Ebook Điện toán đám mây Phần 1 1187504

Điện toán đám mây (Trường Đại học Bách khoa Hà Nội)

HUỲNH QUYẾT THẮNG (Chủ biên)
NGUYỄN HỮU ĐỨC – DOÃN TRUNG TÙNG
NGUYỄN BÌNH MINH – TRẦN VIỆT TRUNG

ĐIỆN TOÁN ĐÁM MÂY

NHÀ XUẤT BẢN BÁCH KHOA HÀ NỘI

Điện toán đám mây
Nhiều tác giả

Table of Contents

MỤC LỤC

LỜI MỞ ĐẦU

Với sự phát triển bùng nổ hiện nay của công nghệ thông tin và ứng dụng trong đời sống, điện toán đám mây trở nên có tầm quan trọng thời sự. Giáo trình ***Điện toán đám mây*** được biên soạn cho đối tượng là học viên cao học các chuyên ngành Công nghệ thông tin. Sinh viên năm cuối của các trường đại học kỹ thuật cũng có thể sử dụng giáo trình như một tài liệu tham khảo để phát triển các ứng dụng cho nghiên cứu, cho đồ án tốt nghiệp.

Các tác giả hy vọng thông qua giáo trình sẽ cung cấp cho người đọc một tiếp cận tổng thể tới các khái niệm cơ bản về điện toán đám mây, các vấn đề về lưu trữ và xử lý dữ liệu, các vấn đề về an toàn và bảo mật, các dịch vụ, kiến trúc dịch vụ, hệ giám sát, một số chủ đề nâng cao gợi mở các vấn đề nghiên cứu hiện nay trong lĩnh vực điện toán đám mây.

Giáo trình là kết quả tổng hợp các nội dung nghiên cứu trong khuôn khổ đề tài tiến sỹ của các tác giả khi học tập tại nước ngoài, một số kết quả nghiên cứu khi triển khai đề tài khoa học công nghệ cấp Nhà nước: "Nghiên cứu làm chủ công nghệ dịch vụ đám mây (tạo lập và cung cấp dịch vụ, cung cấp nội dung số, quản lý truy cập)" mã số KC.01.01/11–15 và các kiến thức, kinh nghiệm qua nhiều năm giảng dạy tại Đại học Bách Khoa Hà Nội. Một số nội dung đã được giảng dạy thử nghiệm cho các khóa thạc sỹ 2012, 2013 của Viện Công nghệ Thông tin & Truyền thông và sau đó đã được chỉnh sửa để phù hợp với sự thay đổi công nghệ.

Giáo trình được xuất bản lần đầu nên không tránh khỏi những khiếm khuyết nhất định. Ngoài ra, do tính chất đặc thù phát triển nhanh chóng của lĩnh vực điện toán đám mây, nên nội dung giáo trình chưa hoàn toàn cập nhật, cô đọng, thiếu các diễn giải chi tiết, nhiều vấn đề chỉ nêu mà chưa minh họa. Chúng tôi mong nhận được nhiều ý kiến đóng góp cụ thể của các bạn độc giả để có thể sửa chữa, bổ sung và làm tốt hơn trong các lần xuất bản sau.

Tập thể tác giả xin bày tỏ sự cám ơn chân thành tới Bộ Khoa học và Công nghệ, Bộ Giáo dục và Đào tạo, Trường Đại học Bách Khoa Hà Nội đã tạo điều kiện để phát triển các nghiên cứu chuyên sâu. Chúng tôi cũng đặc biệt cám ơn các bạn đồng nghiệp ở Viện Công nghệ Thông tin & Truyền thông đã có những góp ý chân thành để giáo trình được hoàn thiện.

Mọi ý kiến đóng góp xin gửi về tập thể tác giả theo địa chỉ sau:

PGS. TS. Huỳnh Quyết Thắng, TS. Nguyễn Hữu Đức

Phòng 504, nhà B1, Viện Công nghệ Thông tin & Truyền thông, Trường Đại học Bách Khoa Hà Nội, số 1 Đại Cồ Việt, Hai Bà Trưng, Hà Nội.

Email: thang.huynhquyet@hust.edu.vn

duc.nguyenhuu@hust.edu.vn

tung.doantrung@hust.edu.vn

minh.nguyenbinh@hust.edu.vn

trung.tranviet@hust.edu.vn

Các tác giả

Chương mở đầu
TỔNG QUAN ĐIỆN TOÁN ĐÁM MÂY

A. LỊCH SỬ RA ĐỜI CỦA ĐIỆN TOÁN ĐÁM MÂY

Khái niệm điện toán đám mây ra đời từ những năm 1950 khi máy chủ tính toán quy mô lớn (large-scale mainframe computers) được triển khai tại một số cơ sở giáo dục và tập đoàn lớn. Tài nguyên tính toán của các hệ thống máy chủ được truy cập từ các máy khách cuối (thin clients, terminal computers), từ đó khai sinh khái niệm "chia sẻ thời gian" (time-sharing) đặc tả việc cho phép nhiều người sử dụng cùng chia sẻ đồng thời một tài nguyên tính toán chung.

Trong những năm 1960 – 1990, xuất hiện luồng tư tưởng coi máy tính hay tài nguyên công nghệ thông tin có thể được tổ chức như hạ tầng dịch vụ công cộng (public utility). Điện toán đám mây hiện tại cung cấp tài nguyên tính toán dưới dạng dịch vụ và tạo cảm giác cho người dùng về một nguồn cung ứng là vô tận. Đặc tính này có thể so sánh tới các đặc tính của ngành công nghiệp tiêu dùng dịch vụ công cộng như điện và nước. Khi sử dụng điện hay nước, người dùng không cần quan tâm tới tài nguyên đến từ đâu, được xử lý, phân phối như thế nào, họ chỉ việc sử dụng dịch vụ và trả tiền cho nhà cung cấp theo lượng tiêu dùng của mình.

Những năm 1990, các công ty viễn thông từ chỗ cung ứng kênh truyền dữ liệu điểm tới điểm (point-to-point data circuits) riêng biệt đã bắt đầu cung ứng các dịch vụ mạng riêng ảo với giá thấp. Thay đổi này tạo tiền đề để các công ty viễn thông sử dụng hạ tầng băng thông mạng hiệu quả hơn. Điện toán đám mây mở rộng khái niệm chia sẻ băng thông mạng này qua việc cho phép chia sẻ cả tài nguyên máy chủ vật lý bằng việc cung cấp các máy chủ ảo.

Amazon cung cấp nền tảng Amazon Web Services (AWS) vào năm 2006, đánh dấu việc thương mại hóa điện toán đám mây. Từ đầu năm 2008, Eucalyptus được giới thiệu là nền tảng điện toán đám mây mã nguồn mở đầu tiên, tương thích với API của AWS. Tính tới thời điểm hiện tại, có rất nhiều các sản phẩm điện toán đám mây được đưa ra như Google App Engine, Microsoft Azure, Nimbus,...

B. KHÁI NIỆM VỀ ĐIỆN TOÁN ĐÁM MÂY

Điện toán đám mây (cloud computing) là một xu hướng công nghệ nổi bật trên thế giới trong những năm gần đây và đã có những bước phát triển nhảy vọt cả về chất lượng, quy mô cung cấp và loại hình dịch vụ, với một loạt các nhà cung cấp nổi tiếng như Google, Amazon, Salesforce, Microsoft,...

Điện toán đám mây là mô hình điện toán mà mọi giải pháp liên quan đến công nghệ thông tin đều được cung cấp dưới dạng các dịch vụ qua mạng Internet, giải phóng người sử dụng khỏi việc phải đầu tư nhân lực, công nghệ và hạ tầng để triển khai hệ thống. Từ đó điện toán đám mây giúp tối giản chi phí và thời gian triển khai, tạo điều kiện cho người sử dụng nền tảng điện toán đám mây tập trung được tối đa nguồn lực vào công việc chuyên môn.

Lợi ích của điện toán đám mây mang lại không chỉ gói gọn trong phạm vi người sử dụng nền tảng điện toán đám mây mà còn từ phía các nhà cung cấp dịch vụ điện toán. Theo những đánh giá của nhóm IBM CloudBurst năm 2009, trên môi trường điện toán phân tán có đến 85% tổng năng lực tính toán trong trạng thái nhàn rỗi, thiết bị lưu trữ tăng 54% mỗi năm, khoảng 70% chi phí được dành cho việc duy trì các hệ thống thông tin. Công nghiệp phần mềm mất đi 40 tỷ USD hằng năm vì việc phân phối sản phẩm không hiệu quả, khoảng 33% khách hàng

phàn nàn về các lỗi bảo mật do các công ty cung cấp dịch vụ. Những thống kê này đều chỉ đến một điểm quan trọng: mô hình hệ thống thông tin hiện tại đã lỗi thời và kém hiệu quả, cần phải chuyển sang một mô hình điện toán mới – đó là điện toán đám mây.

Theo định nghĩa của Viện Quốc gia Tiêu chuẩn và Công nghệ Mỹ (US NIST), điện toán đám mây là mô hình cho phép truy cập trên mạng tới các tài nguyên được chia sẻ (ví dụ: hệ thống mạng, máy chủ, thiết bị lưu trữ, ứng dụng và các dịch vụ) một cách thuận tiện và theo nhu cầu sử dụng. Những tài nguyên này có thể được cung cấp một cách nhanh chóng hoặc thu hồi với chi phí quản lý tối thiểu hoặc tương tác tối thiểu với nhà cung cấp dịch vụ.

C. CÁC ĐẶC TÍNH CỦA ĐIỆN TOÁN ĐÁM MÂY

Định nghĩa của US NIST chứa đựng kiến trúc, an ninh và chiến lược triển khai của đám mây. Năm đặc tính cốt lõi của điện toán đám mây được thể hiện rõ như sau:

– Tự phục vụ theo yêu cầu (on-demand self-service): Khách hàng với nhu cầu tức thời tại những thời điểm thời gian xác định có thể sử dụng các tài nguyên tính toán (như thời gian CPU, không gian lưu trữ mạng, sử dụng phần mềm,...) một cách tự động, không cần tương tác với con người để cấp phát.

– Sự truy cập mạng rộng rãi (broad network access): Những tài nguyên tính toán này được phân phối qua mạng Internet và được các ứng dụng client khác nhau sử dụng với những nền tảng không đồng nhất (như máy tính, điện thoại di động, PDA).

– Tập trung tài nguyên: Những tài nguyên tính toán của nhà cung cấp dịch vụ đám mây được tập trung với mục đích phục vụ đa khách hàng sử dụng mô hình ảo hóa với những tài nguyên vật lý và tài nguyên ảo được cấp phát động theo yêu cầu. Động lực của việc xây dựng một mô hình tập trung tài nguyên tính toán nằm trong hai yếu tố quan trọng: tính quy mô và tính chuyên biệt. Kết quả của mô hình tập trung tài nguyên là những tài nguyên vật lý trở nên trong suốt với người sử dụng. Ví dụ, người sử dụng không được biết vị trí lưu trữ cơ sở dữ liệu của họ trong đám mây.

– Tính mềm dẻo: Đối với người sử dụng, các tài nguyên tính toán được cung cấp tức thời hơn là liên tục, được cung cấp theo nhu cầu để mở rộng hoặc tiết giảm không hạn định tại bất kỳ thời điểm nào.

– Khả năng đo lường: Mặc dù tài nguyên được tập trung và có thể chia sẻ cho nhiều người sử dụng, hạ tầng của đám mây có thể dùng những cơ chế đo lường thích hợp để đo việc sử dụng những tài nguyên đó cho từng cá nhân.

D. SƠ LƯỢC CÁC CÔNG NGHỆ ỨNG DỤNG TRONG ĐIỆN TOÁN ĐÁM MÂY

Công nghệ ảo hoá

Công nghệ ảo hóa (virtualization) là công nghệ quan trọng nhất ứng dụng trong điện toán đám mây. Công nghệ ảo hóa là công nghệ cho phép tạo ra các thực thể ảo có tính năng tương đương như các thực thể vật lý, ví dụ như thiết bị lưu trữ, bộ vi xử lý,... Ảo hóa phần cứng (hardware virtualization) tham chiếu tới việc tạo ra các máy ảo (virtual machine) mà hoạt động với hệ điều hành được cài đặt như một máy tính vật lý thực. Ví dụ, một máy ảo chạy hệ điều hành Ubuntu có thể được tạo ra trên một máy tính thực cài hệ điều hành Windows.

Ảo hoà phần cứng cho phép chia nhỏ tài nguyên vật lý để tối ưu hóa hiệu năng sử dụng.

Điều này được thể hiện qua việc có thể khởi tạo nhiều máy ảo với năng lực tính toán và năng lực lưu trữ bé hơn trên duy nhất một máy chủ vật lý. Máy chủ vật lý được gọi là host machine còn máy ảo (virtual machine) được gọi là máy khách (guest machine). Khái niệm "host" và "guest" được sử dụng để phân biệt phần mềm chạy trên máy tính vật lý hay phần mềm chạy trên máy ảo. Phần mềm hay firmware tạo máy ảo được gọi là hypervisor hay virtual machine manager.

Công nghệ tự động hóa giám sát điều phối tài nguyên (automation, dynamic dynamic orchestration)

Công nghệ giám sát điều phối tài nguyên động là nền tảng để điện toán đám mây thực hiện cam kết chất lượng cung cấp dịch vụ điện toán. Với công nghệ điều phối tài nguyên động, việc lắp đặt thêm hay giảm bớt các tài nguyên máy chủ vật lý hoặc máy chủ lưu trữ dữ liệu được thực hiện tự động để hệ thống điện toán luôn đáp ứng được giao kèo trong hợp đồng dịch vụ đã ký với bên người sử dụng.

Công nghệ tính toán phân tán, hệ phân tán

Điện toán đám mây là một dạng hệ phân tán xuất phát từ yêu cầu cung ứng dịch vụ cho lượng người sử dụng khổng lồ. Tài nguyên tính toán của điện toán đám mây là tổng thể kết hợp của hạ tầng mạng và hàng nghìn máy chủ vật lý phân tán trên một hay nhiều trung tâm dữ liệu số (data centers).

Công nghệ Web 2.0

Web 2.0 là nền tảng công nghệ phát triển các sản phẩm ứng dụng hướng dịch vụ trên nền điện toán đám mây. Công nghệ Web 2.0 phát triển cho phép phát triển giao diện ứng dụng web dễ dàng và nhanh chóng và trên nhiều thiết bị giao diện khác nhau. Web 2.0 phát triển làm xóa đi khoảng cách về thiết kế giao diện giữa ứng dụng máy tính thông thường và ứng dụng trên nền web, cho phép chuyển hóa ứng dụng qua dịch vụ trên nền điện toán đám mây mà không ảnh hưởng đến thói quen người sử dụng.

E. ƯU NHƯỢC ĐIỂM CỦA ĐIỆN TOÁN ĐÁM MÂY

Ưu điểm của điện toán đám mây

Triển khai nhanh chóng: So với phương pháp thông thường triển khai một ứng dụng trên internet, người dùng phải thực hiện một loạt các công việc như mua sắm thiết bị (hoặc thuê thiết bị từ bên thứ ba), cài đặt và cấu hình phần mềm, đưa các ứng dụng vào đám mây, việc sử dụng điện toán đám mây giúp loại bỏ một số công việc đòi hỏi thời gian lớn, ví dụ người dùng chỉ việc quan tâm phát triển triển khai các ứng dụng của mình lên “mây” (internet) khi sử dụng các đám mây nền tảng. Bên cạnh đó, khả năng tăng hoặc giảm sự cung cấp tài nguyên nhanh chóng theo nhu cầu tiêu dùng của ứng dụng tại các thời điểm khác nhau nhờ công nghệ ảo hóa của điện toán đám mây cũng là một trong những đặc điểm vượt trội của công nghệ này, thể hiện khả năng triển khai nhanh đáp ứng đòi hỏi tài nguyên tức thời của ứng dụng.

Giảm chi phí: Chi phí được giảm đáng kể do chi phí vốn đầu tư được chuyển sang chi phí duy trì hoạt động. Điều này làm giảm những khó khăn khi người dùng cần tính toán xử lý các tác vụ trong một lần duy nhất hoặc không thường xuyên do họ có thể đi thuê cơ sở hạ tầng được cung cấp bởi bên thứ ba.

Đa phương tiện truy cập: Sự độc lập giữa thiết bị và vị trí làm cho người dùng có thể truy cập hệ thống bằng cách sử dụng trình duyệt web mà không quan tâm đến vị trí của họ hay thiết bị nào mà họ đang dùng, ví dụ như PC, mobile. Vì [illegible] off-site (được cung cấp bởi đối tác thứ ba) và được truy cập thông qua Internet, do đó người dùng có thể kết nối từ bất

kỳ nơi nào.

Chia sẻ: Việc cho thuê và chia sẻ tài nguyên giữa các người dùng với nhau làm giảm chi phí đầu tư hạ tầng tính toán giữa một phạm vi lớn người dùng. Sự chia sẻ này cũng cho phép tập trung cơ sở hạ tầng để phục vụ các bài toán lớn với chi phí thấp hơn việc đầu tư hệ thống máy chủ tính toán từ đầu.

Khả năng chịu tải nâng cao: Về lý thuyết, tài nguyên tính toán trên đám mây là vô hạn. Việc thêm vào năng lực tính toán để chịu tải cao có thể được thực hiện chỉ bằng các thao tác kích chuột hoặc đã được tự động hoá.

Độ tin cậy: Người sử dụng điện toán đám mây được ký hợp đồng sử dụng với điều khoản chất lượng dịch vụ rất cao ghi sẵn trong hợp đồng. Chất lượng dịch vụ đám mây đơn giản được đánh giá ổn định hơn hệ thống tự triển khai do nền tảng đám mây được thiết kế và bảo trì bởi đội ngũ chuyên gia nhiều kinh nghiệm về hệ thống. Hơn nữa, việc luôn làm việc với hệ thống lớn và gặp nhiều lỗi tương tự nhau nên quá trình khôi phục hệ thống sau thảm họa thông thường là nhanh chóng.

Tính co giãn linh động: Tính co giãn thể hiện sự linh động trong việc cung cấp tài nguyên tính toán theo nhu cầu thực tế của người dùng hoặc các ứng dụng dịch vụ. Theo đó tài nguyên sẽ được đáp ứng một cách tự động sát với nhu cầu tại thời gian thực mà không cần người dùng phải có kỹ năng cho quá trình điều khiển này.

Bảo mật: Tính bảo mật trong điện toán đám mây từ trước đến nay vẫn là câu hỏi lớn cho người dùng tiềm năng. Tuy nhiên, hiện nay, khả năng bảo mật trong môi trường đám mây đã được cải thiện đáng kể, nhờ vào một số lý do chính sau đây: do dữ liệu tập trung trong các đám mây ngày càng lớn nên các nhà cung cấp luôn chú trọng nâng cao công nghệ và đặt ra những rào cản để tăng tính an toàn cho dữ liệu. Bên cạnh đó, các nhà cung cấp đám mây có khả năng dành nhiều nguồn lực cho việc giải quyết các vấn đề bảo mật mà nhiều khách hàng không có đủ chi phí để thực hiện. Các nhà cung cấp sẽ ghi nhớ các nhật ký truy cập, nhưng việc truy cập vào chính bản thân các nhật ký truy cập này có thể cũng rất khó khăn do chính sách của nhà cung cấp đám mây khi người dùng tự mình muốn xác minh rõ hệ thống của mình có an toàn không. Mặc dù vậy, mối quan tâm lo ngại về việc mất quyền điều khiển dữ liệu nhạy cảm cũng ngày càng tăng cao.

Nhược điểm của điện toán đám mây

Chi phí: Giảm chi phí đầu tư ban đầu là ưu điểm của điện toán đám mây. Tuy nhiên, nó cũng là một vấn đề phải tranh cãi khi người sử dụng điện toán đám mây luôn phải duy trì trả phí sử dụng dịch vụ. So với tự chủ đầu tư hạ tầng, người sử dụng điện toán đám mây không có tài sản sau khấu hao chi phí đầu tư.

Các công cụ giám sát và quản lý: Công cụ giám sát và bảo trì chưa hoàn thiện và khả năng giao tiếp với các đám mây là có giới hạn, mặc dù thông báo gần đây của BMC, CA, Novell cho rằng các ứng dụng quản lý trung tâm dữ liệu đang được cải tiến để cung cấp kiểm soát tốt hơn dữ liệu trong điện toán đám mây Amazon EC2 và các dịch vụ đám mây.

Chuẩn hóa đám mây: Chuẩn hóa giao tiếp và thiết kế đám mây chưa được thông qua. Mỗi nền tảng cung cấp các giao diện quản lý và giao tiếp ứng dụng API khác nhau. Hiện nay, các tổ chức như Distributed Management Task Force, Cloud Security Alliance và Open Cloud Consortium đang phát triển các tiêu chuẩn về quản lý tương thích, di chuyển dữ liệu, an ninh và các chức năng khác của điện toán đám mây.

Tính sẵn sàng: Tính sẵn sàng là ưu điểm của đám mây trong lý thuyết. Tuy nhiên, trên thực tế với các đám mây hiện thời, tính sẵn sàng đôi khi không được đảm bảo và cũng là một trở

ngại hiện nay, khi chỉ có một số ít nhà cung cấp dịch vụ cam kết được về sự sẵn sàng và liên tục của dịch vụ, về thời gian sửa chữa và phục hồi dữ liệu.

Vấn đề tuân thủ hợp đồng cũng trở nên phức tạp: Những nhà cung cấp dịch vụ điện toán đám mây có thể chuyển dữ liệu tới quốc gia khác có giá điện rẻ hơn, nhưng luật lỏng lẻo hơn mà người sử dụng dịch vụ điện toán không được thông tin. Điều này hoàn toàn có thể vì đám mây là trong suốt với người dùng.

Tính riêng tư: Hầu hết các hợp đồng thể hiện giao kèo giữa nhà cung cấp và người dùng điện toán đám mây hứa hẹn một viễn cảnh trong đó dữ liệu khách hàng luôn an toàn và riêng tư. Tuy nhiên, tính riêng tư trong điện toán đám mây cũng là một vấn đề đáng quan tâm vì hạ tầng an toàn thông tin cho đám mây hiện vẫn đang là một chủ đề nghiên cứu trong giới khoa học.

Cấp độ dịch vụ: Điện toán đám mây cung cấp dịch vụ theo yêu cầu, tuy nhiên trong thực tế, các gói dịch vụ thường được định nghĩa trước và người sử dụng căn cứ vào nhu cầu và khả năng để chọn dịch vụ sẵn có. Ví dụ, việc tự cấu hình chi tiết thông số các máy ảo hiện tại chưa thực hiện được. Như vậy, khả năng để thích ứng yêu cầu cấp dịch vụ cho các nhu cầu cụ thể của một doanh nghiệp là ít hơn so với các trung tâm dữ liệu xây dựng riêng với mục đích là để tiếp tục mục tiêu nâng cao khả năng kinh doanh của công ty.

Khả năng tích hợp với hạ tầng thông tin sẵn có của tổ chức: Việc tích hợp điện toán đám mây vào hạ tầng sẵn có của khách hàng chưa có mô hình và cách thức thực hiện cụ thể. Các mô hình kết nối đám mây riêng và đám mây thương mại vẫn đang được nghiên cứu.

F. GIỚI THIỆU MỘT SỐ ĐÁM MÂY ĐƯỢC SỬ DỤNG/TRIỂN KHAI PHỔ BIẾN HIỆN NAY

Microsoft Azure: Microsoft Azure là đám mây cung cấp hạ tầng và nền tảng điện toán xây dựng bởi Microsoft và đưa vào khai thác từ 2010. Về mặt hạ tầng, Azure cung cấp các máy chủ ảo có thể chạy hệ điều hành Windows hoặc Unix. Về mặt nền tảng điện toán, Azure hỗ trợ đa ngôn ngữ lập trình cho phép triển khai trên Azure nhiều ứng dụng phát triển trên các công cụ và framework khác nhau. Phổ biến là các ứng dụng viết trên nền .Net của Microsoft.

Amazon Web Service (AWS): AWS được đưa ra vào năm 2006 với khởi đầu là tập lợp các dịch vụ tính toán như dịch vụ máy ảo EC2 và dịch vụ lưu trữ S3. AWS là nền tảng đám mây thương mại đầu tiên và phổ biến nhất hiện nay. Nhiều khách hàng lớn sử dụng AWS có thể nói đến như NASA, Pinterest, Netflix.

Nimbus: Nimbus là đám mây mã nguồn mở cung cấp hạ tầng máy chủ hướng theo dịch vụ thông qua giao diện kết nối dựa trên chuẩn kết nối của AWS.

