

# Distributed and/or Grid-Oriented Approach to BTeV Data Analysis

Talk at Beauty2002

J. N. Butler

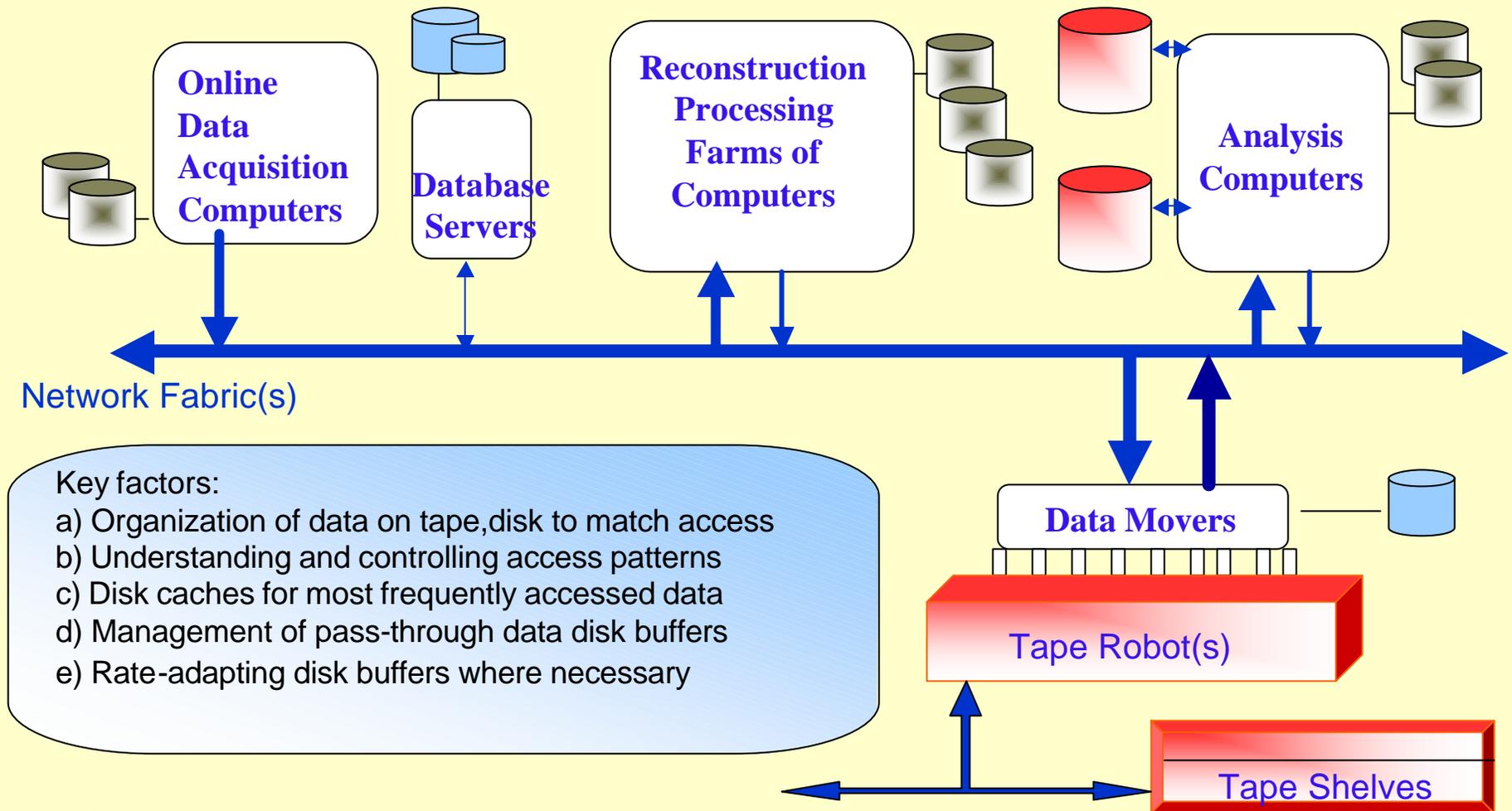
Fermilab

June 18, 2002

# Data Profile for BTeV

Event Size	Total events/year = $4 \times 10^{10}$	Total
<b>50KB</b>	RAW detector measurements	2 Pbytes/year
<b>~50KB</b>	Reconstructed Data - Hits, Tracks, Clusters, Particles	2 Pbytes/year
	Estimate based on 1000 datasets of $10^7$ events each	
<b>~10 KB</b>	Summary Physics Objects	0.1 Pbytes/year
<b>2-5KB</b>	Condensed summary physics data	.02-0.05 Pbytes/ year
<b>~200B</b>	Data Catalog entry	20 Tbytes

# Centralized Data Access Model -Traditional Approach



# The Drawback

- The centralized model ignores the opportunity to expand the resources available to the collaboration by taking advantage of the increased commitment to computing by universities and the increased interest in distributed and GRID computing by funding agencies

# The Problem/Opportunity - I

- **Many** research problems must be addressed by establishing a very expensive facility, which can be, e.g.
  - An accelerator, reactor, or large telescope array;
  - A large database or catalogue; or
  - A collection of satellites, a space-based detector
  - A collection of sensors, e.g. for weather prediction

For these systems, data must be collected, monitored, and stored over large periods of time. It may be impractical physically or unacceptable socially or professionally to consolidate all these activities in a single site

# The Problem/Opportunity - II

- After the data are collected, they typically present a vast array of analysis topics that require a large community to explore. The analysis proceeds best if the people doing it can interact with each others to share methods, techniques, and results.
- Simulation is becoming the LABORATORY of choice for problems where the physical principles are known and the problem is just to do the computing. Every serious research-oriented university will need such a virtual multi-laboratory!

It may be impractical physically or unacceptable socially or professionally to consolidate all these activities in a single site. For this reason, significant resources are pouring into distributing the computing facilities. The resources include money for hardware and support and funding for R&D into methods. The “datagrid” efforts are the most ambitious ones in terms of power and transparency.

