Open Science Grid

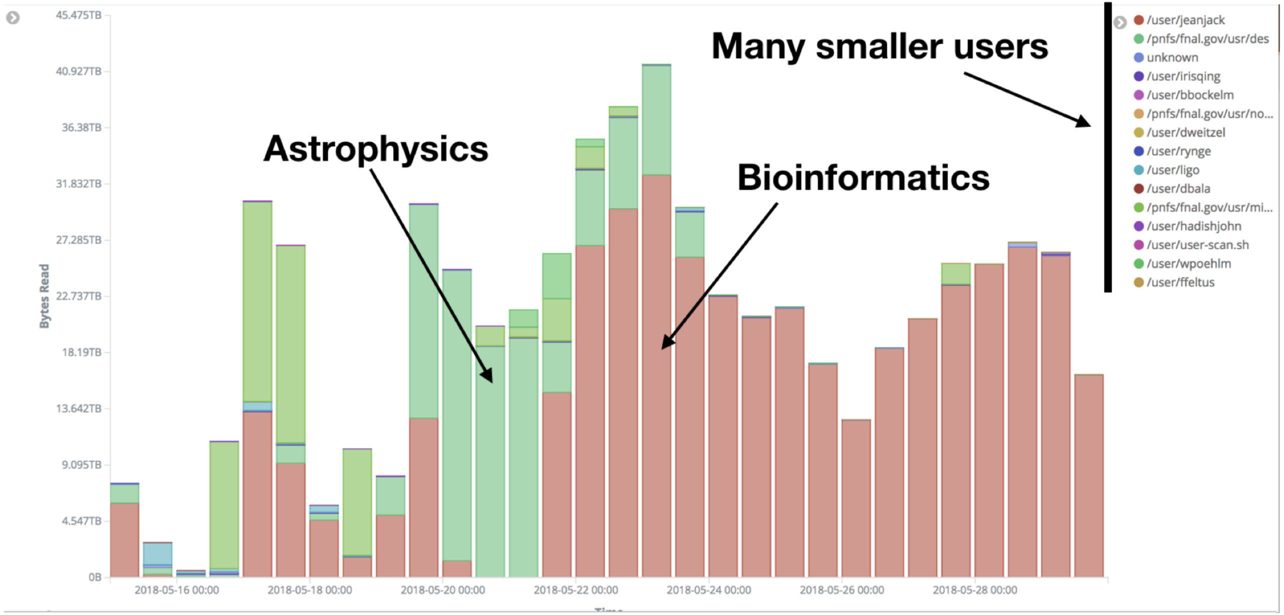
Large Output and Shared File Systems

Thursday PM, Lecture 1

Lauren Michael

StashCache

- Lots of experiments also use StashCache

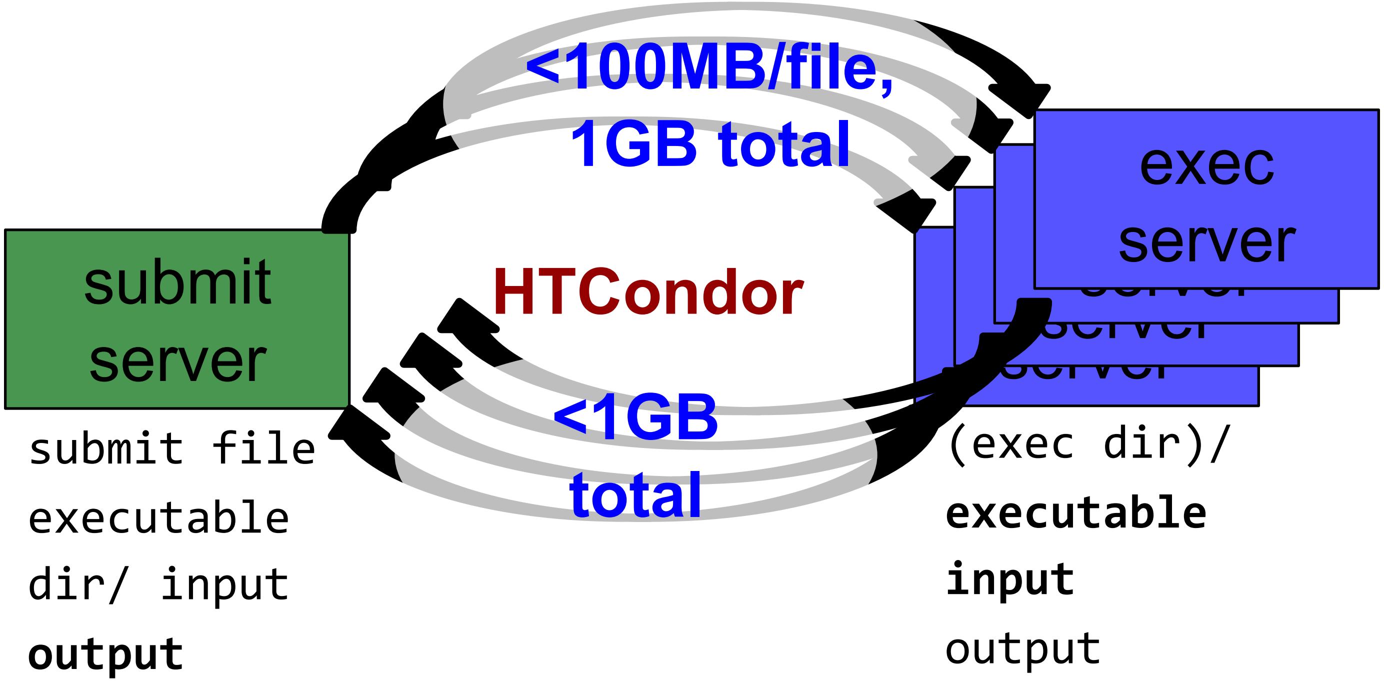


Many smaller users

Astrophysics

Bioinformatics

Per-job transfer limits



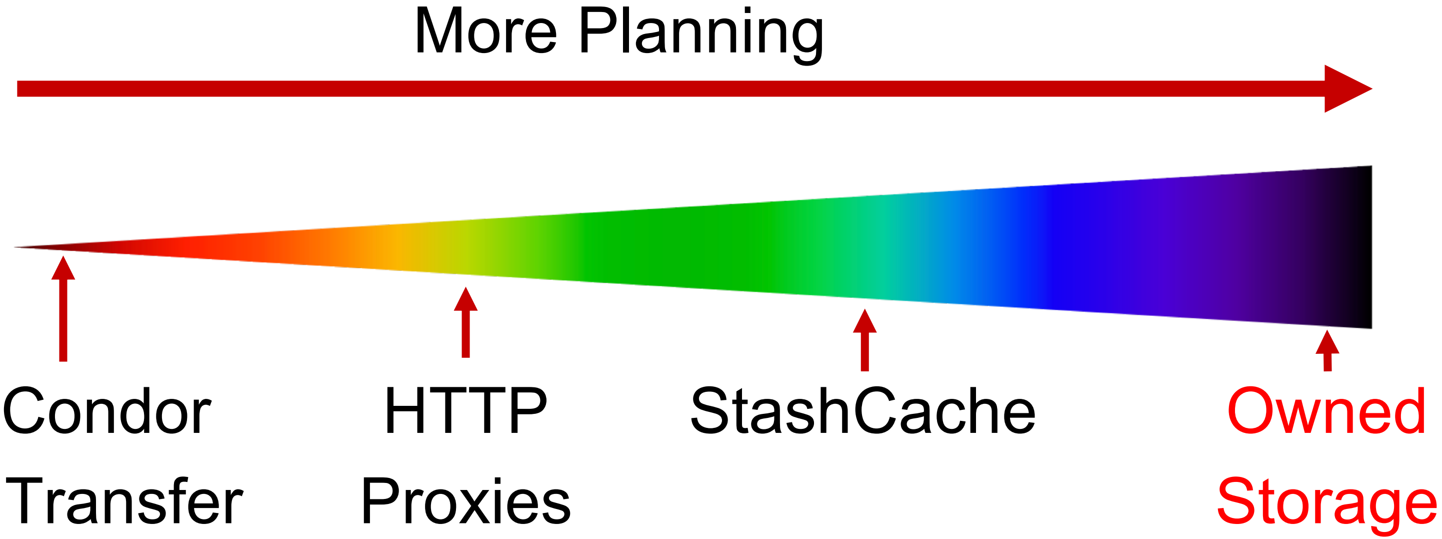
<100MB/file, 1GB total

<1GB total

```
submit file
executable
dir/ input
output
```

```
(exec dir)/
executable
input
output
```

Transfers



More Planning

HTCondor File Transfer

HTTP Proxies

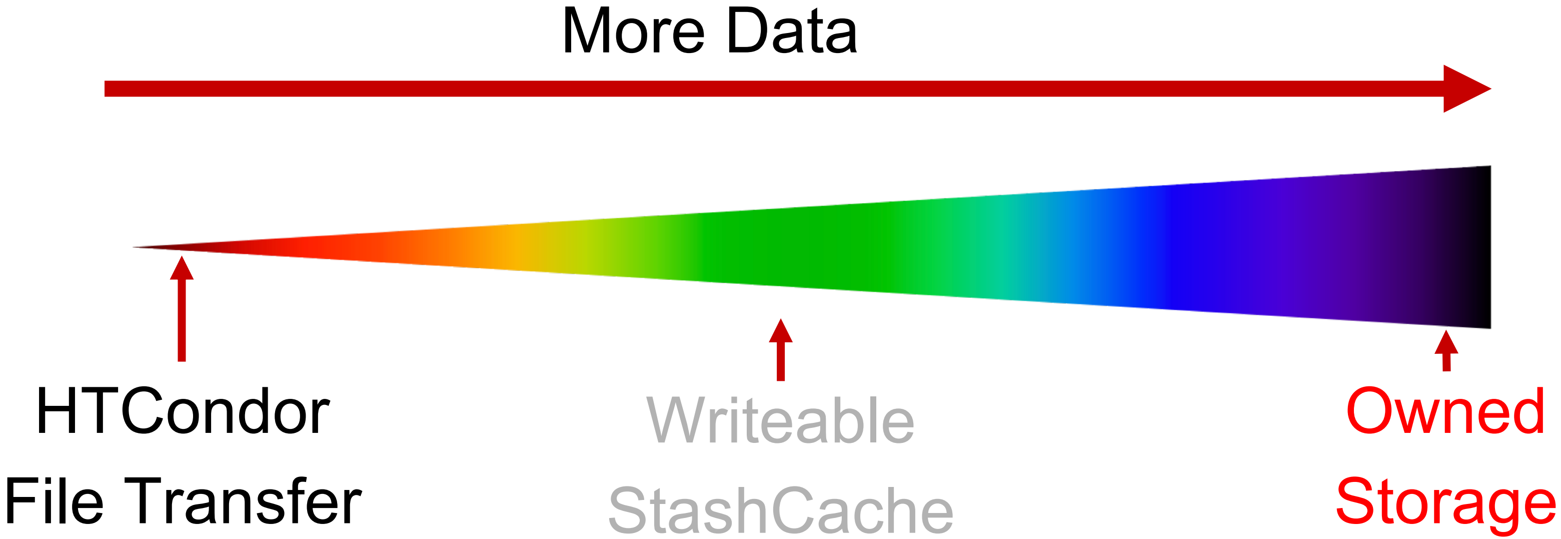
StashCache

Owned Storage

What's Different for Output?

- always unique (right?), so caching won't help
- files not associated with your local username
  - security barriers outside of local context
- security issues with world-writability
  - (versus okay world-readability for input)

Output



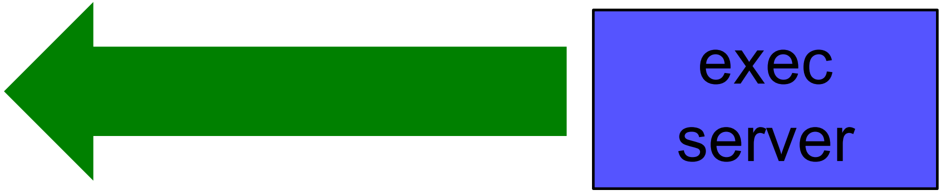
More Data

HTCondor File Transfer

Writeable StashCache

