Evoluzione dei modelli di calcolo per la fisica e impatto sulle reti

Tommaso Boccali
INFN Pisa

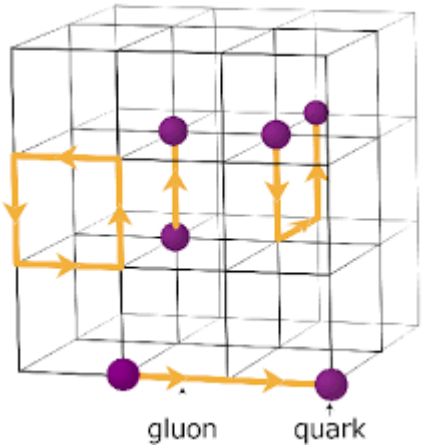
WORKSHOP GARR 2021

NET MAKERS

INFN

70 1951 2021 infn
70 ANNI
DI RICERCA
PER TRACCIARE
IL FUTURO

Esigenze e Modelli di Calcolo per la fisica



Lattice QCD, elementi finiti, simulazion

Difficile condensare in un'unica presentazione, ma possiamo distinguere le esigenze in due categorie fondamentali

1. ***il calcolo della fisica teorica***; principalmente simulazione di modelli a basso / inesistente IO, codice tight scritto in house per cui "adattabile" a sistemi non standard (GPU, KNL, …)
2. ***il calcolo della fisica sperimentale***, che può essere estremamente data intensive e e' tutt'altro che tight (milioni di righe di codice)

Qui parlo sono (di un sottoinsieme) del secondo, piu' rilevante per le reti …

Modelli di calcolo oggi - punti salienti

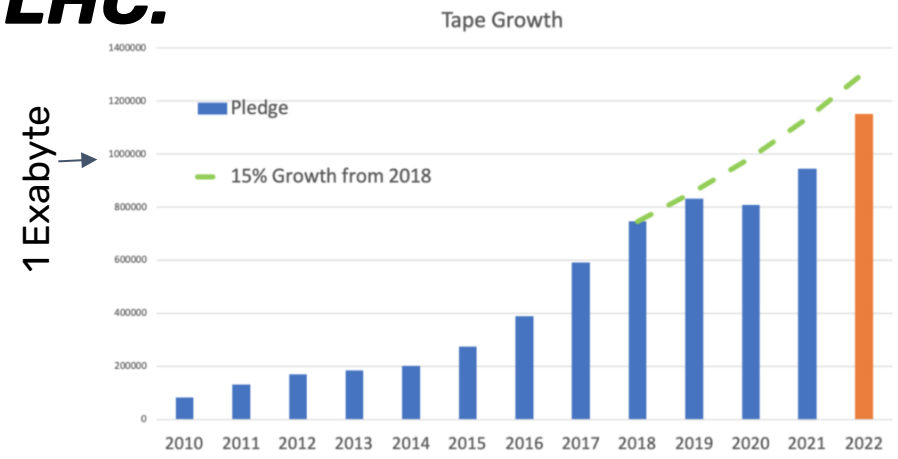
1. Ci sono apparati che “generano” dati, in grande quantita’ (l’Exabyte e’ gia’ un realta’)
2. Ci sono centri che ospitano data repositories, con una certa ridondanza, con capacita’ totale ~ Exabyte per esperimento
3. Ci sono centri che analizzano dati, con capacita’ ~ di centinaio di migiaia di CPUcores per esperimento

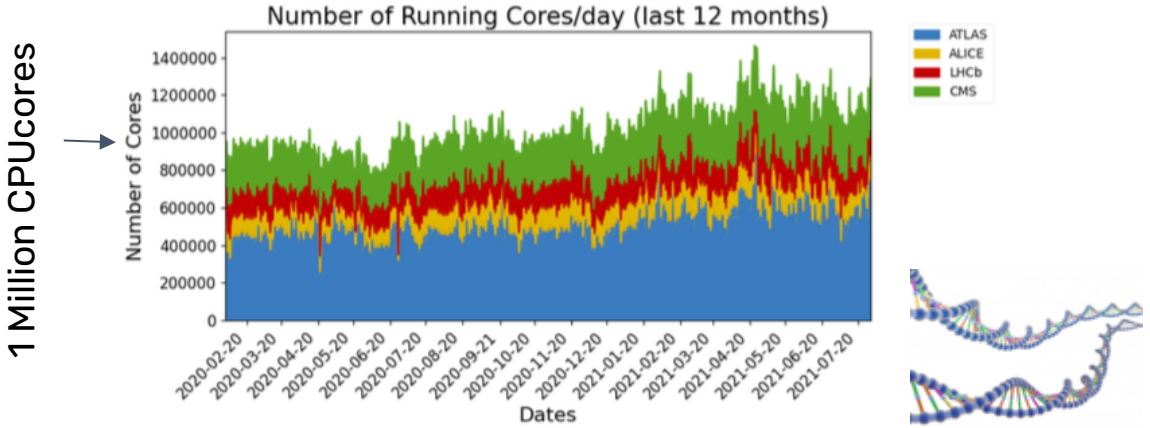
Questi sono ovviamente casi limite (LHC, SKA, Dune, CTA, Mirror Copernicus, …) ma sempre piu’ numerosi.

