Mad II grandeur nature : une expérience de portage

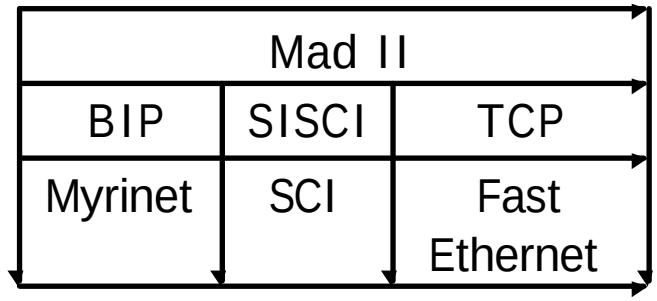
N. Déjean
O. Aumage, L. Bougé

LIP - ENS Lyon
France

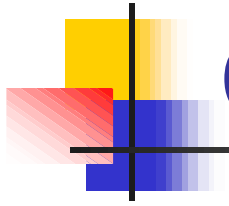
Nicolas.Dejean@ens-lyon.fr

Madeleine II

- Madeleine II
  - Bibliothèque de communication
  - Développée au LIP
  - PM2
- Développements logiciels
  - MPICH
  - DSM-PM2
  - Globus / Nexus

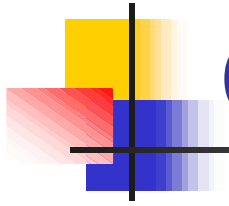


- Tests à grande échelle avec une application réelle



Global Arrays

- Développé à l'EMSL
  - Environmental Molecular Sciences Laboratory
  - Localisé au PNNL
    - Pacific Northwest National Laboratory (USA)
- Dans le domaine public depuis 1994
- Applications de calcul scientifique
  - Chimie quantique (NWChem, GAMESS-UK, Columbus, Molpro, Molcas)
  - Dynamique moléculaire (NWChem)
  - Imagerie informatique, chimie atmosphérique, dynamique des fluides, prévisions financières (Bear Stearn)