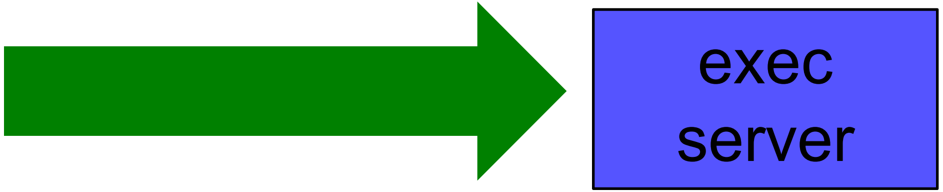
Owned Storage

Output for HTC and OSG



file size	method of delivery
~~words~~	~~within executable or arguments?~~
tiny – 1GB	HTCondor file transfer (up to 1 GB total per-job)
1GB+	**shared file system (local execute servers)**

Large input in HTC and OSG



file size	method of delivery
words	within executable or arguments?
tiny – 10MB per file	HTCondor file transfer (up to 1GB total per-job)
10MB – 1GB, shared	download from web proxy (network-accessible server)
1GB - 20GB, unique or shared file	StashCache (regional replication)
20 GB – TBs, unique or shared	**shared file system (local copy, local execute servers)**

(Local) Shared Filesystems

- data stored on file servers, but network-mounted to local submit and execute servers
- use local user accounts for file permissions
  - Jobs run as YOU!
  - readable (input) and writable (output, most of the time)
- *MOST* perform better with fewer large files (versus many small files of typical HTC)

