

Walchand College of Engineering (Government Aided Autonomous Institute)

Vishrambag, Sangli. 416415



Computer Algorithms

M2-Part 1:Introduction to Parallel programming

24-25

D B Kulkarni

Information Technology Department



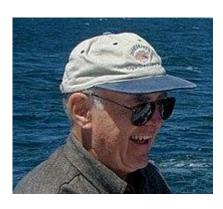
Circuit complexity

How fast is the circuit

Depends on clock speed

First electromechanical general purpose computer, the **Z3**, operated at a frequency of about 5–10 Hz

The first commercial PC, the Altair 8800 (by MITS), used an Intel 8080 CPU with a clock rate of 2 MHz



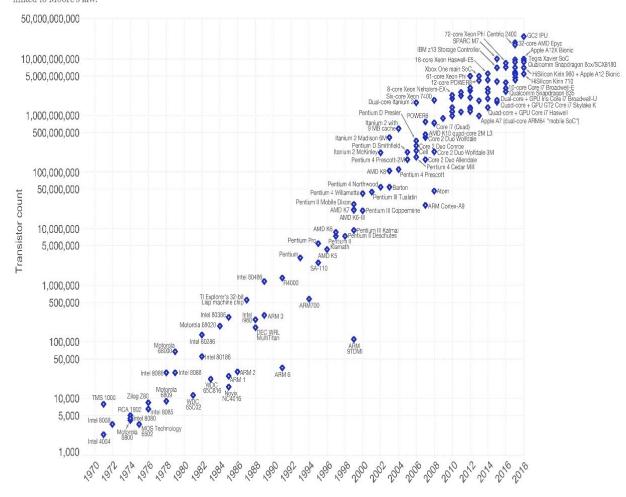
- Transistor count: Moore's

 Iaw is the observation
 that the number of
 transistors in a dense
 integrated circuit (IC)
 doubles about every two
 years
- 1965-75-80----25?

Moore's Law – The number of transistors on integrated circuit chips (1971-2018)



Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.



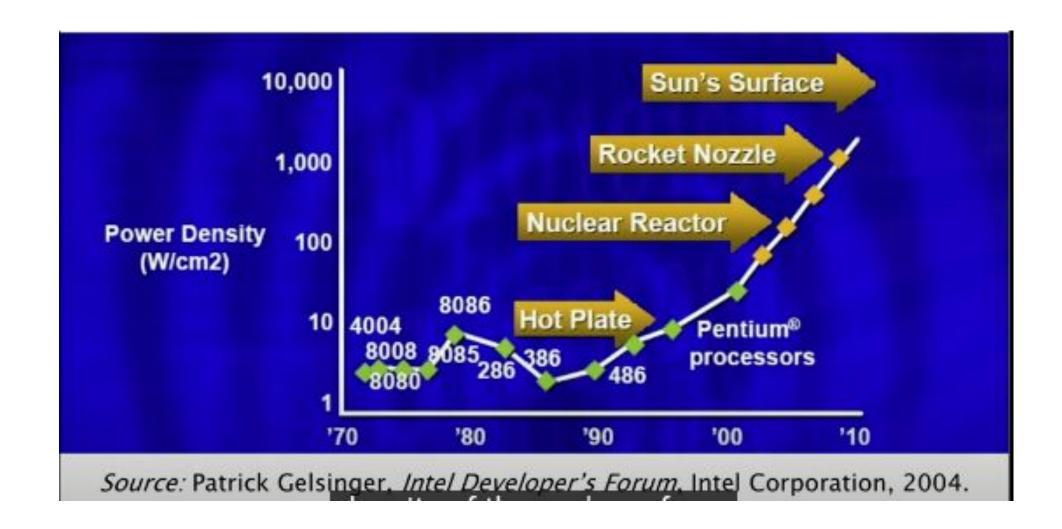
Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Licensed under CC-BY-SA by the author Max Roser.

TY IT 24-25 Sem I DBK Computer Algorithms 2/50



Power Density





Bemchmarks

A benchmark- Act of running a computer program or a set of programs or other operations, in order to assess the relative performance of an object, normally by running a number of standard tests and trials against it.

- NAS Parallel Benchmarks (NPB) developed by the NASA Advanced Supercomputing (NAS) Versions NPB1, NPB2, NPB3
- > LINPACK Benchmark: floating-point rate of execution with a dense system of linear equations.
 - The software used in this experiment is based on two routines from the LINPACK collection
 - DGEFA (Double precision GEneral matrix FActor): Complexities O(n^3)
 - DGESL (Double precision GEneral matrix SoLve): Complexities O((n^2)
- The LINPACK software package has been replaced by LAPACK, which features block algorithms to take advantage of cache memories.
- Matrix computation on distributed memory parallel computers with message passing interprocessing communication is now handled by ScaLAPACK
- > HPL A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers. It is software package that solves a (random) dense linear system in double precision (64 bits) arithmetic on distributed-memory computers

4/50 **Computer Algorithms** TY IT 24-25 Sem I DBK



Supercomputer listing

- List of Supercomputers (rating based)
- TOP500: Based on FLOPS
- 2. GREEN500: Based on FLOPS/watts
- 3. GRAPH500 :Based on TEPS
- Traversed edges per second (TEPS) that can be performed by a supercomputer cluster is a measure of both the communications capabilities and computational power of the machine.

This is in contrast to the more standard metric of floating point operations per second (FLOPS), which does not give any weight to the communication capabilities of the machine.



Processing unit (CPU) speeds

Floating Point Operations Per Second (FLOPS)

- FP16
- FP32 Single precision
- FP64 Double precision

Unit	Notaion	Size	Unit	Notaion	Size
<u>kilo</u> FLOPS	kFLOPS	10 ³	<u>peta</u> FLOPS	PFLOPS	10 ¹⁵
megaFLOPS	MFLOPS	10 ⁶	<u>exa</u> FLOPS	EFLOPS	10 ¹⁸
<u>giga</u> FLOPS	GFLOPS	10 ⁹	<u>zetta</u> FLOPS	ZFLOPS	10 ²¹
<u>tera</u> FLOPS	TFLOPS	10 ¹²	<u>yotta</u> FLOPS	YFLOPS	10 ²⁴

Tensor Processing Unit (TPU) is an Al accelerator application-specific integrated circuit (ASIC) developed by Google specifically for neural network machine learning, particularly using Google's own TensorFlow software. (TPU1, TPU2, TPU3, TPU4, Edge TPU)

TY IT 24-25 Sem I DBK Computer Algorithms 6/50



TOP500-2023

Rank	System	Cores	Rpeak (PFlop/s)	Power (kW)
1	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,679.82	22,703
2	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D. Fujitsu RIKEN Center for Computational Science Japan	7,630,848	537.21	29,899
3	LUMI - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	428.70	6,016

Rank	Rpeak Power (kW) (PFlop/s)
75	13.17



Green500

Rank	TOP50 0 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	255	Henri - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCle, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	34	Frontier TDS - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684
3	12	Adastra - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	921	58.021

India	154	169	Pratyush - Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect, HPE Indian Institute of Tropical Meteorology India	119,232	3.76	1,353	2.781
-------	-----	-----	---	---------	------	-------	-------

TY IT 24-25 Sem I DBK Computer Algorithms 8/50



Parallelism

Data parallel

distributing the data across different nodes, which operate on the data in parallel running the same task on different components of data

Task Parallel

parallelization of computer code across multiple processors in parallel computing environments

distributing tasks—concurrently performed by processes or threads—across different processors

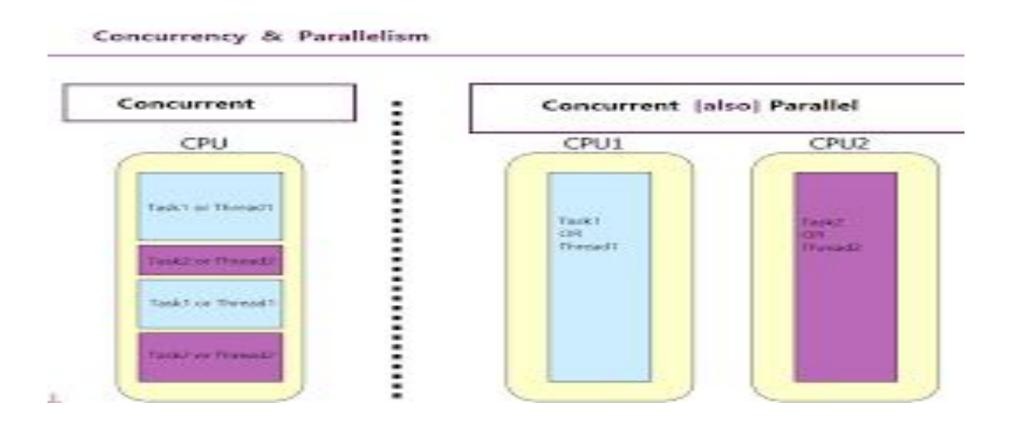
running many different tasks at the same time on the same data.

A common type of task parallelism is pipelining which consists of moving a single set of data through a series of separate tasks where each task can execute independently of the others.

out-of-order execution



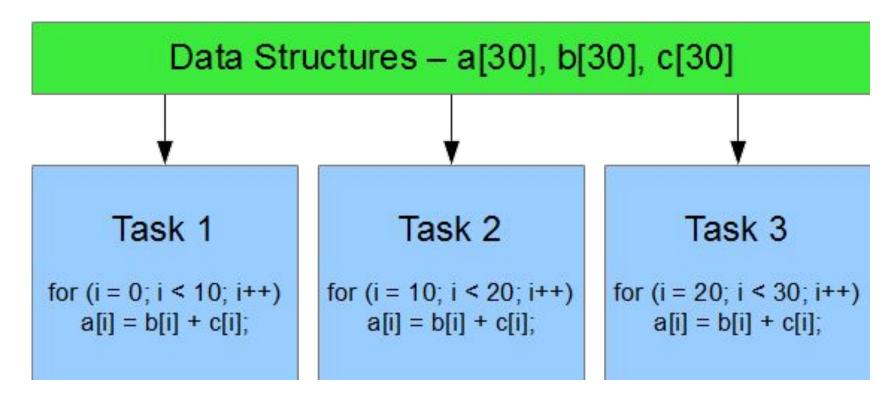
Concurrency and Parallelism



Computer Algorithms TY IT 24-25 Sem I DBK



Data Parallel

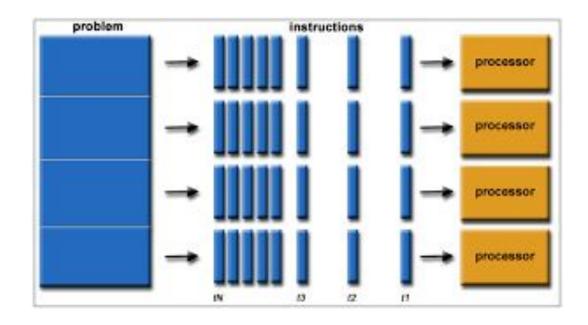


Data parallel

distributing the data across different nodes, which operate on the data in parallel



Task Parallel

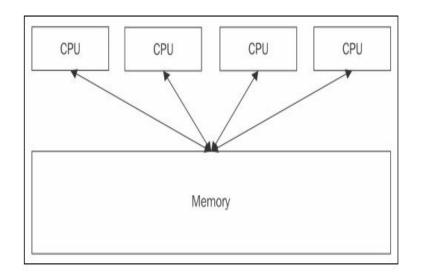


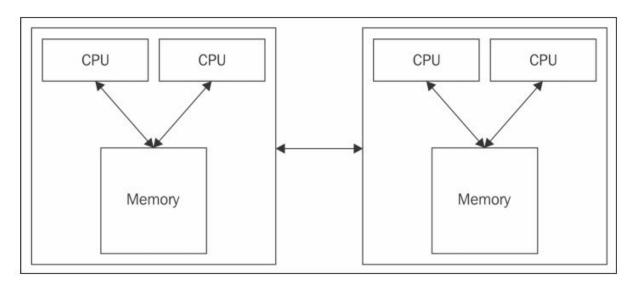
Task parallel

distributing the tasks across different nodes, which completes on the data in parallel



Parallel system





Shared Memory Model

Distributed Memory Model

TY IT 24-25 Sem I DBK Computer Algorithms 13/50



Shared memory models

The shared-memory programming models are used to develop solutions for machine architectures that share one common memory space across a set of processors. The models include:

- the pure shared memory model
- the multithreading models
 - the programmer-controlled model
 - the API controlled model

TY IT 24-25 Sem I DBK Computer Algorithms 14/50



Pure Shared memory models

The pure shared memory programming model identifies data independently of all processors. This model does not associate data with any particular processor. It manages access to shared memory through a system of locks and semaphores.

Advantages:

- no concept of data ownership
- program development is simple

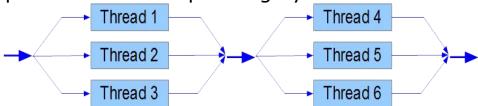
Drawbacks:

- difficult to manage data locality
- programming instructions are low-level



Multithreaded Shared memory models

The multi-threading models divide a part of a process into threads. A thread is an independent stream of instructions that the operating system can schedule independently on the processors. A thread exists within a process and uses that process' resources. Thread-creation requires much less operating-system overhead than process-creation.



The multi-threading models associate each thread with some local data. Each thread can execute concurrently with the other threads. Each thread communicates with other threads through shared memory.

The multi-threading models require synchronization to ensure that concurrently executing threads are not updating the same memory address.

Advantages:

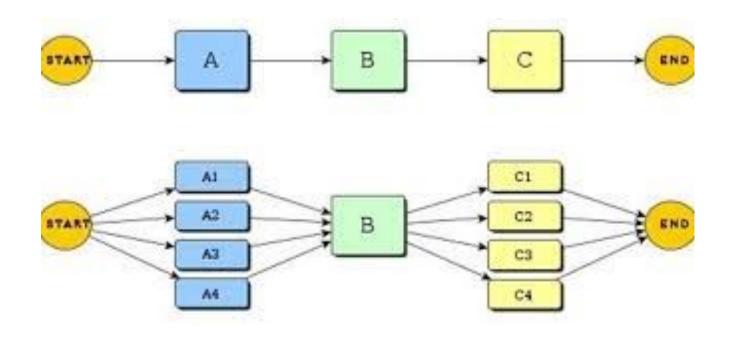
- programmer manages data locality
- run-time system schedules threads

Drawback:

programmer is responsible for determining the parallelism



Shared Memory Model: OpenMP

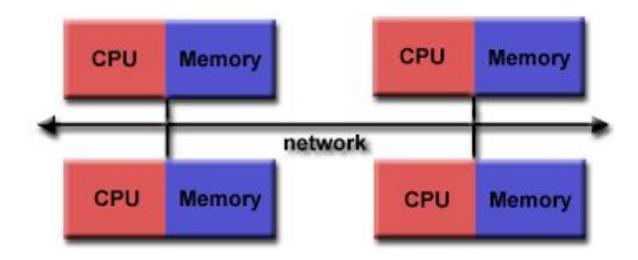


Open Multi Processing (OpenMP) Data parallel

distributing the data across different nodes, which operate on the data in parallel



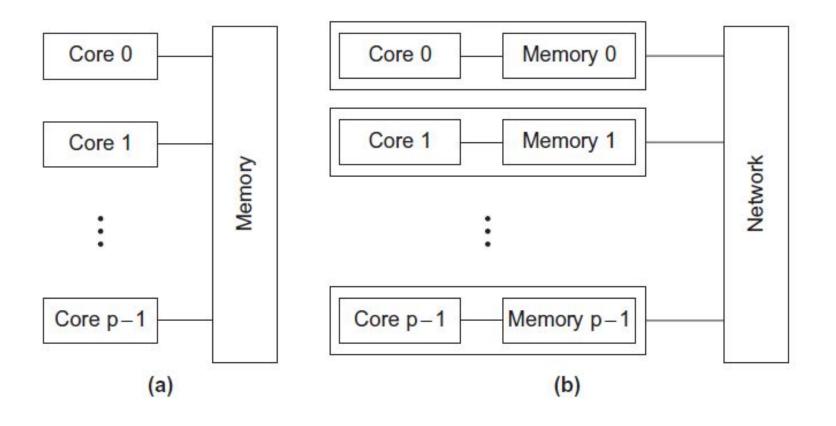
Distributed Memory Model: MPI



Message Passing Interface



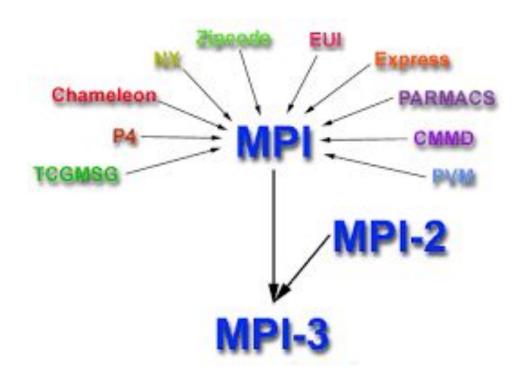
Shared vs Distributed Memory Model



TY IT 24-25 Sem I DBK Computer Algorithms 19/50



MPI



MPI1: May, 1994

MPI2 : July, 1997:focused on process creation and management, one-sided communications, external interfaces and parallel I/O

MPI3 : significant extensions to MPI functionality, including non-blocking collectives, new one-sided communication operations, and Fortran 2008 bindings



Parallel system Performance

Computation- circuit complexity

Communication- commu complexity

TY IT 24-25 Sem I DBK Computer Algorithms 21/50



Flynn's Taxonomy

SISD
Single instruction stream
Single data stream

MISD
Multiple instruction stream
Single data stream

Multiple instruction stream
Single data stream

Multiple instruction stream
Multiple data stream

Multiple data stream

TY IT 24-25 Sem I DBK Computer Algorithms 22/50



Parallel system Performance

Performance Metrics • How do we quantify performance when running in parallel?

Consider execution time T(N,P) measured while running on P "processors" (cores) with problem size/complexity N

- Speedup: $S(N,P) = T(N,1)/T(N,P) = T_s/T_p$ typically S(N,P) < P
- Parallel efficiency: E(N,P)=S(N,P)/P=typically E(N,P) < 1Serial efficiency: typically E(N,1) OR E(N) <= 1
- Scaling: Scaling describes how the runtime of a parallel application changes as the number of processors is increased



Parallel system Performance: Scaling

Two types of scaling: •

1. Strong Scaling (increasing P, constant N): problem size/complexity stays the same as the number of processors increases, decreasing the work per processor

Ideal strong scaling: runtime keeps decreasing in direct proportion to the growing number of processor used

1. Weak Scaling (increasing P , increasing N): problem size/complexity increases at the same rate as the number of processors, keeping the work per processor the same

Ideal weak scaling: runtime stays constant as the problem size gets bigger and bigger

Good strong scaling is generally more relevant for most scientific problems, but more difficult to achieve than good weak scaling



Extent of parallelism

Consider a typical program

- Sections of code that are inherently serial so can't be run in parallel
- Sections of code that could potentially run in parallel •

Suppose serial code accounts for a fraction a of the program's runtime

- Assume the potentially parallel part could be made to run with 100% parallel efficiency, then:
 - Hypothetical runtime in parallel = $T(N,P) = \alpha T(N,1) + (1-\alpha)T(N,1)/P$
 - Hypothetical speedup = $S(N,P) = T(N,1) / T(N,P) = P / (\alpha P + (1-\alpha))$

Speedup fundamentally limited by the serial fraction

• Speedup will always be less than $1/\alpha$ no matter how large P

E.g. for $\alpha = 0.1$:

- hypothetical speedup on 16 processors = S(N,16) = 6.4
- hypothetical speedup on 1024 processors = S(N,1024) = 9.9
- ...
- maximum theoretical speed up is 10.0

Gustafson's Law

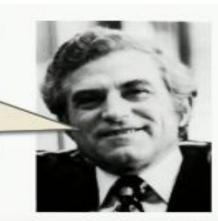
We need larger problems for larger numbers of processors



Amdahl's law

If 50% of your application is parallel and 50% is serial, you can't get more than a factor of 2 speedup, no matter how many processors it runs on.

a-%prog than cant be parallelised p- # processors s=1/(a+(1-a)/p)



Gene M. Amdahl

E.g. for a = 0.1:

- hypothetical speedup on 16 processors = S(N,16) = 1/(0.1+0.9/16) = 6.4
- hypothetical speedup on 1024 processors = S(N,1024) = 9.9
- ...
- maximum theoretical speed up is 10.0

Gustafson's Law

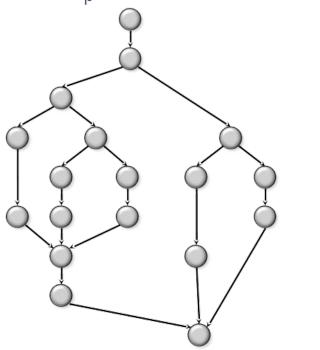
We need larger problems for larger numbers of processors



Work Span Model

Work / Span Model

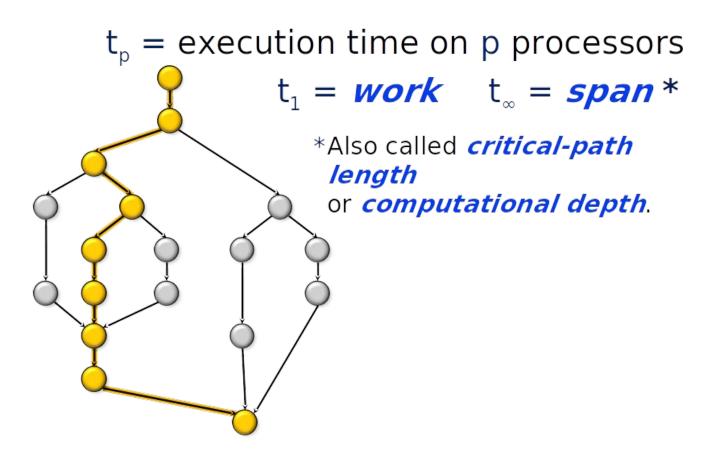
 t_p = execution time on p processors





Work Span Model

Work / Span Model



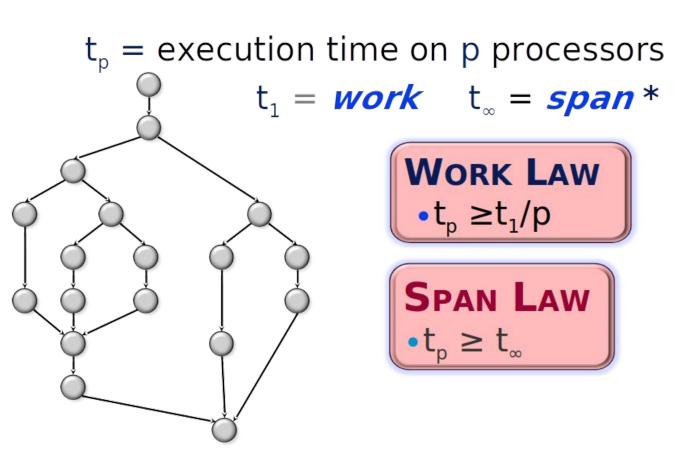
TY IT 24-25 Sem I DBK Computer Algorithms 28/50



Work Span Model

29

Work / Span Model



TY IT 20-21 Sem II PP DBK

3/17/2021



Speedup

```
Def. t_1/t_p = speedup on p processors. 

If t_1/t_p = O(p), we have linear speedup, 

= p, we have perfect linear speedup, 

> p, we have superlinear speedup, 

(which is not possible in this model,because of the Work Law t_p \ge t_1/p)
```

TY IT 24-25 Sem I DBK Computer Algorithms 30/50



Parallel Program Examples

Addition of n numbers

Fibonacci series

Prime number - Finding, generation

prefix sum

Factorial computation

TY IT 24-25 Sem I DBK Computer Algorithms 31/50



MPI Structure (Initilaization)

- int MPI_Init (int *argc, char ***argv)
- int MPI_Comm_size (MPI_Comm comm, int *size)

returns with the size of the group associated with the MPI communicator, comm.

The communicator, MPI_COMM_WORLD is predefined to include all processes. Communicator has two distinguishing characteristics:

Group Name - a unique name given to a collection of processes

Rank - unique integer id within the group: assigned to the process by the system, contiguous and start from 0 within each group

Group and rank serve to uniquely identify each process

Context - the context defines the communication domain and is allocated by the system at run time

int MPI_Comm_rank (MPI_Comm comm, int *rank)

returns with the rank of the processor within the group comm.

TY IT 24-25 Sem I DBK Computer Algorithms 32/50



MPI: Code Implementation Template

1	2	3	4
MPI_Init	MPI_Init	MPI_Init	MPI_Init
	MPI_Comm_Size	MPI_Comm_Size	MPI_Comm_Size
	MPI_Comm_rank	MPI_Comm_rank	MPI_Comm_rank
		MPI_Send	MPI_Send
			MPI_Barrier
		MPI_Recv	MPI_Recv
MPI_Finalize	MPI_Finalize	MPI_Finalize	MPI_Finalize

TY IT 24-25 Sem I DBK Computer Algorithms 33/50



Parallel Paradigm: MPI

```
//first.c
#include "mpi.h"
#include <stdio.h>
int main( int argc, char ** argv )
{
         MPI_Init( &argc, &Sarge );
         printf( "Hello world\n" );
         MPI_Finalize();
         return 0;
}
```

#include "mpi.h" provides basic MPI definitions and types

MPI_Init starts MPI

Note that all non-MPI routines are local; thus the printf run on each process

MPI_Finalize exits MPI

The MPICH implementation provides the commands mpicc for compilation



Simple program

```
How many processors are there? MPI_Comm_size.
Who am I? MPI_Comm_rank.
  The rank is a number between zero and size-1.
Compile program with mpicc and run with mpiexec -n 2 ./a.out
```

```
Process 1
                                                                                                                                  Process 2
int main( int argc, char ** argv )
int rank, size;
MPI_Init( &argc, &argv );
MPI_Comm_rank( MPI_COMM_WORLD, &rank ); MPI_Comm_size( MPI_COMM_WORLD, &size ); printf( "Hello world! I'm %d of %d\n", rank, size );
                                                                                      Hello world! I'm 0 of 2
                                                                                                                                  Hello world! I'm 1 of 2
MPI Finalize();
return 0;
```

Computer Algorithms TY IT 24-25 Sem I DBK



Message passing- Blocking: send and receive

int MPI_Send(void *buf, int count, MPI_Datatype datatype, int dest, int tag, MPI_Comm comm)

int MPI_Recv(void *buf, int count, MPI_Datatype datatype, int source, int tag, MPI_Comm comm, MPI_Status *status)

RED colour: Sender's info Green colour: Receiver's info

Sr No.	Parameter	Information	MPI_Send	MPI_Recv
1	buf	starting address of buffer	I, send buffer	O, recv buffer
2	count	number of elements in buffer	I, send buffer	I
3	datatype	datatype of each buffer element e.g. MPI_INT	I	I
4	int dest, source	rank of process wildcard MPI_ANY_SOURCE	I, dest	I, source
5	tag	message tag wildcard MPI_ANY_TAG	I	I
6	comm	communicator	I	I
7	status	status structure containing a minimum of three entries, specifying the source, tag, and error code of the received message.	NA	O

TY IT 24-25 Sem I DBK Computer Algorithms 36/50



Example of send/receive

```
int main( int argc, char ** argv )
int rank, size;
MPI Init( &argc, &argv );
MPI Comm_rank( MPI_COMM_WORLD,
&rank );
MPI Comm size (MPI COMM WORLD,
&size);
if(rank == 0) {
send data("Hello Worker", size);
else {
receive data(rank);
```

```
void send data(char * data int size)
//int MPI Send(void *buf, int count, MPI Datatype
 datatype, int dest, int tag, MPI Comm comm)
for( int i = 1 : i < size : i++ )
 MPI Send( data, strlen(data)+1, MPI CHAR, i, 0,
 MPI COMM WORLD);
void receive data(int my rank)
char r data[500];
MPI Status status;
//int MPI Recv(void *buf, int count, MPI Datatype
datatype, int source, int tag, MPI Comm comm,
MPI Status *status)
MPI Recv(r data, 500, MPI CHAR, 0, 0,
MPI COMM WORLD, &status);
printf("%s %d\n",r data, my rank);
```



Visualization of send/receive

```
int main( int argc, char ** argv ){
int rank, size;MPI_Init( &argc, &argv );
MPI_Comm_rank( MPI_COMM_WORLD, &rank );
MPI_Comm_size( MPI_COMM_WORLD, &size );
}
```

```
//sender
// rank==0
void send_data(char * data int size)
{
for( int i = 1 ; i < size ; i++ )
MPI_Send( data, strlen(data)+1, MPI_CHAR, i, 0, MPI_COMM_WORLD );
}</pre>
```

```
//Receiver
// rank <> 0
void receive_data(int my_rank)
{
   char r_data[500];
   MPI_Status status;
   MPI_Recv(r_data, 500, MPI_CHAR, 0, 0, MPI_COMM_WORLD, &status);
   printf("%s %d\n",r_data, my_rank);
}
```



int MPI_Barrier(MPI_Comm comm)

Blocks until all members of the group comm have made this call.

This synchronizes all processes.

comm the name of the communicator

int MPI_Finalize(void)

Cleans up all MPI states.

Should be called by all processes.

User must ensure all pending communications complete before calling this routine.

MPI_Wtime ()

returns wall clock (elapsed) time in seconds from some arbitrary initial point

MPI_Wtick ()

returns resolution of MPI_Wtime in seconds

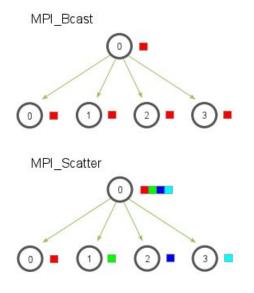
TY IT 24-25 Sem I DBK Computer Algorithms 39/50

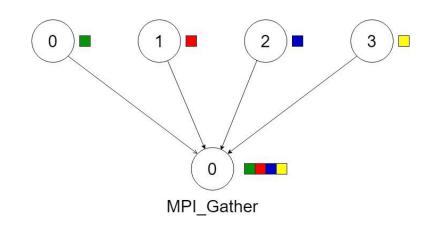


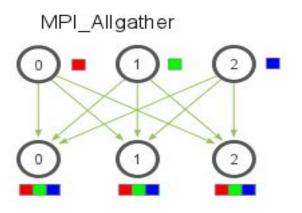
MPI broadcast and all-reduce

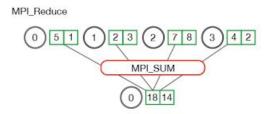
int MPI_Bcast(void* buffer, int count, MPI_Datatype datatype, int root, MPI_Comm comm) int MPI_Allreduce(void *sendbuf, void *recvbuf, int count,MPI_Datatype datatype, MPI_Op op, MPI_Comm comm)

MPI Functions











Other MPI Functions:

- ☐ There are many additional functions, at least 133 functions
- ☐ These functions add flexibility, robustness, efficiency, modularity, or convenience.

TY IT 24-25 Sem I DBK Computer Algorithms 41/50



MPI : Message Transfer

In Message communication, when

> the data buffer is transmitted to the receiver, however, what if the receiver is not ready?

3 Scenarios:

- 1. Wait for receiver to be ready (blocking)
- 2. Copy the message to an internal buffer (on sender, receiver, or elsewhere) and then return from the send call (nonblocking)
- 3. Fail
- > the data buffer is to be received from sender, however, what if the sender has not sent the data

3 Scenarios:

- 1. Wait for sender to send (blocking)
- 2. Instruct the internal buffer for this transfer (on sender, receiver, or elsewhere) and then return from the send call **(nonblocking)**
- 3. Fail

TY IT 24-25 Sem I DBK Computer Algorithms 42/50



Parallel Sum with MPI (Complete code)

```
#include <mpi.h>
                                            #include <stdio.h>
#include <stdlib.h>
                                            #include <unistd.h>
#define n 10
                                            // size of array
int a[] = { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 };
int a2[1000];
                                            // Temporary array for slave process
int main(int argc, char* argv[]) {
    int pid, np, elements per process, n elements recieved; // np -> no. of processes
                                                                                       // pid -> process id
    MPI_Status status;
    MPI_Init(&argc, &argv);
                                           // Creation of parallel processes
    MPI_Comm_rank(MPI_COMM_WORLD, &pid); // find out process ID, how
    MPI_Comm_size(MPI_COMM_WORLD, &np); //many processes were started
    if (pid == 0) {
                                            // master process
        int index, i; elements_per_process = n / np;
        // check if more than 1 processes are run
```

// distributes the portion of array to child processes to calculate their partial sums



Parallel Sum with MPI: Part 1: Intialization

```
#include <mpi.h>
                                     #include <stdio.h>
#include <stdlib.h>
                                     #include <unistd.h>
#define n 10
                                     // size of array
int a[] = { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 };
int a2[1000];
                                     // Temporary array for slave process
int main(int argc, char* argv[]) {
   int pid, np, elements_per_process, n_elements_recieved; // np -> no. of processes and pid ->
process id
// Creation of parallel processes
   MPI_Init(&argc, &argv); MPI_Status status;
   // find out process ID, and how many processes were started
   MPI_Comm_rank(MPI_COMM_WORLD, &pid);
   MPI_Comm_size(MPI_COMM_WORLD, &np);
```

TY IT 24-25 Sem I DBK Computer Algorithms 44/50



Parallel Sum with MPI: Part 2 : Master process : Work distribution

```
// master process to send data and count to slave
   if (pid == 0) { int index, i; elements_per_process = n / np;
   for (i = 1; i < np - 1; i++) { // distributes the portion of array to child processes
                                                                                    index = i *
elements_per_process;
  MPI_Send(&elements_per_process,1, MPI_INT, i, 0, MPI_COMM_WORLD);
  MPI_Send(&a[index],elements_per_process,MPI_INT, i, 0,MPI_COMM_WORLD);}
       for (int sum=0,i = 0; i < elements_per_process; i++)
                                   // master process add its own sub array
           sum += a[i];
                           // collects partial sums from other processes
       int tmp;
       for ( i = 1; i < np; i++)
MPI_Recv(&tmp,1,MPI_INT,MPI_ANY_SOURCE, 0,MPI_COMM_WORLD, &status);
       int sender = status.MPI_SOURCE;
       sum += tmp;
// master process to display result
       printf("Sum of array is: %d\n", sum); // prints the final sum of array }
```

TY IT 24-25 Sem I DBK Computer Algorithms 45/50



Parallel Sum with MPI: Part 3: Slave

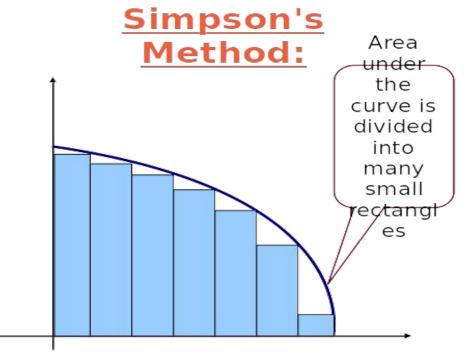
```
else
MPI_Recv(&n_elements_recieved,1, MPI_INT, 0, 0, MPI_COMM_WORLD,&status);
MPI_Recv(&a2, n_elements_recieved, MPI_INT, 0, 0, MPI_COMM_WORLD, &status);
    // calculates its partial sum
    int partial_sum = 0;
    for (int i = 0; i < n_elements_recieved; i++)
    partial_sum += a2[i];
// sends the partial sum to the root process
MPI_Send(&partial_sum, 1, MPI_INT,0, 0, MPI_COMM_WORLD);
MPI Finalize();
    return 0;
```

TY IT 24-25 Sem I DBK Computer Algorithms 46/50



Example: PI Calculation





Area under the curve (Approx.) = Sum of the areas of all rectangles

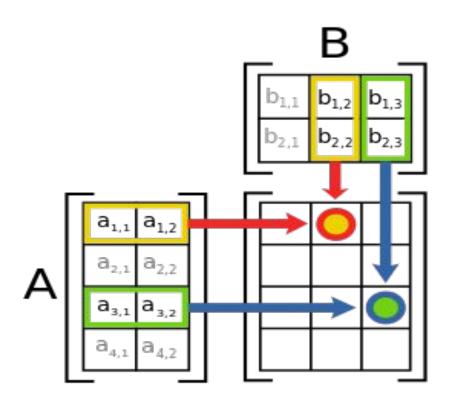
04/26/202 DBK/PP < n

DBK/PP

TY IT 24-25 Sem I DBK Computer Algorithms 47/50



Matrix Multiplication



C=A*B (Compatibility: Rows(A)=Columns(B)



Summary: Parallel Terminology

- Concurrent computing a program is one in which multiple tasks can be in progress at any instant.
- Parallel computing a program is one in which multiple tasks cooperate closely to solve a problem
- Distributed computing a program may need to cooperate with other programs to solve a problem.

TY IT 24-25 Sem I DBK Computer Algorithms 49/50



Thank You

TY IT 24-25 Sem I DBK Computer Algorithms 50/50