Bubble wrapping EasyBuild ...

... and other ways to stage installations

Bart Oldeman

McGill University, Calcul Québec, Digital Research Alliance of Canada

Research Support National Team Software Installation Coordinator

(with Maxime Boissonneault, Charles Coulombe, Doug Roberts (RSNT), Ryan Taylor (CVMFS))

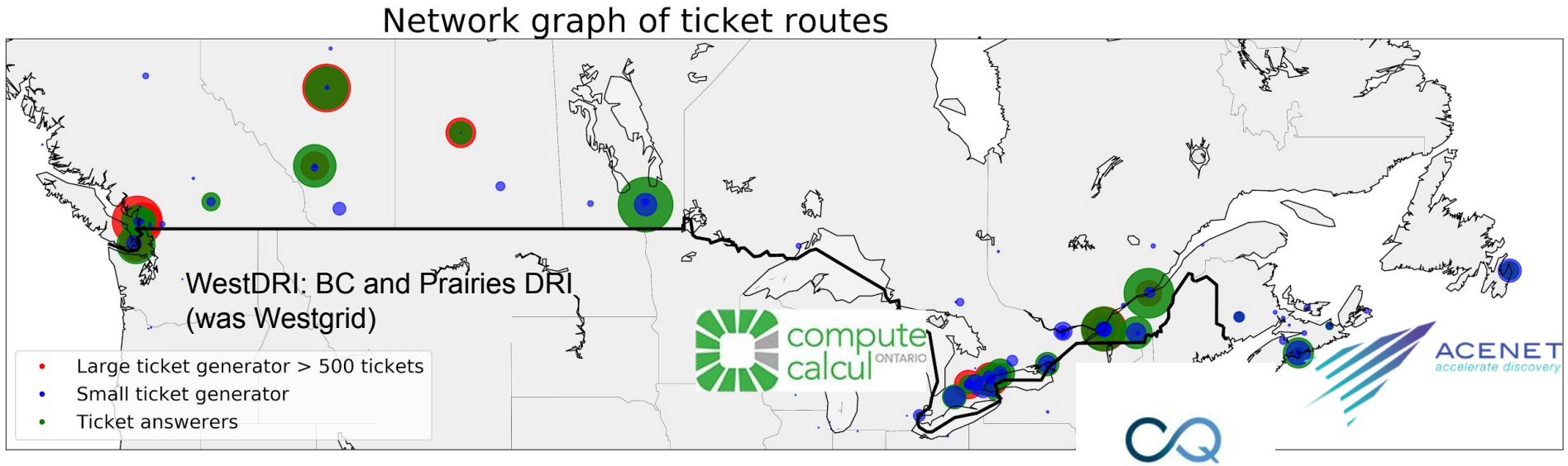
McGill

Calcul Québec

Partenaire régional de l'
Alliance de recherche numérique du Canada

A regional partner of the
Digital Research Alliance of Canada

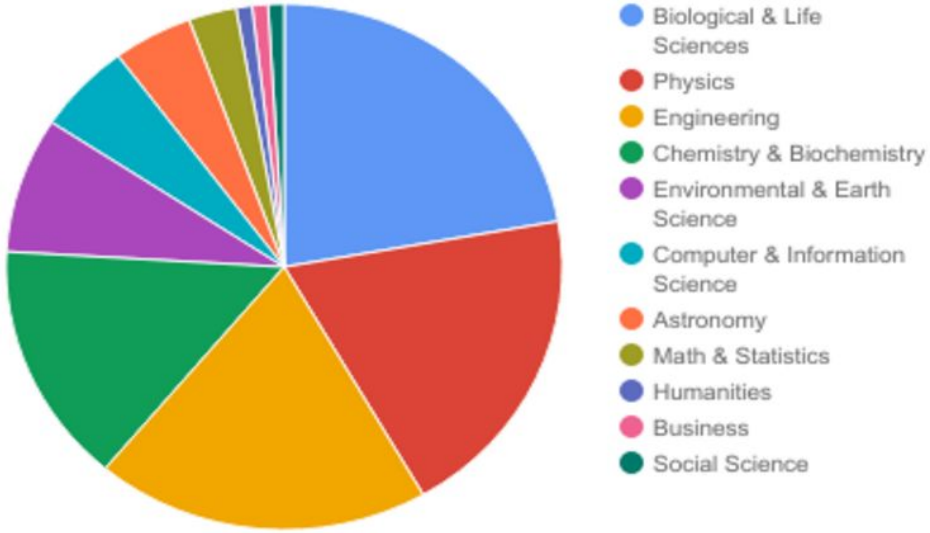
The people



All research disciplines supported

Free access for any researcher at a Canadian institution

- 5 regional consortia
- 38 member institutions
- ~250 technical staff
- ~18,000 user accounts
- 6 clusters, 4 clouds, 815k cores, 2k GPUs, 100s PB storage



