Uso HPC & Cloud per Pledge

G.Donvito – INFN-BARI

D. Spiga – INFN-Perugia

C. Pellegrino – INFN-CNAF

D. Cesini – INFN-CNAF

Use Cases di tipo Cloud supportati

- User interface carrozzate per accesso interattivo (ssh+grafico)
- Cluster k8s managed
  - Low-latency VIRGO
- Istanze IAM di esperimento managed

Use Cases di tipo Cloud supportati

- Ci sono use case che hanno bisogno di cluster complessi di VMs su OpenStack
  - Autoscaling, multi-host, file-system crittografati, reti private, etc
- O più semplici che possono essere schematizzati con uno o pochi docker (al momento usando Mesos, a brevissimo k8s) che viene eseguito ed usato in interattivo dall'utente
  - Il docker è in grado di vedere e usare il file-system condiviso (GPFS) nella farm HPC/HTC
  - Ad oggi è possibile usare tutte le risorse della singola macchina, è ancora necessaria attività per fare multi-host.
  - Molto usato per sfruttare GPU e molte CPU con ottimizzazione bare-metal
  - La gestione del k8s è responsabilità del sito:
    - Minore flessibilità per l'utente, ma zero effort IT da mettere

Use case di tipo CLOUD

Use case supportati oggi sono molto trasversali e coprono attività di R/D fino ad attività core di commissione

- Progetti Esterni (EU o Nazionali)
- Progetti PNRR
- Attività di commissione
  - CSN1 – 2 – 3 e 5

- Solo il CNAF ha PLEDGE su risorse di tipo CLOUD il resto sono nel contesto del catch-all (inKind)

Use Cases di tipo Cloud supportabili

- Cluster K8s on demand
  - Che implica gestione sysadmin
  - Con possibilità di estensione trasparente su HPC/HTC (i.e. Leonardo con Spoke0)
- Cluster Dask (Jupyter + Dask)
  - Con possibilità di estensione "trasparente" [offloading] su risorse NON cloud (HPC/HTC)
  - Use case in consolidamento PNRR (use case AF)

Use Cases di tipo HPC Supportati

- Partizione dedicate di HTCondor per supportare job «fat» che possono usare single-host (fino alla macchina intera, o richiedendo GPU)
- Da qualche mese, HTCondor ha una partizione con nodi multi-host collegati con infiniband
  - È possibile sottomettere job con più di un nodo usando MPICH su infiniband.
- L'allocazione delle risorse è dinamica e flessibile
  - Si allocano solo quelle CPU, GPU, memoria che sono richieste
- Questa possibilità, usando accesso locale al Batch-System, è già usata da diversi esperimenti di GR2 e GR5 soprattutto per richieste multi-cpu e GPU.
  - Nonché dallo use case di Terabit

Use Cases di tipo HPC supportabili

- Oggi:
  - Job single-host fino a 256Core virtuali, e 2-3 GPU A100 e V100
  - Infiniband connected multi-host
- Quando le Bubble saranno in produzione:
  - Job single-host fino a 384Core virtuali, e 4 GPU H100
  - Infiniband connected fino a 8 host

Soluzioni di Storage supporatate/supportabili (Cloud e HPC)

- Storage S3 eventualmente multiple istanze su più siti
  - Federabile attraverso soluzioni basate su Rucio (anche in configurazioni miste Grid+S3)
- Storage di "tipo Grid"
  - Federabile attraverso soluzioni basate su Rucio
  - accedibile da cloud
    - Con protocolli xrootd, webdav …
    - Montabile posix like a user level (no performances)
- Volumi per VM e/o FS per servizi/cluster su cloud
  - Su cloud (OpenStack) abbiamo diverse QoS:
    - Raid // triplice copia
    - SSD // magnetici
    - Backup on demand (tape o offsite)
- Sync&Share on demand
- GPFS (stesso sito):
  - Usabile anche in posix da docker (k8s) [good-performance]
  - Su VMs via NFS (OpenStack) [bad-performance]

Use cases / Richieste progetti ICSC

- 7 richieste per HPC Bubble
- 7 richieste per risorse grid (managed, complessivi 6344 core)
- 15 richieste per risorse cloud (self-managed+Bubble, complessivi 6652 core)
- 6 richieste di storage (grid+cloud)
  - Complessivo disco 1 PB
  - Complessivo tape 5.2 PB

Costi

Costi stimati per HPC/Cloud

CPU		Dati da acquisto CONSIP
€/core	105.64 €	
kW/core	0.0042	
Tempo di ammortamento (giorni)	1095.00	3 anni
€/core/hour ammortamento	0.0040	
€/core/hour consumo elettricità	0.0014	
€/core/hour personale	0.0004	
€/core/hour overhead	0.0006	
€/core/hour totale	**0.0064**	
€/core/mese totale	**4.6086**	