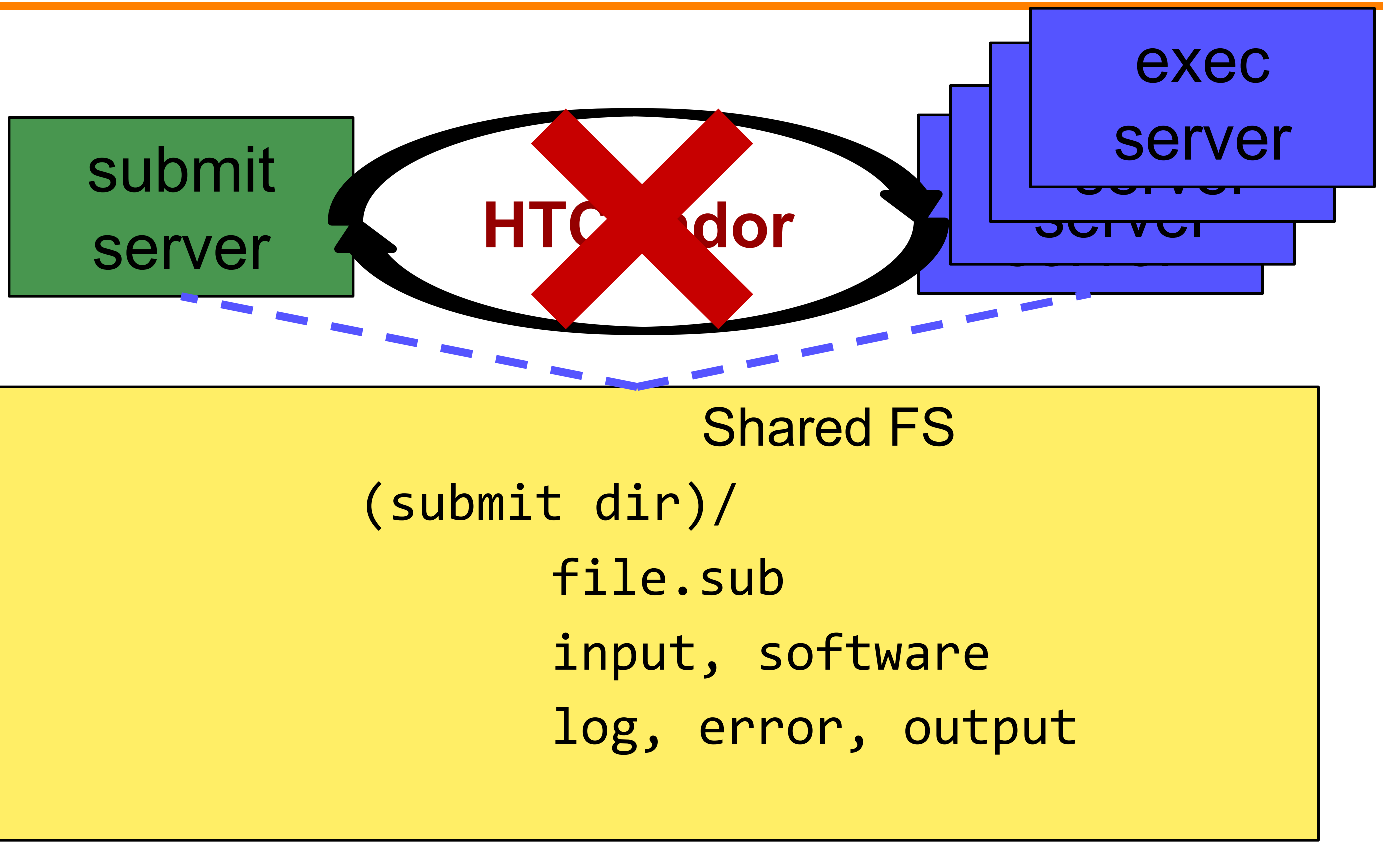
Shared FS Technologies

- *via network mount*
  - NFS
  - AFS
  - Lustre
  - Gluster (may use NFS mount)
  - Isilon (may use NSF mount)
- *distributed file systems (data on many exec servers)*
  - HDFS (Hadoop)
  - CEPH

Shared FS Configurations

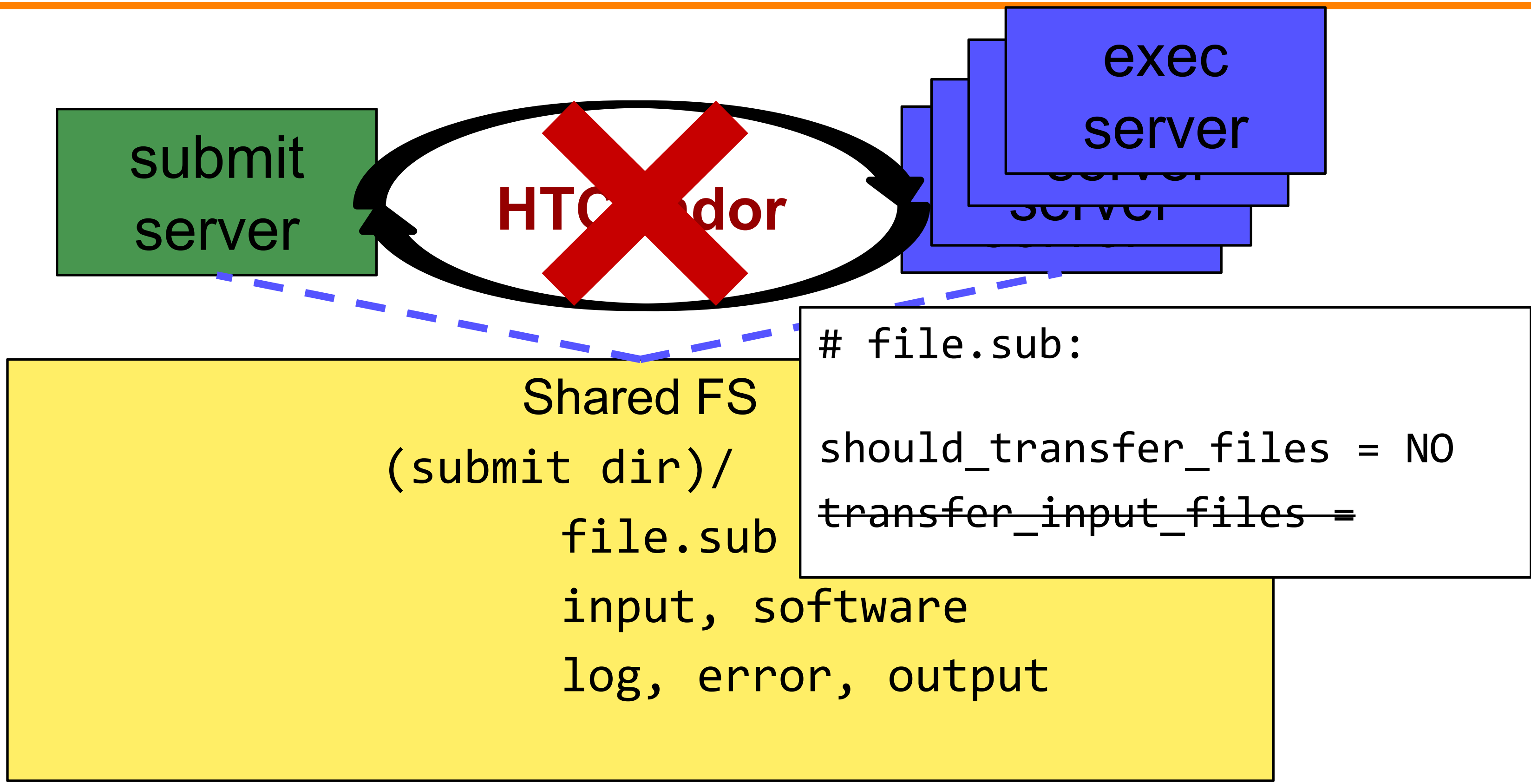
1. Submit directories *WITHIN* the shared filesystem
   - most campus clusters
   - limits HTC capabilities!!
2. Shared filesystem separate from local submission directories
   - supplement local HTC systems
   - treated more as a repository for VERY large data (>GBs)
3. Read-only (input-only) shared filesystem
   - Treated as a repository for VERY large input, only

