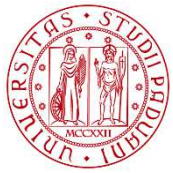

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

ANALOG COMPUTING

Relatore: Prof. Roberto Corvaja

Laureando: Giovanni Ordan

ANNO ACCADEMICO 2023 – 2024

Data di laurea 15/07/2024

Indice

Introduzione

Nel vasto panorama dell'informatica, tra le molteplici tecnologie e metodologie, si trova una categoria di dispositivi che si discosta dall'approccio digitale predominante: i computer analogici. Questi sistemi, basati su principi fisici che operano con grandezze continue anziché discrete, offrono un'alternativa interessante alle architetture digitali tradizionali.

La scelta di esplorare il tema dei computer analogici come argomento per questa tesi è stata motivata dalla crescente attenzione che tale tecnologia ha ricevuto negli ultimi anni. Con la ricerca di soluzioni sempre più efficienti e potenti nell'ambito dell'elaborazione dei dati, i computer analogici hanno suscitato interesse per le loro potenziali capacità di migliorare le prestazioni computazionali.

Nel primo capitolo vengono esaminate le caratteristiche distintive dei computer analogici e le loro differenze rispetto ai sistemi digitali. Viene inoltre fornita una panoramica riguardante la storia dei computer analogici, dagli albori, fino alla loro eclissi verso la fine del Novecento.

Il secondo capitolo espande ulteriormente la comprensione dei computer analogici, concentrando l'attenzione sullo spazio di stato continuo, sulla scala, necessaria per adattare i problemi ai sistemi, ed infine sulla precisione di questi dispositivi.

Nel terzo capitolo vengono presentati due esempi applicativi della tecnologia analogica: la risoluzione di sistemi lineari e Neurogrid, un sistema progettato per simulare modelli neurali in tempo reale su larga scala.

Nel quarto ed ultimo capitolo viene posto in osservazione il campo di ricerca più recente per quanto riguarda l’analog computing, ovvero l’analog in-memory computing. Questo studio si concentrerà su Mythic AI, una delle principali startup nel campo dell'Analog in-Memory Computing, analizzando i loro progressi e le innovazioni che stanno portando nel settore.

Capitolo 1

Storia dell'Analog Computing

1.1. Definizione

L'analog computing prende il nome da un'analogia, o relazione sistematica, tra i processi fisici nel computer e quelli nel sistema che si intende modellare o simulare (il sistema primario). Ad esempio, le quantità elettriche di tensione, corrente, e conduttanza possono essere usate come analoghi della pressione di un fluido, portata e diametro di un tubo. Più nello specifico, nella computazione analogica tradizionale, le quantità fisiche obbediscono alle stesse leggi fisiche e matematiche del sistema primario. Pertanto, le quantità computazionali sono proporzionali alle quantità modellate. Questo è in contrasto con la computazione digitale nella quale le quantità sono rappresentate da stringhe di simboli (e.g. cifre binarie) che non hanno una diretta relazione fisica con le quantità modellate.

In un senso più generale tutto il calcolo è basato sull'analogia, cioè su una relazione sistematica tra gli stati e i processi nel computer e quelli nel sistema primario. In un computer digitale, la relazione è più astratta e complessa di un semplice rapporto di proporzionalità, ma anche un computer analogico così semplice come il regolo calcolatore va oltre la rigorosa proporzione (i.e., la distanza nel regolo è proporzionale al logaritmo del numero). Sia nella computazione analogica che in quella digitale, in verità in tutte le computazioni, la rilevante struttura matematica astratta del problema è realizzata negli stati fisici e nei processi del computer, ma la realizzazione può essere più o meno diretta.

Per questo motivo, nonostante le etimologie dei termini "analogico" e "digitale", nell'uso moderno la principale distinzione tra il calcolo digitale e quello analogico sta nel fatto che il primo opera su rappresentazioni discrete in passaggi discreti, mentre il secondo opera su rappresentazioni continue tramite processi continui. Questa è la distinzione principale che risiede nella topologia degli stati e dei processi; perciò, sarebbe più accurato riferirsi a computazione discreta e continua. Considerando i noti orologi digitali ed analogici possiamo

notare che la differenza principale risiede nella continuità o discretezza della rappresentazione del passare del tempo; il movimento delle due (oppure tre) lancette dell'orologio analogico non imita la rotazione della terra o la posizione del sole relativa ad essa.

1.2. Computer analogici meccanici

Il primo calcolatore analogico di cui si ha traccia è la Macchina di Anticitera (figura 1), risalente al primo secolo a.C. Fu scoperta nel 1900 in un relitto di una nave a largo delle coste dell'isola greca di Anticitera (tra Citera e Creta). Si tratta di un meccanismo composto da 35 ingranaggi in grado di calcolare le previsioni delle eclissi, la posizione e le fasi della luna e infine aveva anche la funzione di calendario lunisolare. Quest'ultimo e la previsione delle eclissi apparivano sulla faccia posteriore, mentre la posizione della luna e la sua fase si potevano vedere sulla faccia superiore.

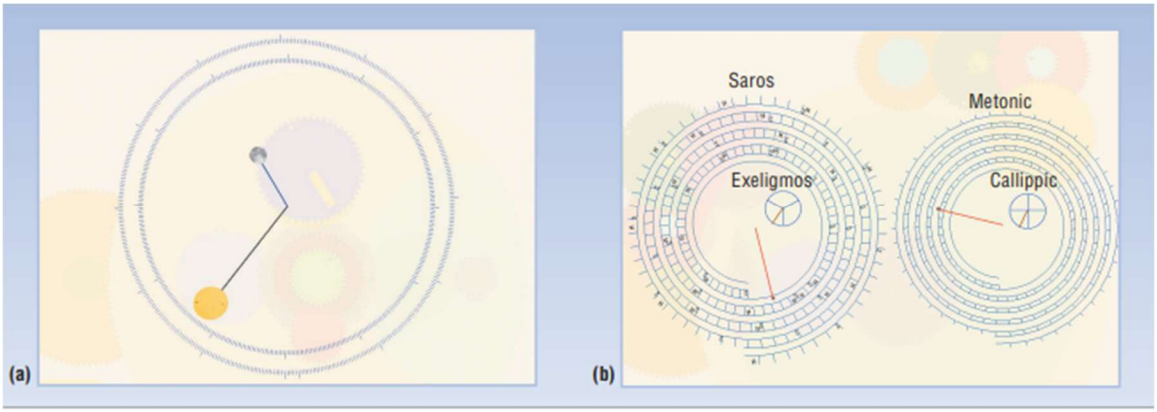

Figura 1: Faccia anteriore e posteriore del Meccanismo di Anticitera.(a) Un quadrante sulla parte anteriore mostra la posizione del Sole durante l'anno sul ciclo zodiacale e un calendario di 365 giorni. (b) La faccia posteriore mostra due quadranti con calendari lunari (raffiguranti il ciclo Metonico e quello Calippico) e due quadranti con le previsioni delle eclissi (raffiguranti il ciclo di Saros e di Exeligmos) [2].

Altri antichi computer analogici meccanici sono l'astrolabio, che veniva usato per determinare la longitudine e aveva anche un'ampia varietà di utilizzi di natura astronomica, e il torqueto, uno strumento medievale che converte misurazioni astronomiche in coordinate equatoriali, orizzontali ed eclittiche.

Una categoria di computer analogici avente scopi particolari è il nomogramma (anche noto come nomografo), che è molto semplice dal punto di vista concettuale ma può essere usato per una vasta gamma di obiettivi. Nella sua forma più comune permette la soluzione di equazioni arbitrarie in tre variabili reali $f(u,v,w) = 0$. Il nomogramma è un grafico o una tabella con scale

per ognuna delle variabili; solitamente queste scale vengono curvate e hanno delle marcature numeriche non uniformi (figura 2). Dati i valori di tutte e due le variabili, viene posta una riga lungo le loro posizioni nelle scale, il valore della terza variabile si può leggere dove la riga attraversa la terza scala. Il punto di forza di questo strumento è quello di migliorare la comprensione intuitiva tramite la visualizzazione delle relazioni tra le variabili in uso e la facilitata analisi della loro variazione ottenibile muovendo la riga.

