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, 中村智昭,後藤雄二, 延與秀人
理研, RIKEN-BNL Research Center
15 July 2008, at ADVnet2008 meeting
Contact： Ichiharaあとriken.jp

RHIC

relativistic heavy ion collider

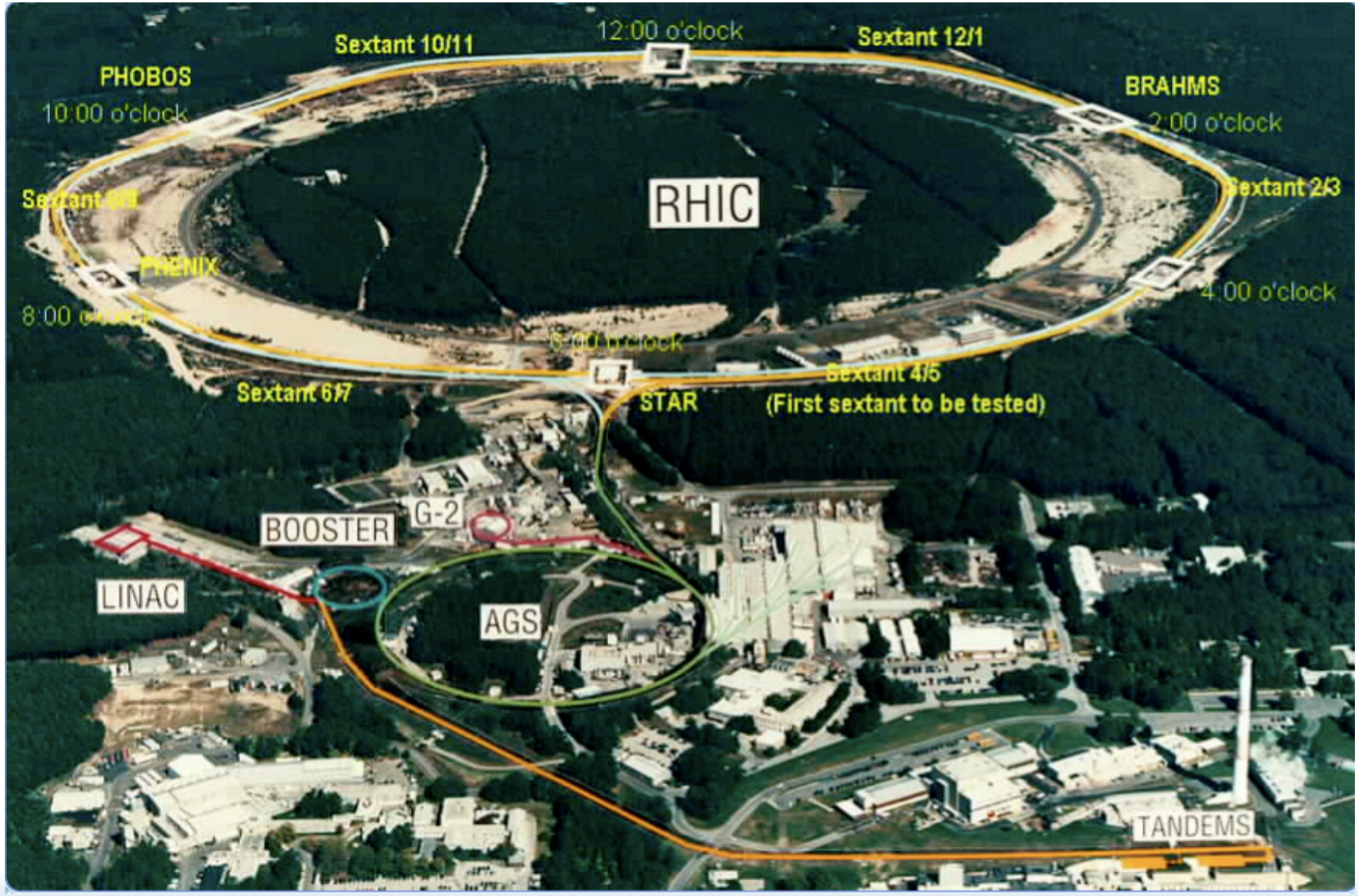
The 2.4-mile circumference RHIC ring is large enough to be seen from space.

The image above shows Long Island, New York, as viewed by the multispectral scanner of the Landsat-4 satellite in July of 1982. At the time the image was taken, tunnel construction was underway for the predecessor project (called 'Isabelle') that would eventually become RHIC. The image at right, where the ring is clearly visible, is an enlargement of the area highlighted above.

< Back to the RHIC home page.

http://www.bnl.gov/rhic/from_space.htm

超高エネルギー衝突型加速器(RHIC) @ BNL Long Island, NY



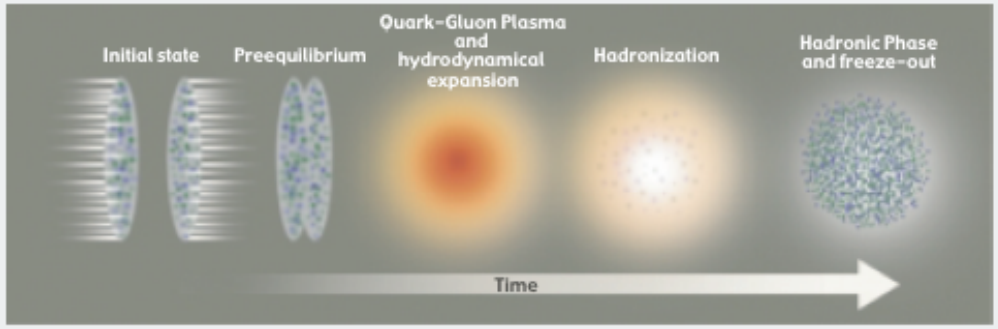
構成: 超伝導電磁石を使用した2重の衝突リング（円周長 3.8 km）
入射: バンデグラフ -> ブースター -> AGS -> RHIC

性能:	金+金 衝突	陽子+陽子衝突
ビームのエネルギー	100 GeV/A	250 GeV
ルミノシティー	2x1026 cm-2s-1	1.4x1031 cm-2s-1
完成	1999 年	

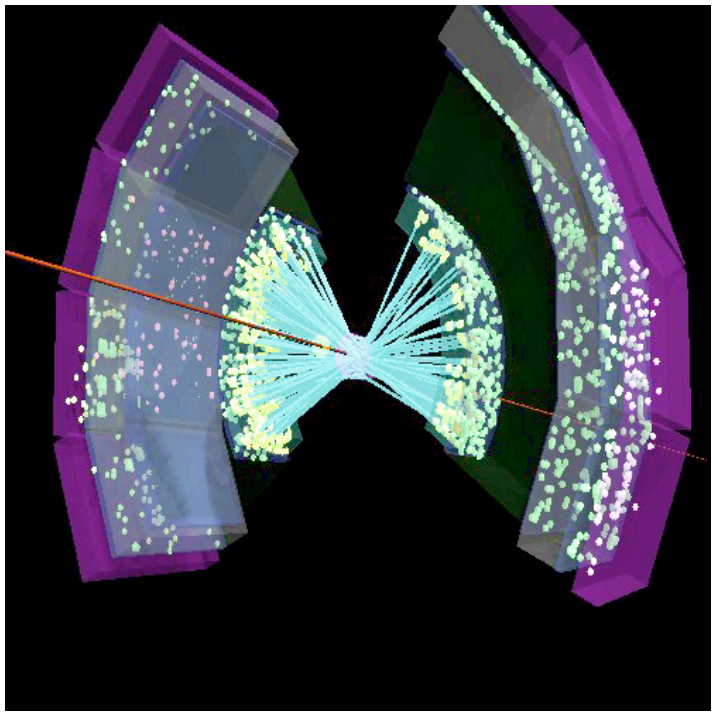
RHICのリング(右回りと左回りの2重リング　円周長 3.8 km)

