Star Asian Computing Center

Pusan National University
In-Kwon YOO

On Behalf of SACC Team♪

J.Lauret, D.Yu, W. Bett, J. Packard♪

E. Dart, E. Hjort♪

S.D.Lee, D.K.Kim, H.W.Kim♪

ESnet4: Networking for the Future of DOE Science

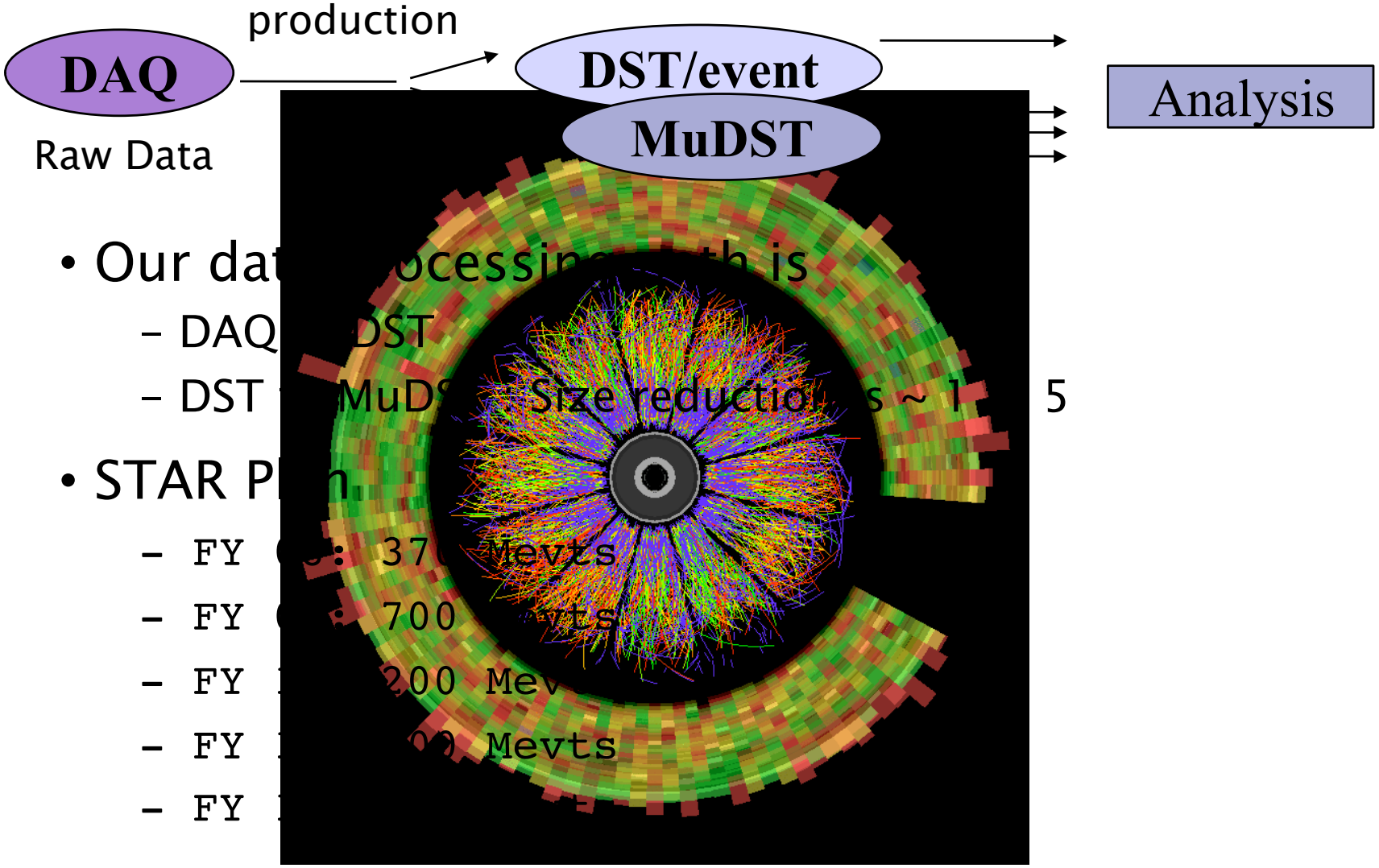
슈퍼컴퓨팅센터

Outline

1. Motivation
   a. STAR Computing Infrastructure
   b. KISTI & Supercomputing Center
2. SACC Project
   a. STAR Asian Hub
   b. SACC Working Group
   c. Computing Resources / Network Research
   d. To-Do List
3. Outlook : Heavy ion Asian Computing Center

JLauret♪

STAR S&C Structure



- Our dat ocessin th is
  - DAQ DST
  - DST MuDS Size reductio s ~ 1 5
- STAR P n
  - FY : 37 Mevts
  - FY : 700 vts
  - FY 200 Mev
  - FY 00 Mevts
  - FY

STAR Computing Sites

JLauret♪

Data transfer should flow from Tier0 to Tier2

Tier 0

- **Tier 1**
  - **Tier2**
    - Would host transient datasets – requires only several 100 GB
    - Mostly for local groups, provide analysis power for specific topics
    - MUST provide cycles (opportunistic) for at least simulation
      - Low requirement of Grid operation support or common project

