Estensione della farm del CNAF su Leonardo

Diego Michelotto (diego.michelotto@cnaf.infn.it)
Andrea Chierici (andrea.chierici@cnaf.infn.it)
Alessandro Pascolini (alessandro.pascolini@cnaf.infn.it)
Daniele Lattanzio (daniele.lattanzio@cnaf.infn.it)

Indice

- Estensione Farm in Passato e Oggi
  - Bari
  - CINECA
- Tecnopolo/Leonardo
  - Tecnopolo
  - Leonardo
- Estensione Farm Leonardo
  - Batch System
  - VM
  - Networking
- Conclusioni

Estensione Farm in Passato

- **PoC** di estensione farm **su cloud provider HNSCiCloud** (2017)
  - Setup molto **complesso** e **manuale**, accesso al T1 tramite **openVPN** → Inefficiente
- **Bari** (2015- 2022)
  - **~ 670 Km**
  - **20 Nodi ~ 11KHS06/KHEPScore23**
    - 2 nodi ospitano DNSMasq per cache DNS + Frontier Squid per cache CVMFS
  - Estensione della rete tramite **L3VPN** implementata da **GARR**
  - storage **GPFS AFM** per cache GPFS a sistemi del T1
    - **Dismesso per inefficienza** del sistema
    - Situazione adatta solo ad **Alice**, che non fa uso di Storage POSIX su GPFS del T1
- **CINECA** (2018 - OGGI)
  - **~12 Km**
  - **504 Nodi ~ 470KHS06/KHEPScore23**
  - Estensione della rete tramite **transponder Infinera su black fiber di Lepida**, al CINECA c'è la rete del T1
  - I nodi sono a tutti gli effetti nel **nostro Batch System** e hanno **acceso diretto a storage** del T1

Tecnopolo e Leonardo

- **Tecnopolo**
  - Ristrutturazione **dell'ex manifattura tabacchi di Bologna**, con fondi europei e regionali a seguito della vincita del bando per ospitare il nuovo centro metereologico europeo **ECMWF**
  - Ospita il nuovo data center **di INFN CNAF** e il data center del **CINECA** dove c'è **Leonardo**

- **Leonardo**
  - Super computer per exa-scale (300 Pflop/s)
  - **Settimo** nella TOP 500 dei super computer (**06/2024**), appena installato nel **11/2023** era **quarto**
  - 2 partizioni - dettagli QUI
    - **3456 nodi booster** 32 Intel core, 512 GB RAM, 4 Nvidia A100 64GB, 2x100 Gb/s Inifiniband
  - **1536 nodi GP** 2x56 Intel core, 512 GB RAM, 3x100 Gb/s Inifiniband, 8TB NVMe
  - **Connettività 100Gb/s Infiniband**
  - **Batch system Slurm**
  - **Storage Lustre**

Estensione Farm Leonardo

- **2800 HEPScore23** per ogni nodo della partizione GP
- **300** nodi assegnati al T1 **840 KHEPScore23**
  - Dal prossimo anno **500 nodi (~1.5 MHEPScore23)**
- Accesso tramite **Slurm**
  - **1 Job → 1 VM**
    - **112 core** (2 socket da 56 core) – No HT sui nodi per rispettare rapporto 1core/4GB RAM
    - **440 GB RAM** con pinning su CPU per rispettare la **topologia fisica del nodo**
    - **1 IPoIB 100Gb/s** con indirizzamento del T1
    - La VM **«appartiene» al batch system** del T1
      - **Accesso diretto allo storage** del T1
    - VM Integrata con tutti i sistemi del T1 (provisioning/configuration/monitoring/ecc.)
- Accesso rete del T1 tramite **Nvidia «Skyway»**
  - Gateway tra Infiniband e Ethernet
  - MTU 2k
  - **2 Skyway**, ciascuno con **8x100Gb** Ethernet collegati al T1 (tot 1.6Tb/s)

```
Welcome to:
                _                                         _
               | |                                       | |
               | |     ___  ___  _ __   __ _ _ __ __| | ___
               | |    / _ \/ _ \| '_ \ / _` | '__/ _` |/ _ \
               | |___|  __/ (_) | | | | (_| | | | (_| | (_) |
               |_____\___|\___/|_| |_|\__,_|_|  \__,_|\___/

*******************************************************************************
* Red Hat Enterprise Linux 8.7 (Ootpa)                                        *
*                                                                             *
* Booster module:                                                             *
* Atos Bull Sequana X2135 "Da Vinci" Blade                                    *
* 3456 compute nodes with:                                                    *
*       - 32 cores Ice Lake at 2.60 GHz                                       *
*       - 4 x NVIDIA Ampere A100 GPUs, 64GB                                   *
*       - 512 GB RAM                                                          *
*                                                                             *
* DataCentric General Purpose module (DCGP):                                  *
* Atos BullSequana X2140 Blade                                                *
* 1536 compute nodes with:                                                    *
*       - 2x56 cores Intel Sapphire Rapids at 2.00 GHz                        *
*       - 512 GB RAM                                                          *
*                                                                             *
* Internal Network: 200G HDR Infiniband Dragonfly+                            *
* SLURM 22.05                                                                 *
*                                                                             *
* For a guide on Leonardo:                                                    *
* https://wiki.u-gov.it/confluence/display/SCAIUS/UG3.2%3A+LEONARDO+UserGuide *
* For support: superc@cineca.it                                               *
*******************************************************************************
IN EVIDENCE:
  - A new personal area $PUBLIC is available to share installations and/or
    data. Please, keep in mind that the $PUBLIC directory is by default open
    to everybody on the cluster, and your files are visible to all users.
  - The automatic cleaning of the $SCRATCH area is NOT active at the moment
  - RCM will be available soon
  - Spack module is available to customize your software environment.
    "module av spack" to list the available versions and
    "module load spack/<version>" to use a specific one
```

