Enabling Grids for E-science in Europe

www.eu-egee.org

NA4 Open Meeting, Catania, 14-16.07.2004

ALICE network stress tests

Roberto Barbera
NA4 Generic Applications Coordinator

Work in collaboration with: P. Cerello, D. Di Bari, G. Donvito (CMS), E. Fragiacomo, A. Fritz, M. Luvisetto, M. Masera, F. Minafra, D. Mura, S. Piano, M. Sitta, J. Švec, R. Turrisi

Contributions from GARR and INFN NetGroup: C. Allocchio, M. Campanella, L. Gaido, S. Lusso, M. Michelotto, S. Spanu, S. Zani, D. De Girolamo

EGEE is a project funded by the European Union under contract IST-2003-508833

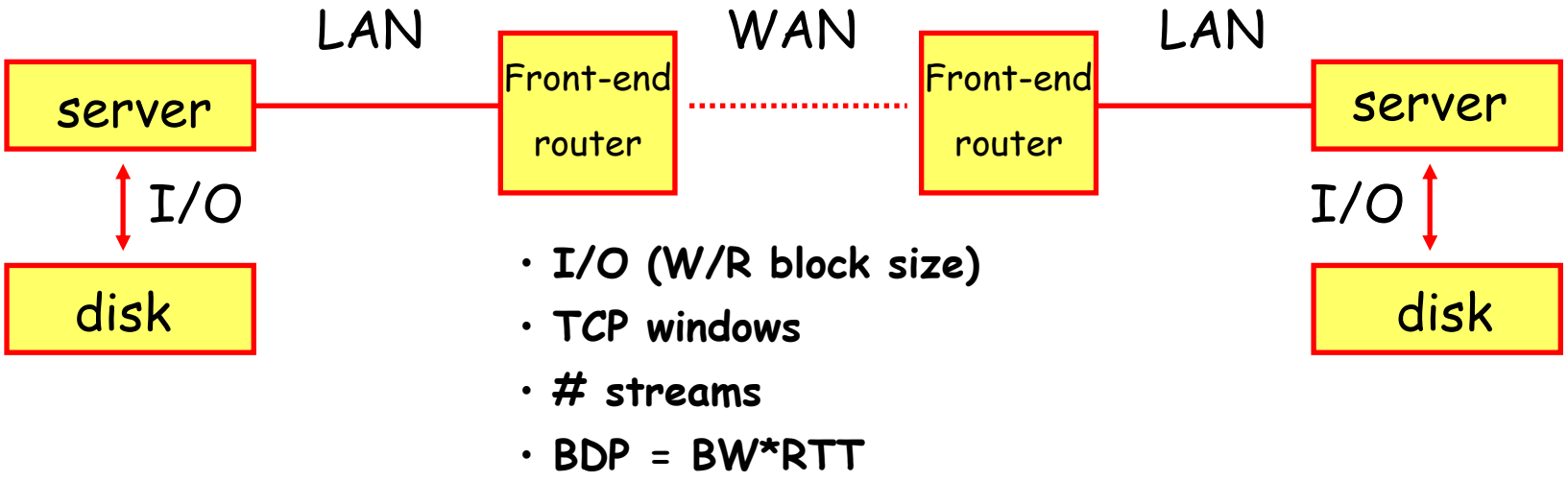
Contents

Enabling Grids for E-science in Europe

- Network stress test 2003
  - Objectives
  - Set-up
  - Results
  - Conclusions
- Physics Data Challenge 2004
  - Phase 1
  - Phase 2
- General conclusions

Objectives

- See if the actual bandwidths can cope with the ALICE needs
- Spot possible bottle-necks out in the point-to-point transfers (I/O↔LAN↔ WAN↔LAN↔I/O)
- Check, with “real” numbers of “real” use cases, if bandwidth attributions foreseen in the next future are adequate



Preparation and benchmark

Enabling Grids for E-science in Europe

- Standard configuration of both the TCP stack and disk I/O parameters in Linux
- SSH keys exchanged among all machines to “secure” file transfers without typing passwords
- Automatic procedure installed on all machines:
  - waits a random time uniformly choosen between 0 and customizable maximum (1 min and 5 mins tried so far)
  - chooses at random on of the other N-1 servers (with a weight proportional to the maximum bandwith of the site that server belongs to)
  - chooses at random one of three files with different sizes (1.6 GB, 0.8 GB, and 0.3 GB)
  - sends back and forth the file using bbFTP with a customizable number of parallel streams (16 and 8 tried so far)
  - checks if any bits gets lost and fills a detailed log file

Test-bed and figures

Enabling Grids for E-science in Europe

Prague

Houston

BA: 3 servers (2 ALICE, 1 CMS)

BO: 6 servers

CA: 2 servers

CNAF: 2 servers

CT: 2 servers

PD: 6 servers

TO: 2 servers

TS: 1 server

Prague: 1 server

Houston: 1 server

Disk access measurements (non reserved access, local disk)

Enabling Grids for E-science in Europe

Bonnie++1.10

Machine	Write (MBytes/s)	Read (MBytes/s)
boalice8.bo.infn.it	5	3
server3.ca.infn.it	43	32
aliserv10.ct.infn.it	57	25
pcalice19.pd.infn.it	5	5
alifarm02.to.infn.it	31	53
alifarm.ts.infn.it	27	34

IOzone-3.164

Machine	Write (MBytes/s)	Read (MBytes/s)
boalice8.bo.infn.it	5	5
server3.ca.infn.it	45	61
aliserv10.ct.infn.it	27	34
alifarm02.to.infn.it	40	59
alifarm.ts.infn.it	28	36

GARR network status at the start-up

Enabling Grids for E-science in Europe

- Bari: 28 Mb/s (BGA: 16 Mb/s)
- Bologna: 32 Mb/s
- Cagliari: 8 Mb/s
- Catania: 34 Mb/s
- CNAF: 1024 Mb/s
- Padova: 155 Mb/s
- Torino: 155 Mb/s (BGA: 70 Mb/s)
- Trieste: 16 Mb/s

Selected results (Bologna)

Enabling Grids for E-science in Europe

INFN Bologna : ATM6/2.103 - Netscape

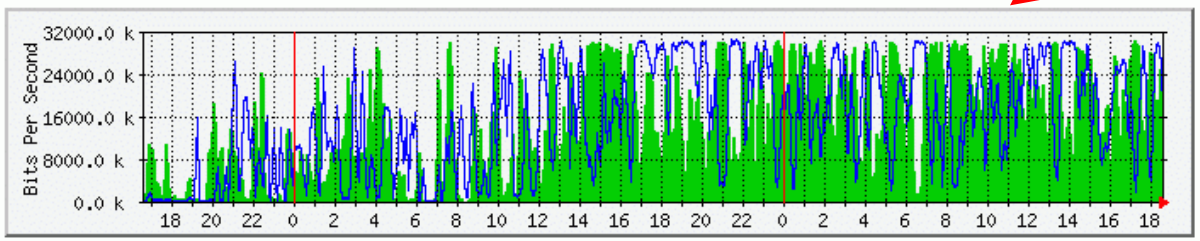
File Edit View Go Communicator Help

Bookmarks Location: http://www.noc.garr.it/mrtg/RTG-BOLOGNA.garr.net/rtg-infnbo.bo.garr.net.html What's Related

The statistics were last updated **Tuesday, 25 February 2003 at 18:38**,
at which time '**12410_Bologna.garr.net**' had been up for **97 days, 7:22:56**.

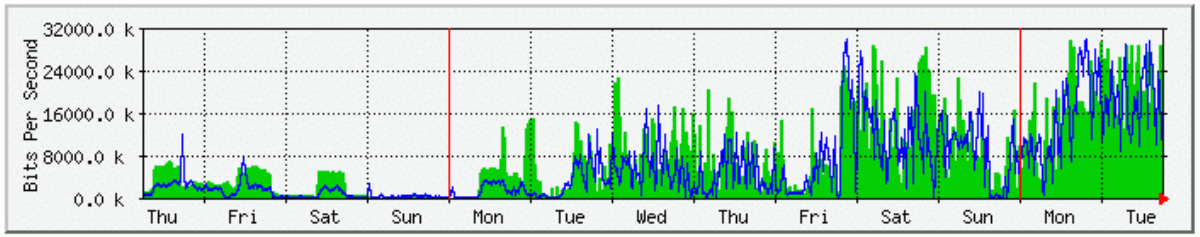
saturated !

