Data and MC Production

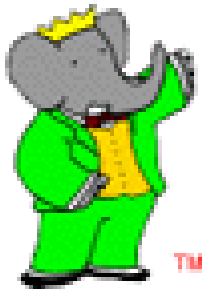
Olya Igonkina

(University of Oregon)

on behalf of PR, Skim, SP groups

- Prompt Reconstruction
- Simulation Production
- Conversion/Skimming effort

- Short overview of the system
- Data processed
- Performance and Development
- Outlook



™ Collaboration meeting, July 8, 2004, SLAC

CM2 Production Era

⇒ A new event-store implementation - the CM2 Kanga event-store

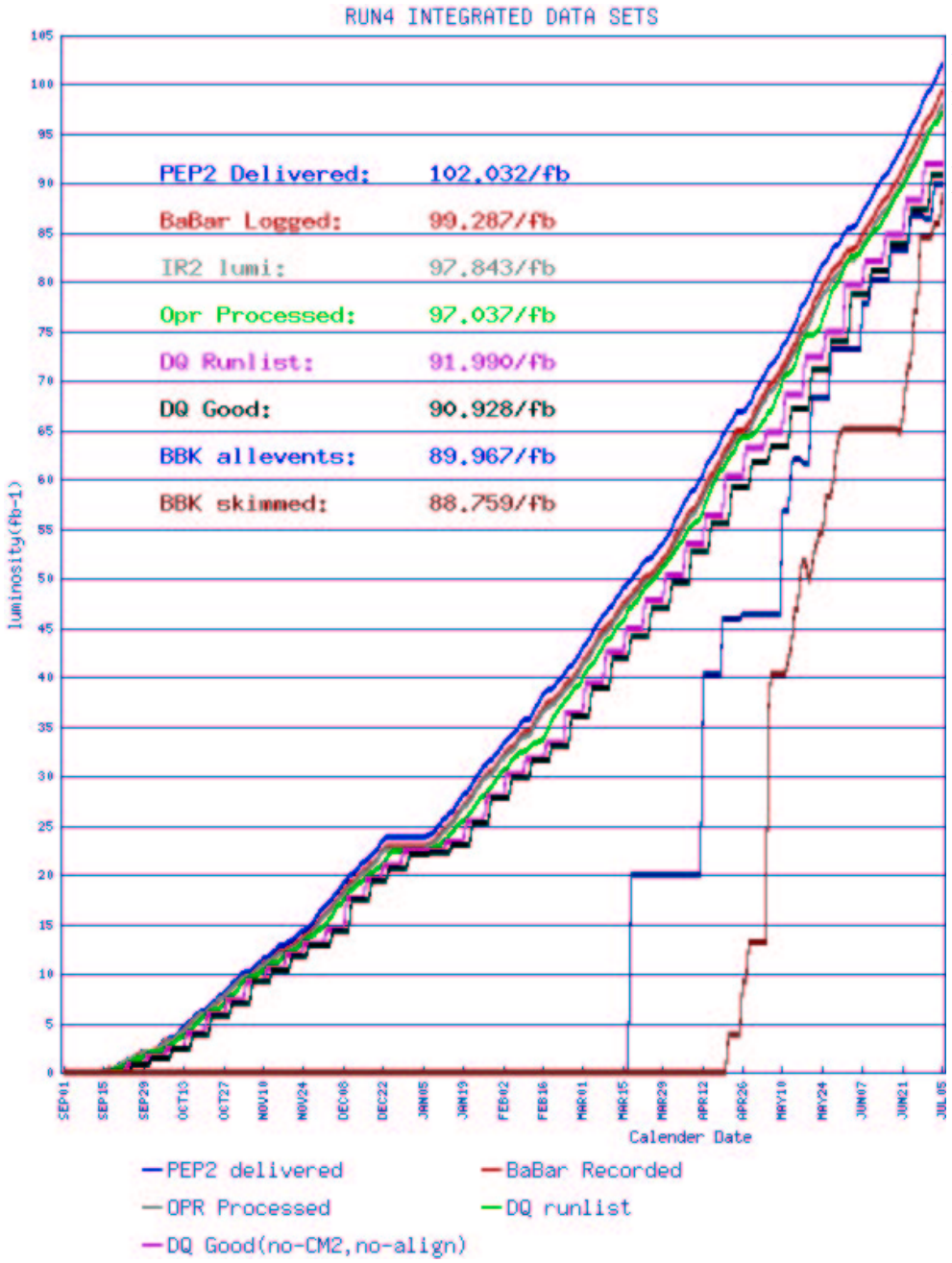
- Run4 data were written out in CM2 format (no more Objectivity event store)
- SP6 was generated in CM2
- Run 1-3 and SP5 were converted to CM2 format
- New skims were defined and created

