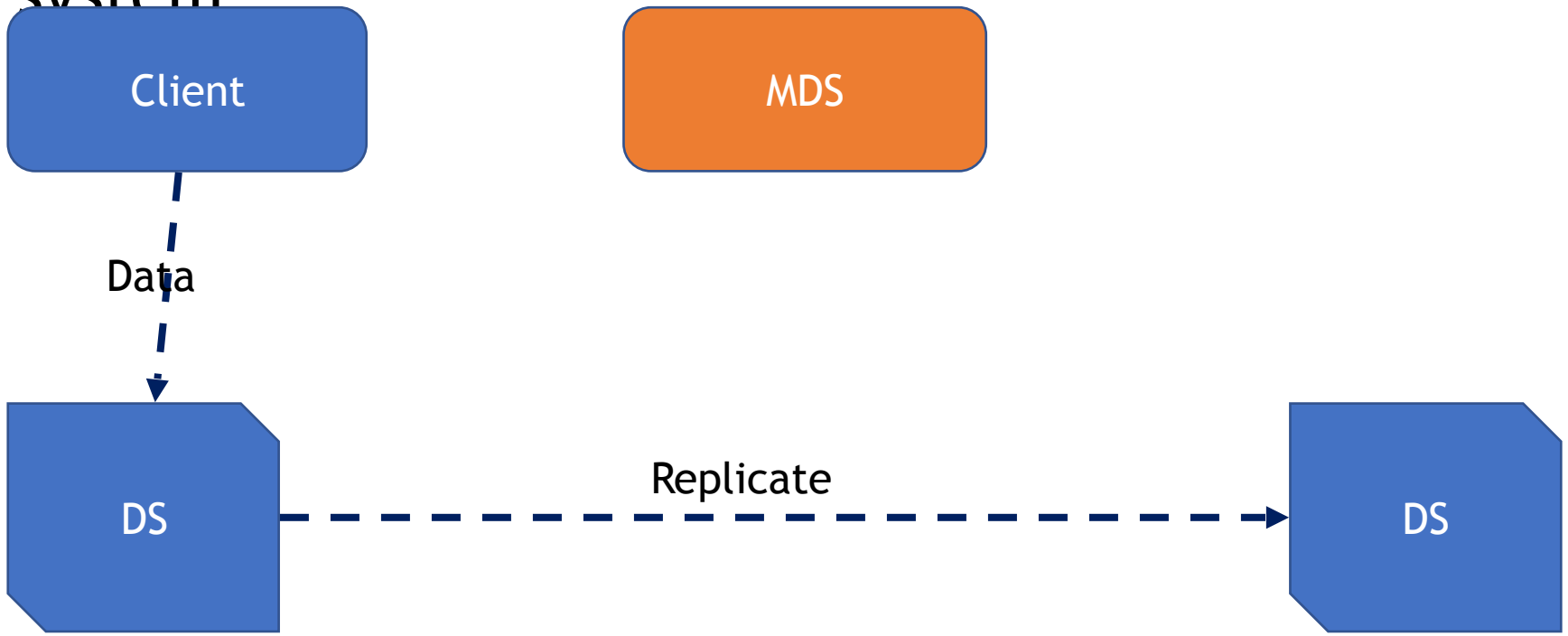
Where to next?
NFSv4 WG
IETF99

Trond Myklebust
trond.myklebust@primarydata.com
Tom Haynes
loghyr@primarydata.com