`Daily' Graph (5 Minute Average)



Max In/s: 30.5 Mbits/s (89.8%) Average In/s: 15.6 Mbits/s (45.8%) Current In/s: 25.7 Mbits/s (75.5%)
Max Out/s: 30.4 Mbits/s (89.5%) Average Out/s: 14.9 Mbits/s (43.9%) Current Out/s: 16.4 Mbits/s (48.2%)

`Weekly' Graph (30 Minute Average)



Max In/s: 29.9 Mbits/s (88.0%) Average In/s: 7507.5 kbits/s (22.1%) Current In/s: 18.1 Mbits/s (53.4%)
Max Out/s: 30.0 Mbits/s (88.1%) Average Out/s: 6138.4 kbits/s (18.1%) Current Out/s: 16.5 Mbits/s (48.5%)

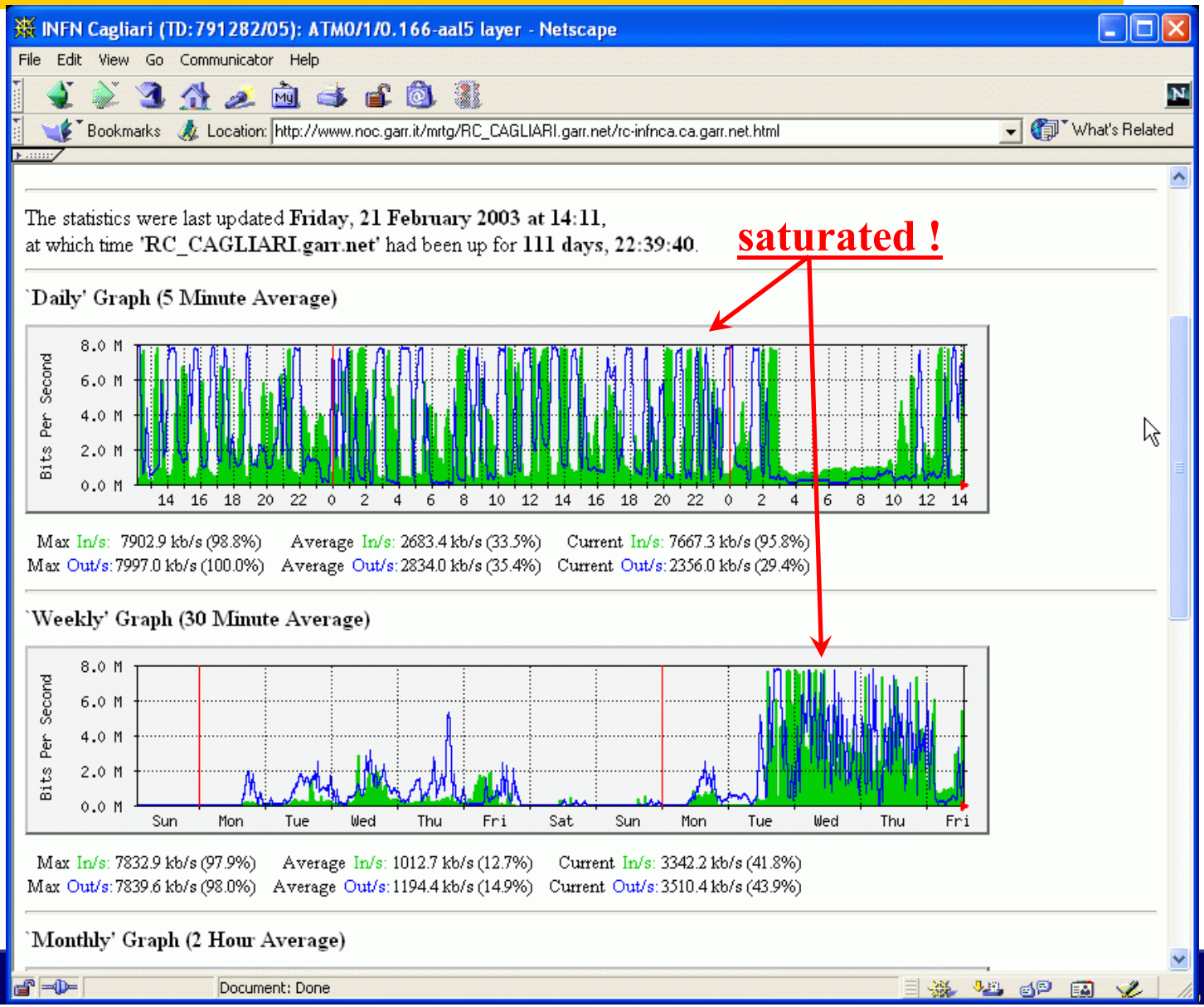
`Monthly' Graph (2 Hour Average)

Document: Done

Official GARR NOC statistics

Selected results (Cagliari)

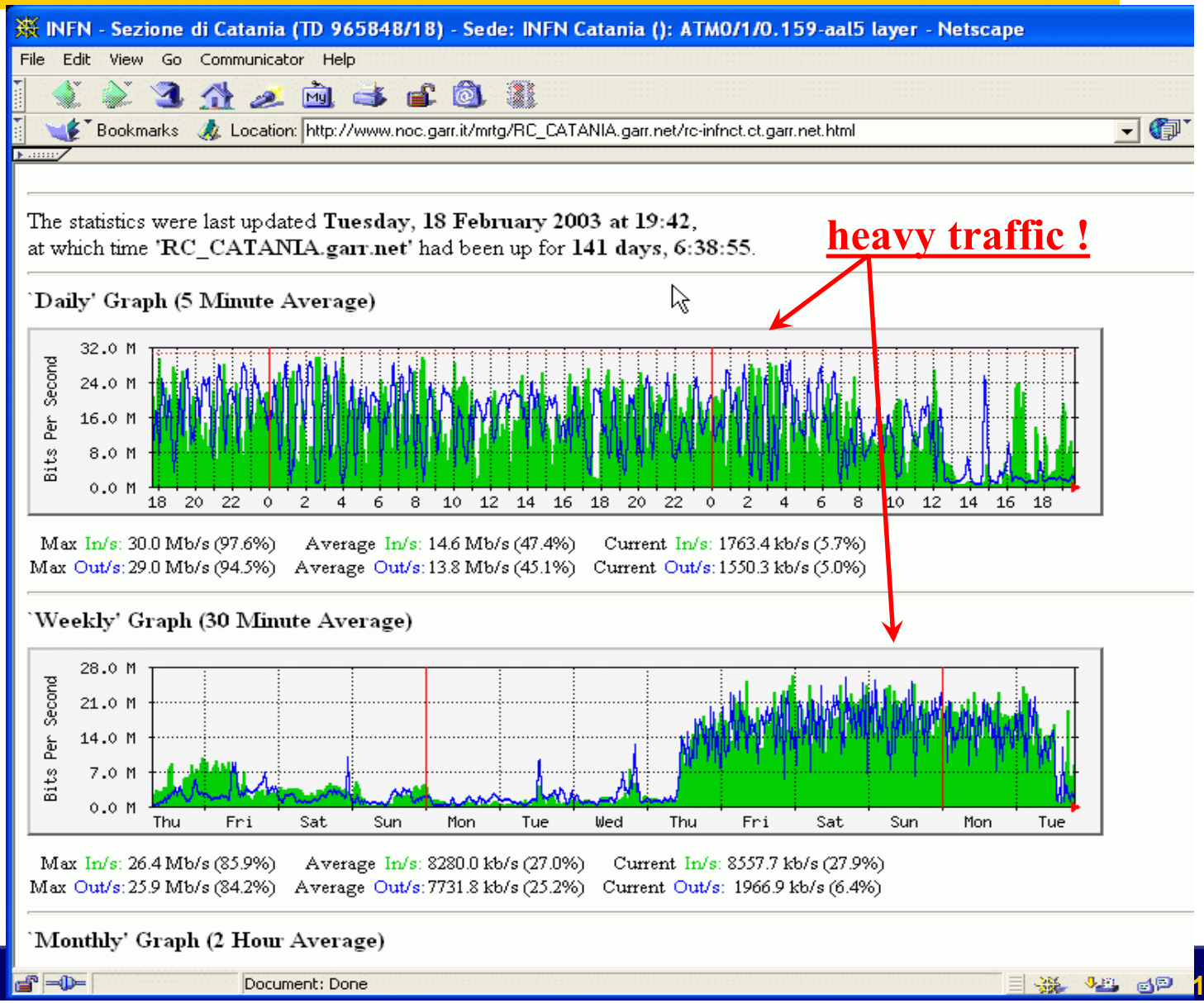
Enabling Grids for E-science in Europe



Official GARR NOC statistics

Selected results (Catania)

Enabling Grids for E-science in Europe



Official GARR NOC statistics

Network bandwidths now

Enabling Grids for E-science in Europe

- Bari: 28 Mb/s (BGA: 16 Mb/s)
- Bologna: 100 Mb/s (BGA: 32 Mb/s)
- Cagliari: 32 Mb/s
- Catania: 34 Mb/s (see later)
- CNAF: 1024 Mb/s
- Padova: 155 Mb/s
- Torino: 155 Mb/s (BGA: 70 Mb/s)
- Trieste: 24 Mb/s