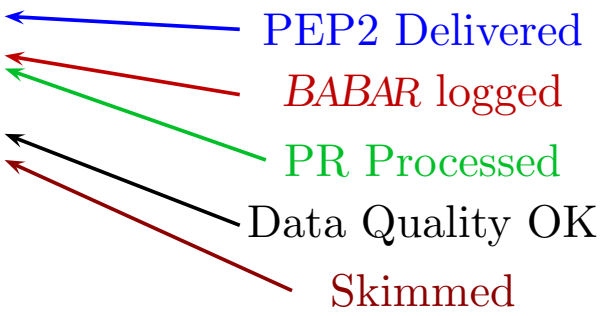
Large improvement in data handling :

Turn around of data (from the time data is taken to the time plots are available to experts) is typically 24 hours. The generation of the signal SP6 is available to user within a week after approval of request.

As with all new developments, large adjustments and tuning of the Production chain were needed. We had to find and fix problems in parallel with accumulating data.

Available Data Set



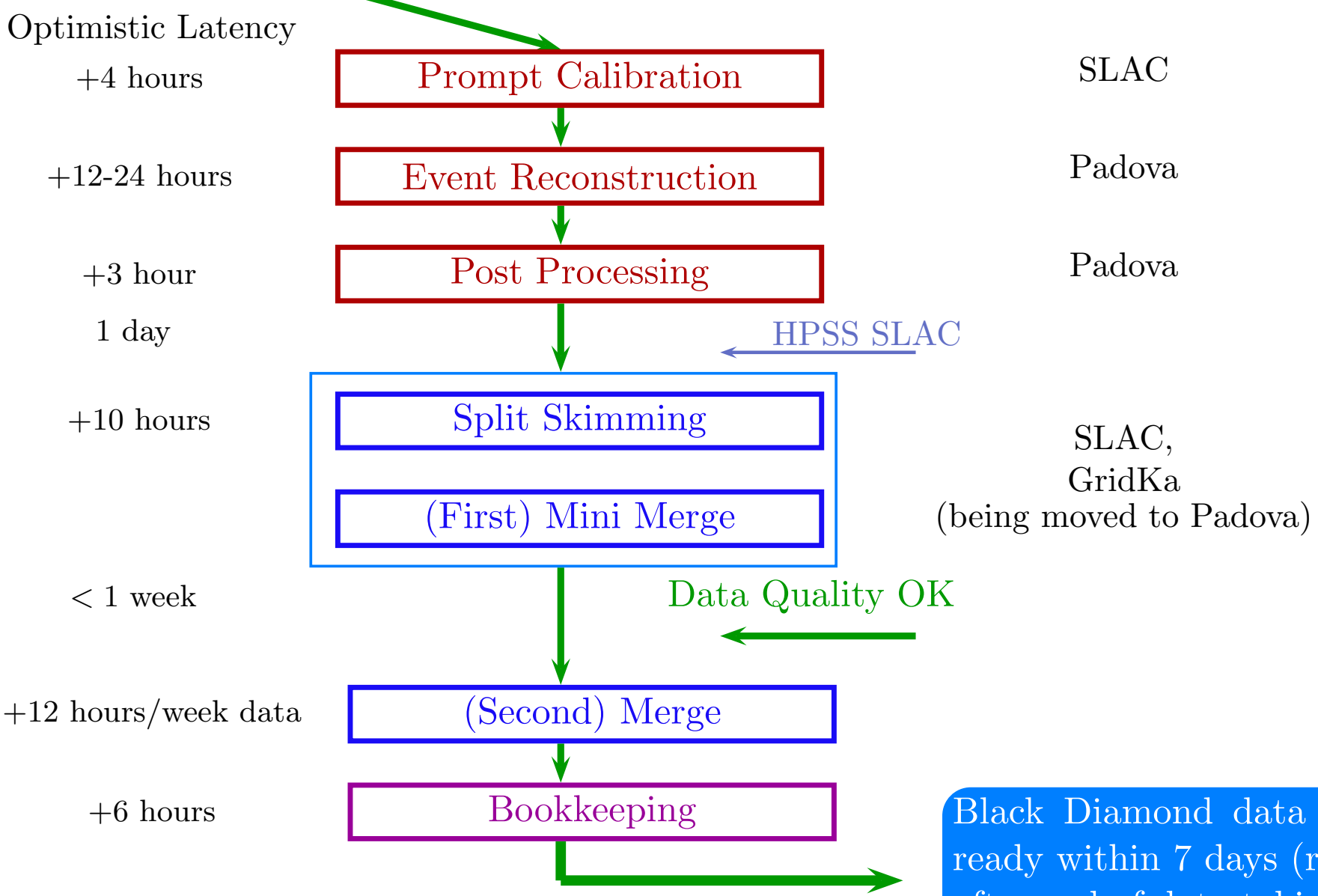


Run 4 data

- 100 fb^{-1}_{July7} are recorded (due to excellent work of PEP-II, *BABAR* detector and online teams).
- *Green Circle* data ($\mathcal{L} = 59.3\ \text{fb}^{-1}$) since end of May.
- *Blue Square* data ($\mathcal{L} = 85.1\ \text{fb}^{-1}$) since July 1 (almost)
- *Black Diamond* data threshold is middle of next week; + 1 week to processes

Run 1-3 is 122.45 fb^{-1} on- and off-peak (*Green Circle*)

Data Production Scheme



Black Diamond data should be ready within 7 days (realistic) after end of data taking

