Quantum computing, System and IC design

Francky Catthoor over computerarchitecturen

“Rekenen in het computergeheugen past in een nieuwe manier van schalen waarmee we nog erg veel energiewinst kunnen boeken”

Francky Catthoor, imec fellow

Veel toekomstige computertoepassingen zullen de rekenkracht nodig hebben die typisch was voor de supercomputers van een paar jaar geleden. Maar we willen die kracht wel graag in kleine, onopvallende apparaatjes gaan stoppen, zoals bijvoorbeeld slimme IoT-sensoren die zeer weinig elektriciteit verbruiken en onmiddellijk reageren. Met de traditionele manier van het schalen van chips – schalen op de laagste niveaus van transistoren, cellen en kleine circuits zoals we al een paar decennia doen – kunnen we nog iets winnen, maar niet genoeg. Dus moeten we naar hogere niveaus gaan kijken, en nieuwe technologieën bedenken die de prestaties en het energieverbruik optimaliseren van specifieke functies of toepassingen.

Die systeem-technologie co-optimalisatie (STCO) is een nog grotendeels onontgonnen terrein. Het zal ons toelaten om veel van het energieverlies terug te winnen dat de voorbije decennia verloren is gegaan doordat de industrie enkel oog had voor het geometrisch schalen, het steeds kleiner maken van onderdelen op een voor de rest gestandaardiseerde chiparchitectuur. Er wordt geschat dat STCO ons voor specifieke rekenopdrachten een energiewinst kan opleveren van meerdere grootteordes. Maar door de complexiteit en de specificiteit van individuele oplossingen kan het nog een hele tijd duren voor we dat potentieel volledig zullen kunnen benutten.

Te veel verkeer

Voor veel moderne toepassingen zit het grootste energieverlies in het heen- en weerschuiven met data tussen de processor en de verschillende niveaus van geheugens. Dat is in het bijzonder waar voor de toepassingen die werken op enorme datasets, zoals de resultaten van DNA-analyses, de verbindingen in een socialemedia-netwerk of de resultaten van digitale hogeresolutie-camera’s. Voor zulke systemen is de energiekost van het verhuizen van de data vele keren groter dan de eigenlijke berekeningen. Bovendien zorgt het in die systemen ook voor vertragingen, en dat op soms erg onvoorspelbare manieren.

Een mogelijk pad om dat energieverlies te verminderen is om chips te ontwerpen waar een deel van de bewerkingen gebeurt op dezelfde fysische locatie als waar de data zijn opgeslagen, dus zonder deze te verplaatsen. Denk bijvoorbeeld aan logische operaties zijn op grote matrices, foutcorrectie op gegeven van draadloze sensoren, of preprocessing op de ruwe data van beeldsensoren.
De techniek wordt 'Computation in Memory' of CIM genoemd. In onderzoeksgroepen werd het al een tijdje bestudeerd, maar de tijd is nu rijp om het aan de praktijk te gaan toetsen.

Met twee scherp van elkaar onderscheiden stroomniveaus wordt een resistief geheugen een gewoon digitaal geheugen. Maar de werkelijke kracht zit hem in het feit dat de weerstand van zo'n geheugenelement vrij kan variëren in functie van de huidige stimulatie en de vorige waarde. Die eigenschap kunnen we gebruiken om te rekenen met alternatieve, niet booleaanse rekenparadigma's.

Geheugens worden functies

Dit jaar heeft imec CIM als potentieel beloftevolle technologie voorgesteld aan zijn partners. Daarbij hebben we ook een nieuwe classificatie van alle mogelijke opties gemaakt. De voorstelling past in onze inspanningen om de klassieke chipschaling meer en meer te gaan aanvullen met DTCO (design-technologie co-optimalisatie) en nog een niveau hoger, met STCO. Het concept werd enthousiast onthaald, in het bijzonder het idee om systemen uit te rusten met specifieke co-processoren om gespecialiseerde taken in het geheugen uit te voeren, de zogenaamde computation in memory accelerators (CIMA).

Verschillende academische groepen hebben al voorstellen voor CIMA-architecturen op tafel gelegd. Maar er zijn grote verschillen in de toepassingen waarvoor ze geschikt zijn, hun positie in de geheugenstructuur, en de manier waarop parallelle berekeningen worden uitgevoerd.

Als we bijvoorbeeld kijken naar waar in de hiërarchie van geheugens de CIM best wordt geplaatst, dan zijn er een paar fundamentele opties. De eerste in om de berekeningen dicht bij het perifere geheugen te doen, bijvoorbeeld geïntegreerd in een hybride geheugenkubus (hybrid memory cube - HMC) of een geheugen met grote bandbreedte (high-bandwidth memory - HBM). Als tweede optie kunnen de berekeningen ook geïmplementeerd worden in de geheugenstack, maar nog altijd buiten de geheugenbanken van de processor, zodat de resultaten door verschillende processoren kunnen gedeeld worden. En tenslotte kan de CIM ook opgezet worden op de geheugenbanken van de hoofdprocessor zelf. Die laatste optie is de meest bestudeerde.

