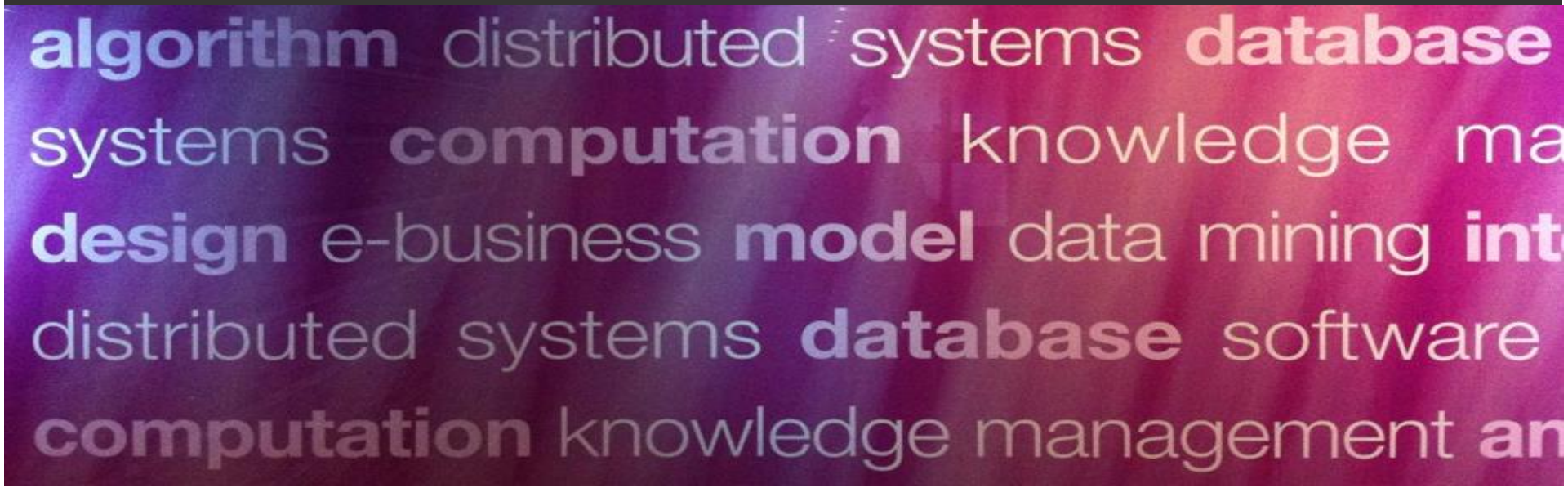
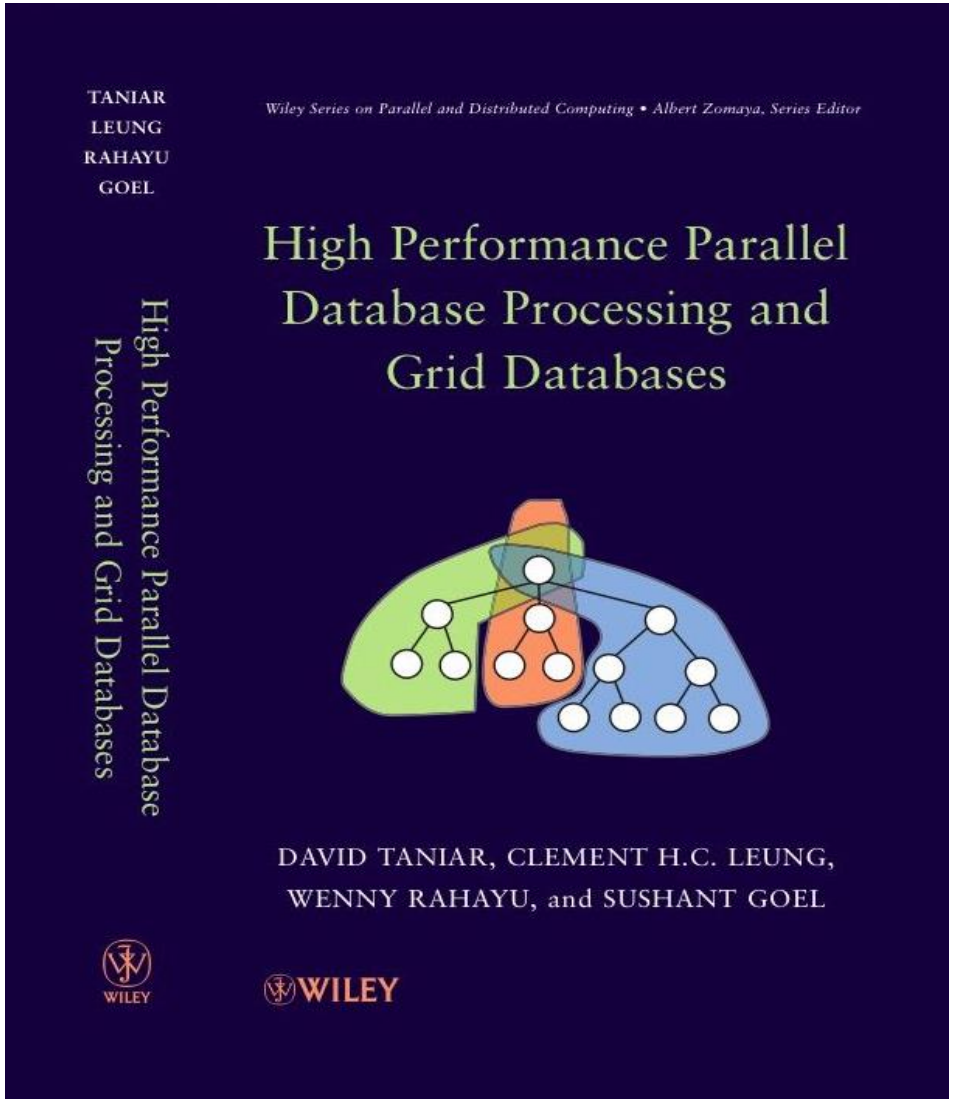
MONASH University

