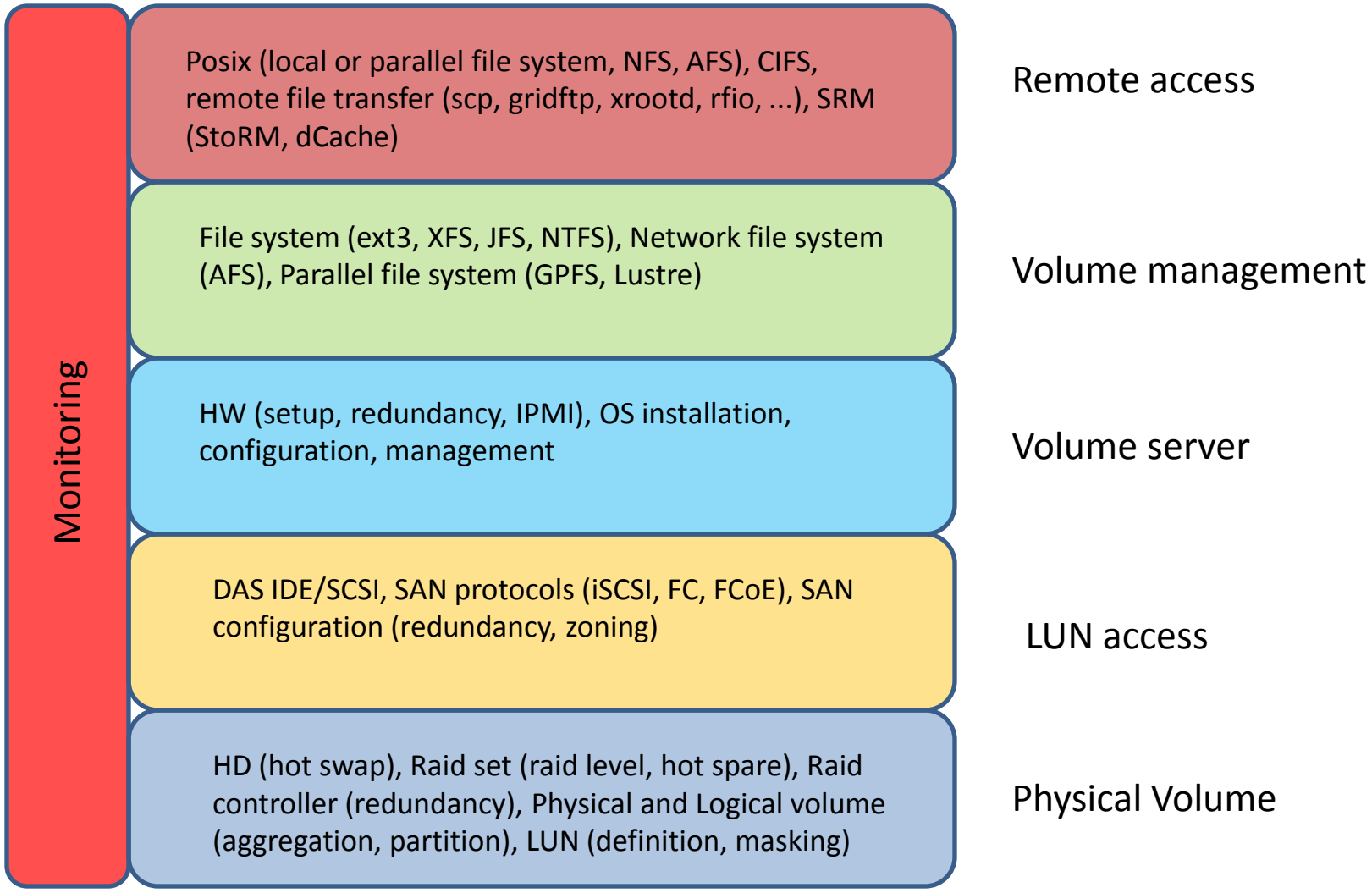
Esperienze in alcune sezioni su integrazione gestione sistemi storage per il calcolo e per i servizi

Alessandro Brunengo

Componenti di un sistema di storage



Infrastruttura di storage

- Opportune scelte tecnologiche permettono di realizzare una struttura omogenea (anche parziale) che puo' essere considerata *infrastruttura*
  - definizione dei requisiti
  - scelta tecnologica implementativa
    - non solo come soddisfare i requisiti, ma valutare
      - difficolta' di gestione
      - capacita di evoluzione (scalabiilita', nuove funzionalita')
      - costi
  - definizione del supporto
    - chi, come, quando, se...
- Deve esserne valutata la convenienza
  - bilanciare vantaggi e svantaggi
    - omogeneita' di gestione della infrastruttura
    - eventuali feature aggiuntive (affidabilita', flessibilita', ...)
    - expertise concentrata in un pool di persone con migliori competenze
    - complessita' di gestione (migliora o peggiora?)
    - costi per l'hardware (spese infrastrutturali, economicita' nell'accorpamento degli acquisti)
    - costi in personale
    - rampa di apprendimento per nuove tecnologie

