Organizzazione analisi per il Run 2a

- In CDF non esistono "analisi ufficiali".
- "physics groups" liberi discutono ed orientano le analisi ed approvano (per acclamazione) i risultati da pubblicare, (sub)working groups "as needed".
  - Trigger & DataSets ✯✯
  - Top ✯
  - ElectroWeak
  - QCD ✯ : ☞di-jet mass ✯
  - B-physics ✯ : ☞reconstruction ✯ ☞B trigger & Dataset ✯✯
  - Exotic (SUSY, Higgs…) ✯

- ✯ convenor italiano ✯partecipanti italiani

Italiani che lavorano con/su tools offline:

- Ancora grande sforzo sull'hardware, molte attivita' sono incentrate sulla messa a punto e monitor dei rivelatori

- Studio del trigger e rivelatore (esp. SVT, SVXII, TOF, calorimetro)
  - 6.9 FTE
- Preparazione algoritmi di ricostruzione
  - 4.7 FTE
- Simulazione e preparazione per analisi dati
  - 20.1 FTE

- Gia' oggi piu' di 30 persone al lavoro

Dettaglio Italiani al lavoro "oggi" (parziale):

Bologna

t,H → multijet 2 FTE
Castro 100% Gresele 100%

min,bias 2.5 FTE
Rimondi 100% Deninno 50%
Zucchelli 50% Moggi 50%

high multiplicity trigger 0.5 FTE
Moggi 50%

Padova

t,H → multijet 3.2 FTE
Azzi 50% Dorigo 60%
Scodellaro 100% Cortiana 100%

Fisica del B 3.2 FTE
Lucchesi 70% Fiori 50%
Rossin 100% DaRonco 100%

tracking 1.5 FTE
Azzi 50% Sidoti 100%

simulazione ed SVT 0.8 FTE
Lucchesi 30% Fiori 50%

Pisa

Fisica del B 2.9 FTE
Punzi 50% Donati 50%
Carosi 20% Tonelli 50%
Belloni 100% Paoletti 10%
Turini 10%

Higgs 3.1 FTE
Pagliarone 100% Vataga 100%
Annovi 50% Dell'Orso 20%
Giannetti 20% Lami 20%

SVT & ISL monitor 4.1 FTE
Punzi 50% Donati 50%
Carosi 80% Chiarelli 100%
Leone 100% Tonelli 50%
Palmonari 30%

Roma

Fisica del B 5.9 FTE
Giagu 60% Rescigno 50%
DeCecco 80% Messina 100%
Vallecorsa 100% Diluise 100%

Energy Flow 3.2 FTE
Sarkar 50% Loverre 100%
Dionisi 70% Margaroli 100%

SVT & TOF monitor 2 FTE
Giagu,Rescigno,Sarkar,DeCecco per la percentuale rimanente

Hardware : Sommario dall'11 Settembre 2000

- Abbiamo un modello :
  - "ntuple" a casa
    - O(1 cpu/user) O(100GB/user)
    - risorse sparse nelle sezioni
  - Produzione n-tuple da data sets O(>1TB) soprattutto a FNAL
    - risparmio denaro e risorse umane
    - accesso a 200TB dati ridotti e 1PB dati totali su robot
    - garanzia riproducibilita' e integrita' software & DB
- Serve un piano di implementazione (oggi) :
  - Cosa
  - Quanto
  - Quanto costa
  - Tempi
  - Modi

In Italia

- Siamo "ancorati" su 1 cpu + 100GB per fisico
- risorse sparse nelle sezioni proporzionalmente alle dimensioni dei gruppi
- condivisione risorse attraverso (a cura di Bologna) :
  - AFS centrale
  - batch condiviso via Condor/Globus
  - Replica locale freeware del DB
- configurazione hardware clonata per quanto possibile dal test bed di Grid (grazie referees!)
  - Preoccupazione sul personale necessario
- sviluppo adiabatico nei prossimi anni (2003+4 ~= 2001+2)
- concentrazione in un unica sede (centro di calcolo INFN) possibile in qualunque momento se c'e' la struttura
- GRID Fnal-Italia ipotizzabile per Run 2b

