ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
BAN QUẢN LÝ DỰ ÁN 11-P04-VIE

Dự án

NGHIÊN CỨU THUỶ TAI DO BIẾN ĐỔI KHÍ HẬU VÀ XÂY DỰNG HỆ THỐNG THÔNG TIN NHIỀU BÊN THAM GIA NHẰM GIẢM THIỂU TÍNH DỄ BỊ TỔN THƯƠNG Ở BẮC TRUNG BỘ VIỆT NAM (CPIS)

Mã số: 11.P04.VIE

(***Thuộc Chương trình thí điểm hợp tác nghiên cứu Việt Nam - Đan hạch 2012-2015***)

BÁO CÁO KẾT QUẢ THỰC HIỆN NĂM 2012-2013

Nội dung 3: ***Phân tích nhu cầu người dùng***

Nhóm nghiên cứu: WP6

Chủ dự án: ***Trường Đại học Khoa học Tự nhiên***
Giám đốc dự án: ***GS. TS. Phan Văn Tân***

Những người thực hiện:

Trưởng nhóm: *ThS. Nguyễn Trung Kiên*
Các thành viên: *TS. Bùi Quang Thành*
CN. Nguyễn Quốc Huy
ThS. Phan Văn Trọng
CN. Đoàn Thị The

Phân tích nhu cầu người dùng

Họ và tên chuyên gia: Đoàn Thị The

1. Mở đầu

Hệ thống thông tin là sản phẩm trực quan và gần gũi với người sử dụng nhất trong khuôn khổ dự án "Nghiên cứu thuỷ tai do biến đổi khí hậu và xây dựng hệ thống thông tin nhiều bên tham gia (CPIS) nhằm giảm thiểu tính dễ bị tổn thương ở Bắc Trung Bộ Việt Nam". Để thiết kế và xây dựng được hệ thống đáp ứng tốt nhất nhu cầu của người dùng, công đoạn khảo sát và phân tích nhu cầu đóng vai trò rất quan trọng. Các nhu cầu sử dụng bao gồm nhu cầu của người dân, nhu cầu của người làm quản lý và nhu cầu của các nhà khoa học. Xét chung lại, các nhu cầu này được chia ra làm hai loại, nhu cầu xây dựng hệ thống GIS và nhu cầu xây dựng hệ thống tính toán mạnh, đáp ứng yêu cầu chạy các mô hình mô phỏng khí hậu, thuỷ văn. Báo cáo này sẽ làm rõ hai nhu cầu trên.

2. GIS và nhu cầu sử dụng

2.1.Khái niệm:

G = Geographic = Địa lý: Dữ liệu dùng trong GIS là dữ liệu địa lý. GIS có thể trình bày dữ liệu dưới dạng bản đồ

I = Information = Thông tin: GIS lưu trữ và xử lý hai loại thông tin: Đặc trưng không gian và thuộc tính

S = System = Hệ thống: GIS là một hệ thống được sử dụng để thực hiện các chức năng khác nhau của thông tin địa lý.

Khái niệm "thông tin" đề cập đến phần dữ liệu được quản lý bởi GIS. Đó là các dữ liệu về thuộc tính và không gian của đối tượng. GIS có tính "hệ thống" tức là hệ thống GIS được xây dựng từ các mô đun. Việc tạo các mô đun giúp thuận lợi trong việc quản lý và hợp nhất.

GIS là một hệ thống có ứng dụng rất lớn. Từ năm 1980 đến nay đã có rất nhiều các định nghĩa được đưa ra, tuy nhiên không có định nghĩa nào khái quát đầy đủ về GIS vì phần lớn chúng đều được xây dựng trên khía cạnh ứng dụng cụ thể trong từng lĩnh vực. Có ba định nghĩa được dùng nhiều nhất

- GIS là một hệ thống thông tin được thiết kế để làm việc với các dữ liệu trong một hệ toạ độ quy chiếu. GIS bao gồm một hệ cơ sở dữ liệu và các phương thức để thao tác với dữ liệu đó.
- GIS là một hệ thống nhằm thu thập, lưu trữ, kiểm tra, tích hợp, thao tác, phân tích và hiển thị dữ liệu được quy chiếu cụ thể vào trái đất.
- GIS là một chương trình máy tính hỗ trợ việc thu thập, lưu trữ, phân tích và hiển thị dữ liệu bản đồ.

Các quan niệm khác nhau về GIS:

- Một bản đồ thông minh
- Một cơ sở dữ liệu kết nối giữa các đặc trưng và thuộc tính

- Các công cụ dùng để phân tích, biên tập, và quản lý dữ liệu địa lý

2.2.Định nghĩa GIS

GIS là một hệ thống dùng để trình bày, lưu trữ, quản lý, và phân tích dữ liệu về các đối tượng trên bề mặt trái đất.

* Thành phần chính của GIS

Gồm 5 Thành phần chính:

- Con người
- Dữ liệu
- Phương pháp phân tích
- Phần mềm
- Phần cứng

Các thành phần này kết hợp với nhau nhằm tự động quản lý và phân phối thông tin thông qua biểu diễn địa lý.

+ Con người

Con người là thành phần quan trọng nhất, là nhân tố thực hiện các thao tác điều hành sự hoạt động của hệ thống GIS.

Người dùng GIS là những người sử dụng các phần mềm GIS để giải quyết các bài toán không gian theo mục đích của họ. Họ thường là những người được đào tạo tốt về lĩnh vực GIS hay là các chuyên gia.

Người xây dựng bản đồ: sử dụng các lớp bản đồ được lấy từ nhiều nguồn khác nhau, chỉnh sửa dữ liệu để tạo ra các bản đồ theo yêu cầu.

Người xuất bản: sử dụng phần mềm GIS để kết xuất ra bản đồ dưới nhiều định dạng xuất khác nhau.

Người phân tích: giải quyết các vấn đề như tìm kiếm, xác định vị trí…

Người xây dựng dữ liệu: là những người chuyên nhập dữ liệu bản đồ bằng các cách khác nhau: vẽ, chuyển đổi từ định dạng khác, truy nhập CSDL…

Người quản trị CSDL: quản lý CSDL GIS và đảm bảo hệ thống vận hành tốt.

