

# BTeV and the Grid

- What is BTeV?
- A “Supercomputer with an Accelerator Running Through It”
- A Quasi-Real Time Grid?
- Use Growing CyberInfrastructure at Universities
- Conclusions



# What is BTeV?



- **BTeV** is an experiment designed to challenge our understanding of the world at its most fundamental levels
- Abundant clues that there is new physics to be discovered
  - Standard Model (SM) is unable to explain baryon asymmetry of the universe and cannot currently explain dark matter or dark energy
  - New theories hypothesize extra dimensions in space or new symmetries (supersymmetry) to solve problems with quantum gravity and divergent couplings at the unification scale
- **Flavor physics** will be an equal partner to **high  $p_t$**  physics in the LHC era... **explore at the high statistics frontier** what can't be explored at the energy frontier.

# What is BTeV?

## Mysteries

Dark Matter

Dominance of Matter over Antimatter

Dark Energy

## Solutions: New Physics

*new particles*

LHC  
New Particles

*new physics, found 1st in either place; specified in both*

*new source of CP violation*

~~CP~~ & Rare  
b & c decays  
**BTeV**

$\nu$   
Mixing

*new models of GUTS & flavor*

*? new forces, dimensions?*

Hubble  
JDEM

figure courtesy of S. Stone

# Requirements

- Large samples of tagged  $B^+$ ,  $B^0$ ,  $B_s$  decays, unbiased  $b$  and  $c$  decays
- Efficient Trigger, well understood acceptance and reconstruction
- Excellent vertex and momentum resolutions
- Excellent particle ID and  $\gamma$ ,  $\pi^0$  reconstruction



Physics Quantity	Decay Mode	Vertex Trig	K/ $\pi$ Sep	$\gamma$ Det	Decay Time $\sigma$
$\sin(2\alpha)$	$B^0 \rightarrow \rho\pi \rightarrow \pi^+\pi^-\pi^0$	✓	✓	✓	
$\cos(2\alpha)$	$B^0 \rightarrow \rho\pi \rightarrow \pi^+\pi^-\pi^0$	✓	✓	✓	
$\sin(\gamma)$	$B_s \rightarrow D_s K^-$	✓	✓		✓
$\sin(\gamma)$	$B^0 \rightarrow D^0 K^-$	✓	✓		
$\sin(2\chi)$	$B_s \rightarrow J/\psi\eta, J/\psi\eta'$		✓	✓	✓
$\sin(2\beta)$	$B^0 \rightarrow J/\psi K_s$				
$\cos(2\beta)$	$B^0 \rightarrow J/\psi K^0, K^0 \rightarrow \pi l \nu$		✓		
$\chi_s$	$B_s \rightarrow D_s \pi^-$	✓	✓		✓
$\Delta\Gamma$ for $B_s$	$B_s \rightarrow J/\psi\eta^{(\prime)}, K^+K^-, D_s\pi$	✓	✓	✓	✓