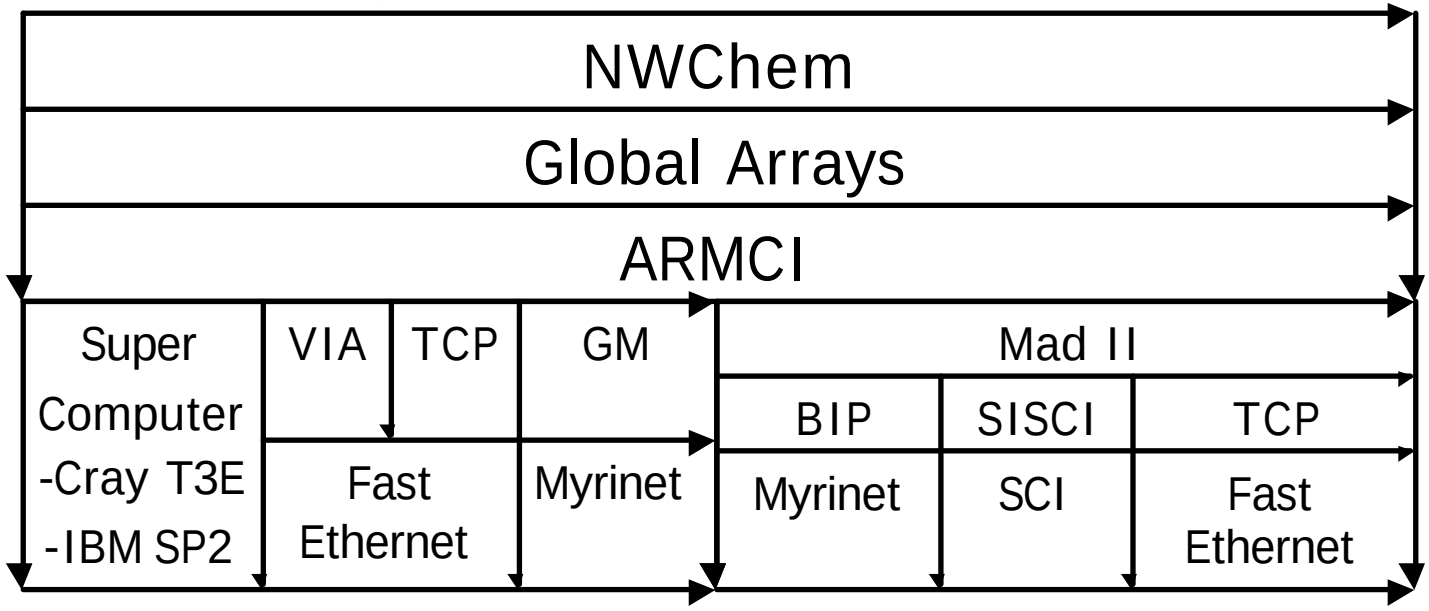


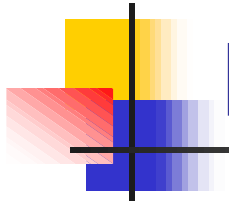
Global Arrays

- Interface de programmation
  - orientée mémoire partagée
  - efficace et portable
  - destinée aux machines parallèles
- Bibliothèques de passage de messages
  - indispensable
  - compatible
  - MPI, PVM, TCGMSG

Objectifs

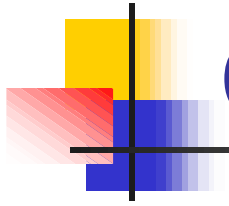
- GA / Mad II
- Application réelle
- Passage à l'échelle





NWChem

- Mécanique quantique et dynamique moléculaire
- Développé par l'EMSL au PNNL
- Architectures disponibles
  - Supercomputers
    - Cray T3D et T3E
    - IBM SP2
    - SGI SMP systems
    - SUN workstations
    - Fujitsu VX/VPP
    - Alpha SMP servers running Tru64 or Linux
    - Alpha SC series running Tru64 or Linux
  - Workstation networks
    - x86-based workstations running Linux
    - x86 Linux Clusters with Giganet switch using the VIA protocol
    - x86 Linux Clusters with Myrinet switch using the GM software
    - HP workstations running HPUX

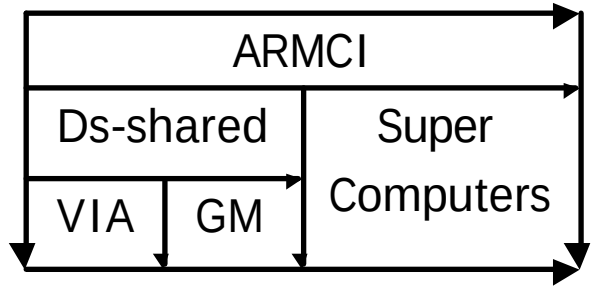


GA / ARMCI

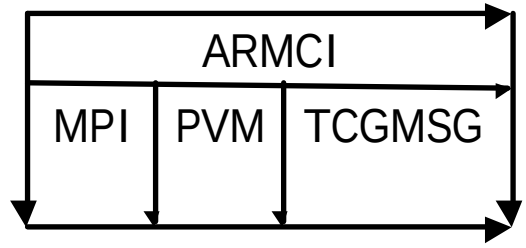
- ARMCI
  - Initialisation
  - Allocation
  - Echange
- GA
  - Initialisation
  - Création
  - Echange
  - Communications globales (Scatter, Gather)
  - Algèbre linéaire
  - Interfaces vers Scalapack et PeIGS
- 1.4 millions de lignes en C et une centaine de lignes en Fortran

ARMCI

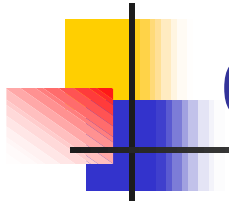
- Aggregate Remote Memory Copy Interface
- Optimisée pour le transfert de données non contiguës
  - Vector
  - Strided
- Interface de portabilité
  - ds-shared



- Nécessite une bibliothèque de passage de messages
  - MPI, PVM, TCGMSG
  - Initialisation, synchronisation



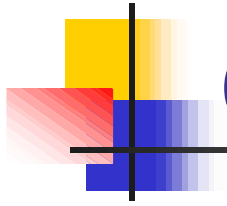
ARMCI / ds-shared + TCGMSG



- ARMCI
  - ARMCI_Init
  - ARMCI_Malloc
  - ARMCI_PutS

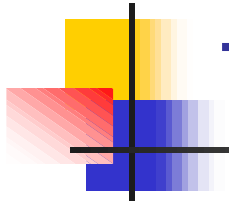
- ds-shared
  - armci_start_server
- 400 lignes de C

