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*: Thiết kế hệ thống CPIS dựa trên GIS*

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

Thiết kế hệ thống CPIS dựa trên GIS

Họ và tên chuyên gia: Phan Văn Trọng

1. Mở đầu

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" thì hệ thống thông tin là sản phẩm trực quan và gần gũi với người sử dụng nhất. Hệ thống này cần được thiết kế để lưu trữ một lượng dữ liệu lớn với nhiều định dạng khác nhau, được thu thập bởi các nhóm nghiên cứu trong suốt quá trình thực hiện cũng như vận hành.Đi cùng với việc lưu trữ, hệ thống cũng cần cung cấp các cơ chế cho các nhà khoa học truy cập vào khối dữ liệu để xử lý, phân tích, trích suất thông tin, tổng hợp lại thành cơ sở dữ liệu tri thức. Và cuối cùng, chức năng quan trọng nhất của hệ thống là truyền tải thông tin, lồng ghép kiến thức bản địa nhằm giúp người dân ứng phó với thuỷ tai do biến đổi khí hậu gây ra nhằm giảm thiểu những tổn thất về tính mạng và tài sản. Trong báo cáo này, chúng tôi sẽ đưa ra thiết kế kiến trúc cơ sởcho hệ thống CPIS.

2. Yêu cầu về hệ thống CPIS

CPIS phục vụ 3 nhóm đối tượng chính: các nhà khoa học, những người làm chính sách và cộng đồng. Hệ thống cần đáp ứng được ba chức năng:

(i) lưu trữ các loại dữ liệu khác nhau;
(ii) giúp các nhà khoa học truy cập, phân tích dữ liệu; và
(iii) truyền tải thông tin đến cộng đồng người sử dụng.

Dữ liệu đầu vào có thể có các loại định dạng khác nhau, bao gồm số liệu định lượng và dữ liệu "kiến thức bản địa" nhận được từ cộng đồng. Dữ liệu có thể ở dạng ảnh, phim, bản vẽ, ảnh vệ tinh, văn bản, sản phẩm dự báo các loại. Lượng dữ liệu rất lớn này sau đó sẽ được số hóa, chuẩn hóa, không gian hóa và tích hợp vào cơ sở tri thức dựa trên GIS.

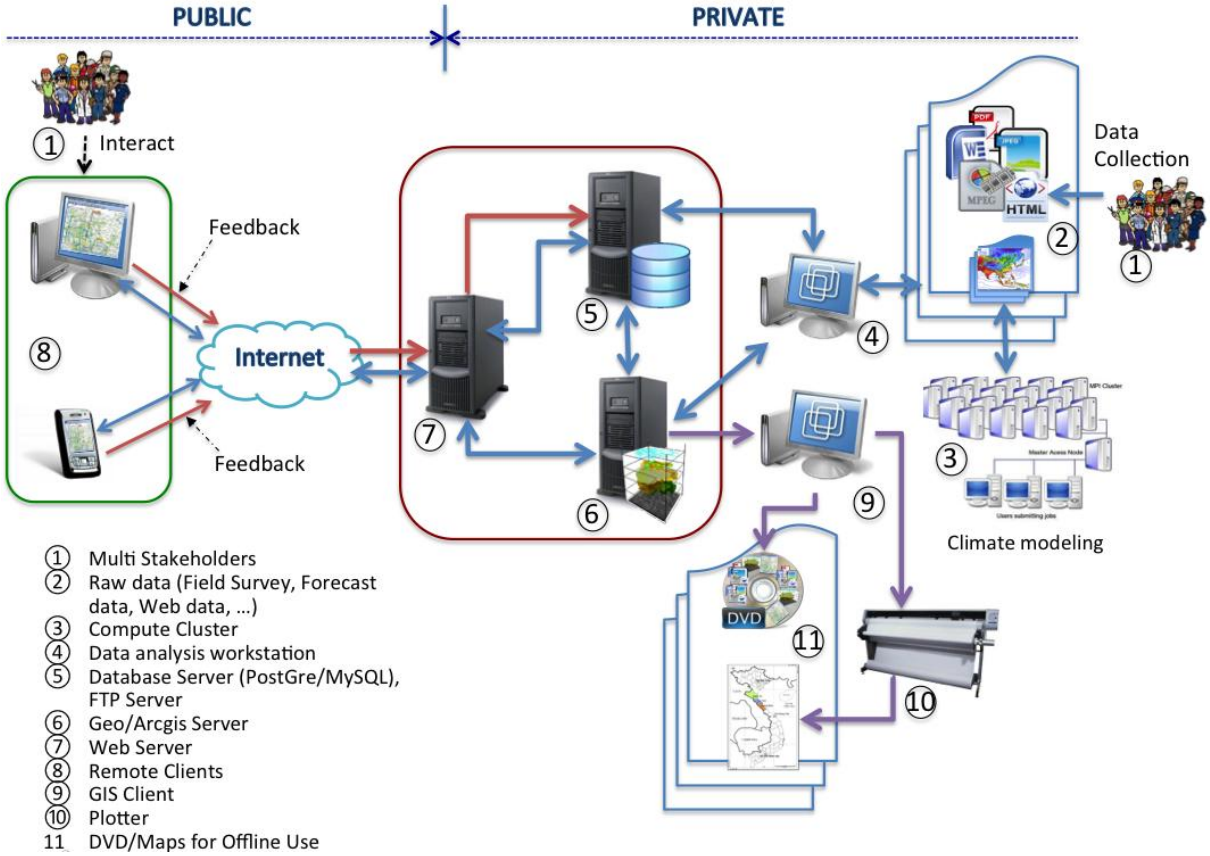
Các nhà khoa học có thể sử dụng các ứng dụng qua mạng nội bộ để truy cập đến hệ thống thông tin, phân tích dữ liệu và trích rút các thông tin cần thiết. Hệ thống cũng cần cho phép người dùng tìm kiếm và tạo tài liệu lưu trữ trên đĩa quang (CD, DVD), bản đồ và sách để gửi cho cộng đồng bị hạn chế trong việc truy cập Internet, nhất là khi có thiên tai xuất hiện.