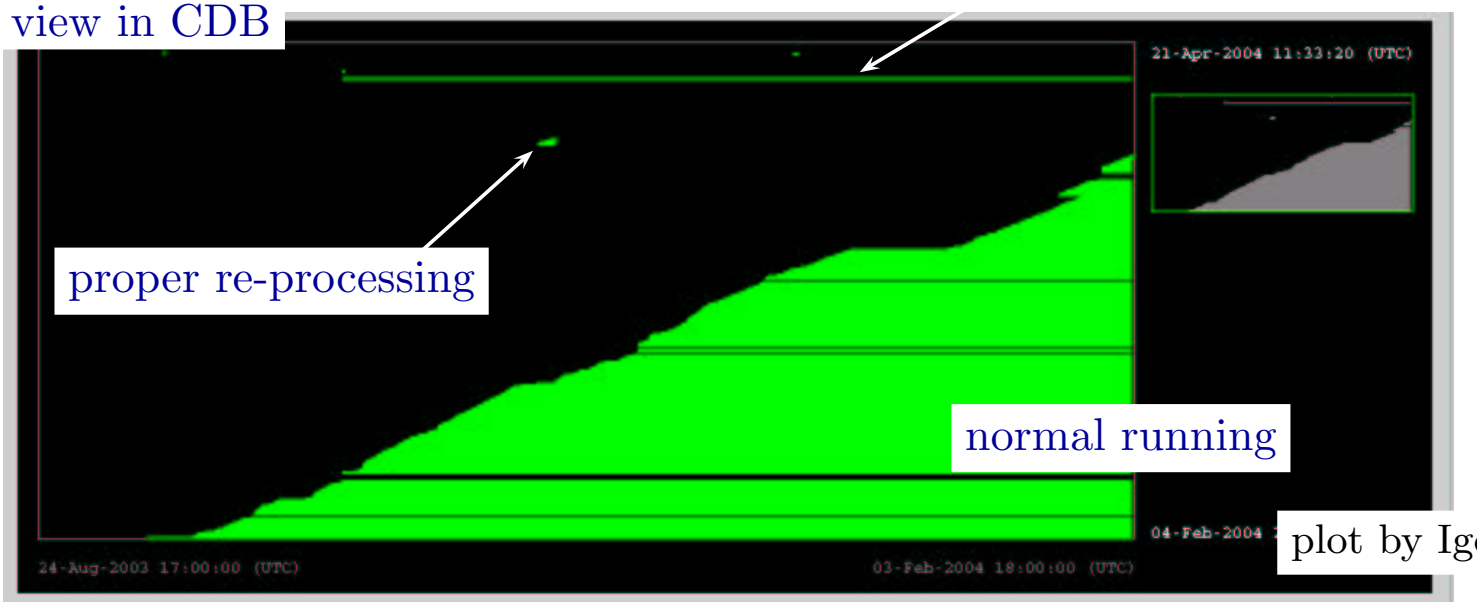
Prompt Reconstruction Overview

- Performs prompt calibration (PC) and event reconstruction (ER) of data
- All data are written out in CM2 format (very positive experience). The output is available to DQG within 1-2 days.
- No more troubles with Objectivity event store, but special care is still needed to manage Conditions DB (specially since it became very transperent to users). Large development is ongoing to stabilize PR interaction with DB. One of them is to introduce additional validation of constants (similar to validation of new release) when new set is loaded into DB (watch for upcomming announcement in IR2PROD-hn).

Illustration of DB problem (long solved)

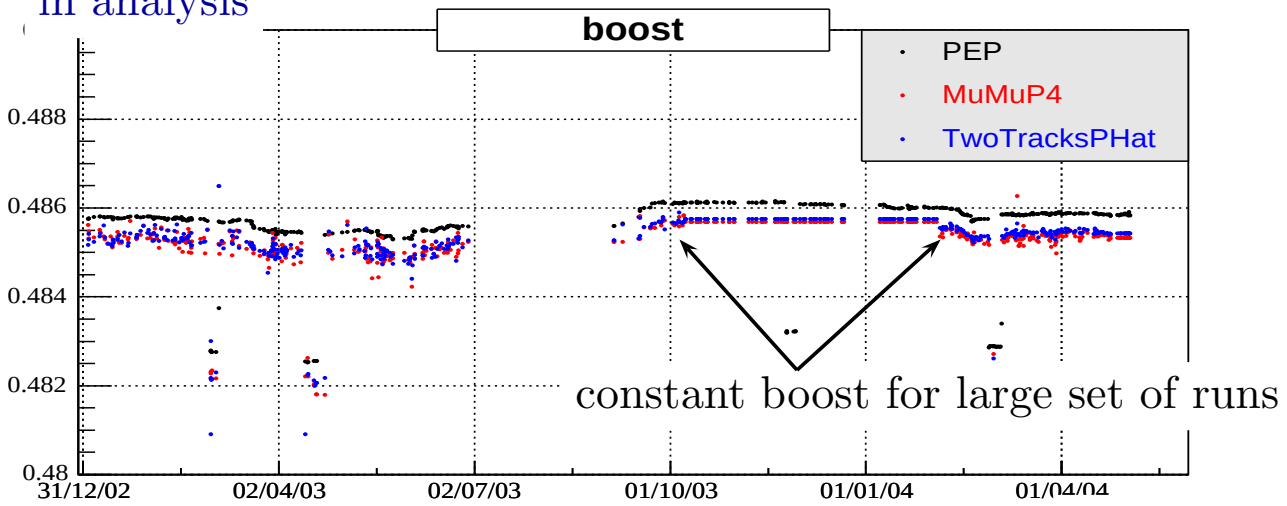
re-processing with wrong settings

view in CDB



plot by Igor Gaponenko

in analysis



plot by Wouter Hulsbergen

Prompt Reconstruction Overview

- Another famous issue is *crashed-elven problem*. Elf (executable for reconstruction) always used to crash on some events. In Objectivity those events were just skipped, while the rest of the run was usable. In CM2 this does not work yet and whole run become unusable. With big effort of Shahram Rahatlou and package coordinators most of the crashes are fixed. Extended collection syntax allows to remove bad run from merged collection. The fix to handle crashes correctly is being worked on and will be in use in Run 5. The number of remaining problematic runs in Run 4 is less than 20.