Attori

- Lo storage (quello in produzione, e quello che arrivera')
- Utenti: i gruppi sperimentali locali, che necessitano di storage e lo finanziano
- Gestore dello storage:
  - il servizio calcolo e reti locale, che gestisce gia' lo storage per i servizi centrali (web, home, profili, scratch, backup, software distrib, ...) e gestisce le infrastrutture di base (connettivita' di rete e servizi relativi, eventuale sala CED)
    - il personale del servizio calcolo e reti ha la caratteristica di avere esperienza sulla gestione di sistemi informatici, anche di storage, che costituisce parte del suo lavoro ordinario
  - personale di esperimento dedicato a quel ruolo
    - personale che generalmente di mestiere fa prevalentemente altro
    - personale che in assenza di una struttura gestionale piu' o meno organizzata offre supporto al solo gruppo sperimentale di appartenenza

Requisiti

- I nostri utenti (gli esperimenti) hanno bisogno essenzialmente di
  - accesso locale posix (tutti gli esperimenti che non usano Grid, ma anche alcuni esperimenti LHC)
  - accesso via interfaccia SRM, e protocolli relativi (GridFTP, xrootd, rfio, dCap, ...)
- altri “utenti” sono i servizi centrali di sezione, che possono avere altre necessita’
  - AFS, CIFS
  - volumi raw da gestire via LVM
  - aree per il backup

Soluzione non integrata

- Nessuna infrastruttura comune
  - l'utente dello storage compra, configura e mantiene il suo hardware
    - l'appoggio ad un servizio si puo' limitare alla disponibilita' di hosting (sala CED), alimentazione e connettivita' locale
  - l'utente mantiene il controllo su
    - scelta tecnologica (idonea a soddisfare i requisiti)
    - acquisto dell'hardware
    - meccanismi e responsailita' di intervento
      - per installazione e configurazione (anche di monitoring)
      - per l'analisi e la soluzione dei problemi
  - ... e deve
    - fare technology tracking
    - dedicare personale sufficiente alla gestione
    - costruirsi e mantenere expertise sulla soluzione scelta

- Spesso l'utente ha bisogno comunque di parziale (totale?) supporto dal servizio calcolo locale
  - il know how piu' ricco di esperienza di solito sta' li'

Soluzione integrata

- Infrastruttura condivisa tra diverse utenze
  - a livello hardware
    - accorpamento degli acquisti
      - minori costi unitari, o migliore qualita' a parita' di costo
    - omogeneita' dell'hardware
      - minore costo per gestione e per il monitoring
  - a livello di accesso al volume logico, essenzialmente tramite Storage Area Network
    - costo infrastrutturale aggiuntivo (se non pre-esistente)
    - costo gestionale aggiuntivo (se non pre-esistente)
    - (notevoli) vantaggi qualitativi: affidabilita' e flessibilita' nella gestione dei volumi

