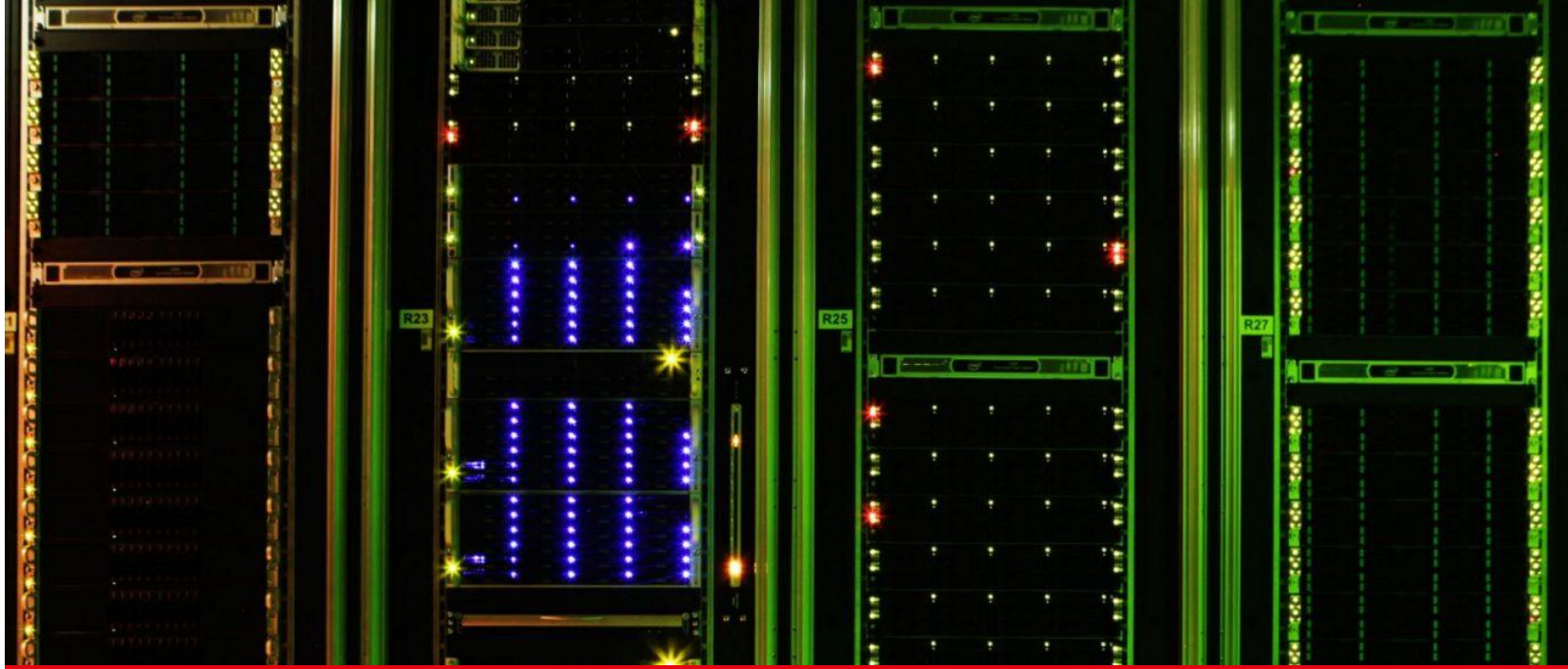
EPFL

EUROfusion



Parallel I/O in GENE-X Using ADIOS

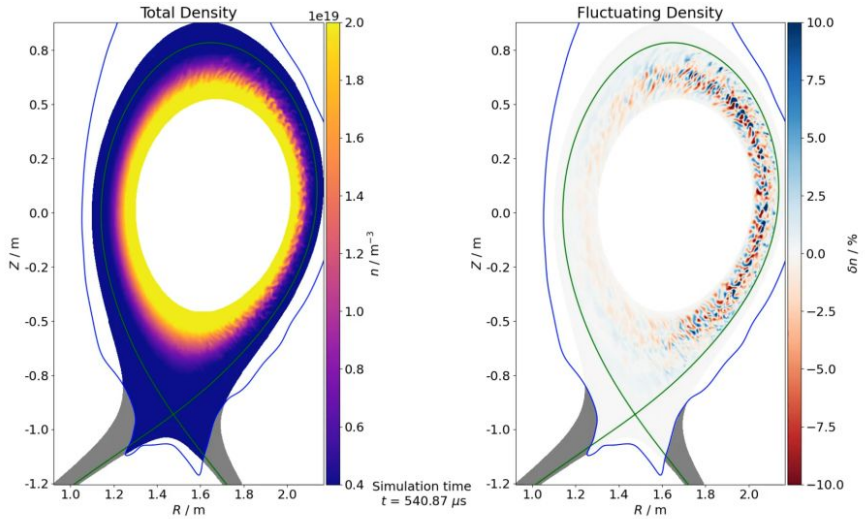
P. Panchal (EPFL),
P. Ulbl, J. Trilaksono, G. Merlo, Y. Ju (Max Planck Institute for Plasma Physics)

