1 Hardware

In questo capitolo mettiamo a punto qualche dettaglio ed aggiungiamo qualche concetto rispetto a ciò che è contenuto nel capitolo precedente.

L'incredibile sviluppo delle tecnologie microelettroniche

Non è questo il testo dove occuparsi in dettaglio dello sviluppo della microelettronica, ma è interessante spendere alcune parole sul più grande fenomeno tecnologico della fine del secondo millennio. Chi ha seguito per più di un anno lo sviluppo dei computer si è stupito della velocità con cui CPU nuove e molto più potenti sostituiscono quelle precedenti. In verità negli ultimi trent'anni lo sviluppo delle tecnologie microelettroniche ha fatto raddoppiare, a parità di dimensioni e costo, ogni anno e mezzo la potenza delle CPU e degli altri circuiti integrati.
Un circuito integrato è una piccola piastrina di silicio sulla quale, attraverso complessi processi chimico-fisici vengono creati dei dispositivi che sono in grado di memorizzare ed elaborare bit di informazione. Questi dispositivi (transistor) possono essere tanto piccoli da essere messi in grandi quantità in una singola scheggia di silicio ("chip"). E' possibile "integrare" decine di milioni di transistor in un singolo circuito. Un circuito che include molti transistor viene detto "**circuito integrato**".
Nel corso del tempo il numero dei transistor che le case produttrici sono state in grado di mettere su un singolo chip è aumentato con velocità straordinaria. Il nome "chip", che significherebbe la sola piastrina di silicio, spesso viene usato impropriamente anche per indicare tutto il circuito integrato, contenitore ("package") e piedini ("pin") inclusi.
Qual è la ragione della sempre incredibile velocità di sviluppo della microelettronica?
Il trucco è che la "filosofia" di base della microelettronica è sempre la stessa, non c'è bisogno di "salti" culturali o nelle tecniche di base per far progredire le prestazioni dei dispositivi. Si tratta "solo" di migliorare le tecnologie, di raffinare i modi di produzione dei circuiti integrati, così da realizzare transistor sempre più piccoli. Transistor piccoli permettono di conciliare esigenze che in molti altri campi della tecnica sono invece in contraddizione. Come nel vecchio detto, è un po' come avere la botte piena e la moglie ubriaca. Vediamo perché: avendo transistor piccoli se ne possono mettere di più sul chip, ottenendo circuiti che possono svolgere funzioni più complesse, oppure svolgere le stesse funzioni più rapidamente. Insieme a questo vantaggio c'è anche l'altro vantaggio che i circuiti cambiano di stato più rapidamente, perché devono muovere meno elettroni per elaborare i bit. Muovere meno elettroni significa erogare minori correnti elettriche, e perciò avere minor consumo di energia e minor riscaldamento del chip. Per cui, in un circolo virtuoso simile ad una retroazione negativa, è possibile metterne di più senza bruciare il circuito.
In sintesi con la sola diminuzione della dimensione dei transistor si ottengono circuiti integrati più "densi" di transistor, e quindi più potenti, più veloci, che consumano di meno e che scaldano di meno. Detto questo, non c'è da meravigliarsi del fatto che in trent'anni la dimensione dei transistor sia andata da 10 mm a 0,1 µm!

1.1 Tipi di computer

Per ragioni storiche e pratiche i computer sono distinti in diversi tipi. Si chiamano "**mainframe**" i grossi e costosi computer centrali, usati da decine o da centinaia di utenti, che risiedono nei Centri di Calcolo. Trent'anni fa erano l'unica alternativa, in quanto ogni computer era tanto costoso da dover essere condiviso fra molti utenti per poterne giustificare la spesa.
Gli utenti utilizzano i mainframe collegandosi con "terminali", dispositivi di Ingresso / Uscita che non possono svolgere alcuna elaborazione. Oggi i mainframe, che hanno software collaudato ed affidabile, vengono usati soprattutto per la gestione degli archivi molto grandi.
Si definiscono "**workstation**" computer che possono servire contemporaneamente più di un utente, collegato con terminali, ma che sono di dimensioni e costo minore di quello dei mainframe ed hanno una minore capacità di archiviazione. Anche se è possibile l'utilizzazione da parte di molti utenti le workstation vengono spesso utilizzate da una sola persona. L'uso più tipico delle workstation è per quelle applicazioni che richiedono una grande potenza di elaborazione dedicata ad un gruppo ristretto di persone, come il CAD, la progettazione, la simulazione o la computer grafica.
Si definiscono "**personal computer**" quei sistemi pensati e prodotti per l'uso da parte di una sola persona, e che non sono perciò destinati a svolgere elaborazioni per più di un utente.
A questo proposito vogliamo fare una precisazione sulla terminologia. Nel seguito e di questo e degli altri volumi delle "serie" quando si userà il termine "personal computer" intenderà trattare di tutti quei computer che non sono destinati ad essere condivisi fra le persone, mentre quando si userà la sigla **PC** si intenderà parlare di un computer 100% compatibile con la linea dei primi personal computer della IBM, chiamata PC.
Spesso quando si parla di computer diversi si usa un termine brutto in lingua italiana, ma che ha un significato specifico in gergo informatico. Esso è "**piattaforma**" (platform). In Informatica per "piattaforma" si intende uno specifico tipo di computer (non solo di CPU). Come esempio si possono prendere i PC, che hanno CPU della famiglia X86 o le workstation Sun, che hanno CPU della famiglia Sparc. Computer diversi con la stessa CPU, come erano per esempio Amiga e Macintosh, sono due piattaforme diverse.

Spesso oltre all'hardware si parla di piattaforma considerando anche il software che "riveste" un certo hardware, per esempio si può parlare di "piattaforma Windows" considerando i PC che usano Windows come sistema operativo o di "piattaforma Linux" considerando i PC che invece usano Linux.

Computer general purpose o dedicati

Un computer **dedicato** ("special purpose") è pensato e costruito per svolgere un compito specifico, mentre uno "**general purpose**" (a scopo generale) è fatto, come ben dice il nome, per essere adattato facilmente agli scopi più vari, cambiando solamente il programma in esecuzione. I computer dedicati sono usati in due casi: quando la complessità del problema da risolvere è talmente grande che necessitano soluzioni "speciali" e non basta un computer general purpose oppure quando il numero dei dispositivi da produrre è talmente alto che piccolissimi risparmi sull'hardware significano vantaggi economici significativi. Pertanto il primo tipo di computer dedicato è impiegato per esempio nella computer grafica molto avanzata o nella simulazione di fenomeni complessi, come la meteorologia.
Il secondo tipo è usato nei sistemi "**embedded**", cioè quei piccoli sistemi che comprendono una CPU e che incontriamo sempre più spesso nella vita di tutti i giorni, come il telefono mobile, la carta del parcheggio "intelligente" o il forno a microonde con il quale ci si può collegare ad Internet.

1.1.1 Organizzazione di un computer

Come già detto in ogni computer si distinguono fisicamente una "scatola", detta "unità centrale", ed altri dispositivi detti "periferiche". Vediamo cosa è contenuto nell'unità centrale, facendo riferimento all'architettura di un PC ma non dimenticando che tutti i tipi di computer del giorno d'oggi sono piuttosto simili.
All'interno dell'Unità Centrale, oltre all'alimentatore, troviamo una scheda elettronica, detta "**scheda madre**" ("motherboard") ed alcune schede "figlie" ("daughter board") che vengono collegate a periferiche interne ed esterne. Sono infatti presenti, fisicamente dentro l'unità centrale, anche alcune unità periferiche quali l'hard disc, il floppy disc ed il CDROM (vedi dopo).

Scheda madre

Tutti i computer hanno una scheda madre. E' la scheda sulla quale sta la CPU, che contiene anche la memoria principale ed i bus di sistema (address bus, data bus ed eventuali altri bus).
In una scheda madre sono inclusi molti sottosistemi. Peraltro, se la scheda è moderna, oltre alla CPU troviamo pochi altri circuiti integrati importanti, che individuiamo facilmente perché hanno molti piedini.
Questi circuiti vengono uniti sotto il nome di "**chipset**" della scheda e concentrano una miriade di funzioni diverse, che un tempo erano realizzate da molti circuiti integrati separati. Le funzioni che il chipset realizza sono relative al funzionamento della totalità del computer.

Un chipset moderno contiene la logica digitale che realizza: timer e contatori per la temporizzazione delle operazioni del software e dell'hardware, logica per la gestione della memoria RAM, per il controllo delle interruzioni e del DMA (vedi Volume 2 di questo testo), i controller dell'hard disc e del floppy disc, porte di comunicazione seriali e parallele, gestione del bus di espansione.

Oltre ad altri piccoli componenti elettronici, che vanno sotto il buffo nome di "glue logic" (logica "che fa da colla"), si notano la memoria di lavoro di tipo RAM, alloggiata in moduli a forma di barretta, comprendenti 8 o 9 circuiti integrati, e quella di tipo ROM, costituita dai circuiti integrati su cui è incollata un'etichetta con la sigla "BIOS" (o qualcosa del genere).
C'è da dire che alcuni computer hanno solo la scheda madre. Ciò accade quando non è prevista l'espansione del sistema durante la sua vita utile e se ne vogliono limitare al massimo i costi.

