RHIC Computing Facility

Bruce G. Gibbard

DOE Nuclear Physics S&T Review

BNL

6-8 July 2005

RHIC Computing Facility (RCF)

- Currently co-located and co-operated with ATLAS Computing Facility (ACF) at BNL

  "US ATLAS Tier 1 Computing Center"

- RCF Mission
  - Online recording of Raw data
  - Production reconstruction of Raw data
  - Primary facility for data selection (mining) and analysis
  - Long term archiving and serving of all data
- RCF Scale
  - Authorized staff of 20 FTE's
  - ~$2M/year equipment ***refresh*** funding (20-25% annual replacement)
    - Addressing obsolescence
    - Resulting in important collateral capacity growth

Experiment Input to RCF

- Weekly Liaison (Operations) Meeting
  - Review recent performance and problems
  - Plan for near term operations
- Annual Series of Meetings RCF/Experiments to Develop Capital Spending Plan
  - Revise estimate of needs for coming run
  - Decide on details of equipment procurement
  - Last such meeting in Fall '04
- Periodic Topical Meetings
  - Example: Linux farm OS upgrade discussions currently underway

Computing Requirements

- Last Comprehensive Long Range Estimate Done in Series of Meetings in Spring '04
  - Produced document ***"RHIC Computing Facility" (5/19/04)***
  - Conclusion consistent with ***"Twenty-Year Planning Study for RHIC Heavy Ion Collider Facility" (BNL -71881-2003)***
- Annual Updates
  - Last Fall
  - Later this Summer
- In longer term, capacity growth associated with equipment ***refresh*** should meet increased capacity requirements but a near term shortfall exists

Summary of Projected Requirements

Summary of Projected Requirements					
				May-04	
	FY '04	FY '05	FY '06	FY '07	FY '08
CPU (kSI2K)	1,958	3,212	4,413	4,413	5,352
Disk (TB)	333	677	1,053	1,053	1,201
Tape (PB)	2.2	4.5	7.8	9.7	13.1

Estimate of Achievable Capacities with Refresh Funding					
				May-04	
	FY '04	FY '05	FY '06	FY '07	FY '08
CPU (kSI2K)	1,200	1,882	2,795	4,810	7,156
Disk (TB)	220	406	665	1,119	1,702
Tape (PB)	4.5	5.6	11.2	11.2	22.5

Modified Requirements Tables					
				Nov-04 updates	
	FY '04	FY '05	FY '06	FY '07	FY '08
CPU (kSI2K)	1,250	2,450	4,413	4,413	5,352
Disk (TB)	280	507	1,053	1,053	1,201
Tape (PB)	1.5	3.1	7.8	9.7	13.1
				Jun-05 updates	
	FY '04	FY '05	FY '06	FY '07	FY '08
CPU (kSI2K)	1,250	2,700	4,413	4,413	5,352
Disk (TB)	280	730	1,053	1,053	1,201
Tape (PB)	1.5	4.3	7.8	9.7	13.1

Actual installed capacity

Funding constrained request

Principal RCF Services

- General Collaboration & User Support
- Compute Services (Linux Farms)
  - Programmatic production processing
  - Individual analysis
- Online Storage (Disk)
  - Working data storage (Read / Write)
  - Data serving (mostly Read)
- Mass Storage (HSM Robotic Tape System)
  - Raw data recording and archiving
  - Derived data archiving
- Grid & Network Services

General Collaboration & User Support

- Collaboration/User Working Environment
  - User accounts, Email, etc.
  - Software Development
  - AFS Domain
  - Databases (MySQL, Oracle)
  - Job submission & resource management
- User interaction
  - Web site – plans for a needed rework
  - Trouble ticket system – actually 3 instances for …
    - RHIC services (~1500 tickets/year)
    - ATLAS services (~400 tickets/year)
    - Grid services (very new)

