

HPC e sistemi di archiviazione per la raccolta ed uso dati SRT.

Evento di consultazione 8 maggio 2020

Ipotesi di soluzione tecnica

Andrea Possenti

Coordinator of the PON-OR8

Gianni Comoretto

Coordinator of the PON-OR6

Riccardo Smareglia

*Responsabile Ufficio ICT e Science
Data Management INAF*

Il calcolo nella Radio Astronomia 2020-2030

Fra tutti gli ambiti dell'astrofisica osservativa, la **radioastronomia** rappresenta il settore con le maggiori richieste prestazionali nel campo dell'acquisizione di dati, del loro immagazzinamento e della potenza di calcolo necessaria al loro pieno sfruttamento.



ANTENNA – FRONT END

→
20-80 GB/s



BACKEND

→
0.5-2 GB/s



HPC + STORAGE

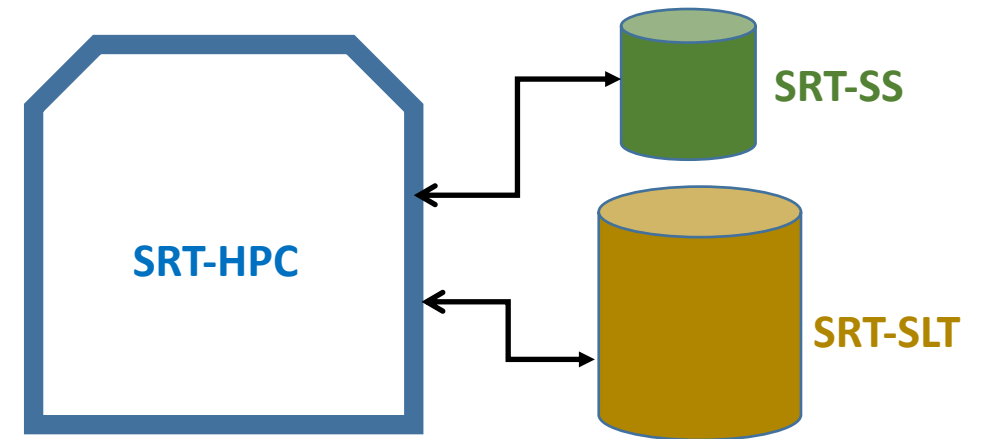
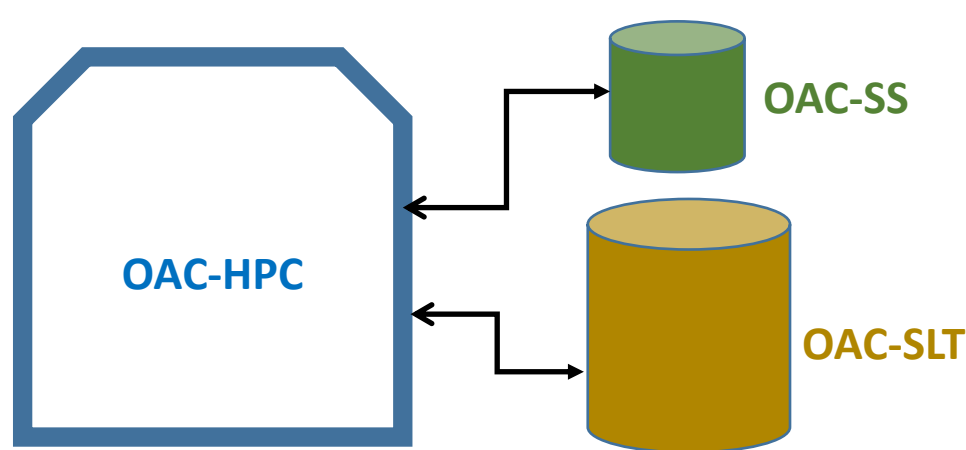
Obiettivo dell'OR8

Fornitura, installazione e avvio delle risorse High Performance Computing, in particolare storage e calcolo massivo, necessarie per l'archiviazione e l'analisi dati ottenuti con SRT.