In-Kwon YOO

STAR Computing Sites

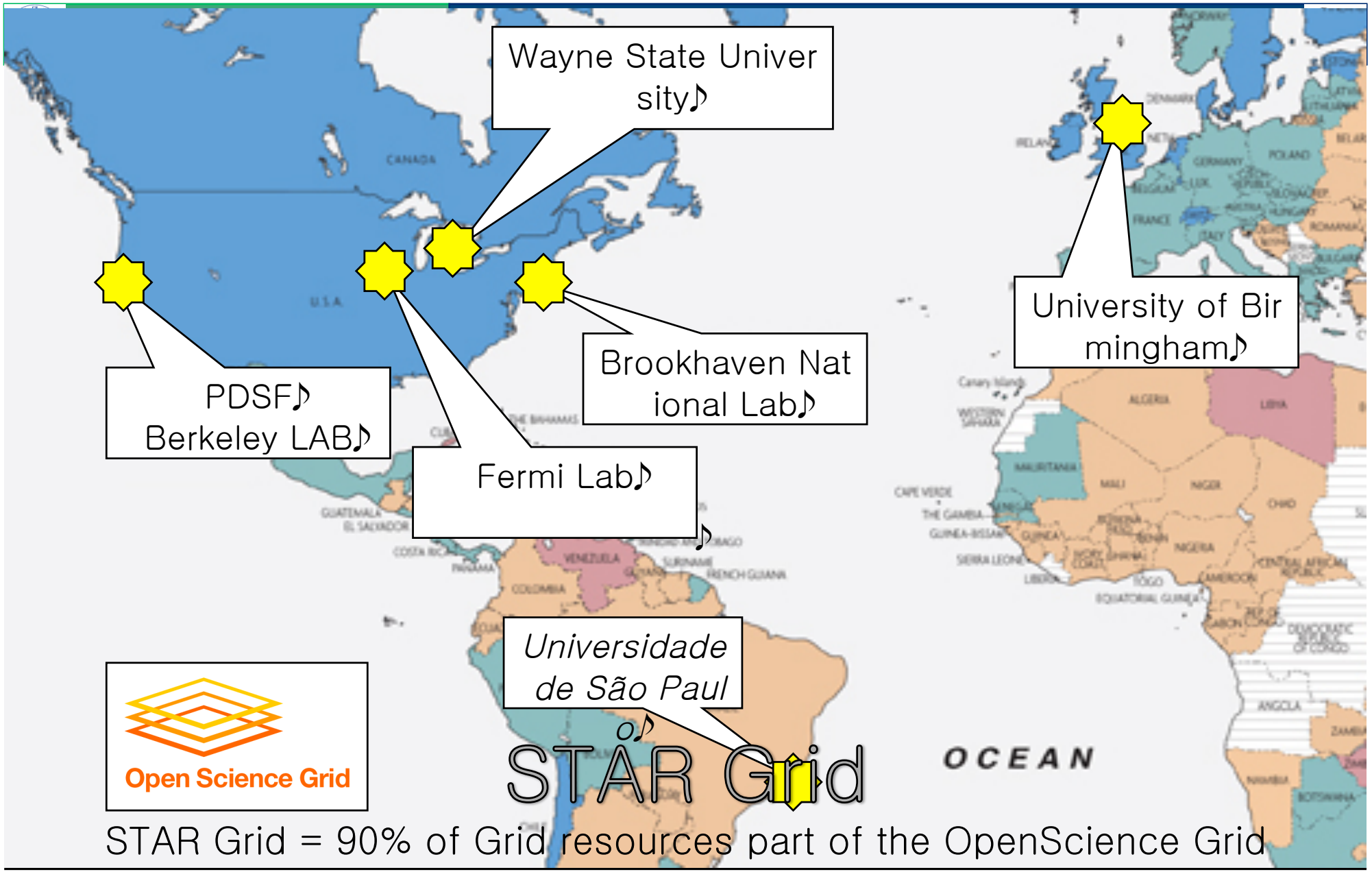
JLauret♪

6 main dedicated sites (STAR software fully installed)

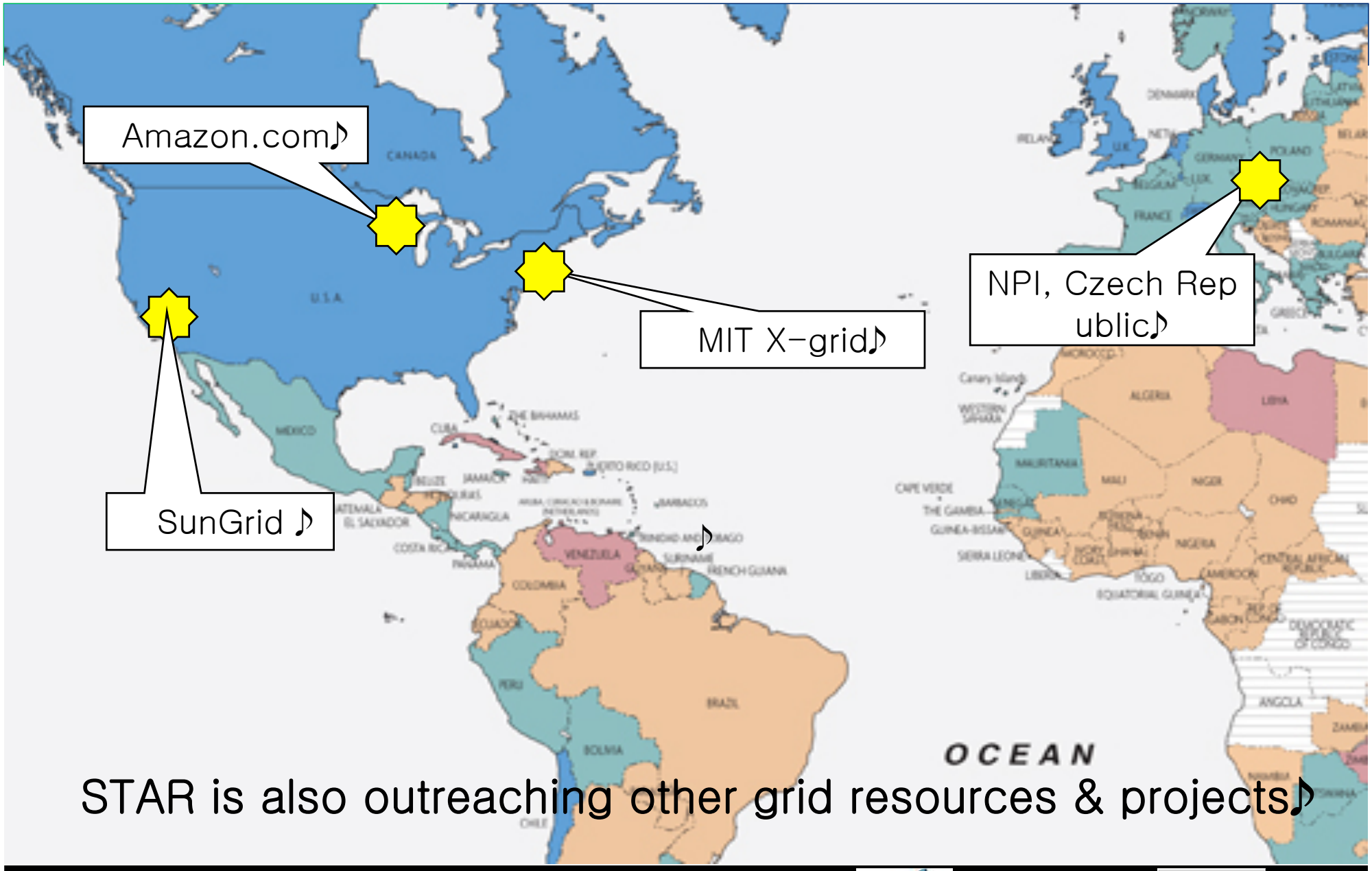
- BNL Tier0
- NERSC/PDSF Tier1
- WSU (Wayne State University) Tier2
- SPU (Sao Paulo U.) Tier2
- BHAM (Birmingham, England) Tier2
- UIC (University of Illinois, Chicago) Tier2

