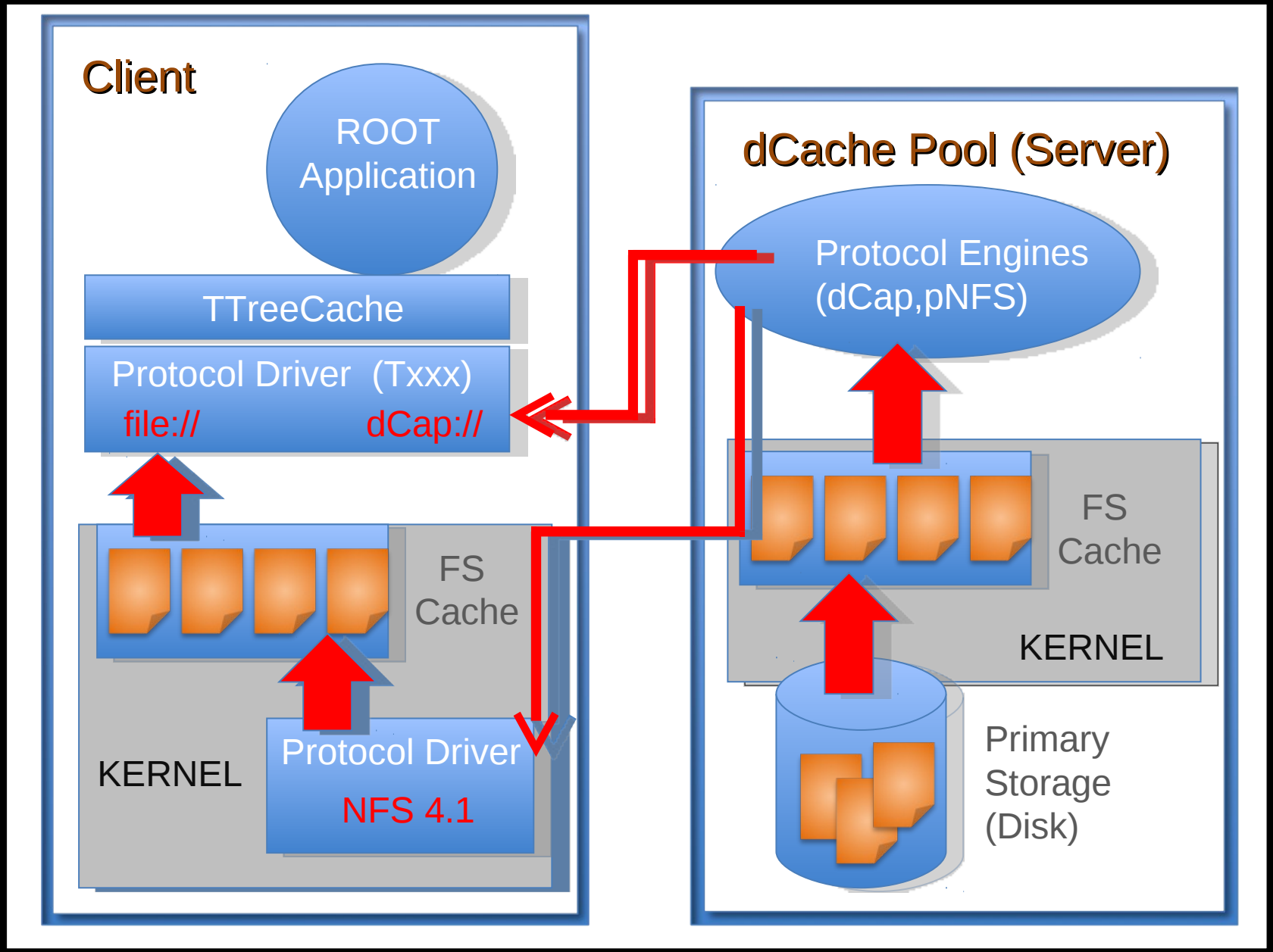
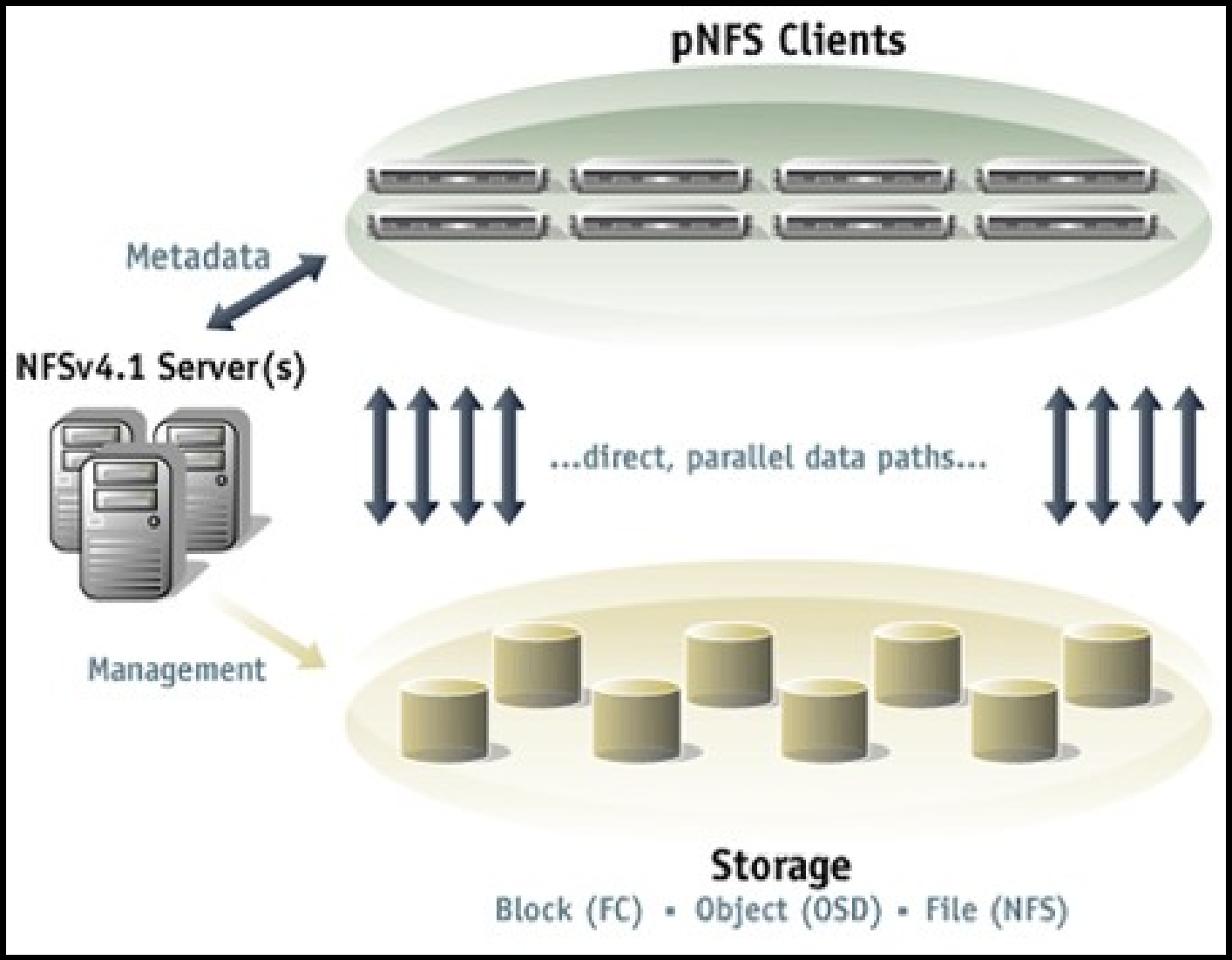


