**BÀI TẬP TRÊN LỚP MÔN HỌC: HỆ PHÂN TÁN**

**CHƯƠNG 7: Sao lưu và thống nhất dữ liệu**

HỌ TÊN SV: Nguyễn Đức Thiên MÃ LỚP: 114175

MSSV: 20168806 MÃ HỌC PHẦN:IT4610Q

***Câu hỏi lý thuyết***

Câu hỏi 1: *Tại sao phải thực hiện sao lưu dữ liệu?*

Lý do để thực hiện sao lưu đó là đảm bảo độ tin cậy, hiệu năng và khả năng mở rộng của hệ phân tán.

- Độ tin cậy: có nhiều bản sao giúp tăng tính sẵn sàng của hệ thống từ đó tăng tính tin cậy của hệ thống, khi một bản dữ liệu bị hỏng thì vẫn còn các bản sao dữ liệu khác backup.

- Hiệu năng::

* + - * Vấn đề chính: Để giữ sự nhất quán giữa các bản sao, cần đảm bảo các xung đột ở các nơi xảy ra được thực hiện theo một thứ tự ở mọi nơi.
* Xung đột:
* Xung đột đọc-ghi: thao tác đọc ghi chạy đồng thời.
* Xung đột ghi-ghi: hai thao tác ghi chạy đồng thời.

- Tính mở rộng: Đảm bảo thứ tự xung đột là một việc tốn kém, làm giảm khả năng mở rộng của hệ phân tán.

Câu hỏi 2: Xét một kho dữ liệu phân tán với 5 tiến trình độc lập P1, P2, P3, P4, và P5. Mỗi tiến trình chỉ tác động lên được bản sao cục bộ riêng của mình. Các bản sao cục bộ kết nối thành kho dữ liệu phân tán. Xét các tiến trình chỉ tương tác (ghi, đọc) lên thành phần dữ liệu x ở bản sao cục bộ riêng của mình. Hoạt động của mô hình ở các thời điểm t tương ứng, các thao tác được thực hiện như sau:

t1: P1 ghi giá trị a

t2: P3 đọc giá trị a

t3: P2 ghi giá trị b và P3 ghi giá trị c

t4: P5 đọc được giá trị b

t5: P4 và P5 đều đọc được giá trị a

t6: P4 đọc được giá trị b

t7: P4 và P5 đọc được giá trị c

(biết rằng ti < ti+1 với i=(0..6)

Câu hỏi:

*a) Mô hình trên có thoả mãn thống nhất nhân quả không? Giải thích.*

*b) Mô hình trên có thoả mãn thống nhất tuần tự không? Giải thích.*

A screenshot of a cell phone

Description automatically generated

Hình : Mô tả hoạt động

a. Mô hình trên có thoả mãn tính chất nhân quả. Vì:

Các thao tác W(x)a, W(x)b, W(x)c xảy ra tại P1, P2, P3 là tương tranh. Tại P3 ta có R(x)a xong rồi mới W(x)c nên giá trị c phụ thuộc nhân quả vào a. Các tiến trình khác đều đọc được thứ tự nhân quả a -> c. Còn giá trị b có thể xảy ra tương tranh khiến cho P4 đọc được b trước rồi mới đọc được a, c.

b. Mô hình trên không thoả mãn tính chất tuần tự. Vì:

Ta thấy P4 và P5 có hai thứ tự khác nhau là a,b,c và b,c,a. Điều này vi phạm tính chất tuần tự.

Câu hỏi 3: *Conit là gì? Nếu đặt kích thước Conit quá lớn thì sẽ gây ra vấn đề gì? Tương tự với kích thước Conit quá nhỏ?*

Conit (consistency unit) là đơn vị dữ liệu dùng để đánh giá sự nhất quán của các bản sao dữ liệu.

Nếu conit quá lớn thì các bản sao dữ liệu sẽ mất nhiều thời gian để đồng bộ giữa các bản sao dữ liệu, từ đó sớm gây ra hiện tượng bất đồng bộ về dữ liệu.

