Realistic, large-scale MC production

M. Moulson, 20 December 2002

Summary presentation for KLOE General Meeting

Outline:

- *Production proposal*
- *Refinements to GEANFI*
- *Background insertion*
- *MC DST's*
- *Production logistics*

General proposal for MC production

500 pb^{-1} of $K_S \to$ all, $K_L \to$ all (about 500M events)

- Can generated and reconstructed in under 2 months
  if done efficiently on IBM farm
- Useful for background studies for all $K_S K_L$ analyses and for $K_S K_L$-contributed background in all other analyses
- Prototype for future production campaigns: K^+K^-, $\phi \to$ all

Goal: Best possible reproduction of time-variable conditions

- State-of-the-art simulation of the detector
  Both in generation and reconstruction
  Run-variable input for $\sqrt{s}$, $\mathbf{p}_\phi$, $\mathbf{x}_\phi$, trigger thresholds, dead channels, etc.
- Realistic machine background, obtained from data run-by-run
- Output in the form of MC DST's

New ϕ-decay generator

Improved radiator function with explicit calculation of matrix elements for all processes with 1 or 2 radiated photons

Photons from both beams

ISR photon tracked by GEANT

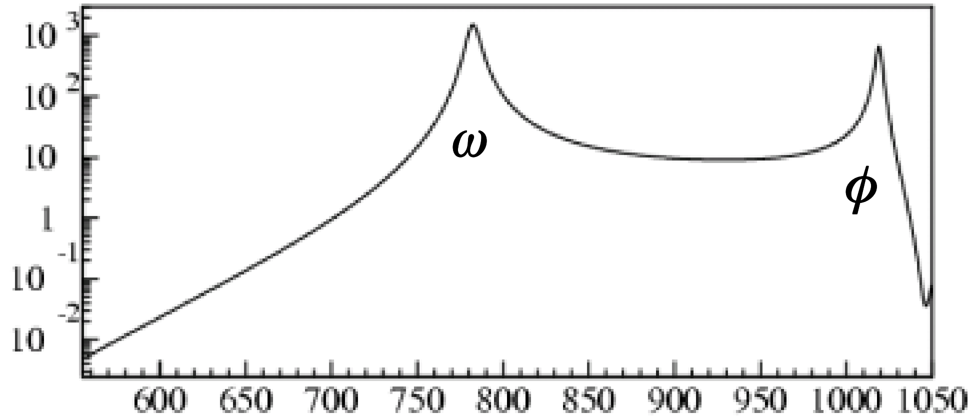
Sampling of $\sqrt{s}$ now precedes choice of ϕ decay mode

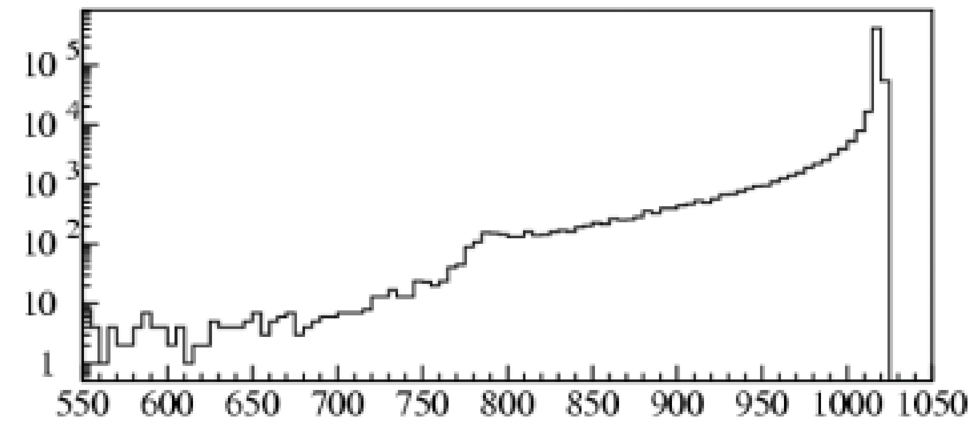
Probabilities for different final states now depend on $\sqrt{s}$

Improved ϕ cross sections:

- Correct treatment of phase space
- ω and ρ terms in 3π cross section

$\sigma(e^+e^- \to \pi^+\pi^-\pi^0)$



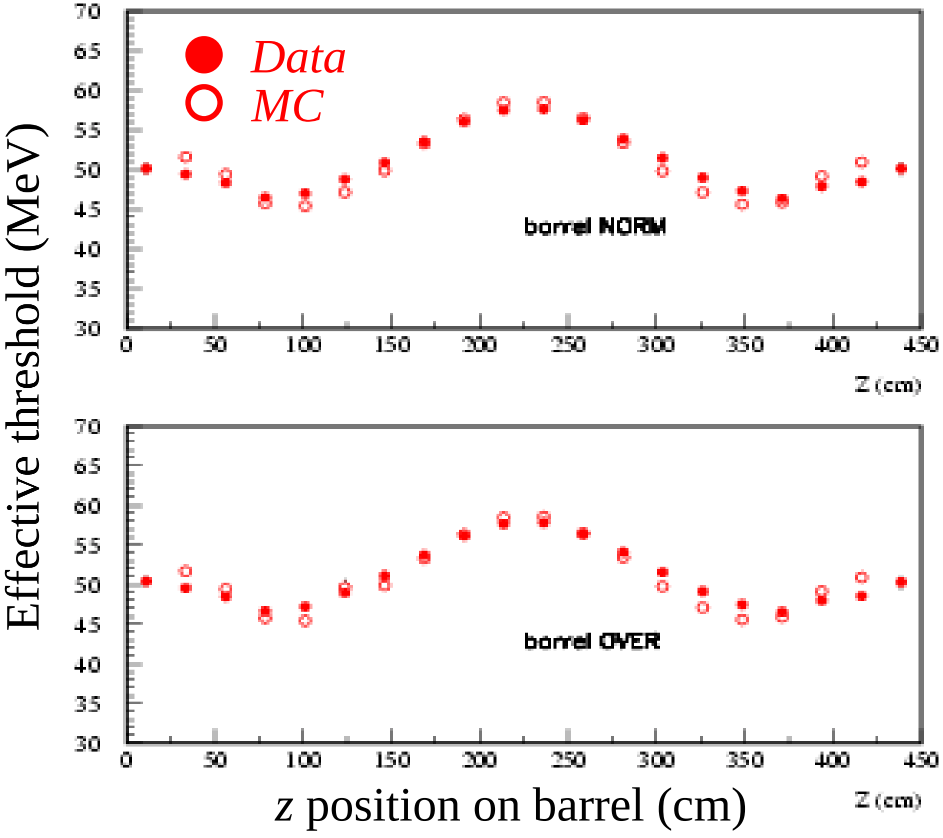


$dN/dM(\pi^+\pi^-\pi^0);\ \sqrt{s} = m_\phi$

Simulation of EmC trigger

Careful tuning of effective thresholds in TSKT (cluster energy needed to fire a sector)

γ's from $K_S \to \pi^0\pi^0$ on barrel



Photons from $K_S \to \pi^0\pi^0$

- Variation of effective threshold with position well within 10% all over detector
- Somewhat better on barrel