Đối với đông đảo cộng đồng người sử dụng, hệ thống cho phép truy cập cơ sở tri thức bằng trình duyệt web (trên PC hoặc điện thoại di động) qua Internet. Website cần được thiết kế để có thể truy cập được theo các mức ưu tiên khác nhau cho các nhà khoa học, các nhà quản lý và cộng đồng địa phương. Giao diện của website phải được xây dựng sao cho thân thiện, mềm dẻo với người sử dụng để họ có thể truy cập và phản hồi thông tin của họ. Thông qua quá trình tinh chỉnh liên tiếp (các phân tích số liệu từ các nhà khoa học, thông tin thẩm định và phản hồi từ người sử dụng) thông tin trong hệ thống sẽ được chứng thực và ngày càng được làm giàu thêm.

Việc lôi cuốn cộng đồng và chính quyền địa phường vào việc xây dựng PIS dựa trên GIS là rất quan trọng để: 1) chia sẻ kiến thức bản địa về hệ thống sản xuất nông nghiệp và thủy sản; 2) biết được đầy đủ phân bố không gian về sự BĐKH và thủy tai cũng như ảnh hưởng của chúng thông qua các bản đồ chuyên đề (bản đồ cộng đồng); 3) khai thác, cập nhật dữ liệu nhờ tương tác website qua Internet.

3. Thiết kế hệ thống CPIS dựa trên GIS

Dựa trên thông tin thu thập về nhu cầu của 3 nhóm hưởng lợi có được qua các bản khảo sát cùng với yêu cầu chung về thệ thống CPIS, chúng tôi đã lên một thiết kế cơ sở cho hệ thống CPIS dựa trên GIS [Hình **1**].



Hình 1. Kiến trúc hệ thống CPIS

Hệ thống có thể được chia thành 2 cụm PUBLIC và PRIVATE.Cụm PRIVATE bao gồm cluster (để chạy các mô hình khí tượng, thuỷ văn), các server dịch vụ (Database, Geo, Web, FTP), các máy trạm xử lý dữ liệu và mạng kết nối.Cụm PUBLIC bao gồm thành phần cơ bản là Website để người dùng tương tác.Các thành phần trong cụm PRIVATE chỉ có thể được truy cập bởi các thành viên thực hiện dự án (các nhà khoa học).Cụm PUBLIC (website) được sử dụng bởi rộng rãi người dùng (cộng đồng, nhà quản lý, nhà khoa học).

Cụm PRIVATE bao gồm các thiết bị và dịch vụ đáp ứng yêu cầu (i) và (ii) của CPIS, tức lưu trữ dữ liệu và cung cấp cơ chế truy cập để xử lý và phân tích. Cụm PUBLIC chỉ gồm các dịch vụ hướng tới người dùng cuối như máy chủ web và website.

3.1.Thiết kế hệ thống lưu trữ

Để lưu trữ được lượng dữ liệu lớn, đặc biệt là dữ liệu dự báo khí hậu, thuỷ văn cho thời gian dài, hệ thống lưu trữ cần đạt dung lượng lên tới hàng chục TB. Hệ thống cũng cần đảm bảo băng thông đọc ghi lớn để giảm thiểu thời gian truy xuất, xử lý và trích tách dữ liệu. Chúng tôi đã thử nghiệm một vài giải pháp lưu trữ để thay thế NFS như GlusterFS [1], GPFS [2] và Lustre [3]. Trong 3 hệ thống tập tin trên, GlusterFS được phát triển sau cùng, tuy có ưu điểm dễ triển khai và sử dụng nhưng do thời gian phát triển chưa lâu, hệ thống vẫn chứa các lỗi có thể dẫn đến mất toàn vẹn dữ liệu. GPFS là một hệ thống hoạt động rất ổn định. Tuy nhiên, do GPFS là phần mềm thương mại, thường được bán kèm với các hệ thống lưu trữ đắt tiền của IBM nên chúng tôi quyết định không sử dụng. Thử nghiệm cuối cùng với Lustre mang lại một kết quả rất khả quan. Lustre cho phép xây dựng những hệ thống lưu trữ cực lớn với băng thông đọc/ghi khó có thể đánh bại; và tuy là phần mềm mã mở, Lustre đã được triển khai và kiểm nghiệm tại nhiều hệ thống tính toán vào hàng lớn nhất thế giới.

