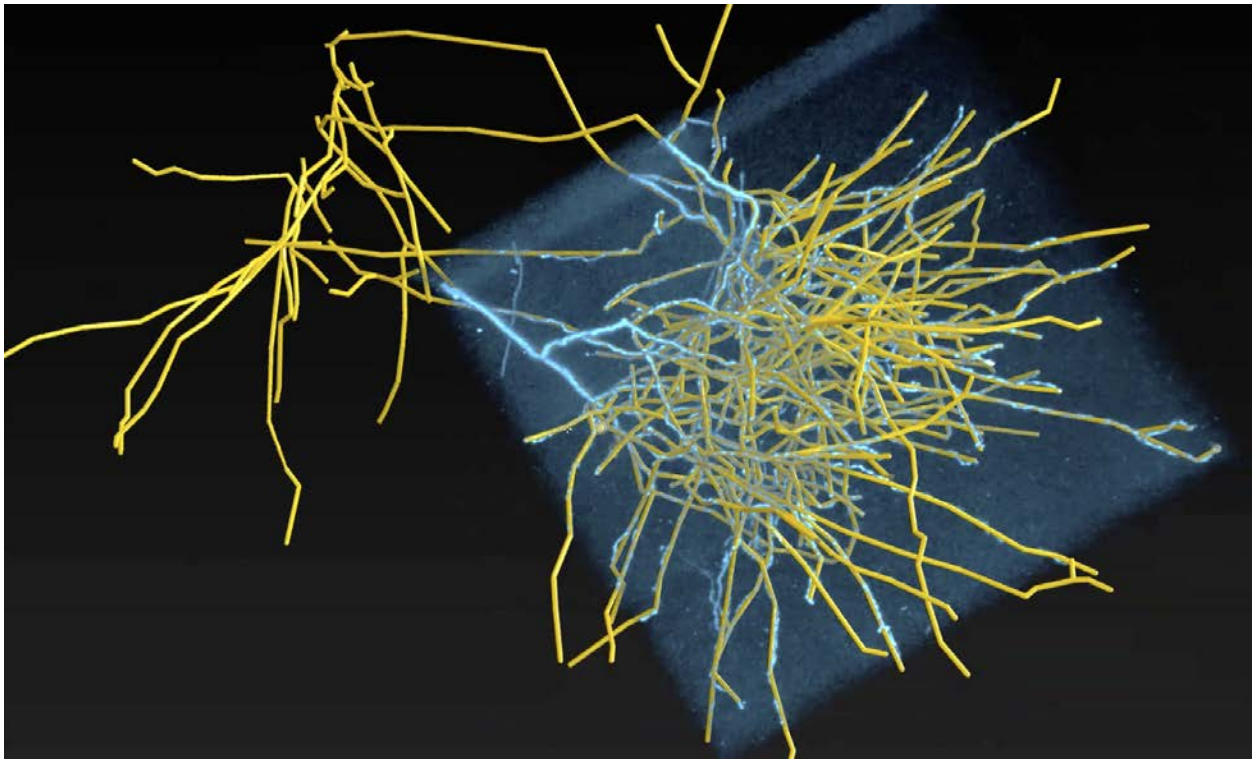
Janelia Research Campus - Open Science Software Initiative

Proposal: janeliaHortaCloud

MouseLight Project Team



BACKGROUND

The MouseLight Workstation and in particular the volumetric Horta[1] neuron tracing environment has become the gold standard for efficient and accurate reconstructions of long range projection neurons in light microscopy data. This software, based on the Janelia Workstation and developed for the MouseLight Team Project, has allowed annotators to efficiently reconstruct entire axonal arbors of individual neurons from whole-brain light microscopy data. There is tremendous value in many of the aspects of this tracing software that was purpose-built for collaborative neuronal tracing and annotation in very large 3D volumes (TB scale data). The primary version of the software was designed as a local instance on a standard hardware server. Though this version has been broadly adopted, there remains a cost and maintenance barrier that we believe limits its broad adoption. We have recently begun developing a cloud-based deployment of Horta which

[1] HORTA (typically stylized as Horta) stands for “How Outstanding Researchers Trace Axons”

we believe will significantly broaden the user base and improve annotators' ability to collaborate across neuroscience labs at different institutions.

