Initial experience with the use of TeraGrid for High Energy Physics research

Andrey Shevel, Roy Lacey

Computing infrastructure in small physics group

- Local cluster (usually small or middle range);
- Remote cluster (or clusters);
- Is it possible to use arbitrary computing cluster with enough computing power?

Distributed computing parameters in our case

- Max number of the computing clusters is about 10.
- Max number of the submitted at the same time Grid jobs is about 2*10**3 or less.
- The amount of the data to be transferred (between group local cluster and remote cluster) for physics analysis is centered about 2 TB/quarter +- 50%.
- We use local cluster file catalog based on slightly re-designed version MAGDA

  http://ram3.chem.sunysb.edu/magdaf/.

Estimates

Let us introduce the variables:

T - total time for the computing task with using only local cluster;

τ - reduced time for the computing with using additional cluster;

t_l - average time for processing of one portion of the data on the local cluster;

t_r - average time which is required to process the one portion of the data on remote cluster;

t_o - the average overhead time which is required to perform any additional operations (for example time for the data transfer) per one portion of the data on remote cluster;

D - total number of the data units which have to be processed;

S - speed: the number of the data portions at one unit of the time;

α - accelerating (speeding up) of the computing (in times) due to use additional cluster;

- for only local cluster:

$$S = \frac{1}{t_l}$$

and total time for the computing is

$$T = \frac{D}{S} = D * t_l$$

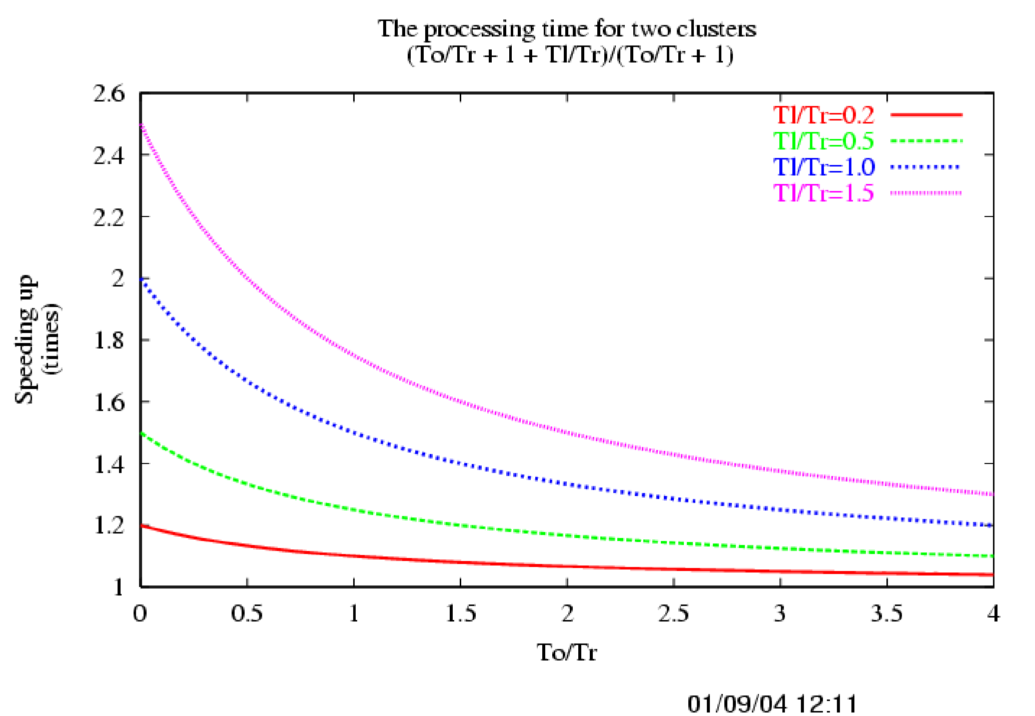
- for two clusters (local + remote)

$$S = \frac{1}{t_l} + \frac{1}{(t_o + t_r)}$$

and total time for the computing is

$$\tau = \frac{D}{(\frac{1}{t_l} + \frac{1}{(t_o+t_r)})}$$

$$\alpha = \frac{\frac{t_o}{t_r} + 1 + \frac{t_l}{t_r}}{\frac{t_o}{t_r} + 1}$$



01/09/04 12:11

Conditions of multi cluster environment

- Computing Clusters have different:
  - computing power;
  - batch job schedulers;
  - details of administrative rules.
- Computing Clusters have common:
  - OS Linux (there are clusters with different Linux versions);
  - Most of clusters have gateways with Globus toolkit.

