


Chương 1: Tổng quan về cơ sở dữ liệu phân tán

Thời lượng: 3 tiết

GV: ThS. Thái Bảo Trân

Nội dung

- 1. Xử lý phân tán**
 - 2. Định nghĩa CSDL phân tán.**
 - 3. Các đặc điểm của CSDL phân tán so với CSDL tập trung.**
 - 4. Tại sao sử dụng CSDL phân tán.**
 - 5. Hệ quản trị CSDL phân tán.**
 - 6. Triển vọng của các hệ cơ sở dữ liệu phân tán.**
 - 7. Phân mảnh dữ liệu (Fragmentation)**
- 

1. Xử lý phân tán

- ▶ Xử lý phân tán là một số **bộ phận xử lý** tự vận hành được liên kết bởi một mạng máy tính và thực hiện các nhiệm vụ mà chúng được phân công.
- ▶ Các bộ phận xử lý là các thiết bị tính toán có thể chạy được một chương trình trên chính nó.

1. Xử lý phân tán (tt)

► Những gì được phân tán:

- **Các thiết bị xử lý**
- **Chức năng:** Nhiều chức năng của hệ thống máy tính có thể được chuyển giao cho các thành phần phần cứng và phần mềm.
- **Dữ liệu:** Dữ liệu được dùng bởi 1 số ứng dụng có thể được phân tán cho 1 số vị trí xử lý.
- **Quyền điều khiển:** Quyền điều khiển việc thực hiện 1 số nhiệm vụ cũng có thể được phân tán.

1. Xử lý phân tán (tt)

► Phân loại các hệ thống xử lý phân tán:

- Mức độ kết nối
- Sự liên quan giữa các thành phần
- Cấu trúc tương giao
- Sự đồng bộ hóa giữa các thành phần.

► Tại sao chúng ta lại thực hiện phân tán ?

- Nhằm thích ứng tốt hơn với việc phân bố rộng rãi của các công ty, xí nghiệp.
- Quan trọng hơn, nhiều ứng dụng hiện tại của công nghệ máy tính cũng được phân tán.

1. Xử lý phân tán (tt)

- ▶ **Ưu điểm cơ bản việc xử lý phân tán:**
 - Tận dụng được sức mạnh tính toán bằng cách sử dụng nhiều bộ phận xử lý một cách tối ưu đòi hỏi phải nghiên cứu các hệ thống phân tán và hệ thống xử lý song song.
 - Giải quyết bài toán theo từng nhóm hoạt động khá độc lập. Do đó, có thể kiểm soát được chi phí phát triển phần mềm.

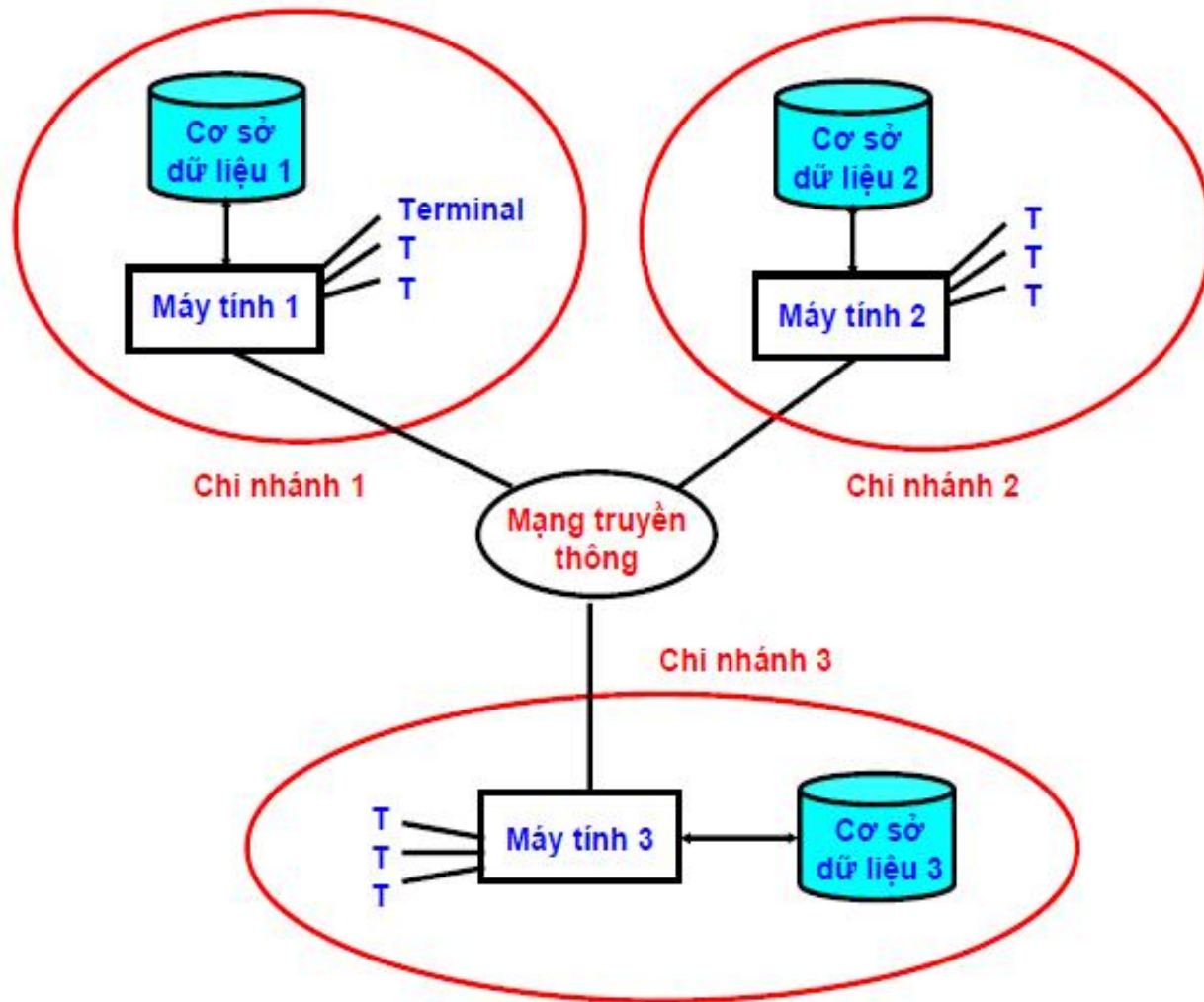
2. Định nghĩa cơ sở dữ liệu phân tán

► Định nghĩa 1:

Cơ sở dữ liệu phân tán (distributed database)