Submit dir within shared FS



```
(submit dir)/
          file.sub
          input, software
          log, error, output
```

Submit dir within shared FS



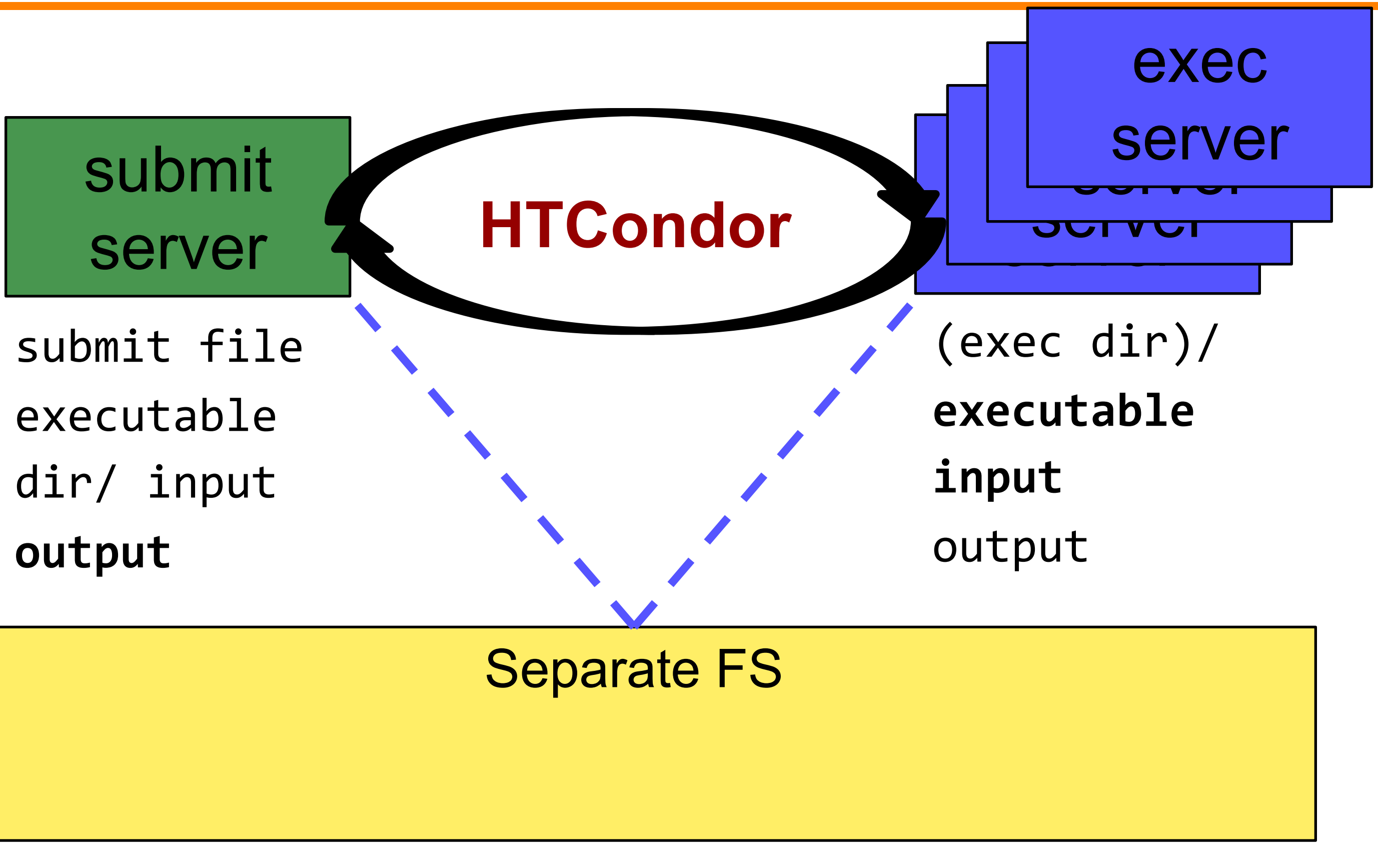
Shared FS

```
(submit dir)/
        file.sub
        input, software
        log, error, output
```

