COMPUTING DEL GRUPPO ITALIANO PER IL RUN II DI CDF
Versione 1.0

Stefano Belforte

April 9, 1999

Contents

1 Quadro generale

Questo documento illustra il piano del gruppo CDF-Italia per l'analisi dei dati che saranno raccolti nel Run-II del Tevatron a partire dall'Aprile 2000. E' il risultato di diversi mesi di discussioni all'interno di tutta la collaborazione CDF, che hanno coinvolto anche la parte non italiana [1] [2] [3] [4]. Pertanto ci aspettiamo che molti dettagli possano certo cambiare nei prossimi mesi o anni, ma la strategia complessiva non sia modificata.

Il punto più importante da considerare è la quantità di dati che CDF raccoglierà nel Run-II e che quindi dovranno essere analizzati [5]. Il Run-II è definito più che da una precisa durata temporale (stimata in due-tre anni) dalla racolta di una luminosità integrata di $2fb^{-1}$, ovvero circa 20 volte quanto raccolto ed analizzato nel Run-I del Tevatron. La quantità di dati RAW e ricostruiti (DST) corrispondente è di circa 1PB (1 PetaByte = 1000 TeraByte = 1 milione di GigaByte).

Questo numero "magico" (1PB) è quanto previsto come produzione totale di dati per un anno di operazione di un esperimento di LHC (CMS od ATLAS). Il nostro compito è quindi affrontare e risolvere con diversi anni di anticipo (e molte meno risorse) un problema dello stesso ordine di grandezza di quanto previsto per LHC. Fortunatamente il nostro lavoro a CDF è facilitato da alcune importanti differenze con LHC:
- partiamo sulla base dell'esperienza del Run-I
- sappiamo abbastanza esattamente cosa fare
- non usiamo approcci futuristici (come Objectivity) al problema del data handling

Il fatto di aver conservato una struttura dei dati di analisi simile a quanto usato in passato e' quello che permette di sfruttare l'esperienza passata per definire la strategia per il Run-II. Gli stadi del processo di analisi dei dati individuati da CDF sono i seguenti[6][8]:

1. i dati RAW sono divisi in circa 8 diverse streams dal Livello 3, spediti su fibra ottica al Feynamm Computer Center (FCC) di Fermilab e lì memorizzati su cassette 8mm (raw data set).

2. tutti gli eventi vengono ricostruiti da una farm di PC's al FCC ed i DST risultanti divisi in circa 40 streams (primary data set) e memorizzati su cassette 8mm. Il volume totale di RAW data e DST e' circa 1PB.

3. parallelamente ai DST, vengono creati files contenenti una quantità ridotta di informazione per ogni evento, sufficiente per la maggior parte delle delle analisi, i Physical Data Set, PADs, (ancora parte del primary data set), anch'essi memorizzati su cassette 8mm. Il volume totale dei PAD sarà di circa 200TB.

4. tutti i dati su cassetta saranno disponibili per la analisi dei menbri della collaborazione che lavorano sui computer del FCC attraverso unità nastro robotizzate.

5. i physics groups produrranno campioni di eventi in formato PAD ulteriormente selezionati che saranno permanentemente disponibili su disco al FCC (secondary data set). Lo spazio disco disponibile sara' circa 20TB.

6. gli utenti finali (i fisici!) lavoreranno prevalentemente a partire da PADs residenti su disco o su robot per produrre campioni ulteriormente selezionati (tertiary data set) o direttamente n-tuple.

7. le n-tuple finali o tertiary data set di piccole dimensioni e riutilizzati molto frequentemente potranno essere copiati su dischi locali negli uffici o nelle istituzioni remote.

Per la realizzazione di questa catena di analisi è prevista [6] una "analysis facility" al FCC basata su un robot con capacità complessiva di circa 1PB, ed un pool di medium-size SMP unix servers (8-cpu tipicamente) con ottima connettività al robot (SCSI locale) ed ad un pool di dischi (Fiber Channel) in configurazione SAN. L'analysis facility al FCC si presenta come molto più evoluta di quanto avessimo nel Run-I e con ottime capacità di scalare in futuro a seconda delle necessità e delle preastazioni del nuovo hardware. Tutti gli ostacoli che in passato hanno impedito un efficiente accesso ai dati al FCC ed hanno quindi spinto a spostare il lavoro di analisi nelle università per problemi di mancanza di CPU o di difficile accesso ai dati, saranno risolti.

Per il Run-II occorre tener ben presente questa "novità" (un ottimo centro di calcolo al FCC) per costruire una strategia per l'analisi in Italia che permetta di essere competitivi con chi lavora al Fermilab.

Infine due note tecniche per quanto riguarda l'accesso ai dati su nastro, che impongono precisi vincoli alla strategia di analisi:
1) i dati saranno scritti in formato ad accesso diretto (ROOT) e potranno essere acceduti solo dopo essere stati copiati da nastro a disco.
2) i nastri saranno cassette tipo 8mm ed avranno capacità O(100GB), analogamente al passato le unità nastro saranno ancora di tipo economico, mentre le cassette stesse saranno molto più costose [7].

Infine, proprio per sfruttare il drammatico aumento del rapporto prestazioni/prezzo dei PC che si acquistano sul mercato di massa, CDF ha deciso di portare su linux tutto il suo software, per cui anche noi in Italia cercheremo di sfruttare al massimo la possibilità di usare mezzi di calcolo economici.

