

MỤC LỤC

CHƯƠNG I.....	1
GIỚI THIỆU VỀ E-SCIENCE.....	1
1.1 Nguồn gốc của e-Science.....	1
1.2 e-Science là gì?	3
1.3 Dùng thuật ngữ eScience, CyberInfrastructure hay eResearch	4
1.4 Những đặc điểm, tính chất của e-Science.....	5
1.5 Những yêu cầu cấp bách về cơ sở hạ tầng phục vụ các phương pháp nghiên cứu khoa học mới.....	7
1.5.1 Khoa học tập trung vào dữ liệu (data-intensive).....	7
1.5.2 Nghiên cứu khoa học dựa trên việc giả lập và mô phỏng.....	8
1.5.3 Truy xuất từ xa đến các công cụ và dữ liệu	9
1.5.4 Tổ chức (Virtual Organization) hay cộng đồng ảo (Virtual Community)	10
1.5.5 Phát triển một cơ sở hạ tầng mới.....	10
1.6 Cơ sở hạ tầng hiện thực cho eScience	11
CHƯƠNG II.....	13
E-SCIENCE TRÊN THẾ GIỚI.....	13
2.1 e-Science tại UK	13
2.1.1 Các nhóm e-Science tại UK.....	13
2.1.2 Các dự án về e-Science của UK.....	15
2.2 e-Science tại US.....	16
2.2.1 Cyberstructure	16
2.2.2 Các dự án về e-Science của US.....	17
2.3 Chi tiết một số dự án tiêu biểu	18
2.3.1 RealityGrid.....	18
2.3.2 Comb-e-Chem.....	19
2.3.3 Distributed aircraft maintenance environment (DAME)	20
2.3.4 myGrid	21
2.3.5 GridPP	21
2.3.6 AstroGrid.....	21
2.3.7 eDiamond	22
2.4 e-Science Core Programme.....	22
2.4.1 Hiện thực hạ tầng mạng kết nối các e-Science Centres	22
2.4.2 Định hướng cho sự phát triển Grid Middleware	23
2.4.3 Interdisciplinary Research Collaboration (IRC)	24
2.4.4 Hỗ trợ các dự án e-Science.....	24
2.4.5 Sự hợp tác toàn cầu	24
2.4.6 Hạ tầng mạng	24
2.4.7 Demonstrator projects	25
CHƯƠNG III.....	26
VAI TRÒ CỦA GRID COMPUTING ĐỐI VỚI TÍNH TOÁN KHOA HỌC.....	26
3.1 Chia sẻ tài nguyên bên trong tổ chức ảo (Virtual Organization).....	26
3.2 Giao tiếp với những công việc đang được thực thi trên hệ thống lưới	27
3.3 Tính toán phân bố	29
3.4 Quản lý dữ liệu.....	31
Tài liệu tham khảo	33

CHƯƠNG I

GIỚI THIỆU VỀ E-SCIENCE

1.1 Nguồn gốc của e-Science

Thuật ngữ e-Science xuất hiện đầu tiên ở Anh từ năm 2000, qua quá trình trao đổi, giao lưu tại nhiều Hội thảo quốc tế đến nay đã trở thành một khái niệm khá phổ biến được một số nước chính thức sử dụng từ năm 2004, trong đó có nhiều quốc gia khu vực Đông nam Á như Singapore, Thailand Theo GS. Tony Hey, giám đốc dự án, thì e-Science chính là một hạ tầng cơ sở cho các ngành khoa học phát triển (e-Infrastructure), nhờ vào khả năng cung cấp kết nối và các dịch vụ tính toán, truy cập thông tin, ứng dụng ngày càng rộng lớn mà CNTT đang đem lại cho các nhà khoa học. Theo một nghĩa nào đó, hạ tầng e-Science có thể đồng nghĩa với mạng lưới (Grid), đó là tập hợp của các dịch vụ trung gian, vận hành trên nền tảng của hệ thống mạng toàn cầu, băng thông rộng, hiệu năng cao, có khả năng hỗ trợ cộng đồng nghiên cứu và các phát minh, sáng tạo.

Từ khi máy tính xuất hiện lần đầu tiên, con người luôn luôn cố gắng cải tiến, nâng cấp, biến nó thành các cỗ máy ngày càng mạnh mẽ. Đây là mong muốn của mọi người sử dụng máy tính cá nhân, và cũng là mong muốn của các tổ chức với những chiếc máy được tạo ra chỉ để dùng cho nghiên cứu khoa học. Một chiếc máy tính, một mình nó, không thể đảm đương việc thu thập, lưu trữ, và phân tích dữ liệu được thu thập bởi một nghiên cứu khoa học. Do đó, sự phát triển của e-Science là rất quan trọng.

Mối liên hệ của các thông tin khoa học đòi hỏi một lượng lớn dữ liệu, đây là điều chỉ có e-Science mới có khả năng đáp ứng. Khi loài người cố gắng khám phá những tri thức mới, hay hiểu rõ về những gì họ đang nghiên cứu, yêu cầu về công nghệ cũng ngày càng cao.

Máy tính giờ đây không chỉ là một công cụ phục vụ khoa học. Nó đã trở thành một phần của khoa học. Khoa học máy tính hiện không chỉ nói về phần cứng hay phần mềm mà còn về các đại dương, ngôi sao, tế bào ung thư, protein và mạng lưới bạn bè. Ken Birman, Giáo sư khoa học máy tính Đại học Cornell (Mỹ), nói ngành học

của ông đang trên đường trở thành “một ngành khoa học của vũ trụ”, một cơ cấu làm nền tảng cho mọi ngành khác, bao gồm các ngành khoa học xã hội.

Bản chất của vấn đề ông Birman khẳng định là máy tính đã biến đổi từ một công cụ phục vụ khoa học thành một phần của khoa học. Và những diễn tiến gần đây trong giới khoa học phần nào đã cho thấy điều này.

“Các nhà sinh vật học hệ thống” tại trường Y Harvard đã phát triển một “ngôn ngữ máy tính” gọi là “Little b” dùng để lập mô hình các tiến trình sinh học. Ngôn ngữ này biết suy luận về dữ liệu sinh học, học hỏi từ nó và tích hợp những gì đã học vào trong những mô hình mới và những dự báo về hành vi của tế bào. Các tác giả gọi loại ngôn ngữ này là một “người cộng tác khoa học”.

Trong khi đó, bộ phận nghiên cứu của Microsoft – Microsoft Research (MSR) – đang hỗ trợ một nhóm trường đại học Mỹ và Canada xây dựng một trạm quan sát khổng lồ dưới biển ở ngoài khơi gần bờ biển bang Washington (Mỹ). Dự án Neptune này sẽ kết nối hàng ngàn bộ cảm biến hóa học, địa lý và sinh vật học trên hơn 1.600 km sợi cáp quang và sẽ liên tục truyền dữ liệu đến các nhà khoa học trong khoảng thời gian đến 10 năm.

Các nhà khoa học sẽ có thể kiểm chứng những học thuyết của mình bằng cách xem xét dữ liệu thu thập được, bên cạnh đó, những công cụ phần mềm mà MSR đang phát triển sẽ tìm kiếm những khuôn mẫu và sự kiện mà các nhà khoa học không tiên liệu được, và gửi cho họ những phát hiện này.

Các nhà nghiên cứu tại trường Y Harvard và Đại học California, trong một cuộc nghiên cứu, đã dùng phương pháp phân tích thống kê để tìm kiếm dữ liệu về bệnh tim của 12.000 người và biết rằng chứng béo phì có vẻ “lây lan” thông qua các mối quan hệ xã hội. Trong khi đó, các nhà khoa học máy tính và sinh vật học cây trồng tại Đại học Cornell phát triển một thuật toán để lập và phân tích bản đồ ba chiều của protein khoai tây.

Những ứng dụng nói trên hầu như không có điểm chung nào, nhưng chúng đại diện cho một loại vấn đề khoa học liên quan đến một khối lượng lớn dữ liệu thực nghiệm phức tạp. Trong thực tế, những loại thông tin thô này quá nhiều đến nỗi các nhà khoa học thường không biết bắt đầu tìm hiểu từ đâu. Khoa học máy tính đang chỉ cho họ đường đi.

1.2 e-Science là gì?

Có nhiều định nghĩa khác nhau về e-Science nên nhóm sẽ trình bày quan điểm của nhiều chuyên gia về e-Science để có được cái nhìn rộng hơn về thuật ngữ mà nhóm đang trình bày.

Theo wikipedia: “e-Science là khoa học đòi hỏi sự tính toán với cường độ cao, được thực thi thông qua môi trường mạng phân tán hay là khoa học sử dụng các tập dữ liệu khổng lồ, đòi hỏi tính toán lưới. e-Science bao gồm các công nghệ cho phép sự cộng tác phân tán”.

Thuật ngữ e-Science được John Taylor, tổng giám đốc của Phòng khoa học và công nghệ của Liên Hiệp Anh đưa ra vào năm 1999 và được dùng để mô tả một dự án với số vốn lớn tại Liên Hiệp Anh, bắt đầu từ tháng 11-2000. Theo Taylor: “e-Science là sự cộng tác toàn cầu trong các lĩnh vực chính của khoa học, và cơ sở hạ tầng tính toán thể hệ kế tiếp cho phép hiện thực điều đó.”

Theo giáo sư Malcolm Atkinson, giám đốc NeSC tại Edinburgh và trung tâm e-Science quốc gia của Liên Hiệp Anh: “e-Science là sự phát triển có hệ thống của các phương pháp nghiên cứu đòi hỏi nhiều sự tính toán”. Theo ông, e-science sẽ thay đổi cách con người làm việc, giúp giải quyết các vấn đề nhanh hơn. Con người sẽ tập trung những nỗ lực của các cộng đồng khoa học, huy động những dữ liệu được chia sẻ và sức mạnh tính toán để đối mặt với những thách thức cấp bách. Một khi các trang web cho phép chúng ta chia sẻ thông tin, khoa học điện tử sẽ cho phép các nhóm nghiên cứu cộng tác để biến dữ liệu thành thông tin và kiến thức.

Giáo sư Jon Kleinberg, Đại học Cornell, nói : “Một xu hướng đang ngày càng trở nên rõ ràng là khoa học máy tính không còn là ngành cung cấp công cụ máy tính cho các nhà khoa học. Nó thực sự trở thành một phần của cách thức các nhà khoa học xây dựng học thuyết và suy nghĩ về những vấn đề của họ.” Theo ông Kleinberg, vai trò của thuật toán máy tính đối với khoa học trong thế kỷ 21 cũng sẽ tương tự như vai trò của toán học (đối với khoa học) trong thế kỷ 20. Giáo sư Kleinberg nói thêm rằng kho dữ liệu khổng lồ trên Internet sẽ thay đổi thực tiễn của những ngành khoa học liên quan đến hành vi con người. Theo lý giải của ông, số lượng dữ liệu khổng lồ và các phương pháp phân tích mới hiện nay đồng nghĩa với việc các nhà khoa học sẽ không còn phải

lập công thức chi tiết về những học thuyết và mô hình rồi kiểm chứng chúng trên dữ liệu thực nghiệm.

