Distributed Systems

Doina Cristina Duma

Workshop SDDS – 18/01/2018

Panoramica

- Personale
- Gestione risorse SDDS (hard & soft)
- Progetti
  - interni/locali CNAF
  - INFN
  - Nazionali
  - Internazionali

Personale

- Vincenzo **Ciaschini** - T.I.
- Alessandro **Costantini** - T.D. (Luglio 2018)
- **Doina Cristina Duma** - T.D. (Dic. 2018, idoneità concorso T.I. 18887)
- Diego **Michelotto** - T.D. (Ago. 2018, vincitore CNAF-T3-657, idoneità concorso T.I. 18887)
- Matteo **Panella** - T.D. (Mar. 2019)

Gestione risorse SDDS

Short term – gestione hardware & software as-is
Long Term:

- Definire procedure chiare per il rinnovo parco macchine
- Ottimizzare uso risorse – vari cluster, varie tecnologie
- Migliorare qualità livello PaaS

Gestione risorse (hard) SDDS (1)

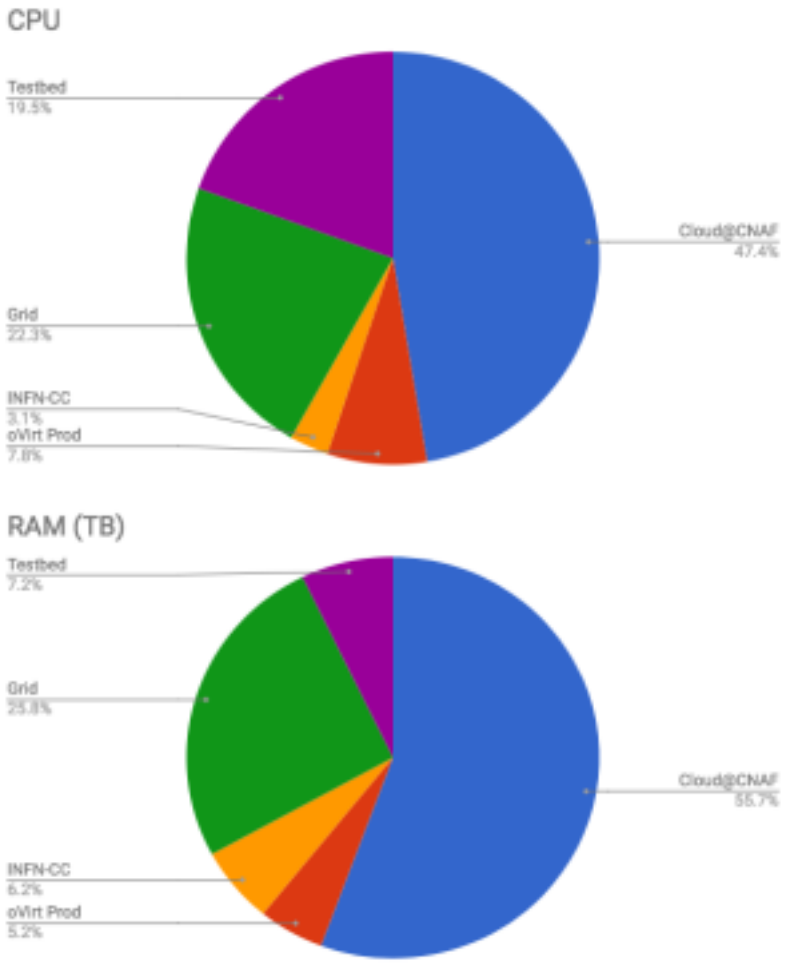
- Totale risorse a disposizione SDDS: **195 server fisici con ~ 3700 core (con HT dove disponibile) e ~ 10TB di RAM**
  - AMD **2012**: 68 server con ~ 1600 core e ~ 5TB di RAM **(PRODUZIONE)** (44 Donate dal Farming nel 2017, Cloud@CNAF + WN IGI BOLOGNA)
  - Intel **2012**: 12 server con ~ 280 core e ~ 500GB di RAM **(PRODUZIONE)** (oVirt di produzione)
  - AMD **2013**: 44 server con ~ 700 core e ~ 2,5TB di RAM **(PRODUZIONE)** (Cloud@CNAF) **(Sotto manutenzione fino agosto 2018)**
  - Intel **2013**: 7 server con ~ 200 core e ~ 400GB di RAM **(PRODUZIONE)** (Cloud@CNAF)
  - Intel **2017**: 3 server con ~ 110 core e ~ 600GB di RAM (INFN-CC) **(Sotto manutenzione fino 2022)**
  - Intel < **2012**: 61 server con ~ 700 core e 1,2TB di RAM (Testbed)

