**Haniye Kashgarani**

**Multicore Programming – Assignment 1**

**Prob 1**

1) Multiprocessing:

Parallel Computing involves executing subtasks at the same time due to each subtask owning exclusively a portion of CPU resources. There is no switching between the tasks and each subtask will reserve some CPU resources.

Also, from a calculation perspective, it is defined as the simultaneous execution of many (millions of) calculations based on the principle of a larger task can be divided into independent smaller tasks known as subtasks which can be executed concurrently.

Parallel computing requires combining an understanding of hardware, software, and parallelism to develop an application.

2)

Multitasking is the illusion in which the processes are not running simultaneously but advance together in the human time scale. This means multiple tasks make progress at the same time, it does not mean that the tasks are running in parallel and simultaneously, but multitasking is about time-slicing between the tasks and switching between the tasks. It is a broader picture of concurrency. OS does multi-tasking, having 4 cores and running 1000 tasks.

There are two basic types of multitasking: preemptive and cooperative.

3)

- SISD (Single Instruction Single Data)
- SIMD (Single Instruction Multiple Data)
- MISD (Multiple Instruction Single Data) and
- MIMD (Multiple Instruction Multiple Data)

4) NUMA stands for Non-Uniform Memory Access.

(Depending on the quality of the Fast Interconnect, a given CPU may not experience the same effective distance or latency to all the M memory blocks. This is known as Memory Affinity, and because of that each memory block will exhibit different access time or non-uniform access time, so the SMP system is not symmetric and it is known as NUMA architecture)

5) What are differences between Moore's law and Dennard Scaling?

Dennard's scaling has more constraints than Moore's law. This is because Moore's law just focuses on the number of transistors, and in Dennard scaling, not only the transistor density but also switching speed and power dissipation are considered. Due to Dennard's scaling, the peak performance of CPUs doubled every 18 months and because of that the application performance would increase automatically without the need for parallel processing; however, based on Moore's law just the transistor density will increase, and the performance will be constant.

6) Cache Lines

7)

- L1 is the lowest cache level and L2 is the second level cache.
- L1 has the fastest access time and L2 is slower than L1 cache memory in access time.
- L1 has the smallest memory size and L2 cache is bigger than L1 cache memory in size.
- L1 closest to the core and registers.
- L1 is used to store both data and instructions and L2 is only used to store data.
- L1 is specific to a core inside the socket and L2 is shared among all cores in a socket. You cannot store instructions because it is common to all cores in a socket.

8) The number of CPUs that can be connected to a single Fast Interconnect and a Single Shared Memory Block is limited due to Coherency issue with Shared Memory Access. Currently, the maximum number of CPUs in a SMP System is Sixty One.

Depending on the quality of the Fast Interconnect a given CPU may not experience the same "effective distance (latency)" to all the M memory blocks, and this is called "memory affinity". So, because of that each memory block in the shared M memory block will exhibit different access time or non-uniform access time. Therefore, SMP system is not perfectly symmetric and it is known as NUMA architecture. Since it's impossible to build memory blocks with the same latency and distance, the number of CPUs that can be connected to a single Fast Interconnect and a single shared memory block is limited due to "coherency" issue with shared memory access. Because of this practical limitation in a single node you cannot have more that 61 CPUs.

9)

When the efficiency is equal to 100% ($\boldsymbol{\eta = 1}$) , the speedup will go up linearly and will be equal to number of computing units (speedup = N). This means if we add more computing units the speedup will go up linearly because parallel time will be $t_p = t_s/N$. This indicates a linear speedup.

In parallel programming we want to have a super linear speedup not a linear speedup. Because if the speedup is linear then there is no need to have parallel program. Super linear speedup is when the speedup bigger than number of computing units (speedup > N), and efficiency is bigger than 100% ($\boldsymbol{\eta > 1}$). This is the goal of parallel programming to achieve super linear speedup.

Super Linear Speedup is possible due to Parallel Programs handling input data differently and following a different execution path from its sequential counterpart.

10) The single chip or multicore processor chips are known as Sockets. A Socket includes few cores inside that, and SMP is formed by few sockets interconnected around a shared memory block.

**Prob 2**

Speedup = 3, N= 6, $(\boldsymbol{1-\alpha})$ **= ?**

$$\boldsymbol{Speedup = \frac{t_s}{t_p} = \frac{1}{(1-\alpha)+\frac{\alpha}{N}} = \frac{1}{(1-\alpha)+\frac{\alpha}{6}} = 3}$$

$\alpha = 0.8$ and $(1-\alpha) = 0.2$

20 percent of the code is executed sequentially, and 80 percent is executed in parallel.

**Prob 3**

Here based on Amdahl's law for achieving 5-fold speedup we need infinite number of cores which is impossible so you cannot achieve a speedup of 5 with this distribution of code.

**Prob 4**

Based on Gustafson-barris's rebuttal

$$\boldsymbol{Speedup} = (\mathbf{1}-\propto) + \boldsymbol{N} \propto$$

A) Here $(\mathbf{1}-\propto) = \mathbf{0.1}$ and $(\propto) = 0.9$, N = 5, speedup = ? ➔Speedup = 4.6
B) If speedup = 9.2, N=? ➔ N ≈10

**Prob 5**

We should buy Machine X with 4 CPUs, each CPU capable of executing the application in 1 hour on its own. Because in parallel programming not always by increasing the number of CPUs you can achieve better speedup.

Based on Scenario 2 of Amdahl's law the speedup achieved when using Machine X compared to Machine Y is about 4.57. So we will have more speed up and we choose the Machine X.