Central Processing Unit (CPU)

1. CPU performs the data processing operations in a computer.
2. Three major parts (Fig. 8-1)

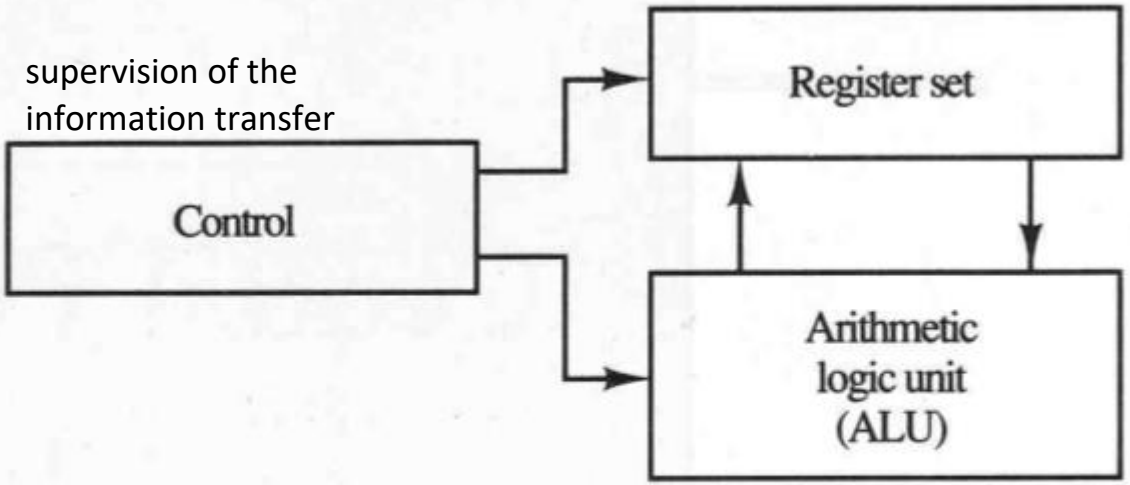


Figure 8-1 Major components of CPU.

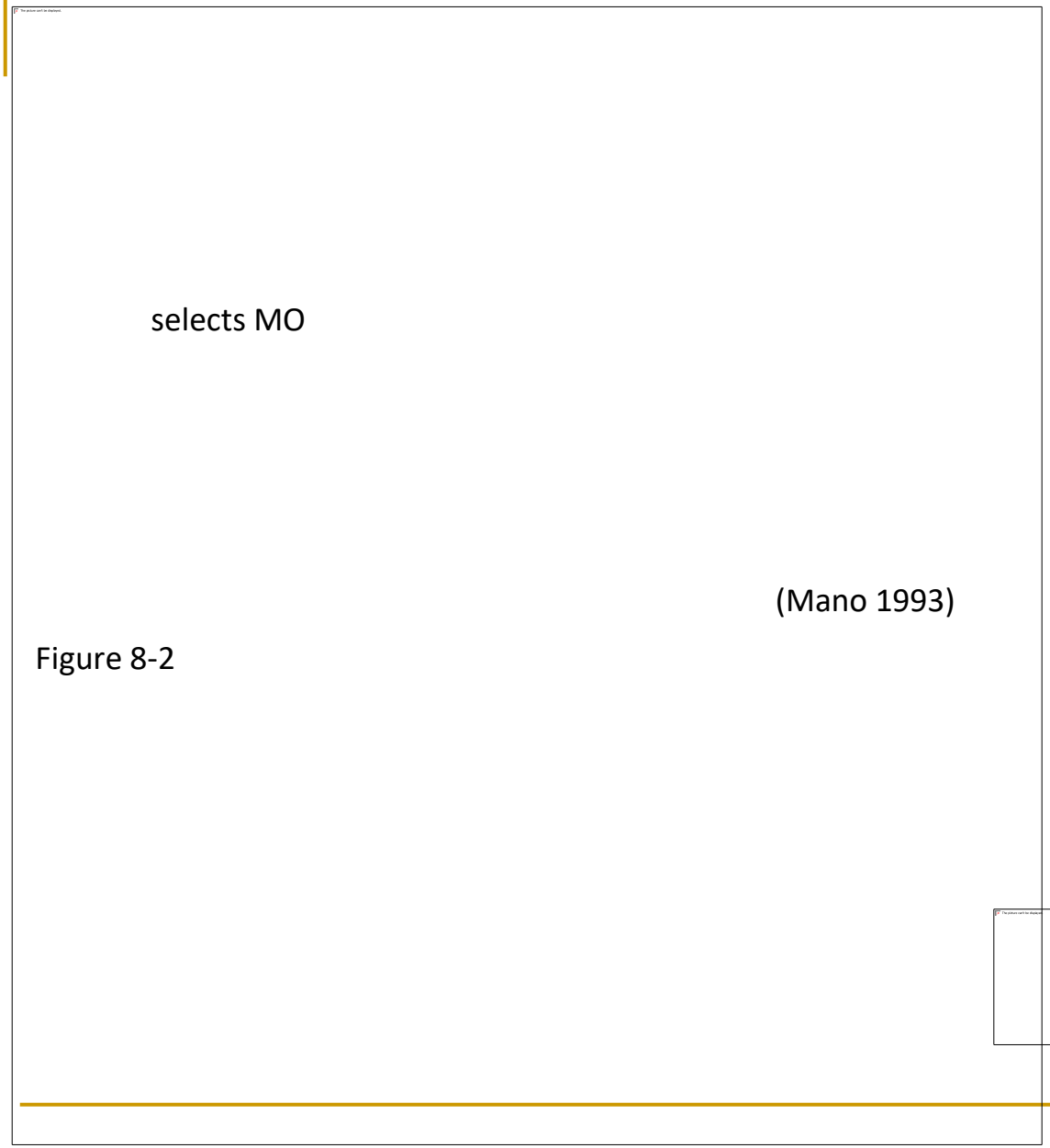
data store

1. Instruction set provides the specification for the design of the CPU.
2. The design of CPU involves choosing the hardware for implementing the machine instructions.
3. A programmer using assembly language has to be aware of the register set, the memory structure, the type of data supported by the instructions, and the function of each instruction.

General Register Organization

1. Storing pointers, counters, return addresses, temporary results, and partial products into the memory is not efficient: referring to memory locations is time consuming: it is more efficient to use registers.
2. If CPU has a large number of registers, a common bus is used to connect the registers.
3. Arithmetic logic unit (ALU) is used to perform various microoperations.
4. A bus organization of 7 CPU registers is shown in Fig. 8-2.

input to ALU (A bus and B bus)

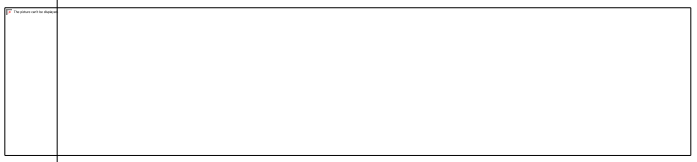


control word

selects MO

(Mano 1993)

