CMS Tier-2 Sites Overview

Ken Bloom

OSG Storage Workshop

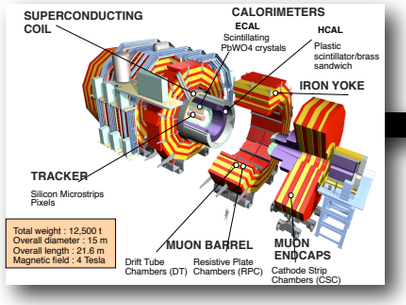
September 22, 2010

It's good to be back home!

CMS distributed computing model

Computing resources are arranged in a hierarchy, with different clusters given different tasks and configured appropriately.

T0 site is at CERN; 7 T1's in 7 nations; ~50 T2 sites in 23 nations (7 in US)!



Tier-0

Prompt Reconstruction,
Calibration Streams
Archival Copy of RAW and
first RECO data

Re-Reconstruction, Skimming
Archival and Served Copy of RECO
Archival Storage for Simulation

Tier-1 Tier-1 Tier-1

Tier-2 Tier-2 Tier-2 Tier-2 Tier-2

Simulated event production
and primary resource for
analysis by physicists

What happens at Tier 2

T2's each have substantial processing (~1500 CPU's) and disk (~400 TB) resources, no tape. Two main activities:

- Simulation jobs that are centrally controlled ("scheduled")
  - Uses about half the CPU, small amount of disk to hold output before transfer to T1 for archiving
- Analysis jobs that are submitted by users ("chaotic")
  - Uses the other half of the CPU and the bulk of the disk, mostly for input datasets (real or simulation), some for output
  - Disk is just a cache; datasets to be refreshed regularly
- All jobs are submitted through grid interfaces; no user logins

Major tasks are thus job hosting, data hosting and data transfer.

A very interesting program to manage:

- Mix of scheduled and chaotic activities
- Elements of national and international politics
- Part of CMS that the largest number of physicists put their hands on!

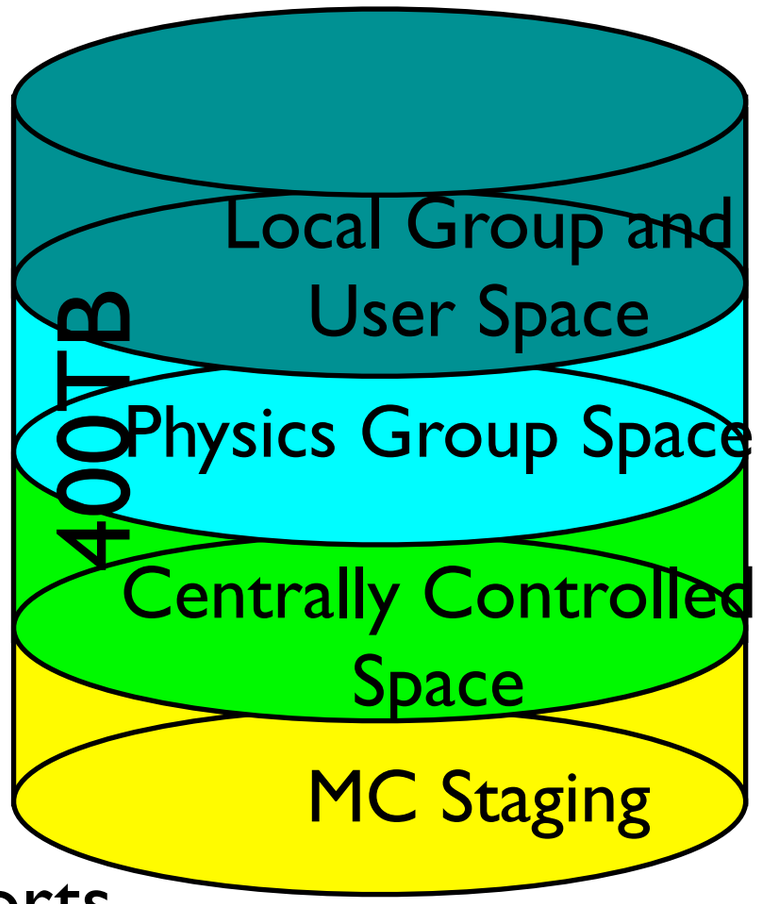
CMS T2 analysis model: data distribution

In CMS, jobs go to the data -- distribute data for efficient resource use.

Nominal T2 storage is 400 TB x ~50 T2 sites = huge!

Some amount set aside for centrally-controlled activities (e.g. distribution of datasets of wide interest) and local activities (e.g. making user-produced files grid accessible).

But bulk is allocated to the various CMS analysis groups for distribution of "their" interesting data.



T2_US_Nebraska Group Usage

Group	Subscribed	Resident
AnalysisOps	85.54 TB	85.54 TB
DataOps	943.35 GB	943.35 GB
FacOps	1008.33 GB	1008.33 GB
b-tagging	22.92 TB	22.92 TB
local	59.53 TB	59.53 TB
top	24.28 TB	24.28 TB
tracker-dpg	8.41 TB	8.41 TB
undefined	17.64 TB	17.64 TB
	220.22 TB	220.22 TB

- 17 such groups in CMS
- Currently no site supports more than 3 groups, no group affiliated with more than 5 sites, manageable number of communication channels
- 7 US T2's support all groups
- Data bookkeeping systems are critical!

Storage evolution (as told by KB)

In the beginning, there was dCache.

- A unifying element of the US T2 program, just like OSG
- A strategic strength for us, because of the excellent dCache knowledge base at Fermilab that we could rely on for support
- dCache was one of the few disk management systems that was proven to work on the scale needed for T2 clusters

But it wasn't a perfect solution for the T2 application.

- dCache is a "disk cache" -- there was a tape backend assumed. Thus limited protections against loss of files on disk.
  - File replication strategy was not super-robust
- But T2 sites don't have tape!
  - Lost files couldn't be recovered from tape
  - Difficult to drain a dCache pool

The Great Storage Revolt of 2009

In 2008-09, sites started searching for alternative solutions.

- Nebraska started developing a Hadoop FS instance that could be interfaced with all the CMS functionality
- UCSD, Caltech, Wisconsin got involved in development work; Wisconsin later decided to stick with dCache (for now)
- Florida HPC center was already investigating the use of Lustre, and T2 center saw useful features
- Meanwhile, many improvements were made in dCache during this period that made it more robust for the T2 application.

So now there is diversity in our storage technology, and we promote it as a strategic strength:

- Nebraska, UCSD, Caltech use Hadoop exclusively
- Florida uses both dCache and Lustre, evolving towards Lustre only
- Wisconsin, Purdue, MIT are happy with dCache

US CMS Tier-2 today

Currently, each of the seven US T2 sites is funded by NSF to support 2 FTE for operations and to purchase $250K/year of hardware ("build to cost").

Since the US is about 1/3 of the CMS headcount, we need to provide about 1/3 of the computing resources, thus US T2 sites are larger than nominal.

Minimum deployment goal for this FY: 7760 HS06 and 570 TB disk per site

- 570 TB is the amount available for data hosting, after accounting for replication etc., so amount of raw disk typically x2 bigger (!)

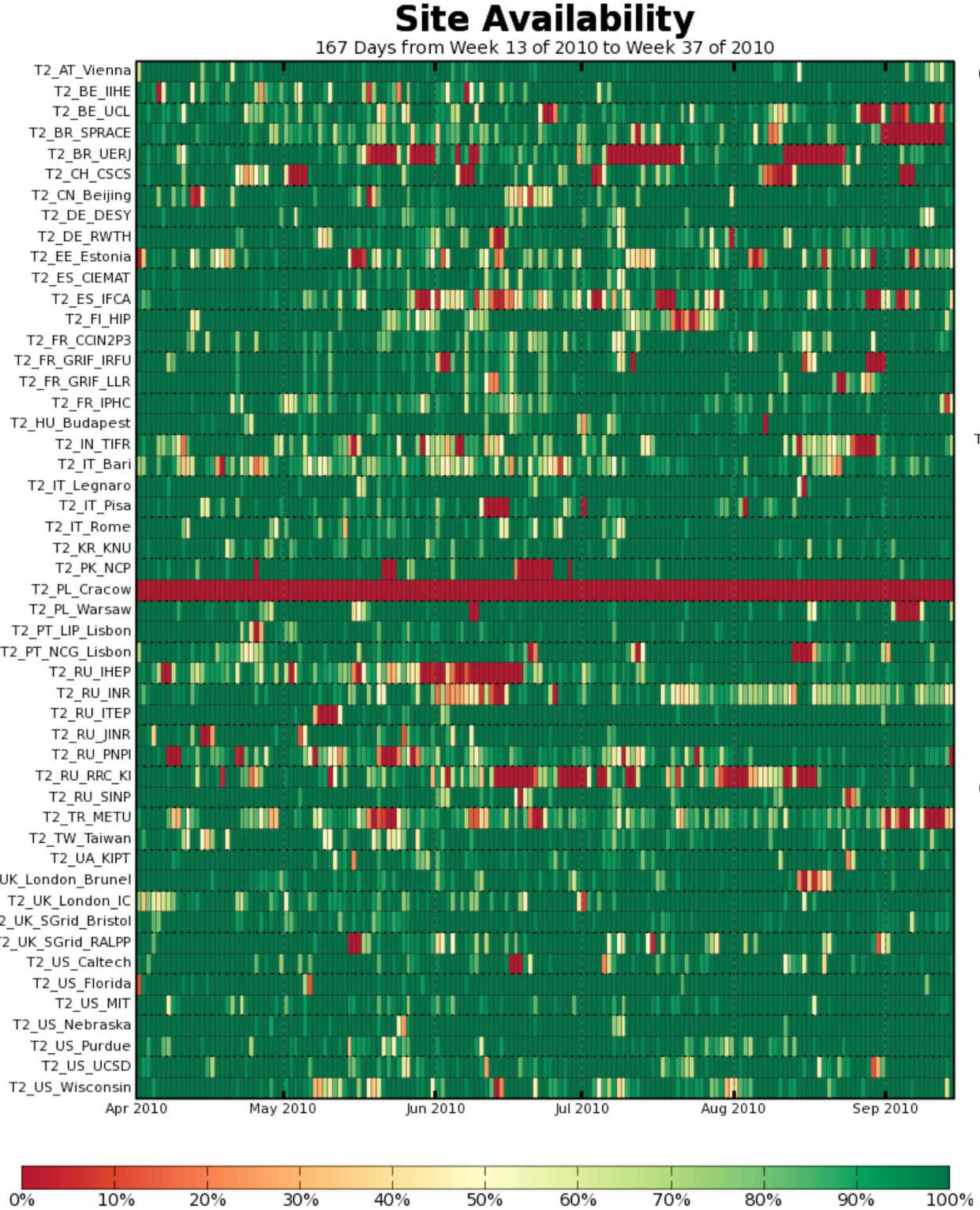
Site	CPU (HS06)	Batch Slots	Space for hosting (TB)	WAN (Gb/s)	Last update
T2_US_Caltech	8450	962	509	20	8/31/10
T2_US_Florida	10655	1348	425	10	3/8/10
T2_US_MIT	9252	1270	395	10	3/8/10
T2_US_Nebraska	18707	1700	800	10	8/31/10
T2_US_Purdue	20103	2012	525	10	3/8/10
T2_US_UCSD	9600	1168	502	10	8/20/10
T2_US_Wisconsin	10400	1220	550	10	8/17/10

