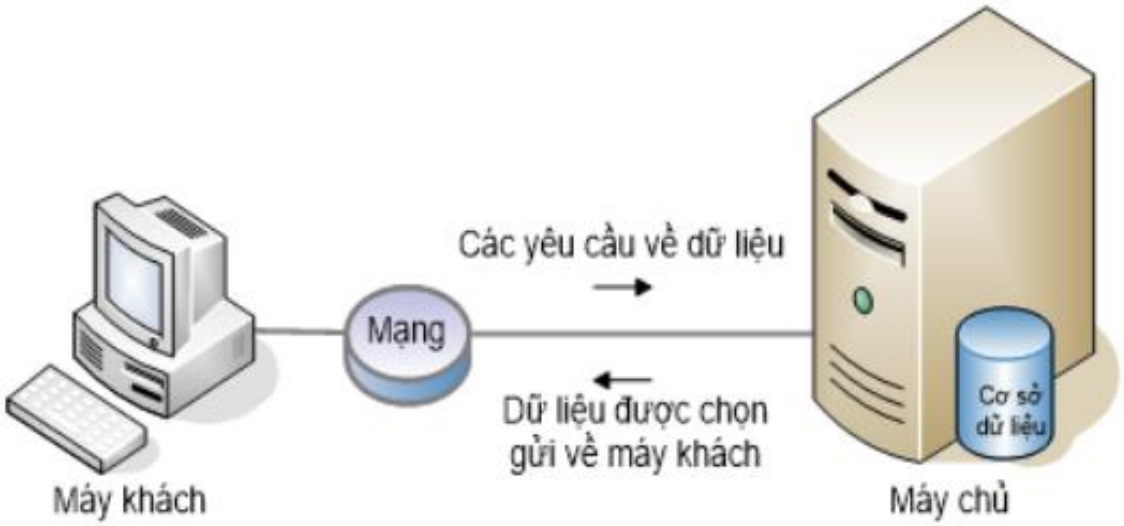
CHƯƠNG 2
Tổng quan về CSDL Phân tán

Nội dung

- Mở đầu
- Định nghĩa CSDL phân tán.
- Các đặc điểm của CSDL phân tán so với CSDL tập trung.
- Các lý do sử dụng CSDL phân tán.
- Hệ quản trị CSDL phân tán.

Mở đầu

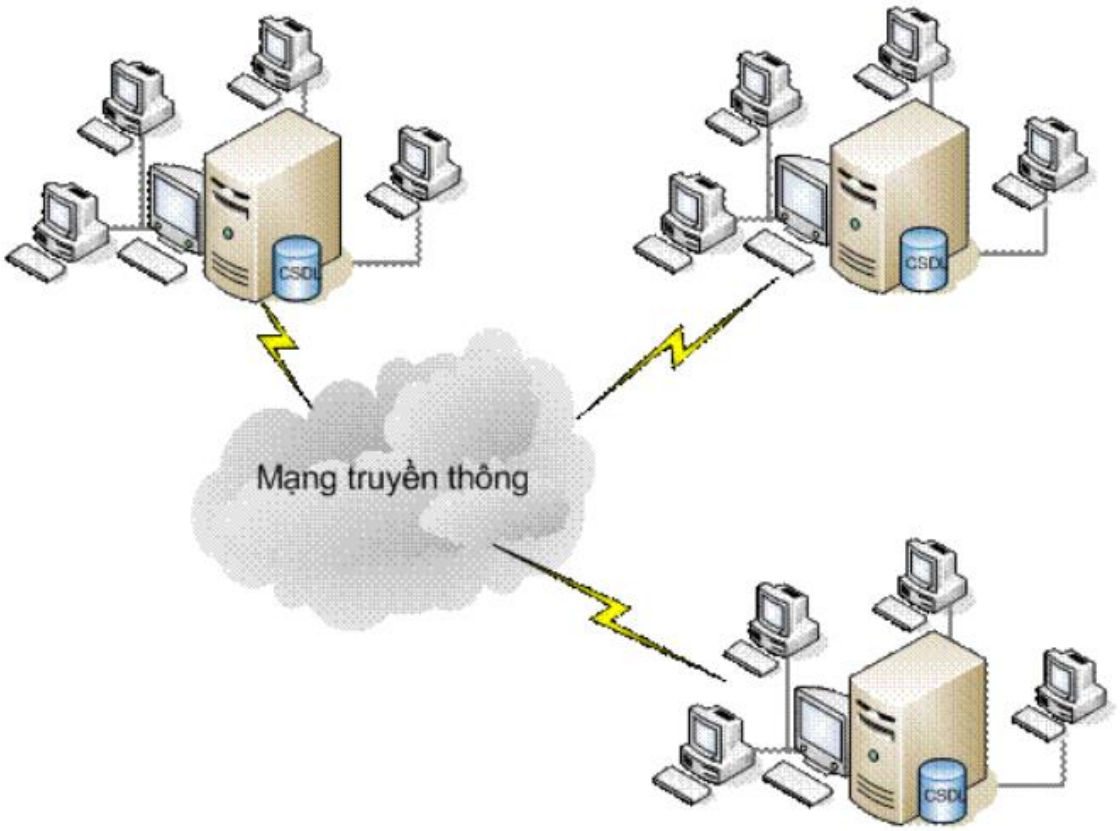


Nhu cầu thu thập, lưu trữ. xử lý và trao đổi thông tin ngày càng tăng, các hệ thống xử lý tập trung đã bộc lộ những nhược điểm sau :

Mở đầu

- Tăng khả năng lưu trữ thông tin là khó khăn, bởi bị giới hạn tối đa của thiết bị nhớ
- Độ sẵn sàng phục vụ của CSDL không cao khi số người sử dụng tăng
- Khả năng tính toán của các máy tính đơn lẻ đang dần tới giới hạn vật lý.
- Mô hình tổ chức lưu trữ, xử lý dữ liệu tập trung không phù hợp cho những tổ chức kinh tế, xã hội có hoạt động rộng lớn, đa quốc gia

Mở đầu



Những nhược điểm này đã được khắc phục khá nhiều trong hệ thống phân tán.

Các hệ thống phân tán sẽ thay thế dần các hệ thống tập trung.

