HL-LHC Data Challenges

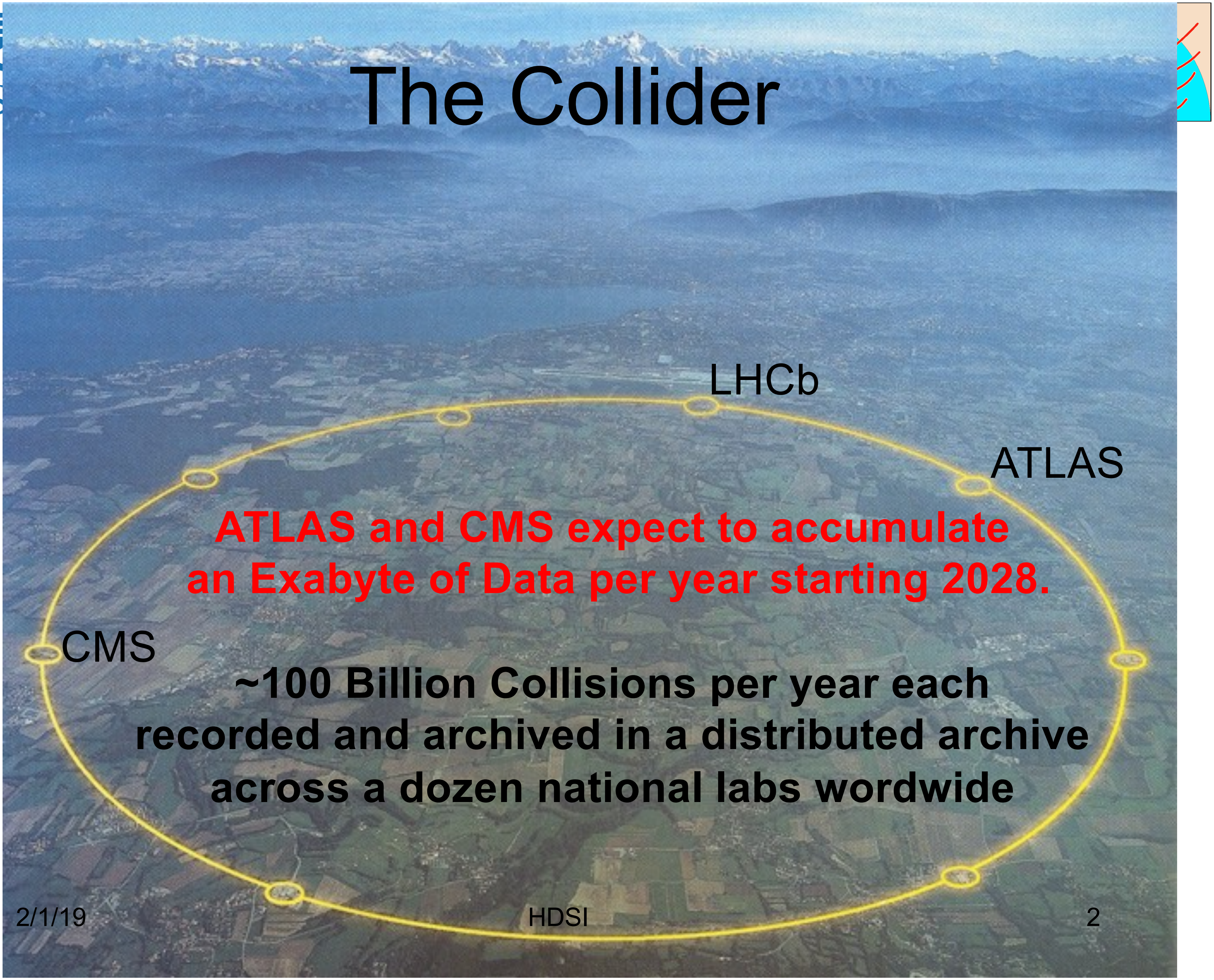
Frank Würthwein

SDSC/UCSD

Huge Data Workshop

April 2020

The Collider



ATLAS and CMS expect to accumulate an Exabyte of Data per year starting 2028.

~100 Billion Collisions per year each recorded and archived in a distributed archive across a dozen national labs wordwide

Challenge 1

- take an exabyte of RAW data from a distributed archive across a dozen national labs worldwide.
- move it for processing to ~100 processing centers
- with probably half a dozen different compute architectures
- producing half an exabyte of output
- bring output back to the archives for custodial storage.

Distributed nature is non-negotiable because no one country is prepared to provide all of the resources to do the job.

This is done once a year. Every 3 years, the processing is 3 times larger. In addition there are similar simulation samples, that also get reprocessed to be consistent with detector data.

This is the easier of the two challenges as it is top down.

Challenge 2

Each of ATLAS and CMS has more than 1000 scientists from a few hundred institutions in more than 50 countries that want to exercise their academic freedom to analyze this data to their hearts contents.

- Data formats designed to support analysis performance from Hz to kHz per CPU with kB to MB of data per collision.
- Datasets have sizes from Millions to Billions of events.
- There are thousands of such datasets.
- Typical analyses require 10 to 100 of these datasets.
- Most analyses access on average <10% of the data per collision
- Access frequency of datasets per month peaks near zero & follows a hyper-exponential distribution (i.e. loooong tails)
- Data access from 100's of locations worldwide.

Innovation & science success depends on academic freedom

See the white paper for more details