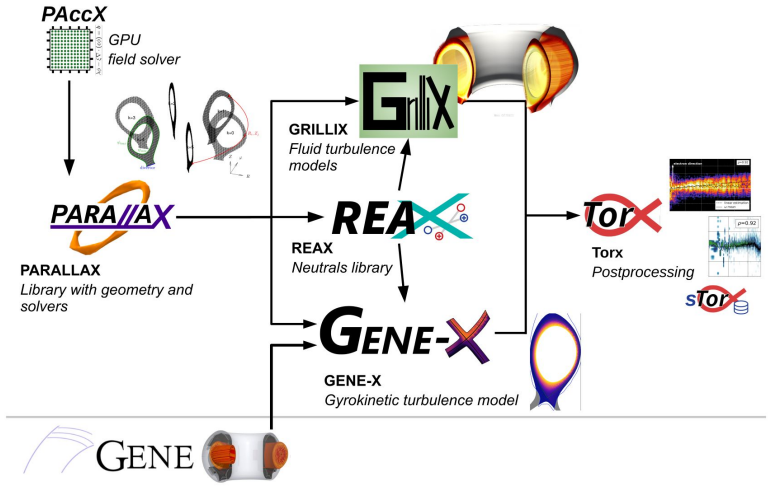
3rd Meeting of EUROfusion HPC ACHs
Tue. 25th of November 2025, Lausanne

GENE-X Overview

GENE-X simulates turbulence in fusion devices

- Gyrokinetic + Electromagnetic equation, Eulerian grid, full-f, collisional
- Fortran 2008. MPI + OMP parallelization:
  - Strong scaling with 93% efficiency up to 512 nodes (~20K cores), good weak scaling

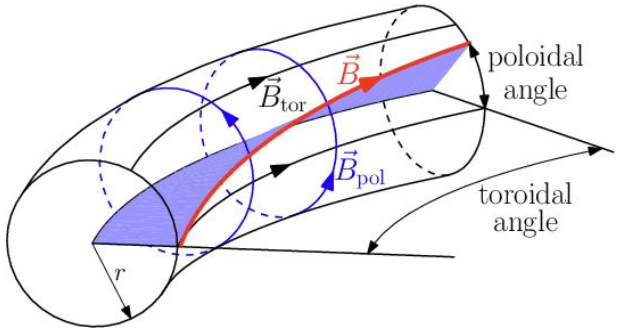


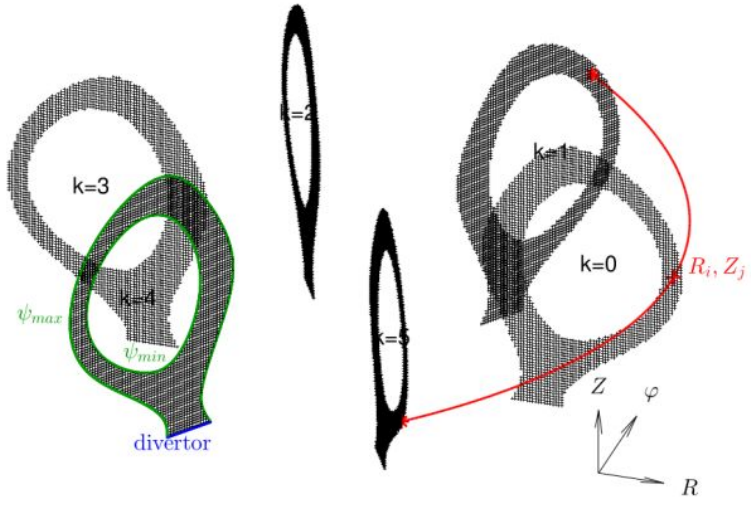


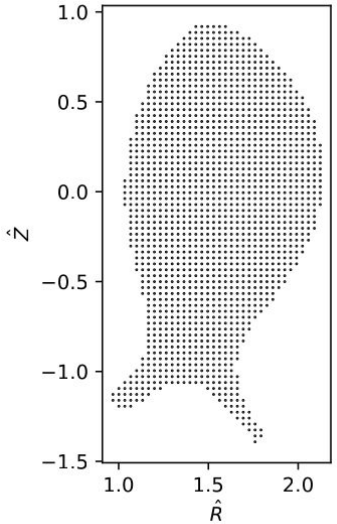
Source: P. Ulbl et al. (2024) “Towards first-principles simulations of the L- to H-mode transition with the global gyrokinetic turbulence code GENE-X”, TSVV4/1 Workshop 2024.