研究目的

1. 原子核衝突による、宇宙初期の高温高密度状態の研究（クォーク・グルオン・プラズマの検証）
2. 核子のスピン構造

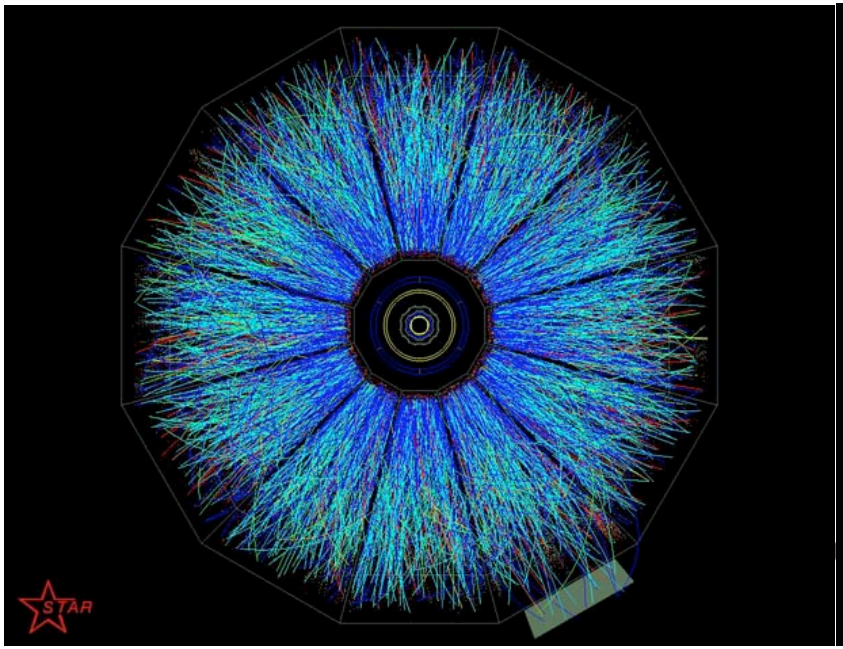


In this schematic illustration two gold nuclei collide and give rise to thousands of quark and gluons, which then equilibrate into a hot cauldron of matter, the quark-gluon plasma. As this plasma cools, it condenses into the ordinary particles seen by the detectors.

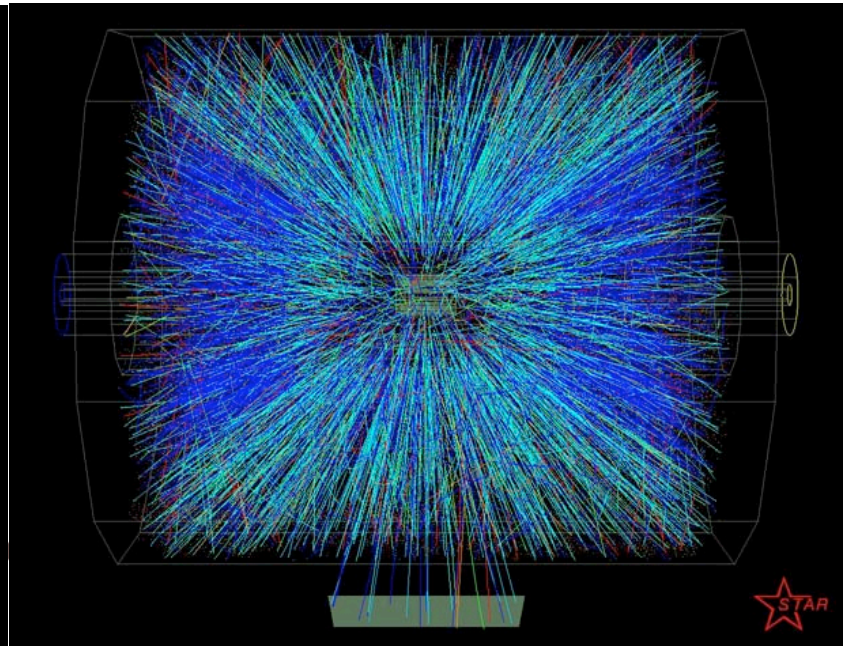


PHENIX event view

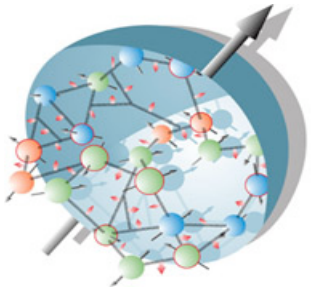
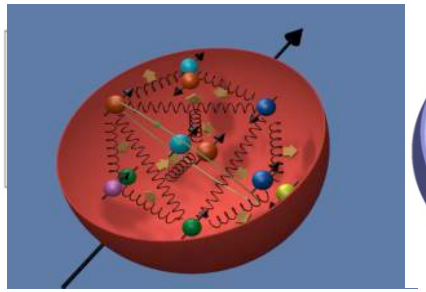
Star front view

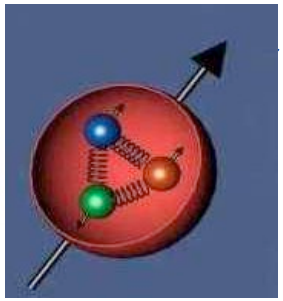
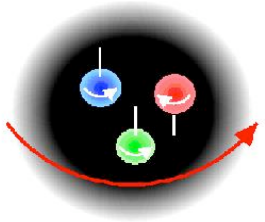
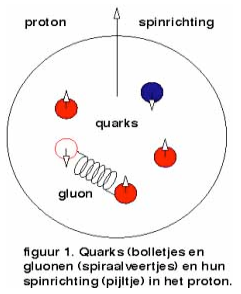


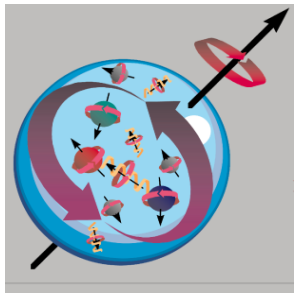
Star side view



（研究目的）核子のスピン構造

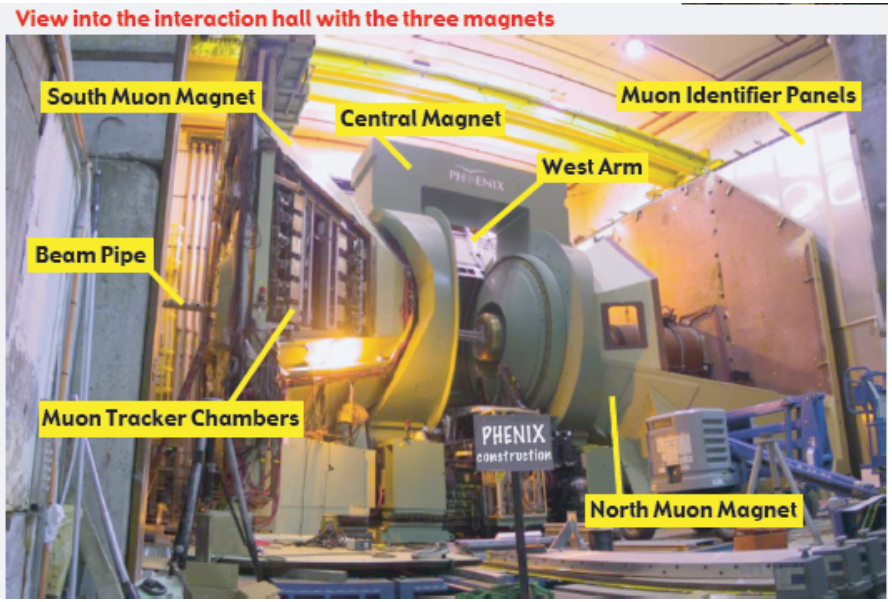
figuur 1. Quarks (bolletjes en gluonen (spiraalveertjes) en hun spinrichting (pijltje) in het proton.


- 核子(陽子、中性子)の内部の多体構造

$$\frac{1}{2} = \frac{1}{2}\Delta\Sigma + \Delta g + L_q + L_g$$

  - クォークスピンの寄与（$\Delta\Sigma$）、グルーオンスピンの寄与（Δg）、軌道角運動量の寄与（L_q, L_g）
- 歴史
  - EMC実験@CERN, 偏極レプトン深非弾性散乱（DIS）実験
    - 小さい$\Delta\Sigma$（Spin Puzzle）, クォークスピンの寄与は30%程度しかない
  - Δg測定実験
    - 偏極レプトンsemi-inclusive DIS実験, 偏極ハドロン衝突実験

