Fermes d'analyse basées sur PROOF

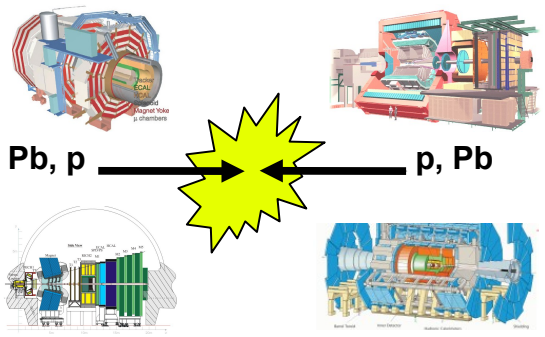
G. Ganis, CERN, PH-SFT

Réunion LCG-France
Annecy, 19 Mai 2009

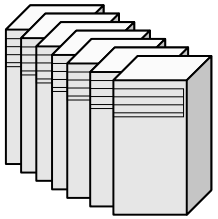
Plan

- Donnée du problème
- PROOF: un rappel
- Installations existantes
- Problématiques adressées
  - Performance
  - Gestion des Datasets
  - Gestion des ressources (scheduling)
  - Intégration avec des systèmes batch

Les données à LHC



~100 Hz
1÷12.5 MB
10 PB / y



MonteCarlo
Production
20÷100% / data

RAW

Reconstruction

ESD

Experiment Reduction

AOD

Individual / Physics Group
Selection / Reduction

DPD

Event Summary Data
0.025÷ 2.5 MB/event
100÷1000 TB / y

Analysis Objects Data
4÷ 250 kB/event
30÷200 TB / y

Derived Physics Data
1÷10 TB / y

Formats utilisés pour l'analyse
au centres Tier 3 / Tier 2

Activités d'Analyse

- Taches interactives: ordinateur de bureau / portable
  - Regarder les outputs, faire des fits simples, visualiser les résultats
- Taches demandant beaucoup d' I/O: lecture des données
  - O(1÷10 TB) données lus effectivement
  - ~10h @ 150 ÷ 250 MB/s (vitesse de lecture typique)
  - ~1h @ ~2 GB/s (10 serveurs ... ou matériel sophistiqué)
- Taches demandant beaucoup de CPU:
  - Simulations “privees” {Full, Fast}
  - Études des systématiques avec des techniques Monte Carlo
  - Fits sophistiqués
  - ...

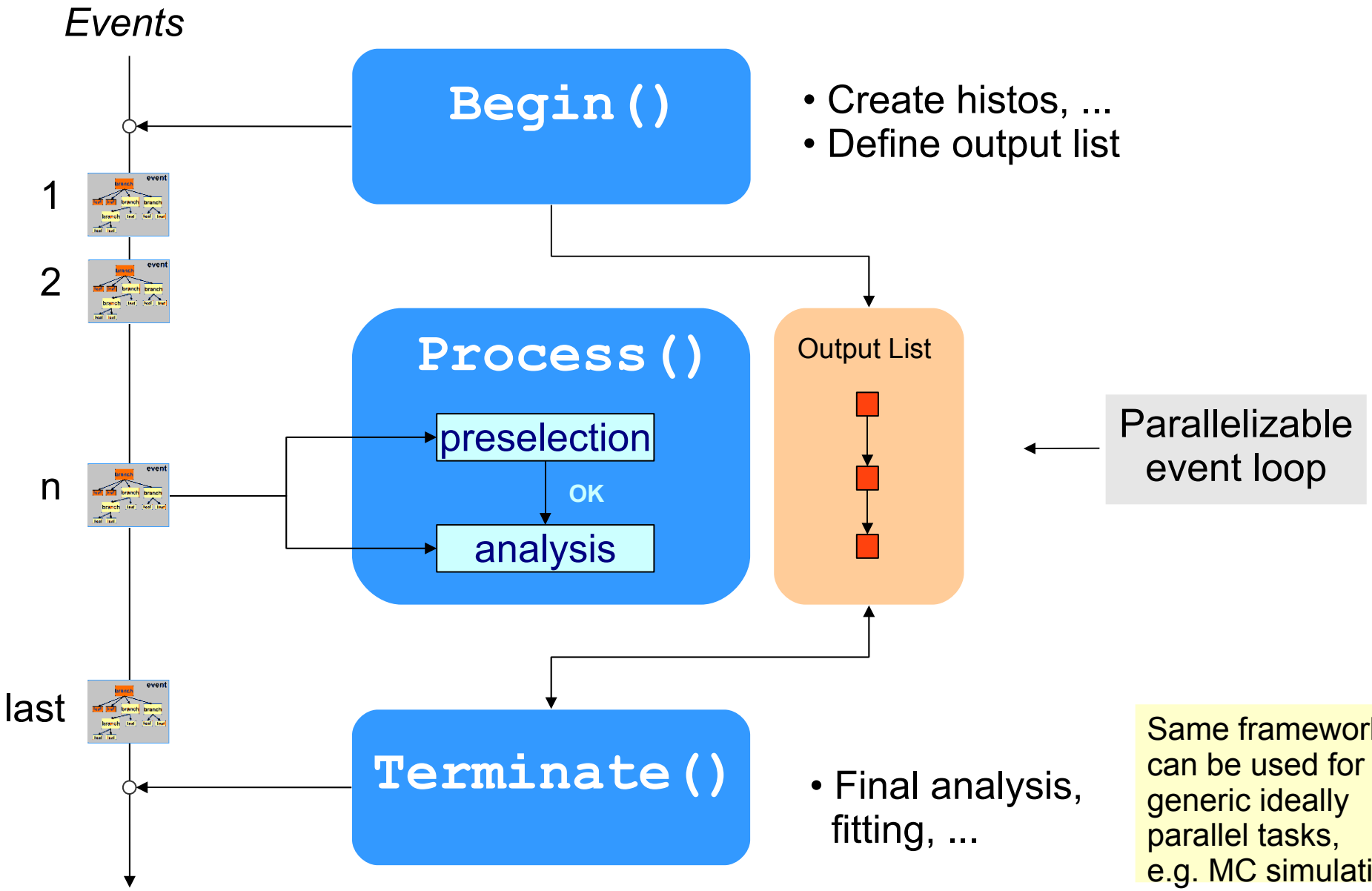
Typiquement il s'agit de taches parellèles triviales: il suffit de diviser en sous-taches pour obtenir l'accélération ideale

PROOF – Parallel ROOT Facility

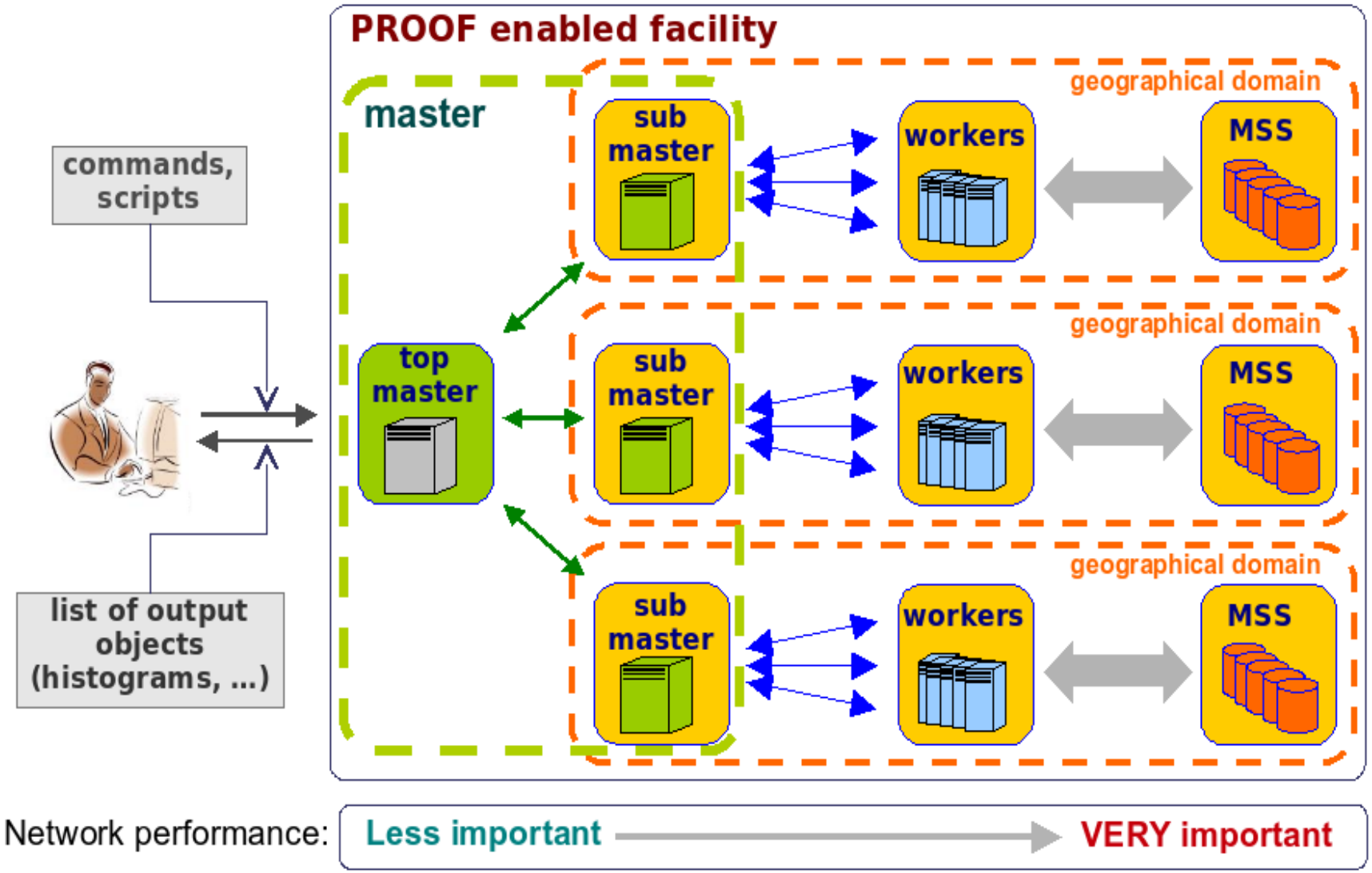
- Coordination du travail de sessions ROOT distribués
  - Architecture pull: répartition de la charge de travail obtenue dynamiquement à niveau d'événement
  - Scalabilité: composante seriale typiquement petite
  - Possibilité d'examiner en temps réel l'état des outputs
  - Transparence: extension de la (ROOT) shell locale
- Parallélisme à processus multiples
  - S'adapte facilement à une ample varieté des cas
- Exploitation de la localité des données
  - Minimisation des transfers
- Utilise XROOTD pour l'infrastructure de communication

