1 Abstract

Explain briefly the paper and what it does.

2 Introduction

Scientific Workflow Management Systems (SWMSs) are an essential tool for automating, managing, and executing complex scientific processes involving large volumes of data and computational tasks[1]. Traditional SWMSs employ a linear sequential approach, in which tasks are performed in a pre-defined order, as defined by the workflow. While this linear method is suitable for certain applications, it might not always be the best choice: processing sequentially can prove inefficient in cases where the next step of the process should adapt to the previous one. For these use-cases a dynamic scheduler is required, of which *Managing Event Oriented Workflows*[2] (MEOW) is one.

Expand on DAGs' inability to adapt. Plagiarize David's thesis.

MEOW employs an event-based scheduler, in which jobs are performed non-linearly (**Better word here**), triggered based on events[2]. By dynamically adapting the execution order based on the outcomes of previous tasks or external factors, MEOW provides a more efficient and flexible solution for processing large volumes of experimental data[3].

- Expand on what "efficient" is
- What work am I doing on MEOW?
- How did it go?
- Introduce the concept of network events.
- **Write this last**

2.1 Problem

In its current implementation, MEOW is able to trigger jobs based on changes to monitored local files. This covers a the range of scenarios where the data processing workflow involves the creation, modification, or removal of files. By monitoring file events, MEOW's event-based scheduler can dynamically execute tasks as soon as the required conditions are met, ensuring efficient and timely processing of the data. Since the file monitor is triggered by changes to local files, MEOW is limited to local workflows.

[1] citation?
[2] citation?
[3] citation?

While file events work well as a trigger on their own, there are several scenarios where a different trigger would be preferred or even required, especially when dealing with distributed systems or remote operations. To address these shortcomings and further enhance MEOW's capabilities, the integration of network event triggers would provide significant benefits in several key use-cases.

Firstly, network event triggers would allow for manual triggering of jobs remotely, without the need for direct access to the monitored files. This is particularly useful in human-in-the-loop scenarios, where human intervention or decision-making is required before proceeding with the subsequent steps in a workflow. While it is possible to manually trigger job using file events by making changes to the monitored directories, this might lead to an already running job accessing the files at the same time, which could cause problems with data integrity.

Secondly, incorporating network event triggers would facilitate seamless communication between parallel runners, ensuring that tasks can efficiently exchange information and updates on their progress, allowing for a better perspective on the whole workflow, greatly improving visibility and control.

Finally, extending MEOW's event-based scheduler to support network event triggers would enable the simple and efficient exchange of data between workflows running on different machines. This feature is particularly valuable in distributed computing environments, where data processing tasks are often split across multiple systems to maximize resource utilization and minimize latency.

Integrating network event triggers into MEOW would provide an advantage specifically in the context of heterogeneous workflows, which incorporate a mix of different tasks running on diverse computing environments. By their nature, these workflows can involve tasks running on different systems, potentially even in different physical locations, which need to exchange data or coordinate their progress. In the figure below, an example heterogeneous workflow is presented.

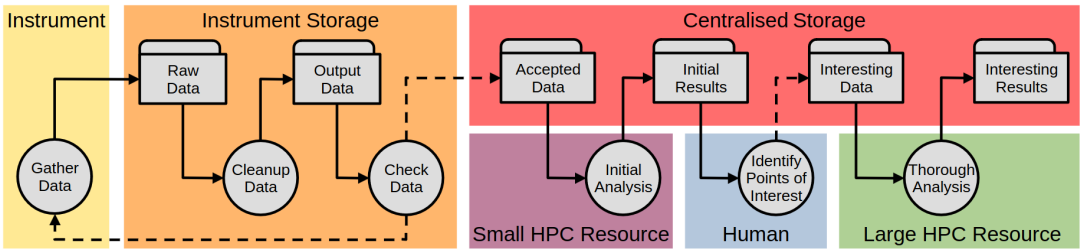

Figure 1: An example of a heterogeneous workflow

The example workflow requires several "halting-points", in which data should be transferred between the instrument, the instrument storage, centralized storage, High Performance Computing (HPC) resources, and a human interaction point. Network events can, for the reasons outlined earlier in the section, be used to prevent the workflow from halting when these points are reached.

2.2 Background

2.2.1 The structure of MEOW

The MEOW event-based scheduler consists of four main components: *monitors*, *handlers*, *the conductor*, and *the runner*.

Monitors listen for triggering events. They are initialized with a number of *patterns*, which describe the triggering event. When a pattern's triggering event occurs, the monitor signals to the conductor that the pattern has been triggered, and schedules a job that has been associated with the pattern.

