Studying the Dynamics and Evolution of Large Elliptical Galaxies at High-redshift (z~0.7) using the LEGA-C spectra and UltraVista photometry

Rachel Bezanson, Department of Physics and Astronomy

Resource Type	SUs Requested	Maximum SUs allowed
MPI	3,200,000	3,200,000
SMP standard (default)	500,000	1,300,000
HTC	0	500,000
GPU	0	100,000

Project Description

‘Testing the Formation of the Largest Elliptical Galaxies’

This project focuses on studying the formation and evolution of these galaxies either by slow channel (cannibalizing neighboring galaxies) or by fast channel (merging of two large spiral galaxies) using the LEGA-C dataset. We want to study the interplay between the history of star formation and the motions of stars in distant galaxies. Although we have made progress on measuring the dynamics of the stars (via the doppler effect), we must make use of more sophisticated modeling techniques to properly explore the range of stellar populations. Using BAGPIPES python package we simulate the galaxy light as a combination of star light and dust and fit them to both photometric data and high-signal-to-noise spectroscopic data using MultiNest nested sampling algorithm and generate star-formation history for each galaxy as well as joint-probability distribution of stellar population parameters like age, mass, metallicity, dust attenuation and nebular emission. Each fit relies upon a few key assumptions about the parametrization (the type of dust, analytic form of star formation history and prior probability distribution); we hope to characterize the robustness of our results to these assumptions by conducting two families of fits with different star-formation histories.

Dr. Rachel Bezanson is the survey scientist of the Large Early Galaxy Astrophysics Census (LEGA-C) survey. It is a 130-night public spectroscopic survey conducted with VIMOS on the Very Large Telescope and has taken ~3200 spectra of galaxies with typical continuum SNR of 20 per angstrom in the redshift range 0.6< z<1.0, each observed for ~20 hours and fully reduced with a custom-built pipeline. These deep and high resolution spectra hold detailed information about the light coming from stars in these very distant galaxies. In fact, the galaxies are so far that light has been traveling for half the age of the Universe to reach our telescopes, giving a valuable glimpse into a much earlier epoch. The primary goal of Bayesian statistical modeling of these spectra is to characterize their average ages, how rapid or extended star formation has been in the past, their dependence on stellar metallicities and chemical enrichment (which is in turn sensitive to the stellar explosions and merger history) and effects of dynamical masses derived from stellar velocity dispersions on these properties. These are some key pieces of information for the study of galaxy evolution, but have so far only been available for representative galaxy samples in the nearby universe, at look-back times of less than 1 billion years. LEGA-C therefore enables for the first time several studies of these questions about galaxy evolution on long cosmological time scales (up to 8 billion years), especially for the oldest and most massive galaxies the pressing questions that could potentially be answered by this dataset are dynamical evolution, mass assembly through merging and the mechanisms that shut down the star formation.