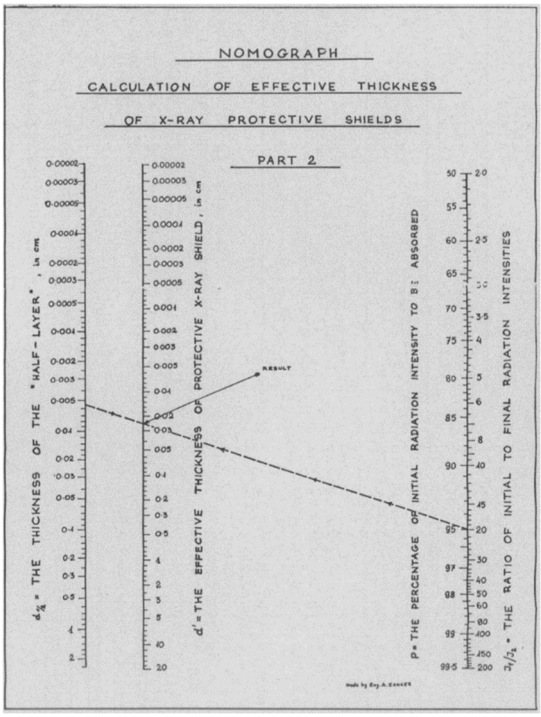

Figura 2: Nomogramma per il calcolo dell'effettivo spessore degli scudi protettivi dai raggi X (1981) [4].

Un altro storico computer analogico meccanico è l'analizzatore differenziale, studiato e realizzato intorno alla seconda metà del 1800 dai fratelli James Thomson (un professore di ingegneria dell'università di Edimburgo) e William Thomson (un professore di fisica dell'università di Glasgow, meglio conosciuto ai giorni nostri come Lord Kelvin) con l'obiettivo di prevedere l'andamento delle maree. Il suo funzionamento verrà approfondito nel secondo capitolo. Gli analizzatori differenziali furono uno dei principali strumenti di calcolo per risolvere problemi con equazioni differenziali durante il ventennio tra il 1930 e il 1950 circa. Vennero prodotte svariate tipologie di macchinari, nelle università e negli stabilimenti di ricerca in Europa e negli Stati Uniti, con un numero sempre crescente di meccanismi di integrazione e sofisticati schemi di interconnessione. Furono ampiamente utilizzati in guerra per fare calcoli balistici al fine di produrre tabelle necessarie per l'artiglieria. Inoltre, i militari li usarono per i puntatori dei cannoni antiaerei e per i mirini. Infine, risultavano molto validi

nel risolvere problemi di meteorologia, propagazione di segnali radio, cristallografia[1], reti elettriche e molti altri campi.

1.3. Computer analogici elettronici

È supposizione comune il fatto che i computer analogici elettronici siano superiori a quelli meccanici, e lo erano per molti aspetti, tra cui velocità, costo, facilità di costruzione, grandezza e portabilità. D'altra parte, gli integratori meccanici (presenti negli analizzatori differenziali) producevano risultati ad alta precisione (0.1% contro 1% dei primi apparecchi elettronici) e avevano una migliore flessibilità dal punto di vista matematico (erano in grado di integrare rispetto ad ogni variabile, non solo il tempo). Comunque, molte importanti applicazioni non richiedevano alta precisione e si concentravano sui sistemi dinamici per i quali l'integrazione rispetto al tempo non era sufficiente.

I computer analogici (sia elettronici che non elettronici) possono essere separati in due categorie: *active-element* e *passive-element* computer; i primi implicano l'utilizzo di una tipologia di amplificatori, i secondi invece no. I computer *passive-element* includevano gli analizzatori di rete (*network analyzers),* che furono sviluppati intorno al 1920 per analizzare la distribuzione di potenza elettrica nelle reti, e rimasero in uso fino agli anni 50. Vennero anche applicati a problemi di termodinamica, progettazione aeronautica ed ingegneria meccanica. Queste reti di sistema o griglie di elementi resistivi o elementi reattivi (ovvero che comprendevano capacità, induttanze e resistenze) venivano usate per modellare la distribuzione spaziale di quantità fisiche come tensione, corrente e potenza (nelle reti di distribuzione elettrica), potenziale elettrico nel vuoto, temperatura (in problemi di diffusione del calore), pressione, portata dei fluidi e ampiezze d'onda. Ovvero, gli analizzatori di rete trattavano equazioni differenziali alle derivate parziali (PDE), mentre i computer *active-element,* come ad esempio l'analizzatore differenziale e i suoi successori elettronici, si concentravano sulle equazioni differenziali ordinarie (ODE) nelle quali il tempo era la variabile indipendente.

I computer analogici elettronici divennero realizzabili dopo l'invenzione dell'amplificatore operazionale ("op amp"). Già negli anni 30, un gruppo di scienziati ai *Bell Telephone Laboratories* (BTL) aveva sviluppato il *DC-coupled feedback-stabilized amplifier,* che era la

[1] La cristallografia è la disciplina scientifica che si occupa dello studio dei cristalli.

base dell'op amp[3]. Nel 1940, quando gli Stati Uniti si stavano preparando ad entrare in guerra (la Seconda Guerra Mondiale), l'ingegnere David Parkinson che allora lavorava ai BTL, fece un sogno nel quale poteva vedere gli amplificatori DC usati per controllare un'arma antiaerea. Di conseguenza, insieme ai suoi colleghi Lovell e Weber, scrisse una serie di articoli nei quali descriveva i circuiti elettronici per "operazionalizzare" addizione, sottrazione, integrazione, differenziazione ecc. Il progetto di costruire un puntatore elettronico portò allo sviluppo e all'affinamento dell'op amp DC adatto al calcolo analogico.

Il lavoro svolto in tempo di guerra ai BTL era concentrato principalmente su applicazioni di controllo dei dispositivi analogici, come ad esempio il puntatore delle armi. Altri ricercatori ai BTL, come ad esempio E. Lakatos [5], erano più interessati nell'applicare gli op amp ad un obiettivo più generale per la scienza e l'ingegneria: questo lavoro risultò nella progettazione del *General Purpose Analog Computer (*GPAC*),* anche noto come "Gypsy", completato nel 1949. Basandosi sul design degli op amp fatto ai BTL, venne svolto del lavoro fondamentale alla Columbia University negli anni Quaranta. In particolare, questa ricerca mostrò come il calcolo analogico potesse essere applicato alla simulazione di sistemi dinamici e alla soluzione di equazioni non lineari.

I computer analogici di uso generico (GPAC) emersero a cavallo tra la fine degli anni Quaranta e l'inizio degli anni Cinquanta. Generalmente fornivano dozzine di integratori, ma più GPAC potevano essere connessi per risolvere problemi più grandi e complessi. Più avanti i GPAC a larga scala arrivarono ad avere fino a 500 amplificatori ed erano in grado di calcolare con una precisione dello 0.01%-0.1%.

Le varianti meccaniche rimasero ancora in uso per un certo periodo di tempo, nonostante i computer analogici elettronici avessero iniziato ad essere utilizzati. I più grandi sviluppatori di questi apparecchi erano gli Stati Uniti, la Russia e il Regno Unito. Inizialmente venivano usati per il design di aerei e per i simulatori di volo; naturalmente la NASA era uno dei maggiori committenti.

1.4. Eclissi dei computer analogici

Una visione comune è che i computer analogici elettronici siano stati dei predecessori dei computer digitali, e che il loro utilizzo sia stato solo un episodio storico, o meglio, una

digressione nell'inevitabile trionfo della tecnologia digitale. Si suppone che l'attuale egemonia del digitale sia una semplice questione di superiorità tecnologica. Tuttavia, la storia è molto più complicata di così: comprende fattori sociali, economici, storici, pedagogici e tecnici. In ogni caso, ci fu un dibattito acceso tra la fine della Seconda Guerra Mondiale e i successivi vent'anni, a proposito dei relativi meriti della computazione analogica e digitale.

Un vantaggio spesso citato dei computer analogici è la velocità. Sebbene i primi computer digitali fossero molto più veloci degli analizzatori differenziali meccanici, rimanevano più lenti (di parecchi ordini di grandezza) rispetto ai computer analogici elettronici. Inoltre, nonostante i computer digitali potessero svolgere rapidamente le operazioni matematiche, i problemi più complessi venivano risolti sequenzialmente, un'operazione alla volta, mentre i computer analogici operavano in parallelo. Venne quindi sostenuto che per i problemi sempre più grandi fosse necessario più tempo per un computer digitale, mentre d'altra parte, i computer analogici potevano richiedere più hardware ma non più tempo. Anche quando la velocità di computazione digitale fu migliorata, la variante analogica mantenne il suo vantaggio per qualche decade, ma questo diminuì progressivamente.

I primi computer analogici elettronici contenevano centinaia di migliaia di tubi a vuoto, rimpiazzati successivamente dai transistor e dovevano essere programmati manualmente collegando i cavi tra i vari componenti. Erano quindi delle macchine molto complesse che richiedevano un personale specializzato per capirle e farle funzionare, questo fu un fattore chiave del loro declino.