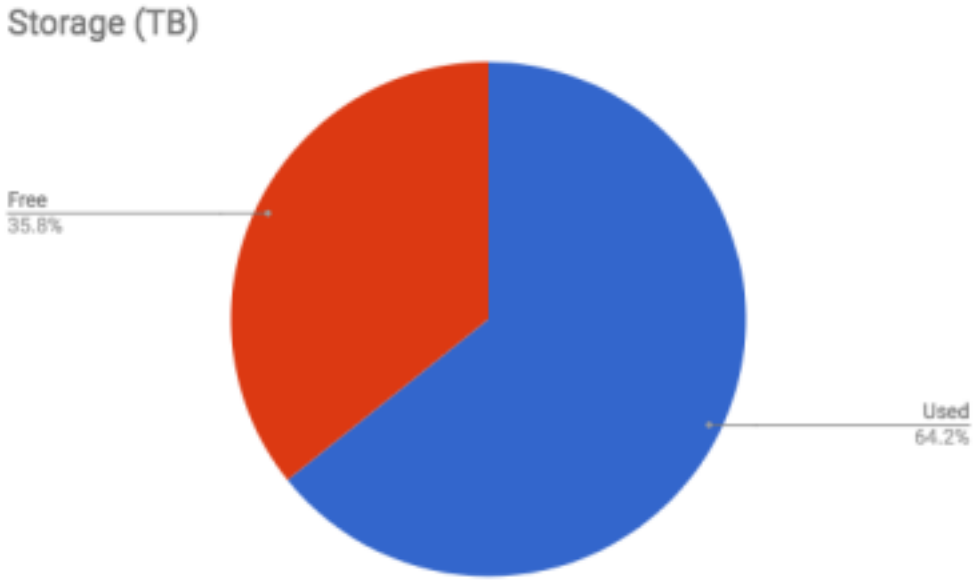
Gestione risorse (hard) SDDS (2)

Storage - total = 386 TB RAW, di cui FREE 138 TB

- **Equallogic** PS6010 16x600GB SAS 15K ~ **9TB** RAW con RAID 50 = **6TB,** di cui 3TB FREE (oVirt di Produzione)
  - **Sotto manutenzione fino a Novembre 2018**
- **Powervault**
  - MD3660i: **60x4TB** NLSAS 7.2K (**Sotto manutenzione fino Aprile 2020**)
  - MD3600e: **10x6TB** NLSAS 7.2K + 10x180GB SSD (**Acquistato da EEE**, **sotto manutenzione fino Gennaio 2020**)
  - **300 TB** NLSAS RAW + **1,8TB** SSD RAW
  - ~ 60TB FREE NLSAS, 100% USED SSD
- Storage **SuperMicro** (INFN-CC) 2xSSD 148GB (Sistema) + 6xSSD 500GB = **3TB** RAW(Ceph) + 12xHDD 6TB = **72TB** RAW (Ceph) (**Sotto manutenzione fino Luglio 2022**)

Gestione risorse (hard) SDDS (3)





Gestione risorse (hard) SDDS (4)

