MỘT SỐ KẾT QUẢ BAN ĐẦU ỨNG DỤNG HỆ THỐNG TÍNH TOÁN LƯỚI TRONG PHÂN LOẠI LỚP PHỦ

NGUYỄN ĐÌNH DƯƠNG

I. MỞ ĐẦU

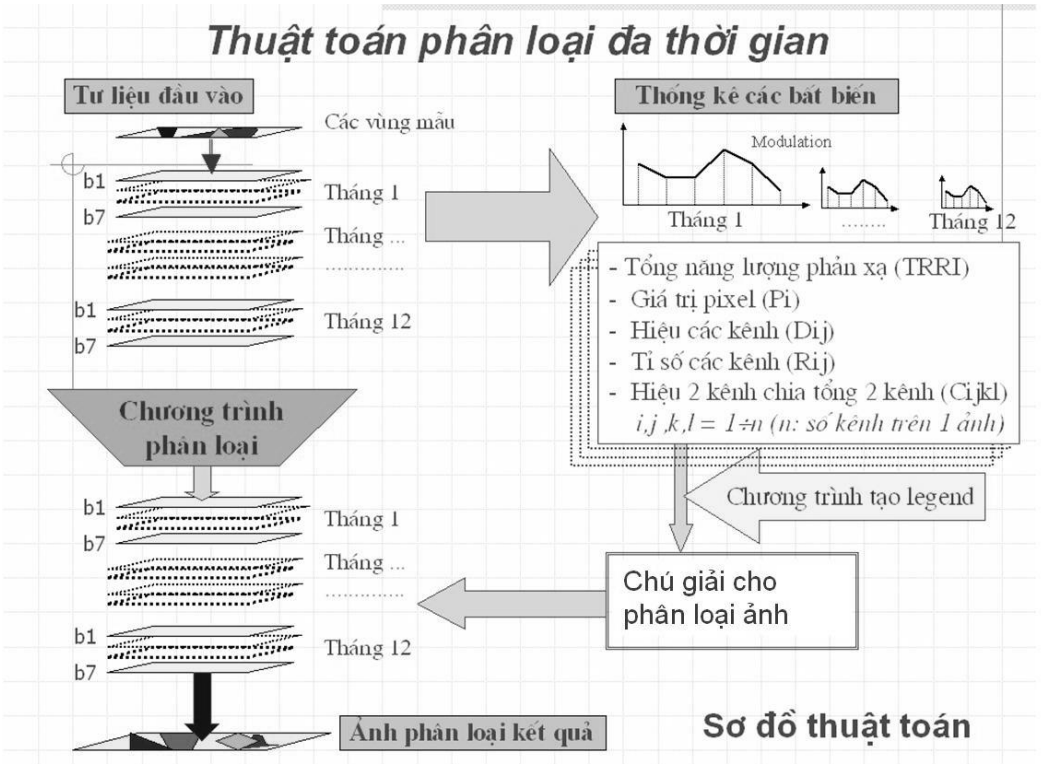
Bản đồ lớp phủ (land cover map) có thể được thành lập bằng nhiều phương pháp. Hai phương pháp thường sử dụng nhiều nhất đó là giải đoán bằng mắt và phân loại có kiểm định. Tuy nhiên trong trường hợp cần xây dựng bản đồ trên diện rộng với tư liệu đa thời gian, các phương pháp truyền thống thường khó áp dụng. Trong khuôn khổ Chương trình nghiên cứu vệ tinh ADEOS-II, tác giả đã phát triển thành công thuật toán GASC, cho phép thành lập bản đồ lớp phủ từ tư liệu ADEOS-II GLI đa phổ đa thời gian [1]. Thuật toán này không giới hạn chủng loại tư liệu đầu vào và có thể được sử dụng cho tư liệu MODIS. Tuy nhiên khi tính toán trên máy PC đơn lẻ và cho dữ liệu toàn bộ Việt Nam, phải cần thời gian từ 6 đến 8 tiếng. Nếu dữ liệu đầu vào là toàn bộ khu vực châu á hay toàn thế giới, thời gian tính sẽ rất lâu.

Giải pháp sử dụng tính toán lưới sẽ là lựa chọn tối ưu trong điều kiện không có kinh phí để trang bị các máy tính lớn. Tính toán lưới là một công nghệ nhằm sử dụng nhiều máy tính nhỏ kết nối trong mạng phân tán, tạo nên một máy tính ảo, người sử dụng làm việc như với một máy tính đơn lẻ. Các máy tính đơn lẻ có thể chạy trên các hệ điều hành khác nhau và được kết nối với nhau qua một hệ thống phần mềm riêng biệt. Với công nghệ này nhiều bài toán có thể giải trong thời gian ngắn và không yêu cầu mỗi phòng thí nghiệm phải đầu tư một hệ thống máy tính lớn. Trong bài báo này tác giả giới thiệu công nghệ tính toán lưới, hệ thống máy tính của Viện Khoa học Công nghiệp tiên tiến Nhật Bản (AIST), tác giả đã kết nối để thử nghiệm và quá trình chuyển đổi chương trình từ hệ điều hành Windows sang hệ tính toán lưới thử nghiệm phân loại lớp phủ cho vùng Việt Nam. Tác giả trao đổi những kinh nghiệm trong việc khai thác hệ thống tính toán lưới và kết quả những thử nghiệm đầu tiên.