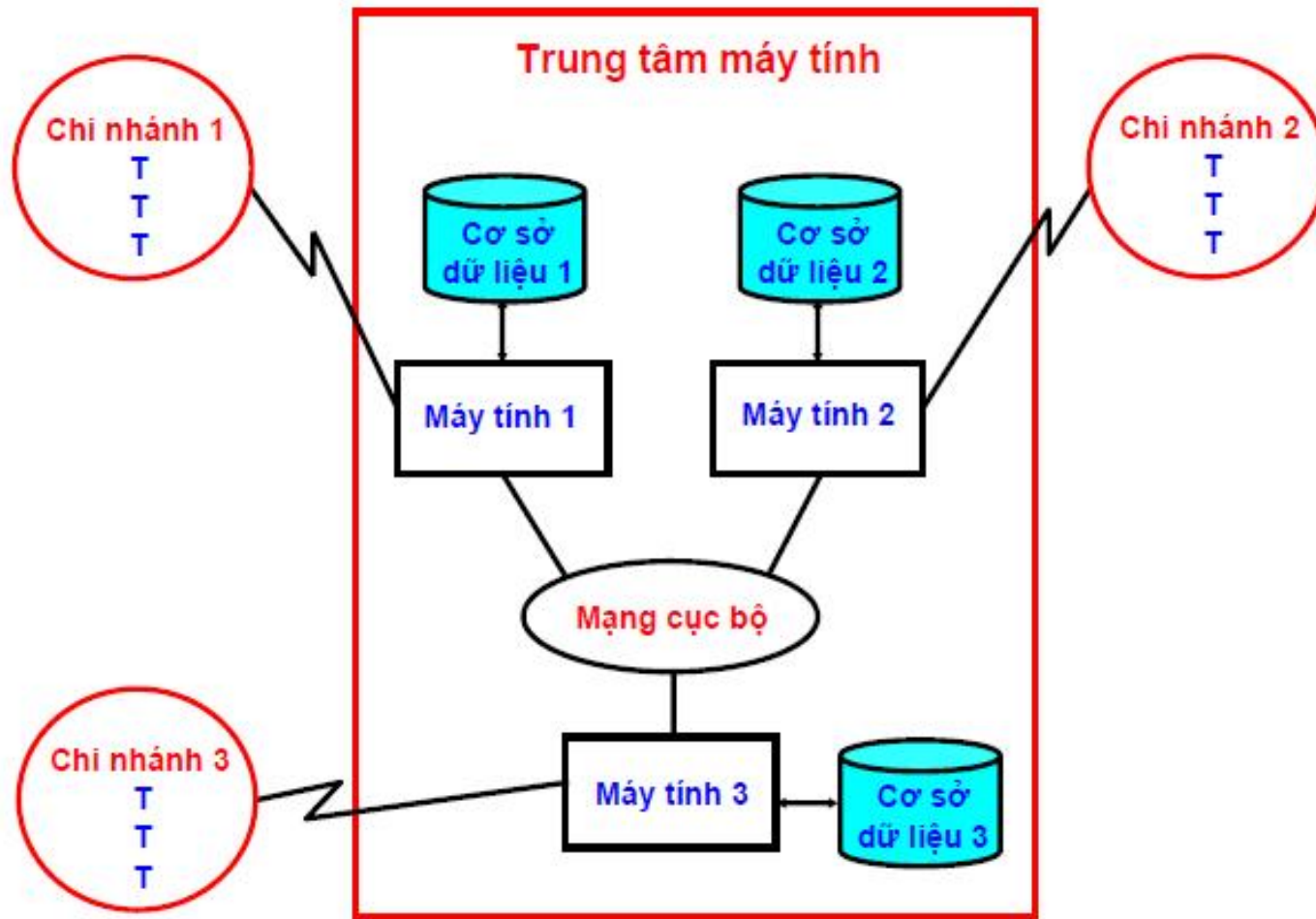
- Là sự tập hợp dữ liệu mà về mặt luận lý chúng thuộc cùng một hệ thống nhưng được đặt ở nhiều nơi (site) của một mạng máy tính.
- **Sự phân tán dữ liệu (data distribution):** dữ liệu phải được phân tán ở nhiều nơi.
- **Sự tương quan luận lý (logical correlation):** dữ liệu của các nơi được sử dụng chung để cùng giải quyết các vấn đề.

2. Định nghĩa cơ sở dữ liệu phân tán



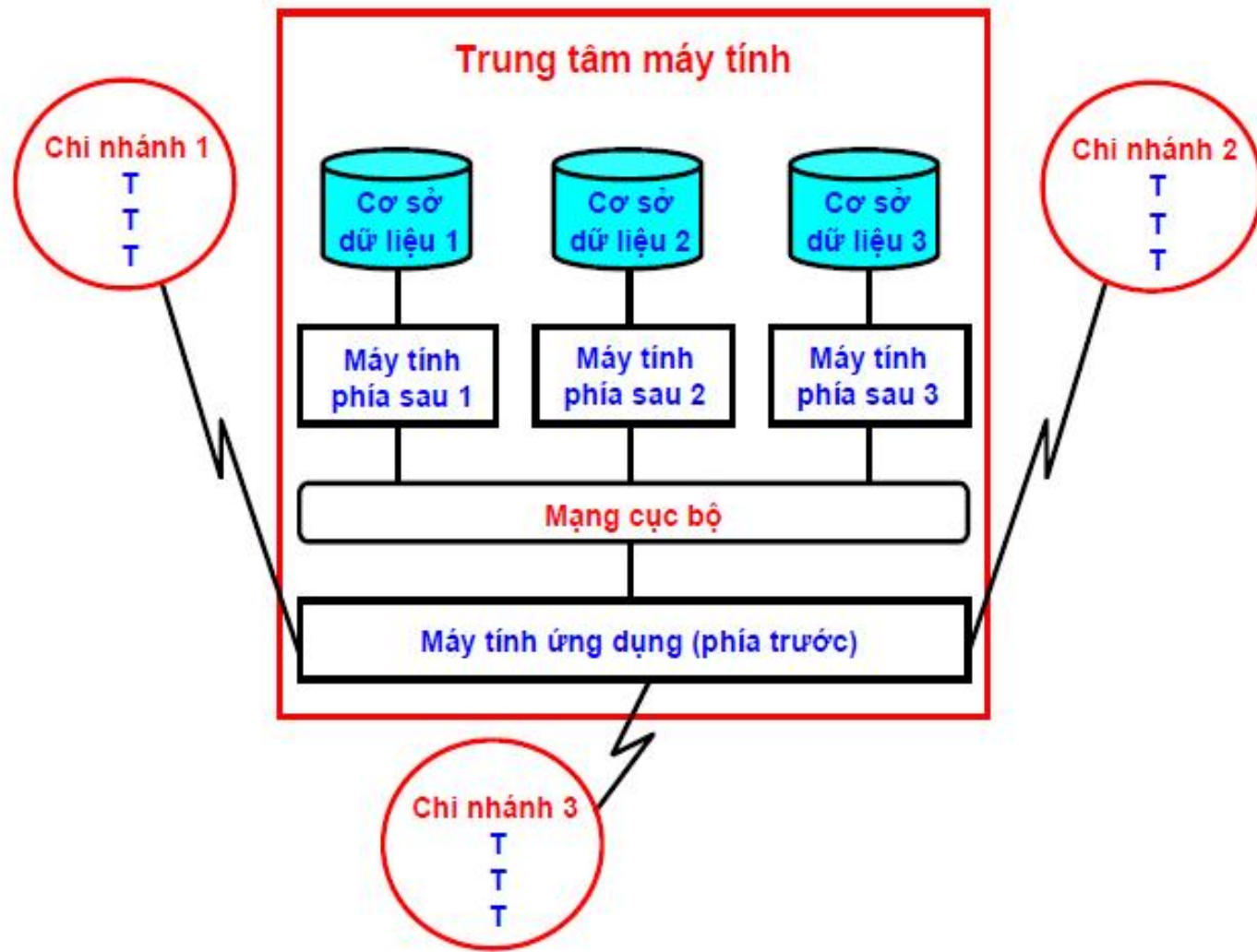
Hình 1.1. Cơ sở dữ liệu phân tán trên một mạng phân tán địa lý.

2. Định nghĩa cơ sở dữ liệu phân tán



Hình 1.2. Cơ sở dữ liệu phân tán trên một mạng cục bộ.