Bandwidth measurements

Enabling Grids for E-science in Europe

Iperf-1.6.3

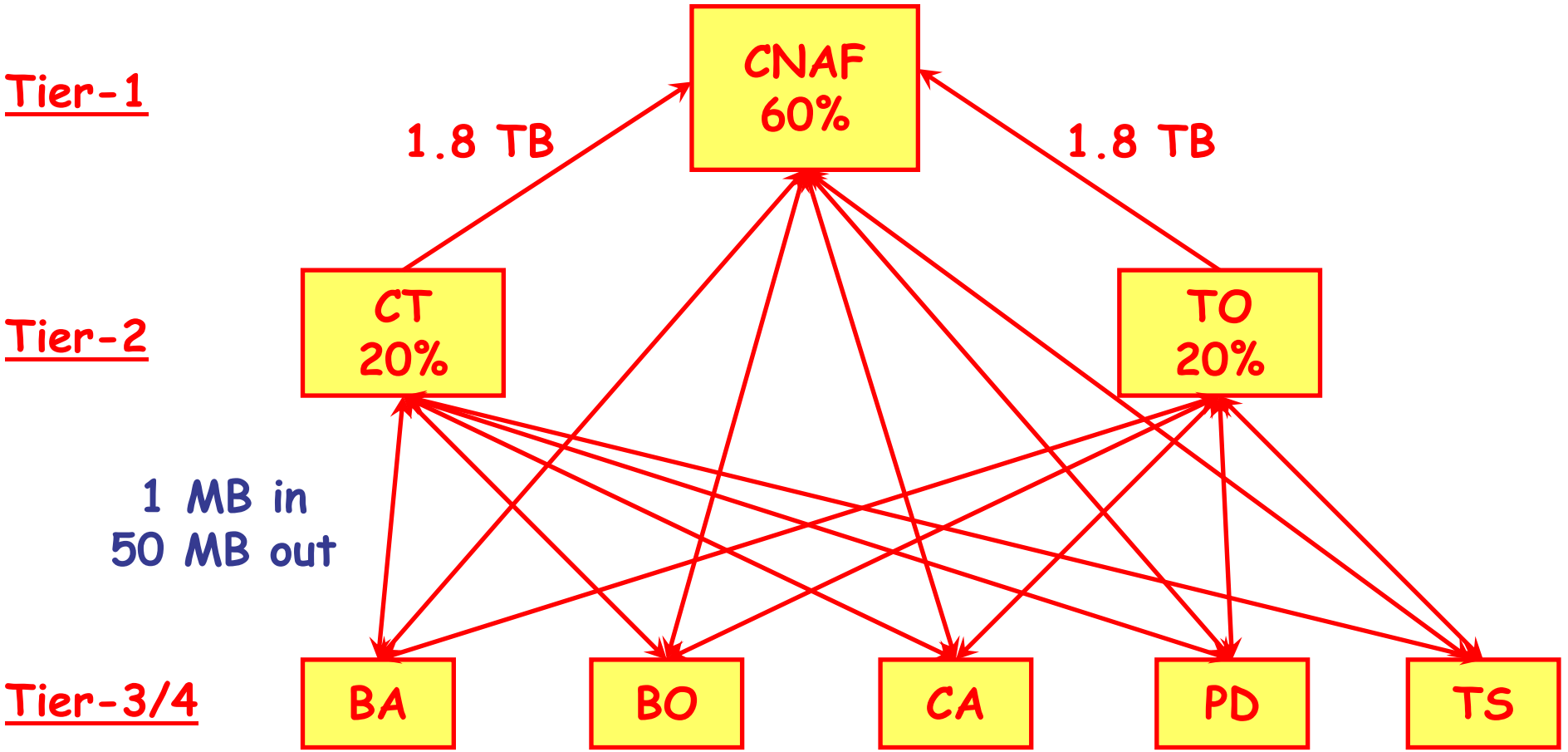
Machine	BW1(Mb/s)	BW2(Mb/s)	BW4 (Mb/s)	BW8(Mb/s)	BW16(Mb/s)	BW32(Mb/s)
boalice8.bo.infn.it	76	77	79	84	86	87
server3.ca.infn.it	12	21	22	21	21	22
aliserv10.ct.infn.it	9	15	18	18	19	20
pcalice19.pd.infn.it	26	51	87	92	93	94
alifarm02.to.infn.it	27	50	57	61	64	69
alifarm.ts.infn.it	14	18	18	18	19	19

Netperf-2.1

Machine	BW1(Mb/s)	BW2(Mb/s)	BW4 (Mb/s)	BW8(Mb/s)	BW16(Mb/s)	BW32(Mb/s)
boalice8.bo.infn.it	30	44	65	80	81	86
server3.ca.infn.it	13	18	22	22	22	23
aliserv10.ct.infn.it	9	16	19	20	22	22
pcalice19.pd.infn.it	26	51	87	92	93	97
alifarm02.to.infn.it	28	41	46	55	61	65
alifarm.ts.infn.it	14	17	18	18	17	19

Multi-tier use case
(HBT prod., 5000 evts., 9 TB)

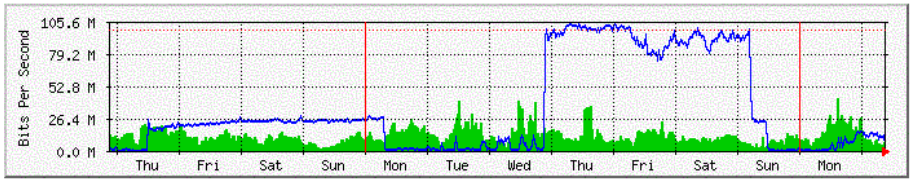
Enabling Grids for E-science in Europe



Results (official GARR-NOC stats.)

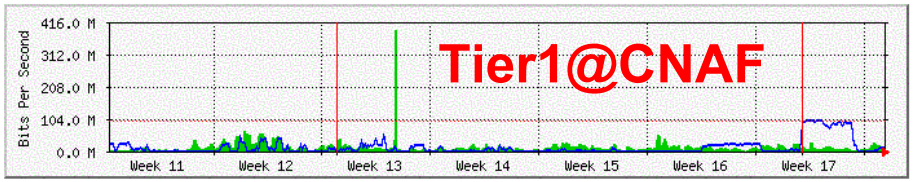
Enabling Grids for E-science in Europe

`Weekly' Graph (30 Minute Average)



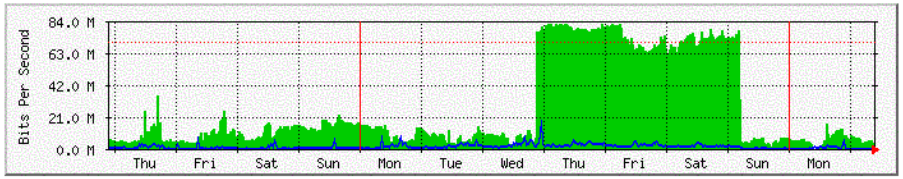
Max In/s: 44.2 Mbits/s (45.0%) Average In/s:13.1 Mbits/s (13.3%) Current In/s:7477.6 kbits/s (7.6%)
Max Out/s:104.3 Mbits/s (106.1%) Average Out/s:34.5 Mbits/s (35.1%) Current Out/s:5699.5 kbits/s (5.8%)

`Monthly' Graph (2 Hour Average)



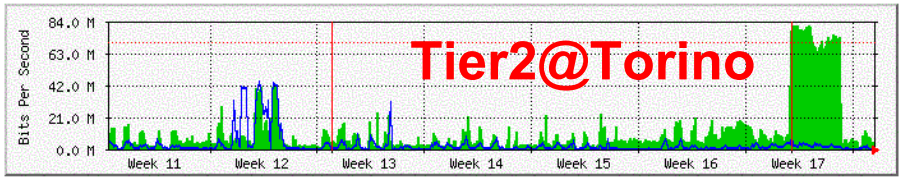
Max In/s:392.6 Mbits/s (399.3%) Average In/s:15.7 Mbits/s (15.9%) Current In/s:5790.6 kbits/s (5.9%)
Max Out/s:103.5 Mbits/s (105.3%) Average Out/s:14.7 Mbits/s (15.0%) Current Out/s:12.0 Mbits/s (12.2%)

`Weekly' Graph (30 Minute Average)



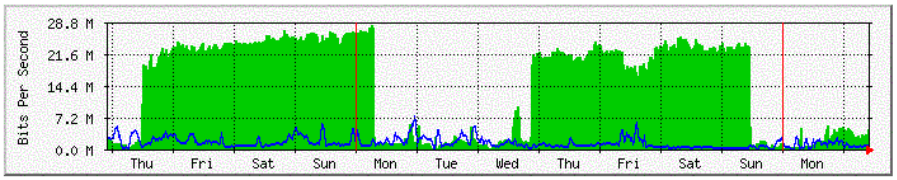
Max In/s:83.0 Mb/s (118.5%) Average In/s:26.9 Mb/s (38.4%) Current In/s:4875.3 kb/s (7.0%)
Max Out/s: 18.8 Mb/s (26.9%) Average Out/s:1967.6 kb/s (2.8%) Current Out/s:1089.3 kb/s (1.6%)

`Monthly' Graph (2 Hour Average)



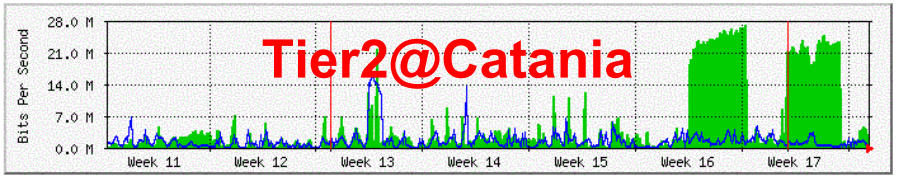
Max In/s:82.2 Mb/s (117.5%) Average In/s:12.0 Mb/s (17.2%) Current In/s:5841.3 kb/s (8.3%)
Max Out/s: 44.5 Mb/s (63.6%) Average Out/s:3398.8 kb/s (4.9%) Current Out/s: 549.5 kb/s (0.8%)

`Weekly' Graph (30 Minute Average)



