Author: Mahzad Khoshlessan
email: mkhoshle@asu.edu
institution: Arizona State University
Author: Ioannis Paraskevakos
email: i.paraskev@rutgers.edu
institution: RADICAL, ECE, Rutgers University, Piscataway, NJ 08854, USA
Author: Shantenu Jha
email: shantenu.jha@rutgers.edu
institution: RADICAL, ECE, Rutgers University, Piscataway, NJ 08854, USA
Author: Oliver Beckstein
email: obeckste@asu.edu
institution: Arizona State University
corresponding:

bibliography: `mdanalysis`

Parallel Analysis in MDAnalysis using the Dask Parallel Computing Library

The analysis of biomolecular computer simulations has become a challenge because the amount of output data is now routinely in the terabyte range. We evaluate if this challenge can be met by a parallel map-reduce approach with the Dask parallel computing library for task-graph based computing coupled with our MDAnalysis Python library for the analysis of molecular dynamics (MD) simulations *Gowers:2016aa, Michaud-Agrawal:2011fu*. We performed a representative performance evaluation, taking into account the highly heterogeneous computing environment that researchers typically work in together with the diversity of existing file formats for MD trajectory data. We found that the the underlying storage system (solid state drives, parallel file systems, or simple spinning platter disks) can be a deciding performance factor that leads to data ingestion becoming the primary bottle neck in the analysis work flow. However, the choice of the data file format can mitigate the effect of the storage system; in particular, the commonly used Gromacs XTC trajectory format, which is highly compressed, can exhibit strong scaling close to ideal due to trading a decrease in global storage access load against an increase in local per-core cpu-intensive decompression. Scaling was tested on single node and multiple nodes on national and local supercomputing resources as well as typical workstations. In summary, we show that, due to the focus on high interoperability in the scientific Python eco system, it is straightforward to implement map-reduce with Dask in MDAnalysis and provide an in-depth analysis of the considerations to obtain good parallel performance on HPC resources.

Keywords

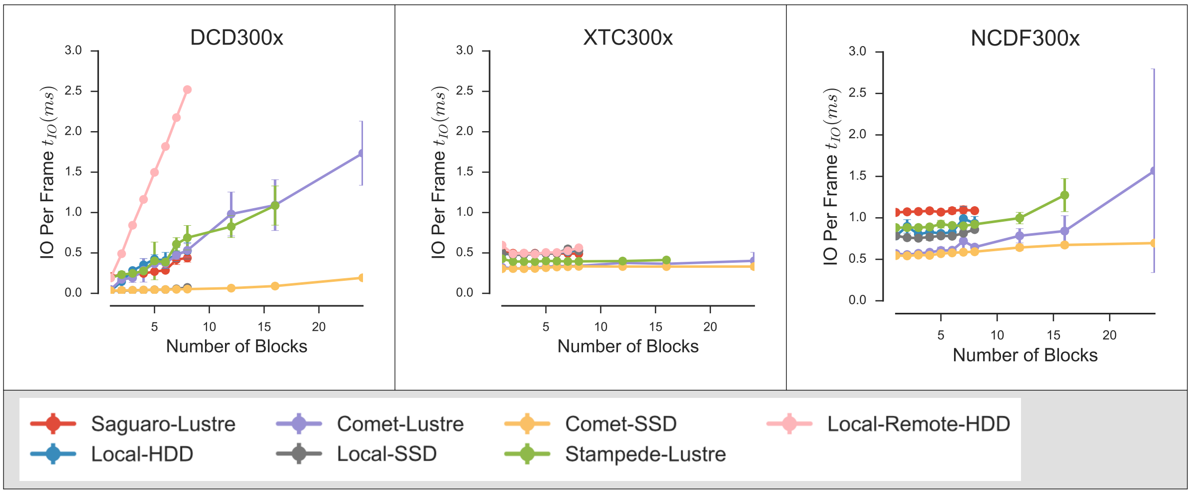
MDAnalysis, High Performance Computing, Dask, Map-Reduce, MPI

