DIRAC for biomed applications

Tristan Glatard
Université de Lyon, CNRS, INSERM, CREATIS, Villeurbanne

with contributions of

Sophie Gallina
Diala Abu Awad
Maxime Pauwels

30.10.2012

biomed VO

World-wide catch-all VO for Life-Sciences

~300 users, 20 countries
~150 sites
Medical imaging, bioinformatics, drug discovery

Mostly supported by opportunistic resources

Very few dedicated sites
Long queuing delays
Volunteer support (and VO admin)

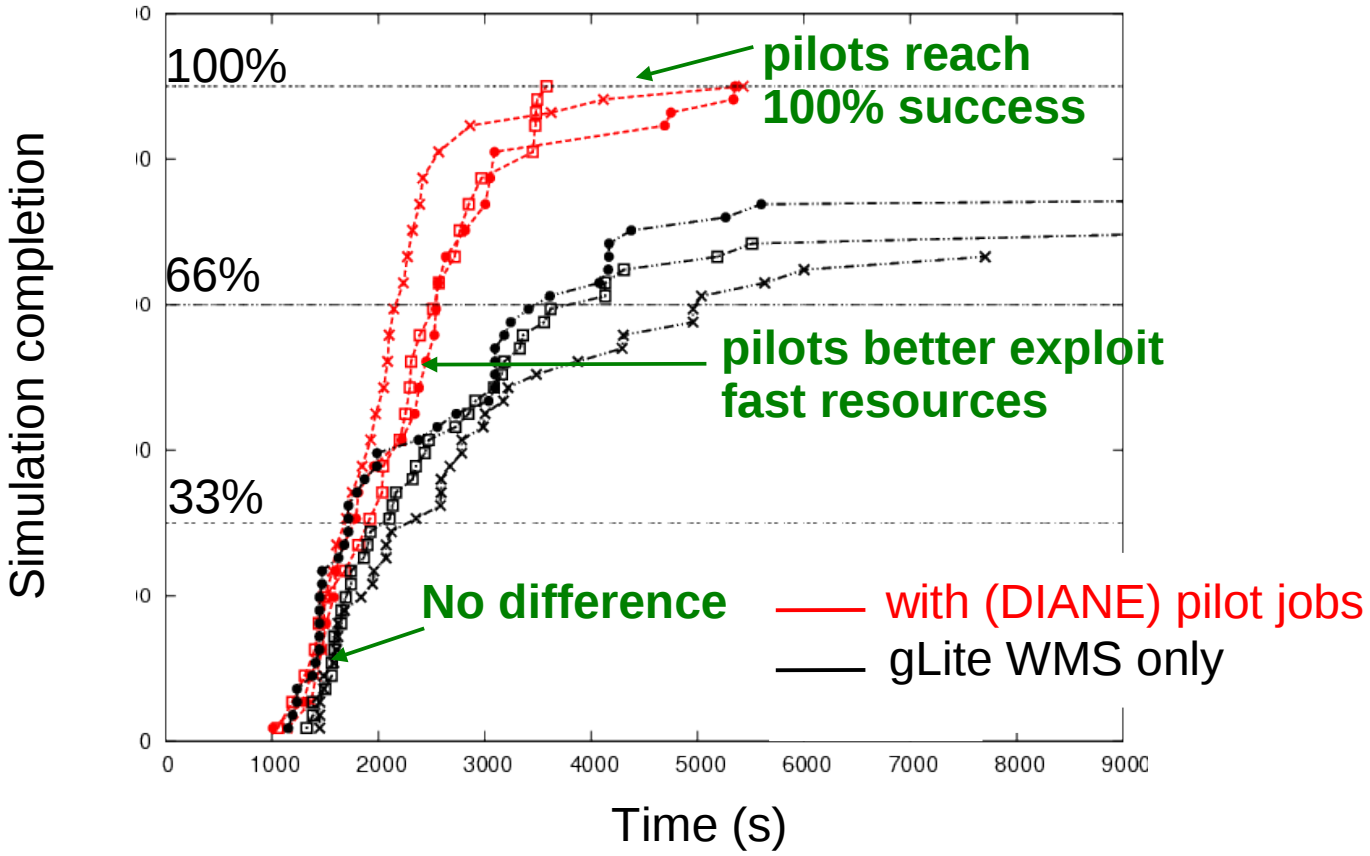
Accessed with heterogeneous tools

Portals, frameworks, workflow engines, gLite clients, pilot-job systems, etc

http://lsgc.org
biomed-users@googlegroups.com
biomed-technical-support@googlegroups.com

Life Science Grid Community

A clear need for pilot jobs



Motivations for using DIRAC in biomed

Facilitate installation of grid clients

Improve usage of resources

Reduce queuing time for short jobs

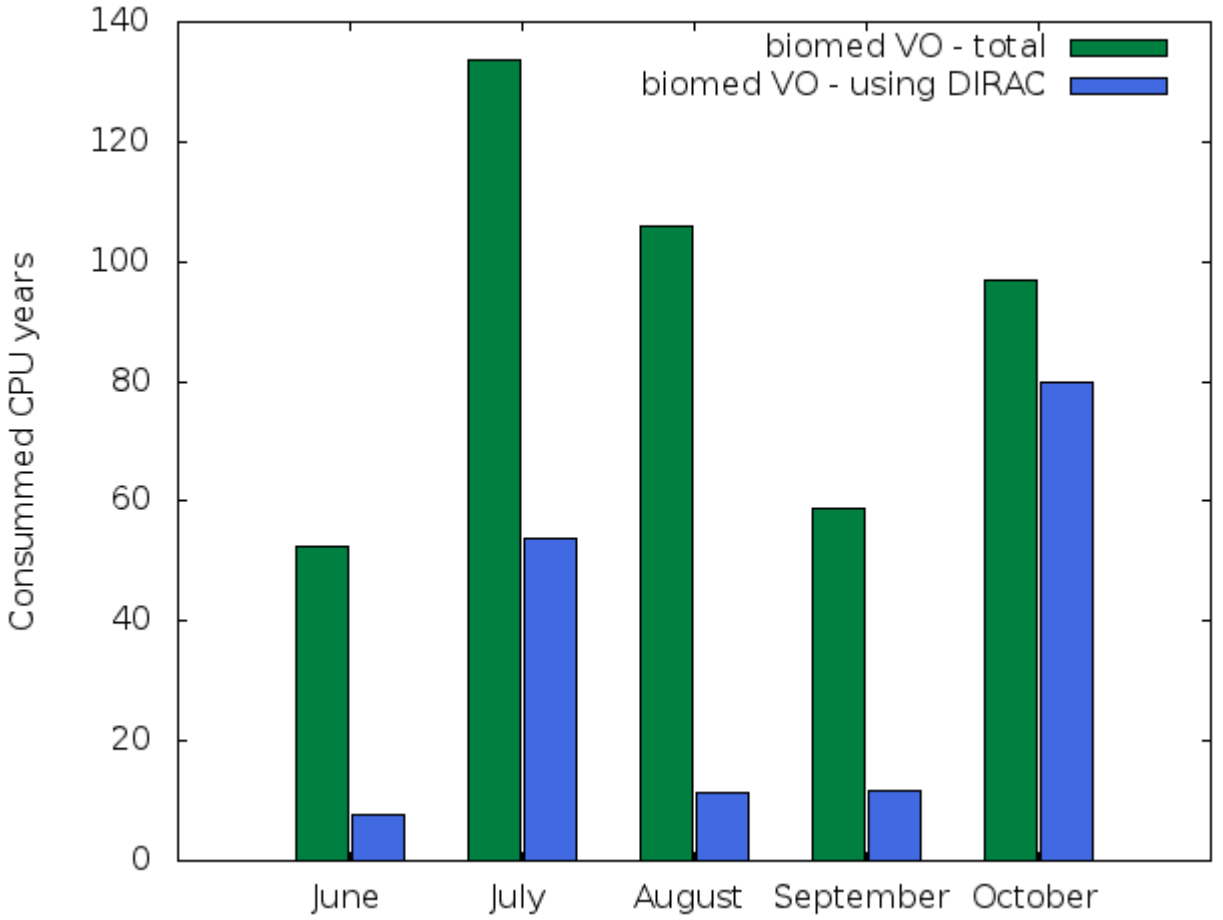
Harmonize pilot-job frameworks in the VO