II. BÀI TOÁN PHÂN LOẠI LỚP PHỦ VÀ NHU CẦU ỨNG DỤNG HỆ THỐNG TÍNH TOÁN LƯỚI

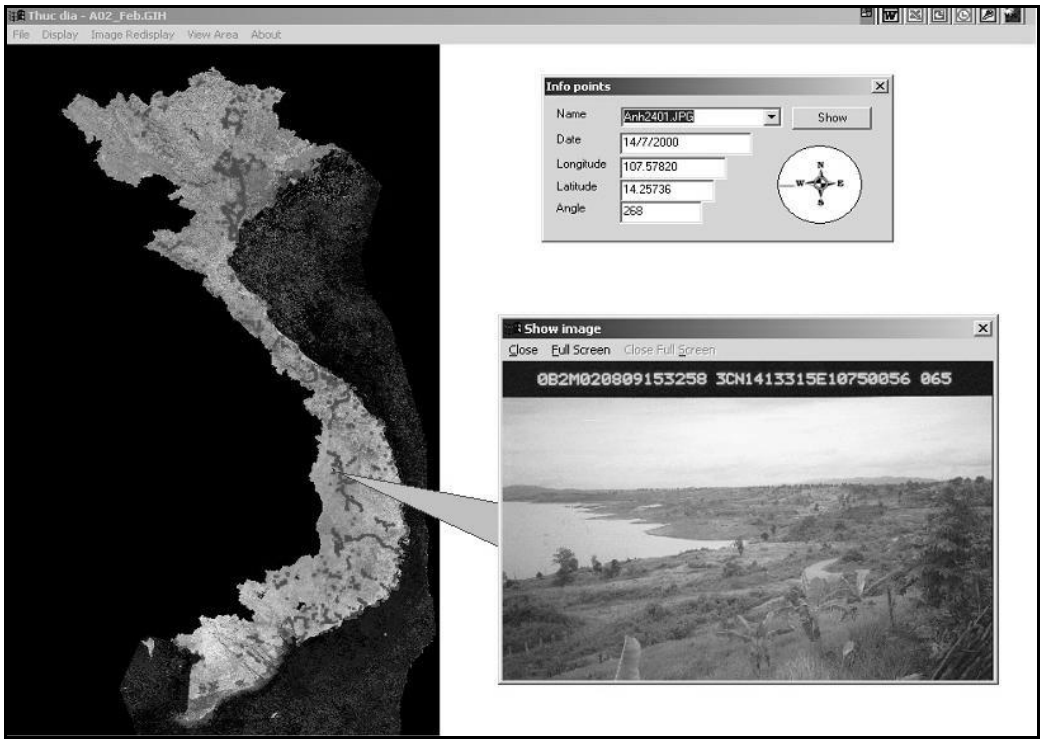
Lớp phủ - Land cover được định nghĩa như lớp phủ vật lý bề mặt Trái Đất bao gồm các loại thực vật : trảng cỏ, cây bụi, rừng thường xanh, rừng rụng lá ... ; các loại đất trống, đá lộ, các loại mặt nước như ao hồ, sông suối, băng tuyết, biển và đại dương. Bản đồ lớp phủ có thể được xây dựng dựa trên hai phương pháp chính là khảo sát thực địa và phân tích tư liệu viễn thám. Phương pháp khảo sát thực địa chỉ có thể thực hiện cho một vùng hẹp, trong trường hợp cần xây dựng bản đồ lớp phủ cho một khu vực rộng lớn, tư liệu viễn thám là nguồn thông tin duy nhất cho phép phân loại các loại hình lớp phủ. Tư liệu viễn thám cung cấp thông tin đa phổ và đa thời gian mà dựa vào đó có thể nhận biết được sự biến động theo mùa, từ đó phân loại được chính xác các thảm thực vật như rừng thường xanh, rừng rụng lá, đất ngập nước theo mùa và đất ngập nước thường xuyên... Các phương pháp phân loại lớp phủ có thể dựa trên các thuật toán truyền thống như phân loại phi kiểm định, phân loại có chọn vùng mẫu kết hợp với sự tương tác giữa người sử dụng và phần mềm. Tuy nhiên trong trường hợp sử dụng tính toán lưới, việc tương tác giữa người sử dụng và hệ thống phần mềm là không thể, do vậy cần có những thuật toán riêng cho môi trường tính toán lưới. Thuật toán GASC - phân loại lớp phủ dựa trên phân tích hình học đường cong phổ phản xạ do Nguyễn Đình Dương phát triển [1] trong khuôn khổ chương trình nghiên cứu vệ tinh ADEOS-II GLI có thể sử dụng hoàn toàn phù hợp cho môi trường tính toán này. *Hình 1* là sơ đồ nguyên lý của thuật toán GASC.

