

Review Calcolo ATLAS Italia

Settembre 2021

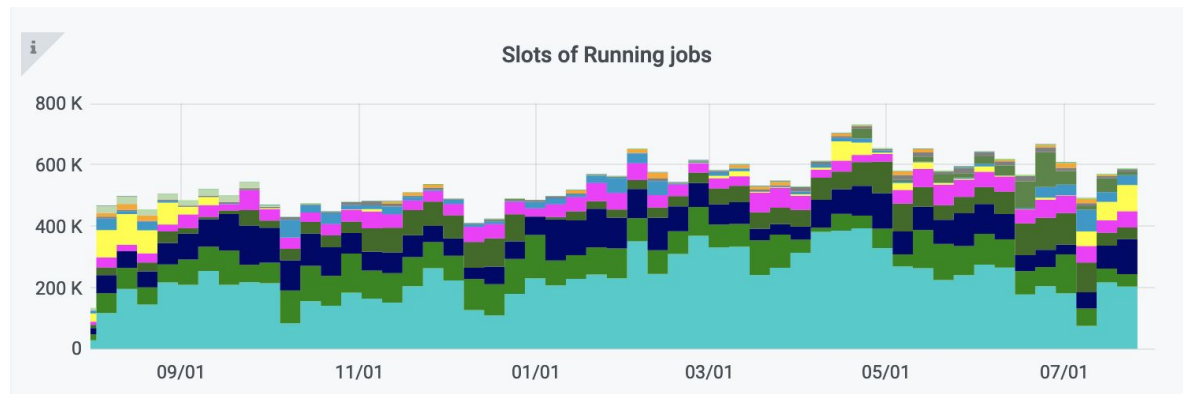
L. Carminati (UniMi), L. Rinaldi (UniBo)
on behalf of the ATLAS-ITALIA computing group

Outline

- ❑ Andamento computing ATLAS e contributo ATLAS Italia (luglio 2020-luglio 2021).
- ❑ Principali sviluppi e novità nel calcolo ATLAS
- ❑ Stato acquisti e gare e richieste per il 2022

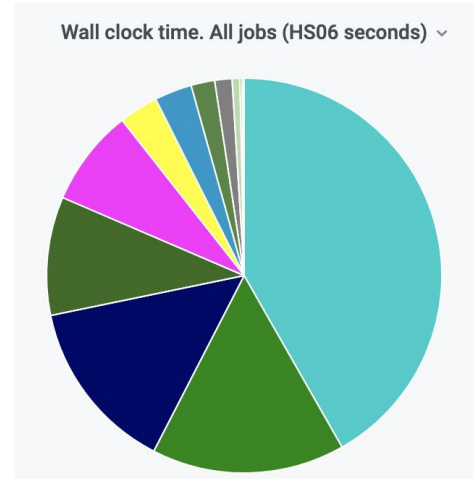
Attività' computing ATLAS (luglio 2020 - luglio 2021) : CPU

- ❑ Circa 500k computing slots ogni giorno per tutto il periodo in esame (pledged and unpledged, all Tiers resources together)
 - ❑ Importante sforzo che ricade su pochi esperti chiave (sia ADC dia DDM)



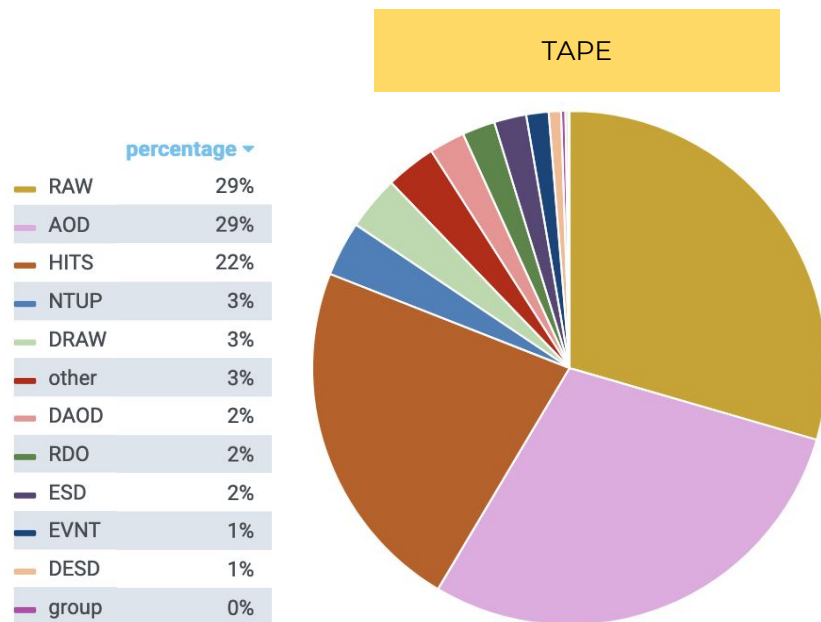
- ❑ Attività' tipica di periodi di analisi (senza data taking): Intensa attività' di produzione MC (~70% includendo generazione, simulazione e ricostruzione) e analisi dati (~25% tra derivations e analisi vera e propria)

	total	percentage
MC Simulation Full	81 Tri	42%
Group Production	31 Tri	16%
MC Reconstruction	27 Tri	14%
MC Event Generation	19 Tri	10%
User Analysis	15 Tri	8%
Data Processing	6 Tri	3%
MC Simulation Fast	6 Tri	3%
MC Resimulation	4 Tri	2%
Group Analysis	3 Tri	1%
COVID	1 Tri	1%
Testing	364 Bil	0%
MC Merge	264 Bil	0%

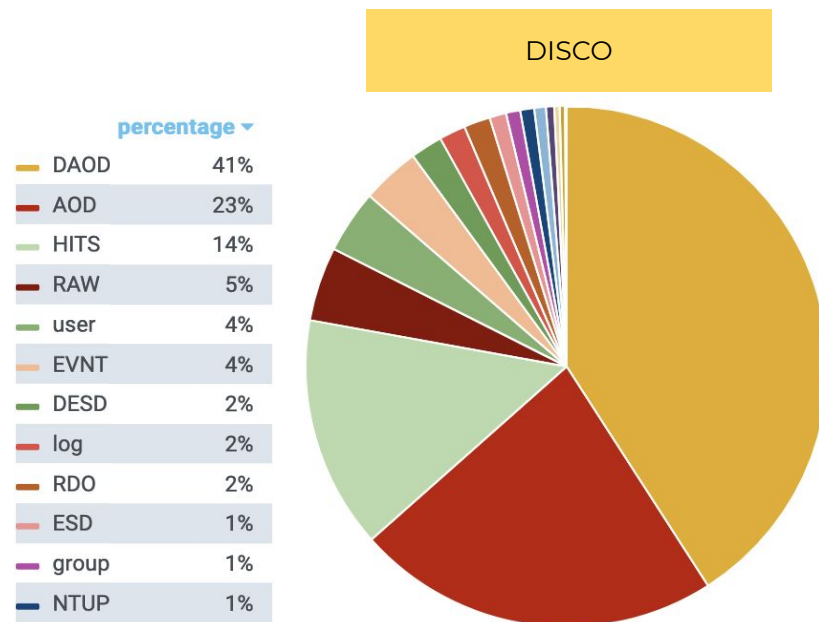


Attività computing ATLAS (luglio 2020 - luglio 2021) : disco e tape

- ❑ Tape (no T0) occupato ~ 180 PB :
principalmente RAW data, AOD e HITS



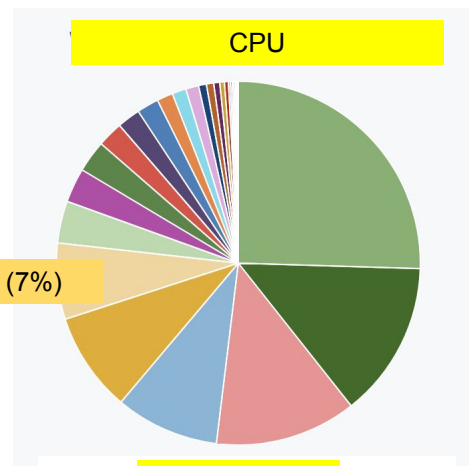
- ❑ Storage occupato (datadisk, calibdisk, groupdisk e localgroupdisk) ai T1/2/3 : ~ 220 PB : principalmente dati derivati (DAOD) e AOD