Nếu các conit nhỏ quá thì hệ thống sẽ phải tốn tài nguyên cho việc quản lý các conit.

Câu hỏi 4: *Tại sao thống nhất nhân quả có tính thống nhất yếu hơn thống nhất tuần tự? Cho ví dụ để làm rõ điều này.*

Thống nhất nhân quả yếu hơn thống nhất tuần tự bởi vì:

- Thống nhất tuần tự yêu cầu cầu đúng thứ tự đọc và ghi.

- Thống nhất nhân quả chỉ yêu cầu cầu đúng thứ tự đọc đối với các thao tác ghi có phụ thuộc nhân quả.

-Thống nhất tuần từ đảm bảo yêu cầu về thứ tự cho tất cả các tiến trình của hệ thống, còn thống nhất nhân quả chỉ đảm bảo cho các tiến trình có tính nhân quả.

Ta có thể lấy luôn *Hình 1* của câu 2 làm ví dụ. Hệ thống đảm bảo tính thống nhất nhân quả nhưng lại không đảm bảo tính thống nhất tuần tự.

Câu hỏi 5: *Vấn đề của mô hình Eventual Consistency là gì? Từ đó đưa ra định nghĩa mô hình thống nhất hướng client.*

Vấn đề của mô hình Eventual Consistency là khi thay đổi các máy chủ bản sao thì dữ liệu chưa kịp cập nhập tại máy chủ bản sao mới.

Một người dùng khi truy cập vào bản sao A và thực hiện thay đổi dữ liệu, sau đó chuyển sang vị trí khác và truy cập vào bản sao B thì dữ liệu vừa thay đổi ở bản sao A có thể chưa được cập nhập đến bản sao B.

Từ đó, mô hình hình thống nhất hướng client được định nghĩa để đảm bảo tính thống nhất cho các truy cập của từng client vào kho dữ liệu của hệ thống.

Câu hỏi 6: *Một ngân hàng quyết định sử dụng dịch vụ CDN (Content Delivery Network) của một công ty mới khởi nghiệp cung cấp.*

*a) Với bước đặt máy chủ, công ty chọn thuật toán chọn đặt các máy chủ bản sao (replica) dựa trên khoảng cách với các chi nhánh ngân hàng. Hãy đề xuất thuật toán chọn đặt k replica với N vị trí có thể đặt máy chủ. Biết rằng đây là thuật toán dựa trên khoảng cách và công ty biết trước các vị trí các chi nhánh ngân hàng.*

*b) Với thuật toán để quản lý nội dung dữ liệu ở các replica, công ty quyết định chọn thuật toán dựa trên bản sao kích hoạt bởi server (server-initiated replicas). Hãy mô tả cơ chế đó với việc xem xét một đơn vị dữ liệu X là thông tin tài khoản một người dùng cùng với 2 ngưỡng là del(X) và rep(X).*

*c) Liên quan đến giao thức đảm bảo thống nhất, công ty quyết định chọn giao thức ghi trên các bản sao (replicated write), tuy nhiên công ty băn khoăn giữa giao thức sao lưu tích cực và giao thức sao lưu dựa trên túc số. Bạn hãy giúp công ty lựa chọn giao thức phù hợp bằng việc so sánh 2 giao thức trên với việc chỉ ra ưu nhược điểm của chúng.*

a. Thuật toán dựa trên khoảng cách giữa các ngân hàng. Chọn ra k địa điểm đặt các máy chủ bản sao, sao cho khoảng cách trung bình từ k máy chủ này đến N chi nhánh là thấp nhất. Có thể dùng các thuật toán quy hoạch động, tham lam, đồ thị hoặc cây.

b. Khi truy cập với đơn vị dữ liệu X tại máy chủ A. Giả sử tồn tại X tại các máy chủ A, B.

- Số lượt truy cập tới máy chủ A đạt dưới ngưỡng del(X): thực hiện xoá bản ghi X tại máy chủ A, điều hướng các truy cập tới máy chủ B.

