GridFTPを使用したPHENX 実験の RIKEN-BNL データ転送

市原卓, 渡邊康, 四日市悟

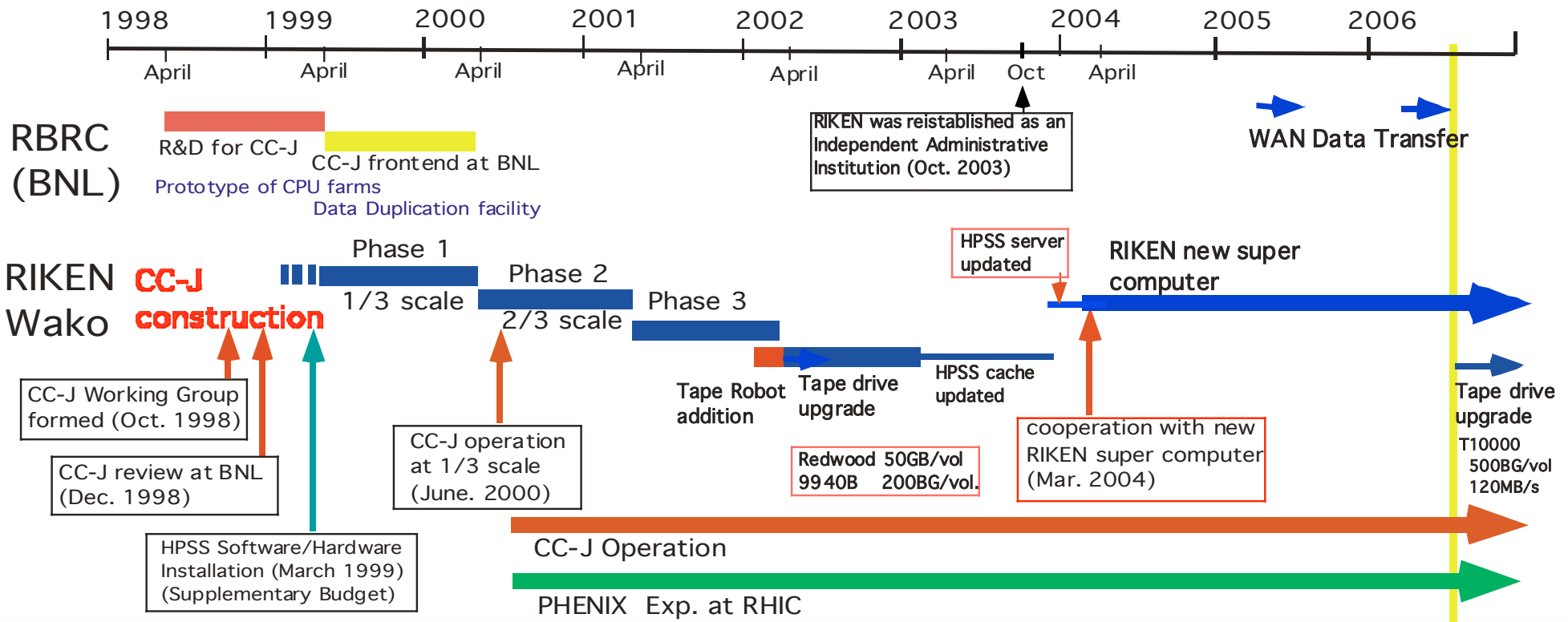
理研, RIKEN-BNL Research Center

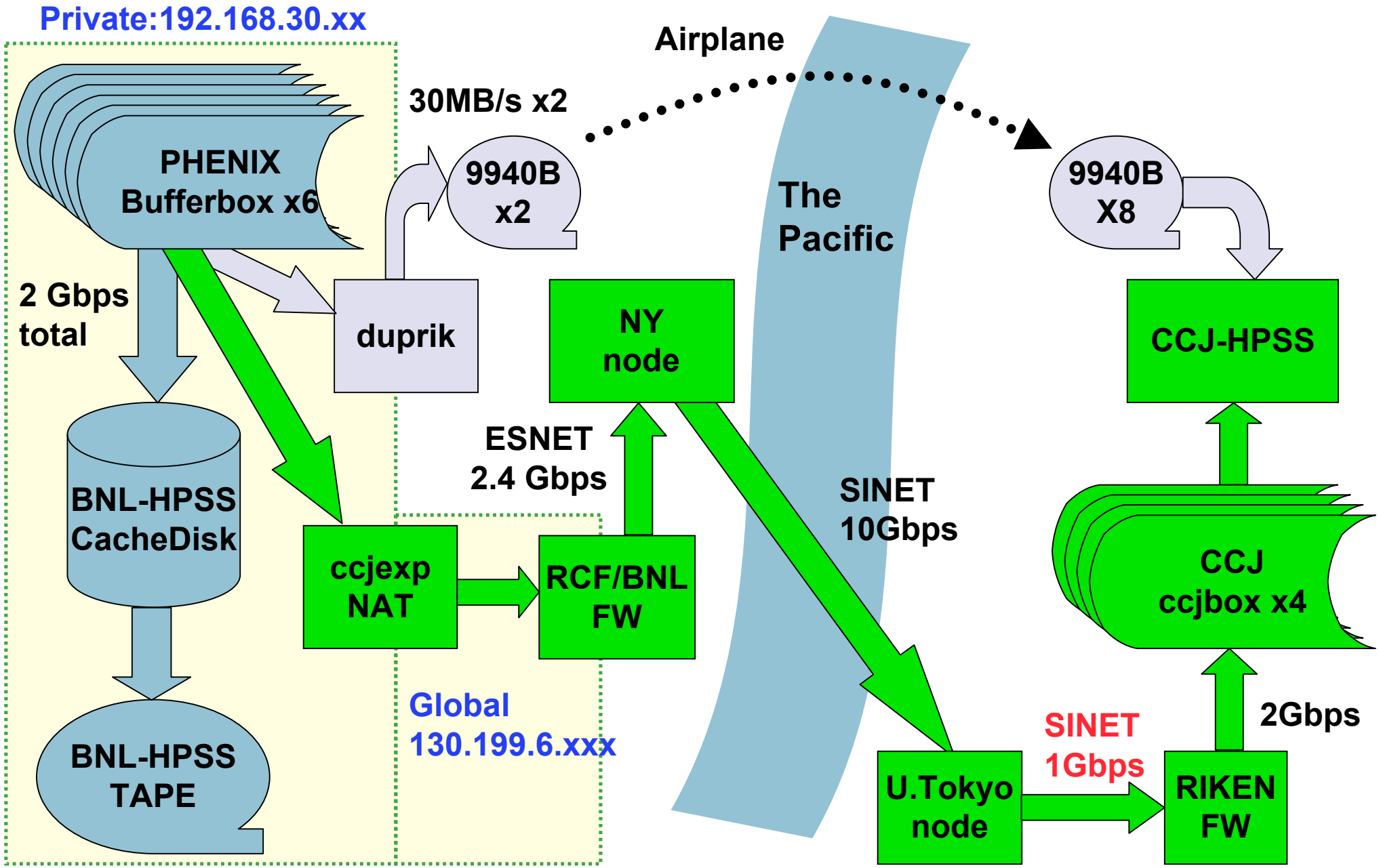
26 July 2006 at ICEPP

RIKEN CCJ : Overview

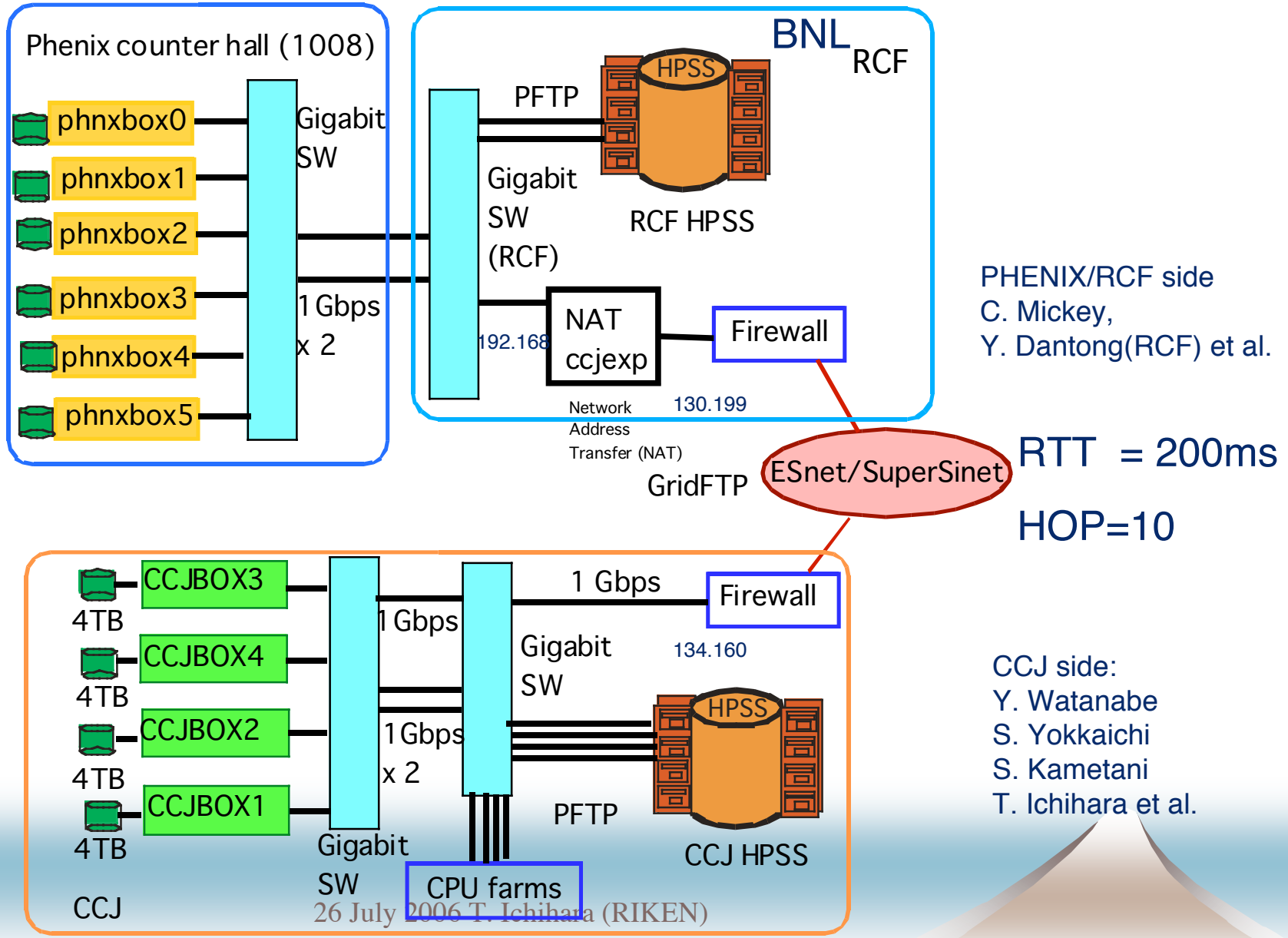
- Scope of CCJ
  - **RHIC スピン物理の解析センター**
  - PHENIX シミュレーション
  - **PHENIXのアジア地域計算センター**
- Size of CCJ
  - 年間取扱うデータ量: 300 TB /year
  - ディスク容量 : ~ 50 TB,
  - テープ容量: ~ 1200 TB capacity (HPSS)
  - CPU 性能 : 256 CPU (Xeon 3.05 GHz) +260 CPU (0.8-2GHz)
- History
  - R&D for the CC-J started in April '98 at RBRC in BNL
  - Construction began in **April '99**
  - CCJ started operation in June **2000**
  - cooperation with a new RIKEN Supercomputer (2004)