Mass Storage

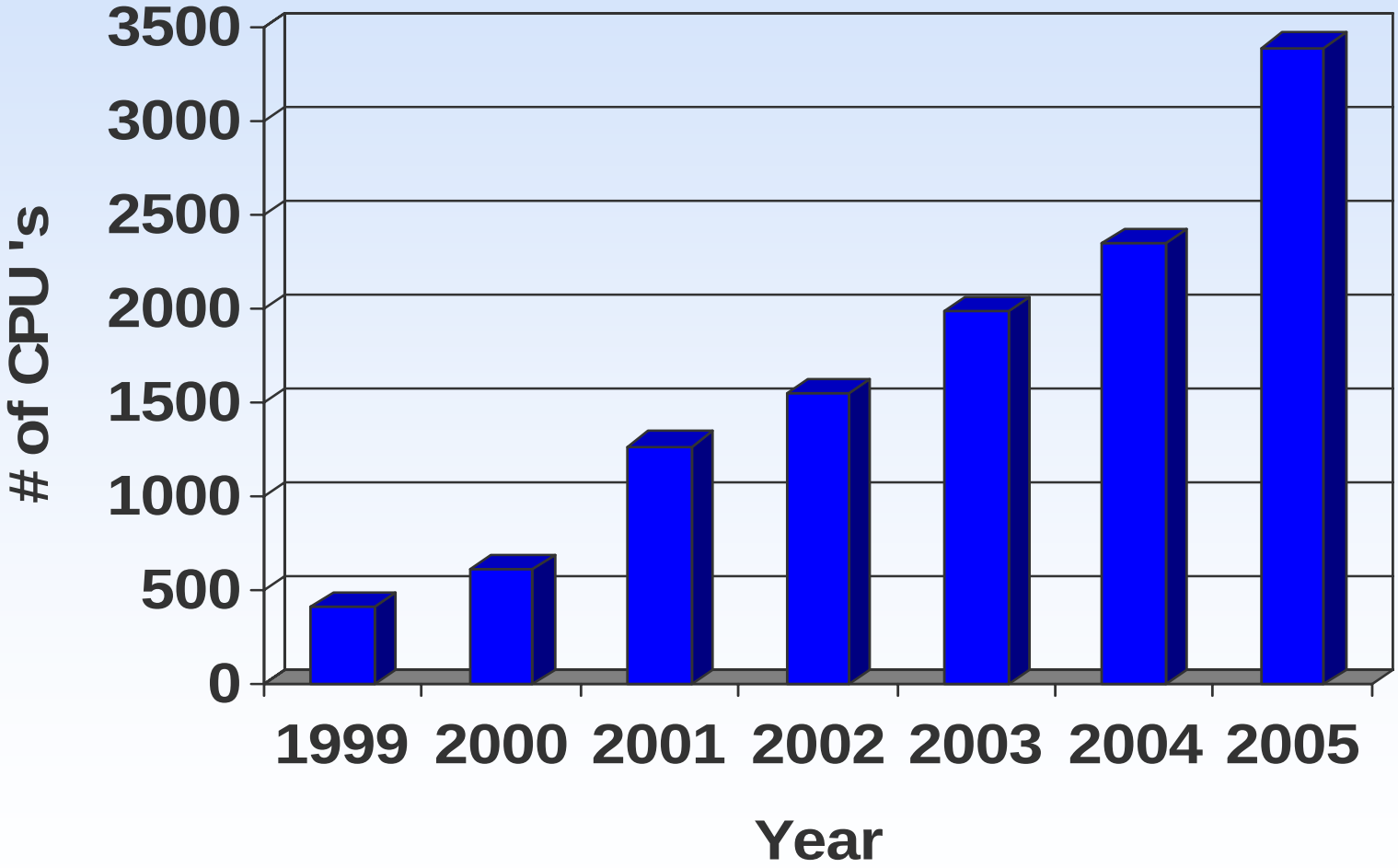
Compute Farm

Online Storage

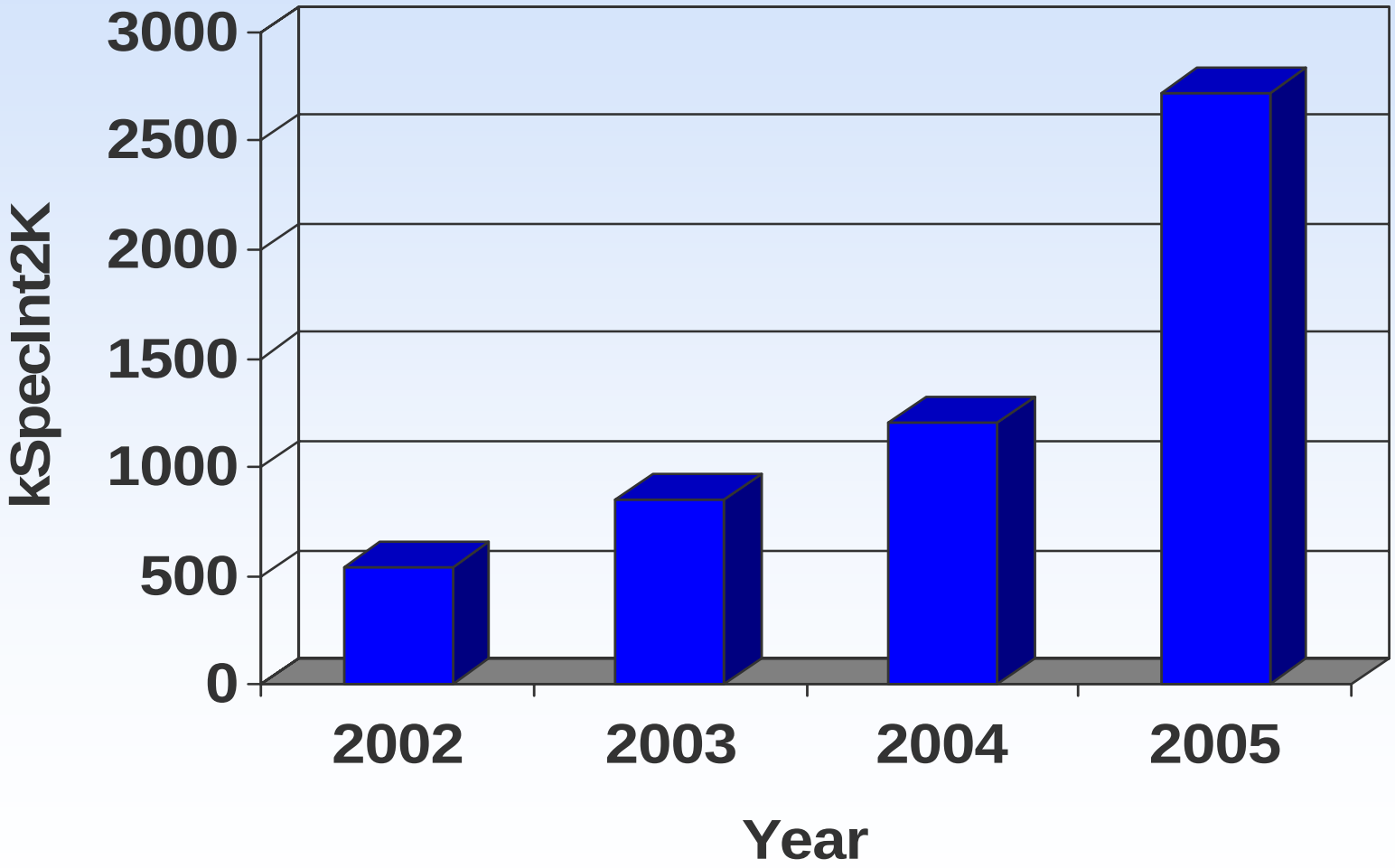
Compute Services

- Intel/Linux dual CPU rack mounted systems
  - 3,400 CPU's for 2,700 kSI2K
  - Condor & LSF resource managers (batch systems)
  - In house developed Condor based resource manager affords optimizing interface to tape HSM for production system
- Expect to satisfy future requirements (beyond 2006) by continuing to follow Moore's law price/performance in commodity market

