Principles of Parallel Algorithm Design
Ananth Grama, Anshul Gupta, George Karypis, and Vipin Kumar

Chapter Overview: Algorithms and Concurrency

• Introduction to Parallel Algorithms
  – Tasks and Decomposition
  – Processes and Mapping
  – Processes Versus Processors

• Decomposition Techniques
  – Recursive Decomposition
  – Data Decomposition
  – Exploratory Decomposition
  – Hybrid Decomposition

• Mapping Techniques for Load Balancing
  – Static and Dynamic Mapping
Preliminaries: Decomposition, Tasks, and Dependency Graphs

- The first step in developing a parallel algorithm is to **decompose the problem into tasks** that can be executed concurrently.
- A given problem may be decomposed into tasks in **many different ways**.
- Tasks may be of same, different, or even interminate sizes.
- A decomposition can be illustrated in the form of a **directed graph** with nodes corresponding to tasks and edges indicating that the result of one task is required for processing the next. Such a graph is called a **task dependency graph**.
Example: Multiplying a Dense Matrix with a Vector

Computation of each element of output vector \( y \) is independent of other elements. Based on this, a dense matrix-vector product can be decomposed into \( n \) tasks. The figure highlights the portion of the matrix and vector accessed by Task 1.

Observations: While tasks share data (namely, the vector \( b \)), they do not have any control dependencies - i.e., no task needs to wait for the (partial) completion of any other. All tasks are of the same size in terms of number of operations. Is this the maximum number of tasks we could decompose this problem into?
Example: Database Query Processing

Consider the execution of the query:

```
MODEL = "CIVIC" AND YEAR = 2001 AND 
(COLOR = "GREEN" OR COLOR = "WHITE")
```

on the following database:

<table>
<thead>
<tr>
<th>ID#</th>
<th>Model</th>
<th>Year</th>
<th>Color</th>
<th>Dealer</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>4523</td>
<td>Civic</td>
<td>2002</td>
<td>Blue</td>
<td>MN</td>
<td>$18,000</td>
</tr>
<tr>
<td>3476</td>
<td>Corolla</td>
<td>1999</td>
<td>White</td>
<td>IL</td>
<td>$15,000</td>
</tr>
<tr>
<td>7623</td>
<td>Camry</td>
<td>2001</td>
<td>Green</td>
<td>NY</td>
<td>$21,000</td>
</tr>
<tr>
<td>9834</td>
<td>Prius</td>
<td>2001</td>
<td>Green</td>
<td>CA</td>
<td>$18,000</td>
</tr>
<tr>
<td>6734</td>
<td>Civic</td>
<td>2001</td>
<td>White</td>
<td>OR</td>
<td>$17,000</td>
</tr>
<tr>
<td>5342</td>
<td>Altima</td>
<td>2001</td>
<td>Green</td>
<td>FL</td>
<td>$19,000</td>
</tr>
<tr>
<td>3845</td>
<td>Maxima</td>
<td>2001</td>
<td>Blue</td>
<td>NY</td>
<td>$22,000</td>
</tr>
<tr>
<td>8354</td>
<td>Accord</td>
<td>2000</td>
<td>Green</td>
<td>VT</td>
<td>$18,000</td>
</tr>
<tr>
<td>4395</td>
<td>Civic</td>
<td>2001</td>
<td>Red</td>
<td>CA</td>
<td>$17,000</td>
</tr>
<tr>
<td>7352</td>
<td>Civic</td>
<td>2002</td>
<td>Red</td>
<td>WA</td>
<td>$18,000</td>
</tr>
</tbody>
</table>
Example: Database Query Processing

The execution of the query can be divided into subtasks in various ways. Each task can be thought of as generating an intermediate table of entries that satisfy a particular clause.

Decomposing the given query into a number of tasks. Edges in this graph denote that the output of one task is needed to accomplish the next.
Example: Database Query Processing

Note that the same problem can be decomposed into subtasks in other ways as well.

An alternate decomposition of the given problem into subtasks, along with their data dependencies. Different task decompositions may lead to significant differences with respect to their eventual parallel performance.
Granularity of Task Decompositions

- The number of tasks into which a problem is decomposed determines its **granularity**.
- Decomposition into a large number of tasks results in **fine-grained decomposition** and that into a small number of tasks results in a **coarse grained decomposition**.

