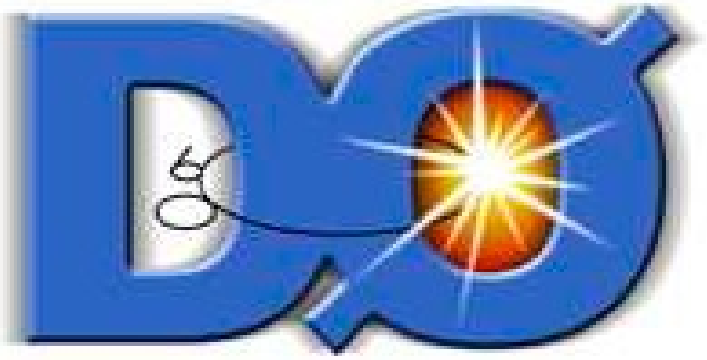
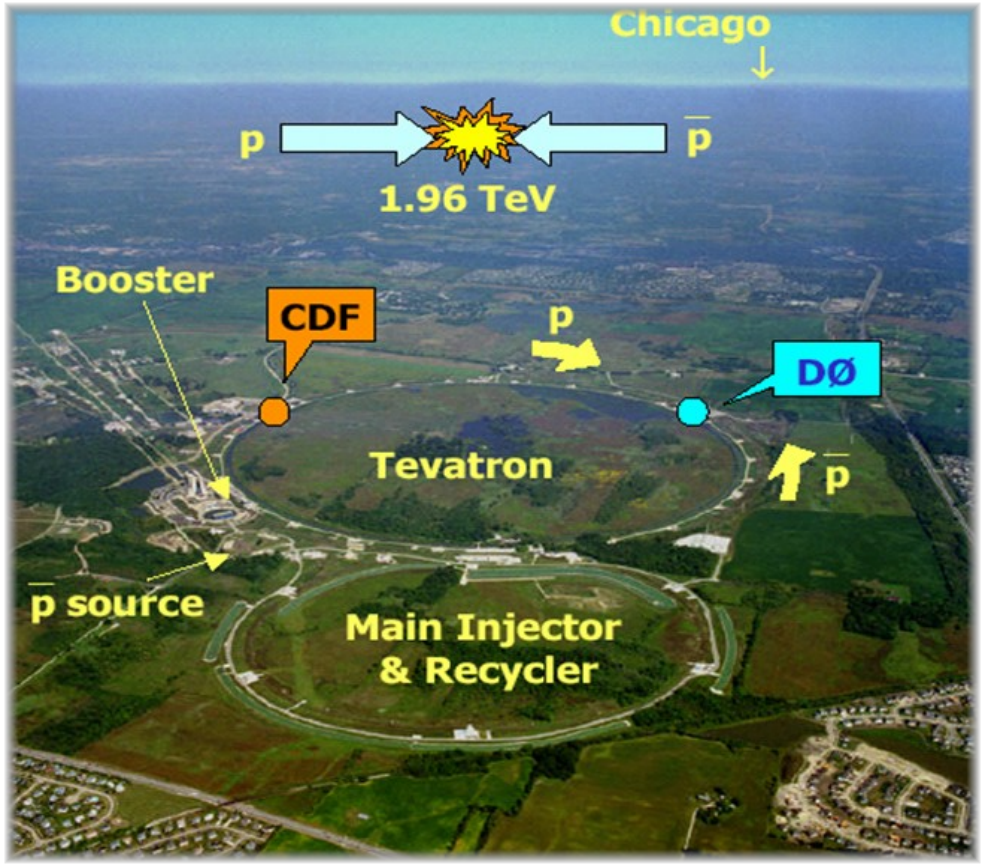


Expérience avec DØ

Jan Stark (LPSC Grenoble)

Atelier sur l'analyse au CC-IN2P3, 17 avril 2008

L'expérience DØ



Manip' en pleine prise de données.