MPI Domain Decomposition

- Time evolution of a 5D distribution f, dims (RZ, phi, vpar, mu, alpha)
- Domain: tensor product of regular grids in all dims
- Spatial mesh: unstructured cartesian mesh for each phi (toroidal angle)
- MPI decomposition in phi, vpar, mu, alpha. OpenMP in RZ, no MPI.
- Typical sizes: (RZ: 1e5-1e7, phi: 32, vpar: 10-100, mu: 10-100, alpha: 2)







Source: P. Ulbl et al. (2024) “Towards first-principles simulations of the L- to H-mode transition with the global gyrokinetic turbulence code GENE-X”, TSVV4/1 Workshop 2024.

Current File I/O

- Spatial mesh
  - All phi planes created/read on all MPI procs
  - Writes only by MPI master to mesh.nc
- 5D state
  - Each MPI proc holds its own chunk of the state vector
  - Each MPI proc i writes its own chunk at time j to part_j/checkpoint_i.nc
- es_pot/A_par (RZ, phi)
  - Reduction to root on all phi sub-communicators. Write by MPI master
- 0/2D moments of 5D state
  - 0D (6 X alpha): collected across one species sub-communicator
  - 2D (RZ, phi, 8, alpha): collected across one phi-species sub-communicator
  - Reduction to MPI master on one sub-communicator and write

Limitations of Current I/O

- Spatial mesh
  - Not all meshes required on all procs
  - Assigned mesh chunks can be generated/read in parallel on each MPI proc
- 5D state
  - Writing generates many directories and files.
  - Restarting the simulation requires the same MPI configuration used during write.
- es_pot/A_par
  - MPI reduction not necessary, parallel I/O can be used on one sub-communicator
- 0/2D moments of 5D state
  - MPI reduction is not necessary, parallel I/O can be used

Previous attempts with parallel netCDF failed due to unknown runtime error

To get around the current limitations, we implement parallel I/O with ADIOS library

ADIOS

Adaptable Input/Output System from Exascale Computing Project (DoE, U.S.)

- Parallel file I/O for distributed arrays of arbitrary dimensions
  - Binary packed format (BP4, BP5)
  - HDF5 support
- Fortran/C/C++ bindings
- “BeginStep/EndStep” abstraction for time series data
- “Deferred Put/Get” abstraction: lazy evaluation to minimize low level ops
- Backends:
  - File: POSIX I/O, stdio, fstream
  - Network: MPI, RDMA

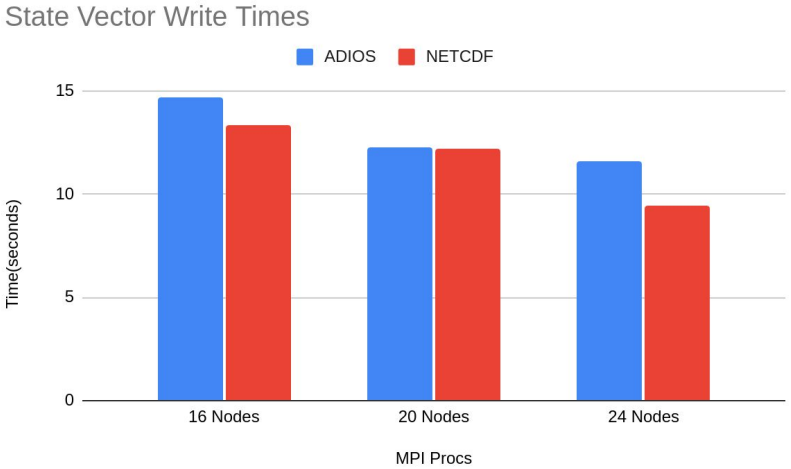
I/O Implementation With ADIOS

5D State Checkpoint

- Data+ghosts stored in contiguous 5D arrays local to each MPI process
- ADIOS write implementation:
  - Determine global/local sizes and offsets based on MPI config
  - Pack data into a continuous buffer for writing
  - Write a new output file or append a new step to an existing file, no parts directories
- ADIOS read implementation:
  - Determine global/local sizes and offsets based on MPI config
  - Read the latest data step into a continuous buffer and unpack the data
- Saved checkpoint is MPI config agnostic - read with different MPI config possible
- Possibility to play with numAggregators, numFiles at runtime
  - Default values: numAggregators = MPI nodes, numFiles = numAggregators

