A fully unprivileged CernVM-FS

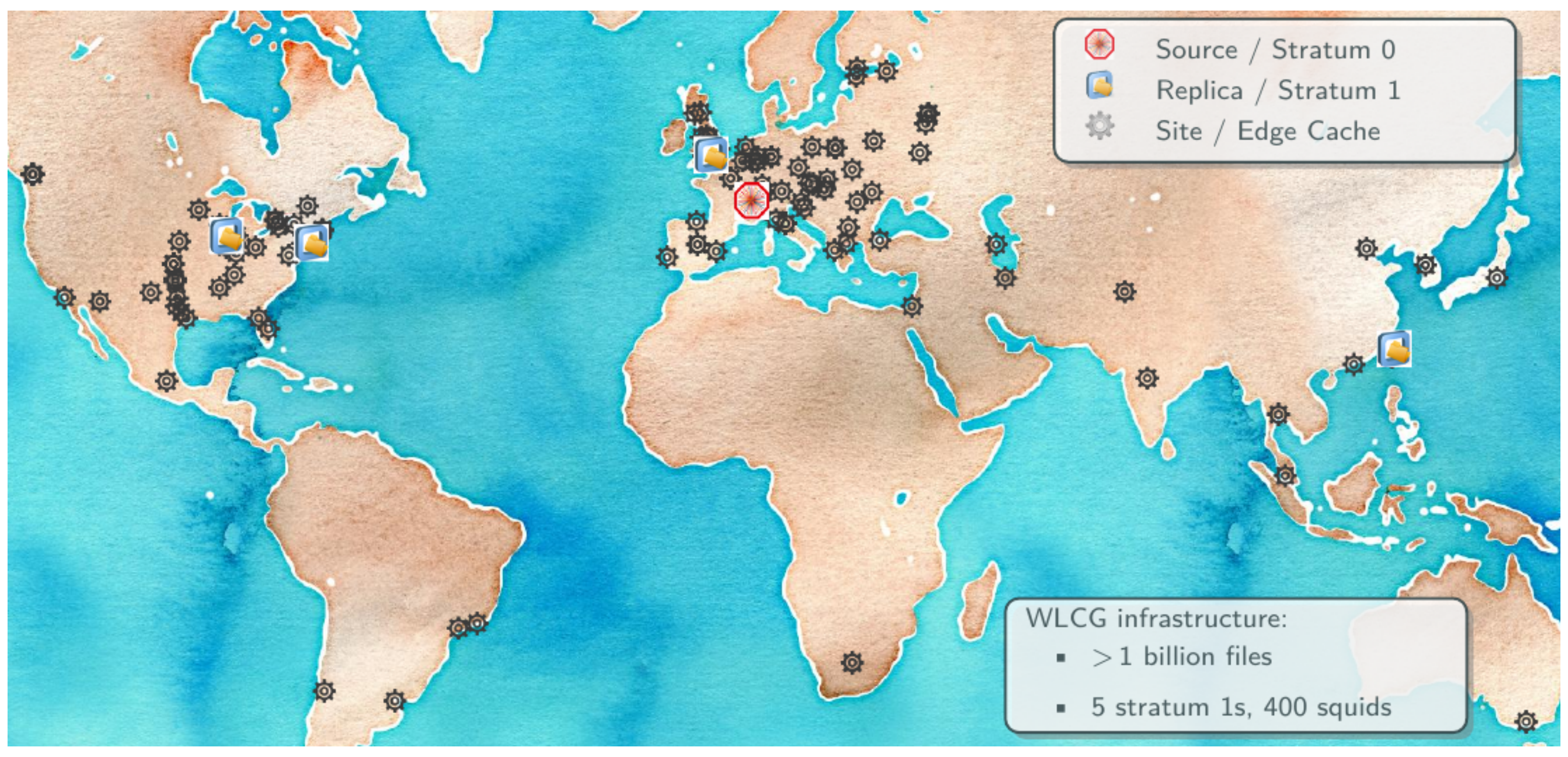
J Blomer[1], D Dykstra[2], G Ganis[1], S Mosciatti[1], J Priessnitz[1]

