Quaderni di informatica 15

Rassegna di tecnologia ed applicazioni degli elaboratori elettronici - Anno VIII - Numero 1 - 1981

Honeywell
Honeywell Information Systems Italia

FPDM-58

Honeywell

Honeywell Information Systems Italia

Quaderni di informatica

Rassegna di tecnologia ed applicazioni degli elaboratori elettronici
Anno VIII - Numero 1 - 1981

Sommario

Direttore Responsabile
F. Filippazzi

Comitato di Redazione
C. Falcetti, P. Lupo,
M. Nobile, G. Occhini
G. Rapelli, F. Sala

Redazione
Honeywell Information Systems Italia
Centro di Ricerca e Progettazione
20010 Pregnana Milanese (Milano)

Stampa
tipolito Maggioni s.a.s.

Autorizzazione del Tribunale di Milano
n. 93 del 20 Marzo 1974

«Quaderni di Informatica» ospita
articoli e contributi di varia provenienza.
La responsabilità delle opinioni
espresse rimane agli Autori.

Sistemi distribuiti di elaborazione: (*) una introduzione

FRANCO FILIPPAZZI
Honeywell Information Systems Italia
Milano

1. Introduzione

La concezione dei sistemi di elaborazione ha registrato una dinamica continua, con trasformazioni profonde in cui si intrecciano componenti scientifiche, tecniche, economiche e organizzative. Questa complessa evoluzione può essere schematizzata, in prima approssimazione, nelle fasi indicate in fig. 1.

All'inizio, i sistemi di elaborazione sono decentrati e senza alcuna capacità di comunicazione. L'elaboratore è una entità a se stante ("stand alone"), il cui uso è possibile solo accedendo direttamente al luogo dove è installata la macchina.

Criteri di economia di scala conducono in uno stadio successivo (circa metà degli anni '60) a criteri di accentramento delle risorse elaborative. Si sviluppano pertanto calcolatori di grosse dimensioni e contemporaneamente prende piede la tele-elaborazione come correttivo alla centralizzazione. L'utente finale può accedere all'elaboratore da lontano, mediante terminali remoti collegati via linea telefonica. I terminali sono però semplici organi di ingresso/uscita, senza capacità di elaborazione, rimanendo tutta l'"intelligenza" concentrata nel grosso elaboratore centrale.

Verso il 1970, l'avvento dei mini-computer induce nelle organizzazioni un riflusso di decentralizzazione. I mini significano infatti una notevole potenza di elaborazione ad un prezzo relativamente basso, tanto da rientrare nei limiti di autorizzazione dei livelli divisionali. I mini si diffondono quindi nell'organizzazione, sfuggendo al controllo del centro di calcolo. È la cosiddetta "rivolta dei mini", un tentativo dell'utente finale di superare la lentezza e la rigidità del servizio centralizzato, acquisendo una propria capacità elaborativa.

Anche questa soluzione mostra però i suoi limiti, perchè ogni computer necessita in generale dei dati forniti dai computer di altre parti dell'organizzazione.

Si arriva così all'ultima fase dell'evoluzione, quella dei sistemi distribuiti. Se prescindiamo dalle realizzazioni sperimentali avvenute nell'ultimo decennio, questa fase si apre sostanzialmente con gli anni '80.

L'informatica distribuita si presenta come un momento di sintesi delle concezioni precedenti, come una soluzione capace di conciliare l'antitesi centralizzazione/decentralizzazione.

Un sistema distribuito è infatti concepito come un insieme di unità dotate di capacità operativa autonoma e al tempo stesso in grado di scambiare mutuamente dati e risorse elaborative attraverso una rete di comunicazione. L'"intelligenza" risulta distribuita nel sistema, non però in un'ottica autarchica, ma di cooperazione ed integrazione delle risorse.

2. Motivazioni dell'approccio distribuito

Da un punto di vista "filosofico", si potrebbe dire che l'informatica distribuita rappresenta la correzione di una tendenza che ha caratterizzato l'elaboratore dalla sua origine ad oggi, cioè la crescita continua di complessità. Questo fenomeno è attribuibile, sostanzialmente, a tre fattori:

- il progresso della tecnologia, che mettendo a disposizione elementi sempre meno costosi, meno ingombranti, più affidabili, ne ha incoraggiato l'impiego;
- i criteri di economia di scala, che hanno spinto alla centralizzazione dell'elaborazione ed a macchine sempre più potenti;
- le nuove "dimensioni" acquisite via via dall'informatica (dalla mono alla multiprogrammazione, dal batch all'interattività, dal sistema indivisibile al time-sharing, e così via).

Tutti questi fattori hanno concorso a rendere l'elaboratore sempre più complesso e quindi sempre più difficile da progettare, da co-

(*) Presentato al 28° Congresso Internazionale per l'Elettronica, Roma, marzo 1981.

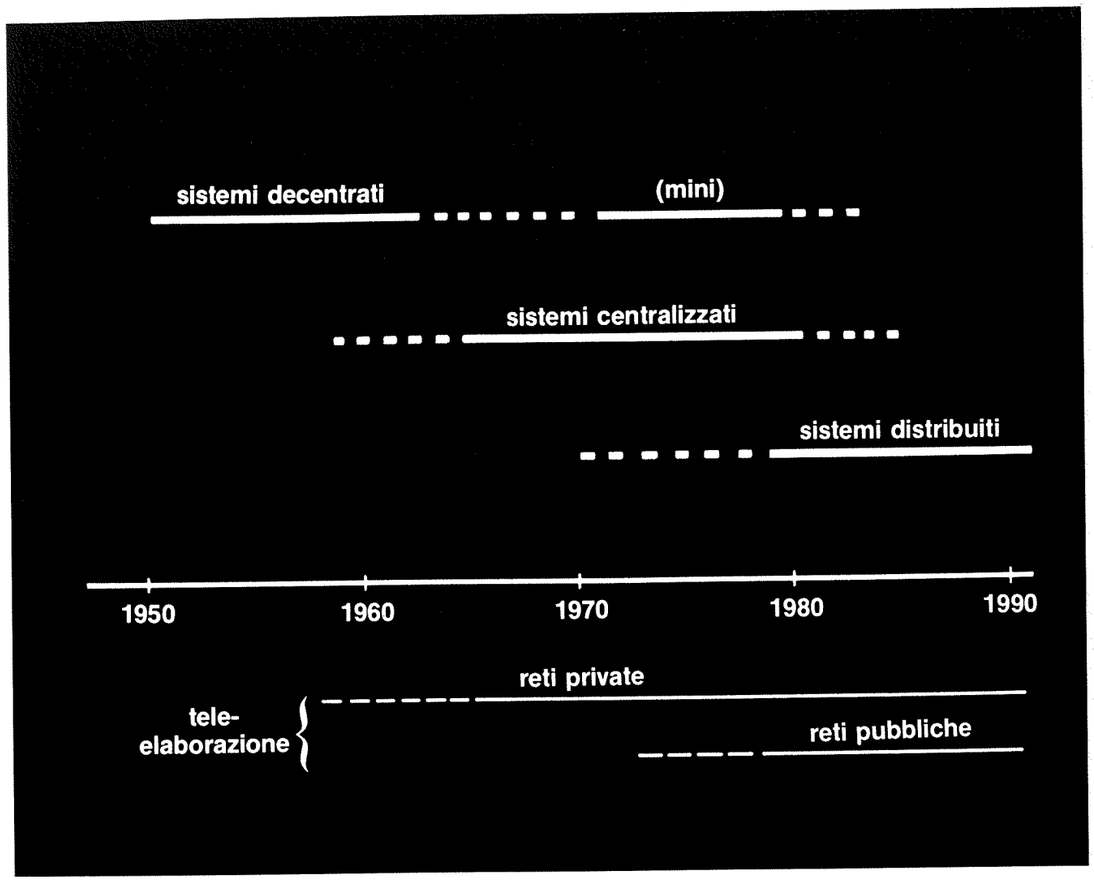


Fig. 1 - Stadi evolutivi dei sistemi di elaborazione.

struire, da collaudare, da riparare, da usare. Questo processo non può continuare indefinitamente, per non arrivare alla "barriera della complessità". Il concetto di sistema distribuito costituisce, sotto questo profilo, una opportuna inversione di tendenza.

Se questo può essere un modo piuttosto astratto di interpretare la svolta verso i sistemi distribuiti, non mancano in materia argomentazioni più precise.

1) L'evoluzione tecnologica, grazie in particolare alla microelettronica, ha consentito di realizzare in piccoli o piccolissimi volumi (mini/microcomputer) rilevanti potenze di calcolo a costi estremamente ridotti, rendendo fattibile una distribuzione capillare della "intelligenza".

Parallelamente, anche il settore delle comunicazioni è progredito, realizzando nuove tecniche di trasmissione, quali la commutazione di pacchetto, che offrono nuovi livelli di efficienza e di qualità del servizio.

Comparativamente però, l'evoluzione dei costi è stata assai più veloce nel settore dell'elaborazione che in quello della trasmissione. Nel primo si è registrato infatti un tasso di riduzione che arriva al 25% annuo, contro un 11% medio del secondo [1]. Il fenomeno è descritto qualitativamente nella fig. 2, che mostra come col tempo si siano invertiti i pesi economici unitari della elaborazione e della trasmissione dei dati. Tale andamento se prima forniva elementi a favore della elaborazione centralizzata, onde ridurne il costo mediante economie di scala, giustifica ora la strategia opposta, cioè l'elaborazione locale, onde minimizzare il volume e quindi il costo della comunicazione.

2) La centralizzazione postulava economie di scala non solo in termini di macchine ma anche di risorse umane. Questi concetti avevano trovato la loro sintesi nella "legge di Grosh", secondo cui l'efficienza operativa cresce col quadrato del costo dell'elaboratore. Questa impostazione non è ora ritenuta più valida, non solo a causa dell'evoluzione tecnologica prima citata, ma anche per una revisione dei criteri organizzativi. La legge di Grosh trova infatti i suoi limiti nella "legge di Parkinson", secondo cui la produttività di una organizzazione aumenta con il logaritmo delle risorse disponibili [2]. La

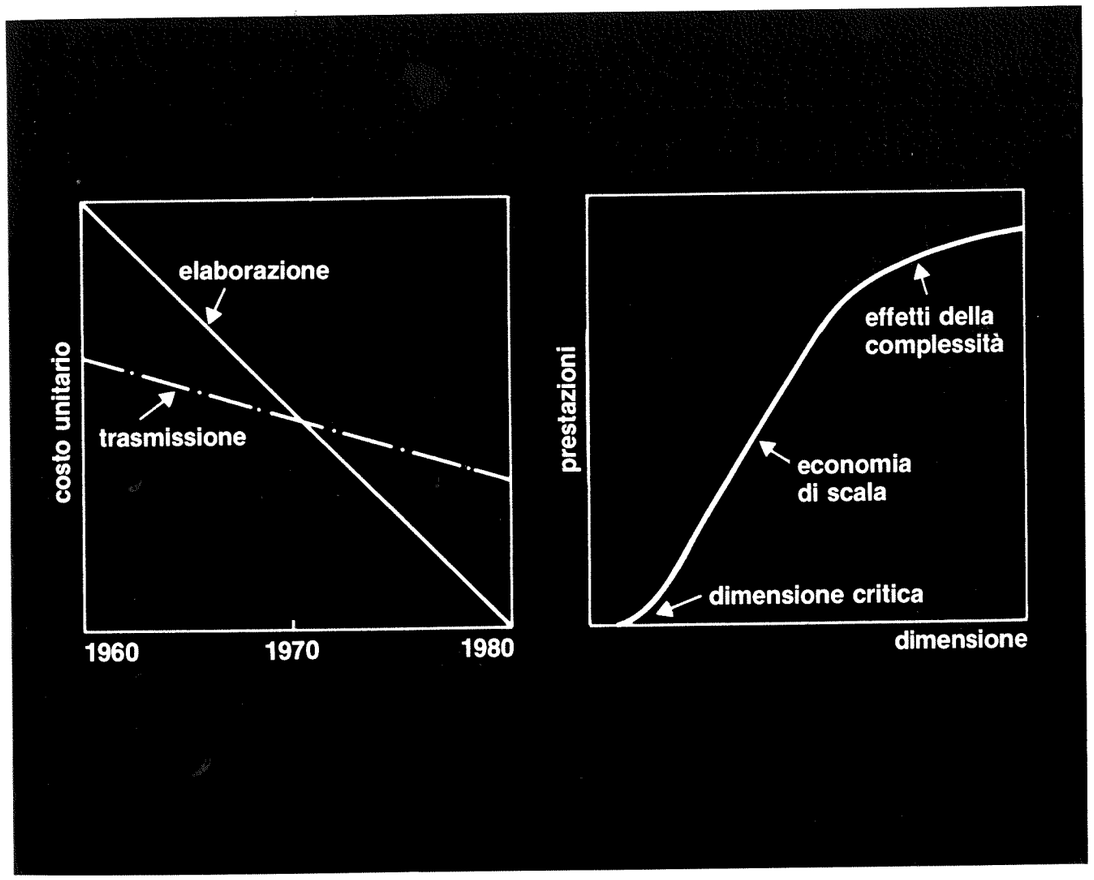


Fig. 2 - Evoluzione del costo di elaborazione e trasmissione dati.

Fig. 3 - Relazione tra prestazioni e dimensione del sistema.

fig. 3 descrive qualitativamente il fenomeno, che implica l'abbandono della centralizzazione tout-court e l'orientamento verso soluzioni di tipo policentrico.

3) L'approccio distribuito consente un salto qualitativo nell'uso dell'informatica.

- Il fatto stesso che le funzioni vitali del sistema non siano concentrate, conferisce alle strutture distribuite una intrinseca "tolleranza al guasto". Il verificarsi di un tale evento non provoca infatti l'arresto del sistema, ma solo una riduzione delle sue prestazioni ("fail-soft"). Per ottenere risultati paragonabili con sistemi centralizzati sarebbe necessario introdurre costose ridondanze. Oltre a ciò, un sistema distribuito è chiaramente meno vulnerabile rispetto ad evenienze catastrofiche, naturali o dolose che siano.
- Essendo l'elaborazione effettuata di norma nel posto stesso di utilizzazione, ne consegue un deciso miglioramento delle prestazioni globali del sistema (tempo medio di risposta, throughput) rispetto al caso di elaborazione centralizzata.
- L'utente finale di un sistema distribuito può usare risorse di cui sarebbe proibitivo disporre in qualunque altra soluzione. Le capacità di elaborazione, i programmi e i dati esistenti nel sistema sono infatti, in linea di principio, patrimonio comune di tutti gli utenti.

Queste considerazioni, che pur non esauriscono l'argomento, sono sufficienti a mostrare come l'approccio distribuito costituisca un reale superamento delle soluzioni precedenti, in linea con l'evoluzione della tecnologia e dei criteri di impiego della risorsa informatica.

3. Tassonomia dei sistemi distribuiti

Il concetto di sistema distribuito presenta una notevole varietà di interpretazioni ed è perciò opportuno fissare le idee.

Un primo criterio di classificazione può essere basato sulla *estensione spaziale* del sistema stesso. Si può in questo senso parlare di tre classi di sistemi distribuiti, come mostrato nella tabella che

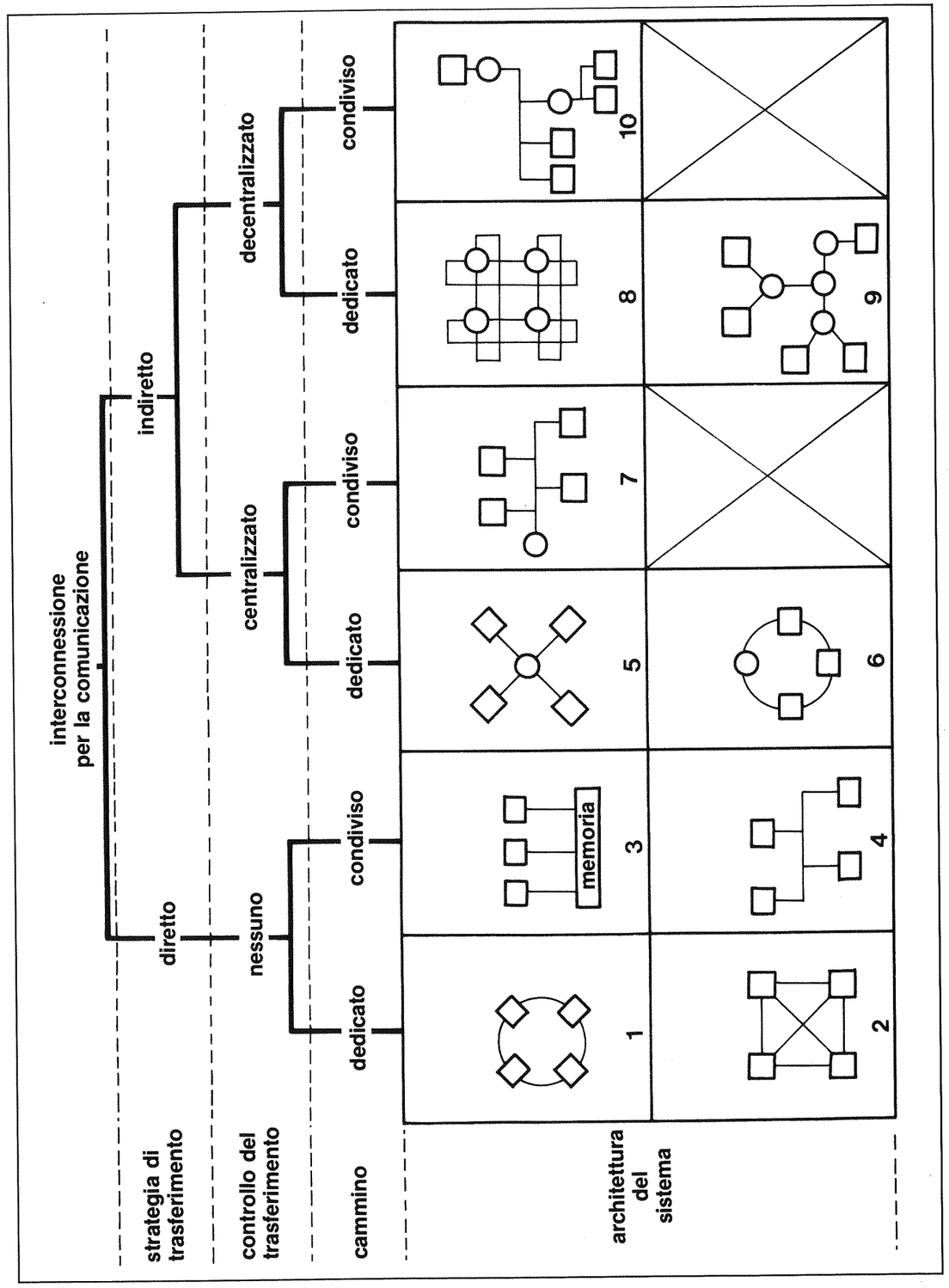


Fig. 4 - Tassonomia dei sistemi distribuiti di elaborazione, in base al metodo di comunicazione.

segue:

Sistema distribuito	Distanza (m)	Veloc. trasf. (bit/sec)
Distribuzione funzionale	<10	10^7-10^9
Distribuzione locale	10-1000	10^4-10^7
Distribuzione geografica	>1000	10^3-10^5

La "distribuzione funzionale" interessa la struttura interna del calcolatore. In sostanza, l'unità di elaborazione (CPU) viene realizzata mediante una pluralità di microelaboratori tra loro cooperanti (CPU "multiprocessor"). La comunicazione tra i moduli del sistema è digitale, presenta elevato parallelismo ed avviene mediante connessioni multiple stampate o cablate.

La "distribuzione locale" fa riferimento a sistemi estesi ad aree limitate, quali possono essere un edificio o gruppi di edifici. La comunicazione tra le varie unità del sistema è tipicamente digitale e seriale su filo o cavo.

La "distribuzione geografica" si riferisce a sistemi sparsi su ampie aree territoriali. La trasmissione è analogica o digitale ed i mezzi impiegati possono essere molteplici, dalla linea telefonica al satellite di comunicazione.

I sistemi della prima classe sono generalmente "ad accoppiamento stretto", nel senso che i vari processor hanno in comune la memoria di lavoro; i sistemi distribuiti localmente o geograficamente sono invece tipicamente "ad accoppiamento lasco", ed in essi lo scambio dei dati comporta un più complesso formalismo.

Se prescindiamo dalla caratteristica spaziale del sistema, esistono altre categorizzazioni possibili a seconda che il concetto di distribuzione sia riferito alle risorse di *elaborazione*, ai *dati* o al *controllo* del sistema.

Un sistema distribuito può essere definito tale se almeno una delle entità citate è distribuita. Una definizione quale proposta da Enslow [3], che includa tutti e tre gli aspetti, è teoricamente rigorosa ma risulta molto restrittiva poichè, in tale accezione, non esisterebbero attualmente veri sistemi distribuiti.

Con riferimento alle risorse di elaborazione, una categorizzazione interessante è quella proposta da Anderson e Jensen [4], che esamina il sistema dal punto di vista della comunicazione (trattamento dei messaggi tra i processor e topologia di interconnessione dell'hardware).

Il modello dà luogo ad un albero, le cui foglie corrispondono a possibili topologie di sistema. Ciò è illustrato nella fig. 4, nella quale i quadrati rappresentano gli elementi di elaborazione mentre i cerchi sono elementi di commutazione. Si hanno le seguenti strutture:

1. ad anello
2. ad interconnessione completa
3. multiprocessor con memoria comune
4. lineare
5. a stella
6. ad anello con controllo centralizzato
7. lineare con controllo centralizzato
8. a maglia regolare
9. a maglia irregolare
10. lineare con finestra

Pur presentando alcune limitazioni, il modello di Anderson e Jensen fornisce tuttora la più valida tassonomia dei sistemi distribuiti. Esso costituisce un utile strumento per analizzare dieci fondamentali concezioni di sistema, permettendo di compararne le più importanti caratteristiche (throughput, modularità, sensibilità al guasto, ecc.) e di identificarne i punti di forza e di debolezza.

Anche se praticamente tutte le alternative previste dal modello sono state realizzate (molto spesso solo in unico esemplare), quelle probabilmente destinate ad avere maggiore seguito sono, a seconda le caratteristiche dell'applicazione, la struttura ad anello (1), a stella (5), a bus (4) ed a maglia irregolare (9). Quest'ultima è la soluzione dominante per i sistemi distribuiti geograficamente e ad essa ci si riferisce nel linguaggio comune quando si parla di "reti di calcolatori".

Se passiamo ora a considerare la distribuzione dei dati, occorre anzitutto dire che l'esistenza di basi di dati nei vari nodi del sistema non costituisce di per sè una "base di dati distribuita". Questa si ha solo se tali archivi sono tra loro correlati logicamente o funzionalmente, in modo da costituire un'unica collezione di dati. In un sistema così fatto, un qualsiasi programma utente può accedere, in modo uniforme e trasparente, ai dati esistenti nei vari punti del sistema. La base di dati può essere soltanto ripartita tra i vari nodi oppure replicata mediante copie parziali (fig. 5), con la possibilità di soluzioni miste. La scelta va fatta in base alle caratteristiche che si desiderano ottenere (tempo di risposta, sicurezza, costo, ecc.) a livello sistema.

Le basi di dati distribuite costituiscono un tema ampio e complesso. I problemi derivano non solo dalle difficoltà tecniche, che pure sono elevate, ma spesso anche dal fatto di dover tenere conto di soluzioni esistenti, quasi sempre tra loro inomogenee. Per queste ragioni, la diffusione di basi di dati distribuite non procede velocemente, anche se non mancano ormai esempi significativi di realizzazione.

Passando infine a considerare il controllo del sistema, questa è certamente la caratteristica più difficile da distribuire. Negli attuali sistemi distribuiti il controllo delle risorse di elaborazione e della base di dati è sostanzialmente di tipo centralizzato o gerarchico. Ciascun componente della struttura è cioè controllato dai membri di livello superiore, ed il controllo complessivo risiede in un nodo centrale, che possiede la visibilità dello stato globale del sistema.

A questa concezione "verticale" del controllo, si oppone quella distribuita od "orizzontale". In un si-

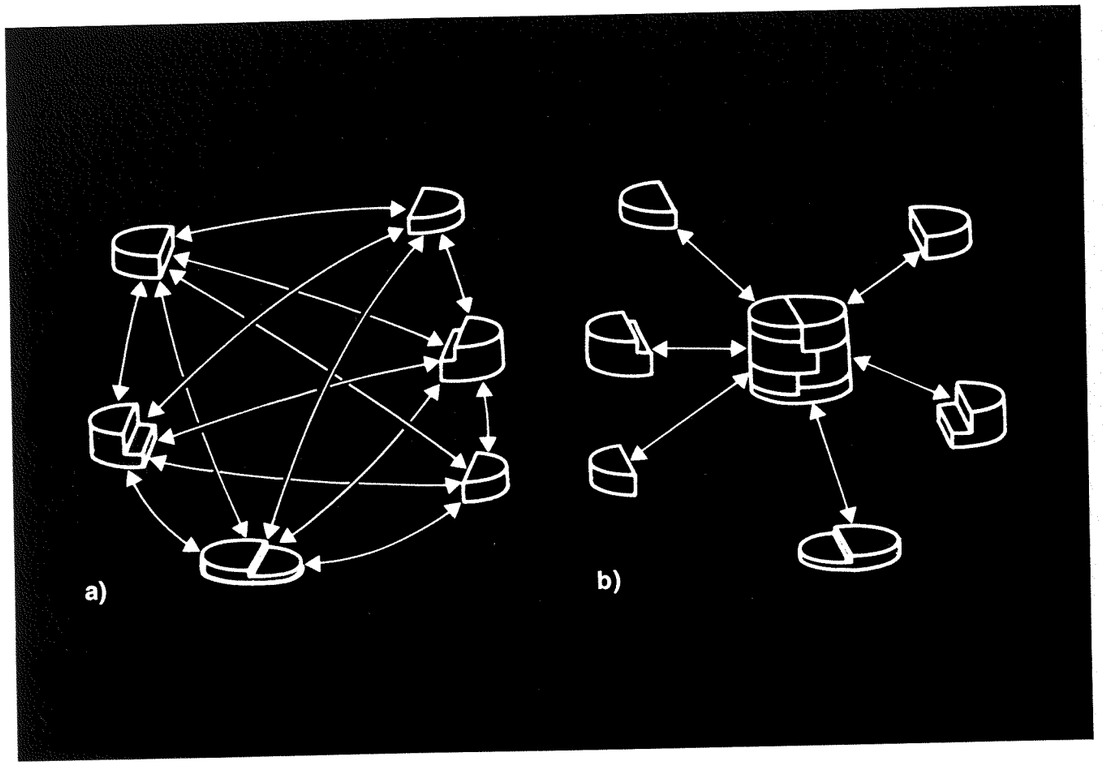


Fig. 5 - Distribuzione dei dati:
a) base di dati ripartita,
b) base di dati replicata.

stema con controllo distribuito, i vari nodi partecipano alla esecuzione dei compiti su un piano di parità logica. Il ruolo di ciascun nodo commuta dinamicamente tra quello di "padrone" e quello di "schiavo" (master/slave) ed il flusso dei dati tra due nodi avviene sotto il controllo di quello che originariamente ha stabilito il collegamento.

L'obiettivo fondamentale che ci si propone con un controllo distribuito è di rendere il sistema meno vulnerabile rispetto al guasto di una singola parte, evento che può risultare invece paralizzante in un sistema distribuito con controllo centralizzato o gerarchico.