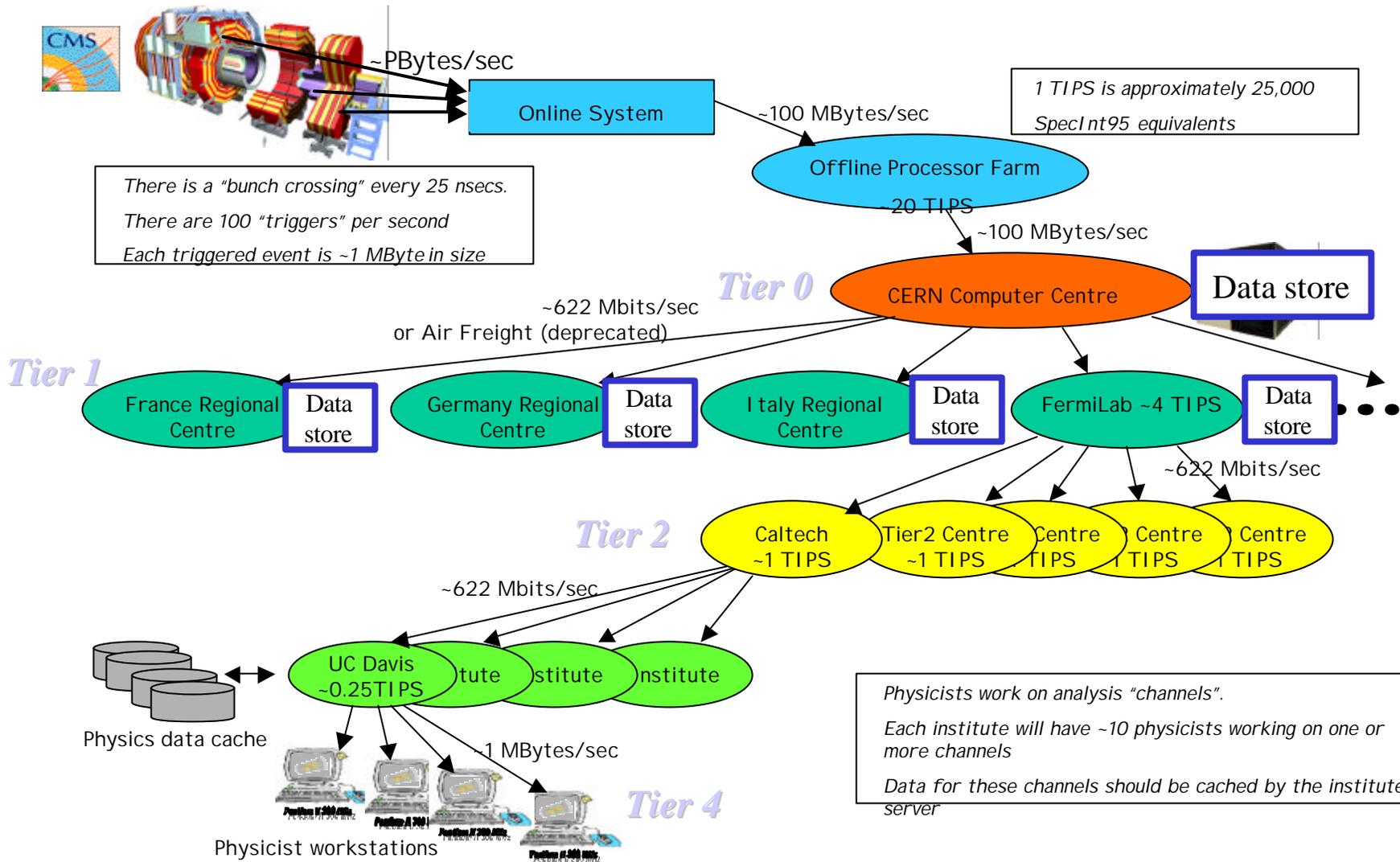
# Approach to a Solution

- Modern networking and computer technology provide the basic building blocks for a different, distributed solution, which addresses some of these issues and will include
  - **A large assemblage of hardware at several sites**
    - **CPU**
    - **Data storage -- disk, tape or their successor, etc. and very high speed data access**
    - **Excellent connectivity**
  - **Excellent software tools to maintain data and data catalogues, distribute, cache, replicate the data as required to provide excellent access with minimal effort for data analysts. This includes data integrity and security.**
  - **Excellent support services including software development personnel, consultants. User support, operations staff, network specialists, database managers, data aides, documentation specialists, etc somewhere in the system**

# A “Conservative” Approach - A Hierarchical Distributed Model

- One method to improve the use of the physics resources and financial resources of the collaborations and nations involved will occur if we adopt a computing model based on a **hierarchy of computing centers. The HEP model follows:**
  - At the top of this hierarchy is a large center which have capability to do all analysis related functions but not the capacity
  - Below this sits a collection of large, multi-service centers with capacities that are a significant fraction , 10-20%, o f the large center. We call these “Tier 1” Regional Centers (RC).
  - Below Tier 1, there may be a set of Tier 2 Regional Centers which provide services to users in a subregion while receiving services from a Tier 1 Regional Center
  - There may also be special “service centers” which provide limited capability such as event simulation

# Hierarchical Distributed Data Concept a la LHC



# Important Points

- This is a distributed model but it not, at least from a purist's point of view a Grid Model.
- A subset of properly working GRID Software should certainly be able to support this model
- This is not just about HARDWARE. **Support** is identified as **the key element** of making all this work. It is essential for an RC to provide a critical mass of user support. Software with the user friendliness and robustness of GRID software will lighten the load
- This is a **commitment that extends over a long time** -- site support, staff and funds for continual hardware evolution and even R&D must be provided. This will probably be funded by an arrangement with the agencies supporting the applications being hosted. GRID approaches will lower the risk due to sites disappearing.

