Forum TERATEC 23

Workshop A6

June 1st, 2023

Trends in supercomputing architectures in the exascale era: heterogeneity, modularity, disaggregation

Architecture, management and administration of large supercomputing centres and of their software stack CEA approach on the path to Exascale

Gilles Wiber

Jean-Philippe Nominé

CEA Département des Sciences de la Simulation et de l’Information

Outline

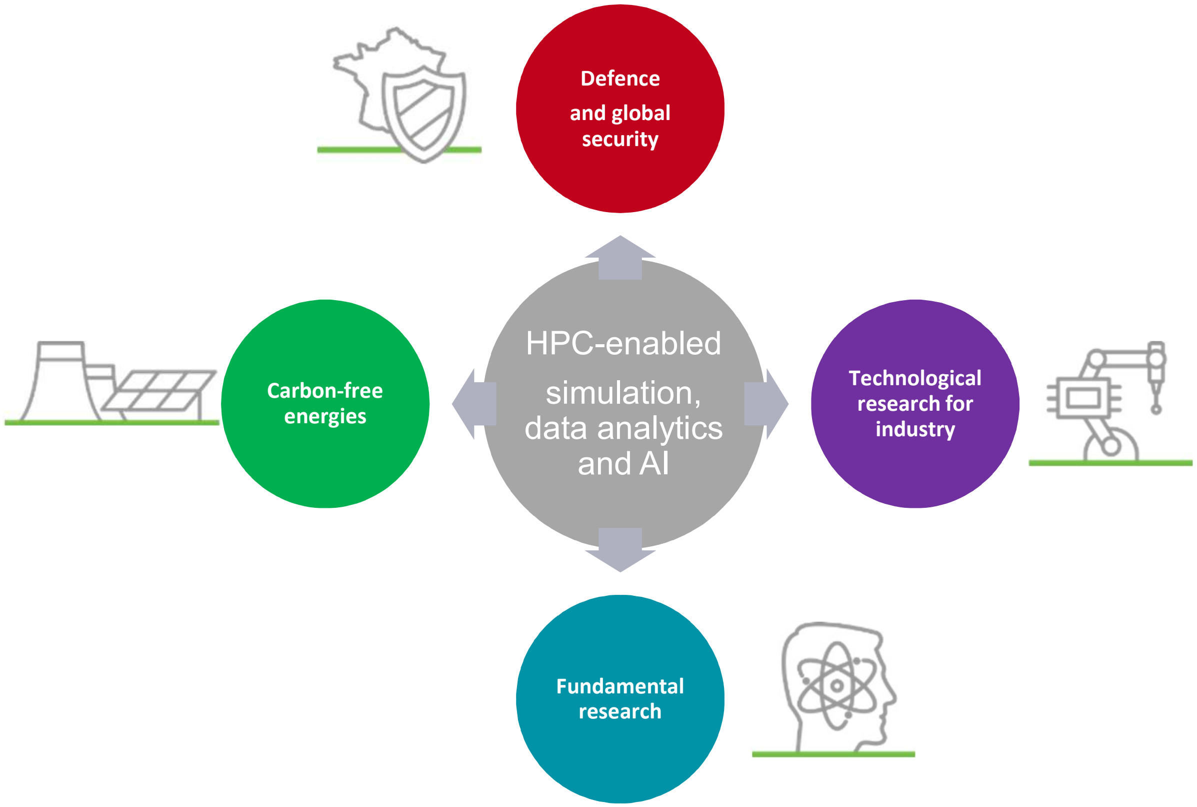
1. CEA supercomputing complex
2. Typical computing centre architecture
3. Data and storage considerations
4. Configuration management and software system stack deployment - OCEAN
5. User services evolution - virtualisation
6. Towards Exascale

This talk is mostly about software => user services

1. **CEA supercomputing complex**
2. Typical computing centre architecture
3. Data and storage considerations
4. Configuration management and software system stack deployment - OCEAN
5. User services evolution
6. Towards Exascale

1. CEA supercomputing complex

CEA experience and expertise in large computer centre design and operations



Defence and global security

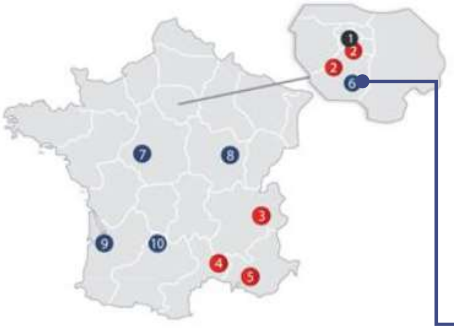
HPC-enabled simulation, data analytics and AI

Carbon-free energies

Technological research for industry

Fundamental research

www.cea.fr
www-hpc.cea.fr/index-en.htm



CEA Supercomputing Complex

Bruyères-le-Châtel
Essonne

TERA/EXA, TGCC facilities

CEA Bruyères Le Châtel
Bull
TERATEC
CEA TGCC
Chem. de la Croix Blanche
Chem. des Petits Bois
Chem. du Petit Ru
Commissariat Energie Atomique
TERA-EXA
itakraft Simon Louis
D152
D152

World-class HPC centres - beyond CEA own needs

As of today: **1 site – 2 facilities – 4 multi-petascale supercomputing centres**

From research to industry, for research and industry

One site: CEA/DIF Bruyères-le-Châtel

One team = HPC division: designing facilities and their infrastructures, co-designing supercomputers, operating them and delivering related services

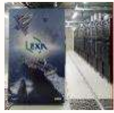
Two facilities

TERA/EXA

TGCC

Four multi-petascale supercomputing systems

TERA+EXA

Defence


- TERA 1000-2 23 Pflops
  2017, Intel KNL+BXI v1.2
  Top500 #42
- EXA-HF 36 Pflops
  2021, BullSequana XH2000, AMD Milan, BXI V2
  Top500 #14

TERA+EXA

Industry (confidential applications)

- CCMD-A 2 Pflops
  2022, BullSequana XH2000, AMD Milan, BXI V2

TGCC

Research

- JOLIOT-CURIE Rome, 12 Pflops
  2019, BullSequana XH2000, AMD Rome, IB HDR
  Top500 #69
- Joliot-Curie SKL, 6,6 PF
  2017, Sequana X1000, Intel SKL, IB EDR
  Top500 #113

TGCC

Industry (20 partners)