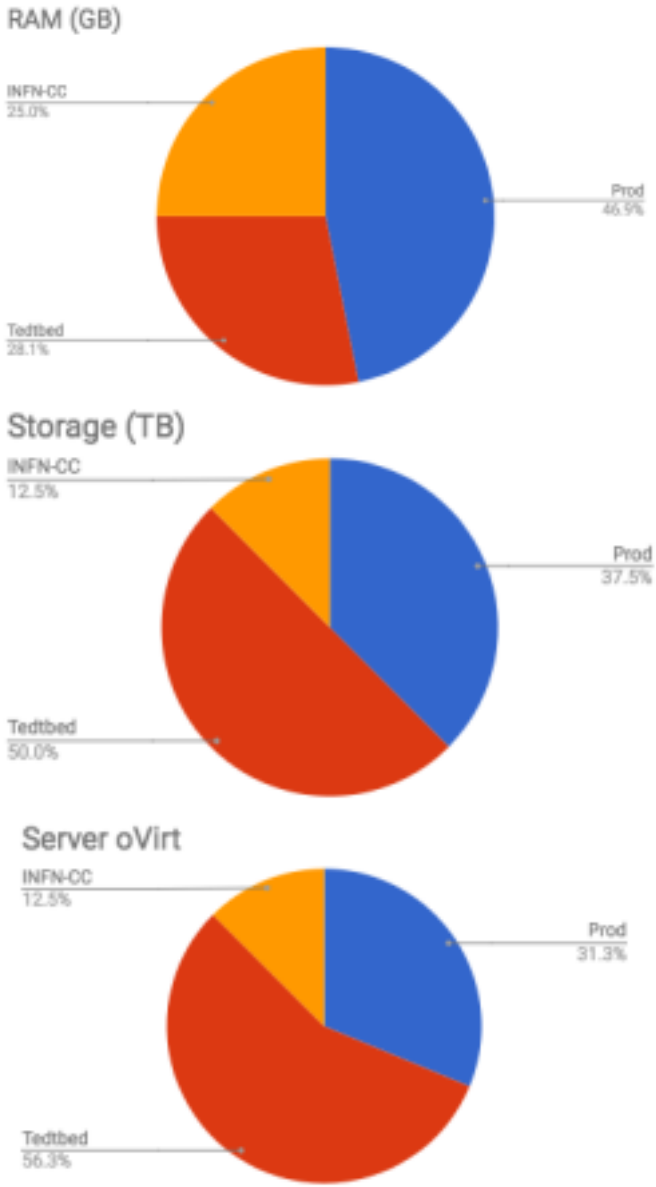
Network

- **Router SDDS**: VDX 6740 48 porte 10Gb/s BASE-SR + 4 porte 40Gb/s BASE-SR (EQL)
- 4 switch 48 porte 1Gb/s BASE-T + uplink 10Gb/s BASE-SR
- 2 switch 24 porte 1Gb/s BASE-T + 2 10Gb/s BASE-T + uplink 10Gb/s BASE-SR (PV)
- 5 switch 48 porte 1Gb/s BASE-T + uplink 1Gb/s BASE-SR
- 1 switch 48 porte 10Gb/s BASE-T + uplink 40Gb/s BASE-SR
- 2 switch 16 porte 1Gb/s BASE-T + uplink 1Gb/s BASE-SR

Gestione risorse (soft) SDDS (1)

oVirt

- DC **Production**
  - 10 Server con 240 core e 480GB RAM
  - 3TB da Equallogic
  - **72 VM** (Servizi GRID, Servizi SDDS, Servizi CNAF, Servizi INDIGO, Servizi Bebop)
- DC **Testbed**
  - 18 Server con 144 core e 288GB RAM
  - 4TB da PowerVault
  - **79 VM** (Vari testbed, Servizi Cloud@CNAF)
- DC **INFN-CC**
  - 4 Server con 64 core e 256GB RAM
  - 1 TB da PowerVault
  - **6 VM** (Servizi INFN-CC)



Gestione risorse (soft) SDDS (2)

Openstack

- Produzione
  - 3 MySQL Percona DB, 5 RabbitMQ (VM), 2 HAProxy + Keepalived (VM)
  - 16TB GPFS da PowerVault (Dischi Dedicati + Metadati su SSD)
  - 2 Controller Node, 2 Network Node
  - **65 Compute Node: ~400 Core, ~4.6TB RAM, 16TB shared Storage e 44TB Local IO Storage**
  - 164 VM
- Testbed Upgrade
  - 3 MySQL Percona DB (VM), 3 RabbitMQ (VM), 2 HAProxy + Keepalived (VM)
  - 2TB GPFS da PowerVault
  - 2 Controller Node, 2 Network Node, 2 Compute node