Un altro importante aspetto fu quello del confronto tra la precisione dell'una o dell'altra tecnologia. I computer analogici solitamente calcolavano con tre o quattro cifre di precisione, ed era molto costoso fare meglio di così, a causa della difficoltà di fabbricazione delle parti ed altri fattori tecnici. Diversamente, i computer digitali potevano effettuare operazioni aritmetiche con molte cifre di precisione, e il costo dell'hardware era approssimativamente proporzionale al numero di cifre. Al contrario, i sostenitori della computazione analogica sostenevano che molti problemi non necessitavano di una così alta precisione, perché le misurazioni venivano fatte con poche cifre significative e i modelli matematici erano solamente delle approssimazioni. Inoltre, essi distinguevano bene tra precisione ed accuratezza, la quale si riferisce alla conformità del calcolo alla realtà fisica, e argomentavano il fatto che la computazione digitale era spesso meno accurata di quella analogica, a causa di limitazioni numeriche (e.g., troncamento ed errori cumulativi nell'integrazione numerica). Ciononostante,

alcune importanti applicazioni, come il calcolo delle traiettorie dei missili, richiedevano molta precisione, e per queste, il calcolo digitale aveva il vantaggio. Infatti, in qualche modo la precisione veniva considerata intrinsecamente desiderabile, anche nelle applicazioni in cui non era così importante, e venne facilmente scambiata per accuratezza.

Ciò che aprì veramente la porta alla rivoluzione digitale fu la scoperta fatta da Claude Shannon all'interno della sua tesi nel 1936[6]. Egli dimostrò che ogni operazione numerica poteva essere eseguita usando i blocchi fondamentali dell'algebra booleana: due valori, vero o falso, anche scritti come uno e zero, e tre operazioni end, or e not. Ciò rese i computer digitali le più versatili e ideali macchine di calcolo. D'altro canto, ogni computer analogico funziona per un singolo tipo di problema. Inoltre, siccome i computer digitali operano su zero e uno, essi sono più resistenti al rumore. Infatti, è necessario un grande errore per confondere un 1 per uno 0 e viceversa, invece, anche piccoli errori nei computer analogici possono crescere e arrivare al punto di sovrastare il segnale.

Era coinvolto anche un fattore sociale, nel senso che i programmi scritti, la precisione e l'esattezza del calcolo digitale venivano associati alla matematica e alla scienza, mentre le operazioni a mano, la variazione dei parametri e le soluzioni approssimative del calcolo analogico venivano associate agli ingegneri, e quindi l'analog computing ereditò "il più basso status delle scienze ingegneristiche *vis-a-vis*"[2]. Pertanto, la posizione del digital computing venne ulteriormente rafforzata nel processo in cui l'ingegneria divenne più matematica e scientifica durante la Seconda Guerra Mondiale.

Già verso la metà degli anni Cinquanta la competizione tra analogico e digitale si era evoluta nell'idea che esse fossero delle tecnologie complementari. Così si arrivò allo sviluppo di una tipologia di sistemi di calcolo ibridi, in parte analogici ed in parte digitali. In alcuni casi ciò prevedeva l'utilizzo di un computer digitale per controllare un computer analogico; veniva usata la logica digitale per collegare gli elementi analogici, impostare i parametri ed ottenere dati. Grazie a questa simbiosi migliorò notevolmente l'accessibilità e la fruibilità dei computer analogici, ma si ebbe lo svantaggio di allontanare l'utilizzatore dal sistema analogico fisico. Negli Stati Uniti il programma di missili balistici intercontinentali stimolò il perfezionamento dei computer ibridi tra gli anni Cinquanta e Sessanta.

[2] Small, JS. 2001. The Analogue Alternative: The electronic analogue computer in Britain and the USA, 1930–1975.

Queste applicazioni richiedevano la velocità del calcolo analogico per simulare i sistemi di controllo a circuito chiuso e la precisione del calcolo digitale per calcolare le traiettorie con accuratezza. Tuttavia, dal 1970 circa gli ibridi vennero spodestati totalmente dai sistemi digitali. Un importante fattore fu il continuo miglioramento della tecnologia digitale, guidato da una vivace industria digitale. D'altra parte, l'inesatta percezione che l'analog computing fosse obsoleto, unita alla mancanza di educazione rispetto ai vantaggi e alle tecniche della computazione analogica, svalutò molto la reputazione di quest'ultima.

Di rilievo fu anche il fatto che i computer digitali vantavano: una semplice programmabilità, facilità nell'immagazzinamento di dati, alta precisione e la capacità di affrontare problemi di ogni grandezza (con il giusto tempo). Le performance di questi computer migliorarono notevolmente durante gli anni Sessanta e la decade successiva grazie anche allo sviluppo della tecnologia MOS per i circuiti integrati che rese possibile il posizionamento di un grande numero di transistor su un singolo chip.

Un'altra argomentazione posta a favore dei computer digitali era che questi fossero multiuso, dal momento che potevano essere usati per l'elaborazione di dati aziendali e applicati ad altri settori, mentre gli analog computer erano monofunzionali, in quanto si limitavano a specifiche applicazioni. Pertanto, le applicazioni aziendali (ed anche i consumatori) hanno motivato gli investimenti dell'industria informatica nella tecnologia digitale, a discapito della tecnologia analogica.

Nonostante sia convinzione comune che i computer analogici siano scomparsi velocemente non appena i computer digitali sono diventati disponibili, questo è inesatto. Sia per usi generali che per usi speciali, i computer analogici hanno continuato ad essere utilizzati per alcune applicazioni specifiche per un periodo molto più lungo di quanto si possa pensare. Ad esempio, un computer analogico elettronico multiuso, l'Anacom, veniva ancora utilizzato nel 1991[7].

Come detto in precedenza, le ragioni per le quali c'è stata un'eclissi dell'analog computing non erano semplicemente collegate ad una superiorità tecnologica; le circostanze furono molto più complesse.

Con l'avvento dei computer digitali, più veloci ed economici, l'era dell'analog computing finì intorno agli anni 80 del Novecento, ma sviluppi recenti hanno generato nuovamente interesse in questa tecnologia.

Capitolo 2

Fondamenti e caratteristiche dei computer analogici

2.1. Spazio di stato continuo

La principale caratteristica che distingue l'analogico dal digitale, come detto nel Capitolo 1, è il fatto che si operi nel continuo per il calcolo analogico e, invece, nel discreto per il digitale. Per questo motivo sarebbe più preciso definire il calcolo analogico e digitale rispettivamente come calcolo continuo e discreto. Oltre a ciò, sin dagli albori ci sono stati computer ibridi che univano spazi di stato e processi continui e discreti. Pertanto, ci sono vari aspetti in cui lo spazio di stato (un insieme di tutte le possibili condizioni (stati) che descrivono completamente il comportamento di un sistema in un dato istante di tempo) può essere continuo.

Nel caso più semplice lo spazio è costituito da un numero finito (solitamente modesto) di variabili, ognuna delle quali contenente una quantità continua (ad esempio tensione, corrente, carica). Nel tradizionale GPAC esse corrispondono alle variabili delle ODE che definiscono i processi computazionali, ognuna avente di solito un significato indipendente nell'analisi del problema. Matematicamente, le variabili vengono scelte in maniera tale da contenere un intervallo limitato di numeri reali. In un senso pratico, comunque, la loro precisione è limitata dal rumore, dalla stabilità e dalla tolleranza del dispositivo.

Un caso diverso è quello delle reti neurali artificiali tradizionali, ovvero dei modelli composti da neuroni artificiali che emulano il funzionamento del cervello umano, nelle quali lo spazio di stato è più grande dal punto di vista delle dimensioni ma più strutturato che nel caso precedente. I neuroni artificiali sono organizzati in uno o più livelli, ognuno dei quali composto da un numero (possibilmente elevato) di neuroni artificiali. Generalmente ogni livello di neuroni è densamente connesso al livello successivo. In generale ogni livello ha un qualche significato

nel dominio del problema, ma i singoli neuroni che li costituiscono invece no (e quindi, nelle descrizioni matematiche, i neuroni sono solitamente numerati piuttosto che denominati).