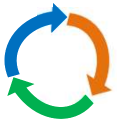
- Topaze-cpu, 4.3 Pflops
  2021, BullSequana XH2000, AMD Milan, IB HDR
  Top500 #140
- Topaze-gpu, 3.7 Pflops
  2021, BullSequana XH2000, NV A100, IB HDR
  Top500 #198
- Since 2018: a QLM 30

+ Experimental centre (incl. 'INTI')

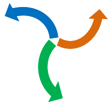
CEA experience and expertise in large computer centre design and operations

- Advance computing at CEA since 1955… (now called HPC/Supercomputing)
- … more recently, 20+ years of experience (co)designing and operating ‘massively parallel era’ large HPC and big data oriented systems and facilities

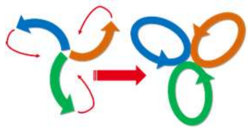
- Dealing with, serving and supporting different user communities
  - Defence, Research (FR+EU), Industry (CCRT)

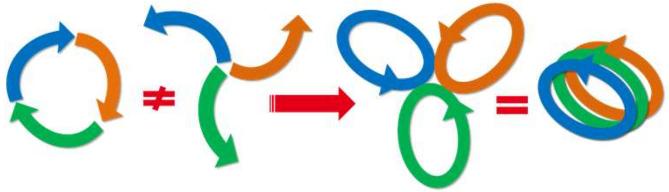


  - Gives a 360° vision of a very wide range of needs, helps steer R&D that matches market needs
  - … but leads to managing several different, although somehow similar, computing centres



- A strong involvement in open source and community developments
- Pooling methods and efforts to operate different but somehow similar computing centres





1. CEA supercomputing complex
2. **Typical computing centre architecture**
3. Data and storage considerations
4. Configuration management and software system stack deployment - OCEAN
5. User services evolution
6. Towards Exascale

2. Typical computing centre architecture

Computing centres services

A computing center has different kinds of capabilities and provides different resources/services – major challenges = flexibility & adaptability

Resources

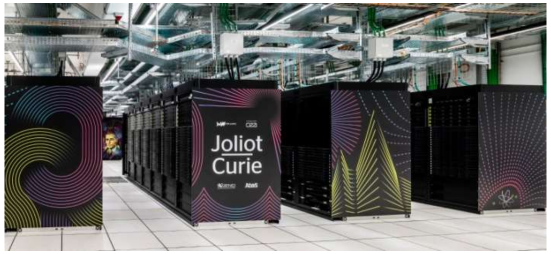
(hard skills)

Compute

I/O

Applications

Exploitation



Security

Capabilities

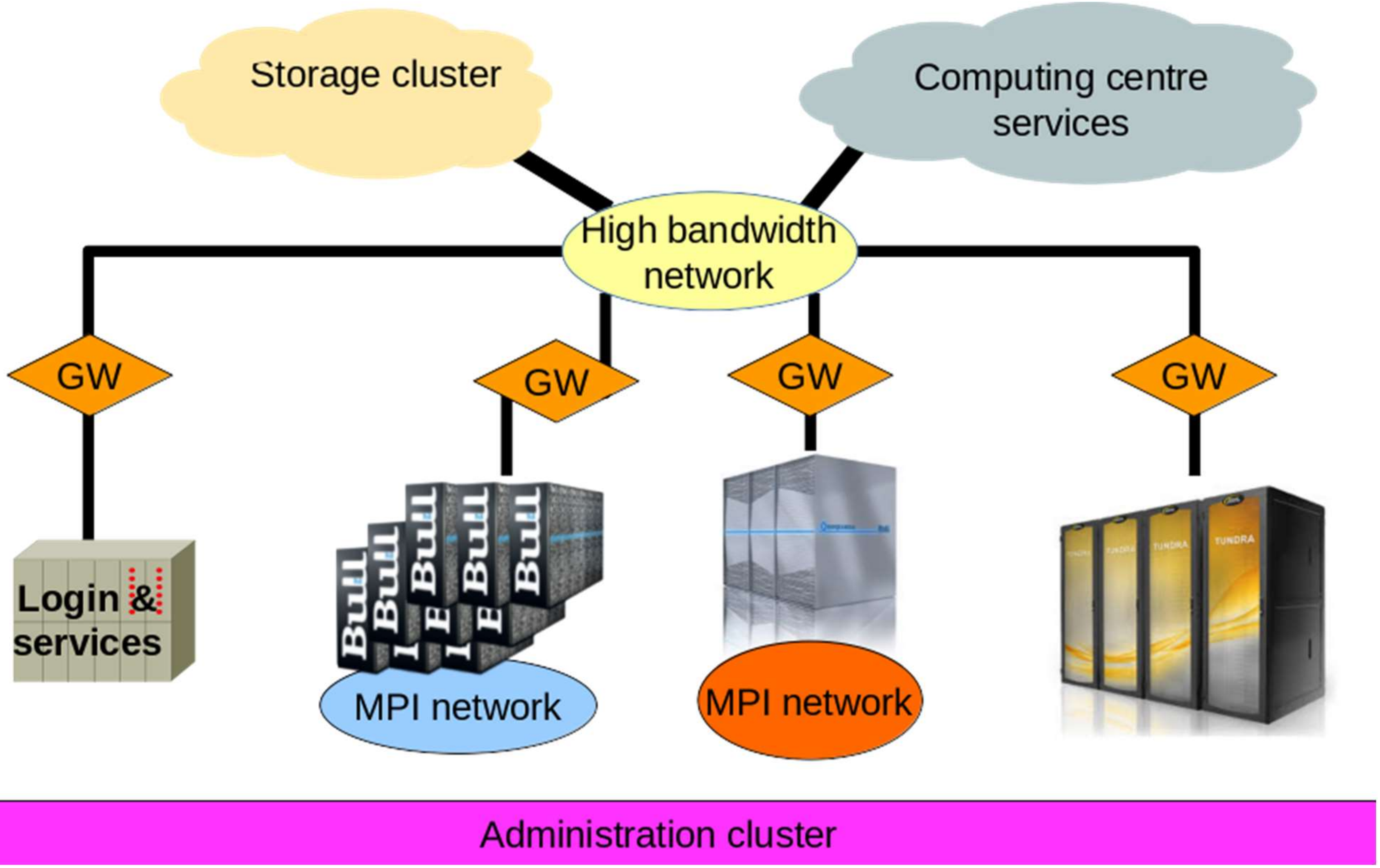
(soft skills)