Pions from $K_S \to \pi^+\pi^-$

- Variation at same level as for γ's
- Effective threshold for MC few % less than for data

Implications for EmC trigger efficiency estimates from MC:

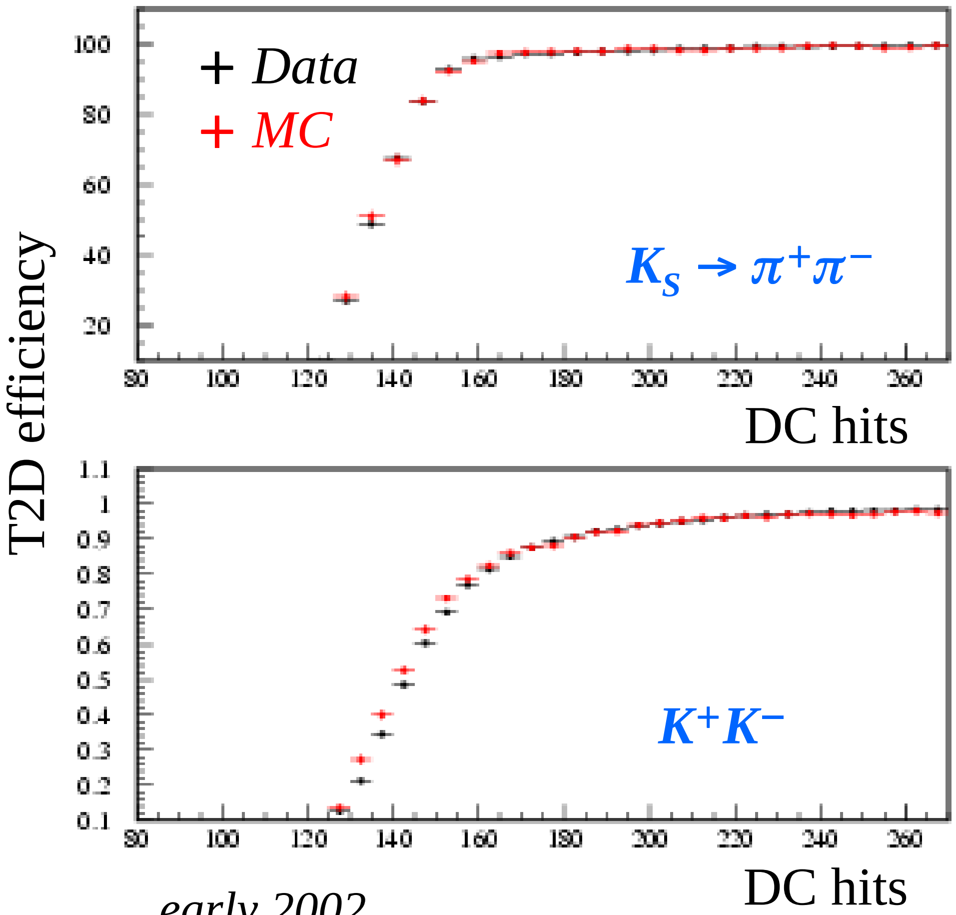
ΔE_{th}(overall) = 5% means

$K_S \to \pi^0\pi^0 \qquad \Delta\varepsilon = 0.5\%$

$K_S \to \pi^+\pi^- \qquad \Delta\varepsilon = 1\%$

Simulation of DC trigger

T2D threshold in TSKT tuned based on effective number of reconstructed DC hits



early 2002

For MC: background from $\gamma\gamma$ events corresponding to data-taking periods added using INSERT

RMS variation of effective threshold over 10 intervals (1 pb^{-1}) from 2001-2002:

K^+K^-	2.5 reconstructed hits
$K_S \rightarrow \pi^+\pi^-$	1.9 reconstructed hits

Implications for DC trigger efficiency estimates from MC:

$K_S \rightarrow \pi^+\pi^-$	$(\Delta\varepsilon)/\varepsilon$ = 0.7%
K_SK_L	$(\Delta\varepsilon)/\varepsilon$ = 1.3%
K^+K^-	$(\Delta\varepsilon)/\varepsilon$ = 2.4%
$\pi^+\pi^-\pi^0$	$(\Delta\varepsilon)/\varepsilon$ = 3.2%

Planned work on EmC response in MC

Survey MC/data differences in 2001-2002 runs

Make diagnostic plots using $\pi^+\pi^-\pi^0$ and $e^+e^- \to e^+e^-\gamma$ samples:

- Energy response and resolution, as function of position
- Timing response and resolution
- Efficiencies
- Splitting and shower fragments

Implement tools to adjust MC response

Probably at reconstruction level:

- Adjustment of energy response in CLUFIXENE
- Threshold simulation
- Simulate “holes” in EmC response

Minimum effort for maximum result:

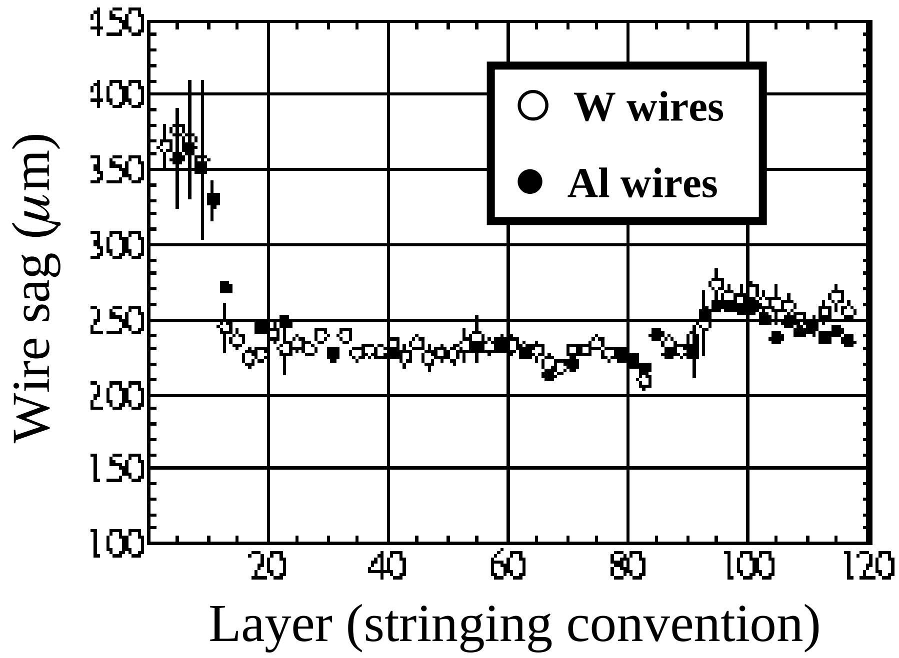
Start with large effects, see how far we get....

