

Le stockage à l'IBMP

Vendu à DataCore ? 😇

JL Evrard - X/Stra - 5 novembre 2024

Une architecture essentiellement bipolaire

Vous avez dit "DataCore" ?

- La première brique tourne autour d'un système Software Defined Storage (SDS) "blocs" basé sur le logiciel DataCore **SANSymphony**
- La seconde brique repose sur un stockage pure objet qui utilise la solution DataCore **Swarm**
- La dernière brique consiste en une automatisation de passage de l'un à l'autre via le logiciel **FileFly** de chez... DataCore

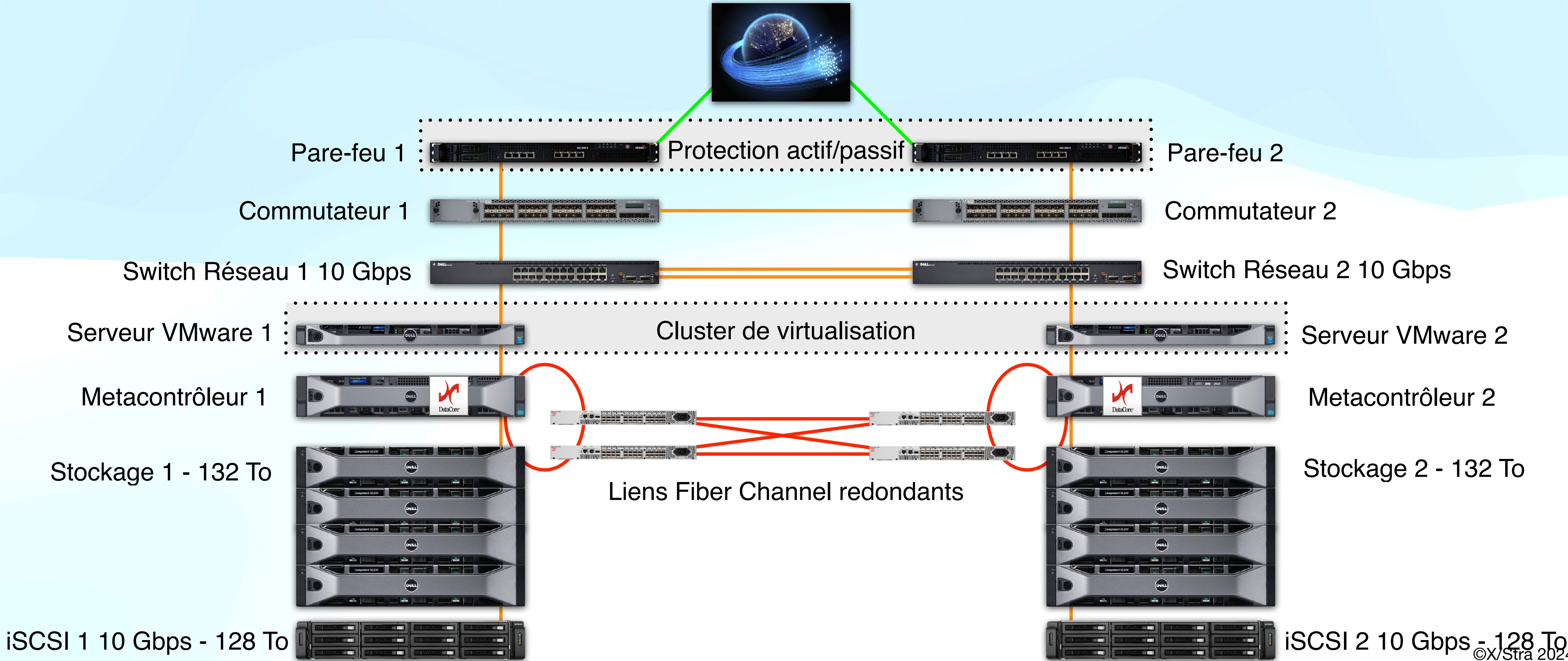
Le bloc selon SANSymphony

Le SDS simple et efficace

- Un choix qui s'est imposé à l'époque (~2015) face à deux vagues :
 - Une numérisation de plus en plus forte des données scientifiques
 - La mise en place de systèmes virtualisés
- La volonté d'offrir un service 24/24 7/7 aux chercheurs avec un maximum de sécurité, mais avec le minimum de backup :
 - volume de données trop important : coût prohibitif pour les sauvegarder

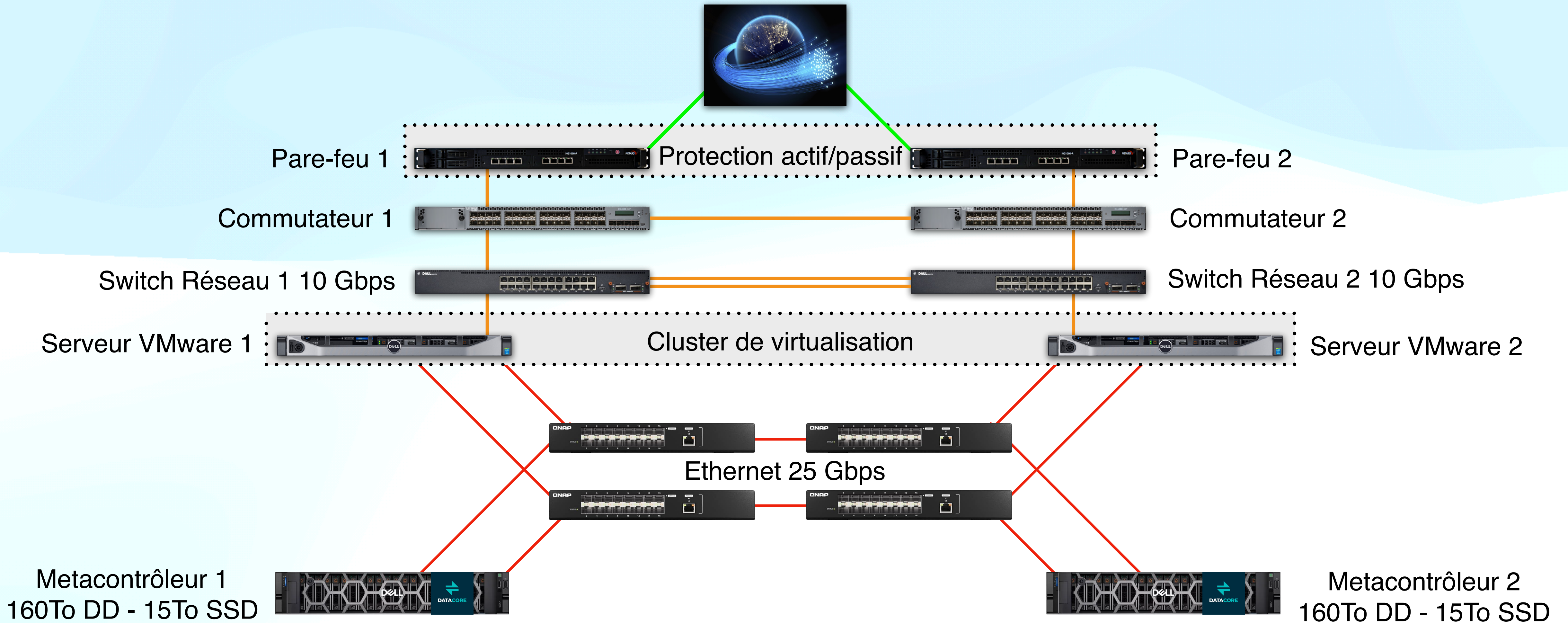
Le bloc selon SANSymphony

Architecture "BioCore" - v1



Le bloc selon SANSymphony

Architecture "BioCore" - v2 (fin 2024)