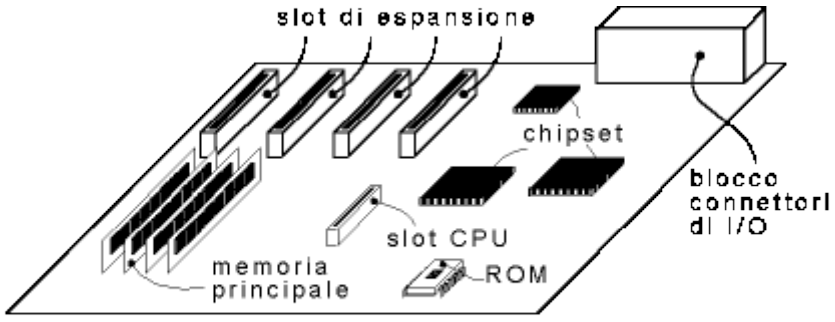


Figura 1: "componenti di una scheda madre"

Bus

Bus in multiplex

In un computer alla Von Neumann prima di scambiare un qualsiasi dato è necessario emettere sull'address bus l'indirizzo della locazione di memoria che è interessata dallo scambio. Per tutto il tempo dello scambio del dato quell'indirizzo viene mantenuto sull'address bus. Dato che comunque l'indirizzo viene sempre emesso PRIMA alcune CPU fanno uso degli stessi piedini sia per il data bus che per l'address bus. Durante la prima fase dell'accesso alla memoria la CPU emette l'indirizzo, poi, sugli stessi piedini scrive o legge il dato corrispondente a quell'indirizzo. In questo modo la CPU può avere meno piedini. Naturalmente sarà necessario aggiungere dei circuiti elettronici esterni ("latch") che "ricordano" l'indirizzo emesso durante la prima fase dell'accesso in memoria e lo continuano a mantenere fino alla fine dell'operazione. Quando data bus e address bus usano gli stessi piedini della CPU si dice che sono "in multiplex" (multiplixed) oppure, con brutto neologismo, "multiplexati".

Bus di espansione

La maggior parte dei computer general purpose prevede l'aggiunta di funzionalità attraverso l'inserzione di schede in opportuni "**slot d'espansione**" (in Inglese slot significa fessura). Le schede di espansione hanno un "pettine" con conduttori elettrici, che si vanno ad accoppiare con gli omologhi contatti contenuti nel connettore delle slot. Perché la scheda possa comunicare con la CPU gli slot devono prevedere l'accesso ai bus di sistema. Sui connettori degli slot devono essere riportati i segnali elettrici relativi al bus dati ed al bus indirizzi usati dalla CPU del computer. Sono inoltre presenti segnali di controllo (control "bus"), di temporizzazione ("clock") e sincronizzazione ("strobe"), che permettono di effettuare lo scambio di informazioni fra la CPU e la scheda di espansione.
Il chipset della scheda madre avrà anche l'importante compito di far funzionare insieme il bus d'espansione ed i bus di sistema.

1.1.2 Parallelismo

Molte delle funzioni di un computer sono svolte in parallelo, cioè contemporaneamente. Tanto maggiore è il numero delle funzioni svolte in parallelo tanto più veloce può essere un computer. Naturalmente ciò ne complica la struttura e quindi significa un aumento del suo costo.
Si parla di "**parallelismo"** per indicare con quante unità in contemporanea sono svolte alcune funzioni.

Parallelismo del Data Bus

E' detto parallelismo del Data Bus il numero di bit trasferiti in parallelo fra CPU e memoria di lavoro attraverso il Data Bus, cioè il suo numero di fili. Più il Data Bus è "largo", maggiore è il numero di bit che possono essere trasferiti alla CPU con un singolo accesso alla memoria.
Abbiamo avuto modo di capire come l'accesso alla memoria sia uno fra i fattori più importanti che limitano la velocità dei computer, per cui è chiaro che la tendenza dello sviluppo dei computer è sempre stata quella di aumentare il parallelismo del Data Bus.
Per esempio si può citare una recente versione IBM della CPU PowerPC, destinata ad essere usata nei mainframe. Essa ha un data bus di 128 byte. Per le comunicazioni con la memoria cache (vedi dopo per la definizione di cache) il parallelismo di quel bus specializzato è addirittura di 256 Bit!

Parallelismo dell'Address Bus

Quanto più alto è il numero di bit dell'Address Bus tanto maggiore è il numero di diverse locazioni di memoria che si possono indirizzare. Nel corso degli anni il costo per bit della memoria centrale è andato diminuendo ad una velocità che continua a stupire anche chi a queste cose è abituato. Così, se nel 1978, quando l'8086 fu introdotto, il suo Address Bus con parallelismo di 20 bit, che gli permetteva una massima memoria di 1 MByte sembrava esageratamente vasto, oggi si comincia a considerare insufficiente la quantità di memoria indirizzabile da un 386 - Pentium, che ha un Address Bus di 32 bit (4 GByte di memoria principale)!
Quindi anche la dimensione dell'Address Bus è andata aumentando nel corso del tempo, fatto non giustificato direttamente dall'aumento delle prestazioni dei computer, ma piuttosto dalla comodità di avere molta memoria.
Il parallelismo dei bus viene indicato negli schemi con una freccia che punta al numero di linee del bus (vedi Figura 2).

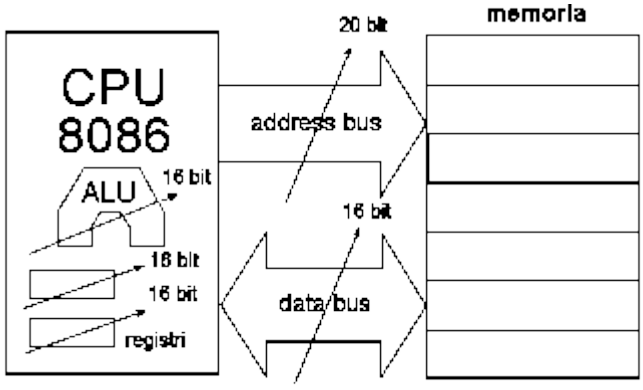


Figura 2: parallelismo nel microprocessore 8086

Non ha senso definire un parallelismo del control "bus", dato che è costituito da tanti fili "singoli" ove le informazioni non viaggiano in parallelo.

Parallelismo della CPU

Il parallelismo della CPU è il numero di bit che essa può elaborare in parallelo. Come tale corrisponde grossomodo al parallelismo della ALU, anche se nelle CPU moderne, che hanno diverse ALU, la cosa non è più automatica.
Il parallelismo della CPU corrisponde anche alla dimensione dei suoi registri generali. L'aumento del parallelismo della CPU dà luogo ad un grande aumento di prestazioni anche se richiede moltissimi transistor in più sul chip. Nel corso del tempo è quindi stato un fenomeno lento ma costante e si è passati dai primi microprocessori da 4 od 8 bit a 16, 32, e recentemente a 64 bit.
Il parallelismo della CPU è il parametro di cui si parla quando si dice che "il 6502 è un microprocessore da 8 bit," o il 68000 è da 16 bit" o anche "Alpha è da 64 bit".

Nome del µprocessore	Anno	Parall. CPU (n. bit)	Parall. Address Bus	Parall. Data Bus	Computer che lo ha usato
4004	1971	4	12	4	Calcolatrici elettroniche (Prima CPU single chip)
8080	1974	8	16	8	Altair 8800 (il primo Personal Computer)
6502	1975	8	16	8	Apple][, Commodore PET e C64
Z 80	1976	8	16	8	Personal computer CP/M
68000	1979	32	24	16	Apple Macintosh, Commodore Amiga, Olivetti M20
8086	1978	16	20	16	IBM PC XT (8088), Olivetti M24
80286	1982	16	24	16	IBM PC AT
80386 - 486 - Pentium	1985 (386)	32	32	64	PC compatibili IBM, workstation non PC
MIPS	1986	32 poi 64			
SPARC UltraSPARC	1987	32 64			
PowerPC	1990	32 poi 64		da 32 a 128	
Alpha	1992	64			
Itanium	2000	64			

Tabella 1: parallelismo di alcuni microprocessori

La vera definizione di "word"
Un tempo era chiamato "word" il numero di bit in parallelo che possono essere elaborati da una CPU. Così esistevano computer con la "parola" da 8 bit, altri da 16, 32 o da 64. Ancora oggi la definizione dovrebbe essere questa, ma nel frattempo le cose sono cambiate. Come abbiamo già illustrato i registri generali e la ALU dell'8086 funzionavano 16 bit. Per cui per l'8086 una "word" era da 16 bit, ed erano chiamate istruzioni su word quelle da 16 bit. Per compatibilità con il codice esistente quelle istruzioni sono state mantenute anche nelle CPU successive, che pure potevano elaborare numeri da 32 bit. Per cui nel mondo 80x86 la parola "word" ha cominciato ad assumere il significato di "numero di 16 bit". La "devastante" supremazia di mercato delle CPU 80X86 ha fatto sì che l'uso si generalizzasse ed ora quando si dice "word" si intende quasi sempre 16 bit. In alcuni casi si è invece mantenuta la vecchia definizione, come nelle CPU PowerPC in cui la word ha 32 bit (peraltro attualmente la prima definizione non vale più neanche per il PowerPC, perché le prime CPU PowerPC lavoravano a 32 bit e la loro "parola" era giustamente di 32 bit, ma le nuove versioni sono a 64 bit ma, per compatibilità, la nomenclatura parla ancora di "word" a 32 bit).

1.1.3 Microprocessori

Sia la CPU che la memoria possono essere costituiti da uno o più circuiti integrati. Al giorno d'oggi la quasi totalità delle CPU è realizzata su un solo componente, mentre la memoria di solito è fatta con diversi chip.
Una CPU realizzata in un solo circuito integrato viene detta "**microprocessore**" (microprocessor).
La prima CPU "on chip", fu l'Intel 4004, un circuito integrato da 2 300 transistor costruito nel 1972 per equipaggiare una calcolatrice tascabile. Per contro le CPU attuali ospitano decine di milioni di transistor.
Al giorno d'oggi i microprocessori integrano anche dispositivi che un tempo non ne facevano parte, come per esempio le unità a virgola mobile.