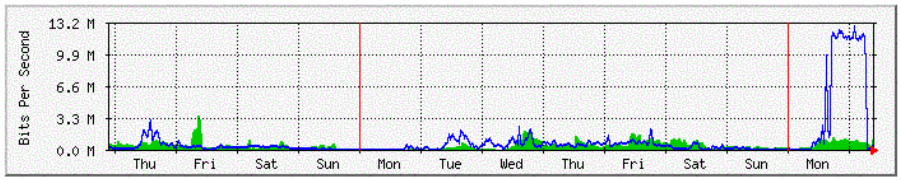
Max In/s: 28.4 Mb/s (92.6%) Average In/s:14.6 Mb/s (47.6%) Current In/s:4786.6 kb/s (15.6%)
Max Out/s:7536.8 kb/s (24.5%) Average Out/s:1635.8 kb/s (5.3%) Current Out/s: 1237.5 kb/s (4.0%)

`Monthly' Graph (2 Hour Average)



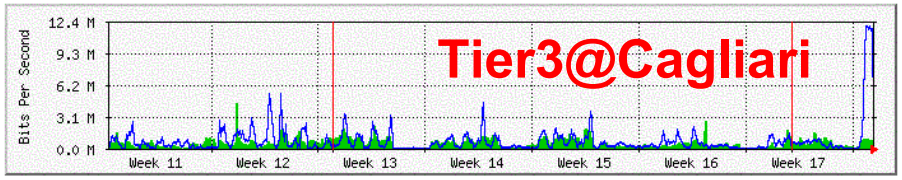
Max In/s:27.2 Mb/s (88.6%) Average In/s:5231.2 kb/s (17.0%) Current In/s:3147.0 kb/s (10.2%)
Max Out/s:16.0 Mb/s (52.2%) Average Out/s: 1666.2 kb/s (5.4%) Current Out/s: 884.2 kb/s (2.9%)

`Weekly' Graph (30 Minute Average)



Max In/s:3590.4 kb/s (22.4%) Average In/s: 524.9 kb/s (3.3%) Current In/s:1116.1 kb/s (7.0%)
Max Out/s: 12.8 Mb/s (80.1%) Average Out/s:1004.9 kb/s (6.3%) Current Out/s: 655.2 kb/s (4.1%)

`Monthly' Graph (2 Hour Average)



Max In/s:4574.4 kb/s (28.6%) Average In/s:604.5 kb/s (3.8%) Current In/s: 714.7 kb/s (4.5%)
Max Out/s: 12.0 Mb/s (75.1%) Average Out/s:799.4 kb/s (5.0%) Current Out/s:2276.4 kb/s (14.2%)

TCP tuning

Enabling Grids for E-science in Europe

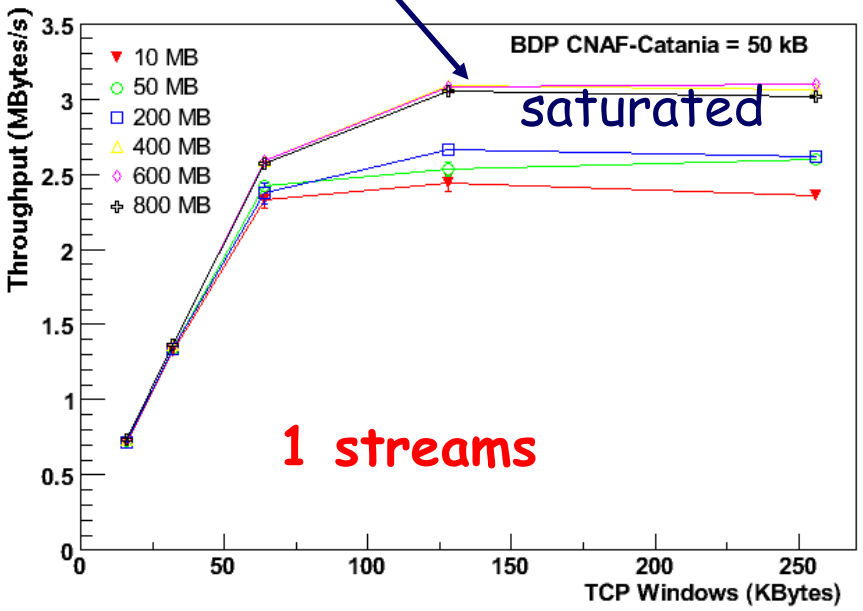
Site	RTT (msec) from CNAF	BW (Mb/s) from CNAF	BDP (MB)
Houston	140	70	1.2
Prague	20	250	0.6
Catania	25	25	0.08

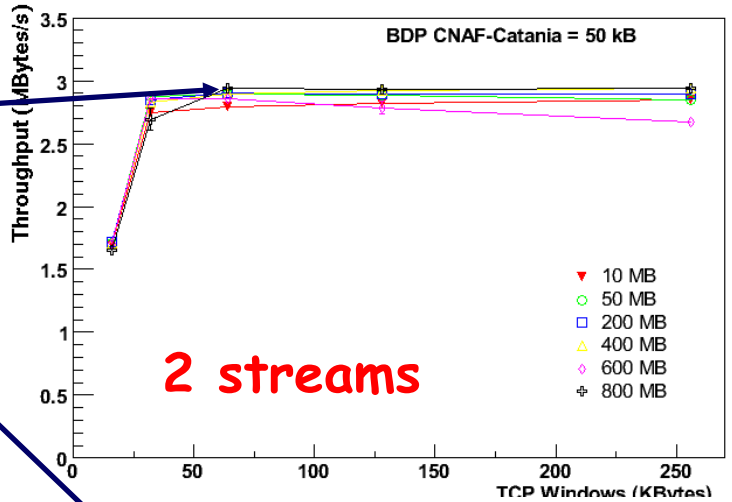
bbFTP vs. # of streams and TCP windows (1/4)

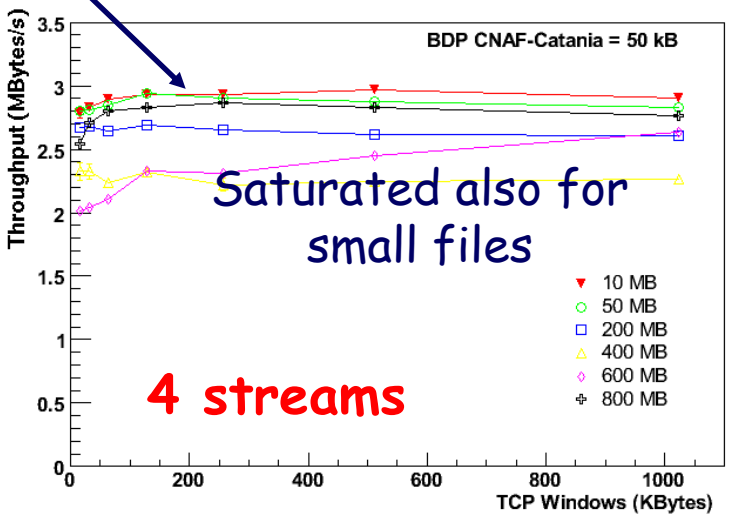
Enabling Grids for E-science in Europe

Catania-CNAF

Max bw measured (iperf) = 25 Mb/s



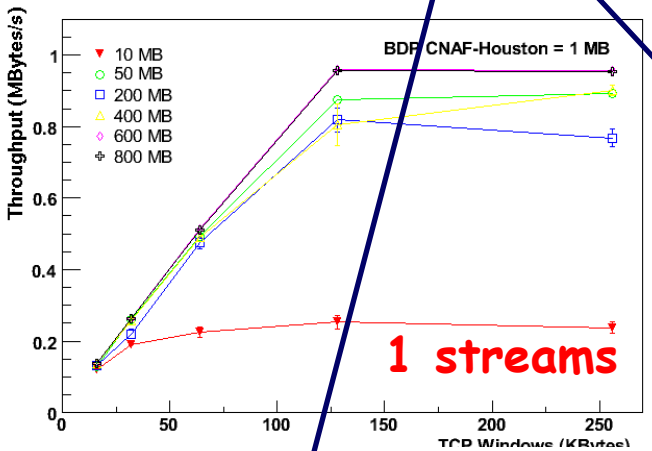


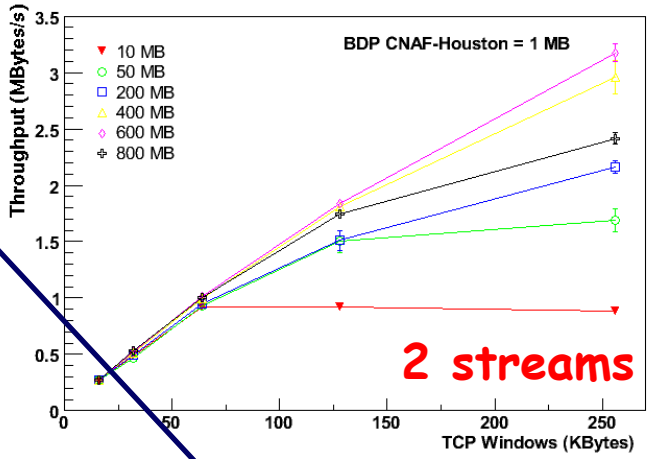


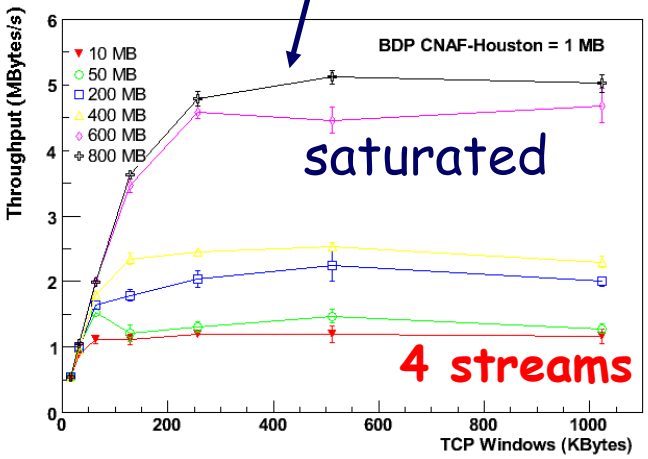
bbFTP vs. # of streams and TCP windows (2/4)