La fornitura è composta da **2 cluster HPC** geograficamente distanti (40 km e completamente indipendenti), denominati rispettivamente cluster OAC e cluster SRT. Le condizioni ambientali dei due CED sono state già presentate

Ognuno dei due cluster è dotato

- ❖ di un proprio **storage di tipo scratch** denominato rispettivamente OAC-SS e SRT-SS
- ❖ di uno **storage long term** denominato rispettivamente OAC-SLT e SRT-SLT



OAC-HPC

nodo tipo
OAC-HPC-A
40 % dei nodi

calcolo parallelo e/o farming
CPU con supporto GPU per calcolo numerico
RAM 1 TB
Infiniband 40 Gbps

nodo tipo
OAC-HPC-B
20 % dei nodi

calcolo parallelo e/o farming
CPU e GPU per grafica interattiva con desktop e
rendering remoto
RAM 1 TB
Infiniband 40 Gbps

nodo tipo
OAC-HPC-C
40 % dei nodi

general purpose
CPU per calcolo e servizi in ambienti virtualizzati
RAM 512 GB

Switch 1/10 Gb/s Ethernet e 40 Gbps Infiniband

I/O alte prestazioni
minimo 0.6 PBy
condiviso fra nodi
o gruppi di nodi

OAC-SS

long term storage
minimo 2.0 PBy

OAC-SLT

SRT-HPC

nodo tipo
SRT-HPC-A
5 nodi

acquisizione dati da strumentazione
flusso dati fino a 40 Gb/s Qsfp
CPU con supporto GPU per calcolo numerico
RAM 1 TB
slots per espansione SSD-NVMEs PCI-Express 4.0
rate di scrittura/lettura simultanea attesi su
filesystem locale:
➤ max 150 MB/s su 100 TB effettivi
➤ da filesystem locale verso SRT-SS: fino a 20 MB/s in simultanea alla scrittura massima

nodo tipo
SRT-HPC-B
max 5 nodi

CPU – GPU per grafica interattiva con desktop e
rendering remoto in real-time
quick-look dati in ingresso da SRT-HPC-A
RAM 1 TB

Switch 1/10/40 Gb/s Ethernet

I/O idoneo a data
transfer da SRT-HPC-A
minimo 0.3 PBy
condiviso fra nodi o
gruppi di nodi