Le bloc selon SANSymphony

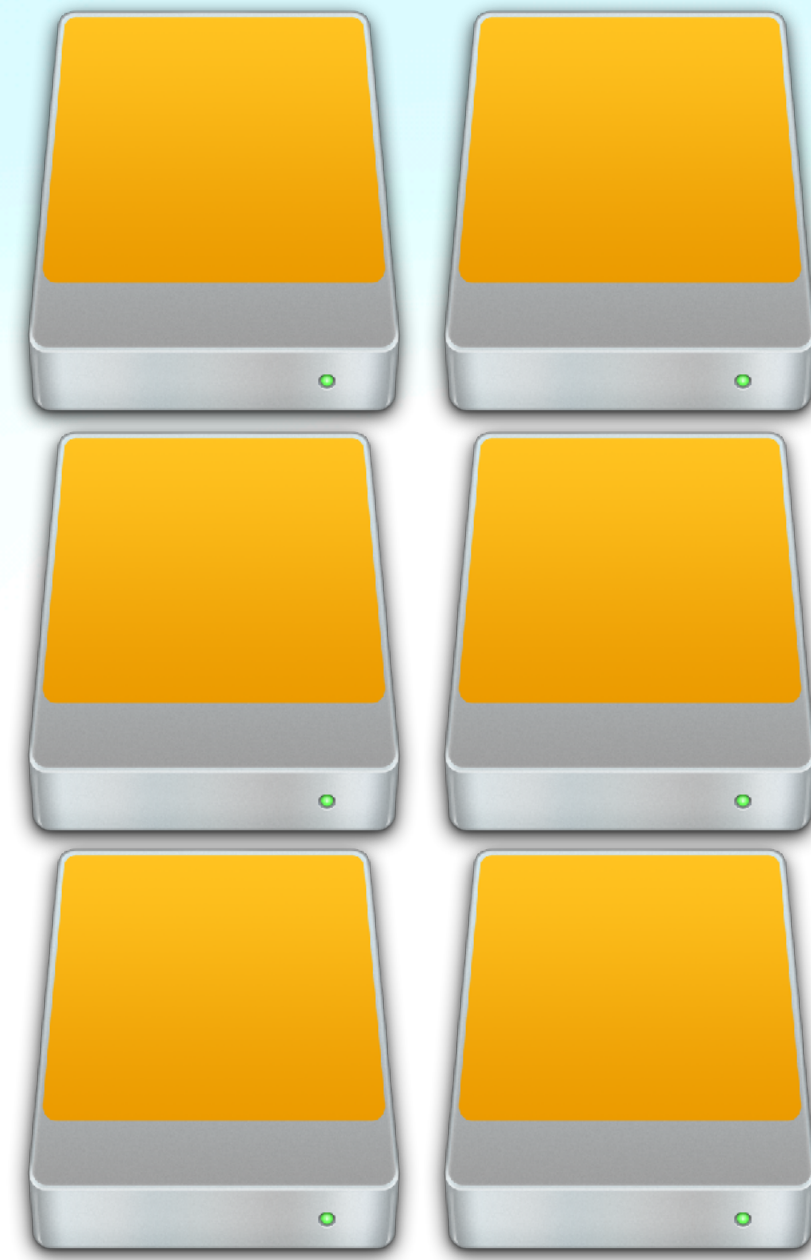
L'architecture logique

Disques physiques

Disques logiques

Pools

Volumes logiques



Le bloc selon SANSymphony

Le logiciel

- SANSymphony nécessite un OS Windows Server (ça fait peur, mais on s'y fait)
- Pour ce qui est du stockage utilisable : si Windows le voit, alors SANSymphony peut l'utiliser (USB, iSCSI, SATA, SAS, FC, NVMe...)
- Toute la gestion se fait en mode graphique à travers une console unique : objectivement, c'est très agréable
- L'intégration première était pour du VMware, mais à ce jour, tous les hyperviseurs du marché sont adressables (**Proxmox**, HyperV,...)

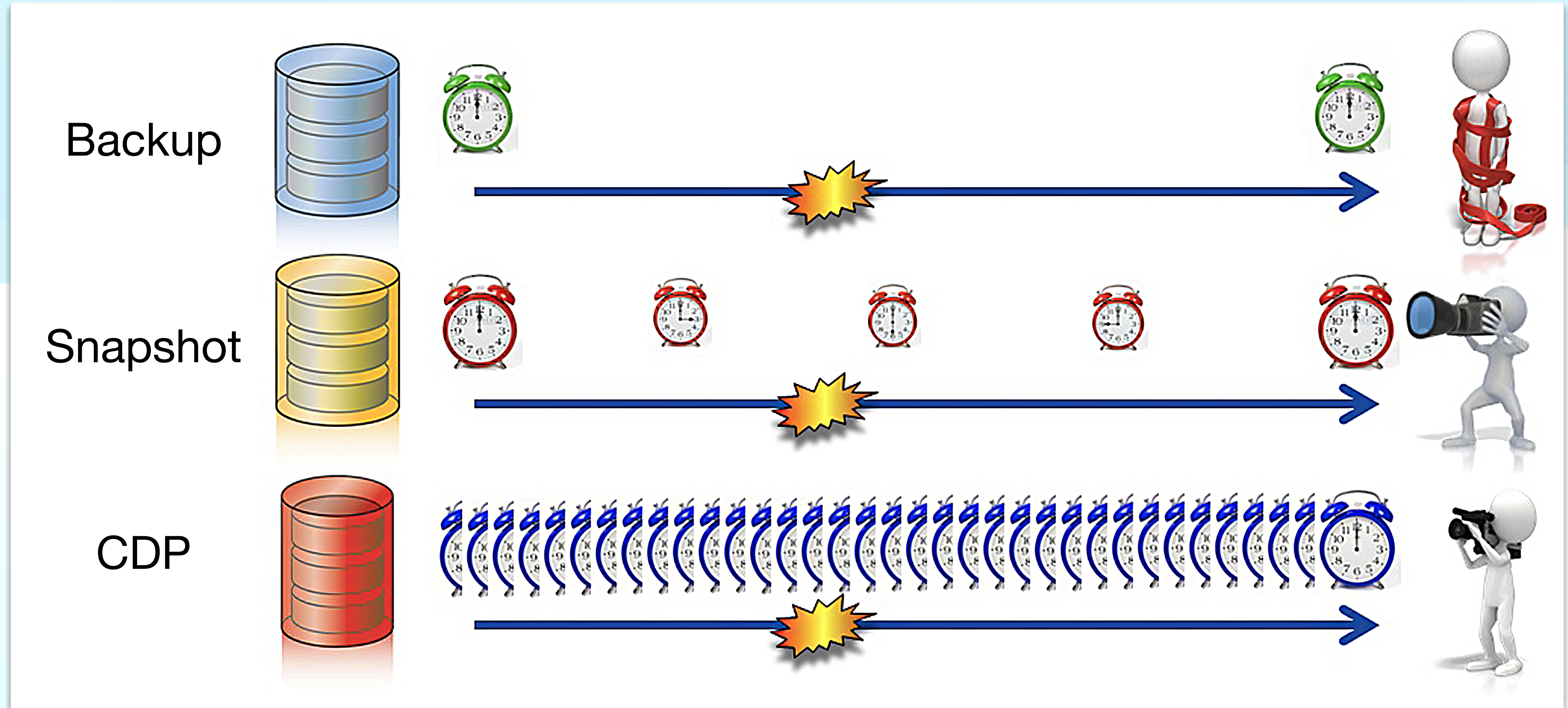
Le bloc selon SANSymphony

Le logiciel

- SANSymphony comporte de très nombreuses fonctionnalités :
 - Réplication synchrone (ou asynchrone)
 - Auto-tiering
 - CDP (Continuous Data Protection) - Snapshot
 - Déduplication
 - Compression
 - Optimisation de Capacité = déduplication & compression