I singoli neuroni artificiali di solito eseguono un semplice calcolo come questo:

$$y = \sigma(s), \quad \text{dove } s = b + \sum_{i=0}^{n} w_i x_i,$$

e dove y è l'attività del neurone, mentre x_1,…, x_n, sono le attività dei neuroni che forniscono gli input, b è il termine del bias, e w_1,…,w_n sono i pesi (o le forze) delle connessioni. Spesso la funzione di attivazione σ è una funzione a valori reali sigmoide, come ad esempio la sigmoide logistica,

$$\sigma(s) = \frac{1}{1+ e^{-s}},$$

in questo caso l'attività del neurone y è un numero reale, ma in alcune applicazioni si usa una funzione discontinua, come ad esempio la funzione di Heaviside,

$$U(s)=\begin{cases}+1 & se \ \ s \geq 0 \\ 0 & se \ \ s < 0\end{cases},$$

nel cui caso l'attività risulta essere una quantità discreta. La sigmoide *saturated-linear* o *piecewise-linear* viene anch'essa utilizzata in alcune circostanze:

$$\sigma(s) = \begin{cases}+1 & se\ s > 1 \\ s & 0 \leq s \leq 1 \\ 0 & se \ \ s \geq 1\end{cases}$$

A prescindere dal fatto che la funzione di attivazione sia continua oppure discreta, il bias *b* e i pesi delle connessioni w_1,…,w_n sono numeri reali, come lo è il "*net input"* $s=\sum_i w_i x_i$. Il calcolo analogico può essere utilizzato per analizzare la combinazione lineare s e la funzione di attivazione $\sigma(s)$, se essa è a valori reali. I bias e i pesi sono solitamente determinati da un algoritmo di apprendimento (e.g., *back-propagation),* il quale è a sua volta un buon candidato per l'implementazione analogica.

Ricapitolando, lo spazio di stato continuo di una rete neurale include i valori di bias e gli input netti dei neuroni e le forze di interconnessione tra i neuroni stessi. Inoltre, sono inclusi i valori di attività dei neuroni, se la funzione di attivazione è una sigmoide a valori reali, come spesso accade. Frequentemente grandi gruppi ("strati") di neuroni (e le connessioni tra i gruppi) hanno

un significato intuitivo nel dominio del problema, ma in genere non hanno questo significato le attività individuali dei neuroni, i valori di bias e i pesi delle interconnessioni.

Se si estrapola il numero di neuroni presenti in un livello al limite del continuo, si ottiene un *campo*, il quale può essere definito come una distribuzione continua (una distribuzione di probabilità nella quale la variabile x può avere qualunque valore continuo, e.g. distribuzione normale) oppure una quantità continua (può assumere un'infinità di valori all'interno di un intervallo specificato, ad esempio velocità, temperatura, pressione). È una approssimazione matematica ragionevole trattare come una massa continua un gruppo di neuroni biologici o artificiali, se il loro numero è sufficientemente grande e se la loro disposizione è significativa (come lo è solitamente nel cervello). I campi sono particolarmente utili nella modellazione di *mappe corticali,* nelle quali l'informazione è rappresentata dallo schema dell'attività su una regione di corteccia neurale.

Il calcolo del campo è applicabile specialmente alla risoluzione di PDE e all'elaborazione di informazioni estese nello spazio, come ad esempio le immagini visive. Alcuni dei primi dispositivi di calcolo analogico erano capaci di eseguire il calcolo dei campi [8]. Per esempio, come menzionato nel primo capitolo, una grande rete di resistori e capacità poteva essere utilizzata per risolvere PDE come, ad esempio, nei problemi di diffusione. In questi casi un insieme di resistori e capacità venne usato per approssimare un campo continuo, mentre in altri casi ad essere spazialmente continuo era il mezzo di calcolo. Quest'ultimo utilizzava fogli conduttori (per campi bidimensionali) o contenitori elettrolitici (per campi a tre dimensioni). Quando venivano applicati a problemi di spazi stazionari, questi computer analogici venivano chiamati *field plotters* o *potential analyzers.*

2.2. Scala per i computer analogici

Un importante aspetto dell'analog computing è la scala, che viene utilizzata per adattare un problema ad un computer analogico. Una prima forma di scalamento è il *time-scaling*, che adegua un problema alla scala dei tempi caratteristica nella quale il computer lavora, che è una conseguenza del suo design e del processo fisico da cui è stato realizzato [9]. Per esempio, potremmo volere che la simulazione proceda su una scala temporale diversa da quella del sistema primario. Perciò una simulazione metereologica oppure economica dovrebbe procedere più veloce del tempo reale per ottenere previsioni utili. Al contrario, potremmo voler rallentare

una simulazione del ripiegamento proteico in modo tale da osservare le diverse fasi del processo. Inoltre, per risultati più accurati è necessario evitare di superare la velocità di risposta massima dei dispositivi analogici, la quale potrebbe determinare una più lenta velocità di simulazione. D'altra parte, un calcolo troppo lento potrebbe risultare inaccurato a causa dell'instabilità (e. g., perdite negli integratori).

Il time scaling riguarda solamente le operazioni dipendenti dal tempo come l'integrazione. Per esempio, supponiamo che *t*, il tempo nel sistema primario, sia correlato a τ, il tempo nel computer, da $\tau = \beta t$. Perciò, un'integrazione $u(t) = \int_0^t v(t')dt'$ nel sistema primario è sostituita nel computer dall'integrazione $u(\tau) = \beta^{-1} \int_0^\tau v(\tau')d\tau'$. Dunque, il time scaling può essere semplicemente ottenuto diminuendo di un fattore β il guadagno in ingresso dell'integratore.

È fondamentale nel calcolo analogico la rappresentazione di una quantità continua del sistema primario, come una quantità continua nel computer. Per esempio, uno spostamento *x* in metri può essere rappresentato da un potenziale V in volt. I due sono collegati da un'*ampiezza* o *fattore di grandezza di scala*, $V = \alpha x$, (con le giuste unità di misura volt/metri), scelto per soddisfare i due criteri. Per quanto riguarda α, deve essere sufficientemente piccola affinché l'intervallo della variabile del problema sia compreso nell'intervallo di valori supportati dal dispositivo. Eccedere l'intervallo operativo del dispositivo può portare a risultati imprecisi (ad esempio forzare un dispositivo lineare ad un comportamento non lineare). D'altra parte, il fattore di scala non dovrebbe essere troppo piccolo, altrimenti una rilevante variazione della variabile del problema sarà minore della risoluzione del dispositivo, portando nuovamente ad inaccuratezza.

In aggiunta alle variabili esplicite del sistema primario, ci sono le variabili implicite, come ad esempio le derivate temporali delle variabili esplicite, e i fattori di scala devono essere scelti anche per loro. Per esempio, oltre allo spostamento di *x*, un problema potrebbe includere velocità $\dot{x}$ e accelerazione $\ddot{x}$. Per questo motivo, i fattori di scala α, α' e α'' devono essere scelti in modo tale che αx, $\alpha' \dot{x}$ e $\alpha'' \ddot{x}$ abbiano un giusto intervallo di variazione (né troppo ampio, né troppo stretto).

Una volta che il fattore di scala è stato scelto, le equazioni del sistema primario vengono regolate in modo tale da ottenere le equazioni del calcolo analogico. Per esempio, se abbiamo scalato $u = \alpha x$ e $v = \alpha' \dot{x}$, allora l'integrazione $x(t) = \int_0^t \dot{x(t')}dt'$ sarà calcolata dall'equazione scalata:

$$u(t) = \frac{\alpha}{\alpha'} \int_0^t v(t')dt'.$$

Ciò viene ottenuto semplicemente impostando l'input gain dell'integratore a $\frac{\alpha}{\alpha'}$.

Praticamente, il *time scaling* e il *magnitude* (grandezza) *scaling* non sono indipendenti. Ad esempio, se le derivate di una variabile sono grandi, allora la variabile stessa può variare rapidamente, e quindi potrebbe essere necessario rallentare il calcolo ai fini di evitare di superare la risposta ad alta frequenza del computer. Invece, delle derivate molto piccole potrebbero richiedere al calcolo di essere eseguito più velocemente per evitare perdite dell'integratore. Fattori di scala appropriati sono determinati considerando sia gli aspetti fisici che quelli matematici di un problema. Ovvero, in primo luogo, la fisica del sistema primario può essere limitata dall' intervallo delle variabili e delle loro derivate. Secondariamente, le analisi delle equazioni matematiche che descrivono il sistema, possono fornire informazioni aggiuntive sul range delle variabili. Per esempio, in qualche caso la frequenza naturale di un sistema può essere stimata dai coefficienti delle equazioni differenziali; il massimo dell'n-esima derivata viene successivamente calcolato come la potenza di n della sua frequenza. In ogni caso, non è necessario avere dei valori accurati per gli intervalli; sono adeguate le stime che riescono ad approssimare l'ordine di grandezza.