Computational Demands : We will be using a state-of-the-art python package called BAGPIPES (Bayesian Analysis of Galaxies for Physical Inference and Parameter Estimation) for modeling the spectral energy distribution of the galaxies. It heavily relies on the Bayesian inference tool and multimodal nested sampling algorithm 'MultiNest'. As Bagpipes is a heavy duty package that requires significant computation time for each galaxy, from the initial resource allocation on the cluster, we were able to fit 10 objects with different node configurations and made an estimate of the resources that will be required to fit the entire sample of 3000 objects. As the spectroscopic resolution and wavelength ranges are approximately the same for all the objects in the catalog, in computation terms, it is expected to not have drastic variation in computation time for the entire sample, but the variations in the signal to noise ratio of each spectrum and the convergence of the sampling will produce some scatter as shown in figure 1. We found that on an average, a complete spectroscopic fit of 1 object on SMP cluster (single node) with 23 cores takes about ~330 CPU hours (average calculated for all runs on node=1) and 1 object on MPI parallelization using python package mpi4py takes about ~440 CPU hours (average calculated for all runs on nodes=2,3,5,10) which is reasonable for a total likelihood evaluations of about ~1.8 million, 500 posterior points and 400 live sampling points from the posterior. Using the scaling factor of 1.0 to convert the CPU hours to SUs for the **MPI cluster** and 1-sigma variation in MPI process, we found that we need an upper limit of about **450 SUs per object for one analytical star-formation history (double-power law)**. The optimum allocation for 3200 objects would be 1,440,000 SUs on MPI that can fit all objects with one analytic star-formation history. We are also interested in analysing the results from one more star-formation history (exponential) for all of them and only some objects with lognormal star-formation history . As the parameter space is approximately the same for all of them, we expect to require 2 times of the above i.e. 2 x 1,440,000 = 2,880,000 SUs on MPI cluster. Including the margin computation resources required for test runs and reruns with different parameter spaces on selected objects in the catalog, we expect to use the entire **3,200,000 SUs allocation on the MPI cluster**. The photometry only models, that are much faster on SMP single node processors and have less data points to fit, require about ~0.64 CPU hours per object for exponential star formation history and ~0.95 CPU hours per object with double-power law star-formation history on SMP cluster. For the entire catalog, we estimate about 0.64x3200=2048 CPU hours and 0.95x3200=3040 CPU hours and hence we will need about 100,000 SUs on the SMP cluster for 5 different star formation histories and allowing more parameters in the fit. We are also planning to run the lognormal star-formation history spectroscopic fits for selected objects on the SMP cluster which will take about 500,000 CPU hours or 400,000 SUs on the SMP with scale factor of 0.8. That sums up to about **500,000 SUs on the SMP cluster with scale factor of 0.8**. We have also found the multi-nodal processing on MPI to be faster than the single node with good efficiency as shown in the Speedup v/s Nodes plot below.