2 Linee guida

Questo lavoro si basa sulla estesa esperienza da noi accumulata nell'analisi dei dati del Run-I. Il gruppo italiano è riuscito in passato a svolgere un ruolo importante nell'analisi dati di CDF ed ha dimostrato come l'analisi di esperimenti con grandi quantità di dati possa essere portata avanti con successo anche in Italia. Da tale esperienza deriviamo alcune raccomandazioni importanti per la prossima volta.
- non eccedere con l'hardware distribuito. La gestione di grossi clusters di workstations e dischi di tipi e performance diversi è estremamente faticosa, fino al punto di sottrarre risorse umane significative al lavoro di analisi. Quando un computer su ogni desktop non basta più è meglio usare piccoli multi-cpu servers.
- evitare di "maneggiare" diverse centinaia di cassette in unità nastro gestite dai singoli utenti. Al di là di ovvie considerazioni di comodità questo è inefficiente e certamente non competitivo con chi ha a disposizione un ampio pool di unità nastro robotizzate.
-migliorare le comunicazioni delle sezioni INFN tra loro e con l'America. Mentre c'è stato successo nell'effettuare analisi interamente in Italia, non si è riusciti ad integrare efficaciemente il lavoro di fisici residenti ai due lati dell'oceano (od anche in diverse città italiane).
- migliorare la capacità di lavorare dall'Italia in modo competitivo con i gruppi residenti a Fermilab, attraverso un rapido accesso a campioni nuovi di dati. Le analisi effetuate in Italia si riferiscono quasi interamente ad analisi separate svolte interamente da un gruppo italiano per lo più su argomenti "non-caldi". Non si è mai potuto essere competitivi sulle analisi di punta portate avanti da chi lavorava presso il laboratorio, sia per la lontananza dai meetings e la difficoltà di interazione con gli esperti residenti a Fermilab che per il ritardo nell'accesso ai nuovi campioni di dati.
- sfruttare maggiormente i mezzi di calcolo a Fermilab. Soprattutto per il Run-Ia la disponibiltà di code batch e spazio disco a FNAL riservati agli italiani è stata fondamentale, ma non sufficiente. Soprattutto bisogna permettere un più facile e rapido accesso dall'Italia ai mezzi di calcolo a Fermilab ed ai campioni di dati residenti là.

Le nostre linee guida sono pertanto:

- piccoli/medi servers per la potenza di calcolo che eccede "un PC su ogni tavolo"
- massimizzare lo spazio disco
- niente grosse farms, ridotto impegno di system management, contare sull'impegno distribuito dei singoli
- avere i nostri computers e dischi a Fermilab, in particolare al FCC con ottimo accesso al robot ed al pool centrale di dischi

- sfruttare al massimo i mezzi di calcolo a Fermilab
- sfruttare e sviluppare i tools di interazione e lavoro a distanza oltre la videoconferenza
- fare affidamento su un costante upgrade della rete con gli USA anzichè replicare in Italia il computing center di CDF al FCC

3 Il quadro temporale

La vera data di inizio del run è ancora incerta, ma la sostanza è invece abbastanza solida se si riassume la schedule di CDF in un periodo di messa a punto che inizia nel 2000 e termina alla metà del 2001, seguito da un periodo di due o tre anni di presa dati in "production-mode". Quindi l'accumulo di dati per analisi di fisica inizierà nel 2001, ma il 2000 sarà comunque un anno importante per la messa a punto del rivelatore e degli algoritimi di trigger (in particolare il livello 3, interamente basato su codice di tipo offline), nonché per lo sviluppo del software offline stesso e la produzione di primi campioni di Montecarlo.

Pertanto lo scenario temporale per il computing richiede una prima installazione di mezzi di calcolo nel 2000 per avviare l'attività e permettere ai membri della collaborazione precedentemente impegnati nella costruzione di dedicarsi maggiormente all'offline, un primo nucleo di hardware più potente nel 2001 per avviare l'analisi vera e propria e successivi upgrades nel 2002 e 2003 in modo da poter fronteggiare la crescente mole di dati.

Lo spirito di questo piano è acquistare l'hardware just in time al momento in cui serve, in modo da sfruttare al massimo lo sviluppo tecnologico, evitando un grosso investimento iniziale. Ciò nonostante bisogna tener presente che qualunque mezzo di calcolo ormai diventa obsoleto in poco più di 3 anni [1], e che l'hardware è si diventato molto economico qualora ci si rivolga al mercato di massa, ma a prezzo di un certo abbassamento qualitativo ed alla mancanza di un vero servizio di manutenzione vecchio stile. Per cui anche i costi di manutenzione imporranno un continuo update del materiale negli anni successivi e la spesa di calcolo non può essere considerata un costo destinato ad esaurirsi con il termine della presa dati.

[1] ogni 3 anni c'è un aumento di circa un fattore 4 in capacità dei dischi e performance della CPU. Diciamo che dopo 3 anni diventa desiderabile rimpiazzare un computer con uno nuovo, e dopo 5 il costo di manutenzione e funzionamento non è più giustificato dalle prestazioni. Inoltre è anche importante la competitività con le altre istituzioni, non si può fare analisi di punta con hardware non adeguato.

4 Stato e prospettive dell'hardware

Esaminiamo quale computing hardware è disponibile al giorno d'oggi e la probabile evoluzione nei prossimi anni, per individuare quali siano le configurazioni di sistema adatte.

In questa analisi è utile tenere presente il confronto con la situazione del Run-I di CDF, che ci può offrire una utile guida. Il numero importante è la quantità di dati, che nel Run-II aumenta di un fattore 20.

4.1 Nastri

La scelta dei nastri, sia media che drives, è effettuata dalla collaborazione completa, anzi sarà una scelta comune a CDF e D0 eseguita di concerto con la Computing Division di Fermilab. La scelta finale è per unità nastro low end (commodity) con costo per drive di pochi MLit (analogo al Run-I) e capacità per nastro di 70-100GByte. Però il costo dei singoli nastri (media cost) sarà molto aumentato da circa 10 a circa 150 KLit per nastro. Questo ribadisce la necessità di evitare le analisi basate su centinaia di cassette 8mm effettuate in Italia in passato.

4.2 Dischi

Assumiamo un fattore 4 di aumento di capacità a parità di costo alle fine del 2000. Per quanto possibile usare IDE. Attuale max capacità = 12GB IDE e 36 GB SCSI. Nessun problema per avere 100 GB su un PC alla fine del 2000 con e.g. 2 dischi IDE.