Tony Hey, Phó chủ tịch bộ phận nghiên cứu bên ngoài của Microsoft, đã nói về e-Science như một tập hợp những công nghệ dùng để hỗ trợ những dự án khoa học có lượng dữ liệu khổng lồ (thường được phân phối), có dữ liệu và nhiều người cộng tác kết nối với nhau, hay có sự tham gia của nhiều ngành khoa học, bao gồm ngành khoa học máy tính. Những dự án này, theo ông, thường rất phức tạp, và các công cụ, thuật toán, học thuyết của khoa học máy tính có thể giúp sắp xếp và làm rõ chúng. Ông Hey cho rằng chúng ta hiện đang tiến vào kỷ nguyên “khoa học tập trung vào dữ liệu” (data-centric science). Bản chất của ngành khoa học này là tập hợp dữ liệu, thường với số lượng lớn và từ nhiều nguồn khác nhau, rồi khai thác chúng để biết được những nội dung vốn sẽ không bao giờ xuất hiện nếu công việc này được làm thủ công hoặc từ việc phân tích bất kỳ một nguồn dữ liệu đơn lẻ nào.

Tony Blair, thủ tướng nước Anh năm 2002 đã nói về eScience như sau: “eScience được định hướng làm cho việc tận dụng nguồn lực to lớn của ngành khoa học máy tính, hệ thống tài nguyên dữ liệu khoa học và những thiết bị thực nghiệm tối tân dễ dàng như Web để truy cập thông tin”.

Tổng kết lại, e-Science là một khái niệm dựa trên những gì khoa học đang làm để phát minh ra những cái mới, cải tiến, phát triển những vấn đề hiện tại trong mọi ngành, lĩnh vực. Khi công nghệ mà các nhà khoa học sử dụng trong nghiên cứu ngày càng quy mô, phức tạp, yêu cầu về khả năng lưu trữ dữ liệu cũng phát triển tương ứng. e-Science là một công cụ cho phép các nhà khoa học lưu trữ, biểu diễn, phân tích và chia sẻ dữ liệu của họ với các nhóm nghiên cứu khác. e-Science giữ một vai trò qua trọng trong mọi mặt của nghiên cứu khoa học, bắt đầu với các nghiên cứu dựa trên các giả thiết tiềm năng, thử nghiệm thông qua mô phỏng, thử nghiệm có điều khiển một cách hệ thống, thu thập dữ liệu từ các bộ phận cấu thành và giải thích các dữ liệu khác biệt, không mong muốn.

1.3 Dùng thuật ngữ eScience, CyberInfrastructure hay eResearch

Thuật ngữ e-Science gần như, nhưng không hoàn toàn, đồng nghĩa với thuật ngữ CyberInfrastructure. Trong khi e-Science xuất xứ từ UK và châu Âu, CyberInfrastructure bắt nguồn từ US. Cả hai thuật ngữ đều đề cập đến việc sử dụng các

công nghệ tính toán dựa trên môi trường mạng để hỗ trợ sự cộng tác và cải tiến các phương pháp trong nghiên cứu khoa học. Trong khi e-Science chú trọng hơn đến nghiên cứu khoa học, thì CyberInfrastructure bao gồm cả các lĩnh vực ngoài khoa học, nhấn mạnh đến sự kết hợp giữa các nguồn tài nguyên siêu tính toán và sự cách tân.

Một số nhà nghiên cứu khác lại thích sử dụng một thuật ngữ khác: e-Research. e-Research là sự mở rộng của e-Science và CyberInfrastructure, bao gồm các lĩnh vực khác như khoa học xã hội và con người. e-Research nhấn mạnh đến việc sử dụng công nghệ thông tin để hỗ trợ các phương thức nghiên cứu hiện tại và tương lai.

Các đặc điểm chính của e-Research bao gồm:

- Sự cộng tác.
- Sử dụng công nghệ tính toán lưới.
- Tập trung vào dữ liệu.

1.4 Những đặc điểm, tính chất của e-Science

- **Lưu trữ:** Một hệ thống cần phải có khả năng lưu trữ và xử lý một lượng dữ liệu khổng lồ một cách hiệu quả với thời gian hợp lý.

- **Quyền sở hữu:** Các bên liên quan cần được bảo lưu quyền sở hữu về những nội dung và khả năng xử lý của họ. Tuy nhiên cũng cần phải cho phép người khác truy cập dưới các điều kiện và hoàn cảnh thích hợp.

- **Nguồn gốc:** Việc lưu trữ các thông tin đáng tin cậy cho phép sử dụng lại các kết quả, thử nghiệm, hay cung cấp bằng chứng về việc có được các thông tin đó.

- **Trong suốt:** Người dùng cần có khả năng tìm ra, truy cập và xử lý các nội dung liên quan bất cứ khi nào chúng xuất hiện trên Grid mà không cần biết nó nằm ở đâu.

- **Cộng đồng:** Phải cho phép hình thành, hoạt động, và giải tán các cộng đồng ảo với những tiêu chuẩn giới hạn thành viên và điều khoản hoạt động.

- **Kết hợp:** Thông tin cần phải được kết hợp từ nhiều nguồn, bằng nhiều cách khác nhau theo nhu cầu của người dùng. Các mô tả về nguồn gốc, nội dung sẽ được dùng để kết hợp nên các thông tin đầy đủ nghĩa.

- **Hội nghị:** Đôi khi việc nhìn thấy các thành viên khác của một hội nghị, các mô hình, sự hiển thị của những gì đang được thảo luận sẽ rất hữu dụng.

- **Chú giải:** Từ việc ghi nhận thông tin cho đến xuất bản các phân tích, cần thiết phải có các chú giải để làm giàu thêm mô tả về các nội dung số. Các siêu nội dung này

có thể áp dụng cho dữ liệu, thông tin, hay tri thức và phụ thuộc vào cách diễn giải quy ước.

- **Quy trình:** Để hỗ trợ quá trình ban hành và tự động hóa các xử lý, hệ thống cần mô tả về các xử lý đó.

- **Thông báo:** lời nhắc về việc có các thông tin mới tới cho phép thông báo cho người dùng và bắt đầu quá trình xử lý tự động.

- **Hỗ trợ quyết định:** Các kỹ thuật viên, nhà khoa học cần được cung cấp các thông tin và gợi ý xác đáng về vấn đề của họ.

- **Bảo lưu tài nguyên:** Cần làm cho quá trình bảo lưu tài nguyên trở nên dễ dàng. Điều này áp dụng cho các dụng cụ thí nghiệm, sự cộng tác (hội nghị...), và sắp xếp tài nguyên cho quá trình mô phỏng.

- **An ninh:** Có những yêu cầu về xác thực, mã hóa, và tính riêng tư với sự tham gia của nhiều tổ chức. Và các yêu cầu này cần được xử lý với sự can thiệp thấp nhất của con người.

- **Tin cậy:** Hệ thống trông có vẻ đáng tin cậy nhưng thật ra có cần xử lý những lỗi và ngoại lệ ở nhiều mức khác nhau, bao gồm cả quy trình thực hiện.

- **Video:** Cả video trực tiếp và được lưu trữ đều có vai trò nhất định, nhất là khi các video này được làm giàu thêm bởi các siêu nội dung liên quan tạm thời.

- **Phòng thí nghiệm thông minh:** Một ví dụ: khi các dụng cụ dò ra các mẫu (như thẻ barcode hay thẻ RFID), nhà khoa học dùng các thiết bị di động để ghi lại, và sự hiện hình hóa có thể được thực thi trong phòng thí nghiệm. Các công cụ từ xa có thể cho biết sự tồn tại của chúng, kết hợp với nhau, và thông báo về nội dung mà chúng nhận được.

- **Tri thức:** Tri thức hoạt động như một bộ phận quan trọng của e-Science. Ví dụ như: tìm kiếm tài liệu, con người, và các thiết kế thực nghiệm trước đó, chú thích cho các phân tích được đăng tải, và thiết lập phòng thí nghiệm cho con người.

- **Sự phát triển:** Hệ thống phải hỗ trợ sự phát triển mang tính cách mạng khi các nội dung và kỹ thuật xử mới hiện hữu.

- **Quy mô:** Quy mô của sự cộng tác khoa học tăng lên cùng với sự phát triển của tính toán, băng thông, khả năng lưu trữ, và độ phức tạp trong mối quan hệ giữa các thông tin.

1.5 Những yêu cầu cấp bách về cơ sở hạ tầng phục vụ các phương pháp nghiên cứu khoa học mới

Có một sự thay đổi lớn trong cách nghiên cứu khoa học so với trước đây mà cần phải có nhiều chuyên gia trong nhiều lĩnh vực khác nhau cùng hợp tác cả về con người và cơ sở vật chất. Các nghiên cứu khoa học ngày nay thường dựa trên việc mô phỏng (simulation base), phân tích dữ liệu (data analysis) và sự hợp tác của những người tiến hành thí nghiệm cũng như các nhà khoa học. Chúng ta sẽ sơ lược qua các phương pháp nghiên cứu mới này.

1.5.1 Khoa học tập trung vào dữ liệu (data-intensive)

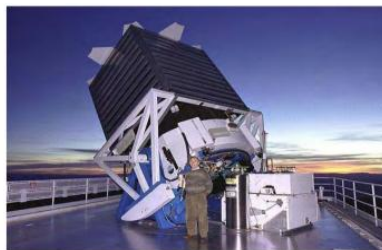
Sự phát triển của khoa học kỹ thuật dẫn đến các thiết bị nghiên cứu ngày càng có độ chính xác cao hơn điều này cũng làm cho khối lượng dữ liệu mà nó sinh ra ngày càng nhiều theo ước tính thì đến vài petabyte dữ liệu sẽ được tạo ra trong các thí nghiệm của thiên văn học, y học, vật lý nguyên tử, năng lượng, môi trường.

Large Hadron Collider (LHC) là một dự án của CERN đặt tại Geneve khi vận hành trở lại vào cuối năm 2009 có thể sinh ra một lượng dữ liệu lên đến vài petabyte hàng năm. LHC là dự án lớn nhất từ trước đến nay trong lĩnh vực vật lý. Mỗi thí nghiệm muốn tiến hành cần phải có đến sự cộng tác của hơn 5000 nhà vật lý trên toàn thế giới. Quá trình phân tích dữ liệu thu thập được cũng cần phải có sự cộng tác của các nhiều tổ chức tham gia mục tiêu, là tìm ra dấu hiệu của Higgs boson.

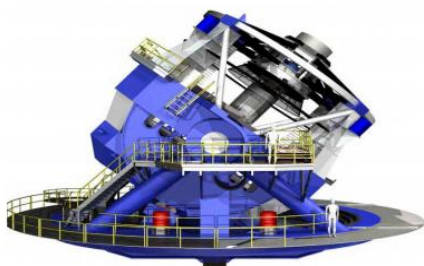
Các thiết bị theo dõi và tiên đoán sự cố trong công nghiệp cũng có thể tạo ra một lượng lớn dữ liệu. Các thiết bị cảm ứng dùng để theo dõi nhiệt độ, áp suất, chấn động trong các mỗi động cơ của hàng ngàn động cơ do Rolls-Royce sản xuất cho các phi cơ của trans-Atlantic có thể sinh ra hàng petabyte dữ liệu hàng năm.