Conçu pour l'exécution interactive de taches trivialement parallèles au centres Tier 2 / 3 et desktops many-core

Parallélisme à niveau d'événement: cadre TSelector



Architecture PROOF



Utiliser PROOF pour faire quoi?

Exécution interactive des taches parallèlisme trivial

- Traitement des données
  - Analyse finale
  - Calibration, premières réconstructions sur un échantillon des données
    - Une réponse rapide est un plus pour les tests de qualité des données
- Taches qui nécessite de beaucoup de CPU
  - Simulations {Full, Fast, Toy} Monte Carlo
  - Boucles dans les fits
  - ...

Installation de PROOF

- PROOF fait partie de ROOT
  - Aucun logiciel additionnel necessaire
- Le service PROOF tourne comme plug-in de XROOTD
  - Le meme XROOTD peut etre utilise pour les fichiers et pour PROOF
    - La porte 1094 pour les fichiers, la porte 1093 pour PROOF
- Fichiers des configuration
  - Partie du fichier de configuration XROOTD
    - Peut etre le meme pour toutes les machines
  - Fichier pour les roles des machines (proof.conf)
    - Bientot obsolete
  - Fichier pour definir les groups d'utilisateurs et leur propriétés
    - Prioritées, quotas, ...

Fermes PROOF ALICE

- CAF: CERN Analysis Facility
  - 112 cores, 35 TB (5 HDD/node en RAID5)
  - Analyse rapide d'échantillons selectionnés, calibration, alignment, réconstruction, simulation rapide
  - 5-10 utilisateurs simultanés (~100 enregistrés)
- GSIAF: GSI, Darmstad
  - 160 cores, 150 TB sous Lustre, 10 GB/s Ethernet to Lustre
  - Analyse de données, calibration TPC
  - 5-10 utilisateurs simultanés
  - Exemple de performance pour la calibration:
    - 1.4 TB processés in 20 min
- Autre fermes de test existe: JINR, Turin, ...

Fermes PROOF ATLAS

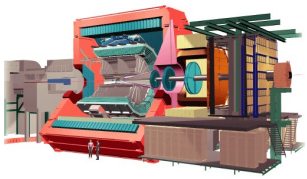
- University of Wisconsin, Madison
  - 200 cores, 100 TB, RAID 5
  - Analyse des donnés (recherche de Higgs)
  - Tests de I/O avec (multi-)RAID, integration ATLAS DM
  - Integration PROOF-Condor, prototypes de fermes d'analyse
  - ~20 utilisateurs enregistrés
- Brookhaven National Lab
  - Prod.: 40 cores, 20 TB HDD; Test: 72 cores, 25 TB HDD, 192 GB SSD
    - Analyse de données, tests d'I/O avec SSD, RAID
  - ~25 utilisateurs enregistrés
- Munich LMU/LRZ
  - 10 AMD Dual CPU, dual core, 2.7 GHz, 8GB RAM
  - Analyse des donnés, tests d'I/O et de scalabilité
- Autre fermes de test: Madrid, UT Arlington, Duke University

Autre installations de PROOF

- NAF: National Analysis Facility at DESY
  - ~900 cores partagés avec du batch gerés par SGE
  - ~80 TB Lustre, dCache, Infiniband to Lustre
  - Analyse des données de ATLAS, CMS, LHCb et ILC
    - PROOF testé par des groups CMS
  - ~300 utilisateurs enregistrés
- CC-IN2P3, Lyon (talk Y. Calas)
  - 160 cores, 17 TB HDD
  - Analyse de données pour LHC
- Purdue University, West Lafayette, USA
  - 24 cores, dCache storage
  - CMS Muon reconstruction
- ...

Schema du ALICE CAF

Courtesy of J.F Grosse-Oetringhaus, CERN



Experiment

Disk Buffer

Tier-1 data export

Tape storage

AliEn SE

CASTOR MSS

Staging on request of physics working groups or single users

CAF computing cluster

Proof master, xrootd redirector

Proof worker | local disk

Proof worker | local disk

Proof worker | local disk

...

Monitorage MonALISA à la CAF

Courtesy of M. Meoni

- A full CAF status table is available on the MonALISA repository

	Machine status					Storage	CPU		Memory		Networking		Hosted files	
Machine	Online	xrootd	cmsd	Load	Proof users	Staging (%)	usr	sys	Used	Free	IN	OUT	Files	Size
1. lxfsrd0506				0.15	3	4	0.848	0.407	2.453 GB	13.21 GB	6.322 Kbps	9.027 Kbps	-	-
2. lxfsrd0507				2.05	4	11	13.23	2.21	1.88 GB	13.78 GB	315.3 Kbps	7.259 Mbps	4285	241.9 GB
3. lxfsrd0508				1.05	4	10	21.13	1.097	2.205 GB	13.46 GB	2.823 Mbps	159.5 Kbps	4141	231.6 GB
4. lxfsrd0509				1.75	4	11	13.06	2.25	2.17 GB	13.49 GB	309 Kbps	6.858 Mbps	4353	242.6 GB
5. lxfsrd0510				2.92	6	10	24.47	1.526	2.239 GB	13.42 GB	1.417 Mbps	2.701 Mbps	4244	232.8 GB
6. lxfsrd0513				3.12	6	11	25.62	1.984	2.219 GB	13.44 GB	968.2 Kbps	2.299 Mbps	4208	233 GB
7. lxfsrd0514				2.05	4	11	19.86	1.689	2.021 GB	13.64 GB	2.646 Mbps	202.9 Kbps	4267	241.5 GB
8. lxfsrd0701				1.05	6	11	9.374	0.911	2.013 GB	13.65 GB	1.746 Mbps	92.12 Kbps	4270	240.9 GB
9. lxfsrd0702				1.75	4	11	21.09	0.898	2.059 GB	13.6 GB	2.253 Mbps	137.9 Kbps	4160	233.9 GB
10. lxfsrd0705				3.66	4	10	24.61	2.621	2.165 GB	13.5 GB	1.092 Mbps	2.992 Mbps	4126	227.3 GB
11. lxfsrd0706				0	0	30	0.015	0.064	5.893 GB	9.77 GB	48.58 Bps	76.8 Bps	13714	685.4 GB
12. lxfsrd0906				1.03	4	11	8.721	0.569	2.257 GB	13.41 GB	1.677 Mbps	82.28 Kbps	4220	236.7 GB
13. lxfsrd1101				1.75	6	11	21	1.407	2.261 GB	13.4 GB	2.379 Mbps	141 Kbps	4329	245.7 GB
14. lxfsrd1111				1.76	4	11	19.69	0.734	2.252 GB	13.41 GB	2.301 Mbps	153.6 Kbps	4211	235.4 GB
15. lxfsrd1114				0.9	4	11	9.352	0.797	2.171 GB	13.49 GB	1.515 Mbps	89.4 Kbps	4221	233.9 GB
Total	**15**									**198.7 GB**	**21.41 Mbps**		**68749**	**3.675 TB**
Average		**1**	**1**	**3.503**	**4.2**	**11.6**	**15.47**	**1.278**	**2.417 GB**	**3.2**	**1.427 Mbps**	**1.543 Mbps**	**4910**	**268.8 GB**

