- 1a. Describe a challenge you can imagine having when writing a program to use **shared** memory.

If the system has a shared-memory architecture, multiple processors compete for access to memory over a shared bus. In such architecture, one of the processors may change the variables to which a second processor has access without the latter being notified. Therefore, the second processor would perform a task on a variable diverse from the one it should be

- 1b. Describe a challenge you can imagine having when writing a program to use **distributed** memory.

In a distributed-memory architecture, each processor has its own local memory. Hence, changes to one of the local memories do not cause changes to the others. However, if one of the processors has to access a local memory different from its own, the required time is longer, leading to higher latency

- 2a. Look up the Top 10 supercomputing list and briefly describe the architecture of the top three machines. List the number of machines of each type on the top 10 (e.g., X are GPU-accelerated, Y contain manycore processors, etc.).

Summit (Oak Ridge): IBM-manufactured supercomputer with 2,414,592 cores and 2,801,664 GB memory. Summit uses 4,608 nodes, each one with 2 IBM POWER9 22C CPUs (multicore processors with 22 cores and 3.07Ghz clock) and 6 NVIDIA Volta V100s GPUs. Each node has 512GB DDR4 + 96GB HBM2 RAM memory GPUs and NVLink for high-speed CPU-CPU and CPU-GPU communication. The node performance is 42TF. The nodes are tied together by a non-blocking fat-tree network of Mellanox EDR Infiniband.

Sierra (LLNL): IBM-manufactured supercomputer with 1,572,480 cores and 1,382,400 GB memory. Sierra has 4,474 nodes equipped with 2 IBM POWER9 22C CPUs (multicore processors with 22 cores and 3.07Ghz clock). Each node has 4 NVIDIA Volta V100s GPUs. The total 1,382,400 GB memory is composed of 256GB CPU memory and 64GB GPU memory per node. The nodes are connected via a Dual-rail Mellanox EDR Infiniband network.

Sunway TaihuLight (Wuxi): NRCPC-manufactured supercomputer with 10,649,600 cores and 1,310,720 GB RAM memory. The system is composed of 40,960 nodes equipped with SW26010 manycore processors with 260 cores each and a 1.45 GHz clock speed. Each processor is composed of 4 Management Processing Elements, 4 Memory Controllers and 4 Computer Processing Elements that have access to 8GB DDR3 memory each.

Of the top 10 supercomputers (TOP500 list):

- 6 are GPU accelerated.
- 3 use manycore processors, while the others have multi-core processors
- 3 have 22-core processors, 2 have 12-core processors, the others have 260-core, 28-core, 68-core,20-core, 24-core.
- Clock speed is 3.1 GHz for 4 supercomputers, the others have processors with clock speed of 1.4, 1.45, 2.2, 2.3, 2.4, 2.6, 2.7 GHz

- 2b. Describe the main characteristics of GPUs, manycore processors, and CPUs - memory, clock speed, structure, etc.

  CPUs' main components are the Arithmetic Logic Unit (ALU) , the control unit (CU) and a processor register.

The CU is the component which receives the information from the outside, decodes it and divides into sequential steps and instructs other components on how to perform such steps.
The ALU is an electronic circuit performing arithmetic and logical operations on input data (operands). The result from ALU can be stored in a registry or in the main memory, at an address generated by the AGU.

In order to reduce time to access the main memory, multi-level CPU caches are adopted, with the L1 cache in proximity of the CPU and whose space is split between data and instructions.
In multicore processors, each core has a L2 (and L3, generally) independent cache.
CPUs have their sequential operations paced by a clock signal produced by an external clock oscillator circuit. As the rate at which the CPU performs instructions is defined by the clock rate, the higher the frequency, the more instructions the CPU can perform in a given instant of time.

GPUs are composed of several multicore processors with **shared** memory and cache.
GPUs typically run at lower clock speeds than CPUs but have much many cores. GPUs have many more ALU units than CPUs. While CPUs are designed for sequential code performance, GPUs are designed for parallel code performance, with single-instruction processors with each core sharing control and instructions with several others.