When taken as a product, Horta is transformative rather than incremental. To our knowledge, there are no other performant solutions which combine state-of-the-art volumetric visualization, advanced features for 3D neuronal annotation, and real-time multi-user collaboration with a set of enterprise-grade backend microservices for moving and processing large amounts of data rapidly and securely. We believe that a small investment now will catalyze Horta as the default tracing environment within the vertebrate neuroscience community and help cement Janelia's impact upon this expanding field.

EVALUATION OF IMPACT

The Horta tracing environment is already in high demand relative to its domain specialization. Several institutions have deployed it, including the Johns Hopkins University, the Chinese Institute for Brain Research, and the Allen Institute for Brain Science. Many others have expressed serious interest, including Cold Spring Harbor Laboratory, German Center for Neurodegenerative Diseases, UCSD, and the Huazhong University of Science and Technology. Several BICCN (Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network) projects are generating and will likely continue to generate large amounts of image data in the coming years that would benefit from a cloud enabled annotation software solution. The BICCN is therefore poised to make decisions on funding such efforts, possibly by scaling or building on their existing resources at the Brain image Library (BIL). This would be a duplicated effort since the Horta software is a mature tool that can already fill this need. Ensuring broad community acceptance at this time will make data collected across multiple labs widely available for analysis. Thus the projected impact of a modest investment and effort to make this THE preferred annotation platform will pay dividends on two fronts. 1) Data collected within other morphology projects will be available to Janelia researchers for analysis. 2) There could be community engagement and resources to fund additional feature development.

Many labs at Janelia will continue to use the current MouseLight software environment for years to come, including the Spruston & Dudman Labs, anticipated new labs and scientists in MCN and the A&A team. These individuals would be better served with a cloud-based system that allows cross-institute collaboration, and more centralized maintenance and upgrades.

CURRENT CHALLENGES

The underlying software has become a mature platform over the past seven years -- a result of a tight software development feedback loop between developers, annotators and biologists -- and has entered a stable, feature-rich phase. It's been tested by tens of thousands of user-hours.

However, it is finely tuned to the MouseLight project's goals and some aspects need to be developed further to ease the barrier to entry and allow for wider adoption:

- The software infrastructure is complex and difficult to install, even when hardware specifications are followed
- It currently requires large capital investments in hardware (~$35K)
- It is limited to a single, non-standard image format produced by MouseLight microscopes

SOLUTION METHOD

We propose to address the challenges directly by wrapping the existing Horta code base with an AWS deployment to create janeliaHortaCloud (jHC), a fully cloud-based collaborative annotation and tracing environment. Two key insights drive the janeliaHortaCloud concept. First, 3D tracing requires massive amounts of data, and moving all of this data to each annotator's computer is cumbersome, often redundant in collaborations, and unnecessary when they are only viewing one projected 2D plane at a time. Second, Horta requires expensive hardware and a complex installation process. Both of these shortcomings are addressed by janeliaHortaCloud.