Il "(wo)manpower"

- Abbiamo una enorme carenza di persone con vocazione di system manager, progettista di centro di calcolo, compratore di hardware etc.
- Ai fisici di CDF piace guardare i dati e fare analisi
- E' una tendenza che preferiamo non scoraggiare anche se va contro la linea atutlae di massimizzare la quantita' di hardware con sistemi fatti in casa quanto possibile.
- Comprendiamo il desiderio di risparmiare quanto possibile, ma non vi stupite se preferiremo acquistare il 20% di hardware in meno anziche' rinunciare a 3 mesi di lavoro di uno studente riunciando a spremere l'ultima ottimizzazione dell'hardware.

A FNAL

- Oggi presentiamo un piano globale, con stime quantitative dei bisogni e dei costi
- Il run 2a e' in corso
  - >50TB gia' su nastro (piu' di tutto il Run 1) due passi della produzione effettuati su tutti i runs di fisica
  - 200pb-1 promessi per l'estate
  - ragionevole 1fb-1 entro il 2002
- non possiamo attendere di aver perso la corsa, abbiamo bisogno della certezza di avere le risorse necessarie quando verrano presi i dati "buoni" e ci sara' la grande corsa per le conferenze (primavera 2002).
- primi bisogni (disco per monitor SVT e fisica del B) tamponati da Trieste (avanzi di sezione e rinucia a tape drives) in copertura con i 2TB finanziati l'anno scorso (attendevamo dischi da 160GB e accordo con Fnal/CD)

Un po' di storia

- Offline computing non e' parte degli MOU (stesso nel Run1).
  CDF = Collider Detector at Fermilab
- FNAL fornisce (entro il budget DOE→Fnal→CD→CDF(=D0))
  - Data storage & ricostruzione
  - Esigenze comuni (batch e disco riservati ai gruppi di fisica)
  - Accesso democratico alle risorse che rimangono
- Come ha funzionato nel Run 1:
  - Gruppi di fisica (riunioni ogni 1~2 settimane) lavorano in modo assembleare con conveners residenti a Fnal
    - ☞ condizionati dai " locali"
  - Carenza di risorse sulle macchine centrali
  - Gran ricorso a VaxStations private (uffici FNAL e Italia)
  - gli Italiani hanno avuto batch e disco propri al FCC
    - ☞ IMPORTANTISSIMI
  - Italia non cosi' utile quando e' arrivato il robot per i nastri

Cosa discutiamo

- L'INFN non e' tenuto a contribuire al computing center di Fermilab
- Fermilab non promette di soddisfare tutte le esigenze di tutti
- Quando le risorse centrali non bastano: ognun per se'

- Qua, oggi e nel futuro parliamo dell'ultima parte:
  - Vogliamo essere attrezzati per competere alla pari (come minimo!) con le grandi universita' e laboratori, americani e non (Chicago, Pennsylvania, LBL, Urbana, UK, Canada …)
  - Non e' una tassa
  - Non e' una percentuale del costo totale
  - Non e' un esercizio in computing d'avanguardia
  - E' quelle che ci serve per essere sicuri di non aver problemi
    - ☞ IL GOAL E' LA FISICA
  - Vogliamo definire una spesa totale fissata per tutto il Run 2a, entro la quale ottimizzare metodo di lavoro e scelte hardware

Stima dei bisogni I: la storia passata

- CDF (FNAL CD) ha fatto una stima nel 1997:
  - 20x gli eventi del Run1, stesso numero di users
  - Stesso bisgno di CPU/evento/user
  - Curare problemi logistici e di scaling/espandibilita'
- previsti 12M$ per la Central Analysis Facility
  - Avuti ~10, "salvati" da ritardo Run2
  - Progetto concluso. 1M$ nel FY 2002 per "upgrade adiabatico" richiesto al laboratorio dalla CD
- Nel frattempo:
  - Ricostruzione ~0.2Hz nel '92, ~0.2Hz nel 2001 (x20)
  - Eseguibile: O(100MByte) (x20)
  - I/O ~3MBytes/sec: disco oggi = nastro ieri
  - "cura": work in progress
- Ma quali sono i veri bisogni ?

