資 料 4－1

失われたデータと研究への影響

平成２５年２月１５日
第一回Belle実験データの一部損失に関する検証委員会

高エネルギー加速器研究機構素粒子原子核研究所
山内正則

この話で明確にしたいこと

- １．何が失われたのか
- ２．その結果として研究成果に与える影響
- ３．調査委員会に今事案の背景として指摘された事項について

失われたデータについて

調査委員会報告書の表１

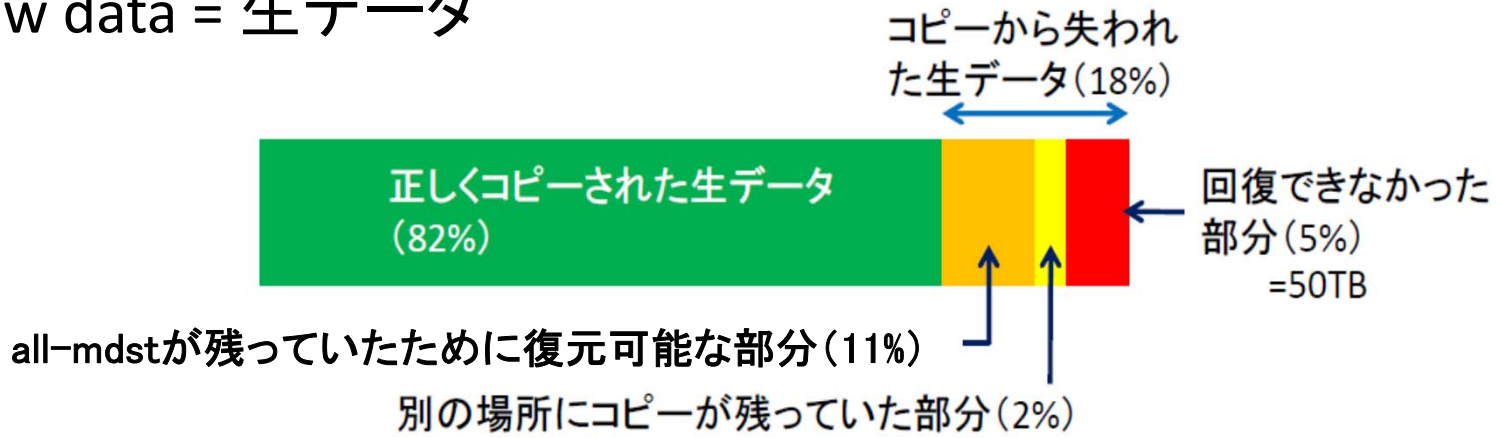
表 1: コピーされなかったファイルの容量と個数でみた比率。users と subdirs は subdirectories 内のファイルも全てコピーするように指定された。

raw data以下すべての合計

種類	コピーされなかったファイルの容量比	コピーされなかったファイルの個数比
全て	530 TB/1820 TB = 29%	203 万/330 万 = 62%
raw data	180 TB/1010 TB = 18%	7.7 万/24.7 万 = 31%
low multi.	8.11 TB/20.4 TB = 40%	0.92 万/2.30 万 = 40%
(dst のうち、low multi のみ)		
DST	98.8 TB/301 TB = 33%	26 万/62 万 = 42%
(all-mdst, skim-mdst, 上の low multi 以外の dst)		
MC generator	22.6 TB/22.6 TB = 100%	5.0 万/5.0 万 = 100%
background	0.85 TB/0.85 TB = 100%	1.2 万/1.2 万 = 100%
users	206 TB/448TB = 46%	151 万/220 万 = 68%
subdirs	11.4 TB/13.4 TB = 85%	11 万/15 万 = 76%