Người thiết kế CSDL: xây dựng các mô hình dữ liệu lôgic và vật lý.

Người phát triển: xây dựng hoặc cải tạo các phần mềm GIS để đáp ứng các nhu cầu cụ thể.

+ Dữliệu

Một cách tổng quát, người ta chia dữ liệu trong GIS thành 2 loại:

- Dữ liệu không gian (spatial) cho ta biết kích thước vật lý và vị trí địa lý của các đối tượng trên bề mặt trái đất.
- Dữ liệu thuộc tính (non-spatial) là các dữ liệu ở dạng văn bản cho ta biết thêm thông tin thuộc tính của đối tượng.

+ Phần cứng

Là các máy tính điện tử: PC, mini Computer, MainFrame … là các thiết bị mạng cần thiết khi triển khai GIS trên môi trường mạng. GIS cũng đòi hỏi các thiết bị ngoại vi đặc biệt cho việc nhập và xuất dữ liệu như: máy số hoá (digitizer), máy vẽ (plotter), máy quét (scanner)…

+Phần mềm

Hệ thống phần mềm GIS rất đa dạng. Mỗi công ty xây dựng GIS đều có hệ phần mềm riêng của mình. Tuy nhiên, có một dạng phần mềm mà các công ty phải xây dựng là hệ quản trị CSDL địa lý. Dạng phần mềm này nhằm mục đích nâng cao khả năng cho các phần mềm CSDL thương mại trong việc: sao lưu dữ liệu, định nghĩa bảng, quản lý các giao dịch do đó ta có thể lưu các dữ liệu đồ địa lý dưới dạng các đối tượng hình học trực tiếp trong các cột của bảng quan hệ và nhiều công việc khác.

2.3.Nhu cầu sử dụng GIS

Việc sử dụng bản đồ giấy thông thường có một loạt các nhược điểm cho người sử dụng trong việc thể hiện, thao tác, xử lý các dữ liệu thông tin, cụ thể như:

- Không có khả năng thay đổi tỷ lệ bản đồ (vì tỷ lệ này là cố định khi bản đồ được in ra).

- Không có khả năng hiển thị lớp thông tin chuyên đề (layer) riêng mà người sử dụng quan tâm.

- Khó khăn trong việc chuyển đổi từ hệ toạ độ này sang hệ toạ độ khác.

-Việc cập nhật thông tin vào trong bản đồ rất khó khăn và mất nhiều thời gian.

- Khó khăn trong việc thực hiện các phân tích về số, về lượng.

- Khu vực quan tâm luôn luôn nằm tại vị trí giao nhau của 4 tấm bản đồ (vấn đề này được biết đến như là ‘luật Murphy’).

- Không có khả năng thay đổi cách hiển thị các đối tượng, các đặc điểm đã được vẽ.

- Sản xuất bản đồ theo nhu cầu riêng vô cùng tốn kém.

Do đó, các nhà nghiên cứu và quản lý tài nguyên dần dần đã nhận ra rằng cần thiết phải cải thiện phương pháp xử lý các thông tin địa lý, điều này đã dẫn tới sự ra đời và đưa GIS vào thực tiễn.

Kỹ thuật GIS là một công nghệ ứng dụng các tiến bộ của khoa học máy tính, (computer based technology) do đó việc sử dụng GIS trong các mục tiêu nghiên cứu so với các phương tiện cổ điển có thể mang lại những hiệu quả cao do:

- Là cách tiết kiệm chi phí và thời gian nhất trong việc lưu trữ số liệu

- Có thể thu thập số liệu với số lượng lớn

- Số liệu lưu trữ có thể được cập nhật hoá một cách dễ dàng

- Chất lượng số liệu được quản lý, xử lý và hiệu chỉnh tốt

Dễ dàng truy cập, phân tích số liệu từ nhiều nguồn và nhiều loại khác nhau

- Tổng hợp một lần được nhiều loại số liệu khác nhau để phân tích và tạo ra nhanh chóng một lớp số liệu tổng hợp mới

GIS làm thay đổi đáng kể tốc độ mà thông tin địa lý được sản xuất, cập nhật và phân phối. GIS cũng làm thay đổi phương pháp phân tích dữ liệu địa lý, GIS có một số điểm thuận lợi chính khi được so sánh với cách quản lý bản đồ giấy là:

- Dễ dàng cập nhật thông tin không gian.

- Tổng hợp hiệu quả nhiều tập hợp dữ liệu thành một cơ sở dữ liệu kết hợp

- Chúng có thể cho ra những kết quả dưới những dạng khác nhau như các bản đồ, biểu bản, và các biểu đồ thống kê,..

- Dữ liệu không gian địa lý được duy trì tốt hơn trong một định dạng tiêu chuẩn.

- Việc xem lại và cập nhật dễ dàng hơn.

- Tìm kiếp, phân tích và miêu tả thuận lợi hơn.

- Dữ liệu có thể chia sẻ và trao đổi.

- Tiết kiệm thời gian và tiền bạc.

- Đưa ra những quyết định tốt và đúng đắn hơn.

GIS lưu giữ thông tin về thế giới thực dưới dạng tập hợp các lớp chuyên đề có thể liên kết với nhau nhờ các đặc điểm địa lý. Điều này đơn giản nhưng vô cùng quan trọng và là một công cụ đa năng đã được chứng minh là rất có giá trị trong việc giải quyết nhiều vấn đề thực tế, từ thiết lập tuyến đường phân phối của các chuyến xe, đến lập báo cáo chi tiết cho các ứng dụng quy hoạch, hay mô phỏng sự lưu thông khí quyển toàn cầu.

Với những tính năng ưu việt, kỹ thuật GIS ngày nay đang được ứng dụng trong nhiều lãnh vực nghiên cứu và quản lý, đặc biệt trong quản lý và quy hoạch sử dụng-khai thác các nguồn tài nguyên một cách bền vững và hợp lý.

