KEKCC のストレージ

2009年4月24日

KEK 共通基盤研究施設 計算科学センター
八代 茂夫

内容

- 新共通計算機システム(データ解析システム KEKCC)の概要
- 仕様策定時にストレージ関係で検討したこと

KEK共通計算機システム

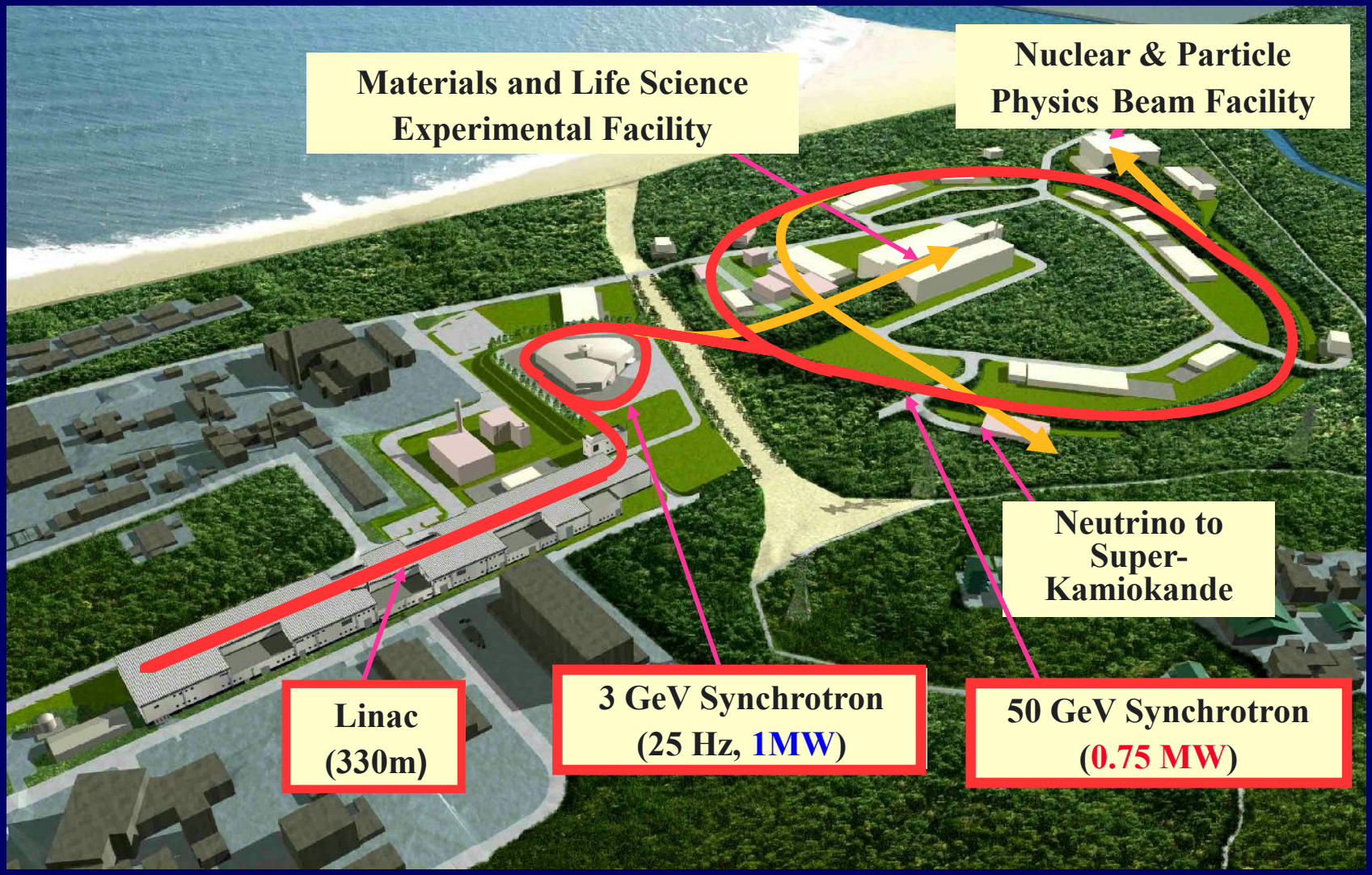
- データ解析システム (KEKCC)
  - KEKのプロジェクト(jparcを含む)のデータの保管および解析ためのシステム
  - ストレージシステム、計算サーバ、並列サーバ、GRIDサーバで構成
- 2009/3に新システム稼動開始

