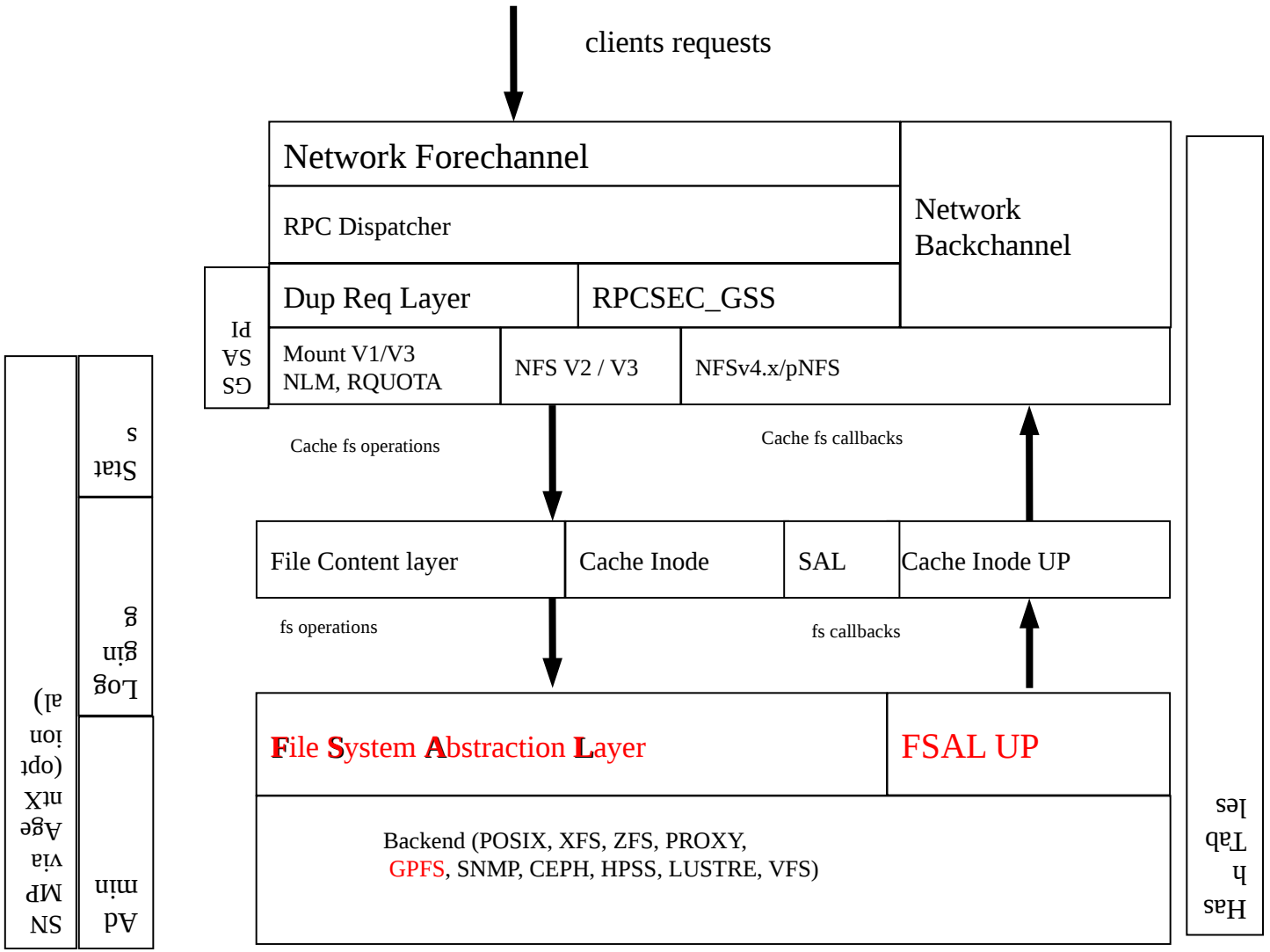
NFS-Ganesha w/GPFS FSAL And Other Use Cases

Marc Eshel, JV, Jim, Philippe
2/26/2013

The modules within NFS-Ganesha



GPFS FSAL

- FSAL to GPFS posix like interfaces
  - Open, close, read, write, remove, ...
  - File Descriptor requires open that can cause
    - Lost shares or op-locks
  - Use File Handle for most operations
  - Use FS support of NFS export ops
  - Convert FH to Linux dentry and use VFS
  - Combine few operation into one
    - Set Share reservation with open
    - Get file attributes and ACLs