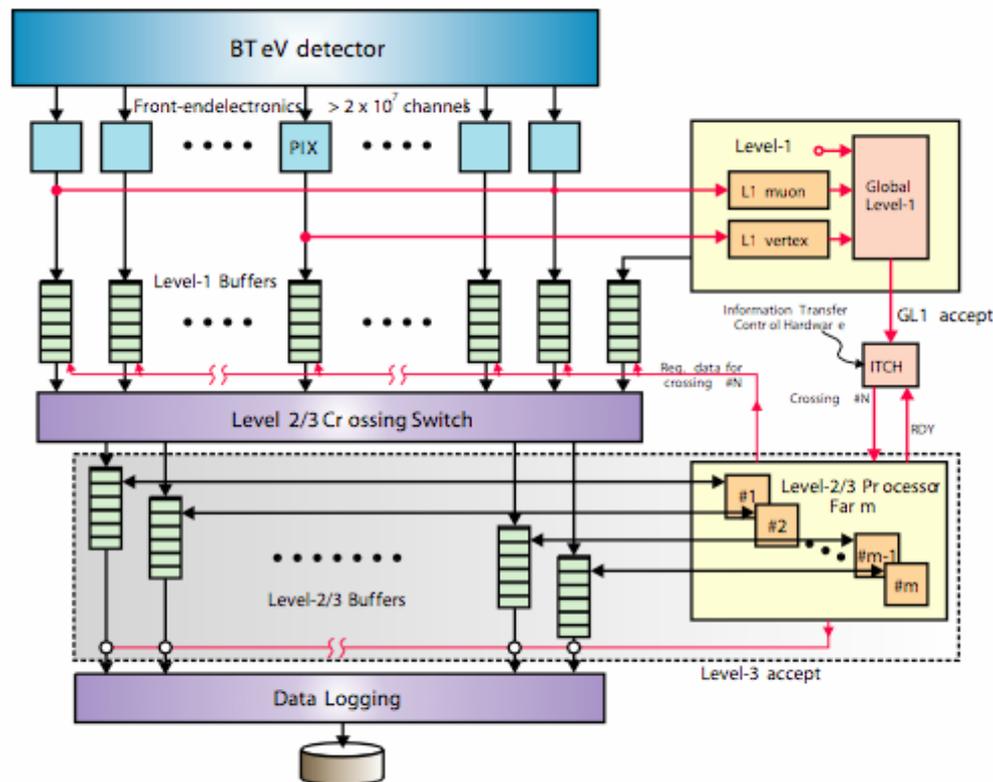
# The Next Generation

- The next (2<sup>nd</sup>) generation of B-factories will be at hadron machines: BTeV and LHC-b
  - both will run in the LHC era.
- Why at hadron machines?
  - $\sim 10^{11}$  b hadrons produced per year ( $10^7$  secs) at  $10^{32}$  cm<sup>-2</sup>s<sup>-1</sup>
  - $e^+e^-$  at  $\Upsilon(4s)$ :  $\sim 10^8$  b produced per year ( $10^7$  secs) at  $10^{34}$  cm<sup>-2</sup>s<sup>-1</sup>
  - Get all varieties of b hadrons produced:  $B_s$ , baryons, etc.
  - Charm rates are 10x larger than b rates...
- Hadron environment is challenging...
  - CDF and D0 are showing the way
- BTeV: trigger on detached vertices at the first trigger level
  - Preserves **widest possible spectrum** of physics – **a requirement.**
  - Must compute on every event!



# A Supercomputer w/ an Accelerator Running Through It

- Input rate: 800 GB/s (2.5 MHz)
- Made possible by 3D pixel space points, low occupancy
- Pipelined w/ 1 TB buffer, no fixed latency
- Level 1: FPGAs & commodity CPUs find detached vertices,  $p_t$
- Level 2/3: 1280 node Linux cluster does fast version of reconstruction
- Output rate: 4 KHz, 200 MB/s
- Output rate: 1—2 Petabytes/yr
- 4 Petabytes/yr total data



# BTeV is a Petascale Expt.



- Even with sophisticated event selection that uses aggressive technology, BTeV will produce  
**Petabytes of data/year**
- And require  
**Petaflops of computing to analyze its data**
- Resources and physicists are geographically dispersed  
*(anticipate significant University based resources)*
- To maximize the quality and rate of scientific discovery by BTeV physicists, all must have equal ability to access and analyze the experiment's data...

**...BTeV Needs the Grid...**

# BTeV Needs the Grid



- **Must build hardware and software infrastructure**  
*BTeV Grid Testbed and Working Group Coming online.*
- **BTeV Analysis Framework is just being designed**  
*Incorporate Grid tools and technology at the design stage.*
- **Benefit from development that is already going on**  
*Don't reinvent the wheel!*
- **Tap into expertise of those who started before us**  
*Participate in iVDGL, demo projects (Grid2003)...*
- In addition, **propose "non-traditional"** (for HEP?) **use:**  
**Quasi Real-Time Grid**

# Initial BTeV Grid Activities

- **Vanderbilt BTeV Group Joined iVDGL as an “external” collaborator**

- Participating in VDT Testers Group



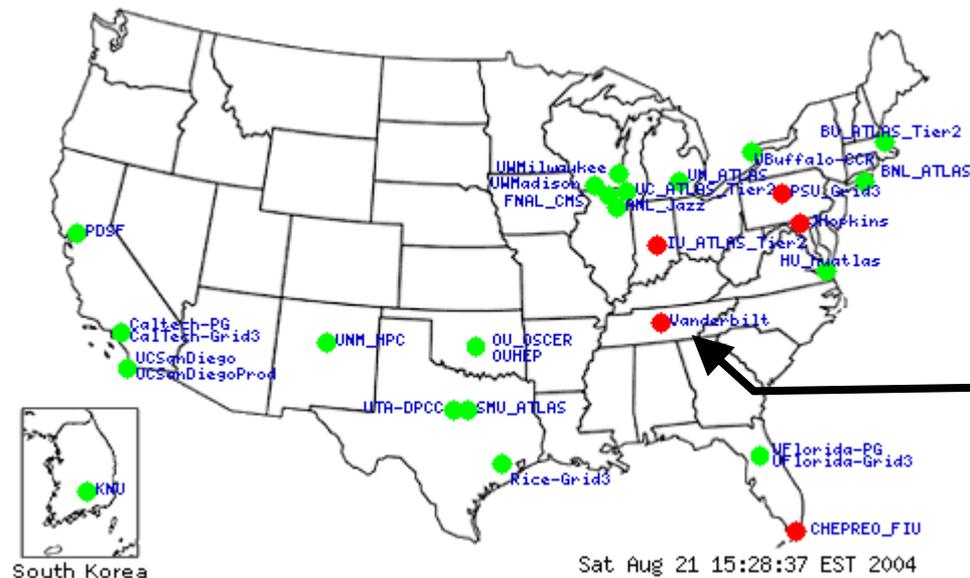
- **BTeV application for Grid2003 demo at SC2003**

- Intergrated BTeV MC with vdt tools
  - Chimera virtual data toolkit
  - Grid portals
- Used to test useability of VDT interface
- Test scalability of tools for large MC production



# ...Initial BTeV Grid Activities

- **Grid3 Site: 10-cpu cluster at Vanderbilt** ACCRE
  - Accomodates use by multiple VOs
  - VDT-toolkit, VO management, monitoring tools...



# ...Initial BTeV Grid Activities



- BTeV Grid Testbed
  - Initial Sites established at Vanderbilt and Fermilab
  - Iowa and Syracuse likely next sites
  - Colorado, Milan (Italy), Virginia within next year.
  - BTeV Grid Working Group with twice monthly meetings.
  - Operations support from Vanderbilt 
- Once established, will use for internal “Data Challenges” and will add to larger Grids

# ...Initial BTeV Grid Activities



- **Storage development with Fermilab, DESY (OSG)**
  - Packaging the Fermilab ENSTORE program (tape library interface)
    - Taking out site dependencies
    - Documentation and Installation scripts / documentation
    - Using on two **ACCRES** tape libraries
  - Adding functionality to dCache (DESY)
  - **ACCRES** using dCache/ENSTORE for HSM, once complete will be used by medical center and other Vanderbilt researchers
    - Developing in-house expertise for future OSG storage development work.

# Proposed Development Projects



- Quasi Real-Time Grid
  - Use Grid accessible resources in experiment trigger
  - Use trigger computational resources for “offline” computing via dynamic reallocation
- Secure, disk-based, widely distributed data storage
  - BTeV is proposing a tapeless storage system for its data
  - Store multiple copies of entire output data set on widely distributed disk storage sites

# Why a Quasi Real-Time Grid?

- **Level 2/3 farm**

- 1280 20-GHz processors
  - split into 8 "highways" (subfarms fed by 8 Level 1 highways)
- performs first pass of "offline" reconstruction
- At peak luminosity processes 50K evts/sec, but this rate falls off greatly during a store (*peak luminosity = twice avg. luminosity*)

- **Two (seemingly contradictory) issues...**

- Excess CPU cycles in L2/3 farm are a significant resource
- Loss of part of the farm (*e.g.* one highway) at a bad time (or for a long time) would lead to significant data loss

- **Break down the offline/online barrier via Grid**

- Dynamically re-allocate L2/3 farm highways for use in offline Grid
- Use resources at remote sites to clear trigger backlogs and explore new triggers

- **Real Time with soft deadlines: Quasi Real-Time...**

# Quasi Real-Time Use Case 1

- **Clearing a backlog or Coping with Excess Rate**
  - If L2/3 farm can't keep up, system will at a minimum do L2 processing, and store kept events for offsite L3 processing
  - Example: one highway dies at peak luminosity
    - Route events to remaining 7 highways
    - Farm could do L2 processing on all events, L3 on about 80%
    - Write remaining 20% needing L3 to disk:  $\sim 1$  TB/hour
    - 250 TB disk in L2/3 farm, so could do this until highway fixed.
    - These events could be processed in real time on Grid resources equivalent to 500 CPUs (and a 250 MB/s network)
    - In 2009, 250 MB/s likely available to some sites, but it is not absolutely necessary that offsite resources keep up unless problem is very long term.
  - This works for other scenarios as well (excess trigger rate,...)
- **Need Grid based tools for initiation, resource discovery, monitoring, validation**

# Quasi Real-Time Use Case 2



- **Exploratory Triggers via the Grid**
  - Physics Triggers that cannot be handled by L2/3 farm
    - CPU intensive, lower priority
  - Similar to previous use case
    - Use cruder trigger algorithm that is fast enough to be included
    - Produces too many events to be included in normal output stream
    - Stage to disk and then to Grid based resources for processing.
  - Delete all but enriched sample on L2/L3 farm, add to output stream
  - Could use to provide special monitoring data streams
- Again, need Grid based tools for initiation, resource discovery, monitoring, validation

# Dynamic Reallocation of L2/3



- **When things are going well, use excess L2/3 cycles for offline analysis**
  - L2/3 farm is a major computational resource for the collaboration
  - Must dynamically predict changing conditions and adapt:  
Active real-time monitoring and resource performance forecasting
  - Preemption?
  - If a job is pre-empted, a decision: wait or migrate?

# Secure Distributed Disk Store



- “Tapes are arguably not the most effective platform for data storage & access across VOs” – Don Petravick
  - Highly unpredictable latency: investigators loose their momentum!
  - High investment and support costs for tape robots
  - Price per GB of disk approaching that of tape
  - Want to spread the data around in any case...
- Multi-petabyte disk-based wide-area secure permanent store
  - Store subsets of full set at multiple institutions
  - Keep three copies at all times of each event (1 FNAL, 2 other places)
  - Back-up not required at each location: backup is other two copies.
  - Use low cost commodity hardware
  - Build on Grid standards & tools

# ...Secure Distributed Store



- Challenges (subject of much ongoing work):
  - Low latency
  - Availability: exist and persist!
    - High bit-error rate for disks
    - Monitor for data loss and corruption
    - "burn in" of disk farms
  - Security
    - Systematic attack from the network
    - Administrative accident/error
    - Large scale failure of a local repository
    - Local disinterest or even withdrawal of service
  - Adherence to policy: balance local and VO requirements
  - Data migration
    - Doing so seamlessly is a challenge.
  - Data proximity
    - Monitor usage to determine access patterns and therefore allocation of data across the Grid

# University Resources are an essential component of BTeV Grid

- Cyberinfrastructure is growing significantly at Universities
  - Obvious this is true in Korea from this conference!
- Funding Agencies being asked to make it a high priority...
- **Increasing importance in new disciplines... & old ones**

“...the exploding technology of computers and networks promises profound changes in the fabric of our world. As seekers of knowledge, researchers will be among those whose lives change the most.  
...Researchers themselves will build this New World largely from the bottom up, by following their curiosity down the various paths of investigation that the new tools have opened. It is unexplored territory.”

Issues for Science and Engineering Researchers in the Digital Age

A report of the National Academy of Sciences (2001)



# An Example: Vanderbilt



*This is not your father's University Computer Center...*

- **Investigator Driven:** maintain a grassroots, bottom-up facility operated by and for Vanderbilt faculty.
- **Application Oriented:** emphasize the *application* of computational resources to important questions in the diverse disciplines of Vanderbilt researchers;
- **Low Barriers:** provide computational services w/ low barriers to participation;
- **Expand the Paradigm:** work with members of the Vanderbilt community to find new and innovative ways to use computing in the humanities, arts, and education;
- **Promote Community:** foster an interacting community of researchers and campus culture that promotes and supports the use of computational tools.

**\$8.3M in Seed Money from the University (Oct 2003)**

**\$1.8M in external funding so far this year**

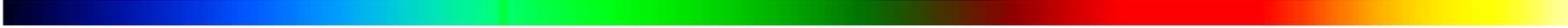
# ACCRE: Investigator Driven

A grassroots, bottom-up project by and for Vanderbilt faculty.



**~76 Active Investigators, 10 Departments, 4 Schools**

# ACCRE Components



- Storage & Backup
- Visualization
- Compute Resources (more in a second)
- Educational Program
  - Establish Scientific Computing Undergraduate Minor and Graduate Certificate programs.
- Pilot Grants for Hardware and Students
  - Allow novice users to gain necessary expertise; compete for funding.
  - See example on next slide...

# Multi-Agent Simulation of Adaptive Supply Networks



- Professor David Dilts, Owen School of Management
- Large-scale distributed “Sim City” approach to growing, complex, adaptive supply networks (such as in the auto industry).
  - “Supply network are complex adaptive systems...”
  - Each firm in the network behaves as a discrete autonomous entity, capable of intelligent, adaptive behavior...
  - Interestingly, these autonomous entities collectively gather to form competitive networks.
  - What are the rules that govern such collective actions from independent decisions? How do networks (collective group of firms) grow and evolve with time?”

