Big Data infrastructure:
What matters and what to expect?

SISC HPC Team, ULB/VUB
Michaël WAUMANS

What matters ?

BigData @ CÉCI : the missing link

- **CÉCI offers :**
  - MPI, HTC, GPUs and high-memory environments
  - not a BigData one

- **2018 BigData survey @ CÉCI revealed a demand :**
  - **95,3%** of the participants wanted “BigData” Softwares
  - **Hardware :** mostly GPU & Memory
  - **Software :**
    - Hadoop (HDFS, Yarn, HBase), Spark, Solr, ELK, Cassandra, Hive, Impala and more
    - Plus ~any library that can be downloaded & always at the latest version

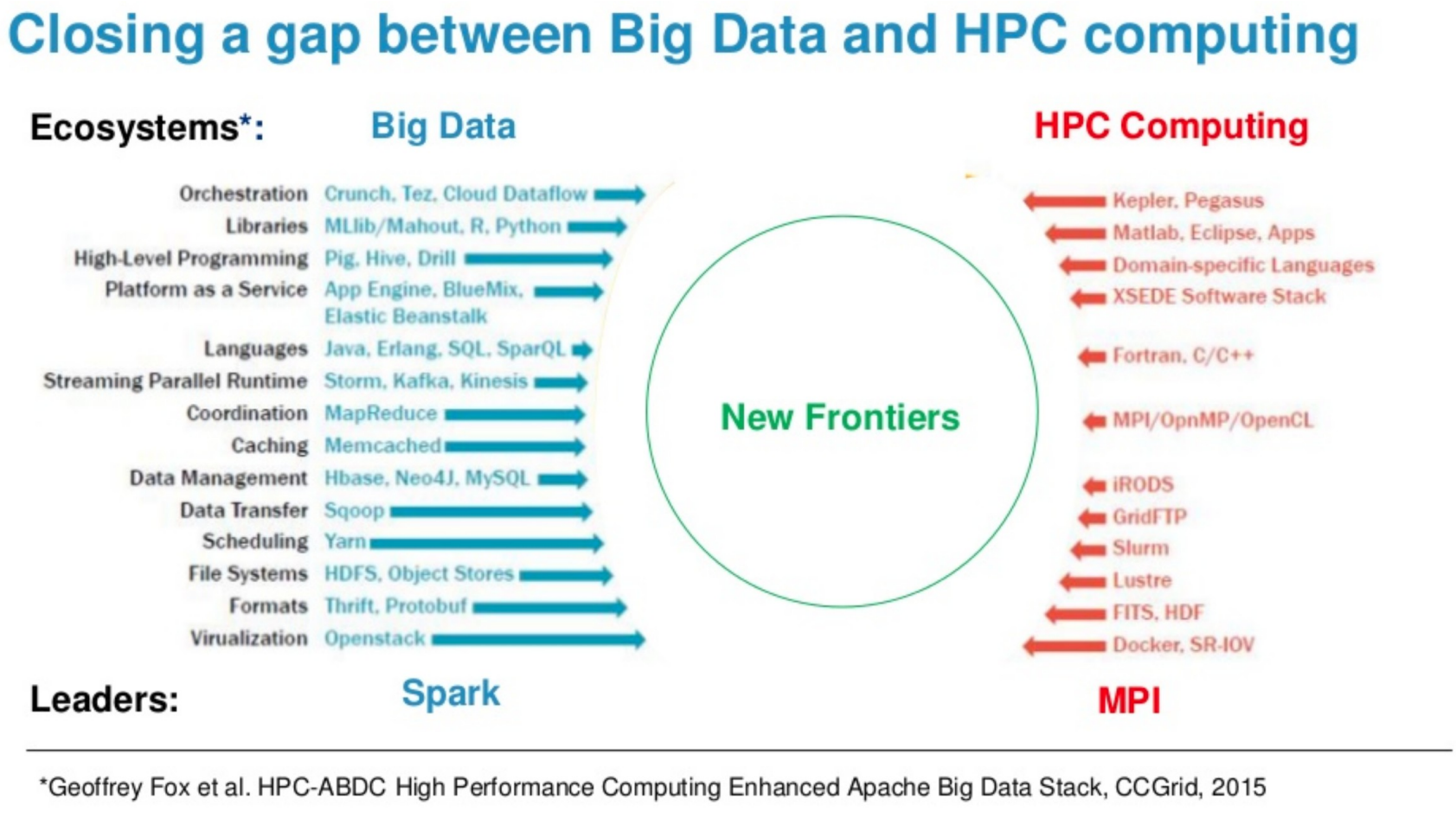
BigData @ CÉCI : the missing link

- **In the 2nd round of CÉCI clusters renewal, Vega2 was targeted as a BigData & HTC offering**

- **Mission :** Build up experience at management and usage levels
- **Means :** Test different configurations and interact with researchers to identify bottlenecks

BigData + HPC = High Performance Data Analysis

- **Remember the speech of Damien Francois ?**



From BigData to HPDA

- **BigData was developed for :**
  - A time when IB and 10+ Gbps networks were too expensive for large scale systems.
  - Taking easily advantage of very heterogeneous hardware configurations (Cores, Memory and Storage)
  - Bringing the computation to the data to avoid network traffic

- **Since then :**
  - High speed network became cheaper
  - Emergence of cloud solutions
  - New/more advanced distributed filesystems
  - Increasing “BigData” software diversity

From BigData to HPDA: SISC Experiment

- **Objectives**
  - Combine some HPC & BigData tech in a single platform
  - Design and build a platform from 2nd hand hardware
  - Target flexibility to accommodate the diversity of the “BigData” ecosystem
  - Evaluate OpenSource distributed filesystems suitable for a HPDA cluster
  - Evaluate OpenSource deployment solutions for maximum automation
  - Place **security as a top priority** in decision processes

From BigData to HPDA: the SISC Experiment

Numerous solutions were tested, discarded or kept...

- Full Hadoop Cluster (Cloudera, HortonWorks,...)
- Full Hadoop Cluster + Kerberos (idem)