Để phục vụ cho dự án này, chúng tôi lựa chọn phần mềm mã nguồn ở Lustre để xây dựng hệ thống lưu trữ đạt dung lượng lớn tới hàng chục TB với băng thông đọc ghi cao tới hàng trăm MB/s.

a) Giới thiệu

Lustre là một kiến trúc lưu trữ được thiết kế cho các bó máy tính (cluster system). Thành phần trung tâm của kiến trúc này là hệ thống tập tin phân tán LustreFS được xây dựng trên nền Linux và cung cấp giao diện hệ thống tập tin UNIX tương thích chuẩn POSIX. Riêng cái tên Lustre cũng đã nói lên mục đích thiết kế và mối liên hệ chặt chẽ với hệ điều hành Linux - Lustre =>Linux + cluster.

Kiến trúc Lustre được sử dụng trong rất nhiều loại bó máy tính cho các ứng dụng khác nhau. Nhưng Lustre được biết đến nhiều nhất với vai trò là hạt nhân các hệ thống lưu trữ trên 7 trong số 10 bó máy tính hiệu năng cao (HPC) lớn nhất thế giới bao gồm cả quán quân TOP500 siêu máy tính - K-Computer [5]. Lustre hỗ trợ các bó máy tính với quy mô khác nhau, từ các hệ thống trong các nhóm làm việc quy mô nhỏ cho tới các hệ thống siêu lớn cỡ hàng chục ngàn node tính toán, hàng chục PB lưu trữ và lưu lượng vào ra (I/O) lên tới hàng trăm GB/s. Rất nhiều trung tâm HPC sử dụng Lustre làm hệ thống tập tin chung cho hàng tá bó máy tính với quy mô chưa từng có.

Khả năng mở rộng của Lustre cho phép giảm thiểu nhu cầu triển khai nhiều hệ thống tập tin cùng lúc (ví dụ mỗi hệ thống cho 1 cluster) khiến cho việc quản trị hệ thống lưu trữ cũng tập trung và đơn giản hơn. Với việc ghép nối nhiều máy chủ cùng làm nhiệm vụ lưu trữ, không chỉ khả năng lưu trữ của cả hệ thống sẽ được cộng gộp mà lưu lượng vào ra cũng tăng tỉ lệ thuận với số lượng máy chủ. Thêm nữa, dung lượng lưu trữ hoặc lưu lượng vào ra của hệ thống hoàn toàn có thể chủ động thay đổi bằng cách ghép thêm máy chủ lưu trữ.

b) Các đặc điểm chính

Các đặc điểm chính của Lustre bao gồm:

- **Khả năng mở rộng quy mô:** Lustre có khả năng mở rộng dung lượng lưu trữ cũng như băng thông đọc ghi tương ứng với số lượng các máy khách (client). Hiện tại, các hệ thống được triển khai trong thực tế có số lượng client lên tới

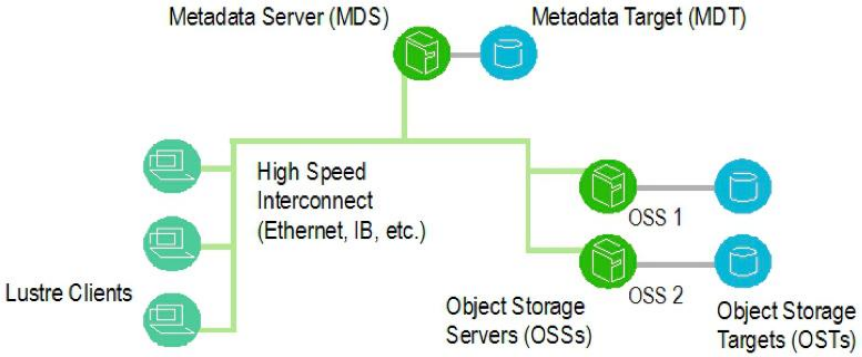
26.000 node và rất nhiều hệ thống có lượng client trong khoảng từ 10.000 - 20.000 node. Một vài hệ thống có khả năng lưu trữ trên 1PB hay tương đương với 2 tỷ tập tin, đã được đưa vào sử dụng từ năm 2006.

- **Hiệu năng:** Các hệ thống Lustre trong môi trường vận hành thực tế hiện tại đạt băng thông lên tới 100GB/s. Băng thông từng ghi nhận được trên 1 node máy khách đạt 2GB/s (tối đa) và băng thông trên 1 máy chủ lưu trữ đạt 2.5GB/s (tối đa). Tổng băng thông trên toàn hệ thống tập tin Spider đặt tại phòng nghiên cứu quốc gia Oak Ridge ở Mỹ từng đạt tới 240GB/s.
- **Tương thích POSIX:** Lustre chạy trên các máy khách thoả mãn mọi yêu cầu của chuẩn POSIX. Trong một cluster, tương thích POSIX có nghĩa là hầu hết các thao tác trên dữ liệu là atomic[1] (không thể phân chia) và client sẽ không bao giờ nhìn thấy dữ liệu hoặc siêu dữ liệu (metadata) bất hợp lệ.
- **Mã nguồn mở:** Mã nguồn Lustre được cấp phép với giấy phép GNU GPL.

c) *Các thành phần của hệ thống Lustre*

Hệ thống tập tin Lustre bao gồm các thành phần cơ bản sau [Hình **2**]:

- **Metadata Server (MDS)** - Máy chủ lưu trữ siêu dữ liệu: MDS server cho phép các Lustre client truy cập được siêu dữ liệu lưu trữ trên một hoặc nhiều MDT. Mỗi MDS quản lý tên tập tin và thư mục trong một (hoặc nhiều) hệ thống tập tin Lustre và xử lý các yêu cầu truy vấn mạng cho một hoặc nhiều MDT cục bộ.



