R&D su modello di analisi a CMS per Run3 e HL-LHC

D. Ciangottini, P. Lenzi, D. Spiga, T. Tedeschi, M. Tracolli

ciangottini@pg.infn.it

Outline

- Analisi CMS oggi
- Criticità per Run3 e HL-LHC
- Analysis Tools Task Force
- Motivazioni per R&D @INFN
- Stato dell'attività
- Descrizione ad alto livello dell'infrastruttura
- Considerazioni e piani

Questo lavoro e' il risultato della collaborazione di diversi profili

Integrazione
- Diego Ciangottini
- Mirco Tracolli

Analisi
- Tommaso Tedeschi
- Piergiulio Lenzi

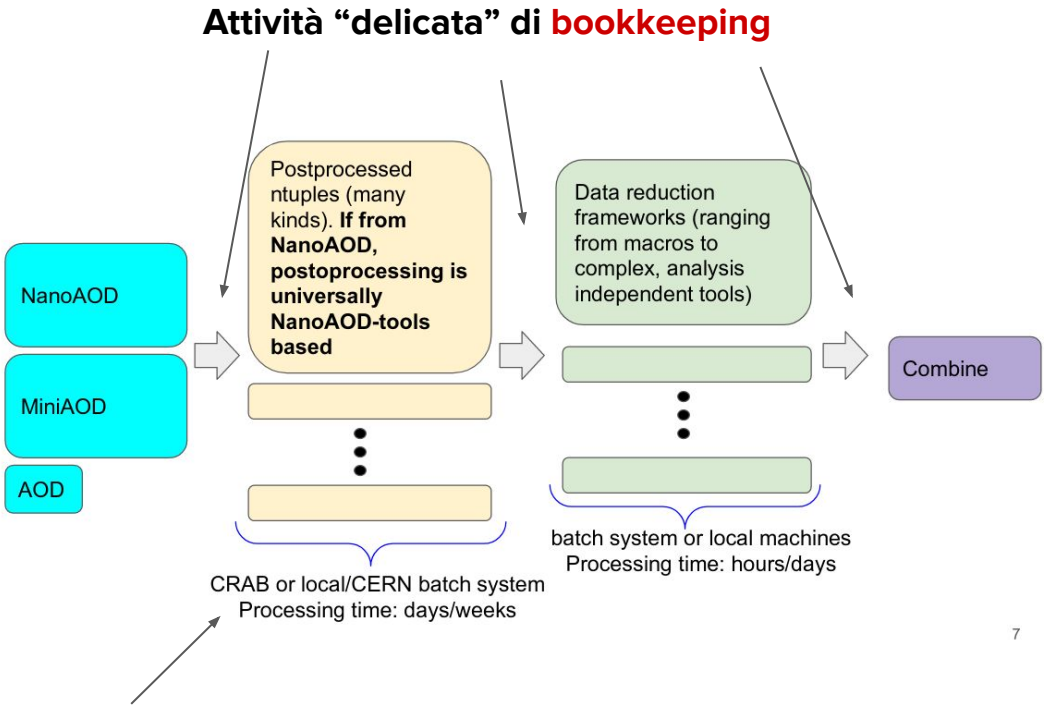
Siti
- Massimo Biasotto
- Massimo Sgaravatto
- Federica Fanzago
- Stefano Nicotri
- Francesco Failla

Idee & brainstorming/ consultancy
- Tommaso Boccali
- … infn -cloud

L’analisi CMS oggi

Un pattern comune in termini di **workflow di analisi per Run2:**

- Prima processamento tipicamente via **CRAB**
  - Ntuplizzazione
  - Produzione privata MC
- Secondo step solitamente “tailored” per ogni analisi su **batch system dedicati** o **macchine locali**



Attualmente circa ~30% del calcolo CMS e’ dedicato a questo step

Problematiche in vista di Run3 e HL-LHC

Questo tipo di approccio **ha dimostrato di funzionare egregiamente** e CMS **NON** può/vuole farne a meno. Si tratterà probabilmente di limitarne lo scopo e **complementarlo** con altre soluzioni.