Accesses various types of resources (not exploited yet)

Find an alternate to LFC (not used yet)

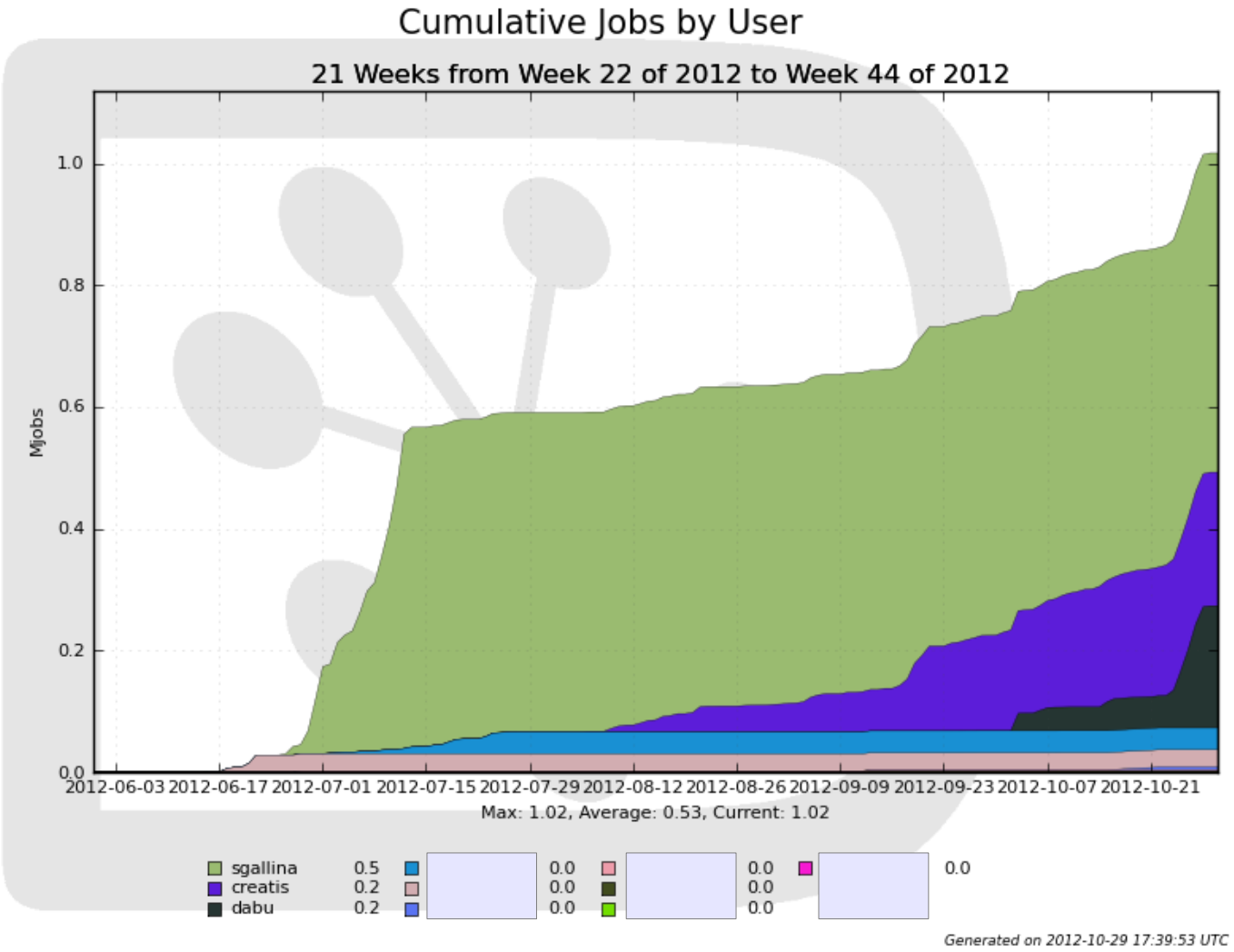
DIRAC in biomed

DIRAC instance provided by France-Grilles since June 2012



source: https://accounting.egi.eu ; https://dirac.france-grilles.fr

DIRAC users in biomed



Reproductive strategies, demography and mutational meltdown

Journées scientifiques mésocentres et France Grilles 2012

Diala Abu Awad[12], Sophie Gallina[1], Cyrille Bonamy[3], Sylvain Billiard[1]

[1]UMR-CNRS 8198

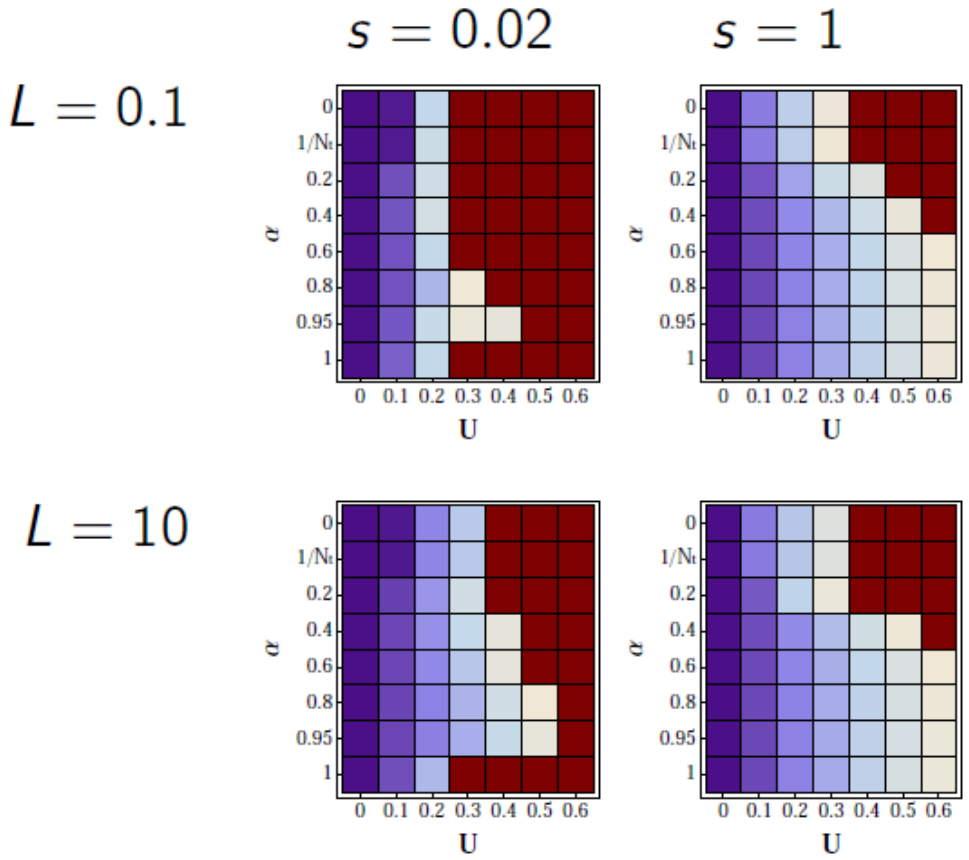
[2]Chaire de Modélisation
Mathématique et Biodiversité