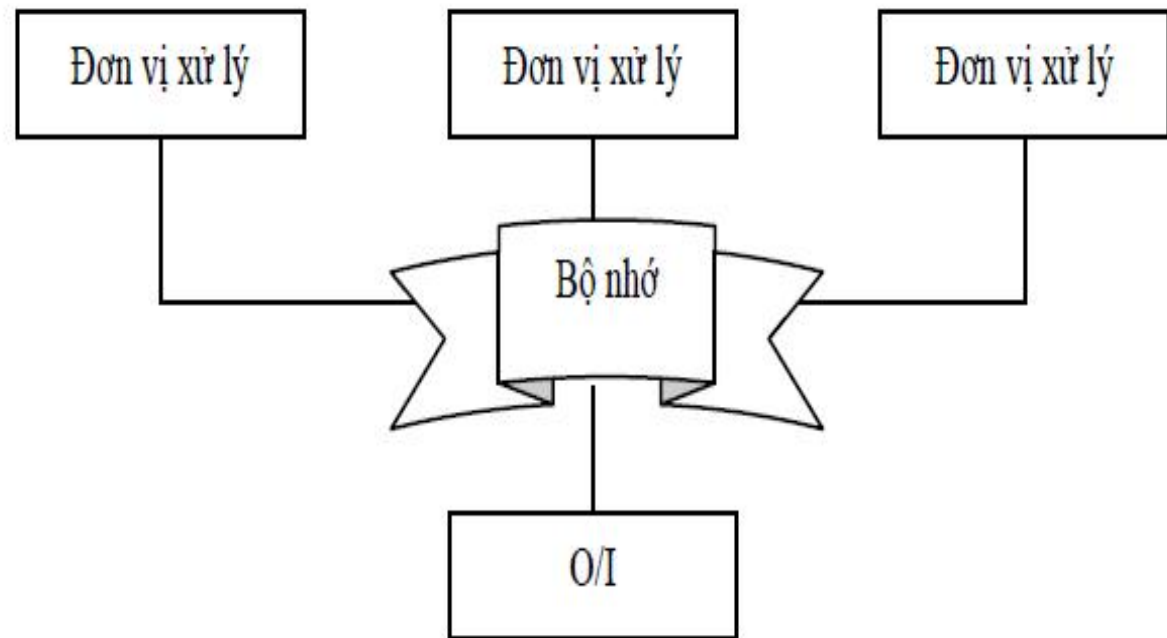
2. Định nghĩa cơ sở dữ liệu phân tán



Hình 1.3. Hệ thống đa xử lý (*multiprocessor system*).

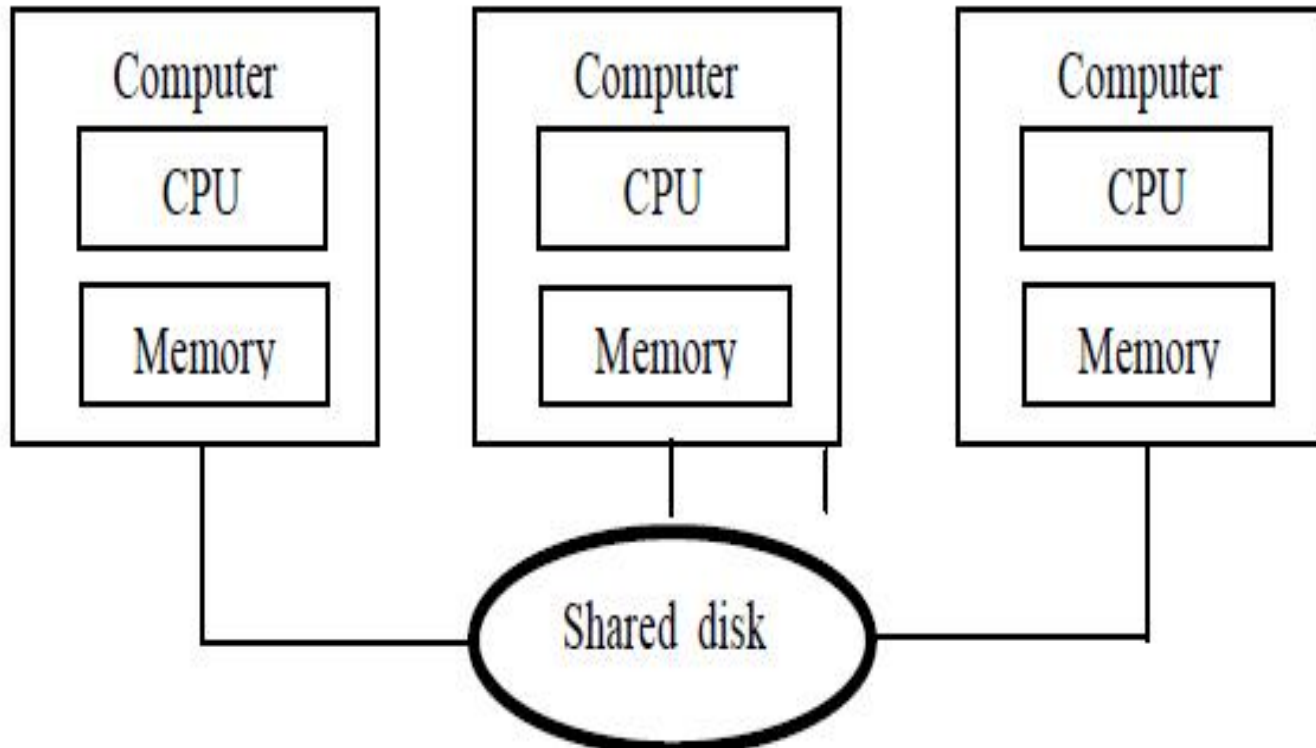
Hệ thống đa xử lý là hệ thống có nhiều đơn vị xử lý cùng dùng chung một dạng bộ nhớ.

Ví dụ 1:



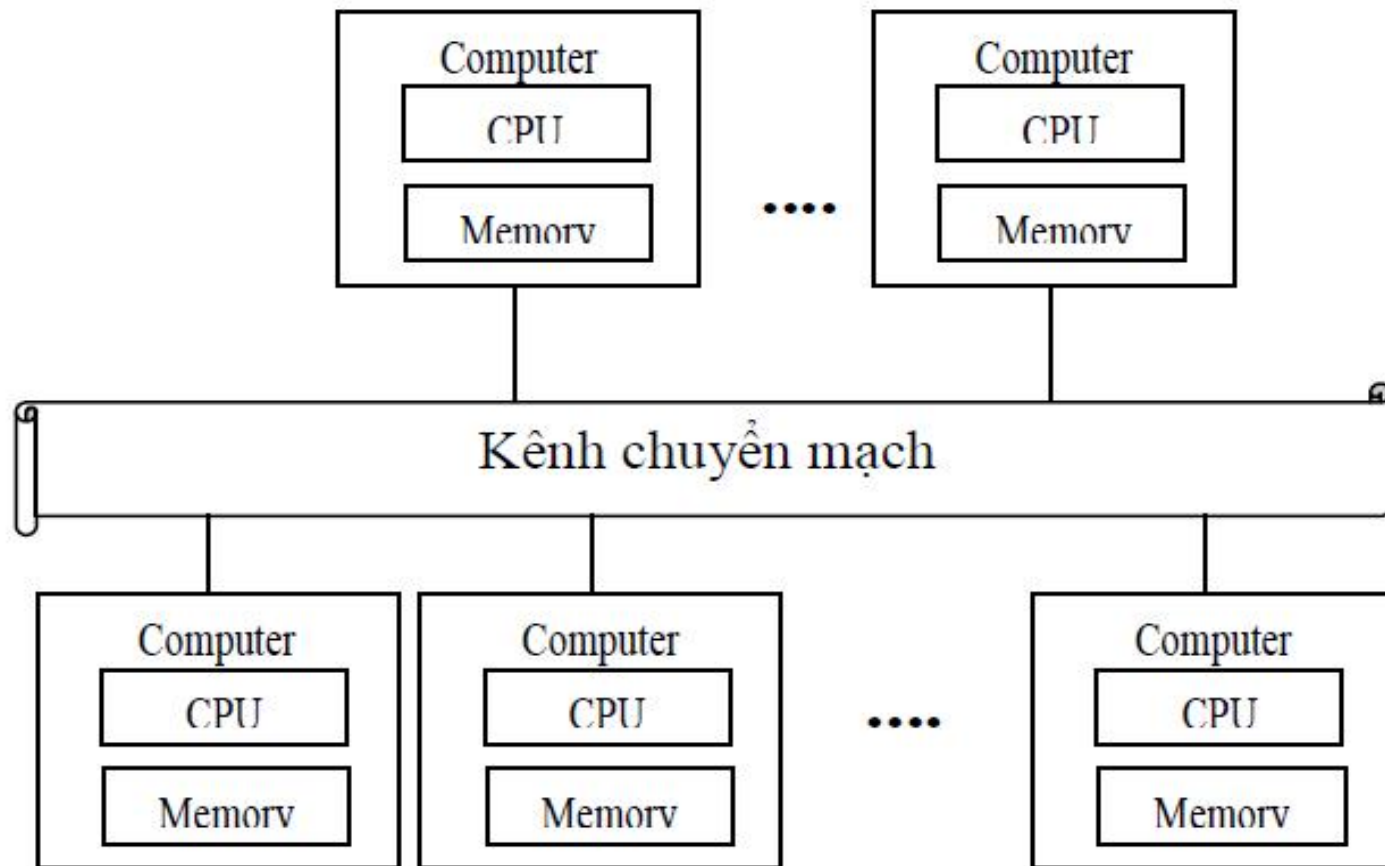
Hệ đa bộ xử lý có bộ nhớ chung

Ví dụ 2:



Hệ đa bộ xử lý có shared disk

Ví dụ 3:



Hệ đa bộ xử lý sở hữu cá nhân

2. Định nghĩa cơ sở dữ liệu phân tán

► 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 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.

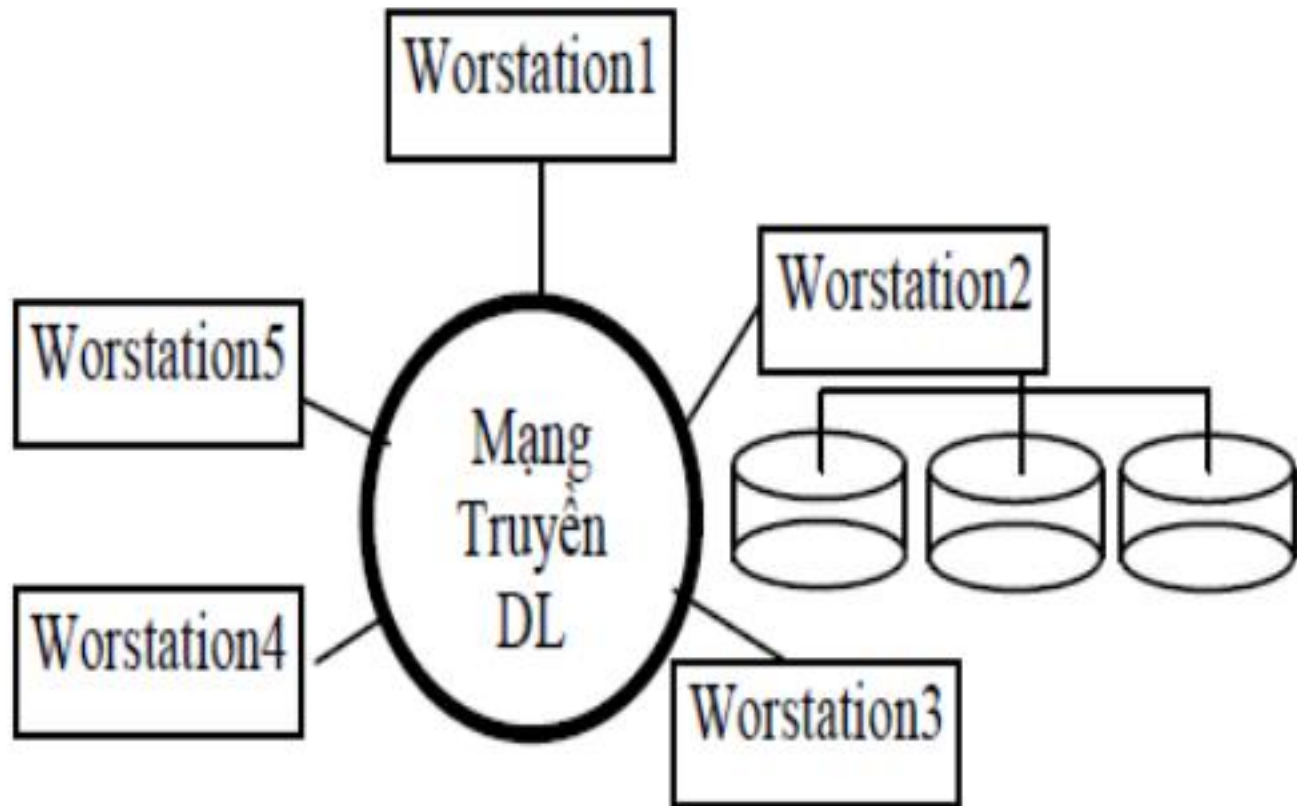
2. Định nghĩa cơ sở dữ liệu phân tán

- ▶ **Sự phân tán dữ liệu (data distribution):** dữ liệu phải được phân tán ở nhiều nơi.
- ▶ **Ứng dụng cục bộ (local application):** ứng dụng được chạy hoàn thành tại một nơi và chỉ sử dụng dữ liệu cục bộ của nơi này.
- ▶ **Ứng dụng toàn cục (hoặc ứng dụng phân tán) (global application / distributed application):** ứng dụng được chạy hoàn thành và sử dụng dữ liệu của ít nhất hai nơi.

