Scalable Big-Data Infrastructure to Enable Petascale Neuroscience

J Matelsky; W Gray Roncal; D Kleissas; P Manavalan; D Pryor; R Hider Jr; T Gion; E Reilly; M Roos; M McLoughlin; B Wester

Johns Hopkins University Applied Physics Laboratory, Laurel, MD • jordan.matelsky@jhuapl.edu • brock.wester@jhuapl.edu

MICrONS at The Johns Hopkins University Applied Physics Laboratory

MICrONS seeks to revolutionize machine learning by reverse-engineering the algorithms of the brain. The program is expressly designed as a dialogue between data science and neuroscience. During the program, multiple 1 mm^3 anatomical and functional datasets will be produced, and used to estimate brain networks requiring petabytes of storage. Over the course of the program, participants will use their improving understanding of the representations, transformations, and learning rules employed by the brain to create ever more capable neurally derived machine learning algorithms. Ultimate computational goals for MICrONS include the ability to perform complex information processing tasks such as one-shot learning, unsupervised clustering, and scene parsing, aiming towards human-like proficiency.

Highly-Scalable Storage

This effort requires efficiently storing and analyzing petabytes of electron microscopy data, which are substantially larger than previously published efforts.

Boss leverages Amazon Web Services to provide cloud-native infrastructure to meet this challenge:

- Efficient spatial database
- Novel memory hierarchy
- Scalable data-ingest service
- Authentication via single-sign-on service

Evaluation

Rapid evaluation of graph reconstructions is required to assess the inferences drawn from petascale data:

- **Neural Reconstruction Integrity** (NRI) metric, a novel approach to directly assess the significance of graph components.
- **CONFIRMS** (Creating Optimized Networks for Informing Reconstruction Metrics and Science) evaluation framework and proofreading tools.
- **CIRCUIT** summer program for undergraduates (Connectomics Institute for Reconstructing Cortex: Understanding Intelligence Together).

Cross-Collaboration

In order to enable data-sharing with systems such as KNOSSOS, NeuroData, and DVID, we have developed a Python package called **intern**:

- Provides a consistent syntax for interacting with different data stores and data standards
- Parallelizes common data access workflows
- Abstracts individual database API designs away from the scientist

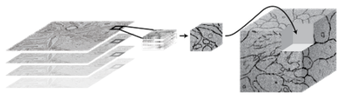
The Boss: Block & Object Storage Service

Motivation

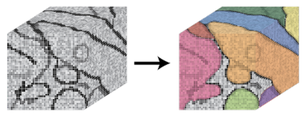
- Image and annotation storage is a common need across connectomics and neuroimaging communities
- New petabyte- scale datasets create unique challenges not common in academic labs that require additional engineering
- Scalability, availability, efficiency and data sharing are key principles for success

Concepts

- The Boss is a multi- dimensional spatial database, provided as a managed service on the Amazon Web Services platform.
- It converts image tiles into a 3D representation that enables efficient arbitrary read/write access:



The Boss stores 64- bit annotation data co-registered to image data:

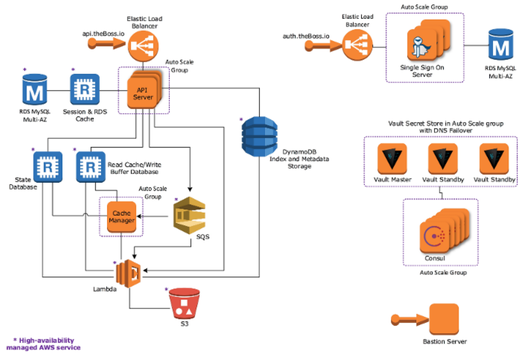


Key Features

- The Boss is built on Amazon Web Services to provide a performant and elastic service while minimizing cost
- A dynamic memory hierarchy architecture migrates data between a fast in- memory database and cheap, durable object storage
- A robust, versioned RESTful API accommodates interaction with the service and arbitrary data access, enabling computer vision in the cloud
- A highly scalable, serverless data ingest process runs completely on AWS
- ASingle Sign- On service provides seamless integration with 3rd party tools (e.g. neuroglancer, proofreading apps)

Status

- The Boss ecosystem is rapidly growing and currently serves over 18 terabytes to 142 users
- Users can autonomously upload data via cutout service or ingest service which supports >4Gbps from multiple users



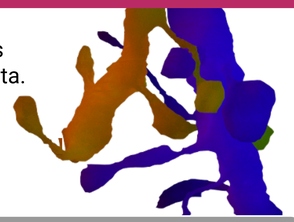
A summary of Boss architecture, leveraging Amazon Web Services infrastructure to provide high-availability, highly-scalable access to large-scale volumetric neuroscience data.