- Many more parameters are available
  - Staging queue, usage of root and log partitions
  - CPU nice and idle status
  - Memory consumption details
  - Number of network sockets

Questions adressées

- Évaluation générale du système
- Performance (surtout ATLAS)
- Gestion des datasets
- Scheduling
  - Prioritées, fair-sharing (ALICE)
  - Traitement des congestions (ATLAS)
- Intégration PROOF/batch

Évaluation générale du système

Plusieurs améliorations internes ont suivi ces tests

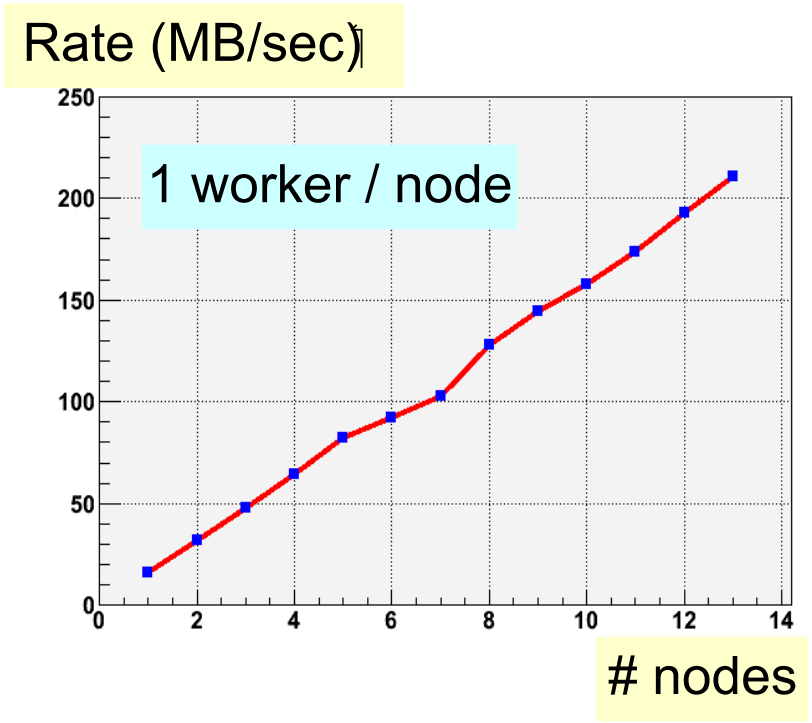
- Stabilité
  - Première version du plug-in XROOTD souffrait de deadlocks
- Interface aux datasets
- Interfaces pour la simulation
- Monitorage de la memoire, accès aus fichiers logs
- Fusion optimisé des gros objets output
- Nouveau distributeur de travail (packetizer)
- Fédération (master multiples, ATLAS)
- Authentication GSI / Authorization (ALICE)
- ...

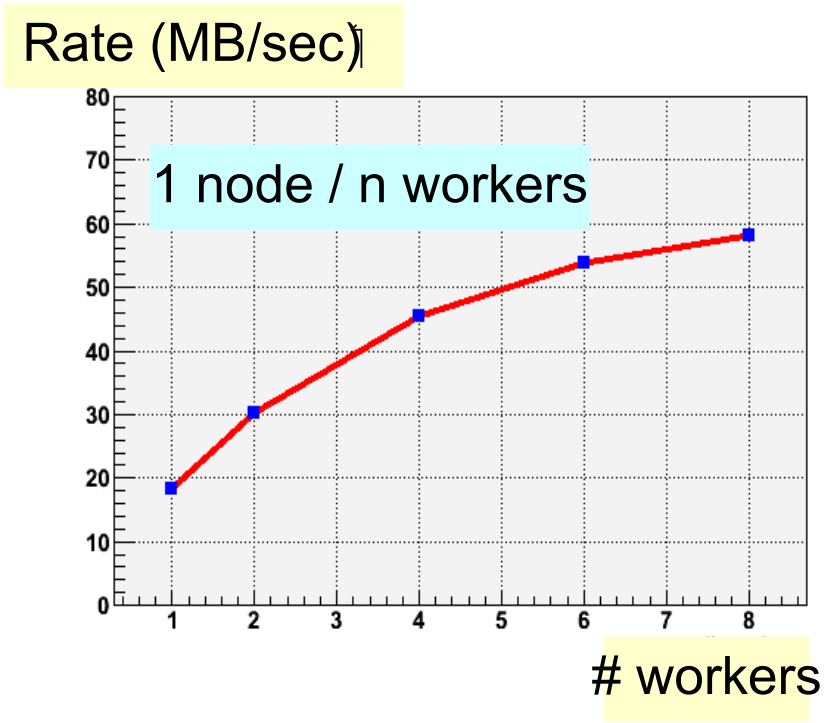
Authentication / Authorization pour ALICE à la CAF

- Authentication
  - Protocole GSI
    - Plug-in XrdSecgsi de XROOTD est utilisé
    - Plusieurs trouvés et fixés
  - Un proxy valide est crée sur chaque worker
    - Le workers peuvent se connecter à AliEn via TGrid::Connect
- Authorization
  - Obtenue en utilisant la LDAP d'ALICE

Performance: scalabilité dans le CAF d' ALICE

- Analyse ESD, ~1 TB (~25% read)
  - ~10 min w/ 26 workers

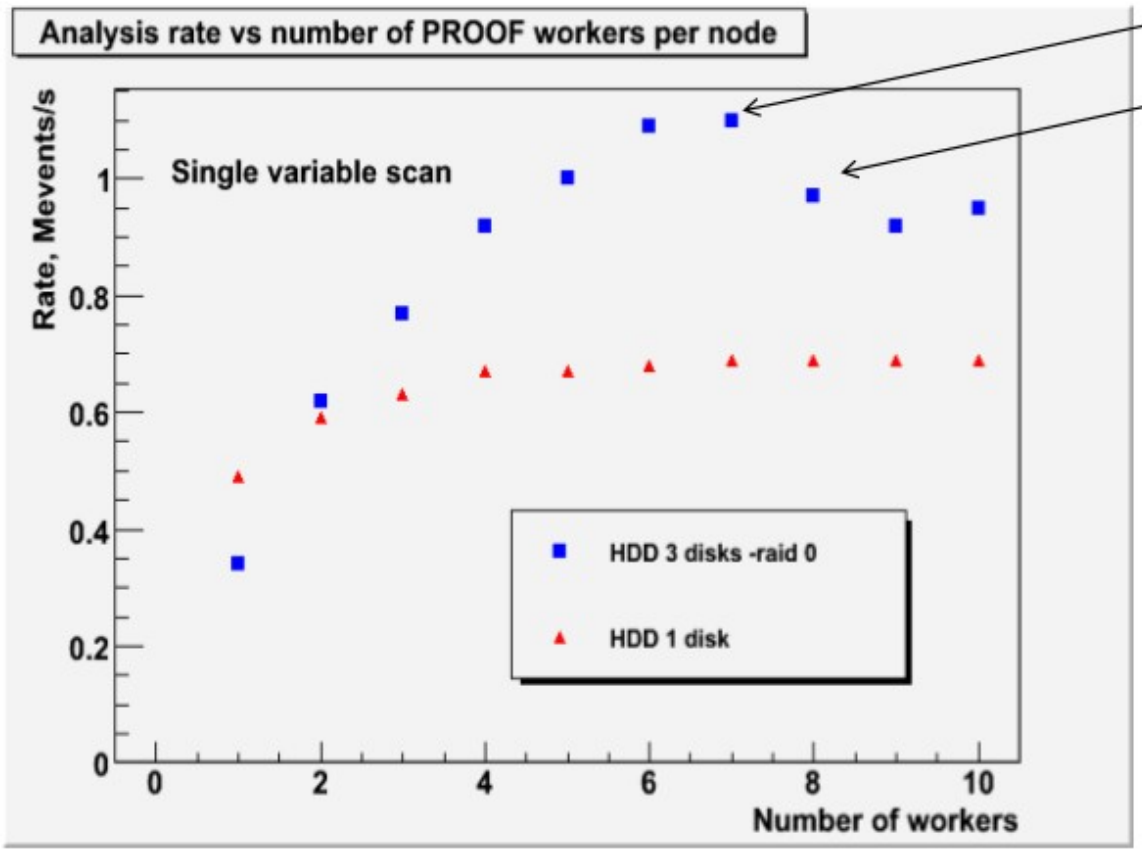




- Au niveau d'une machine: scalabilité limitée par l'hardware I/O
- Le comportement du système est tout de même prévisible