Chúng ta thấy tư liệu đầu vào bao gồm hai phần cơ bản, đó là tư liệu viễn thám đa thời gian và bản chỉ tiêu phân loại cho các đối tượng. Tư liệu đa thời gian xây dựng sao cho phản ảnh được các biến động cơ bản theo mùa của các thảm thực vật tự nhiên cũng



Hình 1. Sơ đồ thuật toán phân loại lớp phủ GASC

cũng như nhân tạo. Với các cây nông nghiệp cơ bản như lúa, ngô..., khoảng cách giữa hai thời điểm dữ liệu ít nhất phải là một tháng mới có thể cho phép tách chúng ra khỏi các đối tượng khác. Tư liệu đầu vào vừa là đa phổ vừa là đa thời gian. Các đối tượng lớp phủ được nhận biết dựa trên một phần tính chất hấp thụ và phản xạ phổ, thể hiện thông qua đường cong phản xạ phổ (reflectance curve) hay mẫu phổ (spectral pattern). Các mẫu phổ đặc trưng cho từng loại hình lớp phủ được xác định dựa trên các bất biến ảnh (image invariant) đã được thống kê từ trước dưới dạng số. Tập hợp các bất biến ảnh cho từng loại hình lớp phủ tạo nên bản chỉ tiêu phân loại. Nguyên tắc xây dựng bản chỉ tiêu phân loại lớp phủ được Nguyễn Đình Dương nêu trong [2]. Đây là điểm khác biệt cơ bản giữa các phương pháp phân loại truyền thống với thuật toán GASC. Trong các phương pháp phân loại thông thường, các đối tượng được nhận biết qua tệp mẫu, kết quả phân loại phụ thuộc chủ yếu vào chất lượng tệp mẫu đã lựa chọn. Tuy nhiên tệp mẫu sẽ không ổn định và thay đổi từ ảnh này đến ảnh khác và đồng thời khó phản ảnh được các tính chất đa thời gian của đối tượng, nên rất khó sử dụng trong thực tế. Nếu phân loại lớp phủ dựa trên yêu cầu của IGBP - Chương trình Địa sinh quyển quốc tế với 17 đối tượng chính và tư liệu MODIS bản chắp 32 ngày thì bản chú giải số có dụng lượng khoảng 1,5 Gb. Trong thực tế, để có được vùng mẫu ổn định, các tệp mẫu được lựa chọn dựa trên thực trạng lớp phủ ngoài thực địa và không dựa trên bất kỳ tính chất nào của tư liệu sử dụng. Tập hợp các vùng mẫu như thế được gọi là dữ liệu mẫu thực địa (ground truth). Trên *hình 2* là dữ liệu mẫu thực địa cho vùng Việt Nam được xây dựng dựa trên ảnh chụp mặt đất có gắn toạ độ (GPS photo). Việc phân loại được thực hiện tuần tự từng điểm ảnh một. Nếu thực hiện trên một máy tính PC với tốc độ CPU 2Mhz và bộ nhớ 2 Gb, thời gian phân loại để đạt được kết quả cho riêng Việt Nam như trên *hình 3* là khoảng 6 giờ. Thông tin về lớp phủ là rất quan trọng cho nghiên cứu môi trường khu vực hoặc toàn cầu, do vậy các vùng nghiên cứu lớn ví dụ cho toàn bộ Đông nam á, châu á hoặc toàn bộ khu vực nhiệt đới cũng rất phổ biến. Với các khu vực nghiên cứu lớn như vậy, việc phân loại lớp phủ sẽ diễn ra rất lâu. Chính vì lẽ đó mà cần các môi trường tính toán phù hợp hơn cho phép rút ngắn thời gian tính toán.



Hình 2. Cơ sở dữ liệu mẫu thực địa

III. HỆ THỐNG GEOGRID CỦA AIST NHẬT BẢN VÀ KHẢ NĂNG KẾT NỐI

Viện Khoa học Công nghiệp tiên tiến Nhật Bản AIST đã có hợp tác song phương chính thức với Viện Khoa học và Công nghệ Việt Nam. Trong khuôn khổ hợp tác này, ngoài các vấn đề trong lĩnh vực sinh học, môi trường, các nhà khoa học Việt Nam tham gia dự án GeoGRID, một mặt có điều kiện sử dụng các tư liệu ảnh vệ tinh và mô hình số độ cao toàn cầu với độ phân giải 15m, mặt khác có thể kết nối với hệ thống tính toán lưới của AIST.