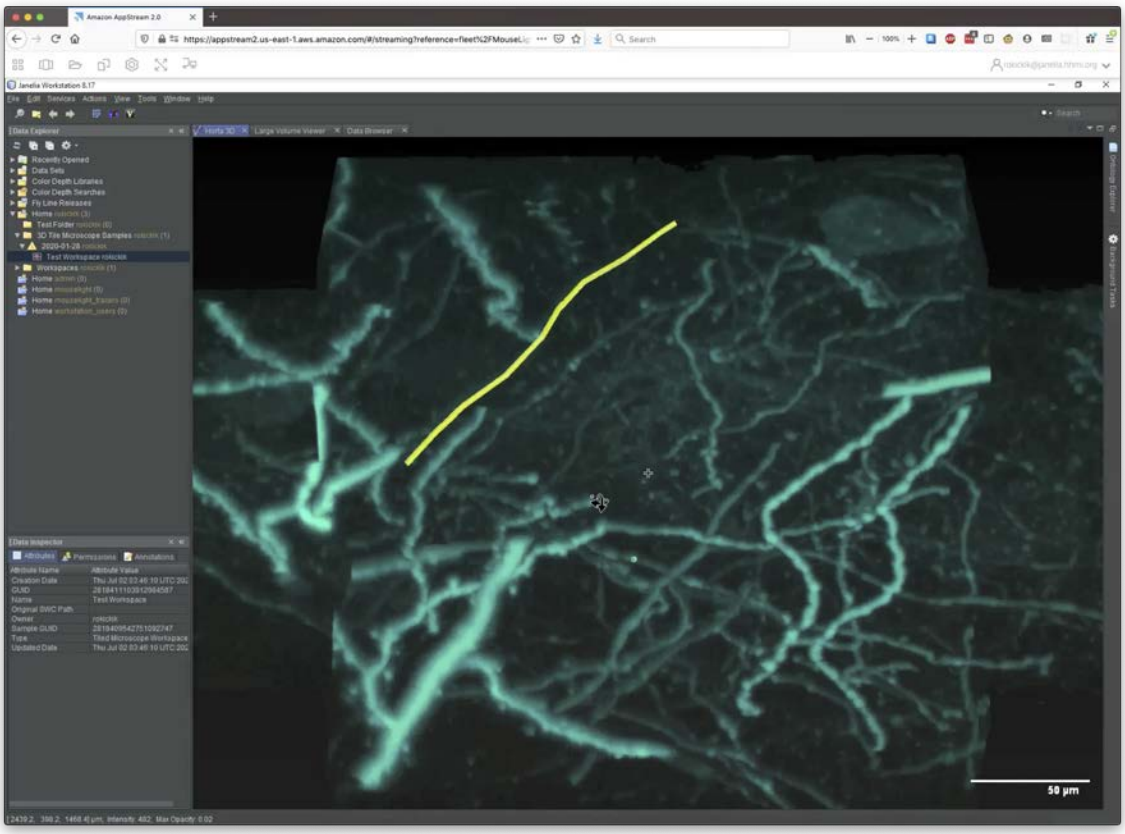


Figure 1: Screenshot of Horta running in a web browser via AppStream

JaneliaHortaCloud takes advantage of recent advances in cloud-based Virtual Desktop Infrastructure (VDI) to perform all 3D rendering in cloud-leased GPUs which are data-adjacent, and only transfer a high-fidelity interactive video stream to each annotator's local compute platform through a web browser (Figure 1).