PHENIX測定装置



- 10ヵ国、42大学•研究機関、約450人の国際共同研究
- 国内（理研、京大、KEK、東大、CNS,筑波大、広島大、東工大、早稲田大、長崎総技大）

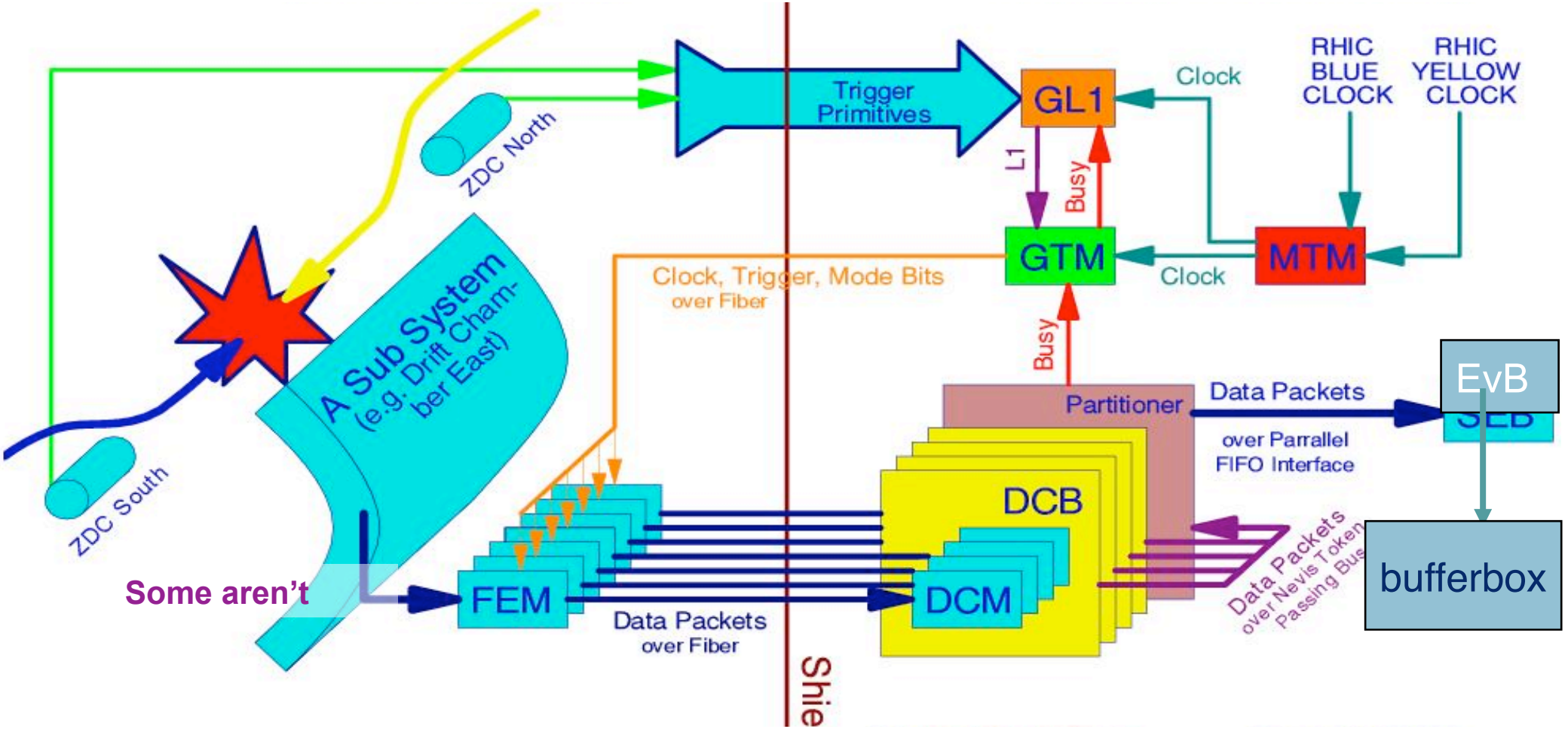
- 検出器のチャンネル数：**40万チャンネル**
- 衝突頻度 10MHz (100nsに1度)
- イベントサイズ
  - 金の原子核同士の衝突 180KB/event
  - 偏極陽子同士の衝突 100KB/event
- トリガーレート
  - 5-12.5kHz
- 実験データ収集量 最大で **800MB/s**
- **（圧縮後 400MB/s）**（設計当初は40MB/s）
  - 生データはBNLにある HPSSにアーカイブされるとともに、偏極陽子+陽子 衝突 実験の生データは理研にWANで準リアルタイムで転送する。

PHENIX ミュオン南電磁石

（阪神大震災直後に、三菱電機神戸製作所で理研が製作）

PHENIX のデータ収集システム概要

Some systems are used to make triggers



A **partition** is one or more granules that receive the same triggers & busies

Regional Computing Center (RIKEN CCJ)

2000年より運用開始

♦ CCJの目的

- **RHIC スピン物理の解析センター（いちはやく実験データを解析）**
- **PHENIXのアジア地域計算センター**
- PHENIX シミュレーション

♦ CCJの規模

- 年間取扱うデータ量: **300 TB /年（毎年、米国から日本へ転送）**
- ディスク容量 : **~135TB**,
- テープロボット容量: ~ **1400 TB (1.4 PB)** for CCJ **(HPSS)**
- CPU 性能 : 256 CPU (Xeon 3.05 GHz) +108 CPU

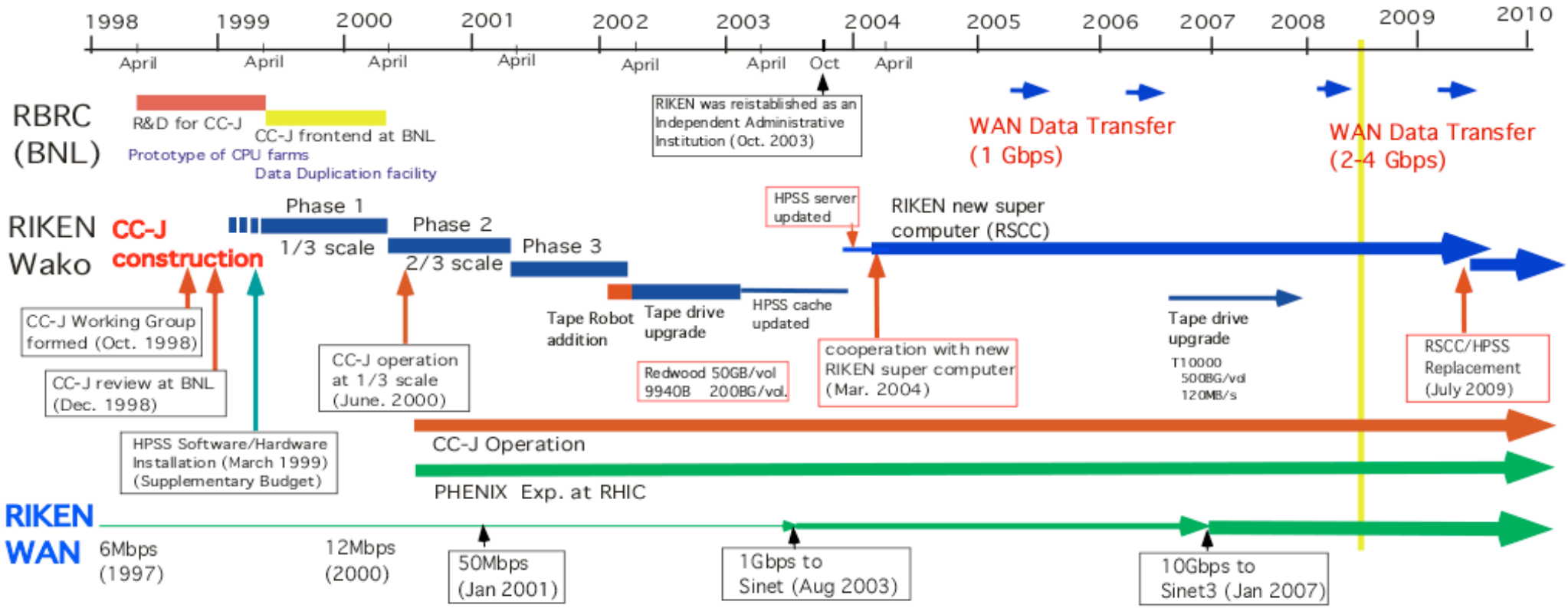
BNL RHIC (Relativistic Heavy Ion Collider) での国際研究協力協定

日米科学技術協力協定(1988年)

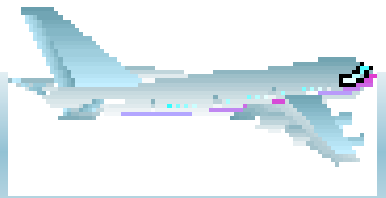
スピン物理研究に関するSTA-DOE実施取極(1995)

「スピン物理」に関する理研-BNL研究協力協定 (1996年9月、5年ごとに更新)

History of the CCJ construction, operation and RIKEN WAN



TISN(1987 64kbps) → STAnet/IMnet(1994)
STAnet → IMNET/APAN (June 1998)
IMNET/APAN→ SINET (Aug. 2003)
SINET→ SINET3 (Jan. 2007)



