MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SLOAN SCHOOL OF MANAGEMENT

15.565 Integrating Information Systems:
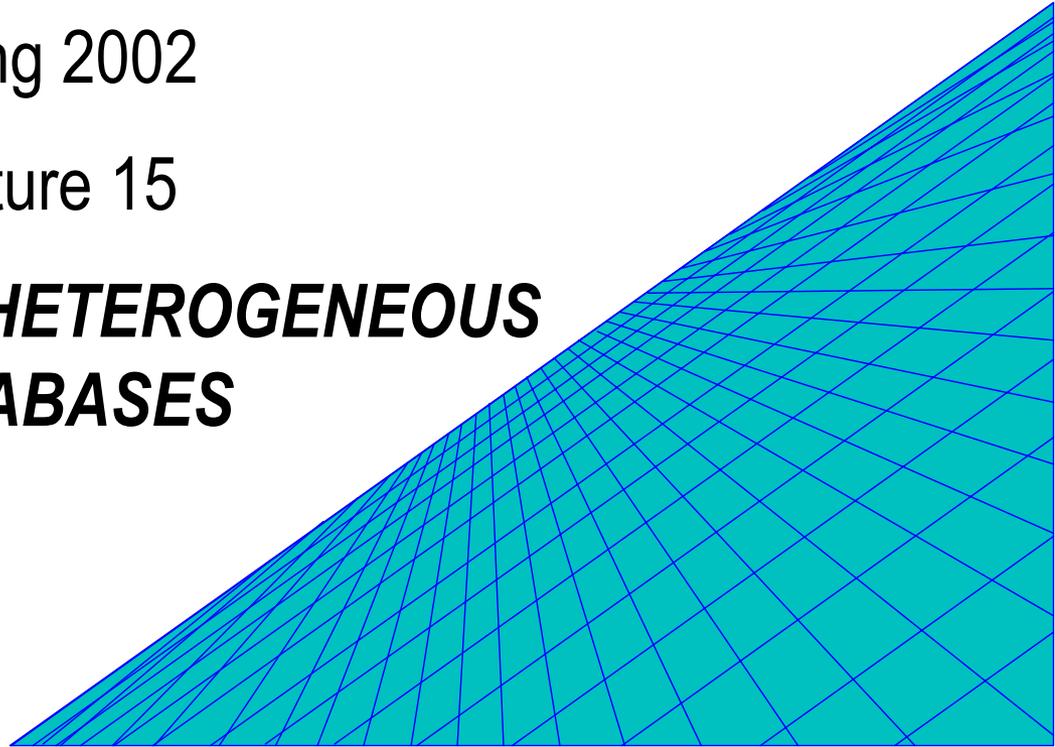**Technology, Strategy, and Organizational Factors**

15.578 Global Information Systems:
Communications & Connectivity Among Information Systems
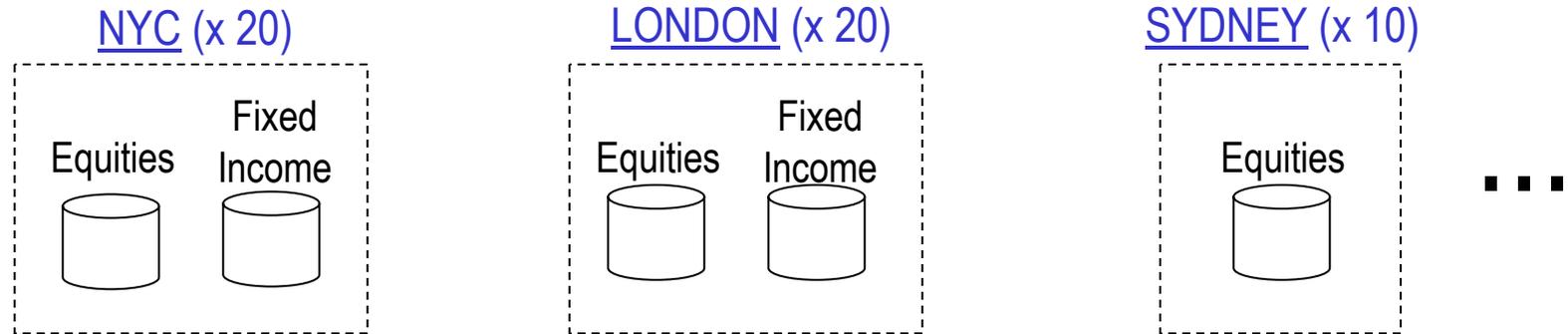
Spring 2002