Stima dei bisogni II: il nostro approccio

- Usare il software/hardware attuali (primavera 2001)
- Individuare un set di risorse che
  - hanno un impatto significativo sul lavoro di analisi:
    - ci si potrebbe anche fare
  - sono realisticamente installabili/usabili "da subito"
  - hanno un costo ragionevole, allineato coll'investimento fatto nella costruzione del rivelatore
- diluire l'acquisto in 3 anni di spesa costante
- ottimizzare le scelte mano a mano che hw e sw migliorano
- sopperire alle esigenze restanti (crescenti) con
  - lo share comune delle risorse di FNAL
  - il miglioramento prestazioni/prezzo dell'hardware
  - l'ottimizzazione della architettura hw/sw

Stima dei bisogni III: l'esercizio

- Dimensione dei data sets dalla Trigger Table finale
- Sviluppo gerarchico dell'analisi di un data set di 10^7 eventi:
  - selezione di un campione ristretto (1TB->100GB): 2 volte
  - studio del campione finale (100GB->1GBntupla) : 10 volte
  - tuning su piccolo % del campione "ogni giorno"
  - simulazione di altrettanti eventi quanti raccolti dal DAQ
  - fare il tutto in 6 mesi
    - 1TB e' una media: top=200GB, B $\rightarrow$ hadron=10TB
    - bastano 2 e 10 iterazioni ? Perche' non 20/100 ?
      - Nel Run 1b circa 3/20, ma rivelatore molto cambiato, B->had tutto nuovo…
    - quanto MC "serve" e quanto dettagliato ?
- Misurato il bisogno di CPU a Marzo 2001 sulla Origin 2000 di CDF e sulla farm di ricostruzione col codice del Run 2 ed i dati veri
- 20 fisici attivi in ogni momento (n.b. 30FTE gia' ora !)

Stima dei bisogni IV: il risultato

- 40 "CPU come quella usata per il test" = 15.000 Mips
  - =1/6 della stima del '97 per tutta la collaborazione
- 15TB disk
- 50 nodi di simulazione (farm)
- Costo: 1M$ (da spendere a Fnal) (cpu-disk-farm=30-40-30%)
  - circa 350KEu all'anno
- cross checks
  - di gran lunga la stima piu' accurata fatta finora a CDF
  - random user/convener: tutto, tutti i giorni, su tutti i dati
  - 900 Mips/user stimato da MIT (basato su numero di jobs in esecuzione ad oggi)
  - 1.6Msterline (=2.6MEu) budget dei gruppi UK (31)
  - 20 nodi farm + tape robot in Canada (17)
  - (n) = #nomi in CDF DB, Italia=(100)

Quanto costa mettere computers a Fnal ?

- Abbiamo un accordo scritto con FNAL CD per possibile installazione
  - FNAL paga overhead, sys.man., manut. licenze ...
  - FNAL fa ricerca di mercato, scelta hw, test configurazioni, gare...
  - INFN riceve risorse equivalenti dedicate in aggiunta al pool comune
  - la politica precisa di utilizzo e' da trattare di volta in volta (acquisti graduali!)
- Pregi
  - Risparmio (IVA , networking, racks etc.)
  - Stefano cerca l'Higgs !!
- Difetti:
  - Le scelte hardware devono adattarsi a quelle del laboratorio
  - Non si puo' aumentare il numero di sys.managers necessari

Che hardware ? Configurazione attuale

- CAF: pool di SMP con dischi FC condivisi in SAN e tape drives SCSI locali:
  - 64x300MHz SGI, 24x900MHz Sun, 4x700MHz Intel
  - 30TB disk + 20TB da UK
  - Vantaggi: scalabilita', facilita' di gestione (limitati dal numero di system managers, 1 pro x 3 anni = 0.5M$), load balancing, accesso trasparente al disk pool, minino overhead di CPU per accesso a dischi e nastri
  - Svantaggi: costo
    - Sun = 10x (24 single PC-s) = 3x (3 Linux 8-cpu)
    - FC disk ~ 30% in piu' di NFS servers PC-based
- FARM: racks di 2xPIII. Strutturabile in "farmlets". Adatta per grossi lavori CPU-bound (almeno 10 nodi x giorni) e limitato I/O. Accede ai dati "attraverso i nastri".