Le bloc selon SANSymphony

Le logiciel



Le bloc selon SANSymphony

Le logiciel

The screenshot displays the DataCore Management Console in 'Advanced View'. The interface is divided into several sections:

- Navigation Bar:** Includes 'Home', 'Common Actions', and 'Virtual Disk Actions'. The 'Virtual Disk Actions' menu contains icons for 'Serve to Hosts', 'Start Reclamation', 'Abort Reclamation', 'Delete', 'Create Rollback', 'Create Snapshot', and 'Create Replication'.
- DataCore Servers:** A tree view on the left showing the hierarchy: IBMP DC Server Group > DC1 > Virtual Disks > Critical CDP > DC_VDisk20_7k_Datastore [21 TB].
- Hosts:** A tree view at the bottom left showing: 10.10.1.7 > IBMP > BIOCORE > newesxi1.biocore.lan and newesxi2.biocore.lan.
- Main Panel:** Displays details for the selected 'Virtual Disk DC_VDisk20_7k_Datastore'.
 - General Info:** Description: 'Volume de stockage N°1 des plateformes sur serveur adslave Windows 2016 (vmfs)'. Size: 21 TB. Reserved space: 0 B. Sector size: 512 B. Storage profile: Normal. Host(s): newesxi1.biocore.lan, newesxi2.biocore.lan.
 - Info Tab:** Shows 'DC2 (Running)' and 'DC1 (Running)' sections. For DC2: Data status is 'Up to date', Host access is 'Read/Write', Mirror link is 'Available, 2 Path(s)', Host(s) are 'newesxi2.biocore.lan (Connected), newesxi1.biocore.lan (Connected)', Storage source is 'DC2_Capactif, 20,03 TB allocated (Online)'. For DC1: Data status is 'Up to date', Host access is 'Read/Write', Mirror link is 'Available, 2 Path(s)', Host(s) are 'newesxi2.biocore.lan (Connected), newesxi1.biocore.lan (Connected)', Storage source is 'DC1_Capactif, 20,03 TB allocated (Online)'. Both sections show 'Tier affinity: 1, 2, 3'.
 - Operations:** A table at the bottom showing recent actions:

| Time | Action | Status |
|-------------------------|--|--------|
| 18/09/2024 13:46:30.870 | Refresh all | Done |
| 18/09/2024 14:04:16.825 | Set virtual disk properties - DC_VDisk1_7k_Datastore_RDM | Done |

Nouveau design hyper-redondé

Seagate Exos AP (2U24)



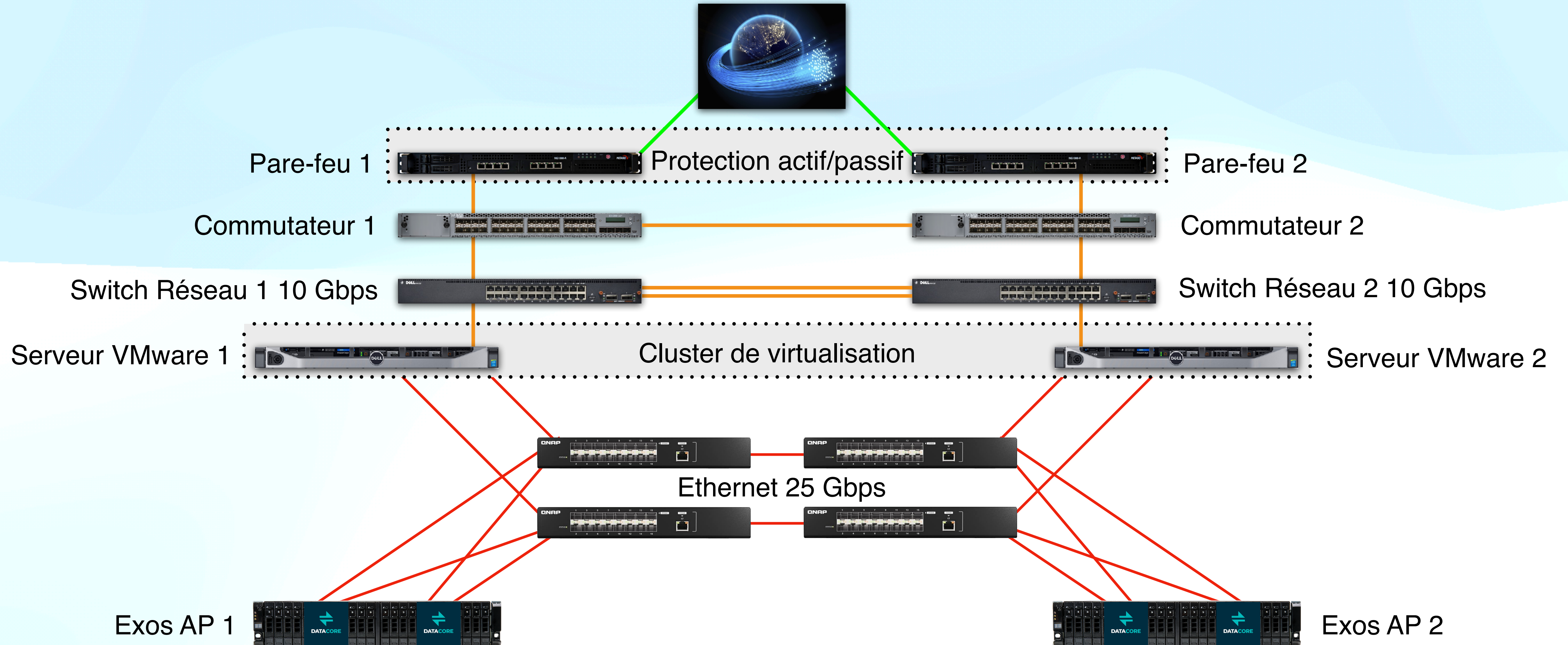
Nouveau design hyper-redondé

Seagate Exos AP (2U24)



Nouveau design hyper-redondé

Seagate Exos AP (2U24)



Nouveau design hyper-redondé

Seagate Exos AP (2U24)

The screenshot displays the management interface for Seagate Exos AP (2U24) servers. The interface is organized into several sections:

- Navigation Bar:** Includes Home, Common Actions, and various tool icons for Virtual Disks, Disk Pools, Hosts, Storage Profiles, Virtual Disk Templates, System Health, Live Performance, Recorded Performance, Replication Performance, Tasks, Reports, Event Log, Alerts, Server Group Connections Panel, Reset Layout, Turn ON Advanced View, Users, Roles, and Help.
- DataCore Servers:** A tree view on the left showing a hierarchy of server groups. Two groups are highlighted with red and green boxes:
 - Site 1 (Red Box):** Contains server groups 'seagate-cio-a' and 'seagate-cio-b'. Each group includes Physical Disks, Capacity Optimization, DataCore Disks, Virtual Disks, Disk Pools, and Server Ports.
 - Site 2 (Green Box):** Contains server groups 'seagate-mar-a' and 'seagate-mar-b'. Each group includes Physical Disks, Capacity Optimization, DataCore Disks, Virtual Disks, Disk Pools, and Server Ports.
- Tasks:** A central panel showing a table of tasks. One task, 'Evacuate On Fail', is listed with a state of 'Idle'. The table includes columns for Name, State, Current action, Last start time, Last stop time, Description, and Enabled.
- Hosts:** A panel on the right showing a list of hosts under the 'ASP-MANAGEMENT' group, including 'esxmgmt-c01.aspserveur.local', 'esxmgmt-c02.aspserveur.local', 'esxmgmt-m01.aspserveur.local', and 'esxmgmt-m02.aspserveur.local'.
- Operations:** A panel at the bottom showing a list of recent operations, all of which are completed successfully. The operations include rescanning various ports on the 'seagate-cio-a' server group.