Lecture 15

DISTRIBUTED HETEROGENEOUS
DATABASES

NEXT GENERATION GLOBAL FOREIGN EXCHANGE SYSTEM (FXX)

- TRADING SYSTEMS

NYC (x 20)

Equities   Fixed Income

LONDON (x 20)

Equities   Fixed Income

SYDNEY (x 10)

Equities

...

- **EXAMPLE QUESTIONS:**
  - WHAT ARE OUR TOTAL HOLDINGS OF "IBM"?
  - WHAT ARE TOTAL ASSETS OWNED BY & EXPOSURES OF "LONG-TERM CAPITAL"?
  - WHAT ARE TOTAL ASSETS & EXPOSURES ASSOCIATED WITH "BRAZIL" ECONOMY?

- **OTHER SIMILAR SITUATIONS:**
  - CUSTOMER RELATIONSHIPS
  - PROFITABILITY ANALYSIS (BY PRODUCT, DIVISION, ETC.)
  - ASSET VISIBILITY (INVENTORY, IN-TRANSIT, ETC.)
  - DATA WAREHOUSES (IN GENERAL)

- **WHY NOT ONE BIG DATABASE?**

POSSIBLE "GLOBAL DATABASE" CONFIGURATIONS

TRADING SYSTEMS GLOBAL DATABASE | TRADING SYSTEMS REGIONAL DATABASES GLOBAL DATABASE | TRADING SYSTEMS VIRTUAL GLOBAL DATABASE
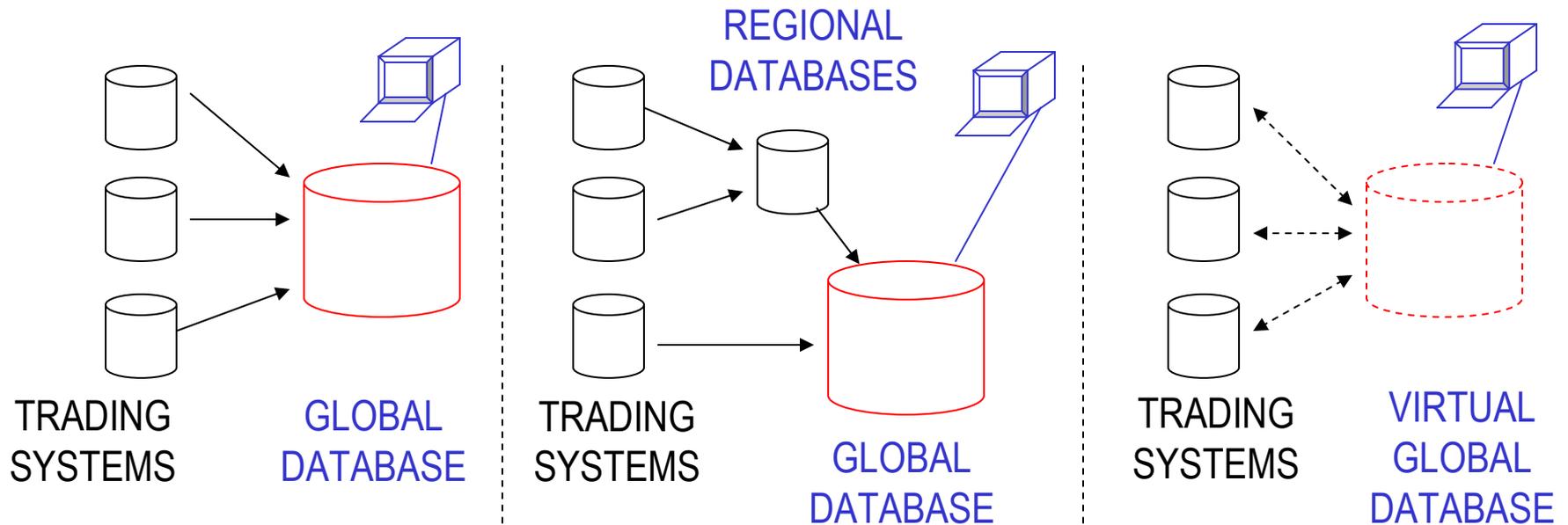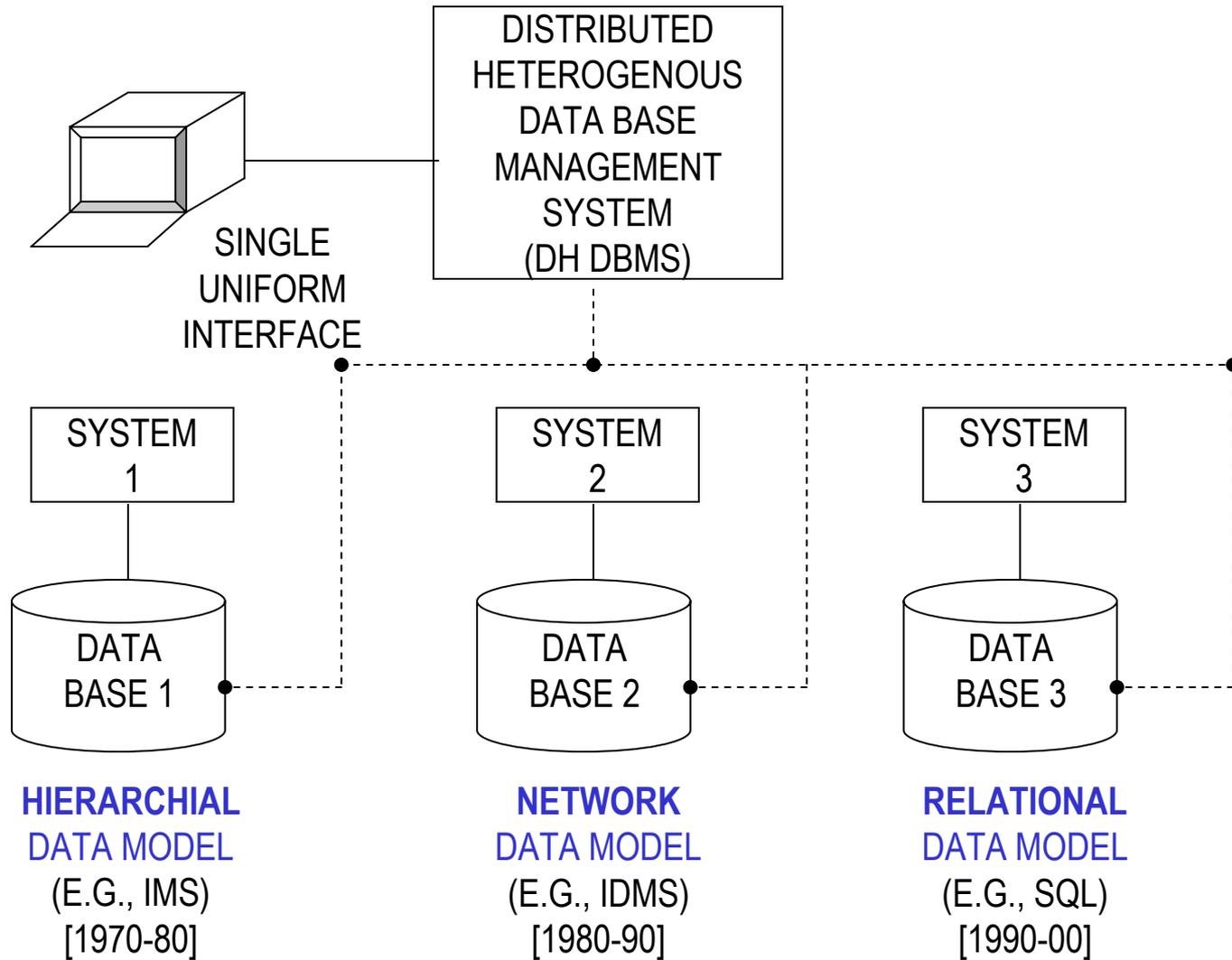
- **ADVANTAGES AND DISADVANTAGES OF EACH APPROACH**
- FURTHER COMPLICATIONS
  - THE TRADING SYSTEMS USE DIFFERING HARDWARE AND OPERATING SYSTEMS (MAINFRAMES, MINI, PC).
  - THE TRADING SYSTEMS USE DIFFERING DATABASE MANAGEMENT SYSTEM (DBMS) SOFTWARE