Usability

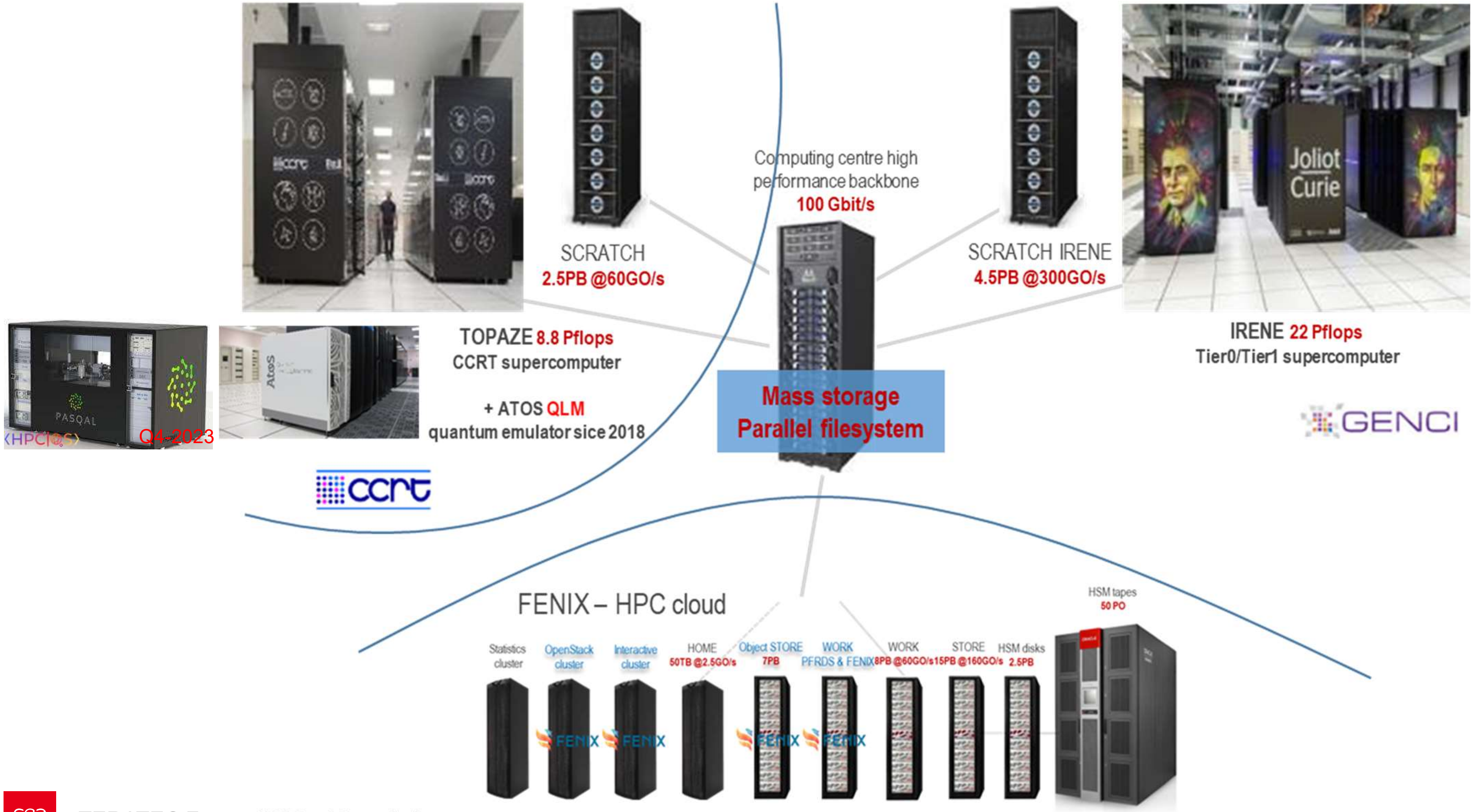
Reliability

Adaptability

Modular concept at CEA



Our state-of-the-art (example of TGCC…)