[1]CERN [2]Fermilab

jblomer@cern.ch

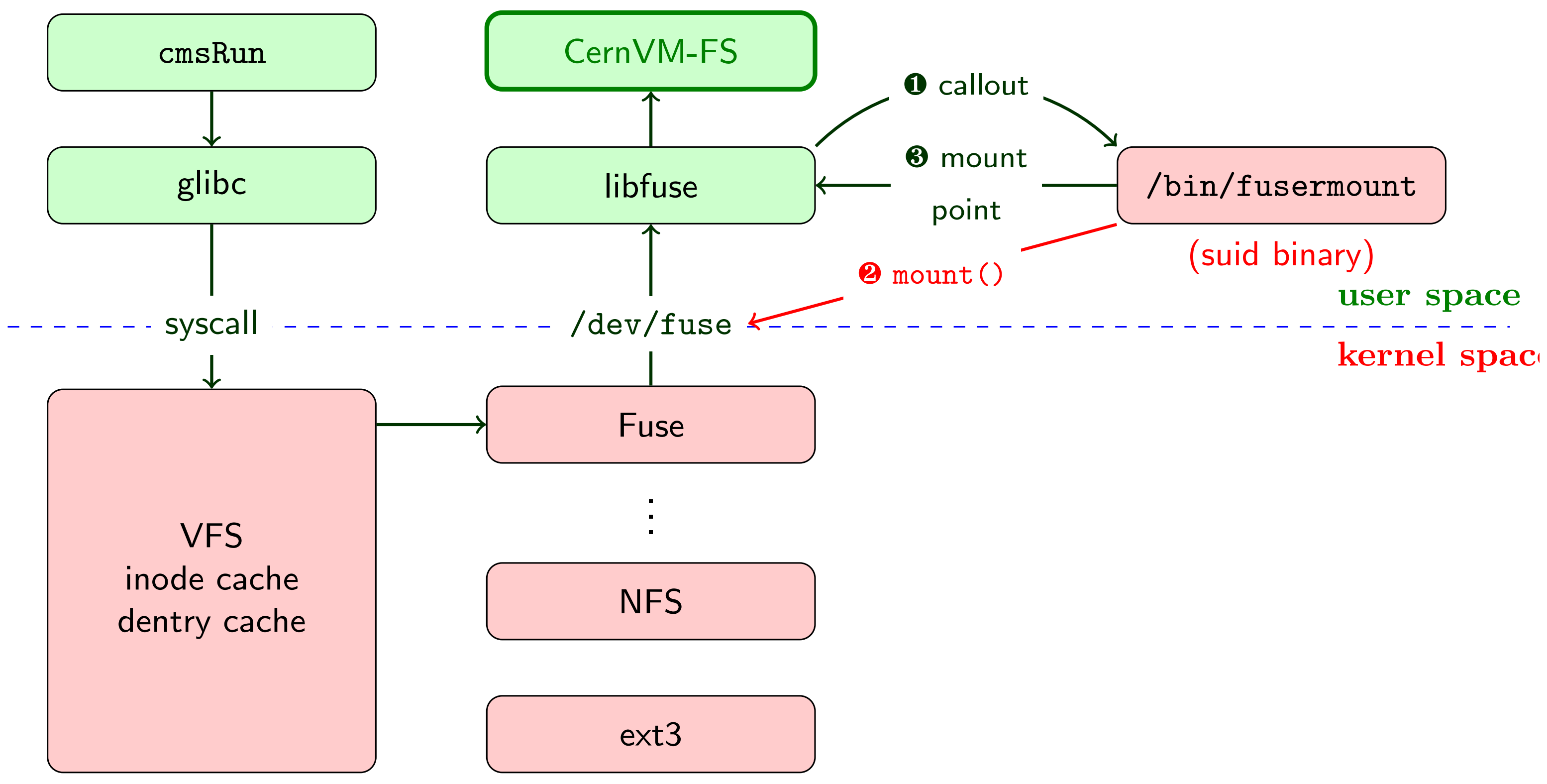
CernVM-FS – Status and Deployment

The CernVM File System provides the **software and container distribution backbone** for most High Energy and Nuclear Physics experiments [1]. Its key features include a POSIX compliant interface, HTTP transport, multi-level caching, versioning, strong consistency, and end-to-end data integrity.



