
# Babar - the last dataset

Douglas A. Smith, Homer Neal, Gregory Dubois-Felsmann

*SLAC National Accelerator Laboratory*

for the Babar computing group


# Babar history

- Babar started measuring colliding beams on Oct 22, 1999, and continued over the course of 7 run cycles until Apr. 7, 2008.
- In current dataset : 541 fb-1 of luminosity, 36842 runs and 9.1e9 events.
- Most of this data (90%) was measured at the Y(4S) resonance, in the last year data measured at Y(3S) and Y(2S) resonances.


# Data Production

- Stages of production
  - Prompt Calibration (PC) - Calibrate conditions using a given number of events from run data.
  - Event reconstruction (ER) - Create data for analysis, and strip background events.
  - Simulation production (SP) - Use conditions and background events to simulate data.
  - Skimming - Skim all good data into multiple streams for separate analysis.


# Many production cycles

- Using even number major release, first results with release 8 in early 2001.
- Done many times over the years, using even major releases until the current release 24.
- Some releases used for complete reprocessings, of all current data, or partial reprocessing on parts of the data, or only certain type of production (because of limited time and resources).


# The Final Reprocessing

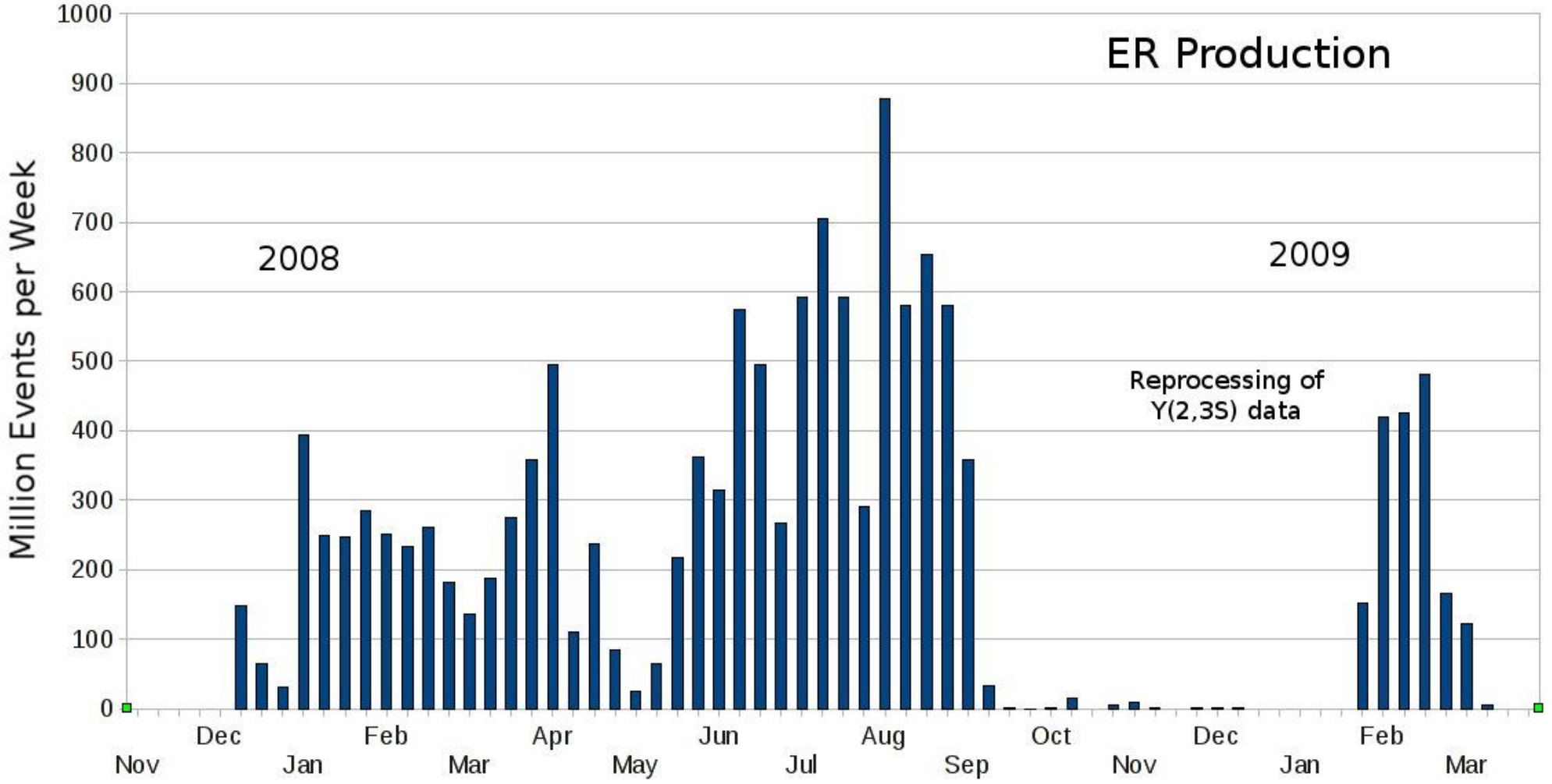
- The "final" reprocessing uses release 24.
- Data is over but there needs to be a good complete dataset for analysis using one release.
- For various time and resource limitations, this was never true for long.  To analyze all data required different releases on different datasets.
- Production of "final" dataset started beginning 2008, mostly finished, more still going on.