Data-Access: Intern

Intern is a Python library for client-side access and manipulation of large-scale volumetric data.

pip install intern

git clone github.com/jhuapl-boss/intern

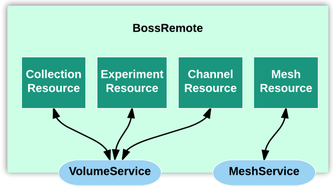


```
from intern.remote.boss import BossRemote
from intern.resource.boss.resource import ChannelResource

boss = BossRemote()
c = ChannelResource("em", "myproject", "mouse1107")
boss.get_cutout(c, 0, [0, 100], [0, 100], [0, 20])
```

Intern simplifies fundamental neuroscience data access from Boss and is extensible to other neuroscience datastores, allowing the focus to remain on research and discovery.

By abstracting API design away from the user, *intern* allows a researcher to use the same simple set of commands to access volumetric data across multiple species of data archives.



Intern uses *Resources* exposed by a *Remote* in order to fulfill *Services* (commonly performed actions, such as volume cutouts or mesh-generation).

References & Links

References

- Kasthuri, N et. al, "Saturated Reconstruction of a Volume of Neocortex." *Cell*, 2015.
- Burns, R et. al "The Open Connectome Project Data Cluster: Scalable Analysis and Vision for High-Throughput Neuroscience." *SSDBM* 2013.
- Reilly, EP, et. al, "Neural Reconstruction Integrity: A novel connectomics metric." *Arxiv* 2017.

Links

IARPA Project Page: https://goo.gl/iWHPhb
IARPA MICrONS Poster: https://goo.gl/BHBgQK
Boss Homepage: https://goo.gl/XMuo4G
KNOSSOS: https://knossostool.org/
DVID: https://github.com/janelia-flyem/dvid

Evaluation

Motivation

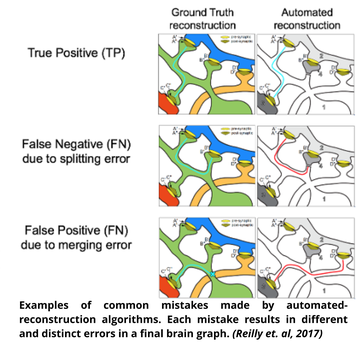
- Manual annotation of neurons and synapses is prohibitively slow
- Automated and semi-automated methods are prone to error
- Metrics are needed to reliably measure reconstruction accuracy

The NRI Metric

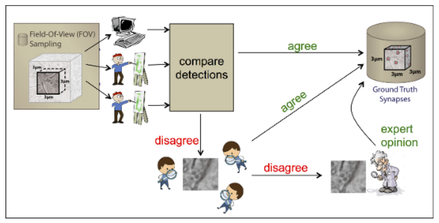
Given matched synapses, Neural Reconstruction Integrity (NRI) considers paths between synapses as a detection problem, rather than considering morphology.

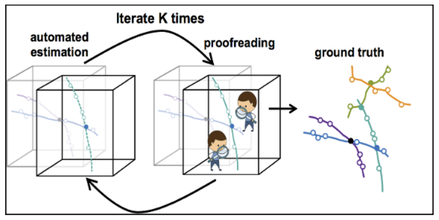
Proofreading Applications

To statistically sample synapses and paths across large volumes, we have developed proofreading tools and a visualization suite, dubbed *Substrate*, that leverage browser-based technologies to dynamically load and render data from the Boss.



Examples of common mistakes made by automated-reconstruction algorithms. Each mistake results in different and distinct errors in a final brain graph. *(Reilly et. al, 2017)*





Proofreaders use purpose-built applications, designed using our graphics API *Substrate*, to ground-truth and detect errors in machine annotations. We show sample synapse (Top Left) and neuron path (Bottom Left) workflows, and Synapse (Top Right) and Neuron path (Bottom Right) Applications.