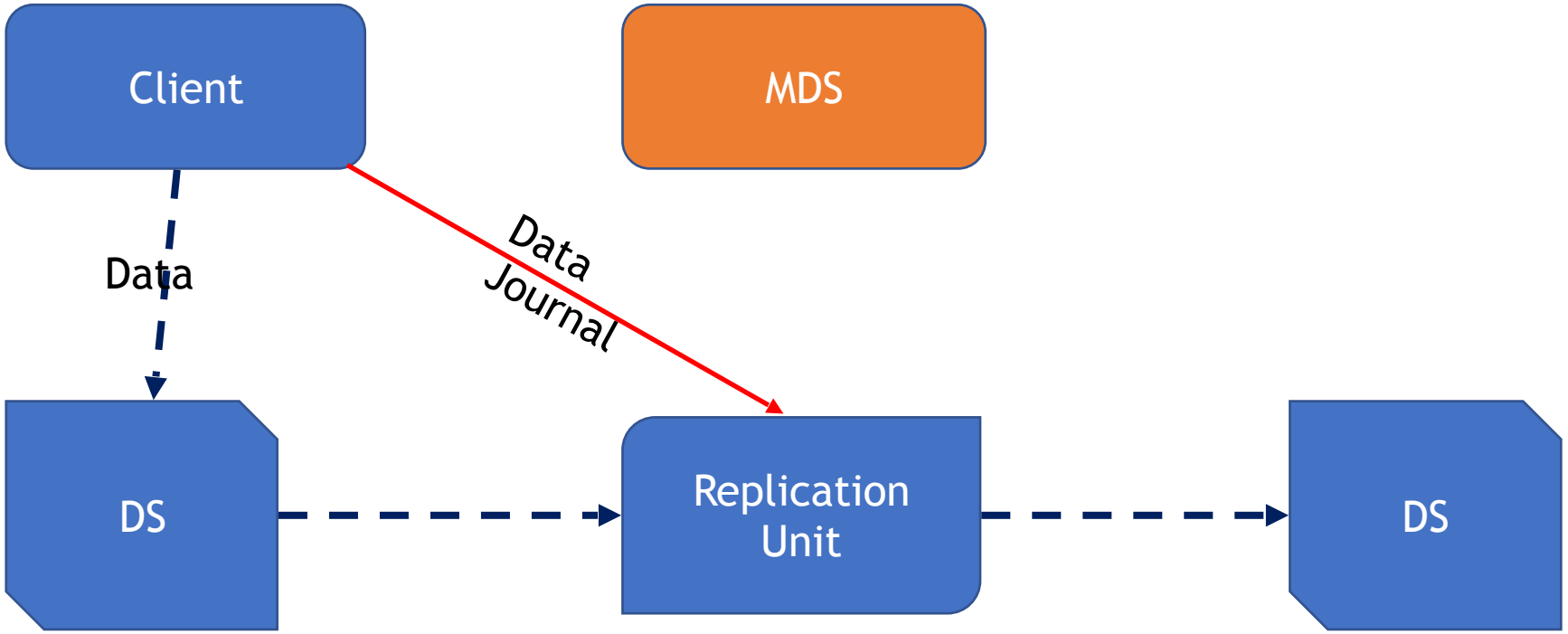
Topic: Scale out & replication

- Looking to address the problem of how to do efficient asynchronous replication of data when using a loosely coupled flexfiles system



Topic: Scale out & replication

- Proposal is to allow lightweight journaling of the data to a ”data replicator” which performs the actual copy.