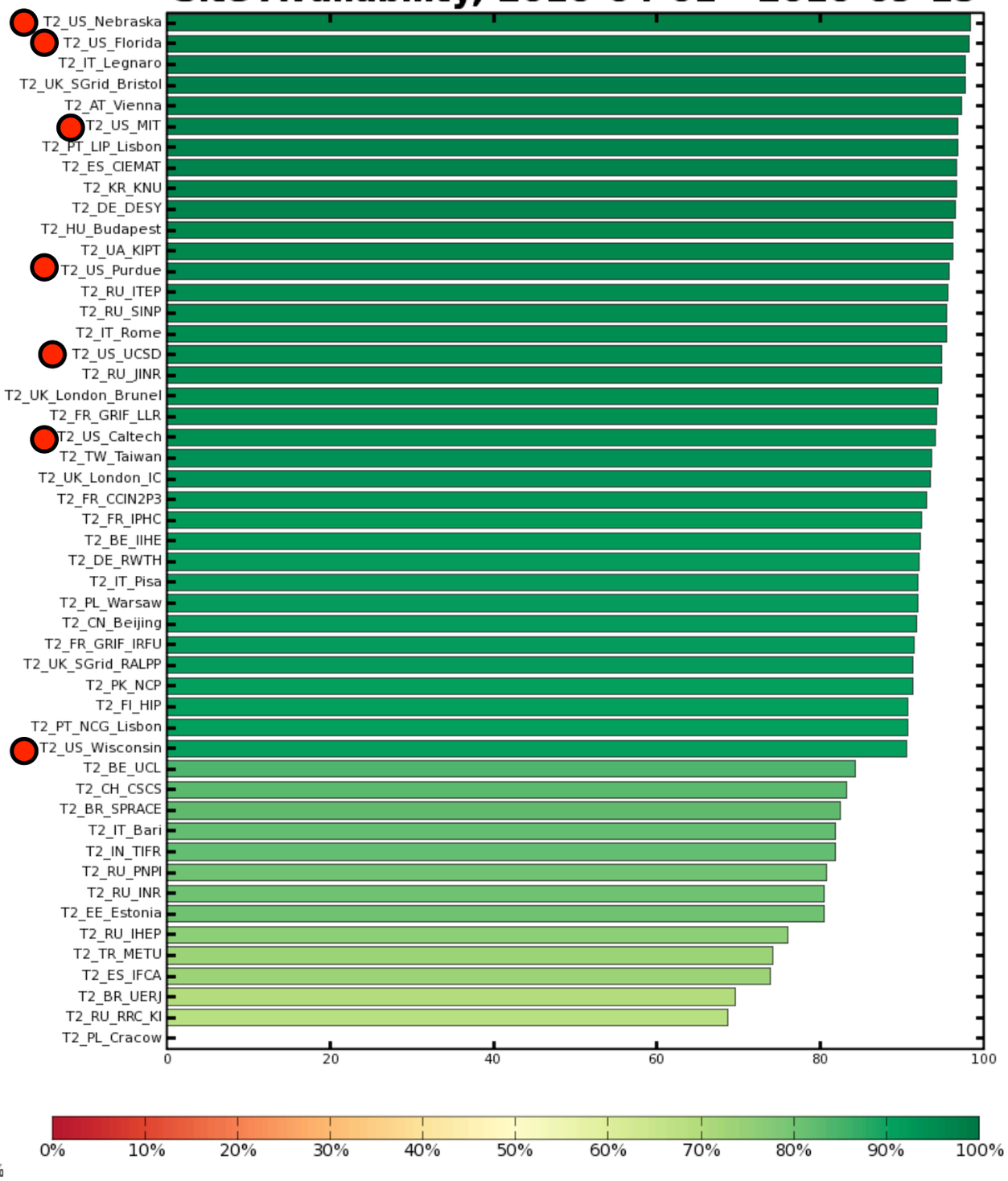
Storage deployment is a bigger challenge than CPU deployment.

Too many slides

What follows is probably too many plots documenting the performance of the CMS T2 sites, in the US and elsewhere, since the 2010 LHC run started in earnest in April.