[3]CRI Université de Lille 1

1er Octobre 2012

Effet du système de reproduction

Taille de population à l'équilibre



- Effet des paramètres génétiques sur la viabilité
- Effet du système de reproduction :
  - Allo-fécondation rarement favorable
  - Avantage des taux d'α intermédiaires

Détection a posteriori de structure génétique des populations hiérarchisée

Maxime Pauwels[1], Adeline Coorneart[1], Sophie Gallina[1], Cyrille Bonamy[2], Jean-François Arnaud[1]

maxime.pauwels@univ-lille1.fr adelinecoornaert@yahoo.fr sophie.gallina@univ-lille1.fr cyrille.bonamy@univ-lille1.fr jean-francois.arnaud@univ-lille1.fr

[1]: Laboratoire GEPV - UMR CNRS 8198
Université Lille1 Bât. SN2
59655 Villeneuve d'Ascq Cedex– France

[2]: Centre de ressources informatiques (CRI)
Université Lille 1
59655 Villeneuve d'Ascq Cedex– France

ENJEUX SCIENTIFIQUES

La génétique des populations s'intéresse à la distribution de la diversité génétique à l'intérieur des espèces biologiques. Elle cherche notamment à identifier des sous-ensembles, appelés populations, entre lesquels les échanges génétiques sont réduits (Hartl & Clark, 2007). La particularité de ces sous-ensembles est de présenter des patrimoine génétiques différents (**Figure 1**). Identifier ces groupes au sein d'espèces biologiques d'intérêt est un enjeu majeur lorsqu'il s'agit, par exemple, de définir des unités sur lesquelles opérer dans le cadre d'un programme de conservation.

Figure 1. L'espèce humaine *Homo sapiens* peut être divisée en un certain nombre de populations génétiquement différenciées. La société américaine de biotechnologie DNA Tribes propose d'identifier, en utilisant quelques marqueurs moléculaires sur un échantillon de votre ADN, l'origine géographique de vos ancêtres.

Plusieurs outils informatiques, dits de regroupement, utilisant la statistique bayésienne sont aujourd'hui disponibles pour déterminer *a posteriori* le nombre et les limites des populations à partir de données de génotypage moléculaire d'un échantillon d'individus.

Nous avons testé l'efficacité d'un de ces outils, implémenté dans le logiciel STRUCTURE (Pritchard *et al.*, 2000) en analysant des jeux de données simulées à l'aide du logiciel NEMO (Guillaume & Rougemont, 2006), sous deux modèles définissant une structuration hiérarchisée de la diversité génétique, c'est-à-dire lorsque un nombre déterminé de populations sont aussi regroupées en un nombre déterminé de groupes de populations génétiquement isolés (**Figure 2**).

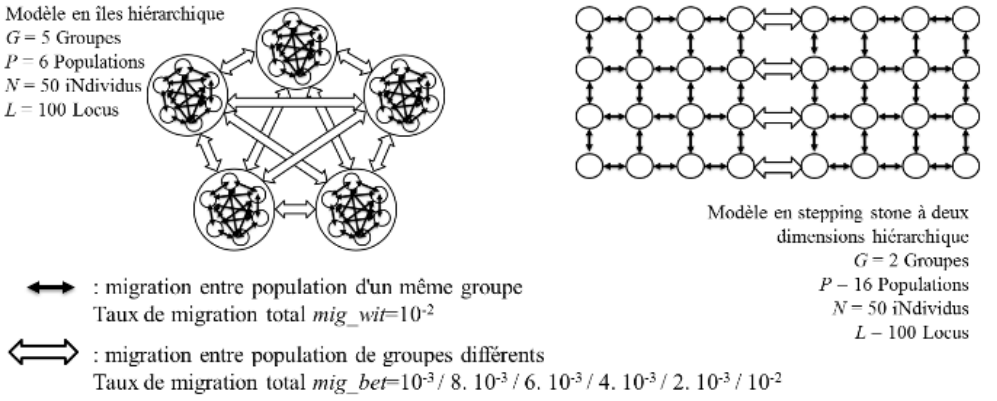
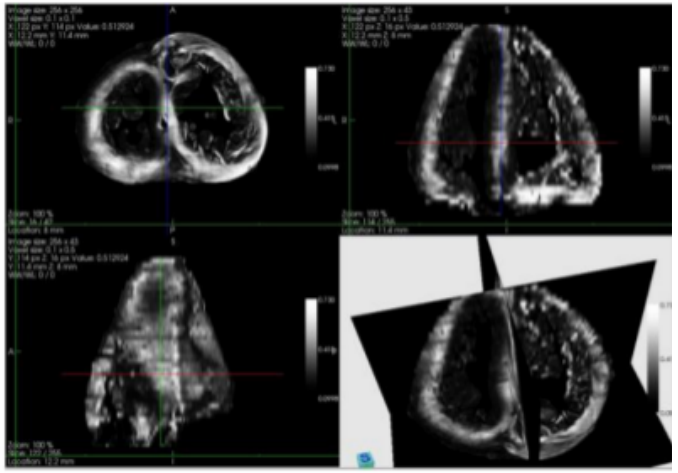
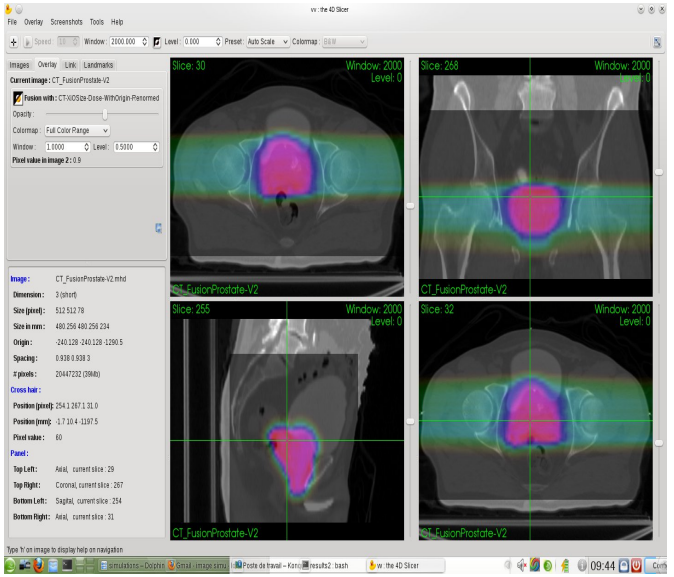


Figure 2. Modèles de simulation utilisés pour générer les données analysées. Le modèle en îles hiérarchique, à gauche, comprend 5 groupes (G, grands cercles) de 6 populations (P, petits cercles) comprenant chacune 50 individus (N). Le modèle en stepping stone à deux dimensions (à droite) comprend 2 groupes de 16 populations (petits cercles) comprenant chacune 50 individus. Dans les deux modèles, les groupes sont définis par des taux de migrations entre populations de groupes différents (flèches blanches) inférieurs aux taux de migration entre populations d'un même groupe (flèches noires). Le nombre de locus simulés (L) est égal à 100.