Hình 2. Các thành phần trong một hệ thống Lustre cơ bản

- **Metadata Target** - Đích siêu dữ liệu (MDT):lưu trữ siêu dữ liệu (như tên tập tin, thư mục, quyền truy cập và phân bố tập tin) trên một MDS. Mỗi hệ thống tập tin có một MDT. Một MDT lưu trữ trên phân vùng lưu trữ chia sẻ có thể được truy cập bởi nhiều MDS nhưng chỉ một MDS được dùng MDT đó tại một thời điểm. Nếu MDS chủ động (active MDS) bị hỏng, MDS thụ động (passive MDS) có thể nắm quyền truy cập MDT và đáp ứng truy vấn từ các client. Mô hình này được gọi là chuyển đổi dự phòng MDS (MDS failover).

[1]Thao tác không thể phân chia (atomic operation) có thể bao gồm nhiều chỉ lệnh CPU, nhưng hoặc tập các chỉ lệnh này phải được thực thi hoàn toàn hoặc không được thực thi. Trong một atomic operation, bộ xử lý có thể đọc và ghi vào một địa chỉ bộ nhớ cùng một lúc mà không một bộ xử lý hay thiết bị vào ra nào khác được phép đọc/ghi cho tới khi thao tác kết thúc. Chính vì vậy, dữ liệu được ghi vào bộ nhớ/hệ thống lưu trữ bởi các thao tác này luôn hợp lệ và điều này rất quan trọng đối với các hệ thống có nhiều CPU và dùng chung không gian lưu trữ như cluster.

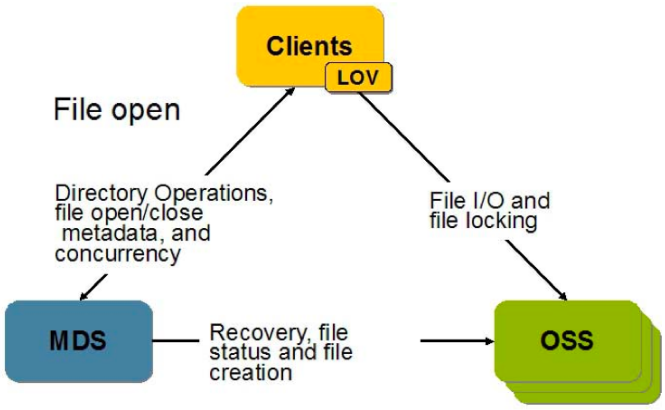
- **Object Storage Server (OSS)** - Máy chủ lưu trữ đối tượng (dữ liệu): OSS cung cấp dịch vụ vào ra tập tin và xử lý truy vấn mạng cho một hoặc vài OST cục bộ. Thông thường một OSS quản lý từ 2 tới 8 OST, với dung lượng một OST lên tới 8TB.
- **Object Storage Target (OST)** - Đích lưu trữ đối tượng (dữ liệu): OST lưu trữ dữ liệu tập tin (hoặc phân đoạn dữ liệu tập tin) dưới dạng các đối tượng dữ liệu trên một hoặc nhiều OSS. Một hệ thống tập tin Lustre riêng lẻ có thể có nhiều OST với mỗi OST lưu trữ một tập con dữ liệu tập tin. Để tăng hiệu năng hệ thống, một tập tin có thể được chia thành các phân đoạn dữ liệu, trải ra trên nhiều OST khác nhau và được quản lý bởi một Logical Object Volume (LOV) – khối đối tượng logic.
- **Lustre Networking (LNET)** – Liên kết mạng Lustre: là một giao diện lập trình ứng dụng (API) quản lý việc vào ra dữ liệu và siêu dữ liệu trên các máy chủ và máy khách. LNET hỗ trợ một vài mạng kết nối thông thường cũng như cao cấp bao gồm Infiniband, TCP/IP, Quadrics Elan, Myrinet và Cray.

  Tính năng chính của LNET bao gồm:
  - RDMA [2] nếu được hỗ trợ bởi mạng kết nối như Elan, Myrinet và Infiniband
  - Hỗ trợ các kiểu kết nối mạng thường dùng như Infiniband và IP
  - Sử dụng máy chủ chuyển đổi dự phòng cho phép tăng khả năng đáp ứng và cung ứng tính năng phục hồi trong suốt
  - Có thể sử dụng hỗn tạp các kiểu kết nối mạng khác nhau với khả năng định tuyến giữa các mạng
- **Lustre clients** –Các máy khách:là các node tính toán, node hiển thị hoặc máy tính để bàn chạy phần mềm Lustre cho phép mount hệ thống tập tin Lustre vào cây thư mục của mình. Không gian tên miền được đồng bộ và đồng nhất trên tất cả các máy khách tại mọi thời điểm. Các máy khách khác nhau có thể cùng lúc ghi vào các phần khác nhau trong cùng một tập tin trong khi các máy khác đọc dữ liệu từ đó ra.