La distribuzione del controllo è un problema di grande difficoltà teorica, poichè implica la creazione di un "sistema operativo distribuito", capace cioè di controllare il sistema nel suo complesso senza la conoscenza di alcuna variabile globale del sistema. È questa un'area ancora ampiamente aperta alla ricerca ed alla sperimentazione.

4. L'architettura dei sistemi distribuiti

Se si considera un qualsiasi sistema distribuito, si distinguono tre aree fondamentali, rispettivamente di utente, di elaborazione e di comunicazione, come illustrato in fig. 6.

La sottorete di comunicazione è qui intesa come quella infrastruttura che permette il collegamento tra le varie unità che sono ad essa allacciate.

La sottorete di comunicazione costituisce un punto fermo del sistema, nel senso che elaboratori e terminali ad essa connessi cambieranno generalmente col tempo, ma la rete rimarrà, anche se espandendosi e modificandosi. Elaboratori e terminali dovranno quindi adeguarsi agli standard della rete di comunicazione, pubblica o privata che sia, e non viceversa.

È evidente da ciò, la grande importanza di standardizzare un insieme di caratteristiche dei sistemi di elaborazione, affinchè ad una rete possano essere collegati senza difficoltà apparati di qualsiasi tipo, dimensione e provenienza, ed essi possano tra loro colloquiare.

L'insieme delle discipline cui attenersi per costruire un sistema di elaborazione, a partire dai moduli hardware e software disponibili, viene usualmente designato come l'"*architettura*" del sistema. Questa definizione è valida in generale e si applica quindi anche al caso che il sistema di elaborazione sia distribuito.

Un fondamento concettuale comune a tutte le architetture distribuite esistenti è la *struttura a strati* (fig. 7).

L'insieme delle funzioni del sistema è suddiviso in strati, o livelli, organizzati gerarchicamente. Gli strati di livello superiore utilizzano le funzioni fornite da quelli inferiori per costruire una funzionalità crescente ed un maggior livello di astrazione. Ogni strato presenta delle "interfacce" precise, che garantiscono il coordinamento e il controllo del flusso dei dati con gli strati adiacenti. Inoltre, ogni strato può dialogare con lo strato di pari livello di altre unità, seguendo un insieme di regole e convenzioni ben definite, chiamate "protocolli". Durante tale dialogo, gli strati a livello inferiore risultano "trasparenti".

Il vantaggio fondamentale della impostazione a strati consiste nella possibilità di isolare le funzionalità dei vari strati e quindi di consentire modifiche al loro interno senza dover intervenire sul resto della struttura.

L'architettura di un sistema distribuito risulta quindi essere, in definitiva, l'insieme organizzato di funzioni, interfacce e protocolli, attraverso cui elaboratori e terminali, quando interconnessi in una rete, possono operare in modo coordinato.

Allo stato delle cose, esiste tutta una varietà di architetture distribuite, generate dai principali costruttori di elaboratori. Si ha così lo SNA (System Network Architecture) della IBM, il DSA (Distributed System Architecture) della Honeywell, il DCA (Distributed Communication Architecture) della Univac, e così via. Tutte queste architetture sono, in pratica, tra loro incompatibili. (*)

Gli enti internazionali di normalizzazione, quali il CCITT, l'ISO, ecc. stanno lavorando da anni su standard per sistemi distribuiti. Attualmente si registra una convergenza di consensi sul modello di architettura proposto originariamente dall'ISO (International Standard Organization) e noto come OSI ("Open Systems Interconnection Reference Model") [6].

Questo modello, come mostrato in fig. 8, definisce 7 strati per l'accesso alle comunicazioni. (Nella figura è indicata anche l'esistenza di strati superiori, per l'accesso alla base di dati, che non sono però ancora definiti).

Come si è già detto, l'adesione agli standard proposti è finora scarsa e comunque limitata ai tre livelli inferiori della architettura. Ad essi si riferisce, ad esempio, lo standard CCITT/ISO X-25 per le reti a commutazione di pacchetto, che pure, di fatto, non risulta inserito in diverse note architetture.

La meta della ampia diffusione di architetture "aperte" appare ancora lontana, a causa più di ragioni commerciali e industriali che non strettamente tecniche. Non mancano tuttavia prese di posizione in tale direzione; ad esempio, lo standard OSI è stato incorporato nella architettura DSA della Honeywell.

(*) La connessione di sistemi non compatibili è possibile introducendo specifici adattatori ("gateway"). È questa evidentemente una soluzione di ripiego.

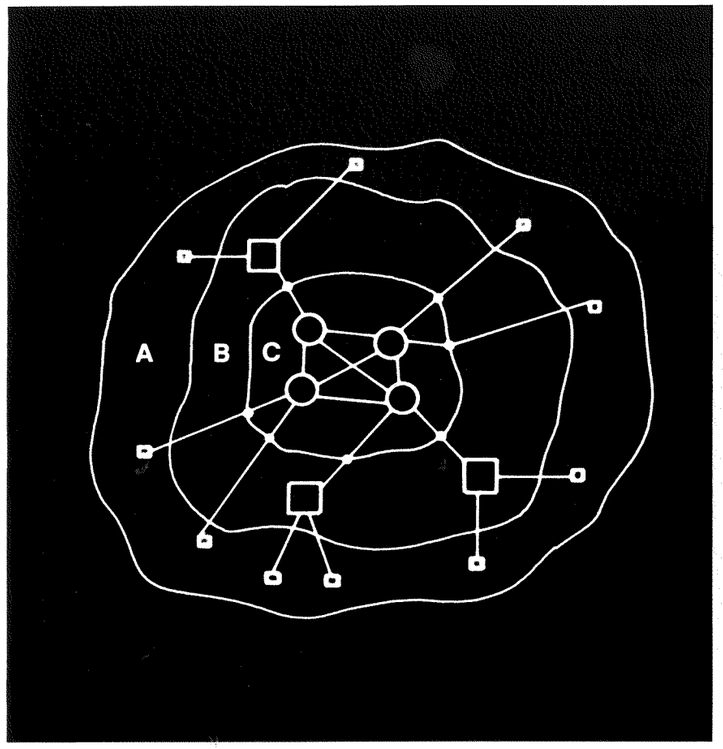


Fig. 6 - Le tre reti fondamentali:
a) rete di utente,
b) rete di elaboratori,
c) sottorete di comunicazione.

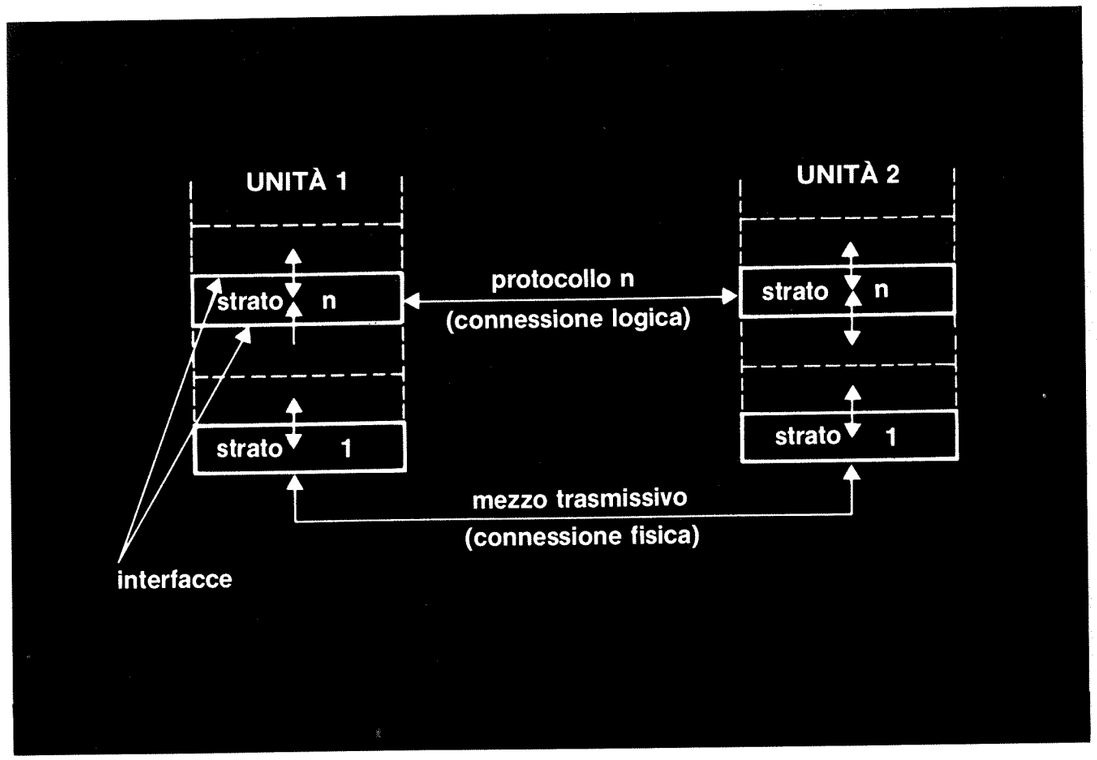


Fig. 7 - Architettura a strati.

5. Le prospettive

Gli anni '70 hanno costituito sostanzialmente la fase di ricerca e di sperimentazione dei sistemi distribuiti, anche se sono ormai numerosi gli esempi di realizzazioni ad opera di università, industrie e grandi utenti, in USA soprattutto ma anche in Europa.

Anche l'Italia ha partecipato a questa evoluzione di fondo dell'informatica. Si può citare, ad esempio, la partecipazione nel progetto di reti continentali, quali l'Euronet; lo sviluppo di progetti di reti locali; la realizzazione di elaboratori con struttura distribuita, arrivati in un caso (DPS 4) alla fabbricazione di larga serie. Tra le iniziative in atto, è da citare il programma Compunet, nell'ambito del Progetto finalizzato informatica del CNR, cui partecipano università ed industrie, col coordinamento del CREI.

Gli anni '80 si presentano ora come la fase di espansione dei sistemi distribuiti, anche se ciò avverrà ancora in modo graduale, dato il numero e la complessità dei problemi coinvolti.

Un problema di fondo è quello degli standard architetturali. Come già accennato, si può dire che una standardizzazione è stata raggiunta, in pratica, solo per i protocolli a livello più basso, quelli cioè che servono per interagire con la sottorete di comunicazione. Lo standard CCITT (incluso nel modello ISO) si è imposto sostanzialmente per il fatto che del CCITT fanno parte le amministrazioni PT di tutto il mondo. La necessità di allacciarsi alle reti pubbliche forza quindi i costruttori ad adeguarsi a tale normativa.

Il problema è invece del tutto aperto per quanto concerne i protocolli di livello più alto, quelli cioè che consentono l'interazione tra processi utente. Non esiste in questo caso una concreta spinta alla standardizzazione, anzi giocano in senso contrario interessi e politiche aziendali. La standardizzazione in questo ambito si prospetta pertanto di difficile attuazione.

Complessivamente quindi, la diffusione di architetture "aperte" va vista in un'ottica di tempi lunghi.

A questo tipo di vincolo, si deve aggiungere il fatto che la progettazione di un sistema distribuito rimane un lavoro molto complesso. La varietà di soluzioni alternative su come configurare il sistema, allocare e dimensionare le risorse, realizzare l'opportuno livello di integrità e riservatezza dei dati, definire una adeguata diagnostica del sistema (che eviti, tra l'altro, palleggiamenti di responsabilità tra il gestore della rete e il fornitore delle apparecchiature di elaborazione), e così via, rendono arduo il processo di ottimizzazione del sistema. In questa luce, indubbia importanza continuerà ad avere la ricerca di metodologie e di strumenti di progettazione, così come

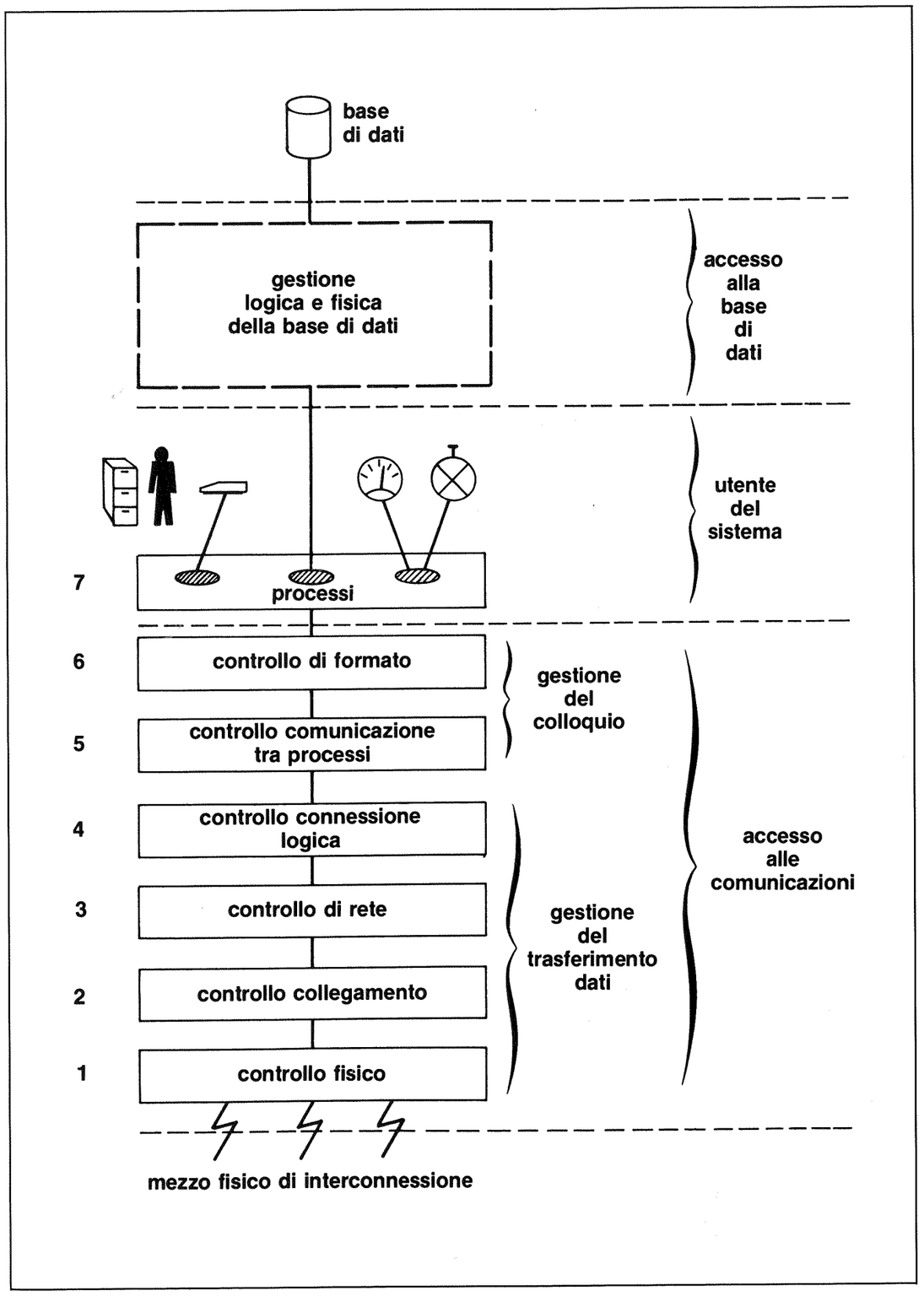


Fig. 8 - Modello di architettura
di sistemi distribuiti (OSI).

un ruolo cruciale avranno le attività di formazione.

A monte di tutto ciò, lo sviluppo dei sistemi distribuiti è condizionato, ovviamente, dalla esistenza di adeguate infrastrutture di comunicazione, in particolare del tipo a commutazione di pacchetto.

L'Italia, a questo riguardo, sta tuttora registrando un sensibile ritardo. Se dai problemi e condizionamenti passiamo alle prospettive, si possono individuare alcune principali linee di sviluppo.

La tecnologia, specie la microelettronica, eserciterà ancora una enorme influenza sullo sviluppo dei sistemi distribuiti. Si può citare, in particolare, il progressivo trasferimento in hardware di funzionalità architetturali oggi realizzate in software.

Ad esempio, si prevede a non lunga scadenza, la realizzazione dell'intero protocollo X-25 in un singolo "chip" di semiconduttore. Si potranno quindi inserire ampie capacità di colloquio anche in apparecchiature di piccole dimensioni e basso costo. Questa possibilità porterà verso una distribuzione capillare della elaborazione, a livello "personale", con apparecchiature inseribili sulla rete di comunicazione mediante una spina, come si fa per l'energia elettrica ("information outlet", la presa informatica).

Sotto un profilo sistemistico, si può dire che nei prossimi anni si andrà verso una "vera" distribuzione dell'intelligenza, cioè verso sistemi in cui il termine distribuito si applica contemporaneamente alle risorse di elaborazione, alle basi di dati ed al controllo.

Un ulteriore orientamento è costituito dalla integrazione nel sistema di tutte le varie forme di comunicazione (testi, dati, voce, immagini). In particolare, nel caso di una azienda ciò significa l'integrazione nella rete di una varietà di funzioni oggi viste separatamente (telex, fac-simile, word processing, ecc.), nonchè di servizi svolti finora mediante altri supporti fisici (es. electronic mail) o addirittura inesistenti (es. teleconferencing).

Il sistema distribuito di elaborazione diventa in tal modo la piú importante risorsa aziendale e come tale andrà pianificato e gestito.

Una ulteriore possibile evoluzione riguarda il concetto di rete pubblica di comunicazione. La nozione di VAN (Value Added Network) potrà cambiare nel senso che il "valore aggiunto" della rete tenderà ad aumentare.

Rappresentativo di questa linea di tendenza, oggi limitata agli USA, è il sistema ACS (Advanced Communication Service) della ATT, attualmente in fase di sviluppo. Si tratta di una rete di comunicazione "superintelligente", che si prende carico di molte funzioni tipiche della rete di calcolatori, sollevandone l'utente.

Il progetto della ATT si inquadra nel nuovo contesto legislativo USA, che ha liberalizzato i settori delle telecomunicazioni e dell'informatica, consentendo ad aziende dell'un settore di entrare nell'altro. In questo quadro di reciprocità, si inserisce il lancio (novembre 1980) del primo satellite di comunicazione della IBM, attraverso la partecipazione al SBS (Satellite Business System). C'è materia per prevedere un duello tra giganti sul terreno, ormai comune, della telematica.

A prescindere dalle strategie industriali, una domanda che viene fatto di porsi è: Quale è il posto migliore per collocare determinate funzioni? Nella rete di comunicazione, negli host computer, nei concentratori o in terminali intelligenti costruiti con microprocessori? La risposta a questa domanda è molto difficile perchè legata alla proiezione di tutta una serie di fattori, che vanno dai costi dell'hardware e del software, allo sviluppo delle reti, all'evoluzione dei servizi.

Ciò che si può dire con sicurezza è che l'incontro (che, come si è visto, contiene anche una dialettica di scontro) tra informatica e telecomunicazioni costituisce un evento di eccezionale portata che apre realmente una nuova epoca, quella della "società dell'informazione".

6. Bibliografia

[1] L. M. BRANSCOMB, "*Computing and Communication - A perspective of the evolving environment*", IEE Compcon, Sept. 1979.

[2] G. OCCHINI, "*L'informatica nella gestione aziendale*", Ed. F. Angeli, 1980.

[3] P.H. ENSLOW, "*What is a Distributed Data Processing System?*", Computer, Jan. 1978.

[4] G.A. ANDERSON, E.D. JENSEN, "*Computer Interconnection: Taxonomy, Characteristics and Examples*", ACM Computer Surveys, Dec. 1975.

[5] P.E. GREEN, "*An Introduction to Network Architecture and Protocols*", IEEE Trans. on Communication, Apr. 1980.

[6] H. ZIMMERMAN, "*OSI reference model - The ISO Model of Architecture for Open System Interconnection*", IEEE Trans. on Communication, Apr. 1980.

Applicazioni dell'elaboratore nel campo dell'intelligenza artificiale

TULLIO CHERSI
ELSAG
Genova

1. Introduzione

Sin dagli inizi dello sviluppo dell'elaboratore elettronico, che coincidono grosso modo con la fine della seconda guerra mondiale, un enorme punto interrogativo è sorto nella mente di chi, come addetto ai lavori o come utente, ha avuto a che fare con questa nuova macchina dell'uomo: il calcolatore è paragonabile oppure no al cervello umano? Ancora oggi il termine "cervello elettronico" è presente nei titoli dei giornali, per lo più in senso derisorio quando tassa qualche operaio per redditi da nababbo oppure commette qualche altro sbaglio. Ma può "sbagliarsi" un elaboratore?

C'è certamente un punto di vista sotto il quale il calcolatore è analogo a un cervello: ambedue elaborano informazioni, che pervengono loro da organi esterni di senso, e ambedue le restituiscono elaborate ad altri organi esterni. Sono cioè "elaboratori" di informazioni e non semplici "calcolatori" (cioè macchine calcolatrici, come il mitico Analytical Engine di Charles Babbage). E poichè gli elaboratori sono macchine costruite dall'uomo in base a componenti sempre diversi e più perfezionati (tubi elettronici, transistori, circuiti integrati, giunzioni Josephson, ecc.), che obbediscono però sempre alla logica binaria dell'algebra di Boole, ecco che l'analogia fra il cervello umano, una "scatola nera" di cui si sa ben poco in quanto ha troppi componenti (circa 10^{11} neuroni, connessi fra di loro in maniera apparentemente caotica), e un elaboratore, di cui si sa tutto perchè lo si è costruito, ha dato vigore a una scuola di pensiero che dice sostanzialmente: studiamo il funzionamento del calcolatore come elaboratore di informazioni e capiremo il funzionamento del cervello. Questo è visto come un particolare elaboratore, costruito con neuroni invece che con transistori, ma basato sulla stessa logica binaria. Ogni neurone è infatti un sistema bistabile, esattamente come un circuito flip-flop.

Questo punto di vista è chiaramente formulato da H.A. Simon e A. Newell, due dei suoi più influenti sostenitori [1]:

"C'è un corpo crescente di indizi che fanno ritenere come i processi informativi elementari usati dal cervello umano nel pensare siano molto simili a un sottoinsieme dei processi informativi elementari che sono incorporati nei codici di istruzioni dei calcolatori di oggi. Come conseguenza, si è trovato che è possibile verificare le teorie sull'elaborazione di informazioni da parte di esseri umani formulando queste teorie come programmi di calcolatori – organizzazioni dei processi elementari informativi – ed esaminando gli output dei calcolatori così programmati. Questo procedimento non assume alcuna similitudine fra il calcolatore e il cervello a livello di hardware, ma solo una similitudine nelle loro capacità di eseguire e organizzare processi elementari di informazione. Da questa ipotesi è nata una fruttuosa collaborazione fra la ricerca nell'"intelligenza artificiale", mirante ad ampliare le capacità degli elaboratori, e la ricerca sulla psicologia dei processi cognitivi umani".

Partendo da queste premesse Simon e Newell giungono a un modello del cervello umano come solutore di problemi: il General Problem Solver. Esso è un programma che tende a simulare il comportamento di un essere umano che si trova di fronte a un problema di cui non conosce il metodo per arrivare alla soluzione, come a esempio in una partita a scacchi.

In effetti, il gioco degli scacchi è uno dei campi in cui si sono fatti i maggiori progressi in fatto di intelligenza artificiale, tanto che sono in vendita calcolatori programmati per giocare a scacchi a vari livelli di difficoltà.

Ma basta, questo, per concludere che il problema dell'intelligenza artificiale è risolto, almeno in via teorica, e che, dato un calcolatore abbastanza potente, potremo simulare il comportamento di un essere umano in qualsiasi circostanza? Un punto di vista del tutto opposto è sostenuto da J. Weizenbaum [2], e, del resto, le previsioni

fatte negli anni Cinquanta e Sessanta su ciò che i calcolatori avrebbero saputo fare negli anni Settanta, a esempio in tema di traduzione automatica, si sono dimostrate del tutto infondate [3].

Ciò nonostante, è bene evitare di saltare a conclusioni premature. C'è tutto un fiorire di ricerche nel campo dell'intelligenza artificiale, che ha portato a sviluppi scientifici importanti e anche alle prime applicazioni industriali. Questo articolo è una breve rassegna di questi sviluppi, con particolare riguardo a ricerche in corso in Italia.

2. Cos'è l'intelligenza artificiale

Secondo D. Marr e H.K. Nishihara, dell'Artificial Intelligence Laboratory del MIT, "l'intelligenza artificiale è (o dovrebbe essere) lo studio dei problemi di elaborazione dell'informazione che hanno le loro radici caratteristiche in qualche aspetto dell'elaborazione biologica delle informazioni" [4]. Abbiamo accennato alla soluzione di problemi e ai processi cognitivi; un altro campo, di vitale importanza per gli esseri viventi, in quanto da esso dipende spesso la loro stessa esistenza, è quello del riconoscimento delle forme, o *pattern recognition*, secondo il termine inglese di uso corrente. Si pensi all'animale che deve riconoscere la preda (o il predatore), all'uomo che deve identificare un volto o una voce, oppure distinguere un missile balistico da un volo di anatre su uno schermo radar... Sono questi processi di riconoscimento in cui il cervello animale ha una abilità superiore rispetto all'elaboratore, che lo batte invece per la velocità con cui elabora dati numerici. Ciò è dovuto, probabilmente, alla natura eminentemente parallela e fortemente ridondante dell'elaborazione delle informazioni del cervello, rispetto a quella eminentemente seriale dell'elaborazione nel calcolatore, almeno quello "classico". Non a caso, i primi elaboratori a struttura parallela (i *parallel processor*) furono sviluppati in gran parte entro programmi di riconoscimento di immagini (ottiche e radar) in tempo reale, per scopi scientifici e anche militari, quali la difesa antiaerea e quella antimissile. Si diceva un tempo che la funzione sviluppa l'organo: evidentemente ciò è vero anche per l'organo artificiale chiamato elaboratore.

Non a caso, quindi, uno dei multielaboratori più avanzati, interamente progettato e costruito in Italia, l'EMMA (Elaboratore Multi Mini Associativo) della Elettronica San Giorgio - Elsag S.p.A. di Genova, è stato sviluppato proprio in relazione a problemi di riconoscimento di caratteri ottici (OCR: Optical Character Recognition) per la lettura automatizzata di indirizzi postali nel SARI (Sistema Automatico Riconoscimento Indirizzi), sviluppato dall'Elsag per le Poste italiane, poi adottato anche da quelle francesi e sperimentato da quelle americane.

EMMA è un multielaboratore a struttura gerarchica, cioè organizzato in "famiglie" costituite ciascuna da un gran numero di moduli base appartenenti a tre sole varietà fondamentali: unità di calcolo, unità di memoria e unità di coordinamento degli scambi di dati. Ciascuna unità è realizzata fisicamente da una scheda a circuito stampato contenente i circuiti integrati e i componenti passivi; le schede sono collegate fra di loro da un "bus" che permette lo scambio dei dati, e tutto l'insieme è modulare ed espandibile secondo le esigenze. La struttura gerarchica di EMMA è concettualmente simile a quella di uno dei più avanzati multielaboratori scientifici americani, il Cm della Carnegie-Mellon University, ma, a differenza di questo, EMMA è un prodotto industriale.