Information Technology

FIT5202 (Volume I)

Week 1c – Introduction to Parallel Databases





Chapter 1 Introduction

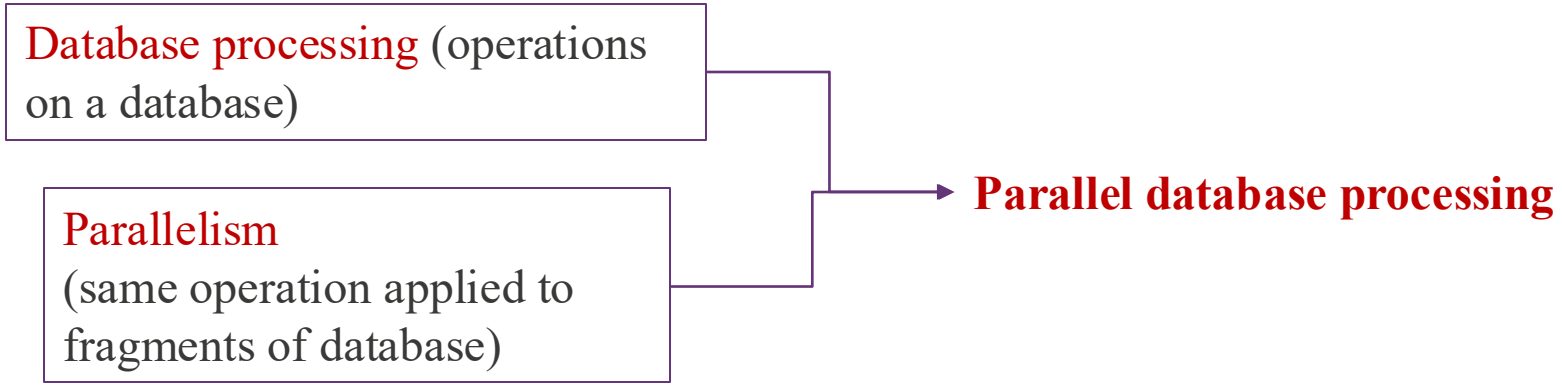
1.1/1.2. A Brief Overview, and Motivations

- An example:
  - If we have 1 petabyte of data, and the processing speed is 1GB/sec
  - How long does it take to process 1 PB of data?