Performance: système multiple HDD en RAID-0

Courtesy of S. Panitkin, BNL



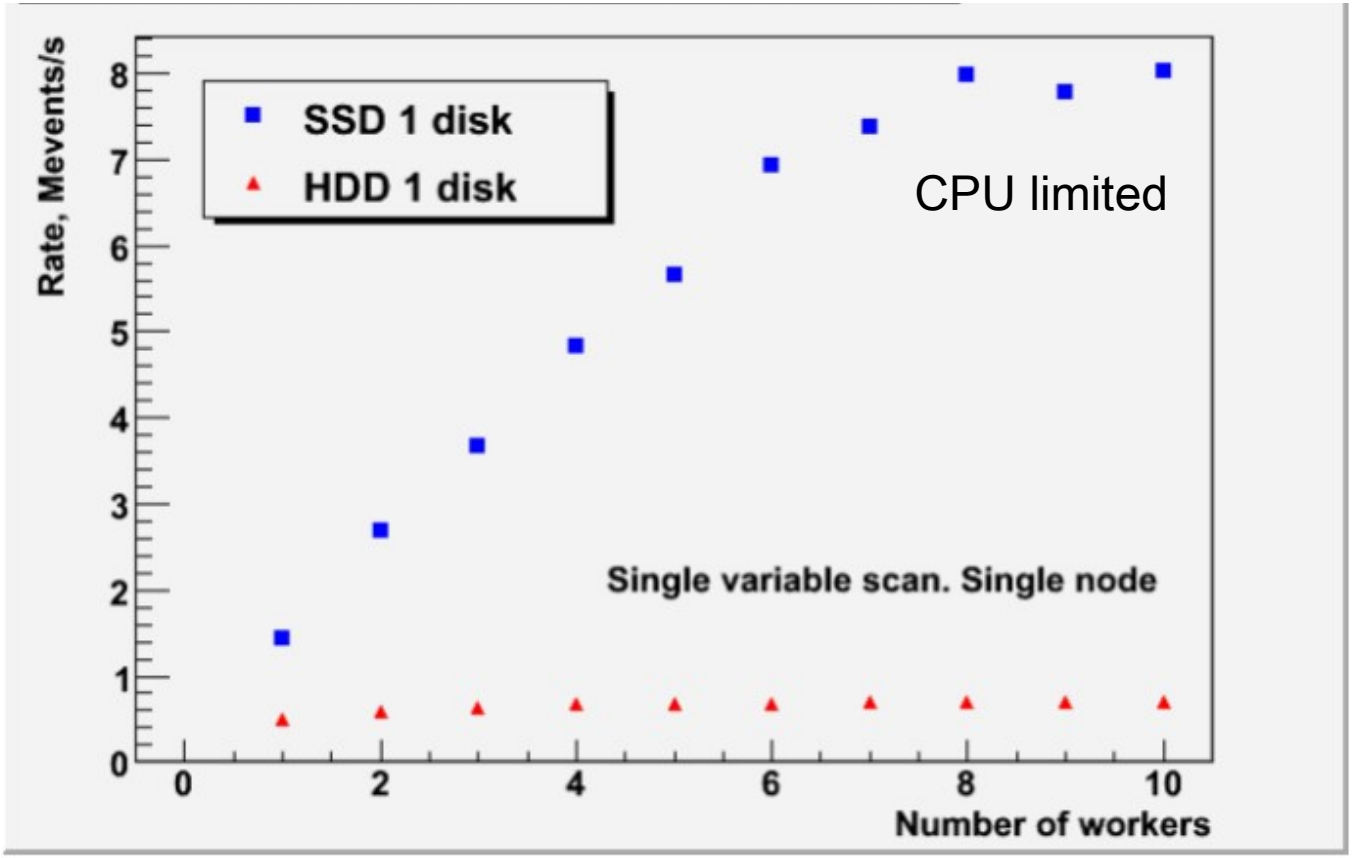
3x750GB disks in RAID 0 (software RAID) vs 1x500GB drive

1 disk shows rather poor scaling in this tests
3 disk raid supports 6 workers?

Sergey Panitkin 12

Performance: utilisant de Solid State Disks (SSD)

Courtesy of S. Panitkin, BNL



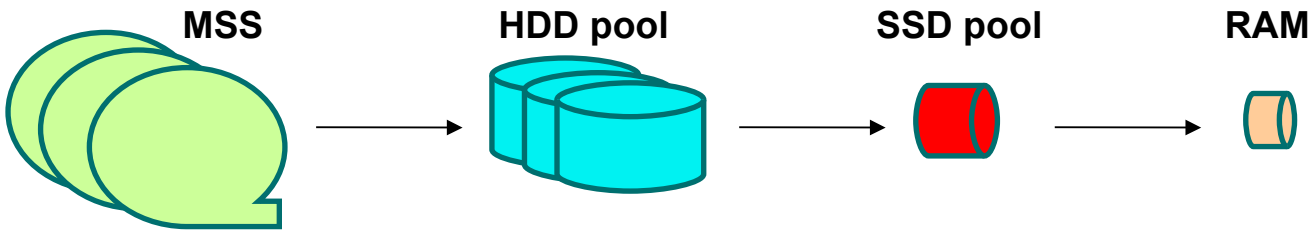
With 1 worker : 5.3M events, 15.8 MB read out of ~3 GB of data on disk
With 8 workers: 42.5M events, 126.5 MB read out of ~24 GB of data

Sergey Panitkin 10

Considerations après les tests d'I/O de ATLAS

- Les tests confirment que
  - Un worker PROOF traiter les donnés à 10-15 Mb/s
  - Un HDD SATA peut alimenter 2-3 workers
  - En RAID on peut 1.5-2 workers per HDD
  - Les SSD ne semblent pas limités par le hardware actuel … mais par le prix

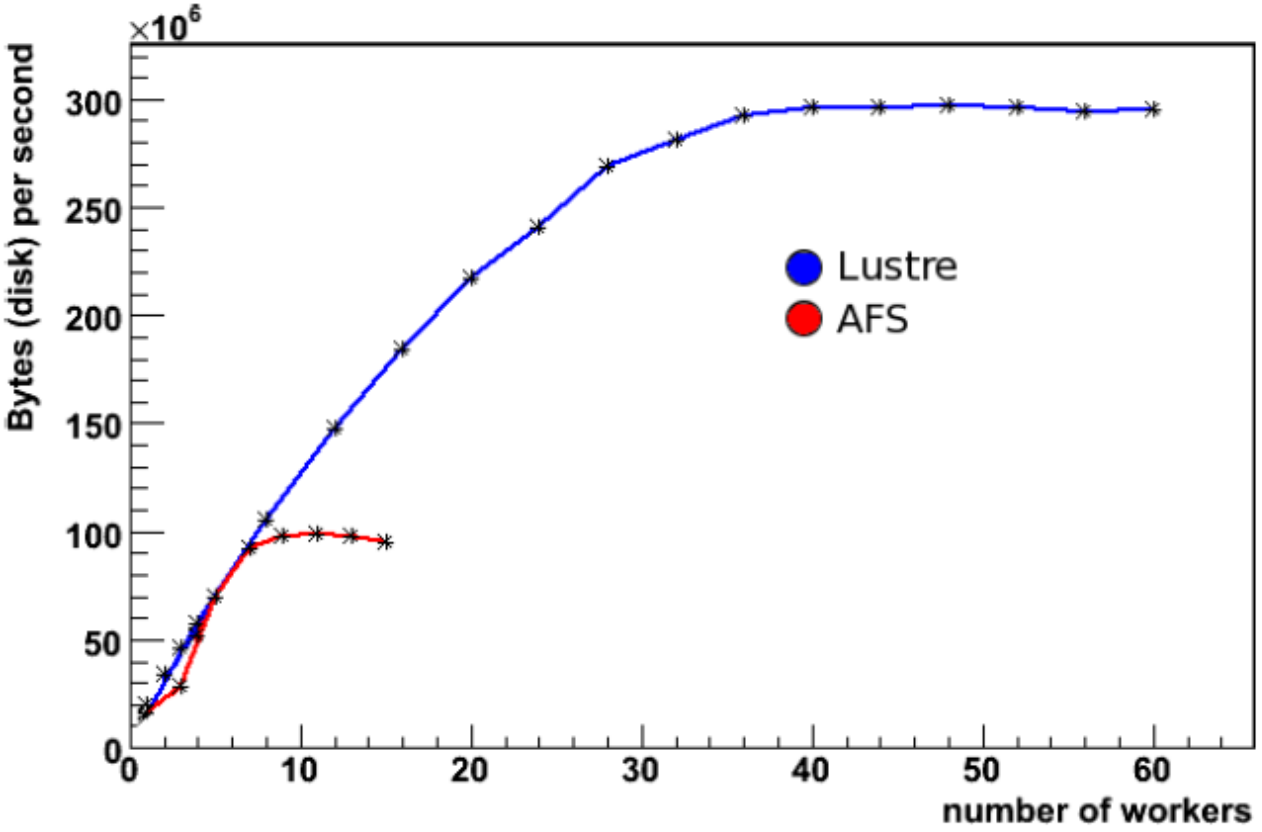
- Un système multi-tier de storage locale semble promettant



pourvu que un système efficace de rémplissage de différent tiers soient développé

