Dal MOSFET planare al chiplet, vediamo come la tecnologia dei semiconduttori è cresciuta per aumentare sempre più la potenza di calcolo, con un occhio di riguardo al consumo energetico e all'efficienza complessiva.

L'elettronica digitale allo stato solido è nata con le famiglie logiche basate su transistor: inizialmente con la tecnologia TTL (Transistor-Transistor Logic, cioè logica a transistor bipolari) e, successivamente, con i MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor, ovvero transistor ad effetto di campo), adottati per ridurre l'ingombro e il consumo di energia.

I dispositivi fondamentali erano le porte logiche OR, AND, XOR, e le loro corrispettive inverse NOR, NAND, XNOR, oltre alla porta NOT.

A partire da questi elementi base, è stato possibile realizzare funzioni logiche sempre più complesse, fino alle unità aritmetico-logiche (ALU), che costituiscono il cuore delle CPU

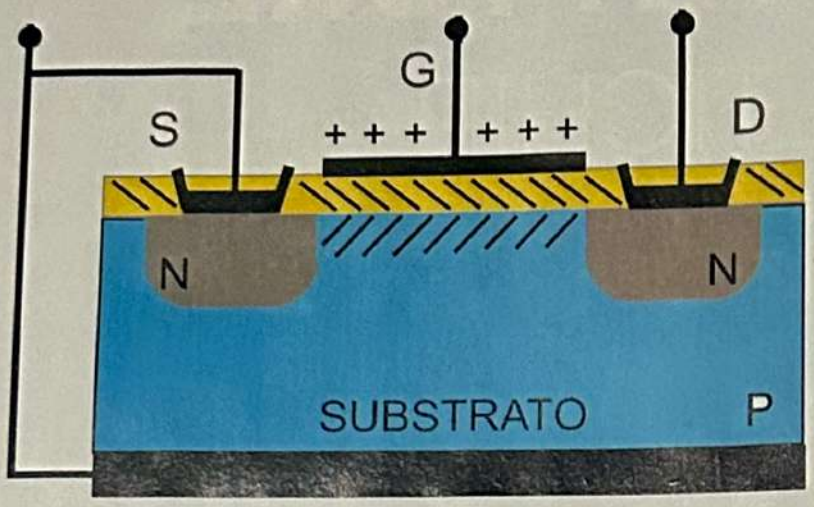


Fig. 1 Tipica struttura del MOSFET, in questo caso a canale N.

(Central Processing Unit), dei microprocessori (MPU) e dei microcontrollori (MCU).

A questi dispositivi logici si sono affiancate le memorie (RAM, ROM, PROM, EPROM, EEPROM e Flash) che condividono le stesse tecnologie di base e che anzi, nelle versioni programmabili e riscrivibili, si basano tutte sulla tecnologia MOSFET o MOS che dir si voglia, perché consente di accumulare permanentemente e rimuovere cariche elettriche.

Benché, per ragioni costruttive, i MOS siano stati per un certo periodo messi in disparte a causa della loro lentezza nella commutazione — dovuta alla natura capacitiva del gate del MOSFET — con il progresso delle tecnologie dei semiconduttori e delle tecniche fotolitografiche è diventato possibile realizzare porte e funzioni logiche basate sul MOSFET come dispositivo fondamentale, capaci di competere con i più veloci transistor bipolari nei tempi di commutazione e di superarli nel contenimento dei consumi e, inevitabilmente, nella riduzione del calore prodotto dai dispositivi, un elemento critico da dover smaltire.

La tecnologia MOS è stata inizialmente sviluppata utilizzando come elemento base il MOSFET a canale N (**Fig. 1**); non a caso si parlava di dispositivi NMOS — come ad esempio le memorie — perché, essendo unipolare, il transistor ad effetto di campo risulta più performante nella versione in cui gli elettroni sono i portatori di carica.

Tuttavia, il singolo MOSFET, al pari del transistor bipolare, per fornire due livelli logici richiede una resistenza di carico che, quando il transistor è in conduzione, deve dissipare una certa potenza.

Fanno eccezione alcune applicazioni, come le memorie NMOS, in cui il transistor funge da semplice interruttore pilotato. Per questa ragione, sebbene sia stata svantaggiosa, è stata creata la soluzione CMOS, ossia a MOSFET complementari, dove due transistor MOS di polarità opposta sono collegati in cascata e pilotati dallo stesso segnale logico, in modo che sia uno solo dei due a condurre per ciascuno stato logico.

La cella CMOS ha il grande vantaggio di consumare energia quasi esclusivamente durante l'istante di commutazione, a causa della natura capacitiva del gate, che può provocare, seppur per un tempo brevissimo, la conduzione simultanea dei due transistor durante il cambio di stato logico.

Di contro, occupa più spazio rispetto a una cella NMOS, che impiega un solo transistor, anche se va considerato lo spazio necessario sul chip per il resistore di pull-up: pertanto, in pratica, un dispositivo logico CMOS occupa poco meno del doppio di un equivalente NMOS.