Nhận xét

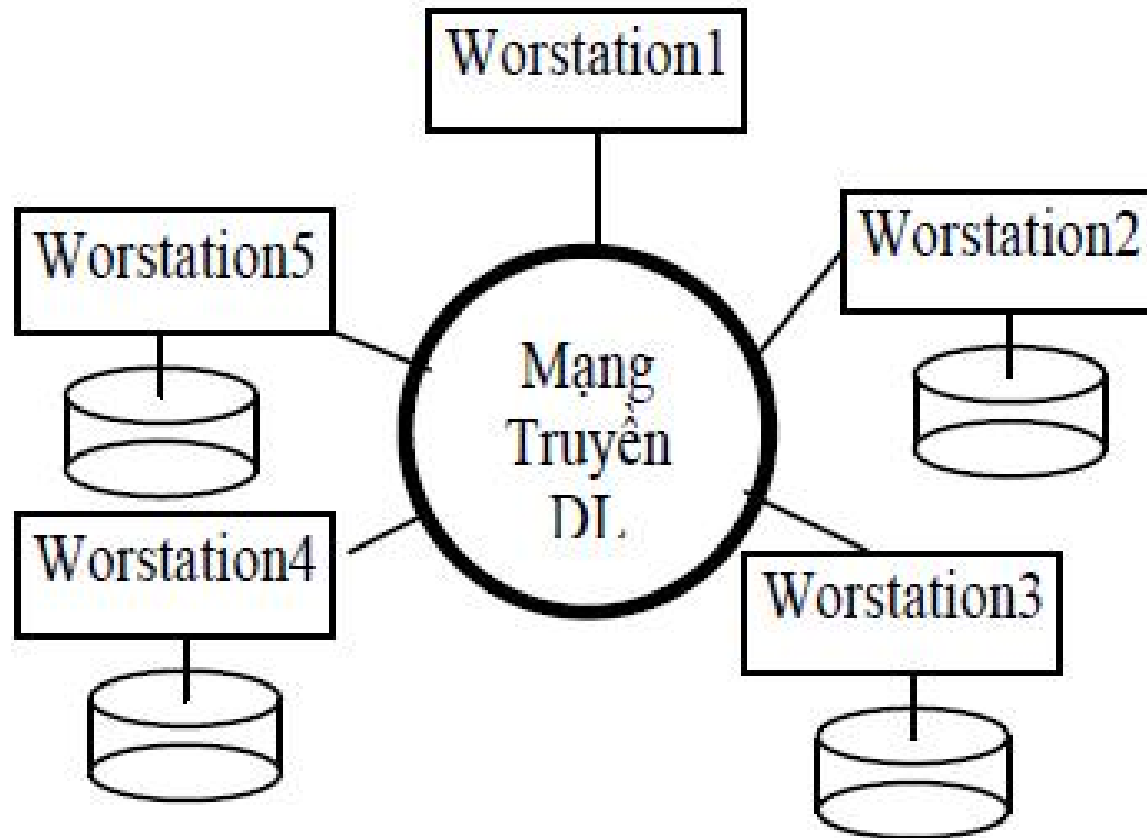
- ▶ Nếu CSDL nằm tại một nút mạng thì nó không phải là DDBS, vì vấn đề quản trị CSDL không khác với quản trị CSDL trong môi trường tập trung kiểu client/server của mạng. (Xem ví dụ 4)
- ▶ Nếu cơ sở dữ liệu được phân tán trên nhiều nút mạng, khi đó CSDL sẽ là cơ sở dữ liệu phân tán. (Xem ví dụ 5)

Ví Dụ 4:



SDL tập trung, không phải DDBS

Ví Dụ 5:



CSDBL được phân tán trên mạng, DDBS

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

► Tính phức tạp

- Các vấn đề của hệ CSDL phân tán phức tạp hơn so với các tập trung.
- Các cấu trúc vật lý phức tạp và truy xuất hiệu quả
 - Cấu trúc vật lý phức tạp để truy xuất hiệu quả.
 - Tối ưu hóa (optimization)
 - *Tối ưu hóa toàn cục (global optimization)*
 - *Tối ưu hóa cục bộ (local optimization)*

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

► Chi phí

- Các hệ phân tán đòi hỏi phải có thêm các thiết bị mới (thiết bị truyền thông chẳng hạn) và như thế làm tăng chi phí phần cứng.
- Thành phần chi phí quan trọng nhất là chi phí về nhân lực. Khi các thiết bị máy tính được xây dựng ở nhiều vị trí khác nhau, chúng đòi hỏi phải có con người điều hành và quản lý.

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

► Phân tán quyền điều khiển

- Điểm này đã được nêu ra trước đây như một ưu điểm của các hệ CSDL phân tán. Không may là sự phân tán lại gây ra các vấn đề đồng bộ hóa dữ liệu làm tăng tính phức tạp.
- Việc điều khiển phân tán có thể trở thành một gánh nặng nếu không có những chiến lược phù hợp để giải quyết chúng.

► Điều khiển tập trung

- Điều khiển tập trung (*centralized control*)
- Người quản trị CSDL cục bộ (*local DBA*)
- Người quản trị CSDL toàn cục (*global DBA*)
- Tính tự trị vị trí (*site autonomy*)

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

► Tính an ninh (bảo mật).

- Một trong những lợi ích chính của các CSDL tập trung là chúng bảo đảm kiểm soát được các truy xuất dữ liệu.
- Vấn đề an ninh trong các hệ CSDL phân tán rõ ràng là phức tạp hơn so với các hệ tập trung.
 - *Thực hiện truy xuất dữ liệu có thẩm quyền.*
 - *Bảo mật CSDL cục bộ.*
 - *Bảo mật mạng truyền thông.*

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

▶ Độc lập dữ liệu

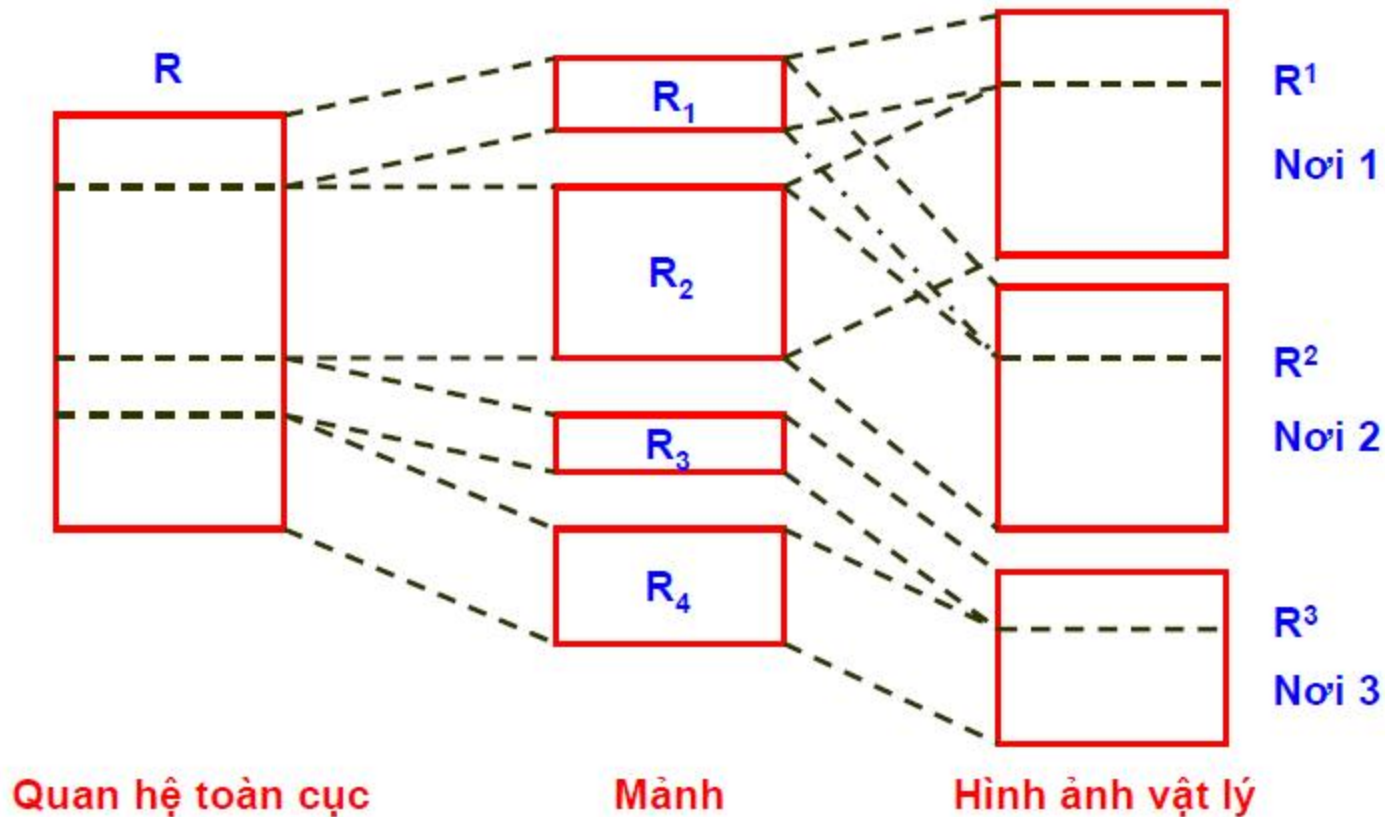
- Độc lập dữ liệu (*data independence*)
- Tính trong suốt dữ liệu (*data transparency*)
- ✓ Trong suốt phân mảnh (*fragmentation transparency*):
 - Không nhìn thấy các mảnh.
 - Nhìn thấy các quan hệ toàn cục (*global relation*).
 - Lược đồ toàn cục (*global schema*).
- ✓ Trong suốt vị trí (*location transparency*)
 - Không nhìn thấy các quan hệ cục bộ.
 - Nhìn thấy các mảnh (*fragment*).
 - Lược đồ phân mảnh (*fragmentation schema*).