Read/Write 5D State: netCDF vs ADIOS

- Test on EPFL Jed, infiniband partition. Read and writes on /scratch
- (RZ, phi, vp, mu, spec) = (270K, 36, 80, 20, 2). 3 MPI configs
  - 16 Nodes: (1, 4, 1, 2, 2)
  - 20 Nodes: (1, 2, 5, 1, 2)
  - 24 Nodes: (1, 3, 2, 2, 2)
- Mean values over all nodes, for 20 runs

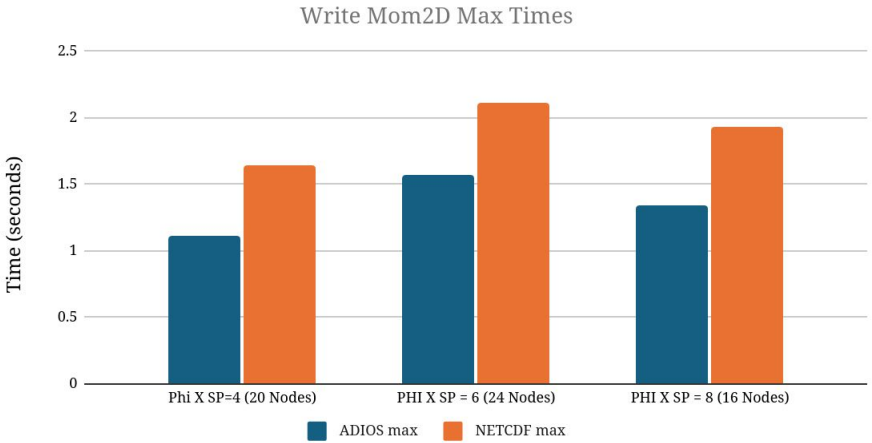


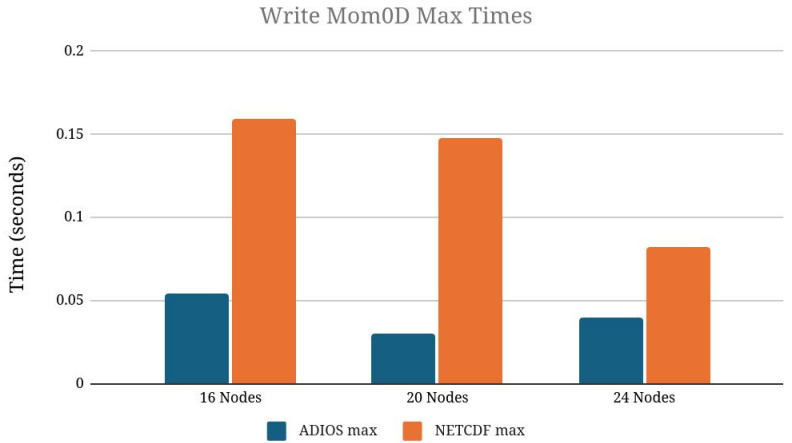
0D and 2D Moments

- Parallel write for moments using ADIOS, no reduction step:
  - Mom 0D: (6 X alpha), write using one species sub-communicator containing master
  - Mom 2D: (RZ, phi, 8, alpha), write using one phi-species sub-communicator
- No read implementation in netCDF or ADIOS

Write 0D/2D Moments: netCDF vs ADIOS

- Test on EPFL Jed, infiniband partition. Writes on /scratch
- Test params: (RZ, phi, vp, mu, spec) = (270K, 36, 80, 20, 2). 3 MPI configs
  - 16 Nodes: (1, 4, 1, 2, 2)
  - 20 Nodes: (1, 2, 5, 1, 2)
  - 24 Nodes: (1, 3, 2, 2, 2)
- Mean values over all nodes, for 20 runs
- Max times over all nodes as not all nodes participate in the exchange