- *early Run 4 data* reprocessing has started Feb 4 and is done now. Production is concentrated on the most recent runs.

- PR runs smoothly, analysing about $800pb^{-1}$ per day (record - 1.4 fb^{-1}/day) Transfer of xtc and root files can become a limiting factor, but it still keeps up with IR2.

SP Overview

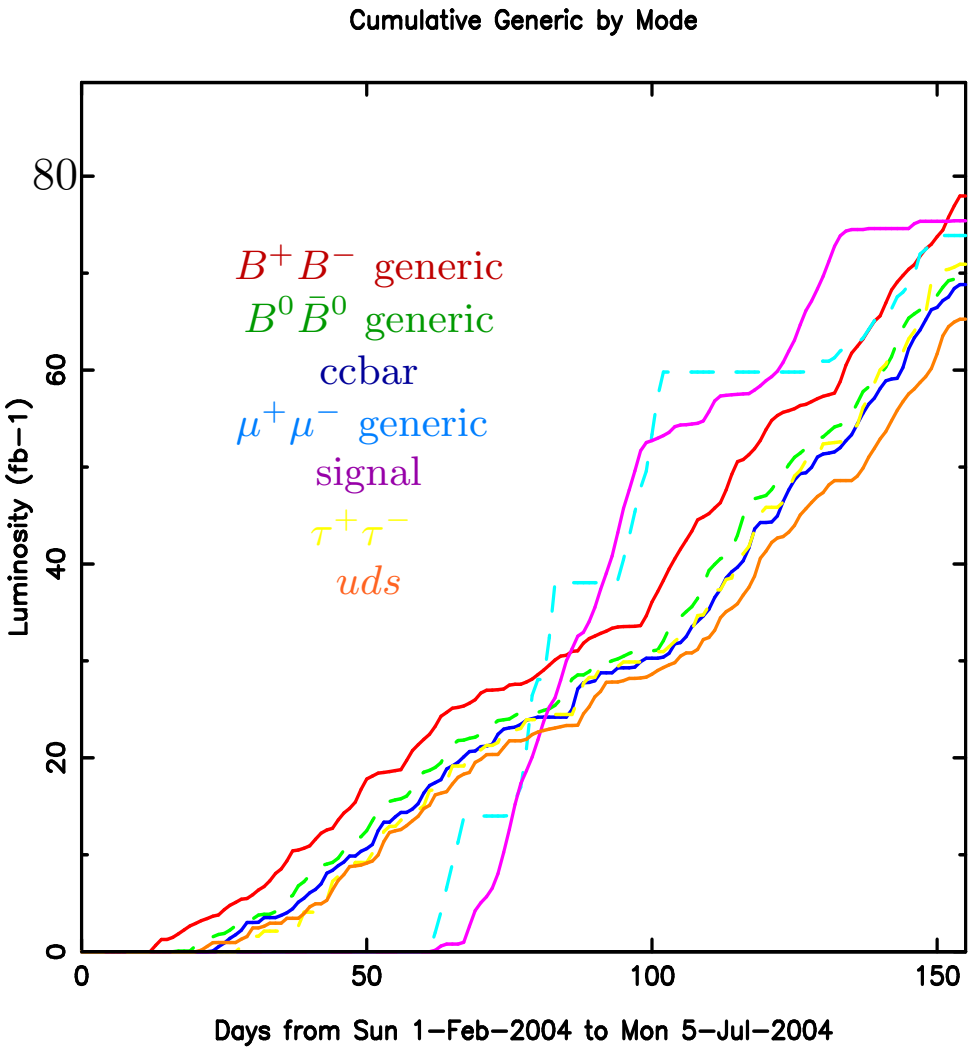
- SP5
  - Main production is done, 2.2 billions events are available
  - Process only signal MC requests for Run 1-3 (SLAC and INFN)
  - Uses Objectivity for generation, then events are converted to CM2 and passed to skimming group.

- SP6
  - Main focus of the production. Need about 1.1 billions events (71% done)
  - Should be done by early September. Signal MC requests (for Run4) will be processed beyond
  - Everything is done within CM2. Great flexibility and fast turn around (within a week for signal request).