LHC(Large Hadron Collider) 60TB/day



Apache Point Telescope SDSS 15TB/day



Large Synoptic Survey Telescope 30TB/day



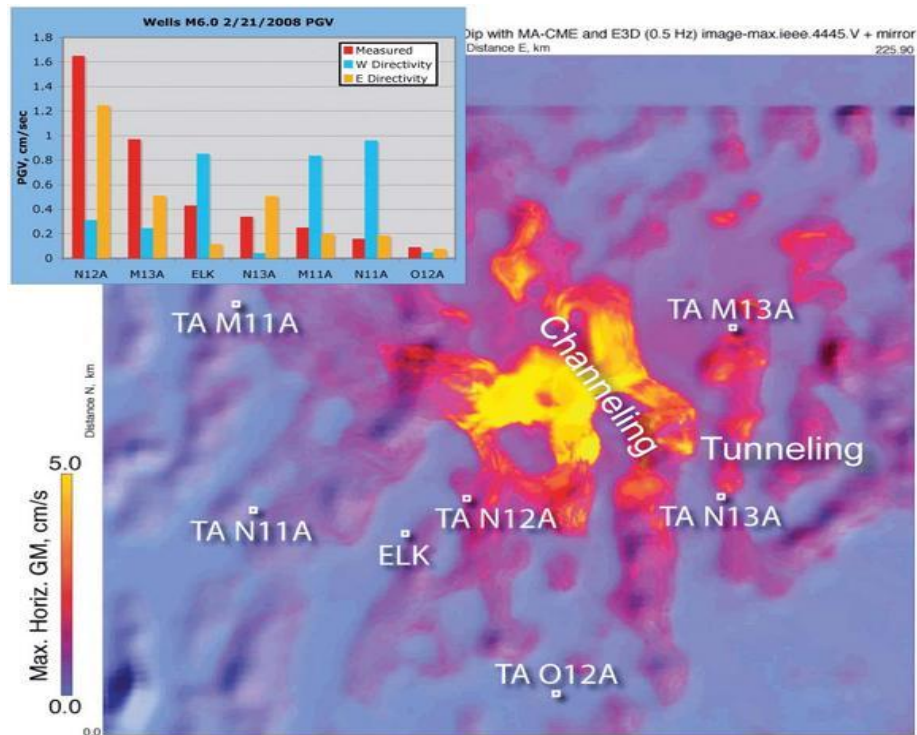
Illumina Genome Analyzer 1TB/day

1.5.2 Nghiên cứu khoa học dựa trên việc giả lập và mô phỏng

Mô phỏng (numerical simulation) là một hướng giải quyết các vấn đề khoa học mà dựa chủ yếu vào việc sử dụng các siêu máy tính để thực hiện mô phỏng các hiện tượng tự nhiên như sự biến đổi của khí hậu hay sự kết hợp của các lỗ đen trong thiên văn học, động đất, lũ, ...

Năm 2003, Japanese Earth Simulator đã thực hiện mô phỏng khí hậu của trái đất với một siêu máy tính có tốc độ xử lý lên đến 40 teraflop/sec để có thể mô phỏng hơn 10km theo chiều rộng và khối lượng dữ liệu sinh ra cho mỗi lần mô phỏng lên đến vài chục terabyte.

Trong lĩnh vực hóa thì các thí nghiệm có thể thực hiện thông qua các máy tính với các tập dữ liệu sẵn có về các hóa chất và những đặc tính của chúng thì việc tiến hành các thí nghiệm có thể thực hiện một cách nhanh chóng. Các phân tử mới có thể được tạo ra từ các thao tác trên máy tính thay vì tiến hành ở phòng thí nghiệm. Comb-e-Chem một trong những dự án (pilot project) thuộc e-Science đã hiện thực ý tưởng này, mục tiêu của dự án là tạo ra các kết hợp mới sau đó sẽ xác định cấu trúc và thuộc tính của các hợp chất mới được tạo ra để tìm kiếm các công thức hóa học mới. Việc tổng hợp sẽ được thực hiện song song và có thể tạo ra hàng trăm nghìn tổ hợp cùng lúc.



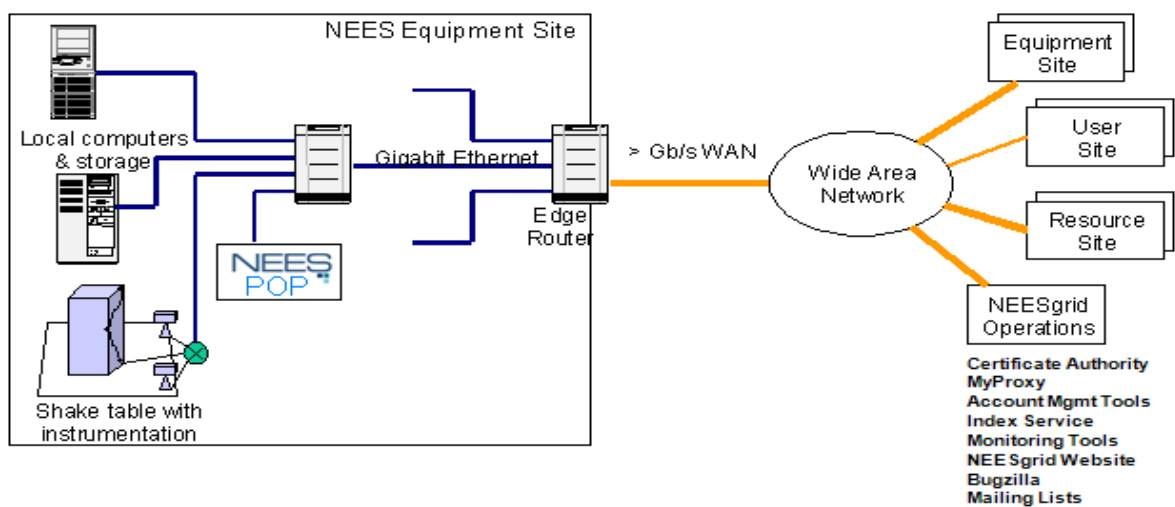
Mô phỏng một trận động đất

1.5.3 Truy xuất từ xa đến các công cụ và dữ liệu

Mô phỏng và phân tích dữ liệu đóng vai trò ngày càng quan trọng trong các lĩnh vực khoa học ngày nay, tuy nhiên, việc tiến hành các thí nghiệm thực tế cũng không thể thiếu được. Ngày càng có nhiều thiết bị thí nghiệm mới được phát triển, bên cạnh đó thì hạ tầng mạng băng thông cũng ngày càng mở rộng, cho phép các nhà nghiên cứu có thể kết hợp các thiết bị này lại với nhau để tiến hành các thí nghiệm phức tạp đòi hỏi sự tham gia của các chuyên gia trong nhiều lĩnh vực. Việc thiết kế, tiến hành và giám sát các thí nghiệm có thể được thực hiện thông qua mạng internet bằng cách truy xuất từ xa vào thiết bị.

Network for Earthquake Engineering Simulation (NEES) là chương trình do NSF phát triển dưới sự chỉ đạo của George E. Brown Jr nhằm tìm cách giảm thiệt hại của các trận động đất thông qua việc sử dụng các công cụ để giả lập các trận động đất và từ đó có thể tìm ra được các cấu trúc cũng như vật liệu mới có thể chịu được các cơn chấn động. NEESgrid được triển khai nhằm liên kết các nhà khoa học ở US có thể chia sẻ cũng như kết hợp các thiết bị thí nghiệm, nguồn dữ liệu và cả nguồn tài nguyên

tính toán. NEESgrid middleware cho phép các nhóm cộng tác với nhau (bao gồm cả các thành viên đăng nhập từ xa) lên kế hoạch, thiết kế, và tiến hành các thực nghiệm sau đó nguồn dữ liệu thu được chia sẻ để xử lý. NEESgrid cho phép các cộng tác viên có thể đăng nhập từ xa để theo dõi cũng như vận hành các thí nghiệm đã được chuẩn bị sẵn, bên cạnh đó họ cũng được cho phép sử dụng các tài nguyên tính toán và các công cụ phân tích mà nguồn mở để xử lý nguồn dữ liệu thu được. NEESgrid hỗ trợ việc chi sẻ dữ liệu thông qua cung cấp nơi lưu trữ, chuẩn cho định dạng dữ liệu và metadata.



Mô hình NEESgrid

1.5.4 Tổ chức (Virtual Organization) hay cộng đồng ảo (Virtual Community)

Phần lớn các nghiên cứu khoa học ngày nay phải cần có sự tham gia của các nhà khoa học trong các lĩnh vực khác nhau và các trung tâm nghiên cứu đặt ở khắp nơi trên thế giới, nên làm thế nào để việc cộng tác này trở nên hiệu quả và ít chịu ảnh hưởng bởi yếu tố địa lý là vấn đề cấp bách cần giải quyết. Từ đó, các cộng đồng khoa học ảo hình thành (VO) dựa trên nền tảng của internet tốc độ cao. Các thành viên trong VO ở các vị trí địa lý khác nhau trên thế giới sẽ đóng góp tài nguyên tính toán và cho phép truy xuất từ xa các các thiết bị thí nghiệm. Các tài nguyên của VO sẽ được kết dính lại với nhau như một thể thống nhất cho các thành viên trong VO sử dụng.

1.5.5 Phát triển một cơ sở hạ tầng mới

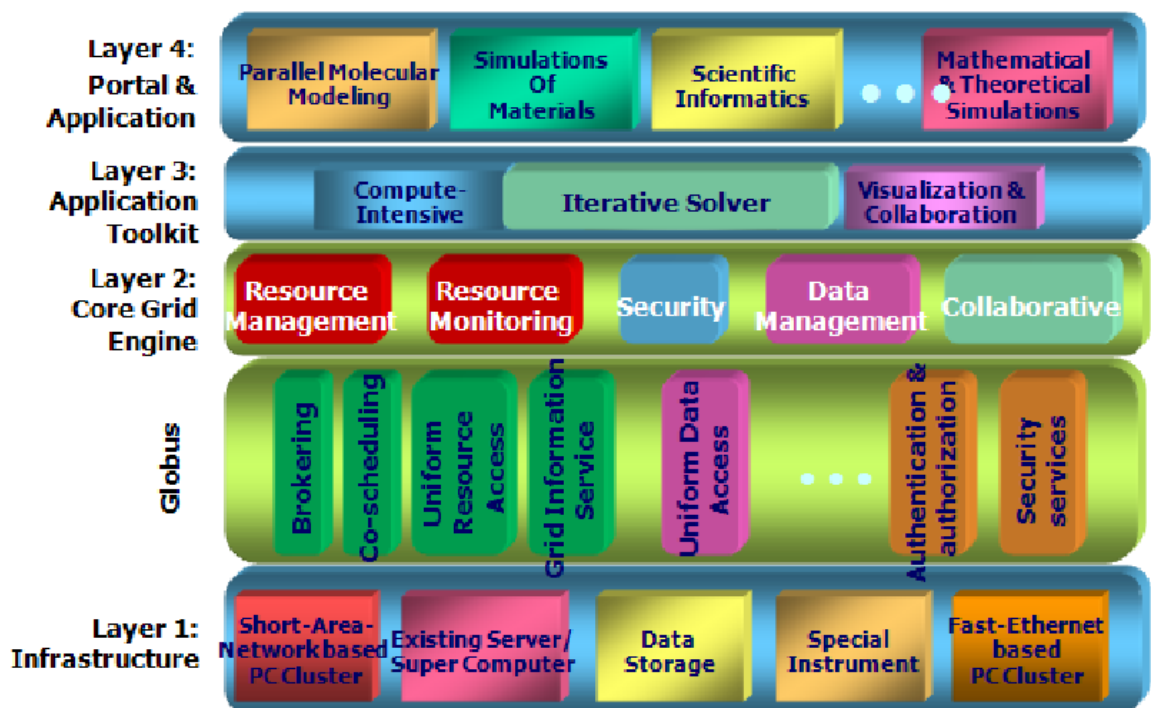
Từ những nhu cầu về tài nguyên tính toán, nơi lưu trữ và làm thế nào để nâng cao hiệu quả sự cộng tác của các nhóm nghiên cứu. Các tổ chức chính phủ đã

nhận thấy sự quan trọng và cần thiết để phát triển cơ sở hạ tầng mới để hỗ trợ quá trình nghiên cứu. Cơ sở hạ tầng mới này phải đáp ứng và giải quyết được các yêu cầu được đưa ra, phải mang tính toàn cầu, hỗ trợ nhiều lĩnh vực khoa học khác nhau, xây dựng Grid middleware và các tiện ích cho nó theo định hướng phục vụ cho các ngành khoa học.

