Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

Utilisation mémoire / LCG FR

renaud.vernet@cc.in2p3.fr

Introduction

- Objectifs
  - Comprendre utilisation mémoire par jobs LHC
  - Dimensionnement workers
  - Vue ‘expérience’ vs vue ‘site’
    - Comment gérer les jobs selon leur mémoire consommée

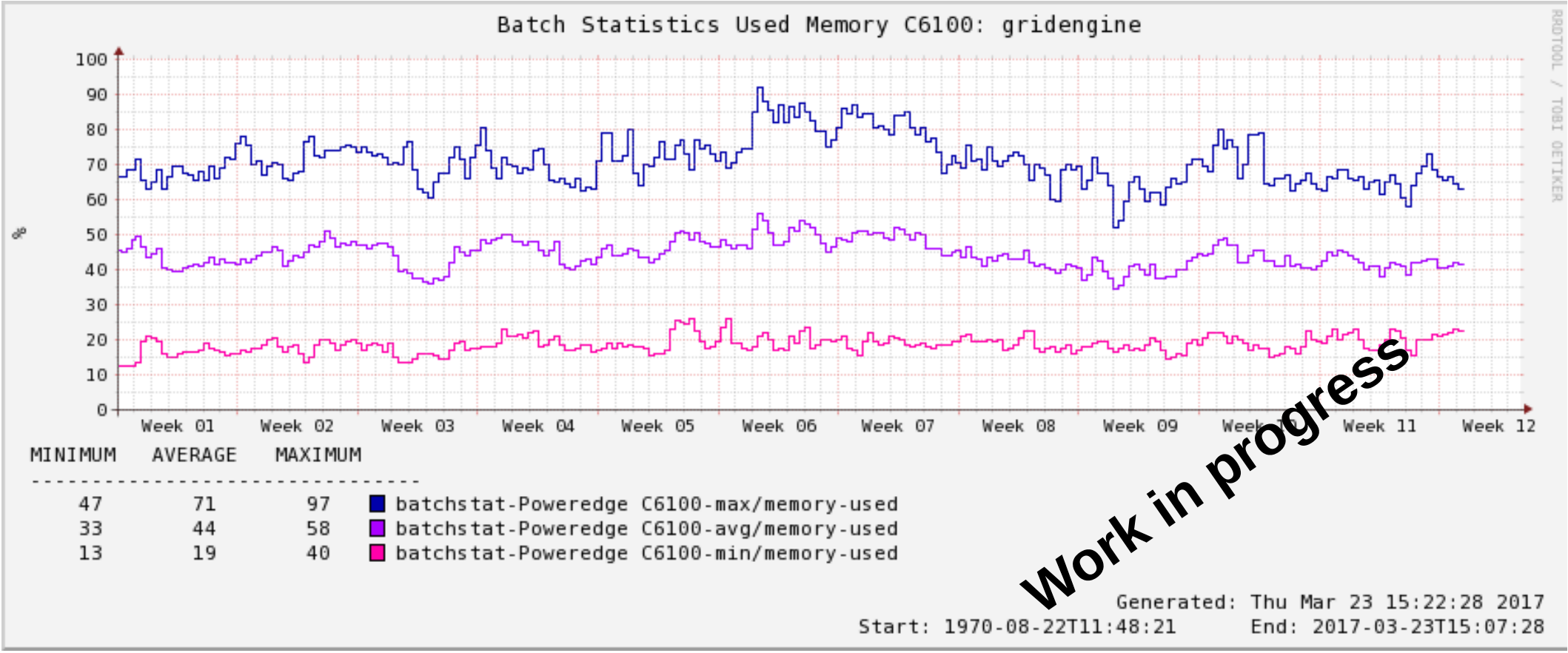
- Petit groupe de travail pour initier la réflexion
  - Edith, Catherine, Victor, Sébastien, Adil, Renaud
  - T1 et T2
  - 4 VO « représentées »
- Point central d’ « information » :
  - http://lcg.in2p3.fr/wiki/index.php?title=MemJobs

Mise en bouche

- Valeur officielle de RAM demandée par WLCG
  - → 2 GB/slot
- Achats de serveurs dans sites LCG France
  - Dépend du site et/ou du modèle de serveur
  - → 2-3 GB/core
- Utilisation de la mémoire par les jobs des VO
  - Entre 0 et 4+ GB/slot, parfois plus
- Limitations en place
  - Aucune, max(VMEM), max(RSS), cgroups
  - Niveau batch, niveau système, par process, par job (arborescence)

