REDDNET & OSG

Tier-3 Analysis with
Distributed Data

Daniel Engh

OSG Storage Forum – Nashville, TN
Sep 21, 2010

Logistical Networking

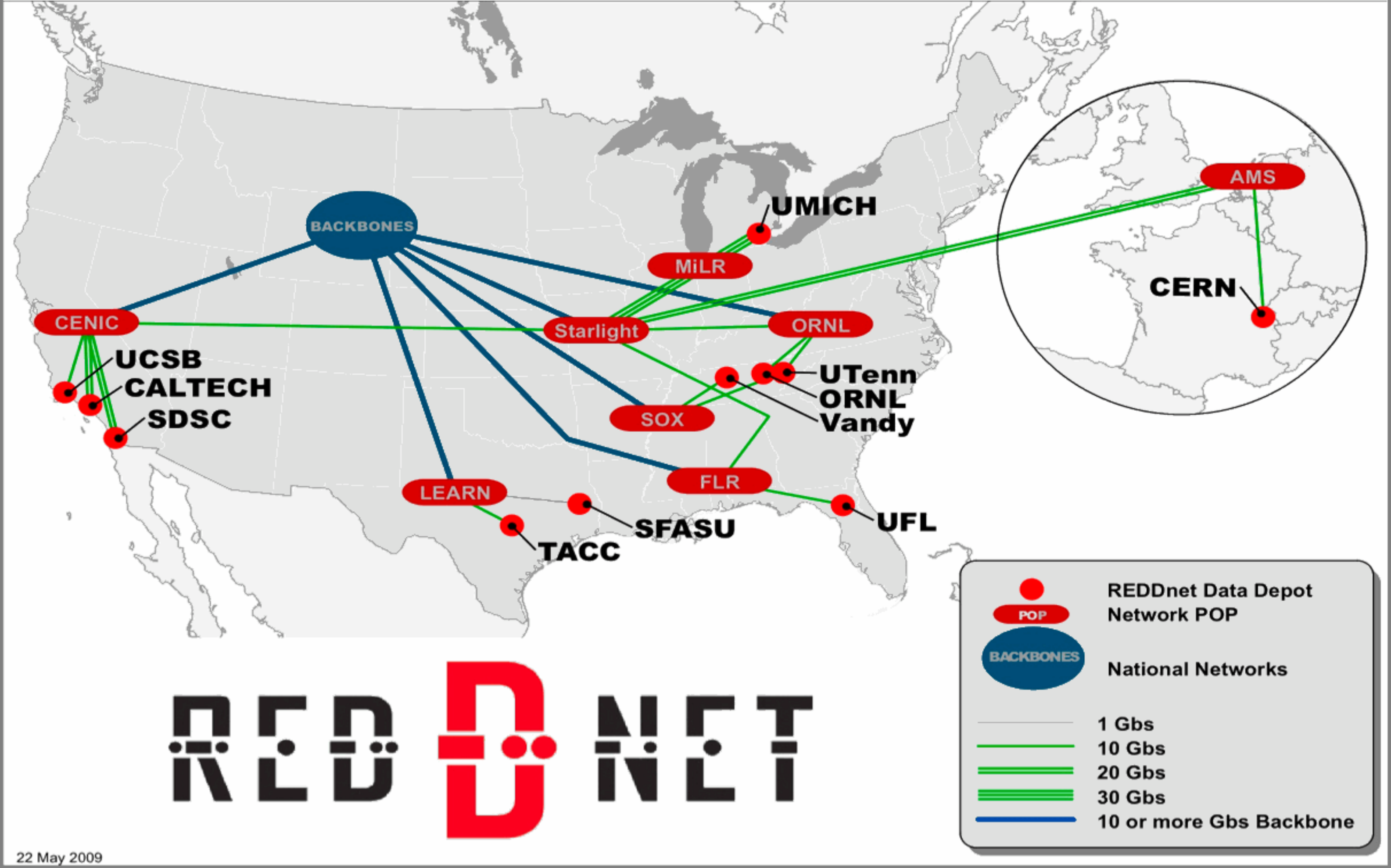
- Designed for Wide Area data access
- Confluence of Data and Networking
  - Think of data as communication not static storage
  - Use layered communications model (like OSI)
    - IBP protocol (like IP)
      - Simple, limited, **-- scalable --**
    - Higher Layers (like TCP, sessions, …)
      - LoDN, Phoebus, PerfSonar, Posix libs
- REDDnet
    - a deployment of LN tools
    - 700+ TB hardware, fast networking

      wins 2010 Internet 2 IDEA award

The Core: IBP Depots and Exnodes

- IBP Depots
  - Simple, basic, limited, distributed
  - Store data blocks (not files)
  - IBP keys – security for each block
  - Best effort (no advance reservation, etc)
  - No info on files, permissions, owners, etc.
- Exnode
  - Assemble your file (like UNIX inode)
  - For each data block:
    - URLs
    - IBP keys (read, write, manage)
    - length, offset