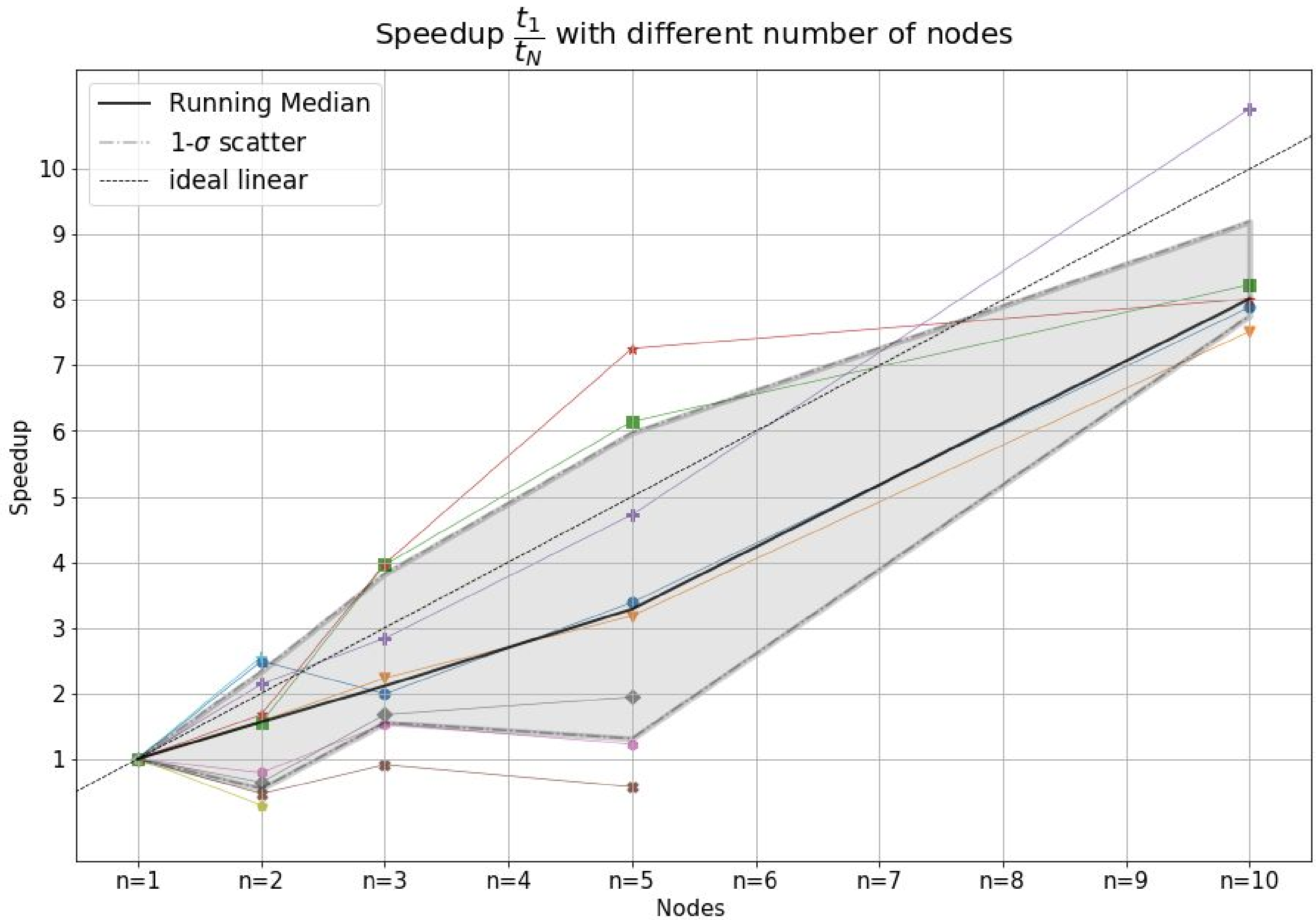

Fig 1 : Every coloured line corresponds to a unique object that was fitted with different number of CPUs
Grey band corresponds to one sigma deviation in speedup at each node for different objects

The plot above is from results of preliminary testing with initial allocation of 30K (10K initial plus 20K later added on request) on the cluster to verify the credibility of running larger samples. The configuration we used for different runs were, SMP cluster: node=1 cores= 23; MPI cluster: node=2,cores=28x2=56; node=3,cores=28x3=84; node=5,cores=28x5=140 and node=10,cores=28x10=280.

As MultiNest is a complete stochastic process with likelihoods calculated for each parameter from a 16-dimensional space and a common prior range allocated to all 3000 objects, it is natural to have some scatter for the speedup with different runs, driven by the variation in the number of likelihood calculations per object to reach convergence, and occasionally encountering some bad fits because of the MCMC live sampling point getting strayed. For such objects we have to re-run the fits with more concise parameter range such that the joint probability distribution does not run into edges. A better intuition of 'good' range of parameter space comes from first run results. From the analyses done for a smaller parameter space before, we found that the bad fits were produced for only 10% of 46-object sample .

For the 400 live sampling points, we can not dedicate more than 400 cores on an object and that puts an upper limit of 14 nodes on the MPI cluster with 28 cores per node. Hence, we expect a similar trend of speedup upto n=14.