I Big Data sono la norma ormai nella fisica sperimentale … alcuni esempi →

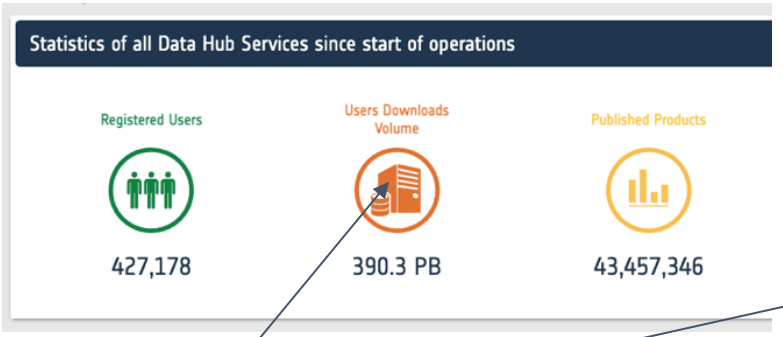
Cosa (LHC / Mirror Copernicus / Dune / SKA) ?

LHC:





Mirror Copernicus:



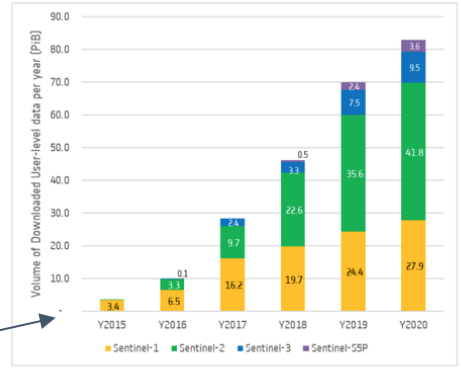


Figure 40: Total volume of user-level data downloaded per year since the start of operations from all of the four hubs, differentiated by mission

data downloaded by clients

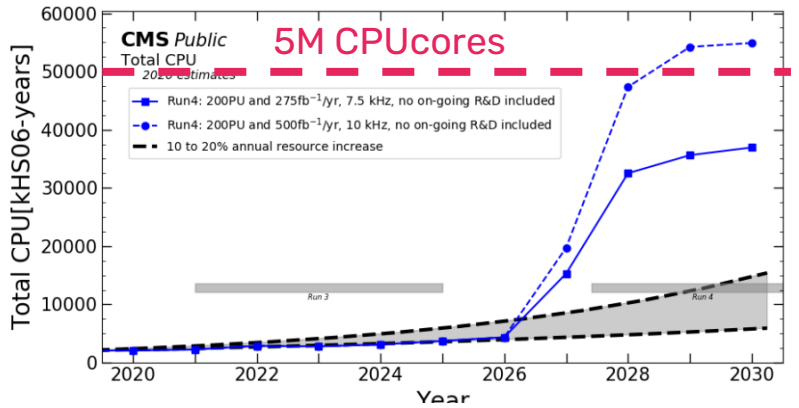
Dune: un singolo evento di supernova @ 200 TB. Il prototipo attuale “solo” 3 GB/s. Il sistema finale 80x

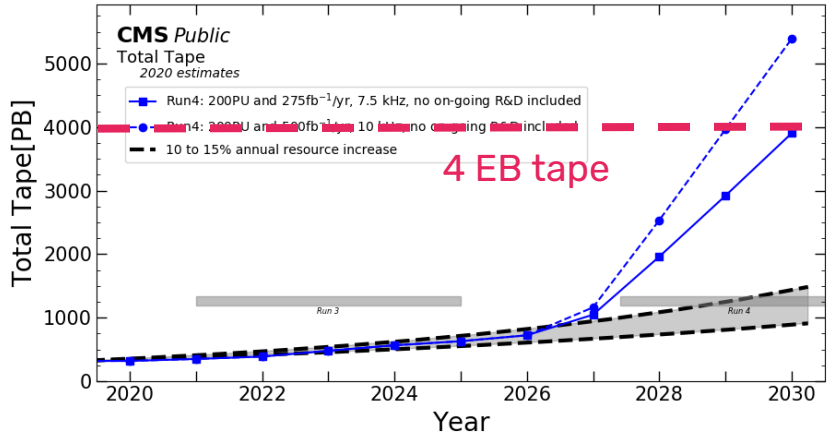
SKA: fino a 2 PB/giorno

Genomics: 1 genoma ~ 100 GB. Uno studio di popolazione di 1 M persone: 100 PB

Cosa (LHC) 202x-203x ?

- LHC ha i piani piu' definiti al momento per il prossimo decennio
- Sia l'acceleratore sia gli esperimenti subiranno un processo di upgrade; un semplice scaling direbbe che le risorse (storage, cpu) dovrebbero crescere fino a 100x
- Chiaramente impossibile ($$). Gia' da ~5 anni grosso lavoro per ridurre le necessita'
- Anche a valle di questo, ci si aspetta che per il 2030 ogni esperimento avra' necessita' di ~ 5 milioni di CPUcores e ~ 5 EB di storage (fra tape e disco)

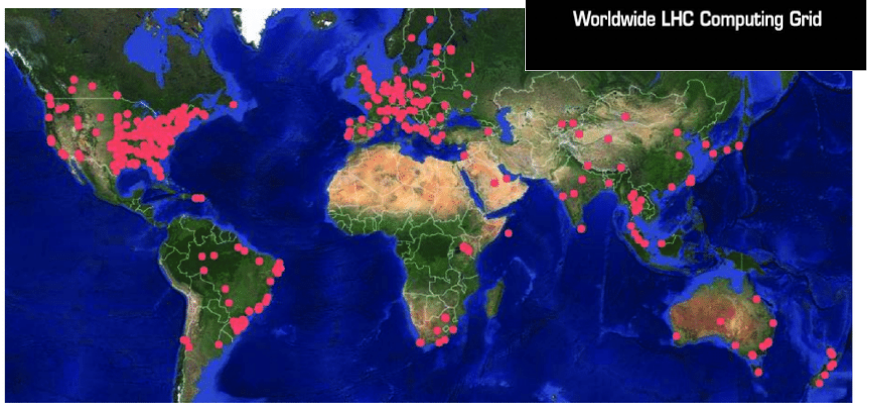




come gestirli? dove metterli? dove comprarli? comprarli?