Manycore processors are those multi-core processors specifically defined for parallel computing. Their throughput and energy consumption are optimized, while latency is penalized vs multicore processors. Typically, they have lower single thread performance then multicore processors, with lower clock rates, but a very high number of processors.

- 2c. Based on what we've talked about so far, postulate challenges of solving the neutron transport equation in a way that would work on all of these architectures.

  Challenges in solution of the transport equation may arise with regards to the parallelization. Indeed, if the solution method involves the subdivision of the domain (e.g. the spatial mesh) among processors, how the division is performed would depend on the system architecture. The programmer, in this case, needs to define a strategy to balance out the tasks of each parallel processor, not to have imbalances in time and computational cost among them.
  Moreover, some systems would be based on shared memory (GPU), while others on distributed memory (manycore).
  Different architecture may cause different bottlenecks in the solution; for instance, in one case delays in program execution may be due to access to memory, while in another the critical resource is the processor computational strength. Hence, there may not be a one fits all optimization strategy and the programmer might have to look for the best strategy in each case.

3. We often measure convergence by comparing one iteration to the previous iteration (rather than the solution, since we presumably don't know what it is). Imagine that you have software that gives the following solution vectors […]

a

$$\left|\left|x_{n-1}\right|\right|_1 = \sum_{i=1}^{5} \left|x_{n-1}^{(i)}\right| = 2.15$$

$$\left|\left|x_n\right|\right|_1 = \sum_{i=1}^{5} \left|x_n^{(i)}\right| = 2.40$$

$$Absolut\ Error: \left|\left|e\right|\right|_1 = \left|\left|x_n - x_{n-1}\right|\right|_1 = \sum_{i=1}^{5} \left|x_n^{(i)} - x_{n-1}^{(i)}\right| = 0.35$$

$$Relative\ Error: \frac{\left|\left|e\right|\right|_1}{\left|\left|x_{n-1}\right|\right|_1} = 0.16$$

b

$$\left|\left|x_{n-1}\right|\right|_2 = \left(\sum_{i=1}^{5} \left|x_{n-1}^{(i)}\right|^2\right)^{0.5} = 1.1630$$

$$\left|\left|x_n\right|\right|_2 = \left(\sum_{i=1}^{5} \left|x_n^{(i)}\right|^2\right)^{0.5} = 1.2186$$

$$Absolut\ Error: \left|\left|e\right|\right|_2 = \left|\left|x_n - x_{n-1}\right|\right|_2 = \left(\sum_{i=1}^{5} \left|x_n^{(i)} - x_{n-1}^{(i)}\right|^2\right)^{0.5} = 0.1658$$

$$Relative\ Error: \frac{\left|\left|e\right|\right|_2}{\left|\left|x_{n-1}\right|\right|_2} = 0.1426$$

c

$$\left|\left|x_{n-1}\right|\right|_\infty = \max_{i=1,\dots 5} \left|x_{n-1}^{(i)}\right| = 0.85$$

$$\left|\left|x_n\right|\right|_\infty = \max_{i=1,\dots 5} \left|x_n^{(i)}\right| = 0.9$$

$$Absolut\ Error: \left|\left|e\right|\right|_\infty = \left|\left|x_n - x_{n-1}\right|\right|_\infty = \max_{i=1,\dots,5} \left|x_n^{(i)} - x_{n-1}^{(i)}\right| = 0.10$$

$$Relative\ Error: \frac{\left|\left|e\right|\right|_\infty}{\left|\left|x_{n-1}\right|\right|_\infty} = 0.1111$$

The least restrictive convergence is that according to the infinity norm.