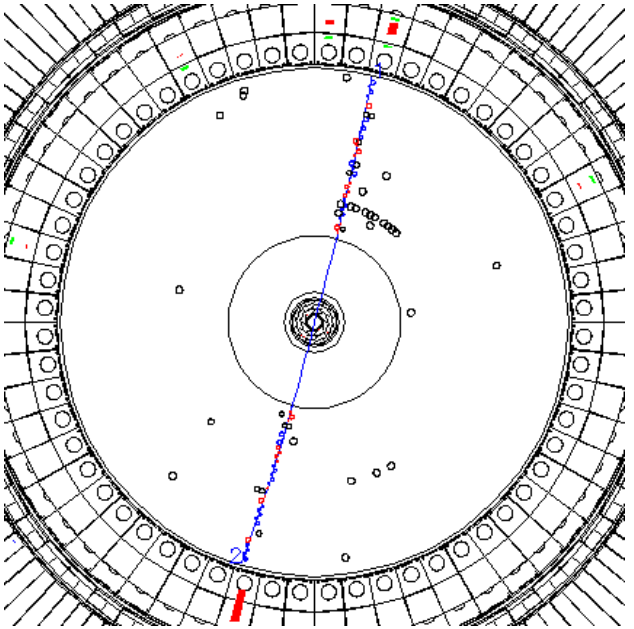
失われたデータについての説明（1）

1. Raw data = 生データ



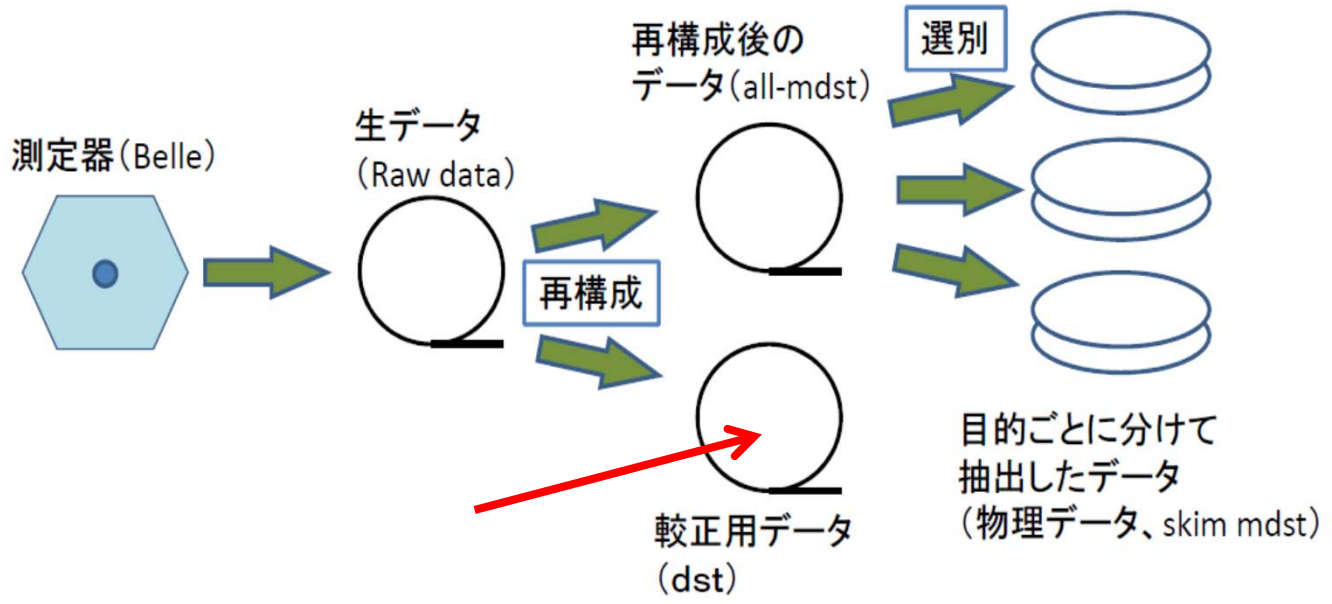
2. Low multi.

後述のDSTのうち、右図のような単純なトポロジーをもった事象だけを集めたもの。物理データではすでにDSTを使った較正が終わっていることから、このサンプルの40%が失われたことによる問題はない。



失われたデータについての説明(2)

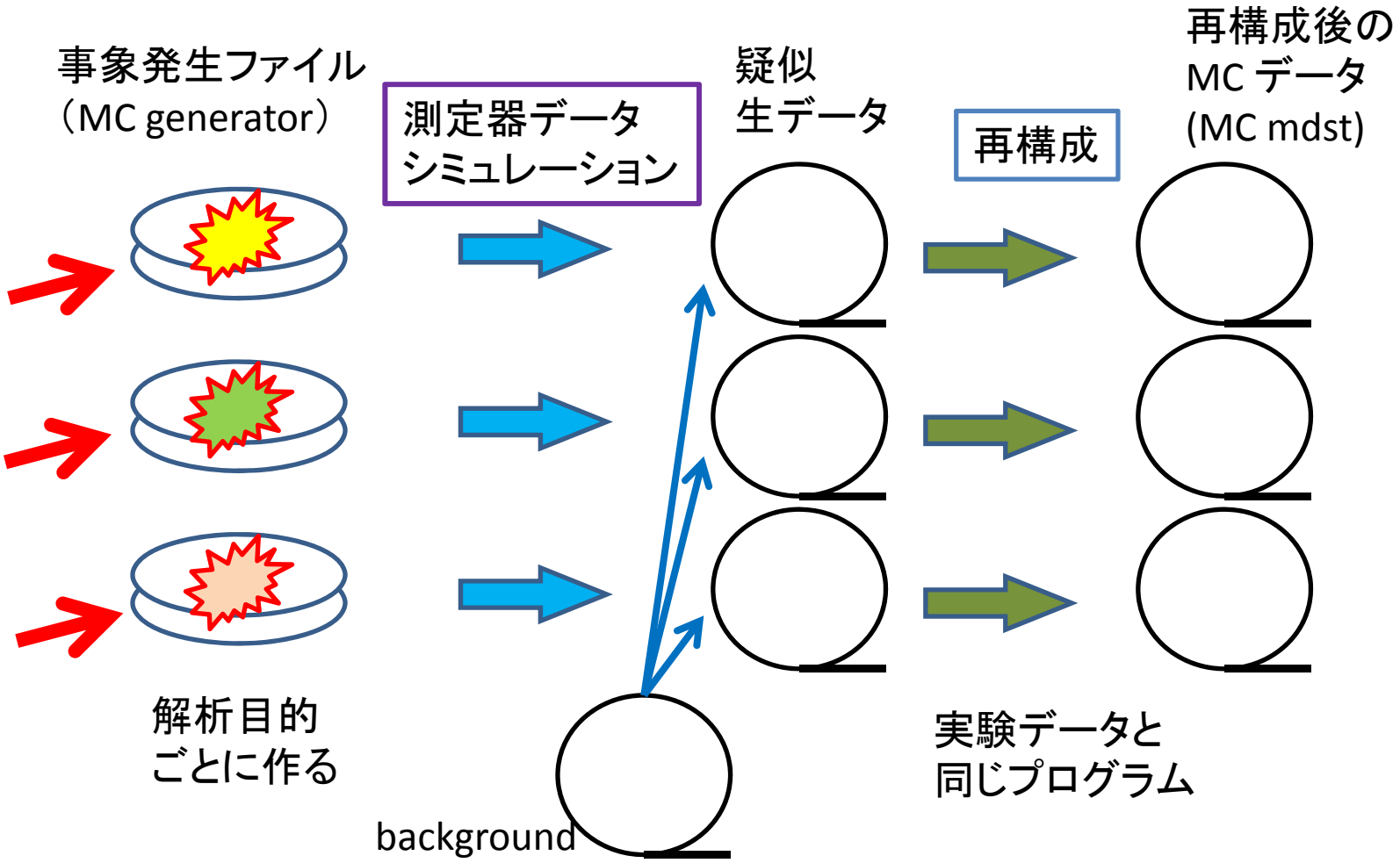
3. DST (データ・サマリー・テープ)＝較正用のデータ



すべての生データに再構成を行った中間データのうちから測定器の較正に使う事象だけを取り出したファイル。「較正用DST」を略してDSTと呼んでいる。物理データではすでにすべての較正が終わっていることから、このサンプルの33%が失われたことによる問題はない。