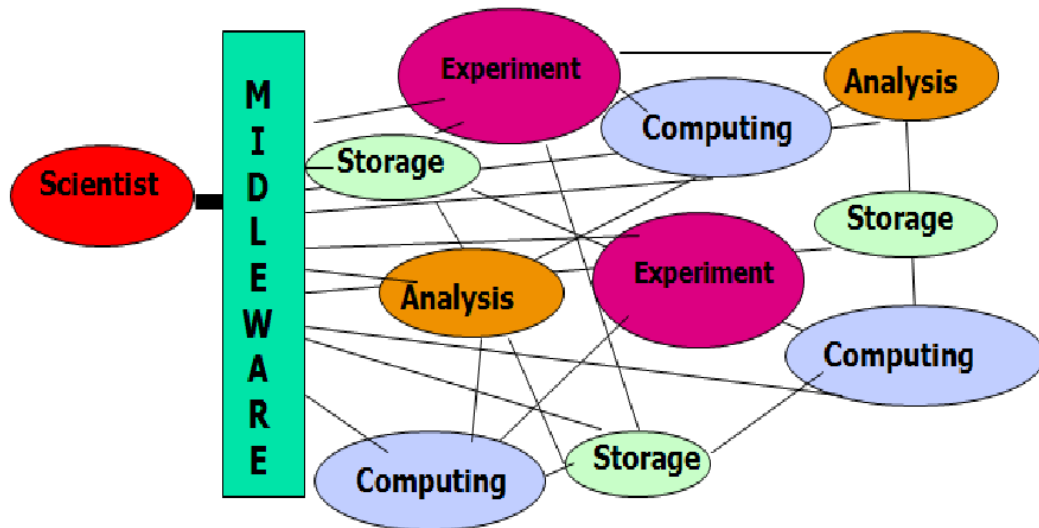
1.6 Cơ sở hạ tầng hiện thực cho eScience

Để hiện thực một hệ thống eScience với những tính chất trên đòi hỏi phải đầu tư xây dựng một cơ sở hạ tầng vật lý hiện đại và một cơ sở hạ tầng phần mềm được thiết kế với quan niệm mới hướng tới khoa học mở và sự hợp tác toàn cầu.

Chính phủ các nước tiên tiến hiện nay đều nhận thấy được tầm quan trọng của việc xây dựng một hệ thống quy mô lớn như eScience để nâng cao vị thế của quốc gia mình. Trong bối cảnh thế giới hiện tại, eScience được tập trung phát triển dựa trên nền tảng hệ thống lưới hiện có.



A eScience Grid based framework



Mô hình phần mềm trong suốt sự phức tạp của hạ tầng vật lý đối với những người làm khoa học

Kiến trúc mới phải hướng tới việc đơn giản hóa hoạt động nghiên cứu khoa học với sự hỗ trợ của hệ thống máy tính tính toán lưới. Phát triển cơ sở hạ tầng phần mềm phải hướng đến việc trong suốt hóa sự phức tạp của hệ thống vật lý đối với các chuyên gia, các nhà khoa học để họ có thể tập trung vào chuyên môn khoa học của mình.

CHƯƠNG II

E-SCIENCE TRÊN THẾ GIỚI

2.1 e-Science tại UK

Có rất nhiều nhóm và tổ chức trên toàn thế giới hình thành nên chương trình e-Science. Các thành viên chính bao gồm các trường đại học, nhóm nghiên cứu và các cộng tác viên khác được tài trợ bởi chính phủ. Mục tiêu của các nhóm như vậy là để phát triển chương trình e-Science nói riêng, và qua đó cũng thúc đẩy sự phát triển của khoa học nói chung. Nghiên cứu về các lĩnh vực như vật lý hạt đòi hỏi một cơ sở hạ tầng ở mức mà chỉ có e-Science mới có thể đáp ứng. Dự án e-Science ở UK là một trong những dự án được đánh giá là khá thành công, được nhiều nơi trên thế giới tìm hiểu và học tập. Dự án e-Science ở UK bao gồm rất nhiều nhóm sẽ được đề cập đến dưới đây.

2.1.1 Các nhóm e-Science tại UK



Các thành viên của dự án e-Science tại UK

Các nhóm trong dự án e-Science ở UK bao gồm:

Science & Technology Facilities Council – STFC - Hội đồng Cơ sở Khoa học và Công nghệ của Anh. Cung cấp các dịch vụ hỗ trợ nghiên cứu khoa học ở UK và cả châu Âu. Bao gồm các nhà nghiên cứu (trong các lĩnh vực như vũ trụ học, kinh doanh) và các phòng thí nghiệm. Mục tiêu của STFC là đáp ứng được yêu cầu của các nhà nghiên cứu trên toàn thế giới, với các yêu cầu rất lớn về tính toán.

Oxford e-Research Centre – OeRC đáp ứng các nhu cầu về tính toán và kỹ thuật cho các nhóm nghiên cứu ở đại học Oxford và cả các nhóm cộng tác khác khi tham gia vào dự án e-Science ở UK. Tầm nhìn của OeRC là hướng tới sự đổi mới, cải tiến quá trình nghiên cứu khoa học.

National e-Science Centre - Các nhà nghiên cứu thuộc Trung tâm Khoa học điện tử quốc gia (NeSC) tại thành phố Edinburgh (Scotland) có kế hoạch phát triển công nghệ máy tính mới có tên gọi là hệ thống mạng Grid. Được xem như là Internet thế hệ 2, NeSC hy vọng Grid sẽ giúp cho các tài nguyên máy tính không lờ trên khắp thế giới có thể được tận dụng và chia sẻ để giúp giải quyết những thách thức lớn nhất trong các lĩnh vực như y tế, vật lý, thiên văn học và chế tạo máy. Ngoài ra, các nhà nghiên cứu cũng hy vọng Grid sẽ đến được với các cá nhân sử dụng máy tính tại nhà hay trong các lĩnh vực thương mại. Điều này cho phép hình thành một phiên bản Internet mới cho phép thực hiện các phép tính phức tạp thay vì chỉ tải dữ liệu. Kế hoạch trên được đưa ra vào thời điểm hơn 1.000 chuyên gia máy tính trên khắp thế giới tập trung tại Edinburgh để tham dự Diễn đàn Toàn cầu về Grid lần thứ 5.

e-Science North West Centre - ESNW là một bộ phận nghiên cứu phần lớn từ đại học Manchester, có liên hệ với các đối tác tài chính. Mục tiêu của họ là cung cấp cơ sở hạ tầng cho Grid, qua đó các nhóm nghiên cứu đa ngành có thể đạt được mục đích của họ. ESNW được thành lập năm 2001, bao gồm các thành viên của School of Computer Science và Research Computing Services.

Cambridge e-Science Centre - CeSC được thành lập năm 2001, dựa trên Centre for Mathematical Sciences. Mục tiêu của CeSC là cung cấp mạng lưới tính toán cho các cộng tác viên nghiên cứu, hỗ trợ các nhà khoa học tại Cambridge, đạt được các cải tiến khoa học thông qua các khái niệm tính toán truyền thống để thu được kỹ thuật lưu trữ mạnh hơn, tốt hơn. Họ cộng tác với nhiều nhóm e-Science khác, bao gồm đại học Lancaster và IBM.

London e-Science Centre - Trung tâm e-Science London (LeSC) bắt đầu hoạt động từ tháng 9/2001, như một thành phần của Chương trình khung UK e-Science. Nhiệm vụ của Trung tâm LeSC là nghiên cứu phát triển một số công nghệ hỗ trợ triển khai e-Science cho các trường đại học, viện nghiên cứu ở London và vùng Đông Nam nước Anh. Cơ sở thành lập LeSC dựa trên Trung tâm nghiên cứu về tính toán song song, một dự án hợp tác giữa Đại học Imperial College và Công ty Fujitsu (1994-2000) và các nhóm nghiên cứu liên quan đến Tính toán Hiệu năng cao của Trung tâm Tính toán song song của Imperial College (thành lập từ năm 1996). Hiện nay Trung tâm LeSC trực thuộc Khoa CNTT (Computing Department) của ĐH Imperial College. Trung tâm LeSC còn có trách nhiệm quản lý toàn bộ hạ tầng mạng e-Science của Đại học Imperial College, bao gồm các máy chủ tính toán mạng lưới, cổng thông tin e-Science Portal, hệ thống truy cập mạng lưới (AccessGrid) có khả năng tổ chức Hội nghị truyền hình trên mạng Internet với với tham gia đồng thời của 15 Trung tâm nghiên cứu khoa học khác nhau thuộc Anh, các quốc gia từ Châu Âu, Bắc Mỹ, Châu Á – Thái Bình Dương đã triển khai hạ tầng AccessGrid dựa trên nền Grid (mạng lưới).

Welsh e-Science Centre - WeSC dựa trên Cardiff School of Computer Science tại đại học Cardiff. Mục tiêu của WeSC là xây dựng và phát triển cơ sở hạ tầng cho e-Science qua đó các nhóm nghiên cứu và cộng tác viên có thể đạt được các yêu cầu tính toán của mình. Mục tiêu của họ còn là phát triển e-Science tại xứ Wales và tây nam UK, và hợp tác với các nhà kinh doanh và nghiên cứu.

Các nhóm e-Science khác – Có rất nhiều các nhóm nghiên cứu e-Science khác tại UK, và tất cả họ đóng một vai trò trong sự phát triển và cải tiến kỹ thuật của dự án e-Science tại UK. Chúng ta có thể kể đến các nhóm như Belfast e-Science Centre, National Grid Service, Lancaster University Centre for e-Science và Southampton e-Science Centre. Công việc của họ có vai trò quan trọng cho tương lai của e-Science và những nỗ lực nghiên cứu mà các nhà nghiên cứu đang thực hiện.

2.1.2 Các dự án về e-Science của UK

- **AstroGrid.** <http://www.astrogrid.org/>. Hỗ trợ bởi UK.e-Science Program. Xây dựng hạ tầng mạng lưới cho Đài thiên văn ảo, tích hợp giao tiếp với CSDL thiên văn và cho phép truy cập từ xa cũng như mô phỏng dữ liệu.

