

Stato ed evoluzione dell'infrastruttura di calcolo INFN

Tommaso Boccali
Gian Paolo Carlino

Luca dell'Agnello

Workshop CCR - Rimini, 14 giugno 2018

Premessa (1)

- Evoluzione delle necessità di calcolo dominata (finora) da LHC
 - Es. al Tier-1 70-80% delle risorse per LHC vs. ~30 altri esperimenti
- Nuovi attori in campo nei prossimi anni con richieste significative
 - Es. Belle2, Juno, KM3, CTA, Euclid, SKA?
- Ma con HL-LHC (2026 -) prevista crescita considerevole richieste
 - Secondo stime assolutamente non definitive prevista crescita di x20
 - Fattore 4 con »flat budget« (i.e. 20% incremento annuo risorse a parità di spesa)
 - Effort necessario da parte comunità WLCG per ridurre tale crescita
 - Software, riduzione del volume dei dati, consolidamento dello storage,
- La nostra infrastruttura di calcolo può reggere tale crescita?

Premessa (2)



- All'interno del C3S (Comitato di Coordinamento Attività calcolo scientifico dell'INFN) è stato effettuato un survey sullo stato e la consistenza dell'infrastruttura di calcolo scientifico dell'INFN
- Metodologia di somministrazione
 - Questionario inviato ai direttori
 - Aggregazione a livello di data center: siti co-locati si riflettono in unico questionario
 - 1 data center CNAF che ospita Tier-1, Tier-2 di LHCb e Tier-3
 - 2 data center per Tier-2 LNL-PD
- Raccolta dati: Giugno – Settembre 2017
 - Interazioni successive con i referenti dei centri di calcolo
 - Alcuni dati raccolti in extremis
 - 33 risposte «pervenute», 29 «validate»

