SSD SCALABILITY PERFORMANCE FOR HEP DATA ANALYSIS USING PROOF-LITE

L. BARBONE [1,2], G. DONVITO[1], A. POMPILI [1,2]

(1): INFN SEZIONE DI BARI, ITALY (2): UNIVERSITA' DEGLI STUDI DI BARI "ALDO MORO", ITALY

HEP data analyses are carried out at Tier-2/Tier-3 computing facilities. The end-user analysis activity can be schematically divided into the following steps:

1. BATCH activity through Grid interface :
initial data format reduction & preliminary event selection by means of specific experimental software framework.
Rare/few revision cycles.
Long execution time.

2. Local BATCH/INTERACTIVE activity :
user-skimmed data collections (organized in ROOT files) further reduced via repeated refinement cycles required for selection tuning/systematic studies.
Reasonable execution time for a single cycle.
Full interactivity desired, but level of interactivity may vary with different solutions adopted at Tier-2/Tier-3.

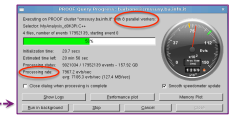
3. Fully interactive tasks :
Histogramming, fitting, plotting,...
Repeated optimization cycles.
Very short response times.

Crucial phase for the analysis and its time schedule !

a) Growth of integrated luminosity collected by LHC experiments (increasing data sets size)
b) Initial event skimming rate can't be increased beyond a certain extent

→ A fully interactive ROOT approach is compromised : too long single cycle duration

PROOF-Lite (*) is a dedicated version of the Parallel ROOT Facility (PROOF) optimized for multi-core multi-user servers or workstations: it can run an interactive analysis task with parallel execution on several CPU resources (workers) allowing:
a) enough short single cycle duration on large ROOT input files
b) real-time feedback (via a GUI dashboard)



(*) G.Ganis et al., Proceedings of Science, XII ACAT, 2008, Erice, Italy

Performance tests

PROOF-Lite performances (linear scalability up to 8 workers, adaptability to changing load) are investigated by means of I/O-*limited* tasks within tests aiming to compare:
1) SSD disks w.r.t. HDD (SAS or SATA) disks
2) a multi-core multi-user server w.r.t. a multi-core workstation

4 ROOT input files (total size: 344GB) locally stored on the servers used in the tests:
- obtained, using *CMSSW_3_8_7_patch2* reconstruction software release, on 2010 CMS data
- for an analysis aiming to extract the $D^0 \rightarrow K^-\pi^+\pi^+\pi^-$ decay mode in *minimum bias events*

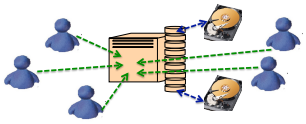
Performance figures of merit used:
1) *speed-up* : $S(n) = \frac{T_1}{T_n}$ with T_n = processing time for n workers
2) *average processing rate* : $< R(n) >= \frac{\#\ \text{processed events}}{T_n}$

ROOT 5.27/06b used in the tests

Hardware configuration of the two test servers (at Bari Tier-2)

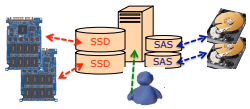
1) *MCIS* : multi-core interactive server for Bari Tier-2 users who are CMS analysts

Two 6-cores CPUs (X5650) with HyperThreading technology providing up to 24 concurrent processes, 24GB of total RAM & one single RAID6 with 9 SATA disks (2TB each), with a mechanics of 7200RPM.



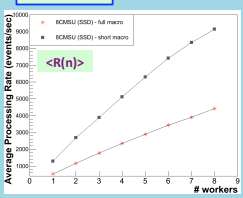
2) *8CMSU* : 8-core server dedicated to a single user (workstation)

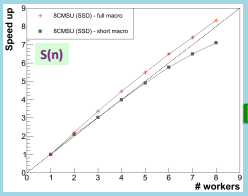
Two 4-cores CPUs (AMD Opteron 2347HE), 8GB of total RAM, RAID1 of 2 SAS drives (146GB each) with a fast mechanics of 10kRPM, RAID0 of two MLC SSD disks (256GB each).



TESTS RESULTS

1.

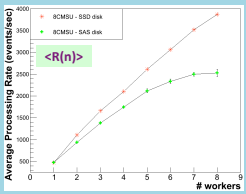


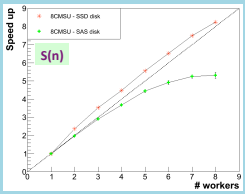


Two versions of the same sequential multi-step selection to extract the physical signal (*short* / *full* macro) are used, by excluding / including heavy histogramming.
They imply less / more I/O-demanding tasks.

As expected, <R(n)> for *short macro* is at least twice that for *full macro*. S(n) is rather linear, better for *full macro*.
In the following tests the *full macro* will be used:
1) task is I/O-limited,
2) task shows linear scalability (by using SSD disks)

2.



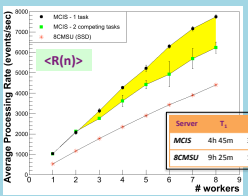


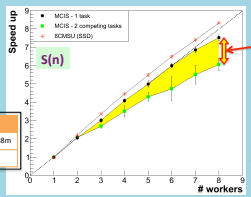
Test done with only 1 ROOT file (100GB): SAS size-limited.

With SAS disks there is a saturation effect for $n > 5$.
This departure from linear scalability reveals that the task executed on SAS disk is completely I/O-bound and CPUs can't be efficiently used (confirmed by direct monitoring).

PROOF-Lite performance is by far better for SSD w.r.t. SAS disks (when both mounted on a good workstation).

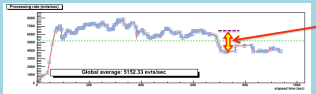
3.





Tests done always with *MCIS* normally/caothically loaded (i.e. never overloaded).
The yellow band represents the performance uncertainty associated to load variability, obtained by continuously running a 4-workers PROOF-Lite task concurrent to the task under test (on a merge ROOT file of the 4 input files).

MCIS hardware configuration prevents from a saturation effect (seen for HDD-SAS).
A performance limitation is behind the corner, starting for $n \geq 8$, as better appears when the concurrent task is executed (linear scalability loss begins for $n \geq 4$).



Dropping rate effect when a concurrent 8-workers PROOF-Lite task begins to run.

This comparison is between 2 realistic alternative solutions at Tier-2/Tier-3.
Technological & economical difference: *MCIS* costs 3 times more than *8CMSU*.
Tier-2/Tier-3 managers should evaluate which choice could better fit the evolving interactivity needs of end-users, taking also into account how much widespread the use of PROOF-Lite currently is and will be in the near future.

Computing in High Energy and Nuclear Physics 2012
May 21-25, 2012 – New York – United States