- Provisioning systems (MAAS, Foreman,...)
- Virtualisation Tools (Ovirt, OpenNebula, MAAS,…)
- Storage (CEPH, ZFS, GlusterFS, combinations,…)
- Other softwares (ELK + Shield → Xpack → SearchGuard, Cassandra + DSE, MongoDB, Cassandra, HyperTable,...)
- Other softwares non strictly “BigData”(Galera, MySQL Cluster, Cockroach, Reddis,...)
- Notebooks & Web Interfaces (Hue, Jupyter, Zeppelin,...)

What we've learned (i.e. what matters, so far)

- **Proprietary hardware is a nightmare**
  - Limits everywhere...
  - Favour an Open Compute Project approach.
- **OpenSource software can be painful**
  - But commercial solutions are very expensive and reduce flexibility. Support can be disappointing too.
  - Local expertise with OpenSource software is a definitive major asset

- **Scheduling is hard, estimating properly its own resources needs too as a user**
  - Evaluating properly the resources required for a job can be hard on HPC/HTC. It's way worse in BigData.
  - **Our choice :** Instead of adding the complexity of BigData scheduling to HPC/HTC, let's try to move it to VM scheduling only
- **Two categories of users (Those who need sercurity (GDPR and so on, and those who don't)**
  - Security is not a developer's concern
  - Unless you pay licenses, provided it offers true security.
  - **Our choice :** Secure the environment, not the software (Hadoop, ELK, Cassandra,…)
- **Three main "user needs"**
  - Oriented towards the DATA (Permanent storage neither filesystem or object storage I.E.: ZOT)
  - Oriented towards the ALGORITHMS (Use of Spark, Storm and such paradigms)
  - Oriented towards BENCHMARK or EXPLORATION (On demand with full control on the services configuration)
  - **Our choice :** Answer all those requests through a single platform
- **Use a virtualisation/cloud platform, not bare metal**
  - Generic BigData compute nodes = impossible (Too many possible combinations...)
  - Redeploying nodes on-demand = complex and slow (idem)
  - OpenSource cloud = OpenStack = too complex and too big (Full Cloud)
  - **Our choice :** MAAS + Ansible + OVN + FreeIPA = isolated virtual secured clusters (Cloud… Kind of)

What to expect?