REDDnet Research & Education Data Depot Network



7/28/09

Mid Layer: File Services

- LoDN and L-Store
  - store exnodes
  - Add file system services
    - Directories, Owner, permissions, xttr
  - Add data placement policies
    - How many replicas, where?
  - Dispatch the data
    - Block-level replication
  - Maintain data integrity
    - Check/repair holes/re-dispatch
  - Maintain data placement policies

High Layer: User Interfaces

- GridFTP/L
  - Standard front-end, standard client tools, GUMS, etc
  - Backend talks to REDDnet service
    - Optimally access fastest (nearest?) data copy
  - Compatibility with:
    - SRM, Bestman, Phedex
- POSIX I/O
  - ROOT/L plugin developed
    - CMSSW compatible
  - Grid-secure, user certs GUMS, etc.
- Site setup: Use just like dCache, for example

File Services Grid Security

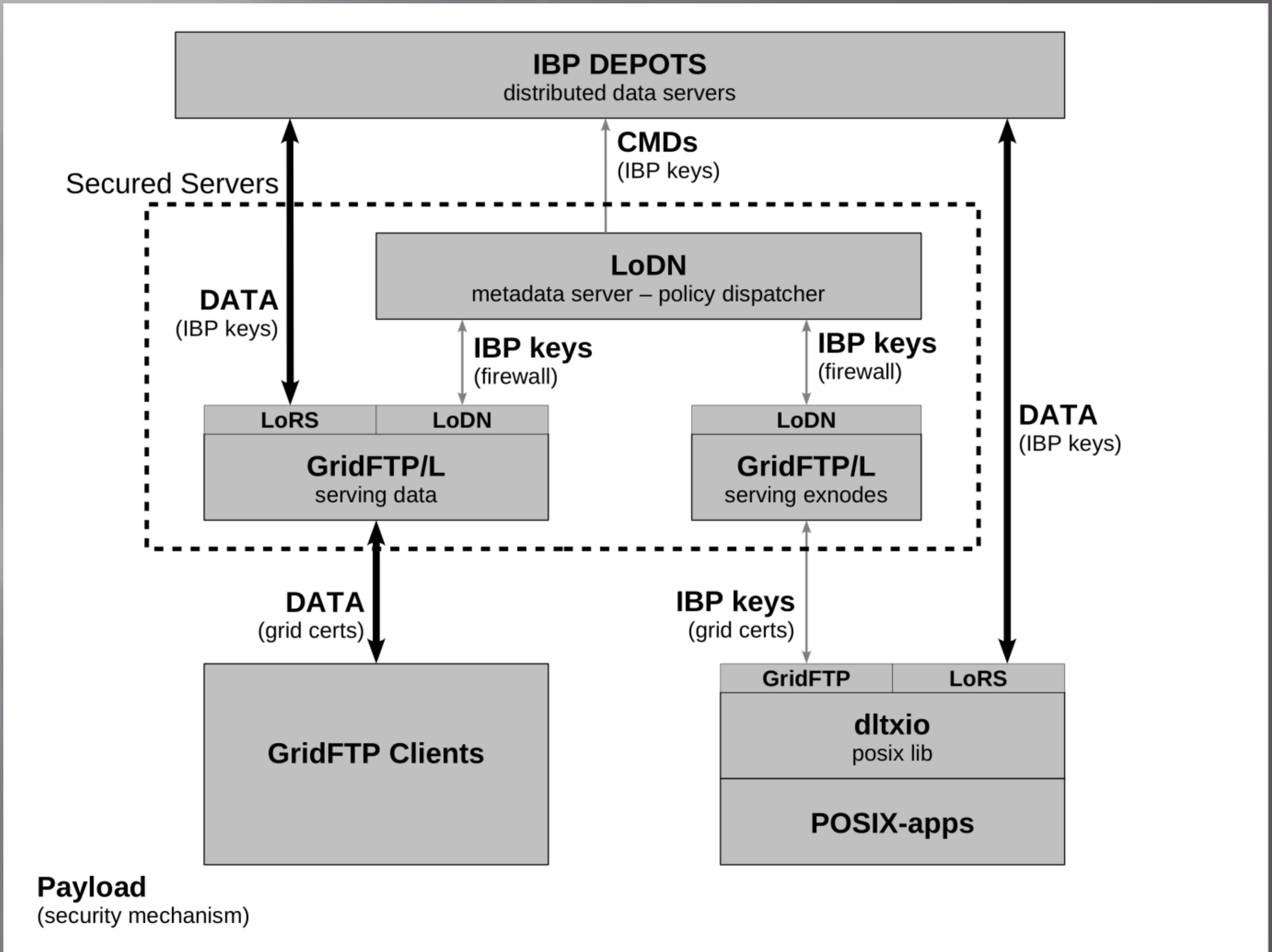
- Each data block secured with IBP keys
  - E.G. Need the read key to read the data
- Access to the exnode = access to data
  - GridFTP/L
    - Default Mode
      - Serves data as usual
      - IBP keys stay in GridFTP backend
    - Exnode mode
      - Serves IBP Keys
      - Small, fast transfer
      - Authenticated, encrypted transfer
      - Used by POSIX lib