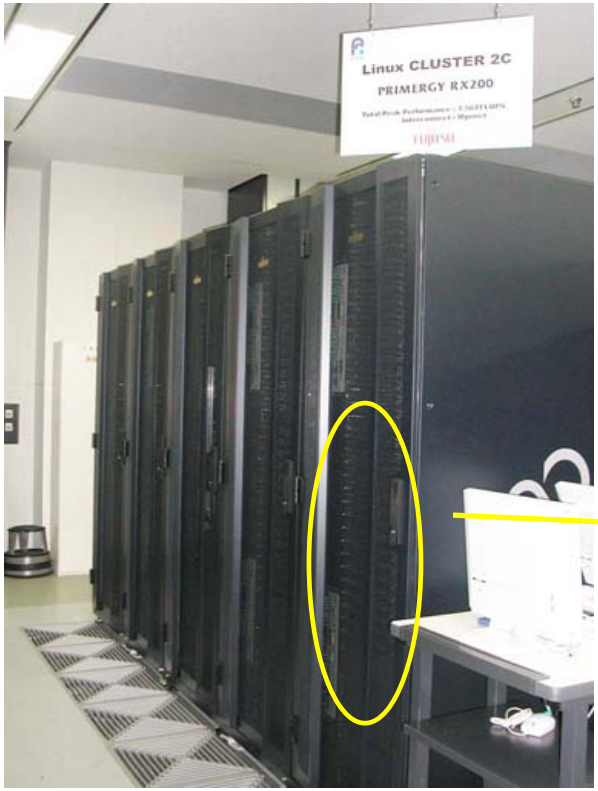
Components of RIKEN CCJ

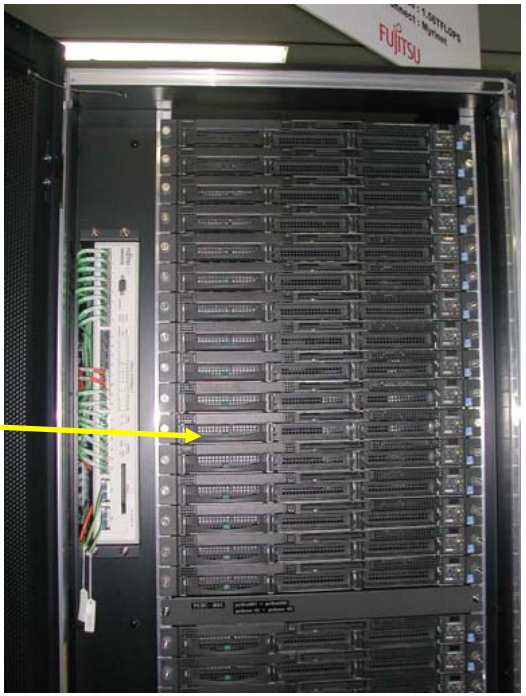
HPSS Server

HPSS

High
Performace
Storage
System

(DOE+IBM)





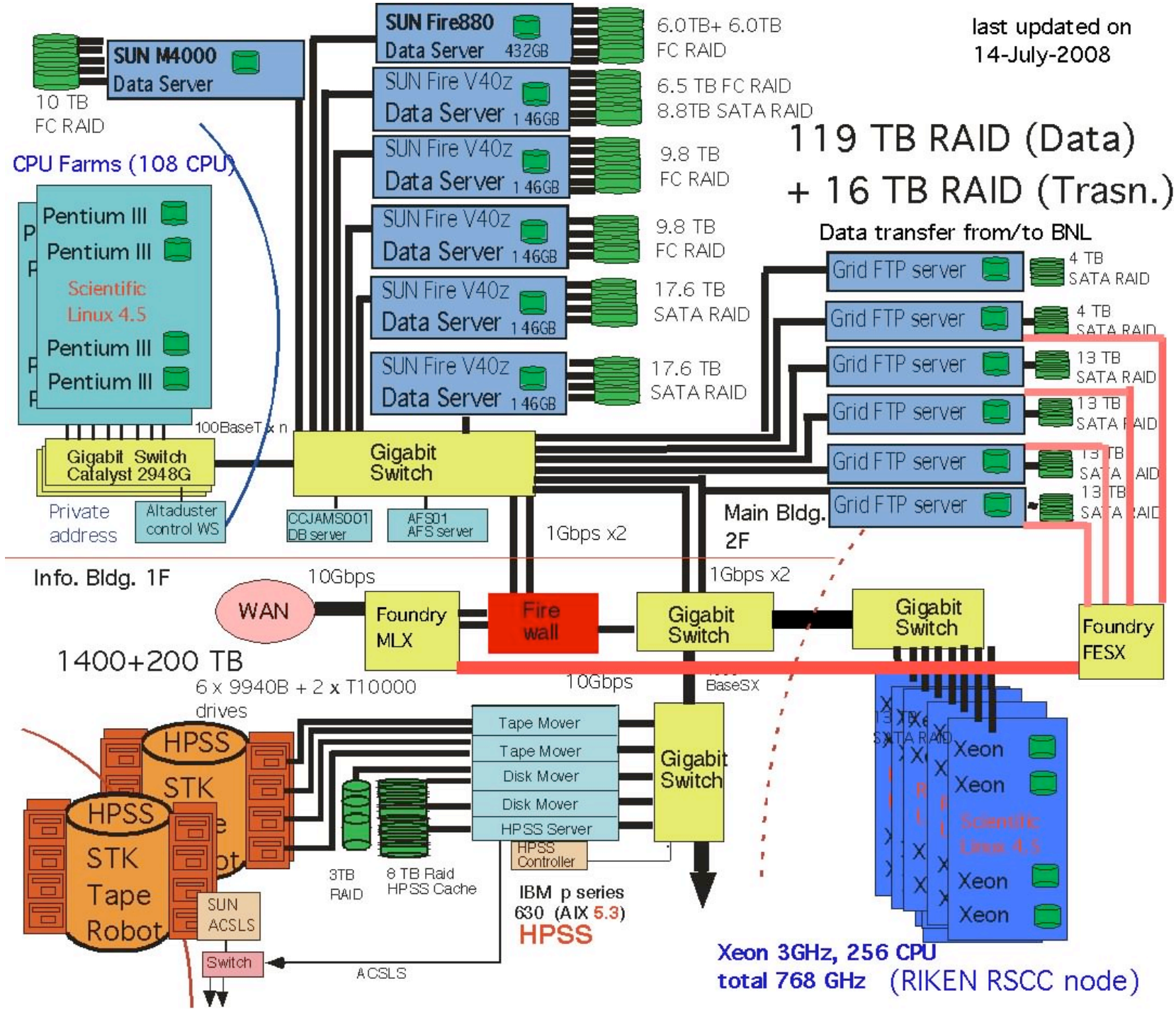
RIKEN common

CCJ

CCJ allocated part of new RIKEN Supercomputer:
128 nodes 256 CPU (Intel Xeon 3 GHz) :
(1/8 of entire system)
(Entire Super computer: 1024 node 2048 CPU)
Next Replace: 2009 July

