Monitoring of the US CMS Xrootd Federation

Matevž Tadel, UCSD

Overview

- What we did about XRD monitoring in the context of the AAA project
  - Things presented are mostly deployed to participating T2 sites, need to add FNAL
- Focus on motivations & technical stuff
- People involved:
  - CMS: Alja, Brian & Matevž
- With help from:
  - XRD: Andy & Gerri
  - ML: Costin & Ramiro

What we monitor & Why?

1. Service availability / basic access checks
   - Are sites / serves alive? Can we access data?
2. XRD summary monitoring stream
   - Extract operational statistics from servers
   - Eventually allow site admins to throttle the usage
3. XRD detailed monitoring stream
   - Follow all file access activity, including data-access patterns
   - Who ... from where ... how → detailed accounting & abuse detection
   - Site-site traffic (data placement, caching strategy) + data-set / file popularity
4. Server health & performance
   - Correlate with other monitoring info to understand problems

I. Service availability

- Probes that periodically connect to redirectors / servers
  - Is service available?
  - Can we authenticate? Try CERN & OSG certs
  - Can we read data? Just grab the first kilobyte
    - Test redirection from top-level meta manager to all sites
      - Use a name-space trick + standard release validation files known to be on all sites for CMSSW checks.
- Currently we use an Nagios instance at UNL - that's also where the top-level redirector is
  - Sites will take responsibility for server monitoring (we plan to use RSV probes at UCSD)
  - Central Nagios will only test redirectors / sites

red workers x4275 (red-workers-x4275)

Host	Status	Services	Actions
node251	UP	4 OK	
node252	UP	4 OK	
node253	UP	4 OK	
node254	UP	4 OK	

Xrootd Redirectors (xrootd-redirectors)

Host	Status	Services	Actions
xrootd-itb.unl.edu	UP	8 OK 2 CRITICAL	
xrootd.unl.edu	UP	13 OK 1 CRITICAL	

Xrootd Servers Caltech (xrootd-servers-caltech)

Host	Status	Services	Actions
cithep160.ultralight.org	UP	2 OK	
cithep172.ultralight.org	UP	2 OK	
cithep230.ultralight.org	UP	2 OK	
cithep251.ultralight.org	UP	2 OK	
gridftp-16-23.ultralight.org	UP	2 OK	

Xrootd Servers Florida (xrootd-servers-florida)

Host	Status	Services	Actions
xrootd1.ihepa.ufl.edu	UP	2 OK	
xrootd2.ihepa.ufl.edu	UP	2 OK	
xrootd3.ihepa.ufl.edu	UP	2 OK	

Xrootd Servers FNAL (xrootd-servers-fnal)

Host	Status	Services	Actions
cmssrv32.fnal.gov	UP	2 OK	

Xrootd Servers MIT (xrootd-servers-mit)

Host	Status	Services	Actions
xrootd1.cmsaf.mit.edu	UP	2 OK	
xrootd2.cmsaf.mit.edu	DOWN	2 CRITICAL	
xrootd3.cmsaf.mit.edu	DOWN	2 CRITICAL	

Xrootd Servers Purdue (xrootd-servers-purdue)

Host	Status	Services	Actions
cmsdbs.rcac.purdue.edu	UP	2 OK	
crabserver.rcac.purdue.edu	UP	2 OK	
xrootd.rcac.purdue.edu	UP	2 OK	

Xrootd Servers UCSD (xrootd-servers-ucsd)

Host	Status	Services	Actions
uaf-3.t2.ucsd.edu	UP	2 OK	
uaf-4.t2.ucsd.edu	UP	2 OK	
uaf-5.t2.ucsd.edu	UP	2 OK	
uaf-6.t2.ucsd.edu	UP	2 CRITICAL	
uaf-7.t2.ucsd.edu	UP	2 OK	
uaf-8.t2.ucsd.edu	UP	2 OK	
uaf-9.t2.ucsd.edu	UP	2 OK	
xrootd.t2.ucsd.edu	UP	2 OK	

Xrootd Servers UNL (xrootd-servers-unl)

Host	Status	Services	Actions
red-gridftp1	UP	8 OK	
red-gridftp10	UP	8 OK	
red-gridftp11	UP	7 OK 1 CRITICAL	
red-gridftp12	UP	8 OK	
red-gridftp2	UP	8 OK	
red-gridftp3	UP	8 OK	
red-gridftp4	UP	8 OK	
red-gridftp5	UP	8 OK	
red-gridftp6	UP	8 OK	
red-gridftp7	UP	8 OK	
red-gridftp8	UP	8 OK	
red-gridftp9	UP	8 OK	
srm	UP	5 OK 1 WARNING	

Nagios@UNL

Xrootd Servers Vanderbilt (xrootd-servers-vanderbilt)

Host	Status	Services	Actions
se2.accre.vanderbilt.edu	UP	2 OK	

Xrootd Servers Wisconsin (xrootd-servers-wisconsin)

Host	Status	Services	Actions
cmsxrootd.hep.wisc.edu	UP	2 OK	

II. Xrd summary monitoring

- Most Xrd instances → pre-processor (perl) → MonALISA

  xrd.report xrootd.t2.ucsd.edu:9931, desire.physics.ucsd.edu:9931 every 30s all sync

  - The pre-processor and ML service both run at UCSD
    - In principle no problem to run both at more sites.
      - We would even prefer this, in fact.
  - Pre-processor:
    - Classification by site (Using ML Cluster namespace)
    - Calculates rates
    - Sends on the desired parameters & rates
  - ML repository & web-interface also at UCSD

