Parc de clusters pour les études d'erreurs de dynamique de faisceau

URIOT Didier

Laboratoire de Modélisation des Accélérateurs et Aimants
CEA/DSM/IRFU/SACM
Le 15 février 2008

Le but de cette note est de présenter l'évolution de l'outil de communication de type client/serveur développé en 2002 permettant de distribuer, sur un parc d'ordinateurs personnels, les calculs de dynamique faisceau dans le cadre d'études d'erreurs très gourmande en puissance de calcul [1]. Cette évolution a pour but d'augmenter significativement la puissance de calcul disponible en incluant dans le parc d'ordinateur les clusters de l'IRFU. Ainsi, ce nouvel outil offre une puissance de calcul capable de répondre aux besoins présents et à venir dans l'étude des linacs haute intensité. Il a été pensé de manière à permettre une abstraction totale de l'hétérogénéité des systèmes. Son utilisation et l'ensemble des procédures d'installation sont présentés ici.

Les besoins

Les études d'erreurs des linacs haute intensité sont des études statistiques reposant sur la simulation d'un grand nombre de linacs différents. Dans ce type d'étude deux grandeurs pertinentes sont à considérer : Le nombre de linacs à simuler et le nombre de particules nécessaires.

- Le nombre de particules doit être suffisant pour permettre d'étudier les pertes dans les machines. Ainsi pour rendre compte correctement d'un Watt perdu dans unes machine comme SPIRAL2 (5 mA, 40 MeV, 200 kW), il faut au minimum 200.000 particules. Une machine comme IFMIF (125 mA, 40 MeV, 4 MW) demande au moins 4 Millions de particules et le futur SPL sera encore plus demandeur.
- Le nombre de machine à simuler est lié au nombre d'éléments la constituant et au nombre d'erreurs différentes pouvant les affecter simultanément (Erreurs de conception, précision

d'alignement, erreurs de mesure…). Plus le linac est complexe, plus le nombre de simulation doit être important de manière à explorer au mieux l'ensemble des combinaisons possibles. Une centaine est un strict minimum, typiquement il faut en simuler plus d'un millier de machines.

En 2002, dans le but de répartir les simulations, avait été mis en place un outil de type client/serveur permettant de réaliser un cluster reposant sur le parc de d'ordinateurs personnels du SACM [1]. Celui-ci avaient permis jusqu'à aujourd'hui de mener des études statistiques sur les linacs en utilisant jusqu'à une quinzaine d'ordinateurs. Ainsi ont été menées, des études de linac incluant 1.000.000 macro-particules transportées sur une centaine de linacs différentes, études IFMIF [2] ou SPIRAL2 [3]. La statistique cumulée atteignait environ une centaine de million de particules et demandait environ 1 semaines de calcul. Les besoins actuels et à venir sont 2 ordres de grandeur au dessus (plus de particules et plus rapidement). Ce qui implique porter le nombre de machine de calcul autour du millier d'unité. La seule façon raisonnable d'atteindre cet objectif a été de faire évoluer les outils de communication, le client et le serveur, afin d'y intégrer les clusters de l'IRFU.

Clusters de calcul de l'IRFU

L'IRFU dispose aujourd'hui de plusieurs clusters de calcul. Le premier d'entre eux est la grille destinée aux dépouillements des expériences LHC. Le second en cours d'installation est spécialement dédié aux calculs interactifs de l'IRFU, MELISSA. Le SACM quand à lui dispose à ce jour en partenariat avec le SAP d'un cluster spécialement dédié aux calculs parallèles, DAPHPC. L'ensemble de ces machines est en constante évolution et met à disposition pour le moment trois à quatre cents nœuds pour la grille LHC (ressources locales prioritairement affectées à l'IRFU), quelques dizaines pour MELISSA et deux centaines pour DAPHPC. L'ensemble des ressources cumulées de ces clusters est d'environ 500 cœurs. Ces ressources sont quasi doublées chaque année. Ces 3 clusters ne présentent pas les mêmes caractéristiques.