Vue globale workers CC

- Mémoire physique utilisée



- Aide de Fabien Wernli
  - → amélioration des mesures en cours

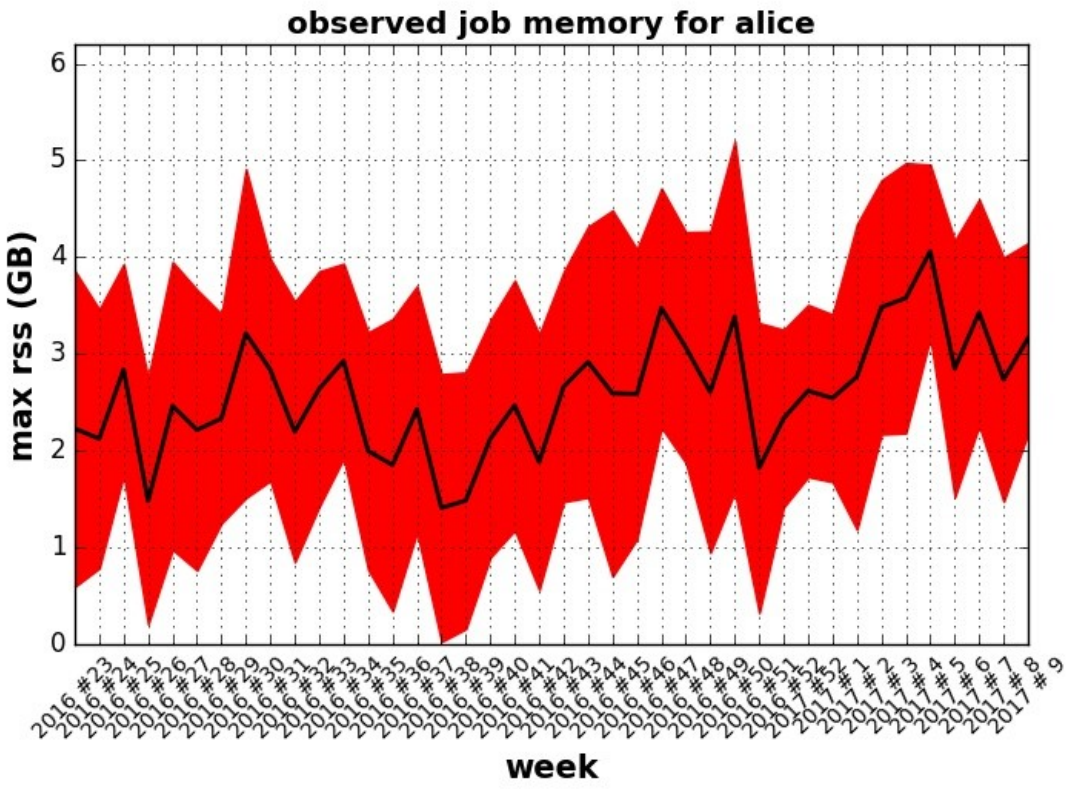
Rappels

- VMEM
  - Mémoire virtuelle
  - Pas directement reliée à la mémoire physique disponible
- RSS
  - Resident set size = privée + partagée
  - Attention aux additions de RSS (!)
- PSS
  - Proportional set size = privée + partagée/n_processes
  - Additions de PSS OK !

- Quelle est la meilleure métrique pour un scheduler ?

Mesures au CCIN2P3

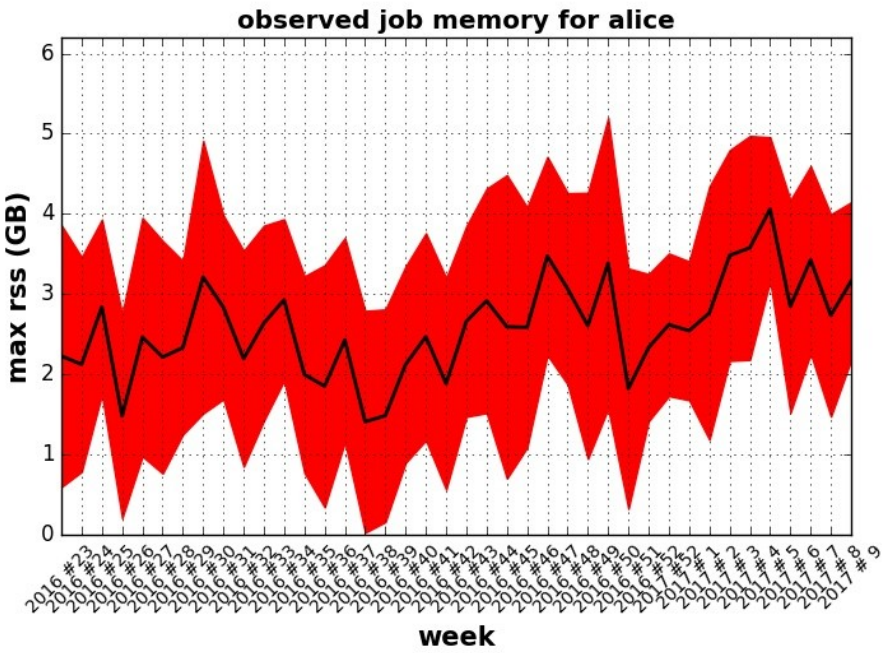
- UGE fournit un accounting
  - Parmi lequel maxvmem : maxrss : maxpss par job

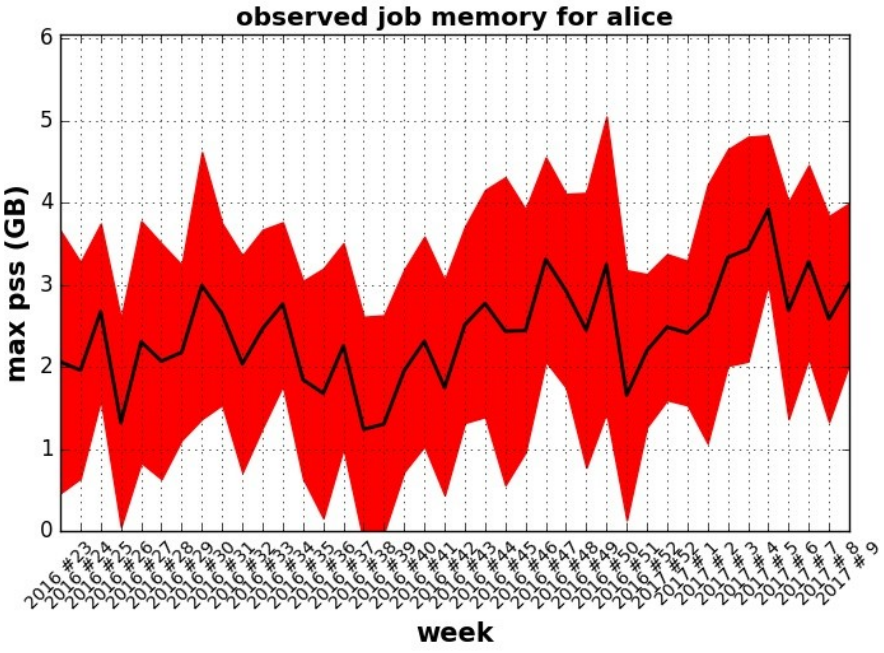


* bins de 1 semaine
* max rss de tous les jobs du bin
  mesuree par UGE (tout le job)