# Significant Computing Issues to be Addressed

- Management of **large scale clusters** in a quasi real time production environment. This system is 10-100 times larger than ones operating today or in near future
  - Mechanisms for monitoring and operation for **manageability, reliability, and fault tolerance**
  - **Software installation** and update on **1000's of computers**
  - Determination of **performance and tuning** of ensemble
  - **Distributed authority and access control (I.e. security and policy issues)**
  - **interconnects**, peripherals, and data/message passing protocols to achieve efficient use
  - **System architecture** to provide the most efficient and **scaleable** system

# The Computational DataGRID

- The idea is to be able to use, in a transparent manner, distributed computing resources to solve a problem as if it were a single workstation. You essentially have constructed for you a virtual, distributed computing facility for each problem -- including a CPU cluster, mass storage system, etc.
- This, in turn, requires software to take a computation, assemble resources over the network to carry it out, split up and distribute the job, permit monitoring and control, allow the results to be collected up, and returned. You are building ad hoc virtual computation farms.
- There are many major issues
  - Security
  - Resource discovery
  - Distributed ownership and local policy
  - System heterogeneity
  - Fault recovery
  - Etc., etc.

Note that this is the same list required to implement the hierarchical distributed model but with much more transparency and an even more distributed “distributed model”

# Fusion of Approaches

- Most HEP jobs can be split up since the issue is analyzing many “independent events”.
- The hierarchical model really has several logical “subcommunities” coexisting at each level. Resources have to be dynamically allocated among them.
- The optimization of resources is best achieved if tasks can actually move among layers to take advantage of available resources wherever they exist
- Transparency is obviously of great value

**It is clear that GRID software, if it achieves its goals, should also satisfy most of the requirements of an optimal distributed hierarchical model. Thus, the two camps have converged: we plan to use the GRID software to implement our “approximate” distributed hierarchical model**

# The BTeV View

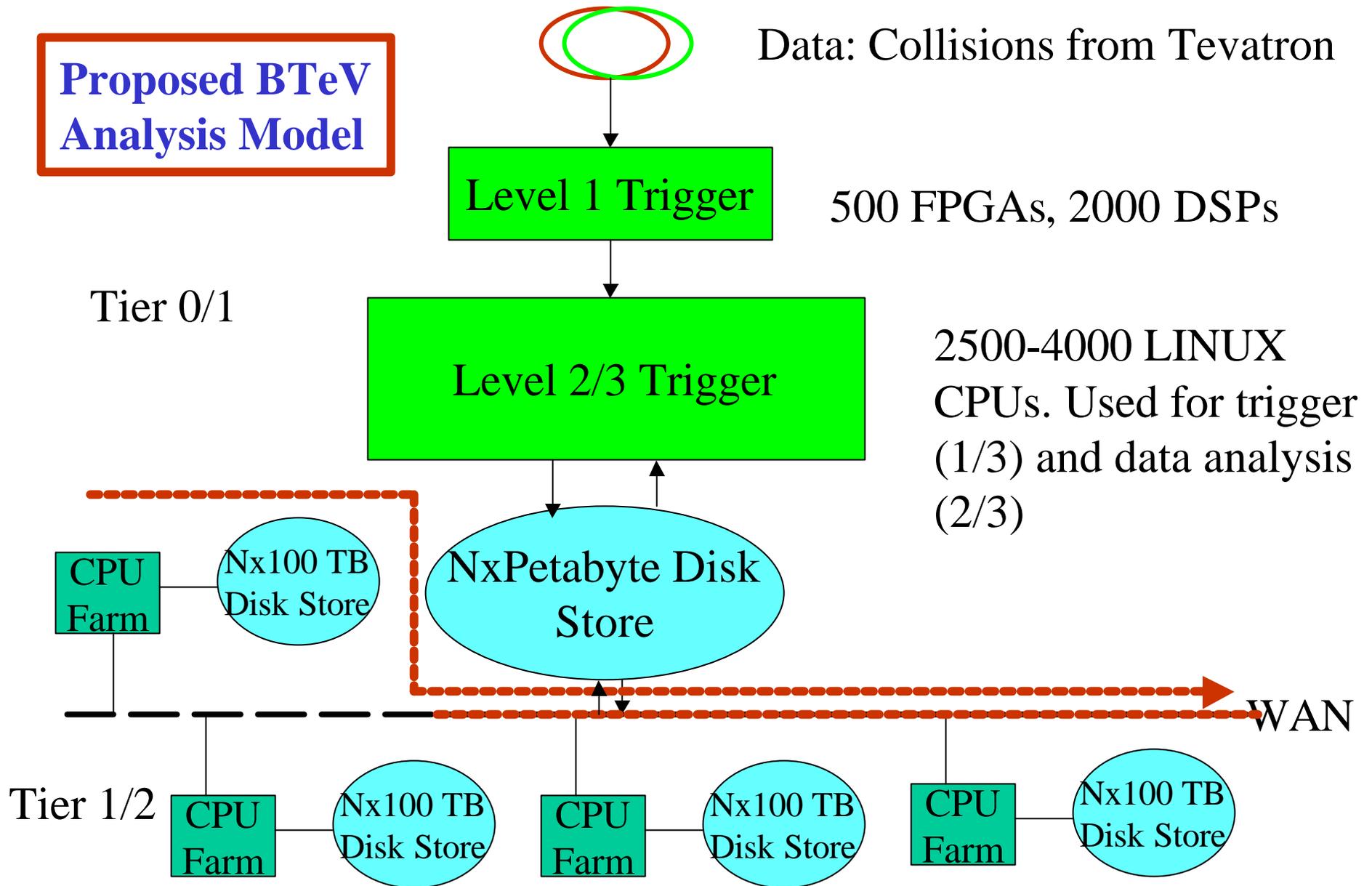
- We have a huge trigger farm which is available for offline reconstruction during periods of low luminosity (late in stores), between stores, during time when beam is down or not scheduled. This amounts to 2/3 availability
  - Have software which permits the dynamic partitioning of the trigger farm among several tasks
    - The NSF-funded “Real Time Embedded Systems” or “RTES” project -- \$5Milliion/5 years -- has as one of its goals to supply this software
  - To get the data into it, let it stay on the farm a long time. Have a disk system of about >1 Pbyte attached to the farm

