Software Heritage

preserves software source code for
present and future generations



Il mirror di ENEA

una replica sperimentale



SOFTWARE HERITAGE

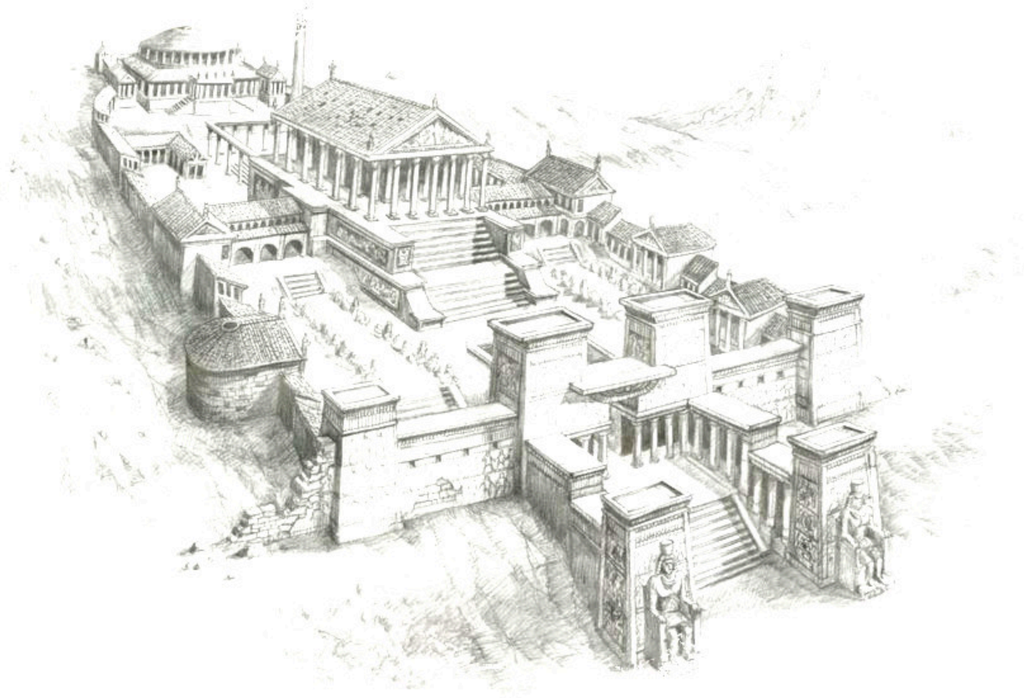
INIZIATIVA CHE SI PROPONE DI CREARE L'ARCHIVIO UNIVERSALE DEL SOFTWARE

Il software rappresenta una componente in rapida crescita della nostra **conoscenza** scientifica, tecnica e culturale

La **conoscenza** è contenuta nel **codice sorgente**

Donald Knuth
considering programs to be *works of literature*

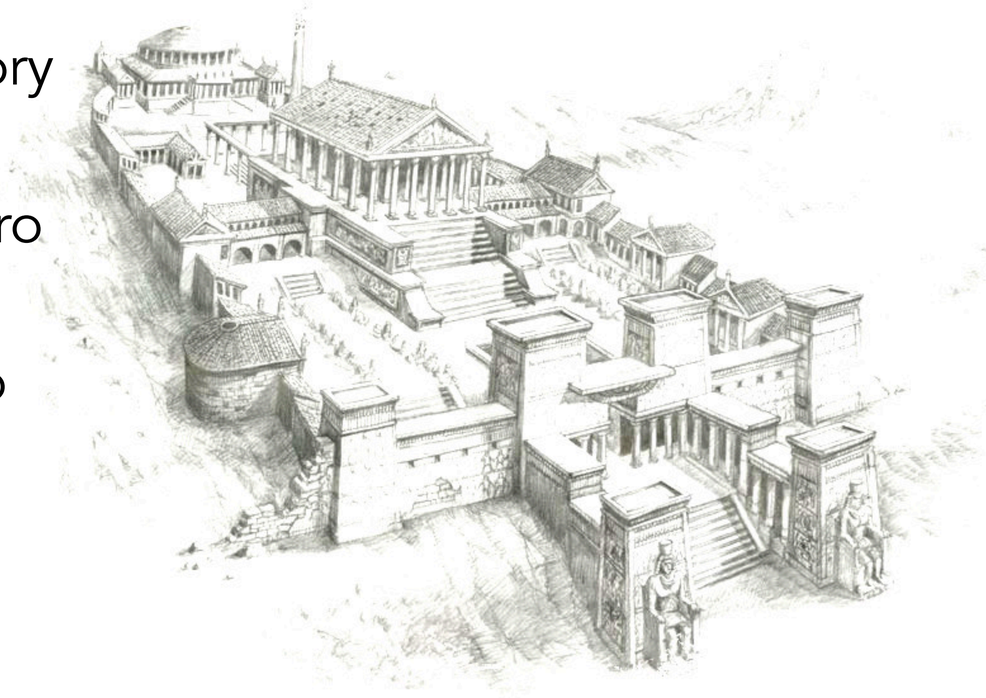
Literate programming. Comput. J., 27(2):97–111, 1984



SOFTWARE HERITAGE

OPERATIVAMENTE SI BASA SU TRE AZIONI PRINCIPALI

- **collect** raccogliere il codice sorgente da diversi repository
- **preserve** memorizzarlo e conservarlo in un archivio sicuro
- **share** rendere accessibile a tutti il software memorizzato



SOFTWARE HERITAGE

- **2015**
  Nasce il progetto su iniziativa di Roberto Di Cosmo e Stefano Zacchiroli (INRIA)

- **2016**
  Presentazione al pubblico con DANS* e Microsoft sponsor

- **2017**
  Accordo INRIA-UNESCO per la conservazione e l'accesso al software

- **2018**
  Apertura dell'archivio all'accesso pubblico

* Dutch National Centre of Expertise and Repository for Research Data

SOFTWARE HERITAGE MIRROR NETWORK

ELEMENTO INDISPENSABILE PER CREARE UNA COMUNITA' CHE ASSICURI:

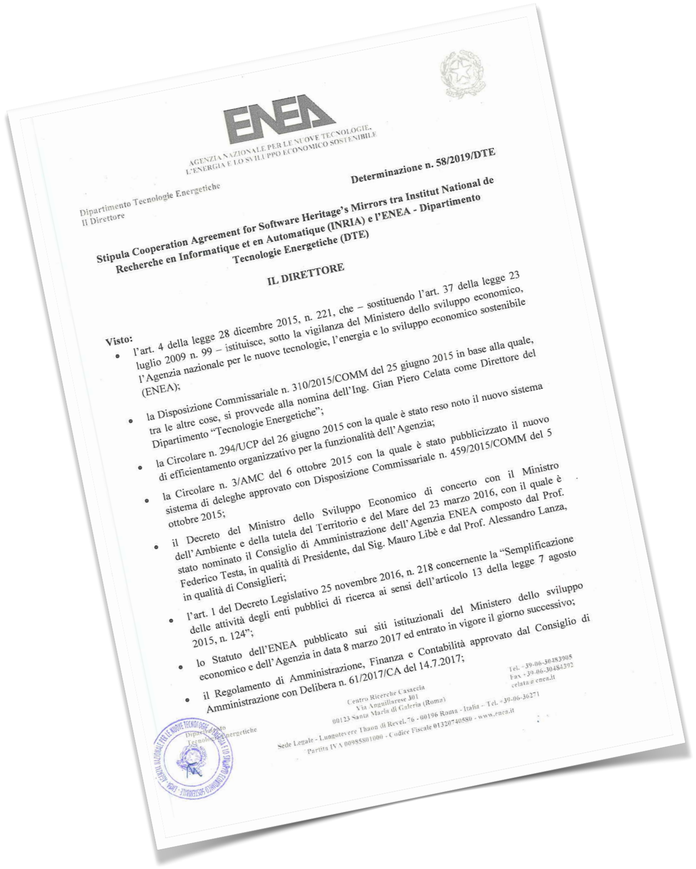


- **sostenibilità** grazie all'aiuto reciproco
- **riduzione del rischio** di perdita dei dati
- **diversificazione** delle soluzioni

SOFTWARE HERITAGE MIRROR NETWORK

2019

Accordo INRIA – ENEA per la creazione di un mirror di SOFTWARE HERITAGE



ENEA

AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE

Dipartimento Tecnologie Energetiche
Il Direttore

Determinazione n. 58/2019/DTE

Stipula Cooperation Agreement for Software Heritage's Mirrors tra Institut National de Recherche en Informatique et en Automatique (INRIA) e l'ENEA - Dipartimento Tecnologie Energetiche (DTE)

IL DIRETTORE

Visto:

- l'art. 4 della legge 28 dicembre 2015, n. 221, che – sostituendo l'art. 37 della legge 23 luglio 2009 n. 99 – istituisce, sotto la vigilanza del Ministero dello sviluppo economico, l'Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile (ENEA);
- la Disposizione Commissariale n. 310/2015/COMM del 25 giugno 2015 in base alla quale, tra le altre cose, si provvede alla nomina dell'Ing. Gian Piero Celata come Direttore del Dipartimento "Tecnologie Energetiche";
- la Circolare n. 294/UCP del 26 giugno 2015 con la quale è stato reso noto il nuovo sistema di efficientamento organizzativo per la funzionalità dell'Agenzia;
- la Circolare n. 3/AMC del 6 ottobre 2015 con la quale è stato pubblicizzato il nuovo sistema di deleghe approvato con Disposizione Commissariale n. 459/2015/COMM del 5 ottobre 2015;
- il Decreto del Ministro dello Sviluppo Economico di concerto con il Ministro dell'Ambiente e della tutela del Territorio e del Mare del 23 marzo 2016, con il quale è stato nominato il Consiglio di Amministrazione dell'Agenzia ENEA composto dal Prof. Federico Testa, in qualità di Presidente, dal Sig. Mauro Libè e dal Prof. Alessandro Lanza, in qualità di Consiglieri;
- l'art. 1 del Decreto Legislativo 25 novembre 2016, n. 218 concernente la "Semplificazione delle attività degli enti pubblici di ricerca ai sensi dell'articolo 13 della legge 7 agosto 2015, n. 124";
- lo Statuto dell'ENEA pubblicato sui siti istituzionali del Ministero dello sviluppo economico e dell'Agenzia in data 8 marzo 2017 ed entrato in vigore il giorno successivo;
- il Regolamento di Amministrazione, Finanza e Contabilità approvato dal Consiglio di Amministrazione con Delibera n. 61/2017/CA del 14.7.2017;

