CDF nel Run II (2000 - 2003 - …)

- Cominciamo dalla fine, i bisogni:
  - **<u>RETE! RETE!! RETE !!!</u>**
    - intercontinentale
    - italiana
    - locale
  - system management: **linux**
  - licenze: compilatore C++: KAI
  - supporto software: Root, C++, Java, SMQL
  - collaborazione a distanza: videoconf + application sharing
  - ancora rete: Qualita' di Servizio

Data set di CDF

- DAQ rate su "nastro" (mediato su un anno) 30 Hz = **10^9 eventi/anno**

- dati raw (stored on tape robot) 250 KB / evento
- dati ricostruiti (DST) (tape robot) 135 KB /evento
- dati compressi per l'analisi (PAD) 60 KB /evento
  - dischi (~30 TB) + robot "8mm"
- replicazione dei dati nelle varie streams ~ 15 %
- dimensione totale dei dati: **500 KB /evento**
  - 10^9 eventi/anno = 500 TB /anno

- **Run II =** 2 fb^{-1} ~ **2 anni** @ $L = 10^{34}$ $\Rightarrow$ **1000 TB = 1 PetaByte**
  - overall PAD data set = 200 TB

Come si prepara Fermilab

- "Come Run-I, con i bugs aggiustati"
- La produzione (raw → DST +PAD) si fa a FNAL
- TUTTI i dati (1PB) sono disponbili a FNAL su nastri "8mm" su robot
- l'analisi si fa sui PAD (raramente DST) per produrre n-tuple
- cluster di multi-cpu unix servers (SGI, Sun, Linux)
  - tape drives : SCSI locale
  - 30 TB disk pool : Fiber Channel Storage Area Network
  - desktop nelle baracche (linux PC con fast-ethernet): Gigabit Ethernet
- PAD/DST copiati su siti remoti per analisi che richiedono ripetuti accessi agli stessi dati:
  - DHL : 1 nastro(100GB) / giorno = 10Mbit/sec
  - WAN : ?????? Difficilmente competitiva in generale, ma molte universita' americane hanno O(155 Mbit/sec)…(UCHI, UIL…)

Quadro Generale per l'Italia

- data set "LHC-like" 5 anni prima:
  - 1 PB di dati. 200 TB di PAD.
  - million-$'s analysis facility a FNAL per gestirli
    - ☞ non pensiamo di replicare il data set in Italia

- l'analisi si sviluppa su data set selezionati in base al trigger
  - alto Pt : 100 ~ 1000 GByte
  - basso Pt: 1 ~ 10 TByte

- il fisico lavora soprattutto sulle n-tuple: 1 ~ 5 Gbyte
  - **n-tupla = pochi minuti di accesso per uso interattivo**
  - compito ideale per un PC desktop
    - ☞ CDF ♥ linux

Il piano italiano per l'accesso ai dati

- Dove stanno i dati. **Struttura a 3 livelli.**
  - Data set completo (1 PByte) : FNAL
  - ~n-tuple (≤ 100GByte): disco locale
  - una n-tuple si crea da un data set ≤ TB
  - data sets intermedi (pochi TByte): server sulla LAN o in una altra sezione INFN (GARR-B) per i data sets di uso frequente.

- Come si accedono:
  - **"BRING THE JOB TO THE DATA"**
  - disco locale: tutto ovvio.
  - ogni altra cosa: batch job per produrre le n-tuple, copia delle n-tuple sul disco locale:
    - ☞ da un server italiano: copia via rete
    - ☞ da FNAL: via rete fino a pochi GBytes, poi DHL

Rete vs. Computers

- Si risparmia un data center italiano da 200TB ($$$ e persone)
- Serve una rete molto migliore di quella attuale.
- 1 ora per copiare una n-tupla = 1 Mbit/sec/user * O (20 users)
  - servono decine di Mbit/sec per CDF
- stima (a Febbraio sembrava credibile) per la b/w con FNAL per CDF:
  - 4 Mbit/sec nel 1999
  - 10 Mbit/sec nel 2000
  - 20 Mbit/sec nel 2001
  - 50 Mbit/sec nel 2002
- Questo e' il cardine del piano, se la rete viene a mancare salta tutto.
- Anche la rete italiana deve crescere ben oltre GARR-B (USA x 5~10)
- Apparentemente nessun problema: 45 Mb/s con gli USA espandibile a 155, e Garr-G (Gbit/s) promesso sulla stampa.