Linux Farm Processor Count

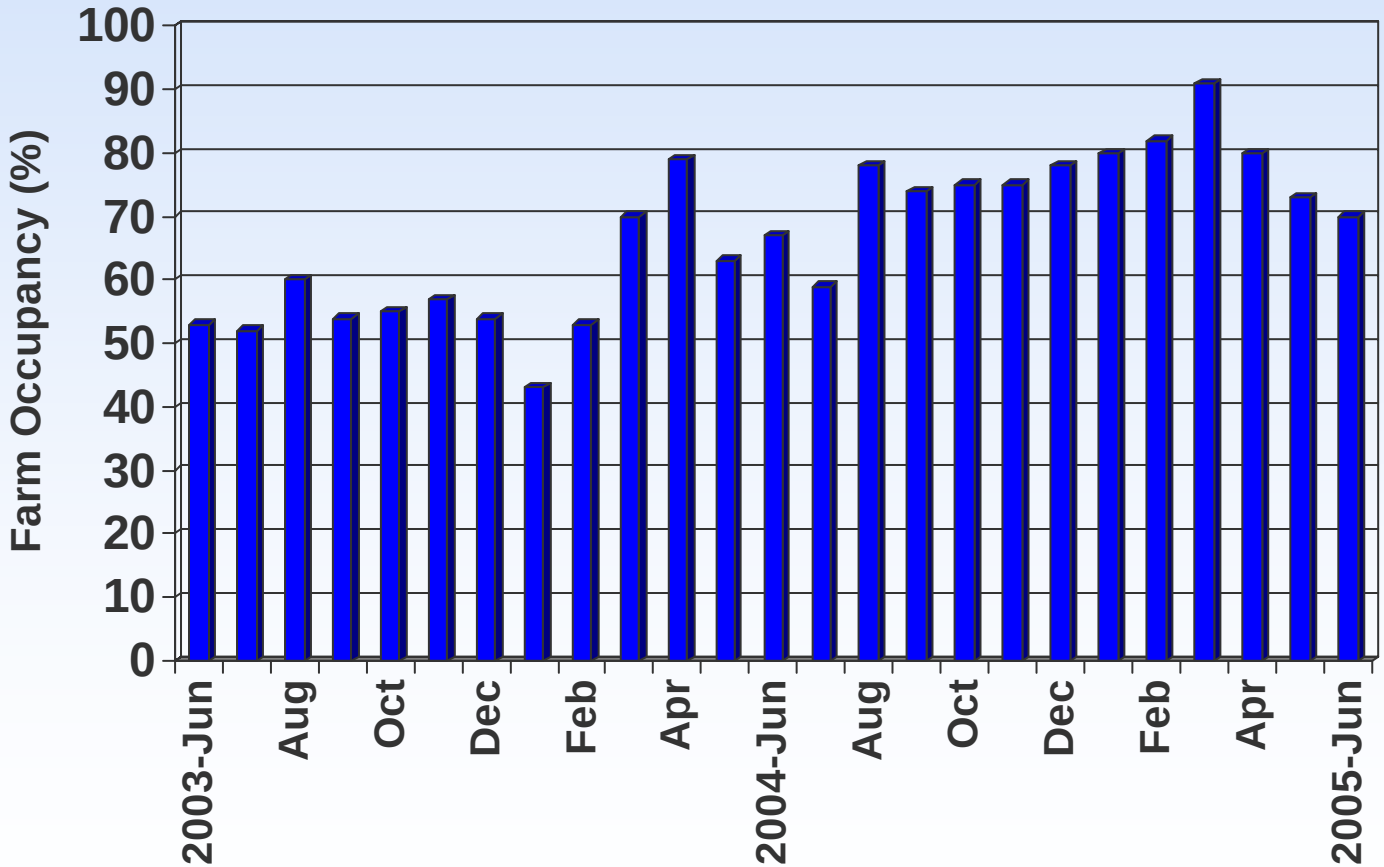


Compute Power



Linux Farm Occupancy

- New machines added in May 2004 and May 2005



Online (Disk) Storage

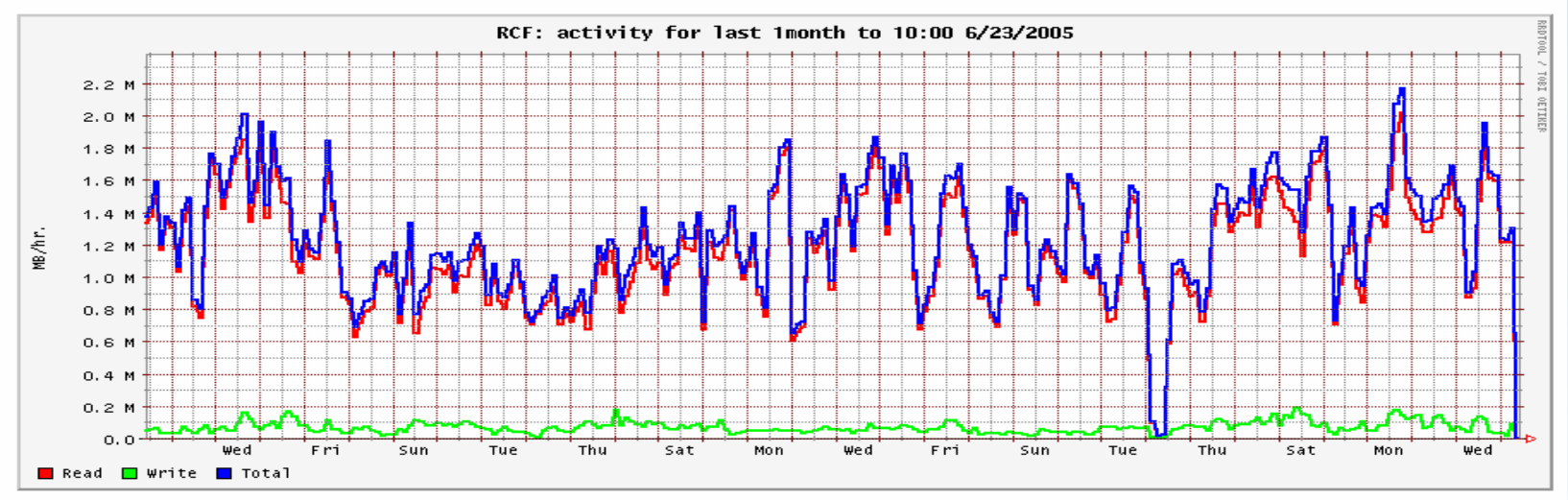
- Historically NFS served central disk
  - SUN served FibreChannel RAID 5