Comprehensive DC geometry review

Beam pipe Inner wall	Correctly simulated according to nominal understanding Continue to study with multiple scattering, *dE*/*dx* (conversions?) Inner wall treated as equivalent thickness of CF in reconstruction
Outer wall	Originally simulated as 0.4 cm CF **2 × 20 μm Al plating and CF struts recently added**
Endplates	Geometry correctly simulated **Adding 50 μm Cu plating to simulate FEE**
Other material	**IR support "legs" added to simulation** Material at borders of endplates not added: Ti screws, gas feedthroughs, DC "feet" Not limiting factor in understanding EmC response at endcap/barrel interface
Global shift	IP/DC/EmC currently simulated as coaxial DC shifted by $\Delta y = -1$ cm in real life EmC shifted by $\Delta y = -0.4$ to –0.7 cm in real life Studying feasibility of including these offsets

MC representation of wire sag

Wire sag on most internal layers of DC much larger (400 μm) than on other layers (250 μm)



In generation:

- Constant sag of 250 μm assumed
- *s*-*t* relations from GARFIELD

In reconstruction:

- Sag not taken into account at all
- *s*-*t* relations for data reflect crosstalk between bins in ϕ
- *s*-*t* relations for MC from GARFIELD

Causes characteristic distortion of track momenta: want to reproduce in MC

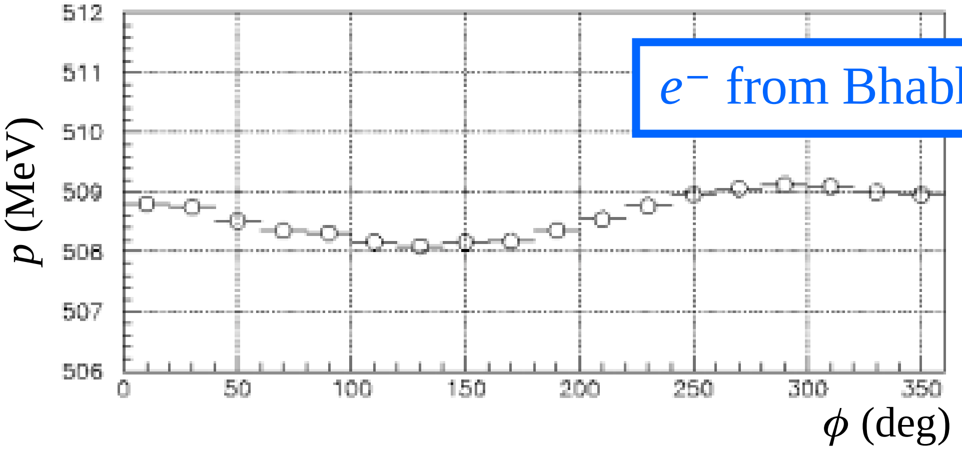
Generation:	Wire sag adjusted to measured value on each layer *s*-*t* relations from GARFIELD
Reconstruction:	Ignore sag, just as in 2001-2002 data reconstruction *s*-*t* relations calibrated using MC cosmic rays Wire sag will be taken into account in future reconstruction

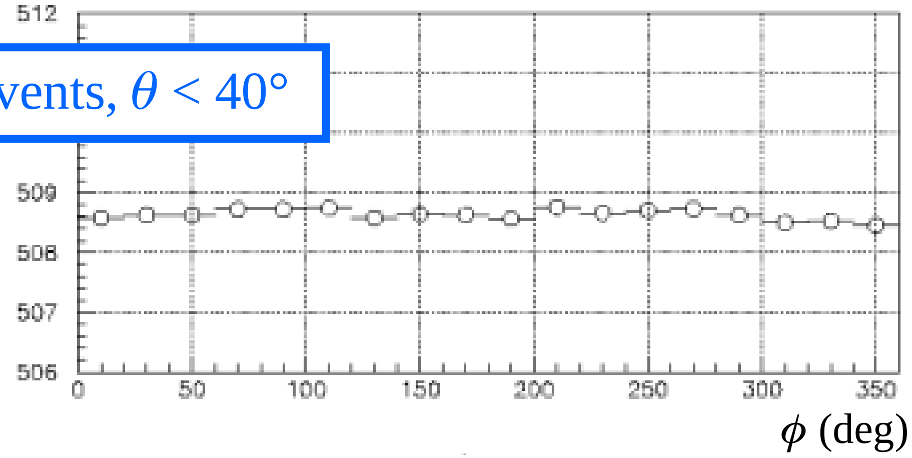
Wire sag and momentum reconstruction

Generated with 250 μm sag, all layers

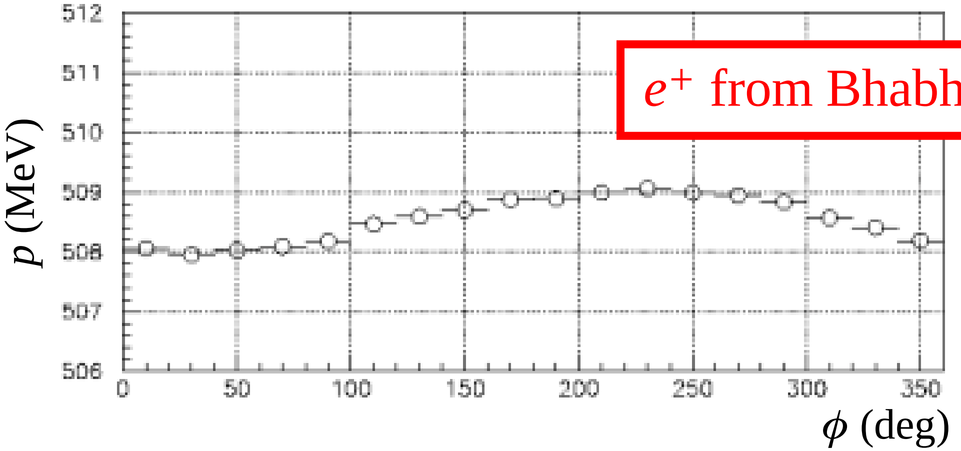
Generated with zero sag, all layers

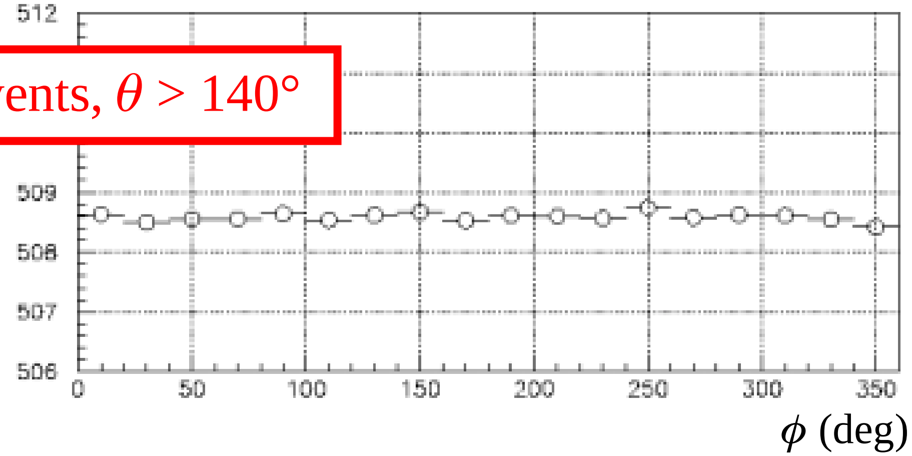
e^- from Bhabha events, $\theta < 40°$





e^+ from Bhabha events, $\theta > 140°$





Background insertion: principles

Previously existing tools for inserting background:

ACCELE: EmC clusters

MBCKADD: DC hits

Both feature selection and insertion phases (modules)

<u>Objectives:</u>

- Complete simulation of background in physical event

  Interplay of hit blocking, t_0 corruption, etc.
- Insert background for both EmC and DC simultaneously
- Unified selection of background events
- Single output file in standard format: compressed YBOS

A/C module for background insertion

New A/C module: INSERT

- Opens background file
- Reads events from the file into secondary YBOS array, reusing events appropriately
- Decompresses (unzips) events in secondary array
- Gracefully handles EOF of the secondary file
- Next step: open/read from secondary file using KID

    Straightforward, just drop-in subroutine replacements.