Famiglie di microprocessori

Le famiglie di CPU più "famose" sono quelle indicate in tabella:

Famiglia	Anno	Computer ove è usata	Sistemi Operativi tipici
80X86, detta anche X86	1978	PC IBM e compatibili	MS-DOS, Windows (tutti), Unix, Linux
68000, detta anche 68k	1979	Tutti i primi Apple Macintosh, Amiga	Apple OS
PA-RISC	1986	Workstation HP	HP-UX (Unix HP), Windows NT
MIPS	1986	Workstation Silicon Graphics	IRIX (Unix Silicon Graphics)
SPARC	1987	Workstation Sun	Solaris (Unix Sun), Linux
Alpha	1992	Workstation e minicomputer Digital	Digital Unix, VMS, ~~Windows NT~~, Linux
PowerPC	1993	Macintosh, IBM RISC, mainframe IBM	Apple OS, AIX (Unix IBM), OS/2, Linux, OS/36
Itanium	2000	PC compatibili	Windows

Tabella 2: alcune famiglie di CPU

1.1.4 Microcontrollori

I microcontrollori sono CPU per scopi speciali, prodotti per finire nel sistema di controllo di un dispositivo "intelligente", come l'ABS di un'automobile, il sistema di controllo di una macchina fotografica o un regolatore di temperatura. Oltre alla CPU integrano svariati dispositivi di I/O, alcuni di essi possono essere considerati dei veri i propri "computer on chip", cioè un solo dispositivo di silicio che contiene tutti i sottosistemi necessari a fare un computer. Nella letteratura tecnica i microcontrollori sono spesso chiamati "MCU" (**M**icro**c**ontroller **U**nit)

I produttori di questi circuiti possono integrare nello stesso chip, oltre alla CPU, altri dispositivi di I/O, che possiedono in "librerie" nei loro sistemi CAD (Computer Aided Design) e possono sviluppare rapidamente in una miriade di versioni, con combinazioni diverse di contatori, memoria RAM o EPROM, convertitori A/D e D/A, interfacce seriali, sistemi di I/O digitale, dispositivi di comunicazione ed altri. Volendo, ed essendo in grado di ordinare milioni di unità, il produttore può realizzare "su ordinazione" un microcontrollore, includendo la CPU ed i dispositivi scelti dal cliente.

Con la miriade di transistor che la tecnologia attuale mette a disposizione a costi sempre più bassi la distinzione fra microcontrollore e microprocessore general purpose è sempre meno definita, dato che anche i microprocessori tradizionali incorporano sempre più dispositivi di I/O.

A volte ritornano

Diversi microprocessori del passato, quando non hanno più trovato applicazione come CPU per i personal computer, hanno avuto una seconda vita come parte nevralgica ("core") di microcontrollori. Oggi anche le CPU fanno parte delle librerie dei produttori di microcontrollori che le possono aggiungere con molta rapidità ai loro prodotti. Molti microcontrollori sono perciò compatibili con i linguaggi macchina di CPU importanti del passato. In questo modo i progettisti di software hanno continuato ad utilizzare le loro conoscenze riguardo alla CPU anche per i sistemi embedded. Si possono portare tre esempi nominando lo Z 80, il 6502 ed anche l'8086 nella sua versione 80186.
Segue una tabella che indica alcune famiglie di microcontrollori tipici:

Microcontrollore	Parallelismo. CPU	Address bus (bit)	Data bus	Note
8051	8	16	8	Basato su accumulatore, architettura Harvard
Z - 180	8	16	8	Compatibile con Z - 80
68HC11	8	16	8	Compatibile con 6800
65C02	8	16	8	Compatibile con 6502
ST6	8	16	8	Basato su accumulatore
ST9	8	22	8	
PIC 16/17	8	16	8	Basato su registri, architettura Harvard
80186	16	24	16	Fa parte della famiglia X86
STPC	32	32	64	PC con 586 "on chip", completo di I/O
!!!!				Playstation
!!!!				Playstation 2
MIPS	64			Nintendo 64
i960				
ARM	32			Architettura RISC load/store

Tabella 3: alcune famiglie di microcontrollori

1.1.5 Clock

Due CPU identiche, con gli stessi transistor disposti nello stesso modo, ma con frequenza di clock diversa avranno velocità di elaborazione diversa. Se una delle due ha frequenza di clock doppia dell'altra andrà a velocità doppia.
Naturalmente questo richiederà memorie più veloci, e quindi il computer in cui viene impiegata la seconda CPU è più costoso, fino al limite per cui non esistono memorie abbastanza veloci.
Si capisce perciò come la tendenza delle case costruttrici nel corso del tempo sia stata quelle di aumentare la frequenza di clock alla quale i nuovi microprocessori potevano lavorare. Naturalmente questo significa chiedere prestazioni elevate anche a tutto il resto del computer, per questo si scelgono soluzioni di compromesso e si realizzano clock diversi per i diversi dispositivi che sono all'interno del computer.

I clock di un computer

Nei computer possono essere presenti diversi segnali di clock. Alcuni possono essere a frequenze basse, per far funzionare orologi e/o contatori utili al sistema. Altri a frequenze non molto minori di quella della CPU, per sincronizzare i dispositivi sul bus di espansione, altri alla frequenza che viene portata anche alla CPU, per l'accesso alla memoria principale e come segnale per la sincronizzazione delle attività interne alla CPU. Per quest'ultimo scopo moltissime CPU odierne usano una frequenza ancora più alta, come spiegato nel prossimo paragrafo.

Moltiplicazione della frequenza di clock nelle CPU

La tecnologia della moltiplicazione del clock nacque quando fu proposta la CPU 486 DX2 che doveva sostituire microprocessori di tecnologia precedente (486DX) sulla stessa scheda madre. Perché questo fosse possibile le due CPU dovevano apparire del tutto identiche ai loro piedini esterni. Per funzionare più velocemente le CPU della serie DX2 includevano un moltiplicatore di frequenza. La CPU più moderna riceveva in ingresso il vecchio segnale di clock e lo moltiplicava internamente per due o per tre. In questo modo internamente funzionava più velocemente, mentre esternamente era identica alla precedente. Naturalmente ogni accesso ai bus esterni era esattamente identico a prima, per cui se questa CPU era due o tre volte più veloce nei calcoli, non altrettanto si poteva dire dell'accesso alla memoria.
Questo tipo di approccio a moltiplicazione di frequenza al giorno d'oggi è generalizzato e viene adottato da molte delle CPU più potenti, che funzionano internamente a frequenze di clock molto più alte di quelle con cui si presentano all'esterno.

1.1.6 Memoria

Trattiamo ora in un certo dettaglio i vari tipi di memoria presenti in un computer, integrando le informazioni date nel precedente capitolo.

Memoria principale

Dimensione della locazione

Nonostante il fatto che al giorno d'oggi ogni trasferimento fra memoria e CPU sia almeno a 32 bit, con casi da 128 bit, l'unità minima di informazione cui si ha accesso con un indirizzo (cioè la locazione), rimane sempre il byte, come ai tempi del 6502.

L'unica spiegazione plausibile è legata alla memorizzazione delle stringhe. Esse sono ancora memorizzate come sequenze di codici ASCII o ANSI a 8 bit, che risiedono in locazioni successive di memoria. Dato che è spesso necessario utilizzare un carattere qualsiasi di una stringa, fa comodo poter indirizzare individualmente un qualsiasi byte della memoria.

kappa, Mega, Giga

Nel gergo informatico è invalso l'uso di prefissi per indicare le quantità di memoria.
Un kByte è il numero di byte indirizzabile da una CPU che abbia un Address bus di parallelismo 10 bit, perciò corrisponde a $2^{10} = 1024$ ed è chiamato k per via del fatto che è "simile" al fattore 1000, che ha prefisso k nel Sistema Internazionale di unita' di misura. Esempio: il 6502, microprocessore del Commodore 64, aveva 16 bit d'indirizzo. La memoria massima che la CPU poteva indirizzare era di 64 kByte, che fra l'altro era tutta presente sulla macchina e che diede il nome al computer.
1 MByte (megabyte) è il numero di byte indirizzabile da una CPU che abbia 20 bit di Address bus (come per esempio l'8086 del primo PC), perciò corrisponde a $2^{20} = 1024 * 1024 = 1\ 048\ 576$. Dunque l'8086 e l'8088 potevano avere al massimo 1 MByte di memoria centrale, invece l'80286, che aveva 24 bit di indirizzo, poteva arrivare a 16 MByte.
1 GByte (gigabyte) corrisponde a 30 bit di indirizzo, cioè $2^{30} = 1024 * 1024 * 1024 = 1\ 073\ 741\ 824$. Il 386, che ha indirizzi a 32 bit, può avere al massimo 4 GByte di memoria centrale.
Continuando a muoversi per multipli di mille, come nel Sistema Internazionale, e di dieci bit in dieci per l'indirizzo, si ha: 2^{40}, che è un Tera (T). Esempio: l'indirizzo virtuale di un 386 è di 46 byte, perciò un 386 può avere al massimo 64 TByte di memoria virtuale (cosa sia la memoria virtuale non è tempo di spiegarlo).

2^{50} è un Peta (P); 2^{60} corrisponde a un EByte (esabyte). Una CPU che abbia indirizzi di 64 bit può indirizzare al massimo 16 EByte di memoria.
Questo uso dei prefissi è piuttosto scorretto dal punto di vista formale perché così in campo informatico si usano gli stessi simboli per quantità diverse rispetto a tutte le altre unità di misura.
Prima o poi le organizzazioni di standardizzazione internazionale realizzeranno norme che imporranno l'utilizzazione dei prefissi basati sui multipli di 1000, come con tutte le altre grandezze, cioè k = 1000, M = 10^6, G = 10^9, T = 10^{12}, P = 10^{15}, E = 10^{18}. Quando questo succederà invece di dire che un Mega è poco più di un milione, diremo che con 20 bit di può coprire un po' più di un Mega, che sarà di nuovo, come era sempre stato anche prima, esattamente un milione.
Quando questo accadrà è difficile dirlo, anche probabilmente ci vorrà parecchio tempo.
Attenzione: gli Americani, che, come si sa, con le unità di misura non ci prendono molto, non stanchi di aver inventato questo pasticcio, usano spesso Mb per indicare megabit e MB, con la B grande, per indicare megabyte. La sola differenza fra maiuscolo e minuscolo è molto spesso del tutto ambigua, e porta facilmente ad errori di 8 volte. In questo testo non si vedrà più né Mb né MB, ma solo Mbit o MByte.