Performance Lustre à NAF

W. Behrenhoff, H. Stadie, Hamburg



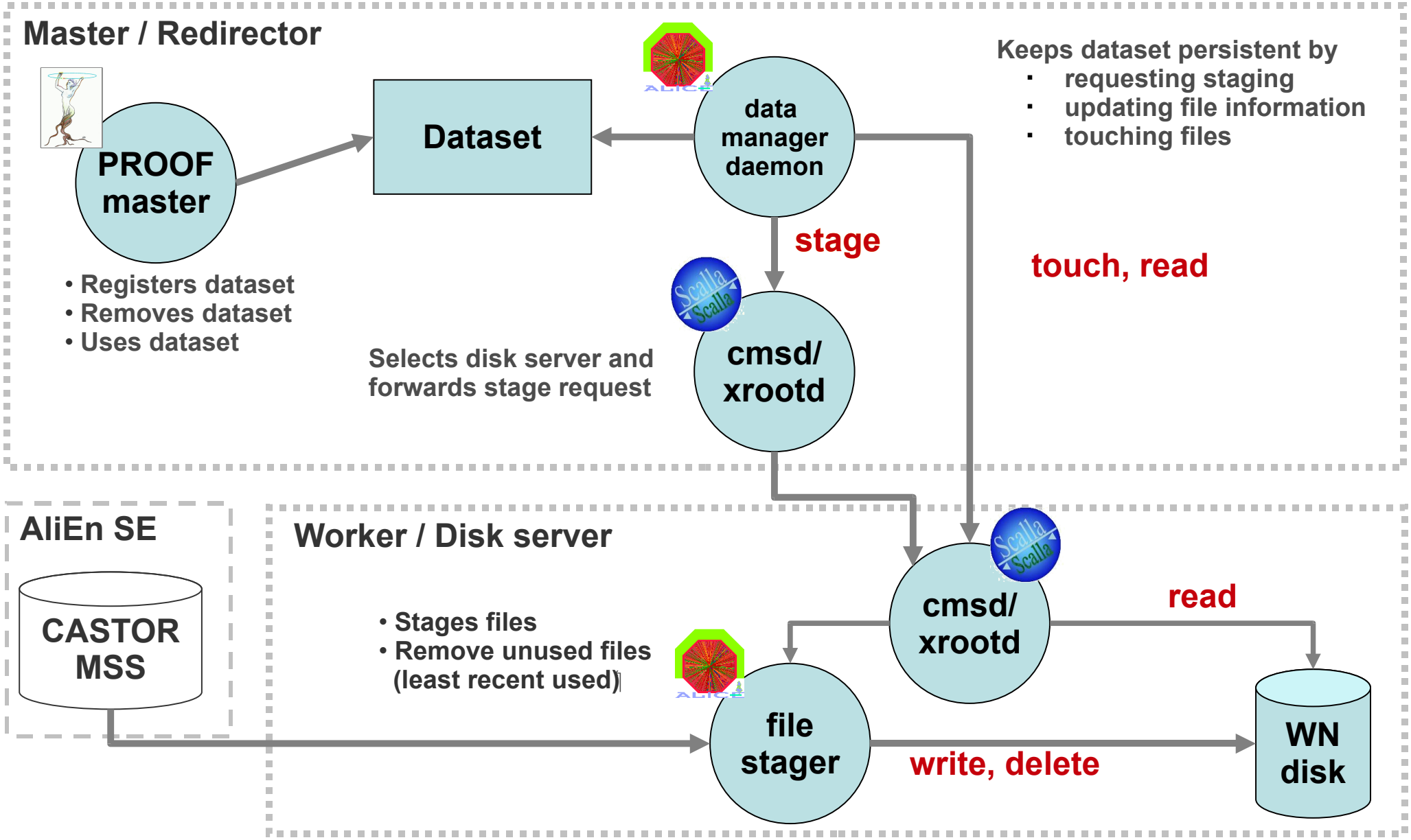
As expected, Lustre via InfiniBand clearly outperforms AFS. Only one file server was used in this test. Thus, the input rate should increase further with additional servers.

Gestion de datasets

- Interface simple et universelle avec les catalogues des experiences.
- Dataset manager
  - Gestion de l'information sous forme d'objet TFileCollection dans des fichiers ROOT sur le master où ailleurs
  - Gestion de base des quota disque
- Developpé par ALICE et le team PROOF
- Utilisé par ATLAS via de scripts ad hoc appelés PQ2-... et qui seront distribués avec ROOT à partir de 5.24 .

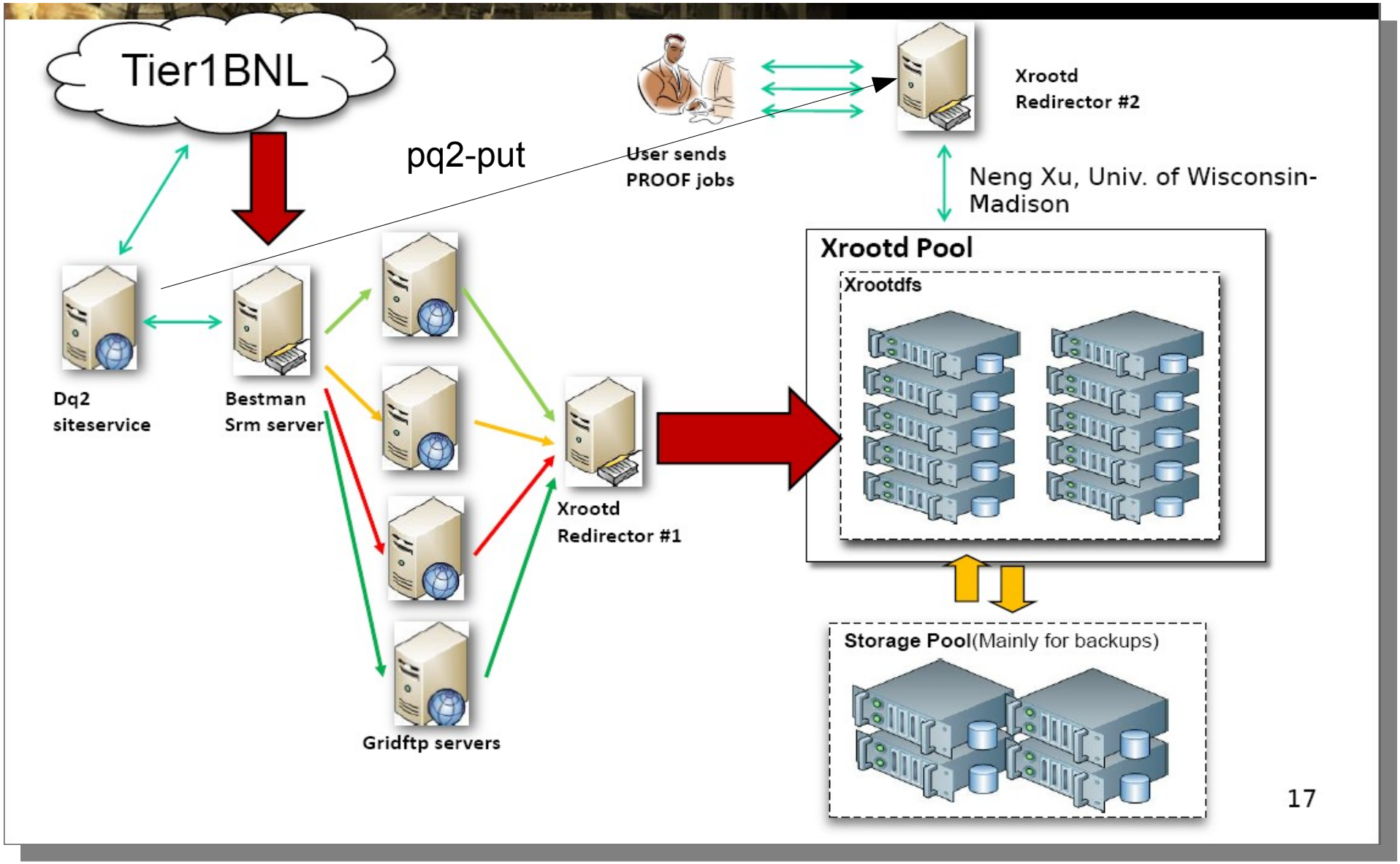
Dataset handling at ALICE CAF

Courtesy of J.F Grosse-Oetringhaus, CERN



Système de peuplement des donnés – UW Madison

Courtesy of N. Xu, S. Panitkin



Ferme PROOF à BNL: transfer des données

Courtesy of N. Xu, S. Panitkin

analysis

acas0420.usatlas.bnl.gov

/data

BNLXRDHDD1

Xrootd/PROOF Farm

DQ2

T0

Panda

Grid transfers

dCache

xrdcp with dq2 registration

Sergey Panitkin

8

Paramètres pour un PROOF cluster

- CPU, RAM
  - Avec des processeurs multi-cores processors, la CPU n'est pas typiquement le bottleneck
  - >~ 2 GB/core est typiquement OK
- Local storage
  - 2÷3 workers pour HDD (~10÷15 MB/s per worker)
  - En RAID, 1.5÷2 workers per HDD
  - SSD peut servir 8 workers sur une machine à 8-core
- Network
  - Pas vraiment un problème si les données sont sur les disques
  - 1 GB/s typiquement suffisant entre les workeurs et vers l'exterieure
  - 10 GB/s vers un serveur exterieur peut servir correctement ~80 workers