local gftp view vs lodn gftp view

LoDN uses grid cert Distinguished Name for ID

```
vpac09:~> uberftp vampire.accre.vanderbilt.edu "ls /home/uscms01/gram*11.log"
220 vampire.accre.vanderbilt.edu GridFTP Server 2.8 (gcc64dbg, 1217607445-63) [VDT patched 4.0.8]
     ready.
230 User uscms01 logged in.
-rw-rw-r--  uscms01  osgusers        192958  Mar 30 13:51  /home/uscms01/gram_job_mgr_30511.log
-rw-rw-r--  uscms01  osgusers         32168  Sep 21 18:02  /home/uscms01/gram_job_mgr_14611.log
-rw-rw-r--  uscms01  osgusers        284858  Mar 29 05:25  /home/uscms01/gram_job_mgr_9311.log
-rw-rw-r--  uscms01  osgusers        128667  Apr  1 07:37  /home/uscms01/gram_job_mgr_8211.log
-rw-rw-r--  uscms01  osgusers        127344  Feb  7 23:09  /home/uscms01/gram_job_mgr_15211.log
vpac09:~> uberftp se2.accre.vanderbilt.edu "ls /home/uscms01/"
220 se2.accre.vanderbilt.edu GridFTP Server 3.19 (gcc64dbg, 1261034258-1) [Globus Toolkit 5.0.0] ready.
230 User uscms01 logged in.
drwxr-xr-x   2 George James 124822      cms         4096 Sep  3 12:34 log1
-rw-------   1 George James 124822      cms           35 Jun  8 12:34 .lesshst
drwxr-xr-x   2 George James 124822      cms         4096 Sep  3 12:35 bin
drwxr-xr-x  60 George James 124822      cms         4096 Nov 25 14:28 ..
drwxr-xr-x   6 George James 124822      cms         4096 Sep  3 12:40 .
drwx------   2 George James 124822      cms         4096 May 29 18:22 .ssh
-rw-------   1 George James 124822      cms         2025 Apr  1 13:10 .bash_history
drwxr-xr-x   3 George James 124822      cms         4096 Jun  8 12:23 .emacs.d
-rw-r--r--   1 George James 124822      cms        36770 Mar 26 15:30 gftp_kill.log
```

REDDnet Grid Security



IBP DEPOTS
distributed data servers

CMDs
(IBP keys)

Secured Servers

LoDN
metadata server – policy dispatcher

DATA
(IBP keys)

IBP keys
(firewall)

IBP keys
(firewall)

LoRS

LoDN

LoDN

DATA
(IBP keys)

GridFTP/L
serving data

GridFTP/L
serving exnodes

DATA
(grid certs)

IBP keys
(grid certs)

GridFTP

LoRS

GridFTP Clients

dltxio
posix lib

POSIX-apps

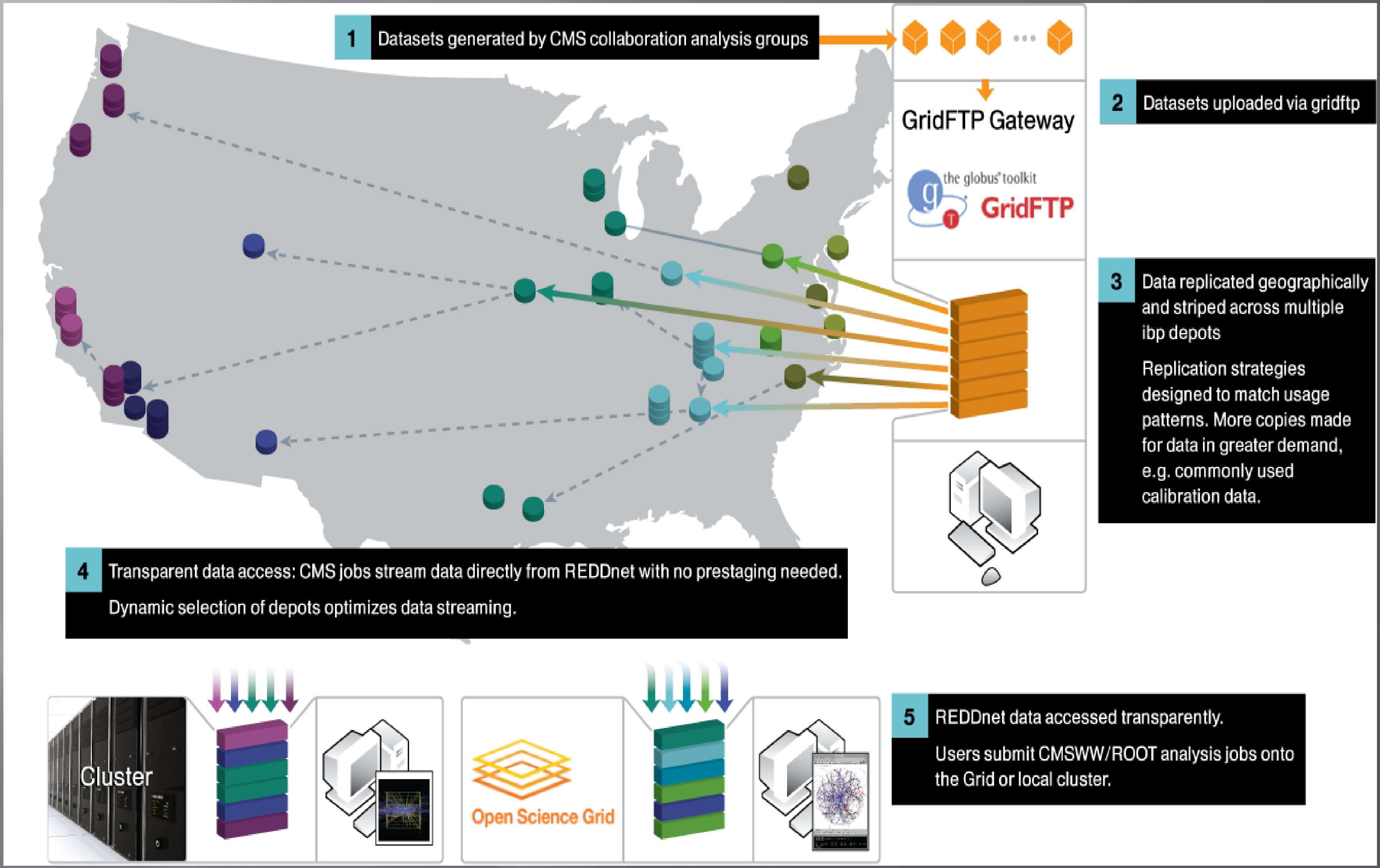
Payload
(security mechanism)

Block-level augmentation

- Big advantages over file replication
  - Smaller units
    - Security embedded within block-level layer (IBP keys)
    - Retries less costly
    - Placement flexibility (space left on device)
  - File semantics reside at higher layer
    - Start, end, size, name
  - File-System semantics & services avoided
    - Files remain in same directory, policy, etc.
    - Owner, group, same
    - The blocks are replicated, the file is “augmented”
    - F.S. security (eg) not used for augmentation

Read Readiness

- Reliable streaming of data
  - Transparent retry
    - Failover to nearest copy, then next nearest, …
  - Jobs won’t fail due to missing data
    - Only slow down
    - Acceptable performance hit for many scenarios
    - Not even noticeable for small enough holes



1 Datasets generated by CMS collaboration analysis groups

2 Datasets uploaded via gridftp

3 Data replicated geographically and striped across multiple ibp depots

Replication strategies designed to match usage patterns. More copies made for data in greater demand, e.g. commonly used calibration data.

4 Transparent data access: CMS jobs stream data directly from REDDnet with no prestaging needed.

Dynamic selection of depots optimizes data streaming.

5 REDDnet data accessed transparently.

Users submit CMSWW/ROOT analysis jobs onto the Grid or local cluster.

How to set up

Entering phase for T3 test community

- Join CMS REDDnet mailing list
- Vanderbilt initially deploys/maintains
  - IBP Depots
  - LoDN/Lstore File Services
  - distributed GridFTP&SRM servers

  Toolkit for easy installation coming…
- User sets up ROOT-based analysis
  - ROOT, Fwlite, CMSSW
  - Manual install 2 REDDnet libs
  - Manually Adapt procedures, scripts.
  - Library will added to CMSSW IO protocol suite

  POSIX lib available for recompiling apps

How to use it

- Request LoDN policy
  - Directory
  - Sites for replication
- Upload/Download data
  - globus-url-copy, uberftp, srm-copy
  - Your standard globus tools will work
  - OSG VO's already set up.
- Stream Data
  - ROOT plugin avail for download.
  - Run data off of local or nearby depots.

Usage examples

- globus-url-copy file:///tmp/myfile gsiftp://se3.accre.vanderbilt.edu/mydir/myfile
- uberftp se3.accre.vanderbilt.edu “ls /mydir”
- ROOT-based analysis
  - Specify physical file name
  - TFile::Open(“lors://se3.accre.vanderbilt.edu/mydir/myfile”);