2.4.Các chức năng, lĩnh vực và cấp độ ứng dụng GIS

** Các chức năng của GIS*

GIS có 3 chức năng chính:

- Hiện thị
- Quản lý
- Phân tích dữ liệu địa lý

Hiển thị là việc nhìn vào dữ liệu trên bản đồ để thu nhận thông tin và xem xét các mối quan hệ. Bản đồ hiển thị có thể được kết hợp với các bản báo cáo, hình ảnh ba chiều, ảnh chụp và những dữ liệu khác (đa phương tiện).

Quản lý là việc tổ chức và cập nhật dữ liệu địa lý làm cho chúng hữu dụng hơn

Phân tích là việc dùng các công cụ GIS tác động vào dữ liệu địa lý để trả lời các câu hỏi và đưa ra quyết định.

Một khi đã có một hệ GIS lưu giữ các thông tin địa lý, có thể bắt đầu hỏi các câu hỏi đơn giản như:

- Ai là chủ mảnh đất ở góc phố?
- Vùng đất dành cho hoạt động công nghiệp ở đâu?

Và các câu hỏi phân tích như:

- Tất cả các vị trí thích hợp cho xây dựng các toà nhà mới nằm ở đâu?
- Nếu xây dựng một đường quốc lộ mới ở đây, giao thông sẽ chịu ảnh hưởng như thế nào

GIS cung cấp cả khả năng hỏi đáp đơn giản "chỉ và nhấn" và các công cụ phân tích tinh vi để cung cấp kịp thời thông tin cho những người quản lý và phân tích.

- *Các ứng dụng của GIS*

  + Lập bản đồ (Mapping),

+ Đo lường (Measuring),

+ Theo dõi (Monitoring),

+ Mô hình hoá (Modeling), và

+ Quản lý (Managing).

- Nhằm mục đích:

+ Theo dõi phân bố không gian

+ Tìm hiểu các quá trình

+ Xây dựng các chiến lược phát triển

Vì GIS được thiết kế như một hệ thống chung để quản lý dữ liệu không gian, nó có rất nhiều ứng dụng trong việc phát triển đô thị và môi trường tự nhiên như là: quy hoạch đô thị, quản lý nhân lực, nông nghiệp, điều hành hệ thống công ích, lộ trình, nhân khẩu, bản đồ, giám sát vùng biển, cứu hoả và bệnh tật. Trong phần lớn các lĩnh vực này, GIS đóng vai trò như là một công cụ hỗ trợ quyết định cho việc lập kế hoạch hoạt động.

a. Chính phủ

- Là người sử dụng chính của GIS: 70-80% công việc của chính quyền địa phương có liên quan đến địa lý.

- Rất nhiều ứng dụng như: Kiểm kê tài sản; Quy hoạch giao thông; Quản lý đất đai; Phát triển kinh tế; Bầu cử; Sức khoẻ cộng đồng

Ví dụ: Tăng doanh thu từ thuế bất động sản, duy trì sổ đăng ký tài sản (dạng số hoá), định giá (Giá thay thế, Giá thị trường so sánh với các cuộc mua bán gần thời gian), giải quyết thắc mắc, khiếu nại của người dân.

b. Môi trường

Theo những chuyên gia GIS kinh nghiệm nhất thì có rất nhiều ứng dụng đã phát triển trong những tổ chức quan tâm đến môi trường. Với mức đơn giản nhất thì người dùng sử dụng GIS để đánh giá môi trường, ví dụ như vị trí và thuộc tính của cây rừng. Ứng dụng GIS với mức phức tạp hơn là dùng khả năng phân tích của GIS để mô hình hóa các tiến trình xói mòn đất sư lan truyền ô nhiễm trong môi trường khí hay nước, hoặc sự phản ứng của một lưu vực sông dưới sự ảnh hưởng của một trận mưa lớn. Nếu những dữ liệu thu thập gắn liền với đối tượng vùng và ứng dụng sử dụng các chức năng phân tích phức tạp thì mô hình dữ liệu dạng ảnh (raster) có khuynh hướng chiếm ưu thế.

- Kiểm kê nguồn tài nguyên thiên nhiên

- Phân tích tác động môi trường

- Giảm nhẹ rủi ro môi trường

- Phát triển bền vững

- Nông nghiệp chính xác

c. Khí tượng thuỷ văn

Trong lĩnh vực này GIS được dùng như là một hệ thống đáp ứng nhanh, phục vụ chống thiên tai như lũ quét ở vùng hạ lưu, xác định tâm bão, dự đoán các luồng chảy, xác định mức độ ngập lụt, từ đó đưa ra các biện pháp phòng chống kịp thời... vì những

ứng dụng này mang tính phân tích phức tạp nên mô hình dữ liệu không gian dạng ảnh (raster) chiếm ưu thế.

d. Nông nghiệp

Những ứng dụng đặc trưng: Giám sát thu hoạch, quản lý sử dụng đất, dự báo về hàng hoá, nghiên cứu về đất trồng, kế hoạch tưới tiêu, kiểm tra nguồn nước.

e. Dịch vụ tài chính

GIS được sử dụng trong lĩnh vực dịch vụ tài chính tương tự như là một ứng dụng đơn lẻ. Nó đã từng được áp dụng cho việc xác định vị trí những chi nhánh mới của Ngân hàng. Hiện nay việc sử dụng GIS đang tăng lên trong lĩnh vực này, nó là một công cụ đánh giá rủi ro và mục đích bảo hiểm, xác định với độ chính xác cao hơn những khu vực có độ rủi ro lớn nhất hay thấp nhất. Lĩnh vực này đòi hỏi những dữ liệu cơ sở khác nhau như là hình thức vi phạm luật pháp, địa chất học, thời tiết và giá trị tài sản.

g. Y tế

Ngoại trừ những ứng dụng đánh gía, quản lý mà GIS hay được dùng, GIS còn có thể áp dụng trong lĩnh vực y tế. Ví dụ như, nó chỉ ra được lộ trình nhanh nhất giữa vị trí hiện tại của xe cấp cứu và bệnh nhân cần cấp cứu, dựa trên cơ sở dữ liệu giao thông. GIS cũng có thể được sử dụng như là một công cụ nghiên cứu dịch bệnh để phân tích nguyên nhân bộc phát và lây lan bệnh tật trong cộng đồng.