Excellentes performances du Tevatron :

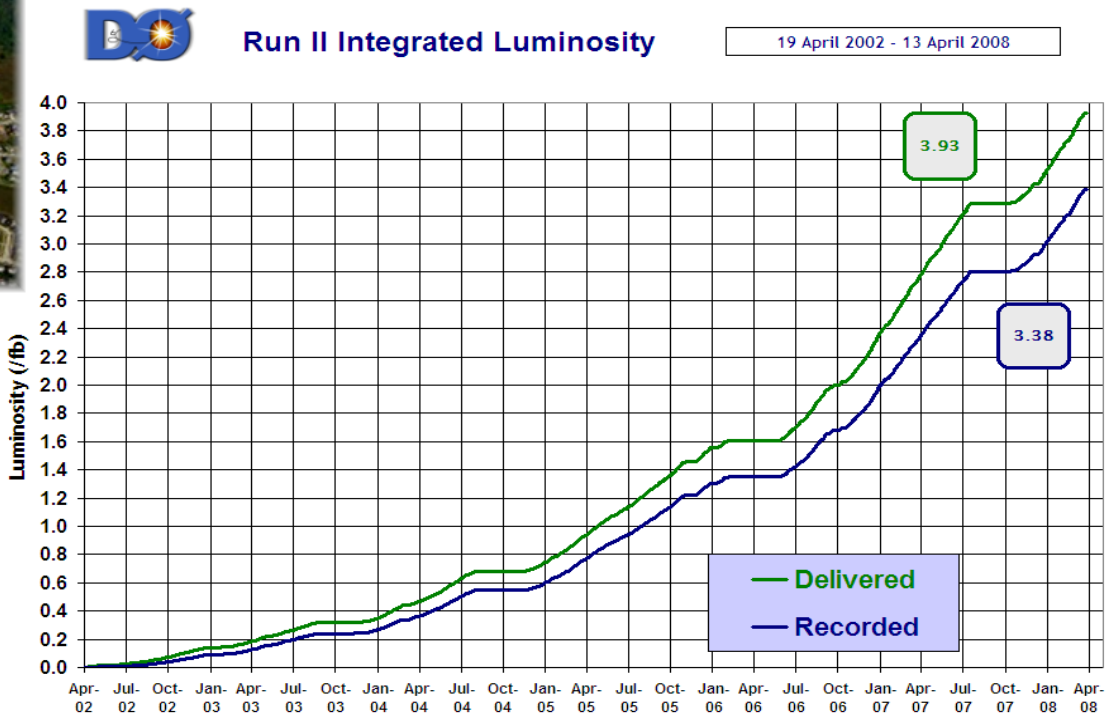
record de luminosité instantanée :

$312 * 10^{30}$ /cm^2/s (mars 2008)

rappel : design initial

$200 * 10^{30}$ /cm^2/s

Lumi intégrée : ~ 150 pb^{-1} / mois



Une véritable «usine à bosons» :

150 pb^{-1} / mois, c'est p.ex.

420 k W -> e ν

40 k Z -> e e

produits par mois

Volume de données, formats de données

RAW

Données brutes : 200 – 250 kB / évènement

Prise de données : 50 – 150 Hz , 5 – 6 M évènements / jour

10 – 30 MB / sec, O(1 TB / jour)

Total jusqu'à présent : ~ 4 milliards d'évènements, ~ 900 TB

reconstruction centralisée à FNAL

TMB

Données reconstruites : 80 – 150 kB / évènement

Format propre à DØ, écrit et lu par le «framework DØ».

Taille assez grande (p.ex. comparé aux RAW) : contient les objets reconstruits (traces, électrons) mais aussi de l'information de base comme l'énergie dans toutes les cellules du calo ou encore les *hits* sur les traces (=> permet une analyse détaillée, récalibrations, etc).

skimming TMB -> TMB ensuite *caffing* centralisés à FNAL

CAF

Format d'analyse : 35 – 60 kB / évènement

Basé sur Root, les objets reconstruits (traces, électrons, ...) sont représentés par des objets C++.

Les centres d'analyse extérieurs (comme p.ex. le CC-IN2P3) importent typiquement (p.ex. via bbftp dans HPSS) des *skims* sélectionnés en format TMB et/ou CAF.

DØ : installations de calcul à FNAL

Enstore

Stockage central

high-speed tape robots plus disk cache

total : ~ 3000 TB; écriture par mois : ~ 250 TB

Reco farm

Reconstruction centralisée des données

normalement la reconstruction se fait de façon centralisée à FNAL

mais : *reprocessings* à l'extérieur (c.f. transparent suivant)

CAB

Ferme d'analyse

permet un accès direct aux données pour les jobs des utilisateurs

mais aussi : *skimming* centralisé

production centralisé de fichiers CAF

reprocessings limités

De plus en plus intégré;
~ 8 THz en 2007

clued0

«Service interactif»

Ensemble des *desktops* Linux qui traînent sur le site de DØ (achetés par les différents instituts)

Plus un ensemble de serveurs NFS Linux «avec plein, plein de disques pas cher (ATA/SATA)» (également achetés par les différents instituts).