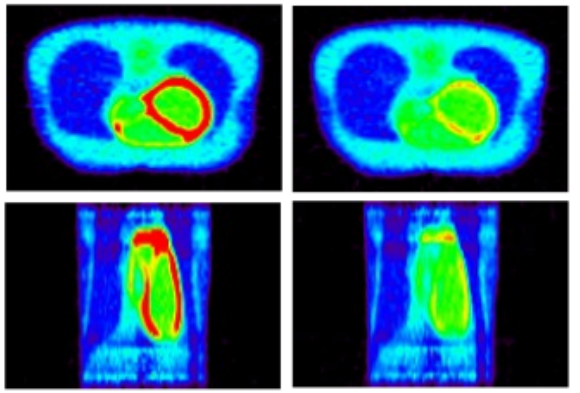
Medical simulation at CREATIS



Simulated diffusion weighted images
[L. Wang, Y. Zhu, I. Magnin] – 8 years

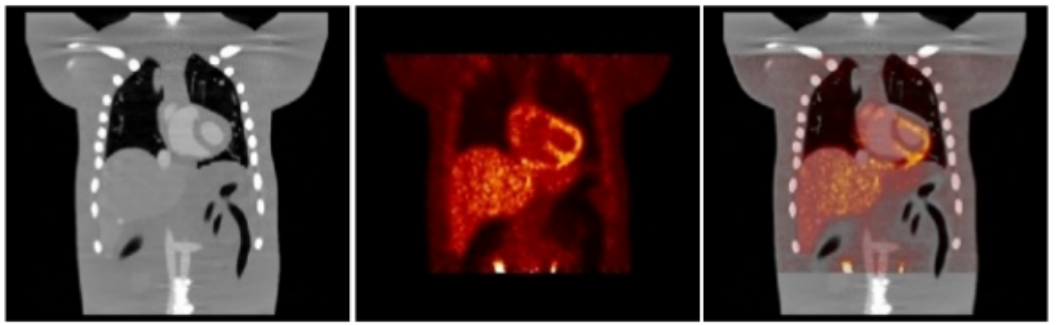


Treatment planning for prostate protontherapy.
[L. Grevillot, D. Sarrut] – 2 months



FDG-PET simulation of a healthy (left) and pathological heart 91 hours

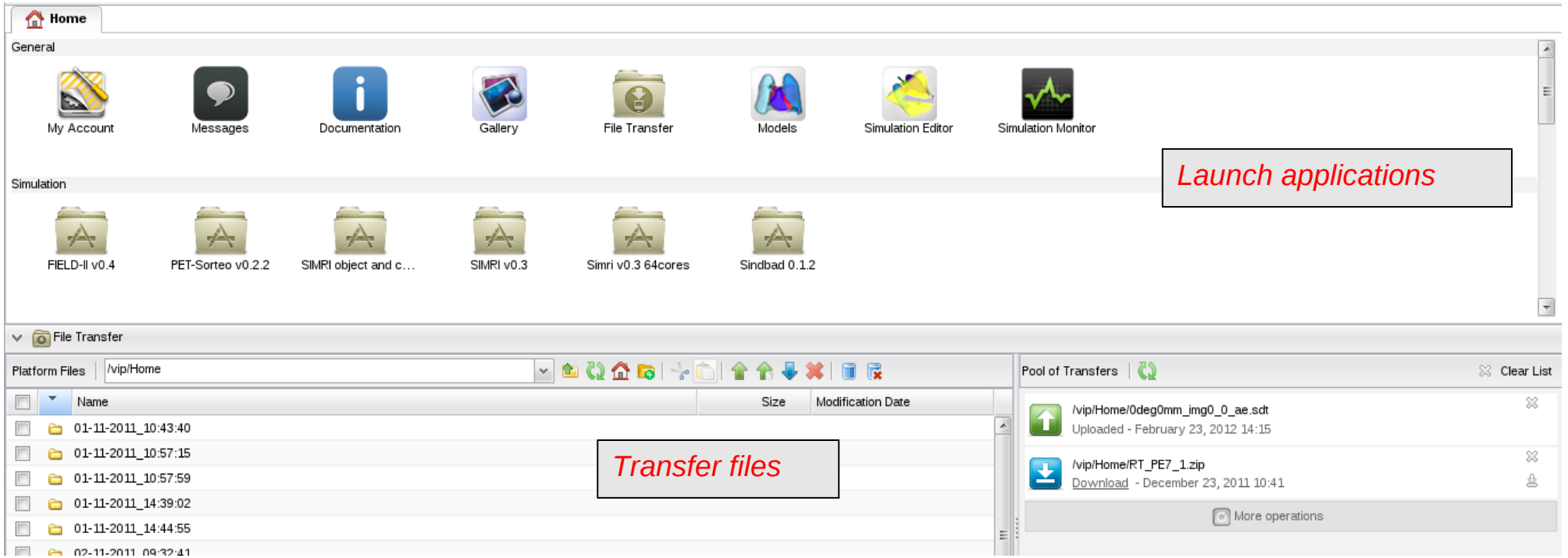
Echography simulation
[O. Bernard, M. Alessandrini] – 42 hours



From left to right: CT, PET and overlaid whole-body simulations 13 hours

Virtual Imaging Platform: web portal

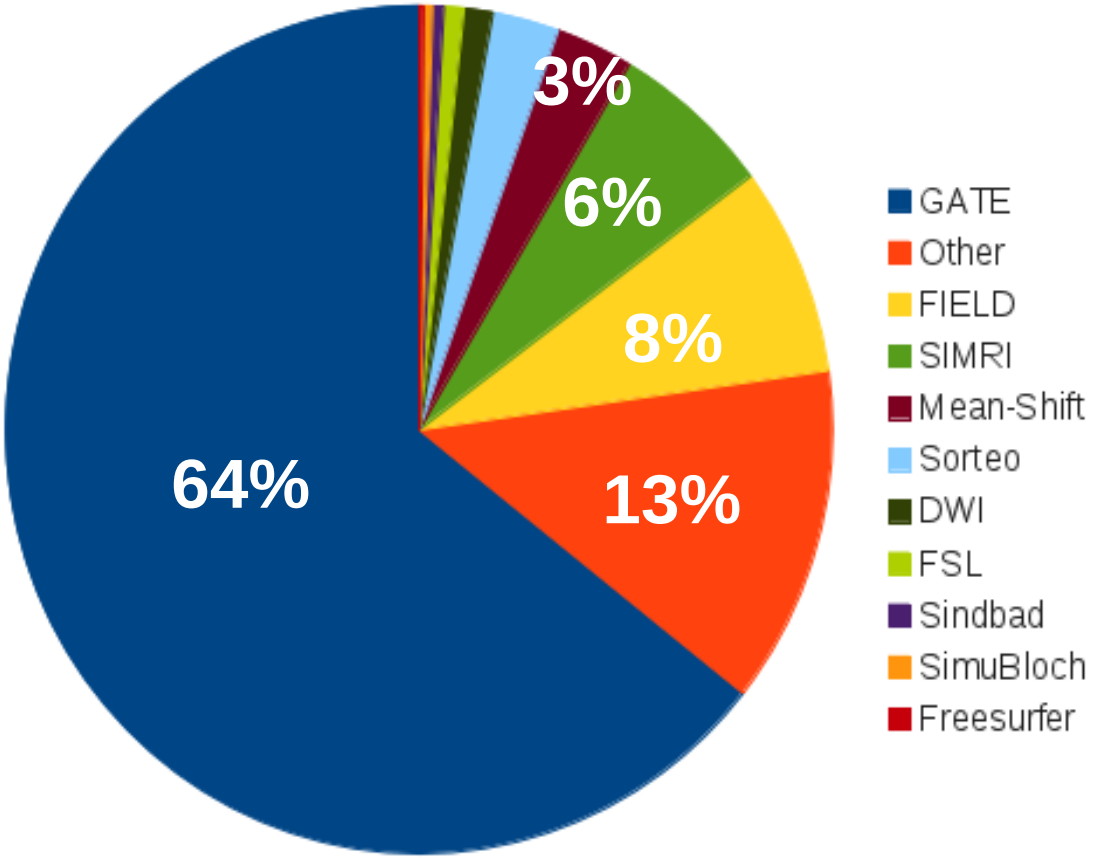
http://vip.creatis.insa-lyon.fr



Heterogeneous workload: heavy simulations, many short jobs, tests, trial-and-error, mistakes