Computing centre high performance backbone **100 Gbit/s**

SCRATCH **2.5PB @60GO/s**

SCRATCH IRENE **4.5PB @300GO/s**

TOPAZE **8.8 Pflops**
CCRT supercomputer

+ ATOS **QLM**
quantum emulator sice 2018

⟨HPC|QS⟩ Q4-2023

IRENE **22 Pflops**
Tier0/Tier1 supercomputer

Mass storage
Parallel filesystem

FENIX – HPC cloud

Statistics cluster

OpenStack cluster

Interactive cluster

HOME **50TB @2.5GO/s**

Object STORE **7PB**

WORK PFRDS & FENIX

WORK **8PB @60GO/s**

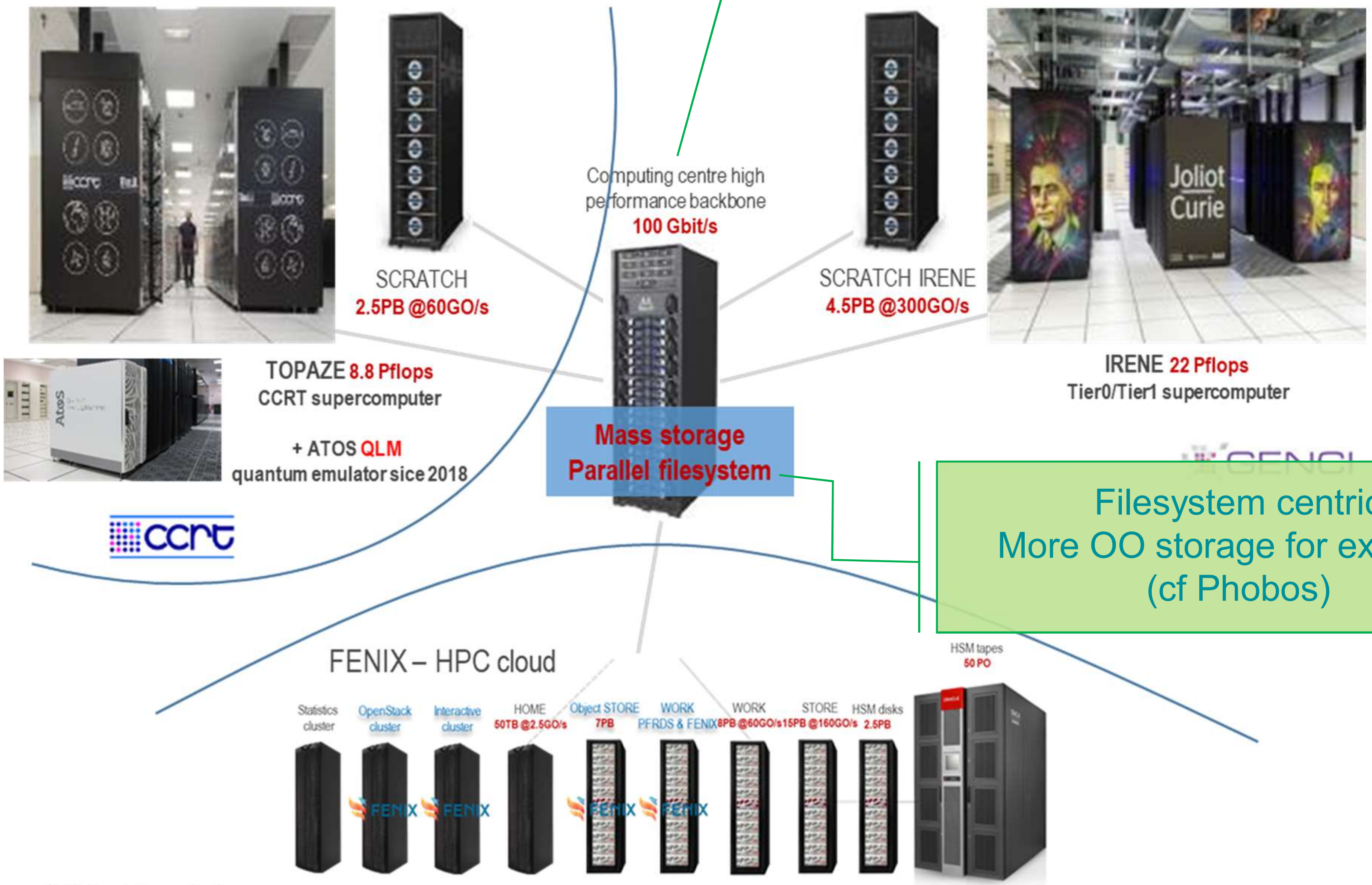
STORE **15PB @160GO/s**

HSM disks **2.5PB**

HSM tapes **50 PO**

Our state-of-the-art…

Backbone



Computing centre high performance backbone
100 Gbit/s

SCRATCH
2.5PB @60GO/s

SCRATCH IRENE
4.5PB @300GO/s

TOPAZE 8.8 Pflops
CCRT supercomputer

+ ATOS QLM
quantum emulator sice 2018

IRENE 22 Pflops
Tier0/Tier1 supercomputer

Mass storage
Parallel filesystem

FENIX – HPC cloud

Statistics cluster

OpenStack cluster

Interactive cluster

HOME
50TB @2.5GO/s

Object STORE
7PB

WORK
PFRDS & FENIX

WORK
8PB @60GO/s

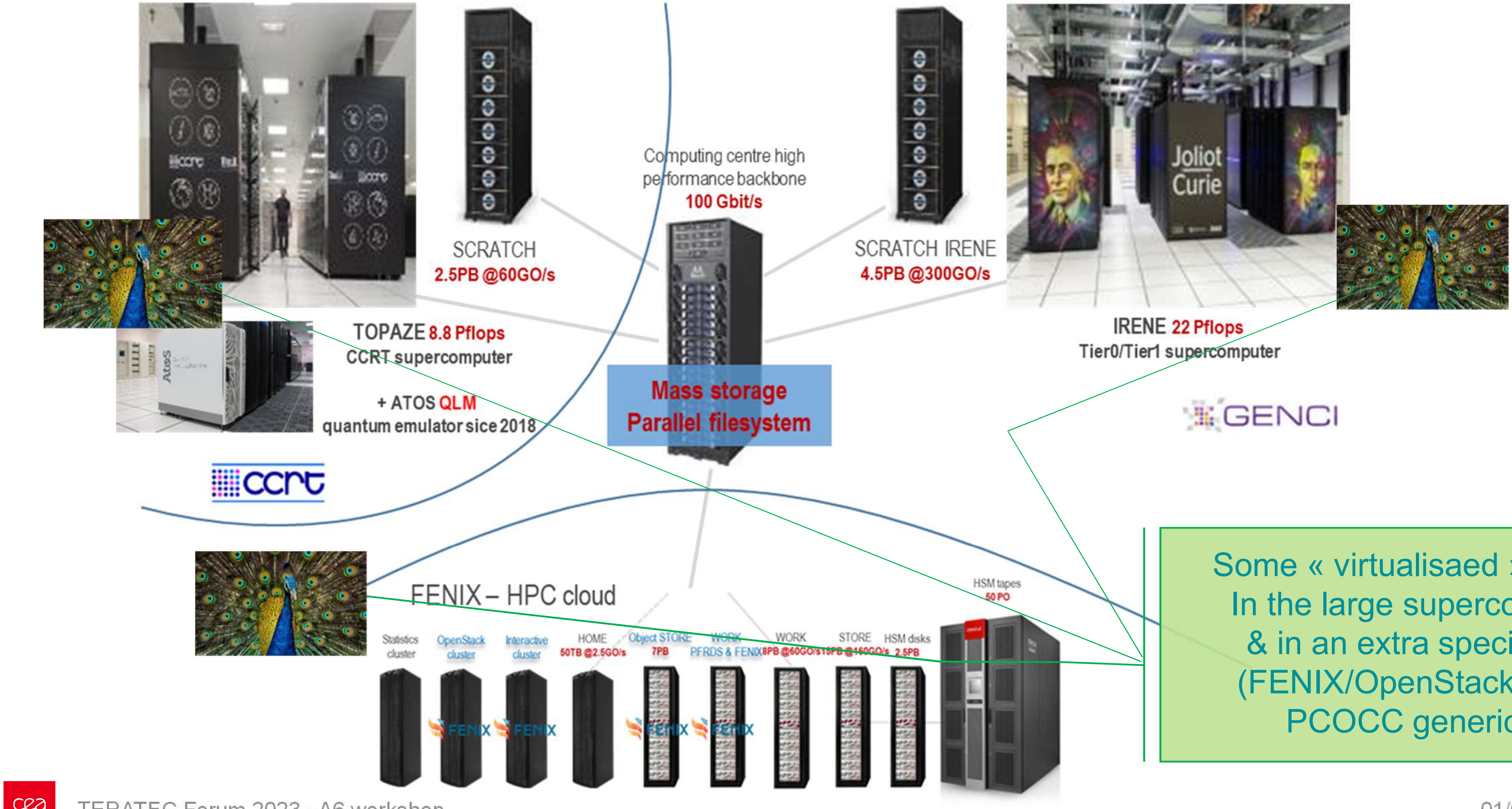
STORE
15PB @160GO/s

HSM disks
2.5PB

HSM tapes
50 PO

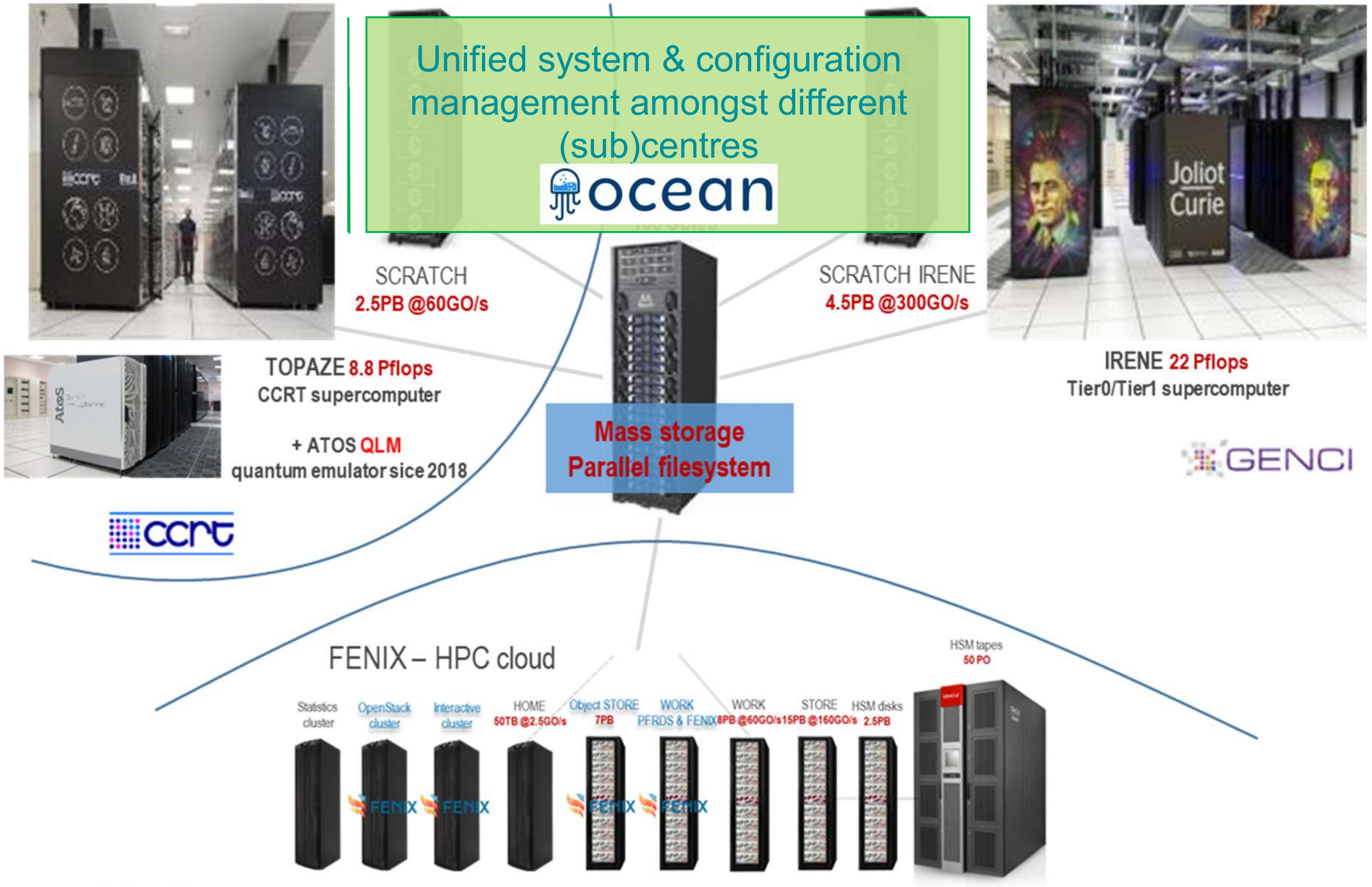
Filesystem centric
More OO storage for exascale
(cf Phobos)

Our state-of-the-art…



Some « virtualisaed » services