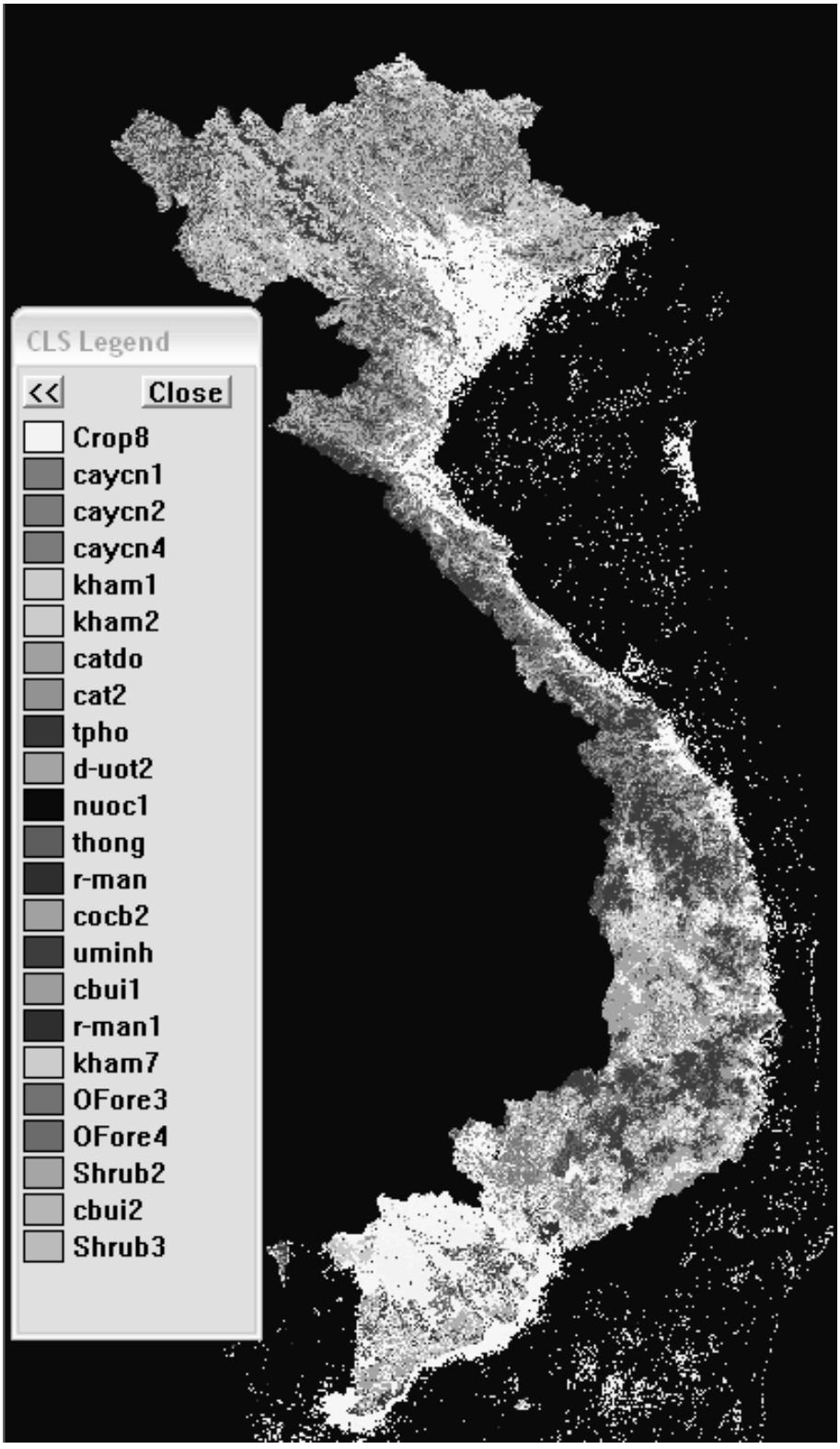
Hệ thống siêu máy tính bó của AIST (Super cluster computers) được xây dựng từ 3.000 CPU tạo nên siêu máy tính bó lớn nhất Nhật Bản (*hình 4*).

Hệ thống siêu máy tính này được tổ chức thành ba tiểu hệ thống : hệ thống P-32 8.59TFLOPS được dành riêng cho tính toán hiệu năng cao, hệ thống M-64 2.72TFLOPS có khả năng lưu trữ dữ liệu cao nhất và hệ thống F-32 3.13TFLOPS dành cho các tính toán song song phức tạp nhất. Hệ thống P-32 và M-64 sử dụng kết nối Myrinet dành riêng cho tính toán với mạng siêu tốc 10-MGbps. Hệ thống lưu trữ có dụng lượng 20 TB dựa trên các ổ đĩa cứng có nhiều ưu thế hơn so với hệ thống sử dụng băng từ. Hệ thống máy tính này được điều khiển bằng hệ điều hành Linux cùng nhiều hệ phụ trợ khác như gFarm, MPI. Các ngôn ngữ lập trình cơ bản là C và FORTRAN. Về cơ bản các ngôn ngữ này tuân thủ theo các ngôn ngữ chuẩn như FORTRAN 77, FOTRAN 90... Tuy nhiên các ngôn ngữ này lại có một số điểm khác biệt so với các ngôn ngữ lập trình trên PC trong hệ điều hành Windows.

