bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Internet of Things Architecture for High Throughput Biology

David F. Parks[1 2a,*], Kateryna Voitiuk[2a], Jinghui Geng[2b], Matthew A. T. Elliott[2a], Matthew G. Keefe[e], Erik A. Jung[b], Ash Robbins[b], Pierre V. Baudin[b], Victoria T. Ly[b], Nico Hawthorne[b], Dylan Yong[b], Sebastian E. Sanso[d], Nick Rezaee[d], Jess Sevetson[a], Spencer T. Seiler[a], Rob Currie[d], Keith B. Hengen[g], Tomasz J. Nowakowski[e,f], Sofie R. Salama[a,c,d], Mircea Teodorescu[b,d], David Haussler[a,c,d]

[a] *Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95060, USA*
[b] *Department of Electrical and Computer Engineering, University of California Santa Cruz, Santa Cruz, CA 95060, USA*
[c] *Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA, 95064, USA*
[d] *UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95060, USA*
[e] *Department of Anatomy, University of California San Francisco, San Francisco, CA 94143, USA*
[f] *The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California San Francisco, San Francisco, CA 94143, USA*
[g] *Department of Biology, Washington University in St. Louis, St. Louis, MO, 63130, USA*

Abstract

New cell culture techniques have led to complex tissue models in biological experiments. For example, 3-D cerebral organoids provide a more realistic model of the human cortical tissue. However, these cell culture experiments are restricted by high costs and limited labor. A massively scalable and cost efficient platform for tissue experiments would benefit genomics, neuroscience, and translational medicine by enabling advanced high throughput tissue screens. Cloud computing and the Internet of Things (IoT) provide new tools for managing multiple experiments in parallel that are remotely controlled through automation. We introduce a cloud-based IoT architecture that takes advantage of these tools to offer an environment where researchers can run thousands of cell culture experiments at once. This technology allows studies with cell cultures to be performed at scales far beyond a single lab setting, democratizing access to advanced tissue models and enabling new avenues of research.

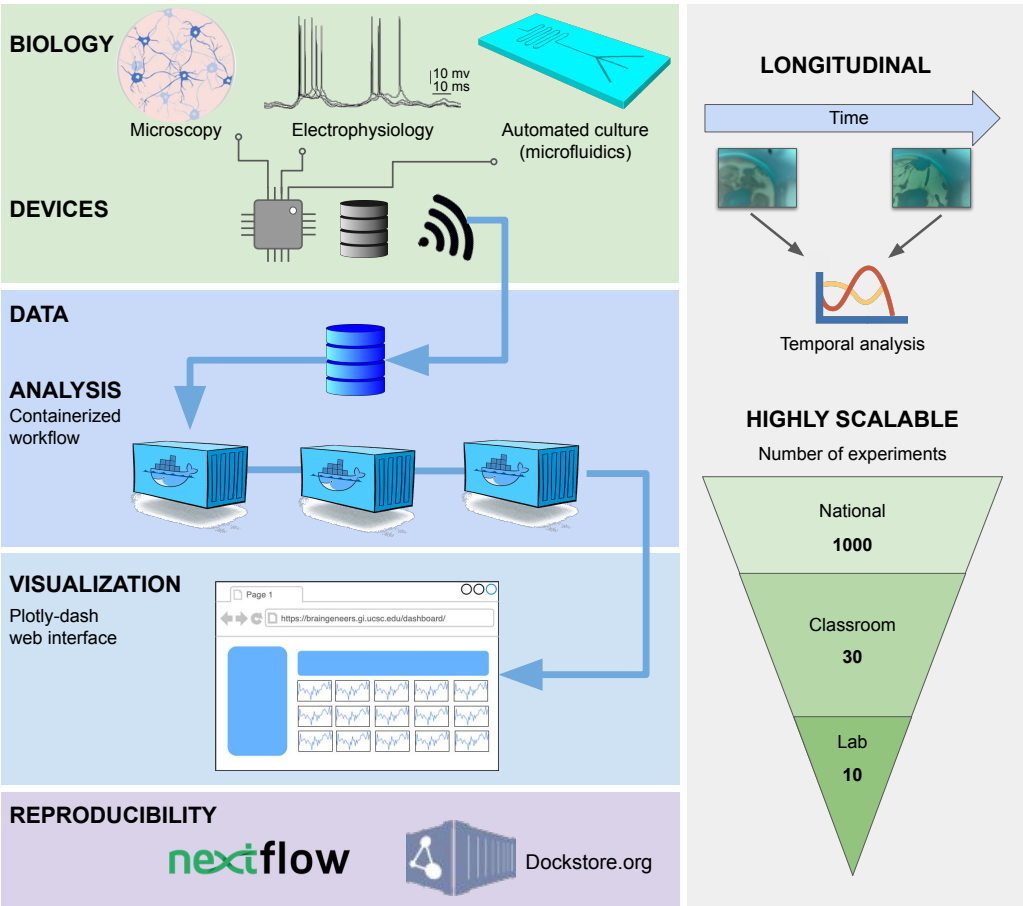
Keywords:
IoT, automated culture, biology, cloud, containerization, electrophysiology, microscopy

[1] Lead Contact
[2] Co-Author
[*] Corresponding author
Email address: dfparks@ucsc.edu (David F. Parks)

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

IoT Cloud Laboratory

Experiments are automated through cloud connected devices to allow scalability, reproducibility, and online monitoring.



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1. Introduction

Recent advances in cell culturing have produced new tissue models with applications to genomics, neurodevelopment, and medicine (Okano, 2002; Furusawa & Kaneko, 2012; Dutta et al., 2017; Rossi et al., 2018). Human cell cultures, especially recently developed 3-D organoid models, enable the study of disease, drug interactions, genetic alterations, and response to neural stimulation in ways infeasible *in vivo* (Kim et al., 2020; Schutgens & Clevers, 2020; Gao et al., 2014). Dozens of cultures are commonly grown (Lancaster et al., 2013, 2017; Paşca et al., 2015) and measured (Széll et al., 2020; Putzeys et al., 2019; Rolston et al., 2009) at a time, but massively parallel longitudinal experiments on organoid models are required to provide insights into organ development and function, and to use them as platforms for drug discovery (Schuster et al., 2020; Khan et al., 2021). Recent complex culturing experiments have focused on the large scale production of samples (Sarrafha et al., 2021), but managing experiments at this scale leads to a host of technical challenges in data management and analysis.

A major computational challenge in long-term longitudinal experiments is the real-time integration of multimodal data. Solving this problem allows experiments to be massively scaled through automation and machine learning. For example, electrophysiology, microscopy, and microfluidics are fundamental methodologies commonly combined in cardiology, gastroenterology, and neuroscience (Ionescu-Zanetti et al., 2005; Huang et al., 2012). Electrophysiology monitors cell health (Bera, 2014; Gamal et al., 2018; Gerasimenko et al., 2020) and neural network activity (Hansel et al., 1995; Beggs & Plenz, 2003; Poli et al., 2015; Huang et al., 2017; Sharf et al., 2021). Microscopy provides information on cell morphology and growth (Stephens & Allan, 2003; Jensen, 2013). Microfluidics devices enable automated media and drug delivery and, via inline sensors, can apply feedback control to maintain desired physiological conditions (Funamoto et al., 2012; Jung et al., 2019).

Each of these methodologies encounters data challenges when scaled. Increasing the number of recording electrodes in electrophysiology experiments makes datasets unwieldy, quickly reaching terabytes per day. High-resolution multi-well microscopy produces large datasets that also must be accessible for real-time image viewing. Nuanced video processing algorithms, data post-processing, and longitudinal cross-system monitoring are essential. Microfluidic devices require carefully timed control systems that integrate data across various biosensors. Combining these methodologies requires data management that ensures future reproducibility and experimental quality control. An open-source architecture for integrating data and automating analysis of massively parallelized experiments needs to be developed.

Advances in IoT and cloud computing offer new solutions to overcome these challenges. IoT is a technology that remotely coordinates communication and management of large numbers of devices that are paired through the cloud. We recently developed affordable, open-source, internet-connected devices ("data acquisition modules") for microscopy (Picroscope) (Ly et al., 2021; Baudin et al., 2021), electrophysiology (Piphys) (Voitiuk et al., 2021), and multiple groups have developed microfluidic devices (Yuan et al., 2018; Ibrahim et al., 2019) with capabilities for integrated biosensors. IoT allows a globally networked fleet of lab devices to be controlled and managed through the cloud.