Per bisogni di disco fino a qualche centinaio di GB basta condividere i dischi locali dei PC, per 1TB o più meglio mettere tutti i dischi su un server. L'esperienza passata indica come unica possibile soluzione sistemi RAID. Per ora solo SCSI, 125GB disponibili ora (8 dischi), 0.5TB nel 2001, per ora stimiamo che siano necessari due RAID arrays "top-of-the-line" per avere 1TB. Stanno appena cominciando ad essere disponibili sistemi RAID basati su dischi IDE a prezzi più competitivi, malgrado non sia facile fare ora stime migliori di cosa potremo comprare nel 2000 o 2001, teniamo $1 \sim 2$ TeraByte come realistico per la quantità di disco che potremo collegare facilmente ad un piccolo server.

Notiamo che FNAL userà home made RAID systems. Noi dobbiamo essere piu' prudenti ed usare prodotti integrati commerciali (Digital, MTI, SUN etc.) che costano circa il doppio.

4.3 Cpu

La potenza aumenta oltre un fattore 20 (3100 vs penthium), ed il costo è enormemente ridotto. Assumiamo che la CPU sia una risorsa abbondante nel Run-II ed i bottleneck sia invece l'I/O dalla rete, da disco, e dalla memoria alla CPU. La CPU che potremo comprare nei prossimi anni al costo di un PII attuale sarà adeguata alle necessità di ogni singolo fisico (in media). Teniamo presente che ad esempio aggiungere una seconda CPU ad un PC desktop può essere fatto molto facilmente al costo di un paio di milioni in un qualunque momento.

4.4 Rete locale (LAN)

Per tutto il Run-II assumiamo 100 Mbits/sec per ogni user. E' allo stesso tempo il massimo offerto dalla tecnologia per i prossimi anni e ottenibile facilmente a basso costo. La CC sta finanziando l'upgrade delle LAN, confidiamo che soddisfi i nostri bisogni. Notare che aumenta solo un fattore 10 dal vecchio thin-wire ethernet da 10 Mbits/sec. E' anche molto più costosa (hubs, switches, fibre ottiche, cablatura a stella...) ma questa infrastuttura è ora esistente in quasi tutte le sezioni e la possiamo dare per scontata.

4.5 Rete estesa (WAN)

Per ora nessun miglioramento dal Run-I: sempre 1.5 Mbit/sec. Ci sono però importanti promesse di miglioramento: Il link GARR-ESNET diventa 10Mbit/sec tra 3 mesi, 45 entro l'estate e 155 possibili. Assumiamo uno sviluppo analogo per la banda utilizzabile da CDF. Come spiegato in seguito, ci attendiamo un grosso miglioramento di efficienza e risparmio di hardware dalla possibilità di sfruttare una banda di rete adeguata con gli USA. Pertanto richiediamo all'INFN di farsi garante di tali prestazioni investendo i fondi adeguati, o preparandosi a finanziare i necessari strumenti di calcolo in Italia in caso contrario:
-end 1999: 4 Mbit/sec
-end 2000: 10 Mbit/sec
-end 2001: 20 Mbit/sec
-end 2002: 50 Mbit/sec
Notiamo che questa è la banda che CDF avrebbe bisogno di utilizzare, il link GARR-ESNET dovrà ovviamente avere capacità maggiore ed essere opportunamente configurato. Inoltre a questi numeri intesi per l'uso di trasferimento dati, dovrà essere aggiunto il traffico necessario al controllo remoto dell'esperimento, comprese attività livello di "Remote Control Room".

Questa strategia "rete al posto di computers" è un pò il cardine di tutto il documento, riteniamo che sia praticabile non solo perché la tecnologia ormai lo consente facilmente e sono numeri in linea con quanto previsto negli altri stati e nelle estrapolazioni dello stesso gruppo reti dell'INFN, ma anche perché il costo di links di tale capacità con gli USA (al momento circa 36 MLit/anno per 1 Mbit/sec[9]) li rende concorrenziali anche da un punto di vista economico.

4.6 Run-I vs. Run-II

Cerchiamo ora di sommarizzare il quadro e trarne indicazioni precise. Mentre per il confronto dei bisogni abbiamo un semplice fattore 20 in mole di dati, non è purtroppo altrettanto semplice confrontare l'evoluzione dei mezzi di calcolo, dato che non è ovvio con quale anno confrontarsi, ed anche 6 mesi di differenza sono significativi data la velocità del mercato. La tabella seguente deve essere presa pertanto come puramente indicativa, qui prendiamo il 1994 come indicativo dell'inizio del RunI e ci confrontiamo con il 2000 (inizio del RunII) sfruttando una piccola estrapolazione di quanto disponibile oggi. Elechiamo nella tabella la "dimensione" del tipico elemento di hardware necessario, ed il costo unitario, ad esempio quanto è grande un tipico disco e quanto costa oggi rispetto al passato.

	1994	adesso	2000	Run-II vs. I	costo
data	$100pb^{-1}$	-	$2fb^{-1}$	x20	-
Tape size	2 GByte	30 GByte	50 $\sim$ 70 GByte	x25	x15
Disk size	2 GByte SCSI	18 GByte IDE	50 GByte IDE	$\geq$ x20	x1/10
CPU	VAX 4100	PII 400Mhz	P3 600MHz	$\geq$ x20	x1/5
Disk I/O	5MByte/s(*)	14MByte/s(*)	20 MByte/s	x4	-
LAN	10Mbit/s	100Mbit/s	100Mbit/s	x10	x10
WAN	1Mbit/s	2Mbit/s	10Mbit/s	x10	x?

(*) misurato al FCC su dati di CDF in situazioni realistiche

Come detto, questa tabella è solo indicativa, ci serve solo per trarre alcune considerazioni che non dipendono dai dettagli:

1. Il problema del tape handling rimane identico, ma con un costo maggiore.
2. CPU e spazio disco saranno disponibili molto piu' facilmente.
3. il vero bottleneck sarà l'I/O, occorre runnare i jobs dove sono i dati, non portare i dati a CPU remote.
4. per la prima volta nella storia di CDF si intravede la possibilità di sfruttare la rete intercontinentale (WAN) come un vero e proprio tool di integrazione in tempo reale del lavoro svolto ai due lati dell'Atlantico.