Soluzione piu' integrata

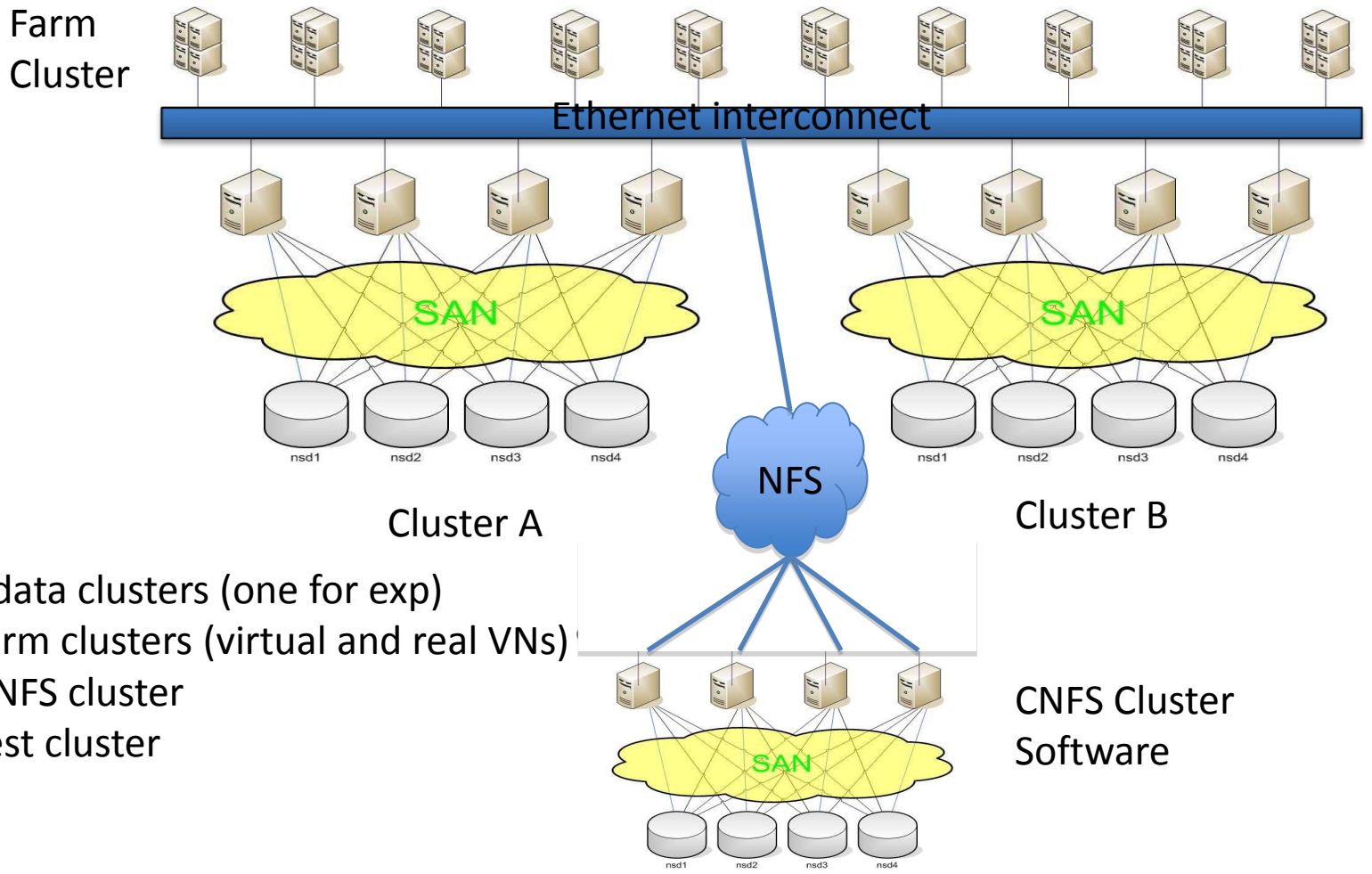
- A livello di file system e disk server
  - scelte opportune permettono migliori
    - prestazioni
    - flessibilita' (espandibilita', migrazione dinamica dei volumi e dei server di riferimento)
    - affidabilita' (ridondanza su metadati, su dati, sul server che rende disponibile il volume)
    - funzionalita' (snapshot, policy per la gestione automatica della migrazione dei dati)
    - ancora omogeneita' di gestione
  - senza perdita di funzionalita'
    - accesso posix
    - accesso remoto tramite i protocolli necessari (NFS, * file transfer, SRM, ...)
  - con qualche costo
    - non e' "chiavi in mano"
- L'infrastruttura rende disponibile all'utente "aree dati", delle dimensioni volute (pagate), accessibili tramite i protocolli desiderati

Qualche esempio di integrazione

Tier1 - CNAF

- Il sistema di storage e' costituito da
  - infrastruttura di accesso allo storage su disco
    - SAN FC
    - dischi gestiti da controller raid che esportano sulla SAN le LUNs
    - disk server dotati di HBA dual head per l'accesso alle LUN
  - infrsatruttura di accesso allo storage su nastro
    - TAN FC
    - GPFS+TSM per la gestione integrata dei dati
  - infrastruttura di gestione dei volumi
    - GPFS
- Su questa infrastruttura si appoggiano le utenze
  - accessi posix via GPFS
  - accessi posix via CNFS ove necessario (aree software)
  - accessi SRM (StoRM) e protocolli connessi

Tier-1 GPFS infrastructure



10 data clusters (one for exp)
2 farm clusters (virtual and real VNs)
1 CNFS cluster
1 test cluster

Tier-1 GPFS infrastructure

- 10 GPFS clusters
  - 1 for each experiment (Diskservers, GridFTP, StoRM, TSM clients)
    - ATLAS: 18 NSD, **426**TB (3xCX3-80), 6 GridFTP
    - CMS: 8 NSD, **472**TB (CX4-960), 4 GridFTP
    - Others: ~36 NSD, 15 FS, ~1PB total (7xCX3-80), 18 GridFTP
  - Experiments' software (all experiments)
    - Clustered NFS
    - 4 NSD, 2.4TB (CX3-80, 22 FC disks, RAID10)
- Worker nodes (real) ~ 500 nodes
  - 8 jobs x node
  - Mounting all FS (statically)
- Worker Nodes (virtual) ~ 3600 nodes
  - 1 job x node
  - Mounting only FS used by job

GEMSS (Grid Enabled MSS)