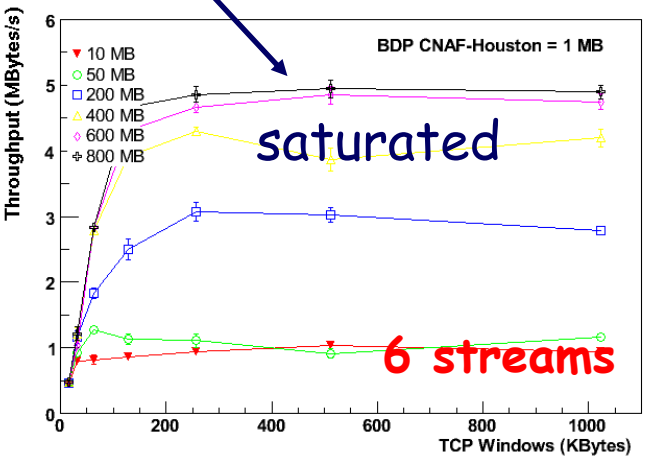
Enabling Grids for E-science in Europe

Houston-CNAF, Max bw measured (iperf) = 50 Mb/s





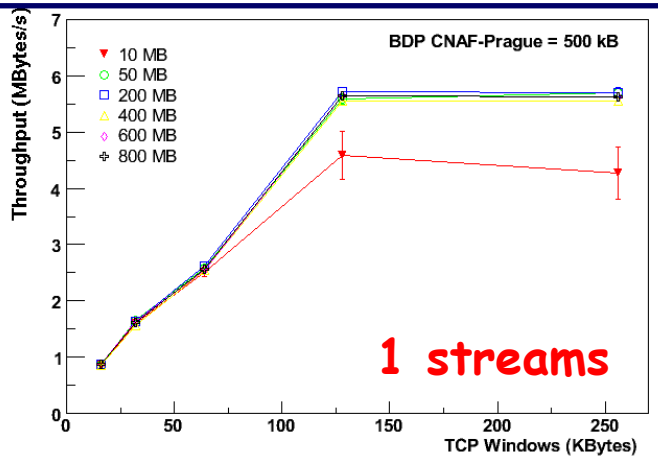


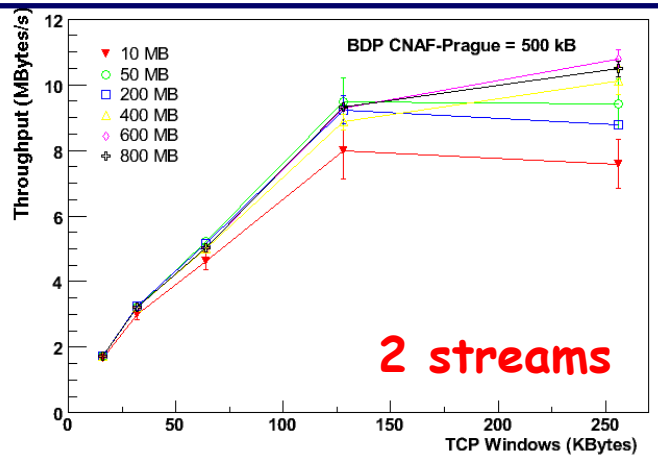


bbFTP vs. # of streams and TCP windows (3/4)

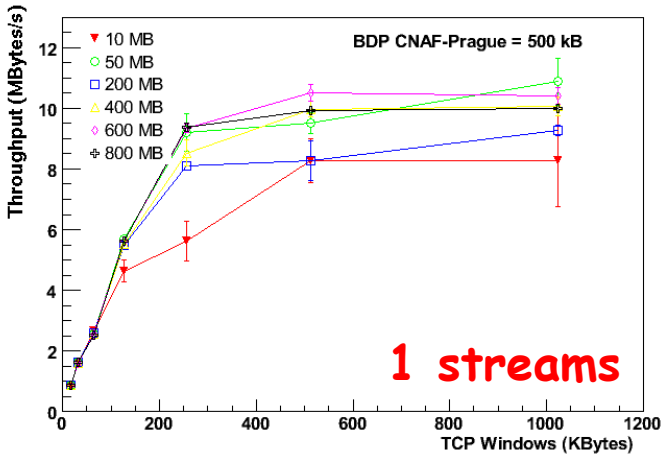
Enabling Grids for E-science in Europe

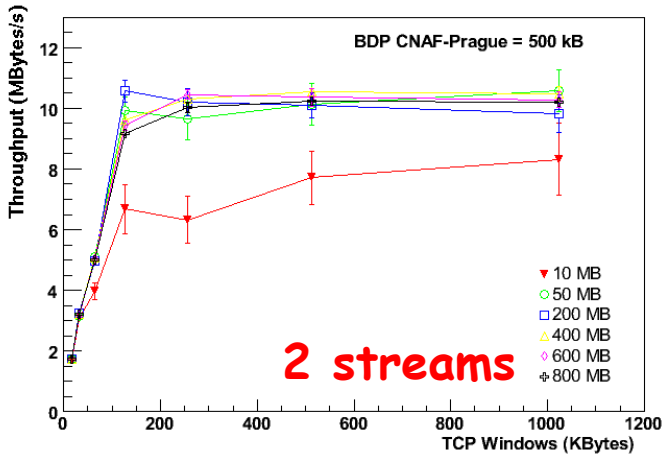
Prague-CNAF, Max bw measured (iperf) = 250 Mb/s





With an very high maximum buffer size (130KB->8 MB)

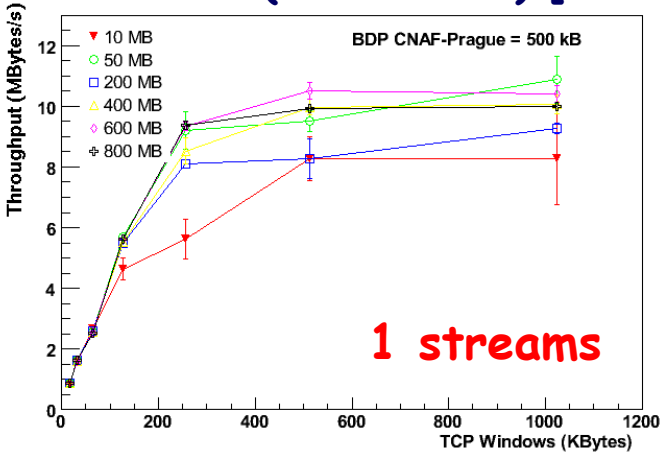


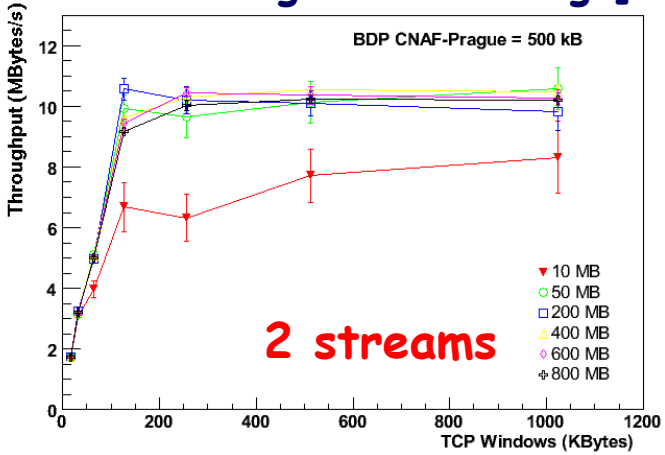


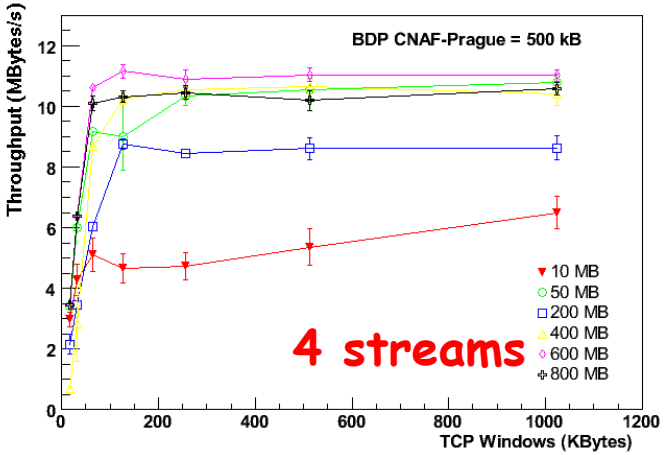
bbFTP vs. # of streams and TCP windows (4/4)