❖ Hệ thống phân tán

- Hệ thống phân tán là tập hợp **các máy tính độc lập kết nối với nhau** thành một mạng máy tính, được cài đặt các hệ cơ sở dữ liệu và các phần mềm hệ thống phân tán tạo khả năng cho nhiều người sử dụng truy nhập chia sẻ nguồn thông tin chung.
- Các máy tính trong hệ thống phân tán có kết nối phần cứng lỏng lẻo, *có nghĩa là không chia sẻ bộ nhớ, chỉ có một hệ điều hành trong toàn bộ hệ thống phân tán*
- Các mạng máy tính được xây dựng dựa trên kỹ thuật Web, ví dụ như mạng Internet, mạng Intranet… là các mạng phân tán

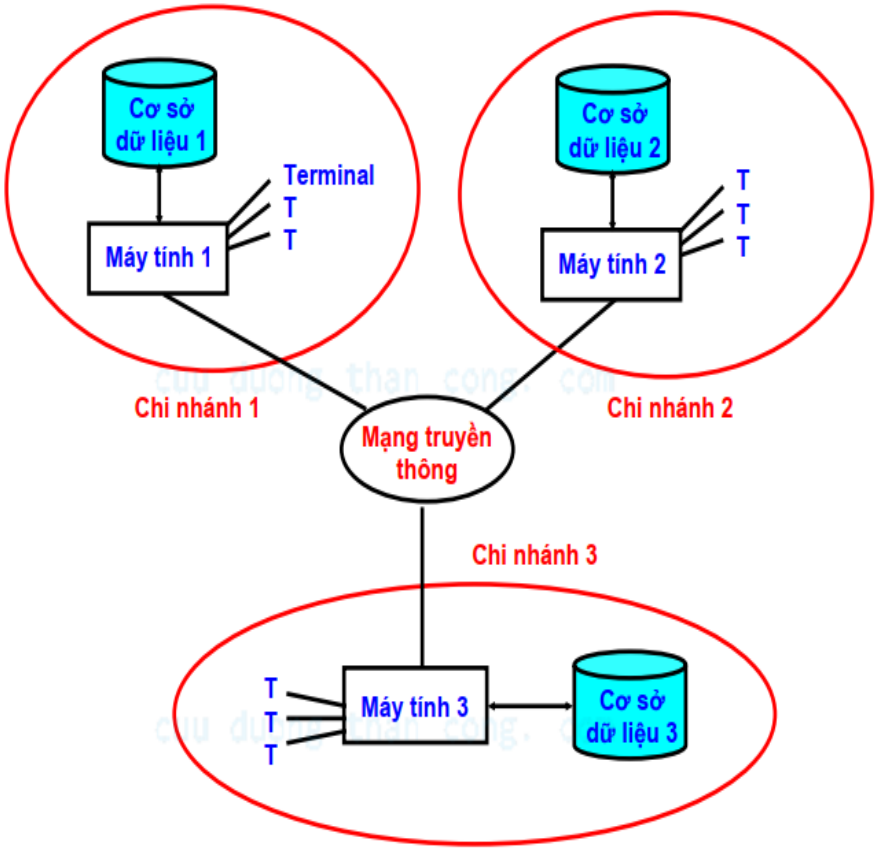
Định nghĩa CSDL phân tán

Định nghĩa 1

Cơ sở dữ liệu phân tán (distributed database) **là sự tập hợp dữ liệu mà về mặt luận lý chúng *thuộc cùng một hệ thống* nhưng *được đặt ở nhiều nơi* (*site*) của một mạng máy tính.**

- ***Sự phân tán dữ liệu (data distribution):*** **dữ liệu phải được phân tán ở nhiều nơi.**
- ***Sự tương quan luận lý (logical correlation):*** **dữ liệu của các nơi có liên hệ mật thiết với nhau, được sử dụng chung để cùng giải quyết một vấn đề.**

Định nghĩa CSDL phân tán



Hình 1.1 Cơ sở dữ liệu phân tán trên 1 mạng phân tán địa lý

-Ngân hàng có **3 chi nhánh ở các nơi khác nhau**.

-Mỗi chi nhánh có máy tính kiểm soát máy rút tiền và 1 CSDL của chi nhánh.

Mỗi máy tính và CSDL tại 1 chi nhánh tạo thành 1 site (nơi) của CSDL phân bố, các máy được nối với nhau qua mạng máy tính truyền thông (communication network)

-Các ứng dụng được thực hiện bởi máy tính của từng chi nhánh → ứng dụng cục bộ (local application),

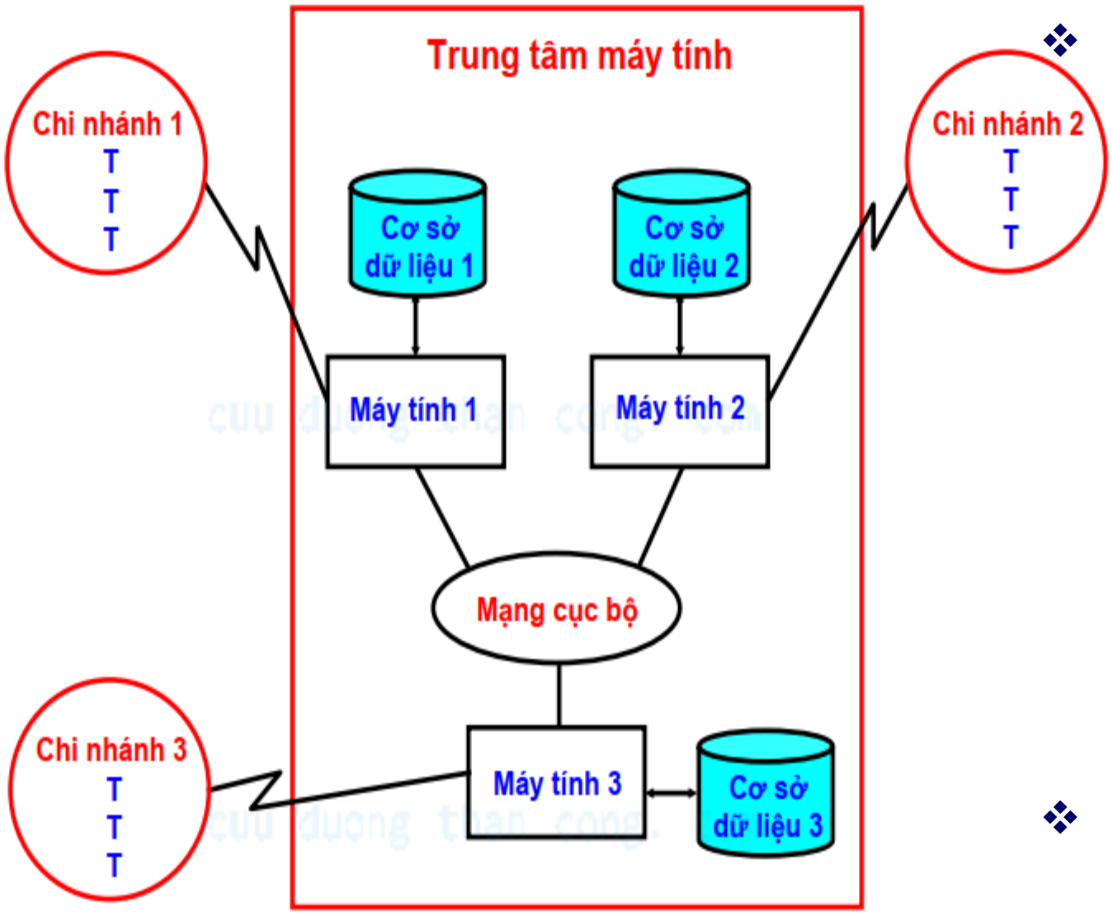
ví dụ ứng dụng ghi nợ và ghi có được thực hiện trên cùng 1 tài khoản trong cùng 1 chi nhánh.