= GPFS+TSM+StoRM+GridFTP

GPFS

- Fast parallel file system
- ILM support
- Interface to external storage pools

TSM

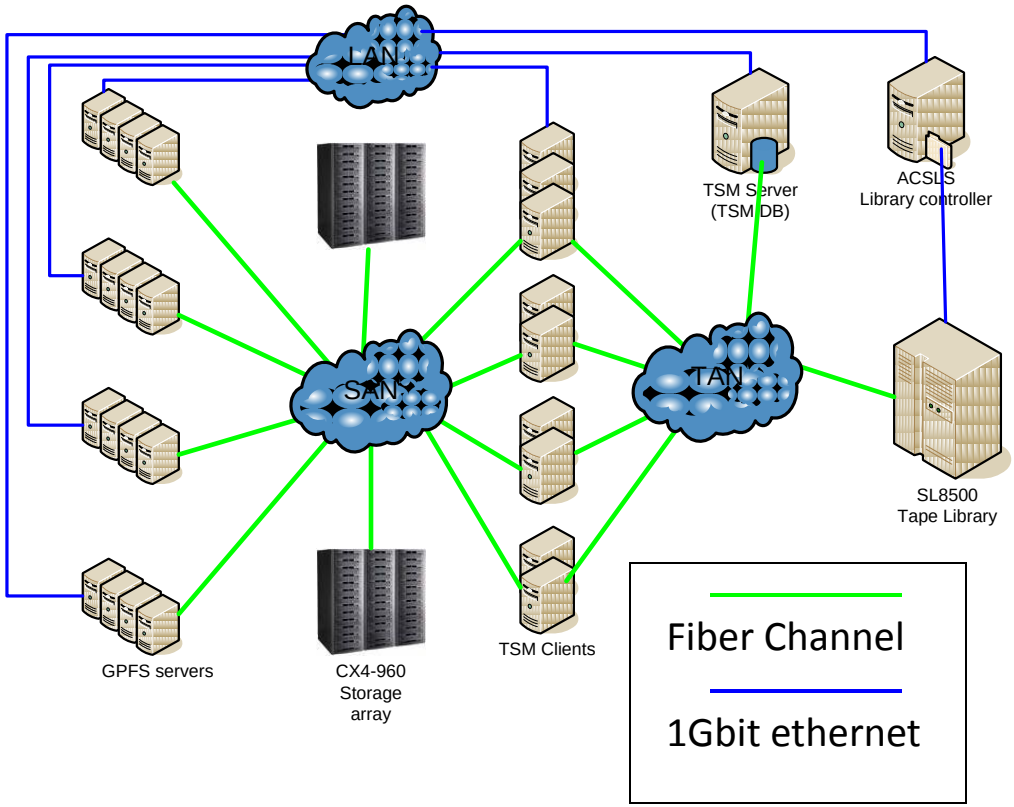
- backup, archive, and space management (HSM)
- DB2 database to track information about server storage, clients, client data, policy, schedules
- **Storage Agent** - LAN-free FC data movement for client operations

StoRM

- SRM interface for Grid users

GridFTP

- Grid file transfer



In production at CNAF as main MSS solution since Q3 2009

INFN Pisa

- Il sistema di storage e' costituito da
  - infrastruttura di accesso allo storage si disco
    - SAN FC
    - dischi gestiti da controller raid che esportano sulla SAN le LUNs
    - disk server dotati di HBA dual head per l'accesso alle LUN
  - infrastruttura di gestione dei volumi: GPFS
- Su questa infrastruttura si appoggiano le utenze (tutti i gruppi sperimentali)
  - accessi posix via GPFS (farm centrale o dedicata per CFD)
  - accessi posix via NFS/CNFS ove necessario
  - accessi SRM (dCache per CMS, StoRM per altri) e protocolli connessi
- Sulla infrastruttura si appoggiano anche alcuni servizi centrali
  - aree AFS (home directory, su LUN dedicate)
  - storage per backup su disco (su GPFS)
  - aree per la distribuzione software (su GPFS)
- In aggiunta vengono ancore utilizzati NAS non integrati nella infrastruttura per utenze SRM/StoRM

INFN Pisa - GPFS

- Unico file system GPFS con separazione delle aree dati via fileset e quota a livello di fileset
  - aree disco dei vari esperimenti non in GRID che hanno farm locali, fanno accesso via NFS;
  - aree disco dedicate al T2 di CMS e quindi ri-servite tramite dCache dalla struttura SRM specifica di esperimento;
  - aree disco di comunità grid diverse da CMS e servite attraverso StoRM;
  - aree disco si servizio dedicate per esempio al back-up dei disk server AFS anche in questo caso attraverso accesso legacy;
  - disco dedicato ad alcuni cluster HPC per calcoli di CFD, in questo caso accesso GPFS nativo
- A livello hardware:
  - 1 sistema di storage DDN9900 con 140TB RAW installati;
  - 2 switch FC Cisco MDS9020 4gbs
  - 4 NSD server con FC a 4gbs e rete GE