5 Data sets e scenari di analisi

Rivediamo la situazione globale: i dati aumentano 20 volte, 1 PetaByte totale, 200 TeraByte di PADs, 2000 cassette 8mm. FNAL avrà un sistema dove tutti i nastri sono su robot, tutti i data set piu' comuni residenti su disco. cpu per tutti.

Una situazione molto migliore del Run-I quando l'analisi si è dovuta spostare sulle VAX-stations dei trailers perché il sistema centrale era intasato.

Nessuno deve comunque lavorare su 1 PB né 200 TB di dati. Il campione complessivo (il PetaByte) è diviso in tanti data sets separati nelle diverse data streams a livello di produzione. Ogni data set può variare tra pochi % del totale ed una frazione di % [8]. Una tipica analisi usa uno o più data sets per il segnale e per gli studi di fondo ed efficienza. Possiamo considerare due casi estremi:

alto Pt: 1% dei dati. e.g. top, W., 2 TeraByte di PAD.

basso Pt: 5-10% dei dati. e.g. $B \rightarrow \pi\pi$, 20 TeraByte di PAD.

Anche assumendo un fattore 10 dai PAD inclusivi di un certo data set al campione usato per l'analisi, si ottengono 200 GB per alto Pt e 1 $\sim$ 2 TB per basso Pt. E' chiaro che quando si superano le centinaia di GB (possibile anche nel caso di analisi ad alto Pt, se la riduzione non viene o non può essere effettuata o se bisogna accedere un grosso campione di calibrazione come i leptoni inclusivi) questi campioni di dati sono troppo grossi per poter essere mantenuti facilmente sullo spazio disco locale di ogni fisico, od anche di ogni sezione. D'altra parte anche l'analisi da nastro è da scartare, sia per il costo dei nastri, che per la inefficienza di lavoro. Ricordiamo che lavorando al FCC questi dati saranno disponibili su dischi on-line strettamente connessi alle CPU con un canale ad alta bandwidth.

La via d'uscita è di valutare quali e quanti dati serve davvero avere in locale, e cosa invece possa essere fatto sfruttando il grosso robot ed il disk pool del FCC. La chiave dello scenario proposto è di **avere sul disco locale solo le n-tuple**. Una n-tupla è un oggetto maldefinito, non solo perchè PAW sarà rimpiazzato da altri tools durante il Run-II, ma perchè quello che rende un data set locale non è il suo formato, ma la sua dimensione. La nostra assunzione è che sia necessario avere su un disco ad accesso locale solo i data sets usati per analisi "interattiva", ovvero per attività come istogrammazioni che abbiano un tempo di risposta al massimo di pochi minuti. Jobs che richiedono un'ora per girare, non hanno nessun bisogno di essere sottomessi su dischi locali, possono essere eseguiti là dove sono i dati per riavere indietro solo pochi Mbytes di istogrammi (che dovrebbe essere immediato con la performance di WAN/LAN ipotizzata).

Quanto sono grossi questi campioni di dati per uso interattivo ? Se una n-tupla è un file tale da poter essere usato per analisi interattiva con programmi PAW-like, con turnaround di uno-due minuti massimo, deve essere di pochi GigaBytes, dato che l'I/O disco-CPU è comunque un bottleneck: anche con 50MByte/sec abbiamo circa 3 GB/min, ma per ora siamo fermi a 10-20 MB/sec. Quindi assumeremo **n-tupla $\simeq$ 1 GigaByte**. Tra due anni sarà facile avere su una workstation (PC) 100GB di disco, pienamente sufficiente per tutte le n-tuple di ogni singolo fisico.

Dove vengono prodotte queste n-tuple ? Tipicamente le n-tuple vengono create a partire da PADs per i quali abbiamo visto prima che si possono prevedere

da qualche centinaio di GB a diversi TB per data set, una tipica analisi può richiedere diversi data set, ed una sezione INFN ci possono essere più analisi in corso allo stesso tempo. Avere i PAD necessari "in casa" vuol dire grosso modo replicare in ogni sezione i 20 $\sim$ 30TB di disco del FCC, che non è sensato. Prevediamo quindi di **produrre le n-tuple al FCC, e copiarle sul PC in Italia**. Con la WAN ipotizzata, si può pensare facilmente che una persona impegnata nell'analisi abbia 1Mbit/sec a disposizione per il trasferimento dati, e quindi copiare la sua n-tupla a 0.5GB/ora. Dato che per produrre una n-tupla da un data set di diverse centinaia di GBytes occorreranno comunque ore, è perfettamente ragionevole che questa attività sia un job batch al FCC che al termine copia la n-tupla sul disco remoto in Italia. L'esperienza del Run-I ha insegnato che nuove n-tuple di un campione completo vengono in genere create non più spesso di una volta ogni qualche giorno, per cui anche poche ore di trasferimento dati non sono un problema. Quindi **n-tuple fino a pochi GBytes viaggiano sulla rete dal FCC all'Italia**. Spesso sarà però necessario trasferire campioni più grossi, decine di GB, se non centinaia. In questo caso è comunque accettabile avere un paio di giorni di latenza, e la soluzione più efficiente è semplicemente spedire i dati per aereo. Al FCC sarà disponibile una veloce "interfaccia con FederalExpress" che permetterà di ricevere anche una cassetta 8mm al giorno in una qualunque sezione INFN con pochi milioni l'anno: **n-tuple di molti GBytes viaggiano in aereo**.