Định nghĩa CSDL phân tán

→Sự khác biệt của CSDLPT và các CSDL cục bộ là tồn tại 1 vài ứng dụng toàn cục (global application) hoặc ứng dụng phân bố (distributed application).

- Ví dụ: ứng dụng chuyển tiền từ 1 tài khoản của CN này vào tài khoản của chi nhánh khác.

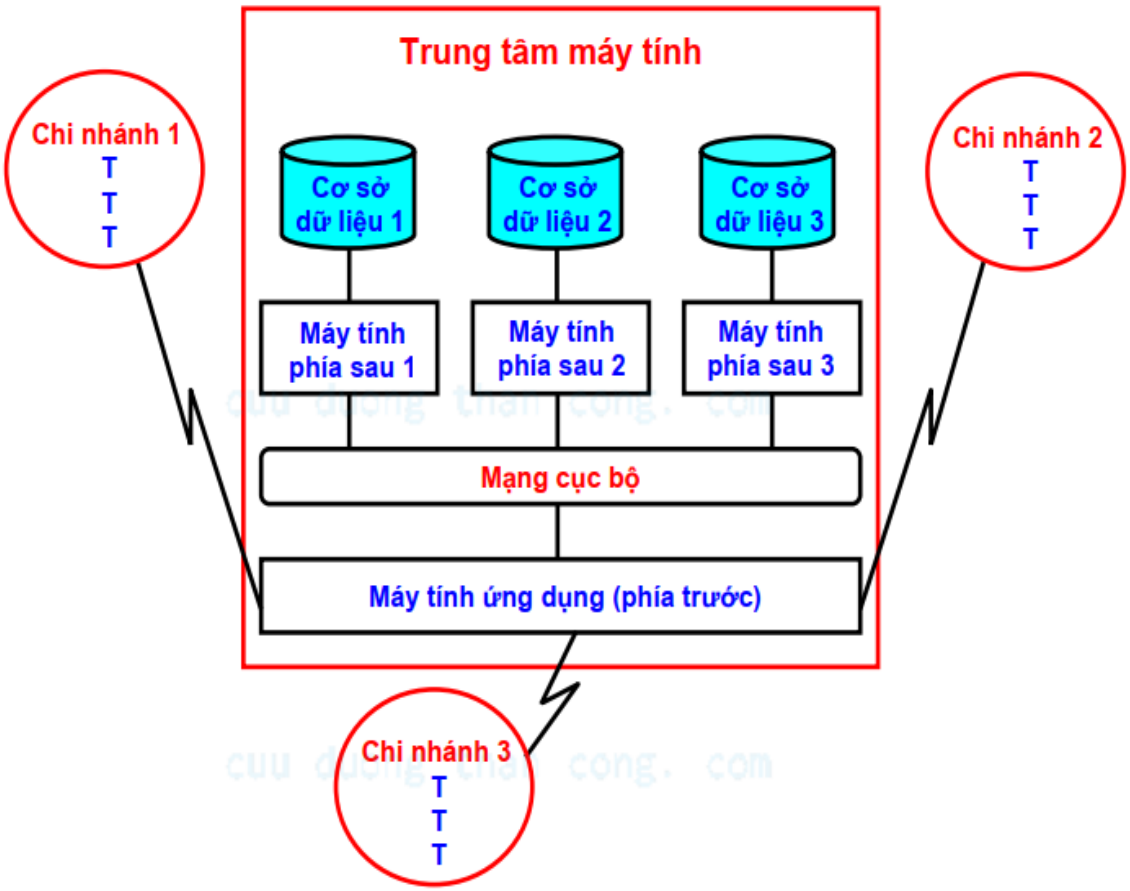
Định nghĩa CSDL phân tán



- Cấu trúc vật lý có thay đổi, tuy nhiên khía cạnh đặc trưng không thay đổi, các ứng dụng vẫn chạy trên các máy tính với CSDL tương ứng, và vẫn gọi là các ứng dụng cục bộ
- **Nếu có các ứng dụng toàn cục thì ví dụ này được xem là 1 CSDL PT**

Hình 1.2: Một CSDL phân tán trên 1 mạng cục bộ

Định nghĩa CSDL phân tán



Hình 1.3: hệ thống đa xử lý

Dữ liệu của các chi nhánh khác nhau được phân bố trên ba máy tính phía sau (backend computer) **chúng thực hiện chức năng quản trị CSDL.**

Ứng dụng được thực hiện bởi 1 máy tính khác, dữ liệu cung cấp từ các máy tính phía sau.

Định nghĩa CSDL phân tán

- Hệ thống này **không được xem** là 1 CSDLPT vì dù dữ liệu phân bố về mặt vật lý, nhưng không phù hợp với quan điểm về ứng dụng vì không có sự tồn tại của ứng dụng cục bộ (các máy tính phía sau không có khả năng thực hiện 1 ứng dụng)

Định nghĩa 2

Cơ sở dữ liệu phân tán là sự tập hợp dữ liệu *được phân tán* trên các máy tính khác nhau của một mạng máy tính. Mỗi nơi của mạng máy tính có khả năng xử lý tự trị (autonomous processing capability) và có thể *thực hiện* các ứng dụng cục bộ. Mỗi nơi cũng *tham gia thực hiện* ít nhất một ứng dụng toàn cục, mà nơi này yêu cầu truy xuất dữ liệu ở nhiều nơi bằng cách dùng hệ thống truyền thông con. (communication subsystem)

Định nghĩa CSDL phân tán

Định nghĩa 2

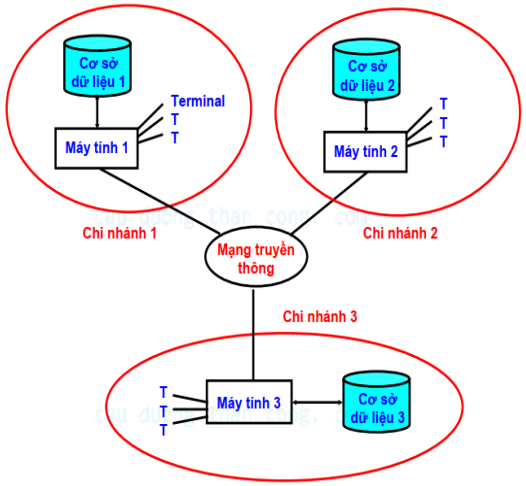
- ***Sự phân tán dữ liệu* (*data distribution*):** dữ liệu phải được phân tán ở nhiều nơi.
- ***Ứng dụng cục bộ* (*local application*):**mỗi nơi phải thực hiện ít nhất 1 ứng dụng cục bộ
- ***Ứng dụng toàn cục* (hoặc ứng dụng phân tán) (*global application / distributed application*):** mỗi nơi phải tham gia vào sự thực hiện của ít nhất 1 ứng dụng toàn cục