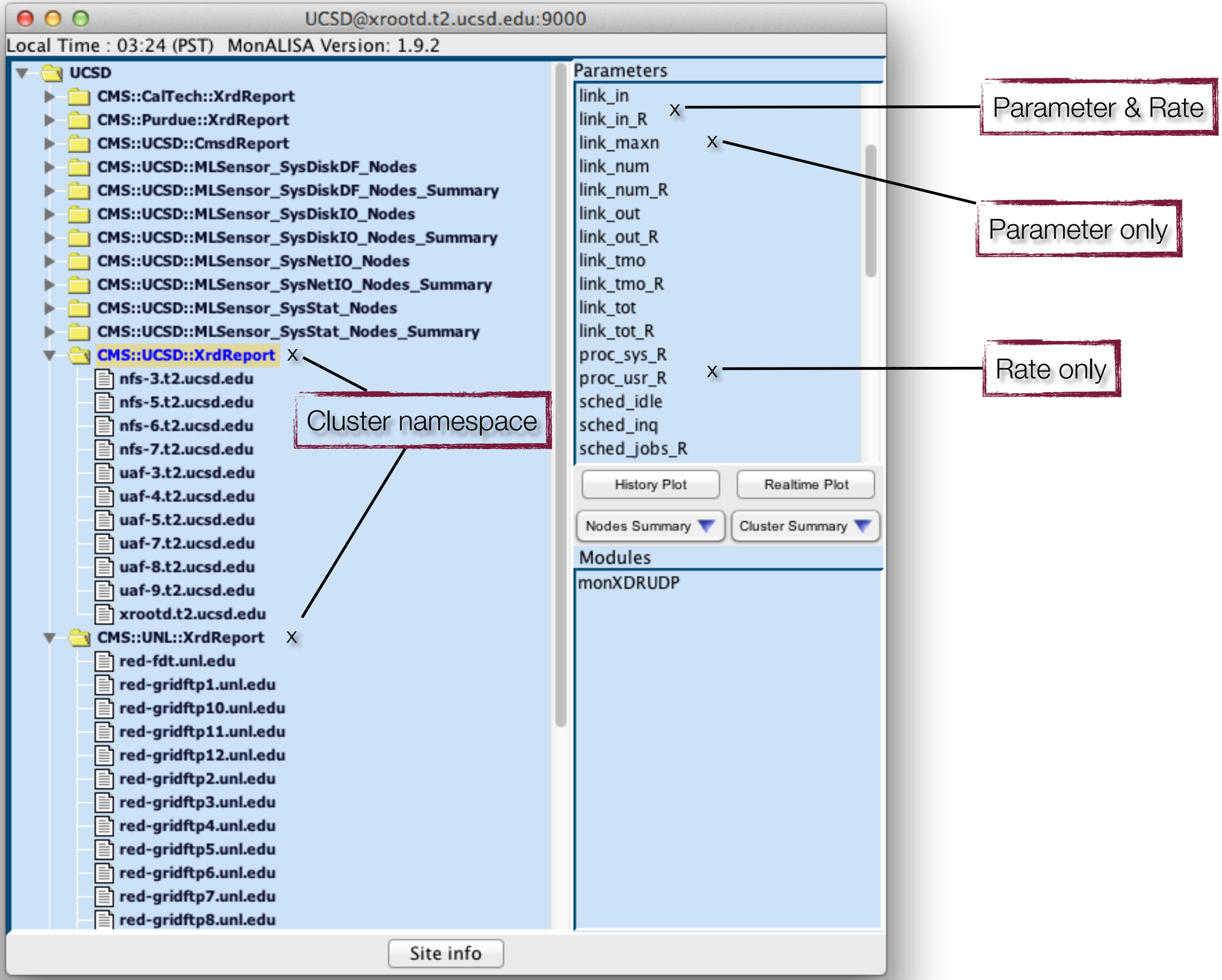
```
$Pgm2Values =
{
  'xrootd' =>
   [
    [ ['buff'],  ['reqs', 'buffs', 'mem'] ],
    [ ['link'],  ['ctime', 'maxn', 'in', 'num', "out", "tmo", "tot"] ],
    # ofs
    # oss
    # ['poll'], ['att', 'en', 'ev', 'int']
    # proc - only as rates
    [ ['sched'], ['idle', 'inq', 'maxinq', 'tcr', 'tde', 'threads', 'tlimr'] ],
   ],
  'cmsd' =>
   [
    # [  ],
   ],
};

$Pgm2Rates =
{
  'xrootd' => [
    [ ['buff'], ['reqs', 'buffs', 'mem'] ],
    [ ['link'], ['in', 'num', "out", "tmo", "tot"] ],
    # ['poll'], ['att', 'en', 'ev', 'int']
    [ ['proc'], ['sys', 'usr'] ],
    [ ['sched'], ['jobs'] ],
    [ ['xrootd'],          ['num', 'dly', 'err', 'rdr'] ],
    [ ['xrootd', 'ops'], ['getf', 'misc', 'open', 'pr', 'putf', 'rd', 'rf', 'sync', 'wr'] ],
    [ ['xrootd', 'lgn'], ['num', 'af', 'au', 'ua'] ],
  ],
  'cmsd'   => [
    [ ['proc'], ['sys', 'usr']  ],
    # [  ],
  ],
};
```

Parameters that go into ML

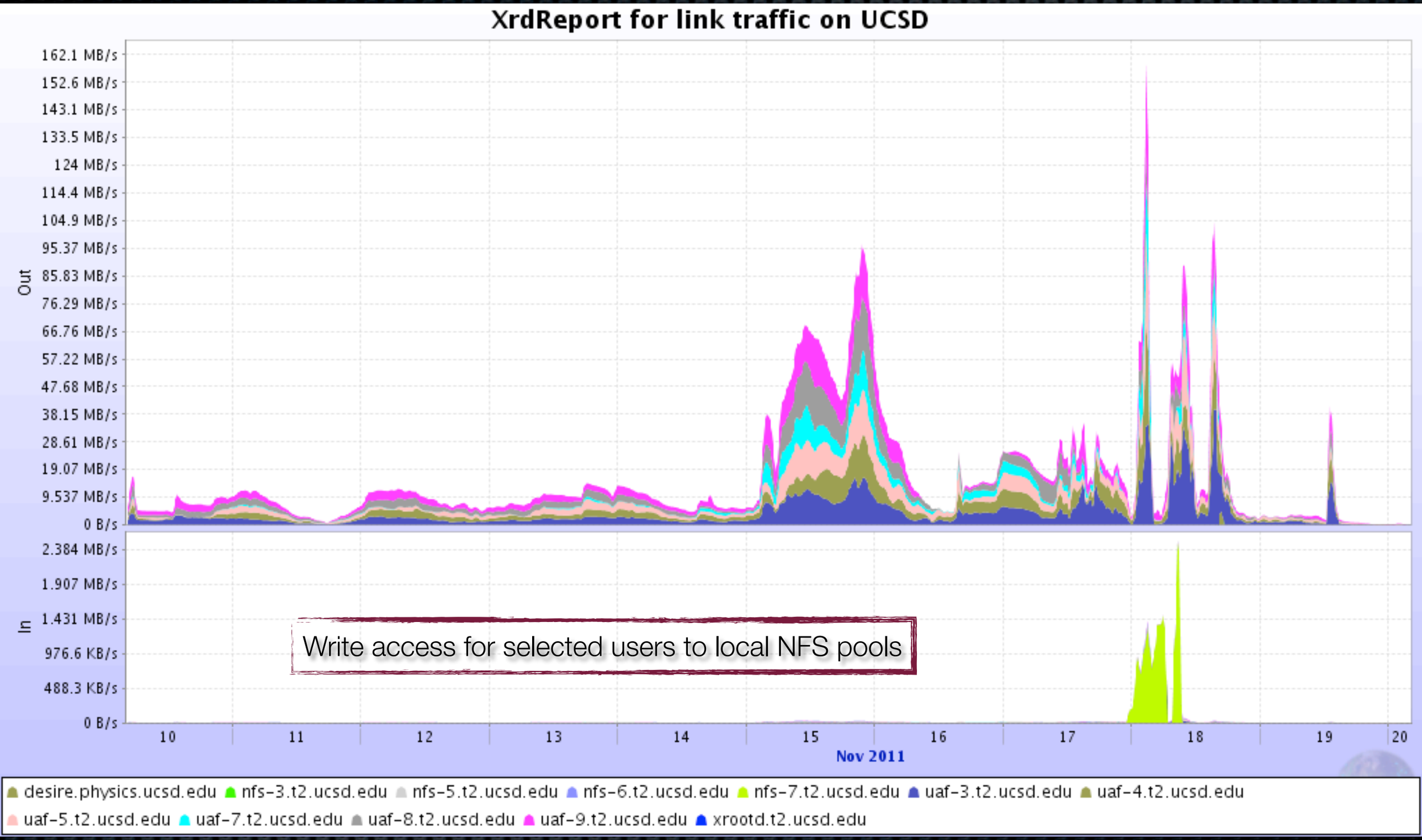
- Original names are kept, joined with '_'
- xrootd / cmsd supported
- Values / rates configured separately
- Trivial to extend