  MDT, OST và Lustre client có thể được cài đặt và chạy trên cùng một node. Tuy nhiên, cấu hình thông thường là MDT chạy trên một node riêng biệt, hai hoặc nhiều hơn OST trên mỗi node OSS và một client trên mỗi node tính toán.
- **Management Server (MGS)** – Máy chủ quản trị: lưu trữ thông tin cấu hình cho tất cả các hệ thống tập tin Lustre trong một bó máy tính. Mỗi đích lưu trữ trong Lustre (MDT, ODT) liên lạc với MGS để cung cấp thông tin, ngược lại, các máy khách Lustre liên lạc với MGS để truy hồi thông tin. Thông thường MGS sử dụng đĩa cục bộ riêng để lưu trữ, tuy nhiên nếu người dùng chỉ định triển khai một hệ thống tập tin Lustre duy nhất, MGS có thể dùng chung ổ cứng với MDT. MGS không được coi là một phần của một hệ thống tập tin

[2]RDMA – Remote Direct Memory Access (truy xuất dữ liệu trực tiếp từ xa) là công nghệ cho phép truy xuất dữ liệu trực tiếp từ bộ nhớ của một máy tính tới bộ nhớ của máy tính khác mà không cần tác động của hệ điều hành hay CPU của cả hai máy tính. Công nghệ này đặc biệt có ích cho các hệ thống hiệu năng cao vì nó cho phép tăng lưu lượng đồng thời giảm độ trễ của mạng kết nối.

riêng biệt, nó cung cấp thông tin cấu hình tất cả các hệ thống tập tin Lustre cho các thành phần khác của Lustre.

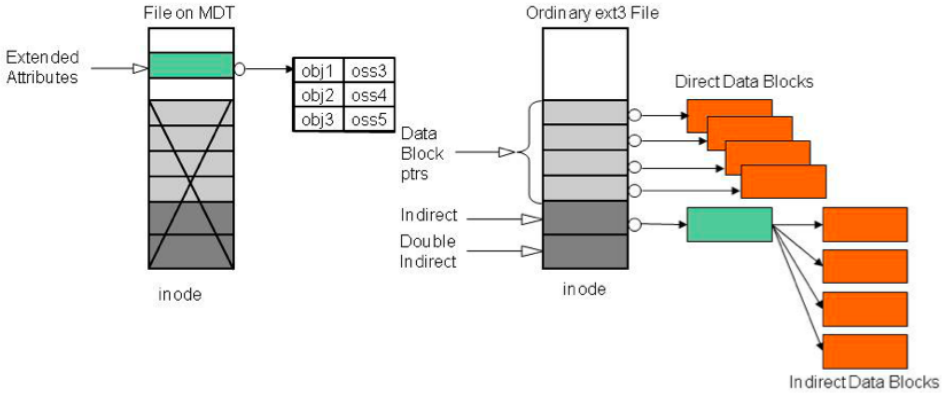


Hình 3. Tương tác giữa các thành phần trong hệ thống tập tin Lustre

Phần mềm Lustre chạy trên các máy khách cung cấp một giao diện cho phép hệ thống tập tin Linux ảo (Linux Virtual File System) trên các máy khách tương tác với các máy chủ Lustre. Mỗi một thành phần phần mềm chạy trên máy chủ (MDT, OST, MGS) tương tác với một thành phần tương ứng chạy trên máy khách: MDC – Metadata Client, OSC – Object Storage Client và MGC – Management Client. Các OSC có quan hệ với nhau sẽ được nhóm lại thành một LOV duy nhất. Các OSC qua LOV sẽ được hiển thị dưới dạng một thiết bị đơn nhất và giúp cho việc truy xuất dữ liệu của các client trở nên trong suốt [Hình **3**].

d) Tập tin trong LustreFS

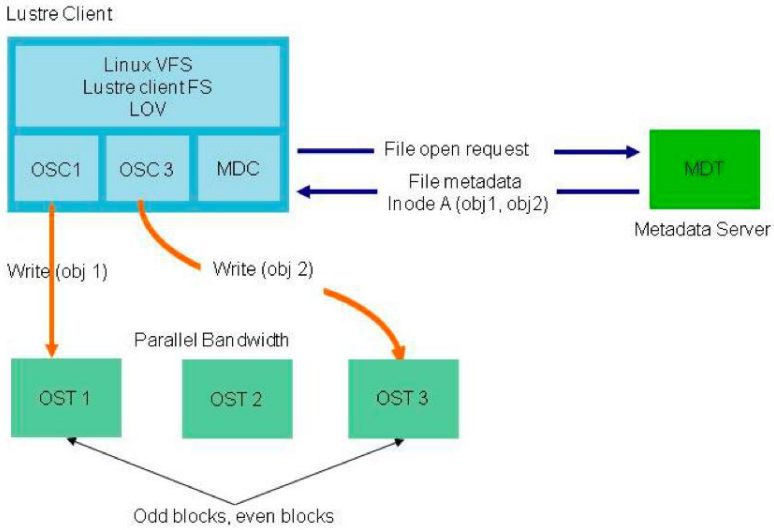
Các hệ thống tập tin UNIX truyền thống sử dụng các inode chứa danh sách đánh số các block (khối) dữ liệu của tập tin. Tương tự, với mỗi tập tin trong hệ thống, LustreFS tạo ra một inode trên MDT. Tuy nhiên, các inode này không trỏ tới bock dữ liệu mà trỏ tới một hoặc nhiều đối tượng dữ liệu tương ứng với các tập tin [Hình **4**]. Các đối tượng này được lưu trữ dưới dạng các tập tin trên OST và chứa dữ liệu (hoặc phân đoạn dữ liệu) của tập tin mà inode trên MDT trỏ tới.