SRT-SS

room: da 2 U a 4 U

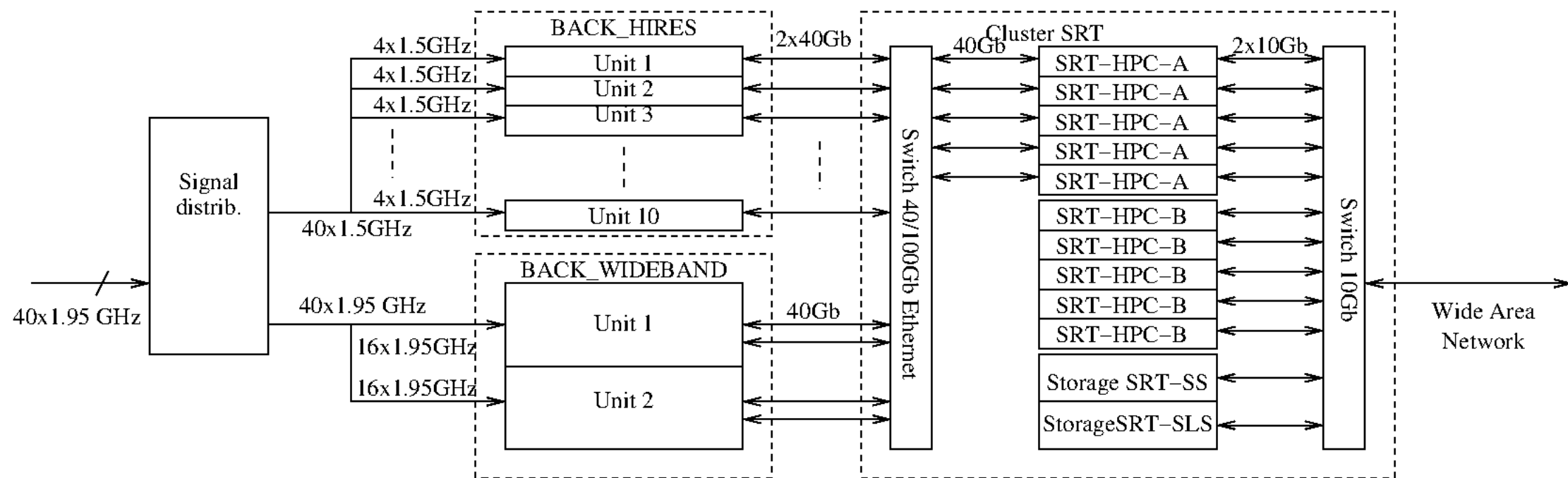
long term storage
minimo 1.0 PBy

SRT-SLT

max room: 12 U
escluse U per
passacavi
cablaggi consolle
etc

Struttura e Inquadramento del cluster SRT-HPC-A + B

Overview



Esempio classico di uso intenso delle risorse

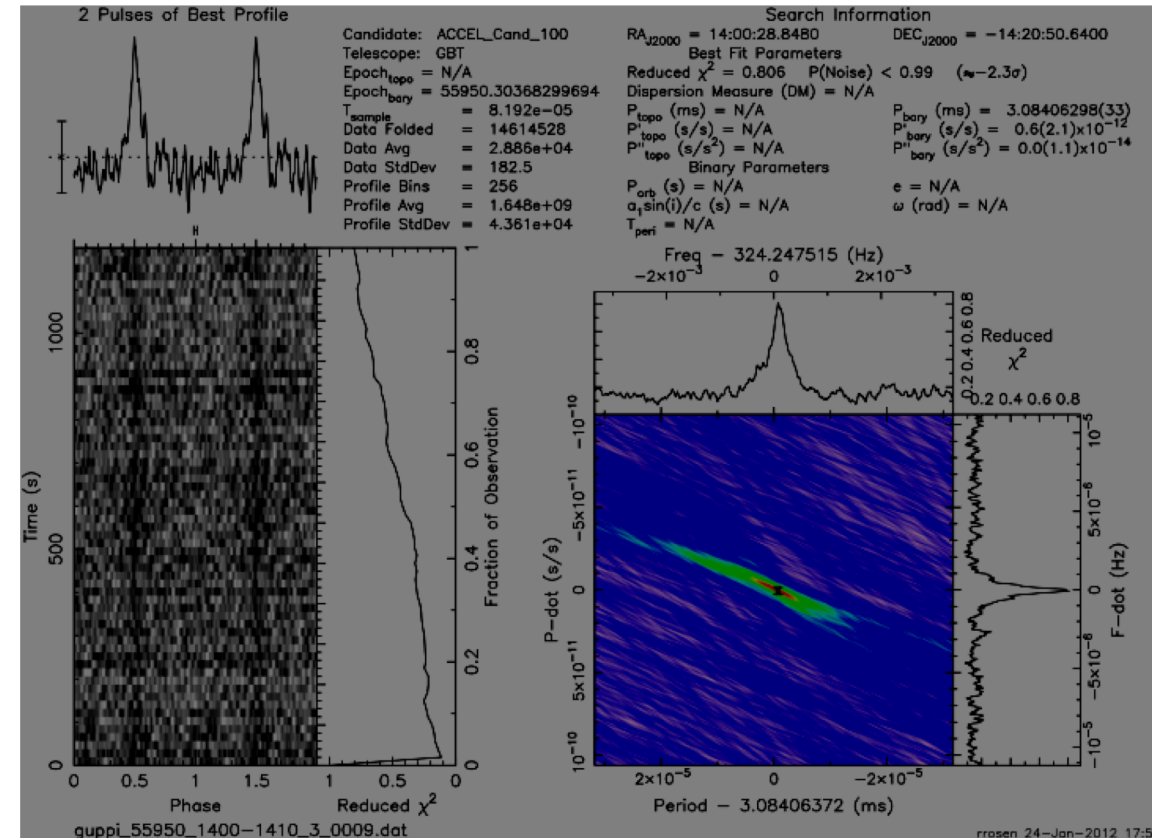
Ricerca di Pulsar

Il contesto di un problema reale

- Input data:
 - Campionamenti rapidi su molti canali, 256 Mbyte/s totali
 - 600s chunks: 170Gb/chunk
- Goal: rivelare impulsi periodici
 - Trasformate di Fourier + ricerca in griglia
 - Algoritmi complessi in 4 stadi successivi
 - 4 buffers, 680 GB minimo
 - Problemi largamente parallelizzabile → GPUs
 - Tipicamente richiesti 1-4 Tflops