- **La grille LHC** est destinée à des calculs multi paramètres, c'est-à-dire, ne demandant pas de communication intensive entre les nœuds, elle est donc très adaptée à notre problématique. Elle n'est pas limitée en théorie aux machines physiquement présente dans les locaux de l'IRFU, mais regroupe théoriquement l'ensemble des ressources des partenaires du LHC (plusieurs dizaines de milliers de machines). Son temps de réponse à une requête (lancer un job) est relativement long, de quelques minutes jusqu'à une dizaine de minutes. Il est très difficile de connaitre la puissance et la mémoire disponible de chaque nœud, compte tenu de son architecture hétéroclite. L'ensemble des ressources et leurs taux d'occupation est visualisable à l'adresse suivante: http://goc.grid.sinica.edu.tw/gstat/GRIF/

- **MELISSA** n'est pas à proprement parlé, un vrai cluster, dans le sens ou elle ne possède pas de machine d'entrée dédiée aux connections des utilisateurs. En effet chaque machine peut être utilisée directement sans passer par une queue de « batch ». Parfaitement adapté à nos besoins, son temps de réponse est de l'ordre de la dizaine de seconde. Elle est composée de 16 machines intégrant des processeurs double cœurs de type P4 cadencés à 2.8 GHz disposant de 1 Giga Octets de mémoire. L'ensemble des ressources et leurs taux d'occupation est visualisable à l'adresse suivante: http://132.166.28.220/ganglia/

- **DAPHPC** est destiné aux calculs parallèles, optimisé pour la communication entre les cœurs, il est composé de 48 nœuds de 4 cœurs (Biprocesseur AMD Dual Core cadencé à

2.6 GHz disposant de 8 Giga Octets de mémoire). Il est très performant, avec des temps de réponse de la seconde. L'ensemble des ressources et leurs taux d'occupation est visualisable à l'adresse suivante: http://daphpc/ganglia/

Objectifs

Le but des nouveaux développements a été de modifier l'outil déjà existant de type client/serveur afin de permettre l'intégration des clusters tout en conservant la possibilité d'intégrer les machines classiques. Celles-ci restent en effet pertinentes pour ce type d'application au regard de l'évolution des PCs. En effet, ceux-ci possèdent maintenant au minimum deux processeurs, parfois 4 et dans quelques années 8. D'autre part compte tenu de la diffusion des codes dans les laboratoires extérieurs, il est essentiels des conserver cette possibilité, les clusters de l'IRFU n'étant évidemment pas accessibles de l'extérieur du CEA. L'ensemble doit bien évidemment rester simple et transparent pour l'utilisateur tout en s'intégrant dans notre outil de simulation, TraceWin [4]. De plus, tous les systèmes d'exploitation restent supportés.

Architecture générale et protocole

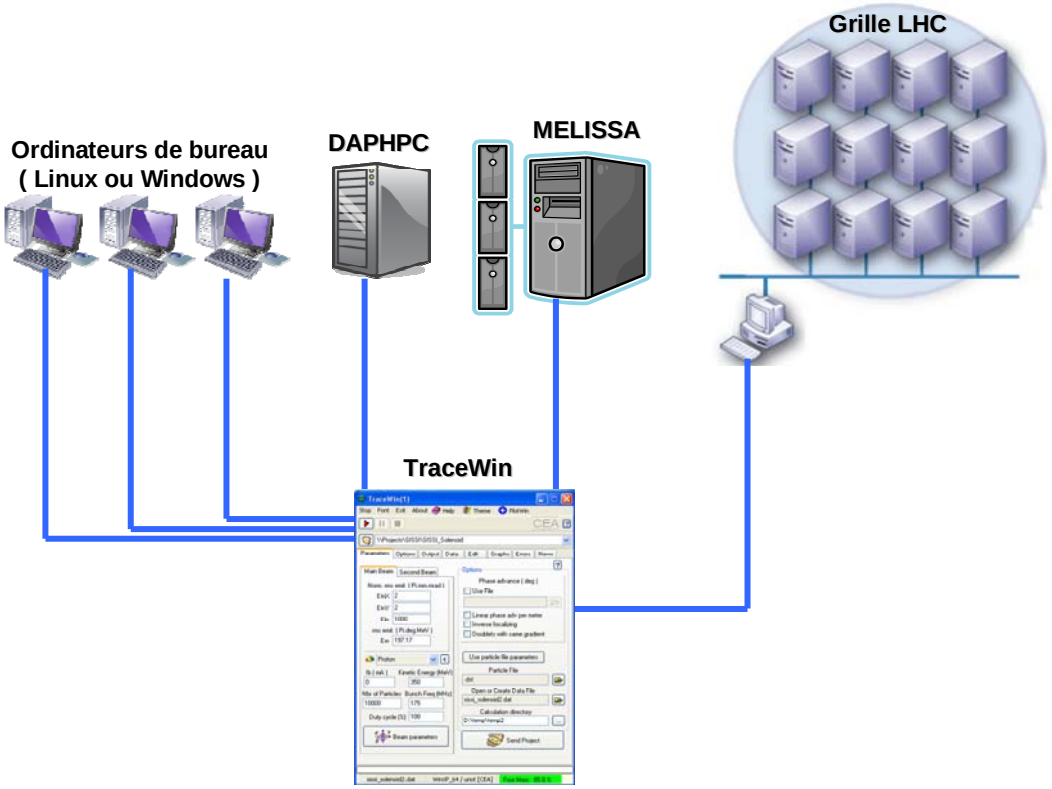
L'architecture générale repose sur le code de simulation, « *TraceWin* ». Il a en charge la gestion de l'ensemble des linacs à simuler et a un rôle de super-client. Il réparti les simulations sur l'ensemble des machines à sa disposition via le code serveur « *twserver* ». Celui-ci doit être installé localement sur l'ensemble du parc. Chacun de ces serveurs traite un ou plusieurs calculs à la fois, les répartissant ou non sur plusieurs cœurs. TraceWin est en charge du dialogue et de la gestion des différents serveurs et ordinateurs du parc. Il les contrôle, les interroge sur leurs disponibilités, transmet les données nécessaires, lance les simulations, rapatrie les résultats en fin de calcul et pour finir compile l'ensemble en données exploitables. L'ensemble de ces étapes se déroule de manière totalement transparente pour l'utilisateur. Un mode surveillance est toutefois possible de manière à contrôler l'ensemble de ces communications.
Ces communications utilisent le protocole TCP/IP pour le transfert des fichiers et nécessite l'ouverture des ports 1024 à 10000. Un certain nombre de commandes supplémentaire ont été développé afin de :

