UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

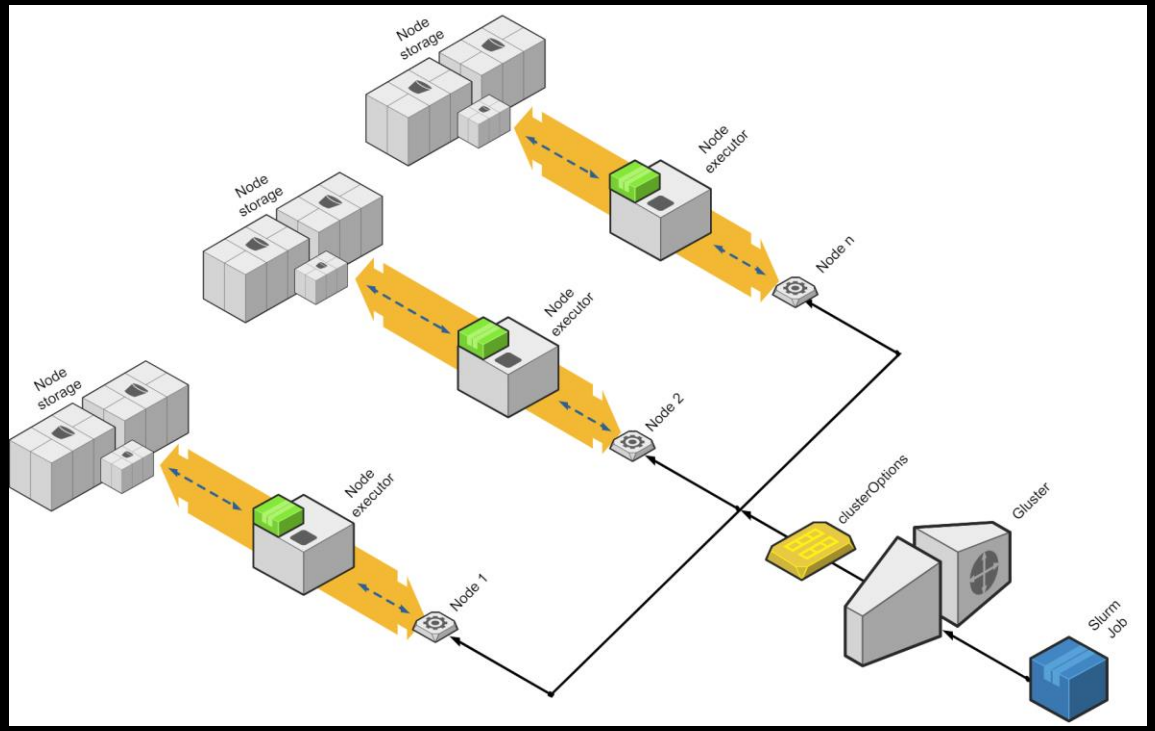
22P63

Location Aware Scientific Workflows

Tristan Lilford & Robin Jonker

SCHOOL OF ELECTRICAL AND INFORMATION ENGINEERING

Nextflow library to enhance workflow execution performance based on locality of the principal data required.



The Nextflow Library design greatly improves the performances of systems utilising a gluster filesystem and a Slurm scheduler. The library identifies on which nodes the data is stored and then determines where best to execute the workflow. If the execution runs where the data is stored, data transfer is reduced, and overall execution time improves. Ask us how we determine the best execution location.

Library Type	Average Execution Time (s)	Difference
Simple	769.364	Executes on available nodes regardless of where the data is stored.
Static	597.063	Sets executer nodes before any processes begin to run. Does not update node selection during runtime.
Dynamic	429.9	Sets executer nodes as the different processes begin, dynamically updating the node selection for each input file

The system was tested whilst the cluster was on both high and low demand. The results are as expected. Under low demand, Static performs best, as node selection is not contested by other jobs, however, under heavy load, Dynamic can adjust its location when nodes are allocated by other jobs.