- TCGMSG
  - pbegin, pend
  - nnodes, nodeid
  - snd, rcv
  - brdcst, sync



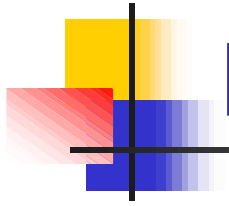
ds-shared / Mad II

- Interface commune au port de VIA et GM
- ds-shared
  - armci_send_req_msg
  - armci_ReadFromDirect
  - armci_call_data_server
  - armci_ WriteToDirect
- Mad II
  - mad_begin_packing, mad_pack, mad_end_packing
  - mad_begin_unpacking, mad_unpack, mad_end_unpacking
- Plusieurs canaux
  - Requêtes
  - Données
  - Passage de messages



TCGMSG / Madeleine II

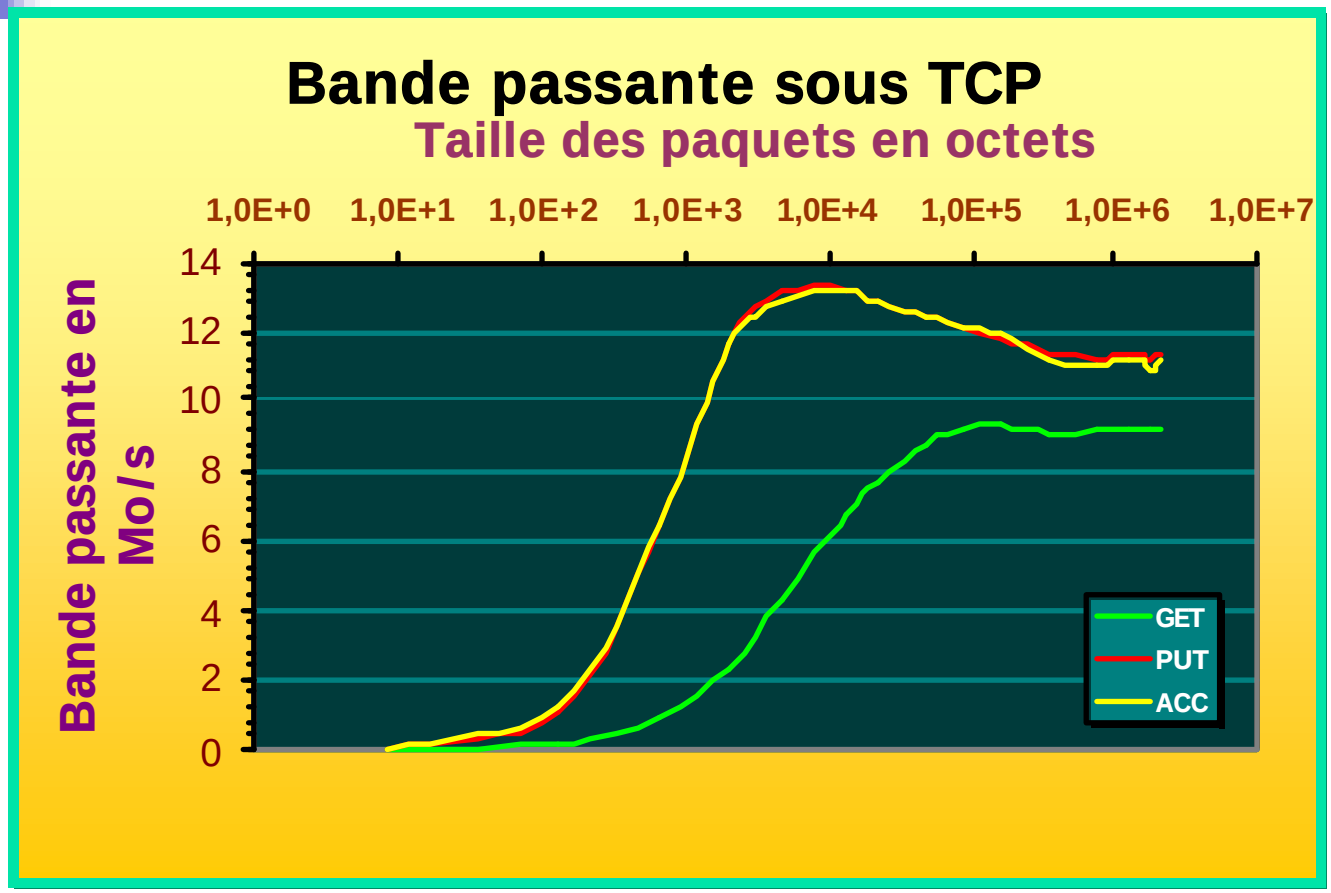
- Initialisation
  - Passage de messages
  - GA / ARMCI
- Fonctionnalités de haut niveau indispensables
  - Réception sélective
  - Synchronisation
  - Diffusion
- Améliorations
  - Définition d'un modèle propre à Madeleine II
  - Ajout du support du modèle dans ARMCI et GA