In the large supercomputers
& in an extra specific area
(FENIX/OpenStack cluster)
PCOCC generic tool

Our state-of-the-art…



Unified system & configuration management amongst different (sub)centres

ocean

SCRATCH 2.5PB @60GO/s

SCRATCH IRENE 4.5PB @300GO/s

TOPAZE 8.8 Pflops CCRT supercomputer

+ ATOS QLM quantum emulator sice 2018

IRENE 22 Pflops Tier0/Tier1 supercomputer

Mass storage Parallel filesystem

FENIX – HPC cloud

Statistics cluster	OpenStack cluster	Interactive cluster	HOME	Object STORE	WORK	WORK	STORE	HSM disks	HSM tapes
			50TB @2.5GO/s	7PB	PFRDS & FENIX	8PB @60GO/s	15PB @160GO/s	2.5PB	50 PO

1. CEA supercomputing complex
2. Typical computing centre architecture
3. **Data and storage considerations**
4. Configuration management and software system stack deployment - OCEAN
5. User services evolution
6. Towards Exascale

3. Data and storage considerations

Beyond Lustre and POSIX

❑ Growing diversity and size of DATA and METADATA

❑ Growing complexity of parallel jobs (data clients…) and workflows comprised of different jobs (e.g. data processing jobs part of workflows)

❑ Hierarchies of memories and media, up to disks and tapes

☞ Object storage a more flexible and agile paradigm

❑ Simple associative addressing (file, key)