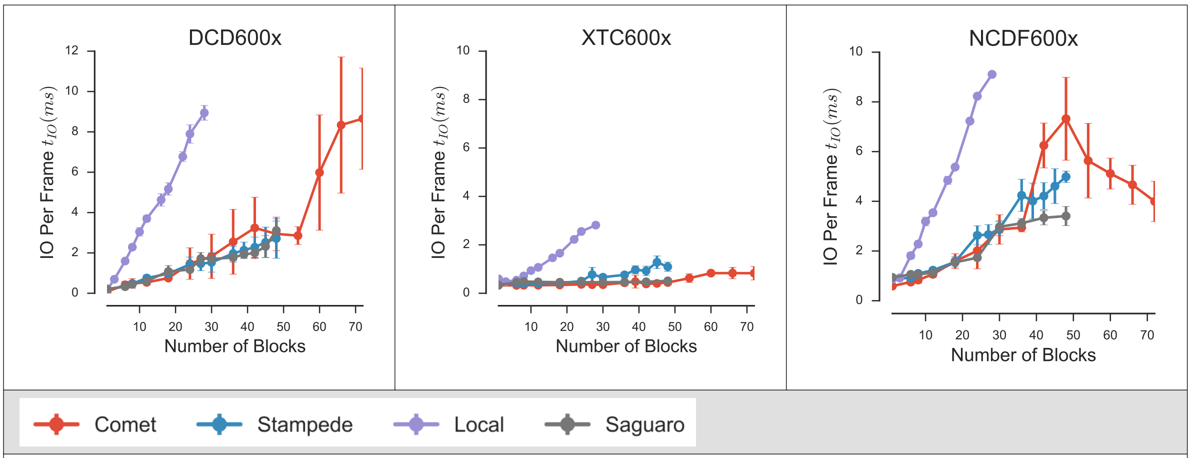
Introduction

MDAnalysis is a Python library that provides users with access to raw simulation data that allows structural and temporal analysis of molecular dynamics (MD) trajectories generated by all major MD simulation packages *Gowers:2016aa, Michaud-Agrawal:2011fu*. The size of these trajectories is growing as the simulation times is being extended from micro-seconds to milli-seconds and larger systems with increasing numbers of atoms are simulated. Thus, the amount of data to be analyzed is growing rapidly (into the terabyte range) and analysis is increasingly becoming a bottleneck *Cheatham:2015qf*. Therefore, there is a need for high performance computing (HPC) approaches to increase the throughput. MDAnalysis does not yet provide a standard interface for parallel analysis; instead, various existing parallel libraries are currently used to parallelize MDAnalysis-based code. Here we evaluate performance for parallel map-reduce type analysis with the Dask parallel computing library for task-graph based distributed computing on HPC and local computing resources. As the computational task we perform an optimal structural superposition of the atoms of a protein to a reference structure by minimizing the RMSD of the atoms. A range of commonly used MD file formats (CHARMM/NAMD DCD, Gromacs XTC, Amber NetCDF) and different trajectory sizes are benchmarked on different HPC resources including national supercomputers (XSEDE TACC Stampede and SDSC Comet), university supercomputers (ASU Research computing center (Saguaro)), and local resources (Gigabit networked multi-core workstations). The tested resources are parallel and heterogeneous with different CPUs, file systems, high speed networks and are suitable for high-performance distributed computing at various levels of parallelization. Such a heterogeneous environment creates a challenging problem for developing high performance programs without the effort required to use low-level, architecture specific parallel programming models for our domain-specific problem. Different storage systems such as solid state drives (SSDs), hard disk drives (HDDs), and the parallel Lustre file system (implemented on top of HDD) are also tested to examine effect of I/O on the performance. The benchmarks are performed both on a single node and across multiple nodes using the multiprocessing and distributed schedulers in Dask library. A protein system of $N = 3341$ atoms per frame but with different number of frames per trajectory was analyzed. We used different trajectory sizes of 50 GB, 150 GB, and 300 GB for Dask multiprocessing and 100 GB, 300 GB, 600 GB for Dask distributed. All results for Dask distributed are obtained across three nodes on different clusters. Results are compared across all file formats, trajectory sizes, and machines. Our results show strong dependency on the storage system because a key problem is competition for access to the same file from multiple processes. However, the exact data access pattern depends on the trajectory file format and a strong dependence on the actual data format arises. Some trajectory formats are more robust against storage system specifics than others. In particular, analysis with the Gromacs XTC format can show strong ideal scaling over multiple nodes because this highly compressed format effectively reduces (global) I/O at the expense of increasing (local) per-core work for decompression. Our results show that there can be other challenges aside from the I/O bottleneck for achieving good speed-up. For instance, with numbers of processes matched to the available cores, contention on the network may slow down individual tasks and lead to poor load balancing and poor overall performance. In order to identify the performance bottlenecks for our Map-Reduce Job, we have tested and examined several other factors including striping, oversubscribing, and the Dask Scheduler. We also compared a subset of systems with an MPI-based implementation (using mpi4py) in order to better understand the effect of using a high-level approach such as the Dask parallel library compared to a lower level one; in particular, we tried to identify possible underlying factors that may lead to low performance. In summary, Dask together with MDAnalysis makes it straightforward to implement parallel analysis of MD trajectories within a map-reduce scheme. We show that obtaining good parallel performance depends on multiple factors such as storage system and trajectory file format and provide guidelines for how to optimize trajectory analysis throughput within the constraints of a heterogeneous research computing environment.