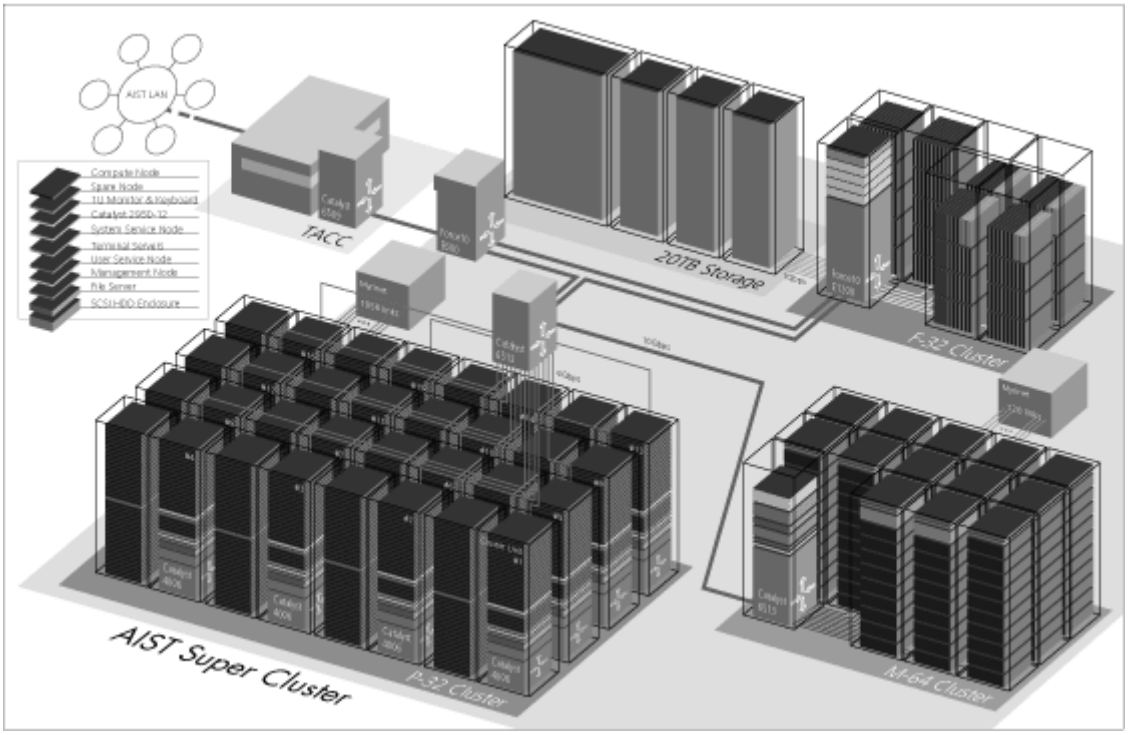
Việc kết nối các máy tính PC trên hệ điều hành Windows với máy tính trong môi trường Linux được thực hiện bằng nhiều giao thức trong đó phổ biến nhất là sử dụng trình PUTTY. Mỗi người sử dụng khi kết nối với Siêu máy tính cần có Keyfingerprint do người quản trị Siêu máy tính cấp và mật khẩu Passkey. Sau khi kết nối thành công, một cửa sổ ở chế độ command line sẽ xuất hiện trên nền màn hình Windows và qua đó chúng ta có thể tiến hành nhập các lệnh của Linux để biên dịch chương trình, tải số liệu và thực hiện tính toán (*hình 5*).

IV. CHUYỂN ĐỔI CHƯƠNG TRÌNH TÍNH TOÁN TỪ MÔI TRƯỜNG WINDOWS SANG LINUX

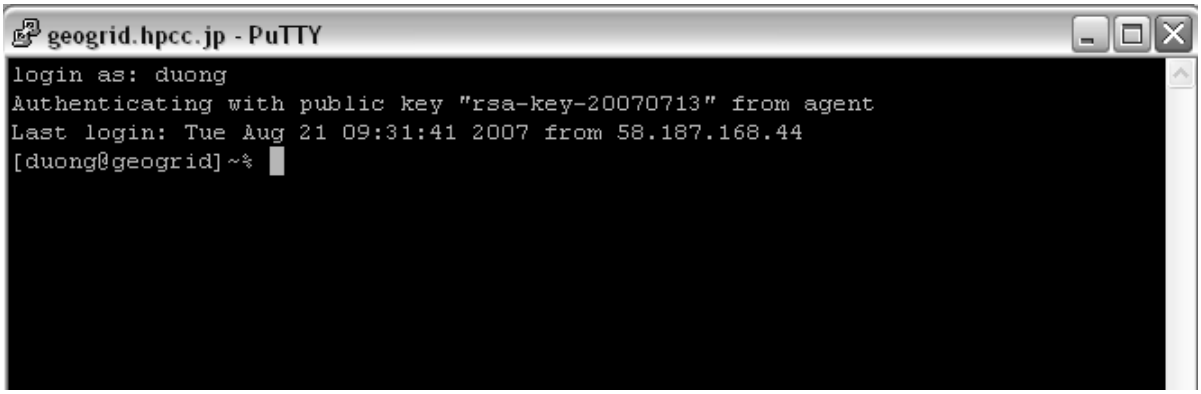
Sau khi kết nối thành công với Siêu máy tính, chúng ta cần tạo không gian làm việc, sử dụng các



