



Lustre & NFS/pNFS

- Oleg Drokin
- Lustre Group



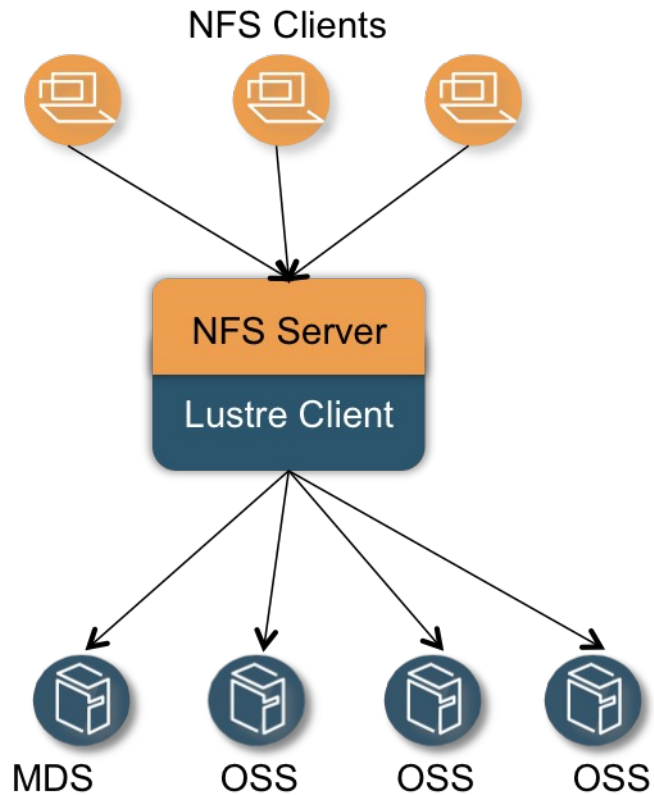
Agenda

- General Lustre NFS export info
- Lustre -> NFS operations translation
- PNFS & Lustre
- Getting the most out of your Lustre-NFS exports (tips)

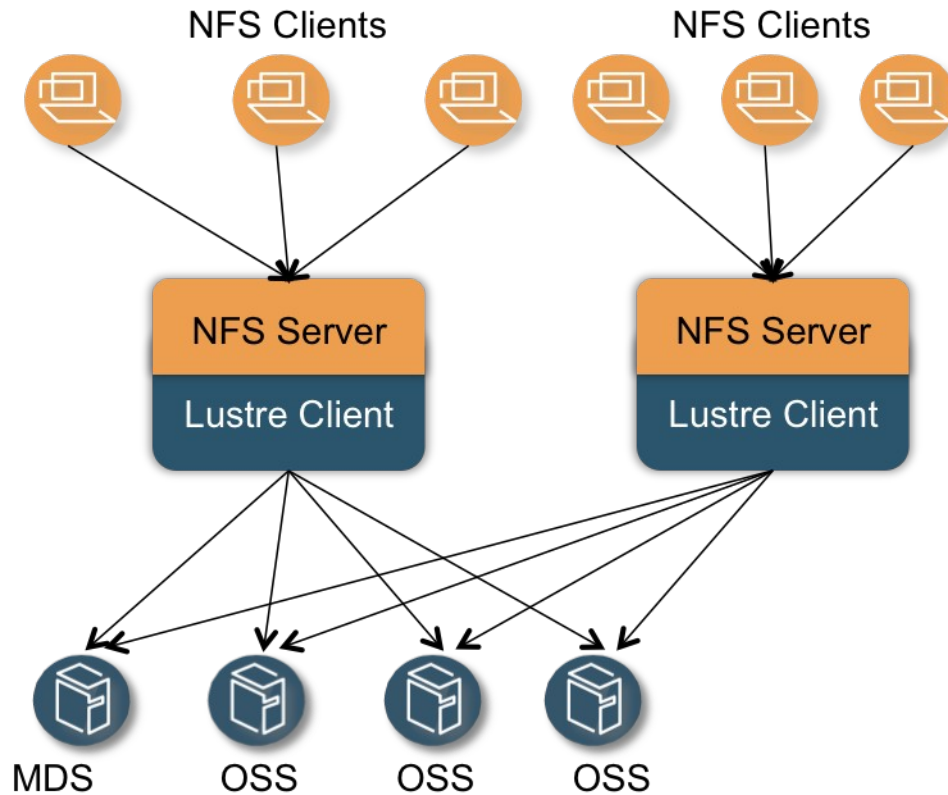
General Info

- Why NFS is important.
- Lustre is a normal local fs from NFS perspective
 - > Just add an entry to `/etc/export` on lustre client/NFS server
 - Older Lustre releases require to specify `fsid`
- Lustre is in fact not your normal local fs, it is distributed and there are possibly many clients
 - > NFS file locking prior to 2.6.22 assumed local fs only
 - > You can export same Lustre FS from several NFS servers to get more bandwidth
 - > NFS servers failover