VIP applications

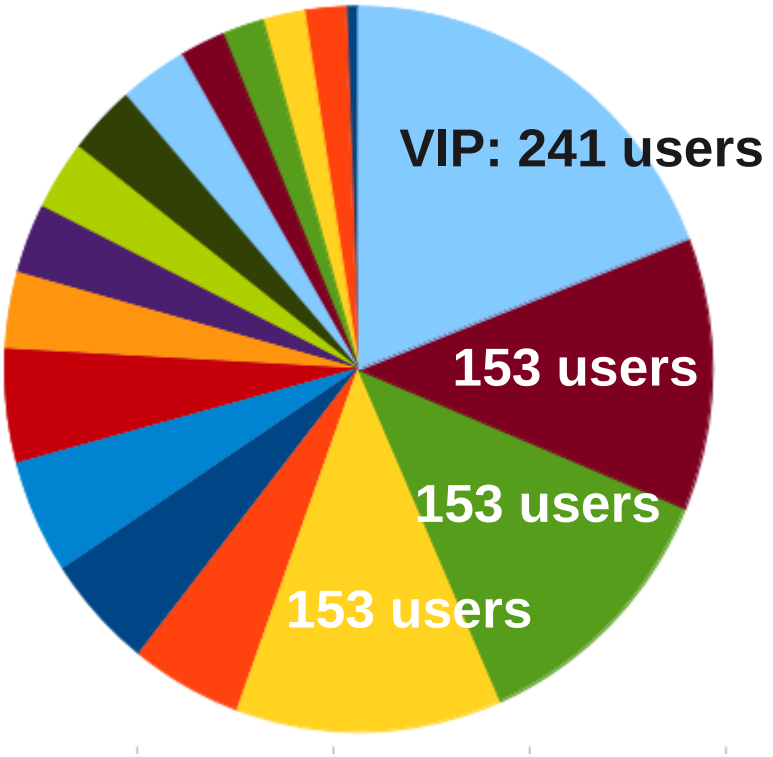
1155 executed simulations during the last year (~3/day)



Repartition of application executions in VIP (Nov 2011 – Oct 2012)

VIP users

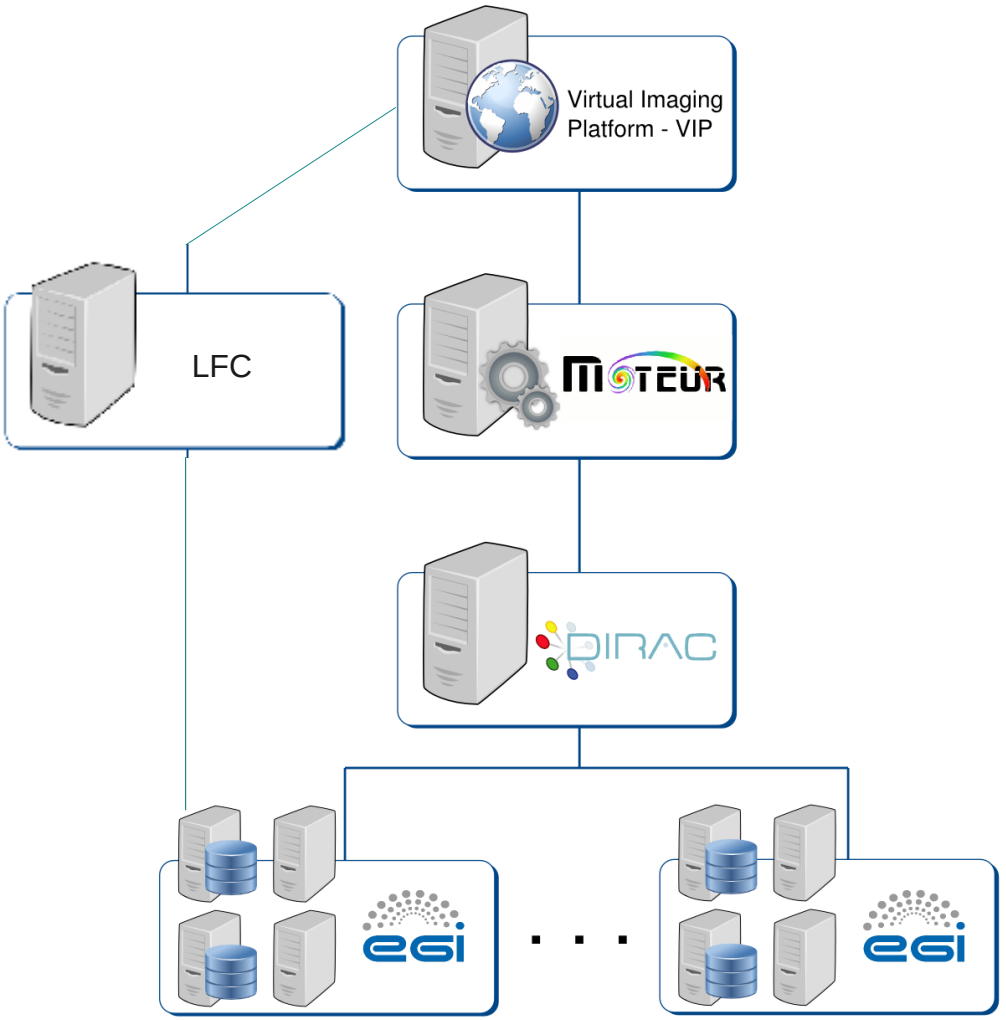
263 registered users from 31 countries



Repartition of portal users on EGI

(Source: **https://wiki.egi.eu/wiki/EGI_robot_certificate_users**)

Virtual Imaging Platform: architecture



Web portal with robot certificate

File transfers, user/group/application management

Workflow engine

Generate jobs, (re-)submit, monitor, replicate

DIRAC

Resource provisioning, job scheduling

Grid resources

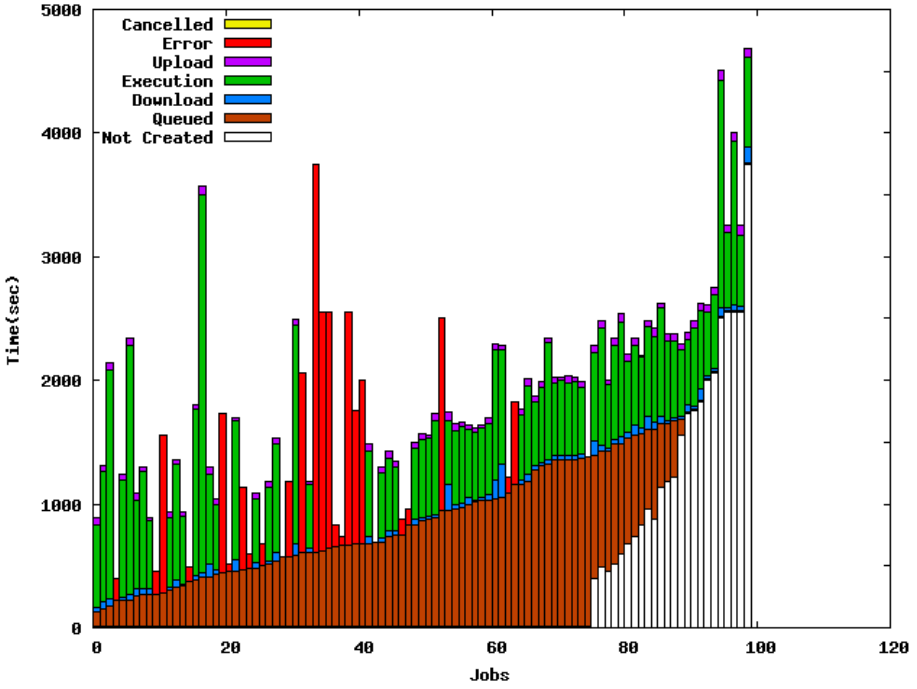
biomed VO

[Camarasu-Pop *et al, 2010*]