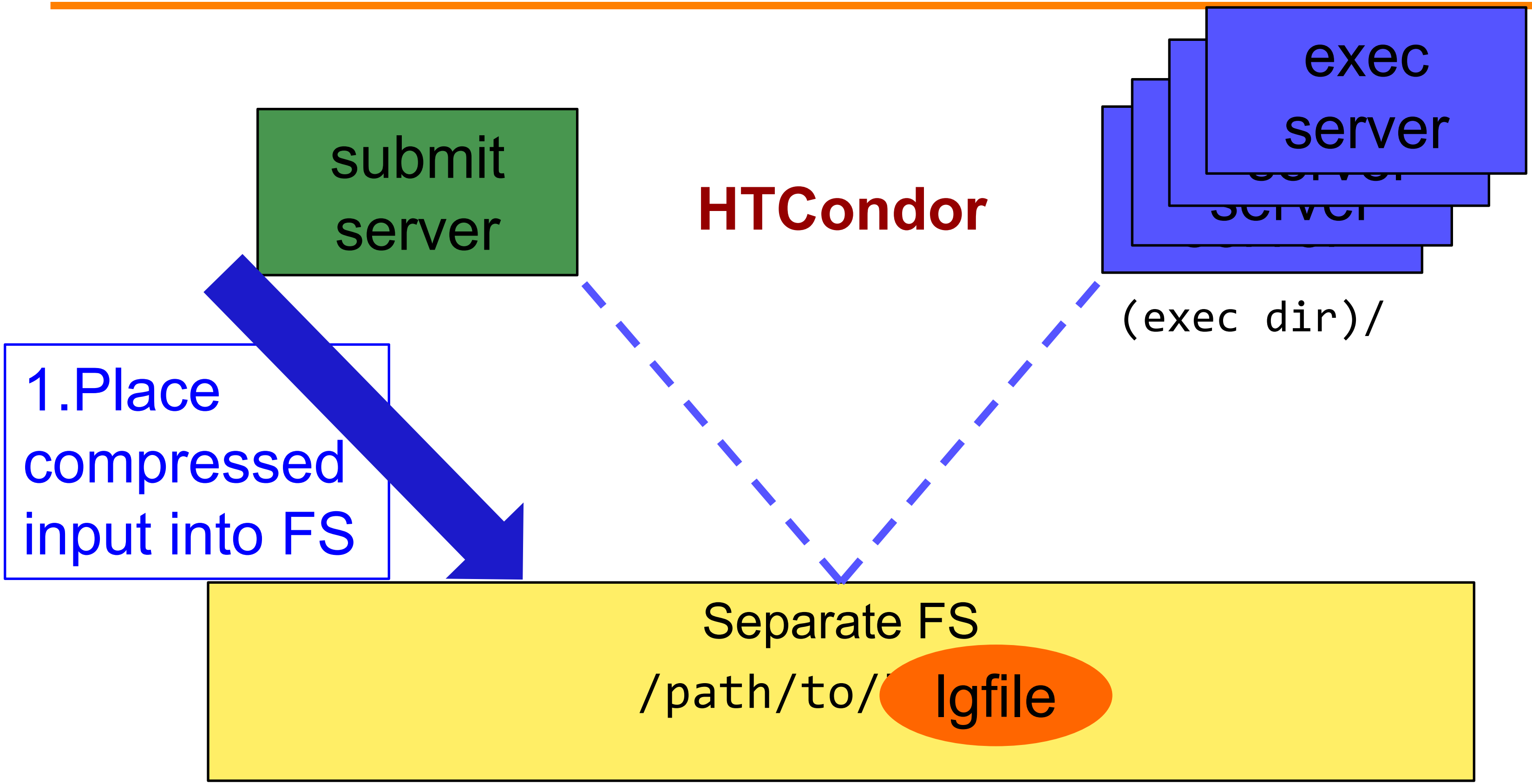
```
# file.sub:

should_transfer_files = NO
```