I continui perfezionamenti tecnici hanno visto affermarsi la tecnologia CMOS in tutta l'elettronica digitale, tanto da soppiantare le logiche a transistor bipolari anche nei circuiti dove la velocità di commutazione era determinante.

La riduzione dei tempi di commutazione (o di propagazione, nel caso dei gate logici) è stata ottenuta sia diminuendo la capacità del gate dei MOSFET, sia accorciando la lunghezza del canale, grazie al continuo affinamento delle tecniche fotolitografiche. L'abbassamento della tensione di lavoro — reso necessario principalmente dalla differenza di potenziale minima tra gate e source per avviare la conduzione — è stato ottenuto utilizzando isolanti di gate più sottili e con una costante dielettrica maggiore, passando così dal biossido di silicio al nitruro di silicio.

Si noti che le dimensioni, la capacità parassita del gate e la tensione di lavoro sono fattori determinanti, soprattutto ai fini della dissipazione di potenza: a parità di corrente che circola nei MOSFET, quanto più bassa è la tensione di alimentazione, tanto minore è la potenza dissipata. Questo ragionamento non si applica tanto ai circuiti integrati elementari, quanto soprattutto alle CPU, che integrano anche centinaia di migliaia di transistor.

Proprio per migliorare tali caratteristiche, la tecnologia di produzione dei dispositivi MOS è stata via-via affinata, passando da quella originaria, ossia la planare, fino ad arrivare alle più avanzate tecnologie odierne, sospinta dalle nuove possibilità offerte dai procedimenti fotolitografici.

Ma anche questi hanno raggiunto i loro limiti, tanto che per integrare un numero sempre maggiore di

non tutte le foundry lo supportano), che consente di realizzare elementi attivi di dimensioni inferiori a un decimo di micron.

Ridurre le dimensioni e con esse la lunghezza del canale di un MOSFET significa non solo poter integrare un maggior numero di dispositivi elementari su un unico chip, ma anche consentire velocità di commutazione più elevate, ovvero, per le CPU, avere frequenze di clock più alte, a tutto vantaggio della rapidità del calcolo, che è un parametro determinante (insieme alla larghezza del bus dati e al numero di istruzioni per ciclo di clock) per la potenza di calcolo.

La maggior velocità di commutazione o propagazione si deve al fatto che più piccolo è il canale del MOSFET, minore è lo spazio che la corrente deve percorrere e siccome la corrente elettrica per convenzione viaggia alla velocità della luce, che è fissa (300.000 km/s) minore è la distanza da percorre, più contenuto è il tempo necessario a propagare un segnale.

Resta però il fatto che sotto una certa dimensione non è possibile scendere, almeno con la tecnologia planare, in quanto a un certo punto la distanza tra drain e source è tanto ridotta da creare dispersioni di corrente e quindi impedire lo spegnimento del transistor; lo stesso vale per lo strato di ossido di gate. Se il limite della risoluzione fotolitografica e della minima dimensione di un MOSFET sono già stati raggiunti, è vero che sono state escogitate altre architetture che permettono di aggirarlo; sono nati così dei MOSFET non più planari, ma sviluppati tridimensionalmente: i FinFET (Fin Field Effect Transistor).

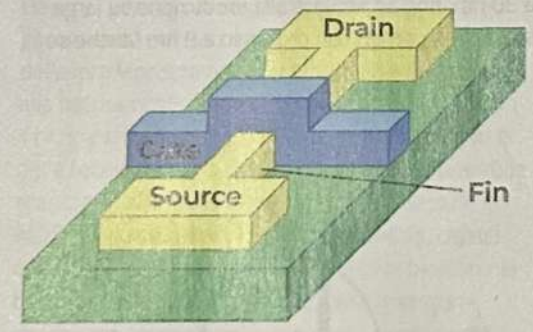


Fig. 3
Struttura del FinFET: il canale è sporgente e chiamato Fin perché ha la forma di una pinna.

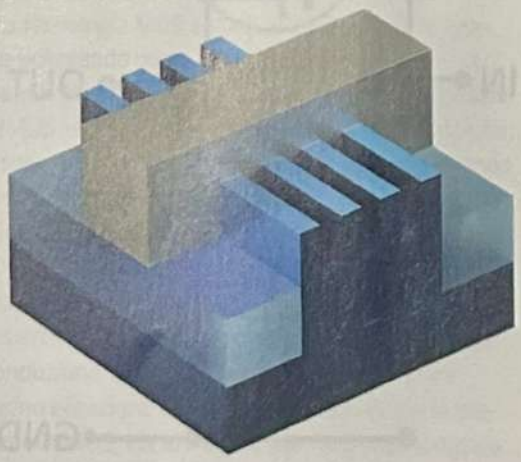


Fig. 4
Struttura di un FinFET a più canali (Fin).

FINFET

Qui la struttura cambia perché gli elettrodi dei transistor ad effetto di campo (gate, drain, source) vengono disposti non più su un piano ma su più dimensioni, in modo che il gate avvolga su tre lati il canale (che è sporgente rispetto alla tecnica planare) e sia più esteso, così da poter meglio controllare il campo elettrico; la caratteristica forma assunta dal canale dà il nome al Fin FET, ossia FET a pinna. La struttura fisica essenziale è quella proposta nella **Fig. 3**.