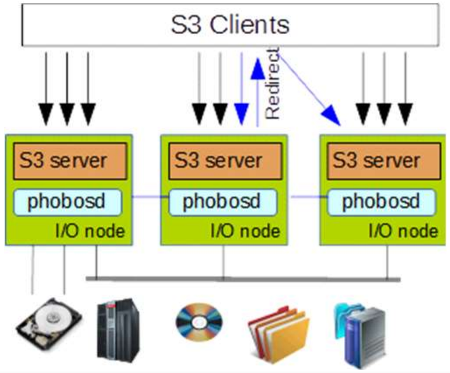
  ❑ Heavily used for streaming in the web

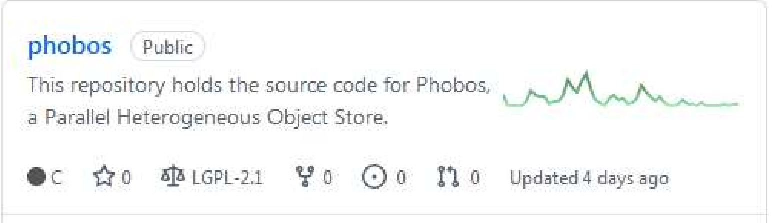
❑ Simpler, lighter

  ❑ Sw/Hw association: ephemeral services related to data nodes (less clients, custom processing)

Phobos by CEA DSSI

- CEA DAM - HPC – Opensources
- PHOBOS=Parallel Heterogeneous OBject Store
- **LGPL v2.1**
- **~50 k source lines, C+Python**



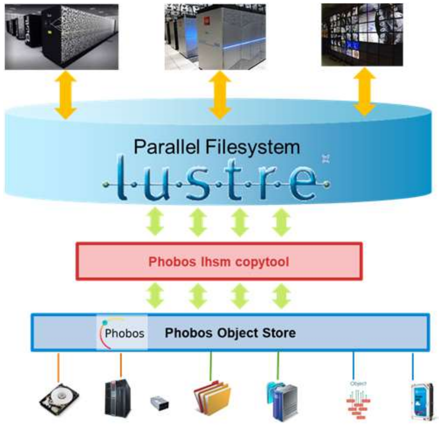


phobos Public

This repository holds the source code for Phobos, a Parallel Heterogeneous Object Store.

● C ☆ 0 LGPL-2.1 0 0 0 Updated 4 days ago

- Developped since 2016, deployment being generalised at CEA supercomputing complex
- Manages all kinds of devices HDD, flash, down to TAPES
  - Currently Lustre for SSD and flash, Phobos for tapes
- C API + command line interface
- Scheduling policies

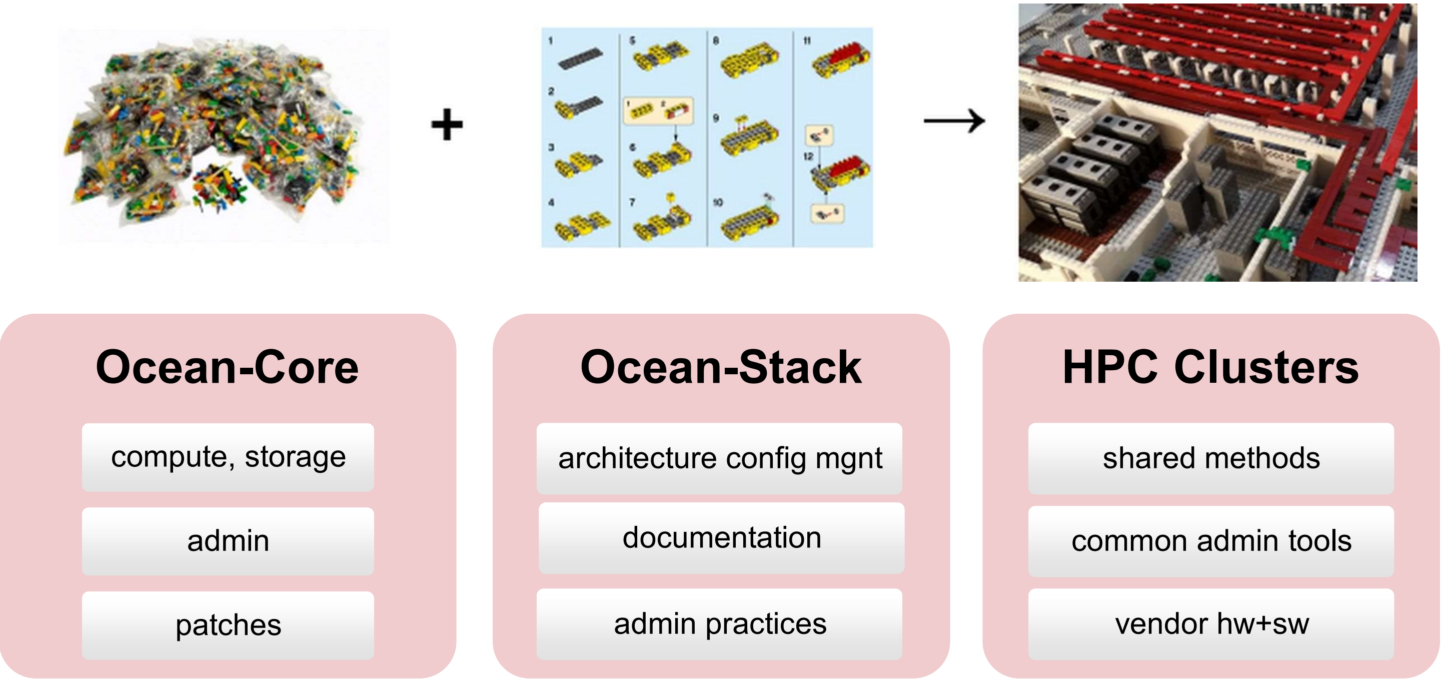


1. CEA supercomputing complex
2. Typical computing centre architecture
3. Data and storage considerations
4. **Configuration management and software system stack deployment - OCEAN**
5. User services evolution
6. Towards Exascale

4. Configuration management and software system stack deployment - ocean

OCEAN

- ❑ Since 2005, going open source/Linux based, then growing number of computing centres to manage
- ❑ Develop common pratices, pool efforts:
    - ❑ Open source basis BUT also a diversity of vendor software to be integrated
    - ❑ 2019: decision to deploy OCEAN in all our computing centres