~~transfer_input_files =~~

Separate shared FS

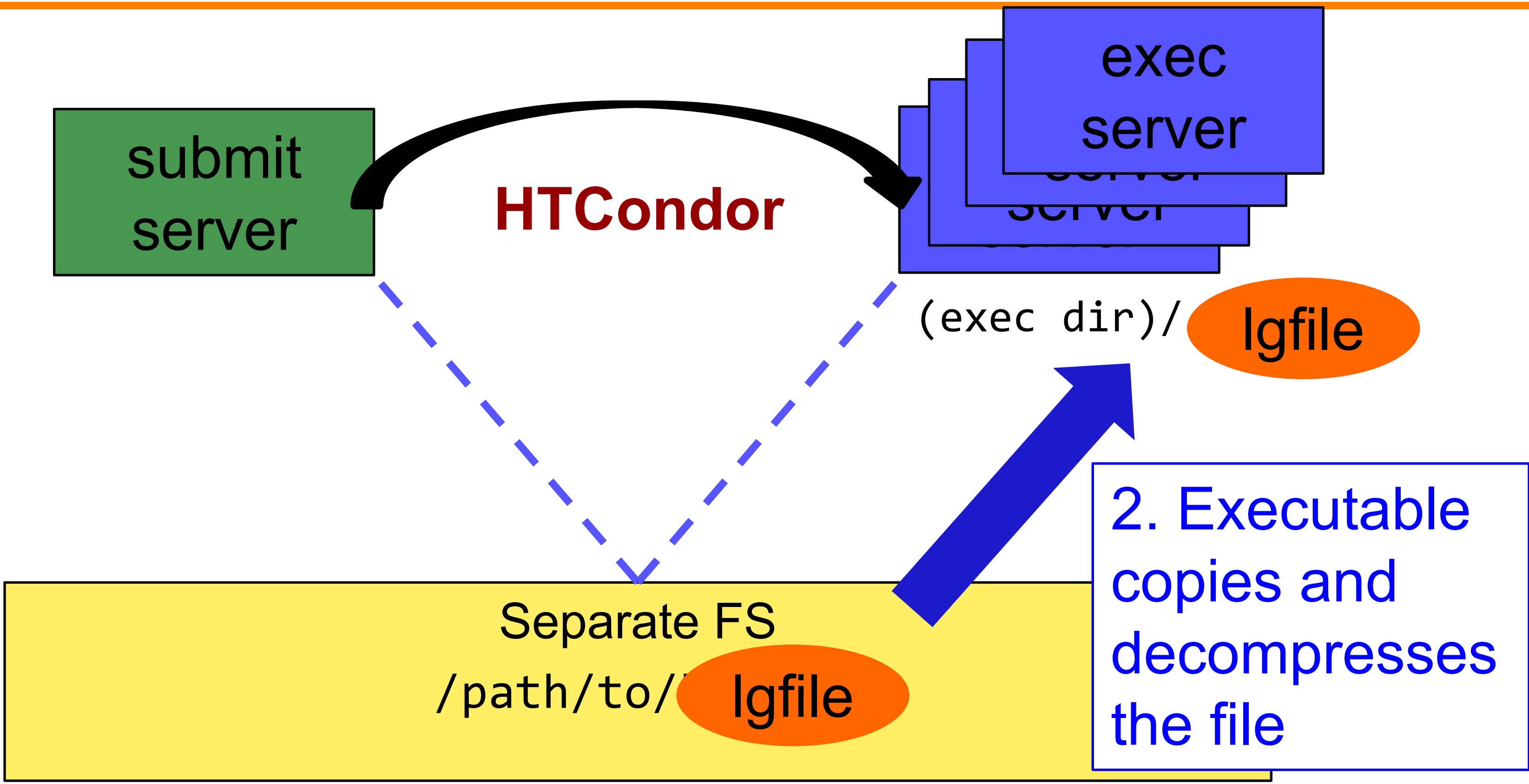


Separate shared FS - Input



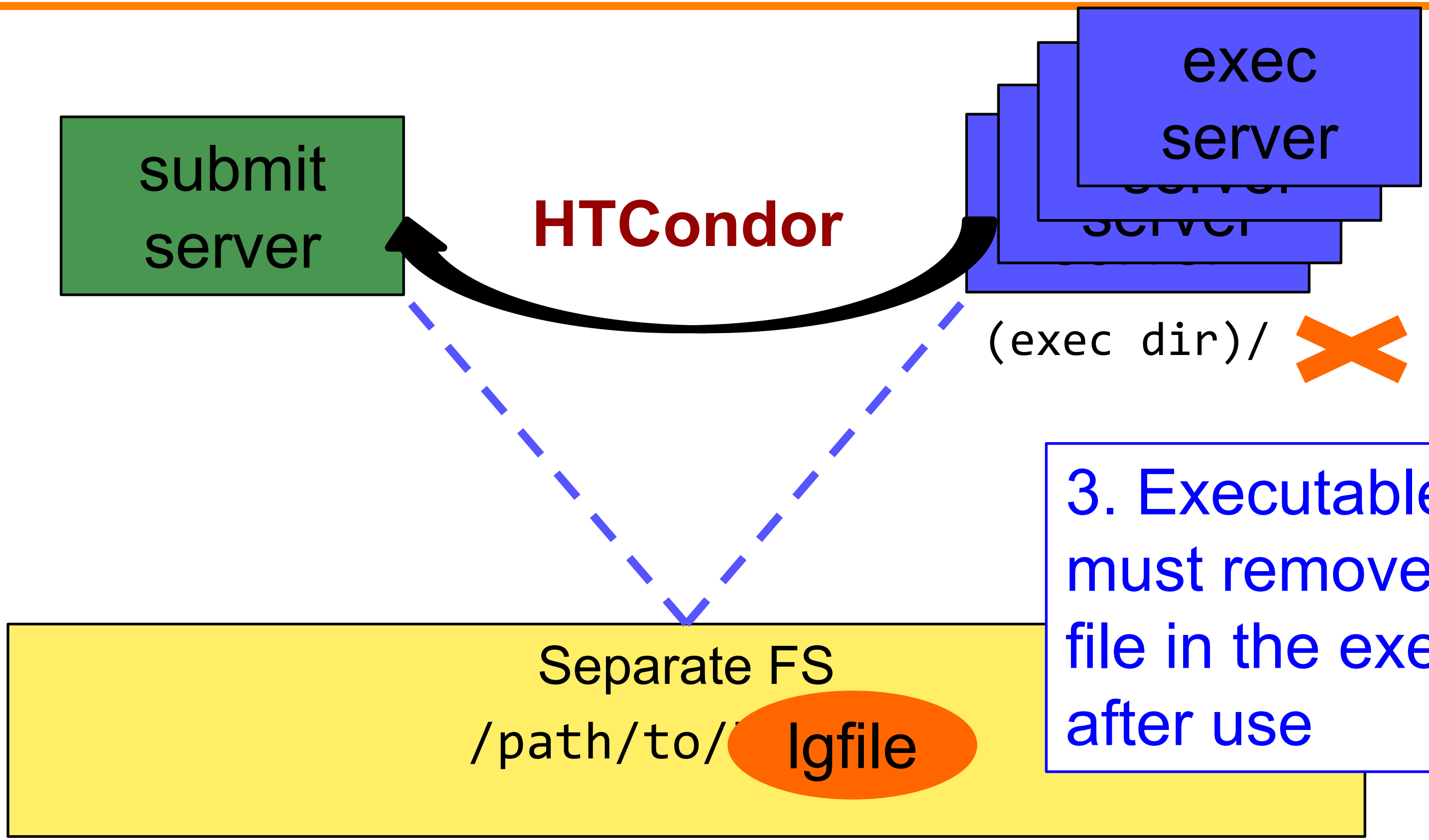
1. Place compressed input into FS

Separate shared FS - Input



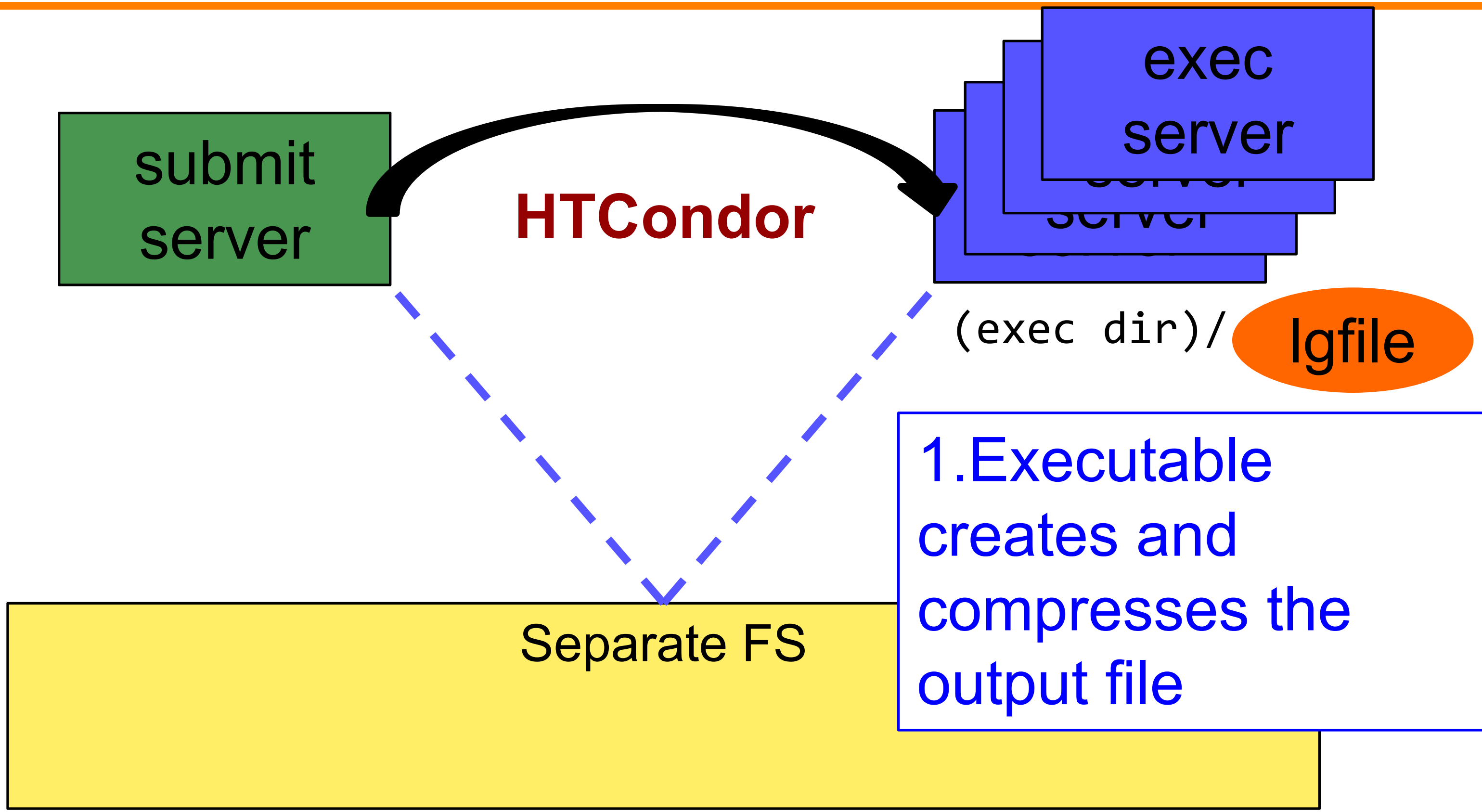
2. Executable copies and decompresses the file