Contributo italiano computing ATLAS (luglio 2020 - luglio 2021)

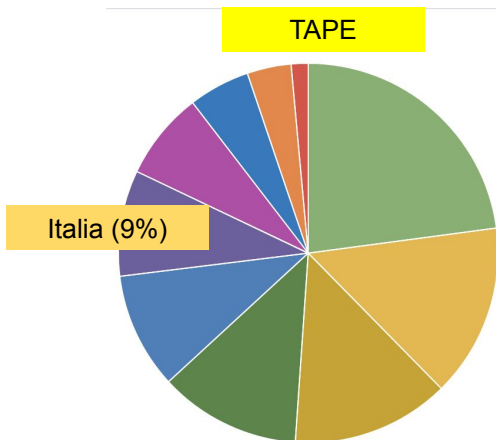
Contributo italiano alle attività di computing di ATLAS :

- ❑ CPU : overall ~ 7% wallclock time totale in HS06-sec (T1+T2+T3, solo risorse grid, T0 escluso)
 - ❑ leggermente sotto share (9%) perché sono inclusi eventuali over-pledge dei siti world-wide
- ❑ Disco occupato (datadisk, calibdisk, groupdisk e localgroupdisk) : overall ~6% (escludendo T0)
 - ❑ leggermente sotto share (8%) risorse 2020/21 in arrivo dalla gara comune
- ❑ Tape (T1 only) occupato : ~9%
 - ❑ perfettamente in linea con lo share (9%)



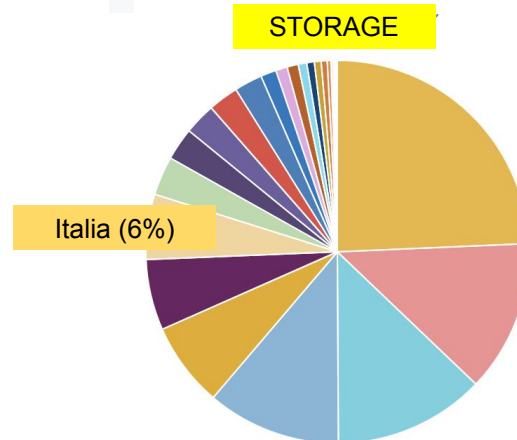
percentage ▾

USA	25%
UK	14%
Germany	13%
France	9%
Canada	9%
Italy	7%
Japan	4%
Spain	3%
Slovenia	3%
Switzerland	2%
Netherlands	2%
Russian Fed	2%



percentage ▾

USA	23%
UK	15%
France	13%
Germany	12%
Canada	10%
Italy	9%
Netherlands	8%
Nordic	5%
Spain	4%
Russian Fe	1%



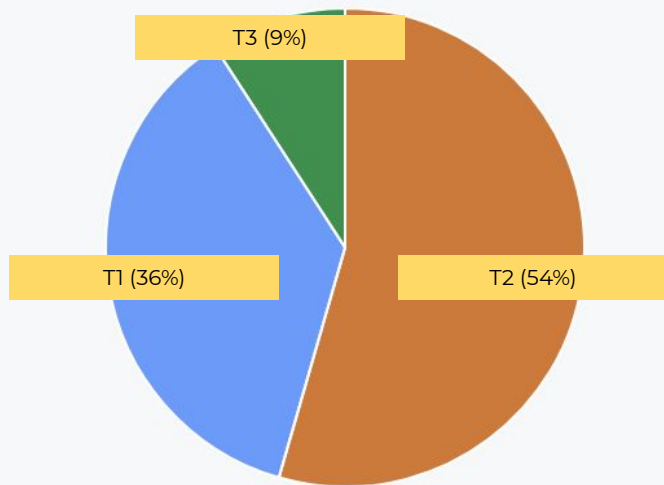
percentage ▾

USA	24%
Germany	13%
UK	13%
France	11%
Canada	7%
Nordic	6%
Italy	6%
Japan	3%
Netherlands	3%
Spain	3%
Taiwan	3%
Russian Fed	2%

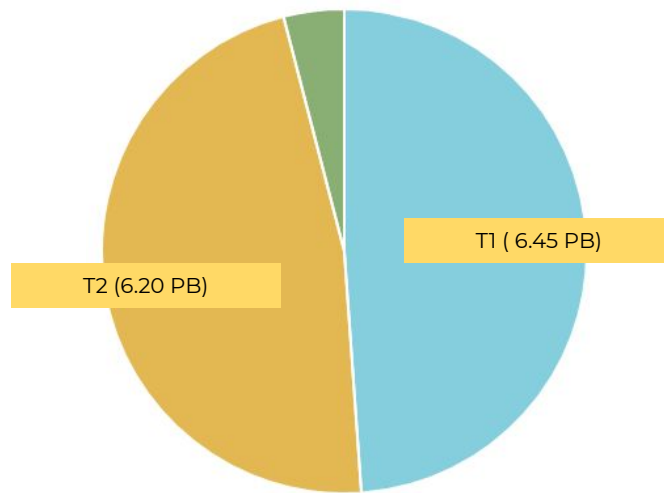
Distribuzione carico lavoro ATLAS in Italia

Lavoro distribuito sul T1 (CNAF), i quattro T2 (Frascati, Milano, Napoli e Roma1) e quattro T3 (Cosenza, Genova, Lecce, Roma3)