Authentication failure counts

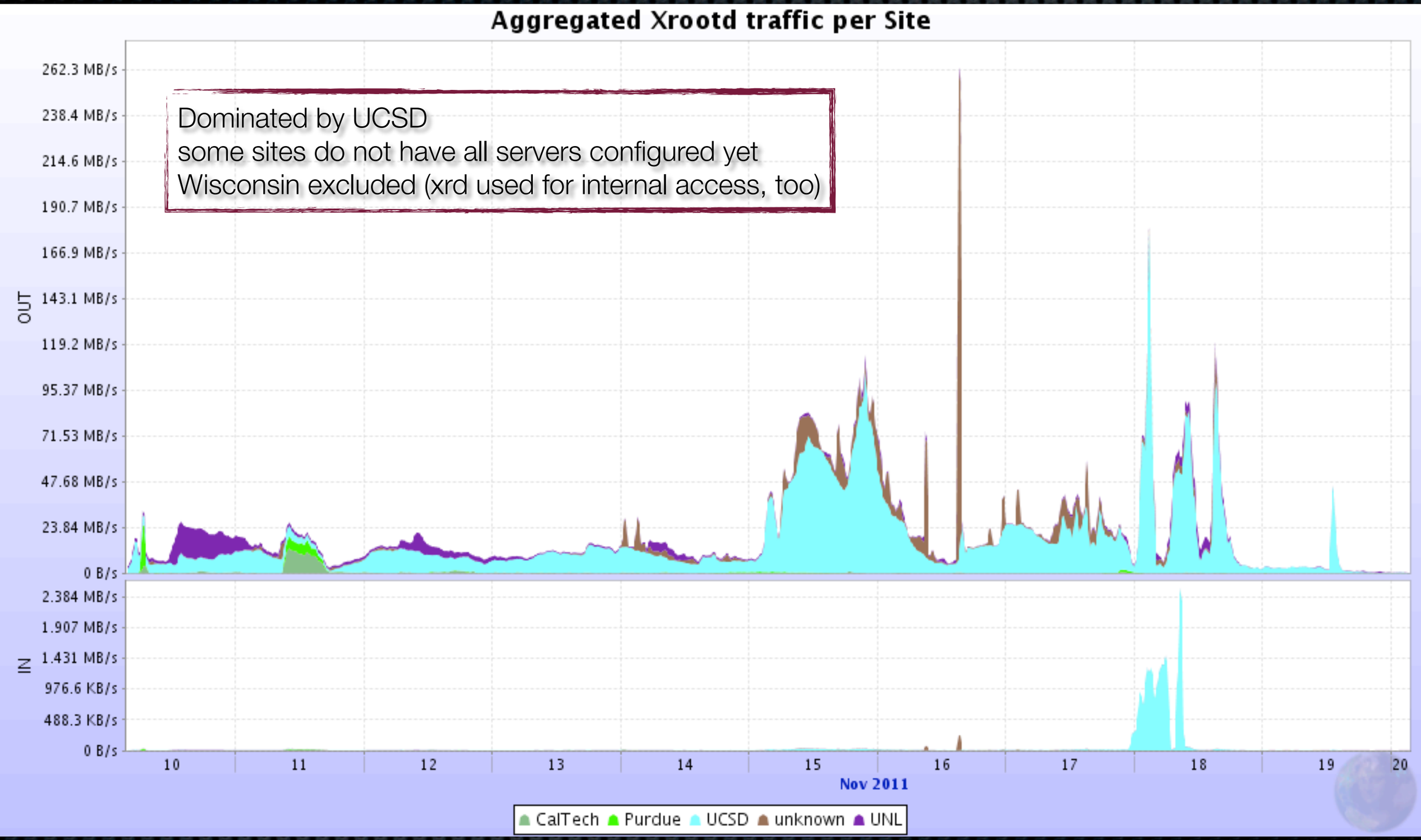
http://www.gled.org/viewvc/var/trunk/xrd-rep-snatcher/
https://svn.gled.org/var/trunk/xrd-rep-snatcher