Privileges for File Systems in User Space

CernVM-FS is implemented as a **file system in user-space (FUSE)** [2] module, which permits its execution without any elavated privileges. Yet, mounting the file system in the first place is handled by a privileged suid helper program that is installed by the fuse package on most systems.


cmsRun
glibc
syscall
CernVM-FS
libfuse
/dev/fuse
❶ callout
❸ mount point
❷ `mount()`
`/bin/fusermount`
(suid binary)
user space
kernel spac
VFS
inode cache
dentry cache
Fuse
⋮
NFS
ext3

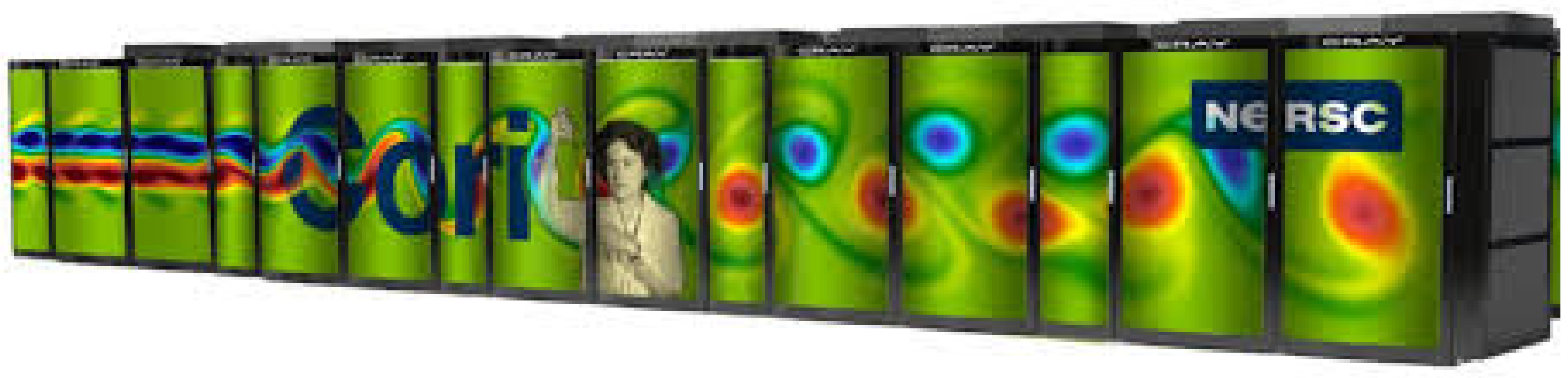
A successful fuse mount returns a file descriptor to /dev/fuse, which is subsequently used by the fully unprivileged *fuse module*.

On-Demand Mounts on Opportunistic Resources

The privileged nature of the `mount` system call is a **serious hindrance to running CernVM-FS on opportunisitic resource and supercomputers**. While the fuse kernel module is a standard Linux facility, the execution of suid binaries is forbidden at some of the biggest supercomputers. Likewise, suid binaries are usually not available in containers.

