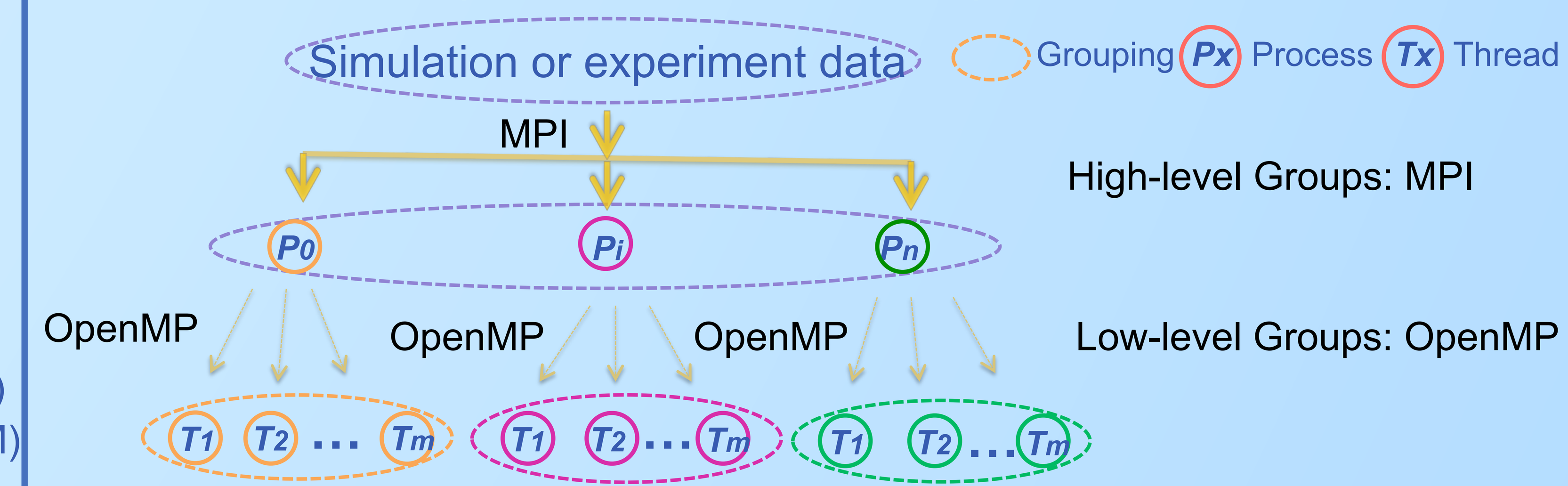
A High-Performance Blob Detection Approach

- ❖ **Our approach:** develop a real-time region outlier detection algorithm for finding and tracking blobs in fusion data streams
- ❖ A novel region outlier detection method to identify blobs:

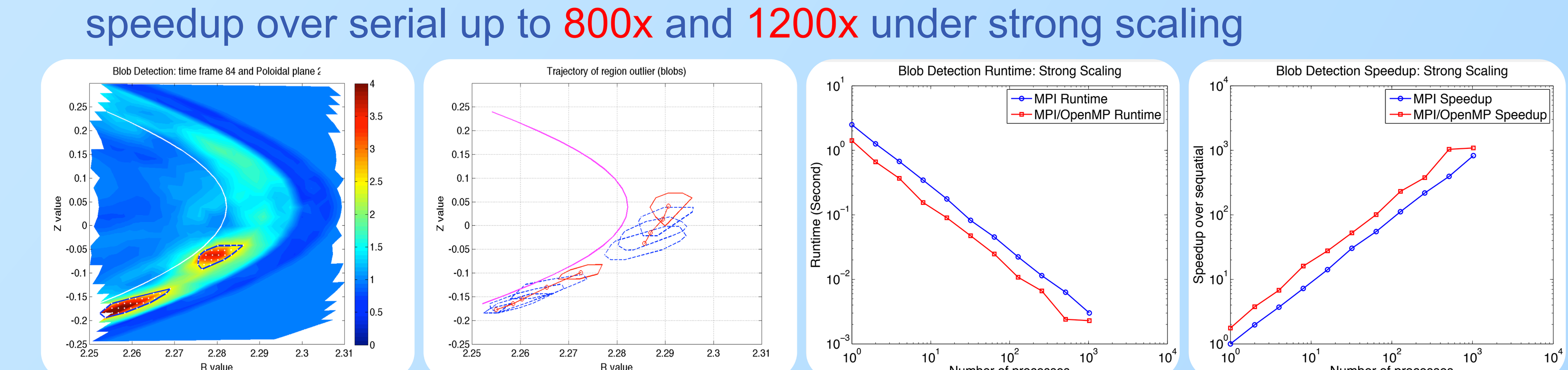


Two-phase region outlier detection algorithm

❖ A real-time blob detection approach using MPI/OpenMP:
1. High-level: use MPI to allocate n processes to process each time frame
2. Low-level: use OpenMP to accelerate the computations with m threads



❖ Experimental setup and results
- Our data sets (33GB) is from the XGC1 simulation containing 1024 time frames which last around 2.5 milliseconds (ms)
- **Most Encouraging results:** complete blob detection in around 2 ms with MPI /OpenMP using 4096 cores and in 3 ms with MPI using 1024 cores
- MPI and MPI/OpenMP achieve linear time scalability in blob detection time with speedup over serial up to 800x and 1200x under strong scaling



Blob detection and tracking examples *Blob detection performance: strong scaling*

Conclusions and Reference

❖ Highlights of PRIMME_SVDS:
- Among the fastest and most robust production level software for computing a small number of singular triplets
- Demonstrates good scalability under both strong and weak scaling
- Free software, available at: https://github.com/primme/primme

❖ Highlights of Blob Detection and Tracking:
- The first work to achieve real-time blob-filaments detection in a few milliseconds and linear-time speed up
- Has been integrated into the ICEE project as data analysis component

[1] Lingfei Wu, et al, **A Preconditioned Hybrid {SVD} Method for Computing Accurately Singular Triplets of Large Matrices**, accepted to SISC, http://arxiv.org/abs/1408.5535, 2014.

[2] Lingfei Wu, et al, **Towards Real-Time Detection and Tracking of Blob-Filaments in Fusion Plasma Big Data**, submitted to IEEE transaction on big data, http://arxiv.org/abs/1505.03532, 2015.

This work is supported by NSF under grants No. CCF 1218349 and ACI SI2-SSE 1440700, and by DOE under a grant No. DE-FC02-12ER41890

WILLIAM & MARY

U.S. DEPARTMENT OF ENERGY