Commodity cloud compute from major companies (Amazon Web Services, Google Cloud Platform, Microsoft Azure, etc.) and academic coalitions (Pacific Research Plat-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

form) (Smarr et al., 2018) provide robust environments to synchronize IoT devices and manage data. Tools like Kubernetes, which coordinate massively distributed computing processes, are well supported. Containerized workflows are used to decouple complicated processes into manageable software programs. Cloud-based data pipelines and containerized applications simplify workflow management and execution (Hasham et al., 2018; Marcus et al., 2007; Hung et al., 2019; Abe et al.; Jansen et al., 2020).

Cloud-controlled IoT systems enable researchers to take advantage of machine learning to optimize scientific studies. Machine learning allows IoT device to be controlled and tuned in real-time using automation. Online data streaming and analysis tools (Wagenaar et al., 2005; Chung et al., 2017; Siegle et al., 2015; Yger et al., 2018; Lee et al., 2020) are combined with learning algorithms to create a faster turnaround between hypothesis, experiment, and re-testing (Raghu & Schmidt, 2020). Cross comparisons of longitudinal recordings from parallel experimental replicates add new possibilities for inference (Dhawale et al., 2017). Such algorithms are easily deployed across all experiments using cloud containers. This automated environment removes batch affects between experiments, because it enforces a consistent standardized protocol.

We introduce a cloud-based IoT architecture where researchers can manage thousands of cell culture experiments in parallel. Scientists access their experiments from anywhere online to monitor results in real-time. Microscopy, electrophysiology, and microfluidics are all controlled through a single integrated environment. Multimodal data across all experiments are saved to ensure future reproducibility. This framework is open source and runs on a freely available compute platform for research institutions, the Pacific Research Platform (PRP) (Smarr et al., 2018), and can also be easily deployed on major commercial cloud platforms.

Providing a massively parallel culturing hub online fundamentally changes the role of complex cell culture experiments. Economies of scale remove constraints on labor and localized equipment. These reduced costs democratize access to the field, allowing more researchers to perform cell culture experiments remotely. Furthermore, the online portal enables complex biological experiments to be incorporated into learning modules targeted at the college and high school level (Hossain et al., 2016; Kim et al., 2016; Beattie et al., 2020). For cerebral organoids, high throughput screens done on genetically altered tissue may provide insights into autism and pediatric glialblastomas. Major advances in medicine, neurodevelopment, and education can be achieved if new technologies in cloud computing and IoT are applied to cell culture experiments.

2. Results

Cost, scalability, maintainability, and scientific reproducibility were the fundamental requirements for our high throughput experimentation software. Low-cost is made possible by cloud computing platforms offering affordable commodity compute and storage resources at supercomputer scales. Scalability and maintainability are achieved through IoT management of devices and software containerization of data analysis processes, which both offer plug-and-play approaches with minimal dependencies between components. Scientific reproducibility is embedded through standards-based workflow definitions using Nextflow and Dockstore.

Figure 1 depicts the high-level overview of the system. Data acquisition modules (devices) execute experiments in the lab. Each module performs a specific task such as

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

electrophysiology, microscopy, and biochemical assay. Users interact with the devices through a web-based user interface, or a lower-level software API. The software API controls devices and enables any program to control the flow of experiments. Logistics of device management, communication, and data storage are handled through the Pacific Research Platform (PRP) and Amazon Web Services (AWS). In the following sections, we describe each component of the architecture.

2.1. Device management, communication, and control using IoT and MQTT

The data acquisition modules are lightweight and general-purpose IoT devices. The IoT devices connect to the various services that support user control, data storage, analysis, and visualization via the MQTT (Message Queuing Telemetry Transport) protocol. MQTT is a well-supported, industry-standard publish-subscribe messaging protocol.

Figure 2 depicts the central role MQTT plays in coordinating data acquisition modules and user interface communications. The MQTT protocol maintains the state and connection status for each device. It also provides a simple, lightweight publish-subscribe platform with defined *topics*. The *topics* are used by devices or user interface components to organize communication. There are two types of *topics*: a *topic* per each device (electrophysiology, microscopy, or any device performing experimental measurements or recording), and a *topic* per each running experiment. Each experiment is also assigned a UUID (Universally Unique IDentifier) which becomes an active *topic* for the period of operation.

An experiment starts when MQTT messages are published to the appropriate experiment and device *topics*. Devices subscribed to those *topics* receive the messages and take the appropriate action. Actions can also be taken automatically based on sensor readings. For example, a temperature sensor that detects overheating can publish an emergency stop message to the appropriate devices. Actions may involve sending users alerts explaining errors or requesting intervention.

2.2. Data storage using Ceph/S3

Figure 3 shows how devices store experimental data. Primary storage and data processing are implemented on the PRP through a distributed commodity compute cluster based on Kubernetes[3] and the Ceph (Weil et al., 2006) distributed file system. Ceph provides a highly scalable S3 interface to a virtually unlimited data store. Ceph/S3 is the primary storage for all datasets, small to terabyte-scale, commonly recorded by electrophysiology, microscopy, and biochemical assays. Our larger parallelized data processing tasks have peaked at over 5 GB/sec of concurrent I/O from S3, demonstrating the substantial scalability of the file system. Access to the Ceph/S3 data store is universally available on the internet, making it an excellent place to share large datasets across institutions.

As a research-oriented compute cluster, the PRP (Pacific Research Platform) does not provide strong SLAs (Service Level Agreements) for the data store. Network outages due to local network, power, or user error can cause temporary service disruptions. No guarantee is made against data loss, though the Ceph filesystem provides mechanisms to guard against common failures such as losing a node or storage media. We mitigate

[3] https://kubernetes.io/

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

against data loss by scheduling a Kubernetes Cron Job with a nightly backup of all data from Ceph/S3 to AWS Deep Glacier, a cloud IaaS (Infrastructure as a Service) service providing a long-term tape storage solution. Also, all data-producing edge devices maintain a local cache that can withstand a temporary service disruption. These design choices come with overhead, but arguably they would be necessary even if Ceph/S3 provided stronger guarantees against service disruption and data loss.

2.3. User interface using Plotly Dash

A Plotly Dash [4] interface is easy to develop and code in Python, a common language for data science. Plotly offers a rich set of interactive plotting functionality, including specialized biology-focused visualizations. Dash provides a template to build user interfaces that implement the Observer Design Pattern (Gamma, 1995) making for an extensible and maintainable environment.

A dedicated server runs a single Plotly Dash instance under which the user interface and visualizations run as a multi-page Plotly Dash web application. This application can plot data from past experiments saved on Ceph/S3, or it can publish MQTT messages to the device or experiment *topics* in real time. Figure 4 shows a how a user would visualize a "Piphys" electrophysiology device streaming data.

2.4. Data streaming using Redis

Real-time streaming and real-time feedback are facilitated through a Redis service. Redis is a high speed database that acts as an inter-server and inter-process communication service. It is straightforward to interact with Redis using many languages, including Bash, Python, and C. Raw data feeds are sent to Redis only when the user is actively interacting with a data stream, e.g., looking at a real-time visualization, the UI client sends MQTT keep-alive messages to keep the data stream active. While MQTT is appropriate for small messages, Redis is the main communication method for larger blocks of data.

Figure 5 introduces a mechanism for handling large-scale real-time data streams. Redis provides common data structures with the inter-process locking required to coordinate between services running on separate devices. It provides a way for a data producer to publish a real-time stream of data, such as an electrophysiology recording, for a consumer of that data, such as the Plotly Dash UI, to coordinate between each other without direct dependencies between them. Data transformations using Redis shown in Figure 5 are discussed in Section 2.6. A Redis stream is effectively a queue that can be capped in length, so that old data is automatically dropped once the maximum size of the stream is reached. Consumers, such as the Plotly Dash website, can send a recurring MQTT message to the relevant data producer to start the data stream and read the data as it is produced. A Redis service interruption merely pauses data visualization. The data producers stream a raw data feed to Redis in real-time while logging data in batches to Ceph/S3. The Ceph/S3 object store remains the primary source for data storage, and the data transfer to Ceph/S3 is resilient to service disruptions. There is no guarantee against the loss of data in the streaming approach, which is why Ceph/S3 is the primary datastore, and the Redis stream is reserved for visualizations that can incur service interruptions without lasting consequences.

[4] https://plotly.com/dash/

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.5. Data processing using containerization and workflow definitions

Final data analysis is performed using containerized workflows built with Docker and Kubernetes and then deployed using Nextflow. Longitudinal electrophysiology, microscopy and microfluidic experiments commonly produce datasets on the multi-terabyte scale. Large scale machine learning especially relies on S3 for reading terabyte scale datasets. Data analytics tasks such as neural voltage signal analysis, machine learning, and image analysis require substantial computing resources and processing in multiple stages.