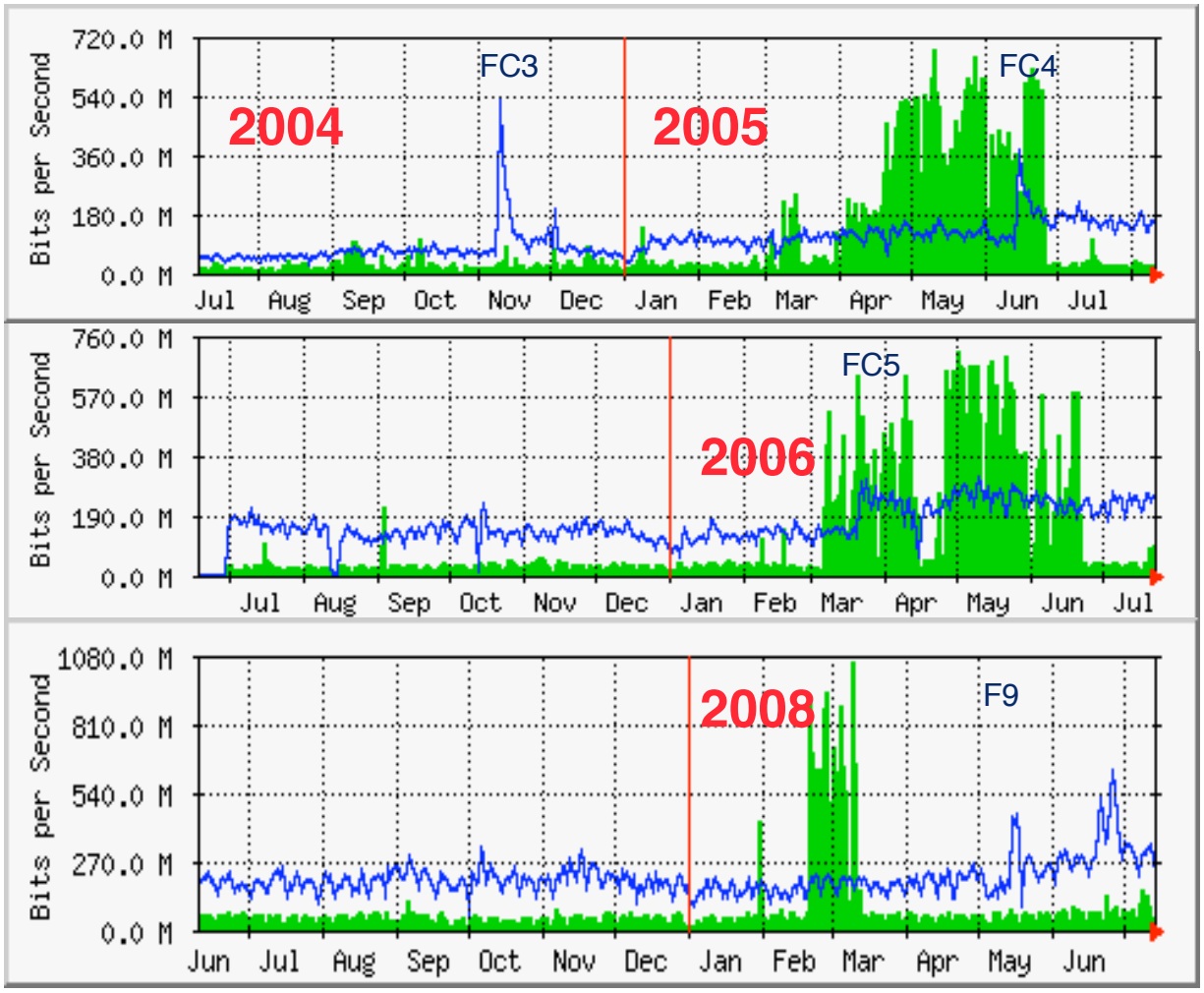
Tape silo [StorageTek(SUN) PowderHorn) 6000 tapes/unit

CCJ の構成

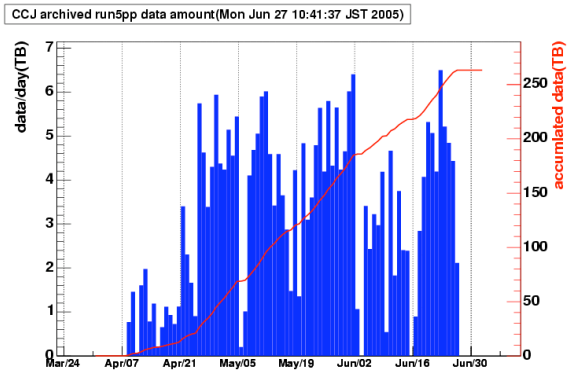

last updated on
14-July-2008
SUN M4000
Data Server
10 TB
FC RAID
SUN Fire880
Data Server 432GB
6.0TB+ 6.0TB
FC RAID
SUN Fire V40z
Data Server 146GB
6.5 TB FC RAID
8.8TB SATA RAID
SUN Fire V40z
Data Server 146GB
9.8 TB
FC RAID
SUN Fire V40z
Data Server 146GB
9.8 TB
FC RAID
SUN Fire V40z
Data Server 146GB
17.6 TB
SATA RAID
SUN Fire V40z
Data Server 146GB
17.6 TB
SATA RAID
119 TB RAID (Data)
+ 16 TB RAID (Trasn.)
Data transfer from/to BNL
CPU Farms (108 CPU)
Pentium III
Pentium III
Scientific
Linux 4.5
Pentium III
Pentium III
100BaseTx n
Gigabit Switch
Catalyst 2948G
Gigabit
Switch
Grid FTP server
4 TB
SATA RAID
Grid FTP server
4 TB
SATA RAID
Grid FTP server
13 TB
SATA RAID
Grid FTP server
13 TB
SATA RAID
Grid FTP server
Grid FTP server
Private
address
Altaduster
control WS
CCJAMS001
DB server
AFS01
AFS server
1Gbps x2
Main Bldg.
2F
Info. Bldg. 1F
10Gbps
WAN
Foundry
MLX
Fire
wall
Gigabit
Switch
Gigabit
Switch
Foundry
FESX
1400+200 TB
6 x 9940B + 2 x T10000
drives
10Gbps
BaseSX
HPSS
STK
HPSS
STK
Tape
Robot
SUN
ACSLS
Switch
ACSLS
Tape Mover
Tape Mover
Disk Mover
Disk Mover
HPSS Server
HPSS
Controller
Gigabit
Switch
3TB
RAID
8 TB Raid
HPSS Cache
IBM p series
630 (AIX 5.3)
HPSS
Xeon
Xeon
Scientific
Linux 4.5
Xeon
Xeon
Xeon 3GHz, 256 CPU
total 768 GHz (RIKEN RSCC node)