Come (LHC / Mirror Copernicus / Dune / SKA)?

Gli esperimenti in corso e quelli pianificati si stanno uniformando ad un unico modello di gestione risorse, quello di WLCG@CERN, che e' un'evoluzione di quello che il CERN usa oggi:

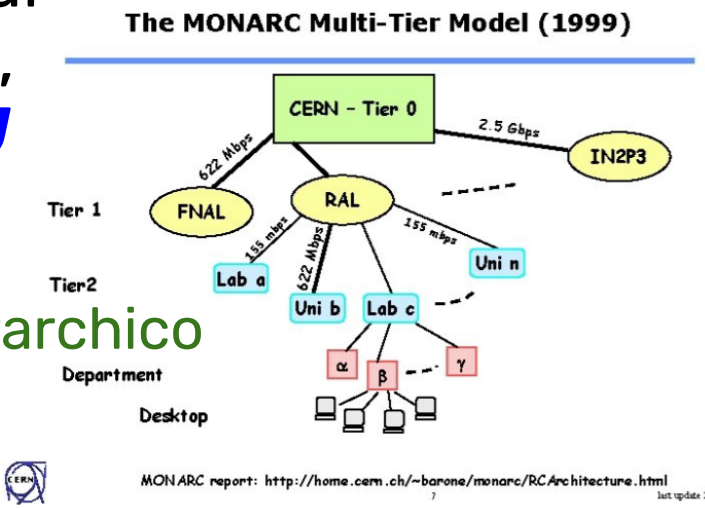


- Modello a ***data lake*** per i dati
- Risorse CPU potenzialmente slegate dagli storage sites, e a vita media anche breve
- Connessioni con ***centri HPC*** (sempre di piu') e Commercial Clouds (meno)
- Da centri di calcolo per l'esperimento X a centri di calcolo ***multi-domain*** su scala nazionale

Rationale: Mettere in sicurezza i dati, usare CPU dove costano meno, usare la rete come colla

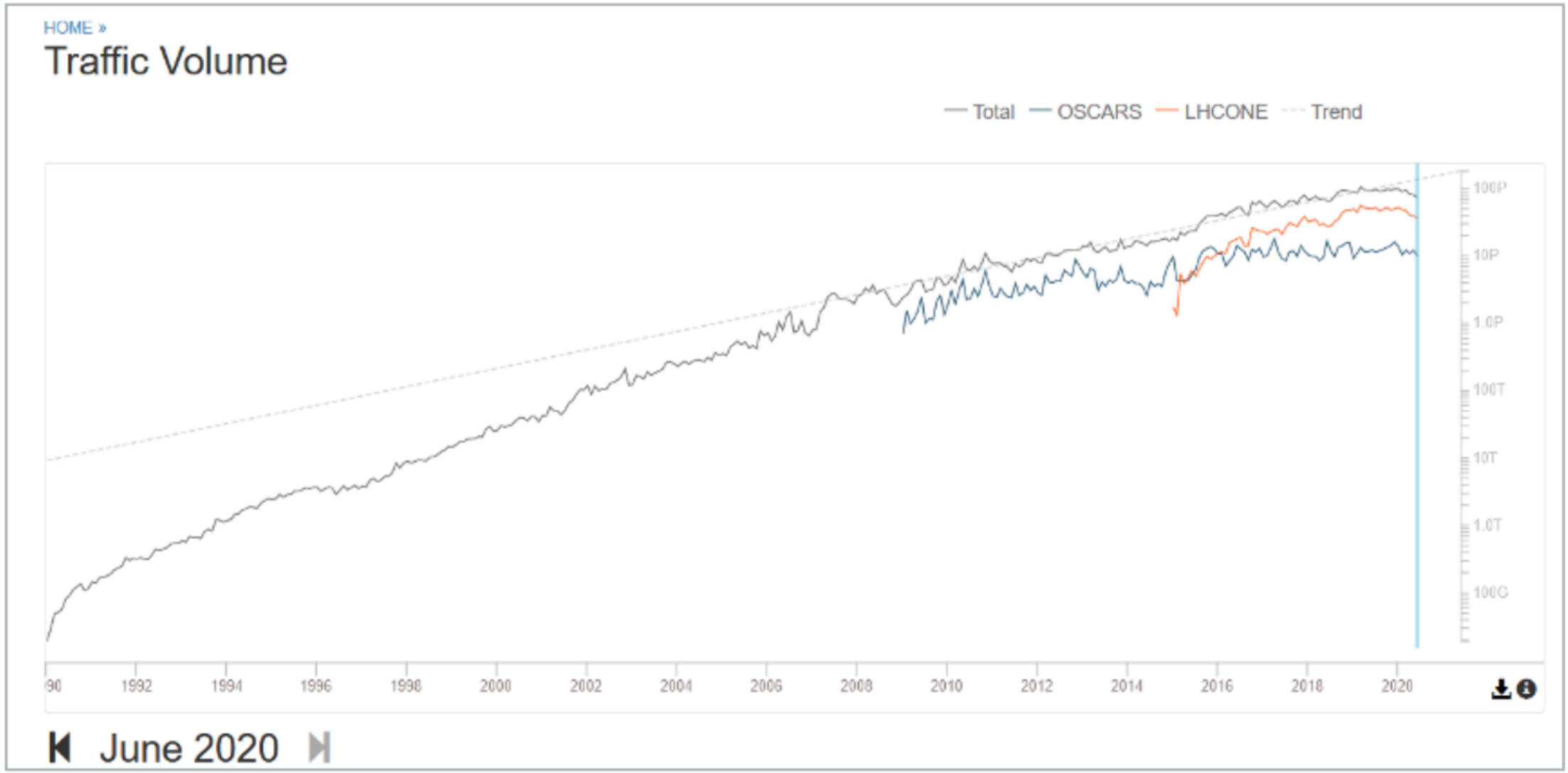
Il modello a data lake (LHC, SKA, Dune, …)

