Calcolo per CDF-Italia per il Run 2

- Nulla e' incerto come il calcolo
- A parte la durata del Run 2
- Run 1 (1985-1993)
  - Vax 780 → Vax Cluster → VM/370 → Vax Stations → big SGI
  - 9trk → Exabyte 8220 → Exabyte 8500 → STK (Redwood)
  - "winchester" → HSC disks → local SCSI
- regole d'oro:
  - flessibilita'
  - acquisti Just In Time
  - la fisica al primo posto: sistemi semplici e robusti
- Una certezza
  - CDF NON HA SPESO ANCORA NULLA PER IL CALCOLO DEL RUN2, solo PC desktop dalle dotazioni.
    L'hardware del Run 1 (Alpha) non e' supportato per il Run 2.

Quadro Generale

- data set "LHC-like" 5 anni prima:
  - 1 PB di dati. 200 TB di Physics Analysis Data **(PAD).**
  - 10 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
  - compito ideale per un PC linux

CDF data sets

- Liz's table

L'analisi: per via gerarchica

- CDF Italia workshop a Padova 6/7 Dicembre 1999:
  - esaminato il flusso di dati in alcune analisi "italiane" del Run 1
  - estrapolato alle dimensioni dei data sets per il Run 2 (x20)

canale	PAD	Ntuple
$t \rightarrow 6j$	0.5 TB	8 GB
B mesons	2.4 TB	~40GB
W/Z H $\rightarrow$ 4j	1.2 TB	20 GB
W $\rightarrow$ lj (include top)	0.4 TB	2 GB
min bias	0.06 TB	20 GB

- altrettanto per il Monte Carlo
- numero di iterazioni delle selezioni:
  - DST $\rightarrow$ PAD ~3 volte (a Fnal) PAD $\rightarrow$ Ntuple ~6 volte (a Fnal)
  - re-tuning dei cuts su piccoli campioni e affinamento Ntuple : ~100

p-pbar : tanti canali, tanta variabilita'

- da 2 a 20 TB di dati
- da 100 ev/sec a 100 sec/ev nella simulazione
- da 1 GeV a 100 GeV di "Pt tipico"
- da pochi mesi (raffinamento di un SUSY search) a molti anni (massa W, sezione d'urto top, "nuova fisica") di lavoro
  - i bisogni cambieranno nel tempo !
  - a CDF non ci sono "analisi istituzionali"
    - ☞ ogni nuovo studente puo' aprire un nuovo capitolo

- **i bisogni (e le richieste) di hardware per l'analisi del gruppo Italiano dovranno seguire l'evoluzione dei gruppi di analisi**

La preparazione al Run2: un esempio (Padova)

- sviluppo software (OO) di tracking in SVXII e COT, con applicazione di algoritmi veloci per la possibile applicazione a livello 3 di trigger.
  - *P. Gatti, A. Ribon, P. Azzi*
- studi di trigger per W/Z+H $\rightarrow$ 4jets. Su dati del RunI, e, per ora, con software del runI, con il pacchetto di estrapolazione al RunII.
  - *M. P. Giordani, L. Scodellaro, P. Azzi, T. Dorigo*
- simulazione del sistema di trigger SVT, anche con layer 00. Applicazioni fisica del B e dei bosoni pesanti. Usando dati e simulazione del RunI, simulazione del RunII, simulazione standalone per SVT.
  - *D. Lucchesi, I. Fiori, T. Dorigo*
- analisi del runI ed estrapolazione al RunII per B_s $\rightarrow$ J/Psi eta.
  - *D. Lucchesi, R. Rossin*
- algoritmi di reti neurali per l'identificazione di b-jets
  - *S.Bettelli, A.Sidoti (Trento)*

Come si prepare Fermilab (bisogni stimati nel 1977)