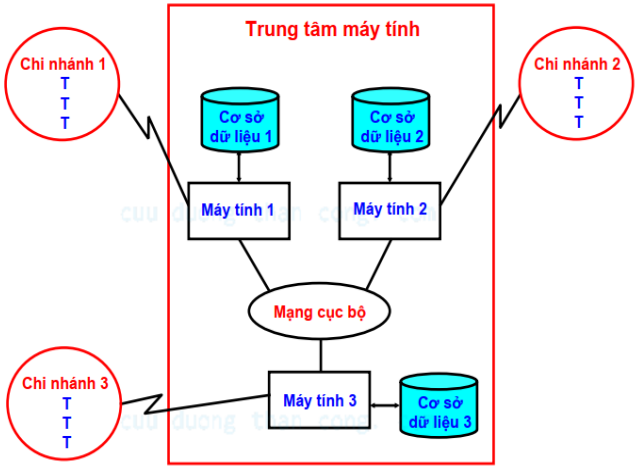
Các đặc điểm của CSDL phân tán so với CSDL tập trung

- Điều khiển tập trung
- Độc lập dữ liệu
- Giảm dư thừa
- Các cấu trúc vật lý phức tạp và truy xuất hiệu quả
- Tính toàn vẹn, phục hồi, điều khiển đồng thời
- Tính riêng biệt và tính bảo mật

❖ Điều khiển tập trung

- **Điều khiển tập trung (centralized control):** đây là vấn đề quan trọng đối với các tổ chức ví dụ điều khiển tập trung trên các tài nguyên thông tin (dữ liệu, tập tin…). Trong CSDLPT việc điều khiển tập trung phụ thuộc vào kiến trúc (ví dụ: 1.2 thích hợp hơn là 1.1 cho việc điều khiển tập trung)





Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Điều khiển tập trung

- **Người quản trị CSDL toàn cục (global database administrator):** chịu trách nhiệm chính về toàn bộ CSDL
- **Người quản trị CSDL cục bộ (local database administrator):** có trách nhiệm về các CSDL cục bộ của họ
- **Tính tự trị vị trí (site autonomy):** người quản trị CSDL cục bộ có thể có mức độ tự trí cao, đến mức không cần người quản trị CSDL toàn cục, sự phối hợp giữa các nơi (intersite coordination) được thực hiện bởi chính những người quản trị cục bộ.

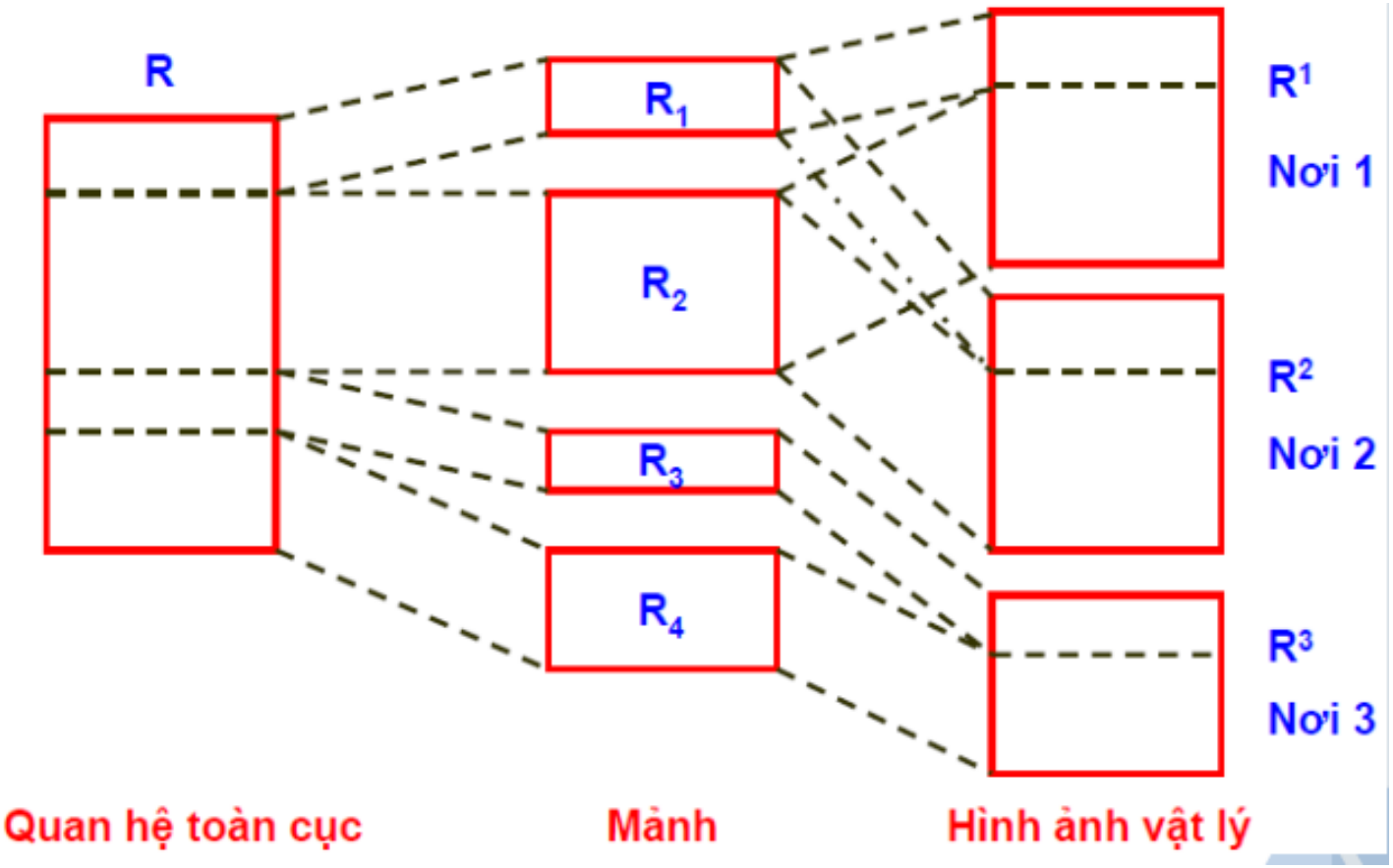
Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Độc lập dữ liệu

- Là tổ chức hiện tại của dữ liệu là trong suốt đối với người lập trình ứng dụng
- Ưu điểm là chương trình không bị ảnh hưởng bởi những thay đổi về tổ chức vật lý của dữ liệu.

