CCR workshop – Stato e prospettive del calcolo scientifico

Cluster Nazionale CSN4
Organizzazione e tipo di accesso

Enrico Mazzoni - Calcolo Scientifico INFN-Pisa

Infrastruttura HW

- 128 nodi Acer Altos 520 2xOpteron 8356 2.3GHz ==> 1024 core totali
- 1 GB di RAM per core (espandibile a 8) ==> 1TB totale
- La configurazione cluster è realizzata tramite rete veloce InfiniBand DDR utilizzando
    - Switch Cisco SFS 7012, 144 porte CX4
    - 128 HCA Cisco SFS HCA320 A1, chipset Mellanox MT25204
- Inoltre i nodi sono connessi alla LAN del Grid Data Center tramite lo switch di core F10 E1200i in tecnologia 1GE
- Per lo storage si sfrutta l'infrastruttura esistente del Grid Data Center basata su due sistemi disco di classe enterpirse
    - DDN S2A9900
    - HDS 9985

L'accesso al disco avviene tramite filesystem GPFS utilizzando la rete InfiniBand tramite il protocollo IPoIB

Configurazione LSF

I requisiti da rispettare nella realizzazione del cluster sono:

- Utilizzo sia seriale che parallelo ottimizzando l'uso delle risorse
- Differenti priorità di utilizzo per vari sottogruppi delle VO theophys
- Possibilità di utilizzo “parassitario” da parte di altre VO

Per questi si utilizza:

- Fair share definito sul gruppo di host per differenziare la priorità dei vari sottogruppi e permetterere l'accesso alle altre VO (others)
- Definizione di 3 code per i vari tipi di job:
  - ***theophys*** per i normali job seriali, definita su tutto il sito
  - ***theompi*** per i soli job paralleli, definita solo sul cluster
  - ***theoshort*** per i job seriali di breve durata, definita solo sul cluster

Configurazione LSF

Analizziamo in dettaglio la configurazione delle code:

- ***theophys***: usata per i job seriali è la coda utilizzata per la VO theophys sul sito INFN-PISA ed insiste su tutti i nodi della farm. Grazie al meccanismo del FS i gruppi che hanno assegnate risorse sul cluster hanno la precedenza su questi nodi
- ***theompi***: usata per i soli job paralleli insiste solo su i nodi del cluster e per potervi accedere è richiesto il ruolo parallel sul VOMS. Ha un runtime massimo di 72 ore ed un reservation time di 8 ore, questo per permettere l'uso delle risorse dedicate al parallelo anche per altri scopi
- ***theoshort***: usata per job seriali di breve durata, non è necessario nessun ruolo ma si deve appartenere ad uno dei gruppi che ha risorse assegnate sul cluster. Insiste su i soliti WN della coda theompi ed ha un runtime massimo di 4 ore. Questa coda fa “backfill” per cui può utilizzare JS che siano stati riservati dalla coda theompi

Da LSF a Grid

Vediamo come si realizza in GRID la struttura disegnata in LSF:

- *Enablempi* è un parametro di sito (alias CE) e non per VO, quindi è stato necessario creare un CE da dedicare solo al supporto MPI, attraverso questo CE passano i job per le code theompi e theoshort
- I normali job seriali passano dai CE standard del sito INFN-PISA
- A livello di VOMS sono stati definiti una serie di sottogruppi della VO theophys rappresentanti le Iniziative Specifiche che hanno richiesto l'allocazione di risorse sul cluster. E' stato definito anche un ruolo (parallel) per accedere alle risorse MPI
- Per job paralleli si dovrà specificare nel JDL il CE dedicato e creare il proxy con il ruolo parallel, il CE mappa questi utenti sulla coda theompi
- Per job seriali corti si dovrà specificare nel JDL il CE dedicato e creare il proxy con uno dei sottogruppi abilitati, il CE mappa questi utenti sulla coda theoshort
- Job seriali standard basta richiedere nel JDL il sito INFN-PISA con un proxy theophys qualsiasi, si finirà nella coda theophys

Storage

Per lo storage si hanno due richieste:

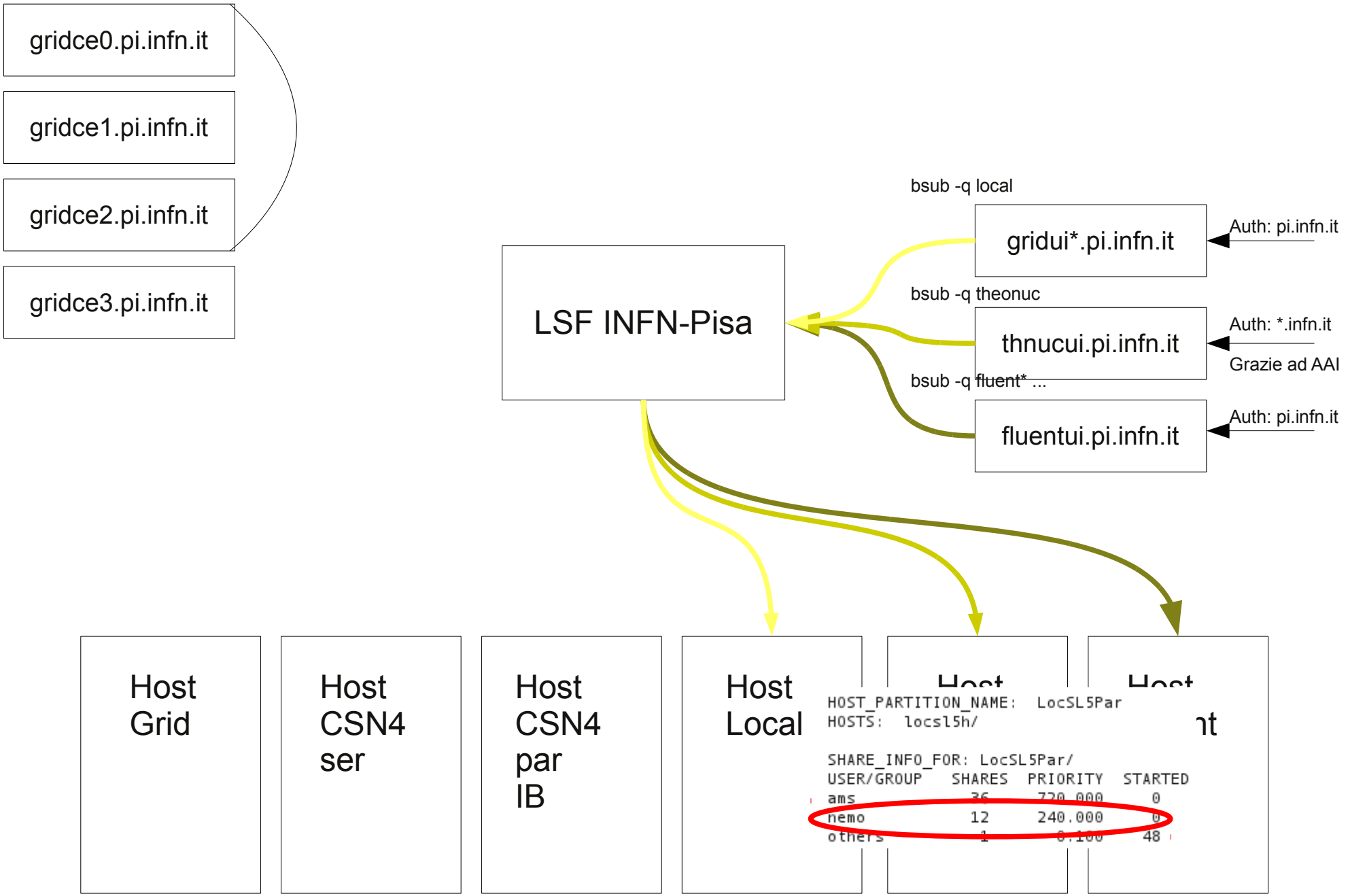
1. Disco per la home condivisa per i job paralleli
2. Disco gestito via SRM per accesso attraverso GRID