NDGF
Nordic DataGrid Facility


pNFS Clients
Metadata
NFSv4.1 Server(s)
...direct, parallel data paths...
Management
Storage
Block (FC) • Object (OSD) • File (NFS)


Sun
ORACLE


EMC²


panasas


IBM


NetApp


LCG


center for
information
technology
integration


European Middleware Initiative

High Latency Link Performance

Proper authentication and authorization

Massive Scalability

Mountable

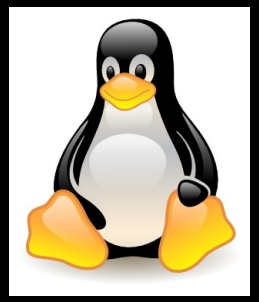
Standardized

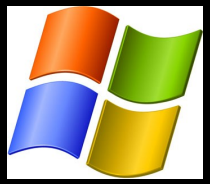
```
commit a4dd8dce14014665862ce7911b38cb2c69e366dd
Merge: b18cae4 411b5e0
Author: Linus Torvalds <torvalds@linux-foundation.org>
Date:   Tue Oct 26 09:52:09 2010 -0700

    Merge branch 'nfs-for-2.6.37' of
git://git.linux-nfs.org/projects/trondmy/nfs-2.6.git

    * 'nfs-for-2.6.37' of git://git.linux-nfs.org/projects/trondmy/nfs-2.6:
      net/sunrpc: Use static const char arrays
      nfs4: fix channel attribute sanity-checks
      NFSv4.1: Use more sensible names for 'initialize_mountpoint'
      NFSv4.1: pnfs: filelayout: add driver's LAYOUTGET and
GETDEVICEINFO infrastructure
      NFSv4.1: pnfs: add LAYOUTGET and GETDEVICEINFO infrastructure
      NFS: client needs to maintain list of inodes with active layouts
      NFS: create and destroy inode's layout cache
      NFSv4.1: pnfs: filelayout: introduce minimal file layout driver
      NFSv4.1: pnfs: full mount/umount infrastructure
      NFS: set layout driver
      NFS: ask for layouttypes during v4 fsinfo call
      NFS: change stateid to be a union
      NFSv4.1: pnfsd, pnfs: protocol level pnfs constants
      SUNRPC: define xdr_decode_opaque_fixed
      NFSD: remove duplicate NFS4_STATEID_SIZE
```