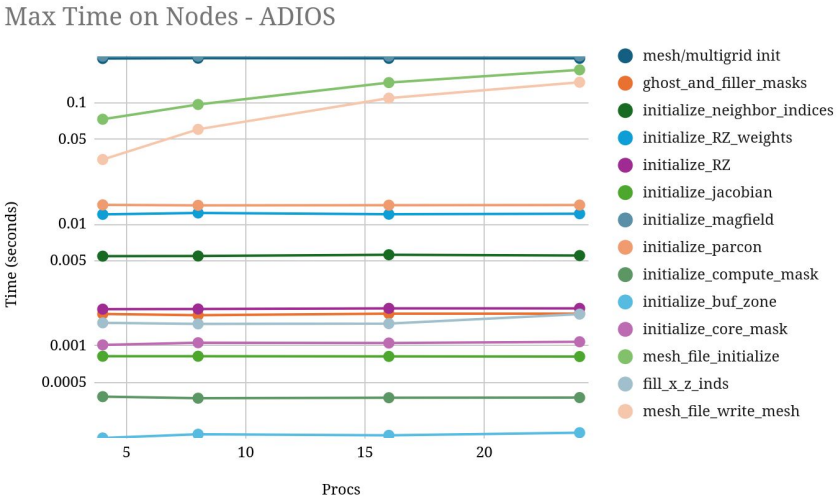


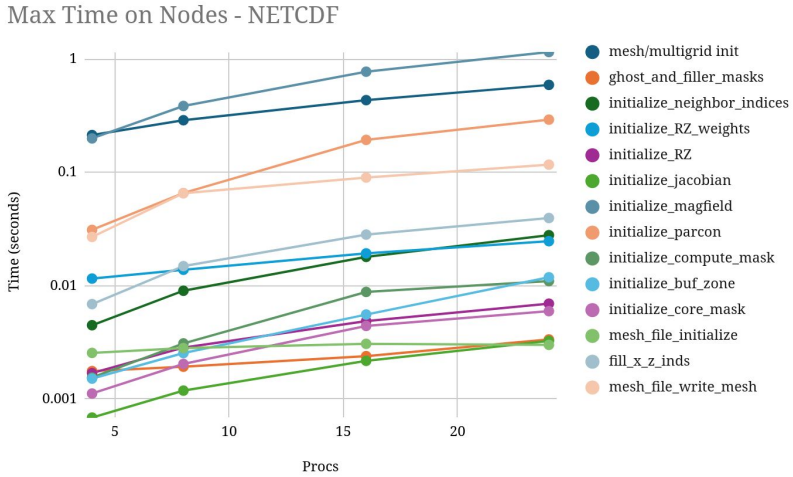
Mesh (Work in Progress)

- Current implementation:
  - Mesh built for all phi on all procs using PARALLAX (builds multigrids)
  - Only MPI master writes the mesh, multigrid info
- ADIOS based implementation:
  - Creating meshes for local phi doesn’t work, mesh data up to phi +2,-2 accessed.
  - Creating meshes for local + ghost phi range on each process.
  - Writing mesh only for local phi range (excluding ghosts) from a process.
  - (RZ, phi) shaped arrays written neatly by ADIOS, e.g. R/Z buffer, magnetic field
- Next steps: writing unstructured, hierarchical data related to multigrid

Mesh Init Time: ADIOS vs netCDF

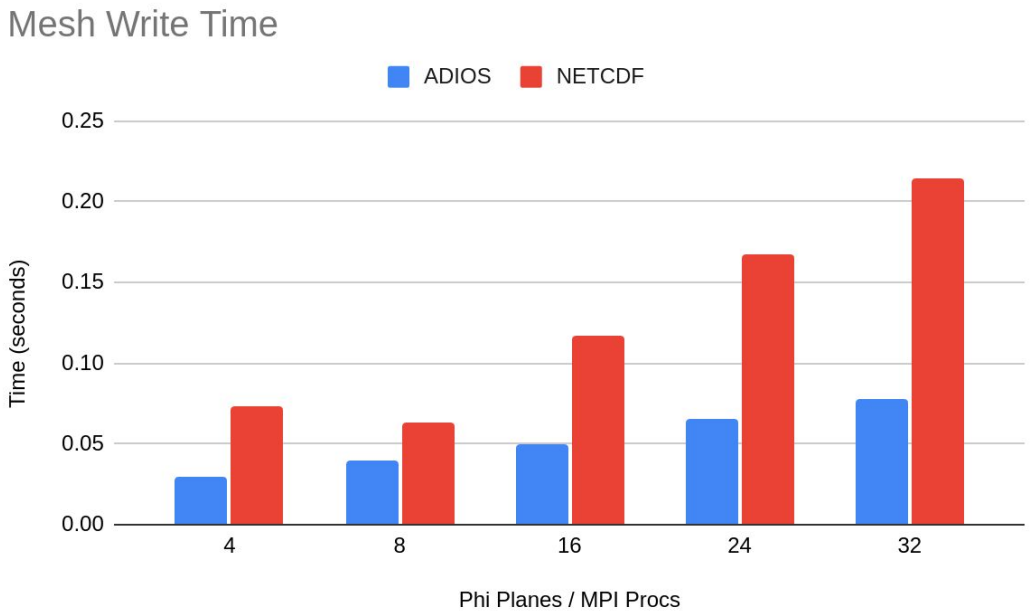
- Tests conducted on Jed cluster, infiniband partition.
- (RZ, phi, vp, mu, spec) = (**270K**, **X**, 40, 16, 2). MPI config: (1, **X**, 1, 1, 1)
- **X**: 4, 8, 16, 24, 32
- ADIOS: average time across all (phi) nodes, 50 runs
- NETCDF: average time on master, 50 runs