- **Comb-e-chem.** <http://www.combechem.org/>. Hỗ trợ bởi UK. e-Science Program. Hỗ trợ tổng hợp của các mẫu phát âm bằng cách kết hợp cấu trúc và nguồn dữ liệu bên trong hệ thống chia sẻ thông tin mạng lưới và mô trường chia sẻ kiến thức.
- **Discovery Net.** Hỗ trợ bởi UK. e-Science Program. Thử nghiệm e-Science để xử lý dữ liệu với khối lượng lớn trong các lĩnh vực hoá sinh, nghiên cứu môi trường, v.v...
- **Distributed Aircraft Maintenance Environment.** <http://www.cs.york.ac.uk/dame/>. Hỗ trợ bởi UK. e-Science Program. Xây dựng hệ thống hỗ trợ quyết định dựa trên nền Grid để bảo trì động cơ máy bay.
- **Grid Enabled Optimisation & Design Search for Engineering (GEODISE).** <http://www.geodise.org/>. Hỗ trợ bởi UK. e-Science Program. Hỗ trợ truy cập dựa trên grid tới các kho tri thức, tối ưu hoá các công cụ tìm kiếm, phân tích độ bền vững của các mẫu thiết kế.
- **GridPP.** www.gridpp.ac.uk. Hỗ trợ bởi UK. e-Science Program. Tạo lập và ứng dụng mô hình tính toán mạng lưới trong nghiên cứu vật lý hạt tử (Particle Physics) của UK.
- **MyGrid.** www.mygrid.org.uk. Hỗ trợ bởi UK. e-Science Program. Phát triển và ứng dụng khoa học điện tử e-Science trong các ứng dụng sinh tin học.
- **Reality Grid.** <http://www.realitygrid.org/>. Hỗ trợ bởi UK. e-Science Program. Hỗ trợ mô hình hoá và mô phỏng thực của các khối vật chất đặc ở mức phân tử và hạt tử.
- **UK Grid Center.** www.grid-support.ac.uk. Hỗ trợ bởi UK. e-Science Program. Trung tâm hỗ trợ các dự án Grid trên lãnh thổ UK.

2.2 e-Science tại US

2.2.1 Cyberstructure

Cyberinfrastructure là thuật ngữ được United States National Science Foundation (NSF) sử dụng tại hội nghị năm 2003 để trả lời cho câu hỏi: NSF, với tư cách là cơ quan chính của chính phủ tài trợ cho các nghiên cứu cơ bản, vượt qua các rào cản hiện tại để đạt được sự đột phá trong tính toán hiệu năng cao, khiến cho nó thật sự hữu dụng cho các nhà khoa học, kỹ sư, nghiên cứu sinh và các công dân của US?

Theo NSF: “Cyberinfrastructure là môi trường nghiên cứu trong đó các dịch vụ: tính toán cao cấp, cộng tác, truy xuất dữ liệu, quản lý thông tin được cung cấp cho người nghiên cứu thông qua hệ thống mạng tốc độ cao. Gồm phần mềm, phần cứng, các công nghệ khác và con người cần thiết để hỗ trợ cho sự khám phá tri thức trong các lĩnh vực khoa học và công nghệ hiện tại và tương lai.”

Một kỷ nguyên mới cho nghiên cứu khoa học và công nghệ đã lộ dạng nhờ sự hỗ trợ của quá trình phát triển không ngừng trong kỹ thuật tính toán, thông tin, truyền thông và nhu cầu nâng cao độ phức tạp lẫn qui mô trong các vấn đề cần giải quyết ngày nay. Kỹ thuật này đã đạt đến mức có thể hiện thực hóa khái niệm cyberinfrastructure: những thể loại mới của môi trường tri thức khoa học, công nghệ, cách thức tổ chức và thực hiện nghiên cứu theo phương pháp mới với hiệu quả cao. Có thể nói: “Nếu cơ sở hạ tầng dân dụng (civil infrastructure) cần thiết cho sự phát triển của kinh tế công nghiệp thì cơ sở hạ tầng công nghệ thông tin (CyberInfrastructure) cần cho sự phát triển kinh tế tri thức trong tương lai.”

2.2.2 Các dự án về e-Science của US

- **Access Grid.** www.accessgrid.org. Hỗ trợ bởi DOE, NSF. Tạo lập hệ thống cộng tác trên mạng, sử dụng các công cụ hiển thị truyền thông.
- **DISCOM.** www.cs.sandia.gov/discom. DOE Defense Programs. Kiến tạo hệ thống tác nghiệp Grid với khả năng truy xuất tới các phòng thí nghiệm chế tạo vũ khí của Bộ năng lượng Mỹ.
- **DOE Science Grid.** www.sciencegrid.org. DOE Office of Science. Kiến tạo hệ thống tác nghiệp Grid với khả năng truy cập tới các tài nguyên và ứng dụng tại các phòng thí nghiệm của Bộ năng lượng Hoa kỳ và các trường đại học tham gia.
- **Earth System Grid (ESG).** www.earthsystemgrid.org. DOE Office of Science. Cung cấp và phân tích các tập hợp mô hình khí hậu lớn phục vụ cộng đồng dự báo thời tiết.
- **Fusion Collaboratory.** www.fusiongrid.org. DOE Office of Science. Kiến tạo phòng thí nghiệm hợp tác tính toán quốc gia cho nghiên cứu khuếch xạ.

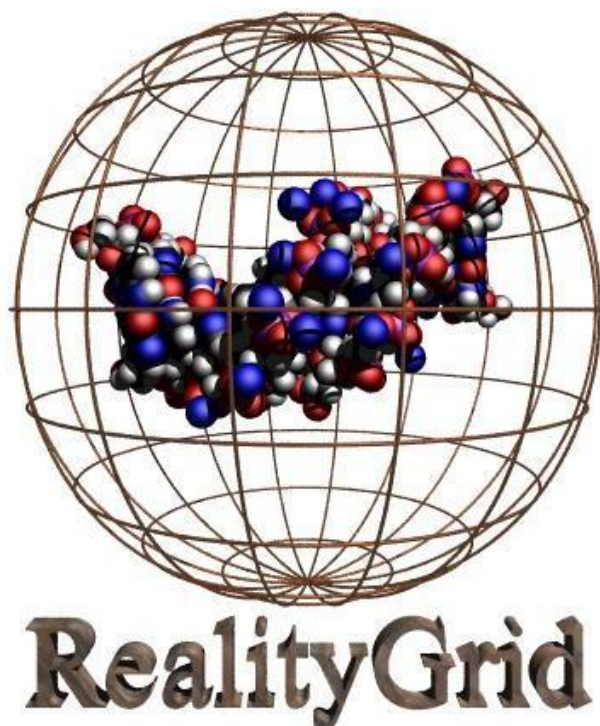
- **Globus.** www.globus.org. DARPA, DOE, NASA, NSF. Nghiên cứu xây dựng hạ tầng và công cụ cho Grid dựa trên cộng đồng, mã nguồn mở, kiến trúc mở của Grid
- **Grid Research Integration Development & Support (GRIDS) Center.** www.grids-center.org. NSF. Tích hợp, triển khai, hỗ trợ của hạ tầng lớp trung cho nghiên cứu – giáo dục của NSF.
- **Grid Application Development Software.** www.hipersoft.rice.edu/grads. NSF. Nghiên cứu phương pháp luận, mô hình khung phát triển ứng dụng mạng lưới.
- **Grid Physics Network (GriPhyN).** www.griphyn.org. NSF. Nghiên cứu – Phát triển công nghệ trong lĩnh vực phân tích dữ liệu của vật lý thực nghiệm: ATLAS, CMS, LIGO, ...
- **Information Power Grid.** www.ipg.nasa.gov. NASA. Kiến tạo và ứng dụng Grid cho khoa học hàng không vũ trụ.
- **Network for Earthquake Eng. Simulation Grid.** www.neesgrid.org. NSF. Kiến tạo và ứng dụng Grid trong nghiên cứu động đất cho kỹ nghệ.
- **National Virtual Observatory (NVO).** www.srl.caltech.edu/nvo. NSF. Kiến tạo và ứng dụng Grid để phân tích xử lý dữ liệu trong thiên văn
- **Particle Physics Data Grid (PPDG).** www.ppdg.net. DOE Office of Science. Kiến tạo và ứng dụng mạng lưới nhằm phân tích dữ liệu trong vật lý năng lượng cao và các thí nghiệm vật lý hạt nhân.
- **Southern California Earthquake Center 2.** www.scec.org. NSF. Hệ thống mô hình hoá địa chất đầy đủ, sử dụng Grid và hệ thống dựa trên tri thức.
- **TeraGrid.** www.teragrid.org. NSF. Hạ tầng mạng kết nối bốn vị trí tài nguyên khoa học lớn của Mỹ với băng thông 40 Gb/s.

2.3 Chi tiết một số dự án tiêu biểu

2.3.1 RealityGrid

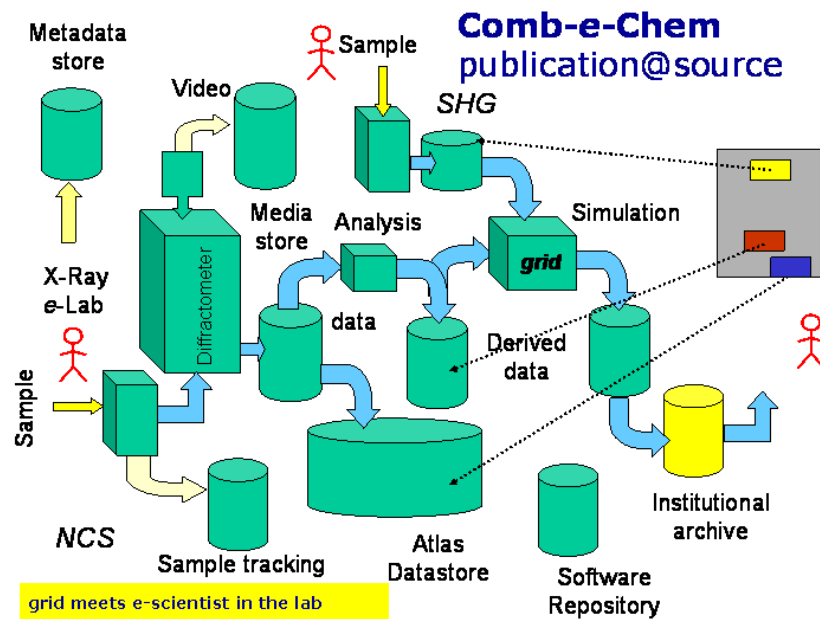
Dự án này được thực hiện dưới sự lãnh đạo của giáo sư Peter Coveney với sự tham gia của các trường Đại học Edinburgh, Loughborough, Manchester và Oxford. Mục tiêu của dự án này là mô hình hóa cấu trúc phân tử của vật rắn và cung cấp một môi trường hỗ trợ việc tạo ra các vật chất mới. Các máy tính hiệu năng cao sẽ được sử

dụng để xây dựng một môi trường trực quan biểu diễn các mô hình này và cung cấp các tiện ích. Đây sẽ là môi trường cho những người làm khoa học trực quan hóa cấu trúc của vật thể, so sánh với thực tế và tích hợp vào dữ liệu thực nghiệm.