The hardware

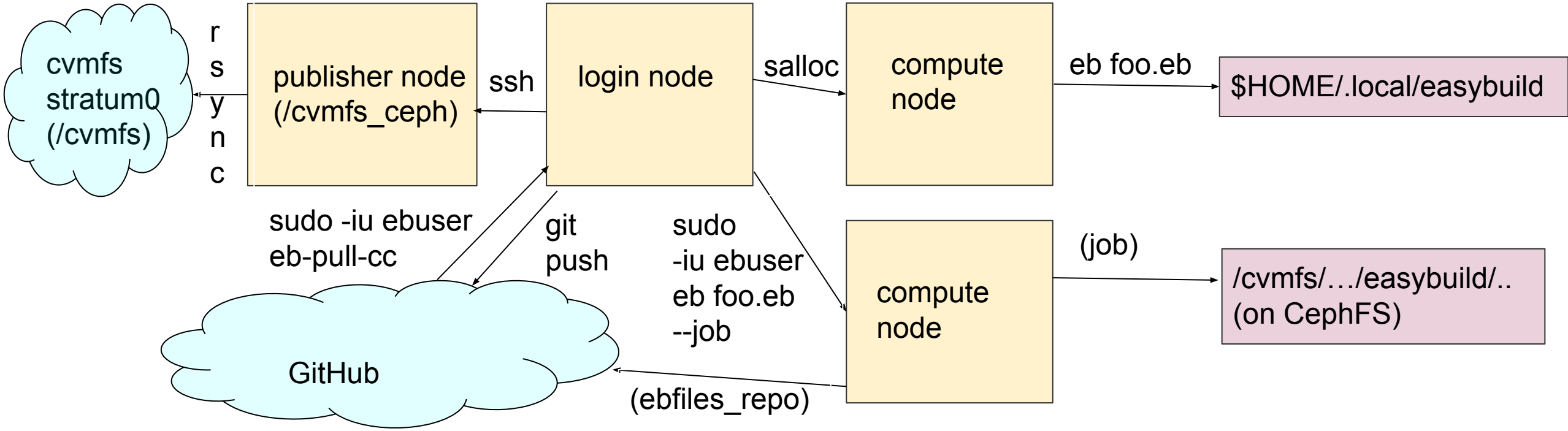
6 major national systems
300K->815K cores,
90->165 PB storage

System	Type	Network	Production
Arbutus (renew)	Cloud	Ethernet	2025 H2
Cedar->Fir	General	NDR IB	2025 H2
Graham->Nibi	General	200 GbE	2025 H2
Niagara-> Trillium	Large MPI	NDR IB	2025 H2
Béluga->Rorqual	General	HDR IB	2025 H2
Narval	General	HDR IB	2021 H2

All clusters use the same (distributed) software stack.

Archimedes, our build cluster

- Magic Castle cluster-in-the-cloud that lives on Arbutus
- login, management, test, Jupyter nodes, a publisher node, compute nodes via Slurm and auto-scaling; sharing /home and /cvmfs (rw-copy) on CephFS
- All managed via puppet, easy to bring back up.

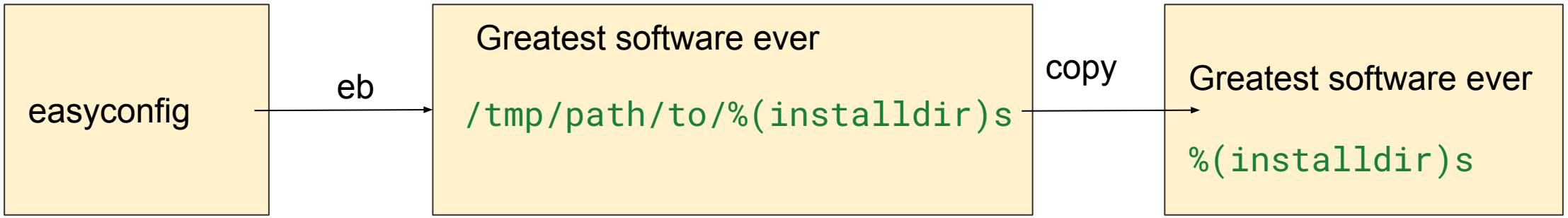


Motivation

- From last year's site talk at EUM in Jülich:
  - `%(installdir)s` on CephFS can be a bottleneck for certain software installations; an EESSI inspired method via local storage, overlayfs, and Apptainer or bubblewrap could speed this up

Staged installations, what are those?

