WestGrid Utilization and Experience

Why TWIST needs a supercomputer

- To obtain a precision of 10^{-3} on determining the Michel parameters, we have determined that we need to look at a dataset of $2x10^8$ events.
- On a 2.4GHz Pentium, original estimates:
  - Simulating one event : 50 ms
  - Analysing one event: 100 ms
- A complete cycle on one event means analysing the raw data, simulating this data and analysing the simulated data for a total of 250 ms or 580 days for $2x10^8$ events (we call this one data set).
- And we need to study many different systematic effects with the full statistics.

The TWIST Triumf cluster

- TWIST has one “cluster” of 15 dual AMD CPUs equivalent to 2.4 MHz Pentium.
- These 30 CPUs can process **one** data set in 20 days. We have 5 grad students that need to work on different studies.
- With only our cluster we could only do limited data set analysis, good enough to find gross problems and improve the software.

GLACIER
The UBC/TRIUMF WestGrid cluster

WestGrid consists of three major compute intensive facilities with different emphasis

- Nexus in Edmonton: 256 proc SGI 3900
- Lattice in Calgary: 144 proc HP(alpha) SC45
- Glacier in Vancouver: 504 dual IBM 3GHz Xeon

Plus a storage facility at Simon Fraser University.

TWIST was chosen as a Beta user for Glacier along with D0 and a group of math & chemistry researchers.

GLACIER main components

- Ice - Compute nodes: dual 3.06 GHz Xeon on blades, 2GB /blade, 80GB disk/blade; 14 blades per crate; 4 x 1 Gb Ethernet channels per crate; 36 crates for a total of 1008 CPUs.
- Nunatak - 3 head nodes for user access, compilation and job submission.
- Moraine – 4 NFS file servers for the large 10 TB disk array split between /global/home (2.25 TB) and /global/scratch (5.75 TB)
- Tivoli – a tape archiving system with robot good for 30 TB.

How TWIST uses GLACIER

- Computing on the nodes is only possible via job submission. Open PBS is the queuing system coupled with MAUI scheduling. It is setup to use the “fair share” scheduling algorithms.
- For MC generation, the data produced is written out to the compute nodes local scratch space.
- MC data is analysed directly from the files on the local nodes. When analysis is finished, data is moved to global scratch and archived.
- Raw data is staged from tape to disk on the TWIST file server located at TRIUMF.
- Analysis produces ROOT trees (similar to HBOOK Ntuples) in the global scratch area. These are returned to the TWIST file server for later analysis.

The TWIST file server

- With the NSERC computing money left over after setting up the TWIST cluster, a 3 TB file server machine has been purchased.
- It has 2 GBit Ethernet links, one for connection to the TWIST cluster, the other connected to the main TRIUMF router.
- In turn the TRIUMF router has a direct GBit Ethernet connection to the BC WestGrid switch

GLACIER
Work in progress

- The purchase order was signed at the end of June 2003.
- Location was chosen in early July at the UBC computing center. Important renovations were needed mainly for AC and power distribution.
- Renovations done by early September but insufficient AC
- Main pieces of hardware had arrived but installation was slow. Main problem: power supplies were not sufficient to power all blades; 144 blades are taken out of service.
- Access was opened to the Beta users in October
- TWIST software was ported and tested on the head nodes by Oct 20th

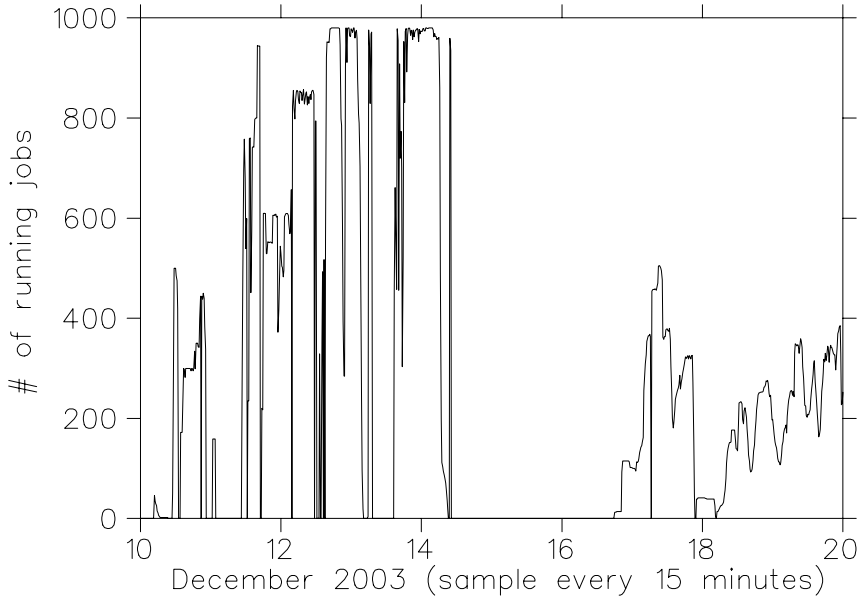
GLACIER
Work in progress (2)