Running computation in the cloud utilizes containerization to package the job. Containerization is the process of packaging up code and all its dependencies, so an application runs reliably in any computing environment. Containers are efficient and lightweight, they share a single host operating system (OS), and each container acts as an independent virtual machine without additional overhead (unlike full hypervisor virtual machines which replicate the OS). The container can be uploaded to a repository (for example, on Docker Hub) and downloaded and run on any computer. This includes servers in a cluster or a local lab computer.

Containers are a popular tool for maintaining independent and reproducible code with minimum overhead. Containerization guarantees that the code written today will work tomorrow, because the code and all of its software dependencies, APIs, and versions are packaged into a binary. While this is a big step forwards towards reproducible scientific work, it is not sufficient on its own.

We introduced Dockstore.org (O'Connor et al., 2017) as the next logical step in scientific reproducibility, building on containerization technology. Dockstore.org is a website dedicated to hosting containerized scientific workflows formalized by workflow definitions. The formal definition of a workflow is its inputs, outputs, steps, dependencies, and the containers they run on. A common workflow language formalizes a containerized software process to ensure that organizations can run each other's software in a standards-compliant manner. Several formal workflow definition languages exist: Nextflow (Di Tommaso et al., 2017), Common Workflow Language (CWL) (Amstutz et al., 2016), and Workflow Description Language (WDL) and are all supported by Dockstore.

Besides being a formalized workflow language, Nextflow provides a workflow runtime engine capable of deploying containerized processes to various platforms such as Kubernetes, AWS, Google Cloud, and Azure. Figure 6 depicts a standard electrophysiology data processing workflow run by Nextflow and deployed to the Kubernetes-based platform on the PRP. All workflows receive a standard UUID (Universally Unique IDentifier) pointer to a dataset, allowing them to find the raw data or pre-processed data produced by a dependent workflow.

2.5.1. Example: electrophysiology data processing workflow

Let us consider an example workflow for an electrophysiology experiment. The goal is to perform spike sorting, or detect the action potentials (spikes) of neurons by analyzing the voltage recordings. The workflow consists of 3 Jobs that occur in sequence:

For Job 1, a subset of the electrophysiology data is scanned to identify active channels. A JSON file with active channel information is recorded to Ceph/S3. This step requires a single task/container to run.

For Job 2, the dataset is converted from its raw 2-byte data (`int16`) format into a 4-byte floating-point format (`float32`) necessary for data analysis. Since the dataset

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

is typically large (commonly in Terabytes), the substantial compute and I/O workload must be distributed across the cluster. One job per data file is launched, downloading the raw 2-byte data file from Ceph/S3 and uploading a 4-byte data file (`float32`) back to a temporary location on Ceph/S3. In the process, the data is separated into individual channels for processing in the next step. Notice that the conversion process must download and re-upload the full dataset because multi-terabyte datasets are too big to fit on local nodes.

For Job 3, the data for active channels is pulled from the distributed filesystem, then spike sorting and spike timing analysis is performed. The results are placed back on the distributed filesystem.

While each job runs in series and depends on the last, there are no dependencies between the jobs other than those involving the data that is posted to the primary datastore Ceph/S3. Each dataset has a unique ID (UUID) which also serves as a location pointer to where data is stored on the Ceph/S3. This UUID is the only parameter passed between jobs.

Beside the illustrated example in this section, Figure 7 shows a more general overview of resources employed and parallelization of the data processing by workflows including electrophysiology, imaging and chemical assays.

2.6. Real-time analysis, data processing, and transformations

Deploying containerized workflows via Nextflow works well for large-scale post-processing and data analysis but does not provide a mechanism for real-time visualizations and experiment control.

The Redis in-memory database service coordinates the real-time exchange of data between many producers and consumers. For example, an electrophysiology recording on 32 channels at 25 kHz will produce a data stream of 1.6 MB/sec, which a user may want to monitor in real-time. Equivalently, a microscopy recording could provide a stream of images for visualization for real-time experimental metrics.

Transformation of data with visual enhancements applied in real-time is often more informative than seeing raw data. Data transformations are performed by containerized processes that read a stream of data and write a new stream of transformed data. For example, a container reads a raw electrophysiology stream and writes a new steam with the bandpass filtered data. After applying the data transformation, a visualization such as a Plotly Dash web page would read the appropriate data stream output. Data transformations have no dependencies other than the Redis stream they read from and can be entirely independent workflows. Transformations can easily be added or changed without changing any other software infrastructure components.

3. Discussion

Our software architecture is designed to support different kinds of data acquisition modules that measure and report data. Here we focus on three types of modules for proof of concept:

1. Electrophysiology – recording and stimulation of neural cell cultures
2. Microscopy – imaging of cell cultures

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3. Automated culture – feeding cells and sampling media for metabolites and RNA expression using a programmable microfluidics system

We will look at each of these data acquisition modules (IoT-based edge devices) in turn and discuss how they interact with the software architecture and user.

For this example, we assume users will interact with devices through the web UI application. Users can be located anywhere on the Internet without concern for where the physical devices are. This facilitates cross-campus and cross-laboratory collaborations. For example, we often perform electrophysiology and microscopy experiments from Santa Cruz on devices located 90 miles away in San Francisco. Experiments do require some manipulation by a researcher at the local site (i.e., placing cell cultures on the devices and performing adjustments if components are misaligned).

To begin an electrophysiology experiment, a user opens the browser with the web application. The application queries AWS IoT service for online electrophysiology devices. The device can be Piphys (Voitiuk et al., 2021) or any platform/recording system whose computer runs the same code that responds to the IoT architecture and can control the system programmatically. When the user selects a device, an MQTT 'ping' message is sent to the relevant device every 30 seconds, indicating that a user is actively monitoring data from that device. As long as the electrophysiology device receives these pings, it will send raw data to its Redis stream. Since the device is responsible for only a single data stream, many users can monitor and interact with the particular device without additional overhead. If the device has not received user messages for at least a minute, it will cease streaming its data. This protocol ensures the proper decoupling of users from devices, and devices are not dependent on a user gracefully shutting down the connection.

As shown in Figure 5, one or more data transformation processes can read the raw data stream and post a processed stream of data, such as real-time spike sorting. The web visualization can display the appropriate transformed data stream for the user.

Stopping the experiment will automatically initiate a batch processing workflow on the Kubernetes compute platform. The workflow can be configured to include job modules such as spike sorting, clustering, and other customized metrics of neural activity.

Microscopy, such as the Picroscope, typically operates at a lower sampling rate and over a longer continuous period than electrophysiology. Microscopy devices record images of cell culture morphology at varying focal layers and time frequency. As with electrophysiology, these images are initially buffered locally and then flushed to the Ceph/S3 filesystem every few minutes. A user will view the data in the same web UI portal as electrophysiology. Since cell culture morphology changes relatively slowly, microscopy visualizations do not require real-time Redis streaming. The user may update the parameters of the microscopy recording with MQTT messages sent to the device *topic* updating the state.

Assay devices support the lifecycle of the cell culture, providing new media and taking regular measurements relevant to the cell culture's health and environmental state. Much like microscopy, most of these measurements are sampled continuously over the lifetime of the culture and are posted directly to Ceph/S3 at regular time intervals. When the user accesses a UI page detailing the lifecycle of the culture, these metrics will be pulled in near real-time from Ceph/S3. The user can update and change metrics related to cells lifecycle by an MQTT message from the UI page to the device to update its state and initiate a change in the device behavior.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3.1. Scaling

In the previous section, we considered one experiment with a few data acquisition modules running in a single lab. This section considers hypothetical studies of tens to thousands of experiments operating simultaneously. Each user will use different features of the devices, and there would be a virtually infinite combinations of features when many devices are deployed. We define three use cases and provide an analysis of these and their assumptions, we call the use cases, Science, Student, and National. We provide a distribution over the devices basic functions that we expect the users will employ in each case. For each case we provide estimates of CPU, Network, and Storage resources required, visualized in Figure 7. Also provided in Figure 7 is an estimate of cloud computing and storage cost based on AWS pricing. The use of the PRP academic compute cluster precludes the majority of these costs and speaks to the value the PRP brings to academic institutions.

In the Science use case, we assume a higher degree of active imaging, electrophysiology, and raw voltage recording. This use case focuses on more resource-intensive lab use in the pursuit of scientific inquiry at high detail. In this configuration, storage requirements are the most significant bottleneck, growing at tens to hundreds of GB of data per hour. We find that tens of devices are appropriate for this use case before resource utilization becomes excessive.