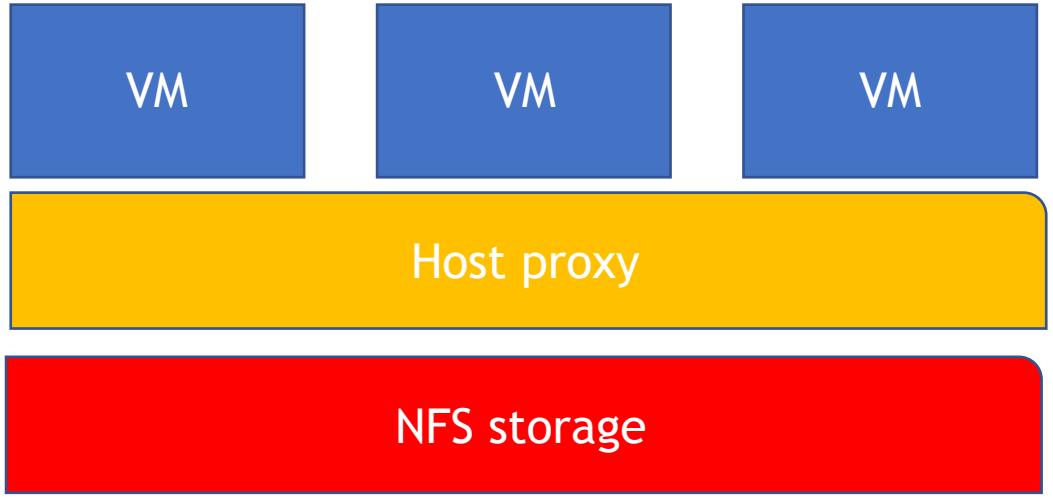


Topic: NFS as a service & quality of service

- One problem to solve is that of allowing certain clients to access a critical set of files/data with higher priority, and to allow the dynamic allocation of resources (e.g. bandwidth) to allow them to do so.
  - Proposal is to add protocol features to allow the MDS to communicate to each client a set of resource limits, and to allow it to change those on the fly without recalling layouts.
  - Resource limits could be set on a per layout, per filesystem, per DS,…
- Another problem is to allow for temporary quiescing of I/O to enable (for instance) group consistent snapshots.

Topic: Virtualisation & VM migration

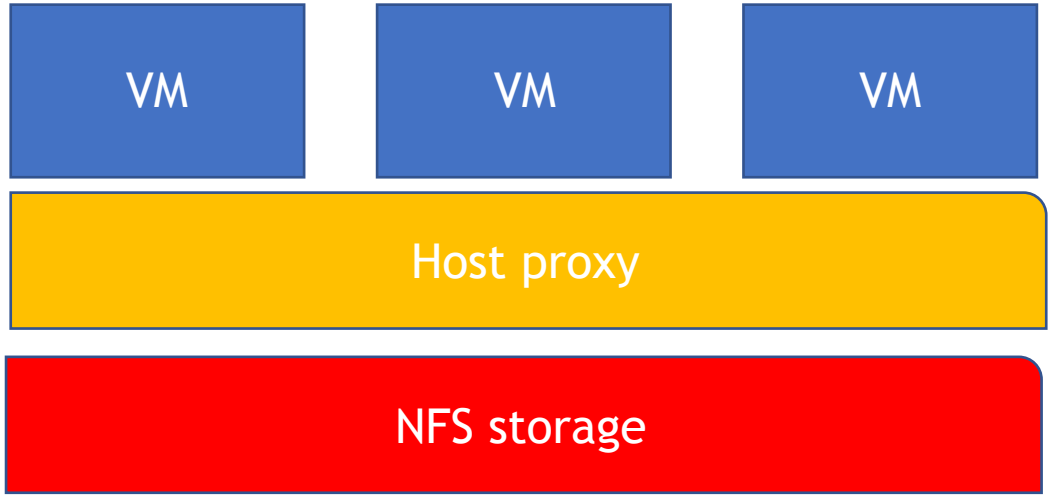
- Looking to enable architectures such as the following:



- Any one VM talks NFS to the host through a zero-config communication channel (e.g. the “vsock” protocol).

