Mô hình chi phí phân tán

- Các hàm chi phí
  - Tổng thời gian (hoặc Tổng chi phí)
    - Giảm từng thành phần chi phí (về mặt thời gian) riêng
    - Thực hiện càng ít thành phần chi phí càng tốt
    - Tối ưu hóa việc sử dụng tài nguyên và tăng thông lượng hệ thống
  - Thời gian đáp ứng
    - Làm song song nhiều việc nhất có thể
    - Có thể tăng tổng thời gian do tổng số các hoạt động tăng

Tổng thời gian

Tổng thời gian = chi phí CPU + chi phí I/O + chi phí truyền thông

Tổng tất cả các yếu tố chi phí

Chi phí CPU = đơn vị chi phí lệnh ∗ số lệnh

Chi phí I/O = đơn vị chi phí I/O ổ đĩa ∗ số I/Os ổ đĩa

Chi phí truyền thông = khởi tạo thông điệp + truyền dẫn

Thời gian đáp ứng

Thời gian đáp ứng = thời gian CPU + thời gian I/O +

thời gian truyền thông

Phải xem xét đến việc thực thi song song

Thời gian CPU = đơn vị thời gian lệnh * số lệnh theo thứ tự

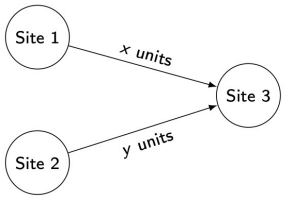
Thời gian I/O = đơn vị thời gian I/O * số I/Os theo thứ tự

Thời gian truyền thông = đơn vị thời gian khởi tạo thông điệp *

số thông điệp theo thứ tự

+ đơn vị thời gian truyền * số byte theo thứ tự

Ví dụ



- Chỉ xem xét chi phí truyền thông
  - Tổng thời gian = 2 × thời gian khởi tạo thông điệp + đơn vị thời gian truyền * (x+y)
  - Thời gian đáp ứng = max {thời gian gửi x từ 1 đến 3, thời gian gửi y từ 2 đến 3}

Thống kê cơ sở dữ liệu

- Yếu tố chi phí sơ cấp: kích thước của các quan hệ trung gian
  - Cần ước lượng kích thước của các quan hệ này
- Làm cho chúng chính xác ➔ tốn kém hơn để duy trì
- Đơn giản hóa giả thiết: phân bố đều các giá trị thuộc tính của một quan hệ.

Thống kê

- Với mỗi quan hệ $R[A_1, A_2, \ldots, A_n]$ được phân mảnh thành $R_1, \ldots, R_r$
  - Độ dài của mỗi thuộc tính: $length(A_i)$
  - Số lượng giá trị riêng cho từng thuộc tính trong mỗi mảnh: $card(\Pi_{A_i} R_j)$
  - Các giá trị lớn nhất và nhỏ nhất trong miền của từng thuộc tính: $min(A_i)$, $max(A_i)$
  - Lực lượng của mỗi miền: $card(dom[A_i])$
- Lực lượng của mỗi mảnh: $card(R_j)$
- Nhân tố lựa chọn của mỗi toán tử trên các quan hệ
  - Xem thống kê tối ưu hóa truy vấn tập trung

Tối ưu hóa truy vấn phân tán

- Cách tiếp cận động
  - Phân tán INGRES
  - Không có ước tính chi phí tĩnh, chỉ có thông tin chi phí thời gian chạy
- Cách tiếp cận tĩnh
  - Hệ thống R*
  - Mô hình chi phí tĩnh
- Cách tiếp cận lai
  - 2-bước

Cách tiếp cận động

1. Thực hiện tất cả các truy vấn đơn quan hệ (ví dụ: phép chọn, chiếu)
2. Giảm truy vấn đa quan hệ để tạo ra các truy vấn con tối giản $q_1 \rightarrow q_2 \rightarrow \ldots \rightarrow q_n$ sao cho chỉ có một quan hệ giữa q_i và q_{i+1}
3. Chọn q_i gồm các mảnh nhỏ nhất để thực thi (gọi MRQ')
4. Tìm chiến lược thực hiện tốt nhất cho MRQ'
   1) Xác định trạm xử lý
   2) Xác định các mảnh để di chuyển
5. Lặp lại bước 3 và bước 4