Handled by INSERT at present:

LRID copied from “background” file into BRID in the “MC” file

DC hits in DTCE in “background” file inserted into “MC” event

Background event selection

Background obtained from recognized physics events and inserted in simulated physics events

→ Sampled evenly as a function of integrated luminosity

Event type should be:

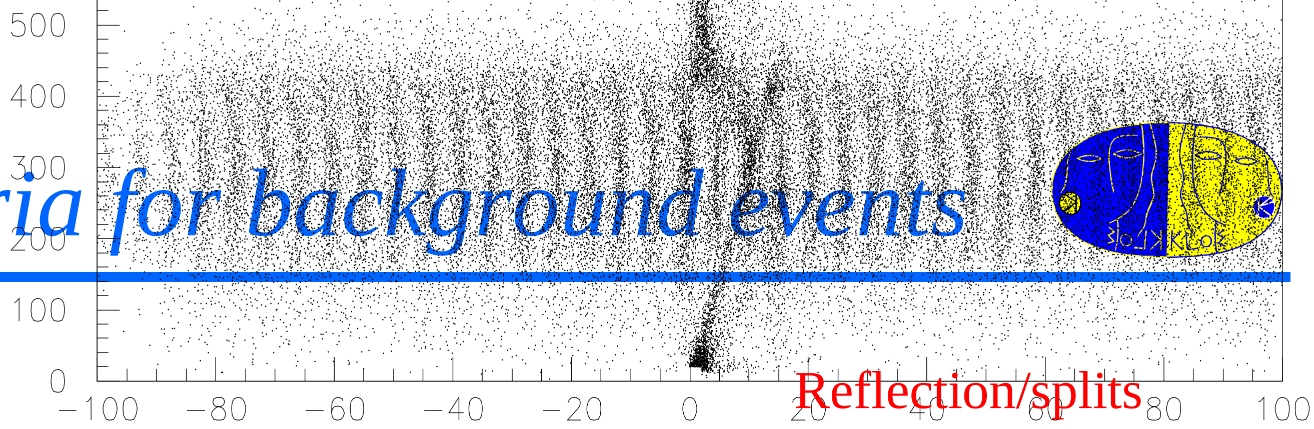
- Relatively abundant
- Easily identified
- Separable from background in DC and EmC

  Isolation easier in EmC: prefer neutral events

Use $\gamma\gamma$ events (solution common to ACCELE and MBCKADD)

Event selection and cluster isolation are closely related problems

Selection criteria for background events



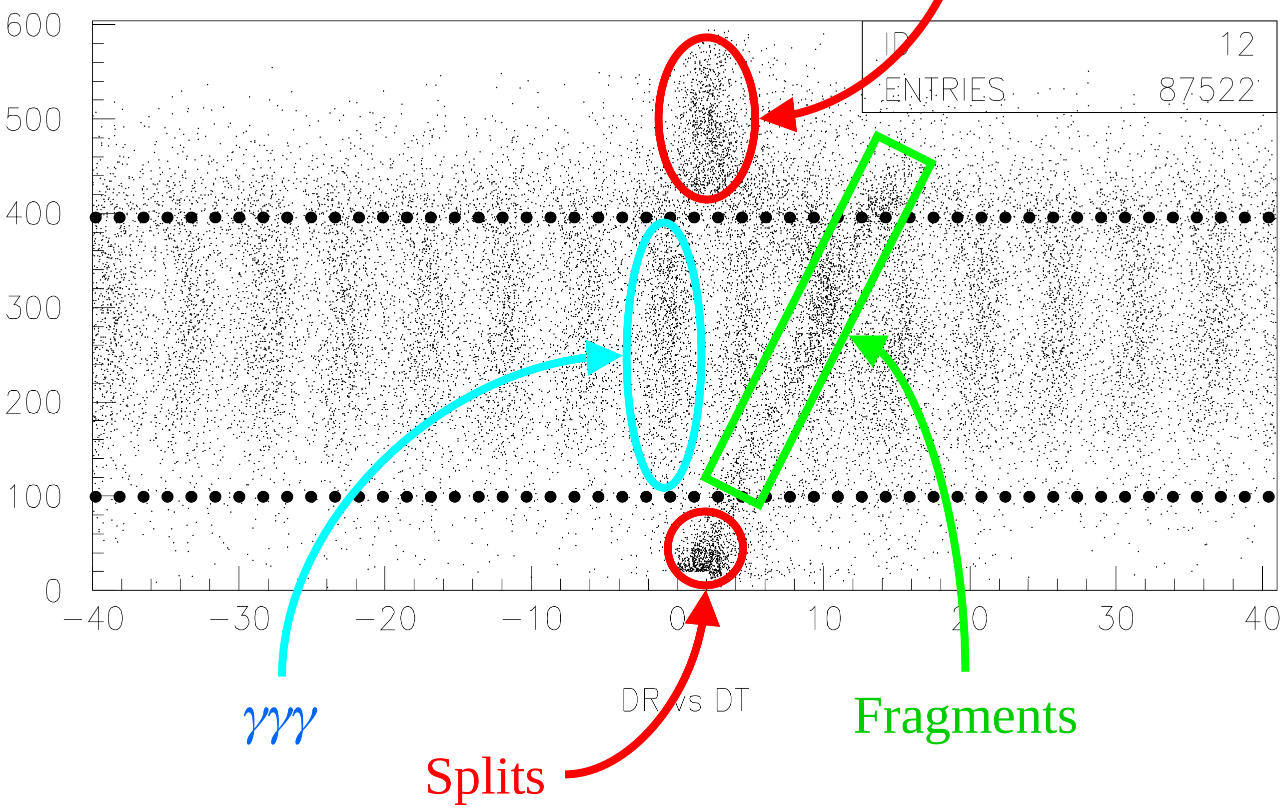
Two clusters in barrel:

- $\Delta t < \min(5\sigma_t, 1\text{ ns})$
- $E_{\rm cl} > 450$ MeV
- $E_{\rm tot} > 900$ MeV
- $|\Delta\phi| > 179°$
- $|\Sigma z| < 10$ cm

Additionally:

$R_{12} > 100$ cm
$R_{12} < 400$ cm
(eliminates splits)