Exporting Structure



Exporting Structure



Lustre NFS exports – work in progress

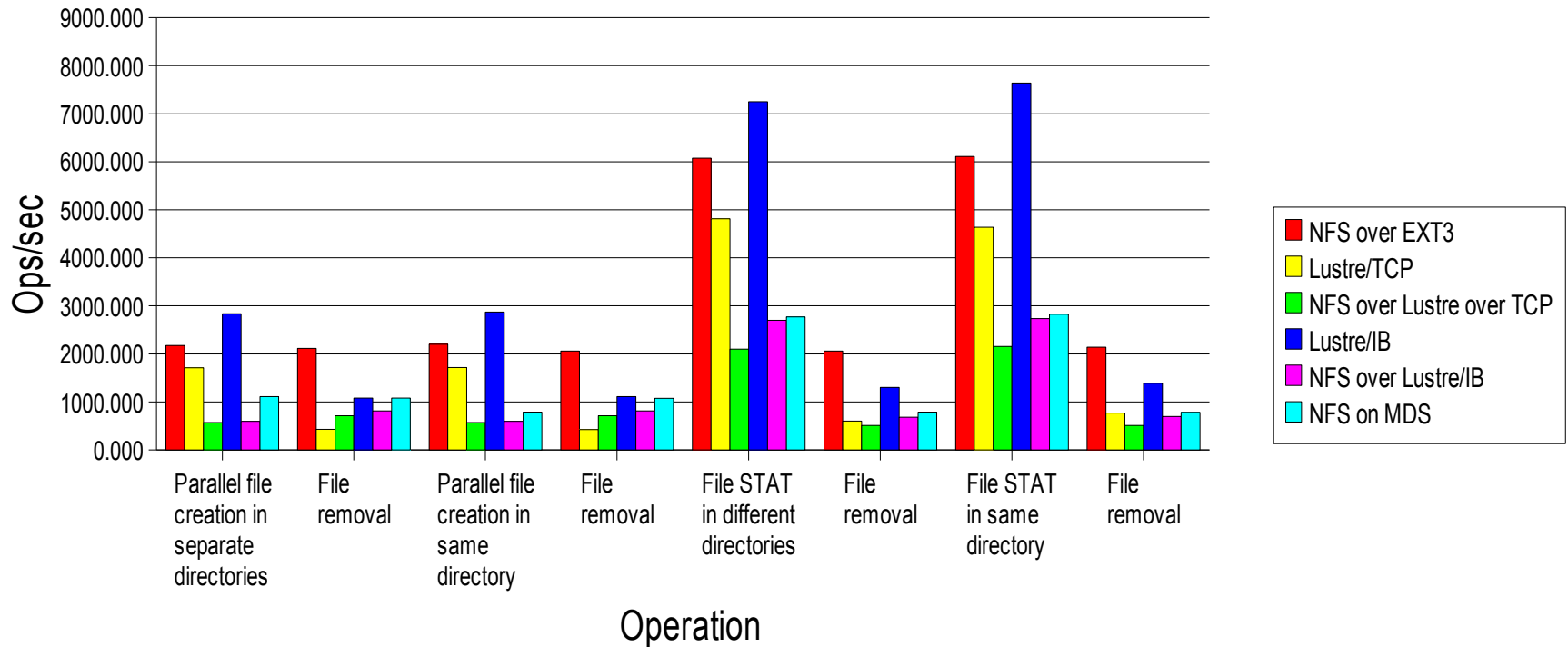
- Improve NFS exports performance
 - > up to 90x write speedup on all 2.6 Linux kernels with Lustre 1.6.5 (done)
 - > Better metadata operations matching (work in progress)
 - > Read path improvements (work in progress)
- New NFS features are integrated into Linux kernel:
 - > Asynchronous file locking (done)
 - > NFS4 delegations (via leases) – (work in progress)
 - > pNFS support (work in progress, prototype available)