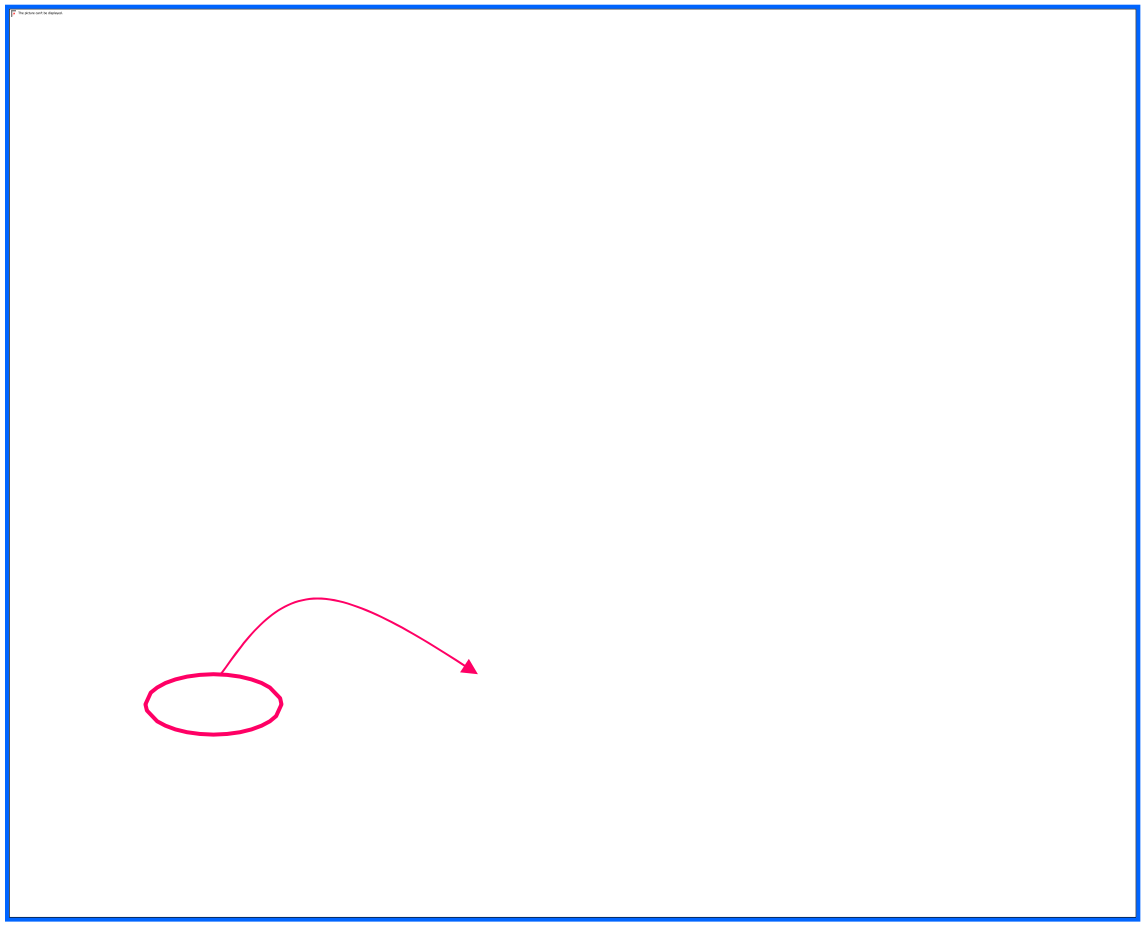
Figure 8-2



- Control unit operates the CPU bus system. For example:

1. SELA is used to place R2 into bus A.
2. SELB is used to place R3 into bus B.
3. OPR selects the arithmetic addition.
4. SELD is used to transfer the result into R1.

4. The buses are implemented with multiplexers or 3-state gates.
5. The state of 14 binary selection inputs specifies a control word.
6. The 14-bit control word (when applied to the selection inputs specify a microoperation).
7. The encoding of register selections in specified in Table 8-1.



Input = external input

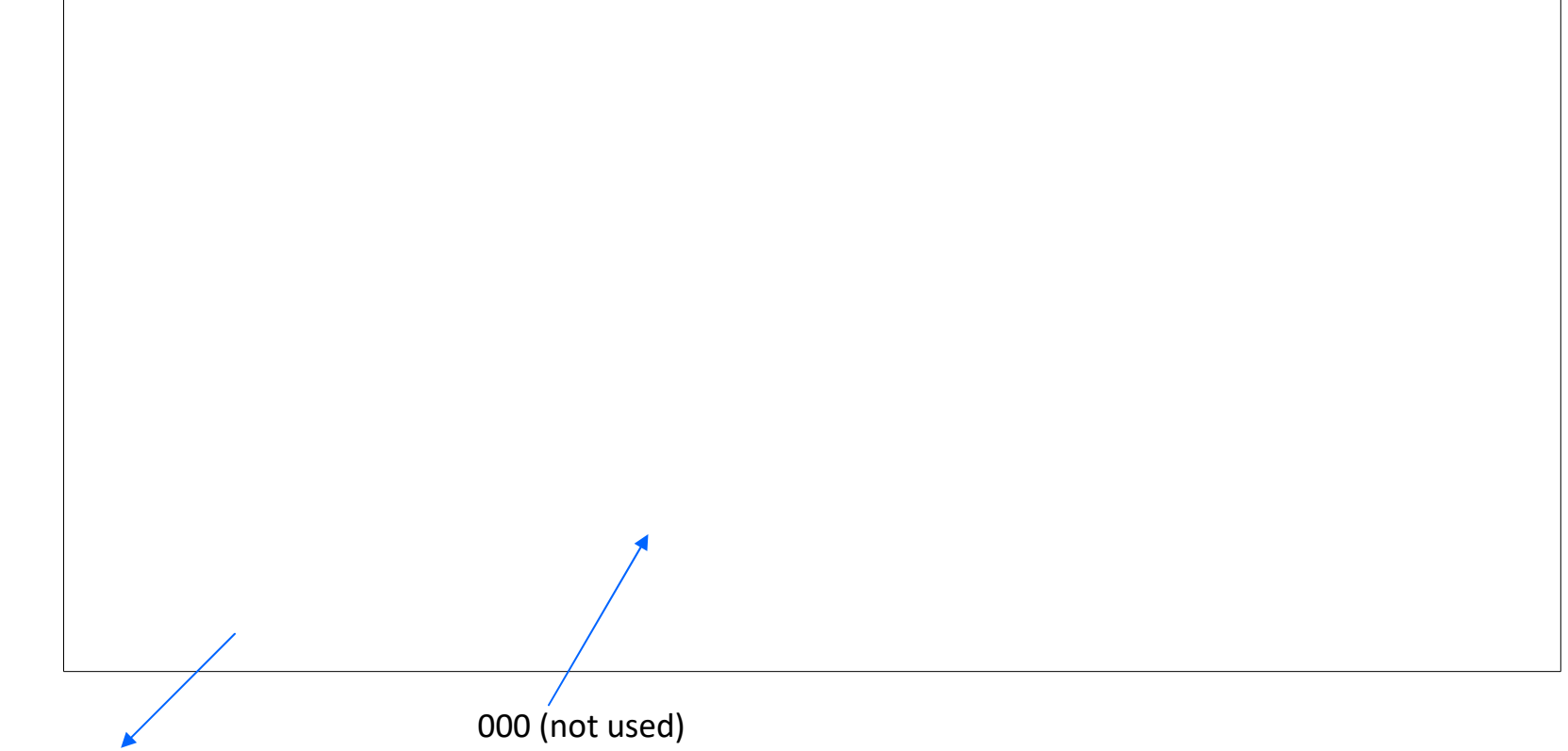
if SELD = 000

then the content of the output bus is available in the external output

1. ALU provides arithmetic and logic operations

8. Encoding of ALU operations (function table for this ALU is listed in Table 4-8):

9. Examples of microoperations and corresponding control words:


000 (not used)
XOR(x,x) = 0

10. The most efficient way to generate control words with a large number of bits is to store them in a memory unit, which is referred to as control memory (control store).

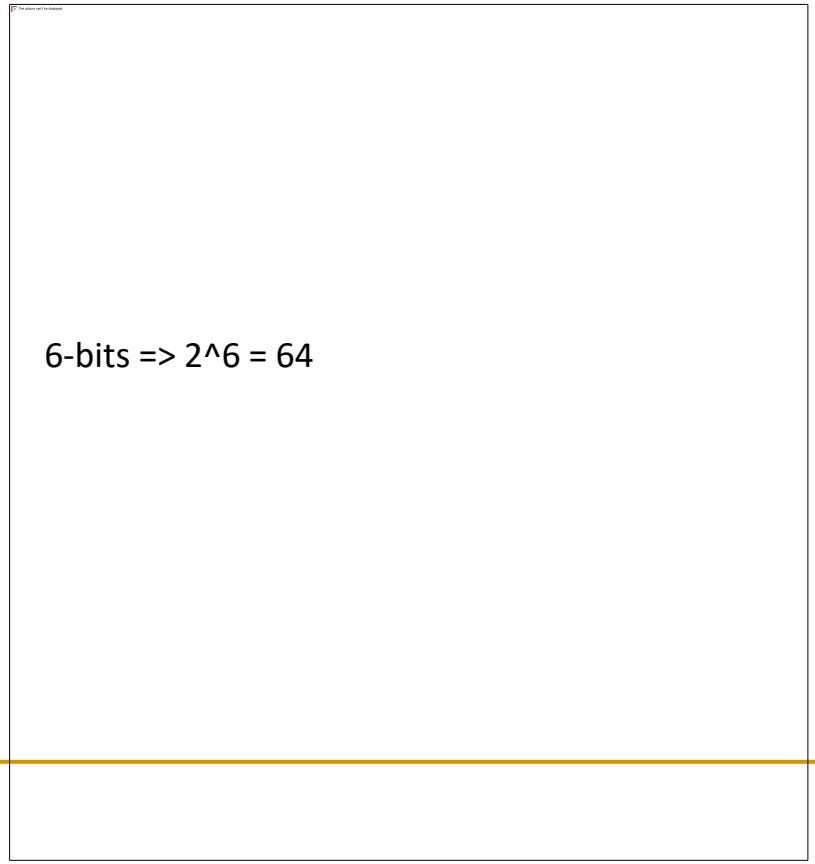
11. By reading consecutive control words from memory, it is possible to initiate the desired sequence of microoperations for the CPU => microprogrammed control.

Stack Organization

1. A stack is a storage device for storing information in such a manner that the item stored last is the first item retrieved (LIFO – last-in, first-out).
2. The stack is a memory unit with an address register called a stack pointer (SP), which always points at the top item in the stack.
3. The two operations of a stack are the insertion (push) and deletion (pull) of items.

4. push-operation increments the SP and pull-operation decrements the SP.

1. Stack can reside in a portion of a large memory unit or it can be organized as a collection of a finite number of (fast) registers.
2. Fig. 8-3: organization of 64-word register stack.



push A
push B
push C

push (performed if stack is not full *i.e.* if FULL = 0):

$SP \leftarrow SP + 1$	Increment stack pointer
$M[SP] \leftarrow DR$	Write item on top of the stack
If $(SP = 0)$ then $(FULL \leftarrow 1)$	Check if stack is full
$EMTY \leftarrow 0$	Mark the stack not empty

pull (performed if stack is not empty *i.e.* if EMTY = 0):

1. Stack can also be implemented with a RAM attached to a CPU
    1. a portion of memory is assigned to a stack operation
    2. a processor register is used as a stack pointer
3. Fig. 8-4 shows how a portion of memory partitioned into three segments: program, data, and stack.
4. Most computer do not provide hardware for checking stack overflow or underflow
    1. if registers are used to store the upper limit (e.g. 3000) and the lower limit (e.g. 4000), then after push SP can be compared against the upper limit register, and after pull against the lower limit register.
2. The advantage of the memory stack is that CPU can refer it without having to specify an address: the address in always in SP and automatically updated during a push or pop instruction.

PC, AR, and SP provide addresses for the memory.

1. PC is used in fetch phase to read instruction from the memory

2. AR is used to read an operand during the execution phase of an instruction.

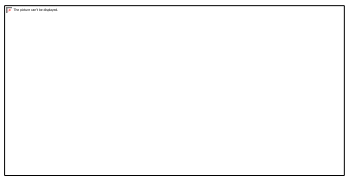
3. SP is used to push or pop items into or from the stack.

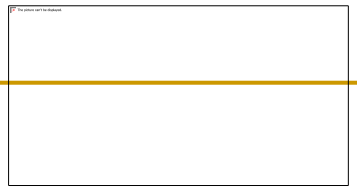
4. <u>In this example</u>, stack grows with decreasing addresses.

5. First item stored is at address 4000.

6. the last address that can be used is 3000.

push:





pull:

(Mano 1993)

1. A stack is effective for evaluating arithmetic expressions

7. Arithmetic operations are usually written in infix notation: each operator resides between the operands, *e.g.*:

(A * B) + (C * D), where x denotes multiplication:

1. A * B and C * D has to be computed and stored.
2. after the two products, sum (A * B) + (C * D) is computed

=> there is no straight forward way to determine the next operationthat is performed.

8. Arithmetic expressions can be presented in prefix notation (also referred to as Polish notation by Polish mathematician Lukasiewicz): operators are placed before the operands.

9. The postfix notation (reverse Polish notation (RPN)) places the operator after the operand.

10. *E.g.*:

A + B , infix notation

+AB , prefix notation

AB+ , postfix notation (RPN)

1. The reverse Polish notation suits of stack manipulation

11. E.g. the expression is

A * B + C * D AB*CD*+

written in RPN as

and is evaluated by scanning from left to right: when operator is found, the operation is performed by using operands on the left side of the operator. The operator and operands are replaced by the result of operation. The scan is continued and the procedure is repeated for every operator:

1. * is found

1. take the two operands from left: A and B
2. compute P = A * B
3. replace operands and operator with the result => PCD*+
4. continue scan
5. * is found

6. take the two operands from left: C and D

7. compute Q = C * D
8. replace operands and operator with the result => PQ+
9. continue scan

11. + is found

1. take the two operands from left: P and Q
2. compute R = P + Q
3. replace operands and operator with the result: R

4. continue scan: no more operators => stop; R is the result of evaluation.

5. The conversion from infix to RPN must take into consideration the operational hierarchy of infix notation:

1. first perform arithmetic inside inner parentheses
2. ..then inside outer parentheses
3. perform multiplication and division before addition and subtraction.

6. E .g.: (A + B)*[C*(D + E) + F] becomes AB+DE+C*F+*
which is computed:

1. P = A+B => PDE+C*F+*
2. Q = D+E => PQC*F+*
3. R = Q * C => PRF+*
4. S = R+F => PS*
5. T = P*S
    1. T represents the result: T = AB+DE+C*F+*

6. RPN is the most efficient method known for evaluating arithmetic expressions.

7. Used *e.g.* in electronic calculators

    1. RPN is an useful way to represent arithmetic expression for a generic arithmetic evaluator.

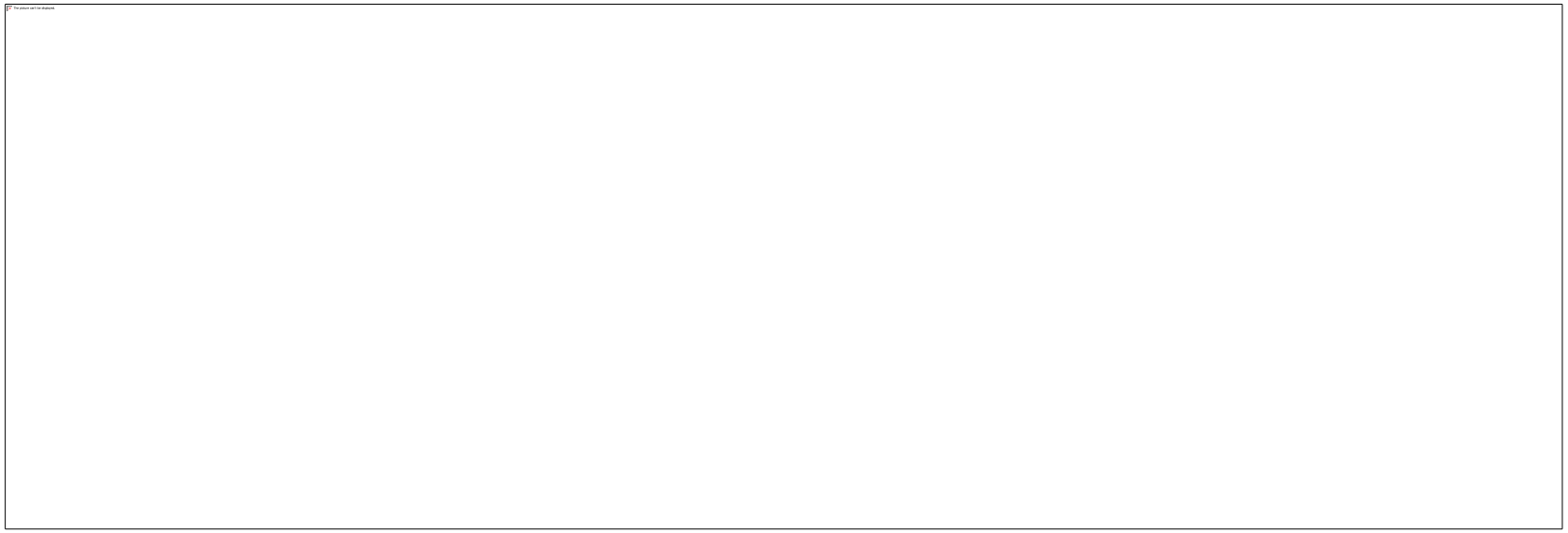
2. Stack is useful for evaluating arithmetic expressions in RPN

    1. operands are pushed into the stack in the order of appearance (in RPN)
    2. the topmost operands are popped from the stack and used for the operation

3. The result is pushed to replace the popped operands

4. Most compilers convert all arithmetic expressions into Polish notation: efficient translation of arithmetic expressions into machine language instructions.

1. 5 ع : (3*4)+(5*6) ← 34*56*+



3 4 * 5 6 * +

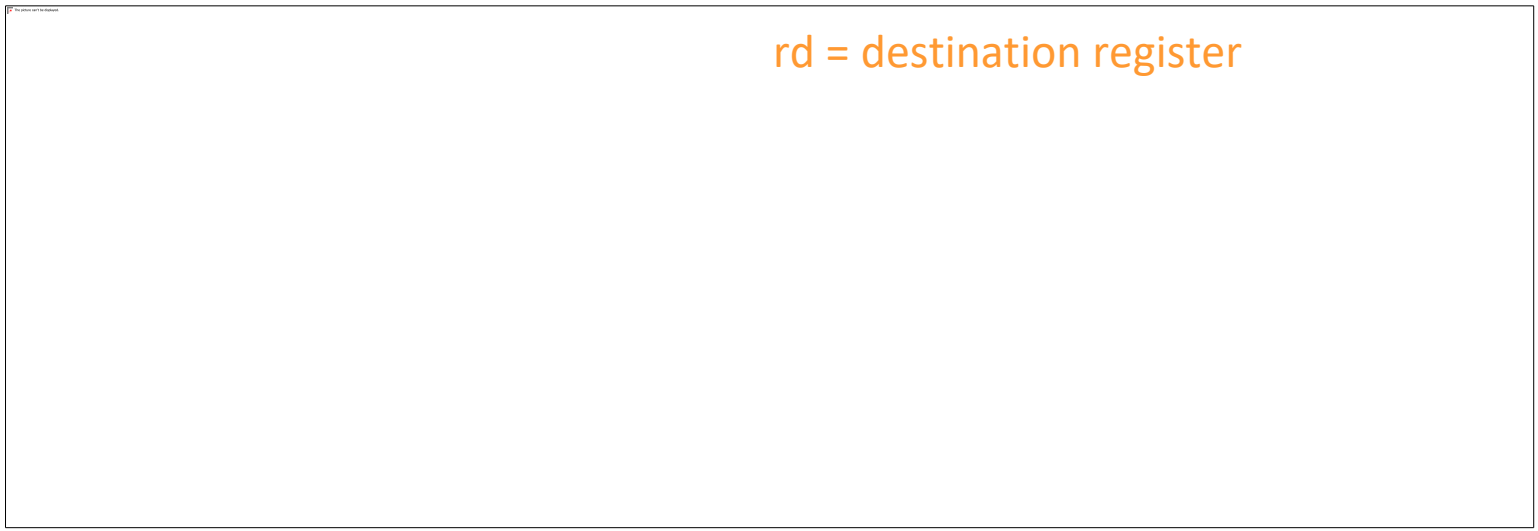
Instruction Formats

1. The typical fields found in instruction formats are:

1. Operation code specifying the operation: add, subtract, complement, etc.
2. Address field designating a memory address or a register
3. Mode field for specifying the way for determining the effective address of an operand.

    1. The number of address fields in the instruction format of a computer depends on the internal organization of its registers.

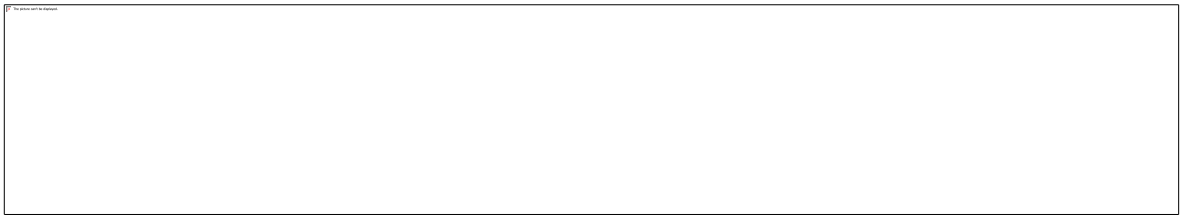
1. E.g.: MIPS (a RISC microprocessor architecture developed by MIPS Computer Systems Inc.)




The picture can't be displayed.

1. Computers may have instructions of several different lengths containing varying number of addresses.

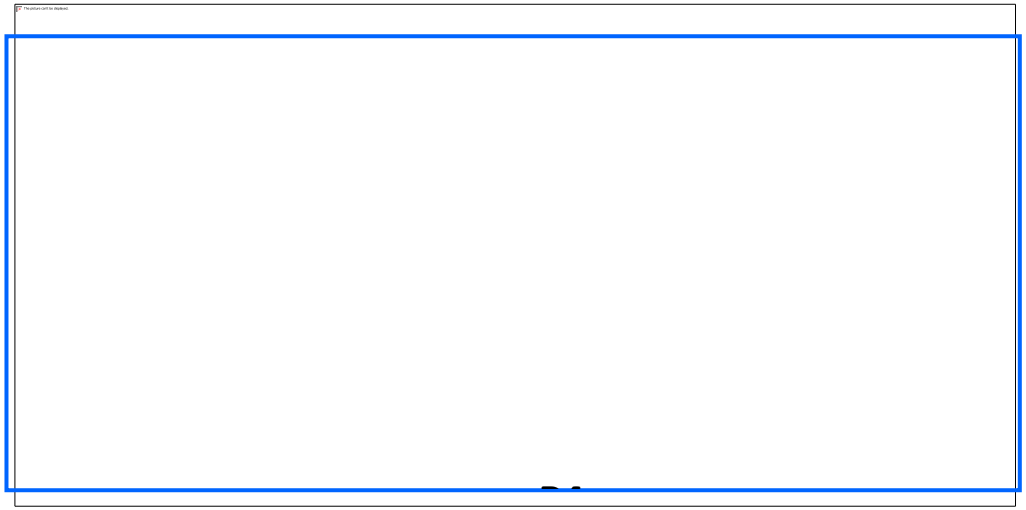
2. E.g. three-address instructions (A+B)*(C+D):



3. E.g. two-address instructions:


The present can be displayed.

1. E.g. RISC instructions:



Addressing Modes

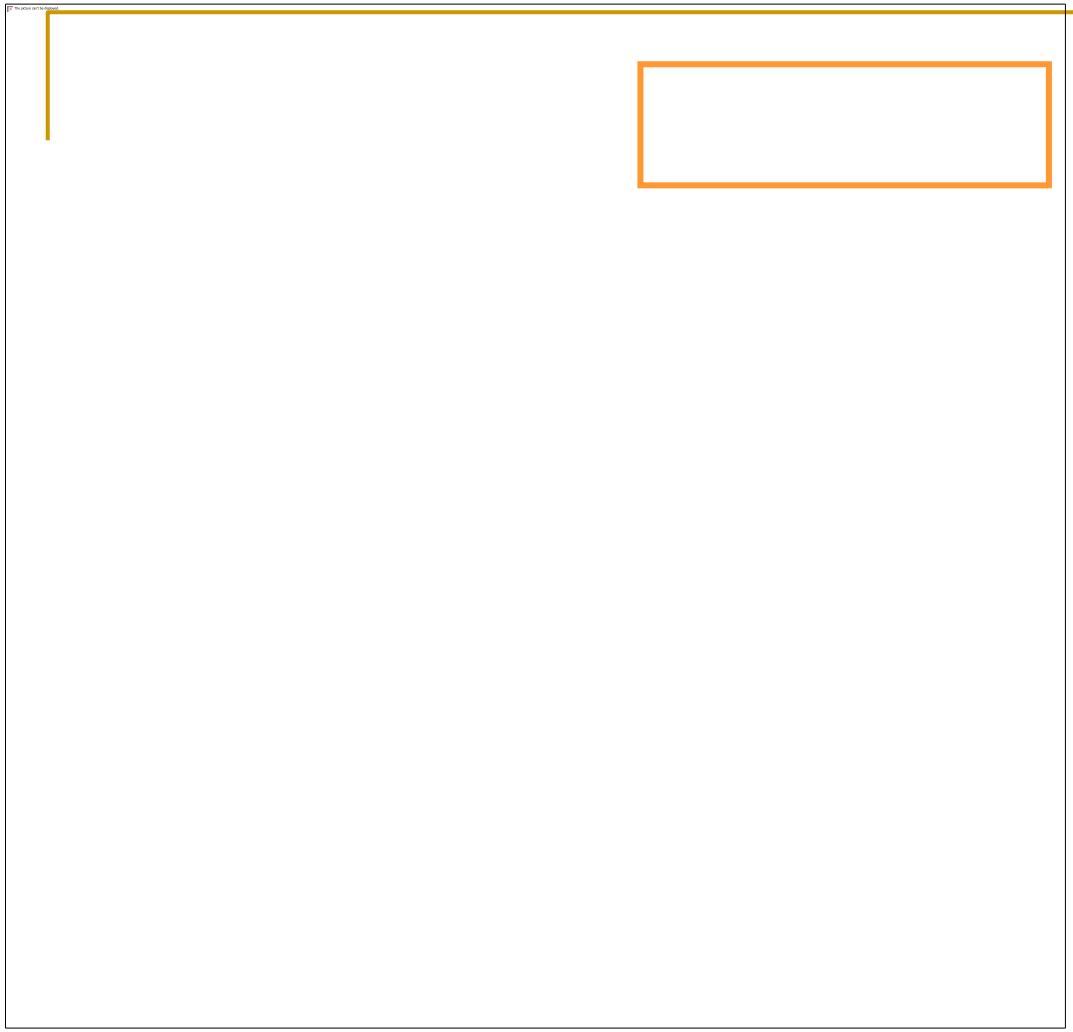
1. The addressing mode specifies a rule for interpreting or modifying the address field of the instruction before the operand is actually referenced.
2. Addressing modes are used:
    1. To provide programming versatility for the user: pointers to memory, counters for loop control, indexing data, etc.
    2. To reduce the number of bits in the addressing field of the instruction.
3. The decoding phase of an instruction cycle determines the addressing mode(s) and the locations (registers and/or memory locations) of operands.
4. Depending on the CPU, an instruction can have more than one address field, and each address field may be associated with its own particular addressing mode.

1. Different addressing modes:

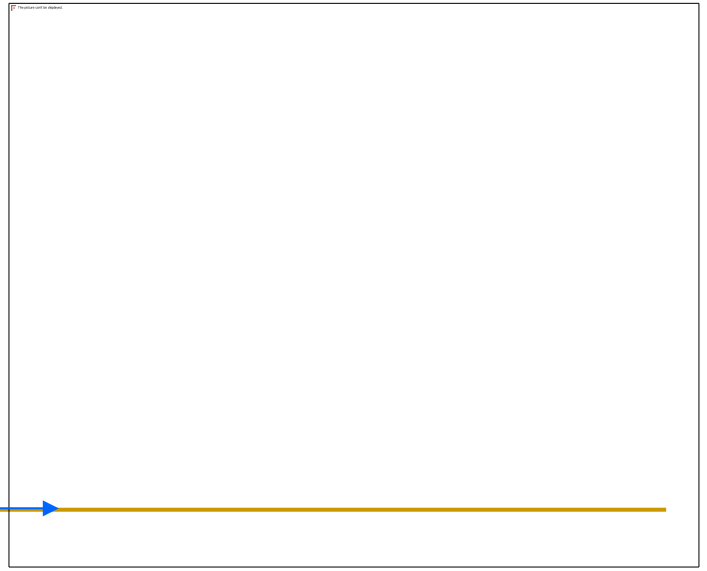
1. Implied mode: the operands are implicitly specified by the instruction, e.g.: “complement accumulator”.
2. Immediate mode: the operand is specified in the instruction (in operand field). Can be used e.g. to initialize register to constant value (=immediate operand).
3. Register mode: operands are in registers that reside within the CPU. The particular register is selected with the register field of the instruction.
4. Register indirect mode: the content of a register specifies the address of the operand in memory.
5. Autoincrement or autodecrement mode: similar to register indirect mode but the content of the register is automatically incremented/decremented after/prior data access.

1. Direct address mode: the effective address is equal to the address part of the instruction. The operand resides in this address.
2. Indirect address mode: the address field gives the address where the effective address is stored in memory.
3. Relative address mode: the content of the program counter (PC) is added to the address part of the instruction in order to obtain the effective address *i.e.* the effective address is relative to the address of the next instruction. This address mode can be used in branch-type instructions when the branch address is in the area surrounding the instruction word.

4. Indexed addressing mode: content of an index register (special CPU register) is added to the address part of the instruction to obtain the effective address. The index register can be incremented for accessing consecutive operands.
5. Base register addressing mode: the content of a base register is added to the address part of the instruction to obtain the effective address. The address part of the instruction gives the displacement relative to the base address.

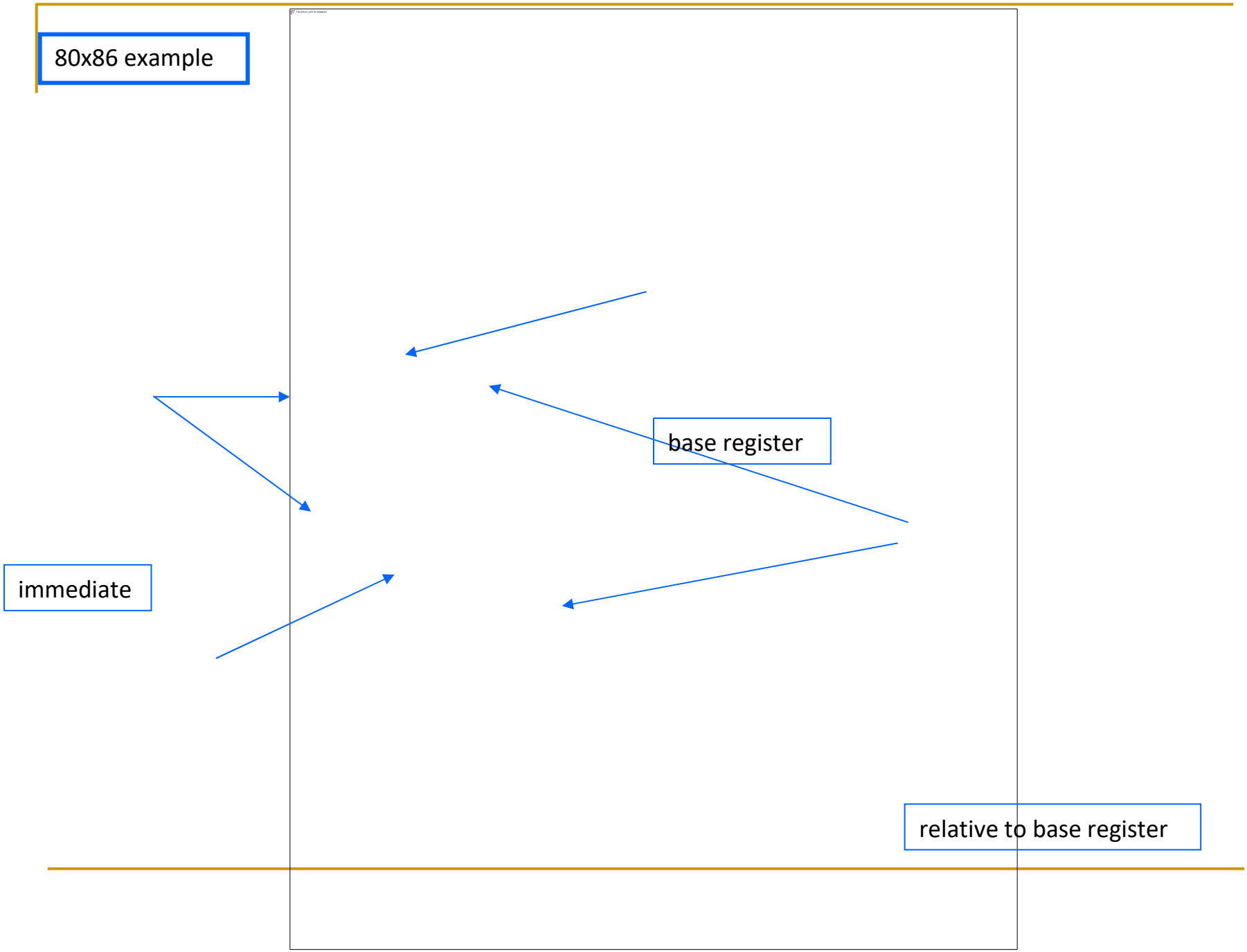


instruction with an address and addressing mode



(Mano 1993)

decrements R1 prior the
execution


80x86 example
base register
immediate
relative to base register

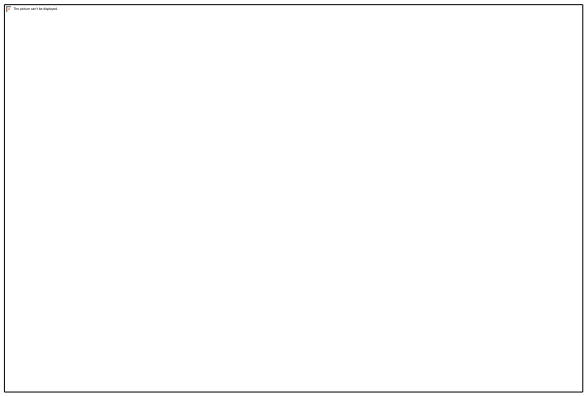
relative address

compile:

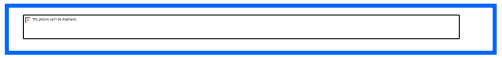
1. **nasm -f obj example2.asm**

Program Control

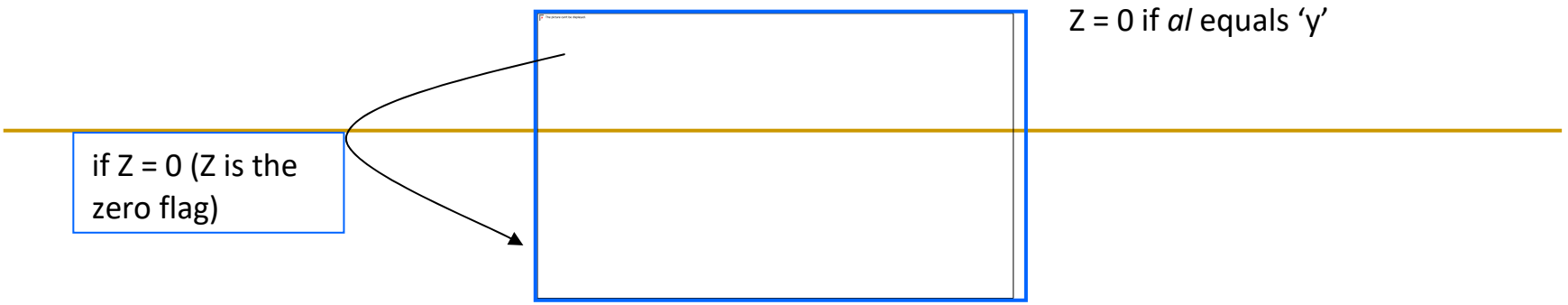
1. Program flow can be altered by instructions that modify the value of the program counter: important feature of a digital computer – provides a control over the program flow and capability for branching to different program segments (blocks of memory).
2. Typical program control instructions:



1. Branch and jump instructions may be conditional or unconditional
    1. An unconditional branch instruction causes a branch to the specific address without any conditions, *e.g.*:



    2. The conditional branch specifies a condition, e.g. branch if zero: only when the condition is met, the program counter is loaded with the branch address, *e.g.*:



6. Compare and test instructions can be used in setting conditions for subsequent conditional branch instructions

1. Compare performs an arithmetic subtraction: result is not saved – only status bit conditions are set as a result of operation.
2. Similarly test performs logical AND of two operands and updates certain status bits.

1. The status register stores the values of the status bits (status register is composed of the status bits).
2. Bits of the status register are modified as a result of an operation performed in the ALU.
3. *E.g.* (8-bit ALU with a 4-bit status register):

C (carry) is set to 1 if end carry C8 is 1. C is cleared to 0 if C8 is 0.

S (sign) is set to 1 the highest-order bit F7 is 1. S is set to 0 if F7 is 0.

Z (zero) is set to 1 if the output of ALU is 0 (all 0's). Z is set to 0 otherwise.

V (overflow) is set to 1 is XOR of the last two carries is equal to 1, and cleared to 0 otherwise.

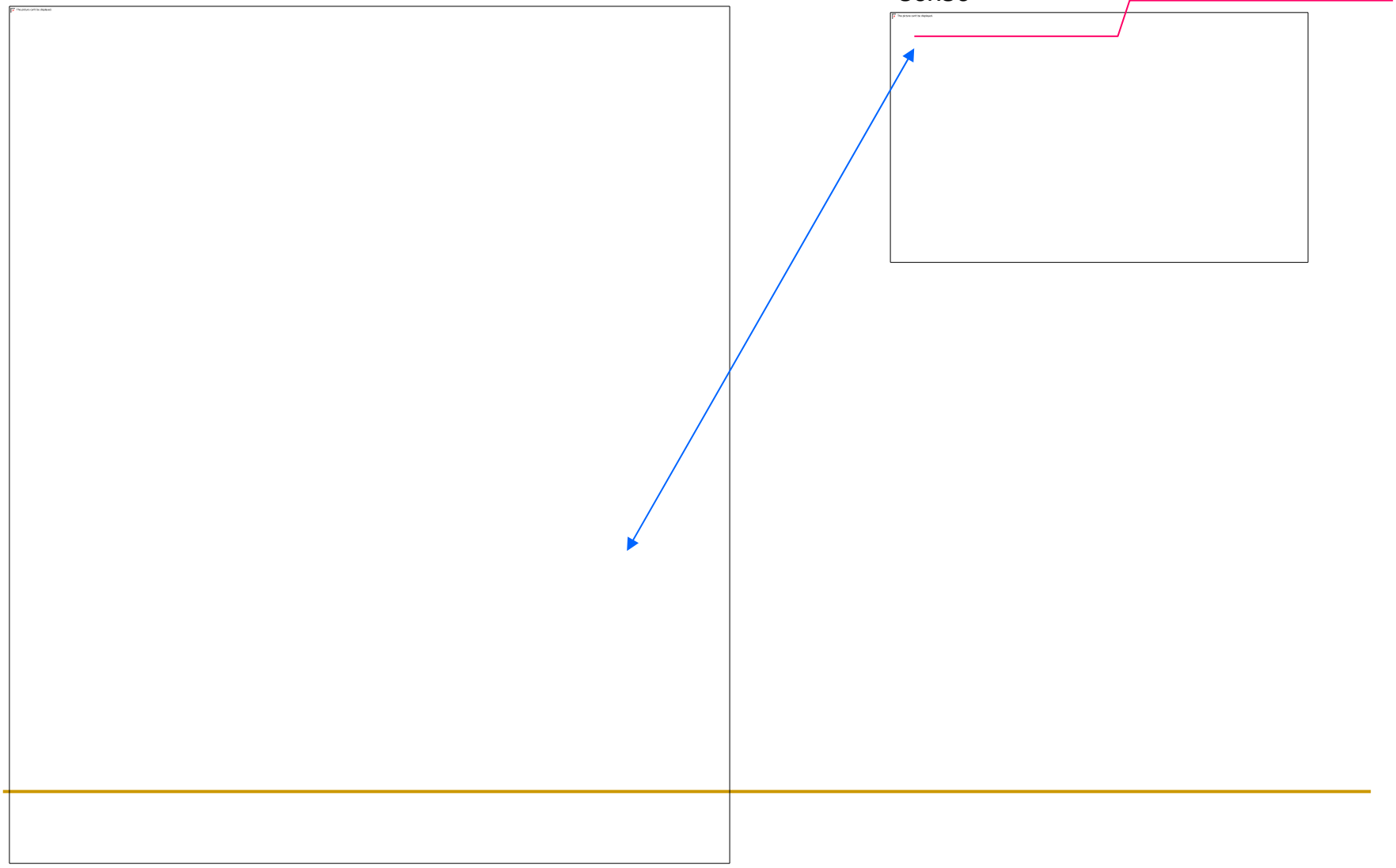
(Mano 1993)

3. Status bits can be checked after ALU operation to determine certain relationships that exist between the values of A and B

    1. V indicates overflow i.e. for 8-bit ALU the result is greater than 127 or less than -127.
    2. If Z is set, the result is zero:

1. we can use e.g. XOR operation to compare two numbers (the result is zero iff *A* = *B*) and Z indicates the result of comparison.
2. A single bit in *A* can be checked with a mask that contains 1 in that particular bit position (others being 0's) and by using AND operation.

3. Conditional branch instructions use the status bits for checking conditions for branching:



1. For subroutine calls, different computers can use a different temporary location for storing the return address
    1. some computers use the first memory location of the subroutine (like the Basic Computer).
    2. some store the return address in a fixed memory location.
    3. some computers use a processor register.
    4. stack memory is yet another possibility (the most efficient way): when a succession of subroutines is called (nested calls), the sequential return addresses can be pushed into the stack. The “return from subroutine” instruction pops the return address (and assigns to program counter) from the top of the stack: we always have the return address for the last called subroutine.

1. Subroutine call (stack based) microoperations:

2. .. and return:

3. By using subroutine stack each return address (in nested calls) can be pushed into the stack without destroying any previous values
    1. *e.g.* in basic computer a recursive subroutine call would destroy the previous return address stored in the first memory location of the subroutine.

2. Program interrupt refers to the transfer of program control from a currently running program to another service program as a result of an external or internal generated request

    1. otherwise similar to subroutine call, except:

1. the interrupt is (usually) initiated by an internal of external signal rather than an execution of an instruction (software interrupts are exceptions).
2. the address of the interrupt service program (routine) is determined by hardware rather than the address field of an instruction: the CPU must possess some form of HW procedure for selecting a branch address servicing the interrupt.
3. Interrupt routine stores all the information (not just PC) necessary to recover the state of the CPU prior the return from the interrupt routine.

1. After the interrupt routine the CPU must return exactly the same state that is was when the interrupt occurred.
2. The state of the CPU at the end of the execute cycle (the interrupt is recognized in this phase) is determined from:

1. The content of PC
2. The content of all processor registers
3. The content of status conditions
    1. status bits (program status word PSW) stored in a separate status register.
    2. contains status information about the state of the CPU: bits from ALU operation, interrupt enable bits, and CPU operation mode (system mode, user mode), for example.

3. Some computer store only program counter (and PSW) prior entering to an interrupt routine
    1. the interrupt routine must take care of storing and restoring the CPU status.
4. CPU does not respond to an interrupt until the end of an instruction execution
    1. in an interrupt is pending control goes to a interrupt cycle.
    2. contents of PC and PSW are pushed onto stack.
    3. the branch address is transferred to PC and new PSW is loaded into the status register.
    4. the interrupt routine can now be executed starting from the branch address (which may contain a branch instruction to a user defined service routine).
    5. the last instruction of the interrupt routine is a “return from interrupt”: the stack is popped to retrieve PWS to status register and return address to PC

    => CPU state is restored and the interrupted program can proceed like nothing had happen.

5. Interrupt types:

1. External interrupts
    1. from I/O, timing, or any other external source.
    2. e.g.: I/O device requesting new data, elapsed time of an event, power failure, etc.
2. Internal interrupts (traps)
    1. from illegal or erroneous use of an instruction or data.
    2. e.g.: overflow, division by zero, invalid operation code, stack overflow, and protection violation.
    3. usually occur as a result of a premature termination of the instruction execution: the service program determines the corrective measure to be taken (*e.g*. terminates the program).
3. Software interrupts
    1. initiated by an instruction (rather than HW signals)
    2. a special call instruction that behaves like an interrupt.
    3. can be used by a programmer to initiate an interrupt routine at any desired point in the program.
    4. can be used for accessing operating system services, for example.

5. *E.g.*: using INT-instruction (i.e. INT 21h) for invoking a DOS interrupt (see 80x86 example code shown earlier):

```
TimePrompt              DB 'Is it after 12 noon (Y/N)?$'
        .
        .
```

pass information to the operating system in order to specify the particular task requested

```
mov     dx,TimePrompt
mov     ah,9
int     21h
```

INT 21,9 documentation (http://members.tripod.com/~oldboard/assembly/int_21-9.html):

INT 21,9 - Print String

```
AH = 09
DS:DX = pointer to string ending in "$"


returns nothing


- outputs character string to STDOUT up to "$"
- backspace is treated as non-destructive
- if Ctrl-Break is detected, INT 23 is executed
```

Reduced Instruction Set Computer (RISC)

1. Instruction set determines the way that machine language programs are constructed.
2. Early computers had small and simple instruction sets in order to minimize the (expensive) hardware needed for their implementation.
3. Today many computers have instructions that include 100 to 200 instructions
    1. variety of data types
    2. large number of addressing modes

3. Complex instruction set computer (CISC) has complex hardware and large instruction set: functions from software to hardware.
4. In contrast, reduced instruction set computer (RISC) uses fewer and simpler instructions which can be executed faster within the CPU.
5. RISC chips require fewer transistors (than CISC), which makes them cheaper to design and produce.
6. There is still considerable controversy among experts about the ultimate value of RISC architectures
    1. Its proponents argue that RISC machines are both cheaper and faster, and are therefore the machines of the future.
    2. Skeptics note that by making the hardware simpler, RISC architectures put a greater burden on the software. They argue that this is not worth the trouble because conventional microprocessors are becoming increasingly fast and cheap anyway.

7. However, CISC and RISC implementations are becoming more and more alike
    1. Many of today's RISC chips support as many instructions as yesterday's CISC chips
    2. Today's CISC chips use many techniques formerly associated with RISC chips.
8. One reason for the trend to provide a complex instruction set is to simplify the translation from high-level to machine language programs.
9. Characteristics of CISC architecture:

1. A large instruction set.
2. Instructions that perform special tasks and are used infrequently.
3. A large variety of addressing modes (5-20 different modes).
4. Variable-length instruction formats.
5. Instructions that manipulate operands in memory.

Characteristics of RISC architecture:

1. Relatively few instructions
    1. mostly register-to-register operations
2. Relatively few addressing modes (because of 1)
3. Memory access limited to load and store instructions.
4. All operations done within the register of the CPU.
5. Fixed-length, easily decoded instruction format
    1. aligned to word boundaries
    2. simplifies control logic
6. Single-cycle instruction execution
    1. fetch, decode, and execute phases for two to three instructions overlap: pipelining.
    2. Memory references may take more clock cycles.

7. Hardwired rather than microprogrammed control
    1. faster

Other RISC characteristics:

1. A large number of register
    1. useful for storing intermediate results and for optimizing operand references: much faster than memory references.
    2. most frequent accessed operands are kept in registers
2. Use of overlapped register windows to speed-up procedure call and return.
3. Efficient instruction pipeline
4. Compiler support for efficient translation of high-level language programs into machine language programs.

1. A characteristic of some RISC processors is their use of overlapped register windows to provide the passing of parameters and avoid need for saving and restoring register values: speeds up procedure calls and returns.

2. Each procedure call results in the allocation of a new window consisting of a set of registers
    1. current window pointer (CWP) is decremented: corresponds ‘save’ in Fig. 8-11.
3. Each return statement increments the CWP
    1. corresponds ‘restore’ in Fig. 8-11.
4. Windows for adjacent procedures (nested calls) have overlapping registers that are shared to provide the passing of parameters and results.
5. Local register can be used for local variables
    1. by using local registers there is no risk of corrupting data of another procedure (e.g. caller).

6. Overlapped registers are used to pass parameters (in) and store results (out).
7. Only one register windows is activated at any given time with a CWP.
8. Each procedure call activates new register window by updating (decrementing) the CWP.
9. To summarize:

1. Register windows provide easy access to a large collection of registers and can reduce the need to save registers in memory.
2. If you write a procedure with more parameters than common registers (reserved for input parameters), you will need to to use the stack for any parameters beyond the amount of reserved number of registers.
3. If your call sequence gets deeper than number of windows (as it probably will in most recursive procedures), you are again forced to use the stack.

NOTE: In Mano 1993, CWP is incremented in a call, and decremented in a return. The example in these slides refers to SPARC processor architecture.

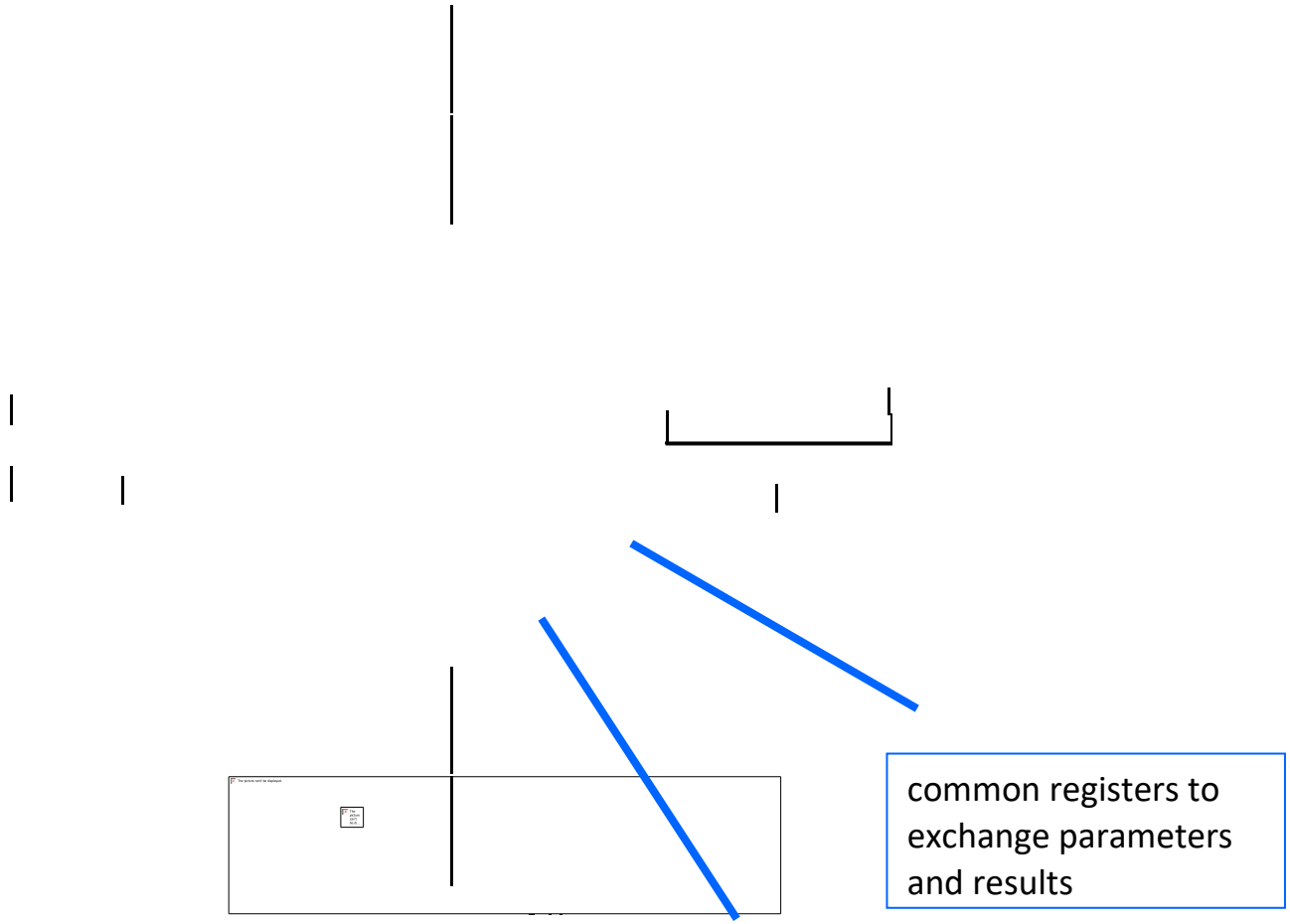
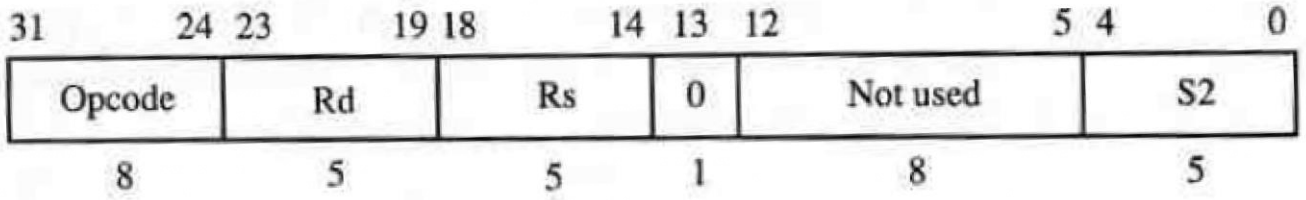


