Simulation Production issues from a Physics Perspective

(based on discussions with Dave Brown, Chih-hisang Cheng, and Fergus Wilson)

Adrian Bevan / Dave Brown

Overview

- The Problem?
  - How do we cope with additional geometries
  - How users access the code / files etc.
- What do we aim for?
  - Scale
  - Scope
  - Documentation/Tools
  - Time-scales
- Resources available?
- Discussion

The Problem

- Physics studies require detailed understanding of signal and background.
  - Analyses take time to develop.
  - Will be ready at different times.
  - Will need background simulation much larger than BaBar.
  - Timescale is fixed by delivery of TDR: i.e. 2010.
- BaBar solution: single point failure that we must not repeat.
  - Filter Tools controls how we select generated/reprocessed data. Any mistake in a single mode brings everything down.
- In an ideal world analyst Joe Blogs shouldn't have to generate all of his generic physics background samples.
  - Move to a more central mode of operation using available GRID enabled resources.

What do we aim for? : Scale

- Physics target is $75ab^{-1}$ in 5 years.
  - $O(15ab^{-1})$ per year.
  - $1ab^{-1}$ corresponds to $\sim 1.1\times10^{9}$ B pairs and similar numbers of $\tau^{+}\tau^{-}$, uds, and ccbar.
    - Bhabha scattering has a σ 40× larger.
  - So 1 year of data taking is approximately equivalent to 15.5billion events per sample.
  - We should be aiming to generate 1 year of generic physics background for studies.
    - This can be used to extrapolate to the $75ab^{-1}$ target in a straightforward way for most cases.
    - Will give confidence to the community that FastSim analyses are considering realistic scenarios.

What do we aim for? : Scale

- 15.5billion events per sample takes a while to generate:

- Example: $\tau \rightarrow 3\mu$ analysis generating events:

$$uds = 24Hz$$
$$ccbar = 21Hz$$
$$B\overline{B} = 16Hz$$
$$bhabha = 13Hz$$

Writing out all events by default – if we discard the generated events, which includes a significant overhead from I/O.

Cob: dual-core 2.0GHz Opteron 270 CPUs, 4GB memory
Fell: dual quad-core 2.66GHz Xeon CPUs, 16GB memory

- Example $B^0 \rightarrow \pi^0\pi^0$ signal MC generation rate ~17 Hz

- Naïve scaling to 15.5 billion events per sample (1yr) requires:
  - ~ 1 week to generate a single generic background sample with 1000 CPUs.

Using V0.0.9 at SLAC
#'s c/o Cedric Weiland

What do we aim for? : Scope

- The golden matrix has many core measurements.
  - Working group conveners have lists of these.
    - e.g. WG1 has identified
      - $(J/\psi, \phi, \eta')K^0$, $\pi^+\pi^-$, $\pi K^0(\gamma)$ + di-leptons

$J/\psi K^0_S$: Purity ~98%, don't care about large generic background production as the signal MC will give us most of the answer.

ϕK^0: Relatively clean (thanks to the ϕ), again, background should not be too much of a problem.

$\pi\pi$: Starts to get to be an issue: need uds (and possibly) ccbar in addition to signal and a handful of exclusive background modes.

τ LFV searches:

Background is vital: Single event sensitivity is proportional to the background rejection ability, and SM signal has BR~10^{-54}.

Also related to the beam polarization issue.

What do we aim for? : Scope

- The golden matrix has many core measurements.
  - Working group conveners have lists of these.
    - e.g. WG1 has identified
      - $(J/\psi, \phi, \eta')K^0$, $\pi^+\pi^-$, $\pi K^0(\gamma)$ + di-leptons
    - Want to co-ordinate MC generation for all Physics WGs.
    - How many modes are in this list?
    - How much CPU required?
    - How much disk required?
    - Packaging the software for GRID submission at all sites we're interested in – how do we do this?
    - How do we achieve this? In a way **o**pen for expansion, and without a single point failure!
      - Need a simple recipe (with QA benchmarks) for users to convert their code into something robust that can be used in production-mode.

What do we aim for? : Documentation/Tools

- Should also set up a list of instructions for individual users to follow in order to generate samples of their own MC.
  - Generic wrappers to PacUserApp
  - Part of the SuperB release

- Need to evolve these simple tools into something sophisticated enough to start doing an organized production.
  - Ultimately want a database to track files produced and location stored (for now is a set of text files linked from the wiki sufficient).

- Different people have different perspectives on analyses (i.e. optimize the silicon tracker, DCH, PID, EMC, IFR).
  - Co-ordinate access to generated samples through some framework.
    - Files are not stored centrally – user access is limited
    - But we should benefit from working group packages giving us access to 'standard selections' to help the DGWG effort.
    - i.e. need a good dialogue between Physics and DGWG.

What do we aim for? : Timescale

- Obviously need to run a production for the TDR:
  - Mid 2010?
- This can't be the first iteration!
  - Must aim for a trial run (soon) with a few modes included.
  - Anticipate a second iteration
    - ~6 months later?
  - This would give us 2 trial runs between now and the TDR production.
- Is this feasible?

What resources are available?

- CNAF (Italy)
- Caltech (US)
- QMUL (UK)
  - Managed by HEP group, so we have ROOT access as well as GRID/Batch access.
    - We can have 10% of the farm + any un-used LHC cycles.
    - Equates to a dedicated resource of about 10Tb of disk + 150CPUs.
- RAL(UK)
  - BaBar CPU allocation is penned in as transferring to SuperB.
    - Hope to be able to use this as soon as job submission works on the GRID.
- + … ?

Discussion

- Computing and Physics have different constraints with regard to this issue…
  - What works?
  - What doesn't work?
  - What other questions do we need to ask?
  - Who's going to be in charge of co-ordinating this?

  - Let's try and reach a compromise solution *this week* that we can test as a starting point.