- Số lượt truy cập tới máy chủ A vượt trên ngưỡng rep(X): thực hiện copy bản ghi X từ máy chủ B đến máy chủ A.

- Số lượt truy cập nằm giữa khoảng del(X) và rep(X): thực hiện đếm số lượt truy cập.

c.

Giao thức sao lưu tích cực:

- Ưu điểm

Đảm bảo tính tuần tự cao.

Sử dụng một tiến trình để cập nhập các bản sao dữ liệu.

Tính sẵn sàng cao, bởi tiến trình luôn thực hiện cập nhập tất cả các bản sao.

- Nhược điểm:

Các yêu cầu của client bị nhân bản lên nhiều lần.

Tính mở rộng không cao.

Giao thức sao lưu dựa trên túc số:

- Ưu điểm:

Đảm bảo tính nhất quán của bản sao, mỗi cập nhập được tính như một giao dịch.

Tính mở rộng cao, có thể chia các máy chủ bản sao thành các nhóm.

- Nhược điểm:

Tính sẵn sàng không cao (truy cập tất cả dữ liệu hoặc không truy cập)

Tốn tài nguyên.

Tuỳ thuộc vào tài nguyên hệ thống cũng như các yêu cầu về tính sẵn sàng, tính mở rộng, tính tuần tự của dữ liệu từ đó lựa chọn phương thức phù hợp.

Câu hỏi 7*: Liên quan đến các mô hình thống nhất hướng dữ liệu và các mô hình thống nhất hướng người dùng:*

*a. Giải thích vắn tắt ý tưởng của 2 loại mô hình thống nhất hướng dữ liệu trên.*

*b. Một công ty startup mới mở chuyên triển khai thương mại hóa dịch vụ CDN (Content Delivery Network) cho 2 loại hình dịch vụ là thư điện tử và WWW. Để đảm bảo thống nhất dữ liệu cho 2 loại dịch vụ đó thì tầng middleware sẽ áp dụng mô hình thống nhất dữ liệu nào (ở câu a) cho mỗi loại dịch vụ trên? Giải thích.*

*c. Công ty đó triển khai 3000 server bản sao vật lý và chọn hình thức sao lưu dữ liệu dựa trên túc số (quorum) với Nw = 1600 và Nr = 1100. Vậy hệ thống đó sẽ tránh được xung đột đọc-ghi và xung đột ghi-ghi hay không? Giải thích.*

a.

Mô hình thống nhất hướng dữ liệu: đảm bảo thống nhất dữ liệu cho các kho dữ liệu. Khi một kho dữ liệu được cập nhập, nó sẽ cập nhập dữ liệu mới đến các kho dữ liệu khác. Các tiến trình truy cập vào kho dữ liệu sẽ cùng nhìn thấy một tuần tự của dữ liệu.

Mô hình thống nhất hướng người dùng: đảm bảo tính thống nhất dữ liệu cho từng người dùng, chỉ cập nhập dữ liệu cho các tiến trình sử dụng đến.

b.

Với dịch vụ WWW thì ta nên dùng mô hình hướng người dùng. Vì người dùng truy cập dịch vụ WWW là người dùng đơn, nên mô hình hướng người dùng sẽ đảm bảo được tính thống nhất dữ liệu hướng đến người dùng đơn này.

Với dịch vụ thư điện tử ta nên sử dụng mô hình hướng dữ liệu. Vì tham gia hệ thống thư điện tử là nhiều người dùng cùng sử dụng một hệ thống,

c.

Để tránh xung đột đọc-ghi cần đảm bảo Nr + Nw > N.

Ta có Nr + Nw = 2700.

N = 3000.

Vậy hệ thống không tránh được xung đột đọc-ghi.

Để tránh xung đột ghi-ghi cần đảm bảo Nw > N/2.

Ta có Nw = 1600, N/2 = 1500.

Vậy hệ thống sẽ tránh được xung đột ghi-ghi.