Mesh Write: netCDF vs ADIOS

- Tests conducted on Jed cluster, standard partition. Write to /home
- (RZ, phi, vp, mu, spec) = (**270K**, **X**, 40, 16, 2). MPI config: (1, **X**, 1, 1, 1)
- **X**: 4, 8, 16, 24, 32
- ADIOS: average time across all (phi) nodes, 50 runs
- NETCDF: average time on master, 50 runs



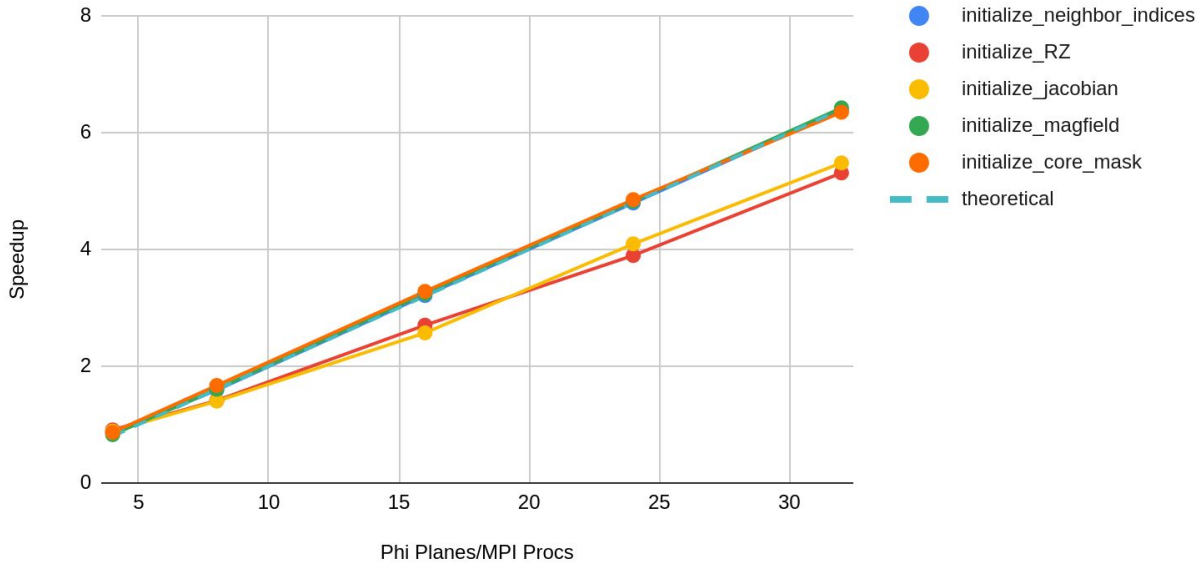
Summary

- Implemented ADIOS based I/O for spatial mesh, 5D state, 0/2D moments
- Speedup in mesh generation, mesh writing
- Slight slowdown in 5D checkpoint writing but speed up in reading
- Speedup in writing 0/2D moments
- Possible to restart simulation with a different MPI configuration
- Next steps: implementing mesh read, multigrid I/O, es_pot/A_par write

Thank you!