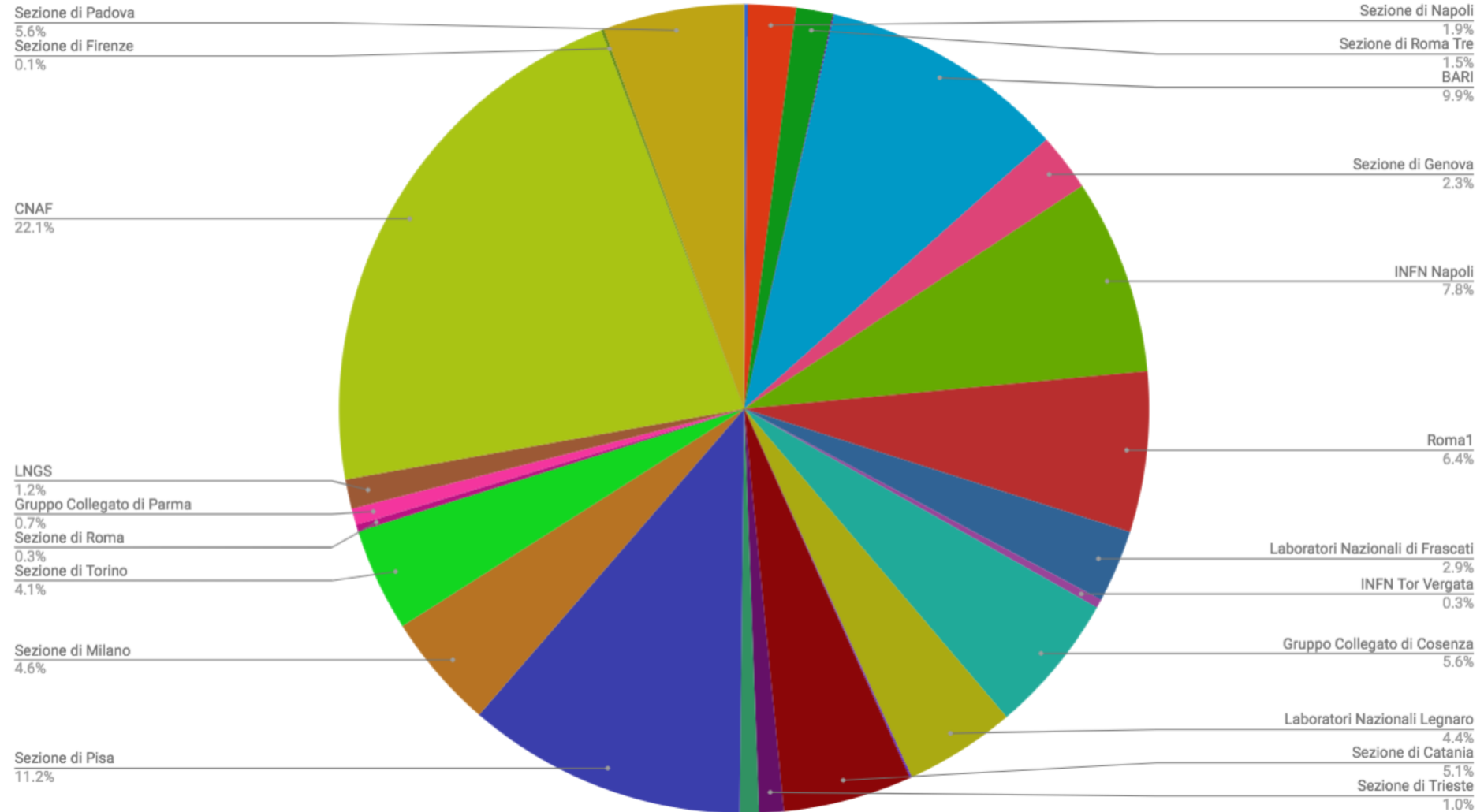
Struttura del questionario

- Definizione del bacino di utenti (esperimenti INFN o esterni);
- Caratteristiche dell'infrastruttura fisica del data center e sua espandibilità;
- Consistenza e caratteristiche delle risorse di CPU, disco e tape installate;
- Tipologia delle connessioni di rete LAN e WAN;
- Consistenza del man-power;
- Eventuali criticità;
- Costi di gestione e manutenzione (inclusa l'eventuale convenzione con altro ente ospitante).
- *Da mandato, non sono state analizzate le performance dei siti e le attività svolte.*

