# XCache deployment@CNAF per CMS

Daniele Spiga

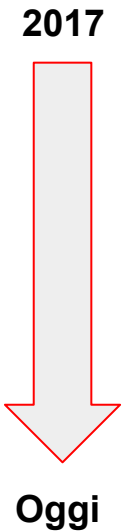
# Cache@CMS-ITA: Cosa abbiamo fatto come INFN negli anni

E’ un aspetto su cui abbiamo investito molto dal 2017

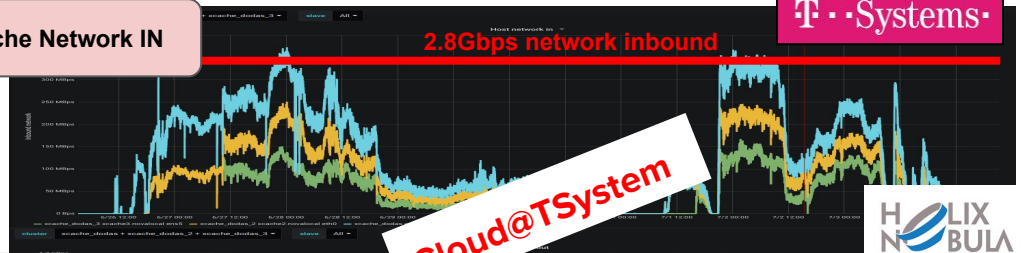
- Ma le prime slides dove si parla di cache circolavano almeno nel 2016 ( Tommaso & Alessandro )

Investimenti CMS-INFN e attività :

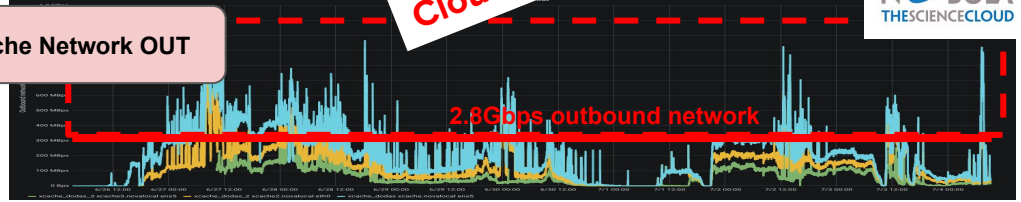
- Borsa INFN per attività di R&D Calcolo per LHC
- Sviluppo fatta nel progetto XDC (Smart Cache)
- Test con il progetto HelixNebula SciCloud
- Sviluppo in corso nel progetto ESCAPE
  - Data lake prototype + cache + token etc…
- Setup “production ready” nell’integrazione @CINECA