Si è tentati a pensare al magnitude scaling come ad un problema unico dell'analog computing, ma prima dell'invenzione dei numeri a virgola mobile era necessario nella programmazione di computer digitali. In ogni caso, è un aspetto essenziale dell'analog computing, in cui i processi fisici sono usati più direttamente per il calcolo rispetto che nel digital computing. Sebbene la necessità di scalare sia stata una fonte di critiche, i sostenitori dell'analogico hanno argomentato che essa sia una fortuna nella sfortuna, perché porta ad una migliore comprensione del sistema primario, che spesso è l'obiettivo del calcolo in primo luogo [10]. I professionisti dell'analog computing sono più inclini ad avere una comprensione intuitiva sia del sistema primario che della sua descrizione matematica.

2.3. Precisione

Il calcolo analogico viene valutato sia in termini di accuratezza che di precisione, ma le due devono essere distinte attentamente. L'**accuratezza** si riferisce prevalentemente alla relazione

tra una simulazione e il sistema primario che sta simulando o, più in generale, la relazione tra i risultati ottenuti dal calcolo e i risultati matematicamente corretti. L'accuratezza è il risultato di molti fattori, tra i quali il modello matematico scelto, il modo in cui viene configurato sul computer e la precisione dei dispositivi di calcolo analogico.

La **precisione**, invece, è un concetto più ampio, che si riferisce alla qualità di una rappresentazione o di un dispositivo di calcolo. Nell'analog computing, la precisione dipende dalla *risoluzione* (finezza delle operazioni) e dalla *stabilità* (assenza di deriva), e può essere misurata come una frazione del valore rappresentato. Pertanto, una precisione dello 0,01% significa che la rappresentazione resterà per un periodo di tempo ragionevole entro lo 0,01% del valore rappresentato. Con lo scopo di confrontare diversi dispositivi analogici, la precisione è solitamente espressa come una frazione della *full-scale variation* (ad esempio, la differenza tra il massimo e il minimo dei valori rappresentati).

È evidente come la precisione dei dispositivi analogici dipenda da diversi fattori. Anzitutto dalla scelta del processo fisico e dal modo in cui esso viene utilizzato nel dispositivo. Per esempio, un'operazione matematica lineare può essere realizzata usando una regione lineare di un processo fisico che non è lineare, ma la realizzazione sarà approssimativa e avrà delle imprecisioni. Dunque, collegato a ciò, gli effetti fisici inevitabili (e.g., carica/scarica di una capacità, dispersione e altre perdite) possono impedire l'implementazione precisa di una funzione matematica. Inoltre, ci sono delle limitazioni fisiche fondamentali alla risoluzione (e.g., diffrazione). Il rumore è inevitabile, sia intrinseco (rumore termico) che estrinseco (radiazione ambientale). Cambiamenti nelle condizioni fisiche dell'ambiente, come la temperatura, possono incidere sui processi fisici e diminuire la precisione. Su scale temporali più lente, i materiali e i componenti invecchiano e le loro caratteristiche fisiche cambiano. Per di più, ci sono sempre dei limiti tecnici ed economici al controllo dei componenti, dei materiali e dei processi di fabbricazione di dispositivi analogici.

La precisione dei dispositivi analogici e digitali dipende da fattori molto diversi tra loro. La precisione di un dispositivo digitale (binario) dipende dal numero di cifre, che influenza il quantitativo di hardware, ma non la sua qualità. Per esempio, un sommatore a 64-bit ha dimensioni doppie rispetto ad un sommatore a 32-bit, ma può essere fabbricato utilizzando gli stessi componenti. Nel caso peggiore, la grandezza di un dispositivo digitale può aumentare con il quadrato del numero di bit di precisione. Questo è causato dal fatto che i dispositivi digitali binari necessitano di rappresentare solo due stati e perciò possono lavorare in

saturazione. Gli standard di fabbricazione sufficienti per il primo bit di precisione sono anche sufficienti per il sessantaquattresimo bit. I dispositivi analogici, al contrario, necessitano di essere in grado di rappresentare precisamente una continuità degli stati. Perciò, la fabbricazione di dispositivi analogici ad alta precisione risulta essere molto più costosa rispetto ai dispositivi a bassa precisione, in quanto la qualità dei componenti, dei materiali e dei processi deve essere controllata molto più meticolosamente. Raddoppiare la precisione di un dispositivo analogico può essere costoso, considerando che il costo di ogni bit di precisione addizionale è incrementale; ossia, il costo è proporzionale al logaritmo della precisione, espresso come una frazione del range completo.

Le considerazioni precedenti possono sembrare un argomento a favore della superiorità della tecnologia digitale rispetto a quella analogica, e infatti, intorno alla metà del ventesimo secolo, sono state un fattore cardine nella "competizione" tra computer digitali ed analogici. Comunque, come venne discusso a quei tempi, molte applicazioni informatiche non richiedevano un'alta precisione. Infatti, in molte applicazioni di ingegneria, i dati di input sono noti essere di solo poche cifre e le equazioni possono essere approssimate o derivate dagli esperimenti. In questi casi, l'alta precisione dei computer digitali non è necessaria e infatti potrebbe essere fuorviante (e.g., se vengono mostrate tutte le 14 cifre di un risultato che è accurato solo fino a 3). Per di più, molte applicazioni nel campo del processamento e controllo delle immagini non richiedono un'alta precisione.

Le ricerche nel campo delle reti neurali artificiali (ANN) hanno dimostrato che il calcolo analogico a bassa precisione risulta essere sufficiente per quasi tutte le applicazioni ANN. In effetti, il processamento delle informazioni neurali nel cervello sembra operare con una precisione molto bassa (forse meno del 10%) [11].

Capitolo 3

Esempi applicativi dei computer analogici

3.1. Risoluzione di sistemi lineari utilizzando computer analogici

Nonostante i computer analogici siano ideali per risolvere le equazioni differenziali, in questa sezione verrà spiegato come essi possono essere utilizzati per risolvere equazioni lineari.

La figura 3 mostra i simboli di programmazione del computer analogico usati in seguito: un divisore di tensione chiamato *coefficiente potenziometrico*, *sommatori* e *integratori*, i quali hanno un amplificatore operazionale ad alto guadagno al centro e rispettivamente un resistore o una capacità nel loro circuito di retroazione negativo.

Le relazioni ingresso-uscita dei tre circuiti di figura 3 sono:

1) $y = \alpha x$ per il coefficiente potenziometrico.
2) $y = \sum_{i=1}^{n} x_i$ per il sommatore.
3) $y = \int \sum_{i=1}^{n} x_i \ dx$ per l'integratore.

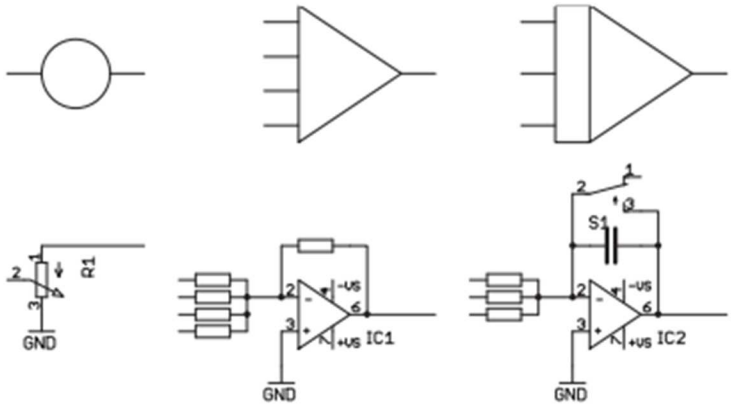

Figura 3: Circuiti di base per potenziometri, sommatori ed integratori [12].

Definendo A, $\vec{b}$ e $\vec{x}$ come segue

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n},\ \vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n,\ \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n ,$$

l'obiettivo è quello di risolvere un sistema di equazioni lineari come

$$A\vec{x} = \vec{b} \tag{1}$$

L'idea alla base è di trasformare un sistema lineare come quello rappresentato sopra (1) in un sistema di equazioni differenziali accoppiate. È importante notare che la risoluzione di tali equazioni tramite l'utilizzo di un computer analogico è un processo continuo per cui non c'è nessun passaggio alla dimensione discreta, come invece accade nei classici algoritmi numerici.

Per questo utilizzo, il vettore soluzione $\vec{x}$ viene inizializzato ad un certo valore e viene calcolato un vettore di errore $\vec{\varepsilon}$ che viene successivamente utilizzato per correggere la prima ipotesi di $\vec{x}$:

$$A\vec{x} - \vec{b} = \vec{\varepsilon} \text{ poi}$$

$$\dot{\vec{x}} = -\vec{\varepsilon}$$

Oppure, dal punto di vista dei componenti del vettore

$$-\dot{x_\iota} = \sum_{j=1}^{n} a_{ij}x_j - b_i,\ \ 1 \le i \le n, \tag{2}$$

dove $\dot{x_\iota}$ rappresenta la derivata temporale di x_i.

