CBP
CENTRE BLAISE PASCAL

HPDA@TheEdge : comment garder la tête hors des nuages ? La station de travail a-t-elle encore un avenir ?

Emmanuel Quemener
IR CBP

UNIVERSITÉ DE LYON

ENS

Où en sommes-nous ?
Un petit inventaire à la Prévert

- Explosion des volumes de données (notamment en biologie)
- *« 90 % des calculs de l'ESR se passent du HPC ! »* (Méso 2016)
- Volonté politique de concentration dans des mésocentres :
  - Des capacités de calcul (dès 2016 avec la labellisation de DataCenters)
  - Des stockages de données (Succès 2017)
- Traitements avec des processus spécifiques complets (*workflows*)
- Évolution souvent difficile des infrastructures internes
- Puissance de calcul brute dans les GPU
- De moins en moins de RH pour l'informatique interne des unités.

Où allons-nous ?
Prendre un certain contre-pied...

- Rapprocher les ressources de traitement de la production
- Adapter les systèmes à la véritable nature des traitements
- Assurer la reproductibilité des traitements
- Rationaliser l'exploitation de ressources locales
- Limiter le temps d'administration système de postes individuels
- Limiter l'empreinte énergétique
- Limiter les délais liés au transfert réseau
- Simplifier l'accès à l'archivage
- Recentrer les informaticiens sur des fonctionnalités spécifiques.

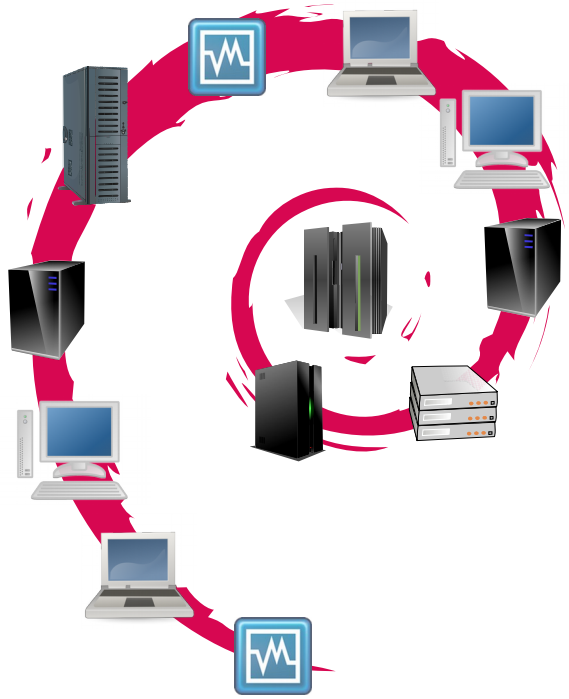
Comment y allons nous ?
En exploitant l'existant, et plus...

- HPDA@TheEdge : traitement efficace à la périphérie
  - *High Performance Data Analysis at the Edge*
- Exploiter **SIDUS** sur **COMOD** :
  - *Single Instance Distributing Universal System*
  - *Compute On My Own Device*
- Déplacer le stockage à la périphérie
- Utiliser la virtualisation disponible depuis 10 ans
  - Et peut-être des versions plus efficaces pour des *Workflows*

Sur les Machines du CBP : SIDUS

Je n'installe pas, je démarre !

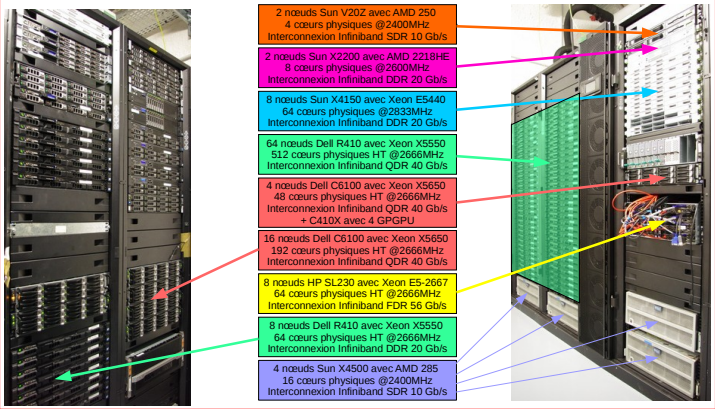
- Quoi ?
  - Déployer un système simplement sur un parc de machines
- Pourquoi ?
  - Assurer l'unicité des configurations
  - Limiter l'empreinte du système sur les disques
- Pour qui ?
  - Étudiants, enseignants, chercheurs, ingénieurs, ...
- Quand & Où ?
  - Centre Blaise Pascal : depuis 2010, près de 200 machines
  - PSMN : depuis 2011, plus de ~500 nœuds (sa propre instance)
  - Laboratoires : Chimie, UMPA, LBMC, IGFL, ISA, CRAL
- Comment ?
  - Utiliser un partage en réseau d'une arborescence
  - Détourner une ruse de LiveCD



« Deux machines ayant démarré SIDUS ne peuvent pas ne pas avoir le même système ! »

Au CBP près de 200 machines
3 SIDUS sur 4 plateaux techniques

Plateau multi-nœuds : 9 grappes
116 nœuds, 4 vitesses réseaux



Plateau myriALUs
Multi-shaders : 72 types de (GP)GPU différents
Accélérateur : 1 Xeon Phi Intel

GPU Gamer : 18

- Nvidia GTX 560 Ti
- Nvidia GTX 680
- Nvidia GTX 690
- Nvidia GTX Titan
- Nvidia GTX 780
- Nvidia GTX 780 Ti
- Nvidia GTX 750
- Nvidia GTX 750 Ti
- Nvidia GTX 960
- Nvidia GTX 970
- Nvidia GTX 980
- Nvidia GTX 980 Ti
- Nvidia GTX 1050 Ti
- Nvidia GTX 1060
- Nvidia GTX 1070
- Nvidia GTX 1080
- Nvidia GTX 1080 Ti
- Nvidia RTX 2080 Ti

GPGPU : 9

- Nvidia Tesla C1060
- Nvidia Tesla M2050
- Nvidia Tesla M2070
- Nvidia Tesla M2090
- Nvidia Tesla K20m
- Nvidia Tesla K40c
- Nvidia Tesla K40m
- Nvidia Tesla K80
- Nvidia Tesla P100