KEKCCのユーザグループ

- 従来からのグループ
  - ILC実験, ATLAS実験, PS実験, 理論, 加速器
  - 放射線遮蔽, PF, BESS実験
- JPARCグループ
  - HADRON実験
  - T2K実験
  - MLF実験

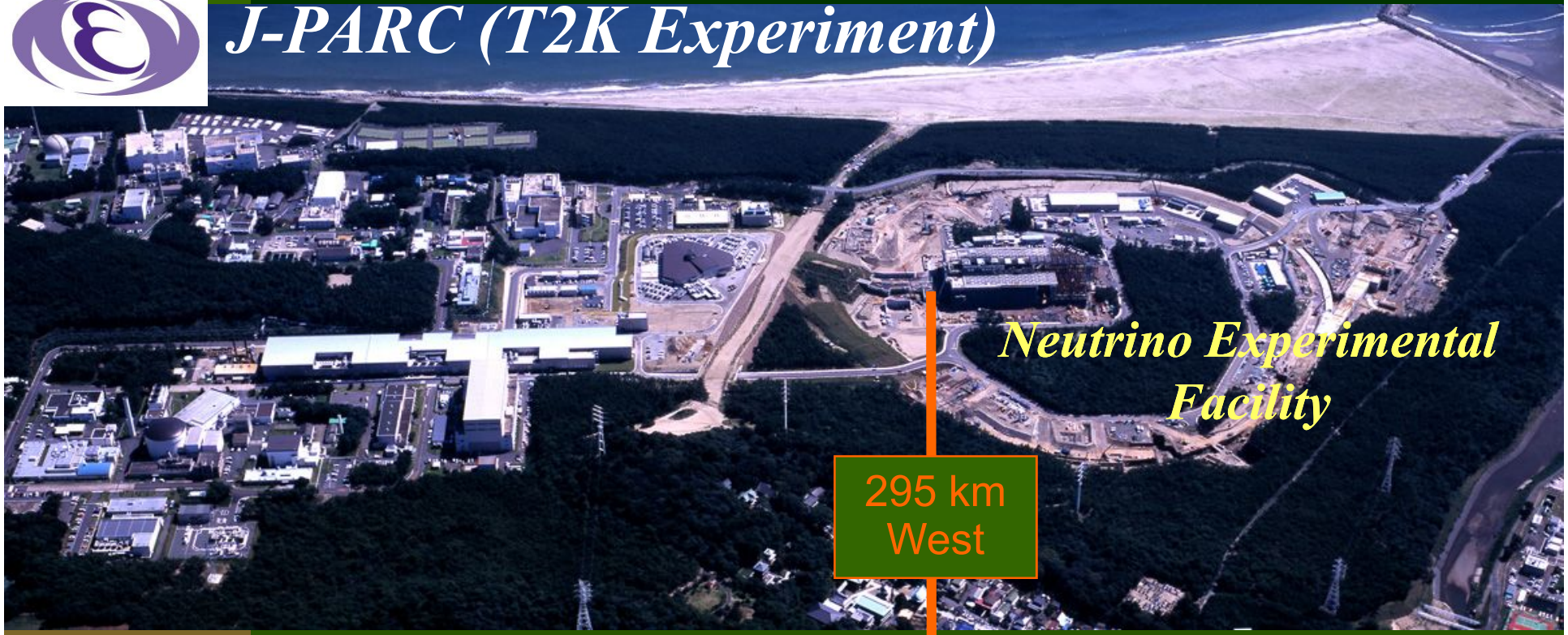
J - PARC (Tokai)