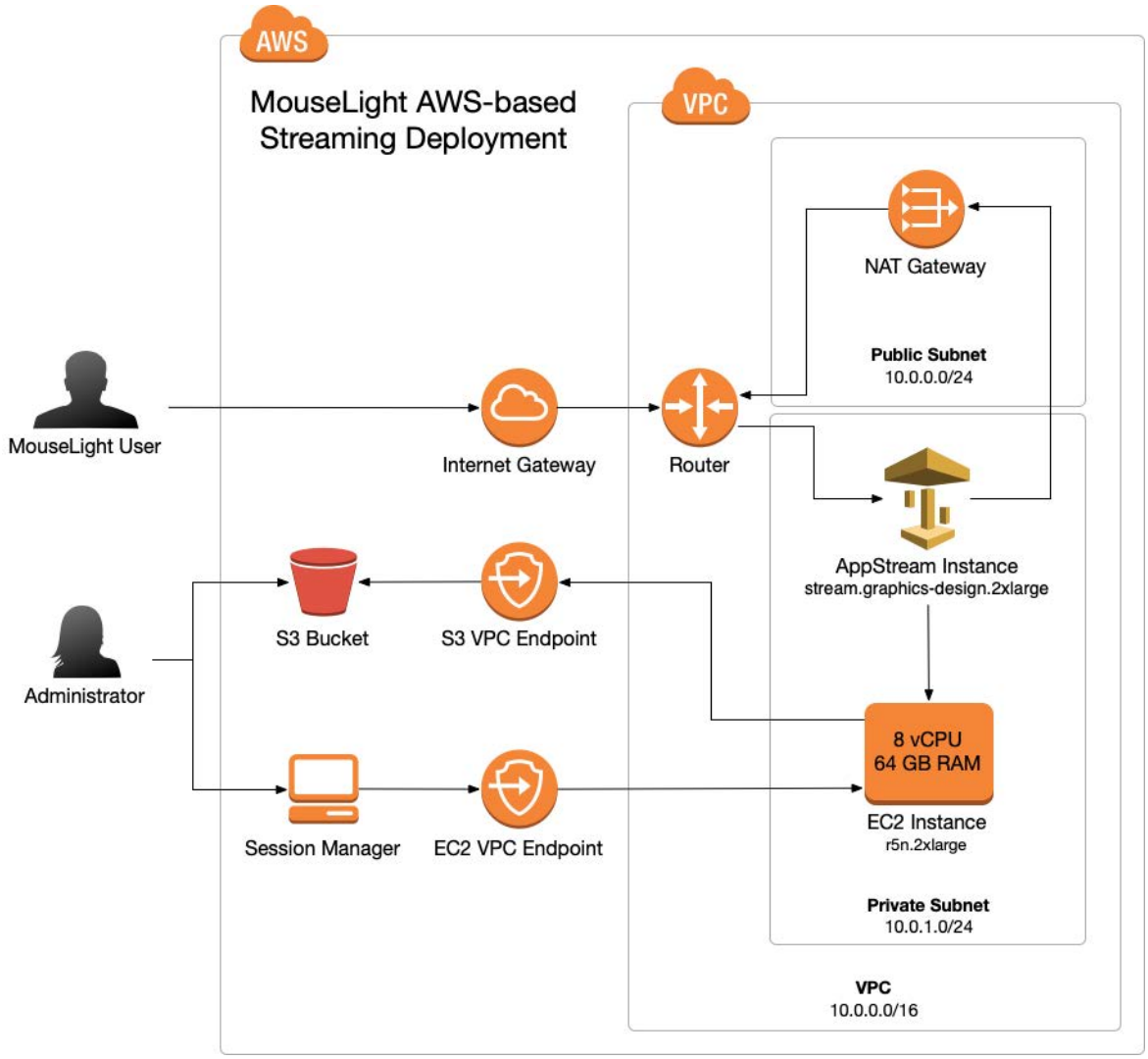


Figure 2: janeliaHortaCloud architecture

Scientific Computing Software has completed a prototype of janeliaHortaCloud on Amazon Web Service (AWS) using AppStream, EC2, and S3 (Figure 2). The entire environment (server and clients) can be installed in a Virtual Private Cloud (VPC), negating most security concerns since all of the backend services will run on machines without public IP addresses. Users can connect to the platform using any web browser on any platform (Mac, Windows, Linux). AppStream supports advanced virtualization of graphics such that all 3D graphics are natively accelerated using any GPU available on AWS. Moreover, all of the hardware specifications can be tightly controlled by

AWS EC2 instance families, eliminating problems associated with supporting different hardware providers at each institution.