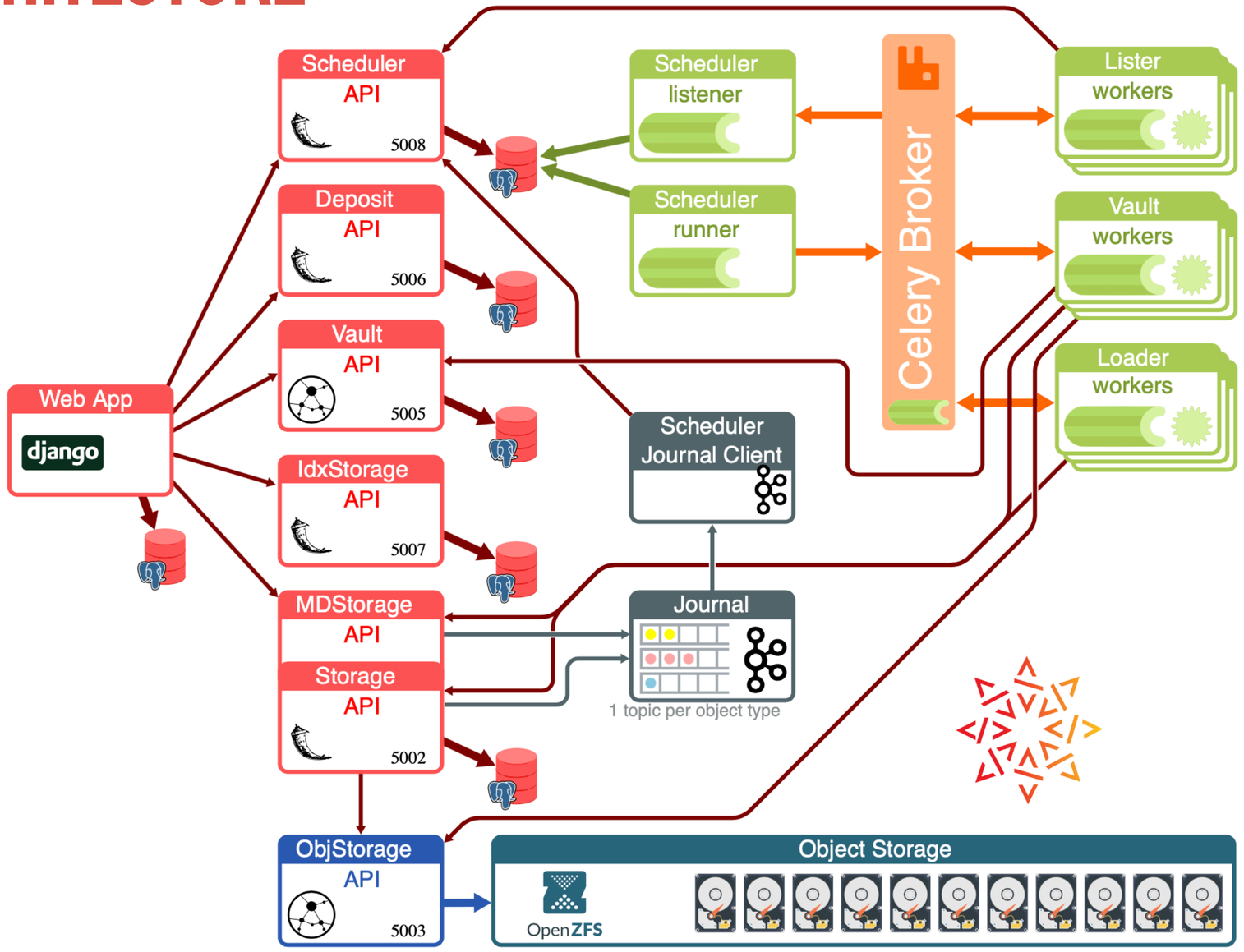
Centro Ricerche Casaccia
Via Anguillarese 301
00123 Santa Maria di Galeria (Roma)
Tel. +39 06 30483968
Fax +39 06 30484397
celata@enea.it

Dipartimento Tecnologie Energetiche
Il Direttore

Sede Legale - Lungotevere Thaon di Revel, 76 - 00196 Roma - Italia - Tel. +39 06 36271
Partita IVA 00985801000 - Codice Fiscale 01320740580 - www.enea.it

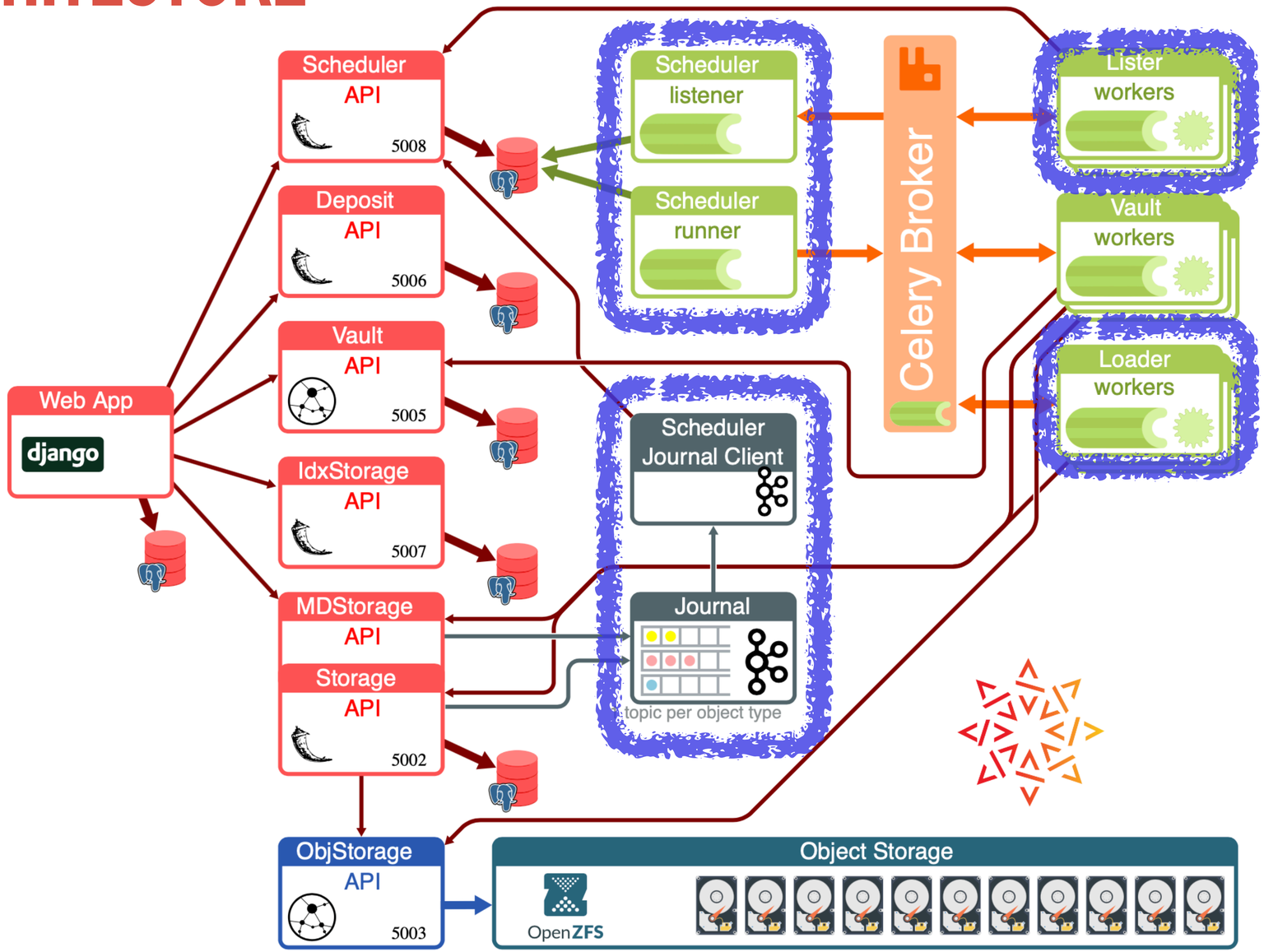
- primo mirror istituzionale
- maggiore resilienza di Software Heritage
- infrastruttura per ricerche in ambito AI
- trampolino di lancio per uno sforzo comune europeo