UCSD@xrootd.t2.ucsd.edu:9000
Local Time : 03:24 (PST) MonALISA Version: 1.9.2
UCSD
CMS::CalTech::XrdReport
CMS::Purdue::XrdReport
CMS::UCSD::CmsdReport
CMS::UCSD::MLSensor_SysDiskDF_Nodes
CMS::UCSD::MLSensor_SysDiskDF_Nodes_Summary
CMS::UCSD::MLSensor_SysDiskIO_Nodes
CMS::UCSD::MLSensor_SysDiskIO_Nodes_Summary
CMS::UCSD::MLSensor_SysNetIO_Nodes
CMS::UCSD::MLSensor_SysNetIO_Nodes_Summary
CMS::UCSD::MLSensor_SysStat_Nodes
CMS::UCSD::MLSensor_SysStat_Nodes_Summary
CMS::UCSD::XrdReport
x
nfs-3.t2.ucsd.edu
nfs-5.t2.ucsd.edu
nfs-6.t2.ucsd.edu
nfs-7.t2.ucsd.edu
uaf-3.t2.ucsd.edu
uaf-4.t2.ucsd.edu
uaf-5.t2.ucsd.edu
uaf-7.t2.ucsd.edu
uaf-8.t2.ucsd.edu
uaf-9.t2.ucsd.edu
xrootd.t2.ucsd.edu
CMS::UNL::XrdReport
X
red-fdt.unl.edu
red-gridftp1.unl.edu
red-gridftp10.unl.edu
red-gridftp11.unl.edu
red-gridftp12.unl.edu
red-gridftp2.unl.edu
red-gridftp3.unl.edu
red-gridftp4.unl.edu
red-gridftp5.unl.edu
red-gridftp6.unl.edu
red-gridftp7.unl.edu
red-gridftp8.unl.edu
Cluster namespace
Parameters
link_in
link_in_R
x
link_maxn
x
link_num
link_num_R
link_out
link_out_R
link_tmo
link_tmo_R
link_tot
link_tot_R
proc_sys_R
proc_usr_R
x
sched_idle
sched_inq
sched_jobs_R
History Plot
Realtime Plot
Nodes Summary
Cluster Summary
Modules
monXDRUDP
Site info
Parameter & Rate
Parameter only
Rate only