h.Chính quyền địa phương

Chính quyền địa phương là một trong những lĩnh vực ứng dụng rộng lớn nhất của GIS, bởi vì đây là một tổ chức sử dụng dữ liệu không gian nhiều nhất. Tất cả các cơ quan của chính quyền địa phương có thể có lợi từ GIS. GIS có thể được sử dụng trong việc tìm kiếm và quản lý thửa đất, thay thế cho việc hồ sơ giấy tờ hiện hành. Nhà cầm quyền địa phương cũng có thể sử dụng GIS trong việc bảo dưỡng nhà cửa và đường giao thông. GIS còn được sử dụng trong các trung tâm điều khiển và quản lý các tình huống khẩn cấp.

i. Bán lẻ và phân phối

Phần lớn siêu thị vùng ngoại ô được xác định vị trí với sự trợ giúp của GIS. GIS thường lưu trữ những dữ liệu về kinh tế-xã hội của khách hàng trong một vùng nào đó. Một vùng thích hợp cho việc xây dựng môt siêu thị có thể được tính toán bởi thời gian đi đến siêu thị, và mô hình hoá ảnh hưởng của những siêu thị cạnh tranh. GIS cũng được dùng cho việc quản lý tài sản và tìm đường phân phối hàng ngắn nhất.

GIS được áp dụng:

+ Tác vụ – xử lý các giao dịch diễn ra hàng ngày (vd. xác định tuyến đường giao hàng)

+ Chiến thuật – phân bổ các nguồn lực giải quyết các vấn đề ngắn và trung hạn (vd. chiến dịch khuyến mại nhằm một đối tượng khách hàng nào đó)

+ Chiến lược – các mục đích và nhiệm vụ dài hạn (vd. lập kế hoạch xây dựng các kho hàng)

k. Giao thông

GIS có khả năng ứng dụng đáng kể trong lĩnh vực vận tải. Việc lập kế hoạch và duy trì cở sở hạ tầng giao thông rõ ràng là một ứng dụng thiết thực, nhưng giờ đây có sự quan tâm đến một lĩnh vực mới là ứng dụng định vị trong vận tải hàng hải, và hải đồ điện tử. Loại hình đặc trưng này đòi hỏi sự hỗ trợ của GIS.

l. Các ngành điện, nước, gas, điện thoại...

Những công ty trong lĩnh vực này là những người dùng GIS linh hoạt nhất, GIS được dùng để xây dựng những cơ sở dữ liệu là cái thường là nhân tố của chiến lược công nghệ thông tin của các công ty trong lĩnh vự này. Dữ liệu vecto thường được dùng trong các lĩnh vực này. những ứng dụng lớn nhất trong lĩnh vực này là Automated Mapping và Facility Management (AM-FM). AM-FM được dùng để quản lý các đặc điểm và vị trí của các cáp, valve... Những ứng dụng này đòi hỏi những bản đồ số với độ chính xác cao.

Một tổ chức dù có nhiệm vụ là lập kế hoạch và bảo dưỡng mạng lưới vận chuyền hay là cung cấp các dịch vụ về nhân lực, hỗ trợ cho các chương trình an toàn công cộng và hỗ trợ trong các trường hợp khẩn cấp, hoặc bảo vệ môi trường, thì công nghệ GIS luôn đóng vai trò cốt yếu bằng cách giúp cho việc quản lý và sử dụng thông tin địa lý một cách hiệu quả nhằm đáp ứng các yêu cầu hoạt động và mục đích chương trình của tổ chức đó.

m. Hậu cần

Có rất nhiều ứng dụng của GIS trong vận tải và hậu cần như: Bảo dưỡng hệ thống cơ sở hạ tầng, kế hoạch đi lại, xác định tuyến đường, theo dõi xe cộ, phương tiện

Ngoài ra, một số ứng dụng cụ thể của GIS thường thấy trong thực tế là:

- Quản lý hệ thống đường phố, bao gồm các chức năng: tìm kiếm địa chỉ khi xác định được vị trí cho địa chỉ phố hoặc tìm vị trí khi biết trước địa chỉ phố. Đường giao thông và sơ đồ; điều khiển đường đi, lập kế hoạch lưu thông xe cộ. Phân tích vị trí, chọn khu vực xây dựng các tiện ích như bãi đỗ xe, ga tàu xe…Lập kế hoạch phát triển giao thông.

- Quản lý giám sát tài nguyên, thiên nhiên, môi trường bao gồm các chức năng: quản lý gió và thuỷ hệ, các nguồn nhân tạo, bình đồ lũ, vùng ngập úng, đất nông nghiệp, tầng ngập nước, rừng, vùng tự nhiên, phân tích tác động môi trường… Xác định ví trí chất thải độc hại. Mô hình hoá nước ngầm và đường ô nhiễm. - -- Quản lý quy hoạch: phân vùng quy hoạch sử dụng đất. Các hiện trạng xu thế môi trường. Quản lý chất lượng nước.

- Quản lý các thiết bị: xác định đường ống ngầm, cáp ngầm. Xác định tải trọng của lưới điện. Duy trì quy hoạch các thiết bị, sử dụng đường điện.

- Phân tích tổng điều tra dân số, lập bản đồ các dịch vụ y tế, bưu điện và nhiều ứng dụng khác.

3. Nhu cầu về tính toán và lưu trữ

Metocean là một cluster cỡ nhỏ nhưng nhu cầu sử dụng lại tương đối lớn. Số lượng người dùng hệ thống thường xuyên lên tới trên 50, trong đó có 10 cán bộ nghiên cứu, 2-3 nghiên cứu sinh, 5-6 học viên cao học, còn lại là sinh viên đại học.