→ moyenne
→ deviation

Mesures au CCIN2P3 : ALICE





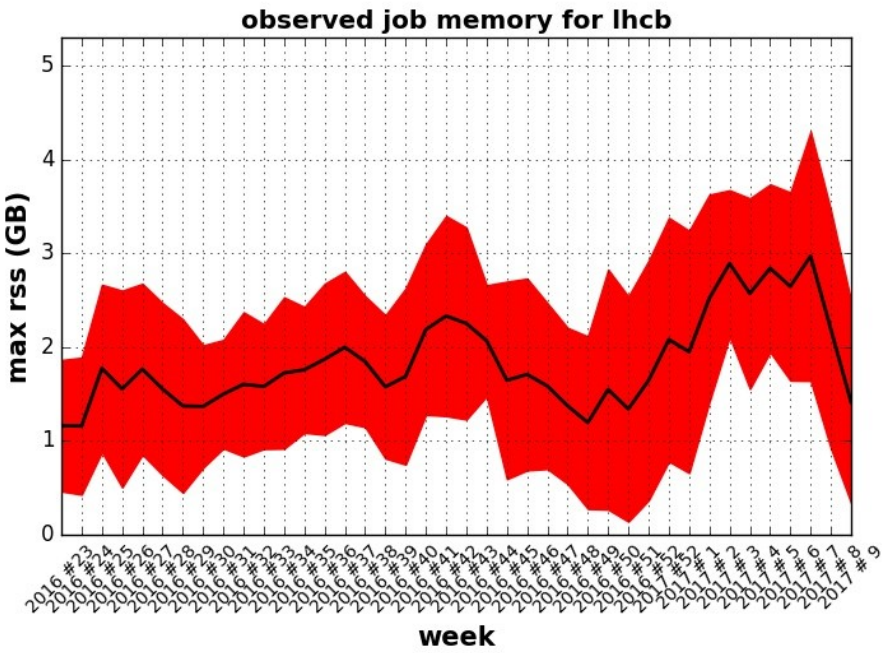
RSS ~ 2.5 GB

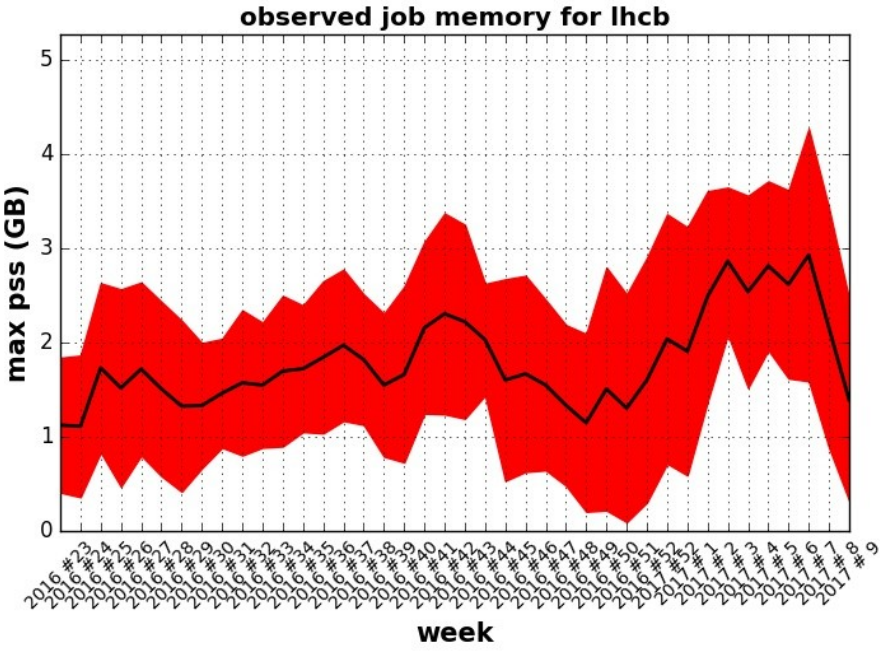
PSS ~ 2.5 GB

Faible différence RSS vs PSS

Jobs single core → peu de mémoire partagée

Mesures au CCIN2P3 : LHCb





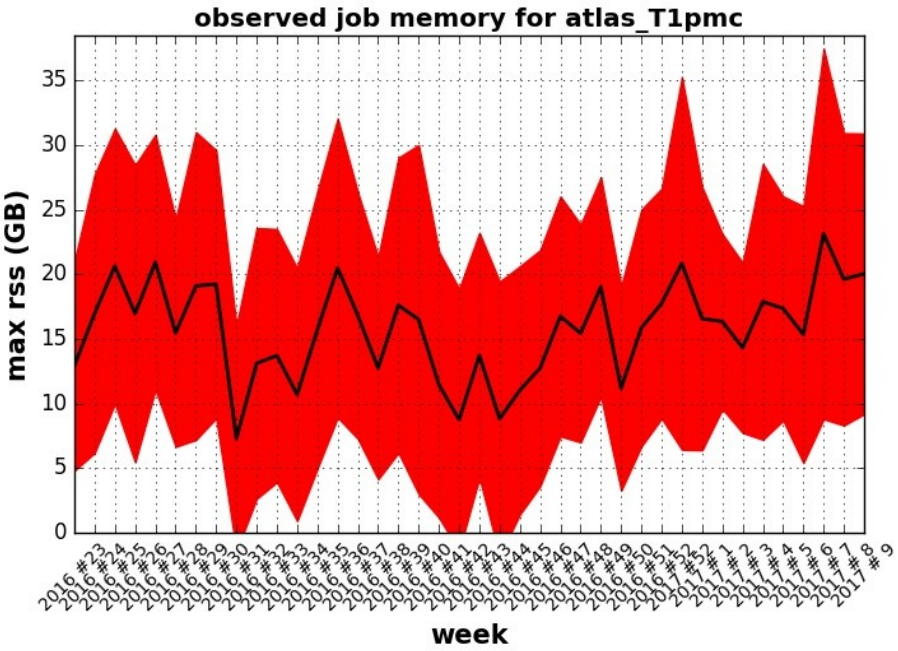
RSS ~ 2 GB