- La decentralizzazione della produzione Ntuple e della produzione MC ha un **costo in termini di efficienza nell'utilizzo delle risorse**
  - N modi diversi e replicati per fare cose molto simili
  - Scarsa ottimizzazione del codice in se
- Il **bookkeeping** necessario nei vari passaggi e' tanto **più complicato quanto più sono i data-taking da analizzare coerentemente**
  - Anche questa una problematica affrontata in N sfumature dagli N framework di analisi/data reduction utilizzati con duplicazione di lavoro "NON-fisica"

Entrambe le problematiche avranno **un impatto già alla fine di Run3**, ma diventeranno **critiche/bloccanti per le stime di HL-LHC** (anche al netto dei possibili miglioramenti offerti da NanoAOD)

In altre parole CMS non puo' permettersi l'**attuale modello di analisi** nello scenario di incremento di dati e di processamento previsto per HL-LHC.

Molti framework per simili obiettivi

Nel tempo **i framework per l'analisi sono diventati un numero considerevole**, pur affrontando delle problematiche molto simili.

Questo ha degli effetti a tutti i livelli:

- Efficienza in termini di **"tempo uomo"**
  - Moltiplicando l'effort necessario per lo sviluppo e il mantenimento di N fw
  - Bug e crosscheck ripetuti in differenti circostanze
  - La condivisione e riproducibilità di una analisi e limitata all'interno del gruppo che usa lo stesso framework (e nemmeno necessariamente vero)
- Efficienza in termini di **uso di risorse**
  - Il codice spesso non garantisce una buona parallelizzazione, consumando di fatto piu risorse di quello di cui avrebbe bisogno
  - NTuple duplicate in maniera non ottimale (e.g. anche solo per aggiungere 2 variabile)

Una task force per l'analisi

Per questi motivi nel 2021 e' stata **istituita in CMS una Task Force per l'analisi** nel finalizzata a definire **raccomandazioni sul modello di analisi**, sia per Run3 che più a lungo termine per HL-LHC

Le raccomandazioni verranno poi, a seconda della tipologia, **sviluppate da gruppi esistenti** (e.g. Coffea e ROOT RDataframe) o da **eventuali nuove aree di coordinamento.** Il focus e' in primis sulla evoluzione dei framework:

- **RDataFrame**: successore di Proof e sviluppato da CERN
- **Coffea:** a guida US, framework di analisi ROOT-less, dove i .root sono letti attraverso tool python/data-science

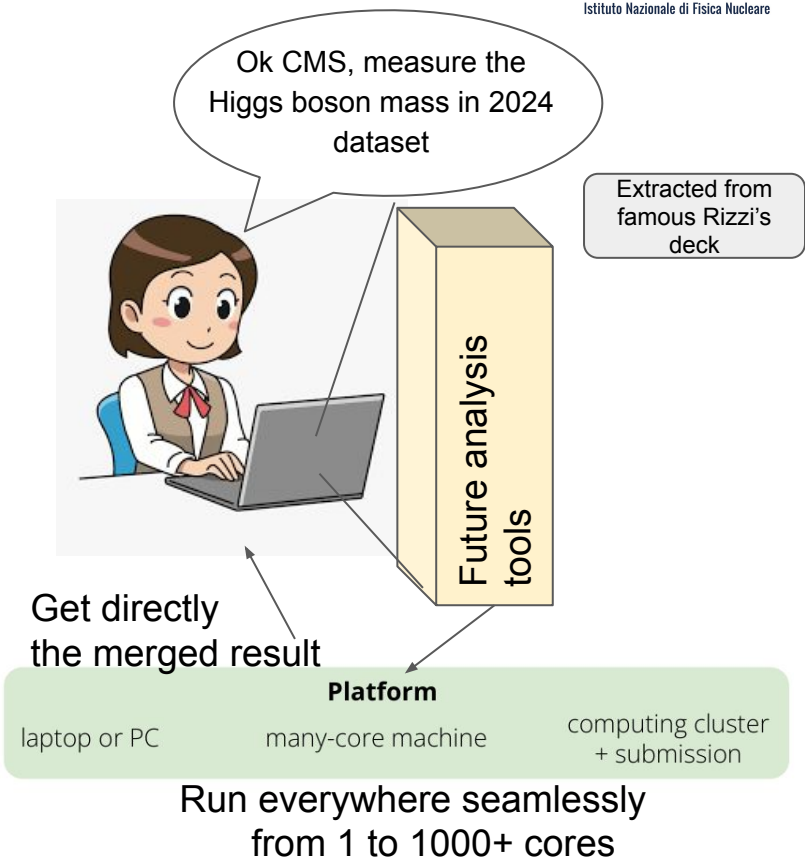
E' presente anche una componente **infrastruttura** dove diverse attività sono in corso con l'idea comune di evolvere verso scenari interattivi o quasi-interattivi (aka **Analysis Facilities**)