With the new values, we have the following results:

$$\left|\left|e\right|\right|_1 = 0.16$$

$$\frac{\left|\left|e\right|\right|_1}{\left|\left|x_{n-1}\right|\right|_1} = 0.0635$$

$$||e||_2 = 0.1058$$

$$\frac{||e||_2}{||x_{n-1}||_2} = 0.0835$$

$$||e||_\infty = 0.10$$

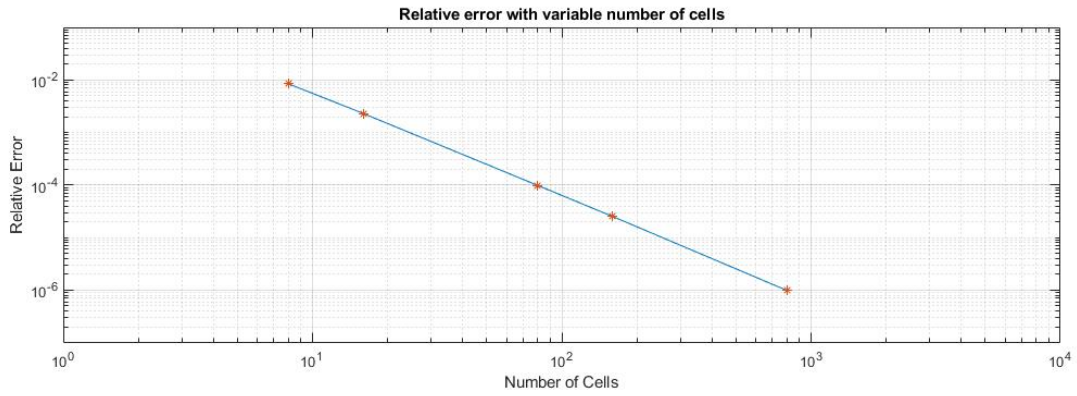
$$\frac{||e||_\infty}{||x_{n-1}||_\infty} = 0.1136$$

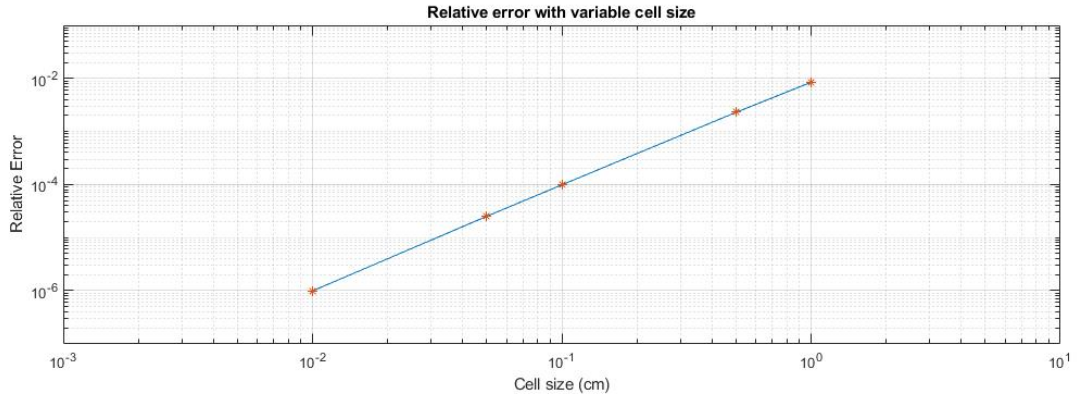
The absolute error continues to be minimized by choosing the infinity norm; however, now the relative error is smaller with the 1-norm. Therefore, even though for any vector in Rn the infinity norm is always smaller than (or equal to) the other norms, when it comes to relative errors we do not have a universal relation, so the choice of the norm to use in the convergence in criteria depends on the specific case

.

4. The relation is $N = 8/h$

Plotting the data:





The relation between N and Relative Error is a decreasing straight line in a LogLog plot, which means that

$$RelErr = \frac{A}{N^B}$$

With A and B to derive from the graph.

Moreover we see that as N doubles, the error is divided by four, hence B=2:

$$RelErr = \frac{A}{N^2}$$

The relation between h and Relative Error is an increasing straight line in a LogLog plot, which means that

$$RelErr = C * h^D$$

Again, here we see that as h halves, the error is divided by 4, so $RelErr = C * h^2$.

(C=A/64, due to the relation between N and h)

6. What are six underlying assumptions in the neutron transport equation? Write at least one sentence explaining what each assumption means and why we need or want to make it.