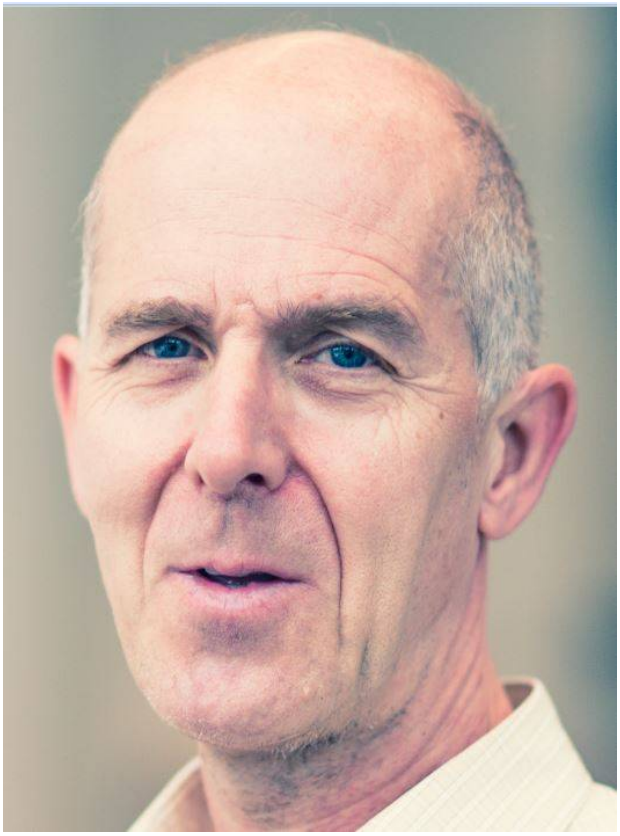
Een andere dimensie wordt gevormd door de manier waarop de parallelle verwerking is georganiseerd. De standaard techniek is om te parallelliseren op het niveau van taken, waarbij elke taak de data heen en weer sluist die het nodig heeft. Maar als we de data in het geheugen houden, of dicht bij het geheugen, dan kunnen we ofwel een parallelle oplossing op het niveau van de gegevens ofwel op het niveau van instructies implementeren. In het eerste geval zullen data (bv. de rijen van een beeld) in parallel door dezelfde functie worden bewerkt (bv. foutcorrectie). In het tweede geval zullen dezelfde gegevens in parallel door verschillende instructies worden bewerkt.

Technologie bouwen en toepassingen kiezen

Nadat imec CIM aan zijn partners heeft voorgesteld, zal het de verschillende voorstellen nu verder uitwerken om onder andere te kijken hoe ze technologisch kunnen worden uitgevoerd, op een manier die past in de plannen en oplossingen van de chipindustrie. De bedoeling is daarbij om microarchitecturen voor CIMA's te ontwikkelen en te demonstreren op geschikte toepassingen, waarbij we alle opties en combinaties van architecturen, circuits, en technologie gaan exploreren.

In dat kader starten we in 2018 met het project MNEMOSENE (computation-in-memory architecture based on resistive devices), een Europees project in het Horizon 2020 programma. De bedoeling is om CIM te ontwikkelen en te demonstreren met resistieve geheugens en voor specifieke toepassingen. Daarbij hoort ook een CIM-simulator gebaseerd op modellen van memristor transistoren en bouwblokken. MNEMOSENE wordt gecoördineerd door de TU Delft, waarmee imec een lange en vruchtbare samenwerking heeft. Samen vormen de partners een consortium dat dit domein een echte boost kan geven.

Rekenen in het computergeheugen past in een nieuwe manier van schalen waarmee we nog erg veel energiewinst kunnen boeken. Dat kan door functies en toepassingen te isoleren en te zien hoe we daarvoor een geoptimaliseerde architectuur kunnen ontwerpen. CIM is een van die architecturen, naast bv. neuromorfische of golfgebaseerde oplossingen of quantumrekenen. Hoe hoger in de hiërarchie we daarbij kunnen gaan, zelfs tot op het niveau van toepassingen, hoe meer winst we kunnen boeken. Op die manier denken we bv. aan een accelerator voor beeldverwerking of voor de verwerking van de gegevens van een DNA-analyse, of een accelerator voor encryptie en beveiliging. Die acceleratoren zullen dan als legoblokken in elkaar worden geplugd, met als resultaat systemen die veel heterogener zijn dan wat we vandaag hebben.



Biografie Francky Catthoor

Francky Catthoor is imec fellow en professor aan het departement elektronica van de KU Leuven. Hij behaalde zijn PhD aan de KU Leuven in 1987 en begon toen bij imec te werken als hoofd van het onderzoek naar systeemsynthese-technieken en architectuur-methodologieën. Sinds 2000 is hij ook betrokken bij het onderzoek naar diep-submicron technologie, biomedische beeldvorming, sensoren, en slimme zonnepanelen. Francky Catthoor was ook associate editor voor meerdere IEEE en ACM publicaties, zoals Transactions on VLSI Signal Processing, Transactions on Multimedia, en ACM TODAES. Hij was programmadirecteur van conferenties zoals ISSS'97 en SIPS'01 en werd in 2005 verkozen tot IEEE fellow.