# Data for analysis

- PC and ER are divided up into different farms.
- PC done at SLAC, ER dist. to SLAC and Padova, with about even computing resources.
- Production size:
  - ~40,000 jobs
  - 9.14e9 events
  - 80 TB of data, in 93,000 files.


TM & © Nelvana

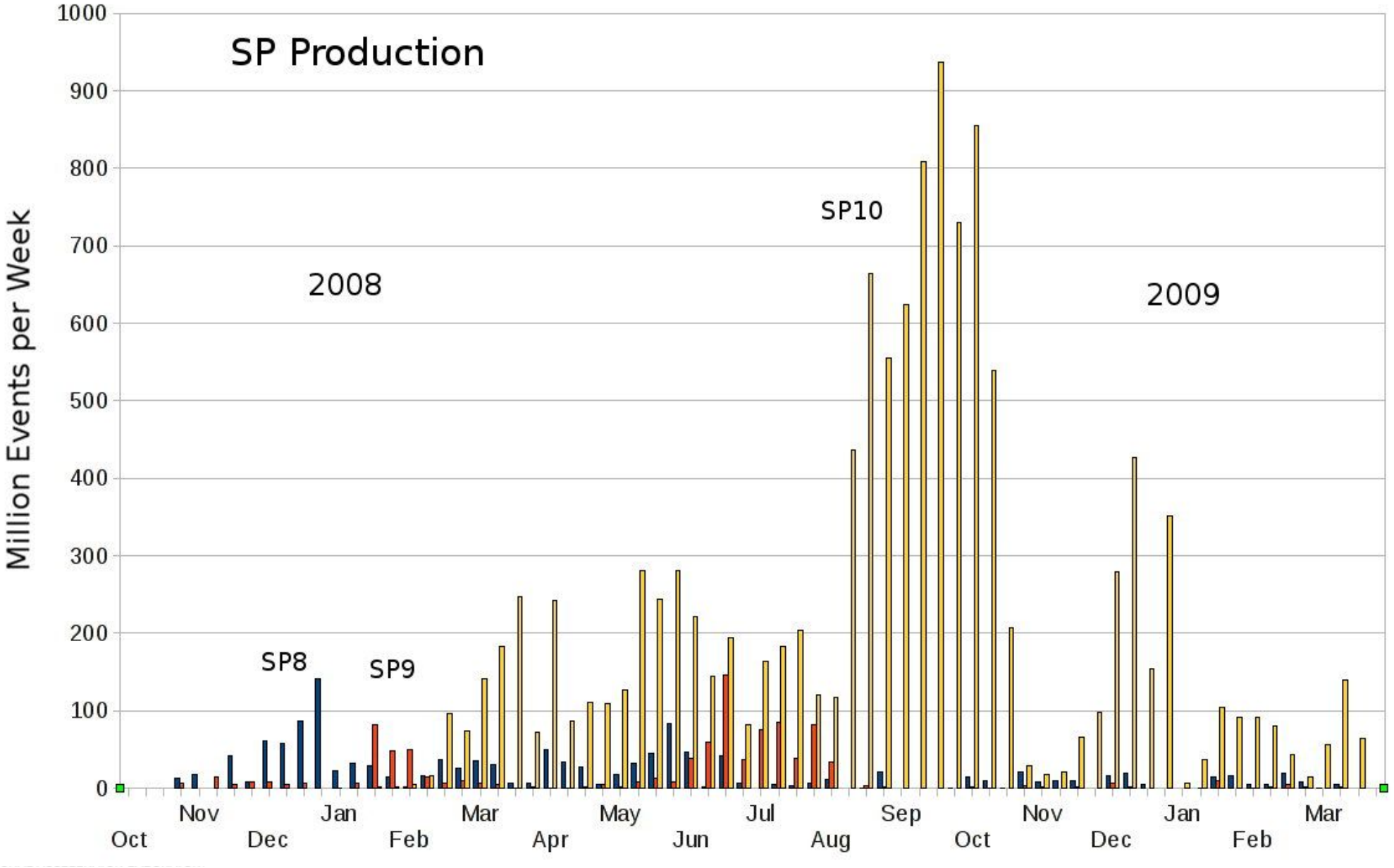
# PR - graph


ER Production
Million Events per Week
1000
900
800
700
600
500
400
300
200
100
0
2008
2009
Reprocessing of
Y(2,3S) data
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar


# Simulation

- Uses conditions and background events as input, needs to wait for these before it can start.
- Produces data from various decay modes, a few thousand in use for each cycle.
- Request 3 times lumi. for B-pair, 1 times for the rest.