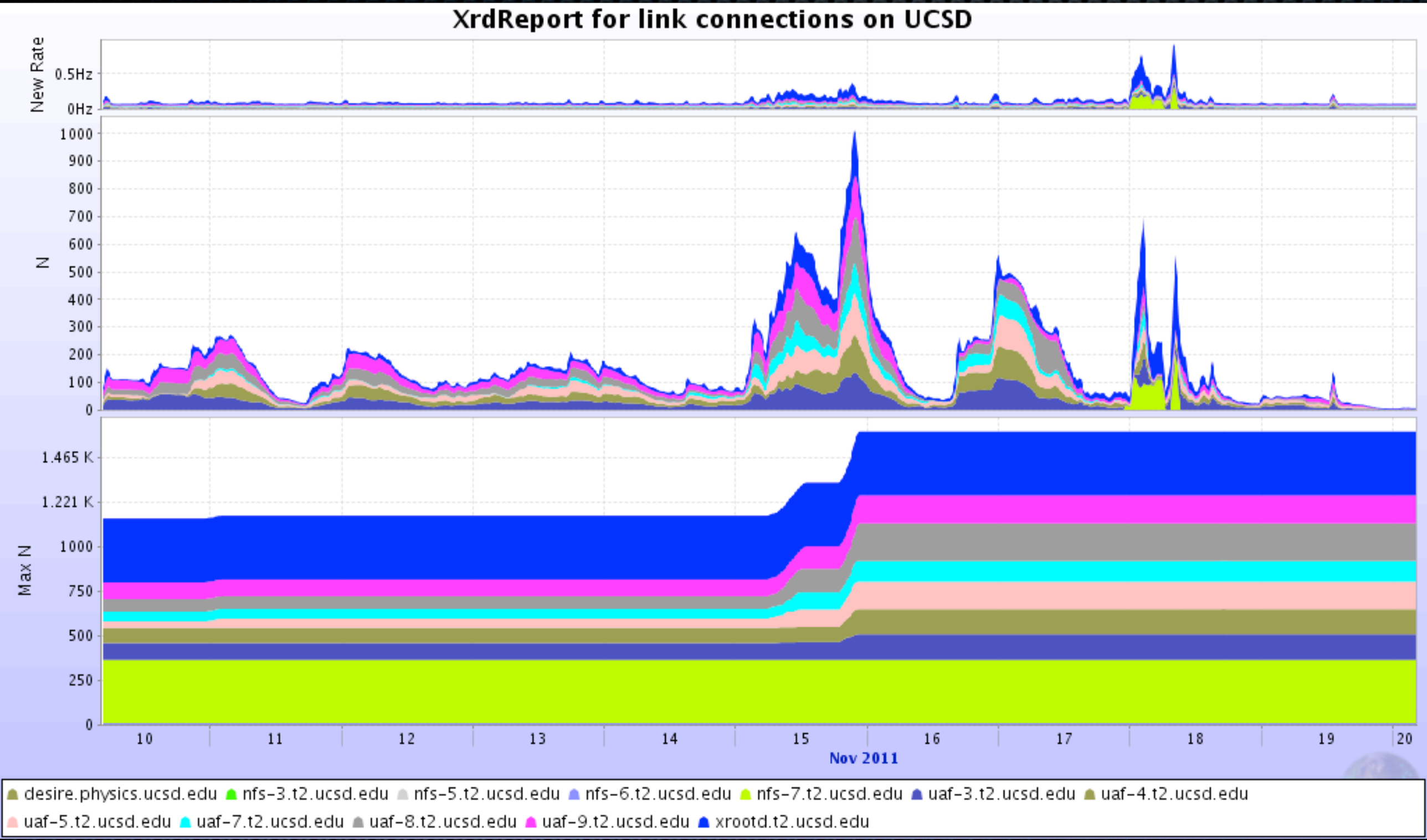
http://xrootd.t2.ucsd.edu/



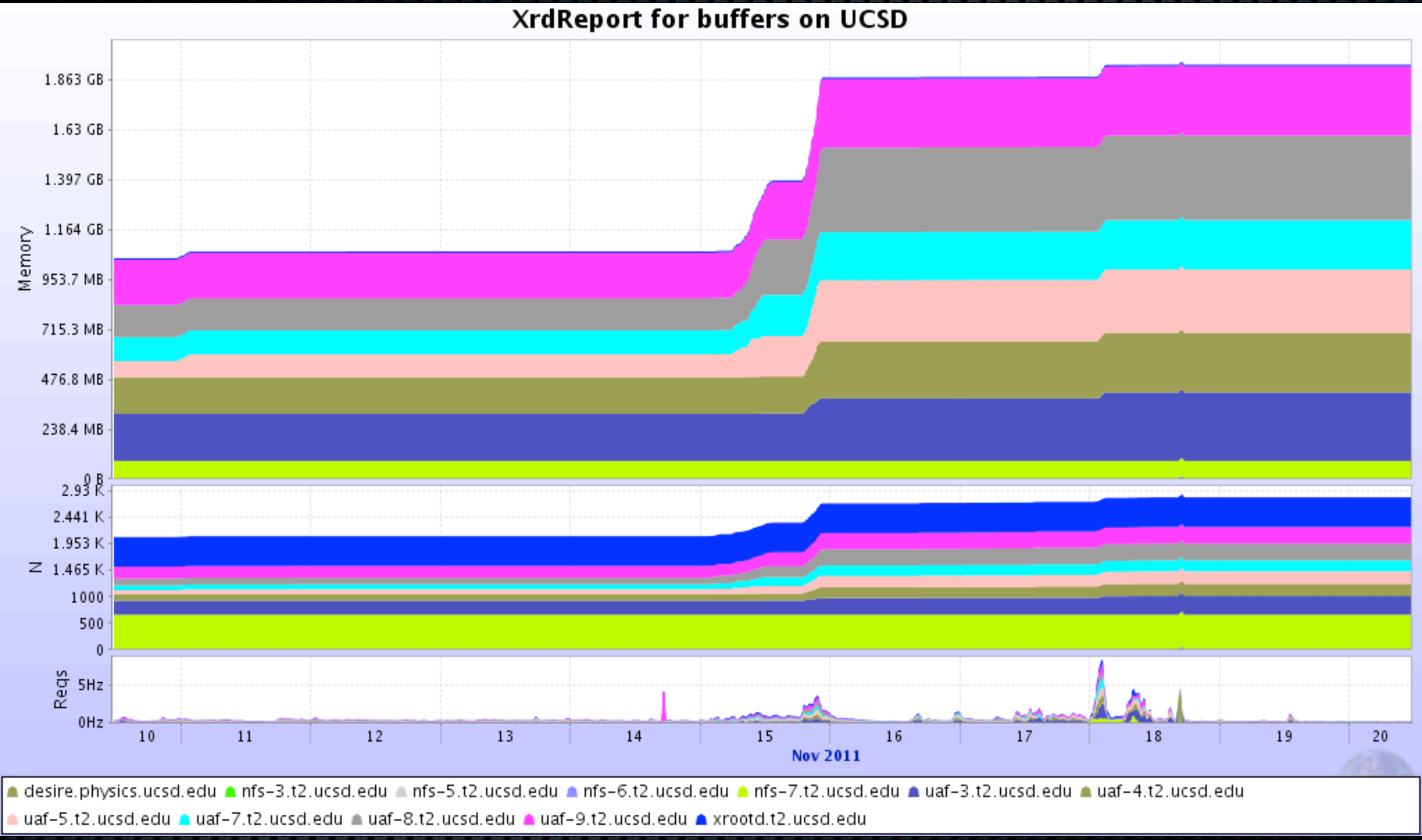
http://xrootd.t2.ucsd.edu/



http://xrootd.t2.ucsd.edu/



http://xrootd.t2.ucsd.edu/



III. Xrd detailed monitoring

- As mentioned, we monitor everything :)
  xrootd.monitor all auth flush io 30s mbuff 1472 window 5s dest files io info user xrootd.t2.ucsd.edu:9930
  - Sessions, file open/close, read/writes
    - Went through quite some trouble to get user DN into monitoring
      - Mostly we use GUMS, but now also works with grid-mapfiles
    - Improved IO trace: flushing, time-stamping, vector reads
- Again, all data collected at UCSD ... so far
- An aside -- multiplexing UDP packets is a pain!
  - We have a proper UDP forwarder almost ready.
  - Increase monitoring buffer size (also needed for redirections)

III. Xrd detailed monitoring

- The plan is to:
  - Store all details into root trees for further analysis
  - What files, data-sets are used (site-to-site matrix) → data placement
  - What fraction of files is actually read, how often → caching!
    - Make real-time 3D animation of data transfers for outreach (X-mass)
- Implementation of monitoring collector:
  - Implemented in Gled framework (ROOT-based):

    http://www.gled.org/
    - Support for multi-threaded execution & object locking
    - Automatic GUI generation based on class definition

CPU usage on CMS::UCSD::MLSe... | Creating Federated Data Stores ... | Scalla Extensions: cms | Xrd open files [258]

