GRID TALK TRANSCRIPT 2010.1

ABSTRACT

This talk will be aimed at an audience with no prior knowledge of any form of grid or distributed computing. It will give a short but concise description of what grid computing is, how it came into form and why; along with the different terminologies/concepts used with current example uses. The second half of the talk will focus on the resources situated within Middlesex University and a demonstration will be given in using these facilities, this is followed by a conclusion of plans to extend its capability. One certain aspect of future growth will be the joining of Middlesex resources to the NGS (National Grid Service), thus a talk on what they are the benefits of joining will also be included.

NS

Hello my name is Vijay, and I work in with the HSSC and EIS to establish Grid computing facilities here at Middlesex. I am here to give you a brief review of what grids is, and what the PLAN is!

NS

What is grid computing well a formal definition is, above…?

The key words here are ‘access to large amounts of data and processing cycles’, as these were the main driving force behind the creation of grid computing.

Also we are given an example of SETI, which basically harvest idle CPU on a volunteer basis to help search for ET.

Note that this is a classical one-way service in that the user can not actually use the SETI grid but can only provide a service.

This definition was taken from IBM; it seems to be a bit old as he claims that Grids will be the next ‘big thing’. I say old because as the concepts of a grid have undergone a branding some what into the term could-computing, more on this later.

NS

This is just a basic overview but we will cover in this talk.

- Introduction to Grid Computing
  - The Need for Grids
  - Concepts
  - Examples of real life use
  - Current status

- What the NGS is.

- Future plans for Grid computing at Middlesex and Grids in general.

NS

What are computers…?

Well put simply they are computational machines- basically maths and their evolution has followed Moore's law.

It stats the every year the capability of computers would double with every year, and it has held true so far…

As you can see some of the latest computers are breaking this trend and exceeding the doubling norm (due to multi core cpus).

Nowadays computers are used to model just about anything.

So much so that it isn't a rare for simulations to take many months to run on a single computer.

Just imagine waiting a year and then finding out you forgot to convert your inches into mm.

Off loading this to a collection of dedicated might be a good idea.

NS

Unfortunately processing power is a bit moreish and people are demanding more of it.

To date the fastest computer in the world is IBM's roadrunner.

It can operate a maximum of Peta FLOPS 10^{15} when compared to a modern quad core CPU operating at around 70 Giga FLOPS 10^{9}.

Making it around 14300 times faster than a top of the line quad core CPU.

Those asides it only costs $133 million, there is a small over head in operating cost, 2.3MW will most likely mean you need to add the operating cost of a small nuclear power station to your budget.

It is a small wonder why they don't let many people experiment with it!

NS

Ok that was an extreme case but even with a road runner on your side some simulation still take weeks if not months to do.

Any way before I go onto grid computing it is important to know about something called clustered computing.

Basically this is joining more than one computer on a local network to serve as one unit.

And this is basically what grid computing is, but over the internet!

However this new found computing resource has basically given rise to two kinds of computing you are most likely to do, and usually comes in 2 flavours…

High-performance computing (HPC)
The main aim here is getting the maximum use of processing power, and to do tasks as quickly as possible.

High-throughput computing
here time is not the main thing here, CPU speed is not important as much as processing lots and lots of data, with jobs existing over long-time scale

The latter is more or less the initial aim of grid computing.

NS

Now with some examples of uses of Grid computing in the biological sense.

Folding at home.

Similar to SETI, Folding at home uses around 250,000 computers to solve protein folding problems, to help research for related diseases like Parkinson's and BSE.

Another is BLAST…

Basic Local Alignment Search Tool, BLAST is a tool that helps find gene sequences, it is traditionally a stand alone application but users have exported it onto a grid environment to enable multiple parallel searches.

Both examples are a classical case of an embarrassingly parallel problems, basically since these tasks are made up of many smaller independent jobs, virtually no work is needed to port these problems onto a grid environment.

NS

However Grids were created by the Physics dept who quiet often had to sieve through and process PBytes of data 1000 TB or 1mill GB.

You may have heard of the LHC,

(well it's a 27 km long circular particle accelerator in CERN, Geneva, Switzerland)

Well they alone produce around 1PB of a year to the best of my belief. And has been one the main driving forces for the conception and creation of grids.

The sheer size of this task and its solving has produced a whole subject of its own Grid Computing, which is built on many other subjects it uses/ derived.

NS

READ THE LIST

NS

Uploading and Downloading large data is mostly done with a program called Grid FTP. This allows for parallel uploads in a grid environment.

Something called data mining is an active research field, basically allowing a program to search data bases for objects of interest.