  $\rightarrow$ **HETEROGENEOUS ENVIRONMENT**

DISTRIBUTED HETEROGENOUS DATA BASE MANAGEMENT SYSTEMS
(VIRTUAL GLOBAL DATABASE)

DISTRIBUTED HETEROGENOUS DATA BASE MANAGEMENT SYSTEM (DH DBMS)

SINGLE UNIFORM INTERFACE

SYSTEM 1

SYSTEM 2

SYSTEM 3

DATA BASE 1

DATA BASE 2

DATA BASE 3

HIERARCHIAL
DATA MODEL
(E.G., IMS)
[1970-80]

NETWORK
DATA MODEL
(E.G., IDMS)
[1980-90]

RELATIONAL
DATA MODEL
(E.G., SQL)
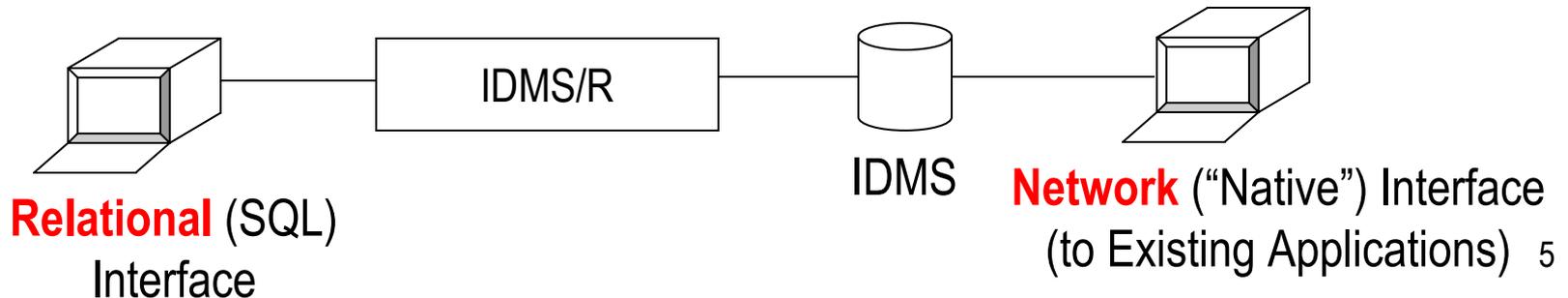[1990-00]

- "LEGACY" SYSTEMS

4

VARIOUS APPROACHES TO DISTRIBUTED DATABASES

1. **BATCH MODE** -- PERIODIC EXTRACTS, NO AD HOC QUERY

   --DOWN-LOAD, SHADOW DATABASES, SNEAKER-NET
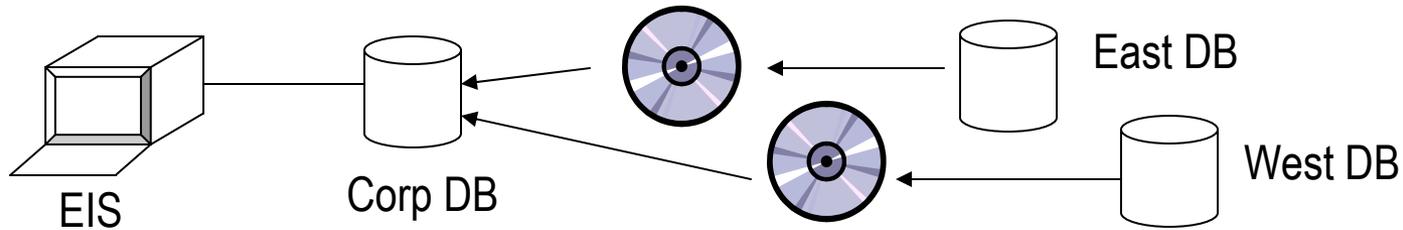
2. **HETEROGENEOUS INTERFACE** -- ONLINE ACCESS TO SINGLE DATABASE