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

▶ Độc lập dữ liệu (tt)

- ✓ Trong suốt nhân bản (replication transparency)
 - *Nhìn thấy các mảnh.*
 - *Không nhìn thấy sự nhân bản của các mảnh.*
- ✓ Trong suốt ánh xạ cục bộ (local mapping transparency)
 - *Nhìn thấy các quan hệ cục bộ (local relation).*
 - *Không nhìn thấy CSDL vật lý.*
- ✓ Trong suốt phân tán (distribution transparency) gồm bốn tính trong suốt trên.

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



Hình 1.4. Các mảnh và các hình ảnh vật lý của một quan hệ toàn cục.

4. Tại sao sử dụng cơ sở dữ liệu phân tán

- ▶ Các lý do về tổ chức và về kinh tế
- ▶ Nhiều tổ chức không được tập trung hóa.
 - Các CSDL hiện tại cần kết nối với nhau
- ▶ Nhiều CSDL đã tồn tại trong một công ty và cần phải thực hiện nhiều ứng dụng toàn cục hơn.
 - Sự lớn mạnh gia tăng
- ▶ Có thêm các đơn vị tổ chức tương đối độc lập.
 - Giảm chi phí truyền thông
- ▶ Nhiều ứng dụng cục bộ làm giảm chi phí truyền thông so với CSDL tập trung.

4. Tại sao sử dụng cơ sở dữ liệu phân tán

- ▶ **Các nghiên cứu về hiệu suất**
 - *Hiệu suất được nâng cao bằng một cơ chế song song hóa.*
 - *Phân mảnh dữ liệu theo ứng dụng, làm cực đại hóa tính cục bộ của ứng dụng.*
- ▶ **Độ tin cậy và tính sẵn sàng**
 - *Vì dư thừa dữ liệu, tính sẵn sàng của dữ liệu (data availability) cao.*
 - *Cần phải bảo đảm độ tin cậy của dữ liệu (data reliability).*

5. Hệ quản trị CSDL phân tán (DDBMS)

- ▶ **Hệ CSDL phân tán – DDB System (DDBS):** là một tập hợp dữ liệu có liên hệ logic và được phân bố trên các nút của một mạng máy tính.
- ▶ **Hệ quản trị CSDL phân tán - DDBS Management System (DDBMS):** một hệ thống phần mềm cho phép quản lý các DDBS và làm cho việc phân tán trở nên vô hình đối với người sử dụng.



6. Triển vọng của các hệ cơ sở dữ liệu phân tán

- ▶ Từ lý do xã hội của việc phi tập trung đến tính hiệu quả kinh tế của CSDL phân tán người ta phân DDBS thành các nhóm triển vọng sau:
 - ❑ *Quản lý dữ liệu phân tán và nhân bản vô hình.*
 - ❑ *Yêu cầu độ tin cậy qua các giao dịch phân tán*
 - ❑ *Nâng cao hiệu năng*

6. Triển vọng của các hệ cơ sở dữ liệu phân tán

❑ Quản lý dữ liệu phân tán và nhân bản vô hình.

- Một hệ thống vô hình là cách cho ẩn đi thao tác cài đặt, cài mà người sử dụng không cần quan tâm đến.
- **Ví dụ:** Một công ty Điện - Toán, có các văn phòng ở Boston , Edmonton, Paris và San Francisco, có một số dự án (project) được thực hiện tại các địa điểm đó, và muốn dùng CSDL để quản lý nhân công (Employee), quản lý dự án và các dữ liệu liên quan khác. Giả sử CSDL là CSDL quan hệ - Relational database (RDB) có thể lưu các bảng sau:

6. Triển vọng của các hệ cơ sở dữ liệu phân tán

EMP(ENO, ENAME, TITLE)

PROJ(PN_o, PNAME, BUDGET)

PAY(TITLE, SAL)

ASG (ENO, PNO, DUR, RESP)

- ▶ Nếu dữ liệu được lưu ở hệ CSDL tập trung và nếu muốn có danh sách tên và lương của các nhân viên đã làm dự án nào đó trên 12 tháng thì câu lệnh truy vấn SQL sẽ là:

SELECT ENAME, SAL

FROM EMP, ASG , PAY

WHERE ASG.DUR > 12 AND EMP.ENO = ASG.ENO

AND PAY.TITLE = EMP.TITLE

▶ **Phân mảnh CSDL (Fragmentation)**

- Vì công ty được phân tán các vị trí khác nhau (Boston, Edmonton, Paris và San Francisco), nên công ty muốn để dữ liệu về các nhân viên các dự án của vị trí nào được lưu ở vị trí đó.
- Do đó cần phải phân hoạch các quan hệ EMP và PROJ và lưu chúng tại các vị trí như đã yêu cầu. Quá trình làm này được gọi là **phân mảnh (fragmentation)**