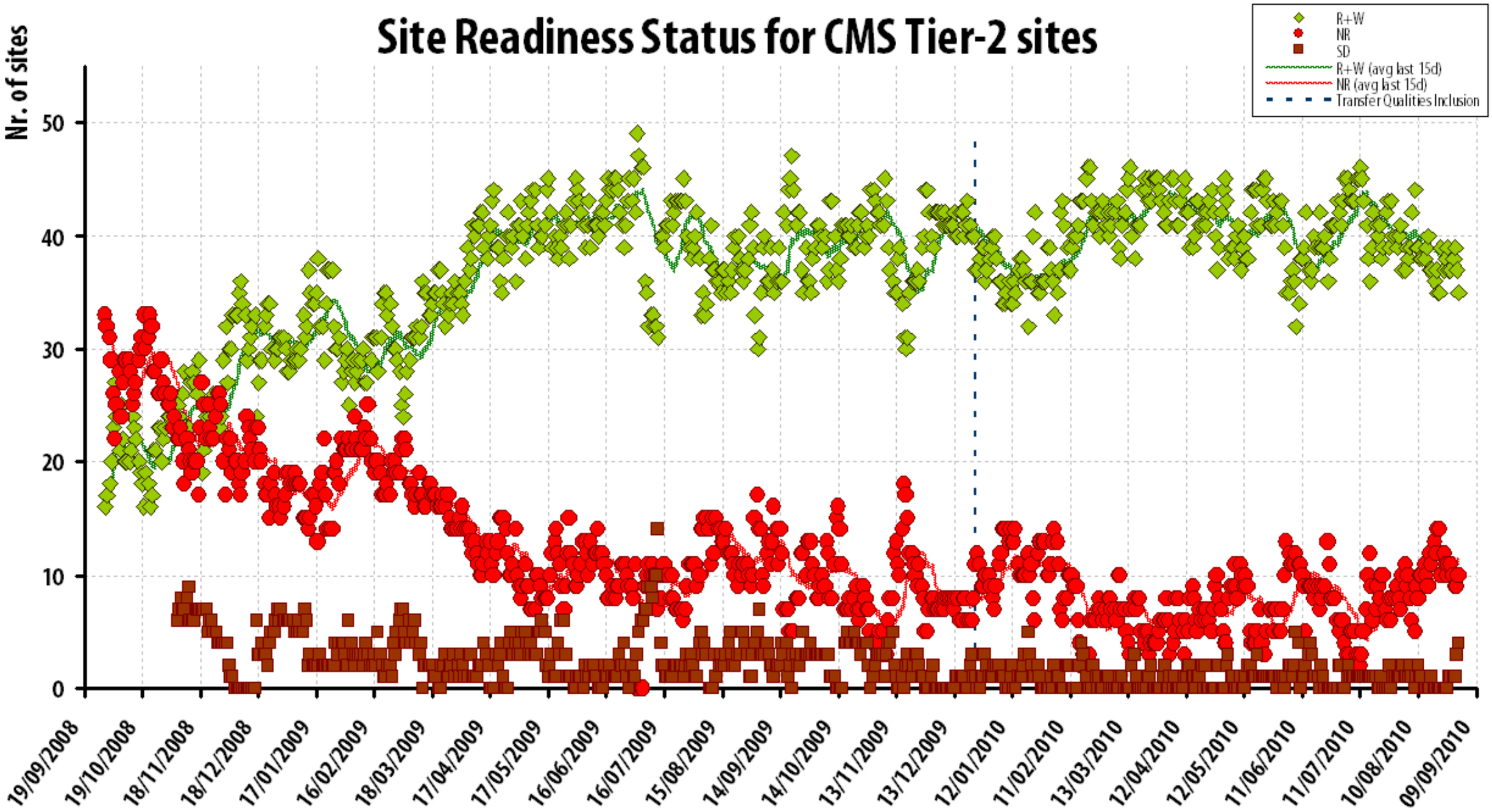
Site availability monitoring



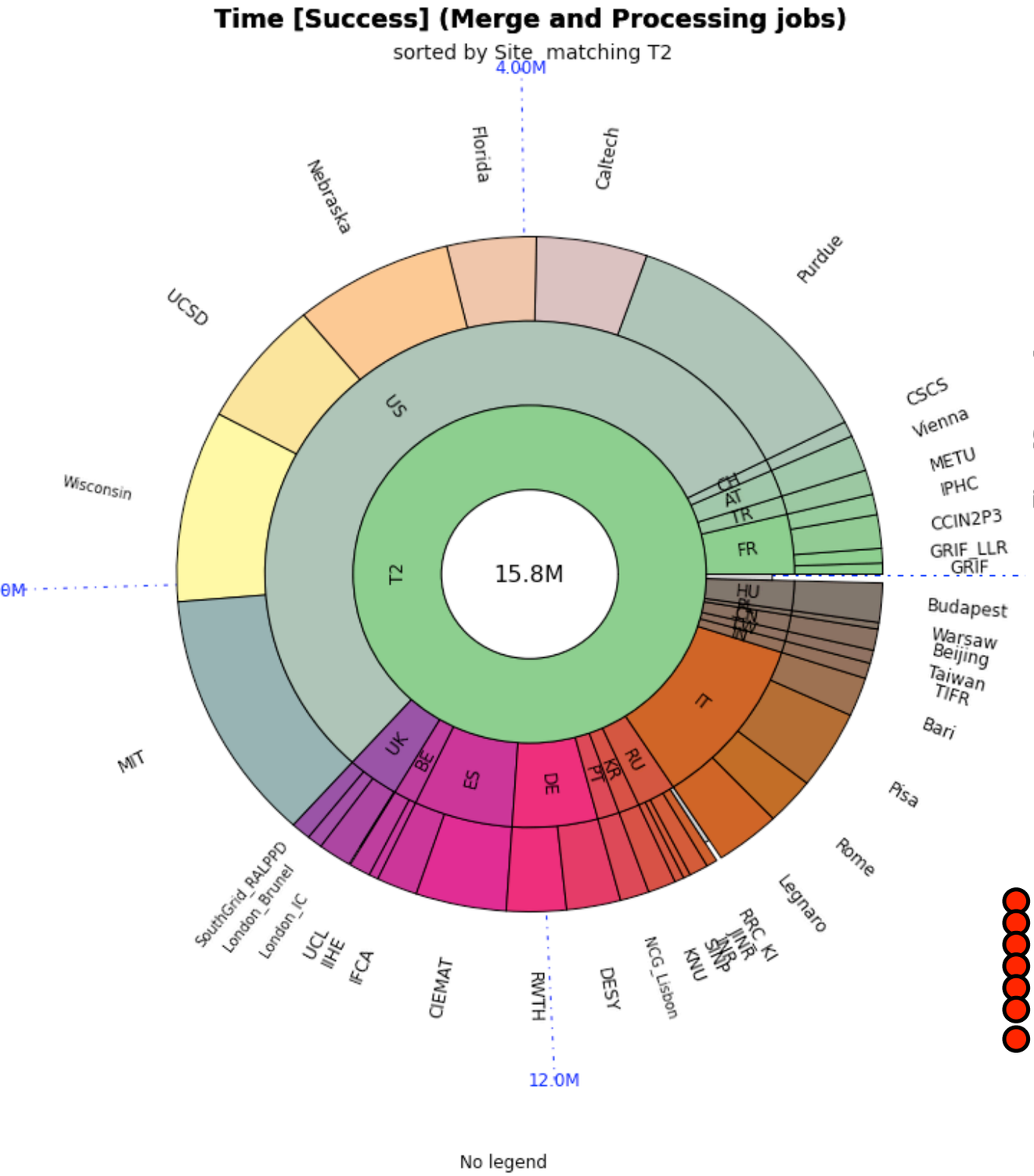


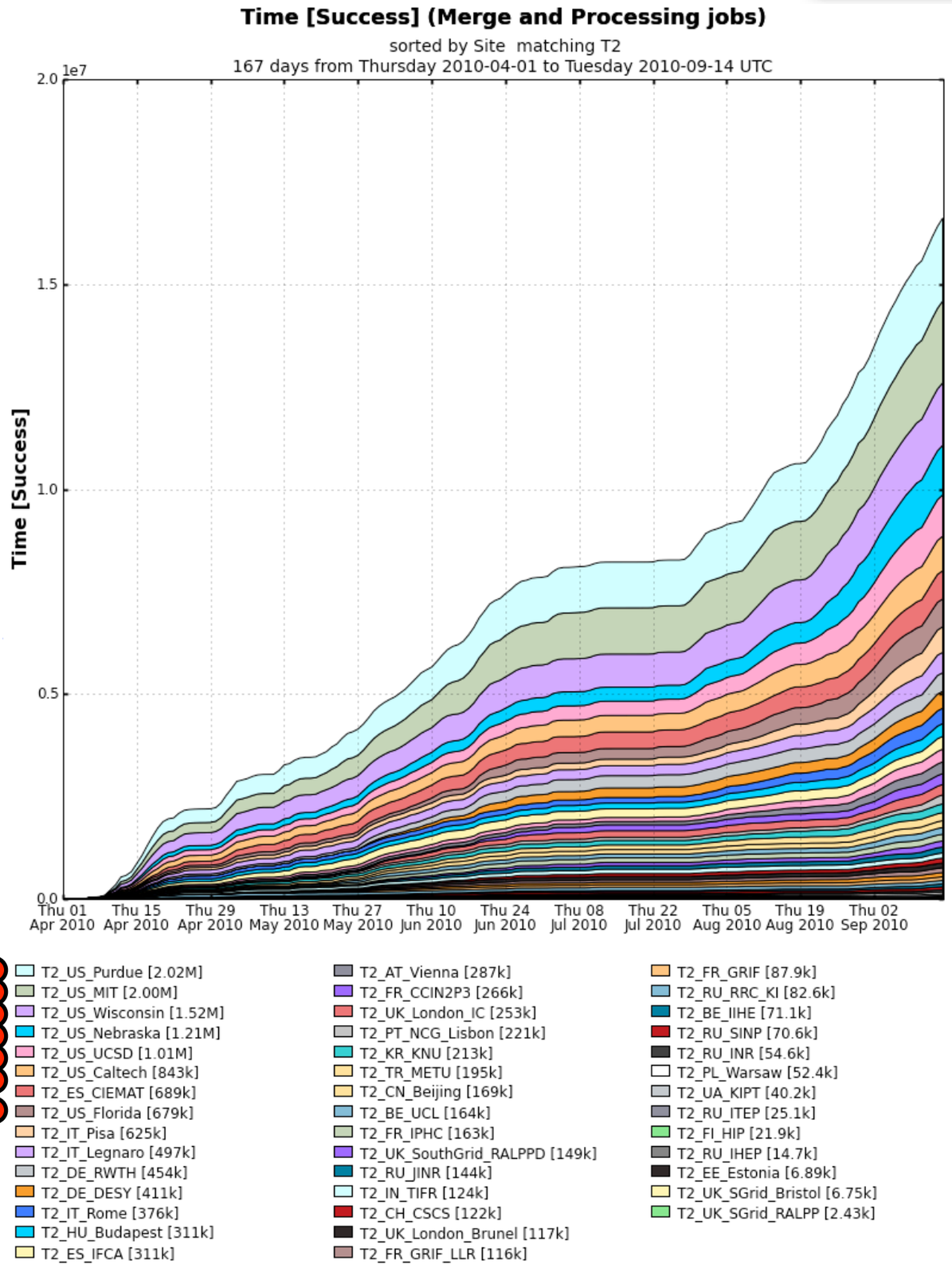
Site readiness

Every day sites are evaluated on the basis of availability tests, success of "robot" job submissions and success of inbound/outbound data transfers