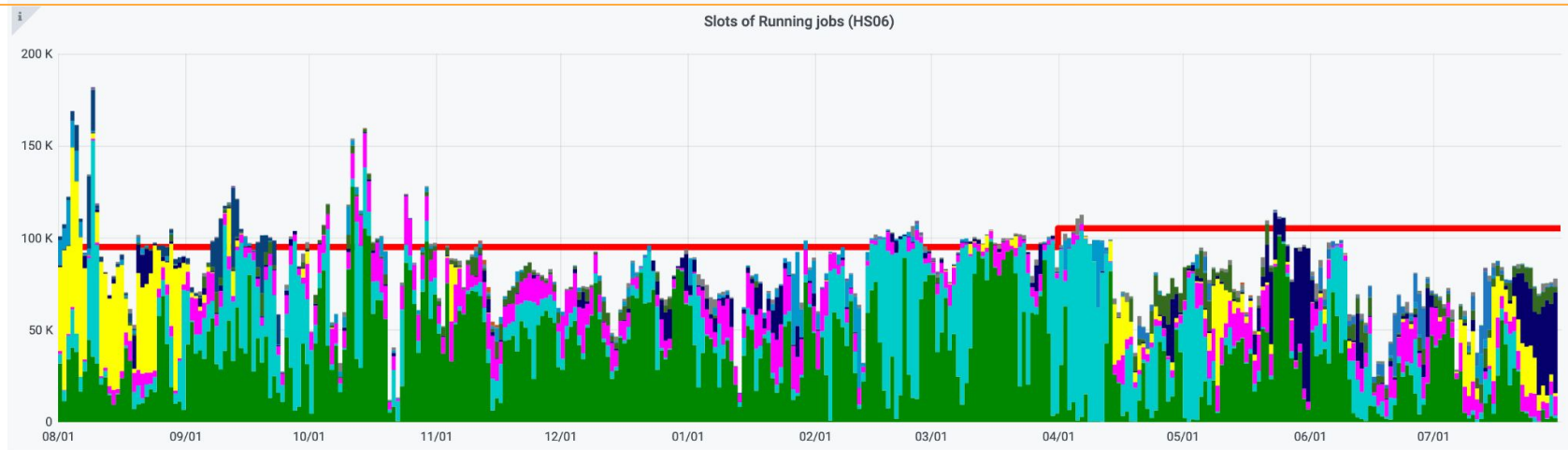
CPU



STORAGE



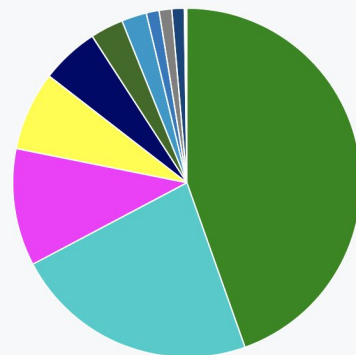
Attività' ATLAS al T1 : CPU



- ❑ Share tipico di un T1 per un periodo di analisi dati
 - ❑ preparazione dati secondari per l'analisi (group production),
 - ❑ analisi (user analysis)
 - ❑ produzione di MC
 - ❑ Reprocessing
- ❑ Average pledge 98.5 kHS06, computing power fornita 82.3 kHS06 (per la prima parte dell'anno nel grafico sottostimato il corepower medio al CNAF)

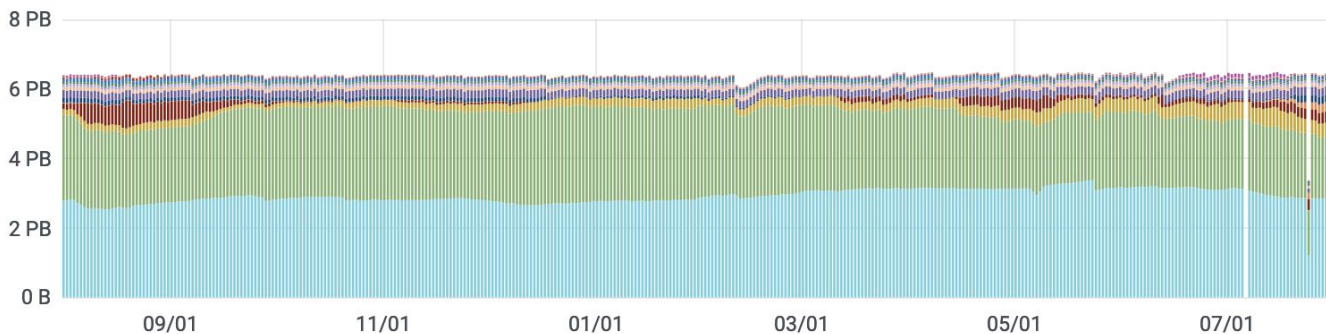
percentage ▾	
Group Production	45%
MC Simulation Full	23%
User Analysis	11%
Data Processing	7%
MC Reconstruction	5%
MC Event Generati	3%
MC Simulation Fas	2%
MC Resimulation	1%
Group Analysis	1%
COVID	1%
Testing	0%
MC Merge	0%