- Linux client since 2.6.32
- Solaris driver available, but not shipped with Solaris yet
- Windows driver exists, but not published yet
- Redhat has builds for Fedora 12, 13, rawhide with pNFS
- Redhat Enterprise Linux is expected to have pNFS in 6.1
- Possibility of backport to SL5



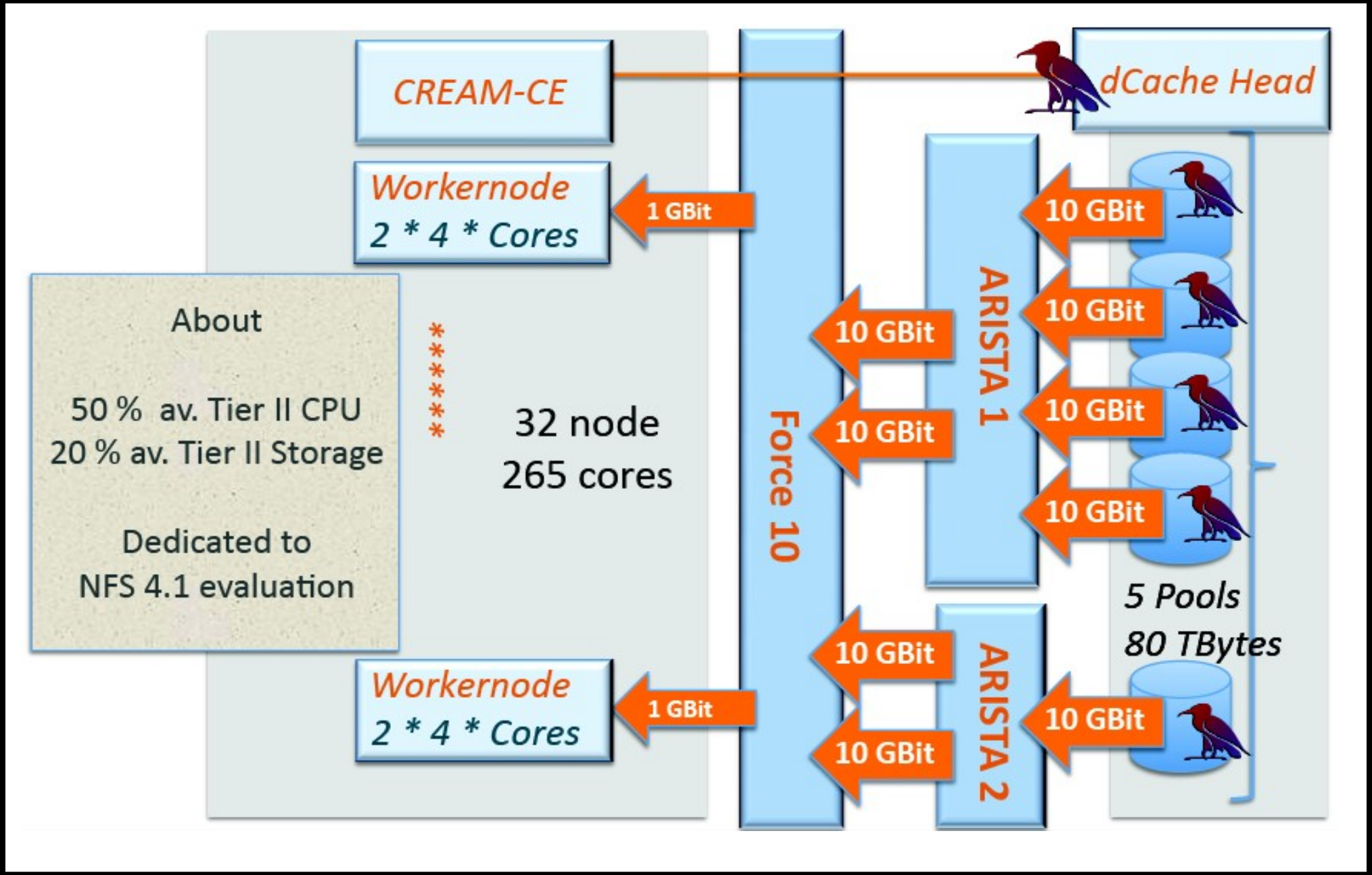


EMI

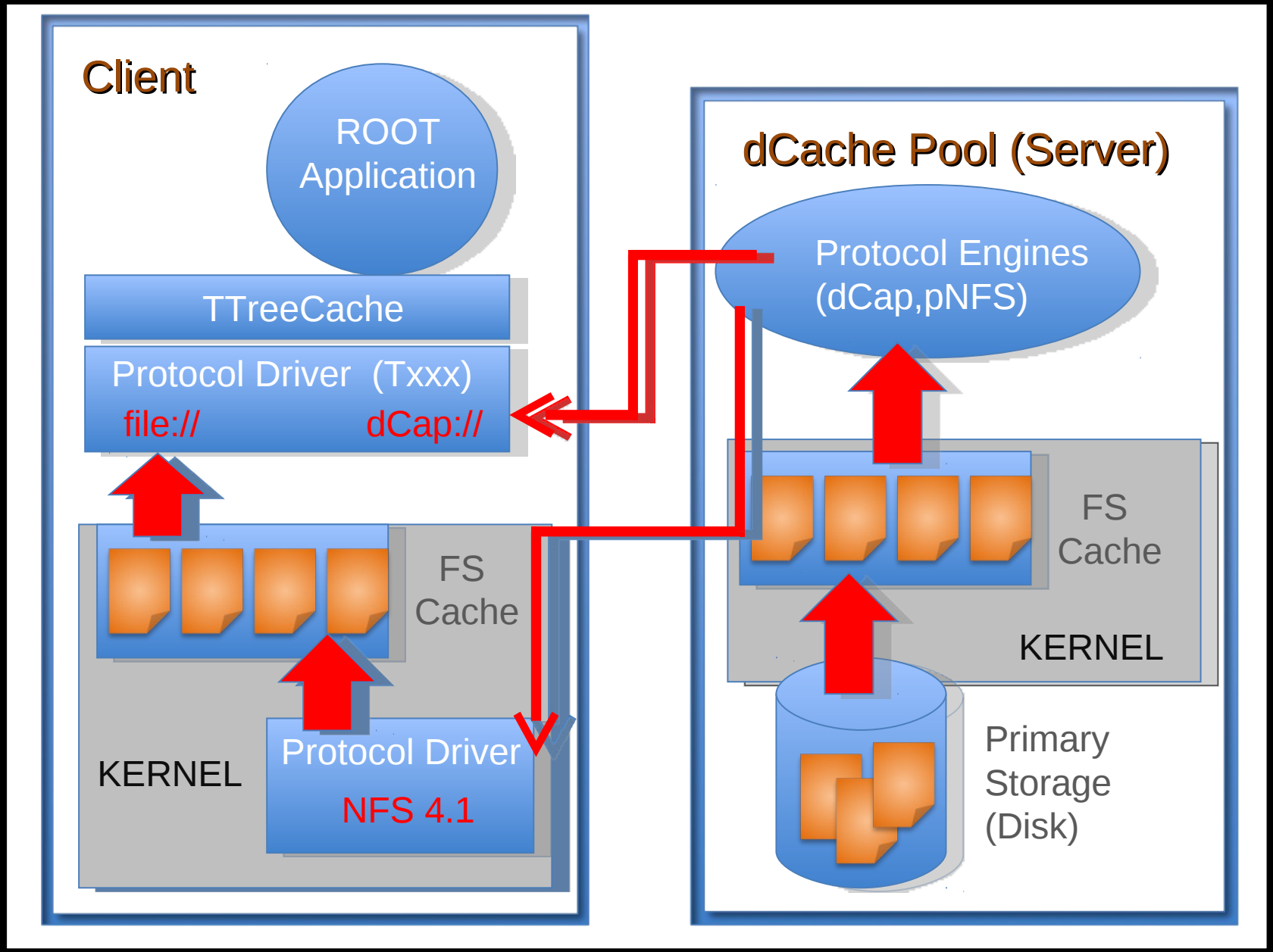
- dCache, DPM and likely Storm
- dCache 1.9.10 has production ready NFS 4.1
- DPM ready any day now

Industry

- Netapp, Panasas, Oracle, EMC, IBM and others have hardware products in the pipeline
- Waiting for broad client availability