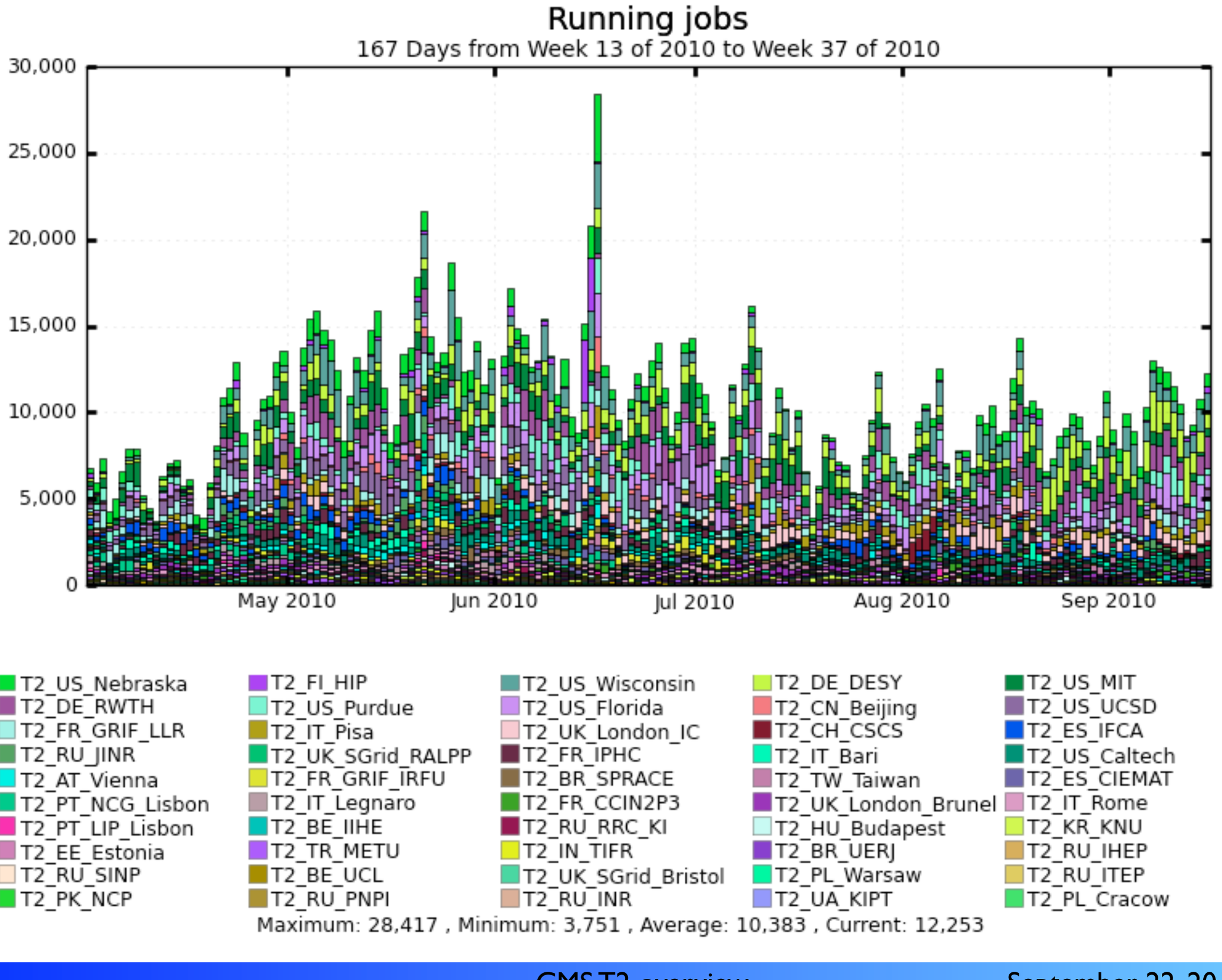


Simulation production

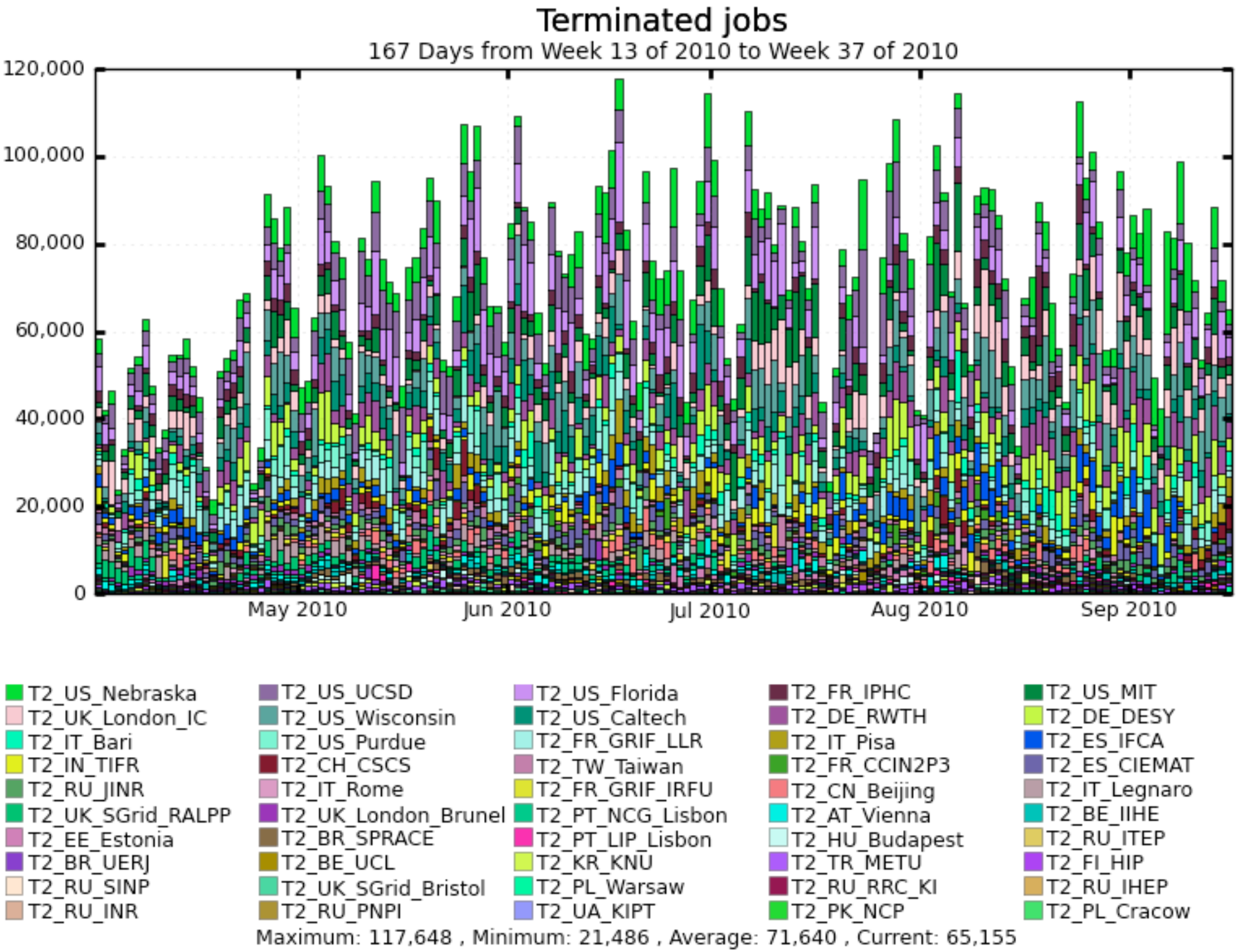




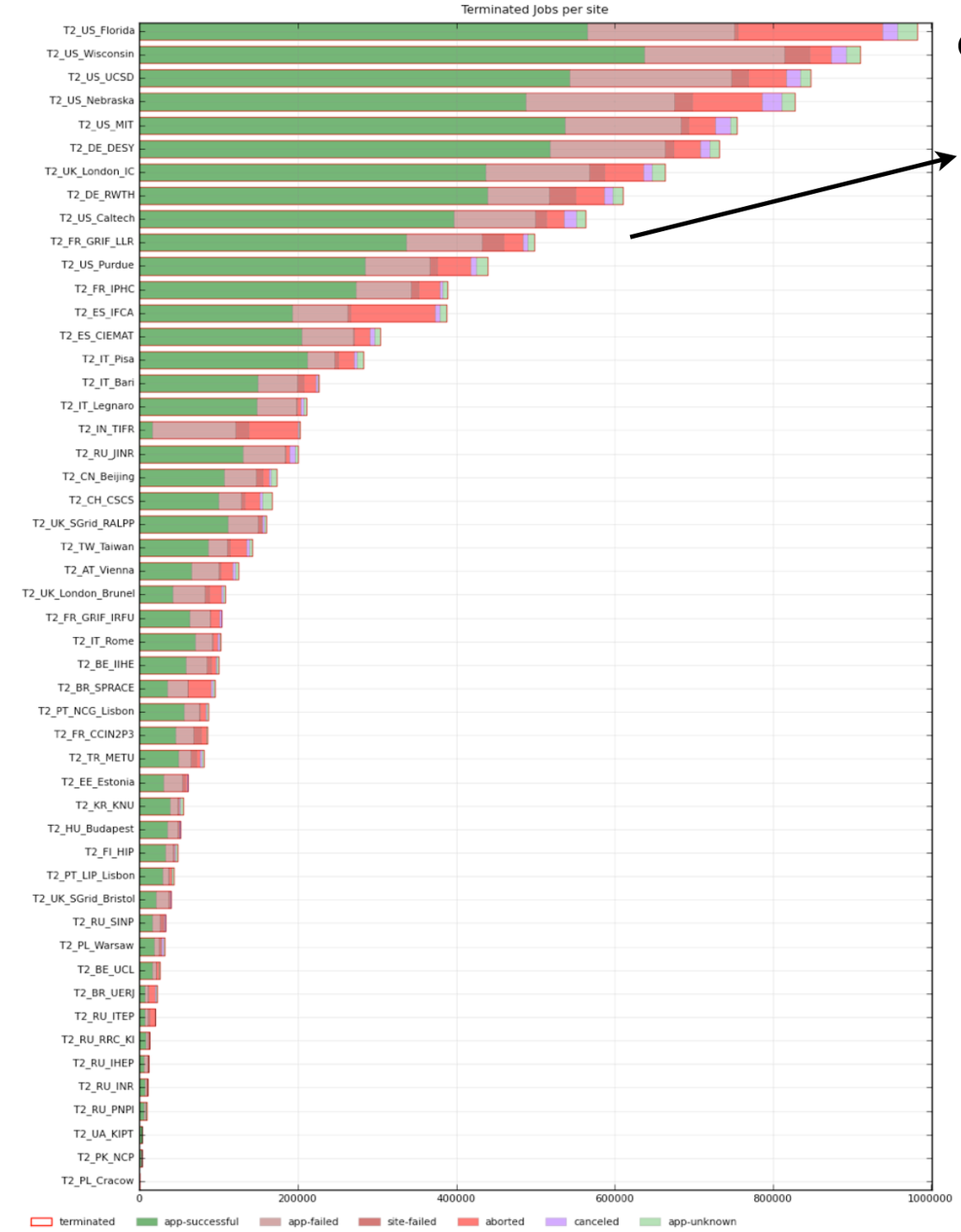
Analysis job hosting

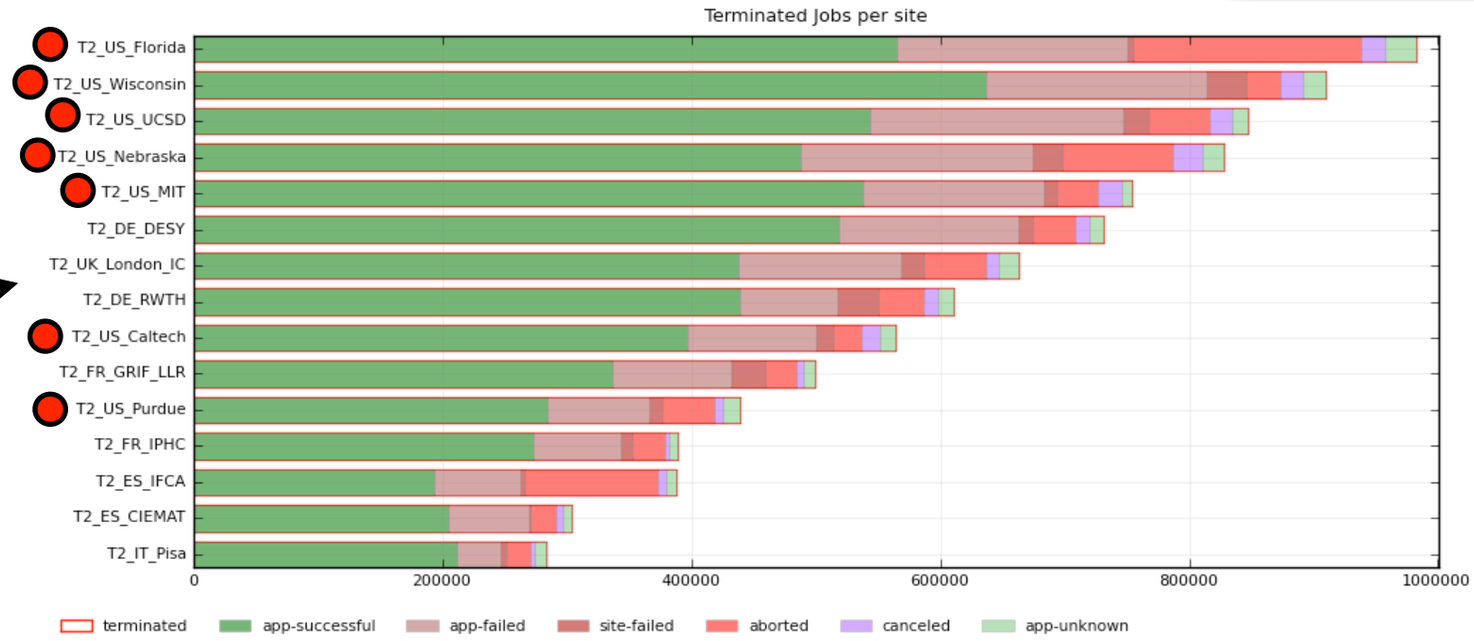