Un set di equazioni come quello in (2) può essere trasformato direttamente in una configurazione per un computer analogico applicando la tecnica di feedback di Kelvin [12]. La figura 4 mostra questo per un sistema avente tre incognite come

$$-\dot{x_1} = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 - b_1$$

$$-\dot{x_2} = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 - b_2$$

$$-\dot{x_3} = a_{31}x_1 + a_{32}x_3 + a_{33}x_3 - b_3$$

L'elemento computazionale principale utilizzato in questo setup, a parte i coefficienti potenziometrici, è un integratore che calcola come uscita l'integrale rispetto al tempo della somma dei suoi input ed esegue anche un cambio di segno implicito.

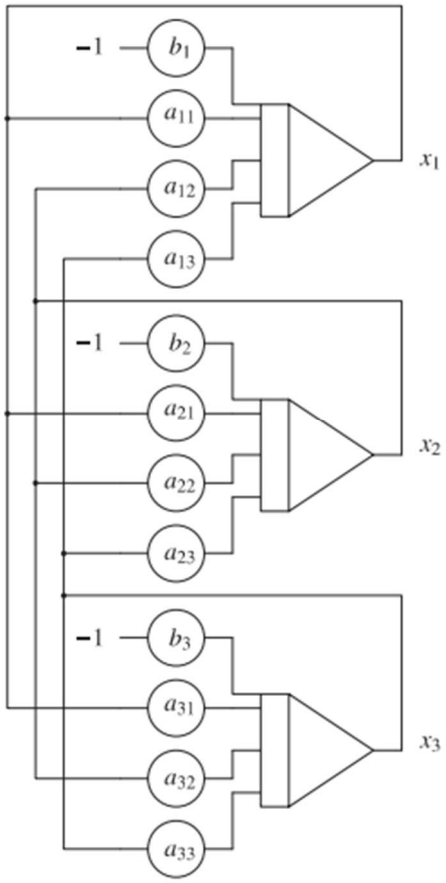

Figura 4: Approccio per risolvere un sistema di equazioni lineari su un computer analogico [12].

Una soluzione intelligente per questo problema è di trasformare $A\vec{x} = \vec{b}$ in un sistema corrispondente di equazioni lineari con una matrice di coefficienti che è sempre definita positiva [13]. Ciò può essere facilmente raggiunto moltiplicando (1) per A^T da sinistra ottenendo

$$A^T A\vec{x} - A^T\vec{b}$$

Un'implementazione puramente analogica può essere raggiunta calcolando l'errore risultante da un'impostazione iniziale di $\vec{x}$ in questo modo

$$A^T A\vec{x} = A^T\vec{b}$$

poi impostando

$$\dot{\vec{x}} = -\vec{\xi}$$

come in precedenza. Raccogliendo A^T si ottiene

$$A^T\left(A\vec{x} - \vec{b}\right) = A^T\vec{\varepsilon} = \vec{\xi} = -\dot{\vec{x}}.$$

Il risultato appena ottenuto può essere diviso in due set di equazioni accoppiate

$$\varepsilon = \sum_{j=1}^{n} a_{ij} x_j - b_i \quad (3)$$

$$\dot{x}_\iota = -\sum_{j=1}^{n} a_{ji} \varepsilon_i \quad (4)$$

con $1 \leq i \leq n$. Queste equazioni possono ora essere trasformate in un setup per un computer analogico come mostrato in figura 5.

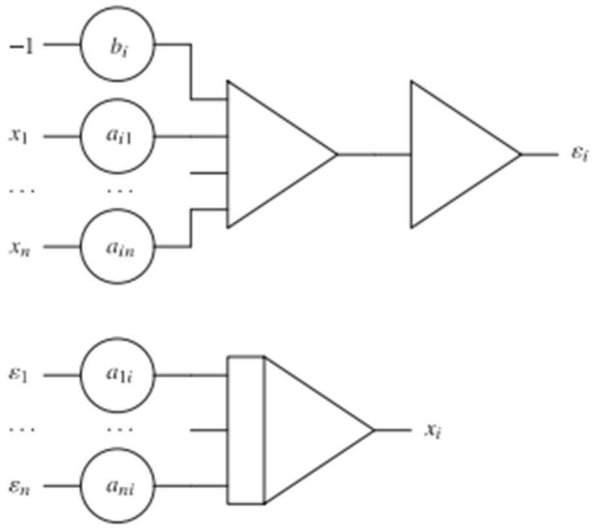

Figura 5: Setup del computer per le equazioni (3) e (4). [12]

In conclusione, l'approccio alla risoluzione di equazioni lineari accoppiate tramite l'utilizzo di un computer analogico descritto in precedenza ha mostrato che talvolta, trasformando un problema semplice in uno più complesso, come ad esempio le equazioni differenziali accoppiate, può non solo fornire nuove interessanti intuizioni ad un problema già ben noto, ma può anche condurre ad un'inaspettata soluzione pratica. Nonostante queste tecniche fossero considerate inadeguate per applicazioni pratiche agli albori dei computer analogici elettronici a causa del grande numero di potenziometri richiesti $(n^2 + n)$, ciò non risulta più essere proibitivo dati i moderni circuiti analogici altamente integrati.

Tralasciando un'implementazione puramente analogica come quella descritta qui, sarà molto interessante utilizzare tale configurazione come parte di un più complesso *computer ibrido* costituito da un computer digitale con programma memorizzato e da un computer analogico. Quest'ultimo sarebbe in grado di portare ad una soluzione approssimativa per un sistema dato di equazioni lineari che potrebbe successivamente, se necessario, essere migliorata ulteriormente da un computer digitale.

3.2. Neurogrid

In questa sezione viene fornito l'esempio di Neurogrid [15], il primo sistema multi-chip ibrido (sia analogico che digitale) in grado di eseguire simulazioni in tempo reale su larga scala di modelli neurali, ovvero dei neuroni e delle loro connessioni.

I sistemi neuromorfici come Neurogrid realizzano la funzione di sistemi neurali biologici tramite emulazione della loro struttura. I suoi progettisti hanno dovuto fare tre scelte di progetto:

1) Se emulare i principali elementi neurali (albero assonale, albero dendritico, sinapsi e soma) con circuiti elettronici dedicati o condivisi;
2) Se implementare questi circuiti in modo analogico oppure digitale;
3) Se interconnettere gli array dei neuroni tramite una rete ad albero o a maglia

Le scelte che sono state fatte sono le seguenti:

1) Tutti gli elementi neurali tranne il soma sono stati emulati con circuiti elettronici condivisi; questa scelta è stata fatta per massimizzare il numero di connessioni sinaptiche;
2) Tutti i circuiti elettronici, eccetto gli alberi assonali, sono stati realizzati in maniera analogica; questa scelta ha permesso di massimizzare l'efficienza energetica;
3) Gli array di neuroni sono stati interconnessi in una rete ad albero; questo è stato fatto per massimizzare la produttività.

Le scelte sopraelencate hanno reso possibile la simulazione in tempo reale di un milione di neuroni con miliardi di connessioni sinaptiche utilizzando 16 Neurocore (figura 6) integrati su una scheda che consuma tre watt.

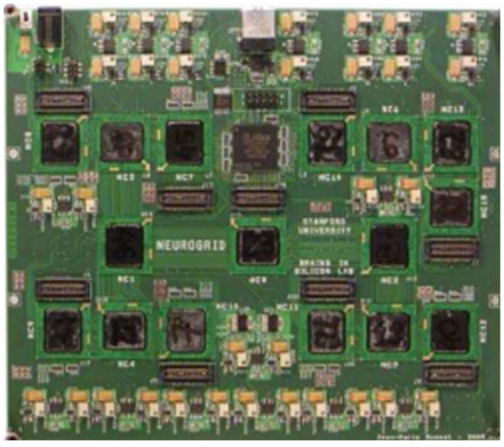

Figura 6: Scheda: ogni strato neurale viene simulato da un massimo di 256 × 256 neuroni di silicio su ciascuno dei 16 Neurocore integrati sulla scheda. [15]

Nell'interfaccia grafica di Neurogrid (figura 7) si può notare sulla sinistra una parte in cui è possibile cambiare i parametri del proprio modello, al centro si può visualizzare l'attività nei vari strati del modello, sulla destra vengono tracciati i raster da uno strato neurale selezionato, infine in basso c'è una sezione in cui è possibile immettere i comandi.