失われたデータについての説明(3)

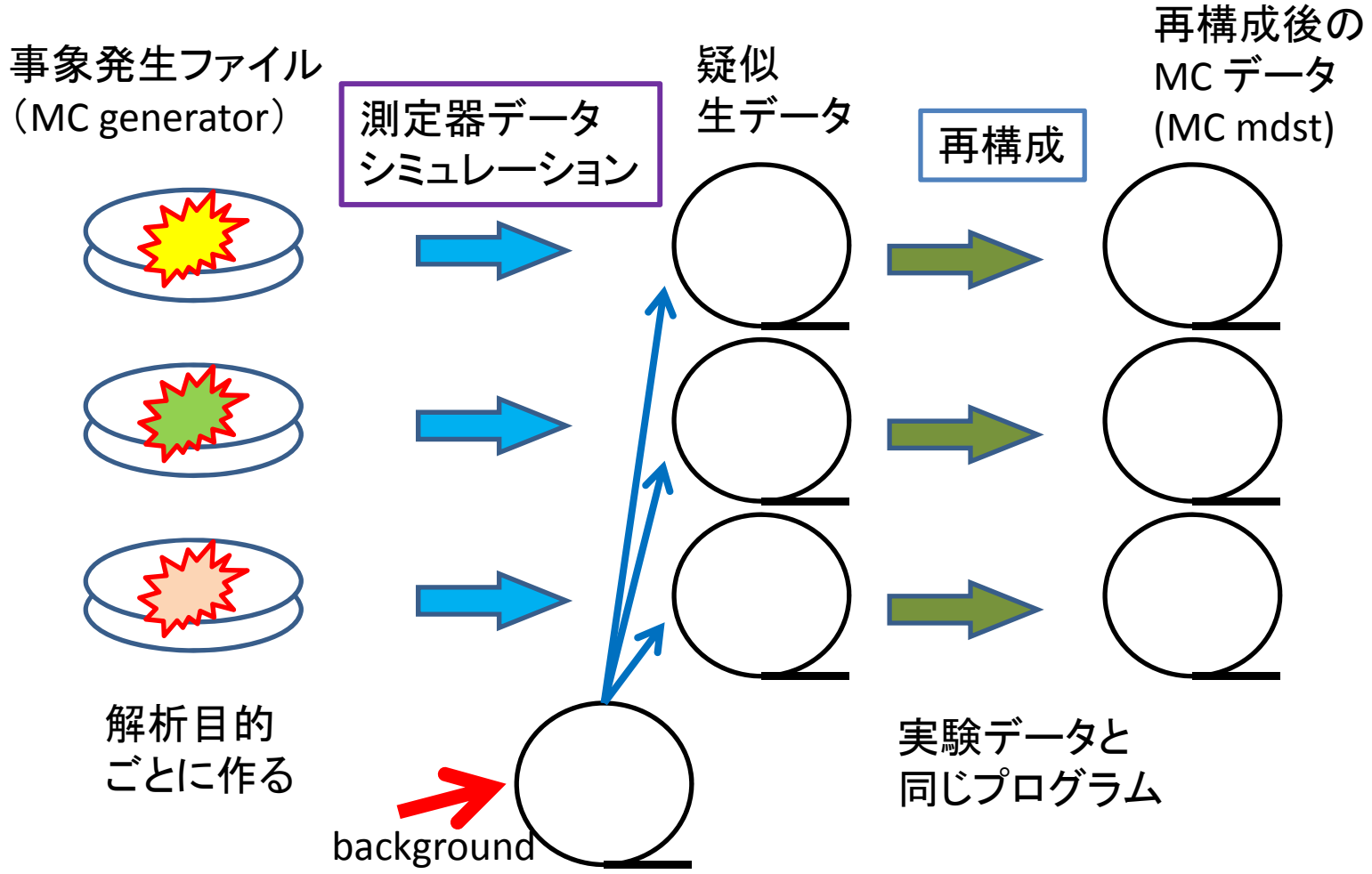
4. MC generator = モンテカルロ法により発生した疑似データのうち、測定器のシミュレーション以前のもの



短時間で作り直すことが可能で、失われたことによる問題はない。

失われたデータについての説明(4)

5. Background = ランダムにシャッターを切って取った事象



疑似生データに混ぜることによって加速器からのノイズなどを含んだ疑似データを作ることが目的。必要量を越えて収集しており、既存のデータから再度選別が可能であり、失われたことによる問題はない。

失われたデータについての説明(5)

6. Users
7. Subdirs

ともに参加研究者が自分のテーマの研究を行うために使っているファイルで以下のような内容がある。

- 特定の目的のために事象を選んで作ったデータファイル
- データ解析用プログラム
- ヒストグラムなど解析の中間データのファイル
- 参考文献をダウンロードしたもの
- プレゼンテーションファイル
- 自著論文などさまざまな個人で使用するデータ