Improved load-balancing for Monte-Carlo applications

Static

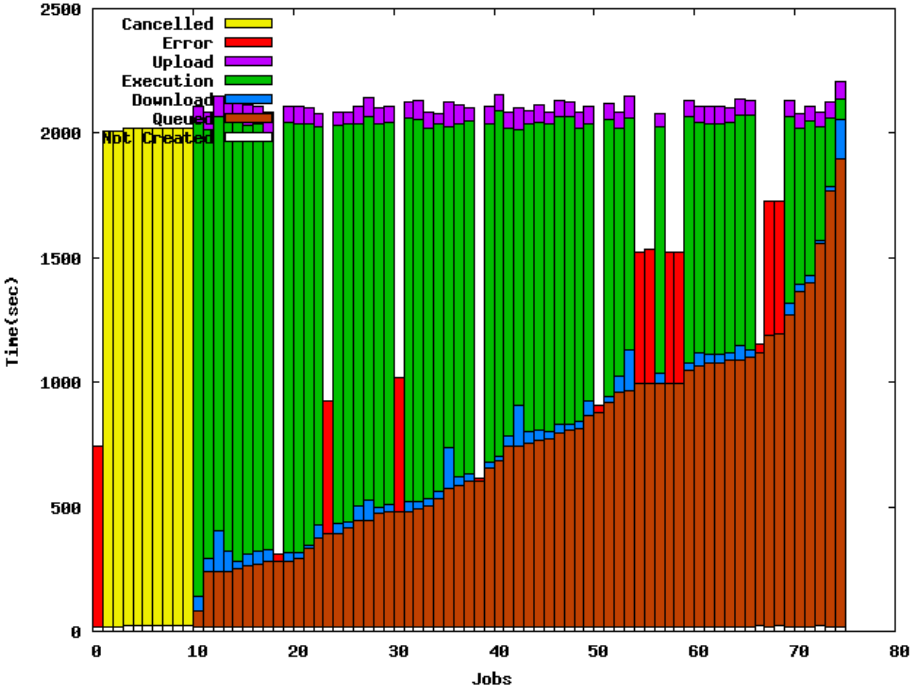
Worker:
Simule P/n events



Dynamic

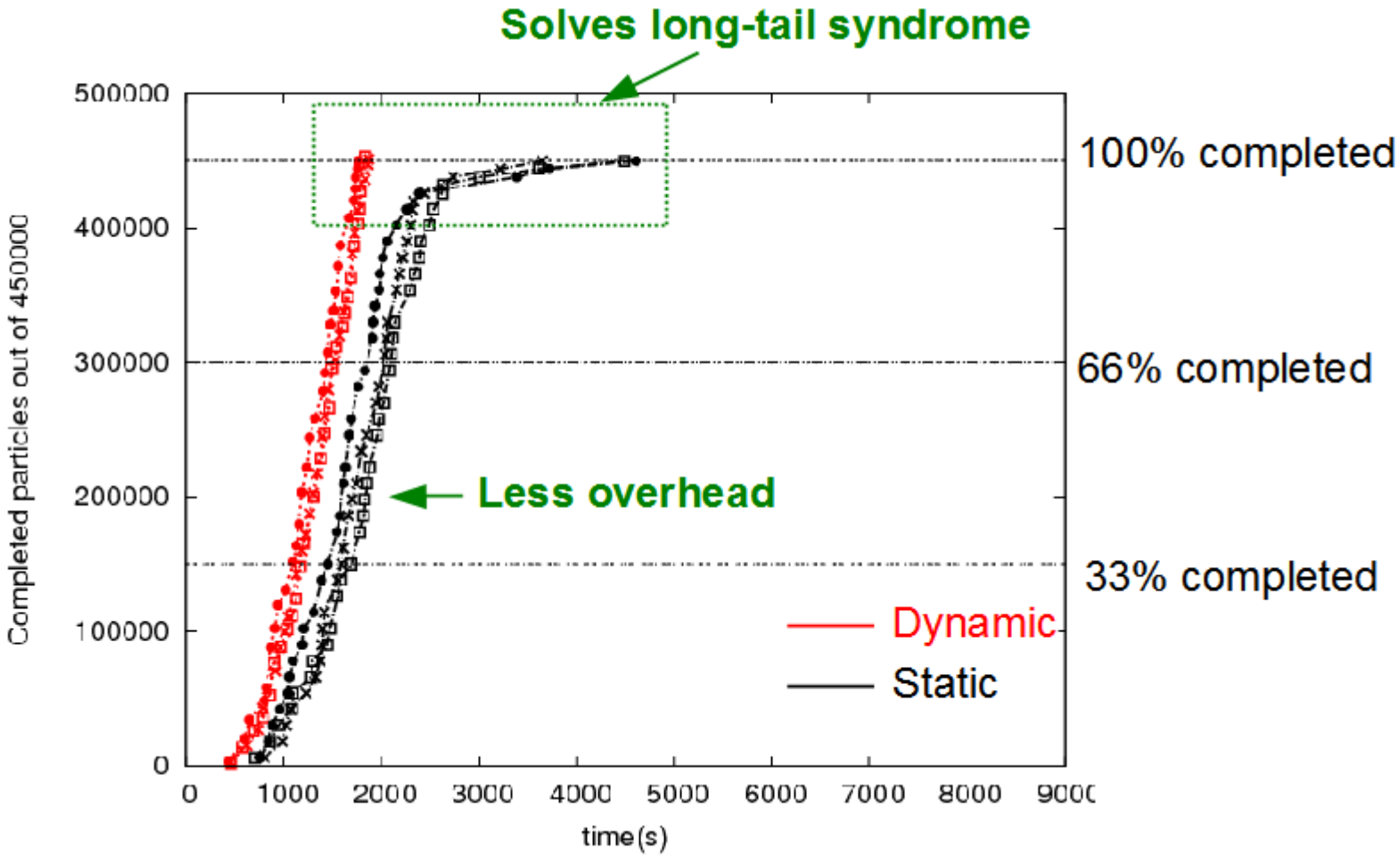
Worker:
While “stop” not received
Simulate 1 event
End while

Master:
While p ≠ P
p ←# simulated events
End while
Stop all workers



[Camarasu-Pop *et al, 2010*]

Dynamic parallelization (results)