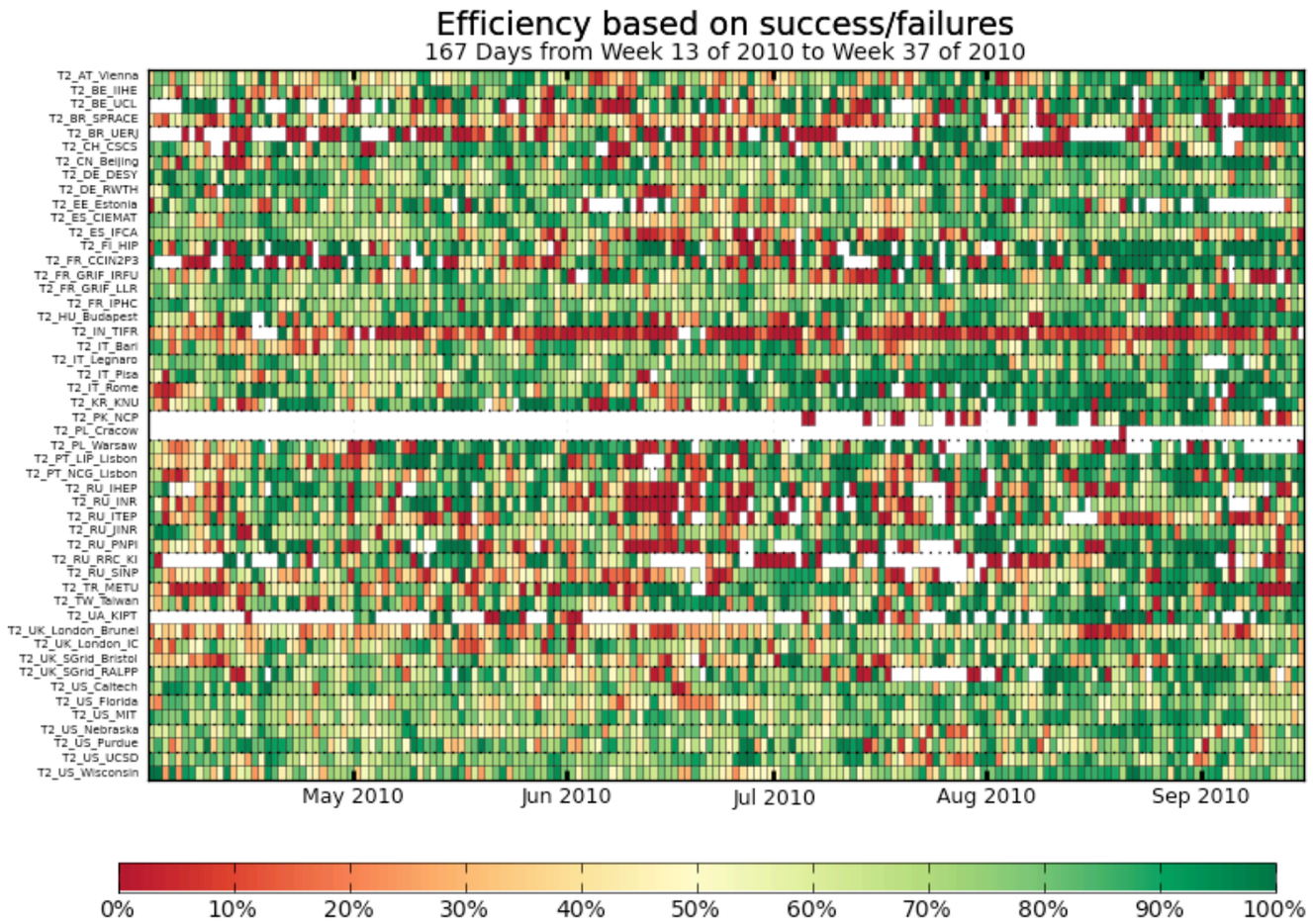
Analysis job hosting



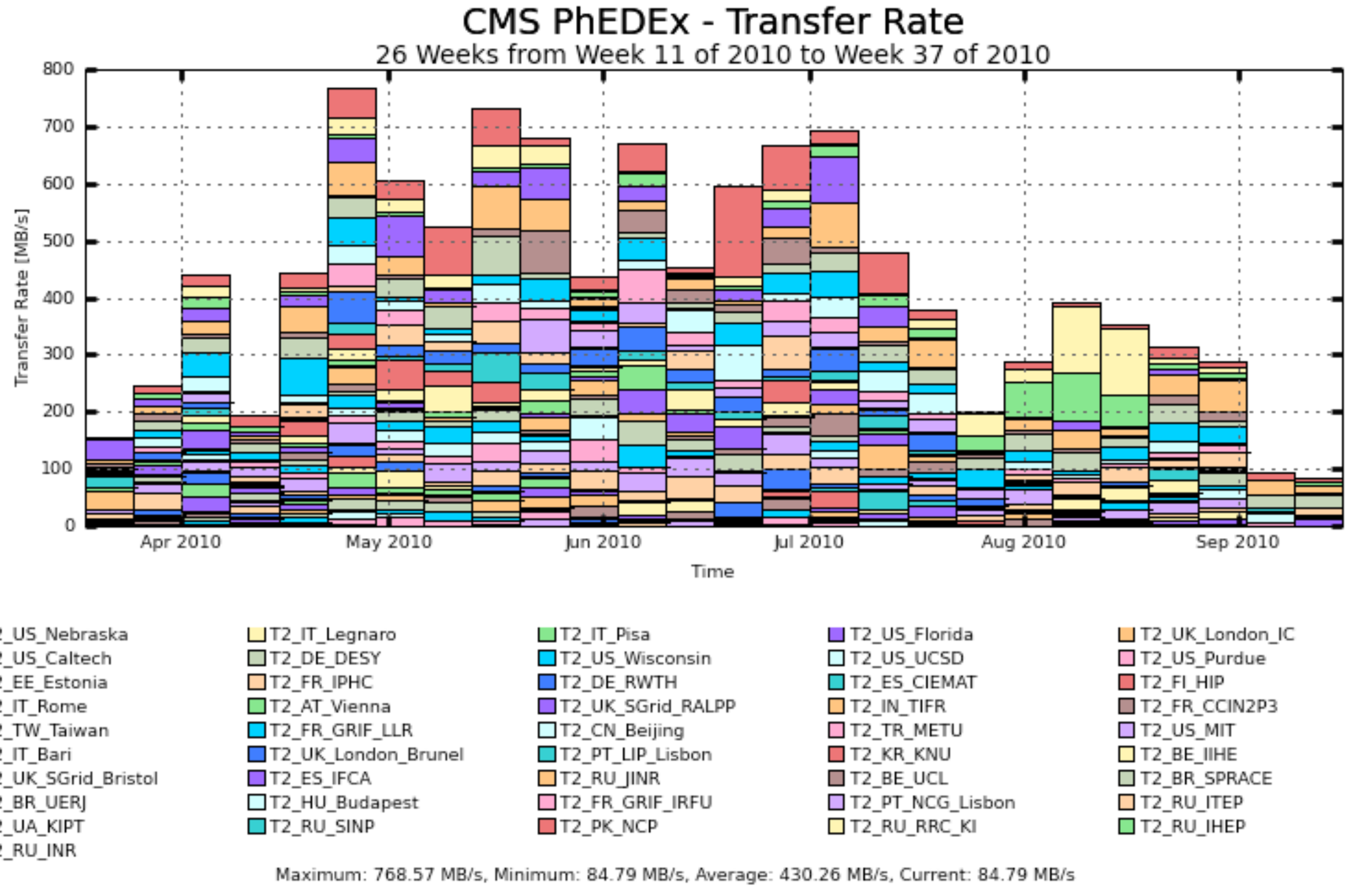
Analysis job hosting



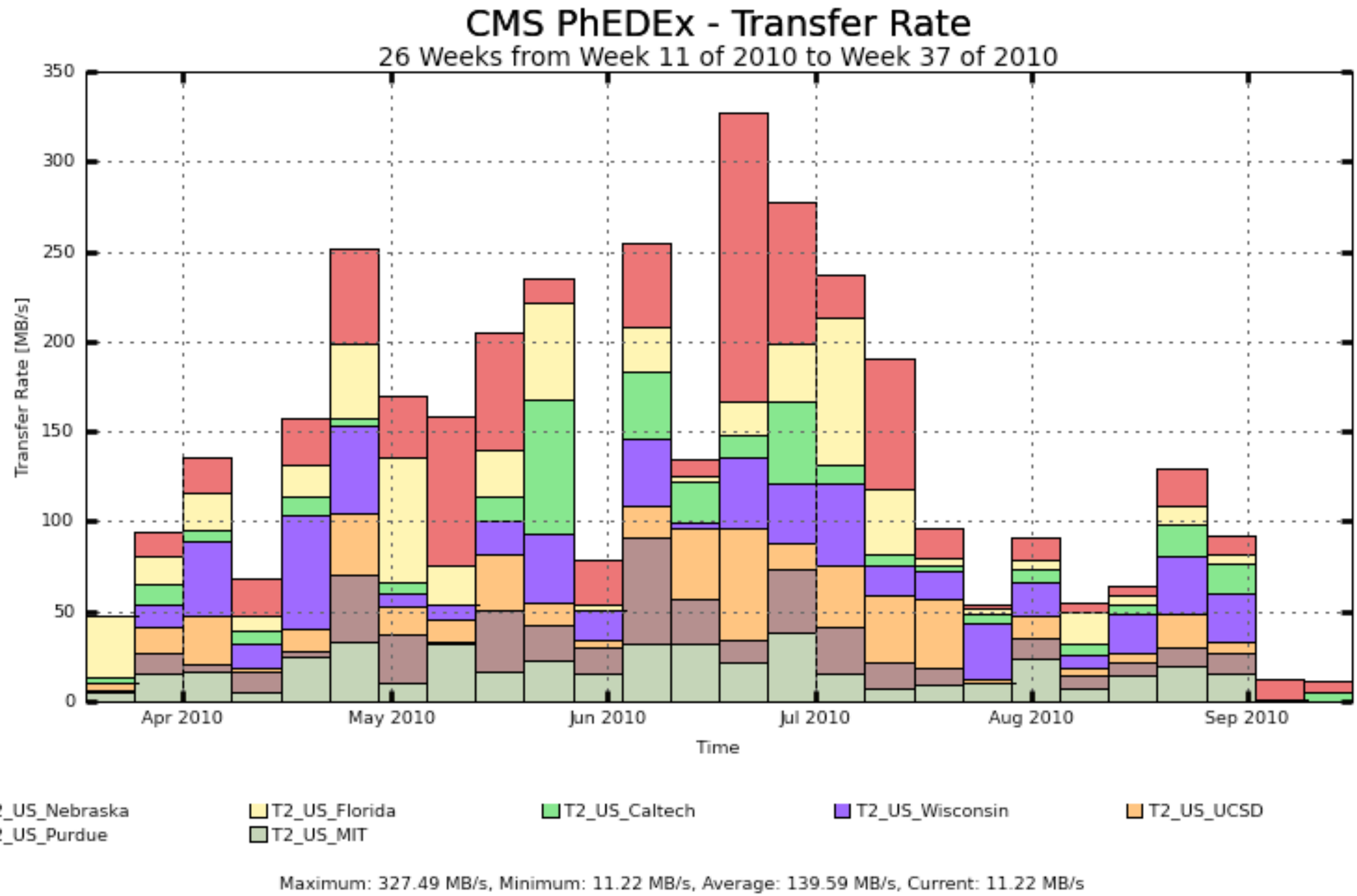




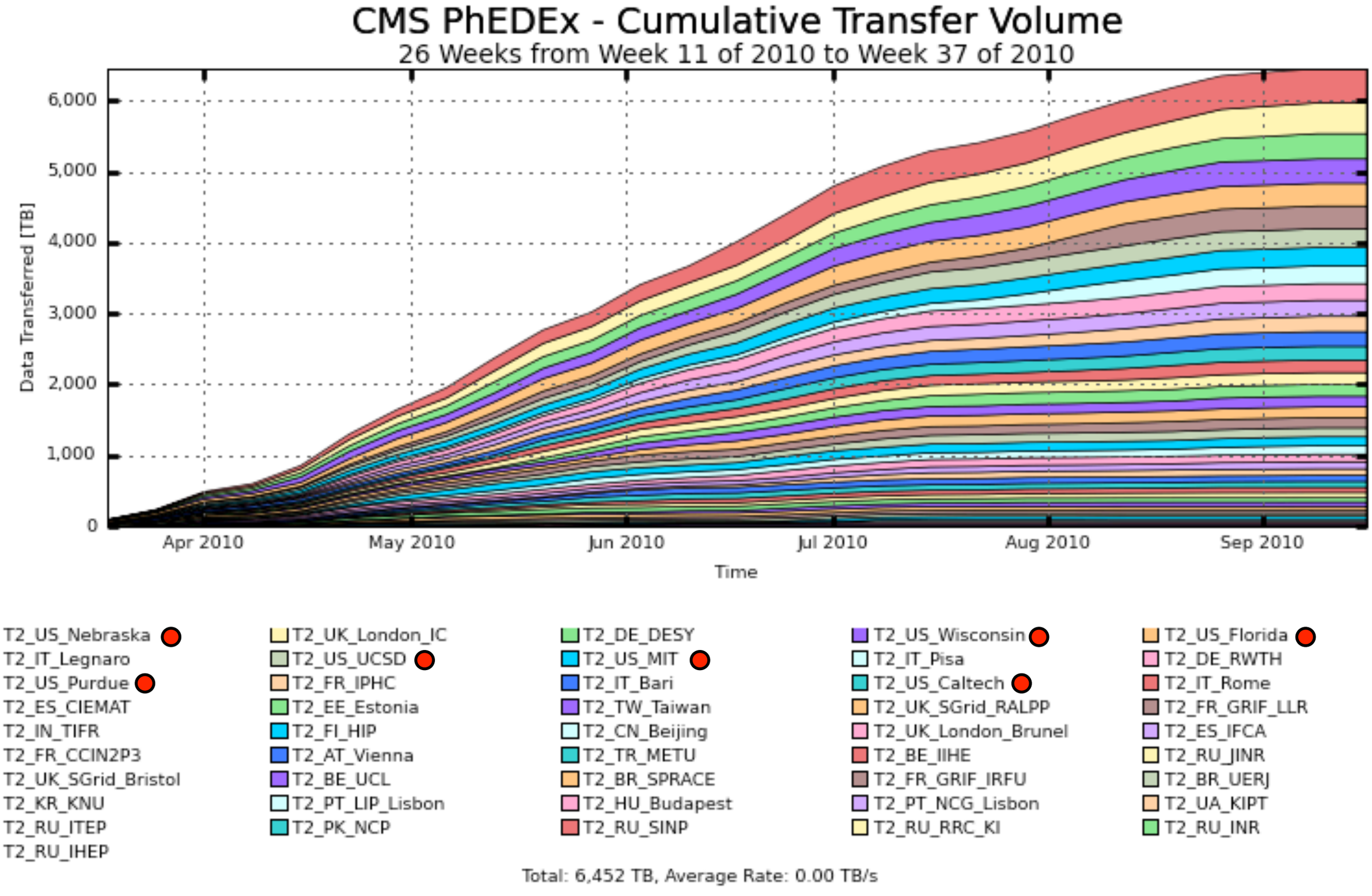