Questo FET tridimensionale è stato sviluppato dall'Università di Berkeley in California e la sua architettura massimizza l'accoppiamento elettrostatico tra gate e canale, ripristinando un controllo efficace anche a lunghezze infinitesime. Inoltre, viene eliminato il silicio in eccesso sotto il canale, fonte di ulteriore dispersione. La corrente fluisce parallelamente alla superficie del wafer all'interno della "pinna" e la struttura attiva si sviluppa in altezza.

Ciò permette al MOSFET realizzato con questa tecnologia, di commutare il 30% più velocemente di uno planare e di avere maggiore efficienza energetica, quindi dissipare meno energia per unità di superficie. Il maggior controllo da parte del gate permette di ridurre la tensione di alimentazione, dissipare meno calore e integrare un numero maggiore di transistor a parità di superficie.

RIBBONFET

La tecnologia RibbonFET rappresenta il cambiamento più significativo nell'architettura dei transistor dopo il transistor FinFET. L'architettura FinFET è stata perfezionata e ottimizzata negli ultimi 15 anni per migliorare le prestazioni e l'efficienza energetica, tuttavia con le geometrie odierne, anche il FinFET ha raggiunto i suoi limiti e non è più in grado di fornire ulteriori miglioramenti in termini di prestazioni o potenza. Il transistor RibbonFET migliora ulteriormente l'elettrostatica del FinFET, avvolgendo il gate del transistor attorno al canale, che in questa configurazione assume la forma di sottili nastri di silicio; quindi un singolo MOSFET

ha più canali in parallelo, ciascuno dei quali è un nastro "annegato" nel gate.
In virtù di ciò, si prevede che RibbonFET offrirà un eccezionale miglioramento dell'efficienza energetica rispetto all'attuale transistor FinFET.
Uno dei chip che utilizza la tecnologia RibbonFET è la serie di processori Intel Xeon (nome in codice Clearwater Forest) che sfrutta la tecnologia RibbonFET di seconda generazione di Intel (processo Intel 18A) per realizzare ciascuno dei chiplet della CPU di elaborazione primaria.

LE SFIDE DELLA MICROELETTRONICA

L'elettronica odierna vede dispositivi sempre più miniaturizzati ma comunque potenti sotto l'aspetto delle funzionalità implementate e, per quanto riguarda i microprocessori, relativamente alla capacità di calcolo; per le memorie, il problema è aumentare sempre più la densità di dati memorizzabili a parità di volume di semiconduttore.
Per affrontare queste sfide, l'industria si è mossa in più direzioni, la prima delle quali è stata cercare di miniaturizzare il più possibile i singoli dispositivi per integrarne il massimo possibile su un chip di semiconduttore; si definisce livello o **scala di integrazione**, proprio la densità di componenti e, per l'esattezza, di transistor o porte logiche. I primi integrati, realizzati circa quarant'anni fa, contenevano pochi dispositivi attivi; la tecnologia odierna consente la realizzazione di circuiti contenenti milioni di transistor.
Le scale di integrazione standardizzate sono:

- SSI (Small Scale of Integration); è la classe più semplice e conta alcune decine di componenti o porte logiche per chip;
- MSI (Medium Scale of Integration); vi appartengono i dispositivi logici complessi (contatori, shift-register, unità aritmetico-logiche) e conta qualche centinaio di componenti elettronici per chip;
- LSI (Large Scale of Integration); è quella di microprocessori, microcontrollori e memorie e conta migliaia di componenti per chip;
- VLSI (Very Large Scale of Integration); comprende i chip contenenti anche più di centomila componenti ed è usata nella realizzazione di memorie ad alta capacità e dei moderni microprocessori e dei microcontrollori più prestanti.

Questa classificazione permette di valutare l'evoluzione della tecnologia dei semiconduttori sotto l'aspetto della concentrazione di funzioni logiche elementari, il cui costante incremento negli anni ha consentito all'industria di sviluppare le CPU di ultima generazione.
Ora siamo tuttavia arrivati al limite fisico in fatto di numero di componenti attivi per chip, non a caso sono nati i dispositivi ripartiti in più chip, detti multi-chip. Questo è dovuto all'aumento del numero di core, all'aumento dei requisiti di I/O e connettività, nonché alla crescita del contenuto IP dell'acceleratore e ad altre funzionalità correlate.