- Two step process:
  - install the software in a different location than `%(installdir)s`
  - copy the reinstalled software to the final location, as atomically as possible
- used by
  - (Almost?) all Linux distributions' package managers
  - EESSI via a build container and unionfs
  - EasyBuild 5.3.0 with `eb --bwrap --experimental`



Staged installations, reason 1

- when reinstalling software with EB on an active cluster, the software is unavailable or in a broken state during the reinstallation process, which might take a long time

Staged installations, reason 2

- On Archimedes software is installed under /cvmfs:
  - /cvmfs is a shared cephfs file system, not CVMFS!
  - it is visible on all compute nodes and the publisher node
- Particularly painful on large binary installations (our restricted repo)
  - We need to patch the binaries with patchelf to link to our compatibility layer glibc, in postinstallcmds involving a slow scan
  - Then the files are read again when pushed to CVMFS
  - A local filesystem is faster for those operations.
  - Examples:
    - MATLAB 2025b.1: 704k files, 23G
    - ANSYS 2025R2.04: 537k files, 155G
    - QuantumATK 2026.03: 105k files, 10G

Staged installations, comparison benchmarks

	`tar xf gcc-11.3.0.tar.gz` (113883 files)	`eb QuantumATK-2026.03.eb` with postinstallcmds scanning/patchelf-ing files
local SSD	6.6s	275s
new shared CephFS	29s	764s
old shared CephFS	900s	7506s

Staged installations, using DESTDIR

- Method commonly used by Linux packaging managers:
- For example, in RPM spec files you often see:
  - ```
    %install
    make install DESTDIR=%{buildroot}
    ```

- DESTDIR is similarly supported by cmake, meson and other common build tools.
- Issue: not all software uses those, and EB deals with some exotic installations.
- See https://dwheeler.com/essays/automating-destdir.html for some background and early workarounds (written in 2011)
    ```

Staged, installations the EESSI way

- `eessi_container.sh` uses unionfs inside apptainer to do something similar (or overlayfs as older default):

```
apptainer shell --contain ... --fusemount
"container:unionfs -o cow -o relaxed_permissions
/tmp/$USER/software.eessi.io=RW:/cvmfs_ro/software.e
essi.io=RO /cvmfs/software.eessi.io"
docker://ghcr.io/boegel/build-node:debian12-amd64r
```

- advantage: consistent environment from apptainer environment

Staged installations, step 2

- Method commonly used by Linux packaging managers:
- For installing package from the staged location some tricks are used, e.g. Gentoo portage:
  - only replaces files that are not (byte for byte) identical
  - uses rename(2) to atomically replace files on the same file system:
    - processes that have the old file open will see the old file
    - its inode stays hidden (see `/proc/<pid>/fd/<fd>`) until the last process that has the file open terminates
- `rsync --delete-after -avc /path/to/bwrap/dir/ /path/to/original/dir`

  works similar but uses checksums instead of Python's `filecmp.cmp`
- or `mv <origdir>{,.backup}; mv <bwrapdir> <origdir>` …

Bubblewrap: a very lightweight container

- Sandboxing tool using Linux cgroups to manipulate mount points (+ IPC, PID, network, UTS) in private user namespace
  - `bwrap --dev-bind / / --bind /path/to/bwrap/software/name/ver /path/to/software/name/ver eb ...`
  - or via overlays, see https://github.com/easybuilders/easybuild-framework/pull/5173
    `bwrap --dev-bind / / --overlay-src /path/to/software --overlay /path/to/bwrap/software /path/to/bwrap/workdir /path/to/software eb ...`
    This can be used if `/path/to/software/name/ver` does not exist and cannot be created (e.g. read-only (cvmfs) file system)

Bubblewrap support: added to EB 5.3.0

- `eb --experimental --bwrap --bwrap-installpath=/path/to/bwrap`
- Contribution from Sam Moors
  - https://github.com/easybuilders/easybuild-framework/pull/5130
  - Needed some easyblock adjustments to clean instead of remove `%(installdir)s` (`rmdir` fails inside `bwrap`)
- This bubble wraps the software path (only!); for modules it doesn't need to, can just use an explicit destination directory using `--installpath-modules`
- Also creates json file `/path/to/bwrap/bwrap_info.json` with information about the directories and flags used.

Getting the result in place (step 2 with cvmfs)

- Our solution, in eb wrapper script:
- `tar cvf /shared_tmp/<tarball> --owner=libuser --directory=<bwrap_installpath> modules software`
  Takes ~1 min with `/shared_tmp` for QuantumATK on new CephFS
- Then this tarball can be unpacked on the publisher node and sync-ed to CVMFS stratum-0 (no tarball ingestion on our end).
- `bwrap_info.json` can be used to obtain information for the tar command
- EESSI does something similar, but with tarball ingestion
- EB has built-in packaging capabilities using FPM (https://docs.easybuild.io/packaging-support), but
  - it packages a bit too early, so files from the easybuild directory (logs, reprod) are missing.

Present state: restricted repo no longer on Ceph!

Demo: Sam Moors (remote)