Accès au soft DØ via NFS.

La vaste majorité du travail d'analyse dans DØ se fait en interactif sur clued0 plus jobs *batch* sur CAB pour le *processing* lourd.

Reprocessing de données

Normalement, la reconstruction des données est faite de façon centralisée à FNAL .

Mais parfois il y a des situations où une bien meilleure compréhension du détecteur et/ou des algorithmes de reconstruction améliorés justifient une nouvelle reconstruction des anciennes données => «*reprocessing*».

Comme la reconstruction de DØ est très gourmande en CPU (p.ex. 110 GHz * sec / évènement à une luminosité de 200E30), FNAL n'a pas les ressources pour *reprocesser*.
=> sites externes.

Deux très grands *reprocessings* de DØ :

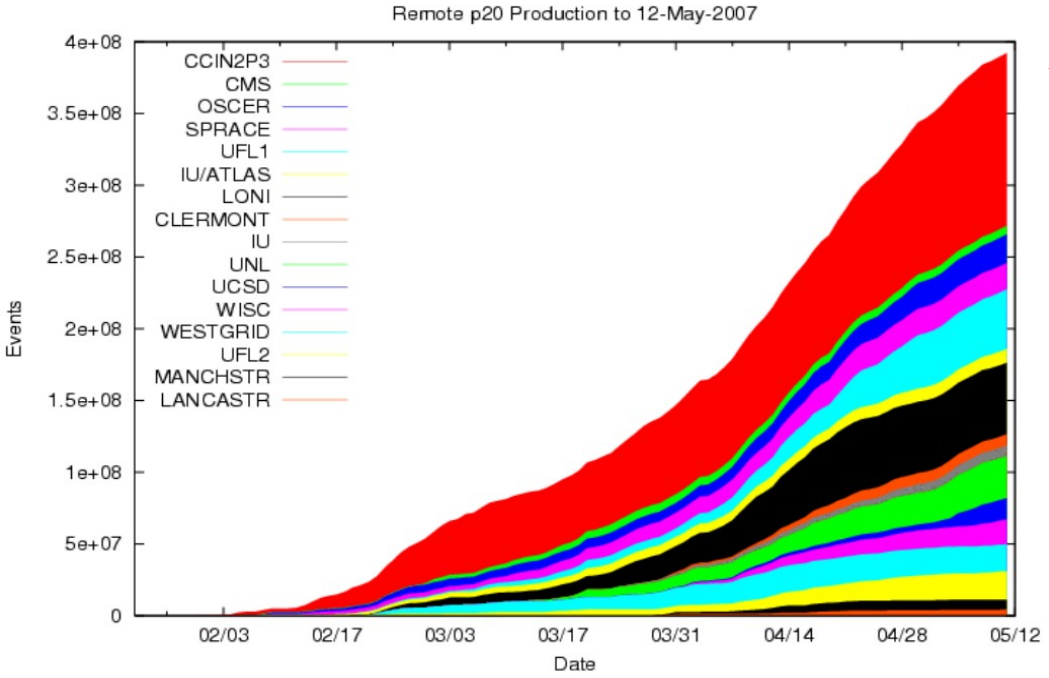
- *reprocessing* des premières données prises directement après l'*upgrade* du détecteur durant l'été 2006,
- *reprocessing* intégral des données en 2005.