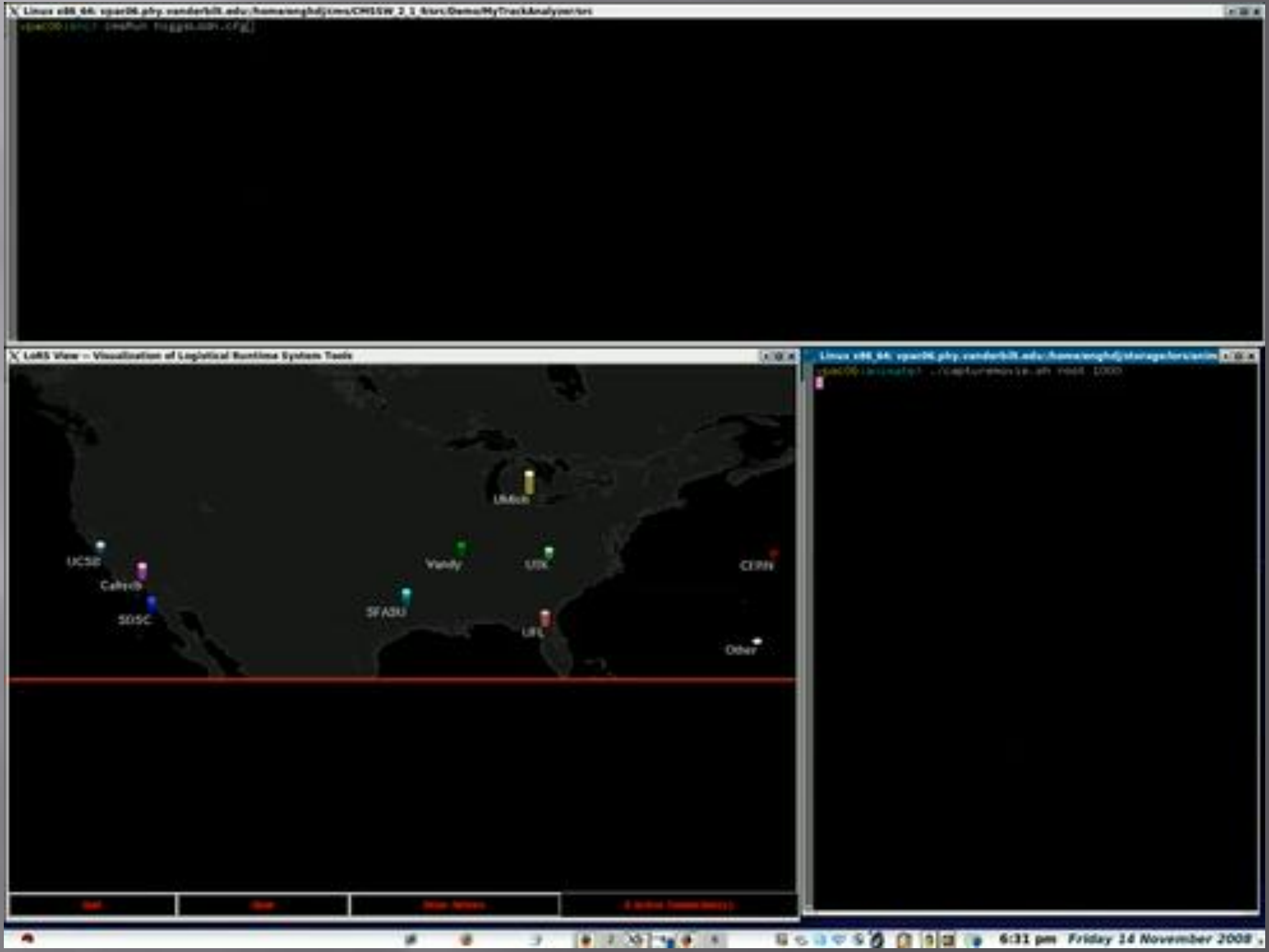
Where to use it

- Vanderbilt Maintains depots and Gateways
- IBP depots Currently at:
  - CERN, Vandy, UFL, Umich, Caltech, SDSC, UCSB, SFASU, TACC, ORNL, UTK
  - 10-15 more sites will be added
    - Who is interested?
    - Email me