Questo schema è ancora incompleto, è troppo limitante pensare di avere in Italia solo poche n-tuple e tenere tutti i dati al FCC. Alcuni data sets saranno spesso riutilizzati e sarà comunque più conveniente averli a disposizione in Italia, come anche ci saranno campioni di simulazione prodotti localmente che dovranno essere analizzati nelle sezioni. Da estrapolazioni dell'attività del Run-I prevediamo la necessità di avere a disposizione in Italia campioni di dati fino a pochi TeraByte. Allo stesso tempo notiamo che per campioni di tali dimensioni intermedie la possibilitá di percorrerli in tempi brevi con jobs di analisi porta naturalmente a richiedere una frequente revisione delle n-tuple prodotte, e quindi al desiderio di avere un accesso più veloce di quanto consentito dal link intercontinentale. Dovendo gestire alcuni TeraBytes di dati, non è pratico pensare a clusters di $\sim$ 10 PC, la scelta migliore sono un numero limitato (2 $\sim$ 3) di array RAID collegati ad un server Unix SMP con 4 CPU's. Questi servers possono essere costruiti attorno a campioni ortogonali di dati, usati per analisi diverse. L'esperienza ha mostrato una naturale tendenza delle sezioni a concentrarsi ognuna su un certo numero di argomenti di analisi, per cui è più conveniente delocalizzare questi sistemi uno presso ogni sezione anzichè costruire un mini centro di calcolo nazionale. Anche banali considerazioni di gestione spingono in questa direzione. Ovviamente storage di dimensioni considerevoli hanno senso solo per campioni condivisi da più persone, per cui tali servers saranno di preferenza situati nelle sezioni con i gruppi più consistenti, anche per limitare l'accesso remoto. Quindi **data sets di pochi TeraBytes risiedono in Italia su dischi RAID connessi a 4-cpu**

SMP servers.

E' nostro interesse tenere il numero di questi sistemi al minimo, circa 3, solo per i dati che assolutamente devono stare in Italia. Anche nell'ipotesi che la rete con gli USA funzioni come previsto, il gruppo italiano si troverà sempre in difficoltà nell'accesso ai dati al FCC rispetto ai gruppi americani che possono contare su connessioni di diverse centinaia di Mbit/s. Per questo ci converrà se possibile localizzare i nostri data servers proprio al FCC.

Anche l'esperienza del Run-I ci ha mostrato che è indispensabile avere adeguate risorse di calcolo a Fermilab con accesso privilegiato per il gruppo italiano, sia CPU, che spazio disco, non solo per quanto riguarda workstations desktop (PC) negli uffici, ma soprattutto code batch e dischi al FCC. Pertanto il terzo ingrediente di questo piano, dopo i PC personali ed i piccoli data servers, è un investimento adeguato in risorse di calcolo riservate al FCC. Possiamo immaginare come in passato di scaricare sul laboratorio il non trascurabile costo di gestione dell'hardware addizionale al FCC in cambio di una assegnazione delle risorse in parte al gruppo italiano ed in parte al pool comune, con un vantaggio reciproco, questa politica di contribuire alla facility di analisi di CDF al Fermilab in cambio delle disponibilità di risorse dedicate si è rivelata particolarmente fruttuosa.

Sostanzialmente il FCC è il posto dove mettiamo la nostra capcità di calcolo per data sets più grandi: **data sets di svariati TeraBytes risiedono al FCC a Fermilab, su dischi RAID connessi via FC a 8-cpu SMPservers**. Notiamo anche che massimizzando la quantità di hardware localizzato a Fermilab anziché in Italia, si possono realizzare notevoli risparmi, sia per la differente situazione di mercato e tassazione, sia perchè in un ambiente con un forte supporto tecnico come il laboratorio, è possibile orientarsi maggiormente verso soluzioni "fatte in casa" anziché preconfezionate da grossi venditori, ad esempio per i sistemi RAID si può spendere fino al 50% in meno. Questa strategia si integra molto bene con la architettura scalabile prevista per l'analysis facility di CDF: un cluster di unix/linux servers SMP (tipicamente 8-cpu) connessi ad un pool di array RAID di dischi. Di nuovo estrapolando dal passato e stimando sulla base del rapporto numerico tra fisici italiani e totali, possiamo prevedere ora tra 2 e quattro servers, e circa 10 TeraBytes di disco.

E' importante tener presente che queste risorse al FCC non sono una aggiunta a quanto previsto in Italia, ma sono risorse di calcolo che sarebbero comunque necessarie in Italia, localizzate al FCC per ottimizzarne il rendimento e realizzare un risparmio di costi. Noi non stiamo rinunciando ad avere computers e dischi con cui analizzare tutti i dati, abbiamo semplicemente concluso che la miglior locazione per alcuni di esse è al FCC dove possono essere integrati nel centro di calcolo di CDF ed avvantaggiarsi dell'accesso al robot ed al pool di dischi comune.

Ricordiamo di nuovo che tutto questo funziona solo se è possibile usare efficaciemente i mezzi di calcolo al FCC anche lavorando dall'Italia, ovvero avere un'ottima connettività per il lavoro interattivo ed una banda per il trasferimento dati che permetta di copiare dell'ordine di un GigaByte all'ora tra Fermilab e

l'Italia. In mancanza, ci sarebbero solo due alternative:
1) fare l'analisi a Fermilab, stando in missione molti mesi l'anno
2) avere in Italia almeno tutti o quasi i dati in formato PAD, ovvero dell'ordine di 100 TeraBytes (invece dei 5 circa previsti)
In entrambi i casi con costi molto elevati e scarsa efficienza.

Infine, sarà comunque necessario avere dei PC negli uffici a Fermilab per poter lavorare quando ci si trova là e per poter sfruttare la grossa capacità di calcolo costituita dall'insieme dei PC di tutte le collaborazioni ad esempio per Monte Carlo. Come in passato ci saranno centinaia di workstations (PC) negli uffici di CDF, in media raramente saturati come uso di CPU, ed utilissimi per analisi che non richiedono intenso accesso ai dati. Far parte di questo cluster è il prerequisito per poterlo sfruttare come polmone di espansione per momentanei picchi di lavoro. Se vogliamo essere brutali, un terminale su ogni scrivania è comunque doveroso, e acquistare qualcosa di meno "intelligente" di un PC di buone capacitá sarebbe stupido.