Fig. 8-11

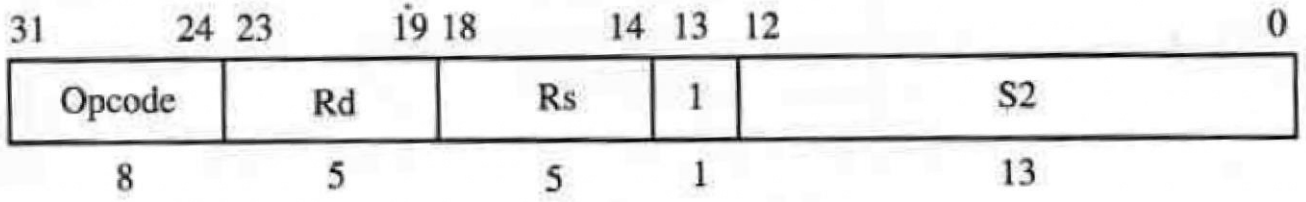
1. Berkeley RISC I

1. One of the first projects for showing the advantages of RISC architecture (in University of California, Berkeley).
    1. 32-bit CPU
    2. 32-bit addresses
    3. 8-, 16-, or 32-bit data
    4. 32-bit fixed length instruction format
        1. suits well for pipelining
    5. 32 instructions
    6. three addressing modes: register, immediate operand, and relative to PC addressing for branch instructions.
    7. <u>138 registers</u>
        1. 10 global and 8 windows of 32 registers.

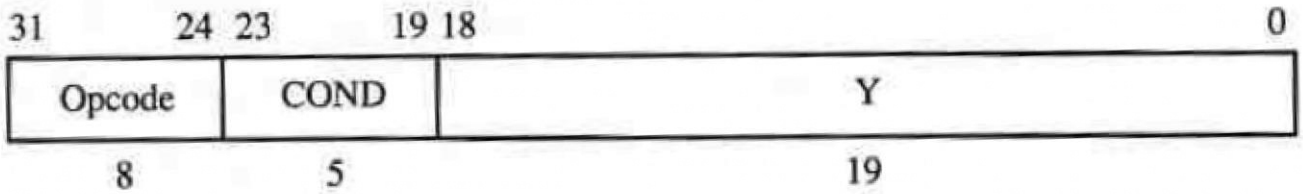
Figure 8-10 Berkeley RISC I instruction formats.



(a) Register mode: (S2 specifies a register)

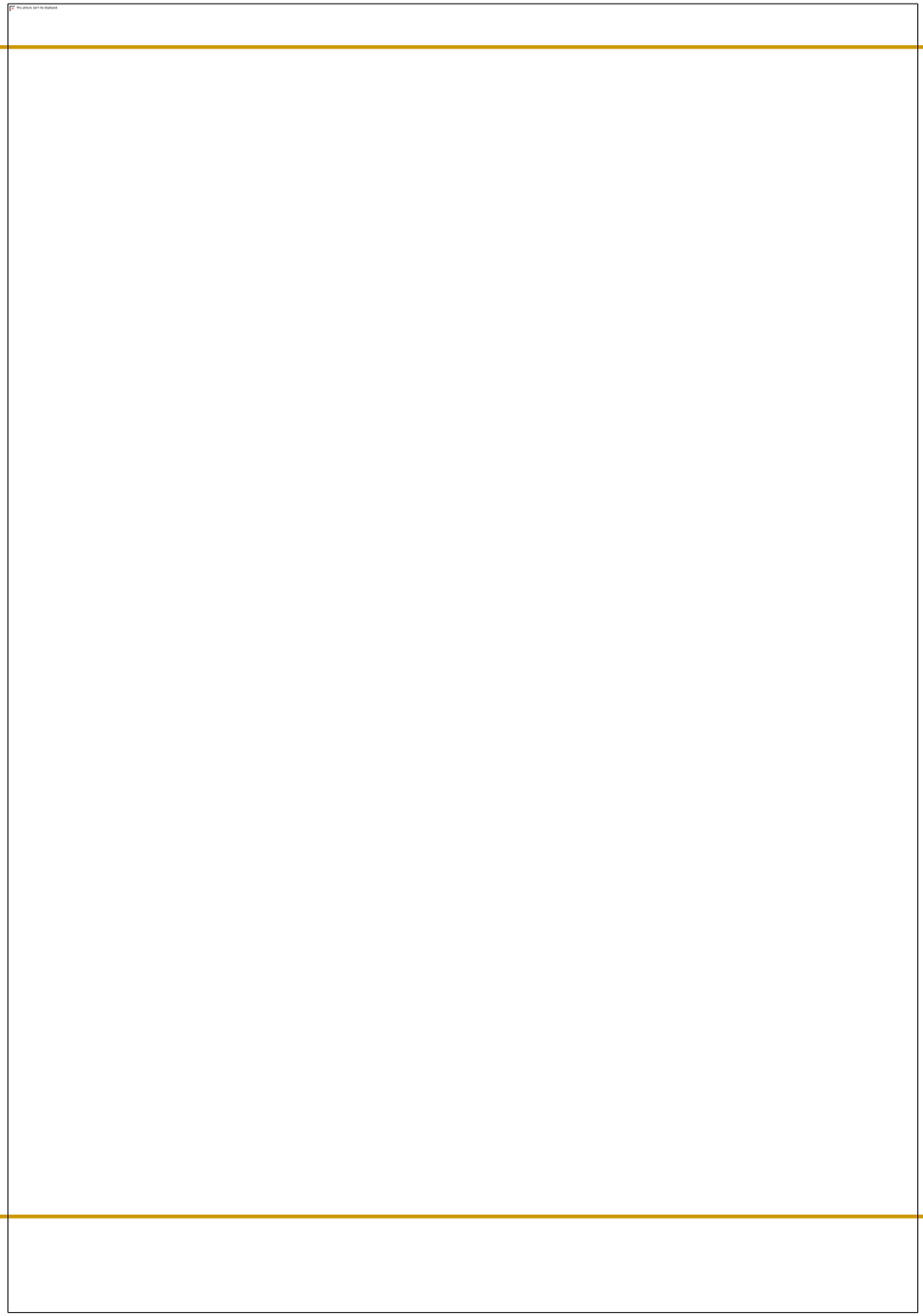


(b) Register–immediate mode: (S2 specifies an operand)



(c) *PC* relative mode:

(Mano 1993)



(Mano 1993)