Effect of I/O Environment

In MDAnalysis library, trajectories from MD simulations are a frame by frame description of the motion of particles as a function of time. To allow the analysis of large trajectories, MDAnalysis only loads a single frame into memory at any time *Gowers:2016aa, Michaud-Agrawal:2011fu*. Some file systems are designed to run on a single CPU while others like Network File System (NFS) which is among distributed file systems are designed to let different processes on multiple computers access a common set of files. These file systems guarantees sequential consistency which means that it prevents any process from reading a file while another process is reading the file. Distributed parallel file systems (Lustre) allow simultaneous access to the file by different processes; however it is very important to have a parallel I/O library; otherwise the file system will process the I/O requests it gets serially, yielding no real benefit from doing parallel I/O. Figure [] show the I/O pattern compared between different file formats. XTC file format takes advantage of in-built compression and as a result has smaller file size as compared to the other formats. In addition, MDAnalysis implements a fast frame scanning algorithm for XTC files. This algorithm computes frame offsets and saves the offsets to disk as a hidden file once the trajectory is read the first time. When a trajectory is loaded again then instead of reading the whole trajectory the offset is used to seek individual frames. As a result, the time it takes a process to load a frame into memory is short. In addition, each frame I/O will be followed by decompressing of that frame as soon as it is loaded into memory. Thus, as soon as the frame is loaded into memory by one process, the file system will let the next process to load another frame into memory. This happens while the first process is decompressing the loaded frame. As a result, the overlapping of the data requests for the same calculation will be less frequent. However, there is no in-built compression for DCD and netCDF file formats and as a result file sizes are larger. This will result in higher I/O time and therefore overlapping of per frame trajectory data access. The I/O time is larger for netCDF file format as compared to DCD file format due to larger file size. This is since netCDF has a more complicated file format. Reading an existing netCDF dataset involves opening the dataset; inquiring about dimensions, variables, and attributes; reading variable data; and closing the dataset [ref]. In fact, netCDF has a very sophisticated format, while DCD has a very simple file format. This is why DCD is showing a weak scaling by increasing parallelism whereas netCDF file format is being scaled reasonably well by increasing parallelism across many cores. Our study showed that SSD can be very helpful (especially for dcd file format) and can improve the performance due to speed up in access time. Also we anticipate that, heavy analyses that take lenger time, per frame trajectory data access happens less often and accession times gradually become staggered across CPUs which can be considered for future studies.