Gestione risorse (soft) SDDS (3)

Grid

- **NGI_IT**
  - Gestione Alias DNS su dominio grid.cnaf.infn.it + Check (sensu) di gestione automatica dell'alias per:
    - **2 LFC per 26 VO**, composto da 2 VM ciascuno con HA via DNS, dati su DB MySQL Percona. (**In dismissione**)
    - **TopBDII** egee-bdii.cnaf.infn.it 5 host (4 VM al CNAF 1 a PD)
    - **WMS** wms-multi.grid.cnaf.infn.it per **31 VO**, 7 host (4 Macchine CNAF, 2 CT, 1 FE) (**In dismissione)**
    - **MyProxy** - myproxy.cnaf.infn.it 2 host con shared filesystem su GlusterFS (2 VM al CNAF)
    - **Squid** - squid.grid.cnaf.infn.it 2 host (2 VM al CNAF)
  - **3 VOMS per 28 VO**, 2 repliche a PD e 1 a NA
  - **Argus** centrale per NGI_IT (1 VM al CNAF)
  - **IGI Portal** (1 Macchina e 3 VM)

Gestione risorse (soft) SDDS (4)

Grid

- **IGI-BOLOGNA**
  - 24 WN ~500 job slot per 5 VO
  - 1 SE/StoRM da 10 TB per 6 VO
  - 1 CE (VM)
  - 1 Argus di sito (VM)
  - 1 SiteBDII composto da 2 VM in HA tramite DNS

VMWare

- 2 CVMFS server

Gestione risorse (soft) SDDS (6)

- Cluster **GPFS**
  - 2 Server, 1 quorum node, 67 client con accesso diretto ai dischi
  - 2 NSD da 8TB da 2 Disk Group da PowerVault ciascuno da 4 Dischi da 4TB in RAID 10 per dati
  - 2 NSD da 180GB da 2 Disk Group da PowerVault ciascuno da 2 Dischi SSD da 180GB in RAID 1 per metadati
  - 12 NSD da 3TB da Volume Group da PowerVault per un totale di **32TB per FS EEE** (su 60TB richiesti dal progetto)
- **GlusterFS**
  - 3 Nodi
  - 6 volumi in replica 3, ogni nodo monta un disco da PowerVault per ogni volume per un totale di 2.3TB per nodo
    - Volumi usati per **MyProxy, Central Logging servizi Grid, Jenkins, MPI shared home**.

Gestione risorse (soft) SDDS (7)

- **CEPH**
  - 3 MON server virtuali
  - 4 OSD server con ciascuno un osd da 2TB
  - Installazione automatizzata del cluster via Puppet
  - Test di valutazione Ceph come storage backend di Openstack
- Testbed **Mesos/Marathon**
  - 3 nodi fisici ciascuno con ruolo mesos-master, mesos-slave, etcd, marathon e chronos.
  - Test di integrazione con Jenkins.
  - Test di autoscaling dei worker HTCondor.
  - Test preliminari del progetto HTMesos (Vedi slide succesive)

Gestione risorse SDDS

Risorse perse causa allagamento

- 4 Compute node Cloud@CNAF
- 1 DB server Cloud@CNAF + 1 lavato e recuperato
- 1 Storage INFN-CC
- 2 Macchine portale IGI (recuperato immagine disco)
- 1 m1000e con 8 M600 (Dell)

Risorse a disposizione dopo l'allagamento

- **Cloud@Bari** - K8s@MW (7 VM), Tool INDIGO IAM + tool collaborativi (4 VM)
- **SDDS@Ferrara** (Ripristinata la produzione di SDDS)
  - Hardware
    - PowerVault + EqualLogic > 300TB RAW
    - 4 Switch
    - 2 Network Node
    - 2 Compute Node
    - 2 GPFS server
    - 1 Backup Server
    - **3 Huawei Farming (24 server con 40 core e 128GB)**
      - 5 nodi oVirt, 3 nodi DB, 1 nodo di servizio, 1 nodo cloud-ui, 9 compute node, 5 Nodi spare
  - Servizi
    - NGI_IT
    - Cloud@CNAF (EEE, FAZIA/US, INDIGO PaaS)
  - Consumi inferiori a 14kW dei 20 a disposizione.