NS

Due to the nature of Grids they are best suited for High through put computing, known as batch processing.

Existing schedulers like Condor aren't grid enabled, but usually serve as the backbone to cluster of nodes at the end of the line.

A layer in between the grid and your condor is needed to grid-iphy your resources.

Talking Globus as an example this software is called GRAM.

This manages the smooth runnings of your jobs, transfer of data (staging) and notifications of completion.

NS

Services on offer and their reliability go hand in hand in that accurate monitoring of each is a solution to both problems.

Sticking to the Globus style, is monitoring and discovery service is their solution.

Each node produces their own monitoring information, which is pushed up to a sub-master index node, which is pushed up again to the main master index, query-able by the grid.

This information is valuable and can be used to assist work flow manager to create and execute complex interdependent tasks.

NS

Deciding who to trust, and prevent identity fraud something call digital certificates have been invented.

These are like digital passports which are tied to your web browser and allow you to access the grid.

NS

In order for all the computers to talk to each other a protocol called SOAP is used which is the basis for Web Services.

It is a plain text language that can be viewed with notepad. This also allows for grid to be used through web pages, and so job submitted are also in this plan text.

But some users still like the command line interface and still exist.

NGS...

NS...

As some of you can see the NGS is a free academic Grid infrastructure, made up of many universities.

There are actually about three times more than what I've got up here; I just didn't have the space to put them all there!

Also this is only a list of universities that provide resources, there are many, many users (I can't recall how many)

But the fact that an NGS certificate will allow access to American Grid is nice extra.

NS...

To be truthful we don't have the man power or money to run a sufficient grid, if its uptake does take off.

Basically as the slides says, there already exist people willing to do these task at no charge to the user.

You can take your research with you.

Leach? We can, nothing wrong with that is it expensive to run a grid. But if you want to have service you would like to have on the grid (obviously there's C++/Java)...

i.e. an application whether free or not, convincing other universities to put it on their Grid is hard and you present a strong case! Where as you local university will grudgingly make an attempt.

NS...

Read the slide.

These are not the only fields using the Grid, since applications such as mat lab/ octave/ C++ easily allow other fields to use the Grid without getting recognition!

NS...

CosmoMC is an Astro-physics application, there's nothing more I can say about this, but you will notice the key word Monte-Carlo.

NS...

Its seems the bio-informatics people have got the ball rolling on using the grid, being second the physics people in terms of access to raw data. Honestly there are tons of little apps so I've cheated and copied the web link!

NS...

Poor mouse, and again with the Monte-Carlo! Radioactive tracer! Used!

GATE software allows to solve complex PET data to produce accurate images.

This proves that grid can be used for image / tomography generation/ reconstruction!

NS...

Not much need to said here, it a particle simulator, is guess there's an irony that you'll need something as large as a grid to solve something so small.

NS...

For anyone who uses allot of data that bears slight relations to each other, these tools in machine learning/ data mining is worth looking at, as this aims to remove the human intervention to make links between these data sets.

NS...

Abacus (I know it's spelt with a q!), a FEM solver- what are they? Just shows that complex tight nit programs can be put on the Grid and given the demand- even commercial ones...

NS...

Blender, apparently this was a commercial application that went free; this has many applications in the multimedia field! It's not all command line.

NS...

Stress that grid users dictate what is on offer.

In reality, the grid should be as versatile as your normal desktop computer, so long as it can run on a computer, it can run on a grid.

Also if it is repetitive than you can almost certainly increase performance too.

NS

Getting a certificate is easy; now that I have been appointed the Certificate verifier for Middlesex, Just go to the website.

Click on Use NGS, How to join and Certificate Home.

NS

Click ‘apply for certificate’.

NS

Click ‘apply for personal certificate’.

NS

Click ‘CA web certificate’.

NS

You may be warned of a security risk, just allow the exception.

NS

This verifies the addition.

NS

Click ‘Request a certificate’

NS

Click ‘User Certificate’

NS

Add another exception

NS

Then fill in your details, chose Middlesex HSSC, and create a 10 digit pin number. (You will need to tell me this when you show me you student/staff ID)

NS

Another Exception

NS

Verify Everything and then click continue.

NS

Like a receipt print of a copy for the record.

NS

You will also get an email confirming application, and to instruct you to provide the photo ID.

NS

Once you ID is confirmed a certificate will be created. The NGS will send you an email with a link to add the certificate to your browser. This has to be the same one you applied in.

There are instructions on how to backup your certificate and move it to another computer.

IE in Windows 7 doesn't seem to work, Fire Fox is best recommended.

NS

How to run job. TO BE DONE!