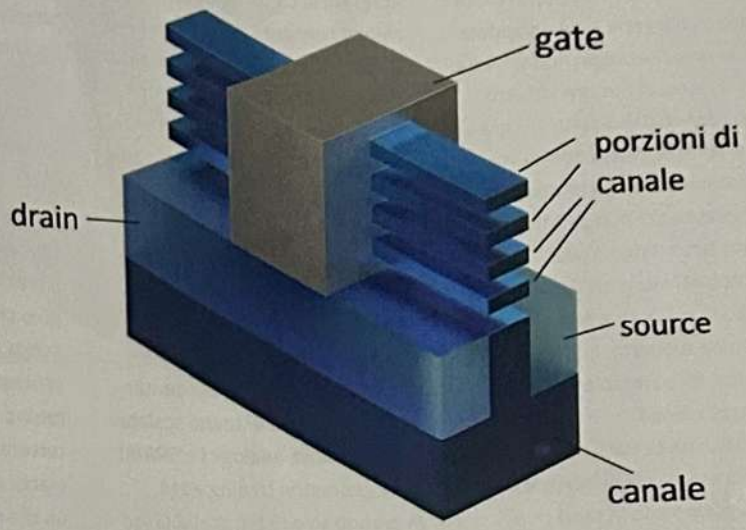


Fig. 5 Struttura caratteristica del RibbonFET.

CHIPLET E ARCHITETTURA MULTI-CHIP

Sempre più spesso, per affrontare i crescenti carichi di lavoro di elaborazione, si ricorre a CPU flessibili, in grado di scalare le prestazioni attraverso core più performanti o una maggiore densità di core. Un'implementazione all'avanguardia di CPU many-core (o multi-core, se si preferisce) richiede oggi un'area di silicio (span) superiore a quella consentita da un singolo reticolo litografico, pari a circa 800 mm^2.
Questo implica che tali architetture hardware non possono più essere realizzate su un unico chip o die e rende necessario adottare un'architettura disaggregata. In essa, le singole CPU — e, in alcuni casi, anche memorie come la cache di primo livello — sono collocate su chip separati, che devono essere opportunamente disposti e interconnessi all'interno di un package specifico, in grado di gestire sia le comunicazioni tra i chip, sia la distribuzione dell'alimentazione.
Entriamo quindi nel concetto di chiplet, intendendo con questo termine l'impiego di singoli chip omo-

DALL'INTEGRATO AL DISAGGREGATO

L'ultima frontiera della microelettronica, dopo il multi-core e il System on Chip, è la disaggregazione, che sostanzialmente significa smontare un chip in tante parti e collegarle assemblandole in un unico package; non si tratta, attenzione, di tornare indietro regredendo dal circuito integrato al componente discreto, ma di scomporre un'architettura complessa che non può trovare posto su un singolo pezzo di wafer, in tanti integrati distinti che possono poi essere combinati su un unico supporto, quasi fossero dei discreti assemblati in un integrato ibrido.
In pratica un circuito integrato a larghissima scala di integrazione viene scomposto in tanti chip (die) chiamati chiplet, che sono più facili da produrre rispetto a chip di grandi dimensioni comparabili con il limite del reticolo litografico. La tecnologia dei chiplet consente di disaggregare l'architettura dividendola in tanti chip uguali (per esempio i core di un microprocessore multi-core) o eterogenei, costituenti ad esempio i blocchi di un integrato molto complesso; ciò si traduce in una grande flessibilità, per i progettisti, nella composizione dell'architettura del sistema.
Ciò consente la disaggregazione per nodo di processo, consentendo di mantenere IP meno scalabili (ad esempio, analogici e SRAM) sulle geometrie trailing edge, migrando solo IP più scalabili (ad esempio, logica digitale) verso geometrie leading edge.
Tecnologie di packaging avanzate come Foveros Direct 3D ed EMIB 3.5D proposte dalla Intel, consentono di fare tutto questo e di creare integrati complessi contenenti tanti chiplet in un singolo package; ulteriori soluzioni Intel come la Foveros Direct 3D consentono inoltre la combinazione di chiplet provenienti da fonti diverse, fornendo ulteriore flessibilità. La dimensione di un singolo chiplet di elaborazione viene scelta per ottimizzare la resa del processo, consentendo al contempo la modularità nell'architettura del prodotto. I chiplet di elaborazione vengono impilati su un tile di base attivo utilizzando Foveros Direct 3D; ogni tile può ospitare IP di logica e memoria per il caching dei dati e il routing dall'I/O ai core e tra i core. Il tile di base può sfruttare progetti precedenti che utilizzavano un nodo di processo di generazione precedente per ridurre i costi di ricerca e sviluppo, fornendo al contempo funzionalità adeguate. I tile I/O possono anche riutilizzare gli investimenti di prodotti precedenti, accelerando i tempi di sviluppo (TAT) e offrendo un significativo vantaggio in termini di costi. Questi ingredienti possono essere combinati e abbinati in prodotti futuri man mano che sorgono esigenze di diversi IP di core del processore e/o funzionalità I/O.

genei o eterogenei combinati in un unico package per implementare architetture multi-chip (o multi-layer) che offrono numerosi vantaggi, tra cui minori costi e migliore efficienza di produzione.
L'architettura multi-chip (System in a Package, Fig. 6) comporta il superamento di diverse problematiche. A differenza di un System on Chip (SoC), in cui più CPU o dispositivi eterogenei sono interconnessi all'interno dello stesso chip (on-chip), in una soluzione multi-chip le CPU necessarie non possono essere realizzate su un singolo die, ma devono essere distribuite su più die — i cosiddetti chiplet — rendendo impossibile l'interconnessione diretta sullo stesso supporto di silicio.
Questo approccio richiede lo sviluppo di tecnologie di packaging avanzate, che prevedano combinazioni di soluzioni 2D, 2,5D e 3D, per massimizzare la larghezza di banda nella comunicazione die-to-die e ridurre al minimo le penalità di latenza.
Inoltre, introduce nuove sfide anche nella distribuzione dell'alimentazione elettrica ai singoli die.