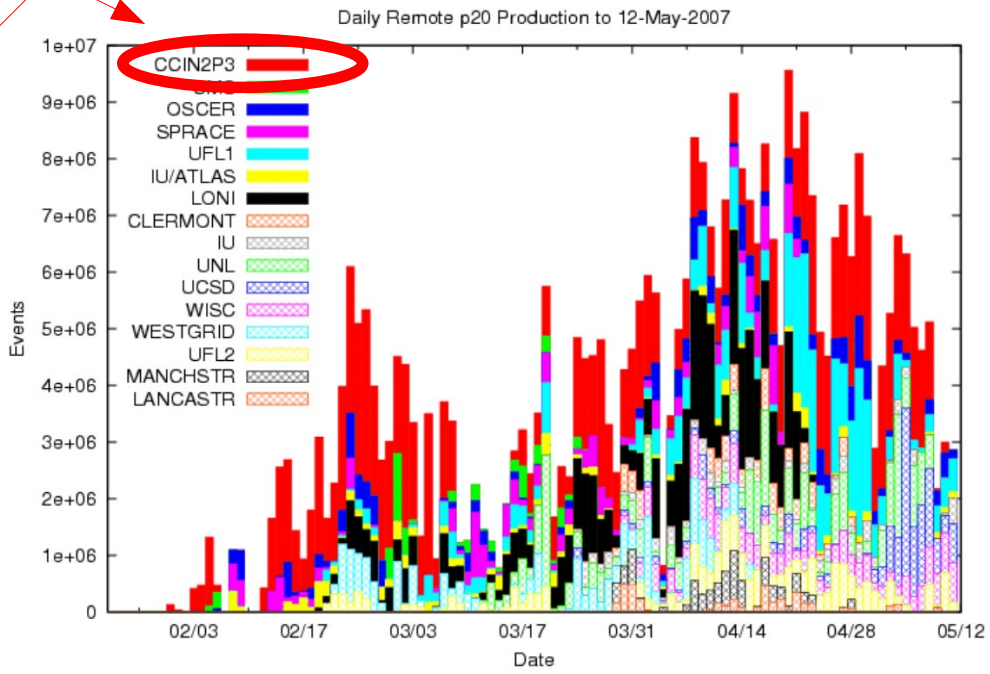
Reprocessing des premières données après upgrade été 2006

La tâche : Entrée : 90 TB données RAW
Sortie : 60 TB de fichiers TMB
500 années de CPU
à accomplir en trois mois (*conférence driven*)

Accomplie :

Contribution substantielle du CC.



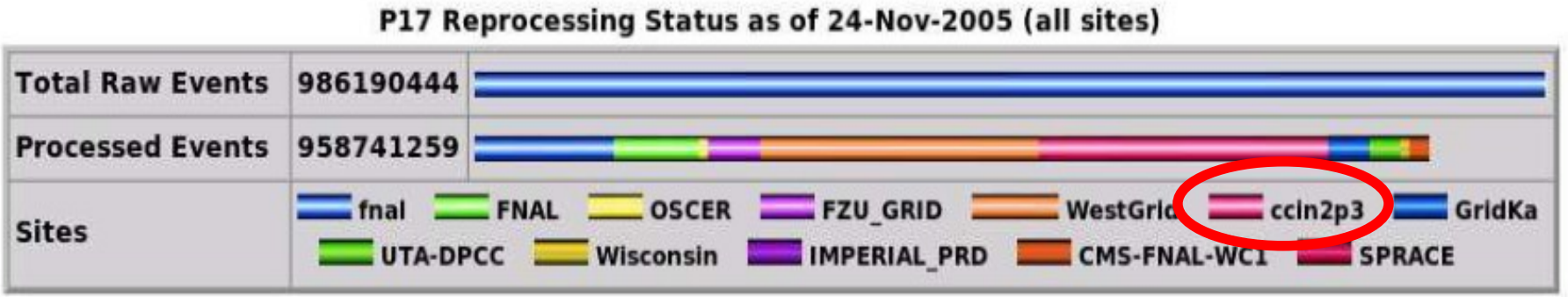


Reprocessing intégral en 2005

La tâche :

	2005 (p17)
Luminosity	$470\,\mathrm{pb}^{-1}$
Events	1G
Rawdata 250kB/Event	250TB
DSTs 150kB/Event	150TB
TMBs 70(20)kB/Event	70TB
Time $50s$/Event	20,000months
(on 1GHz Pentium III)	3400CPUs for 6mths
Remote processing	100%

Elle a été accomplie (en six mois environ) :



Contribution substantielle du CC.

Production Monte Carlo

Une compréhension détaillée des données requiert une simulation Monte Carlo précise et fiable.

Quand la production Monte Carlo prend du retard, les publications prennent du retard.

DØ utilise une simulation détaillée du détecteur basée sur Géant 3.