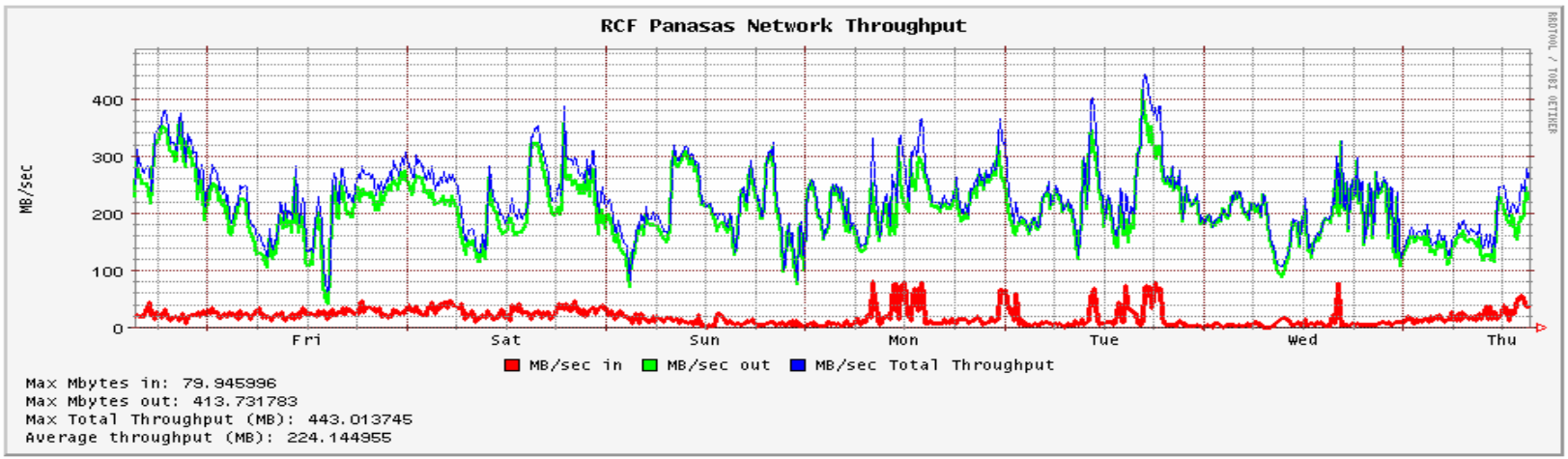
    … augmented by local scratch disk on CPU boxes
- Growing demand is driving disk costs to match and exceed CPU costs
- Strategy is to differentiate disk by function
  - ***Premium*** disk for critical and Read/Write not yet backed up data, of which there is a limited amount
  - ***Inexpensive*** disk for less critical, mostly Read already backed up data, of which there is a great deal

Premium Disk Service

- Bulk of useful disk in current system
- High performance and high availability
- Past technology: Sun served FibreChannel RAID arrays
  - 150 TB with measured serving rates up to 500 MB/sec
- New (superior) technology: Network appliance
  - Panasas, highly scalable, performant, robust commercial product
  - 70 TB with measured serving rates up to 400 MB/sec
- Aggregate premium disk of 220 TB with serving rates up to ~1GB/sec (~85 TB/day)
  - actually being used as “mostly read”