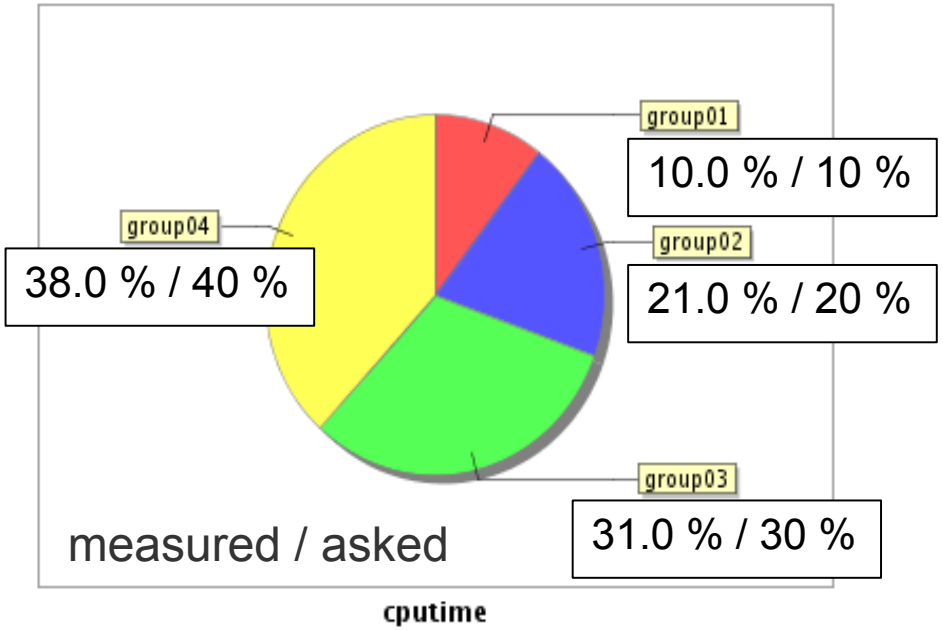
Scheduling

Niveaux de scheduling

- Controller la répartition de la CPU parmi les sessions qui sont actives
  - Fair-sharing
  - Appliquer les politiques des experiences
- Controller combien-de / quelles ressources (workers) sont affectées à un utilisateur
  - Éviter les congestions
  - Appliquer les politiques des experiences

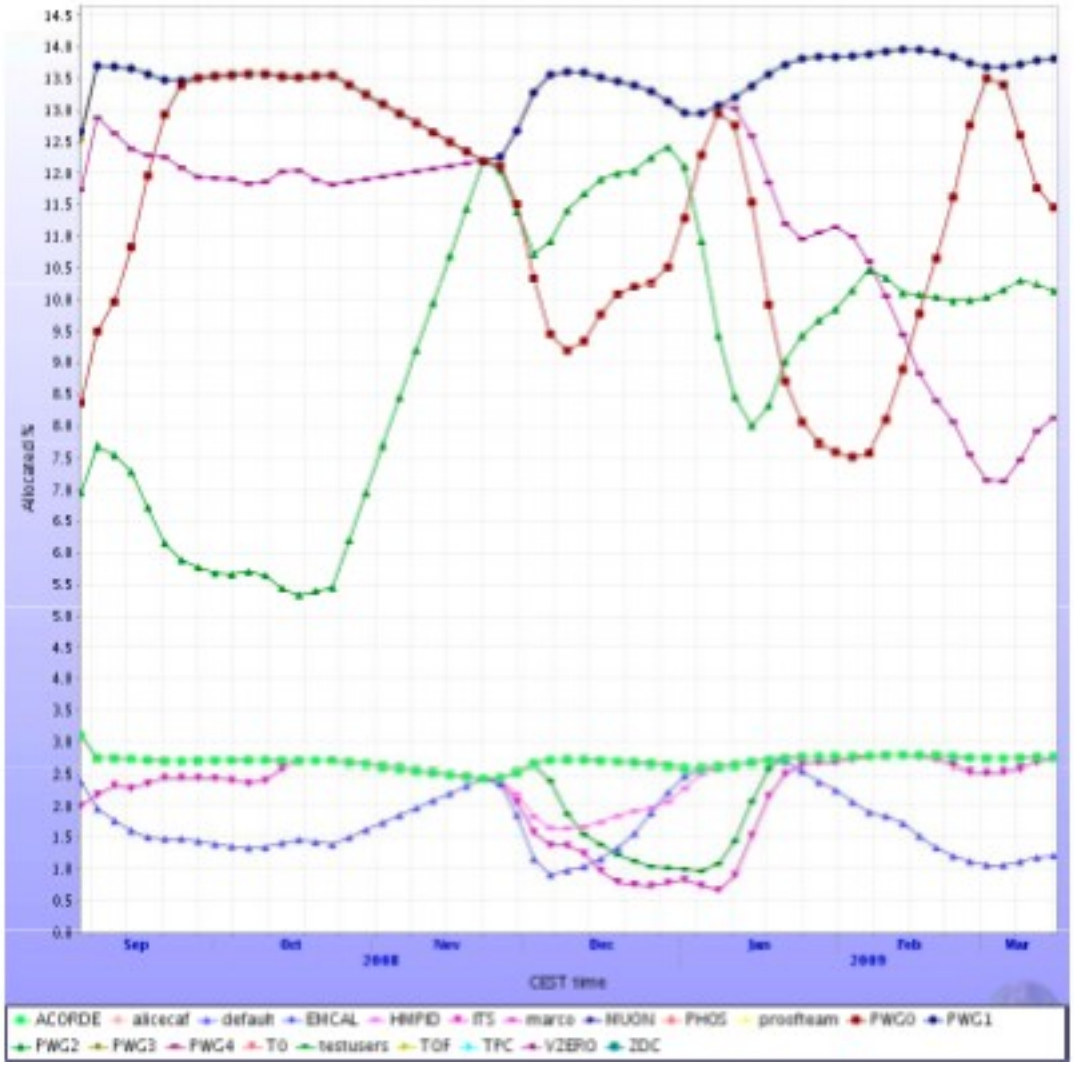
Scheduling: répartition de la CPU

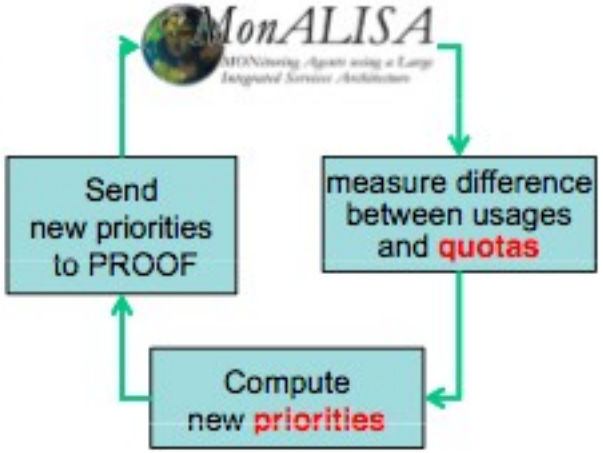
- Basée sur la priorité du group de l'utilisateur
- Technologie de type *renice*
- Controllé
  - Centralement avec un système de monitoring
  - Localement sur chaque worker
- Exemple: stress-test run 4 groups sur 1 jour



Fair-sharing à la CAF

Courtesy of M. Meoni





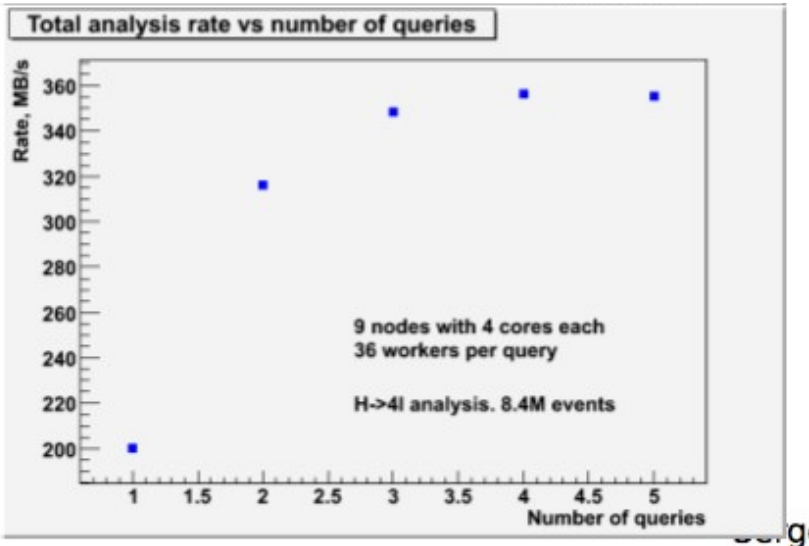
Default group quotas:
- Detectors: x
- PWGs: 5x