History of the CCJ construction and operaton




Private:192.168.30.xx
PHENIX
Bufferbox x6
30MB/s x2
9940B
x2
Airplane
The
Pacific
9940B
X8
2 Gbps
total
duprik
NY
node
CCJ-HPSS
ESNET
2.4 Gbps
SINET
10Gbps
BNL-HPSS
CacheDisk
ccjexp
NAT
RCF/BNL
FW
CCJ
ccjbox x4
Global
130.199.6.xxx
2Gbps
SINET
1Gbps
U.Tokyo
node
RIKEN
FW
BNL-HPSS
TAPE

Overview of Data transfer from PHENIX to CCJ



RIKEN-BNL GridFTP データ転送マシンの環境

•Hardware

- •CCJBOX3
- •CPU: Dual EM64T Xeon 3.6 GHz, 8GB Memory
- •Motherbord: Supermicro X60HE-XG2
- •Gigabit NIC : HP NC7711 (Broadcom BCM95703)

S-ATA Raid5 (4TB) via 2GB FC Host Adapter

- •CCJBOX4
- •Dual Opteon 252 2.6GHz, 8GB Memory
- •Motherboad: Thnder K8W
- •Gigabit NIC : HP NC7711 (Broadcom BCM95703)
- •S-ATA Raid5(4TB) via 2GB FC Host Adapter

•Software

OS: Scientific Linux 4.2 (x86_64) (RHEL4 compatible)

Kernel: 2.6.9-22.0.2.106.unsupportedsmp (CentosPlus) XFS,JFS,ReisrFS

File system : **XFS** (data area) , ext3 (OS part)

Grid environment

The Virtual Data Toolkit (VDT 1.2.4)

(http://vdt.cs.wisc.edu/index.html) (University of Wisconsin-Madison)

The Virtual Data Toolkit (VDT) is an ensemble of **grid middleware** that can be easily installed and configured.

必要な Grid tool一式が pacman-2.129 で簡単にインストールできる

Grid certification

Personal CA, Host CA: **DOE Grid Certificate Service**

http://www.doegrids.org/ Particle Physics Data Grid (PPDG)

Gridftp

/etc/grid-security/grid-mapfile設定、 grid-proxy-init, globus-url-copy

6 年前は　CHEP-2000 presentation (Padova Italy)

WAN performance test

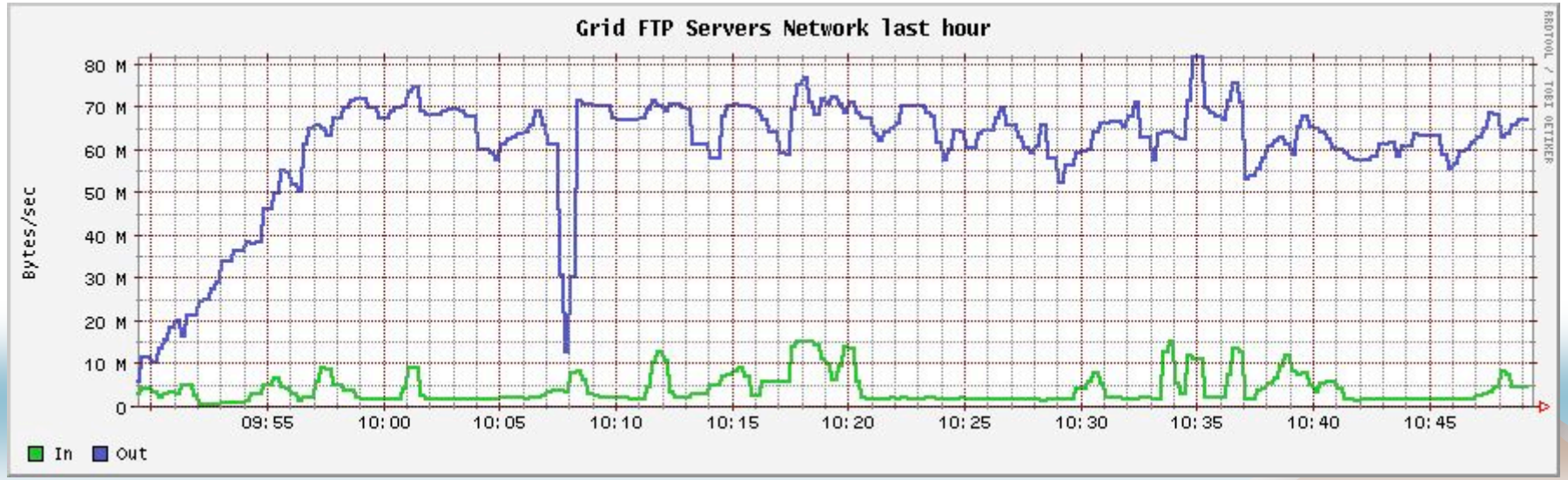
- RIKEN (**12 Mbps**) - IMnet - **APAN (70 Mbps)** -startap- ESnet - BNL (in 2000)
  - Round Trip Time for RIKEN-BNL :170 ms
    - File transfer rate is 47 kB/s for 8 kB TCP widowsize (Solaris default)
    - Large TCP-window size is necessary to obtain high-transfer rate
    - RFC1323 (TCP Extensions for high performance, May 1992) describes the method of using large TCP window-size (> 64 KB)

TCP windowsize	FTP transfer rate (observed)	Theoretical limit For 170 ms RTT
8 kB	41 kB/s	47 kB/s
16 kB	87 kB/s	94 kB/s
32 kB	163 kB/s	188 kB/s
64 kB	288 kB/s	376 kB/s
128 kB	453 kB/s	752 kB/s
256 kB	585 kB/s	1500 kB/s
512 kB	641 kB/s	3010 kB/s

✿ Large ftp performance (641 kB/s = **5 Mbps**) was obtained for a single ftp connection using a large TCP window-size (512 kB) over the pacific ocean (**RTT = 170 ms**)

Network transfer test (Autumn in 2004)

