Appendix 4 Summary of the investigation committee

1. Members of investigation committee

The investigation committee consists of the following four members:
Taku Yamanaka (Osaka University, Professor, Chair)
Hiroshi Sakamoto (University of Tokyo, ICEPP, Professor)
Yoshihide Sakai (KEK, IPNS, Professor, Belle co-spokesperson)
Takashi Sasaki (KEK Computer Research Center, Professor)

2. Summary of the transferred data

The data accumulated in the period between June 1999 and June 2010 by the Belle experiment were stored in the computer system (B computer). In addition, simulation data, produced by the Monte Carlo simulation method, were stored in the same computer system, together with miscellaneous data created by individual users. The data storage of the B computer was based on an HSM system[i] consisting mainly of a 3000 TB storage on magnetic tapes.

3. Summary of the data transfer work

As the lease contract of the B computer was to expire in February 2012, it became necessary to transfer the data from the B computer to the next computer system (new central computer), which had a new storage system. The data transfer was performed in two steps: (1) the data were copied temporarily to another computer, the old common computer between June 2011 and January 2012; (2) the lease contract of the old common computer was extended by several months so that it remained available in a manner overlapping with the new central computer system; and (3) during this period, the data were copied from the old common computer to the new central computer.

4. Data loss and its causes

During the data transfer procedure, some of the data files were not copied from the B computer to the old common computer and a small part of the missing data was found to be non-recoverable. Specifically, 29% of all the data stored in the HSM system was not copied. The data that were not copied included 18% of the raw data taken from the experiment. While many of the data were recovered from backup files stored elsewhere, 5% of the raw data were permanently lost. As for the physics data, duplicate copies existed at three places inside and outside Japan (KEK, Nagoya University, and PNNL in U.S.A.). Thus, no physics data were lost for the analysis of physics phenomena, pertaining to the original research objectives. However, as a sole exception, a small fraction of the special physics data that were used to study hadron resonances was not copied to the data storage outside KEK and thus lost. Parts of the simulated data and other data were also lost, but they could be successfully reproduced.

The direct cause of the data loss was due to errors in assembling the lists of files to copy. Had the work been conducted with a sufficient number of cross-checks, the data loss

could have been avoided. However, due to the lack of sufficient preventive measures taken by the management in the Belle group and the Computer Research Center, the transfer project resulted in data loss. A direct technical cause was in the computer scripts, which were not sufficiently tested for use in assembling file lists. An insufficient amount of human resources allocated for this work for a project of this scale, and the lack of cross-checking in the process of making the lists of files to copy are noted as the other elements that contributed to the occurrence of this incident. In addition, it was speculated that the relatively short period (7 months) assigned to the data transfer process placed further pressure on the limited human resources. The committee recommended that, for important project work in the future, sufficient cross-checking measures should be taken, involving an adequately large number of personnel. Adequate human resources and time should be allocated for successful project execution, and the importance of data preservation should be more seriously recognized.

[i] HSM stands for "Hierarchical Storage Management." In the HSM system, the data that are frequently accessed are stored in the primary media (magnetic disks) and those that are less frequently accessed are stored in the secondary media (tapes). All the files are virtually seen as a single large disk system and file management is done automatically.