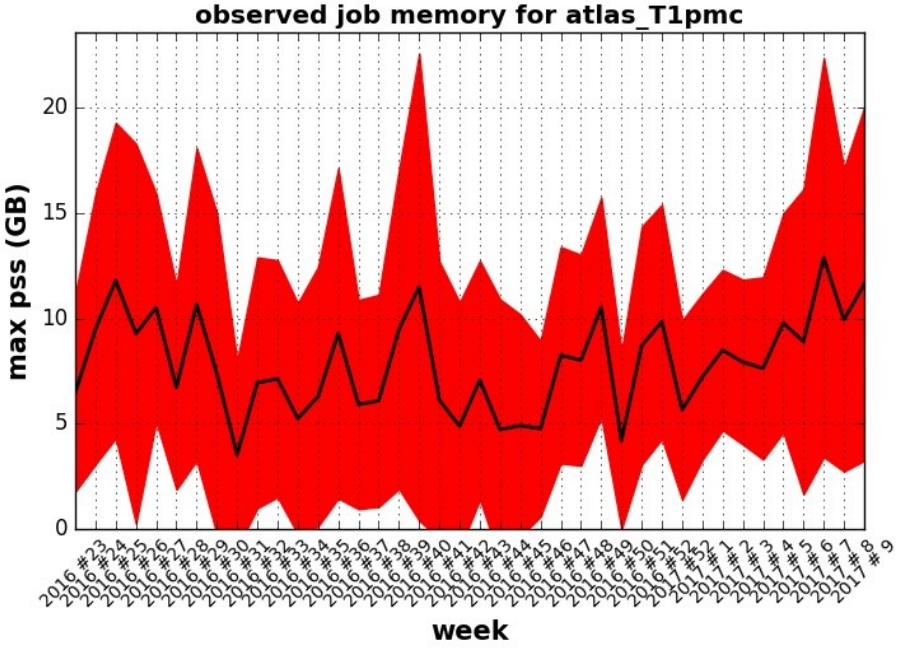
PSS ~ 2 GB

Faible différence RSS vs PSS

Jobs single core → peu de mémoire partagée

Mesures au CCIN2P3 : ATLAS



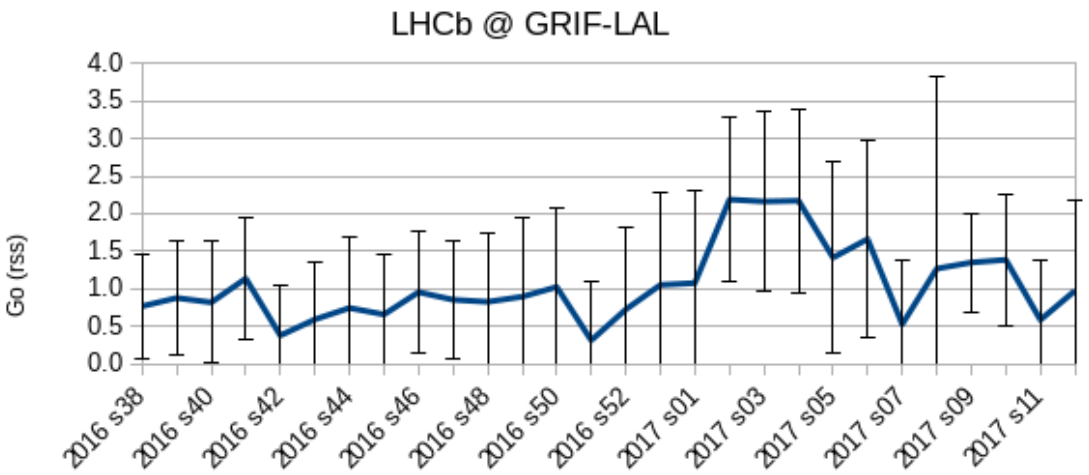


RSS ~ 15 GB

PSS ~ 8 GB

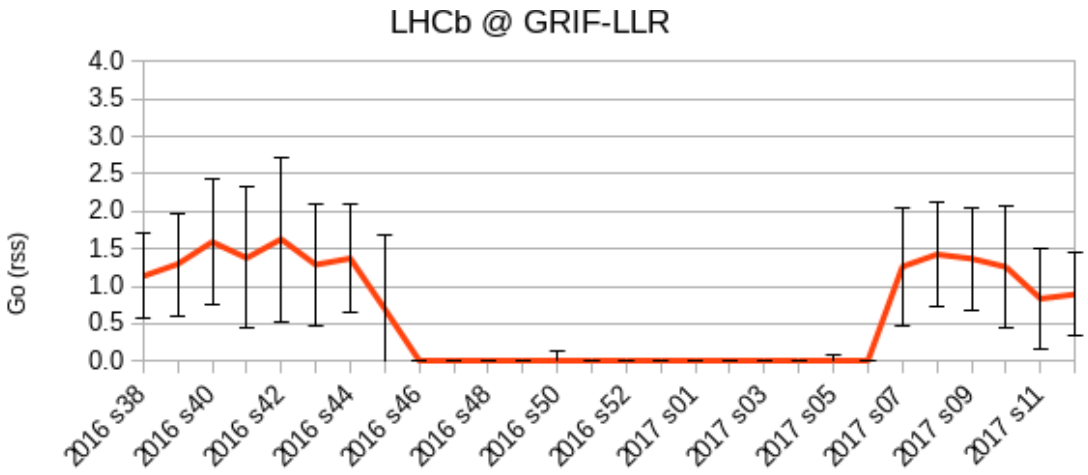
Grosse différence RSS vs PSS

LHCb @ GRIF