- Network bandwidth (RCF-CCJ): 498 Mbps measured with iperf
  - Bottleneck maybe OC12 (622 Mbps) BNL-(ESnet)-NewYork
  - It will be upgraded to OC48 (2.4 Gbps) until Jan/05
- ~60 MB/s maybe confirmed to transfer by network



Transfer rate for single TCP stream

RFC1323 (TCP Extensions for high performance, May 1992) describes the method of using large TCP window-size (> 64 KB)

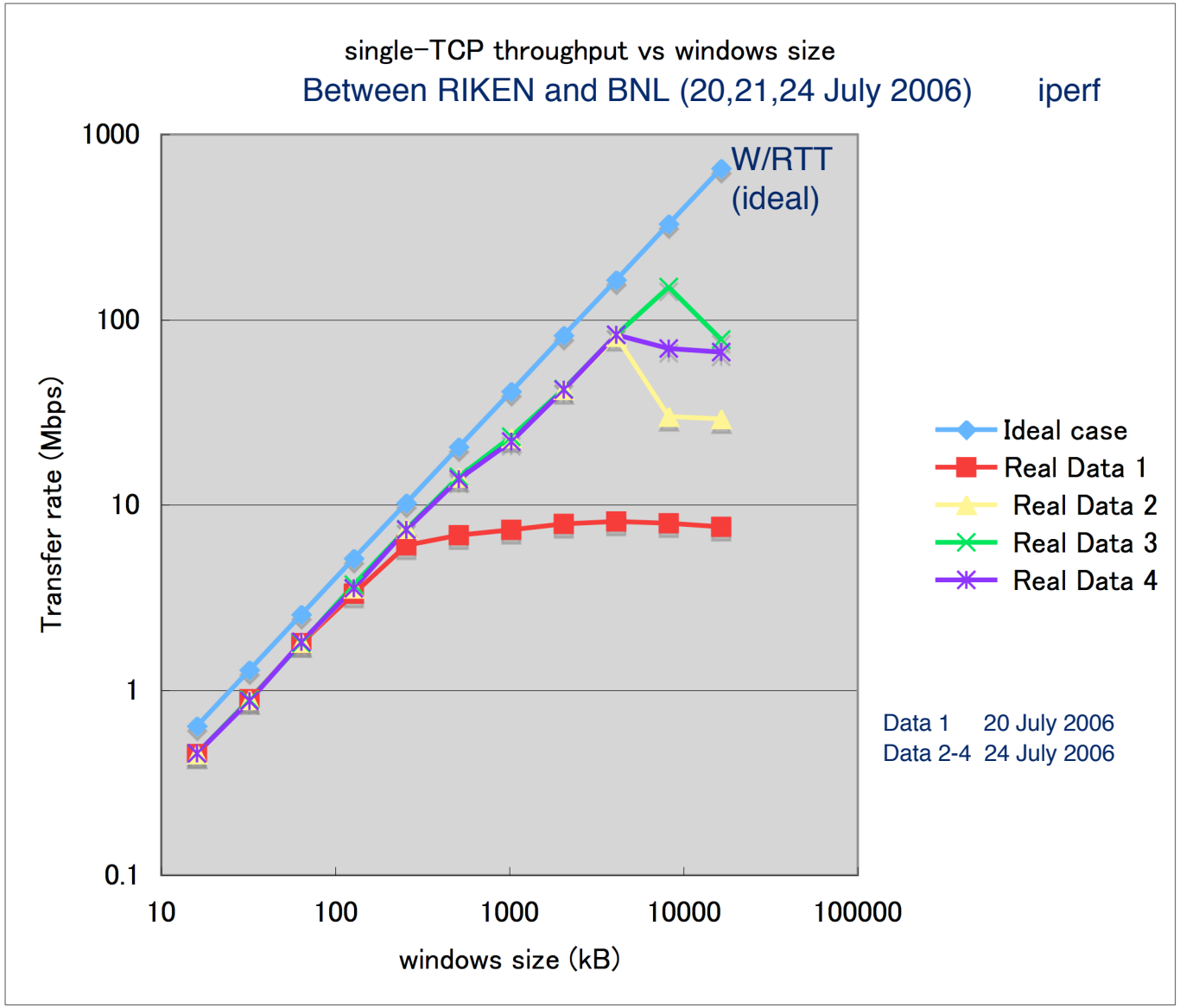
RTT: (RIKEN-BNL): 200ms
Hop between WAN Router :10
RIKEN WAN bandwidth: 1Gbps