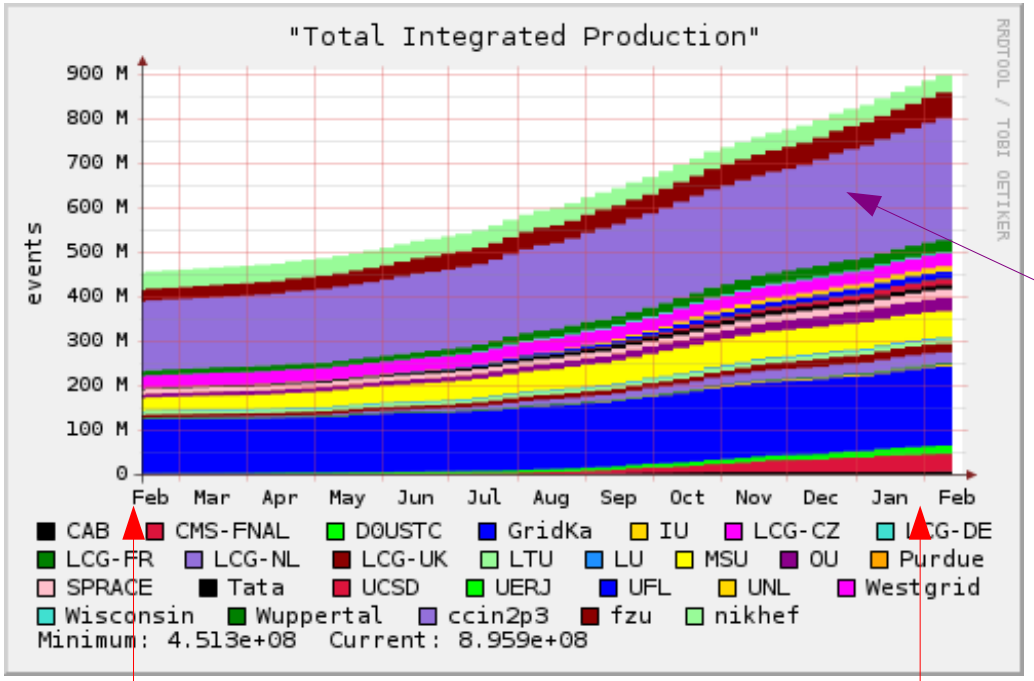
Pour une simulation réaliste du bruit dans le détecteur et des effets des interactions supplémentaires, des vrais données *zero bias* (déclenchement sur des croisements aléatoires) sont superposées au *hard scatter* simulé.

Temps de calcul par évènement simulé : ~240 GHz*sec (évènement complexe, style t-tbar)
(dominé par la partie Géant)

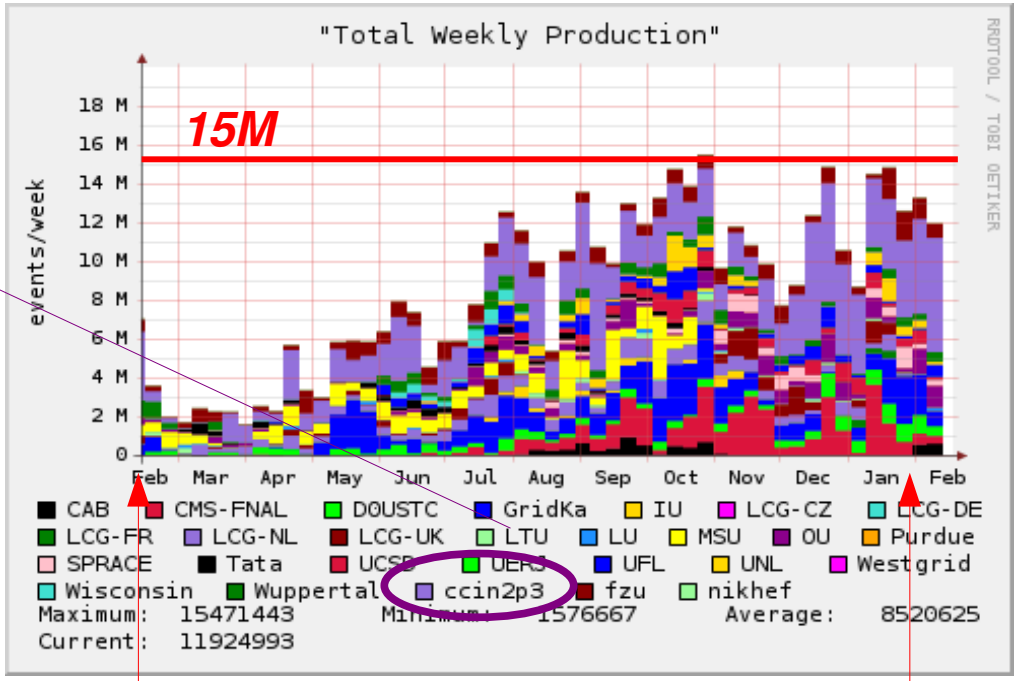
La quasi-totalité du Monte Carlo de DØ est produit *off-site*. Le CC-IN2P3 joue un rôle clé à la fois à cause de la quantité du MC produit et à cause de sa capacité à gérer des demandes non-standard (cf. transparents suivants).