In the Student use case, we imagine a limited number of universities using the devices to teach classes in cell biology on live cultures hosted at a remote lab. In this use case, we assume a scale on the order of hundreds of devices. Users in this scenario will rely heavily on visualizations, including both real-time microscopy and electrophysiology. The lab that hosts hundreds of experiments with the expectation of concurrent access will require additional network bandwidth beyond what is available in a typical lab or office. At least two Gigabit network ports and matching ISP bandwidth would be necessary to support the load. At this scale, limiting electrophysiology recordings to active spiking events is imperative.

Lastly, in the National use case, we consider a scaled-out fleet of thousands to tens of thousands of devices. This case assumes wide-scale adoption by laboratories or secondary education facilities across the country or world. In this case, we cannot reasonably support raw voltage trace recordings. At this scale, the devices would be used for limited filtered voltage trace recordings using on-device spike detection and real-time imaging with direct user feedback. This scale requires substantial cloud computing resources to support the load. It will also require significant wet lab infrastructure at the site(s) housing the biology as well as expenses of cell culture maintenance. However, given this investment, our infrastructure can enable remote experimentation by a large and diverse population.

4. Conclusion

This paper outlines an IoT software architecture that supports the high throughput analysis of electrophysiology, microscopy, and experimental assays on cell cultures and in organoids. This innovation lifts the constraints of having *in vitro* experimentation on a single benchtop. Instead, we emphasize the benefits of having a centralized online hub where thousands of automated experiments are managed in parallel through a portal. Such a system optimizes costs and scales the size of experiments by orders of

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

magnitude. This architecture opens new possibilities for using complex tissue models at a scale necessary for high throughput drug screens and comprehensive gene knockout studies. The architecture is built with open-source design and scientific reproducibility in mind. Use of the open academic PRP compute platform encourages cross institutional collaborations. We hope our technology benefits stem cell biology, neurodevelopment, and medicine by democratizing access to organoid and other complex cell culture experiments for research and education.

5. Acknowledgments

This work is supported by the Schmidt Futures Foundation SF 857 (D.H.). Research reported in this publication was also supported by the National Institute Of Mental Health of the National Institutes of Health under award number R01MH120295 (S.R.S.), the National Science Foundation under award number NSF 2034037 (M.T.), the National Defense Science and Engineering Graduate Fellowship (00002116, M.G.K.), and a gift the William K. Bowes Jr Foundation (T.J.N). K.V. was supported by grant T32HG008345 from the National Human Genome Research Institute (NHGRI), part of National Institutes of Health (NIH), USA. D.H. is a Howard Hughes Medical Institute Investigator.

Through the Pacific Research Platform, this work was supported in part by NSF awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. Thanks to CENIC for the 100Gpbs networks.

Author Contributions

R.C. was the chief architect for the infrastructure design, and D.F.P made it work, creating necessary integration and cross platform communication. V.T.L, P.V.B worked on microscopy. K.V., J.G., M.G.K, A.R., N.H. worked on electrophysiology and analysis and K.B.H advised. S.T.S worked on automated IoT cell culture. D.F.P. and S.E.S. created the web user interface design. D.Y. and N.R. worked on data infrastructure, backups, and automated data processing pipelines. D.H., M.T., S.R.S, and T.J.N. supervised the team and secured funding. D.F.P., K.V., J.G., and M.A.T.E., wrote the manuscript with support from E.A.J.. All authors reviewed the manuscript.

Declaration of Interests

The authors have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

6. Figure Titles and Legends

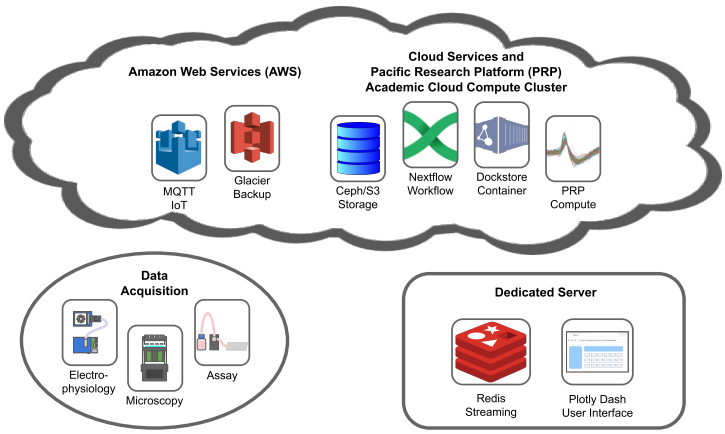


Figure 1: Overview of the software architecture. Data acquisition is performed by lightweight internet-connected devices, often running on low-power compute such as Raspberry Pi, controlling and assaying an experiment. Devices and cloud components communicate using message dispatching (IoT MQTT) with device authentication (IAM). Data is pushed from the devices to a large compute cluster for storage (Ceph/S3) and real-time streaming to visualizations (Redis). Users can control and visualize devices and data through a web interface (Plotly Dash). Real-time algorithmic feedback and batch processing are facilitated using compute containers (PRP), container repositories (Dockstore), and workflow management (Nextflow). All data is backed up to long term storage (Glacier) automatically.

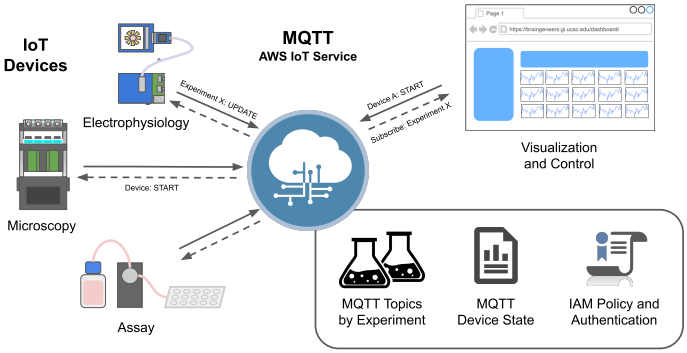


Figure 2: Inter-device MQTT Message Broker. The MQTT message broker provides integration and control over multiple internet-connected instruments. The functionality supports *clients*, data acquisition modules or software applications, to connect and subscribe to topics set by a *publisher*, such as the user interface (UI), with the proper authentication protocols. By doing so, *clients* subscribed to the topic will be informed of the state of each data acquisition module (e.g., start, stop, etc.) and parameter changes throughout an experiment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

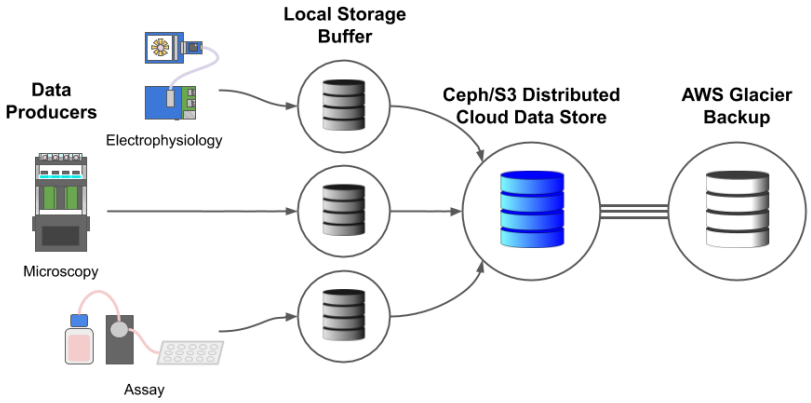


Figure 3: Data storage architecture. Data storage is buffered to the local device before being delivered to cloud S3 storage. Network and cloud service disruptions are expected. With the real-time data feed, interruptions only impact active visualizations of the data, which is acceptable, but the loss of experimental data is not. Each device buffers data to its local storage before making a best-effort attempt to upload it to the S3 distributed object store. Data may be buffered until the local storage is exhausted (typically enough for at least a day). The S3 distributed store is backed up to AWS Glacier to guard against user error (accidental deletion) and the loss of the S3 service. Cloud providers like AWS, GCP, and Azure have strong S3 service level agreements, unlike academic clusters such as the PRP.