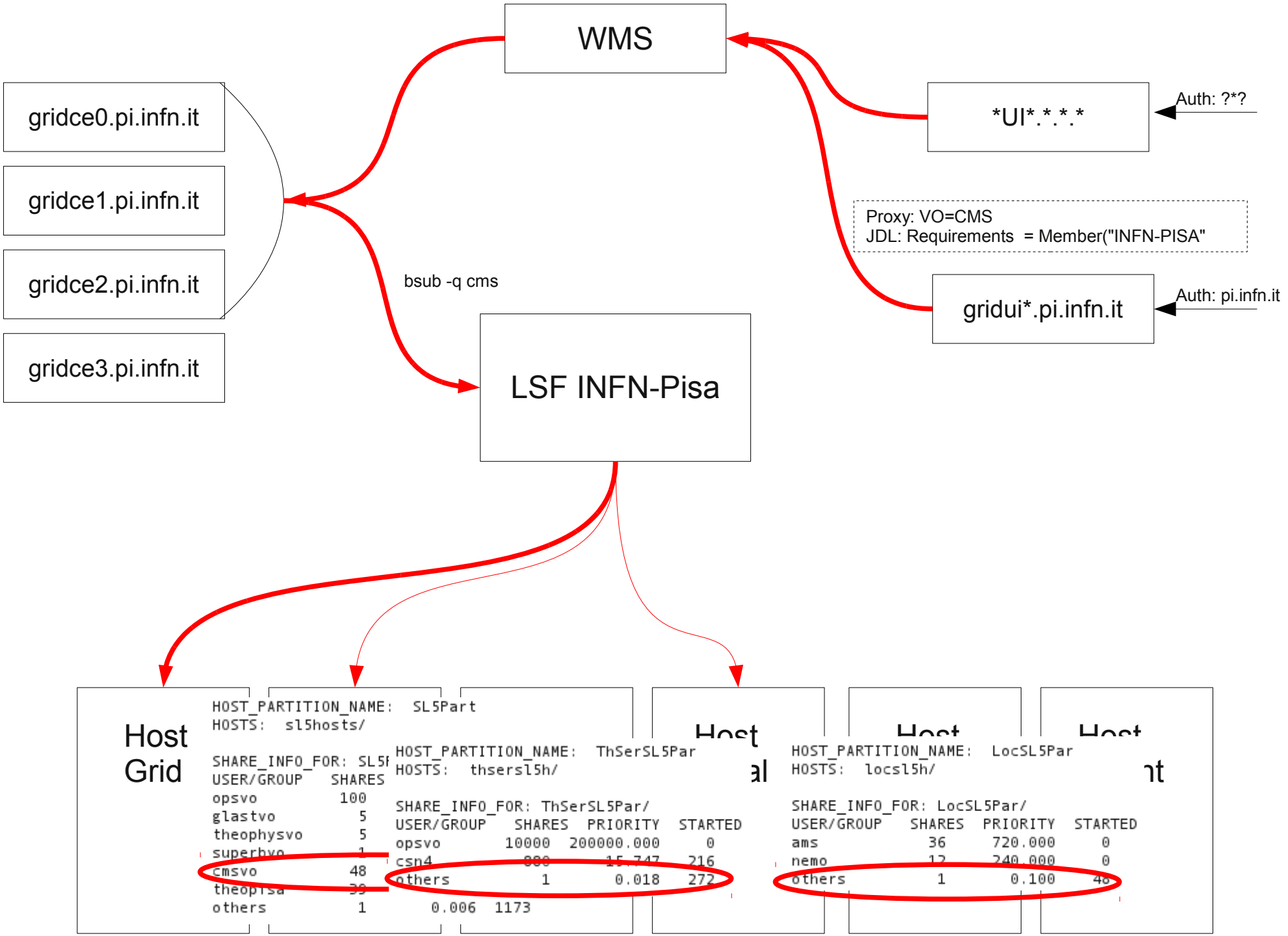
Entrambe le necessità sono state soddisfatte utilizzando l'infrastruttura di storage del Grid Data Center basata su GPFS, allocando 1TB per le home condivise e 9TB gestiti dall'SRM di sito basato su StoRM. In ogni caso è garantito l'accesso Posix ai dati. Si veda CCR-40/2010/P per i dettagli di questa infrastruttura.