Hình 3. Lớp phủ Việt Nam phân loại trên hệ thống tính toán lưới GeoGRID



Hình 4. Hệ thống siêu máy tính bó - Super cluster computers AIST



Hình 5. Cửa sổ kết nối máy tính PC với Siêu máy tính

tài nguyên của Siêu máy tính phục vụ công việc tính toán của mình. Làm việc trên Siêu máy tính đòi hỏi phải tuân thủ các quy định của người quản lý và cần biết các node nào được sử dụng để lưu trữ, chạy thử hay tính toán thực tế. Chương trình phân loại tự động lớp phủ theo thuật toán GASC được viết bằng ngông ngữ Compact FORTRAN 90 với một số điểm khác biệt với gFORTRAN. Để giảm tối đa những sửa đổi có thể, trình biên dịch Intel FORTRAN đã được cài đặt. Intel FORTRAN 90 tương thích hầu như toàn bộ với Compact FORTRAN 90. Tuy nhiên một có số lệnh không tương thích hoàn toàn và cần phải sửa đổi. Có thể kể đến như sau :

- Lệnh đọc biến số tính toán : trong khi các chương trình trên Windows sử dụng hai hàm NARGS và GETARG để xác định số biến số đầu vào thì trên Linux cần thay thế bằng hai lệnh COMMAND_ARGUMENT_COUNT và GET_COMMAND_ARGUMENT.

- Các lệnh mở tệp về cơ bản không khác nhau nhưng đối với chế độ nhị phân thì cần sửa đổi. Trong khi trên hệ điều hành Windows có chế độ nhị phân BINARY, nhưng trên Linux không có và thay thế vào đó là chế độ STREAM. Để mở tệp ảnh ở chế độ nhị phân trên Windows khi sử dụng lệnh OPEN chỉ cần thông báo FORM='BINARY', còn trên Linux cần thêm hai tham biến ACCESS= 'STREAM' và FORM='UNFORMATTED'.

- Trong một số trường hợp cần in lên màn hình quá trình tính toán để kiểm tra, đặc biệt khi chạy thử, khi đó việc điều khiển các thiết bị đầu cuối trên Windows và Linux có phần khác nhau. Đối với các chương trình trên Windows đơn giản cho ký tự + như ký tự đầu tiên trong chuỗi đầu ra. Đối với hệ điều hành Linux thì cần thiết phải khai báo tham biến ADVANCE='NO' trong lệnh WRITE và đồng thời in ký tự đầu tiên bằng ký tự CHAR(13). Các lệnh còn lại phục vụ tính toán hầu như không có gì thay đổi.

Sau khi biên dịch, chương trình có thể chạy bình thường như ở chế độ Command Line trong Windows, tuy nhiên thông thường sẽ phải dùng các lệnh của tính toán lưới để đưa yêu cầu tính toán vào hàng chờ. Trên Siêu máy tính AIST hệ thống tính toán lưới được tổ chức theo Sun GridWare. Phiên bản Sun N1 Grid Engine 6.1 được sử dụng. Các lệnh cơ bản cần nắm vững đó là :

- qsub : đưa một ứng dụng vào hàng chờ
- qstat : kiểm tra trạng thái ứng dụng đang thực hiện
- qmon : đưa một ứng dụng vào hàng chờ tính toán thông qua giao diện đồ họa
- qdel : kết thúc một ứng dụng đang thực hiện
- qmod : tạm dừng hoặc tiếp tục một ứng dụng đang thực hiện

Trong thực tế việc thực hiện tính toán trong môi trường tính toán lưới phức tạp hơn nhiều và người sử dụng cần tìm hiểu kỹ các tài liệu hướng dẫn sử dụng [3].

V. MỘT SỐ KẾT QUẢ THỬ NGHIỆM ĐẦU TIÊN

Trong khuôn khổ đề tài khoa học cơ bản giai đoạn 2006-2008, tác giả thử nghiệm việc áp dụng tính toán lưới trong điều kiện Việt Nam áp dụng cho phân loại lớp phủ từ tư liệu viễn thám đa thời gian. Bên cạnh các vấn đề kỹ thuật liên quan tới lập trình trong môi trường Linux, kết nối giữa PC Windows và Linux và làm việc với môi trường tính toán lưới thông qua một hệ thống cụ thể GridWare nào đó (ví dụ Sun N1 Grid Engine 6.1) cần tìm hiểu hạ tầng cơ sở mạng hiện đang phổ biến ở Việt Nam. Môi trường mạng kết nối Internet phổ biến ở Việt Nam hiện nay là mạng ADSL với các tốc độ truy cập khác nhau cung cấp bởi các nhà phân phối chính là VNPT, FPT và Vietel. Mặc dù tốc độ truy cập không ổn định nhưng về cơ bản cơ sở hạ tầng mạng cho phép triển khai tính toán lưới đối với các bài toán có nhu cầu chuyển tải dữ liệu ít. Tệp dữ liệu đầu vào cho phân loại lớp phủ bằng chương trình Gasc_g07m.f90 bao gồm ba tệp chính :