これらのデータの大きな部分が失われたが、必要なファイルはすでに復旧、再生されている。研究者によってはこのために約2か月を費やす必要があった。

調査委員会報告書の表3の解釈について

別のテープシステム
から復旧後の損失量

データ移行の際に
失われたデータ量

表 3: データの実質的な損失のまとめ。括弧内は Belle 所有の別の HSM から raw data を
復旧する前。

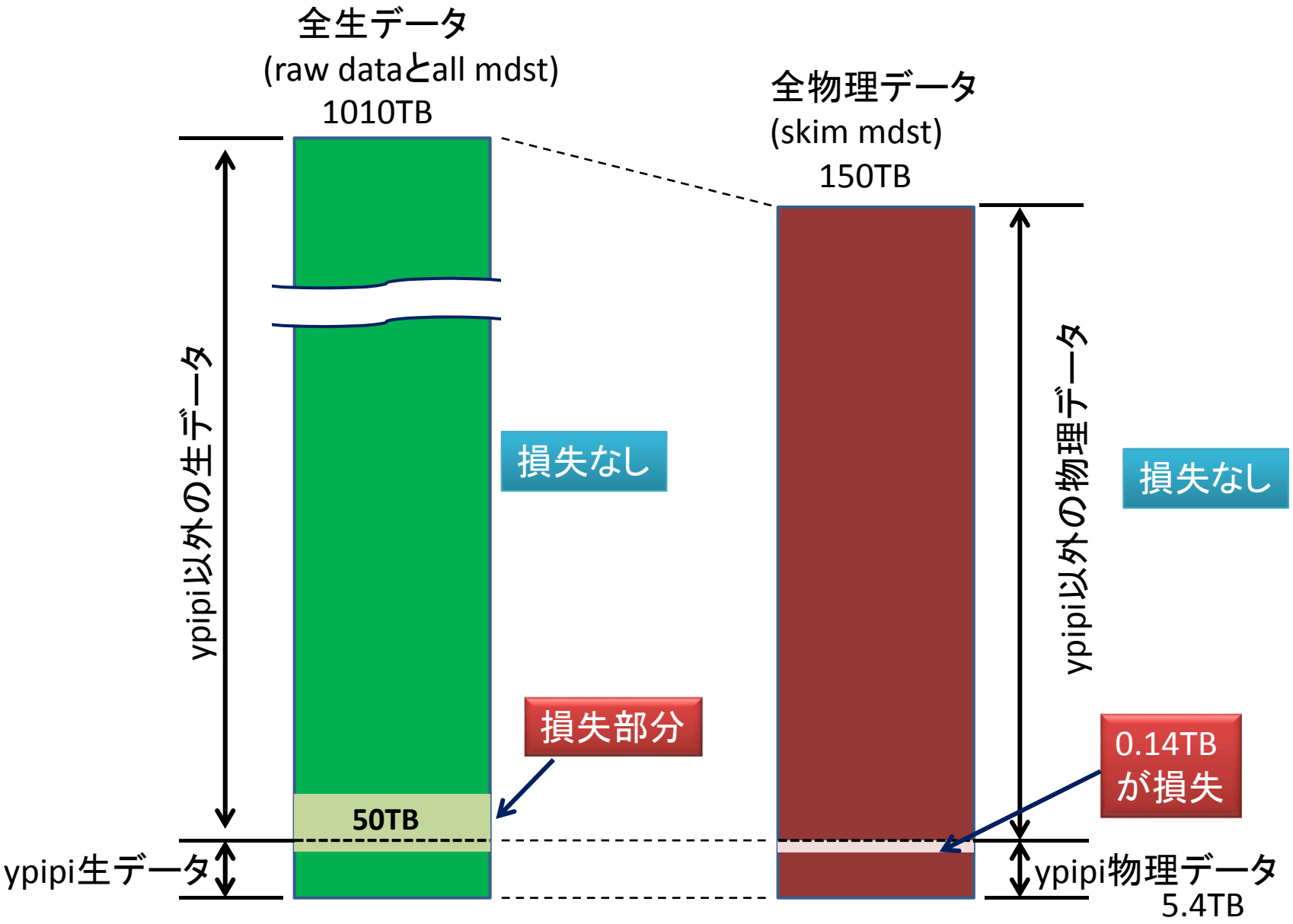
種類	損失量	
raw data	12%	(18%)
raw data と all-mdst 両方	5%	(7%)
skim mdst (ypipi 以外)	0%	
ypipi mdst (scan; 25 点中 6 点で)	30%	
ypipi mdst ($\Upsilon(5S)$ データ)	2.5%	

ypipi以外物理データ
に損失はない

all-mdstに同じ事象がある
ものは実質的に損失となら
ないことから、最終的な生
データの損失量

ypipiにはさらに2種類あって、
scanデータ（0.12TB）の6/25で30%、
すなわちscanデータ全体の7%と、
固定エネルギーでのデータ（5.3TB）
の2.5%が失われた。