+

→

Ocean-Core

- compute, storage
- admin
- patches

Ocean-Stack

- architecture config mgnt
- documentation
- admin practices

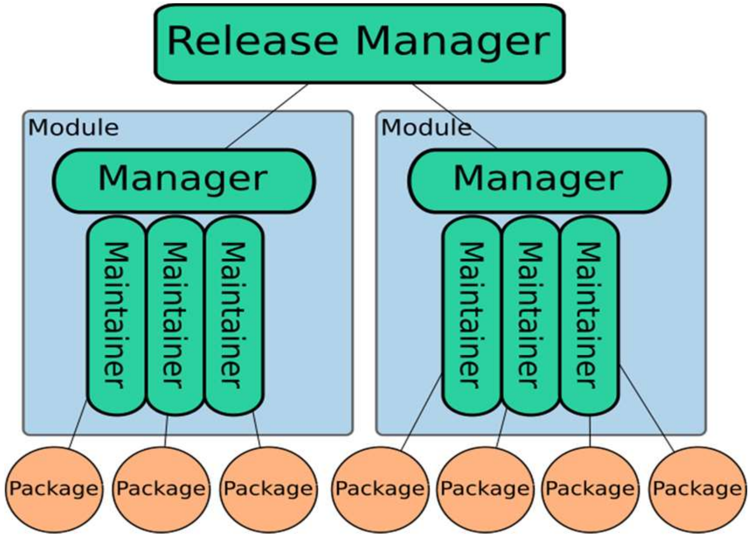
HPC Clusters

- shared methods
- common admin tools
- vendor hw+sw

OCEAN

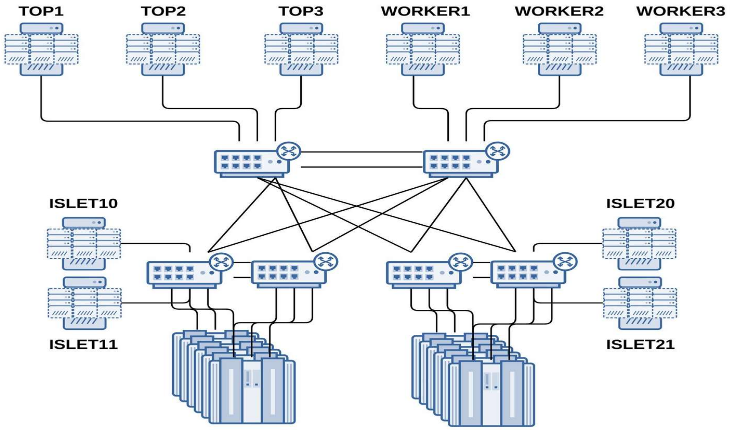
ocean

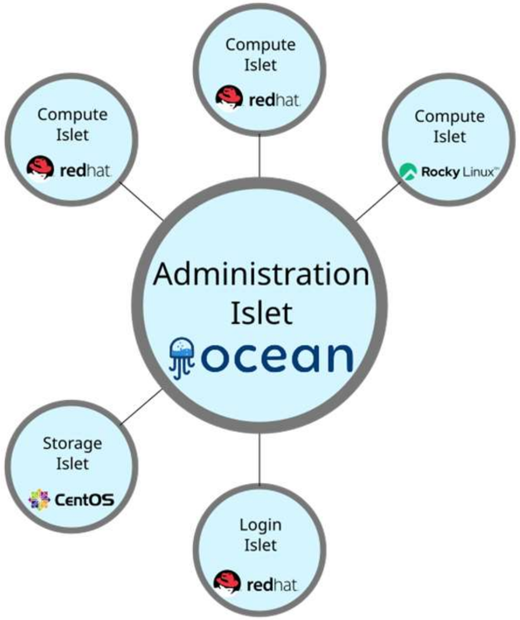
- ❑ OCEAN-core
  - ❑ HPC Linux distribution for compute and storage clusters
- ❑ Ocean encompasses 150+ packages
  - ❑ Incl. key core packages Slurm, Lustre, Puppet, OpenSSH, Qemu (VMs)
- ❑ Continuous integration Git+Jenkins



OCEAN

- OCEAN-stack
  - A cluster = admin islet or versatile islet (compute, storage, login node…)
    - A global admin cluster for each computing centre
  - Configuration management = a database of hw elements and connections IP addresses, core services, nodes images…



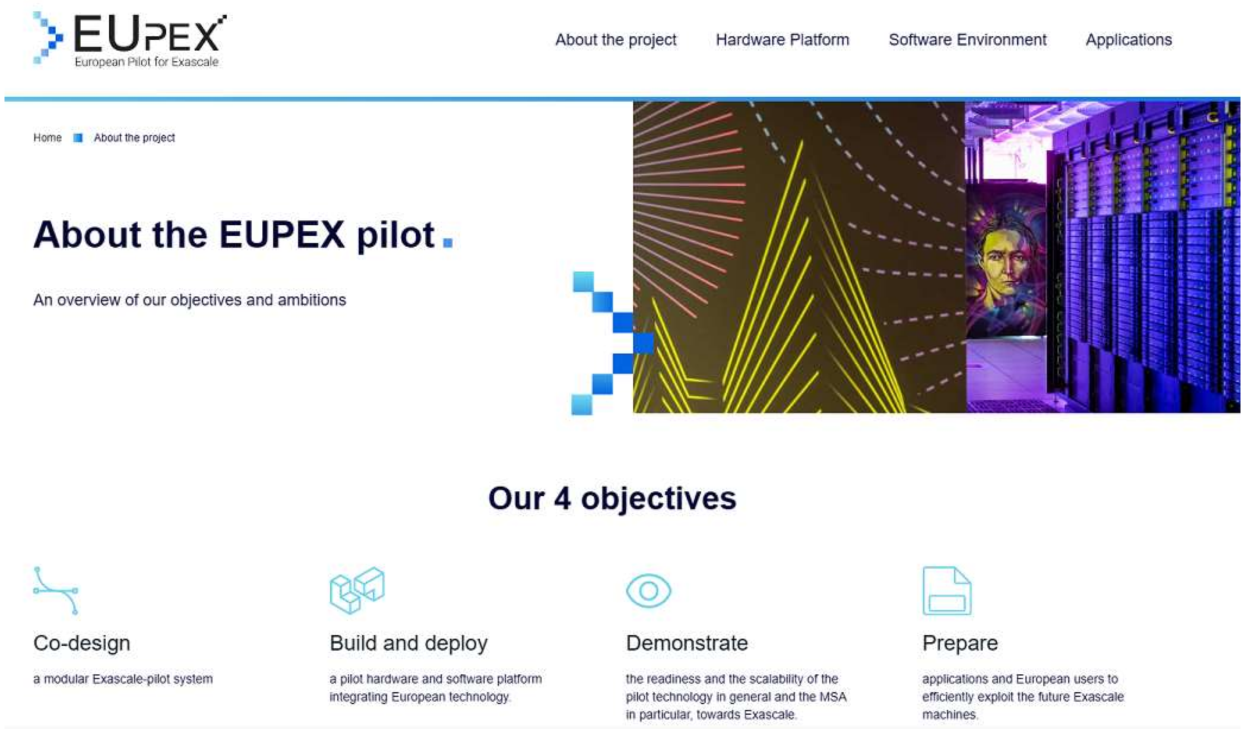


OCEAN

ocean

- ❑ At the core of EUPEX EuroHPC project => demonstrator SiPearl/RHEA+GPU at TGCC