- Connaitre le système d'exploitation,
- disposer de la mémoire libre,
- évaluer l'occupation du processeur (uniquement les machines sous Windows),
- exécuter un programme,
- tuer un programme,
- connaitre le statut du programme exécuté,
- vérifier l'état du serveur.

Le protocole de calcul est le suivant pour chaque machine. Dans un premier temps, elle est interrogée sur sa disponibilité, sa mémoire libre, son système d'exploitation. Si les conditions nécessaires à un calcul sont respectées, Les données d'entrée tel que le fichier de particules, les cartes de champ, la description du linac sont transmises. Puis le code de calcul approprié au système d'exploitation et transmis à son tour. L'ordre est alors envoyé d'exécuter le

programme de simulation et l'état de celui-ci est régulièrement interrogé. Une fois le calcul terminé correctement, les résultats sont rapatriés et un nouveau cycle est relancé.

L'ensemble, une fois installé, forme une infrastructure de calcul formée d'un ensemble hétérogène d'unités de calcul géré de manière totalement transparente pour les utilisateurs.



Architecture générale du super cluster

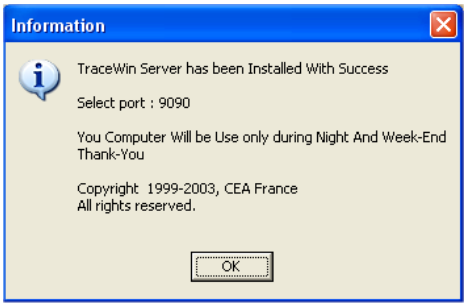
Installation des serveurs

L'ensemble des codes est disponible sur le serveur FTP du CEA à l'emplacement suivant : ftp://ftp.cea.fr/incoming/y2k01/Dynamic_Codes/

La communication du parc hétérogène de machines et de clusters avec TraceWin s'effectue via un client/serveur qu'il faut bien entendu installer sur chacune des unités distantes. Suivant le type d'unité, la procédure est différente.

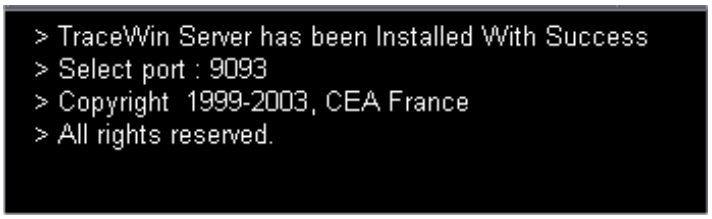
- **Machine de bureau sous système d'exploitation Windows.**
  Lancer l'exécutable « *twserver.exe* » d'un répertoire temporaire. Un répertoire « *C:\TWServer* » sera automatiquement créé dans lequel s'installera le serveur. Ne jamais exécuter le serveur directement à cette position. Une fois installé le message ci-dessous apparait, indiquant le port de communication qui faudra spécifier dans l'interface de gestion de TraceWin. Dans la barre des taches à droite, le serveur apparaitra sous la forme d'un icône similaire à celui de TraceWin. Lorsque celui-ci sera mis à contribution cet icône se mettra à clignoter, indiquant à l'utilisateur de cette machine, qu'elle est utilisée à distance. L'utilisateur a la possibilité de mettre en sommeil le serveur en cliquant sur son