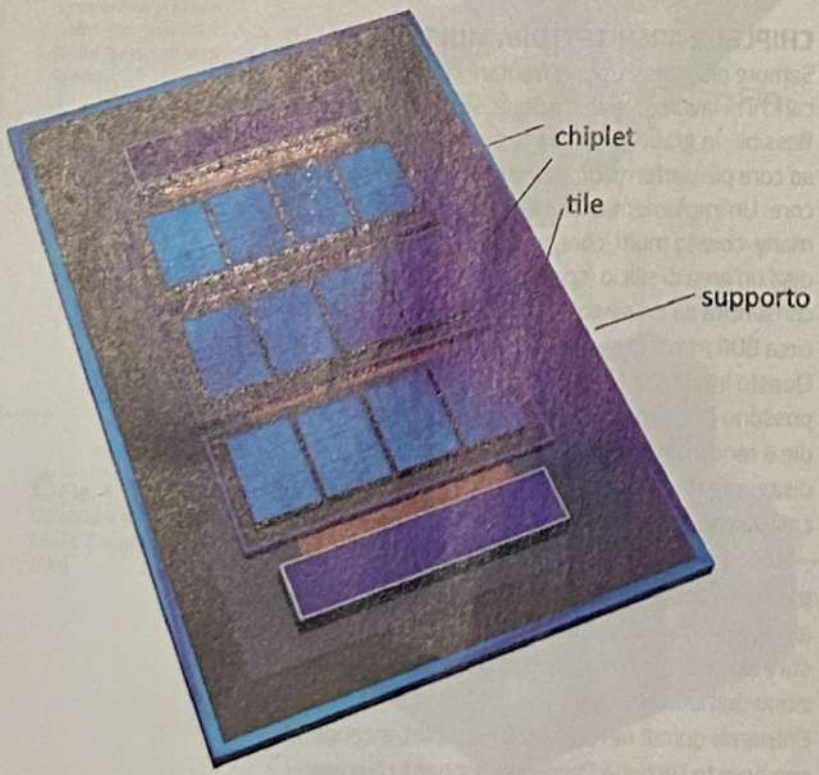
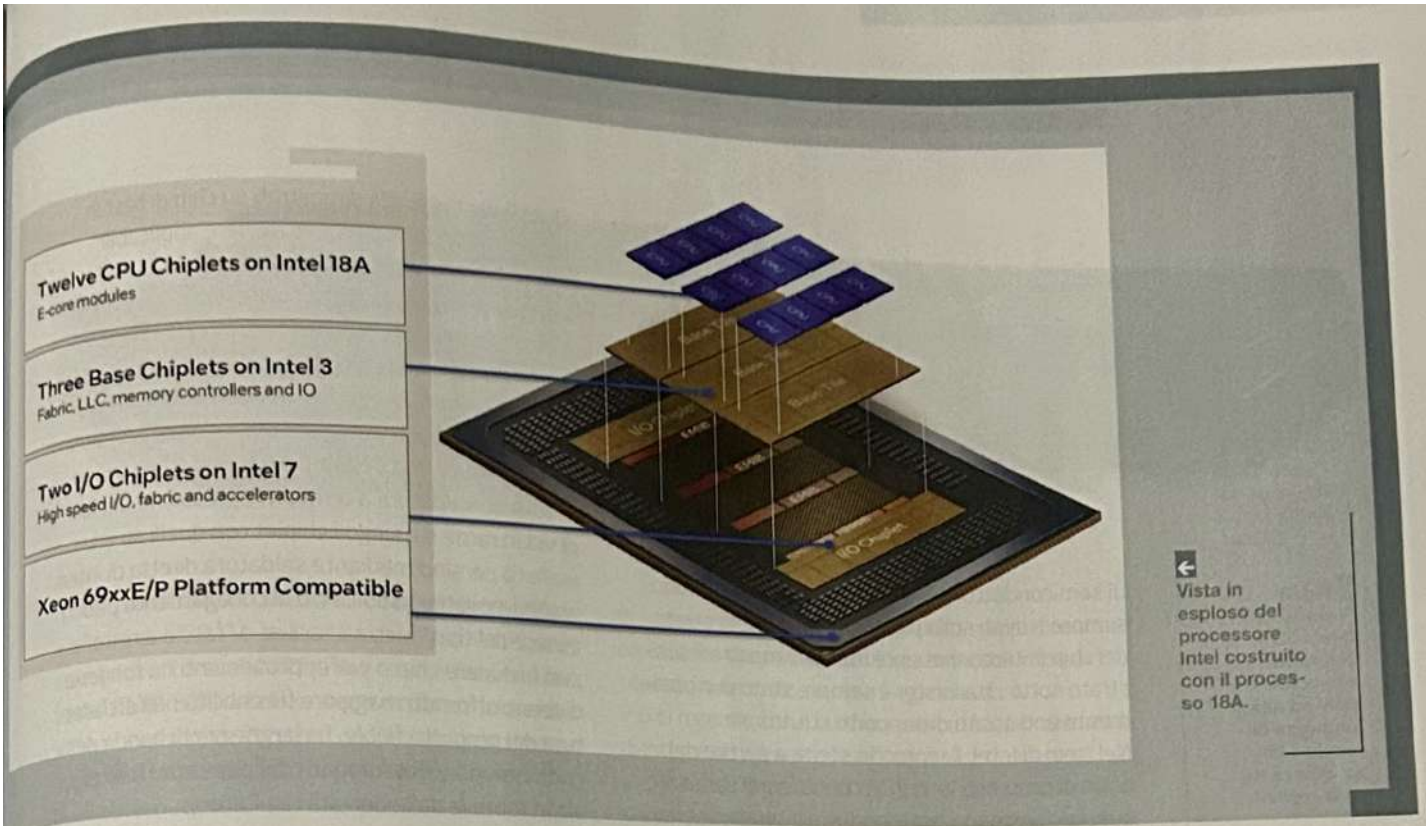