SP production goes well.

SP6 is MC production for Run 4 conditions

3x lumi for $B\bar{B}$, 1x lumi for others



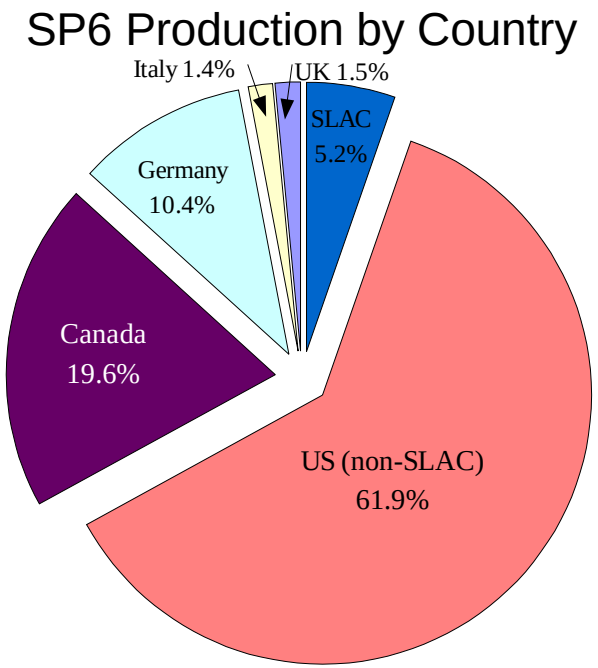
← Sep2003-Apr2004 done

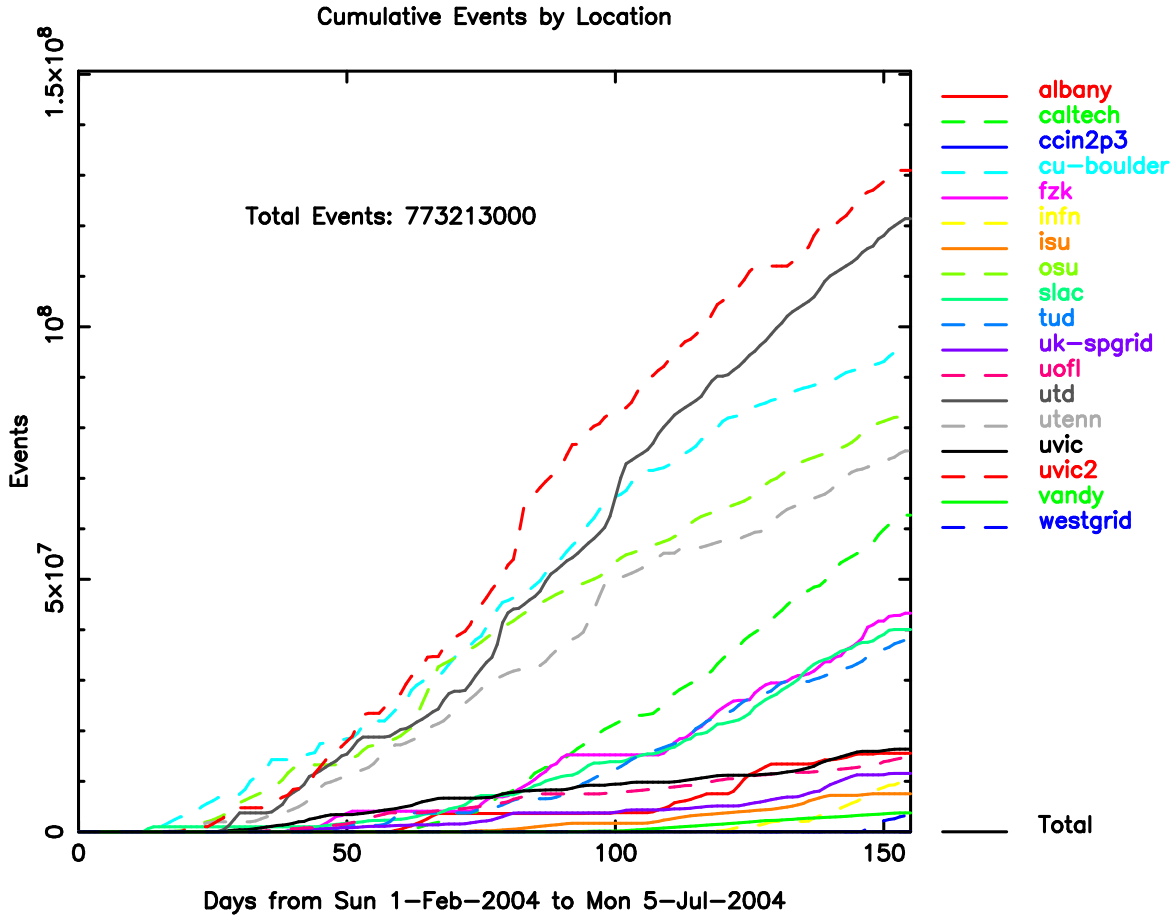
Starting on May2004 simulation

Sample	generated	Lumi	bookkeeping
BBbar	235M	3x 75fb^{-1}	213M
uds	138M	66fb^{-1}	125M
ccbar	90M	69fb^{-1}	81M
tautau	69M	73fb^{-1}	61M
signal	246M	∼75fb^{-1}	237M
Total	778M		717M
Expect	∼1100M		∼ 1100M

Simulation Production Effort

The MC simulation is distributed between 27 sites (30 people, 2000 CPU)