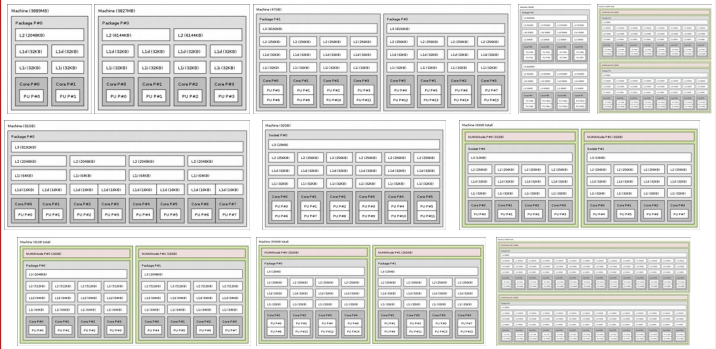
GPU desktop & pro : 27

- NVS 290
- Nvidia FX 4800
- NVS 310
- NVS 315
- Nvidia Quadro 600
- Nvidia Quadro 2000
- Nvidia Quadro 4000
- Nvidia Quadro K2000
- Nvidia Quadro K4000
- Nvidia Quadro K420
- Nvidia Quadro P600
- Nvidia 8400 GS
- Nvidia 8500 GT
- Nvidia 8800 GT
- Nvidia 9500 GT
- Nvidia GT 220
- Nvidia GT 320
- Nvidia GT 430
- Nvidia GT 620
- Nvidia GT 640
- Nvidia GT 710
- Nvidia GT 730
- Nvidia GT 1030
- Nvidia Quadro 2000M
- Nvidia Quadro K4000M
- Nvidia Quadro M2200
- Nvidia MX150

GPU AMD : 18

- HD 4350
- HD 4890
- HD 5850
- HD 5870
- HD 6450
- HD 6670
- Fusion E2-1800 GPU
- HD 7970
- FirePro V5900
- FirePro W5000
- Kaveri A10-7850K GPU
- R7 240
- R9 290
- R9 295X2
- Nano Fury
- R9 Fury
- R9 380
- RX Vega64

Plateau multi-cœurs : petit bestiaire
42 types de CPU différents



Plateau 3IP (prononcez “Trip”)

“Introduction Inductive à l’Informatique et au Parallélisme”

Computhèque

Atelier

- Diagnostics
- Désassemblage
- Tests unitaires
- (Re)Qualification
- Récupération supports

Refuge

- Machines “ouvertes”
- Machines “exotiques”
- Composants obsolètes

Salle de formation

- Ateliers 3IP
- Fête de la science

Depuis 2012, l'approche COMOD

- Vous connaissez le BYOD : Bring Your Own Device
  - Travailler avec son équipement personnel
- COMOD se propose :
  - Disposer en quelques secondes d'un environnement fonctionnel
  - D'exploiter sa machine pour son travail dans des tâches de HPC
  - De se déployer sur une machine complète ou virtuelle (VirtualBox)
- COMOD s'appuie :
  - Sur l'approche SIDUS
  - L'infrastructure de stockage du CBP, mais pas seulement !
  - L'authentification de l'école : le même identifiant, mot de passe

COMOD : et ça fonctionne ?