L'objet selon Swarm

L'objet jusqu'au bout du disque

- Une solution trouvée "par hasard" :
 - ActiveScale (Western Digital)
 - Quantum - Passage en mode locatif...
 - Caringo Swarm - Racheté par DataCore...
- Une solution pensée objet depuis ses fondations : le disque est un objet qui contient des objets (triplication) ou des parties d'objets (Erasure Coding)
- Un OS propriétaire logé en RAM (CastorOS) - Un "essaim" de disques...

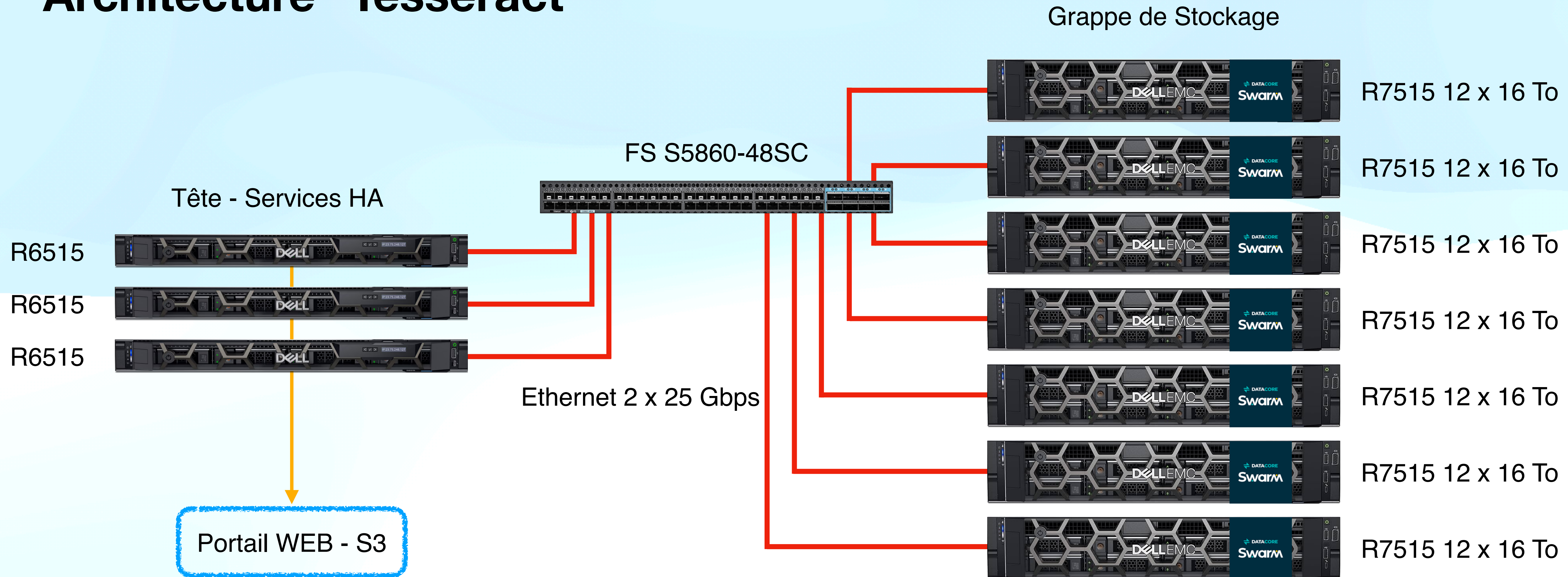
L'objet selon Swarm

La solution du pauvre...

- Besoin d'un "petit" S3 (10~100 To) ? SNS (Single Node Swarm) peut être une solution intéressante :
 - Une appliance capable d'être déployée en moins d'une heure
 - Installation en conteneurs automatisée
 - Possibilité de la raccrocher à un Swarm traditionnel
- Cas d'école : récupérer les données d'un matériel d'acquisition (et transférer vers le S3 "datacenter") sans se poser trop de questions...

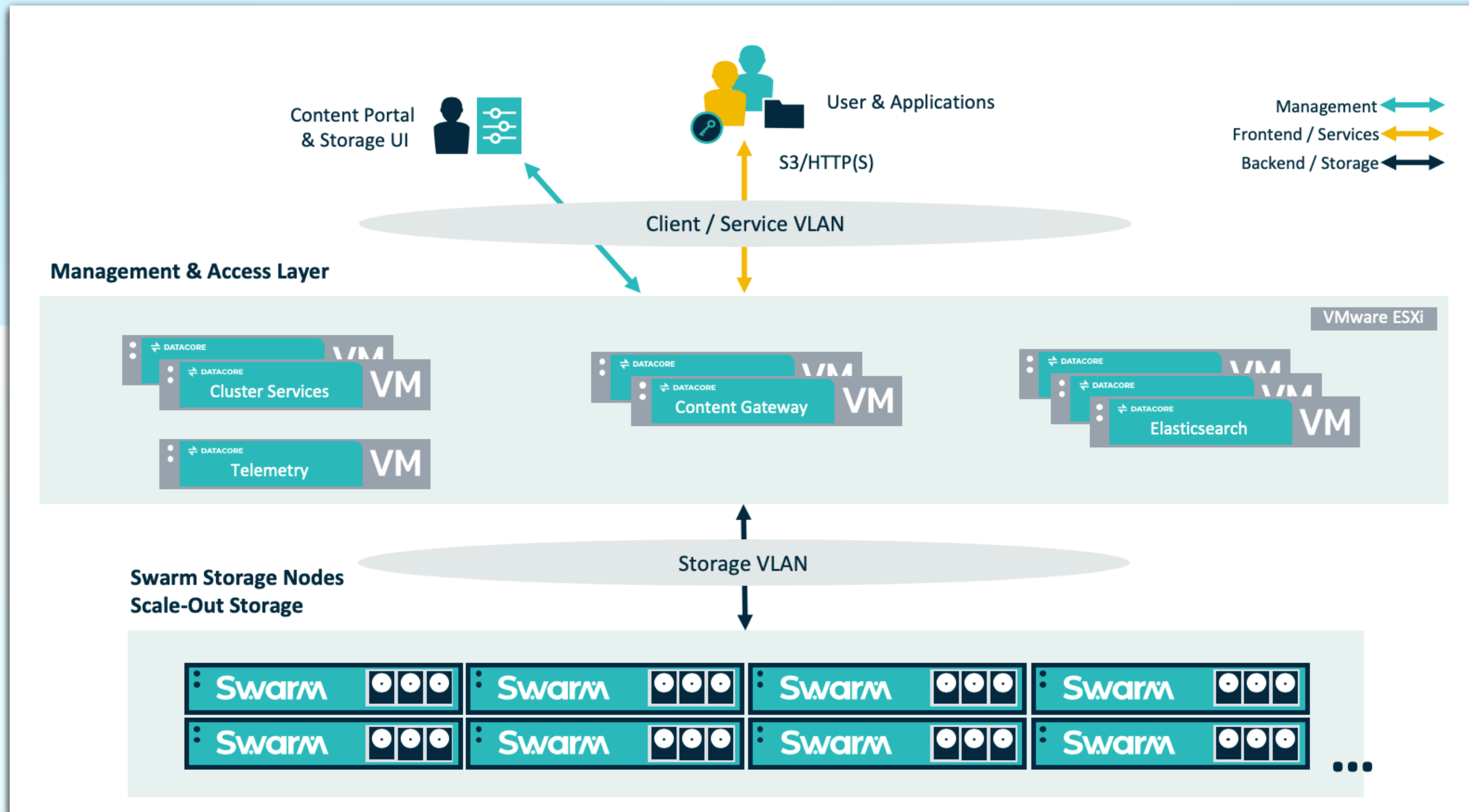
L'objet selon Swarm

Architecture "Tesseract"



