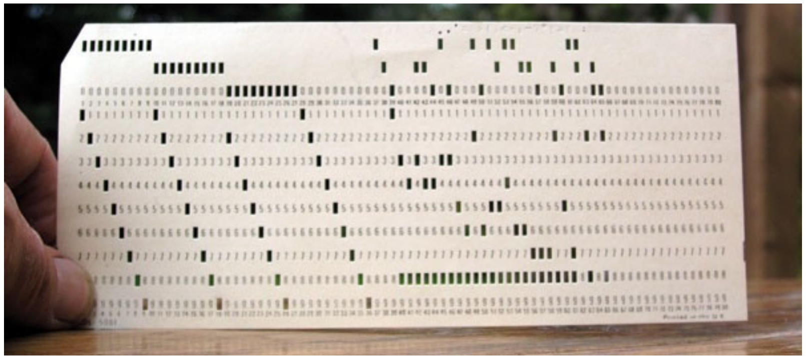
First Generation cont…

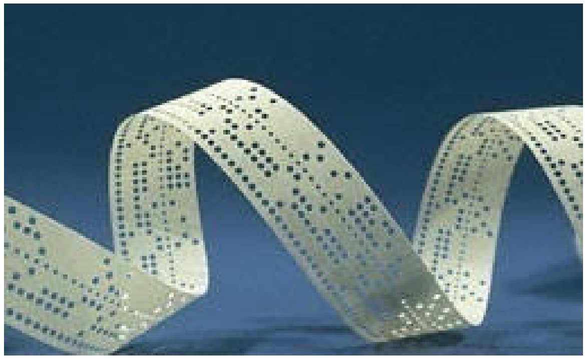
- These computers were basically programmed manually by setting switches, plugging and unplugging cables. You had to interact with the hardware directly.
- Operating systems and Programming languages were unknown (even assembly language was unknown).
- All programming was done in absolute machine language (the lowest-level programming language understood by computers), often by wiring up plugboards.
- The usual mode of operation was for the programmer to sign up for a block of time on the signup sheet on the wall, then come down to the machine room, insert his plugboard into the computer and wait for output.

Plugboard (تخته مدار)

First Generation cont…

- By the early 1950s, the routine had improved somewhat with the introduction of punched cards (also called IBM or Hollerith cards). It was now possible to write programs on cards and read them in instead of using plugboards. The rest of procedure was the same.





Problems of First generation

- These machines were enormous, filling up entire rooms with tens of thousands of vacuum tubes.
- They were very expensive to operate and could be afforded only by very large organizations.
- They used a lot of electricity, generated a lot of heat, which was often the cause of malfunctions.
- Some computers of this generation were:
  - ENIAC
  - EDVAC
  - UNIVAC
  - IBM-701
  - IBM-650

Second Generation (1955-1965); Simple Batch, Spooling

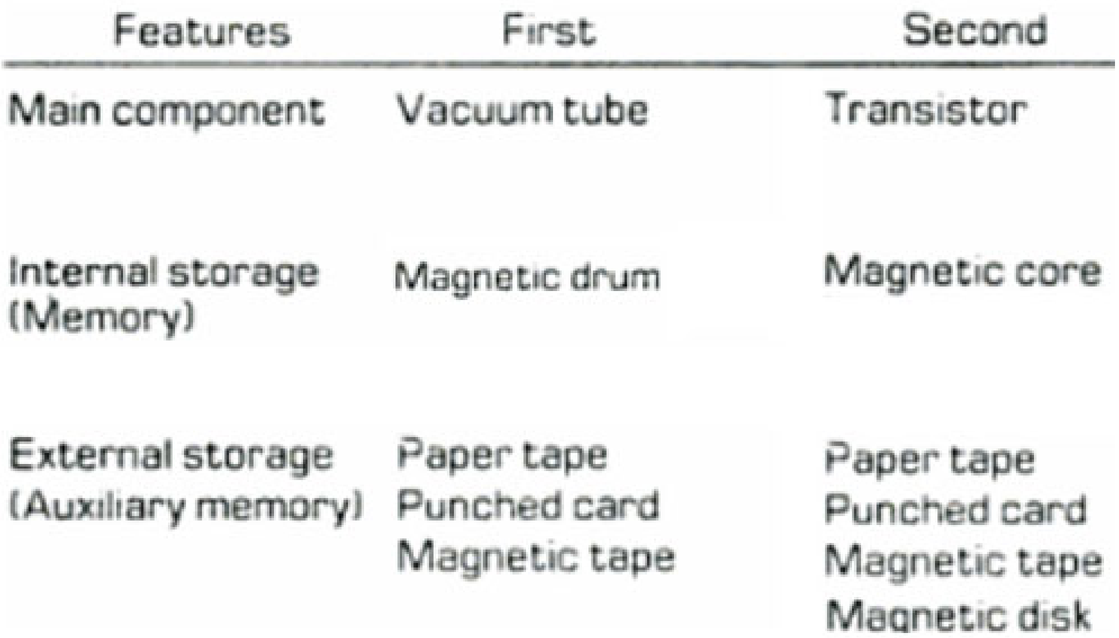
- Using the **transistors** as replacement of vacuum tubes, this generation was cheaper, consumed less power, smaller, more reliable and faster than the first generation machines.
- The Magnetic cores were used as primary memory
- The magnetic tape and magnetic disks as secondary storage devices.

Second Generation cont…

- They still relied on punched cards for input and output.
- It moved from binary machine language to assembly languages, which allowed programmers to specify instructions in words.
- Machine independent languages like COBOL, FORTRAN and ALGOL developed.

Second Generation cont…

- Some magnetic storage devices are:
  - Magnetic Drum
  - Magnetic Core
  - Magnetic Disk
  - Magnetic Tape
  - Floppy Disk

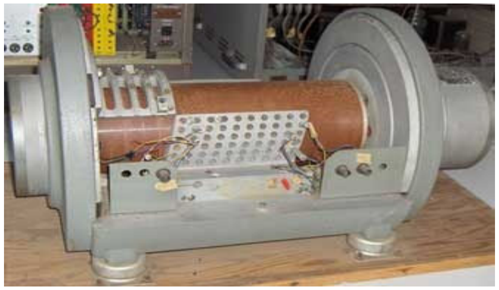


Features	First	Second
Main component	Vacuum tube	Transistor
Internal storage (Memory)	Magnetic drum	Magnetic core
External storage (Auxiliary memory)	Paper tape Punched card Magnetic tape	Paper tape Punched card Magnetic tape Magnetic disk

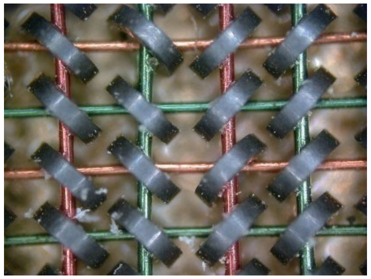
Computer, Audio, Video Magnetic Tapes

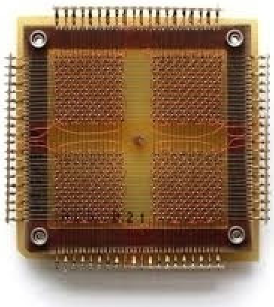
Second Generation cont…

magnetic drum: cylinder with a magnetic surface on which data is stored

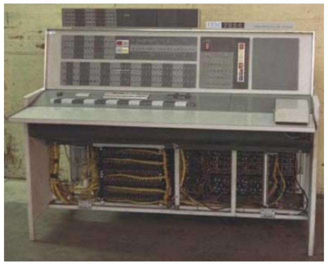


magnetic core: tiny rings of magnetic material string at the intersection of a vertical and horizontal wire





Magnetic Disk



Second Generation cont…

- To run a job (consist of the program, data, and some control card), a programmer would first write the program on paper (in FORTRAN), then punch it on cards. He would then bring the card deck down to the input room and wait until the output was ready.

- An operator would take one of the card decks that had been brought from the input room and read it in. Then he would go over to the printer and take away the output (usually printed papers) and carry it over to the output room, so that the programmer could collect it later.

- OS here called monitor and was so simple and it only move from the finished batch to next new batch (uni-programmed) automatically.

Second Generation cont…

- As we can see, much computer time was wasted while operators put the input tapes or take the output papers while the CPU sat idle for some time.

- Therefore, they looked for ways to reduce the wasted time. The solution generally adopted was the **batch system** (سیستمهای دسته ای).

CPUسریع است ولی دستگاه های I/O کند هستند
وقتی کار جاری برای تکمیل یک عملیات I/Oمنتظر
می‌شود، در این حال CPUبیکار می‌ماند و مجبور است
صبر کند تا عملیات I/Oبه اتمام برسد

Second Generation cont…

- All the programmers leave their programs with the operator. The operator would sort the programs with the same requirement into a group called batch.