Incoming

- Prague Tier2
- KISTI Tier1



STAR Grid = 90% of Grid resources part of the OpenScience Grid



STAR is also outreaching other grid resources & projects♪

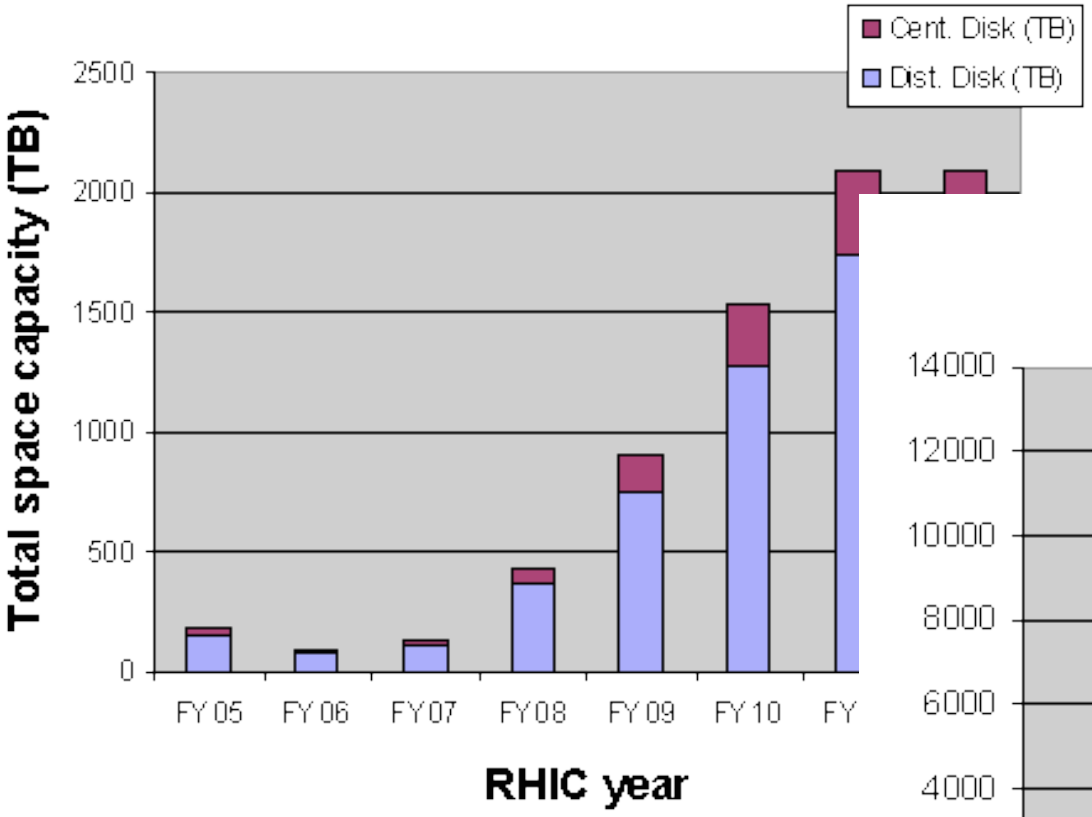
Interoperability / outreach♪ Virtualization♪ VDT extension♪ SRM / DPM / EGEE♪

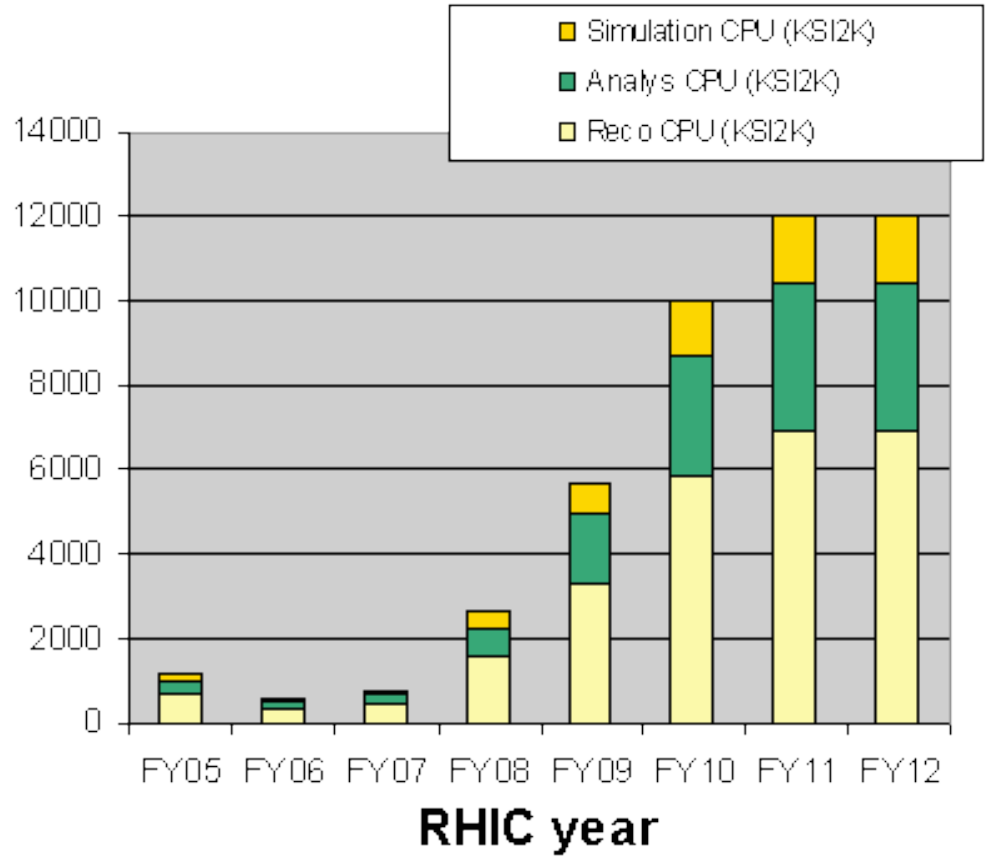
STAR Resource Needs JLauret♪

	FY05	FY06	FY07	FY08	FY09	FY10	FY11	FY12
STAR Requirement								
Real Data Volume (TB)	600	570	560♪	**870**	1720♪	**3000**	4160♪	**4160**
Reco CPU (KSI2K)	660	314	462♪	**1532**	3296♪	**5891**	6960♪	**6960**
Analys CPU (KSI2K)	360	157	210	730	1648	2805	3480	3480
Dist. Disk (TB)	150	71	105	365	749	1275	1740	1740
Cent. Disk (TB)	30	14	21	73	150	255	348	348
Annual Tape Volume (TB)	720	684	672	1044	2064	3600	4992	4992
Tape bandwidth (MB/sec)	200	200	200	200	500	800	1000	1000
WAN bandwidth (Mb/sec)	160	384	528	640	1760	2944	3800	3800
Simulation CPU (KSI2K)	153	71	101	339	742	1304	1566	1566
Simulation Data Volume (TB)	120	114	112	174	344	600	832	832

STAR Resource Needs

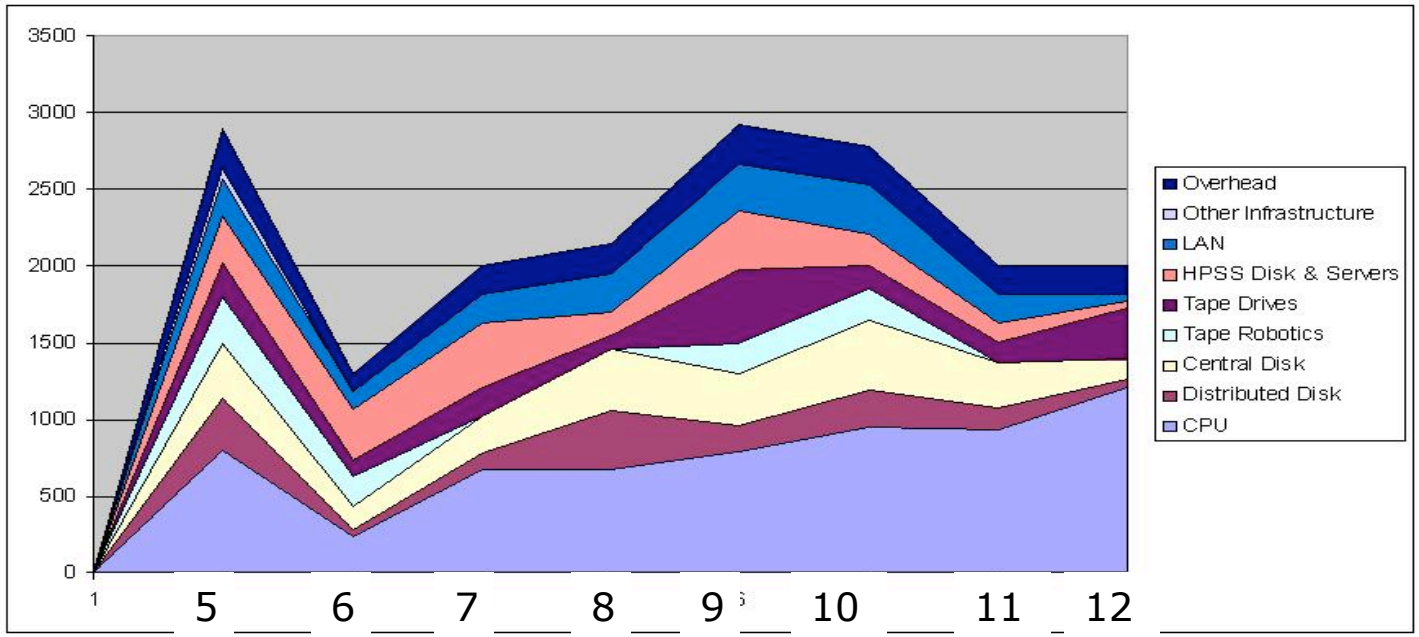
JLauret♪





STAR S&C Cost Analysis

JLauret♪



Observation: Cost seem to go into CPU

- BUT this folds distributed disks (1/2 cost)
- Reduced used of centralized di sk is nonetheless third in cost

