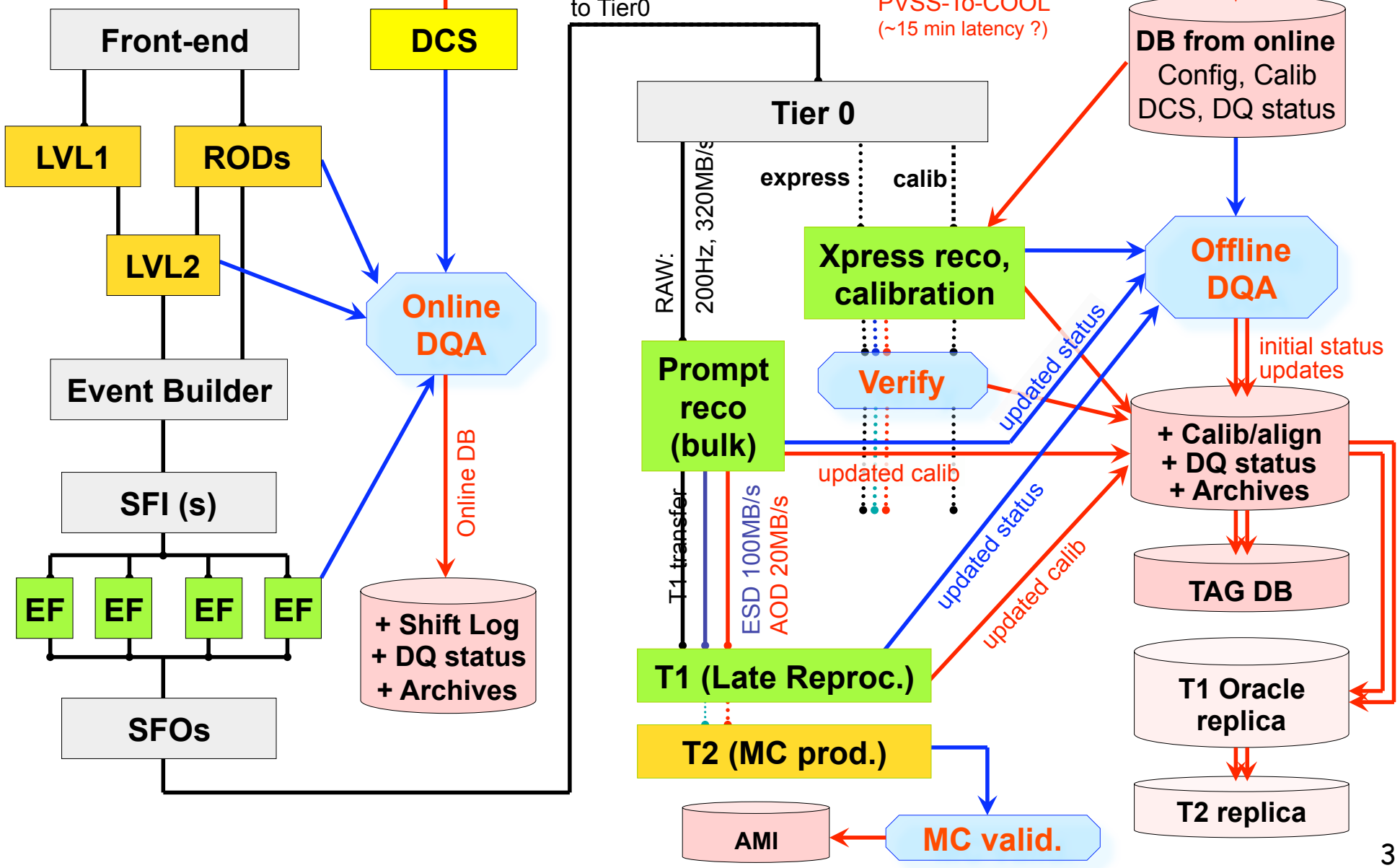
ATLAS
Modello di calcolo per l'analisi

Dario Barberis
Università e INFN Genova
(con il contributo sostanziale di molti altri)

ATLAS event data flow from online to offline

- Events are written in "ByteStream" format by the Event Filter farm in <=2 GB files
  - ~200 Hz trigger rate (independent of luminosity)
  - Events are grouped by "luminosity block" (1-2 minute intervals)
    - One luminosity block can be approximated as having constant luminosity
    - There should be enough information for each luminosity block to be able to calculate the luminosity
  - Average RAW event size is 1.6 MB/event
  - Currently several streams are foreseen (slightly different for early commissioning data):
    - ~5 physics event streams, separated by main trigger signature
      - e.g. muons, electromagnetic, hadronic jets, taus, minimum bias
    - Express stream with "most interesting" events to be processed immediately
      - Initially useful only for monitoring and calibration activities
    - Calibration events
    - "Trouble maker" events (for debugging)
  - Each file contains events belonging to the same trigger stream, luminosity block and SFO (Event Filter Sub-Farm Output unit)
    - ~25 files/minute are produced by the online system
  - Data are transferred to the Tier-0 input buffer at 320 MB/s (average)
    - Files from the same trigger stream and luminosity block are merged
    - Datasets are formed with all files belonging to the same run and trigger stream