SOFTWARE HERTITAGE ARCHITECTURE



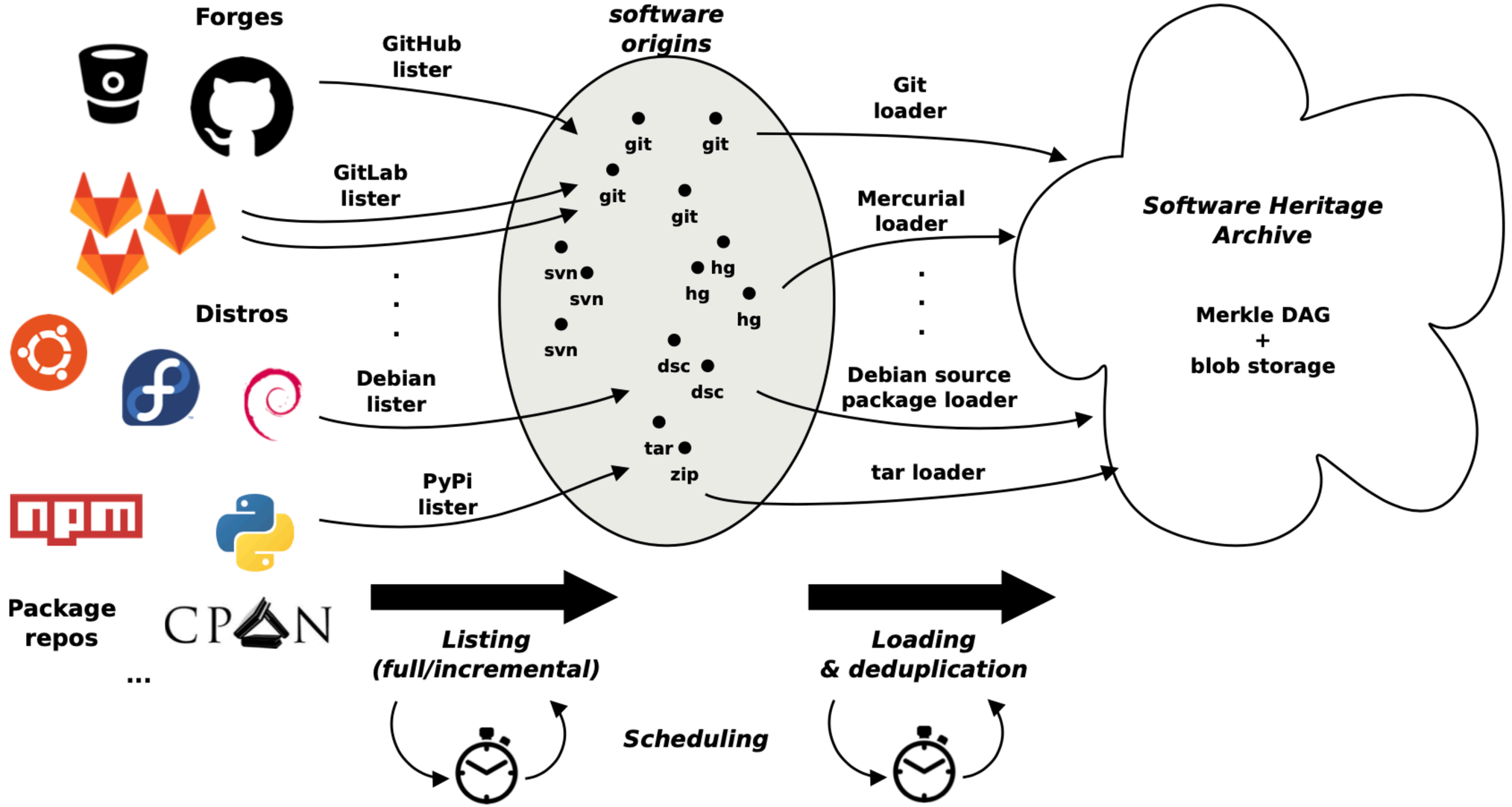
SOFTWARE HERTITAGE ARCHITECTURE

COLLECT



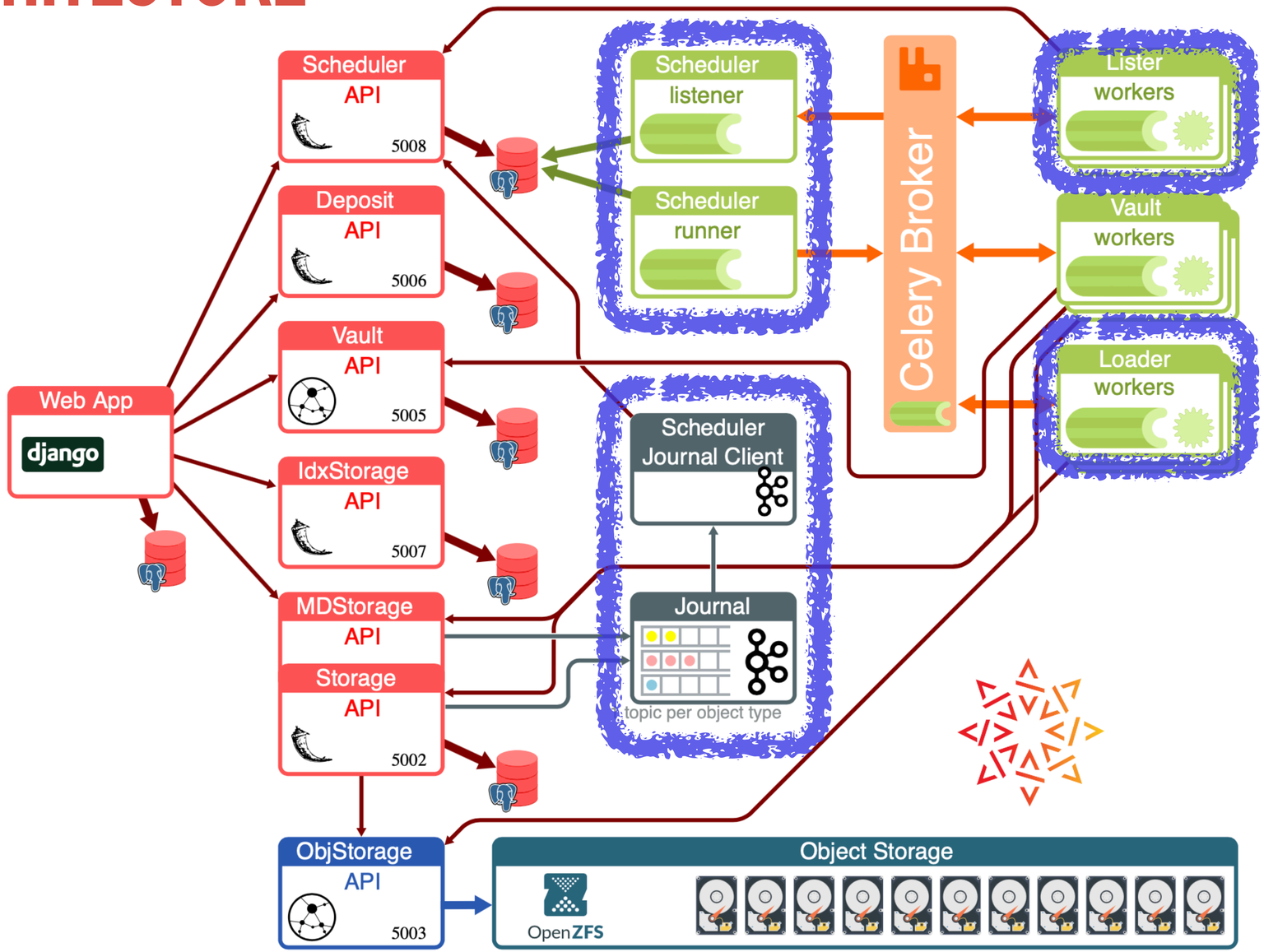
SOFTWARE HERTITAGE ARCHITECTURE

COLLECT



SOFTWARE HERTITAGE ARCHITECTURE

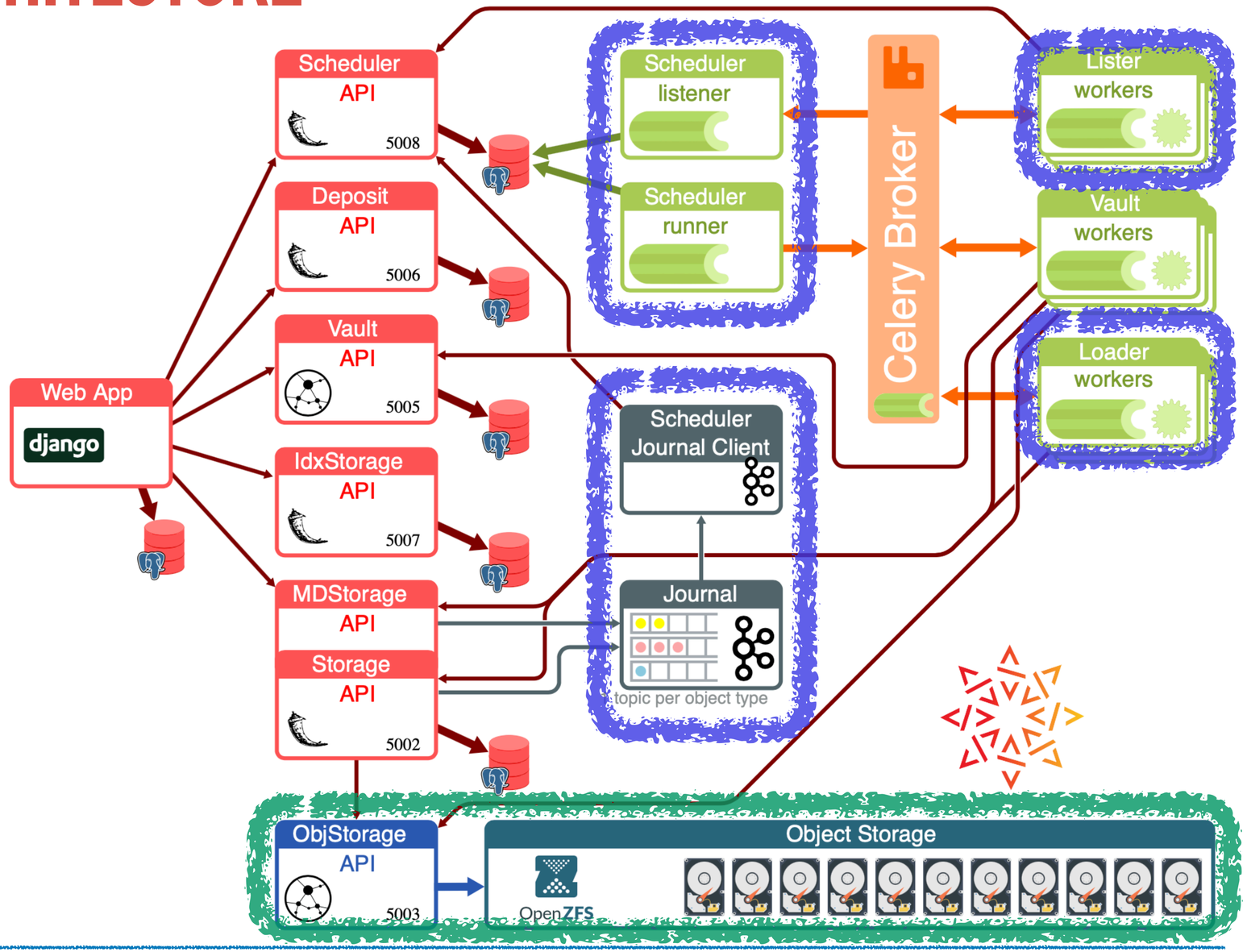
COLLECT



SOFTWARE HERTITAGE ARCHITECTURE

COLLECT

PRESERVE

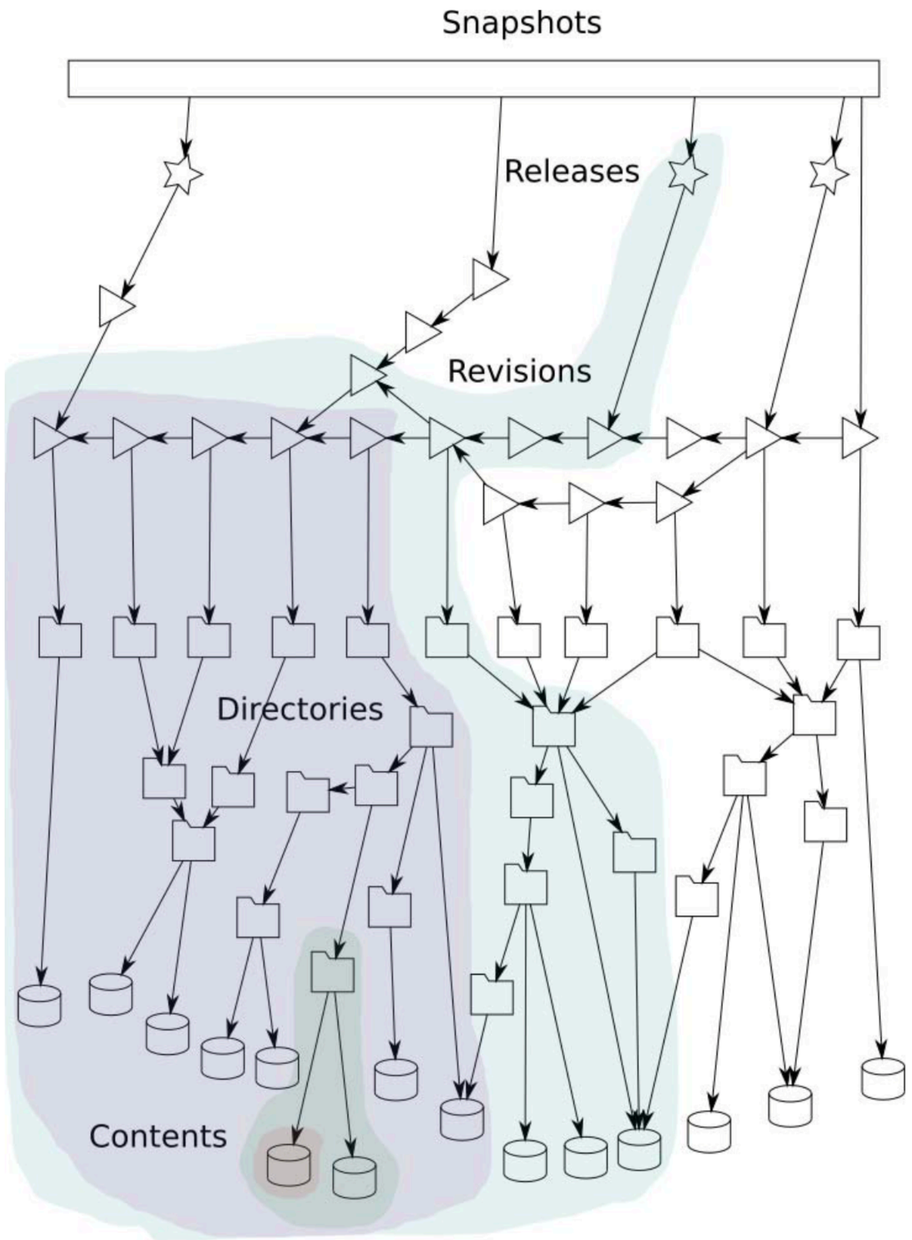


SOFTWARE HERTITAGE ARCHITECTURE

COLLECT

PRESERVE

DATA MODEL: MERKLE DAG



SOFTWARE HERTITAGE ARCHITECTURE

COLLECT

