Physics Analysis and Skims

Shohei NISHIDA

KEK (Belle)

Jan 21, 2004

Super B Factory Workshop in Hawaii

Present Analysis/Skims at Belle

Hadron data

- We have been accumulated $\sim 150\ \mathrm{fb}^{-1}$ data.
- Hadron data ("mdst") size: $1\ \mathrm{fb}^{-1} \sim 50\ \mathrm{GB}$.
- All the hadron data are on disks. Users can read these data through network.
- MC data have been kept on disks and tapes; but we will try to put (new) MCs to disks.
- $\sim 10\ \mathrm{TB}$ to keep hadron data and MC.

Present Analysis/Skims at Belle

Physics Skim

- ~ 20 skims are officially made.
- "index" scheme: only event number is saved. file size is small/negligible.
- Each skim includes $1 \sim 10\%$ of hadron events.
- full_recon skim: 5% of hadron events.
- No skim for systematic study.

`b2dlnu_skim`	`dpart_skim`	`etapk_skim`	`psi_skim`
`b2ulnu_skim`	`dsds_skim`	`frec_skim`	`radb_skim`
`blamc_skim`	`dstarpi_skim`	`hh_skim`	`tau_skim`
`ddk_skim`	`dsubspi_skim`	`icpv_skim`	
`dilep_skim`	`endlep_skim`	`lepton_skim`	
`dmix_kpi_skim`	`etac_skim`	`ppp_skim`	

Present Analysis/Skims at Belle

CPU / Disks

- Main group servers have only ~ 1.5 TB disk space for all users (HSM is also available to users).
- "analysis farm" is newly set up.
  - 6 TB user disks
  - 80 servers (2.8 GHz Xeon, 2 CPU)
- Each insititute has their own CPUs/disks, but not so many institutes can read the hadron data at their institute.

Present Analysis/Skims at Belle

Users' Analysis

- Most users analyze skimmed data, but many users still analyze all hadron data (new mode; systematic study; just don't want to follow the official scheme).
- Roughly 1 day to read $10\ \mathrm{fb}^{-1}$ (per job) if we do not use skim. $\Leftrightarrow$ network access $\sim 10\ \mathrm{MB}/s$
- 3 days $\sim$ 1 week to analyze all the data (except before conferences).
  - it depends on the analysis mode or whether we use skim.
- more time to analyze MC (due to large amount).
  - almost 1 month if we try to read all the MC on tape.
- Many users are still using PAW and hbook. (users' hbook files are sometimes too large...)

Present Analysis/Skims at Belle

CP analysis with $140\ \mathrm{fb}^{-1}$

- Number of events used in the fit.
  - $J/\psi K^0_S$ final CP fit sample: ~ 2000 events
  - ϕK^0_S final CP fit sample: ~ 100 events
  - Control sample for resolution / wrong tag fraction ($D^*\ell\nu, D^{(*)}\pi(\rho), J/\psi K^{(*)}$): $\sim 2\times 10^5$ events
- Time necessary for fit.
  - Final fit: $\sim \mathcal{O}(1)$ sec.
  - resolution/wtag fit (34 parameters) ~ 2 hours. [Xeon 2.8 GHz 29 CPUs]
  - Need to repeat a few hundred times $\Longrightarrow$ 10 days for systematic error.

Present Analysis/Skims at Belle

CP analysis with $140\ \mathrm{fb}^{-1}$ (cont'd)

- Basically, systematic study is time consuming.
- Calculation of significance requires large amount of toy MC
  $\longrightarrow$ another few days.

Analysis/Skims at super B factory

If we just scale to $5\ \mathrm{ab}^{-1}$

- hadron data $\sim 250\ \mathrm{TB}$ (or hadron data + MC $\sim 1\ \mathrm{PB}$).
- ~ 1 year to analyze all the data (?)

$\Longrightarrow$ some improvement necessary both on disks and analysis speed.

Analysis/Skims at super B factory

Disk size

- Technology progress.
  - We will be able to buy a few times larger disks with the same price.
  - But, probably not 50 times larger...
- Need to try to make the data size smaller (?).
  - Smaller data size per event.
  - Tighter hadron criteria.
- Effective use of disk.
  - Distribute hadron data to several institute and share them.
- Just buy more.

Analysis/Skims at super B factory

Analysis speed

- Technology progress.
  - faster CPU.
  - disk I/O & network speed.
- Skim selection criteria.
  - More data $\Longrightarrow$ tighter selection criteria.
  - Start from full reconstruction sample.
- More compact information ("mini-mdst"?)
  - Data that includes only 4-vector, PID, (...) and reference to the usual data ("mdst").

Analysis/Skims at super B factory

CP fit at super B ($5\ \mathrm{ab}^{-1}$)

Time necessary for fit

- Final fit will take $\mathcal{O}(10)$ sec with $1\ GHz$ CPU.
- resolution/wtag fit (34? parameters) $\sim$ 3 days. [Xeon 2.8 GHz 29 CPUs]
- more than 1 year with present CPU power (e.g. Xeon 2.8GHz 150 CPU) and same method.

Solution

- It seems the CPU power is the main problem in this case.
- Most tasks run in parallel. Increasing the number of CPU.
- Change of the analysis method can save CPU power.
  - Unbinned ML fit $\longrightarrow$ Binned ML fit.

Analysis/Skims at super B factory

CP fit at super B ($5\ \mathrm{ab}^{-1}$) (cont'd) Note

- The system tends to be limited by I/O when we extends the system.
- Giga bits ethernet + fast hard drive
- How about Grid?

Analysis/Skims at super B factory

User analysis

- Time for reading all data is now 3 days −− 1 week.
  - The faster, the better.
  - But, 1 week is reasonable (acceptable).
    If we can analyze much faster, we will be careless.
  - Is it possible?
- It will (at least) take more time to analyze all the hadron data.
  - Usage of official skim.
    Perform skim continuously; users submits skim codes anytime when necessary.
  - Users can start from full reconstruction skim for example.

Analysis/Skims at super B factory

User analysis (cont'd)

- Probably, once we can read the data, the situation is not changed so much (I hope).
  - More data / difficult mode $\Longrightarrow$ tight selection criteria.
- But, we may need to through away PAW/hbook.

Summary

Summary

- hadron data size: $10\ \mathrm{TB} \rightarrow 250\ \mathrm{TB}$
- users' analysis
  - tough work to read all the hadron data...
  - continuous skimming scheme.
  - "mini-mdst"
  - ...
- network is always an key issue.
- ...