Compute and data grids: large scale distributed computing
Exercises

1. **Parallel computing**
   Let a program which execution time on a single CPU is T. Suppose that you have N CPUs available to parallelize this program. Define the speed-up of a program. What is a linear speed-up?

   Suppose now that the program is composed of a fraction f ($f \in [0,1]$) of code that can be parallelized and a fraction $1-f$ of non parallelizable code. What is the minimal execution time of the parallelizable program and the optimal speed-up that can be expected? What is the asymptotic behavior when N increases? Describe what are the consequences.

2. **Virtual communities**
   What is a virtual organization? In technical terms, how can it be set up and what are the related properties?

3. **Statistics collection**
   The cumulative density function of the latency (F_R) measured on successful jobs on the EGEE grid is plotted on figure 1.

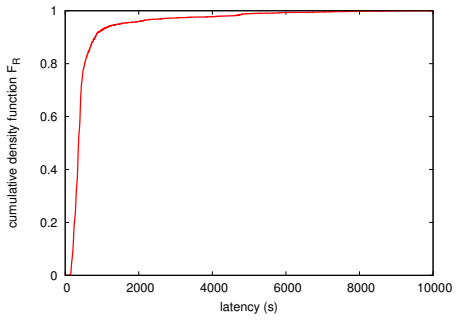

Figure 1: Cumulative density function of the latency

   What is the meaning of this curve (explain the meaning of the value $F_R(2000)$ for example)? What can you infer from the profile of this curve?

4. **Probabilistic modeling**
   One possible strategy to adopt when latency is too high is to cancel jobs that face a latency higher than a threshold t_∞ and to resubmit them again as many time as needed until all jobs pass. The expectation of the total latency, including resubmission, for one job is:

$$E_J(t_\infty) = \frac{1}{\tilde{F_R}(t_\infty)} \int_0^{t_\infty} u \tilde{f_R}(u) du + \frac{t_\infty}{\tilde{F_R}(t_\infty)} - t_\infty$$

   where $\tilde{F_R}$ is the cumulative density function of the latency R of a single job (successful or not) and $\tilde{f_R}$ the probability density function.

   Demonstrate that it can also be expressed as:

$$E_J(t_\infty) = \frac{1}{\tilde{F_R}(t_\infty)} \int_0^{t_\infty} (1 - \tilde{F_R}(u)) du$$

5. **Security**
   Using an asymmetric key system, how can Alice transfer to Bob a document that is both protected from external read and authenticated?

6. **Workload balancing**
   Five computing centers involved in a common scientific activity would like to share their computing resources. These sites, connected through the Internet are geographically spread over Europe. They are each using a batch to control the access to local nodes. Advise the system administrators in these centers to migrate their system towards a grid infrastructure: what are the problems that they are likely to encounter? Is is possible to deploy a single batch manager for all centers? According to you, what is the most appropriate solution?