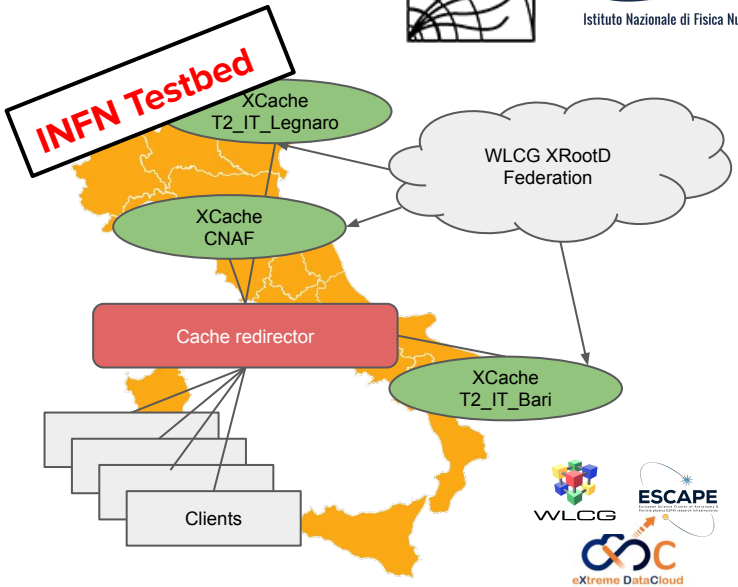
Cache Network IN

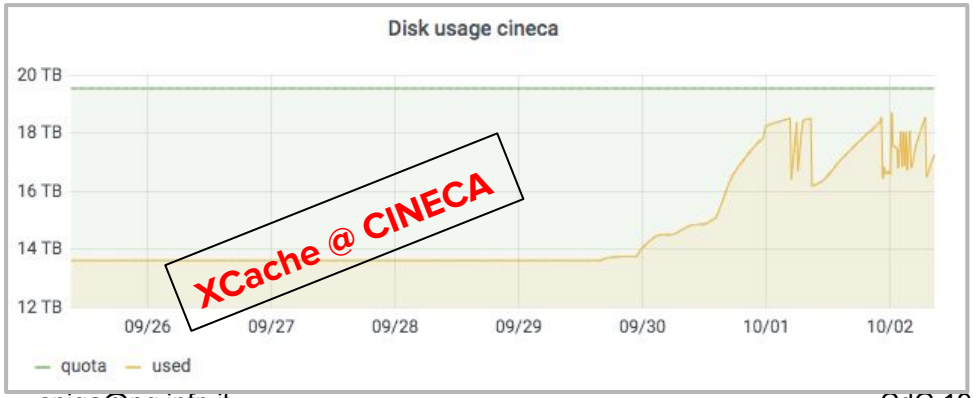



Cache Network OUT

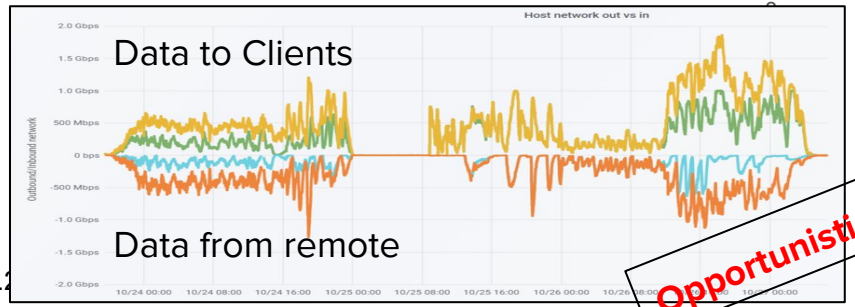



Cloud@TSystem




XCache @ CINECA




Opportunistic

# Cosa abbiamo imparato finora

Sicuramente che sappiamo usare le cache, sappiamo farlo in varie configurazione, e.g. cache gerarchiche, e che le sappiamo “customizzare” ( plugin/configurazioni )

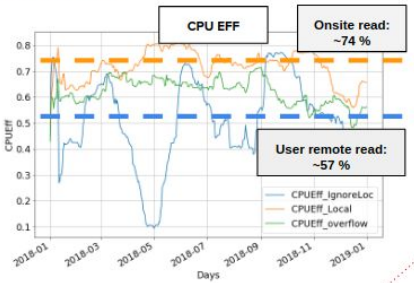
## Cache effect on CPU efficiency @ CMS

- We studied monitoring data of the whole **2018 CMS analysis workflows**
  - Remote data read costs on average **about 15% of CPU time w.r.t. Onsite data reading**

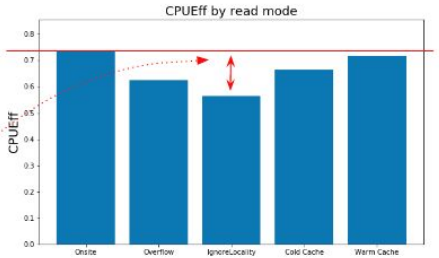
CPUEff = sum(cpu time) / sum (job time)

From data@CERN MONIT hdfs

@Italian Tier2's




Measured on testbed

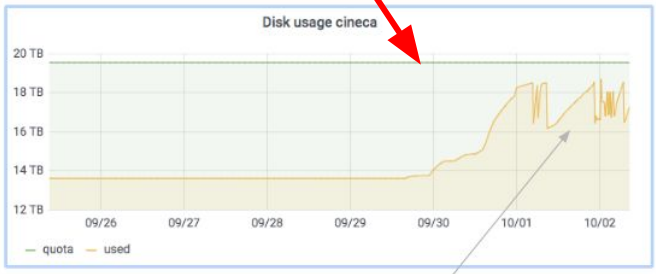



Caches allow to reduces the overall WAN traffic and, makes the processing job that requested the data **more efficient** by reducing I/O wait time for remote data.