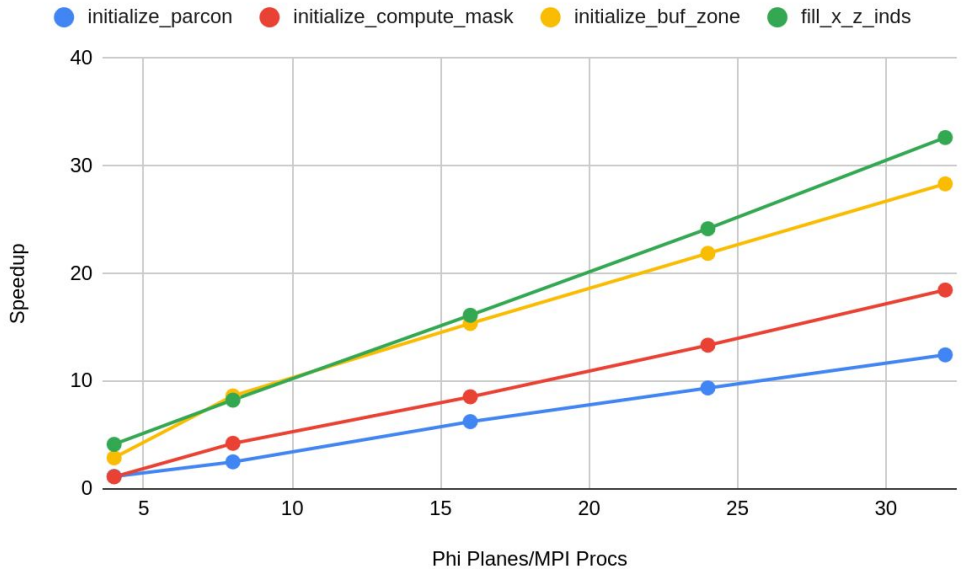
Mesh Init Time: ADIOS Speedup vs netCDF

- Tests conducted on Jed cluster, standard partition.
- (RZ, phi, vp, mu, spec) = (**270K**, **X**, 40, 16, 2). MPI config: (1, **X**, 1, 1, 1)
- **X**: 4, 8, 16, 24, 32
- ADIOS: average time across all (phi) nodes, 50 runs
- NETCDF: average time on master, 50 runs



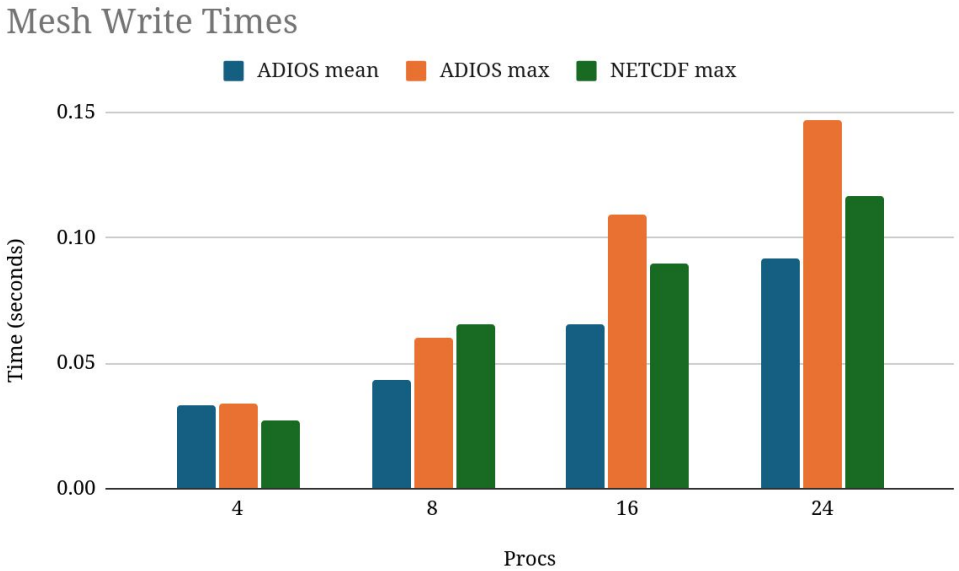
Mesh Init Time: ADIOS Speedup vs netCDF

- Tests conducted on Jed cluster, standard partition.
- (RZ, phi, vp, mu, spec) = (**270K**, **X**, 40, 16, 2). MPI config: (1, **X**, 1, 1, 1)
- **X**: 4, 8, 16, 24, 32
- ADIOS: average time across all (phi) nodes, 50 runs
- NETCDF: average time on master, 50 runs