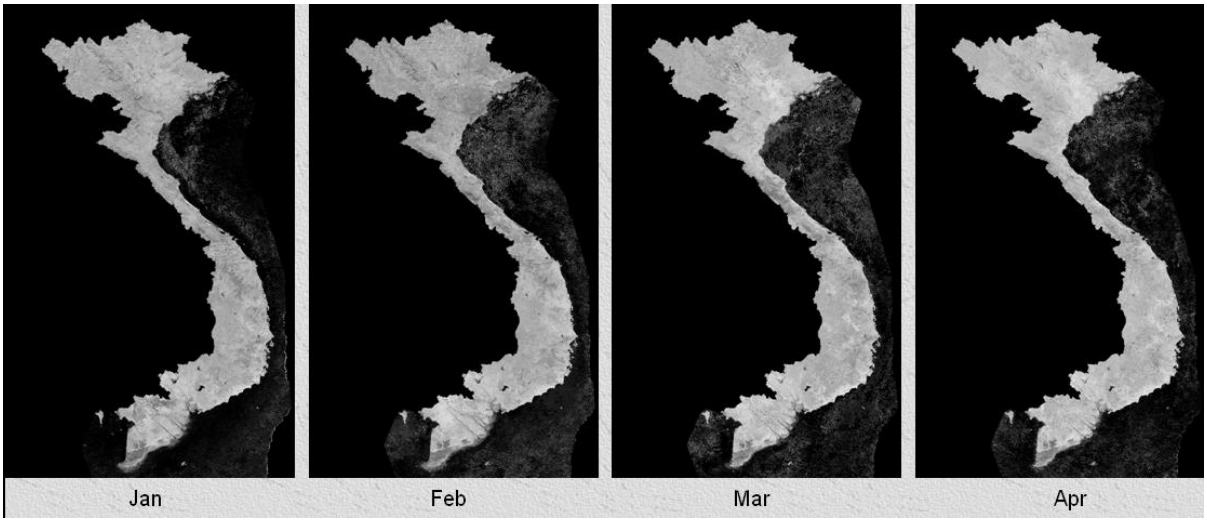
- Tệp điều khiển tính toán
- Tệp dữ liệu ảnh đa thời gian
- Tệp chú giải phân loại số

Tệp điều khiển tính toán cung cấp cho chương trình các thông tin cơ bản như số kênh phổ, số thời điểm tư liệu được sử dụng, số byte cho mỗi điểm ảnh, số hàng số cột cho tệp tư liệu ảnh, tên tệp chú giải, yêu cầu về lọc nhiễu...

Tệp dữ liệu ảnh đa thời gian được sử dụng là bản chắp (composite) tư liệu MODIS 32 ngày do Trường tổng hợp Maryland, Mỹ cung cấp. Số thời điểm tư liệu được sử dụng là 7. Bản chú giải được tính dựa trên cơ sở dữ liệu mẫu thực địa và hệ thống phân loại của IGBP sau khi sửa đổi cho phù hợp với điều kiện Việt Nam. Trên *hình 6* là ví dụ về bản chắp ảnh không mây Việt Nam cho các tháng 1, 2, 3 và 4.

Tệp chú giải phân loại số là tập hợp các bất biến ảnh, dựa theo đó có thể xác định được các loại hình lớp phủ khác nhau. Bản chú giải chứa các thông tin phổ và cả các thông tin biến đổi theo thời gian của mỗi loại hình lớp phủ. Ví dụ về bản chú giải phân loại số bằng chương trình Gasc_g07m.f90 có thể thấy trên *hình 7*. Chúng ta thấy trong bản chú giải này thông tin đầu tiên và quan trọng nhất là biến điệu của đường cong phản xạ phổ cho từng loại hình lớp phủ, kế đến là vec tơ trung bình cho mỗi loại hình lớp phủ theo các thời điểm tư liệu khác nhau, sau đó là chỉ số TRRI (tổng năng lượng phản xạ) cho từng thời điểm và cuối cùng là tổ hợp tỷ số giữa hiệu và tổng của các kênh phổ thành phần.

Sau khi chuyển tải toàn bộ chương trình Gasc_ g07m.f90 lên mạng GeoGRID, biên dịch trong môi trường Linux, cùng với các số liệu cần thiết cho việc



Hình 6. Ví dụ bản chắp tư liệu MODIS không mây

```
Graphical analysis of spectral reflectance curve
Multitemporal version
  5
    1
rung1
rung1
  0  80   0        829
M  11111
  424.8132   324.8661  2300.4592  1356.0774   583.0683
  412.1325   331.7721  1971.6146  1247.6078   570.1266
  536.4344   383.5571  2646.7390  1567.6307   682.7893
  619.6926   450.5620  3005.2449  1700.3873   776.8840
  603.3643   428.8275  3077.7297  1685.1666   751.1578
  584.2820   411.8868  3305.8127  1828.1244   760.9752
  457.3089   335.8925  2643.3738  1502.6892   602.6566
T 01        634.38   1729.38
T 02        654.00   1626.75
T 03        825.75   2020.12
T 04        861.75   2159.75
T 05        886.88   2211.75
T 06        746.62   2245.75
T 07        712.12   1806.38
C 011212 -+  3.0400001E-02  0.2522876
C 011213 -+  9.8039219E-03  5.2481461E-02
```

Hình 7. Trích bản chú giải cho phân loại số lớp phủ bề mặt

phân loại lớp phủ cho vùng đất liền Việt Nam. Trong giai đoạn đầu tiên, việc áp dụng tính toán song song dựa trên MPI (Message Passing Interface) chưa được thử nghiệm nên về cơ bản thời gian tính toán tương tự như trên các máy Windows. Về thực chất mới chỉ có 1 CPU được khai thác. Kết quả tính được trên Linux và trên Windows là hoàn toàn trùng khớp. Trên hình 3 là kết quả phân loại lớp phủ Việt Nam năm 2003 được tính trong môi trường AIST GeoGRID.

