Networking for the Grid: problems and solutions

http://www.grid.ucl.ac.uk/

Saleem Bhatti
Director, UCL e-Science Centre of Excellence
Networks Research Group
http://nrg.cs.ucl.ac.uk/
UCL Grid/HPC - CoE

- http://www.grid.ucl.ac.uk/
- Many projects - UK and EU funding:
  - RealityGrid, EGSO, DataTAG, e-Protein, etc.
  - GRS, MB-NG, UKLIGHT, 46PaQ
- e-Science/Grid Centre of Excellence (CoE) in Networked Systems:
  - http://www.grid.ucl.ac.uk/NETSYS.html
  - high-speed networking, QoS and traffic engineering, performance, network resource control/management, protocol enhancements and evolution, security, complex systems, monitoring and reporting
Funding
Funding, partners, collaborations
Overall scenario (outline)

I want to run X with input data at A and would like the results sent to B by Monday evening.

harmonised CPU + network scheduling

GRS work: network scheduling only

I can do X

I can do X

I can do X
Real example [1]

- ~5000 particles falling onto a surface
- All collisions taken into account in the model.
- Forget the physics - think of the work involved!
- The real models involve ~1,000,000 particles!

thanks to S.Sorensen@cs.ucl.ac.uk
Real example [2]

- ~5,000 particles falling onto a surface
- 18 processors are used in this example
- Processors are colour coded
- Observe colour changes as objects change their “home”

thanks to S.Sorensen@cs.ucl.ac.uk
The Grid networking problem

- Data intensive Grid computing:
  - data Grids vs. computational Grids
  - could be both data and compute intensive
- Data points to highlight the problem:
  - LHC, VLBI: multi Gb/s ($10^9$) to multi Tb/s ($10^{12}$)
  - distribution of data and processing (CPU usage)
  - 33MHz, 32bit PCI ≈ 1Gb/s (reality: ~50% of this)
  - TCP - problems on long delay, high rate links
- Data has to get across net fast - but can’t!
- But what if everyone starts doing this?
- Networking is global, end-to-end problem!
Big data … big problems!

• Particularly relevant to Grid/e-Science
• User in Glasgow wants to access the HGP data
• HGP database:
  • 0.3PB (growing at ~1TB/week)
• SuperJANET4 (SJ4):
  • 10Gb/s backbone (still <2.5Gb/s access in places)
• Extreme case – transfer all of the HGP data
• So, iff user gets all the SJ4 backbone capacity:
  • transfer of HGP data still takes over 55 hours!
  • no one else can use the network at all during this time
• Can’t do it! 😞
Problem: network hierarchy

- **Access network:**
  - low multiplexing
  - low volume of traffic

- **Distribution network:**
  - local level connectivity
  - low multiplexing
  - medium volume of traffic

- **Core network – backbone:**
  - high volume of traffic
  - high multiplexing

- **Different administrative domains**
Problem: administrative domains

- Network QoS reservations require *state* to be set-up, stored, maintained
- State information:
  - what?
  - where?
  - when?
  - how much?
- General problems:
  - signalling
  - scaling
  - (accounting + charging)

A ↔ B: localised scope
A ↔ C: non-localised scope
Problem: mixing traffic

- Example – voice, FTP and WWW traffic through a router:
  - 3 input lines: serviced FCFS at a router
  - 1 output line (1 output buffer)
Problem: modelling traffic

- Poisson Model used for computational convenience, not for accuracy!
  http://math.bu.edu/people/murad/pub/source-printed-version-posted.ps
  http://www.aciri.org/floyd/papers/wsc97.ps
Problem: network traffic profiles
So what can we do about it?