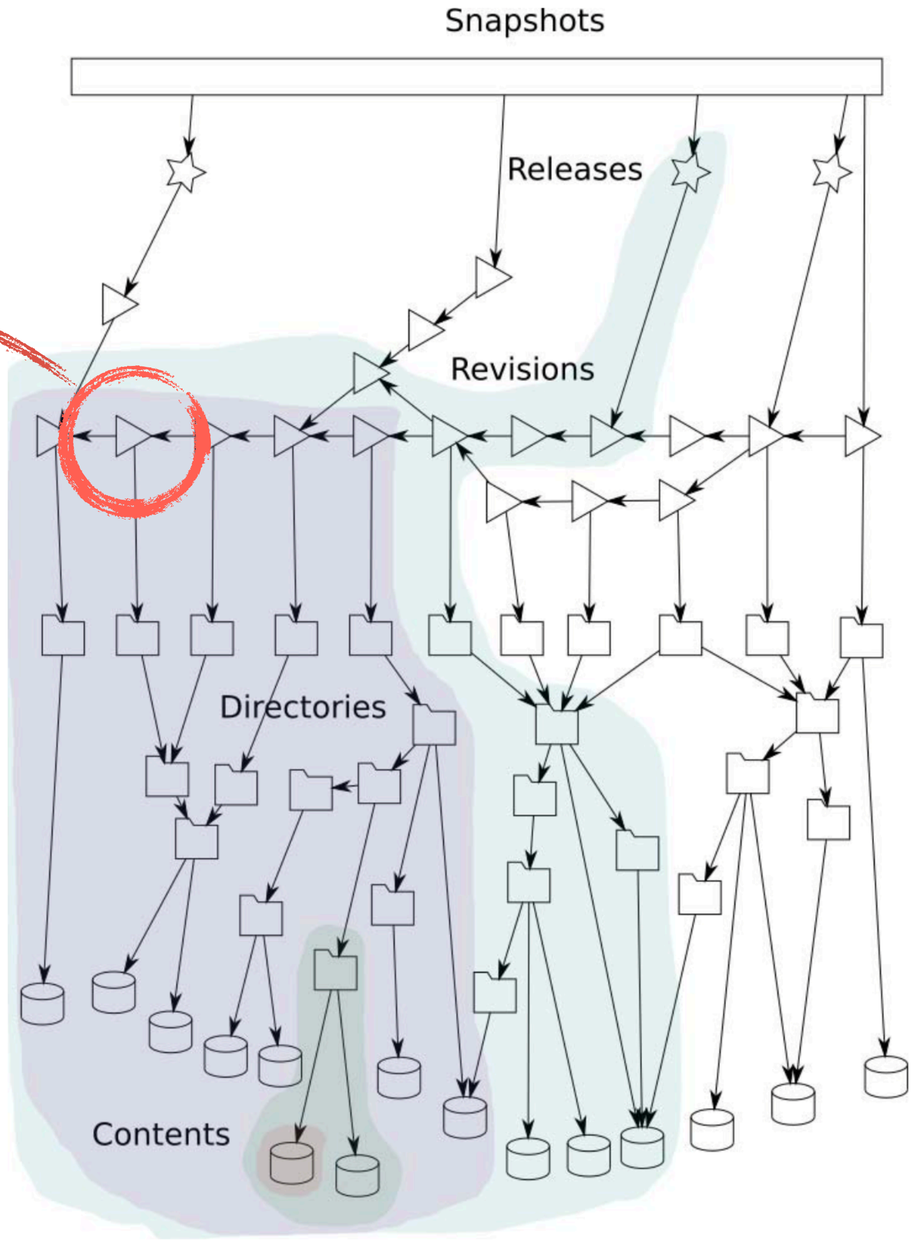
PRESERVE

DATA MODEL: MERKLE DAG

```
directory: fff3cc22cb40f71d26f736c082326e77de0b7692
parent: e4feb05112588741b4764739d6da756c357e1f37
author: Stefano Zacchiroli <zack@upsilon.cc>
date: 1443617461 +0200
committer: Stefano Zacchiroli <zack@upsilon.cc>
commiter_date: 1443617461 +0200
message:
  objstorage: fix tempfile race when adding objects

  Before this change, two workers adding the same
  object will end up racing to write <SHA1>.tmp.
  [...]
```