Khi xem xét kết quả phân loại vùng biển ta thấy có rất nhiều nhiễu gây bởi sự biến động phản xạ của mặt nước. Những nhiễu này có thể được lọc dễ dàng dựa trên một mặt nạ (mask) tạo bởi đường bờ nhằm phân chia vùng đất liền và biển.

KẾT LUẬN

Những kết quả đạt được như đã trình bày trong bài báo này cho thấy thuật toán GASC và chương trình Gasc_mg07m.F90 có thể triển khai cho môi trường tính toán lưới trong phân loại lớp phủ từ tư liệu MODIS. Những kết quả đạt được đầu tiên đã minh chứng khả năng kết nối các máy tính Windows với siêu máy tính thông qua mạng Internet và hạ tầng cơ sở thông tin ADSL của Việt Nam hiện nay. Mặc dù với các môi trường tính toán đơn giản như Windows vẫn có thể giải quyết nhiều bài toán kỹ thuật nhưng với nhiều bài toán phức tạp trên góc độ thuật toán hoặc yêu cầu phải xử lý một khối lượng số liệu rất lớn thì môi trường tính toán lưới với các máy tính bó có khả năng phân chia một bài toán lớn thành nhiều bài toán chạy song song là tối ưu hơn cả.

Việc khai thác môi trường tính toán lưới cho dù trên một mạng cục bộ hay trên mạng diện rộng đều là điều cần thiết đối với nhiều ngành khoa học kỹ thuật nói chung và các khoa học về Trái Đất nói riêng. Mặc dù khi chuyển từ môi trường Windows sang môi trường Linux, GridWare, MPI hoặc GridMPI thông qua kết nối Windows - Linux bằng PUTTY hoặc một giao diện nào đó còn chưa thông dụng và có thể mang đến một số khó khăn, nhưng đó chỉ là những khó khăn hình thức ban đầu và sẽ dễ dàng có thể vượt qua được.

Trong bối cảnh ở Việt Nam bắt đầu hình thành một số trung tâm tính toán hiệu năng cao với các máy tính bó hiện đại thì việc triển khai ứng dụng tính toán lưới nên được đẩy mạnh ứng dụng trong các lĩnh vực khoa học khác nhau. Những kết quả đạt được và giới thiệu trong bài báo này đã gợi mở cho việc hợp tác quốc tế trong trường hợp các cơ sở tính toán của Việt Nam chưa đáp ứng được những yêu cầu kỹ thuật cần thiết về tính toán lưới.

Lời cám ơn : Bài báo công bố các kết quả nghiên cứu của đề tài mã số 700906 thuộc Chương trình khoa học cơ bản. Tác giả cám ơn sự hỗ trợ kinh phí nghiên cứu của chương trình.

TÀI LIỆU DẪN

[1] NGUYEN DINH DUONG, 1997 : Graphical Analysis of Spectral Reflectance Curve. Proceedings of the 18th Asian Conference on Remote Sensing, Kuala Lumpur, Malaysia.

[2] NGUYEN DINH DUONG, 2000 : Land Cover Category Definition by Image Invariants for Automated Classification, International Archives of Photogrammetry and Remote Sensing, XXXIII, B7/3.

[3] SunN1Grid Engine 6.1User's Guide. Sun Microsystem. May 2007.

SUMMARY

A preliminary research result in application of grid computing in classification of land cover

Classification of land cover is one of the most basic analysis of remote sensing data. In case of small study area where table of legend is small and the input data is not so complicated, the classification of land cover could be carried out using a single compu-ter with limited resources. However, in case of larger study area such as country-wide, continental or global coverage as well, the classification of land cover requires faster computer with sophisticated resources or suitable algorithm implementation. In this paper, the author introduces application of grid computing in classification of land cover. The grid computing technology helps us avoiding investment into larger local computers and better exploitation of distributed computing resources over the internet for highly optimal computation needs. The author introduces the Super cluster in the National Institute of Advanced In-dustrial Science and Technology (AIST), the GeoGRID system, connection between Windows PC and the Linux cluster, transferring data and program codes to Linux, compiling and execution of classification of land cover on the Linux cluster. This is only a prelimina-ry research result to demonstrate grid computing application. In the next steps, the use of GridMPI for parallel computing will be applied to achieve high performance of grid computing in classification of land cover.

Ngày nhận bài : 15-01-2008

Viện Địa lý-Viện Khoa học và Công nghệ Việt Nam