Cost is clearly

- Storage (~ ½)
- CPU (1/3rd)
- HPSS & LAN is second

Cluster system

SDLee, HWKim♪

Item	Cluster system	
	Phase 1	Phase 2
Manufacturer & Model	SUN C48	SUN Fusion
Architecture	Cluster	
Processor	AMD Opteron 2GHz (Barcelona)	Intel Xeon 3.3GHz+ (Gainestown)
Operating System	Cent OS	Cent OS
Nodes	188	2,688
CPU cores	3,008 (16/node)	21,504 (8/node)
Rpeak	24TFlops	286TFlops
Memory	6TB	64.5TB
Disk storage	207TB	1PB
Tape storage	422TB	2PB
Interconnection network	Infiniband 4X DDR	Infiniband 4X DDR
Cooling	Chilled water cooling	Chilled water cooling
Delivery date	Jan, 2008	2Q, 2009

SMP system

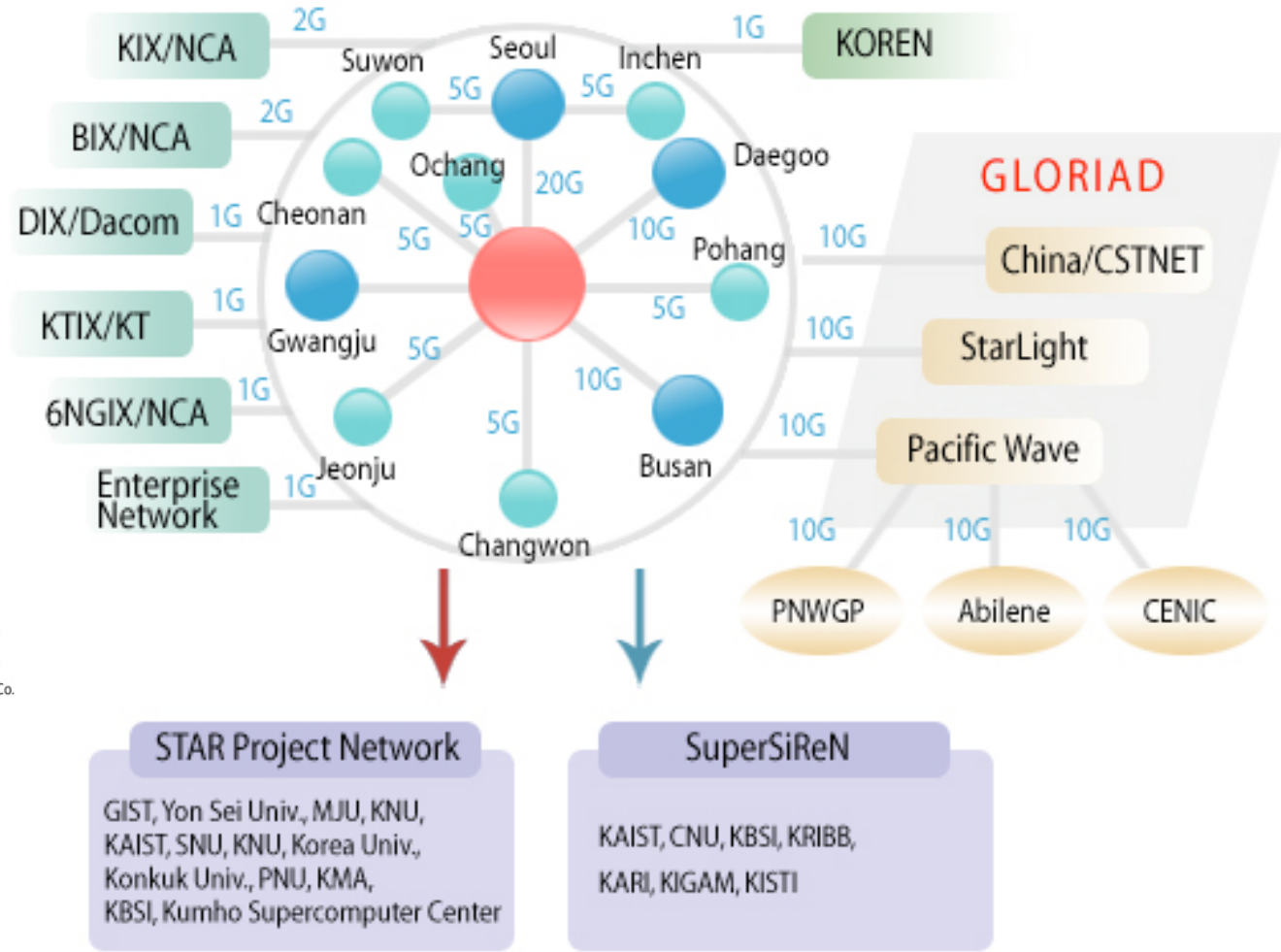
SDLee, HWKim♪

Item	SMP system	
	Phase 1	Phase 2
Manufacturer & Model	IBM p595	IBM p6H
Architecture	SMP	
Processor	POWER5+ 2.3GHz	POWER6 5GHz+
Operating system	AIX 5.3	AIX 5.3+
Nodes	10	24
CPU cores	640 (64/node)	1,536 (64/node)
Rpeak	5.9TFlops	30.7TFlops
Memory	2.6TB	9.2TB
Disk storage	63TB	273TB
Tape storage	-	
Interconnection network	HPS	Infiniband 4X DDR
Cooling	Air-cooling	Air-cooling
Delivery date	Sept, 2007	1Q, 2009