   -- NO DATABASE INTEGRATION

   -- ACCESS USES ONE DATA MODEL; DBMS ANOTHER

   -- E.G., RELATIONAL INTERFACE TO NETWORK SYSTEM (IDMS/R)

5

VARIOUS APPROACHES TO DISTRIBUTED DATABASES (continued)

3. **INTEROPERABILITY** -- ACCESS TO MULTIPLE DATABASES; SIMILAR OR DIFFERENT DBMS'S

   -- NO LOCATION TRANSPARENCY

   -- NON-INTEGRATED SCHEMAS

   DB1
   (IMS: Inventory)

   DB2
   (SQL: Payroll)

   SQL Interface
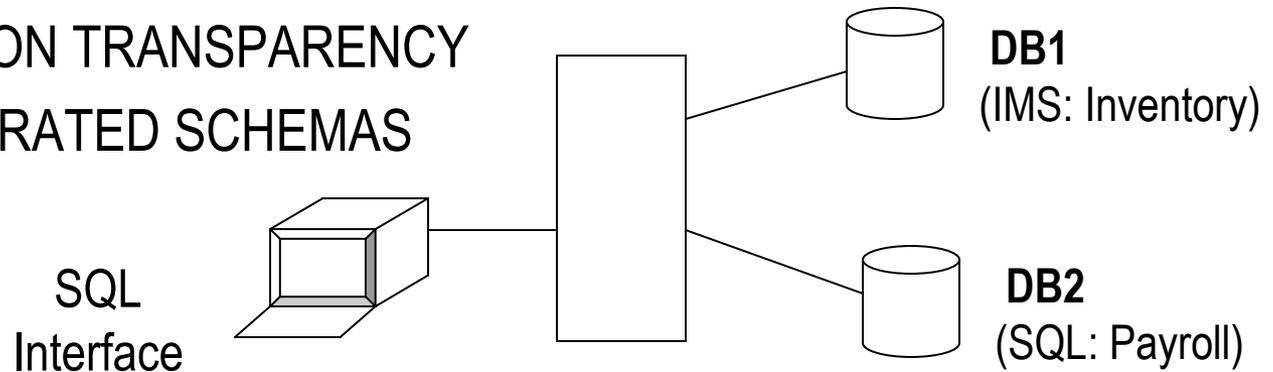
4. **GATEWAY** -- PARTIAL INTEGRATION: MOSTLY HOMOGENEOUS DBMS'S