Separate shared FS - Input



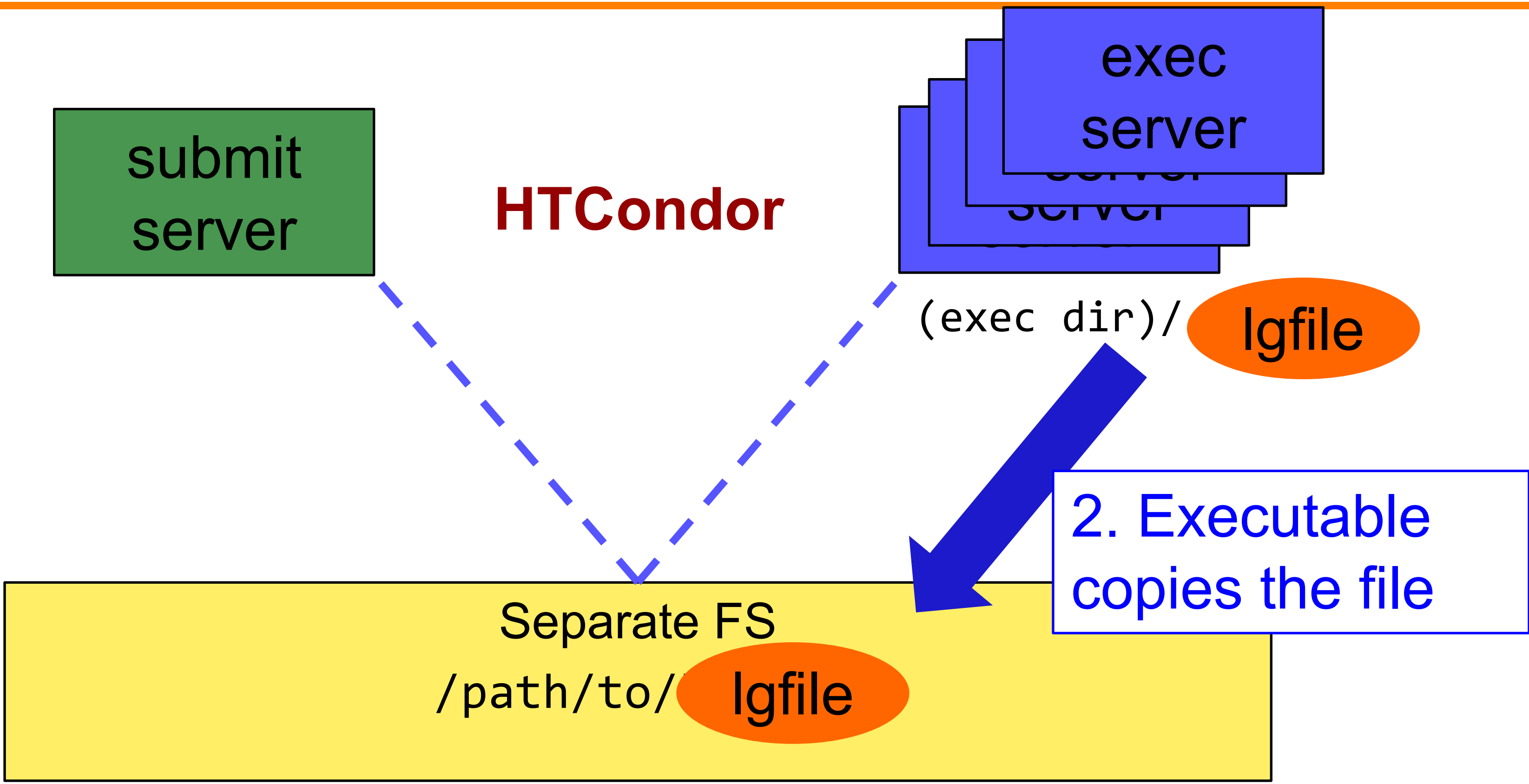
3. Executable must remove the file in the exec dir after use

