Direcţii în evoluţia arhitecturii microprocesoarelor

Mihai Budiu — mailto:mihaib+@cs.cmu.edu
http://www.cs.cmu.edu/~mihaib/

februarie 2000

Subiect: evoluţia microprocesoarelor: trecut, prezent şi viitor

Cunoştinţe necesare: cunoştinţe elementare despre arhitectura calculatoarelor

Cuvinte cheie: microprocesor, SPEC, ILP

Cuprins

Am întîlnit de mai multe ori un banc pe Internet, care spunea că, dacă maşinile ar fi evoluat în aceeaşi măsură cu calculatoarele, acum ar fi mers 120 de kilometri cu benzina dintr-o brichetă, şi ar fi costat cît o pîine. (Un contra-banc, din partea industriei automobilistice, ofensate, zicea apoi că, dacă ar fi evoluat la fel, maşina ar fi refuzat să mai meargă de cîteva ori pe zi, şi ar fi trebuit să o duci înapoi în garaj ca să repornească.)

Adevărul este că progresele făcute de tehnologia calculatoarelor sunt absolut uluitoare; ajunge să îţi cumperi un calculator nou după doi ani ca să fii impresionat de cîştigul de performanţă înregistrat. Cîştigul de performanţă se datorează unor procesoare din ce în ce mai sofisticate şi mai rapide, şi unor memorii de capacităţi din ce în ce mai mari.

Răspunzătoare pentru creşterea exponenţială a performanţei sunt însă în cea mai mare măsură microprocesoarele. În acest articol vom arunca o privire asupra evoluţiei microprocesoarelor de la origini pînă în prezent. Vom încerca apoi să caracterizăm arhitecturile procesoarelor contemporane, şi să extrapolăm din datele la dispoziţie, speculînd despre unele din posibilele evoluţii viitoare.

Trebuie să atrag de la început atenţia că nu am însuşiri paranormale, şi ca atare nu sunt profet. Domeniul tehnologiilor de calcul este extraordinar de volatil, şi se mişcă cu o viteză fantastică; orice previziune este cel puţin hazardată. De altfel în figura 1 ilustrează acest fapt contrapunînd previziunile din urmă cu cîţiva ani ale unei organizaţii extrem de prestigioase, Semiconductor Industry Association (SIA, `http://www.semichips.org/`), cu realitatea. Deci nu vă aşteptaţi de la mine la mai mult.

Am mai publicat în PC Report o serie întreagă de articole despre arhitectura procesoarelor moderne, pe care le voie cita ocazional; toate acestea sunt disponibile din pagina mea de web. Articolul de faţă va fi însă mai superficial. Există o cantitate enormă de informaţie pe web. În acest articol am folosit în mod repetat informaţii de la `http://bwrc.eecs.berkeley.edu/CIC/`, CPU Info Center.

1 Evoluţia microprocesoarelor

1.1 Industria semiconductoarelor: un fenomen economic

Un istoric interesant al diferitelor idei arhitecturale din microprocesoare puteţi găsi pe web la `http://bwrc.eecs.berkeley.edu/CIC/archive/cpu_history.html`.

Primul microprocesor a fost creat de firma Intel în 1971. Numele său era Intel 4004, şi era un procesor pe 4 biţi. Apariţia primului micorprocesor a fost un pas cu uriaşe consecinţe în evoluţia ulterioară a sistemelor de calcul. Diferenţa între microprocesor şi metodele îndeobşte folosite era că procesorul strînge pe o singură pilulă de siliciu toate unităţile funcţionale importante necesare executării programelor; fiind toate strîns integrate, comunicaţia între ele este rapidă şi eficace, permiţînd dintr-o dată un salt calitativ.

Nu mai puţin importantă este reducerea de cost care urmează unei astfel de integrări. Cu siguranţă că principalul motiv al evoluţiei explozive a tehnologiei circuitelor integrate nu este de natură tehnologică, ci economică: spirala preţurilor din ce în ce mai scăzute face echipamentele de calcul din ce în ce mai accesibile, cererea creşte, ducînd la venituri mai ridicate pentru fabricanţi, care investesc mai mult în cercetare/dezvoltare şi linii tehnologice, obţinînd densităţi mai mari, permiţînd integrarea mai multor circuite, şi costuri şi mai joase. Cu toată scăderea de preţ, veniturile globale ale industriei semiconductoarelor au crescut în mod galopant: numai anul trecut vînzările globale au fost de 149 de miliarde de dolari!

Esenţial pentru a menţine această spirală este faptul că echipamentele de calcul măresc enorm productivitatea muncii, direct sau indirect: de aici cererea crescîndă. Iar experţii afirmă că acesta este doar începutul, şi că în viitor fiecare individ va depinde de zeci de dispozitive de calcul în fiecare clipă. Nu suntem prea departe de acest punct: chiar în ziua de azi, o maşină modernă are în medie 15 microprocesoare, care controlează, reglează şi diagnostichează tot felul de parametri, de la injecţie pînă la frîne.

Nu pot să mă abţin să remarc că Statele Unite ale Americii atribuie o treime din creşterea venitului naţional brut în anul trecut doar tehnologiilor informaţionale, care însă ocupă doar 8% din forţa de muncă. În foarte mare măsură, tehnologia informaţiei este responsabilă pentru fenomenala dezvoltare economică pe care Statele Unite o traversează în aceşti ani.