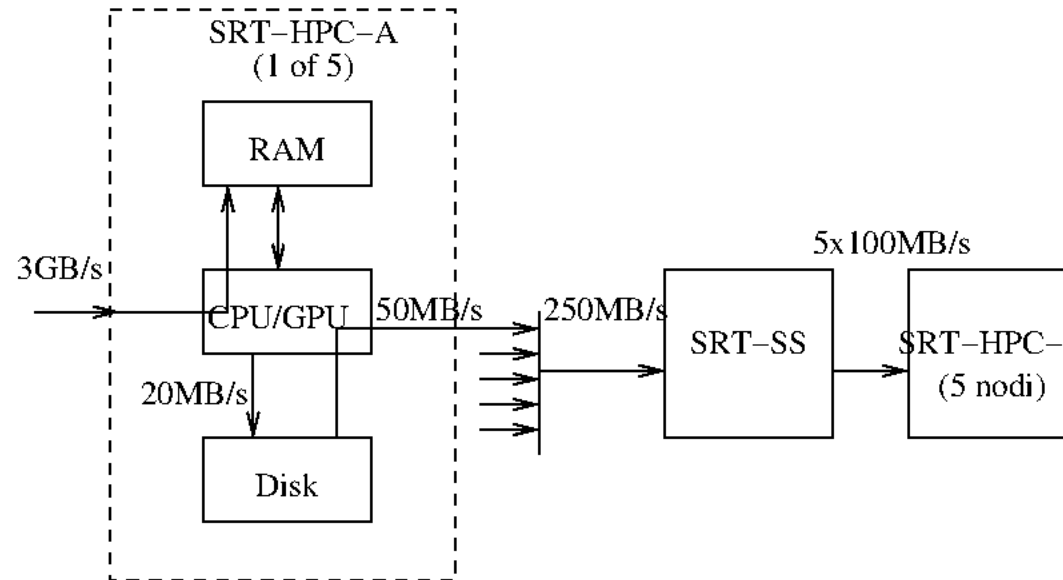
Più rapido è, meglio è

- Più segnali processati per ogni nodo



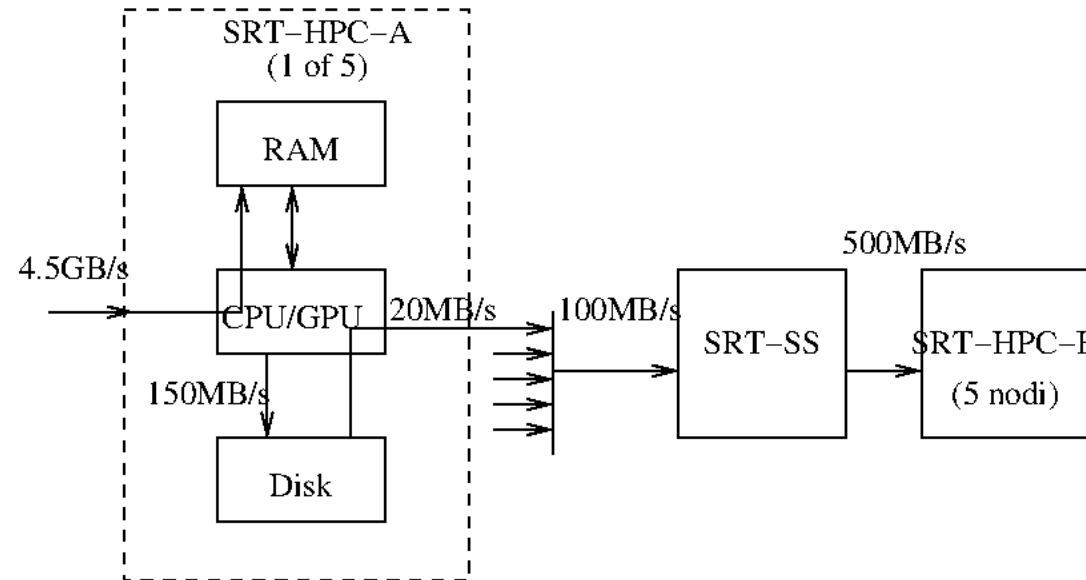
Struttura e Inquadramento del cluster SRT-HPC-A + B

