Ministry of Higher Education and Scientific Research
University Center of Mila
Institute of Mathematics and Computer Science
Department of Computer Science
Master 2 I2A – Big Data
2025/2026

TD 1 – Performance Metrics and Size-Up

Objectives

• Understand how workload scales with problem size.
• Learn to estimate the number of processors needed to keep computation time constant.
• Build intuition for scalability in distributed systems.

Part 1 – Theoretical Exercises

Exercise 1 – Size-Up Analysis

We have a computational problem whose workload depends on the size of the input n. The workload is represented by q(n) (number of operations required).
The computing speed of one processor is V operations per second.
The ideal time to solve a problem of size n on p processors is:
$T_{ideal}(n; p) = q(n) / (p \times V)$

a) Base Case: Compute $T_{ideal}(n_0; 1)$ for a problem of size n_0 on one processor.
b) Keeping Time Constant: Derive the formula for the number of processors p required to solve a larger problem of size n in the same time as n_0 on one processor.
c) Example: If $q(n) \propto n^2$, calculate p for $n = 2n_0$ and for $n = 3n_0$.

Exercise 2 – Comparing Workload Growth

Consider three problems with workloads:
• $q_1(n) \propto n$ (linear)
• $q_2(n) \propto n^2$ (quadratic)
• $q_3(n) \propto n^3$ (cubic)
For each, calculate the number of processors required to keep the time constant when n doubles.
Discuss why higher complexity requires disproportionately more processors.

Part 2 – Applied Big Data Exercises

Exercise A – Storage & Processing Needs

A social media platform logs 500 bytes per event and records 200,000 events per second.

1) Compute the total log size generated in one day (in GB and TB).
2) A single server processes data at 100 MB/s. How long to process one day of logs?

3) If we use 20 servers (ideal scaling), how long would it take?
4) Explain why a single server is impractical for Big Data.

Exercise B – Parallel Processing & Speedup

We need to analyze a dataset of 10 TB.
Each node processes 200 MB/s.
We first use 5 nodes, then 20 nodes.

1) Compute processing time with 5 nodes.
2) Compute processing time with 20 nodes (ideal scaling).
3) Compute the speedup and efficiency when scaling from 5 to 20 nodes, assuming a 10% overhead.

Exercise C – Scalability Challenge

A company increases daily data processed from 5 TB to 25 TB.
Originally, 10 nodes were used, all with the same performance.

1) Assuming ideal scaling, how many nodes are needed to keep processing time constant?
2) If overhead increases by 20% at higher scale, how many nodes are realistically required?
3) Discuss why scaling is harder at higher node counts.