doi:10.1017/S1431927621009909

© Microscopy Society of America 2021

Implementing a storage and compute server to enhance processing of big imaging data.

Jonathan Boyd[1], P. Bradley Goebel[1], Matthias Rust[2] and Christopher Zugates[3]

[1]AstraZeneca, Gaitherburg, Maryland, United States, [2]Arivis AG, Rostock, Mecklenburg-Vorpommern, Germany, [3]Arivis AG, Washington, District of Columbia, United States

The field of microscopy has seen an explosion of new imaging modalities over the past couple years. Some of these transformative techniques include but are not limited to super-resolution microscopy, light sheet microscopy and cryo-electron microscopy. These techniques vary widely in scope and application but all share one common trait, the size of the data requires robust storage, network and computational methodology. Growth in data size creates a series of problems for microscopists. How to store and analyze progressively larger data sets? Using a high-end workstation, a single large data set can often take over a week to complete image analysis. We designed and built an imaging system that has the capacity to store and analyze large data sets. We are working with arivis AG to deliver a turnkey software solution to perform server side image analysis. The software they developed, VisionHub, provides a user-friendly interface to set up advanced workflows for the parallelization of image analysis. The solution has three major components, a large storage component, a 50 TB SSD array and multiple computational servers. Each part of the server work together but perform specific tasks. The large storage component of the server acts as an archival area to hold imaging data that can be easily retrieved for the server based image analysis. The 50 TB SSD array serves as "scratch space" for the large image sets that are being analyzed by the HUB software. Using the SSD array ensures the rapid transfer of data to the compute servers. The computational servers can be leveraged by the HUB to parallelize the image analysis. The speed of our analysis increases by the number of servers used for the analysis compared to workstation based analysis. If we use three computational servers for analysis we see a marked improvement in speed of analysis over a workstation. HUB provides a browser based interface to visualize large data sets and to develop complex imaging workflows. By combining HUB software with this type of system design we aim to provide a roadmap for how to manage and analyze large complex data sets. Moving the storage and analysis of images server side should result in faster analysis and storage of big data in an environment that can scale with the size of the data allowing the new imaging modalities to realize their full potential.

References

1. Tony Hey, Keith Butler, Sam Jackson, and JeyarajanThiyagalingam. Machine learning and big scientific data. Philos Trans A Math Phys Eng Sci. 2020 Mar 6; 378(2166): 20190054.

2. Philip R. Baldwin, Yong Zi Tan, Edward T. Eng, William J. Rice, Alex J. Noble, Carl J. Negro, Michael A. Cianfrocco, Clinton S. Potter, and Bridget Carragher. Big Data in CryoEM: Automated collection, processing and accessibility of EM data. CurrOpin Microbiol. 2018 Jun; 43: 1–8. Published online 2017 Oct 31. doi: 10.1016/j.mib.2017.10.005

3. James Briscoe, Oscar Marín. Review: Looking at neurodevelopment through a big data lens. Science 18 Sep 2020: Vol. 369, Issue 6510, eaaz8627 DOI: 10.1126/science.aaz8627

4. Antony Orth, Diane Schaak, Ethan Schonbrun. Microscopy, Meet Big Data. Volume 4, Issue 3, 22 March 2017, Pages 260-261

https://doi.org/10.1017/S1431927621009909 Published online by Cambridge University Press