- Le richieste al laboratorio "per tutti" (finale/adesso)
  - farm per ricostruzione e (un po' di) MC
    - ☞150/50 PC dual CPU 500MB rack-mount (1.1 GL)
  - robotic tape library da 1pByte per TUTTO il data set
    - ☞ 20 8mm Exabyte Mammoth 2 (drive+media = 2.4 GL)
      - 4 Sony AIT2 per il run di Settembre/Ottobre
    - ☞ Emass/ADIC robot (2 GL)
  - FC-SAN (1.5 GByte/sec) con 30/10 TByte disk (1.5 GL)
    - **+ 10 TB "inglesi"**
  - Central Analysis Facility per "user's analys"
    - ☞ multiprocessor servers per 90K/30K Mips (5 GL)
      - ora: 64-CPU Origin2000 = 1/3 del totale a 1/2 il costo
  - Gibabit Ethernet verso i desktops nelle baracche
  - system management

FCC

- Stephan's picture

Esperienza del Run 1

- L'analisi "ufficiale" e' diretta dai convenor dei gruppi di analisi, full time residents a Fermilab per diversi anni
- Gli Italiani hanno potuto lavorare grazie alla disponibilita' di risorse riservate a Fermilab utilizzabili autonomamete:
  - disco (soprattutto sul sistema centrale)
  - CPU (Vax/Alpha negli uffici e code batch sul sistema centrale)
- Il lavoro in Italia:
  - una pena l'accesso ai nastri, gli ultimi anni si preferiva usare (dall'Italia) il robot STK di Fnal e copiare i dati sulla rete.
  - bene MC e PAW
  - una fatica continua, ma gestibile aggiornamento codice e data base
  - una fatica da non ripetere la gestione di 10 Vax Stations, 30 dischi, 10 unita' nastro di varie taglie/eta'/efficienza
  - **critico l'accesso all'ultima versione, incluso codice e dati in aree "private" ⇒ fondamentale "runnare" a FNAL**

I contro dell'analisi in Italia

- Costo economico: robot, nastri, SAN ... >>1 GLit
- Costo umano: la gestione non e' solo system management "ad horas", ci vuole professionalita' ed esperienza tecnica e dirigenza scientifica (quali dati, quale codice, quali risorse, validazione delle copie locali)
- Stress: si rincorre sempre, i nuovi dati ed i codice sono disponibili sempre prima a Fermilab che in Italia
- Inadeguatezza: si lavora sempre con risorse inferiori a quelle disponibili a Fermilab cercando di adattare strumenti sviluppati per un ambiente diverso
- Insicurezza: c'e' sempre il pericolo che qualcosa rompa la compatibilta' col centro remoto, ogni acquisto porta una nuova versione del sistema operativo...
- **Il vero problema**:
  - **come garantire, a se' ed alla collaborazione, che l'analisi remota e' riproducibile al laboratorio senza effettivamente replicarla al Fermilab.**

La nostra strategia: tenere la CPU vicina ai dati. 3 colonne portanti

- **Fermilab** Feynman Computer Center (**FCC**) per:
  - accesso all'intero data set
  - accesso immediato alla versione corrente di software/dati/DB
  - system management professionale con tools software ed hardware dedicati sviluppati ad hoc
    - ☞ aggiungere un po' di hardware e' ~ trasparente
  - saricare su FNAL il costo di software, manutenzione, operazione
- **risorse locali** in Italia per
  - analisi ripetute su piccoli data set
  - analisi "interattive"
- **wan network** per
  - trasferimento "n-tuple" (~1GByte)
  - remote job entry/control al FCC

Rete vs. Computers

- Si risparmia un data center italiano da 200TB ($$$$!)
- Previsioni di utilizzo per solo CDF per trasferimento dati con gli USA:
  - 4 Mbit/sec nel 2001
  - 10 Mbit/sec nel 2002
  - 20 Mbit/sec nel 2003..
- Serve una rete migliore molto migliore di quella del Run 1
  - Garr - US Research: 45 Mbit/sec ora , 155 planned: ~OK
  - Remote monitoring e lavoro interattivo a FNAL
    - ☞ pochi bytes, ma altra priorita': QOS (R&D in corso con Fermilab in gruppo V (Quadis/GRID)
  - Banda di rete garantita e prenotabile (GRID ?)