Wall clock time. All jobs (HS06 seconds) ▾

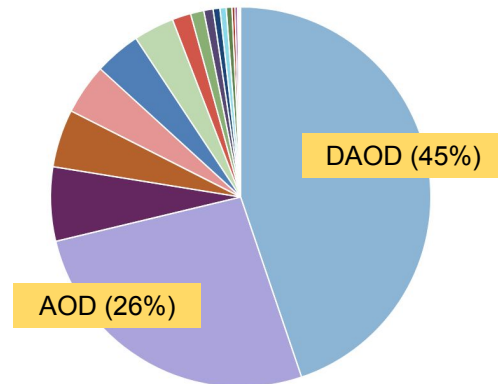


Attività al T1 : occupazione disco e tape

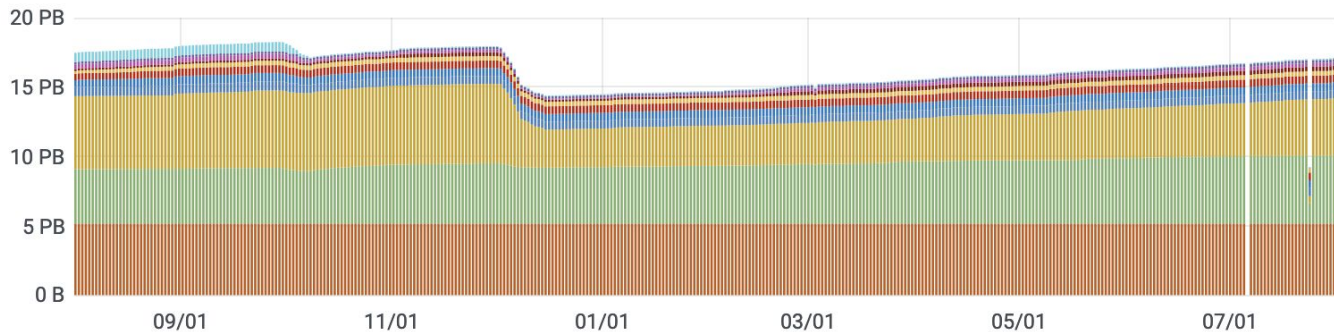
Volume per datatype_grouped



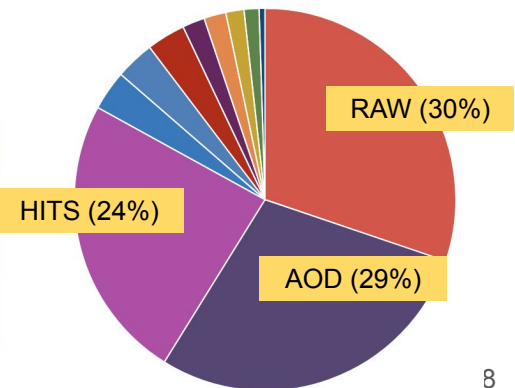
Volume by datatype_grouped ▾



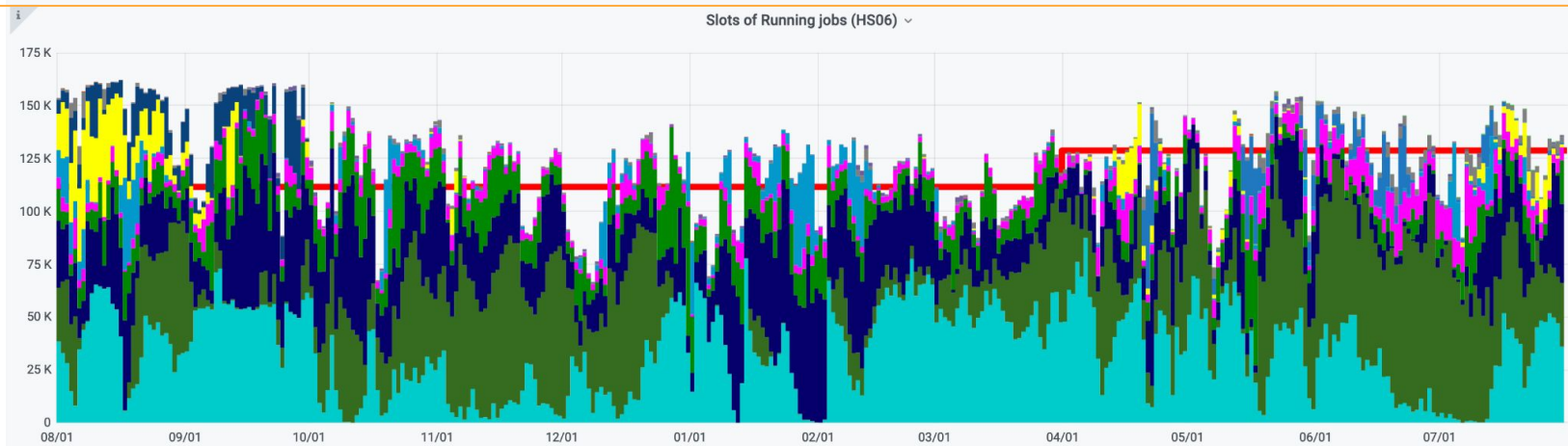
Volume per datatype_grouped



Volume by datatype_grouped ▾

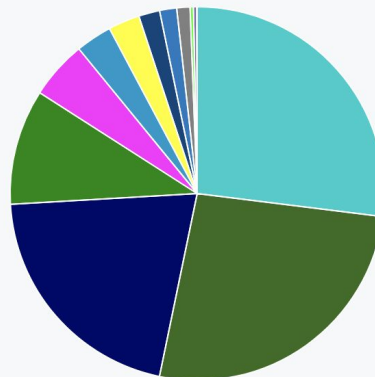


Attività' ATLAS ai T2 : CPU



- ❑ Tipico share T2 :
 - ❑ MC (generation, simulation e reconstruction)
 - ❑ produzione di dati secondari (group production) e analisi utenti
- ❑ Average pledge 117 kHS06, computing power fornita 125 kHS06

Wall clock time. All jobs (HS06 seconds) ▾



percentage ▾

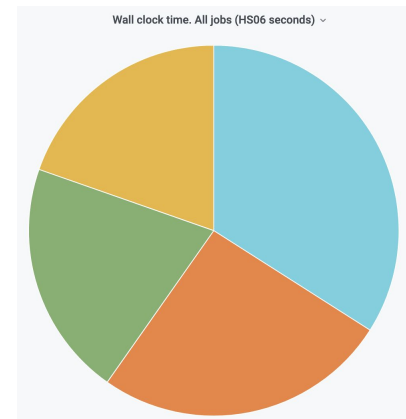
MC Simulation Full	27%
MC Event Generation	26%
MC Reconstruction	21%
Group Production	10%
User Analysis	5%
MC Simulation Fast	3%
Data Processing	3%
COVID	2%
MC Resimulation	1%
Group Analysis	1%
Testing	0%
MC Merge	0%

Attività' ATLAS ai T2 : CPU

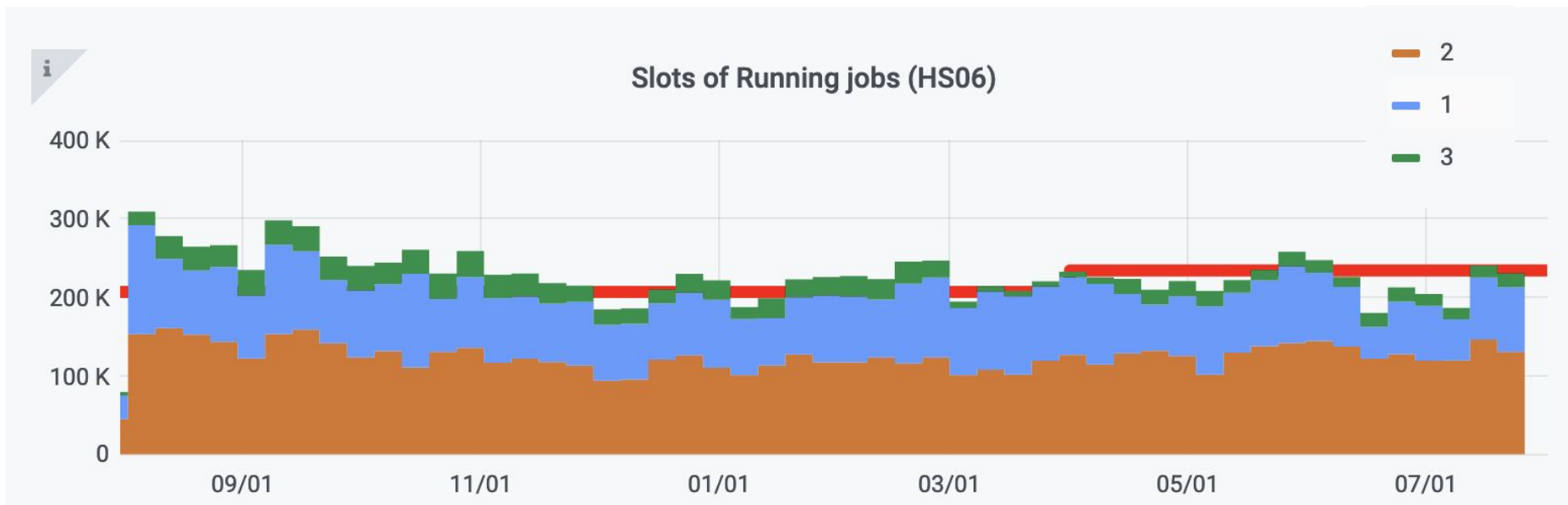
- ❑ In generale buon comportamento dei T2 italiani per l'ultimo anno di attività'
 - ❑ reliability/availability ~ 95 % per tutti i siti (vedi backup)
 - ❑ Efficienza ~ 95%
 - ❑ Valori in linea con i migliori siti ATLAS
- ❑ Importanti interventi configurazioni :
 - ❑ Migrazione di tutti i CE (dismissione creamCE) : HTCondorCE o ArcCE
 - ❑ Transizione a HTTP (invece di gridftp) per Third-party-copy
 - ❑ DPM support will end mid-2024, siti DPM (Roma1, Napoli e Frascati) stanno discutendo i possibili scenari
- ❑ Milano ha affrontato i problemi cronici allo storage con un drastico intervento (storage completamente svuotato, riconfigurato e rimesso online)
 - ❑ Ottima performance ma ristretta attività' a task meno i/o intensive (generazione e simulazione)
 - ❑ Potenziamento degli apparati di rete (richieste CCR lo scorso anno e finanziamenti locali)
 - ❑ Tests per riattivazione analisi in corso