# BTeV View

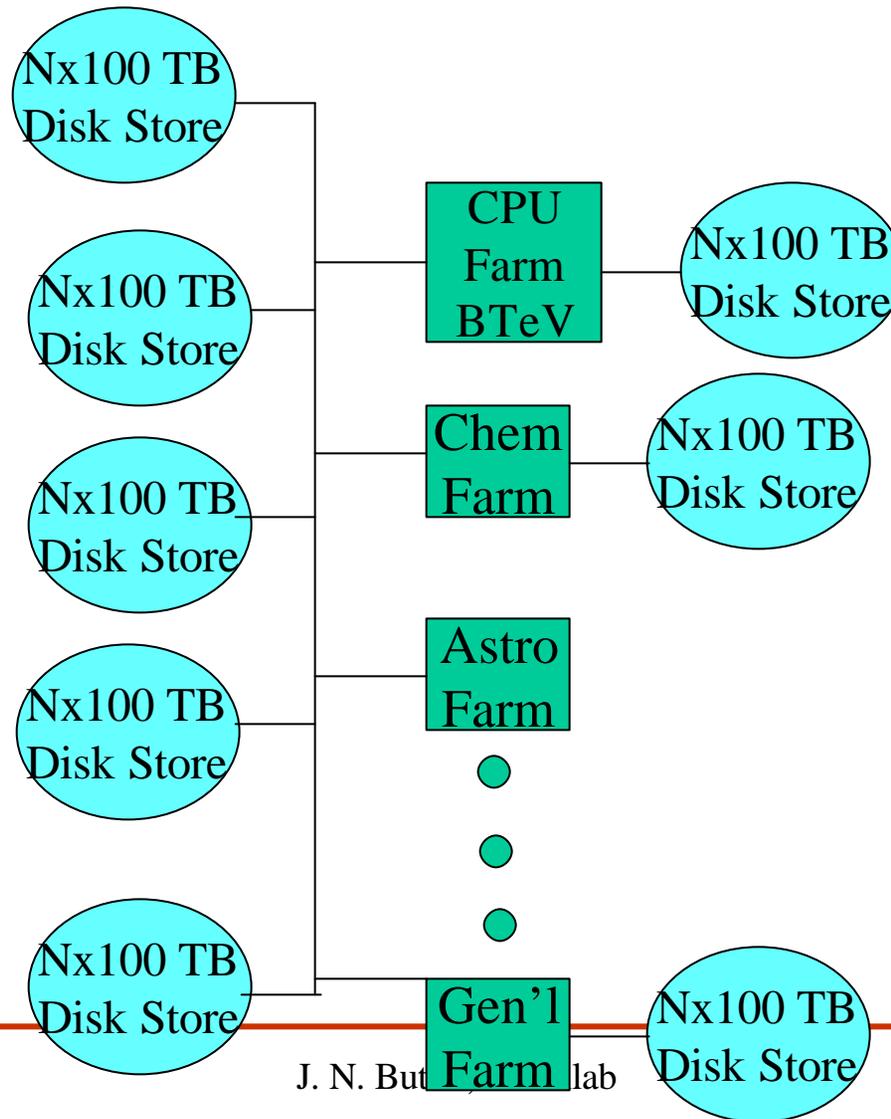
- To use as many offsite facilities as possible to supply resources
  - We have of order one dozen institutions that plan major computation facilities with HEP expected to be major users
    - In most cases, these will be shared facilities, so the software has to exist to assemble “virtual production/analysis” facilities
- BTeV has the concept of a Level 4 trigger, which we “might” implement. It was originally conceived as editing the data within a few months of when it was recorded to “summarize” various low physics interest samples, to apply any final calibrations, fixups, etc. This task can also be done at remote sites

