Point-to-point Architecture topics for discussion

Remote I/O as a data access scenario

- Remote I/O is a scenario that, for the first time, puts the WAN between the data and the executing analysis programs.
  - Previously, the data was staged to the site where the compute resources were located and data access by programs was from local, or at least site-resident, disks
  - inserting the WAN is a change that potentially requires a virtual circuit service to ensure the smooth flow of data between disk and computing system, and therefore the “smooth” job execution needed to make effective use of the compute resources. (Though it may be that raw bandwidth is a bigger issue – see below.)

Remote I/O as a data access scenario

- The simplistic is that each RIO operation involves setting up a circuit between compute system and data system
  - If circuits are set up and torn down when remote files are opened and closed, and if the example is typical, then circuit duration is short – of order 10 minutes
  - This is almost certainly impractical just based on looking at the number of jobs being executed today

Remote I/O as a data access scenario

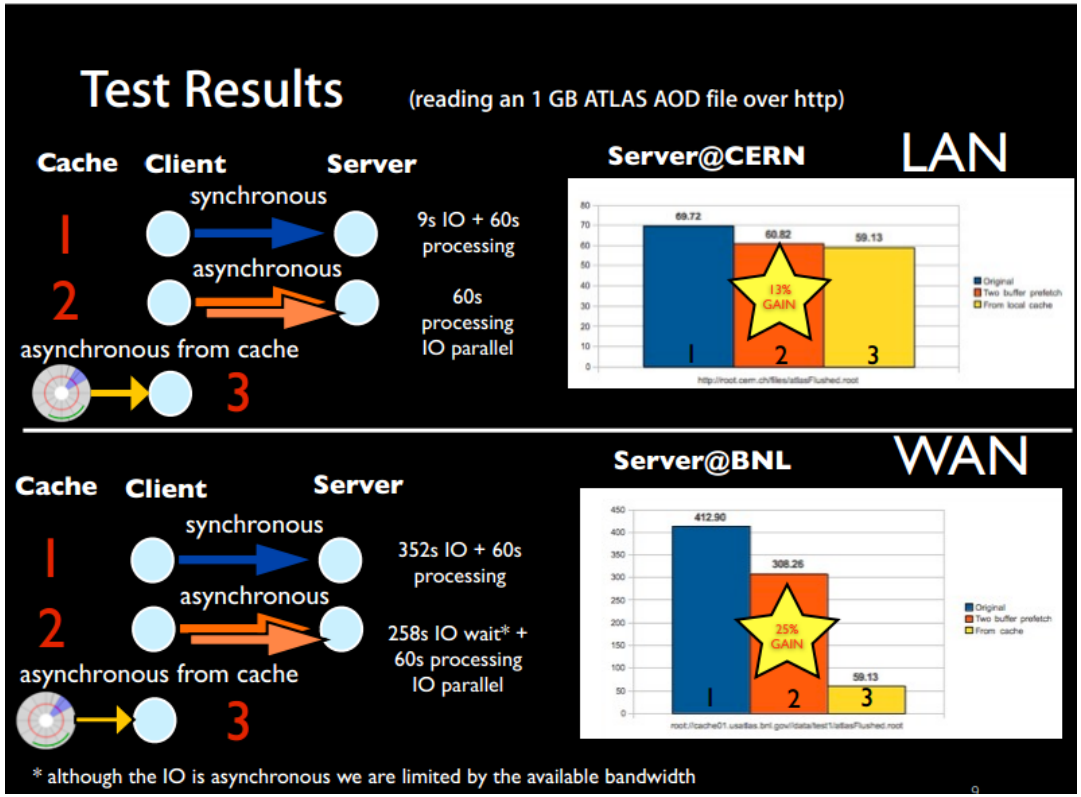
- In regards to the potential intensity and density of circuit usage, what is the potential clustering and its parameters.
  - Is analysis organized around chunks of interesting data, and that a fairly large number of analysis jobs will work on that data for some limited period of time?
  - If this is the case
    - then we would expect a 1 X N sort of clustering, where "1" is the data set and "N" is are the locations of compute resources that will operate on that data. It is likely the case that N is small-ish – maybe of order 10?
    - what is the duration of this "cluster?" That is, how long will the jobs spend processing such a “chunk” of data?
    - How many simultaneous clusters can we expect? That is, how many chunks of data are being analyzed simultaneously? How many different sites will be involved? (Presumably all of the Tier 2 sites.)