Memorizzazione dei numeri

La gran parte delle CPU ha locazioni di 8 bit, mentre il parallelismo del data bus è di solito superiore. Se con una CPU si possono trasferire numeri più grandi della dimensione della locazione è necessario che il numero sia "spezzato" in diverse locazioni contigue di memoria.
Per tornare al modello della cassettiera, illustrato nel capitolo precedente, è come se un singolo cassetto, accessibile usando un solo indirizzo, contenesse due o più compartimenti. La Figura 3 illustra questo modello e mette anche in evidenza la caratteristica della CPU 8086 di poter fare trasferimenti a 8 o a 16 bit.

Al riguardo possiamo già anticipare qualcosa, con l'avvertenza che per tutto il dettaglio si deve attendere i capitoli sull'Assembly.
Nella figura sono illustrate due istruzioni dell'8086; quella a sinistra trasferisce nel registro AH il contenuto di una sola locazione di memoria mentre quella a destra trasferisce nel registro a 16 bit di nome AX due locazioni in un solo accesso. Si noti che le parentesi quadre racchiudono sempre un indirizzo, mentre ciò che "fuori" dalle parentesi quadre è il dato contenuto nel "cassetto".

<FILE>
dueposti.FH5
</FILE>

Figura 3: cassetti a due posti

Dunque le CPU, pur avendo locazioni di 8 bit, possono trasferire ed elaborare numeri binari più grandi in una singola operazione. Questi numeri devono per forza essere spezzati in diverse locazioni.
Perché le cose funzionino bisogna che la CPU usi una convenzione coerente per sapere con che ordine quel numero viene spezzato, cioè se le parti basse dei numeri grandi vanno memorizzate agli indirizzi alti o a quelli bassi.

CPU "little endian" e CPU "big endian"

Le CPU "**little endian**" memorizzano agli indirizzi bassi i byte bassi dei numeri; le CPU "**big endian**" fanno il contrario. Si dice big endian perché la "fine grande" (the big end), cioè la parte alta del numero, viene per prima in memoria (the big end comes first).
La famiglia X86 è little endian, ma esistono processori significativi (Motorola 68k) che sono big endian.
Diverse CPU moderne possono essere impostate per memorizzare i numeri come vuole il programmatore (PowerPC, Alpha

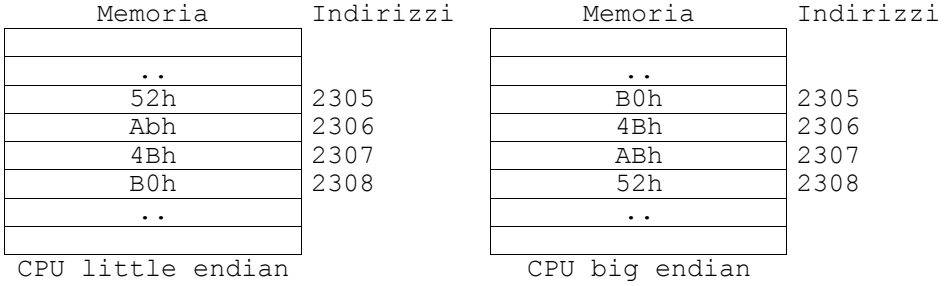


Figura 4: ordine di memorizzazione del numero B04BAB52h con CPU little endian o big endian

Nella figura precedente la memoria ha locazioni di 8 bit; il numero di 32 bit B04BAB52h parte in entrambi i casi dall'indirizzo 2305, ma nel caso little endian ai primi indirizzi corrispondono i byte bassi del numero, il contrario accade con la CPU big endian.

Alcuni strumenti di sviluppo, come i debugger, fanno vedere il contenuto della memoria con i numeri esadecimali, mettendoli in fila da sinistra a destra all'aumentare del loro indirizzo.
Uno strumento del genere visualizzerebbe il numero della Figura 4 in questo modo:

```
2305     52 AB 4B B0  , per una CPU little endian

2305     B0 4B AB 52  , per una CPU big endian
```

Allineamenti in memoria
Il fatto che la dimensione della locazione sia minore del numero di bit trasferibili con una sola istruzione può causare un altro problema, detto di "**allineamento**" (alignment).
Spieghiamoci con un esempio. Supponiamo che l'istruzione:

MOV AX, [315]

Carichi dalla locazione di indirizzo 315 un numero a 16 bit o lo memorizzi nel registro AX. Alcune CPU, come per esempio l'8086, possono richiedere due operazioni in memoria per caricare quel numero, perché nella prima caricano 314 e 315, scartando 314 e nella seconda 316 e 317, scartando 317.
Ciò accade perché la lettura a 16 bit non è "allineata". Per effettuare il trasferimento con un solo accesso alla memoria l'indirizzo avrebbe dovuto essere allineato alla parola ("word aligned"), cioè essere un numero pari.
Generalizzando il concetto, si può dire che succede spesso, ed in tutte le CPU, che se l'indirizzo d'inizio di un numero più grande della locazione non è allineato l'esecuzione dell'istruzione sia inefficiente; alcune CPU, come per esempio quelle della famiglia 68000, vanno in condizione d'errore se l'accesso alla memoria non è allineato.
Da quanto detto si capisce che si possono ottenere aumenti di prestazioni anche notevoli semplicemente allineando i dati in memoria come la CPU usata li preferisce.
Gli allineamenti tipici sono l'allineamento "a word", in cui l'indirizzo d'inizio di un numero deve essere multiplo di 2 (deve essere pari, ossia avere 0 nel bit meno significativo), l'allineamento "a double word", per numeri a 32 bit, nel quale l'indirizzo deve essere multiplo di 4, cioè avere le ultime due cifre uguali a zero, l'allineamento "a quad word", per numeri di 64 bit, con ultime tre cifre zero (multiplo di 8).
Altri tipi di allineamento sono specifici dell'architettura della singola CPU (p.es. l'allineamento a "paragrafo" dell'8086) e hanno i nomi e le lunghezze che l'architettura di quella specifica CPU impone (p.es. l'allineamento a word è diverso nelle famiglie X86 e PowerPC perché è diversa la definizione di "word" nelle due architetture).

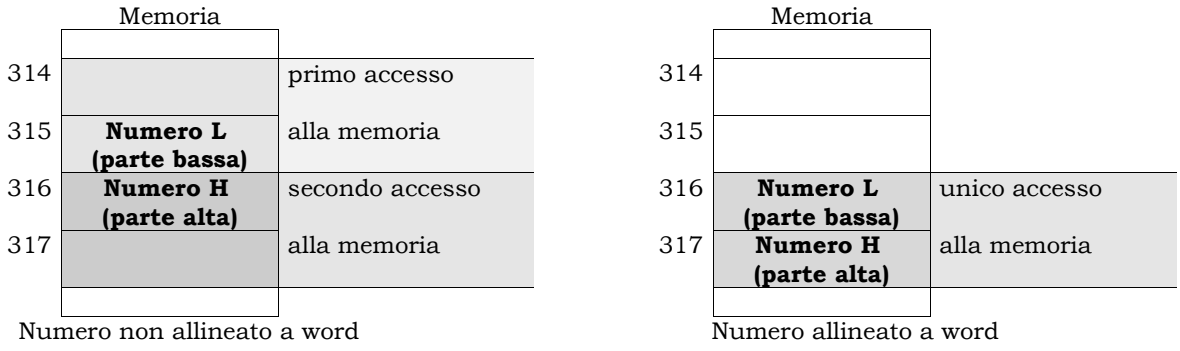


Figura 5: accesso a numeri di 16 bit allineati a word o no

1.1.7 Input/Output

Processori di I/O

Come abbiamo visto tutte le periferiche e gli altri dispositivi collegati ad un computer si "mascherano" da memorie per poter comunicare con la CPU. Spesso fra la CPU ed il dispositivo c'è una vera e propria CPU specializzata, che solleva quella del computer da elaborazioni che hanno a che fare con l'I/O su quel dispositivo.
Possiamo chiamare questi computer dedicati "**processori di I/O**". Un dispositivo che contiene un processore di I/O non ha quindi solo la funzione di leggere o scrivere dati esterni per la CPU, ma ha anche vere e proprie funzioni di elaborazione e svolge calcoli che la CPU del computer può evitare di fare. Questo rende il computer, nella sua globalità, più veloce. Spesso i processori di I/O sono detti anche "controller".

```
Possono considerarsi processori di I/O: la CPU contenuta in ogni stampante grafica, una
scheda grafica acceleratrice, i chip per l'elaborazione dei segnali audio nelle schede
sonore, il controller dell'hard disk. Spesso le CPU contenute nei processori di I/O non
sono normali microprocessori general purpose ma chip dedicati, progettati solo per lo
scopo specifico cui sono destinati.
```

1.2 Memorie

Estendiamo ora il concetto di memoria, rispetto a quanto visto in precedenza.
Consideriamo "memoria" un qualsiasi dispositivo o supporto in grado di mantenere per il tempo voluto una sequenza di bit.
Con questa definizione anche un floppy disk o un foglio di carta è una "memoria". Naturalmente per trasferire alla CPU i bit contenuti in un floppy disk essi devono passare per un dispositivo di I/O e per il data bus. Per far ciò devono essere trattati da un programma che controlla tutto il processo della lettura dal floppy.
Ciò non ci impedisce di considerare il floppy disk, corredato dall'hardware e del software per la sua lettura e scrittura, come una memoria.
Estendendo la definizione di memoria è necessario estendere anche quelle di indirizzo e di locazione. "Locazione" è la minima quantità di bit che si ottiene con un singolo accesso a quel tipo di "memoria", "indirizzo" è il numero che si usa per distinguere quella "locazione" dalle altre.
Naturalmente i nomi che vengono dati a queste "locazioni" ed "indirizzi" sono diversi in base al tipo di "memoria".