Milano solo generazione e simulazione



CPU performance summary (all sites)



Buona performance globalmente dei siti italiani

- ❑ 216 kHS06 pledge, 226 kHS06 delivered : leggero over-pledge dei T2 e il contributo dei T3 hanno compensato l'under-pledge del T1
- ❑ Overall efficiency ~ 95% (in linea con l'efficienza dei migliori siti ATLAS)

Situazione e disco e TAPE

	Disco occupato (PB) (dati luglio 2021)	Disco allocato (PB) (dati luglio 2021)
Frascati	1.40	1.58
Milano	1.10	1.68
Napoli	2.00	2.27
Roma	1.53	1.74
tot.	6.03	7.27

- Pledge 2021 : 9.1 PB
- Gara storage comune : 2.2 PB
- Gara storage Frascati: 0.6 PB
- Obsolescenza : 0.3 PB

	Disco occupato (PB) (dati luglio 2021)	Disco allocato (PB) (dati luglio 2021)	Tape (PB) (dati luglio 2021)
CNAF	6.91	8.48	21.88

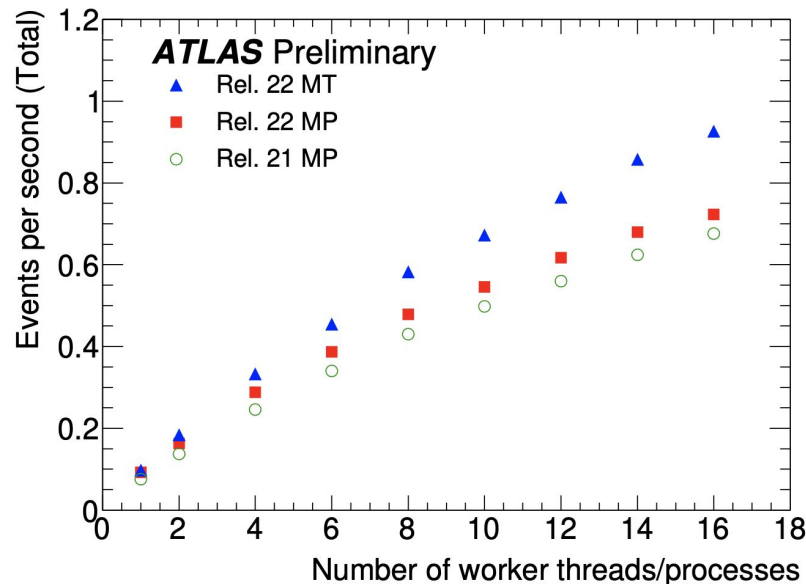
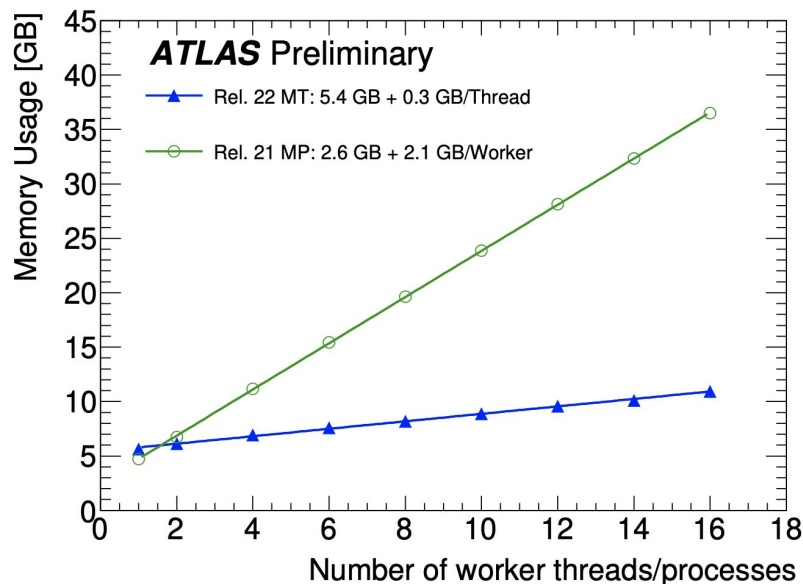
- Pledge 2021 (disco): 9.45 PB
- Pledge 2021 (tape): 21.15

ATLAS Italia partecipa alle attività di Atlas Distributed Computing (ADC) in diversi aspetti

- ❑ Containers (docker, singularity) [A. De Salvo]
- ❑ Database (Frontier, Conditions) [A. De Salvo, L. Rinaldi]
- ❑ Harvester (evoluzione del WFMS di ATLAS) [A. De Salvo]
- ❑ International Computing Board [L. Carminati, L. Rinaldi]
- ❑ "S&C Financial Advisor" [D. Barberis]
- ❑ Installazione del software (CVMFS e distribuzione) [A. De Salvo]
- ❑ ADC Monitoring [dal 2017] [D. Barberis]
- ❑ ASCIG (ATLAS Software & Computing Infrastructure Group) [dal 2019] [D. Barberis]
- ❑ Network infrastructure (LHCONE) [Tutta la federazione italiana dei T1/T2/T3 italiana]
- ❑ distributed computing and storage evolution R&D [G. Carlino, A. De Salvo, A. Doria, E. Vilucchi]
 - ❑ Federazioni di xrootd e HTTPD, DPM, Storage Caching, ...
- ❑ VO management [A. De Salvo, E. Vilucchi]
- ❑ Chair computing speakers committee [L. Perini]
- ❑ EventIndex [D. Barberis]
- ❑ Tier2
 - ❑ Gestione infrastrutture, R&D, etc. [A. De Salvo, A. Doria, D. Rebatto, E. Vilucchi]

AthenaMT (MultiThreading) : uno degli achievement fondamentali degli sviluppi software degli ultimi anni

1. L'ultima release (rel 22) del software di ricostruzione di ATLAS (Athen) che verra' utilizzato di default per reprocessing run2 e per il run3 e' basata su multi-thread
2. Considerevole risparmio di memoria (all threads share the same allocated memory)