Google App Engine (GAE): GAE là nền tảng đám mây mà trên đó Google cung cấp hỗ trợ cho cơ sở dữ liệu, các thư viện lập trình và các môi trường thực thi cho các ngôn ngữ lập trình phổ biến. Các ứng dụng trên các ngôn ngữ hỗ trợ sau khi triển khai trên GAE sẽ chạy trên các máy chủ ảo của Google.

G. NỘI DUNG CƠ BẢN CỦA GIÁO TRÌNH

Giáo trình giới thiệu tổng quan về công nghệ điện toán đám mây, cung cấp những kiến thức cơ bản về điện toán đám mây bao gồm các công nghệ được áp dụng và thành phần thiết kế của nó. Các chủ đề liên quan đến công nghệ nền tảng của điện toán đám mây bao gồm: các trung tâm dữ liệu lớn, công nghệ ảo hóa, các giao thức quản lý và điều khiển các dịch vụ đám mây, cơ sở dữ liệu và lưu trữ trong môi trường đám mây, bảo mật và an toàn khi sử dụng ứng dụng

đám mây, các phương pháp đảm bảo chất lượng dịch vụ, hợp đồng dịch vụ. Các chủ đề liên quan đến việc triển khai và phát triển các dịch vụ đám mây bao gồm các kiến thức về dịch vụ web, các môi trường lập trình trong đám mây, cấu trúc và thiết kế các dịch vụ đám mây.

Trong Chương mở đầu, giáo trình đề cập đến lịch sử ra đời của điện toán đám mây, từ đó đưa ra khái niệm, định nghĩa về nền tảng đám mây. Chương mở đầu tập trung nêu rõ các đặc tính và ưu nhược điểm của nền tảng.

Trong Chương 1, các dạng điện toán đám mây sẽ được phân tích trên cơ sở hiểu rõ trung tâm dữ liệu lớn và công nghệ ảo hoá. Các mô hình mô phỏng môi trường đám mây cũng sẽ được giới thiệu.

Chương 2 giới thiệu một trong những dịch vụ điện toán đám mây cốt lõi là dịch vụ lưu trữ và xử lý dữ liệu. Chương 2 bắt đầu với tổng quan về các hệ quản lý tập tin phân tán, đi sâu vào phân tích hệ thống tập tin HDFS và mô hình tính toán MapReduce. Các cơ sở dữ liệu không quan hệ NoSQL cũng sẽ được đề cập đến.

Chương 3 phân tích khía cạnh bảo mật của điện toán đám mây. Chương gồm ba phần lớn: các thách thức về bảo mật và an toàn dịch vụ đám mây, các phương pháp bảo đảm an toàn cho dịch vụ đám mây và giải pháp thiết kế cấu trúc của đám mây an toàn.

Trong Chương 4, cách thức xây dựng dịch vụ trên nền tảng đám mây điển hình Azure sẽ được giới thiệu cụ thể.

Chương 5 phân tích khía cạnh giám sát và đảm bảo chất lượng các dịch vụ đám mây.

Chương 6 đưa ra các chủ đề nâng cao như đám mây liên bang, mô hình môi giới dịch vụ,...

Chương 1
NỀN TẢNG VÀ PHÂN LOẠI

Trong chương này, chúng ta sẽ làm quen với các khái niệm khái quát chung của nền tảng và phân loại điện toán đám mây. Người đọc sẽ được trang bị các kiến thức về khái niệm và mô hình của Trung tâm dữ liệu, trên cơ sở đó tiếp thu các kiến thức về công nghệ ảo hóa. Một phần nội dung của chương trình bày phân loại các mô hình điện toán đám mây và các kiến trúc, nền tảng điện toán đám mây đang được cung cấp trên thị trường hiện nay. Một số công cụ và giải pháp mô phỏng điện toán đám mây tại các trung tâm nghiên cứu cũng được giới thiệu để người đọc có thể tiếp cận các công cụ này.

1.1. TRUNG TÂM DỮ LIỆU LỚN

Khái niệm Trung tâm dữ liệu

Về mặt kỹ thuật, trung tâm dữ liệu (TTDL) hay data center có nguồn gốc từ các phòng máy tính lớn (main frames) thời sơ khai của ngành công nghiệp máy tính trong những năm 1960. Tại thời kỳ này, các hệ thống máy tính đầu tiên được coi là những thành tựu lớn lao của công nghệ tính toán tự động dựa trên kỹ thuật điện tử (electronic computing). Các hệ thống này là những hệ thống có những yêu cầu cực kỳ phức tạp trong vận hành, bảo trì, và môi trường hoạt động đặc biệt về nguồn điện, điều hòa không khí.

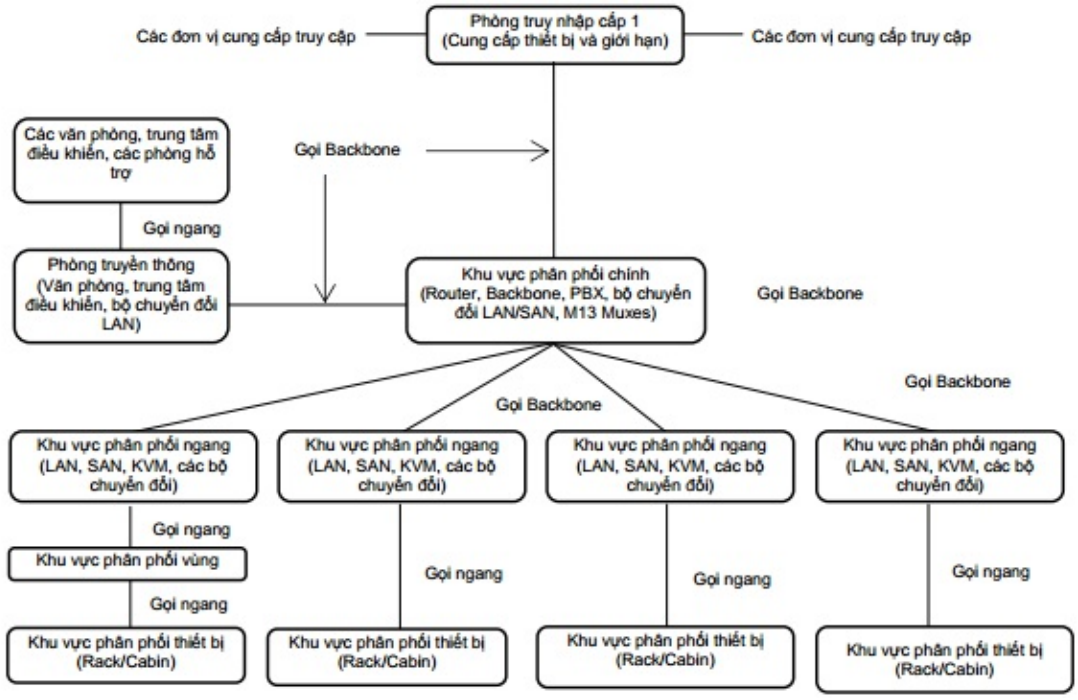
Với sự ra đời của mô hình tính toán "khách – chủ" trong những năm 1990, các máy chủ (server) đã thay thế dần các máy tính lớn trong các phòng máy tính tại các trường đại học và cơ sở nghiên cứu khoa học. Với sự giảm giá mạnh của phần cứng server, các thiết bị kết nối mạng và sự chuẩn hóa các hệ thống cáp mạng, các hệ thống tính toán "client-server" đã dần dần có chỗ đứng riêng trong môi trường của các tổ chức và doanh nghiệp để phục vụ công tác quản lý và kinh doanh. Thuật ngữ TTDL dùng để chỉ các phòng máy tính được thiết kế riêng bên trong tổ chức, doanh nghiệp với những yêu cầu đặc biệt về nguồn điện, điều hòa không khí, cấu trúc liên kết thiết bị và mạng,… bắt đầu đạt được sự công nhận phổ biến trong khoảng thời gian này.

TTDL là giải pháp hoàn chỉnh về một trung tâm điều phối hoạt động, trung tâm lưu trữ, nó có thể cung cấp các ứng dụng cho một tổ chức doanh nghiệp hay phục vụ cho hàng ngàn người cần truy cập, trao đổi thông tin. Mọi hoạt động của TTDL đều có ảnh hưởng rất lớn đến việc kinh doanh, tài chính, cũng như sự sống còn của một doanh nghiệp như các ngân hàng, công ty tài chính, sàn chứng khoán, công ty bảo hiểm,…

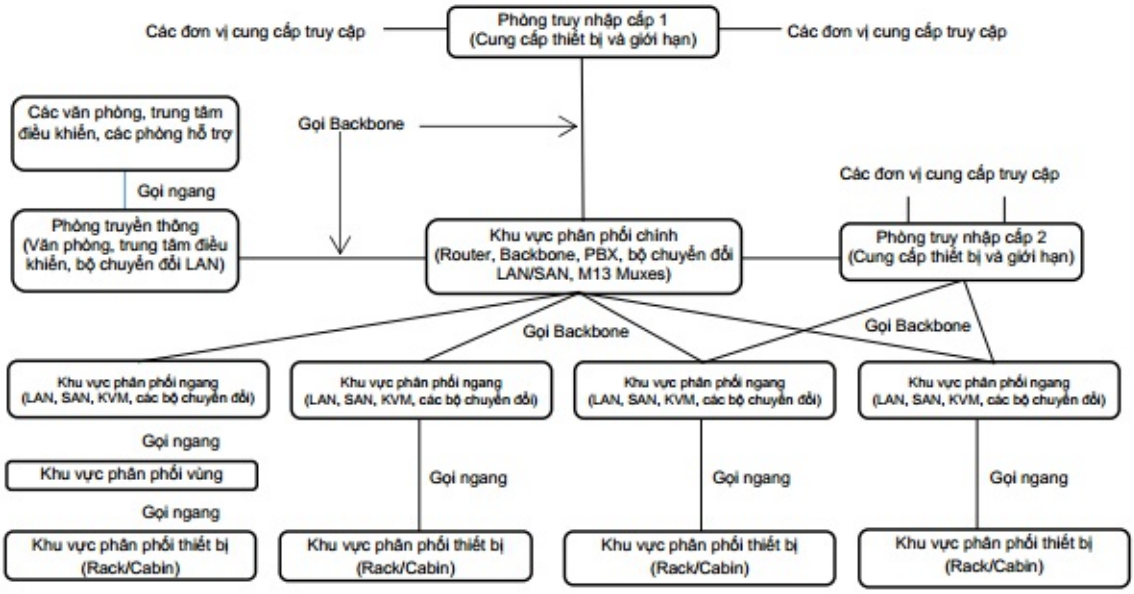
TTDL là một hệ thống máy tính cực kỳ quan trọng và rất dễ bị tổn hại, trong một TTDL chứa một lượng máy chủ, thiết bị lưu trữ rất lớn. Để hoạt động tốt, TTDL cần phải có những hệ thống phụ trợ như nguồn điện, hệ thống làm mát, báo cháy, an ninh bảo mật,..., cho dù kích cỡ như thế nào thì các TTDL vẫn có chung những chức năng là xử lý và lưu trữ dữ liệu. TTDL có thể là một phòng của một tầng lầu, hay một tầng của một toà nhà, hoặc được xây một toà nhà riêng.

Trung tâm dữ liệu đã trải qua nhiều bước phát triển về công nghệ và cấu trúc, cho đến nay tiếp tục phát triển dựa trên điện toán đám mây là xu hướng mới nhất nhằm hiện đại hóa trung tâm dữ liệu, tăng cường hiệu năng tính toán, nâng cao hiệu quả sử dụng năng lượng của thiết bị, giảm chi phí đầu tư và vận hành hạ tầng cho khách hàng.

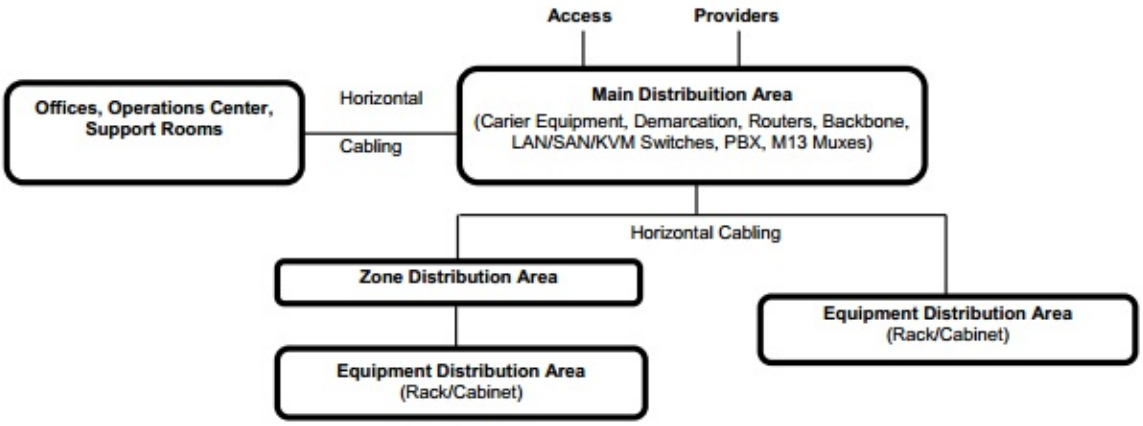
Các mô hình của Trung tâm dữ liệu



Hình 1.1. Mô hình Trung tâm dữ liệu cơ bản



Hình 1.2. Trung tâm dữ liệu có nhiều điểm kết nối đường vào



Hình 1.3. Mô hình trung tâm dữ liệu đơn giản

1.2. CÔNG NGHỆ ẢO HÓA

Khái niệm

Ảo hóa (Virtualization) là công nghệ tiên tiến nhất trong một loạt các cuộc cách mạng công nghệ nhằm tăng mức độ ảo hóa hệ thống, cho phép tăng hiệu suất làm việc của máy tính lên một cấp độ chưa từng có.

Ảo hóa hệ thống tức là tiến hành phân chia một máy chủ thành nhiều máy chủ ảo hoặc kết hợp nhiều máy chủ vật lý thành một máy chủ logic, đối với người sử dụng họ nhận biết và sử dụng các server ảo giống như một máy vật lý độc lập có đủ các tài nguyên cần thiết (bộ vi xử lý, bộ nhớ, kết nối mạng,…), trong khi các server ảo không hề có những tài nguyên độc lập như vậy, nó chỉ sử dụng tài nguyên được gán từ máy chủ vật lý. Ở đây, bản chất thứ nhất là các server ảo sử dụng tài nguyên của máy chủ vật lý, bản chất thứ hai là các server ảo có thể hoạt động như một server vật lý độc lập.

Lợi ích của ảo hóa

Thông thường việc đầu tư cho một trung tâm công nghệ thông tin rất tốn kém. Chi phí mua các máy chủ cấu hình mạnh và các phần mềm bản quyền là rất đắt. Doanh nghiệp luôn luôn muốn cắt giảm và hạn chế tối đa các chi phí không cần thiết trong khi vẫn đáp ứng được năng suất và tính ổn định của hệ thống. Vậy nên việc ứng dụng ảo hóa trở thành nhu cầu cần thiết của bất kỳ doanh nghiệp nào dù lớn hay nhỏ. Thay vì mua 10 máy chủ cho 10 ứng dụng thì chỉ cần mua 1 hoặc 2 máy chủ có hỗ trợ ảo hóa cũng vẫn có thể chạy tốt 10 ứng dụng này. Điều này cho thấy sự khác biệt giữa hệ thống ảo hóa và không ảo hóa. Bên cạnh đó, việc ứng dụng ảo hóa còn đem lại những lợi ích sau:

– Quản lý đơn giản;

– Triển khai nhanh;

– Phục hồi và lưu trữ hệ thống nhanh;

– Cân bằng tải và phân phối tài nguyên linh hoạt;

– Tiết kiệm chi phí;

– Ảo hóa góp phần tăng cường tính liên tục, hạn chế ngắt quãng.

Kiến trúc ảo hóa

Xét về kiến trúc hệ thống, các kiến trúc ảo hóa hệ thống máy chủ có thể ở các dạng chính là: Host–based, Hypervisor–based (còn gọi là bare mental hypervisor, được chia nhỏ làm hai loại là Monothic hypervisor và Microkernel hypervisor), Hybrid. Ngoài ra, tùy theo sản phẩm ảo hóa được triển khai (như VMWare, Microsoft HyperV, Citrix XEN Server) và mức độ ảo hóa cụ thể sẽ khác nhau.

Kiến trúc ảo hóa Hosted–based

Còn gọi là hosted hypervisor, kiến trúc này sử dụng một lớp hypervisor chạy trên nền tảng hệ điều hành, sử dụng các dịch vụ được hệ điều hành cung cấp để phân chia tài nguyên tới các máy ảo. Nếu ta xem hypervisor là một phần mềm riêng biệt, thì các hệ điều hành khách của máy ảo sẽ nằm trên lớp thứ ba so với phần cứng máy chủ.

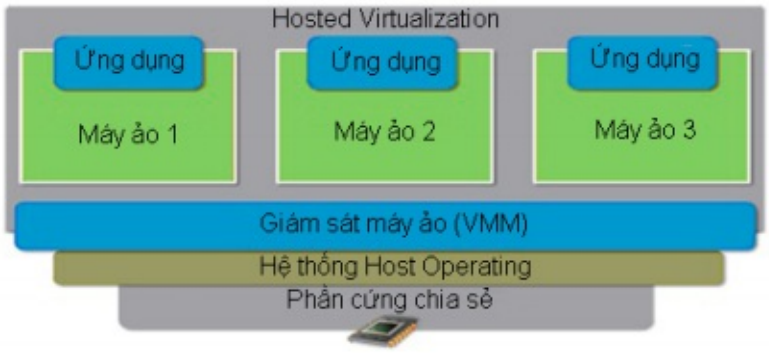
Nhìn vào hình 1.4, ta có thể thấy mô hình này được chia làm bốn lớp hoạt động như sau:

– Nền tảng phần cứng;

– Hệ điều hành Host;

– Hệ thống virtual machine monitor (hypervisor);

– Các ứng dụng máy ảo: sử dụng tài nguyên do hypervisor quản lý.



Hình 1.4. Mô hình ảo hóa Hosted–based

Kiến trúc ảo hóa Hypervisor–based

Hình 1.5. Kiến trúc Hypervisor–based

Trong mô hình này, ta thấy lớp phần mềm hypervisor chạy trực tiếp trên nền tảng phần cứng của máy chủ, không thông qua bất kỳ một hệ điều hành hay một nền tảng nào khác. Qua đó, các hypervisor này có khả năng điều khiển, kiểm soát phần cứng của máy chủ. Đồng thời nó cũng có khả năng quản lý các hệ điều hành chạy trên nó.

Một hệ thống ảo hóa máy chủ sử dụng nền tảng Bare – Mental hypervisor bao gồm ba lớp chính:

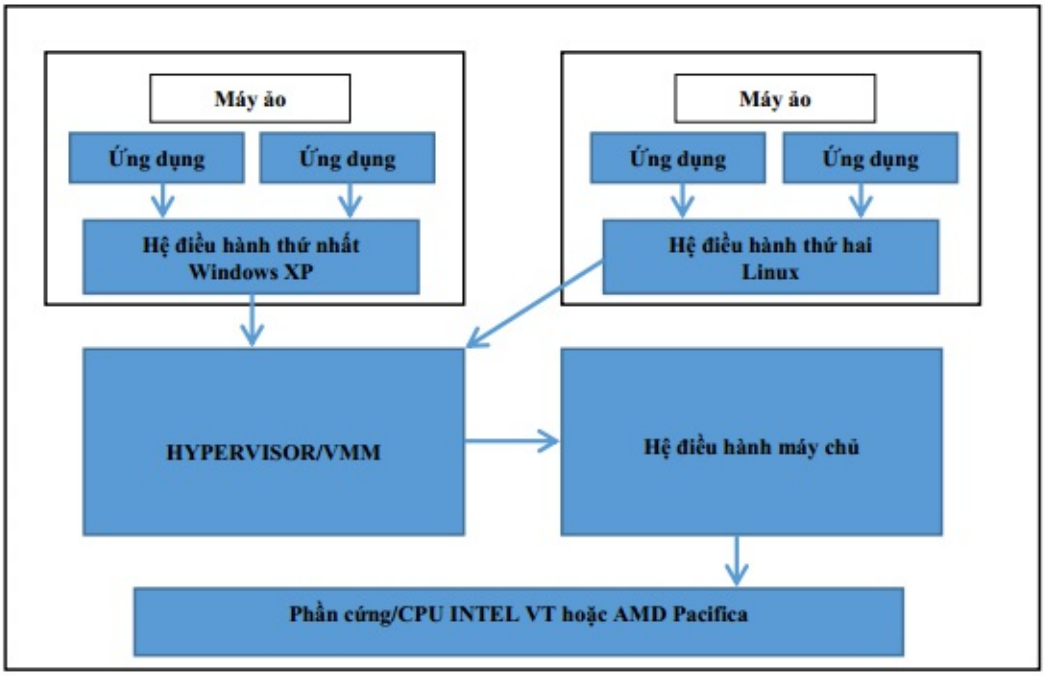
– Nền tảng phần cứng: bao gồm các thiết bị nhập xuất, thiết bị lưu trữ (Hdd, Ram), bộ vi xử lý CPU và các thiết bị khác (các thiết bị mạng, vi xử lý đồ họa, âm thanh…).

– Lớp nền tảng ảo hóa: Virtual Machine Monitor thực hiện việc liên lạc trực tiếp với nền tảng phần cứng phía dưới, quản lý và phân phối tài nguyên cho các hệ điều hành khác nằm trên nó.

– Các ứng dụng máy ảo: các máy ảo này sẽ lấy tài nguyên từ phần cứng, thông qua sự cấp phát và quản lý của hypervisor.

Kiến trúc lai Hybrid

Hybird là một kiểu ảo hóa mới hơn và có nhiều ưu điểm. Trong đó lớp ảo hóa hypervisor chạy song song với hệ điều hành máy chủ. Tuy nhiên, trong cấu trúc ảo hóa này, các máy chủ ảo vẫn phải đi qua hệ điều hành máy chủ để truy cập phần cứng nhưng khác biệt ở chỗ cả hệ điều hành máy chủ và các máy chủ ảo đều chạy trong chế độ hạt nhân.



Hình 1.6. Kiến trúc ảo hóa Hybrid

Ảo hóa trong điện toán đám mây

Trong điện toán đám mây, một trong những vấn đề nền tảng và cơ bản nhất là tính ảo hóa (virtualization) của hạ tầng bên dưới. Trên Linux, các gói phần mềm nguồn mở cung cấp các giải pháp xây dựng các tầng IaaS (đôi khi là cả PaaS) đều đã được đóng gói kèm theo một công nghệ ảo hóa riêng biệt. Ví dụ: Nimbus với Xen (và cả KVM), OpenStack với KVM/QEMU, VMWare với VMWare Hypervisor, OpenVZ Linux Container với công nghệ OpenVZ... Điện toán đám mây và ảo hóa giúp tối ưu hóa tài nguyên về mặt sử dụng năng lượng, sử dụng theo yêu cầu và kèm theo khả năng mở rộng linh hoạt.

Ảo hóa là một phần không thể thiếu trong mọi đám mây dựa trên khả năng trừu tượng hóa và bao đóng. Ảo hóa cung cấp mức độ trừu tượng cần thiết như việc các tài nguyên tính toán, lưu trữ, tài nguyên mạng được đồng nhất thành kho tài nguyên để cấp phát theo nhu cầu. Ảo hóa cung cấp tính bao đóng vì mọi thao tác cài đặt cập nhật trên nguồn tài nguyên ảo hóa chỉ diễn ra trong phạm vi máy ảo mà không ảnh hưởng hay tác động tới các máy ảo khác, tài nguyên khác không được cấp phát.

Công nghệ điện toán đám mây dựa mạnh mẽ vào công nghệ ảo hóa vì các nhân tố sau đây:

– Nhiều ứng dụng có thể chạy trên cùng một server, tài nguyên có thể được sử dụng hiệu quả hơn.

– Khả năng cấu hình cao: Nhiều ứng dụng yêu cầu tài nguyên khác nhau như số lượng core, dung lượng bộ nhớ. Việc cấu hình này khó thực hiện được ở mức độ phần cứng nhưng dễ dàng trong ảo hóa. Ví dụ VMWare.

– Khả năng sẵn sàng của ứng dụng cao: Ảo hóa cung cấp khả năng phục hồi nhanh sau những hư hỏng cũng như khả năng nâng cấp mà không gây ngắt quãng quá trình sử dụng dịch vụ của người dùng.

– Khả năng đáp ứng cao: Ảo hóa cung cấp các cơ chế theo dõi và bảo trì tài nguyên một cách tự động, một số tài nguyên về dữ liệu thông thường có thể được lưu trữ và cho phép dùng lại.

Ứng dụng ảo hóa trong doanh nghiệp

Mục đích ảo hóa máy chủ trong công nghệ điện toán của doanh nghiệp:

– Tiết kiệm được chi phí đầu tư, chi phí duy trì hệ thống.

– Tiết kiệm không gian đặt máy chủ và năng lượng tiêu thụ.

– Giảm thời gian khôi phục sự cố.

– Tạo lập được môi trường kiểm tra chạy thử ứng dụng mà không cần đầu tư thêm hệ thống mới.

– Dễ dàng trong việc mở rộng hệ thống.

– Tạo lập sự tương thích đối với việc sử dụng các chương trình cũ.

Xét ví dụ sau, một bài toán đưa ra cho doanh nghiệp khi họ cần thêm tài nguyên điện toán mới:

a. Lựa chọn 1

– Đầu tư và mở rộng cơ sở hạ tầng của tổ chức.

– Thường xuyên bổ sung thêm máy chủ, thiết bị lưu trữ và kết nối.

b. Lựa chọn 2

– Tập trung hóa và ảo hóa các tài nguyên hiện có.

– Nâng cao mức độ sử dụng tài nguyên vượt qua những hạn chế vật lý.

c. Lựa chọn 3

– Sử dụng cơ sở hạ tầng điện toán đám mây.

– Mở rộng ảo hóa vượt khỏi phạm vi trung tâm dữ liệu doanh nghiệp.

– Thuê các tài nguyên điện toán từ các dịch vụ đám mây.

– Trả tiền theo mức độ sử dụng.

Các bước áp dụng công nghệ ảo hóa

Doanh nghiệp có rất nhiều máy chủ, mỗi máy chủ được đặt ở nhiều nơi khác nhau, vì vậy việc truy xuất hay bảo trì dữ liệu là rất khó khăn. Do đó tất cả các dữ liệu đều được ảo hóa trong đám mây giúp doanh nghiệp giảm thiểu chi phí vận hành bảo trì bảo dưỡng.

Tiếp nhận yêu cầu: ghi nhận lại các thông tin chi tiết về yêu cầu hỗ trợ như: loại yêu cầu, thông tin khách hàng hoặc người yêu cầu, hình thức tiếp nhận (điện thoại, email, chat…).

Phân công người xử lý: Người tiếp nhận có thể trực tiếp xử lý hoặc chuyển cho các bộ phận chức năng xử lý yêu cầu hỗ trợ. Đối với các yêu cầu đơn giản, người xử lý có thể xử lý và nhập

thông tin phản hồi trực tiếp cho yêu cầu. Đối với các yêu cầu phức tạp đòi hỏi nhiều người tham gia xử lý, có thể sử dụng phân hệ quản lý công việc để tiến hành phân công công việc và theo dõi kết quả thực hiện qua hệ thống báo cáo ngày.

Quản lý kho tri thức (knowlegde base): Quản lý các giải pháp xử lý các tình huống sẵn có để người tiếp nhận, xử lý có thể tìm kiếm và trả lời ngay cho các tình huống đã có trên hệ thống cũng như cập nhật giải pháp cho các tình huống đặc trưng vừa xử lý vào kho tri thức chung.

Ví dụ về dịch vụ ứng dụng trên nền tảng hạ tầng ảo hóa – ứng dụng CloudOffice

Quản lý công việc:

– Giao việc, lập kế hoạch thực hiện.

– Báo cáo công việc, theo dõi tiến độ và mức độ hoàn thành công việc.

– Danh sách nhân viên chưa báo cáo.

– Lịch làm việc.

Quản lý Hành chính: Quản lý công văn giấy tờ; Quản lý tài nguyên, tài sản; Quản lý văn phòng phẩm, công cụ dụng cụ; Quản lý chấm công, báo cáo ngày; Quản lý tài liệu dùng chung; Quản lý tin tức nội bộ; Quản lý ngày làm việc; Quản lý thời gian làm việc, nghỉ phép.

CloudHelpdesk – Hỗ trợ khách hàng: Hệ thống cung cấp các tính năng hỗ trợ khách hàng, tiếp nhận và xử lý các vấn đề liên quan đến dịch vụ chăm sóc khách hàng tại các tổ chức, doanh nghiệp. Kiểm soát chặt chẽ quy trình hỗ trợ khách hàng. Qua đó sẽ không còn hiện tượng bỏ quên, không xử lý kịp thời các yêu cầu hỗ trợ, nâng cao năng suất làm việc của nhân viên hỗ trợ khách hàng. Tổng hợp đánh giá được chất lượng sản phẩm, dịch vụ của doanh nghiệp.

CloudHRM – Quản lý nhân sự: Cung cấp các tính năng quản lý xuyên suốt các thông tin về cán bộ từ khi bắt đầu tuyển dụng cho đến khi kết thúc quá trình làm việc. Quy trình quản lý xuyên suốt qua đó đánh giá được đúng năng lực nhân viên, kịp thời ban hành các chính sách thúc đẩy nhân lực phát triển.

Các chức năng của CloudHRM: Quản lý đợt đánh giá, Đánh giá, Tổng hợp báo cáo, Quản lý tuyển dụng, Quản lý hồ sơ, Quản lý hợp đồng, Quản lý nhân viên, Quản lý nhóm nhân viên, Quản lý đào tạo.

CloudAccounting – Quản lý tài chính kế toán: Cung cấp các tính năng phục vụ các nghiệp vụ tài chính kế toán của tổ chức, doanh nghiệp, cập nhật các chế độ tài chính kế toán mới nhất, các quy trình nghiệp vụ kế toán tự động giúp nâng cao hiệu quả bộ phận tài chính kế toán của doanh nghiệp, quy trình quản lý tập trung giúp lãnh đạo kiểm soát được tình hình tài chính, các luồng tiền, công nợ của doanh nghiệp.

Các chức năng chính của CloudAccounting: Quản lý tạm ứng, Quản lý tiền mặt, tiền gửi ngân hàng, Quản lý bán hàng, Quản lý mua hàng, Quản lý lương, Quản lý kho, Quản lý thuế, Quản lý tài sản.

Kết luận

Về cơ bản, mô hình ảo hóa đám mây trong doanh nghiệp đã được đề ra có tính khả thi và đáp ứng được các yêu cầu như:

– Vận dụng lý thuyết về công nghệ ảo hóa: Raid, San, High Availability và những công nghệ liên quan có chức năng hỗ trợ để áp dụng cho doanh nghiệp của mình.

– Vận dụng được các thành phần, cấu trúc và chức năng từng phần của hệ thống ảo hóa. Triển khai mô hình ảo hóa máy chủ có các lợi ích khi ứng dụng mô hình ảo hóa vào trong thực tế như tiết kiệm chi phí, tăng hiệu suất, dễ quản lý,…

1.3. PHÂN LOẠI CÁC MÔ HÌNH ĐIỆN TOÁN ĐÁM MÂY

Điện toán đám mây có ba mô hình cung cấp dịch vụ, tùy theo các đối tượng khách hàng như sau:

Infrastructure as a Service – Dịch vụ hạ tầng: Mô hình dịch vụ này cung cấp cho khách hàng tài nguyên xử lý, lưu trữ, mạng và các tài nguyên máy tính cơ bản khác. Từ đó, khách hàng có thể triển khai và chạy phần mềm tùy ý, bao gồm hệ điều hành và các ứng dụng. Khách hàng không quản lý hoặc kiểm soát các cơ sở hạ tầng điện toán đám mây nằm bên dưới, nhưng có kiểm soát hệ thống điều hành, lưu trữ và các ứng dụng được triển khai đồng thời kiểm soát có giới hạn của các thành phần mạng.

Platform as a Service – Dịch vụ nền tảng: Mô hình dịch vụ này cung cấp cho khách hàng khả năng triển khai trên hạ tầng điện toán đám mây các ứng dụng của họ bằng việc sử dụng các ngôn ngữ lập trình, các thư viện, dịch vụ, công cụ được hỗ trợ từ bên thứ ba. Người dùng không cần quản lý hoặc kiểm soát các cơ sở hạ tầng điện toán đám mây bên dưới như máy chủ ảo, mạng, hệ điều hành, lưu trữ, nhưng có thể cấu hình cho môi trường chạy ứng dụng của họ.

Software as a Service – Dịch vụ phần mềm: Mô hình dịch vụ này cung cấp cho phép khách hàng sử dụng các dịch vụ phần mềm của nhà cung cấp ứng dụng được triển khai trên hạ tầng điện toán đám mây. Các ứng dụng có thể truy cập từ các thiết bị khác nhau thông qua giao diện "mỏng" (thin client interface), chẳng hạn như một trình duyệt web (ví dụ như email trên web), hoặc qua giao diện của chương trình. Khách hàng không quản lý hoặc kiểm soát cơ sở hạ tầng điện toán đám mây nằm bên dưới bao gồm mạng, máy chủ, hệ điều hành, lưu trữ,…, với ngoại lệ có thể thiết lập cấu hình ứng dụng hạn chế người sử dụng cụ thể.

1.4. KIẾN TRÚC ĐÁM MÂY HƯỚNG THỊ TRƯỜNG

Một trong những bài toán quan trọng của điện toán đám mây là định giá cho tài nguyên trên đám mây để cho thuê. Có một số cách định giá cơ bản như sau:

Định giá cố định: Nhà cung cấp sẽ xác định rõ đặc tả về khả năng tính toán cố định (dung lượng bộ nhớ được cấp phát, loại CPU và tốc độ...).

Định giá theo đơn vị sử dụng: được áp dụng phổ biến cho lượng dữ liệu truyền tải, dung lượng bộ nhớ được cấp phát và sử dụng,… cách này uyển chuyển hơn cách trên.

Định giá theo thuê bao: ứng dụng phần lớn trong mô hình dịch vụ phần mềm (SaaS) người dùng sẽ tiên đoán trước định mức sử dụng ứng dụng cloud (cách tính này thường khó đạt được độ chính xác cao).

Với các cách định giá như trên, nhà cung cấp dịch vụ và người sử dụng cần có những thỏa thuận cụ thể về việc áp dụng cách định giá với tài nguyên. Điều này phải được nêu trong SLA (Service Level Agreement) trong đó xác định về yêu cầu chất lượng dịch vụ QoS (Quality of Service). Kiến trúc Market-oriented cloud bao gồm bốn thành phần chủ yếu:

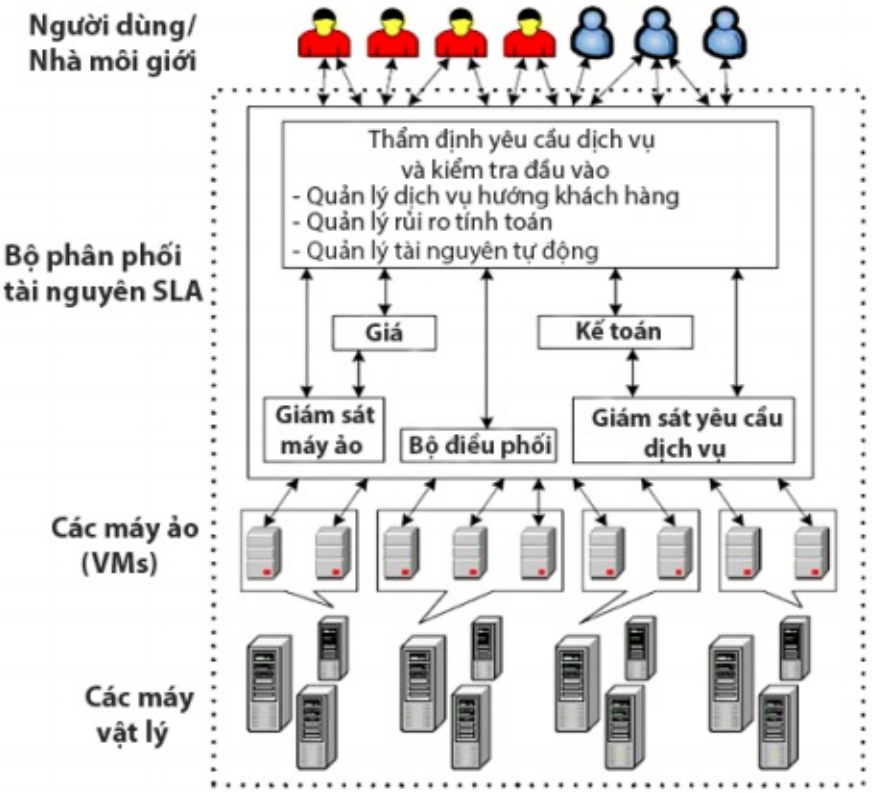
User/Broker (Người dùng/Nhà môi giới): Người dùng hay nhà phân phối sử dụng quyền ủy thác để gửi yêu cầu dịch vụ từ bất kỳ đâu trên thế giới tới trung tâm dữ liệu hay Cloud để được

xử lý.

SLA Resource Allocator (Bộ phân phối tài nguyên SLA): đóng vai trò như một trung gian giữa các nhà cung cấp Cloud với người dùng/nhà môi giới bên ngoài.

VMs (các máy ảo virtual machine): Nhiều máy ảo có thể được mở và tắt tự động trên một máy vật lý để phù hợp với yêu cầu dịch vụ từ phía người dùng/nhà môi giới bên ngoài cũng như phù hợp với việc sử dụng hiệu quả tài nguyên từ phía nhà cung cấp cloud. Cơ chế cung cấp máy ảo, như đã phân tích, là một trong những công nghệ chủ chốt cho phép tham số hóa cấu hình máy chủ ảo để phù hợp với điều luật sử dụng được ký kết.

Physical Machines (các máy vật lý): Các máy vật lý chính là tài nguyên hạ tầng của nhà cung cấp cloud. Các máy tính vật lý này chạy firmware hypervisor để cấp phát và chạy các máy ảo theo yêu cầu.



Hình 1.7. Kiến trúc Market-Oriented Cloud

Đi vào chi tiết, bộ phân phối tài nguyên SLA gồm các thành phần sau đây:

Service Request Examiner and Admission Control: Khi một yêu cầu dịch vụ được gửi lên lần đầu sẽ được phiên dịch thành các yêu cầu về chất lượng dịch vụ QoS trước khi xác định xem yêu cầu đó được chấp nhận hay từ chối. Điều này đảm bảo rằng không có tình trạng quá tải dịch vụ khi các yêu cầu dịch vụ không thể được đáp ứng đầy đủ vì giới hạn của tài nguyên hệ thống sẵn có. Dịch vụ này cần thông tin trạng thái cuối cùng về tình trạng sẵn sàng của tài nguyên (từ cơ chế VM Monitor) và khả năng xử lý tải (từ cơ chế Service Request Monitor) theo thứ tự để quyết định việc phân phối tài nguyên một cách hiệu quả. Sau đó dịch vụ này sẽ phân yên cầu cho các máy ảo VM và xác định đặc tả tài nguyên cho máy ảo được cấp phát.

Pricing: Cung cấp cơ chế quyết định cách các yêu cầu dịch vụ được tính phí sử dụng. Ví dụ như dịch vụ được tính phí sử dụng dựa theo thời gian thực thi các nhiệm vụ, tỷ lệ giá cả (cố định/thay đổi) hay tính sẵn sàng của tài nguyên (sẵn có/yêu cầu). Cơ chế định giá thích hợp có mục đích là nhằm cân bằng chi phí cho người sử dụng và nhà cung cấp dịch vụ cloud.

Accounting: Cung cấp cơ chế theo dõi lưu lượng tài nguyên được sử dụng để tính chi phí cho người dùng.

VM Monitor: Cung cấp cơ chế lưu vết, giám sát những máy ảo đang sử dụng và các thông tin về tài nguyên của chúng.

Dispatcher: Cung cấp cơ chế bắt đầu thực thi việc cấp phát máy ảo cho những yêu cầu dịch vụ đã được chấp nhận.

Service Request Monitor: Cung cấp cơ chế lưu vết tiến trình yêu cầu dịch vụ.

1.5. CÁC CÔNG CỤ MÔ PHỎNG ĐÁM MÂY

Hiện nay, bên cạnh các hệ thống đám mây thương mại, các phần mềm cài đặt đám mây mã nguồn mở còn có một số công cụ có chức năng mô phỏng môi trường đám mây. Các công cụ này có ưu điểm là giúp người sử dụng, các nhà nghiên cứu công nghệ có thể thử nghiệm các sản phẩm công nghệ của mình trên "mây" mà không cần phải tự mình quản lý một đám mây thật. Tiêu biểu nhất trong các công cụ mô phỏng đám mây đó là CloudSim.

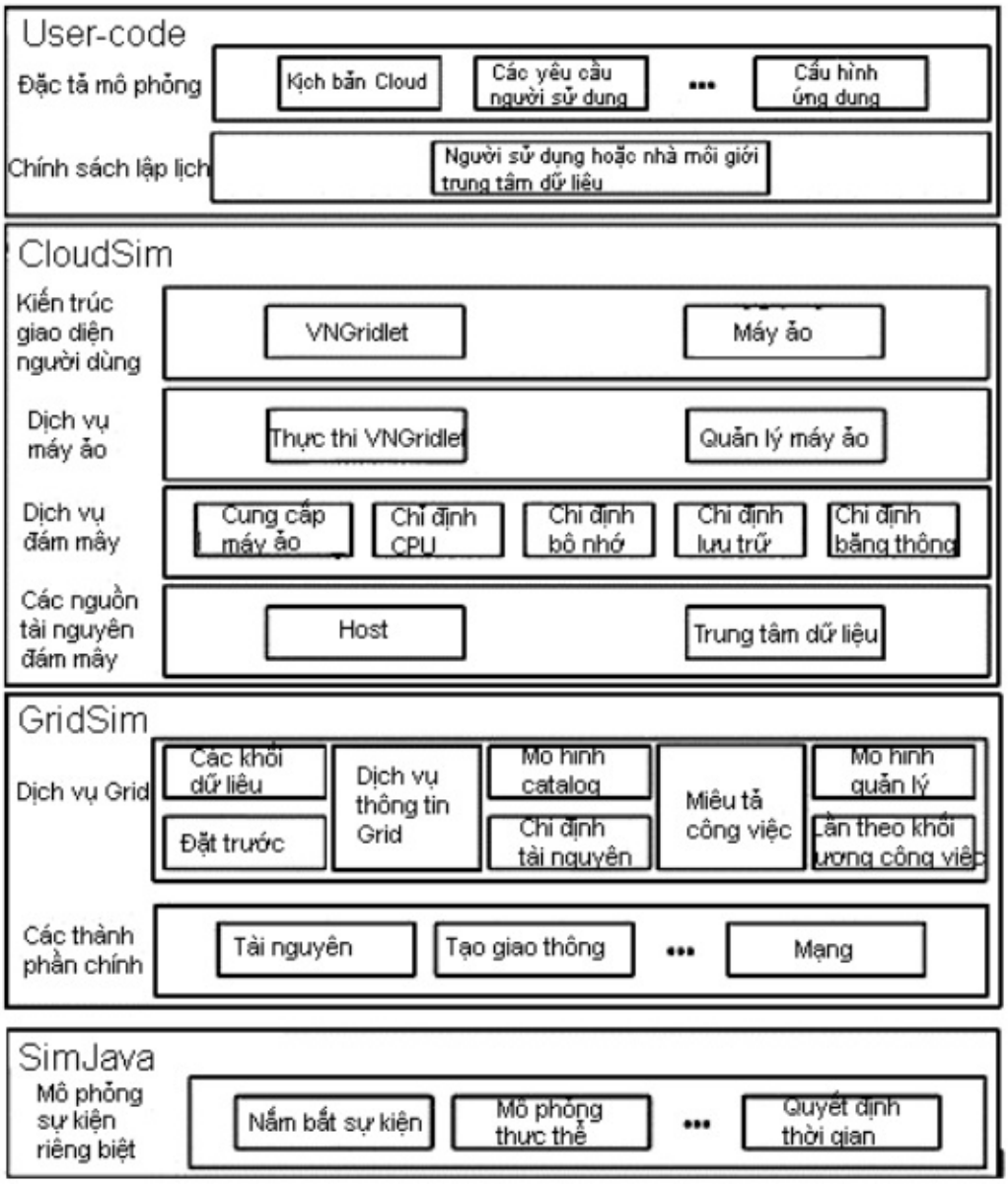
Kiến trúc CloudSim

Một cách tổng thể, CloudSim bao gồm bốn lớp (xem hình 1.8):

SimJava: Mức thấp nhất trong kiến trúc bao gồm những công cụ mô phỏng sự kiện dùng để hiện thực những chức năng cốt lõi cần thiết cho việc mô phỏng ở lớp cao hơn như sắp xếp và xử lý sự kiện, khởi tạo các thành phần, quản lý mô phỏng đồng hồ.

GridSim: Bộ công cụ hỗ trợ các thành phần phần mềm cấp cao hơn để mô hình hóa nhiều nền tảng lưới, bao gồm cả hệ thống mạng và liên kết đồng bộ những thành phần cơ bản của lưới như tài nguyên, tập dữ liệu, dịch vụ giám sát và lưu vết, dịch vụ thông tin.

CloudSim: là phần hiện thực ở mức tiếp theo do việc mở rộng tự động các tính năng cơ bản được cung cấp bởi lớp GridSim. CloudSim cung cấp hỗ trợ cho việc mô hình và mô phỏng hóa môi trường nền tảng Cloud. Lớp CloudSim quản lý việc khởi tạo và thực thi các thực thể cốt lõi (máy ảo, thiết bị lưu trữ, ứng dụng) trong suốt quá trình mô phỏng. Lớp này có khả năng khởi tạo đồng thời và quản lý mở rộng trong suốt với những nền tảng Cloud bao gồm hàng nghìn thành phần hệ thống. Những vấn đề cơ bản như triển khai máy ảo VM dựa trên yêu cầu người dùng, quản lý quá trình thực thi ứng dụng và theo dõi tự động đều được quản lý bởi lớp này.



Hình 1.8. Kiến trúc lớp của CloudSim

User–code: đây là lớp trên cùng của hệ thống mô phỏng cho phép cấu hình những chức năng liên quan đến các máy chủ (số lượng, đặc tả máy chủ ảo), liên quan đến ứng dụng (số lượng các tác vụ và yêu cầu đặc tả), các máy ảo VM, số lượng người dùng. Một người phát triển ứng dụng mô phỏng Cloud có thể tùy chọn và tham số cấu hình ứng dụng, ngữ cảnh thực nghiệm ở lớp này.

Mô hình mô phỏng Cloud

Kiến trúc dịch vụ nền tảng liên quan đến Cloud được mô hình hóa trong chương trình mô phỏng bởi thành phần Data center. Data center được tạo bởi các tập hợp các Host, có trách nhiệm quản lý các máy ảo VM trong chu kỳ sống của chúng. Các Host là các nút trong Cloud: nó được tham số khả năng xử lý của bộ vi xử lý trung tâm CPU (biểu diễn qua đơn vị MIPS = milion of instruction per second), bộ nhớ, khả năng lưu trữ và chính sách định thời để xử lý việc cấp phát lõi tính toán cho các máy ảo. Các thành phần máy Host của nền tảng mô phỏng hỗ trợ mô phỏng vi xử lý một nhân và đa nhân.

Việc phân phối máy ảo phục vụ cho ứng dụng cụ thể nào đó đến các thành phần Host là trách nhiệm của thành phần Virtual Machine Provisioner. Thành phần này cung cấp một tập các phương thức cho người sử dụng, với nhưng chính sách điều phối tài nguyên hướng tới mục tiêu tối ưu hiệu quả sử dụng. Những chính sách mặc định hiện có sẵn rất giản đơn theo hướng

ai đến trước sẽ được phục vụ trước.

Với mỗi thành phần Host, sự cấp phát các vi xử lý CPU tới các máy ảo được thực hiện theo chính sách điều phối cụ thể dựa theo số lượng yêu cầu và số lượng vi xử lý sẵn có. Do vậy, có thể có các chính sách như cấp phát CPU dành riêng cho máy ảo hay phân tán động giữa các máy ảo (chia sẻ theo thời gian).

Mô hình cấp phát máy ảo VM

Một trong những ý tưởng khiến Cloud computing khác biệt với Grid computing là việc triển khai tối đa công nghệ và các công cụ ảo hoá.