File	OpenAgo	ServerDomain	ClientDomain	User	Read [MB]	UpdateAgo
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0003/845B6D13-1007-E111-90F4-003048CFB40C.root	00:00:48	hep.wisc.edu	hep.wisc.edu		0.000	00:00:48
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0003/F6D35AB3-1007-E111-94C4-003048CF94A8.root	00:03:25	hep.wisc.edu	hep.wisc.edu		0.000	00:03:25
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0002/8A37B03A-0A07-E111-BEBB-003048F1C832.root	00:04:48	hep.wisc.edu	hep.wisc.edu		0.000	00:04:48
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0002/BE649E72-0507-E111-BE1F-003048F1110E.root	00:05:28	hep.wisc.edu	hep.wisc.edu		0.000	00:05:28
/store/mc/Summer11/SMS-T1tttt_Mgluino-450to1200_mLSP-50to800_7TeV-Pythia6Z/AODSIM/PU_START42_V11_FSIM-v2/0001/845FD4AA-66F9-E011-9B32-002481E14FFC.root	00:06:44	t2.ucsd.edu	unl.edu	Robert Schoefbeck	69.956	00:00:52
/store/data/Run2011B/L1EGHPF/AOD/PromptReco-v1/000/179/828/7E25F09A-1501-E111-99E8-002481E0D646.root	00:07:16	hep.wisc.edu	hep.wisc.edu		0.000	00:07:16
/store/data/Run2011B/SingleMu/AOD/PromptReco-v1/000/177/878 /D21B34A7-21F1-E011-A627-BCAEC518FF8E.root	00:10:26	unl.edu	xlate.ufl.edu	Gian Piero Di Giovanni	181.147	00:01:29
/store/data/Run2011B/SingleMu/AOD/PromptReco-v1/000/176/309/86448C5C-41E2-E011-B6BC-0030487CD7EE.root	00:10:28	unl.edu	xlate.ufl.edu	Gian Piero Di Giovanni	262.391	00:00:13
/store/mc/Fall11/ZZ_TuneZ2_7TeV_pythia6_tauola/AODSIM /PU_S6_START42_V14B-v1/0000/7CE2911A-92F3-E011-B25F-001A92810AB2.root	00:11:30	hep.wisc.edu	hep.wisc.edu		107.542	00:04:36
/store/data/Run2011B/L1EGHPF/AOD/PromptReco-v1/000/179/828/8690BA5E-8701-E111-95D7-001D09F24EE3.root	00:14:25	hep.wisc.edu	hep.wisc.edu		0.000	00:14:25
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0002/001CD847-0C07-E111-8D69-003048F1BF68.root	00:20:04	hep.wisc.edu	hep.wisc.edu		0.000	00:20:04
/store/mc/Summer11/SMS-T1tttt_Mgluino-450to1200_mLSP-50to800_7TeV-Pythia6Z/AODSIM/PU_START42_V11_FSIM-v2/0001/1C8BEB5B-ADF9-E011-8EA9-0025B3E05C9E.root	00:28:18	t2.ucsd.edu	unl.edu	darren burton	84.113	00:07:12
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0002/B442BA47-0E07-E111-9AE0-003048F1C58C.root	00:32:22	hep.wisc.edu	hep.wisc.edu		0.000	00:32:2
/store/mc/Summer11/DYToEE_M-20_TuneZ2_7TeV-pythia6/AODSIM /PU_S3_START42_V11-v1/0000/3C6B7716-2677-E011-AB06-003048D4DEAC.root	00:35:58	hep.wisc.edu	hep.wisc.edu		312.647	00:04:47
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0002/5E9F8461-0F07-E111-8AF9-003048F11942.root	00:37:18	hep.wisc.edu	hep.wisc.edu		0.000	00:37:1
/store/mc/Summer11/SMS-T1tttt_Mgluino-450to1200_mLSP-50to800_7TeV-Pythia6Z/AODSIM/PU_START42_V11_FSIM-v2/0001/2E8C3310-EAF8-E011-AA68-001A647894A4.root	00:38:01	t2.ucsd.edu	unl.edu	Robert Schoefbeck	313.827	00:01:14
/store/generator/Summer11/DYJetsToLL_M-10To50_TuneZ2_7TeV-madgraph /GEN/START311_V2-v1/0003/E4440E53-1307-E111-A123-003048F11CF0.root	00:38:39	hep.wisc.edu	hep.wisc.edu		0.000	00:38:39

Currently open files

served by the collector

```
#begin
unique_id=1317441001444000
file_lfn=/store/mc/JobRobot/RelValProdTTbar/GEN-SIM-DIGI-RECO/..../Xyzz.root
start_time=1317440972
end_time=1317441001
read_bytes=1130140822
read_operations=269
read_min=1873046
read_max=8388608
read_average=4201266.996283
read_sigma=292199.597812
write_bytes=0
write_operations=0
write_min=0
write_max=0
write_average=0.000000
write_sigma=0.000000
read_bytes_at_close=1130140822
write_bytes_at_close=0
user_dn=/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=matevz/CN=475546/CN=Matevz
Tadel
user_vo=cms
user_role=cmsuser
client_domain=physics.ucsd.edu
client_host=desire
server_username=xrootd
server_domain=t2.ucsd.edu
server_host=uaf-5
#end
```

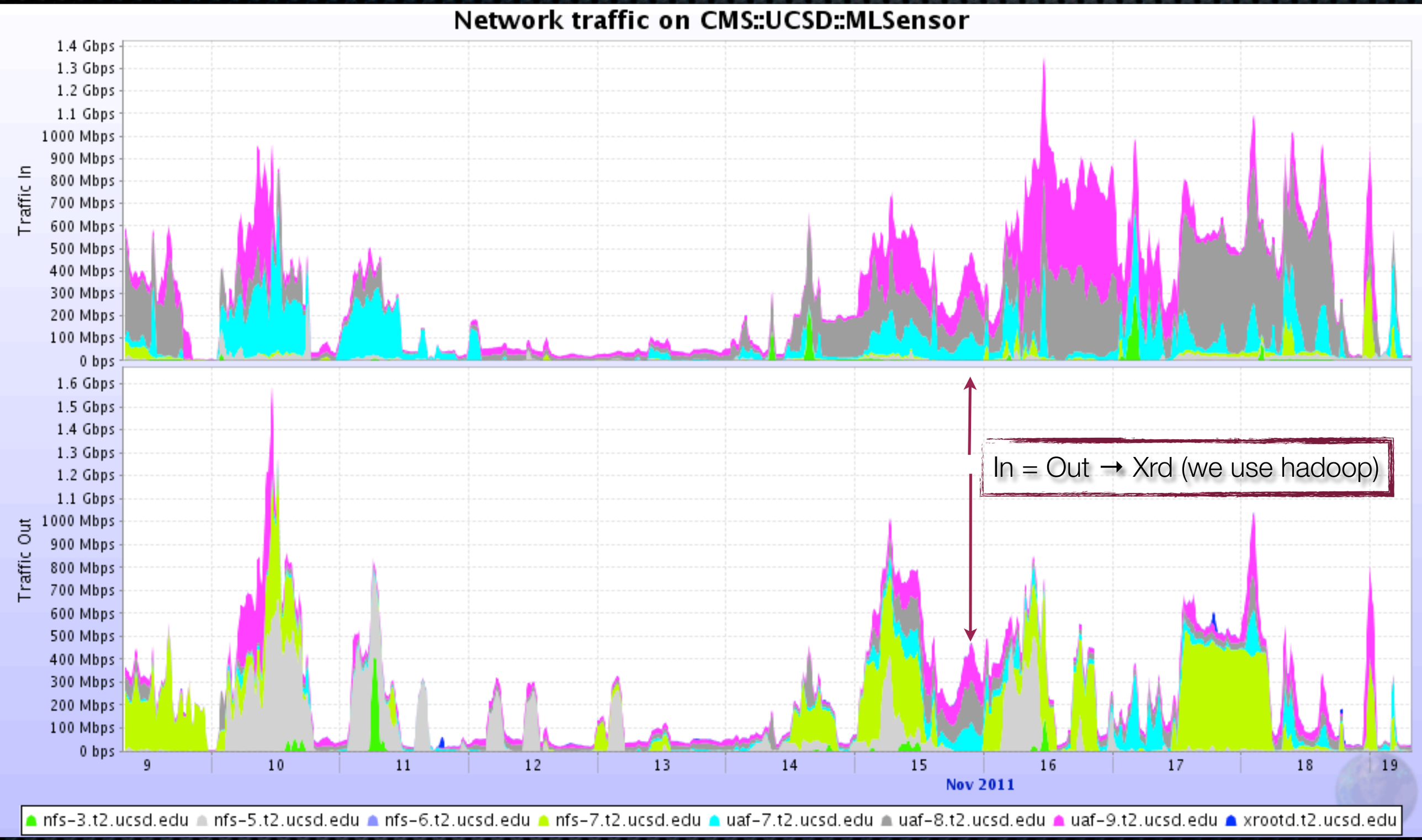
Report at file-close

- note read ops statistics
- this goes to Gratia, too
- maybe will be used by popularity fwk