   -- GATEWAY TO HETEROGENEOUS DBMS'S

   -- NO REAL DISTRIBUTED QUERIES

   IBM DB2

   Oracle/IBM

   Oracle/HP

   SQL Interface

   Oracle*
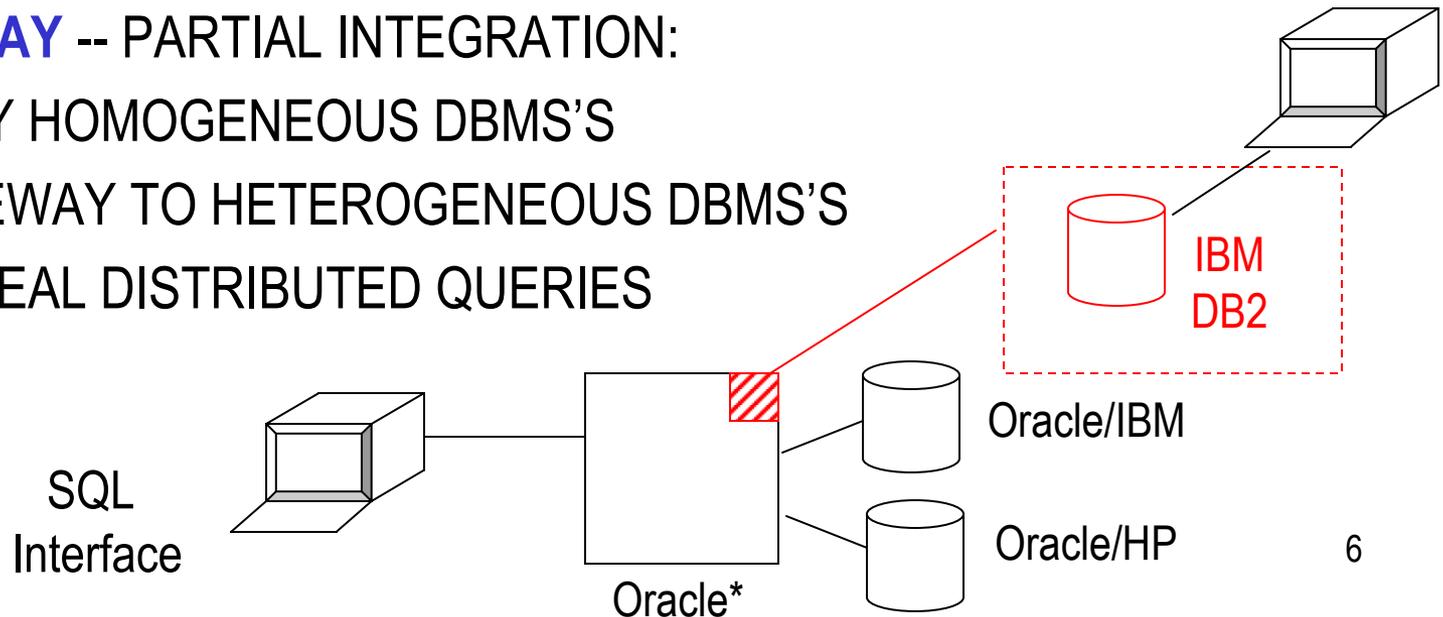
VARIOUS APPROACHES TO DISTRIBUTED DATABASES (continued)
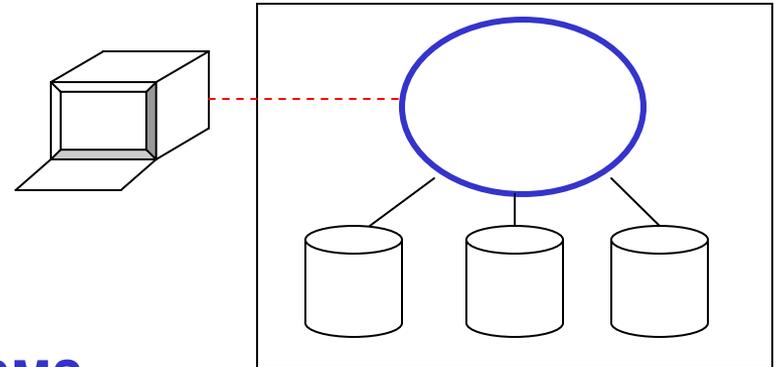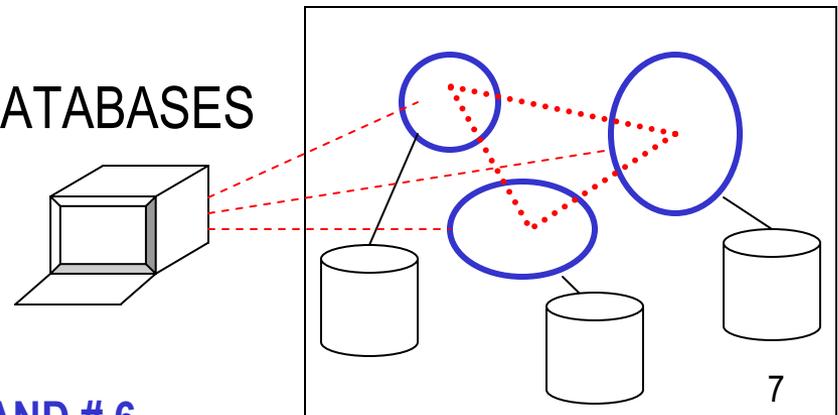
5. HETEROGENEOUS DDBMS