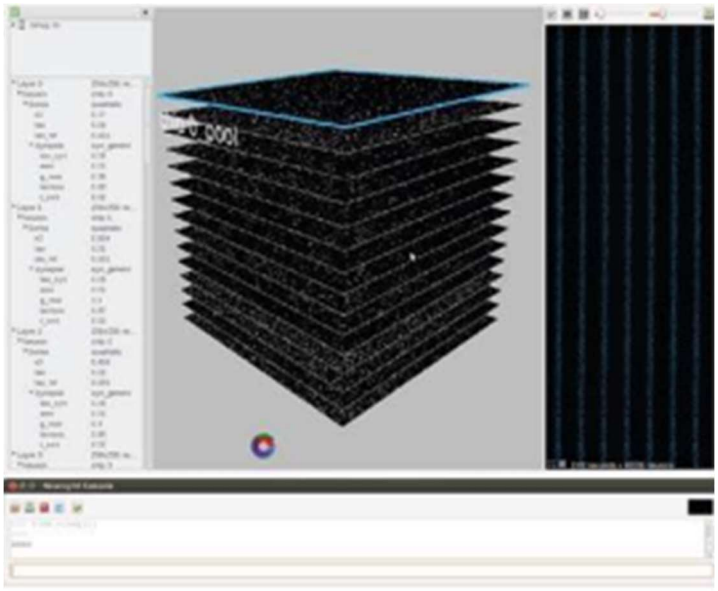

Figura 7: GUI di Neurogrid [15]

Il progetto di Neurogrid utilizza un approccio analogico con l'obbiettivo di ridurre il quantitativo di transistor tramite la condivisione di sinapsi e dei circuiti dell'albero dendritico.

Nella più semplice implementazione completamente analogica, gli elementi neurali di cui abbiamo parlato in precedenza vengono emulati da un filo, una sorgente di corrente commutata, un altro filo ed infine un comparatore [15] (figura 8). Gli impulsi di segnale in entrata sul filo verticale misurano la carica sul filo orizzontale, la cui capacità integra la carica. Il comparatore confronta la tensione risultante con una soglia e attiva un picco in uscita quando la soglia viene superata. Il condensatore viene quindi scaricato e il ciclo ricomincia. La tensione di polarizzazione della sorgente di corrente commutata, che determina il peso sinaptico, viene memorizzata in modo analogico o digitale; quest'ultimo richiede l'utilizzo di un convertitore analogico-digitale.

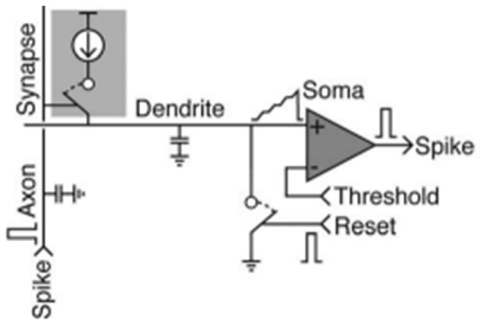

Figura 8: Implementazione analogica del neurone artificiale. [15]

In termini di efficienza energetica si possono confrontare le prestazioni di Neurogrid, PC digitale e cervello umano. Il primo risulta essere ottimo dal punto di vista del consumo, in quanto la sua efficienza energetica è cinque ordini di grandezza migliore rispetto a quella di un personal computer; tuttavia, rimane da quattro a cinque ordini di grandezza peggiore rispetto al cervello umano. Neurogrid, inoltre, utilizza pochi watt per simulare milioni di neuroni, mentre un PC usa centinaia di watt per simulare 2,5 milioni di neuroni. Dal punto di vista della velocità computazionale, invece, Neurogrid riesce a simulare in tempo reale, mentre il personal computer digitale lo fa 9.000 volte più lentamente. Il cervello umano, infine, con 80.000 volte i neuroni di Neurogrid, utilizza solo tre volte la sua potenza. Il raggiungimento di tale livello di efficienza, mantenendo una buona configurabilità e scala è la grande sfida che gli ingegneri neuromorfici devono affrontare.

Capitolo 4

Analog in-Memory Computing

4.1 Introduzione all'Analog in-Memory Computing

Per quanto riguarda l'analog computing il campo applicativo più recente ed interessante è sicuramente quello dell'analog in-memory computing. Quest'ultimo consiste in un metodo di calcolo che sfrutta i dispositivi di memoria analogica, ovvero quelli in grado di memorizzare i dati sotto forma di valori analogici (ad esempio memristori e transistor a gate flottante), per effettuare le operazioni direttamente sui dati analogici in memoria.

Solitamente nella tecnologia digitale i dati vengono memorizzati utilizzando valori discreti (1 e 0), questi ultimi vengono poi memorizzati in celle di memoria DRAM o flash. Per essere elaborati, i dati necessitano di un trasferimento dalla memoria di archiviazione alla CPU, dove vengono processati tramite l'utilizzo di operazioni logiche digitali (porte logiche e le loro combinazioni). D'altro canto, nell'analog in-memory computing le operazioni sui dati vengono eseguite sfruttando le proprietà fisiche dei dispositivi.

L'analog in-memory computing, quindi, effettuando le operazioni di calcolo direttamente in memoria, evita lo spostamento di grandi quantità di dati, con una diretta conseguenza di risparmio dal punto di vista energetico e velocità di calcolo superiore.

Negli ultimi anni stiamo osservando la crescita e lo sviluppo dell'intelligenza artificiale, probabilmente la tecnologia più promettente ed innovativa di questo millennio. Attualmente le reti neurali vengono di norma simulate seguendo dei modelli di calcolo che vengono eseguiti su architetture digitali. Per questo motivo la velocità di calcolo potrebbe non essere ottimale e l'efficienza energetica potrebbe essere ampiamente migliorata tramite l'utilizzo di dispositivi analogici che permettono il calcolo direttamente in memoria.

4.2 Mythic AI

In questo paragrafo verrà approfondita Mythic AI, ossia una delle startup più importanti che si sono concentrate sulla tecnologia dell'analog in-memory computing, vedendone da subito il potenziale. Il loro lavoro si concentra nel creare chip analogici da utilizzare nello studio delle reti neurali.

L'idea di Mythic è quella di permettere il calcolo immediatamente in prossimità delle celle di memoria, aggiungendo semplicemente a queste ultime un'elaborazione locale come si può vedere nella figura 9. L'utilizzo di questa tipologia di architettura da un lato obbliga ad avere una memoria di grandi dimensioni, ma dall'altro permette di avere le stesse prestazioni di una cache L1 [16].

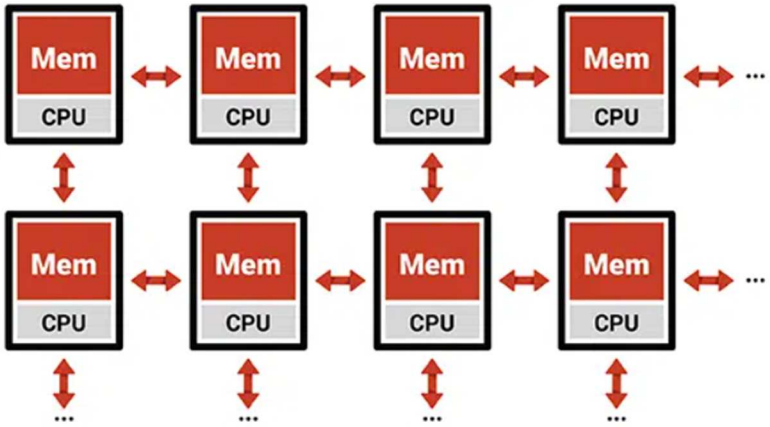

Figura 9: Compute in-memory [16].

Il calcolo in memoria avviene all'interno dell'array di memoria, tramite l'utilizzo di resistori regolabili (o variabili) dando in input una tensione e ricevendo in output le correnti. Questa tipologia di calcolo viene utilizzata per le operazioni della rete neurale principale, dove viene moltiplicato un vettore di input per una matrice dei pesi [16].