![Diagram of task decomposition](image)

A **coarse grained counterpart** to the dense matrix-vector product example. Each task in this example corresponds to the computation of three elements of the result vector.
Degree of Concurrency

- The **number of tasks that can be executed in parallel** is the *degree of concurrency* of a decomposition.
- Since the number of tasks that can be executed in parallel may **change** over program execution, the *maximum degree of concurrency* is the maximum number of such tasks at any point during execution. **What is the maximum degree of concurrency of the database query examples?**
- The **average degree of concurrency** is the average number of tasks that can be processed in parallel over the execution of the program. **Assuming that each tasks in the database example takes identical processing time, what is the average degree of concurrency in each decomposition?**
- The degree of concurrency increases as the decomposition becomes finer in granularity and vice versa.
Critical Path Length

• A directed path in the task dependency graph represents a sequence of tasks that must be processed one after the other.
• The **longest such path determines the shortest time** in which the program can be executed in parallel.
• The length of the longest path in a task dependency graph is called the **critical path length**.
Consider the task dependency graphs of the two database query decompositions:

What are the critical path lengths for the two task dependency graphs? If each task takes 10 time units, what is the shortest parallel execution time for each decomposition? How many processors are needed in each case to achieve this minimum parallel execution time? What is the maximum degree of concurrency?
Limits on Parallel Performance

• It would appear that the parallel time can be made arbitrarily small by making the decomposition finer in granularity.

• There is an inherent bound on how fine the granularity of a computation can be. For example, in the case of multiplying a dense matrix with a vector, there can be no more than $\Theta(n^2)$ concurrent tasks.

• Concurrent tasks may also have to exchange data with other tasks. This results in communication overhead. The tradeoff between the granularity of a decomposition and associated overheads often determines performance bounds.
Task Interaction Graphs

• **Subtasks generally exchange data** with others in a decomposition. For example, even in the trivial decomposition of the dense matrix-vector product, if the vector is not replicated across all tasks, they will have to **communicate elements of the vector**.

• The graph of tasks (nodes) and their interactions/data exchange (edges) is referred to as a **task interaction graph**.

• Note that **task interaction graphs** represent **data dependencies**, whereas **task dependency graphs** represent control dependencies.
Task Interaction Graphs: An Example

Consider the problem of multiplying a sparse matrix $A$ with a vector $b$. The following observations can be made:

- As before, the computation of each element of the result vector can be viewed as an independent task.
- Unlike a dense matrix-vector product though, only non-zero elements of matrix $A$ participate in the computation.
- If, for memory optimality, we also partition $b$ across tasks, then one can see that the task interaction graph of the computation is identical to the graph of the matrix $A$ (the graph for which $A$ represents the adjacency structure).
Decomposition Techniques

So how does one decompose a task into various subtasks?

While there is no single recipe that works for all problems, we present a set of commonly used techniques that apply to broad classes of problems. These include:

- recursive decomposition
- data decomposition
- exploratory decomposition
- speculative decomposition
Recursive Decomposition

- Generally suited to problems that are solved using the **divide-and-conquer strategy**.

- A given problem is first decomposed into a set of sub-problems.

- These sub-problems are recursively decomposed further **until a desired granularity is reached**.
Recursive Decomposition: Example

A classic example of a divide-and-conquer algorithm on which we can apply recursive decomposition is Quicksort.

In this example, once the list has been partitioned around the pivot, each sublist can be processed concurrently (i.e., each sublist represents an independent subtask). This can be repeated recursively.
Recursive Decomposition: Example

The problem of finding the **minimum number** in a given list (or indeed any other associative operation such as sum, AND, etc.) can be fashioned as a divide-and-conquer algorithm. The following algorithm illustrates this.

We first start with a **simple serial loop** for computing the minimum entry in a given list:

1. **procedure** SERIAL_MIN \((A, n)\)
2. **begin**
3. \(min = A[0];\)
4. **for** \(i := 1 \text{ to } n - 1\) **do**
5. \[\text{if } (A[i] < min) \text{ then } min := A[i];\]
6. **endfor**;
7. **return** \(min;\)
8. **end** SERIAL_MIN
We can rewrite the loop as follows:

1. **procedure** RECURSIVE_MIN \((A, n)\)
2. \begin{align*}
    3. & \quad \text{if (} n = 1 \text{) then} \\
    4. & \quad \quad min := A[0] ; \\
    5. & \quad \text{else} \\
    6. & \quad \quad lmin := \text{RECURSIVE}_MIN \left( A, \frac{n}{2} \right) ; \\
    7. & \quad \quad rmin := \text{RECURSIVE}_MIN \left( A[\frac{n}{2}], n - \frac{n}{2} \right) ; \\
    8. & \quad \quad \text{if (} lmin < rmin \text{) then} \\
    9. & \quad \quad \quad min := lmin ; \\
    10. & \quad \quad \text{else} \\
    11. & \quad \quad \quad min := rmin ; \\
    12. & \quad \quad \text{endelse;} \\
    13. & \quad \text{endelse;} \\
    14. & \quad \text{return } min; \\
    15. & \quad \text{end RECURSIVE}_MIN
\end{align*}
Recursive Decomposition: Example

The code in the previous foil can be decomposed naturally using a recursive decomposition strategy. We illustrate this with the following example of finding the minimum number in the set \{4, 9, 1, 7, 8, 11, 2, 12\}. The task dependency graph associated with this computation is as follows:
Data Decomposition

• Identify the **data on which computations are performed**.
• Partition this data across various tasks.
• This partitioning induces a decomposition of the problem.
• Data **can be partitioned in various ways** - this critically impacts performance of a parallel algorithm.
Data Decomposition: Output Data Decomposition

- Often, each element of the **output can be computed independently** of others (but simply as a function of the input).
- A partition of the output across tasks decomposes the problem naturally.
Consider the problem of **multiplying** two $n \times n$ matrices $A$ and $B$ to yield matrix $C$. The output matrix $C$ can be partitioned into four tasks as follows:

\[
\begin{pmatrix}
A_{1,1} & A_{1,2} \\
A_{2,1} & A_{2,2}
\end{pmatrix}
\cdot
\begin{pmatrix}
B_{1,1} & B_{1,2} \\
B_{2,1} & B_{2,2}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
C_{1,1} & C_{1,2} \\
C_{2,1} & C_{2,2}
\end{pmatrix}
\]

Task 1: \( C_{1,1} = A_{1,1}B_{1,1} + A_{1,2}B_{2,1} \)

Task 2: \( C_{1,2} = A_{1,1}B_{1,2} + A_{1,2}B_{2,2} \)

Task 3: \( C_{2,1} = A_{2,1}B_{1,1} + A_{2,2}B_{2,1} \)

Task 4: \( C_{2,2} = A_{2,1}B_{1,2} + A_{2,2}B_{2,2} \)
Output Data Decomposition: Example

A partitioning of output data does not result in a unique decomposition into tasks. For example, for the same problem as in previous foil, with identical output data distribution, we can derive the following two (other) decompositions:

<table>
<thead>
<tr>
<th>Decomposition I</th>
<th>Decomposition II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1: $C_{1,1} = A_{1,1} B_{1,1}$</td>
<td>Task 1: $C_{1,1} = A_{1,1} B_{1,1}$</td>
</tr>
<tr>
<td>Task 2: $C_{1,1} = C_{1,1} + A_{1,2} B_{2,1}$</td>
<td>Task 2: $C_{1,1} = C_{1,1} + A_{1,2} B_{2,1}$</td>
</tr>
<tr>
<td>Task 3: $C_{1,2} = A_{1,1} B_{1,2}$</td>
<td>Task 3: $C_{1,2} = A_{1,2} B_{2,2}$</td>
</tr>
<tr>
<td>Task 4: $C_{1,2} = C_{1,2} + A_{1,2} B_{2,2}$</td>
<td>Task 4: $C_{1,2} = C_{1,2} + A_{1,1} B_{1,2}$</td>
</tr>
<tr>
<td>Task 5: $C_{2,1} = A_{2,1} B_{1,1}$</td>
<td>Task 5: $C_{2,1} = A_{2,2} B_{2,1}$</td>
</tr>
<tr>
<td>Task 6: $C_{2,1} = C_{2,1} + A_{2,2} B_{2,1}$</td>
<td>Task 6: $C_{2,1} = C_{2,1} + A_{2,1} B_{1,1}$</td>
</tr>
<tr>
<td>Task 7: $C_{2,2} = A_{2,1} B_{1,2}$</td>
<td>Task 7: $C_{2,2} = A_{2,1} B_{1,2}$</td>
</tr>
<tr>
<td>Task 8: $C_{2,2} = C_{2,2} + A_{2,2} B_{2,2}$</td>
<td>Task 8: $C_{2,2} = C_{2,2} + A_{2,2} B_{2,2}$</td>
</tr>
</tbody>
</table>
Consider the problem of counting the instances of given itemsets in a database of transactions. In this case, the output (itemset frequencies) can be partitioned across tasks.
Output Data Decomposition: Example

From the previous example, the following observations can be made:

- If the **database of transactions is replicated across the processes**, each task can be independently accomplished with no communication.
- If the **database is partitioned across processes** as well (for reasons of memory utilization), each task first computes partial counts. These **counts are then aggregated** at the appropriate task.
Input Data Partitioning

• Generally applicable if each output can be naturally computed as a function of the input.