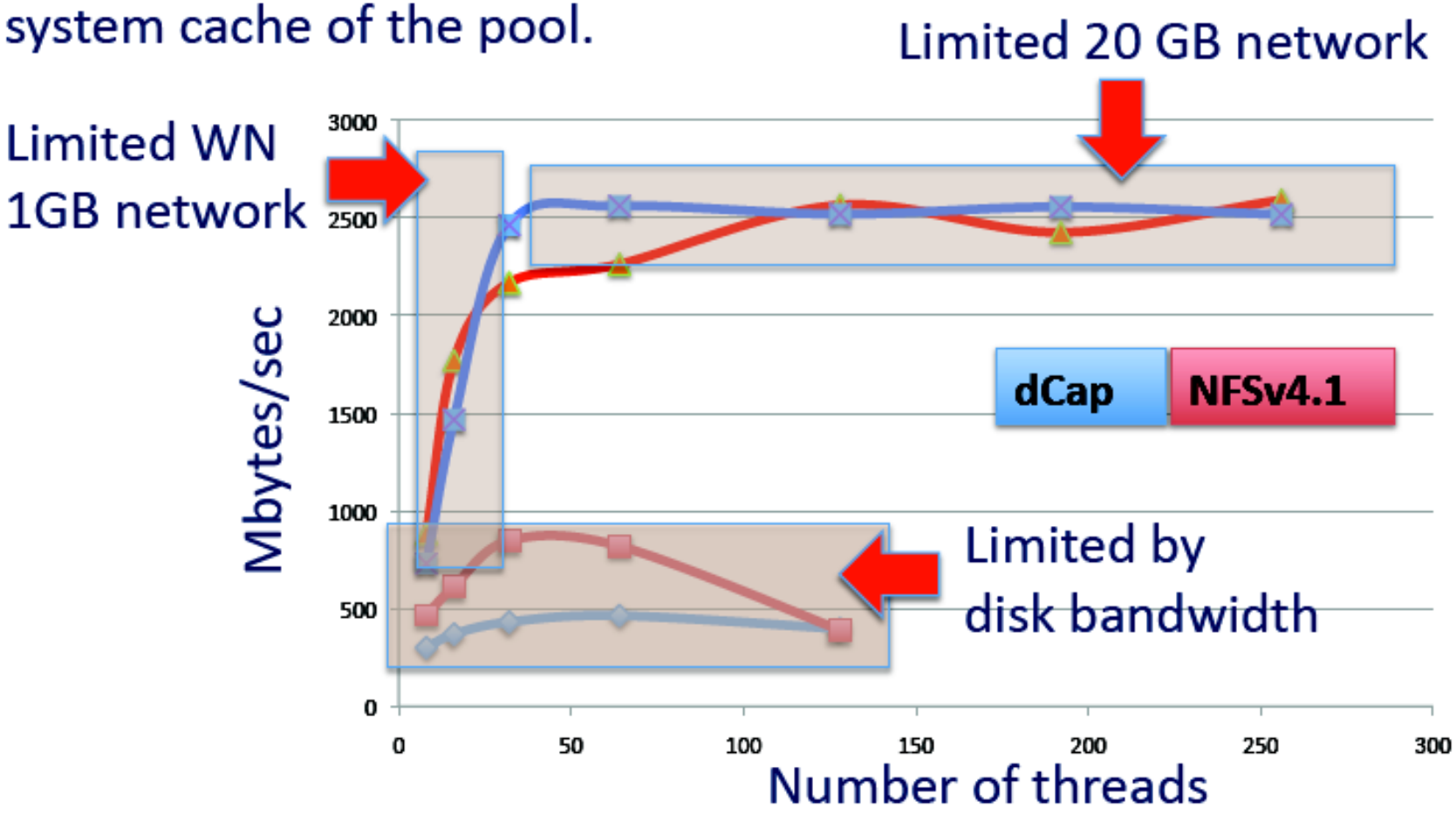


NDGF
Nordic DataGrid Facility



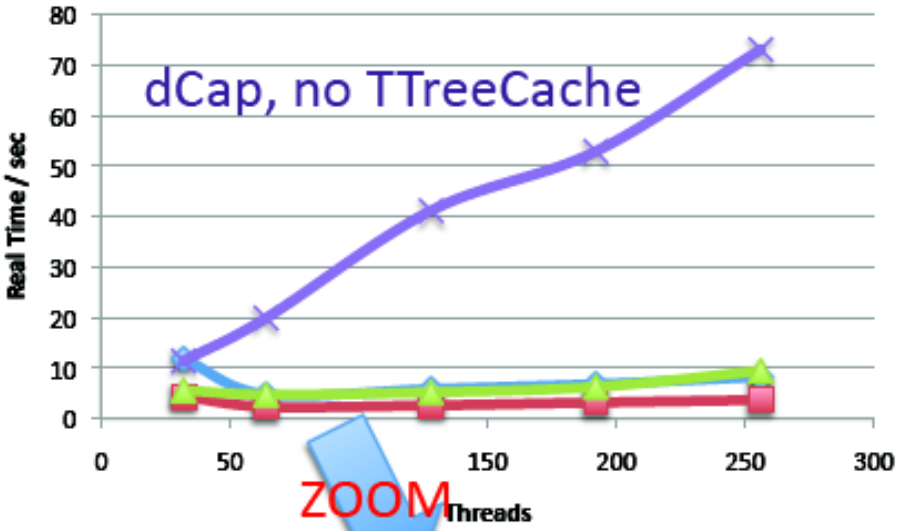
NDGF
Nordic DataGrid Facility

Removing server disk congestion effect by keeping all data in file system cache of the pool.