1.2 Evoluţii cantitative

Să lăsăm acum deoparte economia, şi să aruncăm o privire asupra evoluţiei unor parametri ai procesoarelor de-a lungul timpului. Tabela 1 arată evoluţia generaţiilor succesive ale celei mai proeminente familii de procesoare, ale firmei Intel:

An	Procesor	Ceas (Mhz)	Tranzistori (mii)	Tehnologie (μm)
1971	4004	0,74	2,250	
1972	8008	0,8	3,300	
1974	8080	3,1	4,500	
1978	8086	10	29	2,0
1979	8088	8	29	2,0
1980	80186			1,5
1982	80286	12	134	1,5
1985	80386	33	275	1,0
1989	80486	100	1 600	0,6
1993	Pentium	266	4 500	0,35
1996	PPro	200	5 500	0,35
1997	PII	450	7 500	0,25
1999	PIII	733	28 000	0,18
2000	Merced	800		0,18

Tabelul 1: Evoluţia procesoarelor Intel. Spaţiile albe indică lipsa informaţiei. Merced va fi lansat oficial la cîteva zile de la scrierea acestui articol (pe 17 februarie 2000), deci datele sunt incerte.

Ultima coloană din tabela 1 şi figura 1 arată care este impactul miniaturizării: această coloană indică dimensiunea de bază (feature size), care poate fi văzută ca fiind dimensiunea unui tranzistor. Orice reducere a acestei valori are un impact cuadratic, pentru că suprafaţa creşte cu pătratul laturii. O reducere de la 2 microni la 1,5 (50%) măreşte deci suprafaţa efectivă cu 77% (4/2,25 = 1,77).

Din fericire, reducerea dimensiunilor mai are încă o consecinţă foarte importantă: traseele pe care trebuie să le parcurgă curentul electric între dispozitive devin mai scurte, deci se pot parcurge

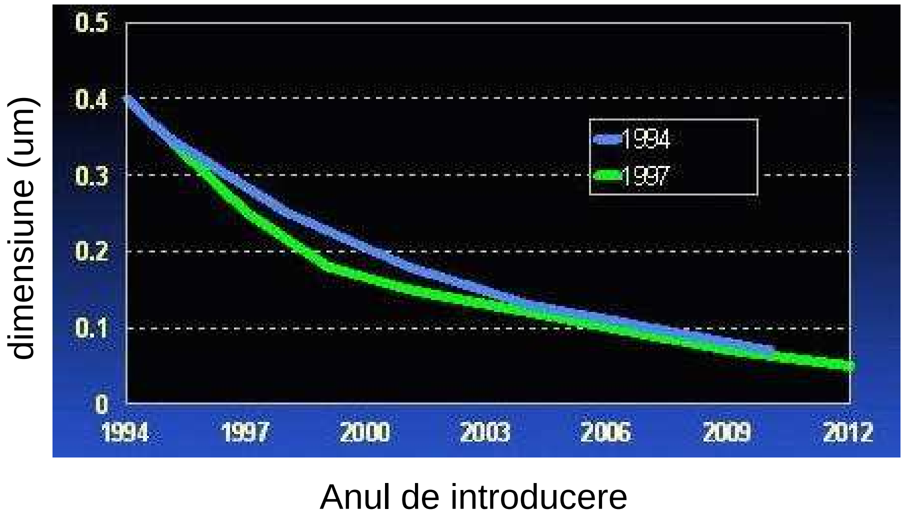

Figura 1: Pronosticurile făcute în 1994 şi 1997 de Semiconductor Industry Association privind miniaturizarea circuitelor integrate digitale. Estimările din 1994 au fost depăşite de realitate, forţînd SIA să re-estimeze evoluţia pentru viitor.

mai rapid. Proiectanţii pot face deci procesorul să funcţioneze cu un ceas mai rapid[1]. Evoluţia frecvenţei ceasurilor se vede în coloana a treia. La ora apariţiei acestui articol procesoarele pentru PCuri la frecvenţe de 1 gigahertz vor fi o realitate foarte apropiată[2].

Faptul că avem siliciu la dispoziţie pentru a implementa mai mulţi tranzistori înseamnă că:

1. Putem muta mai multe circuite auxiliare pe acelaşi cip.

   Evoluţia procesoarelor cunoaşte cîteva salturi calitative: cînd miniaturizarea făcea posibilă integrarea unui nou dispozitiv pe acelaşi circuit integrat, se realiza un salt de performanţă. Astfel, au fost integrate succesiv: unităţi din ce în ce mai mari de procesare (8, 16, 32, acum 64 de biţi), coprocesoare aritmetice, unităţi de management al memoriei, cacheuri de nivel 1 şi chiar 2;

2. Designerii folosesc tranzistorii suplimentari pentru a construi circuite mai sofisticate, care pot executa mai repede şi mai eficient programele. Metoda fundamentală folosită este de a face mai multe lucruri în paralel.

Împreună aceste trei fenomene (viteza ceasului, integrarea pe o singură pastilă şi exploatarea paralelismului) contribuie la creşterea performanţei totale a procesoarelor. Aşa cum am povestit şi cu alte ocazii, măsurarea performanţei unui calculator se face evaluînd sistemul pe mai multe programe (deci performanţa depinde foarte mult şi de compilatorul folosit), care de obicei fac parte din suite de teste standardizate (*benchmark suites*). Cele mai folosite pentru a evalua procesoare sunt

[1] Majoritatea covîrşitoare a procesoarelor contemporane funcţionează în mod *sincron*: întreaga lor funcţionare este orchestrată de un tact de ceas, care garantează că feluritele părţi sunt sincronizate. Din ce în ce mai mult însă se tinde spre scheme cu multiple semnale de ceas, sau chiar scheme asincrone. Nu ne vom ocupa însă de aceste evoluţii în textul acestui articol.