Để cho phép giả lập những chính sách khác nhau, CloudSim hỗ trợ việc cấp phát máy ảo VM ở hai mức: trước tiên tại mức Host và sau đó là mức máy ảo VM. Ở mức đầu tiên, có thể xác định rõ tổng năng lực xử lý của mỗi nhân trong Host sẽ được gán cho mỗi máy ảo. Tại mức tiếp theo, các máy ảo VM sẽ được phân rõ tổng năng lực xử lý cụ thể cho mỗi tác vụ được thực thi. Tại mỗi mức, CloudSim hiện thực chính sách cấp phát tài nguyên theo thời gian và không gian.

Mô hình chợ Cloud

Mô hình chợ Cloud đóng vai trò như người môi giới giữa nhà cung cấp dịch vụ Cloud và khách hàng là điểm nhấn của Cloud computing. Hơn thế nữa, những dịch vụ này cần cơ chế để xác định chi phí dịch vụ và các chính sách về giá.

Mô hinh chính sách, chi phí và giá cả là một ý tưởng được xem xét khi thiết kết chương trình mô phỏng Cloud, bốn thuộc tính được xem xét đến là:

– Chi phí mỗi bộ xử lý;

– Chi phí mỗi đơn vị bộ nhớ;

– Chi phí mỗi đơn vị lưu trữ;

– Chi phí mỗi đơn vị băng thông sử dụng.

Chi phí mỗi đơn vị bộ nhớ và lưu trữ được kèm theo trong quá trình khởi tạo máy ảo. Chi phí mỗi đơn vị băng thông sử dụng có trong quá trình truyền dữ liệu. Bên cạnh đó, các chi phí sử dụng bộ nhớ, lưu trữ và các chi phí liên quan có mối liên hệ với việc sử dụng tài nguyên tính toán. Do vậy, nếu máy ảo VM được tạo mà không có tác vụ nào thực thi trên chúng, thì chỉ có chi phí về bộ nhớ và lưu trữ. Những vấn đề này có thể được thay đổi bởi người dùng.

1.6. CÂU HỎI VÀ BÀI TẬP

1. Nêu các đặc trưng của trung tâm dữ liệu.
2. So sánh các mô hình trung tâm dữ liệu. Đánh giá ưu, nhược điểm của từng mô hình khi áp dụng cho các doanh nghiệp.
3. Trình bày đặc thù của công nghệ ảo hóa và vai trò trong điện toán đám mây.
4. So sánh các kiến trúc ảo hóa. Đánh giá ưu/nhược điểm của từng kiến trúc.
5. Các phân loại mô hình điện toán đám mây có những đặc điểm gì giống và khác nhau.

6. Phân biệt vai trò của các thành phần trong kiến trúc đám mây hướng thị trường.
7. Nêu các đặc điểm của bộ phân phối tài nguyên SLA trong kiến trúc đám mây thị trường. Tìm hiểu và trình bày một số kiến trúc SLA cụ thể.
8. Tìm hiểu và trình bày chi tiết về công cụ mô phỏng đám mây: CloudSim.

Chương 2
LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU

2.1. Ô THỐNG LƯU TRỮ PHÂN TÁN VÀ ĐỒNG NHẤT BỘ NHỚ NFS, AFS

Trong phần này, chúng ta sẽ đi vào tìm hiểu kiến trúc các hệ thống lưu trữ phân tán cơ bản như hệ thống quản lý tập tin phân tán NFS và AFS. Đặc tính chung của các hệ lưu trữ phân tán là nhằm mục đích lưu trữ tập trung và chia sẻ thông tin cho các máy tính trong cùng mạng nội bộ.

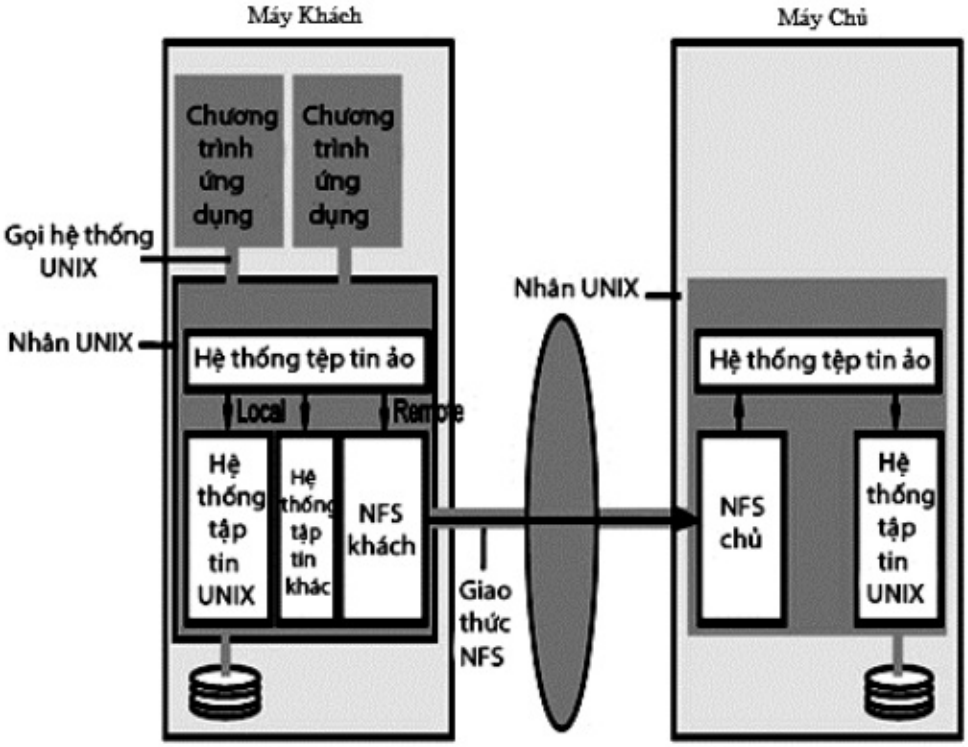
NFS

NFS hay Network File System là kiến trúc hệ thống tập tin phân tán mà một máy chủ trong hệ thống đóng vai trò là máy chủ lưu trữ, cung cấp năng lực lưu trữ của các ổ đĩa cứng cục bộ, hệ thống RAID cho các máy tính khác qua giao thức mạng. NFS là kiến trúc hệ quản lý tập tin phân tán rất phổ biến, được hỗ trợ bởi hầu hết các nền tảng hệ điều hành như Windows, Unix.

Ưu điểm của NFS là tính trong suốt cho người dùng cuối về cách thức truy cập tập tin hay vị trí nơi tập tin được lưu trữ. Hệ thống tập tin NFS được ánh xạ như một thư mục trong hệ thống quản lý tập tin cục bộ và không có sự khác biệt. Yếu điểm của NFS là tính khả mở thấp do mọi thao tác đọc ghi dữ liệu đều thực hiện qua kết nối mạng với máy chủ lưu trữ NFS. Trong trường hợp nếu có truy cập tương tranh vào cùng một tệp, hiệu năng của NFS suy giảm rõ rệt.

AFS

AFS cũng là một hệ thống tập tin phân tán nhằm mục đích chia sẻ tập tin cho một lượng lớn người dùng mạng. So với NFS, AFS có tính khả mở cao hơn, đáp ứng được số lượng người dùng lớn hơn nhờ vào đặc trưng sau đây: Khi truy cập tập tin, toàn bộ tập tin sẽ được sao chép về phía máy người sử dụng và các thao tác đọc ghi được thực hiện trên tập tin đó. Khi tập tin được đóng, nội dung tập tin sẽ được cập nhật về phía máy chủ lưu trữ. Chính vì vậy, quá trình đọc ghi tương tranh là trong suốt đối với từng người sử dụng nhưng tính nhất quán của tập tin không được đảm bảo.



Hình 2.1. Kiến trúc hệ thống tập tin NFS

2.2. HỆ THỐNG LƯU TRỮ HDFS, GFS

HDFS

Hadoop framwork của Apache là một nền tảng dùng để phân tích các tập dữ liệu rất lớn mà không thể xử lý trên được trên một máy chủ duy nhất. Hadoop trừu tượng hóa mô hình tính toán MapReduce, làm nó trở nên dễ tiếp cận hơn với các nhà phát triển. Hadoop có khả năng mở rộng vô số các nút lưu trữ và có thể xử lý tất cả hoạt động và phân phối liên quan đến việc phân loại dữ liệu.

Tổng quan thiết kế của HDFS

HDFS (Hadoop distributed file system) ra đời trên nhu cầu lưu trữ dữ liệu của Nutch, một dự án Search Engine nguồn mở. HDFS kế thừa các đặc tính chung của các hệ thống tập tin phân tán thế hệ trước như độ tin cậy, khả năng mở rộng và hiệu suất hoạt động. HDFS được thiết kế với những giả định như dưới đây:

Thứ nhất, các lỗi về phần cứng sẽ thường xuyên xảy ra. Hệ thống HDFS sẽ chạy trên các cluster với hàng trăm hoặc thậm chí hàng nghìn nút. Các nút này được xây dựng từ các phần cứng thông thường, giá rẻ, tỷ lệ lỗi cao. Chất lượng và số lượng của các thành phần phần cứng như vậy sẽ tất yếu dẫn đến tỷ lệ xảy ra lỗi trên hệ thống cluster cao. Có thể điểm qua một số lỗi như lỗi của ứng dụng, lỗi của hệ điều hành, lỗi đĩa cứng, bộ nhớ, lỗi của các thiết bị kết nối, lỗi mạng, lỗi về nguồn điện… Vì thế, khả năng phát hiện lỗi, chống chịu lỗi và tự động phục hồi phải được tích hợp vào trong hệ thống HDFS.

Thứ hai, do đặc thù lưu trữ dữ liệu có dung lượng lớn, HDFS được thiết kế để tối ưu cho bài toán lưu trữ các tập tin có kích thước lớn hàng GB, thậm chí TB. Để giải quyết bài toán này, dữ liệu của các tập tin lớn sẽ được chia nhỏ thành các khối lớn (ví dụ 64MB) và phân tán trên các nút lưu trữ. So với các hệ thống tập tin khác, HDFS không tối ưu cho bài toán lưu trữ hàng tỉ tập tin nhỏ với kích thước mỗi tập tin chỉ vài KB. Ưu điểm của thiết kế tập tin lớn là giảm tải cho hệ thống quản lý không gian tập tin, giảm thời gian thao tác trên các thư mục hay tìm kiếm tập

tin.

Thứ ba, HDFS ban đầu được thiết kế chỉ cho phép thay đổi nội dung các tập tin được lưu trữ qua phép toán thêm "append" dữ liệu vào cuối tập tin hơn là ghi đè lên dữ liệu hiện có. Việc ghi dữ liệu lên một vị trí ngẫu nhiên trong tập tin không được hỗ trợ. Một khi đã được tạo ra, các tập tin sẽ trở thành file chỉ đọc (read-only). Thiết kế này khác căn bản so với các hệ thống quản lý tập tin truyền thống do khác biệt về mục đích sử dụng. HDFS được thiết kế để tối ưu cho bài toán lưu trữ dữ liệu cho việc phân tích khi mà đầu vào có thể là các tập tin nhật ký logs hay dữ liệu liên tục đến từ các cảm biến. Với đầu vào dữ liệu này thì thao tác ghi ngẫu nhiên hay ghi đè dữ liệu là không cần thiết. Hơn nữa, đơn giản hóa hỗ trợ ghi dữ liệu cũng là nhân tố để HDFS tối ưu và tăng hiệu năng hệ thống.

Ngày nay, Hadoop cluster và HDFS rất phổ biến trên thế giới. Nổi bật nhất là hệ thống của Yahoo với một cluster lên đến 1100 nút với dung lượng HDFS là 12 PB. Các công ty khác như Facebook, Adode, Amazon cũng đã xây dựng các cluster chạy HDFS với dung lượng hàng trăm, hàng nghìn TB.

Kiến trúc HDFS

Giống như các hệ thống tập tin khác, HDFS duy trì một cấu trúc cây phân cấp các tập tin, thư mục mà các tập tin sẽ đóng vai trò là các nút lá. Trong HDFS, vì kích thước mỗi tập tin lớn, mỗi tập tin sẽ được chia ra thành các khối (block) và mỗi khối này sẽ có một block ID để nhận diện. Các khối của cùng một file (trừ khối cuối cùng) sẽ có cùng kích thước và kích thước này được gọi là block size của tập tin đó. Mỗi khối của tập tin sẽ được lưu trữ thành nhiều bản sao (replica) khác nhau vì mục đích an toàn dữ liệu. Các khối được lưu trữ phân tán trên các máy chủ lưu trữ cài HDFS. HDFS có một kiến trúc chủ/khách (master/slave). Trên một cluster chạy HDFS, có hai loại nút (node) là Namenode và Datanode. Một cluster có duy nhất một Namenode và có một hoặc nhiều Datanode.

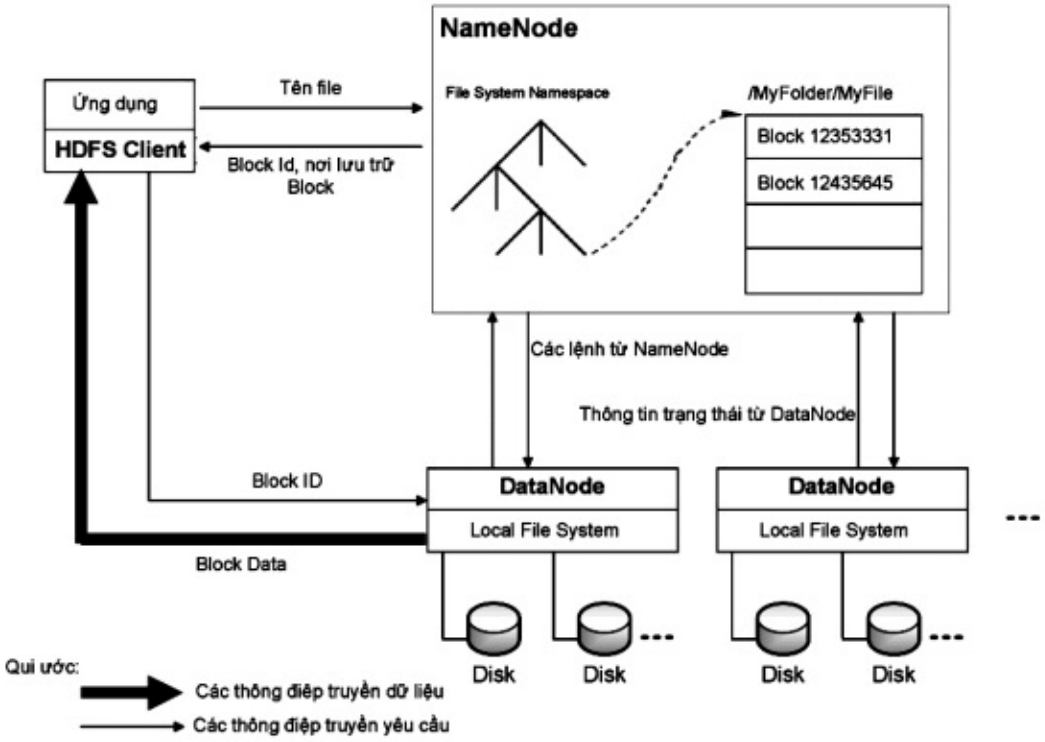
Namenode đóng vai trò là master, chịu trách nhiệm duy trì thông tin về cấu trúc cây phân cấp các tập tin, thư mục của hệ thống tập tin và các siêu dữ liệu (metadata) khác của hệ thống tập tin. Cụ thể, các metadata mà Namenode lưu trữ gồm có:

– *File system namespace (không gian tên tập tin)*: là hình ảnh cây thư mục của hệ thống tập tin tại một thời điểm nào đó. Không gian tên tập tin thể hiện tất cả các file, thư mục có trên hệ thống file và quan hệ giữa chúng.

– *Thông tin để ánh xạ từ tên tập tin ra thành danh sách các khối*: Với mỗi tập tin, ta có một danh sách có thứ tự các khối block của tập tin đó, mỗi khối đại diện bởi Block ID.

– *Nơi lưu trữ các khối*: Các khối được đại diện một Block ID. Với mỗi block, ta có một danh sách các DataNode lưu trữ các bản sao của khối đó.

Các DataNode sẽ chịu trách nhiệm lưu trữ các khối thật sự của từng tập tin của HDFS. Mỗi một khối sẽ được lưu trữ như một tập tin riêng biệt trên hệ thống tập tin cục bộ của DataNode. Trong ngữ cảnh MapReduce, các DataNode này còn có chức năng tính toán với dữ liệu là chính các khối được lưu trữ. Kiến trúc của HDFS được thể hiện trong hình 2.2.



Hình 2.2. Kiến trúc HDFS

Về cơ bản, Namenode sẽ chịu trách nhiệm điều phối các thao tác truy cập (đọc/ghi dữ liệu) của người sử dụng (client) lên hệ thống HDFS. Khi client của hệ thống muốn đọc một tập tin trên hệ thống HDFS, client này sẽ thực hiện một yêu cầu thông qua RPC đến Namenode để lấy các metadata của tập tin cần đọc. Từ metadata này, client sẽ biết được danh sách các block của tệp tin và vị trí của các DataNode chứa các bản sao của từng block. Client sẽ truy cập vào các DataNode để thực hiện các yêu cầu đọc các block.

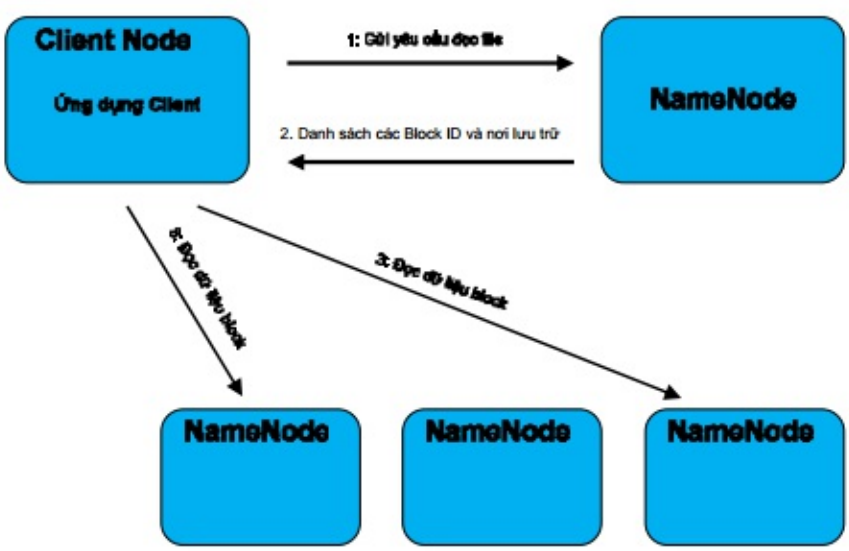
Định kỳ, mỗi DataNode sẽ báo cáo cho NameNode biết về danh sách tất cả các block mà nó đang lưu trữ, NameNode sẽ dựa vào những thông tin này để cập nhật lại các metadata trong nó. Cứ sau mỗi lần cập nhật lại như vậy, metadata trên NameNode sẽ đạt được tính nhất quán với dữ liệu trên các DataNode. Toàn bộ trạng thái của metadata khi đang ở tình trạng thống nhất này được gọi là một checkpoint. Chỉ khi nào metadata ở trạng thái checkpoint mới có thể được nhân bản cho mục đích phục hồi lại NameNode.

NameNode và quá trình tương tác giữa client và HDFS

Việc tồn tại duy nhất một NameNode trên một hệ thống HDFS đã làm đơn giản hóa thiết kế của hệ thống và cho phép NameNode ra những quyết định thông minh trong việc sắp xếp các block dữ liệu lên trên các DataNode dựa vào các kiến thức về môi trường hệ thống như: cấu trúc mạng, băng thông mạng, khả năng của các DataNode... Tuy nhiên, một nhu cầu đặt ra là cần phải tối thiểu hóa sự tham gia của NameNode vào các quá trình đọc/ghi dữ liệu lên hệ thống để tránh tình trạng nút thắt cổ chai (bottle neck). Client sẽ không bao giờ đọc hay ghi dữ liệu lên hệ thống thông qua NameNode. Thay vào đó, client sẽ hỏi NameNode về thông tin các DataNode có các block cần truy cập hoặc các DataNode có thể ghi các block mới. Sau đó, client sẽ lưu trữ lại tạm thời các thông tin này và kết nối trực tiếp với các DataNode để thực hiện các thao tác truy xuất dữ liệu.

Quá trình đọc file

Sơ đồ sau miêu tả quá trình client đọc một tập tin trên HDFS.



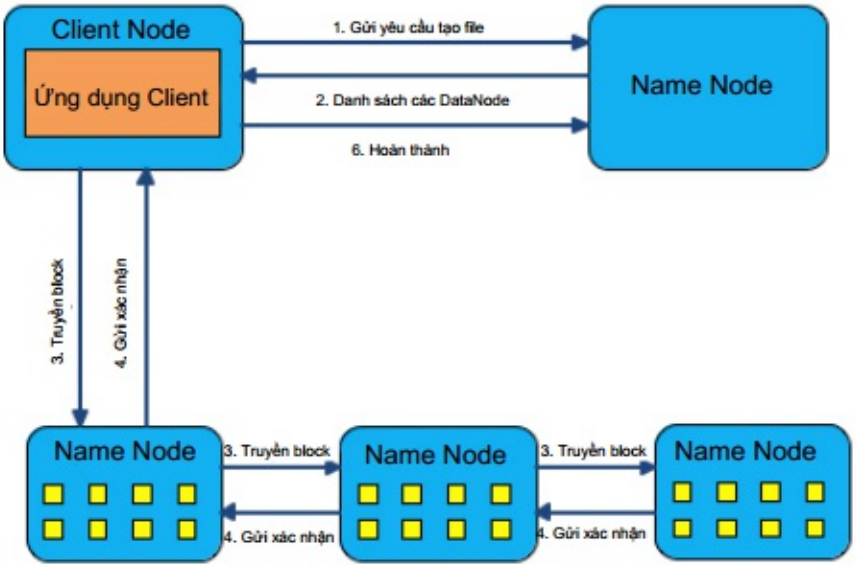
Hình 2.3. Quá trình đọc tập tin trên HDFS

Đầu tiên, client sẽ mở tập tin cần đọc bằng cách gửi yêu cầu đọc tập tin đến NameNode (1). NameNode sẽ thực hiện một số kiểm tra xem file được yêu cầu đọc có tồn tại không. Nếu không có vấn đề gì xảy ra, NameNode sẽ gửi danh sách các block (đại diện bởi Block ID) của tệp tin cùng với địa chỉ các DataNode chứa các bản sao của block này (2). Tiếp theo, client sẽ mở các kết nối tới DataNode, thực hiện một yêu cầu RPC để yêu cầu nhận block cần đọc và đóng kết nối với DataNode (3). Lưu ý là với mỗi block, ta có thể có nhiều DataNode lưu trữ các bản sao của block đó. Client sẽ chỉ đọc bản sao của block từ DataNode theo thứ tự ưu tiên được cấu hình trong hệ thống. Client sẽ thực hiện việc đọc các block lặp đi lặp lại cho đến khi block cuối cùng của file được đọc xong.

Như vậy, trong quá trình một client đọc một file trên HDFS, ta thấy client sẽ trực tiếp kết nối với các DataNode để lấy dữ liệu chứ không cần thực hiện gián tiếp qua NameNode. Điều này sẽ làm giảm đi rất nhiều việc trao đổi dữ liệu giữa client và NameNode, khối lượng luân chuyển dữ liệu sẽ được trải đều ra khắp cluster, tình trạng bottle neck "thắt cổ chai" sẽ không xảy ra. Do đó, cluster chạy HDFS có thể đáp ứng đồng thời nhiều client cùng thao tác tại một thời điểm.

Quá trình ghi file

Quá trình ghi tập tin bắt đầu với việc client sẽ gửi yêu cầu tạo một chỉ mục tập tin (file entry) lên không gian tên của hệ thống tập tin đến NameNode (1). Tập tin mới được tạo sẽ rỗng, tức chưa có một block nào. Sau đó, NameNode sẽ quyết định danh sách các DataNode sẽ chứa các block của tập tin và gửi lại cho client (2). Client sẽ chia dữ liệu của tập tin cần tạo ra thành các block, mỗi block sẽ được lưu ra thành nhiều bản sao trên các DataNode khác nhau (tùy vào chỉ số độ nhân bản của tập tin).



