

Orchestrating Scientific Workflows with Maestro Workflow Conductor https://github.com/LLNL/maestrowf

Francesco Di Natale (LLNL)

Workflow Challenges

- Running HPC/software workflows is complex and error prone.
- External solutions usually employ a wide range of technologies that violate security policies.
- Users often dislike fighting with tools and "black boxes".
- Other tools assume specifics, making it harder for users to develop their own methodologies.
- Using ML in current workflows is extremely difficult.

A Standard Workflow Experience

- Provide the means to facilitate an agile and repeatable approach to scientific workflows.
- Allowing users to focus on exploring the science, not process.
- Workflows should be represented in a manner that is clear, concise, and natural.
- A workflow tool should not assume responsibility for any specific methodology.

Philosophy and Core Principles



Maestro's philosophy is centered around three key principles: 1) documentation of workflows, 2) consistency in communication and representation, and 3) easy repeatability of whole workflows. All three principles are perquisites to reproducibility.

Maestro improves productivity by allowing users to focus on science.



Maestro Study Specifications

Maestro makes use of a markup file called a "study specification". A study contains all the fixed variables, steps, and parameters for running a study. The specification is analogous to the workflow for a physical experiment documented in a lab notebook.



Maestro's study specification is both machine- and human-readable and is analogous to a page in a laboratory notebook that would be used to document a physical experiment.

Automation Made Easy

<pre>vescription: · · · · name: exp_study · · · · description: A simple sample study.</pre>	Variables and Labels			
env: ····variables: ·····OUTPUT_PATH:·./exp_study	Batch Settings	$S_1(P) \longrightarrow S_2(P)$	\rightarrow S ₃ (P [*]) \rightarrow S ₄ (P)	
study: 	Dependencies	Steps		
run: cmd: echo "\$(P)"	Study			
<pre>description: Run step2 description:</pre>				
echo "\$(P)" depends: [s1] - name: s3 description: Run step3 run:	$p_1 \rightarrow s_1$	$_{1}(p_{1}) \longrightarrow S_{2}(p_{1})$	S4(p1)	
<pre>cmd: echo "s3" depends: [s2_*] name: s4 description: Run step4 run:</pre>	arameters ∶	$(p_2) \longrightarrow S_2(p_2) \longrightarrow S_2(p_2)$	$S_3(h-h) \longrightarrow S_4(p_2)$	
global.parameters:	$\begin{array}{c} \square & \cdot \\ & &$	$(p_n) \longrightarrow S_2(p_n)$	S ₄ (p _n)	
P: values: [p_1, p_2, p_3,, p_n] label: p.%	Execution Graph			
	Maestro	¥		Co

Maestro parses the specification and constructs internal representations for the data contained in the study specification. Maestro is then capable of cleverly expanding the workflow into a DAG, which is then automatically monitored and executed on compute resources.





Future Direction and Vision



 Study concept can be used to generalize workflow definitions for other tools to perform UQ, optimization, ML sampling, etc. Additional functionality to add parameters to existing studies, restart failed steps, add steps.

A study is a template for how to run a workflow once and has mechanisms to parameterize them. At a higher level, if we consider a study as a unit, then we can plan workflows around them. This concept means that common problems such as optimization, UQ, and machine learned decision making can influence whether a study needs to be run over more parameters.

Collaborators & Users RECEIVENT OF THE RECEIPTION OF

Enabling Scientific Research

Maestro provides a general and simple way to define and automate computational science workflows. It's helped users achieve multiple high-level strategic initiatives at LLNL and is open sourced for wider community use. Maestro user testimonials:

- "Maestro has made me simulation data-rich for the first time."
- "Maestro allowed me to focus on the engineering aspects rather than the minutia of managing the ensemble of test cases. Its structured output/workspaces also facilitated automating the post-simulation workflow to get the engineering work done faster and to quickly ask new questions as the dataset was explored."