ΔR vs Δt, all cluster pairs



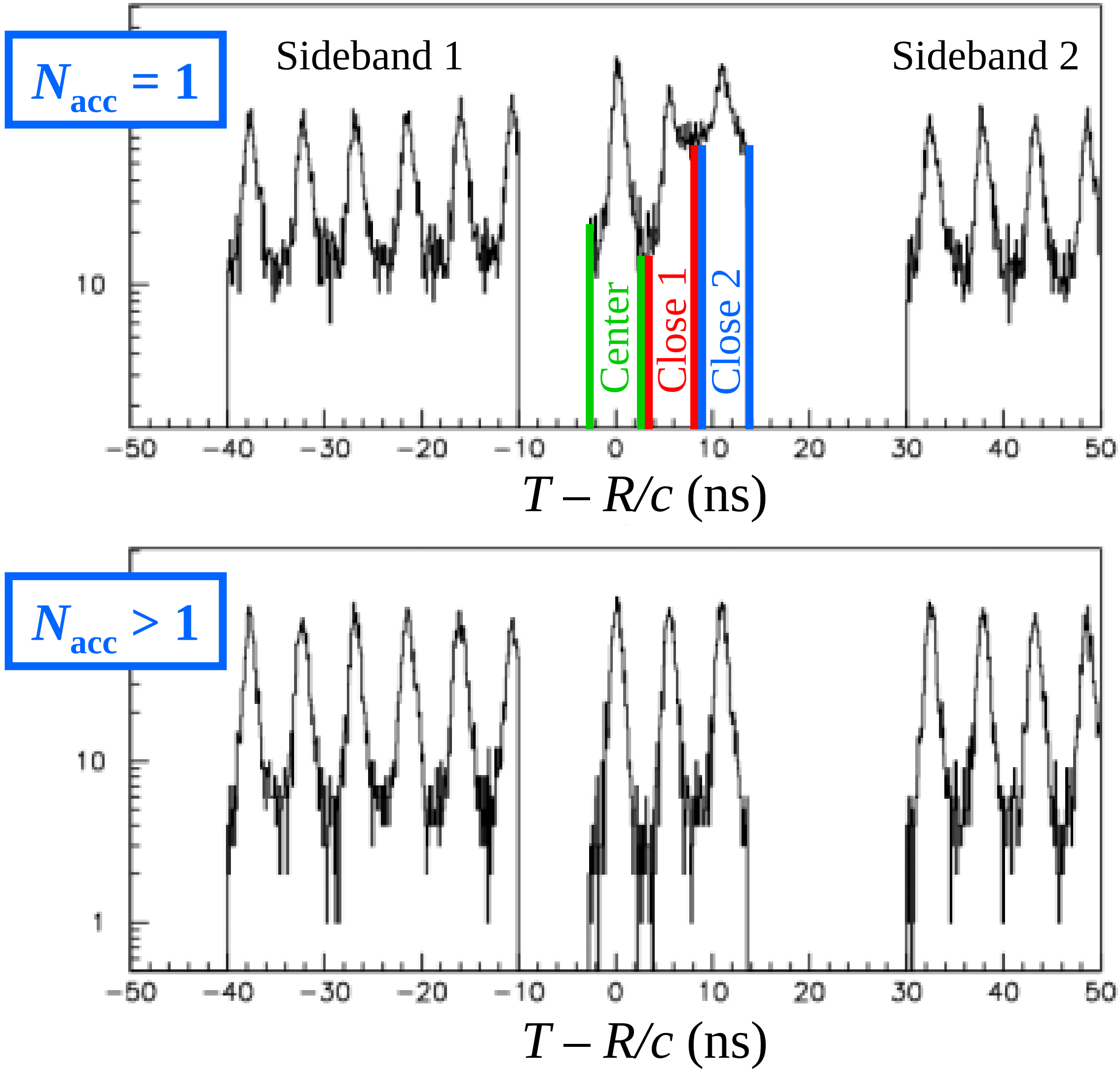
Isolation of background clusters

Excess counts near $\Delta t = 0$ confined to case with only 1 accidental cluster

Previous studies have shown that activity in sidebands 1, 2 has same distributions of E, θ, etc.

