Contribution ID: **229** Type: **Oral**

Raythena: a vertically integrated scheduler for ATLAS applications on heterogeneous distributed resources

Tuesday, 5 November 2019 12:00 (15 minutes)

The ATLAS experiment has successfully integrated High-Performance Computing (HPC) resources in its production system. Unlike the current generation of HPC systems, and the LHC computing grid, the next generation of supercomputers is expected to be extremely heterogeneous in nature: different systems will have radically different architectures, and most of them will provide partitions optimized for different kinds of workloads. In this work we explore the applicability of concepts and tools realized in Ray (the high-performance distributed execution framework targeting large-scale machine learning applications) to ATLAS event throughput optimization on heterogeneous distributed resources, ranging from traditional grid clusters to Exascale computers.

We present a prototype of Raythena, a Ray-based implementation of the ATLAS Event Service (AES), a fine-grained event processing workflow aimed at improving the efficiency of ATLAS workflows on opportunistic resources, specifically HPCs. The AES is implemented as an event processing task farm that distributes packets of events to several worker processes running on multiple nodes. Each worker in the task farm runs an event-processing application (Athena) as a daemon. In Raythena we replaced the event task farm workers with stateful components of Ray called Actors, which process packets of events and return data processing results. In addition to stateful Actors, Raythena also utilizes stateless Tasks for merging intermediate outputs produced by the Actors. The whole system is orchestrated by Ray, which assigns work to Actors and Tasks in a distributed, possibly heterogeneous, environment.

The second thrust of this study is to use Raythena to schedule Gaudi Algorithms (the primary unit of work of ATLAS' Athena framework) across a set of heterogeneous nodes. For ease of testing, we have used the Gaudi execution flow simulator to run a production ATLAS reconstruction scenario consisting of 309 Algorithms, modeled by synthetic CPU burners constrained by the data dependencies, and run for the time duration of the original Algorithms. The Algorithms are wrapped in Ray Actors or Tasks, and communicate via the Ray Global Control Store. This approach allows the processing of a single event to be distributed across more than one node, a functionality currently not supported by the Athena framework. We will discuss Raythena features and performance as a scheduler for ATLAS workflows, comparing them to those offered by Athena.

For all its flexibility, the AES implementation is currently comprised of multiple separate layers that communicate through ad-hoc command-line and file-based interfaces. The goal of Raythena is to integrate these layers through a feature-rich, efficient application framework. Besides increasing usability and robustness, a vertically integrated scheduler will enable us to explore advanced concepts such as dynamically shaping of workflows to exploit currently available resources, particularly on heterogeneous systems.

Consider for promotion

No

Primary authors: MUSKINJA, Miha (Lawrence Berkeley National Lab. (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US)); SHAPOVAL, Illya (Lawrence Berkeley National Laboratory); TSULAIA, Vakho (Lawrence Berkeley National Lab. (US))

Presenter: MUSKINJA, Miha (Lawrence Berkeley National Lab. (US))

Session Classification: Track 5 –Software Development

Track Classification: Track 5 –Software Development