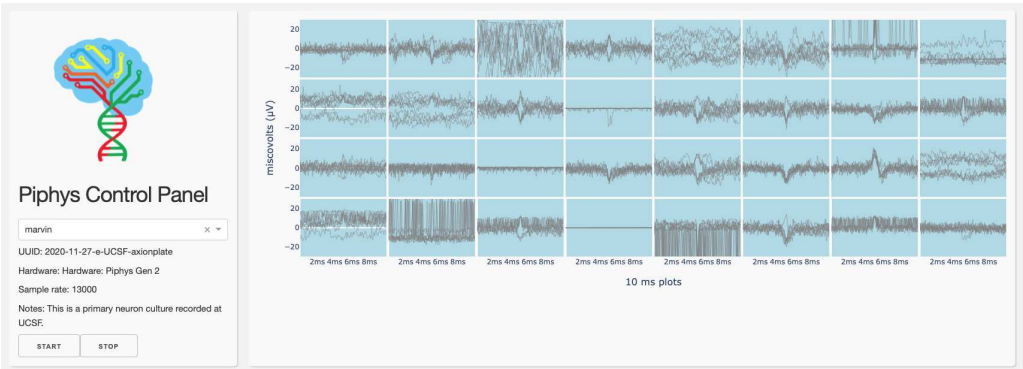


Figure 4: User interface. Example of a dashboard for electrophysiology signals generated by "Piphys" (Voitiuk et al., 2021) IoT device.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

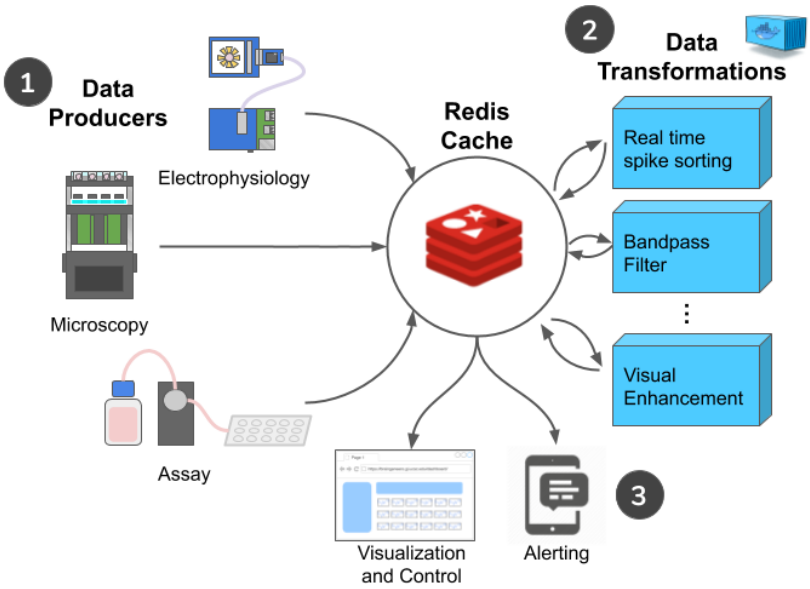


Figure 5: Real-time data visualization: (1) Electrophysiology, Microscopy, and Experimental Assay IoT devices produce real-time data streams on-demand only when a user is connected to a visualization that utilizes that stream. (2) Data transformations process raw data into a variety of helpful forms. Each independently containerized transformation reads a data stream and produces a new data stream. (3) Visualization and alerting notify IoT devices via MQTT that data streams are needed.

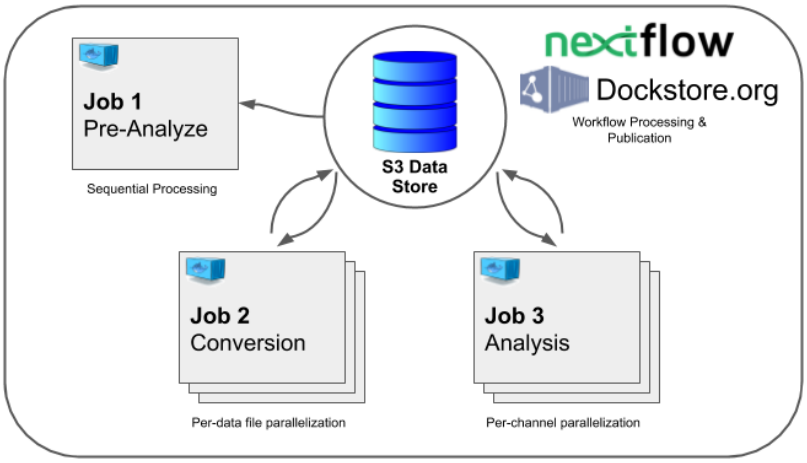


Figure 6: An electrophysiology experiment post data processing workflow. (1) A subset of the data is analyzed to determine which channels are active, (2) raw data for each active channel is converted into the form necessary for data analysis (this step takes advantage of cluster parallelism, splitting tasks by data file), and (3) the data analysis including spike sorting, and other custom analysis tasks, is performed in parallel by an active channel.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

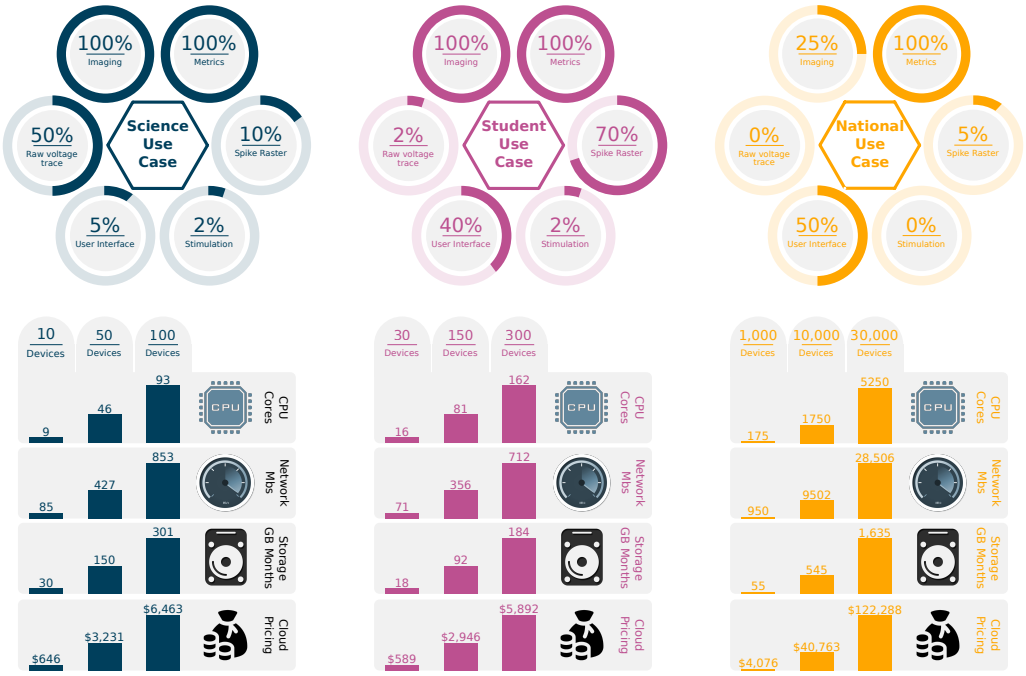


Figure 7: Resource utilization requirements given three use cases: Science, Student, and National scales. The assumed distribution of device functions under each use case is displayed above. Resource utilization for CPU, Network, and Storage are displayed below. An estimate of cloud compute costs are provided on the bottom. The number of active devices varies from fewer in the Science use to many in the National use. We define "% imaging," as the percentage of devices actively recording and storing microscopy images; "% metrics," as the percentage of devices actively recording measurements such as media concentrations and temperatures; "% raw voltage," as the percentage of devices recording and storing full raw voltage traces across all electrophysiology channels; "% spike raster," as the percentage of devices registering only neural spikes events (estimated to be 10% of the raw voltage data); "% UI," as the number of active users on the web interface relative to the total number of devices; and "% stimulation," as the percentage of devices that are actively executing electrode stimulation requests.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7. STAR Methods

7.1. Key Resource Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Python	G. van Rossum, Python tutorial, Technical Report CS-R9526, May 1995.	python.org
C Language	Kernighan, B. W., & Ritchie, D. M. (2006). The C programming language.	open-std.org/jtc1/sc22/wg14/
Tensorflow	10.5281/zenodo.4724125	tensorflow.org
Nextflow	Di Tommaso, P., Chatzou, M., Floden, E. et al., 2017	nextflow.io
Dockstore	O'Connor BD, Yuen D, Chung V et al.	dockstore.org
Redis	Redis Labs	redis.io
Ceph	Sage A. Weil, et al. 2006.	ceph.io
MQTT	MQTT Version 5.0. OASIS. 2019-03-07. Retrieved 2020-12-15.	mqtt.org
AWS IoT	Amazon Web Services, Inc.	aws.amazon.com/iot
Jupyter Notebook	Kluyver et al., 2016	jupyter.org
Docker	Docker, Inc.	docker.com
Kubernetes	The Linux Foundation	kubernetes.io
Pacific Research Platform	San Diego Supercomputer Center	pacificresearchplatform.org
Braingeneers Project	This paper; Braingeneers	braingeneers.ucsc.edu
Braingeneers on PRP	This paper; Email Lead Contact to join	pacificresearchplatform.org /nautilus/namespaces/#braingeneers
BraingeneersPy	This paper; Data and communications utilities on PyPi	github.com /braingeneers/braingeneerspy pypi.org/project/braingeneers
Braingeneers Dashboard Demo	This paper; Braingeneers dashboard	braingeneers.gi.ucsc.edu/dashboard
Picroscope	Ly et al., 2021, Baudin et al., 2021	doi.org/10.1101/2021.05.17.443454, https://arxiv.org/abs/2106.07419
Piphys	Voitiuk et al., 2021	doi.org/10.1101/2021.05.18.444685

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7.2. Resource Availability

7.2.1. Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, David F. Parks (dfparks@ucsc.edu).