Research Networks

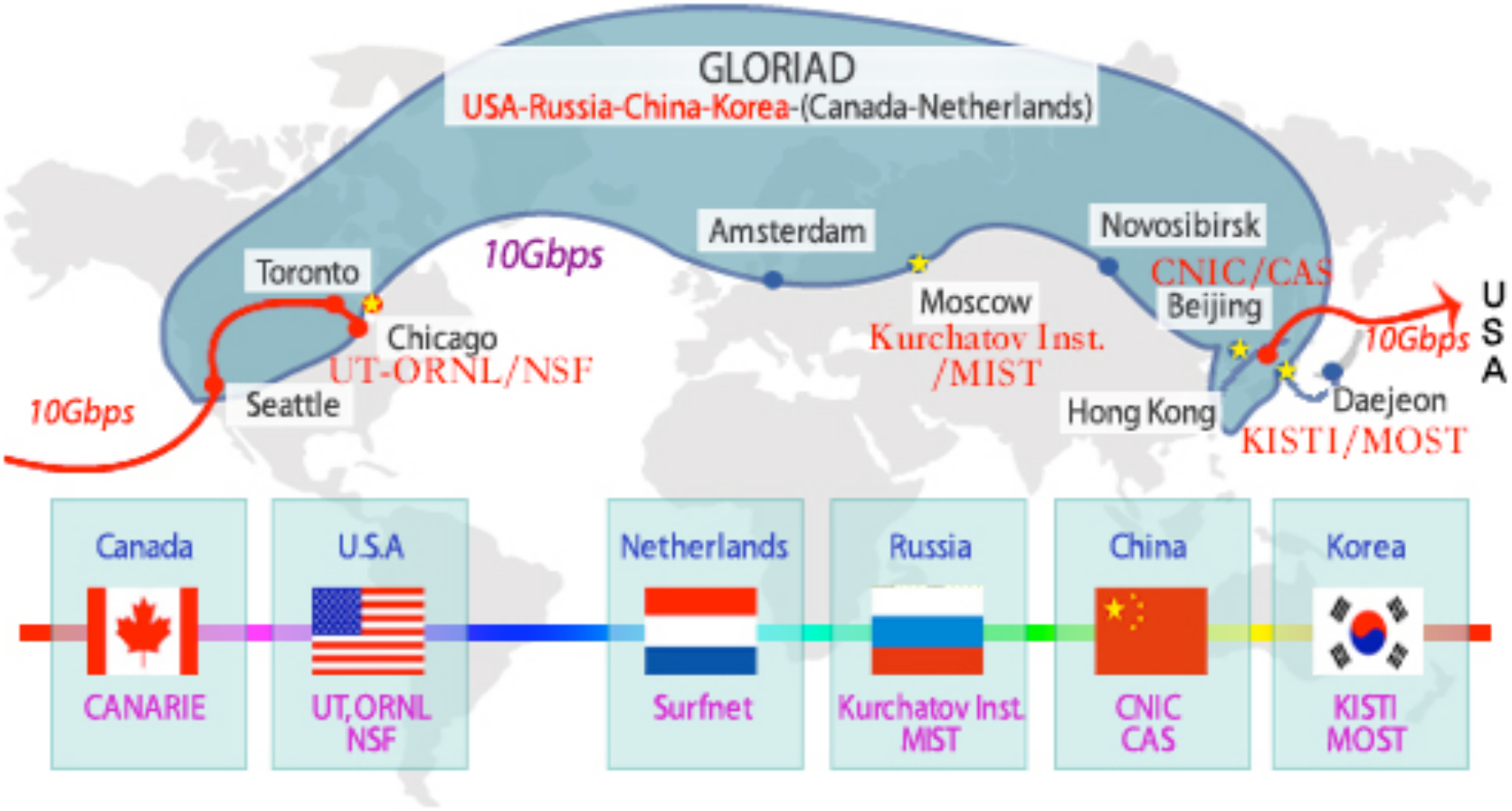
SDLee, HWKim♪

- KREONET



GLORIAD

SDLee, HWKim♪



STAR Asian Hub

Star Asian Computing Center

- Computing Infrastructure with massive data from STAR
  - Frontier Research
  - Maximum Use of IT resources in Korea
    - Data Transfer
    - Cluster Computing with Supercomputer
    - Mass Storage
- Korean Institute for Science and Technology Informations (KISTI @ Daejoen)
  - Korean HUB for GLORIAD + KREONET
  - Super Computing Resources
  - Mass Storage Management

→ Asian Supercomputing HUB :

- BNL – NERSC – KISTI – SSC etc.

SACC Working Group

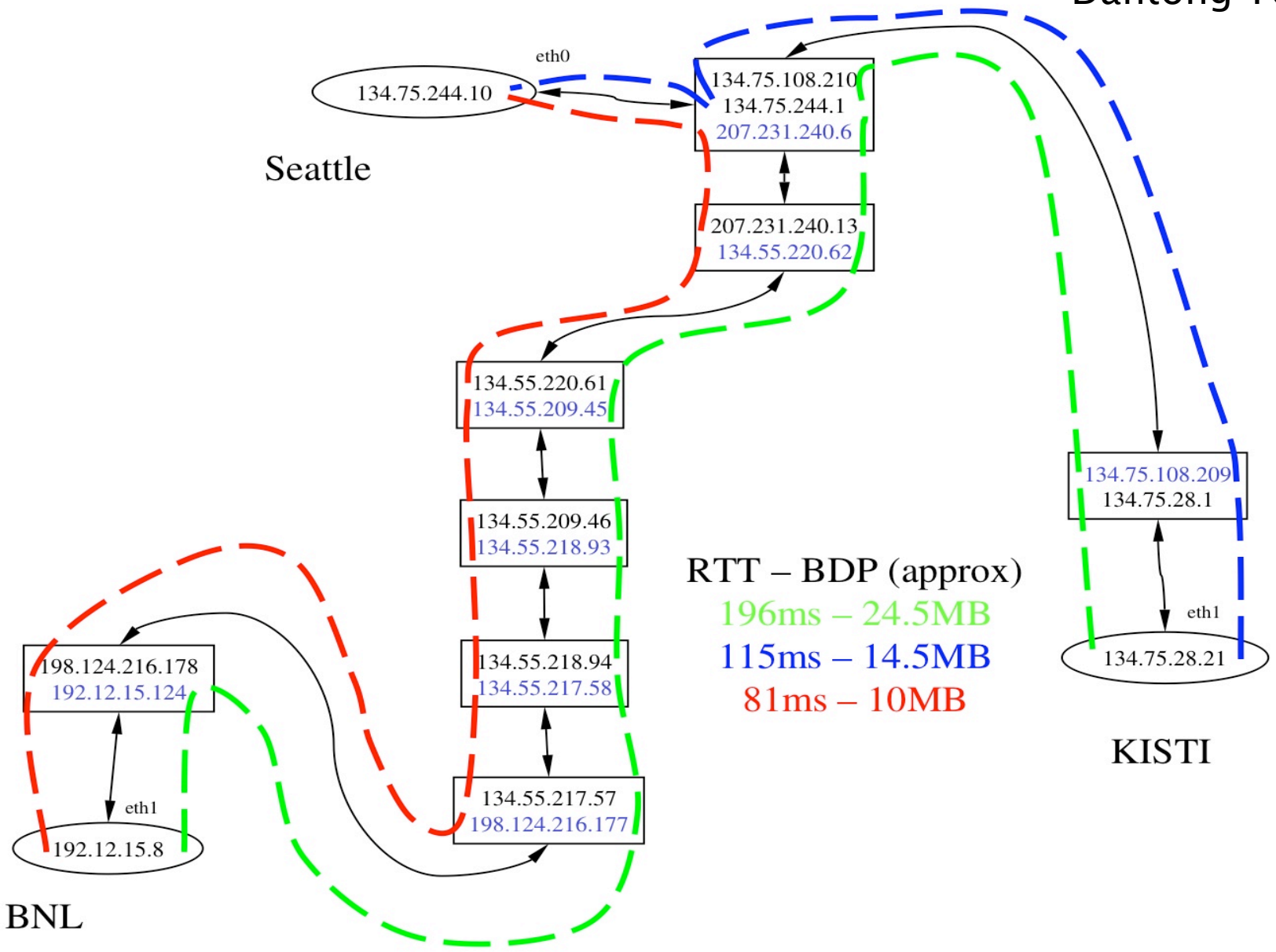
- PNU
  - IKYoo et al.
- KISTI
  - SDLee, DKKim, HWKim
- BNL (STAR)
  - JLauret, DYu, Wbett, Edart, JPackard
- SSC + Tsinghua Univ. ?
  - ZXiao et al. ?