icône. A noter qu'à chaque redémarrage de l'ordinateur le serveur sera automatiquement chargé.



- **Machine de bureau sous système d'exploitation Linux.**

Lancer l'exécutable « *./twserver&* » d'un répertoire temporaire. Un répertoire caché « */home/.TWServer* » sera automatiquement créé dans lequel s'installera le serveur. Ne jamais exécuter le serveur directement à cette position. Une fois installé le message ci-dessous apparaitra indiquant le port de communication à renseigner dans TraceWin. Si on désire que ce serveur soit actif à chaque démarrage du système ajouter l'argument « *auto* » dans la commande d'installation: « *./twserver auto&* »



- **Cluster MILSSA (132.166.28.220)**

Le cluster MELISSA n'ayant pas une gestion par utilisateur, mais par groupe d'utilisateurs, il n'est pas nécessaire d'installer un serveur par utilisateur. Le serveur n'a donc à être installé qu'une unique fois pour le SACM. La procédure consiste à lancer l'exécutable dans le répertoire de travail du SACM, « */home/gpfs/mnt/sacm/* », par la commande « *./twserver melissa* ».
Un répertoire caché « */home/gpfs/mnt/sacm/.TWServer* » sera automatiquement créé dans lequel s'installera le serveur. Ne jamais exécuter le serveur directement à cette position. Une fois installé un message similaire à l'installation sous machine Linux apparaitra.

- **Cluster DAPHPC (132.166.28.212)**

Le cluster DAPHPC ayant une gestion par utilisateur, il est nécessaire d'installer un serveur par utilisateur. Le serveur devra donc être installé chaque fois qu'un nouvel utilisateur voudra inclure ce cluster dans des études via TraceWin. Se placer dans le répertoire de son compte et lancer la commande « *./twserver dpahpc* ». Un répertoire caché « */data/.TWServer* » sera automatiquement créé dans lequel s'installera le serveur.

- **Grille LHC (192.54.208.17)**

Avant toute chose, il est nécessaire d'obtenir un certificat auprès d'une autorité de sécurité. If faut ensuite adhérer à une organisation virtuelle ou VO. Et pour finir avoir un compte sur un point d'entrée de la grille. Une fois l'ensemble de ces procédures achevé, avec l'aide obligatoire de l'équipe en charge de la grille LHC, sur son répertoire courant

« */home/* » lancer la commande « *./twserver grid* ». Un répertoire « /*home/.TWServer* » sera automatiquement créé dans lequel s'installera le serveur. Pour finir, les droits d'accès sont limités dans le temps et cette durée est à spécifier par une commande du type « *grid-proxy-init -valid 120:00* », (Ici, on se donne les droits pour 120 heures par exemple). Chaque utilisateur devra répéter cette procédure.

Les mises à jour des serveurs, quand elles sont nécessaires, suivent exactement la même procédure d'installation, excepté pour les machines bureautiques qui peuvent être mise à jour à distance via l'interface de contrôle de TraceWin.

Interface de contrôle de TraceWin

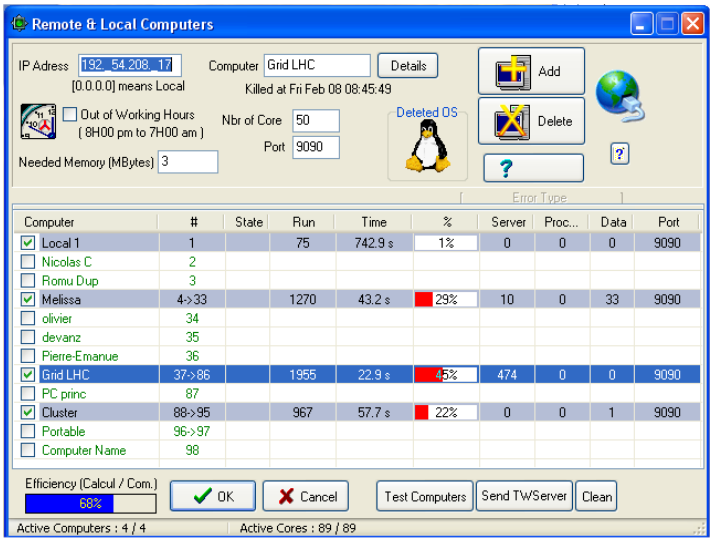
TraceWin possède une interface administrateur permettant aux utilisateurs de spécifier et gérer l'ensemble des ressources disponibles. Un certain nombre de paramètres doit être spécifié pour chaque machine :