Use sidebands to study excess near $\Delta t = 0$

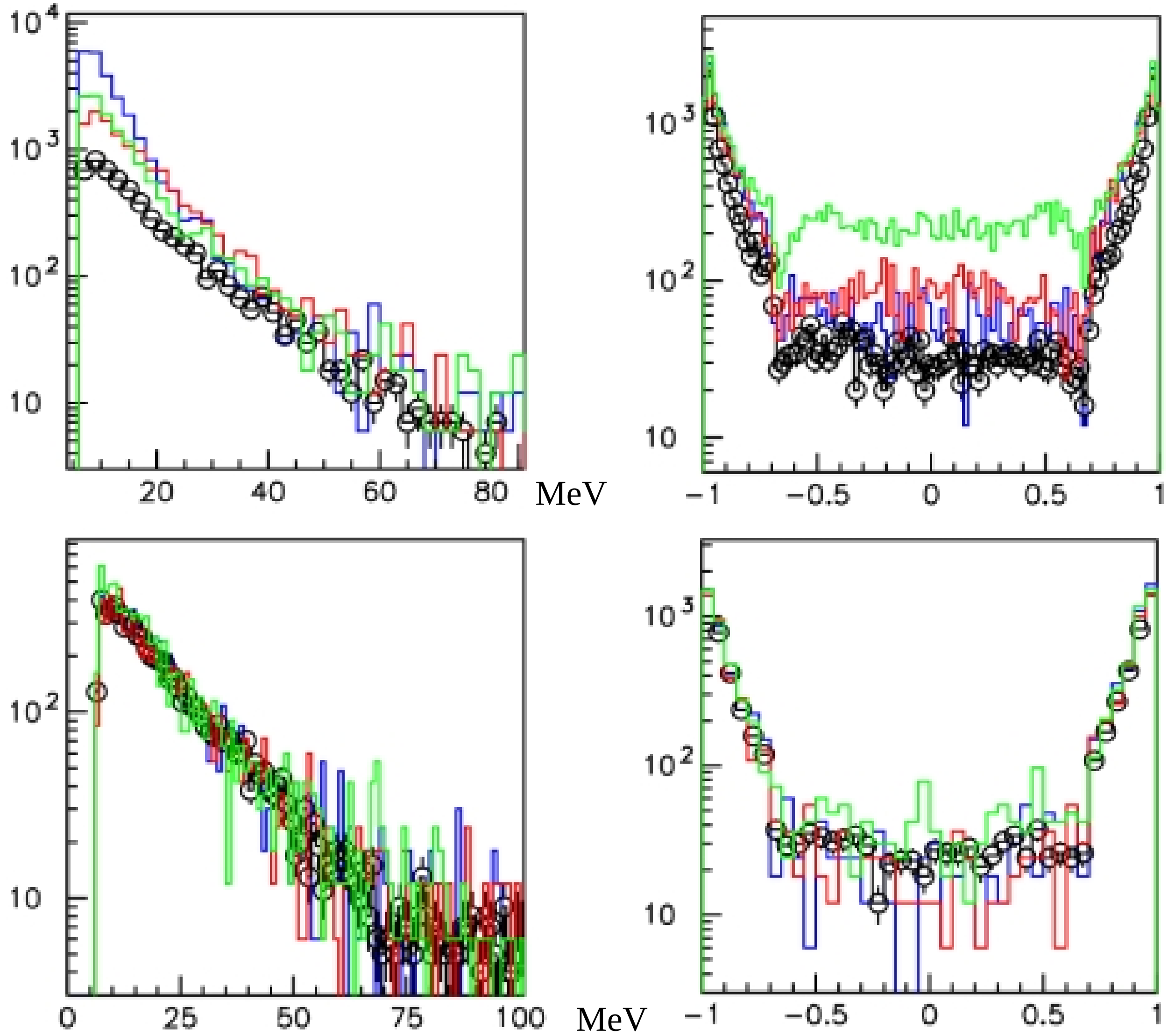
Normalize to width of Δt interval



Analysis of “in-time” clusters

dN/dE
$N_{acc} = 1$

dN/d(cos *θ*)
$N_{acc} = 1$



dN/dE
$N_{acc} > 1$

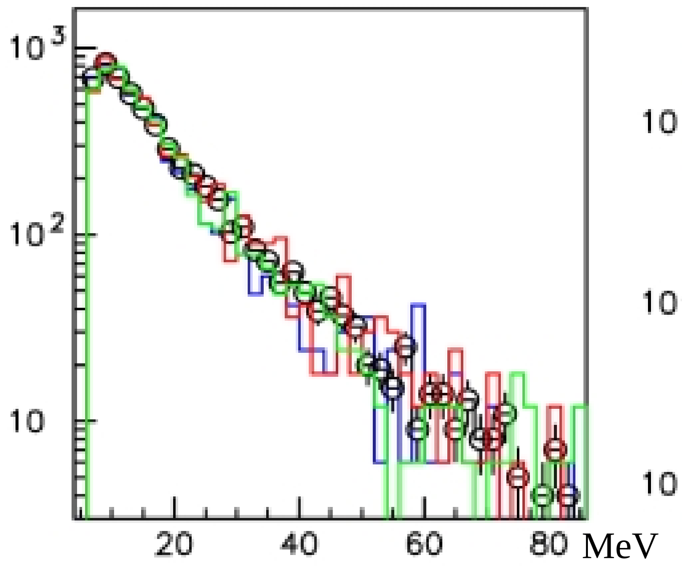
dN/d(cos *θ*)
$N_{acc} > 1$

Reproduction of background distributions

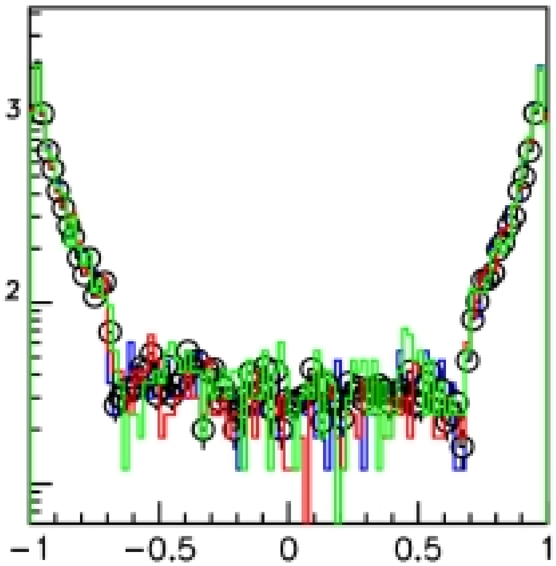
Excess clusters at $\Delta t \approx 0$ for $N_{acc} = 1$ presumably from $\gamma\gamma\gamma$ and cluster fragments:

- Obtain ratio of counts in central region/sidebands as function of $\cos\theta$, E
- Downscale selection of events with $N_{acc} = 1$ according to this ratio

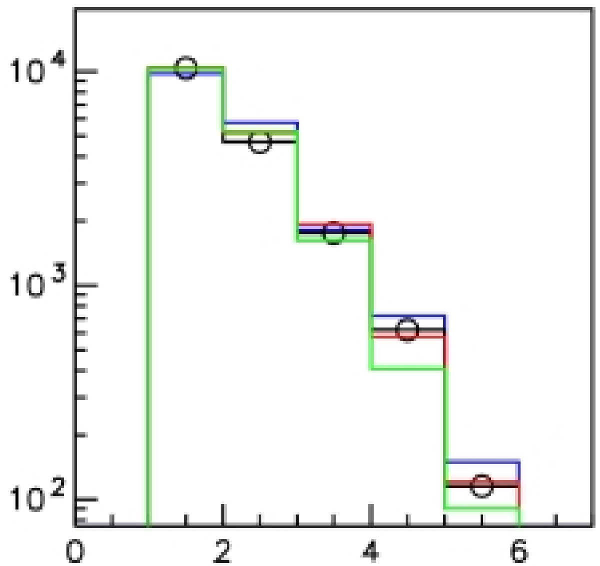
dN/dE, $N_{acc} = 1$



dN/d(cos q), $N_{acc} = 1$



Multiplicity (N_{acc})



Accidental cluster multiplicity reproduced fairly well after this correction

Event selection/EmC background: status

Currently have:

Stable event selection criteria

Statistical separation of accidental clusters/clusters correlated with $\gamma\gamma$

Need to implement:

Event selection module featuring removal of $\gamma\gamma$ clusters on output

Output of DC hits requires no additional work

Cluster superposition code in INSERT

CELE times and energies need to be adjusted for differences between v_{fib}, L_{att} in MC and data

DC background insertion

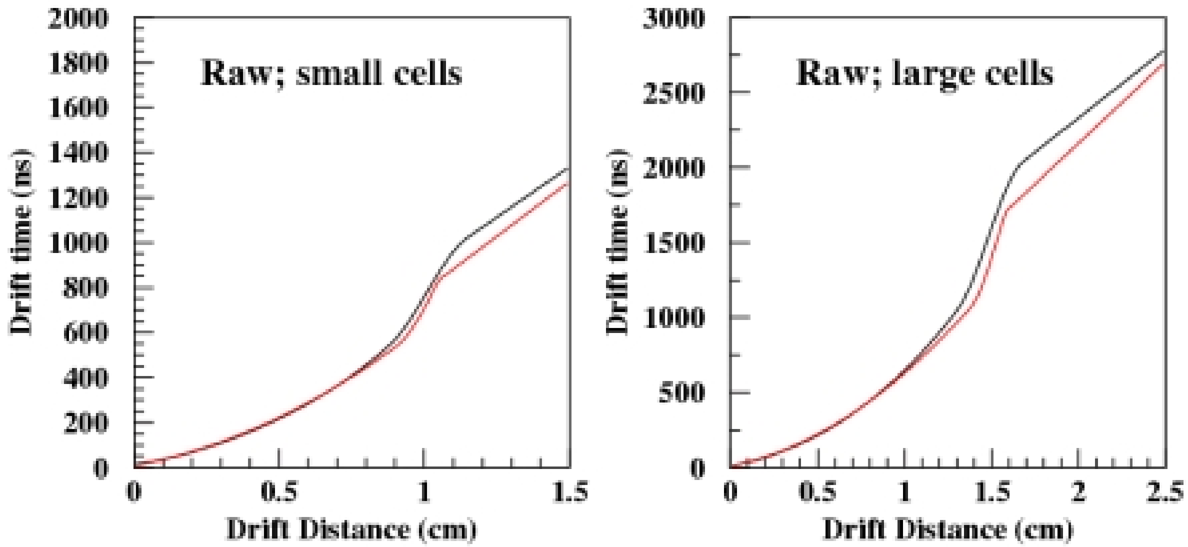
1. Read non t_0-subtracted, non hot-suppressed DTCE on secondary array
2. Read T0GL on secondary array
3. Perform t_0 subtraction, keeping hits with negative times
4. Intercalate MC and background hits:
   Keep hit with earlier time when two hits overlap
5. Suppress negative hits only at end
   Sign of drift distance unusable:
   SMEAR_T0 and DCONVR assume sign carries L/R info
   Negative times not allowed to fluctuate positive because of TSKT