SDLee, HWKim♪

KISTI STAR Computing

- Configuration of a testbed(16 cores) under way
- Eventually SUN cluster 1st shipment (~3,000 cores) will be dedicated to STAR early next year!
- Initial Network status : below 1 Mbps (over 10Gbps line)
- Network Optimization between KISTI and BNL : since 2008-07
- Target Throughput : over 2 Gbps (over LP)
- KISTI's effort
  - Installed 10Gbps NIC equipped server in Seattle
  - Local optimization

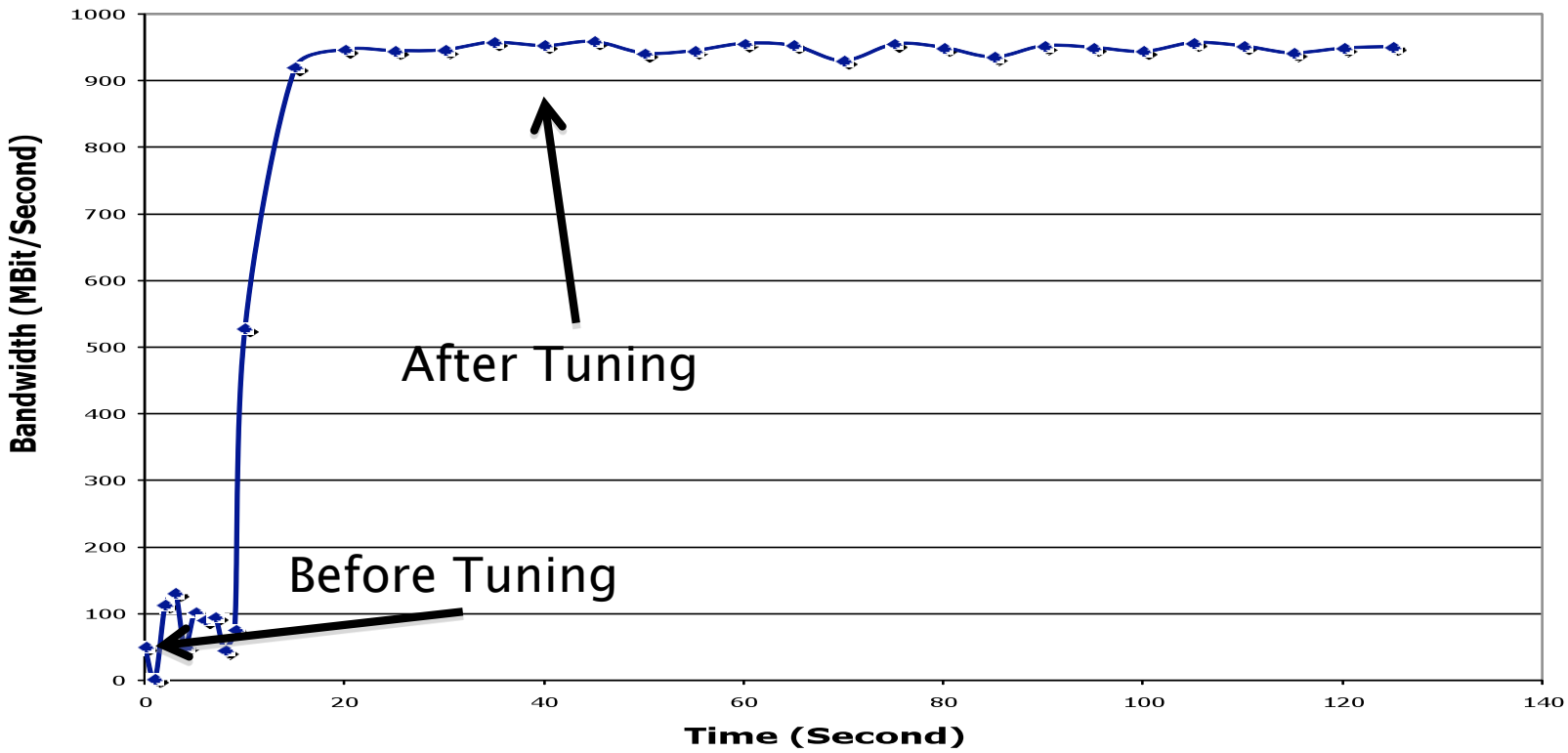
Dantong YU♪


eth0
134.75.244.10
Seattle
134.75.108.210
134.75.244.1
207.231.240.6
207.231.240.13
134.55.220.62
134.55.220.61
134.55.209.45
134.55.209.46
134.55.218.93
134.55.218.94
134.55.217.58
134.55.217.57
198.124.216.177
198.124.216.178
192.12.15.124
eth1
192.12.15.8
BNL
134.75.108.209
134.75.28.1
eth1
134.75.28.21
KISTI
RTT – BDP (approx)
196ms – 24.5MB
115ms – 14.5MB
81ms – 10MB

BNL to KISTI Data Transfer

Dantong YU♪

- Network Tuning improved transfer from BNL to KISTI
  - Identified bottleneck with Kreonet2 peering point with Esnet. 1Gpbs => 10Gbps
  - Network and TCP stack was tuned at KISTI hosts.