Sample NFS & Panasas Rates

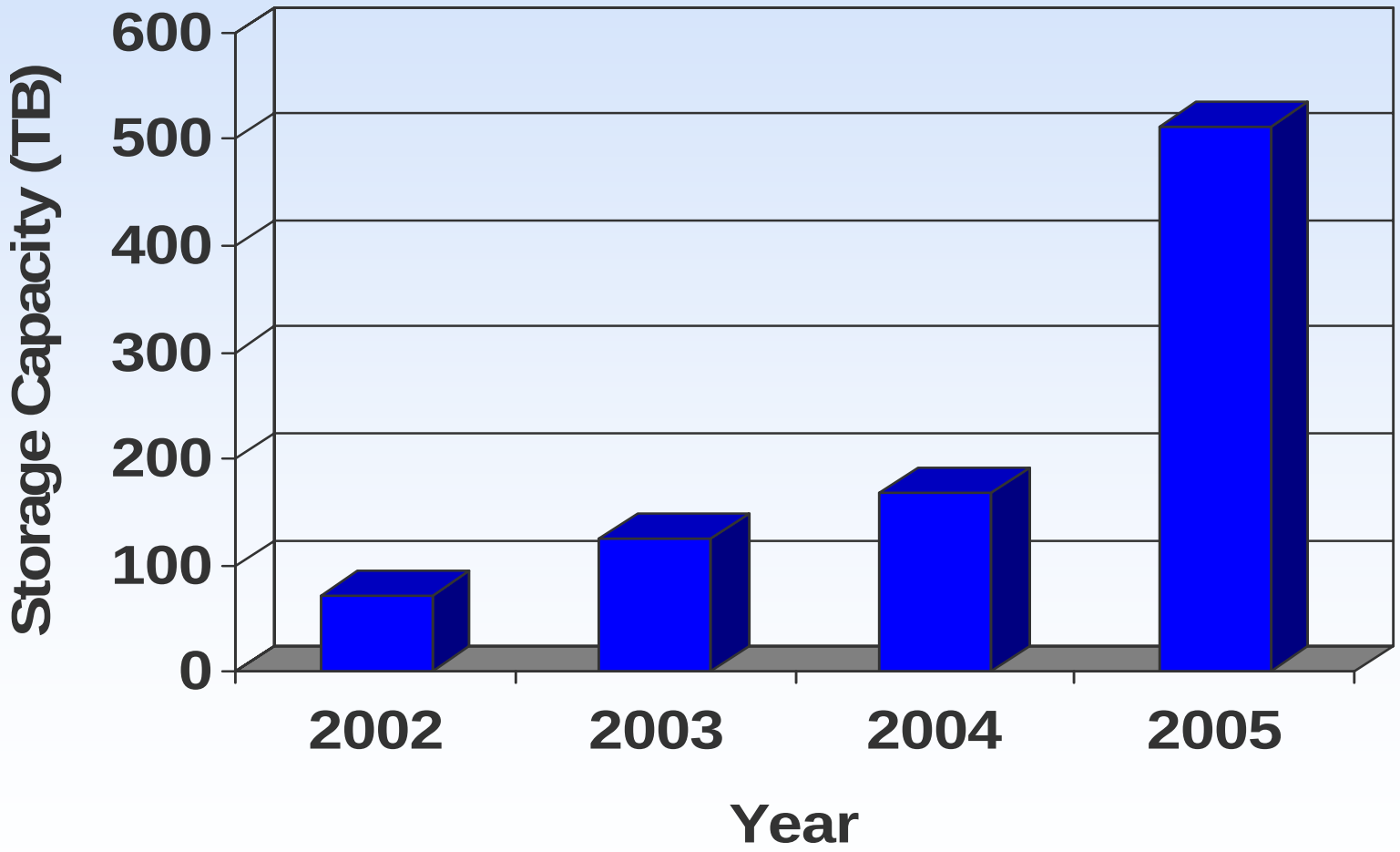




Inexpensive Disk

- Disk deployed on Linux processor farm nodes
  - Currently used in limited ways by various experiment specific mechanisms
- dCache: A centrally supported less limiting solution
  - DESY/Fermilab developed Grid aware product
  - A scalable, robust, unified pool of storage with integral mass store backing, posix-like access, ROOT support
  - ATLAS is current facility pathfinder on this product
    - 100 → 200 TB of Grid available disk in production
  - PHENIX is first RHIC user
    - 22 TB with measured serving rates to ~ 500 MB/sec
    - Near term expansion to 100 TB is possible

Available Storage Capacity in Linux Farm

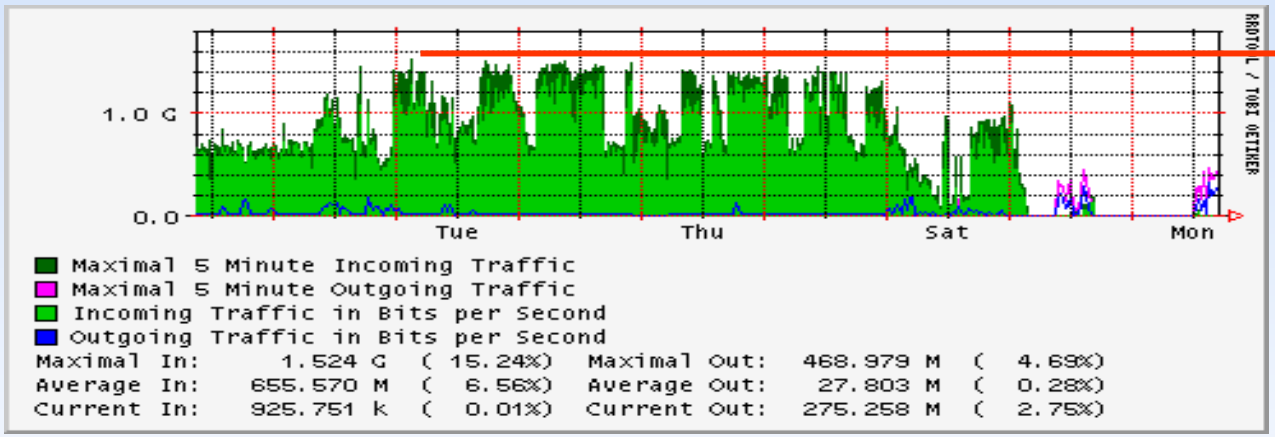


Mass Storage System

- Hierarchical Storage Manager is ***HPSS***
- 4 StorageTek robotic tape libraries
  - 4.25 PB of capacity containing 2.75 PB of data
- 37 StorageTek 9940b tape drives
  - 1.1 GB/sec of aggregate native streaming capacity
- 18.5 TB of disk cache
- In house developed tape access optimization software
- Addition of library and 20 tape drives underway
  - LTO Gen 3 drives: 2 x performance at ½ x price
  - A more commodity-like technology

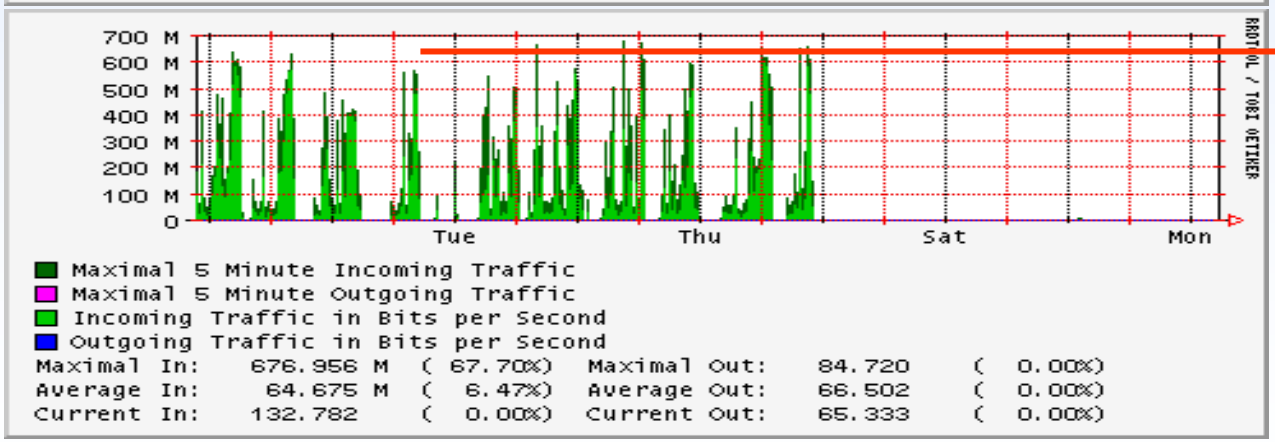
Raw Data Recording in HPSS System

- Aggregate Raw Data Recorded: 700 TBytes
- Aggregate Recording Rates: Up to 300 MBytes/sec