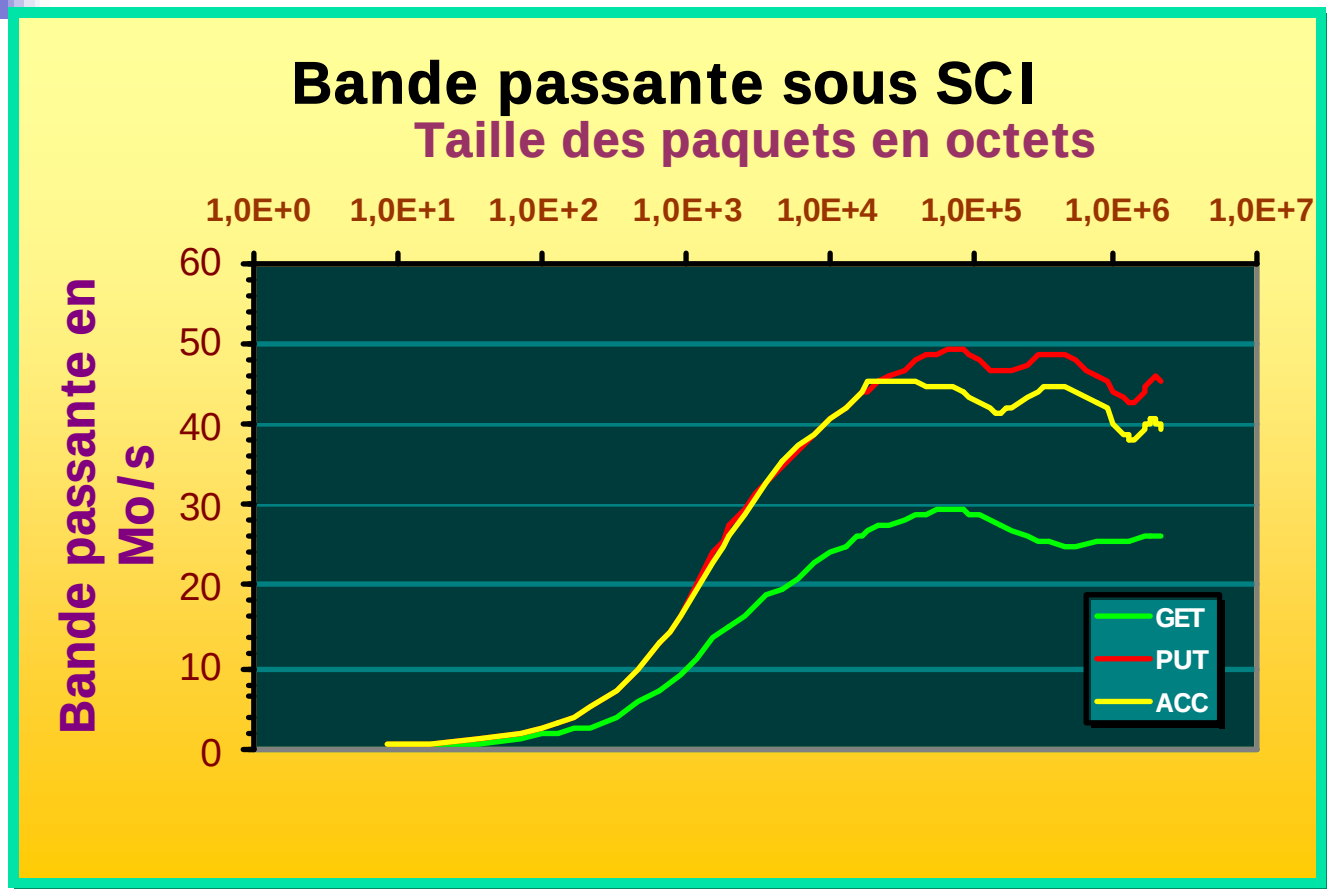
Environnement de tests

- Pentium II 450 MHz / 256 Mo de RAM
  - Interface TCP/Fast-Ethernet 100 Mb/s
  - Interface SISCI/SCI
  - Interface BIP/Myrinet
- 3 courbes
  - PUT
  - GET
  - ACC