Grid

Partecipazione EGI

- Coordinamento NGI_IT, sito IGI-Bologna, Servizi core
- Partecipazione EGI-CSIRT (Computer Security Incident Response Team)
- IGI-Portal - user support
- **Attività**
  - IPv6 – RC readiness plans
  - Decommissionig campaigns - dCache 2.13, WMS EOL -1 Jan. 2018 (calet.org, compchem, theophys, virgo, xenon)
  - Storage Accounting Deployment (DPM&dCache)

Cloud

- **Cloud & CNAF**
  - **Miglioramenti**
    - **Affidabilità -** Deployment nuovi cluster DB, RabbiMQ, separati
    - **Qualità - Aumentato risorse**
      - 20 Hypervisors (compute) con storage locale (40TB)
      - Risorse totali: - 66 hypervisors/compute nodes
  - **Nuovi use case** – FAZIA (~15VM, 256CPU, disco 1,2TB+2 , 5TB (16x160GB)
  - **Attività :**
    - Integrazione INDIGO-IAM
    - Deployment INDIGO PaaS – Orchestrator, Infrastructure Manager, SLA Manager, CMDB,..
    - KUBE = messa in produzione di un cluster Kubernetes per deployment servizi-in-container (IAM, TTS)
    - CEPH come backend per alcuni servizi OpenStack
    - Miglioramento collegamenti connessioni Computes<->PV – 2 nuovi sw per rack 206-10
    - Accounting – Ceilometer+Gnochi
    - Monitoring – definizione nuove probe/check “cloud” per Sensu
    - Upgrade (Mitaka->Newton->Ocata->Pike)

Strumenti per la gestione risorse SDDS

- **Foreman/Puppet - provisioning e configurazione automatica**
  - Contributo alla realizzazione e utilizzo dell'infrastruttura comune per il CNAF (Vedi progetto Bebop)
  - Servizi specifici per SDDS gestiti via foreman-provy
    - DHCP
    - TFTP
    - DNS
- **Sensu, Uchiwa, Grafana - monitoring & allarmistica**
  - Contributo alla realizzazione e utilizzo dell'infrastruttura comune per il CNAF (Vedi progetto Bebop)
  - Servizi specifici per SDDS
    - Sensu-server
    - Uchiwa dashboard
    - Grafana dashboard
    - InfluxDB
- **Rundeck** - Integrato con Foreman, per le risorse SDDS.

Progetti

Progetti - interni/locali - CNAF

- Dynamic Farm
- Data Housing
- WNodeS
- BEBOP

Dynamic Farm

- Permette di usare WN esterni al CNAF come se fossero interni
  - **Collaborazione con T1-Network**
  - Usato per WN su Azure, Aruba, HNSciCloud
  - Requisiti sulle macchine remote minimi (outbound connectivity, subnet su cui mettere le macchine)
  - Sui forniscono anche tool di amministrazione batch remota

DataHousing

- Sistema per permettere l'accesso senza certificato ai nostri server di dati (basati su gridftp)
  - **Collaborazione con User Support**
  - Non richiede installazione
  - Autenticazione basata su Kerberos del CNAF
  - Gestione dei certificati/proxy VOMS completamente nascosta
  - Per lo Rep. Storage un esperimento e' semplicemente una nuova VO
  - Aggiornamento automatico di versione in versione
  - Testato/Usato da cuore, dampe, km3, etc…
  - Correntemente versione 12/client 1.4.5/server

WNodeS

- Creazione on demand di VM per esperimenti con necessita' di setup particolari.
  - Attualmente sospeso causa fine dello use case per il T1
  - Prima della sospensione era in mantenimento
    - Non c'erano bug o problemi noti
    - Non c'era necessita' di nuove feature

BEBOP

Partecipazione al gruppo trasversale del CNAF per la progettazione e realizzazione delle infrastrutture centralizzate per:

- **Provisioning**
  - Utilizzo di Foreman e Puppet per il provisioning e la configurazioni della macchine del CNAF
  - Dimensionamento dei servizio per i numeri del CNAF
  - Alta disponibilità dei servizi
- **Monitoring**
  - Realizzazione delle infrastrutture centrali per il supporto del monitoring basato su Sensu
    - Cluster RabbitMQ per comunicazione tra server e client di sensu
    - Redis + Redis Sentinel archiviazione stato dei check
    - InfluxDB per archiviazione di metriche (Time series DB)
  - Realizzazione di classi puppet per i vari reparti del CNAF per installazione e configurazione di Sensu (1 o più per ogni reparto)
- **Logging**
  - Realizzazione di testbed, ospitati da SDDS, per la realizzazione di una infrastruttura comune al CNAF per archiviazione ed indicizzazione di log basati sullo stack ELK.