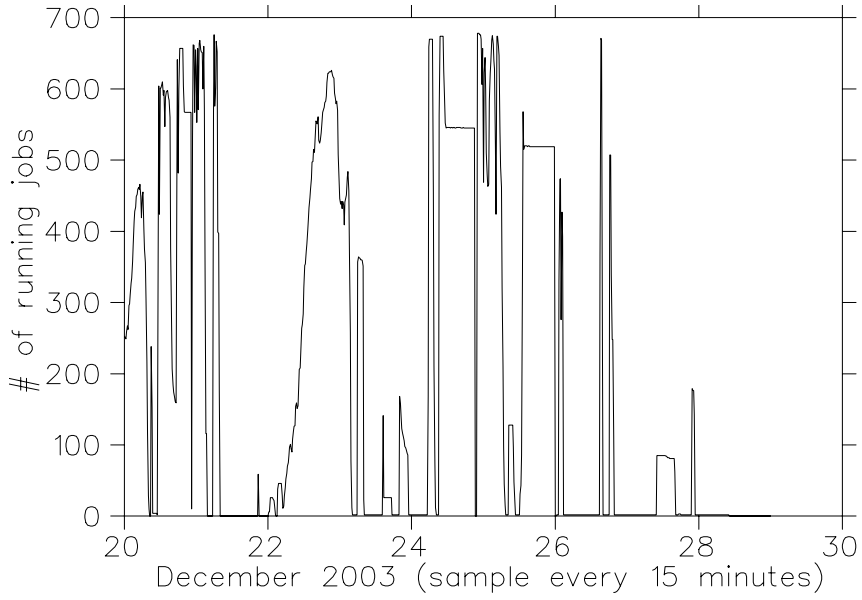
- In early November, TWIST tweaked it's script and started submitting jobs.
- There were lots of problems with the file systems which meant that we lost a lot of jobs and had to write intelligent scripts to redo only the lost jobs. The archiving system has been on/off which creates space limitations for us.
- There were major AC failure and power failure.
- Finally all the hardware issues are resolved and there is enough redundant power for all nodes.
- But there is still a major problem with NFS freezing (on average twice a day) which results in job losts. IBM has been extremely slow in responding to fix the problem.

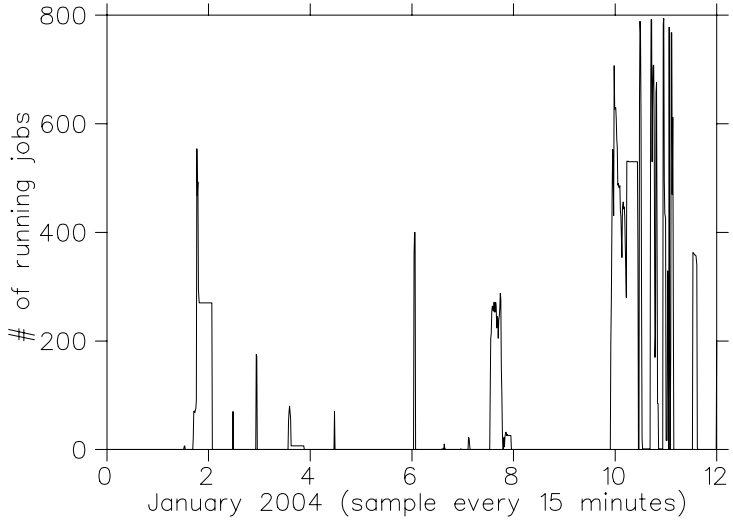
Statistics

- Geant MC data produced: 50 x10^8 events in 51 sets.
- Geant MC processing time: .07 s/event
  - **(4051 CPU days)**
- Raw data transferred: 30 x10^8 events in 15 sets
- # of raw data analysis performed: 32
- Analysis processing time: .025 s/event
  - **(3182 CPU days)**
- Data archived so far:
  - Geant MC data : 5.7 TB
  - Raw data : 6.7 TB
  - ROOT trees (results of analysis): 1.1 TB
  - Total: 13.5 TB

TWIST usage of Glacier in last 30 days







GLACIER in production mode

WestGrid has created a resource allocation committee (RAC) with one representative per institution and chaired by Mike Vetterli. When Glacier opens for production there will be two queues:

- 10% of resources shared by all registered users
- 90% of resources allocated by RAC based on needs. Twist hopes to get a 15% share of compute nodes and use of 50% of local scratch and 25% of global scratch.

Summary

- The power of GLACIER has been made available to TWIST. The cooperation of the WestGrid people has been excellent.
- Despite NFS problems, the performance allows rapid analysis of TWIST data. It makes it possible to think of going to a five fold increase in statistics in the future.