Transfers from T1 to T2



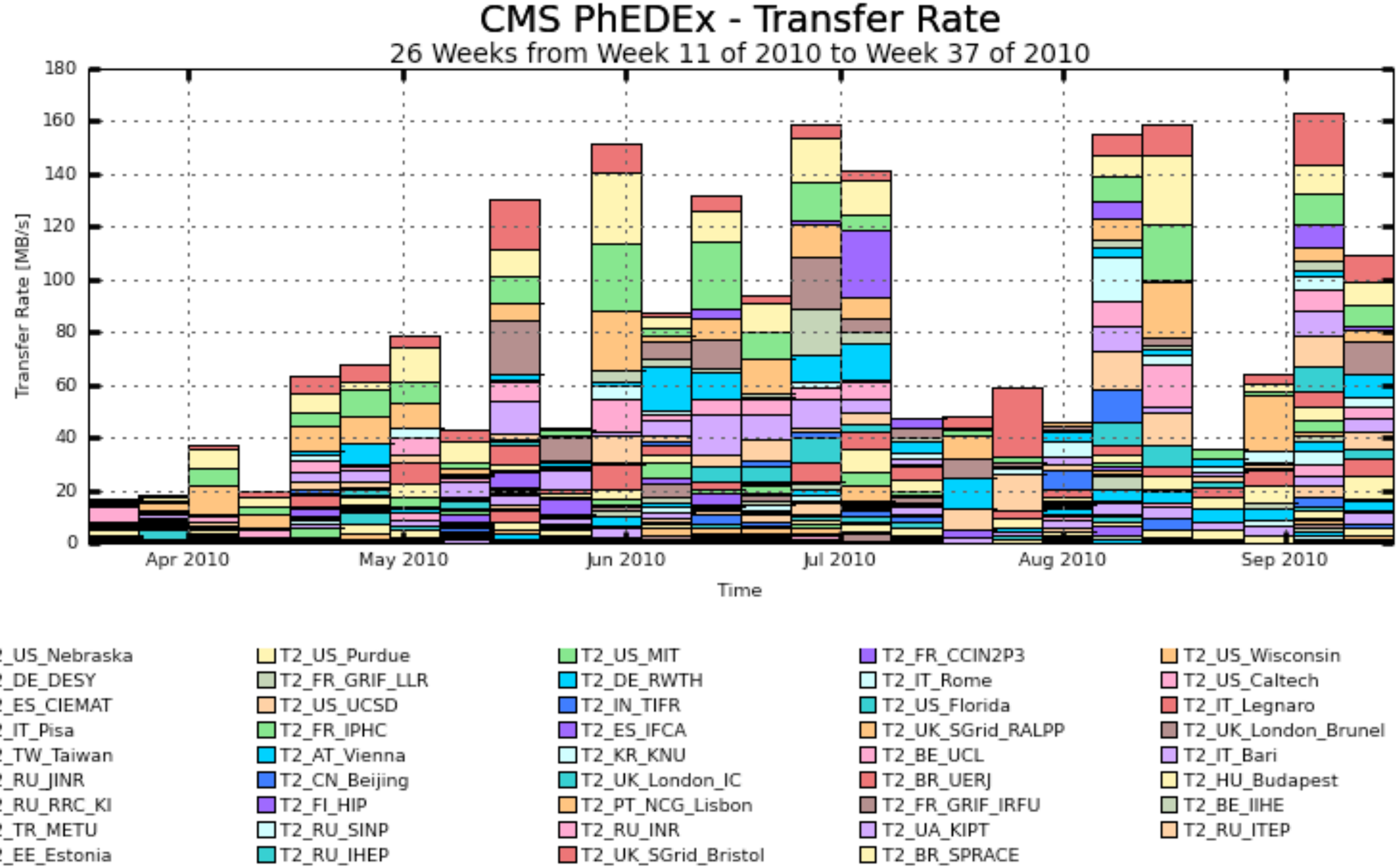
Transfers from T1 to US T2



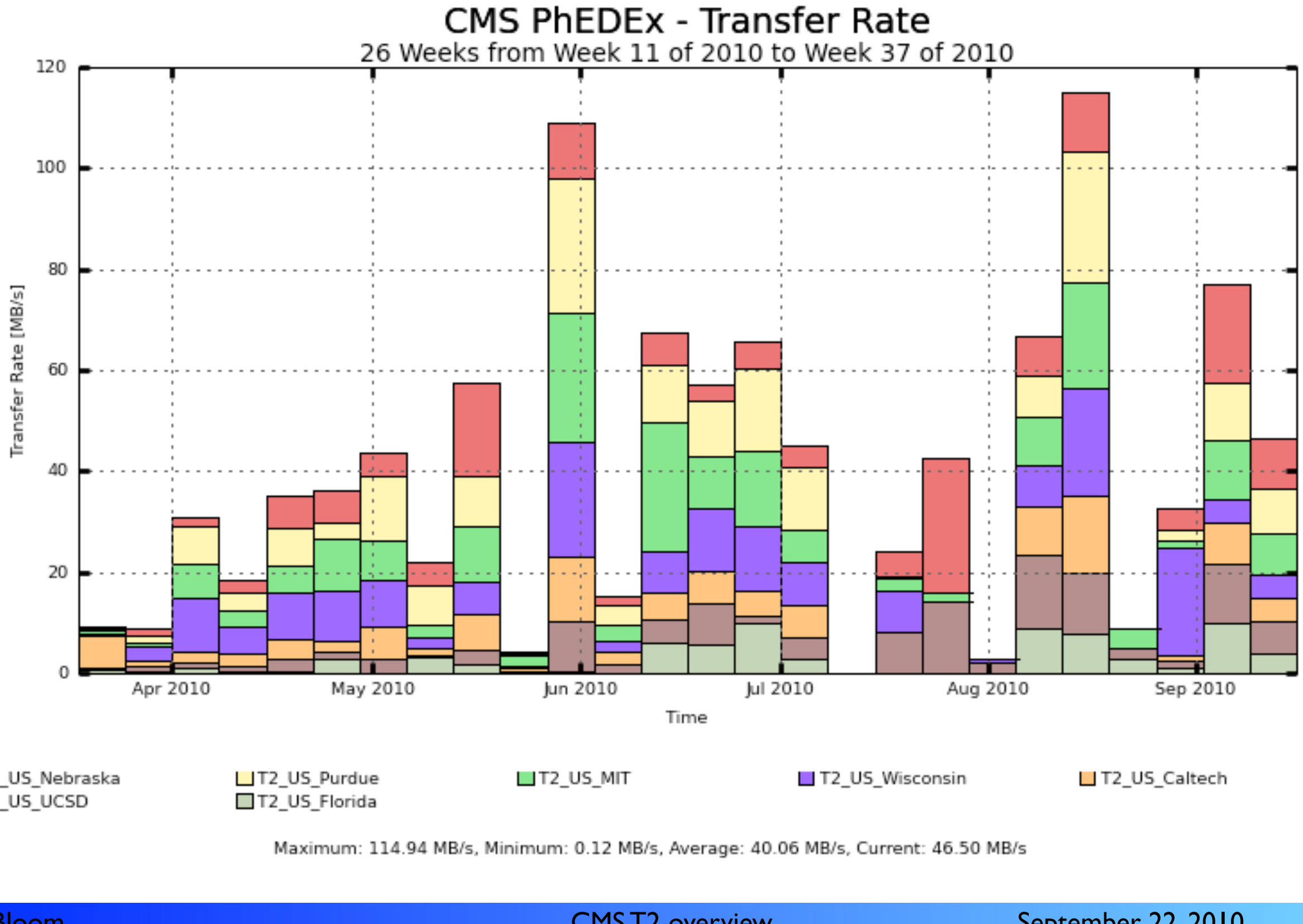
Transfers from T1 to T2



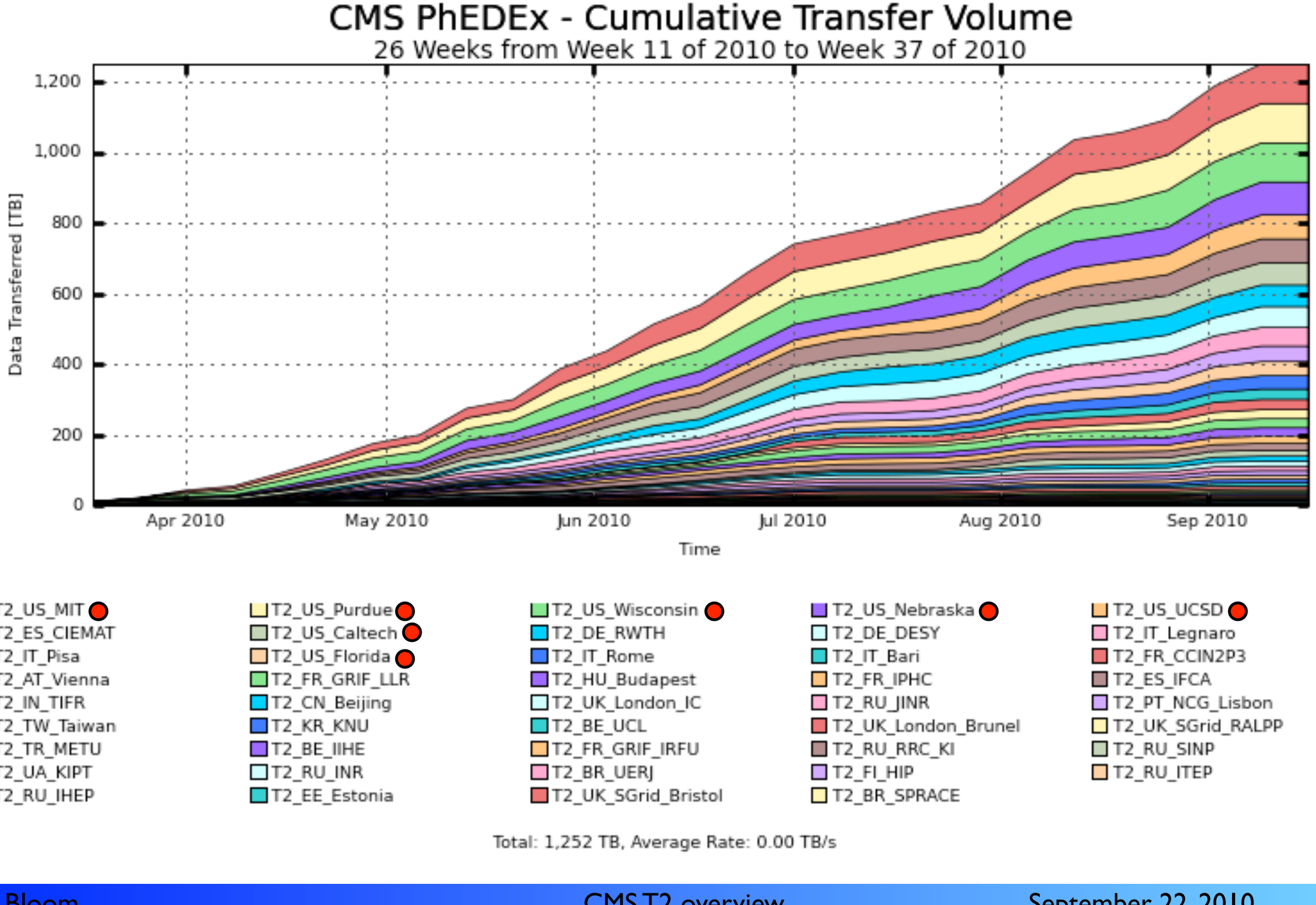