RIKEN WAN traffic とこれまでのWAN実験データ転送量

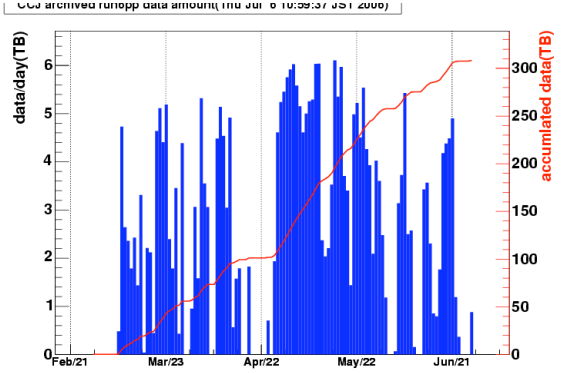
MRTG of RIKEN(Wako) WAN Router



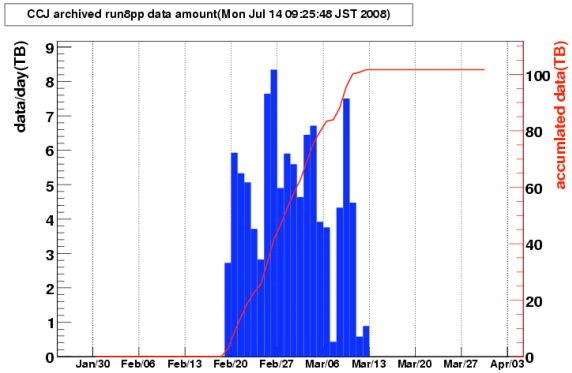
Green : inbound, Blue :outbound traffic



Run5 pp
2005
263 TB



Run6 pp
2006
308 TB



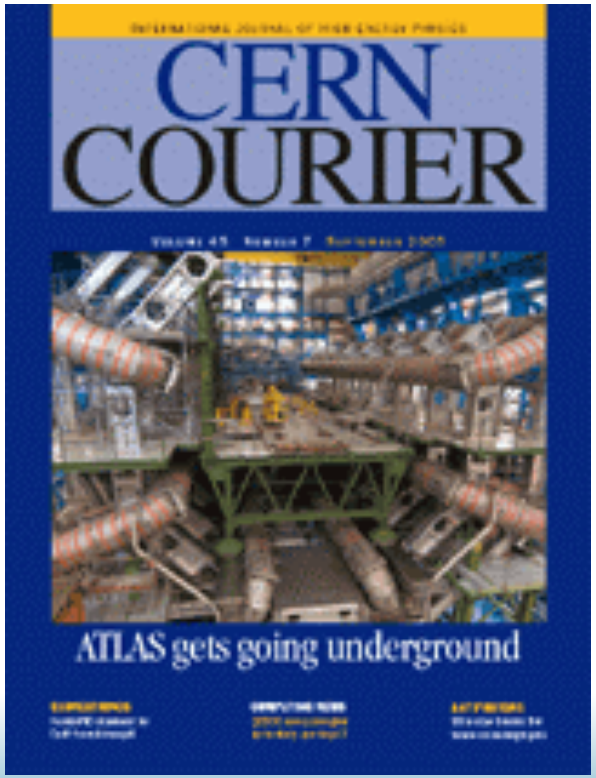
Run8 pp
2008
100 TB

15 July 2008 T. Ichihara (RIK

http://ccjsun.riken.go.jp/ccj/project/run8-transfer/

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

Aug 23 2005



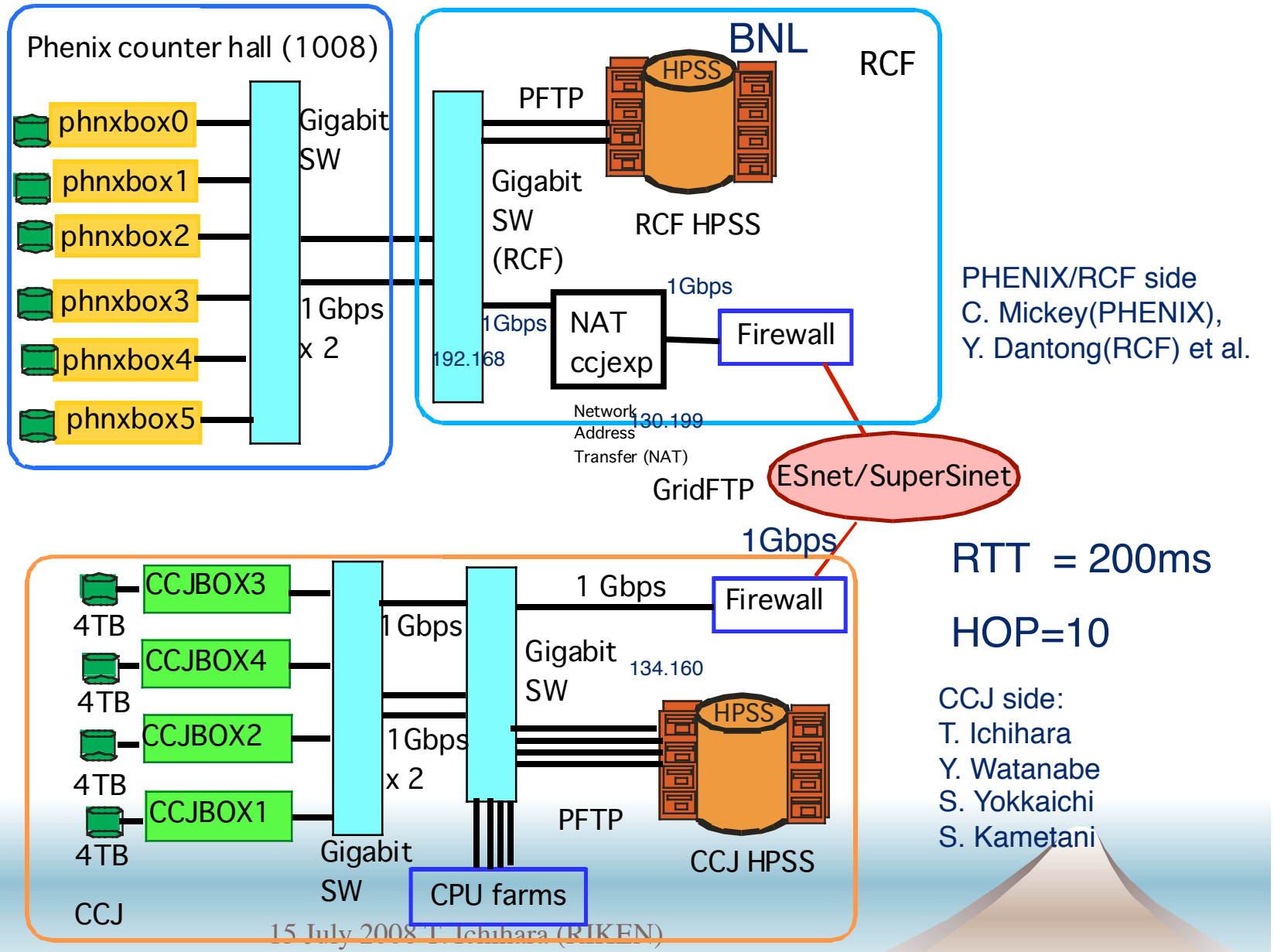
- During the polarized proton-proton run that ended in June at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven, Grid tools were used by the PHENIX experiment to send recently acquired data to a regional computing centre for the experiment in Japan.
- This seems to be the first time that a data transfer of such magnitude was sustained over many weeks in actual production, and was handled as part of routine operation by non-experts.

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

Overview of Data transfer from PHENIX to CCJ (in 2005/2006)

PHENIX

Detector



PHENIX/RCF side
C. Mickey(PHENIX),
Y. Dantong(RCF) et al.

RTT = 200ms

HOP=10

CCJ side:
T. Ichihara
Y. Watanabe
S. Yokkaichi
S. Kametani

RIKEN-BNL GridFTP データ転送マシンの環境(理研側)

- Hardware (ccjbox5-8)
  - 1U Dual-core Opteron 2.6GHz (HP-DL145G3)
  - 4 GB Memory, SAS disk (Soft Raid1), Tigon3 partno(BCM95715) NIC
  - 4 Gbps Dual Fiber–Channel Host Bus Adapter + 13TB SATA RAID6

- Software: Scientific Linux 5.0 (X86_64)
  - File system : **XFS** (data area) , ext3 (OS part)

Grid environment

The Virtual Data Toolkit (v1.8.1)

(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、 grid-proxy-init, globus-url-copy