Laboratoire de chimie

LABORATOIRE DE CHIMIE ENS DE LYON

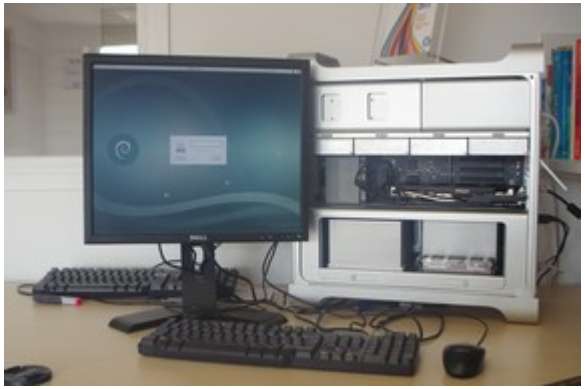
Laboratoire de mathématiques

UMPA ENS DE LYON

Laboratoire d'astrophysique

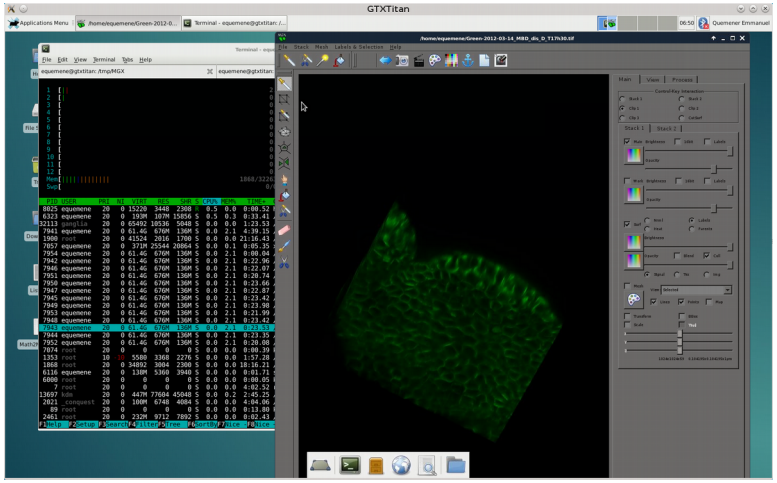
CRAL

&

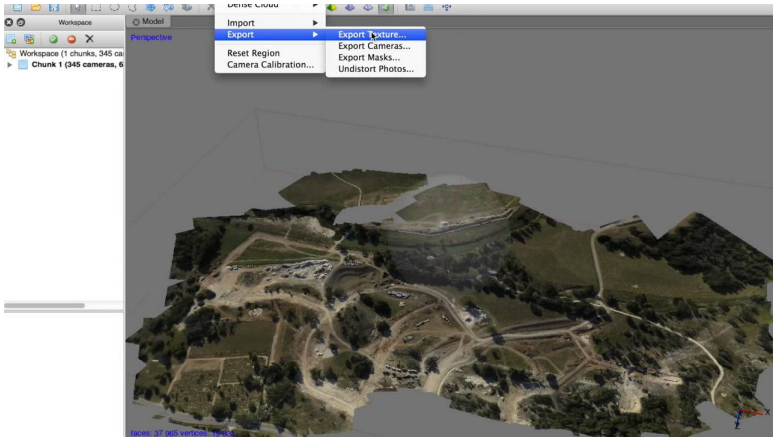


Accès distant pour les traitements
Le couple x2go/VirtualGL

MorphographX



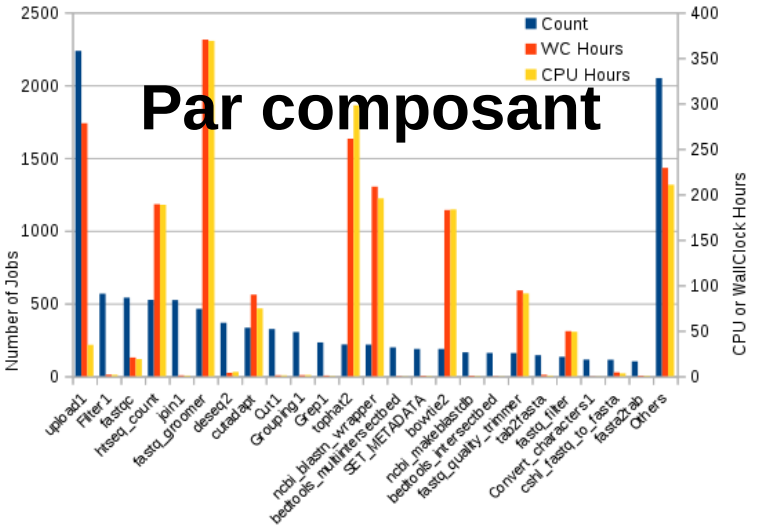
PhotoScan

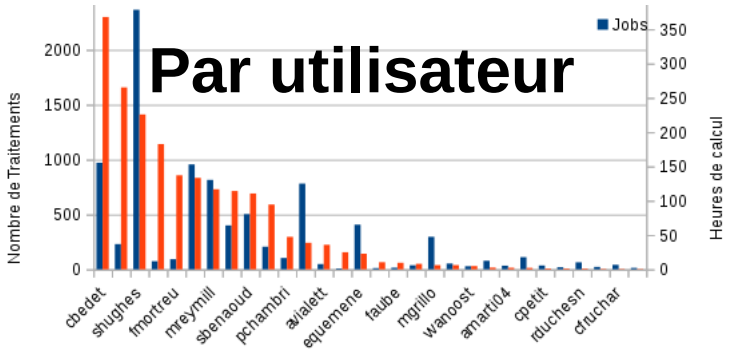


Exploitation massive de CUDA sur des semaines

Études du centre d'essais du CBP
La plateforme Galaxy

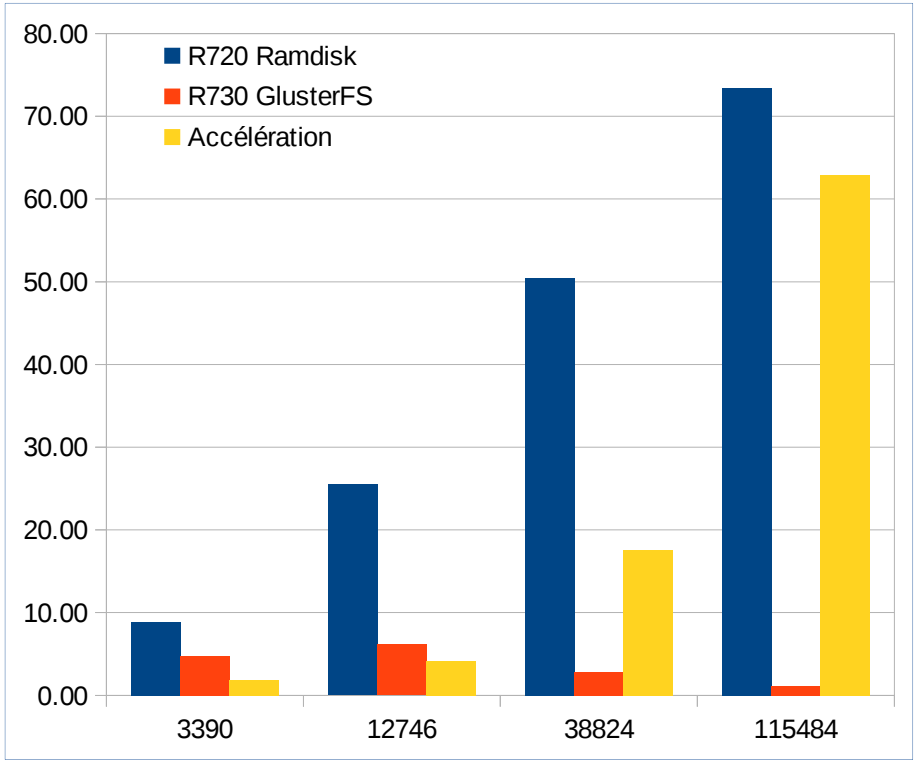
- L'objectif : intégration d'un portail Galaxy au centre de calcul
- Le contexte :
  - explosion du volume de données des biologistes
  - appropriation du centre de calcul difficile
  - standard émergent de portail de traitement
- Le banc d'essai : entre réel & virtuel
  - 1 machine virtuelle pour le portail
  - 1 cluster de 8 nœuds pour les traitements
- Les étapes :
  - version 1.0 2015Q2
  - version 2.0 2017Q3 : IB, volume, PostGreSQL
- Les conclusions :
  - développement en cycle court parfait
  - usages très différents entre laboratoires : pas de panacée...
  - intégration sur infrastructure "centre de calcul" difficile
  - **métrologie fine indispensable pour évaluation des besoins**





Études du centre d'essais du CBP
Les traitements Repeat*

- L'objectif : quelle infrastructure pour ce traitement
- Le contexte :
  - *workflow* trop lourd sur une « petite » station de travail
  - portage sur PSMN entraînant une coupure du service
  - investigations poussées sur équipements du CBP
- Le banc d'essai : plates-formes différentes
  - déploiement sur station de travail avec SSD
  - déploiement sur nœud de calcul + GlusterFS
  - déploiement sur serveur avec Ramdisk
- Les conclusions : une infrastructure nécessaire
  - des traitements générant des millions de petits fichiers
  - un SSD "mort" à la fin de la première campagne
  - un GlusterFS inadapté (comme tout système partagé)
  - **une machine à dédier pour le traitement**