Currently running very stable at 50M events/week (capacity up to 60M events/week).

Skimming and Merging

- Skimming and Merging effort follows the *conversion* to CM2 of Run1-3/SP5 but includes also Run4 and SP6 data.

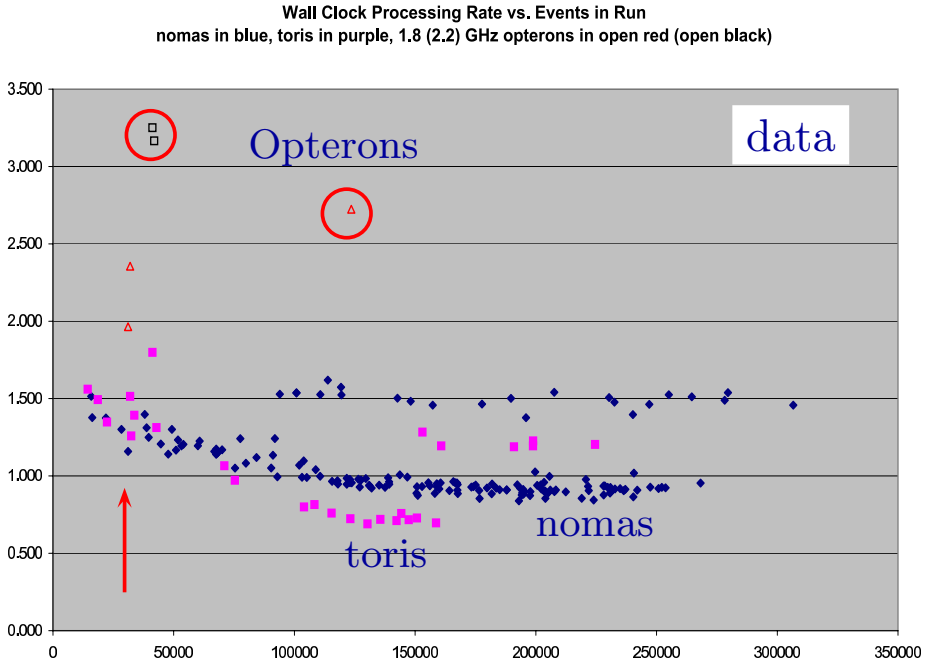
- Data Samples
  - 125 fb^{-1} Run 1-3 data converted, skimmed and merged.
  - 1.4 billion SP5 events converted, 792.3M skimmed and merged (large part is affected by recent *tagbit problem*)
  - 88 fb^{-1} Run 4 data skimmed and merged (25 fb^{-1} is affected by *tagbit problem*)
  - 173M SP6 events skimmed; 149M events merged