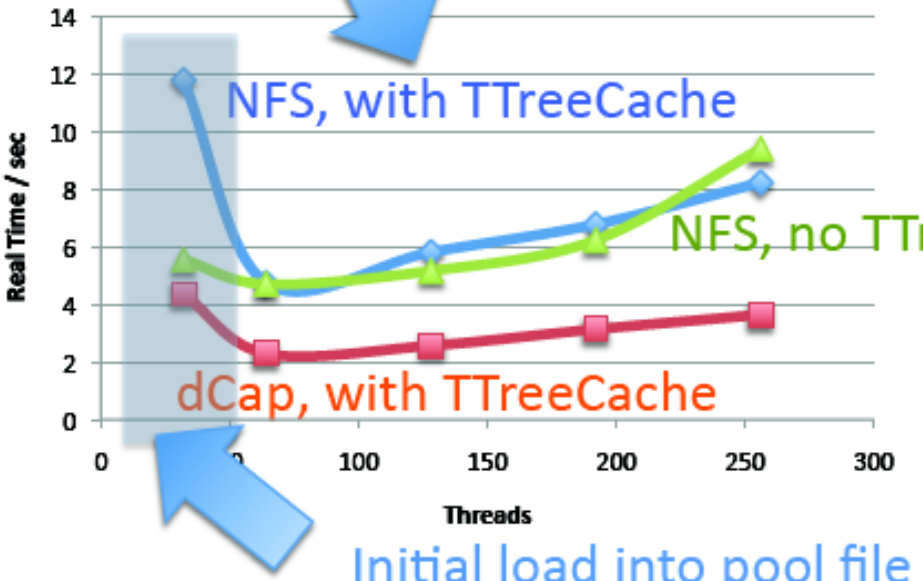


Total throughput doesn't depend on the protocol.

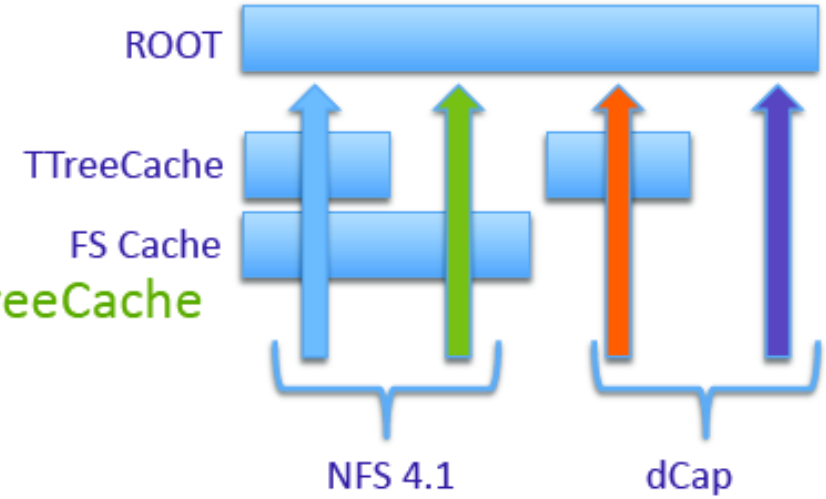
- New ROOT version 5.27.06, compiled with dCap support
- Files provided by René Brun: atlasFlushed.root (re-organized files with optimized buffers) and AOD.067184.big.pool_4.root (some other original file) (optimized: 1GByte, original 1.3 GByte)
- Test script provided by René: simple script reading events: taodr.C
- Different test runs:
  - Reading via NFS or dCap
  - Reading with 60MByte TreeCache, or with 0Byte TreeCache
  - Reading all branches or only 2 branches
  - 32, 64, 128, 192 or 256 jobs running in parallel