Treatment of hot/dead channels needs refinement:
Suppress dead channels in INSERT: TSKT shouldn’t see them
Suppress hot channels in separate A/C module after TSKT:
TSKT sees everything and applies its own hot-channel vetoes

Raw s-t relations for MC and data

Raw *s*-*t* relations for MC/data differ by 100-200 μm at 1-2 cm from wire

Background hits not reconstructed with same radii when inserted

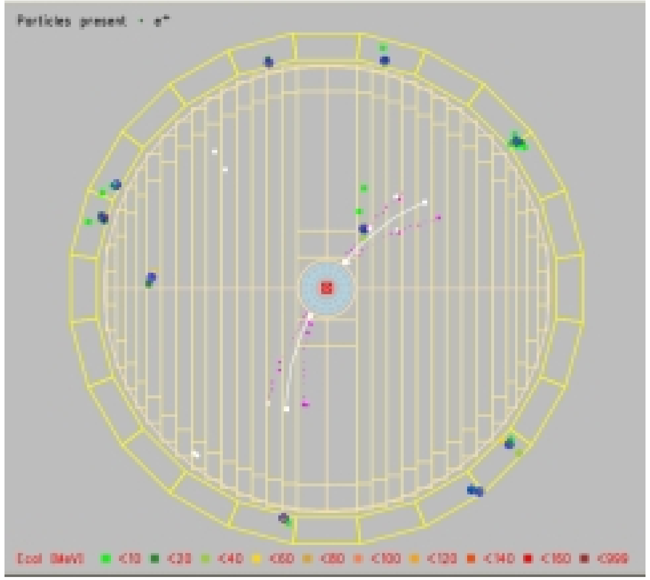


Implications for reconstruction efficiency/quality for inserted *tracks*?

Probable solution:

- Adjust "data" times to "MC" times in INSERT
- Problem: can be done for either raw or fine *s*-*t* relations, but not both
- Only substantial issue remaining to be addressed for DC insertion

Effect of different s-t relations



MC events:

- KS, KL neutrals
- All hit banks dropped

"Background" events:

- Events in **bha** stream
- At least 2 tracks with >20 hits each

Study reproducibility of track reconstruction when inserting tracks into MC events without hits by visual scan of 100 events (200 tracks):

Intact reconstruction	
Perfect	**164**
Split reproduced	**9**
Different reconstruction	
New split	**16**
Split recovered	**5**
Badly reconstructed	**1**
Lost	**1**

About 90% reproducible, with few % excess of split tracks and small losses

(allowing cancellation of new/recovered splits)

MC DST's: development principles

- MC DST's produced from **.mcr** output
- Reconstruction bankset same as for data DST's
- MC-truth bankset highly compressed
  - Most variables for PROD2NTU structures precalculated/stored
  - Most MC hit banks and related link banks dropped
- Code for creating new MC DST banks in MCT library
  - New KLOE offline library; also contains **insert** module
  - Banks defined with header files and descriptions in $K_IMCT
- Bank structures must accommodate presence of background hits
- Existing code in TLS (PROD2NTU) must work out of the box

MC DST's: banks currently present

Headers, etc.	**LRID HEAD EVCL BRIN RUNG**
MC truth	**KINE VERT**
t_0-related	**T0MC T0GL**
Trigger	**TDST CTRG**
EmC recon.	**CLPS CLLS CSPS**
EmC truth	**CFHI CEKA CEKE**
QCAL	**QCAE QWRK QCKA**
DC recon.	**DTFS DVFS**
DC truth	**MDKI MDTF MDCN**
TCA	**TCLO**
Event class.	**ECLS ECLO VNVO INVO KNVO**

MC DST banks for tracking

MKIN: MC details for KINE tracks (20 words per KINE)

- Number of DHIT hits and layer crossings; inner/outermost layer
- x, p at first/last DHIT hits
- Path length and TOF

MDTF: MC truth for DTFS tracks (28 words per DTFS)

- Indices of 3 main KINE contributors; number of hits contributed
- Index of KINE at first/last DTFS hit; layer, x, p for first/last hit
- Layer, x, p for first/last hit from majority KINE contributor

MDCN: MC DC hit count summary bank (10 words/event)

- Substitutes DCNH for MC DST's
- Counters for DHIT and DHRE hits, hits used by PR/TF
- Itemized by small/big cell; generated/background hits

MC DST banks for trigger and EmC

Trigger: Format of TDST bank same for MC/data DST's

- TORTA word:
  L1 type (EmC/DC/both), LET/cosmic multiplicities E/B/W, cosmic veto flag
- T1C, T1D, and T2D times
- Number of L2 DC hits
- Injection and fiducial clock signals not filled in MC DST's

EmC: MC information relatively compact; only discussion is over fate of **CHIT**

CEKE bank created as possible alternative:

- Composition and weight of KINE contributions to cluster elements
- DC banks give similar composition for tracks

CEKA	CELE
Number of KINE contributors	Total energy of KINE contributors
KINE contrib #1	Energy from KINE #1
KINE contrib #2	Energy from KINE #2
…	…
Number of KINE contributors	Total energy of KINE contributors

MC DST's: status and size estimate

Output size estimate:

1000 $K_S \rightarrow$ all, $K_L \rightarrow$ all events

Generated and reconstructed on AIX w/ standard path

.mcr	**23.9 MB** (*i.e.*, KB/evt)
.dst	**4.1 MB**

Very close to a final figure, to compare with 6 KB/evt pessimistically estimated last time

500 pb^{-1} = 500 M evts ≈ 2 TB

Variations:

Standard w/CHIT instead of CEKE	**4.7 MB** (KB/evt)
Standard + QIHI	**4.4 MB**
Standard + QIHI and CEKE $\rightarrow$ CHIT	**4.9 MB**

Proposal for production scheme

Production scheme must satisfy two important criteria:

1. Run-variable conditions must be correctly time-averaged over data set
2. MC output must be able to be divided up while maintaining relevance to a particular set or runs (or standard data set must be defined)

Both satisfied by generating MC files corresponding to actual runs

$\sqrt{s}$, $\mathbf{p}_{\phi}$, dead channels, etc. known by run number:
- Presumed to vary slowly with time
- Run-by-run generation handles averaging over data set:

Background levels highly variable *within* any given run
- Important to time-average background correctly when inserting

Background sampling

Physics events generated with definite cross section

$\phi \to K_S K_L$ events, $\sigma \approx 1050$ nb

Background also taken from events with definite cross section

Recognized $\gamma\gamma$ events, $\sigma \approx 30$ nb

Insert each background event into fixed number of MC events

If $\gamma\gamma$ selection efficiency not significantly dependent on background, obtain temporal profile for background that matches data

Need to simulate chunks of runs for which $\int L\, dt$ available

$\phi \to K_S K_L$ events: max MC file size 25000 evts ≈ **25 nb^{-1}**

Run size in 2001-2002 data **20 → 200 nb^{-1}**

Raw file size: **2-6 nb^{-1} in 2001, 6-15 nb^{-1} in 2002** (have $\int L\, dt$!)