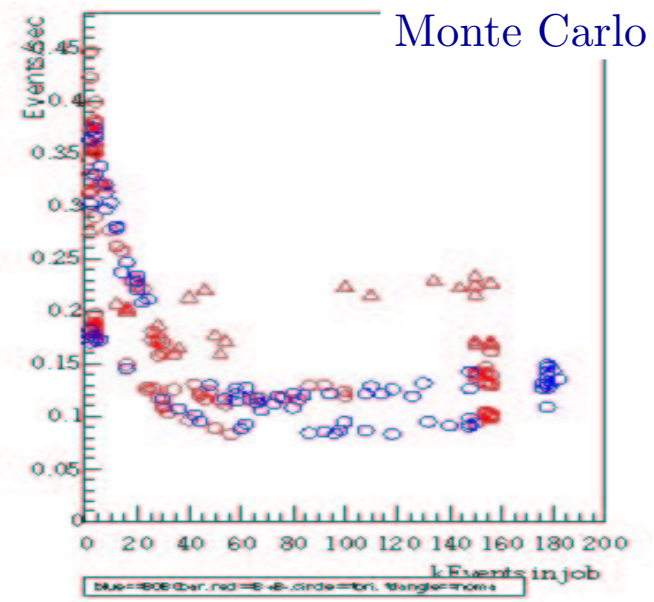
- Projected total data sample : 225 fb^{-1}

CM2 Data (Micro)	8.4 TBytes
CM2 Data (Mini)	19.4 TBytes
CM2 Monte Carlo (Micro)	25.3 TBytes
CM2 Monte Carlo (Mini)	37.0 TBytes
Deep Copy Skims (Data + MC)	129.3 TBytes
Total	219.4 TBytes

Skimming rates

- The processing speed depends on the file size. Split runs into subsamples (split skimming).





- Opteron PC shows very good performance (256 PCs to be ordered).
- More that 1200 CPUs are used at SLAC, 200 at GridKa. IN2P3 will continue to convert remainder of MC.
- Padova is skimming newly reconstructed data. This part of skimming effort is to be included into PR and to be used for Run 5.

SLAC *miniq* batch queue

Latest discovery: *Blue Square* tag problem

- As it was found last week, that new way to skim and merge (with split skimming) resulted in the unreliable tag information within event. This affects latest 25 fb^{-1} of Run 4 data and possibly large part of MC skims.
- The events in the skims are good. All events in the given skim are the one which belongs there and none are lost, but the tag information for these events should not be used.
- it was found that the problem with tags developed during the second merge, while *pre_second_merge* skims show no problem. New release 14.4.3d with a fix was prepared. If no further problem will be found, the *Blue Square* data should be repaired by the end of the week (*BlueSquarePrime*).
- **Today** BbkDatasetTcl will give both *BlueSquare* collections and collections to be included in *BlueSquarePrime* if you ask for *-Run4-OnPeak-R14 data. Should be fixed tonight.
  Make sure you don't analyse both sets simultaneously!
- All users are encouraged (as always!) to look on data and report any problems or bugs or even suspicions.

Manpower

- Production for Run 4 was lead by Howard Nicholson. Now he is going back to his students and Akram Khan will take over. Many thanks to Howard for great organization and support!

- Fred Blanc will pass responsibilty for SP coordination to Dirk Hufnagel (OSU) in autumn.

- PR and Skimming groups are searching for new managers and developers (essential for successfull start up of Run 5). There are many interesting projects, the level of involvement can be varied from marginal to very significant.

Outlook

- PR, SP and Skimming are running all right. The transition to CM2 was quite successfull, in spite of delays and problems. Run 5 promises to be much easier.

- PR keeps up with IR2 data rate, and SP6 keeps up with PR. Main focus of the Skimming group is still SP5 and SP6 skims, although latest Run 4 data will receive special attention.

- Middle of next week will be the end of the on-peak data taking (Black Diamond deadline). Within a week after that the data should be processed by PR, Skimming, Data Quality groups and be ready for analyses.

- preparation for Run 5 has started. August will be used for extensive testing of the system. 16 series are to be exploit for the Run 5 and for SP7, while 18 series are to be prepared to reprocess whole Run 1-5 sample and for SP8. (see Stephen Gowdy talk for details)