Separate shared FS - Output



1.Executable creates and compresses the output file

Separate shared FS - Output



submit server

HTCondor

exec server

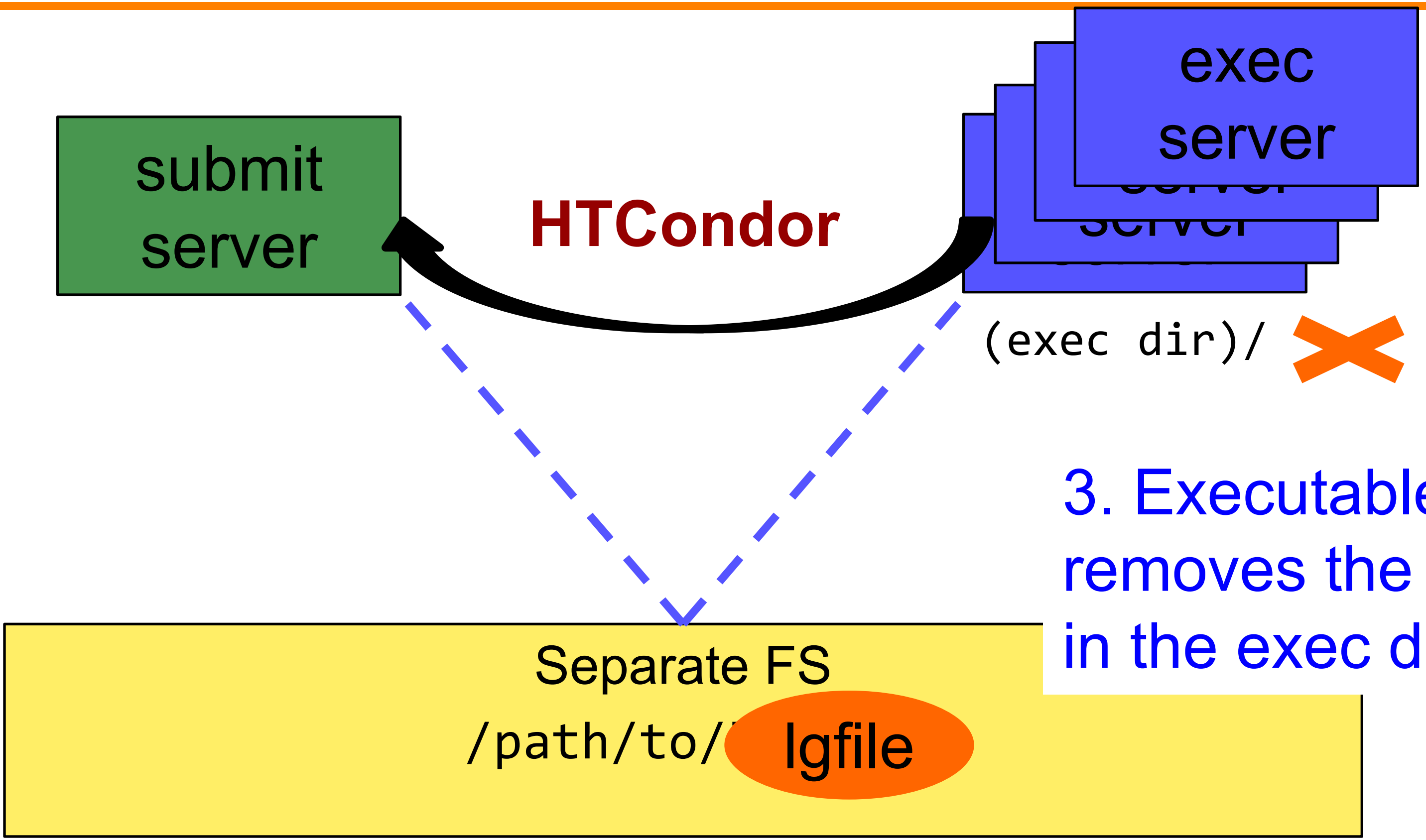
(exec dir)/ lgfile

2. Executable copies the file

Separate FS

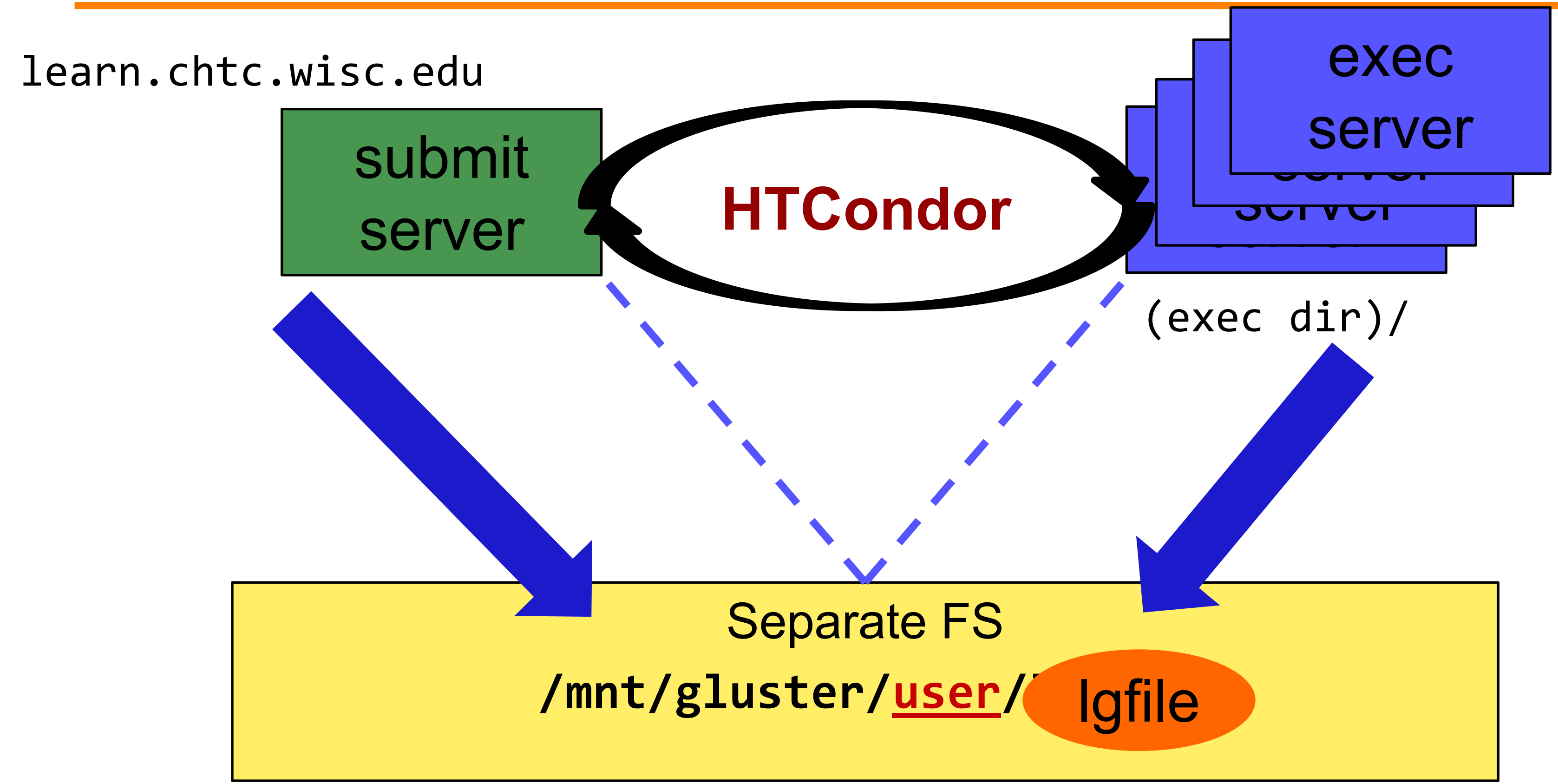
/path/to/lgfile

Separate shared FS - Output



3. Executable removes the file in the exec dir

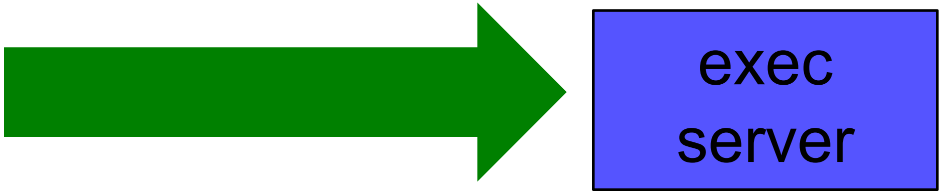
At UW-Madison (Ex. 3.1-3.2)



Shared FS Configurations

1. Submit directories *WITHIN* the shared filesystem
   - most campus clusters
   - limits HTC capabilities!!
2. Shared filesystem separate from local submission directories
   - supplement local HTC systems
   - treated more as a repository for VERY large data (>GBs)
3. Read-only (input-only) shared filesystem
   - Treated as a repository for VERY large input, only

Large input in HTC and OSG



file size	method of delivery
words	within executable or arguments?
tiny – 10MB per file	HTCondor file transfer (up to 1GB total per-job)
10MB – 1GB, shared	download from web proxy (network-accessible server)
1GB - 20GB, unique or shared file	StashCache (regional replication)
20 GB – TBs, unique or shared	**shared file system (local copy, local execute servers)**

Output for HTC and OSG

exec server

file size	method of delivery
~~words~~	~~within executable or arguments?~~
tiny – 1GB	HTCondor file transfer (up to 1 GB total per-job)
1GB+	**shared file system (local execute servers)**

Open Science Grid

Review

Option	Input or Output?	File size limits	Placing files	In-job file movement	Accessibility?
HTCondor file transfer	Both	100 MB/file (in), 1 GB/file (out); 1 GB/tot (either)	via HTCondor submit node	via HTCondor submit file	anywhere HTCondor jobs can run
Web proxy	Shared input only	1 GB/file	specific to VO	HTTP download	anywhere, by anyone
StashCache	Shared and unique input	20 GB/file (will increase!)	via OSG Connect submit server	via `stashcp` command (and module)	OSG-wide (90% of sites), by anyone
Shared filesystem	Input, likely output	TBs (may vary)	via mount location (may vary)	use directly, or copy into/out of execute dir	local cluster, only by YOU (usually)

Exercises

- 3.1 Shared Filesystem for Large Input
- 3.2 Shared Filesystem for Large Output

Questions?

- Next: Exercises 3.1-3.2
- Later: Job workflows