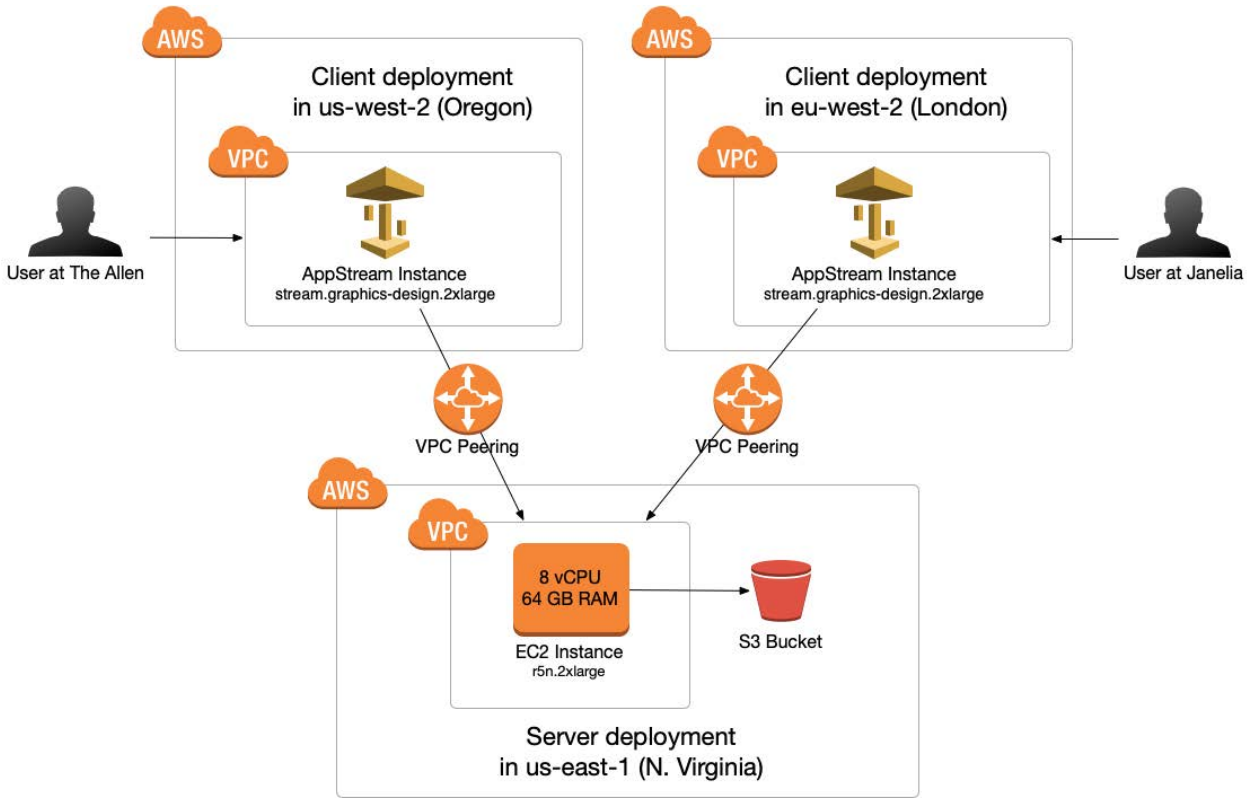


Figure 3: Federated architecture

Using VPC Peering, we can connect multiple AWS VPCs together to form a federated janeliaHortaCloud deployment (Figure 3) where Horta clients can connect to backend servers owned by different institutions. This approach enables cross-institute collaboration and allows each institute to fund its own computational needs.

The existing janeliaHortaCloud prototype is functional and has been evaluated for feasibility by the current MouseLight annotators. They unanimously agree that *the unoptimized prototype cloud deployment is already close in performance (in terms of perceived latency) to our on-premises deployment*. There are many further optimizations that can be made (for example, by scaling the cloud resources or using the high-performance FSx for Lustre file system instead of S3).

The total cost of ownership (TCO) of the cloud-based system has been priced out as similar to the on-premise deployment. This is without accounting for the Systems Team's time at Janelia, so the TCO of an on-premise deployment outside of Janelia is actually much higher than for janeliaHortaCloud.

The proposed project will productionize the janeliaHortaCloud prototype and enable reproducible cloud deployments, as well as extending the system's capabilities to better accommodate non-MouseLight users. In concert with moving Horta into the cloud, we plan to move the software development effort into the open and create an open-source project governance structure that supports involvement from outside parties in future software development and maintenance efforts. The Allen Institute for Neural Dynamics will contribute to this open-source project, and we anticipate that the BICCN may be open to support in the future.