Fig. 6
Molteplici chiplet interconnessi attraverso una combinazione di tecniche di packaging 2D e 3D per dare vita ad un complesso System in a Package.

UN ESEMPIO DA INTEL: IL MULTIPROCESSORE XEON 6

Un modo per affrontare queste sfide è quello proposto da Intel, che trova applicazione nell'ultimo processore Intel Xeon (chiamato Clearwater Forest, che succederà al Sierra Forest) per server ad alte prestazioni, che è basato sulla tecnologia di processo Intel 18A (a 18 nm).
Qui scendono in campo soluzioni tecnologiche avanzate come i RibbonFET, le PowerVia, il Foveros Direct 3D (bonding ibrido per consentire l'impilamento diretto ad alta densità di chip attivi) l'Embedded Multi-die Interconnect Bridge (EMIB) e l'Intel Foundry FCBGA 2D+ (packaging multi-die ad alte prestazioni, economico e con un elevato numero di pin). Il Clearwater Forest è costruito attorno al nuovo core Darkmont, realizzato con il processo Intel 18A, e introduce un'architettura



Vista in esploso del processore Intel costruito con il processo 18A.

dove ogni cluster (modulo) integra quattro core e 4 MB di cache L2 unificata, con un throughput dati di 400 GB/s. Nella configurazione più avanzata, il processore combina 12 compute chiplet (per un totale di 288 core) realizzati con processo 18A, disposti sopra tre base tile prodotte a 3 nm e affiancate da due chiplet I/O su processo Intel 7. Complessivamente si tratta di 17 chiplet, collegati tramite Foveros Direct 3D ed EMIB, in grado di garantire elevata banda interna. I base tile ospitano cache LLC, memory controller e funzioni I/O, mentre i chiplet I/O aggiungono interfacce ad alta velocità, fabric e acceleratori dedicati.

PowerVia

Fin dalla costruzione del primo circuito integrato, ossia da quasi 50 anni, i microscopici fili metallici di bonding, ossia quelli che collegano i pad del chip