[2] Compaq a anunţat procesoare Alpha 20364 la 1Ghz din 1999, iar AMD a demonstrat deja un procesor K7 fără răcire specială la 1.1Ghz.

cele din seria SPEC: Standard Performance Evaluation Corporation (`http://www.specbench.org`). Nu ne vor interesa acum prea tare ce reprezintă numerele acestea; cert e că cu cît sunt mai mari, cu atît e mai bine. Graficul din figura 2 arată evoluţia performanţei procesoarelor în ultimii 10 ani, în termeni SPEC.

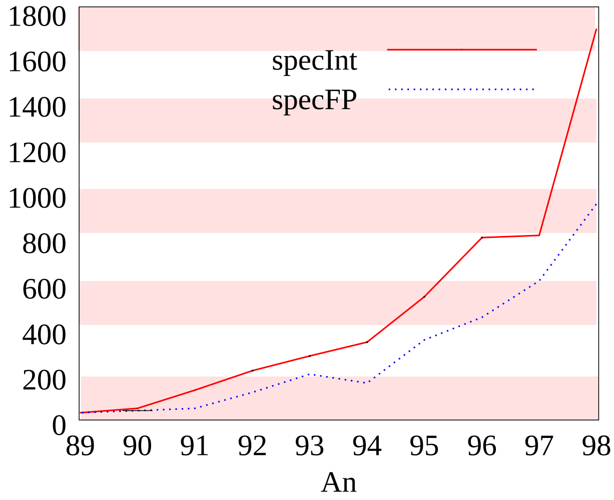

Figura 2: Evoluţia performanţei procesoarelor pe suita de teste SPEC. SpecINT măsoară performanţa programelor de uz general (care folosesc preponderent numere întregi), iar SpecFP pe cea a programelor de calcul ştiinţific (care folosesc mai ales numere în virgulă flotantă).

Evoluţia urmăreşte aproximativ o curbă exponenţială: în fiecare an performanţa creşte cu 60%.

1.3 Tehnologii arhitecturale

Aşa cum am ilustrat în seria mea de articole intitulate "arhitectura avansată a procesoarelor", o mulţime de inovaţii tehnologice au fost introduse una după alta în arhitecturi; de fapt intenţionez să continui această serie şi în viitor, pentru că mai sunt de prezentate şi alte mecanisme importante.

Îmi permit să expun în continuare viziunea profesorului John Hennessy, de la universitatea Stanford, aşa cum a expus-o în prelegerea pe care a dat-o ca invitat la Federated Computer Research Conferences în mai 1999.

Hennessy vede două tehnologii arhitecturale ca fiind esenţiale: exploatarea paralelismului la nivel de instrucţiune (Instruction Level Parallelism, ILP) şi ierarhii sofisticate de memorie (cache-uri). Să spunem cîteva cuvinte despre fiecare:

1.3.1 ILP

Paralelismul la nivel de instrucţiune constă în independenţa instrucţiunilor din programe una de alta, ceea ce ne permite să executăm mai multe instrucţiuni simultan. Am vorbit altădată pe larg despre paralelismul la nivel de instrucţiune; să observăm că toate procesoarele contemporane îl exploatează prin două forme:

- Execuţia pe bandă de asamblare (*pipeline*) a instrucţiunilor succesive;

- Execuţia în paralel a instrucţiunilor independente: procesoarele de tip VLIW (very long instruction word) aleg la compilare care instrucţiuni merg în paralel, iar procesoarele superscalare fac această alegere în timpul execuţiei.

Astfel, în 1985 au apărut primele procesoare cu banda de asamblare, în 1990 primele procesoare de tip VLIW, iar în 1995 procesoare foarte sofisticate superscalare, care pot executa instrucţiunile în ordini foarte diferite de cea din program (*out-of-order execution*).

1.3.2 Cache-uri

Am scris în repetate rînduri despre cache-uri în PC Report (vedeţi de pildă PC Report din martie 1997 şi noiembrie 1998). Aici vom arunca doar o privire superficială asupra lor; scopul nostru este de a înţelege de ce cache-urile joacă un rol fundamental în creşterea performanţei. Figura 3 ne oferă cheia: deşi atît procesoarele cît şi memoriile cresc constant în viteză, creşterea procesoarelor este cu 50% mai rapidă decît a memoriilor. Ca atare există o disparitate crescîndă între nevoile de date (şi instrucţiuni) ale procesorului şi ceea ce memoriile pot oferi. Durata unui acces la memorie ajunge la zeci de cicli de ceas pentru procesoarele contemporane. Întîrzierea accesului este şi mai exacerbată în cazul sistemelor care au mai multe procesoare, în care caz timpii de acces la date pot ajunge la mii de cicli.

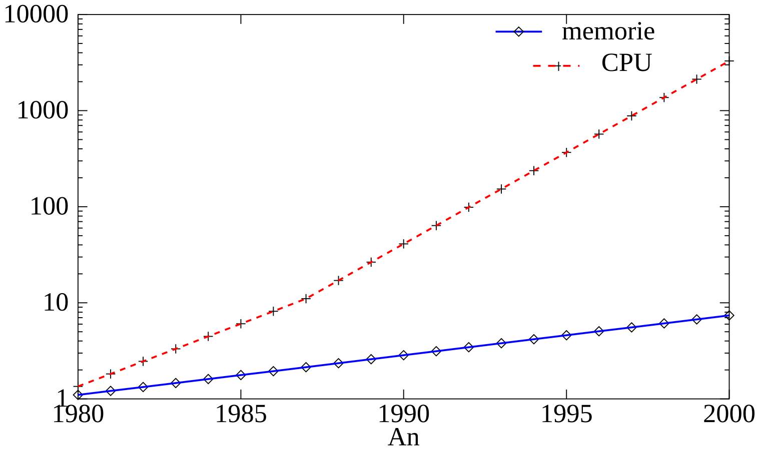

Figura 3: Creşterea disparităţii de viteză între memorii şi procesoare. Ambele valori pleacă de la valoarea normalizată 1, dar procesoarele cresc în performanţă cu 35% pe an pînă în 1987, şi cu 55% pe an după aceea, în timp ce memoriile cresc cu 10% pe an.