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Độc lập dữ liệu



Các mảnh và các hình ảnh vật lý của một quan hệ toàn cục

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Độc lập dữ liệu

➢ Trong suốt dữ liệu

- Trong suốt phân mảnh
- Trong suốt vị trí
- Trong suốt nhân bản
- Trong suốt ánh xạ cục bộ
- Trong suốt phân tán

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Độc lập dữ liệu

➢ Trong suốt phân mảnh

- Không nhìn thấy các mảnh
- Nhìn thấy các quan hệ toàn cục (global relation)
- Lược đồ toàn cục (global schema)

➢ Trong suốt vị trí

- Không nhìn thấy các quan hệ cục bộ
- Nhìn thấy các mảnh (fragment)
- Lược đồ phân mảnh (fragmentation schema)

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Độc lập dữ liệu

- **Trong suốt nhân bản (Replication transparency)**
  - Nhìn thấy các mảnh
  - Không nhìn thấy nhân bản các mảnh
- **Trong suốt trong ánh xạ cục bộ (local mapping transparency)**
  - Nhìn thấy các quan hệ cục bộ (local relation)
  - Không nhìn thấy CSDL vật lý
- **Trong suốt trong phân bố (distribution transparency):** gồm bốn tính trong suốt trên

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Giảm dư thừa

- **Trong CSDL truyền thống, dư thừa dữ liệu được giảm càng nhiều càng tốt:**
  - Không nhất quán dữ liệu (inconsistency)
  - Tiết kiệm vùng nhớ lưu trữ
- **Tuy nhiên đối với CSDL PT thì dư thừa dữ liệu có thể là 1 ưu điểm:**
  - **Tính cục bộ (locality)** của ứng dụng cao (dữ liệu được phân bản ở các nơi mà ứng dụng cần)
  - **Tính sẵn sàng (availability)** của dữ liệu cao (nếu dữ liệu 1 nơi bị hỏng thì không làm ngưng các ứng dụng)

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Giảm dư thừa

- **Nhân bản dữ liệu (data replication)** dữ liệu được lưu trữ thành nhiều bản, cần cân nhắc các ứng dụng:
  - **Ứng dụng chỉ đọc (readonly application):** việc lấy dữ liệu có thể thực hiện trên bất kỳ bản sao nào
  - **Ứng dụng cập nhật (update applicaition):** phải cập nhật nhất quán trên tất cả bản sao.

→Hệ thống có nhiều ứng dụng chỉ đọc và ít ứng dụng cập nhật thì càng có ưu điểm và ngược lại

Các đặc điểm của CSDL phân tán so với CSDL tập trung

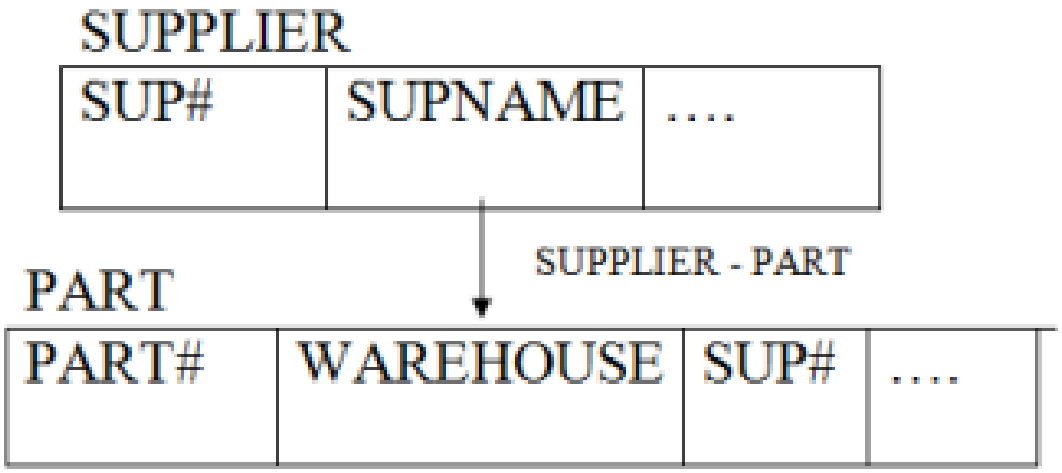
❖ Các cấu trúc vật lý phức tạp và truy xuất hiệu quả

- Trong CSDLTT các cấu trúc truy xuất phức tạp như chỉ mục thứ cấp (index), chuỗi kết nối…các cấu trúc này hỗ trợ cho việc truy xuất hiệu quả
- Trong CSDL phân bố cấu trúc truy xuất phức tạp không phải là một công cụ đúng để truy xuất dữ liệu hiệu quả.

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Các cấu trúc vật lý phức tạp và truy xuất hiệu quả

- Xét lược đồ sau



Các đặc điểm của CSDL phân tán so với CSDL tập trung

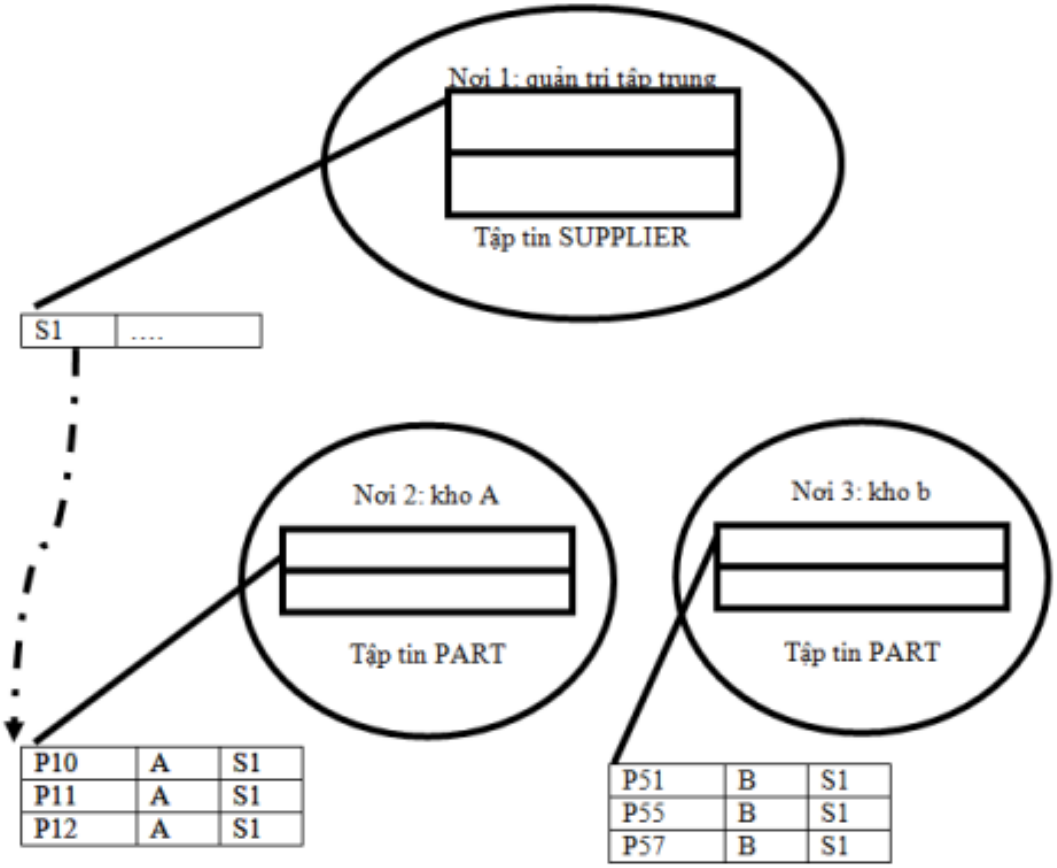
- **Các cấu trúc vật lý phức tạp và truy xuất hiệu quả**
  - **Ứng dụng: "tìm tất cả các mẩu tin của Part được cung cấp bởi nhà cung cấp S1 "**

```
Find SUPPLIER record with SUP# = S1
Repeat until "no more members in set"
        Find next PART record in SUPPLIER-PART set;
Output PART record
```