200 MBytes/sec

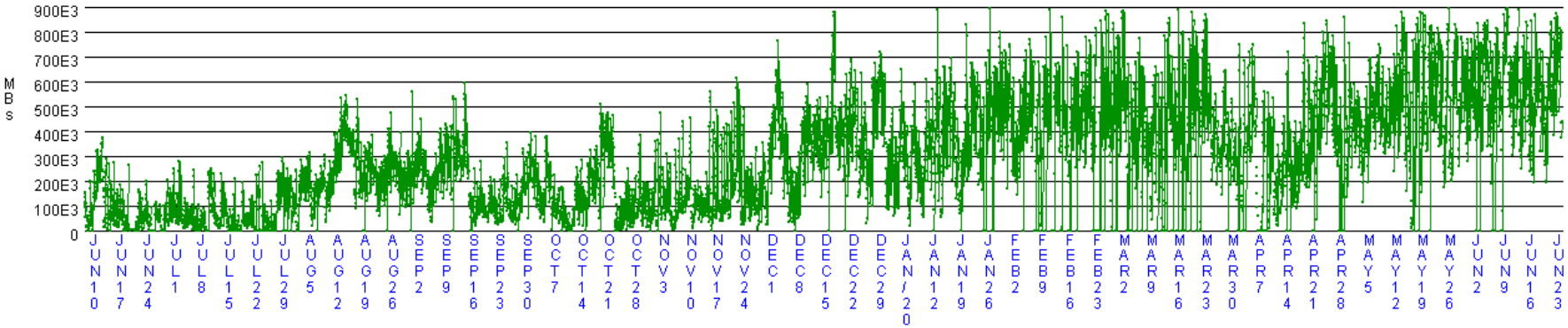
PHENIX



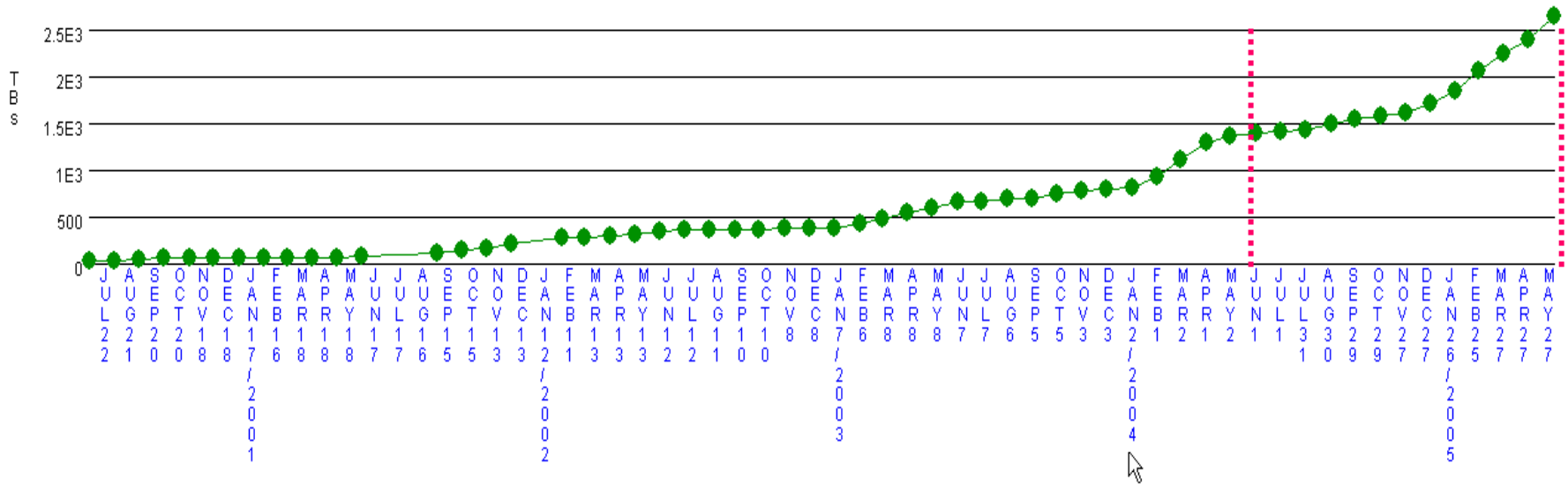
80MBytes/sec

STAR

HPSS Data Handling Rate (MB/sec) over Past Year



Volume of Data (TB) Stored in HPSS over Five Years



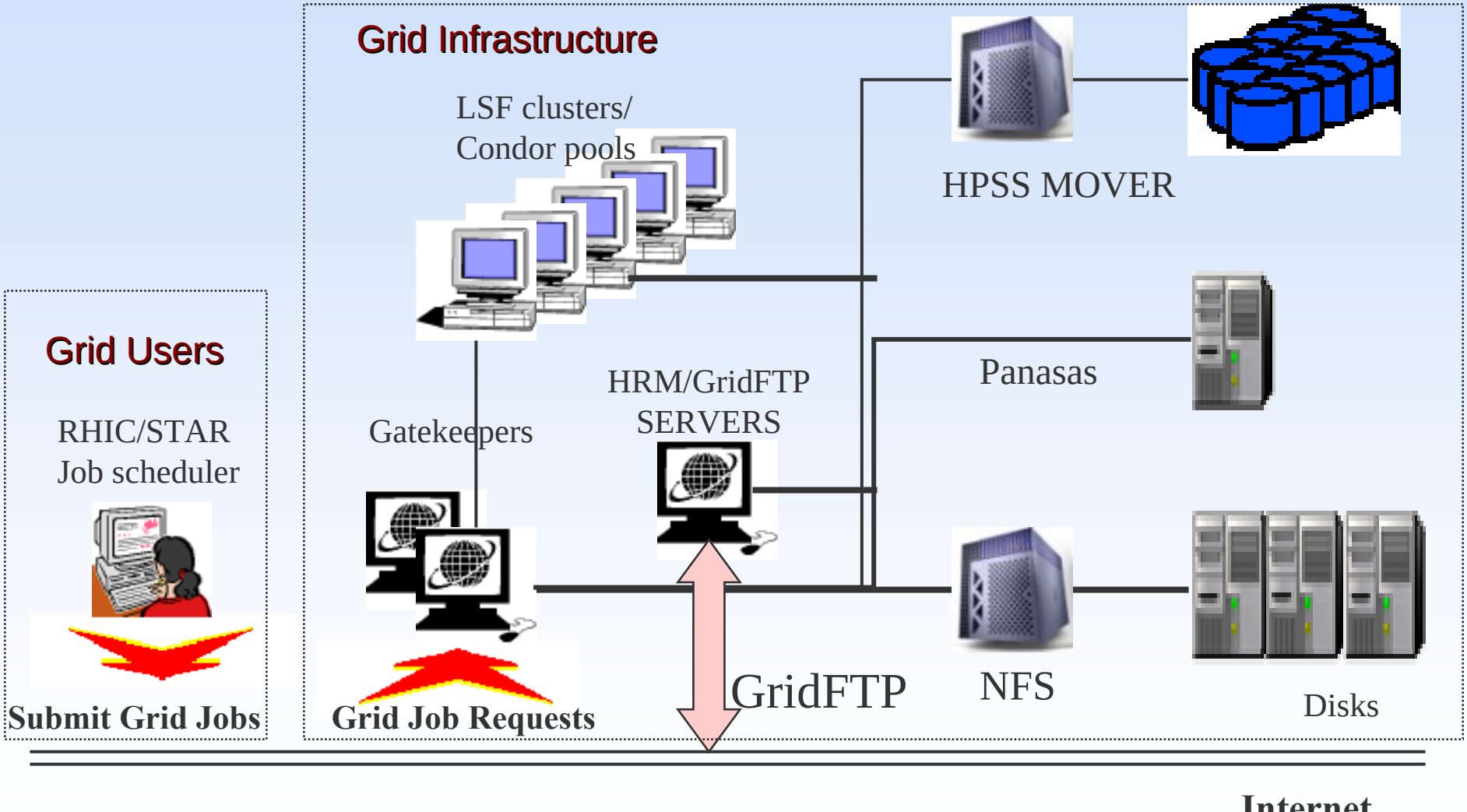
Grid and Network Services

- US ATLAS (centered around its Tier 1 at BNL) with US CMS are the primary drivers of Open Science Grid (OSG)
  - Major US ATLAS Grid production efforts have included
    - ATLAS Data Challenge 2 (May – Dec ’04)
    - ATLAS Rome Workshop Production (Jan – May ’05)
- Same staff now supports RHIC wide area distributed computing with …
  - Basic Grid VDT software deployment on general RHIC & STAR Grid servers
  - Customizes/optimizes servers to improve performance and security
  - Provide server level monitoring
  - Develop/support registration, authorization, and management tools
  - Supports involvement in OSG (In particular by STAR)