1.2.1 Accesso sequenziale o accesso diretto

I diversi tipi di memoria hanno diversi modi di leggere e scrivere le informazioni contenute al loro interno.
Alcuni tipi di memorie sono fatte in modo che l'accesso ad una qualsiasi delle informazioni è possibile solo leggendo prima tutte quelle che la precedono fisicamente. Si pensi per esempio ad una cassetta musicale: non si può ascoltare direttamente la terza canzone, ma bisogna prima passare per la prima e la seconda. Le "memorie" che funzionano in questo modo vengono dette "ad accesso **sequenziale**".
In altri tipi di memorie è possibile avere accesso direttamente alle informazioni che desideriamo, senza passare per quelle che non interessano. Come modello possiamo prendere un libro, nel quale possiamo saltare alla pagina voluta senza dover leggere prima le altre. Questo tipo di memorie viene detto "ad accesso **diretto**".
La memoria principale di un computer è necessariamente ad accesso diretto. L'indirizzo è il modo per stabilire quale parte della memoria principale si vuole usare.

1.2.2 Memorie volatili o non volatili

Alcuni tipi di memoria, di solito i più veloci, hanno lo svantaggio di cancellarsi in assenza di alimentazione. Queste memorie vengono dette "volatili" (evaporano, come l'alcool ;-). Le memorie non volatili invece mantengono l'informazione che hanno memorizzata per tempi molto lunghi, anche dopo spente. Una memoria volatile va alimentata in continuazione per tutto il tempo che la si usa.

1.2.3 "Principio" di località

Abbiamo già visto come un programma sia costituito da una serie di istruzioni che vanno eseguite in sequenza e come queste istruzioni siano fisicamente presenti in memoria centrale, in locazioni successive. Questo significa che se si accede in fase di fetch alla locazione 1000 è molto probabile che si debba accedere anche alla locazione 1001. Questo non accade solo per le aree di memoria che contengono istruzioni, ma anche per le aree che contengono dati. I dati infatti sono sempre sistemati uno vicino all'altro ed accade spesso che alle aree di dati si faccia accesso in sequenza.
Quanto detto corrisponde a quello che viene chiamato "principio di **località** dei programmi **nello spazio**", che si può enunciare in modo informale così: "se un programma accede ad una locazione di memoria allora è probabile che debba accedere anche a quelle vicine".
Un altro principio, mai smentito dai fatti, che riguarda il funzionamento dei programmi è il cosiddetto "principio di **località** dei programmi **nel tempo**", che si può enunciare così "se un programma accede ad una locazione di memoria in un dato istante allora è probabile che accederà di nuovo alla stessa locazione fra poco tempo".
La constatazione di queste due verità porta all'invenzione delle gerarchie di memoria.

1.2.4 Livelli di memoria

I vari tipi di memoria si possono organizzare gerarchicamente, in "livelli", per questioni economiche.
La situazione ideale sarebbe memorizzare ogni dato di cui si ha bisogno nel tipo di memoria più veloce che si ha a disposizione, ma poiché la memoria più veloce costa milioni di volte più di altri tipi di memoria, si sceglie di strutturare la memorizzazione delle informazioni su diversi supporti.
I supporti dei primi livelli sono costituiti da tipi di memoria molto veloci e molto costosi, mentre gli ultimi sono lenti e a buon mercato.
In ordine di velocità decrescente possiamo indicare: registri, cache, memoria di lavoro, hard disk, memorie ottiche, floppy disk, nastri.
Possiamo individuare tre tipi di tecnologie per la memorizzazione: le memorie elettroniche, quelle magnetiche e quelle ottiche.

Memorie elettroniche
Le memorie elettroniche sono basate su semiconduttori. Sono le memorie più veloci. Le tecniche utilizzate per la memorizzazione sono diverse, per cui esistono molti tipi di memorie elettroniche.

A seconda del tipo possono essere sia volatili che non volatili. Quasi sempre sono ad accesso diretto, ma esistono anche memorie elettroniche ad acceso sequenziale (shift register).
A seconda del tipo possono essere sia volatili che non volatili, anche se il primo caso è di gran lunga il più frequente. Quasi sempre sono ad accesso diretto, ma esistono anche memorie elettroniche ad accesso sequenziale (shift register).

Memorie magnetiche
Nelle memorie magnetiche la memorizzazione avviene sfruttando la magnetizzazione di un substrato sensibile ai campi magnetici, che viene "spalmato" sul supporto di memorizzazione. Questo materiale si magnetizza come una calamita e può assumere due stati, uno in cui la magnetizzazione è in un verso e l'altro con il verso contrario (come i poli Nord e Sud di una calamita). Ogni piccola area magnetizzabile può perciò contenere un bit d'informazione. Una volta effettuata la memorizzazione essa si mantiene per anni, a patto naturalmente di non esporre il supporto a campi magnetici molto grandi.
Le memorie magnetiche sono sempre non volatili e possono essere ad accesso diretto o sequenziale.

Memorie ottiche
Nelle memorie ottiche le informazioni sono registrate in modo che un raggio di luce sia riflesso o meno. Un sensore elettronico di luminosità rivela la presenza o meno del raggio riflesso, leggendo in questo modo numeri codificati in binario.
Le memorie ottiche sono non volatili e ad accesso diretto.

Gerarchia di memorizzazione
Vediamo ora la gerarchia della memoria in maggiore dettaglio, in ordine decrescente di velocità.

Registri
Dei registri abbiamo già trattato e sappiamo che funzionano alla velocità della CPU. Peraltro i registri sono sempre pochi e non si possono aumentare, a meno di non cambiare CPU. L'indirizzo di un registro è il suo nome od il numero nel banco di registri da cui lo si vuole prelevare.

Cache
La cache è una memoria elettronica simile alla normale memoria centrale, ma costruita con tecnologie che la rendono più veloce.
La ragione dell'esistenza della cache sta nel principio di località.
Infatti nella cache si memorizzano le ultime locazioni cui la CPU ha acceduto ed i loro "dintorni".
Ogni volta che la CPU deve accedere ad una locazione di memoria controlla prima se essa è già presente nella cache. Se la locazione c'è ne legge il contenuto dalla cache, che è molto più veloce; se non c'è fa accesso alla memoria normale, e trasferisce il contenuto di quella locazione anche nella cache. Poi, nei momenti morti in cui non c'è bisogno di usare la memoria, provvede a caricare nelle cache anche le locazioni vicine a quella appena usata, in modo che se la CPU ne ha bisogno le trovi già nella cache.
Naturalmente ogni volta che scrive qualcosa nella cache dovrà sovrascrivere qualche informazione che c'era prima. La gestione di una cache è quindi un processo complicato sul quale per ora non è il caso di andare più a fondo.
Le CPU moderne hanno molti transistor da utilizzare e qualche decina di migliaia da dedicare alla gestione delle cache si trovano sempre.
La presenza di una cache contribuisce così tanto a far migliorare le prestazioni dei computer che viene a sua volta divisa in livelli.
Vari livelli di cache

In una cache organizzata per livelli la CPU cerca il contenuto della locazione che interessa prima nelle memorie più veloci, poi, via via in quelle più lente.
Quindi il primo posto ove la CPU cerca le informazioni è la cache di primo livello. Se non trova la locazione che cerca, prova nella cache di secondo livello, poi in quella di terzo livello; infine cercherà nella memoria centrale.
La giustificazione di questa complicazione risiede nel fatto che la cache è tanto più veloce quanto è più piccola, per cui la cache di primo livello si trova di solito dentro lo stesso chip della CPU e non può contenere molte informazioni.
La cache di secondo livello può essere "on chip" o su un circuito integrato separato. In alcuni casi è su un chip separato saldato al microprocessore ed incluso nello steso contenitore ("package"). (esempio: Pentium Pro).
La cache di terzo livello è una normale memoria statica, più veloce delle memorie dinamiche che costituiscono la memoria centrale, ed è sempre separata dalla CPU e posizionata sulla scheda madre.

Tutte le tecnologie di cache elettroniche sono volatili.

Il concetto di cache è estensibile a tutti i tipi di memoria, al giorno d'oggi non riguarda quindi solo le memorie elettroniche.

Memoria principale

E' costituita per la gran parte di RAM dinamica. Qui non si vuole entrare nei dettagli elettronici e si fa solo notare che il nome RAM ("**R**andom **A**ccess **M**emory") è improprio. Questo tipo di memoria infatti non fa accesso a casaccio (random = casuale). Il nome "memoria ad accesso diretto" sarebbe più adatto per indicare il fatto che l'acceso non è sequenziale.
La RAM è volatile.
Per memorizzare il primo programma che viene eseguito dal computer all'accensione deve esistere un tipo di memoria centrale non volatile. In realtà ne esistono diversi tipi. Dato che qui non è il caso di entrare nei dettagli elettronici, elencheremo semplicemente i tipi, con un breve commento:
ROM (**R**ead **O**nly **M**emory), memoria a sola lettura, prodotta in fabbrica già scritta come vuole il cliente
PROM (**P**rogrammable **ROM**), memoria a sola lettura, scritta una sola volta direttamente dall'utente
EPROM (**E**rasable **PROM**), memoria elettronicamente a sola lettura, cancellabile mediante esposizione a raggi ultravioletti (il chip va tolto dalla scheda in cui si trova). Può essere programmata e cancellata direttamente dall'utente. Il numero di cancellazioni è limitato.
Le EPROM si distinguono perché hanno una finestra trasparente, attraverso la quale è visibile il chip, che serve per la cancellazione. Nelle apparecchiature "finite" la finestra è solitamente coperta con un adesivo.
EEPROM (o E^2PROM: **E**lectrically **E**rasable **PROM**), memoria non volatile a lettura e scrittura, cancellazione con tecnologia elettrica, senza rimuovere la memoria dal circuito in cui è impiegata. Scrittura molto più lenta della lettura. Numero di scritture limitato. Le EEPROM più moderne: "flash memories" (più veloci e costose delle altre EEPROM).