I resistori regolabili citati in precedenza hanno quindi un ruolo fondamentale nel procedimento di calcolo direttamente in memoria. Essi possono essere implementati utilizzando un transistor a gate flottante (figura 10) che solitamente viene utilizzato nelle celle di memoria flash per memorizzare un valore digitale. Tra source e drain vi è un canale nel quale il passaggio di elettroni è regolato da un campo elettrico che si crea applicando delle tensioni ai due terminali; invece, applicando una tensione elevata al gate gli elettroni che scorrono nel canale tra source e drain vengono attirati nel gate flottante, che si trova sopra al canale oltre uno strato di isolante e rimangono intrappolati al suo interno. Per memorizzare un 1 logico il gate flottante deve

essere vuoto e per leggere il valore è sufficiente applicare una piccola tensione nei terminali di gate e source o drain in modo tale da far scorrere gli elettroni; è possibile misurare anche la corrente tramite un amperometro collegato al drain o al source. Per memorizzare uno 0 logico invece, è necessario che il gate flottante sia riempito di elettroni, per leggere il valore si applicano le stesse tensioni del caso precedente e si misura la corrente, che però in questo caso non ci sarà perché non vi è passaggio di elettroni tra source e drain in quanto essi sono intrappolati nel gate flottante.

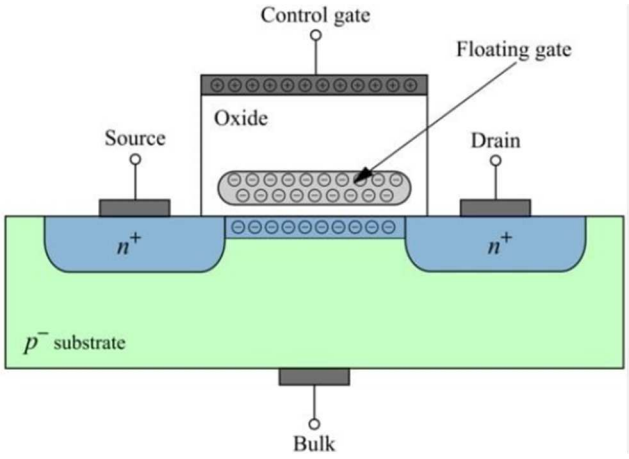

Figura 10: Transistor a gate flottante [17].

L'intuizione di Mythic è quella di utilizzare questi dispositivi non come memorie di 1 e 0 ma come resistori regolabili; ciò avviene posizionando un numero preciso di elettroni (al posto di tutto o niente) all'interno del gate flottante, più è grande il numero di questi elettroni, maggiore sarà la resistenza del canale. Applicando ora una piccola tensione al gate di controllo, la corrente i che scorre nel canale risulta essere $i = {}^{V}/_{R}$ oppure $i = GV$, dove G è la conduttanza. Dunque, una singola cella può essere utilizzata per ottenere il risultato di una moltiplicazione tra conduttanza e tensione applicata, il cui risultato è una corrente misurabile in ampere.

Nell'utilizzo di questa tecnologia per la gestione di una rete neurale vengono scritti i pesi sulle celle di memoria flash come conduttanze, successivamente si inseriscono come tensioni sulle celle i valori di attivazione; la corrente misurata è quindi il prodotto di valore di attivazione per peso. Le celle sono poi collegate tra loro (figura 11) affinché la corrente di ogni moltiplicazione si sommi alle altre, ottenendo come risultato finale la moltiplicazione del vettore di input per la matrice dei pesi.

I due vantaggi principali come detto in 4.1 sono l'efficienza energetica e la velocità di calcolo. La prima è favorita dal calcolo direttamente in memoria, che quindi non richiede lo spostamento

di un grande numero di dati, mentre la seconda è una conseguenza del fatto che le centinaia di migliaia di operazioni che vengono effettuate siano tutte operate in parallelo.

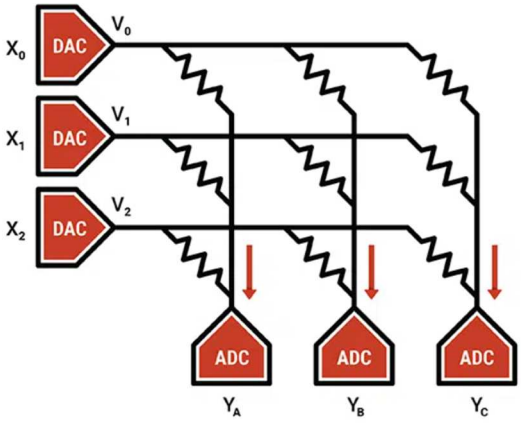

Figura 11: Architettura per moltiplicare i valori di attivazione per i pesi. I resistori sono le celle di memoria flash [17].

Il prodotto principale di Mythic AI si chiama *Mythic Analog Matrix Processor (Mythic AMP)*, il quale è in grado di fornire le risorse computazionali di una GPU consumando una potenza dieci volte minore, il tutto in un singolo chip. Quest'ultimo è progettato in modo tale da avere una serie di componenti di calcolo, al cui interno ognuno presenta un *Mythic Analog Compute Engine (Mythic ACE)* che utilizza convertitori analogico-digitale e memorie flash in combinazione per memorizzare i parametri del modello ed eseguire moltiplicazioni di matrici mantenendo un basso consumo di potenza ed altissime prestazioni [16].

Conclusione

Concludendo, l'analisi dei computer analogici e delle tecnologie emergenti come l'analog in-memory computing evidenzia il loro potenziale nel migliorare le prestazioni e l'efficienza dei sistemi informatici moderni. L'esplorazione della storia e delle applicazioni pratiche di tali dispositivi ha rivelato la loro versatilità e la loro capacità di offrire alternative innovative alle tradizionali architetture digitali.

Inoltre, è importante sottolineare l'importanza della simbiosi tra tecnologie analogiche e digitali nel raggiungere obiettivi di ottimizzazione e innovazione. Storicamente, questa combinazione ha dimostrato di essere estremamente efficace nel soddisfare una vasta gamma di esigenze computazionali, consentendo di sfruttare al meglio le particolarità di entrambe le tipologie di elaborazione.

Guardando al futuro, l'evoluzione dei computer analogici e delle tecnologie analog in-memory offre promettenti opportunità per affrontare sfide complesse e per superare i limiti delle tecnologie attuali. Mythic AI e altre realtà all'avanguardia stanno aprendo nuove strade per l'innovazione, proponendo soluzioni ibride che combinano il meglio dei mondi analogico e digitale.

In conclusione, l'adozione e lo sviluppo di queste tecnologie rappresentano un passo significativo verso il raggiungimento di sistemi informatici sempre più potenti, efficienti e adattabili alle esigenze del mondo moderno.

Bibliografia

[1] Y. Tsividis, "Not your Father's analog computer," in *IEEE Spectrum*, vol. 55, no. 2, pp. 38-43, 2018.

[2] D. Spinellis, "The Antikythera Mechanism: A Computer Science Perspective," in *Computer*, vol. 41, no. 5, pp. 22-27, 2008.

[3] B. J. MacLennan, "A Review of Analog Computing." Department of Electrical Engineering & Computer Science University of Tennessee, Knoxville, 2007.

[4] A. Zanker, "Nomographs to Determine the Effective Thickness of X-Ray Shields," in *IEEE Transactions on Nuclear Science*, vol. 28, no. 1, pp. 985-988, 1981.

[5] J. S. Small, "The Analogue Alternative: The electronic analogue computer in Britain and the USA, 1930–1975", Routledge, 2001.

[6] C. E. Shannon, "A Symbolic Analysis of Relay and Switching Circuits". University of Michigan, 1936.

[7] W. Aspray, "Edwin L. Harder and the Anacom: Analog Computing at Westinghouse." In *IEEE Annals of the History of Computing*, 1993.

[8] T. D. Truitt, A. E. Rogers. "Basics of Analog Computers.", New York: John F. Rider, 1960.

[9] G. R. Peterson, "Basic Analog Computation." New York: Macmillan, 1967.

[10] C. C. Bissell, "A Great Disappearing Act: The Electronic Analogue Computer." in *IEEE Conference on the History of Electronics*, Bletchley, UK, 2004.

[11] J. L. McClelland, D. E. Rumelhart, PDP Research Group, "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models", Cambridge, MA: MIT Press, 1986.

[12] B. Ulmann, D. Killat, "Solving systems of linear equations on analog computers," Kleinheubach Conference, Miltenberg, Germania, 2019

[13] L. D. Kovach, H. F. Meissinger, “Solution of Algebraic Equations, Linear Programming, and Parameter Optimization”

[14] C. Mead, “Analog VLSI and Neural System”, Boston, MA, USA, 1989.

[15] B. V. Benjamin et al., "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," in *Proceedings of the IEEE*, vol. 102, no. 5, pp. 699-716, 2014.

[16] https://mythic.ai/technology/compute-in-memory/ (Ultima consultazione: 29/04/2024)

[17] S. Ilić, A. Jevtić, S. Stanković, G. Ristić, “Floating-Gate MOS Transistor with Dynamic Biasing as a Radiation Sensor.”, 2020.