Highlights (II): simulation optimisation

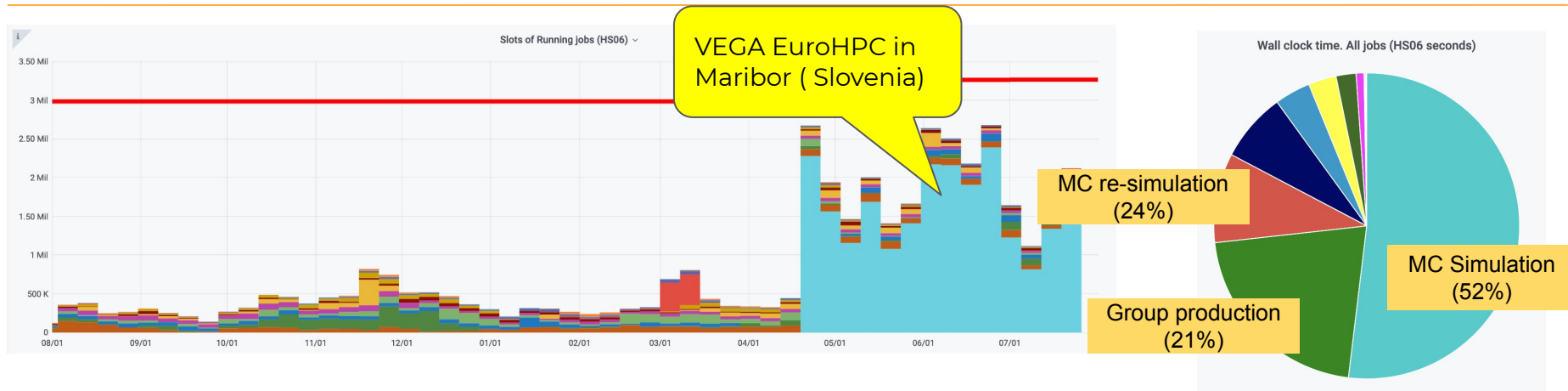
1. A Geant4 (G4) Optimisation Task Force (TF) has been setup for optimising the performance of the ATLAS G4 simulation software:
 - 1.1. Taking advantage of **intrinsic performance optimizations** coming with newer Geant4 versions
 - 1.2. Optimization with **tuning of G4 parameters** (physics models, physics lists per regions)
 - 1.3. **Neutron** and **Photon Russian Roulette + EM range cuts** (ongoing physics validation)
 - 1.4. **Geometry** optimisations (new EMEC variants + R&D on ML guided steppers in geometry)
 - 1.5. **Magnetic field** tailored switch-off
 - 1.6. **Geant4 linking** as static library (a.k.a. Big library)
2. A new workflow, MonteCarlo ReSimulation, was developed to minimize the resources needed to apply physics improvements to already generated FullSim HITS: the resources used by this workflow are 5-10% of the ones which would have been needed if we should have re-run the FullSim completely. This workflow was applied on 16B events.

Stima : si riesce a ottenere un miglioramento della velocità della fullsimulation ~ 20-25%

Highlights (III) : data carousel

- ❑ More dynamic use of the TAPE capabilities: “by ‘data carousel’, we mean an orchestration between workflow management (WFMS), data management (DDM/Rucio) and tape services whereby a bulk production campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X% (X~5%-10%) of inputs onto buffer disk, such that only X% of inputs are pinned on disk at any one time”
 - ❑ Retrieve a fraction of RAW data, process them, delete from disk, start again
- ❑ ATLAS started the Data Carousel R&D in June, 2018, to study the feasibility to get inputs from tape directly, for various ATLAS workflows, such as derivation production and RAW data re-processing.
 - ❑ CNAF is actively (and very efficiently!) participating in the Data Carousel activities
- ❑ Complete Run2 RAW (~18.5PB) reprocessing for long lived particles searches (DRAW_RPVLL) , staged & reprocessed, January~April, 2020
 - ❑ run at real scale: finished on time, no complains from data preparation group.
 - ❑ Much less disk space occupied (~2PB)
- ❑ Run in data carousel mode on AOD : they take 1/3 of ATLAS DISK space, move them to tape !?
 - ❑ Current goal : run derivation production in Data Carousel mode for Y2016-Y2018 datasets, produce DAOD_PHYS and DAOD_PHYSLITE for physics analysis
- ❑ Tape is becoming more and more important in the ATLAS reconstruction workflows

Highlights (IV) : HPC resources



1. Risorse HPC : grande attenzione di ATLAS verso la possibilita' di utilizzare queste risorse (lo sforzo su AthenaMT va anche in questa direzione)
 - 1.1. il successo dipende fortemente dalla configurazione delle macchine (open/closed), dalla possibilita' di intervenire nel processo di definizione del progetto e dell'architettura e dalla disponibilita' di persone esperte dedicate
 - 1.2. su risorse 'open' (accesso dei wn alla rete esterna) si possono girare molti workflows di ATLAS
 - 1.3. Grant su VEGA (EuroHPC in Slovenia) : copre circa 50% del pledge ATLAS, quasi tutti i workflows (anche se principalmente MC)
2. In Italia ATLAS ha partecipato positivamente al grant CINECA 2020
 - 2.1. non chiaro se riusciremo a partecipare al nuovo grant Marconi 100 (non esiste un build del software ATLAS per Power9 serve manpower per costruirlo)

Highlights (V): new analysis model

1. Attualmente l'analysis model prevede ~80 diversi streams (DAOD) basati su skimming-menu diversi e con diverso contenuto a seconda dell'analisi
2. Analysis Model completamente diverso per run3 (e preparazione run 4)
 - 2.1. Solo un tipo di output (DAOD_PHYS) unskimmed con tutte le informazioni principali per il ricalcolo delle correzioni e sistematiche
 - 2.2. DAOD_PHYSLITE, pre-calibrati, possono essere riprodotti a partire da DAOD_PHYS
 - 2.3. open the discussion for analysis facilities

	MC size/event (KB/event)	Data size/event (KB/event)	Nominal MC events/year in run 3 (billion) from CDR*	Nominal data events/year in run 3 (billion) from CDR*	Total size/year/ version (PB)
PHYS	40	30	25	10	1.3
PHYSLITE	15	10			0.475

Default in Run4

Stima : assumendo 4 copie complete di DAOD_PHYS si passa da 60 PB (modello attuale) a ~ 32 PB (nuovo analysis model)

3. Produzione di DAOD_PHYS/PHYSLITE attraverso tape carousel !

Schedule

Intenso piano di lavoro per la comunita' computing per il 2021/2022

Data sample	Activity	2021				2022			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Run 2 Data	DAOD Production and User Analysis								
	DAOD_PHYS Production								
	Reprocessing in Release 22								
Run 2 MC	New Production for Ongoing Analyses								
	DAOD_PHYS Production								
	Reprocessing in Release 22								
2022 Data	Tier-0 Reconstruction								
	Partial Reprocessing								
	Full Reprocessing								
	Delayed stream Reconstruction								
	DAOD Production and User Analysis								
2022 MC	Generation								
	Simulation								
	Reconstruction								
	Re-Reconstruction								
	DAOD Production and User Analysis								
2023 MC	Simulation								
Upgrade MC	Production and Analysis								
2022 Ions	Tier-0 Reconstruction								

1. Task principale attualmente e' la preparazione del reprocessing con l'ultima versione del codice di ricostruzione (Athena rel 22) di tutti i dati e MC del Run2
2. In partenza la preparazione evgen per Run3
3. Novembre finalizzazione Run3 reco configuration
4. novembre/dicembre simulazione ricostruzione

ATLAS		2020			2021		2022		
		C-RSG recomm.	Pledged	Used	C-RSG recomm.	Pledged	Request	2022 req. /2021 C-RSG	
CPU	Tier-0	411	496	569	525	525	550	105%	550
	Tier-1	1057	1129	1338	1170	1243	1356	116%	1300
	Tier-2	1292	1359	2213	1430	1497	1656	116%	1588
	HLT	n/a	n/a	871	n/a	n/a	n/a	n/a	n/a
	Total	2760	2984	4991	3125	3265	3562	114%	3438
	<i>Others</i>			282					
Disk	Tier-0	27.0	27.0	25.0	29.0	29.0	32.0	110%	32
	Tier-1	88.0	99.0	93.0	105.0	116.3	121.0	115%	116
	Tier-2	108.0	108.0	108.0	130.0	127.2	148.0	114%	142
	Total	223.0	234.0	226.0	264.0	272.5	301.0	114%	290
Tape	Tier-0	94.0	94.0	83.0	95.0	95.0	120.0	126%	120
	Tier-1	221.0	225.0	160.0	235.0	241.2	272.0	116%	272
	Total	315.0	319.0	243.0	330.0	336.2	392.0	119%	392