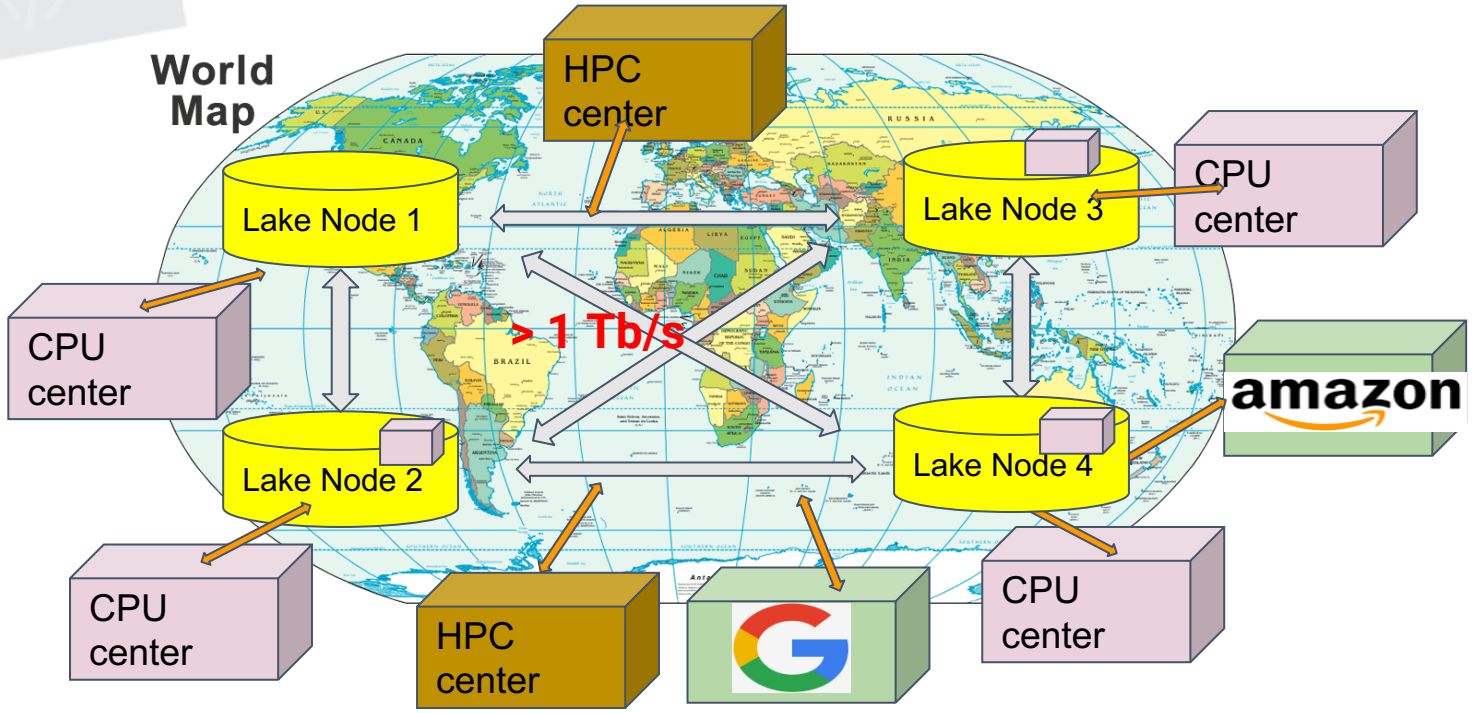
- Il modello Monarc per il calcolo distribuito (alla base di WLCG) prevedeva centri ***owned***, a ***lunga vita media***, e con uno stretto ***bilanciamento fra storage e CPU***
  - i workflow processano ***dati solo locali*** alle CPU
  - ***spostare dati costa*** e deve essere minimizzato
  - gli spostamenti sono su ***pochi link garantiti***, in modo gerarchico
  - (e' una Data-GRID)
- Pero' fa perdere molte opportunita':
  - centri solo storage o solo CPU difficilmente usabili
  - centri che siano disponibili per poco tempo difficilmente utilizzabili
  - necessita' di personale di esperimento "vicino" al centro
  - ***gia' adesso grazie all'esplosione delle reti di ricerca la locality e' sfumata (remote processing, streaming, …)***
- Datalake: ***slegare*** completamente gli aspetti storage e CPU
  - ***approccio per i dati***: sono il vero risultato degli esperimenti, tenerli al sicuro in centri owned; minimizzare le copie del singolo dato (a 1!)
  - ***approccio per le CPU***: usarle dovunque si trovino al miglior costo, anche se disponibili per poco tempo



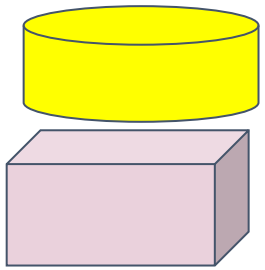
Il modello a data lake (LHC, SKA, Dune, ...)