Transfers from T2 to T1



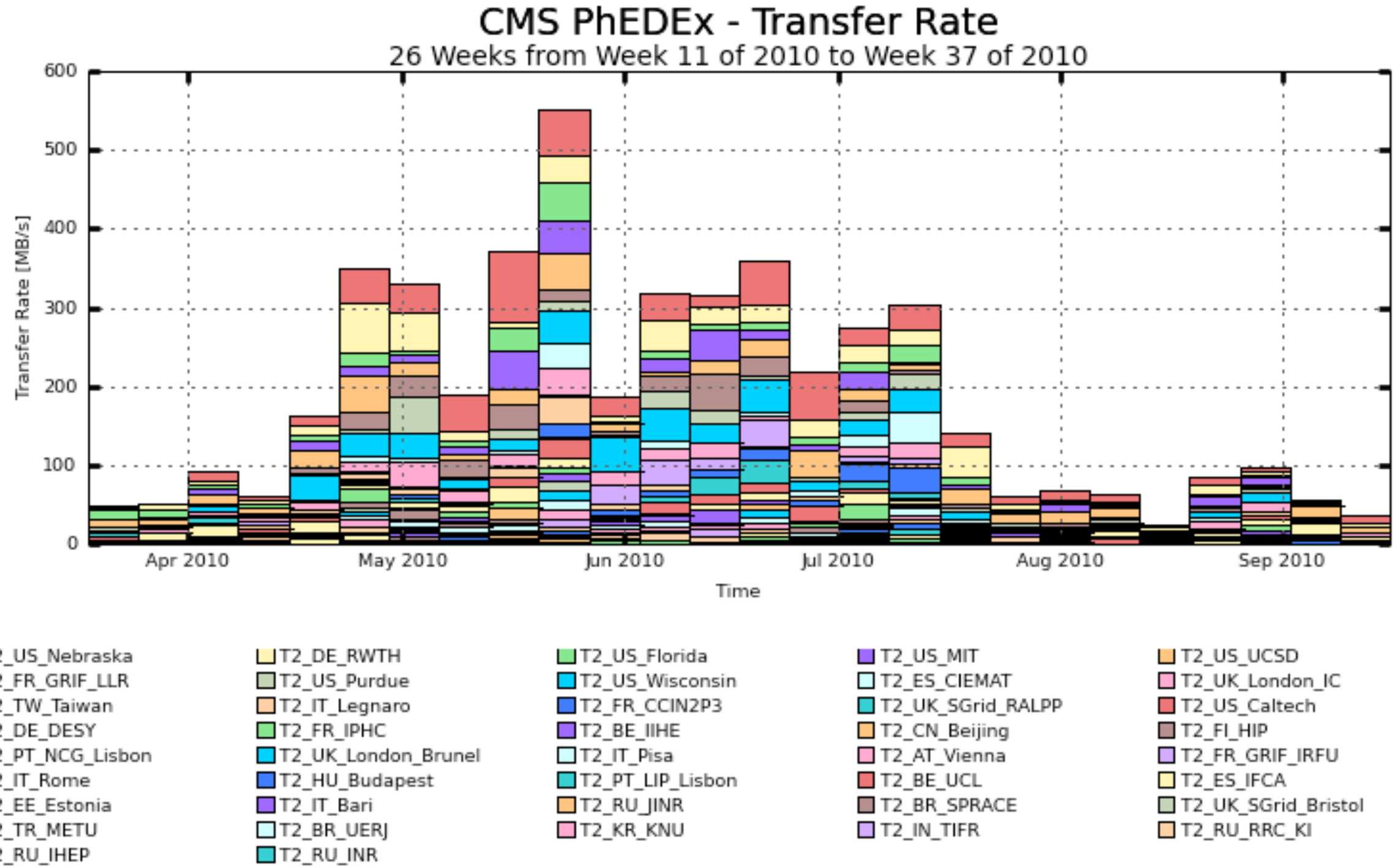
Transfers from US T2 to T1



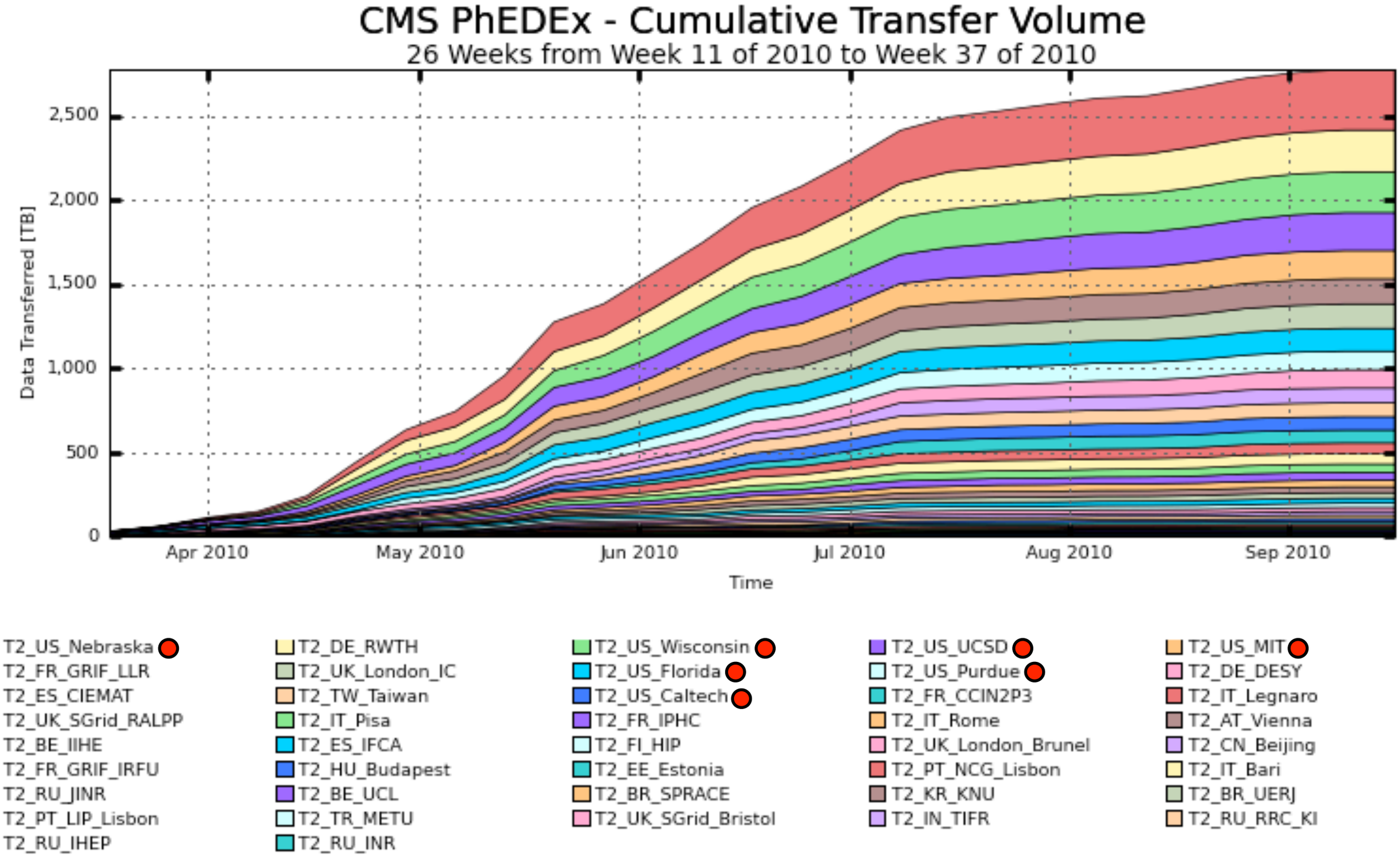
Transfers from T2 to T1



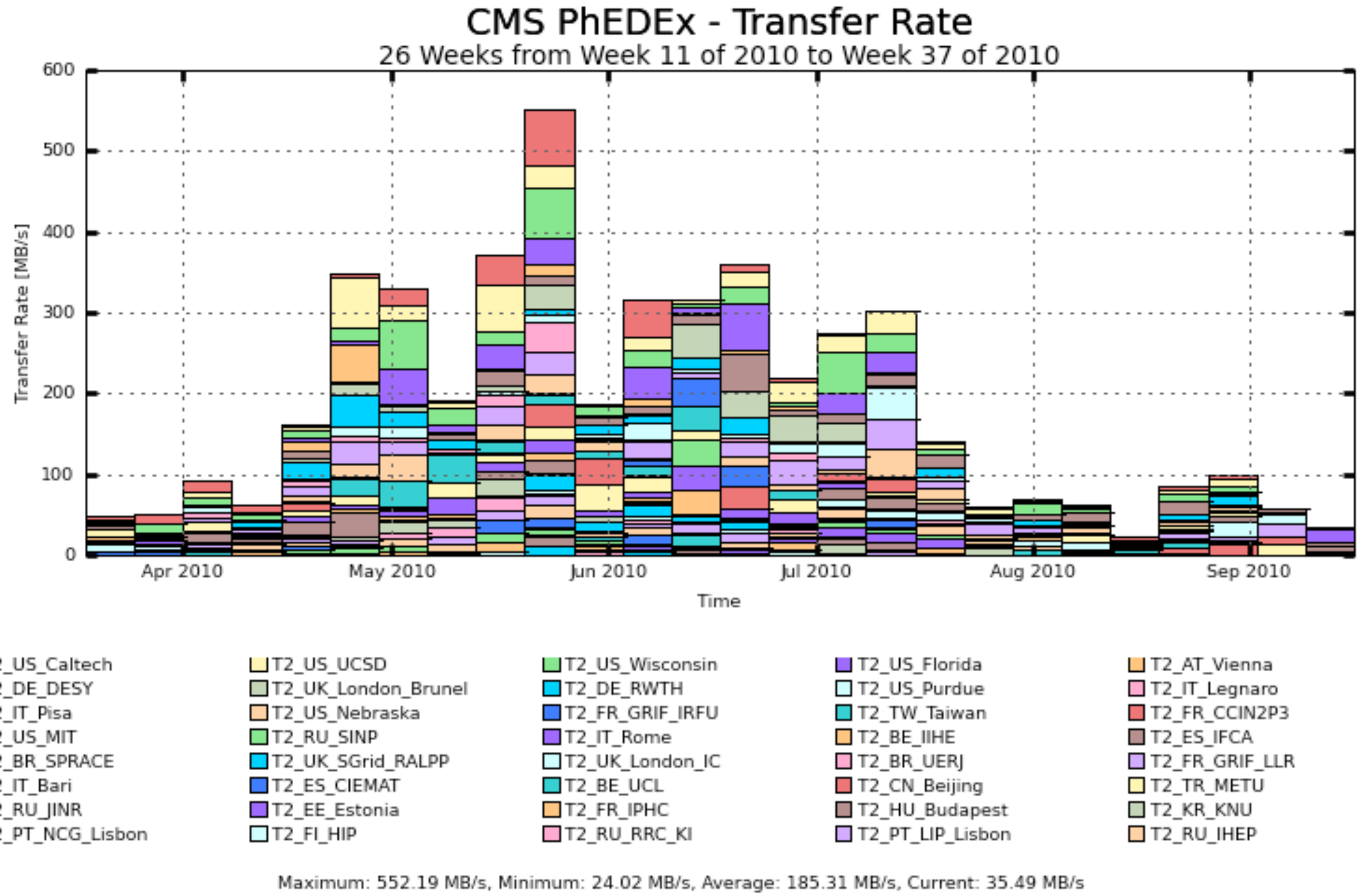