10

Scheduling: affectation de ressources

- Le output combiné du système est maximale quand

  # workers ~ (1÷2) # cores



S. Panitkin

- Aggregate analysis rate saturates at about 3 (full load) queries
- Max analysis rate is about 360 MB/s for a given analysis type

It makes sense to run PROOF farm at optimal number of queries

Analyse H->4l
4.5 M evts Monte Carlo, ~68 GB;
CPU et I/O intensive

gey Panitkin

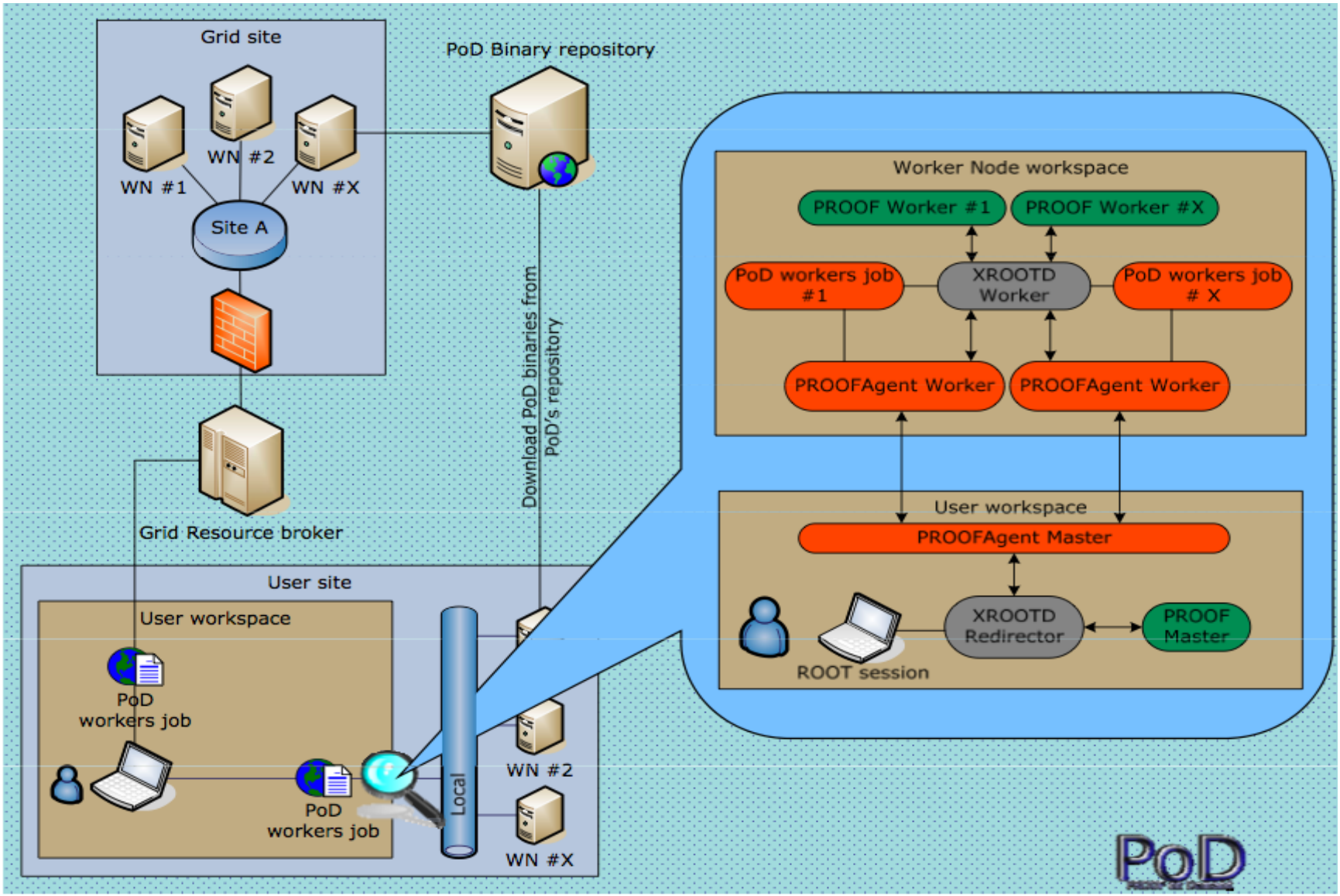
- PROOF permet de controller le nombre de sessions actives
  - Queue interne First-Come-First-Served
  - Sessions suspendues répartent dès que un slot devient libre

Integration avec des systèmes batch

- Model intéressant pour de petits T3 pour remplir de façon constante les machines
- Idée de base: reduire la priorité des jobs batch de façon de favoriser les jobs PROOF
- En realité, on utilise le système batch pour créer un cluster PROOF privée pour chaque utilisateur
- Cela permet de bien gérer les ressources en
  - Utilisant le scheduling du système batch
  - Confinant l'utilisateur dans son éspace
- Models existants:
  - LSF à Darmstadt
  - SGE à DESY
  - Condor à UWM

