Is the time right for a centralised approach to distributing data and tools to where there are needed?

Greg D’Arcy
Research Engagement Strategist, AARNet

> “While software tools and reference datasets are the building blocks for analytic pipelines and workflows, there's a major roadblock in how we manage and provide these to researchers.”

Existing Roadblocks

Access to reference datasets and tools at compute facilities is incomplete and inefficient.
The cumulative effect is reduced time to research.

Installing tools can require complex steps, multiple dependencies, and extra privileges. Technical skills most researchers don't have.

Copies of reference datasets & tool indices needs to be manually duplicated at each node and cannot be easily kept up to date.

This imposes considerable time and effort by the facility system administrators who need to build and install each new release.

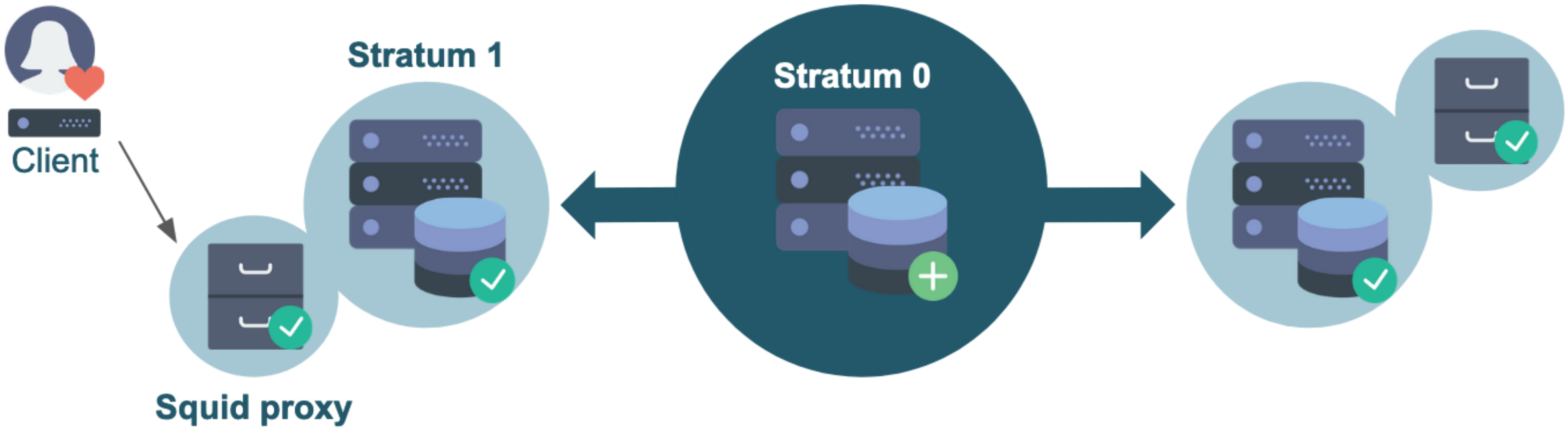
There's a solution to reduce the fragmentation and duplication of effort required to maintain reference data and tool indices across compute facilities.

CernVM-FS

- Through our partnership with the **Australian BioCommons** we have been testing the **Cern Virtual Machine File System (CernVM-FS or CVMFS**) for its potential to provide a distributed file system for software and reference data.
- CernVM-FS was designed to **deliver scientific software** onto virtual machines and physical worker nodes in a fast, scalable and reliable way.
- A national CernVM-FS service has the potential to **accelerate research** by easily deploying tools and reference data to compute facilities.

CernVM-FS: Caching Layers

Files and directories are hosted on standard web servers and are distributed to individual nodes through a **hierarchy of caches**.



The hierarchical structure of **CernVM-FS with multiple caching layers** (Stratum-0, Stratum-1's located at partner sites and local caching proxies) ensures good performance with limited resources.
(European Environment for Scientific Software Installations (EESSI), https://eessi.github.io/docs/filesystem_layer/)

CernVM-FS: The Benefits

Hierarchical caching makes CVMFS **highly scalable** and minimises network traffic.

Reduces **redundancy** as only one copy of software needs to be maintained and can be propagated to and used at multiple sites.

Web proxies reduce the **network latency** for the CernVM-FS clients and the load on the Stratum 1 service.

CernVM-FS: The Benefits

From a researcher's perspective once the file system is mounted it appears **as if the software were locally installed**.

The architecture provides an efficient method for **read-only data sharing** between systems.

CernVM-FS is a good match for container engines to **distribute the container image** contents (sHPC).

Existing Deployments

Where is it being used?

- High Energy Physics
- Medical Sciences
- Physical Sciences
- Space and Earth Sciences
- Biological Sciences

Where to next?

CernVM-FS has potential to support a **national distribution network** capable of delivering tools and reference datasets to where they are needed.

Building a reliable and sustainable production service comes with deployment costs, we are **building a business case to evaluate the feasibility of an Australian distribution system for software and reference data.**

We're interested to talk with research groups/other infrastructure providers with use cases for how such a service might be used.

aarnet

Thank you