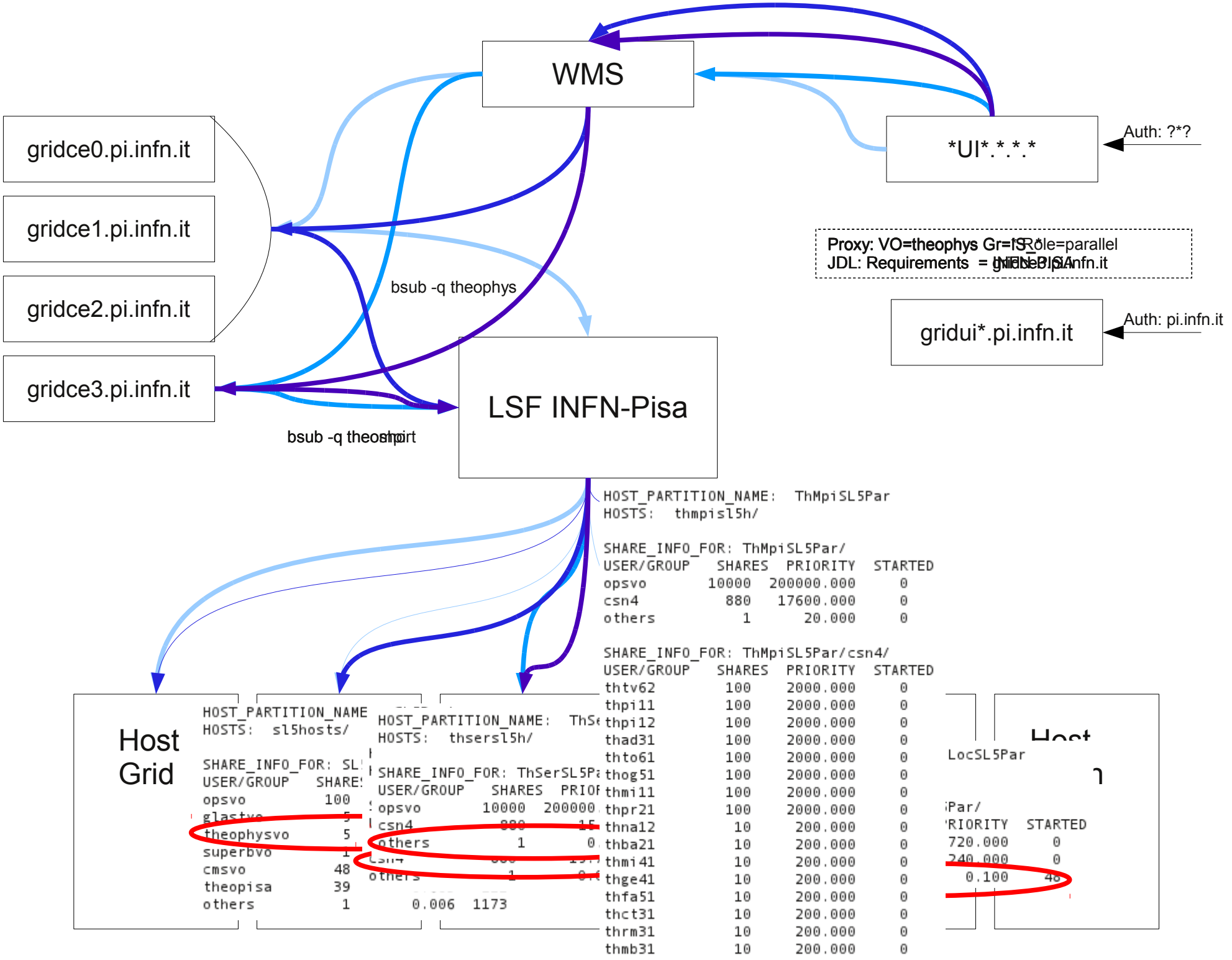
Oltre il cluster CSN4

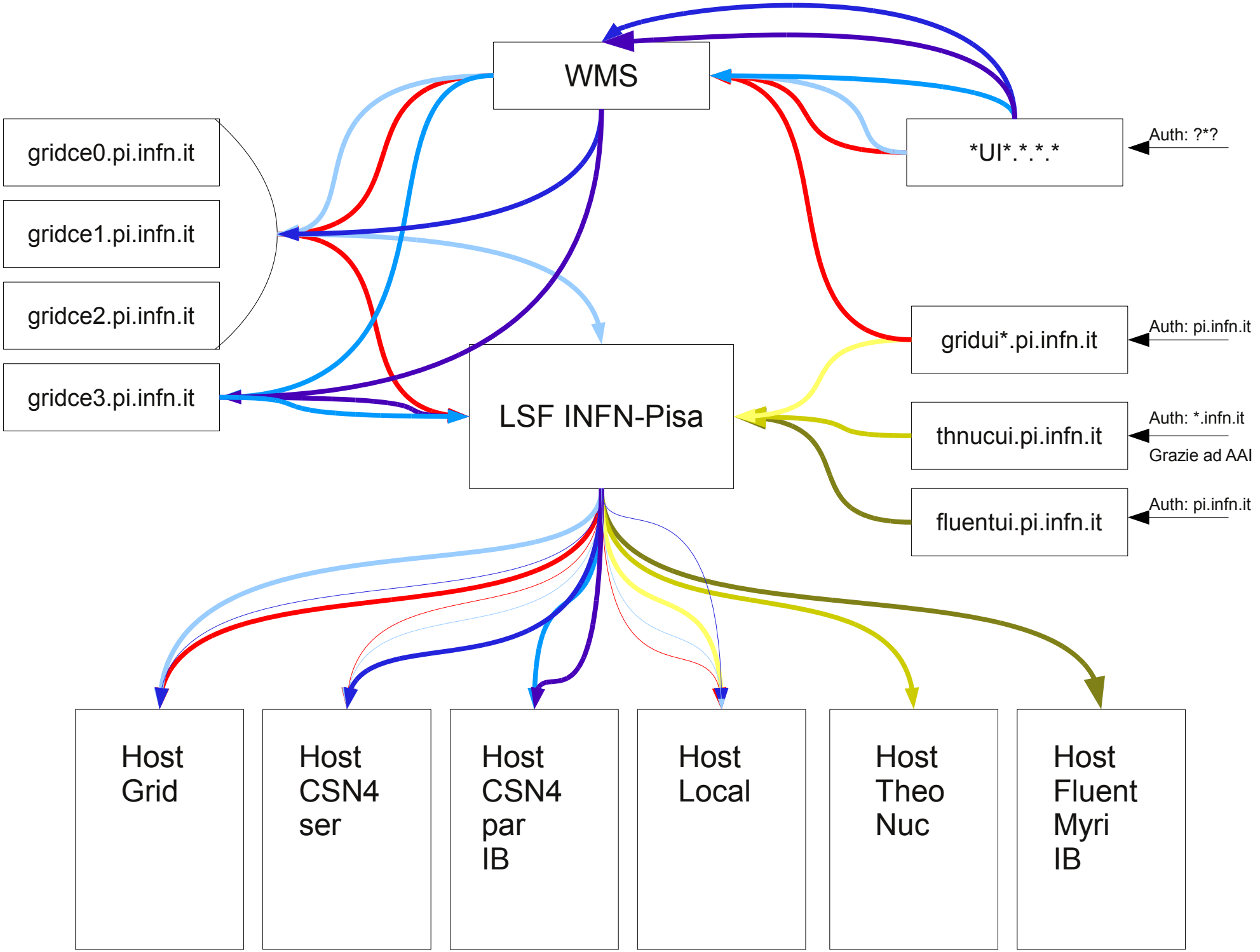
Oltre al cluster CSN4 il Grid Data Center di Pisa ospita il prototipo di un'altra facility nazionale di calcolo scientifico per la comunità afferente alla CSN4.Si tratta di una struttura di calcolo dedicata ai teorici nucleari.

- Facility “memory driven” cioè particolarmente adatta a job “memory bound” piuttosto che “CPU bound” quindi caratterizzata da WN ad elevato rapporto RAM/CORE (>8GB/CORE)
- Facility caratterizzata dall'uso via batch attraverso sottomissione locale (non GRID) di utenti afferenti a sezioni diverse. Questo è reso possibile dalla infrastruttura messa a disposizione da AAI, si tratta della prima applicazione pratica di questo progetto nell'ambito del calcolo scientifico
- La facility sfrutta tutta l'infrastruttura del Grid Data Center, in particolare per quello che riguarda l'accesso al disco.