1.1/1.2. A Brief Overview, and Motivations (cont'd)

- What is parallel processing, and why not just use a faster computer ?
  - Even fast computers have speed limitations
  - Limited by speed of light
  - Other hardware limitations

- Parallel processing divides a large task into smaller subtasks
- Database processing works well with parallelism (coarse-grained parallelism)



1.3. Objectives

- The primary objective of parallel database processing is to gain performance improvement
- Two main measures:
  - Throughput: the number of tasks that can be completed within a given time interval
  - Response time: the amount of time it takes to complete a single task from the time it is submitted
- Metrics:
  - **Speed up**
  - **Scale up**

1.3. Objectives (cont'd)

- **Speed up**
  - Performance improvement gained (in time) because of extra processing elements added
  - **Running a given task in less time** by increasing the degree of parallelism (i.e., process data in parallel using more processors)

- Linear speed up: performance improvement growing linearly with additional resources
- Superlinear speed up
- Sublinear speed up

$$\text{Speed up} = \frac{\text{elapsed time on uniprocessor}}{\text{elapsed time on multiprocessors}}$$

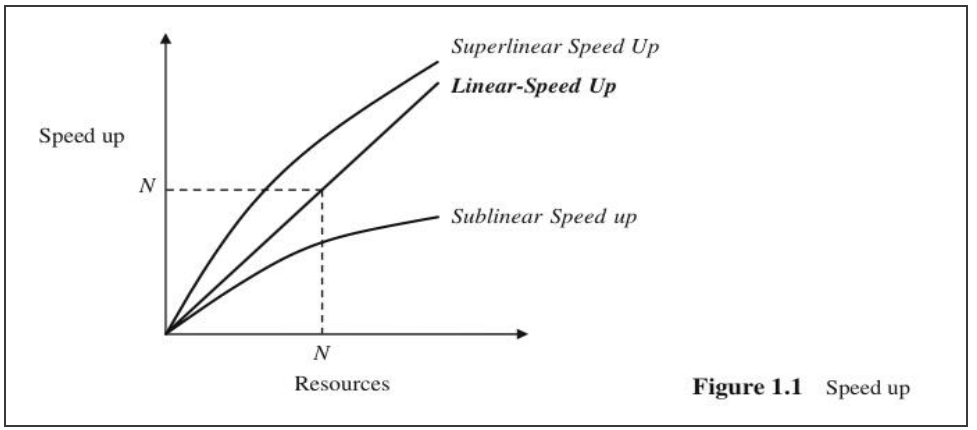


Figure 1.1 Speed up

1.3. Objectives (cont'd)

- **Scale up**
  - Handling of larger tasks by increasing the degree of parallelism
  - The **ability to process larger tasks in the same amount of time** by providing more resources.

- Linear scale up: the ability to maintain the same level of performance when both the workload and the resources are proportionally added
- Transactional scale up (increased number of tasks processed within give time)
- Data scale up (increased size of data processed within give time)

$$\text{Scale up} = \frac{\text{uniprocessor elapsed time on small system}}{\text{multiprocessor elapsed time on larger system}}$$

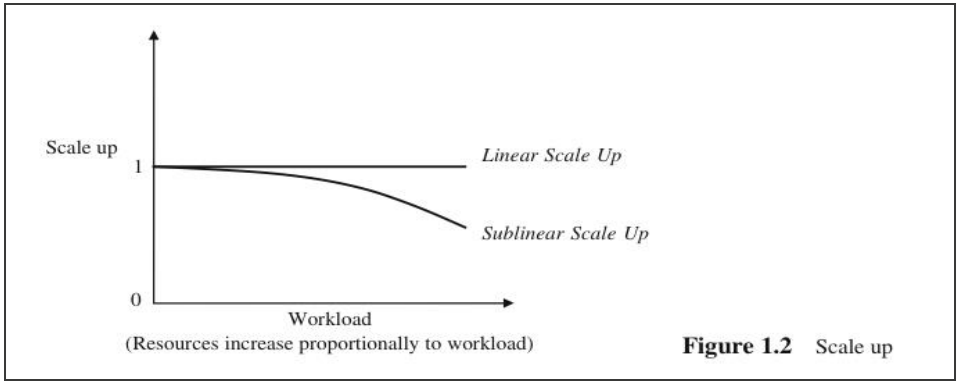


Figure 1.2 Scale up

• **Using the current processing resources, we can finish processing 1TB (one terabyte) of data in 1 hour. Recently the volume of data has increased to 2TB and the management has decided to double up the processing resources. Using the new processing resources, we can finish processing the 2TB in 60 minutes.**

Is this speed up or scale up?

Solution:

Using x resources (current resources), 1TB queries = 60 minutes
When the resources are doubled (e.g. x becomes 2x now), 2TB is completed in 60 minutes.

$$\text{Scale up} = \frac{\text{uniprocessor elapsed time on small system}}{\text{multiprocessor elapsed time on larger system}}$$

Scale up = 60 mins / 60 mins = 1
i.e. Linear Scale up.