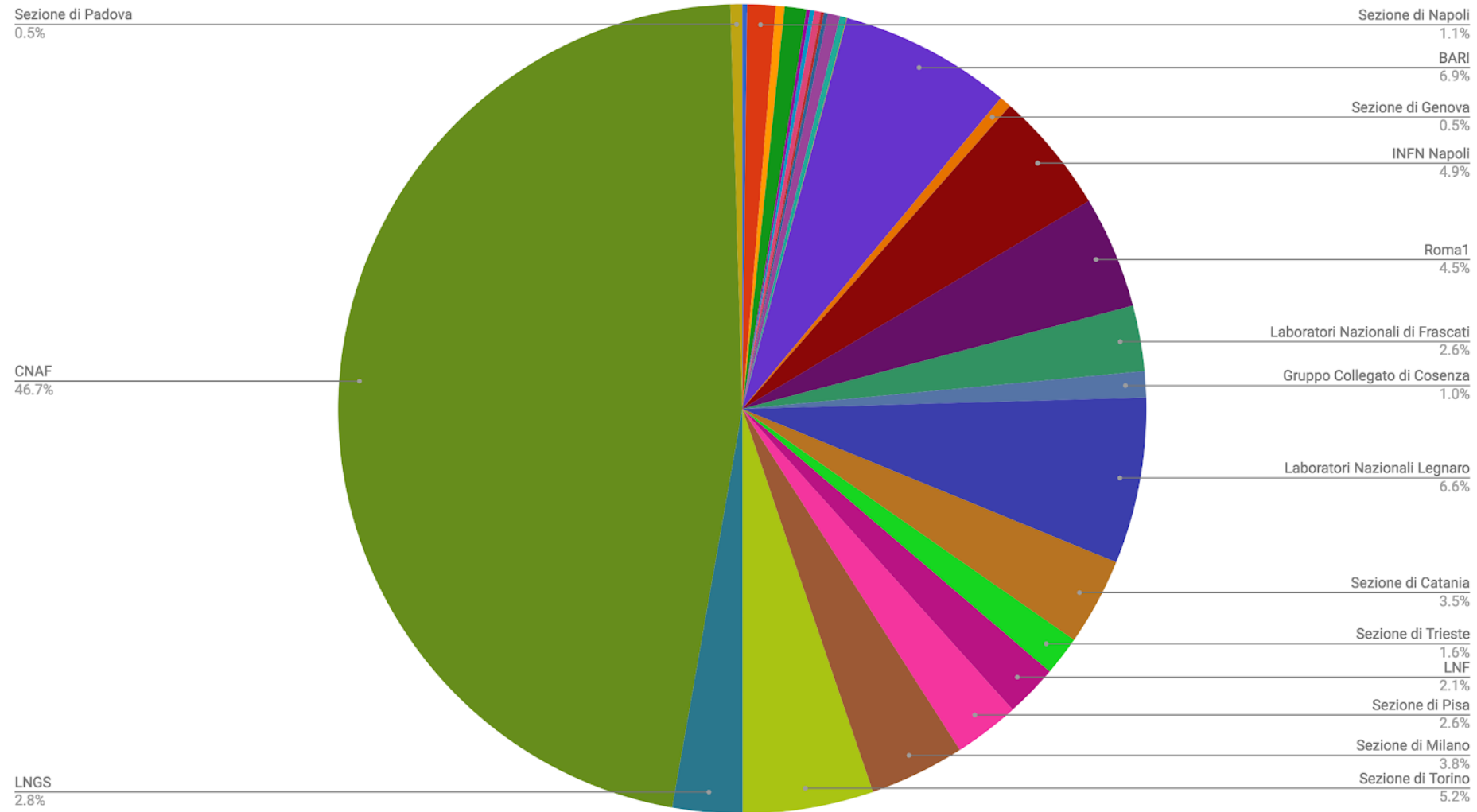
Risultato sintetico

- 29 data center censiti
 - Non considerati installazioni per servizi/sistemi dedicati (es. sistemi di acquisizione di esperimenti, online e trigger)
- CPU: ~70.000 core (corrispondenti a ~800 kHS06)
 - ~55% core coperti da contratto di manutenzione
 - Presenti alcune farm HPC (~400 Tflops)
- Disco: ~57 PB storage disco installato
 - ~91% coperto da contratto di manutenzione
- Tape: ~97.5 PB storage tape installabile (~75.3 PB usati)
- 92 FTE su calcolo (61 FTE supporto tecnico)
 - 32% personale di altri enti (e.g. Università)
- Grande polarizzazione dimensioni siti
 - In prima approssimazione Tier-1/2 di WLCG vs. gli altri