Riassumendo, la strategia per l'analisi così delineata é la seguente:

> *Un PC sulla scrivania di ogni fisico, con disco sufficiente (almeno 100GB) per le n-tuple e piccoli data sets usati frequentemente, ed una unità 8mm per spooling da/a nastro. Usare il FCC per creare questi campioni di uso quotidiano copiando i dati sulla rete o trasferendo cassette 8mm in aereo. Per i data sets di uso frequente da poche centinaia di GB fino a pochi TB, circa 3 server multi-cpu con arrays RAID localizzati nelle sezioni più grosse a cui tutti possano accedere. Per data sets ancora più grandi e per l'accesso alle risorse centrali dell'esperimento, alcuni servers multi-cpu e arrays RAID localizati al FCC. Infine PC sulle scrivanie degli uffici al Fermilab.*

La nostra stragegia con 3 livelli di storage dei dati (disco locale o comunque nella LAN, server all'interno di INFNET, FCC) è sufficientemente elastica e scalabile da non richiedere aggiustamenti in futuro, se non quantitativi.

Una differenza significativa si potrebbe avere, sia per la configurazione hardware, che per l'uso della rete, qualora diventassero disponibili applicazioni più evolute dei programmi di analisi finale, ovvero un equivalente di PAW/ROOT in cui la parte che accede i dati su disco gira remotamente e solo l'istogramma (non il display) è spedito sulla rete ad un browser locale. Una applicazione del genere, magari con implementazione di tecniche di qualità di servizio sulla rete, potrebbe ridurre drasticamente le spese necessarie. Purtroppo un tool del genere non è ora in sviluppo né in CDF né in altre collaborazione ed il gruppo CDF-Italia non ha né il man-power né la competenza per svilupparlo.

Tutti i numeri citati finora per la quantità di spazio disco, cpu etc., sono "a regime" quando tutto il data sample (1PB) sarà stato raccolto. Ovviamente l'hardware sarà acquistato poco alla volta via via che i dati crescono, con un piano pluriennale. Il dettaglio di questo piano è oggetto delle prossime sezioni.

6 l'analisi in Italia

A regime: un PC con 100GB disco per ogni fisico. LAN a 100 Mbits/sec. Pochi (3+-1) servers con 1 $\sim$ 2TB disco per data set comuni, MC etc.

Evoluzione possibile:
2000: 30 PC, 50 GByte disco ognuno.
2001: altri 10 PC, altri 40 dischi IDE, 3 servers con 1TB disco(PD, BO, PI e.g.)
2002: altri dischi e manut/upgrade PC a 2 MLit l'uno, un'altro sistema RAID per ogni server.
2003: altri 50 dischi, altri 10 PC. manut/upgrade servers a 20 MLit l'uno

Bisogna sommare a tutto il 10 per cento annuo del costo di acquisto in modo da costituire un fondo di manutenzione. Per PC e simili l'unica manutenzione è l'acquisto di pezzi nuovi in caso di rottura.

7 l'analisi a fermilab

A regime: PC nei trailers per tutti (15 uffici = 30 PC?), 3 servers nel FCC (in aggiunta a circa 10 previsti dal piano del laboratorio per uso generale), 10 TeraBytes di disco al FCC "a regime" (in aggiunta ai circa 30 previsti per uso generale).

Evoluzione possibile:
2000: 20 PC nei trailers
2001: altri dischi (30 a 200 dollari l'uno), un server nel FCC
2002: altri PC nei trailers/upgrade degli esistenti, circa 1000 dollari per PC, un altro server nel FCC, dischi nel FCC (2TB)
2003: altri dischi (8TB) e altri 2 servers nel FCC

Per la manutenzione facciamo come in Italia assumento un 10%all'anno per rimpiazzare la roba rotta, e speriamo che Fnal si prenda il carico della manutenzione ordinaria e del software.

8 Configurazione e costo dei sistemi considerati

Alcune regole base seguite per identificare l'hardware da acquistare e stimarne i costi:

il PC del 2000: 256MB RAM, CPU uno step sotto il massimo disponibile (in genere si spende la metà e si perde si e no il 20%in performance), 100 GB di dischi EIDE/ATA anziche' SCSI, per massimizzare lo spazio disco a parità di costo anche a patto di perdere qualcosina in performance (ma gli EIDE stanno diventando molto veloci...), adattatore SCSI per l'unità 8mm. Fast

Ethernet a 100Mbit/sec. In previsione di riuscire davvero a migliorare la collaborazione a distanza, dato anche il rapido sviluppo di tools software per questo scopo, penso si debba assolutamente prepararsi per videoconf, costa solo 500MLit. CD riscrivibile (SCSI per lasciare 4 slot IDE ai dischi) per backups, sulla scala di tempi del Run-II saranno disponibili a poco costo DVD riscrivibili da almeno 5 GB. Monitor 19" di buona qualità. Scheda video ed audio "quanto basta".

4-cpu servers Meglio limitarsi a 4 CPU, altrimenti i prezzi esplodono, comunque sufficienti per sostenere una diecina di jobs di analisi. Almeno 512MB RAM per CPU, poco disco locale interno, Ultra2 SCSI per dischi esterni (RAID) per avere circa 1 TeraByte nel 2000. Per questi sistemi vogliamo prodotti solidi ed affidabili, comprati in configurazione completa da un venditore solo. Non abbiamo la forza di system management nè il supporto dai produttori di un laboratorio come FNAL o CERN, per cui bisogna assolutamente evitare il pericolo di dover risolvere da soli problemi di compatibilità tra hardware o software di venditori diversi che si rimpallano le responsabilità. Pertanto dovremo probabilmente escludere sistemi linux più economici e comprare SUN o SGI. Come connessione alla LAN meglio avere direttamente Gigabit Ethernet. Tipicamente 4 unità 8mm per ogni server. Il numero e collocazione dei servers sono da definire anche in base a chi si offrirà volontario per la gestione, una possibiltà è Pisa e Padova perché sono i gruppi più grossi e Bologna perché ha una ottima connessione ad INFNET.