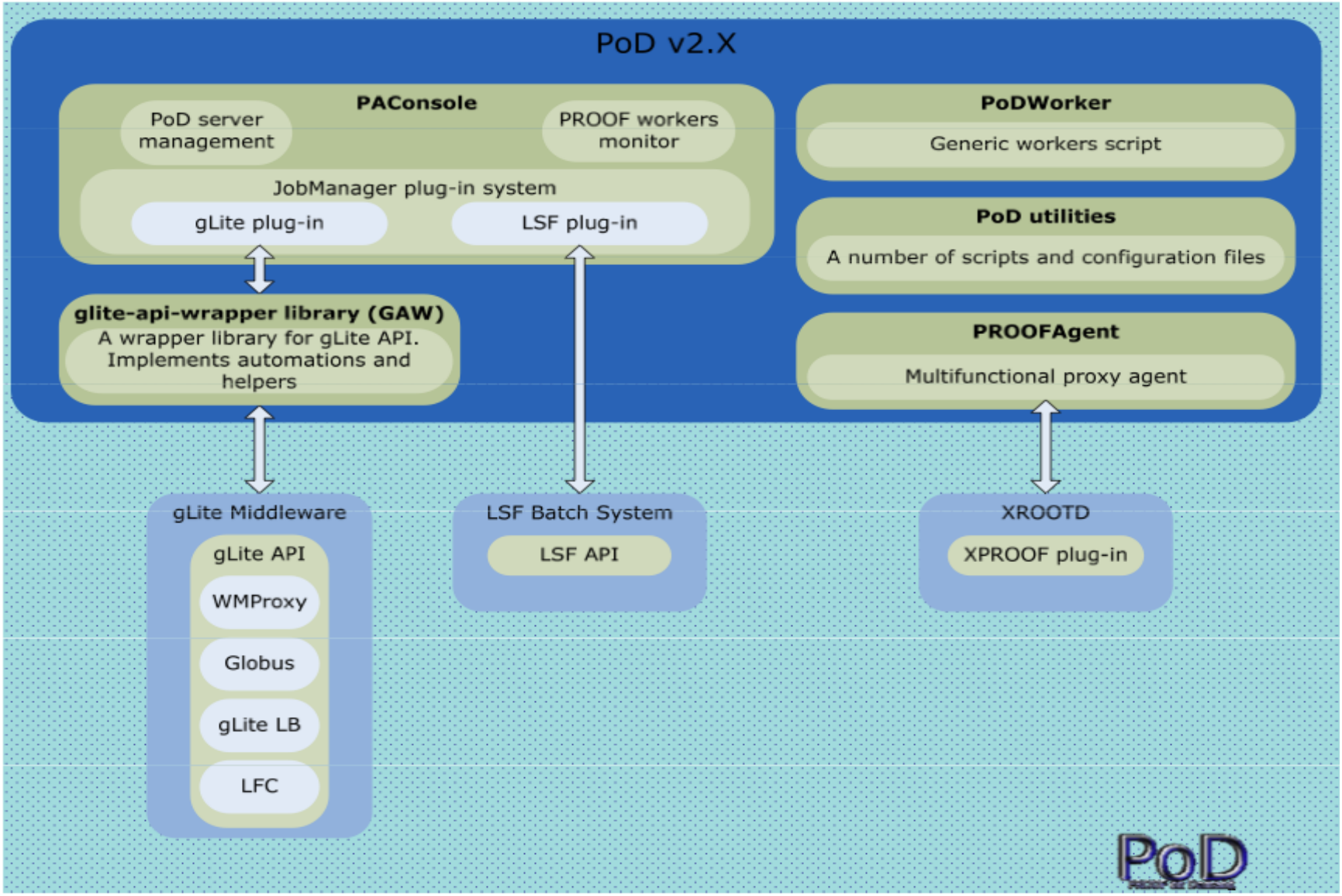
Proof-On-Demand à GSI - Darmstadt

A. Mananof, GSI



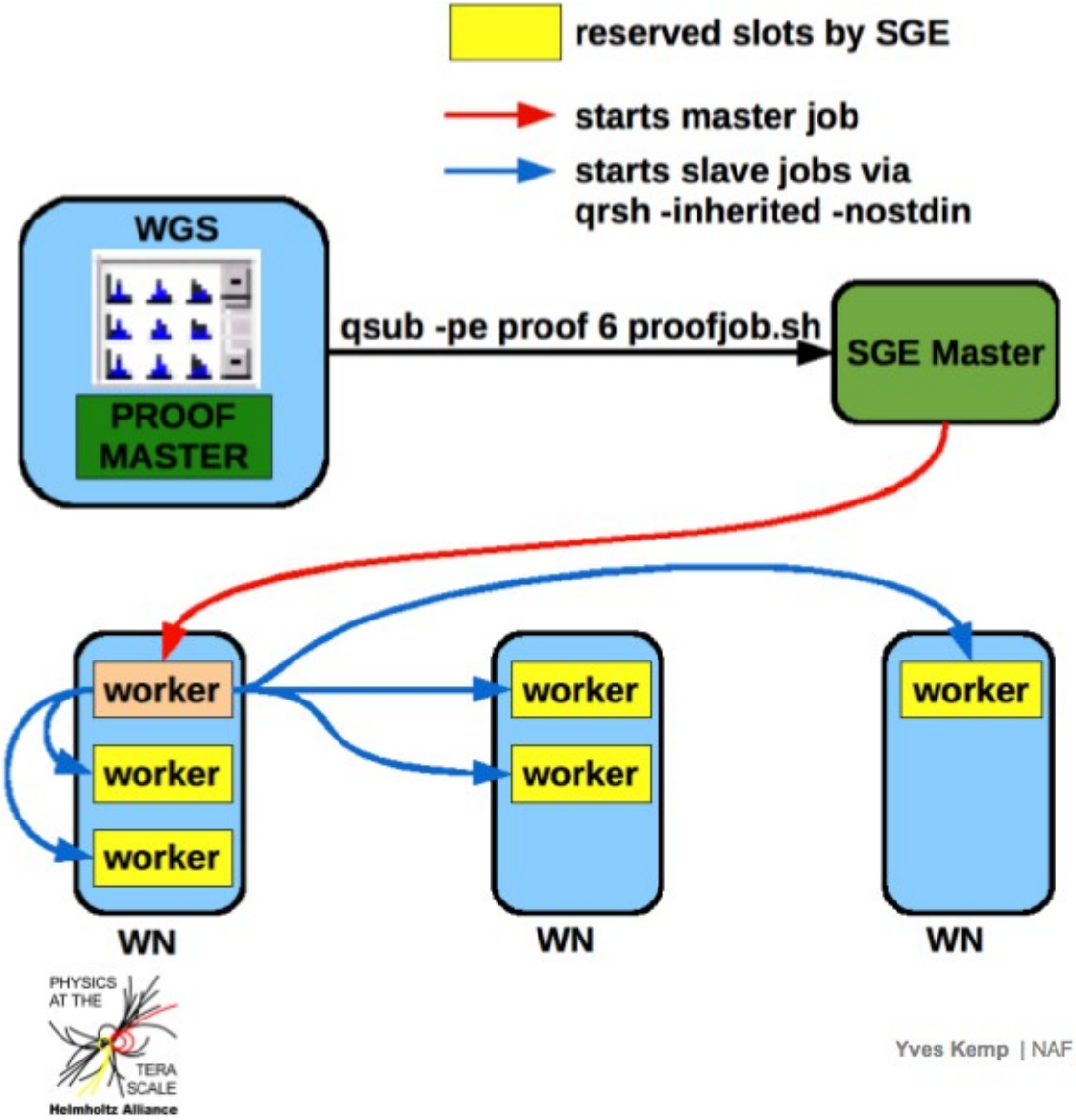
Proof-On-Demand à GSI – Darmstadt (suite)

A. Mananof, GSI



PROOF avec SGE à NAF

Y. Kemp, DESY



- Allows massive parallelization of analysis jobs
- Keep interactive "ROOT prompt"
- Used mainly by CMS (Uni HH)
- Allows for multi-user and multi-group operations
- Accounting & security possible

PROOF&SGE:
Poster ID 66

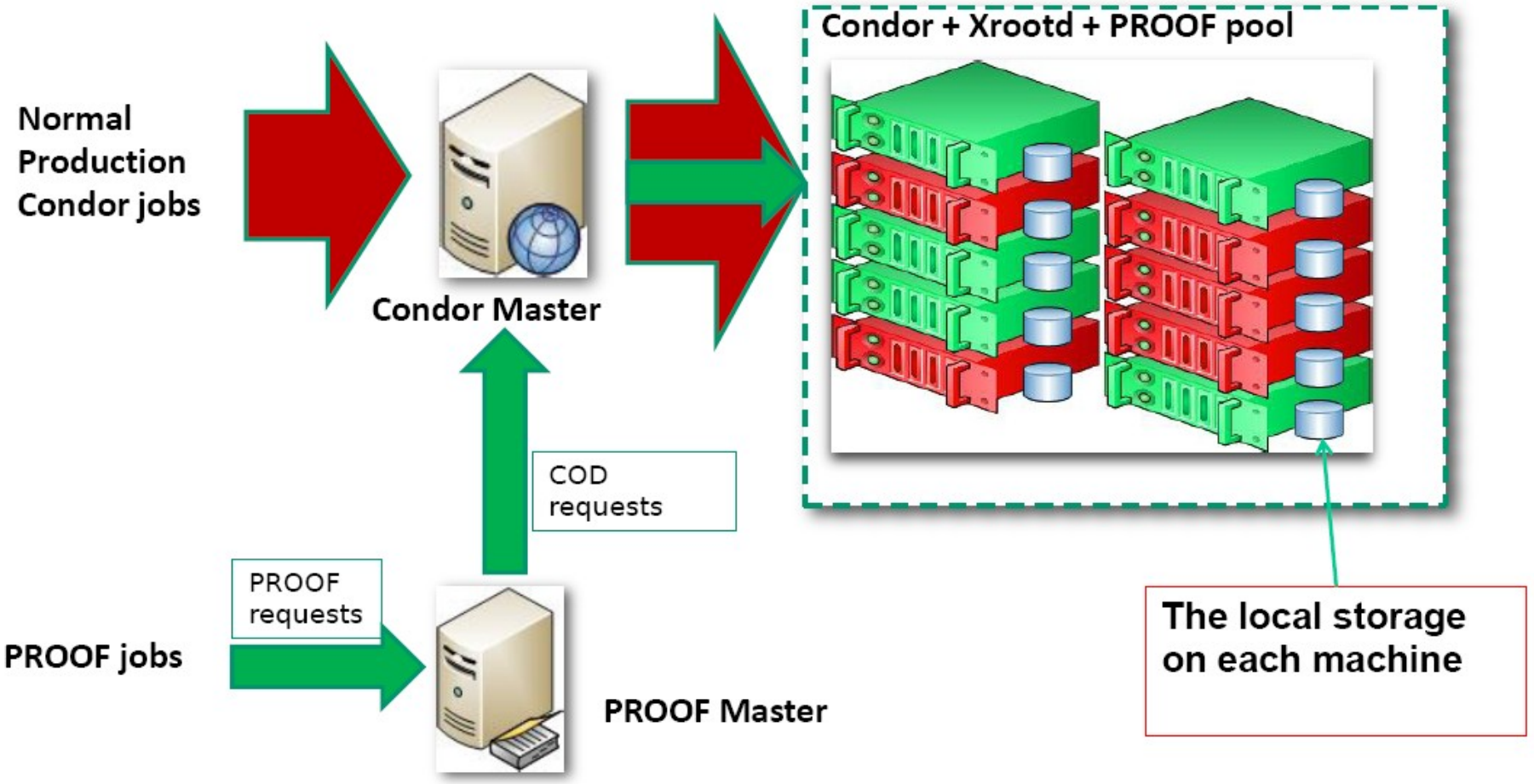
PHYSICS AT THE TERA SCALE
Helmholtz Alliance

Yves Kemp | NAF @ DESY| CHEP 24.3.2009 | Page 9

DESY

Model de base PROOF+COD (Condor)

N. Xu, GG



Sergey Panitkin

20

Résumé

- PROOF: alternative basé sur une architecture *pull* pour l'exploitation des fermes T2/T3 pour l'analyse interactive
- Des prototypes de ferme d'analyse basé sur PROOF ont été testés, surtout par ALICE et ATLAS
- Différents aspects étudiés
  - Fonctionnalité, Performance, Scheduling
  - Résultats valables aussi d'autres contextes (voir I/O)
- Modèles d'intégration avec des systèmes batch
  - Solution si un cluster pur PROOF n'est pas abordable
  - Première impressions très positives

C'est la fin! Merci!

- Nouvelles pages web
  http://root.cern.ch/drupal/content/proof
- Forum PROOF
  http://root.cern.ch/phpBB2