Monte Carlo : taux de production

Le taux de production MC est un facteur limitant pour de nombreuses analyses de physique à DØ.



fév. 2007 fév. 2008



fév. 2007 fév. 2008

Volume total jusqu'à présent (fév. 2008) :

~ 1 milliard d'évènements

dont ~ 450 M entre fév. 2007 et fév. 2008

Taux de production :

maintenant à

10 – 15 M évènements / semaine

augmentation importante au cours des 12 mois derniers

Monte Carlo : requêtes «non-standard»

Beaucoup de sites différents contribuent à la production MC pour DØ (cf. transparent précédent).

Mais le CC-IN2P3 est **unique** en ce qui concerne sa capacité à stocker (de façon permanente, dans **HPSS**) toutes les étapes intermédiaires de la simulation (p.ex. *hard scatter* simulé avant superposition des *zero bias*).

=> permet des études systématiques avec différents types de *zero bias* (lumi ...),

=> permet de garder les fichiers intermédiaires précieux pour les productions très chères en CPU avec des versions spéciales (précises donc lentes) de Géant.

=> ...

«Les requêtes dites non-standard ne sont pas si non-standard», elles arrivent souvent et en volumes non-négligeables.

Elles se sont avérées incontournables pour un grand nombre de développements algorithmiques (notamment pour les hautes luminosités instantanées).

Elles font partie intégrale de certaines analyses de précision, notamment la calibration en énergie des jets, et la mesure de la masse du boson W.

Analyse des données au CC

Un sous-ensemble des données *skimmées* en formats CAF et/ou TMB ainsi que l'ensemble du Monte Carlo est disponible au CC (dans HPSS).
Le sous-ensemble choisi représente les intérêts de physique des analyseurs français.

Le software DØ (les *releases* récentes) est disponible (sur AFS) au CC.

Une façon d'analyser les données qui est pratiquée avec beaucoup de succès au CC est la suivante :

- On tourne une fois (ou une fois tous les N mois) sur les data et le MC pour réduire les données au maximum pour son analyse donnée. Le résultat peut être stocké sous forme de fichiers CAF fortement réduits en nombre d'évènements ou encore sous forme de Ntuples optimisés pour une analyse donnée.
- On prend les données pré-digérées ainsi obtenues, on les sauvegarde dans HPSS, mais on les copie surtout sur un PC sur son bureau (voire son PC portable) pour les analyser.
- On revient au CC seulement quand on se rend compte qu'on a oublié d'inclure une variable importante dans son Ntuple.

Batch et service interactif

Donc, un peu comme dans le cas des productions centralisées (*reprocessing* données, prod MC), nous trouvons que les grands *processings* marchent plutôt bien une fois qu'on a réussi a débougguer les premiers jobs.

Mais – fort heureusement – la physique des particules ne se résume pas à lancer des séries de gros jobs.

Le gros du travail de la plupart des analyseurs se passe en interactif :

- Développer des algorithmes (et donc du code), avec des échantillons limités sur disque.
- Analyser les données pré-digérées en interactif. Il faut une machine interactive rapide et les Ntuples sur disque. On tourne sur les Ntuples pour faire des plots. On réfléchit, on change le code et tourne à nouveau. Itérer des centaines de fois.

  Vu les *datasets* énormes des manips sur collisionneur hadronique et vu les performances (CPU et I/O) des machines modernes, des tailles de plusieurs centaines de GB de données pré-digérées ne sont pas inhabituelles. Un analyseur typique travaille sur plusieurs sous-projets et échantillons, et il a un grand nombre d'analyseurs => il faut plein de disque.

Il me semble que des inefficacités du service interactif au CC, p.ex. dans la compilation du code, et surtout le manque de gros disques «pas chers» est l'une des raisons principale pour les analyseurs de fuir le CC.

En tout cas, ce sont les raisons pour lesquelles je me retrouve souvent à faire mon travail – y compris les grands *processings* ! - à FNAL.
Façon exagérée de le dire : Pourquoi lancer des jobs sur la ferme quand je ne peux pas regarder le résultat ?