INFN Pisa - Considerazioni

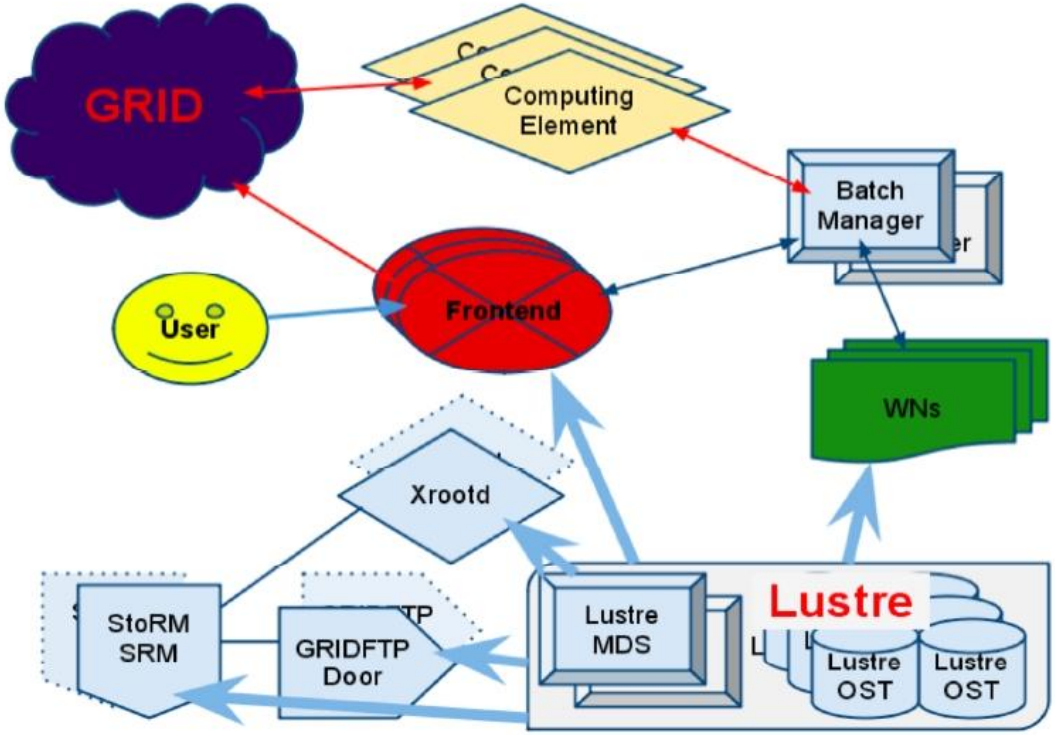
- Problemi risolti dalla infrastruttura comune
  - elevato carico di lavoro per il servizio legato alle diverse peculiarita' di server diversi (comunque anche chi dice di amministrarsi la macchina per proprio conto finirà per chiederti aiuto)
  - lunghi tempi per rendere disponibile nuovo spazio disco agli utenti, dato che era necessario installarlo fisicamente nel sistema, creare le LUN esportarle ai server giusti, creare i filesystem o allarga i volumi nel caso di LVM...
  - data la rigidità del sistema era anche complicato allocare spazio on-spot per esigenze particolari e limitate nel tempo. La richiesta tipica è: ho bisogno di qualche centinaio di GB per 6 mesi per fare una analisi....
  - Nella nuova struttura tutti questi problemi sono scomparsi, in particolare la gestione delle quote per fileset permette di disaccoppiare la gestione fisica del disco (installazione, aggiunta di NSD ecc...) dalla disponibilità dello stesso per gli utenti, riducendo di molto i tempi fra la richiesta di nuovo disco da parte del gruppo e l'effettiva possibilità di utilizzo.

INFN Bari

- Utenze (esperimenti): CMS, Alice, Teorici, Glast/Fermi, Pamela, Totem, Gruppo V
- Diversi gruppi fra quelli elencati hanno necessità di usare file-system con supporto completo a posix, quindi la scelta era quasi unicamente fra GPFS e Lustre
- Il file-system Lustre è quindi unico e condiviso sia sui nodi di frontend che sui WN. I disk server sono misti: SAS, FC, DAS. Marche: SUN, Xyratex, Nexsan e vecchi 3ware.
  - quote e partizioni equivalenti ai fileset
- La gestione dello storage è centralizzata: il team della farm di grid installa e gestisce i disk server di tutti i gruppi in modo da risparmiare sul man power necessario.
- Sia le CPU che lo storage viene acquistato con ordini comuni in modo da ottimizzare le risorse: questo è utile soprattutto nel caso dello storage dove i "quanti" di acquisto sono più alti e un singolo gruppo può non avere le risorse necessarie per effettuare l'acquisto in proprio.