- Son adresse IP,
- son identifiant de manière à la reconnaitre aisément,
- le port sur lequel le serveur à été installé (voir installation des serveurs),
- le nombre de cœurs à utiliser,
- un flag permettant de restreindre l'utilisation de la machine aux heures creuses (pertinent pour les ordinateurs de bureau),
- la mémoire libre minimum en dessous de laquelle la machine ne doit pas être utilisée.

Spécifier une adresse IP = « 0.0.0.0 » permet d'indiquer à TraceWin d'utiliser aussi pour les calculs l'ordinateur local ou il est installé.

De plus, en cours de calcul un certain nombre d'information permet de connaitre l'état de chaque machine.

- Son statut,
- le pourcentage d'utilisation,
- le temps de calcul moyen par cœur,
- le nombre d'erreurs de communications,
- le nombre d'erreurs de calculs.



Interface de contrôle de TraceWin

Remarques:

- Avant chaque campagne de calculs, il est fortement conseillé de lancer la procédure « Test computers », qui permet d'interroger l'ensemble du parc afin de vérifier que tous les serveurs sont correctement installés et opérationnels.

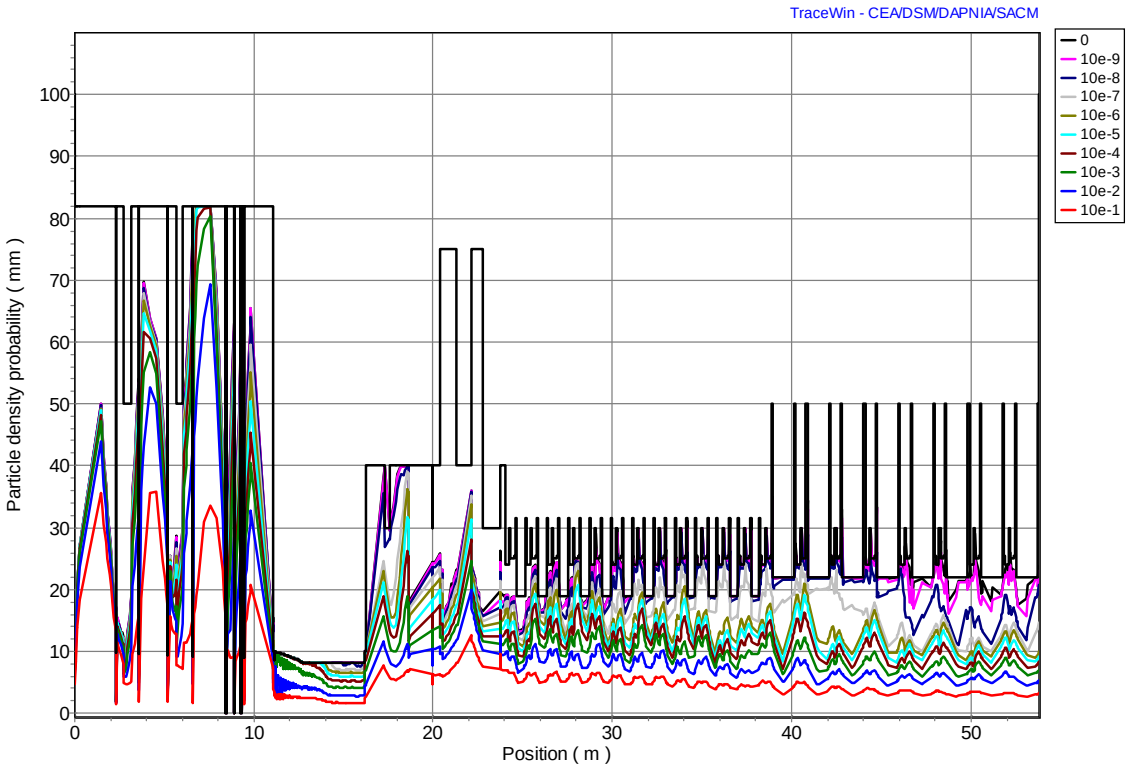
- Compte tenu de la vitesse de transmission des données transitant par le réseau interne, et pour ne pas surcharger les nœuds d'entrée des clusters, les différentes simulations ne sont pas soumises, sur un même cluster, simultanément mais avec un décalage temporel de l'ordre de la dizaine de secondes. Ce qui signifie que cette architecture n'est absolument pas adaptée à des simulations inferieures à quelques minutes. En règle générale, plus les simulations demandent de temps de calcul plus on peut utiliser de nœuds par cluster. Par exemple pour des simulations de 3 minutes, inclure les 300 nœuds de la Grille LHC, ne sert à strictement à rien. Seulement une cinquantaine sera utilisée pour le calcul, alors que les données seront par contre transmises 300 fois, surchargeant le réseau inutilement.