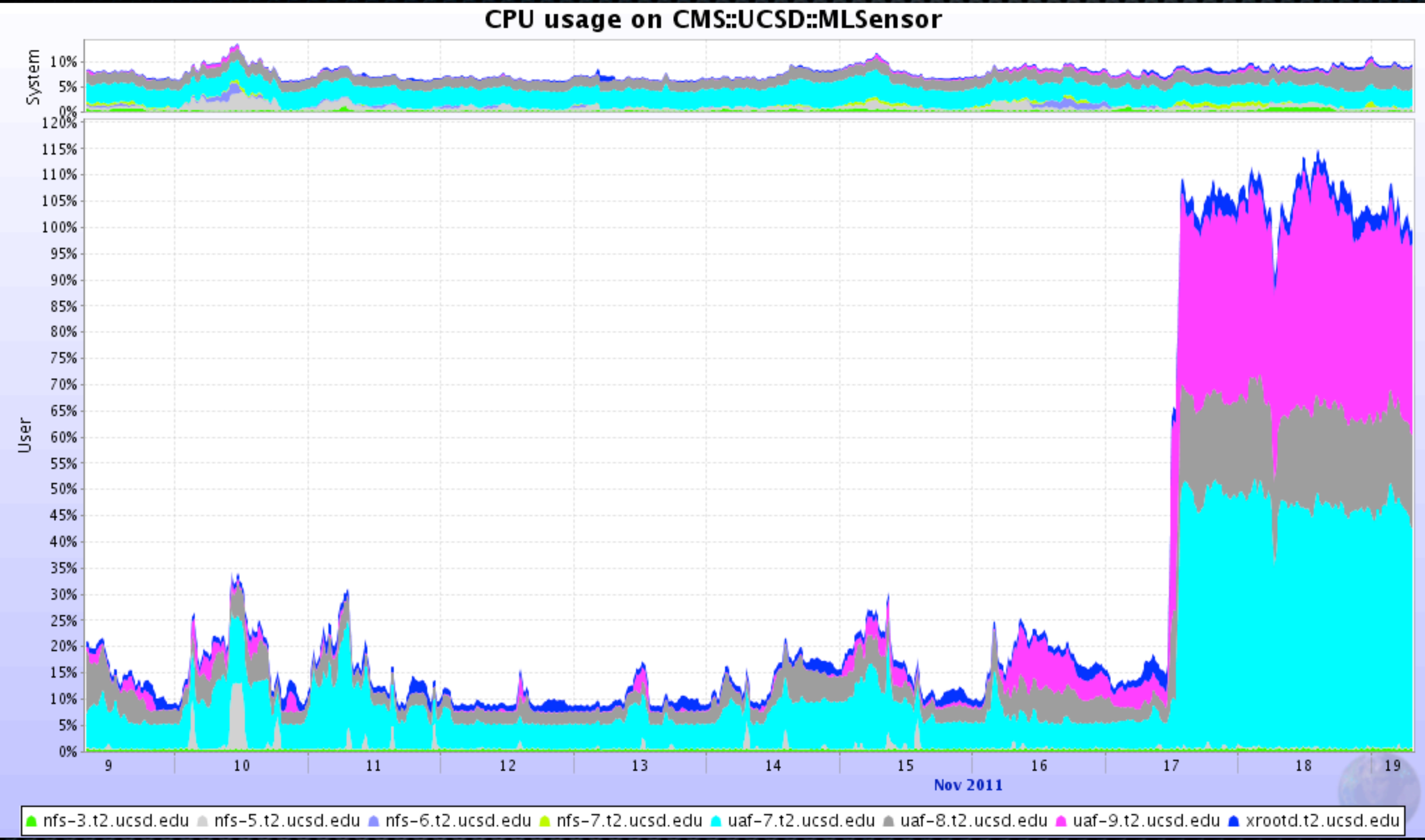
IV. Server health

- MLSensor by MonALISA team
  - Currently only used at UCSD
  - All parameters stored in ML repo
  - Selected graphs available from web interface
    - all NIC traffic
    - CPU / Memory usage
    - load averages
  - Together with summary monitoring info this gives you a rather good idea what is causing trouble

