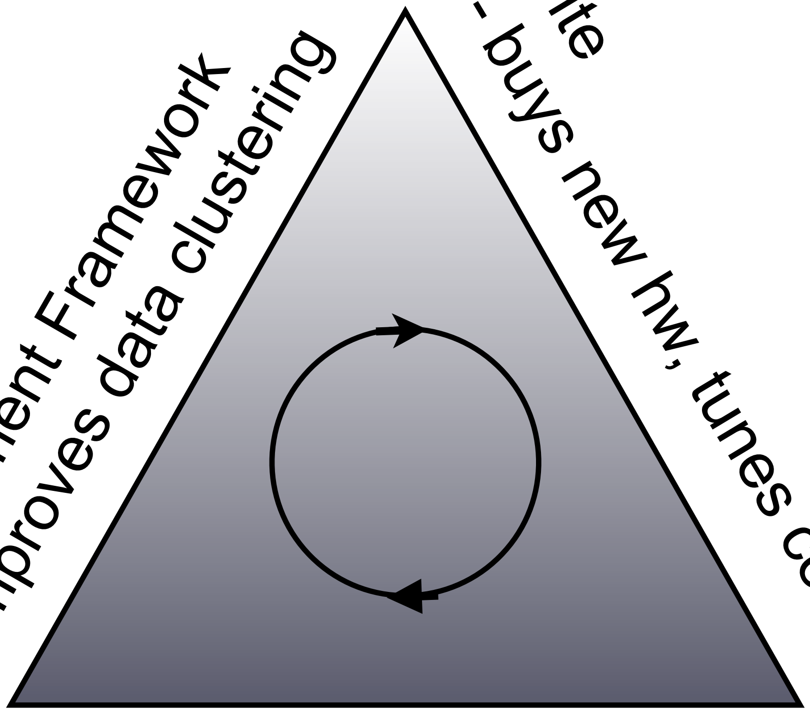
Working Group for I/O Classification and Benchmarking

Dirk Duellmann, CERN

GDB Meeting

12 September 2012

A vicious triangle..



Experiment Framework
- improves data clustering

Site
- buys new hw, tunes config

ROOT / Storage S/W provider
- improves performance

These examples of parallel improvements may converge, but may also just interfere...
- concrete metrics may help to confirm successful improvements and guide the iterations..

Motivation & Mandate

- During the discussions of the Data and Storage Evolution Group several shortcomings in the area of collecting and reproducing realistic workloads for benchmark and optimization purposes have been identified:

1) The real aggregate I/O access pattern against WLCG SEs is not easy to quantify or to reproduce.
2) Sites, experiments and software providers use a variety of tools to address performance optimization and resource planning this including root scripts, HammerCloud, OS level I/O benchmarks.
3) The existing tools do not necessarily use a common approach to define the key metrics nor are benchmark codes and results centrally available from a managed repository.
4) Not all benchmarks can be scaled to run in multi-client mode to obtain the performance of a fully loaded server.
5) In many cases the actual type of access (eg sparseness vs sequential, WN local, site local, WAN federated) is either not documented or not tunable to the changing access approaches of the experiments.

Motivation & Mandate

- We propose to setup a small working group to perform a “market survey”, documenting agreed key metrics, existing tools, pointing out areas where more coherence could be obtained. The document should describe a systematic approach for the different main use-cases for performance analysis using existing tools:

  1) optimisation of existing or planned site installations with respect to an expected I/O workload (eg CPU vs Network vs RAM vs SSD vs Disk cost)
  2) optimisation of experiment I/O layer wrt to local and federated data access
  3) optimisation of SE implementations wrt to an expected I/O load
  4) determination of aggregate I/O patterns of a real job population in order to obtain realistic parameters for 1-3) and in order to identify changes of the real I/O over time

The latter task should involve a survey of the existing monitoring information (from sites & experiments) wrt to key metrics, which would help to validate existing load generators against measured I/O load. It should also investigate the option of logging and replaying I/O patterns in order to create easily deployable workload generators without dependency on experiment software frameworks.