Estensione Farm Leonardo

- **Stato attuale**:
  - Abbiamo **l'immagine della VM** creata a partire dalla configurazione del WN del T1 (HTCondor 23, GPFS, CVMFS, WN Software)
  - Abbiamo **script** per **istanziazione** delle VM tramite **SLURM** di Leonardo
  - VM con accesso general internet tramite **NAT di Leonardo**
  - **Sottomissione** di job dal T1 come **sito esterno** → solo job **senza accesso POSIX**

```
-- Schedd: sn-01t.cr.cnaf.infn.it : <131.154.192.159:9618?... @ 05/21/24 18:49:28
 ID       OWNER            SUBMITTED     RUN_TIME HOST(S)
51401.0   apascolinit1    5/21 18:46   0+00:01:25 slot1_1@cn-leo-003.cr.cnaf.infn.it
```

  - Uso dei **frontier squid del T1 per CVMFS**
  - **Abbiamo 1 collegamento fisico (100Gb/s) tra 1 skyway e il T1**

```
[a07cna00@login05 ~]$ squeue --me -O jobid,name,state,timeused,nodelist,reason
JOBID               NAME                STATE               TIME                NODELIST            REASON
4554439             cn-leo-003          RUNNING             3-00:10:06          lrdn3646            None
4554405             cn-leo-001          RUNNING             3-00:12:54          lrdn3737            None
4554382             cn-leo-002          RUNNING             3-00:15:57          lrdn3861            None
[a07cna00@login05 ~]$
```

- Cosa manca:
  - **Configurazione skyway** (dipendenza da ATOS) e indirizzamento delle VM con rete del T1
  - Completare **collegamento tra skyway e T1** (1,6Tb/s)
  - **Automazione di sottomissione** delle VM

Conclusioni

- Ancora tantissimo lavoro da fare
- Integrazione rallentata da altre attività, non solo lato nostro
  - **Trasloco al Tecnopolo - INFN CNAF**
  - **Manutenzioni Leonardo - CINECA**

```
================================================================================
WARNING:
- The cluster will be under partial maintenance from May 14 to May 16 and from
  June 4 to June 5.
- A limited part of the compute node will not be available.
================================================================================
```

- Da sperimentare **gestione delle manutenzioni programmate di Leonardo** (~ 1 down al mese) senza impattare sulla produzione
- Sperimentati **diversi DOWN** imprevisti
  - 1 dei quali per power cut «involontario» → corruzione FS → DOWN più di 1 settimana

Domande

рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
kiitos
dankie
nanni
nandri
dhanyavad
thank you
gracias
хвала
asante
manana
bedankt
bayarlalaa
gracie
hvala
mauruuru
köszönöm
aciū
dankon
akun
obrigada
tenki
murakoze
enkosi
djiere dieuf
tau
mochchakkeram
dziękuję
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
obrigado
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
tanemirt
grazie
najis tuke
terima kasih
rahmet
kam sah hamnida
rahmat
diolch
dhanyavadagalu
shukriya
mercê
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci