Indian Institute of Technology Guwahati

Department of Computer Science and Engineering

End End Semester Examination Course: CS528 (High Performance Computing) Date: 7th May 2023

Timing: 2.00PM-5.00PM (*Write assumption clearly if you assume anything for answering question*) **Full Marks: 50**

==========

1. **[9 (=2+4+3) Marks]**
   a) Describe the problem $P \mid p_j, tree, a_j=0, d_i=D \mid L_{max}$
   Ans: Homogeneous processor, task with tree dependency and abrbiray execution time. All the tasks arrived at time zero, have end to end deadline D. We need to minimized Lmax, which is maximum lateness.
   b) Solve the above mentioned problem efficiently
   Ans: We need to minimized Lmax, which is maximum lateness. As each task have common deadline D, minimizing C_{max} will ensure the minimizing the L_{max}. Hence we can use any Critical Path heuristics to solve the problem. *HLF/MSF/Levelwise will perform badly because of pj is arbirary.*
   c) Calculate the value of **Work**, **Span** and **Parallelism** for ***N*** nodes balanced k-nary out-tree with execution time of each task is ***p*** unit time.
   Ans: Work =N*p, Span= p * Log_kN, Parallelism=N/(Log_kN)

2. **[7 (=2+2+3) Marks]**
   a) Given the code, we want to run each iteration of the loop in parallel. Is it possible to run the iterations in parallel? Justify your answer with reason.

```
for(i=0;i<N;i++){X[3*i]=X[2*i+1]+5.0;}
```

   Ans: After putting `X[3*i]=X[2*i+1]` in X[a*i+b]=X[c*i+d] format: a=3, b=0, c=2, d=1. GCD(c,a) divides (d-b) for loop dependence: GCD(2,3)=1 divides 1. So there is dependency.

   b) Given a computer system with peak performance of ***20TF/s*** and achievable date bandwidth to the compute is ***10GB/s***. Calculate the expected performance of the following code on the system assuming the size of a float data is ***4B*** and system uses ***write not allocate mode in the cache***.

```
for(i=0;i<N;i++)   //float a[N],b[N],c[N],d[N];
                a[i]=s*b[i]+(c[i]-d[i])*(c[i]-d[i]);
```

   Ans: Ans: Ppeak=20 Tera F/s, I=4F/12B, bs=10GB/s, //No fetch for a[], one time fetch for b[i]. c[i], d[i]
   Expected Perf.= min(Ppeak,I*bs)= min(20TF/s, 4/12 F/B * 10GB/s)
   = min(20 Tera F/s, 3.33 GF/s)= 3.33 GF/s.

   Cases: with I=5F/12B or I=4F/16B or I=4F/20B or I=5F/24B; partial credit is awarded: only one marks
   c) Design an interconnection network with bisection bandwidth ***(lg N) /2*** and diameter ***lg lgN + sqrt([N/lg N]-1)***, where ***N*** is number of node in the network.
   Ans: Hypercube of Mesh: lgN node Hypercube responsible for BS lgN/2, and Diameter of HC is lglgN and for mesh ***sqrt([N/lg N]-1). If we connect middle node of mesh is connected to the hypercube value of Dim of mesh will be sqrt([N/lg N]-1) instead of 2* sqrt([N/lg N]-1). Hypercube of Torus is fine.***

3. [**6 Marks**] Given a set of ***M*** identical processors with different failure rate and given a set of ***N*** tasks with different execution time, schedule these ***N*** tasks to ***M*** processors such that it primarily minimize C_{max} and secondarily maximize the reliability of the schedule for whole task set. Reliability of executing ith task on jth processor is given by $\mathbf{exp(-f_j * t_i)}$ where $\mathbf{t_i}$ is execution of i^{th} task and $\mathbf{f_j}$ is failure rate of j^{th} processor.
   Ans: As the primary objective is Cmax, it can be done in two phases
   - Phase I : Schedule for Cmax using any good approach using the largest processing time (LPT) rule (or ILP for optimal) assuming all processors with zero failure rate.
   - Phase II: Sort the processor index based on load and map the highest loaded processor to the smallest index (processor with lowest failure rate) and so on.

4. [**6 Marks**] An airline name DHAKKAN, always use over booking policy. Suppose cost of ticket per passenger is C and number of seat in airplane is N for the flight. Penalty need to be paid or alternative flight needs to be arranged if the number of turned up passenger is greater than N, and per passenger it is C*(1+d) amount. Assume probability of the passenger turning up for the flight is p. Calculate the number of tickets (one ticket/per passenger) overbooking (in terms of N, C, d, p) the airline perform to maximize the profit from the flight.

Ans: As mentioned in question, per passenger penalty is C*(1+d) and in this case value of d cannot be negative and bound to be some positive value, hence profit decrease if the expected turn out is more than capacity. Assume X ticket got overbooked, profit is maximized if no penalty needed to be paid. Hence (N+X)p –N >=0 and in equality case, (N+X)p-N=0 ==> X=N/p- N as p <1, N/p > N.

5. [**6 (=3+3) Marks**] Suppose ***N*** number of students need to go **from IITG to Sajusajai Stadium Guwahati** and the IITG administration wants to arrange free transport service for the students. Given four types of vehicles in the disposals with capacities 4, 10, 20 and 50 with costs Rs 300, Rs 500, Rs 800 and Rs 1500 respectively. (a) Formulate the problem to be solve mathematically. (b) Design an efficient approach to find appropriate numbers of type of vehicles IITG administration should hire to minimize the cost to transport all the ***N*** students.

Ans (a): Here our main aim is to find value of n1, n2, n3 and n4 the number vehicle of type 1, type 2, type 3 and type 4 respectively, such that

n1*4 + n2*10+ n3*20 + n4*50 >=N and

minimize (n1*300+n2*500+n3*800+n4*1500).

With constraints all value of n1, n2, n3 and n4 are zero or positive integer. Generally bigger vehicle is cheaper per person.

Ans (B) : Cost per person in vehicle type 1: 300/4 =75, type 2: 500/10=50, type 3: 800/20=40 and type 4: 1500/50=30;

if N>100, use (N/100)*2 number of type 4 vehicles and solve for N=N%100;

else enumurate cases for base cases, upto N=100; Or can be done for 50.

This enumurations are

- if N<=4 use one type 1
- if N>4 and N<=10 use one type 2
- if N>10 and N<=20 use one type 3
- if N>20 and N<=24 use one type 3 and one type 1
- if N>24 and N<=30 use one type 3 and one type 2
- if N>=30 and N<=50 use one type 4
- if N>50 and N<=54 use one type 4 and one type 1
- if N>54 and N<=60 use one type 4 and one type 2
- if N>60 and N<=70 use one type 4 and one type 3
- if N>70 and N<=74 use one type 4, one type 3 and one type 1
- if N>74 and N<=80 use one type 4, one type 3 and one type 1
- if N>80 use two type 4

6. [**8=4+4 Marks**] Suppose you are using an EWMA predictor ***E(t) = α*E(t-1) + (1-α) *O(t-1)*** with ***α*** =0.5, where ***E(t)*** and ***O(t)*** are estimated and observed values at time ***t.*** Assume the initial estimated value is 0. The observed values are bounded by 0 and 100. What will be the prediction error in % in the long run?

   a. If the observe values follow a pattern : ***O(t) = (t%6) * 20***;
   b. If the observed values do not follow any pattern but behaves randomly: ***O(t) = rand()%100*** ;

   Ans (a) : In long run it will settle at 2/3*100, which is 66.66% error in prediction.

E(t)	0	0	10	25	42.5	61.25	80.625	40.312	30.15	35.07	47.53	63.76	81.88	40.94
O(t)	0	20	40	**60**	**80**	**100**	**0**	20	40	60	80	**100**	**0**	**20**
Error	0	20	30	35	37.5	41.25	80.625	20.312	10.15	24.93	32.47	36.23	-81.8	20.94

Error continues as 81.9x, 20.9x, 9.5x, 24.7x, 32.7x, 36.2x and repeats. **Which is around 34.31%.**

Code Snippet to test the same:

```
float E=0, O, Err, TE=0;
for(int i=0;i<10000;i++){
    O=(i%6)*20; //O=rand()%100; for other cases
    Err=O-E; printf("i=%d E=%f O=%f Err=%f \n", i, E, O, abs(Err)); TE+= abs(Err);
    E=(E+O)/2;
} printf("Avg Err=%f", TE/10000);
```

Ans (b): Observed value of is randomly distributed and Estimated value suppose be averaged at 50, suppose whole range is divided into 4 parts 0-24, 25-49, 50-74 and 74-99. Average prediction error

=abs(50-Avg(0:24))*1/4+abs(50-Avg(25-49))*1/4+abs(50-Avg(50:74))*1/4+abs(50-Avg(75:99))*1/4
= abs(50-12)*1/4+abs(50-37)*1/4+abs(50-62)*1/4+abs(50-87)*1/4
= 25.5 (which slightly higher than 25) if we consider 4 parts, but if we increase this to 8 parts, 16 parts, 64 parts then the **error converge to 27.25**

7. [**6 (=3+3) Marks**] Given ***N*** webserver tasks and each task with server utilization u_i ***(u_i<1)*** and profit p_i, and there are ***M*** servers, each is powered by renewable/free energy sources and each server has compute power ***U*** (for example 64) but it can utilize U_i (for example 42, which is < 64) amount server freely and this value depends on the amount of free electric power available at that servers. The tasks are non-migratable and one task does not get mapped to multiple servers but one server can host multiple webserver tasks. If a task is mapped to one server and the task gets its required amount of resources (utilization) then the profit of the task gets added to the profit of the system. The system may choose not to allocate/map some tasks. (a) Formulate the problem mathematically. (b) Design an optimal/efficient approach to map these webserver tasks to maximize the profit of the system.
Ans (a) : There are N web server tasks (u_3 and p_i), this needs to mapped to M servers with U_j. This problem can be mapped to LP solution with
Maximize $\sum_i \sum_i X_{ij} p_i$
With constraints C1: $\sum X_{ij} u_i \leq U_j \, for \, all \, j$ ensure the capacity constraints with free energy
C2: $\sum_j X_{ij} \leq 1 \, for \, all \, i$ ensure a task mapped to only one server
C3: $X_{ij} \in \{0,1\}$ ensure task mapping is not partial

Ans (b): One good heuristics : sort tasks by p_i/u_i and mapped to servers one by one.

8. **[2 Marks]** What are the benefits of edge computing over traditional cloud computing?
Ans:Edge server is beneficial for latency sensitive application when cloud latency is extremely high and may not be use full for many IoT applications when task required many real time processing requirement. Load of clous is shared by edge-server. Caching of compute and storage also happens in egde servers.