Non ci addentreremo qui, tuttavia, in problemi di struttura hardware e software dei multielaboratori. Vogliamo solo mettere in evidenza che il fatto di dover affrontare un problema di intelligenza artificiale, secondo la definizione di Marr e Nishihara, ha portato allo sviluppo di un particolare tipo di elaboratore, con caratteristiche di affidabilità, tolleranza ai guasti e ridondanza che ricordano alcune delle caratteristiche del sistema nervoso centrale degli animali superiori e in particolare dell'uomo.

3. Alcune applicazioni del pattern recognition

Il pattern recognition può essere definito come quella tecnica dell'intelligenza artificiale che mira a classificare enti diversi (oggetti, processi, fenomeni) in classi o categorie con misure fatte su questi enti.

I problemi del pattern recognition sono stati generalmente affrontati secondo due metodologie diverse, quella statistica, basata sulla teoria delle decisioni, e quella sintattica.

Nella prima, si tende ad estrarre delle "caratteristiche" (*features*) dai pattern in esame e a costruire uno spazio delle caratteristiche, su cui si opera con metodi statistici, raggruppando i pattern in classi (*cluster*) a esempio con misure di distanza basate su metriche opportunamente definite. Questa linea d'attacco fu storicamente la prima a essere usata ed è quella impiegata nella maggior parte dei sistemi industriali.

La seconda metodologia, quella sintattica, ha un grande interesse concettuale, in quanto è legata alle teorie linguistiche di Noam Chomsky. In essa un pattern complesso viene decomposto in una serie di pattern più semplici, chiamati primitive, esattamente come una frase viene analizzata in componenti elementari che obbediscono a particolari regole sintattiche. Secondo il metodo della "grammatica generativa trasformazionale" di Chomsky, una volta scomposto il pattern nelle sue primitive si cercano delle regole ("produzioni") con cui a partire dal pattern in esame si possono generare tutti i pattern della stessa classe. Dunque due pattern appartengono alla stessa classe quando le frasi che li descrivono sono trasformabili l'una nell'altra. Questa linea d'attacco ha dato rinnovato impulso a studi di linguistica computazionale, ma i suoi sviluppi pratici sono stati finora scarsi.

Nel corso delle ricerche eseguite all'Elsag nel campo del riconoscimento dei caratteri ottici, è stata sviluppata una terza metodologia, quella linguistica-semantica.
Essa si basa sui seguenti concetti fondamentali:

1) il riconoscimento umano sfrutta tutti i livelli informativi, specialmente quelli più alti. Non è quindi indispensabile riconoscere il singolo carattere per identificare correttamente una parola, in quanto il riconoscitore umano fa largamente uso del contesto, specie se la parola è inclusa in una frase;

2) bisogna cercare di mantenere il concetto di misure di distanza, come nella metodologia decisionale, applicandolo però a un livello semantico.

Le fasi successive di un simile metodo di riconoscimento possono essere schematizzate come segue:

a) estrazione delle primitive dal pattern e loro descrizione in termini di vettori di distanza da archetipi conservati in memoria;

b) analisi top-down di un "vocabolario di pattern" per associare stringhe di primitive alle parole del "vocabolario di pattern";

c) calcolo delle distanze delle parole di input in funzione delle distanze definite al punto a), cioè delle distanze input-vocabolario;

d) analisi sintattica delle stringhe di parole nell'ambito della "grammatica" dei pattern per ottenere possibili "frasi corrette". Identificazione di eventuali sottoinsiemi semanticamente corretti;

e) calcolo delle distanze input-frase e input-sottoinsiemi di frase in funzione delle distanze input-vocabolario;

f) riconoscimento delle frasi (o sottoinsiemi di frasi) semanticamente corrette in base a un criterio di distanza minima.

L'applicazione di questo metodo al compito richiesto dal SARI, cioè il riconoscimento di una frase del tipo:

20090 Rodano (Mi)

che è l'ultima riga di un indirizzo postale standardizzato in Italia, ha permesso di leggere più di 30.000 indirizzi l'ora con grandissima affidabilità su posta avente determinati requisiti minimi, cioè indirizzi dattiloscritti o stampati, come in gran parte della posta commerciale.

I sistemi SARI sviluppati per le Poste francesi leggono anche la penultima riga dell'indirizzo, cioè la via e il numero civico, mentre il SARI-USA sviluppato per le Poste americane è un lettore di codice numerico (ed è questo il compito più difficile, in quanto non può fare uso del contesto).

Lo stesso metodo è applicato a sistemi in fase di sviluppo per la lettura di documenti, quali documenti di bancoposta, e per quella di fotogrammi di contatori telefonici. Tutti questi sistemi impiegano multielaboratori EMMA.

4. Il riconoscimento della voce

Un altro problema di pattern recognition su cui molto lavoro è stato già fatto e di cui cominciano ad apparire le prime soluzioni industriali è quello del riconoscimento della voce; un problema correlato, che ha anche risvolti giudiziari, è quello dell'identificazione del parlatore. Qui entriamo in un campo caro agli autori di fantascienza, quello cioè del dialogo diretto uomo-macchina (è d'obbligo la citazione di HAL, il calcolatore parlante di "2001 Odissea nello spazio", il bel film di Kubrick). Se, infatti, potessimo programmare a voce un calcolatore, usando il linguaggio naturale o un suo sottoinsieme, la macchina perderebbe gran parte della sua natura di... macchina; se poi essa a sua volta rispondesse a voce, non sarebbe più una macchina. Ma a che punto stanno le cose?

I primi esperimenti fatti negli anni Cinquanta sul riconoscimento delle vocali e dei numeri avevano suscitato molte speranze, che sono andate gradatamente spegnendosi quando ci si è resi conto delle enormi difficoltà che sorgevano nel riconoscimento di intere parole e, tanto più, del parlato connesso. Solo negli ultimi cinque o sei anni si è avuto un nuovo impulso agli studi sul riconoscimento vocale, dovuto a un progresso nell'analisi acustica del segnale vocale e nella sua rappresentazione parametrica, a una maggiore conoscenza dei fenomeni fonologici (interazione fra fonemi vicini e cambiamento nella loro realizzazione acustica) e linguistici, alla presa di coscienza dell'importanza del contesto sulla comprensione delle singole parole e infine anche al progresso generale degli elaboratori.

I sistemi di riconoscimento della voce possono essere distinti in:

1) sistemi di riconoscimento di parole isolate, in cui le parole sono pronunciate una per volta, con un intervallo minimo di separazione;

2) sistemi di riconoscimento del parlato connesso, in cui nasce il grosso problema di suddividere in parole il segnale continuo, nonchè quello della sparizione o comparsa di fonemi alla giunzione fra le due parole;

3) sistemi di comprensione della parola, in cui si fa uso, per risolvere le ambiguità che nascono dal parlare connesso, delle conoscenze riguardanti il contesto. Questi sistemi presuppongono cioè che il segnale vocale possa mancare di alcune informazioni necessarie per una decodifica univoca del messaggio, per cui occorre ricorrere al contesto. In compenso il criterio di riconoscimento è meno rigido, nel senso che, se il nesso del messaggio è compreso, non importa se qualche fonema o qualche parola non sono stati riconosciuti correttamente.

Si è parlato di più livelli di conoscenza. Infatti, quando ascoltiamo una frase facciamo uso di molte conoscenze, e cioè:

- caratteristiche dei segnali vocali (fonetica)
- variabilità nella pronuncia (fonologia)
- andamenti di accento e intonativi (prosodia)
- pronuncia di ogni parola (lessico)
- struttura grammaticale del linguaggio (sintassi)
- significato di parole e frasi (semantica)

– contesto della conversazione (pragmatica).

I sistemi di riconoscimento di parole isolate esigono che le parole vengano pronunciate isolatamente, con pause di almeno 100 ms.

La velocità massima può essere di 120 parole/minuto per lettori addestrati che leggono cifre isolate, mentre il ritmo medio è di 30/70 parole/minuto.

I sistemi di riconoscimento di parole isolate hanno in genere come strategia quella di confrontare la rappresentazione parametrica della parola ricevuta con quella di tutte le parole di un vocabolario.

L'analisi delle parole è eseguita in genere in termini di analisi dello spettro di potenza del segnale acustico, anche se in effetti si possono esaminare gli andamenti di insiemi diversi di parametri, come:

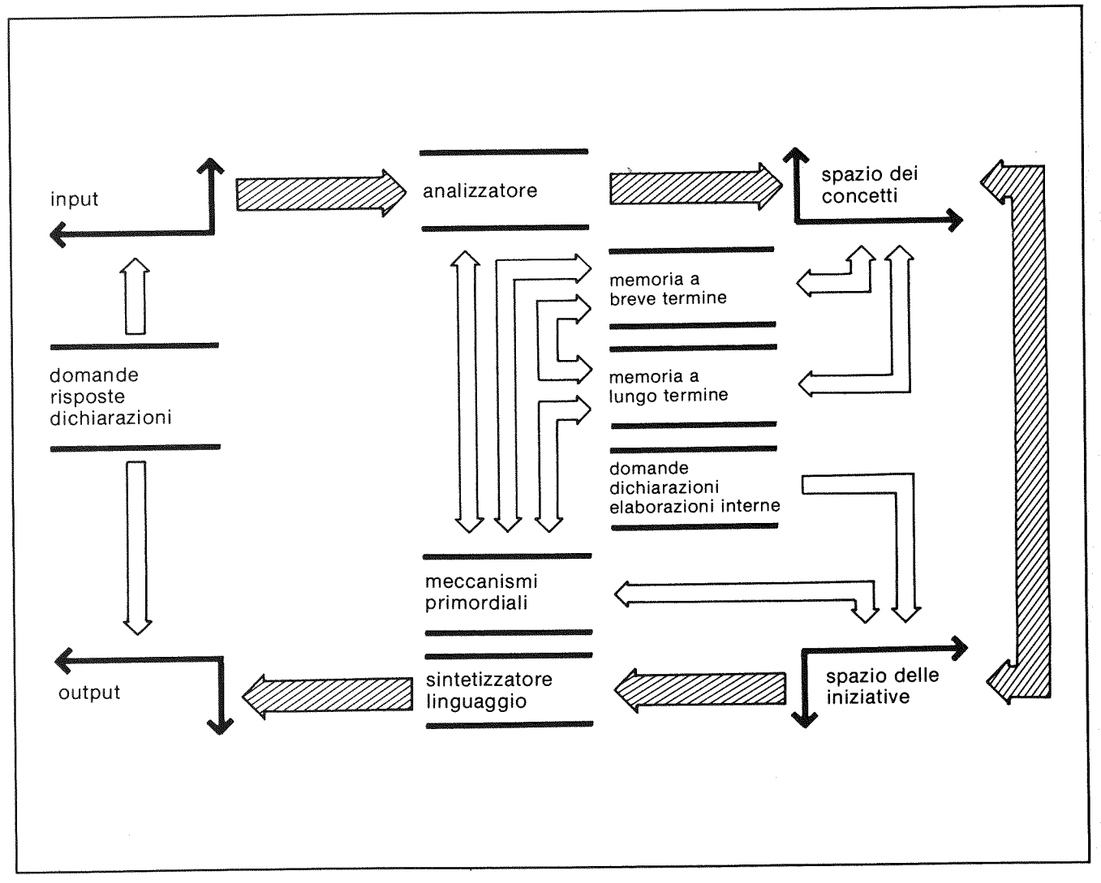
– ampiezze delle uscite di un banco di filtri
– frequenze dei formanti
– coefficienti ricavati dalla predizione lineare
– coefficienti di riflessione del tubo acustico equivalente al tratto vocale.

Da questa analisi si ottiene una rappresentazione parametrica del segnale che impiega da poche migliaia di bit/s a circa 600 bit/s. Essa dev'essere poi normalizzata per tener conto delle variazioni di pronuncia, anche dello stesso parlatore. Infine, per la classificazione finale della parola, viene eseguito il confronto con tutte le descrizioni parametriche di riferimento. Questo confronto porta a una misura della distanza tra la parola sconosciuta e quella di riferimento. La parola di riferimento cui compete la distanza minima viene identificata con quella incognita, purchè questa distanza non superi una certa soglia minima, oltre alla quale si ha un rifiuto.

In generale questi sistemi vengono usati con vocabolari limitati, dell'ordine delle decine di parole, e sono tarati sul singolo parlatore, cosa possibile dato il vocabolario limitato. Le prestazioni che si ottengono raggiungono una percentuale di riconoscimento del 99%, anche in condizioni di lavoro reali.

I sistemi di riconoscimento del parlato connesso devono, in base ancora a una rappresentazione parametrica della voce, eseguire un

Fig. 1 - Schema del sistema LISA (Linguistica Semantica Avanzata).



processo di segmentazione del segnale in unità elementari e assegnare a ciascuna unità il nome appropriato in base al contenuto spettrale del segmento. Dopo questa operazione un confronto con un dizionario fonetico potrà portare all'identificazione delle parole pronunciate. Tuttavia, non solo la suddivisione in parole può non essere univoca, ma segmentazione ed etichettatura possono aver portato a errori come omissioni di fonemi, aggiunta di fonemi o errori nell'assegnazione del nome.

È quindi necessario, al momento della segmentazione ed etichettatura, considerare varie alternative e generare così vari percorsi, ciascuno con un certo grado di affidabilità, così da ottenere la suddivisione più verosimile. In questo processo entrano in gioco tutti i livelli di conoscenza, inclusi quelli più elevati (sintassi, semantica, pragmatica) che non vengono utilizzati nei sistemi di riconoscimento di parole isolate.

I sistemi di comprensione della voce cercano di sfruttare in massimo grado le conoscenze semantiche e pragmatiche per comprendere il significato della frase anche quando l'enunciato presenti qualche errore grammaticale, o in presenza di rumori.

Le applicazioni pratiche dei sistemi di riconoscimento della voce si limitano per lo più a quelli di parole isolate. Esse sono in genere:

1) ingresso dati da parte di addetti a un controllo di qualità;

2) controllo di apparecchiature di trattamento dei materiali;

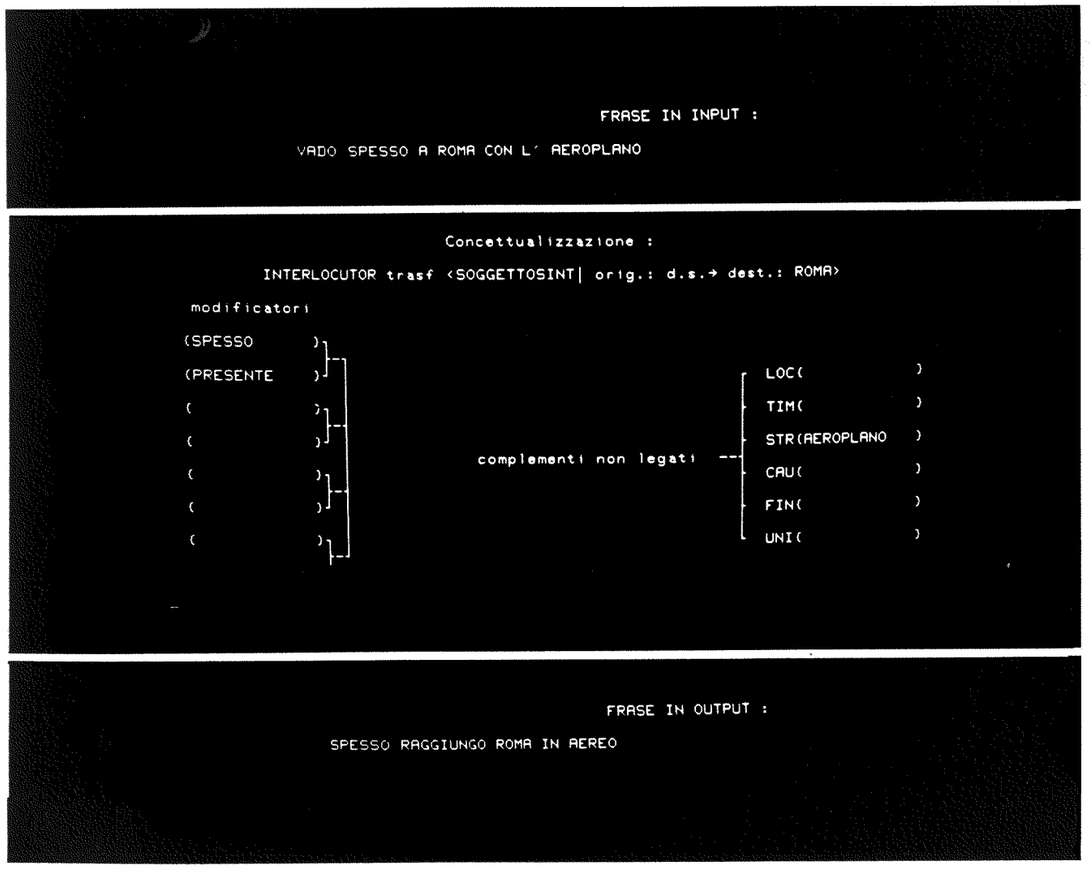
3) programmazione di macchine a controllo numerico;

4) aggiornamento di informazioni.

Un sistema giapponese capace di riconoscere fino a frasi di tre parole è stato applicato alla prenotazione dei posti sulla linea ferroviaria del Tokaido (il famoso rapido Tokio-Osaka), mentre in America è allo studio un sistema di addestramento dei controllori del traffico aereo.

Un riconoscitore di parole isolate sviluppato all'Elsag si basa su di un analizzatore sintattico (parser) realizzato nell'ambito della Direzione Ricerca Centralizzata. Esso permette di comandare a voce un robot simulato, cioè un punto luminoso che si muove sullo schermo di un tubo a raggi catodici entro un labirinto, ed obbedisce a

Fig. 2 - Esperimenti con l'analizzatore ed il sintetizzatore di LISA.



una gamma di comandi quali: vai a sinistra, vai a destra, salta il muro, ecc.

Anche nel campo del riconoscimento della voce è difficile separare il campo del pattern recognition vero e proprio da quello della linguistica computazionale; ciò rende la ricerca ancora più interessante in quanto ogni ricerca sulla macchina diventa anche una ricerca sul suo creatore: l'uomo.

5. Un modello di intelligenza artificiale: LISA

Per facilitare il dialogo diretto uomo-elaboratore, è necessario che l'utente non specializzato sia agevolato nei suoi rapporti con la macchina. Oltre ad agire nel senso di permettere un dialogo a voce, è possibile facilitare ancora l'interazione uomo-macchina se l'uomo si dimentica di avere di fronte una macchina. E poichè una caratteristica spiacevole di queste ultime consiste nella loro ripetitività, cioè nel fatto di reagire in maniera sempre identica a un dato input, all'Elsag si è cercato di sviluppare una macchina che contenesse un quid di arbitrarietà tipicamente umano. Questa idea si è concretata nel progetto LISA (Linguistica Semantica Avanzata).

I vari elementi che compongono il progetto LISA sono rappresentati nella fig. 1. Si tratta in realtà di programmi, cioè di blocchi di software ciascuno dotato di caratteristiche appropriate, che possono essere collegati fra di loro. Due dei blocchi sono stati finora realizzati, l'analizzatore e il sintetizzatore, e collegati fra di loro in via sperimentale, mentre è in fase di realizzazione il modello di intelligenza artificiale. L'analizzatore traduce un input espresso in un sottoinsieme del linguaggio naturale italiano in una rappresentazione semantica, cui, seguendo R.C. Schank [5] si è dato il nome di "concettualizzazione", adatta a essere manipolata dall'intelligenza artificiale.

Gli input ammessi sono frasi enunciative o interrogative semplici, cioè proposizioni principali formate a partire dai vocaboli presenti nella memoria della macchina. Caratteristica fondamentale di questo sistema è l'uso interconnesso di criteri sintattici e semantici, che intervengono assieme nei vari punti dell'analisi.

Gli elementi della concettualizzazione, o "elementi concettuali", sono stati divisi in tre tipi: operatori, concetti, modificatori. I concetti sono tutti gli oggetti pensabili; gli operatori indicano azioni o relazioni fra elementi concettuali; i modificatori servono per meglio specificare il significato degli altri elementi concettuali. Questi elementi si combinano nella concettualizzazione secondo alcuni modi prefissati; l'elemento centrale è l'operatore predicativo (corrispondente al predicato verbale o nominale nella frase in input) al quale è legato un primo insieme di concetti, detti appunto "complementi legati", che costituiscono il nucleo concettuale. La loro definizione dipende dal tipo di operatore.

Altri sei complementi (luogo, strumento, tempo, fine, causa, unione) si associano all'operatore come "complementi non legati"; questi sono prefissati e non dipendono dall'operatore.

Questa concettualizzazione viene analizzata da un simulatore di intelligenza artificiale (in fase di sviluppo) che la elabora o eventualmente genera una nuova concettualizzazione, partendo dalle informazioni contenute in due memorie, una a breve termine e una a lungo termine. Nella prima sono contenute le ultime cinque concettualizzazioni presentate in ingresso al simulatore o da esso generate, nella seconda le informazioni avute in precedenza relative all'ambiente semantico scelto, che è quello delle località geografiche italiane e quello dei possibili utenti di LISA. Una tipica frase di ingresso può essere: "Luigi va spesso a Roma in aereo".

Una volta definita la concettualizzazione da parte del simulatore di intelligenza artificiale, entra in gioco il sintetizzatore linguistico che traduce la concettualizzazione in linguaggio naturale italiano (Fig. 2).

Come si è accennato, è stato effettuato un interessante esperimento, quello di collegare fra di loro l'analizzatore e il sintetizzatore per verificare la loro stabilità semantica. Ciò si effettua nel modo seguente: si pone come input una frase all'analizzatore, che la concettualizza. Questa concettualizzazione viene a sua volta posta come input al sintetizzatore, che ne fa uso per produrre una frase, che a sua volta torna quale input all'analizzatore e così via di seguito. Il sistema si è dimostrato notevolmente stabile, e ciò promette bene per i successivi sviluppi.

6. Conclusioni e prospettive

Nel suo libro *Computer power and human reason*, Weizenbaum cita il fatto seguente. Avente egli realizzato al MIT un programma (ELIZA) [6] che simulava le reazioni di un psicologo analista alle affermazioni di un paziente secondo il metodo di Roger, che consiste nel riprendere essenzialmente le affermazioni del paziente e registrarne le reazioni a quanto da lui stesso affermato, e avendolo fatto provare alla sua segretaria, dopo poche risposte fu da questa pregato di... uscire dalla stanza, in quanto ella doveva confidare al calcolatore qualcosa di troppo personale per essere da lui udito (o visto).

Una reazione simile non si è ancora verificata in alcuno dei ricercatori impegnati nel progetto LISA, forse anche perchè l'ambiente semantico scelto non ha una carica emotiva tanto forte quanto può averne un programma che simuli un psicoterapeuta e le sue domande. Ma lo stesso Weizenbaum fu alquanto sbigottito nel constatare come il suo programma, che per lui era un esperimento di informatica e nulla più, fosse citato a esempio di terapia psicoanalitica finalmente posta a disposizione di tutti grazie ai progressi dell'elettronica.

Esisterà mai, una vera "intelligenza artificiale"? Le opinioni su questo punto variano enormemente.

Un cibernetico come M. Schutzenberger lo nega recisamente. Dall'altro lato del problema, quello cioè che riguarda l'intelligenza umana, un neurofisiologo come K. Pribram fa uso di un modello olografico, basato sulla trasformata di Fourier, per spiegare i meccanismi della memoria umana, su cui è basata gran parte della nostra capacità di apprendere e di riconoscere forme, tipiche manifestazioni dell'intelligenza.

Scrive ancora Weizenbaum [2]:

"C'è un aspetto della mente umana, l'inconscio, che non può essere spiegato dalle primitive che elaborano informazioni, i processi informativi elementari, che associamo con il pensiero formale, i calcoli e la razionalità sistematica. Eppure siamo costretti a usarli per la spiegazione, la descrizione e l'interpretazione scientifiche. Non ci resta altro, quindi, che renderci conto della povertà delle nostre spiegazioni e del loro ambito strettamente limitato. È sbagliato asserire che la scienza potrà rendere conto dell'"uomo intero'. Vi sono cose che vanno al di là dell'ambito della comprensione scientifica".

L'avvenire darà una risposta alla domanda sull'intelligenza artificiale e una sua possibile realizzazione. Intanto le "macchine cibernetiche", dai robot industriali che saldano le scocche delle automobili ai lettori di indirizzi postali che smistano la nostra corrispondenza, sono destinate ad assumere un ruolo crescente nella nostra vita.

7. Bibliografia:

[1] A. NEWELL e H.A. SIMON, *Human problem solving*, Prentice-Hall (1972).

[2] J. WEIZENBAUM, *Computer power and human reason*, Freeman (1976).

[3] H. DREYFUS, *What computers can't do*, Harper and Row (1972).

[4] D. MARR e H.K. NISHIHARA, *Visual information processing: artificial intelligence and the sensorium of sight*, in Technology Review, ottobre 1978 (MIT).

[5] R.C. SHANK e K.M. COLBY (ed.), *Computer models of thought and language*, Freeman (1973).

[6] Si veda: Quaderni di Informatica N. 4 (1975), pag. 27 e seg.

Tecnologia dell'elaboratore: la storia del MOS

TOM HENDRICKSON
Honeywell Inc.
Solid State Electronic Center
Minneapolis (USA)

1. Introduzione

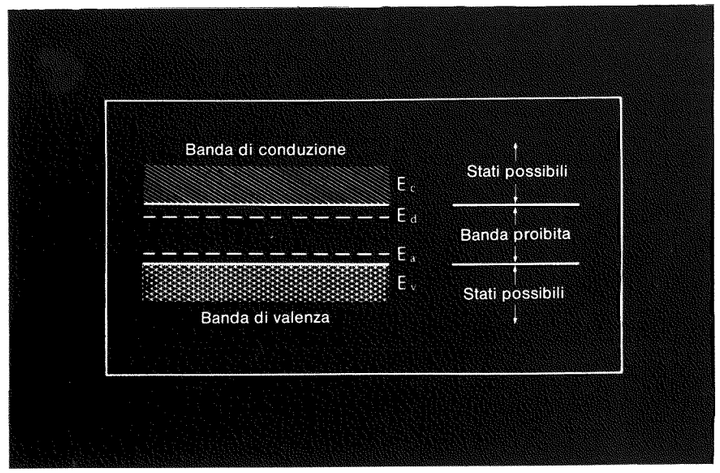
Agli inizi degli anni cinquanta la Philco investì 50 milioni di dollari in una linea di produzione per un dispositivo a stato solido sviluppato dai suoi ricercatori, il transistore a microdiffusione di lega. Nel 1955 i laboratori Bell annunciarono il transistore "mesa" (il precursore del transistore planare) e la Philco si trovò ad avere perso il suo investimento in una tecnologia superata. Di fatto la tecnologia dello stato solido evolse così rapidamente dopo la concessione nel 1928 del primo brevetto sul transistore, che molte società operanti nel settore soffrirono simili perdite. Tutte le società già operanti nel campo delle valvole convenzionali (tubo a vuoto) si dedicarono alla tecnologia dello stato solido, ma la maggior parte di esse fu spinta fuori mercato negli anni 60 dall'avvento di nuove società per lo più concentrate nella valle di Santa Clara, California, che per il seguito divenne più nota come "Silicon Valley", la valle del silicio. Fino al 1974 circa, la tecnologia dominante era quella bipolare, ma ora la tecnologia MOS si sta sostituendo a quella bipolare con la stessa rapidità con cui il transistore mesa soppiantò il transistore a microdiffusione di lega. Questi cambiamenti illustrano drammaticamente la difficoltà di prendere la decisione di seguire una particolare tecnologia a stato solido.

