CHEP’01 Summary - Networking and Grid Computing

Track 7: Local and wide area networking

Track 10:Grid Computing and Toolkits,packages,developments,benchmarks

Local and wide area networking (1 of 3)

- Two driving topics and one paper looking forward to IPV6
  - high throughput WAN performance with five papers from SLAC, LBL and KEK
  - security with three papers from KEK, FNAL and IHEP

- WAN performance talks all about how to get the best out of existing systems which all have bad defaults. They emphasised the gap between ordinary and ‘wizard’ user who can get 5 to 50 times more throughput in WAN transfers. Users should be able to achieve 100Gbytes/day today.

- Best performance by large tcp window sizes, multiple streams and big files. Windows are basically buffer sizes to collect packets at each end. Eg Solaris default is 8KB so with 20 streams SLAC to IN2P3 runs at 5 Mbits/sec while with 512MB windows runs at 95 Mbits/sec. Need root at both ends to change! The plea was that systems should tune themselves e.g. linux 2.4 will grow its window size for you (though not by enough).

Local and wide area networking (2 of 3)

- Performance seen today on research nets is quite good with SLAC to Caltech running at 500Mbits/sec but only 350Mbits/sec over production Internet.Transatlantic production performance is 100-120 Mbits/sec.

- Good talk from Hanuchevsky (SLAC) on the bbcp (babar wide area copy) tool. Has familiar syntax and nice useability features:
  - auto-resume after failure (only uncopied parts resumed)
  - preserves group ownership and time stamps and auto-creates directories
  - can specify file that is a list of files to copy
  - has history log file and gives periodic progress messages
  - has rate limiting function (eg this can be background), checksums and compression

- Another good talk was from KEK where they have built a software/hardware network simulator using a PC as a router between two small Gigabit ethernet clusters. They can vary round trip times, loss percentages, bandwidth, window sizes etc and measure throughput and cpu load on the sources and sinks etc. Gives good comparisson with the real world and seemed a very useful tool.

Local and wide area networking (3 of 3)

- Security talks were less interesting but all sites recognise its increasing importance - all will use standard tools of firewalls, COPS, CRACK, SATAN etc. FNAL spoke again of their move to strong authentication with one sign-on for all services. Based on modified MIT Kerberos 5 or, where not possible, one time passwords eg via smart cards. Full deployment planned by end of this year.

- IPV6 was a tutorial. Was agreed in 1994 to go from 32 to 128 bits address space and allow stateless auto-configuration with the ability to rapidly renumber a site. 3rd generation mobile phone developers are likely to be the first users so the first stage will be upgrading carrier networks. Europe/Asia, where there is a shortage of address space, will probably be before the US.

Grid Computing (1 of 4)

- 9 status reports on first day (I attended networking)
- 4 talks on basic grid tools and one on CMS requirements
- 7 talks on progress in middleware
- Basic tools:

  – Ian Foster gave very general overview of Globus software and perspectives in all areas - architecture, transport, replication, resource discovery, security. Basically steady progress with no major architecture problems.

  – More interesting he reported detailed work on dynamic replication strategies where they have compared performance between no replication, caching that is plain (each request makes local copy to client so fast aging out), to best client (history of busy files maintained for each client) , or cascading (make copy of busy files down to next lowest level), combined caching + cascading or fast spread (each node on the way to a client also gets a copy). The performance metrics were average response time (for a client node to have a copy of a file) and the total bandwidth required to accomplish this. Fast spread was found to be best in the case of total randomness in access patterns while cascading was best in the case of geographic locality of access.

Grid Computing (2 of 4)

- I gave B.Tierney talk comparing rfio and gsiftp over WAN. Gsiftp is the high performance (supports multiple streams) secure file transfer protocol part of the Globus toolkit. Same results as for other WAN talks regarding window sizes, parallel streams and file sizes. gsiftp reaches much better performance than rfio for intermediate file sizes (5-50MB) but is only a little better for files ge 100MB.

- Nice talk from INFN on a practical evaluation of Globus. Recommends improvements needed in packaging, performance, security and access control and registration. All being addressed. GSI security model found to be satisfactory but needs tools for group management.

- Koen Holtman of Caltech gave a much appreciated talk on the CMS requirements for the Grid. Between December 2000 and July 2001 CMS conducted a major requirements and consensus building effort resulting in a series of requirements documents for the GriPhyN, PPDG and EU DataGrid projects in which CMS is involved as a ‘customer’. It specifies the software components that should be delivered to CMS for it to build a Data Grid system it intends to operate from December 2003 and gives a comprehensive overview of this system.

Grid Computing (3 of 4)

- Middleware
  - report from INFN on the EU WP1 work integrating grid tools to build a computing resource broker. Very thorough - data flow clear between components of user interface, logging and bookeeping, resource broker, computing elements, storage elements, replica catalog and information services.
  - Report from R.McClatchey on query and analysis processing models
    - physicist develops and registers algorithm and submits query locally
    - query handler decomposes it and locates data
    - algorithms are executed where their data resides and results are returned to query handler for presentation to the physicist
  - J.Legrand reported an interesting prototype agent-based system to gather, disseminate and coordinate configuration and time dependent state information in the Grid system as a whole. Written in Java using the JINI support for distributed applications. JINI seems a very productive tool but with an unknown future.
  - Rademakers presented the parallel root facility (PROOF) for parallel interactive analysis across clusters of heterogeneous computers. This was a very practical implementation of parallelism where the user root client session creates a master server on the remote cluster and this in turn creates slave servers. Still in a prototype phase but looks promising with good real time reductions for analysis queries across a given number of dedicated nodes.

Grid Computing (4 of 4)

- Morita reported on a KEK project called GFARM, a grid middleware project to solve peta-byte scale data intensive analysis. Large logical files are distributed across multiple PCs. A user job runs in parallel on the nodes where a fragment of the data file resides with intercommunication using a lightweight rpc mechanism and special gfarm daemons running on each participating node.

- From Liverpool Patel reported on an extension of their existing Mont-Carlo Array processor to make it Grid aware. The MAP system consists of 300 commodity PCs plus 6 master nodes. Like the KEK project a user job is transparently run in parallel over all nodes allowing for small differences in initialisation parameters. They are now building G-MAP, a Grid aware MAP control software allowing remote job preparation and submission using the Globus toolkit for authentication and communication. Another very promising project.

- Finally IHEP reported on their local implementation of farm monitoring with a highly flexible and portable cluster monitoring system using a dynamic class loading technique and a publish/unpublish and subscribe protocol and directory service.

Grid Computing - Conclusions

- activities are still mainly concentrated on strategies, architectures and tests
- there is general adoption of the Globus layered architecture and basic services
- new middleware tools are starting to appear and be used but there are some parallel developments so strong coordination will be needed to avoid divergent solutions (especially between continents and different scientific fields)
- we need to plan carefully the next iteration of Grid middleware development
- there is in general a good collaboration between the existing EU and US Grid projects
- experiments are getting on top of Grid activities
- finally -the HEP Grid infancy begun at the Padova CHEP conference has now ended at CHEP 2001 as evidenced by the recent creation of the LHC Computing Grid project.