▶ **Nhân bản (Replication)**

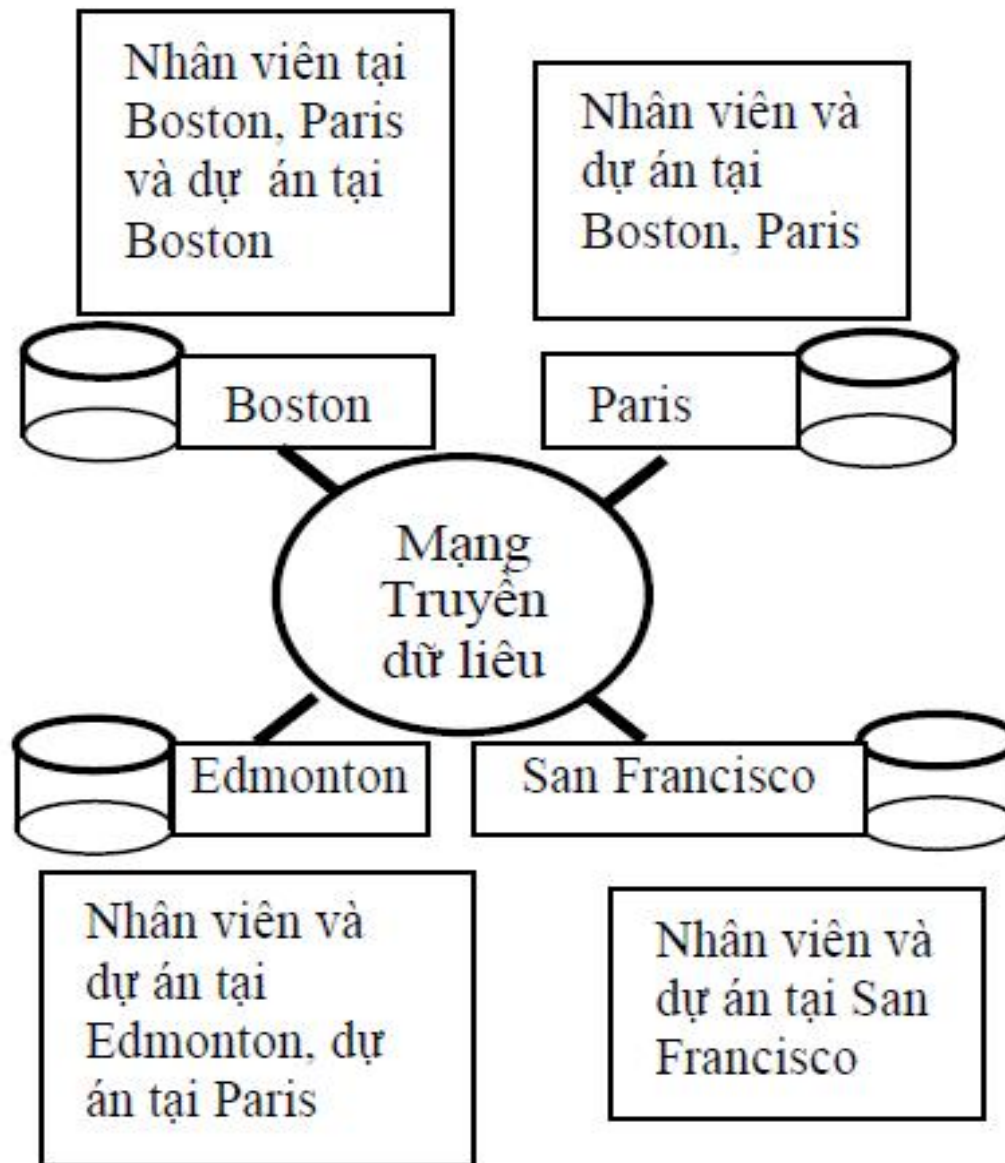
Do yêu cầu về hiệu quả và độ tin cậy cho CSDL nên một phần dữ liệu đã được phân mảnh ở trên cần phải được lưu (bản sao) tại một số vị trí khác. Quá trình này **bản.**

► Ví dụ:

Do tính chất phân tán của công việc kinh doanh của công ty, người ta muốn đặt dữ liệu về các nhân viên của văn phòng ở Edmonton được lưu ở Edmonton, nhân viên của văn phòng ở Boston được lưu ở Boston v.v...

Như vậy cần có một quá trình phân hoạch mỗi quan hệ này và lưu các phân hoạch tại các vị trí khác nhau. Kết quả phân mảnh và nhân bản trên được thể hiện trên hình sau:





▶ Truy xuất vô hình

Tuy CSDL được phân tán và phân tán trên các nút mạng, nhưng người sử dụng vẫn có thể vẫn tin như vẫn tin trên CSDL tập trung tại điểm nút thực hiện truy vấn. Việc che giấu này được gọi là **truy xuất vô hình**.

▶ Vô hình liên kết mạng

Người sử dụng được tách ra khỏi mọi chi tiết hoạt động của mạng, thậm chí là không biết có sự hiện diện của mạng nếu được - nghĩa là người sử dụng không biết là mình đang làm việc với CSDL tập trung hay phân tán. Kiểu vô hình này được gọi là **vô hình kết mạng (Network transparency)** hoặc là **vô hình phân tán (Distributed transparency)**.

Vô hình liên kết mạng có thể được chia thành hai loại; vô hình vị trí và vô hình đặt tên.

▶ **Vô hình nhân bản**

Vô hình nhân bản làm cho người sử dụng không thể biết họ đang làm việc với CSDL gốc hay với các bản nhân bản. Vậy vô hình nhân bản làm nhiệm vụ của DBMS.

▶ **Vô hình phân mảnh (Fragmentation transparency)**

- Vô hình phân mảnh làm cho người sử dụng không cần tham gia vào việc phân mảnh và không thể biết họ đang làm việc với CSDL gốc hay với các mảnh đã được phân mảnh từ CSDL gốc.
- **Phân mảnh** là chia CSDL thành các mảnh dữ liệu (Fragment) nhỏ hơn và xử lý mỗi mảnh nhận được như một CSDL độc lập - tức là như một quan hệ. Phân mảnh chỉ được thực hiện khi nó tăng hiệu quả, và có độ tin cậy. Có hai kiểu phân mảnh cơ bản là phân mảnh ngang (Horizontal phân mảnh dọc (Vertical fragmentation)).

6. Triển vọng của các hệ cơ sở dữ liệu phân tán

- ❑ **Yêu cầu độ tin cậy qua các giao dịch phân tán**
 - Một giao dịch (transaction) là một đơn vị tính toán cơ bản, nhất quán và tin cậy được, bao gồm một loạt các thao tác CSDL như các hành động nguyên tử (atomic action).
 - Nó biến đổi CSDL từ trạng thái nhất quán này sang trạng thái nhất quán khác ngay cả khi có một số lượng giao dịch được thực hiện đồng thời và ngay cả khi có sự cố xảy ra.

6. Triển vọng của các hệ cơ sở dữ liệu phân tán

❑ Nâng cao hiệu năng

Hiệu năng của DBMS phân tán sẽ được nâng cao dựa vào hai điều kiện:

- Một DBMS phân tán có khả năng phân mảnh CSDL mức khái niệm, cho phép dữ liệu ở gần nơi sử dụng (cũng được gọi là cục bộ hóa dữ liệu - data localization).
- Tính chất song hành của các hệ phân tán có thể được tận dụng để thực hiện song hành liên vấn tin và nội vấn tin. Song hành liên vấn tin là khả năng thực hiện cùng một lúc nhiều câu vấn tin còn song hành nội vấn tin là tách một câu vấn tin thành các câu vấn tin con, mỗi câu sẽ được thực hiện tại một vị trí và truy xuất các phần khác nhau của CSDL phân tán.