パケットロス、ボトムネックのない理想的な場合
Throughput= WindowSize/RTT

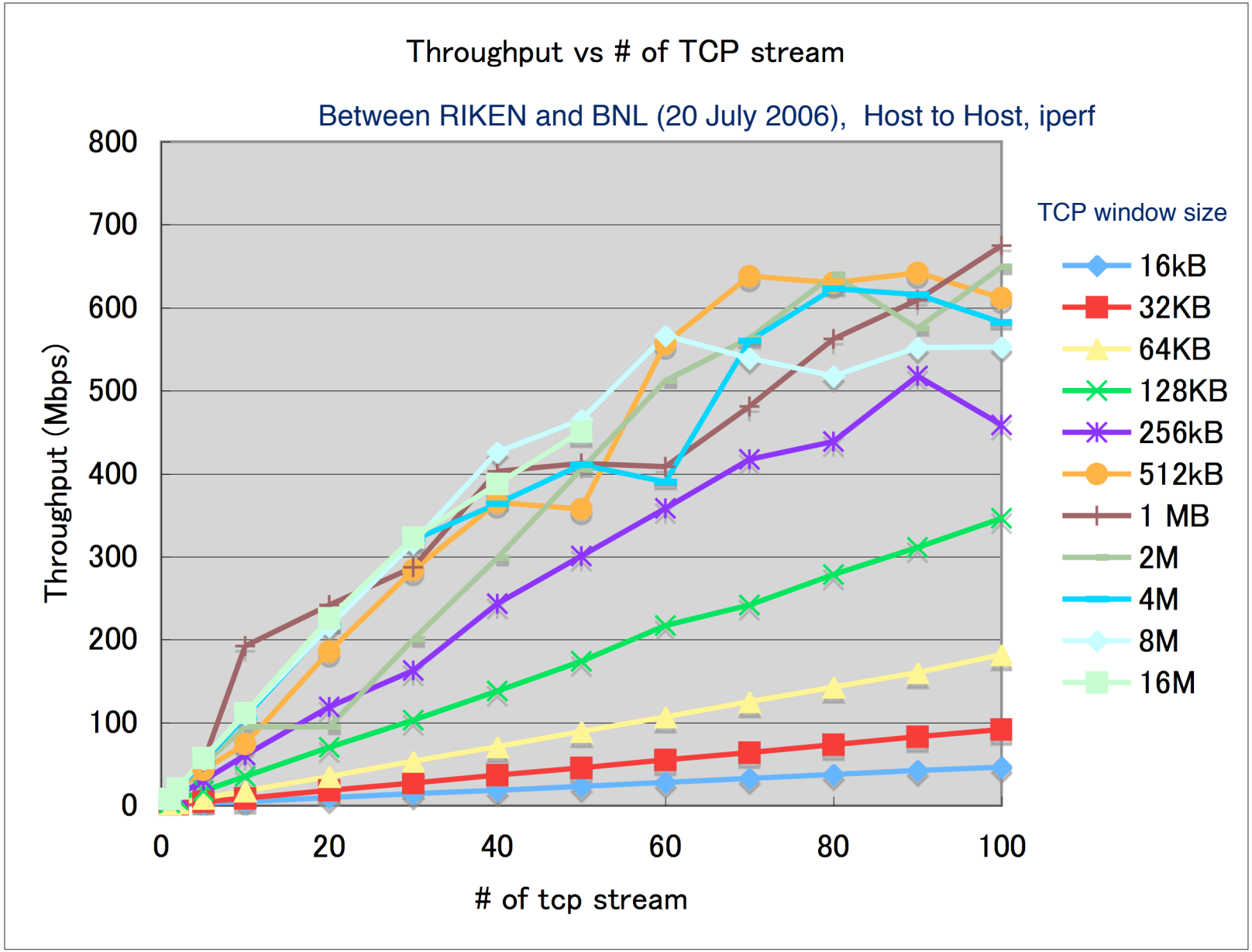
現実のネットワーク
(RIKEN-BNL 間)

Single TCP streamでは
TCP window sizeを増やしていくと 256KB ぐらいまではリニアにスループットが増大するがそれ以上はあるところで飽和し、込み具合で飽和点は変動する

Single TCP 転送の限界



Transfer rate for parallel tcp stream



Ganglia Monitor

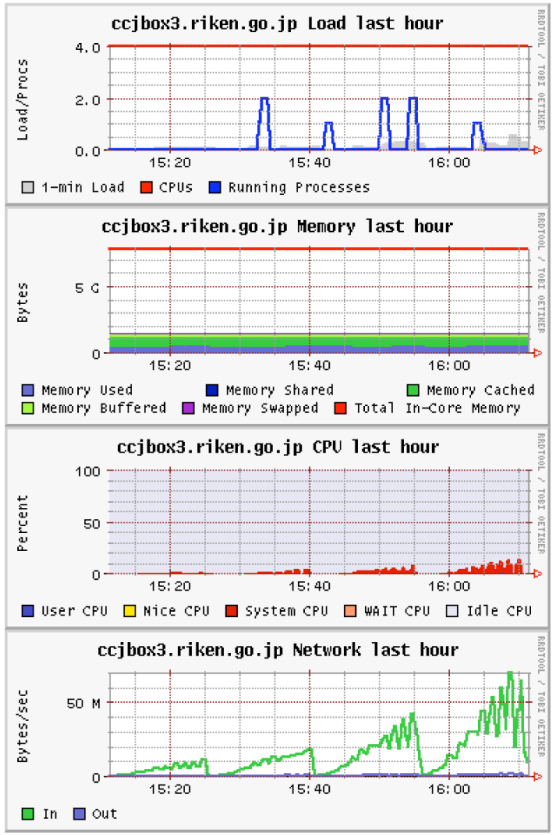
Host Report for Thu, 20 Jul 2006 16:11:26 +0900 | Get Fresh Data | Last hour | Node View

RIKEN CCJ Grid > CCJ Grid Data Servers > ccjbox3.riken.go.jp

ccjbox3.riken.go.jp Overview

This host is up and running.

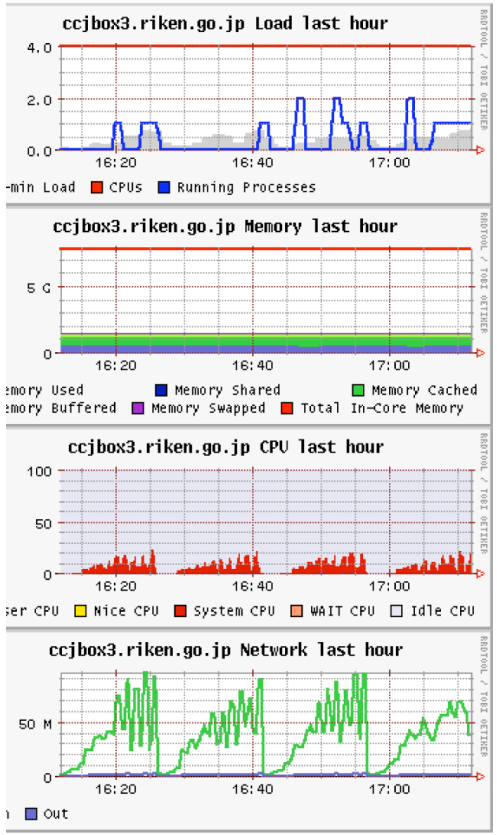
Time and String Metrics	
boottime	Wed, 22 Mar 2006 10:34:34 +0900
gexec	OFF
gmond_started	Fri, 12 May 2006 18:04:25 +0900
last_reported	0 days, 0:00:03
machine_type	x86_64
os_name	Linux
os_release	2.6.11.7
uptime	120 days, 5:36:49