L'objet selon Swarm

Architecture logique



L'objet selon Swarm

Architecture logique

The screenshot displays the VMware ESXi vSphere interface. The main window shows a list of virtual machines (VMs) under the host 'tesseract1.biocore.lan'. The VMs are listed in a table with columns for name, state, used space, OS version, host name, CPU usage, and memory usage. The VMs include CSN, GATEWAY1, ELSEARCH1, HAPROXY1, Windows Server 2019, SCS, and NODETEST. The bottom section shows a 'Tâches récentes' (Recent Tasks) table, which is currently empty.

| Machine virtuelle | État | Espace utilisé | SE invité | Nom d'hôte | CPU d'hôte | Mémoire d'hôte |
|---------------------|---------|----------------|---|------------|------------|----------------|
| CSN | Norm... | 50 Go | CentOS 6 (64 bits) | Inconnu | 0 MHz | 0 Mo |
| GATEWAY1 | Norm... | 100 Go | CentOS 7 (64 bits) | Inconnu | 0 MHz | 0 Mo |
| ELSEARCH1 | Norm... | 2 To | CentOS 7 (64 bits) | Inconnu | 0 MHz | 0 Mo |
| HAPROXY1 | Norm... | 50 Go | CentOS 7 (64 bits) | Inconnu | 0 MHz | 0 Mo |
| Windows Server 2019 | Norm... | 200,1 Go | Microsoft Windows Server 2019 (64 bits) | Inconnu | 0 MHz | 0 Mo |
| SCS | Norm... | 4,36 Go | CentOS 4/5/6/7 (64 bits) | Inconnu | 0 MHz | 0 Mo |
| GATEWAY01 | Norm... | 10,69 Go | Red Hat Enterprise Linux 8 (64 bits) | gateway01 | 39 MHz | 5,07 Go |
| HAPROXY01 | Norm... | 10,69 Go | Red Hat Enterprise Linux 8 (64 bits) | haproxy01 | 38 MHz | 3,87 Go |
| ELSEARCH01 | Norm... | 66,54 Go | Red Hat Enterprise Linux 8 (64 bits) | elsearch01 | 38 MHz | 63,21 Go |
| SCS01 | Norm... | 7,15 Go | Red Hat Enterprise Linux 8 (64 bits) | scs01 | 67 MHz | 3,74 Go |
| NODETEST | Norm... | 480 Go | CentOS 7 (64 bits) | Inconnu | 0 MHz | 0 Mo |

L'objet selon Swarm

Portails WEB

The screenshot displays the DATACORE Swarm web interface for a cluster named 'clustercsn.biocore.lan'. The interface is organized into several sections:

- Header:** Includes the DATACORE Swarm logo, the cluster name, and the user 'dcsadmin'.
- Navigation:** A sidebar on the left contains 'Tableau de bord', 'Cluster', 'Rapports', and 'Paramètres'.
- Cluster Overview:** Shows the cluster name 'tesseract.ibmp.unistra.fr' and its last update time '2024-09-18 15:44:03 CEST'. A 'SANTÉ' (Health) section features a green circular gauge and a table with the following data:

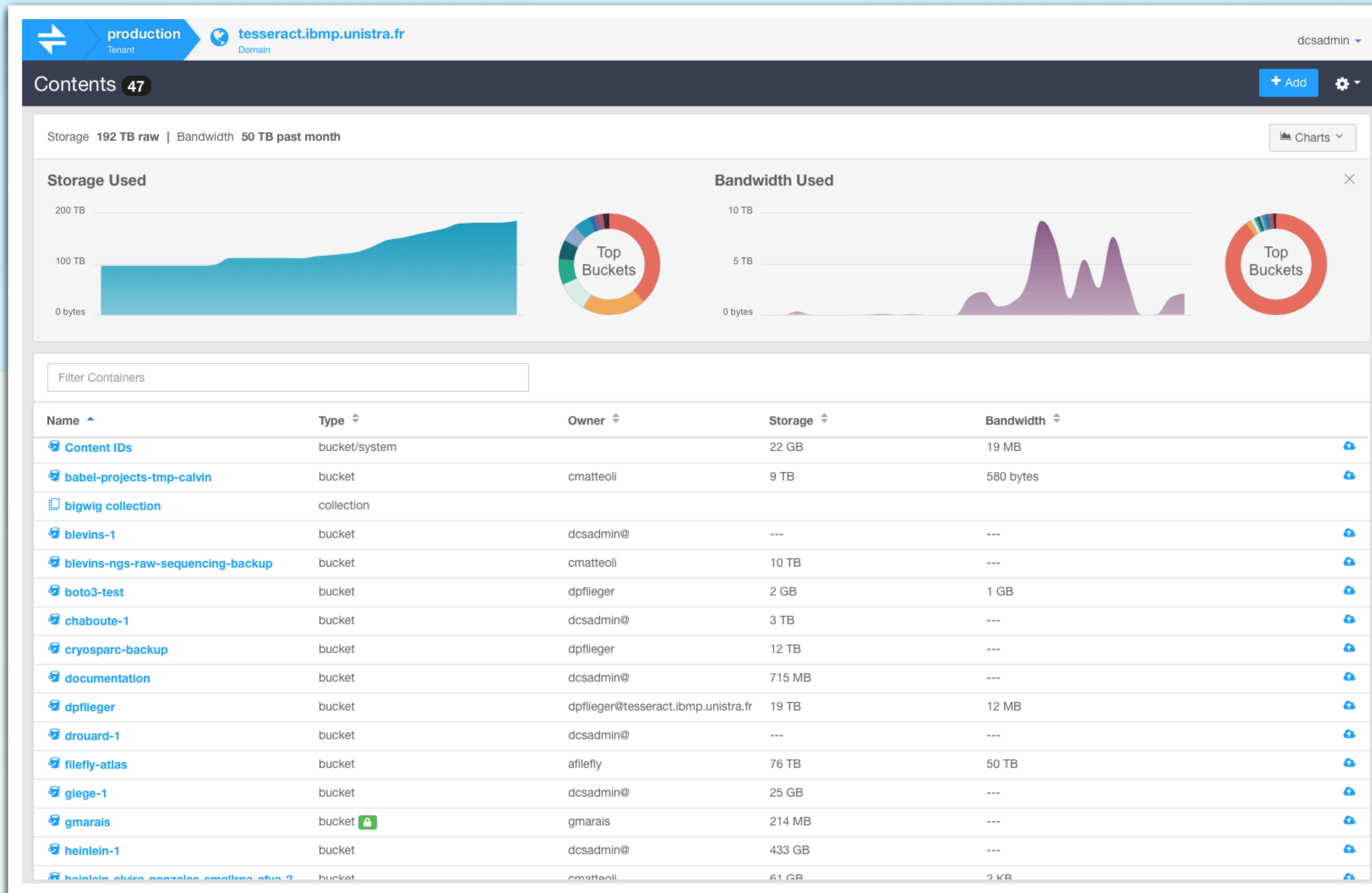
| Cluster | Status |
|-------------|--------|
| Subclusters | 1 |
| Châssis | 7 |
| Disques | 84 |
- Usage Metrics:** A 'SON UTILISATION' (Usage) section includes a 'Espace Disque' (Disk Space) gauge showing 15% usage (194.5 TB used of 1.3 PB, 1.0 PB available, 48.6 TB pinned) and an 'Indice de flux' (Flow Index) gauge showing 1.0 G of 15.0 G.
- Elasticsearch:** A section for 'ELASTICSEARCH' shows the cluster name 'elasticsearchcluster.biocore.lab' in a 'Vert' (Green) state. Below it is a table with the following data:

| Horodatage | Nombre de nœuds | Fragments actifs | Initialisation des fragments | Tons non attribués | Tâches en attente |
|------------|-----------------|------------------|------------------------------|--------------------|-------------------|
| 13:42:03 | 3 | 224 | 0 | 0 | 0 |
- Search Feed:** A 'SEARCH FEED ID 0' section shows a 'swarmfeed' in an 'Actif' (Active) state. It includes a table for 'Événements file d'attente' (Queue Events) with the following data:

| État | Nombre | État | Nombre |
|--------------|--------|--------------|--------|
| En attendant | 0 | En attendant | 0 |
| Traitement | 16 | Traitement | 0 |
| Réessayer | 0 | Réessayer | 0 |

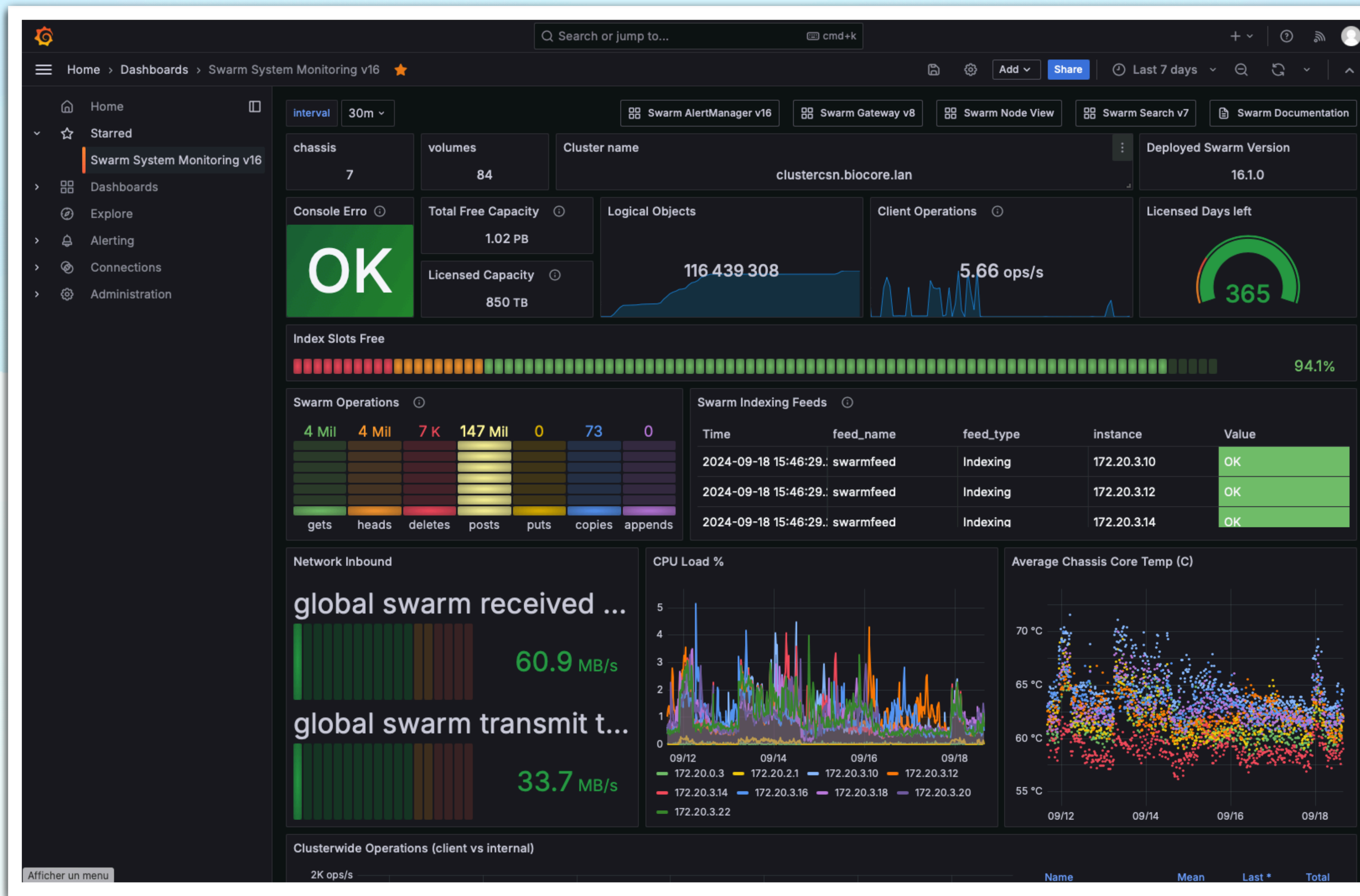
L'objet selon Swarm

Portails WEB



L'objet selon Swarm

Portails WEB



L'objet selon Swarm

Pour les utilisateurs

- Pas forcément "évident" de changer sa façon de faire
- Pour les "gros" utilisateurs :
 - La ligne de commande AWS
- Pour les personnels "lambda" :
 - Mountain Duck
- Idéalement un portail WEB dédié qui reste à faire...
- Et si on automatisait certaines choses ?