Topic: Virtualisation & VM migration

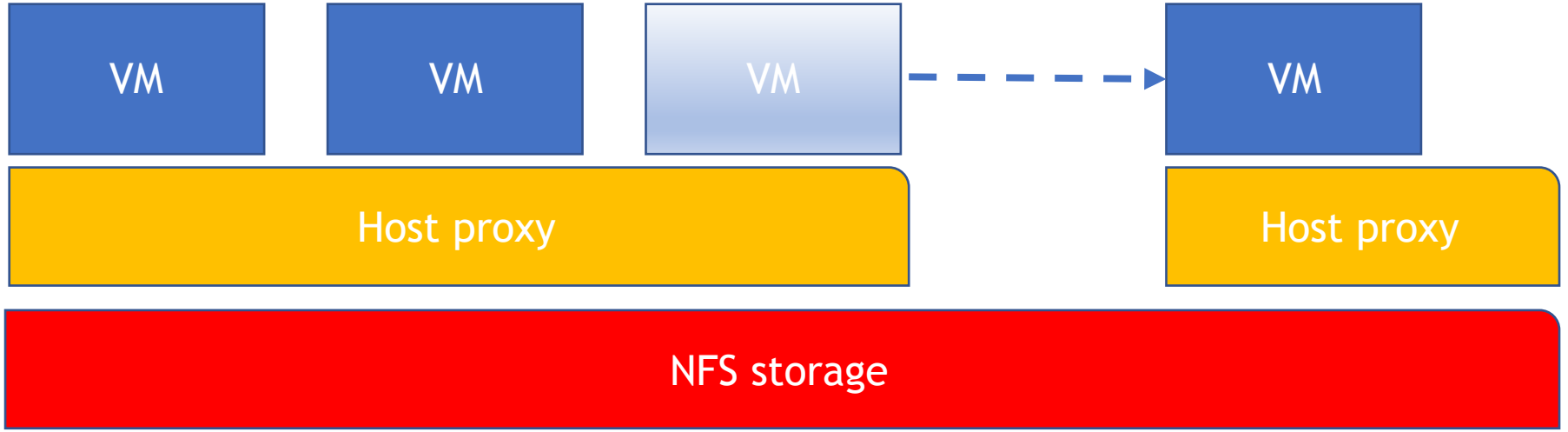
- Looking to enable architectures such as the following:



- The host acts as a proxy for the underlying NFS storage, insulating the clients from the details of the IP protocol.

Topic: Virtualisation & VM migration

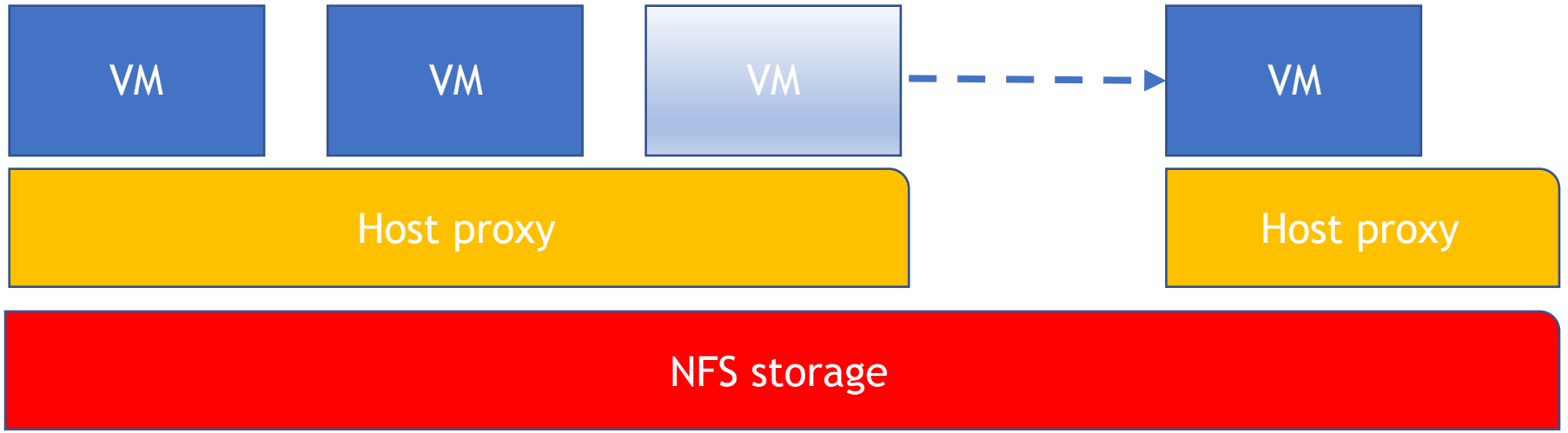
- Looking to enable architectures such as the following:



- Want to allow VMs to be transparently migrated to new host that is proxying the same NFS storage.