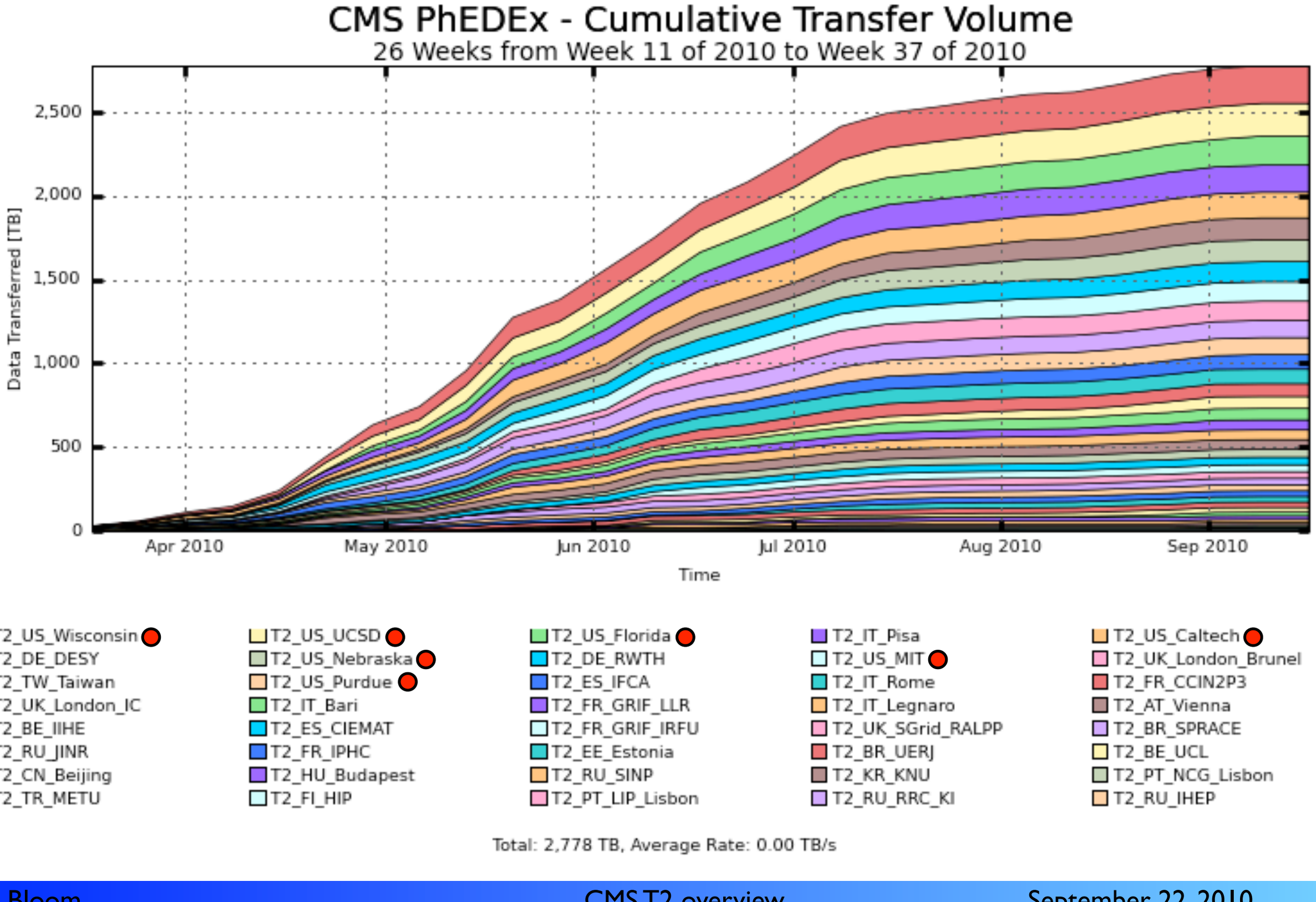
Transfers from T2 to T2 by source



Transfers from T2 to T2 by source



Transfers from T2 to T2 by destination



Transfers from T2 to T2 by destination