- jupyterHUB come interfaccia
- Distribuzione payload su risorse distribuite via Dask

L'obiettivo… in estrema sintesi

- **Spingere sull'utilizzo di NanoAOD**
  - Coprire almeno O(**75%**) delle analisi
- **Misurare/stimare il guadagno atteso** rispetto all'attuale 30% di risorse dedicate grazie ai nuovi paradigmi
  - Obiettivo avvicinarsi all'**ordine del MHz (Mevts/s)**
- Evitare di ri-inventare la ruota e concentrarsi su un **codice "dichiarativo"**
  - Dichiari solo quello che vuoi fare ad alto livello
- Il **framework pensa a tradurlo** nella maniera **piu' efficiente**
  - E.g. data splitting o multi-threading
- Con **~0 effort scalo la mia analisi su risorse distribuite**

Di fatto **2 livelli distinti** e complementari di problemi, da una parte **i framework di analisi** e dall'altra **il "provisioning" delle risorse**



Cosa facciamo in Italia: CMS@INFN

A fine 2021 e' partita un'attività con l'obiettivo di costruire **un prototipo di una infrastruttura per l'analisi dove effettuare misure e sviluppare il modello.**

Mettendo a fattor comune:

- le **esperienze tecniche accumulate negli anni**
  - Sia in termini di framework di analisi che di calcolo distribuito
- **sinergie con attività/progetti in corso**
  - Con la possibilità di contribuire con **prime misure "sul campo"**

Con alcuni valori aggiunti in mente per l'implementazione tecnica:

- **Indipendenza dal framework di analisi**
  - E.g. nel primo caso d'uso abbiamo effettuato un test con RDataFrame per interazioni avute con esperti ROOT. Ma niente vieta l'eventuale utilizzo di altri fw come Coffea
- Ambiente possibilmente **CMS-agnostic**
  - Il software di esperimento e' importato via immagini containerizzate e cvmfs

Un modello per accesso alle risorse… efficiente

L’idea e’ la creazione di una infrastruttura che permetta:

- l’**implementazione di un “continuum” di risorse federate** via HTCondor e accedute in maniera “legacy” o ~interattiva
- **Grid, Cloud, HPC non e’ piu’ un problema visible all’utente**
  - Di fatto si nasconde ed ottimizza l’accesso alle risorse di calcolo senza (o quasi) l’intervento da parte di chi fa analisi
- In aggiunta, questo scenario permette per definizione un’integrazione di **risorse opportunistiche**, ma **anche di cluster/macchine “private”** affianco a quelle **“pledged”**

Ma soprattutto possiamo così **misurare, attraverso benchmark comparativi, il guadagno rispetto ad approcci “legacy”**