Accesso ai dati dall'Italia

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

- Come si accedono:
  - 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 Federal Express

CPU/Disco

- I bisogni di data storage sono chiari
- Le necessita' di CPU non sono mai stimabili correttamente
  - il software e' un gas libero, l'espansione delle eseguibile (memoria occupata, CPU, spazio disco) e' limitata dalle risorse disponibili
  - in ogni caso la logistica deve occupare poco tempo, la fisica tanto
    - ☞ niente home-made-farms
    - ☞ **un job per task per persona = una CPU per fisico**
    - ☞ **corrisponde all'estrapolazione dal Run 1**
  - bisogna comunque riparlarne quando i numeri sono solidificati
    - ☞ ora il software sta appena prendendo forma (CDF tradition):
      - 20 eventi/secondo solo lettura senza tracking
      - 3 eventi/secondo per generazione (ISAJET)
      - 1 evento/minuto per simulazione, lo stesso per ricostruzione
    - ☞ siamo lontanti dal goal.

Corollario

- L'analisi satura qualunque set di risorse
- Non esiste "la cifra giusta"
- Piu computers = meno fatica (se la gestione e' semplice)
- Dato un qualunque totale di risorse finanziare
  - spendere efficientemente
  - risparmiare quanto possibile
  - comprare piu' CPU e piu' disco
    - meglio avere piu' hardware a FNAL dove l'investimento e' piu' efficiente e si fatica di meno
    - ovviamente fino a che totale < 10 GLit.

Quale hardware

- ITALIA (ntuple)
  - sezioni piccole: 1 PC + 100GB disco + 8mm
  - sezioni grandi: 4-CPU servers, 4 8mm drives, ~1 TByte RAID
    - ~7 MLit/anno per ogni fisico che fa analisi **(LINUX !)**
- ITALIA (PADs)
  - ~5 TB disco RAID, 8mm, CPU "da stimare"
- FNAL
  - servers nel Computing Center connessi a tape robot e RAID disk pool (garantiscono accesso immediato al data set completo)
    - 10 TB disk, ~15% di 90.000 Mips = 15.000 Mips
    - PC negli uffici: come in Italia ~1.5ML/anno/tavolo
- SIMULAZIONE (PC FARMs)
  - Italia o Fnal: stesso costo (O(100MLit)), cambia la fatica, da discutere al momento del bisogno

Il Piano

- Sistema distribuito, espandibile. Dettagli adattati/bili.
- Constraints di unitarieta':
  - quadro temporale: computers = acquisto just-in-time
    - nulla nel 2000
    - poco nel 2001
    - molto nel 2002
    - altrettanto nel 2003
    - upgrades periodici e manutenzione (10% all'anno)
  - regole di somma (solito caveat sulla CPU):
    - ~50 CPU in Italia (servers + desktops da analisi)
    - due PC per ogni ufficio a FNAL: ~30 tot
    - hardware al FNAL Comp.Center comune per tutti:
      ~ 15000 Mips (450SpecInt95) + ~ 10 TByte RAID
    - drives 8mm (nastro) in Italia e Fermilab (~20+10)

Indicativamente, se le stime di CPU di 3 anni fa (estrapol. dal Run 1) reggono e se in Italia possiamo stare con Linux

- 2001: anno zero, partiamo piano sfruttando l'hardware "di Fermilab"
  - 200 MLit in Italia, 100MLit a Fnal (2TB disco)
  - eventuali aggiunte da discutere in corso d'anno

- da qua in poi si puo' sbagliare (in meno) di un fattore ≥ 2
  - 2002: il primo fb-1:
    - Italia: 3TB disco, CPU da definire O(~300 MLit)
    - Fnal: 3TB disco, 5.000 Mips O(~500 MLit)
    - 10~20 nodi per Monte Carlo Farm (anywhere) O(~100MLit)
  - 2003: completato 1 PetaByte di dati
    - Italia: 4TB disco, altra CPU O(~300 MLit)
    - Fnal: 5TB disco, 10.000 Mips O(~500 MLit)
    - 10~20 nodi MLit MC Farm (anywhere) O(~100MLit)
  - 2004: analisi a pieno ritmo
    - sara' in discesa... se non arrivano nuovi dati !