回復不可能なデータの図示



ypipiの一部を失ったことによる影響

- ypipiについては2種類のデータが失われた。

加速器のエネルギーを少しずつ変えながら採ったデータ

新しいハドロンの中心にエネルギーを固定して採ったデータ

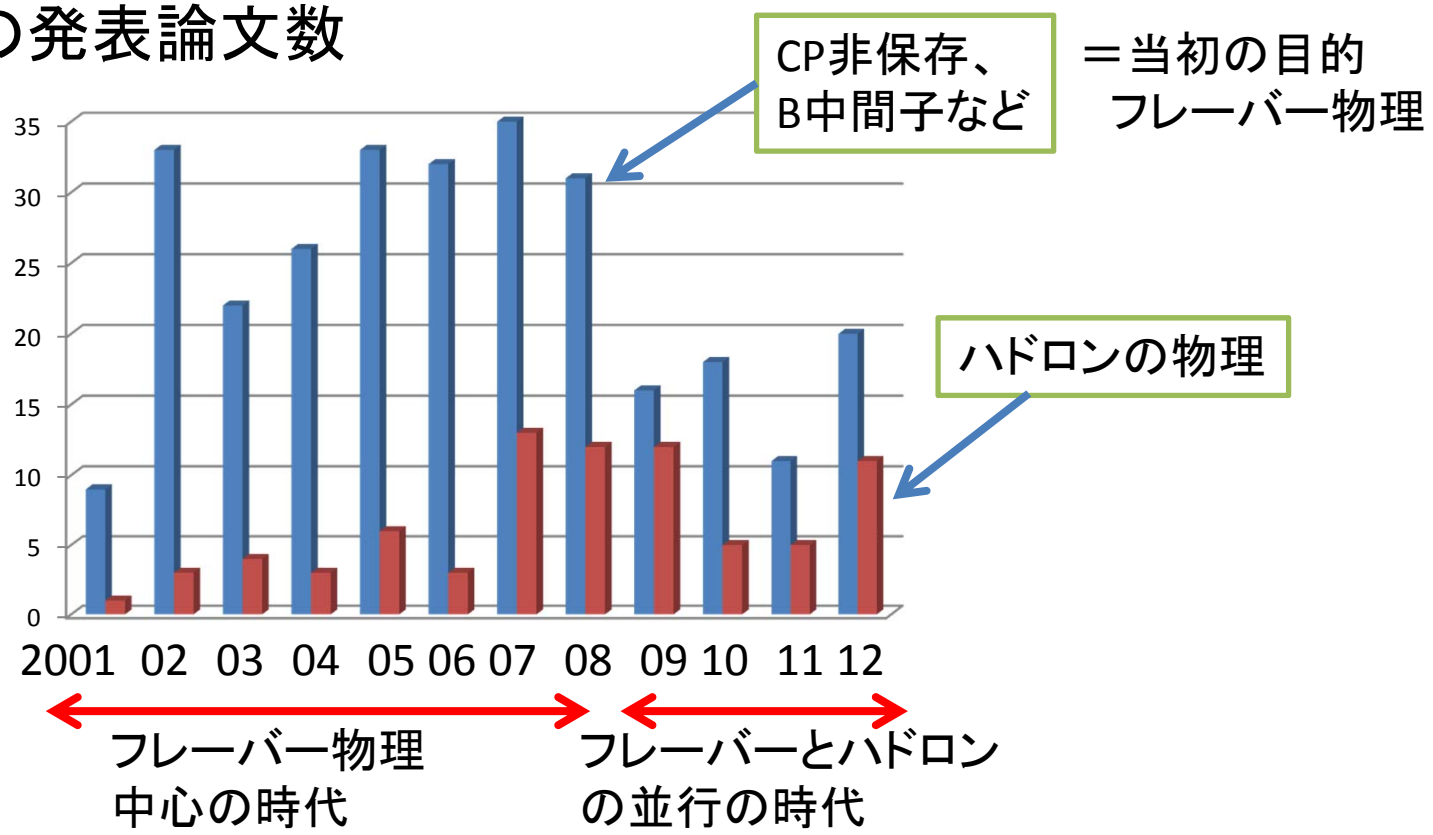
- この結果、スキャンデータの7%を失ったことによって、測定結果の統計誤差が4%ほど悪化することが予想される。

仮にYb粒子の何かのパラメータを測定したとして、100.0±10.0という結果が期待されるところが、この損失のため100.0±10.4という結果になる、というのがこの損害である。

なぜ4%か？	一般に統計誤差は統計量の平方根に反比例するので、$\sqrt{1+0.07}-1\sim0.04$.
4%の影響は	実際上は4%の統計誤差の悪化によって成果が左右されることはない。

ypipiデータを収集し研究対象とした経緯

- 各年の発表論文数



- フレーバー物理の大きな成果は2008年頃までに概ね達成されており、その先にはSuperKEKBが必要との認識が広がった。参加する学生に学位を取得させるためにハドロン物理の研究が盛んになった。（2007年頃）
- その一環としてエネルギースキャンを行いypipiデータを収集した。