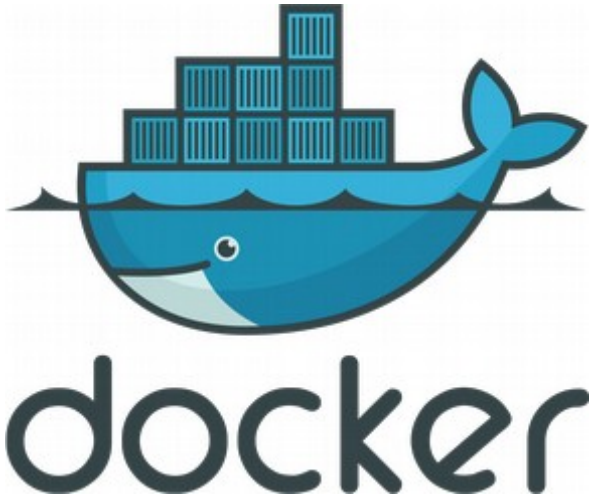
Des études du centre d'essais Émergence de solutions

- L'usage n'est pas le traitement : la donnée compte
  - A chaque usage son « système » optimal : métrologie
- Une « distance » entre expérience & traitement à réduire
  - Solution par la relocalisation des processus de traitement
- La nécessité de « boîtes noires » de *Workflows*
  - Simplification d'accès à l'exécution
  - Solution par les conteneurs
- Le stockage avec archivage intégré
  - ZFS comme socle de stockage local (mdadm+lvm+...)

Pourquoi la station de travail ? Séparer pour mieux traiter...

- PUE : 1.00 par définition : *done...*
- Session graphique distante : *done...*
- Machines virtuelles & conteneurs : *done...*
- Sécurité & ses piliers :
  - Disponibilité : celle de SIDUS
  - Intégrité : *« une seule racine pour les gouverner tous »*
  - Confidentialité : accès par liste, configuration à la volée de groupe
  - Traçabilité : « nature » SIDUS & syslog distant
- Stockage local : fin 2018, 42TB nets pour moins de 2k€

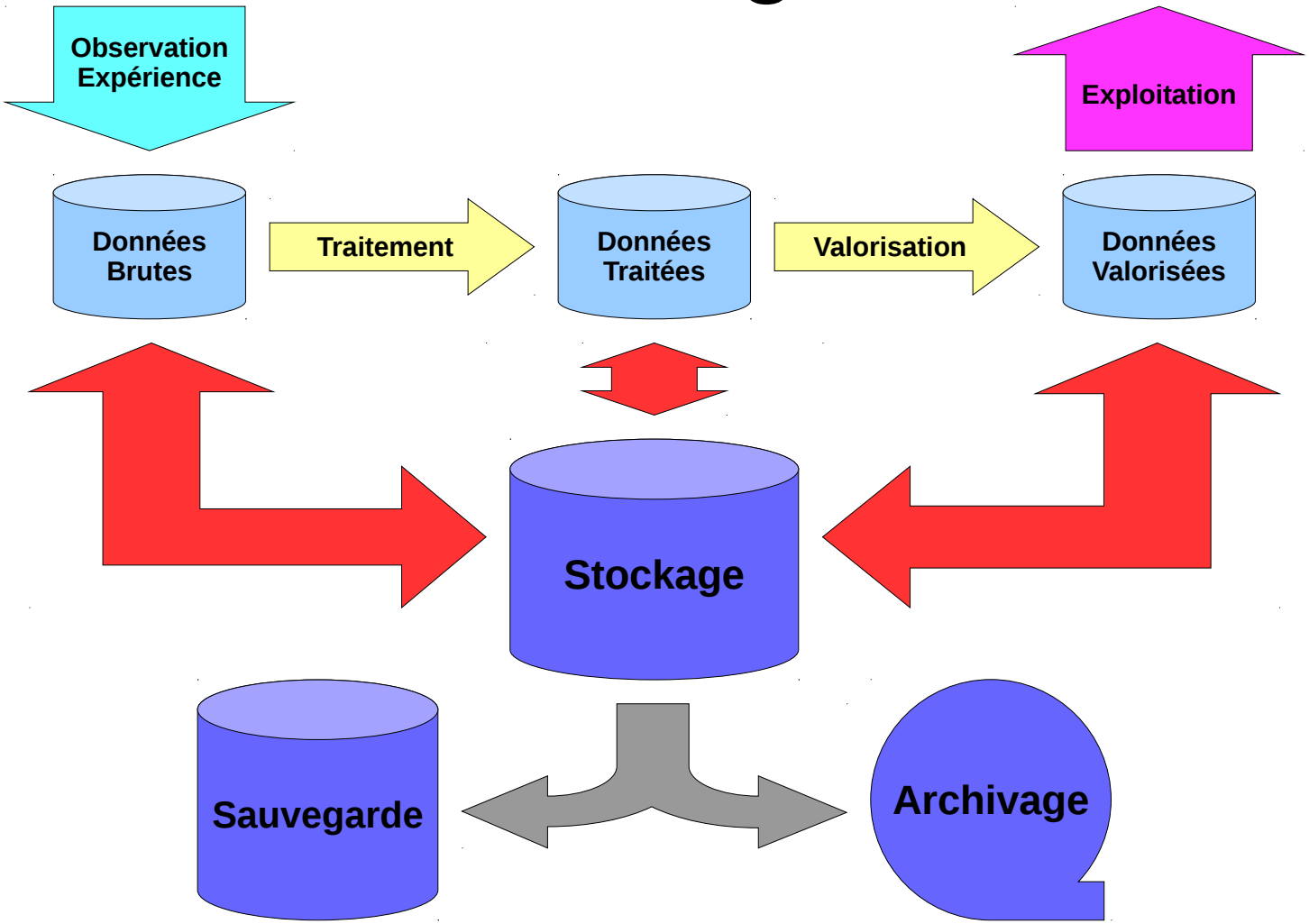
La virtualisation : les solutions VirtualBox & Docker