Hệ thống tính toán được sử dụng vào nhiều mục đích, trong đó chủ yếu là chạy các mô hình số trị, bao gồm cả những mô hình sử dụng cho đề tài dự báo mùa và các đề tài khác. Các mô hình hiện đang khai thác thường xuyên trên hệ thống gồm có:

- Mô hình MM5 cho dự báo thời tiết 210x210 điểm lưới với bước lưới 36km. MM5 được cấu hình chạy 4 lần 1 ngày (cách nhau 6h), mỗi lần chạy trên 1 node 8 nhân xử lý (80 GFlops) - 8 tiến trình MPI - mất 2h. Mỗi lần chạy, MM5 sinh ra 1 GB số liệu, tuy nhiên lượng dữ liệu này không cần lưu trữ lại sau khi đã được phân tích. Tổng thời gian chạy mô hình MM5 dự báo thời tiết (tính trên 1 node) trong 3 năm đề tài là: 8.760h.
- Mô hình HRM cho dự báo thời tiết với cấu hình miền tính tương đương mô hình MM5 ở trên. HRM cũng được thiết lập để chạy 4 lần 1 ngày, thời gian chạy 1 lần trên 1 node tính cũng khoảng 2h. Tổng thời gian chạy HRM trên 1 node trong 3 năm cũng khoảng 8.760h.
- Mô hình WRF cho dự báo thời tiết với cấu hình như 2 mô hình trên. WRF được cấu hình chạy 4 lần/ngày, mỗi lần chạy 20 member để phục vụ dự báo tổ hợp. Mỗi member cần 20 phút để hoàn thành và thời gian chạy cả 20 member trên 1 node 8 tiến trình MPI là khoảng 7h. Tổng thời gian chạy WRF trong 3 năm là 30.660h.
- Mô hình MM5CL (MM5 được chỉnh sửa cho dự báo khí hậu) được thiết lập để chạy cho 2 kịch bản A1B và A2 cho cả dự báo lại và dự báo trong khoảng thời gian từ năm 2001 đến năm 2100 với đầu vào từ mô hình CCSM. Miền tính có 144x130 điểm lưới, độ phân giải 36km, bước thời gian tích phân là 90 giây và bước ghi dữ liệu là 360 phút. MM5CL-CCSM mất 5h cho dự báo 1 tháng (trên 1 node 8 tiến trình MPI) và sinh ra 1 GB số liệu. Tổng thời gian chạy MM5CL (chỉ sử dụng 1 node tính - 80 GFlops) với dữ liệu CCSM là: 200 * 12 * 5 = 12.000h tức 500 ngày hay 1,4 năm. Tổng lượng dữ liệu do mô hình sinh ra và cần lưu trữ để phân tích sau này là khoảng 2,5 TB.
- Mô hình MM5CL chạy dự báo mùa với dữ liệu đầu vào từ mô hình CFS: MM5CL được cấu hình để chạy với miền tính giống trường hợp chạy với đầu vào CCSM. MM5CL-CFS được chạy 4 lần 1 tháng, mỗi lần chạy dự báo cho 7 tháng, thời gian chạy trên 1 node 8 CPU với 8 tiến trình MPI là 35h và sinh ra 6 GB dữ liệu. Tổng thời gian cần thiết để chạy MM5CL-CFS trên 1 node trong cả thời gian đề tài - 3 năm là khoảng 5.000h. Tổng lượng dữ liệu cần lưu trữ trong suốt 3 năm vào khoảng 865 GB.
- Mô hình RegCM4 chạy dự báo mùa với đầu vào CFS: RegCM4-CFS được cấu hình chạy 4 lần/tháng, mỗi lần chạy dự báo cho 7 tháng, thời gian chạy trên 1 node 8 CPU - 8 tiến trình MPI là khoảng 50h và sinh ra 16 GB dữ liệu. Tổng thời gian cần chạy RegCM4-CFS (trên 1 node) trong suốt 3 năm thực hiện đề tài là: 7.200h. Tổng lượng dữ liệu sinh ra là khoảng 2,3 TB.
- Mô hình RegCM4 chạy dò xoáy thuận nhiệt đới (TC Detect): RegCM4-TC được chạy 1 tháng 1 lần, dò xoáy thuận nhiệt đới cho 12 tháng tiếp theo. Mỗi lần chạy trên 1 node - 8 tiến trình MPI mất khoảng 140h và sinh ra 69 GB số liệu. Lượng dữ liệu này được phân tích luôn và không cần lưu trữ lâu dài. Tổng thời gian cần thiết để chạy RegCM4-TC trong suốt chiều dài đề tài là khoảng 5.000h.

- Mô hình CAM chạy dự báo khí hậu cho khoảng thời gian từ năm 1971 tới năm 2100 với 2 kịch bản A1B, A2 cho quy mô toàn cầu và khu vực. Thời gian chạy cho 1 tháng khoảng 120 phút, tổng thời gian cần chạy là khoảng 6.000h với tổng lượng dữ liệu sinh ra là 15 TB.

Ngoài lượng dữ liệu do mô hình sinh ra như đã nói ở trên, hệ thống còn lưu trữ dữ liệu từ các mô hình như CCSM, ROM, ECHAM, dữ liệu tái phân tích, ... Lượng dữ liệu này được sinh ra từ các lần chạy mô hình trong các đề tài trước, hoặc được tải từ internet và cần được lưu trữ lâu dài để tái sử dụng. Tổng dung lượng các dữ liệu này vào khoảng 40 TB.

Như vậy, tổng thời gian chạy các mô hình trong suốt 3 năm, tính trên 1 node 8 nhân CPU (80 GFlops) là khoảng 83.380h. Khi chạy trên hệ thống có 10 node tính toán, tổng thời gian cần thiết để chạy các mô hình chỉ còn: 8.338h, nhưng do hiệu suất của hệ thống chỉ đạt 60%, thời gian thực tế sẽ vào khoảng 13.900h, tức 579 ngày hay 1,6 năm. Tổng lượng dữ liệu cần lưu trữ là khoảng 60 TB.