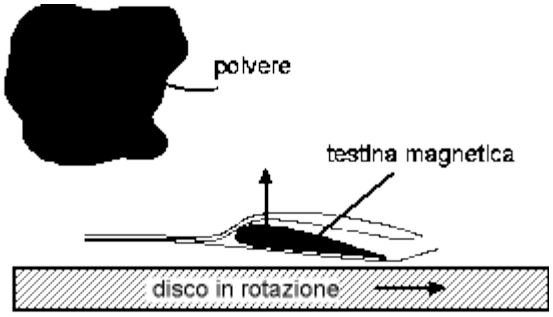
Memorie di massa

Le memorie non elettroniche vengono anche chiamate "memorie di massa" (mass storage). Fin dall'inizio della storia dei computer sono state tipicamente in tecnologia magnetica. In tempi relativamente recenti si è assistito all'ingresso di dispositivi in tecnologia ottica o mista (magneto - ottica).
Il nome di questo tipo di memorie viene dal fatto che, essendo il loro costo per bit molto più basso di quello delle memorie elettroniche, sono sempre state usate per memorizzare, a bassa velocità ed a basso costo, grandi "masse" di dati. Dato che i dati da memorizzare sono molti non è pensabile perderli tutti allo spegnimento del computer. Perciò le tecniche per le memorie di massa sono sempre non volatili.
Di seguito vengono presentati i diversi tipi di memorie di massa, dalla più veloce alla più lenta.

Hard disk

L'hard disk è una pila di piatti metallici posti in rapida rotazione. Sui dischi viene depositato un sottilissimo strato di materiale magnetico.
Una o più testine magnetiche, in grado di leggere e scrivere le informazioni binarie sul disco, si possono muovere radialmente (lungo un raggio del piatto). In questo modo si può accedere ad informazioni situate a diverse distanze dal centro del disco.
<FILE>
disco.FH5
</FILE>
Figura 6: (a) hard disk con 8 piatti (b) piatto con due testine
Le informazioni sono memorizzate magneticamente in cerchi concentrici, detti "cilindri" (cylinder) o tracce, spesso un bit per ogni piatto. In questo modo è possibile leggere o scrivere tanti bit contemporaneamente quanti sono il doppio dei piatti (possono essere utilizzate entrambe le superfici).
Ogni cilindro è suddiviso in "settori", ciascuno dei quali contiene delle informazioni per la sincronizzazione, cioè una serie di bit, scritti sulla superficie del disco, che indicano all'elettronica di lettura e scrittura quando è il momento per leggere informazioni significative. Per creare i settori sulla superficie del disco si deve eseguire quella che si chiama "formattazione a basso livello". Attualmente la formattazione a basso livello viene fatta direttamente dal produttore.
I settori vengono raggruppati dal software di sistema (Sistema Operativo) in "**cluster**" (significa "mazzo" o "grappolo", ma non viene mai tradotto). Il Sistema Operativo del computer numera i cluster cominciando da zero ed usa il numero di cluster come "indirizzo" delle informazioni contenute nell'hard disk. La formattazione ordinaria, come potrebbe essere FORMAT C: in MSDOS, crea i cluster sul disco rigido.
L'hard disk è un sorprendente pezzo di tecnologia meccanica che si è evoluto ad una velocità che non ha nulla da invidiare con le sempre incredibili velocità dell'innovazione elettronica. Le sue due specifiche più importanti sono la velocità di accesso e la capienza, cioè il numero di byte che vi si possono memorizzare. La rapidità dell'evoluzione è dovuta anche al fatto che, come nella microelettronica, la miniaturizzazione dà vantaggi sia di velocità che di densità di memorizzazione. Se l'area in cui si memorizza un bit è più piccola, nello stesso disco potranno stare più dati ma contemporaneamente, se ogni bit prende meno posto, a parità di velocità di rotazione del disco nello stesso tempo passano più bit.
E' anche chiaro che più veloce gira un hard disk più rapido sarà l'accesso alle informazioni in esso memorizzate. Questo ha portato all'aumento delle velocità di rotazione ed ai problemi tecnologici ad essa correlati, quali la resistenza meccanica e la deformabilità dei piatti dovuta alla forza centrifuga. L'altro parametro che influenza la velocità di accesso è la rapidità con cui la testina si muove fra un cilindro e l'altro.
Un'altra specifica che richiede alta tecnologia è la densità di memorizzazione dell'informazione. Essa dipende dalla sensibilità della testina ai campi magnetici e dalla sua vicinanza alla superficie. Nelle testine sono stati sperimentati materiali che presentano effetti fisici strani, l'ultimo è la "magnetostrittività" (yuk! 8-), che permettono di realizzarle molto

piccole e sensibili. Perché la testina possa essere molto vicina alla superficie essa viene sagomata con una forma aerodinamica in modo che l'aria che le soffia intorno (in verità è il disco che gira, ma tutto è relativo), dia un effetto portante che fa alzare la testina alla giusta altezza, come un aereo. Quest'altezza è così piccola che un granello di polvere sarebbe come un enorme masso e se dovesse impattare con la testina potrebbe distruggerla.



Lo stesso effetto avrebbe una rugosità o residui di lavorazione sulla superficie dei piatti, che sono gli oggetti con la più sofisticata lavorazione meccanica che esistono. Ovvio che la produzione avviene in ambienti di camera bianca, ove l'abbattimento delle polveri deve essere estremo.
All'interno di un hard disk viene creata un'atmosfera rarefatta, la pressione è bassa per diminuire gli attriti ed aumentare la velocità, ma non è un vuoto spinto, per poter far sollevare la testina.
Data la grande velocità di rotazione (da 3000 a 10000 giri/min) e la necessità di lavorazioni superficiali molto fini, il materiale di cui è costruito un hard disk deve essere rigido. Ecco il perché del nome "disco rigido", anche in contrasto con il "disco moscio", floppy disk.

CDROM

Le tecnologie ottiche di memorizzazione originano dallo sforzo tecnologico che l'industria dell'elettronica di consumo profuse per la realizzazione dell'audio digitale e che si concretizzò nel Compact Disk (CD). Alla sua uscita il CD era un esempio di tecnologia sofisticata a basso costo, ed aveva molte caratteristiche che a quel tempo erano assolutamente allo "stato dell'arte".

Il suono è registrato in un CD audio in forma numerica. Un lettore di CD audio deve perciò comprendere un convertitore digitale – analogico dato che il segnale sonoro finale è di tipo analogico, mentre la sua memorizzazione avviene in forma digitale.

Dopo diversi anni dallo standard per i CD audio venne emanata una normativa anche per i CD dati (1984), che prevedeva la sola lettura del supporto. Per questo i dischi CD dati vennero chiamati CDROM.
Un CDROM può contenere 640 MByte di dati. La norma ISO 9660 stabilisce quale deve essere il formato dei dati memorizzati su un CDROM.
I dati in un CD audio sono memorizzati in una traccia a spirale, analoga a quella dei dischi audio in vinile. In questo sono diversi dagli hard disk, nei quali le tracce sono circolari. I lettori per CDROM sono compatibili con i precedenti CD audio, per cui le tracce sono a spirale.
Un CD è costituito da un supporto in materiale plastico trasparente su un lato del quale è incollata una pellicola di materiale metallico, sulla quale vengono memorizzati i numeri. La superficie della pellicola può essere divisa in miliardi di piccoli punti, che possono memorizzare un bit ciascuno. La memorizzazione è definitiva e può avvenire con grande densità.
La memorizzazione avviene facendo in modo che segmenti di traccia riflettano una luce infrarossa, di una lunghezza d'onda 780 nm, emessa da un laser. Altri tratti della traccia invece assorbono quella luce, rendendo possibile la codifica binaria di numeri.
Le parti della traccia in cui il substrato è riflettente vengono detti (????land o pit????), quelli in cui è assorbente (????pit o land????)

La codifica effettiva dei numeri è un po' strana. Il valore del bit a 1 è rappresentato da una transizione, sia da pit a land che da land a pit, mentre la lunghezza dei segmenti di pit o di land indica il numero degli zero.

Per molto tempo l'unico modo di produrre i CDROM era lo stesso usato per i CD audio: produzione di una copia "master", con apparecchiature del costo di milioni di EUROe riproduzione dal master di molte copie da pochi centesimi l'una.
In tempi recenti è divenuto possibile realizzare CDROM con apparecchiature dal prezzo abbordabile per tutte le aziende ed anche per le persone. Dato che questi dispositivi producono una copia master, vengono detti "masterizzatori".
Un disco ottico scrivibile viene detto CD-R, ove R sta per "recordable" (registrabile).
Il laser di un masterizzatore può funzionare con due potenze. Alla potenza minore funziona come lettore, alla potenza maggiore è in grado di degradare la superficie di un substrato dorato, che si deteriora in modo irreversibile nei punti ove

il laser viene "sparato" con maggiore potenza. In questo modo vengono "scavati" i ????pit o land???, nei punti ove il laser ha operato a maggiore potenza la superficie non è più riflettente.
In tempi recenti sono stati introdotti CD dati riscrivibili (CD-RW). In questo caso il laser può funzionare ad una terza potenza, intermedia, alla quale è in grado di cancellare il supporto, che poi può essere riscritto rimettendo il laser alla sua massima potenza. Lo sforzo successivo dell'industria dell'elettronica di consumo è stato la realizzazione di dischi che contengano, in dischi della stessa dimensione del CD, sia audio che video. Dopo una lunga battaglia sugli standard, si è giunti alla definizione del DVD (Digital Versatile Disk), che esiste sia in forma ROM che RAM. I DVD-RAM possono contenere fino a 5,2 GByte, mentre i DVD-ROM 17 GByte. Esistono diversi formati, a singola e a doppia faccia, che hanno diverse capacità.
Per soddisfare le necessità di memorizzazione di quantità di dati molto grandi esistono i "juke box" di CDROM (o di DVD), che possono contenere centinaia di CD caricabili in modo totalmente automatico, per cui è possibile l'accesso "in linea" a molti TByte di dati.

Floppy disk
Il nome di questo dispositivo viene dal fatto che il supporto magnetico è di plastica, ed è morbido. Viene stampato e rivestito di materiale magnetico senza subire altre lavorazioni e costa quindi molto poco. La velocità di rotazione di un floppy disc è di 360 giri/min. E' il classico supporto per lo scambio dei dati fra i personal computer. Rimane un po' come un dinosauro vivente, dato che per oltre dieci anni non ha subito evoluzioni tecnologiche. Oggi l'unico formato per i floppy disk usato estesamente è quello che permette di memorizzare 1,4 MByte di dati.

Esistono diversi altri tipi di dischi, che non hanno raccolto consenso oppure che sono stati venduti da un solo produttore e quindi non sono divenuti standard.
I più rilevanti sono i floppy "grossi" come lo Zip (capacità 120 MByte) o gli hard disk rimovibili come i JAZ (capacità 2 GByte).

Nastri
I nastri magnetici sono stati la memoria di massa dei primi computer. Hanno il grande svantaggio dell'accesso sequenziale; per leggere un'informazione qualsiasi è necessario leggere prima tutte quelle che la precedono fisicamente sul nastro. Il costo per bit dei nastri è ancora il più basso, per cui vengono ancora usati, soprattutto per il salvataggio di sicurezza dei dati degli hard disk, operazione che viene detta "**backup**".
Spesso i nastri per il backup vengono detti "**data streamer**" ed hanno molti formati, come per esempio il DAT, anch'esso di provenienza audio digitale (infatti la sigla significa Digital Audio Tape).

La parola "**stream**" si ritrova spesso in Informatica; letteralmente significa "corrente" (di un corso d'acqua), in gergo si usa per indicare un insieme di dati che fluisce continuamente, senza interruzioni significative, da una sorgente ad una destinazione.

Livello	1	2	3	4
Nome	Registri	Cache	Memoria principale	Dischi
Tempo di accesso	2 - 5 ns	3 - 10 ns	8 - 400 ns	5 ms
Velocità di trasferimento (GByte/s)	4 - 32	0,8 - 5	0,4 - 2	0,0004 - 0,032
Gestito da	Compilatore	Hardware	Sistema Operativo	S.O. o utente finale
Ricopiato in:	Cache	Memoria principale	Dischi	Nastri o CDROM

Tabella 4: tabella riepilogativa dei vari livelli della memoria

1.2.5 Prospettive delle memorie di massa

Nuove tecnologie non elettroniche si preparano a prendere il posto dell'elettronica convenzionale quando essa sarà si sarà spinta ai suoi limiti fisici. Per le memorie di massa si stanno studiando sistemi in tre dimensioni, come gli ologrammi, che possono aumentare la densità di memorizzazione di molte migliaia di volte. Le più alte densità di informazione potranno essere raggiunte con l'elettronica molecolare o l'elettronica quantica, in grado di memorizzare i bit nei singoli atomi. Esistono sistemi sperimentali che sono in grado di realizzare "registri" di quattro bit con atomi. Se questi risultati scientifici daranno il via ad una serie di nuove tecnologie, esse potranno spingere la densità di memorizzazione fino a 10^{12} bit/mm^2, che significa il contenuto di 30 hard disk da 4 GByte nello spazio di un'unghia (di un bambino).
La velocità con cui i risultati della ricerca di base vengono convertiti in prodotti commercializzabili è sempre imprevedibile, per cui non si può sapere se e quando queste prospettive si concretizzeranno.

1.3 Unità di ingresso e uscita

1.3.1 Tastiera

Una tastiera è una matrice di contatti elettici. Individuando quale contatto elettrico viene chiuso si sa quale tasto è stato premuto. Nelle tastiere ordinarie il contatto è un normale interruttore, che può presentare qualche problema di usura, dato che possono scoccare piccole scintille che nel lungo periodo possono danneggiare il contatto.

Le migliori tastiere sono induttive, senza contatto elettrico; la pressione del tasto spinge un piccolo magnete permanente in mezzo ad un giogo magnetico, che ne rileva la presenza e chiude un contatto elettronicamente, senza usura dovuta a scintille od all'azione meccanica.
Esistono particolari tastiere per ambienti industriali, ermetiche alle polveri ed ai liquidi ed altre "antivandalismo" che hanno anche una particolare robustezza. In questi casi di solito la pressione del dito fa cambiare la resistenza della superficie del tasto, che è sempre percorsa da una piccolissima corrente. Vedendo dove cambia la corrente si sa quale tasto è stato premuto.
Le tastiere dei PC sono asservite ad un piccolo microcontrollore che individua il tasto e spedisce al computer un codice "già pronto" del carattere premuto.

1.3.2 Mouse

Il mouse è fatto con una pallina che fa ruotare due rulli, uno le coordinate X, l'altro per le Y. La rotazione dei rulli viene misurata da un sensore di angolo, di solito di tipo ottico. L'elettronica all'interno del mouse raccoglie la variazione di angolo dai due sensori e ne spedisce un codice al computer attraverso un'interfaccia seriale.

1.3.3 Scanner

Lo scanner è una telecamera a immagine fissa collegata al computer.
L'immagine del foglio da riprodurre viene proiettata su un particolare circuito integrato sensibile alla luce (CCD, **C**harghe **C**oupled **D**evice). Il CCD fornisce un flusso di numeri binari che corrispondono all'intensità luminosa di ciascuno dei suoi punti. In questo modo l'immagine viene convertita da una firma analogica ad una numerica.
Il flusso di dati digitali viene spedito al computer, che li può visualizzare ed elaborare.

1.3.4 Lettori di codici a barre

Il codice a barre è il più tipico sistema per l'identificazione degli oggetti. Questo libro ha un codice a barre che lo identifica.
Il codice a barre è costituito da una serie di righe nere in campo bianco, che vengono "spazzolate" da un raggio, emesso da una sorgente laser o LED. Un sensore fotoelettrico misura quanta della luce emessa dalla sorgente torna indietro riflessa.
Quando la luce riflessa è nulla od è poca, siamo in corrispondenza di una riga nera, se la luce che torna è molta siamo sopra ad un fondo bianco.
La larghezza delle righe nere e la distanza fra di esse è diversa in base al carattere codificato. Così, assumendo una velocità di scansione delle barre relativamente costante, è possibile leggere ciò che è scritto interpretando il segnale On - Off che proviene dal sensore di luminosità.
Esistono codici a barre in due dimensioni, che sono in grado di codificare molte più informazioni sulla stessa area. Con i codici a barre a due dimensioni è possibile memorizzare sull'etichetta non solo il numero dell'oggetto, ma anche un breve testo che lo riguarda.
Esistono diversi standard che stabiliscono come devono essere fatte le etichette, di solito i lettori di codici a barre sono in grado di capire automaticamente quale standard segue il codice e ad interpretare così il suo contenuto.
Come esempio non si può fare a meno di citare i lettori di codici a barre che stanno alla cassa dei supermercati più trafficati (laser scanner). Sono sistemi a laser rosso, che sparano il raggio "dappertutto" e sono in grado di rivelare la riflessione anche in ambienti con molta luce e con l'etichetta in posizione ben diversa da quella ideale.

Figura 7: codice a barre ad una e due dimensioni.

Altri sistemi per l'identificazione degli oggetti

Il codice a barre richiede che il lettore e l'oggetto siano in vista.
Se ciò non è possibile bisogna ricorrere ai RFID (**R**adio **F**requency **I**dentification system)
Un sistema RF è costituito da un "lettore", che è sostanzialmente una radio ricetrasmittente, e da un piccolo "**transponder**" che viene messo sull'oggetto.
Il transponder, che normalmente rimane spento, ha una piccola antenna integrata, tramite la quale lo fa accendere quando passa in prossimità del lettore. Una volta "sveglio" il transponder comunica via radio al lettore un numero, che la casa costruttrice scrive quando produce il dispositivo e che garantisce come univoco.
In questo modo il lettore potrà identificare, senza possibilità di contraffazione, l'oggetto che gli passa vicino.
Il transponder RF prende l'alimentazione elettrica che gli permette di funzionare dallo stesso segnale di attivazione, che gli proviene dal lettore. In questo modo si può anche evitare di dotarlo di una batteria.

Sistemi di identificazione RF sono usati per misurare i tempi sul giro delle F1 (accoppiati ad un traguardo laser), nei sistemi antitaccheggio dei supermercati, per misurare esattamente i tempi di tutti gli "atleti" della maratona di Londra (150 000!), per identificare i bidoni della spazzatura, in modo da far pagare la tassa rifiuti in funzione della spazzatura prodotta.

1.3.5 Monitor

Il monitor è la periferica grafica attraverso la quale viene espressa la gran parte degli output di un computer. Nei computer moderni i monitor sono dispositivi in grado di mostrare sofisticate immagini a colori, ma nella storia dell'Informatica non è stato sempre così; per lungo tempo il monitor è stato un dispositivo in bianco e nero che visualizzava un insieme di caratteri ristretto e non aveva possibilità grafiche.

CRT

I monitor usuali sono tubi a raggi catodici (**C**athode **R**ay **T**ube).
In un tubo riempito di gas rarefatto (a pressione bassissima, quasi vuoto) un "raggio" costituito di elettroni, prodotto da un filamento riscaldato, può essere "spostato" per mezzo di un "giogo" magnetico (yoke).
Alla fine del suo percorso il "pennello" di elettroni colpisce il rivestimento interno del cinescopio, che è ricoperto da una sostanza fluorescente ("fosfori"). I fosfori assorbono l'energia degli elettroni che li colpiscono e la riemettono sotto forma di luce.
Pilotando elettronicamente il giogo è possibile far disegnare al raggio righe sul cinescopio. E' poi possibile pilotare l'energia del fascio di elettroni, in modo che i fosfori si accendano con l'intensità luminosa voluta.
Con molte righe entro le quali l'intensità luminosa cambia a piacimento si possono formare le immagini volute.
I monitor a colori hanno tre tipi di fosfori. Infatti i colori vengono generati sommando la luce emessa da tre puntini dei colori primari "addittivi": Rosso, Verde, e Blu (RGB **R**ed, **G**reeen, **B**lue).

LCD

1.3.6 Stampanti

Verranno presentate in ordine di qualità di stampa, che è anche un ordine "vagamente" cronologico.

Stampanti ad impatto

Le stampanti ad impatto funzionano come una macchina per scrivere. Un nastro inchiostrato che sta sopra al foglio di carta viene colpito da un "martello" che ha la forma del carattere da stampare. In questo modo l'inchiostro viene trasferito al foglio.
Le varie tecnologie vengono distinte per il "martello" che impatta con il nastro inchiostrato. Le stampanti "a margherita" od "a rullo" hanno un elemento imprimente diverso per ogni carattere, mentre le stampanti a matrice di aghi hanno una piccola griglia di aghi molto fini che possono essere "sparati" o meno sul nastro inchiostrato, realizzando "a puntini" il carattere da stampare. Le stampanti a margherita od a rullo hanno la migliore qualità, mentre quelle a matrice di aghi sono più economiche e veloci (meno lente!).
Dato che i caratteri sono realizzati "a punti" le stampanti ad aghi possono essere utilizzate anche per stampare grafica, basta che il software controlli gli aghi della stampante per realizzare, a forza di puntini, il disegno che si vuole.
Con nastri a tre o quattro colori le stampanti a matrice di aghi possono stampare a colori.
Per la lentezza e la rumorosità le stampanti ad impatto sono usate al giorno d'oggi solo in quei settori nei quali è indispensabile ottenere più copie di un documento in una sola passata. Si pensi per esempio alla tenuta della contabilità delle aziende, per la quale la legge richiede che siano stampate diverse copie "in carta carbone" dello stesso originale.

Nelle stampanti "termiche", stampanti non propriamente "ad impatto", c'è una matrice di punti incandescenti che "brucia" il foglio di carta, che quindi deve essere termosensibile. Per quanto ci sia lo svantaggio importante che la "carta termica" tende a cancellarsi con il tempo, le stampanti termica sono molto economiche e robuste, per cui sono ancora usate, per esempio nei registratori di cassa o nelle bilance elettroniche.

Stampanti a getto

Come le stampanti a matrice di aghi, anche le stampanti a getto realizzano i caratteri "a puntini".
La testina di una stampante a getto ha molte centinaia di microugelli, di dimensioni piccolissime. Ciascun ugello è sotto il controllo della CPU della stampante, che può decidere in ogni istante se da ogni ugello deve essere "sputata" una microgoccia di inchiostro.
La testina si muove sul foglio e vi lascia milioni di gocce d'inchiostro, nei punti giusti, in modo che vengano realizzati i caratteri e la grafica desiderati.
Dato che gli ugelli di una stampante a getto possono essere centinaia di volte più piccoli degli aghi di una stampante ad aghi, la stampa a getto d'inchiostro (ink jet) dà risultati di una qualità enormemente superiore. Con testine che contengono inchiostro a tre o quattro colori è possibile stampare a colori.
Nelle stampanti i colori di base usati devono essere diversi da quelli dei CRT. Infatti il processo di colorazione è diverso. Il pigmento che viene depositato sul foglio riflette la luce, non la emette come nei CRT. Per esempio, il pigmento

rosso riflette solo la luce rossa, per cui ciò che è ricoperto di quel pigmento appare rosso. Perciò per generare tutti i colori bisogna mescolare i tre colori primari "sottrattivi": Ciano (azzurro), Magenta (rosso– viola) e Giallo (CMY **C**yan, **M**agenta **Y**ellow). Mescolando la massima quantità di ciano, magenta e giallo viene il nero. Dato però che il nero prodotto in questo modo è un po' sbiadito ed è "costoso" perché usa la massima quantità di inchiostro colorato, spesso le stampanti hanno anche un inchiostro nero, che usano nelle parti di testo nero ed in quelle a colori scure (CMYK **C**yan, **M**agenta **Y**ellow, Blac**k**).
Le stampanti a getto d'inchiostro hanno, rispetto alle stampanti laser, lo svantaggio dell'alto costo per foglio stampato e della lentezza di stampa. Inoltre il foglio stampato è soggetto a rapido deterioramento se esposto alla luce diretta del sole e l'inchiostro si scioglie se il foglio viene bagnato. Il vantaggio è che hanno un minore costo d'acquisto e possono produrre stampe a colori.
Altre tecnologie di caratteristiche analoghe a quelle delle ink jet sono il getto di cera, ed il trasferimento termico, che produce eccellenti ma costosissime stampe a colori, dato che per ogni stampa vengono usate quattro pellicole colorate che vengono "fuse" sul foglio in modo da realizzare tutti i colori possibili.

Stampanti laser

Una stampante laser è una fotocopiatrice controllata da un computer.
Infatti il processo di stampa è identico a quello di una fotocopiatrice, cambia solo "cosa" si stampa.
L'inchiostro di una fotocopiatrice, detto "toner", è fatto in modo da poter essere nebulizzato in microparticelle finissime ed è sensibile ai campi elettrostatici.
La luce che proviene dal piano di esposizione, dove c'è l'originale, viene mandata su un rullo elettrostatico, fotosensibile, detto "tamburo" (drum, appunto). Il tamburo si carica di elettricità statica nei punti in cui l'originale è nero, mentre rimane scarico dove esso è bianco. Poi il tamburo gira e passa sopra al serbatoio del toner (inchiostro nero). Il toner si attacca al tamburo solo nei punti dove esso è impressionato (punti neri dell'originale). Nel tamburo c'è quindi l'immagine dell'originale "fatta di toner".
A questo punto il tamburo continua il suo giro ed incontra il foglio. Il toner viene trasferito dal tamburo al foglio, così che ora sul foglio c'è l'immagine dell'originale.
Il lavoro non è finito, perché il toner è sul foglio ma è instabile, basterebbe un soffio per farlo volare via.
Il foglio, con la sua immagine fatta di toner, viene quindi mandato dentro ad un forno, in cui raggiunge per un brevissimo istante la temperatura di 200 gradi. Con il riscaldamento il toner si fissa in modo definitivo al foglio e la fotocopia può essere espulsa calda calda. Il tamburo è pronto per fare un altro giro e trasferire un'altra immagine.
Nella stampante laser la luce, invece di provenire dal piano di esposizione della fotocopiatrice, proviene da un laser che, controllato dalla CPU della stampante, "disegna" sul tamburo tutti i punti del foglio che vuole stampare in nero. Una volta che il tamburo è stato impressionato il resto del processo è identico a quello della fotocopiatrice.
Una stampante laser produce stampe in bianco e nero di eccellente qualità e che non "temono" l'acqua, la luce ed i graffi, dato che l'inchiostro è stato fissato a caldo sul foglio. Inoltre può essere velocissima.
Rispetto ad una stampante a getto d'inchiostro è più costosa come prezzo di acquisto, ma ha costi di esercizio minori se si fanno molte stampe, dato che la carta ed il toner costano meno per foglio stampato.
Alcune stampanti di questo tipo invece di usare un laser per impressionare il tamburo usano una fila di LED.
Esistono anche stampanti laser a colori, che solo da poco danno risultati comparabili con quelle a getto d'inchiostro, ma ad un costo di acquisto decisamente superiore (anche in questo caso i costi di esercizio sono minori).

1.3.7 Periferiche multimediali

Oggi sempre più il personal computer è un dispositivo multimediale, che deve accettare i suoi ingressi e produrre le sue uscite su una grande quantità di periferiche quali per esempio il microfono e la scheda sonora, tramite i quali si possono impartire al computer comandi vocali, la telecamera o la Web – Cam per le teleconferenze, la macchina fotografica digitale, la scheda grafica.

Famiglia	Computer ove è usata	Sistemi Operativi tipici	Dimens. locazione	Dimens. "word"	n. registri gnerali
X86	PC IBM e compatibili	MS-DOS, Windows, Unix	8 bit	16 bit	
68k	Tutti i primi Apple Macintosh	Apple OS	8 bit	32 bit	
PowerPC	Macintosh nuovi, IBM RISC	Apple OS, AIX, OS/2			
PA-RISC	Workstation HP	HP-Ux, Windows NT			
SPARC	Workstation Sun	Solaris			
MIPS	Workstation Silicon Graphics	????			
Apha	Workstation e mini-computer Digital	Digital Unix, VMS, Windows NT			

Esempio finale

Problemi ed esercitazioni

Domande