Di fatto, la tecnologia MOS è molto più vecchia di quella bipolare. Era già noto nel ventennio 1920-1940 che una corrente poteva essere controllata in un semiconduttore mediante l'applicazione di un campo elettrico (proprietà su cui si basa la tecnologia MOS) mentre il transistore bipolare non fu sviluppato che nel 1948. Tuttavia fino a pochi anni fa, insuperabili problemi di produzione impedirono ai dispositivi MOS di conquistare una significativa parte del mercato dei dispositivi a stato solido. Ora che i maggiori problemi di produzione sono stati superati la tecnologia MOS promette di divenire quella dominante per i dispositivi a stato solido. Una breve storia del transistore MOS può dare un'idea dello sviluppo della tecnologia a stato solido così imprevedibile da tradire qualsiasi sforzo di previsione.

2. Fondamenti di tecnologia

L'esperienza di ogni giorno ci insegna che la resistenza dei materiali al passaggio di una corrente elettrica varia ampiamente. Poichè la resistenza varia anche, indipendentemente dal materiale, con la sezione e la lunghezza del conduttore, si definisce una quantità, *la resistività*, che è la resistenza di un materiale in Ohm-cm per un elemento di sezione e lunghezza unitaria elettricamente omogeneo. La resistività varia per più di 40 ordini di grandezza dai superconduttori agli isolanti di alta qualità. La resistività di ogni materiale è funzione della differenza tra l'energia posseduta dagli elettroni liberi e quella dagli elettroni legati nel materiale. Quando più atomi si condividono gli stessi elettroni, come è nel caso dei cristalli, ogni elettrone è sottoposto a forze più complesse di quelle esercitate su di lui in un singolo atomo e i livelli di energia dell'atomo vengono suddivisi in una pluralità di livelli energetici possibili, dando luogo a bande di energia usate per descrivere i possibili livelli energetici degli elettroni in un solido. I metalli sono caratterizzati da bande energetiche contigue, parzialmente occupate da elettroni: gli elettroni che si trovano ai più bassi livelli energetici di queste bande richiedono solo un modesto incremento di energia per salire nella banda, ossia per svincolarsi dal loro legame e per divenire elettroni liberi. Nei semiconduttori e negli isolanti invece, una ampia banda proibita separa i livelli energetici occupati da elettroni dai livelli energetici non occupati dagli elettroni. In altre parole, la banda di valenza è separata dalla banda di conduzione per effetto della presenza di una banda proibita, di ampiezza caratteristica per ogni materiale. L'ampiezza della banda proibita determina la quantità di energia che è necessario fornire a un elettrone per liberarlo dal reticolo cristallino e renderlo disponibile come portatore di carica elettrica. Dunque gli isolanti hanno bande proibite di ampiezza più larga di quella dei semiconduttori. Tipicamente gli isolanti hanno bande proibite pari a diversi eV (electron-Volt), mentre per i semiconduttori l'ampiezza della banda proibita è approssimativamente di un eV.

Fig. 1 - Diagramma delle bande di energia in un semiconduttore: esso mostra il limite inferiore della banda di conduzione (Ec), il livello energetico dei donatori (Ed) entro la banda proibita, il livello energetico degli accettori (Ea) e il limite superiore della banda di valenza (Ev).



In un semiconduttore intrinseco (ossia un semiconduttore puro, che non è stato drogato) l'agitazione termica degli atomi contigui libera qualche elettrone dal reticolo cristallino dandogli la possibilità di muoversi nel materiale e perciò di trasferire una carica elettrica. Ogni transizione di un elettrone da una banda di valenza a quella di conduzione, libera due portatori di cariche, l'elettrone e la "lacuna" o buco ossia il legame vuoto lasciato privo di carica elettrica nel reticolo cristallino, quando un elettrone se si svincola dal legame. Le "lacune" sono una astrazione teorica e non delle particelle e corrispondono a uno squilibrio elettrico tra elettroni dell'atomo e protoni del nucleo atomico. Le "lacune" sono mobili perchè un atomo del reticolo privo di un elettrone nella banda di valenza può sottrarre un elettrone della banda di valenza a un atomo contiguo, riempiendo la sua lacuna ma provocando una "lacuna" nell'atomo contiguo. Pertanto la corrente in un conduttore è determinata da elettroni che si muovono in una direzione nella bamda di conduzione e da "lacune", corrispondenti a cariche positive, che si muovono in direzione opposta nella banda di valenza.

In un semiconduttore intrinseco, la sola sorgente di portatori di carica è l'eccitazione termica e così gli elettroni e le lacune esistono in numero eguale. È tuttavia possibile "drogare" il materiale semiconduttore, silicio per esempio, con elementi che abbiano un numero di elettroni di valenza maggiore (elementi donatori) o minore (elementi accettori) di quelli del semiconduttore (quattro nel caso del silicio). Se l'elemento usato per drogare il silicio ha cinque elettroni di valenza (fosforo e arsenico sono comunemente usati come donatori) ogni atomo di drogante si inserirà nel reticolo cristallino con un elettrone sovrabbondante rispetto a quelli richiesti per soddisfare i quattro legami di covalenza richiesti dagli atomi di silicio contigui. Questo elettrone sovrabbondante è solo debolmente trattenuto dalla carica positiva in eccesso del nucleo atomico dell'elemento drogante. Perciò una quantità di energia ridotta, inferiore a quella normalmente necessaria per eccitare un elettrone dalla banda di valenza a quella di conduzione è richiesta in questo caso per eccitare l'elettrone in eccesso. In effetti, gli atomi di impurezze contenute nel semiconduttore aggiungono uno stato energetico possibile nella banda proibita, al di sotto della banda di conduzione del cristallo semiconduttore (Fig. 1). La maggioranza dei portatori elettrici nel silicio così drogato è costituita da elettroni e il silicio è detto di tipo n (n sta per negativo). Quando il drogante ha tre elettroni di valenza (boro), uno dei quattro legami covalenti degli atomi contigui di silicio rimane insoddisfatto. In questo caso una quantità di energia ridotta è necessaria per eccitare un elettrone dalla banda di valenza al livello energetico libero possibile prodotto dall'atomo di impurezza accettore, in conclusione muovendo una lacuna. Dunque il silicio drogato con boro può essere immaginato come un materiale in cui un livello energetico possibile è presente nella banda proibita poco al di sopra della banda di valenza e dove i portatori di carica elettrica sono in maggioranza positivi (lacune). Il silicio così drogato è detto di tipo p.

3. Il diodo

Se si deposita il silicio di tipo n su del silicio di tipo p, i portatori di carica all'interfaccia (o giunzione) si attirano reciprocamente e attraversano la giunzione in direzioni opposte ricombinandosi con i portatori di polarità opposta e lasciando dietro di sè immobili gli atomi ionizzati degli elementi droganti.

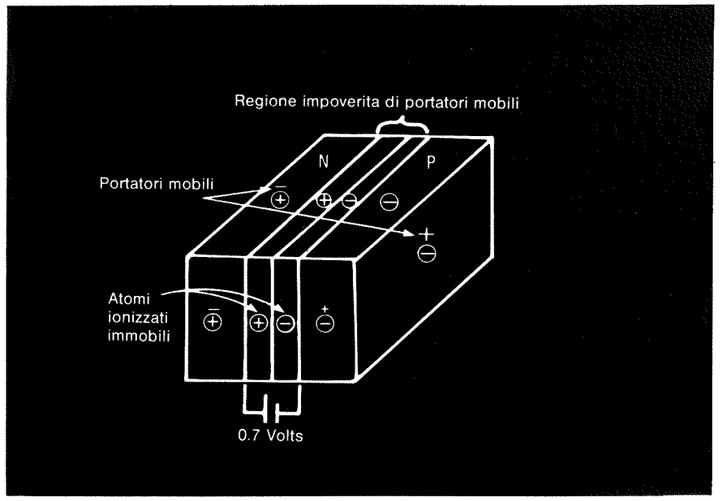


Fig. 2 - Un diodo al silicio consiste di due regioni giustapposte di silicio drogato con impurezze. Quando una regione di silicio di tipo n è in contatto con una regione di tipo p, gli elettroni liberi nel silicio di tipo n e le lacune nel silicio di tipo p attraversano la giunzione e si ricombinano con i portatori di polarità opposta provocando l'impoverimento in portatori mobili della regione prossima alla giunzione e lasciando solo degli ioni immobili. Questi ioni immobili generano una differenza di potenziale di circa 0,7 Volt tra le regioni di tipo n e p, che impedisce il flusso di portatori.

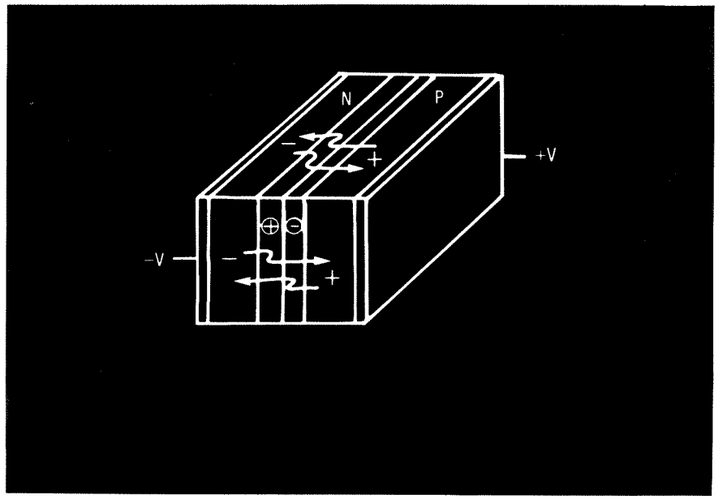


Fig. 3 - Se una tensione negativa è applicata alla regione n di un diodo al silicio, gli elettroni liberi nel silicio di tipo n e le lacune nel silicio di tipo p vengono attirati verso la regione di polarità opposta riducendo l'ampiezza della zona di impoverimento.
Quando la tensione applicata è di circa 0,) Volt, il campo elettrico è sufficientemente intenso per forzare i portatori attraverso la giunzione; il flusso di corrente nel diodo cresce rapidamente con la tensione e si dice che la giunzione è polarizzata direttamente.

Gli atomi donatori ionizzati possiederanno una piccola carica positiva netta e gli atomi accettori ionizzati possiederanno una piccola carica negativa netta. La carica localizzata così prodotta induce una differenza di potenziale tra le regioni di tipo n e p che blocca il flusso ulteriore di portatori di carica (Fig. 2). Nel silicio ciò avviene quando alla giunzione vi sono atomi ionizzati sufficienti a produrre una differenza di potenziale di 0,7 Volt tra le regioni di tipo n e p (è necessaria cioè una energia di 0,7 eV per forzare un elettrone attraverso la giunzione dalla regione n nella regione p). Se ora applichiamo una polarizzazione positiva alla regione di tipo n e una polarizzazione negativa alla regione di tipo p, i portatori di carica maggioritari saranno attratti lontano dalla giunzione, la zona di svuotamento, ossia la zona priva di portatori, alla giunzione aumenterà la sua ampiezza e nella giunzione non fluirà alcuna corrente (per essere precisi vi sarà una bassissima corrente inversa causata dal movimento attraverso la giunzione di lacune generate per eccitazione termica nel silicio di tipo n e di elettroni generati per eccitazione termica nel silicio di tipo p, ossia dai portatori minoritari). In queste condizioni si dice che il diodo è *polarizzato inversamente*. Se invece applichiamo una tensione negativa alla regione di tipo n e una tensione positiva alla regione di tipo p, portatori maggioritari attraverseranno la giunzione in ambedue le direzioni (purchè la tensione applicata sia maggiore di 0,7 Volt). Ne risulta allora una corrente relativamente intensa e si dice che il diodo è *polarizzato direttamente* (Fig. 3).

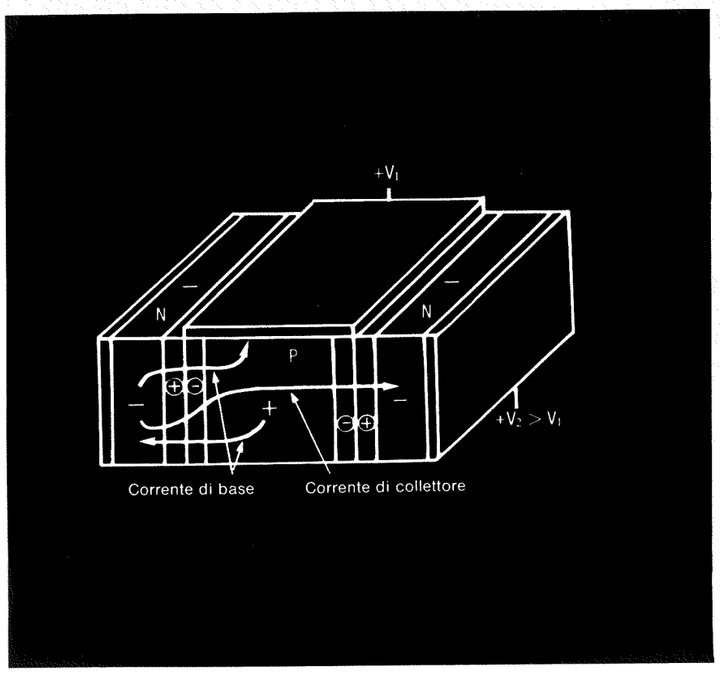
Un diodo non è passibile di amplificazione e perciò è un elemento passivo anzichè attivo. Però, a differenza di altri elementi circuitali è asimmetrico, presentando una bassa resistenza per segnali elettrici di una polarità e una resistenza elevata per segnali di polarità

opposta. Esso è l'equivalente a stato solido e porta lo stesso nome del diodo a vuoto. L'equivalente del triodo è invece il transistore, che è l'elemento attivo costruttivo di base dei circuiti elettronici.

4. Il transistore bipolare

I primi transistori erano costituiti dall'accoppiamento di due diodi contrapposti e davano luogo a dispositivi sia di tipo npn sia di tipo pnp. In ambedue i casi la regione da cui proviene la corrente è denominata emettitore, la regione intermedia base e la regione da cui viene estratta la corrente collettore. Se il primo diodo in un transistore npn (il diodo emettitore/base) è polarizzato direttamente, una corrente fluisce nella giunzione emettitore/base. Se la base è sufficientemente sottile, la maggior parte dei portatori di carica si muove attraverso la base e raggiunge la giunzione base/collettore. Se questa giunzione è polarizzata inversamente (ossia se il collettore è a potenziale elettrico più alto del potenziale di base) gli elettroni che arrivano alla giunzione base/collettore vengono attirati attraverso la giunzione nel collettore. Occorre ricordare che una piccolissima corrente fluisce attraverso un diodo polarizzato inversamente, dovuta ai portatori minoritari generati per agitazione termica. In questo caso invece portatori minoritari (elettroni) sono iniettati nella base in abbondanza attraverso la giunzione emettitore/base. Gli elettroni presenti nella base e le lacune presenti nel collettore vengono attratti attraverso la giunzione base/collettore dando luogo a una corrente rilevante (Fig. 4). Tuttavia alcuni degli elettroni iniettati nella base dall'emettitore si ricombinano con le lacune della base prima di raggiungere la giunzione di collettore, riducendo il numero di portatori di carica nella base. Poichè ambedue le giunzioni sono polarizzate in modo che nessuna delle regioni di tipo n può fornire lacune alla base, queste devono essere fornite da una connessione esterna con la base. Il rapporto tra la corrente che fluisce attraverso la base nel collettore ed esce da questo, e la piccola corrente che alimenta la base dall'esterno è il guadagno in corrente del transistore. Nei transistori moderni questo guadagno varia da 100 a 1000.

Fig. 4 - Un transistore bipolare NPN consiste di un diodo polarizzato direttamente giustapposto a un diodo polarizzato inversamente, in altre parole di un "sandwich" di strati di silicio, rispettivamente di tipo n, p, n. La maggior parte degli elettroni forzati nella regione di tipo p dal campo applicato alla giunzione polarizzata direttamente, attraversano tale regione e sono attirati attraverso la seconda giunzione polarizzata inversamente. La sorgente di elettroni è denominata emettitore, la regione di tipo p è denominata base, e la seconda regione di tipo n è detta collettore.



5. Il MOSFET

I transistori di tipo npn e pnp sono detti transistori bipolari perchè portatori di ambedue le polarità sono richiesti nel loro funzionamento. Vi sono tuttavia, transistori unipolari che fanno uso di un solo tipo di portatori. Il tipo più comune di transistore unipolare è il Transistore a effetto di campo Metallo-Ossido-Semiconduttore, o Metal Oxide Semiconductor Field Effect Transistor, da cui l'acronimo anglosassone MOSFET. Il funzionamento del MOSFET sfrutta l'azione di controllo che può essere esercitata da un campo elettrico esterno su un semiconduttore. Si consideri un condensatore costituito da una piastra metallica (o "porta") separata da una sbarretta di silicio p da un sottile strato isolante. Se una tensione negativa è applicata alla piastra, le lacune nel silicio sono attirate alla sua superficie sotto la "porta". Se una tensione positiva è applicata alla porta, le lacune sono respinte dalla superficie (la superficie è *impoverita* di portatori maggioritari). Se la tensione è aumentata, i portatori minoritari, ossia elettroni generati termicamente, saranno attirati alla

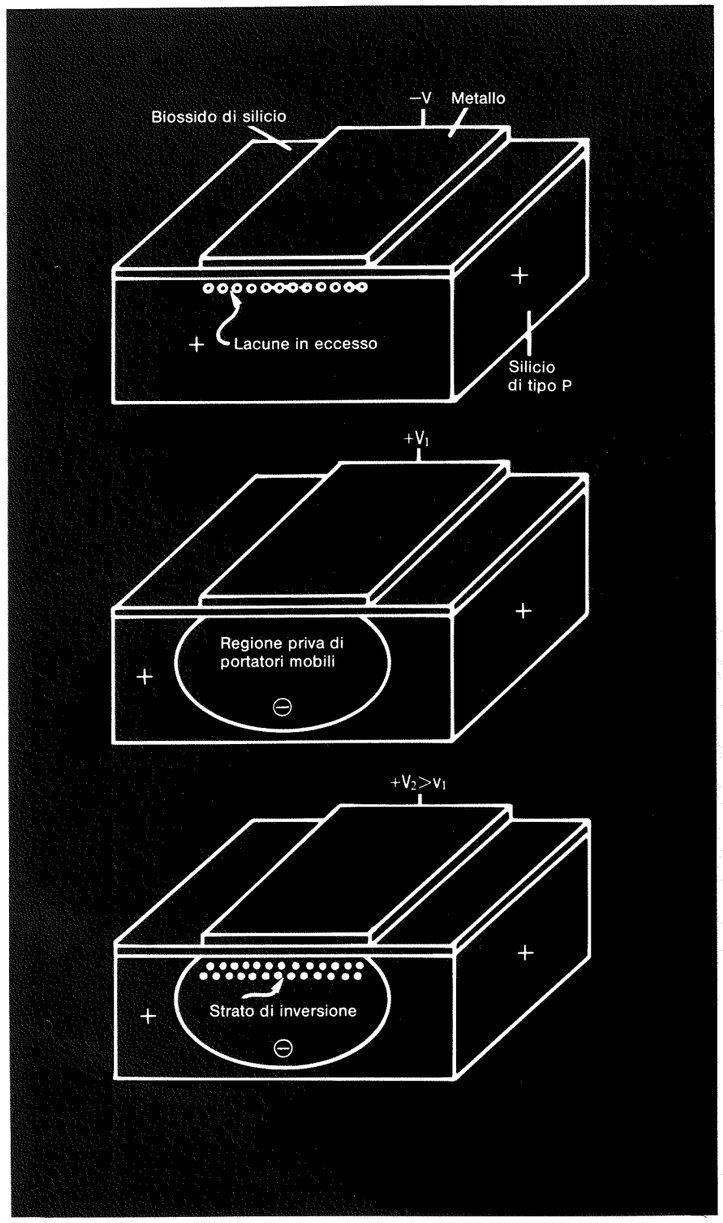


Fig. 5 - Un elettrodo isolato può essere usato per modulare i portatori alla superficie di un dispositivo al silicio.
Se il silicio è di tipo p e una tensione negativa è applicata all'elettrodo delle lacune in eccesso vengono attirate nella regione immediatamente sottostante all'elettrodo. Questo fenomeno è chiamato arricchimento. Se la tensione applicata è debolmente positiva, le lacune vengono allontanate dalla regione immediatamente sottostante all'elettrodo. Questo fenomeno è chiamato impoverimento. Se la tensione applicata è sufficientemente positiva, gli elettroni liberi, ossia i portatori minoritari generati per agitazione termica nel silicio, vengono attirati nella regione sottostante all'elettrodo. Questo fenomeno è chiamato inversione.

superficie creando un sottile *strato di inversione* (Fig. 5).
Se vi sono delle aree di silicio n alle estremità di questo condensatore o porta, si ottiene un FET o canale n del tipo ad arricchimento (il transistore è detto ad arricchimento perchè in esso la corrente fluisce solo quando la regione di canale è invertita o arricchita di portatori minoritari per inversione). Se una regione di tipo n, la sorgente, è connessa al substrato di tipo p e l'altra regione di tipo n, lo scarico, a una tensione positiva e se ancora la porta è tenuta a potenziale eguale o inferiore al potenziale del substrato, lo scarico è in effetti un diodo polarizzato inversamente e nessuna corrente fluisce nel dispositivo. Se invece la tensione applicata alla porta è sufficientemente positiva da creare uno strato di inversione alla superficie del canale, una corrente può fluire dalla sorgente allo scarico e più questa tensione è positiva tanto maggiore risulta il flusso di corrente (Fig. 6).
La tensione a cui la superficie comincia a presentare una inversione è detta *tensione di soglia.* Una transistore MOS presenta un guadagno di corrente praticamente infinito perchè la corrente nel dispositivo è modulata dalla tensione di porta e la porta è isolata dal circui-

Fig. 6 - Un Mosfet a canale n e ad arricchimento non conduce se la tensione applicata all'elettrodo di porta è negativa o inferiore alla tensione di soglia. Se la tensione è eguale a, o maggiore della tensione di soglia la regione di silicio sotto la porta viene invertita, creando un canale che consente un flusso di corrente dalla sorgente allo scarico, se una tensione positiva è applicata tra scarico e sorgente. Il canale risulta sagomato a cuneo per effetto del campo elettrico impresso dalla porta e del campo elettrico applicato tra sorgente e scarico.
Con P+ e N+ si indicano regioni fortemente drogate.

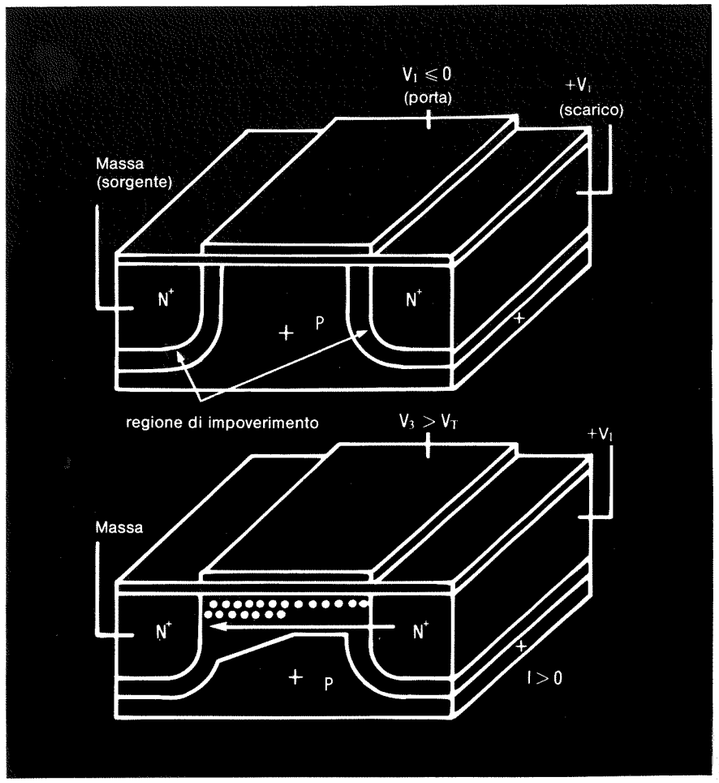
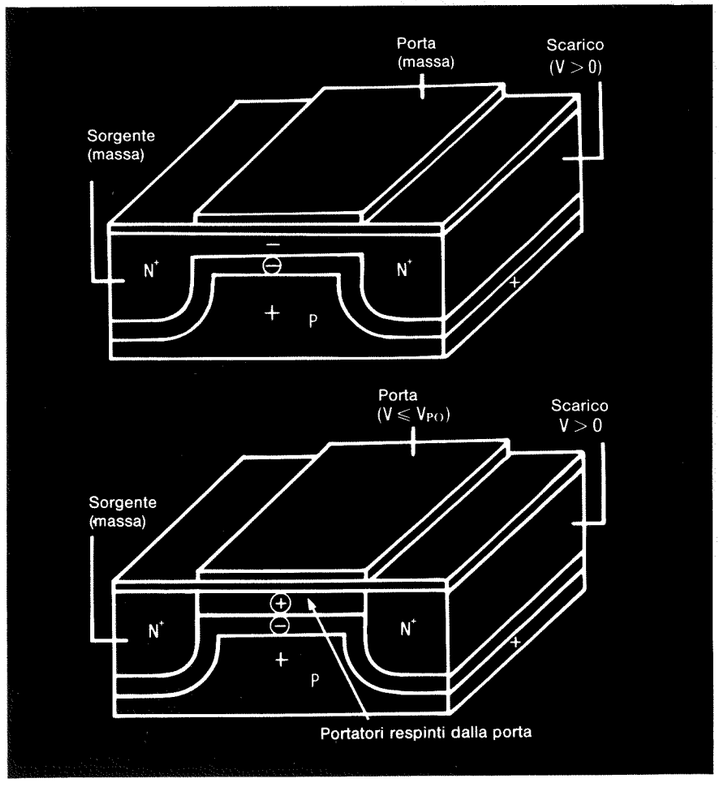


Fig. 7 - Un Mosfet a canale n e ad impoverimento conduce finchè una tensione negativa eguale o maggiore a quella necessaria per svuotare di elettroni il sottile canale di tipo n sotto la porta non è applicata all'elettrodo di porta.



to di corrente. Per questa ragione un dispositivo MOS è caratterizzato non tanto dal guadagno di corrente, bensì dalla sua transconduttanza che è definita come il rapporto tra le variazioni di corrente indotte e le variazioni di tensione alla porta che le inducono.

Un transistore MOS può essere ottenuto anche usando un sottile strato di silicio, sottostante la porta, dello stesso tipo della sorgente e dello scarico, strato che costituisce in effetti uno strato di inversione permanente. Questo transistore è normalmente in conduzione ed è detto a *impoverimento* (depletion) perchè in esso la corrente fluisce dalla sorgente allo scarico finchè la tensione applicata alla porta non è sufficiente per respingere i portatori di carica via dalla regione di canale ossia per impoverirlo (Fig. 7). Dispositivi a impoverimento sono frequentemente usati come carichi nei circuiti elettronici a stato solido in luogo di resistori a stato solido che sono difficili da realizzare. Per esempio i MOS a impoverimento sono normalmente usati nei MOS veloci per portare lo scarico, in assenza di segnale impresso alla porta, al livello della tensione di alimentazione, ossia come resistori di "pull up".