Hình 2.4. Quá trình tạo và ghi dữ liệu lên tập tin HDFS

Client gửi block cho DataNode thứ nhất, DataNode thứ nhất sau khi nhận được block sẽ tiến hành lưu lại bản sao thứ nhất của block. Tiếp theo DataNode thứ nhất sẽ gửi block này cho DataNode thứ hai để lưu ra bản sao thứ hai của block. Tương tự, DataNode thứ hai sẽ gửi packet cho DataNode thứ ba. Cứ như vậy, các DataNode cũng lưu các bản sao của một block. Quá trình này được hình dung như là một ống dẫn dữ liệu *data pile*. Sau khi DataNode cuối cùng nhận được thành công block, nó sẽ gửi lại cho DataNode thứ hai một gói xác nhận rằng đã lưu thành công (4). Và gói thứ hai lại gửi gói xác nhận tình trạng thành công của hai DataNode về DataNode thứ nhất.

Client sẽ nhận được các báo cáo xác nhận từ DataNode thứ nhất cho tình trạng thành công của tất cả DataNode trên *data pile*. Nếu có bất kỳ một DataNode nào bị lỗi trong quá trình ghi dữ liệu, client sẽ tiến hành xác nhận lại các DataNode đã lưu thành công bản sao của block và thực hiện một hành vi ghi lại block lên trên DataNode bị lỗi.

Sau khi tất cả các block của tập tin đều đã được ghi lên các DataNode, client sẽ thực hiện một thông điệp báo cho NameNode nhằm cập nhật lại danh sách các block của tập tin vừa tạo. Thông tin ánh xạ từ BlockID sang danh sách các DataNode lưu trữ sẽ được NameNode tự động cập nhật qua giao thức điểm danh mà định kỳ các DataNode sẽ gửi báo cáo cho NameNode danh sách các block mà nó quản lý.

Như vậy, cũng giống như trong quá trình đọc, client sẽ trực tiếp ghi dữ liệu lên các DataNode mà không cần phải thông qua NameNode. Một đặc điểm nổi trội nữa là khi client ghi một block với chỉ số nhân bản là n, tức nó cần ghi n block lên n DataNode, nhờ cơ chế luân chuyển block dữ liệu qua ống dẫn (pipe) nên lưu lượng dữ liệu cần write từ client sẽ giảm đi n lần, phân đều ra các DataNode trên cluster.

Kích thước khối block

Như ta đã biết, trên đĩa cứng, đơn vị lưu trữ dữ liệu nhỏ nhất không phải là byte, bit hay KB mà là một block. Kích thước block (block size) của đĩa cứng sẽ quy định lượng dữ liệu nhỏ nhất mà ta có thể đọc/ghi lên đĩa. Các hệ thống tập tin như của Windows hay Unix cũng sử dụng block như là đơn vị trao đổi dữ liệu nhỏ nhất, block size trên các file system này thường là khoảng nhiều lần block size trên đĩa cứng.

HDFS cũng chia file ra thành các block, và mỗi block này sẽ được lưu trữ trên Datanode thành một file riêng biệt trên hệ thống file local của nó. Đây cũng là đơn vị trao đổi dữ liệu nhỏ

nhất giữa HDFS và client của nó. Block size là một trong những điểm quan trọng trong thiết kế HDFS. Block size mặc định của HDFS là 64 MB, nhưng thông thường trên các hệ thống lớn, người ta dùng block size là 128 MB, lớn hơn block size của các hệ thống file truyền thống rất nhiều.

Việc sử dụng block size lớn, tức sẽ giảm số lượng block của một tập tin, mang lại một số thuận lợi. Đầu tiên, nó sẽ làm giảm nhu cầu tương tác với NameNode của client vì việc đọc/ghi trên một block chỉ cần một lần tương tác với NameNode để lấy Block ID và nơi lưu block đó. Thứ hai, với block size lớn, client sẽ phải tương tác với DataNode ít hơn. Mỗi lần client cần đọc một BlockID trên DataNode, client phải tạo một kết nối TCP/IP đến DataNode. Việc giảm số lượng block cần đọc sẽ giảm số lượng kết nối cần tạo, client sẽ thường làm việc với một kết nối bền vững hơn là tạo nhiều kết nối. Thứ ba, việc giảm số lượng block của một tập tin sẽ làm giảm khối lượng metadata trên NameNode. Điều này giúp MasterNode có thể đưa toàn bộ metadata vào bộ nhớ chính mà không cần phải lưu trữ trên đĩa cứng.

Mặt khác, việc sử dụng block size lớn sẽ dẫn đến việc một tập tin nhỏ chỉ có một vài block, thường là chỉ có một. Điều này dẫn đến việc các DataNode lưu block này sẽ trở thành điểm nóng khi có nhiều client cùng truy cập vào tập tin. Tuy nhiên, hệ thống HDFS đa phần chỉ làm việc trên các file có kích thước lớn với nhiều block như đã đề cập ở phần trên nên sự bất cập này là không đáng kể trong thực tiễn.

GFS

GFS hay Google file system là hệ thống tập tin phân tán phát triển bởi Google và ra đời trước HDFS. GFS có kiến trúc tương tự HDFS, là hình mẫu để cộng đồng phát triển nên HDFS. GFS thường được cấu hình với MasterNode và các shadow Master nhằm mục đích chịu lỗi. Trong quá trình hoạt động, nếu MasterNode gặp sự cố, một shadow Master sẽ được lựa chọn thay thế MasterNode. Quá trình này hoàn toàn trong suốt với client và người sử dụng. Ngoài ra, trong quá trình lưu trữ các block, GFS sử dụng các kỹ thuật kiểm tra lỗi lưu trữ như checksum nhằm phát hiện và khôi phục block bị lỗi một cách nhanh chóng. GFS là nền tảng phát triển các hệ thống khác của Google như BigTable hay Pregel.

2.3. CƠ SỞ DỮ LIỆU NOSQL

Các hệ quản trị dữ liệu quan hệ (RDBMS) từ lâu đã được xem như là giải pháp số một trong lưu trữ và truy vấn dữ liệu cấu trúc trong nhiều thập kỷ qua. RDBMS cung cấp mô hình dữ liệu quan hệ mà với mô hình này có thể đặc tả hầu hết mối quan hệ giữa các tập dữ liệu lưu trữ. Một trong những đặc tính cơ bản của RDBMS là bảo đảm ngữ nghĩa ACID cho phép xử lý các giao dịch dữ liệu một cách tin cậy. Điều này giải phóng ứng dụng khỏi khối lượng công việc khổng lồ đảm bảo tính toàn vẹn của dữ liệu trong truy cập tương tranh.

Khi lượng dữ liệu được lưu trữ ngày càng lớn và vượt ra khỏi giới hạn xử lý của một máy chủ RDBMS duy nhất, rất nhiều các kỹ thuật được đưa ra để mở rộng hiệu năng của hệ thống RDBMS. Một trong những kỹ thuật phổ biến là “database sharding” thực hiện bằng cách chia cơ sở dữ liệu tổng thể thành các phần “shards” và phân tán trên cụm RDBMS cluster. Tuy nhiên, do độ phức tạp cao của cơ chế bảo đảm ACID, các cụm RDBMS cluster thường chỉ được triển khai trên quy mô nhỏ vài chục nốt.

Từ năm 2005, cộng đồng nghiên cứu về cơ sở dữ liệu trên thế giới đồng ý rằng sự thống trị của RDBMS đã kết thúc và kêu gọi cần phải thiết kế các cơ sở dữ liệu chuyên biệt hóa với tính khả mở cao để phù hợp với nhu cầu thực tế. Lớp các hệ cơ sở dữ liệu mới này được gọi dưới tên chung là NoSQL. Hiện tại có các hệ cơ sở NoSQL điển hình như Amazon Dynamo, Cassandra, CouchDB,…

Đặc điểm chung của các hệ NoSQL là tính khả mở cao, so với RDBMS có những khác biệt cơ

bản như dưới đây:

Mô hình dữ liệu đơn giản hoá: NoSQL không tổ chức dữ liệu dưới các bảng quan hệ. NoSQL có mô hình tổ chức dữ liệu dưới bốn nhóm chính: key/value, hướng văn bản (document-oriented), hệ cột (column-familly store) và cơ sở dữ liệu đồ thị (Graph database).

Cơ sở dữ liệu (DBMS) key/value cung cấp mô hình dữ liệu và giao diện ứng dụng API đơn giản nhất giống với bảng băm (hashtable). Ứng với mỗi khóa, DBMS key/value chỉ cho phép đọc, ghi và xóa giá trị được định danh bởi khóa đó. Ví dụ kiểu NoSQL key/value là: Amazon Dynamo, Riak.

DBMS hướng văn bản được thiết kế để quản trị dữ liệu nửa cấu trúc (semistructured data) tổ chức dưới dạng tập hợp các văn bản. Giống như DBMS key/value, mỗi văn bản trong DBMS được định danh bằng một khóa duy nhất. Tuy nhiên, vì văn bản có cấu trúc gồm nhiều thuộc tính, DBMS hướng văn bản cho phép tạo chỉ mục tìm kiếm cho các trường thuộc tính này. Ví dụ của DBMS hướng văn bản bao gồm: CouchDB và MongoDB.

DBMS hệ cột tổ chức cấu trúc dữ liệu dưới dạng các bảng ánh xạ đa chiều thưa như mô hình Google Bigtable. Về cơ bản, dữ liệu tồn tại dưới dạng bảng như trong RDBMS nhưng các dòng trong bảng không nhất thiết có cùng tập các cột. Hơn nữa, DBMS hệ cột này phần lớn không hỗ trợ phép toán JOIN như trong DBMS.

Độ phức tạp được đơn giản hóa: RDBMS với đặc tính ACID đảm bảo dữ liệu luôn được toàn vẹn và nhất quán trong giao dịch. Tuy nhiên, đặc tính này trong thực tiễn với các ứng dụng internet như hiện nay trở nên không cần thiết. NoSQL lựa chọn tính khả mở thay vì cam kết bảo đảm ACID. Để có thể mở rộng tới mô hình triển khai phân tán trên hàng ngàn máy chủ lưu trữ, NoSQL phần lớn chỉ hỗ trợ mô hình nhất quán sau cùng (eventual consistency model) như là các thao tác đọc có thể được trả về dữ liệu cũ chưa được cập nhật với thay đổi mới với điều kiện đảm bảo sau cùng thì các thao tác đọc luôn nhận về dữ liệu mới nhất đã cập nhật.

Mô hình mở rộng ngang trên phần cứng phổ thông: Không giống như RDBMS, các DBMS NoSQL có tính khả mở cao trên các phần cứng phổ thông. Các nốt lưu trữ của NoSQL có thể gia nhập hay ra khỏi hệ thống tùy theo nhu cầu mà không làm ảnh hưởng đến sự hoạt động của hệ thống.

2.4. ĐIỆN TOÁN ĐÁM MÂY VÀ DỮ LIỆU LỚN

Các kho lưu trữ dữ liệu phân tán là thành phần không thể thiếu trong điện toán đám mây. Ngoài tính năng lưu trữ dữ liệu tập trung cho người sử dụng điện toán đám mây, các kho lưu trữ dữ liệu này còn dảm nhận vai trò lưu trữ và cung cấp các tập tin ảnh máy ảo phục vụ cho chính nền tảng ảo hóa của điện toán đám mây. Trong chương này, chúng ta sẽ đi vào tìm hiểu thiết kế các kho lưu trữ Openstack Swift, Amazon S3, thiết kế của các hệ quản trị dữ liệu mới phổ biến trên nền điện toán đám mây và phân tích cụ thể mô hình tính toán MapReduce/Hadoop.

Một đặc điểm chung của hệ quản trị dữ liệu cho điện toán đám mây là giao diện tương tác HTTP API, không sử dụng giao diện hệ thống quản lý tập tin thông thường.

Amazon S3

Amazon S3 là dịch vụ kho lưu trữ dữ liệu trên nền điện toán đám mây Amazon Web Service, được đưa ra giới thiệu vào năm 2006. S3 cung cấp giao diện tương tác ứng dụng API đơn giản cho phép lưu trữ và truy cập dữ liệu bất cứ khi nào và ở bất cứ đâu có kết nối Internet. Các nhà phát triển ứng dụng hay người sử dụng S3 không phải trả bất kỳ khoản phí cài đặt nào, chỉ trả phí dựa trên dung lượng lưu trữ và băng thông sử dụng.

Khi ra công bố, Amazon tính phí cho người sử dụng là 0,15 USD một gigabyte mỗi tháng, với chi phí thêm cho băng thông được sử dụng để gửi và nhận dữ liệu, tính phí cho mỗi yêu cầu (nhận hay gửi). Amazon tuyên bố ngay chính Amazon cũng sử dụng nền tảng S3 với tính khả mở cao để chạy mạng lưới thương mại điện tử toàn cầu của họ.

Amazon S3 được báo cáo đã lưu trữ hơn một nghìn tỷ thực thể tính đến tháng 6 năm 2012. Con số này tăng lên từ 102 tỷ thực thể trong tháng 3 năm 2010, 64 tỷ thực thể trong tháng 8 năm 2009, 52 tỷ tháng 3 năm 2009, 29 tỷ trong tháng 10 năm 2008, 14 tỷ trong tháng 1 năm 2008 và 10 tỷ trong tháng 10 năm 2007. S3 được sử dụng cho web hosting, image hosting và lưu trữ cho các hệ thống backup. Hợp đồng dịch vụ cho S3 kèm theo một đảm bảo 99,9% thời gian hoạt động hằng tháng, tương đương với khoảng 43 phút thời gian chết mỗi tháng.

Các khái niệm cơ sở trong Amazon S3

Đối tượng dữ liệu (Objects): Các đối tượng dữ liệu được coi như khái niệm các tập tin trong hệ thống quản lý tập tin thông thường. Mỗi đối tượng được lưu trữ với siêu dữ liệu đặc tả đi kèm như ngày chỉnh sửa, ngày khởi tạo,… Số lượng các đối tượng mà người sử dụng có thể lưu trữ là vô hạn và mỗi đối tượng có thể chứa 5 TB dữ liệu.

Thùng (buckets): Mỗi đối tượng được lưu trữ trong một thùng, được hiểu như là một thư mục trên hệ quản lý tập tin. Tuy nhiên, khác với hệ quản lý tập tin, các thùng chỉ chứa các đối tượng dữ liệu, không chứa các thùng con.

Khóa (keys): Mối đối tượng dữ liệu trong Amazon S3 được định danh bởi một khóa duy nhất ứng với thùng có chứa đối tượng đó. S3 hỗ trợ tính phiên bản, do vậy định danh của mỗi một phiên bản (version) của đối tượng dữ liệu được xây dựng từ tên thùng, khóa và mã phiên bản.

Vùng địa lý kho lưu trữ (Regions): Người sử dụng S3 khởi tạo thùng theo khu vực địa lý nơi triển khai hệ thống S3. Việc cho phép lựa chọn vùng lưu trữ là để tối ưu độ trễ đường truyền tăng tốc độ truy cập. S3 hiện nay được triển khai trên các vùng địa lý như Mỹ, châu Âu, khu vực châu Á,…

Thiết kế

Amazon không công bố thông tin chi tiết thiết kế của S3. Theo Amazon, thiết kế của S3 nhằm mục đích cung cấp khả năng mở rộng, tính sẵn sàng cao và độ trễ thấp với chi phí.

S3 lưu trữ các đối tượng hay thực thể dữ liệu kích thước lên đến 5 terabyte. Các đối tượng được tổ chức thành các thùng bucket (mỗi thùng thuộc sở hữu của một dịch vụ Web Amazon hoặc tài khoản AWS). Mỗi thực thể được định danh trong mỗi nhóm bằng một khóa duy nhất gán với người sử dụng. Các thùng và các đối tượng có thể được tạo ra, được liệt kê và lấy ra bằng cách sử dụng hoặc một giao diện HTTP kiểu REST hoặc giao diện SOAP. Ngoài ra, các đối tượng có thể được tải về bằng cách sử dụng giao diện giao thức BitTorrent.

S3 hỗ trợ các giao thức bảo mật, chính sách quyền truy cập đến các đối tượng và các thùng. Tên và khóa thùng được lựa chọn nên các đối tượng có thể được đánh địa chỉ thông qua URL:

http://s3.amazonaws.com/bucket/key

http://bucket.s3.amazonaws.com/key

Vì các đối tượng có thể truy cập qua giao thức HTTP, S3 có thể được sử dụng để thay thế đáng kể cơ sở hạ tầng lưu trữ web tĩnh hiện có. Cơ chế xác thực AWS Amazon cho phép chủ sở hữu nhóm tạo ra một URL xác thực với thời gian tồn tại được định sẵn. Ví dụ như, người sử dụng có thể xây dựng một URL mà có thể được giao cho một bên thứ ba để truy cập trong một

thời gian như trong 30 phút tiếp theo, hoặc 24 giờ tiếp theo.

Mỗi mục trong một thùng cũng có thể được phục vụ như là một nguồn cấp dữ liệu BitTorrent. Các kho dữ liệu của S3 có thể hoạt động như một máy chủ lưu trữ nguồn cho một torrent và client BitTorrent bất kỳ có thể lấy lại tập tin qua giao thức BitTorrent. Điều này làm giảm đáng kể chi phí băng thông cho việc tải xuống của các thực thể phổ biến vì các client BitTorrent trong quá trình tải về đối tượng dữ liệu cũng đóng góp vào việc làm trung gian đưa dữ liệu tới các client khác.

Dịch vụ lưu trữ web tĩnh

Tính đến ngày 18 tháng 2 năm 2011, S3 Amazon cung cấp các tùy chọn để lưu trữ các trang web tĩnh với sự hỗ trợ tài liệu trang chủ chỉ mục (index.html) và hỗ trợ tài liệu báo lỗi. Sự hỗ trợ này đã được thêm vào như một kết quả từ yêu cầu của người sử dụng từ năm 2006. Ví dụ, trong trường hợp Amazon S3 đã được cấu hình với các bản ghi CNAME để lưu trữ http://subdomain.example.com/. Trong quá khứ, một người truy cập vào URL này sẽ chỉ là một danh sách định dạng XML của các đối tượng thay vì một trang chủ index để phù hợp với khách truy cập ngẫu nhiên. Tuy nhiên, ngày nay, các trang web lưu trữ trên S3 có thể chỉ định một trang chủ mặc định để hiển thị và một trang khác để hiển thị trong trường hợp trang mặc định này không tồn tại.

Ưu điểm và tính năng của S3

Tính ổn định: Người sử dụng S3 được ký thỏa thuận cung cấp dịch vụ với mức duy trì tính sẵn sàng của hệ thống đạt 99,99%. S3 được thiết kế chịu lỗi và phục hồi hệ thống nhanh trong thời gian tối thiểu.

Tính đơn giản, dễ dùng: S3 cung cấp giao diện kết nối Rest API và các thư viện giao tiếp trên các ngôn ngữ phổ biến như JAVA, Python,...

Tính mở rộng: S3 chỉ tính phí dịch vụ trên dung lượng và băng thông sử dụng. Người sử dụng có toàn quyền tăng hoặc giảm số lượng các đối tượng dữ liệu được lưu trữ để tối ưu chi phí sử dụng.

Tính rẻ: Theo tính toán thống kê, chi phí sử dụng S3 rất cạnh tranh với các giải pháp tự thiết kế lưu trữ trên máy chủ riêng.

OpenStack Swift

OpenStack Swift là hệ thống kho lưu trữ đối tượng có tính khả mở cao, được thiết kế để lưu trữ khối lượng lớn các dữ liệu phi cấu trúc với chi phí thấp thông qua giao diện ứng dụng Restful. Hệ thống OpenStack Swift có thể triển khai mở rộng từ một vài nốt (node) lưu trữ với dung lượng lưu trữ giới hạn lên tới hàng ngàn nốt lưu trữ phân tán với dung lượng tổng thể lên tới hàng ngàn Petabytes. OpenStack Swift được thiết kế chịu lỗi, trong hệ thống không có bất cứ thành phần nào là điểm chết duy nhất (single point of failure) mà hỏng hóc của thành phần này kéo theo việc dừng hoạt động của toàn bộ hệ thống. Như các kho lưu trữ điện toán đám mây khác, OpenStack swift được thiết kế để có thể lưu trữ và phục vụ dữ liệu nội dung đồng thời nhiều người dùng dịch vụ.

Đặc trưng của OpenStack Swift

OpenStack Swift lưu trữ dữ liệu dưới dạng các đối tượng dữ liệu được truy xuất qua định danh URL.

Tất cả các đối tượng lưu trữ đều được nhân bản ra nhiều bản vùng chịu lỗi được định nghĩa

như là nhóm các ổ đĩa cứng, nhóm các nốt lưu trữ,… Mục đích của việc nhân bản ra các vùng chịu lỗi khác nhau là để giúp đối tượng lưu trữ vẫn có thể truy cập tới được khi có sự cố trên một hoặc hai vùng chịu lỗi.

Tất cả các đối tượng có thông tin siêu dữ liệu đi kèm như quyền người sử dụng, thời gian khởi tạo...

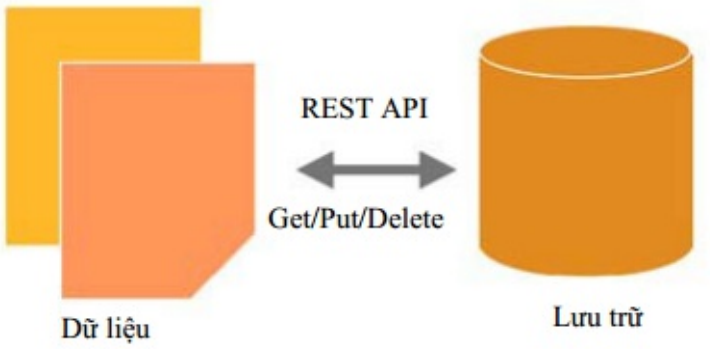
OpenStack Swift cung cấp giao tiếp RestFul API để truy xuất đối tượng dữ liệu.

Đối tượng dữ liệu được phân tán trên hệ thống mà không phụ thuộc vào định danh URL của đối tượng.

OpenStack Swift mở rộng hay thu hẹp hệ thống bằng việc thêm bớt các nốt lưu trữ nhưng không ảnh hưởng đến hiệu năng của hệ thống. Quá trình thêm bớt các nốt là hoàn toàn trong suốt và không ảnh hưởng đến hoạt động của toàn hệ thống.

Các nốt lưu trữ là phần cứng máy chủ phổ biến như Dell, HP,...

Các lập trình viên hoặc người sử dụng nền tảng OpenStack Swift có thể sử dụng trực tiếp giao tiếp API cung cấp bởi OpenStack Swift, hoặc sử dụng các thư viện giao tiếp trên các ngôn ngữ lập trình phổ biến như Java, Python, Ruby và C#.



Hình 2.5. Truy cập dữ liệu qua Rest API

Xử lý dữ liệu lớn MapReduce/Hadoop

Theo thống kê, lượng dữ liệu được tạo ra hiện nay tăng gấp đôi sau chu kỳ 2 năm. Đóng góp vào sự tăng trưởng khổng lồ này là các văn bản, tập tin video, audio được chia sẻ trên các mạng xã hội như twitter, facebook. Ngoài ra còn phải kể đến dữ liệu sinh ra trong nghiên cứu khoa học và các loại dữ liệu lấy từ thiết bị cảm biến lắp đặt khắp nơi trên thế giới. Trong ngữ cảnh này, từ khóa "big data" hay dữ liệu lớn trở thành từ khóa được nhắc đến nhiều nhất trong giới công nghệ. Dữ liệu lớn đặt ra nhiều thách thức cho việc thiết kế lại hệ thống lưu trữ và xử lý, cũng như mang lại nhiều lợi ích nếu được khai thác như một nguồn tài nguyên trong kỷ nguyên số.

Dữ liệu lớn được đặc tả bằng 4V: Volume, Velocity, Variety và Value. Volume nói tới dung lượng khổng lồ của dữ liệu lớn, velocity nói tới lớp dữ liệu lớn đòi hỏi phải xử lý nhanh và liên tục. Variety tham chiếu tới sự đa dạng của các loại dữ liệu cấu trúc và phi cấu trúc tồn tại hiện nay. Value là về giá trị quý báu có thể có được khi khai thác dữ liệu lớn.

Một vấn đề cần phải thấy, đó là điện toán đám mây là nền tảng cơ bản cho phép lưu trữ và khai thác dữ liệu lớn trở nên phổ biến. Như đã phân tích, với ứng dụng điện toán đám mây, các doanh nghiệp tổ chức có thể thuê máy chủ hoặc dịch vụ lưu trữ phân tích dữ liệu lớn với chi phí thấp mà không phải đầu tư hạ tầng và con người để quản trị hệ thống. Trong phần này,

chúng ta sẽ đi vào tìm hiểu xử lý dữ liệu lớn với Hadoop MapReduce.