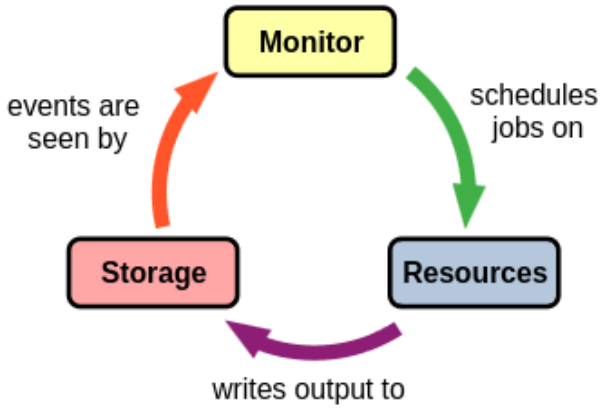

Figure 2: **Redo this to fit with the current version.** The monitor's role in MEOW's event-based system.

I haven't used "Resources" to describe the job queue. Should I do that or should I rephrase the diagram to be more in line with the rest of the project?

Handlers perform actions and jobs on behalf of the scheduler. They are initialized with a number of *recipes*, which describe the action to be taken. The handler starts a job when signal to do so by the conductor.

The conductor handles the jobs queue. It is initialized with a number of rules, which a pattern paired with a recipe. When a monitor sends it a triggered pattern, the rules are checked for that pattern. If one or more rules contain that pattern, the corresponding recipes are triggered in their handler.

Finally, the runner is the main program that orchestrates all these components. Each instance of the runner incorporates at least one instance of a monitor, handler, and conductor.

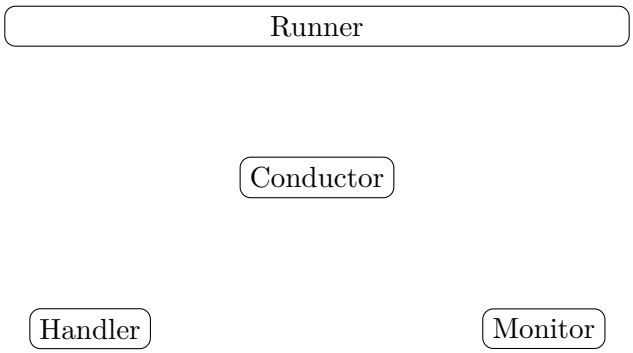

Figure 3: **WIP.** How the elements of MEOW interact.

2.2.2 The `meow_base` codebase

`meow_base`[3] is an implementation of MEOW written in python. It is written to be modular, using base classes for each element in order to ease the implementation of additional handlers, monitors, etc.

How much should I include here?

- The runner (brief)
- Conductors (brief)
- Recipes and handlers (brief)
- File event monitor (Watchdog)
- Events (important to clarify how file events work since I refer to it in the method section)
- Testing (brief)

2.2.3 The `socket` library

The `socket` library[1], included in the Python Standard Library, serves as an interface for the Berkeley sockets API. The Berkeley sockets API, originally developed for the Unix operating system, has become the standard for network communication across multiple platforms. It allows programs to create 'sockets', which are endpoints in a network communication path, for the purpose of sending and receiving data.

Many other libraries and modules focusing on transferring data exist for Python, some of which may be better in certain MEOW use-cases. The `ssl` library, in specific, allows for ssl-encrypted communication, which may be a requirement in workflows with sensitive data. However, implementing network triggers using the `socket` library will provide MEOW with a fundamental implementation of network events, which can later be expanded or improved with other features.

In my project, all sockets use the Transmission Control Protocol (TCP), which ensures safe data transfer by enforcing a stable connection between the sender and receiver.

I make use of the following socket methods, which have the same names and functions in the `socket` library and the Berkeley sockets API:

- `bind()`: Associates the socket with a given local IP address and port. It also reserves the port locally.
- `listen()`: Puts the socket in a listening state, where it waits for a sender to request a TCP connection to the socket.
- `accept()`: Accepts the incoming TCP connection request, creating a connection.
- `recv()`: Receives data from the given socket.
- `close()`: Closes a connection to a given socket.

During testing of the monitor, the following methods are used to send data to the running monitor:

- `connect()`: Sends a TCP connection request to a listening socket.
- `sendall()`: Sends data to a socket.

3 Method

To address the identified limitations of MEOW and to expand its capabilities, I will be incorporating network event triggers into the existing event-based scheduler, to supplement the current file-based event triggers. My method focuses on leveraging Python's socket library to enable the processing of network events. The following subsections detail the specific methodologies employed in expanding the codebase, the design of the network event trigger mechanism, and the integration of this mechanism into the existing MEOW system.