ATLAS C-RSG recommended 2022

	ATLAS 2022 pledge	Italy share 2022	ATLAS IT 2022 pledge	Richieste 2022	Richieste 2022 (k€)	ATLAS 2021 pledge	Italy share 2021	ATLASIT 2021 pledge
CPU T1 (HS06)	1300000	0,09	117000	11700	117000	1170000	0,09	105300
Disco T1 (TB)	116000	0,09	10440	990	118800	105000	0,09	9450
Tape T1 (TB)	272000	0,09	24480	3330	59940	235000	0,09	21150
Totale (k€)	295740							

- ❑ Richieste calcolate rispetto al pledge del T1 (non rispetto alle risorse effettive attualmente disponibili)
- ❑ Mantenuto lo stesso share per ATLAS-IT (9%)
- ❑ Obsolescenza non inclusa
- ❑ Fattori di conversione : 120 €/TBn (disco), 10 €/HS06 (cpu), 18 €/TBn (tape)

ATLAS C-RSG recommended 2022

	ATLAS 2022 pledge	Italy share 2022	ATLAS IT 2022 pledge	ATLAS 2021 pledge	Italy share 2021	ATLASIT 2021 pledge
CPU T2 (HS06)	1588000	0,09	142920	1430000	0,09	128700
Disco T2 (TB)	142000	0,07	9940	130000	0,07	9100

- ❑ Mantenuto lo stesso share ATLAS-IT (9% CPU e 7% disco)

- ❑ Storage: richiesta 2020 era 0.9 PB e forti preoccupazioni per il prezzo storage ed 'esiguita" della gara.
 - ❑ Gara unica da 2.2 PB per tutte le risorse disco 2020 (0.9 PB) + 2021 (1.3 PB) .
 - ❑ La commissione ha analizzato le offerte, risparmio del ~37% rispetto alla base gara : ~ 90 euro / TBn rispetto ad un riferimento di 140 euro / TBn (iva inclusa). Incartamento inviato in amministrazione centrale
 - ❑ Acquisto singolo a Frascati per ulteriori 600 TB in coda alla finalizzazione della gara comune (espansione).
 - ❑ Risorse idealmente online entro fine anno

- ❑ CPU : si e' deciso di spostare l'acquisto delle risorse 2020+2021 al 2021 in attesa attivazione convenzione CONSIP per CPU (procedura piu' semplice).
 - ❑ Convenzione attivata fine 2020 : risparmio stimato (non ancora verificate le effettive performance delle macchine) ~10-20% (riferimento : 10 euro / HS06 iva inclusa)
 - ❑ leggero ritardo per l'attesa sblocco sj 2021, i siti stanno procedendo all'acquisto.
 - ❑ Frascati ordine partito, Milano e Roma1 in fase di finalizzazione
 - ❑ Risorse online entro fine anno

Richieste dettagliate T2 : CPU

CPU	Frascati	Milano	Napoli	Roma1	Totale	Pledged
Stima dopo acquisti 2021 (**) (HS06)	36812	35781	34097	35204	141894	128700
Obsolescenza 2022 (HS06) (***)	0	8433 (8433/0)	12667 (6900/5767)	13800 (7666/6134)	34900 (23000/11900)	
Delta 2022 (HS06)	0	0	0	0	0	
Totale richiesta (HS06)	0	8433	12667	13800	34900	
Totale CPU nel 2022 (HS06)	36812	35781	34097	35204	141894	142920

(**) Ancora accessi a Napoli 11 kHS06 obsoleti da 3 anni (in attesa di rimpiazzo PON). Acquisti CPU su CONSIP ancora in corso (39.2 kHS06 stimati)

(***) in parentesi (obsolescenza primo semestre 2022 / obsolescenza secondo semestre 2022)

- ❑ Lo scorso anno chiesti 11 kHS06 per Napoli (formalmente messi su PON nel referaggio 2019 ma non ancora arrivati) che erano fuori garanzia. Primi 4320 HS06 da Ibisco on line a Napoli
- ❑ Risparmio di gara CPU atteso ordine 20%. La richiesta di anticipo dell'anno scorso e il risparmio di gara 24 compensano il delta di quest'anno

Richieste dettagliate T2 : disco

STORAGE	Frascati	Milano	Napoli	Roma1	Totale	Pledged
Stima dopo acquisti 2021 (*)	2839	2432	2355	2432	10058	9100
Obsolescenza 2022 (TB)	998	998	922	998	3916	
Delta 2022 (TB)	0	0	0	0	0	
Totale richiesta 2022 (TB)	998	998	922	998	3916	
Totale disco nel 2022 (TB)	2839	2432	2355	2432	10058	9940

(*) Gara storage unica (2020+2021) 2.2 PB appena chiusa

- ❑ Alcune imprecisioni nelle richieste 2021 (valutazione non corretta di una obsolescenza e conteggio inconsistente dei localgroupdisk in alcuni siti) ha portato ad una richiesta maggiore del pledge che ora compensiamo
- ❑ Scorporati i localgroupdisk (180 TB) per sito indispensabili per l'analisi da parte degli utenti italiani
- ❑ Tutta l'obsolescenza dischi nella prima meta' del 2022

Richieste finanziarie T2

	Frascati	Milano	Napoli	Roma1	Totale
Richiesta CPU (€)	0	84330	126670	138000	349000
Richiesta Disco (€)	119760	119760	110640	119760	469920
Richiesta totale (€)	119760	204090	237310	257760	818920
Overhead server (€) (*)	8383	14286	16612	18043	57324
Totale	128143	218376	253922	275803	876244
Totale (no OH per sito PON)					859633
Totale (no sito PON)					622323

(*) Algoritmo per calcolo overhead : “server” : 7% del totale CPU + disco

❑ Fattori di conversione : 120 €/TBn (disco), 10 €/HS06 (cpu), 18 €/TBn (tape)

	Frascati	Milano	Napoli	Roma1	Totale
Overhead server (k€) (*)	8383	14286	16612	18043	57324
Overhead rete (k€) (*)	5988	11048	13132	14268	44436