Porting the application software to run on remote cluster

- The porting of application physics software in binary form is presumably most easiest port method:
  - copying over AFS to mirror directory structure on remote cluster (by cron job);
  - preparing PACMAN packages for specific class of tasks (e.g. specific simulation).

User job scripts

- One of the important issues for multi cluster environment is to make sure that all user scripts run on all available clusters with same results;
- Simple way to do that:
  - to eliminate from the existing user job scripts all specifics and make them neutral to the cluster environment (to do that we introduced pre assigned environment variables like
    - GRIDVM_JOB_SUBMIT, GRIDVM_BATCH_SYSTEM, etc.

Terminology

- *Typical user* means physicist who has no administrative privileges on any cluster and is not an expert in computing architectures.
- *Qualified computing cluster* means the cluster which does fit specific requirements: available disk space, installed software, etc.
- *Major data sets* means physics data (mostly reasonable volume from 100’s of GB and much more).
- *Minor data sets* means set of scripts, parameters, etc (mostly small volumes may be less than 100 MB or so).
- *Master job* (script) which (is submitted by user) will submit so called job production (many *sub-jobs)* with default (for remote cluster) batch system.
- *Parent-job* means any job which submits sub-jobs.
- *Sub-job* means the jobs which were submitted by *master job* or *parent-job*.
- Generally speaking the characteristics (*parent* and *sub*) are relative. The *sub-job* might be observed as *parent* to the submitted sub-*sub-job*.

The job submission scenario at remote Grid cluster

- User needs just *qualified computing cluster*: i.e. with enough available disk space, specific version for compiler and related software pieces.
- Before job production:
  - To copy/replicate the *major data sets* (physics data – several TBs) to remote cluster.
  - To copy/replicate the *minor data sets* (scripts, parameters, etc. – about 1 GB) to remote cluster.
  - To guarantee that remote cluster with required environment is functioning properly by the set of test tools (scripts).
- To start the *master job* (script) which will submit many *sub-jobs* with default (for remote cluster) batch system. Also master job might deploy the software components required for *sub-jobs.*
- During production (job run) and after all jobs were accomplished (after job production):
  - To watch the jobs with monitoring system.
  - To copy the result data from remote cluster to target destination (desktop or RCF).

TeraGrid

- TeraGrid project serial number (PSN) is ‘TG-PHY050000T’;
- TeraGrid is large organization with many powerful clusters (http://www.teragrid.org).
- Initially we get the access to the cluster
  - *tg-login1.ncsa.teragrid.org*;
  - hardware architecture – Intel Itanium 2, IA64, 1.3 GHz;
  - software platform: SuSE SLES 8 (powered by UnitedLinux 1.0) (ia64);
  - data transfer for end user (between rserver1.i2net.sunysb.edu and tg-login1.ncsa.teragrid.org) speed is varied from 1 to 5 MB/sec;
- TeraGrid has Mass Storage System (MSS) to keep large volume of data (many TBs)
  - uberftp mss.ncsa.uiuc.edu
  - Average data transfer speed is about 11 MB/sec.

TegaGrid cluster types

- After quick testing we have found out that this platform is not what we need especially if we plan to save efforts to port our software.
- We formulated more correctly what we would need:
  - Intel Pentium Xeon 3 or 4, IA 32; main memory about 1 GB;
  - OS type = Linux RedHat (preferably Scientific Linux);
  - 2-4 TB of disk space as minimum.

TeraGrid help and other features

- TeraGrid help is available by mail help@tergrid.org. The tech answer with the question number we have received quite fast (in minutes). But more consistent info about the matter of the question might come during days or even weeks.
- In one day MSS became unavailable for normal operation and we got a lot of errors, but the information about the upgrade done on MSS was received only two days or more later. It is not bad to have the mailing list with information about such the events like upgrade, etc.

Summary

- The Teragrid is large scale computing facility which is available for researchers.
- The Teragrid consists of many different clusters with different hardware platforms, HPSSs, OSes, other important features.
- The detailed specification is required to choose the right cluster .
- TeraGrid testing for HEP application has to be continued.