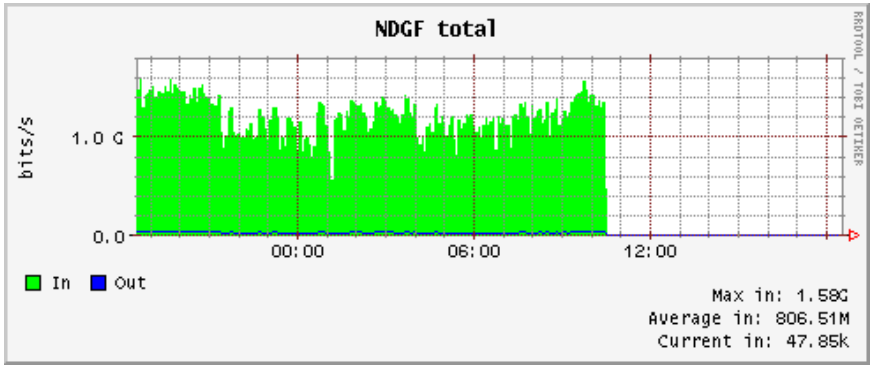
NDGF SC3 performance

On behalf of the Nordic Data Grid Facility Service Challenge 3 Team, in particular Lars Malinowsky (Stockhlom), Leif Nixon (Linkoping), Anders Waanaanen(Copenhagen) and Oxana Smirnova (Lund)

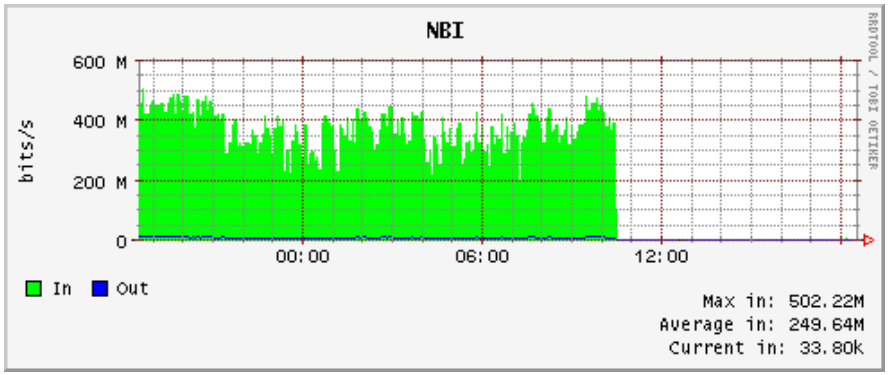
Tord Ekelof, Uppsala University

NDGF throughput 18 July 2005

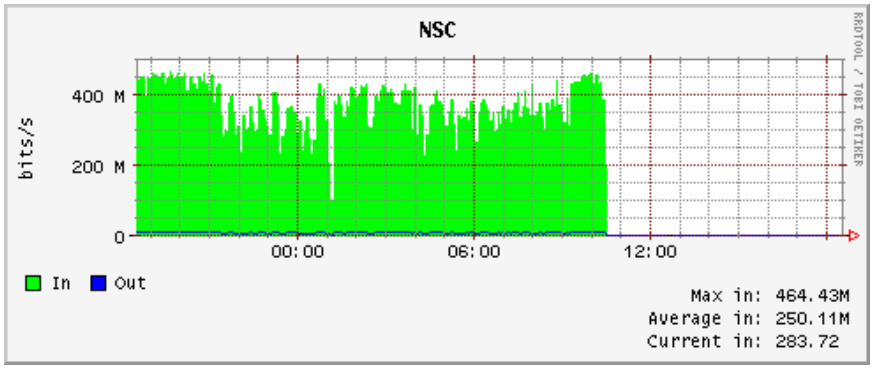
Total



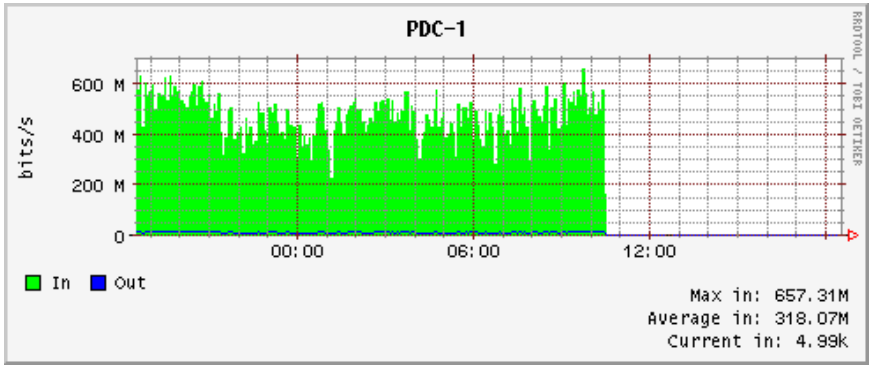
Copenhagen



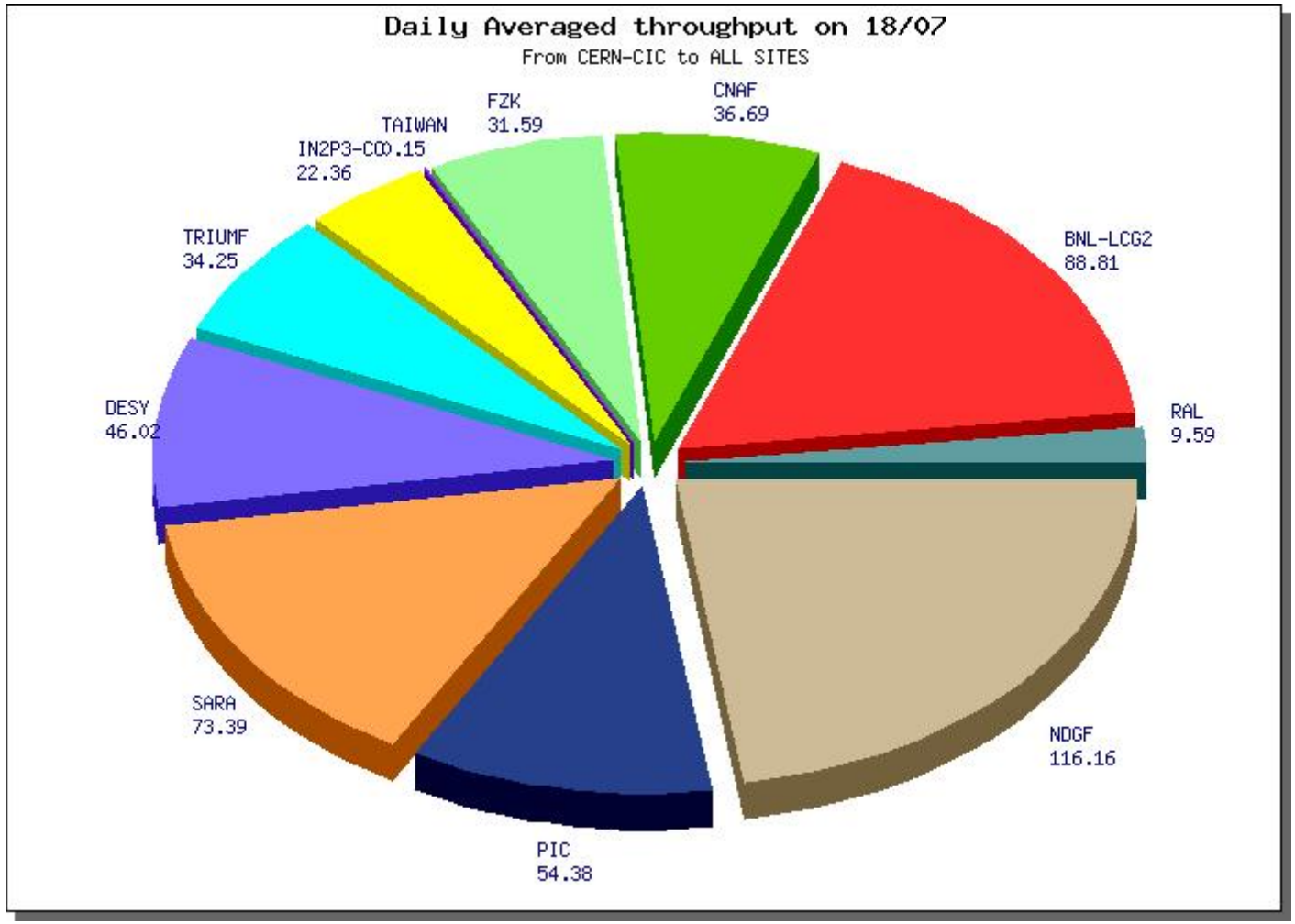
Linkoping



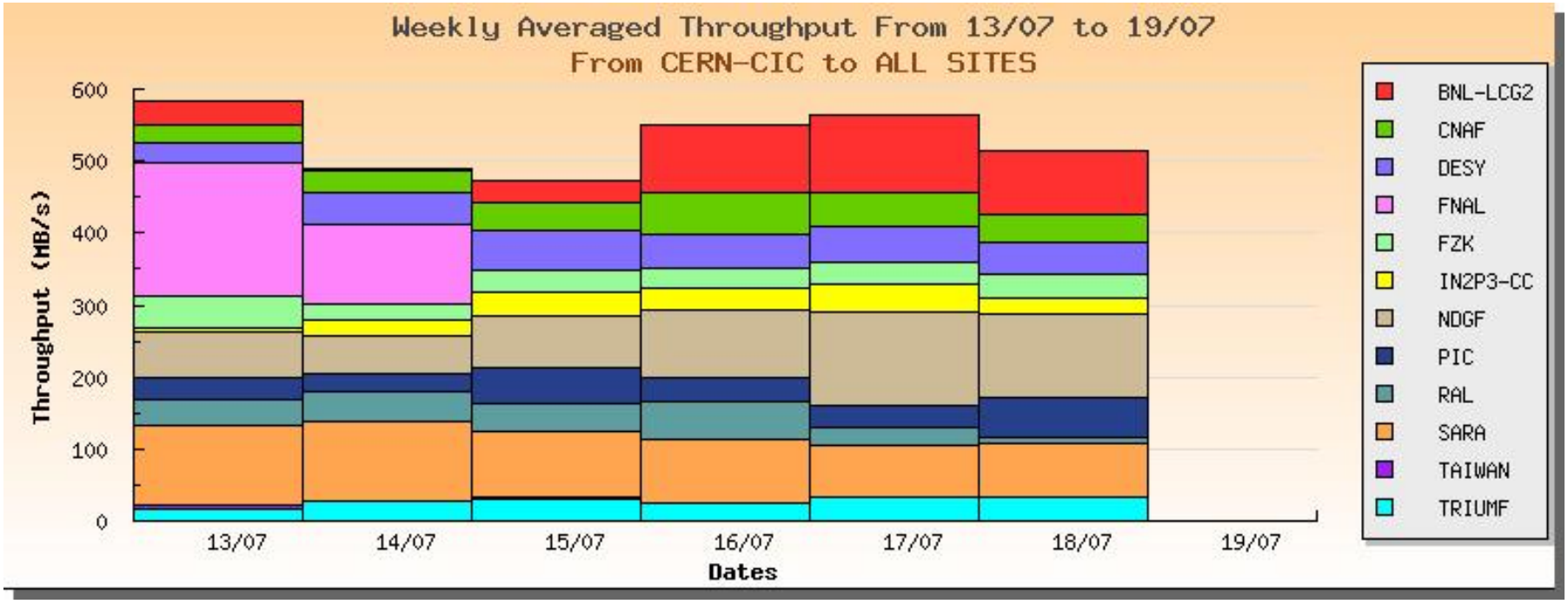
Stockholm



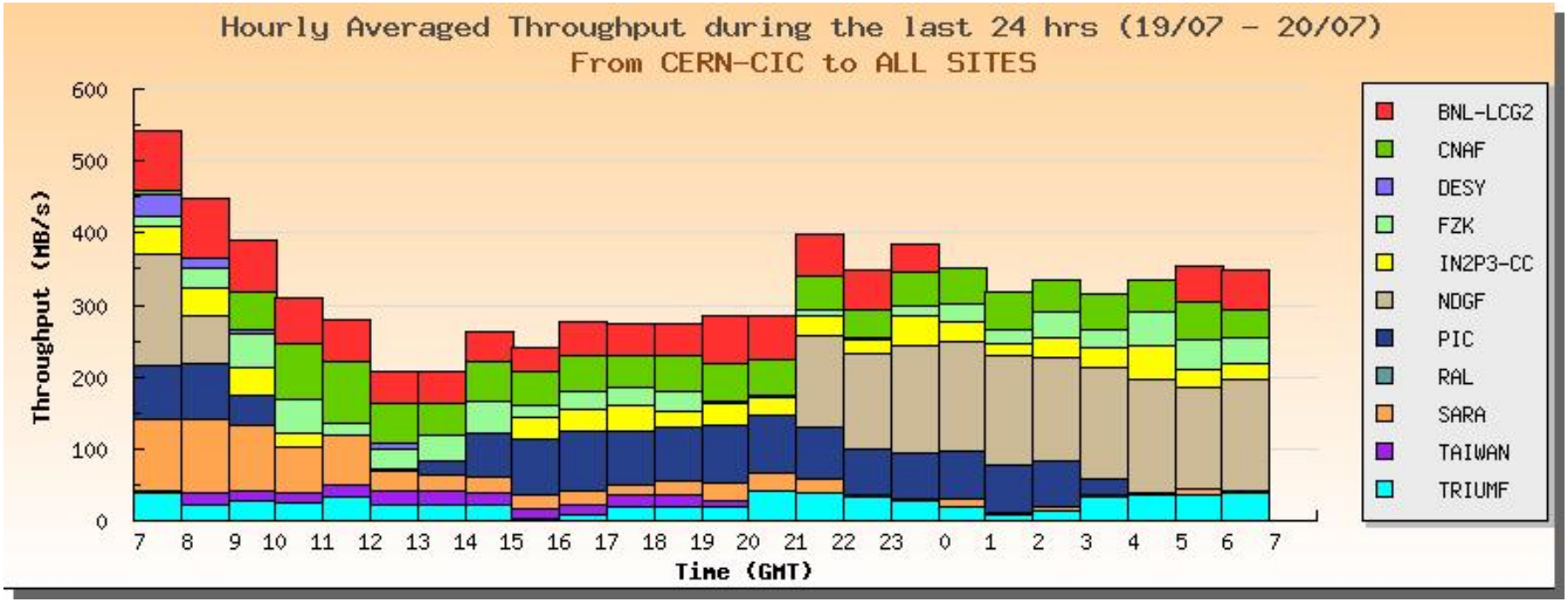
SC3 average throughput 18 July 2005



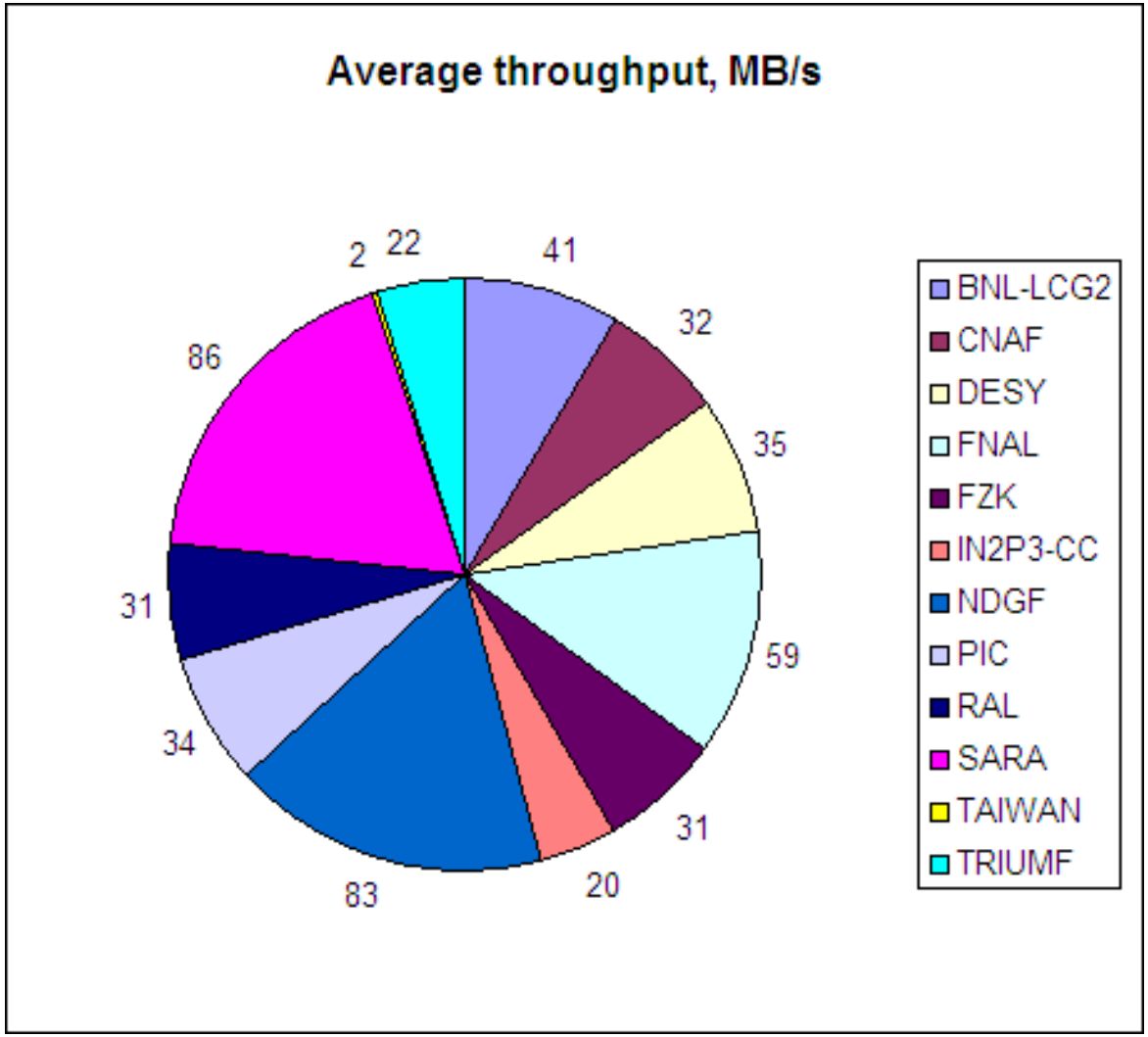
Last weeks SC3 throughput



The effect of the cut fiber at NBI yesterday…



Average during first SC3 week 11-18 July 2005



NDGF Tier 1 set-up

The NDGF organization for Service Challenge 3 was created on very short notice (preparations started around 10 June 2005).

The overall set-up is somewhat ad hoc –
we decided to use whatever tools and solutions that seemed to give the best chances of fulfilling the short-term goals for the throughput tests.

We did not really have much choice in this, as the longer-term requirements and goals were, and still are, somewhat hazy.

We wound up with NBI (Copenhagen), NSC (Linkoping) and PDC (Stockholm) pooling resources; one person at each site and in total four storage servers with approx. 14 TB disk. (More hardware might be added as we go.)