GPU		Dati da gara Terabit - no iva
€/GPU	19283.50	1/4 del costo del nodo con 4 GPU, da aggiungere switch
kW/GPU	700.00	consumo della sola GPU a piena potenza
Tempo di ammortamento (giorni)	1095.00	3 anni
€/GPU/hour ammortamento	0.7338	
€/GPU/hour consumo elettricità	0.2385	
€/GPU/hour personale	0.0080	
€/GPU/hour overhead	0.0137	
€/GPU/hour totale	0.9939	
€/GPU/mese totale	715.6339	

Fornitura HPC Bubbles

Gara "HPC Bubbles"

- Accordo Quadro Nazionale
  - Listino prezzi per nodi + accessori
  - 2 anni di validità
  - Lotto1
    - CPU, GPU, FPGA
  - Lotto2
    - Storage
  - Sedi Coinvolte: CNAF, BARI, MI-BI, PI, TO, LNGS, NA, RM1, PD/LNL
- **Stato gara**
  - **Ordini inviati a parte 6/5**
    - **MI e LNL anticipati su capienza ordinaria**
  - **HW arrivato**
    - **CNAF**
  - **HW installato**
    - **PD, Torino**
  - **Mancano su 6/5**
    - **CT L1+L2, LNFESA L2, ROMA1 L1, NA L2**

Quantità nodi con fondi Terabit-ICSC-DARE

	Nodo CPU	Nodo GPU	Nodo FPGA Xilinx	Nodo FPGA Terasic	Nodo storage
BA	24	6	0	0	32
CNAF	26	30	2	2	52
MIB	0	0	2	2	0
NA	18	1	2	0	8
PD	6	6	0	0	0
PI	20	0	0	0	0
RM1	12	0	0	0	0
TO	14	6	0	0	0
LNGS	0	6	0	0	12
CT	12	0	0	0	8
LNF	12	0	0	0	0
LNFESA	8	6	0	0	6
LNL	4	0	0	0	0
MI	4	0	0	0	0
TOTALE	160	61	6	4	118

Core: 30 kcore fisici
Circa 34 HS/core

GPU: 244 NVIDIA H100
40 FPGA
InfiniBAnd 400Gbs

45 PB RAW

HPC Bubbles

 Nodo CPU	192 core fisici 1.5TB RAM DDR5 IB NDR 400G 20TBL (SSD) + dischi di sistema
 Nodo GPU	Come CPU + 4x NVIDIA H100 SXM5 con minimo 80GB e memoria HBM2e
 Nodo FPGA	32core RAM 768GB DDR5 IB NDR 440G 4 x XILINX U55C o 4 x TerasicP0701
 Nodo Storage (CEPH Bricks)	64 core fisici 1TB RAM DDR5 384 TBL HDD + 25.6 TBL NVMe
 Accessori	Switch IB, Switch ETH Cavi IB, Cavi ETH Transceiver vari Assistenza 3+2

Provisioning Bubble: soluzioni valutabili

- Bubble integrata in un sistema Batch System o di dipo HTCondor o di tipo Slurm
  - Accesso locale
  - Sarà possibile accederle da remoto in entrambe i casi o in uno dei due?
  - Accesso dati da storage grid come sempre
- Bubble integrata in un sistema Openstack e/o K8s
  - Qui esisterà la possibilità di istanziare servizi self managed o saranno solo servizi managed centralmente ?
  - Accesso dati ?
  - Multi host?
- Bubble integrata in batch system ma accessibile da interfaccia cloud attraverso il meccanismo dell'offloading.
  - Questo significa che deve esistere almeno una istanza cloud con k8s e che k8s si può estendere dinamicamente sulla bubble. L'utente vede cloud
  - Accesso dati da storage grid come sopra
  - Accesso dati da storage cloud via protocollo S3 o mount posix a user level (no performances)

Richieste 2025
HPC + Cloud

Richieste 2025 – Cloud e HPC

[DarkSide] CPU-Cloud: 1000 HS06	10.00
[Limadou] CPU-Cloud: 200 HS06	2.00
[QUAX] CPU-Cloud: 300 HS06	3.00
[DarkSide] DISK-Cloud: 100 TB	10.00
[Limadou] DISK-Cloud: 10 TB	1.00
[Auger - DataCenter] GPU: 8640 GPU hours	6.00
[CYGNO] GPU: 2400 GPU hours	1.50
[Limadou] GPU: 2037 GPU hours	1.50
[XRO] GPU: 4320 GPU hours	3.00
[ET] HPC: 100k core*hours	0.50
[Euclid] HPC: 7M core*hours	26.00
[LSPE] HPC: 500k core*hours	2.00
[QUBIC] HPC: 200k core*hours	0.50