- **Build a new and better network (of course)!**
  - ... well ... at least the core
  - very high capacity ($\text{Gb/s} \rightarrow \text{Tb/s} \rightarrow \text{Pb/s} \rightarrow \text{Eb/s}$)
  - users can have access from their desktop
  - provide (QoS-)controlled access
- **Two broad problems to consider:**
  - **control**: how do we mix different types of traffic and still control the traffic flows in the network sensibly?
  - **capacity**: what happens when you run a very high capacity network with very high capacity access links?
- This talk highlights some of the **research** issues:
  - there are also **operational** issues! (but that’s SEP 😊)
Problem space - networks [1]

from http://www.ja.net/
Problem space - networks [2]

from http://www.ja.net/
Problem space - networks [3]

from http://www.dante.net/geant/
UKLIGHT - networking research

- High-speed networking research:
  - no production/service traffic
  - high-speed optical
  - ~£4.6M from HEFCE
- http://www.ja.net/development/UKLight/
- Connectivity to other national high-speed networks:
  - global research infrastructure
- UK/UKERNA founding member of GLIF:
  - http://www.glif.is/
UKLIGHT connectivity

Global Participation: GLIF at StarLight

Source: Kees Neggers, SURFnet

from http://www.glif.is/
Project links

MB-NG - http://www.mb-ng.net/
GRS - http://www.cs.ucl.ac.uk/research/grs/
UKLIGHT - http://www.ja.net/development/UKLight/
EGEE - http://public.eu-egee.org/
GEANT - http://www.dante.net/geant/
46PaQ - TBA
Project links - info

- **MB-NG:**
  - core network: capacity + QoS
- **GRS:**
  - edge-edge/site-site QoS control
- **46PaQ:**
  - performance and QoS monitoring
- **EGEE/GEANT:**
  - international Grid connectivity
- **UKLIGHT:**
  - international high-speed networking research
GRS project outline [1]

- Mar 2002 – Sep 2004
- 3 Phases:
  - 3 incremental development phases
  - currently at **Phase 2**
GRS project outline [2]

• Architecture for dynamically configurable network reservations system
• Micro-management of flows at sites:
  • in this case DIFFSERV aggregates
• Focus on state management and signalling
• Assume DIFFSERV network (for now):
  • architecture will not be restricted to DIFFSERV
  • assume BE and EF per-hop behaviours
• Initial development on Linux (using tc):
  • architecture not restricted to Linux
  • current work-in-progress to port to CiscoIOS
Outline architecture

application

requests

“scheduling harmonisation middleware”

requests

(to CPU scheduling)

NRSE

data packets

configure routers

application

NRSE

Computer Science
### General problem space

<table>
<thead>
<tr>
<th></th>
<th>Homogenous: bottleneck at edge</th>
<th>Heterogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>single domain</td>
<td>multiple domain</td>
</tr>
<tr>
<td>Dynamic reservations</td>
<td></td>
<td>multiple domain</td>
</tr>
<tr>
<td>Advance reservations</td>
<td>current GRS work</td>
<td>current GRS work</td>
</tr>
<tr>
<td></td>
<td>difficult</td>
<td>very difficult</td>
</tr>
</tbody>
</table>

**New in GRS:**
- Reservation types: real-time & non-real-time
- Application paradigms: notifications and deadlines
Approach

• Assume:
  • end-users are willing to co-operate
  • highly de-centralised
  • users form a community
  • similar properties to peer-to-peer (p2p) systems

• NRS users form a community:
  • share resources between sites
  • network scheduling is between sites in the community
  • micro-management of flows at sites
NRSE Design (in progress)

user (human or application)

NRS-M

NRS-A

scheduling control (schedule management) (reservation invocation)

NRS-L

NRS-P

RD

local management

network elements (e.g. routers, switches, etc.)

other NRSEs
Scheduling control principle [1]

Real-time request

- Data rate
- Start time
- End time

Non-real-time request

- Data rate
- Start time
- End time

Fixed

Same area
Scheduling control principle [2]

available capacity (e.g. EF)

MAX

population this area with requests for jobs (schedule management)

scheduling policy for allocation between real-time and non-real-time jobs can be controlled locally
Current status: testing on SJ4dev

SuperJANET4 development network
Application synchronisation

Deadlines
- File transfers
- Use with non-real-time reservations

Notifications
- Event-driven synchronisation:
  - reservation-begin and reservation-finish
- Notifications for:
  - QoS violations
  - administrator intervention
  - SLA changes ...

Needs re-design of APIs and applications
Future

• NRSE:
  • extend to “full” network reservation platform
  • scheduling policies
  • management interface
• 46PaQ:
  • IPv4 + IPv6 Performance and QoS
  • QoS and monitoring deployment and use
• General signalling platforms and systems:
  • state management
  • optical and hybrid-optical
Questions?

A good way to get answers 🍺🍺