Eupex - Home



EUPEX European Pilot for Exascale

About the project | Hardware Platform | Software Environment | Applications

Home ■ About the project

About the EUPEX pilot.

An overview of our objectives and ambitions

Our 4 objectives

Co-design
a modular Exascale-pilot system

Build and deploy
a pilot hardware and software platform integrating European technology.

Demonstrate
the readiness and the scalability of the pilot technology in general and the MSA in particular, towards Exascale.

Prepare
applications and European users to efficiently exploit the future Exascale machines.

1. CEA supercomputing complex
2. Typical computing centre architecture
3. Data and storage considerations
4. Configuration management and software system stack deployment - OCEAN
5. **User services evolution**
6. Towards Exascale

5. User services evolution - Virtualisation

Virtualisation, PCOCC…

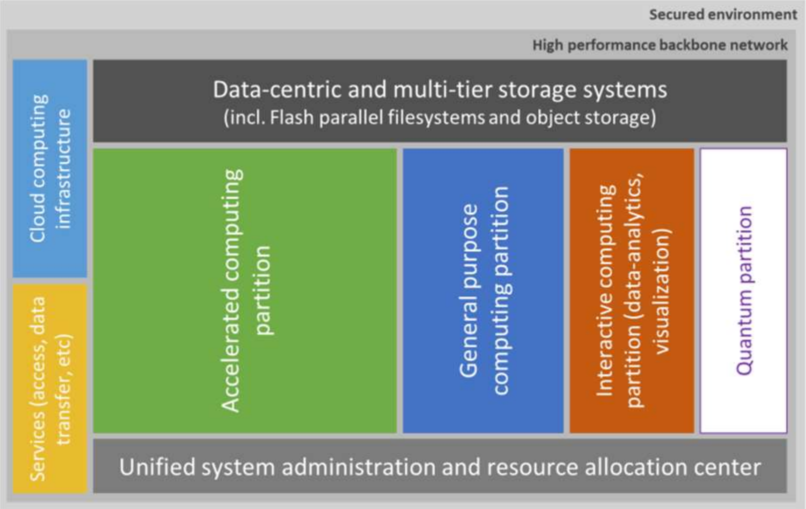
- User needs are more and more diverse + jobs/workflows have more and more dependencies/components + expected flexibility
- Give users a more customized and direct control => virtualisation
- PCOCC = Private Cloud On a Compute Cluster
  - Started in 2013 at CEA DSSI: same interface to manage VIRTUAL MACHINES and CONTAINERS
  - Simple, lightweight (low overhead): a virtual cluster launched with a single command
  - Same stable API for launching a VM or a container
  - Slurm-based
  - Since 2018: support containers (OCI standard format)
- ~30000 lines of Python + Rust; open source licence GPLv3
  - Already a few uses outside CEA

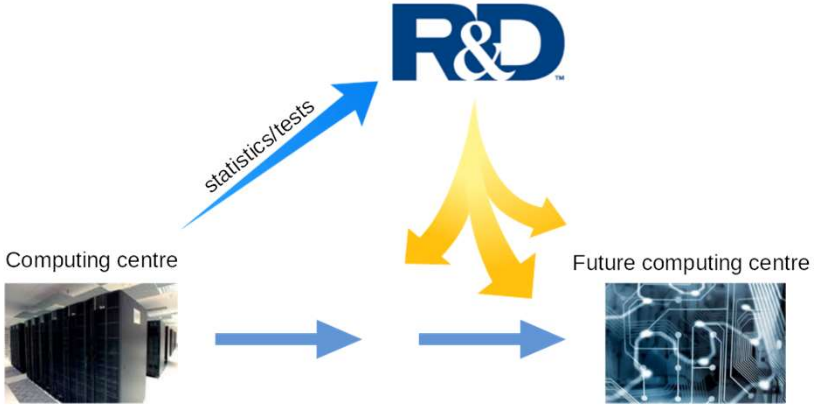
1. CEA supercomputing complex
2. Typical computing centre architecture
3. Data and storage considerations
4. Configuration management and software system stack deployment - OCEAN
5. User services evolution
6. **Towards Exascale**

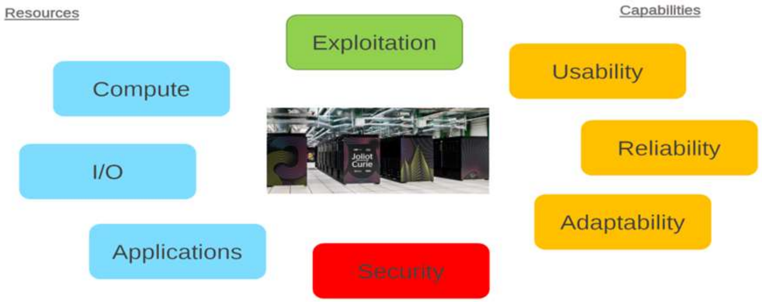
6. Towards Exascale

Application in 2025 to French Exascale Project

Go on applying and consolidating the aforementioned methods & tools to all our computing centres, and in particular to Jules Verne project with EuroHPC (and GENCI) in 2025


Secured environment
High performance backbone network
Cloud computing infrastructure
Services (access, data transfer, etc)
Data-centric and multi-tier storage systems
(incl. Flash parallel filesystems and object storage)
Accelerated computing partition
General purpose computing partition
Interactive computing partition (data-analytics, visualization)
Quantum partition
Unified system administration and resource allocation center





This talk was mostly about software
=> 3 key ingredients => user services