Có thể nhận thấy rằng, năng lực lưu trữ của hệ thống hiện tại không đủ để phục vụ nhu cầu tính toán của nhóm nghiên cứu. Thiết bị lưu trữ có độ ổn định và độ tin cậy không cao, dễ dẫn đến mất dữ liệu. Tốc độ đọc ghi vào thiết bị lưu trữ thấp, dẫn đến thời gian đọc/ghi và phân tích dữ liệu cao. Hệ thống lưu trữ cần phải được bổ sung về dung lượng, nâng cấp về phần cứng để tăng độ tin cậy và giảm thời gian truy cập dữ liệu.

Năng lực tính toán, nếu chỉ nhìn vào các con số là đủ để phục vụ nhu cầu tính toán của nhóm nghiên cứu. Tuy nhiên, các con số tính toán trên chưa tính đến sự ổn định của các yếu tố liên quan như điện lưới, hệ thống làm mát hay chính sự ổn định của thiết bị lưu trữ và tính toán. Hiệu suất của hệ thống không cao dẫn đến thời gian thực sự sử dụng thiết bị thấp, gây lãng phí tài nguyên. Các mô hình tính toán có độ ưu tiên khác nhau nhưng chưa được sắp xếp hợp lý. Các mô hình dự báo khí hậu khi chạy có thể chiếm dụng tài nguyên hệ thống trong khoảng thời gian dài, dẫn tới các mô hình dự báo thời tiết (có độ ưu tiên cao hơn) không được chạy đúng thời điểm. Chính vì vậy, hệ thống tính toán cần được bổ sung thêm node tính, áp dụng các phương pháp tăng hiệu suất sử dụng và có cơ chế phân bổ tài nguyên thích hợp. Chúng tôi sẽ lần lượt trình bày giải pháp cho các vấn đề trên ở phần tiếp theo.

3.1.Nhu cầu tối ưu hoá hệ thống tính toán

Để tăng cường năng lực tính toán thì cách đơn giản nhất là nâng số lượng node tính. Tuy nhiên, do quỹ đầu tư cho phần cứng của đề tài có hạn, chúng tôi cần tìm các phương án tối ưu hoá cho hệ thống để có thể sử dụng tối đa khả năng của hệ thống hiện tại trong khi hạn chế trang bị thêm phần cứng. Khi hiệu suất sử dụng của hệ thống tăng thì năng lực tính toán thực tế của hệ thống cũng sẽ tăng, mặc dù năng lực tính toán lý thuyết có thể không thay đổi.

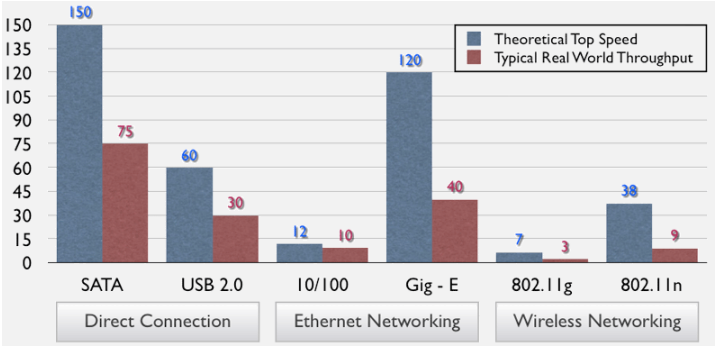
a) Mạng kết nối - điểm yếu của hệ thống cluster

Hệ thống sau khi được bổ sung 2 server lưu trữ LustreFS có tất cả 106 nhân vi xử lý (processor core) và 144 GB RAM. Do các node lưu trữ kể cả MDS và OSS đều có

thể sử dụng làm node tính toán cho các tác vụ đặc biệt, tổng năng lực tính toán (theo lý thuyết) của toàn bộ cluster sẽ là:

Rpeak = (11 x 8 x 2.5 + 2 x 8 x 2.4 + 2 x 3.0) x 4 = 1057.6 GFlops (~ 1 TFlops)

Năng lực tính toán của hệ thống tính toán hiệu năng cao không chỉ phụ thuộc vào tổng năng lực tính toán của các CPU (R_{peak}) như cách tính ở trên. Với các bài toán thuộc dạng embarrasingly parallel, thời gian tính toán trên CPU lớn hơn rất nhiều so với thời gian trao đổi thông tin giữa các CPU, tổng khối lượng tính toán trong 1 giây trên tất cả các CPU có thể tiệm cận với R_{peak}. Với các bài toán có yêu cầu lớn về truyền thông liên node, liên CPU, hiệu năng tính toán thực sự của hệ thống phụ thuộc rất lớn vào hiệu quả của mạng kết nối giữa các node. Trong đo đạc hiệu năng tính toán của một cluster, chỉ số có ý nghĩa quan trọng hơn R_{peak} chính là tổng số phép tính dấu phẩy động tối đa mà cluster có thể thực thi được trong 1 giây - R_{max}. Chỉ số này thường được đo bằng một gói phần mềm đánh giá được thiết kế đặc biệt cho các hệ thống tính toán phân tán - HPL [1]. Tỉ số giữa R_{max} và R_{peak} cho biết hiệu suất sử dụng của hệ thống. Với các hệ thống trong danh sách TOP500, chỉ số này thay đổi trong khoảng từ 45% cho tới hơn 90%. Các hệ thống có hiệu suất thấp đồng nghĩa với khả năng tăng quy mô tính toán thấp, càng tăng số lượng node tính không đồng nghĩa với giảm thời gian tính. Thực tế cho thấy rằng các hệ thống kết nối bởi mạng Ethernet 1 Gbps chỉ có hiệu suất trung bình đạt khoảng xấp xỉ 50%. Khi băng thông kết nối giữa các node tính toán tăng, thời gian trao đổi thông tin giữa các node sẽ giảm, cũng có nghĩa hiệu suất sử dụng CPU tăng và làm cho hiệu suất của toàn bộ hệ thống tăng. Việc tăng hiệu suất tính toán lên sẽ đồng nghĩa với việc giảm số lượng tài nguyên tham gia vào tính toán mà vẫn đảm bảo được hiệu quả công việc. Điều này cũng có nghĩa giảm yêu cầu về đầu tư phần cứng và ngân sách cho việc vận hành và bảo trì hệ thống.