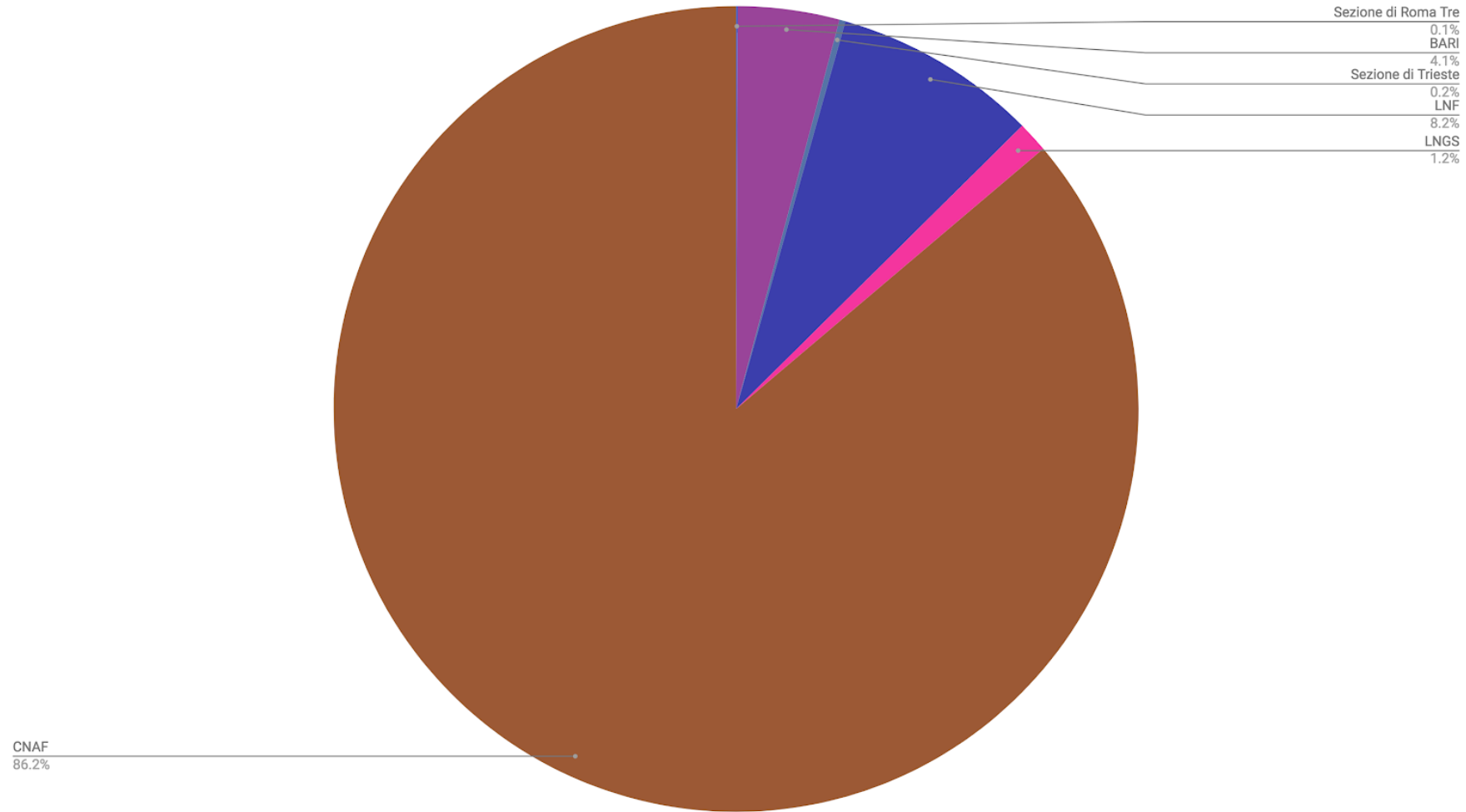
Slot di Calcolo



Disk



Tape

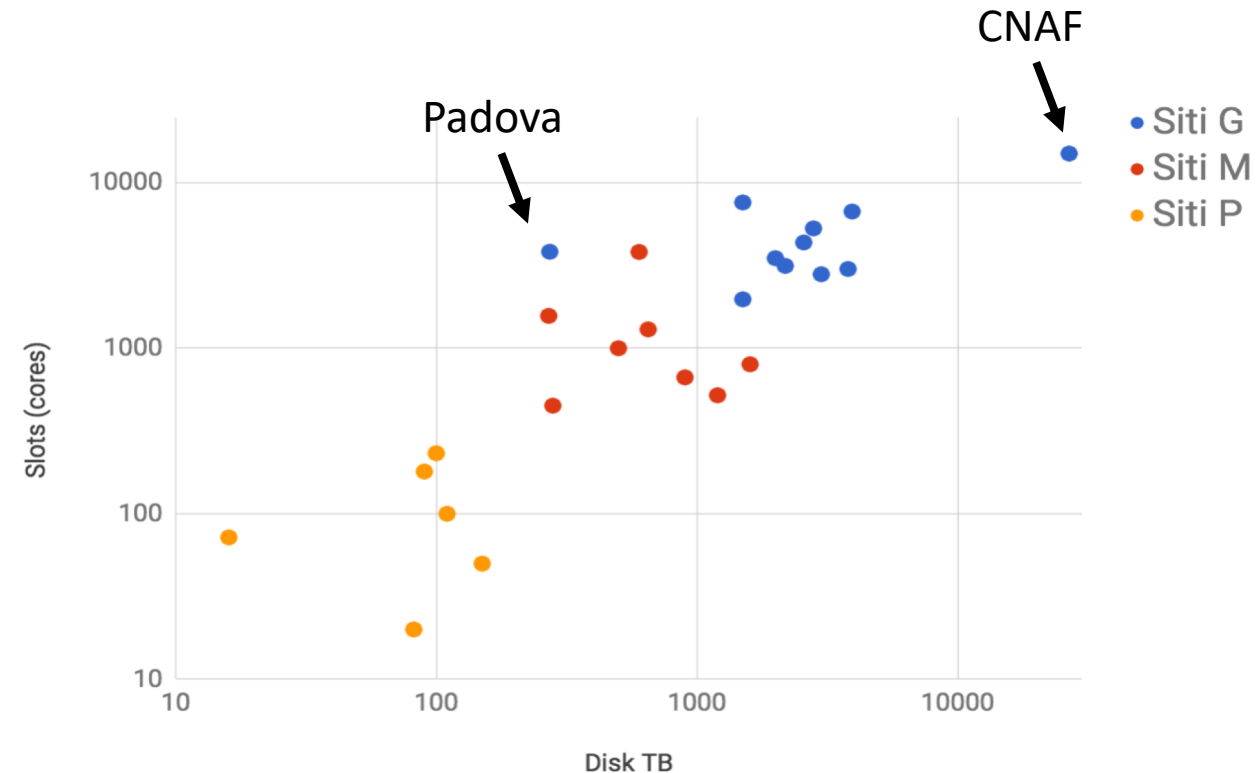


Siti piccoli, medi e grandi: le soglie

- Sito “grande” - data center con farm > 1000 core OR storage > 750 TB-N.
 - Soglia equivalente a circa metà del disco e CPU installate in un Tier-2 medio italiano di WLCG.
 - Rientrano tutti i firmatari di MoU WLCG
- Sito “piccolo” - data center con farm < 200 core OR storage < 100 TB
 - Risorse installabili in 4 RU totali e con potenza elettrica necessaria inferiore ai 2 kW
- Sito “medio” - I rimanenti
 - Sito INFN Padova (storage < 750 TB-N) promosso a “grande” in quanto parte del Tier-2 di Legnaro

Siti piccoli, medi e grandi: la lista

- 11 siti grandi, 8 medi, 10 piccoli
- CNAF, Torino, Pisa, Milano, Legnaro, Padova, Roma1 (Tier-2), Bari, Napoli (Tier-2), Catania, Frascati (Tier-2)
- LNGS, Napoli, Roma3, Genova, Cosenza, Trieste, Frascati - KLOE, Parma
- Milano Bicocca, LNS, Pavia, Roma2, Perugia, Tor Vergata (2 siti), Ferrara, Firenze, Roma1 (non Tier-2)