INFN Bari

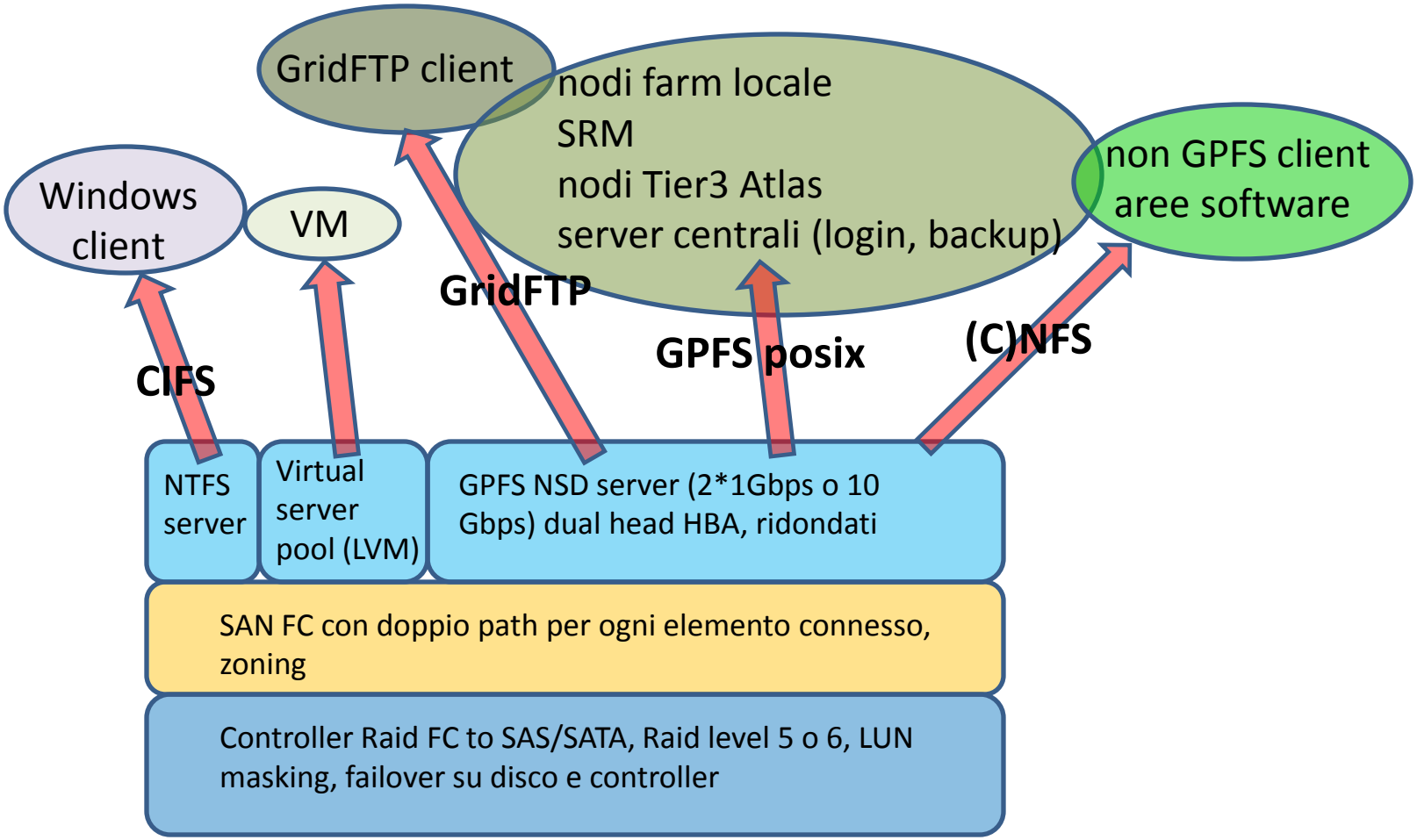
Schema logico della farm



INFN Genova

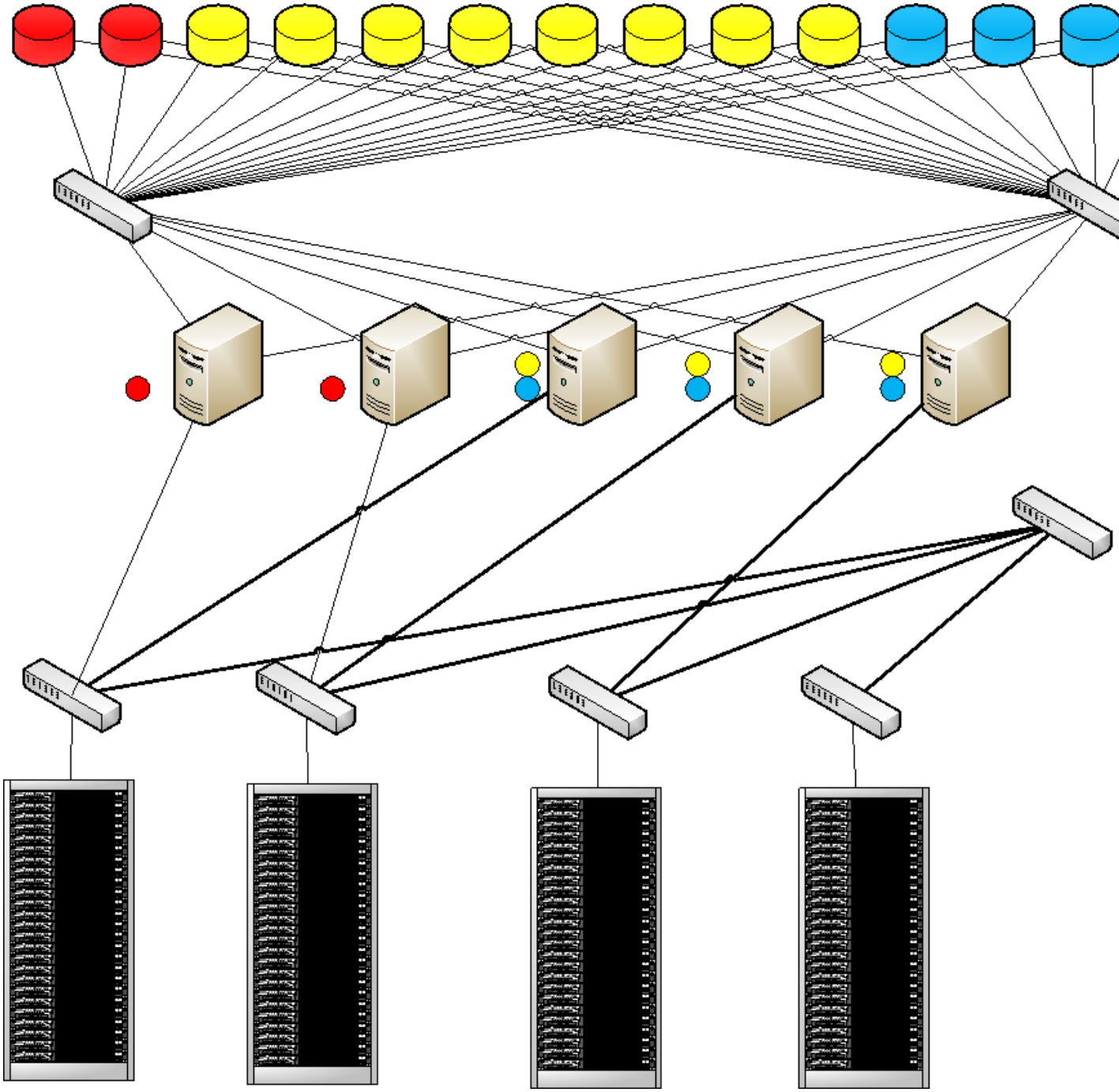
- Utenze (esperimenti): Atlas, Borex, JLAB12, Magic5, Totem, Gruppo4
- Tutti hanno necessità di usare file-system con supporto completo a posix
  - file system GPFS, con gestione di fileset e quote (110 TB)
  - export di aree dati GPFS via CNFS per le aree software o verso client non GPFS
- Accesso alle LUN via SAN FC switched ridondata
- La gestione dello storage è operata dal servizio calcolo (dischi, SAN e server)
- Ordini di acquisto accorpati e gestiti dal servizio calcolo