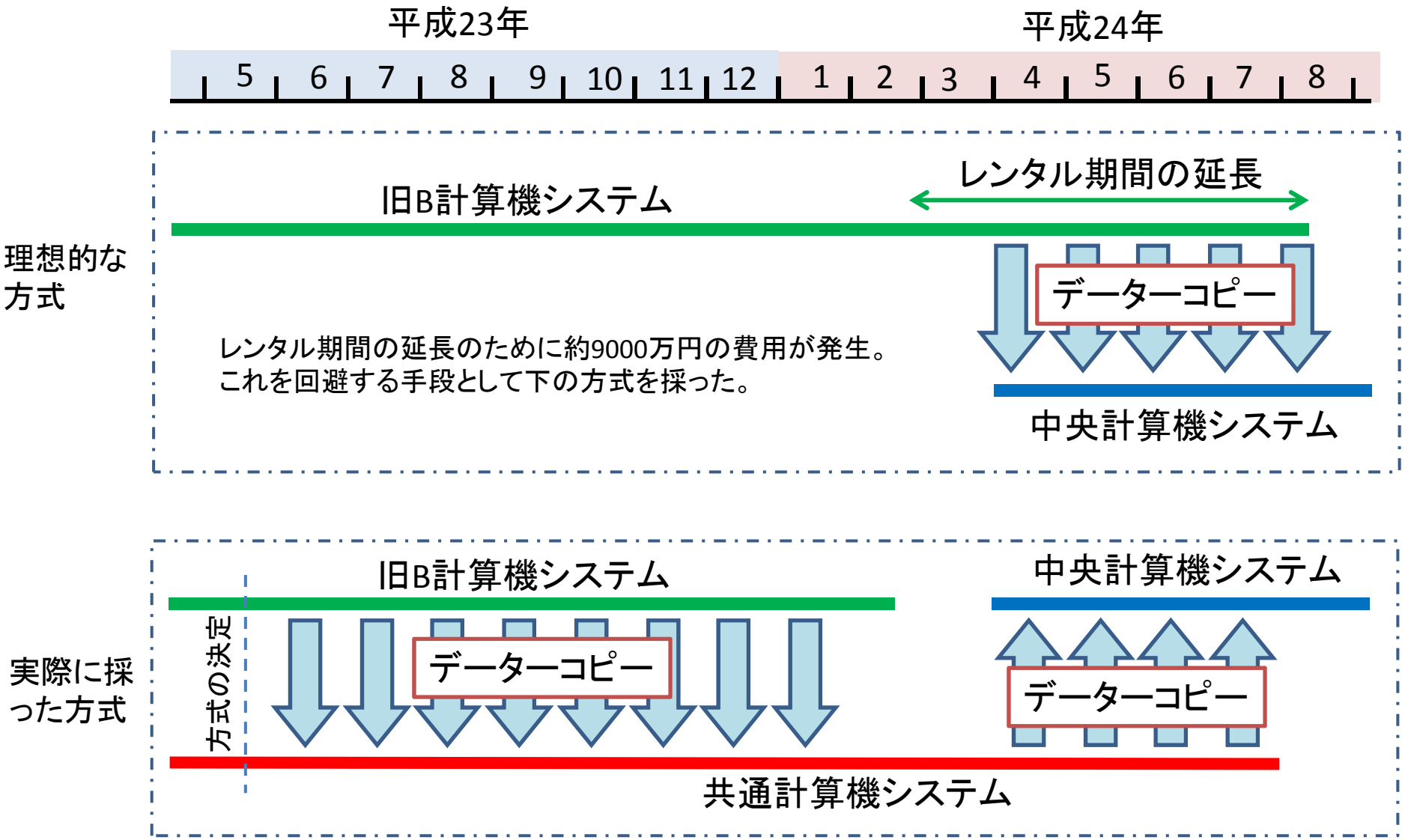
発表論文の内訳

	これまでに発表した論文数	今後発表予定の論文数
当初の研究計画に関する成果	299	~20
ハドロンの研究など副産物的成果	76	~20

今回のypipiデータ損失によって、このうち一編について統計精度の僅かな悪化が起こる。

データ損失が発生した背景①

手順の複雑さと作業期間が短かったこと



データ移行方式決定の経緯

- H23年2月頃　Belle計算機担当者から山内（当時素核研副所長）に計算機入れ替えに伴うデータ移行の方式案について説明があり、計算機のレンタル契約を延長するために数千万円規模の経費が必要となることが説明された。

- Belle側担当者とレンタル会社と協議の結果、レンタル期間の延長は一年単位となるために、その費用は一億円に及ぶことが山内に報告された。

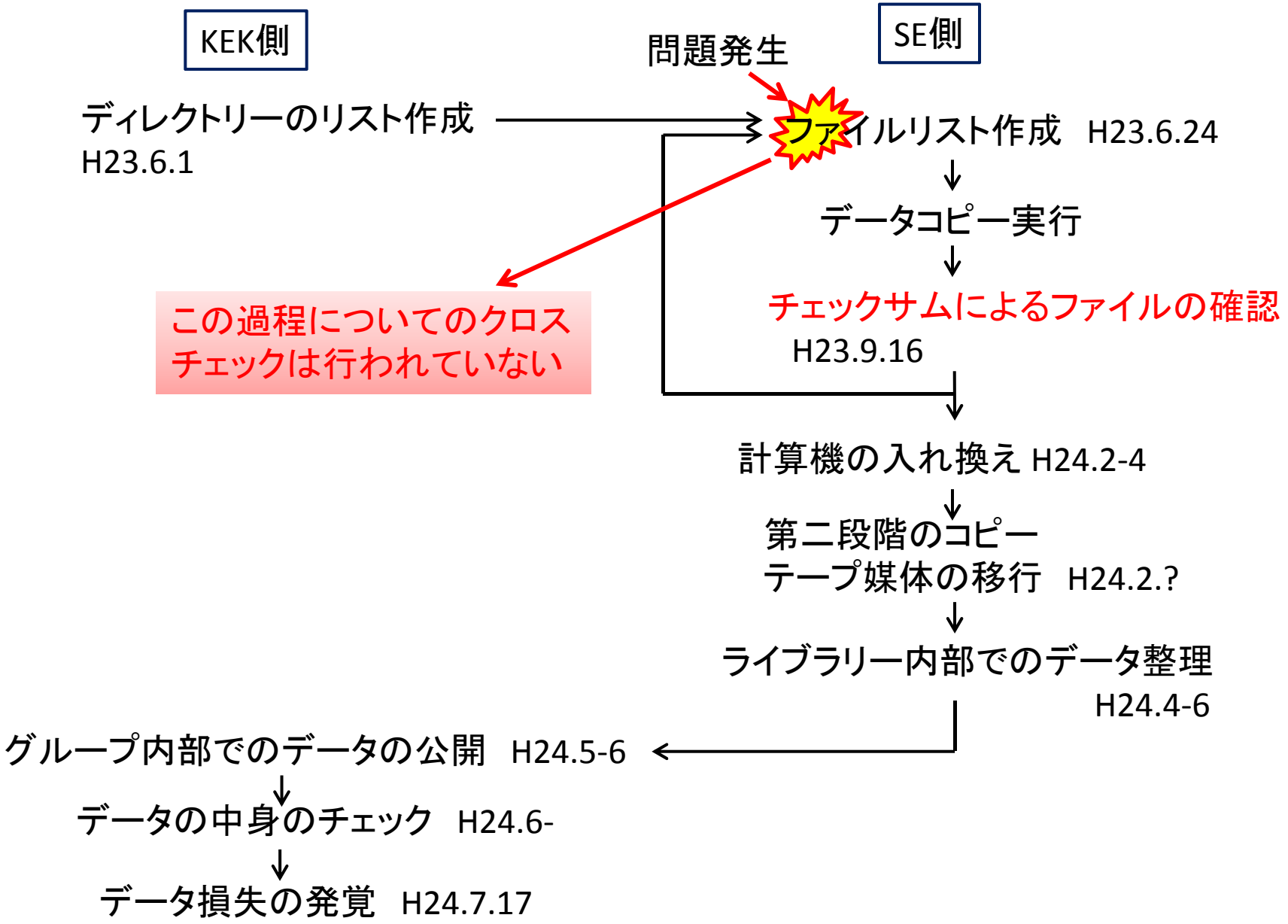
- 山内が西川素核研所長（当時）と相談した結果、Belle予算に余裕がなく、他からも回す余裕がないことから一億円の出費を回避するようデータ移行方式を再検討すべしという結論になり、その旨Belleに伝えた。