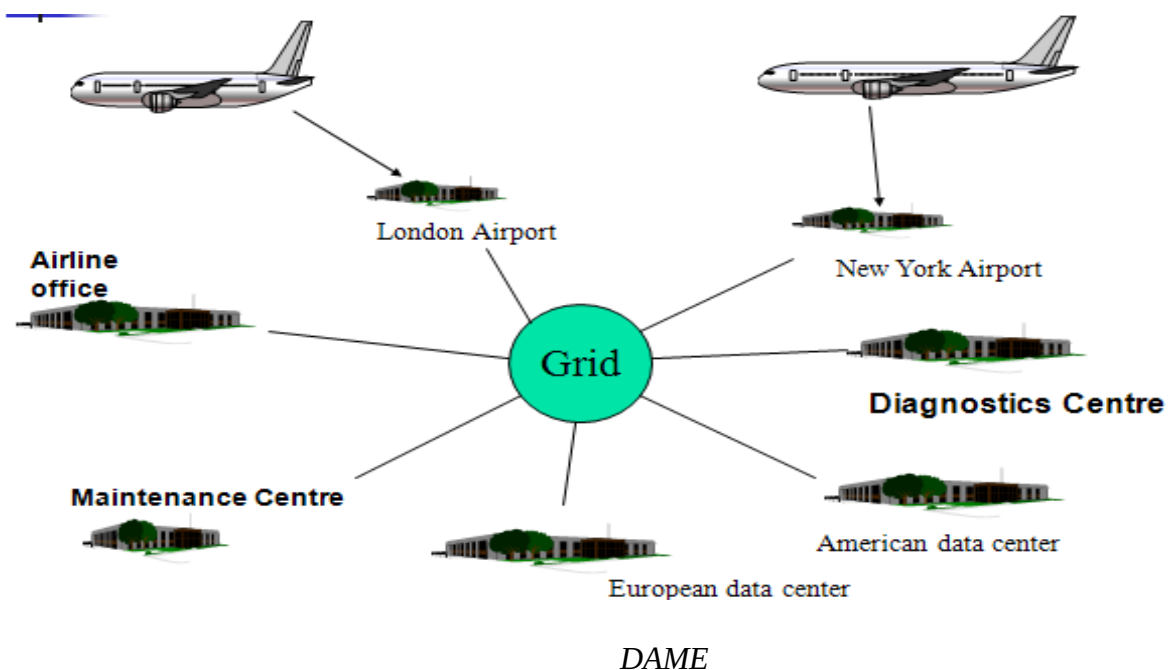
2.3.2 Comb-e-Chem

Comb-e-Chem được thực hiện dưới sự cộng tác của các trường Đại học Southampton và Bristol, do Dr. Jeremy Frey lãnh đạo. Comb-e-Chem chủ yếu dùng để tạo ra các hợp chất mới thông qua việc áp dụng các phương pháp kết hợp giữa các hóa chất với nhau. Phương pháp kết hợp để tìm ra các hợp chất mới từ một nguồn dữ liệu về các tính chất hóa học, dữ liệu được tích lũy ngày càng lớn. Để đạt được hiệu quả cao thì một platform sử dụng hạ tầng Grid được sử dụng kết hợp các cấu trúc được sinh ra và nguồn dữ liệu tri thức được đã được lưu trữ. Platform được xây dựng phải hỗ trợ việc chọn lọc dữ liệu, bao gồm việc tạo ra dữ liệu cũng như khả năng xử lý dữ liệu. Ngoài ra, Comb-e-Chem còn cung cấp giao diện để tương tác với người sử dụng cho phép thực hiện các bước mô hình hóa, thiết kế và thực hiện các thí nghiệm từ xa (online accessible).



2.3.3 Distributed aircraft maintenance environment (DAME)

Đây là dự án theo dõi độ an toàn của các chuyến bay thông bằng cách lắp đặt các thiết bị cảm ứng ở động cơ máy bay. Các thiết bị này sẽ liên tục gửi dữ liệu về cho các trạm (Grid node) ở mặt đất để tiến hành xử lý và cho biết thông tin hiện tại (real-time) của máy bay. Đây là dự án có sự tài trợ của Rolls-Royce. Chi tiết hơn của dự án này sẽ được trình bày ở phần sau của báo cáo.



2.3.4 myGrid

Dự án này có sự tham gia của nhiều thành viên nhất bao gồm các trường Đại học Manchester, Southampton, Nottingham, Newcastle, và Sheffield hợp tác với viện công nghệ sinh học Châu Âu. Mục tiêu của myGrid là phát triển một hạ tầng (infrastructure) phía bên hạ tầng của Grid (Grid infrastructure) nhằm giúp cho các nhà khoa học có thể dễ dàng sử dụng các hệ thống tài nguyên được phân bố trên Grid. Một workbench riêng cho các nhà khoa học (e-Scientist's workbench) sẽ được phát triển, workbench sẽ thay đổi tùy theo nhu cầu của thực nghiệm cứu. Vì myGrid được thiết kế và phát triển nhằm cho các ứng dụng về công nghệ sinh học nên phải cung cấp hai môi trường ứng dụng, một dùng để hỗ trợ quá trình phân tích dữ liệu gen, hai là dùng để giải thích cho các mẫu dữ liệu. GSK, AstraZeneca, IBM and SUN là các doanh nghiệp cộng tác phát triển dự án myGrid.

2.3.5 GridPP

GridPP được hình thành dựa trên sự hợp tác của các nhà vật lý học và các chuyên gia trong lĩnh vực khoa học máy tính của UK và CERN. GridPP phát triển từ năm 2001 hướng theo 3 mục tiêu chính, thứ nhất là phát triển một ứng dụng cho phép các nhà vật lý nguyên tử có thể chạy (run) các jobs trên Grid, hai là viết một middleware dùng để quản lý và phân phối các công việc tính toán (computing tasks) trên Grid bên cạnh đó cũng phải chú trọng đến vấn đề bảo mật, mục tiêu còn lại của GridPP là triển khai một hạ tầng tính toán (computing infrastructure) với các sites đặt ở UK như là một prototype về Grid cho LHC của CERN.

2.3.6 AstroGrid

AstroGrid được xây dựng với sự tham gia của các nhà thiên văn học và các nhà khoa học máy tính ở các trường Đại học Edinburgh, Leicester, Cambridge, Queens Belfast, UCL và Manchester. Mục tiêu của AstroGrid là xây dựng một cơ sở hạ tầng cho Grid hỗ trợ “Virtual Observatory” (VO). VO cung cấp một giao diện giống như ứng dụng trên desktop giúp cho các nhà thiên văn có thể khảo sát và đánh dấu tài nguyên trên toàn thế giới, tìm kiếm, lưu trữ và chia sẻ dữ liệu.

2.3.7 eDiamond

Đại học Tổng hợp Oxford cùng IBM và chính phủ Anh hợp tác xây dựng một mạng tính toán hiện đại, được gọi là eDiamond, giúp sớm kiểm tra và chẩn đoán bệnh ung thư vú cũng như cung cấp nhiều thông tin hơn cho các chuyên gia y tế để điều trị căn bệnh này. Với khoản đầu tư gần 6 triệu USD từ IBM và chính phủ Anh, dự án eDiamond là một phần trong phát kiến Khoa học điện tử (e-Science) trị giá 118 triệu euro của nước này. Đây sẽ là mạng đầu tiên được xây dựng toàn bộ bằng công nghệ hiện có đã được thương mại hoá mà không cần phải chỉnh sửa hay thay đổi gì thêm, sử dụng phần mềm do hãng Mirada Solutions phát triển có thể chuẩn hoá các hình ảnh số hoá mới và có sẵn về bệnh ung thư vú.

Bên cạnh việc cho phép các bệnh viện lưu trữ và chia sẻ hình ảnh điện tử về bệnh ung thư vú, eDiamond sẽ cung cấp cho các bác sĩ những công cụ cùng khả năng phân tích tiên tiến nhằm chuẩn đoán bệnh tốt hơn. Hình ảnh về ung thư vú sẽ được lưu trữ và khai thác, cho phép bác sĩ phát triển những phương pháp điều trị mới thông qua thực hiện những nghiên cứu chuyên sâu nhằm xác định ảnh hưởng của môi trường và phong cách sống tới sự phát triển của căn bệnh này.

2.4 e-Science Core Programme

e-Science Core Programme được phát triển như là một cơ sở hạ tầng Grid chung cho các dự án được miêu tả ở trên (pilot projects). Với sự cộng tác của các nhà khoa học, khoa học máy tính, và các doanh nghiệp nhằm xây dựng và phổ biến một framework có thể đáp ứng được các ứng dụng trong nhiều lĩnh vực khoa học khác nhau cũng , thích hợp với các ứng dụng trong công nghiệp.

Cấu trúc của e-Science Core Programme bao gồm sáu thành phần:

2.4.1 Hiện thực hạ tầng mạng kết nối các e-Science Centres

Các e-Science Centres được phân bố trên khắp UK như hình bên dưới:



Các trung tâm e-Science này rất quan trọng cho toàn dự án e-Science vì nó giữ các vai trò:

- Phân bổ tài nguyên tính toán, dữ liệu và cài đặt các dịch vụ chuẩn (standard) và cơ bản để phục vụ cho UK e-Science Grid.
- Thu hút nguồn đầu tư từ cộng tác của công nghiệp để xây dựng Grid middleware.
- Phổ biến các kiến thức về Grid ra cộng đồng.

2.4.2 Định hướng cho sự phát triển Grid Middleware

Nhằm thuyết phục các công ty và tổ chức tài trợ cho sự phát triển của Grid middleware bằng cách định hướng xây dựng các middleware mới sẽ có nhiều chức năng hay dịch vụ mà các doanh nghiệp có thể sử dụng.

Irving Wladawsky-Berger: “all of our systems will be enabled to work with the Grid, and all of our middleware will integrate with the software”.

Tuy nhiên, công việc quan trọng nhất của middleware này là hỗ trợ các yêu cầu nảy sinh từ các dự án pilot như khả năng tính toán, không gian lưu trữ và phải đáp ứng được các chức năng của một middleware cần có. Middleware này sẽ được phát triển dưới dạng mã nguồn mở với sự cộng tác của cộng đồng Globus (www.globus.org) và có sự thảo luận với IBM, Oracle, Microsoft về các vấn đề đặc biệt là về DBMS.

2.4.3 Interdisciplinary Research Collaboration (IRC)

IRC được tài trợ bởi EPSRC, BBSRC, MRC and the MoD dưới sự cộng tác của các trường Đại học Cambridge, London và Bristol. Nguồn tài trợ của IRC được chia ra cho nhiều dự án trong nhiều lĩnh vực như: môi trường, vật lý, y tế,...

- Advanced Grid Interfaces for Environmental e-science in the Lab and in the Field
- CoAKTinG: Collaborative Advanced Knowledge Technologies in the grid.
- Grid enabled knowledge services: môi trường giải quyết các vấn đề cộng tác trong thông tin y tế.
- Grid-Based Medical Devices For Everyday Health
- MIAS - Grid. A Medical Image and Signal Research Grid

2.4.4 Hỗ trợ các dự án e-Science

Trung tâm hỗ trợ Grid ra đời nhằm mục đích hỗ trợ kỹ thuật cho các ứng dụng e-Science. Trung tâm này sẽ hỗ trợ kỹ thuật cho cũng như các dự án e-Science và “Grid Stater Kit”, đây là bộ công cụ bao gồm Globus Toolkit, Condor và Storage Resource Broker middleware. Nhóm hỗ trợ sẽ trả lời và giải quyết các vấn đề liên quan đến Grid trong phạm vi toàn UK. Ngoài ra nhóm hỗ trợ còn thực hiện công việc đánh giá, kiểm thử và nâng cấp các middleware của “Grid Stater Kit”.