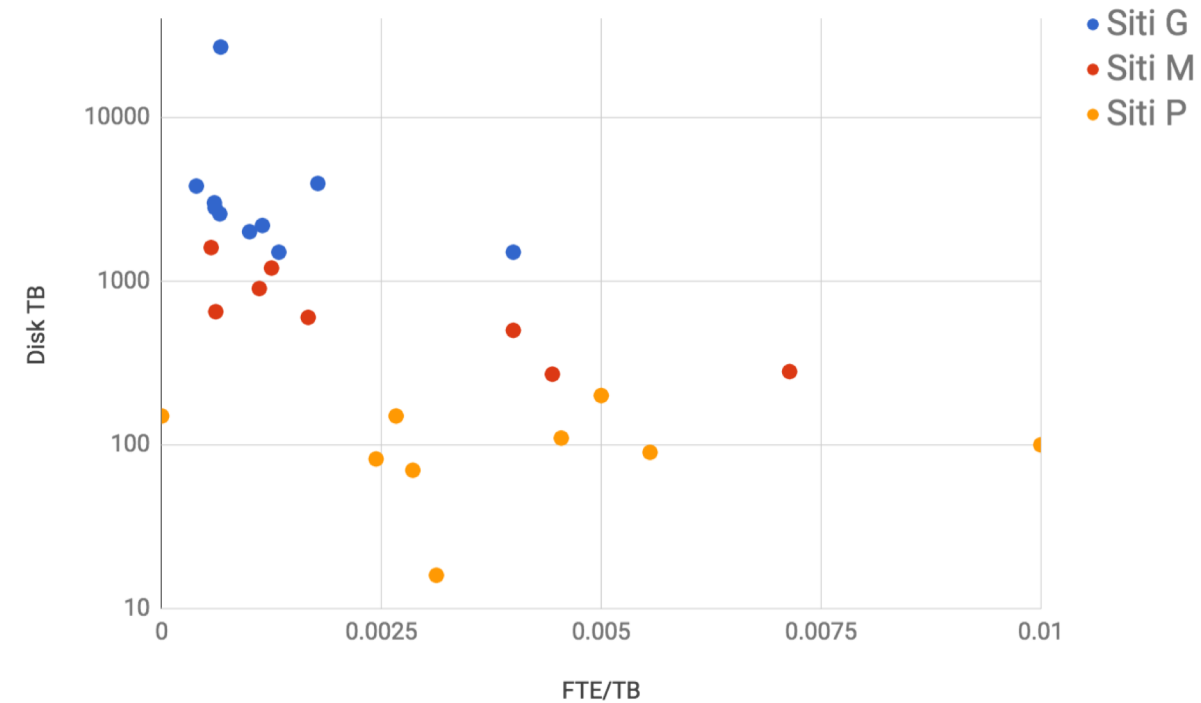
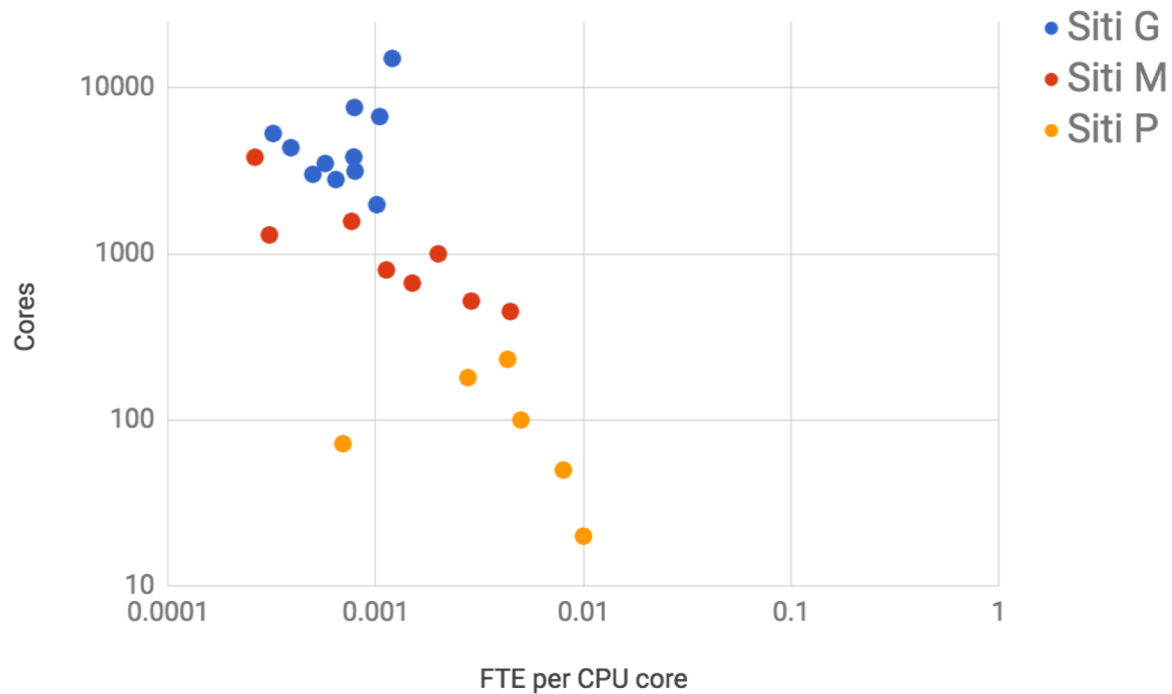


Distribuzione risorse e FTE

- Nei centri medi+grandi risultano concentrati il 95% dei core ed il 96% del disco; la maggior parte dei tape è ovviamente al CNAF
- 15% degli FTE (sia tecnici sia totali) è presente nei “siti piccoli”
 - $\sim \frac{1}{3}$ degli FTE non-INFN è nei siti “piccoli”.
- Stima «grossolana» del peso degli FTE effettuata tramite il rapporto tra FTE e risorse («efficienza»)
 - «Efficienza» (ovviamente) migliore nei siti grandi

<i>Efficienza</i>	grandi	medi	piccoli
<i>FTE/Slot</i>	0.0011	0.0017	0.0155
<i>FTE/Disk_TB</i>	0.0009	0.0017	0.0040

FTE vs. risorse



- Nota: la SLA non è omogenea
 - $SLA_{Tier-1} > SLA_{Tier-2} > SLA_{altri}$ (best effort?)