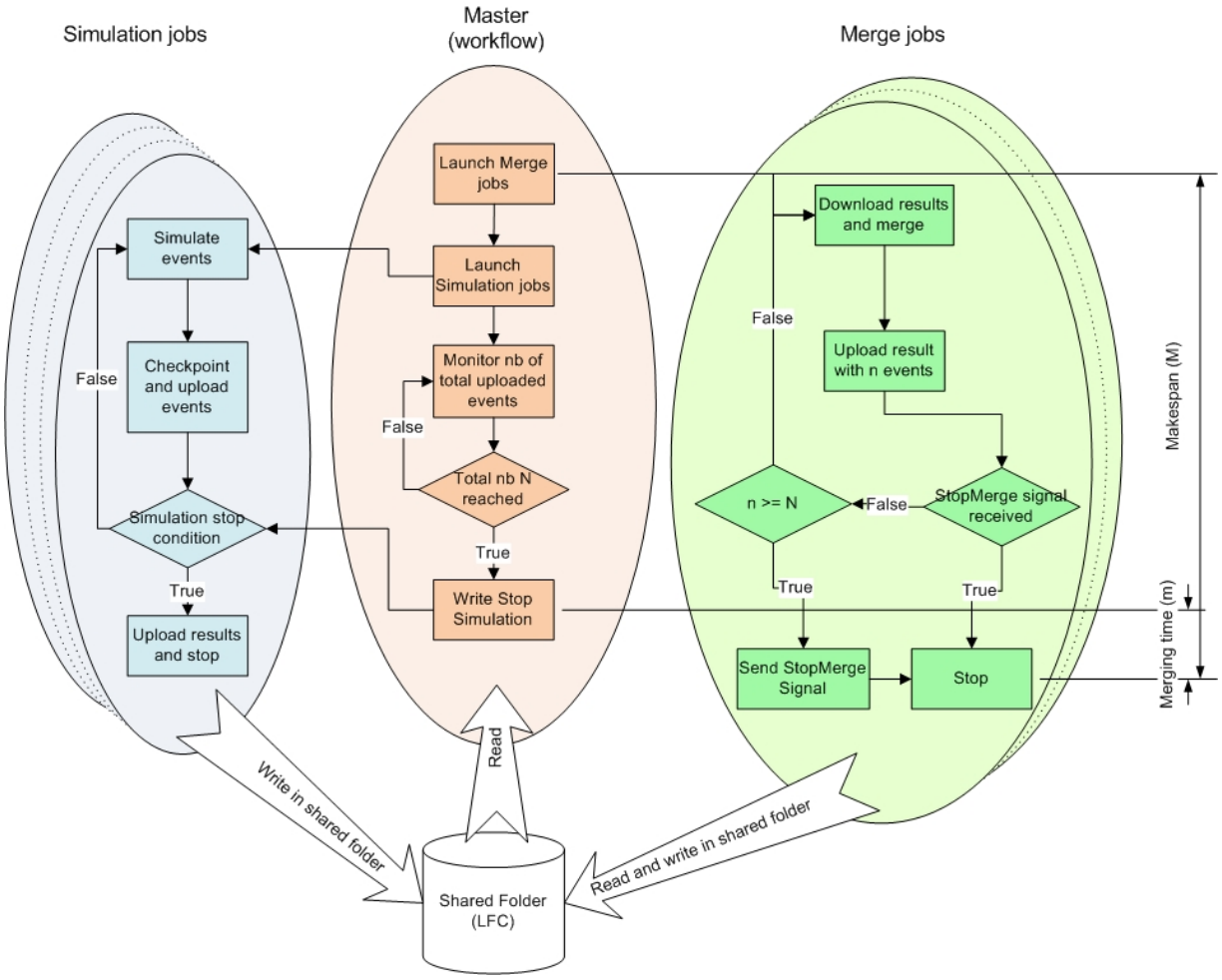
[Camarasu-Pop *et al, 2012*]

Merging of partial results for Monte-Carlo simulations

Multiple parallel merge tasks
(merge is commutative and associative)

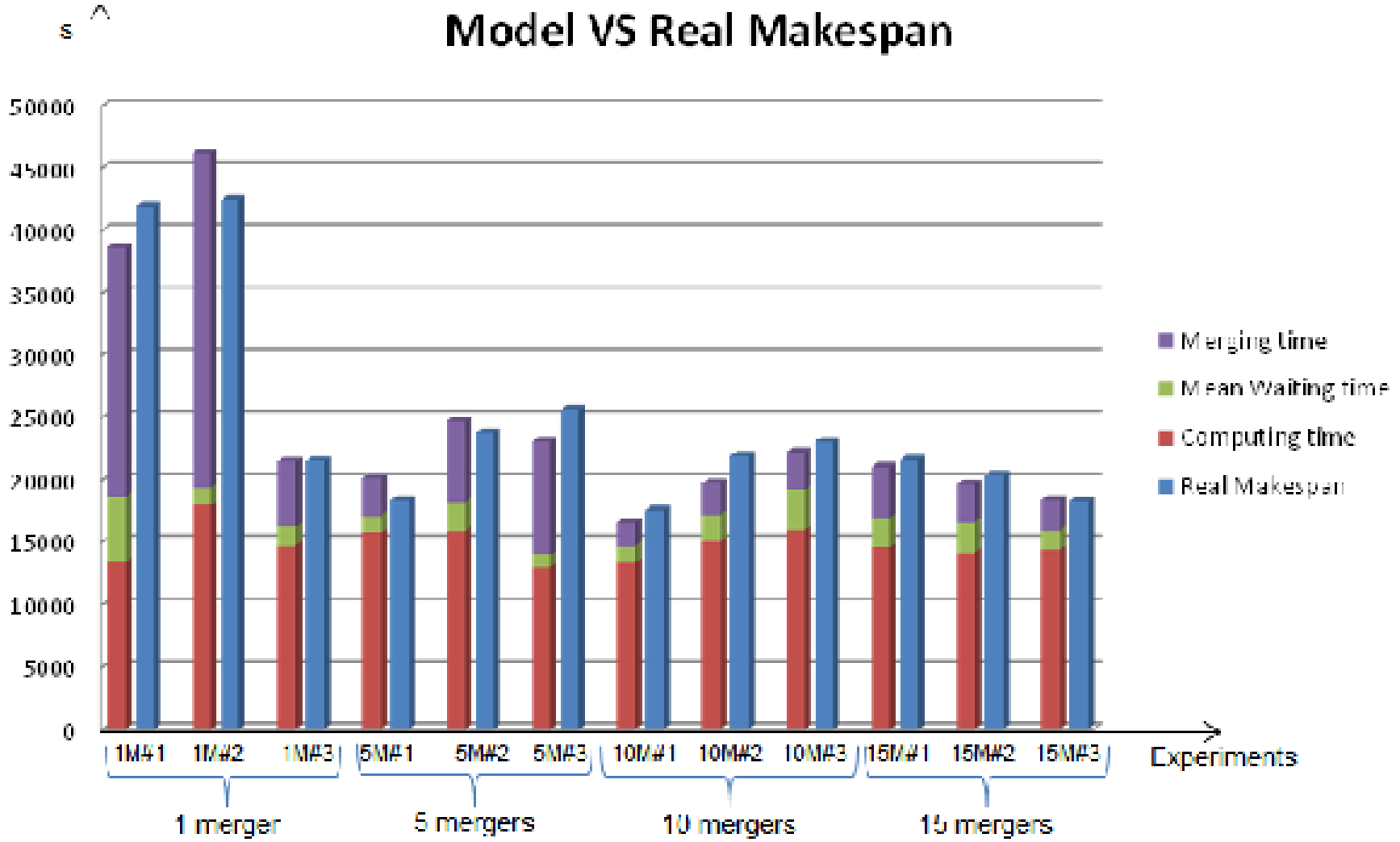
Incremental merging
From the beginning of the simulation; requires checkpointing

Checkpointing
Frequency adjusted to merging throughput



[Camarasu-Pop *et al, 2012*]

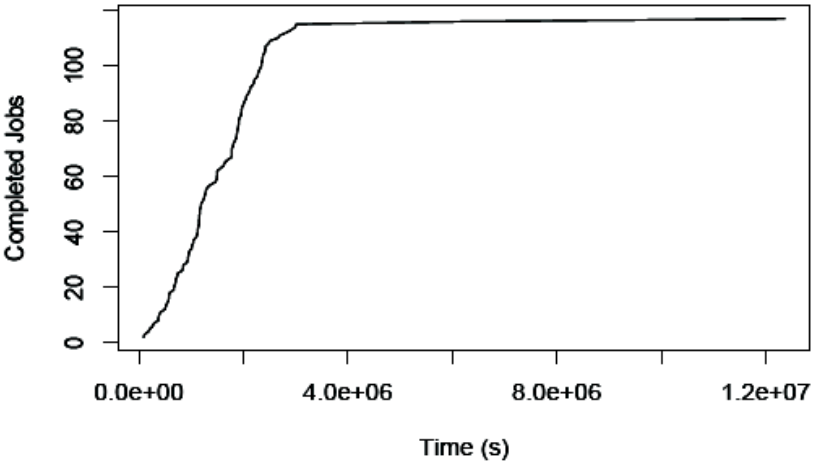
Parallel merging (results)