RAID preferiamo andare verso array RAID anzichè i vecchi "mucchi" di dischi in scatole singole per la maggior efficienza di gestione, le performace molto migliori, ed appunto la possibilità di sfruttare la ridondandza per gestire eventuali rotture. Con questa tecnologia si apre anche la possibiltà di avere i più economici dischi EIDE montati in un sistema SCSI e quindi avere performance notevoli a costi ridotti. Questi sistemi misti cominciano appena ora ad essere disponibili, e per adesso possiamo fare stime di prezzi realistiche solo per sistemi SCSI-SCSI. Notiamo che al momento attuale un raid SCSI può contenere al massimo 8 dischi da 18.8 GByte, quindi anche assumento da qui al 2001 un fattore quattro di crescita dei dischi, avremo circa 0.5 TeraByte per ogni sistema.

nastri 8mm un centinaio di cassette 8mm per sezione a regime, un numero maggiore dove ci sono i piccoli data servers (non di più se no il costo è eccessivo, i nastri spediti dagli USA vanno rispediti indietro e riciclati). Stimiamo i costi di spedizione cassette assumendo come unità una spedizione al giorno per ogni giorno feriale dell'anno, una spedizione (0.5 libbre) può contenere 2 cassette 8mm.

Per stimare i costi di hardware che verrà acquistato tra diversi anni, abbiamo assunto che vorremo comunque acquistare top of line CPU, che ci sarà uno incremento di un fattore quattro nei dischi a parità di costo, ed un aumento di un fattore due o tre nella RAM a parità di costo, tutto tra oggi e due anni da ora. Pertanto usiamo un sistema configurato oggi con un quarto dello spazio disco e circa metà della RAM richiesta a regime per avere la stima dei costi finali. Questo permette di avere costi veri basati su offerte di venditori di hardware. I costi sono nella prima pagina dell'appendice. A questo punto sono TUTTI costi veri, ottenuti da offerte di venditori al marzo 1999. Se vogliamo cambiare il totale, bisogna modificare la configurazione.

9 piano finanziario 2000-2003

Le pagine due e tre dell'appendice sono un possibile esercizio di piano finanziario. Non e' facile fare stime accurate, l'analisi si può fare con tanti mezzi di calcolo, come con pochi, e' difficile dire se 1TB e' un numero magico, se ne servono 2 o tre... Ho cercato sostanzialmente di fare una stima approssimata per eccesso, un pò un limite superiore approssimato, vediamo ora con l'aiuto di tutti di ridurlo a numeri veramente difendibili. Notate che la divisione tra anni è molto indicativa, l'importante è arrivare ad una configurazione a regime sensata con un profilo di spesa ragionevole. Solo per semplicità mia ho alternato un anno di acquisto di PC in Italia ed uno a Fermilab. Infine per la divisione tra sezioni, è ovvio che ho tirato ad indovinare solo per vedere se le cose fanno senso. Spetta ad ognuno farsi i suoi conti per la sua sezione e difendere le proprie richieste, ovviamente c'è una qualche regola di somma per cui il totale di PC a Fermilab non può superare il numero di tavoli (=uffici*2) e quello in Italia non supera il numero di persone attive in CDF...

References

[1] S.Belforte, talk all'Incontro sulle prospettive di calcolo a Pisa, 12/5/1998. http://www.pi.infn.it/ccl/980512/cdf

[2] S.Belforte, talk agli Incontri di Infnet a Bologna, 19/1/999. http://www.cnaf.infn.it/INFNet/cdf/sld001.htm oppure http://www.ts.infn.it/ belforte/presentations/infnet_19jan99/talk_4upbw.ps.gz

[3] S.Belforte, talk al Cdf Computing Workshop del 10/2/1999. http://www.ts.infn.it/ belforte/presentations/ccw_10feb99/handouts.ps.gz

[4] S.Lammel, talk al Cdf Computing Workshop del 10/2/1999. http://www-cdf.fnal.gov/upgrades/cdfdh/talks/19990210_stephan.ps

[5] L.Buckley et al. CDF Run II Data Handling (Data Volume Estimates and Equipment Requirements). CDF note 4072, 1977. http://www-cdf.fnal.gov/upgrades/cdfdh/doc/r2dh/r2dh_dvol.ps

[6] S.Lammel et al. CDF Run II Data Handling Central Analysis System Hardware Architecture. CDF note 4707, 23/1/1999. http://www-cdf.fnal.gov/upgrades/cdfdh/doc/hardware/hard_arch.ps

[7] Serial Media Working Group First Run II Report. CDF note 4560. http://www-cdf.fnal.gov/upgrades/cdfdh/doc/hardware/smwg_public.ps

[8] L. Buckley, Run II Data access overview. Talk al CDF computing review del 26/1/1999. Pag. 5. http://www-cdf.fnal.gov/upgrades/computing/offline_minutes/buckley_vr_talk_jan_99.ps

[9] Costo attuale del link DANTE tra Europa e US. A. Ghiselli, private communication.

10 Appendice: 3 pagine da Excel coi dettagli finanaziari

In attesa di un formato migliore seguono 3 pagine postscritp con uno spreadsheet Excel che contiene tutti i conti per un possibile piano finanziario. Queste pagine non appaiono in ghostview, ma vengono stampate quando inviate questo documento in stampa.

Piano finanziaro per il calcolo di CDF per il RunII

Premessa: costi unitari (MLit). IVA inclusa in Italia. Net price in US.