Din această cauză se construiesc cache-uri, care sunt memorii mai mici şi mai rapide, care se plasează între procesor şi memoria principală, şi în care sunt aduse datele pentru prelucrare. Proiectanţii au reuşit să sporească eficacitatea cache-urilor folosind două metode:

- Prin folosirea unor cache-uri din ce în ce mai mari, plasate din ce în ce mai aproape de procesor. Această evoluţie este clar vizibilă:

    1. primele procesoare nu aveau nici un fel de cache, pentru că memoriile erau suficient de rapide pentru a le servi cu date.
    2. În 1980 au apărut cache-uri (L1) sub forma unor circuite speciale, care

3. în 1984 au fost integrate pe aceeaşi pilulă de siliciu cu procesorul central,
4. după care (1986) a apărut un al doilea nivel de cache (L2), mai mare şi ceva mai lent, care
5. în procesoarele moderne (1995) este la rîndul lui adesea integrat cu circuitul microprocesorului, pentru a permite un acces rapid.
6. Au apărut nivele terţiare de cache (1999).

- Pe de altă parte metodele de management ale cache-urilor sunt din ce în ce mai sofisticate:

  1. au apărut cache-uri care servesc procesorul de îndată ce primul cuvînt a sosit, chiar dacă restul sunt pe drum (*early restart*, 1992),
  2. cache-uri care nu blochează procesorul cînd datele lipsesc, ci îi permit să continue execuţia (*non-blocking*, 1994),
  3. şi tot felul de alte tehnologii sofisticate, pe care le-am expus în alte părţi (cache-uri victimă, buffere de scriere, instrucţiuni speciale de management al cache-ului (prefetching), etc.);
  4. Tot aici se cuvine să menţionăm multiprocesoarele simetrice şi protocoalele de coerenţă ale cache-urilor pentru astfel de sisteme; toate procesoarele moderne sunt construite pentru a fi folosite în sisteme multi-procesor, şi includ astfel de dispozitive.

2 Arhitecturile contemporane

Astfel, urmărind evoluţia arhitecturilor, am ajuns pînă în ziua de azi. Vom încerca să caracterizăm sumar starea arhitecturilor, după care vom arunca o privire asupra unora din direcţiile viitoare.

2.1 Hardware şi software

Istoria modernă a procesoarelor contrapune două paradigme pentru creşterea performanţei, bazate pe software şi respectiv pe hardware. Aparent un articol despre arhitectura procesoarelor nu are nimic de-a face cu software-ul. Nimic mai greşit: la ora aceasta există o simbioză totală între hardware şi software; procesoarele se proiectează odată cu compilatoarele care le folosesc, şi relaţia dintre ele este foarte strînsă: compilatorul trebuie să genereze cod care să exploateze caracteristicile arhitecturale, altfel codul generat va fi foarte ineficace.

Metodele de creştere a performanţei cu ajutorul compilatoarelor se numesc şi *statice*, pentru că programul este analizat şi optimizat o singură dată, înainte de a fi pornit în execuţie. Metodele bazate pe hardware se numesc *dinamice*, pentru că sunt aplicate în timp ce programul se execută.

Istoria arhitecturilor contrapune mereu cele două paradigme: de exemplu dezbaterea iniţială RISC/CISC[3] era de aceeaşi natură, ca şi dezbaterea între superscalar şi VLIW, pe care am menţionat-o deja în acest text.

De fapt, aşa cum menţionam şi în alte articole (de exemplu în PC Report din iunie 1999), există lucruri care se pot face numai static, şi există lucruri care se pot face numai dinamic. Aşa

[3] În anii '80 a apărut ideea de a face procesoarele mult mai simple pentru a le permite să meargă mai repede. Astfel de arhitecturi au fost numite RISC: Reduced Instruction Set Computer, prin contrast cu celelalte, Complex ISC.

că de fapt, chiar arhitecturile care pornesc la una din extreme, tind să conveargă către folosirea unui amestec de trăsături din ambele domenii:

- la ora actuală distincţia RISC/CISC aproape că s-a estompat:
  - Pentium, un procesor tipic CISC, de fapt traduce în mod automat instrucţiunile în instrucţiuni de tip RISC în hardware, după care le execută;
  - toate procesoarele RISC au căpătat extensii la setul de instrucţiuni (gen CISC) pentru a le mări eficacitatea; de pildă toate procesoarele au extensii speciale pentru multimedia.
- De asemenea, graniţele dintre super-scalar şi VLIW tind să se estompeze, fiecare împrumutînd din tehnologiile celuilalt.

Cu siguranţă că un model mixt este preferabil, pentru că poate lua ce e mai bun din fiecare tehnologie.

2.2 Crusoe

Se cuvine să atragem atenţia asupra unei recrudescenţe a "luptei" sistemelor pure: anul acesta compania Transmeta a anunţat apariţia unui nou procesor, numit Crusoe, care exploatează la maximum tehnologiile statice (compilarea). Compania Transmeta a făcut mare vîlvă, nu atît prin procesorul lor, care poate simula alte procesoare, inclusiv cele ale firmei Intel, ci prin faptul că angajează pe cel mai faimos programator al planetei, Linus Torvalds, creatorul sistemului de operare Linux.