Risorse SDDS: - 1 Sensu Server, 1 InfluxDB, 15 VM oVirt + 1,5TB da PowerVault per Logging, Provisioning e Logging upgrade testbed su Cloud@CNAF 10VM

Progetti - INFN

INFN-CC

- a geographically distributed private cloud infrastructure, based on OpenStack, that has recently been deployed in three of the major INFN data-centres in Italy, fully redundant and resilient architecture to provide critical network services for the INFN community
  - deploy a PaaS layer by adopting services developed within the EU funded project INDIGO-DataCloud
- **Attività**
  - progettazione dell'architettura
  - Provisioning e manutenzione delle risorse & servizi:
    - Foreman/Puppet, openVPN, DNS, DHCP, HAProxy, Cluster MySQL, CEPH, Zabbix (server + proxy), Logging (syslog-ng)
    - Openstack (Keystone, Glance, RabbitMQ)
- **Collaborazione - SDDS, Sistemi Nazionali, Rete**

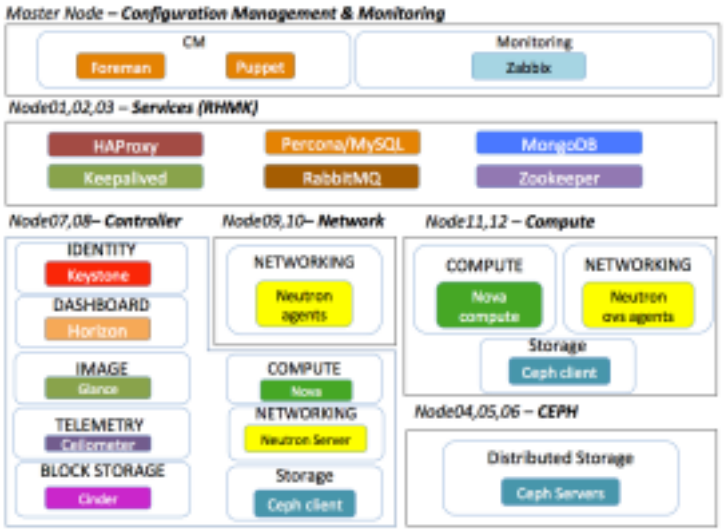
HTMESOS

- Implementazione di un servizio di “Batch Farm on-demand” a livello PaaS basato su HTCondor e Mesos.
  - i vari daemon HTCondor - Docker container - preconfigurati e distribuiti come Long Running Service (LRS) attraverso Marathon
- **Risultato**: INFN-TO -> **Open Computing Cluster per Advanced Data Manipulation (OCCAM)**, una struttura HPC volta a fornire un'infrastruttura flessibile e multifunzionale (Poster @ INDIGO Summit 2017)
- **Attività**
  - Contributi alla progettazione architettonica e alla creazione di contenitori, config. con ansible => **DODAS**
  - Provisioning e manutenzione delle risorse

Progetti Nazionali

- OCP - finito
- Cagliari2020 - possibile

OCP



Progetto finanziato dal MIUR ,a guida INFN, partecipazione pubblico/privato - finito