PC - Italia	7	400Mhz PII, 25 GB IDE disks, 128 MB RAM, SCSI adaptor, 100 Mbit ethernet, R/W CD, 16-bit audio board, microfone, casse, 4MB video board, 19" color monitor, scheda acquisizione video, videocamera, 3 anni garanzia
drive 8mm - Italia	5	8mm drive scelto da FNAL/CDF (SONY AIT o Exabyte Mammoth)
PC con 8mm drive - Italia	**12**	
PC - FNAL	6	stessa configurazione che in Italia
drive 8mm - FNAL	4	stesso che Italia. 2000-2500$
PC con 8mm drive - FNAL	**10**	
dischi addizionali - Italia	**0.8**	top of line IDE disk
dischi addizionali - FNAL	**0.6**	top of line IDE disk
UNIX server - Italia	100	4-CPU SMP server, SUN or SGI, 512 MB RAM per CPU, 25 GB local disk, four 100Mbit ethernet ports or Gigabit ethernet, SCSI controllers Ultra2 per dischi e normale per 8mm.
UNIX server + 4 8mm - Italia	**120**	
linux server - Italia (**STIMA !**)	60	configurato come sopra, ma basato su chips Intel a partire da un server NT a 4 CPU
linux server + 4 8mm - Italia	**80**	
UNIX/linux server - FNAL	50	configurato come in Italia, meno costoso. FNAL estimate =20~25 K$
UNIX/linux server + 4 8mm - FNAL	**66**	
SCSI-RAID disks - Italia	**53**	125 GB SCSI RAID array con 8 top of line SCSI disks. 0.5 TeraByte alla fine del 2000, raddoppio ogni 1.5 anni
FC-RAID disks - FNAL	**30**	125 GB FiberChannel RAID array con 8 top of line SCSI disks. 0.5 TeraByte alla fine del 2000, x2 ogni 1.5 anni. FNAL est=15K$ (home made)
nastri 8mm (l'uno)	**0.15**	Fnal estimate = 75$ / tape
spediz. nastri (1 anno)	**13**	0.5lb FedExp US to Italy @ 25.5$/day * 5days/week * 52weeks

Esercizio di un possibile scenario, l'incertezza sul dettaglio aumenta cogli anni										
2000	**BO**	**LNF**	**PD**	**PI**	**PV**	**RM**	**UD**	**TS**	**TOT**	**Mlit**
PC-Italia	4	4	7	7	1	4	2	1	30	360
PC-FNAL	3	3	4	4	1	2	2	1	20	200
dischi PC Italia									0	0
dischi PC FNAL									0	0
server Italia									0	0
server FNAL									0	0
RAID Italia									0	0
RAID FNAL									0	0
nastri 8mm	30	30	30	30	30	30	30	30	240	36
spedizione nastri (anni)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	11
TOT 2000										**607**
2001										
PC-Italia	1	1	2	3	1	1		1	10	120
PC-FNAL									0	0
dischi PC Italia	5	5	9	10	2	5	2	2	40	32
dischi PC FNAL									0	0
server Italia	1		1	1					3	360
server FNAL									1	66
RAID Italia	1		1	1					3	159
RAID FNAL									2	60
nastri 8mm	50	30	50	50	20	30	25	25	280	42
spedizione nastri (anni)	0.5	0.4	0.5	0.5	0.2	0.4	0.2	0.2	2.9	38
manutenzioni										60
TOT 2001										**937**
2002										
PC-Italia									0	0
PC-FNAL	2	1	2	3		1		1	10	100
dischi PC Italia	5	5	9	10	3	4	2	2	40	32
dischi PC FNAL	5	4	6	8	1	3	2	2	31	19
server Italia									0	0
server FNAL									1	66
RAID Italia	1		1	1					3	159
RAID FNAL									4	120
nastri 8mm	50	30	50	50	20	30	20	20	270	41
spedizione nastri (anni)	1	0.5	1	1	0.4	0.5	0.5	0.5	5.4	71
manutenzioni										144
TOT 2002										**752**
2003										
PC-Italia	1	1	2	2	1	1	1	1	10	120
PC-FNAL									0	0
dischi PC Italia	8	8	12	15	3	6	4	4	60	48
dischi PC FNAL	5	4	6	8	1	3	2	2	31	19
server Italia									0	0
server FNAL									2	132
RAID Italia	1		1	1					3	159
RAID FNAL									6	180
nastri 8mm	50	20	50	50	20	30	25	25	270	41
spedizione nastri (anni)	0.5	0.2	0.5	0.5	0.2	0.4	0.2	0.2	2.7	36
manutenzioni										198
TOT 2003										**933**
GRAND TOT										**3229**

situazione a regime alla fine del 2003

	BO	LNF	PD	PI	PV	RM	UD	TS	TOT
desktops									
PC Italia	6	6	11	12	3	6	3	3	**50**
dischi acquistati per PC	5	5	4.7	4.9	4.7	4.5	4.7	4.7	4.8
PC FNAL	5	4	6	7	1	3	2	2	**30**
dischi acquistati per PC	4	4	4	4.29	4	4	4	4	4.07
Computing Center									
SMP servers Italia	1	0	1	1	0	0	0	0	**3**
con RAID arrays totali	3	0	3	3	0	0	0	0	9
SMP servers FNAL									**4**
con RAID arrays totali									12
nastri 8mm	180	110	180	180	90	120	100	100	**1072**

breackout della spesa totale per grossi capitoli ovvero piano finanziaro globale

	ITALIA			**FERMILAB**						**GRAN**
	desk	Cmp.Cntr.		desk	Cmp.Cntr.		mant	tapes	trasp	**TOT**
	top	CPU	disk	top	CPU	disk				
2000	360	0	0	200	0	0		36	11	607
2001	152	360	159	0	66	60	60	42	38	937
2002	32	0	159	119	66	120	144	41	71	752
2003	168	0	159	19	132	180	198	41	36	933
totali parziali		360	477		264	360				
totale	712	837		338	624		402	160	156	**3229**
totali Italia/USA/misc.	1549			962			718			

divisione per capitoli	**2000**	**2001**	**2002**	**2003**	**TOT**
invent. (hw Ita.+mant)	360	731	335	525	1951
inv. / c.a. (hw @ FNAL)	200	126	305	331	962
consumi (nastri 8mm)	36	42	41	41	160
trasporti (spediz. nastri)	11	38	71	36	156
TOTAL	607	937	752	933	**3229**

se il link internet con fnal non e' sufficiente a copiare le ntuple, serve una quantita' maggiore di PADs in Italia, quindi 5 servers invece di 3 e 4TB disco su ognuno invece di 1.5, in totale 5x8=40 RAID arrays invece di 3x3=9. Per un costo aggiuntivo di circa 2 mliardi.

	no.	Mlit
additional servers	2	240
additional RAID arrays	31	1643
TOTALE		**1883**