Contribution ID: **61** Type: **presentation**

The Design of a Distributed Key-Value Store for Petascale Hot Storage in Data Acquisition Systems

Thursday, 12 July 2018 11:45 (15 minutes)

Data acquisition (DAQ) systems for high energy physics experiments readout data from a large number of electronic components, typically over thousands of point to point links. They are thus inherently distributed systems. Traditionally, an important stage in the data acquisition chain has always been the so called *event building*: data fragments coming from different sensors are identified as belonging to the same physical *event* and are physically assembled in a single memory location on one host. The complete events cached on *event builder nodes* are then served to farms of processors for analysis and filtering of data. In this work we propose a new approach - logical event building with hot storage. Data fragments are stored in a large distributed key-value store without any physical event building. Fragments belonging to one event can be then queried directly by the processes carrying out the data analysis and filtering when needed. We analyze the advantages of this approach. Among them are optimized usage of network and storage resources, and foremost increased CPU efficiency in the computing farm. The latter is possible thanks to the decoupling of the lifetime of the analysis/filtering processes from the changing event rate due to duty cycle of the accelerator. Then, we present the design and initial performance evaluation of *FogKV* - a distributed key-value store for high-bandwidth data acquisition systems. We discuss key design choices, including hybrid NVDIMM/SSD storage backend with buffering and RDMA transport. They are required in order to meet bandwidth and storage requirements of the high luminosity upgrades of the LHC at CERN, after which data will be produced at a rate of 6 TB/s. The storage required to be able to keep data for up to twenty four hours is in the order of 500 PB. We present how single node performance is to scale to meet these requirements.

Primary authors: Dr JERECZEK, Grzegorz (Intel Corporation); LEHMANN MIOTTO, Giovanna (CERN); MOMMSEN, Remi (Fermi National Accelerator Lab. (US)); LOVE, Jeremy Robert (Argonne National Laboratory (US)); LE GOFF, Fabrice (CERN); Mr MACIEJEWSKI, Maciej (Intel Corporation); Mr LEBIODA, Pawel (Intel Corporation); Mr MAKOWSKI, Pawel (Intel Corporation); Mr PELPLINSKI, Piotr (Intel Corporation); Mr RADTKE, Jakub (Intel Corporation); Mrs SZYCHOWSKA, Malgorzata (Intel Corporation); Mrs WISZ, Aleksandra (Intel Corporation)

Presenter: Dr JERECZEK, Grzegorz (Intel Corporation)

Session Classification: T1 - Online computing

Track Classification: Track 1 - Online computing