Virtual data

Richard Mount (SLAC) – February 1, 2017

The Concept

Derived or simulated data are transparently derived or simulated on demand if not already instantiated on physical storage. When storage approaches capacity, the least valuable physical data are deleted.

History

The concept has been around in HEP since at least the early 1990s. While it has been used as a successful buzz word in attracting funding, it has never been implemented – if you discount the poor-man's version below.

Motivation for Reexamination

The relative cost of storage is likely to increase

- Technical slowdowns with the evolution of storage are even worse than those for CPU.
- We will get opportunistic or free (to HEP) access to CPU, but likely never to reliable persistent storage.
- In experimental particle physics, the raw data cannot usually be re-created and so should be preserved indefinitely. However, storage occupancy is dominated by derived and simulated data.

Solution 1 – Poor Man's Virtual Data

- Lifecycle management of all derived data (when space is tight, delete the data with least probable value, where value takes into account importance to physics and cost and difficulty of re-creation.)

Solution 2 – True Virtual Data

- All derived data are created automatically (data and code provenance is registered in a persistent database that is used to support all instantiations of the derived data).
- Some derived data may be virtual from the start, only being instantiated if needed.
- Data lifecycle management takes place as above, with the addition of automatic re-instantiation when required.
- Gets into deep technical and cost trouble when it is necessary to run old code on old hardware and operating systems. The assumption that transformations are defined only by the code we write and preserve is sadly untrue.

Some Research Topics

- **Data Value**: how to calculate the value of physically stored data and compare it with the cost of its storage.

- **Virtual-Data-Ready Provenance**: Capturing, storing, identifying and retrieving all information needed to create derived or simulated data. Assumes that running on currently available systems will work and produce correct data.
- **Strategies for Older Data:** Including
    - **Deterministic Processing – Insulating from hardware, compilers and run-time environments**: What would be required to ensure adequate invariance of processing or simulation if our code remained invariant.
    - **Feasibility and cost of virtualization**: Where does it make economic sense to provide secure virtualized, or even emulated, versions of obsolete hardware and software environments.