- Jobs defined for 8-10k events each.
- Distributed to production sites, ~14, depending on ability to produce.
- Production size:
  - 1.94M jobs, 12.89B events, 1341 years of cpu time.
  - 166Tb of data, stored in 250k files.

# Simulation graph

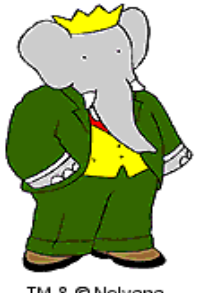



# Skimming

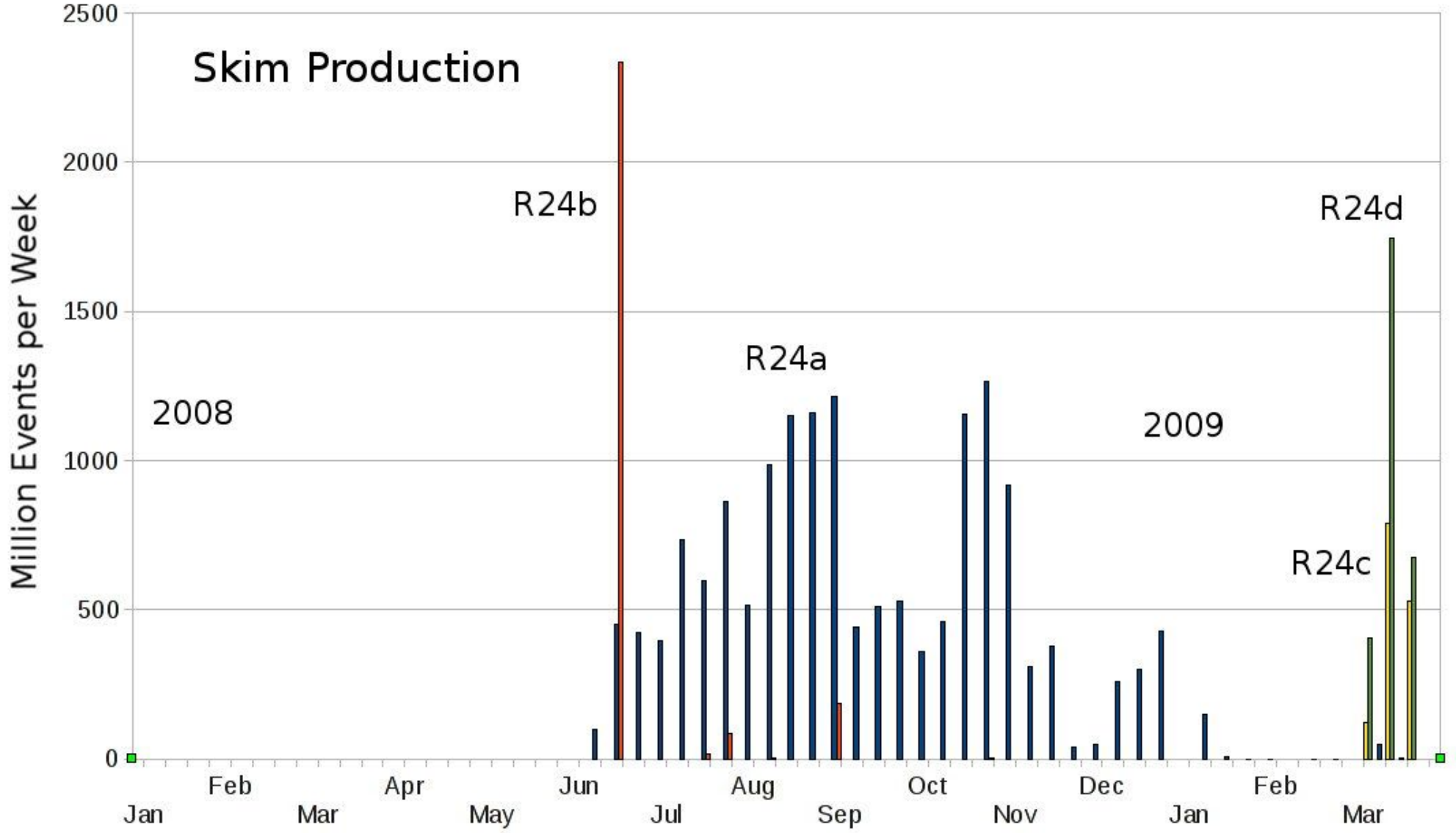
- Data is skimmed for various analyzes.
- Some skimming copies part of data, some only points to events, some skims add data to the events.
- This can multiply the data size, and greatly multiply number of files, need to be careful to keep these as low as possible for final dataset.
- Skimming is done in cycles, each of which runs over the all good ER and SP data.