Daniel.Engh@vanderbilt.edu

- We'll bring more info to CMS/OSG T3 regular mtgs

CMSSW data streaming



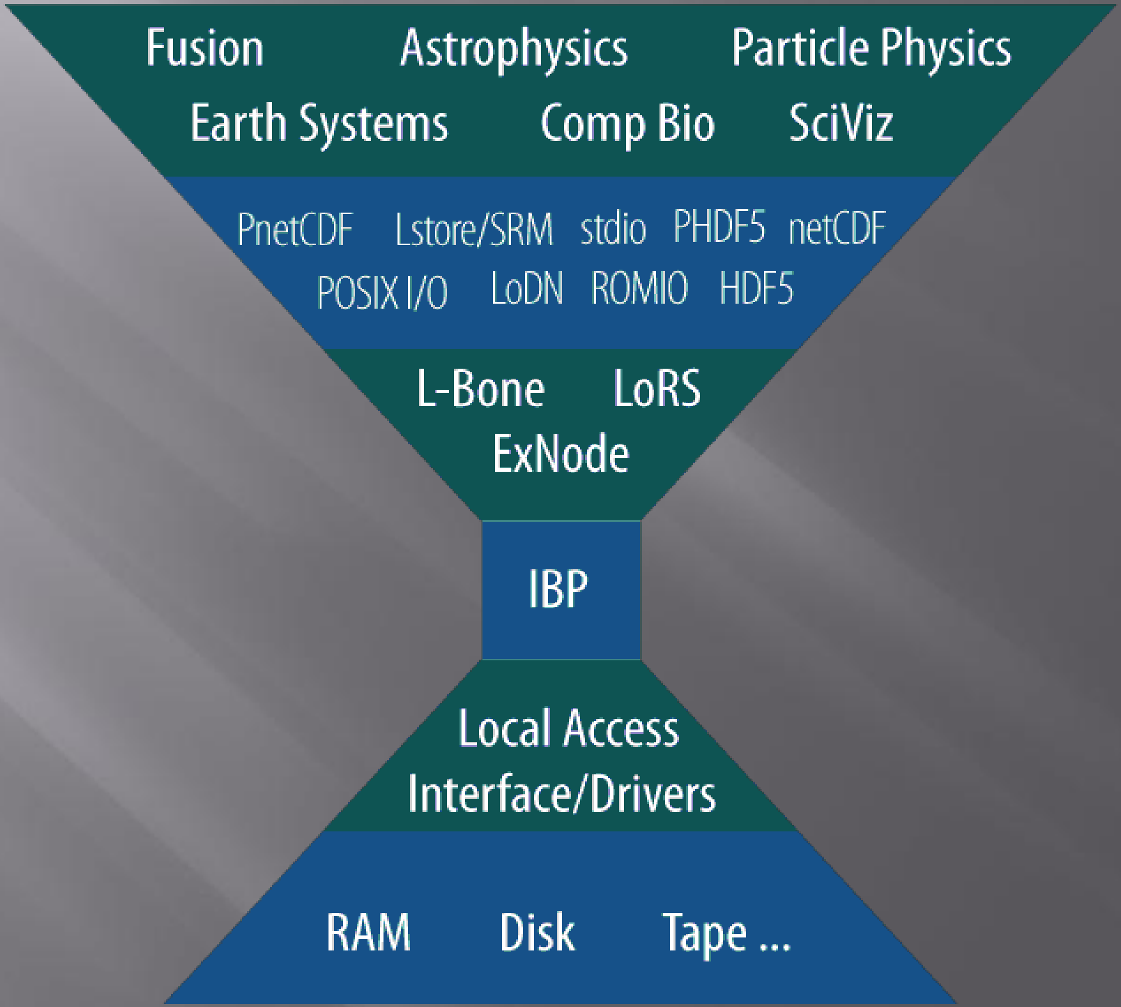
Linux x86_64: vpac06.phy.vanderbilt.edu:/home/enghdj/cms/CMSSW_2_1_9/src/Demo/MyTrackAnalyzer/src
LoRS View -- Visualization of Logistical Runtime System Tools
Linux x86_64: vpac06.phy.vanderbilt.edu:/home/enghdj/storagelors/anim
UMich
UCSD
Caltech
SDSC
Vandy
UTK
CERN
SFASU
UFL
Other
6:31 pm Friday 14 November 2008

```
vpac06(src)> cmsRun higgs.cdn.cfg
```

```
vpac06(animate)> ./capturemovie.sh root 1000
```

