White Paper to LSN Workshop on Huge Data
Frank Wuerthwein
UCSD, SDSC, OSG

HL-LHC Data Challenges

During the HL-LHC era (~2028-38) the two general purpose experiments ATLAS and CMS are expected to take roughly 80 Billion collision events per year of data taking, plus another roughly 160 billion simulation events. The corresponding data volumes are estimated at roughly an Exabyte of new data per year. Data will be curated by each collaboration independently, and then served to their members, O(1000) scientists that will want to derive science results across hundreds of independent analyses, each leading to peer reviewed publications in major journals.

There are data and networking challenges of a wide variety. To curate the data, i.e. produce high level data formats from low level or RAW data, is a very compute intensive task that is organized centrally by the global collaborations. E.g. at the end of each annual year running period, more than half an Exabyte of RAW data per experiment will need to be processed across whatever resources the collaborations have access to globally. 30-40% of the RAW data is likely going to be archived in one of two tape archives at FNAL and BNL. These hundreds of Petabytes will have to be restaged from tape, and staged in to input buffers at Universities, National Facilities, and possibly Cloud processing centers. Ideally, the collaborations will want to maximally scale out vertically to reduce processing time and thus accelerate time to insight as soon as final calibrations and software are available for end of year processing. The output will come in multiple formats, ranging from the most flexible but largest (AOD roughly 1/3 to ¼ the size of RAW) to the most reduced and easiest to analyze (NANOAOD roughly 1/500 the size of RAW).

The curated science quality data may be thought of as two dimensional, one dimension being the series of independent collision events, and the second dimension the physics objects per event. The science at the LHC is to a significant extend a matter of counting statistics. I.e. for a science result, we count the number of events that satisfy certain criteria, based on a combination of objects in the event. In addition, probability distributions for a variety of characteristics in filtered events are determined, and ultimately published. Both counts and distributions need to be efficiency corrected to be most useful for publication. The latter requires simulations. A standard workflow thus includes event filtering based on a small subset of objects in the event, followed by more detailed analysis of the characteristics of the filtered events. Science thus leads to very sparsely read curated date (10-20% read fractions are typical). Moreover, the data, especially the simulated data, is not all equally popular, and in fact, the popularity is likely to be predictable from its Metadata. The number of times a given file is accessed in a given time period peaks at or near zero, and extends hyper-exponentially towards larger number of reads. All of the above leads to orders of magnitudes differences in the working set per day, month, and year. Here we define the working set to be the data accessed in aggregate by the collective of all scientists during some time period in some region, i.e. summed over a set of nearby processing centers. This in turn allows for “Content Delivery Networks” (CDN) that include domain science aware caching layers in order to perform cost trade-offs between the required disk space, including data replication at processing centers, and the

bandwidth of the networks that connect them. Better networks allow for less investment in caching space at processing centers because shorter time periods for the working set in a cache can be supported without decreases in CPU utilization, i.e. stalling in wait for data.
The shear volume of data and processing required to curate the data are major cost drivers for the HL-LHC. To fit into finite budgets requires significant R&D investment into exploiting as many of the potential cost saving ideas as possible. Now is a good time to engage the community as directions for this “data and networking research” are becoming more clear, and initial prototype CDNs have been built that have a plug-in architecture to allow for experimentation at multiple layers, all the way from the transfer layer (e.g. exploration of NDN), to caching algorithms, even on the fly reformatting, a kind of “virtual data”, and intelligent storage systems are being considered. All of this can be experimented with under production conditions due to devOps deployments using containers and container orchestration, e.g. via Kubernetes. The community thus has adopted an agile infrastructure model ripe for experimentation.

In summary, we see two types of opportunities, some in the area of efficient scheduled bulk data movement for large scale processing during curation, i.e. the processing of more RAW into more derived data. And a second set of opportunities in the area of intelligent CDNs for analysis of the curated data by large numbers of scientists. Furthermore, the community is ready for experimentation in that its core production infrastructure is agile and flexible. Finally, the community needs to experiment because it can otherwise not afford its future computing operations. Significant cost savings need to be found from more optimal use of computing, storage, and networking.