Topic: Virtualisation & VM migration

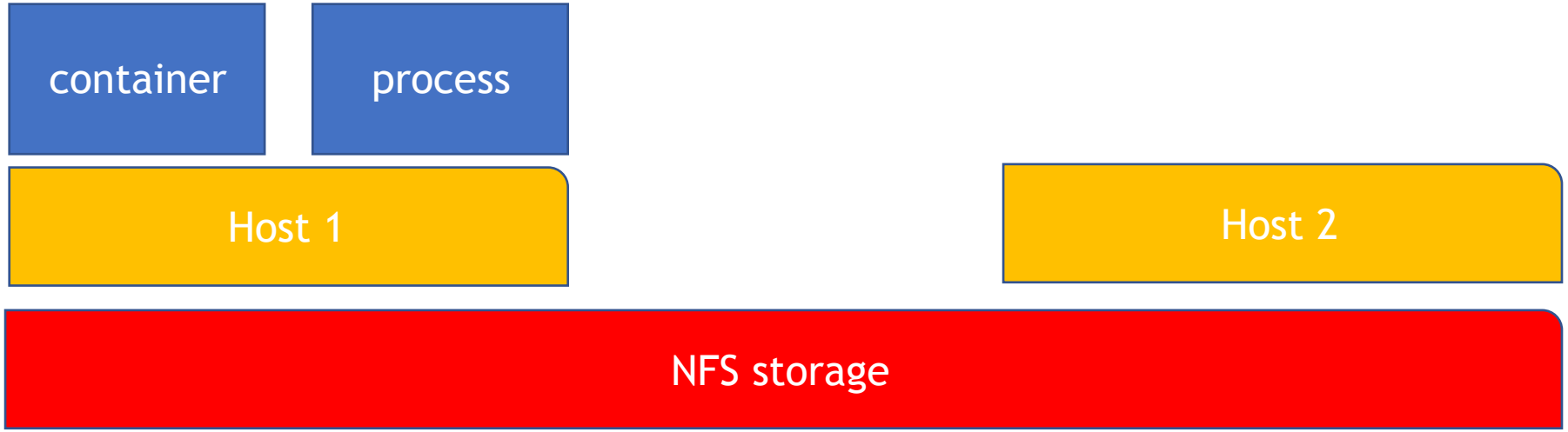
- Looking to enable architectures such as the following:



- Proposal is to include vsock channels as a supported transport for RPC, and NFSv4.

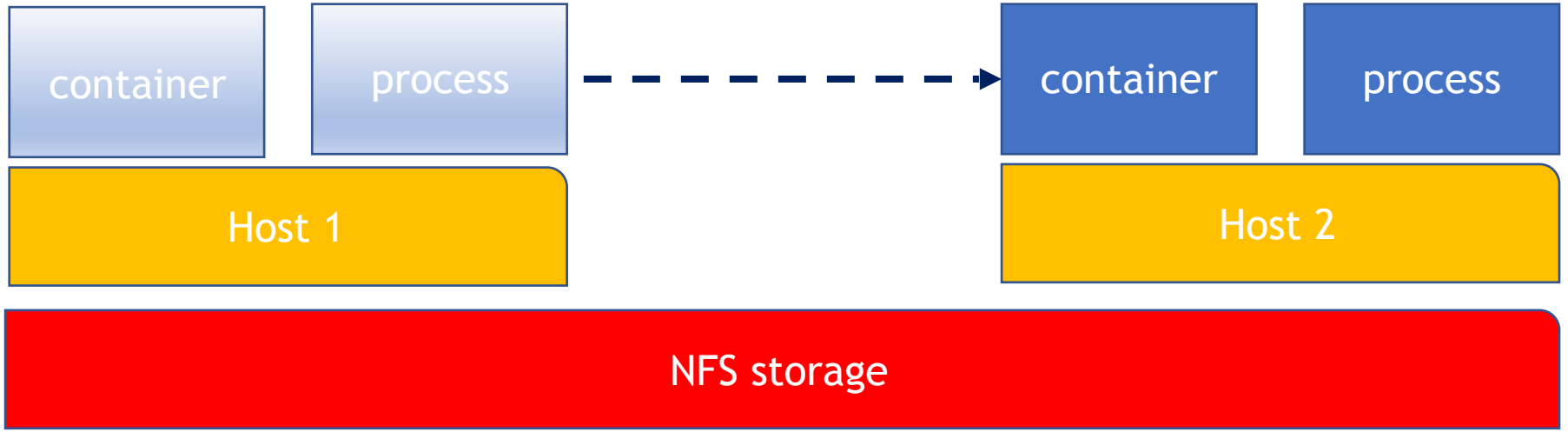
Topic: Container and process migration

- Also want to migrate containers and even individual processes:



Topic: Container and process migration

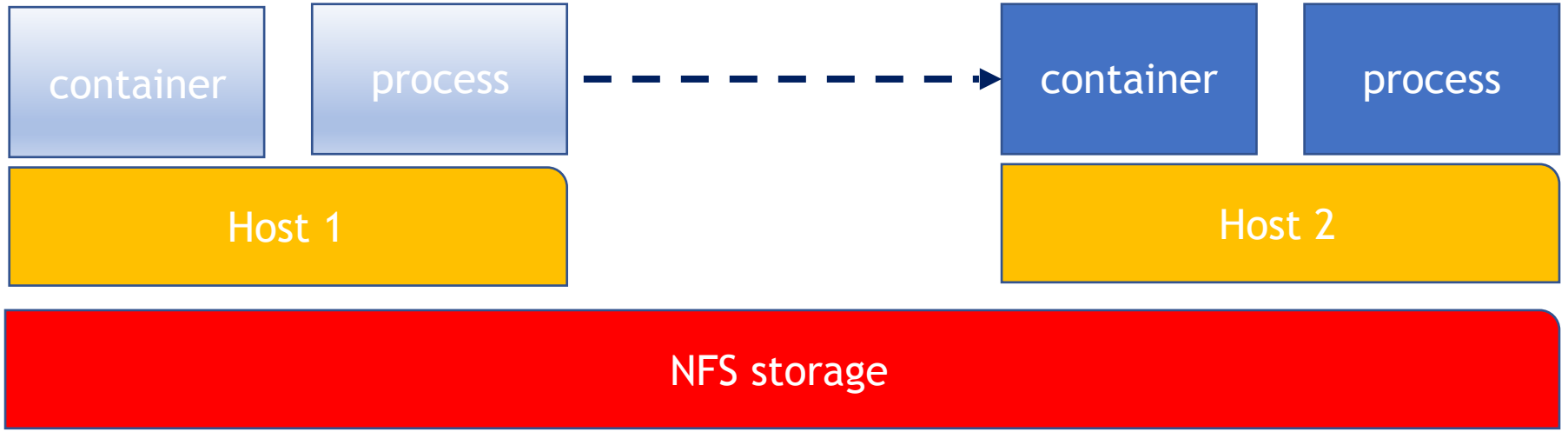
- Also want to migrate containers and even individual processes:



- Question is how to migrate file/lock state in these conditions.

Topic: Container and process migration

- Also want to migrate containers and even individual processes:



- Question is how to migrate file/lock state in these conditions.
- Defining a lease per process is onerous…

Topic: Misc NFSv4 protocol bugs and issues

- OPEN returns either an open stateid or an open stateid + a delegation. Why do we need both?
  - Add an OPEN share flag to specify “I’d like a delegation, or an open stateid if the delegation is not available”?
- READDIR is still hard to implement for filesystems because it is stateless. In practice that means cookies need to be persistent.
- Atomic append support? Still a useful functionality for logs and journals.

Topic: Misc RPC issues

- It would be useful to add a system of annotations to the RPC header. Two annotations come immediately to mind.
  - Tracing support. Allow the caller of an RPC to annotate the call with debugging/tracing information, such as process pid, etc.
  - Bulk data annotations. Allow the RPC call to specify data areas within the call which correspond to bulk data (e.g. NFS READ or WRITE data) and that might benefit from special treatment when reading from the socket (e.g. copying into aligned buffers to enable zero-copy placement into a page cache etc.)