### 

Alla Commissione Scientifica Nazionale 1 del INFN

Agosto 2016.

Oggetto: richiesta di finanziamento di un progetto di data preservation dei dati raccolti da CDF durante il Run1.

La conservazione a lungo termine dei dati scientifici e la loro condivisione sono ormai parte integrante di qualsiasi esperimento scientifico. Conservare i dati e renderli fruibili non solo agli scienziati ma a tutta la società, per esempio a scopo educativo, è ormai richiesta obbligata da parte degli enti finanziatori.

Nella fisica delle alte energie il tema è però particolarmente impegnativo, data la grande mole di dati da conservare e il fatto che ogni esperimento possiede un software specifico per la loro analisi. Nel caso di esperimenti che hanno terminato da tempo la presa dati, le difficoltà sono anche maggiori per la necessità di preservare software non più supportato e per il rischio di perdere informazioni sui dati con il restringersi delle collaborazioni.

E’ quindi importante implementare progetti di conservazione dei dati a lungo termine quando la collaborazione è ancora viva ed è più facile reperire informazioni relative ai dati e al codice dell’esperimento.

L’INFN in collaborazione con CDF e Fermilab ha implementato con successo un progetto di conservazione a lungo termine dei dati del Run2 di CDF (2002-2011). Più di 4 TB di dati grezzi e ntuple (circa 9.6 fb-1 di luminosità integrata alla energia di 1.96 TeV nel centro di massa) sono stati copiati da Fermilab al centro di calcolo CNAF dell’INFN e possono essere ora là analizzati.

In questo documento presentiamo un nuovo progetto per conservare i dati raccolti durante il Run1 di CDF (1992-1995). Si tratta di circa circa 110 pb-1 di luminosità integrata a 1,8 TeV nel centro di massa, più circa 4 pb-1 a 630 GeV.

Dopo aver discusso le motivazioni scientifiche alla base del progetto, verranno presentate nel dettaglio le attività previste con una stima dei costi, ottenute sulla base dei test di fattibilità eseguiti al CNAF.

### Motivazione scientifica

I dati del Run1 sono di importanza storica oltre che scientifica perchè contengono gli eventi che hanno portato alla scoperta del quark top.

Si tratta di un campione unico, a disposizione della sola collaborazione CDF, dato che a suo tempo la collaborazione D0 decise di di non conservare i suoi dati del Run1.

Inoltre, come per i dati del Run2, si tratta del risultato di collisioni protone-antiprotone: ad oggi non è prevista la costruzione di un altro collisionatore protone-antiprotone, quindi i dati del Tevatron sono unici anche per stato iniziale, oltre che per energia di produzione.

Una parte della informazione contenuta nei dati del Run1 di CDF non è stata studiata a fondo a suo tempo perchè la collaborazione era tesa a nuove scoperte e pressata dall’urgenza delle modifiche da apportare al rivelatore in vista del Run2. Questo va tenuto presente perchè alcune misure interessanti potrebbero essere riprese in futuro, come ad esempio i processi inclusivi e semi-inclusivi di QCD e di fisica EW, i cui rapporti relativi e dipendenza dalla energia, che possono essere misurati più rapidamente, rivestono rilevante interesse per la teoria. Un canale particolare poco studiato è stato la produzione centrale in rapidità di stati adronici in eventi di “doppio Pomeron”. Inoltre, sarebbe interessante misurare la struttura dell’ “underlying event” in eventi di W, Z a 630 GeV, al fine di disporre di un ampio “braccio di leva” per calibrare i modelli che descrivono l’andamento di questo fondo in funzione dell’energia, da 630 GeV fino all`energia di LHC.

Infine non si deve sottovalutare l’utilizzo dei dati a scopo educativo: la individuazione dei rari eventi di top nei dati del Run1 sarebbe una straordinaria palestra per studenti di laurea in fisica delle alte energie, che si avviano alla ricerca di fenomeni complessi e rari di nuova fisica al LHC. L’utilizzazione a fini didattici dei dati di LHC per la ricerca del bosone di Higgs è gia` da tempo iniziata al CERN[[1]](#footnote-0). Sarebbe motivo di orgoglio per l’INFN e la collaborazine CDF, e anche un meritato riconoscimento del valore della scoperta, se anche ritrovare i rari eventi di coppie di quark top nel Run1 di CDF fosse una palestra di avvio alla ricerca dei futuri studenti in HEP.

Descrizione del progetto

#### I circa 40 TB di dati del Run1 (calibrazioni incluse) sono contenuti in circa 4.000 nastri.

Il software usato per la produzione di eventi simulati e per l’analisi dei dati durante il Run1 è scritto in Fortran e attualmente conservato in diverse copie. Il codice è completo, comprendendo anche database, tabelle di trigger ecc.

Per i test di fattibilità descritti nel seguito, i dati del Run1 sono stati riprocessati con il codice del Run2. Il codice del Run2, per il quale CDF ha ora una release su sistema operativo SL6 destinata ad essere preservata nel futuro a lungo termine, può essere quindi utilizzato per lavorare con i dati del Run1.

Analogamente, per la produzione di eventi simulati è in fase di studio la possibilità di usare il codice del Run2, adattandolo al Run1.

#### Test di fattibilità

#### Risorse necessarie

I circa 40 TB di dati del Run1 (calibrazioni incluse) sono contenuti in circa 4.000 nastri. Questi possono essere trascritti su tre T10.000 da 8.5 TB ciascuno. Il costo totale e` di circa Euro 650, IVA inclusa. Volendo farne per sicurezza doppia copia e aggiungendo un minimo di contingenza, si arriva ad un costo totale di Euro 1.500.

Il CNAF ha ricevuto una offerta di Euro 2.147 (IVA inclusa) per due mammout autoloader da 7 nastri ciascuno, rigenerati ma garantiti per un anno. Ne servirebbero 4 per tenerne almeno tre in produzione continua. Aggiungendo qualche cassetta per pulizia ed un minimo di contingenza il costo e` di circa Euro 6.000.

Fermilab ha accettato di occuparsi della spedizione dei nastri al CNAF (via nave) e di pagarne il costo. Pertanto, il finanziamento richiesto alla CSN1 e` di Euro 7500, da destinarsi al CNAF e da rendere disponibili appena possibile.

Le prove fatte hanno mostrato che ogni cassetta viene letta in un tempo variabile fra 3 e 4 ore, per cui una carica di 7 cassette in un autodrive automatizzato verrebbe smaltita in 24 ore. Con tre autodrives initerrottamente al lavoro si leggerebbero 21 cassette al giorno. Assumendo una efficenza del 50% per tener conto delle perdite di tempo per pulizia, guasti, riparazioni, si può stimare che il lavoro sarebbe compiuto in circa 400 giorni lavorativi, ossia in meno di due anni solari. Se fosse possibile iniziare il lavoro quest`anno esso dovrebbe essere finito entro il 2018.

David Toback ed io stesso, attuali spokespersons di CDF, siamo estremamente interessati alla felice conclusione di questo progetto e saremo grati al INFN ed al CNAF per il loro supporto.

Giorgiob, Silvia A

1. si veda ad esempio<https://indico.cern.ch/event/277650/session/3/contribution/298/attachments/505999/698584/INCNF_2014_Kourkoumelis.pdf>). [↑](#footnote-ref-0)