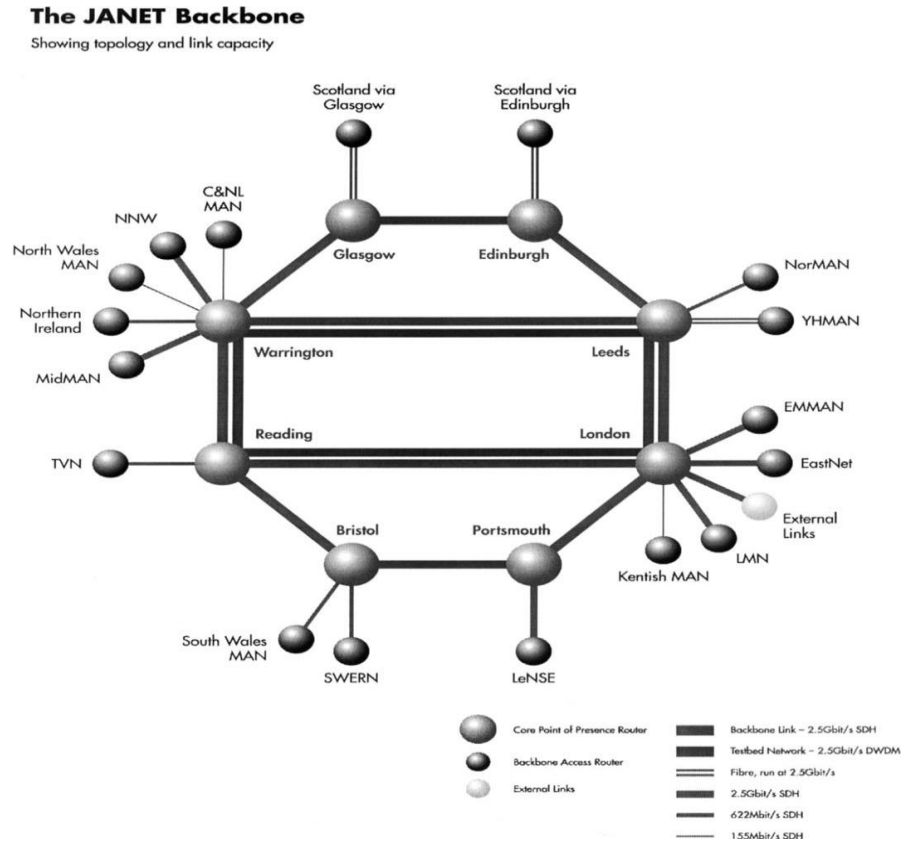
2.4.5 Sự hợp tác toàn cầu

e-Science phải hỗ trợ cộng đồng khoa học UK cộng tác với các nhà khoa học trên toàn thế giới, đây là mục tiêu quan trọng nhất mà hạ tầng cũng như các ứng dụng Grid phải đạt được. Vào thời điểm hiện tại thì e-Science đã tham gia cộng tác với Argonne National Laboratory, San Diego Supercomputing Center and NCSA in the USA. Những người phát triển e-Science luôn mong muốn có thêm nhiều sự cộng tác như thế trong tương lai.

2.4.6 Hạ tầng mạng

Các ứng dụng e-Science sẽ dựa vào SuperJANET4 để truyền tải dữ liệu. SuperJANET4 hiện tại có băng thông lên đến 10Gbps. Grid Network Team (GNT) ra đời nhằm để thu thập thông tin về nhu cầu sử dụng băng thông của các dự án e-

Science, giám sát các sự cố về mạng có thể xảy ra (bottleneck,..) , đảm bảo chất lượng dịch vụ cho các dự án e-Science. Ngoài ra việc cân bằng tải hay phân phối băng thông một cách hợp lý cũng được e-Science kết hợp với EU DataGrid project thực hiện.



Backbone network of e-Science

2.4.7 Demonstrator projects

Ngoài ra e-Science cũng thực hiện các dự án ngắn (short-term project) nhằm thể hiện khả năng của Grid trên nhiều lĩnh vực khác nhau. Các dự án này có ý nghĩa trong việc chứng minh khả năng của Grid cũng như tiềm năng của e-Science nhằm khuyến khích mở rộng các ứng dụng của e-Science.

CHƯƠNG III

VAI TRÒ CỦA GRID COMPUTING ĐỐI VỚI TÍNH TOÁN KHOA HỌC

3.1 Chia sẻ tài nguyên bên trong tổ chức ảo (Virtual Organization)

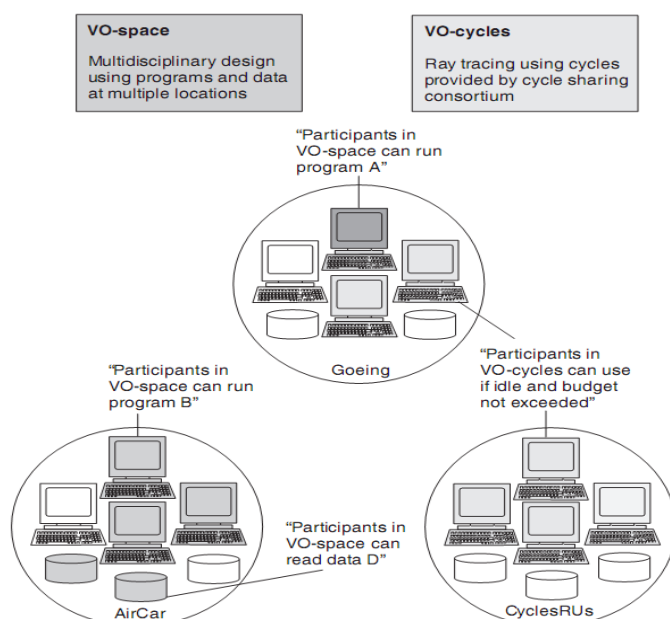
Một tổ chức ảo là một tập hợp các tài nguyên được quản lý một cách độc lập, những được hợp tác, tập hợp lại để cộng đồng dùng nó giải quyết một mục tiêu chung. Việc tổ chức và sử dụng tài nguyên một cách hiệu quả rất quan trọng đối với cộng đồng người sử dụng này.

Kịch bản sử dụng đơn giản và quan trọng nhất trong tính toán khoa học dựa trên Grid là khám phá tài nguyên (resource discovery) và nhờ hệ thống lưới thực thi một công việc nào đó (job submission). Một nhà khoa học ở Châu Âu đã chuẩn bị một file thực thi và các tham số đầu vào để giả lập việc trộn lẫn vào nhau giữa các lỗ đen, việc giả lập sử dụng nhiều tài nguyên phân bố ở khắp nơi trên thế giới, thông thường việc truy xuất tài nguyên ở mỗi nơi cần có một tài khoản, password, hàng đợi, hệ thống file... khác nhau. Điều này gây khó khăn rất lớn cho các nhà khoa học có thể tiến hành các thí nghiệm, các mô phỏng với quy mô lớn. Ngay cả việc đơn giản là chọn một nơi nào đó để thực thi quá trình mô phỏng cũng đã rất phức tạp.

Nhưng với công nghệ Grid quá trình truy xuất những tài nguyên này sẽ trở nên đơn giản. Việc xác nhận dựa trên certificate cho phép người sử dụng truy xuất tất cả hệ thống trong tổ chức ảo với chỉ một lần log-in, ID và password khi log-in vào tổ chức ảo sẽ được ánh xạ vào tài khoản của người sử dụng đó. Đồng thời mô hình Grid cũng sẽ loại bỏ đi những đặc tính riêng của từng tài nguyên, tạo ra interface chung cho hệ thống file, cách nén dữ liệu... và do đó cho phép người sử dụng dùng các câu lệnh truy xuất thống nhất cho tất cả các tài nguyên. Hiện tại Globus Toolkit đã phát triển tốt và vững chắc các quá trình trên.

Việc xây dựng những cổng thông tin dựa trên web (web-based portal) cho phép người dùng truy xuất dễ dàng và thân thiện tất cả tài nguyên trong một tổ chức ảo. Sau

khi log-in vào, người dùng về nguyên tắc có thể nhìn thấy được tất cả các tài nguyên sẵn có và có thể submit một công việc trên một máy nào.



Mối quan hệ chia sẻ bên trong tổ chức ảo

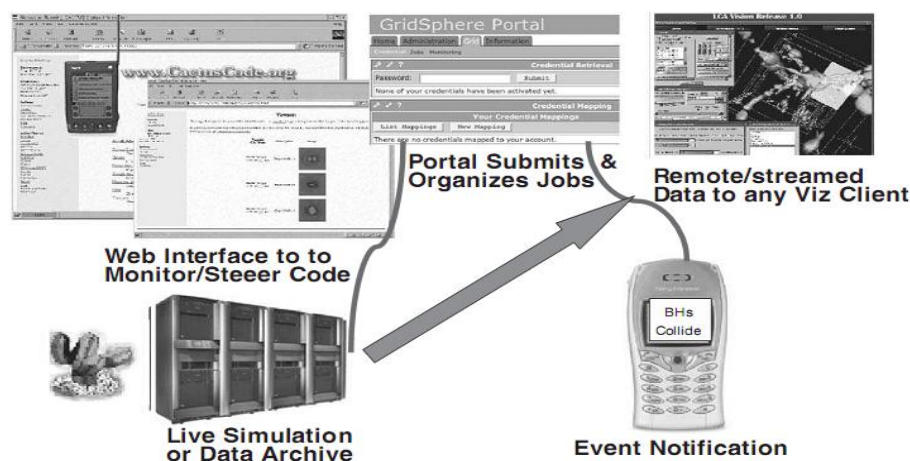
3.2 Giao tiếp với những công việc đang được thực thi trên hệ thống lưới

Trước đây, một nhà khoa học khi muốn chạy một ứng dụng mô phỏng lớn thì trước tiên phải thông báo và chuyển công việc vào một hệ thống trên một máy ở xa và sau đó cứ một khoảng thời gian lại log in vào máy từ xa đó để kiểm tra trạng thái của công việc. Công việc đó có thể vẫn còn trong hàng đợi hoặc có thể công việc đó đang được thực thi và cũng thường xuyên xảy ra trường hợp công việc đó đã bị kết thúc vì một lý do nào đó. Có những trường hợp công việc nằm trong hàng đợi cả ngày, nhưng khi được thực thi thì thất bại chỉ vì một lỗi đơn giản trong quá trình khởi động hay có những công việc đã chạy cả 100.000 giờ của CPU nhưng cuối cùng lại trả ra kết quả sai chỉ vì tham số đầu vào sai. Những trường hợp như thế đã làm nảy sinh nhu cầu cần phải phát triển cơ chế giao tiếp, quản lý và chuyển hướng những công việc tốt hơn, đòi hỏi độ phức tạp hơn.

Cactus là một ứng dụng Grid cho phép người dùng thông qua trình duyệt web có thể giao tiếp với các công việc đang thực thi. HTTPD Cactus thorn là một web server cho phép thể hiện tất cả thông tin của một công việc đang thực thi bao gồm chương

trình con đang tích cực, phiên bản, thời gian ước lượng hoàn tất, file dữ liệu nào được ghi. Các tham số được khai báo *steerable* có thể được thay đổi thông qua giao diện web, tần suất I/O, variable output, downsampling... cũng có thể được thay đổi nếu cần, cho phép người sử dụng sửa lại chương trình mà không phải bắt đầu lại. Khi một công việc bắt đầu, URL của nó được phát tán đến một công thông tin (portal), công thông tin này sẽ thông báo đến một nhóm người cùng hợp tác trong công việc này thông qua email hoặc tin nhắn SMS cùng với những thông tin để họ có thể truy xuất quá trình mô phỏng từ trình duyệt. Những sự kiện mô phỏng quan trọng, như việc trộn lẫn hai lỗ đen có thể được lập trình để kích hoạt thông báo đến một người sử dụng nào đó hoặc là một nhóm người cùng cộng tác trong việc giả lập đó.

Công thông tin vì vậy trở thành công cụ tổ chức cho khoa học tính toán hợp tác. Công việc có thể được phân loại theo trạng thái, theo chủ đề hoặc theo sự cộng tác. Đường liên kết (hyperlink) đến dữ liệu sinh ra từ quá trình giả lập. Khi nhưng đường liên kết này được click vào dữ liệu bên trong một file được download về từ máy ở xa và một chương trình client thích hợp sẽ đọc dữ liệu này và hiển thị trực quan cho người dùng. Nếu dữ liệu nằm trong bộ nhớ cấp phát cho quá trình mô phỏng thì thông qua socket chương trình client có thể hiển thị trực quan kết quả cho người dùng.