ZOOM

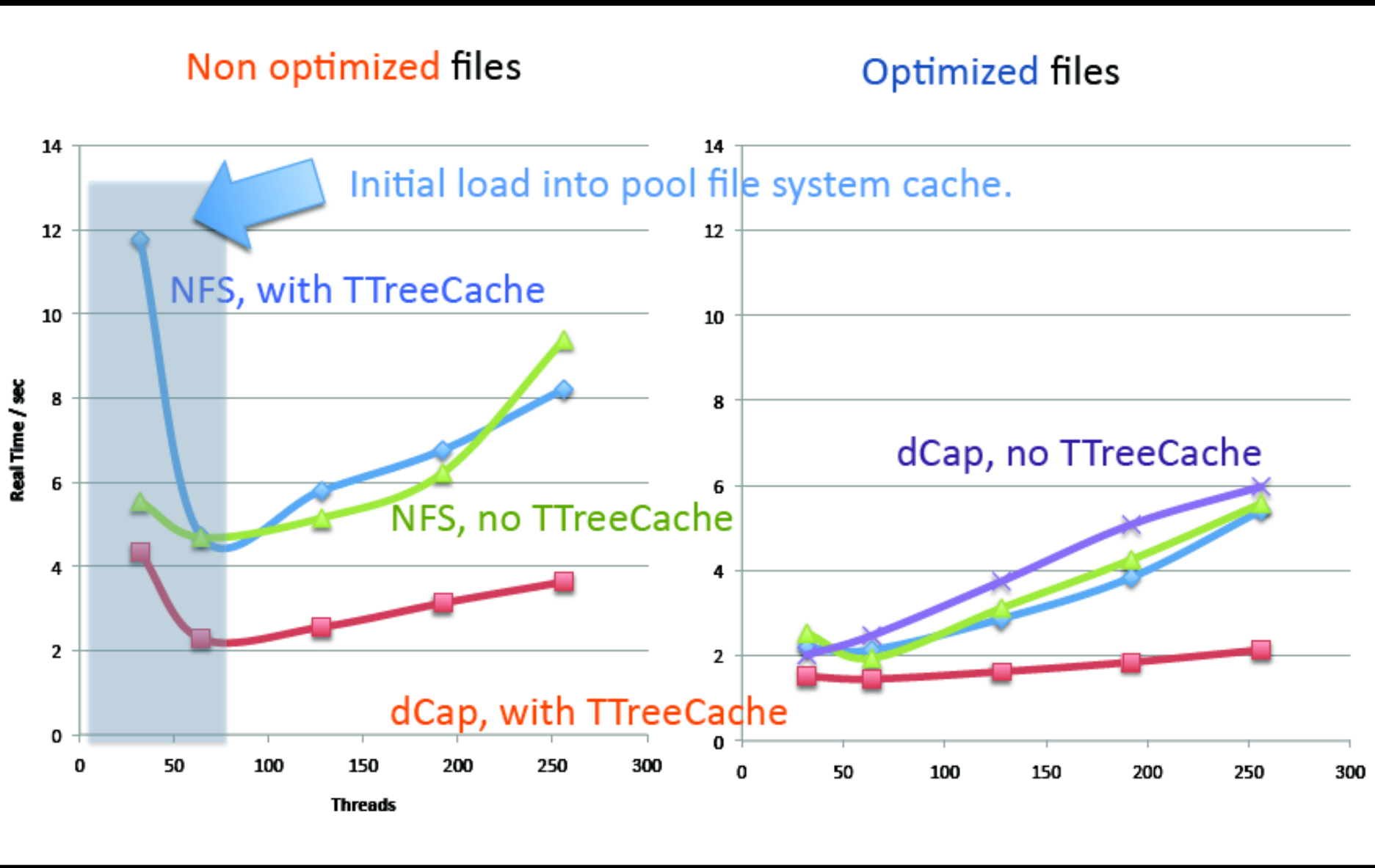


- ✓Non optimized files
- ✓Reading only 2 trees.
- ✓TTreeCache does vector read with dCap.
- ✓VR = fadvise disabled in ROOT for NFS.