7.2.2. Materials availability

This study did not generate new unique reagents.

7.2.3. Data and code availability

All data reported in this paper will be shared by the lead contact upon request.

All original code has been deposited at https://github.com/braingeneers/braingeneerspy and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

7.3. Method Details

Pacific Research Platform

The Pacific Research Platform is an academic compute cluster run by the UCSD Supercomputer Center. It is described as a "potluck supercomputer," hosting and consolidating computing resource components of a large scale across many academic institutions. Compute resources are organized in public and private areas of the cluster. If a lab or institutions hosts compute resources on the cluster, they can choose to limit access to those resources, make them cluster-wide public, or make unused compute available for cluster-wide use. Most academic institutions can access the cluster's public resources via existing integration with CILogon. Institutions also benefit from hosting existing resources with the PRP cluster by taking advantage of the automated cluster administration and existing DevOps team. The cluster currently hosts over 7000 CPU cores, 500 GPUs, 35 TB of memory, and 2.5 PB of storage, as well as accelerator devices such as FPGAs.

Workflow

Docker, Kubernetes, Nextflow, and Dockstore are employed as workflow management tools. Docker enables a Container (effectively a virtual machine with less overhead) to be packaged with all software and configuration in a binary package that does not require re-installation to run on a new machine. Kubernetes is an industry scale scheduler for Docker Containers, which is hosted by the PRP. Nextflow provides standardized workflow definitions, which are stored and distributed on the Dockstore website.

Data Management

Large-scale data transfer and storage are achieved using a Redis server for real-time data transfer and Ceph distributed filesystem deployed on the PRP cluster. Redis is free open source software that enables distributed computing and high bandwidth data transfer. Ceph is a distributed filesystem that provides an S3 interface. The S3 interface is compatible with AWS/S3 but is not related to Amazon services. Tools such as the AWS command line tools and Rclone can be used to access Ceph via the same interface as defined by AWS. The reason an S3 filesystem is valuable is that it is a scalable

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

internet-accessible data store; therefore, data can be easily accessed by anyone with the appropriate credentials, within a lab, or between labs and institutions.

IoT

For IoT, Raspberry PI devices comprise most edge computing infrastructure, and they communicate over the MQTT protocol using the AWS IoT implementation of MQTT and related IoT services. Raspberry PI devices are registered on the AWS IOT service and communicate over the publish/subscribe MQTT protocol.

Frontend

On the frontend, Plotly/Dash is employed for real-time visualizations, and Jupyter Hub is employed for data analysis and data visualization tasks. Plotly/Dash is a web-hosted extension to the open-source Plotly plotting library, providing programmatic and real-time interaction with data for visualization purposes. Jupyter Hub provides a hosted Jupyter Notebook environment in which data analysis takes place. These services are hosted on a local lab server; they interact with the PRP/S3 for access to data and Redis and MQTT for communication with the cluster and real-time data.

References

Abe, T., Kinsella, I., Saxena, S., Paninski, L., & Cunningham, J. P. (). Neuroscience Cloud Analysis As a Service, . (p. 32).

Amstutz, P., Crusoe, M. R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S., & Stojanovic, L. (2016). Common Workflow Language, v1.0, . URL: https://www.research.manchester.ac.uk/portal/en/publications/common-workflow-language-v10(741919f5-d0ab-4557-9763-b811e911423b).html. doi:10.6084/m9.figshare.3115156.v2. Publisher: figshare.

Baudin, P. V., Ly, V. T., Pansodtee, P., Jung, E. A., Currie, R., Hoffman, R., Willsey, H. R., Pollen, A. A., Nowakowski, T. J., Haussler, D., Mostajo-Radji, M. A., Salama, S., & Teodorescu, M. (2021). Low cost cloud based remote microscopy for biological sciences. *arXiv:2106.07419 [cs, eess]*, . URL: http://arxiv.org/abs/2106.07419. ArXiv: 2106.07419.

Beattie, R., Hippenmeyer, S., & Pauler, F. M. (2020). SCOPES: Sparking Curiosity Through Open-Source Platforms in Education and Science. *Frontiers in Education*, *5*. URL: https://www.frontiersin.org/articles/10.3389/feduc.2020.00048/full. doi:10.3389/feduc.2020.00048. Publisher: Frontiers.

Beggs, J. M., & Plenz, D. (2003). Neuronal Avalanches in Neocortical Circuits. *Journal of Neuroscience*, *23*, 11167–11177. URL: https://www.jneurosci.org/content/23/35/11167. doi:10.1523/JNEUROSCI.23-35-11167.2003. Publisher: Society for Neuroscience Section: Behavioral/Systems/Cognitive.

Bera, T. K. (2014). Bioelectrical Impedance Methods for Noninvasive Health Monitoring: A Review. *Journal of Medical Engineering*, *2014*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782691/. doi:10.1155/2014/381251.

Chung, J. E., Magland, J. F., Barnett, A. H., Tolosa, V. M., Tooker, A. C., Lee, K. Y., Shah, K. G., Felix, S. H., Frank, L. M., & Greengard, L. F. (2017). A Fully Automated Approach to Spike Sorting. *Neuron*, *95*, 1381–1394.e6. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627317307456. doi:10.1016/j.neuron.2017.08.030.

Dhawale, A. K., Poddar, R., Wolff, S. B., Normand, V. A., Kopelowitz, E., & Ölveczky, B. P. (2017). Automated long-term recording and analysis of neural activity in behaving animals. *eLife*, *6*, e27702. URL: https://doi.org/10.7554/eLife.27702. doi:10.7554/eLife.27702. Publisher: eLife Sciences Publications, Ltd.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*, 316–319. URL: https://www.nature.com/articles/nbt.3820/. doi:10.1038/nbt.3820. Number: 4 Publisher: Nature Publishing Group.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dutta, D., Heo, I., & Clevers, H. (2017). Disease Modeling in Stem Cell-Derived 3D Organoid Systems. *Trends in Molecular Medicine*, *23*, 393–410. URL: https://www.sciencedirect.com/science/article/pii/S1471491417300278. doi:10.1016/j.molmed.2017.02.007.

Funamoto, K., K. Zervantonakis, I., Liu, Y., J. Ochs, C., Kim, C., & D. Kamm, R. (2012). A novel microfluidic platform for high-resolution imaging of a three-dimensional cell culture under a controlled hypoxic environment. *Lab on a Chip*, *12*, 4855–4863. URL: https://pubs.rsc.org/en/content/articlelanding/2012/lc/c2lc40306d. doi:10.1039/C2LC40306D. Publisher: Royal Society of Chemistry.

Furusawa, C., & Kaneko, K. (2012). A Dynamical-Systems View of Stem Cell Biology. *Science*, *338*, 215–217. URL: https://science.sciencemag.org/content/338/6104/215. doi:10.1126/science.1224311. Publisher: American Association for the Advancement of Science Section: Perspective.

Gamal, W., Wu, H., Underwood, I., Jia, J., Smith, S., & Bagnaninchi, P. O. (2018). Impedance-based cellular assays for regenerative medicine. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*, 20170226. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2017.0226. doi:10.1098/rstb.2017.0226. Publisher: Royal Society.

Gamma, E. (1995). *Design patterns : elements of reusable object-oriented software*. Reading, Mass. : Addison-Wesley. URL: http://archive.org/details/designpatternsel00gamm.