GPFS FSAL UP_CALLS

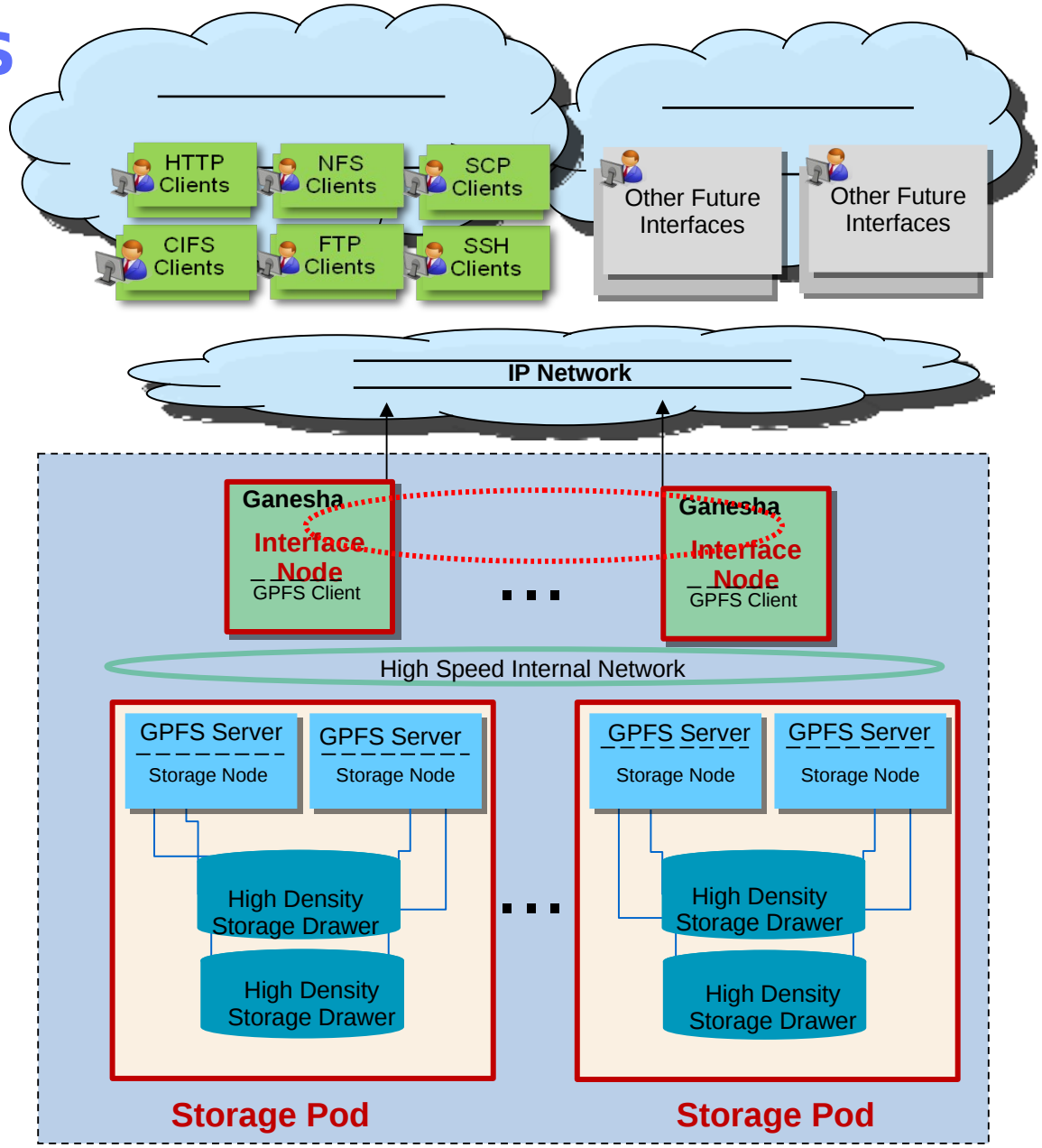
- Ganesha has an Inode cache and other state like Delegations and pNFS Layouts
- One thread per exported FS to maintain state in cluster environment or other access to same fs
- GPFS to FSAL up-calls
  - Inode invalidate or update, grant locks
  - Recall Delegations or pNFS Layouts
- Immediate up-calls, queued, or both
  - Inode invalidate is immediate
  - Layout recall is both immediate and queued

GPFS FSAL pNFS

- FSAL to GPFS pNFS interface
  - On MDS Layout get/return/commit
    - Get device-info device-list
    - Recall Layout device-info
  - On DS read/write/commit
- MDS DS backchannel communication
  - FH given by MDS include its ip address
  - DS check with MDS before I/O
    - Open state and Layout information
  - MDS will notify DS when file is closed