Hình 4. Inode trong MDS trỏ tới các đối tượng, inode trong ext3 trỏ tới dữ liệu

[Hình **5**] minh hoạ thao tác mở một tập tin để ghi dữ liệu trong Lustre. Client thông qua MDC, yêu cầu MDS mở một tập tin. MDS trả lại cho client danh sách địa chỉ các đối tượng trên các OST tương ứng. Client sử dụng các thông tin này, tương tác trực tiếp với các OSS, ghi các phân đoạn dữ liệu vào các đối tượng với sự trợ giúp của

LOV và OSC. Các OSC trên client chạy đồng thời và ghi dữ liệu song song lên các OST.



Hình 5. Thao tác mở và ghi dữ liệu trong Lustre

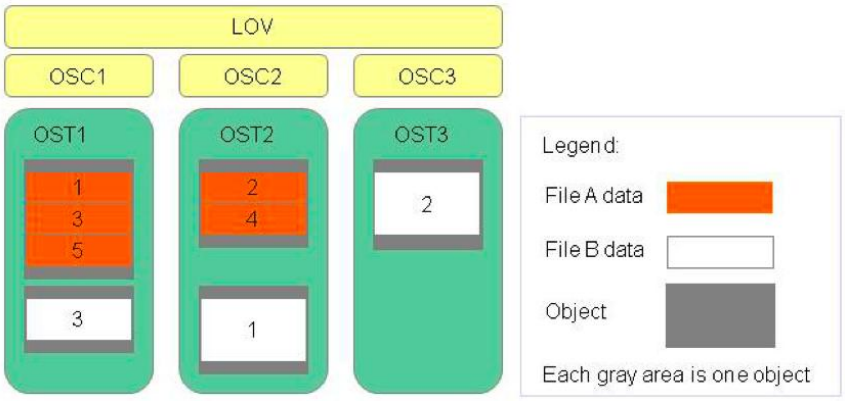
Nếu chỉ có duy nhất một đối tượng dữ liệu gắn với một inode trên MDT, đối tượng đó chứa tất cả dữ liệu của tập tin mà inode trỏ tới. Nếu có nhiều đối tượng tương ứng với một tập tin, dữ liệu tập tin đó đã được trải ra trên các đối tượng.

MDS chứa thông tin phân bố của tất cả các tập tin, vị trí và số lượng các phân đoạn dữ liệu của tập tin trên các OST. Các client nhận "sơ đồ" phân bố các phân đoạn dữ liệu của tập tin từ MDS sau đó liên lạc trực tiếp với các OST tương ứng để thực hiện các thao tác vào ra với các phân đoạn dữ liệu.

Với cách tổ chức dữ liệu như trên, dung lượng lưu trữ của một hệ thống tập tin Lustre bằng với tổng số dung lượng lưu trữ của các OST. Băng thông đọc ghi tổng cộng của hệ thống bằng với tổng băng thông giữa OSS và các đích ghi dữ liệu. Cả khả năng lưu trữ và băng thông vào ra tổng cộng của toàn bộ hệ thống đều có thể tăng lên một cách đơn giản với việc tăng số lượng các OSS.

Phân đoạn tập tin

Việc phân đoạn cho phép các phần dữ liệu của một tập tin được lưu trữ trên các OST khác nhau. Mô hình này mô phỏng công nghệ lưu trữ kiểu RAID 0, trong đó, dữ liệu được chia ra thành nhiều phần lưu trữ trên một số lượng các đối tượng khác nhau, số lượng đối tượng này được gọi là stripe_count. Mỗi đối tượng chứa một hoặc một vài phân đoạn dữ liệu. Khi đoạn dữ liệu được ghi vào một đối tượng vượt quá một ngưỡng nhất định - stripe_size, đoạn dữ liệu kế tiếp của tập tin sẽ được lưu trữ trên đích (OST) tiếp theo. [Hình **6**] minh hoạ phân bố dữ liệu của 2 tập tin A và B với stripe_count khác nhau (A=2, B=3) và strip_size khác nhau.



Hình 6. Phân đoạn tập tin với các độ lớn phân đoạn khác nhau

Việc phân đoạn tập tin có một vài lợi ích. Thứ nhất, dung lượng lớn nhất của tập tin sẽ không bị giới hạn bởi dung lượng của một đích lưu trữ đơn. Lustre có thể tách dữ liệu ra để lưu trữ trên 160 OST với mỗi OST cho phép sử dụng tối đa 8 TB[3] dung lượng cho mỗi tập tin; tức là mỗi tập tin có thể sử dụng hơn 1PB dung lượng đĩa cứng. Lợi ích khác là băng thông vào/ra của một tập tin bằng tổng băng thông vào/ra của các phân đoạn dữ liệu tập tin lưu trong các OST, tức là tối đa có thể tương đương với tổng băng thông vào/ra của 160 máy chủ.