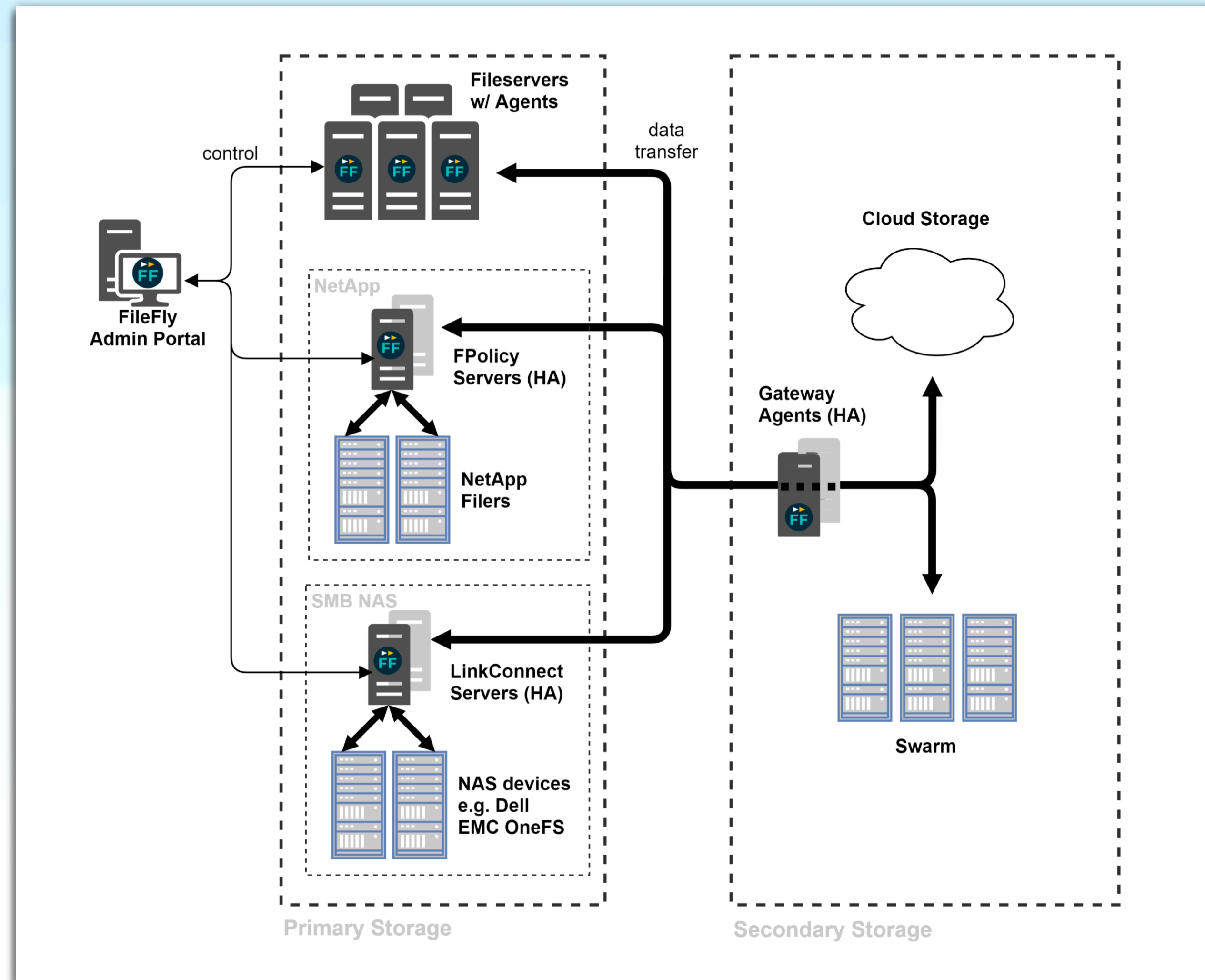
La solution FileFly

Le meilleur des deux mondes ?

- Il est très compliqué de forcer les chercheurs à trier les données chaudes et froides
- Par ailleurs, des données "froides" peuvent redevenir "chaudes"
- **FileFly** permet de déplacer automatiquement des données d'un stockage bloc (SMB) vers un stockage objet selon des critères temporels
- Le déplacement est totalement transparent pour l'utilisateur : les fichiers sont remplacés par des "alias" qui seront automatiquement regarnis s'il sont demandés

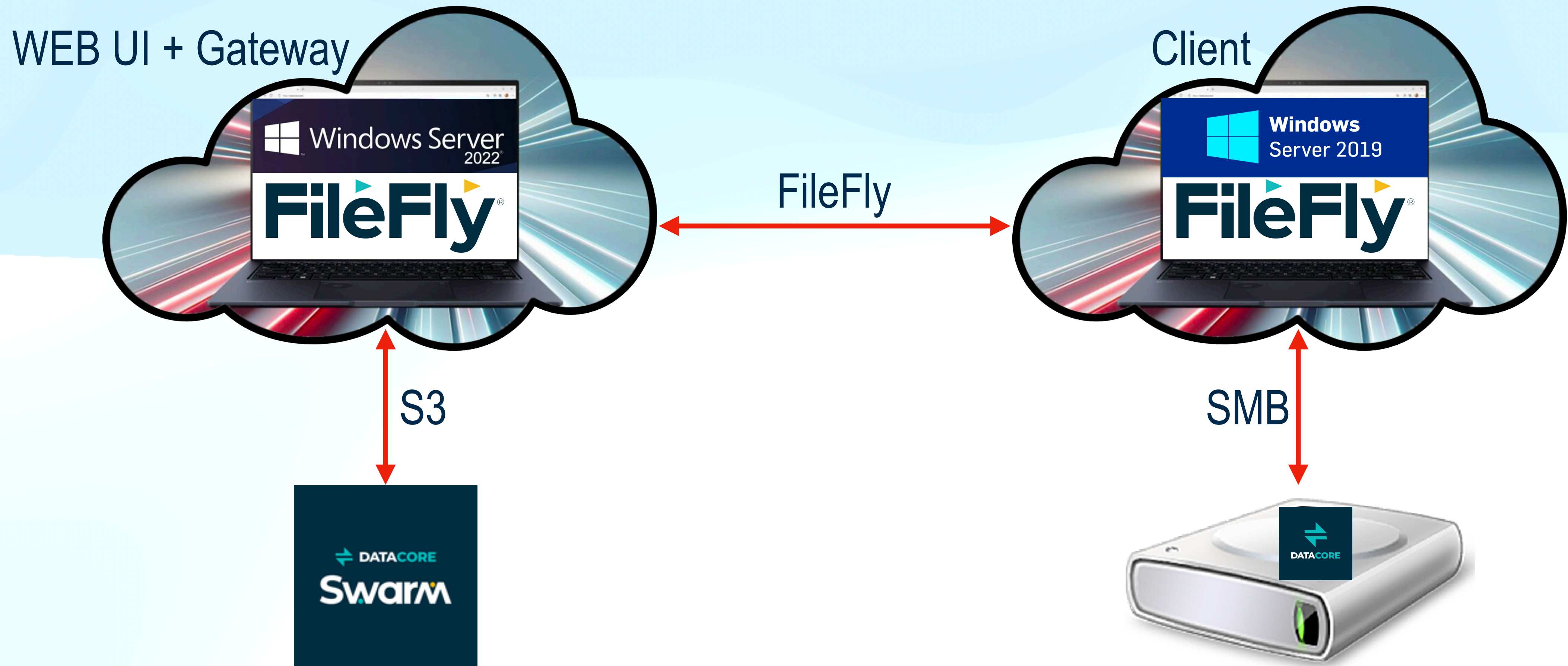
La solution FileFly

Le schéma logique (complet)