A try at HPDA : Management

Master node 1

Software Defined Networking

Provisioning

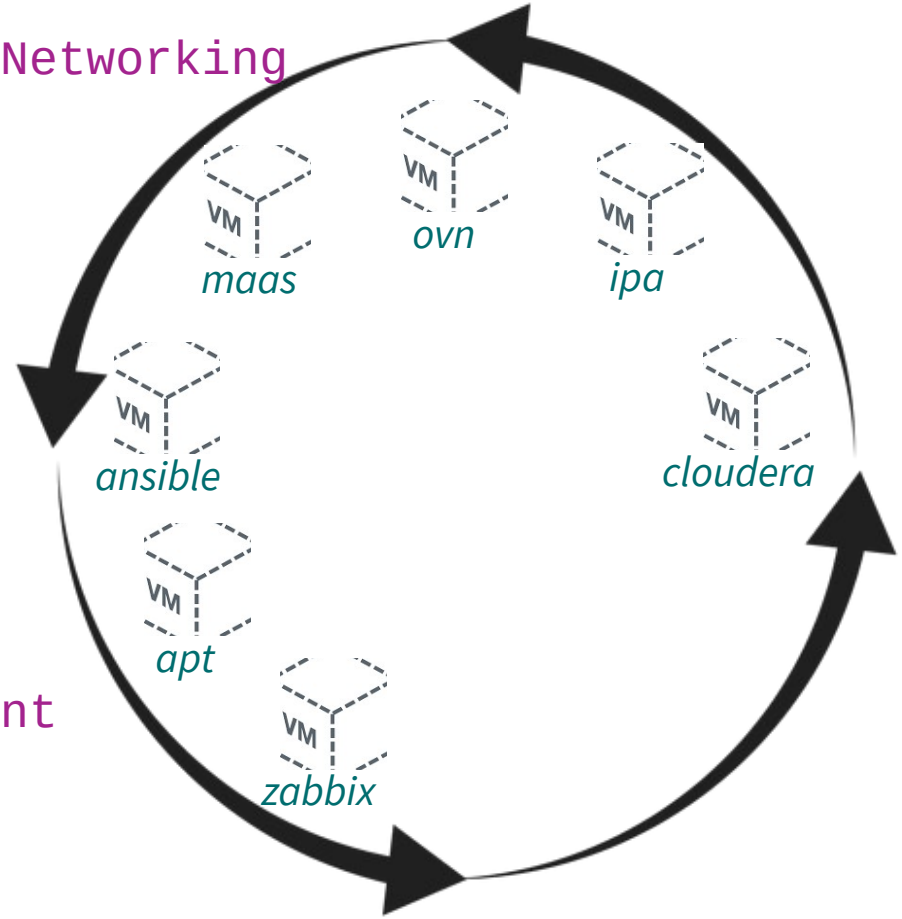
Orchestration

Caching

Monitoring