• In many cases, this is the only natural decomposition because the output is not clearly known a-priori (e.g., the problem of finding the minimum in a list, sorting a given list, etc.).

• A task is associated with each input data partition. The task performs as much of the computation with its part of the data. Subsequent processing combines these partial results.
Input Data Partitioning: Example

In the database counting example, the **input (i.e., the transaction set) can be partitioned**. This induces a task decomposition in which each task **generates partial counts** for all itemsets. These are combined subsequently for **aggregate counts**.

<table>
<thead>
<tr>
<th>Database Transactions</th>
<th>Items</th>
<th>Itemset Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, C, E, G, H</td>
<td>A, B, C</td>
<td>1</td>
</tr>
<tr>
<td>B, D, E, F, K, L</td>
<td>D, E</td>
<td>2</td>
</tr>
<tr>
<td>A, B, F, H, L</td>
<td>C, F, G</td>
<td>0</td>
</tr>
<tr>
<td>D, E, F, H</td>
<td>A, E</td>
<td>1</td>
</tr>
<tr>
<td>F, G, H, K</td>
<td>C, D</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D, K</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>B, C, F</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C, D, K</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database Transactions</th>
<th>Items</th>
<th>Itemset Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, C</td>
<td>A, B, C</td>
<td>0</td>
</tr>
<tr>
<td>D, E</td>
<td>D, E</td>
<td>1</td>
</tr>
<tr>
<td>C, F, G</td>
<td>C, F, G</td>
<td>0</td>
</tr>
<tr>
<td>A, E</td>
<td>A, E</td>
<td>1</td>
</tr>
<tr>
<td>C, D</td>
<td>C, D</td>
<td>1</td>
</tr>
<tr>
<td>G, H, L</td>
<td>G, H, L</td>
<td>1</td>
</tr>
<tr>
<td>D, E, F, K, L</td>
<td>D, E, F, K, L</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C, D, K</td>
<td>0</td>
</tr>
</tbody>
</table>
Often input and output data decomposition can be combined for a higher degree of concurrency. For the itemset counting example, the transaction set (input) and itemset counts (output) can both be decomposed as follows:
Intermediate Data Partitioning

- Computation can often be viewed as a sequence of transformation from the input to the output data.
- In these cases, it is often beneficial to use one of the intermediate stages as a basis for decomposition.
Let us revisit the example of **dense matrix multiplication**. We first show how we can visualize this computation in terms of **intermediate matrices** $D$.
Intermediate Data Partitioning: Example

A decomposition of intermediate data structure leads to the following decomposition into 8 + 4 tasks:

Stage I

\[
\begin{pmatrix}
A_{1,1} & A_{1,2} \\
A_{2,1} & A_{2,2}
\end{pmatrix}
\cdot
\begin{pmatrix}
B_{1,1} & B_{1,2} \\
B_{2,1} & B_{2,2}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
D_{1,1,1} & D_{1,1,2} \\
D_{1,2,1} & D_{1,2,2} \\
D_{2,1,1} & D_{2,1,2} \\
D_{2,2,1} & D_{2,2,2}
\end{pmatrix}
\]

Stage II

\[
\begin{pmatrix}
D_{1,1,1} & D_{1,1,2} \\
D_{1,2,1} & D_{1,2,2}
\end{pmatrix}
\quad +
\quad \begin{pmatrix}
D_{2,1,1} & D_{2,1,2} \\
D_{2,2,1} & D_{2,2,2}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
C_{1,1} & C_{1,2} \\
C_{2,1} & C_{2,2}
\end{pmatrix}
\]

Task 01: \( D_{1,1,1} = A_{1,1} B_{1,1} \)

Task 03: \( D_{1,1,2} = A_{1,1} B_{1,2} \)

Task 05: \( D_{1,2,1} = A_{2,1} B_{1,1} \)