Vi sono MOSFET a canale p o canale n, così come vi sono transistori bipolari di tipo npn e pnp. I dispositivi bipolari pnp e i MOSFET a canale p furono i primi dispositivi solidi distribuiti sul mercato perchè erano meno sensibili a contaminazione degli altri dispositivi e perciò più facili da produrre. Tuttavia la maggior parte dei circuiti moderni è realizzata con transistori npn e MOSFET a canale n. Ciò è dovuto alla maggiore velocità di questi dispositivi perchè gli elettroni (portatori maggioritari) si spostano nel semiconduttore con una velocità da due o tre volte superiore a quella delle lacune, almeno in campi elettrici modesti. (La relazione tra gradiente elettrico e velocità non è rigorosamente lineare perchè in campi elettrici intensi i fenomeni di diffusione limitano la velocità che può essere raggiunta dagli elettroni).

In aggiunta ai MOS di tipo p ed n c'è un terzo tipo di transistore a effetto di campo. Si tratta del FET a giunzione o JFET (Fig. 8). Come si è accennato con riferimento ai diodi, l'ampiezza della zona di impoverimento ad una giunzione può essere variata modulando la tensione di polarizzazione inversa applicata alla giunzione. Il transistore JFET fa uso di questo meccanismo per controllare la corrente attraverso una regione delimitata da una o più giunzioni pn. La porta di controllo in un JFET è costituita da una regione fortemente drogata di polarità opposta a quella del canale e delle regioni di sorgente e scarico. Una corrente fluisce dalla sorgente allo scarico attraverso il canale se c'è una tensione applicata tra sorgente e scarico. Tuttavia se la porta di controllo è polarizzata inversamente, la zona di impoverimento formata alla giunzione pn sotto la porta, riduce la sezione

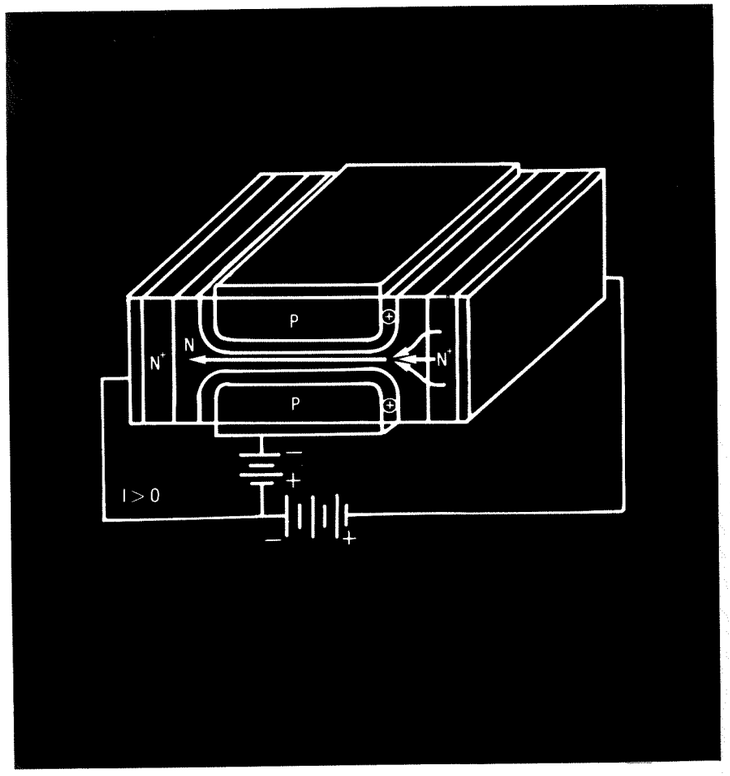


Fig. 8 - Un JFET o transistore a effetto di campo a giunzione. Il flusso di corrente è modulato per mezzo di due giunzioni. Quando le due giunzioni sono polarizzate inversamente la regione di impoverimento si allarga svuotando di conduttori il canale e in definitiva bloccando la corrente.

passante del canale attraverso cui fluisce la corrente. Quanto maggiore è la polarizzazione tanto minore è la corrente. È così possibile modulare la corrente con la modica spesa della corrente di controllo dovuta ai portatori minoritari che fluiscono attraverso la giunzione polarizzata inversamente.

6. La storia del MOSFET

La storia del transistore MOS non è la storia di una evoluzione continua da transistori MOS rudimentali fino ai dispositivi moderni, ma la storia di ricerche su altri dispositivi che più o meno casualmente diedero la soluzione a problemi che bloccavano la realizzazione di transistori MOS effettivamente operanti. L'esposizione che segue è organizzata cronologicamente, anzichè logicamente; è perciò opportuno iniziare evidenziando i tre problemi tecnici fondamentali che dovettero essere risolti prima che un transistore MOS potesse essere realizzato.

Il primo è il problema di produrre germanio e silicio sufficientemente puri per evitare la ricombinazione dei portatori di carica o il loro intrappolamento in difetti o impurezze del materiale.

Il secondo problema è quello di produrre giunzioni pn con una ragionevole tensione disruptiva (a tensioni elevate una corrente elevata fluisce anche in una giunzione polarizzata inversamente perchè molti portatori di carica vengono generati per effetti speciali, come l'effetto tunnel e la ionizzazione a valanga) e che abbiano al tempo stesso una ragionevolmente bassa corrente inversa in condizioni di polarizzazione inversa a valori normali di tensione.

Il terzo problema è la produzione di un isolante "pulito" per isolare la porta del canale semiconduttore, problema che non poteva essere risolto senza qualche comprensione teorica dei meccanismi di carica superficiale presenti all'interfaccia ossido/semiconduttore. Inoltre, dobbiamo aggiungere che il punto di partenza era in un certo senso deviante in quanto i primi dispositivi erano rudimentali transistori JFET o MESFET (FET a Metallo Semiconduttore).

7. Il problema Lilienfeld

Il dispositivo MOS è concettualmente semplice, ma la produzione di un MOSFET richiede una buona competenza nella fisica dello stato solido. Non è sorprendente che la storia iniziale del MOS sia la storia di brevetti concessi per dispositivi fondamentalmente mitici, che risultavano operanti sulla carta ma non in laboratorio. L'effetto di campo (o il controllo della conduttanza laterale in un cristallo di silicio per mezzo di un campo elettrico applicato trasversalmente alla superficie del silicio) fu riconosciuto già nel 1925 da Julius Edgar Lilienfeld [1] a cui fu concesso un brevetto per qualcosa di rassomigliante a un FET a porta isolata nel 1933 e un secondo brevetto per qualcosa rassomigliante a un JFET a giunzione nel 1935. (Nello stesso anno fu concesso un brevetto inglese a Oskar Heil [2] per un amplificatore basato sull'effetto di campo sebbene ancora sia dubbio che questo amplificatore abbia mai funzionato).

C'è un notevole disaccordo sulla stima e considerazione da tributare a Lilienfeld. Il punto di vista più critico è che: "Lilienfeld presentò più domande di brevetto per un dispositivo complesso vagamente simile a un transistore. Era una struttura multistrato costituita da strati di metallo e di materiale semiconduttore. La questione se tale dispositivo avrebbe potuto funzionare è interessante e appare di tanto in tanto nella letteratura scientifica. Lilienfeld condusse una vita alquanto oscura e certamente non ebbe i mezzi e le risorse di laboratorio per provare a sviluppare le sue idee. Forse il corso della storia sarebbe stato diverso se egli avesse lavorato in un grosso laboratorio industriale, ma è più verosimile che le sue idee non fossero attuabili semplicemente per l'ignoranza della fisica dello stato solido a quei tempi. In alcuni settori è possibile inventare dispositivi ed attuarli senza eccessivi fondamenti teorici; nel campo degli amplificatori a stato solido ciò si è dimostrato del tutto impossibile". [3].

Vale la pena di guardare ad uno dei dispositivi inventati da Lilienfeld per illustrare il problema che si presenta agli storici della scienza. Il dispositivo brevettato nel 1935 (Fig. 9) consisteva in due strati di un *semiconduttore*, il solfuro di rame con interposto uno strato metallico di magnesio. Lilienfeld suggerì che il metallo potesse essere così sottile che i due strati semiconduttori potessero venire in contatto tra loro attraverso piccoli fori e che il metallo potesse essere trattato in modo da non risultare in contatto ohmico con il semiconduttore. In altre parole l'interfaccia metallo semiconduttore doveva operare come un *diodo polarizzato inversamente* in modo che la corrente non fluisse dal metallo nel semiconduttore. Pertanto se un collegamento elettrico veniva stabilito con ambedue gli strati di semiconduttore, una corrente passante attraverso gli strati avrebbe potuto essere interrotta applicando una tensione positiva allo strato metallico. La tensione applicata avrebbe ampliato la *regione di impoverimento* in corrispondenza delle due *giunzioni*, bloccando la corrente (tutti i termini in corsivo sono naturalmente anacronismi e si può dubitare che Lilienfeld abbia effettivamente compreso i mezzi di modulazione di corrente che proponeva).

Sfortunatamente per Lilienfeld il suo genio creativo era andato ben oltre la sofisticazione tecnologica del suo tempo. Non era possibile, per esempio, ottenere una distribuzione uniforme di aperture nel metallo e non era perciò possibile dimostrare che l'intensità della corrente che attraversava il metallo era direttamente dipendente dalla tensione di polarizzazione applicata. In conseguenza di ciò e di altre difficoltà pratiche, nonchè in relazione al rapido sviluppo dei tubi a vuoto, il lavoro di Lilienfeld non stimolò ulteriori ricerche.

8. Schockley e i Laboratori Bell

Il primo, vero dispositivo a stato

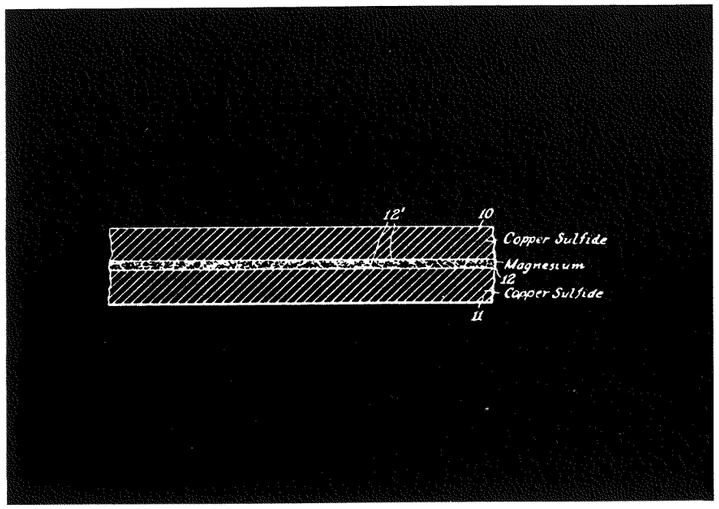


Fig. 9 - (*Dal Brevetto USA n. 1.900.018*). - Nei decenni 1920 e 30 furono concessi a Julius Edgard Lilienfeld dei brevetti per dispositivi che sarebbero stati dei transistori a effetto di campo se avessero funzionato. Un dispositivo brevettato nel 1935 consisteva di due strati di semiconduttore, ossidulo di rame, con interposto un sottile strato di magnesio. Lilienfeld suggerì che una tensione positiva applicata al metallo avrebbe potuto essere usata per bloccare un flusso di corrente attraverso gli strati di semiconduttore. Tuttavia è solo con i lavori di Schottky sulle giunzioni metallo/semiconduttore che si ottenne una comprensione teorica dei principi su cui si basava tale dispositivo.

solido fu il raddrizzatore a baffo di gatto usato nelle radio a cristallo e inventato nel 1874 da un certo Ferdinand Braun, professore di fisica a Marburg. Egli scoprì che il contatto tra un filo di metallo e il minerale noto come galena (solfuro di piombo) aveva proprietà raddrizzanti, ossia conduceva la corrente in un solo senso. Sebbene l'invenzione di Braun sia stata usata largamente essa non fu spiegata se non tardi nel ventesimo secolo. Negli anni '30 qualche attività di ricerca era svolta nei Laboratori Bell intesa a sviluppare un amplificatore a stato solido. Mervin Kelly, a quei tempi direttore della Ricerca ai Laboratori Bell, era d'avviso che i tubi a vuoto usati nelle centrali di commutazione telefonica erano troppo costosi e poco affidabili. Egli si convinse che un sostituto elettronico per il tubo a vuoto dovesse essere trovato e organizzò la ricerca dei Laboratori Bell verso questo scopo. Il suo più ardente discepolo era William P. Schockley.

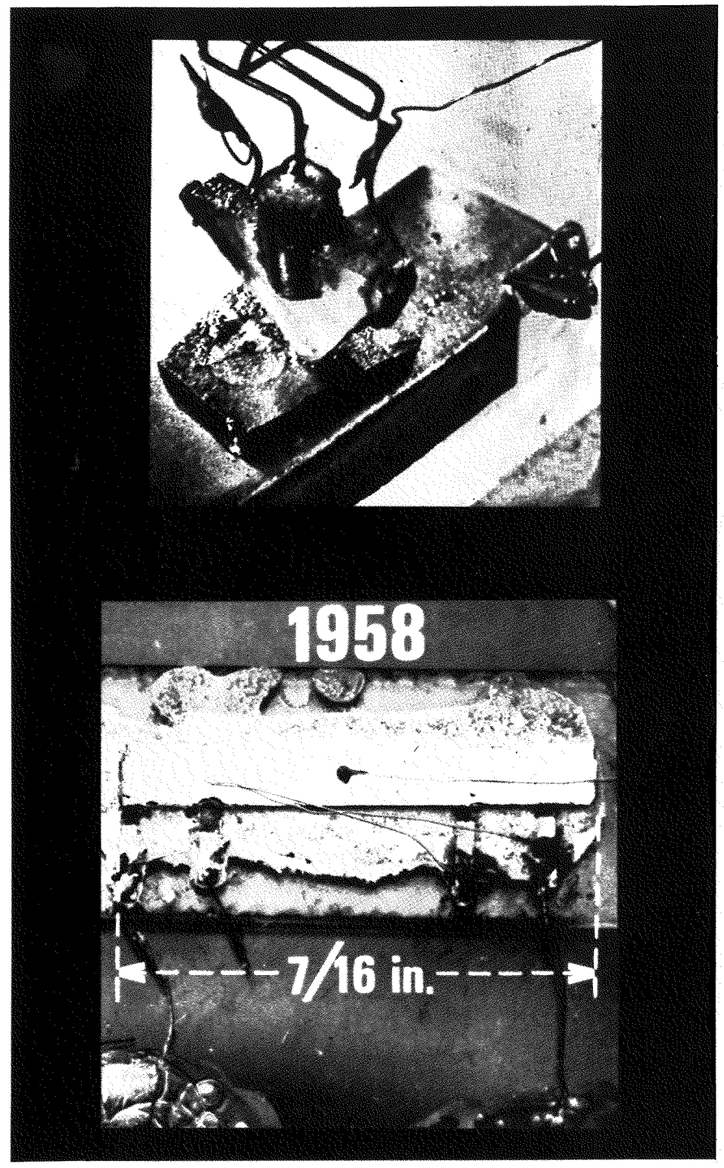
Schockley era interessato alla teoria che Walter Schottky aveva proposto per spiegare le proprietà rettificanti delle giunzioni metallo/semiconduttore. Schockley credeva che la carica spaziale o strato di impoverimento, che nell'interpretazione di Schottky si formava alla giunzione, avrebbe potuto essere usata per controllare la conduttività del semiconduttore a una certa distanza dal contatto, così come nel triodo si può controllare la corrente. Nel 1939 Schockley e Walter Brattain stavano sperimentando con raddrizzatori ad ossido di rame (prodotti in larghi volumi negli anni '30 a dispetto della più completa ignoranza del loro principio di funzionamento) cercando di inserire una sottile griglia di controllo nello strato di ossido sul rame, che potesse essere usata per controllare il flusso di corrente nel semiconduttore. Fondamentalmente essi cercavano di realizzare un transistore a effetto di campo. Tuttavia, tutti i loro tentativi furono un fallimento totale.

Allo stesso tempo altri scienziati dei Laboratori Bell stavano lavorando sui metodi di purificazione del silicio poichè c'era molto interesse sulla possibilità di usare il silicio per rilevare radio onde. "Due dei metallurgisti, J.H. Scaff e H.C. Theurer, avevano scoperto che mediante la fusione del silicio sotto vuoto, era possibile ottenere lingotti relativamente puri, sebbene qualcuno con proprietà rettificanti in un senso, qualcuno nell'altro e qualcuno non rettificante del tutto. Essi chiamarono di tipo n il materiale che conduceva meglio quando polarizzato negativamente, di tipo p quello che conduceva meglio nella direzione opposta. Infine Scaff e Theuerer scoprirono che ciò che distingueva il silicio di tipo n da quello di tipo p erano le tracce di impurezze ivi contenute. La concentrazione di impurezze che drogavano il silicio p o n era così bassa che a quel tempo sfuggiva persino all'esame spettroscopico. Secondo Brattain, fu l'odore dei lingotti di silicio che uscivano dal forno che indusse Scaff a sospettare la presenza di una contaminazione da fosforo". [4].

In aggiunta ci fu qualche lavoro avviato ai Laboratori Bell sulle giunzioni pn. Nel 1940 una fetta di silicio con un confine particolarmente netto tra regioni di tipo n e p faceva il suo turno di sperimentazione. Della luce proiettata sulla giunzione causò l'emissione di elettroni, come ci si attendeva, che furono poi attirati attraverso la giunzione dal campo creato dalla regione di impoverimento. Ma la luce produsse una forza elettro motrice alla giunzione dieci volte più grande di quella che ci si aspettava. Alcuni dei ricercatori e tra questi Walter Brattain furono così increduli che dapprima sospettarono qualche scherzo [5].

Durante la guerra, la maggior parte delle ricerche sui semiconduttori ai Laboratori Bell furono subordinate a ricerche nei semiconduttori orientate verso i radar per rivelazione sottomarina, organizzate dal Radiation Laboratory al MIT

Fig. 10 - Il transistore a punta di contatto (in alto), fu il primo transistore funzionante. Esso consisteva di due contatti elettrici, estremamente ravvicinati (emettitore e collettore) sulla superficie di una piastrina di silicio (la base) e fu sviluppato durante ricerche e ai Laboratori Bell, relativamente al problema degli stati di superficie che avevano frustrato gli sforzi dei primi ricercatori nel campo dello stato solido. Perchè il dispositivo funzionasse era necessario che i due contatti fossero ravvicinati.
Ingegnosamente, Walter Brattain, dei laboratori Bell, incollò una foglia d'oro sul bordo di un triangolo di polistirolo e tagliò accuratamente con un rasoio la foglia d'oro all'apice del triangolo, creando due contatti elettrici ravvicinati. Il triangolo di polistirolo era premuto sulla base da una molla, realizzata con un fermaglio per carte. È interessante notare che la base di un transistore di oggi, che per nessuna ragione può essere considerata una base, nei primi dispositivi costituiva di fatto una base di appoggio.
L'oscillatore (in basso), è uno dei primi se non il primo Circuito Integrato. Le larghe sbarrette sono sezioni di una fetta di semiconduttore, contenente transistori mesa e funziona come supporto per resistori, capacitori e transistori a stato solido.



(Istituto di Tecnologia del Massachusetts). Il lavoro più interessante con i semiconduttori fatto durante la guerra, fu quello svolto a Purdue da due studenti neolaureati che stavano studiando la resistenza di contatti a punta sul germanio. Questi due studenti furono a un pelo dall'inventare il transistore a stato solido. Dopo la guerra William Schockley e altri ricercatori dei Laboratori Bell visitarono Purdue per prendere conoscenza dei lavori ivi svolti e poi tornarono ai Laboratori Bell per riprendere il loro lavoro di sviluppo di un transistore a effetto di campo. Questa volta però essi lavoravano col germanio, le cui proprietà erano molto meglio comprese di quelle del rame e dell'ossido di rame e che inoltre si era dimostrato più facilmente purificabile del silicio. Ma questi sforzi non ebbero un successo migliore di quelli fatti prima della guerra.

John Bardeen e Walter Brattain ebbero allora il compito di verificare perchè i dispositivi costruiti in queste ricerche non funzionavano. La teoria di Schottky dell'effetto di campo assumeva che il numero di elettroni liberi in un semiconduttore fosse lo stesso alla superficie e all'interno. Bardeen suggerì che gli elettroni alla superficie fossero intrappolati e che perciò non fossero disponibili come portatori di carica quando un campo elettrico era applicato al semiconduttore. Bardeen e Brattain decisero di esplorare questi strati su-

perficiali. Uno dei dispositivi che essi realizzarono per questo scopo consisteva di due elettrodi metallici, assai vicini l'uno all'altro su un cristallo di germanio. Con loro sorpresa, quando cercarono di fare misure di corrente nel cristallo con uno degli elettrodi di contatto polarizzato direttamente e l'altro polarizzato inversamente, essi osservarono un leggero guadagno di corrente. Essi avevano inventato quello che noi conosciamo come transistore a punta di contatto (Fig. 10). Le sonde metalliche funzionarono come emettitore e collettore di tipo p di un transistore bipolare pnp.

Il loro successo fece sì che Schockley perse temporaneamente il suo interesse nel transistore a effetto di campo. Alla fine degli anni '40 egli aveva sviluppato una teoria del transistore a giunzione, anche se la pubblicazione di una relazione su tale teoria fu rifiutata dall'editore di *Physical Review* sostenendo che le argomentazioni di quanto-meccanica utilizzate in supporto non erano sufficientemente rigorose [6]. Tuttavia un *transistore a giunzione* funzionante fu realizzato solo nel 1952 e solo allora perchè le ricerche sui materiali condotte durante la guerra avevano notevolmente aumentato le capacità di controllo sul drogaggio dei semiconduttori. Negli anni 50 furono prodotti su scala industriale sia transistori a punta di contatto sia transistori a giunzione, questi ultimi con molto maggior successo. Tuttavia la resa di produzione, anche del transistore a giunzione era così bassa che si suggerì in modo più o meno faceto l'esistenza di un mitico elemento che "uccideva" i transistori il "deathnio" [7].

Nel 1948 Schockley ed un altro scienziato dei Laboratori Bell, Gerald L. Pearson [8], finalmente dimostrarono che un debole campo, applicato a una barra di semiconduttore di tipo n a basso drogaggio mediante elettrodi leggermente distanziati dal semiconduttore, poteva essere usato per modulare la conduttanza nella regione superficiale della barra. Pearson e Schockley miravano alla modulazione di portatori di carica maggioritari per effetto di un campo elettrico applicato. Sfortunatamente vi sono severe restrizioni geometriche per dispositivi basati su questo principio. Il rapporto tra l'area di superficie perpendicolare al campo elettrico di controllo e il volume del materiale di cui si desidera modulare la conduttanza deve essere molto alto se si vuole conseguire una modulazione apprezzabile [9]. Brattain era scettico sulla possibilità che i primi lavori di Schockley con i raddrizzatori a ossido di rame potessero produrre un transistore perchè si rese conto di questo problema. A quel tempo non esisteva alcuna tecnologia che consentisse la produzione di dispositivi con geometriche ridotte.

Per esempio i transistori a giunzione di Schockley erano transistori ottenuti per accrescimento. Mentre il cristallo di germanio era estratto dal crogiuolo caldo, una pastiglia di materiale drogante era aggiunta al bagno di fusione per modificare il cristallo in tipo p. Poi un'altra pastiglia era aggiunta per produrre uno strato sottile di germanio n e infine una terza per produrre un secondo strato di tipo p. Chiaramente non è possibile produrre dispositivi di dimensioni ridotte con questi mezzi ed anche ora è difficile realizzare dispositivi a effetto di campo a portatori maggioritari.

9. Il problema degli stati di superficie

Ironicamente lo sfondamento nella tecnologia dell'effetto di campo fu reso possibile da ricerche sulle prestazioni aberranti di alcuni dei primi transistori bipolari. Si scoprì che la conduttanza tra emettitore e collettore di alcuni dei transistori npn, quando la base non era connessa, era molto maggiore di quanto ci si potesse aspettare sulla base dell'impedenza di una giunzione inversamente polarizzata e di qualche dispersione attraverso la giunzione. Ipotesi tentative di interpretazione della causa di queste perdite variavano dal drogaggio accidentale della regione di base alla contaminazione della regione di base per effetto di una metallizzazione sovrastante. Nel 1953 W.L. Brown [10] stabilì che il comportamento quantitativo di questi transistori bipolari aberranti era consistente con l'ipotesi che la dispersione di corrente tra l'emettitore e il collettore fosse dovuta a un canale di ioni, (come lo stato di inversione era allora chiamato) indotto nella regione di base del transistore (Fig. 11). Inoltre lui ed altri avevano elaborato la teoria della distorsione o curvatura delle bande di conduzione e valenza in corrispondenza della giunzione e del modo in cui questa curvatura è influenzata da un campo elettrico.

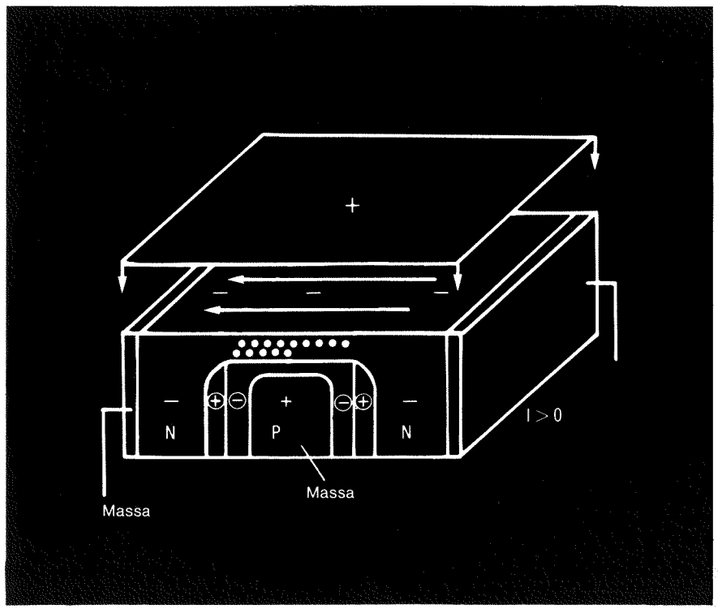
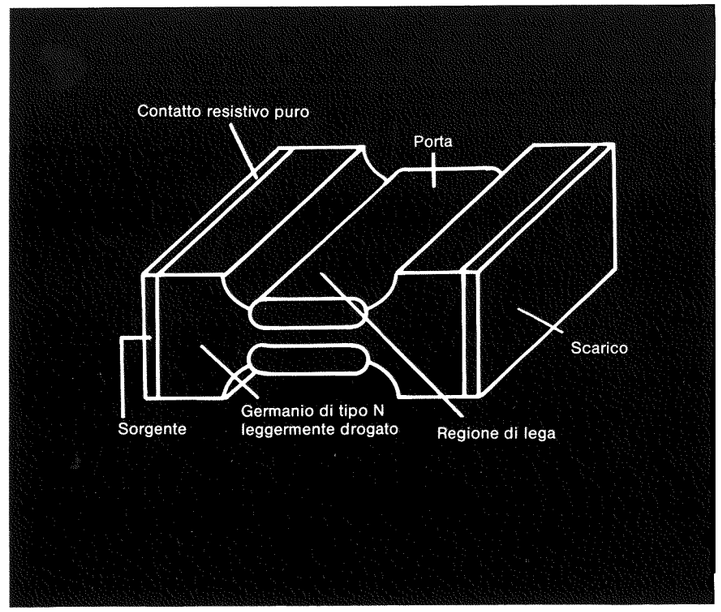


Fig. 11 - Ricerche sulle prestazioni aberranti di transistori bipolari servirono a risolvere i maggiori problemi che impedivano lo sviluppo dei MOSFET. Con la base a massa, nessuna corrente dovrebbe fluire tra emettitore e collettore di un transistore bipolare, ma molti dei primi dispositivi bipolari avevano invece delle correnti di perdita significative. W. L. Brown propose che queste perdite fossero dovute a una contaminazione accidentale della superficie del transistore con ioni positivi che in effetti creavano un canale permanente attraverso il quale poteva scorrere la corrente. Brown studiò anche la curvatura delle bande energetiche alla giunzione, determinata da tale contaminazione.