7. Phân mảnh dữ liệu (Fragmentation)

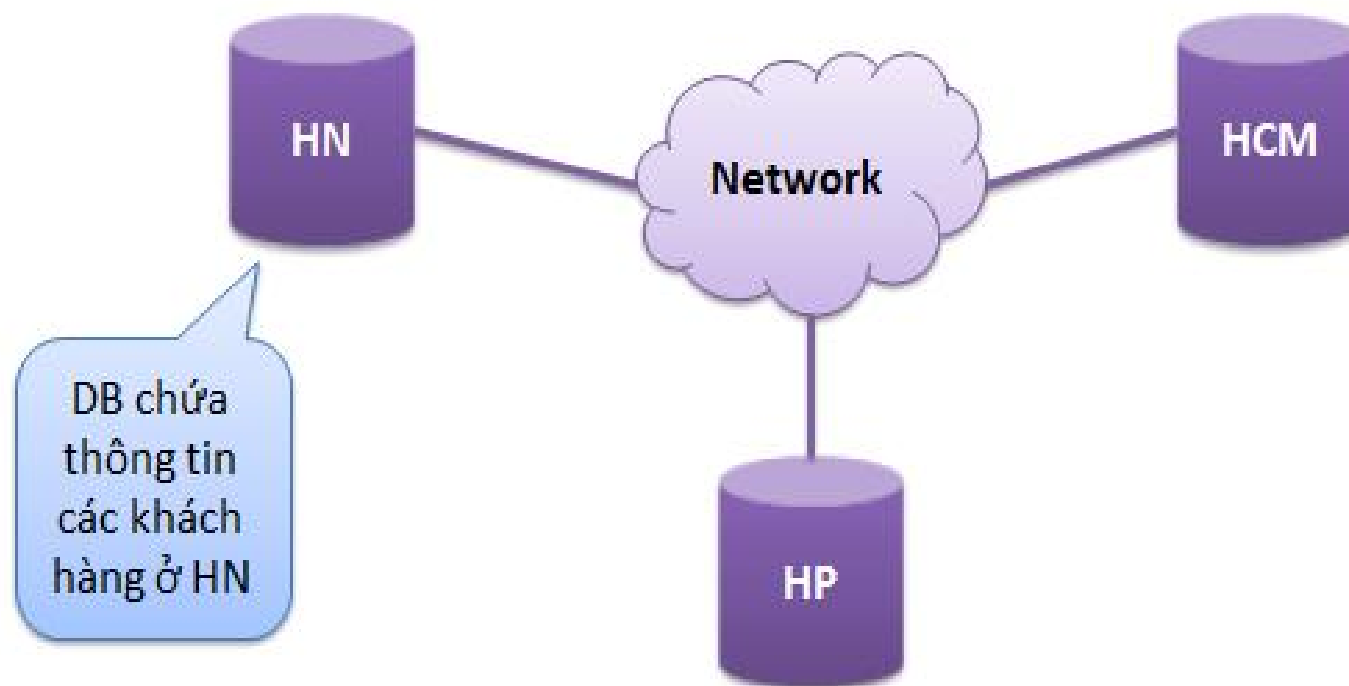
- ▶ Việc phân tán dữ liệu được thực hiện trên cơ sở cấp phát các tập tin cho các nút trên một mạng máy tính. Các nút mạng thường nằm ở các vị trí địa lý khác nhau trải rộng trên một diện tích lớn. Do vậy để tối ưu việc khai thác thông tin thì dữ liệu không thể để tập trung mà phải **phân tán trên các nút của mạng**.
- ▶ Một quan hệ không phải là một đơn vị truy xuất dữ liệu tốt nhất. (Ví dụ)
- ▶ Do vậy phân rã một quan hệ thành nhiều mảnh, mỗi mảnh được xử lý như một đơn vị sẽ cho phép thực hiện nhiều giao dịch đồng thời. Một câu truy vấn ban đầu có thể được chia ra thành một tập các truy vấn con, các truy vấn này có thể được thực hiện song song trên các mảnh sẽ giúp cải thiện tốc độ hoạt động của hệ thống.

- ▶ **Thể hiện của các quan hệ chính là các bảng**, vì thế vấn đề là tìm những cách khác nhau để chia một bảng thành nhiều bảng nhỏ hơn.
- ▶ **Có hai phương pháp khác nhau:** Chia bảng theo chiều dọc và chia bảng theo chiều ngang.
- **Chia dọc** ta được các quan hệ con mà mỗi quan hệ chứa một tập con các thuộc tính của quan hệ gốc – gọi là phân mảnh dọc.
- **Chia ngang** một quan hệ ta được các quan hệ con mà mỗi quan hệ chứa một số bộ của quan hệ gốc – gọi là phân mảnh ngang.
- Ngoài ra còn có một khả năng **hỗn hợp**, đó là phân h phân mảnh ngang và dọc.

Review

- ▶ **Cơ sở dữ liệu phân tán (Distributed Database)** là cơ sở dữ liệu được phân mảnh và được lưu trữ trên các trạm trong hệ thống mạng.
- ▶ Cơ sở dữ liệu phân tán là quan trọng trong kinh tế, tổ chức và kỹ thuật với nhiều lý do khác nhau. Chúng có thể được cài đặt trên một mạng máy tính có phạm vi rộng lớn hoặc nhỏ bé.
- ▶ Hiện nay, các DDBMSs thương mại đều tích hợp các ứng dụng phân tán nên rất tiện cho người sử dụng.

Review



Review

- ▶ **Hệ quản trị cơ sở dữ liệu phân tán (DDBMS)**
 - Cho phép người dùng tạo, sử dụng csdl.
 - Đảm bảo an ninh (cấp phát quyền, 1 nhóm người được sử dụng, ...)
 - Đảm bảo tính trong suốt của csdl (Transperence)
 - Người dùng sử dụng như csdl tập trung.
 - Truy vấn tập trung → Truy vấn phân tán.
- ▶ **Các ứng dụng:**
 - ▶ Ứng dụng cục bộ: Chỉ quan tâm tới dữ liệu ở 1 trạm.
 - ▶ Ứng dụng toàn cục: Liên quan đến nhiều trạm.

Review: Ưu nhược điểm của cơ sở dữ liệu phân tán

▶ Ưu điểm:

- Dữ liệu gần với nơi xử lý → Hiệu suất cao.
- Tính sẵn sàng của hệ thống cao: Nếu một trạm bị lỗi sẽ không ảnh hưởng tới các trạm khác trong hệ thống.
- Việc tăng các trạm sử dụng trong hệ thống là đơn giản nên việc mở rộng CSDL là dễ dàng.

▶ Nhược điểm:

- ▶ Lưu trữ: Ngoài lược đồ CSDL như trong CSDL tập trung (Thuộc tính, kiểu dữ liệu, ...) còn thêm các lược đồ phân đoạn CSDL, lược đồ định vị CSDL (cho biết các đoạn được lưu trữ ở đâu).
- ▶ Xử lý: Truy vấn tập trung là đơn giản còn truy vấn phân tán phức tạp.
- ▶ An toàn: CSDL được lưu trữ ở nhiều nơi nảy sinh vấn đề: đảm bảo an toàn dữ liệu khi truyền qua mạng.

Review: So sánh ưu nhược điểm của CSDL phân tán

• Ưu điểm

- ▶ **Chia sẻ dữ liệu và điều khiển phân tán:** Người sử dụng tại một vị trí này có thể truy xuất dữ liệu (được phép) ở vị trí khác. Hơn nữa việc quản trị cơ sở dữ liệu có thể được phân tán và thực hiện tự quản tại mỗi vị trí.
- ▶ **Độ tin cậy và tính sẵn sàng:** Nếu một vị trí bị hỏng thì các vị trí còn lại trong hệ thống cơ sở dữ liệu phân tán vẫn tiếp tục hoạt động. Nếu dữ liệu được nhân bản ở một số vị trí thì một giao dịch cần truy xuất một mục dữ liệu có thể tìm thấy ở bất kỳ vị trí nào trong số vị trí đó. Như thế sự cố tại một vị trí không ảnh hưởng đến hệ thống.
- ▶ **Tăng tốc độ xử lý truy vấn:** Nếu một truy vấn cần dữ liệu ở một số vị trí thì có thể chia câu truy vấn đó thành các câu truy vấn con và thực hiện tại các vị trí.

Review: So sánh ưu nhược điểm của CSDL phân tán

• **Nhược điểm**

- ✓ **Chi phí phát triển phần mềm:** Việc phát triển một hệ thống cơ sở dữ liệu phân tán khá phức tạp vì thế cần chi phí lớn.
- ✓ **Khó phát hiện lỗi:** Việc phát hiện lỗi và đảm bảo tính đúng đắn của các thuật toán song song sẽ rất khó khăn.
- ✓ **Chi phí xử lý tăng:** Sự trao đổi các thông báo và xử lý phối hợp giữa các vị trí sẽ tăng chi phí xử lý hơn trong các hệ thống tập trung.