3.1 Design of the network event pattern

In the implementation of a pattern for network events, a key consideration was to integrate it seamlessly with the existing MEOW codebase. This required designing the pattern to behave similarly to the file event pattern when interacting with other elements of the scheduler. A central principle in this design was maintaining the loose coupling between patterns and recipes, minimizing direct dependencies between separate components. While this might not be possible for every theoretical recipe and pattern, designing for it could greatly improve future compatibility.

Network event patterns are initialized with a triggering port, analogous to the triggering path used in file event patterns. This approach inherently limits the number of unique patterns to the number of ports that can be opened on the machine. However, given the large number of potential ports, this constraint is unlikely to present a practical issue. An alternative approach could have involved triggering patterns using a part of the sent message, essentially acting as a "header". However, this would complicate the process since the monitor is otherwise designed to receive raw data.

To keep the implementation as straightforward as possible and to allow for future enhancements, I opted for simplicity over complexity in this initial design.

Once the network monitor is started, it opens sockets that start listening on the each of the ports specified in the patterns it was initialized with. This is consistent with the behavior of the file event monitor, which monitors the triggering paths of the patterns it was initialized with.

3.2 Integrating network events into the existing codebase

The data received by the network monitor is written to a temporary file, a design choice that serves two purposes.

Firstly, this method is a practical solution for managing memory usage during data transfer, particularly for large data sets. By writing received data directly to a file, we bypass the need to store the entire file in memory at once, effectively addressing potential memory limitations.

Secondly, this approach allows the leveraging of existing infrastructure built for file events. The newly written temporary file is passed as the ”triggering path” of the event, mirroring the behavior of file events. This approach allows network events to utilize the recipes initially designed for file events without modification, preserving the principle of loose coupling. This integration maintains the overall flexibility and efficiency of MEOW while extending its capabilities to handle network events.

3.3 Testing

4 Results

Does it work? How well?

4.1 Discussion

With the hindsight of the results, what could I have done better?

4.2 Future Work

What should someone do if they want to fix my mistakes, or expand on them further.

- Implementation of the other options mentioned when discussing the socket library.
- Triggering on a header item in addition to port

Give context to following paragraph.

One specific example of a use-case where network event triggers could prove useful is the workflow for The Brain Imaging Data Structure (BIDS). The BIDS workflow requires data to be sent between multiple machines and validated by a user. Network event triggers could streamline this process by automatically initiating data transfer tasks when specific conditions are met, thereby reducing the need for manual management. Additionally, network triggers could facilitate user validation by allowing users to manually prompt the continuation of the workflow through specific network requests, simplifying the user's role in the validation process.

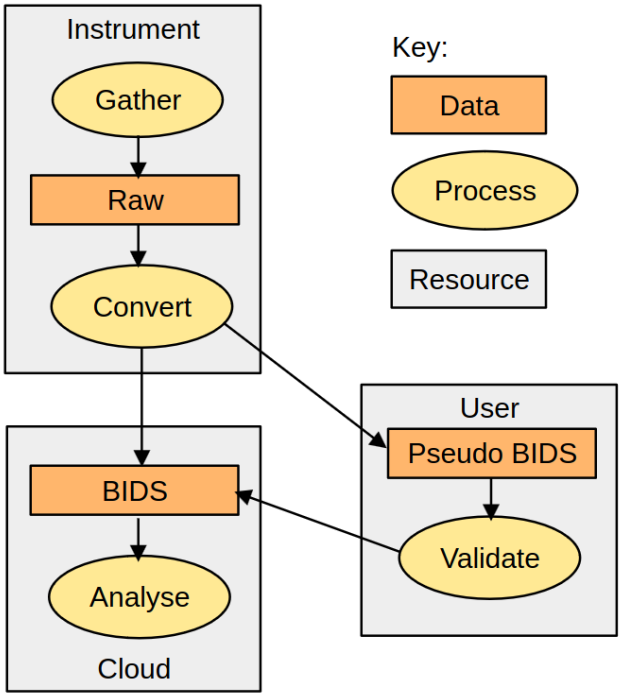

Figure 4: **Temp.** The structure of the BIDS workflow. Data is transferred to user, and to the cloud.

5 Conclusion

Did I succeed in what I wanted to do?

References

[1] Python documentation. *socket - Low-level networking interface*. https://docs.python.org/3/library/socket.html.

[2] David Marchant. "MEOW - Enabling Dynamic Scheduling of Scientific Analysis". PhD thesis. University of Copenhagen, May 2021.

[3] David Marchant. *meow_base*. https://github.com/PatchOfScotland/meow_base. 2023.