Host Report for Thu, 20 Jul 2006 17:11:57 +0900 | Get Fresh Data | hour | Node View

Grid Data Servers > ccjbox3.riken.go.jp

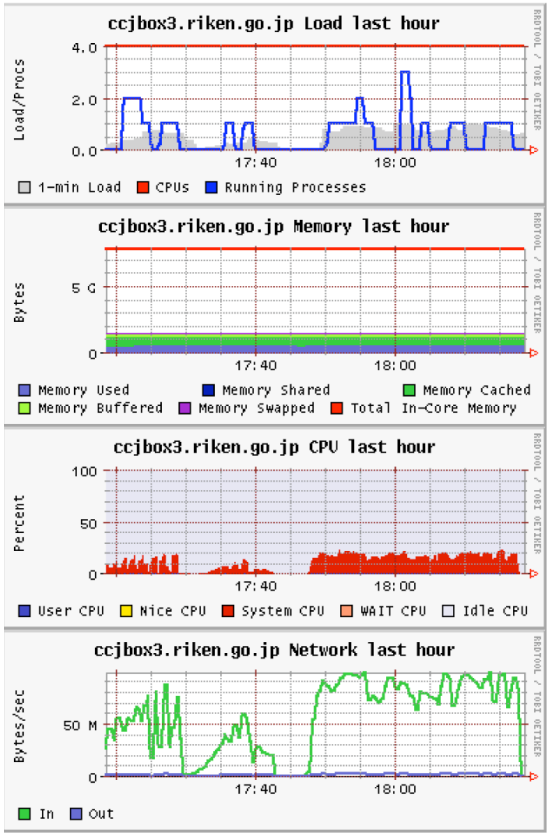
3.riken.go.jp Overview



Host Report for Thu, 20 Jul 2006 18:18:38 +0900 | Get Fresh Data | Last day | Node View

CJ Grid Data Servers > ccjbox3.riken.go.jp

box3.riken.go.jp Overview



TCP window size 32k 64k 128k 256k 512k 1M 2M 4M 8M 16M 1M-80P

Data Transfer (GridFTP) between BNL and RIKEN

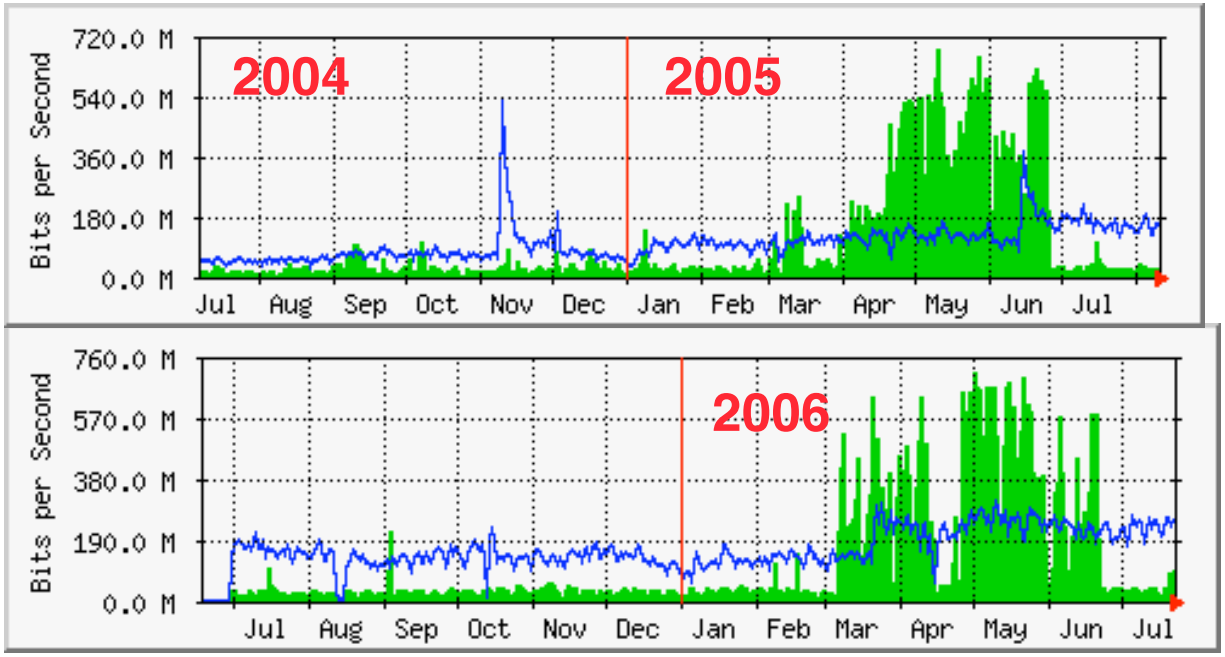
- Single GridFTP でのデータ転送
  - Tcp windows size ： 1-8 MB
  - TCP parallel 度 ： 80 tcp streams/1 GridFTP
  - Memory to Memory (60-90MB/s), Disk to Disk 40-60MB/s

- 実際のデータ転送
  - データ有りの時： 2 - 4 個のGridFTPを同実行
  - 1 Grid ftp : Disk to disk 40-50MB/s
  - 3 Grid ftp : Disk to disk 70-100MB/s