Fig. 12 - Il primo transistore a effetto di campo era un JFET piuttosto che un MOSFET. Esso fu costruito da Ian Ross e George Dacey dei Laboratori Bell. Essi risolsero i problemi dimensionali che ostacolavano lo sviluppo di FET a portatori minoritari mediante l'attacco chimico di una sbarra di germanio leggermente drogato e formando una lega di indio e germanio da ambedue le facce del canale sottile ottenuto per attacco chimico. Essi furono in grado di ottenere proporzioni del dispositivo tali che la corrente nel canale era influenzata in modo apprezzabile dalla tensione applicata all'indio.



Ian Ross dei Laboratori Bell si rese conto che se Brown aveva ragione sarebbe stato possibile sfruttare il problema della contaminazione ionica drogando deliberatamente la regione di canale di un dispositivo. Nel 1952 la General Electric realizzò un dispositivo che usava un campo elettrico per controllare la conduzione in una barra di germanio. Nel 1953 Ian Ross e George Dacey [11] spronati dal lavoro teorico di Schockley realizzarono un dispositivo simile (Fig. 12). Ambedue le estremità di una barretta di Germanio erano fortemente drogate e la zona intermedia era assottigliata mediante attacco chimico fino al punto che la corrente nel dispositivo risultava modulata da un campo elettrico relativamente debole. Poi dell'indio era sovrapposto al sottile canale e fuso in lega col Germanio formando in effetti una giunzione pn sul canale. Se la lega era polarizzata negativamente, lo strato di impoverimento pn nel canale si espandeva, riducendo la sezione del canale attraverso cui la corrente poteva fluire dall'emettitore al collettore.

Questo dispositivo era di fatto un rudimentale JFET piuttosto che un MOSFET. È in un derivato di questo dispositivo che la Philco fece grossi investimenti con gli infelici risultati già menzionati.

10. Il problema dell'isolamento della porta di controllo

Nel 1955 Ross [12] suggerì che la contaminazione ionica o zona di lega avrebbe potuto essere sostituita da un elettrodo intenzionalmente collocato in prossimità della regione di base di un transistore bipolare. Egli propose anche che "se lo spazio tra l'elettrodo di controllo e un isolante sulla regione di superficie della base fosse stato riempito con un materiale ferroelettrico, si sarebbe potuto stabilire un canale in maniera controllata e mantenerlo permanentemente finchè la porta di controllo non fosse stata eccitata per invertire la sua polarizzazione, così bloccando e rimuovendo il canale. Allora un misuratore di impedenza connesso tra emettitore o sorgente e collettore o drenaggio avrebbe indicato se tale canale era presente o meno ad ogni dato istante" [13]. Ciò che egli proponeva era ciò che oggi chiameremmo una cella di memoria elettrostatica a lettura non distruttiva (Fig. 13).

Sebbene tutte le memorie a semiconduttore con lettura non distruttiva attualmente disponibili siano memorie MOS, esse non sono discendenti lineari del dispositivo di Ross. Il suo materiale fer-

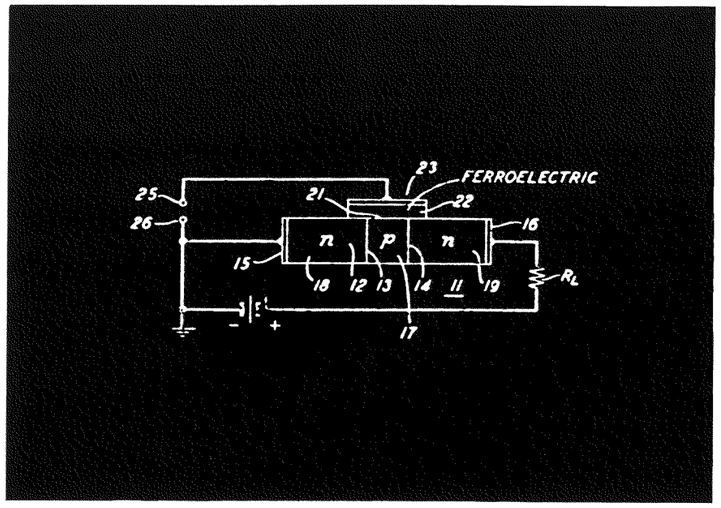


Fig. 13 - La porta dei moderni MOSFET fu inventata da Ian M. Ross che propose che l'effetto di campo indotto da contaminazione sulla superficie di un transistore bipolare, potesse essere riprodotto da un elettrodo separato dal semiconduttore per mezzo di un sottile strato isolante. Inoltre egli suggerì che se la regione tra il semiconduttore e la porta di un FET ad accumulazione di cariche fosse stata riempita con materiale ferroelettrico (che mantiene una polarizzazione elettrica) sarebbe stato possibile "leggere" il transistore, stabilendo se era stato creato o meno in precedenza un canale conduttore. In altre parole il dispositivo avrebbe funzionato come memoria MOS non volatile.

roelettrico doveva essere depositato sul semiconduttore mediante la tecnica di "sputtering" o spruzzamento catodico, che può causare danni al reticolo cristallino all'interfaccia semiconduttore isolante, provocando un funzionamento irregolare del transistore. L'alluminio, il metallo ora usato per realizzare gli elettrodi delle porte MOS può essere evaporato e perciò non causa gli stessi danni.

Ross aveva suggerito che per isolare l'elettrodo dal silicio nel suo FET si poteva usare uno spazio vuoto o un isolante liquido come nitrobenzolo o cianuro di etilene. Tuttavia questi due isolanti non sono stabili e determinano correnti di perdita nel dielettrico troppo elevate per essere in pratica usati come isolatori della porta di controllo.

Durante gli anni 1950 il silicio cominciò a soppiantare il germanio, come materiale preferito per transistori, perchè è stabile in un campo di temperature più ampio e perchè prometteva di poter essere controllato più facilmente. Dopo l'invenzione di Ross ci fu un'ampia ricerca industriale per l'individuazione di un isolatore compatibile con il silicio, che avesse una costante dielettrica elevata, una rigidità dielettrica pure elevata e potesse essere prodotto in forma relativamente pura. A posteriori può sembrare ovvio che il biossido di silicio avrebbe soddisfatto questi requisiti. Tuttavia a quel tempo nessuno sapeva come accrescere biossido di silicio mentre si sapeva che i legami di valenza non saturati all'interfaccia silicio/biossido di silicio erano estremamente efficaci nel catturare contaminanti ionici che degradavano le prestazioni del transistore. In aggiunta, non era possibile formare strati di ossido di spessore determinato perchè non si poteva controllare in maniera sufficientemente precisa la temperatura delle fornaci per la formazione dell'ossido.

Tuttavia tutto un corpo di conoscenze sulle proprietà della superficie del silicio e di quelle dell'interfaccia silicio/biossido di silicio si stava rapidamente sviluppando, insieme a tecniche di produzione di strutture stabili e riproducibili.

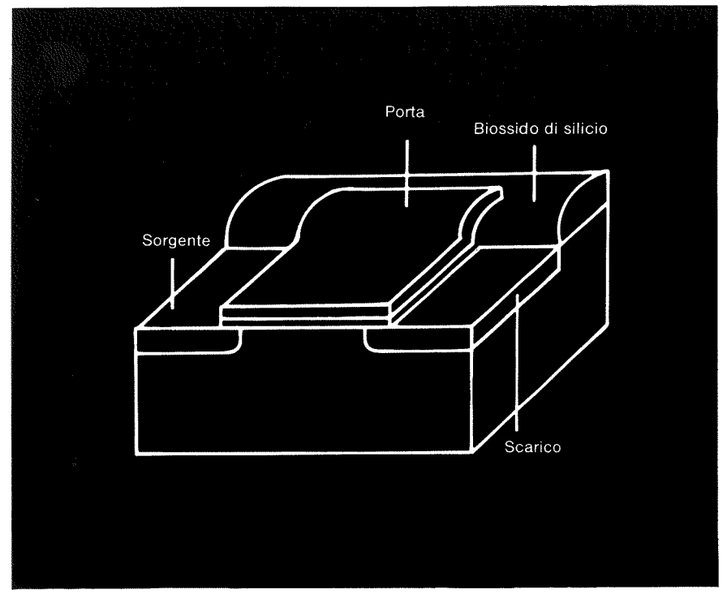
Finalmente nel 1959 Atalla [14] riferì che il silicio poteva essere passivato (ossidato) in fornace a temperatura moderata (da 925°C a 1050°C) e che l'ossido risultante soddisfaceva tutti i requisiti richiesti. (La costante dielettrica relativa del biossido di silicio è 4,5 e la rigidità dielettrica è di $8 \div 9 \times 10^6$ Volt/cm.).

11. Il primo MOSFET

Nel 1960 Dawon Kahng e John Atalla [15] del Laboratori Bell, proposero una struttura di silicio in cui una piastra isolata di campo, o porta, era usata per indurre un canale di superficie conduttivo tra due giunzioni superficiali pn. Essi brevettarono un dispositivo che può essere considerato il primo vero MOSFET (Fig. 14). Tuttavia la possibilità di realizzare industrialmente il dispositivo non fu dimostrata che due anni più tardi quando Stephen Berstein e Frederick P. Heimen [16] della RCA produssero un MOSFET funzionante con sorgente e scarico fortemente drogati e con SiO_2 come isolatore di porta. Questo fu il primo FET a portatori minoritari: la corrente fluiva quando veniva formato uno strato di inversione nella regione di canale.

Una delle ironie nella storia del MOS è che il primo FET funzionante non era il MOSFET di Berstein e Heimen ma un FET a film sottile descritto in un articolo pubblicato da P.K. Weiner [17] nel 1961. Weiner fu in grado di superare le limitazioni geometriche nei FET a portatori maggioritari, discusse in precedenza, sfruttando i recenti sviluppi della tecnologia a film sottili. Tuttavia il suo FET rimase principalmente una curiosità di laboratorio perchè le tecniche del silicio planare (l'introduzione di drogaggi in aree limitate del silicio attraverso una superficie del semiconduttore) erano molto meno costose delle tecniche a film sottile. In aggiunta i sottili film di

Fig. 14 - Il primo MOSFET fu realizzato nel 1960 da Dawon Kahng e John Atalla dei Laboratori Bell.



semiconduttore che venivano depositati erano policristallini e i transistori così fatti avevano prestazioni inferiori a quelle dei transistori bipolari ottenuti da una barra di silicio monocristallina.

12. La peste da Sodio

Sebbene la maggior parte delle conoscenze tecnologiche necessarie per la produzione del MOS fosse disponibile nei primi anni del '60, vi erano ancora molti problemi da risolvere. Per esempio la tensione di soglia della porta di molti tra i primi MOSFET variava nel tempo e con la temperatura. Ci vollero diversi anni per eliminare questo genere di fluttuazioni, determinate da contaminazione, particolarmente contaminazione dell'interfaccia ossido/silicio. La RCA e la Fairchild Camera ebbero qualche successo nel prevenire la contaminazione ma erano troppo impegnate dalla tecnologia bipolare per investire tempo e risorse sostanziali nei MOS. Questo consentì a due piccole Società, la General Microelectronics e la General Instruments, di cominciare a produrre circuiti MOS. Esse pure ebbero dei problemi. Finalmente negli ultimi anni '60 si scoprì che il sodio, (che ha un atomo abbastanza piccolo da risultare mobile nel reticolo del silicio e del biossido di silicio) era una delle maggiori cause di contaminazione di interfaccia. Poichè il sudore, la saliva, i capelli, la pelle e l'aria respirata contengono tutti quantità rilevanti di sodio, rigorosi standard di pulizia dovettero essere adottati nelle industrie. Ora tutte le case produttrici di MOS depurano l'aria e l'acqua degli ambienti di lavoro e impongono agli operatori di usare indumenti protettivi e di pulire periodicamente tutto l'equipaggiamento. Anche le sostanze chimiche e i gas che vengono usati nei processi produttivi sono controllati per rivelare la presenza di contaminanti, ciò che diede luogo a una industria chimica di grado elettronico.

Sebbene questi accorgimenti si siano dimostrati relativamente efficaci apparve subito evidente che non tutta la contaminazione da sodio poteva essere eliminata e che c'erano altri contaminanti altrettanto difficili da rimuovere. Nel 1964 William Miller [18] propose una tecnica che poteva essere usata per passivare i contaminanti che era impossibile escludere. Egli suggerì che uno strato di materiale di copertura che avesse dei legami aperti e che perciò potesse legare chimicamente i contaminanti, avrebbe potuto essere usato per tenere i contaminanti lontano dall'interfaccia silicio/ossido. Poichè i contaminanti che danno problemi sono quelli altamente mobili, molti di essi avrebbero incontrato questo strato durante il primo ciclo di riscaldamento successivo alla deposizione di questo strato (chiamato agente di cattura o "getter"). Oggigiorno vi sono comunemente in uso due versioni di questa tecnica. Una fa uso di passivazione con vetro fosfosilicato di copertura che attira e cattura contaminanti come il sodio. Questo processo può essere reso più efficace mediante l'applicazione di un debole campo elettrico di polarità ben definita che attira i contaminanti su una interfaccia di questo strato e non sull'altra. Il secondo tipo di passivazione, o cattura dal retro, è principalmente usato per neutralizzare impurezze metalliche residue lasciate nel silicio dopo il processo di raffinazione. Un metodo di cattura dal retro involve il danneggiamento intenzionale del reticolo cristallino sulla faccia posteriore della fetta di silicio, creando punti di legame per gli atomi metallici.

13. Ulteriori affinamenti

Nel 1963 Wanlass [19] introdusse una tecnica per produrre MO-

SFET di tipo p ed n sulla stessa pastiglia di silicio (chip). Egli suggerì che un transistore a canale di tipo p poteva essere formato su un substrato di tipo n in maniera convenzionale e che un canale n poteva poi essere formato sullo stesso chip creando nel substrato una regione di tipo p. Con questa soluzione si riduce la densità di impaccamento dei componenti attivi sul substrato, inconveniente che è però compensato dall'ottenimento di circuiti a bassissima dissipazione di potenza. Questa tecnologia, chiamata a MOS complementari o CMOS, fu la prima tecnologia a semiconduttori combinata (ossia il primo processo in cui due dispositivi diversi furono realizzati sul medesimo substrato).

La bassa dissipazione dei CMOS li rende attraenti per applicazioni militari e per strumentazione. La ragione per cui i CMOS sono circuiti a bassa potenza può essere spiegata confrontando un invertitore CMOS con un dispositivo MOS sia di tipo p che di tipo n (Fig. 15). Un invertitore CMOS consiste di due MOSFET, rispettivamente a canale p ed n, connessi in serie ad una sorgente di alimentazione con un nodo comune di ingresso e un nodo di uscita al punto intermedio di connessione tra i due FET. Una tensione positiva di porta, pone in conduzione il canale FET a canale n e blocca il FET a canale p, mentre una tensione negativa di porta pone in blocco il FET a canale n e in conduzione il FET a canale p. Pertanto non c'è passaggio continuo di corrente dalla sorgente di tensione verso massa, ma solo verso il carico, salvo che durante le operazioni di commutazione. Per confronto, un invertitore MOS di tipo n assorbe potenza quando la tensione di porta è positiva e un invertitore MOS di tipo p assorbe potenza quando la tensione di porta è negativa. Così un invertitore MOS a canale p o n assorbe più potenza di un CMOS.
Nel 1968 J.C. Sarace [20] suggerì che uno strato di silicio policristallino depositato sul dielettrico di porta avrebbe potuto essere usato come elettrodo di porta. La tecnologia "a porta di Silicio", come oggi è chiamata, presenta diversi van-

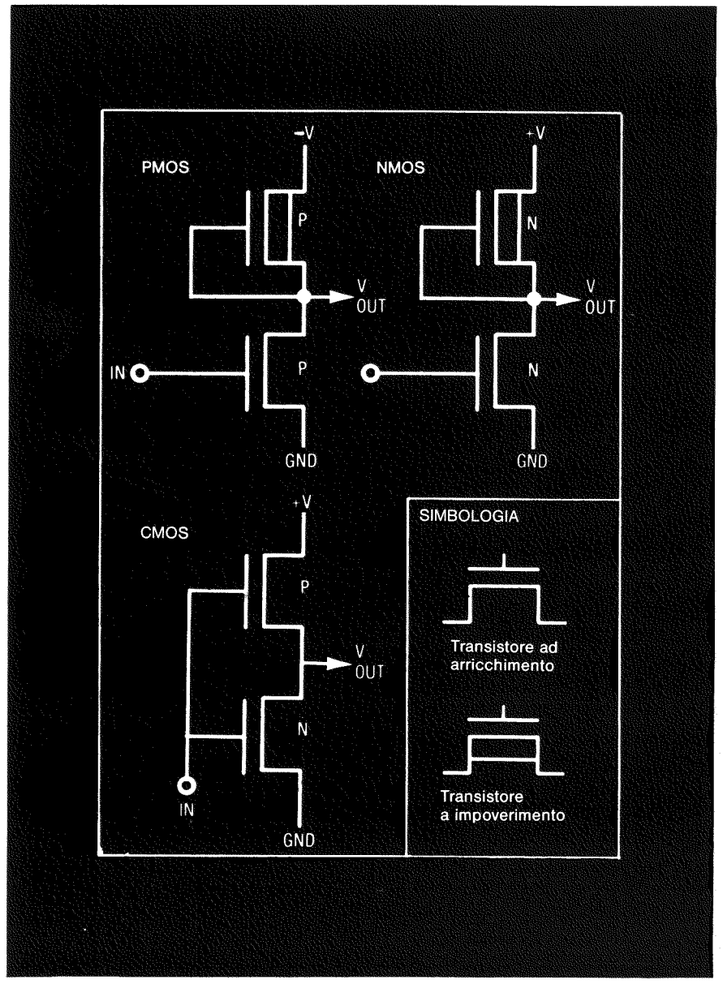


Fig. 15 - I vantaggi dei CMOS rispetto ai MOS a canale N o P, per quanto riguarda la potenza assorbita possono essere illustrati confrontando un invertitore realizzato con queste differenti tecnologie.
Nell'invertitore CMOS, le porte di ambedue i transistori sono connesse a un ingresso comune. Poichè i transistori richiedono segnali di polarità opposta per entrare in conduzione, i due transistori, in serie, non sono mai contemporaneamente in conduzione e una esigua corrente fluisce tra la sorgente di alimentazione e la massa. Invece sia nel caso dei MOS n, come in quelli di tipo p, ambedue i transistori sono in conduzione in determinate circostanze e perciò assorbono maggior potenza.

Fig. 16 - I MOS a porta policristallina presentano due vantaggi rispetto agli altri dispositivi MOS. La porta può essere usata come maschera per la diffusione delle regioni di sorgente e di scarico, con la conseguenza che le dimensioni del dispositivo possono essere ridotte e si minimizzano le capacità parassite derivanti da eventuali sovrapposizioni tra porta e regioni di sorgente/scarico.
Inoltre poichè il silicio policristallino può essere ossidato, si può formare su di questo un altro strato di interconnessioni aumentando ulteriormente la possibilità di realizzare più dispositivi interconnessi sul chip.

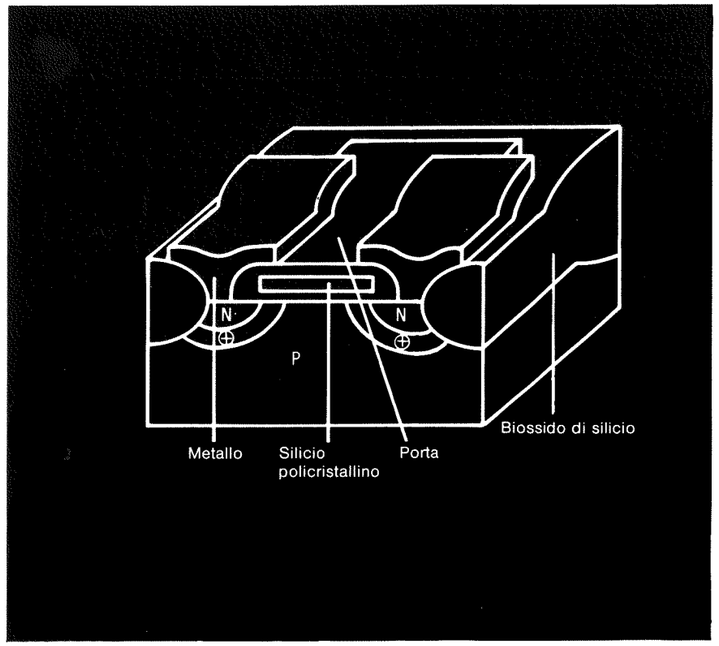
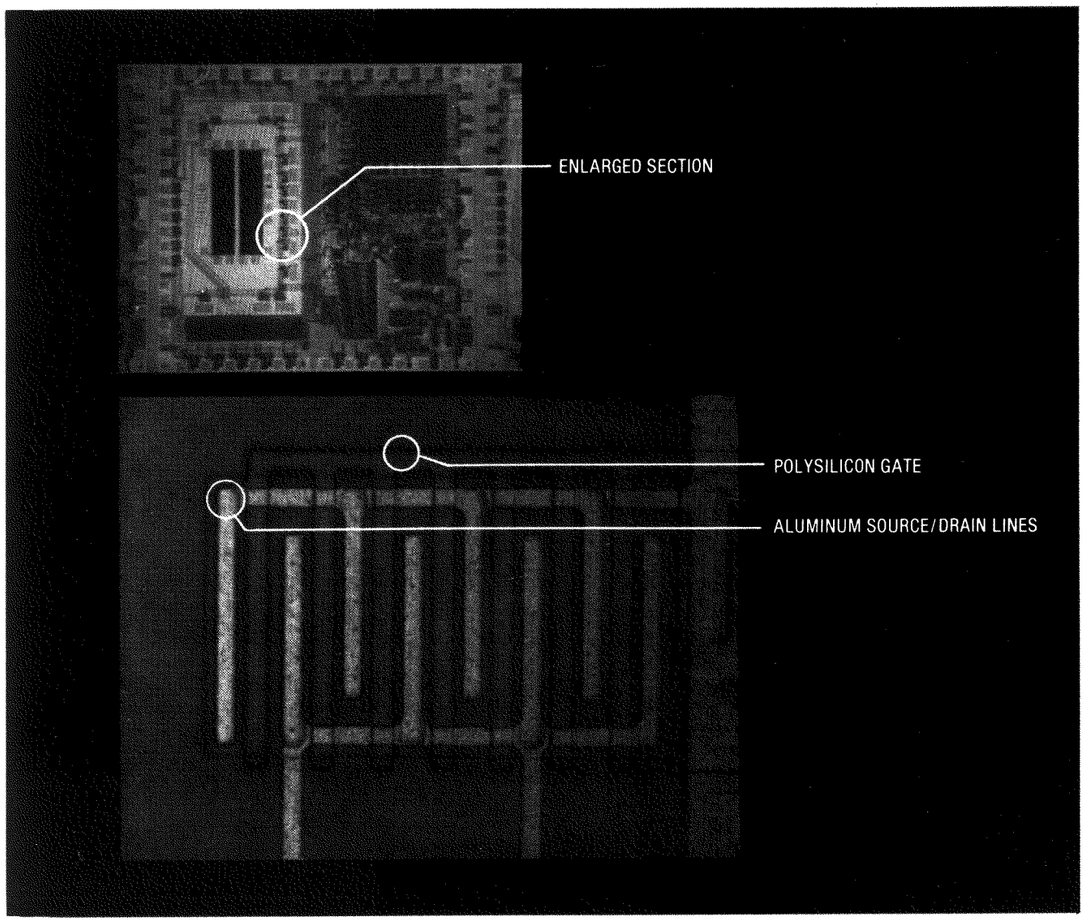


Fig. 17 - Il chip illustrato è un microprocessore NMOS disegnato al Solid-State Electronic Center della Honeywell Inc.
L'ingrandimento mostra un transistore di uscita a struttura interdigitata, così denominata per la configurazione interallacciata del silicio policristallino della porta e dei collegamenti di alluminio per la sorgente e lo scarico.



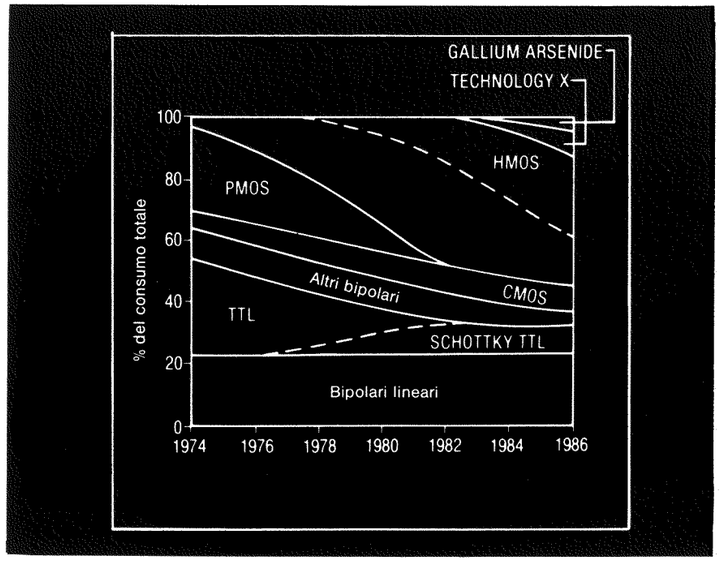


Fig. 18 - I MOSFET si stanno rapidamente sostituendo ad altri dispositivi sul mercato dei semiconduttori. Quanto più viene usata l'integrazione su larga scala di componenti in un chip, tanto più i MOS risultano avvantaggiati.
(per cortesia di: *Integrated Circuit Engineering*).

taggi sulla tecnologia a porta metallica. Per un verso la diffusione delle regioni di sorgente e scarico in un circuito a porta di silicio può essere fatta dopo che l'elettrodo di porta è stato formato. Ciò non è possibile con elettrodi di porta metallici per via dei problemi di contaminazione metallica causati da fasi di processo ad alta temperatura, posteriori alla deposizione del metallo. Il silicio policristallino, a differenza dei metalli, non diffonde nel substrato a temperature elevate (sopra i 500°C). Una porta di silicio policristallino può operare come maschera in un successivo processo di drogaggio per diffusione (il silicio policristallino arresta la diffusione dei droganti).

Ciò consente un preciso allineamento della porta con le regioni di sorgente e di scarico. Ciò significa pure che i dispositivi di questo tipo occupano meno spazio e possono essere impaccati più densamente. Anche le capacità parassite dovute a sovrapposizioni tra porta e le regioni di sorgente e di scarico vengono enormemente ridotte (Fig. 16). Inoltre il silicio policristallino può essere ossidato, e ciò comporta che è possibile formare un altro strato di interconnessioni sopra il silicio policristallino, aumentando ulteriormente la densità dei dispositivi nel chip.