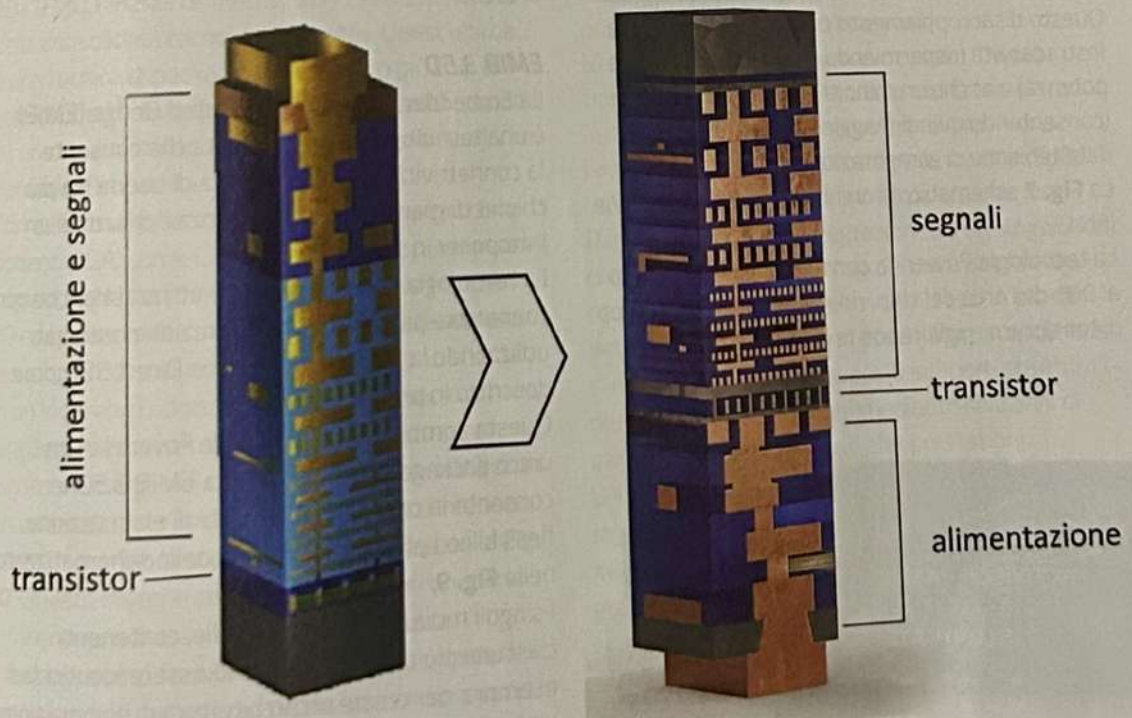


Fig. 7 La soluzione PowerVia introduce interconnessioni metalliche sotto lo strato del transistor, disaccoppiando per la prima volta l'instradamento del segnale e l'erogazione di potenza.

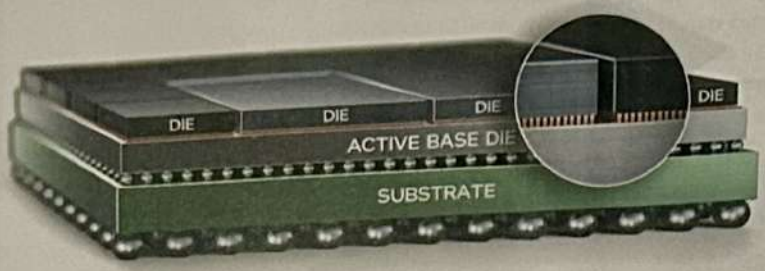


Fig. 8
Foveros Direct 3D consente interconnessioni ad alta larghezza di banda e bassa latenza tra chip impilati.

di semiconduttore ai pin del contenitore, si sono sempre trovati sulla parte superiore dello strato dei chip (interconnessioni frontali), mentre il substrato sotto i transistor è sempre stato principalmente uno strato di supporto strutturale.
Nel caso di Intel, l'approccio scelto a partire dal nodo di processo Intel 20A consiste nel cambiare questo paradigma per introdurre interconnessioni metalliche sotto lo strato sul quale vengono realizzati i transistor (interconnessioni posteriori).
Nel sistema tradizionale, l'architettura di interconnessione frontale era condivisa tra i fili per instradare i segnali elettrici tra i transistor e i fili per fornire alimentazione ai transistor.
Con l'introduzione della tecnologia PowerVia su Intel 20A, il routing del segnale e l'alimentazione verso i chip sono stati separati, consentendo di ottimizzare l'architettura di interconnessione frontale per il routing dei segnali scambiati tra i die, implementando appositamente una nuova architettura di interconnessione posteriore ottimizzata in modo indipendente per l'erogazione di potenza.
Questo disaccoppiamento consente una migliore instradabilità (risparmiando così spazio sul chip e potenza) e anche un minore droop della tensione (consentendo quindi maggiori prestazioni a una data tensione di alimentazione).
La **Fig. 7** schematizza l'architettura della PowerVia Intel.
La tecnologia PowerVia consente di sfruttare fino al 90% dell'area del chip, riducendo del 30% il droop di tensione e migliorando le prestazioni del 6%.

Fig. 9
La combinazione di EMIB e Foveros consente la creazione di sistemi flessibili ed eterogenei con una superficie totale di silicio significativamente maggiore all'interno di un singolo package.

Questi vantaggi, già dimostrati su chip di test in fase di sviluppo, dovrebbero essere applicabili anche a livello di produzione industriale.
Per quanto riguarda l'interconnessione dei chiplet all'interno dello stesso package, Intel ha messo a punto una tecnologia chiamata Foveros Direct 3D, che consente il collegamento diretto di uno o più chiplet a una base attiva per creare moduli di sistema complessi. Il collegamento "diretto" si ottiene mediante saldatura a termocompressione dei fori di via in rame sui singoli chiplet con quelli su un wafer o persino mediante saldatura diretta di interi wafer impilati uno sull'altro. Il collegamento può essere del tipo "faccia a faccia" o "faccia a retro" e può includere chip o wafer provenienti da fonderie diverse, offrendo maggiore flessibilità nell'architettura del prodotto finale. La larghezza di banda della connessione è determinata dal passo del foro di via in rame (e dalla densità risultante).
La prima generazione di Foveros Direct 3D utilizzerà la saldatura in rame con un passo di 9 µm, mentre la seconda generazione ridurrà il passo a soli 3 µm. L'unione tra chiplet e base attiva avviene mediante sfere di lega saldante simili a quelle dei componenti BGA, un po' più piccole di quelle applicate alla base del substrato del package per la saldatura su circuito stampato.
Questa unità di chiplet CPU posizionata su una grande cache locale, diventa un modulo di elaborazione completo, che può quindi essere replicato per aumentare la capacità di elaborazione e creare uno stack SKU basato sul numero di core e sui requisiti di cache.