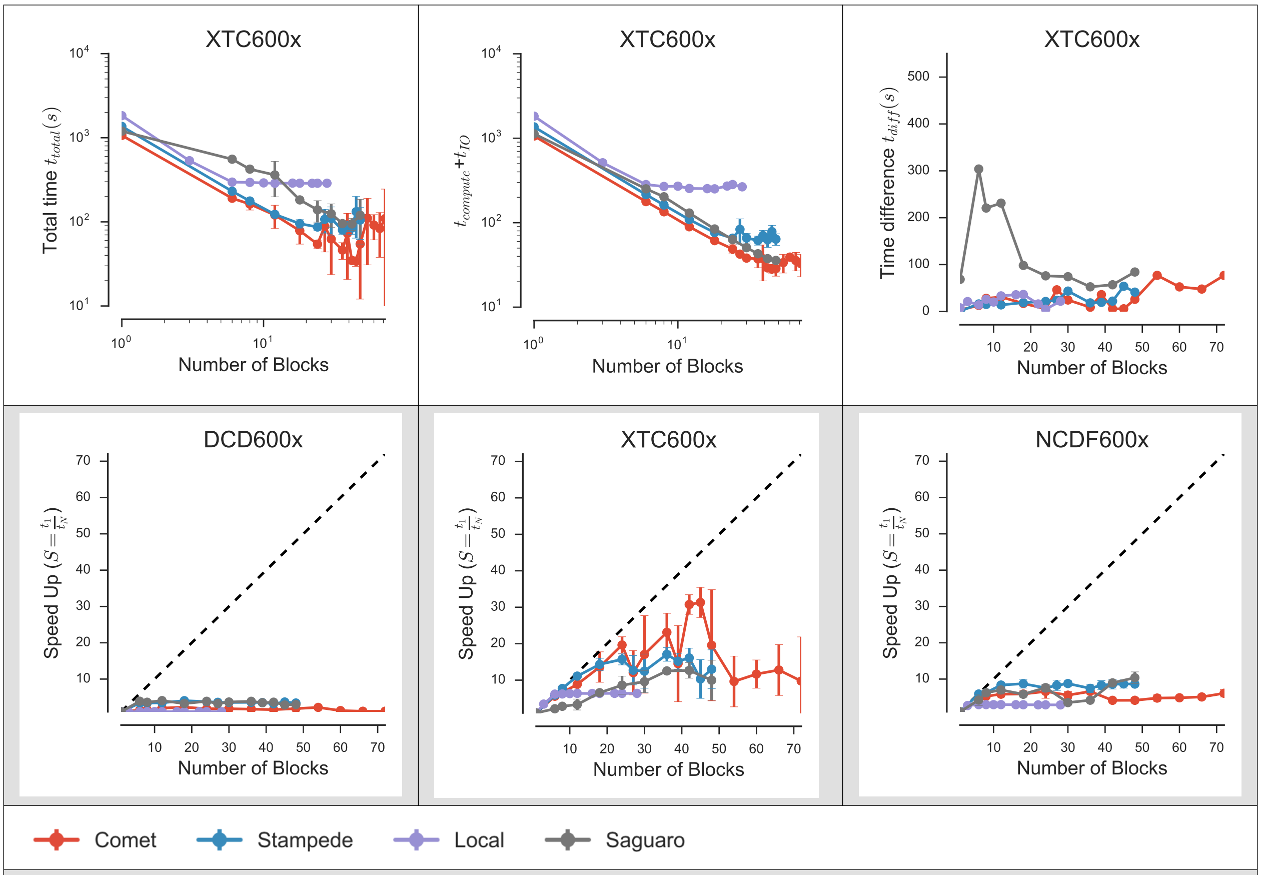




Comparison of IO time between top) 300x (multiprocessing) and bottom) 600X (distributed) trajectory for all file formats