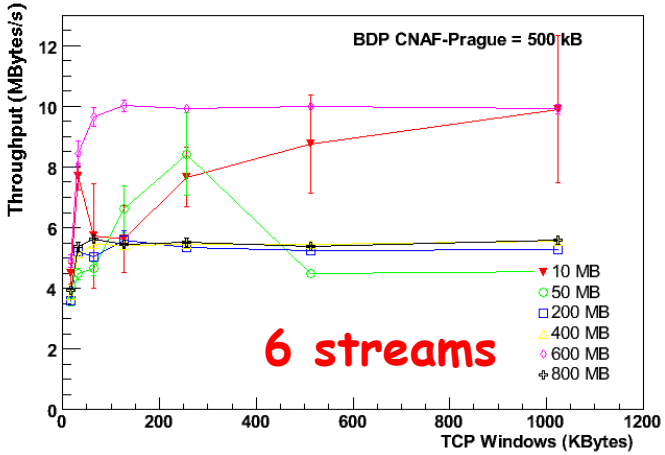
Enabling Grids for E-science in Europe

Prague-CNAF with 1, 2, 4 and 6 streams with a very large maximum buffer size (8MB » BDP) [Ref: http://www-didc.lbl.gov/TCP-tuning/]









Bottleneck at I/O level

Some conclusions (1/2)

Enabling Grids for E-science in Europe

- First “real” multi-site/multi-server stress-test of the Italian GARR network
- Actual bandwidths resulted strongly inadequate if we especially consider all ALICE sites “as a whole” and the present number of servers already available by now
- Useful information on the actual farm architecture (limits of NFS in case of many parallel threads and big files)
- Big “perturbation” and interest inside both INFN NetGroup and GARR with prompt and excellent feed-back and support
- Strong and “incredibly” fast bandwith upgrades in many sites made by the GARR NOC
- Mapping of the testbed on a multi-tier topology does not seem to pose major problems for Tier-3’s

Some conclusions (2/2)

Enabling Grids for E-science in Europe

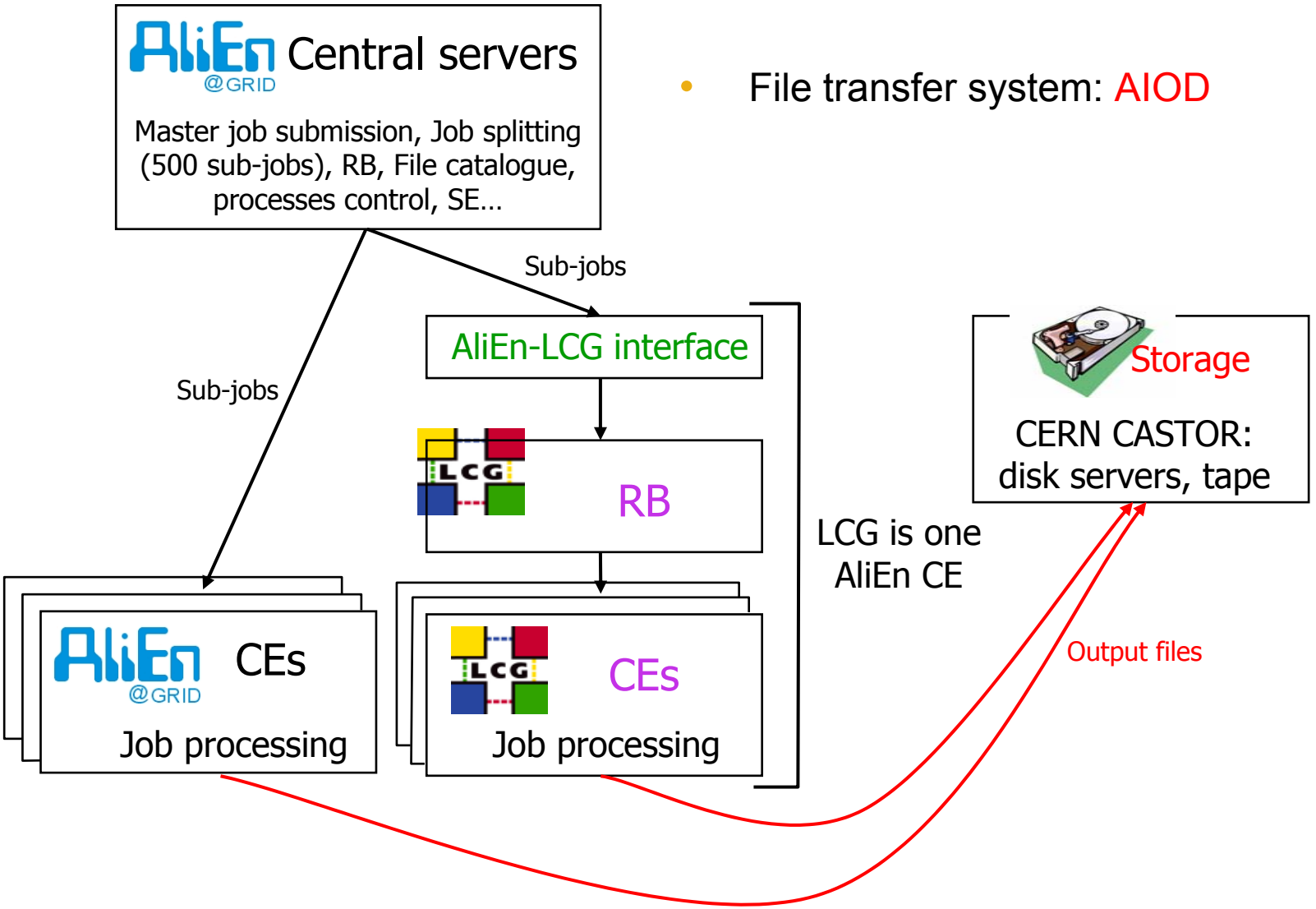
- Throughput measurements have been obtained for several file sizes as a function of TCP windows and with variable number of streams.
- Up to now, increasing TCP maximum buffers size does not seem to give evident advantages with respect to multistreaming, but tests are not finished
- The research of the best set of number of streams and TCP windows for a given file size and network path could help us to optimize our data transfers. In this way we are able to determine our effective capability to use network and then what could be our reasonable requests.
- It would be really useful the partecipation of USA sites with machines Gbit-connected in order to study TCP tunning in details. A good news is that OSC recently joined the testbed.

Purpose, structure and principles of ALICE PDC04

Enabling Grids for E-science in Europe

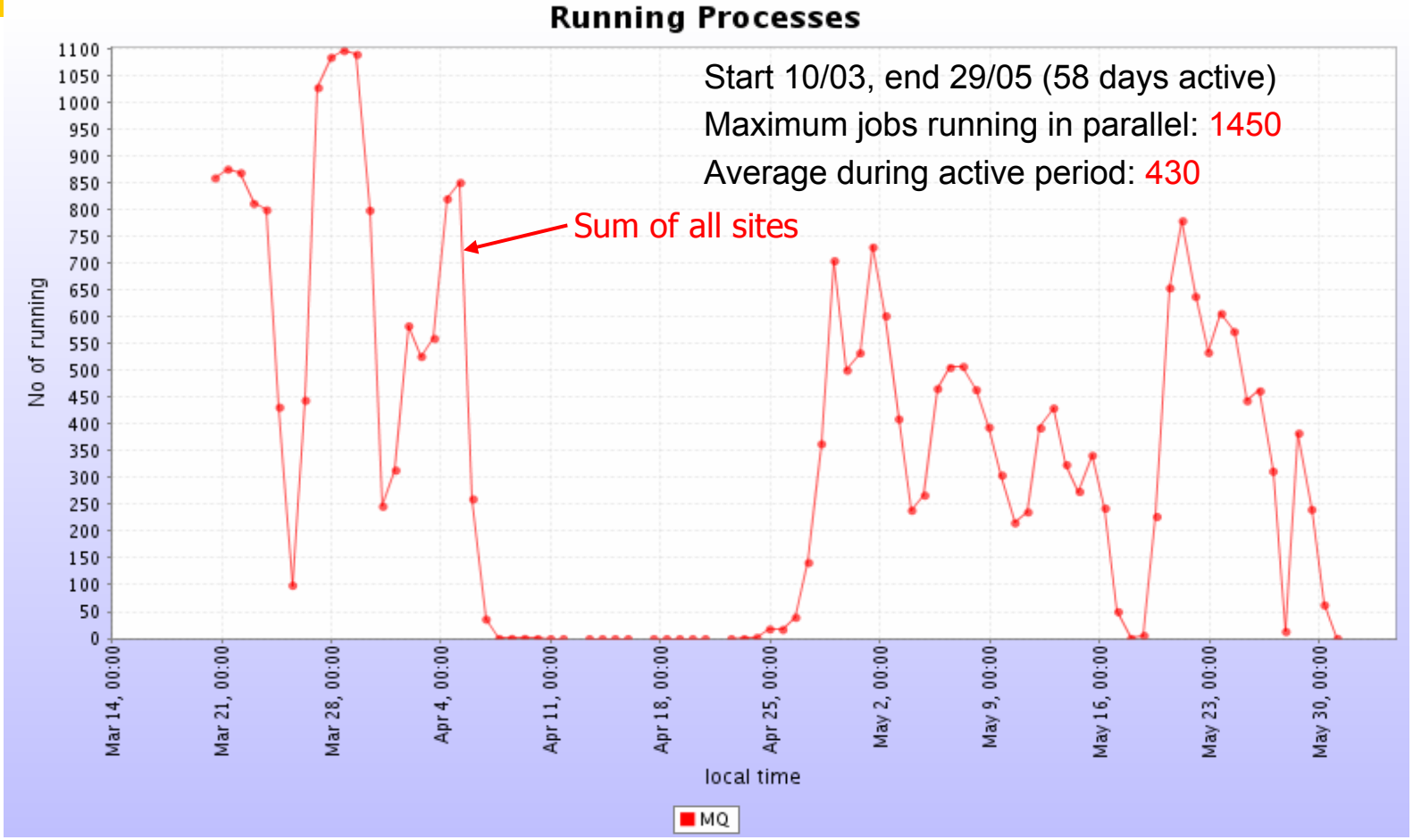
- Test and validate the ALICE Offline computing model:
  - Produce and analyse ~10% of the data sample collected in a standard data-taking year
  - Use the entire (complicated) system: AliEn, AliROOT, LCG, Proof…
  - Dual purpose: test of the software and ***physics analysis*** of the produced data for the Alice PPR
- Structure:
  - Logically divided in three phases:
    - Phase 1 - Production of underlying Pb+Pb events with different centralities (impact parameters) + production of p+p events
    - Phase 2 - Mixing of signal events with different physics content into the underlying Pb+Pb events (underlying events are reused several times)
    - Phase 3 – Distributed analysis (not discussed in this talk)
- Principles:
  - True GRID data production and analysis: all jobs are run on the GRID, using only ***AliEn*** for access and control of native computing resources and, through an interface, the LCG resources
  - In phase 3 ***AliEn+Proof*** ***(ARDA - A Realisation of Distributed Analysis for LHC)***