# Skimming more...

- Number of streams per cycle changes for each cycle. For R24 there are now 5 production cycles, each about 30-40 production streams.
- Data as produced and good is given to skim tasks, these tasks are distributed to SLAC, GridKa, and Manchester. For resources reasons most (~90%) of skimming done at SLAC.



# Skimming graphs




# Current state of dataset

- Current dataset of good for analysis data

| Coll. | Type | Events (B) | Files | Size (TB) |
|---|---|---|---|---|
| 115 | Bkg | 0.012 | 266 | 0.240 |
| 1100 | Nonevent | 0.0 | 1100 | 0.112 |
| 104271 | PR | 11.810 | 160677 | 101.55 |
| 128468 | PRskims | 55.663 | 166972 | 86.26 |
| 82142 | SP | 10.064 | 196517 | 137.04 |
| 198250 | SPskims | 59.392 | 273585 | 142.46 |
| 514346 | | 136.949 | 799117 | 467.67 |


# Current production

- ER mostly done (>99%), small problem jobs getting done, code fixes to handle run data that crashes executables.
- Main SP request complete, now producing better set of simulation for last year's Y(2,3S) datasets, done soon.
- Two new skim cycles (R24c, R24d) recently started, and all other cycles are open, and data is skimmed as it is produced.
- Partly presented at winter conferences, current effort to present full dataset at summer conferences.


# Near future production

- Requests for a larger simulation dataset, there are resources for use until summer. Hopefully double or more certain datasets, and will get used in analysis as needed.
- New simulation for new decay modes. This goes on for about 2 years after the first large scale requests are complete.
- New skim cycles will get defined as there are changes to analysis code, defining new streams. And all skim cycles will need to run over data as produced.


# Future of dataset

- Directly accessible for 2-3 years, for analysis.  Mostly this will all fall back to SLAC.
- Accessible in some manner for 5 years, this is up to discussion:
  - Current supported build platforms (SL3, SL4) will go away before long.
  - Updates for SL5 that will give us a couple more years.
  - People are investigating virtuallized platforms, releases will lose man power to update to SL6.
- After this needs to get stored somewhere, the dataset is a large scale expensive investment.


# Summary

- At the end a "final" dataset is needed.
- ER done now for this, SP almost done, and skimming will keep up.
- More to get done, large scale production will drop off to small amounts by fall.
- Data will be accessible by batch farms at SLAC for next 2-3 years.
- Some ability to run on data for 5 years is an interesting technical problem.
- After that, not sure, require some maintenance for some time...