revision_id: 64a783216c1ec69dcb267449c0bbf5e54f7c4d6d



SOFTWARE HERTITAGE ARCHITECTURE

COLLECT

PRESERVE

DATA MODEL: MERKLE DAG

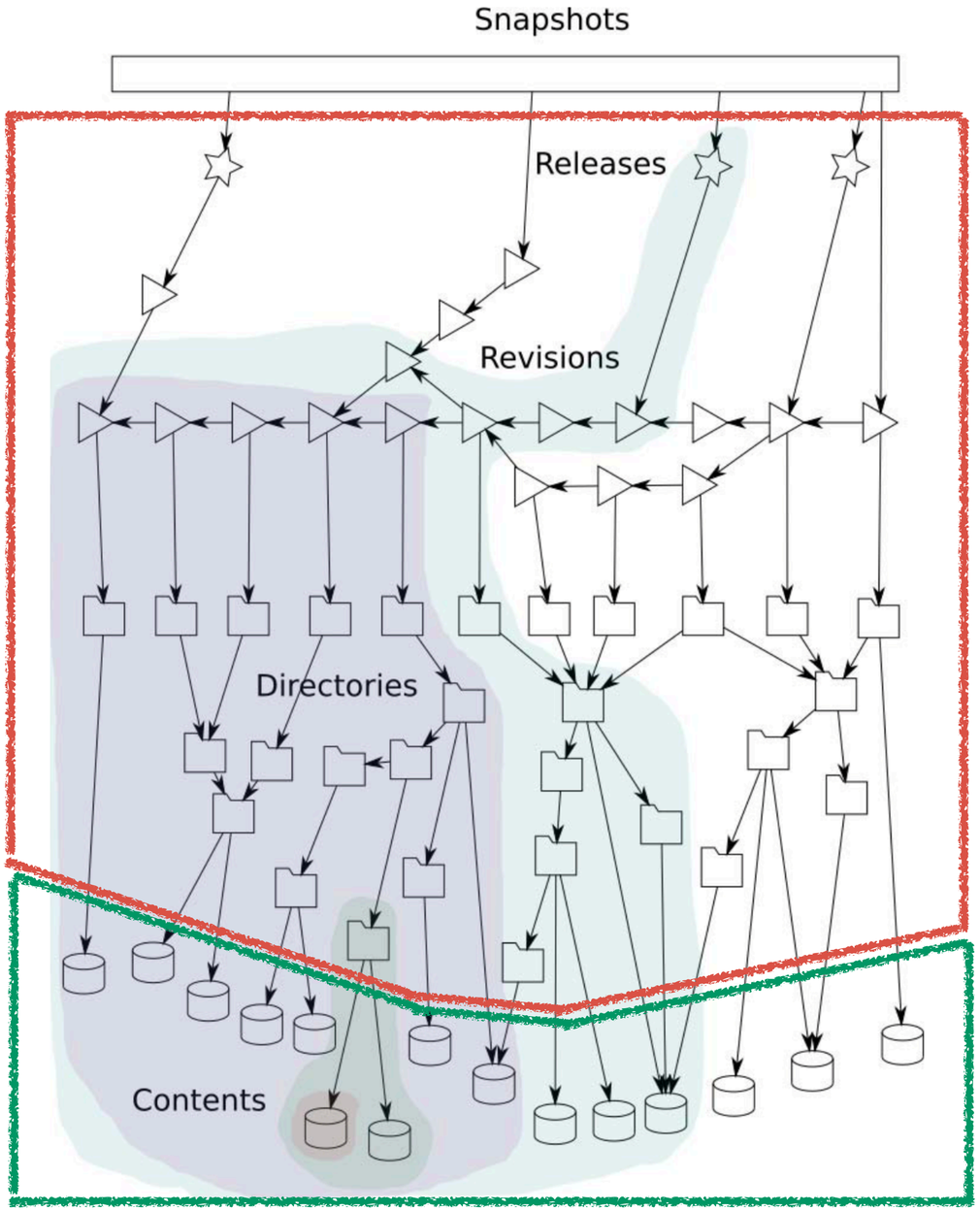


NODES



RAW DATA

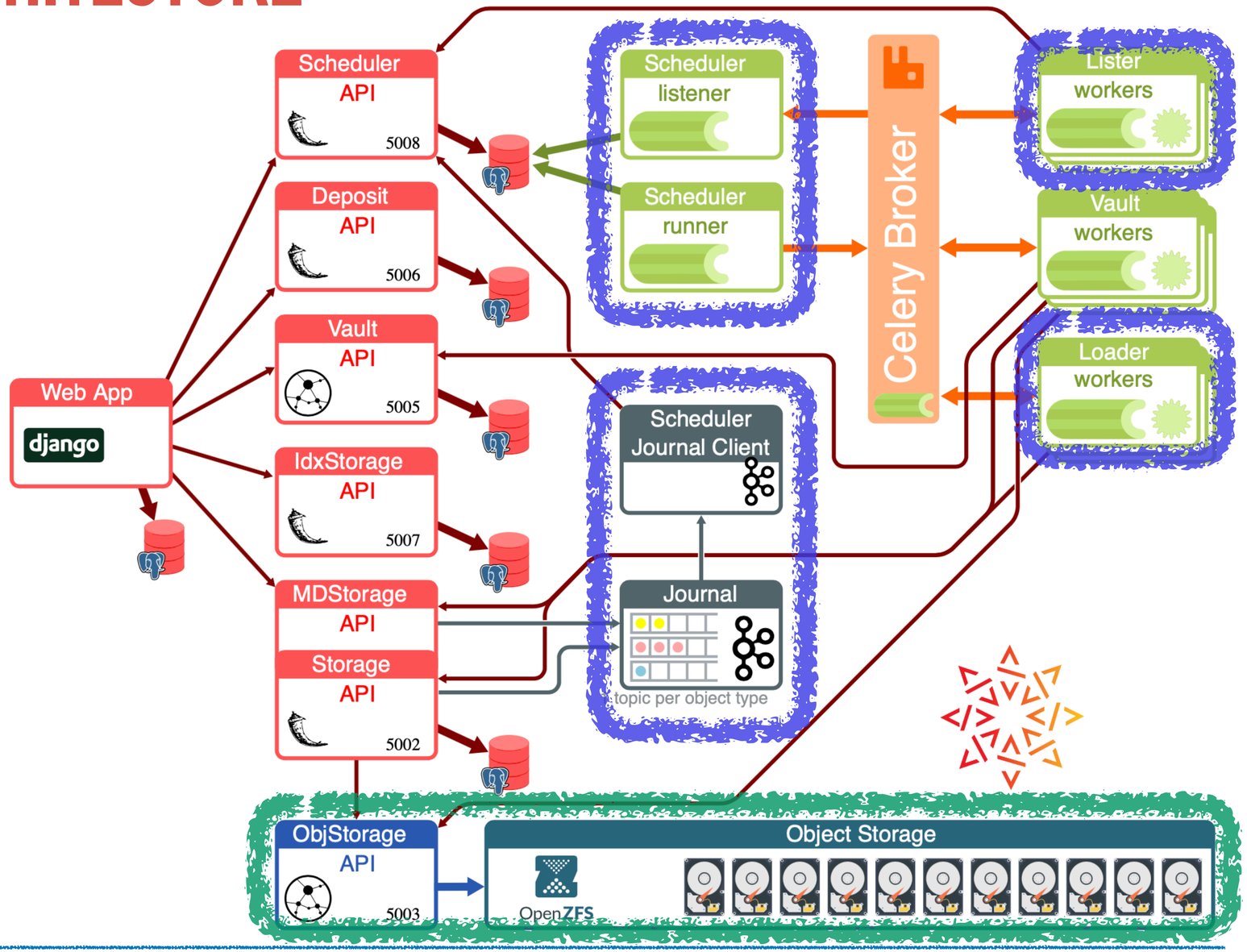
blobs



SOFTWARE HERTITAGE ARCHITECTURE

COLLECT

PRESERVE

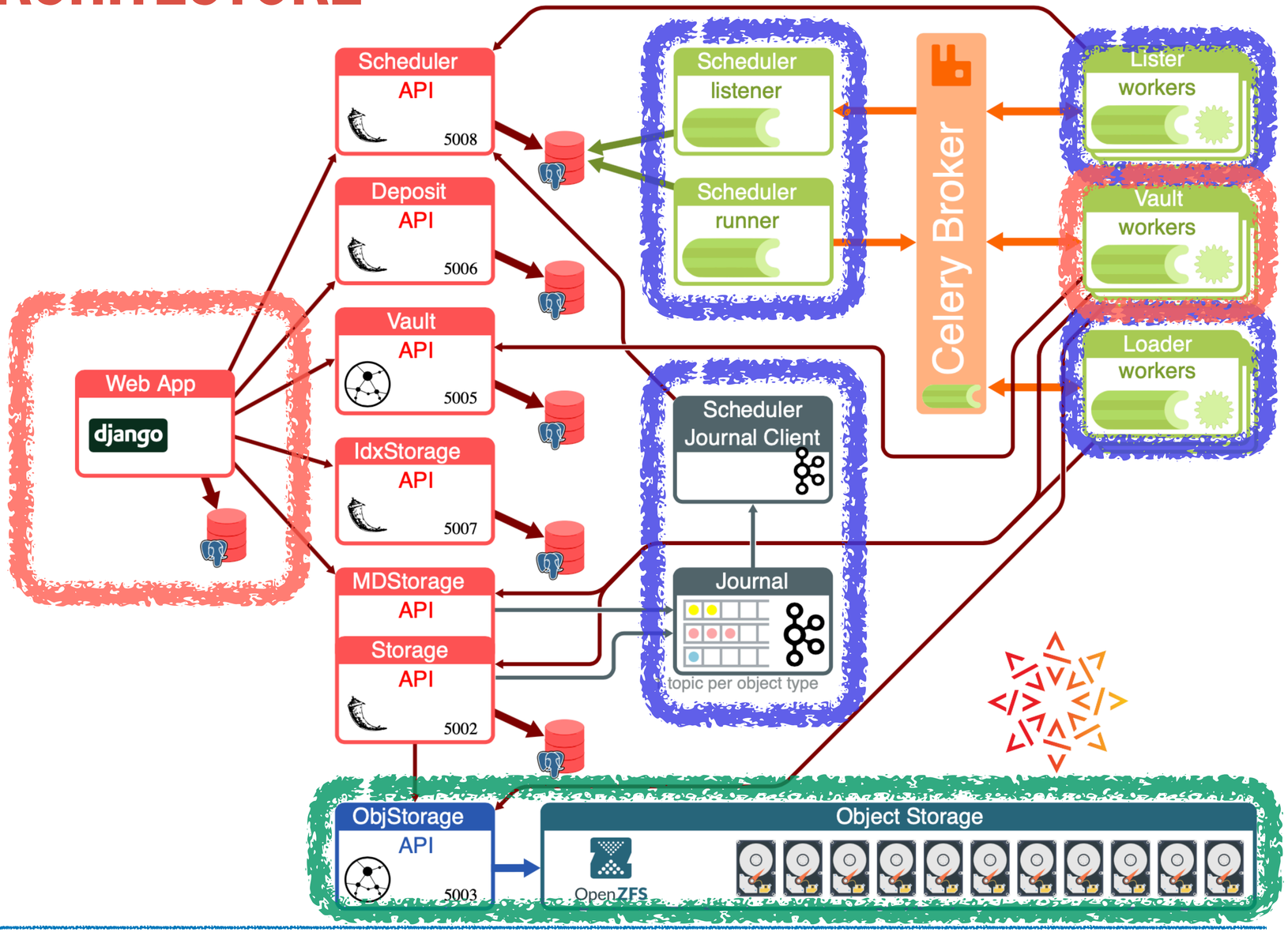


SOFTWARE HERTITAGE ARCHITECTURE

COLLECT

PRESERVE

SHARE

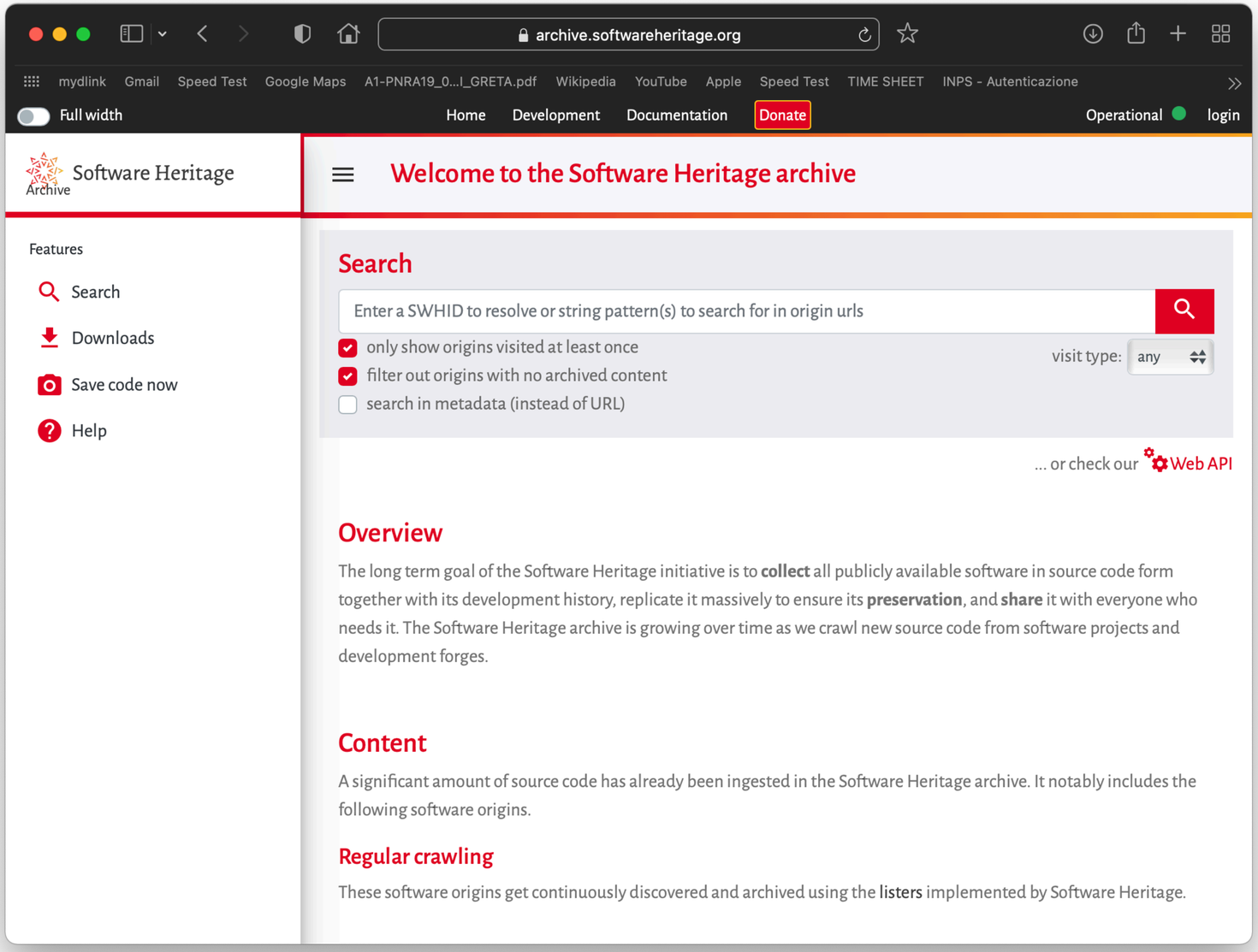


SOFTWARE HERTITAGE ARCHITECTURE

COLLECT

PRESERVE

SHARE



SOFTWARE HERTITAGE TODAY

REGULAR CROWLING

Bitbucket
2,036,678 origins

git
5,414 origins

R
18,823 origins

debian
35,972 origins

5,866 origins

GitHub
125,134,487 origins

GitLab
985,086 origins

Guix
10948 origins

GNU
354 origins

heptapod
983 origins

launchpad
20,398 origins

NixOS
30688 origins

npm
1,725,286 origins

opam
3,516 origins

python Package Index