Initial Work Areas

- Site Optimisation
  - collect relevant existing tools and make them available together with measured results
  - documentation on how interpret obtained results
  - provide an automatic framework to execute tests and collect results (often multi-client)
- Storage System s/w Optimisation
  - provide standard benchmarks and data to software provides interested
  - provide an automatic framework to execute test and collect results
- I/O classification & benchmark tuning
  - define key metrics and obtain distribution from production monitoring
  - regularly retune existing benchmarks to match the behaviour of the real job population

Non-Focus Areas

- Compare different storage implementations against each other
  - this is already done in a systematic way by the Hepix Storage WG

- Experiment Framework & ROOT I/O layer Optimisation
  - this work is already taking place as part of the ROOT I/O workshops
    - Commonality between experiments is essential!
  - for me these is a key exchange forum, which may not have gotten the right attention yet
    - additional discussion space in this area can be offered if there is interest

One Example Question

- What medium term strategy do we follow for the resource balance between WN storage and storage cluster?
- Copy-local
  - all random I/O takes place on WN
  - storage systems are optimised for put/get
  - heavy, but short network connections to store
  - easier integration of simple storage (eg S3)
  - => need WN IOPS and volume to scale with core growth
- LAN-access
  - WN storage volume and IOPS less relevant
  - random access scalability of back-end important
  - many long-term connections
- Right now both are happening depending on experiment and site

Another one...

- Which job types can be run efficiently in a federated and/or cached environment?
- Naively the following metrics should be sufficient to estimate this:
  - fraction of data read / file size
  - total number of reads
  - total number of regions per vector-read
  - integrated seek distance / file size
- Many of these numbers are now becoming available in several places
  - AAA monitoring (not only for xroot)
  - detailed logs of EOS usage at CERN
- Interpreting them will be work, but provide real benefits

Required Participation

- Site performance expert
  - to define relevant metrics
  - to help simplifying benchmarks deployment
- Storage system tuning experts
  - to define the key metrics need for strategy decisions in existing storage packages
- ROOT I/O system expert
  - owner of at least one very popular benchmark (Rene’s script with ATLAS nTuple)
  - document parameters and explain the various statistics (TTreeStat)
- Experiment performance expert
  - present experiment performance evaluation and monitoring frameworks
  - how can sites / storage sw providers interpret existing results?

Meetings & Documentation

- Kick-off meeting next week
  - Doodle poll will go out to old DM & SM TEG lists
  - http://www.doodle.com/q8r2whm8qx8wbma4
- From then on
  - new list wlcg-wg-storage-benchmarking@cern.ch
    - please sign up (but expect to get work)
- Twiki to collect minutes and recipes
- First report
  - October pre-GDB in Annecy

Relationship to other working groups and activities

- Federation WG
  - participating directly (bi-directional)
- Storage Protocols WG
  - participating directly (bi-directional)
- CERN Cloud storage evaluation
  - evaluate existing multi-client benchmark framework used for Cloud storage and EOS evaluation
- ROOT I/O optimisation workshops
  - participate actively

DSS

Status after initial discussions

CERN IT Department

- ROOT (Fons/Philippe)
  - signed up for description of performance stats
  - ownership and periodical review of Rene’s script and suitable parameters
- DPM (Oliver/Ricardo)
  - interest in shared benchmark environment
  - DPM perfsuit
- Site testing via modified HC (Wahid/Ilja)
- Still missing
  - AAA monitoring (Brian/Matevz)

Summary

- A lot of relevant work on benchmarking is already going on
  - new and more aggregated metrics collections are becoming available
- WG will try to pull existing work together
  - document existing tools and results for the different use cases
- Provide a forum
  - for increasing the commonality between used metrics across experiments and sites
  - extracting information from the data
- In contact with several key players
  - but still completing the list (please sign-up)