gridce0.pi.infn.it
gridce1.pi.infn.it
gridce2.pi.infn.it
gridce3.pi.infn.it
LSF INFN-Pisa
bsub -q local
gridui*.pi.infn.it
Auth: pi.infn.it
bsub -q theonuc
thnucui.pi.infn.it
Auth: *.infn.it
Grazie ad AAI
bsub -q fluent* ...
fluentui.pi.infn.it
Auth: pi.infn.it
Host Grid
Host CSN4 ser
Host CSN4 par IB
Host Local
HOST_PARTITION_NAME: LocSL5Par
HOSTS: locsl5h/
SHARE_INFO_FOR: LocSL5Par/
USER/GROUP SHARES PRIORITY STARTED
ams 36 720.000 0
nemo 12 240.000 0
others 1 0.100 48


WMS
gridce0.pi.infn.it
gridce1.pi.infn.it
gridce2.pi.infn.it
gridce3.pi.infn.it
UI.*.*.*
Auth: ?*?
Proxy: VO=CMS
JDL: Requirements = Member("INFN-PISA"
gridui*.pi.infn.it
Auth: pi.infn.it
bsub -q cms
LSF INFN-Pisa
Host
Grid
HOST_PARTITION_NAME: SL5Part
HOSTS: sl5hosts/
SHARE_INFO_FOR: SL5
USER/GROUP SHARES
opsvo 100
glastvo 5
theophysvo 5
superbvo 1
cmsvo 48
theopisa 39
others 1 0.006 1173
HOST_PARTITION_NAME: ThSerSL5Par
HOSTS: thsersl5h/
SHARE_INFO_FOR: ThSerSL5Par/
USER/GROUP SHARES PRIORITY STARTED
opsvo 10000 200000.000 0
csn4 800 15.747 216
others 1 0.018 272
HOST_PARTITION_NAME: LocSL5Par
HOSTS: locsl5h/
SHARE_INFO_FOR: LocSL5Par/
USER/GROUP SHARES PRIORITY STARTED
ams 36 720.000 0
nemo 12 240.000 0
others 1 0.100 48


WMS
UI.*.*.*
Auth: ?*?
gridce0.pi.infn.it
gridce1.pi.infn.it
gridce2.pi.infn.it
gridce3.pi.infn.it
bsub -q theophys
Proxy: VO=theophys Gr=lS_Role=parallel
JDL: Requirements = gridce3.pi.infn.it
gridui*.pi.infn.it
Auth: pi.infn.it
bsub -q theophpar
LSF INFN-Pisa
Host Grid
HOST_PARTITION_NAME:
HOSTS: sl5hosts/
SHARE_INFO_FOR: SL
USER/GROUP SHARE
opsvo 100
theophysvo 5
superbvo 1
cmsvo 48
theopisa 39
others 1
0.006 1173
HOST_PARTITION_NAME: ThSe
HOSTS: thsersl5h/
SHARE_INFO_FOR: ThSerSL5Pa
USER/GROUP SHARES PRIOF
opsvo 10000 200000
csn4 880 15
others 1 0
HOST_PARTITION_NAME: ThMpiSL5Par
HOSTS: thmpisl5h/
SHARE_INFO_FOR: ThMpiSL5Par/
USER/GROUP SHARES PRIORITY STARTED
opsvo 10000 200000.000 0
csn4 880 17600.000 0
others 1 20.000 0
SHARE_INFO_FOR: ThMpiSL5Par/csn4/
USER/GROUP SHARES PRIORITY STARTED
thtv62 100 2000.000 0
thpi11 100 2000.000 0
thpi12 100 2000.000 0
thad31 100 2000.000 0
thto61 100 2000.000 0
thog51 100 2000.000 0
thmi11 100 2000.000 0
thpr21 100 2000.000 0
thna12 10 200.000 0
thba21 10 200.000 0
thmi41 10 200.000 0
thge41 10 200.000 0
thfa51 10 200.000 0
thct31 10 200.000 0
thrm31 10 200.000 0
thmb31 10 200.000 0
LocSL5Par
Par/
RIORITY STARTED
720.000 0
240.000 0
0.100 48


WMS
UI.*.*.*
Auth: ?*?
gridce0.pi.infn.it
gridce1.pi.infn.it
gridce2.pi.infn.it
gridce3.pi.infn.it
LSF INFN-Pisa
gridui*.pi.infn.it
Auth: pi.infn.it
thnucui.pi.infn.it
Auth: *.infn.it
Grazie ad AAI
fluentui.pi.infn.it
Auth: pi.infn.it
Host Grid
Host CSN4 ser
Host CSN4 par IB
Host Local
Host Theo Nuc
Host Fluent Myri IB