Transmeta a lansat Crusoe cu mare pompă în luna ianuarie; compania predică întoarcerea la simplitate (care a fost sugerată atît de curentul RISC, cît şi de modelele VLIW), în care hardware-ul este simplu şi rapid, iar compilatorul face greul. Echipa care a lucrat la Transmeta este compusă din mare măsură din ingineri plecaţi de la IBM: IBM a lucrat la o versiune de procesor PowerPC care putea face exact acelaşi lucru: putea executa în mod nativ cod x86 (adică compatibil Intel), dar proiectul lor a fost întrerupt deşi era într-o stare foarte avansată, aparent din motive de marketing.

Cît de serios este acest nou competitor?

Din păcate atuurile lui Crusoe nu sunt prea clare:

- cipul nu are un ceas mai rapid decît procesoarele Intel (versiunile de Crusoe disponibile acum merg doar la 400Mhz, comparativ cu Pentium, care ajunge la 800);
- cipul într-adevăr consumă mult mai puţină energie, şi are nevoie de mult mai puţină răcire. Transmeta afirmă că asta-l face ideal pentru laptop-uri. Din păcate, principalul consumator de energie într-un laptop nu este procesorul, ci ecranul şi discul, aşa încît avantajele noului cip vor fi marginale;
- Crusoe se bucură de compatibilitate cu setul de instrucţiuni x86; dar pentru platformele pe care x86 este dominant (desktop, laptop, chiar şi server) am văzut că performanţa lui este insuficientă. Dacă Crusoe vrea să concureze pentru celelalte pieţe, de procesoare integrate (embedded computing), atunci are de-a face cu alţi competitori formidabili, ca procesoarele de semnal de la Motorola, Texas Instruments şi Intel (ARM), asupra cărora nu este clar cîte avantaje are.

Probabil că pentru a rămîne viabil Crusoe va trebui să se metamorfozeze şi să devină mai complicat, folosind şi o serie de mecanisme dinamice de creştere a performanţei.

În definitiv există o singură resursă aproape gratuită şi care este în cantităţi suficiente: numărul de tranzistori. Datorită miniaturizării numărul de tranzistori disponibili pentru design creşte enorm; de aceea simplitatea cu orice preţ (aşa cum o încarnează Crusoe) nu este neapărat o calitate.

3 Constrîngerile fizice

Proiectanţii de microprocesoare se lovesc în ziua de azi de mai multe dificultăţi. Nici una dintre ele nu e insurmontabilă, dar soluţiile sunt din ce în ce mai grele. Vom arunca o privire asupra unora dintre ele; încercarea de a extrapola impactul acestor bariere în viitor va sugera apoi soluţii pentru depăşirea lor.

Accesul la memorie. După cum am văzut în figura 3, în ultimii 10 ani viteza memoriilor a crescut cu 10% pe an, în timp ce viteza procesoarelor a crescut cu o rata de 60%. Toate motivele ne îndeamnă să credem că această disparitate va continua să se accentueze, şi că preţul relativ al unui acces la memorie (măsurat în cicli de ceas) va continua să crească.

Putere. Un alt factor care limitează evoluţia circuitelor integrate este consumul de putere; în urmă cu 15 ani un procesor consuma 2 waţi; astăzi un procesor ca Alpha 80364 consumă 100W; de aici rezultă limitări pentru ceas (puterea consumată creşte cu frecvenţa ceasului), şi necesitatea unor dispozitive speciale de răcire.

Din fericire tehnologia lucrează în direcţia favorabilă: miniaturizarea duce la scăderea puterii necesare. Un alt factor care duce la scăderea puterii consumate este scăderea tensiunilor de alimentare.

Interesant este faptul că, deşi dimensiunea tranzistorilor a scăzut într-una, dimensiunile circuitelor fabricate au crescut: foamea neostoită a designerilor cere suprafeţe din ce în ce mai mari ale plăcuţelor de siliciu; de aceea puterea consumată a crescut şi ea.

Complexitate. Un factor deloc neglijabil este complexitatea enormă a circuitelor. Procesoarele cele mai moderne au peste 25 de milioane de tranzistoare, iar în cîţiva ani designerii vor avea la dispoziţie un miliard. Astfel de circuite sunt foarte greu de verificat şi testat. La ora actuală o companie ca Intel cheltuieşte 40% din budget pentru proiectare şi dezvoltare, şi 60% pentru verificare şi testare!

O altă problemă importantă este legată de liniile tehnologice de fabricaţie: o astfel de instalaţie costă la ora actuală două miliarde de dolari. Puţine companii îşi pot permite investiţii de asemenea anvergură pentru o tehnologie care se schimbă în 3 ani!

Sîrmele. E clar că miniaturizarea nu va putea continua în acelaşi ritm exponenţial: peste ceva vreme am ajunge la necesitatea de a face un tranzistor mai mic decît un atom, ceea ce e evident imposibil. Dar chiar înainte de a atinge un astfel de prag, vom avea alte probleme de înfruntat.

O analiză extrem de interesantă a fost făcută de Mark Horowitz, profesor la universitatea Stanford, într-un articol intitulat "Viitorul sîrmelor". Articolul porneşte de la caracteristicile electrice ale semiconductorilor, şi analizează o serie de scenarii posibile pentru tehnologiile de fabricaţie. Textul ia în considerare tot felul de factori, cum ar fi geometria sîrmelor, capacităţi şi rezistenţe, disiparea puterii, etc. Vom ignora toate aceste detalii, însă vom privi una din concluziile la care autorul ajunge.