ESTIMATED EFFORT

The total estimated effort is 6.5 FTE months, assuming the development is performed by Scientific Computing staff who are already familiar with the codebases. Depending upon availability of funds, we may be able to reduce the scope of work funded through the OSSI process. The order below is our estimate of rank order.

Automated Deployment (2.0 FTE mo)

We will develop automated deployment scripts for all AWS components using AWS CDK. Repeatable deployments are necessary to reliably maintain this system going forward and will also allow other groups outside Janelia to deploy the services to their own AWS accounts, forming the basis for a federated deployment model.

User Management (1.5 FTE mo)

Horta includes a user manager, but its UI does not make sense to integrate with the cloud. Instead, we will develop a higher-level web-based user management dashboard to allow administrators to add/remove users from the system, controlling both the AppStream user pool and the Workstation user database. It will be deployed automatically as part of AWS automated services deployment. In the future, this component could be extended to provide additional administrative capabilities such as adding/removing users from groups, scaling AppStream resources, or sharing samples.

Open Governance Structure (1.0 FTE mo)

After this initial build-out, and assuming the software gains traction outside of Janelia, we will establish an open (i.e. public) project governance structure which provides:

- Decision-making process for change management
- A venue for collaboration on engineering and architecture of the software
- Code standards, code review, and pull-request process

- Bug triage and assignment
- Code base management responsibilities and point-of-contacts
- Infrastructure for continuous integration and deployment

It is our intention to develop a broad user community where other software developers can contribute resources to add functionality and for ongoing maintenance efforts. The BICCN is keenly interested in using jHC for community annotations and would likely provide funding once deployed and used in the cloud.

Data Management (0.5 FTE mo)

The MouseLight imagery was only partially traced, and a lot of unmined information remains buried within it. MouseLight will make all of this data public so that others can continue to mine it. Recently, AWS generously agreed to host 120 TB of data shared publicly as part of the AWS Open Data program. We will upload ~40 MouseLight samples to the cloud and make the following data available for each sample:

- KTX image data
- Automated segmentation data (when available) (SWC)
- Traced neuron data (curated segmentation) (SWC)
- Registration files (NRRD)
- Annotated volume labels (OBJ)
- Whole sample thumbnail MIPs (when available) (PNG)

We will also research the available tools and establish a reliable method for allowing data admins to upload samples to private S3 buckets, so that pre-published data can be traced in the janeliaHortaCloud system without being made prematurely accessible.

Imagery Import (1.0 FTE mo)

We will extend the system to support the import of additional image formats, beginning with n5/zarr. Current data imports must be performed by system administrators, but the new import functionality will be user-accessible from the Workstation UI, and work seamlessly with data on S3. In the future, the application could be modified to work directly with next-generation file formats (NGFF), but that is out of scope for this initial effort.

Cloud Integration (0.5 FTE mo)

We will develop an automated sample synchronization in the Horta JACS backend so that any image samples available on S3 are automatically imported into the Workstation. We will import

sample MIPs so that they are shown in the Workstation as preview images, to make the large amount of data easier to browse.

FUTURE WORK

We are proposing this effort as a way to jump start an open science project that brings together scientists and engineers from multiple institutions around an already successful software project originally created to facilitate Janelia research. The software is unique and has the potential to attract additional developers and external funding if it is launched successfully as an open project. The cloud-based software architecture is also transformative and takes advantage of bleeding-edge innovation heretofore not exploited for scientific research, and could be a contribution in itself.

If the launch is successful, the initial six month build-out could be just the beginning. We expect to learn a lot of lessons during this implementation, allowing us to make further improvements and refinements to the cloud architecture in the future. We also anticipate the need to integrate different file formats to keep pace with internal and external imaging methodological advances. The Horta tool could also be extended in multiple ways, for example to support annotation and analysis of Lightsheet and expansion microscopy (ExM) imagery, and possibly other types of annotations besides neuron tracing.