Gao, D., Vela, I., Sboner, A., Iaquinta, P. J., Karthaus, W. R., Gopalan, A., Dowling, C., Wanjala, J. N., Undvall, E. A., Arora, V. K., Wongvipat, J., Kossai, M., Ramazanoglu, S., Barboza, L. P., Di, W., Cao, Z., Zhang, Q. F., Sirota, I., Ran, L., MacDonald, T. Y., Beltran, H., Mosquera, J.-M., Touijer, K. A., Scardino, P. T., Laudone, V. P., Curtis, K. R., Rathkopf, D. E., Morris, M. J., Danila, D. C., Slovin, S. F., Solomon, S. B., Eastham, J. A., Chi, P., Carver, B., Rubin, M. A., Scher, H. I., Clevers, H., Sawyers, C. L., & Chen, Y. (2014). Organoid Cultures Derived from Patients with Advanced Prostate Cancer. *Cell*, *159*, 176 – 187. URL: http://www.sciencedirect.com/science/article/pii/S0092867414010472. doi:https://doi.org/10.1016/j.cell.2014.08.016.

Gerasimenko, T., Nikulin, S., Zakharova, G., Poloznikov, A., Petrov, V., Baranova, A., & Tonevitsky, A. (2020). Impedance Spectroscopy as a Tool for Monitoring Performance in 3D Models of Epithelial Tissues. *Frontiers in Bioengineering and Biotechnology*, *7*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6992543/. doi:10.3389/fbioe.2019.00474.

Hansel, D., Mato, G., & Meunier, C. (1995). Synchrony in Excitatory Neural Networks. *Neural Computation*, *7*, 307–337. URL: https://doi.org/10.1162/neco.1995.7.2.307. doi:10.1162/neco.1995.7.2.307. Publisher: MIT Press.

Hasham, K., Munir, K., & McClatchey, R. (2018). Cloud infrastructure provenance collection and management to reproduce scientific workflows execution. *Future Generation Computer Systems*, *86*, 799–820. URL: http://www.sciencedirect.com/science/article/pii/S0167739X17314917. doi:10.1016/j.future.2017.07.015.

Hossain, Z., Bumbacher, E. W., Chung, A. M., Kim, H., Litton, C., Walter, A. D., Pradhan, S. N., Jona, K., Blikstein, P., & Riedel-Kruse, I. H. (2016). Interactive and scalable biology cloud experimentation for scientific inquiry and education. *Nature Biotechnology*, *34*, 1293–1298. URL: https://www.nature.com/articles/nbt.3747. doi:10.1038/nbt.3747. Number: 12 Publisher: Nature Publishing Group.

Huang, Y., C. Williams, J., & M. Johnson, S. (2012). Brain slice on a chip: opportunities and challenges of applying microfluidic technology to intact tissues. *Lab on a Chip*, *12*, 2103–2117. URL: https://pubs.rsc.org/en/content/articlelanding/2012/lc/c2lc21142d. doi:10.1039/C2LC21142D. Publisher: Royal Society of Chemistry.

Huang, Y.-T., Chang, Y.-L., Chen, C.-C., Lai, P.-Y., & Chan, C. K. (2017). Positive feedback and synchronized bursts in neuronal cultures. *PLOS ONE*, *12*, e0187276. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187276. doi:10.1371/journal.pone.0187276. Publisher: Public Library of Science.

Hung, L.-H., Hu, J., Meiss, T., Ingersoll, A., Lloyd, W., Kristiyanto, D., Xiong, Y., Sobie, E., & Yeung, K. Y. (2019). Building Containerized Workflows Using the BioDepot-Workflow-Builder. *Cell Systems*, *9*, 508–514.e3. URL: https://www.sciencedirect.com/science/article/pii/S2405471219302765. doi:10.1016/j.cels.2019.08.007.

Ibrahim, M., Gorlatova, M., & Chakrabarty, K. (2019). The Internet of Microfluidic Things: Perspectives on System Architecture and Design Challenges: Invited Paper. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (pp. 1–8). doi:10.1109/ICCAD45719.2019.8942080 iSSN: 1558-2434.

Ionescu-Zanetti, C., Shaw, R. M., Seo, J., Jan, Y.-N., Jan, L. Y., & Lee, L. P. (2005). Mammalian electrophysiology on a microfluidic platform. *Proceedings of the National Academy of Sciences*, *102*, 9112–9117. URL: https://www.pnas.org/content/102/26/9112. doi:10.1073/pnas.0503418102.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ISBN: 9780503418106 Publisher: National Academy of Sciences Section: Physical Sciences.

Jansen, C., Annuscheit, J., Schilling, B., Strohmenger, K., Witt, M., Bartusch, F., Herta, C., Hufnagl, P., & Krefting, D. (2020). Curious Containers: A framework for computational reproducibility in life sciences with support for Deep Learning applications. *Future Generation Computer Systems*, *112*, 209–227. URL: https://www.sciencedirect.com/science/article/pii/S0167739X19318096. doi:10.1016/j.future.2020.05.007.

Jensen, E. C. (2013). Overview of Live-Cell Imaging: Requirements and Methods Used. *The Anatomical Record*, *296*, 1–8. URL: https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/ar.22554. doi:https://doi.org/10.1002/ar.22554. _eprint: https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/ar.22554.

Jung, D. J., Shin, T. H., Kim, M., Sung, C. O., Jang, S. J., & Jeong, G. S. (2019). A one-stop microfluidic-based lung cancer organoid culture platform for testing drug sensitivity. *Lab on a Chip*, *19*, 2854–2865. URL: https://pubs.rsc.org/en/content/articlelanding/2019/lc/c9lc00496c. doi:10.1039/C9LC00496C. Publisher: The Royal Society of Chemistry.

Khan, I., Prabhakar, A., Delepine, C., Tsang, H., Pham, V., & Sur, M. (2021). A low-cost 3D printed microfluidic bioreactor and imaging chamber for live-organoid imaging. *Biomicrofluidics*, *15*, 024105. URL: https://aip.scitation.org/doi/10.1063/5.0041027. doi:10.1063/5.0041027. Publisher: American Institute of Physics.

Kim, H., Gerber, L. C., Chiu, D., Lee, S. A., Cira, N. J., Xia, S. Y., & Riedel-Kruse, I. H. (2016). LudusScope: Accessible Interactive Smartphone Microscopy for Life-Science Education. *PLOS ONE*, *11*, e0162602. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0162602. doi:10.1371/journal.pone.0162602. Publisher: Public Library of Science.

Kim, J., Koo, B.-K., & Knoblich, J. A. (2020). Human organoids: model systems for human biology and medicine. *Nature Reviews Molecular Cell Biology*, *21*, 571–584. URL: https://www.nature.com/articles/s41580-020-0259-3. doi:10.1038/s41580-020-0259-3.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & development team, J. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. (pp. 87–90). IOS Press. URL: https://eprints.soton.ac.uk/403913/. doi:10.3233/978-1-61499-649-1-87.

Lancaster, M. A., Corsini, N. S., Wolfinger, S., Gustafson, E. H., Phillips, A., Burkard, T. R., Otani, T., Livesey, F. J., & Knoblich, J. A. (2017). Guided self-organization and cortical plate formation in human brain organoids. *Nature biotechnology*, *35*, 659–666. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5824977/. doi:10.1038/nbt.3906.

Lancaster, M. A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L. S., Hurles, M. E., Homfray, T., Penninger, J. M., Jackson, A. P., & Knoblich, J. A. (2013). Cerebral organoids model human brain development and microcephaly. *Nature*, *501*, 373–379. URL: https://www.nature.com/articles/nature12517. doi:10.1038/nature12517. Number: 7467 Publisher: Nature Publishing Group.

Lee, J., Mitelut, C., Shokri, H., Kinsella, I., Dethe, N., Wu, S., Li, K., Reyes, E. B., Turcu, D., Batty, E., Kim, Y. J., Brackbill, N., Kling, A., Goetz, G., Chichilnisky, E. J., Carlson, D., & Paninski, L. (2020). YASS: Yet Another Spike Sorter applied to large-scale multi-electrode array recordings in primate retina. *bioRxiv*, (p. 2020.03.18.997924). URL: https://www.biorxiv.org/content/10.1101/2020.03.18.997924v1. doi:10.1101/2020.03.18.997924. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Ly, V. T., Baudin, P. V., Pansodtee, P., Jung, E. A., Voitiuk, K., Rosen, Y., Willsey, H. R., Mantalas, G. L., Seiler, S. T., Selberg, J. A., Cordero, S. A., Ross, J. M., Pollen, A. A., Nowakowski, T. J., Haussler, D., Mostajo-Radji, M. A., Salama, S., & Teodorescu, M. (2021). Development of a Low-Cost System for Simultaneous Longitudinal Biological Imaging. *bioRxiv*, (p. 2021.05.17.443454). URL: https://www.biorxiv.org/content/10.1101/2021.05.17.443454v1. doi:10.1101/2021.05.17.443454. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Marcus, D. S., Olsen, T. R., Ramaratnam, M., & Buckner, R. L. (2007). The extensible neuroimaging archive toolkit. *Neuroinformatics*, *5*, 11–33. URL: https://doi.org/10.1385/NI:5:1:11. doi:10.1385/NI:5:1:11.