Un exemple

Prenons un code d'analyse typique de DØ et compilons-le en interactif sur clued0 (FNAL) et sur ccali (CC).

Le code choisi ici est le code officiel du groupe d'analyse «masse du boson W» pour lire les fichiers CAF et remplir des histogrammes. Un aurait pu choisir n'importe quel autre code CAF; ils suivent tous le même principe et utilisent les mêmes outils communs.

Le code est distribué sous forme de 8 *packages*, dont deux avec les outils communs pour analyser les fichiers CAF, trois avec les outils communs pour le traitement de la qualité des données, et trois avec le code spécifique à la masse du W :

```
wmass_analysis v00-00-31
wmass_util v00-00-24
wmass_blinding_util v00-00-02
cafe p18-br-121
cafe_sam p18-br-07
dq_util v02-01-01
caf_dq v02-01-01
dq_defs v2006-11-30
```

De plus, ce code dépend d'un certain nombre de *packages* dans la *release* D0 (p.ex. définition du format CAF) que nous ne recompilons pas (utilisation de la *release*).

clued0 :

«mon» desktop (affe-clued0, 2x3.2 GHz, 3 GB RAM)
Le code et les librairies générées sont stockés par affe-clued0 via NFS sur un serveur disque.

ccali :

ccali21 (à 2h du mat' pour être le seul utilisateur)
Le code et les librairies générées sont stockés par ccali21 sur /sps/d0.

Résultat (*wall-clock time*) :

clued0 : 11 mins ccali : 27 mins

(ce qui est déjà une amélioration énorme par rapport à l'année dernière ou le temps ccali était typiquement plus de 60 mins)

Miscellaneous

Collection de commentaires qui me semblent tomber de façon récurrente dans les discussions avec les autres utilisateurs DØ du CC :

- Très positif : on peut avoir beaucoup de CPU sur la ferme batch

- Très positif : stockage semi-permanent dans HPSS; grand et facile à utiliser (bien plus facile que Enstore à FNAL)

- Souci : «il n'y a pas de disque au CC» - en particulier pour l'analyse en interactif; l'introduction de /sps était une aide énorme, mais c'est bien plus petit que nos disques «pas cher» sur clued0.

- Souci : utilisons-nous AFS pour des tâches pour lesquelles il n'est pas fait ?

- Souci : disponibilité et fiabilité du CC.

  Les *downtimes* (soit du CC complet, soit des services individuels) sont perçues comme trop importantes, les *downtimes* non-programmées trop fréquentes. Les informations disponibles pendant une *downtime* non-programmée sont perçues comme insuffisantes («je n'arrive pas à accéder à /sps, est-ce un problème connu en cours de réparation ?»; page web avec problèmes connus ??). Une *downtime* pendant le week-end est toujours une *downtime*.

  C'est vrai que FNAL a des *downtimes* programmées régulières (premier mardi de tous les mois) mais en dehors de ça, ça tourne assez rond.

Conclusions

Le CC est une partie intégrale du *computing* de DØ, ses contributions sont bien appréciées par **toute** la collaboration (quand on dit dans les couloirs de FNAL «j'ai lancé un job à Lyon», les gens savent tout de suite «ah oui, c'est la ferme qui fait le gros de notre Monte Carlo»).

Les tâches centralisées (prod MC, reprocessing données) représentent le gros de l'utilisation. Je pense que c'est en partie parce que, dans ce domaine, un petit nombre d'individus déterminés (vous les connaissez) peut avoir un grand impact.

Le travail d'analyse finale est moins développé. Il y a des exemples de groupes français d'analyse qui font la quasi-totalité de leurs grand *processings* au CC.
Mais le gros de l'analyse se fait à FNAL. Il y a une très forte demande de machines IN2P3 installées a FNAL (service interactif et serveurs disque «pas chers»).

Je pense que c'est en partie parce que l'analyse est plus «commode» à FNAL (outils debouggués par des centaines d'utilisateurs, bonne disponibilité, ...). Mais je pense que des ajustements au parc disque et au service interactif pourraient avoir un impact important sur l'efficacité de l'analyse (parce que les grands outils comme la ferme batch et HPSS sont bien là au CC et dans quelques cas bien mieux que à FNAL [p.ex. HPSS vs. Enstore]).