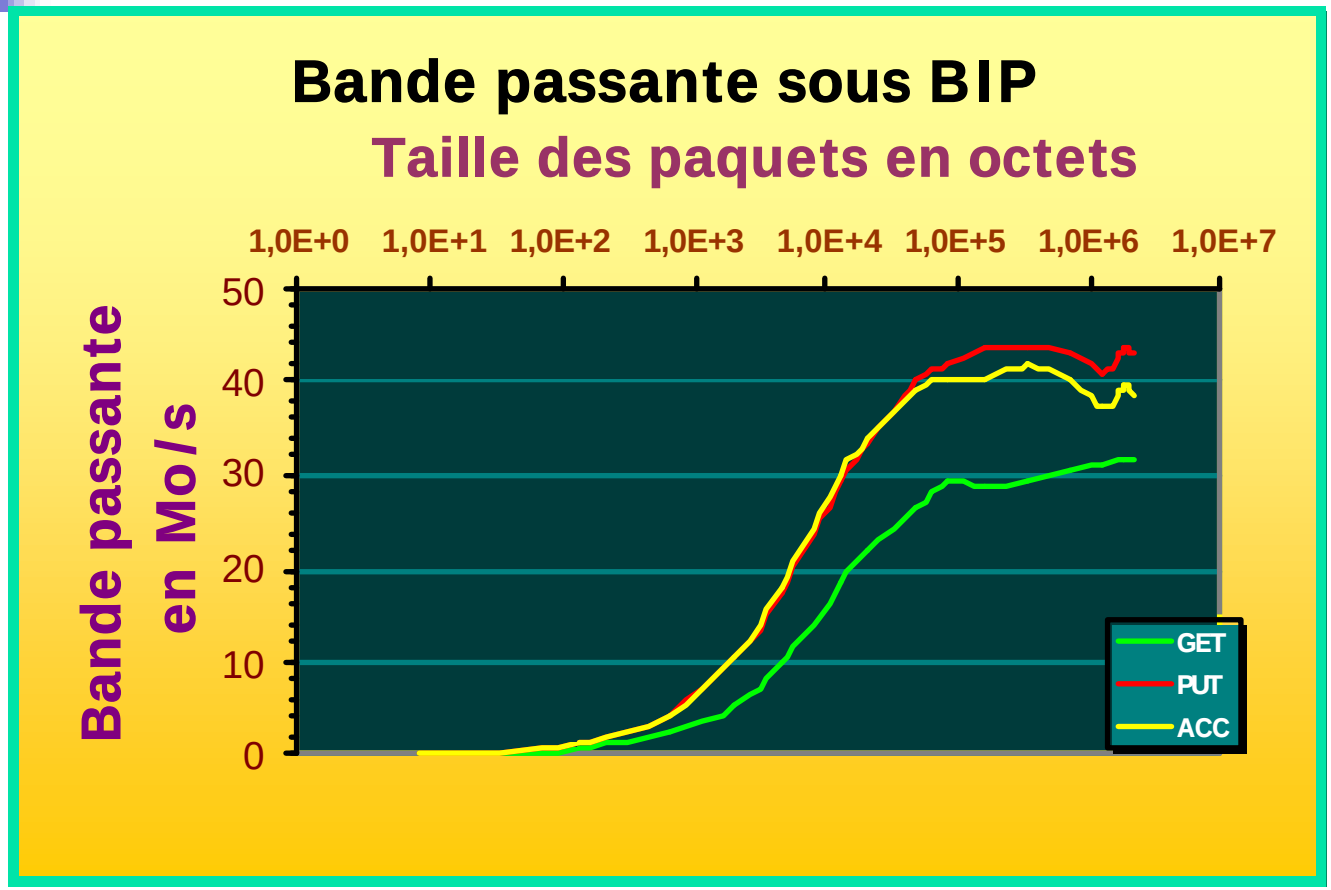
Performances - TCP

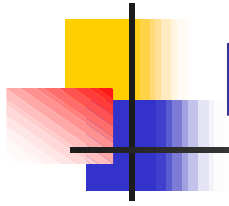


Performances - SCI



Performances - BIP

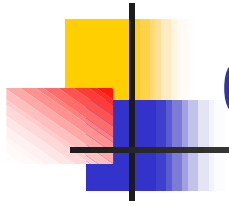




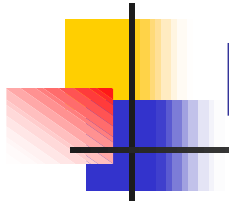
Performances - conclusions

- Bande passante
  - TCP/FastEthernet
    - 11 Mo/s
  - SISCI
    - 50 Mo/s
  - BIP/Myrinet
    - 45 Mo/s
- Sans optimisation
- Fluctuations au-delà de 400 ko

Intégration des threads Posix dans Madeleine II



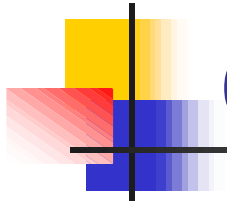
- ARMCI {version ds-shared} est un système multi-threadé
  - Support des threads natifs
    - Solaris
  - Support des threads Posix
- Madeleine supporte seulement les threads Marcel
  - Développés de concert dans PM2
- Support des threads Posix par Madeleine II
  - ARMCI supporte déjà les Pthreads
  - Les threads Marcel sont de niveau utilisateur



Liaison C / Fortran

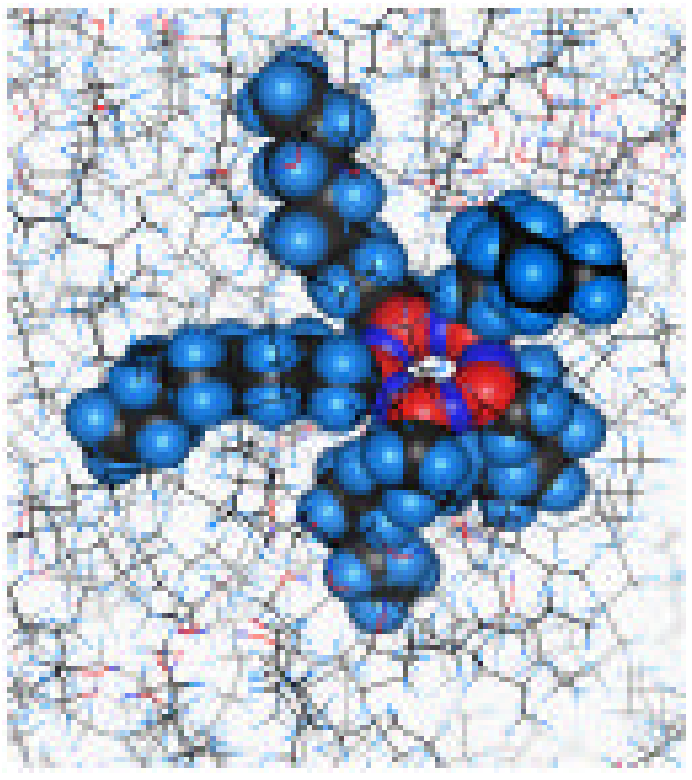
- GA est écrit en C
- Interfaces
  - C
  - Fortran
- Programmes le plus souvent en Fortran
  - Partage des symboles
    - Arguments de la ligne de commande
- Dépendance au système cible
  - Architecture matérielle
  - Système d'exploitation
  - Compilateur

Modification du support de BIP dans Madeleine II

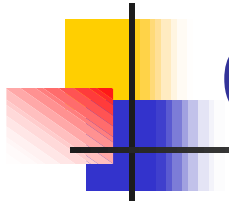


- BIP
  - Interface pour les cartes Myrinet
  - Développé au LIP
  - Bonnes performances
    - Bande passante 120 Mo
    - Latence 7 µs
- Madeleine II utilise BIP
  - Support des cartes Myrinet
  - Taille des messages limitée à 1 Mo

NWChem / GA / Mad II



- 180 000 Lignes de C
- 900 000 Lignes de Fortran
- Compilation
  - 1 heure
  - Pentium II 350 MHz
  - 128 Mo RAM



Conclusion

- La faisabilité est démontrée
  - Le développement en couches est très efficace
  - Niveau logiciel de Madeleine II
- Simplifications
  - TCGMSG
  - Pas de support SMP
  - Pas d'optimisations
- A faire
  - Tests larges
  - Adapter d'autres programmes