Il "mio" problema con la rete

- A chi chiedere ?
  - Da oltre un anno i nostri bisogni sono "noti"…
  - chi ci deve rispondere ?
- Quale e' la schedule (=documento ufficiale) dell'INFN per la rete ? Come evolve ? Quali sono i problemi e le soluzioni ?
- Esempio: dove sono i 10 Mbit/sec con Esnet attraverso TEN-155 ?
- Unica cosa certa da anni: INFN-FNAL =1.5Mbit/sec, links sempre piu' saturi (e.g. da PI e TS).

- Se la rete non si materializza, che succede ?
  - Copiamo 200 TB di dati in Italia ? In ogni sezione ?
  - Andiamo a fare l'analisi a Fermilab ?
  - Lasciamo "scoprire" tutto agli americani ?
  - In ogni caso, servono molti soldi, vanno previsti per tempo.

Analisi in Italia

- Cosa si fa in Italia:
  - sviluppo codice
  - Monte Carlo (generazione e simulazione)
  - analisi di n-tuple e piccoli data set
- Gli strumenti necessari:
  - PC, nastri SCSI 8mm, disk servers (4-CPU + 1~2 TB RAID)
  - 100Mbit switched ethernet
  - linux, C++, Root, SMQL, batch
  - ancora possibile l'uso di sistemi operativi proprietari (Solaris, IRIX) per i disk servers.

Supporto software

✪ **La politica della collaborazione e' di fornire pacchetti pronti all'uso alle istituzioni. "Ci basta poco: tanto system management".**

- **Non sottovalutare linux** (molti PC, diversi tra loro, compatibilita' hardware/software puo' essere una difficolta'. I desktop di CDF non sono terminali, ma macchine Unix in continuo bisogno di software maintenance).
- **CDF e' OO** (C++ per l'offline, Java per l'online), ma non ci attendiamo un supporto particolare nelle sezioni. Expertise in sede e' una bella cosa, si acquisisce sviluppando codice.
- Distribuzione del codice nun usa AFS, basata su ftp (la rete !!)
- Distribuzione del database: **SMQL**. Competenza locale desiderabile.
- **Root**: sara' comunque distribuito da Fermilab. " " ".
- Licenze: C++ compiler (KAI): ~50 Mlit divisi in piu' anni.
- Batch system ancora un'incognita. LSF ?

Videoconferenze etc.

- CDF e' stato il pioniere del "telelink", siamo ancora della stessa idea.
- Videoconferenze su Codec/ISDN sono ancora fondamentali.
- M-bone poco usato, tentativi sempre deludenti, molto meglio il viva-voce … forse per colpa della rete ?
- Bisogno ancora insoddisfatto: **collaborazione a distanza**
  - ➢ lavorare attraverso la rete come nella stessa stanza
  - ➢ application sharing (windows sharing)
    - ☞ le ns. applicazioni, non Power-Point !

Oltre i Mbit/sec: Qualita' di Servizio

- La rete sara' satura di dati copiati avanti e indietro.
- E' pero' vitale:
  - sottomettere e controllare jobs a FNAL che creano le n-tuple
  - partecipare al controllo e monitoraggio dell'esperimento (dalla manutenzione dell'hardware alla remote control room)

- Priorita' sul link per il lavoro interattivo.

- Bassa latency per applicazioni particolari:
  - client-server con ridotto flusso di dati, usato e.g. per "i turni":
    - ☞ separare il display/browser dall'hardware/analisi
    - ☞ server @ Fermilab, client locale in Italia
  - R & D in corso (QUADIS in Gr V). Support it !

Una proposta di sviluppo software

- Come ridurre il traffico di tutti questi GigaByte di n-tuples che viaggiano sulla rete e garantirsi comunque l'accesso immediato ai dati piu' recenti per un'analisi competitiva ?

- Una provocazione: tutto l'hardware a Fermilab, solo X-Term in Italia. Serve circa la stessa banda di rete, ma l'accesso ai dati e' garantito.

- Una alternativa intelligente, ma ora impossibile:
  - "run the job where the data is". Dal batch alla analisi interattiva:
    - ☞ estendere e.g. il meccanismo client-server di Root per lavorare su n-tuple remote inviando localmente i comandi di selezione e ricevendo indietro solo l'istogramma (= 100 numeri reali) invece di tutta la schermata. Fit, plot, confronti etc. si fanno usando la CPU locale.