Mesh Write: netCDF vs ADIOS

- Tests conducted on Jed cluster, standard partition. Write to /home
- (RZ, phi, vp, mu, spec) = (**270K**, **X**, 40, 16, 2). MPI config: (1, **X**, 1, 1, 1)
- **X**: 4, 8, 16, 24, 32
- ADIOS: average time across all (phi) nodes, 50 runs
- NETCDF: average time on master, 50 runs



Mesh Init Time: ADIOS vs netCDF

adios					
	4	8	16	24	32
init_first_part	0.854958572	0.85705134	0.8563431516	0.8509147401	0.8497804452
initialize_RZ_domain	0.0000158475	0.00000885787	0.0000088520425	0.000008648669167	0.00000864859125
ghost_and_filler_masks	0.001920958	0.001863052923	0.001882407162	0.001828187382	0.00185232923
initialize_neighbor_indices	0.019397679	0.01817873297	0.01784290545	0.01788911952	0.01790065111
initialize_RZ_weights	0.06060214	0.06036289596	0.0602911834	0.06035665752	0.06016798763
initialize_RZ	0.002864023	0.0027012437	0.002593224014	0.002609399882	0.002582642998
initialize_vp_bsg	0.00316822225	0.00319293121	0.003201397196	0.00322651878	0.003221028761
initialize_jacobian	0.00095901075	0.000996864455	0.001038795685	0.0009725199533	0.0009614840256
initialize_magfield	0.4599644618	0.4595015625	0.4623392187	0.4646287057	0.4651894842
initialize_parcon	0.05584953	0.05184922977	0.05154521931	0.05168482959	0.05172708945
initialize_compute_mask	0.00206808825	0.0009552213425	0.0009796418087	0.0009841211433	0.0009773515287
initialize_buf_zone	0.00087600475	0.00046512104	0.000473519455	0.0004710492208	0.0004756471263
initialize_core_mask	0.00380840575	0.00354589999	0.003560439711	0.003569056783	0.003576467656
set_bsg_flags	0.000000636	0.0000006328125	0.0000006045975	0.0000006069841667	0.0000006567675
mesh_file_initialize	0.1151438505	0.1110297351	0.1477438678	0.1920398449	0.2559843296
fill_x_z_inds	0.0059884025	0.005810388898	0.005826820852	0.005849304162	0.005877692043
write_parallax	0.1612477895	0.1995761752	0.3329392955	0.4209620545	0.5678550485
mesh_file_write_mesh	0.0299476095	0.03957648549	0.04971448752	0.06587861621	0.07772146888

Mesh Init Time: ADIOS vs netCDF

netcdf					
	4	8	16	24	32
init_first_part	0.8322822097	0.9122345613	1.055840945	1.195848919	1.336582226
initialize_RZ_domain	0.00001540015	0.00000907994	0.00000931147375	0.000009376845833	0.00000931687375
ghost_and_filler_masks	0.002110708945	0.0021019893	0.002553898026	0.003006189112	0.003520346348
initialize_neighbor_indices	0.01674484507	0.02892006132	0.05731287841	0.08577917885	0.1143764987
initialize_RZ_weights	0.06017954972	0.06270466019	0.06914717026	0.07638419853	0.08806714
initialize_RZ	0.002604098615	0.00381579288	0.007007808245	0.01017058755	0.01371829844
initialize_vp_bsg	0.00322387016	0.003159375345	0.003211614508	0.00321811802	0.003207477323
initialize_jacobian	0.000852546775	0.001396652035	0.002670922839	0.003982811676	0.005272102918
initialize_magfield	0.3803545202	0.7387782497	1.499956439	2.241742322	2.989207912
initialize_parcon	0.06474757303	0.1301970669	0.3222760405	0.4847445977	0.6447891202
initialize_compute_mask	0.002330084765	0.004043642805	0.008381771997	0.01314121542	0.01807350873
initialize_buf_zone	0.00253728759	0.004020999653	0.007279010421	0.01030884915	0.0134821693
initialize_core_mask	0.003305013865	0.005925970655	0.01168614078	0.01732904997	0.02271425116
set_bsg_flags	0.000000683905	0.000000768595	0.00000072729125	0.0000007276341667	0.00000080250625
mesh_file_initialize	0.001217482015	0.000328629465	0.00023541758	0.0001231867367	0.0001159309106
fill_x_z_inds	0.02480439837	0.04797807351	0.0940138437	0.1414209391	0.191869545
write_parallax	0.02277475939	0.01223597045	0.00602845312	0.004210326838	0.003095064771
mesh_file_write_mesh	0.01835569328	0.007844673083	0.007353392543	0.006972715696	0.006729743577