- **ricerca, sviluppo e sperimentazione** di nuove soluzioni tecnologiche open, interoperabili utilizzabili on-demand nell'ambito del Cloud Computing per le Pubbliche Amministrazioni
- **ultimi deliverables e SAL (stato av. Lavori) per tutti I semestri**
- Validazione piattaforma OCP, in particolare la parte IaaS,

preso le enti/regioni sperimentatrici

Attività

- **Automazione**
  - Sviluppo di strumenti e procedure per installazione automatica di tutti i componenti IaaS e loro upgrade
    - **AutomaticOCP: Know-how applicato per l'upgrade di Cloud@CNAF (Juno->Mitka)**
- Monitoring applicativo - **interfaccia web** per il monitoring applicativo
- Supporto e documentazione
  - Supporto per le attività di installazione e configurazione IaaS nei testbed regionali
  - Supporto alle attività di **mantenimento e risoluzione problematiche** per i testbed regionali
  - Scrittura **documentazione** e mantenimento (ove necessario)
    - Guide installazione, Documenti di validazione, Presentazioni Enti e Regioni
- Corsi OpenStack

Cagliari2020

Progetto finanziato dal MIUR a guida Crs4 (Centro di Ricerca, Sviluppo e Studi della Sardegna) che ha tra i capofila INFN-Cagliari, Partecipazione pubblico/privato

- Sviluppo di strumenti e soluzioni tecnologiche per l'ottimizzazione della mobilità urbana
  - Def. requisiti per soluzioni cloud Iaas e SaaS
  - **Sviluppo di specifici strumenti per l'integrazione su piattaforme cloud delle applicazioni per l'analisi e l'elaborazione dei dati e supporto all'operatività**.
    - Studio di soluzioni cloud coerenti con le esigenze di progetto & Porting delle applicazioni su infrastrutture IaaS/Saas

Attività

- installazione di una infrastruttura cloud OpenStack via **AutomaticOCP**
- descrizione dei deliverable INDIGO su requirement tecnologici (IaaS, PaaS e SaaS)
- integrazione di INDIGO **IAM** in OpenStack, installazione/uso **TTS** (WaTTS & WaTTSon)
- **PaaS** - installazione IM, Orchestrator
- installazione di **Oneprovider** (agganciato alla onezone del CNAF) e prove di Onedata per accesso dati - eventualmente data replication
- istanziazione di cluster Mesos / batch system on-demand via TOSCA template (DODAS o simile)
- **udocker** - uso; portali - FG

Progetti Internazionali

- INDIGO - DataCloud - finito
- EOSCpilot - in corso
- XDC, DEEP, EOSChub - nuovi

INDIGO-DataCloud

“Develop an open source data and computing platform - targeted at scientific communities - helping developers, resources providers, e-infrastructures and scientific communities to overcome current challenges in the Cloud computing, storage and network areas”

Attività:

- Software lifecycle management
  - Release - CI, repositories, documentation
- Support Services - OpenProject, agenda, owncloud ...
- Preview testbed
  - PaaS
  - Onedata - CMS/DODAS, *ICCU - MuseID (e-Cultural Science Gateway)*

EOSCpilot

The European Open Science Cloud for Research Pilot Project

- Develop **demonstrators** of **integrated services** and **infrastructures** in a number of scientific domains, showcasing interoperability and its benefits;

Objectives:

- Research and Data Interoperability
- Infrastructure interoperability - Demonstrate with multi-infrastructure, multi-community pilots

Key Output: - The design of a future EOSC based on federated interoperable services meeting the needs of the thematic research domains and wider user base

Tasks:

- T6.1: gap analysis & interoperability architecture [CNRS lead]
- T6.2: EOSC Research and Data interoperability [ELIXIR lead]
- T6.3: Interoperability pilots [INFN lead]

Nuovi progetti

eXtreme – DataCloud
- developing scalable technologies for federating storage resources and managing data in highly distributed computing environments, as required by the most demanding, data intensive research experiments in Europe and worldwide

DEEP-Hybrid – DataCloud
- support intensive computing techniques that require specialized HPC hardware tto explore very large data
- deploy under the common label of “DEEP as a Service” a set of building blocks that enable the easy development of applications requiring these techniques: - deep learning using neural networks; parallel post-processing of very large data; analysis of massive online data streams.

EOSCHub
- Integrating and managing services for the European Open Science Cloud

Attività
- Preparazione proposta progetto XDC & DEEP
- XDC & DEEP
  - WP1 - Project management
  - WP3 - SW management, SQA, Pilot testbed (inclusi tool – collaborativi)
- EOSCHub - supporto DODAS (Thematic Service)