Autorul observă că în general sîrmele vor evolua în sensul dorit: vor deveni mai scurte, iar viteza de transmisiune a informaţiei nu va încetini relativ la dimensiunea circuitului. Deci dacă am lua un microprocesor de astăzi şi l-am reduce la scară, sîrmele nu ar constitui un impediment în funcţionarea sa corectă.

Problema apare însă din faptul că de fapt suprafaţa circuitelor nu scade, din cauză ca designerii adaugă noi module. O mare problemă sunt sîrmele care *traversează mai multe module*. Lungimea acestora rămîne practic constantă, în milimetri. Ori, cum viteza ceasului creşte mereu, asta înseamnă că semnalele electrice *nu mai au timp* să parcurgă sîrmele de la un capăt la altul. La 1Ghz lumina în vid străbate $3 * 10^8 * 10^{-9}$m = 30cm. Dar viteza luminii în solide este mai mică, iar viteza de propagare scade dramatic cu numărul numărul de "consumatori" ai sîrmei (adică o sîrmă conectată la 3 circuite e mult mai lentă decît una cuplată la doar două). De asemenea, liniile lungi de transmisiune vor avea nevoie de amplificatoare, care încetinesc substanţial semnalul.

Asta înseamnă că circuitele viitorului *nu vor mai putea comunica prin semnale globale*: pur şi simplu va fi imposibil pentru o sîrmă să unească diferitele părţi ale circuitului. Aceasta este o consecinţă de cea mai mare importanţă pentru arhitecturile viitoare!

Zgomotul În fine, pe măsură ce tranzistorii sunt mai mici, sîrmele sunt mai subţiri şi consumul de putere este mai mare circuitele sunt mai sensibile la zgomot, fie el termic, din mediu (ex. radiaţii cosmice) sau, în curînd, chiar efecte cuantice! Fenomenele de transport din semiconductori pe care se bazează tranzistorul sunt fenomene statistice: ori cînd dimensiunile devin atît de mici încît numai cîţiva electroni produc semnalele, statistica nu mai operează, şi excepţiile încep să apară.

4 Viitorul

În această ultimă secţiune vom încerca să discernem ce ne oferă viitorul. Marile companii lucrează simultan la mai multe generaţii ale unui procesor, cu echipe independente, ca atare ceva din ceea ce viitorul ne rezervă poate fi observat în produsele comerciale în curs de proiectare, în măsura în care companiile dezvăluie astfel de informaţii.

4.1 Generaţia următoare

Cercetarea în arhitectura procesoarelor este efervescentă, atît în industrie cît şi în universităţi; este absolut imposibil de urmărit întregul peisaj. Iată însă unele dintre direcţii:

4.1.1 Evoluţie incrementală

Un efort substanţial este în continuare depus în a perfecţiona tehnicile care în ultimii 15 ani au servit atît de bine arhitectura, pe care le-am descris mai sus: exploatarea paralelismului la nivel de instrucţiune şi ierarhiile de memorie.

Iată unele dintre tendinţe:

- Trace cache: este un cache pentru instrucţiuni care, în loc de a păstra instrucţiunile în ordinea adreselor lor, le menţine în ordinea în care este probabil să fie executate.

- Execuţia speculativă şi predicţia valorilor: într-un articol anterior din PC Report (din iulie 1999) am văzut că principala limitare în calea execuţiei paralele a instrucţiunilor sunt *dependenţele* între instrucţiuni: una are nevoie de rezultatul alteia pentru a se executa. Ori dacă prima instrucţiune durează mult, atunci a doua nu se poate executa nici ea. Soluţia ades folosită este de a *ghici* valoarea rezultată, şi de a executa şi instrucţiunea dependentă. Cînd rezultatul primei instrucţiuni soseşte este comparat cu cel ghicit (prezis); dacă predicţia a fost corectă, toate sunt bune, altfel instrucţiunea dependentă este re-executată.

  Există felurite forme de predicţie a valorilor, unele folosite deja de multă vreme (cum ar fi predicţia salturilor, pe care am descris-o în PC Report din august 1999), dar este plauzibil ca scheme din ce în ce mai sofisticate să-şi facă apariţia.

- Execuţia predicată, care este deja folosită de procesoare de prelucrare de semnal ca Texas Instruments C6X, şi care va fi una din trăsăturile fundamentale ale noii arhitecturi de la Intel, Merced. Execuţia predicată evită execuţia instrucţiunilor de salt (care au un efect negativ asupra performanţei, aşa cum am arătat în articolul mai sus-citat) şi preferă să execute instrucţiuni în mod inutil după care să arunce rezultatele la gunoi (de exemplu, cînd avem o structură de genul `if-then-else`, o arhitectură predicată poate executa ambele ramuri ale condiţiei, dar va permite numai uneia dintre ele să-şi facă efectele vizibile).

4.1.2 Multi-procesoare

Aşa cum am văzut, proiectanţii tind să înghesuie din ce în ce mai multe circuite pe aceeaşi pilulă de siliciu. O evoluţie naturală este de a face saltul de la mai multe procesoare legate printr-o magistrală comună (ca în cazul sistemelor cu multiprocesoare simetrice, pe care le-am descris în PC Report din noiembrie 1998) în procesoare strîns cuplate, pe aceeaşi pilulă de siliciu.

De fapt astfel de scheme există deja: procesorul pentru mainframes de la IBM S/390 are două nuclee identice, care execută sincron acelaşi program: în cazul în care rezultatele nu sunt identice se execută o excepţie şi programul este reluat. Acesta este un exemplu în care mai multe resurse sunt folosite pentru o fiabilitate sporită, dar IBM a anunţat că viitorul lor procesor G5 va conţine două nuclee independente pe aceeaşi pilulă, permiţînd realizarea unor sisteme multi-procesor cu un singur cip.