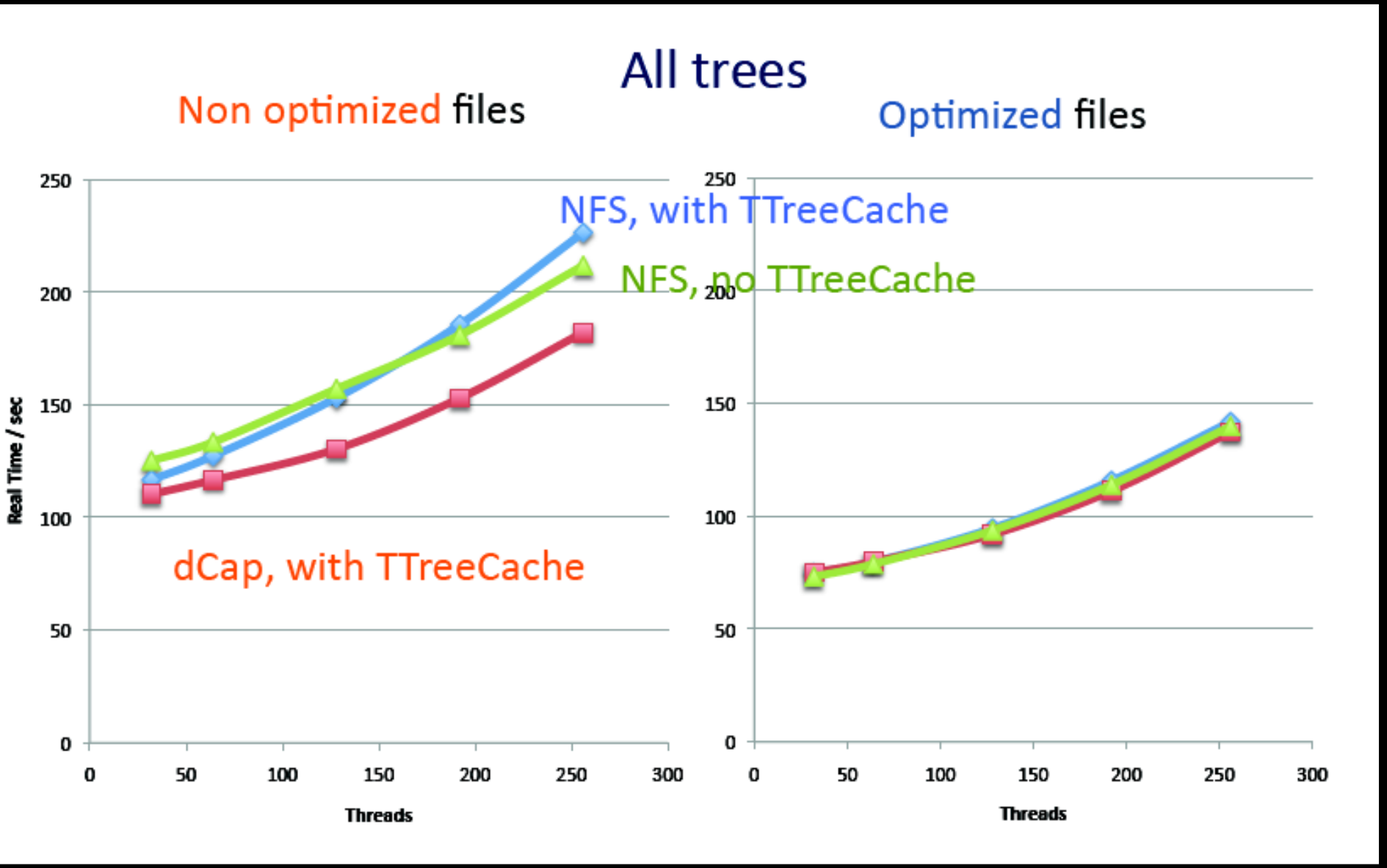


Initial load into pool file system cache.

NDGF
Nordic DataGrid Facility



NDGF
Nordic DataGrid Facility



NDGF
Nordic DataGrid Facility

	FTP	DCAP	xrootd	HTTP	WebDAV	NFS 4.1	SRM
Clear text password	x		o	1.9.11	1.9.11		
Digest password	o		o	1.9.12	1.9.12		
Kerberos	x	x	o	o	o	x	
TLS/SSL	o		o	x	x	?	1.9.11
GSI	x	x	1.9.11	o	o		x
Anonymous	x	x	x	x	x	x	
Custom			x	o			

x=supported by dCache
o=not implemented by dCache

	FTP	DCAP	xrootd	HTTP	WebDAV	NFS 4.1	SRM
File read	x	x	x	x	x	x	
File write	x	x	x	1.9.6	x	x	
Name space ops	x	x	1.9.9		x	x	x
Listing	x	x	1.9.9	?	x	x	x
Explicit staging		x	1.9.11				x
Pinning							x
ACL modification					o	x	o
Space management							x
Protocol negotiation	?						x
Async/out of order		x	x			x	x
Vector read		x	x	o	o	x	
Locking						NA	

x=supported by dCache
o=not implemented by dCache

	FTP	DCAP	xrootd	HTTP	WebDAV	NFS 4.1	SRM
Explicit data channel	x	x				x	
Mountable	Fuse		Fuse		Fuse	x	
Multiple streams	x		o			o	
Transport protocols	tcp udt	tcp	tcp	tcp	tcp	tcp udp rdma etc	tcp
Standards	rfc 959 rfc 2228 rfc 2389 rfc 3659 gfd.20 gfd.47			rfc 2616 rfc 2617 rfc 2818	rfc 3744 rfc 4918	rfc 3530 rfc 5661	gfd.129

x=supported by dCache
o=not implemented by dCache

ROOT	FTP	DCAP	xrootd	HTTP	WebDAV	NFS 4.1	SRM
… protocol	gfal:	dcap://	root://	http://	http://	file://	gfal:
… hidden cache	?		x				?
… async open			x				
… vector read		x	x	x	x	fadvise	

x=supported by dCache
o=not implemented by dCache

- Stability is much better than expected : Production ready.
- Kernel situation : short term solution for SL5 would be available, if we want.
- pNFS is partially already in 2.6.37
- Performance already comparable with existing solutions.
- Nevertheless : more evaluation on ROOT framework interaction needed. (vector read, fadvise)
- Efforts will continue within the EMI/dCache.org framework.

Thanks to

- Yves Kemp, Dmitri Ozerov (evaluation)
- Patrick Fuhrmann (I stole his slides)
- Tigran Mkrtchyan (he implemented it all!)
- Réne Brun (providing data files and script)

Slides

- Patrick: 2, 11, 12, 13, 14, 15 16, 17, 24
- www.pnfs.com: 4