I. A) Particle are point objects; B) $\lambda = h/mv$ is small compared to the atomic diameter. *Explanation*: A*) Rotation and vibration can be disregarded, as the object become 0 dimensional. B) Heisenberg's uncertainty principle can be disregarded for macroscopic objects, whose states can be fully described by location and velocity vectors at any time.*

II. Neutral particles travel in straight lines between collisions. *Explanation: Particle trajectory is fully determined by its previous collision/ emission. Actions-at-a-distance (except from Electromagnetic force) do not need to be modeled and deflection are due to short-range forces. So the Ω that a particle has when it enters the n+1 collision is the same it had at the exit channel of the nth collision.*

III. Particle-Particle collisions are negligible. *Explanation: The described particles (i.e. neutrons) cannot interact with each-other. From a mathematical perspective, this simplifies the transport equation as it makes it linear, i.e. parameters such as cross sections, ν, χ, etc do not depend on the flux (but on the independent variables).*

IV. Material properties are isotropic*. Explanation: The value of a material property does not depend on the direction of the particle is approaching it. This cuts the dependence of the properties on the direction of the incident particles (Note: this holds also for the scattering cross section, as it depends only on $\Omega \cdot \Omega'$ and not on the individual values).*

V. Material composition is time independent. *Explanation: Material properties (for example cross sections) do not change in time. We do that to simplify the equation and reduce the variables they depend on. (Cross sections can be moved out of the integral when integrating over time).*

VI. Quantities are expected values. *Explanation: we do not consider fluctuations of quantities around their mean, and we remove their statistical dependence on time. Thus, we do not consider density distribution (which should be integrated over their statistical distribution) but exact values.*

6. Consider the transport equation:
a) Briefly describe what each term in the Transport Equation physically represents.

A. TIME DEPENDENCE: The term represents the time variation of the neutron angular density
B. STREAMING TERM: The term represents the loss rate of particles (of energy E and direction $\underline{\Omega}$) through the boundaries

C. TOTAL INTERACTION TERM: Term representing the total loss rate of particles (of energy E and direction $\underline{\Omega}$) because of interactions (i.e. parasitic absorptions, absorptions for fission, outscattering)
D. EXTERNAL SOURCE TERM: Term representing particle source rate due to an external source (of particles of energy E and direction $\underline{\Omega}$)
E. INSCATTERING SOURCE RATE: Term representing the source rate of particles for that specific energy and direction due to scattering from all other energies and directions
F. FISSION SOURCE RATE: Source rate of particles of that given energy and direction because of fissions generated by particles of any energy and direction

b) Rewrite the time independent form of the equation to include azimuthal symmetry. Show the steps needed to get there.

See scanned PDF.

1) TIME INDEPENDENT T.E.:

$$\frac{\partial \cdot}{\partial t} = 0 \Rightarrow S = S(\bar{r}, E, \hat{\Omega}) \;,\; \psi = \psi(\bar{r}, E, \hat{\Omega})$$

$$\hat{\Omega}\cdot\nabla\psi + \Sigma_T \psi = S + \int_0^{\infty} dE' \int_{4\pi} d\hat{\Omega}'\, \Sigma_S(\bar{r}, E' \to E, \hat{\Omega}' \to \hat{\Omega})\, \psi(r, E, \hat{\Omega}') + \frac{\chi(E)}{4\pi} \int_0^{+\infty} dE'\, \nu\Sigma_F(\bar{r}, E') \int_{4\pi} d\hat{\Omega}'\, \psi(\bar{r}, E', \hat{\Omega}')$$

2) STUDY $\hat{\Omega}$

$$d\hat{\Omega} = d\varphi \underbrace{\sin\vartheta\, d\vartheta}_{d\mu}\, d\varphi = d\varphi\, d\mu$$

$$\mu = \cos\vartheta \equiv \hat{\Omega}\, z$$

$$\int_{4\pi} d\hat{\Omega} = \int_0^{2\pi} d\varphi \int_0^{\pi} \sin\vartheta\, d\vartheta = \int_0^{2\pi} d\varphi \int_{-1}^{1} d\mu = 4\pi$$