MapReduce là một trong những công nghệ tạo khả năng của cuộc cách mạng dữ liệu lớn, một mô hình lập trình và công cụ được Google đưa ra giới thiệu vào năm 2004. Mapreduce có thể hiểu là một phương thức thực thi để giúp các ứng dụng có thể xử lý nhanh một lượng dữ liệu lớn trên môi trường phân tán. Các máy tính này sẽ hoạt động song song nhưng độc lập với nhau, mục đích là làm rút ngắn thời gian xử lý toàn bộ dữ liệu.

Ưu điểm của MapReduce

– Xử lý tốt bài toán về lượng dữ liệu lớn có các tác vụ phân tích và tính toán phức tạp mà không thể xử lý tốt trên môi trường một máy tính toán.

– Giải thuật MapReduce gồm hai bước cơ bản, Map và Reduce đơn giản nhưng có thể đặc tả hầu hết các bài toán trên môi trường phân tán.

– Mô hình thực hiện MapReduce có khả năng chịu lỗi cao. Khi có lỗi xảy ra tại một nút tính toán trong hệ thống, tính toán tại nút đó sẽ được thực hiện lại mà không ảnh hưởng tới tính toán trên các nút khác.

– Tính toán MapReduce được phân tán trên các nút lưu trữ. So với các mô hình tính toán khác mà dữ liệu được sao chép đến các nút tính toán và thực hiện trên các nút đó, mô hình tính toán MapReduce khác biệt ở chỗ mã chương trình được sao chép tới các nút lưu trữ để thực thi. Đây là một trong những điểm mấu chốt tiên tiến của MapReduce vì quan điểm di chuyển mã chương trình tiết kiệm và hiệu quả hơn di chuyển dữ liệu mà có thể lên tới hàng TB.

– Nền tảng tính toán MapReduce được thiết kế để thực thi với các máy chủ phổ thông, không cần năng lực tính toán và lưu trữ lớn như mô hình tính toán song song MPI. Điều này đạt được nhờ vào thiết kế chịu lỗi cao.

Nguyên tắc hoạt động của MapReduce

MapReduce hoạt động trên một nguyên tắc đơn giản. Các phép toán lấy đầu vào là một tập các cặp khóa/giá trị và đưa ra một tập khóa/giá trị đầu ra. MapReduce biểu diễn tính toán chỉ bằng hai hàm: Map và Reduce.

Hàm Map, lấy đầu vào là một cặp khóa/giá trị và đưa đầu ra là một tập các cặp khóa/giá trị trung gian. Các cặp khóa/giá trị trung gian này sau đó được gộp lại và các cặp khóa/giá trị trung gian có cùng khóa sẽ được chuyển cho hàm Reduce. Từ đó hàm Reduce tính toán trên các cặp này nhằm đưa ra các giá trị tổng quan hơn là kết quả cuối cùng.

Tính toán MapReduce trong thực tiễn

Với nguyên tắc hoạt động của MapReduce giới thiệu bên trên, có nhiều phương pháp hiện thực hóa mô hình tính toán MapReduce như phát triển nền tảng trên các máy tính chia sẻ bộ nhớ trong, trên các máy tính đa xử lý NUMA, trên cluster tính toán... Trong phần này, chúng ta sẽ đi vào tìm hiểu tính toán MapReduce triển khai trên cluster. Một cluster gồm hàng trăm tới hàng ngàn máy chủ phổ thông kết nối với nhau qua mạng LAN thông thường 100 Mbs hoặc 1 Gbs. Phần cứng lưu trữ trên các máy chủ này không đòi hỏi hiệu năng cao, chỉ là các đĩa cứng thông thường kết nối qua chuẩn IDE. Đơn vị công việc trong chương trình MapReduce gọi là gói công việc (job), mỗi job gồm nhiều tác vụ được điều chuyển thông qua hệ thống phân phối chung tới các máy chủ thuộc cluster.

Tác vụ Map được thực hiện phân tán trên các nút lưu trữ. Quá trình phân tán được thực hiện tự động thông qua việc dữ liệu đầu vào được chia nhỏ. Tác vụ Reduce cũng được phân tán

thông qua việc các cặp khóa/giá trị trung gian được nhóm lại thành các cặp có khóa giống nhau. Luồng thực thi một gói công việc MapReduce bao gồm các bước sau (hình 2.6):

– MapReduce cluster về cơ bản gồm một nút Master và các nút Worker. Nút Master làm nhiệm vụ quản lý và điều tiết các Worker.

– Thư viện MapReduce phía người sử dụng gửi mã biên dịch của tác vụ Map tới các nút tính toán thường là các nút chứa dữ liệu của tệp tin đầu vào. Số lượng các tác vụ Map được chạy song song thường bằng số các block của tệp đầu vào.

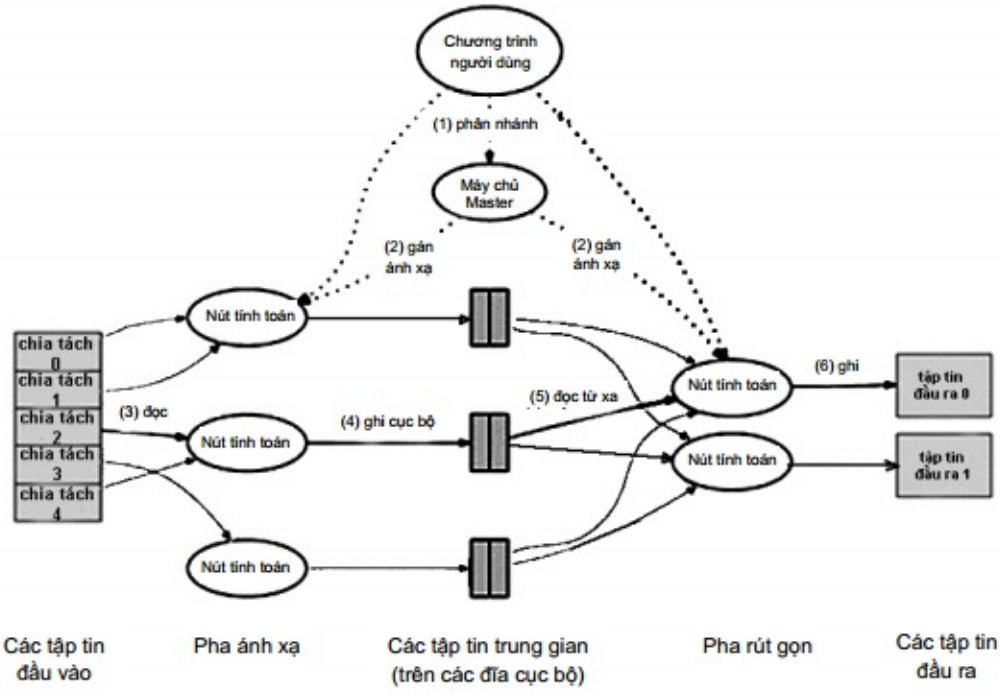
– Các nút tính toán là các worker đọc dữ liệu từ tệp tin đầu vào, thường là từ chính block được lưu trữ trên ổ đĩa cục bộ. Tác vụ Map sinh ra tập các cặp khóa/giá trị trung gian lưu trên bộ nhớ trong.

– Theo định kỳ được cấu hình, các cặp khóa/giá trị trung gian này được ghi vào ổ đĩa cứng cục bộ, chia thành R nhóm (R là số lượng các nút chạy tác vụ Reduce). Vị trí của các cặp này được thông báo cho máy chủ Master để Master làm nhiệm vụ đưa lại địa chỉ này cho các máy chủ làm tác vụ Reduce.

– Khi Worker thực hiện tác vụ Reduce được thông báo về vị trí các cặp khóa/giá trị trung gian, các Worker này sử dụng RPC để đọc dữ liệu này. Khi quá trình đọc kết thúc, Worker sắp xếp lại các cặp khóa/giá trị trung gian theo khóa.

– Worker thực hiện Reduce tính toán tuần tự trên các dữ liệu được sắp xếp. Đầu ra của tác vụ Reduce được ghi vào tệp tin đầu ra.

– Khi tất cả các tác vụ Map và Reduce được kết thúc, Master thông báo lại kết quả cho chương trình của người sử dụng.



Hình 2.6. Mô tả quá trình thực thi gói công việc

2.5. CÂU HỎI VÀ BÀI TẬP

1. Nêu các đặc điểm của hệ thống lưu trữ phân tán.

2. Nêu các đặc điểm của hệ thống lưu trữ HDFS.
3. Phân biệt và so sánh cơ sỡ dữ liệu NOSQL và hệ cơ sở dữ liệu quan hệ.
4. Tìm hiểu về kiến trúc Amazon S3 và so sánh với OpenStack swift.
5. So sánh kỹ thuật lưu trữ trên điện toán đám mây với kỹ thuật lưu trữ truyền thống.
6. Phân tích ưu nhược điểm của HDFS.
7. So sánh mô hình thực thi MapReduce và mô hình thực thi MPI.
8. Phân tích nhược điểm của MapReduce.
9. Giới thiệu các xu thế tính toán cải tiến MapReduce.

Chương 3
AN TOÀN VÀ BẢO MẬT

Trong một hệ thống thông tin, an toàn và bảo mật (ATBM) là sự đảm bảo tính bí mật, tính toàn vẹn và tính sẵn dùng của hệ thống dưới những đe dọa đến từ các sự cố phần cứng/phần mềm hoặc đến từ sự tấn công có chủ đích của con người. Với các hệ thống đám mây, nơi dữ liệu và các tiến trình xử lý thông tin của người dùng được giao phó cho sự quản lý của các nhà cung cấp dịch vụ, thì vấn đề đảm bảo tính an toàn và bảo mật càng trở nên cấp thiết. Chương này sẽ trình bày những vấn đề đặt ra về ATBM và giới thiệu một số giải pháp đảm bảo tính ATBM cho các hệ thống đám mây. Các chủ đề chính của chương bao gồm: Các vấn đề về ATBM trong điện toán đám mây; Một số phương pháp đảm bảo tính ATBM cho các dịch vụ đám mây; Giải pháp thiết kế kiến trúc hệ thống đám mây nhằm đảm bảo ATBM.

3.1. CÁC VẤN ĐỀ VỀ AN TOÀN VÀ BẢO MẬT TRONG ĐIỆN TOÁN ĐÁM MÂY

Điện toán đám mây được xem như giải pháp giúp khách hàng tiết kiệm được nhiều chi phí đầu tư cũng như công sức quản lý và vận hành hệ thống. Tuy vậy, do tính chất phân tán và trực tuyến, tích hợp nhiều tầng dịch vụ với nhiều công nghệ đặc thù, giải pháp này đồng thời cũng bộc lộ nhiều nguy cơ mới về ATBM. Năm 2009, tổ chức khảo sát và phân tích thị trường International Data Corporation (IDC) tiến hành khảo sát ý kiến của các giám đốc thông tin (CIO) từ nhiều công ty hàng đầu về những trở ngại trong việc chuyển đổi sang mô hình dịch vụ đám mây. Kết quả khảo sát (minh họa trong hình 3.1) cho thấy vấn đề ATBM đang là lo lắng hàng đầu của các tổ chức thuê dịch vụ đám mây.

Trong phần này, chúng tôi trước tiên tóm tắt một số vấn đề liên quan tới ATBM trên các tầng dịch vụ khác nhau của điện toán đám mây. Sau đó giới thiệu một số lỗ hổng ATBM và nguy cơ về ATBM trên các hệ thống đám mây.

Các vấn đề về an toàn và bảo mật trên các tầng dịch vụ đám mây

Như đã đề cập đến trong chương 1, mô hình điện toán đám mây cung cấp ba tầng dịch vụ chính: dịch vụ phần mềm (SaaS), dịch vụ nền tảng (PaaS) và dịch vụ hạ tầng (IaaS).

Với SaaS, gánh nặng về ATBM thuộc về phía nhà cung cấp dịch vụ. Các dịch vụ phần mềm thường đặt trọng tâm vào việc tích hợp chức năng đồng thời tối thiểu hóa khả năng can thiệp và mở rộng của người sử dụng. Trong khi đó, các dịch vụ nền tảng hỗ trợ nhiều hơn khả năng can thiệp và mở rộng. Và dưới cùng, các dịch vụ hạ tầng cho phép người sử dụng có khả năng can thiệp và đồng thời chịu trách nhiệm lớn nhất về ATBM.

Hình 3.1. Kết quả khảo sát của IDC về những quan ngại của khách hàng với mô hình

điện toán đám mây

An toàn và bảo mật trong các dịch vụ phần mềm

SaaS cung cấp các dịch vụ phần mềm theo nhu cầu như thư điện tử, hội thảo trực tuyến, ERP, CRM,… Nội dung tiếp theo sẽ trình bày một số vấn đề về an toàn và bảo mật trong tầng dịch vụ này.

Vấn đề 1. Bảo mật ứng dụng

Người sử dụng thường truy nhập các ứng dụng đám mây thông qua trình duyệt web. Sai sót trong các trang web có thể tạo nên những lỗ hổng của dịch vụ SaaS. Tin tặc từ đó có thể gây thương tổn tới các máy tính của người sử dụng để thực hiện các hành vi ác ý hoặc ăn trộm các thông tin nhạy cảm. ATBM trong dịch vụ SaaS không khác với trong các ứng dụng web thông thường. Có thể tham khảo thêm về những vấn đề ATBM ứng dụng web thông qua bài viết "The most critical web application security risks" do OWASP (Open Web Application Security Project) xuất bản.

Vấn đề 2. Nhiều người thuê đồng thời (multi-tenancy)

Các dịch vụ SaaS có thể được xây dựng theo ba mô hình:

– Mô hình khả mở (scalability model): Mỗi người sử dụng được cấp một thể hiện đã chuyên biệt hóa của phần mềm. Mặc dù có nhiều nhược điểm, nhưng mô hình này lại không tạo nên những vấn đề lớn về an toàn và bảo mật.

– Mô hình cấu hình qua siêu dữ liệu (configurability via metadata): Mỗi người cũng có một thể hiện riêng của phần mềm, tuy nhiên các thể hiện này dùng chung một mã nguồn, sự khác biệt chỉ là cấu hình của phần mềm thông qua các siêu dữ liệu.

– Mô hình nhiều người thuê đồng thời: Trong mô hình này, một thể hiện duy nhất của ứng dụng được chia sẻ cho nhiều người thuê. Khi đó tài nguyên sẽ được sử dụng hiệu quả (mặc dù tính khả mở sẽ giảm đi). Trong mô hình này, do dữ liệu của các người dùng được lưu trữ trên cùng một cơ sở dữ liệu nên nguy cơ về rò rỉ dữ liệu có thể xảy ra.

Vấn đề 3. Bảo mật dữ liệu

Trong SaaS, dữ liệu thường được xử lý dưới dạng bản rõ và được lưu trữ trên đám mây. Nhà cung cấp dịch vụ SaaS sẽ phải chịu trách nhiệm về bảo mật dữ liệu trong khi chúng được xử lý và lưu trữ. Việc sao lưu dữ liệu rất phổ biến trong các hệ thống đám mây cũng tạo nên những vấn đề phát sinh cho bảo mật dữ liệu, nhất là khi nhà cung cấp dịch vụ ký hợp đồng sao lưu lại với một đối tác thứ ba không đáng tin cậy.

Vấn đề 4. Truy cập dịch vụ

Việc các dịch vụ SaaS hỗ trợ khả năng truy cập thông qua trình duyệt mang lại nhiều thuận lợi, ví dụ như chúng có thể dễ dàng được truy cập từ các thiết bị kết nối mạng khác PC như điện thoại hay máy tính bảng. Tuy nhiên, điều này lại tạo nên những nguy cơ mới về ATBM như các phần mềm ăn trộm dữ liệu trên di động, mạng Wifi không an toàn, kho ứng dụng không an toàn,…

An toàn và bảo mật trong các dịch vụ nền tảng

PaaS hỗ trợ việc xây dựng các ứng dụng đám mây mà không cần quan tâm tới vấn đề thiết lập và duy trì hạ tầng phần cứng hay môi trường phần mềm. Vấn đề ATBM trong PaaS liên quan tới hai khía cạnh: bảo mật trong nội tại của dịch vụ PaaS và bảo mật trong phần mềm của

khách hàng triển khai trên nền dịch vụ PaaS. Nội dung tiếp theo giới thiệu một số vấn đề về ATBM trong tầng dịch vụ nền tảng.

Vấn đề 5. An toàn và bảo mật của bên thứ ba

Dịch vụ PaaS thường không chỉ cung cấp môi trường phát triển ứng dụng của nhà cung cấp dịch vụ mà đôi khi cho phép sử dụng những dịch vụ mạng của bên thứ ba. Những dịch vụ này thường được đóng gói dưới dạng thành phần trộn (mashup). Chính vì vậy, ATBM trong các dịch vụ PaaS cũng phụ thuộc vào ATBM của chính các mashup này.

Vấn đề 6. Vòng đời của ứng dụng

Giống như các loại hình ứng dụng khác, các ứng dụng trên dịch vụ đám mây cũng có thể liên tục nâng cấp. Việc nâng cấp ứng dụng đòi hỏi nhà cung cấp dịch vụ PaaS phải hỗ trợ tốt cho những thay đổi của ứng dụng. Đồng thời, người phát triển cũng cần phải lưu ý rằng sự thay đổi của các thành phần ứng dụng trong quá trình nâng cấp đôi khi gây ra các vấn đề về ATBM.

An toàn và bảo mật trong các dịch vụ hạ tầng

IaaS cung cấp một vùng chứa tài nguyên như máy chủ, mạng, kho lưu trữ bao gồm cả các tài nguyên được ảo hóa. Khách hàng có toàn quyền kiểm soát và quản lý các tài nguyên họ thuê được. Các nhà cung cấp dịch vụ IaaS phải bảo vệ hệ thống khỏi những ảnh hưởng liên quan tới vấn đề ATBM phát sinh từ các tài nguyên cho thuê của khách hàng. Nội dung tiếp theo liệt kê một số vấn đề về ATBM trong tầng dịch vụ IaaS.

Vấn đề 7. Ảo hóa

Công nghệ ảo hóa cho phép người sử dụng dễ dàng tạo lập, sao chép, chia sẻ, di trú và phục hồi các máy ảo trên đó thực thi các ứng dụng. Công nghệ này tạo nên một tầng phần mềm mới trong kiến trúc phần mềm của hệ thống. Chính vì vậy nó cũng mang đến những nguy cơ mới về ATBM.

Vấn đề 8. Giám sát máy ảo

Thành phần giám sát máy ảo (virtual machine monitor – VMM) hay còn gọi là supervisor có trách nhiệm giám sát và quản lý các máy ảo được tạo trên máy vật lý. Chính vì vậy, nếu VMM bị tổn thương, các máy ảo do nó quản lý cũng có thể bị tổn thương. Di trú máy ảo từ một VMM này sang một VMM khác cũng tạo nên những nguy cơ mới về ATBM.

Vấn đề 9. Tài nguyên chia sẻ

Các máy ảo trên cùng một hệ thống chia sẻ một số tài nguyên chung như CPU, RAM, thiết bị vào ra,... Việc chia sẻ này có thể làm giảm tính ATBM của mỗi máy ảo. Ví dụ, một máy ảo có thể đánh cắp thông tin của máy ảo khác thông qua bộ nhớ chia sẻ. Hơn nữa nếu khai thác một số kênh giao tiếp ngầm giữa các máy ảo, các máy ảo có thể bỏ qua mọi quy tắc bảo mật của VMM.

Vấn đề 10. Kho ảnh máy ảo công cộng

Trong môi trường IaaS, ảnh máy ảo là một mẫu sẵn có để tạo nên các máy ảo. Một hệ thống đám mây có thể cung cấp một số ảnh máy ảo trong một kho công cộng để người dùng có thể dễ dàng tạo nên máy ảo theo nhu cầu của mình. Đôi khi hệ thống đám mây có thể cho phép người sử dụng tự tải ảnh máy ảo của mình lên kho công cộng này. Điều này tạo nên một nguy cơ về bảo mật khi tin tặc tải lên những ảnh máy ảo có chứa mã độc và người sử dụng có thể dùng những ảnh này để tạo máy ảo của họ. Hơn nữa, qua việc tải ảnh máy ảo lên kho công cộng, người dùng cũng có nguy cơ mất đi những dữ liệu nhạy cảm của mình trong ảnh máy ảo đã tải.

Ảnh máy ảo cũng tạo ra nguy cơ về bảo mật khi chúng không được vá lỗi giống như các hệ thống đang vận hành.

Vấn đề 11. Phục hồi máy ảo

Người sử dụng có thể phục hồi máy ảo về một trạng thái đã được lưu trữ trước đó. Tuy nhiên, nguy cơ về ATBM lại phát sinh khi những lỗi được vá mới không áp dụng cho trạng thái máy ảo cũ.

Vấn đề 12. Mạng ảo

Mạng ảo có thể được chia sẻ bởi nhiều người thuê trong một vùng chứa tài nguyên. Khi đó, các vấn đề ATBM có thể phát sinh giữa các người thuê chia sẻ chung mạng ảo này như việc một máy ảo có thể nghe trộm các bản tin gửi cho máy ảo khác trên cùng mạng.

Một số lỗ hổng về an toàn và bảo mật trong các hệ thống đám mây

Để giải quyết những vấn đề về ATBM đã đặt ra như trong phần trước, công việc đầu tiên của các nhà cung cấp dịch vụ đám mây là phải xác định được những lỗ hổng có thể tồn tại về ATBM trong hệ thống của mình. Bảng 3.1 tổng kết một số lỗ hổng điển hình về ATBM trong các hệ thống đám mây.

Bảng 3.1. Các lỗ hổng an toàn bảo mật trong hệ thống đám mây

STT	*Lỗ hổng*	*Mô tả*
1	Giao diện người sử dụng và API được cung cấp không an toàn	Phần lớn các nhà cung cấp dịch vụ đám mây cung cấp dịch vụ thông qua các giao diện HTTP, SOAP, hoặc REST. Những vấn đề bảo mật gắn với các giao diện này bao gồm: – Chứng nhận sử dụng hợp lệ yếu; – Việc kiểm tra xác quyền không đủ; – Thiếu kiểm tra tính hợp lệ của dữ liệu.
2	Tài nguyên phân bổ không giới hạn	– Hệ thống có thể gặp phải tình huống dành sẵn quá nhiều tài nguyên nếu như việc đánh giá về tài nguyên sử dụng không chính xác.
3	Các lỗ hổng liên quan tới dữ liệu	– Khả năng phân tách yếu cho những dữ liệu có nguồn gốc khác nhau (chẳng hạn của các đối tượng cạnh tranh, hoặc của tin tặc) dẫn đến chúng dễ bị đánh cắp. – Dữ liệu không được xóa hoàn toàn. – Dữ liệu được sao lưu bởi một bên thứ ba không tin cậy. – Người sử dụng thường không biết vị trí lưu trữ dữ liệu. – Dữ liệu thường được lưu trữ, lưu chuyển và xử lý dưới dạng bản rõ.

4	Lỗ hổng trong máy ảo	– Tồn tại những kênh giao tiếp không tường minh. – Cấp phát và giải phóng tài nguyên không hạn chế trong máy ảo. – Di trú không được kiểm soát. – Không kiểm soát các snapshot có thể dẫn đến rò rỉ dữ liệu. – Các máy ảo có IP có thể quan sát được bên trong đám mây, do vậy tin tặc có thể định vị được một máy ảo cần tấn công.
5	Lỗ hổng trong ảnh máy ảo	– Ảnh máy ảo được đặt trên kho lưu trữ công cộng một cách không kiểm soát. – Ảnh máy ảo không thể vá lỗi vì chúng không hoạt động.
6	Lỗ hổng trong hypervisor	– Khả năng cấu hình linh động của hypervisor có thể khiến chúng bị khai thác.
7	Lỗ hổng trong mạng ảo	– Lỗ hổng khi chia sẻ các cầu nối ảo giữa các máy ảo.

Những nguy cơ về an toàn và bảo mật trong các hệ thống đám mây

Tháng 11 năm 2008, liên minh an toàn bảo mật trong điện toán đám mây (Cloud Security Alliance – CSA) được thành lập dưới hình thức một tổ chức phi lợi nhuận. Nhiệm vụ chính của CSA là xác định các vấn đề liên quan tới ATBM đám mây, sau đó cung cấp những kinh nghiệm và giải pháp hỗ trợ giải quyết các vấn đề đó. Tổ chức này đã nhận được sự ủng hộ của trên một trăm hai mươi nhà cung cấp dịch vụ đám mây, bao gồm những nhà cung cấp hàng đầu như Google, Amazon hay Saleforce.