- Il r
  Wl
  e c
  -
  -
  -
  -
- Pe
  -
  -
  -
  -
- Da
  -
  - centri owned; minimizzare le copie del singolo dato (a 1!)
  - ***approccio per le CPU***: usarle dovunque si trovino al miglior costo, anche se disponibili per poco tempo





1. Preparare una struttura di **storage centers affidabili e interconnessi** gestito da sistemi automatici; se la rete e' almeno al Tbps, la visione verso l'esterno (gli utenti) e' quella dell'**oggetto unico**
2. i centri storage possono avere CPU locali, ma non e' una richiesta
3. centri CPU owned si attaccano al nodo piu' "vicino" del lake e processano i dati o direttamente (streaming) o **mediati da caches istanziate anche automaticamente**
4. centri CPU non owned, anche a bassa vita media (un grant …) si possono attaccare nello stesso modo



Risorse Storage

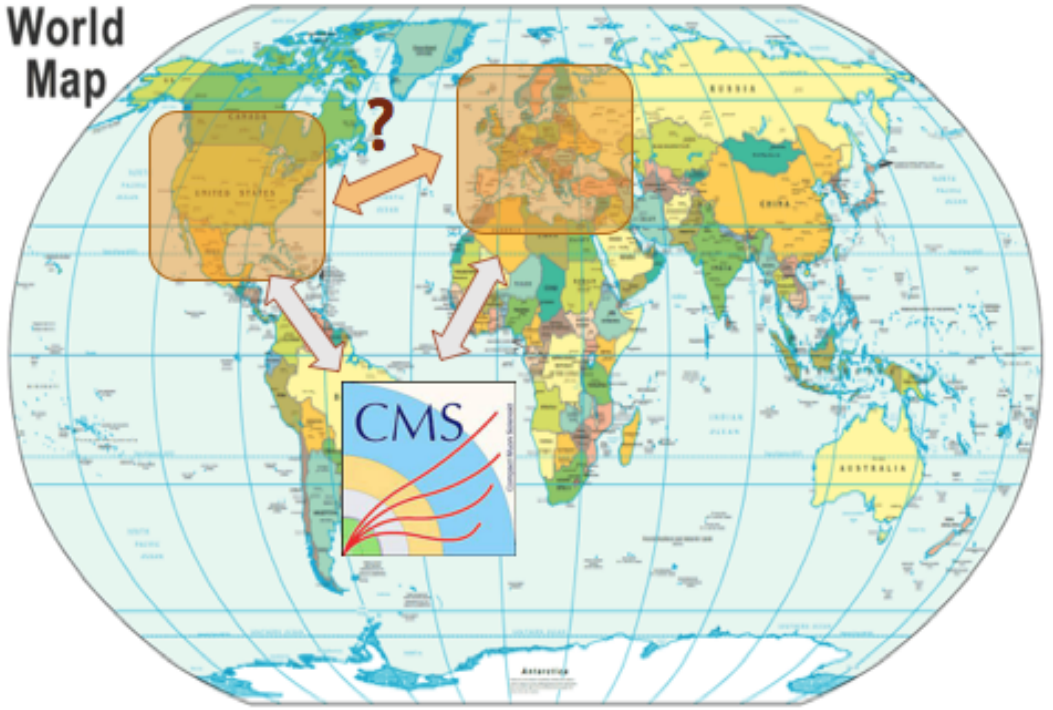
Risorse CPU

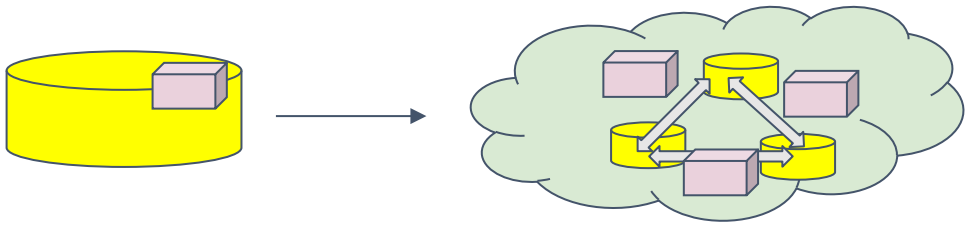
Ovviamente, un oggetto di questo tipo ha senso solo se

1. si possono stabilire **reti multi Tbps** anche su scala extra continentale
2. si possono **accendere** anche **on demand path di rete** verso centri di ricerca, HPC, Cloud Commerciali

Piu' nel dettaglio, cosa serve / cosa faremo?