409,270 origins

SOURCEFORGE
341,756 origins

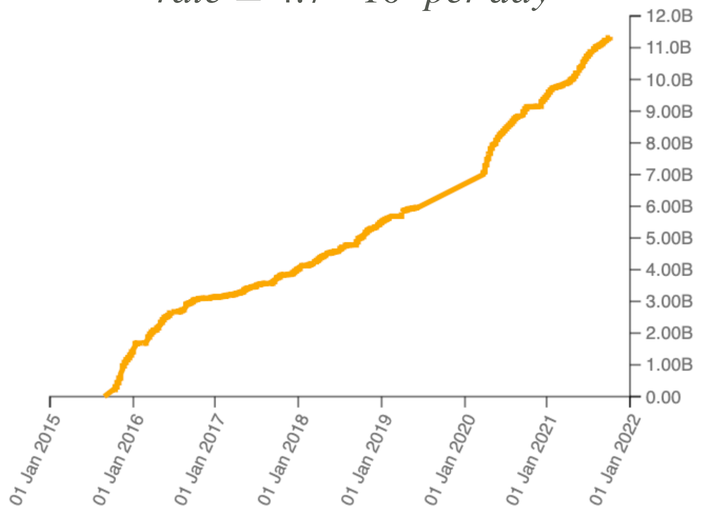
SOFTWARE HERTITAGE TODAY

SIZE

Source files

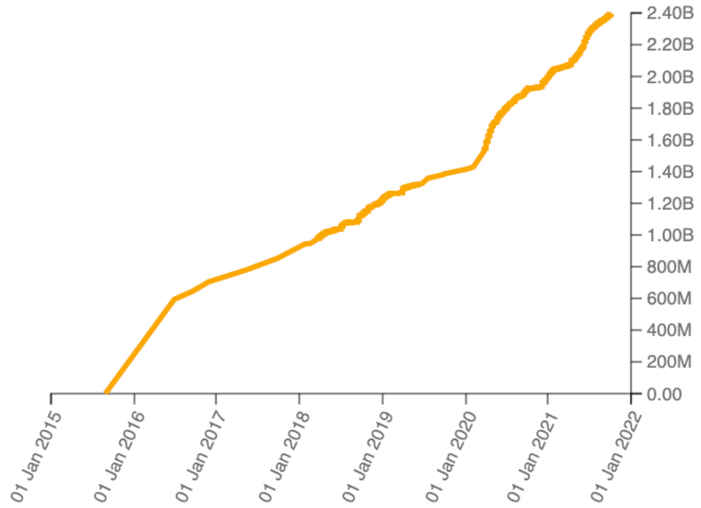
11.345.609.469

rate $\backsimeq 4.7 * 10^6$ *per day*



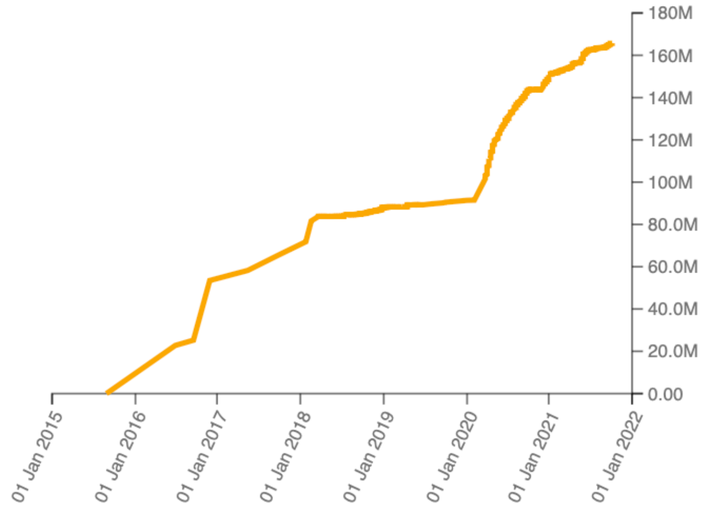
Commits

2.396.353.337



Projects

165.746.040

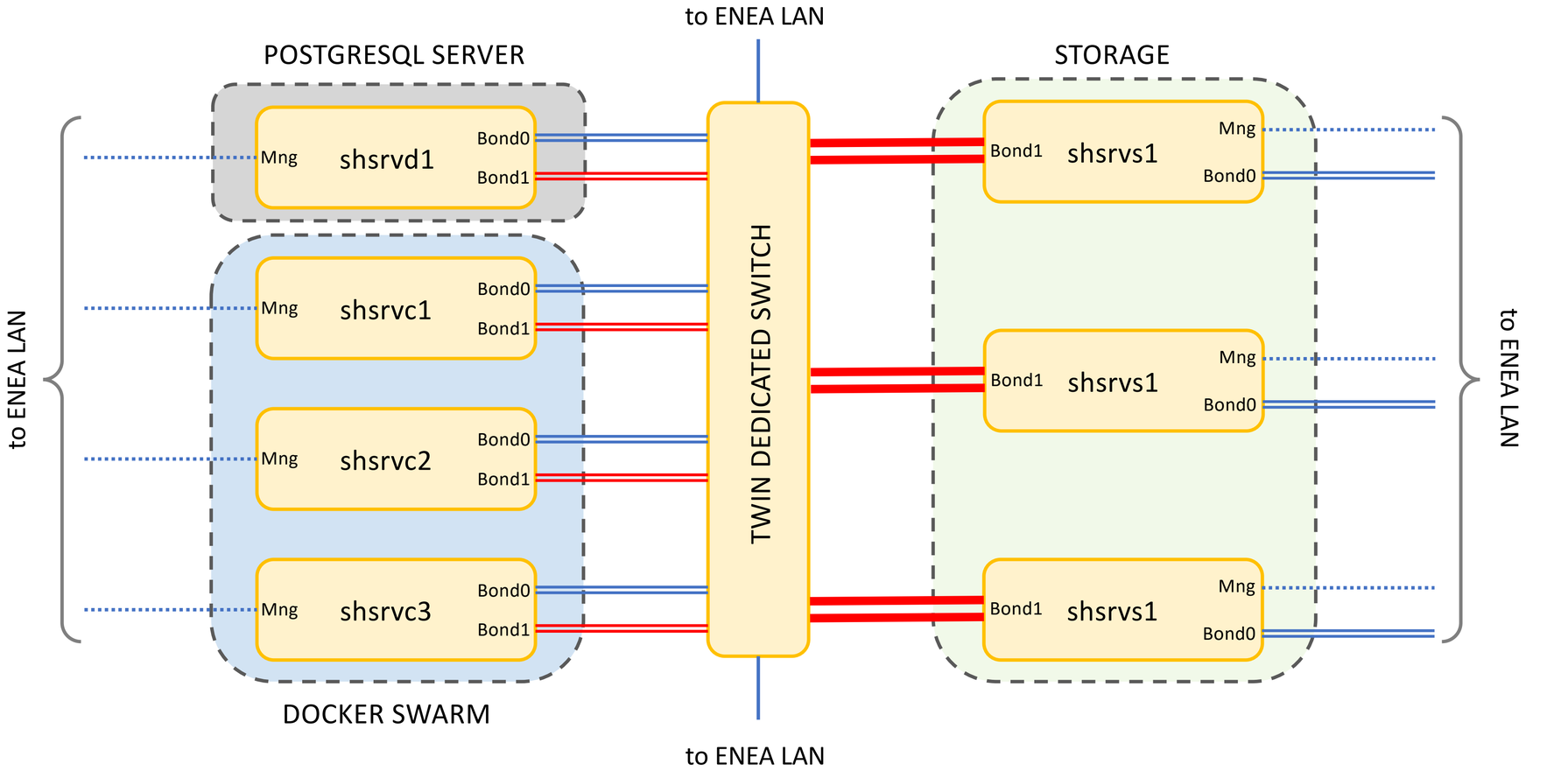


COMPONENTI HW MIRROR ENEA

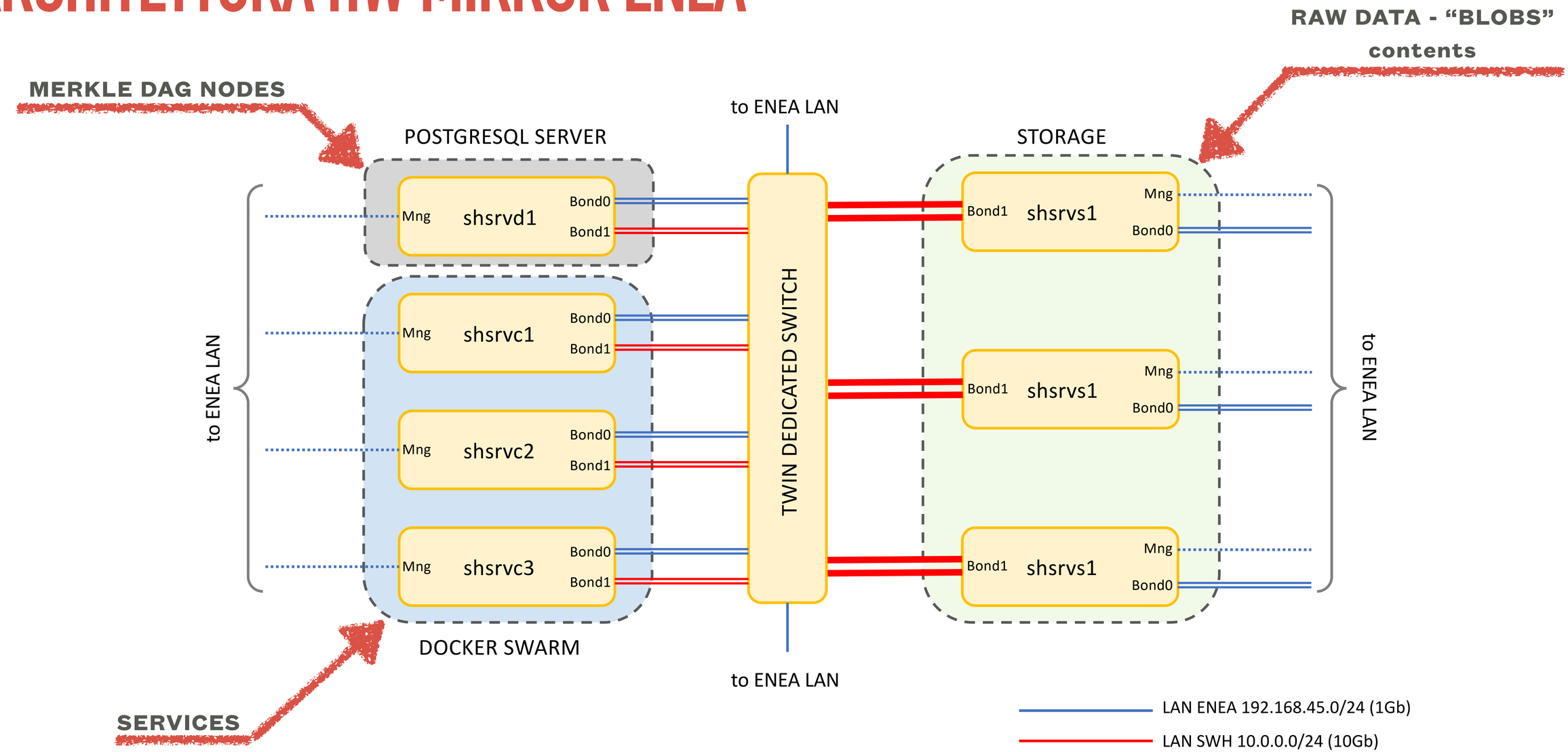
Server	Model	RAM (GB)	HD (TB)	SO
shsrvd1	Dell PowerEdge R640	256	5.6 (SSD)	Ubuntu 20.04 LTS
shsrvc1	Supermicro SYS-6028TR-HTR	64	1.8	Ubuntu 20.04 LTS
shsrvc2	Supermicro SYS-6028TR-HTR	64	0.9	Ubuntu 20.04 LTS
shsrvc3	Supermicro SYS-6028TR-HTR	64	0.9	Ubuntu 20.04 LTS
shsrvs1	Supermicro SSG-6049P-E1CR24H	92	218	Ubuntu 20.04 LTS
shsrvs2	Supermicro SSG-6049P-E1CR24H	92	218	Ubuntu 20.04 LTS
shsrvs3	Supermicro SSG-6049P-E1CR24H	92	218	Ubuntu 20.04 LTS