Effect of File Format

Figures [] to [] summarizes speedups and parallel efficiencies for 300X and 600X trajectories and all file formats for multiprocessing and distributed scheduler respectively. According to Figures [] DCD file format does not scale at all by increasing parallelism across different cores. This is due to the overlapping of the data access requests from different processes. XTC file format express reasonably well scaling with the increase in parallelism up to the limit of 24 (single node) for all trajectory sizes for all machines (multiprocessing scheduler) and Comet and Stampede (for distributed scheduler). The NCDF file format scales very well up to 8 cores for all trajectory sizes. For XTC file format, the I/O time is leveled up to 50 cores and compute time also remains level across parallelism up to 72 cores. Therefore, it was expected to achieve speed up, across parallelism up to 50 cores However, this amount is reduced to 20 cores as can also be observed in speed up plots. Based on the present result, there is a difference between job execution time, and total compute and I/O time averaged over all processes. This difference increases with increase in trajectory size for all file formats for all machines (Not shown here). This time difference is much smaller for Comet and Stampede as compared to other machines. In order to find the underlying reasons for this difference, web interface of Dask is used to obtain information about the amount of time spent on the communication between workers, and different computations at the worker level in the Map-reduce job. Because Dask parallel computing library is too high level, it is really hard to obtain detail information about each task at different levels. The difference between job execution time and total compute and I/O time measured inside our code is very small for the results obtained using multiprocessing scheduler; however, it is considerable for the results obtained using distributed scheduler. In order to obtain more insight on the underlying network behavior both at the worker level and communication level and in order be able to see where this difference originates from we have used the web interface of the Dask library. This web interface is launched whenever Dask scheduler is launched. Table ref{tab:time-comparison} summarizes the average and max total compute and I/O time measured through our code, max total compute and I/O time measured using the web interface and job execution time for each of the cases tested. As seen from the tests performed on ASU Saguaro, there is a very small difference between maximum total compute and I/O time and job execution time. This difference is mostly due to communications performed in the reduction process. In addition, maximum total compute and I/O time measured using the web interface and our code are very close. As can be seen from the results, due to different reasons, some tasks (so-called Stragglers) are considerably slower than the others, delaying the completion of the job.



Comparison of IO time between top) 300x (multiprocessing) and bottom) 600X (distributed) trajectory for all file formats

Summary of the measured times for different calculations, tested on different machines for 600X trajectory and XTC file format. N_{cores} is the number of cores used in each test, average total compute and I/O time is the I/O plus compute time for all frames per process averaged across all processes, max total compute and I/O time is the I/O plus compute time for all frames for the slowest process measured through the code, max total compute and I/O time measured using web interface is the I/O plus compute time for all frames for the slowest process measured through web interface. :label:`timecomparison`

Resource	N_{cores}	Average total compute and I/O time(s)	Max total compute and I/O time(s)	Max total compute and I/O time measured using web interface(s)	Job executio time(s)
Local	24	93.83	110.58	110.43	111.83
Local	28	86.54	111.54	111.24	112.81
SDSC Comet	30	37.79	41.11	41.12	42.23
SDSC Comet	54	36.15	43.58	104.25	105.1