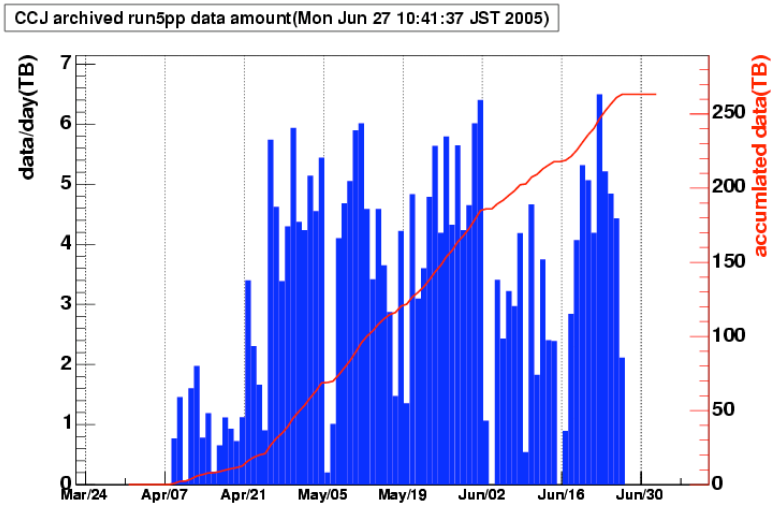
  - RUN 5 (2005) 260 TB
  - RUN 6 (2006) 310 TB

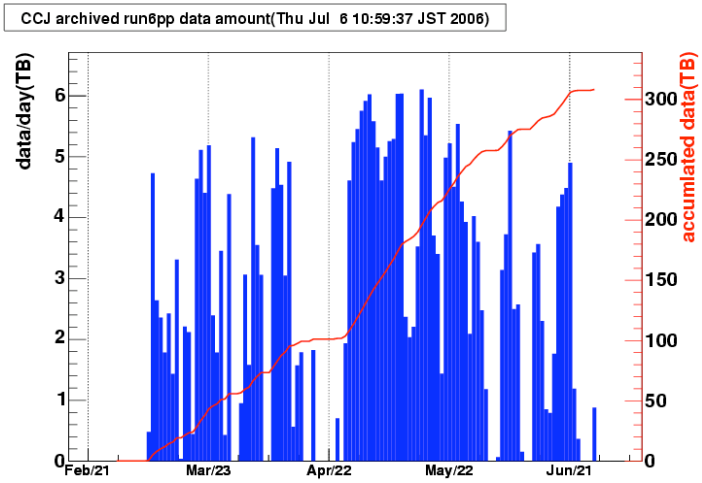
RIKEN WAN traffic and transferred data

MRTG of RIKEN(Wako) WAN Router

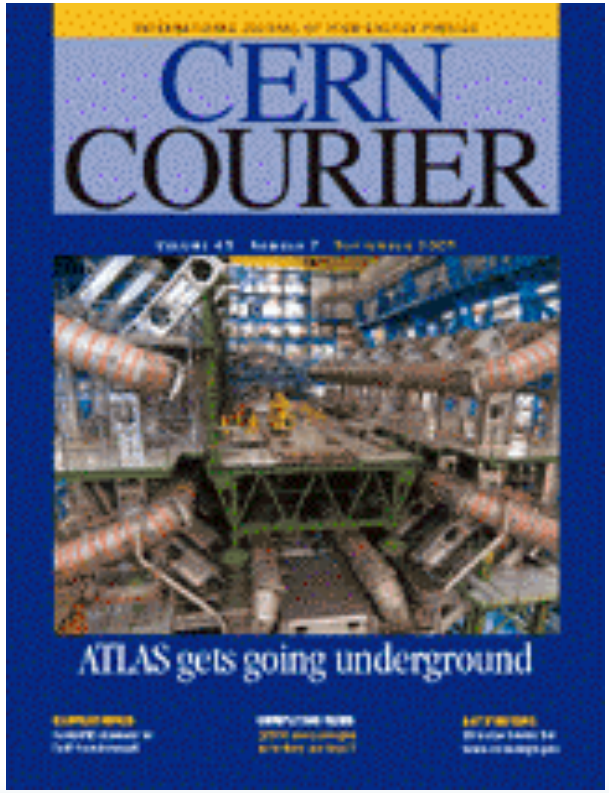


Green : inbound, Blue :outbound traffic





PHENIX experiment uses Grid to transfer 270 TB of data to Japan

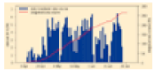


- This seems to be that a data transfer of such magnitude was sustained over many weeks

  - http://www.cerncourier.com/main/article/45/7/15

PHENIX experiment uses Grid to transfer 270 TB of data to Japan

During the polarized proton-proton run that ended in June at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven, Grid tools were used by the PHENIX experiment to send recently acquired data to a regional computing centre for the experiment in Japan. Brookhaven National Laboratory, on Long Island, New York, is home to the RHIC/ATLAS Computing Facility (RCF/ACF), which is the main computing centre for experiments at RHIC and a Tier-1 computing centre for ATLAS. The PHENIX regional computing centre in Japan (CCJ) is at the RIKEN research centre on its Wako campus close to Tokyo.



Data transfer

Going into the polarized proton-proton run, PHENIX faced the challenge that the RCF would be busy reconstructing and analysing gold-gold and copper-copper data recorded in 2004 and 2005. The enormous polarized proton-proton data set was transferred to Japan to make use of the substantial computing resources at CCJ, which is comparable to the PHENIX portion of the RCF.

The PHENIX data acquisition can sustain a peak data rate of up to 600 MB/s, and runs at a typical rate of 250 MB/s while beam is stored in RHIC. The data were buffered at the experimental site before being transferred and archived in the RCF tape library. A 35 TB disk-storage system (about 60 h at typical data rates) allowed PHENIX to archive and transfer data at a lower steady rate, taking advantage of various breaks in the flood of data. A transfer rate of 60 MB/s sustained steadily around the clock was able to keep up with the incoming data stream.

Initially, PHENIX had planned to transfer the polarized proton-proton data by physically transporting tape cartridges to CCJ. During the early part of the run, however, it was found that network transfer rates of 700-750 Mbits/s could be achieved. A dedicated network path was established from the PHENIX counting house to the BNL perimeter network, and the tape option became a fall-back solution. In the end, not a single tape was shipped.

The principal tool used for the transfer was GridFtp, which proved to be very stable. Brookhaven has a high-speed connection (OC48) to ESNET, which is connected to a transpacific line (10 Gbit/s) served by SINET in Japan. Apart from two half-day outages of ESNET, the transfers continued around the clock for the entire 11 week run.