3) EXPRESS TERMS AS FUNCTIONS OF μ and ASSUME INDEPENDENCE FROM φ

$$\psi = \psi(\bar{r}, \hat{\Omega}, E) \;;\; \psi\, d\hat{\Omega} = \psi(r, \hat{\Omega}, E)\, d\mu\, d\varphi = \psi(\bar{r}, \mu, E)\, d\mu\, d\varphi$$

$\Sigma_S \to \Sigma_S(\bar{r}, E \to E, \hat{\Omega}'\cdot\hat{\Omega}) = \Sigma_S(\bar{r}, E' \to E, \mu)$, i.e. scattering depends only on the cosine of scattering angle

$$\int_{4\pi} d\mu\, \psi = \int_0^{2\pi} d\varphi \int_{-1}^{1} \psi(\bar{r}, \mu, E)\, d\mu = 2\pi \int_{-1}^{1} d\mu\, \psi(\bar{r}, \mu, E) = 2\pi \cdot \phi(\bar{r}, E)$$

(HP) 1D $\Rightarrow \bar{r} \to z$

The streaming term becomes:

$$\hat{\Omega}\cdot\nabla\psi(\bar{r}, E, \Omega) \to \Omega_z \frac{\partial\psi}{\partial z}(z, E, \mu) = \mu \frac{\partial\psi}{\partial z}$$

4) REWRITE THE EQN:

$$\mu \frac{\partial\psi(z,E)}{\partial z} + \Sigma_t(z,E)\,\psi = S(z,E) + \int_0^{\infty} dE' \cdot 2\pi \int_{-1}^{1} d\mu\, \Sigma_S(z, \mu, E' \to E)\, \psi(z, \mu E') + \frac{\chi(E)}{4\pi} \int_0^{+\infty} dE'\, \nu\Sigma_F\, \phi(z,E')$$

Actually, we do not need the 1D explicit requirement; indeed

$\Omega_x = \sin\vartheta \cos\varphi$

$\Omega_y = \sin\vartheta \sin\varphi$

Now, as we are in azimuthal symmetry, the physical behaviour (and its mathematical form) shall not change with different values of φ.

Consider $\varphi = 0 \Rightarrow \Omega_y = 0 \Rightarrow \hat{\Omega}\cdot\nabla\psi = \sin\vartheta \cdot 1 \cdot \frac{\partial\psi}{\partial x} + \Omega_z \frac{\partial\psi}{\partial z}$

Consider $\varphi = \pi/2 \Rightarrow \Omega_x = 0 \Rightarrow \hat{\Omega}\cdot\nabla\psi = \sin\vartheta \cdot 1 \frac{\partial\psi}{\partial y} + \Omega_z \frac{\partial\psi}{\partial z}$

~~$\frac{\partial\psi}{\partial y} = \frac{\partial\psi}{\partial z}$~~ Since we can continue this reasoning with whatever value of φ, it is then clear that the requirement ~~or~~ to have a constant $\hat{\Omega}\cdot\nabla\psi \ \forall\varphi$ is $\frac{\partial\psi}{\partial x} = \frac{\partial\psi}{\partial y} = 0$

Therefore, also without requiring the problem to be 1D, we get

$$\hat{\Omega}\cdot\nabla\psi \underset{\text{AZIMUTHAL SYM}}{\overset{\text{HP}}{=}} \mu \frac{\partial\psi}{\partial z}(\bar{r}, E, \mu)$$

and we can rewrite the TT as

$$\mu \frac{\partial}{\partial z}\psi(\bar{r}, E, \mu) + \Sigma_t(\bar{r}, E)\psi(\bar{r}, E, \mu) = S(\bar{r}, E, \mu) + 2\pi \int_0^{+\infty} dE' \int_{-1}^{1} \Sigma_s(\bar{r}, E' \to E, \mu)\psi \, d\mu$$

$$+ \frac{\chi(E)}{2\pi} \int_0^{\infty} dE' \, \nu\Sigma_F(\bar{r}, E') \, \phi(\bar{r}, E)$$