- Difficilmente la visione idilliaca della pagina precedente sara' realizzata. E' piu' probabile che avremo datalake multipli (US e EU almeno) che devono pero' interoperare (politica)
- Le maggiori nazioni vorranno un nodo del datalake, ma per varie ragioni (politica) il singolo nodo sara' internamente "distribuito" con una singola interfaccia esterna; una sorta di delega di responsabilita' a tutti i livelli

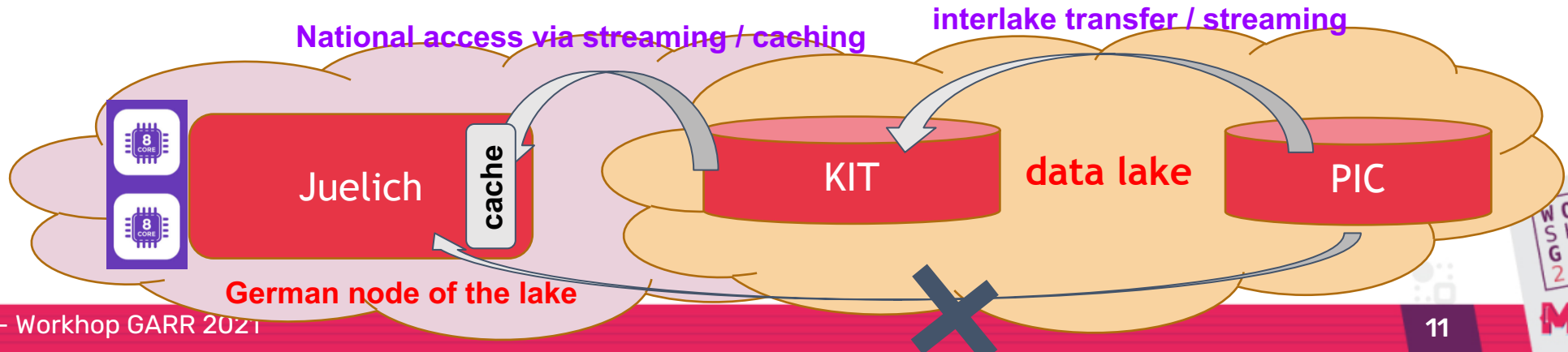




Almeno ne mondo LHC, e' probabile che il nodo della data nazione si identifichi logicamente con il Tier-1 della stessa nazione. Quello che succede fisicamente e' responsabilita' della singola nazione, e deve essere implementato mediante la sua NREN
(CNAF, PIC, KIT, RAL, FNAL, Lyon, JINR, BNL, SurfSARA, TRIUMF, ASGC, …)

Come avviene un processing a larga scala

- Il modello non e' solo di collegamento, ma anche "di responsabilita'"
  - Il centro X in Italia visto come "estensione" del Tier-1 italiano (CNAF), in cui lo user support dell'esperimento si fa carico anche di validare / dare supporto all'utilizzo
- Esempio pratico: nel caso in cui l'HPC a Juelich(DE) sia disponibile per analizzare dati presenti a PIC(ES) il flusso di dati a livello logico e'
  - Juelich e' connesso a Tbps con KIT(DE), nodo del datalake tedesco, dalla NREN tedesca; il MW del nodo tedesco decidera' se istanziare una cache a Juelich o servire dati in streaming da KIT
  - KIT e' connesso con PIC nel lake, e il sw centrale del lake decidera' se sia meglio fornire i dati in streaming a KIT e/o spostare dei dati PIC → KIT
  - Per noi "lo storage e' centrale", e i collegamenti fra i centri storage sono quelli prioritari
- Dal punto di vista del lake, e' come se il processing avvenisse fra PIC e KIT; la connessione nazionale KIT → Juelich e' un "dettaglio interno tedesco"



Ma la rete? Ov

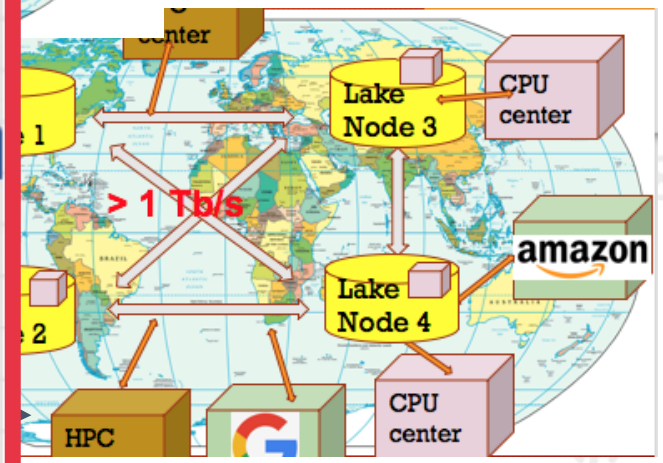
Ci sono (almeno) 4 tipi di r

1. a livello alto, la rete fr (intercontinentale)
2. a livello medio, la rete data lake (a livello co
3. a livello nazionale, la che compongono un
4. a livello nazionale, la centri opportunistici

Per esempio, ecco la sch data challenge da qui al 2 considerata sia da un no l'esterno, sia verso la stru

LHC Network Needs (Gbps) Flexible Scenario in 2027	
T1	
CA-TRIUMF	393
DE-KIT	1247
ES-PIC	370
FR-CCIN2P3	1124
IT-INFN-CNAF	1345
KR-KISTI-GSDC	99
NDGF	285
NL-T1	376
NRC-KI-T1	247
UK-T1-RAL	1183
RU-JINR-T1	207
US-T1-BNL	906
US-FNAL-CMS	1817
(atlantic link)	2723
Sum	9600

SOLO PER LHC!