- Computing models of RHIC experiments predate the Grid
  - Unlike ATLAS, they were not based on Grid computing
  - The desire to utilize substantial distributed resources is now driving evolution of these models toward Grid computing

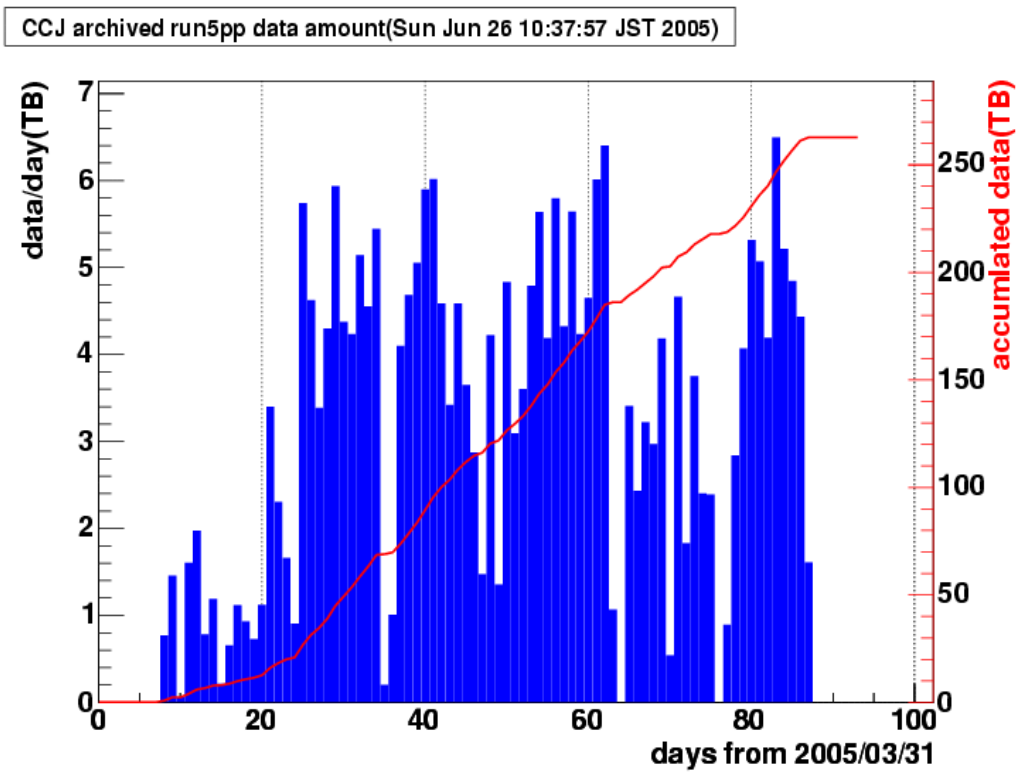
RHIC Distributed Computing

- Remote Facilities
  - In original planning identified primarily as source of Monte Carlo data
  - Now including substantial analysis activity, not unlike ATLAS model
  - Some of the sites now operational are:
    - STAR
      - NERSC/PDSF, LBNL
      - Sa Paulo, Brazil
      - Wayne State University
    - PHENIX
      - RIKEN, Wako campus, Japan
      - Center for High Performance Computing, University of New Mexico
      - IN2P3, Lyon, France
      - VAMPIRE cluster, Vanderbilt University
- Grid Computing Service Priorities
  - Initial strong focus on data movement (1st STAR, now PHENIX at high rate)
  - Evolving to distributed work flow (only STAR so far)