Lo stato attuale

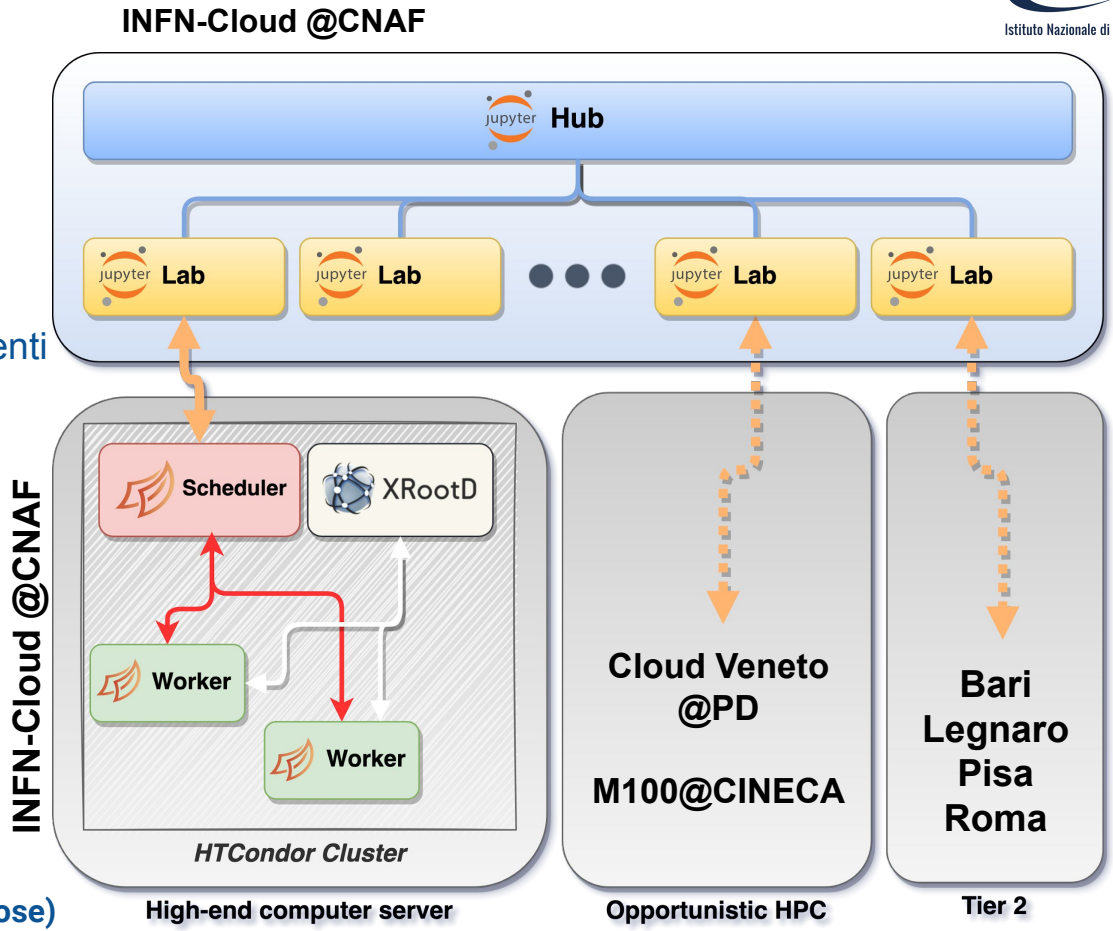
Un primo testbed è attualmente in funzione ed integrato con risorse "spare" fra T2 e Cloud

Il cluster dedicato all'entrypoint per gli utenti è instanziato @CNAF (INFN-Cloud)

- JupyterHUB per spawn di JupyterLAB on-demand

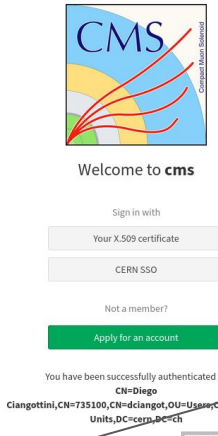
I T2 a Legnaro e Roma al momento dedicano O(100) core come "seed" per i primi test con analisi CMS - incrementabili on-demand in caso di necessita'