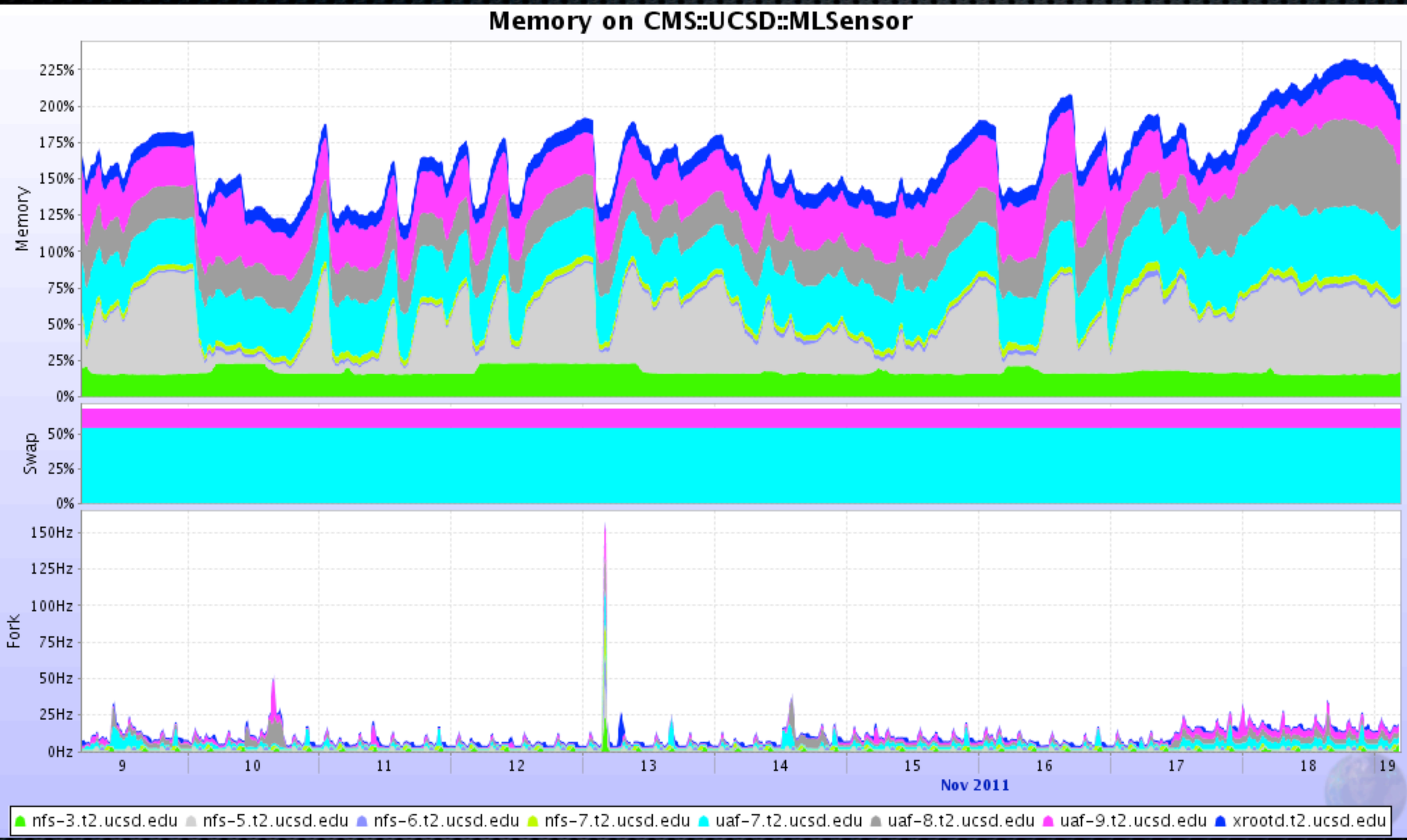
http://xrootd.t2.ucsd.edu/



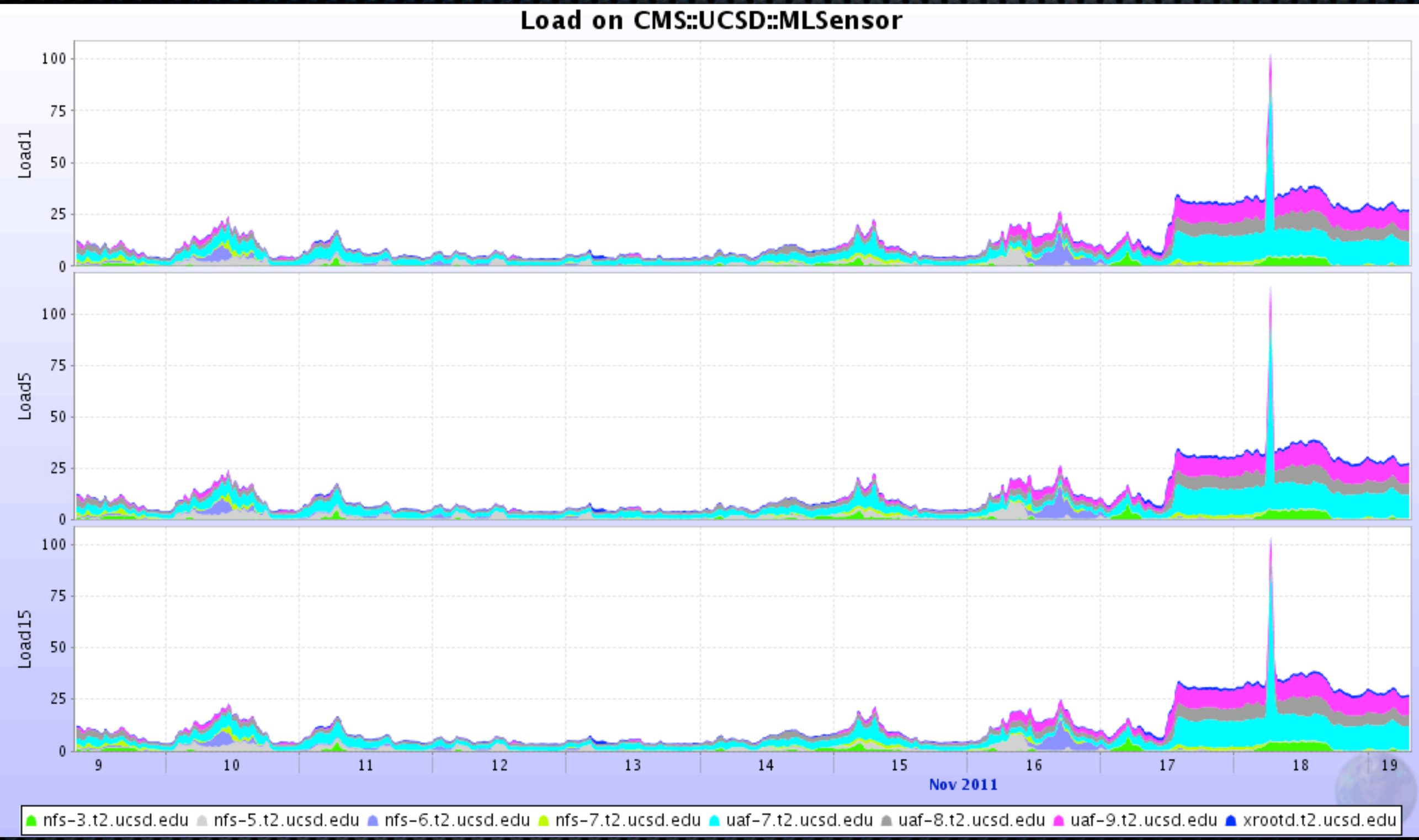
http://xrootd.t2.ucsd.edu/



http://xrootd.t2.ucsd.edu/



http://xrootd.t2.ucsd.edu/



Conclusion

- We have a rather complete monitoring system
  - Todo / plans:
    - Redirection monitoring
    - Connect with CMSSW monitoring
      - Use custom messages to send job id?
    - Storing detailed monitoring data into root files
    - Caching proxy - see how this works
    - Fancy stuff: 3D visualization, playback
- Xrd monitoring tested all the way through … fixed & improved
  - Can guarantee it is OK :)