Nel 1967 la tecnologia MOS aveva progredito al punto che si fece il tentativo di produrre un calcolatore da tavolo con transistori MOS. L'insuccesso di questo tentativo, troppo pubblicizzato, causò una certa apprensione tra gli utilizzatori dei MOS negli anni seguenti. Nel 1969 l'industria dei MOS ricevette una spinta sostanziale dai contratti di sviluppo firmati dalla società Viatron, un produttore di terminali e processori a basso costo. Sfortunatamente questa società fallì ma a quel tempo il potenziale del MOS per questo tipo di applicazioni era dimostrato. Con il 1970 le società AMI, Intel, Mosteck ed altre, erano ben avviate sul cammino di creare un mercato per memorie MOS ad accesso casuale e poi per interi microprocessori (Fig. 17) e i MOS cominciarono a togliere quote di mercato ai dispositivi bipolari. L'espansione del mercato MOS ha coperto nel 1980 il 55% del fabbisogno globale di dispositivi a stato solido ed una ulteriore espansione è ragionevolmente prevedibile nei prossimi anni (Fig. 18).

14. Bibliografia

[1] JULIUS E. LILIENFELD, U.S. Patent 1,745,175; U.S. Patent 1,877,140; U.S. Patent 1,900,018.

[2] OSKAR HEIL, British Patent 439,457.

[3] ERNEST BRAUN and STUART MAC DONALD, *Revolution in Miniature*, New York; Cambridge University Press, 1978, pp. 29-30

[4] "*The Solid State Era*", Electronics, Vol. 53, No. 9, p. 223, April 17, 1980.

[5] BRAUN and MAC DONALD, p. 43.

[6] Ibid., p. 50.

[7] Ibid., p. 76.

[8] W. SHOCKLEY and G.L. PEARSON, "*Modulation of Conduction of Thin Films of Semiconductors by Surface Charges*", Physical Review, Vol. 74, p. 232, 1948.

[9] DAWON KAHNG, "*An Historical Perspective on the Development of MOS Transistors and Related Devices*", IEEE Transactions on Electron Devices, Vol. ED-23, No. 7, p. 655, July 1976.

[10] W.L. BROWN, "*n-Type Surface Conductivity on p-Type Germanium*", Physical Review, Vol. 91, p. 518, 1953.

[11] Phone call with GEORGE DACEY.

[12], [13] I.M. ROSS, U.S. Patent 2,791,760.

[14] M.M. ATALLA, U.S. Patent 3,206,670.

[15] D. KAHNG and M.M. ATALLA, "*Silicon-Silicon Dioxide Field Induced Surface Devices*", presented at the IRE-AIEE Solid State Devices Research Conference, 1960.

[16] WILLIAM C. HITTINGER, "*Metal-Oxide-Semiconductor Technology*", Scientific American, Vol. 229, No. 2, P. 50, August 1973.

[17] P.K. WEINER, "*An Evaporated Thin Film Triode*", presented at the IRE-AIEE Devices Research Conference, 1961.

[18] W.H. MILLER and F. BASON, U.S. Patent 3,343,049.

[19] F.M. WANLASS and C.T. SAH, "*Nanowatt Logic Using Field Effect Metal Oxide Semiconductor Triodes*", ISSOC Digest, pp. 32-33, Feb. 1963.

[20] J.C. SARACE et al., "*Metal-Nitride-Oxide-Silicon Field Effect Transistors with Self-Aligned Gate*", Journal of Solid State Electronics, Vol. II, pp. 653-660, 1968.

Un modello previsionale per il mass-marketing: applicazione al settore EDP

FABRIZIO AGNESI,
RAFFAELLO PIERI,
FULVIA SALA

Honeywell Information Systems Italia
Servizio Ricerche e Sviluppo Marketing
Milano

1. Introduzione

Le conseguenze della rapidissima evoluzione tecnologica che caratterizza il mercato EDP sono ormai un argomento fin troppo dibattuto. Le implicazioni macro e microeconomiche, sociali, politiche, culturali, etc. di tale fenomeno hanno costituito l'oggetto di migliaia di articoli, tavole rotonde, interventi; si può dire che l'argomento sia "stantio" quasi quanto lo sono, sulle terze pagine dei giornali, le "analisi" sul ritorno al privato. Forse il marketing è l'unico punto di vista dal quale il fenomeno sia stato scarsamente studiato; mancano soprattutto metodologie *operative* che consentano di approcciare in modo nuovo una situazione nuova.
Questa lacuna non deve stupire: il marketing industriale è ancora, particolarmente in Italia, in una fase embrionale, sia per difficoltà obiettive (1), sia per la diffusa diffidenza e/o incomprensione del top management, soprattutto commerciale (2).
L'evoluzione tecnologica del settore EDP ha determinato, grazie ad una drastica riduzione di prezzo e ad una maggiore facilità d'uso del calcolatore, un allargamento del mercato potenziale-effettivo (e ciò è arcinoto); meno scontata, ma comunque diffusa è la convinzione che il mercato vada di conseguenza affrontato in termini nuovi, ossia in termini di mass-marketing. Bisogna riconoscere che tale parola, mass-marketing, stà ormai entrando nel vocabolario usuale dei quadri delle aziende che operano nel settore; esiste però una certa confusione in merito; i più tendono ad identificare-limitare il mass-marketing con un'intensificazione delle attività promozionali di "massa".

Scopo di questo articolo è quello, accennato molto concisamente al problema nei suoi termini generali, di illustrare un modello pratico per la determinazione delle probabilità di vendita di un prodotto a clienti che precedentemente non ne facevano uso; verranno in seguito descritte alcune possibilità di utilizzo di tali probabilità – oltre, come ovvio, che a scopi previsionali – per la razionalizzazione degli investimenti marketing e l'individuazione dei fattori che influenzano eventuali scarti di "produttività" fra filiali.
Il modello è stato concepito (e poi applicato) per il mercato dei nuovi utenti di elaboratori "general purpose" e di conseguenza a tale mercato si farà soprattutto riferimento; siamo però convinti che la metodologia generale sia applicabile ad altri beni strumentali a condizione che il numero di potenziali – effettivi clienti e la relativa complessità del prodotto giustifichino un approccio basato sul mass-marketing.
Ad esempio, nel campo dei grandi calcolatori la metodologia sviluppata non è applicabile perchè, oltre a trattarsi di un mercato di sostituzione, il mercato di clienti potenziali è ristretto (nell'ordine delle centinaia) e la complessità, non solo del prodotto, ma anche delle modalità di impiego e delle esigenze prospettate dagli utilizzatori è elevata.

2. Mass marketing e market segmentation

L'evoluzione di un mercato da ristretto a mercato di massa offre agli operatori del settore ampie possibilità (si pensi all'automobile ed al fordismo), ma comporta anche notevoli difficoltà. Innanzitutto bisogna capire che la situazione è mutata e poi avere la volontà – capacità di adeguarsi alle nuove condizioni (3).
È necessario far nascere una nuova mentalità aziendale, adeguare la struttura organizzativa (4), soprattutto adottare nuovi strumenti di marketing o modificare le modalità di utilizzo di quelli già esistenti.

L'importanza del funzionario commerciale, ad esempio, rimane, anche nella nuova situazione, immutata, cambiano però profonda-

mente i contenuti del ruolo; esso deve non più cercare il cliente, ma lavorare, con adeguati supporti centralizzati, affinchè il cliente cerchi lui. Non è possibile vendere con profitto un prodotto da 30-50 milioni con le stesse tecniche di un prodotto il cui prezzo è (od era) dieci volte maggiore.

L'offerta di software applicativo e genericamente di servizi deve essere incrementata e razionalizzata sia in termini di standardizzazione (per contenere i costi), sia in termini di differenziazione (per soddisfare esigenze diverse in un mercato non omogeneo).

L'utilizzo dei media, il direct mail, i seminari, le dimostrazioni, in generale tutte le attività promozionali devono essere sviluppate tenendo però presente che queste non costituiscono, come taluni pensano, l'essenza del mass-marketing, bensì uno, sia pure importante, dei suoi strumenti.

La gestione di tali strumenti non può essere demandata, se non in parte, alla struttura commerciale; essa richiede la costituzione, o comunque il potenziamento, di una funzione specifica centralizzata che oltre agli aspetti operativi segua anche e soprattutto le fasi di ricerca, pianificazione e controllo di tutte le attività di supporto commerciale o di marketing.

Normalmente, almeno nelle aziende di maggiori dimensioni, già esiste una Direzione a cui sono demandati istituzionalmente i compiti di promozione e supporto (e talora di pianificazione commerciale-finanziaria); viceversa è quasi sempre assente una funzione di analisi del mercato (Market Research and Development, nella terminologia anglosassone) che in un certo senso orienti o comunque fornisca tutte le informazioni quantitativo-qualitative necessarie alle altre funzioni di marketing per operare razionalmente.

L'analisi del mercato porta, quasi sempre, alla necessità di una segmentazione. Un mercato molto ristretto può essere considerato come un insieme di singoli clienti; un mercato di massa, viceversa, deve essere affrontato in termini di segmenti, ossia di gruppi omogenei di clienti.

Il primo punto da chiarire è cosa si intenda per omogeneità, definire cioè quali sono i criteri in base ai quali due clienti vengono considerati omogenei. Tali criteri discendono dagli obiettivi conoscitivi specifici che si vogliono conseguire dalla market segmentation e variano quindi in funzione della singola azienda e del mercato in cui opera.

Senza voler proporre una casistica generale – che soddisferebbe forse a esigenze classificatorie, ma avrebbe scarsa utilità pratica – ci limiteremo a citare, a mo' d'esempio, gli obiettivi della segmentazione del mercato dei nuovi utenti di calcolatori "general purpose": individuazione di nuovi prodotti e/o servizi, determinazione della potenzialità - redditività del mercato.

È ovvio che esisteva (a priori), la convinzione che il mercato non fosse omogeneo in termini di requirements, e di potenzialità, e che la conoscenza degli elementi che differenziano ciascun segmento avesse una utilità operativa; in particolare che tale conoscenza potesse servire, oltre che a scopi generali di pianificazione, a orientare le scelte di investimento (prodotti applicativi, promozione, risorse commerciali etc.) in e fra diversi segmenti.

Definiti, in termini generali, i criteri di omogeneità come portato di obiettivi conoscitivi finalizzati a determinate utilizzazioni (od obiettivi pragmatici), è necessario selezionare alcune variabili (e/o mutabili) che "misurino" tale omogeneità, ossia caratteristiche quantitative (e/o qualitative) proprie di tutti gli elementi (clienti potenziali o effettivi) del mercato oggetto di segmentazione.

La selezione di tali variabili può avvenire sulla base delle conoscenze che già si hanno del mercato o con indagini ad hoc, sulla base dell'esperienza e del "fiuto" o di strumenti statistici (6); in ogni caso debbono essere selezionate quelle variabili che "spiegano" l'omogeneità-disomogeneità fra gli elementi del mercato.

È di importanza fondamentale contenere il numero delle variabili considerate e dei loro valori discriminanti, ossia dei valori in base ai quali gli elementi del mercato vengono "assegnati" all'uno od all'altro dei segmenti, in quanto il numero di questi ultimi è dato dal prodotto del numero di valori discriminanti considerati.

Maggiore è il numero di segmenti in cui viene suddiviso il mercato e maggiori sono le difficoltà ed i costi di reperimento, elaborazione ed utilizzo delle informazioni; d'altro canto aggregare elementi disomogenei rispetto agli obiettivi della market segmentation porta a risultati privi di utilità (7). Si scontrano qui due esigenze contrastanti, che vanno in ogni caso mediate dato che la migliore segmentazione è quella che assicura contemporaneamente la massima omogeneità all'interno del segmento e la massima disomogeneità fra segmenti.

3. Un modello probabilistico delle potenzialità di mercato

La determinazione delle potenzialità dei diversi segmenti di mercato (e complessiva) rappresenta spesso una necessità, sia che ciò costituisca uno degli obiettivi della segmentazione, sia che i segmenti siano definiti in base ad altri criteri.

Senza conoscere le potenzialità dei segmenti è infatti impossibile prendere decisioni razionali circa l'opportunità e la distribuzione degli investimenti in prodotti, promozione, personale, etc., in quanto non si hanno elementi per valutarne la redditività.

L'utilizzo del concetto di probabilità di acquisizione di una nuova referenza (cliente) consente di esprimere in termini formalizzati e quantificati le potenzialità del mercato.

L'approccio probabilistico è ancora scarsamente diffuso (8), probabilmente a causa delle supposte difficoltà che esso presenterebbe; la metodologia qui proposta si caratterizza per l'estrema semplicità sia teorica sia di calcolo e utilizzo.

In sintesi, si tratta di stimare le probabilità applicando un modello (relazione funzionale) a delle frequenze (o probabilità empiriche) rilevate nel corso di ricerche di mercato o sulla base di statistiche di vendita.

Le frequenze sono il rapporto fra eventi favorevoli, ossia numero di nuove referenze acquisite in un certo periodo prefissato, ed eventi possibili, ossia numero di aziende che non utilizzano un calcolatore all'inizio del periodo e che potenzialmente potrebbero acquistarlo. Il modello è stato applicato sia alle sole nuove referenze HISI sia a quelle complessive, considerando nelle frequenze al denominatore solo le aziende che per dimensione avessero una probabilità non prossima allo zero di meccanizzarsi; sono cioè state escluse tutte le aziende con un numero di dipendenti inferiore ad una soglia minima stabilita sulla base di precedenti indagini.

Le variabili di segmentazione, in funzione delle quali sono poi state calcolate le probabilità, erano il settore economico ed il numero di dipendenti (espresso in classi); l'esperienza e precedenti indagini ci avevano infatti convinto che esiste una stretta relazione fra tali variabili, i requirements di prodotto (soprattutto applicativo) e le potenzialità di meccanizzazione.

In ogni caso il modello permette di verificare a posteriori se le variabili prese in considerazione sono significative o no, ossia se "spiegano" la meccanizzazione.

Una più rigorosa illustrazione del modello probabilistico renderà più formale le annotazioni precedenti e chiarirà l'esposizione del metodo (9).

Modello probabilistico

Le imprese del mercato possono essere classificate sia rispetto ai settori in cui operano sia rispetto alla dimensione (N° di addetti). Ipotizzando una certa influenza di questi fattori sulla probabilità di 1ª meccanizzazione si può studiare come varia tale probabilità al variare delle combinazioni di livelli dei fattori presi in esame.

Tab. 1

i Settori \ J Classe di addetti		1		2		3	
		B_1		B_2		B_3	
		K = 1	K = 2	K = 1	K = 2	K = 1	K = 2
1	A_1	N_{111}	N_{112}	N_{121}	N_{122}	N_{131}	N_{132}
2	A_2	N_{211}	N_{212}	N_{221}	N_{222}	N_{231}	N_{232}
3	A_3	N_{311}	N_{312}	N_{321}	N_{322}	N_{331}	N_{332}

Per esempio nella tabella a doppia entrata (vedi tab 1) sono riportati sia il n° di imprese meccanizzatesi per la prima volta durante l'anno sia quelle non ancora meccanizzatesi alla fine dell'anno.

i denota i livelli del fattore "Settore"
i = 1 Settore A_1;
i = 2 Settore A_2;
i = 3 Settore A_3.

J denota i livelli del fattore "Classe di addetti"
J = 1 Classe di addetti B_1;
J = 2 Classe di addetti B_2;
J = 3 Classe di addetti B_3.

K denota la prima meccanizzazione
K = 1 Prima meccanizzazione avvenuta nell'anno
K = 2 Prima meccanizzazione non ancora avvenuta alla fine dell'anno.

N_{iJ1} è il n° di imprese meccanizzatesi la prima volta nel corso dell'anno. i = 1,2,3 J = 1,2,3

N_{iJ2} è il n° di imprese non ancora meccanizzatesi alla fine dell'anno. i = 1,2,3 J = 1,2,3.

Con i dati della Tabella 1 si può costruire una matrice di probabilità di meccanizzazione (vedi tab. 2) durante l'anno al variare dei livelli sia del settore che della classe di addetti.

Tale probabilità empirica è intesa come rapporto tra eventi verificatesi ed esposti all'evento.

$P_{iJ1} = N_{iJ1} / (N_{iJ1} + N_{iJ2})$ ed ha il significato di "Probabilità che hanno le imprese che appartengono al settore A_i e alla classe di addetti B_j di meccanizzarsi per la prima volta durante un certo anno".

L'ultima colonna rappresenta come varia la probabilità al variare dei settori indipendentemente dalla classe di addetti.

$$P_{i01} = \sum_{J=1}^{3} N_{iJ1} / \left(\sum_{j=1}^{3} N_{iJ1} + \sum_{j=1}^{3} N_{iJ2}\right)$$

L'ultima riga rappresenta come varia la probabilità al variare delle classi di addetti indipendentemente dai settori.

$$P_{0J1} = \sum_{i=1}^{3} N_{iJ1} / \left(\sum_{i=1}^{3} N_{iJ1} + \sum_{i=1}^{3} N_{iJ2}\right)$$

Tab. 2

i Settori	J Classe di addetti	1	2	3	A
		B_1	B_2	B_3	
		K = 1	K = 1	K = 1	K = 1
1	A_1	P_{111}	P_{121}	P_{131}	P_{101}
2	A_2	P_{211}	P_{221}	P_{231}	P_{201}
3	A_3	P_{311}	P_{321}	P_{331}	P_{301}
	B	P_{011}	P_{021}	P_{031}	P_{001}

P_{001} rappresenta la probabilità di meccanizzarsi indipendentemente dal settore e dalla classe di appartenenza (media generale).

Queste probabilità sono legate al tempo per due aspetti.

Il primo si riferisce all'epoca dell'indagine (un certo anno di calendario), l'altro si riferisce alla durata in cui l'evento meccanizzazione si può manifestare (3 mesi, 6 mesi, 1 anno ecc.).

Per esempio nell'anno 1979 la probabilità di meccanizzazione annua è stata per il settore A_1 con classe di addetti B_2,$P_{1,2,1}$.

Le varie probabilità sin qui menzionate sono dette empiriche e quindi come tali affette da errori dovuti al caso.

Per esempio durante l'indagine possono esserci state delle condizioni casuali che hanno favorito o no alcuni settori o alcune classi di addetti.

L'assunzione di un modello probabilistico, che consideri i dati dell'indagine, almeno nelle loro relazioni, come campione temporale di un arco più vasto di tempo, permette, se verificato, una stima delle probabilità teoriche.

In altre parole la ricerca del legame mediante il modello probabilistico sfrutta globalmente l'informazione dei dati attenuando l'influenza del caso.

Si sono considerati quindi due modelli; uno additivo ed uno moltiplicativo nelle probabilità:

$$\text{Modello 1: } P_{iJ1} = \mu + \alpha_i + J\beta_J$$
$$\Sigma\alpha_i = \Sigma J\beta_J = 0$$

$$\text{Modello 2: } P_{iJ1} = \mu \cdot \alpha_i\beta_J^J$$
$$\pi\alpha_i = \pi\beta_J^J = 1$$

In questo modello abbiamo considerato la probabilità di meccanizzarsi condizionata alla appartenenza ad un dato settore e ad una data classe di addetti. La distribuzione in ogni combinazione di livelli si è considerata Binomiale:

$$\forall_{ij}\ P_{iJ1} + P_{iJ2} = 1$$

P_{iJ1} = Probabilità di meccanizzarsi
P_{iJ2} = Probabilità di non meccanizzarsi
per la combinazione dei livelli i.J
μ è la media aritmetica semplice dei P_{iJ1}
α_i è l'effetto del livello i del fattore settore
β_J è l'effetto del livello del fattore classe di addetti.

Presentiamo solamente il modello 1 dato che il modello 2 si ottiene mediante la trasformazione logaritmica dei dati.

In notazione matriciale il modello 1 può essere scritto:

$$P_1 = X^T\underline{\beta} \text{ dove}$$

$$X^T\underline{\beta} = \begin{bmatrix} 1&1&0&0&1&0&0 \\ 1&1&0&0&0&2&0 \\ 1&1&0&0&0&0&3 \\ 1&0&1&0&1&0&0 \\ 1&0&1&0&0&2&0 \\ 1&0&1&0&0&0&3 \\ 1&0&0&1&1&0&0 \\ 1&0&0&1&0&2&0 \\ 1&0&0&1&0&0&3 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

La stima $\underline{\hat{\beta}}$ di $\underline{\beta}$ mediante i minimi quadrati è data da

$$\underline{\hat{\beta}} = (XX' + HH')\ X'\ \underline{P}_1 \text{ dove}$$

$\underline{P}_1$ è il vettore delle probabilità stimate dalle osservazioni e

$$H' = \begin{bmatrix} 0&1&1&1&0&0&0 \\ 0&0&0&0&1&2&3 \end{bmatrix}$$

è la matrice corrispondente ai vincoli

otteniamo così le stime di

$$\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$$

che permettono la stima dei $\hat{P}_{iJ1}$ (valori teorici)

$$\hat{P}_{iJ1} = (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_J)$$

Verifica del modello

– Calcolo del N° teorico di imprese meccanizzate.

Mediante la matrice delle probabilità teoriche scaturite dal modello si calcola il N° teorico di imprese meccanizzatesi nell'anno distinte per settore e classe di addetti

$$\hat{N}_{iJ1} = (N_{iJ1} + N_{iJ2}) \cdot \hat{P}_{iJ1}$$
$$i = 1,2,3$$
$$J = 1,2,3$$

Test del chi quadro

La verifica della validità del modello è ottenuta dal confronto dei dati empirici con quelli teorici – mediante il test del CHI QUADRO (X^2)

dove il X^2 calcolato risulta essere:

$$X^2 = \sum_{ij} \frac{(N_{iJ1} - \hat{P}_{iJ1} \cdot N_{iJ0})^2}{(\hat{P}_{iJ1})\ (1 - \hat{P}_{iJ1})\ N_{iJ0}}$$
$$N_{iJ0} = N_{iJ1} + N_{iJ2}$$

e quello teorico ($\hat{\chi}_2$) viene dedotto dalle tavole con 4 Gradi di Libertà* al livello di fiducia ($1 - \alpha$)** prefissato per sostenere in termini statistici la significatività o meno del modello scelto.

Il Modello è respinto se:

$$\chi^2 > \hat{\chi}^2_{1-\alpha,4}$$

Viene accettato o meglio non vi sono motivi per respingerlo se:

$$\chi^2 < \hat{\chi}^2_{1-\alpha,4}$$

**) I Gradi di Libertà (G.L.) rappresentano il parametro delle distribuzioni CHI QUADRO e dipendono dal N° dei livelli dei fattori.*

***) 90%, 95%, 99%.*

Risulta evidente che la scelta tra il modello additivo (1) e quello moltiplicativo (2) viene basata sul confronto dei due X^2 calcolati, preferendo il modello a cui compete il X^2 calcolato più basso.

Esempio Numerico

Un esempio numerico chiarirà sia l'applicazione del modello che alcuni utilizzi.

Nelle tabelle 3 e 4 (input del modello) sono riportate rispettivamente le "Nuove Referenze 80" ed il n° di imprese non ancora meccanizzate distinte per Settore (A_1, A_2, A_3, A_4, A_5, A_6) e per Classe di addetti (B_1, B_2, B_3, B_4).

Tab. 3 Nuove Referenze

Settore \ Classe addetti	B_1	B_2	B_3	B_4	TOT.
A_1	15	30	20	10	75
A_2	13	35	30	10	88
A_3	18	55	42	21	136
A_4	4	15	20	10	49
A_5	22	60	52	30	164
A_6	10	32	25	10	77
TOTALE	82	227	189	91	589

Tab. 4 Mercato Aperto

Settore \ Classe addetti	B_1	B_2	B_3	B_4	TOT.
A_1	13000	1800	700	150	15650
A_2	2500	500	300	50	3350
A_3	1500	400	200	30	2130
A_4	600	200	140	20	960
A_5	4000	700	400	100	5200
A_6	6000	1000	500	100	7600
TOT.	27600	4600	2240	450	34890

Nel tabulato A sono riportati oltre agli input già visti nelle tab. 3 e 4 i seguenti output:

1) Matrice delle probabilità empiriche
   es. $P_{A_1,B_3} = 20/700 = 0.02857$
2) I parametri del modello moltiplicativo (il modello moltiplicativo in questo caso interpola meglio i dati) che risultano essere rispettivamente:

$\mu = 0.0466573$ — Valore base

Effetto differenziato dei Settori:
$\alpha_1 = 0.2965105$
$\alpha_2 = 1.1133036$
$\alpha_3 = 2.6747371$
$\alpha_4 = 1.6568808$
$\alpha_5 = 1.4034628$
$\alpha_6 = 0.4870493$

Effetto differenziato della classe di addetti:
$\beta_1 = 0.0867607$
$\beta_2 = 1.1017504$
$\beta_3 = 1.248094$
$\beta_4 = 1.4865934$

dove risulta il peso che hanno i livelli dei due fattori (>1 influenze positive; <1 influenze negative)

3) Le probabilità teoriche di penetrazione

$$\text{es. } P_{A_1,B_3} = \quad = 0.0466573 \cdot 0.2965105 \cdot 1.248094^3 = 0.0269$$

4) Le nuove referenze teoriche ($\hat{NR}$)

$$\text{es. } \hat{NR}_{A_1,B_3} = \hat{P}_{A_1,B_3} \cdot N_{A_1,B_3} = 0.0269 \cdot 700 = 18.83$$

5) Il valore del CHI QUADRO empirico

$$\text{es. } X^2 = \frac{(15{,}60 - 15)^2 + \ldots\ldots +}{0.0012 \cdot 0.09988 \cdot 13000 + + (11{,}10 - 10)^2} + \ldots\ldots + 0.11098 \cdot 0.88902 \cdot 100 = 8.84$$

con 15 Gradi di Libertà $(6-1)\cdot(4-1)$.

Il valore del $\hat{X}^2$ teorico al 95% di fiducia risulta essere 25 (dedotto dalle tavole) e quindi il Modello moltiplicativo può essere accettato.

Tabulato A

Nuove referenze

15.	30.	20.	10.
13.	35.	30.	10.
18.	55.	42.	21.
4.	15.	20.	10.
22.	60.	52.	30.
10.	32.	25.	10.

Mercato

13000.	1800.	700.	150.
2500.	500.	300.	50.
1500.	400.	200.	30.
600.	200.	140.	20.
4000.	700.	400.	100.
6000.	1000.	500.	100.

1

Tassi di penetrazione

0.00115	0.01667	0.02857	0.06667
0.00520	0.07000	0.10000	0.20000
0.01200	0.13750	0.21000	0.70000
0.00667	0.07500	0.14286	0.50000
0.00550	0.08571	0.13000	0.30000
0.00167	0.03200	0.05000	0.10000

Modello moltiplicativo:
P (I,J) = MU * ALFA (I) * BETA (J)^J

2

Parametri espressi in ln

-3.0649267	-1.2156726	0.1073318
0.9838511	0.5049368	0.3389426
-0.7193899	-2.4446019	0.0969002
0.2216176	0.3964872	

3

Tassi teorici di penetrazione

0.00120	0.01679	0.02690	0.06757
0.00451	0.06305	0.10099	0.25369
0.01083	0.15148	0.24263	0.60949
0.00671	0.09384	0.15030	0.37755
0.00568	0.07909	0.12731	0.31981
0.00197	0.02758	0.04418	0.11098

4

Nuove referenze teoriche

15.60	30.23	18.83	10.13
11.27	31.53	30.30	12.68
16.24	60.59	48.53	18.28
4.02	18.77	21.04	7.55
22.72	55.64	50.92	31.98
11.83	27.58	22.09	11.10

5

CHI-QUADRATO

= 8.841330

Nella tabella 5 viene riportato il confronto tra $\hat{NR}$ teoriche e quelle effettuate in 5 zone geografiche.