Hình 1. So sánh tốc độ truy cập lý thuyết và thực tế

Với các bài toán khí tượng và khí hậu, ngoài yêu cầu về băng thông kết nối giữa các node tính toán thì tốc độ đọc/ghi dữ liệu từ hệ thống lưu trữ cũng đóng một phần vô cùng quan trọng. Tốc độ đọc/ghi thấp đồng nghĩa với việc thời gian nhàn rỗi của CPU cao và cũng có nghĩa hiệu suất sử dụng hệ thống giảm, nhất là khi khối dữ liệu đọc ghi đạt tới hàng GB. [Hình 1] tổng hợp một số thí nghiệm so sánh tốc độ truy cập dữ liệu giữa lý thuyết và thực tế được thực hiện bởi trang AppleInsider[1][3]. Từ đồ thị, ta có thể thấy rằng thông lượng trung bình thực tế ghi nhận được trên mạng gigabit chỉ đạt được 40 MB/s, tốc độ này nhỏ hơn khá nhiều so với băng thông trung bình của ổ

[1]Thí nghiệm này thực hiện trên thiết bị Time Capsule của Apple nhưng kết quả này có thể coi tương đương với các thiết bị phần cứng thông dụng khác

cứng SATA - 75 MB/s. Điều này cho thấy nếu các node tính toán đọc ghi vào một ổ lưu trữ trên đĩa cứng SATA chia sẻ qua NFS, thông lượng đọc ghi thực tế sẽ không sử dụng hết khả năng của ổ lưu trữ. Hệ thống Metocean sử dụng một số ổ RAID cả cứng và mềm để làm nhiệm vụ lưu trữ chung, thử tốc độ truy xuất dữ liệu trên một ổ software RAID 5 với 4 ổ SATA 2 TB bằng công cụ hdparm cho kết quả như [Hình 2].

```
[root@cluster ~]# hdparm -tT /dev/md_d4p1

/dev/md_d4p1:
 Timing cached reads:   19576 MB in  2.00 seconds = 9808.30 MB/sec
 Timing buffered disk reads:  414 MB in  3.00 seconds = 137.89 MB/sec
```

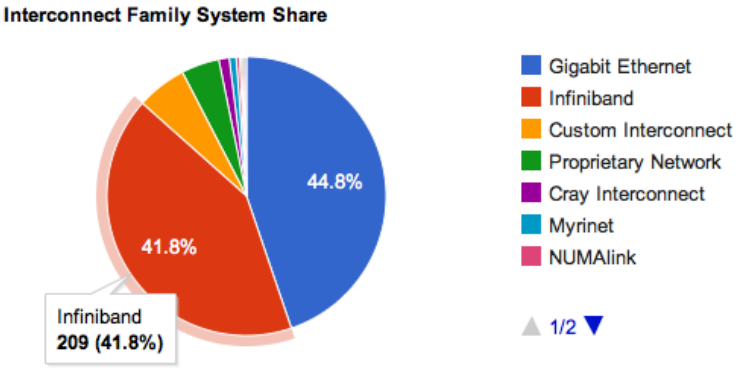
Hình 2. Thử nghiệm tốc độ truy xuất dữ liệu trên 1 Software RAID

Kết quả thử nghiệm trên cho thấy, tốc độ đọc dữ liệu qua cache (do hệ điều hành điều khiển) đạt tới 9.8 GB/s, tốc độ đọc dữ liệu trực tiếp từ ổ lưu trữ cũng lên tới gần 138 MB/s. Các tốc độ này vượt xa lưu lượng mạng trung bình trên mạng gigabit và có thể khẳng định rằng mạng kết nối chính là điểm yếu gây "nghẽn cổ chai" cho hệ thống Metocean.

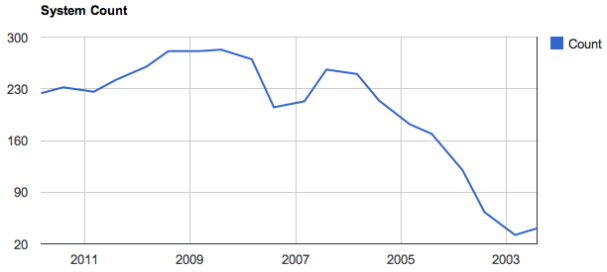
b) Giải pháp thay thế mạng 1 gigabit trong cluster

Khi mới phát triển, các cluster thường được trang bị mạng kết nối Ethernet với tốc độ 100Mbps. Tới quãng những năm 2004, khi thiết bị mạng 1 gigabit trở nên thông dụng và giá thành ngày càng hạ thì số lượng các cluster nằm trong TOP500 của thế giới sử dụng mạng GEthernet tăng nhanh chóng [Hình 4]. Tuy nhiên, cùng với sự phát triển như vũ bão của khoa học, nhu cầu tính toán ngày càng tăng cả về khối lượng và tốc độ thì mạng GEthernet với băng thông hạn chế và độ trễ cao ngày càng bộc lộ nhiều nhược điểm. Có nhiều giải pháp đã được đưa ra như nâng băng thông của mạng Ethernet lên 10Gbps, thiết kế các mạng kết nối chuyên dụng như trong các máy MPP hay tạo ra các chuẩn kết nối mới. Trong số các chuẩn mới này, Infiniband được ưa chuộng hơn cả nhờ băng thông cao (cỡ hàng chục Gbps), độ trễ nhỏ (cỡ vài trăm nano giây, so với vài trăm micro giây của Ethernet), dễ triển khai và đặc biệt là hỗ trợ RDMA[2]. Với chi phí sản xuất ngày càng giảm, Infiniband ngày càng góp mặt nhiều hơn trên các hệ thống trong TOP500 [Hình 5]. Theo danh sách được cập nhật mới nhất (tháng 11 năm 2011), số lượng các hệ thống tính toán trong 500 hệ thống mạnh nhất sử dụng Infiniband chiếm tới 41.8% [Hình 3] và xu hướng sẽ tăng tiếp trong những năm tới.