- 旧共通計算機のレンタル契約を延長してこれを経由すれば2000万円程度でできる方法がある旨、当時の計算科学センター長から提案を受け、山内は了承した。

- H23年5月　Belleグループ、計算科学センター、関連会社が協議し、データ移行方式を決定した。

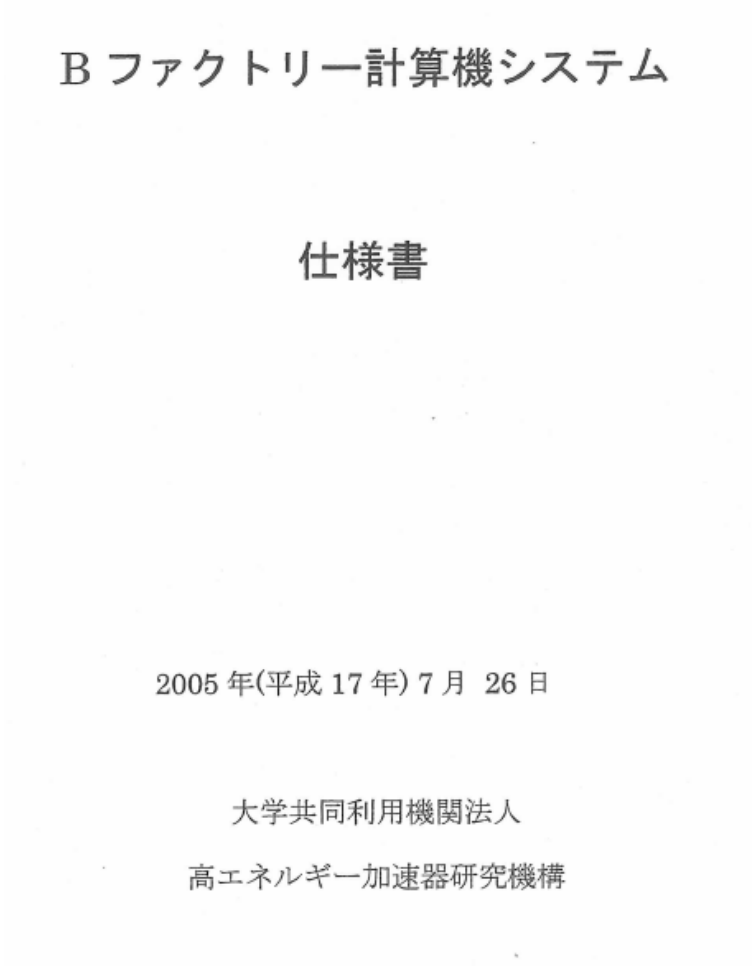
データ損失が発生した背景③

作業のクロスチェックが不十分であったこと



データ損失が発生した背景②

人員配置が不十分で責任分担が不明確であったこと



B ファクトリー計算機システム

仕様書

2005 年(平成 17 年) 7 月 26 日

大学共同利用機関法人

高エネルギー加速器研究機構

2.9 移行に関する要件

現有システムから本システムへの移行に際して協力を行うこと。現有システムのユーザーアカウント及びユーザーホームディレクトリー、HSMシステムのファイル移行に協力すること。なお、現有システムが撤去されてから、本システムが稼働するまでの期間は約2週間を予定している。又、本システムから次期導入のシステムへの移行に協力すること。大容量ストレージシステムのテープライブラリ装置に保存されるデータの後継システムへの移行に協力すること。

実際にはKEK職員2名とN社、I社のシステムエンジニア2-3名で作業が行われた。2名のKEK職員は他に多くの仕事を抱えつつこの作業に従事した。

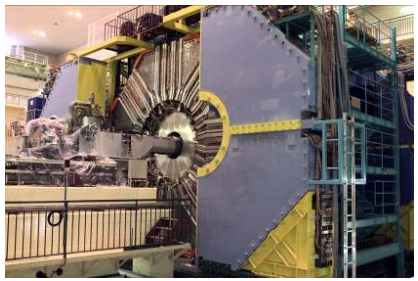
「本機構職員の監督のもとで」という一文を入れた上でしかるべく監督する体制を取るということがされていなかった。

旧B計算機システムのレンタル契約の際の仕様

関連するKEKの組織

注：今回の事案に直接関係のない部局等は大幅に省略してあります。

機構長、機構執行部



予算配分、人員配置を通して研究グループを統括する。

素核研

加速器施設

共通施設

.....

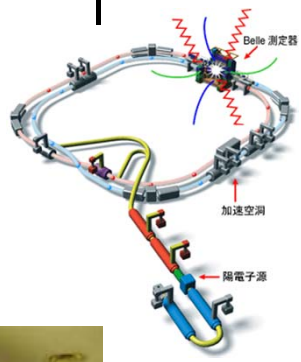
Belle グループ

22名

.....

研究グループB

研究グループA



計算科学センタ

.....

国内外の研究者
20カ国、400名

Belleコラボレーション

計算機にかかわる専門家としてシステムの運営、会社との仲介等を行う。

運営への協力

運営

ユーザー

計算機システム

まとめ

- Belle実験は平成22年夏までに当初予定された研究プログラムを終了し、375編の論文として発表した

- 失ったデータ
  - Belle実験が11年間に収集した生データ1010TBの5%
  - ypipiと呼ばれる特殊な研究のための物理データのうち2.6%
  - この他にも参加研究者が作成したデータなどが失われたが再生可能

- 研究への影響
  - 生データの損失による影響はない
  - ypipiの測定精度がやや低下（統計誤差の悪化）するが、成果を左右する程度のものではない

- データ損失が発生した背景：調査委員会による指摘事項
  - データコピーの方式の複雑さ
  - 十分な人員と作業期間を確保できなかったこと
  - 責任分担が明確でなかったこと
  - クロスチェックが不十分であったこと