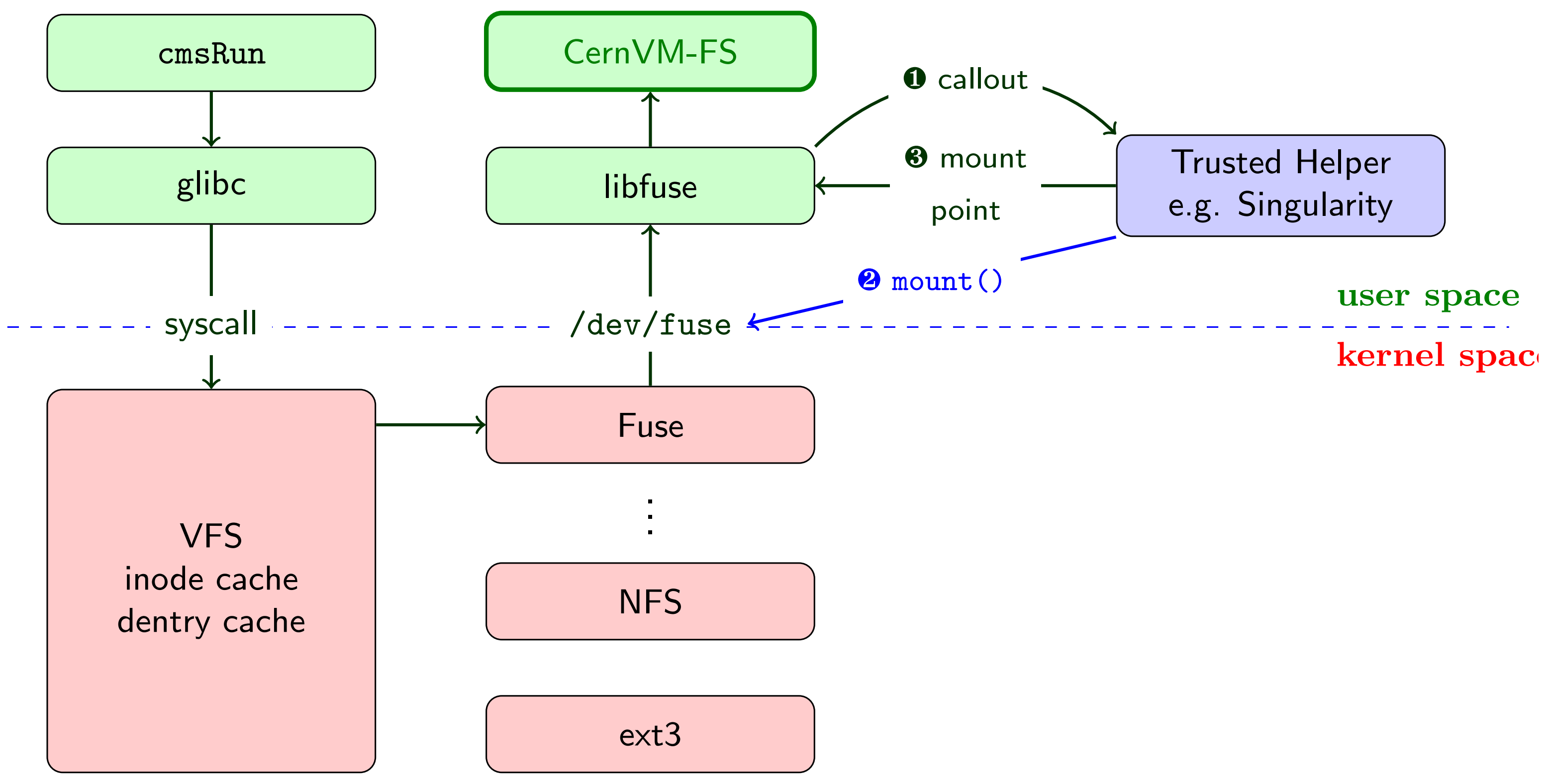
Recent FUSE feature were integrated with CernVM-FS in order to

- enable fully unprivileged mounting of FUSE file systems
- outsourcing `mount()` to a trusted, external process.



New Feature: Pre-mounted File System

With the new FUSE3 libraries, the task of mounting /dev/fuse can be handed to a trusted, external helper. Support for mounting /dev/fuse has been added to Singularity, which runs as a trusted process on many supercomputers. Fuse 3 support has been added to CernVM-FS. FUSE3 libraries have been backported to EL6 and EL7 platforms.


cmsRun
glibc
syscall
CernVM-FS
libfuse
/dev/fuse
❶ callout
❸ mount point
❷ `mount()`
Trusted Helper
e.g. Singularity
user space
kernel spac
VFS
inode cache
dentry cache
Fuse
⋮
NFS
ext3

Pre-mounting is implemented in **Singularity** 3.4 and **CernVM-FS** 2.7 (tagged)!

New Feature: Namespace Mounts with FUSE 3

- Explain name space call chain for unprivileged mounts
- Point out that this is a kernel-level feature available with EL8

Namespace mounts enable CernVM-FS in unprivileged containers!

Application ❶: "Universal Pilot"

- generally usable "super pilot" consisting of the pilot code bundled with singularity and cvmfs
- Refer to `cvmfsexec`

Application ❷: On-Demand Publishing

- Describe that client is required to publish for the r/o layer
- Show the prototype `cvmfs enter`

References

[1] *Towards a serverless CernVM-FS*, EPJ Web Conf **214** (2019)

[2] *To FUSE or Not to FUSE: Performance of User-Space File Systems*, Proc. 15th USENIX conf. on File and Storage Technologies (FAST'17)