Structure of Phase 1 event production



- File transfer system: AIOD

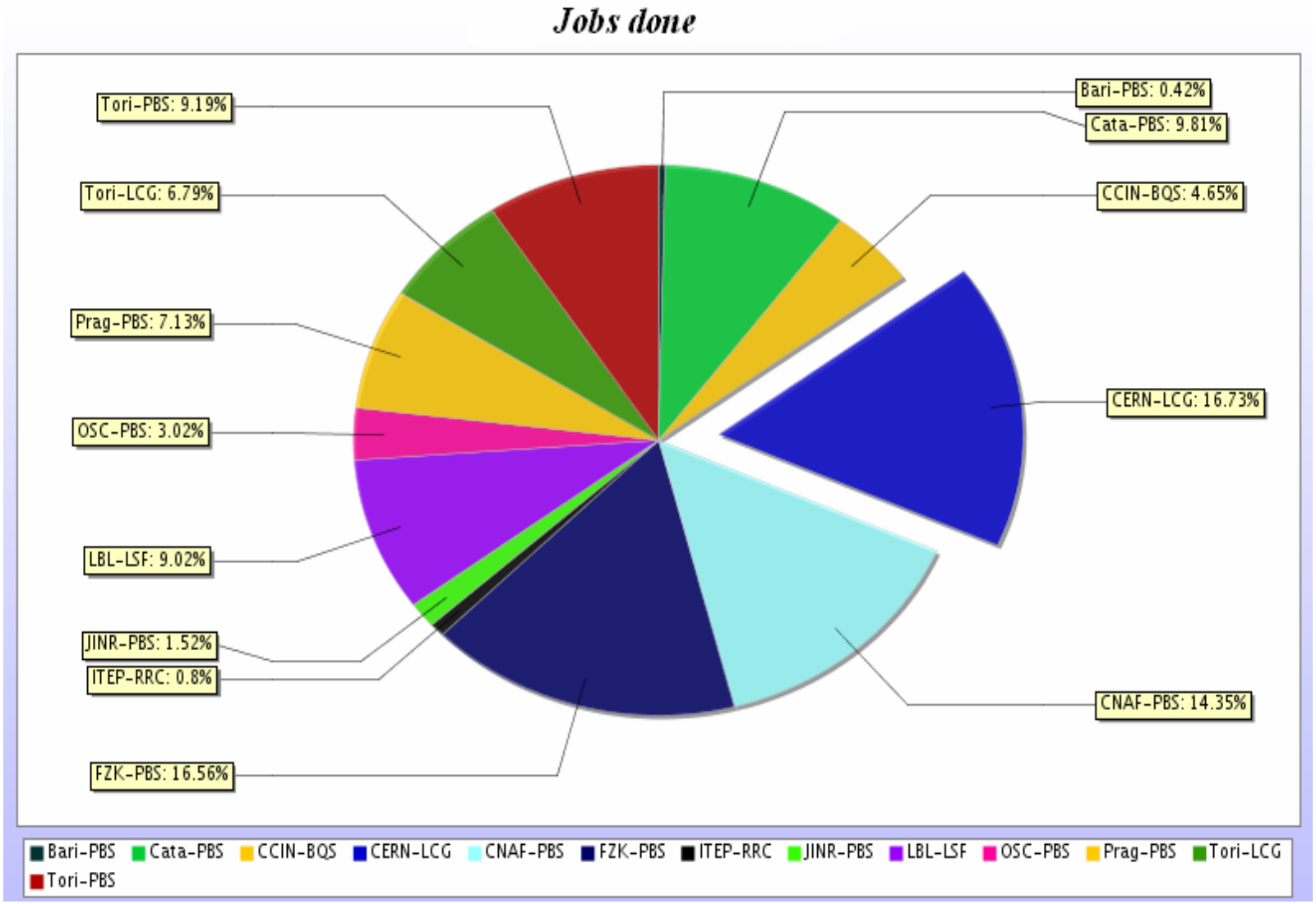
Phase 1 running history



- CEs: Bari, Catania, Lyon, CERN-LCG, CNAF, Karlsruhe, Houston (Itanium), IHEP, ITEP, JINR, LBL, OSC, Nantes, Torino, Torino-LCG (grid.it), Pakistan, Valencia (+12 others)

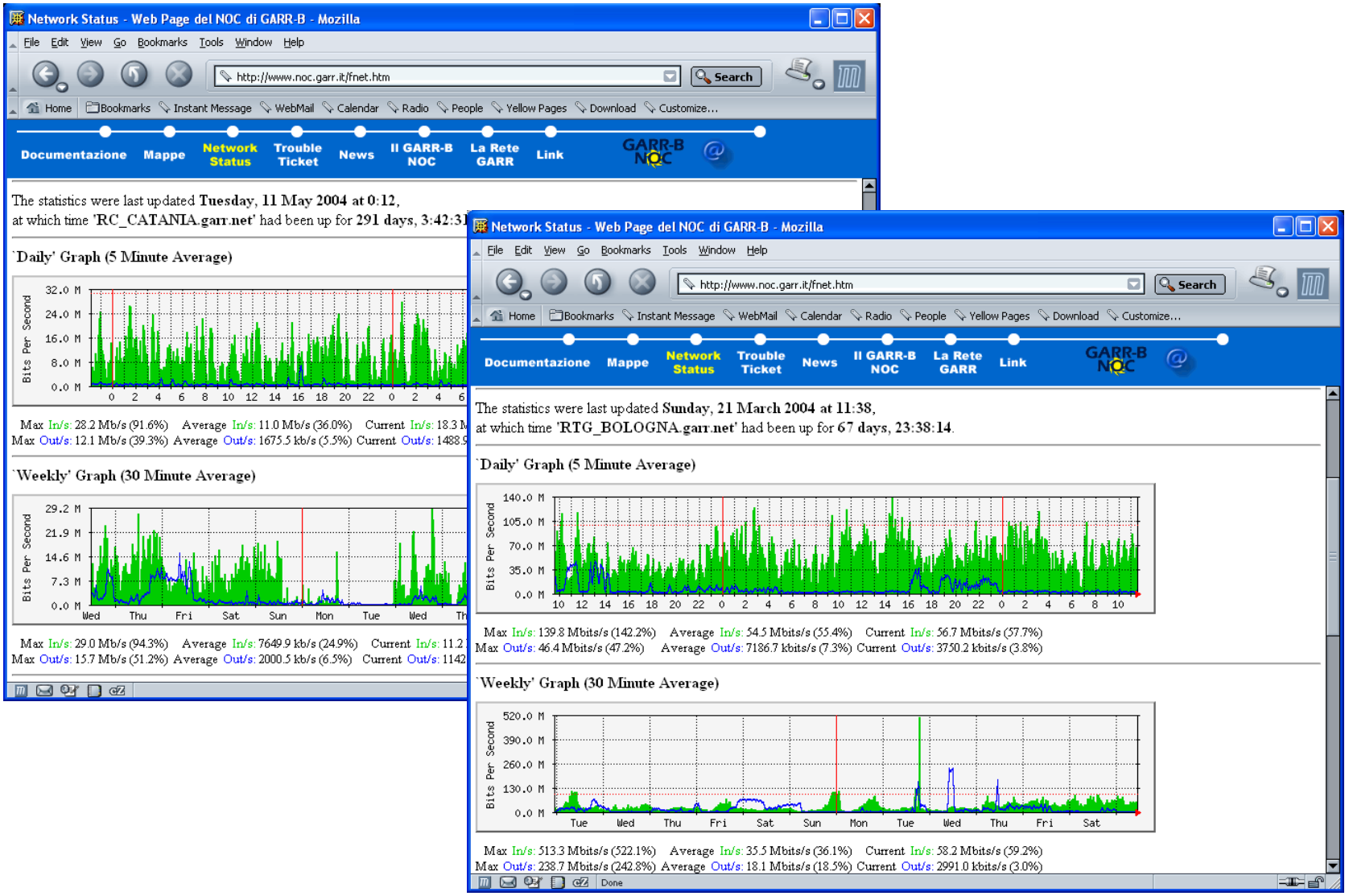
Resource contribution of individual sites (Phase 1)