Approximately 270 TB of data (representing 6.8 billion polarized proton-proton collisions) were transferred to CCJ. After a few days of fine-tuning the transfer parameters, the transfers became part of the regular data-handling operation of the PHENIX shift crews, requiring experts to intervene only occasionally.

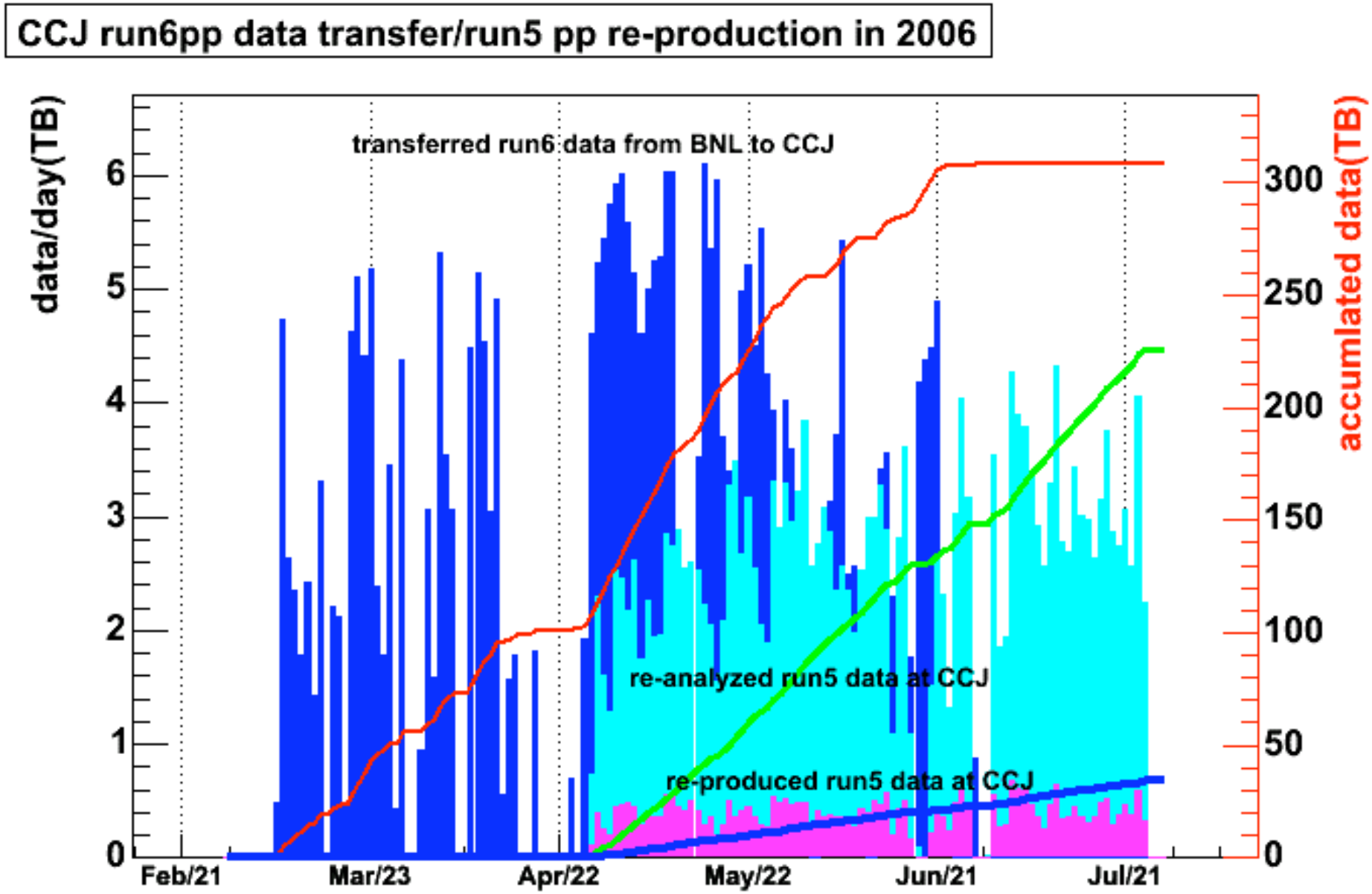
This seems to be the first time that a data transfer of such magnitude was sustained over many weeks in actual production, and was handled as part of routine operation by non-experts. The successful completion of this large-scale transfer project demonstrates both the maturity of today's Grid tools and the real feasibility of integrating remote resources into the data-handling and processing chain of large experiments.

Author:

/etc/sysctl.conf のサンプル(初期値)
(suggested by Dangong Yu @RCF BNL)

/etc/sysctl.conf

- net.ipv4.tcp_rmem = 262144 1048576 8388608
- # sets min/default/max TCP read buffer, default 4096 87380 174760
- net.ipv4.tcp_wmem = 262144 1048576 8388608
- # sets min/pressure/max TCP write buffer, default 4096 16384 131072
- net.ipv4.tcp_mem = 262144 1048576 8388608
- # sets min/pressure/max TCP buffer space, default 31744 32256 32768
- ### CORE settings (mostly for socket and UDP effect)
- net.core.rmem_max = 4194304
- # maximum receive socket buffer size,default 131071
- net.core.wmem_max = 4194304
- # maximum send socket buffer size, default 131071
- net.core.rmem_default = 1048576
- # default receive socket buffer size, default 65535
- net.core.wmem_default = 1048576
- # default send socket buffer size, default 65535
- net.core.optmem_max = 1048576
- # maximum amount of option memory buffers, default 10240
- net.core.netdev_max_backlog = 100000
- # number of unprocessed input packets before kernel starts dropping them, default 300


CCJ run6pp data transfer/run5 pp re-production in 2006
data/day(TB)
accumlated data(TB)
transferred run6 data from BNL to CCJ
re-analyzed run5 data at CCJ
re-produced run5 data at CCJ
0
1
2
3
4
5
6
0
50
100
150
200
250
300
Feb/21
Mar/23
Apr/22
May/22
Jun/21
Jul/21