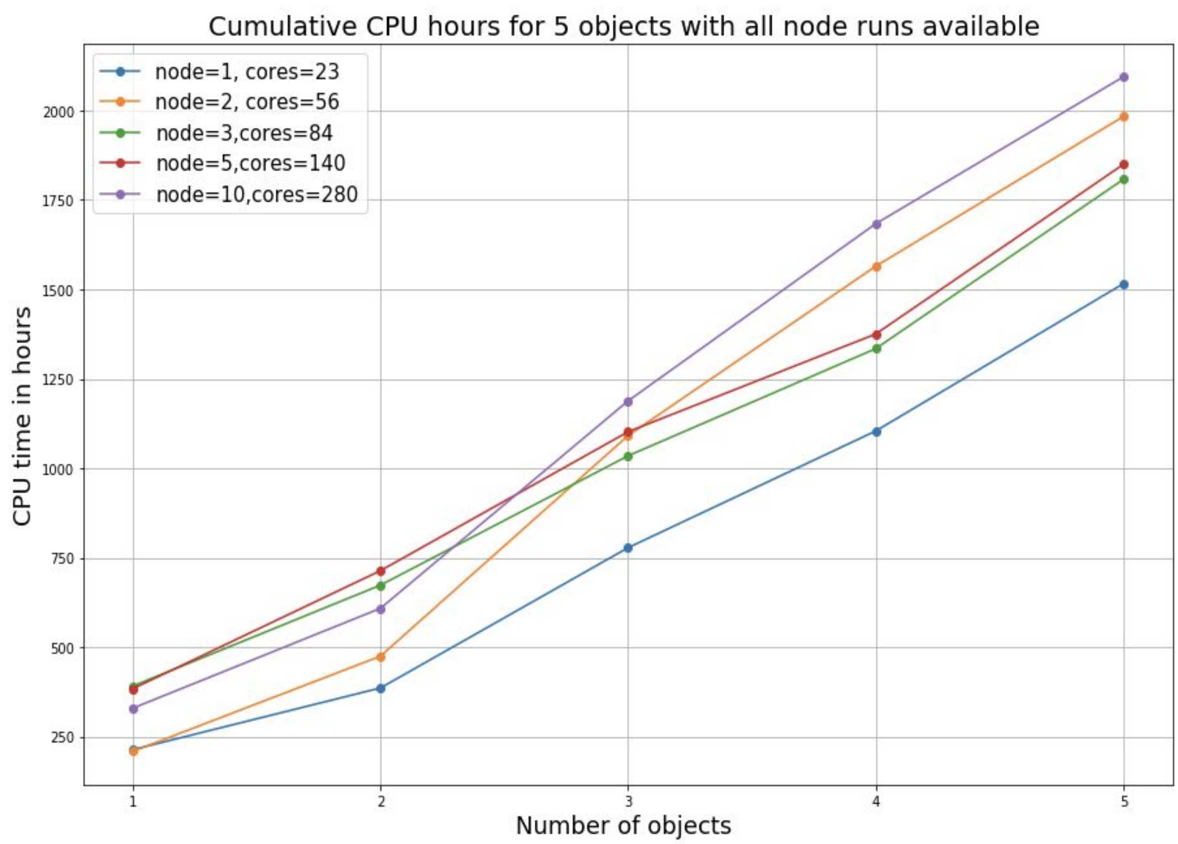

Fig 2 : Each colored line corresponds to a unique node configuration used to run different number of objects

Funding Sources

- Title: ‘Dancing of the Stars: Testing the Formation of the Largest Elliptical Galaxies’
- Duration : September, 2019 - August, 2021
- Amount : $150,000
- PI/co-PI : R. Bezanson
- One to two sentence description : Probing the star formation histories and stellar kinematics in galaxies at half the age of the Universe.
- Agency or other entity : The Kaufman Foundation (New Investigator Award)

Involvement of CRC Consultants

All consultants have been dedicatedly involved in this project by addressing issues in the computational errors via help tickets. We are especially thankful to Kim Wong and Barry Moore for their timely responses on queries, getting codes working on the cluster and giving suggestions for the most optimized way of running a big batch of jobs efficiently.

Publications acknowledging use of CRC (or SaM) resources over the past year

The LEGA-C data release DR3 came out in 2018 and so far many papers have been published on it from different collaborators. I mention below the ones for which Dr. Bezanson is the first author. We are implementing this computation intensive bayesian technique of spectral analysis for the first time and hence have no previous record of using CRC resources.

1. SPATIALLY RESOLVED STELLAR KINEMATICS FROM LEGA-C: INCREASED ROTATIONAL SUPPORT IN Z~0.8 QUIESCENT GALAXIES by Bezanson et. al. (https://arxiv.org/pdf/1804.02402.pdf)
2. 1D Kinematics from stars and ionized gas at z ~ 0.8 from the LEGA-C spectroscopic survey of massive galaxies by Bezanson et. al. (https://arxiv.org/pdf/1811.07900.pdf)