Uso con high rate standard: e.g. pulsar search su 1 beam



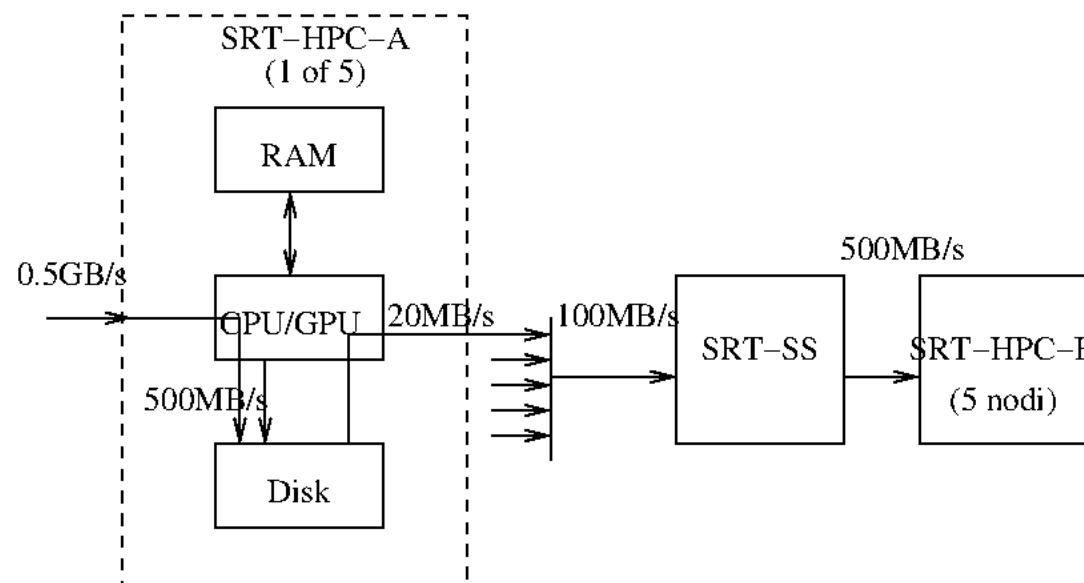
Struttura e Inquadramento del cluster SRT-HPC-A + B

Uso con high rate massimo: e.g. pulsar search su 7 beams



Struttura e Inquadramento del cluster SRT-HPC-A + B

Uso ultra spinto in rate di scrittura: e.g. *baseband recording*



Richiede SSD-NVME con PCI-E 4.0

Richieste ulteriori: tutti i nodi OAC-HPC/SS/SLT & SRT-HPC/SS/SLT

- ✓ **RAM per nodo espandibile fino a 2 TB**
- ✓ **spazio disco locale scratch per nodo (escluso OS): minimo 8 TB su dischi SSD, minimo 40TB su HDD**
- ✓ **schede di rete minimo dual-port 1/10 GbE**
- ✓ **accessori per gestione e allestimento cluster e storage: cablaggi, console kvm, passacavi etc**
- ✓ **specificare i consumi elettrici per ciascuno dei sottosistemi proposti**
- ✓ **indicare le prestazioni dei dischi SSD proposti in termini di velocità di scrittura e lettura sequenziale, e gli rpm per i dischi HDD proposti.**

Spazi a disposizione

spazio totale a disposizione:

- ✓ SRT-HPC-A: **max 27U** compresi apparati di rete e cablaggi/accessori, in un rack fornito dalla stazione appaltante, che conterrà anche 5 schede skarab
- ✓ SRT-HPC-B + SRT-SS + SRT-SLT + accessori: **max 60U suddivisi in max 3 rack** forniti dalla stazione appaltante
- ✓ OAC-HPC + OAC-SS + OAC-SLT: rack 47U (oggetto di questa fornitura) affiancati in linea, per una **lunghezza massima di 6 m**