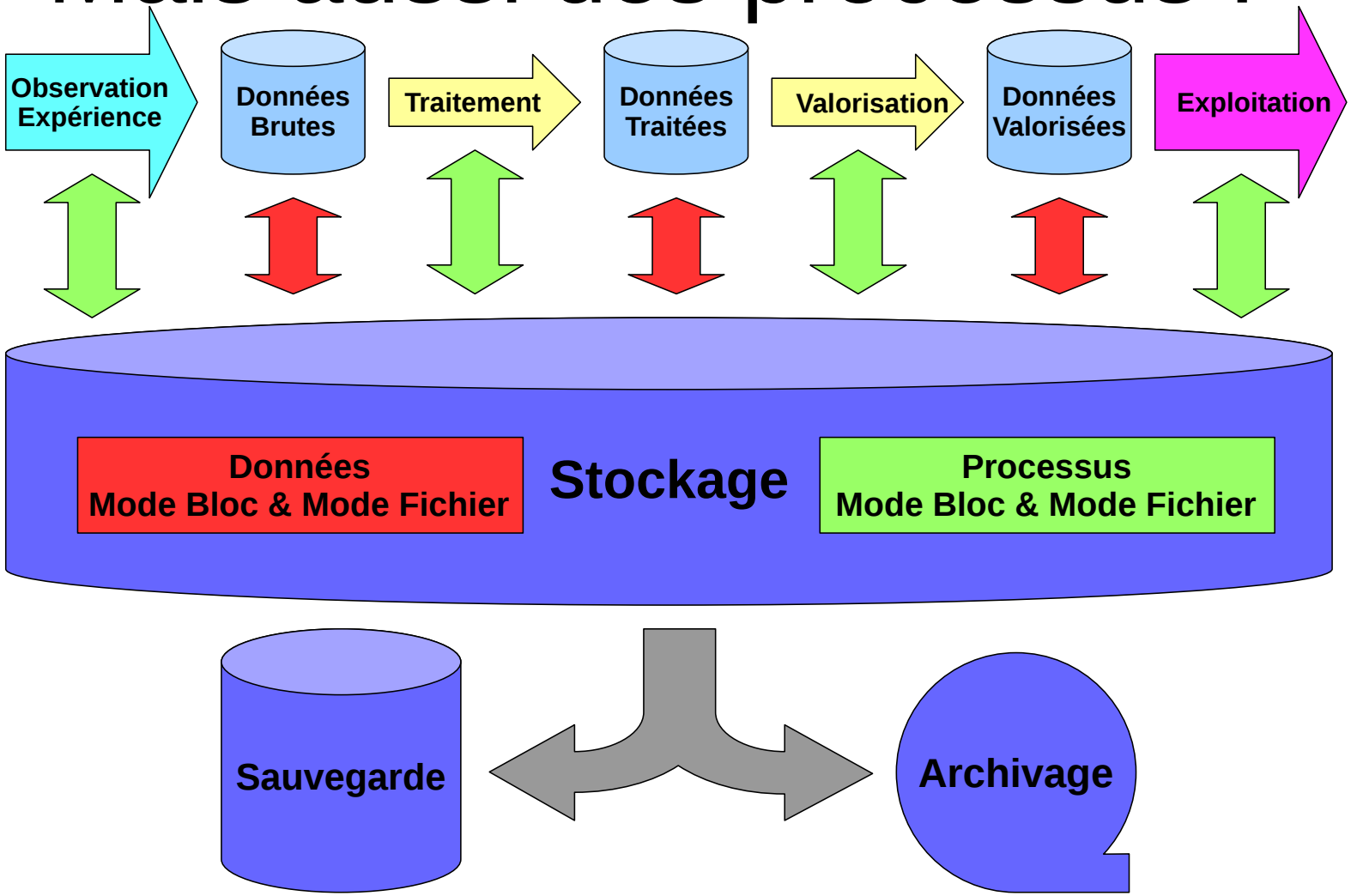
- Précautions :
  - PAS dans son $HOME
  - Réservations de ressources
  - Partages de dossiers

- Précautions :
  - d'ABORD connaître GNU/Linux
  - Pas de persistance implicite
  - Partages de volumes

Une étude à l'ENS-Lyon en 2010
Besoins de stockage des labos...



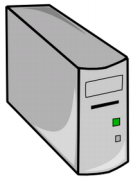
Le cycle de la donnée…
Mais aussi des processus !



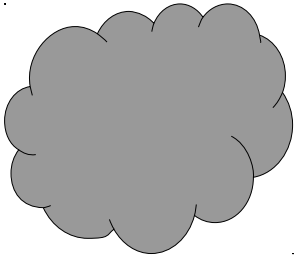
Comparons 4 approches de HPDA (sans insérer la sécurité DICT)

La station basique

+ Ressources dédiées
+ Traitement « localisé »
- Administration atomique
- Scalabilité
- Criticité



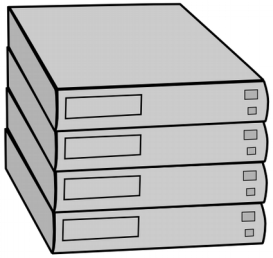
Le « cloud »



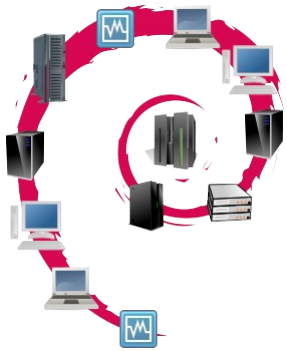
+/- Personnalisation
+ Scalabilité
- Ressource matérielle partagée
- Extensibilité des volumes
- Traitement délocalisé (transfert)

Le centre de Calcul

+ Environnement unifié
+ Scalabilité
+ Ressources dédiées au job
+/- Dossiers distribués
- Appropriation & intégration



La station SIDUS



+ Administration centralisée
+ Personnalisation persistante
+ Traitement localisé
+ Stockage local polyvalent
- SIDUS à disposition...