A user can submit jobs from a portal, and jobs are registered with the portal when they start, allowing access by all participating collaborators. Jobs can be monitored, and steered or their output data visualized, either live or from data archives. Important events in the life of a job trigger notification to various types of devices.

Công thông tin hỗ trợ dịch vụ tính toán lưới

Những công nghệ này đang được sử dụng rất nhiều trong các dự án khoa học. Tuy nhiên vẫn còn một số vấn đề cần giải quyết đối với những công nghệ này.

Thứ nhất vấn đề firewall, thường mâu thuẫn với nhu cầu người dùng muốn giao tiếp trực tiếp với dữ liệu từ xa hoặc mô phỏng từ bất kỳ vị trí nào.

Thứ hai những ứng dụng như trên cần được cải tiến để khai thác cơ chế Web và những dịch vụ lưới cho phép những ứng dụng có thể thông báo cho nhau, không chỉ thông báo cho cổng thông tin địa lý mà còn có thể thông báo cho bất kỳ ứng dụng hay server thông tin nào theo ý muốn, trao đổi dữ liệu, tiếp xúc với broker tài nguyên để tìm một tài nguyên mới, khởi động những ứng dụng khác, thông báo cho người sử dụng hoặc những ứng dụng khác khi một sự kiện nào đó xảy ra.

3.3 Tính toán phân bố

Chúng ta sẽ mô tả các phương thức khác nhau cho một ứng dụng phân bố thông qua 4 khái niệm liên quan đến việc sử dụng tài nguyên :

- task farming
- metacomputing
- migration
- spawning

Task farming là bước tiếp cận mà trong đó những task độc lập nhau hoặc là kết nối lỏng lẻo sẽ thực thi trên những tài nguyên rải rác trong tổ chức ảo. Những tasks này đòi hỏi rất ít hoặc thậm chí không có sự giao tiếp dữ liệu giữa các task và thường trả về dữ liệu rất ít. Task farming có thể được dùng cho việc nghiên cứu tìm ra các tham số phù hợp cho một bài toán, các tham số này thay đổi với rất nhiều khả năng khác nhau để tìm ra một giải pháp hứa hẹn nhất trước khi đi sâu vào nghiên cứu. Ví dụ việc nghiên cứu hiện tượng lực hấp dẫn hình thành nên lỗ đen, việc nghiên cứu này đòi hỏi phải biết một cách chính xác biên độ của sóng lúc ban đầu. Một thay đổi nhỏ trên giá trị biên độ này sẽ quyết định sóng bị hút bởi lỗ đen hay bị phân tán. Ngoài ra còn rất nhiều ví dụ khác trong khoa học cần tìm những giá trị có độ chính xác tương đối cao. Grid task farming tận dụng công nghệ lưới để khám phá những tài nguyên thích hợp trong một tổ chức ảo, thực thi một tập các task, nhận kết quả trả về, tất cả diễn ra trong một thời gian ngắn

Metacomputing là việc phân bố một hoặc nhiều task kết nối chặt đến một số lượng máy rất lớn. Metacomputing có thể được sử dụng để tăng không chỉ khả năng tính toán của hệ thống mà còn tăng tính sẵn có của hệ thống. Ví dụ quá trình tính toán

cần 1024 bộ xử lý nhưng không có một máy nào có sẵn chừng đó bộ xử lý, thì có thể dùng 4 máy với 256 bộ xử lý mỗi máy để tính toán. Sự khả thi của metacomputing đã được chứng minh với những ứng dụng thực tế. Ví dụ năm 2001, mô phỏng sự va chạm giữa các lỗ đen sử dụng nhiều máy từ xa, chạy trên những hệ điều hành khác nhau, sử dụng những kỹ thuật có khả năng thích nghi để có thể tự động điều chỉnh những message gửi thông qua môi trường mạng giữa các máy, làm tăng hiệu suất từ 15% đến trên 70% thời gian thực thi việc mô phỏng. Tầng trao đổi message dựa trên Grid cho phép những ứng dụng dựa vào Calculus có thể được chạy mà không cần sự điều chỉnh nào. Những thí nghiệm như thế cho thấy rằng metacomputing dựa trên Grid có thể được chạy với độ hiệu quả cao thậm chí cho những giả lập có tính kết nối cao như phương trình Einstein (đòi hỏi nhiều sự giao tiếp). Khi công nghệ Grid được triển khai đến nhiều phía khác nhau, những khả năng này có thể trở thành chế độ hoạt động bình thường của hệ thống.

Migration là việc di chuyển quá trình giả lập từ phía này sang phía khác hoặc có thể là di chuyển process mô phỏng từ máy này sang máy khác. Ví dụ do sự độn độ làm chậm quá trình mô phỏng hay quá trình mô phỏng cần nhiều bộ nhớ hơn để giải quyết một bài toán đang phát triển, những nhu cầu đó dẫn đến quá trình mô phỏng cần phải di chuyển đến một tài nguyên khác thích hợp hơn. Việc di chuyển này cần sự hỗ trợ của broker dịch vụ tài nguyên, nếu một tài nguyên mới được tìm thấy phù hợp với yêu cầu, broker dịch vụ sẽ thông báo, để quá trình mô phỏng được chuyển sang môi trường mới.

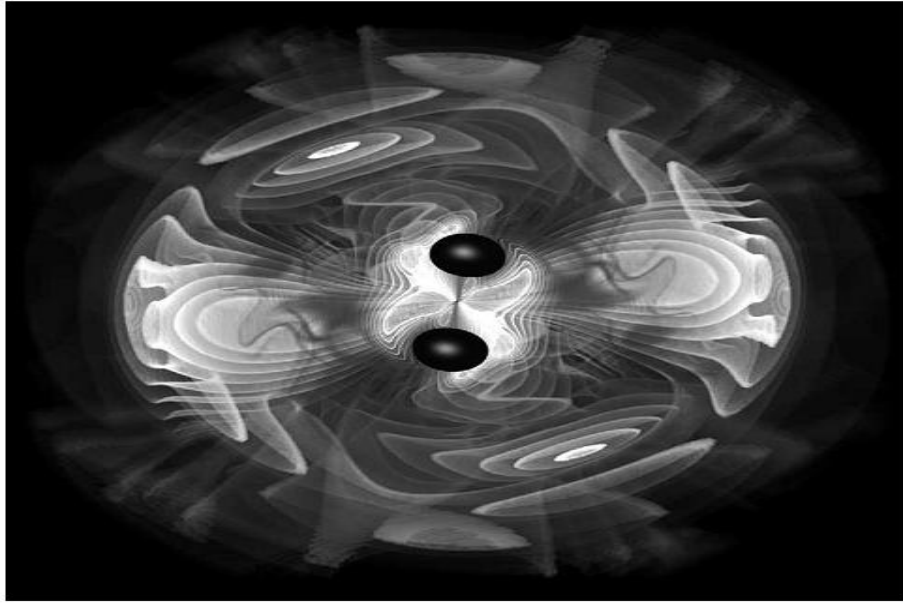
Spawning là một biến dạng của Migration. Trong Spawning, chỉ một phần ứng dụng được di chuyển đến một tài nguyên ở xa. Ví dụ khi mô phỏng sự độn độ của các lỗ đen, nhiệm vụ phân tích cần xác định vị trí của lỗ đen và tính toán lực hấp dẫn phát ra. Những nhiệm vụ này cần rất nhiều thời gian và có thể không cần phải trả kết quả về cho chương trình giả lập chính hoặc những nhiệm vụ này dễ dàng song song hóa và do đó có thể được di chuyển đến một tài nguyên khác, cho phép tài nguyên chính tập trung trong việc giải quyết vấn đề mô phỏng chính, trong mô phỏng lỗ đen năm 2001, việc mô phỏng được chạy ở Đức nhưng nhiệm vụ phân tích được di chuyển đến tài nguyên ở Châu Âu, Châu Á và Bắc Mỹ.

3.4 Quản lý dữ liệu

Một ví dụ cho thấy vấn đề quản lý dữ liệu là cần thiết và quan trọng thế nào đối với việc nghiên cứu khoa học là quá trình mô phỏng 3D trong khoa học. Việc mô phỏng 3D sinh ra một lượng lớn dữ liệu, dữ liệu này phải được phân tích, hiển thị trực quan và lưu trữ cho những lần sử dụng sau. Mỗi một lần giả lập có thể sinh ra hàng trăm file với những định dạng khác nhau, việc khám phá và thao tác trên những file này thì rất phức tạp bởi vì người sử dụng chạy trên những máy tính khác nhau với những hệ thống file, khả năng lưu trữ khác nhau.

Đối với ứng dụng lưới, vấn đề quản lý dữ liệu còn khó khăn hơn. Người dùng thậm chí còn không biết quá trình mô phỏng đang chạy trên máy nào hay sự mô phỏng có thể được di chuyển giữa các tài nguyên khác nhau và dữ liệu chứa ở nhiều nơi khác nhau. Thậm chí khi biết được vị trí của file dữ liệu thì cũng không thể di chuyển đến một máy local để thuận tiện cho việc xử lý vì kích thước quá lớn của nó. Mặc dù công nghệ Grid giúp ích rất nhiều cho nhóm các nhà khoa học trong việc thực thi, giám sát và điều chỉnh những mô phỏng này từ xa trong khi những mô phỏng này đang được chạy, nhưng một khối lượng dữ liệu khổng lồ được sinh ra từ quá trình mô phỏng không thể gửi tới từng thành phần tham gia cho việc phân tích, thay vì vậy một nhóm các nhà khoa học có thể phải bay từ Berlin sang Mỹ để nhìn thấy và lấy những kết quả này về.

GridFTP server là một công cụ cho việc thao tác dữ liệu từ xa đã được sử dụng bởi các nhà thiên văn học. GridFTP server là sự mở rộng của dự án Đức GriKSL. Khi GridFTP server chạy trên hệ thống file của máy đang có dữ liệu, nó cho phép file dữ liệu HDF5 được phân tích với hệ thống hiển thị trực quan ở local. Bất kỳ hệ thống hiển thị trực quan nào tích hợp GridFTP client cùng với chương trình đọc file HDF5 đều có thể hiển thị dữ liệu từ xa. Những chương trình đọc file HDF5 có sẵn trong các phần mềm như OpenDX, Amira. Những công cụ này đã giải quyết được vấn đề dữ liệu ở xa, và có thể được sử dụng để tạo cái nhìn trực quan của dữ liệu ở cách xa hàng ngàn km.



*Sử dụng GridFTP để có cái nhìn trực quan kết quả thu được
từ sự mô phỏng 2 lỗ đen đụng độ nhau*

Tài liệu tham khảo

- [1] eScience and Grid Computing, *Prof. Mike Giles, Oxford Centre for Computational Finance City Seminar, 12/02/03*
- [2] What is Grid Computing?, *Richard Hopkins, NGS Induction – Rutherford Appleton Laboratory, 2nd / 3rd November 2005*
- [3] What is e-Science and Grid computing? *Dave Berry, NeSC*
- [4] The Encyclopedia Wikipedia, <http://en.wikipedia.org/wiki/>
- [5] NSF office of Cyberinfrastructure, <http://www.nsf.gov/>
- [6] E-Science Grid, <http://www.escience-grid.org.uk/>
- [7] Distributed Aircraft Maintenance Environment , <http://www.cs.york.ac.uk/dame/>
- [8] eDiaMoND, <http://www.ediamond.ox.ac.uk/>
- [9] Grid Enabled Optimisation and Design Search for Engineering (GEODISE), <http://www.geodise.org>