Extra Slides

LN Layers



Fusion Astrophysics Particle Physics
Earth Systems Comp Bio SciViz

PnetCDF Lstore/SRM stdio PHDF5 netCDF
POSIX I/O LoDN ROMIO HDF5

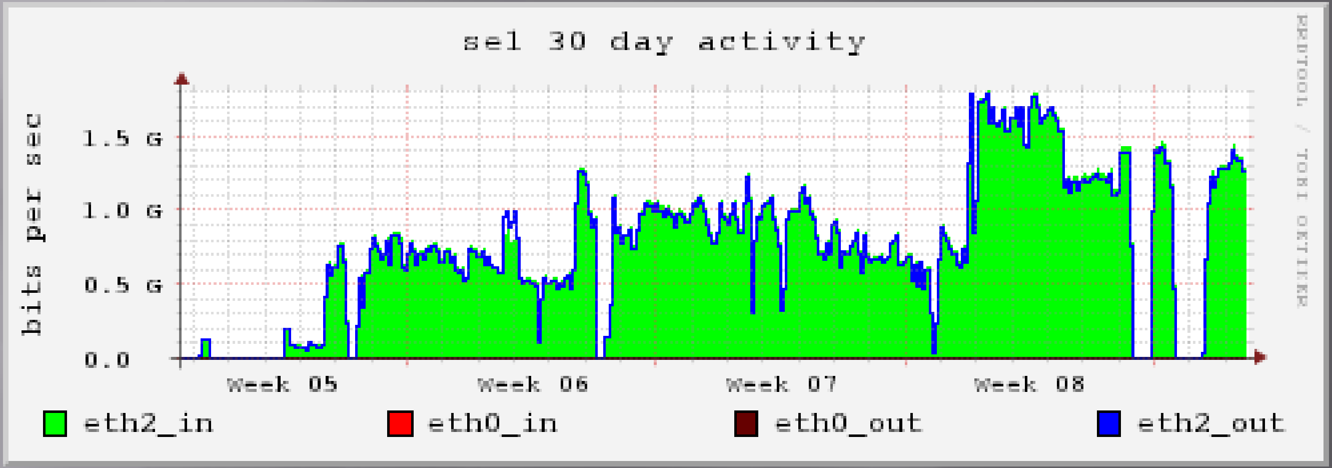
L-Bone LoRS
ExNode

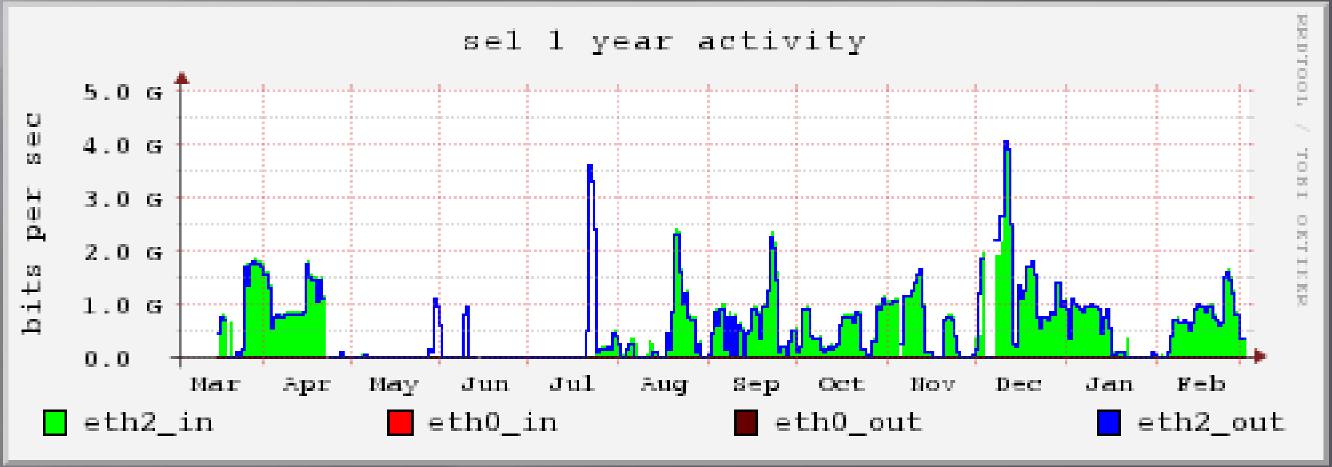
IBP

Local Access
Interface/Drivers

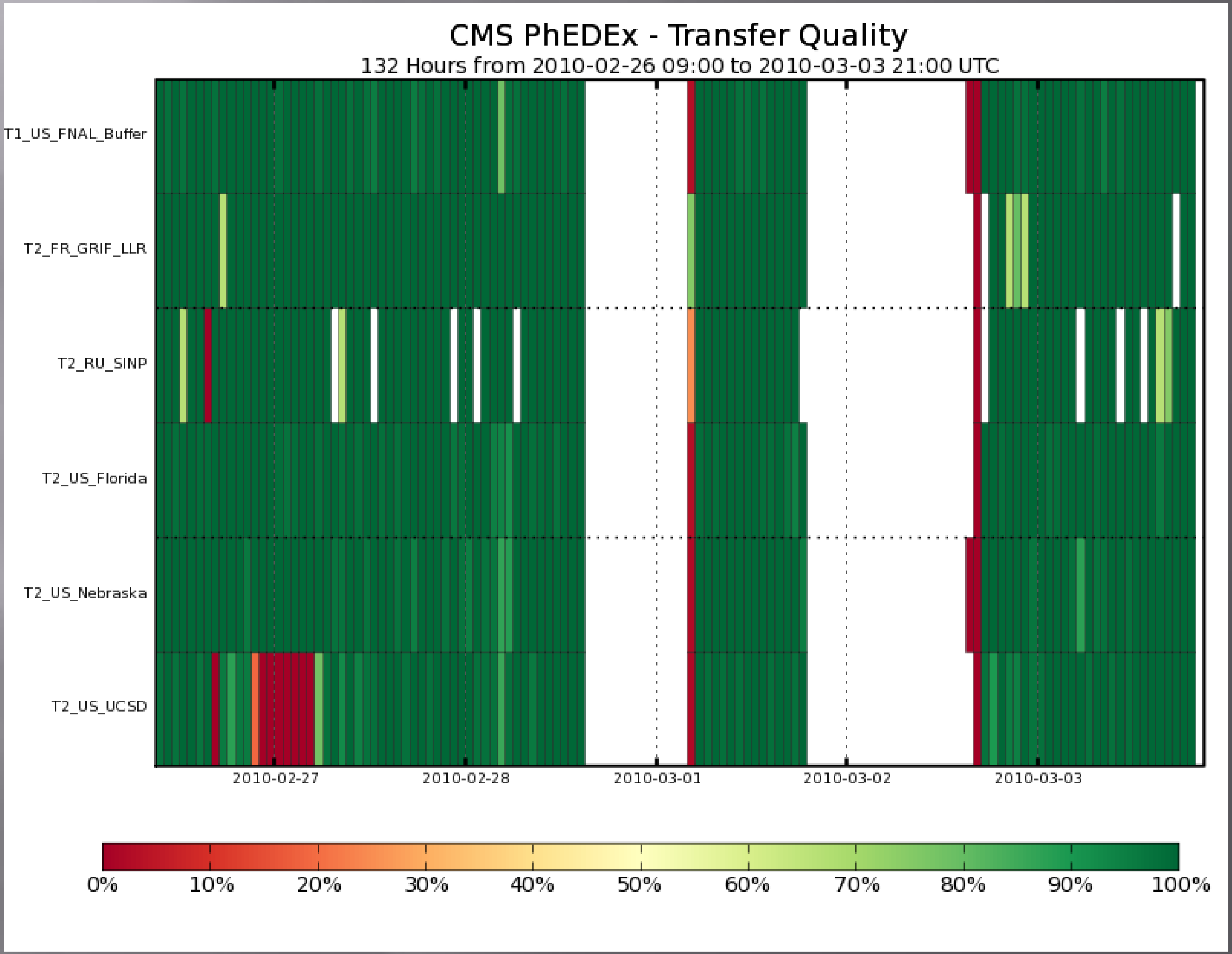
RAM Disk Tape ...

