International Symposium on Grids & Clouds 2018 (ISGC 2018) in conjunction with Frontiers in Computational Drug Discovery (FCDD)

Contribution ID: **6** Type: **Oral Presentation**

Data intensive ATLAS workflows in the Cloud

Friday, 23 March 2018 10:50 (30 minutes)

From 2025 onwards, the ATLAS collaboration at the Large Hadron Collider (LHC) at CERN will experience a massive increase in data quantity as well as complexity (High-Luminosity LHC). Including mitigating factors, the prevalent computing power by that time will only fulfil one tenth of the requirement.

This contribution will focus on Cloud computing as an approach to help overcome this challenge by providing flexible hardware that can be configured to the specific needs of a workflow. Experience with Cloud computing exists, but there is a large uncertainty if and to which degree it can be able to reduce the burden by 2025. In order to understand and quantify the benefits of Cloud computing, the "Workflow and Infrastructure Model" was created. It estimates the viability of Cloud computing by combining different inputs from the workflow side with infrastructure specifications. The model delivers metrics that enable the comparison of different Cloud configurations as well as different Cloud offerings with each other. A wide range of results can be obtained - from the required bandwidth over the workflow duration to the cost per workflow - making the model useful for fields outside of physics as well. In the High Energy Physics (HEP) use case, a workload is quantifiable by individual bunch crossings within the detector ('events'). A powerful metric that can be derived from that is EC = 'Events per Cost'. Comparing EC values with each other immediately points to the best Cloud offering for HEP workflows, maximising the physics throughput while minimising the cost.

Instead of using generic benchmarks, the model uses reference workflows in order to obtain infrastructure parameters. The workflow parameters are obtained from running the workflow on a reference machine. The model linearly combines different job aspects such as the machine specific CPU time in order to get the results for one workflow on one machine, which is then extrapolated to a whole Cloud infrastructure. Limiting factors to the predictions of the model are therefore fluctuations within the workflows (varying data complexity, software updates) as well as within the infrastructure ("noisy neighbours").

Finally the usefulness and accuracy of the model will be demonstrated by the real-world experience gathered during the latest CERN Cloud procurement, which included several commercial Cloud providers. The results encompass recommendations regarding the desirability to commission storage in the Cloud, in conjunction with a simple analytical model of the system, and correlated with questions about the network bandwidth and the type of storage to utilise.

Primary author: Mr RZEHORZ, Gerhard (CERN, University of Göttingen)

Co-authors: Prof. QUADT, Arnulf (University of Göttingen); Dr KAWAMURA, Gen (University of Göttingen); Dr KEEBLE, Oliver (CERN)

Presenter: Mr RZEHORZ, Gerhard (CERN, University of Göttingen)

Session Classification: Infrastructure Clouds & Virtualisation Session

Track Classification: Infrastructure Clouds and Virtualisation