ARCHITETTURA HW MIRROR ENEA

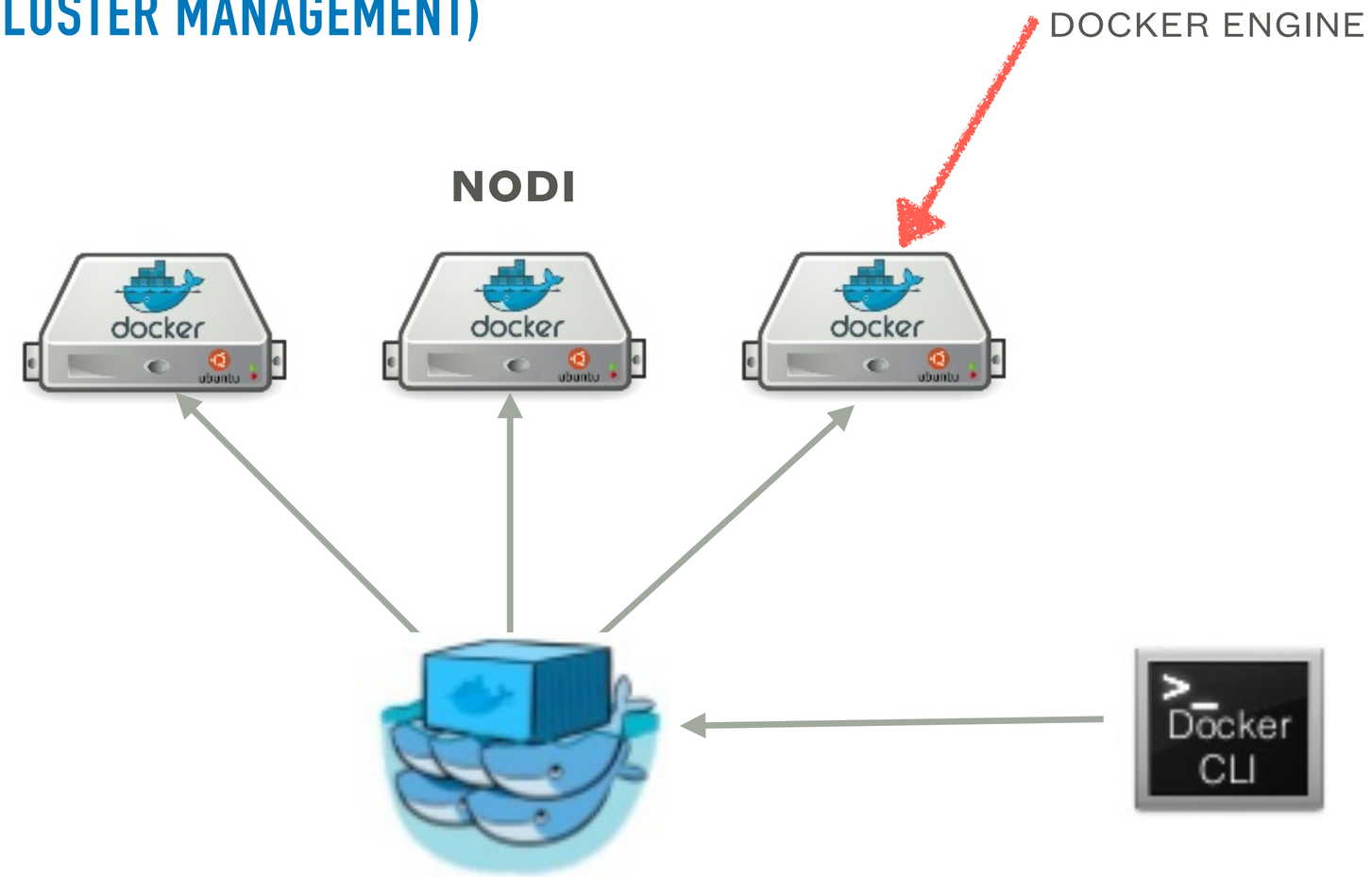


ARCHITETTURA HW MIRROR ENEA



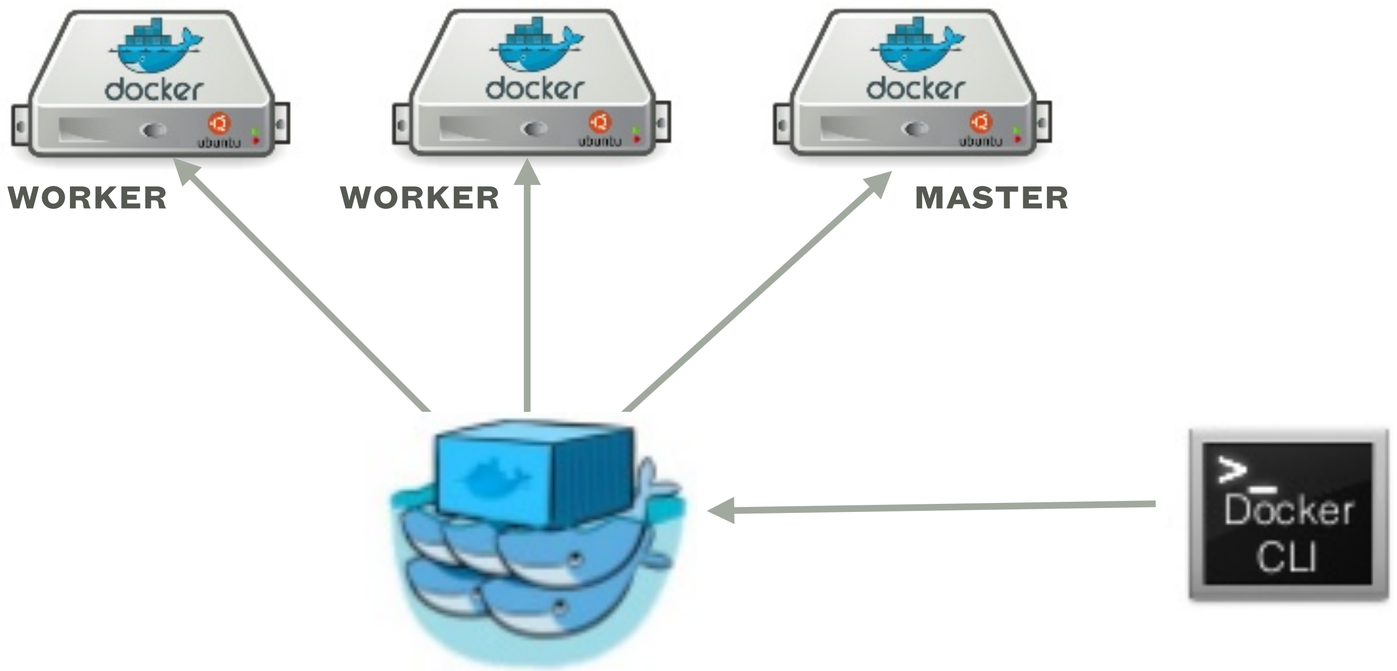
DOCKER SWARM

TOOL DI ORCHESTRAZIONE (CLUSTER MANAGEMENT)



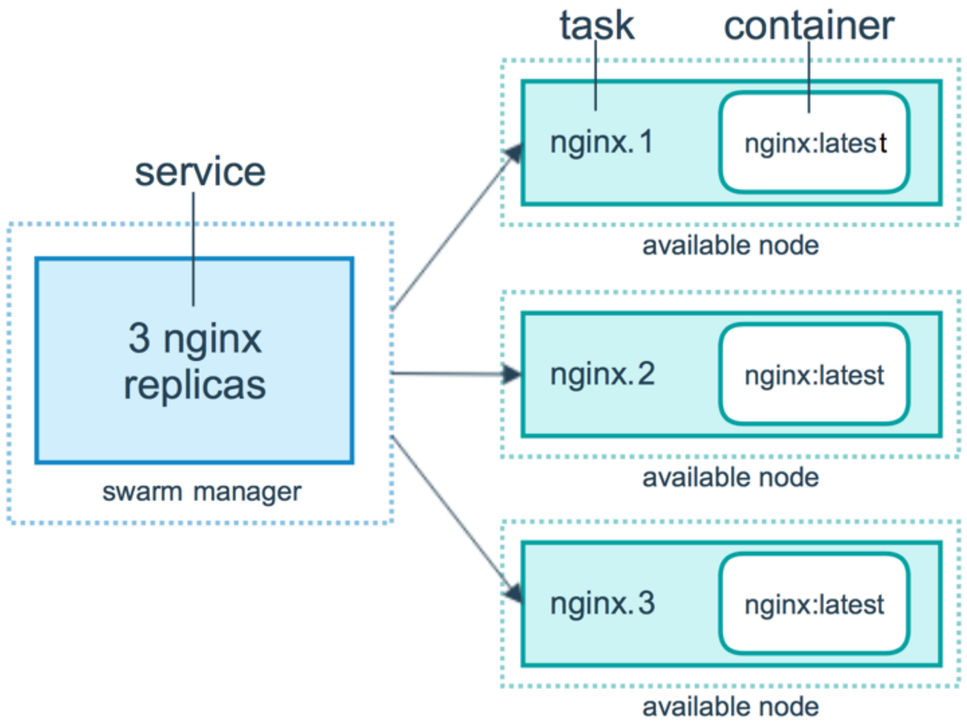
DOCKER SWARM

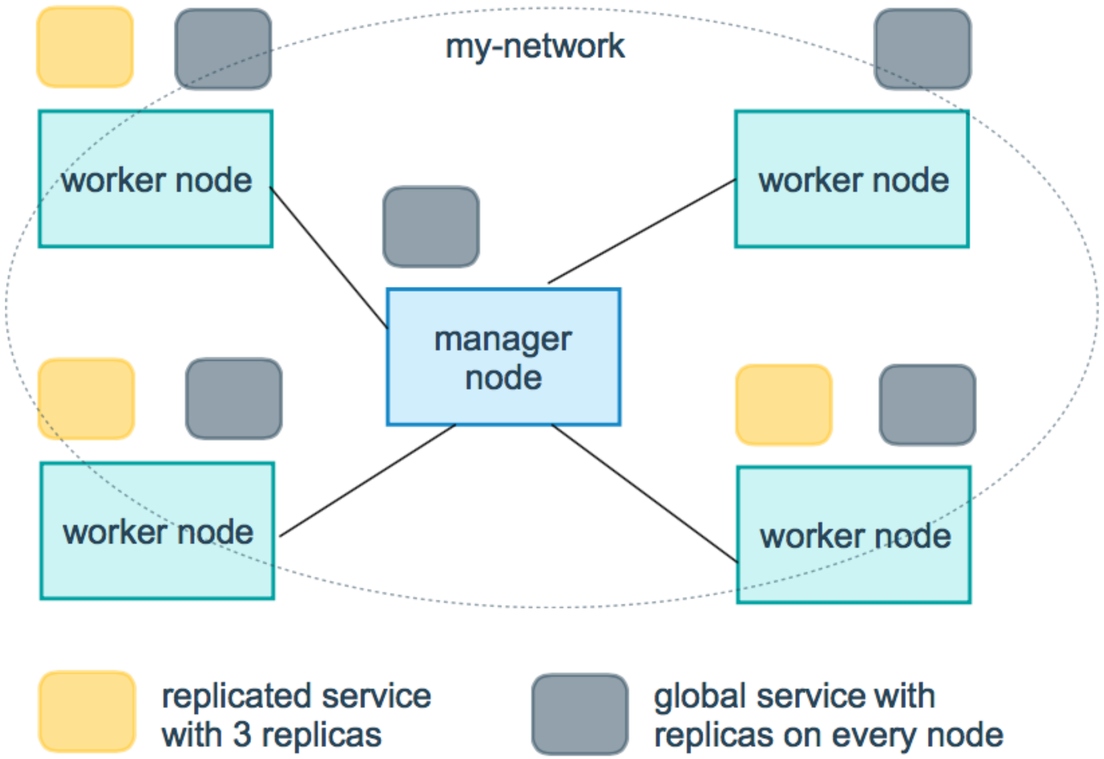
TOOL DI ORCHESTRAZIONE (CLUSTER MANAGEMENT)



DOCKER SWARM

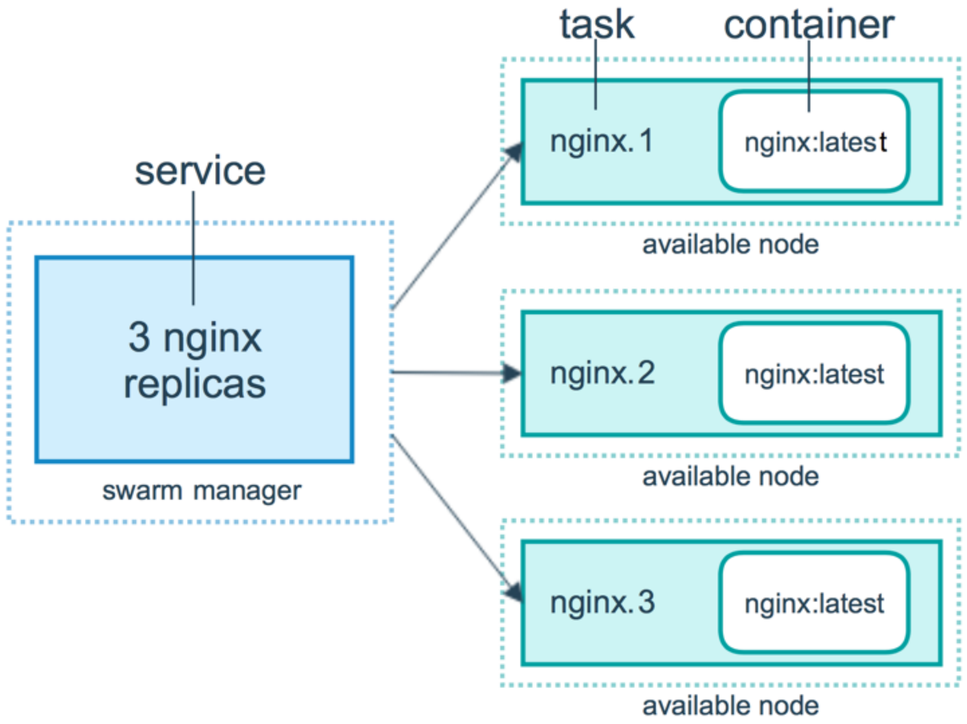
TOOL DI ORCHESTRAZIONE (CLUSTER MANAGEMENT)