Enabling Grids for E-science in Europe



Network use during Phase 1

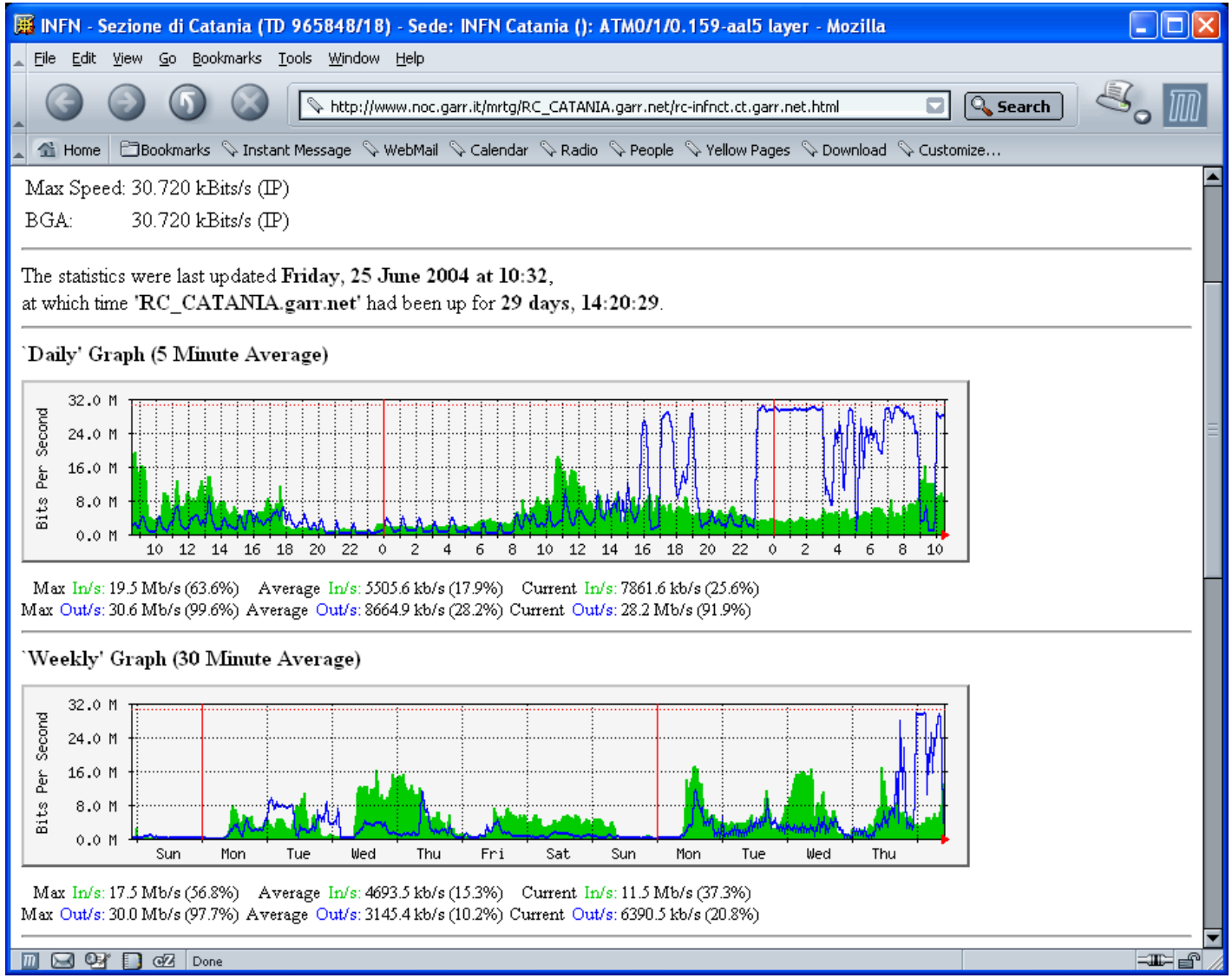
Enabling Grids for E-science in Europe



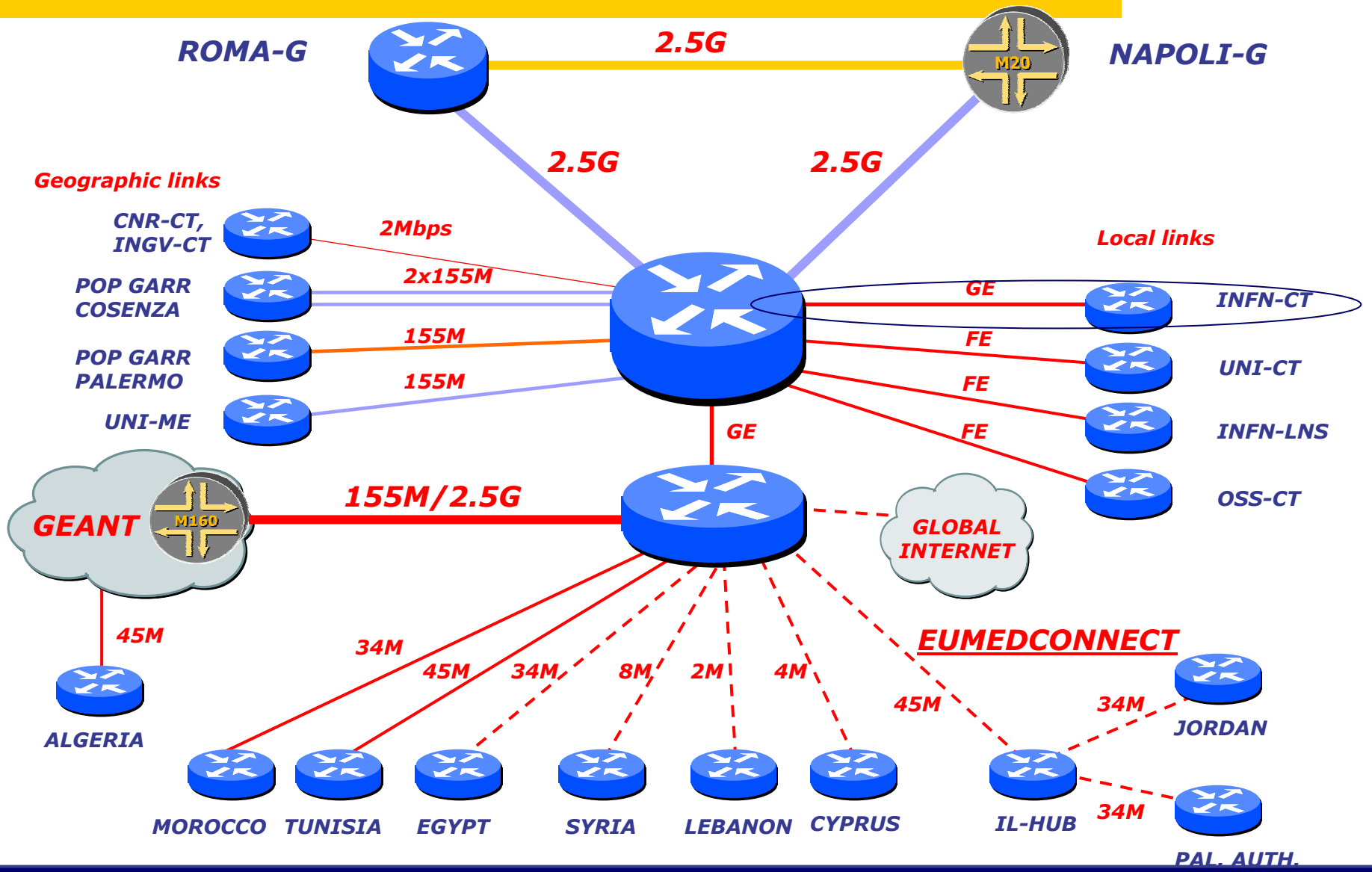
Official GARR NOC statistics

Network use during Phase 2

Enabling Grids for E-science in Europe



GARR-G connections at Catania
(INFN-CT local link operational from July, 21!)



General conclusions

Enabling Grids for E-science in Europe

- Virtual Organizations' planned and chaotic activities have big impacts on networks and strongly rely on their robustness and reliability.
- Network not only means the high bandwidth of international links but also, and more importantly, reliable end-to-(many)ends connections ("last mile" problems should be addressed and hopefully solved).
- The concept of Grid Network Element (emerging in the grid information schemas) should be pursued and implemented as soon as possible.
- Scientific "collaboratories" are very dynamical as a function of both space and time so best effort and over-provisioning are not always the best solutions. **Quality of Service and Bandwidth-on-Demand will be key issues of future networks**.