Truy xuất trên từng mẩu tin

Các đặc điểm của CSDL phân tán so với CSDL tập trung

- **Các cấu trúc vật lý phức tạp và truy xuất hiệu quả**



Các đặc điểm của CSDL phân tán so với CSDL tập trung

- **Các cấu trúc vật lý phức tạp và truy xuất hiệu quả**

Find all: gom tất cả các truy xuất được thực hiện trên cùng một nơi

At site 1:

Lấy dữ liệu nhà cung cấp S1: Gởi đến site 2 và 3

At site 2 và 3

Thực hiện song song

Find all PARTS records having

SUP#=S1

Send result to site 1

At site 1

Trộn kết quả từ site 2 và 3

Output kết quả

❖ Các cấu trúc vật lý phức tạp và truy xuất hiệu quả

Các cấu trúc vật lý phức tạp và truy xuất hiệu quả là 1vấn đế quan trọng trong CSDLPT

- Vấn đề này được phân chia thành 2 loại:
  - **Tối ưu hóa toàn cục (global optimization):** xác định dữ liệu nào phải được truy xuất tại nơi nào, tập tin dữ liệu nào phải được truyền giữa các nơi nào.
  - **Tối ưu hóa cục bộ (local optimization):** xác định CSDL cục bộ được truy xuất như thế nào tại mỗi nơi.

Các đặc điểm của CSDL phân tán so với CSDL tập trung

- **Tính toàn vẹn, phục hồi, điều khiển đồng thời**

Đây là các vấn đề có liên quan chặt chẽ với nhau, giải pháp là sử dụng các giao tác (transaction)

- **Giao tác** là một đơn vị thực hiện nguyên tố, nghĩa là một chuỗi các tác vụ hoặc tất cả đều được thực hiện hoặc tất cả đều không được thực hiện.

- **Giao tác toàn cục (global transaction):** là một ứng dụng toàn cục (ví dụ ứng dụng chuyển quỹ trong hình 1.1 ghi nợ và ghi có)

→ giao tác đảm bảo tính toàn vẹn của CSDL, chuyển CSDL từ trạng thái nhất quán này sang trạng thái nhất quán khác, hoặc trở về trạng thái ban đầu.

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Tính toàn vẹn, phục hồi, điều khiển đồng thời

- **2 vấn đề ảnh hưởng đến tính nguyên tố:**
  - **Sự hư hỏng:** hệ thống ngưng hoạt động khi đang thực hiện giao dịch
  - **Tính đồng thời:** thực hiện đồng thời các giao dịch khác nhau có thể cho phép 1 giao dịch chuyển sang 1 trạng thái không nhất quán được tạo ra bởi các giao dịch khác
- **Điều khiển đồng thời:** đảm bảo tính nguyên tố của giao dịch khi có sự thực hiện đồng thời các giao dịch khác → vấn đề đồng bộ hóa trong CSDL phân bố

Các đặc điểm của CSDL phân tán so với CSDL tập trung

❖ Tính riêng biệt và tính bảo mật

- **Thực hiện truy xuất dữ liệu cho thẩm quyền: giống như các CSDL truyền thống**
- **Tính bảo mật CSDL cục bộ:** CSDLPT có mức độ tự trị vị trí cao, cần thực hiện các bảo vệ riêng biệt.
- **Bảo mật mạng truyền thông:** CSDLPT được triển khai trong các mạng truyền thông do đó cũng cần được bảo vệ.

Tại sao phải sử dụng CSDL PT

Thảo luận nhóm

- **Tại sao nên sử dụng CSDL Phân tán?**

Tại sao phải sử dụng CSDL PT

- **Các lý do về tổ chức và kinh tế:** nhiều tổ chức không được tập trung hóa, vấn đề chi phí khi đầu tư các trung tâm máy tính lớn
- **Sự kết nối của các CSDL hiện tại:** có càng nhiều ứng dụng toàn cục kết nối các CSDL đã tồn tại, tốn ít nguồn lực hơn so với việc đầu tư xây dựng CSDL tập trung lại từ đầu.

Tại sao phải sử dụng CSDL PT

- **Sự lớn mạnh gia tăng của các tổ chức:** có thêm các đơn vị tổ chức mới (chi nhánh, kho…), CSDL phân tán hỗ trợ sự lớn mạnh và giảm thiểu ảnh hưởng ít nhất đến các đơn vị đã tồn tại.
- **Giảm chi phí truyền thông:** với trường hợp CSDL phân bố về mặt địa lý, thì nhiều ứng dụng cục bộ sẽ làm giảm chi phí truyền thông → cực đại hóa tính cục bộ của các ứng dụng là mục tiêu chính của CSDLPT