INFN Genova


GridFTP client
nodi farm locale
SRM
nodi Tier3 Atlas
server centrali (login, backup)
non GPFS client
aree software
Windows client
VM
GridFTP
CIFS
GPFS posix
(C)NFS
NTFS server
Virtual server pool (LVM)
GPFS NSD server (2*1Gbps o 10 Gbps) dual head HBA, ridondati
SAN FC con doppio path per ogni elemento connesso, zoning
Controller Raid FC to SAS/SATA, Raid level 5 o 6, LUN masking, failover su disco e controller

INFN Trieste

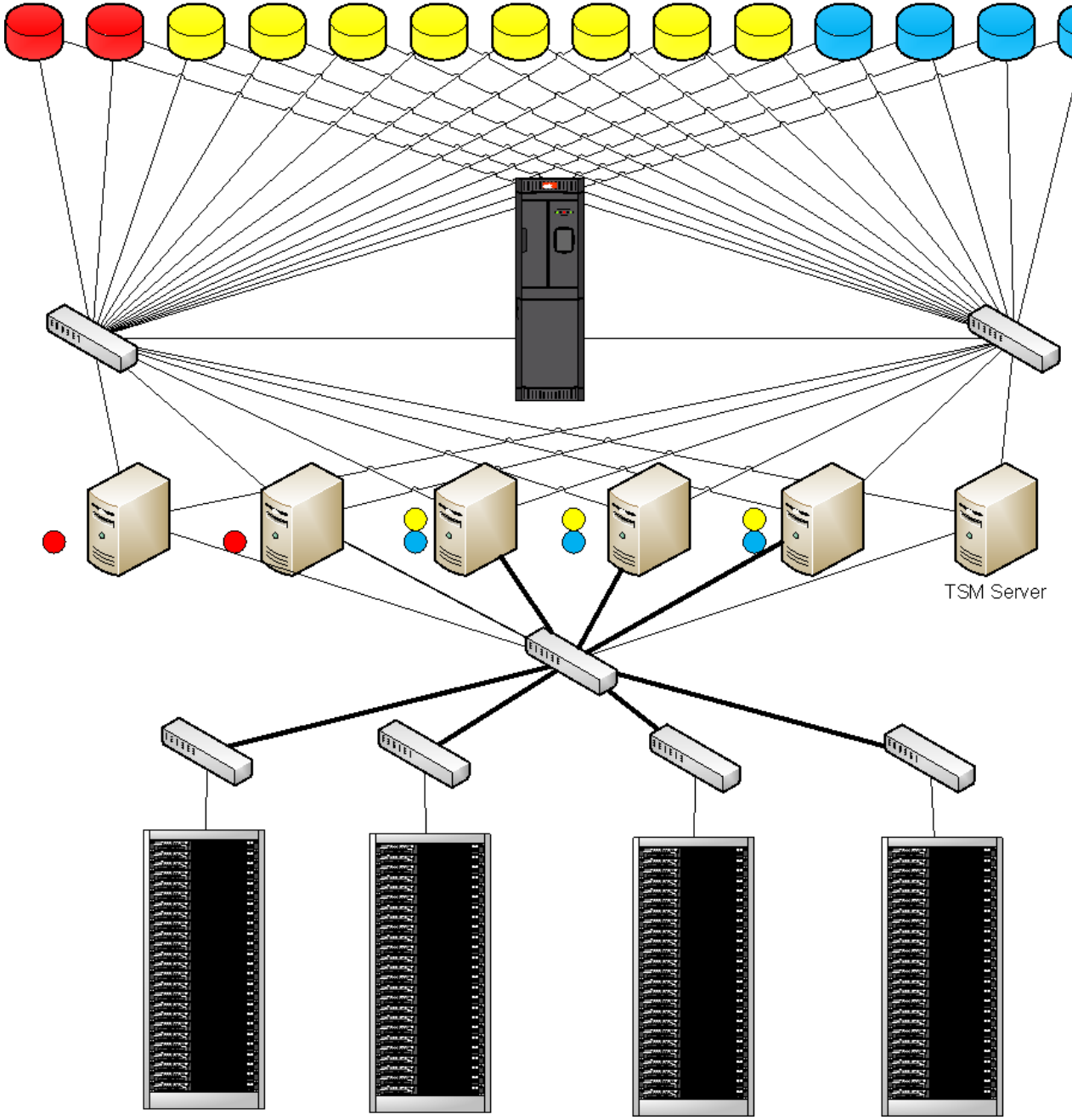
- Tutti i volumi di esperimento appartengono ad un unico filesystem GPFS condiviso e gestito con le quote
- una parte di GPFS serve come base per STORM (3TB) dei complessivi 160 TB
- La gestione e' fatta dal servizio e vengono programmati gli acquisti cercando di armonizzare le richieste di esperimento. Questa soluzione offre solo vantaggi per la scelta del materiale acquisito e perche' consente di fornire anche un overbooking dello spazio perche' non sempre tutti gli esperimenti utilizzano a pieno quello che loro hanno acquistato e quindi questo ottimizza le risorse.
- Si pianifica di utilizzare la Tape Library per realizzare una soluzione HSM per i dati usando Tivoli come al CNAF.
- Tutto questo consente la massima ottimizzazione che con il contemporaneo utilizzo di assegnazione dinamica degli slot LSF insieme al checkpointing ci ha portato ad un utilizzo intensivo della CPU e alla completa integrazione di GRID e calcolo locale degli esperimenti.