Challenges for Good HPC Performance

There is a caveat needs to be added here that all results were obtained during normal, multi-user, production periods on all machines. In fact, the time the jobs take to run are affected by the other jobs on the system. This is true even when the job is the only one using a particular node, which was the case in the present study. There are shared resources such as network filesystems that all the nodes use. The high speed interconnect that enables parallel jobs to run is also a shared resource. The more jobs are running on the cluster, the more contention there is for these resources. As a result, the same job runs at different times will take a different amount of time to complete. In addition, remarkable fluctuations in I/O time across different processes is observed through monitoring network behavior using Dask web interface which kind of confirms this issue. These fluctuations differ in each repeat and are dependent on the hardware and network. Another caveat needs to be added here is that jobs may also be scheduled to run on different nodes at different times. For example, our local machine in Beckstein's lab has also a heterogenous environment. This problem together with the others mentioned above further complicates any attempts at benchmarking. Therefore, this makes it really hard to optimize codes, since it is hard to determine whether any changes in the code are having a positive effect. This is because the margin of error introduced by the non-deterministic aspects of the cluster's environment is greater than the performance improvements the changes might produce. There is also variability in network latency, in addition to the variability in underlying hardware in each machine. This causes the results to vary significantly across different machines. Because our Map-reduce job is pleasantly parallel, all of our processes have the same amount of work to do. Therefore, observing these stragglers is unexpected and the following sections in the present study aim to identify the reason for which we are seeing these stragglers.

Performance Optimization

In the present section, we have tested different features of our computing environment to see if we can identify the reason for those stragglers and improve performance by avoiding the stragglers. Lustre Striping, oversubscribingi, scheduler throughput are tested to examine their effect on the performance. In addition, scheduler plugin is used to validate our observation using web interface. In fact, we create a plugin that performs logging whenever a task changes state. Through the scheduler plugin we will be able to get lots of information about a task whenever it finishes computing.

Effect of Lustre Striping

As discussed before, the overlapping of data requests from different processes can lead to higher I/O time and as a result poor performance. This is especially strongly affecting our results since our compute per frame is not heavy and as a result the overlapping of data requests is more frequent. The effect on the performance is strongly dependent on file format and some formats like XTC file formats which take advantage of in-built decompression are less affected by the contention from many data requests from many processes. However, when extending to more than one node, even XTC files were affected by this as is also shown in the previous section. In Lustre, a copy of the shared file can be in different physical storage devices (OSTs). Single shared files can have a stripe count equal to the number of nodes or processes which access the file. In the present study we set the stripe count equal to three which is equal to number of nodes. This may be helpful to improve performance, since all the processes from each node will have a copy of the file and as a result the contention due to many data requests will decrease. Figures [] and [] show the speed up and I/O time plots obtained for XTC file format (600X) when striping is activated. As can be seen, IO time is level across parallelism up to 72 cores which means that striping is helpful for decresing IO time. However, we are still seeing these stragglers and the overal speed-up is not improved.

Effect of Oversubscribing

One useful way to robust our code to uncertainty in computations is to submit many more tasks than the numer of cores. This may allow Dask to load balance appropriately, and as a result avoiding the stragglers. In order for this we set the number of tasks to be three times the number of workers. Striping is also activated and is set to three which is also equal to number of nodes. Figures [] and [] show the speed up and I/O time plots obtained for XTC file format (600X). As can be seen, we are still seeing these stragglers and the overal speed-up is not improved. In order to see if the calculation is load balanced and the same amount of load is assigned to each worker by the scheduler, scheduler pluging is used to get detailed information about a task and to also validate our observationis obtained from web-interface. The results from scheduler pluging is described in the following section.

Scheduler Plugin Results

Examining Scheduler Throughput

Comparison of Performance of Map-Reduce Job Between MPI for Python and Dask Frameworks

Based on the results presented in previous sections, it turned out that the stragglers are not because of the network, shared resources or scheduler throughput. Lustre striping improves I/O time; however, the job computation is still delayed and as a result lead to poor speed-up when extended to multiple nodes. In order to make sure if the stragglers are created because of scheduler overhead in Dask framework we have tried to measure the performance of our Map-Reduce job using MPI-based implementation. This will let us figure out whether the stragglers observed in the present benchmark using Dask parallel libray are as a result of scheduler overhead or the environment itself.

Acknowledgments

MK and IP were supported by grant ACI-1443054 from the National Science Foundation. SJ and OB were supported in part by grant ACI-1443054 from the National Science Foundation. Computational resources were in part provided by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575 (allocation MCB130177 to OB and allocation TG-MCB090174 to SJ).

References