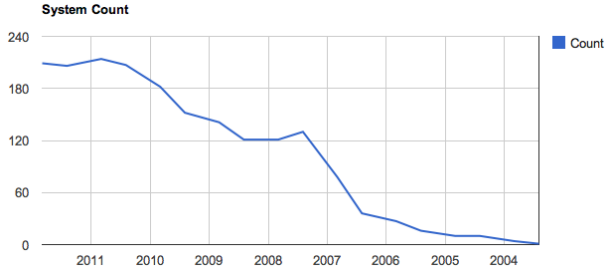
[2]RDMA - Remote Direct Memory Access (Truy cập bộ nhớ trực tiếp từ xa) là phương thức truy cập dữ liệu trực tiếp từ CPU của một máy đến bộ nhớ của máy tính khác. Khi một chương trình thực hiện yêu cầu đọc/ghi RDMA, dữ liệu sẽ được truyền trực tiếp qua mạng mà không cần tới sự điều khiển của Hệ điều hành hay CPU nên độ trễ giảm, tốc độ truyền tin tăng cao. Phương thức này đặc biệt hữu ích đối với các hệ thống tính toán hiệu năng cao



Hình 3. Tương quan các mạng kết nối cluster trong TOP500



Hình 4. Xu hướng giảm dần các cluster sử dụng mạng gigabit

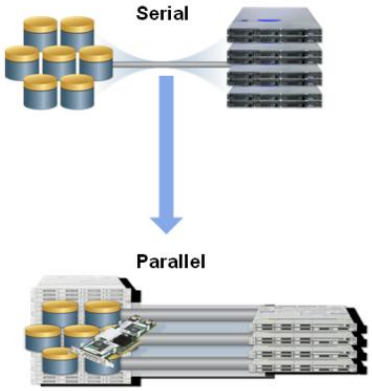


Hình 5. Xu hướng tăng nhanh các cluster sử dụng mạng Infiniband

c) Tối ưu hoá mạng kết nối cho hệ thống METOCEAN Cluster

Hệ thống Metocean là một cluster cỡ nhỏ với tổng số node hiện có là 14 node (sau khi bổ sung server cho hệ thống lưu trữ LustreFS). Các node này vừa đóng vai trò node tính toán, vừa đóng vai trò node lưu trữ. Có hai không gian lưu trữ được xây dựng trên hệ thống: không gian thứ nhất là LustreFS cho băng thông đọc ghi cao, độ trễ thấp, được dùng làm nơi chứa dữ liệu vào ra cho các chương trình tính toán song song. Không gian lưu trữ thứ hai là HDFS với đặc điểm độ trễ cao, tốc độ ghi thấp, được sử dụng làm nơi cất trữ số liệu lâu dài và sao lưu. Các dữ liệu được lưu chuyển trên mạng kết nối thuộc 3 kiểu chính: thông tin quản trị hệ thống, dữ liệu vào ra hệ thống lưu trữ và các thông điệp giữa các tiến trình song song.

Trong các dòng dữ liệu lưu chuyển trên mạng kết nối thì dữ liệu chứa thông điệp trao đổi giữa các tiến trình song song và dữ liệu vào ra hệ thống LustreFS đòi hỏi phải được truyền đi với tốc độ cao. Thông tin quản trị cluster và dữ liệu vào ra hệ thống lưu trữ đám mây HDFS chỉ cần được truyền trên mạng kết nối có tốc độ vừa phải. Để tránh tình trạng nghẽn mạng, hai dòng dữ liệu này được tách ra để truyền trên hai mạng kết nối khác nhau.



Hình 6. Ethernet vs Infiniband

Thông tin quản trị cluster và dòng dữ liệu vào ra hệ thống lưu trữ HDFS được cấu hình để truyền di trên mạng 1 Gbps. Kết nối để phục vụ truyền thông điệp giữa các tiến trình song song và vào ra hệ thống LustreFS được đặt trên mạng Infiniband. Infiniband có nhiều chuẩn với băng thông rất khác nhau như SDR, DDR, QDR, FDR, ... Thiết bị hỗ trợ chuẩn SDR, DDR và QDR đã được sản xuất và sử dụng đại trà trong thực tế. Bộ chuyển mạch và bộ giao tiếp FDR cũng đã được giới thiệu mới đây, hỗ trợ băng thông lên tới 56 Gbps. Việc chọn chuẩn Infiniband để sử dụng cho hệ thống không chỉ phụ thuộc vào tốc độ của thiết bị Infiniband mà còn phải phụ thuộc vào sự tương thích giữa các thiết bị này với phần cứng máy tính sẵn có trong hệ thống.

4. Kết luận

Trong báo cáo này, chúng tôi trình bày GIS và tổng quan nhu cầu sử dụng trong hệ thống. Nhu cầu tối ưu hoá hệ thống tính toán, bao gồm tối ưu hoá mạng kết nối và tối ưu hoá hệ thống lưu trữ cũng được xem xét tới, cùng với những gợi mở về giải pháp xử lý. Các nhu cầu sử dụng là kết quả phân tích khảo sát người dùng tiềm năng của hệ thống nhằm tạo cơ sở cho bước thiết kế và triển khai hệ thống sau này.

5. Tài liệu tham khảo

[1]. A portable Implementation of the High performance Linpack Benchmark for Distributed-Memory Computers, http://www.netlib.org/benchmark/hpl

[2]. Danh sách TOP500 siêu máy tính mạnh nhất thế giới, 11-2011, http://www.top500.org/lists/2011/11

[3]. Hadoop Distributed File System, http://hadoop.apache.org/hdfs

[4]. Exploring Time Capsule: Theoretical speed vs Practical throughput, AppleInsider, http://www.appleinsider.com/articles/08/03/28/exploring_time_capsule_theoretical_speed_vs_practical_throughput.html

[5]. Intel Server System SR2520SAXR, http://ark.intel.com/products/48648/Intel-Server-System-SR2520SAXR

[6]. Server Board s5000vsa, http://www.intel.com/cd/channel/reseller/apac/eng/products/server/boards/dp/s5000vsa/feature/index.htm

[7]. Berkeley Lab Checkpoint/Restart, https://ftg.lbl.gov/projects/CheckpointRestart