The 10Gb/s project

Alastair Basden / Jeremy Yates

Durham University
Institute for Computational Cosmology

DiRAC
High Performance Computing Facility

ExCALIBUR 10

Trans-atlantic big data

- Aiming to transfer large (~50TB) datasets from US facilities
  - Achieving 10Gb/s rates
  - 4TB/hour
  - End-to-end transfers
- This seems straightforward
  - But we have very little control

Alastair Basden
DiRAC / Durham University

Projects

- ORNL to Cambridge (UKAEA and DiRAC)
  - Large scale wavefunction data (50TB-1PB)
    - Generated in US for fusion and lattice QCD
- Cambridge to Durham (DiRAC)
  - Offsite backups to tape
- Edinburgh to Durham and Cambridge (DiRAC)
  - Movement of wave function data from TURSA
    - Lattice QCD
- Durham to collaborators (VIRGO consortium)
  - Download of large cosmology datasets

Alastair Basden
DiRAC / Durham University

Route taken

- From disk (parallel file system)
  - Typically multiple files, some multiple-striped
  - via RAM, fabric
- To Data Transfer Node RAM
- To NIC
- Via Switches/cables/etc
- To NIC
- To RAM
- To disk (parallel file system)
  - via fabric, RAM

Alastair Basden
DiRAC / Durham University

Bottlenecks

- Multiple bottlenecks if not optimised
- Aiming to have the external network as the sole bottleneck
  - Node to disk transfers should not be a problem
  - Parallel file system: capable of line rate from a node
    - 200Gb/s HDR InfiniBand

Alastair Basden
DiRAC / Durham University

Logistics

- Users (with an account at both ends)
- System administrators
- Data transfer experts
- Multiple transfer software options
  - Globus
  - BBCP
  - FTS3

Generic case

- No control over external projects who use the facilities
  - Globus slected by the projects
  - FTS3 not yet ready for use with tokens
    - We will aim to use it when it becomes available

Optimisations

- Optimise the various components of the IO stack
  - Lustre striping
  - Ceph placement
  - TCP parameters
    - iperf3, tcpmon
  - Switch configurations

Alastair Basden
DiRAC / Durham University

Current status

- Link to Durham is a 4x 10G bonded link to JANET
  - Max 10Gb/s for single transfer
- MTU settings require changing
  - Central IT change request process
  - Collaboration with JISC
  - Currently seems to be differing MTU sizes
- DoS detection may be hampering things
- 8Gb/s transfers between Cambridge and Durham are achievable
- 25Gb/s transfers between Cambridge and GEANT London
- 400Mb/s from the US Frontier system (peak around 800Mb/s)
- ORNL to London: 6.4Gb/s (1500 MTU)

Work completed

- Network cards upgraded to 100G
- Globus installed at Cambridge
  - Already present at Durham
  - Free version, so some performance tuning not available
    - To be determined: whether a paid licence is necessary (£20-30k)
- Network maps/architecture diagrams understood

Some settings updated

- Ring buffer size:
  - /usr/sbin/ethtool -G p2p1 rx 8192 tx 8192
- sysctl.conf:
  - net.core.rmem_max = 268435456
  - net.core.wmem_max = 268435456
  - # min/default/max TCP read and write buffers
  - # increase Linux autotuning TCP buffer limit
  - net.ipv4.tcp_rmem = 4096 87380 268435456
  - net.ipv4.tcp_wmem = 4096 65536 268435456
  - # increase the length of the processor input queue - number of unprocessed input packets (Mellanox value too)
  - net.core.netdev_max_backlog=250000
  - # recommended default congestion control
  - net.ipv4.tcp_congestion_control=cubic
  - # don't cache ssthresh from previous connection
  - net.ipv4.tcp_no_metrics_save=1
  - # recommended for hosts with jumbo frames enabled so can deal with MTU 1500 ESnet also recommend enable
  - net.ipv4.tcp_mtu_probing=1
  - # enable timestamps
  - net.ipv4.tcp_timestamps=1
  - # enable SACK allows specification which bytes have been lost and which bytes received - only retransmit the lost bytes
  - # ESnet also recommend enable
  - net.ipv4.tcp_sack=1
  - # enable window scaling - allow larger TCP Receive Window
  - net.ipv4.tcp_adv_win_scale=1

Related:

- LSST transfers from NERSC to CSD3
- PerfSonar node installed at Durham

Ongoing

- Further work required to achieve 10Gb/s
  - May require a link upgrade at Durham
    - (requires a demonstrated need to JISC)
- Expected to be achievable
- May require additional software
  - On-the-fly tuning of congestion/buffers
- And other tunings
  - To/from file not yet optimised