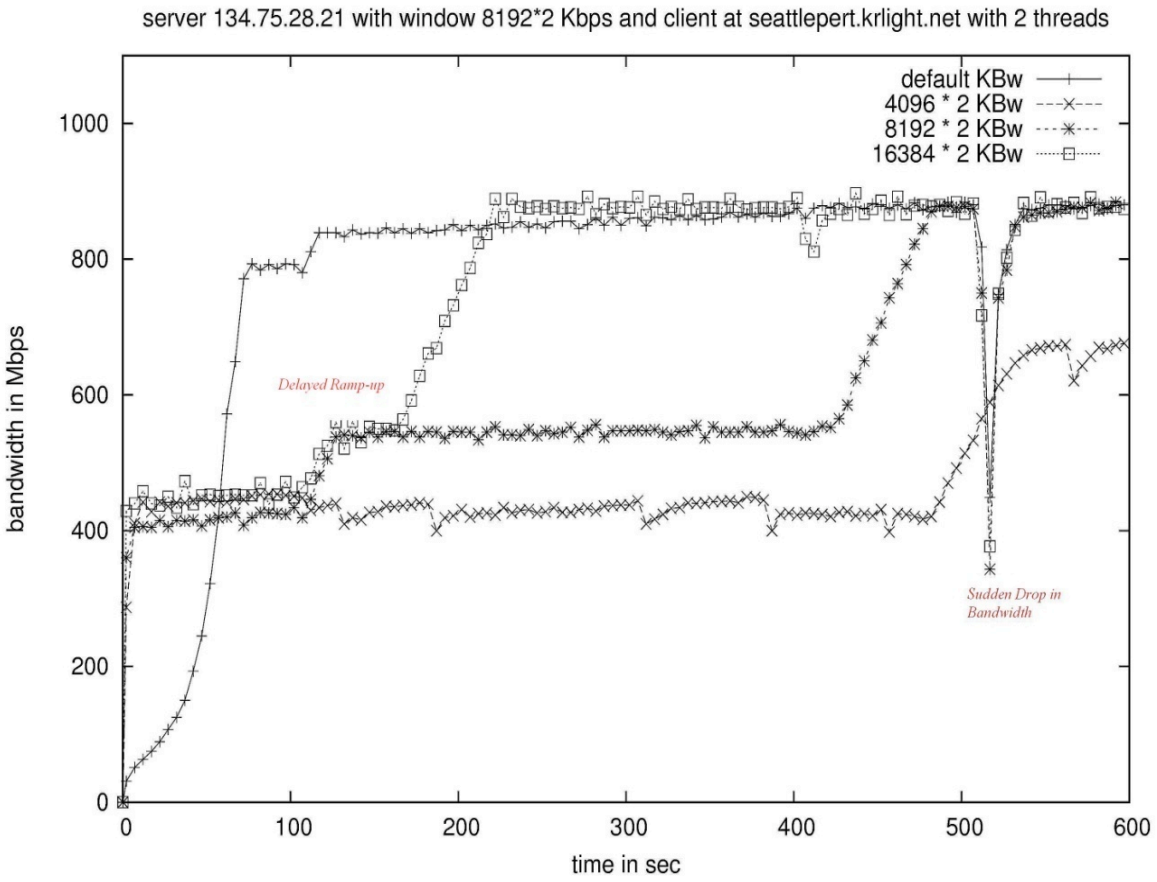
Dantong YU♪

Network Research

There are two kinds of network events that r equire some further e xamination.

1.Delayed ramp up.

2.Sudden Drop in Ban dwidth.



STAR Data Transfer Status

Dantong YU♪

- Performance between BNL and KISTI : not symmetrical.
- A bottleneck from KISTI back to BNL.
  - Packet drops at BNL receiving host. (will be replaced).
  - Old Data Transfer Nodes at BNL: being replaced
  - Findings being corrected:
    - High performance TCP parameters
  - Findings are still under investigation
    - TCP slow ramp up, and performance sudden drop.
- test/tune GridFtp tools : in next 4 weeks

STAR Data Transfer Plan

Dantong YU♪

- Replace the old data transfer nodes
  - 1 Gbps per node, with expansion slots for 10Gbps.
  - Large local disk for intermediate cache.
- Deploy OSG BestMan for these nodes.
- RACF firewall will be rearchitectured.
- Data transfer performance should be only limited by the local disk buffer at both ends.

To do list

- KISTI needs to finish the testbed preparation
- STAR Software should be installed and tested
- KISTI net people need to set up lightpath between KISTI and BNL eventually
- BNL net people need to set up a host for the end-to-end test and measure the throughput
- We need to complete “a real mass data-transfer” from BNL to KISTI sooner or later!
- Start Production test at KISTI

Outlook towards HACC

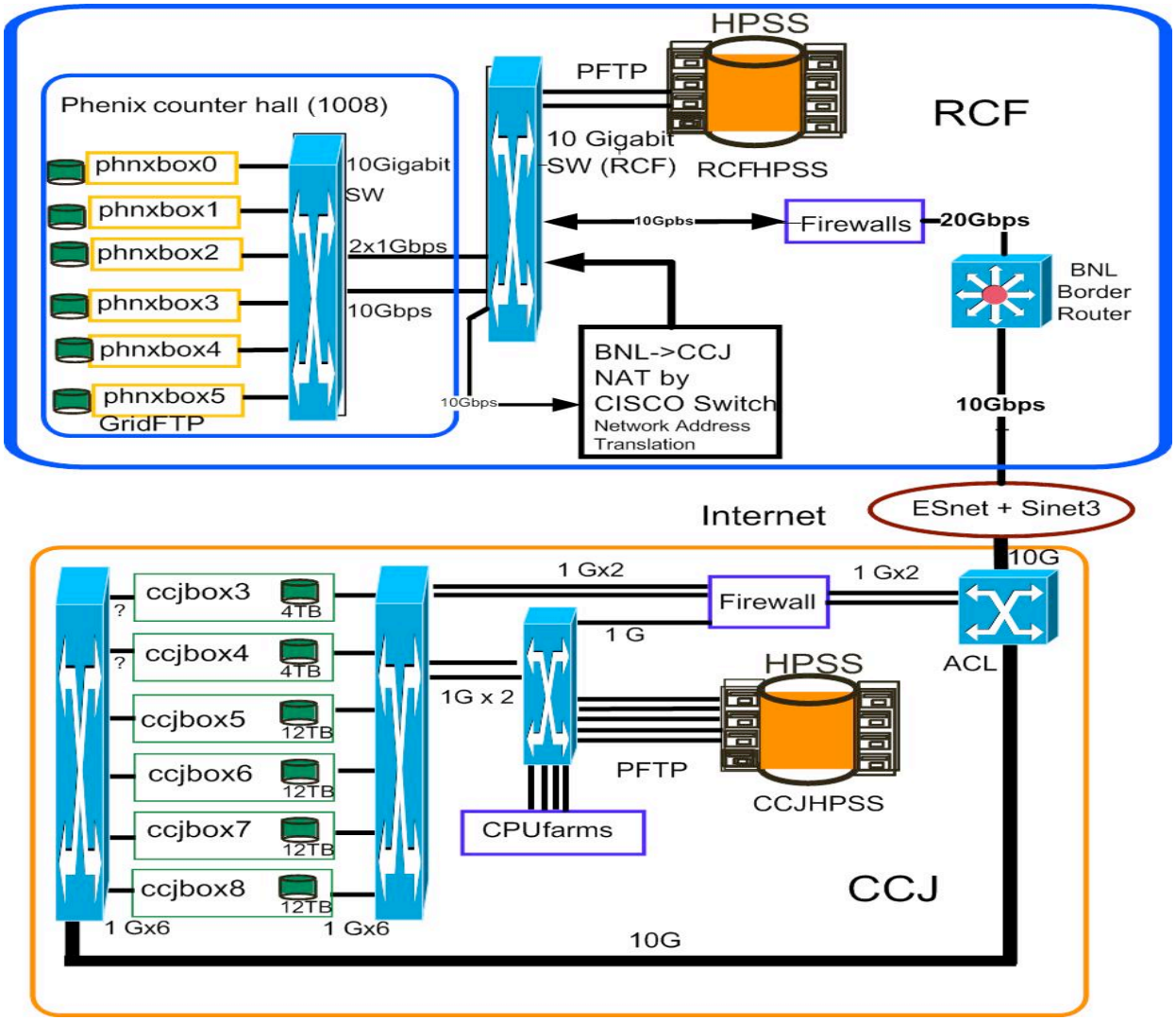
- STAR Asian Computing Center (SACC) (2008 – 2011)
  - Experimental Data from International Facility
    - Computational Infrastructure for LHC / Galaxity
  - Asian Hub for International Coworking
  - Frontier Research
- Heavy ion Analysis Computing Center (HACC) (2011–)
  - Extend to Other project (HIM) ?
  - Extend to ATHIC ?
  - Dedicated Resources for Heavy ion Analysis Computing