Authentication

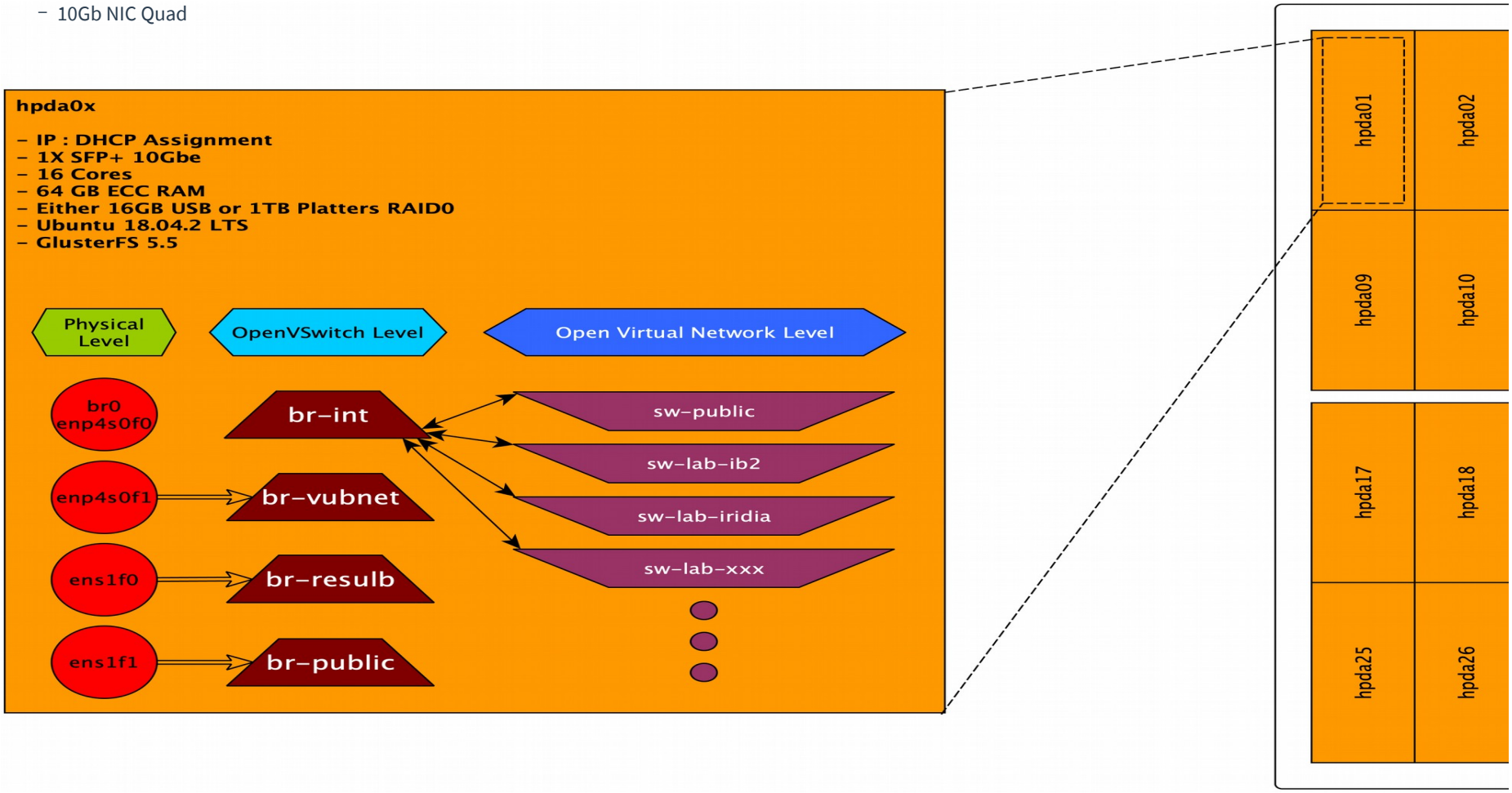
Clusters Management



Master node 2

A try at HPDA : Hypervisors

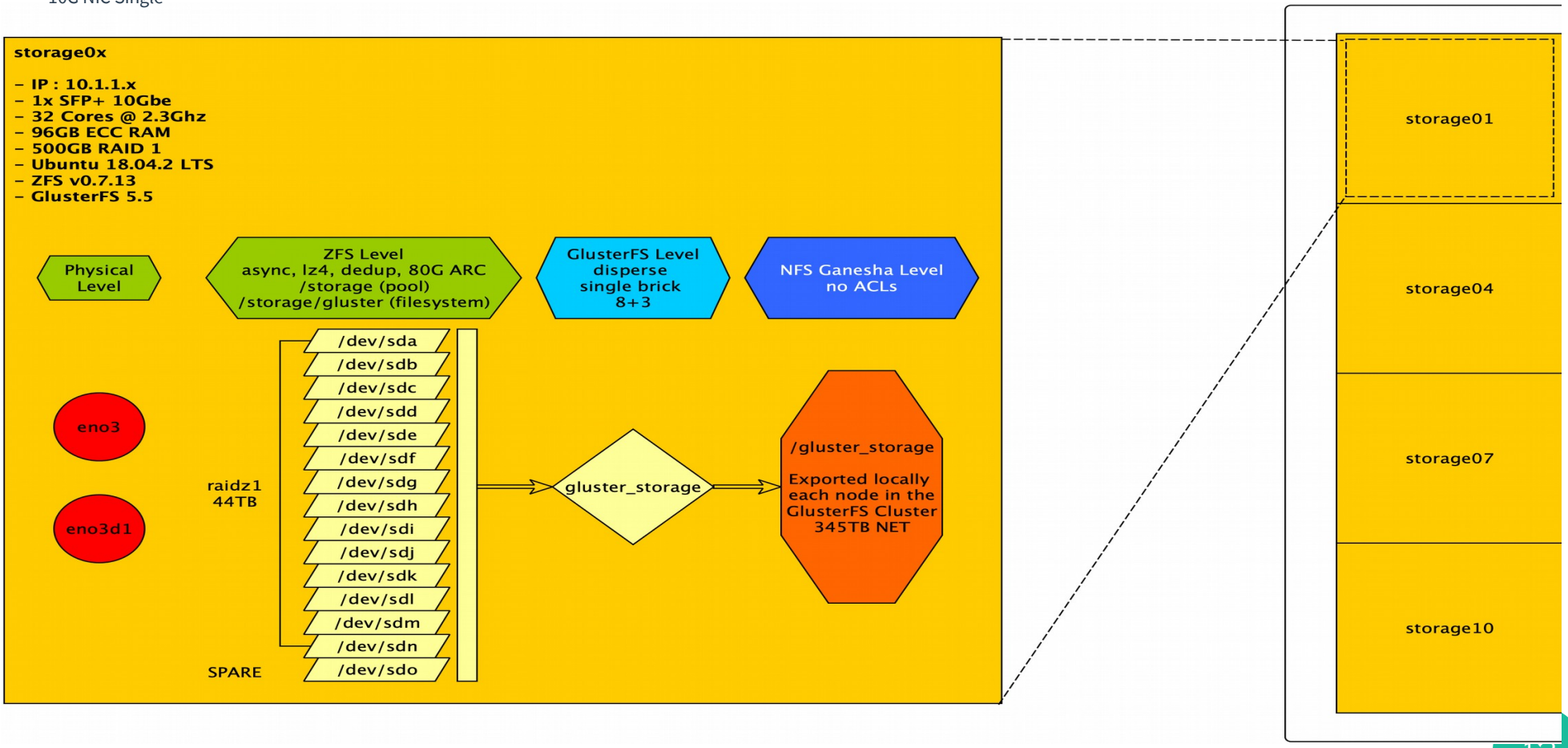
- **Hypervisor base on HP BL465c G7**
  - 512 Cores Opteron 62xx [16 Each]
  - 2048GB DDR3 ECC [64GB Each]
  - 32GB USB / 1TB RAID0
  - 10Gb NIC Quad