Năm 2013, trong tài liệu "The Notorious Nine: Cloud Computing Top Threats in 2013", CSA công bố 9 nguy cơ lớn nhất về ATBM trong các hệ thống đám mây. Các nguy cơ này bao gồm:

– *Rò rỉ dữ liệu*. Rò rỉ dữ liệu là việc dữ liệu của người dùng hoặc tổ chức thuê dịch vụ đám mây bị thất thoát vào tay những đối tượng không mong đợi. Đây có lẽ là một trong những đe dọa nghiêm trọng nhất đối các tổ chức sử dụng dịch vụ. Trong một hệ thống đám mây, việc tích hợp những công nghệ mới tạo nên những nguy cơ mới về thất thoát dữ liệu. Ví dụ, vào tháng 11 năm 2012, các nghiên cứu viên ở trường Đại học North Carolina, trường Đại học Wisconsin và tổ chức RSA đã công bố một công trình, trong đó mô tả phương thức sử dụng một máy ảo để trích xuất các khóa riêng tư từ máy ảo khác trên cùng một máy vật lý.

– *Mất mát dữ liệu*. Mất mát dữ liệu là việc dữ liệu của người dùng hoặc tổ chức thuê dịch vụ bị phá hủy hoặc không thể truy nhập được. Đối với khách hàng, mất mát dữ liệu là một điều tồi tệ, nó không những khiến khách hàng mất đi thông tin mà đôi khi khiến các hoạt động của khách hàng trên hệ thống dịch vụ bị gián đoạn hoặc thậm chí sụp đổ. Nguyên nhân cho việc mất mát dữ liệu có thể đến từ những tấn công của tin tặc; từ trục trặc của hệ thống phần mềm/phần cứng; hoặc do các thảm họa như cháy nổ, động đất. Đôi khi mất mát dữ liệu còn do phía người dùng, ví dụ như khi người dùng gửi bản mã hóa của dữ liệu lên đám mây nhưng lại quên mất khóa để giải mã chúng.

– *Bị đánh cắp tài khoản hoặc thất thoát dịch vụ*. Hiện tượng người sử dụng bị đánh cắp tài khoản hoặc thất thoát dịch vụ có thể nói khá phổ biến trong các loại hình dịch vụ trực tuyến.

Nhiều người sử dụng bị đánh cắp tài khoản do "dính bẫy" phishing hoặc do các lỗ hổng phần mềm trong hệ thống bị tin tặc khai thác. Môi trường điện toán đám mây đang là miền đất mới cho các kỹ thuật tấn công dạng này. Khi quyền truy nhập của một hệ thống dịch vụ đám mây rơi vào tay tin tặc, chúng có thể can thiệp vào các hoạt động của hệ thống, thay đổi các giao dịch của hệ thống, dẫn hướng khách hàng của hệ thống tới những liên kết của chúng, biến tài nguyên của khách hàng trên đám mây thành một căn cứ tấn công mới của chúng,…

Một ví dụ điển hình là sự kiện dịch vụ đám mây của Amazon gặp lỗi XSS (Cross-site Scripting) vào tháng 4 năm 2010. Lỗi này khiến khách hàng mất đi quyền truy nhập vào hệ thống và tài nguyên của khách hàng trên hệ thống trở thành các botnet của mạng lưới tấn công Zeus.

– *Giao diện và API không an toàn*. Các nhà cung cấp dịch vụ đám mây thường cung cấp cho khách hàng một tập giao diện phần mềm (API) nhằm giúp khách hàng có thể quản lý và tương tác với dịch vụ một cách tự động. Các API được tổ chức thành nhiều nhóm theo từng tầng dịch vụ. API thuộc các tầng khác nhau phụ thuộc vào nhau giống như sự phụ thuộc giữa các tầng dịch vụ. Khi lỗ hổng bảo mật trong các API bị tin tặc khai thác, tính ATBM của hệ thống sẽ bị xâm phạm. Lỗ hổng trong các API tầng thấp sẽ ảnh hưởng đến các API thuộc tầng cao hơn. Do vậy, vấn đề ATBM của hệ thống đám mây gắn bó mật thiết tới việc bảo mật cho các API này.

Bên cạnh đó, các tổ chức sử dụng dịch vụ đám mây đôi khi tự xây dựng những tầng dịch vụ mới cho khách hàng của họ dựa trên các API của nhà cung cấp dịch vụ đám mây. Điều này càng làm tăng thêm những rủi ro về ATBM từ hệ thống API của đám mây.

– *Từ chối dịch vụ*. Tấn công từ chối dịch vụ là cách thức hạn chế khả năng truy nhập vào dữ liệu và ứng dụng của người sử dụng dịch vụ. Phương thức thường dùng trong việc tấn công từ chối dịch vụ là việc tạo ra một số lượng yêu cầu lớn bất thường tới các dịch vụ bị tấn công khiến cho tài nguyên hệ thống (RAM, CPU, HDD, băng thông,…) cạn kiệt. Khi đó hệ thống trở nên chậm chạp, đáp ứng kém hoặc không đáp ứng được các yêu cầu từ khách hàng khiến cho họ bất bình và quay lưng lại với dịch vụ.

– *Nguy cơ từ bên trong*. Nguy cơ từ bên trong ám chỉ những nguy cơ đến từ các cá nhân có ác ý nằm trong tổ chức cung cấp dịch vụ, ví dụ như một quản trị viên của hệ thống đám mây. Khi các đối tượng này có quyền truy nhập vào mạng, dữ liệu, các máy chủ của hệ thống đám mây, các dữ liệu quan trọng của khách hàng có thể bị đánh cắp; ứng dụng của khách hàng có thể bị sửa đổi khiến chúng vận hành theo chiều hướng gây thiệt hại tới khách hàng.

– *Sự lạm dụng dịch vụ đám mây*. Một trong những lợi ích mà điện toán đám mây mang lại là nó cho phép một tổ chức nhỏ cũng có khả năng sử dụng một hạ tầng lớn. Một tổ chức nhỏ có thể gặp khó khăn khi xây dựng và duy trì hàng chục ngàn máy chủ. Tuy nhiên, với mô hình điện toán đám mây, họ lại có thể thuê được chúng trong một khoảng thời gian nhất định. Với nguyên tắc như vậy, một tin tặc có thể thuê một hệ thống hàng chục ngàn máy chủ của dịch vụ hạ tầng đám mây để nhằm mục đích xấu như giải mã dữ liệu, tấn công DDoS, hay phát tán các thông tin hoặc phần mềm độc hại.

– *Khảo sát không đầy đủ*. Nhiều doanh nghiệp lựa chọn việc chuyển đổi sang sử dụng các dịch vụ đám mây do những hứa hẹn về việc giảm chi phí đầu tư, tăng hiệu quả vận hành,… Tuy nhiên, doanh nghiệp sẽ chịu nhiều rủi ro tiềm tàng nếu họ thiếu sự hiểu biết về môi trường công nghệ mới này. Ví dụ khi một doanh nghiệp chuyển đổi một hệ thống ứng dụng đang vận hành trên một mạng cục bộ lên đám mây. Nếu hệ thống ứng dụng này áp dụng một số giả định về chính sách ATBM cho mạng cục bộ thì khi chuyển đổi ứng dụng lên môi trường điện toán đám mây, các chính sách này sẽ không còn hiệu lực nữa và khi đó ứng dụng sẽ nằm dưới nguy cơ mất an toàn.

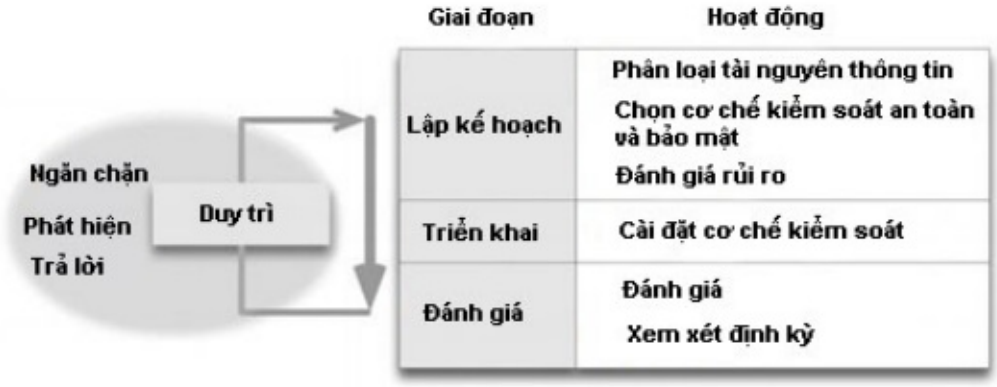
– *Lỗ hổng trong các công nghệ sử dụng chung*. Hệ thống đám mây thường cung cấp các dịch

vụ một cách linh hoạt thông qua việc chia sẻ hạ tầng, nền tảng và ứng dụng. Tuy nhiên, trong hệ thống thường có một số thành phần (chủ yếu từ hạ tầng như bộ đệm cache của CPU, bộ xử lý đồ họa GPU,…) không được thiết kế cho việc chia sẻ này. Các đặc điểm này khi bị tin tặc khai thác có thể tạo nên những lỗ hổng bảo mật mới.

Có thể nói, đa phần những nguy cơ kể trên đến từ những công nghệ cấu thành nên hệ thống đám mây như dịch vụ web, trình duyệt web, ảo hóa,…

3.2. MỘT SỐ PHƯƠNG PHÁP ĐẢM BẢO AN TOÀN CHO DỊCH VỤ ĐÁM MÂY

Để đảm bảo an toàn và bảo mật cho hệ thống đám mây, các nhà quản lý dịch vụ đám mây cần những chiến lược và quy trình hoàn chỉnh thay vì áp dụng những kỹ thuật ứng phó đơn lẻ, rời rạc. Nếu chúng ta xem xét các sự cố an toàn và bảo mật là một dạng rủi ro với hệ thống thì việc đảm bảo an toàn và bảo mật cho hệ thống có thể được thực hiện theo một quy trình quản lý rủi ro như trong hình 3.2.



Hình 3.2. Quy trình quản lý rủi ro về an toàn và bảo mật

Các bước thực hiện chính trong quy trình bao gồm:

Bước 1. Lập kế hoạch: Mục tiêu của bước này là nhận định những nguy cơ về an toàn và bảo mật; xác định các *cơ chế kiểm soát an toàn và bảo mật* (security controls) hiệu quả nhằm giải quyết các nguy cơ; lên kế hoạch cho việc thực hiện các cơ chế kiểm soát an toàn và bảo mật này.

Bước 2. Triển khai: Bao gồm việc cài đặt và cấu hình cho các cơ chế kiểm soát an toàn và bảo mật.

Bước 3. Đánh giá: Đánh giá tính hiệu quả của của các cơ chế kiểm soát và định kỳ xem xét tính đầy đủ của cơ chế kiểm soát.

Bước 4. Duy trì: Khi hệ thống và các cơ chế kiểm soát đã vận hành, cần thường xuyên cập nhật những thông tin mới về các nguy cơ ATBM.

Cơ chế kiểm soát an toàn và bảo mật (security controls) được hiểu như một kỹ thuật, một hướng dẫn hay một trình tự được định nghĩa tường minh giúp ích cho việc phát hiện, ngăn chặn, hoặc giải quyết các sự cố về an toàn bảo mật.

Năm 2013, liên minh an toàn và bảo mật trong điện toán đám mây (CSA) xuất bản tài liệu CSA Cloud Control Matrix phiên bản 3.0 (viết tắt là CSA CCM v3.0). Tài liệu này đề xuất một tập hợp bao gồm hơn một trăm hai mươi cơ chế kiểm soát an toàn và bảo mật nhằm trợ giúp các

nhà cung cấp dịch vụ đám mây dễ dàng ứng phó với các nguy cơ về ATBM.

Trong khuôn khổ cuốn sách này, chúng tôi không có ý định giới thiệu lại tất cả các cơ chế kiểm soát đó. Thay vì vậy, cuốn sách giới thiệu một số biện pháp đảm bảo ATBM được áp dụng phổ biến trong các hệ thống đám mây.

Bảo mật trung tâm dữ liệu

Bảo mật mức vật lý: Các công ty như Google, Microsoft, Yahoo, Amazon và một số nhà khai thác trung tâm dữ liệu đã có nhiều năm kinh nghiệm trong việc thiết kế, xây dựng và vận hành các trung tâm dữ liệu quy mô lớn. Những kinh nghiệm này đã được áp dụng cho chính nền tảng cơ sở hạ tầng điện toán đám mây của họ. Kỹ thuật tiên tiến trong việc bảo mật mức vật lý là đặt các trung tâm dữ liệu tại các cơ sở khó nhận biết với những khoảng sân rộng và vành đai kiểm soát được đặt theo tiêu chuẩn quân sự cùng với các biên giới tự nhiên khác. Các tòa nhà này nằm trong khu dân cư không đặt biển báo hoặc đánh dấu, giúp cho chúng càng trở nên khó nhận biết. Truy cập vật lý được các nhân viên bảo vệ chuyên nghiệp kiểm soát chặt chẽ ở cả vành đai kiểm soát và tại các lối vào với các phương tiên giám sát như camera, các hệ thống phát hiện xâm nhập và các thiết bị điện tử khác.

Những nhân viên được cấp phép phải sử dụng phương pháp xác thực hai bước không ít hơn ba lần mới có thể truy cập vào tầng trung tâm dữ liệu. Thông thường, tất cả khách tham quan hay các nhà thầu phải xuất trình căn cước và phải đăng ký. Sau đó họ tiếp tục được hộ tống bởi đội ngũ nhân viên được cấp phép.

Các công ty cung cấp dịch vụ đám mây đôi khi thiết lập trung tâm dữ liệu với mức độ tiên tiến vượt xa so với các trung tâm dữ liệu của các công ty dịch vụ tài chính. Máy chủ của các trung tâm dữ liệu này được đặt vào hầm trú ẩn kiên cố không dễ dàng vượt qua như chúng ta vẫn thấy trong các bộ phim gián điệp. Trong trung tâm dữ liệu Fort Knox của Salesforce.com, các nhân viên an ninh áp dụng phương pháp tuần tra vòng tròn, sử dụng máy quét sinh trắc học năm cấp độ, hay thiết kế lồng bẫy có thể rơi xuống khi chứng thực không thành công. Hình 3.3 minh họa một số biện pháp bảo mật vật lý.

Hình 3.3. Bảo mật mức vật lý

Để tránh các cuộc tấn công nội bộ, hệ thống ghi nhật ký và kiểm tra phân tích cho các kết nối cục bộ được kích hoạt thường xuyên. AICPA (American Institute of Certified Public Accountants) cung cấp những tiêu chuẩn kỹ thuật liên quan tới bảo mật kể trên trong chứng nhận SAS 70.

Chứng nhận SAS 70: Phần lớn các đám mây công cộng đều cần chứng nhận này. Chứng nhận này không phải là một danh mục để kiểm tra tại một thời điểm. Nó yêu cầu các tiêu chuẩn phải được duy trì trong ít nhất 6 tháng kể từ khi bắt đầu đăng ký. Thông thường chi phí để đạt được chứng nhận này rất lớn mà chỉ các nhà cung cấp hàng đầu mới đạt được.

Các biện pháp kiểm soát truy nhập

Tiếp theo vấn đề bảo mật mức vật lý là các kỹ thuật kiểm soát những đối tượng có thể truy

nhập vào đám mây. Dĩ nhiên điều này là cực kỳ cần thiết, bởi vì thiếu nó, tin tặc có thể truy nhập vào các máy chủ của người sử dụng, đánh cắp thông tin hoặc sử dụng chúng cho các mục đích xấu. Chúng ta hãy lấy ví dụ về cách thức kiểm soát truy nhập của Amazon Web Services (cũng tương tự như với một số đám mây khác). Cách thức kiểm soát này được thức hiện qua nhiều bước, thường bắt đầu với thông tin thẻ tín dụng của khách hàng.

Xác nhận bằng hóa đơn thanh toán: Nhiều dịch vụ thương mại điện tử sử dụng hóa đơn thanh toán cho mục đích xác thực với người dùng. Ở môi trường trực tuyến, hóa đơn thanh toán thường gắn liền với thẻ tín dụng của khách hàng. Tuy nhiên thẻ tín dụng thì thường không có nhiều thông tin gắn với khách hàng nên một số biện pháp khác có thể được áp dụng.

Kiểm tra định danh qua điện thoại: Mức độ tiếp theo của kỹ thuật kiểm soát truy cập là phải xác định đúng đối tượng truy cập. Để tránh rủi ro trong việc xác nhận, một hình thức xác nhận qua các kênh liên lạc khác như điện thoại là cần thiết. Thông thường nhà cung cấp sẽ liên hệ với khách hàng và yêu cầu khách hàng trả lời số PIN được hiển thị trên trình duyệt.

Giấy phép truy nhập: Hình thức giấy phép truy nhập đơn giản nhất chính là mật khẩu. Khách hàng có thể lựa chọn cho mình một mật khẩu mạnh hoặc có thể lựa chọn những giấy phép truy nhập nhiều bước như RSA SecurID. Người sử dụng cần dùng giấy phép truy nhập khi họ muốn sử dụng dịch vụ trực tiếp. Trong trường hợp người sử dụng dịch vụ qua API, họ cần phải có khóa truy nhập.

Khóa truy nhập: Để gọi bất kỳ API nào của hệ thống đám mây, người sử dụng phải có một khóa truy nhập. Khóa này được cung cấp cho người sử dụng trong quá trình thiết lập tài khoản. Người sử dụng cần bảo vệ khóa truy nhập này để tránh sự rò rỉ dịch vụ.

Giấy phép X.509: Giấy phép X.509 dựa trên ý tưởng về hạ tầng khóa công khai (PKI). Một giấy phép X.509 bao gồm một giấy phép (chứa khóa công khai và nội dung cấp phép) và một khóa bí mật. Giấy phép được sử dụng mỗi khi tiêu thụ dịch vụ, trong đó khóa bí mật được sử dụng để sinh ra chữ ký số cho mỗi yêu cầu dịch vụ. Dĩ nhiên, khóa bí mật cần phải được giữ kín và không được phép chia sẻ. Tuy nhiên, do giấy phép X.509 thường được các nhà cung cấp sinh ra và chuyển cho người sử dụng nên không thể đảm bảo 100% rằng khóa bí mật không bị rò rỉ.

Để sử dụng giấy phép X.509, khi yêu cầu dịch vụ, người sử dụng sinh chữ ký số bằng khóa bí mật của mình, sau đó gắn chữ ký số, giấy phép với yêu cầu dịch vụ. Khi hệ thống nhận được yêu cầu, nó sẽ sử dụng khóa công khai trong giấy phép để giải mã chữ ký số và chứng thực người dùng. Hệ thống cũng sử dụng giấy phép để khẳng định các yêu cầu đặt ra là hợp lệ.

Cặp khóa: Cặp khóa là yếu tố quan trọng nhất trong việc truy nhập vào các thể hiện của AWS. Mỗi dịch vụ cần một cặp khóa riêng biệt. Cặp khóa cho phép hệ thống đảm bảo người dùng hợp lệ. Mặc dù không thể thay thế cặp khóa, tuy nhiên người sử dụng có thể đăng ký nhiều cặp khóa.

AWS tạo cặp khóa bằng AWS Management Console nếu người sử dụng không tự tạo ra cho mình. Khóa bí mật sẽ được gửi đến người sử dụng và sau đó hệ thống sẽ không lưu trữ lại chúng.

Bảo mật dữ liệu và mạng

Bảo mật hệ điều hành: Bảo mật mức hệ thống có nhiều cấp độ: bảo mật cho hệ điều hành của máy chủ vật lý; bảo mật cho hệ điều hành của các máy ảo chạy trên nó; tường lửa và bảo mật cho các API.

Để bảo mật cho máy chủ vật lý, Amazon yêu cầu người quản trị sử dụng khóa SSH để truy nhập vào các máy *bastion*. Bastion là các máy được thiết kế đặc biệt và không cho phép người

sử dụng truy nhập tới chúng. Sau khi đã truy nhập được vào bastion, người quản trị có thể thực hiện một số lệnh với mức ưu tiên cao lên các máy chủ vật lý. Khi người quản trị đã hoàn tất công việc, quyền truy nhập của họ vào các máy bastion sẽ bị rút.

Bảo mật mạng: Các đám mây công cộng thường cung cấp một hệ thống tường lửa để ngăn chặn các truy nhập trái phép. Hệ thống tường lửa nội bộ được sử dụng để kiểm soát sự trao đổi nội tại bên trong đám mây. Thông thường người sử dụng cần định nghĩa tường minh các cổng cần mở cho các giao dịch nội bộ này. Việc kiểm soát và thay đổi các luật tường lửa do mỗi máy ảo tự đảm nhận, tuy nhiên hệ thống đám mây sẽ yêu cầu giấy phép X.509 khi người sử dụng thực hiện các thay đổi này trên máy ảo. Trong mô hình cung cấp dịch vụ của Amazon EC2, quản trị viên của hệ thống đám mây và quản trị viên của các máy ảo là hai đối tượng khác nhau. AWS khuyến khích người sử dụng tự định nghĩa thêm các luật tường lửa cho các máy ảo của mình.

Thông thường, tường lửa cho mỗi máy ảo mặc định sẽ từ chối mọi kết nối tới các cổng, người sử dụng sẽ phải cân nhắc cẩn thận cho việc mở cổng nào phù hợp với ứng dụng của mình. Các đám mây công cộng thường là đích ngắm của các tấn công trên mạng internet như DDoS.

Bảo mật cho môi trường cộng sinh: Trong hệ thống đám mây Amazon EC2, một máy ảo không thể chạy dưới chế độ hỗn tạp (promiscuous mode) để có thể "ngửi" gói tin từ các máy ảo khác. Kể cả khi người sử dụng có ý thiết lập chế độ hỗn tạp này cho máy ảo thì các gói tin tới các máy ảo khác cũng không thể gửi đến máy ảo đó được. Chính vì vậy, các phương pháp tấn công theo kiểu ARP cache poisoning không có hiệu lực trong Amazon EC2. Tuy nhiên, một khuyến cáo chung cho khách hàng là họ nên mã hóa những giao dịch qua mạng quan trọng cho dù chúng đã được bảo vệ cẩn thận bởi EC2.

Các nhà cung cấp dịch vụ đám mây cũng thường cung cấp không gian lưu trữ trên một kho dữ liệu dùng chung. Các đối tượng được lưu trữ thường được kèm theo mã băm MD5 để xác nhận tính toàn vẹn. Không gian lưu trữ cho từng người sử dụng cũng được ảo hóa thành các đĩa ảo và chúng thường được xóa mỗi khi khởi tạo. Chính vì vậy, vấn đề rò rỉ dữ liệu do sử dụng chung không gian lưu trữ vật lý không quá lo ngại. Tuy vậy, các nhà cung cấp dịch vụ đám mây thường khuyến cáo người sử dụng hệ thống tệp được mã hóa trên các thiết bị lưu trữ ảo này.

Bảo mật cho các hệ thống giám sát: Các hệ thống giám sát thường được sử dụng để bật/tắt các máy ảo, thay đổi tham số về tường lửa,... Mọi hành động này đều yêu cầu giấy phép X.509. Hơn nữa, khi các hành vi này được thực hiện qua các API, có thể bổ sung thêm một tầng bảo mật nữa bằng cách mã hóa các gói tin, ví dụ sử dụng SSL. Khuyến cáo của các nhà cung cấp dịch vụ là nên luôn luôn sử dụng kênh SSL cho việc thực thi các API của nhà cung cấp dịch vụ.

Bảo mật lưu trữ dữ liệu: Các dịch vụ lưu trữ đám mây thường kiểm soát quyền truy nhập dữ liệu thông qua một danh sách kiểm soát truy nhập (ACL – access control list). Với ACL, người sử dụng có toàn quyền kiểm soát tới những đối tượng được phép sử dụng dịch vụ của họ.

Một lo lắng khác cho vấn đề bảo mật dữ liệu là chúng có thể bị đánh cắp trong quá trình truyền thông giữa máy của người sử dụng dịch vụ và đám mây. Khi đó các API được bảo vệ bởi SSL sẽ là giải pháp cần thiết. Khuyến cáo chung với người dùng là trong mọi trường hợp, nên mã hóa dữ liệu trước khi gửi đến lưu trữ trên đám mây.