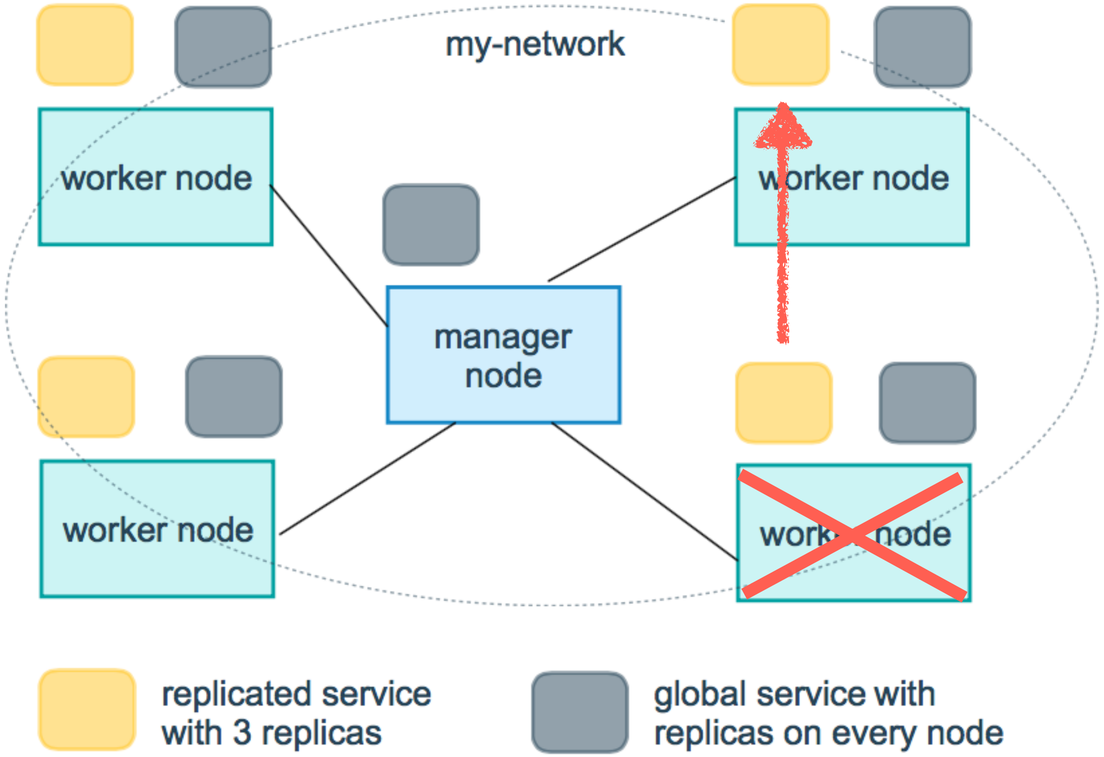




DOCKER SWARM

TOOL DI ORCHESTRAZIONE (CLUSTER MANAGEMENT)





STORAGE

SEAWEEDFS file system distribuito, altamente scalabile

- **scalabilità**
  la capacità si aumentata facilmente aggiungendo volumi

- **ottimizzato per enormi quantità di piccoli file**
  overhead per ciascun file 40 bytes

- **accesso veloce**
  i client accedono direttamente ai volumi
  lettura di un file con un solo disk seek
  metadati su server dedicato



SEAWEEDFS

master server

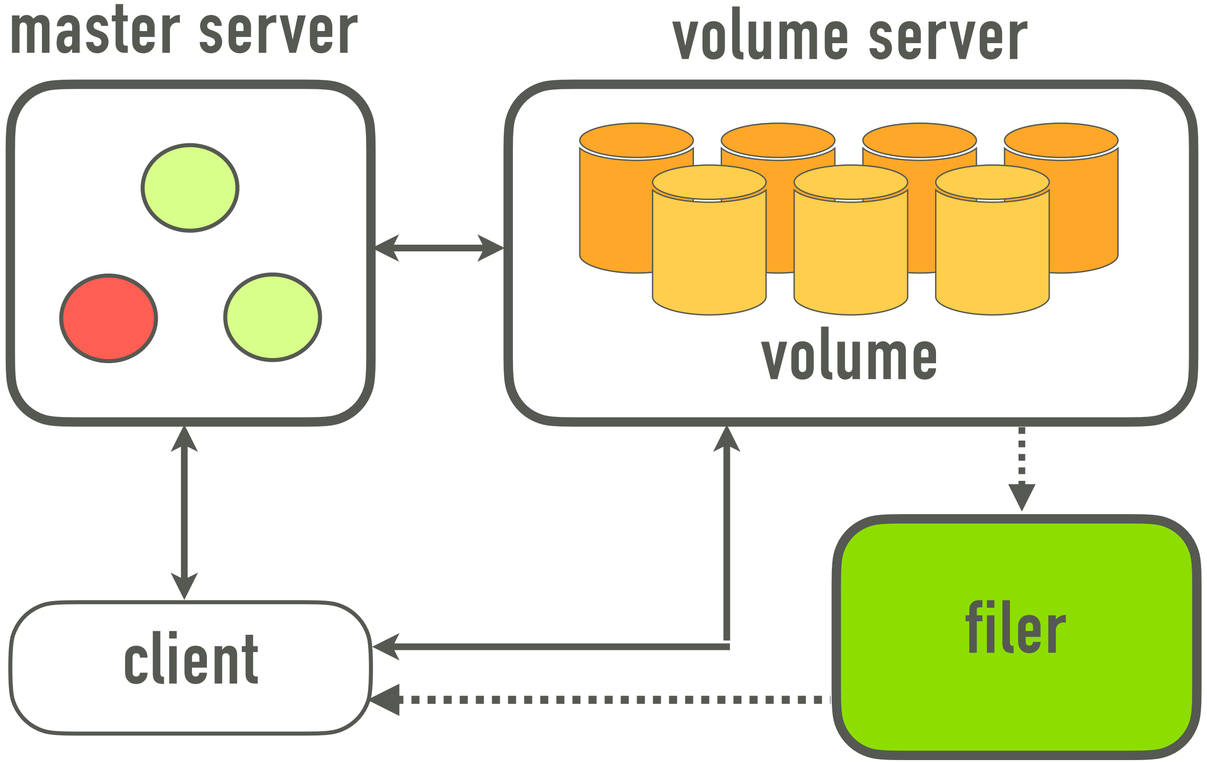
server cluster
tiene traccia dei volumi nel volume server

volume server

contiene svariati volumi logici (30 GB)
lo spazio disco è preallocato
la scrittura è eseguita in append

volume

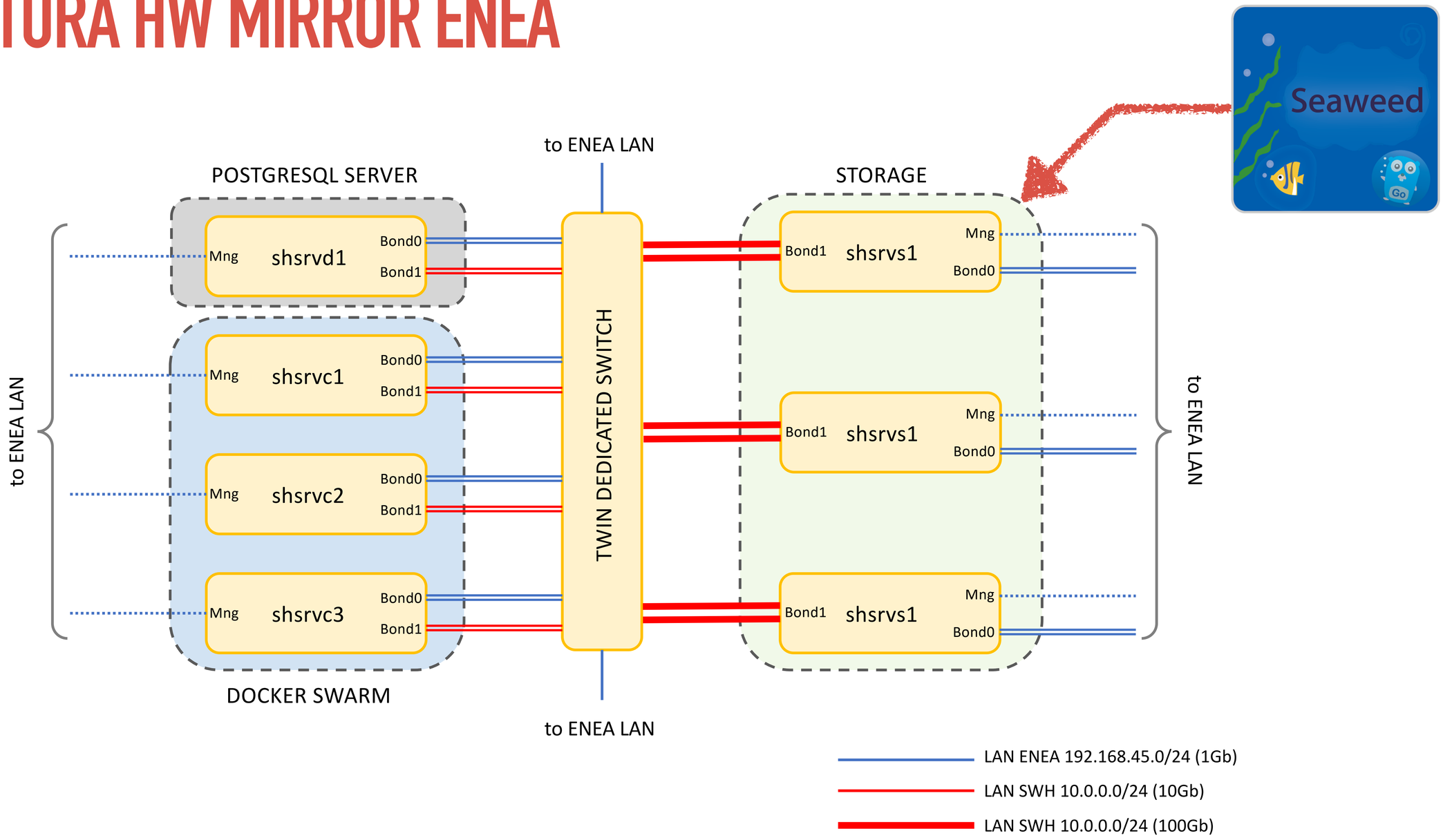
contiene un elevato numero di file
i file sono indicizzati con offset
lettura di un file con un solo seek



filer (opzionale)

nome dei file e directory
struttura memorizzata separatamente (SQL, NoSQL, LevelDB)

ARCHITETTURA HW MIRROR ENEA



SEAWEEDFS

SCRITTURA FILE

1. il client richiede al master server di allocare **volume** e **id** per il file **/home/user/myfile.txt**
   curl http://192.168.88.129:9333/dir/assign

2. la **risposta** del server
   {"fid":"**14,0c3657515f**","url":"192.168.88.129:9222","publicUrl":"192.168.88.129:9222","count":1}

3. il client scrive **direttamente** il file nel volume assegnato
   curl -F file=@/home/user/myfile.txt 192.168.88.129:9222/**14,0c3657515f**

LETTURA FILE

il client richiede **direttamente** il file
wget 192.168.88.129:9222/**14,0c3657515f**

SEAWEEDFS

FILER & MOUNT POINT

se installato, il filer consente di montare i volumi
e accedere dal file system del server alla struttura delle directories

```
weed mount -filer=192.168.88.129:8888 -dir=/seaweedfs/data/mount &
ls /seaweedfs/data/mount
```

MIRROR TODAY

1. completata l'installazione dell'hardware
2. installati e configurati SO, Docker Swarm, SeaWeedFS
3. creato il playbook di Ansible per l'upgrade automatico dei SO dei server
4. installato il SW di Software Heritage del mirror
5. iniziata la migrazione del DB Postgesql con il grafo di Merkel

TO DO

1. migrazione dei file con i codici sorgente (blobs)
2. benchmark prestazioni di SeaWeedFS

GRAZIE