補足説明

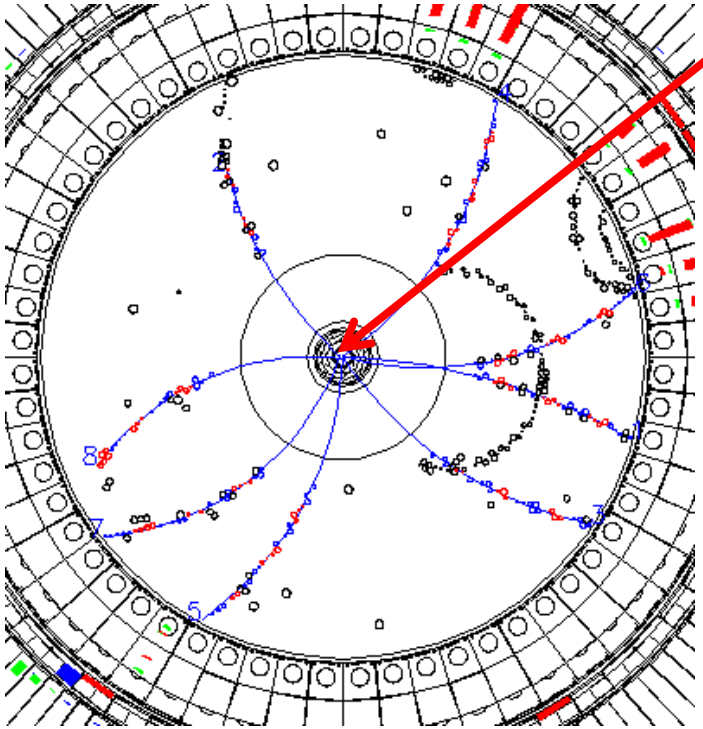
生データの損失によってこの研究の成果が損なわれないことの説明（1）

- Bファクトリー実験が当初目的とした成果はすでに達成されており、375編の論文として発表されている。文科省の審議会などで次期計画への移行も認められている。
  - B中間子におけるCP非対称性の発見
  - 小林・益川理論の証明
  - 新しい物理法則の探求
  - 　等々

- 大部分の物理データ（ypipiの損失部分を除く）は保存されており、上記に関連した別の角度からの研究が必要になった場合でも問題なく対応可能である。

生データの損失によってこの研究の成果が損なわれないことの説明(2)

- 生データに想定していない新しい現象が潜んでいるとすれば、生データの5%の損失は成果の損失になるのではないか。



反応が起こるのは衝突点のごくごく近傍のみ

記録されている飛跡は反応ではなく、反応によって発生した安定粒子が飛散する様子。このような事象は可能な限りバイアスなしに収集、物理データとして保存されている。

想定外の新現象があっても測定されるのは同じように飛散する安定粒子群なので、見かけは区別がつかないであろう。
→ 物理データとして抽出されているはず。

生データの損失によってこの研究の成果が損なわれないことの説明(3)

- 物理データに何らかの問題が見つかり、生データに戻って解析をやり直す必要があるかもしれない。

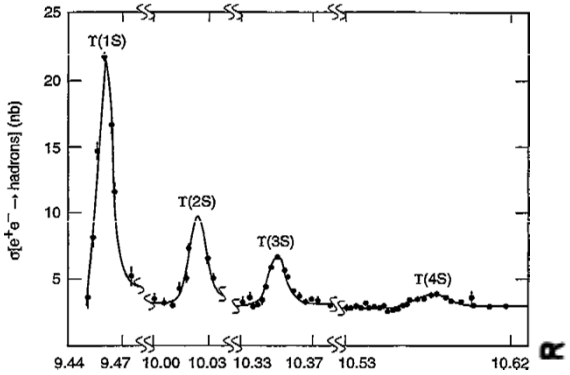
➢Belle実験が発表した結果の多くは競争相手であるBaBar実験(アメリカ、SLAC)によってクロスチェックされており、物理データに問題があるとは非常に考えにくい。

➢データ解析が正しく行われていることはモンテカルロ法による疑似データを用いて非常に慎重に確認されている。

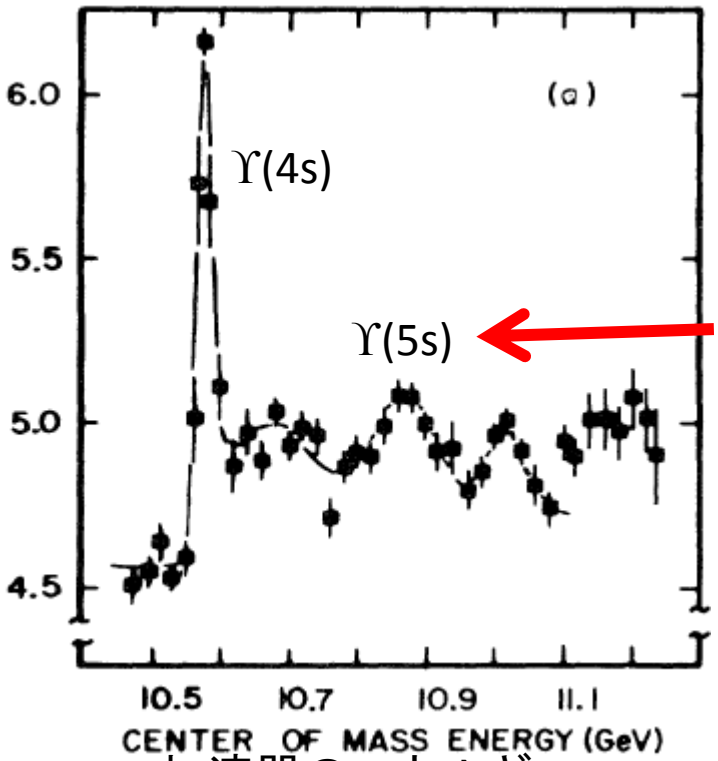
➢実験開始直後には生データに戻ることはしばしば必要となるのが通例だが、実験開始11年を経て生データに戻らなければならない事態は非常に考えにくい。実際、今回の損失がなくても生データからall-mdstを作り直す必要も予定もなかった。

Ypipi (= $\Upsilon\pi\pi$) の測定

加速器のエネルギーを変えながら衝突反応の起こりやすさを測ると凸凹が見える



縦軸は衝突反応がおこる確率



加速器のエネルギー

この粒子が単独の粒子ではないことを発見

重なっているもう一つの粒子が$\Upsilon\pi\pi$に壊れる様子を調べるためにエネルギーを少しずつ変えながらデータを採った。

→Ypipiデータ

Belle実験では当初狙った成果に加えて副産物もいくつか得られており、この粒子の“重なり”もその一つ。

詳しく調べるために特別に収集したデータの2.6%を失った。