Mặt khác, phân đoạn tập tin cũng có thể tăng nguy cơ mất mát dữ liệu khi xảy ra lỗi hệ thống: ví dụ trường hợp trải nội dung tập tin ra khắp các máy chủ, nếu 1 máy chủ OSS bị sập, một phần nhỏ của tất cả các tập tin trong OSS đó sẽ mất, kéo theo tất cả các tập tin có thể sẽ không truy cập được. Để so sánh, nếu mỗi tập tin chỉ có duy nhất một phân đoạn dữ liệu (tức được lưu trữ trên 1 OSS duy nhất), khi OSS sập, tập tin đó sẽ bị mất hoàn toàn; tuy nhiên, các tập tin lưu trên OSS khác sẽ không bị ảnh hưởng.

e) Yêu cầu về thiết bị lưu trữ

Siêu dữ liệu trên MDS và đối tượng dữ liệu trên các OSS được lưu trữ dưới dạng các tập tin trong các thiết bị lưu trữ. Các thiết bị này được kết nối với các máy chủ MDS và OSS, được phân vùng và định dạng với một hệ thống tập tin nhất định (ví dụ: ext3, ext4).

Lưu trữ trên OSS

Mỗi OSS có thể quản lý nhiều OST, thông thường từ 2-8 OST với dung lượng tối đa cho mỗi OST là 8TB. Dạng truy xuất dữ liệu đặc trưng trên các OSS là truyền một lượng dữ liệu lớn với băng thông cao. Do Lustre lưu trữ dữ liệu trên hệ thống theo mô hình RAID 0, nó không có cơ chế để đảm bảo dữ liệu được toàn vẹn trong trường hợp một hoặc một vài ổ cứng lưu trữ bị hỏng. Chính vì vậy, với hệ thống lưu trữ lớn, để đảm bảo sự toàn vẹn của dữ liệu, các thiết bị được lựa chọn sử dụng cần có tuổi thọ cao, tần suất xuất hiện lỗi nhỏ và các OST bắt buộc phải được đặt trên các phân vùng RAID 5 (hoặc RAID 6 hoặc thậm chí RAID 1).

Lưu trữ trên MDS

Máy chủ MDS chứa siêu dữ liệu của hệ thống Lustre. Theo khuyến nghị từ nhóm phát triển Lustre, dung lượng MDT cần đạt từ 1-2% dung lượng lưu trữ của cả hệ

[3]Với Lustre 1.8.2, 16TB trên RHEL5

thống. Dạng truy xuất dữ liệu trên MDS khác với OSS: yêu cầu thời gian tìm kiếm thấp với rất nhiều lệnh đọc ghi lượng nhỏ dữ liệu. Mô hình lưu trữ dạng RAID 0+1 sẽ là lý tưởng cho MDT.

f) Năng lực lưu trữ của hệ thống Lustre

Năng lực lưu trữ của hệ thống bằng tổng năng lực lưu trữ của các đích dữ liệu (OST).

Ví dụ, một hệ thống có 64 OSS, mỗi OSS quản lý 2 OST dung lượng 8 TB sẽ cho phép lưu trữ tối đa 64x2x8 = 1 PB dữ liệu. Nếu mỗi OSS sử dụng 16 ổ cứng SATA 1 TB với tốc độ đọc/ghi trung bình 50 MB/s, băng thông vào ra tổng cộng trên 16 ổ cứng đạt 800 MB/s. Nếu OSS này tham gia vào mạng kết nối hỗ trợ băng thông tương đương (ví dụ Infiniband), lưu lượng vào/ra trên OSS cũng có thể đạt tới 800 MB/s. Băng thông tổng cộng trên toàn bộ hệ thống có thể lên tới 64x800 = 50 GB/s.

3.2. Thiết kế hệ thống WebGIS

Hệ thống phần mềm WebGIS được lựa chọn dựa trên tiêu chí dễ sử dụng, dễ tuỳ chỉnh, có nhiều ứng dụng thực tế cũng như có một cộng đồng phát triển mạnh mẽ. Trong dự án này, chúng tôi lựa chọn kiến trúc WebGIS dựa trên các thành phần phần mềm mã nguồn mở và hoạt động theo mô hình máy chủ - máy khách [Hình **7**].

Kiến trúc này có 3 lớp:

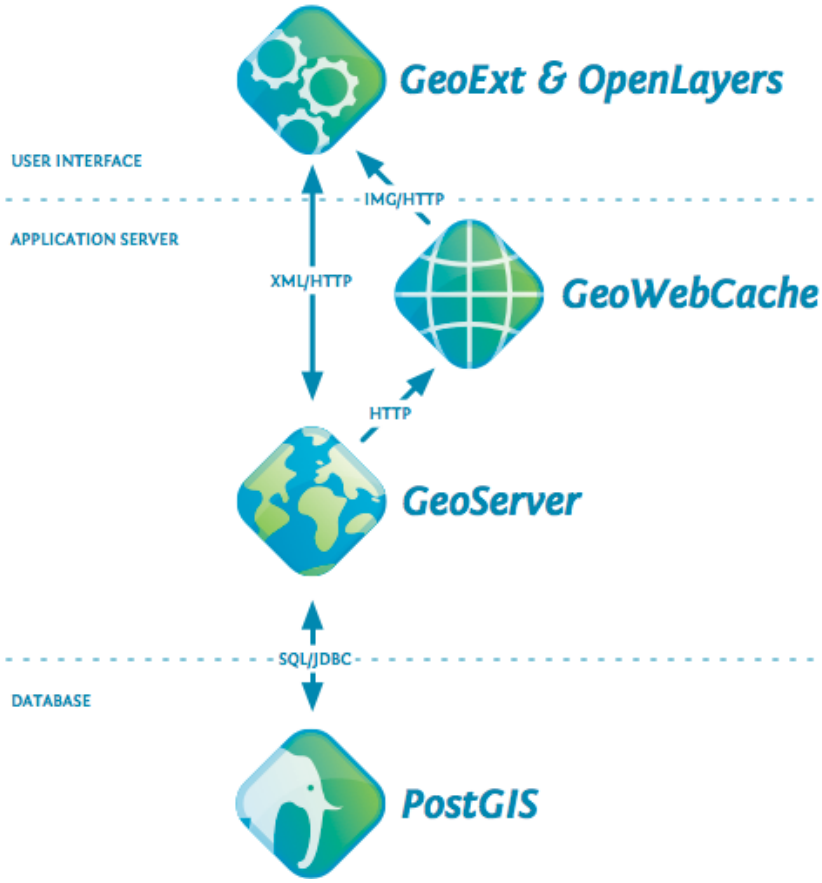
Lớp thứ nhất – Cơ sở dữ liệu: Cơ sở dữ liệu không gian được quản trị bởi hệ quản trị cơ sở dữ liệu không gian PostGreSQL/PostGIS.