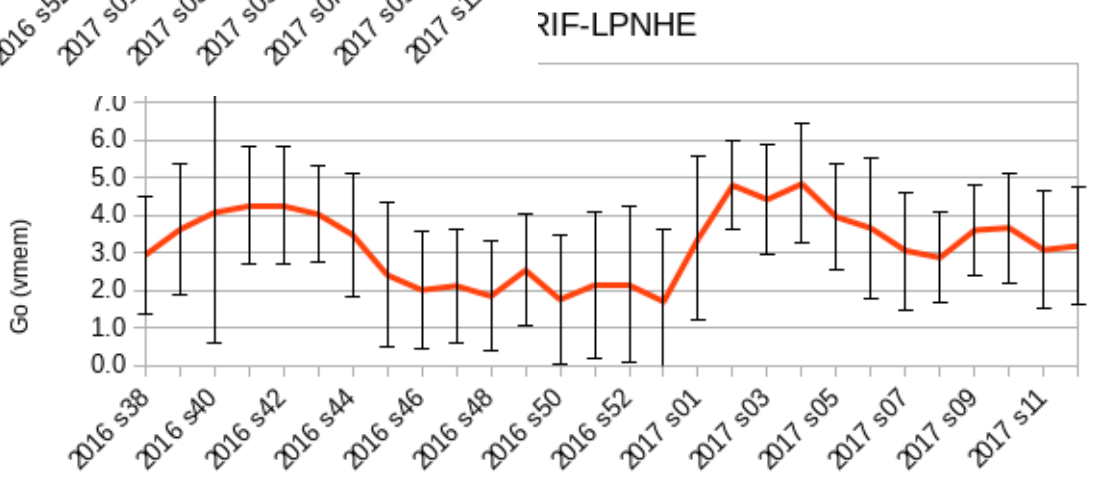


LAL

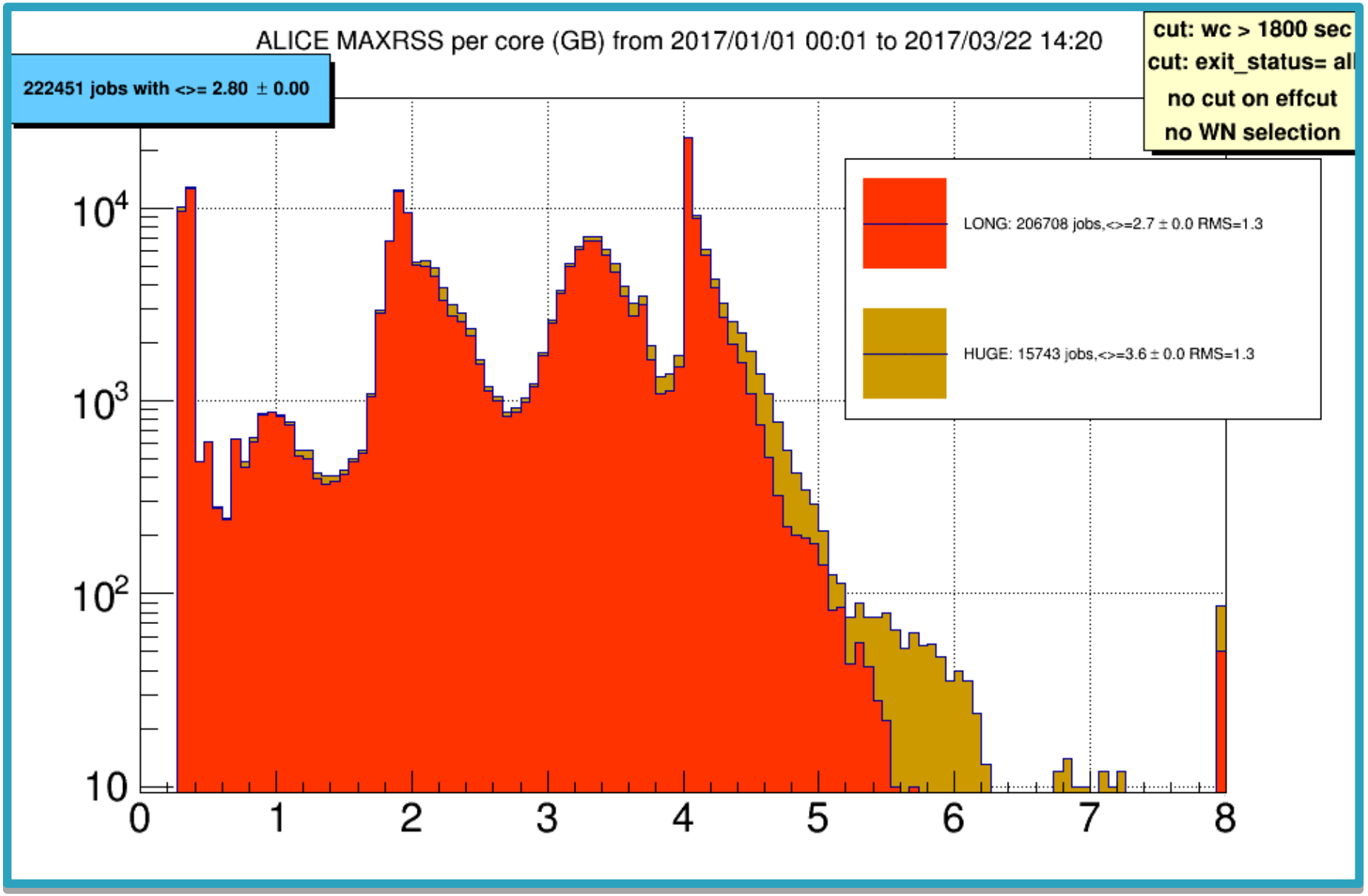
LLR



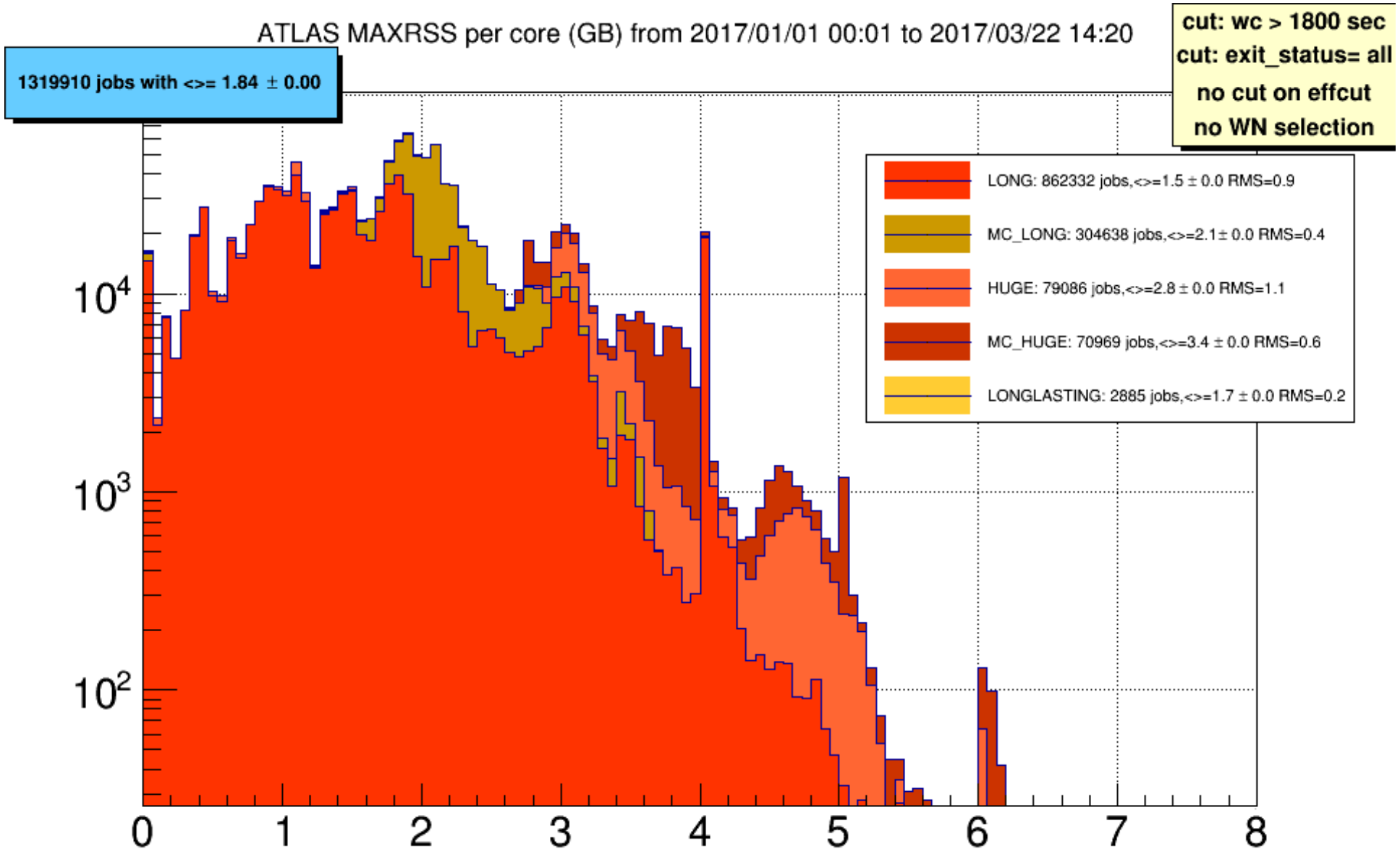
LPNHE



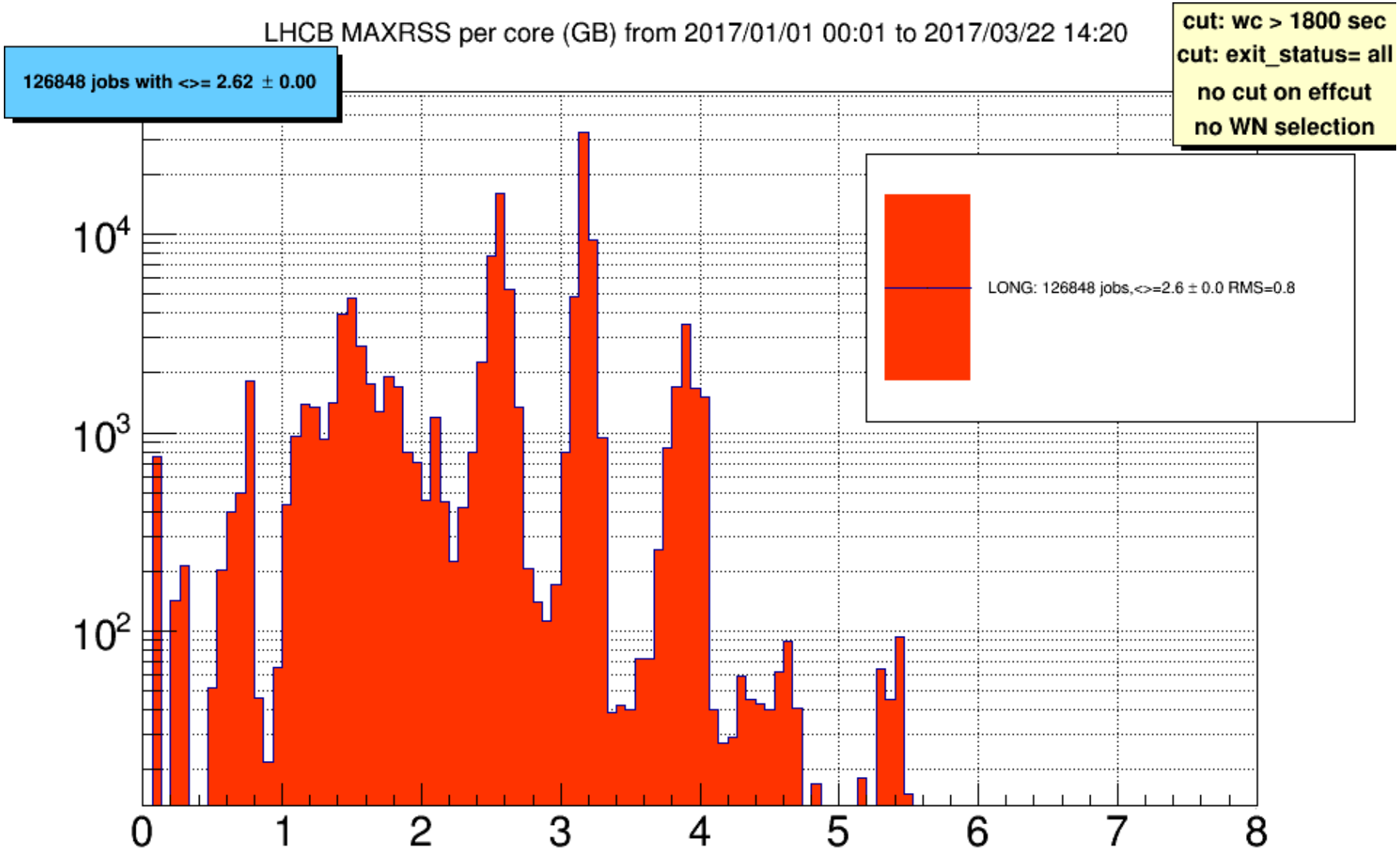
dN/dRSS {ALICE}



dN/dRSS {ATLAS}



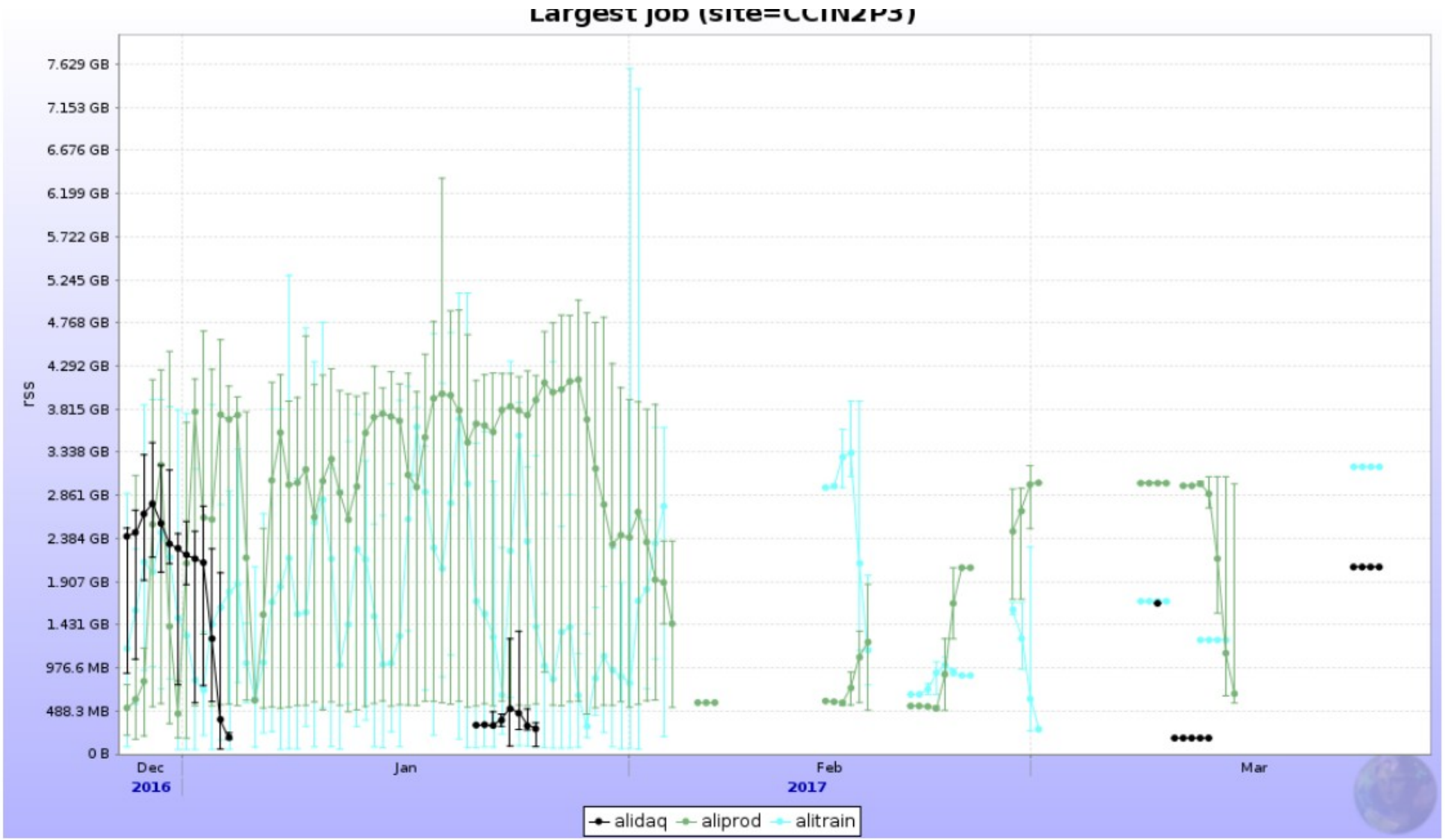
dN/dRSS {LHCb}


LHCB MAXRSS per core (GB) from 2017/01/01 00:01 to 2017/03/22 14:20