Dantong YU♪

BNL PHENIX WAN Data Transfer

Dantong YU♪

PHENIX Data Transfer Infrastructure

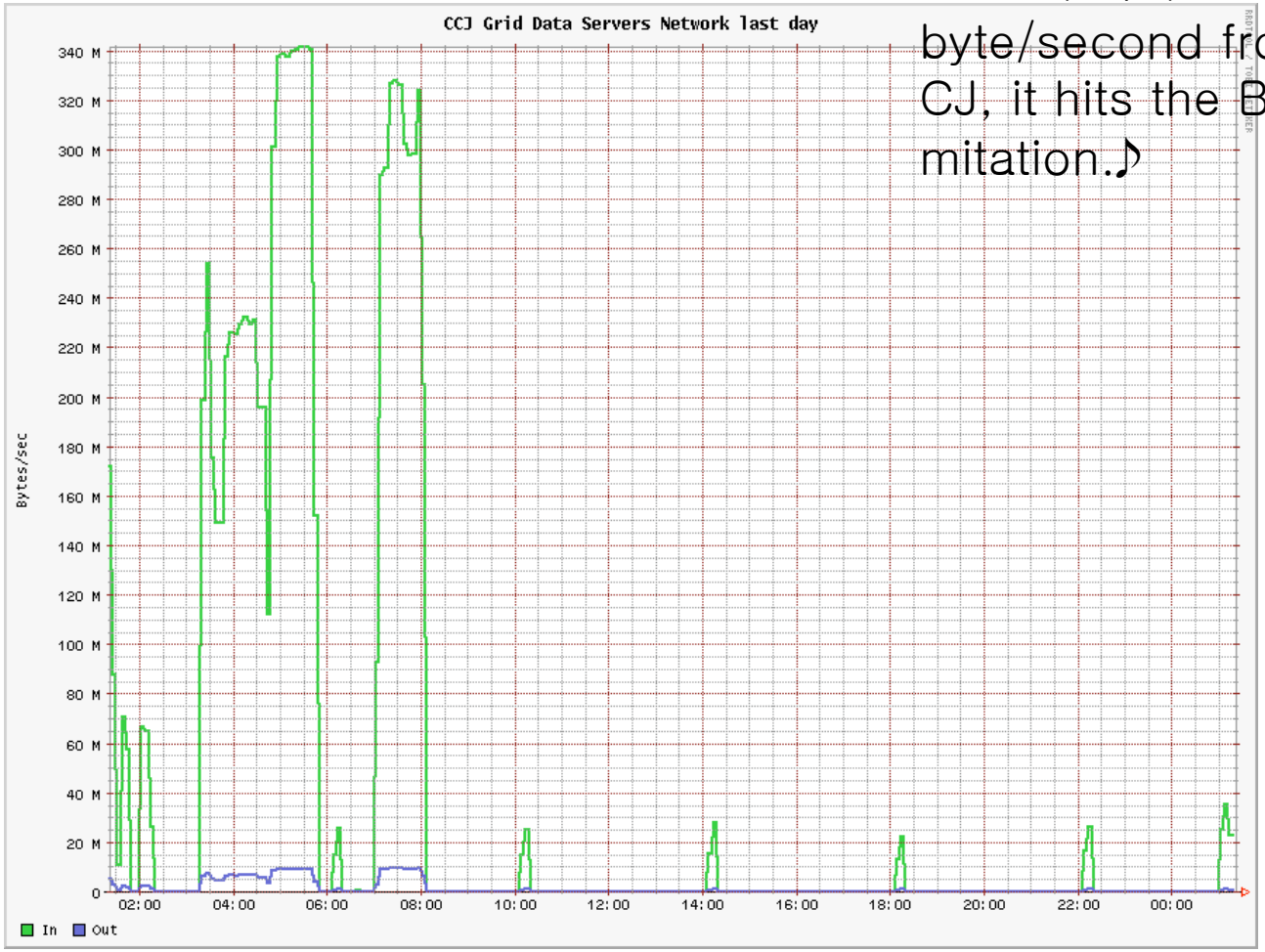


Computer Platforms on Both Ends

Dantong YU♪

- BNL: Multiple 3.0Ghz dual CPU nodes with Intel copper gigabit network. Local drives connected by Raid Controller. There are PHENIX on-line hosts.
- CCJ Site: 8 dual-core AMD Opteron based hosts, each with multiple Tera bytes SATA drives connected with a RAID controller. Each one has one gigabit broadcom network card.
- The LAN on both ends are 10Gbps.
- The data transfer tool is GridFtp.

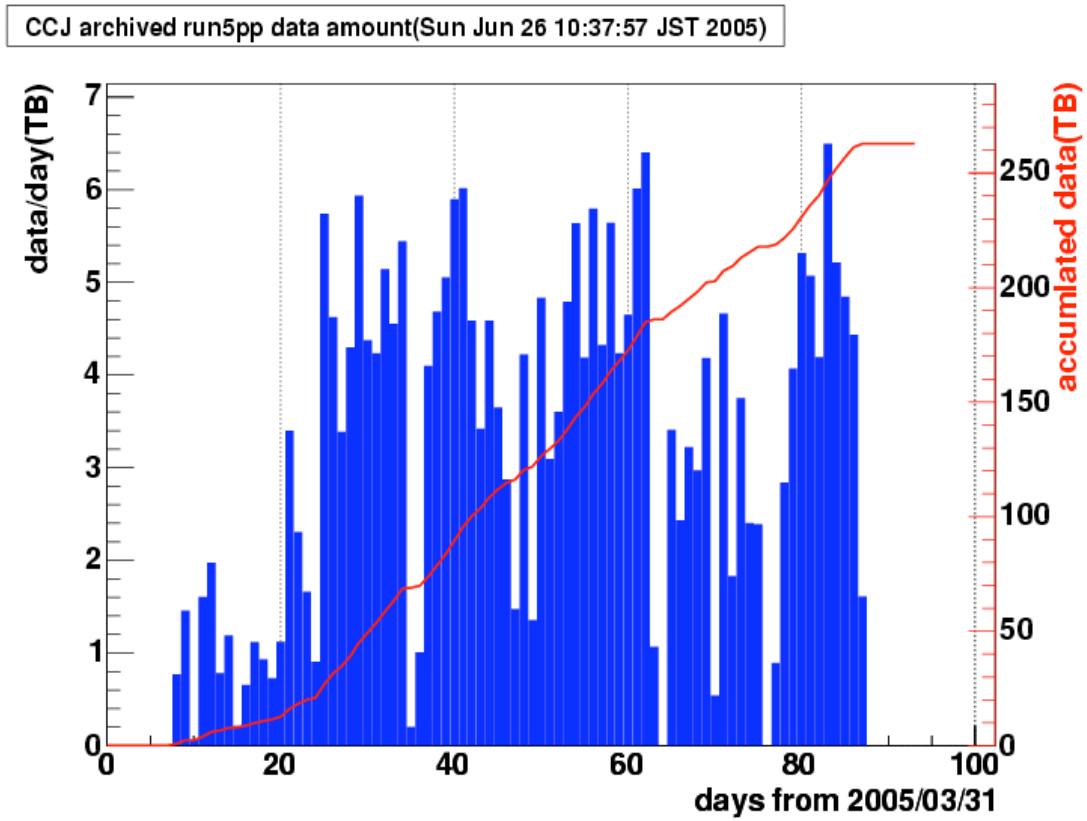
PHENIX to CCJ Data Transfer Test at the beginning of 2008 (Mega Byte/second)



We can transfer up to 340M byte/second from BNL to CCJ, it hits the BNL firewall limitation.♪

Data Transfer to CCJ, 2005

Dantong YU♪



Courtesy of Y. Watanabe

✳2005 RHIC run ended on June 24, Above shows the last day of RHIC Run.

✳Total data transfer to CCJ (Computer Center in Japan) is 260 TB (polarized p+p raw data)

✳100% data transferred via WAN, Tool used here: GridFtp. No 747 involved.

✳Average Data Rate: 60~90MB/second, Peak Performance: 100 Mbytes/second recorded in Ganglia Plot! About 5TB/day!

Data Transfer to CCJ, 2006

Dantong YU♪