Cas d’usage : séquenceur Illumina
Du process aux approches

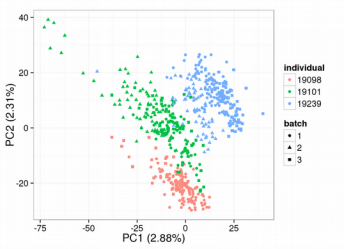
Transfert

Données Brutes

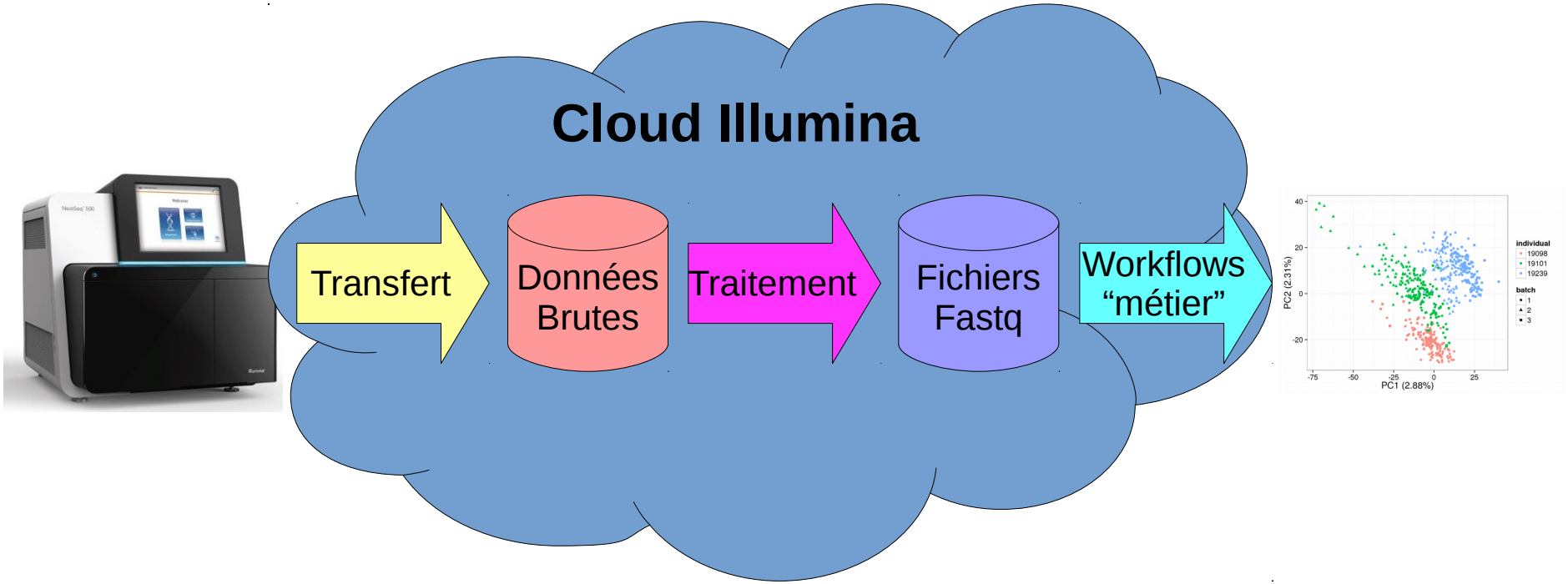
Traitement

Fichiers Fastq

Workflows “métier”

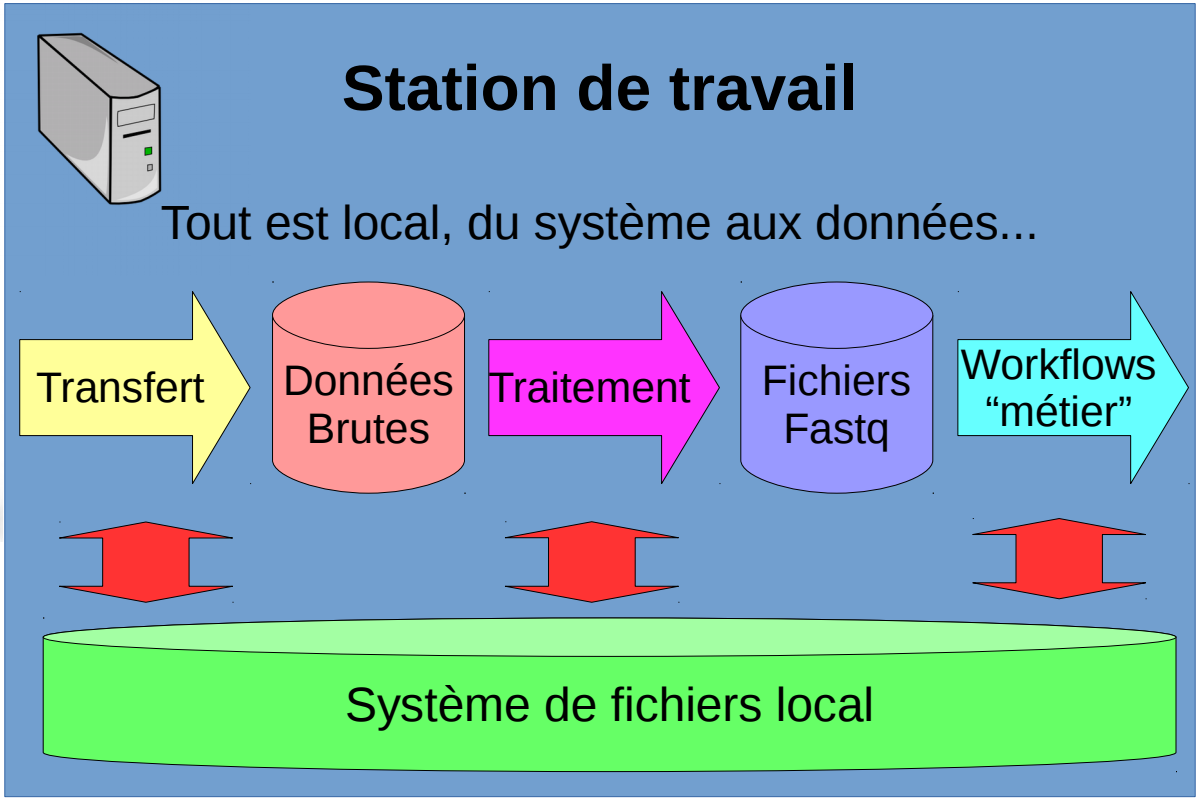


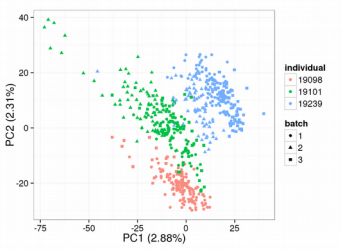
Cas d'usage : séquenceur Illumina
Approche « cloud » Illumina



- Externalisation complète : la première dose est offerte...