3.3. THẾT KẾ KIẾN TRÚC HỆ THỐNG ĐÁM MÂY NHẰM ĐẢM BẢO AN TOÀN BẢO MẬT

Thiết kế kiến trúc là bước quan trọng trong quy trình xây dựng một hệ thống phức tạp. Mục

tiêu chính của bước này là xác định được một (hoặc nhiều) cấu trúc tổng thể của hệ thống với những thành phần và mối quan hệ giữa chúng.

Trong phần này, chúng ta tập trung quan tâm tới cấu trúc vật lý (trung tâm dữ liệu, mạng kết nối,...) và cấu trúc thành phần phần mềm (các phân hệ) của hệ thống đám mây trong mục tiêu đáp ứng tính an toàn và bảo mật. Đầu tiên, chúng ta sẽ nhận định một số yêu cầu kiến trúc liên quan tới an toàn và bảo mật. Sau đó giới thiệu một số mẫu kiến trúc điển hình cho an toàn và bảo mật đám mây. Phần cuối cùng giới thiệu một số ví dụ về kiến trúc các hệ thống đám mây.

Những yêu cầu an toàn và bảo mật cho kiến trúc đám mây

Một trong những mục tiêu cho việc thiết kế kiến trúc là việc đảm bảo sự đáp ứng của hệ thống với những yêu cầu đặt ra, trong đó bao hàm cả các yêu cầu về an toàn và bảo mật. Các yêu cầu này thường xuất phát từ một số yếu tố cần cân nhắc như chi phí, độ tin cậy, hiệu năng, các ràng buộc pháp lý,... Nội dung tiếp sau đây tóm tắt một số yêu cầu bảo mật cho các hệ thống đám mây.

Yêu cầu bảo mật mức vật lý

Hệ thống đám mây được xây dựng từ một hoặc một vài trung tâm dữ liệu. Việc đảm bảo ATBM cho các trung tâm dữ liệu này cũng chính là một yêu cầu quan trọng cho hệ thống đám mây. Công việc này chủ yếu liên quan tới hai nhóm yêu cầu:

– Phát hiện và phòng chống sự thâm nhập trái phép vào trung tâm dữ liệu, các thiết bị phần cứng.

– Bảo vệ hệ thống khỏi các thảm họa tự nhiên.

Yêu cầu bảo mật với các thành phần hệ thống

Quản lý định danh: Quản lý định danh là chìa khóa của việc đảm bảo ATBM của hệ thống. Thông tin về định danh phải chính xác và sẵn sàng cho các thành phần khác của hệ thống. Những yêu cầu cho thành phần này bao gồm:

– Phải có cơ chế kiểm soát để đảm bảo tính bí mật, tính toàn vẹn và tính sẵn dùng của thông tin định danh.

– Phân hệ quản lý định danh cũng phải được sử dụng cho mục đính chứng thực người dùng của hệ thống đám mây (thường với tải yêu cầu cao).

– Cân nhắc cơ chế sử dụng hoặc tương tác với các hệ thống quản lý định danh của bên thứ ba.

– Kiểm tra định danh của người sử dụng khi đăng ký khớp các yêu cầu của pháp luật.

– Lưu thông tin định danh của người sử dụng, kể cả khi họ rút khỏi hệ thống, phục vụ cho công tác kiểm tra, báo cáo (với các cơ quan pháp luật).

– Khi một định danh được xóa bỏ, sau đó tái sử dụng, cần đảm bảo người sử dụng mới không thể truy cập vào các thông tin của định danh trước đó.

Quản lý truy cập. Quản lý truy cập là thành phần sử dụng thông tin định danh để cho phép và đặt ràng buộc với các truy cập dịch vụ đám mây. Các yêu cầu liên quan tới quản lý truy cập bao gồm:

– Quản trị viên của đám mây chỉ có quyền truy cập hạn chế tới dữ liệu của khách hàng. Quyền truy cập này phải được ràng buộc chặt chẽ và được công bố rõ ràng trong hợp đồng sử dụng dịch vụ (SLA).

– Cần có cơ chế chứng thực nhiều bước cho những thao tác yêu cầu mức ưu tiên cao. Cần có cơ chế xác quyền đủ mạnh để đảm bảo các thao tác này không ảnh hưởng trên toàn đám mây.

– Không cho phép chia sẻ một số tài khoản đặc biệt (ví dụ tài khoản root), thay vì vậy, sử dụng các cơ chế khác như sudo.

– Cài đặt các cơ chế như LPP (least privilege principal) khi gán quyền truy nhập hay RBAC (role-based access control) để thiết lập các ràng buộc truy nhập.

– Thiết lập danh sách trắng (white list) về IP cho các quản trị viên.

Quản lý khóa. Trong đám mây, việc mã hóa dữ liệu là phương tiện chính để đảm bảo an toàn thông tin. Quản lý khóa là phân hệ phục vụ công tác lưu trữ và quản lý khóa cho việc mã hóa và giải mã dữ liệu của người sử dụng. Yêu cầu chính cho phân hệ này là:

– Có cơ chế kiểm soát và giới hạn các truy cập vào khóa.

– Với mô hình đám mây có cơ sở hạ tầng trên nhiều trung tâm dữ liệu, cần đảm bảo việc hủy bỏ khóa phải có hiệu lực tức thì trên các trạm.

– Đảm bảo việc khôi phục cho các khóa khi có lỗi.

– Mã hóa dữ liệu và máy ảo khi cần thiết.

Ghi nhận sự kiện và thống kê. Các sự kiện liên quan tới an toàn bảo mật của mạng và hệ thống cần được ghi nhận (logs) và thống kê cho nhu cầu kiểm tra, đánh giá. Những yêu cầu chính cho phân hệ này là:

– Ghi nhận sự kiện ở nhiều mức: từ các thành phần hạ tầng vật lý như máy chủ vật lý, mạng vật lý, tới những thành phần ảo hóa như máy ảo, mạng ảo.

– Các sự kiện được ghi nhận với đầy đủ thông tin để phân tích: thời gian, hệ thống, người dùng truy cập,...

– Các sự kiện cần được ghi nhận gần tức thời.

– Thông tin ghi nhận cần liên tục và tập trung.

– Thông tin ghi nhận cần được duy trì cho tới khi chúng không còn cần thiết.

– Thông tin ghi nhận được có thể được cung cấp tới khách hàng như một dạng dịch vụ.

– Tất cả các thao tác ghi nhận đều phải đảm bảo tính bí mật, nhất quán và sẵn sàng của thông tin ghi nhận được.

Giám sát bảo mật. Giám sát bảo mật là phân hệ liên quan tới việc khai thác các thông tin ghi nhận (logs), thông tin giám sát mạng hay thông tin bảo mật từ hệ thống giám sát vật lý. Yêu cầu cho phân hệ này bao gồm:

– Là dạng dịch vụ có tính sẵn sàng cao, có thể truy cập cục bộ hoặc từ xa trên một kênh bảo mật.

– Bao gồm một số chức năng chính:

+ Cảnh báo sự cố bảo mật dựa trên phân tích tự động các thông tin thu thập được.

+ Gửi cảnh báo bằng nhiều phương tiện như email, sms.

+ Cho phép người quản trị khai thác và phát hiện nguyên nhân của các sự cố.

+ Có cơ chế phát hiện xâm nhập hoặc bất thường của hệ thống.

+ Cho phép khách hàng có thể tự xây dựng cơ chế cảnh báo khi sử dụng PaaS hoặc IaaS.

Quản lý sự cố. Quản lý sự cố và phản ứng khi có sự cố là công tác quan trọng với bảo mật hệ thống. Các yêu cầu cho công tác này bao gồm:

– Có quy trình đầy đủ cho việc phát hiện, ghi nhận và xử lý sự cố.

– Có các cơ chế hỗ trợ người sử dụng thông báo về sự cố.

– Việc kiểm tra sự cố cần được thực hiện thường xuyên.

Kiểm tra an toàn và vá lỗi. Đây là công tác được thực hiện mỗi khi triển khai hoặc nâng cấp một dịch vụ mới. Các yêu cầu cho công việc này bao gồm:

– Có môi trường cô lập để phát triển, kiểm tra và điều chỉnh trước khi đưa dịch vụ vào sử dụng.

– Có quy trình vá lỗi cho các thành phần của hệ thống.

– Theo dõi thường xuyên các lỗ hổng bảo mật.

Kiểm soát mạng và hệ thống. Hệ thống kiểm soát mạng và các máy chủ được áp dụng cho cả các hạ tầng vật lý và hạ tầng ảo. Các yêu cầu cho hệ thống này bao gồm:

– Đảm bảo khả năng cô lập, khả năng cấu hình và tính bảo mật cho các thành phần bảo mật.

– Đảm bảo khả năng cô lập về mạng cho các vùng chức năng của hệ thống đám mây.

– Phân tách truy nhập thiết bị vật lý với thiết bị ảo.

– Phân tách vùng thiết bị ảo của các khách hàng khác nhau.

– Đảm bảo tính nhất quán của máy ảo, hệ điều hành,... cho ứng dụng của khách hàng.

Quản lý cấu hình. Trong một hệ thống đám mây với hạ tầng linh động, việc duy trì một danh sách thông tin về các tài nguyên của hệ thống và cấu hình của chúng là cần thiết. Các yêu cầu cho công tác này bao gồm:

– Sử dụng một hệ thống cơ sở dữ liệu cấu hình CMDB.

– Phân loại các tài nguyên theo chức năng, tính nhạy cảm, độ quan trọng,…

Các yêu tố kiến trúc và mẫu bảo mật

Phòng ngự chiều sâu (defence in-depth). Thuật ngữ phòng ngự chiều sâu lần đầu tiên được đề cập đến trong lĩnh vực an ninh mạng và máy tính là trong bài báo “Information warfare and

dynamic information defence" vào năm 1996. Tiếp cận này trước đó được gọi bằng nhiều tên trong đó có "phòng ngự theo lớp" (layered defence). Tư tưởng chung của tiếp cận này là sử dụng nhiều tầng kiểm soát bảo mật để tạo nên một giải pháp đầy đủ, hoàn chỉnh hơn.

Trên quan điểm kiến trúc, kỹ thuật phòng ngự theo chiều sâu có thể được xem như một mẫu thiết kế hiệu quả cho vấn đề bảo mật. Ứng dụng của mẫu này có thể thấy ở nhiều hệ thống thực tiễn. Ví dụ như phòng ngự chiều sâu cho phân hệ kiểm soát truy nhập bao gồm nhiều lớp: lớp 1 – mạng riêng ảo (VPN); lớp 2 – bộ định tuyến cổng vào với cơ chế lọc IP; lớp 3 – token bảo mật.

Hũ mật ong (honeypots). "Hũ mật ong" là một kỹ thuật bẫy nổi tiếng. Trong một hệ thống mạng doanh nghiệp, "hũ mật ong" tạo nên một hệ thống không tồn tại hoặc không có giá trị nhằm thu hút sự tấn công. Khi đã thu hút thành công, "hũ mật ong" lại được sử dụng để quan sát, phân tích và cảnh báo. Dù thế nào thì kỹ thuật này cũng khiến cho bên tấn công tiêu phí thời gian và sức lực.

Hộp cát (sandbox). Hộp cát (sandbox) là một lớp trừu tượng nằm giữa phần mềm với hệ điều hành nhằm tạo môi trường độc lập cho việc thực thi ứng dụng. Tác dụng của hộp cát cũng giống như hypervisor trong việc cung cấp các máy ảo. Với hộp cát, hệ thống có thêm một tầng bảo vệ theo mô hình phòng ngự chiều sâu.

Cô lập máy ảo. Hạ tầng chuyển mạch trong một hệ thống đám mây không thể cô lập được các gói tin truyền thông giữa các máy ảo nằm trên cùng một môi trường phần cứng. Do vậy, nếu các gói tin không được mã hóa, máy ảo có thể theo dõi, quan sát các gói tin gửi đến/gửi đi từ máy ảo khác trong cùng một mạng. Cô lập máy ảo là kỹ thuật:

– Ứng dụng công nghệ ảo hóa để cô lập các máy ảo trong cùng một mạng vật lý;

– Mã hóa các gói tin gửi đến/gửi đi từ máy ảo;

– Kiểm soát truy cập đến máy ảo, đặc biệt là các cổng dịch vụ;

– Lọc gói tin đến máy ảo qua các cơ chế tường lửa.

Tạo dư thừa và đảm bảo tính sẵn sàng. Một trong những mẫu thiết kế thường được ứng dụng rộng rãi trong việc đảm bảo tính sẵn sàng cao của dịch vụ, trong đó bao gồm cả các dịch vụ bảo mật như quản lý định danh, quản lý truy cập,… là tạo dư thừa cho những thành phần hệ thống, bao gồm máy chủ, thiết bị mạng,… Tùy thuộc vào mức độ đảm bảo tính sẵn sàng mà kiến trúc có thể thiết lập dư thừa tương ứng. Tuy nhiên, cần lưu ý rằng sự dư thừa bao giờ cũng kéo theo những hệ quả như: tăng chi phí, tăng độ phức tạp của hệ thống.

Nội dung mục này giới thiệu một số kiến trúc đám mây điển hình, trong đó có bao hàm các thành phần đảm bảo tính an toàn và bảo mật.

Kiến trúc đám mây cung cấp dịch vụ PaaS (dịch vụ định danh, dịch vụ cơ sở dữ liệu)

Hình 3.4 giới thiệu kiến trúc một hệ thống đám mây cung cấp các dịch vụ PaaS (dịch vụ định danh, dịch vụ cơ sở dữ liệu).

Trong kiến trúc này, người sử dụng thông thường truy nhập vào dịch vụ của hệ thống thông qua mạng công cộng. Bên cạnh đó, hệ thống cũng cung cấp một mạng riêng biệt – mạng OOB (out-of-band) nhằm phục vụ công tác quản trị. Việc kiểm soát truy nhập vào mạng OOB này có thể được thực hiện thông qua một danh sách IP trắng – IP của các quản trị viên hệ thống. Thêm vào đó, quản trị viên cần thực hiện xác thực mỗi khi thao tác. Cơ chế xác thực hai bước (token và pin) có thể giúp hệ thống trở nên an toàn. Đây cũng là ví dụ về việc áp dụng cơ chế phòng

ngự chiều sâu.

Hệ thống mạng cục bộ chia làm ba mạng chính:

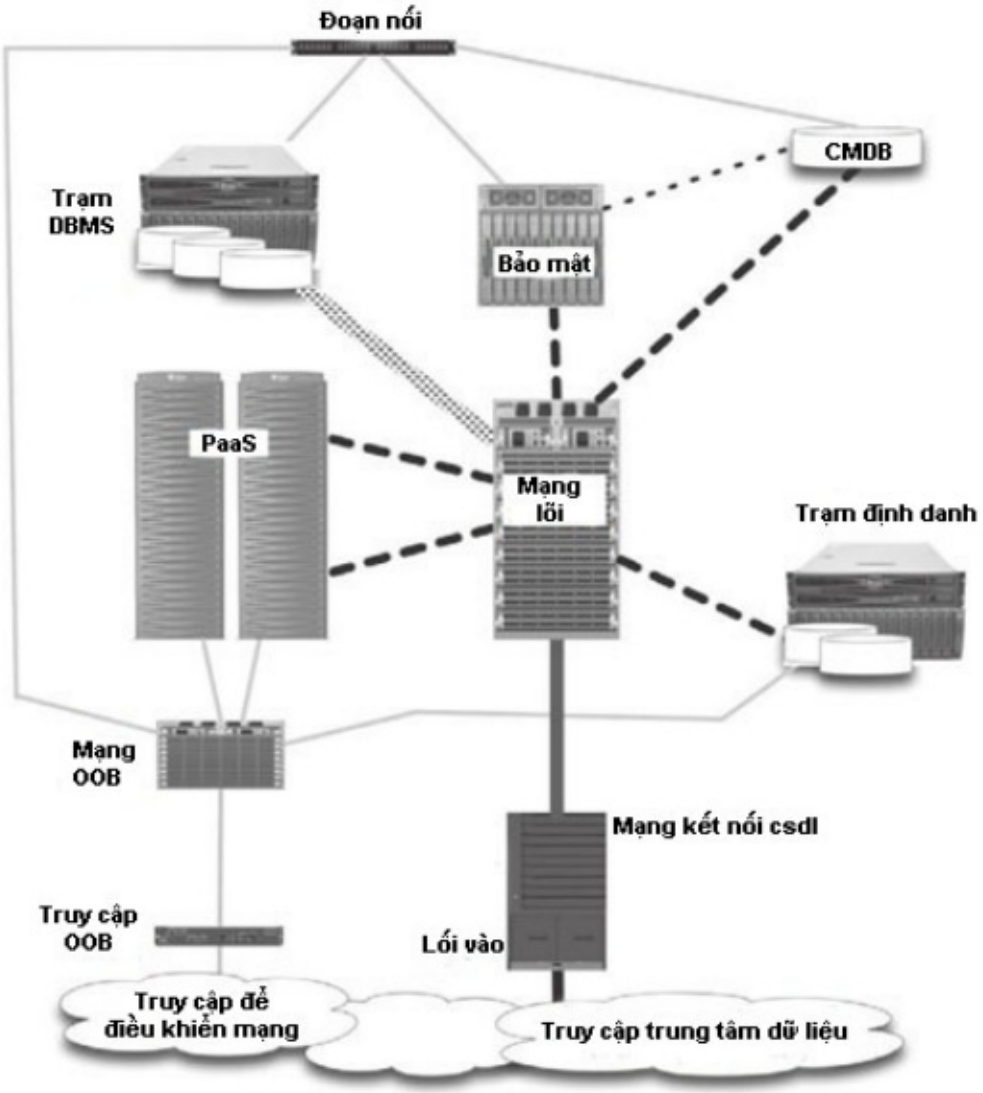
– Mạng OOB: sử dụng để quản trị các thành phần khác trong hệ thống.

– Mạng lõi: sử dụng để cung cấp dịch vụ.

– Mạng kết nối với cơ sở dữ liệu: bao gồm nhiều kết nối đảm bảo tính sẵn sàng.

Các thành phần của hệ thống cũng được thiết kế để cô lập các dịch vụ khác nhau của hệ thống như dịch vụ cơ sở dữ liệu và dịch vụ định danh. Một thành phần quản trị cấu hình (CMDB) cũng được đưa vào trong kiến trúc nhằm quản lý cấu hình các tài nguyên cung cấp bởi hệ thống.

Một số kiến trúc đám mây điển hình đáp ứng yêu cầu an toàn và bảo mật



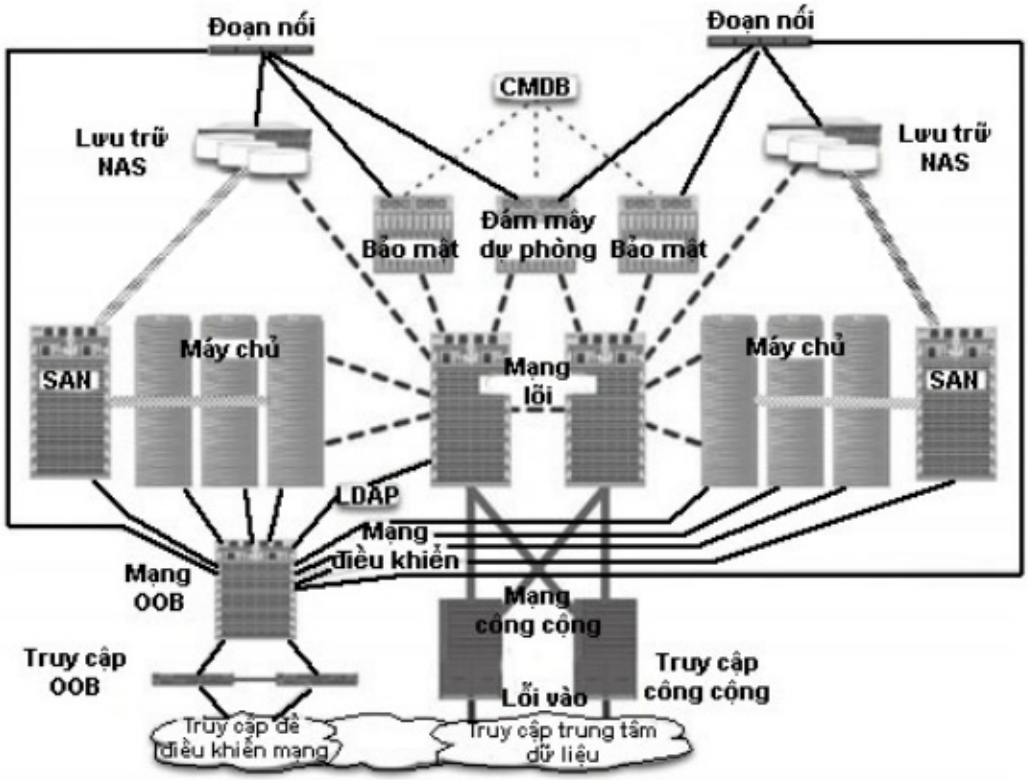
Hình 3.4. Kiến trúc đám mây cung cấp dịch vụ định danh và dịch vụ cơ sở dữ liệu

Kiến trúc đám mây cung cấp kho lưu trữ và dịch vụ tính toán

Hình 3.5 minh họa một ví dụ khác về kiến trúc đám mây với các loại hình dịch vụ cung cấp là dịch vụ tính toán và dịch vụ lưu trữ dữ liệu. Trong kiến trúc này, hệ thống đám mây cung cấp một số lượng lớn tài nguyên tính toán được ảo hóa trên các máy chủ, cũng như các kho lưu trữ trên các thiết bị SAN. Hơn nữa, kiến trúc này hỗ trợ tính sẵn sàng cao cho người sử dụng thông

qua việc tạo lập dư thừa cho mạng kết nối công cộng và mạng OOB. Để đảm bảo tính sẵn sàng cho việc truy nhập vào kho lưu trữ SAN, mạng SAN được thiết lập với các kết nối giữa kho lưu trữ SAN và các máy chủ tính toán.

Để đảm bảo tính sẵn sàng cao, bản thân các máy chủ và kho lưu trữ SAN cũng được thiết kế dư thừa. Ở đây chúng ta có thể áp dụng một số chiến lược khác nhau để cân đối giữa chi phí đầu tư và tính sẵn sàng của hệ thống.



Hình 3.5. Kiến trúc đám mây cung cấp dịch vụ tính toán và lưu trữ

Với các tài nguyên tính toán, để đảm bảo khả năng đáp ứng tốt cho dịch vụ, hệ thống cần có những thiết kế cho việc dành sẵn tài nguyên (provision). Các máy chủ phục vụ cho việc này được thiết kế độc lập. Việc dành sẵn tài nguyên này cũng đòi hỏi sự kiểm soát và quản lý của một phân hệ quản lý cấu hình tài nguyên CMDB. Kết nối trên sơ đồ đã thể hiện điều này.

Một điểm đặc biệt khác trong sơ đồ kiến trúc của hình là hai phân hệ bảo mật được thiết kế theo mẫu dư thừa. Các phân hệ này cung cấp những dịch vụ bảo mật như:

– Jumphost và VPN: cho phép tạo lập các mạng riêng ảo và cho phép các quản trị viên có thể truy nhập trực tiếp vào các máy chủ cần quản lý.

– Trung tâm điều hành bảo mật: cho phép giám sát các vấn đề liên quan tới an toàn bảo mật, quét các lỗi bảo mật của hệ thống, phân tích nguyên nhân và báo cáo.

– Ghi nhật ký về các sự kiện của hệ thống và cảnh báo nếu có.

– Giám sát thông tin mạng (quét lỗ hổng, giám sát băng thông mạng,...)

Có thể nói thành phần bảo mật này là công cụ giám sát bảo mật chính của các quản trị viên trong hệ thống đám mây.

3.4. CÂU HỎI VÀ BÀI TẬP

1. Tại sao người sử dụng dịch vụ đám mây thường lo ngại về vấn đề an toàn – bảo mật?
2. Hãy trình bày những nguy cơ về an toàn bảo mật trong mô hình đa người thuê (multi-tenancy)?
3. Công nghệ ảo hóa có tạo nên những nguy cơ mới về an toàn bảo mật không? Hãy trình bày những vấn đề an toàn bảo mật với công nghệ ảo hóa.
4. Nêu các đặc điểm chính của biện pháp bảo mật mức vật lý.
5. Nêu các đặc điểm chính của biện pháp bảo mật dữ liệu.
6. Chiến lược/mẫu phòng ngự chiều sâu là gì? Cho ví dụ minh họa.
7. Chiến lược/mẫu "hũ mật ong" là gì? Cho ví dụ minh họa.
8. Chiến lược/mẫu tạo dư thừa là gì? Cho ví dụ minh họa.