Potenza elettrica

- Consumo totale comunicato: ~ 3.4 MW
 - 2.96 MW (1.3 MW solo al CNAF) nei centri grandi, 204 kW nei siti medi e 187 kW in quelli piccoli
 - Costo totale stimato: 5 M€/anno (0,17 €/kWh)
 - Metà dei consumi coperti da convenzioni con Università ospitante:
 - Convenzione a consumo: 61 kW
 - Convenzione forfettario: 1.6 MW (risparmio stimato: 2.4 M€/anno)
- L'utilizzo di centri di calcolo con corrente non direttamente pagata da INFN è perciò valutabile in un risparmio di ~ 2.4 M€/anno.
 - **Attenzione: situazione potrebbe cambiare rinegoziando le convenzioni**

Dimensione data center ed espandibilità

- Attualmente sono installati 504 rack che possono aumentare, con l'infrastruttura attuale, fino a 752.
- La superficie utile totale attuale è pari a 4400 mq. espandibile, senza interventi “drastici” (come la realizzazione di nuove sale), fino a 4800 m²:
 - espandibilità totale: $4800/4400 \times 752/504 = 160\%$.
- Ritenendo valida la legge di Moore (ottimistico?) anche per il prossimo decennio (+20%/anno), da oggi (2018) al 2026 i centri INFN potrebbero scalare significativamente (a meno di altri limiti)
 - $1.2^8 \times 1.6 = 6.999 \sim 7$
- I centri piccoli hanno il 15% dello spazio per apparati IT a fronte del 5% delle risorse.
 - dimensione minima allocata indipendentemente dalle risorse installate;
 - minore densità delle risorse rispetto ai grossi centri, acquisite puntando anche alla compattezza dei sistemi.
- **Caveat**
 - Non ovvio impatto su infrastruttura esistente aumento di un fattore 7 (es. densità di calore nei rack)
 - Es. al CNAF non possibile avere più di 20 kW/rack
 - Da valutare inoltre effettiva espandibilità siti piccoli
- Ulteriore possibilità è il nuovo data center (3000 mq) in progettazione nell'area dell'ex Manifattura Tabacchi a Bologna (*tecnopolo*)

Alcuni spunti di riflessione (1)

- I siti grandi (fra cui il CNAF, molto grande e con SLA più stringente) hanno costi di gestione in termini di FTE/risorsa simili.
 - Da capire se ulteriore condensazione possa portare risparmi significativi
 - Da considerare comunque che alcuni Tier-2 non pagano, al momento, la corrente elettrica
- Per crescere nei siti grandi oltre il fattore 7 (ma probabilmente anche prima) necessari cambiamenti drastici nell'infrastruttura
 - Anche CNAF potrebbe avere problemi senza Tecnopolo
- Non ovvio come gestire la situazione per il personale

Alcuni spunti di riflessione (2)

- Siti medi: tipicamente dipendenti da particolari situazioni locali
- Siti piccoli: è conveniente mantenere in vita infrastrutture non molto efficienti (FTE/risorse)?
 - 5% risorse totali....
 - Contributo significativo in FTE non INFN
 - Da valutare anche in termini di affidabilità ed uso
- Siti piccoli sostituibili con una Cloud nazionale?
 - Unica o federata
 - Probabilmente da considerare creazione di gruppo di gestione a livello nazionale per coinvolgere personale delle varie sezioni

Alcuni spunti di riflessione (3)

- Per HL-LHC in discussione in ambito WLCG evoluzione verso consolidamento storage (modello Data Lake)
 - Alcuni grandi centri O(10 MW) di potenza con dati (e CPU) interconnessi ad alta velocità (nx100 Gbit)
 - Implementazione a livello di infrastruttura di funzionalità gestite ora a livello applicativo
 - Es. Replica automatica dati, QoS, data curation (e.g. self-healing)
 - Razionalizzazione repliche (risparmio spazio disco?)
 - Centri satellite con CPU e cache
- Modello applicabile anche ad altri esperimenti
- Un possibile scenario a livello INFN
 - Tier-1 (tecnopolo) con storage e CPU
 - Tier-2 con CPU e cache
 - Un altro centro per DR per servizi importanti, dati di exp INFN
 - CPU opportunistica

Data lake