SPES-MED - CPU 3000HS06 (Cloud?)	30.00

ARTEMIS - GPU: 2000 GPU-hours GPU A100 (> 32 GB di VRAM) per training modelli AI	1.50
AIM_MIA - GPU: 9000 GPU-hours suddivise su 2 x GPU A100 (> 40GB di VRAM) per training modelli AI	6.50
Geant4INFN - GPU: 10000 GPU-hours tramite piattaforma AI_INFN	7.00
GEANT4INFN - CPU: 120k core-hour per simulazioni geant4_DNA richiesta una macchina con almeno 128 GB di RAM e accesso tramite coda condor su tier1	1.50
GEANT4INFN - Storage 2 TB accessibile posix da infn cloud	0.50
BRAINSTAIN - CPU: 400k core-hours	4.50
BRAINSTAIN - GPU: 10k GPU-hours	7.00
BRAINSTAIN - Storage: 8 TB	1.00
FRIDA - CPU: 1M core-hours (2GB/core) per attivitaÃ¢ÂÂ Roma1: 200k core-hours (6GB/core) per attivitaÃ¢ÂÂ Roma3-TIFPA	13.50
FRIDA - GPU: 1440 GPU-hours (GPU NVIDIA A100 con VRAM ~ 10 - 20 GB) per simulaz. FAST MC	1.00
SPRITZ - CPU: 125000 core-hours per simulazioni multithread (40-80 cores) con shared-memory (256-512 GB).	1.50
SPRITZ - Storage: 150 TB con protocolli di accesso standard basati su ssh	15.00
PLASMA4BEAM2 - Storage: 100 TB	10.00
PLASMA4BEAM2 - CPU: 500000 core-hours	6.00
SEGNAR - CPU: 100000 core-hours	1.00
FUSION - CPU: 240000 core-hour	3.00
FUSION - GPU: 20000 GPU-hours	14.00
FUSION - Storage: 1 TB	0.50
VITA - CPU: 1400000 core-hours	16.00
VITA - GPU: 3000 GPU-hour	2.00
VITA - Storage: 700 TB (una parte per dati sensibili)	70.00
MIRO - CPU: 1.2M core-hours [500k core-hours per simulaz. MC (Catania); 500k core-hours per simulaz. dinamica molecolare (Pisa); 200k core-hours per simulaz. MC e analitiche multiscala (TIFPA)]	13.50
MIRO - GPU: 2200 GPU-hours GPU NVIDIA A100 (> 32 GB di VRAM) per training modelli AI	1.50
SEGNAR - Storage: 10 TB	1.00

Al fine di indirizzare le richieste...

- Cloud
  - Serve solo IaaS: esempio VM da gestire come sysadmin alla HERD/AMS ?
  - Serve un servizio "complesso": esempio un Jupyter con accesso a batch e/o storage?
  - Il servizio deve essere mantenuto centralmemte (servizio di tipo SaaS)?
  - O la gestione sarà autonoma (sysadmin)
    - Sarà un servizio multi-tenancy? Se si serve integrarlo con un token issuer di esperimento?
  - Si prevede un uso interattivo?
    - Un Jupyter o simili su singola macchina, magari anche con tante risorse richieste?
  - Serve accedere a dati presenti in storage esistenti (i.e. ti tipo grid?) o storage cloud S3?
    - Ci si aspetta protocollo posix?
- HPC
  - Serve una macchina carrozzata con disco veloce e GPU per utilizzo interattivo?
  - Serve un sistema batch con accesso a storage, MPI?
  - Servono risorse HPC accessibili da servizi cloud? (i.e. training MLFlow? Dask? hyperparameter optimization?

Backup-slides

Use cases / CLOUD

- AI4EOSC
- AI_INFN
- ALICE
- AMS-02
- CCL-BOLOGNA / Tier 3
- CYGNO
- DARKSIDE
- DODAS
- EA-CCS
- EGIFED
- EGIOPS
- ESCAPE
- FERMI
- HERD
- ICSC-CN1-SPOKE2-ANALYSISFACILITY
- INACTIVE-SARCOV2
- INCIPIT
- I-NERGY
- INFN Cloud Catchall
- IOT
- IOTWINS
- NGS-MM
- QUAX
- SmartChain
- SouthCloud
- STARWAI
- SUPER
- SISINFO
- XDC
- VIRGO

Federati INFN-Cloud