- In batch processing (unlike transaction processing) there is **no** interaction (from start to completion) between the user and the job while that job is executing. For example during a database run, we can not open, edit and save it.

- The job is prepared and submitted, and at some later time, the output appears. The delay between job submission and job completion is called turnaround time (زمان برگشت گردش یا زمان پاسخ).

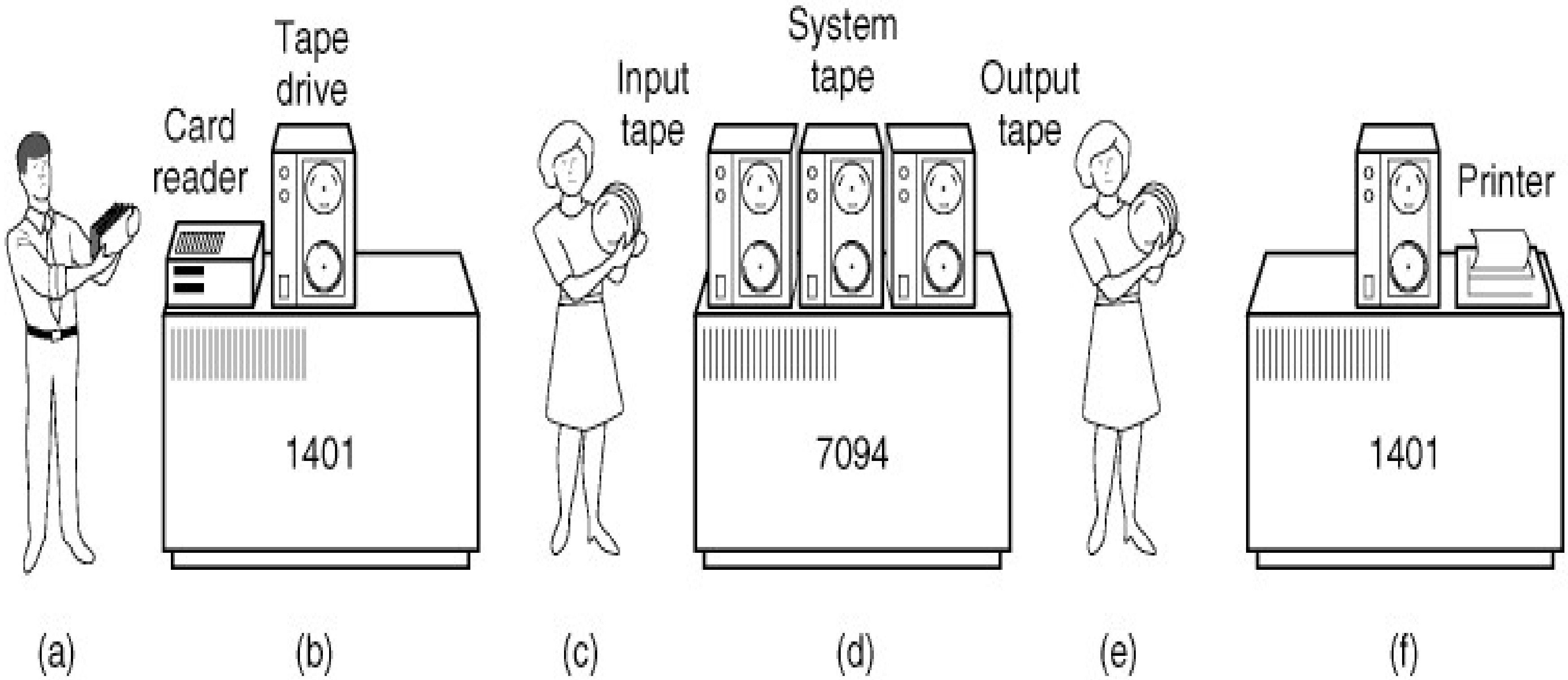
Second Generation cont…

- The suitability of this type of processing is in programs with large computation time with no need of user interaction/involvement. Users are not required to wait while the job is being processed. They can submit their programs to operators and return later to collect the results.

- Some examples of such programs include payroll (ليست حقوق), bill, forecasting, statistical analysis, and large scientific number crunching programs (quite simple but takes a long time).

Second Generation cont…

- Since it is faster to read from a magnetic tape than from a deck of cards, it became common for computer centers to have two computers:
  - Less powerful computer IBM 1401
  - Powerful main computer IBM 7094

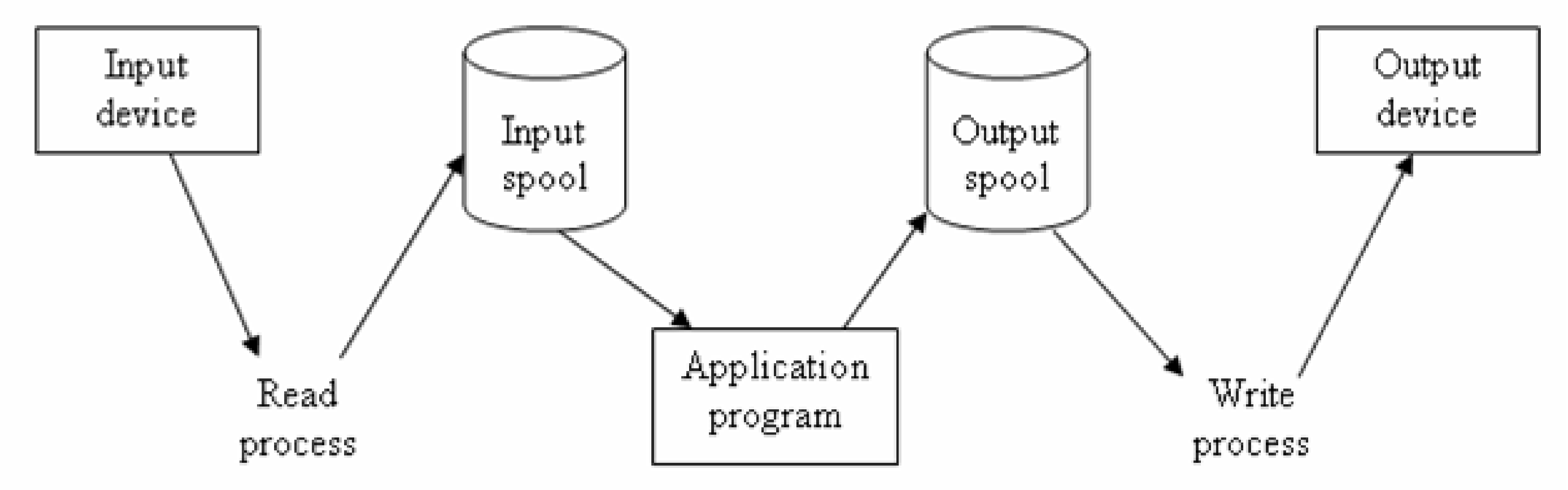


Second Generation cont…

- (a) Programmers bring punched cards to the IBM 1401 computer system. The 1401 was very good at reading cards, copying tapes, and printing output, but not at all good at numerical calculations.
- (b) The1401 reads a batch of jobs onto tape.
- (c) The tape then rewound and brought into the machine room, where it was mounted on a tape drive on IBM 7094.
- (d) The operator then loaded a special program (the ancestor of today's operating system), which read the first job from tape and ran it. It runs a job and writes the output on another tape (instead of being printed). After finishing each job, it automatically runs the next job.
- (e) After finishing the whole batch, the operator removed the input and output tapes, replaced the input tape with the next batch, and carries the output tape to the IBM 1401 for printing off-line (i.e., not connected to the main computer).
- (f) The 1401 prints output. This is called Off-line Spooling.

Spool

- Since CPU is much faster than I/O devices, so CPU puts data for I/O devices in a buffer which can be directly accessed by I/O devices without requiring CPU involvement.
- To spool a job means to store it on a buffer (tape) so that it can be printed or processed by another program later at a more convenient time when necessary. Spooling helps to solve the problems of speed mismatch among different devices. For example CPU operates at a very high speed than the printer.



Spool cont…

- In Spooling all devices (CPU and I/O) can work at a time and no one is idle.

- While the output in the last room is printed, in the first room the next input tape is getting ready and in the second room, the OS is running the job. So all devices are working together which is called spool.

- Since CPU is much faster than I/O devices, the spooling is possible using buffer.

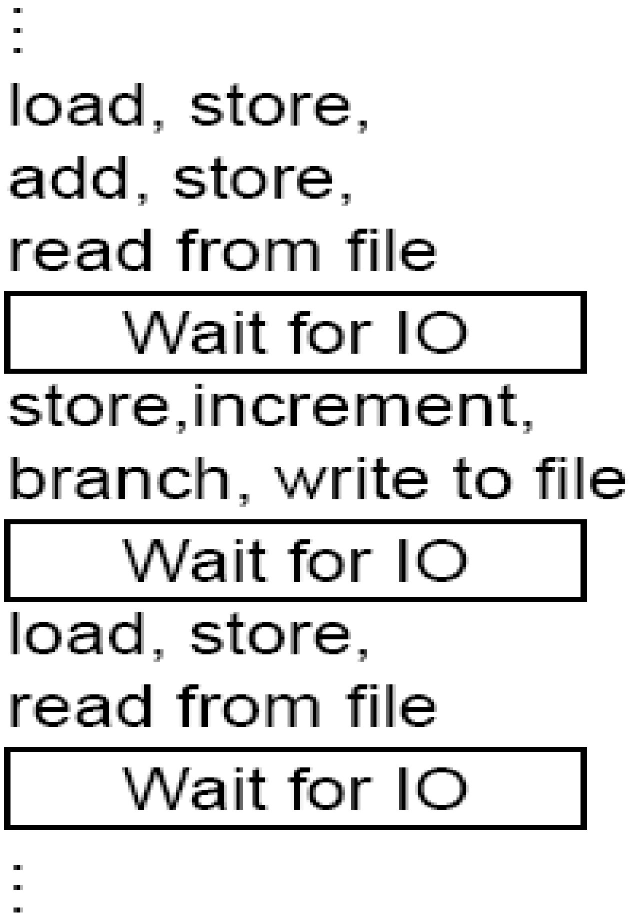
Spool= simultaneous peripheral operations

Offline spool

- When the I/O devices are not directly connected to the computer, it is called offline spooling.

Second Generation cont…

- CPU burst means CPU’s operations on the process.
- IO Burst means interaction of the process with the I/O operations.



⋮

load, store,
add, store,
read from file

Wait for IO

store,increment,
branch, write to file

Wait for IO

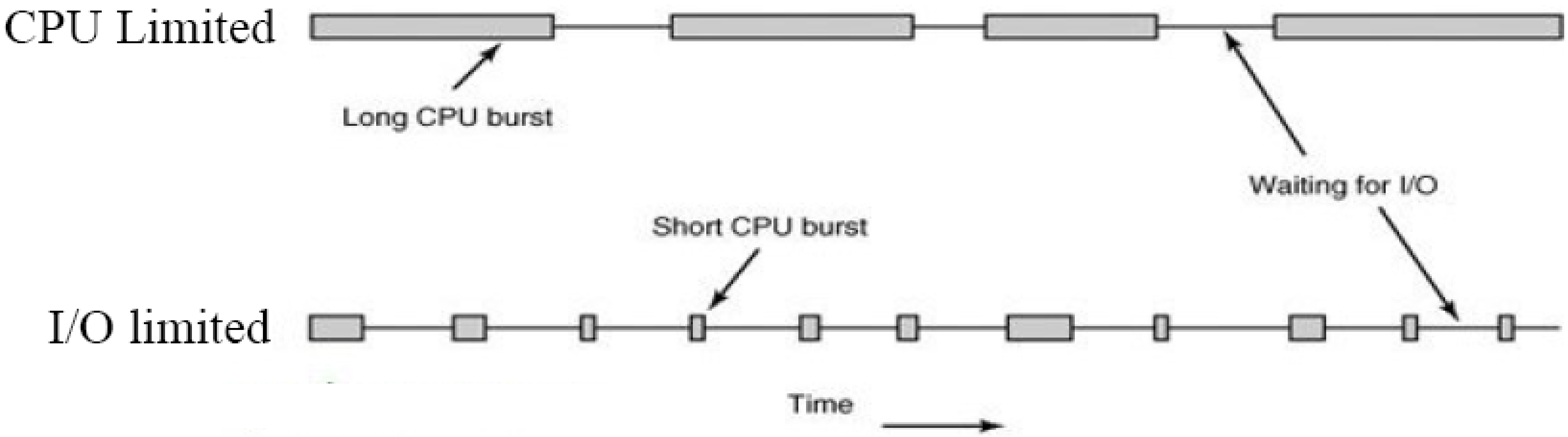
load, store,
read from file

Wait for IO

⋮

Second Generation cont…

- In general we can divide programs into two groups:
  - CPU bound/limited (تنگنای محاسباتی): programs that perform lots of computation and do little IO. Tend to have a few long CPU bursts.
  - IO Bound/limited: programs that perform lots of IO operations. Each IO operation is followed by a short CPU burst to process the IO, then more IO happens.



Problems of second generation

- Problem1: In the batch processing, the system first prepares a batch and after that it will execute all the jobs in that batch. But the main problem is that if a job requires an input or output operation, then it is not possible. The reason is that the jobs were generally submitted on punched cards and magnetic tapes and users were not present to **interact** with their jobs were run.

  With CPU-bound jobs, IO is infrequent, so this waste time is not significant. However, with IO-bound data processing jobs, the IO wait time is significant.

Problems of second generation cont…

- Problem2: time wasted when CPU remains idle during:
  - preparing the batch on the 1401
  - replacing the input tape with the next batch on the 7094

  This wasted time caused poor utilization of the CPU resulting in high **turnaround times** hence low throughput (Turnaround time is a time between submitting a job and receiving the output).

- Problem3: developing and maintaining two completely different product lines (1401 and 7094) was expensive for the manufacturers. In addition, many new computer customers initially needed a small machine.

Third Generation (1965-1980); IC, multiprogramming, online spooling, timesharing

- The third generation computer is marked by the use of Integrated Circuits (IC) in place of transistors. A single IC has many transistors, resistors and capacitors along with the associated circuitry. The IC was invented by Jack Kilby.
- This development made computers smaller in size, reliable and efficient.
- Instead of punched cards and printouts, users interacted with third generation computers through keyboards and monitors and interfaced with an operating system.
- Computers for the first time became accessible to a mass audience because they were smaller and cheaper than their predecessors.

Commodore PET, 1977

Third Generation cont…

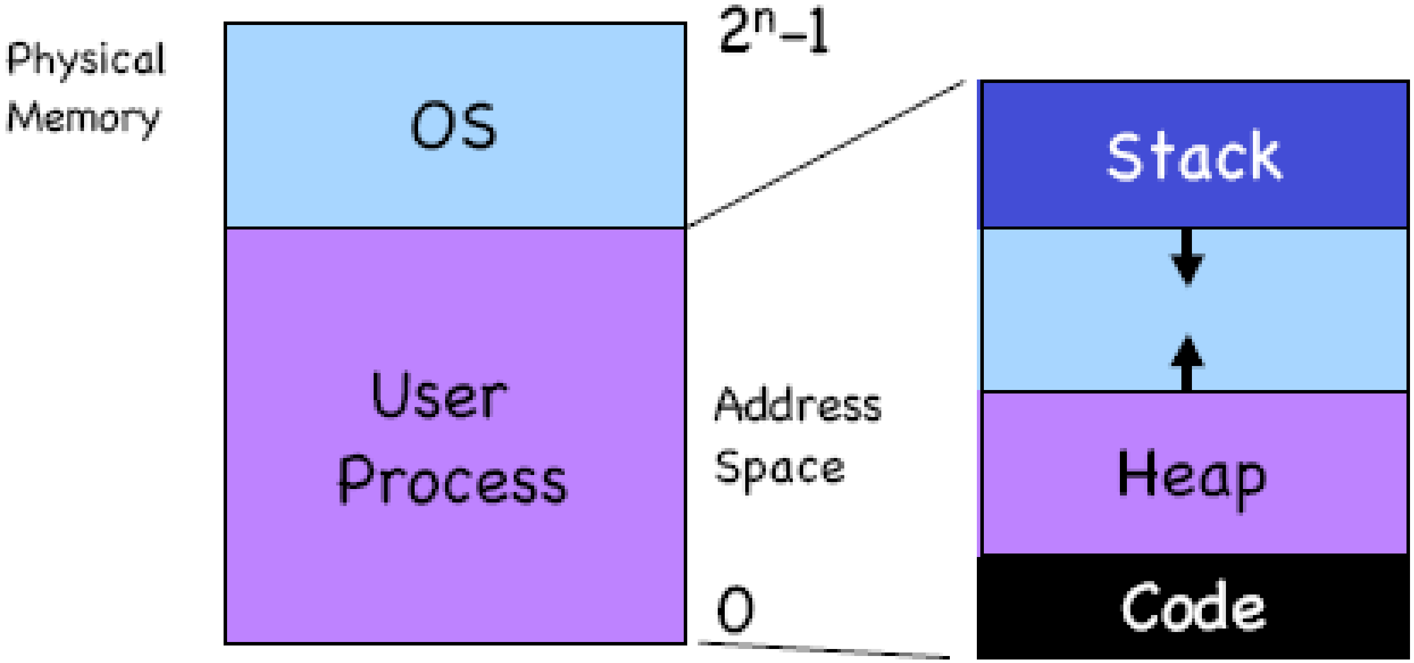
- IBM attempted to solve the problem3 by introducing the 360 series which does both kinds of jobs (IO-bound and CPU-bound)
- For the other two problems something had to be done to avoid having the (expensive) CPU be idle so much. Solution to these problems:
  - Batch Multiprogramming /cooperative multitasking/non-preemptive multitasking/multiprogramming without timesharing. Examples are Contiki, TinyOS
  - Online spooling/spooling
  - Time-sharing Multiprogramming/preemptive multitasking /multitasking (اشتراک زمانی). Examples are: RIOT, FreeRTOS,

Preemptive means Prevent →

preemptive: پیشگیرانه-قابل پس گرفتن
non-preemptive: غیر قابل پس گرفتن

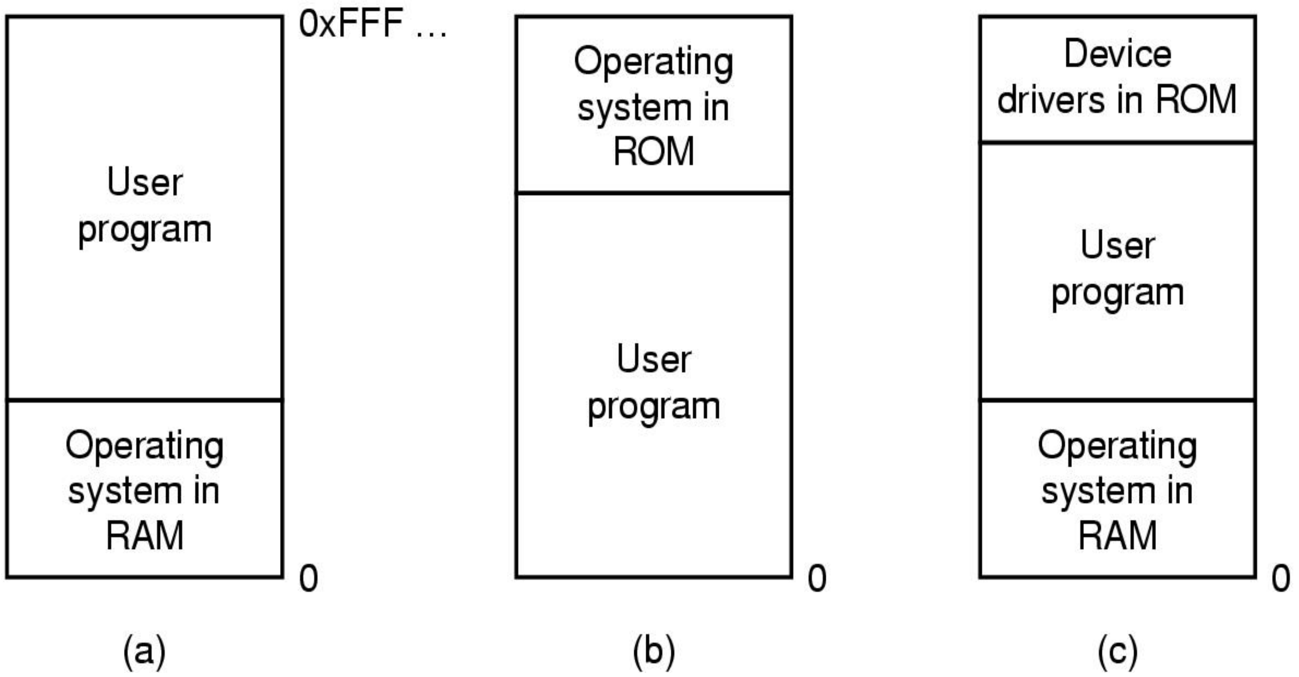
Uni-Programmed OS

- Uni-Programmed systems run just one program at a time, sharing the memory between that program and the operating system.
- Typical example: DOS



Uni-Programmed OS cont…

- There are three variations:
  - The OS is at the bottom of RAM, Fig.(a)
  - The OS is in the ROM, Fig.(b)
  - The device drivers is in the ROM and the rest of the system in RAM down below, Fig.(c)



Uni-Programmed OS cont…

- Disadvantages of Uni-Programmed OS are:
  - Only one program runs at a time
  - Process may destroy OS
  - Not cost effective. When a single process is doing I/O, the CPU is inactive. Multiprogramming makes more effective use of the CPU.

- Uni-programming does not support time-sharing and provides poor CPU utilizations when the workload includes many I/O bound processes.

- Then multiprogramming came. They support multiple batch jobs (batch multiprogramming) or multiple interactive jobs (time-sharing) in memory at once thus need memory management.

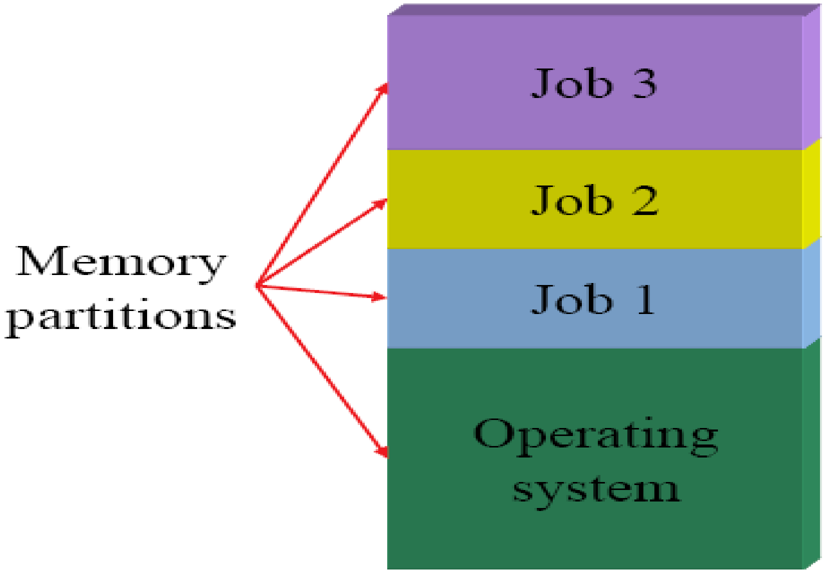
Multiprogramming

- Multiprogramming means the OS keeps multiple programs in the main memory at the same time and the CPU switches among them to run them.
- **Multiprogramming = OS + several jobs in memory (apparently) running at the same time.**
-

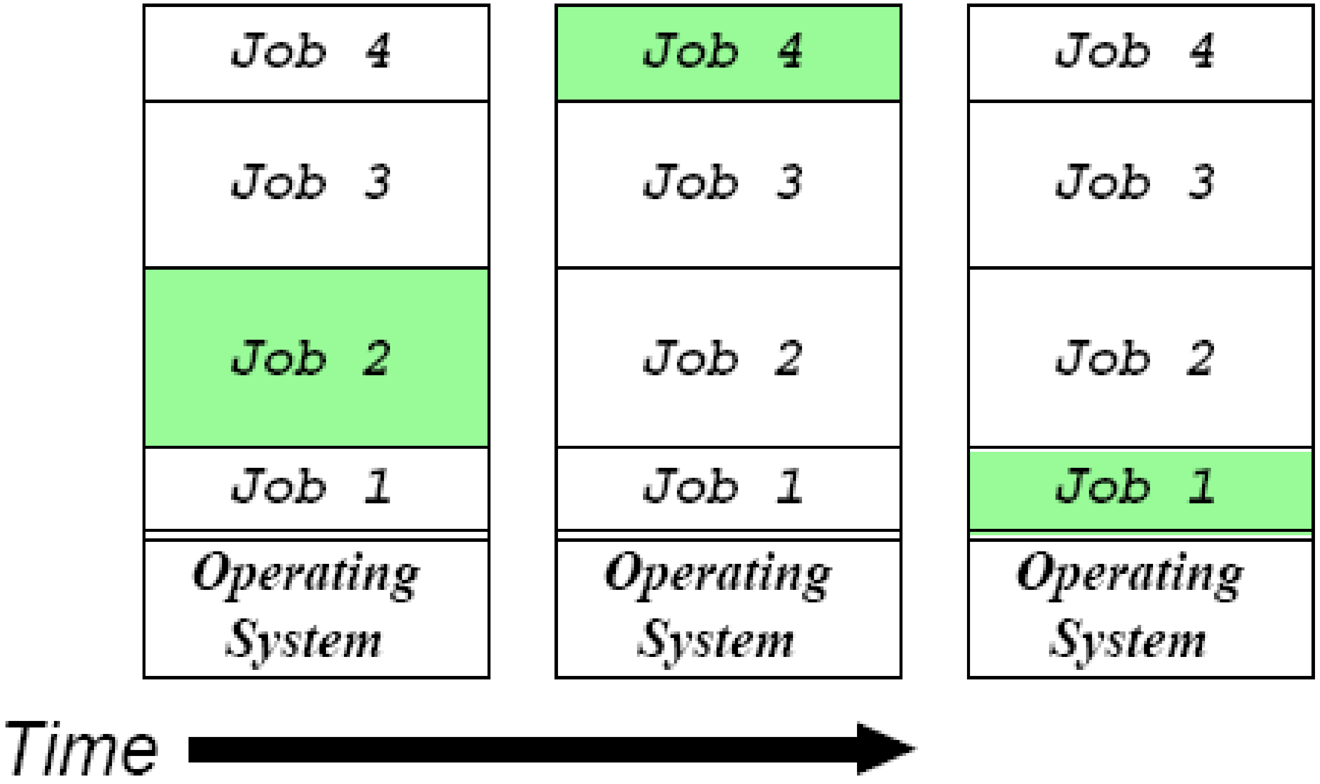
Batch Multiprogramming

- Memory was divided into several partitions and each job was given a partition in memory. A job would then be given the CPU for execution. When that job needs an I/O operation and releases CPU, the CPU switches to another job that is kept in memory (the CPU is allocated to a job only if the job is in memory) until it also needed the services on an I/O device. Therefore, the CPU is always busy.

- For example a multiprogramming system with three jobs in memory is shown as follow:



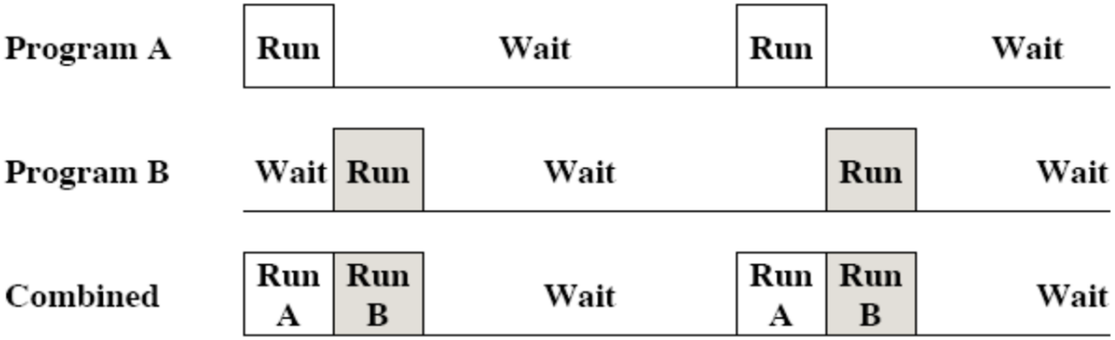
Batch Multiprogramming cont…

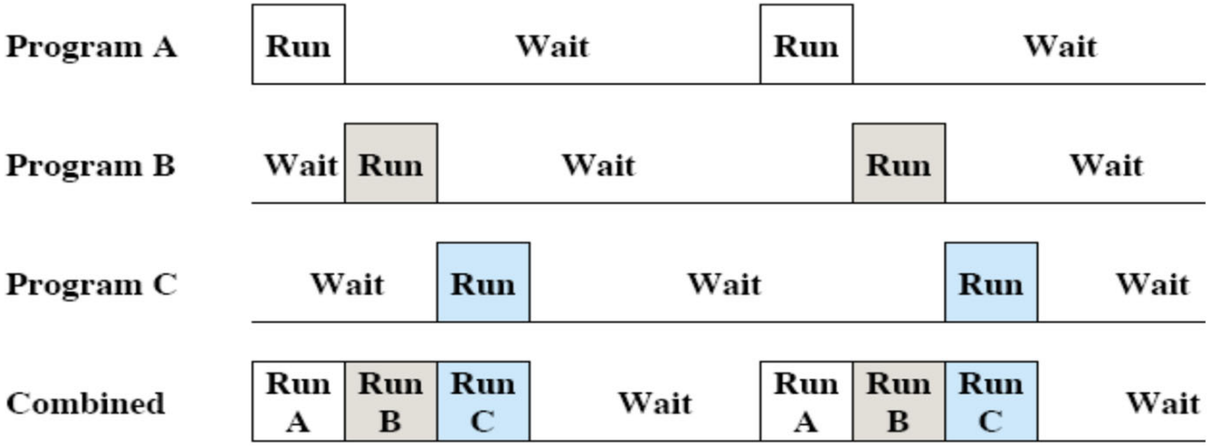


Multiprogramming makes sure that the CPU always has something to execute, thus increases the CPU utilization/usage.

Batch Multiprogramming cont…

- The act of reassigning a CPU from one program to another one is called a context switch (context means state). When context switches occur frequently enough the illusion of parallelism is achieved.





Batch Multiprogramming cont…

- Since, several jobs are kept in main memory at the same time, and the CPU is multiplexed among them, this requires memory management and protection and necessary hardware and software support:
  - Resource Allocation: OS manages different types of resources such as main memory, CPU cycles, file storage, etc. If there are more than one jobs running at the same time, then resources must be allocated to each of them. Resource Allocation means to decide which process gets which resource when.

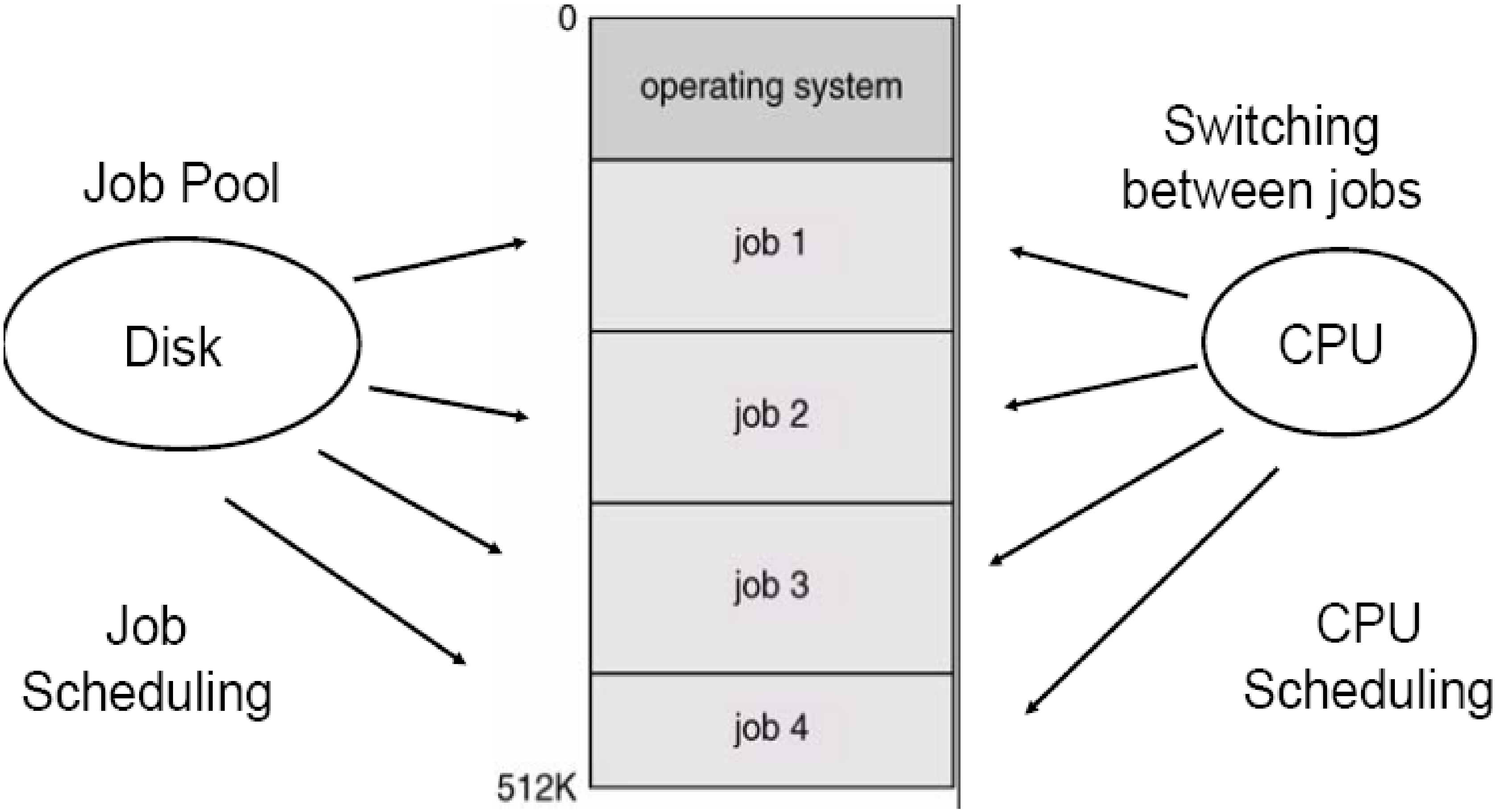
Batch Multiprogramming cont…

- Job Scheduler: If several jobs are ready to bring from disk to memory and there is not enough room for all of them, then the system chooses jobs among them (Job Scheduling).

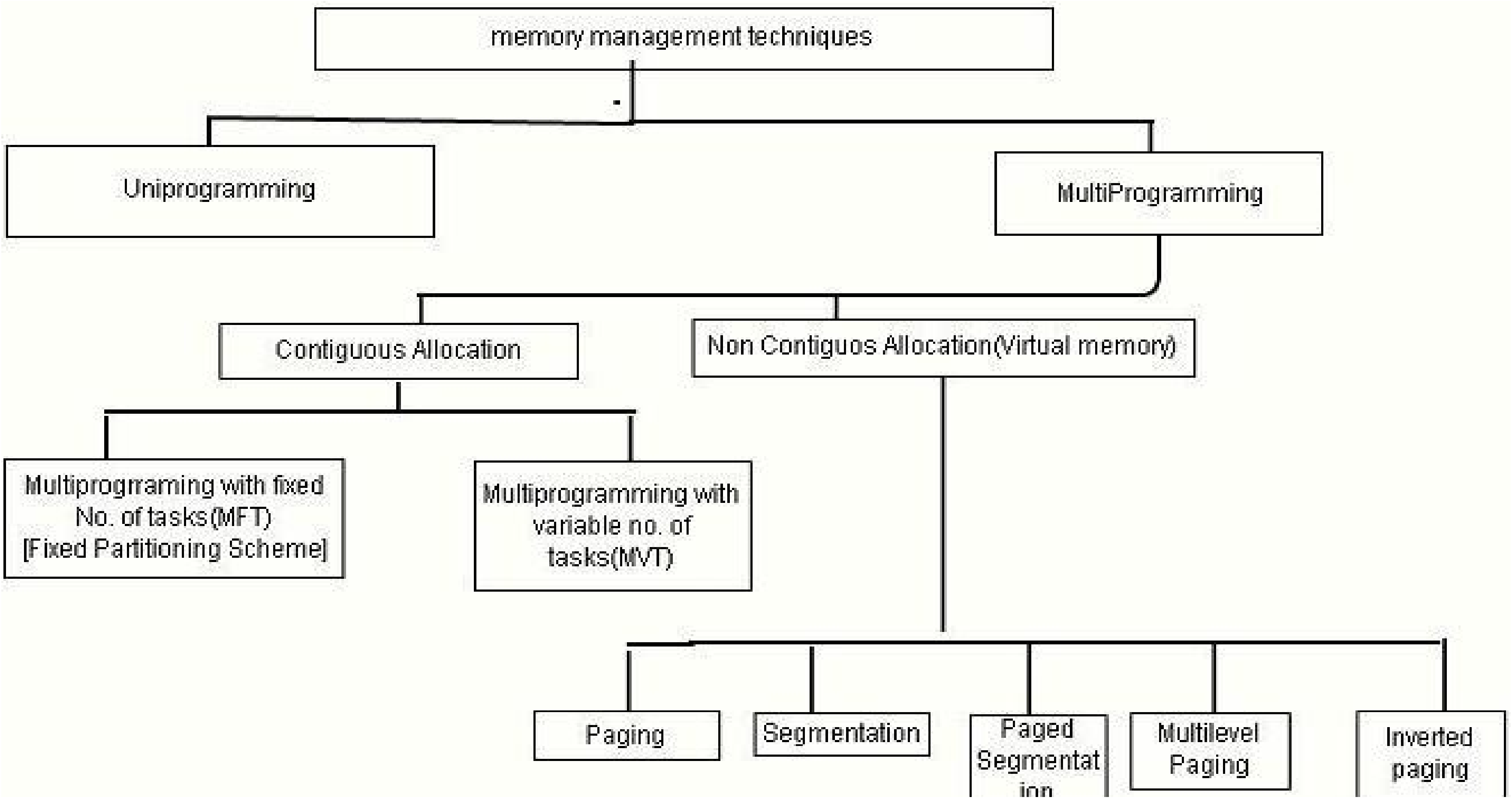
- CPU scheduler: If several jobs are ready to run in memory, the system must choose among them (CPU Scheduling).

- Memory manager: Having several jobs in memory at the same time requires memory management to allocate the memory to them.

Batch Multiprogramming cont…



Batch Multiprogramming cont…


memory management techniques
Uniprogramming
MultiProgramming
Contiguous Allocation
Non Contiguos Allocation(Virtual memory)
Multiprograming with fixed No. of tasks(MFT) [Fixed Partitioning Scheme]
Multiprogramming with variable no. of tasks(MVT)
Paging
Segmentation
Paged Segmentation
Multilevel Paging
Inverted paging

Batch Multiprogramming cont…

- In Multiprogramming Batch System the OS never initiates a context switch from the running program to another program.

- In fact since the running program is controlling the CPU, only this running program can offer control to another program(either it voluntarily releases the CPU or when it needs for an I/O operation).

- That is why it is called cooperative because all other programs must cooperate for it to work.

Advantage

- It improves CPU utilization because it keeps the CPU busy as long as there are processes ready to execute.

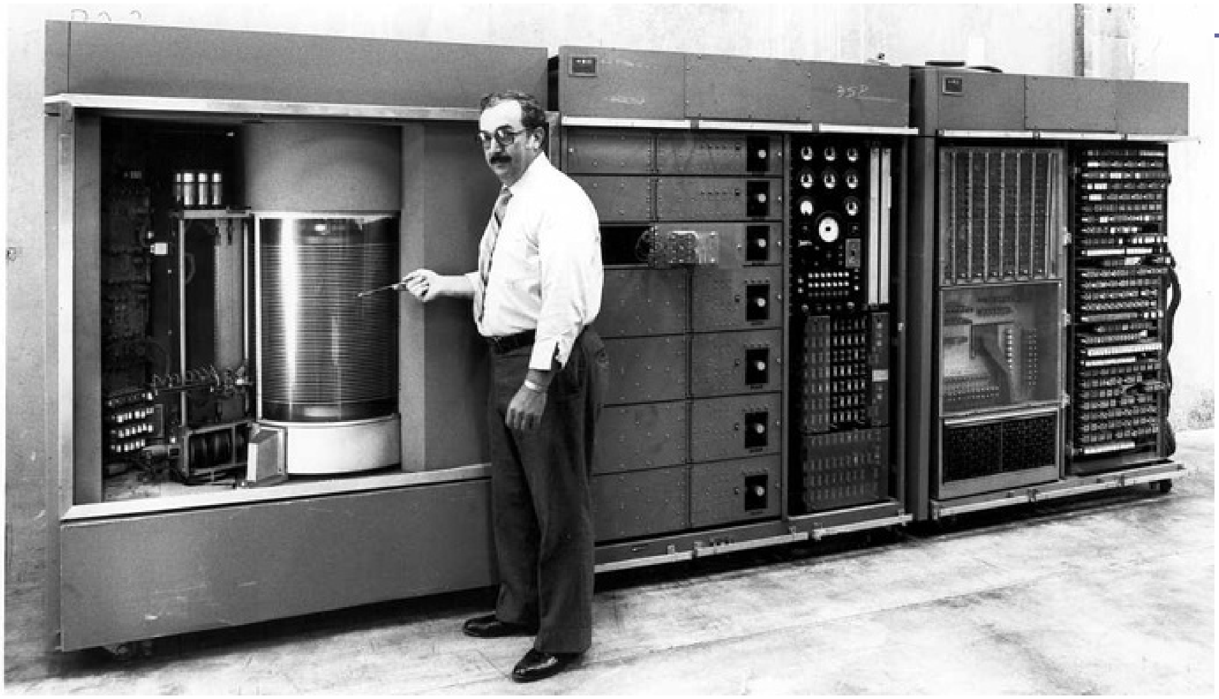
Disadvantages

- If there are N ready processes and all of those are highly CPU bound, one program might wait all the other N-1 ones to complete before executing.
- In order for Multiprogramming Batch System to function properly, the OS must be able to load multiple programs into separate areas of the main memory and provide the required protection to avoid the chance of one process being modified by another one.
- Other problems when having multiple programs in memory is fragmentation as programs enter or leave the main memory.
- A large programs may not fit at once in memory which can be solved by using paging and virtual memory.

Online spooling

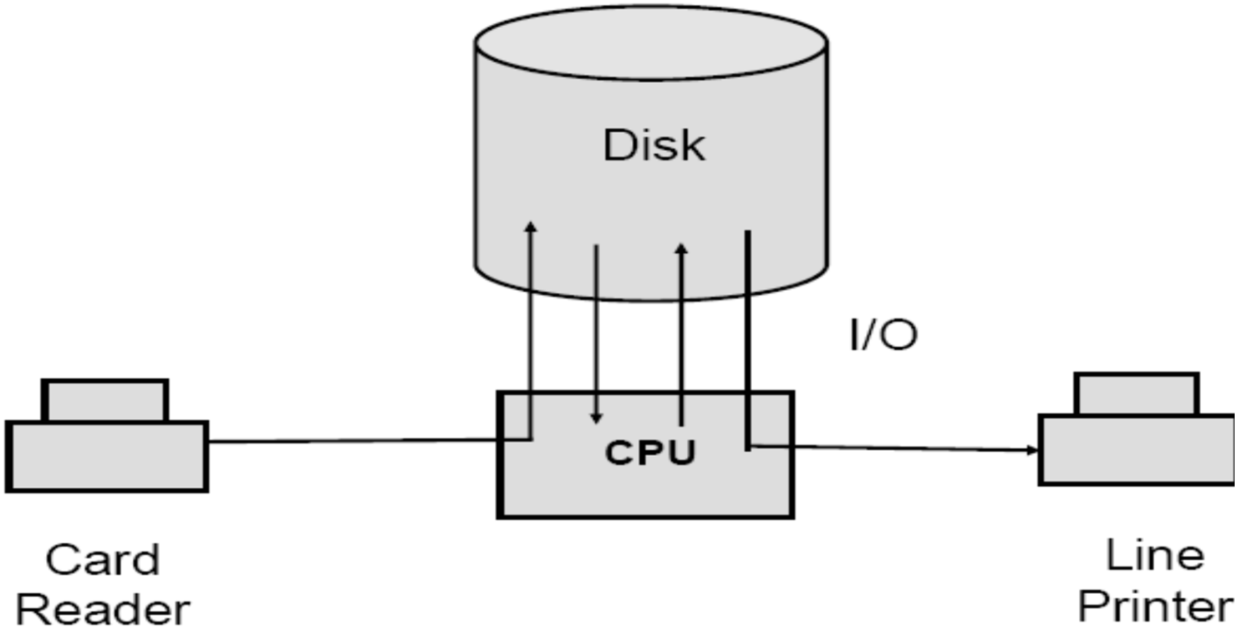
- With invention of Hard Disks, the spooling stores job on a buffer (which is Disk instead of tape).
- In online spooling, disks are directly connected to computer (online). In this case, output of data processing is available immediately.
- Disks are faster than tapes.

First hard disk (IBM, 1956, 1000Kg, 5MB≈ 64,000 punched cards)



Online spooling cont…

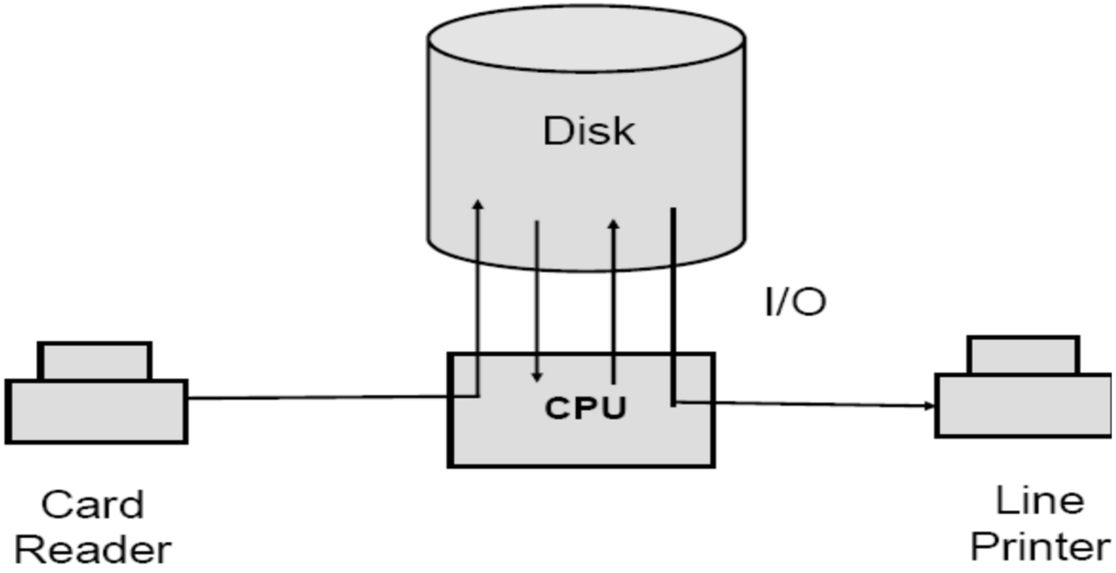
- In spooling, a high-speed device like a disk is placed between a running program and a low-speed IO device.
  - Input Spooling: is the technique of reading jobs (for example from cards) onto the disk so that when the currently executing processes are finished, quickly there will be work waiting for the CPU.



Online spooling cont…

- Output Spooling: instead of writing directly to a printer, the outputs are written to the disk. This way, while other programs can be initiated sooner, when the printer becomes available, the outputs go to the buffer of printer to print.

- With spooling, the 1401 was no longer needed, and tapes disappeared.



Spooling Today

- The spoolsv.exe file is described as the Spooler SubSystem App or Windows Print Spooler Service and is the main component of the printing interfaces. The spoolsv.exe file is initialized when the computer starts, and it runs in the background until the computer is turned off.

  The process spoolsv.exe transfers the data in a buffer. If the printer needs the data, it will retrieve it from the buffer. While the spoolsv.exe file is storing the data in the buffer, the user can run other operations.

- The spoolsv.exe process is also responsible for queuing printing tasks. Through this function, the user does not need to wait for each printing task to be completed one after the other.

Time-sharing Multiprogramming

- A Batch Multiprogramming system improves utilization but does not support interaction: the task itself decides when to give up the CPU (either when it completed or needs I/O). No other program can run until the non-preemptive program has given up control of the CPU.

- Time-sharing extends Multiprogramming to support multiple interactive jobs: the OS can take control of the CPU without the task's cooperation.

Time-sharing Multiprogramming cont…

- The reason why it is called time-sharing is that the processor's time is shared among multiple programs simultaneously.

- Time-sharing is a variant of multiprogramming technique.

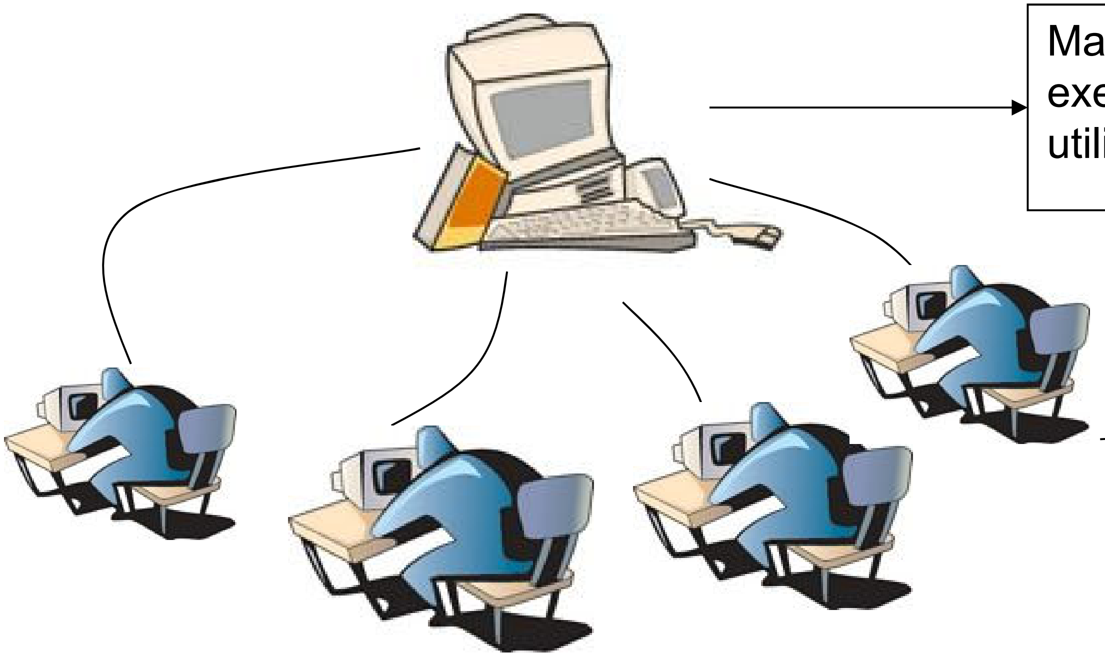
- Time-sharing system can run several programs at the same time, so it is also a multiprogramming system. But not all multiprogramming OS are time-sharing systems.

Time-sharing Multiprogramming cont…

- Time sharing refers to the allocation of computer resources in time slots to several programs/users simultaneously.

- For example, in a mainframe computer that has many users logged on to it, each user uses the resources of the mainframe (memory, CPU). Thus, the users feel that they are exclusive user of the CPU while they are not.

Time-sharing Multiprogramming cont…



Main computer; having a CPU executing processes by utilization of the OS, (e.g. UNIX).

Terminals are connected to the main computer and used for input and output. No processing is made. They do not have CPUs.

Time-sharing in mainframe means multiple users have terminals (not computers) connected to a *main computer* and execute their task in the main computer.

How Time-sharing works

- Timesharing system allows many users to share computer resources. Each user is allocated a **fixed tiny slice of time** (e.g. two milliseconds). The computer performs whatever operations it can for that user until the time-slice ends then utilizes the next allocated time for the other users. The time allowed is extremely small and the users **feel** that they each have their own CPU.

- In case program does not complete within time slot, then extra time slot will not be given to it. For example, if a program with a time slot of 5 seconds is not completed in 5 second and requires 1 extra second then it will be executed in next execution cycle but time slot will not be extended (unlike real time sharing OS in which time slot can be extended).

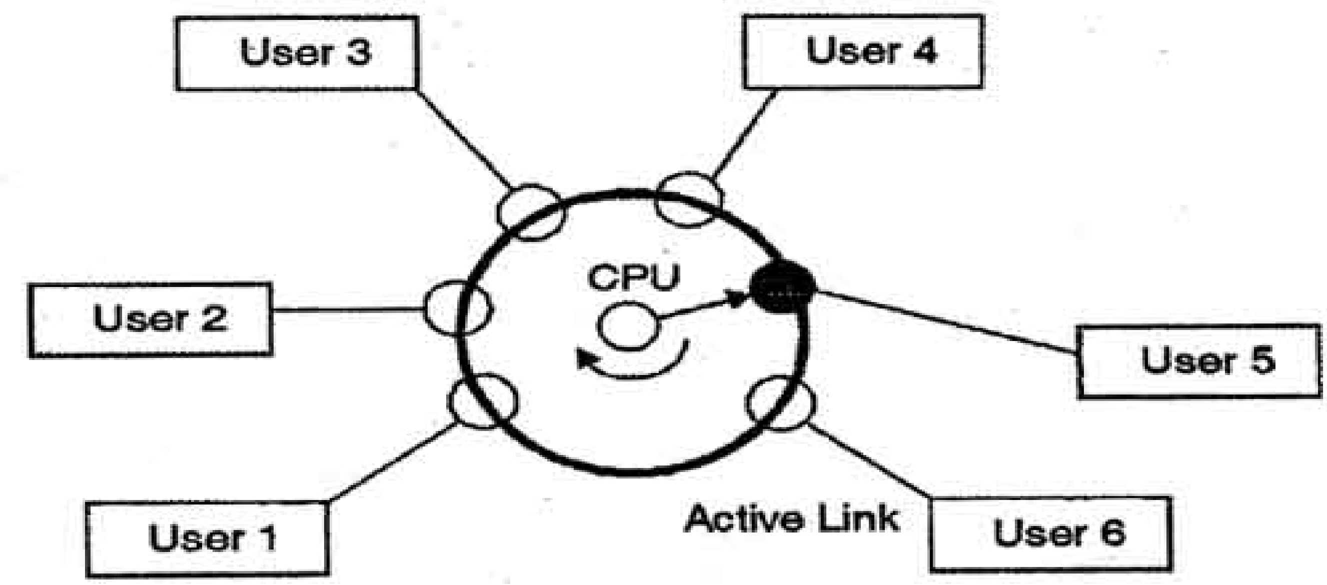
How Time-sharing works cont…

- Time-sharing works because of the difference between the few milliseconds (at least) between a user’s keystrokes and great difference in the speed of the terminal and the computer.

- Since each action or command in a time-shared system take a very small fraction of time, only a little CPU time is needed for each user. As the CPU switches rapidly from one user to another user, each user is given impression that he has his own computer, while it is actually one computer shared among many users.

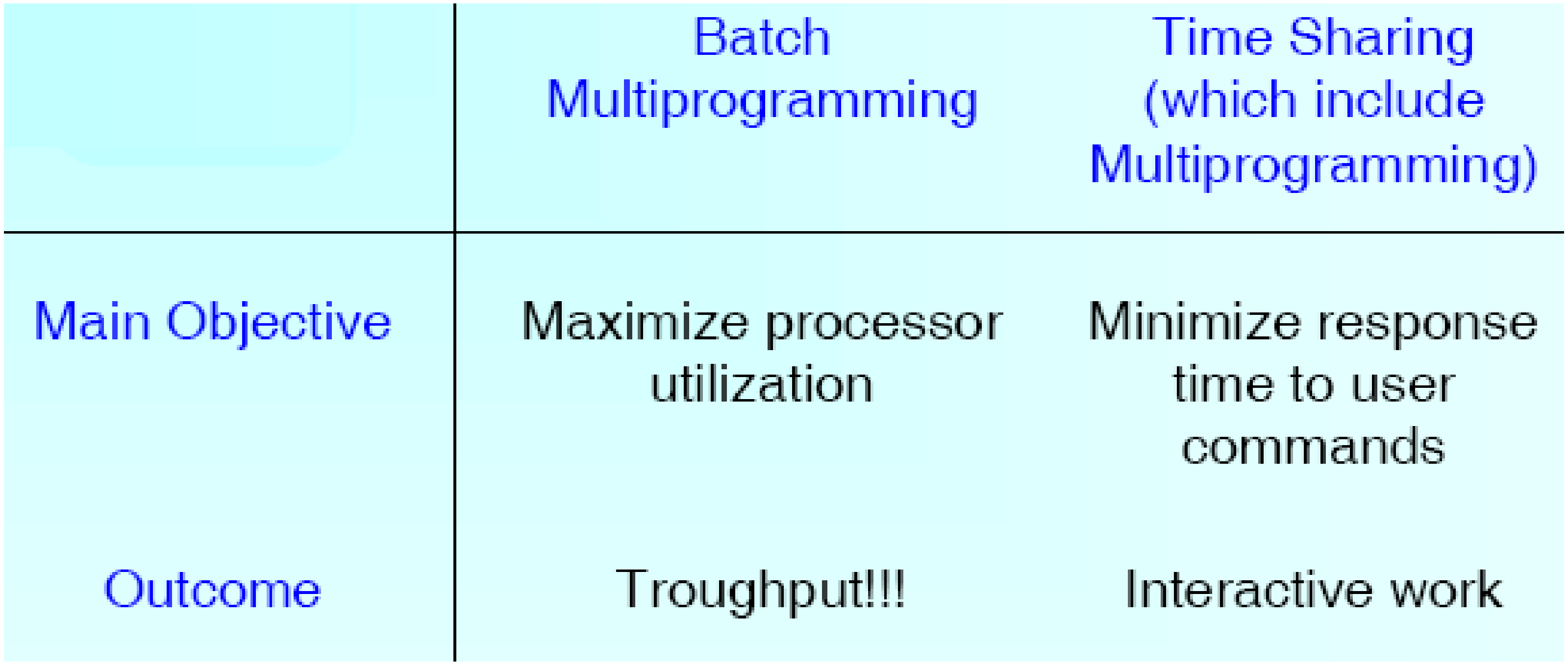
Time-sharing example

- In below figure the user6 is active, user5 is in **ready** status, and other users are in **waiting** state. As soon as the time slice of user6 is completed, the control moves on to the next ready user i.e. user5.



Preemptive multitasking operating systems include Cloud Computing, Linux and other Unix-like systems, Microsoft Windows NT/2000/XP, Mac OS X and OS/2 (IBM)

Comparison



	Batch Multiprogramming	Time Sharing (which include Multiprogramming)
Main Objective	Maximize processor utilization	Minimize response time to user commands
Outcome	Troughput!!!	Interactive work

The main differences between multiprogramming and multitasking is:

- In multiprogramming, a user cannot interact (everything is decided by OS, like picking the next program etc...)
- where as in multitasking, a user can interact with the system (you can type a letter, while the other task of printing is done)"

Fourth Generation (1980-present); PC

- The microprocessor brought the fourth generation of computers, as thousands of IC were built onto a single **silicon chip**.
- The fourth generation of computers is marked around 1980 by the use of Very Large Scale Integration (VLSI) circuits. VLSI circuits having about 5000 transistors and other circuit elements and their associated circuits on a single chip made it possible to have microcomputers of fourth generation. Fourth Generation computers became more powerful, compact, reliable, and affordable. As a result, it gave rise to personal computer (PC) revolution.

Fifth Generation

- In the fifth generation, the VLSI technology became ULSI (Ultra Large Scale Integration) technology, resulting in the production of microprocessor chips having **ten million** electronic components.
- This generation is based on parallel processing hardware and AI (Artificial Intelligence) software.

Other types of operating system

- Real-Time Operating System (RTOS)
- Network Operating System (NOS)
- Parallel processing
- Multithreading Operating System