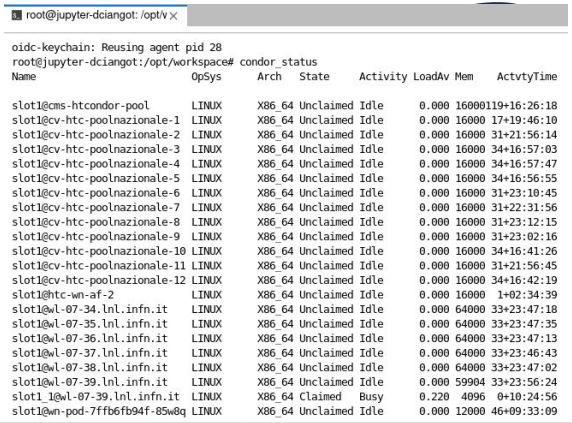
Quasi tutti Tier2 hanno comunque dimostrato la fattibilità dell'integrazione (ricette ~battle tested basate su docker-compose)

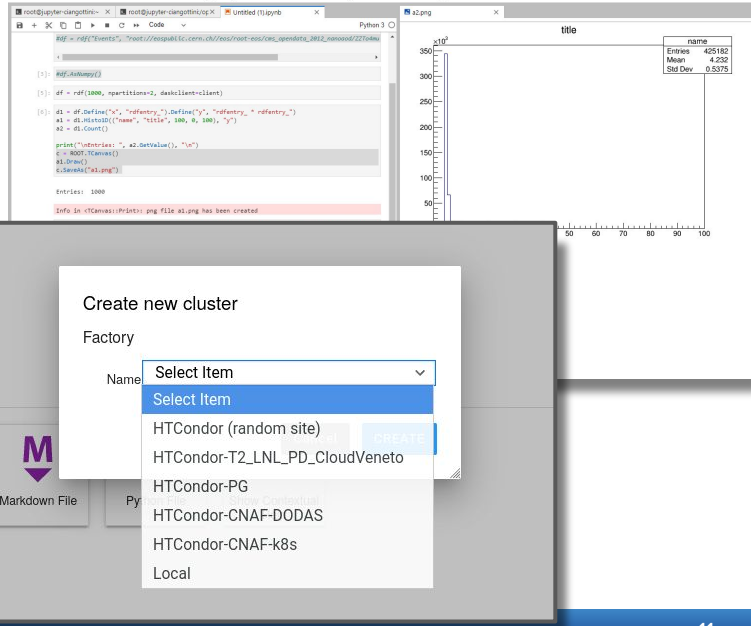


Come funziona…

- **Unico entrypoint** con JupyterHUB con **autenticazione via CMS-IAM**
- Una volta entrato l'utente ha a disposizione **accesso al pool HTCondor dedicato**
  - Da poter usare anche "as is"
- **Kernel python configurabili** per poter lanciare notebook interattivi con il software desiderato
- Possibilità di **andare distribuiti sulle risorse e.g. di un T2 via DASK con pochi click** su una dashboard







Gestione ambiente software dell'utente

Il disegno per permettere ad un utente di **utilizzare il proprio ambiente** sulla infrastruttura si puo' riassumere con:

1. Creazione dell'immagine (docker, o cmq OCI compliant) con il **mio software al interno**
   a. Caricandolo in un registro (dockerhub, gitlab etc)
2. **Sincronizzazione in un repository CVMFS**
   a. Aggiungendo il nome dell'immagine in una lista ad esempio su github (vedi DUCC)
3. Caricamento, nel mio ambiente jupyterLab, di un **kernel "singularity-based"** come se fosse il proprio kernel python di default
   a. Più efficiente dal punto di vista del trasferimento dati, di caricare il "LAB" direttamente dal container utente (idealmente di diversi GB)
4. I **nodi remoti saranno capaci di avviare la stessa immagine** montando a loro volta il repo CVMFS