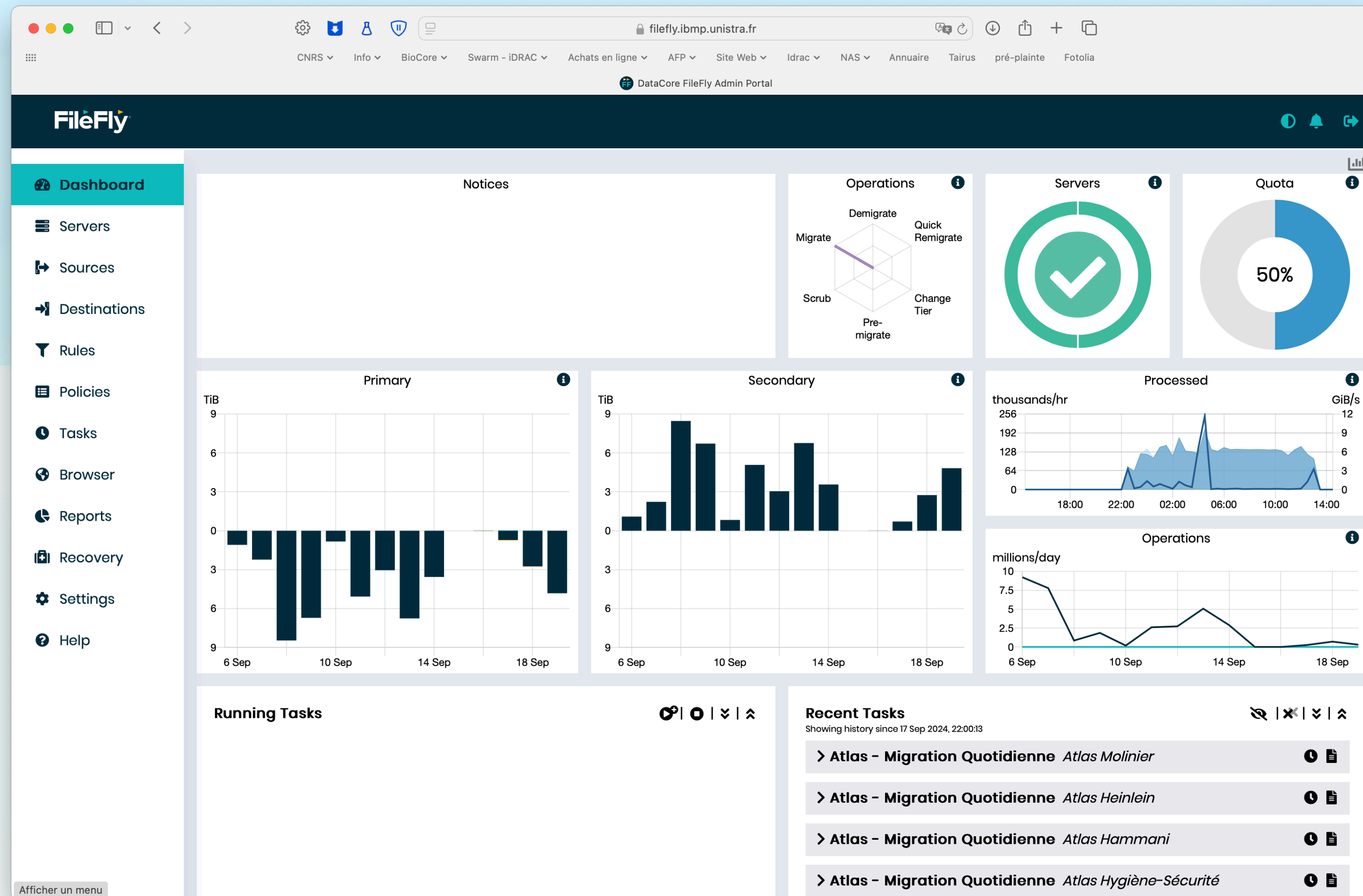
La solution FileFly

L'implémentation simplifiée IBMP



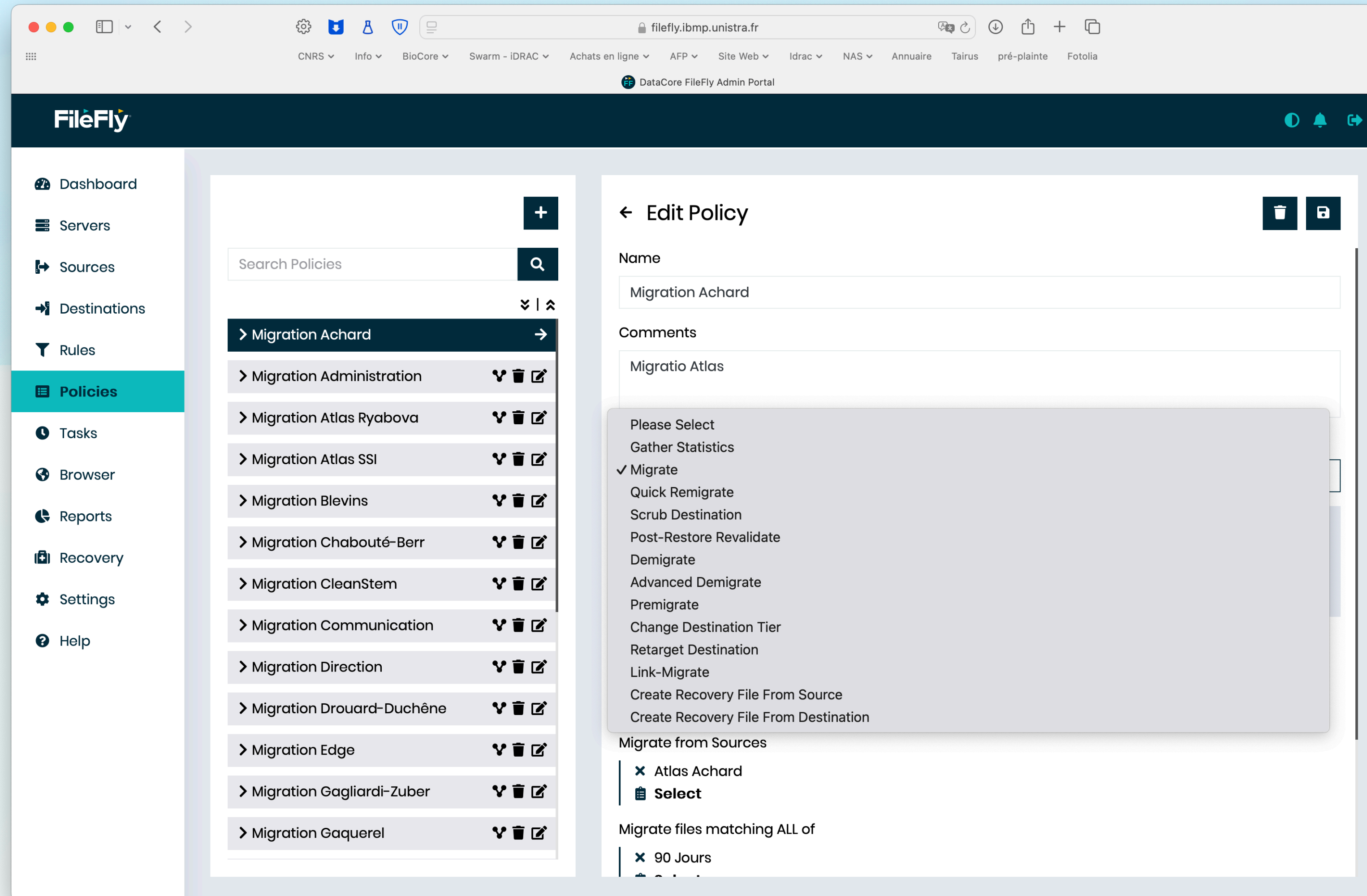
La solution FileFly

Une interface de gestion simple



La solution FileFly

Une interface de gestion simple



Le stockage à l'IBMP

Le bilan...

- Certes un investissement, mais la politique des licences "à vie" est un grand plus face aux velléités des acteurs du marché, et au final pas si cher...
- Des systèmes hyper-fiabiles qui nous ont sauvé la mise face à des pannes matérielles sévères (baie RAID qui part en sucette par exemple)
- Une simplicité au quotidien qui fait que la mobilisation humaine reste minimale : pas besoin d'avoir une personne dédiée aux systèmes
- L'automatisation SMB <-> S3 est intéressante pour "désaturer" le système le plus coûteux
- La démocratisation du stockage objet reste un chantier à conduire...

Des questions ?
...ou une démo ?