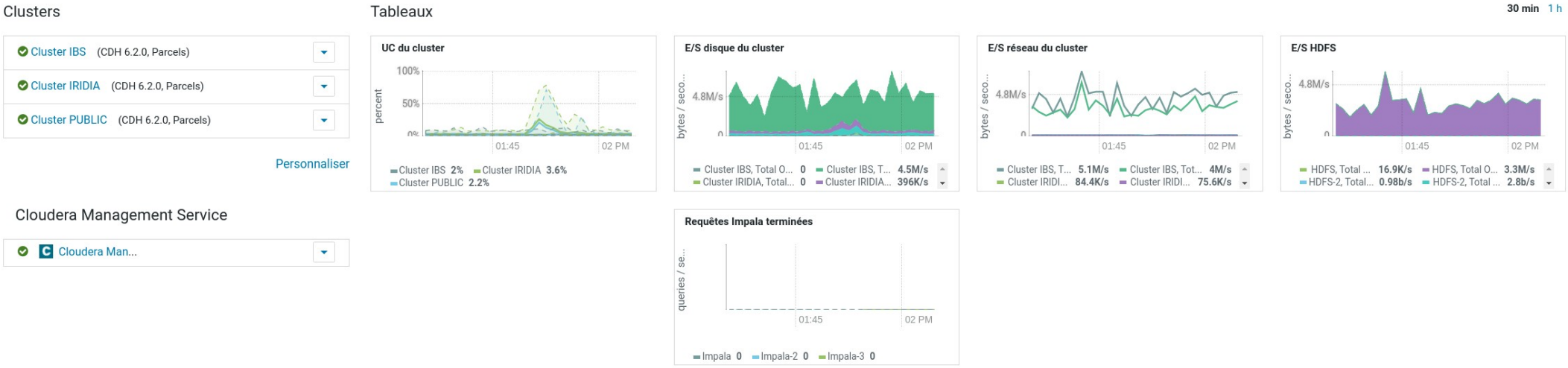
A try at HPDA : Storage

- **Storage node based on HP SL4540 G8**
  - 192 Cores @ 2.3Ghz [16 Each]
  - 1152GB ECC RAM [96GB Each]
  - 660TB RAW [60TB Each]
  - 500GB RAID1
  - 10G NIC Single

350 TB NET



What to expect ? Right now ?



3 Virtual clusters fully loaded with :

- HDFS, Hbase, Yarn, ZooKeeper, Hive, Impala, Solr, ELK, Cassandra, Hue, Zeppelin, JupyterHub, Kafka, and more…

- IBS - IB^2 laboratory (80TB)
- IRIDIA laboratory (80TB)
- **PUBLIC cluster for ULB/VUB Course that is open to CECI users for testing (80TB)**

Let’s be clear.

- **Infrastructure in its early stage !**
  - Accounts are local and are provided on demand only
    - Interested ? Contact us at **hpc@ulb.ac.be**
    - Please provide use your feedback, we need it :-)

- **Experimental stuff !**
  - Infrastructure built from recycled hardware :
    - NO 24/7 uptime warranty
    - NO data warranty either, keep copies
    - NO optimized compilation whatsoever
    - Limited storage and compute capacity
    - All is “best effort”
  - Opportunities to test BigData workflows .
  - Any library may be installed from repo fast

What to expect tomorrow?

- **HPC and BigData convergence: still a long way to go.**
  - (Any) BigData jobs scheduling in HPDA will require significant time and efforts investments.
  - Software deployment: in SysAdmin hands and will be step-by-step automated.
  - HPC offloading to HPDA: possible.

- **Infrastructure: needs to be ambitious.**
  - 10, 40 or more Gbps network, or Infiniband. Like HPC.
  - A lot of memory: OS level + virtualisation level + software level + caches = 128 – 512 GB per node. More than HPC.
  - Mixture of CPU and GPU cores per node. Convergence with HPC but without GPGPUs.
  - Favour continuous investments vs one big purchase every 4-6 years, i.e. more flexibility .

Questions ?