- A la fin de chaque campagne de calcul ou en cas d'arrêt par l'utilisateur, TraceWin se charge de nettoyer chaque nœuds un par un. Cette procédure peut être très longue, mais il est fortement conseillé de ne pas l'interrompre.

- Le nombre de nœuds simultanément géré par l'interface de contrôle ne peut excéder 1000 unités.

Première exploitation

Les premiers résultats issus du super-cluster concernent une étude d'erreurs de l'accélérateur SPIRAL2. Les simulations ont été effectuées de la source à la fin du linac supraconducteur. 1.000.000 ont été transportées dans 1400 accélérateurs différents, produisant ainsi une statistique supérieure à 1 milliard de particules [5]. Pour rappel, le faisceau de SPIRAL2, 5 mA de deutons à 88 MHz comprend environ 350.000.000 de particules par paquet. L'ensemble de ces simulations a occupé 200 cœurs sur une journée.



Densité de probabilité des particules le long de SPIRAL2
(La courbe rouge représente 90% du faisceau, la bleu 99%, etc...)

Futurs développements nécessaires

La plupart des nouvelles machines de calculs ainsi que les clusters possèdent un système d'exploitation 64 octets. Le code de calcul distribué est actuellement compilé en 32 octets de manière à être compatible avec l'ensemble du parc. Cette limitation impose une limite aux nombre maximum de particules transportables simultanément dans TraceWin autour de 10.000.000 de particules. Bien que cette barrière puisse être contournée par des artifices informatiques, il faudra travailler en 64 octets quand l'ensemble des machines le permettra, de manière à simplifier et surtout accélérer les simulations.
Une autre évolution inévitable concerne le protocole de transfert de fichiers. Actuellement développé autour du protocole de base TCP/IP protégé par un nom d'utilisateur et un mot de passe. Il sera nécessaire de faire évoluer le client et le serveur vers un protocole plus sécurisé tel que SFTP ou SSH.

Conclusions

La configuration actuelle de cet outil permet de répondre efficacement aux besoins actuels. Dors et déjà dimensionnée pour les futurs besoins, elle profitera naturellement de l'évolution constante des machines personnels et des clusters de calcul. Le serveur développé dans le cadre de cette application a une vocation générale et pourrait sans difficulté être utilisé pour d'autre type de besoin. Les syntaxes exactes des protocoles spécifiques de communication permettant de le contrôler sont évidemment disponibles pour l'ensemble de la communauté.

Remerciements

L’ensemble de ces développements n’aurait pas été possible sans de l’équipe en charge de la grille LHC, en particulier Christine LEROY. Jean-Francois LECOINTE qui a volontiers adapté la gestion de MELISSA aux mes besoins spécifiques. Romuald DUPERRIER qui m’a patiemment initié à l’utilisation de DPAHPC.

Références

[1] : http://www-dapnia.cea.fr/Phocea/file.php?class=std&&file=Doc/Publications/Archives/dapnia-02-72.pdf

[2] : http://prst-ab.aps.org/abstract/PRSTAB/v9/i4/e044202

[3] : http://accelconf.web.cern.ch/AccelConf/e04/PAPERS/WEPLT078.PDF

[4] : R. Duperrier, N. Pichoff, D. Uriot, “CEA Saclay, codes Review for High Intensity Linacs Computations” , ICCS2002, Amsterdam

[5] : https://edms.in2p3.fr/edms/doc.info?cookie=2091966&document_id=I-012039&version=1