Tại sao phải sử dụng CSDL PT

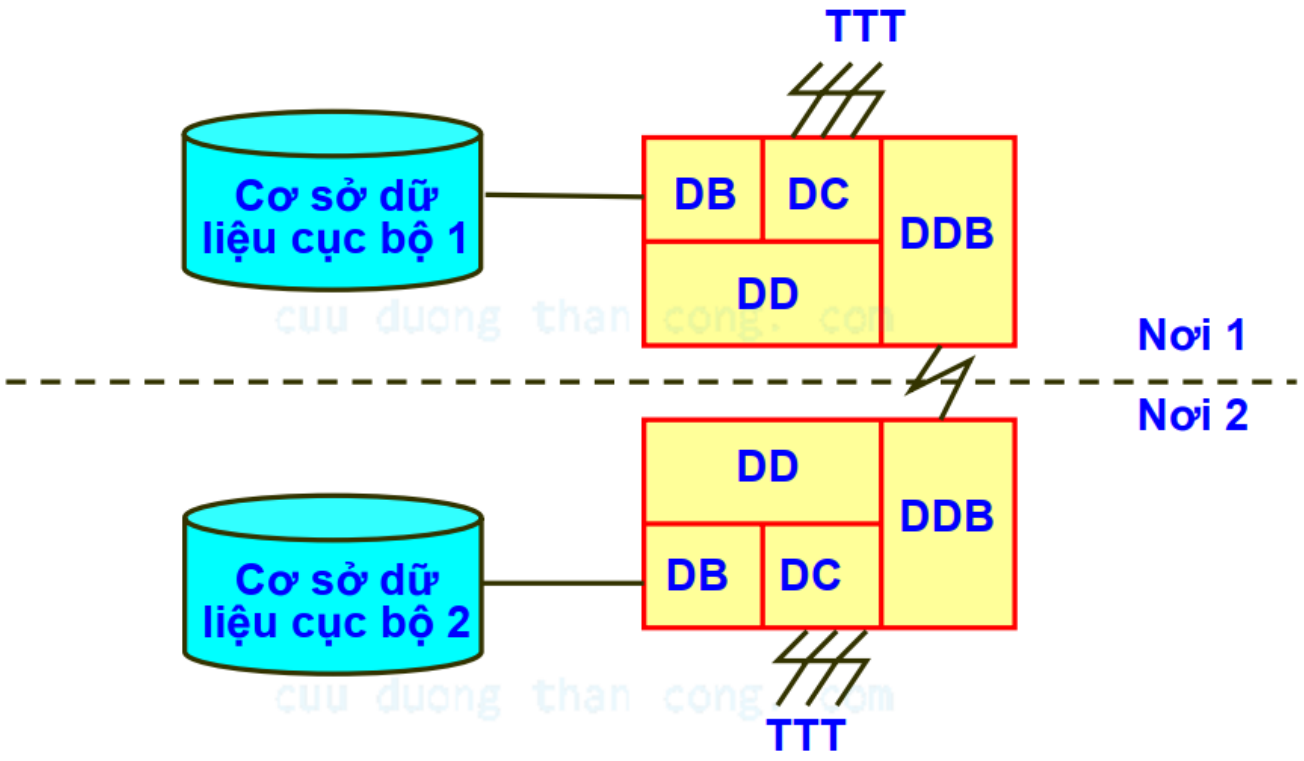
- **Các nghiên cứu về mặt hiệu suất:** khả năng xử lý tự trí đối với các ứng dụng cục bộ làm tăng hiệu quả, giảm hiện tượng thắt cổ chai các vấn đề như mạng truyền thông hoặc các dịch vụ chung của toàn bộ hệ thống
- **Độ tin cậy và sẵn sáng:** sự dư thừa dữ liệu trong CSDLPT giúp đạt được tính sẵn sàng vá độ tin cậy cao. Mặc dù các hư hỏng trong CSDL PT có thể xảy ra nhiều hơn (do khả năng xử lý tự trị tại các nơi khác nhau) nhưng ảnh hưởng sẽ bị hạn chế, hiếm khi hệ thống bị ngưng hoàn toàn..

Tại sao phải sử dụng CSDL PT

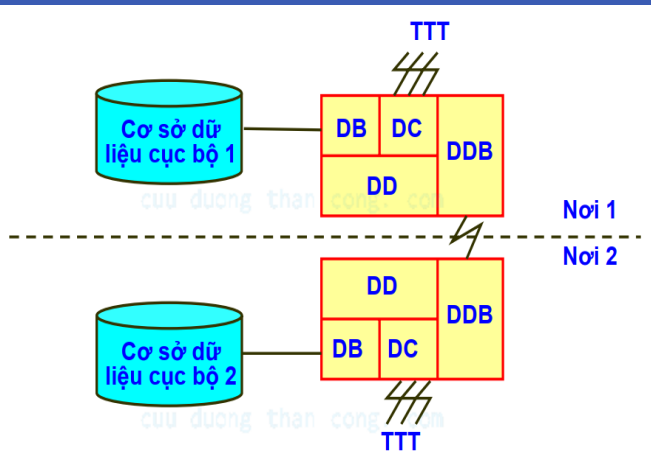
Thảo luận nhóm

- Các động cơ không phải là mới, vậy tại sao sự phát triển của CSDLPT lại mới bắt đầu??

❖ Các thành phần của DDBMS



Hệ quản trị CSDL PT (Distributed DBMS)

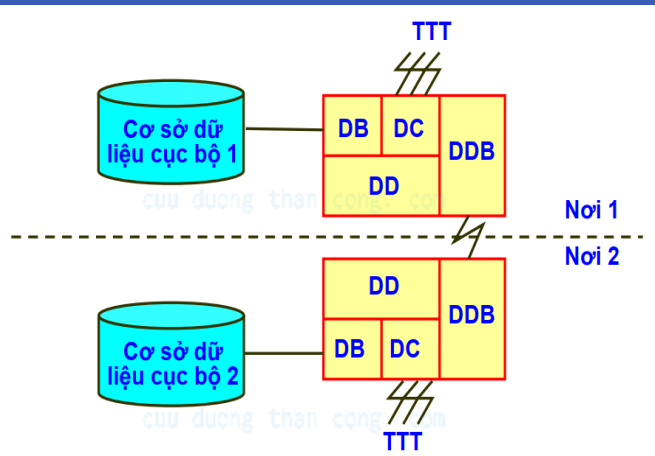


❖ Các thành phần của DDBMS

- ***Truyền thông dữ liệu:* DC–*Data Communication***
  - Nhận yêu cầu truy xuất dữ liệu của ứng dụng chạy tại thiết bị đầu cuối
  - Trả kết quả về cho ứng dụng.
- ***Quản trị CSDL:* DB – *DataBase management***
  - Quản lý CSDL
  - Thực hiện các yêu cầu của ứng dụng: xử lý dữ liệu (*data processing*).

❖ Các thành phần của DDBMS

- ***Từ điển dữ liệu:* DD – *Data Dictionary***
  - Lưu trữ thông tin về các đối tượng dữ liệu trong CSDL.
  - Lưu trữ thông tin về sự phân tán dữ liệu tại các nơi.
- ***CSDL phân tán* DDB – *Distributed DataBase***
  - Liên lạc giữa các nơi: gửi yêu cầu và nhận kết quả.