Each site has a 1 Gb/s connection to the general NORDUnet 10 Gb/s research network, which in turn is connected through a 10 Gb/s link via GÉANT to CERN.

NORDUnet can easily accommodate our bandwidth usage, even if we manage to run at the maximum 3 Gb/s that our local links allow us.

Software-wise we chose DPM as our Storage Resource Manager (SRM) implementation, because it is a relatively lightweight and simple piece of 'software, and because its components are relatively loosely coupled, which suits our distributed set-up.

The main disadvantage is that DPM by design cannot manage tape storage. (It is, after all, called the Disk Pool Manager).

Each site runs one or several gridftp servers.

The DPM server, which handles the SRM interface, manages the pool of gridftp servers and keeps track of which files are stored where, runs at NBI.

When the File Transfer Service (FTS) at CERN wants to transfer a file, it sends an SRM PUT request to the DPM server and receives a reply telling it which gridftp server to contact.

The actual data transfer is then made directly to the gridftp server.

Experiences

DPM has turned out to work reasonably well, considering it is newly written software.

Under normal operations it has worked reliably. It has sometimes failed to cope with anomalous situations, like full filesystems or downed gridftp servers, resulting in failed transfers and strange error messages.

The simplest way for us to recover from such situations has been to wipe the storage areas, clear the DPM database and start over from a clean slate.

We had been hoping that the DPM components were loosely coupled enough to run in a distributed fashion.

While it can be made to work, it is unfortunately not currently suitable for production use in a distributed environment.

For example, grid identities must be mapped to the same pool users on all servers, and those pool users must have the same numerical UID on all servers.

This makes DPM unfeasible for usage over organizational borders.

Furthermore, the control channel between the DPM server and the gridftp servers uses insecure, unencrypted communication.

We have continuously monitored the system performance and adjusted the system tuning parameters, trying to maintain maximum performance.

This proved an almost hopeless task, since so many different factors, most of them outside our control, influenced the performance.

Performance

Since start of the throughput tests on July 11, NDGF has run continuous file transfers around the clock, with only short interruptions.

Some interruptions have been planned, to allow further adjustments to the set-up, others have been unplanned, due to outages at CERN, and hardware failures.

The two longest interruptions were due to failure of the cooling system at NSC, lasting a few hours, and a cut fiber at NBI, lasting 12 hours (yesterday).

Currently, peak network utilization has been 1.67 Gb/s, and peak data throughput (measured as successful file transfers during one hour) has been 160 MB/s (i.e. 1.28 Gb/s).

We have been running at >100 MB/s data throughput for extended periods. (The discrepancy between network bandwidth utilization and data throughput is due to various protocol overheads and to failed transfers.)

Concluding remark

Considering that we have never defined an explicit target throughput for NDGF, having always used hedged phrases like "participation on a reduced scale" and "best effort" and saying "we might reach 150 MB/s, we might get stuck at 50 MB/s", the achieved performance may be seen as quite satisfactory.