Lớp thứ hai – Máy chủ ứng dụng: Máy chủ này chạy 2 dịch vụ là GeoServer và GeoWebCache:

- GeoServerđảm nhận nhiệm vụ tạo lập bản đồ theo yêu cầu gửi tới từ máy khách. GeoServer là một máy chủ mã nguồn mở với mục đích kết nối những thông tin địa lý có sẵn tới các Geoweb (trang Web địa lý) sử dụng chuẩn mở. Được bắt đầu bởi một tổ chức phi lợi nhuận có tên The Open Planning Project (TOPP), nhằm mục đích hỗ trợ việc xử lý thông tin không gian địa lý với chất lượng cao, đơn giản trong sử dụng, là phần mềm mã nguồn mở nhằm cung cấp và chia sẻ dữ liệu. Được kỳ vọng sẽ trở thành một phương thức đơn giản để kết nối những nguồn thông tin có sẵn từ Google Earth, NASA World Wind nhằm tạo ra các dịch vụ Webmap như Google Maps, Windows Live Local và Yahoo Maps. Các cấu hình cho GeoServer cần tương thích chuẩn OGC (Open Geospatial Consortium).

- GeoWebCache tăng tốc quá trình truyền dữ liệu và ảnh bằng cách chia vùng cần truyền thành nhiều phần nhỏ kết hợp với sử dụng bộ đệm để tăng tốc độ truyền.

Lớp trên cùng – Giao diện người dùng: người dùng truy cập hệ thống bằng trình duyệt web. Giao diện này cần được thiết kế để dễ sử dụng và thân thiện với người dùng, đặc biệt với người dân tại vùng chịu ảnh hưởng bởi thuỷ tai – những người có ít kiến thức cũng như kinh nghiệm sử dụng máy tính, mạng internet.



Hình 7. Kiến trúc hệ thống WebGIS

3.3. Thiết kế hệ thống truyền thông

Hệ thống truyền thông bao gồm mạng kết nối và website. Mạng kết nối cần đạt băng thông cao và có tính dự phòng với các thiết bị mạng ổn định, có khả năng hoạt động bền bỉ 24/7. Website về thực chất là một hệ quản trị nội dung (CMS)

Các tính năng hệ thống CMS cần được xây dựng:

- Xây dựng dựa trên các thành phần mã nguồn mở
- Sử dụng Hệ điều hành Linux, ngôn ngữ lập trình PHP, hệ quản trị cơ sở dữ liệu MySQL và máy chủ web Apache phiên bản mới nhất
- Tương thích cho các trình duyệt thông dụng: IE, Firefox, Chrome
- Các tính năng cần có trong phần quản trị nội dung:
    - Phân quyền người dùng: 3 nhóm người dùng sẽ có quyền khác nhau trong việc truy cập vào nguồn số liệu và công bố các thông tin trên website.
    - Trình soạn thảo trên nền web.
    - Quản trị ngôn ngữ (Tiếng Anh/Tiếng Việt)
    - Cho phép tải dữ liệu lên máy chủ hoặc xuống máy tính cá nhân
    - Quản trị tài liệu
    - ...

4. Kết luận

Hệ thống thông tin là kết quả trực quan và gần gũi người sử dụng nhất của 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".

Hệ thống này cần đảm bảo lưu trữ được lượng dữ liệu cực lớn với băng thông vào ra cao, cho phép các nhà khoa học truy cập từ xa để phân tích, trích xuất thông tin cũng như truyền tải tri thức tới người sử dụng. Trong tài liệu này, chúng tôi trình bày kiến trúc cơ sở của giải pháp hệ thống thông tin dựa trên GIS với việc sử dụng Lustre làm hệ thống lưu trữ, GeoServer cùng các thành phần đi kèm làm máy chủ thông tin địa lý và hệ CMS làm phương tiện quản lý và truyền tải thông tin.

5. Tài liệu tham khảo

[1] http://www.gluster.org

[2] http://www-03.ibm.com/systems/software/gpfs

[3] http://www.lustre.org

[4] http://www.top500.org/list/2011/11

[5] Fujitsu Limited, *An overview of Fujitsu's Lustre Based File System*, June 24, 2011

[6] Nathan Rutman, Xyratex, *Rock-hard Lustre: Trends in scalability and Quality*, 2012

[7] Sun Oracle, *Lustre File System - Operations Manual - Version 1.8*, December 2010