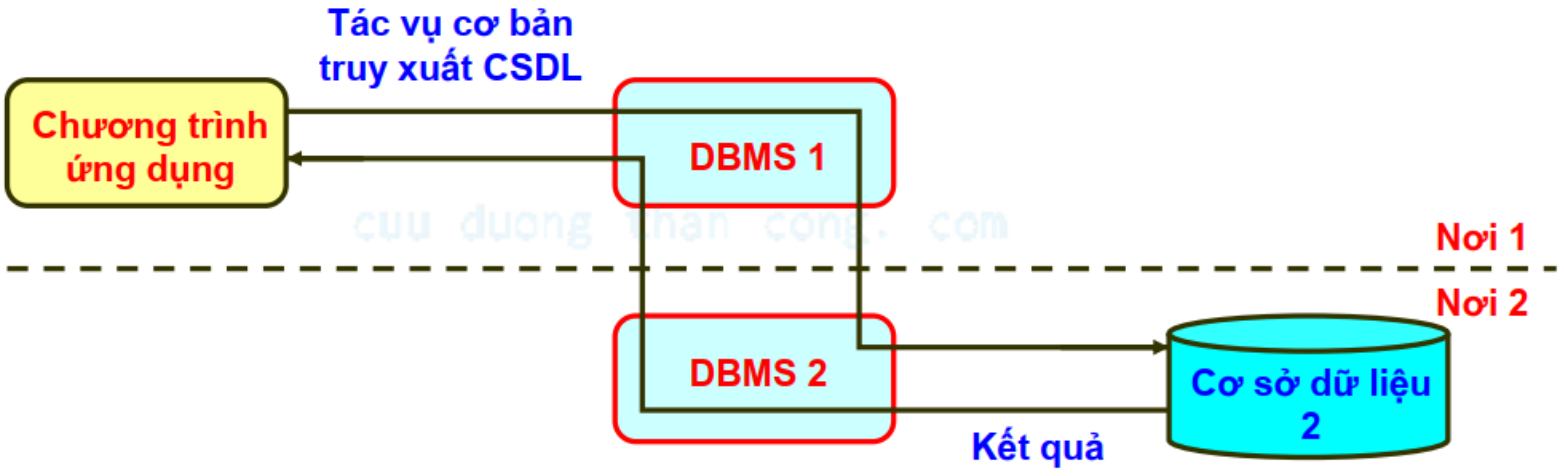


Hệ quản trị CSDL PT (Distributed DBMS)

❖ Các chức năng tiêu biểu của DDBMS

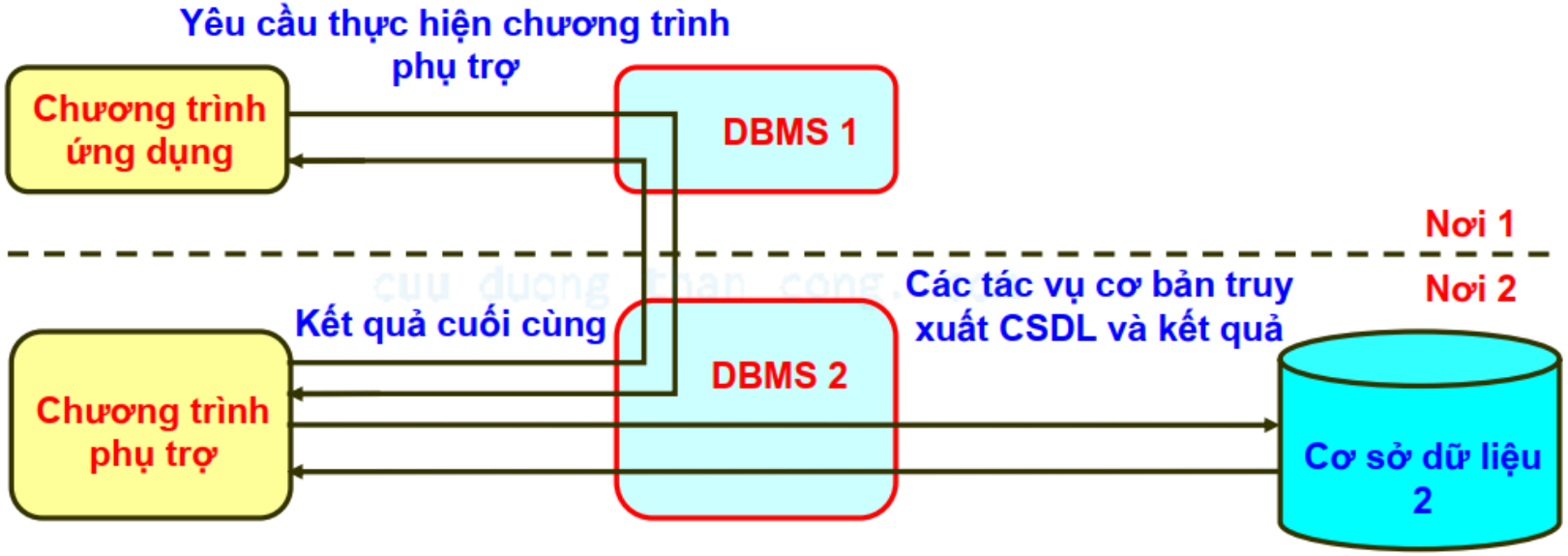
- Truy xuất CSDL từ xa bởi chương trình ứng dụng
- Hỗ trợ một số mức trong suốt phân tán: tùy vào các hệ thống khác nhau
- Hỗ trợ cho việc quản trị CSDL phân tán: giám sát CSDL, thu thập thông tin về việc sử dụng, cái nhìn toàn cục về các tập tin dữ liệu tại các nơi khác nhau
- Hỗ trợ cho việc điều khiển tương tranh và phục hồi các giao tác phân tán.

- **Truy xuất CSDL từ xa bởi 1 ứng dụng: được thực hiện bởi 1 trong 2 cách**



Truy xuất từ xa thông qua các tác vụ cơ bản của DBMS

- **Truy xuất CSDL từ xa bởi 1 ứng dụng: được thực hiện bởi 1 trong 2 cách**



Truy xuất từ xa thông qua một chương trình phụ trợ

Hệ quản trị CSDL PT (Distributed DBMS)

- **Tính đồng nhất (homogeneous) và không đồng nhất (heterogeneous):**
  - Phần cứng
  - Hệ điều hành
  - Các DBMS cục bộ

→ tập trung vào các DBMS

Hệ quản trị CSDL PT (Distributed DBMS)

- **DDBMS đồng nhất:**
  - Các DBMS tại mỗi nơi là giống nhau
- **DDBMS không đồng nhất:**
  - Có ít nhất 2 DBMS khác nhau.
  - Vấn đề chuyển đổi mô hình dữ liệu khác nhau của các DBMS cục bộ khác nhau → vấn đề khó.