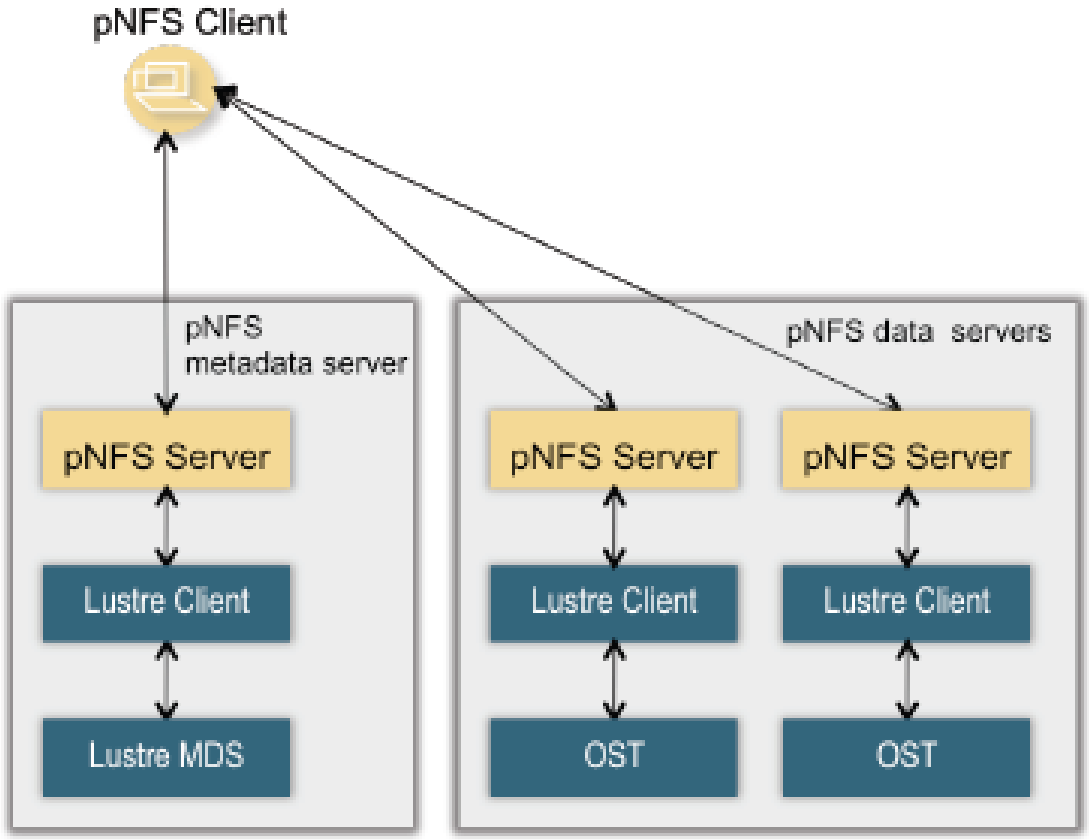
New pNFS layout for Lustre

February 21, 2012
Cthon 2012 – Santa Clara

Lustre & pNFS

- pNFS is somewhat similar with Lustre in architecture; use layouts
  - Lustre pNFS exports support (pNFS native) File Layout, so there is no need for extra layout drivers on all of your pNFS clients to get benefits.
- To minimize Lustre-network latency with pNFS
  - Put NFS MDS on lustre MDS and NFS Data servers on Lustre OSTs.
  - It is also possible to have dedicated pNFS servers separate from Lustre server nodes (at some performance penalty).

Lustre/pNFS exporting structure



Efficiency of I/O

- Reexport write bandwidth is good, we can actually saturate NFS link provided that there is enough bandwidth on a Lustre side.
- Reexport read bandwidth is mostly good, not ideal but in general we saw around 50% of write speeds.
- Reexport read bandwidth is not 100% consistent yet, mostly believed to be due to readaheads conflicts.

Alternative implementation options

- Improve the file layout by removing Lustre client overhead
- Implement object layout adapted for Lustre for data servers and ? For MDS
- New layout based on Lustre layout

New Lustre specific pNFS layout

- Goal to use the Lustre server unmodified
- Put a shim pNFS layer for layout translation on top of Lustre client that will be included in the Linux kernel; use same caching as Lustre
- Add a shim layer to the Lustre MDS that will allow multiple/cluster pNFS MDS's
- Improve HA and allow POSIX support that Lustre doesn't
- Improve MD operations and small files

Plan of action

- Write new pNFS layout draft
- Port Lustre client to Linux kernel
- CITI will analyze alternatives with EMC and others