I primi test

credits to T. Tedeschi

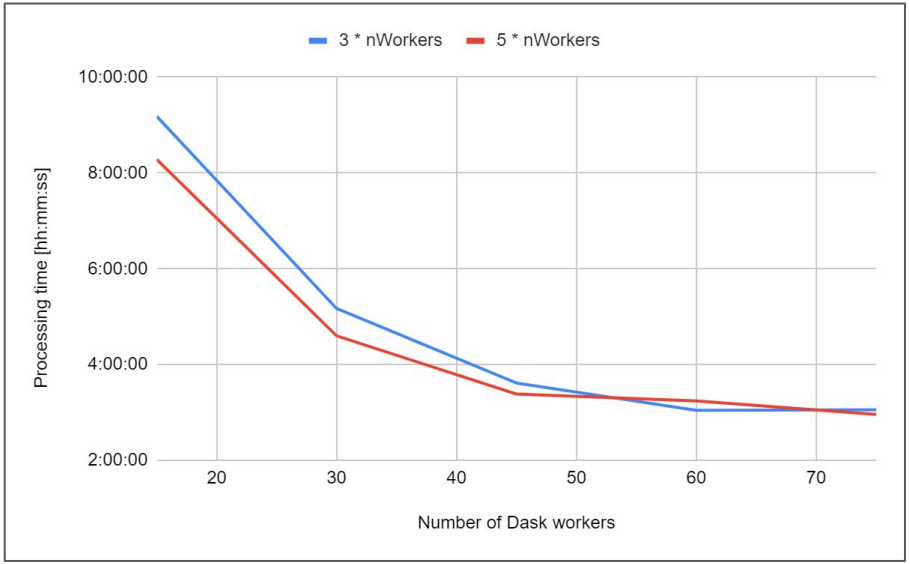
E’ iniziato un processo di **“benchmarking”** per ottenere delle stime sui benefici. Oltre ad offrire i primi feedback **in uno scenario dove siamo tra i primi a poter “giocare”**

- Il primo passo è stato prendere un’analisi esistente in modalita’ Run-2 e **riprodurla su framework RDataFrame**
- A questo punto stiamo raccogliendo i primi risultati **comparando sulla stessa infrastruttura l’approccio batch con quello quasi-interattivo**

L’analisi usata è VBS SSWW che è basata su **~10^9 evt NanoAOD**

Un preliminare giro di test mostra un **risparmio in termini di tempi per l’analisi e2e incoraggianti** (ordine 40%)

Preselected ReReco 2017 data + MC (9605 files, 1.5 TB, 953 million events)



Anche l’andamento all’aumentare dei nodi sembrerebbe quello atteso

Workshop su analisi CMS Italia

Siamo quindi al punto dove è importante **consolidare ed ampliare la base di utenza** per ricevere feedback ed indirizzare gli sviluppi successivi.

A questo proposito a Marzo 2022 e' stato organizzato un WS con utenti allo scopo di:

- Coordinare la **raccolta di requirements e feedback** per l'attività di R&D da parte di chi fa analisi
- Estendere i test ad altre analisi con **altri casi d'uso**
  - grazie alla attività iniziata durante il workshop abbiamo iniziato interazione con volontari

Piergiulio Lenzi sta coordinando l'attività degli "early adopter" per l'attività di benchmarking.

Sviluppi in discussione

Ci sono discussioni in corso per attività sinergiche ad altri progetti e finalizzate a:

- Fornire **una “shared home directory”**, per gli utenti e condivisa su tutti i WNs
  - Un’idea è quella di valutare la strada Ceph-FS
- Integrazione con **DataManagement via plugin RUCIO**
  - Fondamentale per spostare input ed output in maniera efficiente e semplice
  - Propedeutico per una integrazione alla “data-lake”
- Come gia detto prima, è fondamentale l’estensione della base utente per feedback