  -- <u>SINGLE</u> GLOBAL SCHEMA

  -- COMMON DATA MODEL AND QUERY LANGUAGE

  -- INTEGRATED SCHEMA

  -- <u>TRUE DISTRIBUTED QUERY</u>

  -- E.G., MULTIBASE

6. FEDERATED HETEROGENEOUS DDBMS

  -- <u>MULTIPLE</u> SCHEMAS

  -- AUTONOMY OF COMPONENT DATABASES

  -- <u>TRUE DISTRIBUTED QUERY</u>

TRUE HETEROGENEOUS DATABASES = # 5 AND # 6

TASKS PERFORMED BY DHDBMS

1. ACCEPT STANDARD QUERY AND UPDATE REQUESTS

2. AUTOMATICALLY DIVIDE UP QUERY FOR INDIVIDUAL DATABASES

3. TRANSFORM INTO SUBQUERIES IN APPROPRIATE LOCAL DBMS LANGUAGE

4. FORMULATE SEQUENCE OF EXECUTION AND DATA MOVEMENT (Optimization)

5. RESOLVE INCOMPATIBILITIES BETWEEN DATABASES (E.G., DOLLARS vs. POUNDS vs. YEN)

6. RESOLVE INCONSISTENCIES IN COPIES OF SAME INFORMATION (E.G., 77 MASS AVE. vs. 50 MEMORIAL DRIVE)

7. COMBINE RETRIEVED DATA INTO CONSOLIDATED RESPONSE

COMMENTS AND CONSTRAINTS ON A DHDBMS

1. COMPONENTS ARE <u>EXISTING</u> DBMS SYSTEMS, NOT DESIGNED FOR DDBMS PURPOSES

2. COMPONENTS CANNOT BE EASILY MODIFIED
   – WHY?

3. MUST BE ABLE TO ADD AND DELETE COMPONENTS TO DHDBMS

4. NOTION OF "NON-INTRUSIVE"

THREE SCHEMA APPROACH TO DHDBMS

USER VIEWS (EXTERNAL SCHEMATA)

VIEWS

PAYROLL	PERSONNEL	PROJECT

GLOBAL DATA MODEL
(CONCEPTUAL SCHEMA)

INDEPENDENT SCHEMAS

HQ   DIV 1   DIV 2

IMS   IDMS   SQL

LOCAL FILE AND DBMS SYSTEMS
(INTERNAL SCHEMATA)

10

MULTIBASE EXAMPLE

Want: "SUM OF ALL SALARIES"

(GLOBAL QUERY)

GLOBAL DATA MANAGER

(GLOBAL / COMMON LANGUAGE)

(CONSOLIDATED RESPONSE)

(SINGLE-SITE QUERY)

Network

LOCAL DATABASE INTERFACE (LDI)

NYC, USA

NETWORK INTERFACE

OPTIMIZER

(FORMATTED DATA)

DATA FORMATTER

TRANSLATOR

(RAW DATA)

HOST INTERFACE

LOCAL DBMS 1

DB 1 (IMS)

Salary ($)

TOKYO, JAPAN

NETWORK INTERFACE

OPTIMIZER

DATA FORMATTER

TRANSLATOR

HOST INTERFACE

LOCAL DBMS N

DB 2 (SQL)

Salary

11

MULTIBASE CAPABILITIES

- MULTIBASE:
    - UNIFORM, RETRIEVAL INTERFACE (DAPLEX)
    - SINGLE QUERY LANGUAGE
    - PRE-EXISTING, HETEROGENEOUS DISTRIBUTED DATABASES
- OBJECTIVES: GENERALITY, COMPATIBILITY, AND EXTENSIBILITY
- NON-INTRUSIVE: NO CHANGES NEEDED TO UNDERLYING DBMS'S, FILE SYSTEMS, OR APPLICATIONS
- CAPABILITIES:
    - SCHEMA INTEGRATION -- DEFINE GLOBAL VIEW FROM LOCAL SCHEMAS
    - DATA INCOMPATIBILITY HANDLING -- MAINTAIN AND ENFORCE RULES FOR RECONCILING DIFFERENCES
    - QUERY OPTIMIZATION -- OPTIMIZE LOCALLY AND GLOBALLY (COMPENSATING FOR LOCAL SYSTEM LIMITATIONS)
    - QUERY TRANSLATION -- TRANSLATE FROM DAPLEX TO LOCAL QUERY LANGUAGES

12

SCHEMA INCOMPATIBILITIES

- PROBLEMS (*TO BE DISCUSSED IN NEXT SESSION*)

  -- DIFFERENCES IN:

  - -- NAMING CONVENTIONS

  - -- UNDERLYING DATA STRUCTURES

  - -- REPRESENTATIONS (Male/Female, Man/Woman, M/F, 0/1)