**Proposed BTeV  
Analysis Model**

Data: Collisions from Tevatron



# Extended View



Want to be able to flexibly use any portion that is unused and allow any “BTeV” resource that is unused to be used by others. Want to extend This both ways To non-BTeV sites

# Tapes?

- BTeV will facilitate this by having even primary datasets on disk. We plan one primary dataset on disk at FNAL and perhaps 2 or more “copies” distributed over say 5 major sites, for access and redundancy purpose
- We may not write tape at all. There is, of course, the WORN (Write Once Read Never) crowd which wants a tape copy in case of disaster. This would not need to operate at high access speeds for analysis, since it is intended never to be read.

# Where might BTeV do something “unique”?

- **Break down barrier between online/trigger computing and offline, by allowing farm nodes to do simulation and analysis when trigger is not running or nodes have unused cycles**
  - **Break down barrier between offline and trigger by allowing “Level 4” (possibly even part of Level 3) to run offsite if required.**
  - **Proposal would**
    - **Develop the key missing pieces of Grid software to do this**
    - **Ask for prototype systems to conduct large scale, meaningful tests**
    - **Eventually get production systems and some support for major (equivalent to Tier 2 sites)**
  - **Software tasks would be**
    - **Extended from RTES of pre-emption scheduling, policy software**
    - **Very powerful interactive user interface to define jobs, acquire resources, launch and monitor jobs**
    - **Extended fault tolerance**
-

# Conclusion

- ❖ The distributed data analysis problem is a major challenge. The large number of people simultaneously analyzing such a large dataset may be the model of the collaborative, networked environment in which the research of the future will be done.
- ❖ This problem is at the top of the list for IT challenges for many of the agencies that fund science
- ❖ To get started, you need participation from the University Computer Science community as well as from one or more applications communities.
- ❖ Initial successes will then attract other applications communities
- ❖ BTeV offers significant advantages for a prototype application
- ❖ Fermilab's strong participation in these efforts for its own experimental program, which includes BTeV, and for its participation in the LHC project will provide a significant advantage as well
- ❖ The BTeV Model breaks down many of the distinctions between “real-time/online” and “offline” and blurs the distinctions in the hierarchy so the various “centers” are much more equal.
- ❖ If the grid is turns to to be too ambitious, the distributed hierarchical model is our “Plan B”