Che hardware ? Alternative

- Non conosciamo un esempio funzionante di architettura basata "su PC's" che collega 200 processori a 50TB di disco e 1PB di nastro per un unico gruppo di utenti
- ci sono molti casi "vicini", e.g. PDSF a LBL (150 nodi, 7TB)
  - scalano ? quale e' il costo della connettivita' ?
  - quanto sono difficili da replicare ? (expertise locale)
- ci sono proposte rivoluzionarie
  - MOSIX (migrazione automatica del processo ai dati)
- si puo' forzare una maggior utilizzazione di risorse comuni
- si puo' migliorare il codice, soprattutto I/O:
  - ROOT multibranching, puff on demand
- si puo' modificare la struttura dei dati
  - Standard ntuple in ROOT con stessa info dei "dati"

Managing delle incertezze

- La performance del codice e' inaccettabile
  - italiani in prima linea nello spingere per una soluzione
  - a giugno creato un comitato per "raccogliere i lamenti" (tra altri tasks): ICRB. Chairman: Stefano Belforte
  - l'architettura del CAF era stata pensata per ottimizzare l'I/O, ma siamo CPU limited
    - acquisto seconda 24-cpu Sun rimandato (400K$)
    - CAF review committe per
      - benchmarking
      - suggerire alternativa alla seconda Sun
      - riesaminare tutto il sistema
      - membri italiani: Stefano Belforte, Flavia Donno
- la soluzione non e' una esplosione del budget, ma un cambiamento di rotta
- vogliamo fare di piu' con gli stessi soldi

Risoluzione delle incertezze

- CPU di analisi e disco:
  - Report del CAF review committee entro 15 novembre
  - prevedibile una "direttiva" al collaboration meeting di Gennaio
  - Definizione dei primi acquisti italiani per la CAF entro Marzo. Possibile usarli per indirizzare l'evoluzione del CAF
  - Continuare il ciclo:
    - verifica sul campo, evoluzione mercato -> nuove scelte
- CPU di simulazione
  - Definizione della logistica (validation, bookeeping) in corso
  - Iniziare il lavoro sulla farm attuale, dimostrare la fattibilita'/convenienza
  - Acquistare nodi nostri "in corso d'opera"

Implementazione del piano

- "Esercizio" presentato ai referees il 1 Aprile 2001
- Aggiunta la stima dei costi il 1 Maggio 2001
- Sottoposto alla Commissione Nazionale oggi 11 Sett 2001
- La richiesta alla commissione:
  - fissare una envelope di spesa totale per i 3 anni 2002/3/4 entro la quale ci possiamo muovere
  - approvare i finanziamenti periodicamente dopo discussione (con i referees) delle scelte specifiche
  - tener presente >6 mesi dalla approvazione all'utilizzo
- Il nostro impegno
  - non chiedere altro, non spendere per altre voci
  - ottimizzare gli acquisti per prestazioni/prezzo
  - continuare la pressione in CDF per architetture e/o tools di analisi piu' efficienti

Conclusioni

- Abbiamo stimato in modo realistico una configurazione hardware che ha un impatto significativo sulla capacita' di analisi del gruppo italiano
- Ora come ora queste risorse non sono sufficienti per fare l'analisi, m facciamo affidamento su:
  - X1.5 da evoluzione hw X1.5 da sharing risorse comuni
  - Il resto dal'ottimizzazione del lavoro e della spesa, anche indirizzando l'evoluzione del centro di calcolo di FNAL
- Contiamo di realizzare questa configurazione in 3 anni con un profilo di spesa costante e contenuto.
- Chiediamo alla commissione una rapida tabella di marcia:
  - All'inizio ci sono pochi dati, ma molto da capire e poca riduzione dei campioni, non possiamo aspettare $2pb^{-1}$
  - All'inizio bisogna usare la configurazione attuale