O'Connor, B. D., Yuen, D., Chung, V., Duncan, A. G., Liu, X. K., Patricia, J., Paten, B., Stein, L., & Ferretti, V. (2017). The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. *F1000Research*, *6*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333608/. doi:10.12688/f1000research.10137.1.

Okano, H. (2002). Stem cell biology of the central nervous system. *Journal of Neuroscience Research*, *69*, 698–707. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jnr.10343. doi:https://doi.org/10.1002/jnr.10343. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jnr.10343.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Paşca, A. M., Sloan, S. A., Clarke, L. E., Tian, Y., Makinson, C. D., Huber, N., Kim, C. H., Park, J.-Y., O'Rourke, N. A., Nguyen, K. D., Smith, S. J., Huguenard, J. R., Geschwind, D. H., Barres, B. A., & Paşca, S. P. (2015). Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nature Methods*, *12*, 671–678. doi:10.1038/nmeth.3415.

Poli, D., Pastore, V. P., & Massobrio, P. (2015). Functional connectivity in in vitro neuronal assemblies. *Frontiers in Neural Circuits*, *9*. URL: https://www.frontiersin.org/articles/10.3389/fncir.2015.00057/full. doi:10.3389/fncir.2015.00057. Publisher: Frontiers.

Putzeys, J., Raducanu, B. C., Carton, A., De Ceulaer, J., Karsh, B., Siegle, J. H., Van Helleputte, N., Harris, T. D., Dutta, B., Musa, S., & Mora Lopez, C. (2019). Neuropixels Data-Acquisition System: A Scalable Platform for Parallel Recording of 10 000+ Electrophysiological Signals. *IEEE Transactions on Biomedical Circuits and Systems*, *13*, 1635–1644. doi:10.1109/TBCAS.2019.2943077.

Raghu, M., & Schmidt, E. (2020). A Survey of Deep Learning for Scientific Discovery. *arXiv:2003.11755 [cs, stat]*, . URL: http://arxiv.org/abs/2003.11755. ArXiv: 2003.11755.

Rolston, J. D., Gross, R. E., & Potter, S. M. (2009). NeuroRighter: Closed-loop multielectrode stimulation and recording for freely moving animals and cell cultures. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6489–6492). doi:10.1109/IEMBS.2009.5333589 iSSN: 1558-4615.

Rossi, G., Manfrin, A., & Lutolf, M. P. (2018). Progress and potential in organoid research. *Nature Reviews Genetics*, *19*, 671–687. URL: https://www.nature.com/articles/s41576-018-0051-9. doi:10.1038/s41576-018-0051-9.

Sarrafha, L., Parfitt, G. M., Reyes, R., Goldman, C., Coccia, E., Kareva, T., & Ahfeldt, T. (2021). High-throughput generation of midbrain dopaminergic neuron organoids from reporter human pluripotent stem cells. *STAR Protocols*, *2*, 100463. URL: https://www.sciencedirect.com/science/article/pii/S2666166721001702. doi:10.1016/j.xpro.2021.100463.

Schuster, B., Junkin, M., Kashaf, S. S., Romero-Calvo, I., Kirby, K., Matthews, J., Weber, C. R., Rzhetsky, A., White, K. P., & Tay, S. (2020). Automated microfluidic platform for dynamic and combinatorial drug screening of tumor organoids. *Nature Communications*, *11*, 5271. URL: https://www.nature.com/articles/s41467-020-19058-4. doi:10.1038/s41467-020-19058-4. Number: 1 Publisher: Nature Publishing Group.

Schutgens, F., & Clevers, H. (2020). Human Organoids: Tools for Understanding Biology and Treating Diseases. *Annual Review of Pathology: Mechanisms of Disease*, *15*, 211–234. URL: https://www.annualreviews.org/doi/10.1146/annurev-pathmechdis-012419-032611. doi:10.1146/annurev-pathmechdis-012419-032611.

Sharf, T., Molen, T. v. d., Guzman, E., Glasauer, S. M. K., Luna, G., Cheng, Z., Audouard, M., Ranasinghe, K. G., Kudo, K., Nagarajan, S. S., Tovar, K. R., Petzold, L. R., Hansma, P. K., & Kosik, K. S. (2021). Intrinsic network activity in human brain organoids. *bioRxiv*, (p. 2021.01.28.428643). URL: https://www.biorxiv.org/content/10.1101/2021.01.28.428643v1. doi:10.1101/2021.01.28.428643. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Siegle, J. H., Hale, G. J., Newman, J. P., & Voigts, J. (2015). Neural ensemble communities: open-source approaches to hardware for large-scale electrophysiology. *Current Opinion in Neurobiology*, *32*, 53–59. URL: http://www.sciencedirect.com/science/article/pii/S0959438814002268. doi:10.1016/j.conb.2014.11.004.

Smarr, L., Crittenden, C., DeFanti, T., Graham, J., Mishin, D., Moore, R., Papadopoulos, P., & Würthwein, F. (2018). The Pacific Research Platform: Making High-Speed Networking a Reality for the Scientist. In *Proceedings of the Practice and Experience on Advanced Research Computing* PEARC '18 (pp. 1–8). Pittsburgh, PA, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3219104.3219108. doi:10.1145/3219104.3219108.

Stephens, D. J., & Allan, V. J. (2003). Light Microscopy Techniques for Live Cell Imaging. *Science*, *300*, 82–86. URL: https://science.sciencemag.org/content/300/5616/82. doi:10.1126/science.1082160. Publisher: American Association for the Advancement of Science Section: Special Reviews.

Széll, A., Martínez-Bellver, S., Hegedüs, P., & Hangya, B. (2020). OPETH: Open Source Solution for Real-Time Peri-Event Time Histogram Based on Open Ephys. *Frontiers in Neuroinformatics*, *14*. URL: https://www.frontiersin.org/articles/10.3389/fninf.2020.00021/full. doi:10.3389/fninf.2020.00021.

Voitiuk, K., Geng, J., Keefe, M. G., Parks, D. F., Sanso, S. E., Hawthorne, N., Freeman, D. B., Mostajo-Radji, M. A., Nowakowski, T. J., Salama, S. R., Teodorescu, M., & Haussler, D. (2021). Light-weight Electrophysiology Hardware and Software Platform for Cloud-Based Neural Recording Experiments. *bioRxiv*, (p. 2021.05.18.444685). URL: https://www.biorxiv.org/content/10.1101/2021.05.18.444685v2. doi:10.1101/2021.05.18.444685. Publisher: Cold Spring Harbor Laboratory

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.29.453595; this version posted August 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Section: New Results.

Wagenaar, D., DeMarse, T., & Potter, S. (2005). MeaBench: A toolset for multi-electrode data acquisition and on-line analysis. In *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005.* (pp. 518–521). doi:10.1109/CNE.2005.1419673 iSSN: 1948-3554.

Weil, S., Brandt, S. A., Miller, E. L., Long, D. D. E., & Maltzahn, C. (2006). Ceph: A Scalable, High-Performance Distributed File System. In *Proceedings of the 7th Conference on Operating Systems Design and Implementation (OSDI '06)*.

Yger, P., Spampinato, G. L., Esposito, E., Lefebvre, B., Deny, S., Gardella, C., Stimberg, M., Jetter, F., Zeck, G., Picaud, S., Duebel, J., & Marre, O. (2018). A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. *eLife*, *7*, e34518. URL: https://doi.org/10.7554/eLife.34518. doi:10.7554/eLife.34518. Publisher: eLife Sciences Publications, Ltd.

Yuan, Y., Feng, S., Alahi, M. E. E., Nag, A., Afsarimanesh, N., Zhang, H., & He, S. (2018). Development of an Internet of Things Based Electrochemical Microfluidic System for Free Calcium Detection. *Applied Sciences*, *8*, 1357. URL: https://www.mdpi.com/2076-3417/8/8/1357. doi:10.3390/app8081357. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.