Legenda
Metadata (rosso)
Data (/gpfs) (giallo)
Data (/scratch) (blu)

Pro:
Divisione metadata/data sia disco (SAS/SATA) che servers
Load balacing/failover FC

Cons:
Utilizzo del bonding, distanza WN-Servers non uniforme, switch FC a 2Gb



Futuri sviluppi (inizio lavori maggio/giugno)

Installazione switch FC 8Gb
Infrastruttura Ethernet 10Gb per server dati e switch di rack
Installazione HSM (GPFS+TSM) 2 LTO4 + 100 slots

Considerazioni sull'esperienza

- Molti punti in comune per le tecnologie:
  - quasi tutti basano l'architettura di accesso al volume logico su SAN FC
    - soluzione flessibile (separazione controller/server, possibilta' di controllo di accesso tramite zoning), scalabile in modo progressivo, performante (switching), affidabile (grazie al multipath failover proprietario o meno)
  - tutti fanno largo uso di file system parallelo
    - accesso posix locale e remoto, con tutte le caratteristiche necessarie (quote a livello utente o gruppo, partizionamento tramite fileset con definizione di quote per fileset, ACL) e non (snapshot, ridondanza dati e/o metadati, espandibilita' dinamica, migrazione on line, integrabilita' con tape storage)
    - prestazioni (pur con limiti sulla tipologia di accesso)
    - possibilita' di export via NFS o CNFS
- Tutti riescono a servire le esigenze di tutti gli esperimenti
  - sia per accesso diretto (posix)
  - che per accesso via SRM
- L'infrastruttura fino a livello di SAN e' utilizzata da utenze con requisiti incompatibili con il file system (essenzialmente servizi centrali)
  - CIFS, AFS, LVM: il file system parallelo non e' una panacea..

Feedback dell'utenza

- Vantaggi su molti aspetti:
  - flessibilita' e affidabilita' si traducono in minore down-time: l'utente e' felice
  - l'accorpamento degli acquisti e la migliore qualita' del sistema di storage sopperisce al sovraprezzo della infrastruttura: l'utente e' felice
  - l'esistenza di un pool di gestione stabile su un sistema omogeneo permette di sviluppare un migliore know-how con tempi di risposta piu' brevi : l'utente e' felice
- Feedback negativo: nessuno

Considerazioni conclusive

- L'utilizzo di una infrastruttura omogenea per lo storage ha dei costi
  - hardware per l'infrastruttura stessa
  - apprendimento nella gestione di tecnologie nuove (SAN, file system parallelo)
  - dipendenza dello storage di tutti gli utenti da una unica infrastruttura
    - richiede gestione e monitoring opportuni
- e dei vantaggi
  - qualita' migliori
  - omogeneita' della tecnologia
    - possibilita' degli esperimenti di sfruttare il supporto del servizio calcolo senza un sovraccarico eccessivo (a regime)
  - migliore know how in un pool di persone dedicato

Ringraziamenti

- Grazie per la collaborazione e per le slide ricevute da
  - Vladimir Sapunenko (CNAF)
  - Enrico Mazoni (Pisa)
  - Giacinto Donvito (Bari)
  - Alessandro Tirel (Trieste)