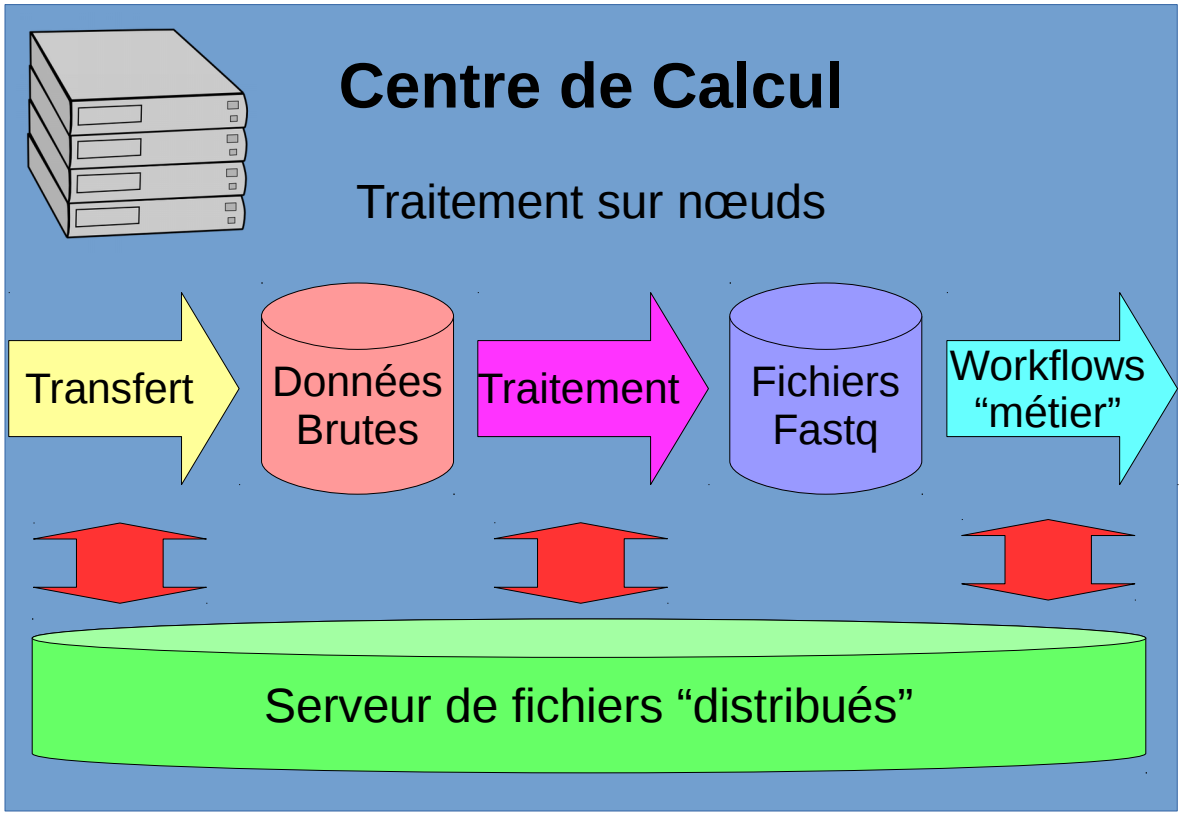
Cas d'usage : séquenceur Illumina
Du process aux approches

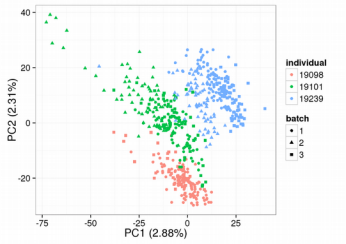




- Même avec le stockage le *SPOF*, c'est le système...

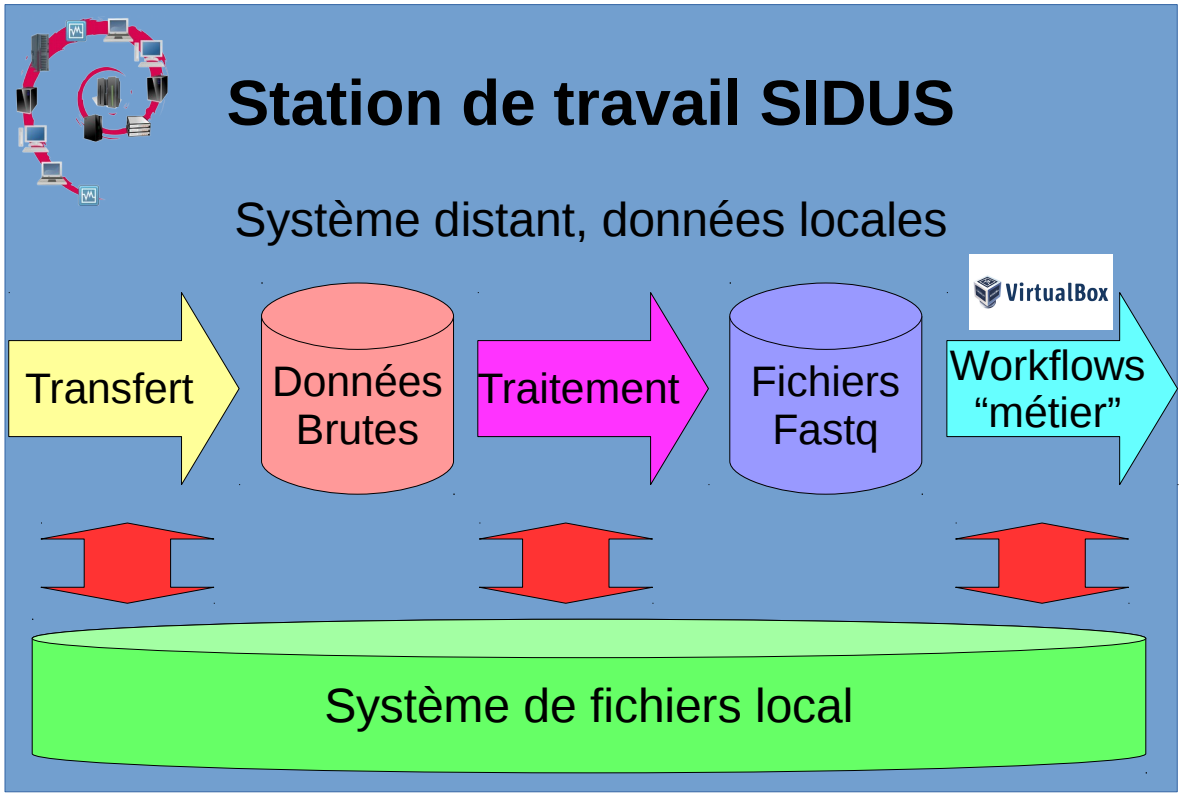
Cas d'usage : séquenceur Illumina
Approche « Centre de Calcul »