Data flow from DAQ to offline



Distribuzione dei dati per l'analisi

- I dataset prodotti al Tier-0 o ai Tier-1 dalle campagne di reprocessing vengono distribuiti come segue:
  - RAW: una copia su nastro al CERN, una copia su nastro fra tutti i Tier-1, dati più recenti su disco ai Tier-1
  - ESD: una copia su nastro al sito di produzione, 2 copie su disco fra tutti i Tier-1 (più una copia completa a BNL su disco non-pledged)
  - AOD e dESD (skimmed ESD): una copia su nastro al sito di produzione, 2 copie su disco fra tutti i Tier-1 e 10 copie fra tutti i Tier-2
  - Copie aggiuntive (secondarie) vengono distribuite finché c'è spazio disco disponibile e rimosse in seguito per far posto a dati nuovi
- Il modello di analisi prevede che:
  - I job di analisi di gruppo (che di solito sono selezioni su grandi quantità di eventi) vengano eseguiti ai Tier-1 e i loro output vengano salvati nelle aree GROUPDISK allocate ad ogni gruppo
  - I job di analisi vengano eseguiti ai Tier-2e i loro output vengano salvati nelle aree LOCALGROUPDISK allocate in ogni sito ("home dir" di ogni utente Grid)
  - Per l'analisi interattiva finale, i dati necessari (tipicamente ntuple) devono essere copiati alla struttura locale (Tier-3) ed analizzati con Root o Proof

Tipologia di lavoro di analisi

- La catena completa di un'analisi di fisica in generale comprende:
  a) la simulazione di piccoli campioni di eventi di segnale e di fondo per il processo che si vuole studiare
  b) la generazione in grande scala di eventi simulati
  c) la selezione di eventi reali che potenzialmente corrispondono al processo in questione
  d) l'analisi finale degli eventi selezionati
- Le attività a) e d) vengono tipicamente svolte localmente (Tier-3), mentre le attività b) e c) sono naturalmente svolte sulla Grid (Tier-2 e, per le simulazioni, anche Tier-1).
  - L'attività a) necessita un'installazione completa di Athena, similmente allo sviluppo del software
  - L'attività d) invece necessita solamente l'accesso ai dati selezionati (normalmente ntuple che devono venire trasferite nel Tier-3 per essere analizzati) e un'installazione locale efficiente per l'analisi (di solito basata su Proof).
- Possiamo quindi considerare solo due gruppi di attività:
  - Attività basate su Athena che necessitano un'installazione completa del software di ATLAS e accesso al database e ai file di calibrazione;
  - Attività basate su Proof che necessitano di un accesso veloce ai dati per l'analisi.

ATLAS-Italia Tier-3 Task Force

- Definire un modello italiano per i Tier-3 di ATLAS
  - Funzionalità
  - Dimensione
  - Esigenze operative
- Complementare ed estendere il lavoro del gruppo di lavoro sui Tier-3 recentemente formato per tutto ATLAS
  - Guidato da M. Lamanna (CERN)
  - Con contributi importanti dalla comunità italiana
- La task force ha analizzato in dettaglio i seguenti aspetti dei Tier-3:
  - Tipologia del lavoro svolto localmente
  - Risorse necessarie in relazione al numero di persone attive
  - Configurazione dei siti
    - Tier-3 co-locato con un Tier-2
    - Tier-3 integrato in INFN-Grid
    - Tier-3 per solo accesso locale
- Rapporto preliminare: inizio maggio
- Rapporto finale: fine giugno

Risorse di calcolo (ordine di grandezza)

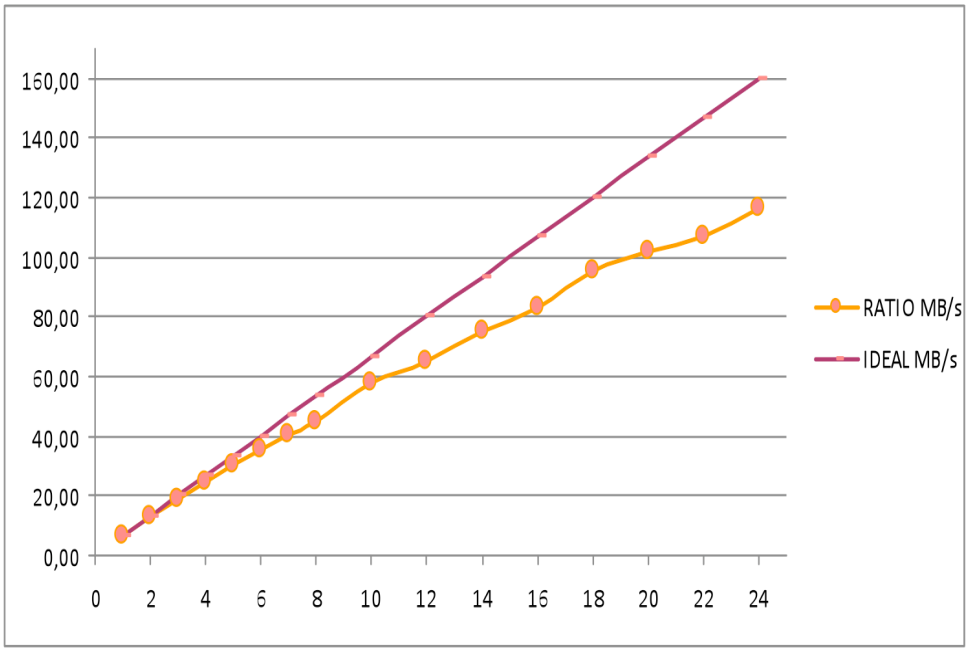
- Sviluppo e test di codice in ambiente Athena, per ogni utente attivo:
  - 2 TB di spazio disco per files di test
  - 1 macchina (8 cores) in media in un sistema condiviso
- Analisi finale con Root/Proof, per ogni utente attivo:
  - 2 TB di spazio disco per ntuple
  - Almeno 1 macchina (8 cores) in media in un sistema condiviso
- Maggiori informazioni nel rapporto finale
- Una componente estremamente importante di un Tier-3 è la rete interna, che definisce la banda passante fra i dischi e le CPU.
  - Una macchina con 8 cores che gira Proof può assorbire fino a circa 100 MB/s di dati, che saturano la porta Ethernet da 1 Gb/s.
  - Chiaramente lo storage deve poter fornire dati a molto più di una singola macchina, per cui l'infrastrutture di rete interna deve essere basata su link a 10 Gb/s ed avere una topologia ben progettata per eliminare eventuali colli di bottiglia
    - con dati dello stesso tipo distribuiti su parecchi disk server

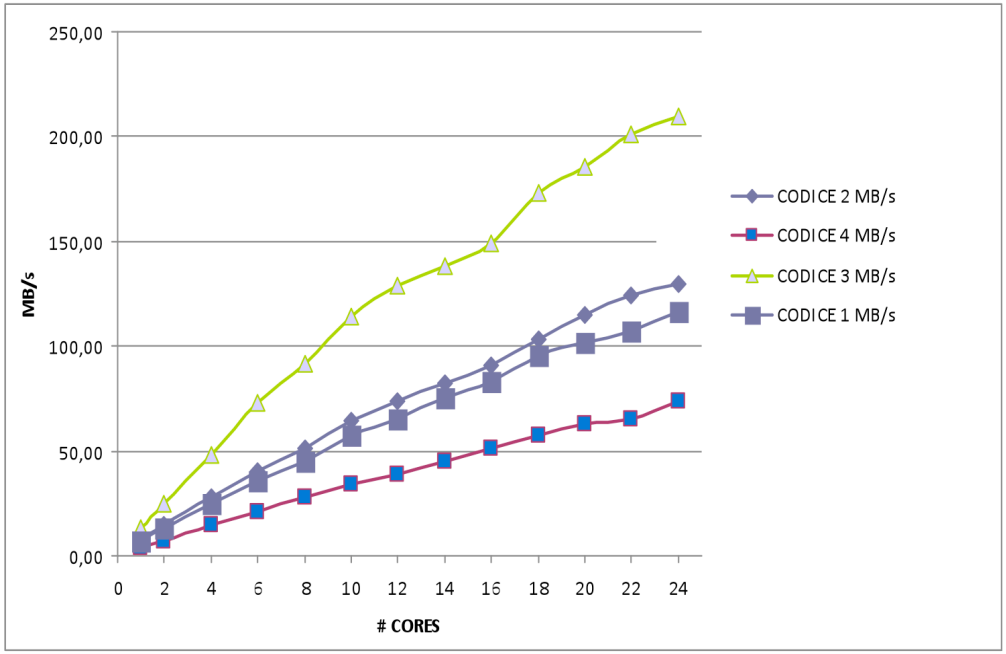
Siti che supportano Athena

- Questa slide si riferisce sia ai Tier-2 che ai Tier-3 Grid-enabled
  - Storage con GPFS/StoRM (consigliato per i Tier-3): MI, RM3, UD/TS, GE
  - Storage con DPM (solo Tier-2): RM1, NA, LNF
- Software di ATLAS distribuito automaticamenta da RM1 con sw-mgr
- Storage configurato con 3 "space tokens" (storage con catalogo centrale):
  - SCRATCHDISK per l'output dei job Grid e per file transfer
  - HOTDISK per i files di calibrazione e DB release
  - LOCALGROUPDISK per dataset di test e per uso locale
- Spazio disco per gli utenti locali su file system ad accesso diretto
  - I siti con GPFS possono avere un sistema integrato fra lo storage Grid e quello locale
    - Vantaggio notevole per system admins e utenti
- Descrizione dettagliata della configurazione hardware installata a Genova nel rapporto preliminare
  - I test stanno iniziando questa settimana, coordinati con RM3 e UD/TS

Siti che supportano Proof (1)

- In principio tutti i Tier-3 supportano l'analisi finale e perciò Proof
- PROOF è pensato per agire su ntuple di Root
  - l'utente le ha prodotte con i job di analisi distribuita all'interno di Athena girando sui dati (in formato AOD o ESD generalmente) nei vari Tier-2 e le ha poi copiate nello Storage Element locale usando i vari tools di GRID
- Test sono stati fatti a Milano con contributi importanti da UD/TS e PI, utilizzando 4 tipi diversi di codice:





Siti che supportano Proof (2)

- Conclusioni preliminari:
  1) L'uso di PROOF Lite risulta scalare bene (entro un 10% dall'idealità) con il numero dei cores fino a circa 5-6 cores, ma poi ci si discosta sempre più dall'idealità, arrivando a un fattore oscillante tra 15.8 e 19.8 per i diversi tipi di codice testati nella versione a 24 cores.
  2) Con una macro di analisi di media complessità, la velocità di lettura dallo storage non risulta essere limitata né dalla velocità intrinseca del disco né da quella dell'unico switch di rete presente nella configurazione utilizzata. Inoltre, la parallelizzazione dell'analisi realizzata da PROOF può garantire una velocità di lettura nettamente superiore a quella che garantirebbe lo switch della rete nella versione "single core".
  3) La scalabilità di PROOF migliora con la complessità del codice di analisi (e quindi con il consumo di CPU).
  4) Dal confronto tra i 4 tipi di codice emerge come gran parte della differenza tra una macro complessa e una iper-semplice sia dovuta alla lettura e all'allocazione in memoria dei vari branches dell'ntupla piuttosto che alla complessità del calcolo richiesto dalla macro. E' pertanto opportuno, soprattutto con ntuple molto grandi, leggere solamente i branches effettivamente utilizzati nel codice di analisi.

Siti integrati Athena/Proof

- L'integrazione fra parte interattiva e parte batch/Grid in un sito è molto più agevole se lo storage è basato su di un file system con protocollo Posix (come GPFS o Lustre) piuttosto che su uno storage element come DPM.
- In questo campo, un lavoro interessante e potenzialmente molto utile ancora da fare è lo studio del cosiddetto "proof on demand".
  - L'idea base è di poter fornire proof slaves sui worker nodes del sistema batch, lanciando un job batch che configura il worker node su cui gira come un proof slave, si connette al proof master (che invece sarà una delle macchine interattive dedicate) e contribuisce al carico di lavoro di Proof. Quando il lavoro è terminato, il job finisce e il worker node ridiviene disponibile per un altro job batch.
- La gestione di un sistema come quello qui descritto è tutt'altro che ovvia.
  - Per cui una fase estesa di test di configurazioni e performance deve essere prevista a breve termine.

Lavoro ancora da fare

- Test di funzionalità e performance dei Tier-3 "Grid-enabled":
  - Definizione di dataset e configurazioni per i test comuni
  - Sottomissione di job Athena in modo batch locale (test di funzionalità)
  - Sottomissione di job Athena in modo Grid (test di funzionalità)
  - Sottomissione di test automatizzati HammerCloud (test di throughput fra storage e CPU farm)
- Test di sistema integrato Grid/locale in tutti i siti
  - Performance di Proof con dati in input nello storage element
    - Particolarmente per i siti con DPM
- Definizione del modello di supporto per siti e utenti
- Più a lungo termine:
  - Proof on demand