# ACCRE Compute Resources



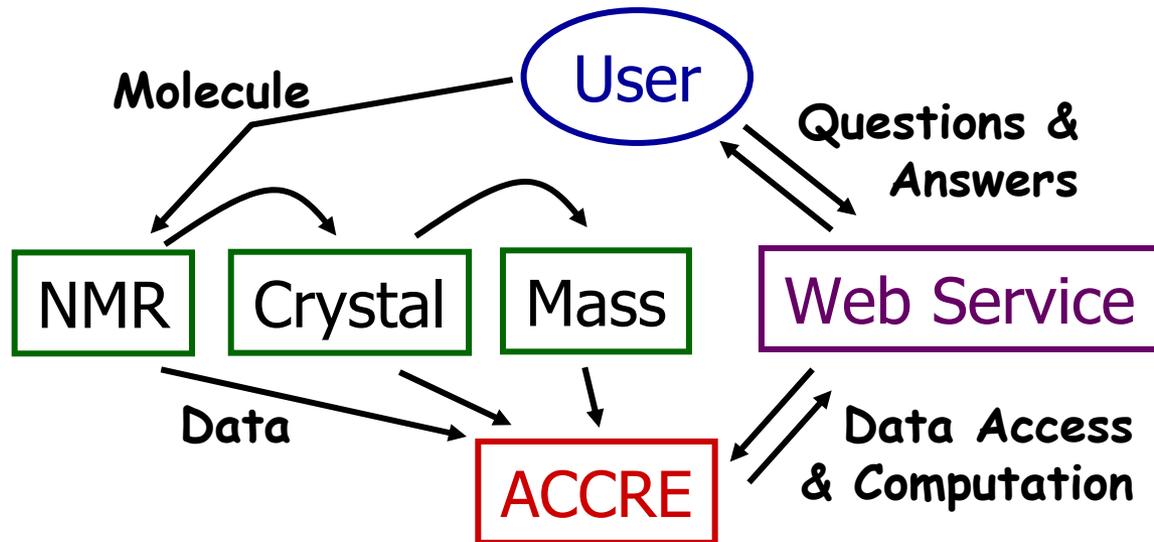
- Eventual cluster size (estimate): 2000 CPUs
  - Use fat tree architecture (interconnected sub-clusters).
- Plan is to replace 1/3 of the CPUs each year
  - Old hardware removed from cluster when maintenance time/cost exceeds benefit
- 2 types of nodes depending on application:
  - Loosely-coupled: Tasks are inherently single CPU. Just lots of them! Use commodity networking to interconnect these nodes.
  - Tightly-coupled: Job too large for a single machine. Use high-performance interconnects, such as Myrinet.
- Actual user demand will determine:
  - numbers of CPUs purchased
  - relative fraction of the 2 types (loosely-coupled vs. tightly-coupled)

# A New Breed of User



- Medical Center / Biologist
- Generating lots of data
  - Some can generate a Terabyte/day
  - Currently have no good place/method currently to store it...
  - They develop simple analysis models, and then can't go back and re-run when they want to make a change because their data is too hard to access, etc.
- These are small, single investigator projects. They don't have the **time**, **inclination**, or **personnel** to devote to figuring out what to do (how to store the data properly, how to build the interface to analyze it multiple times, etc.)

# User Services Model



- User has a biological molecule he wants to understand
- Campus “Facilities” will analyze it (NMR, crystallography, mass spectrometer,...)
- Facilities store data at ACCRE, give User an “access code”
- ACCRE created Web Service allows user to access and analyze his data, then ask new questions and repeat...

# ...Initial BTeV Grid Activities

- **Storage development with Fermilab, DESY (OSG)**
  - Packaging the Fermilab ENSTORE program (tape library interface)
    - Taking out site dependencies
    - Documentation and Installation scripts / documentation
    - Using on two **ACCRES** tape libraries
  - Adding functionality to dCache (DESY)
  - **ACCRES** using dCache/ENSTORE for HSM, once complete will be used by medical center and other Vanderbilt researchers
    - Developing in-house expertise for future OSG storage development work.

[Talked about this earlier]

# Conclusions



- BTeV needs the Grid: it is a Petascale experiment with widely distributed resources and users
- BTeV plans to take advantage of the growing cyberinfrastructure at Universities, etc.
- BTeV plans to use the Grid aggressively in its online system: a quasi real-time Grid
- BTeV's Grid efforts are in their infancy: as is development of their offline (and online) analysis software framework
- Now is the time to join this effort! **Build this Grid with your vision and hard work.** Two jobs at Vanderbilt:
  - Postdoc/research faculty, CS or Physics, working on Grid
  - Postdoc in physics working on analysis framework and Grid