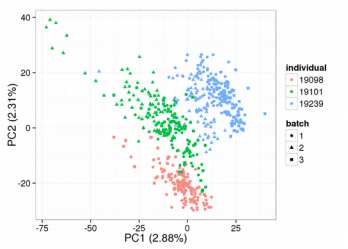




- Transfert direct difficile (protocole CIFS)

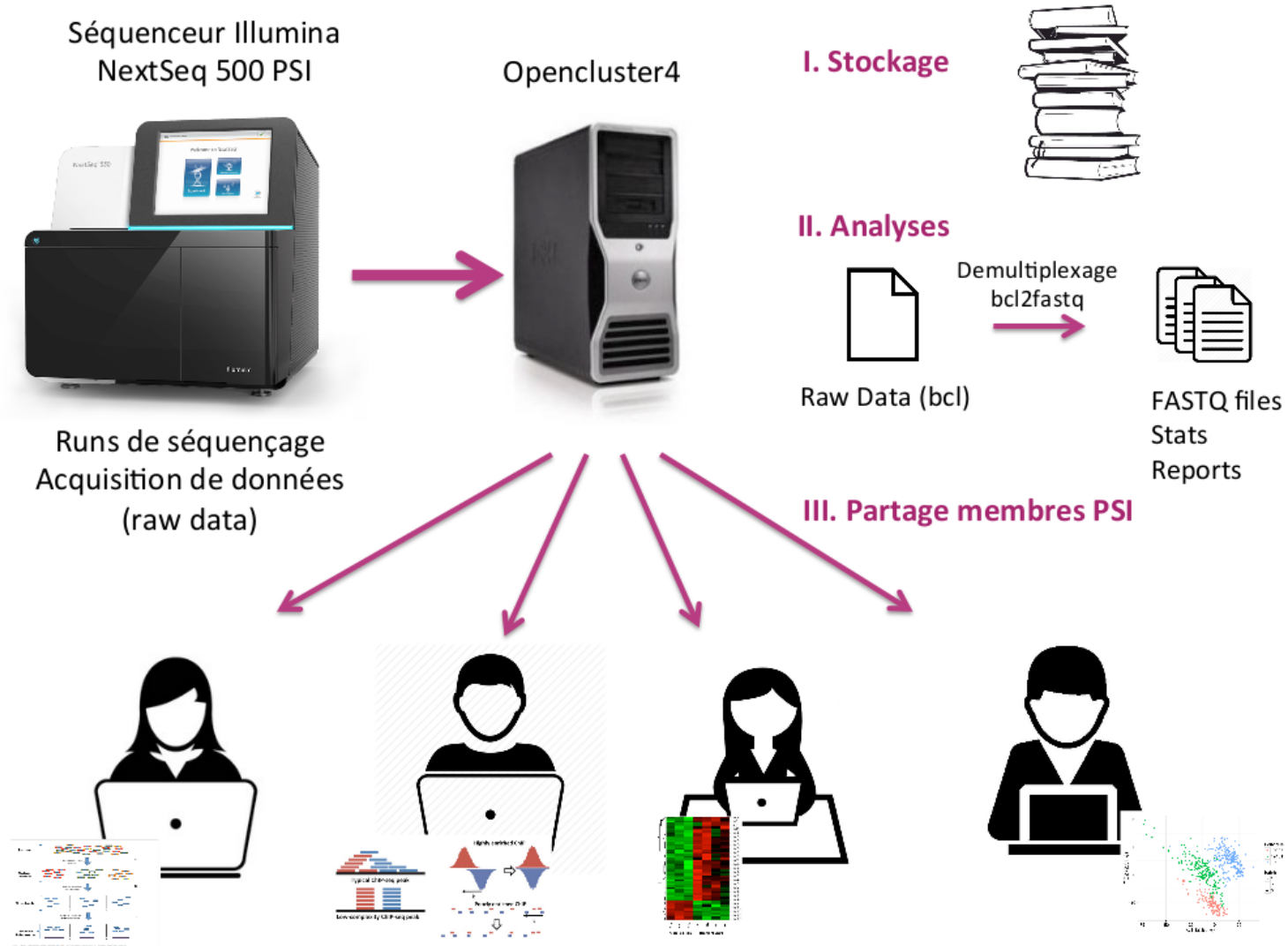
Cas d'usage : séquenceur Illumina
Approche « HPDA@TheEdge »





- ***WIP*** : pipeline de Illumina uniquement sous Windows 10

HPDA@TheEdge Vu par le labo



Cas d'usage : séquenceur Illumina
Le match (à 4h du matin...)

	Cluster CBP			HPDA@The Edge	Speed Up
	Transfert	**Traitement**	**Total**	**Traitement**	
User time	0.19	2332.2	2332.39	1365.38	1.71
System time	12.18	54.06	66.24	31.32	2.11
Elapsed time	**160.06**	**186.06**	**346.12**	**96.9**	**3.57**
% CPU this job got	7.00%	1282.00%		1441.00%	
File inputs	16482448	16165880	32648328	18431667	1.77
File outputs	0	0	0	17509833	0.00

- Meilleur cas possible : infrastructure cluster CBP « vide »
- Autour de 4x plus rapide (mais processeur HPDA x2.5)

En conclusion

« *Refaites vos choix !* » (L'éveil d'Endymion, Simmons)

- HPDA@TheEdge avec SIDUS :
  - Beaucoup d'avantages, peu d'inconvénients
- La virtualisation : en cas de besoin, sécurisation minimale
  - Traitement efficace avec des outils déjà intégrés (conteneurs)
- Le stockage local :
  - Stockage local tremplin vers l'archivage : *Data Management Plan* à la station
- L'accès distant : x2go/VirtualGL
  - Une occasion d'anticiper le traitement sur GPU et voir son évolution
- *Bref, essayez, et jugez ensuite (comme le PSMN ;-)) !*