Generate MC files corresponding to raw files?

Production scheme: examples

	$K_S K_L$	$\phi \to$ all	$K_S \to \pi^+\pi^-$ $K_L \to \pi^+\pi^-$
σ (nb)	1050	3100	1.4
Max evts/1GB file	25000	30000	40000
Max L per file (nb^{-1})	25	9.7	28000
Files/200 nb^{-1} run	8 full length	20 full length	0.008 (1 file ~320 evts)
If 15 raw files of ~13 nb^{-1} each…	OK 15 files ~55% full	Must split MC files 30 files ~50% full	15 files ~20 evts each Must group raws
Background reuse factor	35	100 → 50 (raw files used twice)	0.033

To split MC files across raw files: background from entire raw file used for each corresponding MC file; reuse factor adjusted accordingly

Requirements on DB2

Background can be treated as a datarec stream

New DB requirements for MC runs/files:

- Runs are generated for each raw file in data set
- Additional complications from grouping raw files/splitting MC files

New DB2 tables in logger schema for official MC production

- Link MC runs with background files used for reconstruction
- New tables only supplement existing tables
  - → Fully backward compatible

Note:

- MC run number will not correspond to simulated physical run number
- Correspondence will be available from database

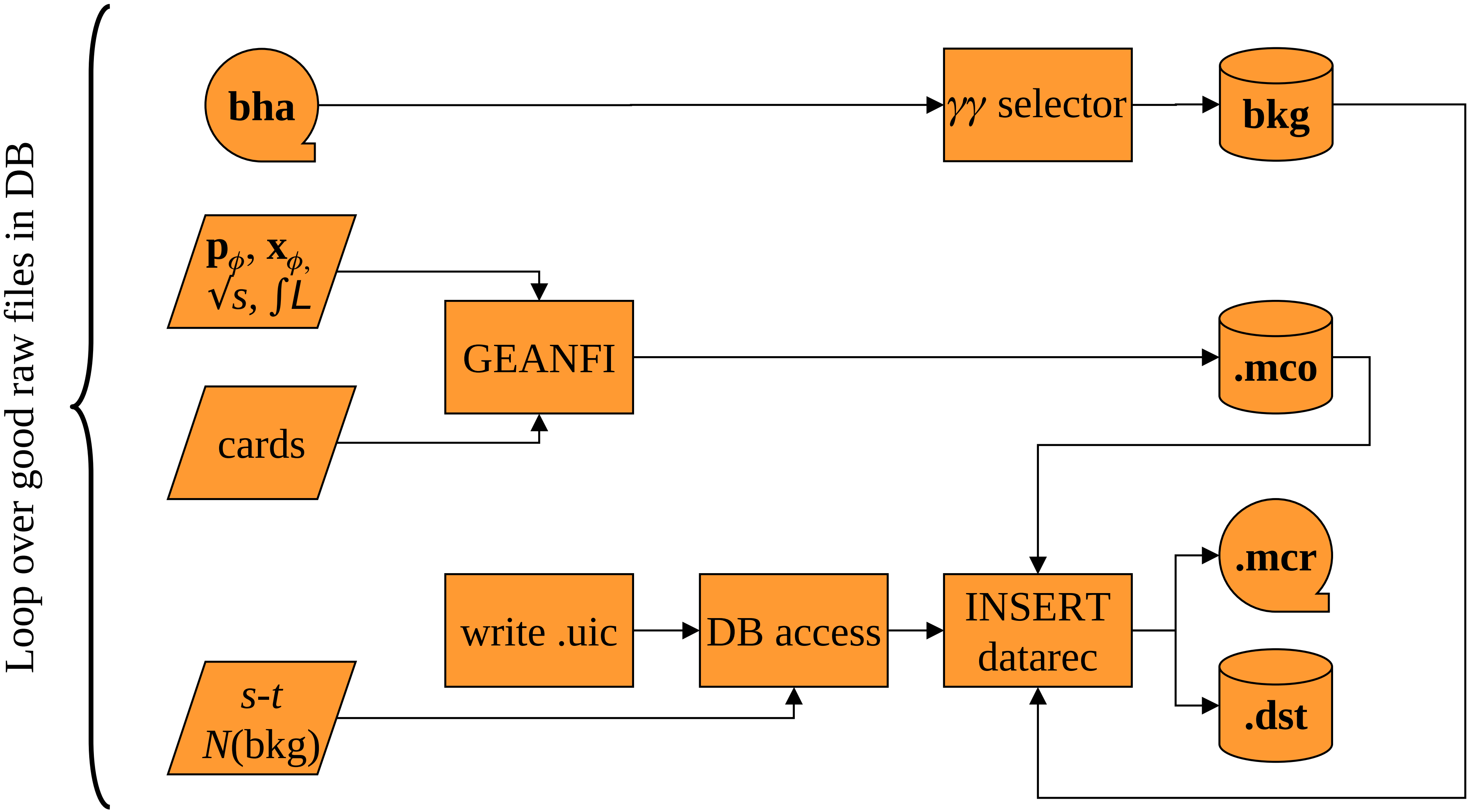
Database modifications

Example of new table to extend information in logger.mc_runs:

logger.mc_runs One entry per MC run	**logger.mc_runs_raws** One entry per MC run and background file	
MCCard_ID, MCRun_Nr	MCCard_ID, MCRun_Nr	Primary keys identifying MC run
	Bkg_Run_Nr, Bkg_Version, Bkg_Offline_ID, Bkg_Datarec_Nr, Bkg_Stream_ID, Bkg_GB_Nr	Primary keys of associated background, can be used to index: • **logger.datarec_logger** • **logger.datarec_raws** → **logger.raw_logger**

Create views *e.g.*, to allow MC files to be selected by physical run number

Complete production flowchart



Combine or separate neutral kaon runs?

Combined production $K_S \to$ all, $K_L \to$ all	Separate production $K_S \to \pi^0\pi^0$, $K_L \to$ all $K_S \to \pi^+\pi^-$, $K_L \to$ all Differentiated by K_L decay in DC?	Combined generation $K_S \to$ all, $K_L \to$ all Streaming to dst by MC truth
Simpler to produce	Simpler to analyze	Simple to produce and analyze (if no reprocessing)
Fewer files (if file length unsaturated)	Smaller files	Smaller files
Less disk turnover? (if people cooperate)	Less disk turnover? (if event subset dominates interest)	Less disk turnover? (if event subset dominates interest)
	Lighter disk access (if event subset dominates interest)	Lighter disk access (if event subset dominates interest)
No need to prioritize	Possible to prioritize	No need to prioritize
Naturally treats rare channels		Rare channels treated well in generation Problems with zero-length files
Well-suited for background studies (rare K_S decays, non-K_SK_L physics)		Acceptable compromise for background studies (mechanically more running, total volume and content of data set unchanged)