[Ferreira da Silva *et al, 2012*]

Task replication (for all simulations)

Problem: a few tasks delay the execution



Detection: task duration VS median

$$p(t_i, \tilde{t}) = \frac{t_i}{\tilde{t} + t_i}$$

Action: replicate these tasks

Warning: don't replicate if queued replicas cancel slow replicas

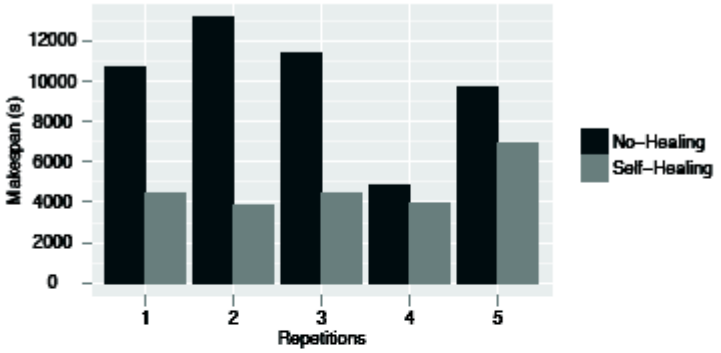
Input: Set of replicas R of a task i

01. rep = true
02. **for** $r \in R$ **do**
03. **for** $j \in R$ **do**
04. **if** $p(t_r, t_j) > \tau$ **and** j is a step further than r **then**
05. abort r
06. **done**
07. **if** r is started **and** $p(t_r, \tilde{t}) \leq \tau$ **then**
08. rep = false
09. **done**
10. **if** rep == true **then**
11. replicate r

[Ferreira da Silva *et al, 2012*]

Task replication (results)

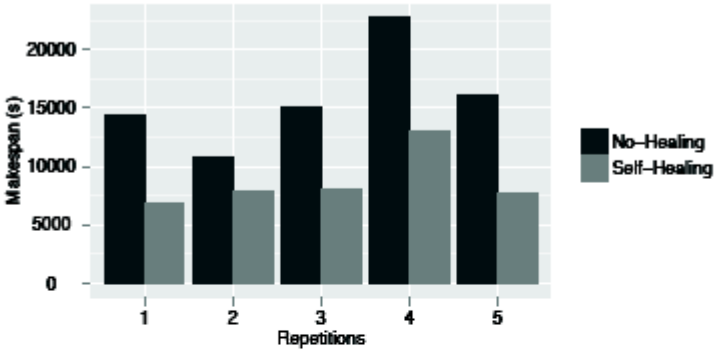
FIELD-II/pasa



speeds up execution up to 4

Repetition	w
1	–0.10
2	–0.15
3	–0.09
4	0.05
5	–0.26

Mean-Shift/hs3



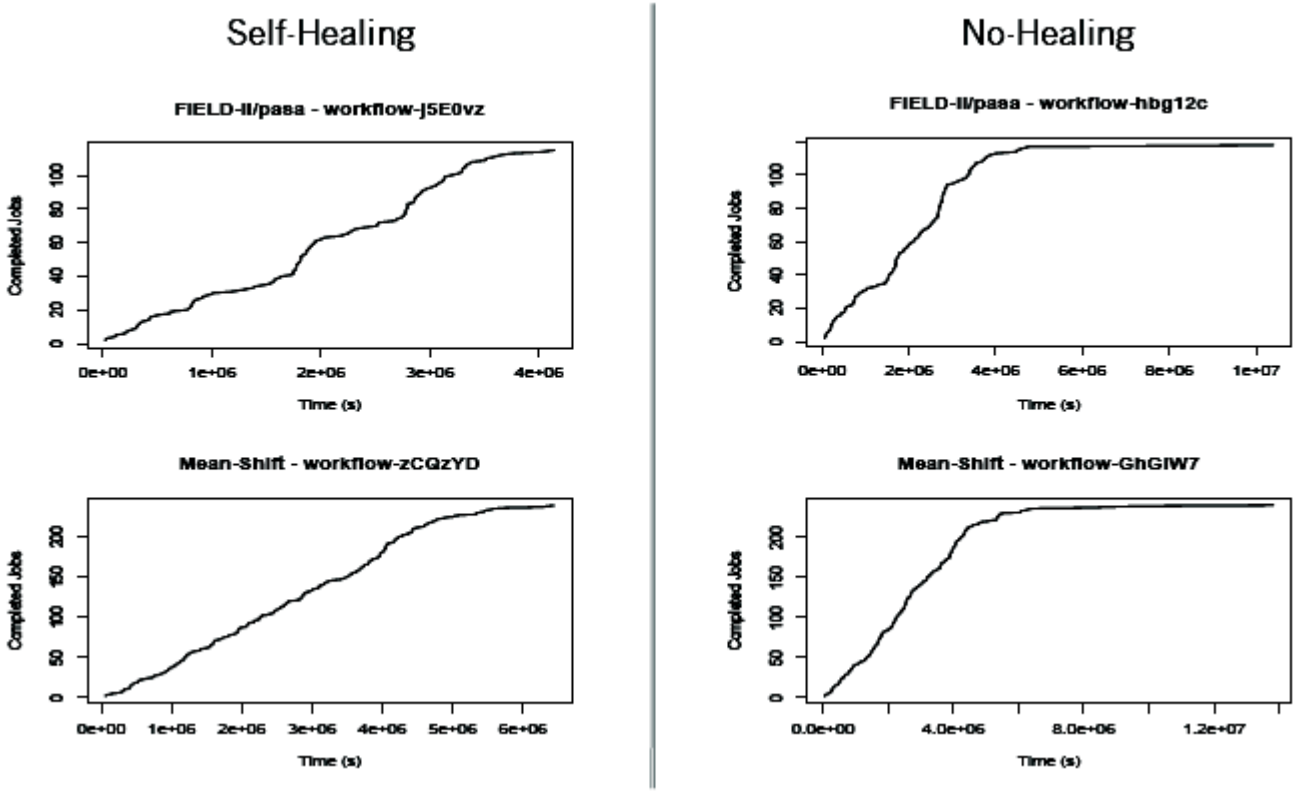
speeds up execution up to 2.6

Repetition	w
1	–0.02
2	–0.20
3	–0.02
4	–0.02
5	–0.01

Self-Healing process reduced resource consumption up to 26% when compared to the *No-Healing* execution

[Ferreira da Silva *et al, 2012*]

Task replication (results cont'd)



Conclusion: DIRAC in biomed

Allows fast startup of newcomers
The recommended solution to start in the VO

Improved performance compared to gLite WMS

Homogeneizes VO tools

Better usage of available resources

Aggregates non-grid resources
(desktop grids on-going)

Future work and interests

Test DFC

Java API

MPI

Conclusion: DIRAC in VIP

Efficient, robust, scalable service at low administration cost

Hardly any operational issue coming from DIRAC

Workflow manager on top

Task resubmission upon failure, replication, load-balancing

Merging partial results remains a problem

Looking into fairness among workflow executions

Thank you!

glatard@creatis.insa-lyon.fr

Creatis

Inserm