Vanderbilt GridFTP/L Gateway

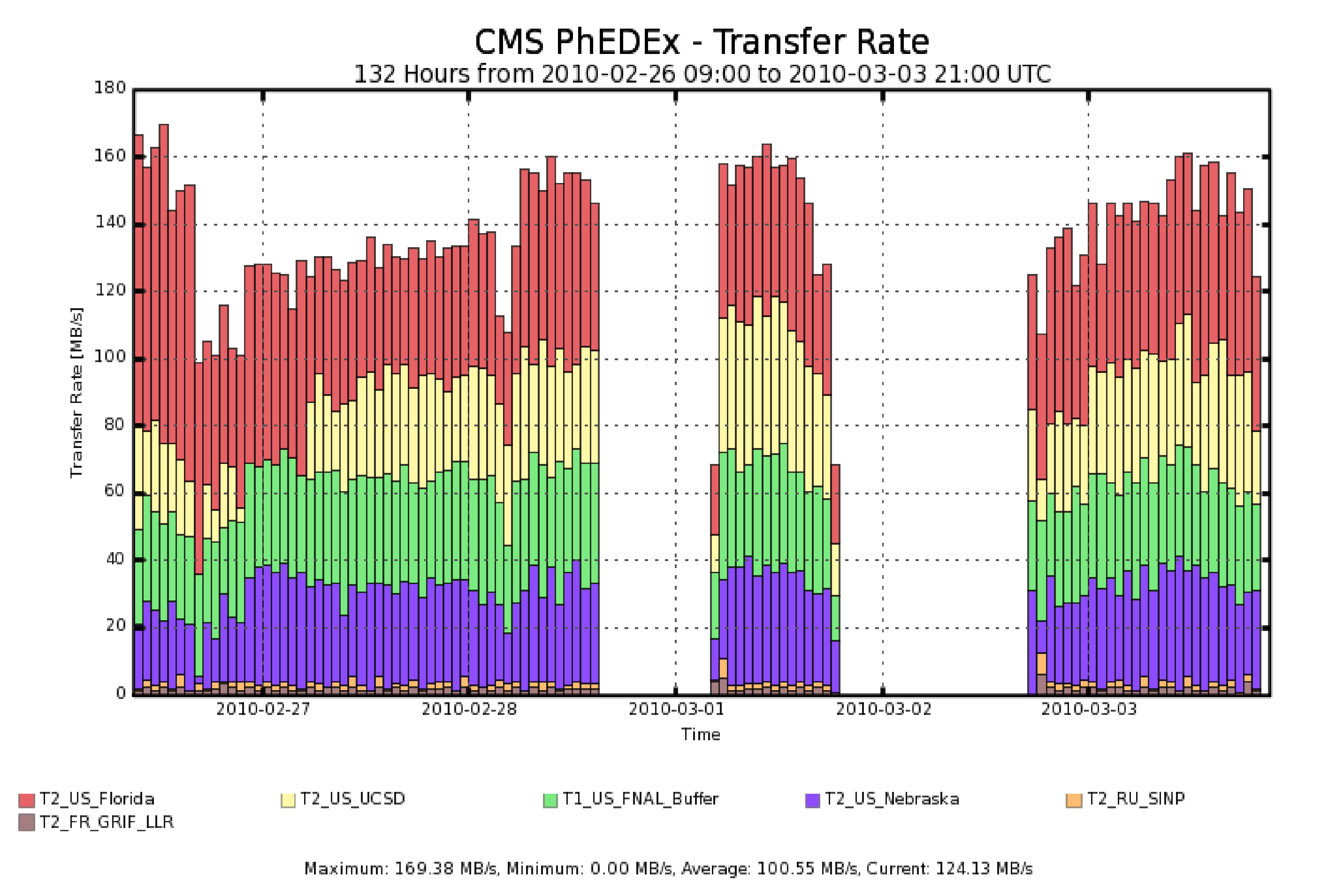




CMS PhEDEx monitors REDDnet



CMS PhEDEx monitors REDDnet


CMS PhEDEx - Transfer Rate
132 Hours from 2010-02-26 09:00 to 2010-03-03 21:00 UTC
Transfer Rate [MB/s]
Time
T2_US_Florida
T2_US_UCSD
T1_US_FNAL_Buffer
T2_US_Nebraska
T2_RU_SINP
T2_FR_GRIF_LLR
Maximum: 169.38 MB/s, Minimum: 0.00 MB/s, Average: 100.55 MB/s, Current: 124.13 MB/s