Cách tiếp cận tĩnh

- Hàm chi phí bao gồm xử lý cục bộ và truyền dẫn
- Chỉ xem xét các kết nối
- Tìm kiếm "đầy đủ"
- Biên dịch

Cách tiếp cận tĩnh – Thực hiện các phép kết nối

- Dịch chuyển toàn bộ
  - Truyền dữ liệu lớn hơn
  - Số lượng các thông điệp nhỏ hơn
  - Tốt hơn nếu các quan hệ nhỏ
- Dịch chuyển khi cần
  - Số lượng các thông điệp = *O*(Lực lượng của quan hệ ngoài)
  - Dữ liệu truyền trên mỗi thông điệp là tối thiểu
  - Tốt hơn nếu quan hệ lớn và phép chọn tốt

Cách tiếp cận tĩnh – Phân chia theo chiều dọc & kết nối

1. Di chuyển các bộ quan hệ ngoài đến vị trí của quan hệ bên trong
   (a) Truy xuất các bộ dữ liệu bên ngoài
   (b) Gửi chúng đến trạm quan hệ bên trong
   (c) Thực hiện kết nối khi chúng đến nơi

Tổng chi phí = chi phí (truy xuất các bộ bên ngoài đủ điều kiện)
+ số các bộ bên ngoài được tìm nạp * chi phí (truy xuất các bộ bên ngoài đủ điều kiện)
+ chi phí thông điệp * (số các bộ bên ngoài được tìm nạp * kích thước bộ bên ngoài trung bình) / kích thước thông điệp.

Cách tiếp cận tĩnh – Phân chia theo chiều dọc & kết nối

2. Di chuyển quan hệ bên trong đến trạm của quan hệ bên ngoài

Không thể kết nối khi chúng đến nơi; chúng cần được lưu trữ

Tổng chi phí = chi phí (truy xuất các bộ bên ngoài đủ điều kiện)
+ số các bộ bên ngoài được tìm nạp * chi phí (truy xuất các bộ bên trong phù hợp từ bộ lưu trữ tạm thời)
+ chi phí (truy xuất các bộ bên ngoài đủ điều kiện)
+ chi phí (lưu trữ tất cả các bộ bên trong đủ điều kiện trong bộ lưu trữ tạm thời)
+ chi phí thông điệp * số các bộ bên trong được tìm nạp * kích thước bộ bên trong trung bình / kích thước thông điệp

Cách tiếp cận tĩnh – Phân chia theo chiều dọc & kết nối

3. Di chuyển cả quan hệ bên trong và bên ngoài sang một trạm khác

Tổng chi phí = chi phí (truy xuất các bộ bên ngoài đủ điều kiện)
+ chi phí (truy xuất các bộ bên trong đủ điều kiện)
+ chi phí (lưu trữ các bộ bên trong bộ lưu trữ)
+ chi phí thông điệp * (số các bộ bên ngoài được tìm nạp * kích thước bộ bên ngoài trung bình) / kích thước thông điệp
+ chi phí thông điệp * (số các bộ bên trong được tìm nạp * kích thước bộ bên trong trung bình) / kích thước thông điệp
+ số các bộ bên ngoài được tìm nạp * chi phí (truy xuất các bộ bên trong từ bộ lưu trữ tạm thời)

Cách tiếp cận tĩnh – Phân chia theo chiều dọc & kết nối

4. Tìm nạp các bộ dữ liệu bên trong khi có yêu cầu
   (a) Truy xuất các bộ đủ điều kiện tại trạm quan hệ bên ngoài
   (b) Gửi yêu cầu chứa (các) giá trị cột kết nối cho các bộ bên ngoài tới trạm quan hệ bên trong
   (c) Truy xuất các bộ bên trong phù hợp tại trạm quan hệ bên trong
   (d) Gửi các bộ bên trong phù hợp tới trạm quan hệ bên ngoài
   (e) Kết nối khi chúng đi đến

Tổng chi phí = chi phí (truy xuất các bộ bên ngoài đủ điều kiện)
+ chi phí thông điệp * (số các bộ bên ngoài được tìm nạp)
+ số các bộ bên ngoài được tìm nạp * số các bộ bên trong được tìm nạp * kích thước bộ bên trong trung bình * (chi phí thông điệp / kích thước thông điệp).
+ số các bộ bên ngoài được tìm nạp * chi phí (truy xuất các bộ bên trong phù hợp cho một giá trị bên ngoài)