Coffee break

A. OTTIMIZZAZIONE NODI COMPUTAZIONALI

- a. Dovendo girare sia applicazioni batch pesantemente parallele che applicazioni di visualizzazione grafica interattiva remota, possiamo scegliere un unico tipo di GPU per tutti i subcluster per ottimizzare costi e scalabilità o è preferibile differenziare le GPU per applicazione nei tre subcluster ?
- b. E' preferibile avere più GPU (uguali) su un unico nodo o aumentare il numero di nodi, ove non ci siano vincoli di spazio per allestimento rack ?
- c. E' preferibile adottare soluzioni dual CPU con molte decine di core ciascuna o su soluzioni CPU che pareggino il numero totale di thread aumentando il numero di nodi ?
- d. E' possibile gestire sullo stesso nodo due aree scratch raid distinte, ciascuna delle quali costruita con dischi SSD e HDD ? I controller che gestiscono i raid possono/deve essere fisicamente distinti per ottimizzare le prestazioni di entrambe le tecnologie ?
- e. Esistono soluzioni "all in one" in cui vengano condivisi in un unico case unità di tipo blade o slim node, e storage che condividano alimentazione, ventilazione network interna etc già ottimizzate ?
- f. Nel caso del subcluster SRT-HPC-A, poniamo di voler potenziare ulteriormente la connettività della rete per indirizzare il/i flusso/i dati vs la/le aree disco ad alte prestazioni. Allo stato attuale dell'arte è preferibile organizzare il nodo con un'unica area disco ed una scheda 100 Gb/s o raddoppiare la scheda di rete a 40 Gb/s e separare le aree disco ?

B. OTTIMIZZAZIONE AREE STORAGE SCRATCH LOCALI E CONDIVISE

- a. Posto che i due cluster HPC hanno nodi (organizzati in sub-cluster) che sono eterogenei sia come hardware (diversa dotazione di RAM, eventuali GPU, dischi etc.), sia per la natura degli applicativi, è preferibile predisporre una sola area scratch comune per tutti i nodi afferenti o avere aree diverse (eventualmente più piccole), ottimizzate sulle tipologie di nodo e sui relativi applicativi, in termini di rapporto prestazioni/prezzo?
- b. Quale tipo di storage tra NAS e SAN consente maggiore flessibilità gestionale e accessibilità via rete (ed esempio volendo utilizzare lo stesso storage come front end del cluster o lo stesso filesystem dei nodi) e integrazione in un cluster HPC di nodi “open source” ?
- c. Che prestazioni massime in scrittura/lettura (1 lettura 1 scrittura simultanea) si possono ottenere da un'area scratch costruita utilizzando ad esempio 24 HDD SATA3 7200 rpm in raid linear, rispetto allo stesso numero di dischi SSD e con il filesystem parallelo proposto?

C. LONG TERM MAINTENANCE

- a. Quale tipo di storage tra NAS e SAN consente maggiori possibilità di espansione futura, scalabilità e upgrade, anche oltre le tempistiche di garanzia standard (ad esempio mediante sostituzione dei dischi interni COTS)?
- b. Quale tipo di servizi accessori (supporto software, interventi in garanzia, manutenzioni periodiche etc...) deve essere attivato a supporto di una durata almeno quinquennale della fornitura ?
- c. Espandibilità: indicare se nel sistema proposto possono essere integrati dischi, memorie, schede ecc... non proprietarie, anche in sostituzione di parti offerte al termine del periodo di garanzia.

D. INFRASTRUTTURA

Posto di avere nelle due sedi un datacenter con sistema di raffreddamento *esclusivamente a sala, senza aria forzata e con rack in linea*

- A parità di configurazione, quale case tra 1/2/4U garantisce migliore efficacia nella dissipazione termica?
- E' preferibile collocare i nodi nei rack popolando consecutivamente tutte le U (eventualmente chiudendo gli spazi liberi con pannelli ciechi) oppure lasciare una unità U libera per favorire la circolazione dell'aria tra un nodo e l'altro ?
- Esistono sistemi di scambio aria calda/fredda e/o estrattori da collocare su rack non refrigerati con sistemi in-row, e con porte aerate? Come vanno dimensionati rispetto al popolamento del rack e alla potenza elettrica assorbita?
- Come vanno collocati i rack rispetto alla posizione delle uscite dell'aria fredda dell'impianto di condizionamento?