					-CMS ds (Gbps) enario in 7	Alice Network Needs (Gbps) Minimal Scenario in 2027	LHCb Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027
					196	0	0	196	393
					473	85	66	624	1247
					169	0	16	185	370
					448	56	58	562	1124
					472	105	95	673	1345
					0	50	0	50	99
					111	31	0	142	285
					143	12	34	188	376
					51	51	21	123	247
UK-T1-RAL	15	9	3	27	472	10	109	591	1183
RU-JINR-T1	0	5	0	0	103	0	0	103	207
US-T1-BNL	23	0	0	0	453	0	0	453	906
US-FNAL-CMS	0	45	0	0	909	0	0	909	1817
(atlantic link)					1362	0	0	1362	2723
Sum	100	100	100	100	4000	400	400	4800	9600

Quindi, nel prossimo decennio servira'

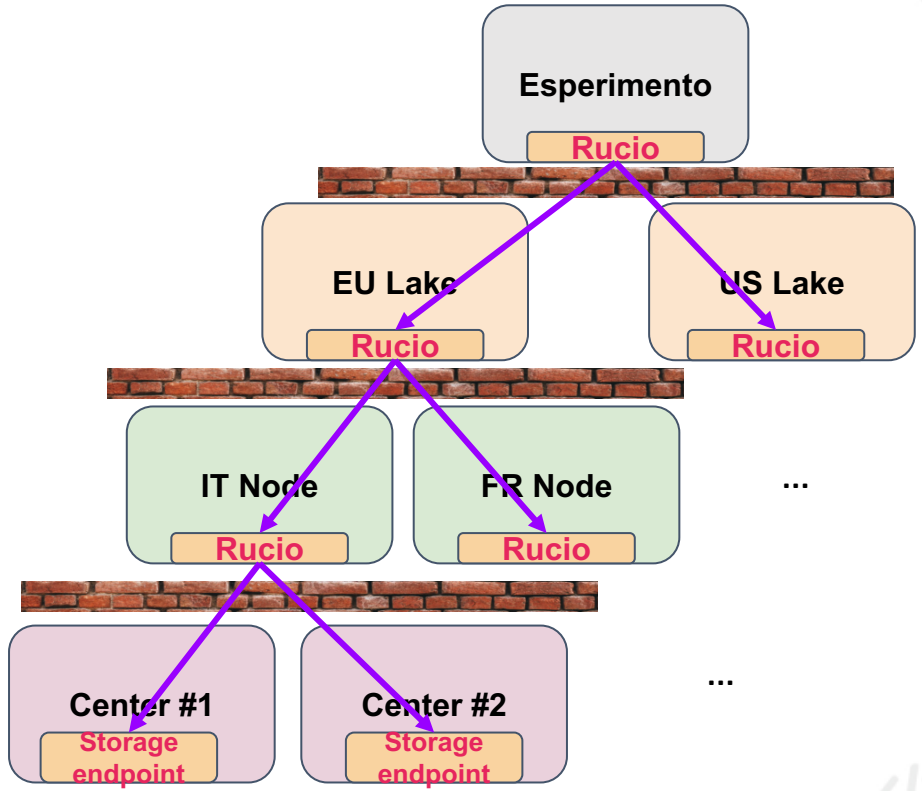
Limitandosi alla rete della fisica delle alte energie

1. Una connessione verso l'esterno superiore al Tbps. Se via Geant o anche via DCI con il CERN da decidere / testare.
2. Una connessione interna nazionale globalmente superiore al Tbps. 1 Tbps comunque necessario per i nodi principali (CNAF, Bari, Roma, … ?), Anche questo se via backbone o DCI "interessa poco"
3. Una capacità simile verso i grossi "*fornitori di cicli di calcolo*", in primis il CINECA. Per noi comunque da vedere come "leaf" del nodo CNAF almeno logicamente se non fisicamente.
4. Capacita' di accendere on demand network paths verso "opportunità di calcolo" se si presentassero (Aruba? TIM Cloud? nodi della "futura cloud nazionale" ?)

Moltiplicare questo per gli altri ambiti con richieste "simili" (Astroparticle almeno…)

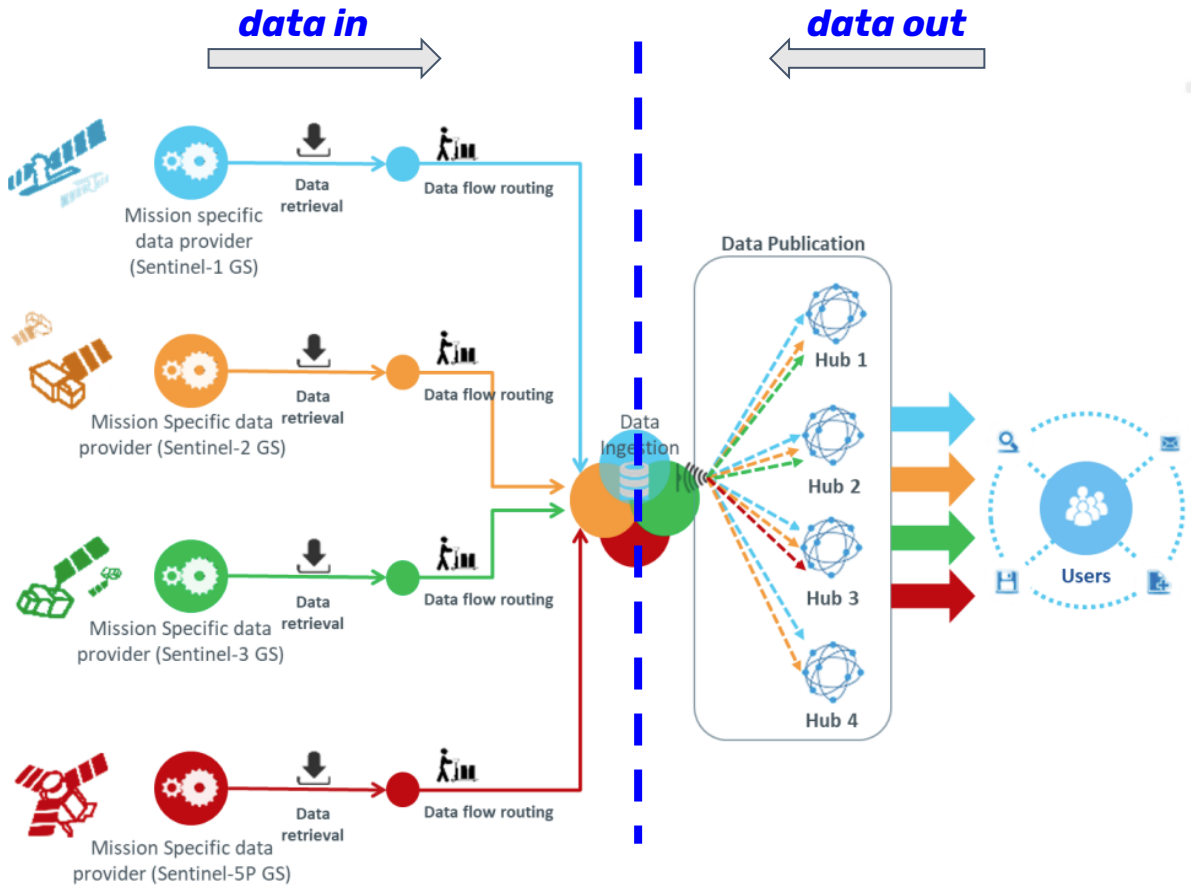
Un po' di idee sparse ...