RHIC Grid Configuration at BNL



PHENIX Data Transfer to CCJ



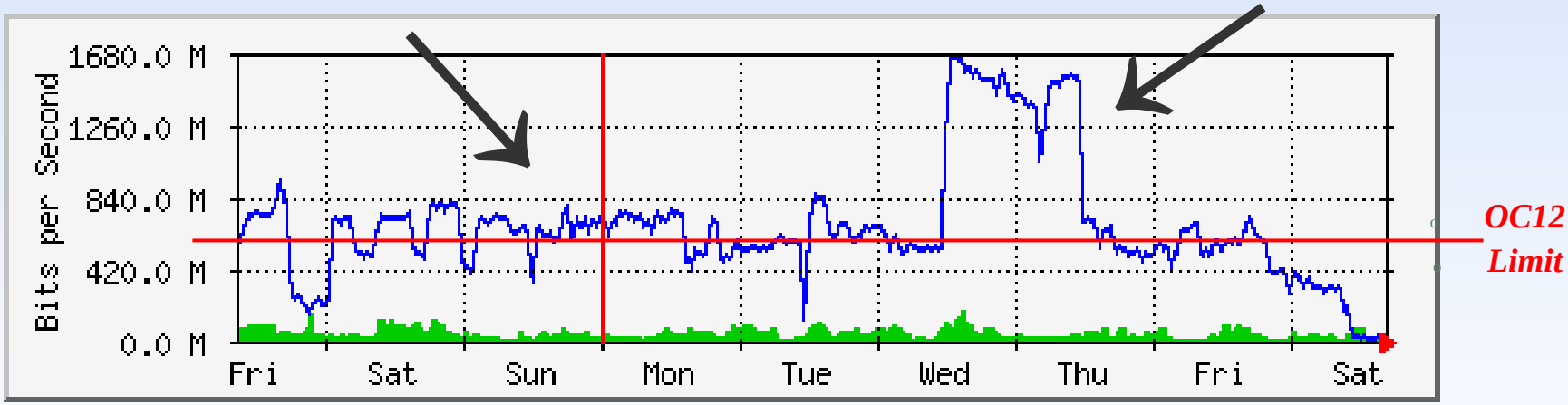
Courtesy of Y. Watanabe

- Total data transferred to CCJ (Computer Center in Japan) was 260 TB of Raw Data
- Average transfer rate was 60~90 MB/sec, ***~5TB/day!***
- RCF supplied Network/Grid expertise included:
  - Installation of Grid transfer and monitoring tools
  - Trouble shooting and tuning of network path
  - Integration dCache/SRM (Storage Resource Manager) for future transfers

Wide Area Networking

- In Nov '04 ESnet upgraded BNL connectivity
  OC12 (622 Mb/sec) ➔ OC48 (2.5 Gb/sec)
- Already being heavily used

PHENIX sustain transfer to Riken CCJ **ATLAS Service Challenge Test**



- Timely further upgrades, pressed for by ATLAS, are problematic from a funding perspective
- Also clearly a RHIC issue with collision of two very bandwidth hungry programs

Physical Infrastructure

- To date only modest improvement in physical infrastructure required
- Approaching limit of capacities in existing space
  - … chilled water, cooled air, UPS power, distributed power
  - Additional cooling and power will be required for 2006 additions
- Also approaching limit of computer space itself
  - Space for only one more round of procurements (2006) exists
  - By 2007 additional space must be identified and renovated
  - While not fully decide, 3rd floor of the collider center, now office space but on raised floor, is a prime candidate
- Working now to get these needs into GPP planning process
- Power cost projections show that, for total facility, there will likely be a rise in 5 year from ~ $200k now to ~ $1.5M
  - Technology advances, 64 bit architectures and dual core CPU's, are intended to counter this trend but will not effectively ameliorate it

Cyber Security

- Facility is a firewall protected enclave within the BNL firewall protect site
- Good working relationship with BNL ITD cyber security
- Facility has a ~single sign-on Kerberos base authentication infrastructure
- Do almost all of the things that could reasonably be asked to assure cyber security – need and are recruiting more help
- Only cyber security lapses of the last year have been at the user level; a result of externally sniffed passwords
- None of these resulted in root level comprise of any facility computer or result in the utilization of facility resources against other sites
- A concern is possible future conflict between Grid requirements, regulatory requirements, and an effective single facility cyber security policy/architecture

Conclusions

- Plans to evolve and expand facility services to meet expected needs exist
  - Based on to date successful adjustment of technical directions while remaining within the mainstream of the HENP computing
  - They fail to fit within equipment refresh funding guideline in FY ’06 but appearing to fit within such funding in the out years
- Physical infrastructure issues are of particular concern
  - Most critically, the need to establishing new space within 2 years
  - In long term power costs will significantly increase
- The Grid is likely to significantly change future computing
  - Based on ATLAS pathfinder experience, RCF should be well positioned to exploit its advantages
  - RCF should not jeopardize a successful central facility approach in pursuit of a Grid solution until it is fully successfully demonstrated
- WAN bandwidth is the critical infrastructure of the Grid
  - RHIC (NP) must cooperate with ATLAS (HEP) to assure sufficiency