Task 07: \( D_{1,2,2} = A_{2,1} B_{1,2} \)

Task 09: \( C_{1,1} = D_{1,1,1} + D_{2,1,1} \)

Task 11: \( C_{2,1} = D_{1,2,1} + D_{2,2,1} \)

Task 02: \( D_{2,1,1} = A_{1,2} B_{2,1} \)

Task 04: \( D_{2,1,2} = A_{1,2} B_{2,2} \)

Task 06: \( D_{2,2,1} = A_{2,2} B_{2,1} \)

Task 08: \( D_{2,2,2} = A_{2,2} B_{2,2} \)

Task 10: \( C_{1,2} = D_{1,1,2} + D_{2,1,2} \)

Task 12: \( C_{2,2} = D_{1,2,2} + D_{2,2,2} \)
Intermediate Data Partitioning: Example

The task dependency graph for the decomposition (shown in previous foil) into 12 tasks is as follows:
Exploratory Decomposition

- In many cases, the decomposition of the problem goes hand-in-hand with its execution.
- These problems typically involve the exploration (search) of a state space of solutions.
- Problems in this class include a variety of discrete optimization problems (0/1 integer programming, QAP, etc.), theorem proving, game playing, etc.
A simple application of exploratory decomposition is in the solution to a 15 puzzle (a tile puzzle). We show a sequence of three moves that transform a given initial state (a) to desired final state (d).

Of course, the problem of computing the solution, in general, is much more difficult than in this simple example.
Exploratory Decomposition: Example

The state space can be explored by generating various successor states of the current state and to view them as independent tasks.
Exploratory Decomposition: Anomalous Computations

- In many instances of exploratory decomposition, the decomposition technique may change the amount of work done by the parallel formulation.
- This change results in super- or sub-linear speedups.

![Diagram of exploratory decomposition](image)
Speculative Decomposition

- In some applications, dependencies between tasks are not known a-priori.
- For such applications, it is impossible to identify independent tasks.
- There are generally two approaches to dealing with such applications: conservative approaches, which identify independent tasks only when they are guaranteed to not have dependencies, and, optimistic approaches, which schedule tasks even when they may potentially be erroneous.
- Conservative approaches may yield little concurrency and optimistic approaches may require roll-back mechanism in the case of an error.
Speculative Decomposition: Example

Another example is the simulation of a network of nodes (for instance, an assembly line or a computer network through which packets pass). The task is to simulate the behavior of this network for various inputs and node delay parameters (note that networks may become unstable for certain values of service rates, queue sizes, etc.).
Hybrid Decompositions

Often, a **mix of decomposition techniques** is necessary for decomposing a problem. Consider the following examples:

- In quicksort, recursive decomposition alone limits concurrency *(Why?)*. A mix of data and recursive decompositions is more desirable.
- In discrete event simulation, there might be concurrency in task processing. A mix of speculative decomposition and data decomposition may work well.
- Even for simple problems like **finding a minimum** of a list of numbers, a **mix of data and recursive decomposition** works well.

![Diagram of hybrid decompositions]

3 7 2 9

11 4 5 8

7 10 6 13

1 19 3 9

Data decomposition

Recursive decomposition

Recursive decomposition
Characteristics of Tasks

Once a problem has been decomposed into independent tasks, the characteristics of these tasks critically impact choice and performance of parallel algorithms. Relevant task characteristics include:

• Task generation.
• Task sizes.
• Size of data associated with tasks.
Task Generation

• **Static task generation**: Concurrent tasks can be identified *a-priori*. Typical matrix operations, graph algorithms, image processing applications, and other *regularly structured problems* fall in this class. These can typically be decomposed using *data or recursive decomposition techniques*.

• **Dynamic task generation**: Tasks are *generated as we perform computation*. A classic example of this is in game playing - each 15 puzzle board is generated from the previous one. These applications are typically decomposed using *exploratory or speculative decompositions*. 
Task Sizes

- Task sizes may be **uniform** (i.e., all tasks are the same size) or non-uniform.
- **Non-uniform** task sizes may be such that they can be determined (or estimated) a-priori or not.
- Examples in this class include discrete optimization problems, in which it is difficult to estimate the effective size of a state space.
Mapping Techniques

- Once a problem has been decomposed into concurrent tasks, these must be mapped to processes (that can be executed on a parallel platform).
- Mappings must **minimize overheads**.
- Primary overheads are **communication** and **idling**.
- Minimizing these overheads often represents **contradicting objectives**.
- Assigning all work to **one processor trivially minimizes communication** at the expense of significant idling.
Mapping Techniques for Minimum Idling

Mapping must simultaneously minimize idling and load balance. Merely balancing load does not minimize idling.
Mapping Techniques for Minimum Idling

Mapping techniques can be static or dynamic.