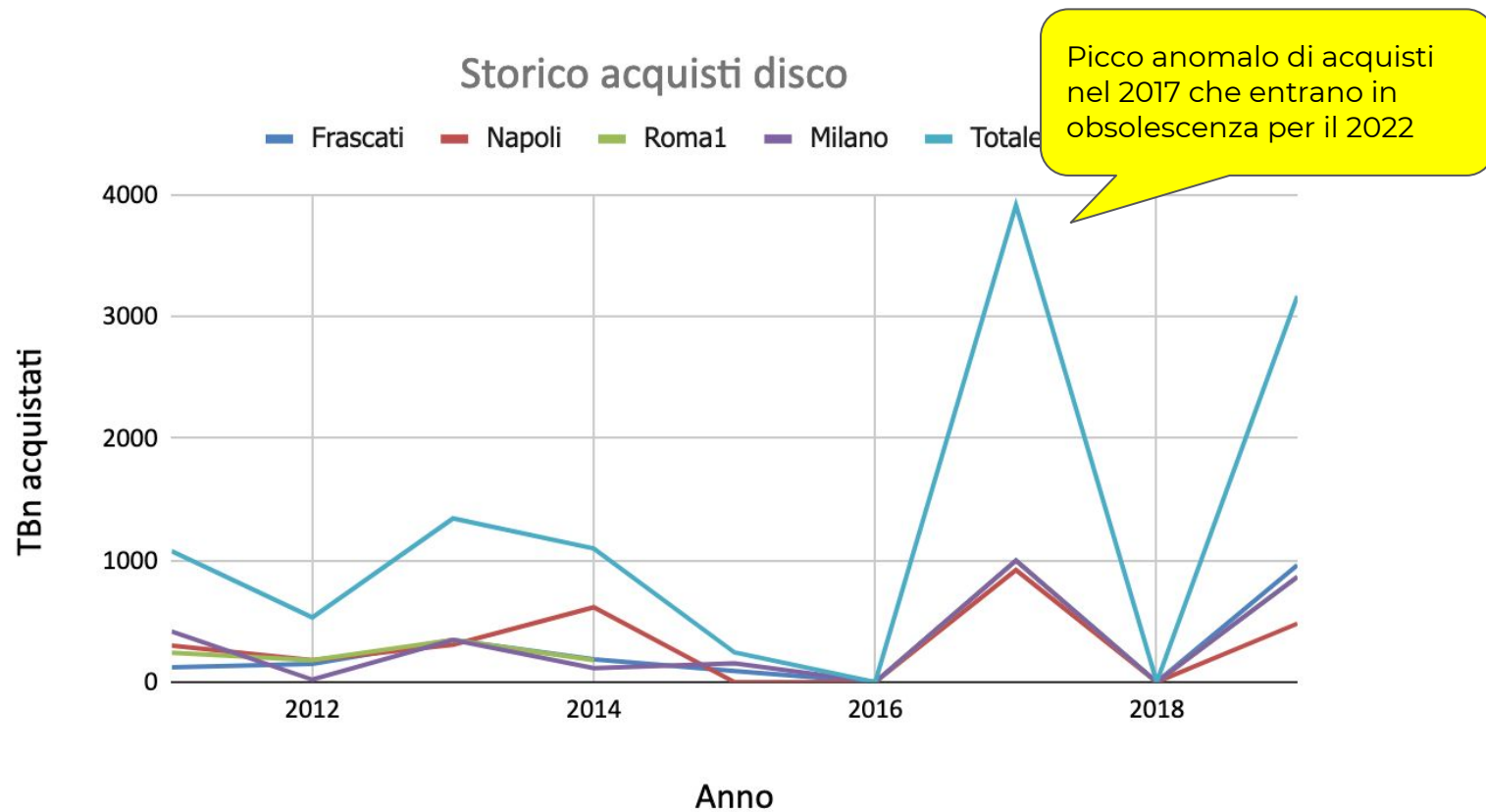
- ❑ Overhead 'server' importanti per il funzionamento dei servizi dei siti.
 - ❑ cifra minima critica stimata 10 k€ per Milano e Roma1 (tot. 20 k€)
- ❑ Overhead rete (44.5 k€) (*) non richiesti in questo referaggio ma singoli siti hanno preparato richieste specifiche in CCR per ~ 17 k€ totali
 - ❑ Roma1 : 13.8 k (3 Switch 10Gbps con uplink a 40 Gbps)
 - ❑ Milano : 3 keuro (3 switch ToR)

(*) Algoritmo per calcolo overhead :

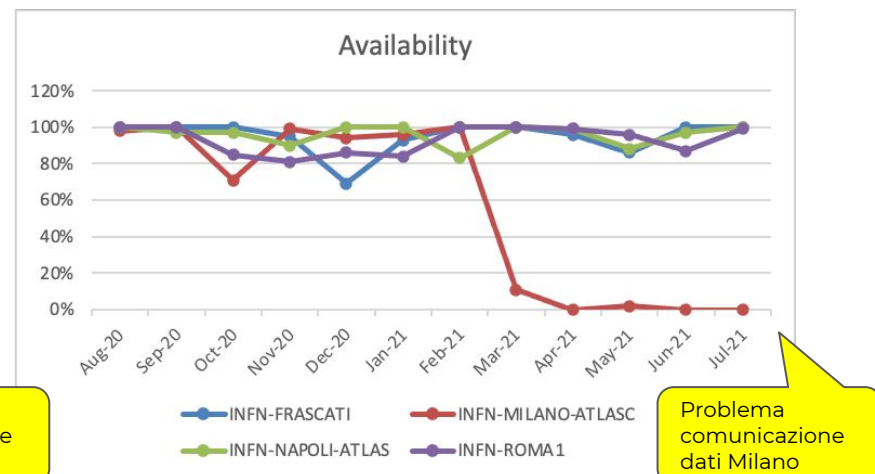
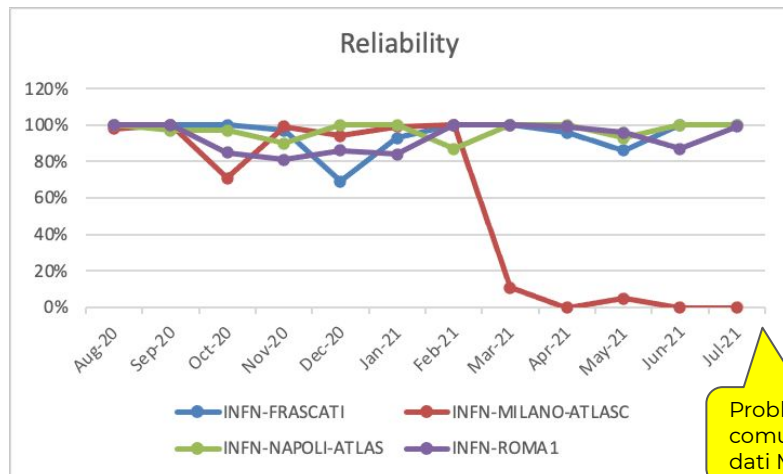
- ❑ "server" : 7% del totale CPU + disco
- ❑ "rete" : 6% CPU e 5% disco

Conclusioni

1. I siti italiani hanno lavorato in modo efficiente per tutto il periodo agosto 2020-luglio 2021 garantendo le risorse promesse ad ATLAS. Siti affidabili e con efficienza paragonabile ai migliori centri ATLAS
 - 1.1. Milano ha funzionato molto bene ma su workflows meno i/o intensive. Riattivazione analisi in corso (dopo l'upgrade degli apparati di rete)
2. Varie attività in corso dentro l'esperimento per ottimizzare l'uso delle risorse (G4 optimisation, nuovo analysis model, tape carousel, AthenaMT, integrazione di risorse HPC)
3. Inteso piano di lavoro per il 2021/2022 : reprocessing di dati e MC Run2, preparazione nuovi MC per Run3 e processamento dei dati Run3
4. Presentata la situazione delle risorse disponibili (gare) e le richieste per il 2022



Affidabilit  dei siti italiani



	Frascati	Milano	Napoli	Roma
Average Reliability (01/08/2020 - 31/07/2021)	95%	94% (*)	97%	93%
Average Availability (01/08/2020 - 31/07/2021)	95%	94% (*)	96%	93%

(*) (01/08/2020 - 28/02/2021)