/etc/sysctl.conf のサンプル
(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

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

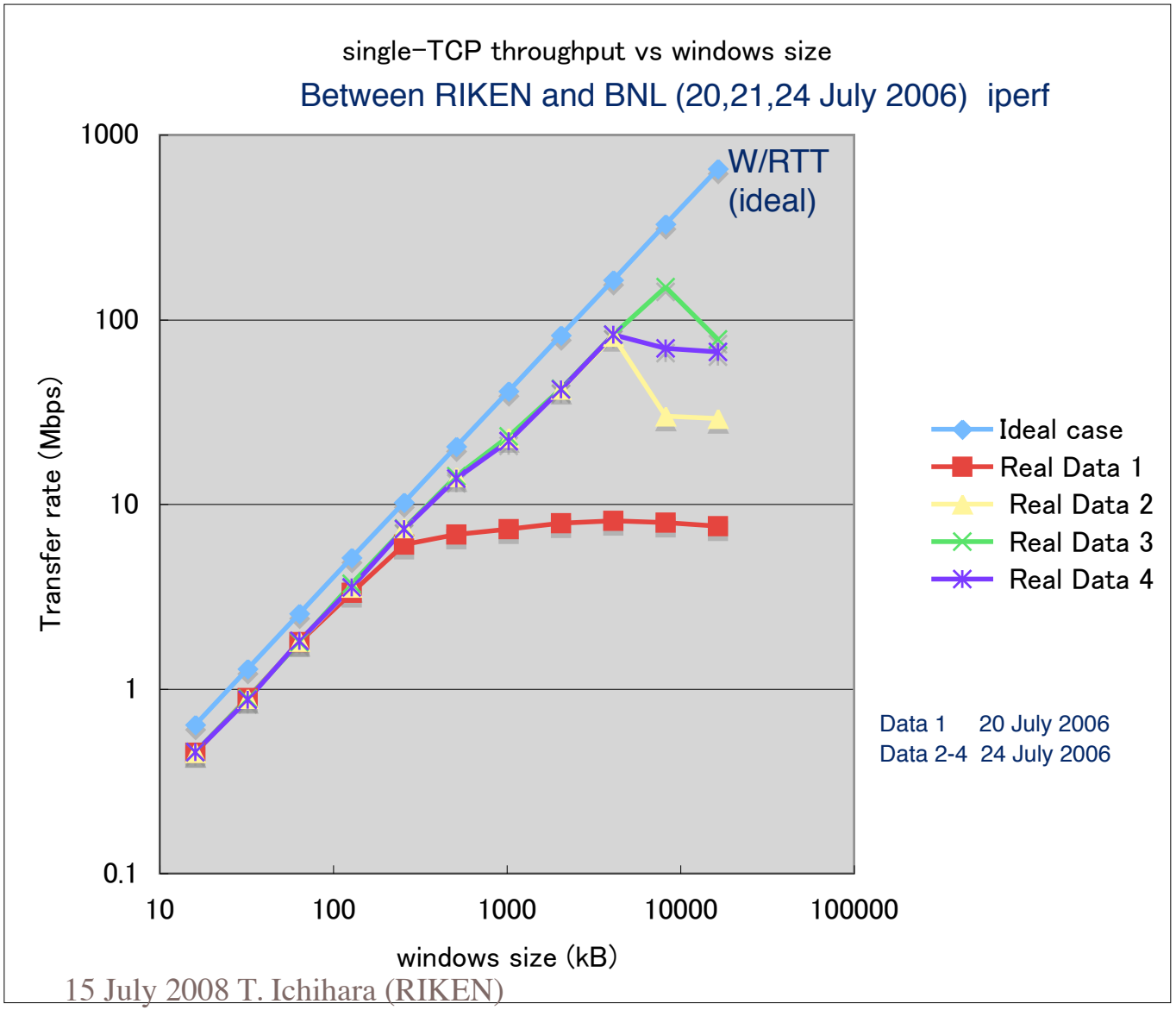
パケットロス、ボトムネックのない理想的な場合

Throughput= WindowSize/RTT

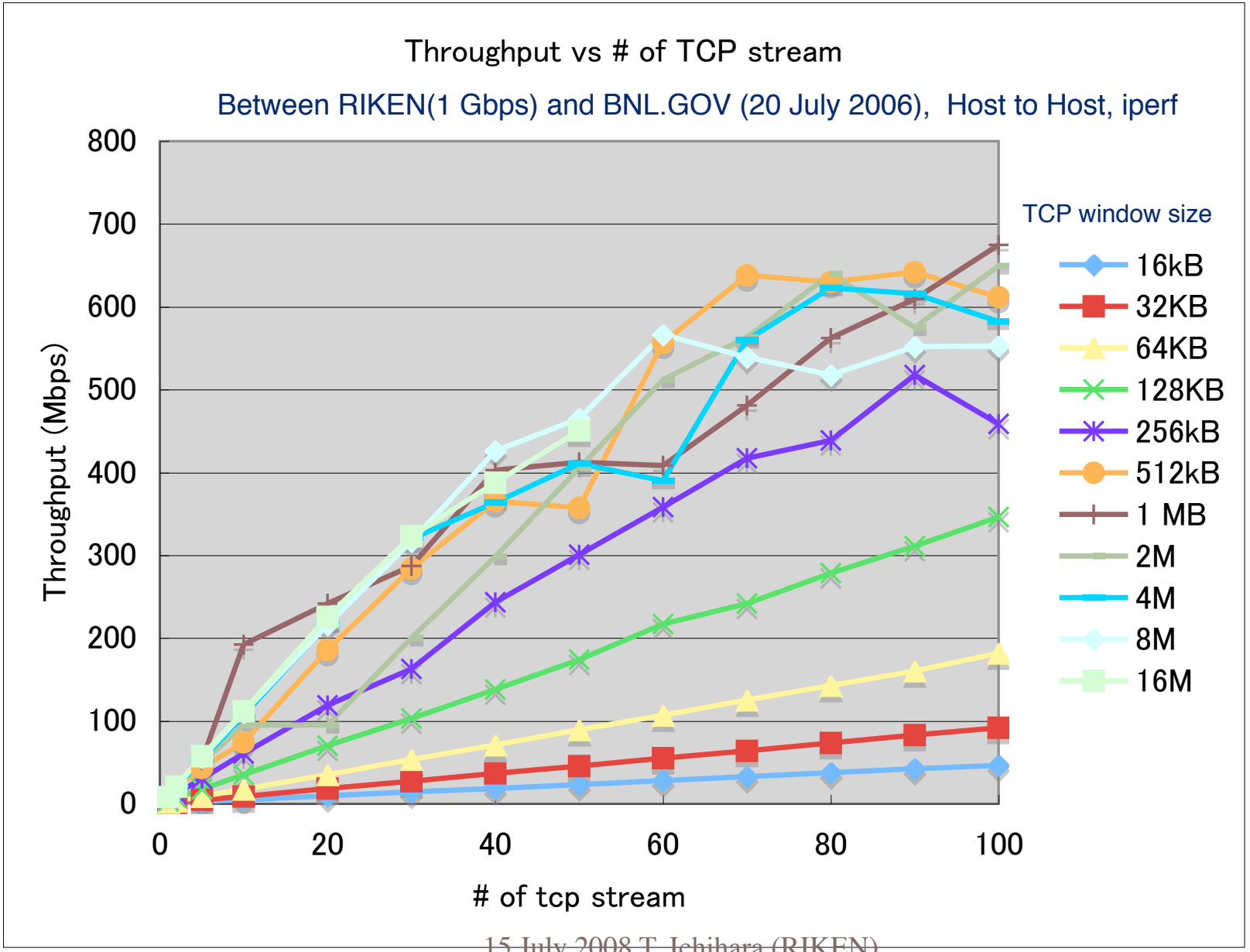
現実のネットワーク (RIKEN-BNL 間)

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



理研側　（2007–2008）の改善

- SINET3接続(2007年1月）**10Gbps**
- CCJ マシン室まで 10 GBpsを引き延ばす (2007.11に完了）
  - **Sinet3(10GBps) → Foundry MLX → Foundry FESX**
  - **No Firewall**
  - Firewall機能（WAN RouterでのACL+各serverでのiptables)
- CCJデータ転送用新Buffer Boxを4台増強（2007.11に完了）
- 理研所内LAN更新(10GExN Backbone LAN) 2009年2月
- スパコン(CPU farm)更新 2009年春

BNL側

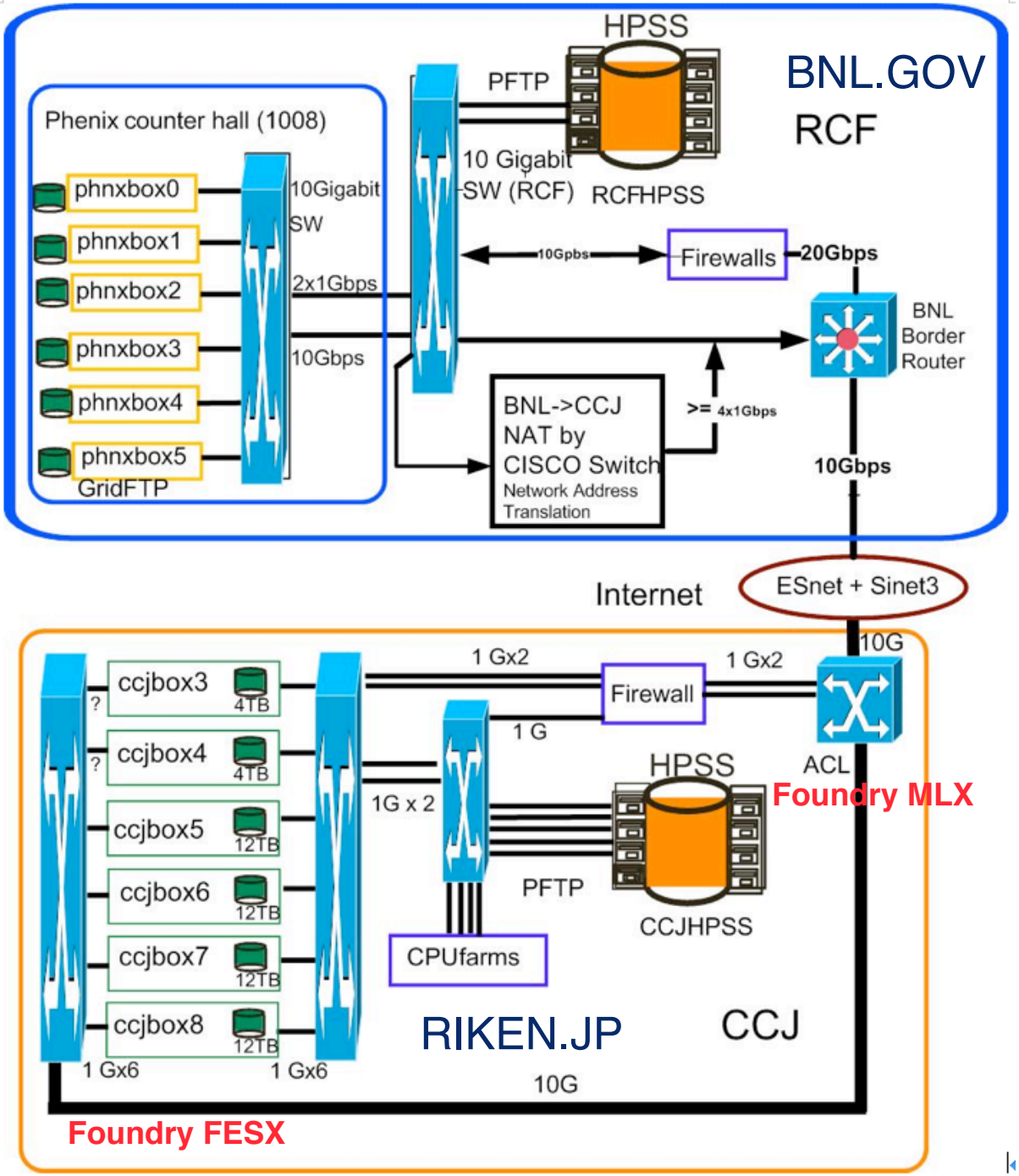
- 2006年に所内LAN更新(Catalyst 6513, 20 GBps Backbone)
  - 20 GBps LAN for Production
    - Cisco Firewall Service Module(FWSM)
      - 5*1Gbps（実際は 最大で2.4Gbps程度）
  - 20 GBps LAN for **LHCOPN** (No Firewall)