Buon impatto sull’efficienza

Possiamo ottimizzare le configurazioni




**20 TB SSD cache is smallish, especially when doing analysis (we need better cache policies)**

# Le Cache @CMS

Non è un interesse INFN. US-CMS ha investito molto sviluppando il progetto SoCal, e ora anche il CERN ( vedi dopo ) e anche gli Spagnoli ci stanno guardando

Di cosa fanno caching?

- /*/*/MINIAOD
- /*/*/MINIAODSIM

(Circa 8 PB Totale)

## SoCal Cache idea: Two caches become one

US - CMS

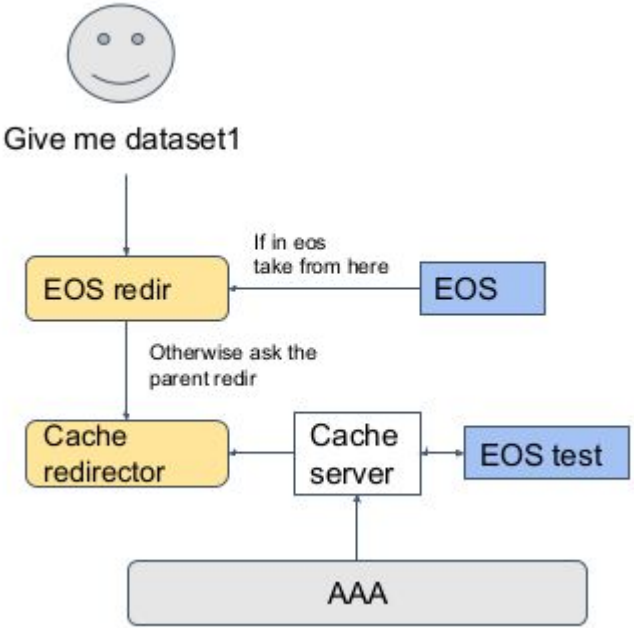

- 120 Miles
- 100 Gbit/sec
- 3ms

| | UCSD | Caltech |
|---|---|---|
| **Nodes** | 11 (+1 JBOD) | 2* |
| **Disk Capacity per node** | 12 x 2TB = 24TB (+ 48 x 11TB) | 30 x 6TB (HGST Ultrastar 7K6000) |
| **Network Card per node** | 10 Gbps (+ 40 Gbps) | 40 Gbps |
| **Total Disk Capacity** | 264 TB (+ 528 TB) = **792 TB** | **360 TB*** |
| **TOTAL** | 792 TB + 360 TB* = **1,152 TB** | |

# CMS-CERN e XCache

CMS ha proposto di esporre alla comunità, un'opzione per accedere a NanoAOD e i MiniAOD non presenti :

- Tradotto: schierare un'istanza xCache al CERN e testarla come soluzione per leggere tutti i Nano e i Mini che non sono su EOS
  - Successivamente fare comparazioni tra varie configurazioni staging vs proxy




**Inoltre: O&C ha proposto a CMS-Ita di gestire questa iniziativa e di farlo in collaborazione con le nostre attività**

- Tradotto: fare quindi un setup al CNAF simile (ma non identico, prox slide)

Il lavoro al CERN è partito (per mano INFN) e lo stato dei lavori è abbastanza avanzato ( setup iniziale pronto e in fase di puppettizzazione )

# Cosa vogliamo fare noi ( un po di numeri )

La proposta è di implementare la seguente architettura:

- Schierare una cache al Tier1 ( cache di livello 2) con una quantità di disco dell'ordine di 50 TB, per i NanoAOD
  - Successivamente capiamo cosa vogliamo fare coi Mini
- Questa deve alimentare una serie di caches ( primo livello ) eventualmente veloci e distribuite dove si fa l'analisi
  - In linea di principio "dove si fa analisi" può essere "qualunque" cosa: una singola macchina o un cluster dislocati nei centri di calcolo italiani (vedi dopo)

**Cosa vorremmo chiedere al Tier1**: 2 server dedicati, che possano essere anche VM con xrootd ( con 50TB di disco anche sottratto alle pledges )

- Come detto, le configurazioni andranno aggiustate man mano quindi non mi aspetto di partire con un setup standard

# Perchè ci interessa ( prossimo step )

Tutto questo sistema è funzionale all' attività di R&D sulle infrastrutture di analisi ( ultimo miglio ) e valutarne le prestazioni.

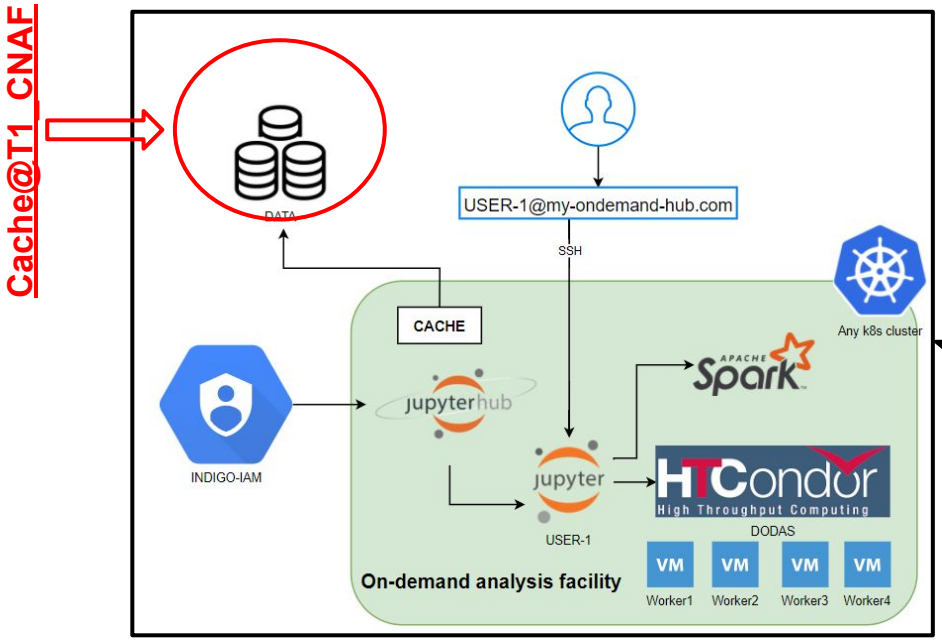
Stiamo investigando un approccio interattivo (che aiuti e stimoli anche l'adozione dell'ecosistema python). In particolare vogliamo cercare di:

- Minimizzare l'uso della GRID per questa fase  del data processing ( latency )
- Massimizzare il throughput, cosa che ci interessa tanto più quanto più usiamo i NanoAOD ( Run3 e HL-LHC )

Vorremmo/dovremmo poi coinvolgere ricercatori che facciano attività vera...

# Dove siamo

Abbiamo varie idee, ma anche alcuni prototipi pronti ad esser integrati per testare il sistema di cache descritto.




Gli scenari che vogliamo esplorare sono sia quello di "singola macchina" che quello distribuito per fare scale out

- Scale out su risorse "locali" e veloci

Schema dell'architettura che stiamo implementando

# Infine… integrazione con “DataLake”

Qui collassano tante attività che portate avanti in vari contesti

- ESCAPE
- INFN-Cloud
- ..

NOTA: tanti contributi ( tanto CNAF ) su molti aspetti