Remote I/O as a data access scenario

- **Mix of remote I/O and bulk data transfer**
  - What will be the mix of RIO-based access and bulk data transfer?
  - The change to remote I/O is, to a certain extent, aimed at lessening the use of the bulk data transfers that use GridFTP, so is addressing GridFTP addressing a dying horse?
  - How much bulk transfer will there be in a mature RIO scenario, and between what mix of sites?

Remote I/O as a data access scenario

- Performance
  - Some of the ATLAS initial testing using HTTP as the remote I/O protocol shows a 40x decrease in program I/O throughput compared to local disk



Remote I/O as a data access scenario

- Is this example typical of the amount of data consumed by an analysis program?
- Is this typical of the amount of data read compared to the amount of CPU time used?
- Will circuits be used primarily to secure bandwidth?
  - (If so, this exercise may be more about the available underlying network bandwidth rather than about circuits, *per se*.)

- If circuit setup fails, does this program proceed and automatically fallback to the VRF access that the computing systems is embedded in (or the general IP infrastructure)?

Remote I/O as a data access scenario

- Interactions of VRF and circuits (? – maybe an operations issue)
  - It may be that aggregations of compute and disk resources feeding a small number of long-lived circuits (site-to-site, cluster-to-cluster, etc.) could meet all user requirements.
    - This would reduce the “intensity” of use of the service and obviate the need for users to deal directly with circuits.
  - Could, on the other hand, the existing LHCONE VRF environment make it difficult to aggregate resources in this way?
    - By way of example, in ESnet many OSCARS circuits are used for interconnecting routers that, at the site end, aggregate resources, e.g. by routing for a cluster on a LAN.
    - This is the sort of thing that might be hard to do if that cluster is also accessed via the VRF because of how the address space is managed

Point-to-point Architecture topics for discussion

The LHCOPN as a circuit scenario

- Requirements
  - guaranteed delivery of data over a long period of time since the source at CERN is essentially a continuous, real-time data source;
  - long-term use of substantial fractions of the link capacity;
  - a well-understood and deterministic backup / fail-over mechanism;
  - guaranteed capacity that does not impact other uses of the network;
  - a clear cost model with long-term capacity associated with specific sites (the Tire 1s);
  - a mechanism that is easily integrated into a production operations environment (specifically, the LCG trouble ticket system) that monitors the circuit health and has established trouble-shooting, resolution responsibility, and provides for problem tracking and reporting

The LHCOPN as a circuit scenario

- There has been discussion of moving the LHCOPN to a virtual circuit service for several reasons:
  - VCs can be moved around on an underlying physical infrastructure to better use available capacity, and potentially, to provide greater robustness in the face of physical circuit outages;
  - VCs have the potential to allow for sharing of a physical link when the VC is idle or used less than the committed bandwidth.
- Therefore these are requirements (?) for circuits
  - Topological flexibility
  - Circuit implementation that allows sharing the underlying physical link
    - That is, b/w committed to, but not used by the circuit, are available for other traffic

The LHCOPN as a circuit scenario

- Other useful semantics
  - Although the virtual circuits are rate-limited at the ingress
    - to limit utilization to that requested by users
    - they are permitted to burst above the allocated bandwidth if idle capacity is available
    - Must be done without interfering with other circuits, or other uses of the link, such as general IP traffic, by, for example, marking the over-allocation bandwidth as low- priority traffic
  - User can request a second circuit that is diversely routed from the first circuit.
    - In order to provide high reliability for backup circuit ……
- Why is this interesting?
  - The rise of a general infrastructure that is 100G / link, using dedicated 10G links for T0 – T1 becomes increasingly inefficient
  - Shifting the OPN circuits to virtual circuits on the general (or LHCONE) infrastructure could facilitate sharing while meeting the minimum required guaranteed OPN bandwidth

Point-to-point Architecture topics for discussion

Cost models, allocation management

- The reserved bandwidth of a circuit is a scarce commodity
  - this commodity must be manageable
    - From the view of the network providers
    - What sorts of manageability does a user community require
      - What does the user community need to control in terms of circuit creation?