NFS to Lustre metadata ops translation

NFS metadata operation	No. of Lustre metadata operations
Lookup	2+
mknod/mkdir/symlink	2
open+create	1(2 pre 2.6.15)
open	1
readdir	1
readdirplus	$1 + \text{Numentries} * (\text{numstripes} + 1)$
unlink	2
rename	3
statfs	1

Metadata comparison

Metadata performance



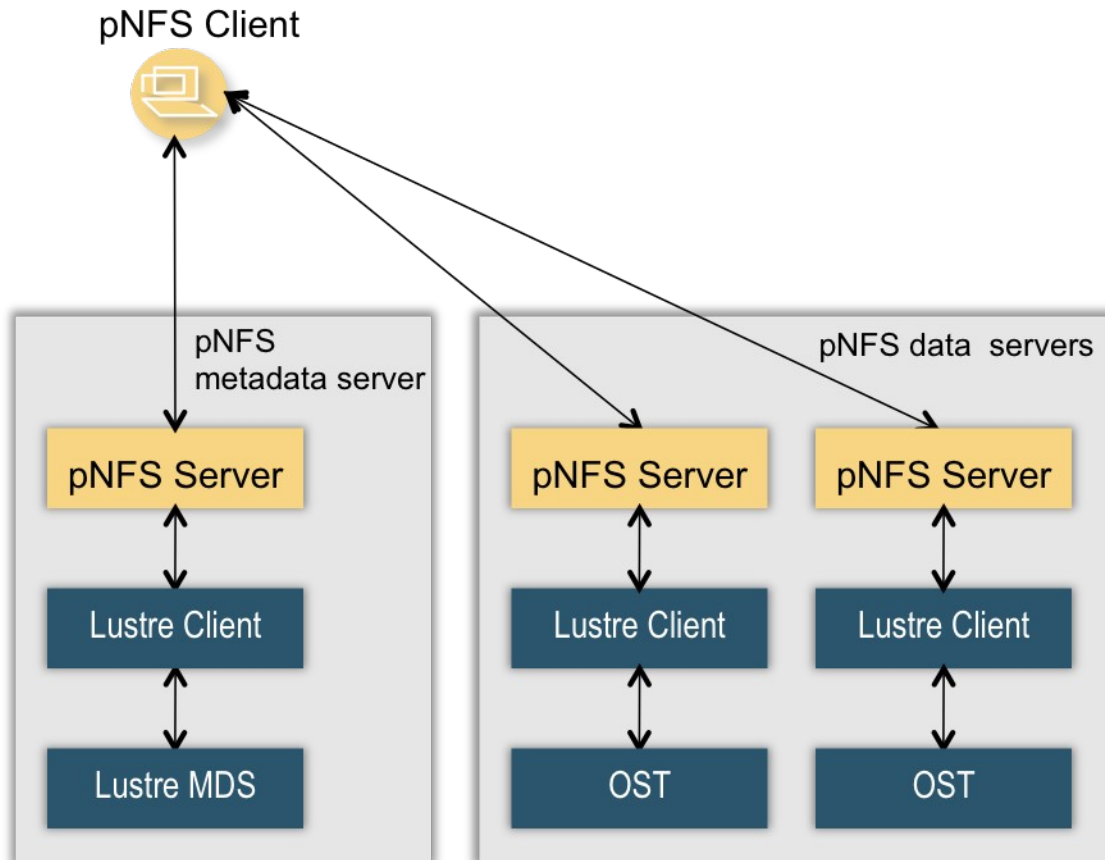
pNFS – why and how

- One NFS server is a bottleneck
 - > One disk is even bigger bottleneck
- Spreading NFS workload across several servers is cumbersome with NFSv3 and impossible with NFSv4
- pNFS – more scalable NFS
 - > Different files and parts of files on different NFS server nodes, even non-NFS nodes
- pNFS is currently work in progress, there is no stable implementation of the protocol itself yet
 - > <http://www.citi.umich.edu/projects/asci/pnfs/linux/>

Lustre & pNFS

- pNFS is somewhat similar with Lustre in architecture
- 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. - work in progress.

Other performance tips

- Disable Lustre debug - it really hurts with NFS
- Readahead (significant impact on reads, but no rule for now, requires some experimentation)
- Async operations – can give you extra 10% on writes at the expense of data safeness
 - > “async” option in /etc/exports
- Export from several nodes and let your clients do DNS round robin (or static assignment) to avoid bottlenecks
- Run Lustre and NFS on different physical networks

Questions?



Oleg.Drokin@sun.com