EMIB 3.5D

La Embedded Multi-die Integrated Bridge (EMIB) è una tecnologia Intel collaudata che consente la connettività ad alta larghezza di banda tra più chiplet di grandi dimensioni senza l'utilizzo di un interposer in silicio.
La tecnologia EMIB può essere utilizzata anche per connettere più moduli di elaborazione realizzati utilizzando la tecnologia Foveros Direct 3D, come descritto in precedenza.
Questa combinazione di EMIB e Foveros in un unico package è stata chiamata EMIB 3.5D e consente la creazione di sistemi di elaborazione flessibili ed eterogenei, come quello schematizzato nella **Fig. 9**.
I singoli moduli o tile (mattonelle, contenenti ciascuna più chiplet...) possono essere identici (ad esempio, per creare un'architettura di elaborazione scalabile) oppure diversi tra loro (ad esempio, per

Fig. 10 Architettura Foundry FCBGA 2D+: Una "patch" ad alta densità è inserita tra un chip attivo (in alto) e un interposer simile a un PCB (in basso).

connettere moduli di elaborazione con tile di I/O o con moduli DRAM) nell'ottica di un'architettura eterogenea. La scalabilità e la flessibilità consentite dall'EMIB 3.5D Intel consentono la creazione di sistemi in package (System on Package) con una superficie totale in silicio di gran lunga superiore a quella ottenuta con i soli interposer in silicio.
La tecnologia Intel EMIB di seconda generazione (bump pitch scalato da 55 micron a 45 micron) in sviluppo, permetterà ottenere una connettività ad alta larghezza di banda con chiplet Foveros Direct 3D o chiplet I/O multipli.

Intel Foundry FCBGA 2D+: il package è tutto

Oltre alle ampie possibilità offerte dal packaging 3D avanzato, Intel dispone anche di architetture e tecniche di progettazione specifiche per fornire soluzioni di packaging ottimizzate in termini di costi di produzione.
Una di queste è l'architettura Intel Foundry FCBGA 2D+ (Flip Chip Ball Grid Array 2D+), evoluzione della consolidata tecnologia flip chip. Quest'ultima è una tecnica di packaging in cui uno o più chip di semiconduttore vengono montati capovolti e saldati su un circuito stampato multistrato (multilayer), che riproduce l'impronta del dispositivo.
Con questo approccio si realizzano numerosi componenti BGA, come CPU, GPU e chipset per personal computer. Nell'architettura Intel Foundry FCBGA 2D+, le funzionalità più sofisticate (e costose) della tecnologia del substrato organico vengono sfruttate in un ingombro ridotto (un substrato "patch" ad alta densità) e assemblate su un interposer (con ingombro maggiore) che sfrutta le funzionalità di un "circuito stampato" o di un PCB a un costo inferiore; il tutto, secondo la struttura esemplificata nella **Fig. 10**. Questo insieme composito (package-on-package) viene quindi assemblato su una scheda.
I vantaggi complessivi in termini di riduzione dei costi derivanti dall'utilizzo di tale architettura per i processori Intel Xeon possono facilmente raggiungere centinaia di milioni di dollari. Intel implementa con successo questa tecnologia nella sua linea di prodotti Intel Xeon da diverse generazioni.
Più di recente, con il continuo aumento delle velocità di interconnessione e la difficoltà nel superare le implicazioni di perdita di margine (discontinuità nel percorso elettrico) dovute ai margini elettrici, sono stati sviluppati progressi nei materiali e tecniche di progettazione che possono contribuire a raggiungere velocità simili a quelle di PCIe Gen6, DDR5 e MR DIMM.

CONCLUSIONI

Le crescenti esigenze in termini di potenza computazionale e densità di calcolo hanno spinto la microelettronica oltre limiti un tempo ritenuti insuperabili, passando dalle tecnologie di realizzazione dei transistor planari e planari-epitassiali alle più avanzate architetture tridimensionali, fino all'introduzione dei chiplet e dei moduli multi-chip, che oggi rappresentano lo stato dell'arte nella progettazione dei circuiti integrati.
Nel caso specifico dei microprocessori, l'evoluzione ha riguardato sia il calcolo multicore — con la comparsa di più CPU su un unico chip — sia l'integrazione di sistemi eterogenei su chip (SoC), capaci di gestire funzioni molto diverse tra loro.
Raggiunto il limite di integrazione su un singolo chip imposto dalla fotolitografia, si è passati alla scomposizione di dispositivi complessi in più die, detti chiplet, che vengono assemblati, interconnessi e alimentati tramite appositi supporti strutturali, per poi essere incapsulati mediante soluzioni di packaging innovative e ad alte prestazioni.
Un esempio concreto di tecnologia multi-chip è rappresentato dal recentissimo processore Intel Xeon Clearwater Forest, già analizzato nei paragrafi precedenti, che mostra chiaramente come si realizza un'implementazione disaggregata e scalabile di dispositivi di calcolo a semiconduttore complessi. □