BNL-RIKEN PHENIX実験データ転送

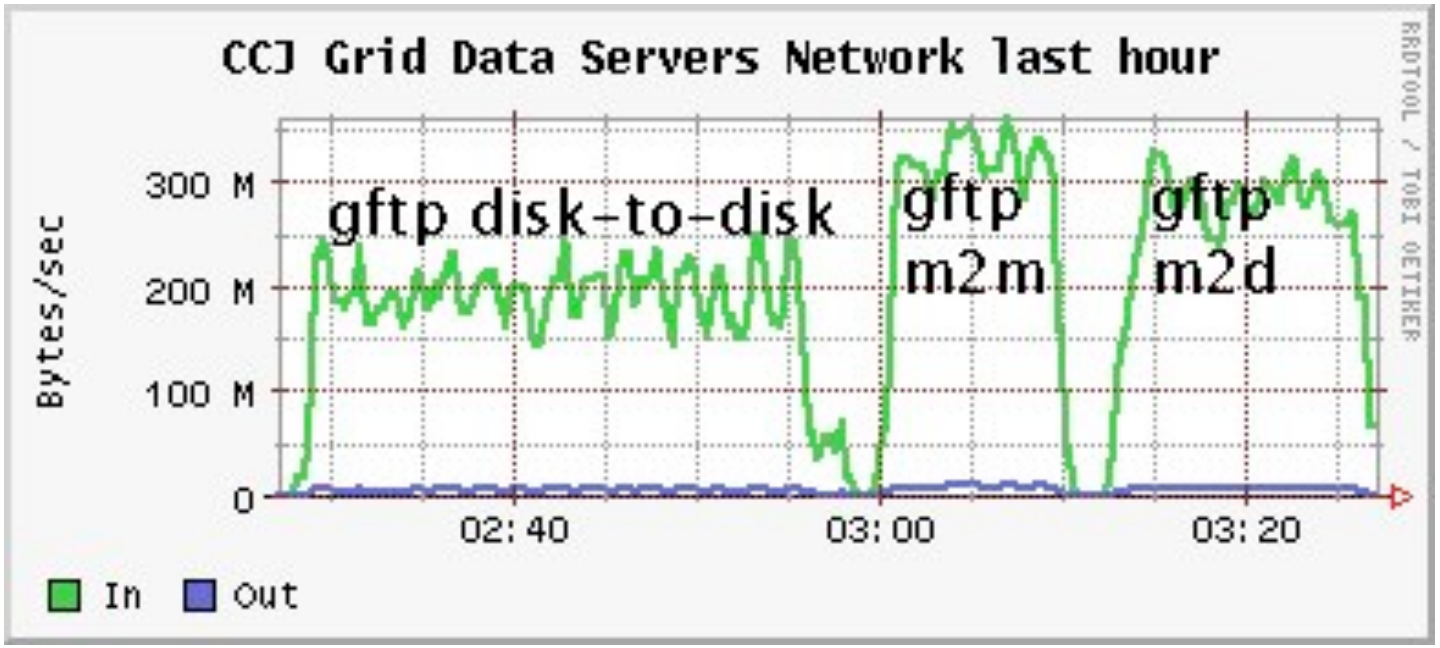
2008年 初冬／春　～2Gbpsを目標設定

2008年の
Configuration

10Gbps WAN
No Firewall



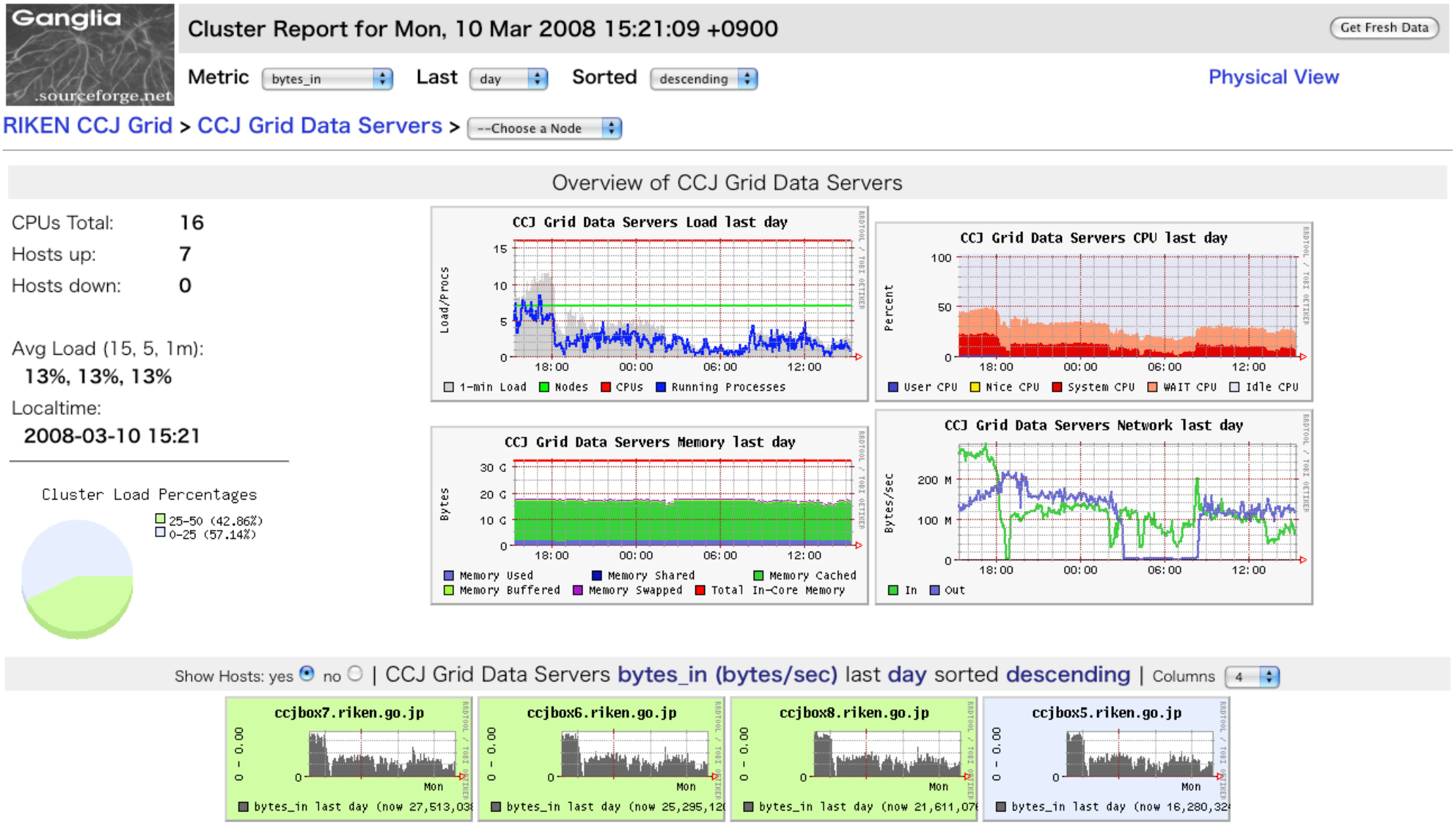
2008年2月 BNL-RIKEN Gridftpのテスト



325MB/s （2.6 Gbps) memory to memory (BNL to RIKEN)
300 MB/s （2.4 Gbps) memory to disk (BNL to RIKEN)
200MB/s （1.6 Gbps) disk to disk (BNL to RIKEN)
[disk of BNL is busy and slow]

4 parallel gridftp transfers from phenix0-4 to ccjbox5-8x

2008年3月10日の 実験中のBNL-RIKEN WAN転送



まとめ

- 理研では BNL PHENIX実験のため、Regional Computing Center (RIKEN CCJ)を2000年より運用開始

- 2005年より日米間でWAN+GridFTPを用いたデータ転送を実施
  2005年（263 TB）、2006年(308 TB)、2008年(100 TB) をWANで転送

- 2008年は日米間で250MB/s sustainedでdisk-to-diskのデータ転送可
  - 10Gbps WAN/LAN（Peak 4 Gbpsをめどに準備）
  - No Firewall (WAN Switch：IP-address baseのACL)

- 今後もWANを用いて日米間でデータ転送の予定(0.3-1 PB/year)
  （毎年2−3ヵ月の実験期間中に,理研−BNL間で 2−4Gbps程度の帯域の利用予定）

- （要望）国際間（特に日米間）および、国内（他のISPを含む）での十分な帯域確保、安定な接続性、障害発生時の迂回路の確保,日米間等の主要なトラフィックの開示