1. Quale sistema di gestione del (dei) lake? A momento candidato migliore e' ***Rucio*** (gia' accettato per LHC, SKA, Dune, ….). Ci immaginiamo fino a 3 Rucio in cascata per delimitare le responsabilità
2. Rucio deve implementare la QoS a tutti i livelli, sempre in modo compartimentato: se la richiesta e' "mantieni 2 copie del dato", il sistema deve reagire a problemi in modo automatico
3. l'istanziazione delle caches e' un aspetto interessante: devono essere vicine ai siti, domain / protocol specific, o non domain specific a livello infrastrutturale, per esempio nei POP?



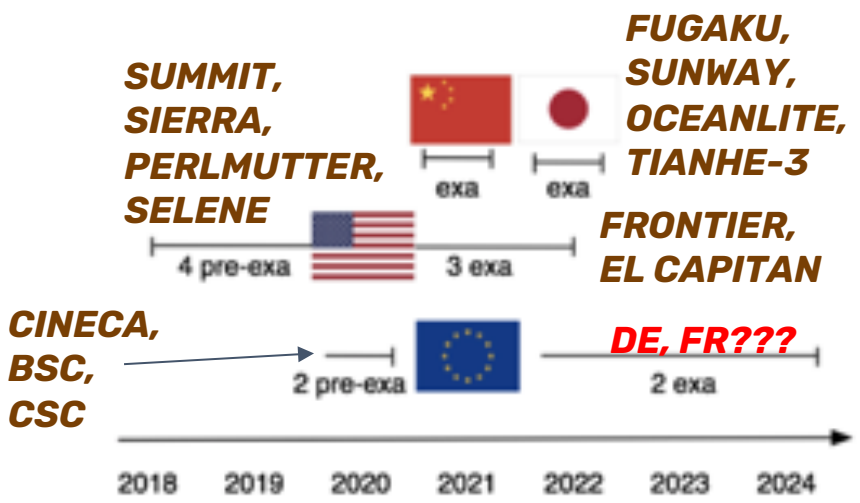
Un modello piu' classico: Mirror Copernicus

- Modello piu' classico, con netta distinzione fra *ingress* e *egress* (2 strutture logicamente diverse)
- Pero' nulla vieta di implementarlo fisicamente sui nodi del data lake, non e' necessaria un'infrastruttura parallela



L'uso dei centri HPC … una sfida!

- Un singolo centro Exascale (se uno riuscisse a convertire *Flops* in *HepSPEC06*) sarebbe nel 2030 in grado di soddisfare tutte le richieste di calcolo di LHC
- Convertire: utilizzare in modo ottimale GPUs, FPGAs, TPUs etc
- Problema non banale: codici della fisica sperimentale O(10M) righe di codice, 10+anni di sviluppo. Non si riscrivono, e di certo non si riscrivono ogni pochi anni
- Grande interesse in un tipo di programmazione generica, con backend di esecuzione molteplici e future-proof. ***Write once (more), run forever (beh…)***



kokkos alpaka

SYCL™ oneAPI

DATA PARALLEL C++

Conclusioni

- Per la fisica, ovviamente il calcolo non e' un fine ma un mezzo, spesso (dolorosamente) necessario
- Dalle necessita' nascono opportunita': il calcolo distribuito, la GRID, l'uso di Cloud Commerciali e sistemi HPC
- Nel futuro prossimo la rete (se possibile) sara' ancora piu' importante per la riuscita delle nostre ricerche: da *"mezzo su cui spostare files"* a *"collante fisico di infrastrutture remote"* per farle sembrare uniche.
- Una delle poche regole che ho sempre visto rispettate nel calcolo scientifico a queste scale: quando possibile, fare overprovisioning di rete. Le idee su come usarlo sono sempre arrivate (remote streaming, modello full mesh, centri storageless, …)