J-PARC = Japan Proton Accelerator Research Complex



Joint Project between KEK and JAEA

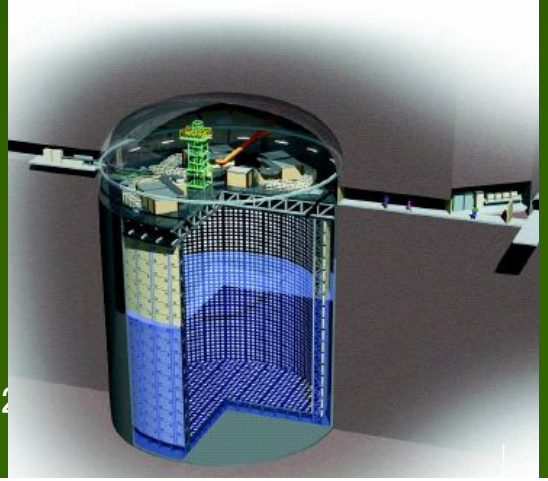
J-PARC (T2K Experiment)

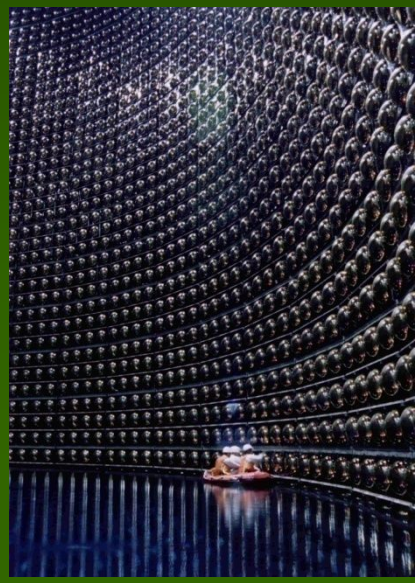


Neutrino Experimental Facility

295 km West

Super-Kamiokande





Central Computing System (Mar. 2009–)

Computing Server
84 IBM System x3550
Intel Xeon-QX 5460 x2
MEM:16GB

GPFS/API/VFS

API/VFS *GPFS* *GPFS* *GPFS*
API/VFS

GRID Systems
LCG System
Naregi System
iRODS/SRB System

HPSS
IBM TS3500(3PB)
IBM 3592 Tape Drive
IBM DS4800(10TB)

Disk Storage
IBM DS4800(205TB)
LTO4 Tape Drive

Work Server
4 x3550
Xeon-QX 5460 x2
MEM:16GB/node

Parallel Server
4 x3550
Xeon-QX 5460 x2
MEM:32GB/node

API/FTP
CIFS/NFSv4

CIFS/NFSv4 *SSH* *SSH*

Tokai campus
J-PARC

FW

FW

KEK intra

User's host

KEK-FW

SINET3

KEKCCのサーバ(1)

- 計算サーバ、並列サーバ
  - IBM System x3550　88台
  - Intel Xeon-QX 5460 x2, 8コア/ノード
    - 処理能力 2112 SPECint2006 (旧システムの約3倍)
  - LSF
- ストレージシステム
  - 磁気ディスクシステム　205TB
    - 旧システムの約4倍
  - 大容量ストレージシステム　3PetaBytes
    - 旧システムの約10倍
    - HPSSを継続使用

KEKCCのサーバ(2)

- Gridシステム
  - LCG, Naregi
    - LCGはCERNで開発されているミドルウエア
    - LSFで計算サーバ利用
    - HPSSにアクセス
  - iRODS(i Rule Oriented Data Systems)
    - SDSC開発のデータ管理システム

HPSS

- 階層型ストレージシステム
  - アメリカエネルギー省研究所とIBM Houstonとで開発され、IBM Houston がサポートを行なっている
  - データの保管先として磁気テープを利用
- 機器構成
  - Tape library: TS3500 3PetaBytes
  - Tape drives: 10 3592E06, 1TB/vol, 160MB/s
  - Cache disk : 10TB