- **Static Mapping**: Tasks are mapped to processes a-priori. For this to work, we must have a good estimate of the size of each task. Even in these cases, the problem may be NP complete.

- **Dynamic Mapping**: Tasks are mapped to processes at runtime. This may be because the tasks are generated at runtime, or that their sizes are not known.

Other factors that determine the choice of techniques include the size of data associated with a task and the nature of underlying domain.
Determining the optimal mapping of tasks is an NP-complete problem in general. Some examples:

- Mapping of tasks with dependencies on a single processor is solvable in polynomial time.
- Mapping of tasks without dependencies on a parallel processors (even 2) is NP-complete.
- The same problem but with uniform task size can be solved in polynomial time.
- If we add dependencies the problem becomes NP-complete.
Schemes for Static Mapping

- Mappings based on data partitioning.
- Mappings based on task graph partitioning.
- Hybrid mappings.
Mappings Based on Data Partitioning

We can combine data partitioning with the "owner-computes" rule to partition the computation into subtasks. The simplest data decomposition schemes for dense matrices are **1-D block distribution** schemes.

**row-wise distribution**

<table>
<thead>
<tr>
<th>$P_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
</tr>
<tr>
<td>$P_2$</td>
</tr>
<tr>
<td>$P_3$</td>
</tr>
<tr>
<td>$P_4$</td>
</tr>
<tr>
<td>$P_5$</td>
</tr>
<tr>
<td>$P_6$</td>
</tr>
<tr>
<td>$P_7$</td>
</tr>
</tbody>
</table>

**column-wise distribution**

| $P_0$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ | $P_7$ |
Block Array Distribution Schemes

Block distribution schemes can be generalized to higher dimensions as well.

<table>
<thead>
<tr>
<th>$P_0$</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_4$</td>
<td>$P_5$</td>
<td>$P_6$</td>
<td>$P_7$</td>
</tr>
<tr>
<td>$P_8$</td>
<td>$P_9$</td>
<td>$P_{10}$</td>
<td>$P_{11}$</td>
</tr>
<tr>
<td>$P_{12}$</td>
<td>$P_{13}$</td>
<td>$P_{14}$</td>
<td>$P_{15}$</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>$P_0$</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
<th>$P_6$</th>
<th>$P_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_8$</td>
<td>$P_9$</td>
<td>$P_{10}$</td>
<td>$P_{11}$</td>
<td>$P_{12}$</td>
<td>$P_{13}$</td>
<td>$P_{14}$</td>
<td>$P_{15}$</td>
</tr>
</tbody>
</table>

(b)
For multiplying two dense matrices $A$ and $B$, we can partition the output matrix $C$ using a block decomposition.

For load balance, we give **each task the same number of elements** of $C$. (Note that each element of $C$ corresponds to a single dot product.)

The choice of precise decomposition (1-D or 2-D) is determined by the associated communication overhead.

In general, higher dimension decomposition allows the use of larger number of processes.
Data Sharing in Dense Matrix Multiplication

(a)

(b)
Cyclic and Block Cyclic Distributions

- If the **amount of computation** associated with data items varies, a block decomposition may lead to **significant load imbalances**.
- A simple example of this is in **LU decomposition** (or Gaussian Elimination) of dense matrices.
LU Factorization of a Dense Matrix

A decomposition of LU factorization into 14 tasks - notice the significant load imbalance.

\[
\begin{pmatrix}
A_{1,1} & A_{1,2} & A_{1,3} \\
A_{2,1} & A_{2,2} & A_{2,3} \\
A_{3,1} & A_{3,2} & A_{3,3}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
L_{1,1} & 0 & 0 \\
L_{2,1} & L_{2,2} & 0 \\
L_{3,1} & L_{3,2} & L_{3,3}
\end{pmatrix}
.\begin{pmatrix}
U_{1,1} & U_{1,2} & U_{1,3} \\
0 & U_{2,2} & U_{2,3} \\
0 & 0 & U_{3,3}
\end{pmatrix}
\]