Le Nuove Referenze teoriche attribuite ad ogni zona sono state ottenute applicando le probabilità teoriche di penetrazione alla struttura del mercato di ogni singola zona.

Le $\hat{NR}$ così ottenute rappresentano anche le potenzialità di mercato di ogni zona e tale potenzialità potrà essere resa relativa ponendo la zona di massima potenzialità uguale a 100.

Risulta chiaro che una più precisa valutazione della potenzialità di mercato di ogni zona dovrà tener conto anche di altri fattori, come la concorrenza, le software houses etc.

Le relazioni ed il peso di fattori non considerati potranno essere valutate poi dai dati effettivi che incorporano già l'influenza di tali fattori oltre a quella delle strutture del mercato già analizzate.

Tab. 5

Zona	Nuove referenze			Potenzialità
	$\hat{NR}$	NR	Δ%	
1	100	90	+11.1	34.6
2	200	260	−23.1	100
3	89	95	− 6.3	36.5
4	150	105	+42.8	40.3
5	50	39	+28.2	15

4. Applicazioni delle probabilità di meccanizzazione

Le possibilità di utilizzo delle probabilità di meccanizzazione così calcolate sono numerose; la prima e più ovvia riguarda le attività di pianificazione (budget vendite, long range plan, etc.). Per scopi previsivi ciò che importa maggiormente sono i valori assoluti delle probabilità; se è possibile che tali valori, riferiti al mercato nel suo complesso, non varino in misura notevole un anno con l'altro (e le variazioni possono essere comunque "dominate" se si dispone di serie storiche che pongano in luce trend e/o cicli) è anche plausibile che le probabilità del singolo venditore, soprattutto se la sua quota di mercato è bassa, subiscano oscillazioni rilevanti in presenza di eventuali modificazioni della posizione competitiva. Ciò significa che le probabilità, così come ogni estrapolazione di dati storici, debbono essere utilizzate con molta cautela a fini previsionali ed integrate da altri strumenti.

I valori relativi delle probabilità, che dovrebbero essere più stabili, possono essere utilizzati per comparare i risultati del singolo fornitore con quelli della concorrenza, per verificare cioè, calcolando magari indici di dissomiglianza, se esistono differenze significative fra le distribuzioni normalizzate delle probabilità relative ai diversi segmenti; l'individuazione di segmenti in cui le probabilità relative del singolo fornitore sono superiori o inferiori a quelle della concorrenza, e soprattutto un'attenta analisi delle cause, può consigliare l'adozione di provvedimenti che migliorino la situazione competitiva dell'azienda.

Un'altra possibilità di utilizzo delle probabilità, come già accennato, riguarda l'elaborazione di criteri che orientino le decisioni di investimento commerciale.

A questo fine vanno considerate:

A) due misure delle potenzialità del segmento; la probabilità di meccanizzazione ed il numero teorico di nuove referenze che, come si è visto, è il prodotto della probabilità per il numero di imprese non ancora meccanizzatesi.

B) due tipi di investimento: "a costi variabili", dove il costo è proporzionale al numero di aziende del segmento per il quale l'investimento è sostenuto, "a costi fissi", dove il costo è indipendente dal numero di aziende.

Un generico direct mail è un esempio del primo tipo di investimento, in quanto il costo è determinato dal numero di aziende a cui si decide di spedire il materiale promozionale; lo sviluppo di un pacchetto applicativo per uno specifico settore è un esempio del secondo tipo di investimento, in quanto il costo è indipendente dal numero di aziende potenziali utilizzatrici.

Nel selezionare i segmenti in cui effettuare investimenti "a costi variabili" si dovranno privilegiare quelli che presentano una maggiore probabilità di meccanizzazione, viceversa per investimenti "a costo fisso" si dovranno privilegiare i segmenti che hanno il maggior numero teorico di nuove referenze.

I motivi sono ovvi; è più "produttivo" inviare del materiale promozionale ad un'industria, poniamo, farmaceutica, il cui segmento ha una probabilità di meccanizzazione 0.05, piuttosto che ad una industria del settore meccanico, il cui segmento ha una probabilità di meccanizzazione 0.01, d'altro canto sarà più produttivo sviluppare un pacchetto per quest'ultimo segmento se il suo numero teorico di nuove referenze è 100 e quello dell'industria farmaceutica è 40.

Esistono però investimenti i cui costi sono in parte "fissi" ed in parte "variabili"; ad esempio, l'invio di una brochure di settore, tipo "la HISI e l'industria dei gelati" presenta dei "costi variabili" (spese di spedizione e parte del costo di stampa) e dei "costi fissi" legati all'ideazione, stesura e stampa della pubblicazione.

In questi casi, peraltro frequenti, si potrà ricorrere ad una semplice formula che permette di determinare quello che abbiamo chiamato il "costo per potenziale cliente" dell'investimento in esame:

K_i = costo per potenziale cliente del segmento i − mo $= \dfrac{F + VN_i}{p_i N_i}$

ove F = costo fisso
V = costo variabile (unitario)
N_i = numerosità del segmento i − mo
p_i = probabilità di meccanizzazione del segmento i − mo

Si supponga che si voglia scegliere se inviare una brochure di settore al segmento S_{44} (settore 4, classe di addetti 4, vedi esempio precedente) o S_{52}; sia F = 2000 e V = 2, si avrà che

$$K_{44} = \frac{2000 + 2 \cdot 20}{0.377 \cdot 20} \simeq 270$$

$$K_{52} = \frac{2000 + 2 \cdot 700}{0.079 \cdot 700} \simeq 61$$

Si sceglierà ovviamente il settore S_{52} perchè il "costo per potenziale cliente" è nettamente inferiore; se fosse V = 50 sarebbe $K_{44} \simeq 397$ e $K_{52} \simeq 665$ e la scelta cadrebbe sul settore S_{44}.

F e V di uno stesso tipo di investimento potrebbero variare leggermente da segmento a segmento o addirittura riferirsi ad alternative fra investimenti diversi; in questo caso nella formula basta attribuire un indice i − mo anche ad F e V, ma la significatività del risultato può essere profondamente inficiata da limiti che sono d'altro canto presenti anche nel caso più semplice in cui F e V sono uguali per tutti i segmenti.

Perchè la formula possa essere applicata si deve infatti ipotizzare che "l'efficacia" dell'investimento sia circa eguale in tutti i segmenti, o meglio, che essa sia determinata esclusivamente da probabilità e numero teorico di N.R. e l'influenza di ogni altro fattore non vari al variare del segmento. Ma se, ad esempio, l'industria farmaceutica è subissata da direct mailing sicchè la probabilità di "cestinazione" è 10 volte maggiore rispetto a quella dell'industria meccanica, converrà spedire materiale promozionale a quest'ultima anche se la probabilità di meccanizzazione è 5 volte inferiore.

Tuttavia in molti casi pratici si può assumere che l'"efficacia" sia analoga oppure calcolare empiricamente un coefficiente di ponderazione che attribuisca maggiore peso ai segmenti per i quali l'"efficacia" è ritenuta maggiore.

Ad esempio, nel decidere lo sviluppo di un pacchetto si possono pesare i denominatori $p_i N_i$ moltiplicandoli per la frequenza di ricorso a software houses rilevata negli anni precedenti, sicchè se il 20% delle NR. del segmento A e l'80% di B ricorre a consulenti per lo sviluppo delle applicazioni, B avrà un peso proporzionalmente maggiore di A ($p_a N_a \cdot 0.2$, $p_b N_b \cdot 0{,}8$).

Conviene sottolineare che il metodo proposto può costituire una risposta al quesito "avendo deciso di investire una certa somma in una data iniziativa, quale distribuzione fra segmenti è la più conveniente?"; nulla ci dice però circa il quanto investire ed in quali iniziative. Poche aziende − e forse nel settore degli elaboratori, nessuna − sono in grado di stimare se il loro budget promozionale è un terzo, il doppio, oppure è prossimo a quello "ottimale"; poche aziende sanno valutare in termini quantitativi se sia più conveniente investire nello sviluppo di un nuovo prodotto o in risorse commerciali aggiuntive.

Normalmente le imprese determinano il budget dei costi secondo criteri che, volendo usare un eufemismo, potremmo definire "empirici": più X% rispetto al budget dell'anno precedente, oppure Y% del fatturato, e così via.

Gli investimenti "discreti" − ossia gli investimenti sostenuti una tantum, ad esempio per lo sviluppo di un nuovo prodotto − sono spesso oggetto di più attente e formalizzate valutazioni; gli investimenti "continui" − ossia quelli il cui costo può variare con una certa libertà, ad esempio investimenti in promozione od in risorse commerciali, sono viceversa stabiliti in modo del tutto arbitrario.

Per questa seconda categoria di investimenti la risposta teorica alla domanda "quanto spendere?" è semplicissima: incrementare la spesa sino a che il costo marginale eguagli il profitto marginale (11).

Ma quale è il profitto marginale, ossia come calcolare quale siano gli incrementi delle vendite determinati (o determinabili) da uno specifico investimento?

Misurare l'elasticità della domanda rispetto al prezzo, alle spese promozionali, al numero di funzionari commerciali, etc., non è sicuramente un obiettivo facile.

Ma, quale che siano i criteri utilizzati (12), un budget di investimenti marketing, articolato secondo un ceto mix di iniziative, viene in ogni caso deciso all'interno di tutte le aziende.

A questo punto, e solo a questo punto, la metodologia proposta può essere utile per orientare la ripartizione del budget fra i diversi segmenti di mercato.

Un'ulteriore applicazione delle probabilità di acquisizione di nuove referenze riguarda l'analisi dei risultati conseguiti dalle singole filiali.

Moltiplicando le probabilità di meccanizzazione dei singoli segmenti per il numero di aziende di tali segmenti presenti nell'area geografica di competenza della filiale, si ottiene il numero teorico di nuove referenze che la filiale avrebbe dovuto conseguire se si fosse "comportata in media", se cioè si fosse allineata a quelli che sono i risultati complessivi nazionali.

Numero teorico e numero effettivo di nuove referenze possono grosso modo coincidere oppure, come nel caso da noi esaminato, presentare sensibili scostamenti. In quest'ultima evenienza si formeranno, di norma, tre gruppi di filiali, quello in media, quelli sopra e sotto la media. È importante sottolineare che questa classificazione non comporta giudizi sulla efficienza delle filiali; i risultati conseguiti dalle filiali vengono rapportati a teoriche potenzialità di mercato, definite in funzione di due o più variabili (nel nostro caso, il settore e la classe di addetti) che compaiono nel modello probabilistico, ma possono esistere numerose altre variabili che "spiegano" gli scostamenti. Ed è proprio l'individuazione di queste ultime variabili, specie se endogene rispetto all'azienda, che può consentire di migliorare i risultati complessivi.

Se le filiali sono composte da un numero non eccessivamente ristretto di persone, si può presumere che le differenti capacità dei singoli individui non influenzino i risultati: le differenze positive e negative grosso modo si compenseranno.

La prima variabile da prendere in considerazione è il numero di funzionari commerciali (ed eventualmente tecnici, personale di supporto, etc.) presenti in ogni filiale.

Se considerando non più il numero assoluto di nuove referenze teoriche ed effettive, ma il numero per funzionario commerciale, le differenze scompaiono, vorrà dire che queste dipendono da una squilibrata ripartizione delle risorse. Squilibrata si, ma forse non irrazionale perchè le filiali "sovradimensionate" conseguono, in termini relativi, gli stessi risultati di quelle "sottodimensionate", si può presumere quindi che una redistribuzione delle risorse non porterebbe a benefici apprezzabili, e che, risorse aggiuntive incrementerebbero il volume delle vendite senza diminuire il rapporto vendite per funzionario commerciale (se il profitto per funzionario fosse positivo converrebbe, salvo considerazioni di altra natura, incrementare il numero delle risorse commerciali in quanto il profitto crescerebbe proporzionalmente).

Nel caso il numero di venditori non spieghi le differenze, o lo spieghi solo in parte, dovranno essere prese in considerazione altre variabili, quelle che la conoscenza che si ha del settore fa ritenere essere influenti e per le quali si hanno a disposizione informazioni quantitative attendibili.

Nel vagliare l'influenza di tali variabili converrà utilizzare tecniche statistiche, quali l'analisi della varianza, la regressione o l'analisi fattoriale, che consentono di misurare l'affidabilità delle conclusioni a cui si perviene.

Per quanto riguarda l'applicazione di questa metodologia al caso delle nuove referenze HISI, sono state prese in considerazione numerose variabili quali la dimensione del preesistente parco clienti, la promozione, la concorrenza,il tempo, l'anzianità professionale dei funzionari commerciali, etc. ed essendo l'analisi tutt'ora in corso, non possiamo anticiparne i risultati.

Note

(1) È un dato di fatto che metodologie ed esperienze sono più sviluppate nell'area del consumer marketing, dove il marketing è nato, rispetto all'area dell'industrial marketing.
Webster individua quattro ordini di difficoltà nell'applicazione di tecniche sviluppate per il mercato dei beni di largo consumo a quello dei beni strumentali: a) maggiore rigidità di manovra delle variabili di marketing (si pensi ad esempio ai tempi necessari per sviluppare - modificare un prodotto), b) complessità del prodotto, c) base di clienti potenziali - effettivi più ristretta, d) difficoltà di misurazione delle variabili di marketing. F.E. WEBSTER Jr., "*Management Science in Industrial Marketing*", The Journal of Marketing, vol. 42 (Gennaio 1978), pp. 22-25.
Gli stessi problemi giustificano la relativa arretratezza del marketing industriale; in particolare i punti b) e c), quanto a d) è una conseguenza dei due punti precedenti ed a) riteniamo che influisca solo parzialmente.

(2) Il top management, soprattutto se commerciale o di estrazione commerciale, è abituato a pensare ed agire in termini di selling più che di marketing; ciò porta, fra l'altro, a considerare i problemi di mercato dal punto di vista del venditore più che del compratore.
T. LEVITT, "*Marketing Myopia*", Harward Business Review, sett.-ott. 1975.

(3) Immaginare una opportunità di mercato ossia uno stato alternativo al presente richiede sempre un notevole sforzo di immaginazione; la forza d'inerzia è una grandezza psicologica oltre che fisica. P. KOTLER, "*Marketing Management*", Prentice Hall, 2ª ediz., 1972, p. 59.

(4) M. HAMAN, "*Reorganize your Company around its Markets*", Harward Business Review, nov.-dic. 1974, pp. 63-65.
La HISI ha, ad esempio, riorganizzato la propria struttura di marketing ponendo l'accento non solo e non tanto sulla linea di prodotto, quanto sul macro-settore economico delle aziende clienti.

(5) Per una trattazione di carattere generale, si veda: H. ASSAEL, A.M. ROSCOE Jr., "*Approaches to Market Segmentation Analysis*", The Journal of Marketing, vol. 40 (Ottobre 1976), pp. 67-76.

(6) Cluster analysis e analisi discriminante. M.G. KENDALL, The basic problems of cluster, in T. Cacoullos (ed.), "*Discriminant analysis and applications*", Academic Press, New York 1973. M.G. KENDALL, Discrimination and classification, in P.R. Krishnaia (ed.), "*Multivariate analysis*", Academic Press, New York 1966.

(7) Un interessante approccio per la determinazione di appropriati livelli di aggregazione nei piani di market segmentation è proposto da: F.W. WINTER, "*A Cost - Benefit Approach to Market Segmentation*", The Journal of Marketing, vol. 43 (Fall 1979), pp. 103-111.
In caso di obiettivi plurimi di market segmentation, la definizione delle variabili e quindi dei segmenti deve tener conto di tale pluralità; così, nell'esempio del mercato dei nuovi utenti di elaboratori, se consideressimo solo l'obiettivo "potenzialità del mercato" potremmo aggregare aziende di settori diversissimi fra loro, se viceversa è presente anche l'obiettivo "requirements di mercato" dovremo distinguere, poniamo, i grossisti farmaceutici dalle industrie meccaniche pur se presentano le stesse potenzialità.

(8) Fra gli altri, si possono citare due esempi di utilizzo del concetto di probabilità, l'uno che parte da un modello della decisione d'acquisto (industrial buying behavior) e che quindi segmenta il mercato in funzione di una tipologia di formazione delle decisioni di acquisto delle aziende, l'altro che determina le probabilità a partire dalle intenzioni di acquisto dichiarate nel corso di interviste.
J.M. CHOFFRAY, G.L. LILIEN, "*Assessing Response to Industrial Marketing Strategy*", The Journal of Marketing, vol. 42 (April 1978), pp. 20-31. D.G. MORRISON, "*Purchase intentions and Purchase Behavior*", The Journal of Marketing, vol. 43 (Spring 1979), pp. 65-74.

(9) Per una visione d'assieme di problematiche statistiche, si consiglia: E. PARZEN, "*La moderna teoria delle probabilità e le sue applicazioni*", Franco Angeli, Milano 1978.
J.L. HODGES Jr., E.L. LEHMAN, "*I concetti fondamentali della probabilità e della statistica*", 2 vol., Il Mulino, Bologna 1971.
Più attinente allo specifico modello proposto è: W. MENDENHALL, "*Introduction to linear models and the design and analysis of experiments*", Wadsworth, New York 1968.

(10) Tali informazioni possono essere desunte sia dai dati del censimento (ormai obsoleti) sia da archivi nominativi acquistabili presso società specializzate.

(11) Nel caso si considerino più segmenti, le spese debbono essere ripartite in modo che i profitti marginali siano eguali per tutti i segmenti.
N.K. DHALLA, W.H. MAHATOO, "*Expanding the Scope of Segmentation Research*", The Journal of Marketing, vol. 40 (April 1976), pp. 34-41.

(12) Alcuni criteri non "empirici" sono stati elaborati ed applicati anche nel campo dei beni strumentali; si veda, ad esempio: G.L. LILIEN, J.D.C. LITTLE, "*The Advisor Project: a Study of Industrial Marketing Budgets*", Sloan Management Review, vol. 17 (Spring 1976), pp. 17-31. C.A. BESWICK, D.W. CRAVENS, "*A Multistage Decision Model for Salesforce Management*", Journal of Marketing Research, vol. 14 (May 1977), pp. 135-144.

Collana dei Quaderni di informatica

Novità

Giulio Occhini
L'informatica nella gestione aziendale
Aspetti e prospettive di impiego

I ritmi con cui ha in questi anni progredito l'informatica hanno creato un divario tra le sue reali potenzialità applicative e la consapevolezza, da parte di molti dirigenti, di tali potenzialità, nonchè un ritardo nell'acquisizione da parte delle aziende dei criteri secondo cui avvalersi dell'informatica stessa nell'ambito organizzativo-gestionale. Ne è prova la persistente difficoltà a trasformare gli obiettivi dell'impresa in precisi requisiti del suo sistema informativo automatizzato, e quindi a sviluppare programmi d'investimento in questa area veramente finalizzati e controllati anche dal punto di vista economico.
Questo volume si rivolge a quanti desiderano documentarsi in modo approfondito in questo campo: non solo, quindi, agli specialisti del settore – che vi troveranno discussi temi economici e organizzativi generalmente trascurati nella letteratura loro familiare – ma anche, e soprattutto, a chi nelle aziende ha il compito d'indirizzare e controllare le modalità d'impiego della risorsa informatica e a chi tale risorsa intende utilizzare come supporto allo svolgimento della propria attività.
Partendo da questi presupposti, l'opera offre per la prima volta un quadro organico e completo dei problemi connessi ai diversi aspetti organizzativi e gestionali relativi all'informatica, iniziando dalla creazione di basi e banche di dati e proseguendo con le possibilità d'impiego dei terminali e problemi di creazione di reti di elaborazione, per affrontare quindi i complessi rapporti tra sistema informativo e organizzazione e definire poi i criteri di scelta delle aree di automazione, di controllo della redditività informatica, di analisi del sistema informativo aziendale, ecc. Un particolare spazio è dedicato ai problemi della sicurezza del sistema informativo, alle prospettive per i prossimi anni, all'impiego dei «personal computers». L'accento è posto soprattutto, di volta in volta, sugli aspetti concettuali più che su quelli tecnologici, pur non trascurando di dare tutte le informazioni di base necessarie al «non specialista» e dedicando particolare attenzione a evidenziare i risvolti pratico-operativi. Ne risulta, quindi, un'opera equilibrata, indispensabile sia a quanti già occupano posizioni di responsabilità sia a quanti vogliono fare carriera nel mondo aziendale: l'informatica, infatti, appare destinata ad occupare un posto fondamentale nel bagagliaio culturale di chiunque operi nel mondo economico, perdendo in misura sempre maggiore la sua connotazione di disciplina riservata agli addetti ai lavori.
G. Occhini è dirigente responsabile della ricerca applicativa alla Honeywell Information Systems Italia. Laureato in fisica all'Univ. di Milano, opera da oltre 20 anni nel settore dell'informatica. Ha partecipato a numerose realizzazioni e ha al suo attivo diversi contributi sugli aspetti economico-organizzativi del processo di automazione. È docente di sistemi informativi alla Scuola di Direzione Aziendale dell'Univ. Bocconi.

Sommario

Basi e banche di dati: Introduzione - Realtà e rappresentazione informatica - Record, campi e flussi - Entità e attributi - Indipendenza logica e fisica - Schema concettuale - Schema esterno - Schema interno - Ridondanza e integrazione - Strutture di dati - Integrazione - Vantaggi della centralizzazione - Aspetti di progettazione - Aspetti operativi ed organizzativi - Linguaggio di definizione dati - Linguaggio di trattamento dati - Linguaggio di definizione menoria - Il dizionario dei dati - Amministrazione della base - Struttura della base - Diritti di accesso - Affidabilità - Priorità e situazioni di stallo - L'opinione degli utenti - Indipendenza dei dati - Integrità dei dati - Interattività - Ridondanza dei dati - Riservatezza e sicurezza - Inefficienza delle procedure - Costi addizionali - Riqualificazione del personale - Linguaggi utente - Introduzione alle banche di dati - Banche di dati e azienda - Esempi - Banche americane - Banche italiane - Aspetti operativi.

Dal terminale alle reti di elaborazione: Introduzione - Struttura di un sistema di elaborazione e comunicazione - Gestione della rete - Modalità di collegamento - Apparecchiature terminali - Telescrivente-stampante - Video-terminale - Terminali grafici - Terminali a lettura ottica/magnetica - Apparecchiature di preparazione dati - Terminali per applicazioni speciali - Terminali a risposta fonica - Reti di elaborazione - Reti chiuse e aperte - La rete Transpac - Protocolli e modalità di trasmissione - Commutazione di pacchetto e di circuito - Prestazioni di targa - Criteri tariffari - La rete Euronet - Servizi di informatica e di comunicazione - Il televisore interattivo - "Telematica" e servizio postale - Stati Uniti - Europa - Controlli doganali per il trasferimento dei dati - Misure legislative per la protezione dei dati personali.

Sistema informativo e organizzazione: Introduzione - Obiettivi di un sistema informativo - Struttura organizzativa - Struttura a livelli - Struttura gerarchica - Tipologia delle esigenze informative - Informazioni interne ed esterne - Livello operativo - Livello tattico - Livello strategico - Logica del processo decisionale - Fasi del processo - Decisioni programmabili e non - Decisioni e automazione - Struttura del sistema informativo - Articolazione del sistema - Rapporto specialista-utente - Livelli e modalità di automazione - Le ragioni di un insuccesso - Aspetti di gestione e di controllo - Automazione amministrativa - Automazione intersettoriale - Livelli di gestione - Gestione dei costi - Meccanismi di allocazione dei costi - Valore aggiunto di elaborazione - Società di servizi - Dimensione economica - La situazione oggi - Le prospettive.

Criteri di scelta delle aree di automazione: Introduzione - I termini del problema - Una metodologia generale - La analisi di bilancio - Gli indicatori di gestione - Conclusioni.

Redditività informatica: l'analisi costo/benefici: Introduzione - Il ciclo di vita del progetto - L'analisi costo/benefici - Flusso di cassa - Classificazione dei progetti - Uno schema di riferimento - Tecniche finanziarie - La valutazione dei costi - I prodotti applicativi - Costi di sviluppo e gestione - La valutazione dei benefici - Aree di riduzione costi - Miglioramenti gestionali - Esempi di valutazione costo/benefici - L'informatica nel settore amministrativo - L'informatica nella gestione - L'informatica nella azienda - Conclusioni.

Metodologia di analisi del sistema informativo aziendale: Introduzione - La risorsa informatica dell'azienda - Ruolo dell'analista di sistema informativo - Struttura della metodologia - Analisi di base - Analisi funzionale - Carta delle attività - Interfaccia totale - Interfaccia funzionale - Disegno del prototipo - Specifiche del sistema - Conclusioni.

La sicurezza del sistema informativo: Introduzione - Aspetti generali della sicurezza fisica e logica - Rischio fisico - Rischio logico - Crimine informatico - Funzione di auditing - Strategia di intervento - La responsabilità della sicurezza - Ruoli e sfere di azione - Cosa va protetto - Raccolta dati e analisi del rischio - Valutazione dei rischi - Valutazione dei danni - - Criteri alternativi per l'analisi del rischio - Misure di protezione - Protezione fisica - Protezione logica - Danni accidentali e danni intenzionali - Strategia combinata - Un esempio di applicazione - Identificazione dell'utente - Impronte digitali - Geometria della mano - Analisi della voce - Firma - Costo della sicurezza - Criteri di allocazione - Conclusioni.

Problemi e prospettive del settore informatico: Introduzione - Indicatori di settore - Il processo di innovazione tecnologica - Unità centrale e memorie - Il microprocessore - Le periferiche - Conseguenze sugli utenti - Il problema del software - Domanda e offerta - La produttività - Tipologia dei programmi - La manutenzione - Interventi di tipo organizzativo - Pianificazione dei progetti - Motivazione dei gruppi - Documentazione - Formazione e ruoli - Tecniche e metodologie - Modalità di sviluppo dei sistemi di software - Linguaggi di comando e orientati - Ipotesi di evoluzione - La crisi del concetto di economia di scala - L'entropia del software - Il costo della ge-

stione - Interventi correttivi - Affidabilità del sistema - Aspetto sistemistico - Livelli di multiplazione - Gestione della rete e della base di dati - Aspetto logistico - Costi di elaborazione e di comunicazione - Sistemi distribuiti - Aspetti sistemistici - Architetture di sistemi distribuiti - Aspetti organizzativi - Un modello di distribuzione - Aspetti economici.

L'automazione dell'ufficio: Introduzione - La struttura del sistema - Commutazione elettronica privata - Trattamento dei documenti - Il sistema integrato - Il lavoro nell'ufficio di domani - Aspetti economici - Considerazioni generali - Lavoro segretariale ed esecutivo - Lavoro direttivo e professionale - Aspetti organizzativi e sociali - Stato applicativo e tendenze di ricerca - Conclusioni.

L'elaborazione personale e le sue prospettive: Introduzione - L'elaborazione personale, oggi - Evoluzione hardware - Il problema del software - Operatori e canali di distribuzione - Costruttori di hardware - Società di software e progettazione - Distributori - Funzionamento del negozio - L'attività editoriale - Situazione e tendenze del mercato - Prospettive di impiego domestico dell'elaboratore personale - Conclusione.

Bibliografia.