Lezioni imparate durante questa prima fase…

- L’**integrazione di nuove risorse** si riduce a seguire istruzioni per la creazione di N container via docker-compose
  - E.g. nessuna richiesta special dal punto di vista di inbound connectivity
  - O(30min) di lavoro effettivo
- Una **home condivisa** è un aspetto fondamentale a cui far fronte
- Ogni singola componente utilizzata **NON ha nulla di CMS-specifico**
  - Interessante per valutare uno scopo multi-esperimento magari
- Discutendo sia all’interno della task force che nei forum (vedi dopo) è di estremo interesse il **sapere sfruttare risorse in siti geo-distribuiti ed integrarli con requirement minimi**
  - **peculiarita’ del nostro setup,** grazie ad R&D fatto a livello DASK
- Sebbene il tutto sia intuitivo/facile da gestire, **non c’è al momento un modello di supporto**
- L’intero deployment è basato su tecnologie rodate e pensate per cloud. Di fatto è quindi **una soluzione “INFN-cloud compliant”**
  - può essere un sistema esportato verso altri casi d’uso in quell'infrastruttura

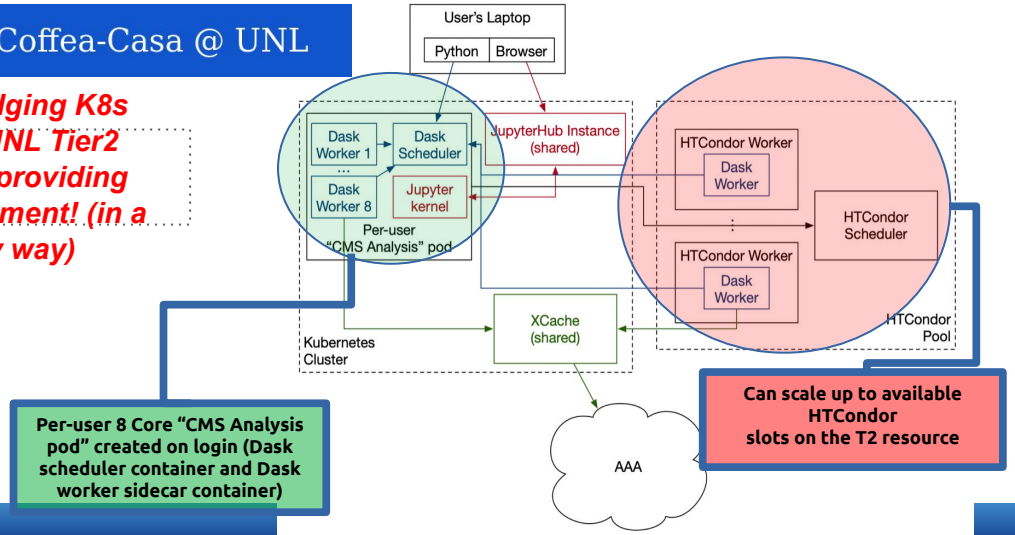
Contesto “cross-experiment” e attivita’ US

- HSF AF Forum
  - HSF ha avviato un forum dedicato alle analysis facility, oltre al gia esistente WG sui tool
  - Lo scopo è raccogliere e condividere esperienza tra i vari gruppi che lavorano ad R&D legati all’analisi
  - Uno degli obiettivi del forum è la stesura di un whitepaper in questo ambito
- GDB

Al momento, oltre all'attività INFN, lo sviluppo più avanzato nel contesto dei modelli di analisi è lato US.

Analysis Facility Coffea-Casa @ UNL

We are easily bridging K8s resources with UNL Tier2 resources, while providing interactive environment! (in a token-friendly way)



CMSAF @T2 Nebraska “Coffea-casa”
https://coffea.casa

OpenData AF @T2 Nebraska “Coffea-casa”
https://coffea-opendata.casa

BACKUP

Recap attivita’ e discussioni 1/2

- **In CMS:**
  - Analysis Task Force [REF]
  - CMS O&C week [REF]
    - Presentazioni e sessioni dedicate alla discussione delle attivita’ di R&D in corso e ai report da altri forum
  - Meeting settimanali CMS-IT calcolo [REF]
    - Discussioni tecniche per l’implementazione di un prototipo INFN
  - CMS-IT workshop analisi [REF
    - Raccolta feedback dai vari gruppi di analisi e offerta di test dell’infrastruttura a primi volontari (vedi dopo)