1: \( A_{1,1} \rightarrow L_{1,1}U_{1,1} \)

2: \( L_{2,1} = A_{2,1}U_{1,1}^{-1} \)

3: \( L_{3,1} = A_{3,1}U_{1,1}^{-1} \)

4: \( U_{1,2} = L_{1,1}^{-1}A_{1,2} \)

5: \( U_{1,3} = L_{1,1}^{-1}A_{1,3} \)

6: \( A_{2,2} = A_{2,2} - L_{2,1}U_{1,2} \)

7: \( A_{3,2} = A_{3,2} - L_{3,1}U_{1,2} \)

8: \( A_{2,3} = A_{2,3} - L_{2,1}U_{1,2} \)

9: \( A_{3,3} = A_{3,3} - L_{3,1}U_{1,3} \)

10: \( A_{2,2} \rightarrow L_{2,2}U_{2,2} \)

11: \( L_{3,2} = A_{3,2}U_{2,2}^{-1} \)

12: \( U_{2,3} = L_{2,2}^{-1}A_{2,3} \)

13: \( A_{3,3} = A_{3,3} - L_{3,2}U_{2,3} \)

14: \( A_{3,3} \rightarrow L_{3,3}U_{3,3} \)
LU Factorization of a Dense Matrix

A **serial column-based algorithm** to factor a nonsingular matrix $A$ into a lower-triangular matrix $L$ and an upper-triangular matrix $U$.

```plaintext
1. procedure COL_LU (A)
2. begin
3.   for k := 1 to n do
4.     for j := k + 1 to n do
6.     endfor;
7.   for j := k + 1 to n do
8.     for i := k + 1 to n do
10.    endfor;
11.   endfor;
12. /*
    After this iteration, column $A[k + 1 : n, k]$ is logically the kth
column of $L$ and row $A[k, k : n]$ is logically the kth row of $U$.
    */
13. end COL_LU
```
Block-Cyclic Distribution for Gaussian Elimination

The active part of the matrix in Gaussian Elimination changes. By assigning blocks in a block-cyclic fashion, each processor receives blocks from different parts of the matrix.

Block Cyclic Distributions

- Variation of the block distribution scheme that can be used to alleviate the load-imbalance and idling problems.
- Partition an array into many more blocks than the number of available processes.
- Blocks are assigned to processes in a round-robin manner so that each process gets several non-adjacent blocks.
**Block-Cyclic Distribution**

- A **cyclic distribution** is a special case in which block size is one.
- A **block distribution** is a special case in which block size is $\frac{n}{p}$, where $n$ is the dimension of the matrix and $p$ is the number of processes.

![Diagram](image)
Mappings Based on Task Partitioning

• Partitioning a given task-dependency graph across processes.
• Determining an optimal mapping for a general task-dependency graph is an **NP-complete problem.**
• Excellent **heuristics** exist for structured graphs.
Task Partitioning: Mapping a Binary Tree Dependency Graph

Example illustrates the dependency graph of one view of quick-sort and how it can be assigned to processes in a hypercube.
Task Partitioning: Mapping a Sparse Graph

Sparse graph for computing a sparse matrix-vector product and its mapping.

- **Process 0**
  - **C0 = (1, 2, 6, 9)**

- **Process 1**
  - **C1 = (0, 5, 6)**

- **Process 2**
  - **C2 = (1, 2, 4, 5, 7, 8)**
Schemes for Dynamic Mapping

• Dynamic mapping is sometimes also referred to as **dynamic load balancing**, since load balancing is the primary motivation for dynamic mapping.

• Dynamic mapping schemes can be **centralized** or **distributed**.
Centralized Dynamic Mapping

- Processes are designated as masters or slaves.
- When a process runs out of work, it requests the master for more work.
- When the number of processes increases, the master may become the bottleneck.
- To alleviate this, a process may pick up a number of tasks (a chunk) at one time. This is called Chunk scheduling.
- Selecting large chunk sizes may lead to significant load imbalances as well.
- A number of schemes have been used to gradually decrease chunk size as the computation progresses.
Distributed Dynamic Mapping

• Each process can send or **receive work from other processes**.
• This **alleviates the bottleneck** in centralized schemes.
• There are four critical questions: how are sensing and receiving processes paired together, who initiates work transfer, how much work is transferred, and when is a transfer triggered?
• Answers to these questions are generally **application specific**. We will look at some of these techniques later in this class.