  - -- SCALES / UNITS

  -- MISSING DATA

  -- CONFLICTING DATA VALUES

- SOLUTIONS: (MULTIBASE)

  - A VIEW MECHANISM, WITH RENAMING, LOGICAL RESTRUCTURING, AND SCALE CONVERSION, CAN SOLVE "DIFFERENCE" PROBLEMS

  - AUXILIARY DATABASE CAN BE ATTACHED TO THE UNIFIED SCHEMA TO SOLVE OTHER TWO TYPES OF PROBLEMS

QUERY PROCESSING

- PROBLEMS
    - -- USE PARALLEL PROCESSING & MINIMIZE PROCESSING AND COMMUNICATIONS COSTS
    - -- ISSUES:
        - -- LOCAL QUERY COSTS DIFFER AT DIFFERENT SITES
        - -- MANY DBMS'S DON'T SUPPORT CREATION OF TEMPORARY DATABASES BY GDM
        - -- LOCAL QUERYING CAPABILITIES MAY BE LIMITED
        - -- LOCAL DBMS'S HAVE VARYING OPTIMIZATION CAPABILITIES
- SOLUTIONS / ALTERNATIVES
    - -- **ALTERNATIVE 1: GLOBAL QUERY --> SMALLEST POSSIBLE SINGLE SITE SUBQUERIES**
        - -- COULD BE MULTIPLE SUBQUERIES AT A SITE, MERGE RESULTS AT GDM
        - -- MORE GDM WORK AND COMMUNICATION, BUT SIMPLE
    - -- **ALTERNATIVE 2: GLOBAL QUERY --> LARGEST POSSIBLE SINGLE SITE SUBQUERIES**
        - -- LESS MERGING WORK, LESS COMMUNICATION
        - -- MORE LDI SOPHISTICATION NEEDED
    - -- **ALTERNATIVE 3: GENERATE EFFICIENT PLAN, LDI'S PARTICIPATE IN GLOBAL OPTIMIZE**
        - -- SEND PARTIAL RESULTS TO GDM OR OTHER LDIS
        - -- LDI'S NEED TO SUPPORT SORTING, DUPLICATE REMOVAL, MERGING TEMPORARIES
        - -- LESS WORK BY GDM, LESS COMMUNICATION
        - -- COMPLEX GDM AND LDI'S

14

TRANSACTION MANAGEMENT (COMPLEX UPDATES)

- PROBLEMS
    - -- DIFFERENT DBMS'S -- DIFFERENT CONCURRENCY CONTROL METHODS
    - -- IMPLEMENTATION DIFFERENCES:
        - -- DEADLOCK HANDLING & LOCKING GRANULARITY
    - -- COMMIT PROTOCOLS MAY VARY
    - -- GLOBAL DEADLOCK DETECTION DIFFICULT:
        - -- LOCAL PROCESSES DON'T KNOW ABOUT NON-LOCAL TRANSACTIONS
        - -- GLOBAL PROCESSES DON'T KNOW ABOUT LOCAL TRANSACTIONS
- SOLUTIONS/ALTERNATIVES
    - -- ALLOW READS ONLY--NO WRITES. NO CONCURRENCY CONTROL PROBLEM.
    - -- ALLOW SINGLE SITE UPDATES ONLY
        - -- USE LOCAL CONCURRENCY CONTROL MECHANISMS
    - -- ASSUME ALL DBMS'S PROVIDE LOCKING AND TWO-PHASE COMMIT.
        - -- DETECT CONFLICTS BY TIME OUT.
    - -- ALLOW MULTI-SITE UPDATES.
        - -- MAY REQUIRE MODIFICATIONS TO LOCAL DBMS'S AND OS'S.

15

SUMMARY

- DISTRIBUTED HETEROGENEOUS DATA BASES ARE A REALITY IN MOST ORGANIZATIONS -- FOR BETTER OR WORSE.

- MOST APPROACHES TO HANDLE AGGREGATION AND COORDINATED ACCESS -- HAVE BEEN AD-HOC.

- DISTRIBUTED HETEROGENEOUS DATABASE MANAGEMENT SYSTEMS (DHDBMS) ARE EMERGING TO ADDRESS THESE NEEDS.