126848 jobs with <>= 2.62 ± 0.00
cut: wc > 1800 sec
cut: exit_status= all
no cut on effcut
no WN selection
LONG: 126848 jobs,<>=2.6 ± 0.0 RMS=0.8
10^4
10^3
10^2
0
1
2
3
4
5
6
7
8

Que souhaitent les VO ?

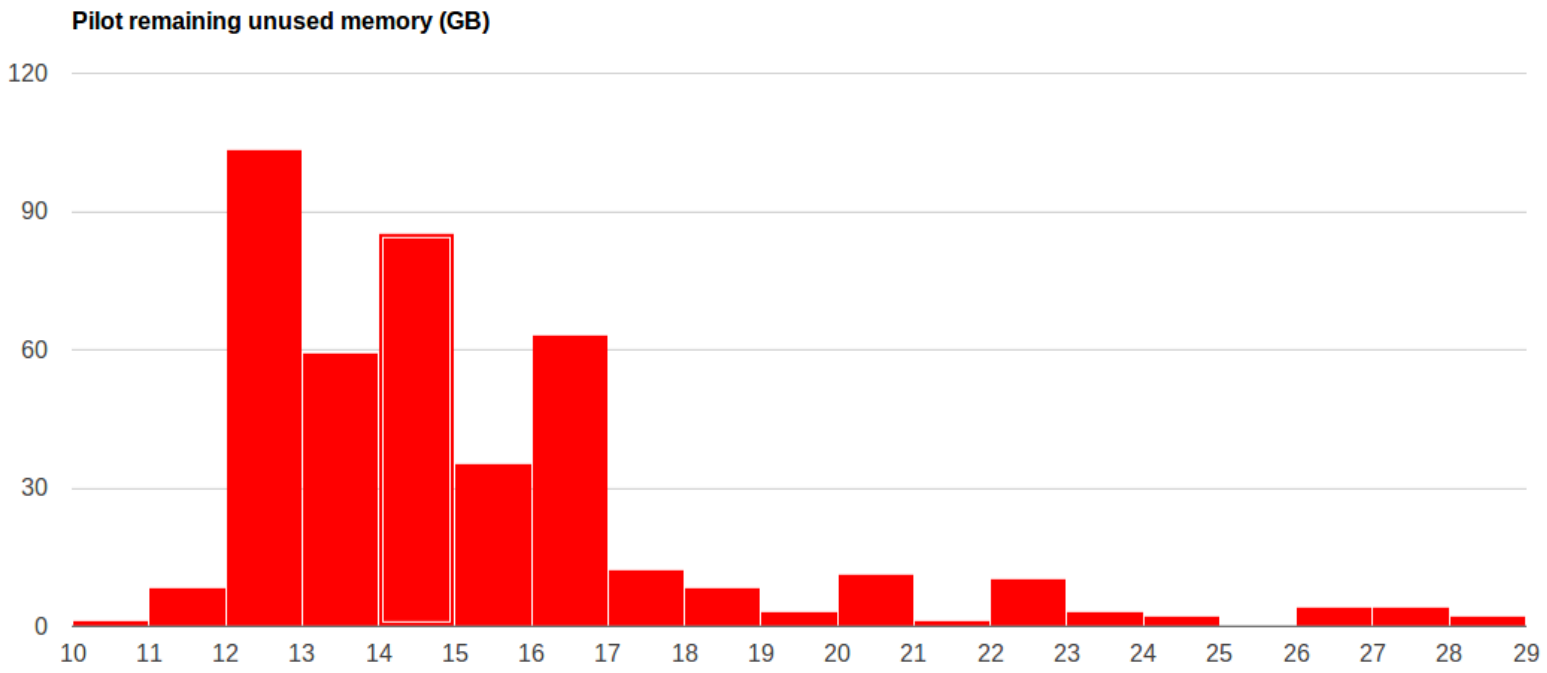
- CMS & LHCb
  - Pas de limitation sur memoire : OK
  - Si limitation, préférer cgroups
  - Pas de limite VMEM, pas de SWAP
- ALICE
  - Comme veut le site mais pas VMEM
- ATLAS
  - PSS idéal
  - Particulièrement intéressant dans certains cas

Monitoring ALICE



Monitoring CMS

- CMS : memoire ‘inutilisee’ par job



Bilan

- On peut mesurer des choses sur nos sites
  - Outils locaux
  - Dashboards VO
- Outils communs vs diversite
  - Pas de dashboard commun
  - Configurations différentes selon sites
  - VO supportees vs LRMS vs pratiques locales au site

- Peut-on / doit-on converger sur ces questions ?
  - Prix du GB/slot de RAM ?
  - Besoins des sites ?
- Etude préliminaire
  - Peut être poursuivie, ou pas.

Backup