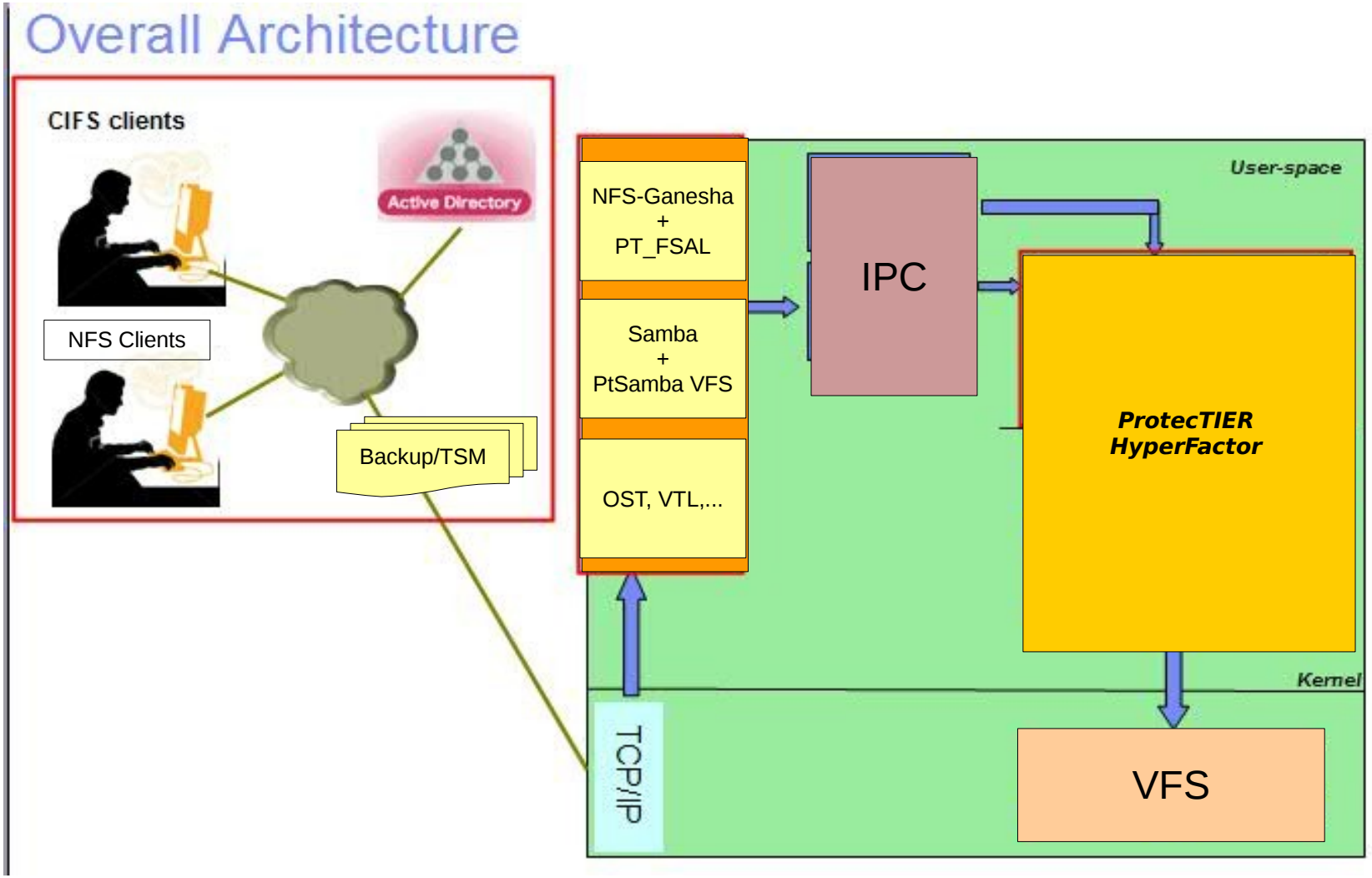
Use Case - SoNAS

High level design of Cloud Storage infrastructure



- 'Lego-like' modular design allows online independent scaling of I/O capacity and Storage capacity

UseCase - ProtecTIER



Panasas Distributed Storage

- Current Product (ActiveStore 14)
  - DirectFlow == pre-pNFS distributed object storage
  - Object storage (pre-OSD-OSC iSCSI)
- Director Blade
  - PanFS meta-data server
  - NFS V3 + CIFS
- Storage Blade
  - OSD (Direct Flow + PanFS)
  - 2x 4T disks (+ 400GB SSD)

Panasas pNFS

- DirectFlow + NFS 4.0 +4.1 pNFS
- Director Blade
  - PanFS meta-data server
  - Ganesha (NFSv3, NFSv4.0, 4.1 metadata)
- Storage Blade
  - DirectFlow OSD
  - OSD-OSC iSCSI
  - pNFS data server w/ object layout storage

Ganesha @ CEA

- Used as a NFSv3/NLM server in part fo a large cluster at the TGCC/CCRT facility to export LUSTRE.
- Used at CEA (at the TERA facility) since 2006 for exporting the namespace of a HSM to compute clusters. - Now runs infornt of a petaflopic class cluster.
- Future compute machines will see an explosion of the number of CPU cores due to “manycore” model like Xeon Phi (and future evolution).
  - A single server is no more reasonable => need for “IO proxies”
  - The ratio of memory per core will drastically reduce => need for lightweight protocols
  - Currently, CEA is developing IO Proxy model in Ganesha by
    - Implementing 9p.2000L as a IO protocol (the protocol has fast and cacheless clients)
    - Using RDMA as transport protocol (zero copy is easy in 9p.2000L)

NFS-Ganesha pointers

- NFS-Ganesha is available under the terms of the LGPLv3 license

- NFS-Ganesha is hosted on SourceForge:
  - Project homepage
    - http://nfs-ganesha.sourceforge.net
  - Download page
    - http://sourceforge.net/projects/nfs-ganesha/files
  - Mailing lists
    - nfs-ganesha-devel@lists.sourceforge.net
    - nfs-ganesha-support@lists.sourceforge.net
    - nfs-ganesha-announce@lists.sourceforge.net