HPSSの概略図

Tape library: TS3500
3000vols,
max. 3PetaBytes

Tape drives: 10 3592E06
1TB/vol, 160MB/s

Cache disk
10TB

5 movers

Core server

2 VFS
servers

2 SMB/NFS
servers

Work servers
(VFS server)

API, pftp

kftp, ftp

CIFS

SCP

HPSSの利点

- 運用
  - 全データを1つのネーム空間で提供
  - 高速のファイル転送機能
  - 運用に影響を与えないrepack機能
- 省電力
  - UPS、空調

ストレージ機器への要求

- 信頼性
  - 機器の故障
  - 長期保存
- 性能
- 価格

磁気テープ装置

- ハイエンド
  - IBM 359x/TSファミリ
  - Sun StorageTek Tシリーズ
- ミッドレンジ
  - LTO, SONY AIT
- エントリ
  - DAT
- 性能の差は？ 信頼性の差は？

磁気テープ装置の比較

- 性能
  - 転送性能
    - IBM3592: 160MB/s, T10000: 120MB/s, LTO4: 120MB/s
  - ロード+サーチ
    - IBM3592: 49sec, T10000: 46sec, LTO4: 46sec?
- エラーレイト
  - T10000: 10^{-19} , LTO4: 10^{-17}
  - IBMは非公開
- リポジショニング現象を低減する技術
  - T10000、IBM3592
  - LTO4はなし
- 媒体を保護する技術
  - ヘッドの技術(IBM3592)、走行速度の抑制(T10000)
  - ヘリカルスキャン方式はヘッドおよび媒体に過酷

磁気テープライブラリのアクセス時間

- アクセスの最長時間
  - ソフトウェアの time out との関係に要注意
  - StorageTek SL8500
    - ロボット時間＋エレベータ時間＋ロボット時間
  - IBM TS3500
    - ロボット時間

SATAディスクvs. FC/SCSIディスク

- 性能の差
  - 記録密度、円盤数、ヘッド数 ➔ ヘッドの動作
  - 回転数
- 大まかな応答時間
  - 平均 seek時間 +平均回転待ち時間
  - 15Krpm FC　3.6ms+2.0ms
  - 10Krpm FC　4.7ms+3.0ms
  - 7200rpm SATA　8.0ms+4.2ms
    http://enterprisezine.jp/article/detail/157?p=3
  - ランダムな処理で差が出る。

SATAディスクvs. FC/SCSIディスク

- 耐久性
  - FC/SCSI系とATA系の稼動想定時間
    - SCSI系HDD = 24時間×365日
    - ATA系HDD = 8時間×300日

KEKCCの磁気ディスク

- 初めてSATA の RAIDを導入
  - ホーム領域はFCのRAID
  - データ領域はSATA の RAID
- SATA
  - RAIDで性能の問題を吸収できるか?
  - RAIDで耐久性の問題を吸収できるか?

ストレージに関する課題

- FWを越えてのデータ共有
  - 特定の少数サイトとのデータ共有
    - JPARCからのデータ転送
    - 共同利用の機関とのデータ共有
  - 特定多数のサイトへのデータ提供
    - 小規模な実験のデータ
    - 簡便な機能
- NFS v4
  - Kerberosによるユーザ認証に期待
  - Linux のクライアントの安定性

将来の課題

- システム更新時のデータ移行
  - 移行に要する時間
  - 運用への影響

磁気テープ装置 vs.磁気ディスク

	所要電力 kVA	設置面積m^2
磁気テープ 3PB	16	10
磁気ディスク 0.2PB	28	4
磁気ディスク 1PB	114	16
磁気ディスク 2PB	222	30
磁気ディスク 3PB	329	44

- 前頁の表について
  - 値は現行システムでの概算値
  - 磁気テープ装置/磁気ディスク装置とサーバ部分の合算値
  - 磁気ディスク装置の値1PB, 2PB, 3PBは現行システム200TBから算出。サーバ部分は同一として磁気ディスクのみを単純に増加させた