4.1.3 Multithreading

O evoluţie naturală ar fi la a exploata alte forme de paralelism decît cel la nivel de instrucţiune (ILP). Calculatoarele moderne exploatează excelent paralelismul la nivel de proces, dar există forme intermediare, şi trebuie să ne aşteptăm să vedem arhitecturi din ce în ce mai orientate spre acestea:

- Paralelism la nivel de buclă: în care iteraţii succesive ale unei bucle sunt executate în paralel.
- Paralelism la nivel de thread; despre multithreading am scris un articol amplu în PC Report din ianuarie 1997.

Există o sumă de inovaţii arhitecturale legate de aceste tehnologii, încă ne-integrate în produse comerciale. Să privim cîteva dintre ele:

- Thread-level data speculation: este o metodă de a implementa paralelismul la nivel de buclă lansînd cîte un thread pentru fiecare iteraţie a buclei. De exemplu, proiectul STAMPede de la Universitatea Carnegie Mellon, condus de profesorul Todd Mowry explorează această alternativă (`http://www.cs.cmu.edu/~tcm/STAMPede.html`).

- Simultaneous multi-threading, propus la universitatea Washington din Seattle în 1995. Această tehnologie menţine starea fiecărui thread în hardware şi permite comutarea rapidă între thread-uri. Putem distinge două variante, ca în figura 4: într-una din variante în fiecare ciclu de ceas putem executa instrucţiuni dintr-un alt thread, şi alta, în care în fiecare ciclu, instrucţiuni din thread-uri diferite candidează pentru unităţi funcţionale diferite.

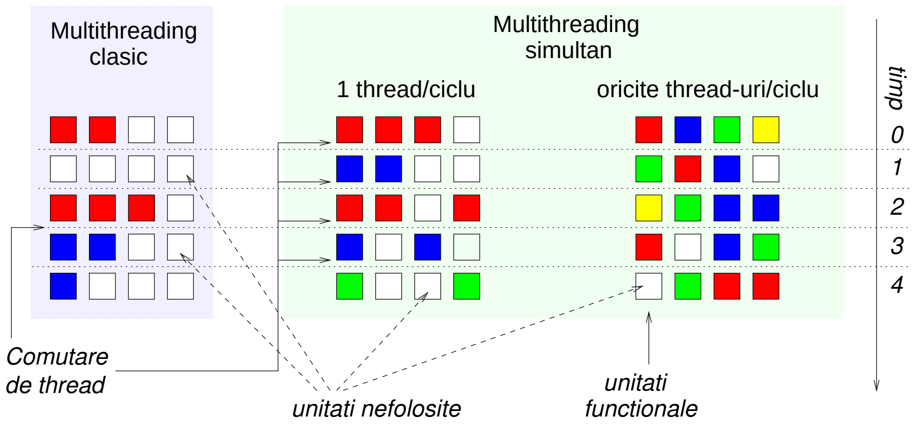

Figura 4: Multithreading clasic şi multithreading simultan. Fiecare culoare simbolizează o instrucţiune dintr-un thread diferit; alb înseamnă "nici o instrucţiune". Schemele clasice de multithreading comută între execuţia a două thread-uri numai arar. În multithreading simultan comutarea este ajutată substanţial de hardware, şi se poate face chiar la fiecare ciclu de ceas (centru), sau chiar instrucţiuni din thread-uri diferite se pot executa simultan (dreapta).

Avantajul unor astfel de scheme este că, dacă un thread execută care au nevoie de mult timp (de exemplu accese la memorie), putem executa alte thread-uri care sunt gata de execuţie, folosind mai eficient unităţile funcţionale ale procesorului.

4.2 Dincolo de legea lui Moore

În fine, voi încheia acest articol cu o privire extrem de sumară asupra unor proiecte de cercetare ambiţioase, care încearcă să privească nu numai în viitorul imediat, ci să anticipeze peisajul calculatoarelor peste zece ani şi mai mult. La acea dată barierele impuse de fizică vor fi atinse, aşa că trebuie să ne aşteptăm la o încetinire a vertiginoasei creşteri de performanţă. Dar chiar şi aşa, numărul uriaş de resurse puse la dispoziţie trebuie să fie folosit cumva; iată unele posibilităţi:

4.2.1 IRAM, Smart Memory

Proiectul IRAM (Intelligent RAM) de la Berkeley este condus de David Patterson (`http://iram.cs.berkeley.edu`) şi îşi propune integrarea tehnologiilor de fabricaţie a memoriilor şi procesoarelor (la ora actuală liniile de fabricaţie sunt complet diferite). IRAM încearcă să evite disparitatea de acces la memorii împingînd o multitudine de procesoare micuţe printre celulele de memorie, unde pot lucra independent.

Un proiect foarte asemănător este cel de la Stanford, al profesorului Mark Horowitz, numit Smart Memories (`http://velox.stanford.edu/smart_memories/`).

Aceste proiecte încearcă să depăşească problema accesului lent la memorie prin distribuirea unităţilor de procesare printre memorii, astfel încît accesul să fie paralel şi rapid. De asemenea, asemenea scheme, în care calculul este distribuit în multe unităţi independente, face ca impactul "sîrmelor lungi" să fie redus.

4.2.2 RAW

Proiectul RAW de la MIT (`http://www.cag.lcs.mit.edu/raw/`) atacă problema dintr-un cu totul alt punct de vedere: maşina RAW constă din foarte multe procesoare relativ simple construite pe aceeaşi pilulă de siliciu. Aceste procesoare cooperează pentru a executa o singură aplicaţie, care este paralelizată automat de compilator.

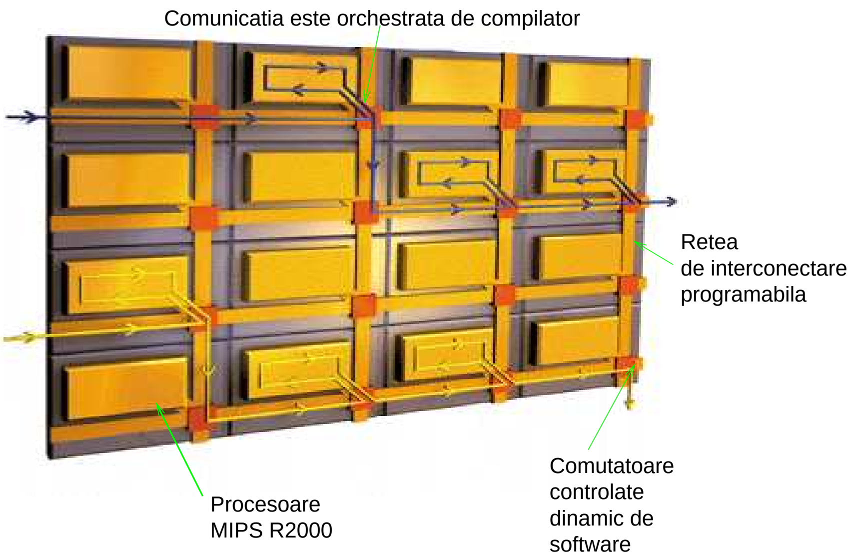

Figura 5: Schema maşinii RAW de la MIT: foarte multe procesoare simple pe aceeaşi pastilă de siliciu, legate printr-o reţea de interconectare rapidă comutată, aflată sub controlul software-ului. Comunicaţia este planificată cît mai ades posibil de compilator.

4.2.3 Imagine

Un proiect foarte interesant este Imagine (`http://cva.stanford.edu/imagine/cva_imagine.html`), la universitatea Stanford este orientat mai curînd pe procesarea de semnal; conducătorul este William Dally. Proiectul acesta propune un nou model de programare, orientat spre multimedia, în care paralelismul datelor este făcut explicit prin noţiunea de *flux* (stream). De exemplu, pentru a afişa scene complicate pe ecran, prelucrarea transformă datele dintr-un flux de obiecte într-un flux de poligoane, care devin un flux de triunghiuri, apoi un flux de pixeli, etc.

4.2.4 PipeRench şi Brass

În fine, menţionez două proiecte care încearcă să îmbine hardware-ul reconfigurabil cu procesoarele tradiţionale: proiectul Brass de la Berkeley (`http://http.cs.berkeley.edu/projects/brass/`), condus de John Wawrzynek, şi proiectul PipeRench de la Carnegie Mellon, condus de Seth Goldstein şi Herman Schmit (`http://www.ece.cmu.edu/research/piperench/`).

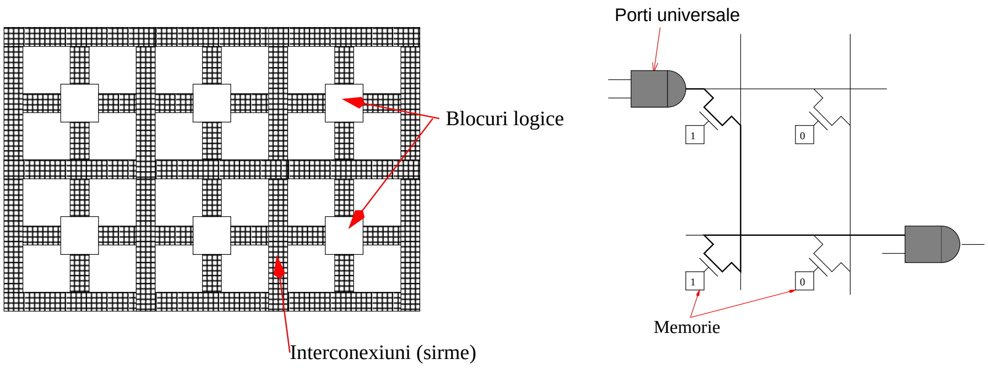

Figura 6: Hardware-ul reconfigurabil constă din porţi universale, care pot implementa orice funcţie, care sunt legate între ele prin sîrme cuplate prin comutatoare. Figura din dreapta arată un detaliu de implementare despre figura din stînga.

Am vorbit altă dată pe larg despre hardware-ul reconfigurabil (figura 6) (PC Report din iulie 1998). Fiecare poartă universală poate fi programată să execute orice funcţie logică, iar comutatoarele pot cupla şi decupla sîrmele. Porţile universale se pot implementa din mici celule RAM. Astfel se pot sintetiza unităţi funcţionale extrem de complexe, care pot opera uneori mult mai eficient decît un procesor de uz general.

5 Concluzii

Am văzut în acest articol că performanţa microprocesoarelor s-a situat pe o curbă exponenţială în toţi cei treizeci de ani de la crearea lor. Am văzut că miniaturizarea şi tehnici de design contribuie în mod egal la aceste spectaculoase creşteri. De asemenea, am văzut că creşterea aceasta se apropie de sfîrşit, datorită unor bariere fizice fundamentale. În fine, am încercat să profeţim unele din tehnologiile care-şi vor face apariţia în generaţiile următoare de procesoare.